Cape Peninsula
University of Technology

LEVERAGING BIG DATA FOR COMPETITIVE ADVANTAGE IN A MEDIA ORGANISATION

by

Cecil Kabu Nartey

Thesis submitted in fulfilment of the requirements for the degree

Master of Technology: Information Technology

In the Faculty of Informatics and Design

at the Cape Peninsula University of Technology

Supervisor: Dr Andre de la Harpe
Co-supervisor: Mr Cecil Van der Watt

Cape Town

Date submitted: March 2015

DECLARATION

I, Cecil Kabu Nartey, declare that the contents of this dissertation represent my own unaided work, and that the dissertation has not previously been submitted for academic examination towards any qualification. Furthermore, it represents my own opinions and not necessarily those of the Cape Peninsula University of Technology.

Signed **Date**

ABSTRACT

Data sources often emerge with the potential to transform, drive and allow deriving never-envisaged business value. These data sources change the way business enacts and models value generation. As a result, sellers are compelled to capture value by collecting data about business elements that drive change. Some of these elements, such as the customer and products, generate data as part of transactions which necessitates placement of the business element at the centre of the organisation's data curation journey. This is in order to reveal changes and how these elements affect the business model. Data in business represents information translated into a format convenient for transfer. Data holds the relevant markers needed to measure business elements and provide the relevant metrics to monitor, steer and forecast business to attain enterprise goals. Data forms the building blocks of information within an organisation, allowing for knowledge and facts to be obtained. At its lowest level of abstraction, it provides a platform from which insights and knowledge can be derived as a direct extract for business decision-making as these decisions steer business into profitable situations. Because of this, organisations have had to adapt or change their business models to derive business value for sustainability, profitability and transformation.

An organisation's business model reflects a conceptual representation on how the organisation obtains and delivers value to prospective customers (the service beneficiary). In the process of delivering value to the service beneficiaries, data is generated. Generated data leads to business knowledge which can be leveraged to re-engineer the business model. The business model dictates which information and technology assets are needed for a balanced, profitable and optimised operation. The information assets represent value holding documented facts. Information assets go hand in hand with technology assets. The technology assets within an organisation are the technologies (computers, communications and databases) that support the automation of well-defined tasks as the organisation seeks to remain relevant to its clientele. What has become apparent is the fact that companies find it difficult to leverage the opportunities that data, and for that matter Big Data (BD), offers them. A data curation journey enables a seller to strategise and collect insightful data to influence how business may be conducted in a sustainable and profitable way while positioning the curating firm in a state of 'information advantage'.

While much of the discussion surrounding the concept of BD has focused on programming models (such as *Hadoop*) and technology innovations usually referred to as disruptive technologies (such as *The Internet of Things* and *Automation of Knowledge Work*), the real driver of technology and business is BD economics, which is the combination of open source

data management and advanced analytics software coupled with commodity-based, scale-out architectures which are comparatively cheaper than prevalent sustainable technologies known to industry. Hadoop, though hugely misconstrued, is not an integration platform; it is a model the helps determine data value while it brings on-board an optimised way of curating data cheaply as part of the integration architecture.

The objectives of the study were to explore how BD can be used to utilise the opportunities it offers the organisation, such as leveraging insights to enable business for transformation. This is accomplished by assessing the level of BD integration with the business model using the BD Business Model Maturation Index. Guidelines with subsequent recommendations are proposed for curation procedures aimed at improving the curation process. A qualitative research methodology was adopted. The research design outlines the research as a single case study; it outlines the philosophy as interpretivist, the approach as data collection through interviews, and the strategy as a review of the method of analysis deployed in the study.

Themes that emerged from categorised data indicate the diverging of business elements into primary business elements and secondary supporting business elements. Furthermore, results show that data curation still hinges firmly on traditional data curation processes which diminish the benefits associated with BD curation. Results suggest a guided data curation process optimised by persistence hybridisation as an enabler to gain information advantage. The research also evaluated the level of integration of BD into the case business model to extrapolate results leading to guidelines and recommendations for BD curation.

Keywords: Big Data, data curation, business model, competitive advantage, data monetisation, polyglot persistence.

ACKNOWLEDGEMENTS

I wish to thank:

- Dr Andre de la Harpe for his invaluable support
- Cecil van der Watt for his great contribution towards this work
- Andre, Stephan and Johan of Media24 corporate and the entire staff
- Nick, Sheldon, Allan and the entire staff of Spree for the interviews and their time
- Lashwin Naidoo and the DB teams at Tag Worldwide in South Africa and India for their support

The financial assistance of the Cape Peninsula University of Technology (CPUT) towards this research is acknowledged. Opinions expressed in this thesis and the conclusions arrived at, are those of the author, and are not necessarily to be attributed to CPUT.

DEDICATION

To my brother and friend, matured (RIP)
and my mother and sister

For (whomever)

TABLE OF CONTENTS

DECLARATION ..... **i**
ABSTRACT ..... **ii**
ACKNOWLEDGEMENTS ..... **iv**
DEDICATION ..... **v**
LIST OF FIGURES ..... **xi**
LIST OF TABLES ..... **xiv**
GLOSSARY ..... **xvi**
CHAPTER ONE: INTRODUCTION TO THE STUDY **1**
1.1 Introduction 1
1.2 Research context 4
1.3 The research problem 7
1.4 Aim 8
1.5 Research questions and sub-questions 9
1.5.1 Research questions 9
1.5.2 Research sub-questions 9
1.6 Objectives of the Study 10
1.7 Research methodology 11
1.7.1 Research philosophy 12
1.7.2 Research approach 14
1.7.3 Research strategy 15
1.8 Introduction to status of the research area 21
1.8.1 Traditional data curation 23
1.8.2 NoSQL databases and Big Data 24
1.8.3 Traditional data flow in organisations 25
1.8.4 Competitive advantage with Big Data 26
1.9 Delineation of the research 27
CHAPTER TWO: LITERATURE REVIEW **30**
2.1 Introduction 30
2.2 Databases 32
2.3 The new industrial asset 36
2.4 E-commerce 45
2.5 E-commerce concepts in retail 47
2.6 Business models 48
2.6.1 Competition and strategy 51
2.6.2 E-commerce revenue models 52
2.6.3 Internet marketing models 53

2.7 Big Data value generation 55
2.7.1 Structured data 58
2.7.2 Unstructured data 58
2.7.3 Data velocity 58
2.7.4 Predictive analytics 59
2.8 Traditional e-commerce data flow 62
2.9 Contextualising Big Data 65
2.9.1 Ethics: Big Data privacy 67
2.10 Dealing with the data deluge 69
2.10.1 Operationalising Big Data 72
2.10.2 Data models in business models 73
2.10.3 Challenges and trends in favour of NoSQL 75
2.10.4 NoSQL databases 78
2.10.5 Categorising NoSQL databases 79
2.10.6 Hybridisation of persistence platforms 82
2.11 Vending Big Data 85
2.12 Big Data and NoSQL implementations 89
2.13 Solution Engineering 91
2.13.1 Big Data journey 94
2.13.2 Big Data strategy 96
2.13.3 Decision Theory for competition 97
2.13.4 Leveraging weblog 98
2.14 Polyglot persistence 104
2.15 Summary 106
CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY 108
3.1 Introduction 108
3.2 Research design 109
3.3 Research philosophy 112
3.3.1 Research philosophy 112
3.3.2 Ontology 113
3.3.3 Epistemology 114
3.3.4 Interpretive and critical approach 115
3.4 Research strategy 116
3.5 Research methods 117
3.5.1 Qualitative and quantitative research 118
3.6 Units of analysis 121
3.7 Data collection 122
3.7.1 Primary data 122

3.7.2 Secondary data 122
3.7.3 Interviews 122
3.8 Data analysis 126
3.8.1 Hermeneutics 127
3.8.2 Conversation analysis 127
3.8.3 Content analysis 128
3.9 Ethical considerations 128
3.10 Research contribution 129
3.11 Summary 130
CHAPTER FOUR: RESEARCH FINDINGS 131
4.1 Introduction 131
4.2 The case 132
4.2.1 Departments 134
4.3 Case summary 140
4.4 Interviews 141
4.4.1 Pre-interviews 142
4.4.2 Business interviews 147
4.4.3 Technical Interviews 158
4.5 Themes development 166
4.5.1 General themes discussion 166
4.5.2 Themes 170
4.6 Headline findings by department 173
4.6.1 Findings: Business Management department 173
4.6.2 Important findings 174
4.6.3 Findings: Technical development team 174
4.6.4 Findings: Merchandising department 174
4.6.5 Findings: Marketing department 175
4.6.6 Findings: Social Media and Marketing department 175
4.6.7 Findings: Business Intelligence department 176
4.6.8 Findings: Supply Chain and Customer Support department 177
4.7 Summary 178
CHAPTER FIVE: DISCUSSION 180
5.1 Introduction 180
5.2 Themes developed 181
5.2.1 Data as an asset 181
5.2.2 Profiling the customer 185
5.2.3 Developing and satisfying customer needs 186
5.2.4 Planning 186

5.2.5 Decision-making ... 187
5.2.6 Competitive advantage ... 188
5.2.7 Marketing optimisation ... 189
5.2.8 Service delivery ... 189
5.2.9 Sales management ... 190
5.2.10 Analytics of data ... 191
5.2.11 Building strategies ... 192
5.2.12 Transforming the business model ... 193
5.3 Headline findings: Department level discussion ... 194
5.3.1 Headline Findings: Business Management ... 194
5.3.2 Headline Findings: Merchandising ... 195
5.3.3 Headline findings: Social Media and Marketing ... 196
5.3.4 Headline findings: Supply Chain and Customer Support department ... 197
5.3.5 Headline Findings: Business Intelligence ... 199
5.3.6 Headline Findings: Technical development ... 200
5.4 Answering the research questions ... 201
5.4.1 Answering Research Question One ... 201
5.4.2 Answering Research Question Two ... 205
5.5 Data curation guidelines ... 207
5.6 Summary ... 214
CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS ... 217
6.1 Introduction ... 217
6.2 Conclusion ... 218
6.3 Recommendations ... 222
6.4 Future research ... 227
6.4.1 Polyglot persistence and Business Intelligence ... 227
6.4.2 Big Data and fraud ... 228
6.4.3 Increase site traffic does not warrant sales improvement ... 228
6.4.4 New scalable software architectures for challenges ... 228
6.4.5 Evaluating the effectiveness of growth hacking techniques in marketing ... 229
6.4.6 Big Data monetisation as part of data supply chain ... 229
6.4.7 Big Data service refinery ... 230
6.4.8 Business model, dig data engineering and framework ... 230
6.5 Reflection ... 231
6.6 Summary ... 232
REFERENCES ... 233
ANNEXURE A: Interview Findings ... - 1 -
ANNEXURE B: Data Analysis ... - 58 -

ANNEXURE C: E-commerce Retail Concepts ..... - 64 -
ANNEXURE D: Letter of Consent for Interview ..... - 65 -

LIST OF FIGURES

Figure 1.1: Chapter One layout - Introduction 1
Figure 1.2: Chapter layout – Research context 4
Figure 1.3: Chapter layout – The research problem 7
Figure 1.4: Chapter layout – Research methodology 11
Figure 1.5: Philosophical paradigms 21
Figure 1.6: Traditional data curation architecture 26
Figure 1.7: Chapter layout – Delineation of the research 27
Figure 1.8: Chapter layout – Contribution of the research 28
Figure 2.1: Chapter Two layout – Literature review 30
Figure 2.2: Chapter layout - Databases 32
Figure 2.3: Evolution of Big Data 34
Figure 2.4: Chapter layout – The new industrial soil 36
Figure 2.5: Four Vs of Big Data 37
Figure 2.6: Performance against data complexity of database 39
Figure 2.7: Chapter layout – E-commerce 45
Figure 2.8: Chapter layout – Business model 48
Figure 2.9: Business model components 50
Figure 2.10: Chapter layout – Big Data value generation 55
Figure 2.11: Commerce view of the customer 57
Figure 2.12: Big Data journey roadmap 60
Figure 2.13: Heat map of hotel home page prototype 61
Figure 2.14: Chapter layout – Traditional e-commerce data flow 62
Figure 2.15: Chapter layout – Big Data in context 65
Figure 2.16: Chapter layout – Dealing with the data deluge 69
Figure 2.17: The Big Data framework 70
Figure 2.18: Dimensions of Big Data insight generation 72
Figure 2.19: Database scaling to size against complexity 78
Figure 2.20: Graphical representation of columns store database 80
Figure 2.21: A Big Data ecosystem with Hadoop 84
Figure 2.22: Chapter layout – Vending Big Data 85
Figure 2.23: Chapter layout – Solution Engineering 91
Figure 2.24: Solution Engineering phases 93
Figure 2.25: Big Data Business Model Maturation Index 95
Figure 2.26: Big Data strategy document 97
Figure 2.27: Data pyramid depicting information worth 101

Figure 2.28: Chapter layout – Polyglot persistence 104
Figure 2.29: Diggs polyglot persistence conceptual architecture 105
Figure 2.30: Chapter layout - Summary 106
Figure 3.1: Chapter Three layout - Research design and methodology 108
Figure 3.2: Chapter layout – Research design 109
Figure 3.3: The research onion 111
Figure 3.4: Research philosophy 112
Figure 3.5: Chapter layout - Research strategy 116
Figure 3.6: Chapter layout - Research methods 117
Figure 3.7: Chapter layout – Units of analysis 121
Figure 3.8: Types of interviews 123
Figure 3.9: Chapter layout – Data analysis 126
Figure 3.10: Chapter layout – Research contribution 129
Figure 4.1: Chapter Four layout – Research findings 131
Figure 4.2: Chapter layout – The Case 132
Figure 4.3: Naspers company structure 133
Figure 4.4: Chapter layout – Summary of business overview 140
Figure 4.5: Chapter layout - Interviews 141
Figure 4.6: Interview diagram 142
Figure 4.7: Chapter layout – Themes development 166
Figure 4.8: Chapter layout – Headline findings by department 173
Figure 4.9: Chapter layout - Summary 178
Figure 5.1: Chapter Five layout - Discussion 180
Figure 5.2: Chapter layout – Themes development 181
Figure 5.3: Chapter layout – Themes discussion 194
Figure 5.4: Chapter layout – Answering the research questions 201
Figure 5.5: Chapter layout – Data curation guidelines 207
Figure 5.6: Business model canvas 209
Figure 5.7: Enterprise-wide data curation guidelines 211
Figure 5.8: Forceful value proposition 212
Figure 5.9: Ingest and evaluate data against goals 213
Figure 5.10: Information assets of the organisation 213
Figure 5.11: Technology assets of the organisation 214
Figure 5.12: Chapter layout – Summary overview of discussion 214
Figure 6.1: Chapter Six layout – Conclusion and recommendations 217
Figure 6.2: Chapter layout - Conclusion 218
Figure 6.3: Chapter layout - Recommendations 222
Figure 6.4: Chapter layout – Future research 227

Figure 6.5: Conceptual view of a polyglot persistence layer 228
Figure 6.6: Chapter layout – Reflection 231
Figure 6.7: Chapter layout – Summary 232

LIST OF TABLES

Table 1.1: Research questions, sub-questions and objectives 10
Table 1.2: Philosophical paradigms 12
Table 2.1: More definitions of Big Data 40
Table 2.2: Internet business models 51
Table 2.3: Revenue models for the online market space 53
Table 2.4: Summary of trends in favour of NoSQL 75
Table 2.5: Comparison of a relational database to a NoSQL database 76
Table 2.6: Categories of NoSQL databases 79
Table 2.7: Hadoop supporting toolset 84
Table 2.8: Phases of Big Data business Maturation Index 95
Table 2.9: Components of Decision Theory 98
Table 3.1: Categories developed from coding 115
Table 3.2: Qualitative research techniques 120
Table 3.3: Guidelines for qualitative research 124
Table 3.4: Pre-interview question 125
Table 3.5: Sample business interview questionnaire for participant 12 125
Table 3.6: Sample business interview questionnaire for participant 2 125
Table 4.1: Functional departments at Spree 134
Table 4.2: Different data required at department levels 137
Table 4.3: Magento Big Data extensions 139
Table 4.4: Research questions 143
Table 4.5: Business interviews 146
Table 4.6: Technical interviews 147
Table 5.1: Summary of Business Management headline findings 195
Table 5.2: Summary of Merchandising headline findings 196
Table 5.3: Summary of Social Media and Marketing headline findings 197
Table 5.4: Summary of Supply Chain and Customer Support headline findings 199
Table 5.5: Summary of Business Intelligence headline findings 200
Table 5.6: Summary of Technical department headline findings 200
Table 5.7: Summary of answers to Research Question One sub-questions 201
Table 5.8: Summary of answers to Research Question Two sub-questions 205
Table 6.1: Research Question One sub-questions and findings 220
Table 6.2: Research Question Two sub-questions and findings 221
Table 6.3: Big Data curation guidelines 223
Table 7.1: Pre-interview findings ..... - 1 -

Table 7.2: Findings of business interviews ..... - 2 -
Table 7.3: Summary of business interview findings ..... - 6 -
Table 7.4: Technical participant response indicating unclear status of Big Data ..... - 8 -
Table 7.5: Findings from technical interviews ..... - 8 -
Table 7.6: Summary of technical interview findings ..... - 11 -
Table 7.7: Research questions and related findings ..... - 13 -
Table 7.8: Diggs polyglot persistence architecture terms and descriptions ..... - 18 -
Table 7.9: Sample Interview Supply Chain and Customer Service division ..... - 18 -
Table 7.10: Sample Interview Business Intelligence ..... - 27 -
Table 7.11: Sample Interview Business Management ..... - 34 -
Table 7.12: Sample Interview Merchandising ..... - 43 -
Table 7.13: Sample Interview Social Media and Marketing ..... - 47 -
Table 7.14: Emerged research themes ..... - 57 -
Table 7.15: Summary of responds to Spree having Big Data ..... - 57 -
Table 8.1: Pre-interview question and summarised responses by department ..... - 58 -
Table 8.2: Themes development ..... - 59 -
Table 8.3: Establish themes frequency ..... - 60 -
Table 8.4: Abridged themes frequency ..... - 61 -
Table 8.5: Data analysis of emerged themes ..... - 62 -
Table 8.6: Pie chart grouping of emerged themes ..... - 63 -
Table 9.1: E-commerce retail concepts ..... - 64 -

GLOSSARY

Terms/Acronyms/ Abbreviations	Definition/Explanation
ACID	ACID stands for atomicity, consistency and durability. Following Boyce-Codd's principles, relational database management systems adopted the ACID approach for data design. In essence, the relational database systems ensure atomicity (a transaction is all or nothing), consistency (only valid data is written to the database), isolation (all transactions are happening serially and the data is correct) and durability (what you write is what you get) (Mohanty, Jagadeesh & Srivatsa, 2013:39).
BI	Business Intelligence as business term: A set of tools and techniques for performing data management, for example extract-transform-load (ETL). Business Intelligence by department: A department responsible for the management of data through its life-cycle of benefits.
BM	Business Management as business term: The act of allocating business resources to accomplish desired goals and objectives. Business Management by department: A decision group responsible for the tactical (day-to-day) running of the organisation such as controlling, monitoring, planning and decision-making. For example, making business decisions leveraging metadata insight.
Business model	Business model reflects how an organisation operates to accomplish its goals (Teece, 2010:2).
Business planning	An intermediate level of strategic planning, typically focusing on the individual strategies that will lead to the broader organisational strategies. It may include the creation of competitive differentiation, minimising of costs or vertical integration (Stubbs, 2014:206).
CAP	CAP stands for consistency, availability, and partition-tolerance. CAP theory came into existence because of several shortcomings that emerged when the ACID principles were extended to large distributed data systems.
Clickstream	The path of pages and links followed by a user during a visit to a web site (Rho, Moon, Kim & Yang, 2004:65).
Competitive advantage	Competitive advantage is obtained when an organisation develops or acquires a set of attributes or executes actions that make it outperform its rivals (Wang, 2014). Competitive advantage is a strategic advantage held by one organisation that cannot be matched by its competitors. This advantage may or may not be sustainable and, if not, may eventually be replicated by its competitors (Stubbs, 2014:207).
Commodity computers	These are off-the-shelf computers readily available for purchase used in parallel computing to get a high level of computation at low cost.

Terms/Acronyms/ Abbreviations	Definition/Explanation
Cross selling	Cross-selling is an old and valuable technique used by salespeople to increase order size and to transform single-product buyers into multi-product ones (Kamakura, 2007:41). A process by which new, non-overlapping products are sold to existing Customers (Stubbs, 2014:207).
Customer	A customer is a person who buys goods or services from a shop or a business.
Data curation	Data curation is the active and continual management of data through its life-cycle of interest and usefulness to enable data discovery and retrieval, quality maintenance, value addition, and provision for reuse over time.
Data monetisation	The process of capturing data from appropriate sources, storing and managing the data, performing analytics to identify key trends and themes, to establish potential insights for consumption by business and customers. The process culminates in selling curated data or analysed data to improve business performance.
Dark transaction data	An organisation's pool of transaction data collected over a protracted period of time.
Data curation journey or data development	Data development is the analysis, design, implementation, deployment and maintenance of data solutions to maximise the value of the data resources to the enterprise (Mosley, Brackett, Earley & Henderson, 2009:87).
Decision-making	Decision-making is the process of choosing what to do by considering the possible consequences of different choices (Brockman & Russel, 2009:1).
Denormalisation	Denormalisation is the process of attempting to optimise the read performance of a database by adding redundant data or by grouping data.
Disruptive technologies	These are technologies that demonstrate a rapid rate of change in capabilities in terms of price and or performance relative to substitutes and alternative approaches. The technologies experience breakthroughs that drive accelerated change rates and discontinuous capability improvements. They ultimately force business model reengineering.
Edge (graph database)	A connection between two vertices (nodes).
Enabling initiative	A business analytics initiative focused on creating processes or assets needed for a planned growth initiative or to deliver evolutionary efficiency improvements (Stubbs, 2014:209).
Field replaceable unit (FRU)	A FRU is an assembly part that can be replaced quickly and easily from a computer system without having to take the system to a repair facility.
Graph	A network of linked vertex and edge objects.

Terms/Acronyms/ Abbreviations	Definition/Explanation
Growth hacking	Growth hacking is a term that describes the use of passion and focus to push a metric through use of a testable and scalable methodology. This approach allows a growth hacker to market and distribute an idea based on an impression of user behaviour the hacker has.
Hadoop	Hadoop distributed file system: the storage layer of Hadoop. It is a distributed, scalable, java-based file system adept at storing larger volumes of unstructured data.
Heat map	A heat map is a pictorial color-coding representation of data where colours are used to represent values contained in a matrix.
Horizontal scaling	The process of increasing the power of a system by simply adding more actors to the system through network/cluster connectivity between systems.
Lean start-up	A method for developing businesses and products, first proposed by Eric Ries. This method adopts validated learnings, business hypothesis-driven experimentation and iterative product releases which aim to sidestep initial large project funding, expensive project launches and mitigate failure (Ries, 2011:103).
Market basket analysis	A study of items purchased together in a single transactional sale or multiple sequential transactions.
Marketspace	An information- and communication-based electronic exchange environment which is an online space that facilitates bi-directional commerce. Sellers can list their goods and buyers can list their needs.
Online marketplace	Online marketplace is a type of e-commerce site where information on product and inventory is provided by multiple third parties where transactions are processed by the marketplace operator. Online marketplaces are the primary type of multichannel for e-commerce.
Overlaps	To have an area or range in common with patterns observation.
Pattern	An intelligible or regular form discernible in the way in which something happens.
Product	A product is an item offered for sale; it may be a service or an item, physical or virtual or in cyber form.
Properties (graph database)	A data item that belongs to a particular edge or node.
Sales	A sale is the exchange of a commodity for monetary value.
Scalability	The ability of a system or process to handle growing amount of work in a capable manner.
Search engine optimisation	Search engine optimisation is the process of affecting the visibility of a website or a web page in a search engine's natural or un-paid (organic) search results.
Seller	A person or company who buys and sells commodities for profit.
Service	A service is a system supplying a public need such as support of a product sale.

Terms/Acronyms/ Abbreviations	Definition/Explanation
Segmentation	A process by which entities within a population are grouped into segments that have common characteristics. This grouping process may be manually, algorithmically or statistically based and will often take into account anywhere from a handful to hundreds of common attributes across all the entities (Stubbs, 2014:217).
Sentiment analysis	The use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.
Structured data	Data that can be organised in a predefined manner (Stubbs, 2014:17). The primary advantage of structured data is its ease of analysis.
SKU	A stock keeping unit (SKU) is a unique representation of a product or service in inventory management offered for sale; it embodies all attributes related to the particular item such as manufacturer reference, product description, size, colour, packaging information, price category and warranty terms.
Sustainable technologies	Sustainable technologies support and improve the performance of mainstream or established products along the dimensions of performance and delivery.
Trends	A direction in which something is turning. It is a discernible pattern of change. A trend is driven by one or more factors, referred to as drivers that cause change.
Unstructured data	Data that cannot fit cleanly into a predefined structure.
Up-selling	Involves the increase of order volume either by the sales of more units of the same purchased item, or the upgrading into a more expensive version of the purchased item (Kamakura, 2007:42). A process by which customers are upgraded to more expensive products, replacing their existing products (Stubbs, 2014:217).
Vertex (graph database)	A single node in a graph structure.
Vertical scaling	The process of taking existing actors in a system and increasing their individual power.
Weight (graph database)	A quantity associated with an edge.

CHAPTER ONE: INTRODUCTION TO THE STUDY

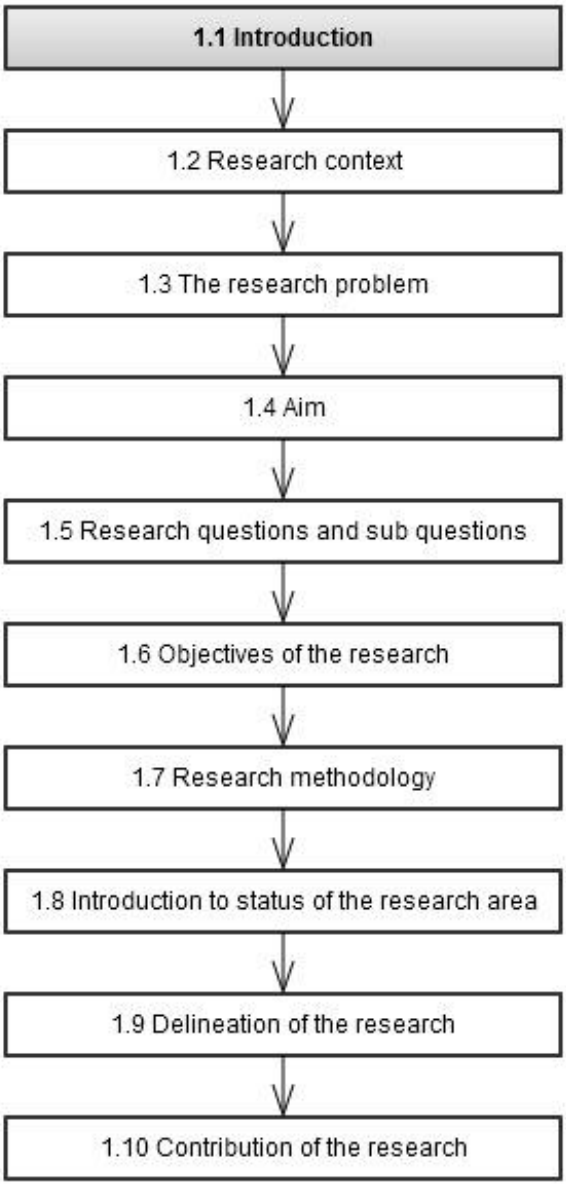


Figure 1.1: Chapter One layout - Introduction

1.1 Introduction

> Data represents the lowest level of abstraction from which information and then knowledge can be derived (Malicke, 2012:15).

Data forms the building blocks of information within an organisation allowing for knowledge and facts to be obtained. Drawing from the ontological (fundamental category of reality) and epistemological (nature and scope) essence of data, data represents the object of knowledge as presented to the mind.

Data is information at its simplest. Information is comprised of contextualised data which eventually forms the basis of knowledge. Significantly, data correlates the basis of information and derived knowledge crucial to an organisation's basic existence; hence a driver of the data gathering trends in industry. According to Davenport and Prusak (2005), data is not knowledge, neither is it information, though it forms the basis of both. Confusion as to what data, information and knowledge mean has been the cause of business expenditure that failed to yield the needed revenue. Mosley *et al.* (2009:2) distinguish data from information by stating that "data is the representation of facts as text, numbers, graphics, images, sound or video while information is data in context". According to Setzer (2006), data provides a common way of representing and transmitting information, while knowledge is a subjective and a practical representation of contextualised data.

Khatri and Brown (2010) state that organisations invest in information assets to curate data to leverage insight for service delivery. Generated insight from curated data facilitate the formation of impressions which are the basis of process intelligence and business model re-engineering for competitiveness, profitability and sustainability (Khatri & Brown, 2010). According to Teece (2010), an organisation's business model reflects a conceptual representation of how the organisation obtains value. That is, when a business is formed by default, it implicitly or explicitly adapts a particular business model that delineates the design or architecture of value creation, service delivery and data capture mechanisms it may employ. The significance of the business model then lies with its description and communication of the manner in which the enterprise will deliver value to customers, attract customers with the intention of paying for value, and further converting the acquired value to profits (Teece, 2010). In the process of delivering value to the service beneficiary, data is generated which leads to business knowledge that can be leveraged to optimise the business model as a form of insight monetisation. The business model dictates which information and technology assets are needed for a successful operation. The information assets within an organisation are documented facts that have value and go hand in hand with the technology assets. The technology assets within an organisation are the technologies (computers, communications and databases) that support the automation of well-defined tasks.

According to Schmarzo (2013), every now and then new sources of data emerge with the potential to transform, drive and allow deriving never envisaged business value especially with customers, products, and transactions. These data sources change the way business orchestrates and models value generation. As a result, retailers over time have become compelled to place business elements such as the customer firmly at the centre of their decision journeys in order to reveal influencers along the retail path to purchase, while

revealing opportunities for seamless and personalised retail experiences (Hritzuk, Esquero, Jones & Burke, 2013).

Schmarzo (2013) further states that web clicks have become the new knowledge currency that enables online sellers to gain a competitive advantage over their brick-and-mortar counterparts, mainly due to the insights within the weblogs. The insights gained allow sellers to influence customers purchase choices using technologies like recommendation engines. The insights spur organisations to change their business models to integrate their structured and unstructured data, forming the basis of Big Data (BD) which Mohanty *et al*. (2013:1) define as "a combination of transactional and interactive data". While relational technologies have mastered the art of managing volumes of transaction data, these technologies where not built for the terabytes to petabytes of industry-generated data. Together with data type diversity, complexity and volume attributes, these datasets (BD) become too strenuous for existing technology stacks in the form of warehouses and business technology architectures. Also, there is the interactive side of it which adds to the complexity of BD to pose a challenge to traditional data management (Schmarzo, 2013).

Facebook, Google, Twitter and Yahoo are companies that have always dealt in the realm of BD (Forsyth, 2012). These companies are "necessitated to know how to deal with BD, as data explosion is on an upward trend due to its origin" (Forsyth, 2012:15). According to Forsyth (2012), the data deluge is made worse by the increasing proliferation of technologies able to generated data. These technologies include sensors, weblogs, imaging technology and data streaming devices. McCallum (2012) refers to this data as "bad data". Bad data is data that gets in the way, disruptively, yet holds the potential to transform a business. This is the data that wastes the consumers' time in that they attempt to put the bits and pieces together, causing huge frustration. McCallum (2012) states that the real costs associated with BD are not storage but rather getting value from the stored data. Furthermore, storing data correctly for retrieval is a challenge.

This research investigates the use of BD in a media organisation, the integration of BD in the business model, and how BD can be leveraged. The research is carried out first by ascertaining on a higher level what factors affect business to leverage BD and how BD can be leveraged for business enablement.

1.2 Research context

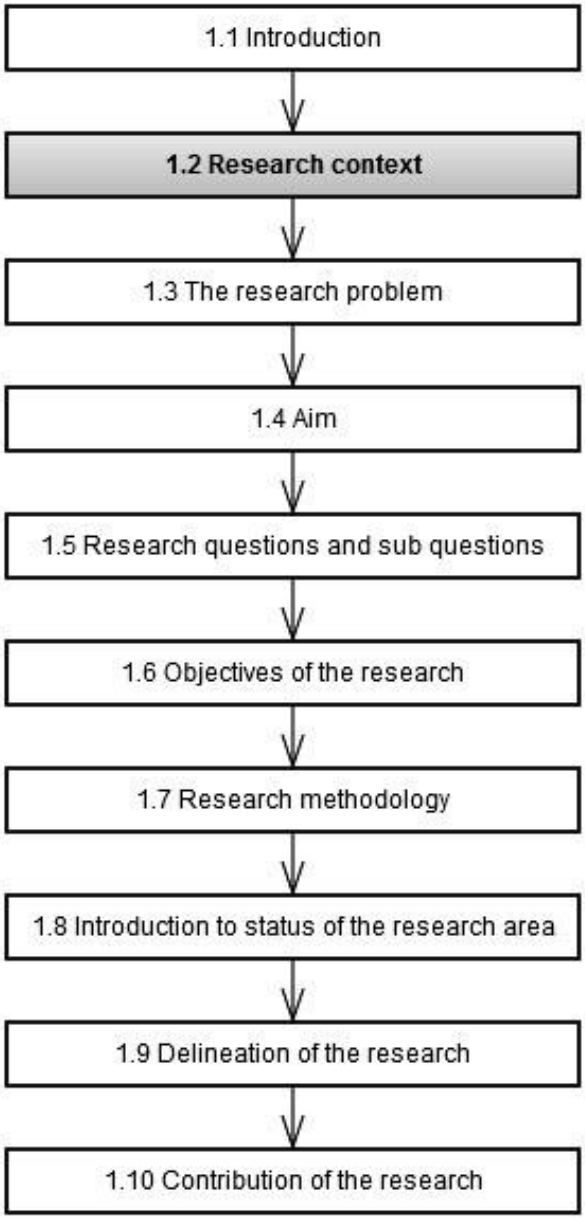


Figure 1.2: Chapter layout – Research context

The term BD, according to Schroeck, Shockley, Smart, Romero-Morales and Tufano (2012), is pervasive and causes confusion as it has been used to cone many concepts including huge quantities of data, social media analytics, next generation data management capabilities and real-time data. According to Schroeck *et al.* (2012), vast quantities of data are being created in many organisations. However, very few organisations have documented structures based on insight into data curation which stems from a mastery of process. The lack of mastery into curation processes to augment key business processes hinders the eventual leveraging of BD curated opportunities in the form of monetisation or organisational

metamorphosis, to afford organisations the needed competitiveness. This in turn prevents gaining much needed implicit or explicit quantifiable benefits that parallels data curation as a performed organisational process. Although the opportunities of curated data are viewed as a direct business benefit of organisation-wide data development and business enabler, leveraging these opportunities still remains a challenge for many organisations, especially start-ups and businesses lacking the needed capacity and skills (McCallum, 2012).

Mosley *et al.* (2009) define data development as the analysis, design, implementation, deployment and maintenance of data solutions to maximise the value of data resources for the enterprise. The data development process focuses on stating data requirements, designing and implementing data solution components, and includes primary components such as databases and related data structures. Information, according to Mosley *et al*. (2009), is data in context that has relevance and timeliness, and that for an organisation to capitalise on the full value of curated data, it is required that the organisation is aware, merit-wise, of the benefits of data development and management through its life-cycle of beneficence. Although many organisational heads will concur as to the opportunities curated data may offer, many still lack the needed insight to commence a structured curation process and to continue the data management process through its life-cycle. The process of data curation comes with many activities such as information discovery, retrieval and provision for use and re-use, and data quality management (Mosley *et al.,* 2009).

According to Jahnke, Asher and Keralis (2012:1), data curation is comprised of a set of activities that includes preserving, maintaining, archiving and depositing of data to keep it secure, intact and accessible for reuse. They further assert that data curation or management is an on-going process throughout the life-cycle of an organisation, and the pivotal role of data to business facets such as decision-making, mandates data administration through its life-cycle. The data curation processes of maintaining, preserving, archiving and depositing add value to data throughout its life-cycle, making the data relevant. Some of the practices that underlie data curation include archiving, authentication, management, preservation, retrieval, use and reuse, representation, visualisation, metadata and provision for organisational functions such as decision-making and predictive analysis.

The impact of the data deluge on society by multimedia communication is mostly exacerbated by the diffusion of digital processing algorithms and hardware (Pereira, 2007). According to Pereira (2007), information is primarily disseminated through text, video and music. For example, most users of smart phones access the Internet via their phones. As a result, users are able to access information posted on the Internet much quicker and easily, multiplying the dissemination of data.

The Massachusetts Institute of Technology (MIT) Computer Science and Artificial Intelligence Laboratory (CSAIL) has been exploring the solutions and challenges surrounding the emergent field of BD. This initiative is called *bigdata@CSAIL*. The initiative brings together leaders from academia, industry and government to explore and develop techniques for processing, capturing, analysing, storing and sharing BD with the primary goal of making BD more useful to society as a whole (Hardesty, 2012).

Agrawal, Bernstein, Bertino, Davidson, Dayal *et al*. (2012) state that BD integration is a major value creator. Lumpkin (2013) asserts that integrating Big Data into corporate information architectures provides deeper insight into what customers are thinking and how business operates. Forsyth (2012) supports this by stating that a report by Mackenzie Group International (MGI), after a study of five markets which include healthcare, retail, public sector, manufacturing and personal-location data, came to the conclusion that BD can generate value in each sector. For example, a retailer using BD or embracing BD could increase its operating margin by more or less 60 per cent, providing motivation for the continual capture of data. According to Forsyth (2012), organisations curating BD will outperform their competitors by about 20 per cent in the next three years, making BD curation an enabler for business. Manyika, Chui, Brown, Bughin, Dobbs, *et al.* (2011) mention that the use of BD has become the key way for leading companies to outperform their peers. For example, the UK branch of Tesco has used BD to capture market share from its local competitors. There are many other examples in the financial services and insurance sector (Schmarzo, 2013).

According to Becla and Wang (2005), more data enables new functionality and capability; as a result, many leading organisations are transforming their thinking on data, treating data more as an enterprise asset than operational cost to be minimised. Big Data administration is a long-term movement to extract much more sophisticated inferences from large datasets, ones that are much less clean and devoid of structure compared to data in traditional relational databases. Forsyth (2012:5) mentions that:

> Big Data applies to datasets that are large, complex and dynamic (or a combination thereof) for which there is a requirement to capture, manage and process the dataset in its entirety, such that it is not possible to process the data using traditional software tools and analytic techniques within tolerable time frames.

In addition, Mayer-Schönberger and Cukier (2013) mention the BD concept looking at the 3Vs which represents volume, velocity and variety, the primary attributes of BD. Volume reflects the breadth and depth of data being collected for varied reasons. Velocity represents the point of interaction and speed of data as it enters and exits the organisation. Variety reflects the variety of data types such as emails, web pages, books and videos.

The dimensions have increased in recent years to include accuracy and value (Hitzler & Janowicz, 2013). According to Buhl, Röglinger, Moser and Heidemann (2013), when dealing with large datasets some of the data is bound to be 'dirty' which raises the question: "How clean is the data that is good enough for analysis as this affects the reliability of results?" The fifth dimension, value, needs to be understood as data has become a company asset. For an organisation, the storage of these varied data objects is critical for the overall system performance and functionality. This is due to two principal constraints, timing and size (Vakali & Terzi, 2002; Sathi, 2012), as the rate of assimilation of data and the increasing data volumes remain a challenge for organisations seeking to leverage insight from BD.

1.3 The research problem

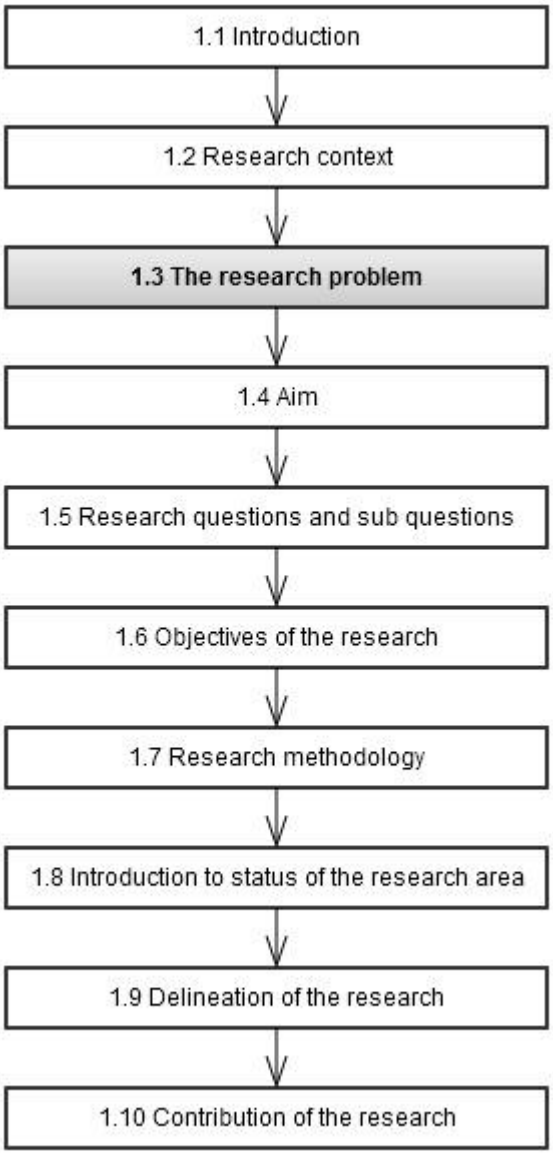


Figure 1.3: Chapter layout – The research problem

Khatri and Brown (2010:4) state that “many organisations do not know what data they have, how critical that data is, the sources that exist for critical data or the redundancy degree of their data assets”. In order to manage the inventory data as well as related sources, data curators need to develop an understanding of the types and sources of data, the storage requirements and growth trends. According to Dumbill (2012), organisational datasets grow so large and complex that they are difficult to work with using traditional database management tools. The key factors affecting this data include “volume, velocity and variability” of the data. Data within an organisation may be categorised as being structured, unstructured or semi-structured. Structured data forms about 20 per cent of the total amount of information whereas unstructured data forms up to about 80 per cent of data within the organisation, as mentioned by Kolluru (2012).

Leblanc (2011) states that business is at an inflection point where it is faced with having to answer questions and resolve challenges in order to drive business transformation forward. Prajapati (2013:62) mentions that most of the tools that are best suited for curating BD originate from open source initiatives such as the R and Hadoop integrated programing environment. This as a result provides the research community with a huge opportunity as no investment in software licenses may be needed.

According to Manyika *et al.* (2011), the benefits of gaining knowledge from BD are manifold, for example an organisation becoming ‘information advantaged’ through BD curation and leveraging insights thereof to attain a competitive edge that quickly translates into a differentiation potential such as gaining process optimisation (a form of monetisation). Monetisation happens in two forms: process improvements and/or revenue generation from curated data. Organisations struggle for many reasons to leverage BD for insights and knowledge. The challenges faced are further exacerbated by the inability of traditional data processing tools to adequately process data, necessitating the ability to overcome the imposed data challenges such as increasing data sizes, velocity of incoming data, and the need for new tools for curation and analytics. Companies find it difficult to leverage the opportunities BD offers them in terms of monetising the content of curated data.

Problem Statement:

Companies find it difficult to leverage the opportunities Big Data offers them.

1.4 Aim

The aim of this research is to explore how BD can be leveraged to utilise the opportunities BD offers the organisation, such as leveraging insights to gain competitive advantage.

Guidelines are proposed on curation procedures, aimed at improving the curation process to ultimately create a competitive advantage for the data curating organisation. The created guidelines will enable meaningful conversation between technical data curators, operation level managers, technical personnel, and strategic and tactical decision-making workers.

1.5 Research questions and sub-questions

The following sections (1.5.1, 1.5.2 and 1.6) focus on the research questions, sub-questions and objectives of the research. These questions were asked in order to support the research process considering two basic interrogatives: what and how. Table 1.1 presents the research questions and sub-questions for ease of readability. Subsequent to each main question are the sub-questions, methods of analysis and objectives of the questions.

1.5.1 Research questions

The two main research questions are:

i) What are the factors affecting business to leverage Big Data for competitive advantage?
ii) How can Big Data be leveraged in a media organisation for business to gain a competitive advantage?

1.5.2 Research sub-questions

The research sub-questions are stated below:

Research sub-questions in support of Main Question One:

a) What is business doing to leverage Big Data to gain a competitive advantage?
b) What is the business's view of Big Data in terms of competitive advantage?
c) What are the policies and strategies for leveraging Big Data?
d) What information do businesses want to get from Big Data?
e) What kind of data is being curated as part of Big Data?

Research sub-questions in support of Main Question Two:

a) How can Big Data be utilised effectively for competitive advantage?
b) How can business implement Big Data curation?

Table 1.1: Research questions, sub-questions and objectives

Problem statement	**Companies find it difficult to leverage the opportunities Big Data offers them in terms of monetising the content of garnered data.**	
Research question 1 (RQ 1)	What are the factors affecting business to leverage Big Data for competitive advantage?	
Research question 2 (RQ 2)	How can Big Data be leveraged in a media organisation for business to gain a competitive advantage?	
Research sub-question (RSQ)	**Objectives**	**Research method(s)**
RSQ 1.1 What is business doing to leverage Big Data to gain a competitive advantage?	Identify current procedures and schemes in place for BD curation.	• Case study • Interview
RSQ 1.2 What is the business view of Big Data in terms of competitive advantage?	Identify businesses' view of BD and how this data is used after curation for competitive advantage.	• Documents review • Case study • Interview
RSQ 1.3 What are the policies and strategies for leveraging Big Data?	Identify policies and strategies in the organisation for leveraging BD.	• Documents review • Case study • Interview
RSQ 1.4 What information do businesses want to get from Big Data?	Determine what business data to curate.	• Documents review • Case study • Interview
RSQ 1.5 What kind of data is being curated as part of Big Data?	Identifying types of data and data models deployed.	• Literature • Case study with interviews
RSQ 2.1 How can Big Data be utilised effectively for competitive advantage?	Ascertain the aims of curation, better strategy and policy implementation.	• Interview with data curators and senior staff • Interviews with industry experts at the fore front of technical data curation
RSQ 2.2 How can business implement Big Data curation?	Identifying curation types and implementations. Ascertain the aims of curation.	• Interviews • Interviews with technical staff

1.6 Objectives of the Study

The objectives of the study are to:

a) Ascertain how the organisation is using Big Data.
b) Identify factors contributing towards curating Big Data.
c) Evaluate the status quo of the organisational data against the Big Data maturation index.
d) Determine how knowledge of the identified factors can be used to propose curation guidelines and recommendations for curation of Big Data.

These objectives are aimed at improving and optimising curation processes in a media organisation to gain a competitive edge as a means of monetising insights from data and to foster process intelligence and continuous transformation.

1.7 Research methodology

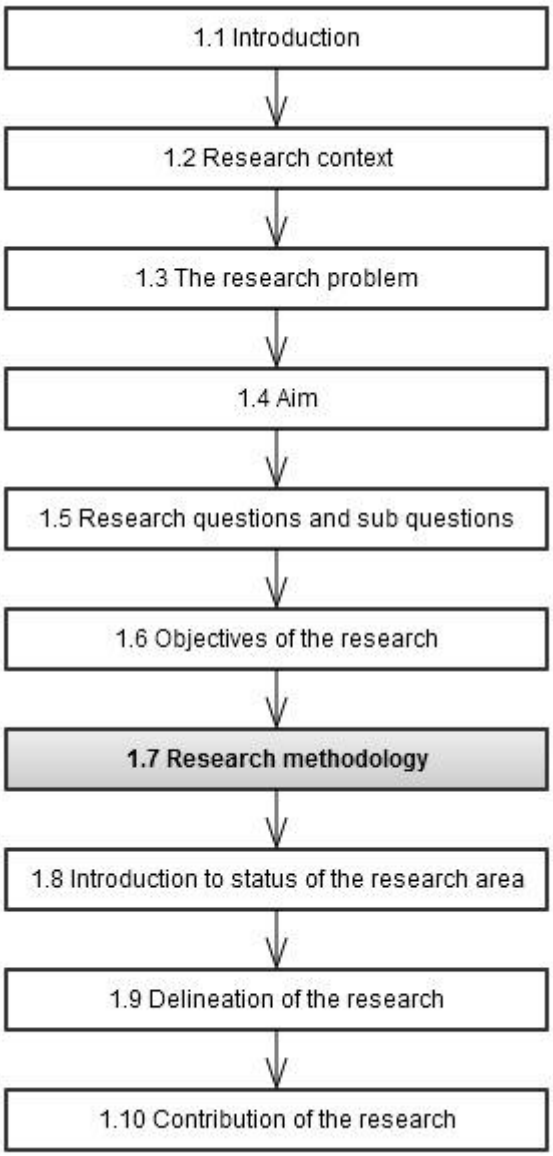


Figure 1.4: Chapter layout – Research methodology

All research is dependent upon underlying philosophical assumptions regarding what constitutes valid research and what research methods are valid for creating knowledge in a given study. This accentuates the significance of knowing the philosophical assumptions that underlie a chosen research methodology. Flowers (2009:1) states that “it is vital to consider the different research paradigms and matters of epistemology and ontology as these contextualise the perceptions, beliefs, assumptions and nature of reality and truth”.

These variables influence the way in which the research is carried out from design through to conclusion.

1.7.1 Research philosophy

According to Flowers (2009), the overall philosophy is the choice between two alternatives—positivist or phenomenological. The basic beliefs differentiating positivist from phenomenological is the view of the surrounding world in which research is conducted. The positivist perceives the world as being objective and external and the phenomenologist views the world as being socially constructed and subjective. Table 1.2 briefly delineates differences between the two paradigms.

Table 1.2: Philosophical paradigms

(Source: Gray, 2013:25)

Variable	Positivist	Phenomenological
Basic	The observer is independent; science is free.	Observer is part of what is observed; science is driven by human interest.
Focus	Focus on facts; look for causality and fundamental laws; reduce phenomenon to simplest elements; formulate hypotheses and then test them.	Focus on meanings; try to understand what is happening; look at the totality of each situation; develop ideas through induction from data.
Preferred methods	Operationalisation of concepts so that they can be measured.	Depend on multiple methods to establish views of phenomenon.

According to Wahyuni (2012), the philosophical backgrounds normally remain implicit in most research as they affect the research practice. This is supported by Creswell (2013), Neuman (2011) and Saunders, Lewis and Thornhill (2009). Predominantly, the authors suggest the essence of questioning the research paradigm to be applied when conducting the research study as this influences the nature of the research. Relative to the given research problem, this research follows the phenomenological paradigm which is inclined towards producing qualitative data. The data is rich and subjective and the research focuses on a natural setting rather than a controlled environment such as a laboratory, which Peterson (2014) denotes as deductive. The validity of results is high but affected by low reliability. This can be controlled by triangulation. According to Peterson (2014), the research process is comprised of three dimensions, namely ontological, epistemological and methodological. These dimensions define the research paradigm which is a system of interrelated practice and thinking that shapes the nature of enquiry. A paradigm is a research culture with a set of beliefs, values and assumptions that a community of researchers share regarding the conduct and nature of research (Walker, 2010).

1.7.1.1 Ontology

According to Peterson (2014), ontological assumptions ask questions relating to the nature of reality. It encompasses claims about what exists, what units make up this reality, what reality looks like and how these units interact with each other. The main focus of ontology is the question of how the world is built. Ontology can be viewed as the nature of the social context a researcher might acquire knowledge about. This infers two basic distinctions:

- There is a real world independent from our knowledge (foundationalism)
- There is no real world but the world is socially and discursively constructed

The latter implies a dependence on culture or particular time. Foundationalism refers to independence from knowledge upon which life is built. Ontology views the construction of reality as intersubjective through meanings and understanding that are developed socially and experientially. Ontology is based on the assumption that subject and object cannot be separated; the values become inherent in the different phases of the research, implying a negotiation of what is true. In this approach findings come to light as investigation proceeds, which is a salient feature of subjectivism. This research adopts a subjective paradigm.

1.7.1.2 Epistemology

“Epistemology has to do with the nature and forms of knowledge” (Scotlant, 2012:9). The emphasis, according to Scotland (2012), is on how knowledge can be created, acquired and communicated. Epistemology describes the theory of knowledge, the view of what we can know about the world and how a researcher is able to know it. Epistemology (the science of knowledge) focuses on views about the most appropriate ways of enquiring into the nature of the world (Easterby-Smith, Thorpe & Jackson, 2008), what is knowledge, the sources and limits of knowledge (Eriksson & Kovalainen, 2008). It is affected by the social constructions of reality.

According to Adam (2014:4), “epistemology is a dynamic process that centres on the knowledge-gathering process and involves developing better theories than competing theories”. In short, epistemology is process driven and focuses primarily on how a researcher comes to know what he knows. Epistemology inherently brings across a level of certainty relative to conclusions drawn from the research process (Adam, 2014). Epistemology (theory of reality) and ontology (reality) in itself are connected such that the researcher’s epistemological stances will indicate the best process to adopt in order to answer the research questions. According to Saunders *et al.* (2009), epistemology can be seen from an interpretivist or positivist perspective—interpretivist being qualitative and positivist being quantitative in approach.

This research will follow an interpretivist stance. Interpretive research assumes that people associate and create their own subjective and intersubjective meanings as they interact with the world around them, by attempting to understand phenomena through meanings participants assign.

1.7.2 Research approach

According to Adam (2014), the foundations on which scientific researchers work is their ontological and epistemological positions. The foundations underlie the approach and methodology as the path taken by the researcher is shaped. The research methodology is a strategy of enquiry which moves from underlying assumptions to research design and data collection (Myers, 2009). The most common classifications of research are qualitative and quantitative. Quantitative research relies on numerical inferences while qualitative research draws its meaning and essence from rich textual descriptions (see Section 3.8). Wahyuni (2012:72) describes methodology "as a model to conduct a research within the context of a particular paradigm as it encapsulates the philosophy and methods (tools, techniques and procedures) required to embark on the study".

Wahyuni (2012) clarifies the difference between research methodology and method by relating the two to a map (domain) and steps to travel between two distinct places (Jonker & Pennink, 2010). The research method specifies the research tools, procedures and techniques (for example interviews and data collection methods required to gather data), and methods of analysis, which imply that it forms the practical component required to establish the research process. The methodology theorises the approach with important ideas relevant to the research while the methods implement the research through well-thought-out tools that form the design of the research. Reasoning or direction of theorising underlining research methods can be divided into inductive and deductive approaches (Neuman, 2011), that is, theorising can commence with abstract thinking leading to eventual concrete evidence through logical connections. Alternatively, one may start with specific observations of empirical evidence that leads to abstract ideas from generalisations from evidence.

1.7.2.1 Deductive

A deductive form of reasoning is logic based on existing theories of stories and storytelling as knowledge sharing practices. According to Gray (2013:19), a deductive approach uses a "theory to generate a working hypothesis concerning relationships between variables, the hypothesis is operationalised and tested leading to acceptance or rejection based on the evidence". Saunders *et al.* (2009) assert that researchers using a deductive approach develop a theory and hypothesis, and then design a research strategy to test the hypothesis.

The researcher starts with theories and suppositions and then systematically tests the implications.

1.7.2.2 Inductive

An inductive approach does not require a hypothesis to commence research. A scientific hypothesis is based on a background theory which typically assumes the form of a proposition of which the validity hinges on empirical confirmation, otherwise a hypothesis, as indicated by Bendassolli (2013), is nothing but imaginative speculation. Furthermore, Bendassolli (2013) mentions that qualitative researchers contend that their work does not consist of proposing and testing hypotheses. The focus of inductive data analysis is to draw insight, and provide better understanding of the interaction of mutually formative influences and to explain the realities of interacting variables, experiences and the participant (Bendassolli, 2013). The research design for this study is a descriptive and interpretive case study that is analysed through qualitative inductive methods.

1.7.3 Research strategy

The research strategy indicates the type of study to be undertaken to provide acceptable answers to the identified research problem and sub-problems. There are many different strategies or designs which include surveys, case studies, grounded theory, action research, modelling, experiments and ethnography, among others. The chosen strategy for this study is the case study. According to Wahyuni (2012:72), a case study is "a research method that facilitates a deep investigation of a real-life contemporary phenomenon in its natural context".

1.7.3.1 Case study

Yin (2014:2) defines a case study as "an investigation into a contemporary phenomenon ('the case') in its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident". The aim of case study research is to comprehend the complexity of the case by understanding its activity within important circumstances. Pickard (2013) classifies case studies into holistic and embedded categories. Both Pickard (2013) and Yin (2014) describe a case study as being shaped by a thorough qualitative approach that relies on narrative, phenomenological descriptions. According to Hartley (2004:326), the primary distinction in designing case studies is between holistic (single-case) and embedded designs (multiple-case). Holistic case study allows the researcher to understand one unique critical case while embedded case study on the other hand involves more than a single unit or object of analysis and is not limited to qualitative research alone (Pickard, 2013).

Pickard (2013) further mentions that themes and hypotheses remain significant but should remain subordinate to the understanding of the case. According to Yin (2014:2), a researcher "may use case study when answering how or why questions with little or no control over the events and preferably a real-life context".

The focus of a case study is on a contemporary phenomenon which traverses a real-life context and boundaries without a clear evident context. In a case study, there are theoretical propositions which guide the collection and analysis of data; these are mentioned as methods which cover qualitative or quantitative designs and single or multiple-case studies. According to Neuman (2011), case studies fall under explanatory, exploratory and descriptive categories.

McLeod (2008) differs by asserting that a case study has no scientific value, results cannot be generalised to the wider public, and it is difficult to replicate. This is due to the near impossibility of ascribing causation in a single case where no pre-test is available and few variables are measured at post-test. This approach to scientific study lacks the very principles of scientific enquiry as produced knowledge lacks diversity of the source of information. McLeod (2008) further asserts that because a case study deals with only one person, event or group, the conclusions drawn may not apply elsewhere. According to Hyett, Kenny and Dickson-Swift (2014), case studies are described as normally lacking systematic data handling, time limits which involve lengthy timelines for writing reports, and the lack of basis for scientific generalisation due to the subjective nature of human observations. In light of this pessimism, what then is an acceptable case study according to standard systematic scientific research?

Contrary to McLeod's (2008) notion, Yin (2010) mentions that case studies provide rich contextual knowledge; even if impressionistic, it may lead to intelligent assumptions. Such a knowledge source may enhance a particular case study as the researcher's concerted effort to acquire knowledge about a particular setting may be succeeded by measuring many variables at post-test. Yin (2014) describes a case study as an intensive investigation of a single unit which involves the examination of multiple variables. The primary focus of a case study is emphasised on an individual unit. A case study may involve at least three steps in design, which include (Yin, 2010:40-45):

i) Define the case
ii) Select one of four types of case study designs
iii) Using theory in design work

These design steps are further expounded in Chapter Three. The object of analysis in a case study, according to Yin (2014), falls under holistic or embedded designs as the flexibility of the case design depends on selecting cases different from those initially identified but not in changing the objective of the study. A holistic design includes a single unit of analysis while an embedded design includes multiple units of analysis.

Neuman (2011) defines units of analysis as units, cases or parts of social life that are under consideration. This research is a single case study with a holistic design preference.

1.7.3.2 Units of analysis

Units of analysis are the things examined in order to construct summary descriptions of all such units and to explain the differences among them. The units of analysis in this study are the departments in the organisation, and the units of observation are the employees curating and using data in diverse ways. According to Wahyuni (2013:73), qualitative researchers "should get involved in a communication with the practitioners in order to understand the current state of real-world practices". Participants were chosen based on recommendations from department heads on the basis of the participants' interaction with data and overall knowledge of data.

1.7.3.3 Data collection

Qualitative data collection, according to Hyett *et al.* (2014), includes data gathered primarily in the form of spoken or written language rather than numbers. Data sources used for this research include interviews with industry experts, leaders, technical personnel at the forefront of data curation in the organisation, observation of current curation processes, and documents analysis so that relevant evidence supporting the derived guidelines may be drawn. Semi-structured interviews were used to explore relevant avenues during the interviews. The interview guide (set of questions) has been compiled to be flexible with varying levels of difficulty.

a) Primary data

Primary data is unpublished data gathered directly by the researcher (Myers, 2009:3). Interviews were used to gather data. Face-to-face and semi-structured interviews were conducted with target participants on an individual bases. Occasionally, there were repeat interviews for further clarification. These in-depth interviews assisted in gathering data on current curation practices and how the organisation is leveraging data for business enablement. Questions were predominantly open-ended as it gave room for further explanation or outlining of opinions by participants.

b) Secondary data

Secondary data is previously published data such as newspaper articles (Myers, 2009:3). Secondary data was gathered through literature analyses from sources such as reports, publications, library books, web portals and company documents. The secondary data is useful for initial impression formation about the chosen research topic. It also augments the primary data in identifying the information and technological requirements of BD and its curation. It helps to clarify and put into context the views and perceptions of data curators within the organisation.

1.7.3.4 Data analysis

This section focuses on interpreting the collected data for the purpose of making informed decisions on the ideas, interests, theories and questions that prefigured the research study. According to Flick (2013:4), "qualitative data analysis is the classification and interpretation of linguistic (or visual) material to make statements about implicit and explicit dimensions and structures of meaning-making in the material and what is represented in it". Chenail (2012:7) defines qualitative data as the rigorous process of selecting qualitatively distinct data, articulating the qualitative meaning ascribed to those units, and commenting on the qualitative similarities and differences noted between and among the distinct units of data. Data analysis forms a central step in qualitative research. Qualitative data analysis can have several aims (Flick, 2013), which are firstly to describe a phenomenon in some or greater detail. The second is the comparison of several cases and what they have in common or what the differences are between them. The second aim may take into account the conditions upon which the differences exist. The third aim is to develop a theory of the phenomenon being studied.

Taylor-Powell and Renner (2003:1) mention that qualitative data analysis (QDA) requires creativity, discipline and a systematic approach, as there is no single best way, This is a fluid process, requiring the researcher to go back and forth through data. This is a five-step process, which include:

i) Getting to know the data
ii) Focusing the analysis
iii) Categorising information
iv) Identifying patterns and connections within and between categories
v) Interpretation

Getting to know your data (Taylor-Powell & Renner, 2003), means reading and re-reading the text (transcriptions) to form impressions; these impressions may prove worthy later in the

research process. This is followed by an introspection of analysis either by question or by topic, time period or event. To focus by question means organising the data by question to look across all respondents and their answers, in order to identify consistencies and differences (Taylor-Powell & Renner, 2003; Flick, 2013).

Categorising information involves a coding process that brings meaning to words before themes or pattern identification occurs. Coding implies a coherent organisation of text as a way to contextually group data. Identifying patterns and connections within and between categories lead to the formation of themes by assembling all the data pertaining to a particular category. The key idea emerges as a theme. Oosthuizen and Phil (2012:60) mention that when qualitative data is compressed excessively, the very point of maintaining integrity of the narrative during analysis may be lost as findings are abstracted to categories and then to themes. In the process of abstracting data, categories cancel out each other where similarities occur. However, an abundance of categories imply an inability to cancel overlapping categories. In qualitative data analysis, abstraction and eventual cancelation of overlapping categories occur until saturation of the data is reached. Saturation point often signals a completion of the study, culminating in a judgement of diminishing returns and stopping further sampling. This is a complex process. Jansen (2010) mentions that seemingly simple study results are a sign of incomplete analysis.

Step five implies the attachment of meaning and significance to the data. According to Mayring (2014), the essence of content analysis is to extract material from a huge amount of text. To finally analyse the research data, a content analysis approach to qualitative data will be used.

1.7.3.5 Hermeneutics

As stated by Myers (2009), the primary focus of hermeneutics is the analysis of text to achieve coherent explanations. Qualitative research data content is thick textual data; it is subject to diverse interpretations. Hermeneutics provide a way of interpreting text to bring across the true meaning while mitigating bias and misconception. This method, according to Mayring (2014), draws on how scientific methods of analysis can be used to address worldly experiences and interpretations.

1.7.3.6 Conversation analysis

Conversation analysis is used in the analysis of day-to-day conversations (Flick, 2013). According to Hutchby and Wooffitt (2008), this includes the analysis of naturally occurring talk (often the result of transcribed audio and video recordings) and seeks to specify the formal principles and mechanisms with which participants express themselves in social

interactions. Conversation analysis focuses on context; it was originally limited to study everyday conversations or family conversations, but the success of this analysis approach culminated in extending it to meetings and interviews. Conversation analysis was used to analyse interviews in this study.

1.7.3.7 Content analysis

Content analysis refers to "the study of recorded human communications" (Babbie, 2001:304). According to Elo and Kynga (2008), content analysis is well known for its ability to deal with a full variety of evidence such as documents and interviews. Neuman (2011:49) defines content analysis as "a technique for examining the content of information and symbols contained in a written communication media". According to Elo and Kynga (2008), content analysis as a technique allows for making inferences by systematically and objectively identifying special characteristics of messages. Content analysis was first used as a method for analysing hymns, newspapers, magazines, articles, advertisements and political speeches (Neuman, 2011). The focus of content analysis as a qualitative data analysis tool for written, verbal or visual communication messages is on the elements of validity, and this has its basis in the trustworthiness of the analysis process, that is, the results should be described in sufficient detail so that readers have a good understanding of how analysis was carried out, clearly stating its strengths and limitations. Content analysis approach to qualitative data will be used to analyse the collected data.

1.7.3.8 Data presentation

The analysed data is presented as a result of the research. The knowledge created from this research is presented in the form of data curation guidelines for the media organisation seeking to integrate BD, and also with recommendations to optimise the process of curation for optimal benefits. Summary values and graphical presentations are used to present the data from the descriptive analysis.

1.8 Introduction to status of the research area

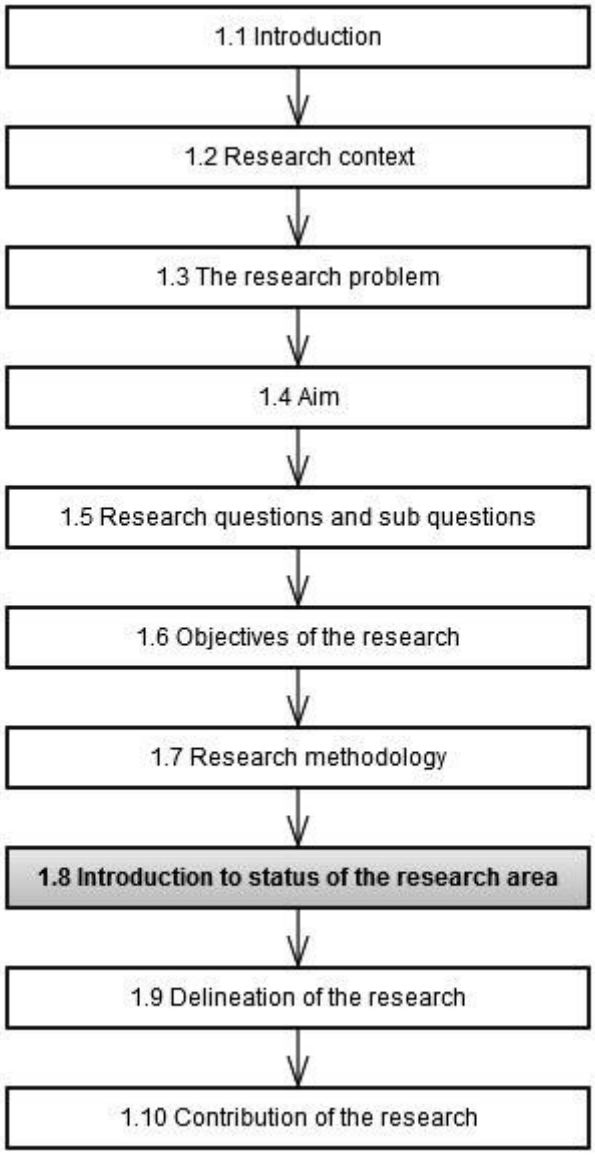


Figure 1.5: Philosophical paradigms

According to Cuzzocrea, Song and Davids (2011), large amounts of unstructured data stem from different sources. The data influx is exacerbated by the ubiquity of the Internet, especially with most businesses becoming digital. An (2012) also mentions that the influx of data surpasses the processing capabilities of traditional curation and processing methods. According to Schmarzo (2013), traditional IT infrastructure is incapable of managing, analysing and acting on BD. This requires new approaches to managing the emergent data influx. According to Jacobs (2009), a visible trend emerging with BD is that the data deluge forces data curators to look beyond tried-and-trusted methods that are prevalent at the time. Schmarzo (2013) affirms the adequacy of traditional data curation processes and the inability

of traditional curation tools by mentioning that Internet companies such as Google and Yahoo explored traditional tools from traditional data and Business Intelligence (BI) vendors by even attempting to tinker with their software kernels to accommodate real-time analysis across hundreds of terabytes and petabytes of data, but that was futile and will not scale to meet data technology demands.

For the purpose of this research, the definition of Manyika *et al.* (2011:1) will be adopted as the primary definition of BD. They define BD as *"*datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse*".*

Manyika *et al*. (2011) in their definition introduce an assumption based on intentional subjectivity to incorporate dynamism as to what a BD set should be in order to be considered BD. This assumption is on the basis that technology is constantly advancing and that datasets currently considered as big are still subject to grow over time. Furthermore, due to the different applicability and diverse adaptability the definition of BD might change. This is borne from comments by Jacobs (2009) and Stonebraker (2010) that the term BD has seen many different definitions since its inception; the definition will continue to change. This definition was chosen because it focuses on the essence of BD as an asset to the organisation and refrains from putting a restriction on size and required technology. It makes the definition applicable to both start-up companies that might not have the large datasets and organisations that have been curating data for many years already.

Another definition that may aid in contextualising and elucidating BD with respect to the proposed case study is the definition by Gualtieri. Gualtieri (2012:10) defines BD as:

> The frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.

According to Forsyth (2012), the benefits of BD curation vary widely, just as the benefits of BD analytics are countless. Chaudhuri (2012) mentions that some of the advantages of BD curation from which an organisation stands to benefit include insight generation, massive value gains in the form of data monetisation, profiling for targeted marketing, and identification of business transformation datasets.

Cuzzocrea *et al*. (2011) state that there are many and diverse underlying factors common within the domain of applications generating BD, for example the rate of data entry into the organisation and the need to generate insight; these factors have mandated the curation of BD. Apparent with these trends is the fact that organisations seeking to leverage the opportunities of BD face the challenge of not knowing what the process entails and how to

approach and launch the BD journey while mitigating risk through know-how and reliable information.

As the size of data increases, the challenges associated with scalability and performance of curation platforms become eminent, thus creating hurdles that hinder data curators and organisations seeking to curate BD (Stonebraker, 2010; Seeger, 2009). According to Stonebraker (2010), curators are seeking alternatives due to the inability of traditional methods. This decision is taken from the perspective of performance, scalability and flexibility of NoSQL-driven databases over the traditional curatorial approach of relational implementations (Stonebraker, 2010).

1.8.1 Traditional data curation

According to Perdue (2012), the term NoSQL was coined in 1998 meaning 'not only SQL'. The term NoSQL is many times misconstrued as meaning the exclusion of standard Structured Query Language (SQL). Perdue (2012) continues by stating that term implies the co-existence of both SQL and NoSQL within some implementations with each playing its own part.

SQL was invented in the 1970s at IBM by Edgar F Codd. SQL is designed to handle structured data (Seeger, 2009). According to Seeger (2009), SQL is much more than just the ability to store and retrieve data, so much so that its implementation in some instances turns out to be overkill. SQL allows people to construct powerful queries that are able to analyse and sub-sample huge amounts of structured data (Seeger, 2009), but with the increasing size of data and different data forms, the insufficiency of SQL to process these large and different datasets became more eminent, spurring the need for alternative processing forms. These data needs arose with companies having to deal with large quantities of data which traditional relational database management (RDBMS) could not cope with (Jacobs, 2009; Stonebraker, 2010). This is not to say that RDBMS upon which SQL runs was inadequate. Zikopoulos (2013:15) affirms this by stating that “Big Data solutions are not a replacement for existing warehouse solutions”. NoSQL databases arose from the data management needs of the many different Internet companies, especially digital media companies such as Google, Digg, LinkedIn, Yahoo and Amazon (Perdue, 2012). These companies sought to address the need for colossal data size management. The colossal data affected their business models and for them to deliver service or stand out in industry required adequate control over the data, thereby circumventing the unaddressed limitations of traditional curation tools.

Pokorny (2013) states that RDBMS have properties normally referred to as ACID. This acronym stands for atomicity (A), consistency (C), isolation (I) and durability (D).

Atomicity, according to Ben-Gan, Sarka and Talmage (2012:413), implies the complete failure or complete success of a transaction. There is no halfway processing. No inconsistencies are acceptable by any means. *Consistency* denotes the fact that data violating a predefined constraint or rule must not be persisted with for integrity sake. *Isolation* becomes relevant where data is accessed concurrently, and *durability* implies that once a transactional operation is confirmed, it is assured. Some of these features, for example consistency, are mandatory in certain environments such as in the stock market as given data must always be correct and consistent (Ben-Gan *et al*., 2012). In contrast to ACID properties, Pokorny (2013) states that databases that do not fully implement ACID can only eventually become consistent as is the case with NoSQL databases. Nayak, Poriya and Poojary (2013) posit that NoSQL does not guarantee ACID properties but instead guarantees that BASE (Basically Available, Soft State, Eventually Consistent) properties make it compliant with the CAP (Consistency, Availability and Partition Tolerance) theorem.

Databases exhibiting eventual consistency have properties that include consistency (C), availability (A) and partition tolerance (P), CAP in short. According to Pokorny (2013), *consistency* means every one reading from the database will see the latest written information. *Availability* implies that every operation is expected to terminate in an intended way and can be accomplished by increasing the number of deployed servers. *Partition tolerance* means that, should parts of the database network be inaccessible, it should still be possible to write to the database. Some examples of NoSQL databases include Cassandra, Redis, MongoDB, CouchDB, BigTable and Tokyo DB.

1.8.2 NoSQL databases and Big Data

The forms of these different datasets and diversity of challenges needing redress culminated in the arrival of the NoSQL era, where data presented in varied forms are stored in databases different from the traditional relational databases. NoSQL databases come in many different forms, optimised for different applications, with many varying advantages over relational approaches. But, common industry implementations see the deployment of persistence technologies that take advantage of the strengths of different databases. With data influx rapidly increasing, many organisations continuously collect data with the aim of monetising data and possibly gaining insights never before available in the past (Schmarzo, 2013). It is essential to point out that data unprocessed is not monetised as a source of value; rather, it is the insights derived and associated with the monetisation effort aimed at yielding value for the curator as a monetisation proposition that generates value. The monetisation of data may take different forms such as packaging insights for reselling, integrating insights to create better products which may be classified as next generation

products, and leveraging the insights to optimise processes and engage customers better for profitability.

The culmination of these data accrual results in gigantic datasets that defy or challenge industry analytical tools necessitating organisations like Amazon, YouTube, Facebook, Flickr and many others to create their own database systems (Schmarzo, 2013).

1.8.3 Traditional data flow in organisations

Many organisations which are curating data have and still are operating with data architectures based on the Online Transaction Process (OLTP) which is based on relational database technology (Kimball & Ross, 2013). Schmarzo (2013) mentions that about 95 per cent of companies are still curating data with technologies firmly hinged on relational architecture. This architecture worked well for organisations both large and small, however with the current technology trends and data influx, the volume of data continues to grow as warehouses are populated with increasingly atomic data and updated at a high frequency. According to Kimball and Ross (2013), over the years the industry has witnessed business data grow from megabytes, to gigabytes, to terabytes and to petabytes.

As systems in these organisations evolved and data sizes increased, many of the organisations ended up with silos of data. These silos are described as a “bowl of spaghetti” by Mosley *et al*. (2009:65) who indicate that the information infrastructures of many organisations:

> …is a bowl of spaghetti due to the convolution of relatively simple isolated applications, supporting with tactical approaches to moving and sharing between these silos has rather increased the relative complexity, sky rocketing the cost of maintaining, heightening gaining insight, collating and retrieval of data for information.

The effect of fragmented organisational data is shared by Baird (2013) who indicates that the traditional organisation is divided into a series of departments, each focused on a specific function with distinct operational boundaries. These systems, besides other complexity issues such as scalability and performance, hinder the time to value creation from data collation from the different systems in a unified way. Figure 1.6 depicts three different data sources feeding data into a traditional relational data warehouse through extract-transform-load processes (ETL) anchored with logic to clean, collate and aggregate data in the data warehouse.

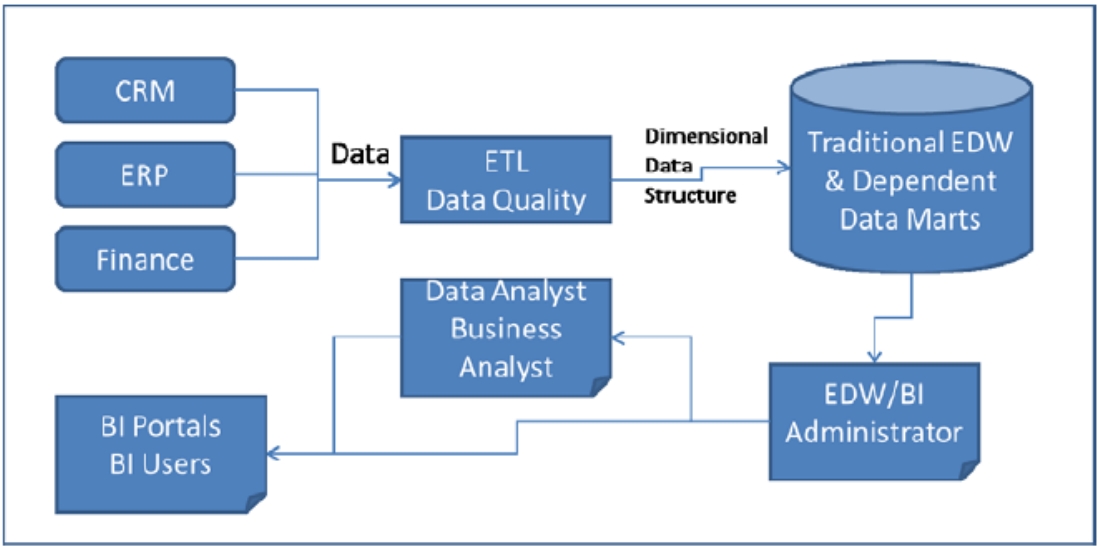


Figure 1.6: Traditional data curation architecture

(Source: Mohanty et al., 2013:109)

The increasing data sizes result in scalability and performance challenges for some data companies with large volumes of data, hence the need for new data management approaches. The business model of many of these companies require the integration of semi-structured data (logs) and unstructured data (consumer comments, documents), real-time data integration for data insight generation and the deployment of predictive analytics. This poses a challenge for traditional approaches of data management. Companies functioning within the realm of BD, for example Yahoo, Amazon, Google and Facebook, sought alternative approaches that ended up in the creation of a whole new ecosystem of tools for data curation (Vaish, 2013).

1.8.4 Competitive advantage with Big Data

According to Stubbs (2014:207), competitive advantage is a "strategic advantage held by one organisation that cannot be matched by its competitors". This advantage may or may not be sustainable and, if not, may eventually be replicated by its competitors. As mentioned by Mosley *et al*. (2009), data is the life blood of the 21st century economy. That is, in a world where BD has become the norm and data affords the curator a competitive advantage, achieving the technical ability to curate and monetise BD is no longer an option but a mandate.

According to Zikopoulos (2013:15), many organisations find it difficult to envision how enterprise data assets or resources can power their business initiatives. It is especially true for organisations that are unaware of the data they have, what kind of questions the data can answer and what kind of decisions can be made leveraging BD.

Competitive differentiation is created when a firm has products or services that distinguish it in a way that its service or product is better, relevant, superior and adaptable in comparison to its competitors. Stubbs (2014) asserts that business analytics is the catalyst that unlocks value from data. Data has the potential to derive and transform business models through information advantage, placing data among one of the essential business elements. It is the dominant force of competitive differentiation (Stubbs, 2014:22). As a result of the benefits of working with large datasets, organisations are seeking ways to acquire even more data they can leverage to gain more value.

1.9 Delineation of the research

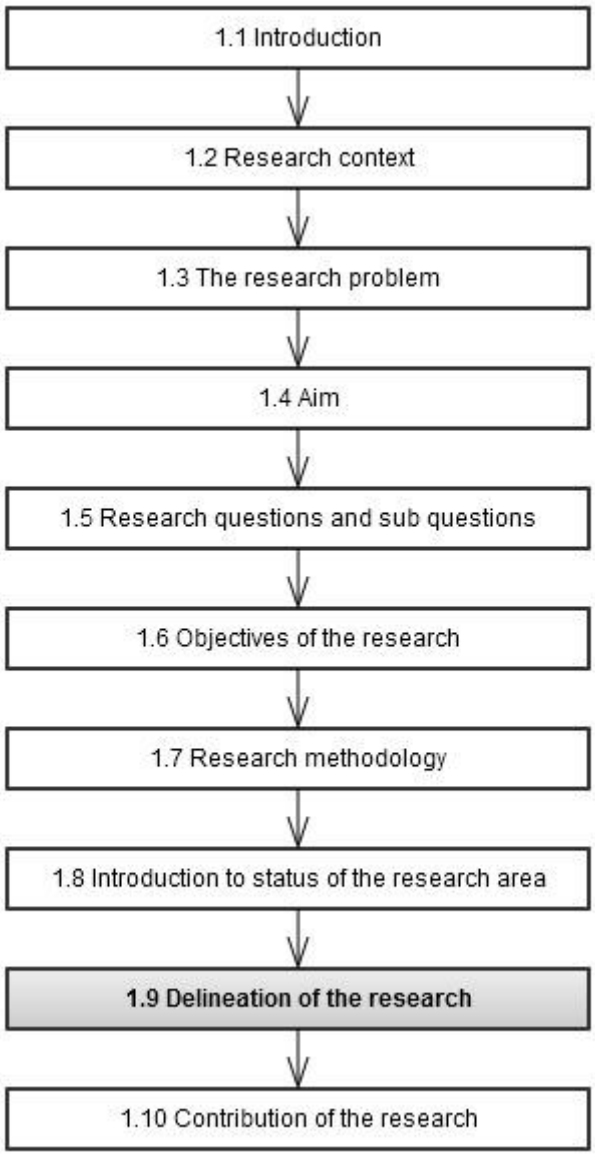


Figure 1.7: Chapter layout – Delineation of the research

This research has not instantiated or implemented any particular system for the curation of BD due to time constraints and scoping of the work. The created guidelines or recommendations in a production environment to ascertain its usability and feasibility have not been tested. In no way does the researcher attempt to generalise the findings and more research needs to be done on different companies before the theory can be tested.

1.10 Contribution of the research

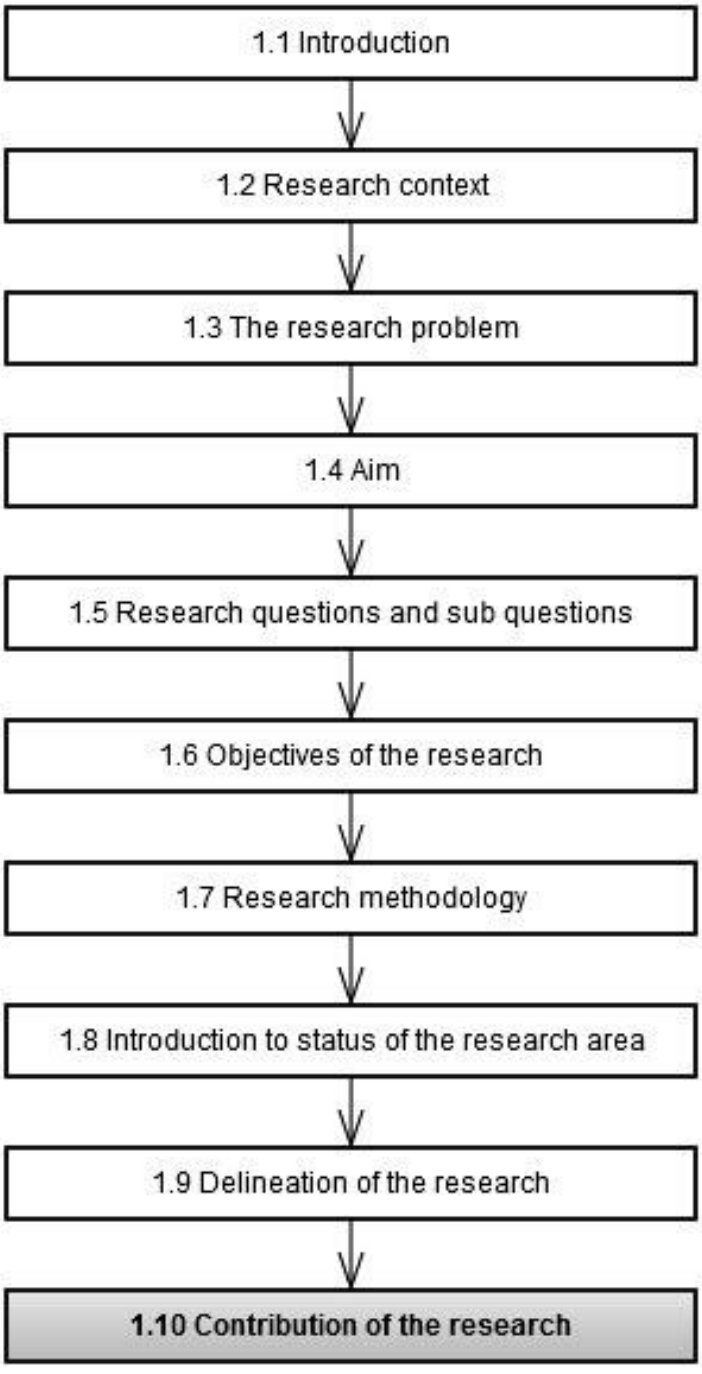


Figure 1.8: Chapter layout – Contribution of the research

The contribution of this research is to create BD curation guidelines adoptable by multimedia organisations both as a start-up and already functional organisation. The research will also identify factors that may cause the media organisation to curate BD. Manyika *et al*. (2011) is of the opinion that organisations curating BD to gain a competitive advantage will outperform their counterparts in the same industry by 20 per cent. This research sought to ascertain the current curation practices and how data is leveraged to gain a competitive advantage, and further to draw insights applicable to the creation of improved guidelines and recommendations that may help improve the current curation practices. With the increased importance and interest in BD, the research contributes to the body of knowledge in terms of the BD discipline.

The research is presented in the following way: The introduction is followed by a literature review in Chapter 2. Chapter 3 presents the research methodology. This is followed by the research findings and a discussion of the findings in terms of the research questions and thematic analysis. The thesis ends with conclusions, recommendations and a reflection on the research conducted. For the convenience and ease of the reader, annexures are added with detailed data on interviewees and the data analysis.

The next chapter discusses BD, databases, E-commerce, business models, BD privacy, NoSQL implementations, solution engineering and polyglot persistence.

CHAPTER TWO: LITERATURE REVIEW

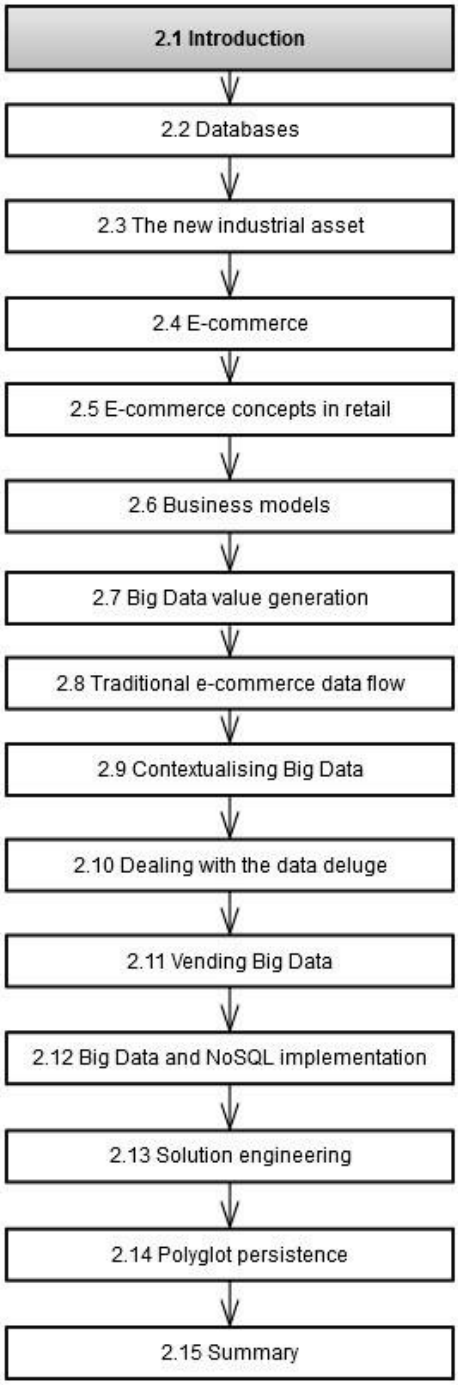


Figure 2.1: Chapter Two layout – Literature review

2.1 Introduction

Data sources have emerged that hold business transformation potential. Sathi (2012) identifies some of these data sources as web and social media, machine to machine, big transaction data, biometric, and human generated data.

Businesses and organisations have had to change their business models to drive and derive business value. In the light of transaction data, Schmarzo (2013) mentions that the 1980s observed point-of-sale data emerging to change the balance of power between consumer package goods manufacturers such as Procter & Gamble, Unilever and Frito Lay, and retailers such as Walmart, Tesco and Vons. The data comprised detailed transaction data about products. It provided retailers unique insights into product sales, customer buying patterns and general market trends that were previously not available. This data progressively and quickly increased in volume while improving insight on how organisations conducted their business. Big Data, as a source of insight, has driven business transformation in many organisations (Manyika *et al*., 2011). According to Das (2012), the world's volume of data doubles every 18 months. This data influx, often referred to as "data deluge", creates a challenge for business leaders; however, despite the challenges, business executives delight in the value of the data deluge (Kelton Research, 2010).

Analysing large amounts of data generated by different applications is critical for gaining a competitive advantage and improving customer experience (Agrawal, Das & Abadi, 2011). According to Agrawal *et al.* (2011), the ability to analyse these large amounts of data is one of the game-changers for business as it enables decision-making and improves the reliability of business forecasts. Improved and better business forecasts reduce uncertainty in decision-making and improve competitive positioning (Kelton Research, 2010). For example, Walmart made a significant capital investment in software to track customer behavioural patterns in real-time which gives them insight, and it is this information they share with product manufacturers. This improves their relationship and influence in the retail market, allowing them to exert pressure for better supplies and manufacturing. Schmarzo (2013) mentions that as Walmart's influence grew, so did their power to nearly dictate the price, volume, delivery and packaging of products. The benefits of working with large storage digital media are numerous. The increasing data deluge and many diverse data forms is forcing industry, especially IT, towards a trend that requires a form of processing that parallels this explosion in data volumes (Austin & Mitcham, 2007; Agrawal, Ailamaki, Bernstein, Brewer, Carey e*t al*., 2008; Manyika *et al*., 2011; Stonebraker, 2010).

According to Jahnke *et al.* (2012), organisations involved in data curation face several important challenges, most of which are not unique to digital research data. Some of the challenges mentioned include inadequate access to network storage, data loss over poor organisational structure, and scale of the data or increasing volume of data which seems further exacerbated by challenges such as selection or retrieval and digital preservation. Agrawal *et al.* (2012) state that much of the data in organisations are not structured naturally, for example, tweets and blogs are semi-structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search.

This later becomes a major challenge when attempting to transform such content into a structured format. In addition, this data become unmanageable because standard tools and procedures are not designed to search and analyse massive datasets (Jacobs, 2009). Such data can range from structured to unstructured data (Pokorny, 2013). Unstructured data is not suited for analytics because analytical tools need some form of structure to analyse data (Pokorny, 2013).

2.2 Databases

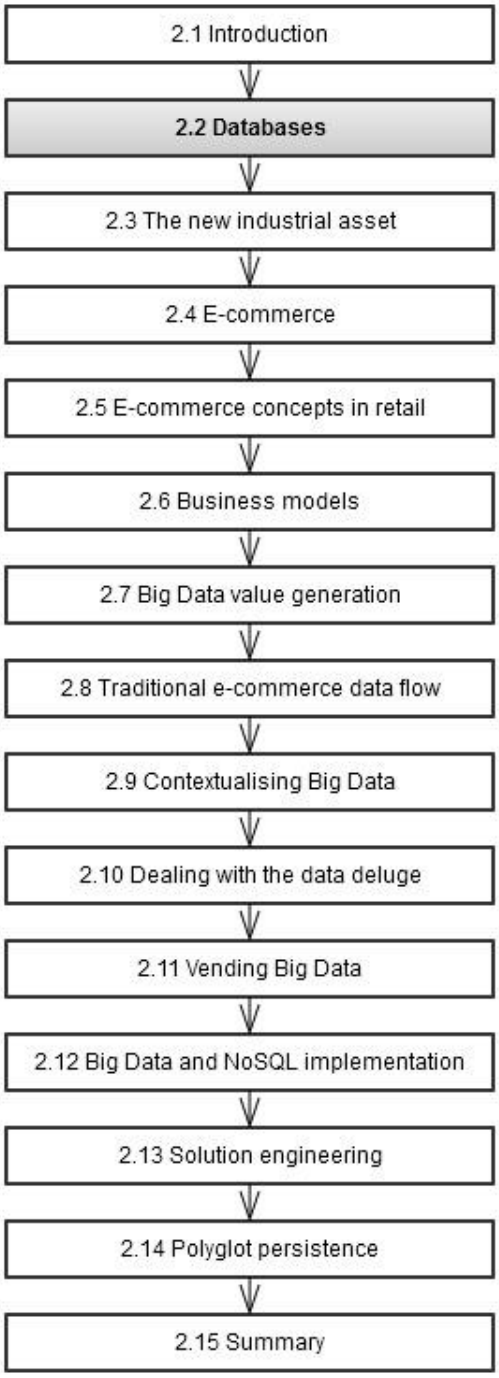


Figure 2.2: Chapter layout - Databases

Piateski and Frawley (1991) state that knowledge acquisition from databases through discovering useful and interesting information has many different approaches, including an inductive approach, Bayesian statistics and semantic query optimisation. There are other forms such as knowledge acquisition with expert systems, information theory and type-1 fuzzy sets. With the myriad forms of information sources, databases have become a means and mechanism for collectively storing, organising and retrieving data and information. Databases are systems designed to offer enterprises an organised mechanism for storing, managing and retrieving information (Stephens, 2008). According to Stephens (2008), many programs that seem at first glance to have little to do with traditional business-oriented data, use databases to make processing easy. The essence then lies in the ability to serve as a repository of information that the business application manages and displays for use (Stephens, 2008:4). Stephens explains that a database is a shared collection of logically related data, designed to meet the information needs of multiple users in an organisation. More often than not, the term database is often erroneously referred to as a synonym for a database management system (DBMS), which is the software system responsible for the control of the database. The use of databases in organisations is hugely important for data management. The justification for data as an enterprise asset hinges on its use as a decision tool for organisations. A database provides the information needed to steer the organisation into successful situations as a resource. Organisations are unable to make good and effective decisions without contextualised data attained in time and in the correct format without proper data management.

According to Stephens (2008), databases have originated from simple forms such as file systems to complex forms such as object-relational databases, hierarchical, network, relational object databases, and the current emergent forms such as the NoSQL databases.

The concept of a relational database was introduced by Codd (1970) in his seminal paper, *A relational model of data for large shared data banks*, which was published in 1970, with the precise meaning as to what a relational database implies. Codd (1970) describes a relational table structure to be comprised of:

- Columns and rows
- A row is a sequence of values such that the n^{th} value of the row corresponds to the n^{th} column of the table
- Each row is identified by a primary key
- A result table is returned by a *select* statement

Codd's (as cited by Halpin & Morgan, 2010) inference of the database includes the "Codd 12 rules", but over time, many other implementations of the relational model did not conform to

all of his rules. At a minimum it presented data to a user as a relation. However, it gradually expanded to include many other classes. These relational database systems have remained the dominant choice for both transactional and analytical applications but with the influx in the current data deluge experienced across industries, many other approaches to dealing with the data deluge have emerged, including the NoSQL approach. Relational databases deployed structured query language (SQL) as a means of retrieving data from the database. Figure 2.3 shows a progression of computing timelines with increasing data size complexities.

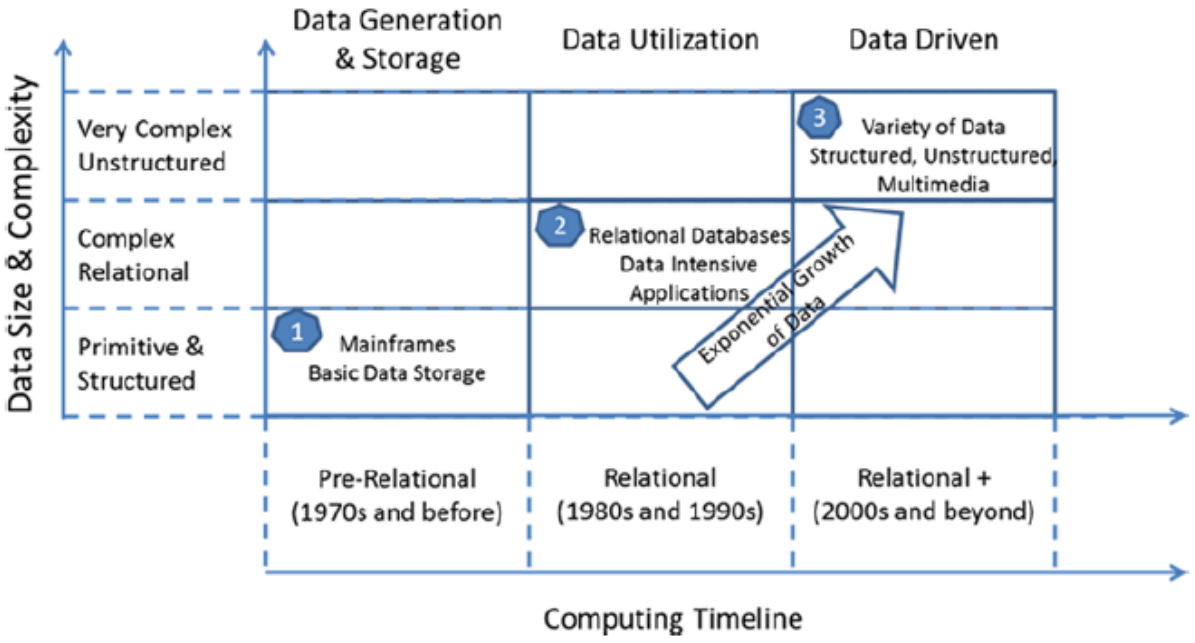


Figure 2.3: Evolution of Big Data

(Source: Mohanty et al., 2013:12)

Relational databases, though a market choice, have enjoyed dominance in storage over the years. Specific characteristics such as enforcing data integrity afforded the user better security where necessary, and similarly, data could be managed to avoid duplication (Kimball & Ross, 2013).

Relational databases assume a well-defined structure in data; however, it can be dense and largely uniform (Tawari, 2011:5). As a prerequisite for the properties of data storage, an upfront definition of a schema with specified inter-relations and systematic referencing is required. It also assumes that the indexes for relational databases are consistently defined on specific datasets which promotes faster querying (Tawari, 2011:5). As data forms evolved, predefining data structures for storage became less desirable in developing systems. These also became more intensified by the massive datasets on curation today.

Superficially, a curator may decide to drop constraints, de-normalise tables, and/or relax transactional guarantees to facilitate scaling. Yet, after these modifications, the RDBMS distinctly evades its original definition and as such, the data set begins to look like any other NoSQL database (Stonebraker, 2010).

Over time, the demands of data culminated in a means to alleviate the problems imposed by the curation of large sparse data. But this demand took a sharp down-turn away from the integrity, flexible indexing and querying, quickly turning most implementers to NoSQL databases (Tawari, 2011). In the absence of SQL, however, the NoSQL implementation is a desirable feature with many implementers working hard to bridge this gap. Some implementers have turned to the use of persistence polygenism to adopt the power or strength of each implementation. NoSQL has become a generic term for databases that do not follow the relational principles of RDBMS, specifically the ACID nature which favours speed, scalability and the likes for companies dealing with massive datasets such as Google, Facebook, Yahoo and Twitter.

Regarding classifications of databases, many authors (including Stonebraker, 2010; Kimball & Ross, 2013) have classified databases based on application (mode of use), placing them under:

- Analytic
- Operational

Another way of classifying databases is according to the data model. A data model is the intangible form in which data is stored. The inference here though relates to the structure of the database more relevant and relative to a theoretical idea. This, however, is an abstract idea, as it is a concept allowing for conceptualisation and visualisation of the database from a global perspective, but it promotes a quick and easy way of communicating data structures, storage and retrieval (Tawari, 2011). The highest standard of transactional integrity in RDBMS is described by a database's conformance to ACID, meaning Atomicity, Consistency, Isolation and Durability (Tiwari, 2011) as described in Section 1.8.2.

According to Tiwari (2011), resource unavailability in long-running transactions is a challenge that appears quite frequently in high-availability scenarios. This is made worse in less tolerant situations of resource unavailability and outage; these and many other factors signify the inability of relational databases to singly meet the pressing demands of data processing, curation or management. As such, the user may need to seek other alternatives to the process data management.

2.3 The new industrial asset

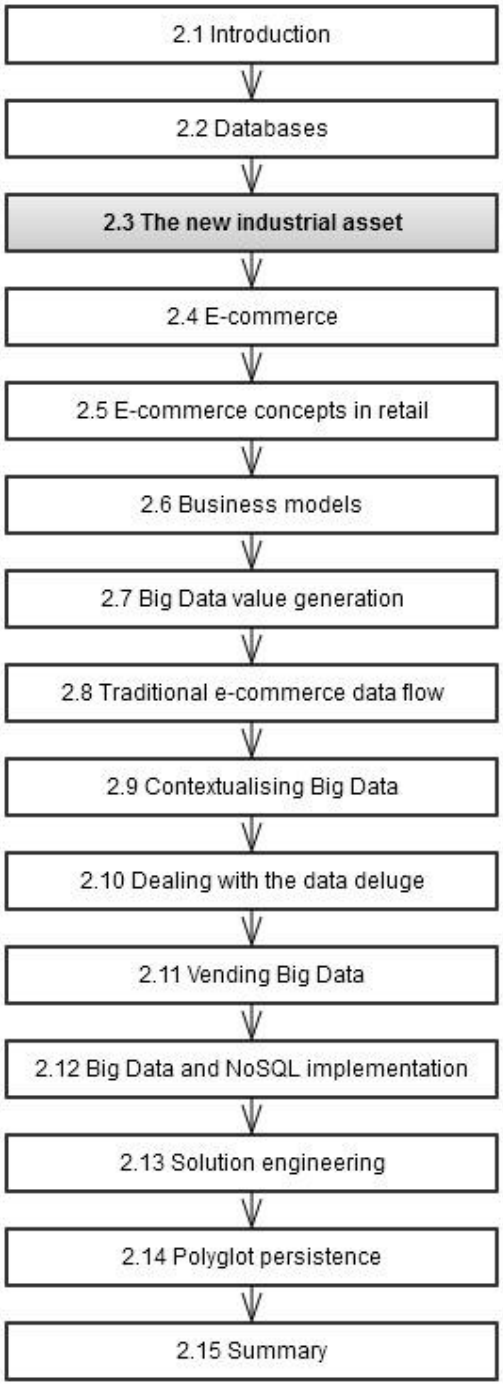


Figure 2.4: Chapter layout – The new industrial soil

Cuzzocrea *et al*. (2011) are of the opinion that large amounts of unstructured data originate from high-performance applications. These fall into a wide range of a heterogeneous family of application scenarios, with sources such as scientific computing, social networking, eGovernment applications and medical information.

According to Jacobs (2009), Big Data, in the context of information visualisation, is viewed as datasets that are too big to fit onto a screen. This means BD constitutes datasets that cannot be handled or processed in a straight-forward manner. Jacobs further states that a visible trend emerging with BD is the unchangeable fact that this data deluge forces data curators to look beyond tried-and-trusted methods that are prevalent at the time. The varied definitions of BD indicate the diversity of its applicable context and source of data. For the purpose of this research, the definitions by Manyika *et al.* (2011) and Jacobs (2009) will be adopted as primary definitions.

The working definition for this research is:

Big Data is the continuous accrual of structured and unstructured data in an organisation for the purpose of gaining knowledge with direct cognisance to the four (4) Vs.

Figure 2.5 highlights the four (4) Vs as posited by Louwers (2013).

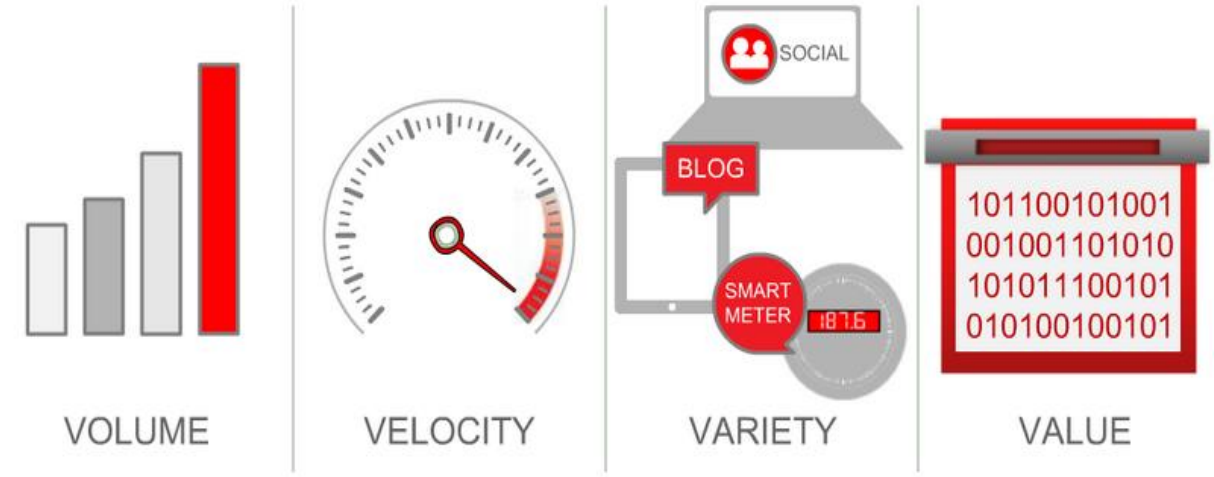


Figure 2.5: Four Vs of Big Data

(Source: Louwers, 2013)

Four underlying factors common within the domain of applications generating BD include (Cuzzocrea *et al*., 2011):

- Large scale data
- Scalability issues
- Advanced extraction-transforming-loading (ETL)
- Designing and developing easy and interpretable analytics over BD repositories for the derivation of intelligence and extraction of useful knowledge

According to Marciano (2012), data curation as described by the Graduate School of Library and Information Science (GSLIS) is:

> …the active management of data through its life-cycle involving a number of approaches and techniques including data management, digital archiving, and long-term preservation.

The process of BD curation culminates in datasets which force curators to look beyond traditional curation methods (Jacobs, 2009; Floyer, 2012). Big Data curation (or BD administration) may provide an organisation with insight through which the organisation may drive a business forward (LeBlanc, 2012). According to Forsyth (2012), the benefits of BD curation are many, just as the benefits of BD analytics are numerous and varied. Some of the advantages of BD curation and analytics that may benefit an organisation include (Chaudhuri, 2012):

- Insight from exploring text and semi-structured data
- Real-time business analytics, inferring a limited time-gap between acquisition of data and time to act
- Ability to experiment with deep analytics beyond functionality offered by the traditional business intelligence (BI) stack
- Availability of low-cost, highly scalable analytics platforms; a strategic and tactical asset, a source of knowledge for competitive analysis and decision-making in business

What is apparent is the voluminous amount of data captured to facilitate BD analytics. According to Gelber (2012), the way vendors and enterprise executives are structuring the conversation around BD is misleading, as the underlying conception of value generation, relative to BD, is whether an industry needs better data or voluminous data. Too much information is captured for the purpose of analytics; however, forecasting the future with a level of reliability without footprints in data may be questionable (Gelber, 2012). Voluminous data with quality will improve the level of reliability of generated insights. According to Manyika *et al*. (2011):

> If US health care could use BD creatively and effectively to drive efficiency and quality, we estimate that the potential value from data in the health sector could be more than $300 billion in value every year. In the developed economies of Europe, we estimate that government administration could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using BD.

Pokorny (2011) states that IT systems have for decades relied on vertical scaling which used huge expensive servers, shared nothing, and required higher levels of user-skills.

These systems were in some cases also unreliable. This situation therefore influenced data curators and organisations to move data towards a new approach to collecting and managing data (Stonebraker, 2010:10; Seeger, 2009). According to Stonebraker (2010), curators seek alternatives to advance this new approach from the perspective of performance and flexibility. A common argument from data curation proponents sometimes include: “I started off with MySQL for data storage needs and over time found performance to be inadequate as in the case of Wordnik” (Tonytam, 2010). Furthermore, according to Tonytam (2010), Wordnik experienced huge performance degradation while using a normal relational database system for operation. Figure 2.6 highlights performance degradation with increased complexity of data in data storage systems. From this diagram, it can be inferred that RDBMS have a peaked performance with structured data.

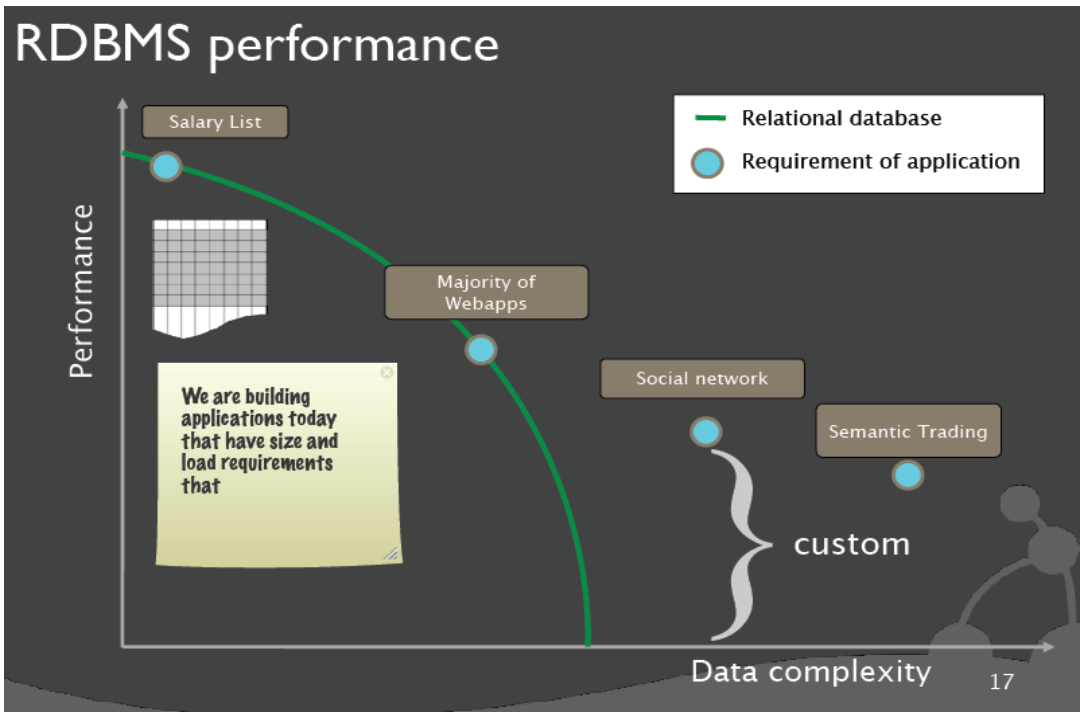


Figure 2.6: Performance against data complexity of database

(Source: Ivarsson, 2010:17)

As a result, users reverted to NoSQL databases for performance and flexibility benefits, although Pokorny (2011) rejects the idea that SQL databases do not always scale (ability to handle work as it expands). Agrawal *et al.* (2011) state that the need for applications to scale out to thousands of commodity machines and the need for schema-less data storage and access methods in applications, led to the birth of NoSQL databases. Zikopoulos, Deroos, Parasuraman, Deutsh, Corrigan *et al.* (2013:xviii) define BD as:

> …adopting new technologies that enable the storage, processing, and analysis of data that was previously ignored. It is also about the adoption of new business processes and analytics approaches that take advantage of that data.

Jacobs (2009:44) however, provides what he terms a meta-definition of BD, which he defines as:

> Data whose size forces curators to look beyond the tried-and-true methods that is prevalent at the time.

Table 2.1: More definitions of Big Data

Author(s)	Definition
(Rele, 2012)	"High volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization."
(Gualtieri, 2012)	"The frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers."
(Forsyth, 2012:5)	"Big Data applies to datasets that are large, complex and dynamic (or a combination thereof) and for which there is a requirement to capture, manage and process the dataset in its entirety, such that it is not possible to process the data using traditional software tools and analytic techniques within tolerable time frames."
Adduci, Blue, Chiarello, Chickering, Mavroyiannis *et al.* *(*2011)	"It is not a precise term; rather it's a characterisation of the never-ending accumulation of all kinds of data, most of it unstructured. It describes datasets that are growing exponentially and that are too large, too raw or too unstructured for analysis using relational database techniques. Whether terabytes or petabytes, the precise amount is less the issue than where the data ends up and how it is used."
Ramanathan and Raja (2013:53)	"The term adopted by the market to describe extreme information management and processing issues which exceed the capability of traditional information technology along one or multiple dimensions to support the use of the information assets."
Cuzzocrea *et al.* (2011:101)	"Enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from eGovernment applications to medical information systems, and so forth. Data stored in the underlying layer of all these application scenarios."
Primesberger (2011:1)	"Big Data is the increasing number of jumbo-size enterprise datasets—and all the technology needed to create, store, network, analyse, archive and retrieve them."
Mayer- Schönberger and Cukier (2013:6)	"Things that one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organisations, the relationship between citizens and governments and more."
Boyd and Crawford (2012:3)	"A cultural, technological, and scholarly phenomenon that rests on the interplay of: (1) Technology: maximising computation power and algorithmic accuracy to gather, analyse, link, and compare large datasets. (2) Analysis: drawing on large datasets to identify patterns in order to make economic, social, technical, and legal claims. (3) Mythology: widespread belief that large datasets of a higher form of intelligence and knowledge that can generate insights previously impossible, with the aura of truth, objectivity, and accuracy."

Taking a step-wise progression in data growth to accentuate and elucidate the need for different curatorial approaches, Jacobs (2009) mentions that in the early 1980s, datasets increased to the point where it necessitated the need for a robotic tape monkey to swap thousands of tapes in and out. In the 1990s, any data that transcended the bounds of Microsoft Excel and desktop PC required alternative software to process or analyse and was viewed as BD. At present, BD might be too large to be placed in a relational database and be analysed with the help of a desktop statistics package or data requiring software running in parallel on tens, hundreds, or even thousands of servers. As such, most definitions of BD seem to imply colossal data at hand.

Dumbill (2012) states that Big Data is data that exceeds the processing capacity of conventional database systems; the data may be colossal, thus implying volume needs, or moves too fast, thus implying the need for velocity issues to be addressed, or does not fit the structures of the varied database architectures. To gain value from this data, the user must find alternative ways to process it, which might be the complete opposite of existing mundane processing methods. According to Hardesty (2012), a BD definition encapsulates techniques for organising and making sense of the huge amounts of data generated from web users and networked sensors.

Rele (2012) takes a more global view to defining BD. He suggests that the “Big” should be removed from the term as it would boil down to the understanding of how to use all data to meet customer needs so they could grow their own business. Furthermore, Howe (2013) states that BD is any data that is expensive to manage and hard to extract value from—this definition hinges on the key idea that big is a relative terminology, implying ‘difficult data’. According to Howe (2013), BD is not really about size or volume; rather, it has more to do with the degree of associated challenges relative to insight generation. Howe continues by stating that the early adopters of BD collected data for a specific purpose and eventually saw this data being used for unrelated purposes other than originally intended.

The leveraging of new value BD streams initially poses a challenge to many organisations. BD, with the many varied cases and domains of applicability, has become the source of competitive advantage in enterprises since the 1980s (Forsyth, 2012; Jacobs, 2009; Schmarzo, 2013) mainly due to varied applicability and associated benefits. This is especially true with the confluence of enterprise Information technology (IT), cloud computing, social media, eScience and media in combination with mobility and emerging social trends which are re-shaping the technology industry (Floyer, 2012).

Floyer (2012) states that the emergence of BD as information source for organisations has facilitated a better understanding of customer needs, business management (BM) and

operations optimisation. This has created a form of knowledge which enables the prediction of potential customer dissatisfaction with the service when service providers fail maintain the required standards. Dissatisfied customers might not stay customers for long; if business managers were aware of customer dissatisfaction they might be able to take action to ensure customer retention. Such pertinent insights are obtained through steadily garnering and subjecting data to analytics for continuous insight generation (Forsyth, 2012). Schmarzo (2013) mentions that many organisations are putting BD and advanced analytics to use to improve procurement, manufacturing, products development, marketing, sales and store operation.

Some BD sets use cases, as mentioned by Prajapati (2013) and Sathi (2012), for which many organisations are collecting and managing BD to reap benefits. These include:

a) A 360 degree view of customer data

This is the most common use of BD with retailers (Leidwinger, 2013), as online retailers seek to ascertain what customers are doing on their site, that is, what pages they view the most, where they linger the most, when they exit and how long they stay. This data is collected as part of weblogs which is unstructured data, and when combined with structured (transactional data) and social media, it yields a 360 degree view of data.

b) Internet of Things

This is the interconnection of devices managed by hardware, sensor and information security companies (Manyika *et al.,* 2011). These are networks of low-cost sensors and actuators that collect and monitor data from these devices for decision-making and process optimisation. The significance lies in how businesses and the public sector organisations are able to manage assets, optimise performances and create new business revenue such as in monetising insights (ibid).

c) Big Data service refinery

Big Data service refinery, which Leidwinger (2013) refers to as data warehouse augmentation, is the process of increasing corporate data store efficiency by breaking down silos across data stores and sources as a way to optimise data warehousing, Take for example a global financial institution moving from next-day to same-day reporting for its corporate banking customers. The organisation with proprietary software receives data from different sources, processes the data and stores it in Hadoop. The data is then extracted for subsequent reporting and analytics in a more structured manner. This facilitates an intra-day use of data which is more agile than previous, longer frequency based (Leidwinger, 2013).

d) Fraud detection with BD

Financial institutions are turning more to analytics to predict and prevent fraud in real-time (Sensmeier, 2013). As inconvenient as this may sound, especially for customers making random purchases, it is a necessary inconvenience to prevent losses stemming from fraud. Traditionally, fraud detection has focused on detecting the use of known bad Internet protocol (IP) addresses or unusual login patterns. Fraud, according to Hsu (2013), has changed so much that it is miniscule compared to routine purchases which can be flagged as outliers. Such transactions, which are mostly small and seemingly harmless, look very much like any other transaction. Fraud detection in such instances is possible because of the interception of the transaction at the point of authorisation through the analysis of point-of-sale data, geo-location, authorisation and transactional data. The main challenge lies in surveying the various data sources.

Fraud detection software look for spikes or anomalies in transactions that may indicate and flag transactions as fraudulent, but how this is done without disrupting service to valuable customers is a challenge (Sensmeier, 2013). Integrated models and algorithms process massive amounts of structured and unstructured data from different sources to find patterns of fraud and anomalies that help predict customer behaviour. These are predominantly real-time BD architectures (RTBDA). They deploy a multi-tiered architecture which is comprised of data, analytics, integration and decision-making for prediction.

e) Big Data monetisation and business metamorphosis

Harnessing the potential of BD to yield new revenue sources in an economy saturated with opportunities, increased regulations and shrinking revenues is a position many organisations generating data are opting for (Bohé, Hong, Macdonald & Paice, 2013). Schmarzo (2013) defines BD monetisation as the leveraging of BD for net new revenue opportunities which include:

- Packaging customer, products and marketing insights for sale
- Creating intelligent products from analytics
- Leveraging actionable insights and personalised recommendations based on customer behaviour and the re-engineering of an organisation's business model

Adriaenssens (2013) discusses how monetisation may improve business performance. He mentions two methods: sell the collected data in the market or analyse data to improve business performance. Data monetisation allows curating organisations to offer service beneficiaries more value-added services such as analytics and report packages flanked with attributes such as increasing timeliness, accessibility, quality and completeness of data.

Launching a BD journey might be a huge and overwhelming task but this should not be the case as the opportunities to leverage BD are all in the footprints of curated data. According to Bohé *et al.* (2013), business managers should look inward to assess priorities and available options, and then look outward to identify market opportunities and scale up offerings to meet these needs.

A well-grounded data monetisation platform paves the way for the curating organisation to transition into business metamorphosis (Schmarzo, 2013). This is the ultimate goal for organisations that want to leverage captured insights about business elements such as customer usage or behaviour patterns and market trends to adapt business to winning trends.

Adduci *et al*. (2011:6) assert that organisations which introduce the full scope of BD management issues to their IT strategies obtain the ability to outperform competition, as demonstrated in the case of Facebook. BD enables business models, products and services to be transformed based on proven insights. According to Manyika *et al.* (2011), organisations that incorporate BD "will begin to outperform their unprepared competitors within their industry sectors by 20% in every available financial metric". Manyika also mention that these figures serve to reinforce the claim of a need to harness BD and use it to create revenue within any and every business industry.

Among Kelly's (2012) findings:

> …in a recent survey by Oracle of 300 executives, nearly a third—29%—of respondents graded their current Big Data management and analytic capabilities as D or F. Nearly all respondents—93%—agreed with the statement that their organisation was leaving revenue on the table by not doing a better job of leveraging Big Data.

It appears therefore that revenue is lost by the inefficient management of data. This therefore begs the question, how is it that value may be drawn from BD to create or prevent losses in revenue? Parise (2012) has found a way to make use of data by developing the *Big Data Framework*. The framework arranges BD into four different quadrants, first separating data into two categories: transactional (structured) and non-transactional (unstructured) data. These categories assist analysts to determine the strategy that will be implemented in mining useful information from the organisation's BD. The framework is further discussed in Section 2.10.

2.4 E-commerce

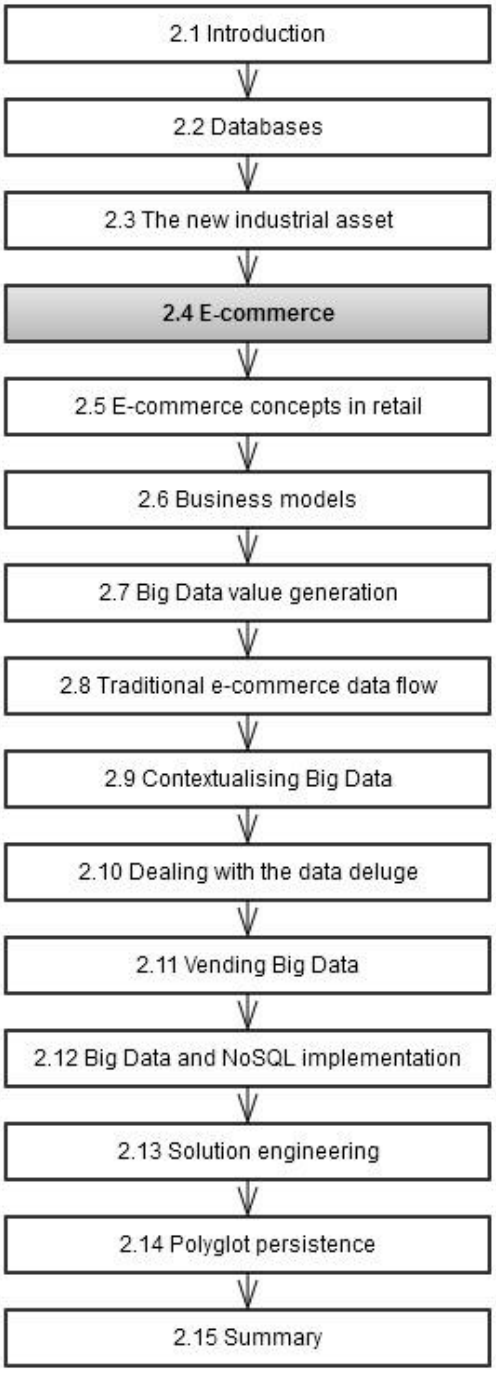


Figure 2.7: Chapter layout – E-commerce

E-commerce involves digitally-enabled commercial transactions conducted online. In the most popular sense, e-commerce can be defined as digitally-enabled commercial transactions, with people and organisations as the proponents. The viability and sustainability of Internet business and transactions originated from benefits such as reduced costs for consumers who search for services and products, and also the simplicity, speed and accuracy of price discovery (Engström & Salehi-Sangari, 2007).

For the seller, these benefits may include an extended shop front or visibility and lower cost of market entry among many distinct significant factors such as the ubiquity of Internet technology, global reach, richness, propensity for personalisation or customisation, availability of observing universal standards and above all, the opportunity to interact socially with an increased number of like-minded people. According to Engström and Salehi-Sangari (2007), business initiatives from an organisation with well-defined measurable business critical factors form the bases of e-business enablement.

Engström and Salehi-Sangari (2007) define e-business as an intra-organisational enablement of business processes and transactions excluding any value exchange, which, in other words, implies total exclusion of commercial transactions across organisational boundaries. E-commerce and e-business intersect at the value exchange point, where business links with the external environment in service exchange for value development. E-commerce extends the boundaries of the organisation by improving the business's discoverability and visibility. The Internet provides a market space that allows sellers the opportunity of reaching a wide variety of consumers with content rich presentation, limitless complexity and ability to easily surmount the perceived trade-offs relative to traditional business presentation. In a study to determine why South Africans used or will not use the Internet for online shopping, Swardt (2008) mentions several factors in favour of online shopping, namely security, reliability, convenience, ease of finding products and ability to make informed purchases. The study however mentions the inability of consumers to feel and touch products before purchase as one major disadvantage to e-commerce. Ben-Shabat and Gada (20012:8) also mention that once an a customer has a sense of a product and likes it, he or she will readily convert into an online regular shopper, implying that this disadvantage will fall away with time and especially with standards and presentation of content.

Lopez (2009:3) mentions three evolutionary phases in the history of e-commerce as being innovation, consolidation and reinvention. Lopez (2009) states that the innovation phase took place from 1995 to 2000 with the availability of rich content, and that context information was available to online buyers and sellers alike. The innovation phase progressed into consolidation where many more traditional businesses made use of the Internet with content rich information and limitless information presentation patterns and trends. The transformation phase further progressed into re-invigoration as social trends and data dissemination reached its peak. Significantly, invigoration spurred many business transformations with many businesses developing and altering business models to provision for pertinent gains thereof (Adduci *et al.,* 2011).

What is appropriate about e-commerce is the vastness of applicable disciplines, and according to Laudon and Traver (2002:39-42) and Turban and King (2003:8), this is limitless and multi-disciplined, and it is not singly covered by any one academic discipline.

Understanding e-commerce requires becoming aware of business concepts such as consumer behaviour, business models, industry structures, market, decision-making, e-commerce retail concepts, organisation and industry value chains, information, data as an organisational asset and BD as an augmenter. Davenport and Prusak (2005) suggest that people need to learn the impact it has on society as global e-commerce can have consequences for individuals concerning their intellectual property and privacy rights. Public policy issues such as equal access, equity, content control and taxation will require a level of evaluation. Further, with all the progression and emerging technologies, it has become almost mandatory for thriving businesses to change business models and strategies to remain relevant, viable, sustainable and profitable.

2.5 E-commerce concepts in retail

See Annexure C, Table 9.1, for retail concepts.

2.6 Business models

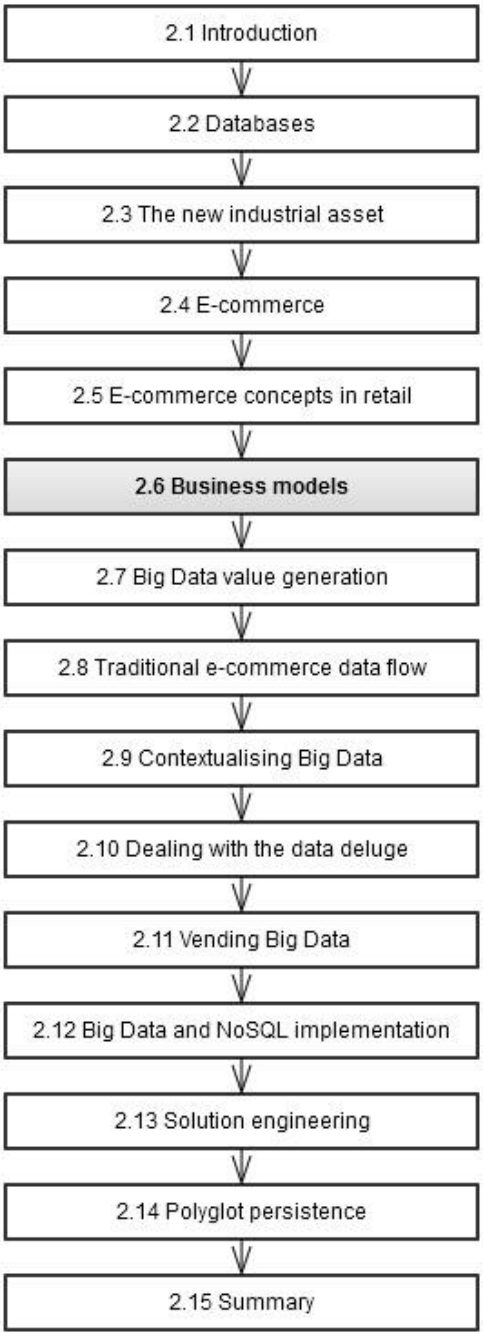


Figure 2.8: Chapter layout – Business model

A business model may imply three aspects (Linder & Cantrell, 2000:1): (i) components of business models; (ii) real operating business models; and (iii) change models. Normally in business, when people mention the term ‘business model’, the inference is components of business models, real operating models or change models (Linder & Cantrell. 2000). To be precise, a business model is the organisation’s core logic for creating value. Change logic represents the core logic for how a firm will change over time in order to remain profitable.

The business model encompasses a plan for the successful operation of a business, identifying sources of revenue, the intended customer base, products, and details of financing. Schmarzo (2013) mentions that many people have attributed the success of companies such as Facebook, Amazon and Google to the way in which they used new technologies—not just for effective and efficient operation but also in the engineering of new business models. According to Linder and Cantrell (2000:2), companies succeed by choosing an effective business model which they execute and relentlessly renew to retain their distinctiveness as fierce competitors. This process of renewing and remaining relevant is hinged on the company's ability to change their business model at a pace that matches the dynamism in their operating market space.

From an analysis of the literature (Linder & Cantrell 2000; Schmarzo, 2013) it becomes clear that business models translates to the underlying principles of value creation within an organisation—that is what an organisation does and how it does what it does to generate revenue while remaining sustainable. According to Shafer, Smith and Linder (2005), Sun Microsystems made the strategic decision to gain ground in the industry with standardised chips and software using proprietary hardware and software. This gave them considerable success in the industry as IT leaders. In a research paper that interviewed 70 executives from 40 companies regarding the company's core logic for creating and capturing value which the authors (Linder & Cantrell, 2000) describe as being the business model, it was found that these executives were confused about what the business model represented. Schmarzo (2013) aligns a business model to business processes by stating that a business model is a set of planned activities or business processes designed to result in a profit in the business marketplace. Organisations outline details of a business model in the business plan stating its value creation process, benefits and strategies. Figure 2.9 presents a diagrammatic layout of a business model as presented by Shafer, Smith and Linder (2005). The diagram shows the different sections of a business model, including strategic choices, value network, create and capture value.

The four components in Figure 2.9 allow the organisation to create a value proposition, identify sources of competition (competitors) within the market, and create an encompassing strategy to deliver and promote services and products. This process also identifies the competitive niche the business intends on exploiting in the market space, leveraging the capability of the delivery team in applying skills and strategies to deploy the outlined strategies. For example, some e-commerce companies such as Amazon have an unparalleled selection and convenience system that functions hand-in-hand with a recommendation engine for better sales promotion and upsell/cross-sell.

The successful implementation of strong and robust strategies gives any organisation a significant competitive edge (Maas, 2013). According to Maas (2013:3), well-formulated strategies may produce superior performance for organisations when successfully implemented.

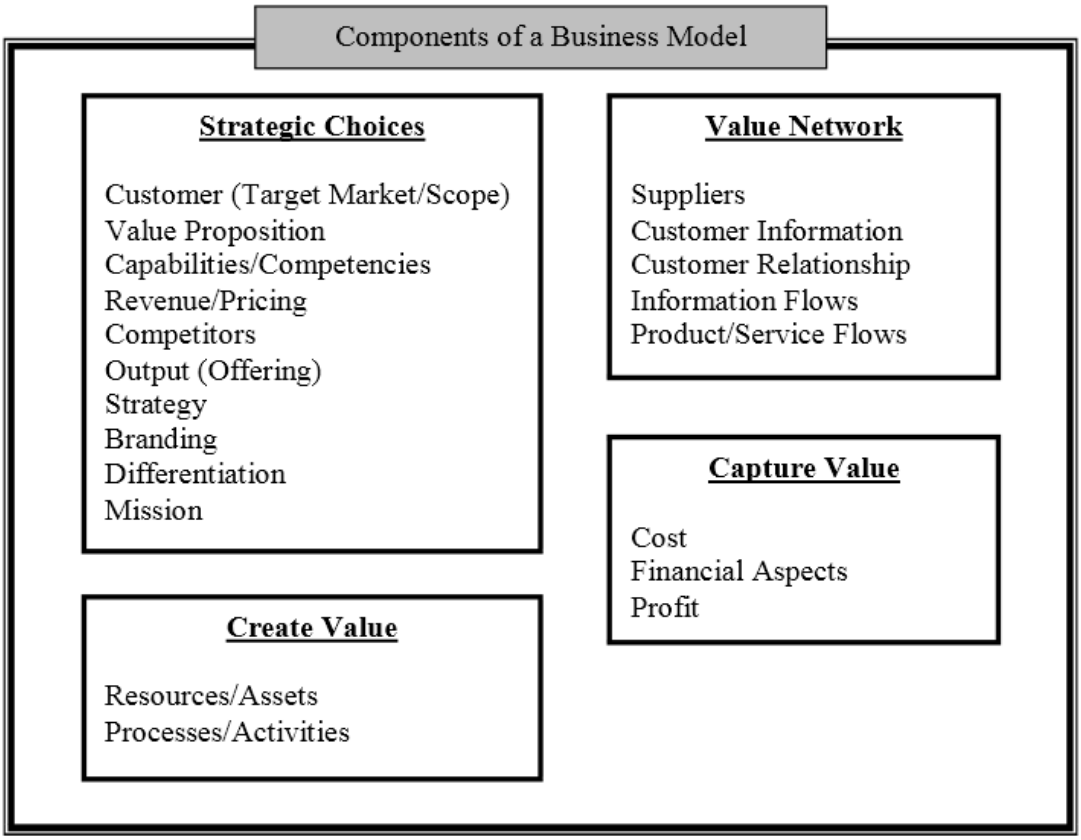


Figure 2.9: Business model components

(Source: Shafer, Smith & Linder, 2005:43)

E-commerce business models and value generation methods are varied and many. Categorising business models is virtually impossible as every businesses approach may be distinct. As there many variations of business models, overlapping trends and daily emergence of new ideas, the dissimilarity of business models will continue to evolve. These variations are also heightened by the technologies used in the core of the business model. According to Laudon and Traver (2009), the type of e-commerce technology used can also affect the classification of a business model. On a broader spectrum, e-commerce business models fall under the categories as presented in Table 2.2. As there are overlaps, emergence and diversity, it is quite common to find the same business falling under different model categories with respect to value generation. For example, eBay is essentially a consumer-to-consumer (C2C) marketplace, but also functions as a business-to-consumer (B2C) market player. Table 2.2 presents a layout of Internet business models.

Table 2.2: Internet business models

(Source: Laudon & Traver, 2009:88)

Generic business model	Description	Sample web site
Business-to-business (B2B)	Business sells products or services to other businesses	http://www.ge.com/ http://www.salesforce.com/ http://www.paychex.com/
Business-to-consumer (B2C)	Business sells products or services to consumers	www.Mbaobao.com www.kalahari.co.za www.okbuy.com
Consumer-to-consumer (C2C)	Consumer or person-to-person transaction	Gumtree.co.za www.kijiji.co.za

2.6.1 Competition and strategy

The vast amount of information available on the Internet to consumers to drive their decision-making is forcing retailers to place the consumer at the centre of their decision journeys; this is in order to uncover consumer decision influencers. Decision influencers are the factors consumers consider while deciding what and why a particular product meets their needs. Sellers research buyers in terms of consumer decision-making prior to purchase in order to expose hidden opportunities for the seller for better value creation. These opportunities for value creation provide business with insight to personalise and engineer a seamless retail experience for their consumers. Hritzuk *et al*. (2013) mention consumer influencers to include personalisation, product information availability, enrichment and validation.

In addition, Hritzuk *et al.* (2013) also highlight that consumers consider how a product fits into their life to satisfy specific needs; knowledge of how the consumer decide may help the seller personalise tools and platforms for display of products. The information to influence the consumer's decision relates what information the seller makes available to aid consumer decision-making. Adduci *et al.* (2011) refer to this as the *proposition of value* which is the concise information that appeals to the consumer's strongest sense of decision drivers. According to Wang (2014), satisfying the buyer's needs is at the core of success in any business endeavour. Adduci *et al.* (2011) assert further that value proposition is a positioning statement that explains what a seller provides to a consumer and how uniquely the service provider does this. This involves identifying the very problem the solution addresses, the urgency of the problem and the solution state of the problem resolved.

Adduci *et al.* (2011:2) define information advantage as:

> …cultivating the mindset, skills, processes and technologies to use information to operate more efficiently, increase customer loyalty, grow market share and create business opportunities that were not possible before.

As Information has become a powerful source of competitive advantage, perhaps comparable with capital assets and human talents, it has enabled organisations to revolutionise how they compete and do business. The integration of information advantage into an organisation's competitive strategy and business model fosters the engineering of an insightful and innovative global level value proposition. This allows the organisation to excel in decisive and predetermined ways favourable for development.

According to Wang (2014:34), a competitive advantage is "a key determinant of superior performance", at the heart of a competitive strategy, and achieving competitive advantage requires an organisation to make a choice in terms of how it seeks to attain, and the scope within which it will attain this advantage. A common scenario seen across industries is the concept of a strategic mediocrity that even culminates in a below-average performance (Porter, 1998). This is when an organisation decides to be all things to all people. Taking a stance as to how to attain a competitive advantage is mandatory for above average performance and gaining a competitive edge.

The business-to-business (B2B) category is broadly grouped into e-retailers, content providers, service providers and portals. Rappa (2001) however, categorises Internet business models differently though. From a broad perspective, Rappa acknowledges the constant evolution of business models to include brokerage, advertising, informediary, seller, manufacturer, affiliate community, subscription and utility. A visible trend is how a particular business based on its services may fall under different categories simultaneously, implying the overlapping concept as described by Laudon and Traver (2009). Business models relate directly to revenue generation of organisations, and most often than not changes in business models are driven by business strategic decisions to effect change in the direction of a business transaction. This is to improve their value creation which also directly affects the organisational IT infrastructure.

The next section focuses on the different forms of business revenue generation models applicable to the online market space.

2.6.2 E-commerce revenue models

According to Nikov (2012), the web provides a platform for revenue generation in the online market space by changing the basis of competition among rivals. The basis of competition and approaches businesses take to address opportunities has culminated in a myriad of business models. Proponents leveraging the strength of the chosen model propel organisations to develop strategies that allow them to take full advantage of the discovered niche for profitability.

Nikov (2012:65) indicates five factors that are changing revenue generation in the online market space, namely “the barriers to entry, strength of suppliers, bargaining power of buyers and also the threat of new substitute products”.

Table 2.3 describes Internet revenue models and sample sites used in the online market space.

Table 2.3: Revenue models for the online market space

(Source: Zarrella, 2010)

Revenue models	Description	Sample web sites
Advertising revenue model	Paid advertising to derive profit, it hinges on sizable viewership or niche market viewership through segmentation.	www.yahoo.com
Subscription revenue model	Restrict content based on subscription fee for revenue generation.	www.ancestry.com www.yahoo.com
Sale revenue model	Profit generation comes from the sale of goods, information and or services.	www.gumtree.co.za www.ebay.com, www.amazon.com www.aliexpress.com
Sale revenue	Profit generation comes from the sale of goods, information and or services.	www.aliexpress.com
Affiliate revenue model	Derive revenue or a percentage of the sale from steering traffic that ends in sales.	www.amazon.com

The advertising revenue model derives profit from delivering paid advertising to a segmented group of users on the premise of relevance to a niche market group. The aim is to segment users for marketing while convincing advertisers of marketing to the sought after segment. The Affiliate revenue model operates on the premise of steering traffic from affiliates. Referred traffic that ends in a sale attracts profit which is a percentage of the sale or a referral fee. The other models are descriptive as per the Table 2.5. The next section focuses on Internet marketing models to drive profitability and sales.

2.6.3 Internet marketing models

Internet marketing refers to advertising and marketing efforts that depend on web and email to derive and drive sales from e-commerce. Many organisations use Internet marketing and online advertising in conjunction with traditional advertising to augment sales. Some of the traditional advertising types include radio, television, newspapers and magazines. According to Zarrella (2010:21), Internet marketing can be divided into:

- Web marketing
- Email marketing
- Social media marketing

2.6.3.1 Web marketing

Web marketing, as mentioned by Zarrella (2010:30), comprises of e-commerce websites, affiliate marketing, online advertising on search engines, promotional informative websites, and organic search engine results derived from search engine optimisation. Web marketing is similar to traditional marketing with operations aimed at improving the organisation's visibility through exposure and communication. It focuses on the five marketing constants which are people, price, product, place and promotion. According to Alt and Zimmermann (2001), target market knowledge is crucial and gaining knowledge about the target market starts with a thorough analysis of the current customer base. It is also essential to ascertain the Internet accessibility profile of the target market which encapsulates the technological capabilities of the market target group, means of access, attitudes and psychographics as these all affect marketing.

2.6.3.2 Email marketing

Email marketing consists of promotional marketing and advertising which is achieved through email messages both to current and prospective customers. GroupOn is a company with a business model based on emailing and advertising.

2.6.3.3 Social media marketing

Social media marketing focuses on marketing and advertising efforts based on social networking sites such as Facebook, YouTube, Digg, Instagram and Twitter. Some of the marketing efforts here may be referred to as viral marketing. According to Chiu, Hsieh, Kao and Lee (2007), viral marketing is an online marketing technique that allows electronic information dissemination about the organisations' goods or services from one Internet user to another. This is to encourage exponential growth in the exposure of messages (Eckler & Rodgers, 2010).

2.7 Big Data value generation

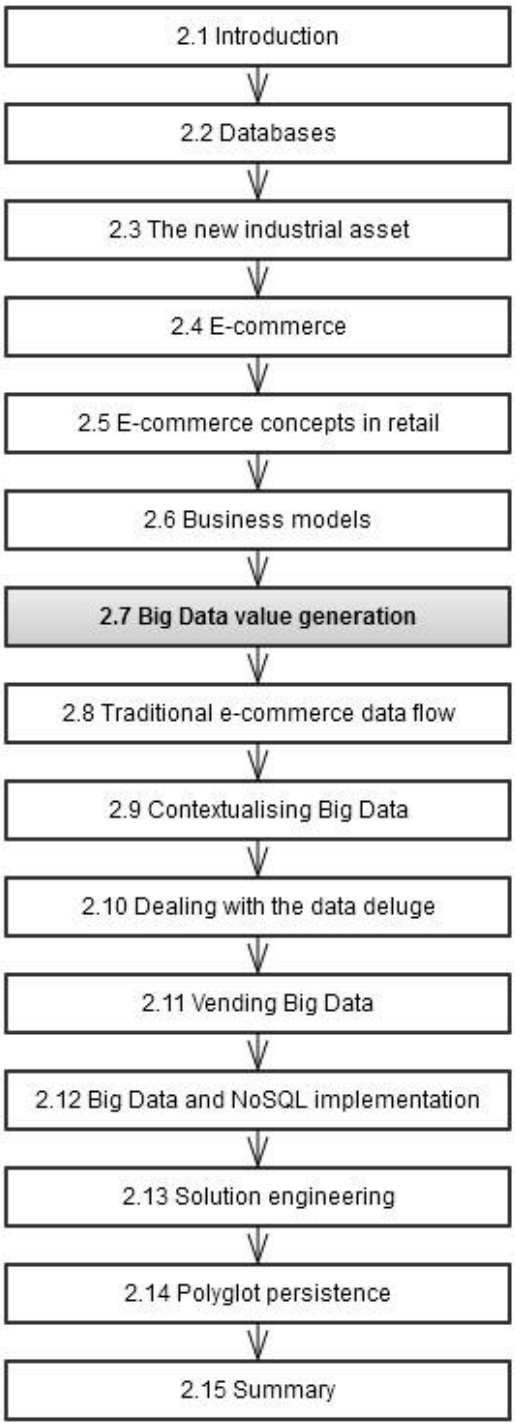


Figure 2.10: Chapter layout – Big Data value generation

Organisations ask questions of data to investigate revenue opportunities that may drive business transformation for sections such as sales and marketing. These questions are enabling and risk mitigating, if not averting or circumventing, as questions are designed to furnish the organisation and decision-makers with answers.

These responses may steer the business into a profitable situation and sustainability, promoting growth and competitiveness (Schmarzo, 2013). The answers sought by questioning data affect the many different facets of a business such as products and procurements, customers and profiling, marketing, sales, customer support and even the entire business model as the business model encapsulate the value creation process. For example, an organisation may ask questions such as:

- Who are our most important customers?
- What are our most important products?
- What are our most successful campaigns?
- What are my best performing channels?

The first question refers to a classification of customers based on certain criteria such as purchases (spending power) and loyalty to the brand or organisation. The answers obtained provide the organisation with insight that could be significant to improve customer experience. Improvements become even more visible in the form of optimised business processes and improved metrics or critical success factors. This is because actionable information, gained from leveraging curated data, is put to use to effect organisational change. Mosley *et al.* (2009:91) state that managing data and information as tangible assets in the economy is significant and relevant for modern economies, and any organisation which relies on data and information as the driving force or crux of survivability, especially when seeking to remain relevant in a chosen sector of business, must curate data as a leading asset and an integral part of the business model.

Curating data as a leading tool has resulted in many organisations' business model transformations which have enabled them to tap into opportunities.

According to Nickols (2012:1), the kind of problems addressed in complex organisations:

> is usually affected by indirect change, you don't change the problem, rather you change something else and the state of the problem changes as a result, that is you change something "here" to culminate desired results "over there".

This process is analogous to the unfreeze-change-refreeze model inferred by Kaminski (2011), who takes on a more simplified approach to implementing change. He explained his theory through the use of defrosting a large ice cube, changing it into what is desired, for example a cone, and then refreezing it. This process is iterative until desired results are obtained.

The commerce view of data provides a 360 degree view of customers. Knowing what customers' intents are through browsing patterns may allow a seller to propose items through site personalisation and email target marketing. On a wider scope, the customer data in relation to data collected from social media sites may enable businesses to determine the range of influence and degree of advocacy for each customer. The data when aggregated helps calculate the range of influence. Hritzuk *et al.* (2013) put forward that retailers place consumers firmly at the centre of their own decision journeys in order to reveal influencers along the retail path to purchase, while revealing opportunities in the form of seamless and personalised retail experiences. Sathi (2012) refers to this process of placing the customer in the heart of decision journeys as significant for profitability.

According to Schmarzo (2013), BD enables sellers to answer questions, allowing them to fine-tune and accelerate their skills in identifying sections for value generation and what specific business processes could be optimised or leveraged for business value creation. Manyika *et al*. (2011) propose BD as the foundation for progress and innovation for the foreseeable future as it propels an intelligent future and economy of smart items including smart cars, smart buildings, better education, even smart apparels, productivity gains across the economy and much better ways of championing customer experiences or interactivity.

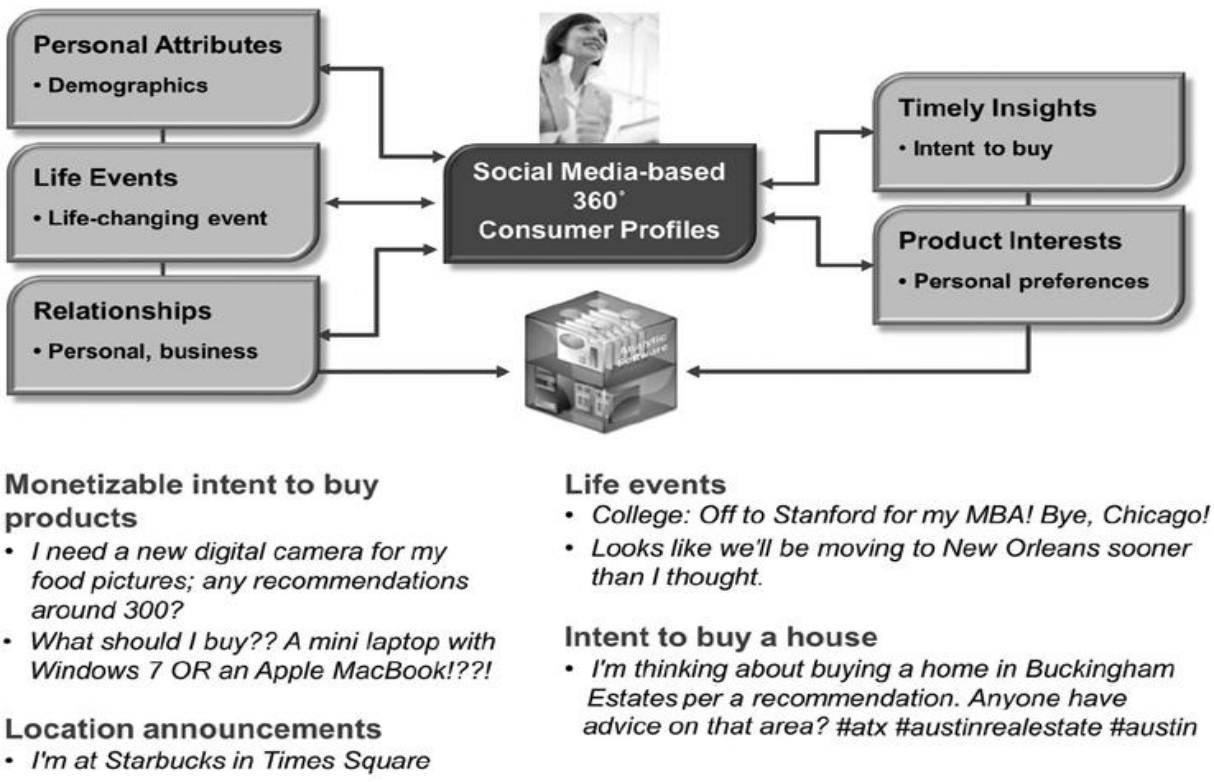


Figure 2.11: Commerce view of the customer

(Source: Sathi, 2012:42)

Schmarzo (2013) mentions four BD business drivers applicable to business, including customers, products, operations, markets and many more areas to improve decision-making and value generation. These value drivers include structured data, unstructured data, data velocity and predictive analysis.

2.7.1 Structured data

Structured data are datasets whose inclusion in a database are seamless and easily searchable using straightforward algorithmic search tools and processes due to the degree of organisation (Ganis, 2013). Access to structured data enables a high degree of trustworthiness to the questions business may ask to facilitate decision-making. According to Schmarzo (2013), the insight gleaned from structured data can be metaphorically related to the lowest hanging fruits for most organisations, implying the results of efforts coupled with decisions and strategy. Structured data is synonymous to transactional data or business data. Structured data is mostly stored easily in schema-based databases as the forms are of a high order.

2.7.2 Unstructured data

A major business transformer is the ability to derive insight from structured data coupled with unstructured data brought together for analysis. For example, customer comments and returns provide insight that could be leveraged to find faults and improve customer service and order problem resolutions. According to Ganis (2013), the lack of structure makes compilation of unstructured data a time and energy-consuming task. Steiner (2009:1) defines unstructured data as:

> …a machine or human generated information where the data do not easily conform to standard data structures (such as rows and columns) and where the understanding of data is not readily accessible without machine based or human intervention.

Mohanty *et al*. (2013) define unstructured data as information that either does not have a pre-defined data model or is not organised in a pre-defined manner.

2.7.3 Data velocity

Davenport (2013) mentions that the velocity attribute of BD involves streams of data, structured record creation, and availability for access and delivery. Velocity refers to both how fast data is being produced and how fast the data must be processed to meet demand. The ability to provide real-time data access offers an organisation many profitable opportunities.

One of the greatest benefits of mobile technologies is its ability to provide extremely precise, real-time, geolocation information. This information is pivotal to the organisation's ability to influence customer decisions from a seller point of view as insights originating from low-latency data access.

2.7.4 Predictive analytics

Predictive analytics is a section of data mining that focuses on the extraction of information from data to forecast envisaged behavioural patterns and trends with emphasis on prediction. Data mining, on the other hand, allows for exploratory data analysis with little or no human interaction, deploying computationally feasible techniques to accentuate on interesting but previously unknown structures. Predictive analytics introduces new sets of verbs into the analysis and data leveraging environment. These verbs enable sellers and data curators to ask new questions that bring across a new dimension of insight generation. Some of these verbs include predict, forecast, score, recommend, optimise, classify, estimate and cluster. Similar to identifying customer behavioural trends, Hritzuk *et al.* (2013) point out the need of consumers in the context of championing decisions within the domain of consumers' purchases, including personalisation, information availability, enrichment and validation.

Personalisation is encapsulated in the form of a consumer asking how this product may fit into their life. In this context, the consumer is pondering what might work and what the product or service would do for them. The organisation needs to strike a balance to tailor a solution to meet their consumers' needs, products range composition (by analysing market baskets to promote purchase), yield, cross-selling and upselling (to push profitability). Information centres on the question, "how do the facts stack up?" The sellers respond by getting more information about products, including features and benefits. One of the benefits of the e-commerce market space is the availability of information for consumers to compare and make decisions. The seller providing adequate information and capabilities of the product will facilitate closing a sale. Enrichment is attained when the customer has a better tactile sense of the product or service being considered. Ultimately, how does the customer feel after purchase? Business needs to be sustainable. The sustainability happens through repeat selling which implies gaining and retaining customers. According to Villars and Vesset (2014), BD analytics is big business that spurs value generation but cutting through the hype requires being able to collect, store and analyse more granular information about products, people (customers) and transactions.

The concept of BD value generation is about the use of real-time and historical data to promote the business environment while optimising profitability. There are many online tools

that may facilitate the BD journey, ultimately to the point of compressing the time to value generation such as big sheets by IBM and Googles Dremel (Mohanty *et al.,* 2013). This value generation process, as mentioned by Villars and Vesset (2014), is influenced by the collaboration of tools available for data and insight generation; insights gained through data analysis may be used to shape system design to promote and improve system usability. According to Șaovă and Raduteanu (2013), a heat map is an important tool that provides feedback on customer activity which may direct the design team on improvement measures for the website. Schmarzo (2013), on the other hand, proposes brainstorming available data and new sources of data through the creation of possible use cases. The then use cases include interrogations about where and how to leverage BD from new sources such as customers, products and operational data, to augment advanced and predictive analytics to power targeted business initiatives. Mohanty *et al.* (2013) propose BD curation through a three-prong approach as indicated in Figure 2.12. That is, the diagram identifies where BD can change the game, build future state capability scenarios and define benefits and roadmaps.

Figure 2.12: Big Data journey roadmap

(Source: Mohanty et al., 2013:20)

According to Mohanty *et al*. (2013), the BD journey starts with identifying sections of business where BD can be a game changer, defining what benefits can be canvassed and what possible strategy or plan the organisation may follow to reach desired goals. Strategies to accomplishing BD goals are numerous, and they all hinge on the concept of curating data to generate insight for desired business transformation. In contrast, Șaovă and Raduteanu (2013) propose the concept of a heat map as being a process to leverage customer desirability to improve systems or redesign their systems for achieving gains as a subset of data visualisation. Figure 2.12 illustrates heat maps on a seller website depicting user activity concentration through colour codes on the site.

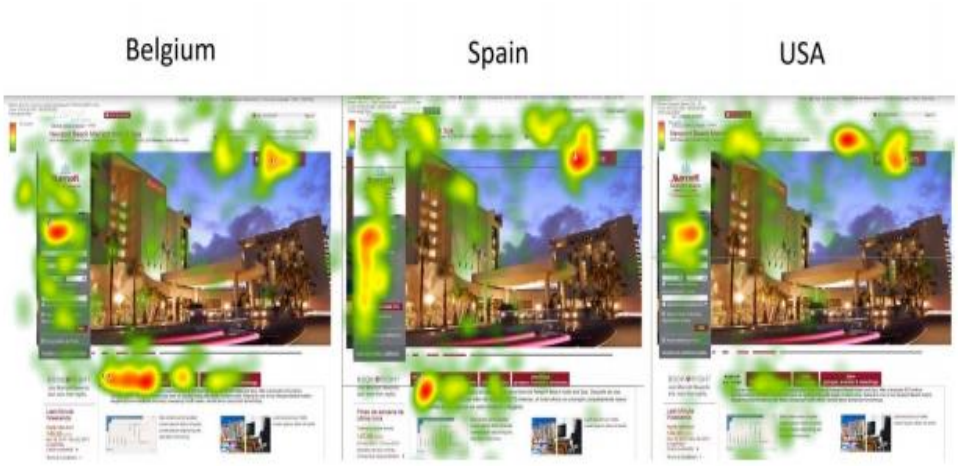


Figure 2.13: Heat map of hotel home page prototype

(Source: Keller, 2011:1)

In a study by the International Usability Testing Partnership (IUTP) and Marriott International to elicit differences in user behaviour patterns between countries, cultures and gender groups, it came to light that culture had an evident impact on user behaviour and expectations (Keller, 2011:4). Language translation did very little to meet customer requirements of users from different countries. The evidence showed that to avoid losing customers, it is important to design culture specific web pages to satisfy different user needs. The researchers created a one-page hotel website prototype that tested 510 users from 17 different countries. The aim was to track clicking behaviour patterns and focus duration of different website areas. Prajapati (2013) highlights that the web server logs provide information about web requests, such as URLs, date, time and protocol. From this information, it is possible to ascertain peak load hours of the website from web server logs and scale configuration based on traffic to the site.

Web analytics with website statistics, according to Prajapati (2013:60), is a rich information source about the visitor's metadata, including the source, campaign, visitor type, location, keyword search, pages requested, browser, total time spent on pages and based on site configuration where the customer's mouse focuses the most. To augment these analytics and to gain value, analysers not only have to click through metrics but also have to ascertain customer preferences and purchasing patterns by asking the data a series of questions that normally reveal latent obscure patterns.

2.8 Traditional e-commerce data flow

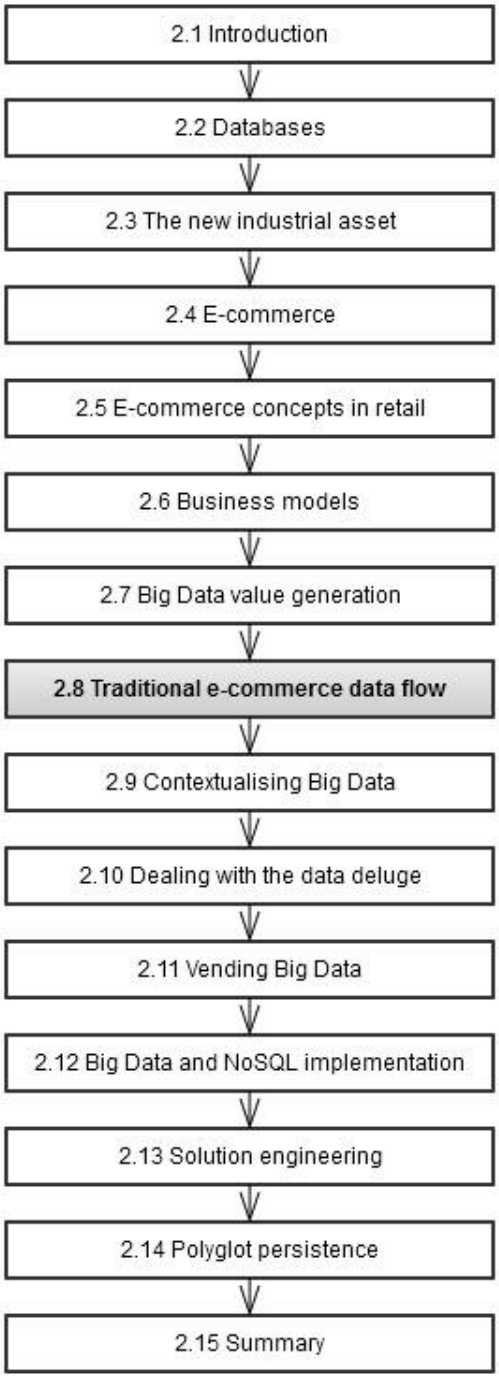


Figure 2.14: Chapter layout – Traditional e-commerce data flow

About a decade ago, an e-commerce website in operation traditionally would have a form of database persistence layer to capture data such as user information, new purchases, new products and stock-keeping units (SKUs) entered for display from inventory. By default, this data would be entered into a relational database by company employees.

The data is further ported into a data warehouse for varied operational reasons, including reporting, decision-making and to provide a general overview of the businesses performance at a given time. According to Sen (2012) and (Prajapati, 2013), as patronage and competition increased over time, sellers sought the opportunity to integrate user behaviour for site optimisation by collecting and processing weblogs and performing web analytics with website statistics based on visitors metadata, including browser information, visitor location, keyword search, requested information and total time spent on a page.

Sellers started capturing weblog data (for example clickstream) for the value vested therein (Rho *et al*., 2004). Grace, Maheswari and Nagamalai (2011) propose that web browsers and web servers communicate using the stateless hypertext transfer protocol (HTTP). In the header section of an HTTP request, a web server records the attribute “value pairs” in the log file; technically the log file has fields with information about each browser request to the server.

According to Grace, Maheswari and Nagamalai (2011), weblog files are rife with insight about user activities, and by combining what can be inferred from this data coupled with how the information was obtained, the user behaviour or customer activity can be analysed. For example, a consumer on the Amazon website who intends on buying a book peruses the site looking at different books and then ultimately decides on a particular book to purchase. The web servers capture the footprint of the customer’s every negotiation on the web page which can be used as leverage to suggest other related items.

Sen (2012) indicates that colossal amounts of data generated are from the logs previously thrown out, but operational intelligence organisations scoop some of this data for insight generation through traditional extract-transform-load (ETL) processes. Sen (2012) further highlights that this has changed in recent times with storage cost plummeting, and as such, operational infrastructural advancement organisations are storing more of this type of data. The aim is to answer questions that sellers could previously not answer.

The data available to businesses for use is not only about logs; it is also concerned with the general concept of new data sources which are generally very non-traditional, large in volume and flanked with attributes such as fine granularity and velocity (Sen, 2012). The possession of the data provides a curator with information advantage and spurs competition as curators also become more prepared for economic pressures through the use of new analytics technologies to answer questions for better business insights.

According to Sathi (2012), this tsunami of data hitting organisations is a disruptive force in that leading organisations are curating the data for their advantage. Start-ups and other organisations appreciate the need for curation but these organisations most often than not are hindered by lack of know-how or needed technology provision.

Many organisations are aware that leveraging BD could change the game for organisations, especially for businesses seeking to answer questions such as:

- What are other organisations doing with BD?
- How do we build a strategic plan for BD analytics?
- How does BD change our analytics architecture?

According to Mohanty *et al.* (2013) BD is now more than a marketing term as many organisations are assessing ways to make better business decisions using data. The applicability of BD in industry is varied from web 2 companies such as Facebook and Google to fortune 500 companies. Mohanty et al. (2013) continue to mention that many of the success stories have come about from companies enabling and creating data services with analytic innovations and integrating a culture of innovation to create and propagate new database solutions to enhance existing solutions. These enhancements complement traditional and old methods that already exist.

2.9 Contextualising Big Data

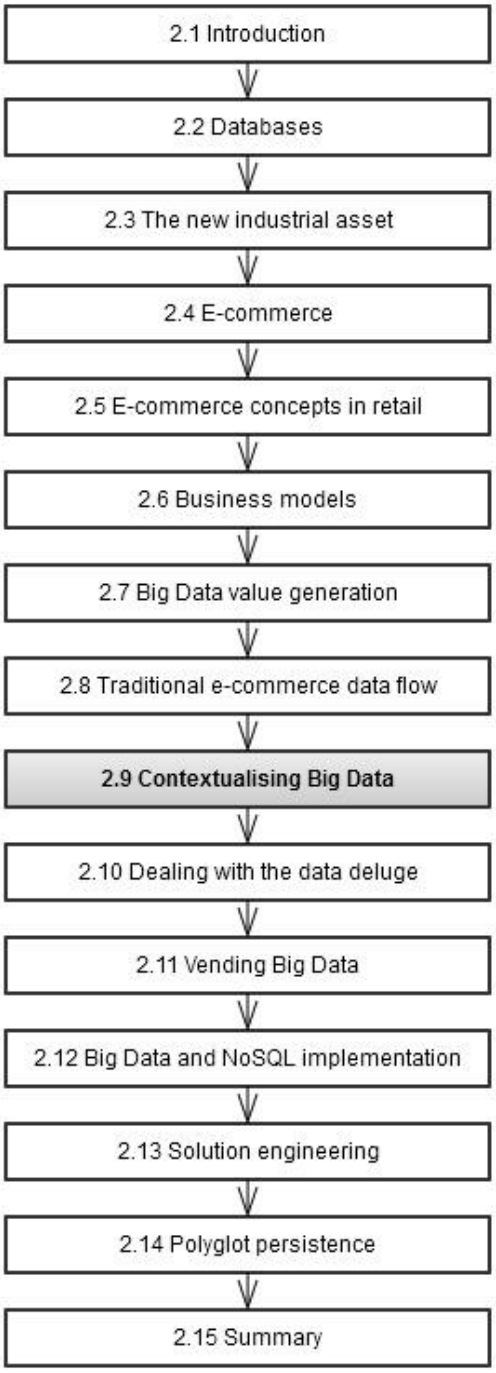


Figure 2.15: Chapter layout – Big Data in context

Zikopoulos *et al.* (2013) indicates analytics as one of the core benefits of BD curation. Without mining the data for its value it has no use, but with the help of analytics its importance becomes apparent. This is true for small and medium enterprises, as according to Zikopoulos *et al.* (2013) and Angeles (2014), small companies have Big Data, and companies have been analysing their existing data, and Big Data sizes with varied sources and forms simply take that analysis to a higher level (giving the business the opportunity to

broaden curation and use of data). The adoption of analytics has caused a split in the industry into two groups: the leaders and the followers. The information BD offers to an organisation with the capacity to make good use of it (through analytics, data collection and integration processes), gives the firm a competitive advantage over organisations that do not have this capacity. In a nutshell, the quest for BD is directly attributable to analytics, which has evolved from being a business initiative to a business imperative (Zikopoulos *et al*., 2013).

In a collaborative study carried out by IBM's Institute of Business Value and MIT's Sloan Management Review (LaValle, Hopkins, Lesser, Shockley & Kruschwhitz, 2011:2), in which their findings are documented in a paper entitled *Analytics: The new path to value: How the smartest organisations are embedding analytics*, the following results were published:

- Top performing businesses tended to apply BD analytics across their organisation in comparison to their lower performing counterparts who tended to rely on intuition
- The overall top business challenge identified by the majority of the respondents (61%) was that of achieving innovation to drive competitive differentiation

A study by LaValle *et al.* (2011) shows a link between performance and the "competitive value of analytics", bringing the authors to the conclusion that:

> …organisations who strongly agreed that the use of business information and analytics differentiates them within their industry were twice as likely to be top performers as lower performers.

LaValle *et al.* (2011) also distinguish between organisations according to the capabilities they hold when it comes to leveraging BD analytics. These capabilities are classified into aspiration, experienced and transformed, based on the scientific nature and adaptability of data by the organisation to support processes and decisions. *Aspirational* relates to making the least use of analytics and possessing the fewest of the resources (people, processes and tools) that would enable them to take advantage of BD analytics. The *experienced* organisations would be those who have automated their processes and whose objectives have moved past cost management to pursuing organisational optimisation, and who have some experience in the use of analytics. Lastly, *transformed* organisations are those that have all their resources (people, processes and tools) in place, are fully automated and have successfully used analytics to optimise their organisation. Transformed organisations are at a stage where they are focused on "driving customer profitability and making targeted investments in niche analytics as they keep pushing the organizational envelope"; they were three times more likely to outperform their aspirational industry peers (LaValle *et al.,* 2011). Schmarzo (2013) infers to this phase as transforming and in-business metamorphosis.

According to Manyika *et al*. (2011), ignoring such insightful sources of data means:

> …missing out on content from over 650 million websites on the Internet, 340 million tweets per day, 30 billion pieces of content posted to Facebook monthly; over 150 million LinkedIn users, and much more; at 40 per cent annual growth.

This reflects the sheer volume of information that is available to organisations. A possible limitation to the ability to collect, convert and analyse BD could be the potential cost and need to invest in additional expertise. But, it is possible that the value of harnessing this information could far outweigh the cost of financial, technical and human capital invested (Manyika *et al.,* 2011). A construct called the *Big Data Framework*, developed by Parise (2012), enables the mining of BD for the benefit of an organisation. But, Parise (2012) states that one of the challenges of statistical and analytical techniques employed to gain information from unstructured data is that there is a "shortage of qualified technical data personnel with expertise in the latest business analytical techniques". In addition, they infer that privacy issues may arise where consumers may feel that the methods of obtaining this information are equivalent to "spying", implying ethical consideration issues.

In an article, Duhigg (2013) documents how Target (an American retail company, founded in 1902), used a pregnancy prediction model to create advertisements that would influence customers to extend their shopping beyond baby products to other non-baby products that the brand also sold.

However, Duhigg (2013) indicates the upset that this caused consumers when they discovered that their spending habits were being closely monitored to enhance the company's marketing campaign. Thus, in order to silence these concerns, the organisation took to "including advertisements that were not baby-related so the baby ads would look random" (Parise, 2012).

2.9.1 Ethics: Big Data privacy

Davey (2013:2) posits that:

> …the ordinary consumer has a right to products that are produced and marketed ethically, sustainably and transparently. Secondly, communities have a right to manage and protect their own environment and natural resources. Thirdly, society's mandatory need for business models that put people and the planet first instead of the normally occurring short-sighted quest for monetary gains that propels business operations, these form the core principles of the existence of the Centre for Media Justice.

Parise (2012) asserts the privacy concerns of consumers as an issue that needs to be addressed, especially as data curation for insights about consumer shopping behaviour

remains a trend that will continue. From an economic stand-point, business needs the data collection to be competitive, except this happens at the expense of the consumer's privacy.

Walmart, a market leader in retail, transformed their business model by leveraging the benefits of curated data to become a market leader (Schmarzo, 2013). Davey (2013) proposes that Walmart's comprehensive analysis efforts with BD or colossal amounts of information about consumers in complex ways, allowed them to track consumers on- and offline especially through mobile technologies. Walmart's CEO, Bill Simon, affirms this in a Walmart 2013 September report (p.2) by stating that "our ability to pull data together is unmatched". According to Davey (2013), the fundamental rights of the customer are to control their personal data and how the data is used, by giving them the right to know how the data is being collected and used, by being transparent, by giving the customer choices and most importantly by being fair.

Crawford and Schultz (2014) describe a phenomenon called the "predictive privacy harm" which emphasises the real threat of a possible disclosure that could emanate from predictive analytics with unforeseen consequences. The authors mention that the potential for predictive modelling from BD, and the possible unintended consequences could be devastating to society. Crawford and Schultz (2014) cite the case of a retailer's analytics system identifying a high school-age female who fell pregnant. The child kept the pregnancy from her father, however, the father eventually realised the situation through the post without the daughter's disclosure. According to Crawford and Schultz (2014), such a situation could be devastating. On the other hand, Davis and Patterson (2012) argue that BD is ethically neutral as technology offers the ability to connect information and innovate new products and services for both profit and greater social good. Davis and Patterson (2012) refer to BD as being ethically neutral and they argue that BD does not come with a value framework.

BD also does not come with an in-built perspective of what is right or wrong. However, Davis and Patterson (2012) argue that ethical concerns become polarised in that they are targeted as good, bad, right and wrong.

According to Davis and Patterson (2012), the sheer size and omnipresence of BD technology is forcing new questions into play about identities, evolution of personal privacy, what it means to own data, and how online data trails influence reputation both online and offline. BD product design, sales, development and management actions expand an organisation's influence over individual lives in ways that may change privacy, reputation, ownership and identity.

Davis and Patterson (2012) also mention that the solution to this dilemma lies in organisations being responsible, meaning that:

> …a responsible organization is an organization that is concerned both with handling data in a way that aligns with its values and with being perceive by others to handle data in such a manner.

2.10 Dealing with the data deluge

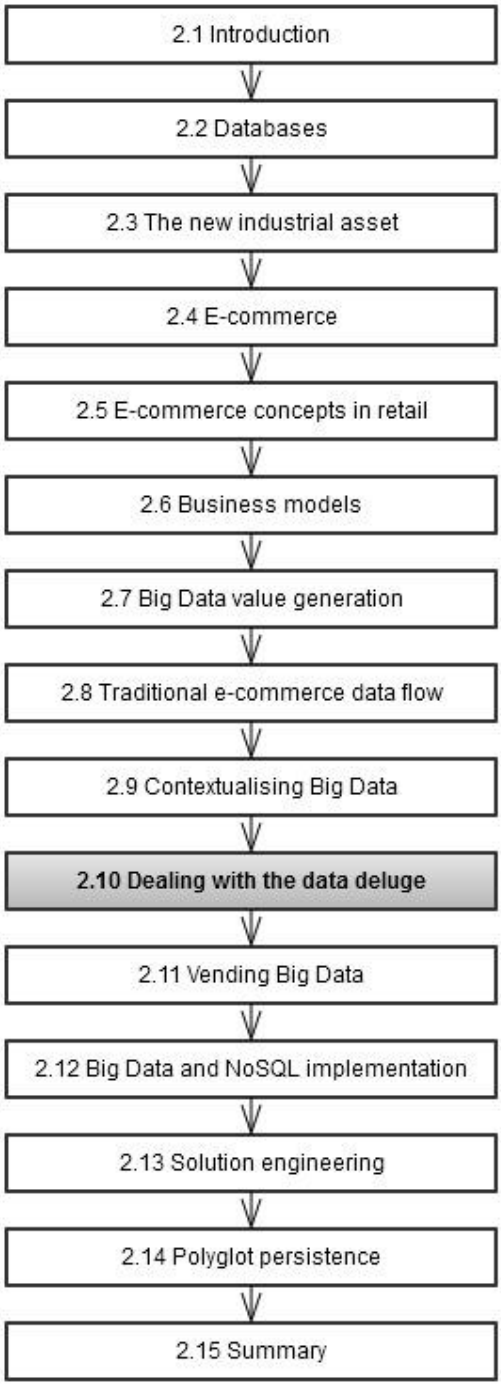


Figure 2.16: Chapter layout – Dealing with the data deluge

Inktomi Corporation, an American company based in California which provided software for Internet service providers, was a market leader in the provision of search results (Vaish, 2013). It displaced AltaVista as the leading web crawler-based search engine and was later succeeded by Google. Inktomi's software was incorporated in the widely used HotBot search engine which was perhaps one of the first true search engines. The limitations of traditional RDBMS became eminent especially relative to scalability, parallelisation and cost demands on traditional data processing, specifically taking into consideration the cross-referencing of data as compared to the chunked, transaction data normally fed to RDBMS. The most affected organisations were the likes of Yahoo, Facebook, Twitter and Google. To overcome these challenges, Google for example, had to create a whole ecosystem of computing which comprised of:

- Google File System (GFS): Distributed file system
- Chubby: Distributed coordination system
- MapReduce: Parallel execution system
- Big Data: Column-oriented database

Parise (2012), from a conceptual approach, proposes that the *Big Data Framework* allows a curator to visualise data as per Figure 2.17. The framework enables BD to be arranged into four different quadrants, with emphasis on a layered categorisation: transactional (structured) and non-transactional (unstructured) data. These categories assist analysts to determine the strategy they need to implement to mine useful information from the organisation's BD.

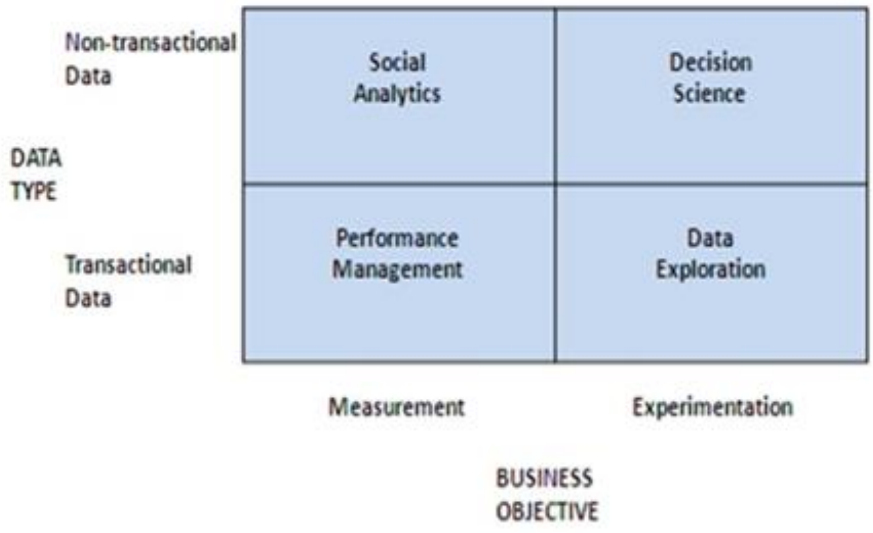


Figure 2.17: The Big Data framework

(Source: Parise, 2012)

The *Big Data Framework* employs four strategies to sort through BD, as seen in Figure 2.17. These are performance management, data exploration, social analytics and decision science. Within this framework, the strategies that most organisations are currently concerned with are decision science and social analytics because the data that falls into these categories possess the qualities of being both non-transactional and voluminous in character. Pederson (2012) says that 90 per cent of the Internet exists as unstructured data, comprising of PDF files, message boards, tweets, entire web pages, emails, MSWord documents, audio-visual formats, and other forms of content that do not have fields identifying specific data points. To make use of unstructured data, the data must be "deconstructed then enriched with metadata—transforming the data from unstructured to semi-structured before it can be pushed through analytics" Pederson (2013:3). Structured data makes up the remaining 10 per cent of BD and transactional information can be retrieved for immediate meaningful use within an organisation. This data is captured through the normal course of a business's operations, for example through sales, and thus these transactions are in a database that has a structure or schema, described as transactional data. This structured data denotes then, that in its standard form it requires no or little deconstructing before pushing it through analytics.

"Social analytics measure three broad categories: awareness, engagement, and word-of-mouth or reach" (Parise, 2012). Social analytics makes use of data which exists on social media platforms. Social media data combines with non-transactional data to measure success among consumers with regard to an organisation's external and internal social digital campaigns and activities. Depending on the consumers' responses, managers are able to target their campaigns to their audience more accurately. Parise (2012) gives the example that "low Facebook engagement may mean more interesting and interactive content needs to be created". In addition, Parise (2012) mentions how recent advancements in social measurement techniques allow companies to track their digital footprint in the social media world, with companies such as Peer Index and Klout possessing the capabilities to measure a digital user's social influence.

The Decision Sciences Strategy enables a company to measure consumer sentiment about its product offering; however Parise (2012) cites that:

> …while technology has helped companies scale the listening process involving social Big Data, the accuracy of listening tools is nowhere near perfect. Manual work is needed to "train" these technologies on company- and industry-specific keywords with regard to textual and sentiment analysis. Another good practice is to initially do parallel manual and listening tool analysis to understand the accuracy of the tool and determine ways to improve its effectiveness.

The challenge therefore, with analysing BD through the Decision Sciences Strategy, is that despite it being an automated process, bypassing the required manual work will only diminish the efficiency of listening tools. Parise (2012) points out that it is most beneficial to integrate the different BD strategies rather than to implement a single strategy in obtaining value from the big dataset (structured and unstructured).

2.10.1 Operationalising Big Data

Organisations have for many decades used business intelligence to derive insight for competitive advantage through the application of analytics, albeit mainly to structured data in relational databases. Drawing insight from unstructured data such as videos, audio, clickstream, text, weblogs and email is different in comparison to structured data due to the new set of challenges the data presents, for example the volume, the rate at which the data is generated and captured and the disparate forms of the data.

This is different for BD, by characterisation, due to the new set of challenges this data presents, such as the volume of data, the rate at which data is being generated and the different forms of the data, including videos, audio, clickstream, text, weblogs and email. This phenomenon is described as outperforming the capabilities of traditional storage and analytic solutions, whereby new approaches to unlock the potential of BD are needed (Sathi, 2012). This is depicted in Figure 2.18 which describes the different data sources, the colossal data sizes and the technologies leading to new insights that have become available.

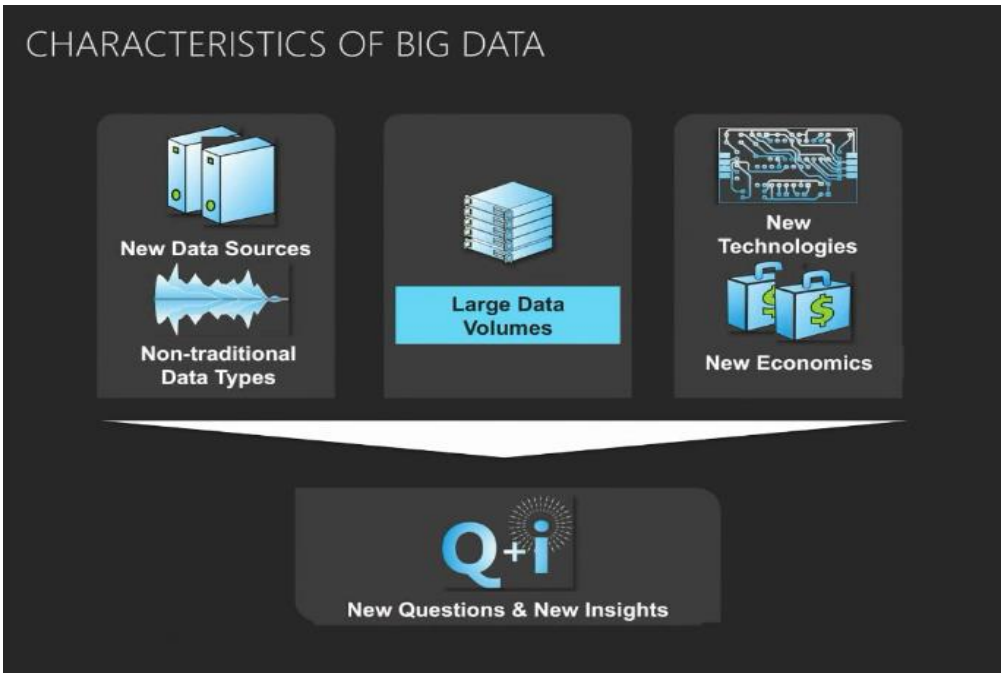


Figure 2.18: Dimensions of Big Data insight generation

(Source: Sen, 2012)

The most significant impact BD can have on an organisation is its ability to upgrade existing business processes and uncover new monetisation opportunities from key business elements such as customers, products, campaigns and operations.

Schmarzo (2013) states that BD brings to the table a collection of applications (ways to use or deploy) in the business domain which include personalised marketing, location based services, predictive maintenance attribution analysis and machine behavioural analytics. Big Data is about business transformation in moving business from retrospective batch business monitoring hind sights to predictive, real-time business optimisation insights. Predictive analytics enable a business to operationalise BD by:

- Optimising business processes and reducing operational costs
- Engaging deeper with customers for better customer experiences
- Identifying new products and market opportunities
- Reducing risk through identification and problem mitigation

Industry experts use the term predictive analytics broadly to imply predictive and prescriptive analytics. Predictive analytics looks into the future to provide insight into what will happen and includes what-if scenarios and risk assessment. The focus is mainly forecasting, hypothesis testing, risk modelling and propensity modelling. Prescriptive analytics is focused on understanding what will happen based on different alternatives and scenarios, and choosing the best options and optimising what is ahead. Sample use cases may include best action-related offers, customer cross-channel optimisation, portfolio and business optimisation and risk management.

2.10.2 Data models in business models

A business model refers to an organisation's core logic for creating value (Linder & Cantrell, 2000:1). According to Teece (2010), as mentioned in Section 1.1, when a business is formed, it either explicitly or implicitly employs a particular business model that describes the design or architecture of value creation, delivery and capture mechanisms. Linder and Cantrell (2000) mention that the ability to distinguish and communicate business models (see Section 2.6) will improve the organisation's focus, establish a frame for competing in an agile way and position the organisation to thrive despite industry discontinuities. According to Stubbs (2014:69), this is a continuous process referred to as functional innovation. Stubbs (2014) describes functional innovation as when an organisation has extended its perspective from a one-off to continual improvement of the business.

Functional innovation occurs when the organisation understands how the business works and develops the ability to measure, search and deliver gains continuously. Although continuing to deliver, gains depend on continuous improvement and senior executive commitment, which implies new improved business models centred on information (Stubbs, 2014).

Such dependence on information relates to data development which Mosley *et al.* (2009) describe as defining data requirements and specifications that may be organised into logical data models which are a fundamental part of the core business strategy. According to Mosley, in order to create an enterprise data architecture that delivers, the enterprise needs to first define its information needs. An enterprise data model is a way of capturing and defining enterprise information needs and data requirements. An organisation's business model, coupled with information use (access and retrieval), forms the heart or rationale behind systems architected to deliver insight for business processes, not to mention the need for persistence optimisation based on the organisation business model which is developed as part of data classification.

Adamopoulos (2013:2) defines data classification (DC) as a process that groups and refines data based on shared characteristics with the aim of facilitating a key storage objective. For example, to create a comprehensive storage or data process strategy, the organisation may choose to classify data on the basis of business priority. An organisation may use a classification optimised for data retrieval as the primary goal to store data (Adamopoulos, 2013). Explicitly optimising data retrieval might be the key rationale for the adoption of a particular data classification to aid a storage approach for data analysis, decision-making or *ad hoc* querying (Pokorny, 2011).

Organisations such as financial institutions may adopt optimised information systems for transactions, hence the deployment of a tiered system that scales up well. Media organisations (for example Facebook) may adopt optimised systems for information retrieval, hence deploying an unstructured data solution such as Cassandra, or better still, a polyglot persistence implementation (Mohanty *et al.*, 2013). According to Mohanty, the nature of a business model directly affects or depicts which systems are implemented, as these are in turn affected by the data models that have been employed on the basis of asset information.

Implicit in BD curation is the use of data models. Merson (2009:ix) defines a data model as:

> …the description of the structure of the data handle in information systems and persisted by a database management system, it has a set of symbols and text to precisely explain a subset of real information to improve communication within the organisation leading to a more flexible and stable application environment.

Modell (2007) describes a data model as a picture which depicts how data is to be arranged to serve a specific purpose. For decades, organisations have used relational databases as repositories for reliable data storage, but for last three decades, there has been a shift in data storage which has been fuelled by advances in technology and demands in software (Stonebraker, 2010; Jacobs, 2009).

A common data model is consistent, integrated and logical, and it defines the characteristics of information stored in the configuration management database. The model states how this data is organised relative to real-world entities and the relationship among entities. In other words, this can be called the ontological representation. One of its main advantages is making data management clear and useful for consumption. These definitions directly relate to containment or classification as put forward by Adamopoulos (2013) referring to storage and access of data in the organisation.

In computer storage, containment is not a zero sum, as storage items could be easily duplicated at cost which may be trivial but expensive in unstructured environments when data reaches the volumes of terabytes (Whitehead, 2002). In a real life scenario, an object can only be at one place at a time within or outside a particular containment. Whitehead (2002) mentions containment-suitable terminologies such as uniformity, support analysis, graphic formalism, utility and cross-discipline to explain the concept of storage. These terms will not be described in this research as it is outside the scope of requirements.

2.10.3 Challenges and trends in favour of NoSQL

According to Ivarsson (2010), the four trends that encouraged the growth of NoSQL are detailed in Table 2.4.

Table 2.4: Summary of trends in favour of NoSQL

Trend	Description
Size	From 2005 to 2020, the digital universe will grow by a factor of 300, that is, from 130 Exabyte to 40,000 Exabyte (Gantz & Reinsel, 2012:1).
Data connectedness	Data has evolved with interlinks, for example Text documents, Hypertext, RSS, Wikis.
Data structure (semi-structured)	Individualisation of content; decentralisation of content.
Architecture of applications	Evolution of applications.

Table 2.5 below highlights differences between a relational database and a typical NoSQL database such as MongoDB.

Table 2.5: Comparison of a relational database to a NoSQL database

(Source: Nayak et al., 2013)

Relational Database (RDBMS)	NoSQL Database (MongoDB)
Tables, views	Collections
Rows	JSON Document
Index	Index
Join	Embedded document
Partition	Shard

The form of different datasets and diversity of challenges needing redress influenced the start of the NoSQL era. However, this data is considered disruptive (Sathi, 2012). Sathi (2012) continues by stating that technological innovations are classified as being sustaining or disruptive. In the category of sustaining technologies, well-run organisations invest to continually improve performance of established products along the dimensions of performance that customers have valued. On the other hand, disruptive technologies bring to the market different value propositions which are different from the norm.

Disruptive technologies perform comparatively well in relation to established products in the market (Sathi, 2012); however, such technologies are typically cheaper, simpler, smaller and possibly more convenient to use. Big data and the curation thereof is known to be a disruption to normal data curation as organisations for decades used traditional methods and tools for data management. The disruption comes from the known characteristics of BD.

The disruption, relative to BD, emanates from the characteristics of the data as it defies traditional data curation processes that have been used by organisations for decades. Schmarzo (2013:40) indicates that about 95 per cent of organisations are still predominantly using traditional curation processes, thereby placing them in the monitoring phase of the business model BD Maturation Index.

Relational databases (RD) differ from non-relational approaches in that RD uses SQL for defining and manipulating data. Many NoSQL databases use the unstructured query language (UnQL) which is focused on the collection of documents. There are significant differences between the two approaches; RD scale vertically, requiring an upgrade in field replaceable units (FRU) such as random access memory (RAM) and central processing unit (CPU) to scale up and improve performance, while most NoSQL databases scale horizontally with increasing commodity machines. Doubling the number of machines will double processing power. Relative to transaction properties, SQL databases focus on ACID (Atomicity, Consistency, Location, Durability) properties while NoSQL databases rely on

BASE (Basically Available, Soft State, Eventually Consistent) properties, which makes it compliant with the CAP theorem.

What is apparent is the lack of industry standards relative to defining the unstructured data forms (Pokorny, 2011). To clarify this further, an entity relational diagram provides a standard way of visually conceptualising a relational database. As stated by Ivarsson (2010), the different forms of NoSQL databases are best suited for different storage jobs within the non-relational environment as per performance. This is displayed in Figure 2.19 as to how the different NoSQL databases scale to complexity.

Key value stores allow for storage of schema-less data whereas the key facilitates identification of the data and the actual data is stored in the value. According to Seeger (2009):

> The data itself is usually some kind of primitive of the programming language (a string, an integer, and an array) or an object that is being marshalled by the programming languages bindings to the key value store. This replaces the need for fixed data model and makes the requirement for properly formatted data less strict.

Pokorny (2011:5) states that, “relative to the differences among NoSQL databases, a unified query standard is not apparent”. There is no grounded theory of NoSQL databases with the exception of the works of Meijer and Bierman (2011), relative to their mathematical data model for most common NoSQL databases.

The culmination of this data accrual results in gigantic datasets that challenge industry analytical tools, forcing organisations such as Amazon, YouTube, Facebook, Flickr and many others to create their own database ecosystem.

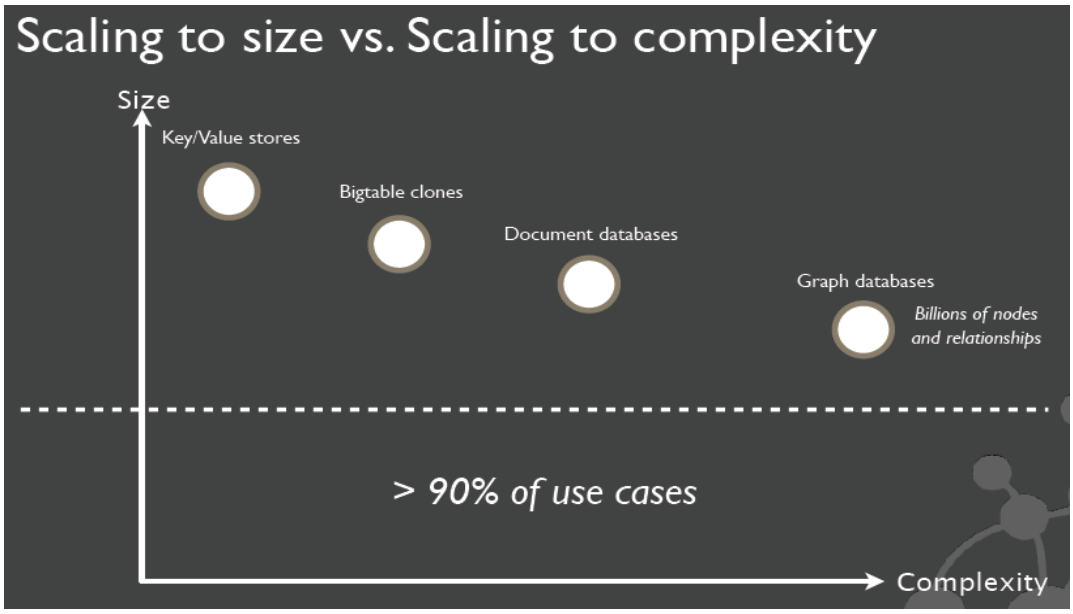


Figure 2.19: Database scaling to size against complexity

(Source: Ivarsson, 2010:28)

2.10.4 NoSQL databases

NoSQL comes in many forms and has quickly gained ground in the production environment (Rees, 2010). In brief, some of the factors that describe NoSQL databases include, but are not limited to, the following:

- Ease of use in conventional load-balanced clusters
- Persistent data (not just caches)
- Scale-to-available memory with no need for a fixed schema
- Schema migration without downtime

There are many more characteristics which are beyond the scope of this research. However, other more common characteristics mentioned by Nayak *et al.* (2013) include NoSQL DBs which use a proprietary query system rather than a standard query language. They are ACID within a node of the cluster and eventually consistent across the cluster. They can be used in a web application to promote customisation of fields and have no need for schema changes. Fields in NoSQL DB can still be indexed irrespective of the lack of an initial schema. Other attributes salient in NoSQL DBs include their use as a caching layer; they can be stored as binary files and used to attach metadata to specific files. Some NoSQL databases can be used to serve full web applications, for example couch DB.

The convenience of NoSQL implementations lies also in the simplicity of use; very common in today's web applications is the use of NoSQL to serve CSS, HTML and JavaScript directly. Further permissions can then be used to control who can read and write data to the

database. NoSQL databases can be used to build new applications from scratch as well as augment the capabilities of relational databases as in polyglot persistence which is described in Section 2.14.

2.10.5 Categorising NoSQL databases

NoSQL databases fall into four broad categories based on their operating principles, as listed in Table 2.6. There are other variations of NoSQL databases not mentioned in Table 2.6. The variations are discussed further in Sections 2.10.5.1 to 2.10.5.7.

Table 2.6: Categories of NoSQL databases

(Source: Nayak et al., 2013)

Category Name	Description (Most salient features)	Example
Key value (based on Amazon dynamo paper)	• Scaling • Designed for massive data loads • Data model based on key value pair • Dynamo ring partitioning and replication	Voldemort Dynomite Tokyo DB
Document store	• Similar to key-value stores • Motivated by Lotus • Data model: Collection of Key value collections	CouchDB MongoDB Redis
Big Table (Based on Google's Big Table paper)	• Like column oriented RDBMS but handles semi-structured data better. • Data model: Columns – column family-ACL • Datums keyed by rows, columns, time, index • Row range: Row- tablet- distribution	HBase Hypertable Cassandra
Graph document	• Focus on modelling the structure of data • Scales to complexity of data • Based on mathematical graph theory • Data model: Nodes • Relationship/edges between nodes • Key value pairs on nodes	Neo4j Allegro graphDB Sones graphDB

2.10.5.1 Column store

According to Vaish (2013:26), "a column family store, stores data as columns as opposed to rows that is prominent in RDBMS". This is highlighted in Figure 2.20. Nayak *et al.* (2013) indicate that this range of databases is a hybrid row/column store. Unlike pure relational databases they share the column-by-column concept, except they do not store data using tables; rather, they store massive distributed architectures.

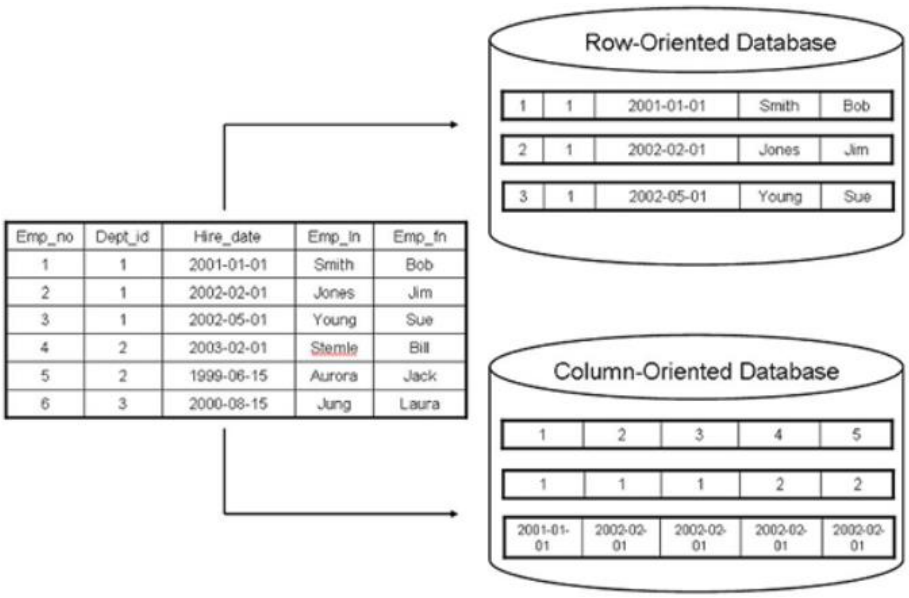


Figure 2.20: Graphical representation of columns store database

(Source: Bi Insider, 2014)

2.10.5.2 Document stores

A document store is a document-oriented database optimised for the retrieval, storage and management of document-oriented data. It is also referred to as semi-structured data.

There are various implementations of the document-oriented database forms but common to these different implementations is the standard encoding schemas used widely across the board. Encoding, by definition, is the conversion of data into code for storage. The structured formats include, though not limited to, XML, JSON, YAML, BSON or any other structure or format that can be queried, such as PDF and MS office documents (MSWord and Excel). These are usually organised into collections and in conformance with the principles of NoSQL, individual documents can have unique structures, implying the possibility of different fields in different documents. Each document has a specific unique key for the retrieval of a document. This is augmented by the possibility of querying a particular document using a field contained in the document (Mohanty *et al*., 2013).

Some examples of document stores include ArangoDB, BaseX, Cassandra, Cloudant, Clusterpoint, CouchDB, eXist, FleetDB, Jackrabbit, Informix, Inquire, Lotus Notes, MarkLogic, MongoDB, MUMPS DB, OrientDB, RavenDB, Redis, RethinkDB, Rocket U2 and Sqrrl Enterprise.

Documents stored within a document-oriented database are similar to the rows of a relational database management system (RDBMS), but these are not bound by predefined structures

as the rows may vary. That is, strict adherence to this predefined structure is not mandated or enforced by a standard schema, section wide organisation of data, slots, parts or keys.

The above mentioned documents share a similar basic structure but each has different elements with no empty fields unlike the case with empty fields in a RDBMS. This approach allows adding new records without conforming to a predefined structure. Document stores are organised around keys, data retrieval and organisation. A unique key maps a value in the database; this is in the form of a simple string, uniform resource identifier (URI) or a path allowing for retrieval of a value. An index in the database which is an integral part also allows for the quick retrieval of documents. Some implementations may allow data to be queried with more than one key, permitting evasion of a single point of failure as a draw back. This database category is used alongside relational databases for caching.

Another apt characteristic is the provision of an API or a query language for the retrieval of information from the database, with differences with respect to implementation based on the way the document is organised. The organisation of documents takes the form of collections, tags, non-visible metadata, directories or hierarchies and buckets.

2.10.5.3 Key value store

Key value databases are considered one of the most simplistic of the NoSQL domain databases. The key value databases are extremely efficient and have a powerful model; as with many of the databases it allows for a schema-less data storage. According to Hadjigeorgiou (2013:5), “in a key value store, data is stored as values with a key assigned to each value similarly to hash-tables”. This selection of databases can be found in the implementation of cloud mobile applications, point of sale systems, and also in applications that manage factory control and information systems.

2.10.5.4 BigTable/tabular

BigTable is a high performance, proprietary and compressed data storage system built on the Google file system, Chubby Lock Service and other Google technologies (Burrows. 2006). It started in 2004 and is used in many Google applications such as web indexing. MapReduce is often used for generating and modifying data stored in Google Maps, BigTable and Google Book Search (Burrows, 2006). Named after Google’s proprietary BigTable implementation with each row possibly having many different columns, it maps to two arbitrary string values, row key and column key and timestamp allowing for version and creating an arbitrary associated byte array in the form of a three dimensional map. It is designed to scale into the petabyte range across hundreds or thousands of machines with

the ability to add many more machines. The time dimension allows for versioning and garbage collection.

2.10.5.5 Graph/interconnected nodes

A graph is a collection of vertices (nodes) and edges, or stated in lay man terms, a graph is a set of nodes and connecting relationships. A graph represents entities such as a person (John); entities are nouns while edges represent the relationship to the corresponding entity pair such as “knows” (verb). For example, John knows Jerry. Jerry knows Jane. John knows Jane. Some frequently used graph-related terminologies, including vertex, edge, properties, weight and graph; these are defined in the glossary.

Graph databases store data in the form of a graph; a graph consists of nodes and edges (Nayak *et al.,* 2013). Nodes act as objects while edges act as the relationship between the objects, which is best represented as interconnected nodes visualised as a series of road intersections. Nayak *et al.* (2013) state that graph databases use a technique called Index Free Adjacency which implies that every node consists of a direct pointer which points to the adjacent node. The main emphasis is on the connection between data; graph databases are schema-less and optimised for semi-structured data. It scales quite easily and is whiteboard-friendly and ACID compliant with rollback back support. An example of this graph database is Neo4j.

2.10.5.6 Object database

Object-oriented databases, according to Nayak *et al.* (2013), can be considered as a combination of object-oriented programming and database principles. It is tightly integrated with the object-oriented programming language, with the database acting as a persistence layer which stores objects directly. Objects can also be linked together through pointers; one of its primary advantages is the ease of integration into modern software development processes when the Software development life-cycle (SDLC) is agile. By recommendation, object databases should be used in environments with complex object relationships.

2.10.6 Hybridisation of persistence platforms

The co-existence of relational databases and NoSQL databases has been the focus of many database organisations seeking to optimise persistence for gain. According to Schmarzo (2013), about 95 per cent of organisations are still using relational database systems for data management. While NoSQL is presented as an alternative to the widely used relational database, Nayak *et al.* (2013) advise that it does not completely replace SQL but rather complements it in such a way that both can and should co-exist for optimal delivery where

necessary. Devlin, Rogers and Myers (2012) of Enterprise Management Associates (EMA) call this co-existence a hybrid data ecosystem, while other authors may refer to these implementations as a polyglot persistence. EMA is a leading industry analyst and consulting firm that specialises in providing deep insight across the full spectrum of IT and data management technologies.

Devlin *et al.* (2012:11) state that a complete hybrid data management ecosystem may comprise of:

- An enterprise federated data warehouse
- Data marts
- Operational data stores
- Analytical database platforms/appliances
- NoSQL data store platforms
- Data discovery platforms
- Cloud-based data solutions
- Hadoop and its sub-projects

In a hybrid ecosystem, each platform supports a combination of business requirements and provides solutions to processing challenges. This persistence and insight generation approach circumvents the restriction of a single traditional database which is more prevalent. Furthermore, the hybrid or polyglot brings to the table the strengths of different systems through a seamless connectivity and application to a different, yet common course (Devlin *et al.,* 2012).

Dumbill (2012) states that Apache Hadoop has been the driving force behind the growth of the BD industry. Its primary essence is its ability to process large amounts of data at low cost, regardless of structure.

Figure 2.21 demonstrates a BD ecosystem with Hadoop. It has weblogs, transactional data and relational data as the different data sources load into the Hadoop MapReduce system. These are then further ported with the right structure into data marts, data warehouses and operational systems, provisioned for consumption.

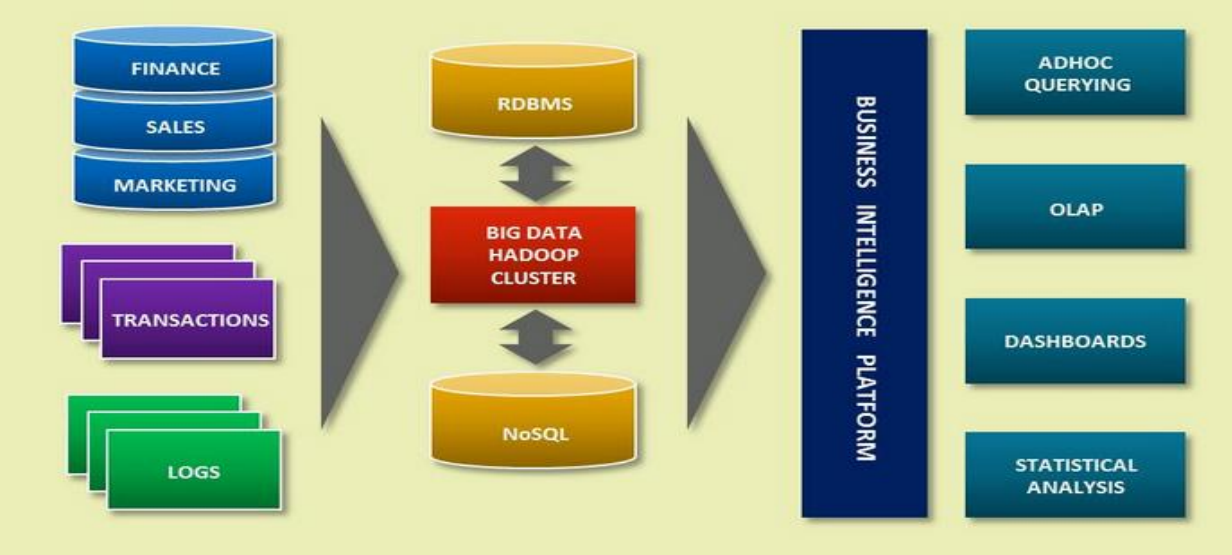


Figure 2.21: A Big Data ecosystem with Hadoop

(Source: Fryersolution, 2013)

The core of the Hadoop processing system is MapReduce. MapReduce is a framework that brings on board the ability to query a dataset, break it up and run it parallel across multiple data nodes. The distribution of computation solves the problem of data too large to fit onto a single machine. The Hadoop ecosystem is comprised of tools indicated in Table 2.7. Tools such as *Scoop* are connectivity tools for moving data from non-Hadoop data stores, such as a relation database, into Hadoop.

Table 2.7: Hadoop supporting toolset

(Source: Mohanty et al., 2013:199)

Tool	Deployment
Ambari	Deployment, configuration and monitoring
Flume	Collection and import of log and event data
HBase	Column-oriented database scaling to billions of rows
HCatalog	Schema and data type sharing over pig, Hive and MapReduce
HDFS	Distributed redundant file system for Hadoop
Hive	Data warehouse with SQL-like access
Mahout	Library of machine learning and data mining algorithms
MapReduce	Parallel computation on server clusters
Pig	High-level programming language for Hadoop computations
Oozie	Orchestration and workflow management
Sqoop	Imports data from relational databases
Whirr	Cloud-agnostic deployment of clusters
Zookeeper	Configuration management and coordination

2.11 Vending Big Data

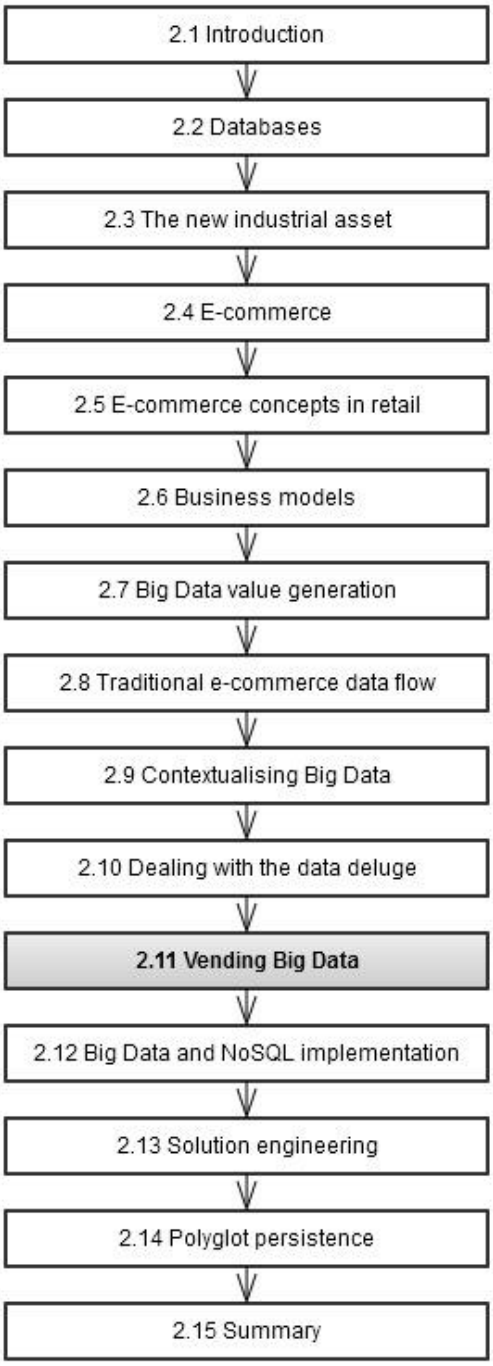


Figure 2.22: Chapter layout – Vending Big Data

A case study was conducted by IBM that focused on how BD may assist media companies to assist customers through sentiment analysis (See glossary). IBM uses BigSheets to assist in the collection and analysis of data. According to Laningham (2010), BigSheets is a cloud application used to perform *ad hoc* analytics at web-scale, both on structured and unstructured content.

It leverages Apache Hadoop and Map Reduce technologies to give business users a tool for exploring BD, gain insights and observations and be able to execute data analytics without IT support.

BigSheets, according to Laningham (2010), effectively integrates gigabytes, terabytes or petabytes of unstructured data. BigSheets further collects a wide range of unstructured web data, extracts and enriches the data, and allows the user to explore and visualise this data in specific user defined contexts (such as ManyEyes), and is a component of IBM InfoSphere Big Insights solution.

In the aforementioned case study, BigSheets was used to assist in measuring consumer sentiment towards Box Office offerings and to predict which films would succeed according to results obtained from the data analytics. The type of data being captured in the instance of this particular case study was social media. Laningham (2010) lists the advantages of this platform as being able to (i) understand consumer sentiment before a film is released in box office, and (ii) measure efficacy of the marketing effort of studios, as studios will acquire knowledge about consumer interest. Insights could help studios with *"*go/no-go decisions on everything from the breadth of distribution to whether it made sense to invest additional marketing muscle to push a movie over the public's awareness tipping point".

> Being able to retrieve data from all available repositories in an organization is a key part of doing analysis involving Big Data, especially for exploratory analysis (Zikopoulos *et al*., 2013:165).

One of the challenges Zikopoulos *et al.* (2013) mention is that of collecting relevant information from BD in an organisation. Different kinds of information are collected into silos, and therefore there is a need to integrate data from all repositories so that it can be of use during the analytics process. An example of such a challenge would be when the user wants to retrieve customer information from transactional databases, email servers and call-centre engagement logs in order to organise the information to form a holistic picture of single customer behaviour.

Generally, different applications are used to store the various forms of information. Zikopoulos *et al.* (2013) cite the IBM InfoSphere Data Explorer as being a critical component in the IBM Big Data platform. This enables the user to access all forms of information in a "single integrated view", thus eliminating the work that has to go into trying to organise information in its various formats in a relevant fashion.

It is starting to become apparent that BD's value is drawn from its ability to be manipulated and used as a decision-making tool for business strategy. What is more significant is that if

used alone, BD could prove just as detrimental as staying with traditional methods of business management in a rapidly evolving world (global) market.

Leveraging BD is enabling media companies to compete outside of traditional media markets. Bruell (2013) states that:

> New data capabilities already are supporting work outside traditional media. During a client-planning exercise, for example, Annalect used its audience-segmentation tool to create insights for the client's research, creative and media teams.

According to Bruell (2013), Publicis Vivaki (a media company) which supports its holding company's media agency operations, has set up a data-driven marketing task force. They are so determined to harness the power of BD that they are “thinking through how best to organise across all of Public IS" and which tools to invest in (Bruell, 2013). One example is a product called Skyscraper, which collects, cleans and verifies data, helping agencies negotiate the rights to access external data. It means that "12 different agencies do not have to negotiate their own deals".

It is becoming apparent that media companies need to move away from a traditional approach in operations to a more technologically based approach in order to ensure that they are correctly placed in the market. Thus, data which once served as a satellite element adjacent to any business must shift in priority and become core to business functions (Bruell, 2013).

The initial step to creating value from Big Data is to create a “Big Data Plan”. According to Biesdorf, Court and Willmott (2013), despite being able to list the benefits of why an organisation should start to make use of Big Data to create value, it is even more advisable to create a plan to allow technological change to integrate holistically and beneficially into an organisation. Biesdorf, Court and Willmott (2013) thus purport that a successful plan should contain three core elements:

- Data
- Analytical models
- Tools

In summary, the Big Data plan needs to detail how the organisation intends to gather and integrate data. Next, for data to be used successfully in the organisation, the plan needs to allow for the creation of analytics models for optimisation and prediction based on the data.

Furthermore, the method of operation and the individuals who manage and use these models would need to be identified where these models would be used in the organisation.

Lastly, tools that translate the output of these models into action and strategy for use by members of the organisation would need to be developed.

According to Vogt (2013), BD is imperative to media companies' success. He indicates that while there is an unlimited amount of data available to media companies for making decisions and ensuring that value added services are provided for clients, media companies need to "focus on obtaining unique relevant data, respecting the privacy of consumers and analysing data to deliver engaging content, insight or advertising for consumers at the right time".

When employing a consumer-focused model, Vogt (2013) emphasises that there is a shift from media to the relevant use of analysed data to managing the quality of content offered to the consumer, and thereafter, the delivery of that content. Software vendors through their product offerings give evidence to the value of mining data to any organisation. According to Dijcks (2013:7), there are four steps to effectively mine BD and that these include, from beginning to end:

- Acquiring and organising data
- Discovering and analysing results
- Predicting and planning findings
- Making better decisions

> The Oracle NoSQL Database, with its "No Single Point of Failure" architecture is one of the many solutions when data access is "simple" in nature and application demands exceed the volume or latency capability of traditional data management solutions (Dijcks, 2013:7).

Forsyth (2012) highlights how Oracle and Cisco are among the first software vendors to develop and distribute software that would assist in the effective mining of BD: Oracle with their NoSQL database and Cisco with their unified computing system (UCS).

> Together, this solution provides a platform for quick deployment along with predictable throughput and latency for most demanding applications (Forsyth, 2012:4).

Research by software and IT companies therefore show that the IT industry has put in much thought into the value that leveraging BD has in increasing an organisation's reach within their specific target markets.

2.12 Big Data and NoSQL implementations

The opportunities of leveraging BD have been at the forefront of discussions in most sections of industry, especially in e-commerce retail. This seems to be driven by the competitive advantage organisations obtain from these databases and which they subject to analytics. Agrawal *et al.* (2011) state that BD, as a feasible concept, is due mainly to the explosion of data and disparity of usage scenarios, coupled with the major shifts in computing hardware and platforms, as agreed by the Claremont research. Among the opportunities, researchers concur that the greatest shared challenge is not engineering BD but rather harnessing the opportunities thereof (Bizer, Boncz, Brodie & Erling, 2012).

Tonytam (2010) refers to Wordnik, an online dictionary launched in 2009, as a language and resource centre. Wordnik's business operation is aimed at the discovery of new words, concepts and meanings by applying natural language processing to newly published content from feeds such as The Wall Street Journal, Forbes and Flickr, including audio from MacMillan. According to Tonytam (2010), Wordnik has grown to the point of adding about 5 000 to 8 000 citations per second, signifying a tremendous growth and need for expansion to parallel the growth size. In the beginning of operation there were a little less than a million records with about 50 records updated per second. Tonytam (2010) states that the growth potential degraded performance drastically. The system performance stalled badly to the point that operation was a nightmare. Over time, it became apparent that what was needed was a migration to NoSQL for its raw speed, as the source of the performance debacle was rooted in the lack of speed, flexibility and scalability; hence a migration to MongoDB was imperative.

Stonebraker (2010) has a different view of this situation and puts forward that blinding performance has nothing to do with SQL systems, but to rather eradicate overheads within RDBMS as the way forward to accomplishing an improved performance. Many authors however disagree as this is against the concept of using the right tool for the right job (Sathi, 2012; Mohanty *et al.,* 2013; Zikopoulos *et al.*, 2013). It entails sacrificing the ACID concepts of relational databases for gains which is a direct contradiction of Codd's concept of a relational database.

According to Tonytam (2010), the raw speed of MongoDB was clearly visible and a measurable performance of 1000 records per second on MongoDB. From a low performance of 50 records per second on MySQL, it was a turning point in operation and a right move in terms of decision-making. Furthermore, MySQL required a large volume of conditional logic coding for multiple schemas, whereas there was a registered reduction in coding by about 75 per cent.

Another example of leveraging BD is the case of Watson, an IBM artificial intelligence project with the ability to imitate human behaviour. Watson was developed by IBM as an open-domain question answering system. It draws knowledge from BD as its information source. Watson participated in Jeopardy, a contest among Watson and human counterparts. Watson analysed a question posed during the competition and through an intrinsic algorithm it generated an answer while building a knowledge base. In the process, Watson improved its knowledge base over time.

Watson had access to 200 million pages of structured and unstructured data content consuming four terabytes of disk storage which included the full Wikipedia text, but he was disconnected from the Internet (Jackson, 2011). The data is resident in memory to facilitate a quicker access rate, as information resident on a hard disk is slower to access compared to memory.

At the end of the Jeopardy contest, Watson outdid its human counterparts by far.

2.13 Solution Engineering

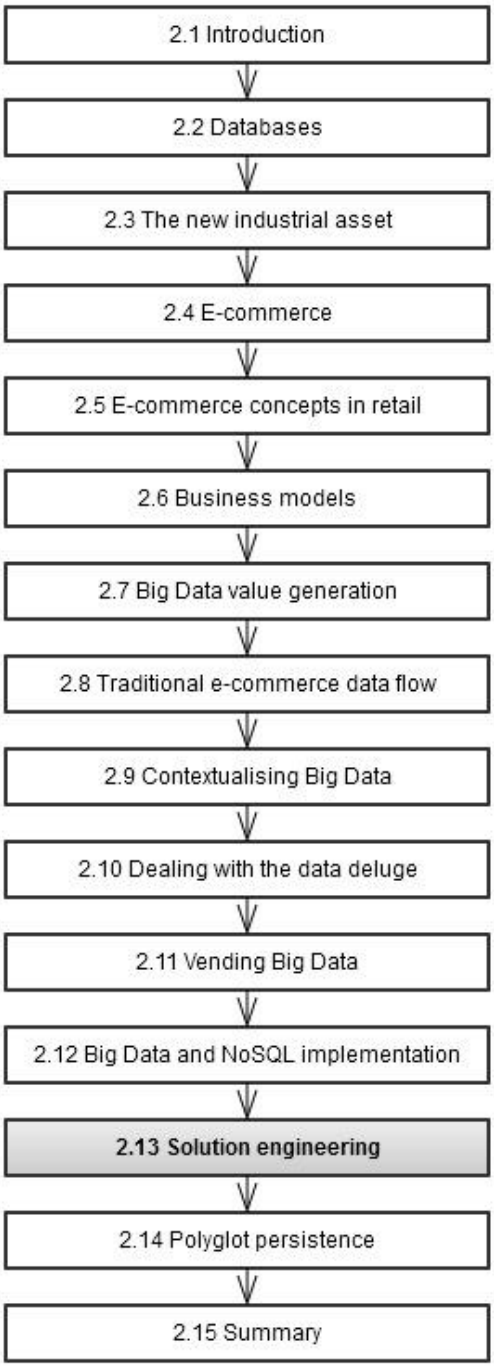


Figure 2.23: Chapter layout – Solution Engineering

Big Data is disruptive and a risk for organisations when curators or organisations are not fully aware of the underpinnings of BD curation. This is true especially for a young organisation seeking information advantage, however, it may be a risky venture to pursue, but its benefits are a myriad and most often far surpassing the initial hurdles or problems posed. According to Grant (2010), solving a problem is a matter of transforming the problem state into the solved state and that the transformation is accomplished by setting up and achieving three

types of goal states: transform, reduce and apply. These form the basis of Solution Engineering—a disciplined, systematic way of getting at all three categories of knowledge required to solve a problem. It is especially useful in obtaining knowledge of the structures in which problems are typically embedded. *Transform* deals with identifying differences between the two states, that is, what is and what should be. *Reduce* refers to identifying ways of eliminating these differences to close the gap, while *apply* implies putting into play operators that eliminate the differences.

According to Nickols (2012:4), there are three categories of knowledge and information needed to facilitate this process successfully, namely knowledge of ends, structure and means. *Knowledge of ends* implies goals, aims and objectives of the problem solving effort, a view of the goal and the final solved state. *Structure* refers to the cause-effect relationships, while *means* implies the methods and techniques for affecting change. According to Schmarzo (2013:152):

> …solution engineering identifies and breaks down an organisation key business initiatives into its business enabling capabilities and supporting technology components in order to support an organization's decision-making and data monetization efforts.

Nickols (2012) asserts that engineering a solution lies in having adequate knowledge of the structure of the larger situation in which the problem is embedded. Knowledge of this structure facilitates engineering the solution with the value of perfect information.

Lin (2013:20) defines the value of perfect information as:

> …the difference between the maximum payoff under conditions of certainty and the maximum payoff under uncertainty.

In other words, the Rand monetary value of information is called the expected value of perfect information written as EVPI. The processes involved in engineering a solution, according to Nickols (2012), is divided into two phases, namely the investigation and intervention phase as indicated in Figure 2.24. Going through these phases with the aim of solving a problem is Solution Engineering. Nickols (2012) further adds that a problem exists when action is needed, but the action to be taken is unknown, implying that somehow someone has to figure out what must be done.

Figure 2.24: Solution Engineering phases

(Source: Nickols, 2012:5)

The phases of engineering a solution as shown in Figure 2.24 are investigation and intervention. Investigation focuses on the structure and the problem description, stating whether the problem is static, dynamic or wicked. Properly identifying the dynamics of the problem is the key to solving the problem, and as it brings across the needed understanding, it helps modularise the problem intelligently, and allows for the progression and assimilation of the design and implementation of the intervention. Intervention deals with following the steps listed in the investigation phase to attain a better understanding, and implementing the steps listed in the intervention to accomplish the desired solution. Ultimately, the point of concern is to configure and carry out action steps like an algorithm scientifically that produces desired results by making changes at the point of intervention.

Schmarzo (2013:152), on the order hand, defines a six-step process for engineering a solution by solving organisational problems. These steps include:

a) Understanding how the organisation makes money
b) Identifying the organisation's key business initiatives
c) Brainstorming the impact of BD
d) Breaking down the business initiative into use cases
e) Proving out the use case
f) Designing and implementing the solution (such as BD curation solution)

Schmarzo (2013) mentions that time should be invested to identify the organisation's strategic business entities such as customers and products. Understanding how they fit into the bigger picture and their integration into the organisation's business initiative allows for understanding the bigger structure. According to Nickols (2012), designing a solution or algorithm to accomplish solving a problem becomes more intelligent and doable when armed with this information.

2.13.1 Big Data journey

Big Data curation is a multi-phase multi-year journey that many data-driven organisations are pushing to empower, transform and derive value from the organisation, especially to gain a competitive edge (Sathi, 2012). Schmarzo (2013) poses a question which many companies seeking to curate BD ask: how do I compare with others in terms of my organisation's adoption of BD for business enablement? According to Sathi (2012), it is important to have a strategic vision that aligns with industry direction and responds well to the disruptive forces with the aim of reaching a target that makes a substantial impact on the organisation. Sathi (2012) confirms that the BD journey can be accomplished by a number of iterative phases with stepwise advancements in capability, via well-defined yet small steps that are guarded to alleviate risk, leading to a successful investigation with a defined output.

Big Data curation or integration into a business model is considered a journey by many authors (Schmarzo, 2013; Dijcks, 2013; Stubbs, 2014). It is mandatory for company executives and data curators to realise that the integration or the adaption of the resulting tactics depend on their own company's readiness. In so doing, the organisation endorses a data-driven culture that data curators, on the forefront or curation, enforce. Mosley *et al.* (2009) list data-specific topics that are significant for firms pursuing a data-driven approach to differentiation in the long-term as a strategic plan. These include the need for data specialisation, a solid knowledge of data architectures, metadata, data quality and correction processes, data curatorship and administration, master data management hubs, and matching algorithms.

Sathi (2012) mentions the importance of selecting and designing short-term and measurable impact data projects with benefiting areas, such as product engineering and operations for quick and early results. This is a process of strategising BD as a short- term project.

According to Schmarzo (2013), it is imperative for an organisation seeking to integrate BD, to ascertain where it stands with respect to exploiting BD and advanced analytics. He mentions five distinct phases as depicted in Figure 2.25.

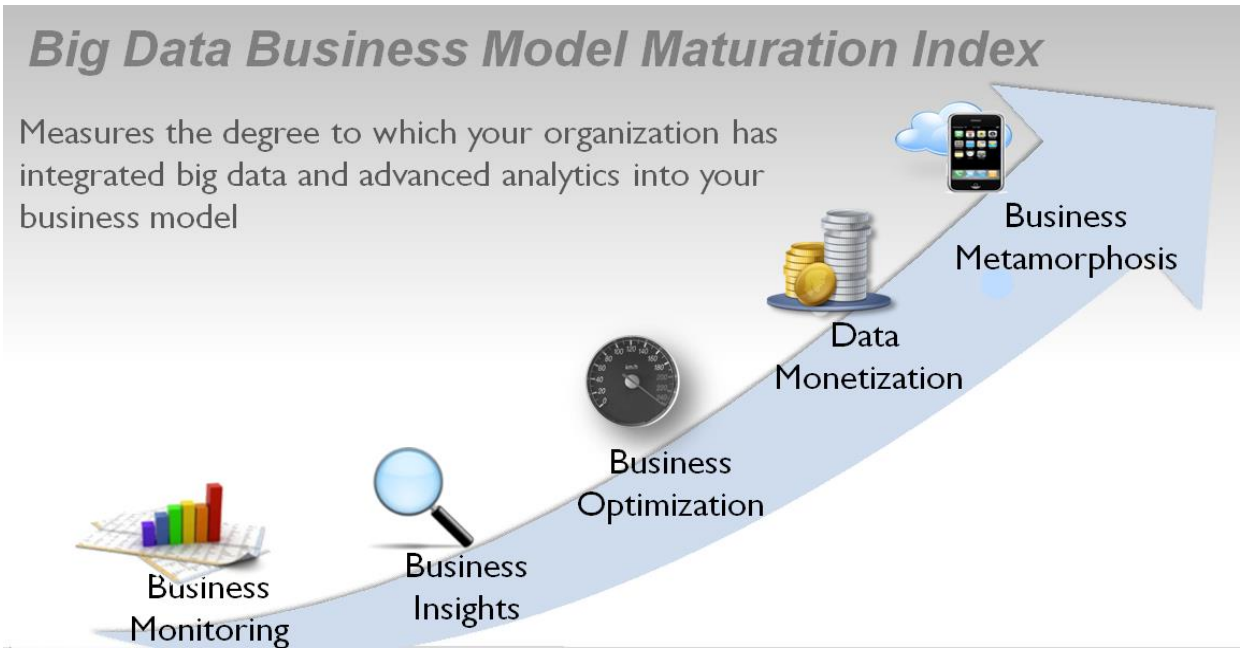


Figure 2.25: Big Data Business Model Maturation Index

(Source: Schmarzo, 2013:16)

The phases include business monitoring, business insights, business optimisation, data monetisation and business metamorphosis. Schmarzo's (2013) model is used in this research as a gauging model or control to ascertain where an organisation stands prior to commencing curation for an intelligent take-off into the BD integration process. This model is appropriate for this research due to its platform-independent approach to addressing BD curation.

Table 2.8: Phases of Big Data Business Maturation Index

(Source: Schmarzo, 2013:10)

Phase	Description
Business monitoring	Organisation deploys traditional curation processes to manage and monitor business using traditional reporting to progress measurement.
Business insights	Business insights serve as a business monitoring process however this phase integrates with unstructured data sources, advanced statistics, predictive analytics, data mining, coupled with real-time data feeds to identify actionable business insights for integration into business processes.
Business optimisation	Optimisation of business processes with insights from analytics in an automated way.
Data monetisation	Leveraging BD for new revenue opportunities, for example using a smart phone to glean data for trends to influence consumer behaviour.
Business metamorphosis	Leveraging data to transform business models into new services in a new market space.

In contrast, Sathi (2012) outlines these five steps as part of a business maturity model that enables an organisation to track its way through a process of stepwise iterations to maturity with increasing capabilities. These phases include *ad hoc*, foundational, competitive, differentiation and breakaway. The author mentions that this model is an important tool for developing an enterprise-wide analytical roadmap which also provides a basis for comparative analyses with other models on the BD curation front.

2.13.2 Big Data strategy

Strategising BD curation can be viewed as part of Solution Engineering. Every curation process geared toward a long-term or short-term project is aligned to the overall business strategy to furnish the organisation with relevant data to address a need. Curation projects can be categorised into short-term or long-term projects and flanked with similar attributes for feeding the data back into the system to augment decision-making. For example, curating data to profile customers to improve customer engagement on the site can be a long-term project, but also segmented into smaller short-term projects to evaluate the success as mentioned in Section 2.13.1. This process involves identifying the business initiatives that support the short-term curation process for the given business strategy. This is followed by identification of anticipated outcomes and critical success factors as actionable data which is generated in subsequent phases of the process.

The next step then involves the task of embarking on actionable data. Actionable data from social media platform sources is reactive and needs to be acted upon quite quickly. This is followed by identification of sources of data both internal and external to the organisation. The richness of the data will provide better insight for executing the proposed solution. For a project of this kind, possible data sources may include customer data, transaction data, marketing data, contact data and social media data (Schmarzo, 2013).

This document associates a data curation process to a particular business initiative that enables a curator to quickly define how the curation process should support the business. The critical factors and outcomes help measure the success of the curation process, while the tasks state the plans or the solution which is engineered to accomplish the set goals. The final part of the document relates to the different data sources that a curator may look at to collect data to satisfy the need for information (Schmarzo, 2013).

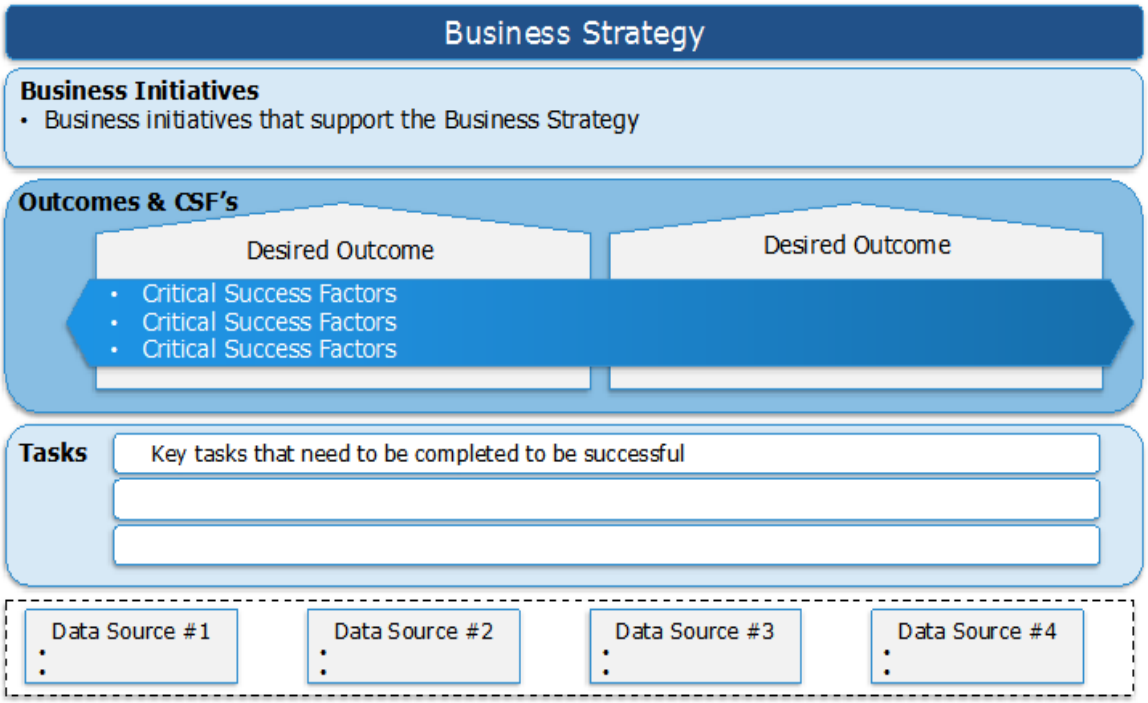


Figure 2.26: Big Data strategy document

(Source: Schmarzo, 2013:66)

2.13.3 Decision Theory for competition

Rosenbush and Totty (2013) emphasise that BD is changing the whole equation for business decision-making, as it allows proponents the thought process of selecting a logical choice from available options. Decision-making has roots in Decision Theory. LaValle (2013:437) describes decision-making as the selection of choices between alternatives involving risk and uncertainty because it addresses the problem of making decisions about a situation where there is a certain level of uncertainty. Many decision-makers, especially in organisations, are sometimes caught in a dilemma as to what decision to make. According to Schmarzo (2013), Google's dominance of the advertising industry came, not by understanding or perfecting advertising techniques, but by applying analytics to massive, detailed data sources to identify what works, without the need to worry about why it worked. The essence thereof is BD, and advanced analytics may yield insights about a business without it, thereby reducing the absolute need for heavy statistical modelling which is typically required with sampling datasets.

According to Lin (2013:20), there are three components of any decision:

a) The choices available or the alternatives
b) The state of nature (future demand)
c) The payoffs

The choices available or the alternatives are choices the decision-maker finds himself having to juggle for the best option based on the presented variables, including the state of nature and the payoffs. Normally this becomes easier with a smaller set of choices; the state of nature defines the uncontrollable future events which normally are out of their control. The payoff symbolises the outcome based on which decision is made; it is needed to compare each combination of decision alternative and its state of nature (Lin 2013). Table 2.9 summarises the states of Decision Theory.

Table 2.9: Components of Decision Theory

(Source: Nikov, 2012:58)

Component	Explanatory events
Event	• Uncertainty regarding future demand • State of nature (future demand) is unknown • The decision-maker has no control over state of nature
Act	• Two or more courses of action are available to the decision-maker • The decision-maker must evaluate alternatives • The decision-maker selects a course of action based on certain criteria • Depending on the set of circumstances, these criteria may be quantitative, psychological and social, to name a few.
Outcome Payoff Consequence	• Profit • Break-even • Loss

Mastering the concept of data-driven decision-making enables an organisation to advance in so far as they have the opportunity to continuously improve and learn through new ideas and constant feedback. Nikov (2012) declares that as decision-making allows for continuous improvement, it calls for community accountability towards requirements, and fosters a close organisational culture through unified learning.

2.13.4 Leveraging weblog

Weblog, according Sathi (2012), creates a new market where customer data from many contacts points can be secretly collected in many different forms. Significantly, clickstream data can be categorised anonymously and even repackaged to sell to others as a way of monetising data from analytics and BD curation. Sathi (2012) further adds that automation provides industry with a great opportunity to use sensors to collect data in every step of the customer-facing process, such as clickstreams in the use of web, making weblog data a big source of insight for consumer behaviour analytics.

2.13.4.1 Clickstream

Clickstream data is a form of unstructured data (BD) which, once obtained, can be analysed to demonstrate or predict the behaviour of users of Internet (online) technologies.

Analysts make use of clickstream data because it provides information about the sequence of pages or the path viewed by users as they navigate the website (Li, Sun, Alan & Montgomery, 2011).

In their proposition of making use of a dynamic Multinomial Probit Model of web browsing, Li *et al*. (2011) state that the intention of research is to show that the sequence of web viewings may be useful in forecasting a user's online path. Clickstream analysis (see Section 2.6), according to Schubert and Koch (2002), enables one to profile the kind of consumer who would potentially be interested in a website's product. This includes the nature of products that are most popular with visitors, the sources of referrals, the season, time or date that people take interest in that site's products, buying trends/patterns and lastly, whether there is likelihood that consumers review new releases and references first before sifting through product offerings.

One cannot therefore discuss BD analytics without making reference to clickstream analysis and the role it plays in assisting the leveraging of BD to create an organisation's competitive advantage.

2.13.4.2 Monetising Big Data

According to Banerjee, Bolze, Mcnamara and O'Reilly (2011), there is a great benefit in leveraging BD as other methods, especially traditional methods of selling, have been exhausted in the increasingly digitised era with which businesses exist. The Mobile industry, for example, is one such industry that has a greater advantage due to its automatic access to extremely large amounts of customer data.

There are latent opportunities that lie hidden in an organisation's data, one of which was demonstrated by OnStar and GMAC Inc. in 2007 by using data obtained via telemetry (Banerjee *et al*., 2011). The auto companies, in collaboration with an insurance company, created marketing strategies that leverage information from BD curation to create an insurance product that offers lower insurance premiums to drivers who travel fewer miles than the average consumer. Thus, there appears to be new marketing avenues available, which would otherwise be more difficult to engage without the real-time capabilities of Big Data technology.

According to Bohé, Hong, Macdonald and Paice (2013), mobile phones have become the consumer's main source of information due to versatility and accessibility, thus giving mobile operators the impetus to make use of the information harnessed from mobile devices. Mobile operators have a vast range of information available to them, which is beneficial to an array of industries simply because of the diversity of users and their needs.

Bohé *et al.* (2013) mention retail, advertising, marketing, the public sector, financial services, healthcare and other customer-facing businesses as standing to benefit from data extracted and curated from mobile operators.

In itself, the opportunity for data monetisation is vast and thus requires a more focused and detailed approach when it comes to how an organisation will create value from data. There exists therefore the need for a business to carefully strategise to enhance its ability to leverage Big Data. Bohé *et al.* (2013) suggests that an organisation needs to begin by identifying where they want to position themselves in the value market, and then ensure that they have amassed the appropriate IT solutions and analytics software before identifying collaborators and determining a plan of action.

Figure 2.27 is a diagrammatic representation of how value obtained from data is affected by both its nature and its corresponding volume (occurring data). According to Banerjee *et al*. (2011), it is possible to gain greater value at the insight and transactions level, which is the highest level up the pyramid. In this, there is an attempt to encourage business owners or decision-makers to apply careful thought before effecting a BD strategy. In other words, an organisation needs to assess its needs by deciding what kind of data they require, firstly, and secondly, by defining for themselves how they intend to use that data.

Hockenson (2013) highlights the importance of understanding the shortcomings of Big Data by mentioning the need to monitor an implemented *Big Data Framework* with strategies for curation, and also the need to ensure that security is maintained. Furthermore, she mentions that while Big Data is useful for post-analysis, it is crucial in identifying datasets which enable an organisation to predict future consumer behaviours. In answering the "how?" of harnessing the power of Big Data, Weinberger (2012:5) proposes the use of SBAs (Search Based Applications), because they possess the following benefits:

- Sentiment analysis
- Site visitor or user segmentation
- Better personalisation

> One of the main ways in which SBAs differ from Business Intelligence (BI) is in their ability to process and analyse all forms of content, as well as the behaviour of users across that content (Weinberger, 2012:5).

Therefore, SBA's add to the quantitative values normally acquired through business intelligence by simultaneously displaying qualitative explanations for the numeric data generated within a single view (Rogge, 2011).

So expansive is the data collection process that Rogge emphasises the need for any SBA's to possess service-oriented architectures (SOAs), thus enabling "integrated decision tools for each user, enabling rapid deployment and simple integration within the company's information ecosystem" (Rogge, 2011).

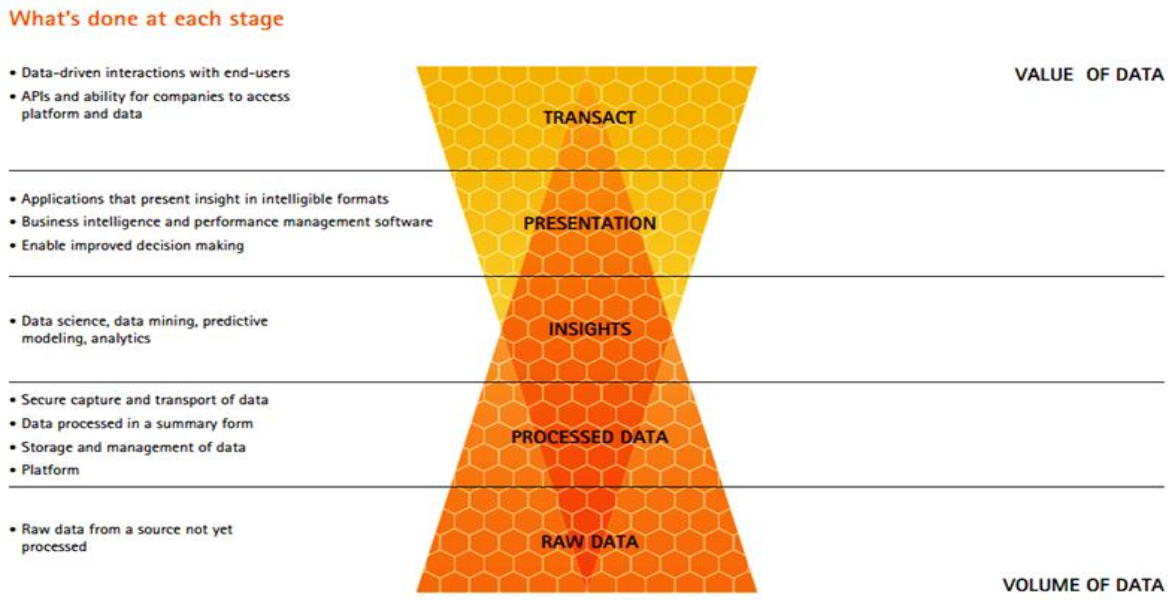


Figure 2.27: Data pyramid depicting information worth

(Source: Bohé et al., 2013:4)

The wealth of external opportunities is illustrated through the data pyramid (Figure 2.27), which overlays the stages of data processing and types of output (depicted by the colour orange) onto the value that data creates (depicted by the colour yellow). At the bottom of the chart, mobile operators can, if they choose, sell only the raw data they have in hand. A good example is the collaboration of Ford Motor Company with INRIX, a leading provider of traffic and navigation services, which now provides Ford with real-time traffic information and enhanced routing for all Ford vehicles with Ford SYNC (an in-car communications and entertainment system developed by Ford and using Microsoft technology). Of course, being a pure data provider is a relatively low-value option, minimising the opportunity to provide enhanced services.

Moving up the pyramid as data is further processed, operators begin to create applications and insight products that sit on top of the data which Bohé *et al.* (2013) refer to as data monetisation.

These offerings help generate new value-added services and create platforms for data-driven transactions such as advertisement targeting, retail payments and the provision of customer insights.

The recent launch of Telefónica Dynamic Insights is a good example. This service collects and aggregates anonymous customer data in real-time to understand how segments of the population behave as a group, thereby helping local governments and businesses make better decisions. A retailer thinking of opening a new store, for example, can see how many customers visit a given location each day by time, gender and age. In addition, Precision Market Insights from Verizon generates analytics-driven behavioural insights based on mobile engagement, location and demographics information, creating a 360-degree view of the consumer.

For outdoor advertisers, for example, such insights can measure the effectiveness of outdoor advertising units, validating the impact and reach of specific advertising campaigns. At the top of the pyramid, mobile operators create platforms on which customers become part of a transactional ecosystem. 02's Priority Moments application, for example, provides tailored, exclusive offers to its mobile customers based on detailed data provided with the customer's consent. The proliferation in BD trends is fuelled continually by the benefits of working with large datasets, thus allowing analysts to spot trends which may include combating crime, preventing diseases, forecasting and making business decisions (Hebner, 2012). For example, in clinical decision support (CDS) systems, the five tenets to effective CDS implementation, as cited in Posey (2010:4-5), include:

i) get the right information,
ii) to the right person,
iii) in the right format,
iv) through the right channel,
v) at the right point in the work flow.

A practical application here is in the administration of a drug to a patient allergic to the treatment. With the right decision support system implementation, backed by the right data source, such a situation can be circumvented. According to Manyika *et al.* (2011), multimedia and individuals with smartphones and who are on social network sites will continue to fuel

the exponential growth of BD. The current data limits are in the order of terabytes, exabytes and zettabytes accentuating the continual growth of data volumes.

Manyika *et al.* (2011) mention that the key to improving decision-making, minimising risk and unearthing valuable insights as part of leveraging BD, is in implementing strategies that allow data to be stored and used in an automated, cheap and reliable way (possibly based on what is deemed salient to the organisation, for example, a tax agency may depend on a BD analytics engine to flag candidates for further examination).

Schmarzo (2013) states that businesses need to recognise that there is no silver bullet when it comes to data storage and protection; in reality, a suite of solutions aligned to various functions is required. This data deluge requires that organisations do not only contend with volume but with a variety of data formats which, as a result, seek diversification as a means of surmounting the challenges which also require an effective data management strategy. Ivarsson (2010) states that the era of a single DBMS as a storage repository is over, and that what seem more practical and already in use, is polyglot persistence. This is in support of the old adage: use the right tool for the right job.

Agrawal *et al.* (2012) state that since data is already in digital format, curators stand a better chance of influencing the creation to facilitate later linkage automatically for better data analysis, organisation, retrieval and modelling, and by doing so, future challenges are more adeptly surmounted.

The use of relational DBMS's has been pervasive in traditional systems (Castrejon-Castillo, Vargas-Solar, Collet & Lozano, 2014). Polyglot persistence, like polyglot programming, is all about choosing the right persistence option for the task at hand; it express the need that a single data storage scheme may not suit all the needs of an application or the organisation (Castrejon-Castillo *et al.,* 2014). As described by Ivarsson (2010), a single application may use relational database structured needs, and further, where the needs of the application requires storage which is best suited to a NoSQL database system, this is implemented.

2.14 Polyglot persistence

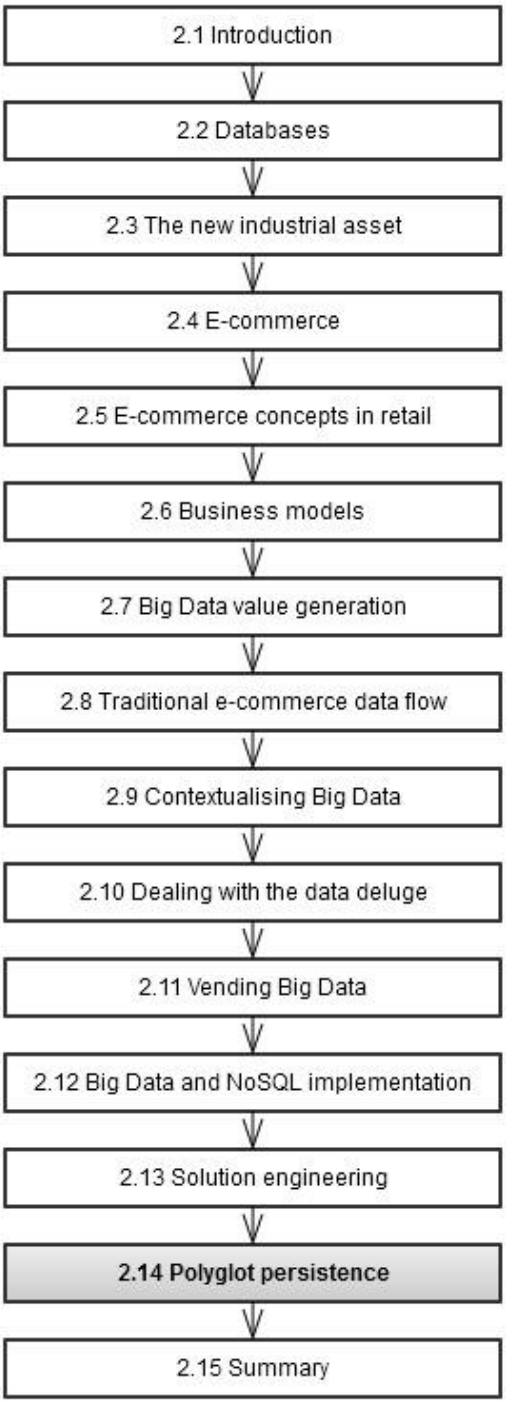


Figure 2.28: Chapter layout – Polyglot persistence

The way forward for persistence polygenism is the concept of leveraging the power of many different systems as part of the complete data curation architecture. Some authors (for example Davenport, 2013; Dumbill, 2012) refer to this as a hybrid BD ecosystem, as this concept of curation moves away from focusing on one type of system. This is because the curation platform focuses on the essential, optimised capabilities of the different available

systems to form an ecosystem of tools which form the basis of the BD architecture for data curation.

The evolution of NoSQL and databases beyond RDBMS has encouraged organisations to look beyond the traditional curation architecture. Along the same lines, it has become evident that one database does not fit all sizes and knowledge, and adoption of more than one database is a wise strategy. The knowledge and use of multiple database products and methodologies is popularly now being called polyglot persistence. Figure 2.29 demonstrates a conceptual view of a persistence polyglot called the Diggs persistence conceptual architecture, indicating the use of both relational and non-relational databases to feed curated data into the Diggs website (Mohanty *et al*., 2013).

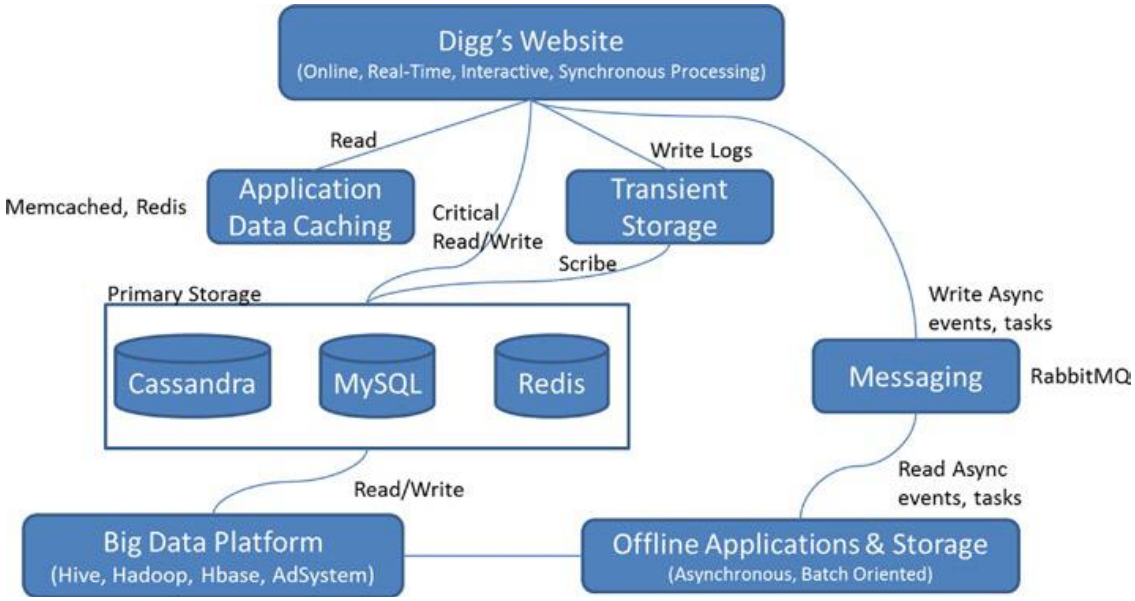


Figure 2.29: Diggs polyglot persistence conceptual architecture

(Source: Mohanty et al., 2013:100)

NoSQL databases come in many shapes, sizes, and forms and as a feature-based comparison is the first way to topically group them together. Often, solutions for many problems easily map to desired features (see Section 2.13). Mohanty *et al*. (2013:99) define polyglot persistence as "the use of both an RDBMS and one or more NoSQL databases as the database management layer for modern applications". Figure 2.29 is a typical example of a persistence polyglot used by Digg, a social news website. Digg's business model or approach to business allows users to recommend, share and discover web content. Its differentiating lies in presenting the consumer community with popular web trends from around the Internet in an aggregated fashion.

The conceptual view of the persistence layer is comprised of a Facebook connect, the Digg dialog, DigBar, Digg ApI and Digg App (further details in Annexure 7.8). Polyglot persistence is essential in today's data curation approaches, but it introduces unforeseen complexities in creating, implementing and managing the environment. It becomes vital to understand, with adequate internalisation, how each facet of the implementation works to achieve consistency and availability.

2.15 Summary

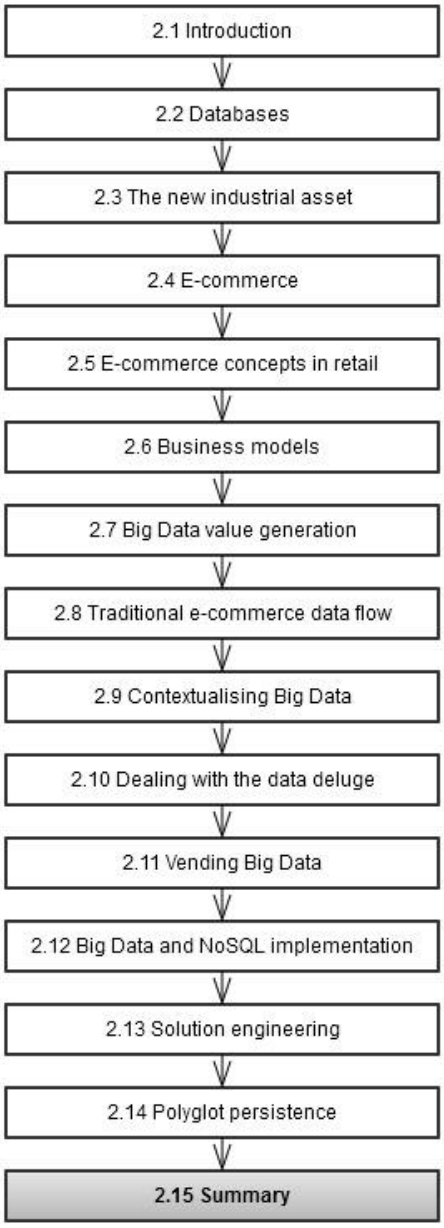


Figure 2.30: Chapter layout - Summary

The definition of BD encompasses velocity, volume, and variety. Over time, it evolved to include a fourth attribute being accuracy. The growth of Big Data can be attributed to forces on the online market such as consumers, products, process automation, technology

advancements and data monetisation which are the centre pillars of data curation. This chapter reviewed Internet marketing models, BD value generation, traditional e-commerce data flow, and addressed the application of big data by contextualising the phenomenon.

This chapter looked at the disruption as opportunities to encourage organisations to achieve economic gain once certain criteria were met. Dealing with the data deluge is not an issue that can be addressed singly by any particular platform. It is rather an ecosystem of tools which some organisations call the hybrid or polyglot—Solution Engineering as a means of modularising a problem situation, intelligently throwing light on it and drawing what steps are needed to accomplish the solution state in a real life setting, and polyglot as the way forward, for organisations seeking to benefit from the influx using BD curation.

In the next chapter, Chapter Three, the research design and methodology of the study will be described. This will be done by indicating the chosen philosophical paradigm by which the research was conducted.

CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY

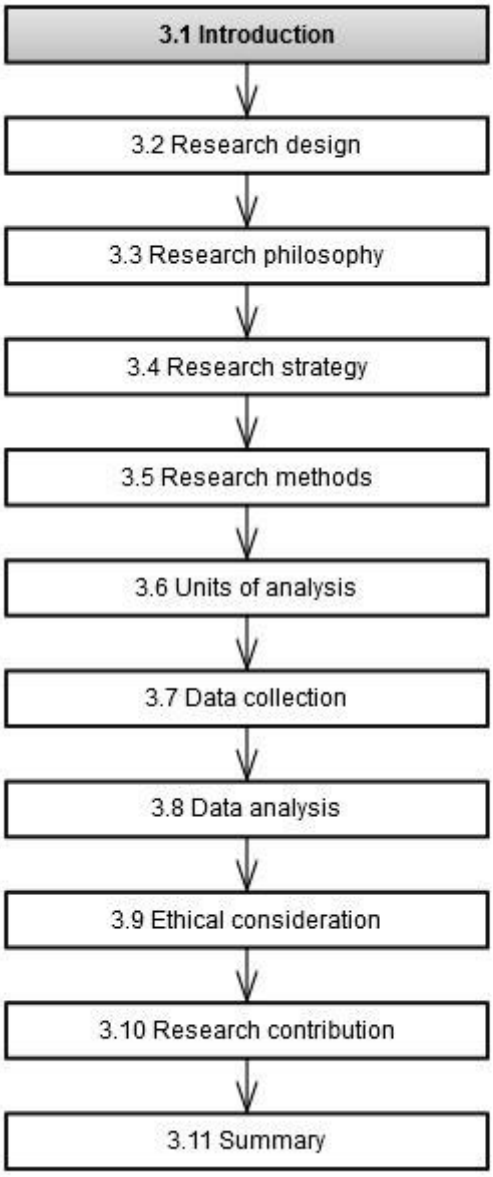


Figure 3.1: Chapter Three layout - Research design and methodology

3.1 Introduction

This chapter covers the research methodology. It provides an outline of the research as a single case study with emphasis on the philosophy, research approach, strategies, interviews, data collection and review of the analysis method deployed in the study.

This section also covers ethical considerations and the contribution of the research to the body of knowledge. The chapter ends with a summary of the entire chapter.

Upon approval of the research proposal and ethical clearance to conduct this study, phone calls were made and emails were sent to department heads at Spree informing them of the research and possible benefits to Naspers Media24 and Spree to seek collaboration and canvass participatory support. Approval to commence with the research was given by the university's Faculty of Informatics and Design and also by the management of Spree, formerly Touchlab, at the inception of this research study.

3.2 Research design

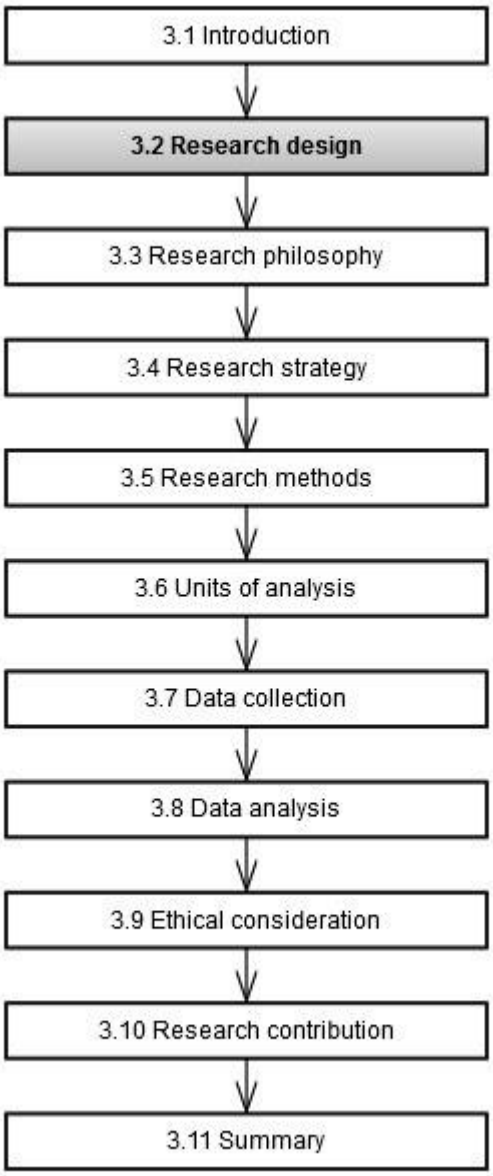


Figure 3.2: Chapter layout – Research design

Qualitative researchers are interested in learning the meaning people have formed to make sense of their world and the experience they have of the world around them (Merriam, 2009:13). Parkinson and Drislane (2011) contextualise qualitative research by asserting that research is using methods such as participant observation or case studies which result in a narrative, descriptive account of a setting or practice. The process involves acquiring scientific knowledge by means of various objective methods and procedures which are scientific and lead to verifiable conclusions. The methods stipulate procedures for drawing, measuring, collecting and analysing data.

There are many methods and techniques but the aim of the research determines which methodology is eventually chosen as the means of creating knowledge. Interpretivist research contends that only through the subjective interpretation and intervention in reality can the world be fully comprehended, as described in Section 1.7.1.

Research methodology allows researchers to explain and analyse research methods in an organised and systematic way, it explains the logic behind research methods and techniques (Welman, Kruger & Mitchell, 2006; Parkinson & Drislane, 2011). According to Neuman (2011), research methodology is a broader term that encapsulates the entire research process, including its social-organisational context, philosophical assumptions, ethical principles and political impact of new knowledge from the research enterprise. Research, according to these authors, is highly relevant for understanding social life in general and the decisions people make. The process yields knowledge that is organised into theories and grounded empirical data. Neuman (2011:9) refers to this theory as:

> …coherent, systematic, consistent and interconnected ideas used to condense and organized knowledge.

The research design involves choosing methods that will apply to answering the research questions. The design can be thought of as the blueprint of the research as it details and even determines the methodology, participant sampling and data collection. According to Saunders *et al.* (2009:132), as depicted in Figure 3.3:

> …it is essential to clearly identify explicitly with the different layers of the ring the chosen philosophy, which implies the approach through to data collection and analysis.

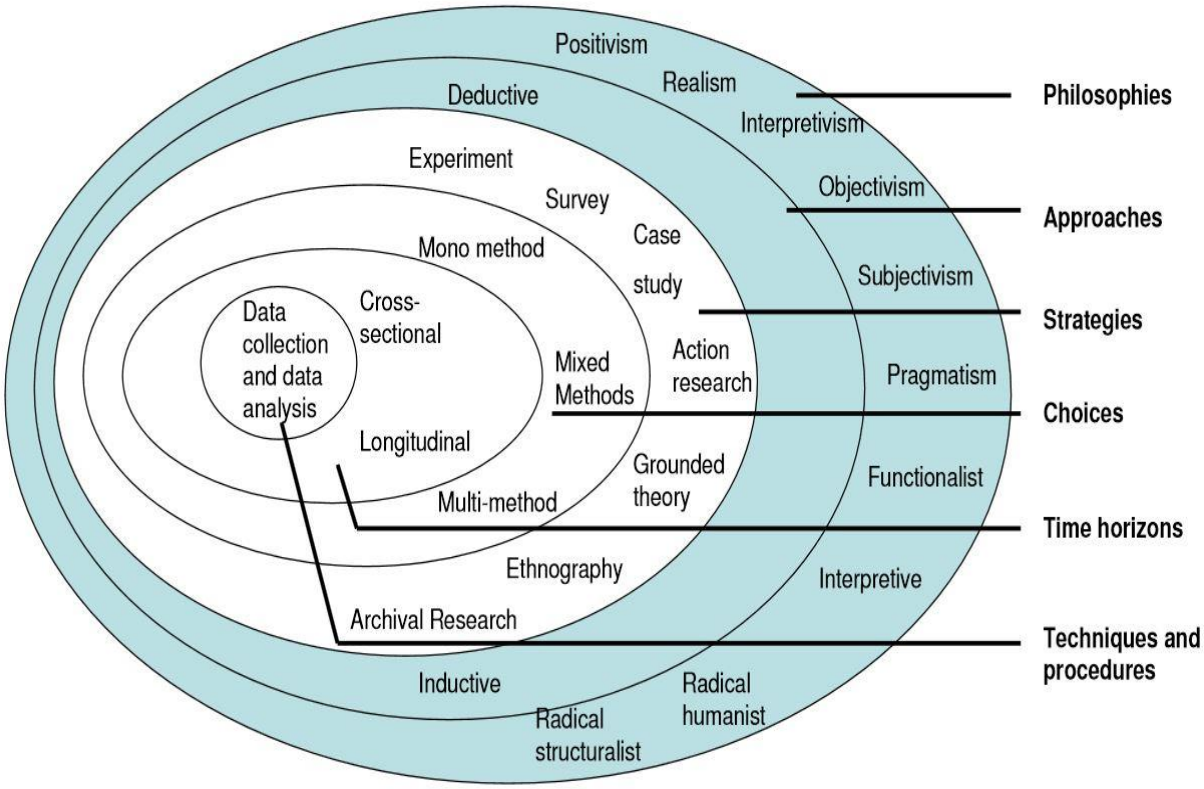


Figure 3.3: The research onion

(Saunders et al., 2009:132)

The following sections focus on the philosophical stance, approach and strategy.

3.3 Research philosophy

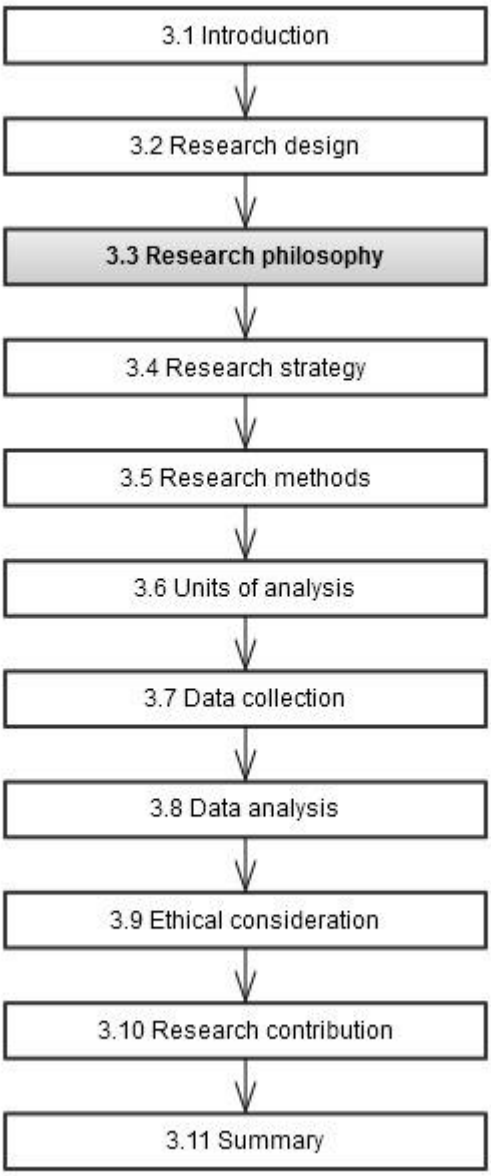


Figure 3.4: Research philosophy

3.3.1 Research philosophy

Researchers differ in their assumptions about what is important to study, what can be known, what tools and designs are preferred, and what standards may be used to judge the quality of the research (Everest, 2014). Research philosophy is a scientific and systematic search for knowledge existence, reason, values, mind and language (Peterson, 2014); it is a belief about the manner in which data about a phenomenon should be gathered.

According to Bandaranayake (2012), research philosophy is an over-arching term relating to the development of knowledge and the nature of that knowledge. As stated earlier, this research adopted an interpretivist position. The primary aim of this research is to explore and understand what perspectives, perceptions, thoughts and feelings exist throughout the organisation concerning BD curation for insight, interpreting emerged findings along the lines of academic research and theories to create knowledge. The philosophical underpinnings of research, which are ontology, epistemology and methodology, categorise researchers (Everest, 2014:8). These were discussed in Sections 1.7, 1.7.1.1, 1.7.1.2 and further in Sections 3.3.2, 3.3.3 and 3.3.4.

This study is inductive rather than deductive, and builds theory rather than theory testing, given that this case study is highly contextual. Consequently, the opportunity to generalise is far removed. Since an interpretivist stance places emphasis on communication and language, this approach seems particularly suited to the focus of this study.

3.3.2 Ontology

Ontology is the study or science of being. This definition originates from the social sciences to encompass claims about what things exist, how the things look, the units, and how the units interact with each other (Neuman, 2011:92). Ontology concerns the issue of what exists or the fundamental nature of reality. There are two basic positions within ontology, namely realist and nominalist (Neuman, 2011). The realist postulates that the world is organised into pre-existing categories waiting to be discovered; the realist believes that the “real world” exists independently of humans and their interpretations. The nominalist assumes that humans never directly experience a reality out there, and that what researchers consider to be world, occurs through a lens or a series of subjective interpretations. Ontology is concerned with how the world is built, that is, what the nature is of the social and political context that knowledge might be acquired about.

These different stances lead to epistemology which describes how the researcher knows the world around him or what makes the claim to be true, and how the researcher can learn or know how the world is rooted in a specific ontological assumption. The ontological stance of this research presupposes and draws on an interpretive paradigm. Ontology assumes that knowledge creation transitions from observation to understanding from the external to the internal point of view, meaning the research is resistant to the naturalisation of the social world.

3.3.3 Epistemology

Epistemology is an area of philosophy concerned with the creation of knowledge. It focuses on how we know what we know or what are the most valid ways to reach the truth (Eriksson & Kovalainen, 2008; Neuman, 2011:93).

Reaching truth from an epistemological stance requires inquiring if knowledge as facts can be externalised and communicated as facts or rather experienced personally. Epistemology closely relates ontology and its considerations of what comprises reality. It considers views of ways of enquiring into the nature of the world, what knowledge is, and its sources and limits. In short, epistemology includes what the researcher has to do to produce knowledge and what scientific knowledge will evolve into once produced. This knowledge can be produced viewing the world from the two stated beliefs, the realist and the nominalist. Tharakan (2006:16) divides epistemology into two positions, namely positivism and anti-positivism. From the stance of the positivist, hypotheses are used to prove or falsify claims while anti-positivism refutes gaining knowledge by being an external observer of social activities. Anti-positivism argues that one needs to become involved personally in the activities in order to gain an understanding from the inside rather than the outside (Peterson, 2014). This research study followed a phenomenological paradigm, this is, a qualitative approach that furnishes the research process with rich and subjective information from a natural setting as opposed to controlled settings as suggested by proponents of a deductive approach.

Saunders *et al*. (2009) state that researchers should use a deductive approach in which the researcher develops a theory and hypothesis, and design a research strategy to test the hypothesis or inductive approach, where data is collected and a possible theory developed as a result of the data analysis. Saunders *et al*. (2009) further suggest that an inductive approach is more suited to the interpretivist. According to Thomas (2003:2), an inductive approach is:

> …a systematic procedure for analysing qualitative data where the analysis is guided by specific objectives which include condensing extensive and varied raw text data into a brief, summary format; establishment of clear links which are transparent and defensible between the research objectives and the summary findings derived from the raw data; and the establishment of a model or theory about the underlying structure evident in the text (raw data).

An inductive approach is driven by underlying assumptions which include the development of categories from the raw data into a framework that captures the key themes. Through a range of techniques, the trustworthiness of findings can be assessed using independent replication of the research, triangulation within the project, and feedback from participants,

among others (Thomas, 2003). The development of categories within this inductive research is summarised in Table 3.1.

Table 3.1: Categories developed from coding

Phase	Description
Label of category	Inherent name use to describe category.
Description	Description of key characteristics, scope and limitations.
Text associated with categories	Text examples coded into categories.
Link	Indicate relationship with other categories.
Type of model where category is embedded	Incorporating category system in a model or framework.

This study took an inductive approach since data was gathered and analysed to develop guidelines on data curation and recommendations on BD curation.

3.3.4 Interpretive and critical approach

According to Orlikowski and Baroudi (1991), interpretive proponents assume that people create and associate their own subjective and inter-subjective meanings as they interact with the world around them. Interpretive researchers thus attempt to understand phenomena through accessing the meanings participants assign to them.

According to Hatch and Cunliffe (2006), Interpretivism is anti-positivism since it is contended that there is a fundamental difference between the subject matter of natural and social sciences. Meaning is constructed and over time reconstructed through experience, resulting in many different interpretations. This forms the basis of a myriad of interpretations that create a social reality in which people act. In this research paradigm it becomes a standing pillar for people to discover the meaning and context of factors that influence, determine and affect different individuals.

3.4 Research strategy

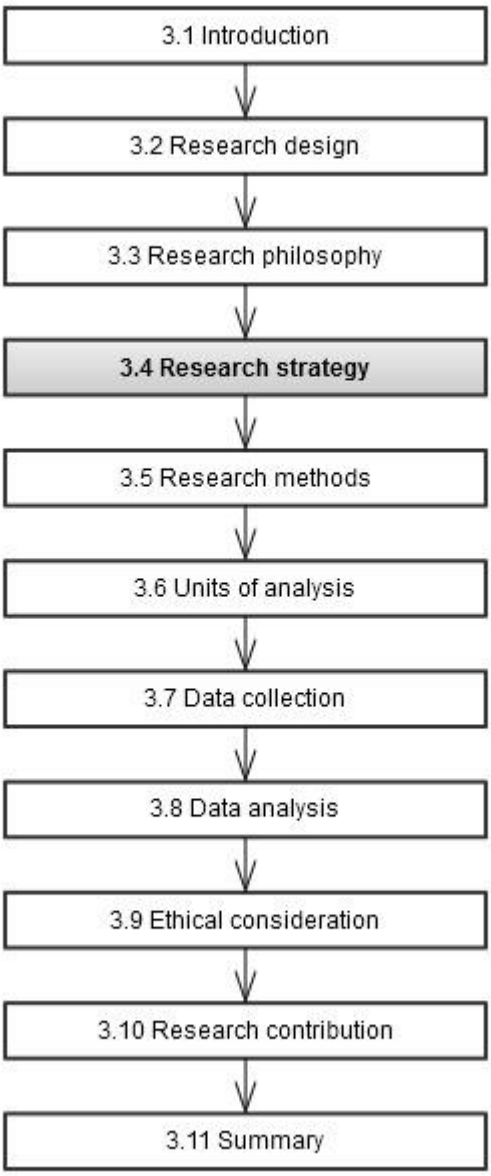


Figure 3.5: Chapter layout - Research strategy

Peterson (2014) mentions that:

> …there are many different research strategies deployable in a particular scientific research study; the choice of selection depends on the ontological and epistemological stance of the research.

This research is qualitative, based on a case study. A case study is used as the research covers the data curation practices of a particular company and how this can be improved with the influx of data in industry. According to Yin (2014:2), a case study is a strategy for doing research which involves an empirical investigation of a particular contemporary

phenomenon within its real life context using multiple sources of evidence. This is affirmed by Yin (2003) as suitable for studying complex social phenomena. This research is a single case study of Spree, a subsidiary of Naspers media24. The company was selected based on convenience as the researcher was an employee of the company.

3.5 Research methods

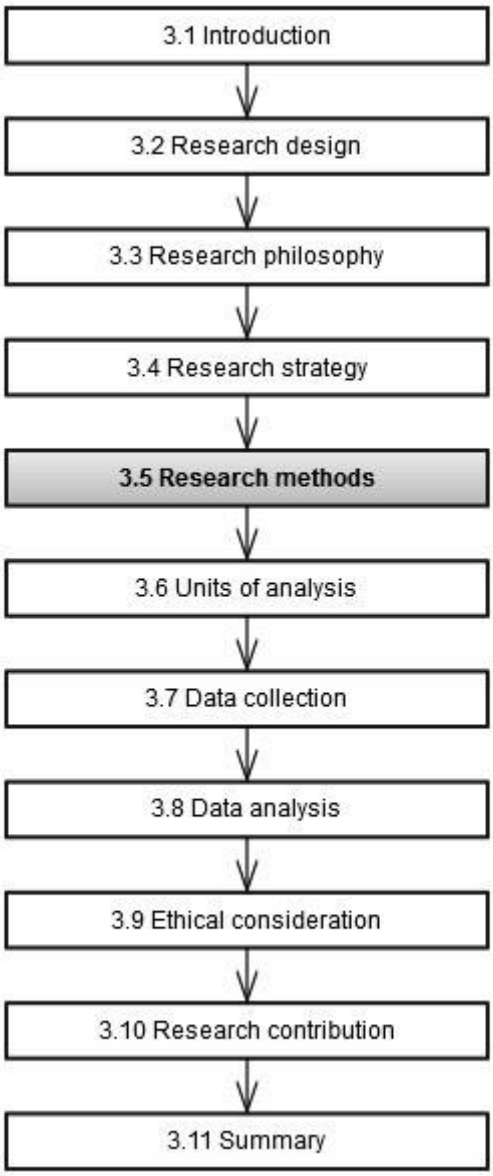


Figure 3.6: Chapter layout - Research methods

Section 3.8.1 discusses qualitative and quantitative research and justifies choosing a qualitative method as a means for this study.

Zikmund, Babin, Carr and Griffin (2013:5) categorise scientific research methods to be "qualitative and quantitative methods" and define research as "scientific methods of searching the truth concerning a phenomenon". For research to be relevant and applicable,

it is mandatory to match the study to the appropriate scientific methods and use the right tools in the search process.

3.5.1 Qualitative and quantitative research

Quantitative research examines relationships among variables—a means of testing objective theories. This research design method has an element of numerical representation and manipulation of observations mainly for describing and explaining what the phenomenon reflects. The registered variables are computed instrumentally for analysis using statistical methods. What is apparent in this kind of research design is its implicit protection against bias as stated by Creswell and Clark (2007). This approach encompasses (i) testing theories deductively, (ii) control of alternative theories, and (iii) the ability to replicate findings.

Qualitative research on the other hand, as presented by Creswell and Clark (2007), allows for the exploration and understanding that may be ascribed to a social or human problem. According to Nkwi, Nyamongo and Ryan (2001:1), qualitative research involves any research that uses data which does not indicate ordinal values. The emphasis here is on the type of data generated, which is non-numerical, and involves analysis and interpretation of observation, purposely for discovering patterns of relationships and underlying meanings. Researchers who subscribe to this design support a way of looking at research that honours an inductive style, permeating meaning and rapidly demystifying the complexity of topics through meaning-making. The aim is to elucidate, making sense of the area of interest, while increasing the knowledge available about the study.

In conducting qualitative research, the primary goal of the researcher is to describe and help understand rather than explain. It can also be viewed from a broad methodological approach in sharing certain principles or logic. According to Nkwi *et al.* (2001), qualitative research is conducted in a natural setting of social actors with the focus on process rather than outcome. The insider view is emphasised. The primary aim is to ascertain in-depth descriptions and understanding of actions and events by paying specific attention to social action in terms of its specific motive rather than a generalisation of theory pertinent to the study.

In the setting of qualitative research, naturalism is the style and theory of representation based on the accurate depiction of detail and process. Naturalism also encompasses many other factors such as the insider view, description and understanding, contextual interest, idiographic strategy and the inter-subjective nature of the research. The inductive nature is a prime factor which sets it apart and establishes the strengths of the approach. The emphasis on a natural setting in qualitative research indicates its suitability to a social process given the time inclusive factor, meaning the researcher is well-positioned to ascertain in-depth

information as it occurs rather than base the study on a reconstruction of the occurrence. This strengthens the insider view by placing the research in context. According to Babbie and Mouton (2009:271), seeing the research through the eyes of the actors relates to the phenomenological roots of qualitative research. That is, phenomenologists view human behaviour as a product of how people interpret their world. The role of the researcher here as the phenomenologist is to capture the interpretation by placing himself in the shoes of the actor.

Description and understanding places the study in context relevant for contribution to the body of knowledge through clear descriptions. The focus here shifts quickly from counting and quantifying patterns to descriptions and meaning-making. This is summarised in the rich elaborate descriptions originating from qualitative research. To stay true to the categories and concepts of the actors, the qualitative researcher becomes couched in everyday terminologies of the actors rather than to abstract theoretical constructs foreign or not easily decipherable by the actors. This aligns with the contextual interest and preference for understanding events, actions and processes in context (Neuman, 2011).

The interest in context by a qualitative researcher aligns the researcher more to an idiographic rather than nomothetic research strategy as per the works of German hermeneutist, Wilhelm Windelband. He states that, in their quest for knowledge of reality, the empirical sciences either seek the general in the form of the law of nature, or the particular in the form of a historically defined structure. Researchers are concerned with a form that invariably remains constant. That is, scientific thought is nomothetic while the latter is idiographic, describing a qualitative approach (Kaplan & Maxwell, 2005).

Implicit in a qualitative study are elements of analytical strategies such as grounded theories and discourse analysis, which are inductive in nature. The qualitative researcher inductively builds and develops interpretations which are based on first-order descriptions of events rather than deductively derived research hypotheses. According to Kaplan and Maxwell (2005), analysing textual or pictorial data is of prime importance as data can be produced by a range of techniques including in-depth interviews, focus groups, participant observation in which the researcher participates, audiotaping naturally occurring conversations, among others.

The primary data of this research study was collected by means of interviews and literature analysis. The research interviews spanned the organisational hierarchy from top management responsible for strategic decision-making to every day tactical decision-making staff to ascertain especially how participants use curated data to guide decision-making. Interviews were less structured and focused on revealing meanings and new insight.

Techniques used in the research focused on helping discover the primary themes; these themes hold human motivations and document activities that are in most instances thorough and complete (Zikmund *et al.* 2013). Table 3.2 presents a tabulation of common qualitative research techniques.

Table 3.2: Qualitative research techniques

Technique	Description
Focus groups	Group discussion, largely unstructured, coordinated and controlled by a trained moderator. Widely used because it is easy to execute and quite adept at generating better insight.
Observation	This technique allows for observation and recording of events in a natural setting without disturbing participants.
Conservations	Dialogue with targeted respondents that get recorded by researcher. Very unstructured.
Depth interviews	Face to face interview between researcher and respondent. Leads to deeper probing and better insights.
Semi-structured interviews	This technique makes use of flexible open ended questions where a respondent can answer providing with brief descriptions or explanations in a verbose manner. Results can easily be contextualised and interpreted.
Cartoon tests	This technique uses images as the respondents generate ideas out of cartoons or ambiguous images.

Some of the advantages of using the qualitative research method include the provision of rich textual descriptions hence understanding that the phenomenon is contextualised in great detail and depth.

The selected methods for this research were (i) qualitative research through observation, and (ii) in-depth interviews by means of semi-structured questionnaires.

3.6 Units of analysis

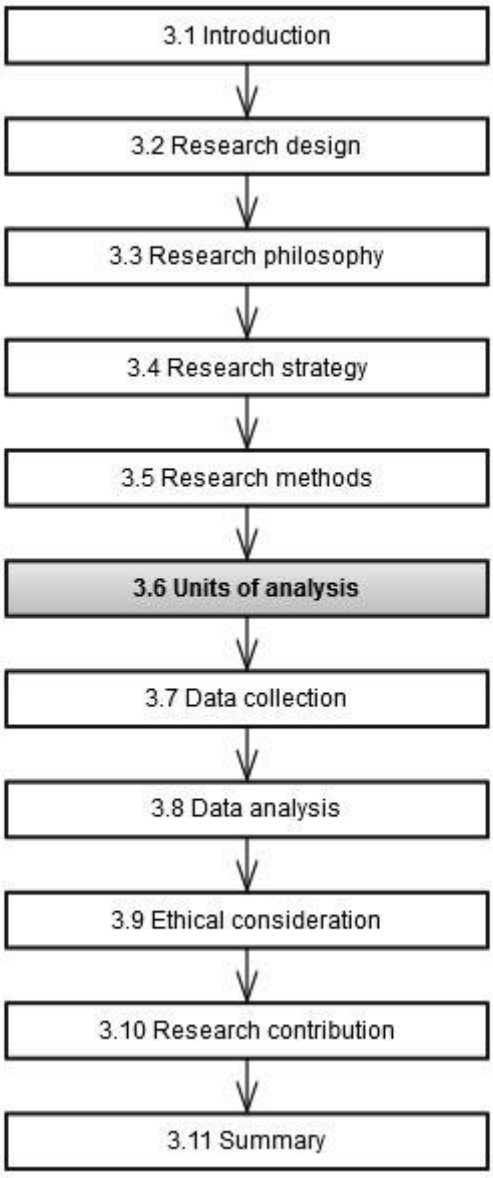


Figure 3.7: Chapter layout – Units of analysis

The units of analysis, as stated by Yin (2014:55), are those things we examine in order to construct summary descriptions and to explain differences among them. The **units of analysis** in this study are the departments in Spree, while Spree is the **case**. The **units of observation** are the employees. Participants for interviews were selected purposively as a common sampling strategy according to preselected criteria relevant to the research by the various departments on the bases of interaction with data and foreknowledge of data. It is important to mention that participants sample size was not fixed prior to data collection as this is depended also on the availability of the participants and time allocated.

Interviews were immediately and progressively transcribed to afford the researcher the ability to determine theoretical saturation, which is a point in data collection where no new data or insights are generated.

3.7 Data collection

Neuman (2011) divides data collection into two categories based on the type of data collected—quantitative which is data collected in the form of numbers and qualitative based on data collected in the form of words or pictures. Qualitative data comes in a vast array of forms which include photos, maps, interviews, observations and documents. Qualitative data can be divided into field research and historical comparative research. Data from this research, apart from literature analysis, is obtained through interviews, documents, reports and observation of participants.

3.7.1 Primary data

Primary sources are unpublished data which the researcher gathers from participants or the organisation directly (Kadam, Shaikh & Parab, 2013). These sources may include transcripts from interviews, observations, and documents collected from the case organisation in the form of scheduled subscription and *ad hoc* reports for different departments, versions of entity relationship diagrams on the Magento database called the Magento entity attribute value diagram (EAV), and a traditional data curation diagram forming a guide to traditional curation. Interviews and observations were used to gather primary data in this study.

3.7.2 Secondary data

Secondary data in this study was acquired through literature studies from sources such as publications, research institutions, Internet websites and reports. The secondary data formed the bases of the research as it provided the basic understanding of terms and concepts in the field of BD, data curation and the current industrial view of data relative to retail and competition. Secondary data is useful to augment and serve as a control for primary data in identifying technology implications from a literature perspective. It supports the textual data obtained from primary data and even validates the data.

3.7.3 Interviews

An interview is defined as a purposeful discussion between two or more people (Saunders *et al*., 2009). The qualitative interview is the most common and one of the most important data gathering tools in qualitative research (Myers & Newman, 2007). Interviews are pre-planned face-to-face structured or unstructured questioning of participants, conducted in groups or

individually. According to Saunders, Lewis and Thornhill (2007), interviews may either be structured or unstructured. Structured interviews have pre-organised questions which are planned ahead of time whereas the later has topic areas to explore but questions or order of questions are pre-planned. Figure 3.8 illustrates interviews as discussed by Saunders *et al.* (2007:313).

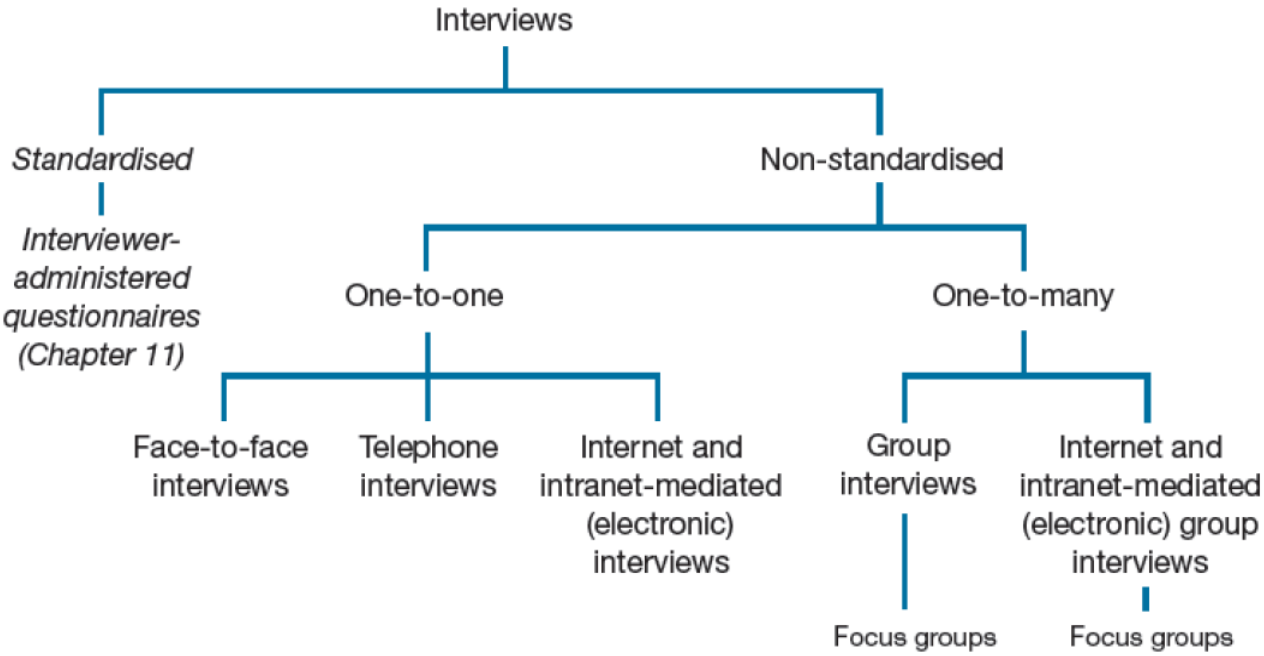


Figure 3.8: Types of interviews

(Source: Saunders et al., 2007:313)

The duration of an interview in this research averaged more or less an hour, with interviewees giving consent to record the interviews. Interviews were conducted on a one-on-one basis in a semi-structured setting; some interviews had multiple participants while others were a follow up to get more clarification. Interviews were held at the convenience of the interviewees at a prescheduled date, time and venue which were mainly the offices of the interviewees or meeting areas of the organisation. The recorded interviews were later transcribed and used as part of the primary sources of data for analysis.

3.7.3.1 Guidelines for qualitative research interview

Myers and Newman (2007) mention seven guidelines for carrying out qualitative research as tabulated below with brief a description of each guideline.

Table 3.3: Guidelines for qualitative research

(Source: Myers & Newman, 2007:16)

Number	Guideline	Explanation
1	Situate researcher as actor	Set researcher as actor. Researcher may consider asking the following questions to provide reason and context to answers: • Who are you? • What is your involvement in the organisation? • Level of experience, gender, age and possibly nationality This helps the reader to ascertain the validity of the research.
2	Minimising social noise	Interviewer should ensure interview occurs in a setting where participants are well set and comfortable. This constitutes minimising noise. According to Myer, *et al.* (2007), this improves disclosure.
3	Represent various voices	This includes enforcing objectivity by representing in a fair manner all voices within the interview setting.
4	Everyone is an interpreter	Allow interpretations to emerge without bias.
5	Use mirroring in questions and answers	Focus on subject while allow participants to mean what they intend to bring across by not distracting them or imposing thoughts upon them.
6	Flexibility	Be flexible while leading the conversation in an unstructured way. Allow discussion to evolve and be ready for surprises.
7	Confidentiality of disclosures	Provision to keep records and transcripts confidential.

The research to a wider extent ensured that the guidelines were followed to facilitate interviews. Firstly, dissonance was minimised to the point of insignificance as all interviews were conducted in a setting where the interviewees were comfortable and away from any possible interruptions. Interviewees were asked about their involvement in the organisation and their experience; consent forms were signed and other recommended guidelines were followed.

3.7.3.2 Explanation of the guidelines

The interview guide forms the base set of questions used as a mechanism to gather information from interviewees. These questions were broken down into the pre-interview questionnaire which allowed for scaling and assignment of users to business or technical interview groups for suitability. At the start of interviewing, participants expressed discomfort and unawareness to the seemingly technicality of the whole BD concept. Technical questions were addressed to technical data curators and business personnel who were non-technical participants. Management level participants with no technical orientation addressed non-technical questions.

3.7.3.3 Interview questionnaire

Interview questions were formulated to enable the researcher to gain first-hand information about the organisation and participants. Participants interviewed were from the six previously mentioned departments with varying knowledge of data. Big Data and data were sometimes used interchangeably. In some situations, it was imperative to clearly distinguish between BD and data as participants insisted they had data but not BD. Interview guidelines are presented in Tables 3.4, 3.5 and 3.6. Due to the semi-structured natured of interviews, interview questions differed per participant as a base to draw as much as possible insight from interviews.

Table 3.4: Pre-interview question

Interview number	Pre-interview description
1	What is your tactical and strategic involvement with BD?

Table 3.5: Sample business interview questionnaire for participant 12

Interview number	Business interview description
1	What type of data do you curate?
2	How does it affect the Touchlab environment and introduce a competitive edge?
3	Is Touchlab curating BD?
4	How is data or BD contributing to growth of your domain?
5	Are there any policies or strategies outlined for the curation of BD?
6	Do you have access to real-time data for decision-making and how does real-time data affect decision-making?
7	What kinds of decisions do you need to make to be competitive and remain competitive?
8	How important is a data model to communicate ideas?
9	Relative to destination thinking, how do you relate data and results, statistical results, and identifiable results?

Table 3.6: Sample business interview questionnaire for participant 2

Interview number	Business interview description
1	What kinds of data do you curate to provision business?
2	How will improving the timely availability of data improve decision-making?
3	What data challenges do you have?
4	Are you curating BD and how?
5	What are some of the KPIs across departments?
6	Is the organisation curating BD?
7	What technologies are in place for the curation of BD?
8	How will data curation contribute to the growth of this company?
9	What is the businesses' view of BD in terms of competitive advantage?
10	Assuming we should have a BD curation platform implemented, what are some of the policies that will be in place?

Interview number	Business interview description
11	What data curation modules do we have in place?
12	Does the organisation have a data curation framework as a guideline for data curation?
13	How often does the organisation do an audit of data for a single view of data?
14	Relative to destination thinking, how do you relate data and results in terms of statistical results and identifiable results?

3.8 Data analysis

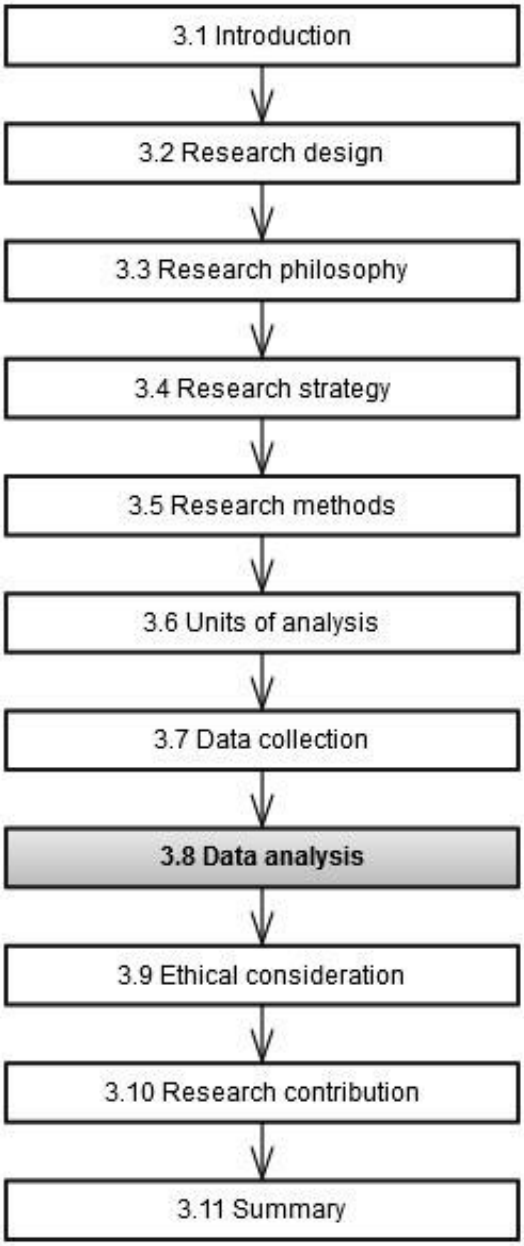


Figure 3.9: Chapter layout – Data analysis

The research data was summarised from transcripts, and categorised and developed into themes from findings. Data analysis provides the researcher the opportunity to review raw data, organise and order the data into useful contextualised facts called information for presentation. This process forms the basis of knowledge creation as data is broken into manageable patterns and trends with correlations that bring across meaning and relevance while revealing tangibility and connections. The following data analysis methods were used:

3.8.1 Hermeneutics

As posited by Myers (2009), the primary focus of hermeneutics is the analysis of text to achieve coherence in presenting research findings. Qualitative content analysis is mainly about textual data, and textual data as subject to interpretation might imply that every researcher might interoperate text differently. Hermeneutics provide a way of understanding and interpreting text in a philosophical way. The researcher in this case study is faced with having to objectively interpret text from interviewees through meaning formation with context application as participants from different departments might use different terminologies, it is important to bring out the true meaning as implied. This method according, to Kinsella (2006), aims to see what scientific methods of analysis can do with worldly experiences. Hermeneutics was used as a part of data analysis.

3.8.2 Conversation analysis

According to Gray (2013), conversation analysis (CA) includes the analysis of natural texts (often the result of transcribed tape recordings). In this research all interviews where recorded and transcribed for textual analysis. CA specifies the formal principles and mechanisms of how participants express themselves in social interactions; the reason of interest for this method of analysis is its alignment with dissonance minimisation in Research Guideline 2 as mentioned in section 3.10.4.1. CA goes beyond analysing research data into aiding the interview process as elimination or complete avoidance of noise is essential to facilitate gaining information from interview participants. This forms part of the social setting, created to host the actor; it allows the researcher to establish a level of validity through the introduction of the actors. CA uses audio recordings and videos made from naturally occurring interaction yielding descriptions of recurrent structures and practices of social interaction. This research made use of audio recordings with prior consent of participants and hand written notes to collect data. Peräkylä (2008:1) mentions three dimensions of applicability which are focused on action, structures of explication and the investigation of intersubjective understanding.

3.8.3 Content analysis

According to Stemler (2001), CA is “a technique for making inferences by objectively and systematically identifying specified characteristics of messages”. CA is mainly a coding operation with coding being the process of transforming raw data into a standardised form (Babbie & Mouton*.,* 2001). Its main advantage is that it is applicable to large volumes of textual data and different textual sources can be dealt with and used in corroborating evidence (Hoskins & Mariano, 2004). According to Elo and Kynga (2008), content analysis is any technique for making inferences by systematically and objectively identifying special characteristics of messages. The main focus of content analysis is on elements of validity and this has its basis in the trustworthiness of the analysis process as stated by Elo and Kynga (2008). According to Kohlbacher (2005) as well as Yin (2003), CA is a valid tool to analyse qualitative research data from a case study as the case study will provide a multi-dimensional perspective that may create a shared view of the situation being studied; therefore case studies provide the opportunity for a holistic view of a process. Kohlbacher (2005) mentions that besides validity, the strengths of content analysis in qualitative data lie in the openness and ability to deal with complexity and integration of context. It also provides the ability to integrate different material or evidence in support of findings; furthermore, because classical content analysis has its basis in quantitative analysis of data, content analysis preserves the advantage of classical quantitative content analysis thus including the integration of quantitative steps of analysis. This research deploys content analysis for coding of text into classifications and subsequently into findings and themes.

3.9 Ethical considerations

According to Stemler (2001), ethical issues in social research are both important and often ambiguous, spurring a necessity for strict compliance to the code of ethics of scientific research and of the involved organisations. Stemler (2001) quote Webster New World Dictionary as defining ethics as “conforming to the standards of conduct of a given profession or group.” Some of the ethical issues pertinent to this scientific research include anonymity and confidentiality, responsibility to sponsors and ethics committee.

This research study adheres to the code of ethics of Cape Peninsula University of Technology as well as Naspers Media24. This is in the form of signing the necessary legally binding documents (Non-Disclosure Agreement) stating the rights of the organisation and the Cape Peninsula University of Technology as the sponsors of this research study.

3.10 Research contribution

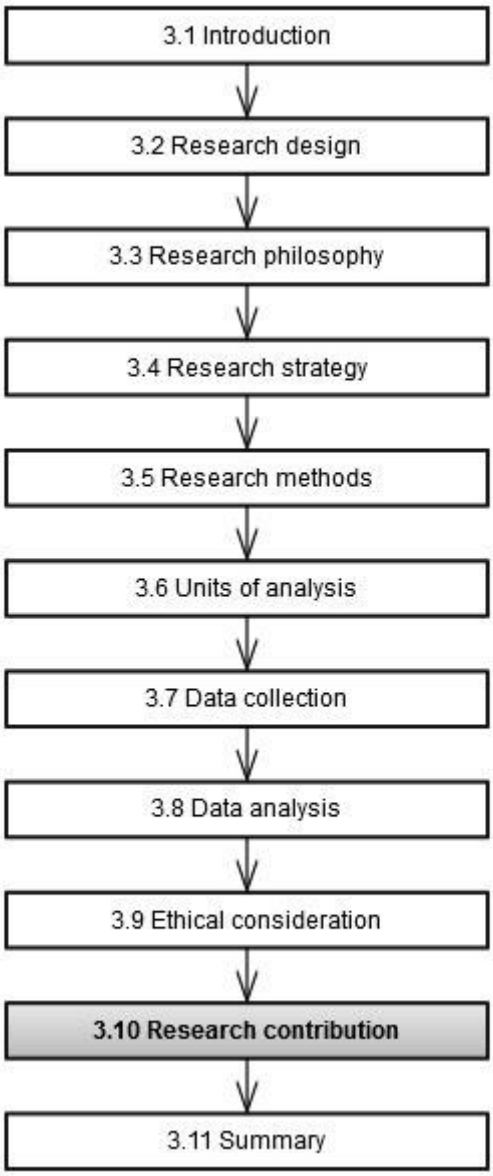


Figure 3.10: Chapter layout – Research contribution

The contribution of this research is to create guidelines for the curation of BD in a multimedia organisation. The research ascertains the current curation practices and how data is leveraged to gain a competitive advantage and further draw insights leading to guidelines and recommendations that will help improve the current data curation practices. This study enables meaningful conversation between technical data curators, operation level managers, technical personnel, and strategic and tactical workers.

3.11 Summary

A subjectivist view was taken in this study to make sense of the environment to help generate insightful meanings and conclusions. Interpretivism formed the epistemological stance for this research. An inductive approach has been taken as data was gathered and analysed to create curation guidelines and recommendations. A case study was used as a research strategy to focus on the phenomenon. The case study chosen was Spree and the units of analysis were the employees. In-depth interviews, a literature analysis and observation were conducted to collect data for the study. Hermeneutics, conversation analysis and narrative analysis were used to analyse data so that useful information could be extracted.

In the next chapter the findings of the study will be discussed. This will be done by using the research questions, sub-research questions and interview questions as guidelines.

CHAPTER FOUR: RESEARCH FINDINGS

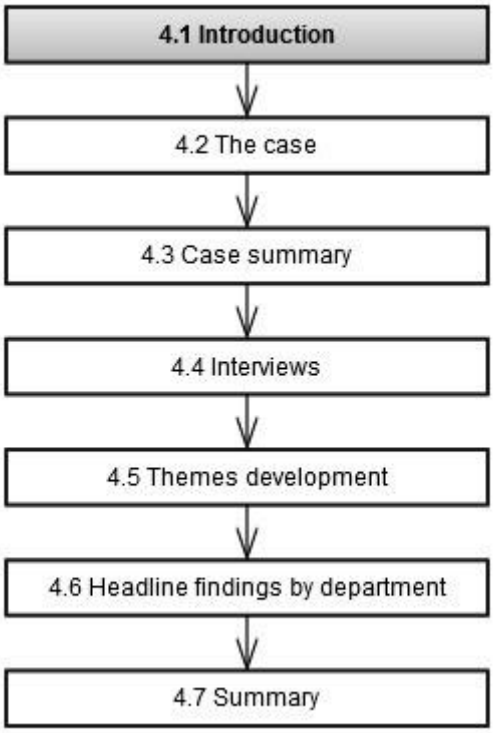


Figure 0.1: Chapter Four layout – Research findings

4.1 Introduction

Many findings emerged from pre-interviews, business interviews and technical interviews. As part of pre-interview findings, all participants acknowledged data as being one of the most important business assets as it no longer showed to be a business by-product but a pivotal business component that has taken centre stage both in decision-making and in business processes. Participants concurred that data driven decision-making has become a necessity for the successful operation of any business unit such as the department. Decisions made based on evidence from curated data should improve over time to value creation and enable departments to make quicker decisions. This is because the analysis of data reveals customers' needs and wants. For many organisations, insight from curated data is used to promote decision-making and as such, the first step to profitability.

Chapter Four is structured as follows: introduction, Spree a subsidiary of Naspers Media24 as the case, the interviews, the presentation of the findings, categories, the development of emerged themes and summary.

4.2 The case

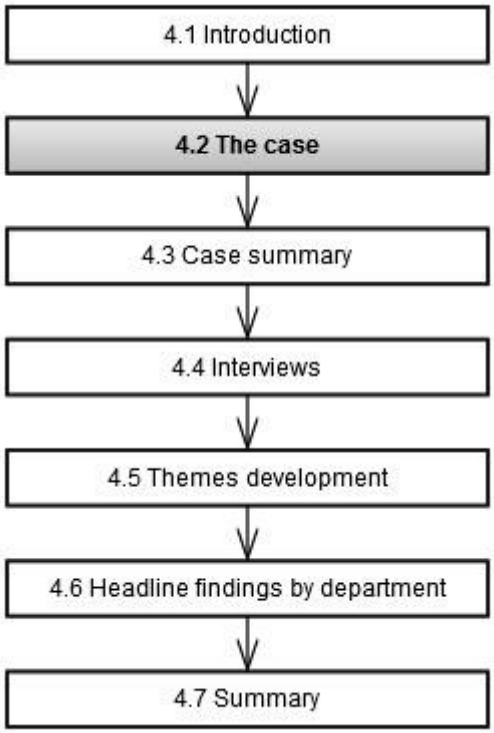


Figure 0.2: Chapter layout – The Case

Naspers Media24 is a global platform operator with interest and primary operations in Internet services, especially e-commerce, pay television and print media as depicted in Figure 4.3. Naspers Media24 provides services in over 130 countries worldwide with business operations in Africa, China, Brazil, Latin America, Europe, Russia, India and Asia. The strength of this multinational multimedia group of e-commerce and media brands organisation lies in being able to identify consumer needs, manage entrepreneurs, evaluate technology trends, engineer software, manage start-up companies and tailor solutions to meet consumer needs. The group aspires to being a strong operator in the e-commerce market space.

Touchlab is a subsidiary of Naspers Media24 responsible for the development of mobile applications and software systems in general. After many applications that did not generate revenue as anticipated, the organisation ventured into the Internet business (e-commerce) and one of the products that emerged successfully is the Sarie online shop called “Sarie Winkel”. Sarie is a well-known monthly magazine for women in South Africa. The Sarie online shop functioned for nearly two years as a proof of concept and as an incubation project. After proving successful and assuring viability, stakeholders then re-engineered the Sarie online shop into a fully-fledged online store called Spree.

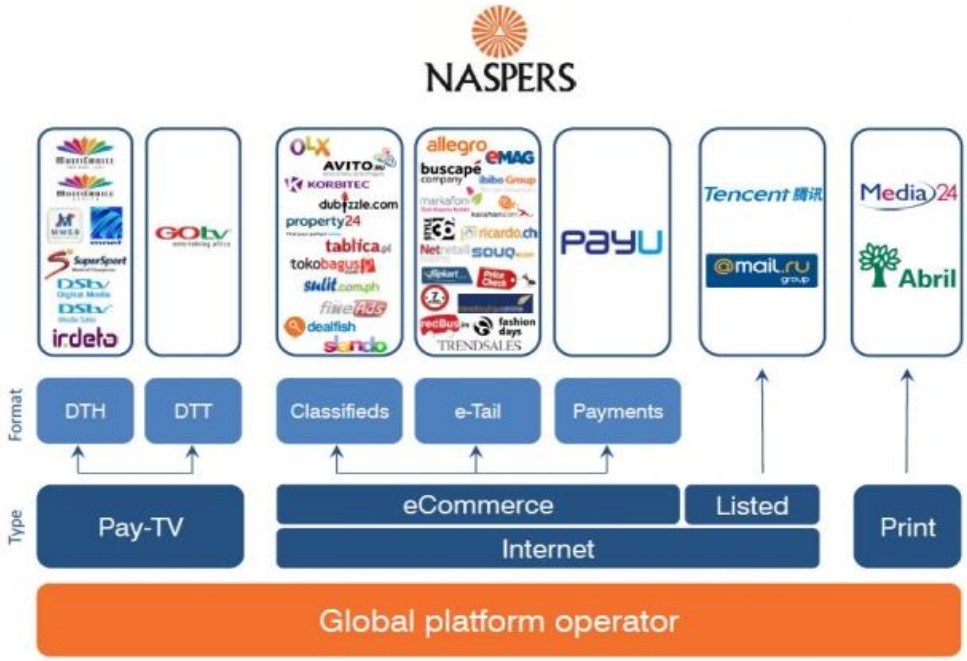


Figure 0.3: Naspers company structure

(Source: Naspers Media24, 2013)

Spree as a multimedia company has a primary business focus on Internet and e-commerce. It responsible for the e-commerce store www.spree.co.za. Spree has many departments, including an in-house technology team responsible for the development and maintenance of the e-commerce store (online software system). The Spree technology team comprises of 20 technical personnel with varied skills and experience levels ranging from front-end design executives responsible for the user interface (UI) to server side-developers with expertise in PHP and databases. The UI team focuses on the design and usability of the system to create a seamless and intuitive user experience (USX). Prior to the launch of Spree, most of the data serving as operational data which was validated for business monitoring came from the "Sarie Winkel" online store. The data is user browsing or transaction-related data, transactional and day-to-day operational data.

Members of the technical team are purposively chosen due to the complexity of the Magento platform ensuring that members have the required technical expertise. The look and feel of the system conforms to modern design standards with design essentials as stated below by a member of the design team during interviewing (Annexure A, Table 7.9):

> We aim to give the customer a real world experience with intuitiveness when they come to the site… The system status must be clearly visible so customers don't abandon shopping carts. Some additional heuristics also include user control and freedom, flexibility and efficiency of use with controls responding and giving indication of status, and very importantly aesthetics and minimalistic design; these elements are extremely important to obtain a good user experience.

The aim of the UI team is to promote seamlessness and intuitiveness in usability while maintaining a minimalistic design, yet to keep it secure and modernised. The quality assurance (QA) team have a responsibility to test the outcome of any task which affects the system in terms of upgrades and bug fixes to minimise any potential negative impact while upholding desired set standards.

The business intelligence (BI) department functions as a separate entity but they work closely with the technical developers as their primary responsibility is to ascertain in lay terms and translate into technical terms what each department needs to operate, especially for making decisions. The team mediates between Business Management and technical developers as data curators. The team's responsibility is to translate business requirements into a format that the development team can understand to enable them to capture such data and to provide the business intelligence (BI) team with this data. This curated data is made available in the form of processed reports and raw data.

Some of the departments that depend on the data include Supply Chain and Customer Services; Social Media and Marketing; Merchandising; Business Management; technical team of developers; and Business Intelligence and Customer Support as part of customer services. From a BI perspective, the aim is to present data to business in a correlated way to align with critical success factors. The data presents metrics, KPIs and dimensions in the form of reports, dashboards and raw data, all in support of business processes.

The following section highlights the departments at Spree.

4.2.1 Departments

Table 4.1 summarises departments at Spree with brief functional description regarding how each department functions as part of the business model. The unit of analysis is at department level.

Table 0.1: Functional departments at Spree

Department	Description/functions
Supply Chain, Customer Service and Support	Warehousing; deliveries; customer support; management of returns; this is a business and customer-facing department.
Business Management	General management; identification of business cases; conceptualising and identifying what data to log prior to building the system; this department is business-facing.
Marketing and Social Media	Marketing and advertising; this department is especially customer-facing, carrying across to customers products made available by business for consumption.
Business Intelligence	Building data structures for data curation; reporting and data provisioning; this department is business-facing.

Department	Description/functions
Technology or Technical	Development of Spree system; this is a business and customer-facing department.
Merchandising	This department is business-facing but has a responsibility that faces the customer directly; insight into its duties determines the organisation's success or demise.

Table 4.1 was compiled progressively through interviews with department heads and Business Management.

4.2.2.1 Business Management

The Business Management department has a primary responsibility to steer the business into profitable situations from a strategic and tactical decision-making perspective. Strategic decisions are the long-term decisions needed to delineate where business is heading while tactical decisions are the day-to-day business decisions required to keep business running successfully. Business Management provides figures or metric information to stakeholders, direction to operate, and tactical decision makers with most of its information from the BI department. One of its responsibilities is to identify the business case in a business scenario (business idea) and what data is required and curatable for business departments to make the idea a success. The business depends on BI for policies, strategies, guidelines and data governance which all form part of data management within the organisation.

4.2.2.2 Marketing and Social Media

Marketing is predominantly customer-facing, but also has a business-facing component, as its primary responsibility is to shape customer expectations through content building, advertisements, promotions and loyalty programs, and to also shape the customer's expectation through media such as online, television, newspaper advertisements, magazine and radio. Through these diverse communication methods, the customer forms an impression of the products and services Spree offers. Aligning targeted communication to a well-segmented customer group with a well thought-out service structure may lead the organisation into building loyal customers with an affinity towards the brand.

According to P4 (Annexure A, Table 7.9), the organisation has up to 300 customer contacts whereas Amazon has one million plus customer contacts per week. P4 (Annexure A, Table 7.9) exerts that:

> …like I said, this is not an old business; this is the difference between Spree and Amazon, we have up to about 300 contacts in a week. By contacts I mean email, phone calls, etc. At Amazon we are talking millions of contacts in a week. Amazon has customer service associates all over the world, thousands and thousands of them. So, the metric used there literally is defects per million opportunities.

> The goal was a Six Sigma goal. The Six Sigma states that the quality you are trying to achieve is less than or equal to 3.4 defects per million opportunities—that's the goal. So for every million orders, I mean transactions with customer, you are allowed 3.4 defects.

Spree defines a defect as any item sent to a customer that may have a valid shortcoming causing dissatisfaction. Customer contacts imply contacts to Spree through email, phone calls, complaints and returns, among others. A percentage of customer contacts pertains to returns, the lower the returns, the better, as the organisation constantly strives towards a lower defect per million measured. Currently, this is a metric that seems inapplicable in the domain of Spree due to it being a start-up; however, it is still worth noting.

4.2.2.3 Supply Chain and Customer Service

The Supply chain and Customer Services department at Spree is responsible for logistics management, which is comprised of planning, execution, design, control and monitoring of supply chain activities with the objective of creating net value, building a competitive and synchronising supply with demand, and measuring performance. This department views data as inbound or outbound with the generation of stock keeping units (SKU) as the linking medium. SKUs are generated when products have been ordered from suppliers and entered into the system. The generated SKU number which is a 20 character identifier allows the product to be entered into the Magento system and this is how the product is tracked throughout the system until it is sold. The product transitions from being a product to becoming an 'actual'; an actual is a product with price tag assigned. The product is then grouped as *configurable* and *simple* based on whether it is a virtual or sellable product. A SKU is either simple or configurable. A simple SKU uniquely identifies a sellable item on an Internet web product page, such as a clothing item. The unique number has attributes representing the manufacturer, colour, size and price category. A simple SKU is lower in hierarchy to a configurable item which is a virtual item assigned for grouping of items. The configurable SKU is virtual and allows for grouping and categorising simple items.

4.2.2.4 Technology

The Technology department is responsible for translating business concepts into a software system by building, monitoring and maintaining the software system as directives are provided from Business Management. The team collects and provides data to the BI department as a business requirement. It collects both transactional and transaction-related data and ports the data to Magento and Google Analytics. Most of the curation done by the BI department is on a Microsoft platform optimised for read, implying de-normalised tables to enhance reporting.

4.2.2.5 Merchandising

The Merchandising department attends to items that are sold on the site by Spree. They determine which items sell the best and which items not in order to manage stock based on trends and buying patterns. Most importantly, merchandisers identify sellable items. The manufacturers or wholesalers establish a working relationship with the seller by negotiating a unit stock item price and delivery to Spree. What is most important to the department is attaining a balance between selling and buying and being able to align customer buying to merchandise buying from manufacturers. The closer the alignment, the better the revenue generated. This balance comes from gaining insight into sales data and customer demand patterns for products.

4.2.2.6 Business Intelligence

The BI department is business-facing. It is responsible for provisioning data in the form of reports and raw data for other departments. The data required by the different departments are communicated to BI for curation. Most of the data required by these departments are identified and being logged through Magento on the back-end. Where a particular dataset is not part of curation, the data is modelled, collected and provisioned for the department. BI supports all other departments by meeting their data needs in the form of aggregated data, fine grain data for decision-making and raw data for manipulation. The varied departments use the data to support decision-making and day-to-day operations of departmental functions.

To provide data to the various departments, BI first provides data to the management team and thereafter to department heads and other data users by mediating between them and the technical team of developers. Table 4.2 summarises the data required by various departments for operational decisions.

Table 0.2: Different data required at department levels

Department	Data	Practical use
Business Intelligence	KPI; metrics; sales	Forecasting
Marketing and Social Media	Traffic, conversion; spend; campaign data; products data; results of AB testing	Customer segmentation or profiling; message packaging
Supply Chain and Customer Care	Customer data, buying pattern, delivery information, customer returns	Forecasting; improved service delivery
Technical	Data for the improvement of system performance	Recommendation engine to improve system personalisation; system improvement
Business Management	CSF; sales; GMV; rate of conversion; aggregated data	Profitability and sustainability
Merchandising	Customer buying; transactional data; transaction-related data	Smart buying; sustainability and brand awareness; relevance

Some of the many responsibilities of the department are to design and build the structures required for data curation. These include the data marts, data warehouses, tables to house the data, *ad hoc,* standardised and subscription reports. Some department heads capable of manipulating data prefer raw data to be made available. Up until the writing of this paper, Spree was at the business monitoring stage of the BD Business Maturity Index (see Section 2.13.1).

As a functional mediator, the BI department initiates the BI roadmap as a project to help centralise data received from the many different sources. The BI department exists to provide official support to Business Management. Data at Spree is decentralised and fragmented (data silos).

Some of the sources of data include:

- Magento
    - Site data and metadata on a Magento platform
    - Sales
    - Transactional and transaction-related data (browsing data)
    - Inventory data
- Google Analytics
- Bookmaster (On The Dot – a subsidiary of Media24)
- Affiliate data from magazines
    - Magazine print data
    - Magazine subscription and demographics data
    - Magazine likes, dislikes, marketing and advertising data
    - Magazine sales data
    - Magazine analytics data
- Call centre data comprised of customer sentiments and returns
- TV and banner advertising data with marketing

These data sources, besides being decentralised, are also not readily available and accessible by data consumers for analysis and decision-making on a tactical and operational level. From a strategic and tactical perspective, and also from an operational point of view, the data can be requested and provided within a period notification of 24 hours in aggregated format with the exception of magazine data. Participant (P) 2 (Annexure A, Table 7.10) stated that:

> …we are pretty much a start-up still and we are testing out new ideas, and we immediately get to see how they are doing. That is why we are almost very much looking for real-time information. The best we can get now is one day's delay; we will like it more real-time to make decisions right away.

There are only 19 magazines in the entire Naspers setup in South Africa. Of the 19, nine are affiliates with large volumes of data by virtue of longevity of operation. The data is not accessible to Spree as magazines operate as separate entities, although they are closely affiliated.

Decision-makers need data to make smart decisions but the data is spread across three silos which are sometimes out of synchronisation. For example, sales data from Magento does not always match data from Google Analytics (GA). This jeopardises the accuracy and integrity of the reports. The BI department believes initiating data centralisation to collate all the data in a central enterprise data warehouse for retrieval and easy access will help consolidate the data and provide better data governance. Business thinks of data centralisation as the first step to a long-term data journey of curation for analytics, as it is anticipated that the answers will be found in the data.

Departments depend on scheduled reports, but the reports are normally late. Data on the Magento platform is ported into a Microsoft 2012 SQL database. Scheduled reports are sent to recipients in the form of daily sales, weekly sales and monthly-collated sales with metrics such as sign ups, site subscriptions, sales figures and inventory, among others. Processed data are sometimes presented by means of dashboards which provide an overview of performance over time. From reporting services, the data is sent to recipients daily to reflect performance metrics. The data centralisation starts with the aim to collate all data onto one platform, being an enterprise relational environment. This is done to include analysis, data auditing and automation. The three silos are Magento, Google Analytics store and Bookmaster (a publishing-specific Integrated Enterprise Management business software solution). Bookmaster is used by On the Dot (OTD) to track deliveries and dispatches.

The Magento framework is a database persistence layer that uses the entity-attribute-value (EAV) structure to store data. Magento and partner companies offer a number of extensions that specialise in analysing BD (macro and micro data). These extensions are tabled below in Table 4.3 with stated areas of specialisation. Spree currently store data using the Magento framework, making it easy for the organisation to integrate the extensions for insight generation.

Table 0.3: Magento Big Data extensions

Magento extension	Focus
Lexity	Analytics
RJ Metrics	Analytics
Terapeak	Analytics
AddShoppers	Social analytics
Windsor Circle	Segmentation and targeting

Magento extension	Focus
Bronto	Segmentation and targeting
Springbot	Segmentation and targeting
Optimise	A/B testing
Fanplayr	Offers

4.3 Case summary

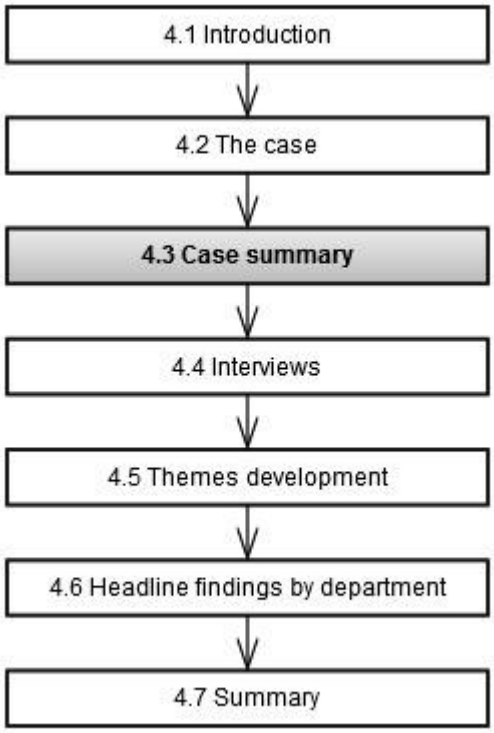


Figure 0.4: Chapter layout – Summary of business overview

Naspers Media24 is the research entity while Spree is the case. As an incubation project, “Sarie Winkel” has matured through the phases to become Spree, the e-commerce store. There are many departments at Spree, but for scoping and relevance the research focused on data-centric departments which include Business Management, Marketing and Social Media, Merchandising, Supply Chain and Customer Service, and Business Intelligence. These departments need data to make informed decisions. The data is made available by BI. BI uses traditional data curation toolsets to curate and provision data to the various departments.

4.4 Interviews

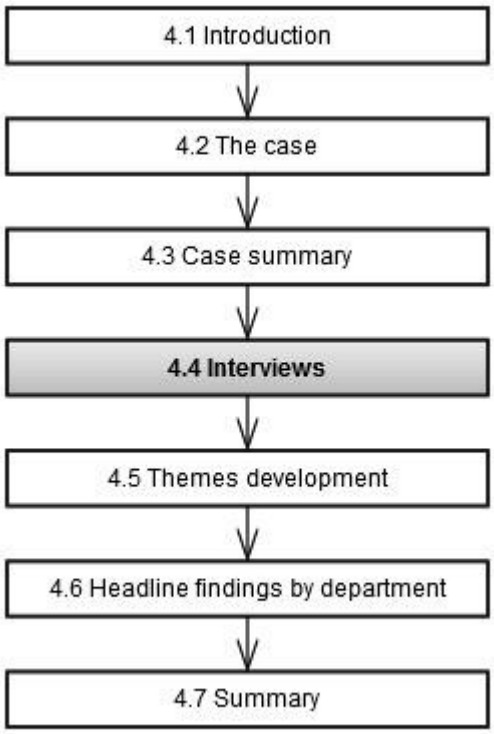


Figure 0.5: Chapter layout - Interviews

For the reader's convenience the problem statement, research questions and aim are restated.

Problem statement

Companies find it difficult to leverage the opportunities Big Data offers them in terms of monetising the content of curated data.

Research Question

RQ1 What are the factors affecting business to leverage Big Data for competitive advantage?

RQ2 How can Big Data be leveraged in a media organisation to gain a competitive advantage?

Aim

This study aims to explore the opportunities the research entity may obtain curating Big Data for competitive advantage. The study further aims to propose Big Data curation guidelines and recommendations for the curation of Big Data.

4.4.1 Pre-interviews

Prior to interviewing participants, an observation that emerged was the differences in data usage skills levels and affiliation to data within the organisation as a whole and in departments. This prompted a redesign of the interview framework to include a pre-interview phase that assessed:

i) How participants used data.
ii) Their involvement with data and how data affected their department.
iii) Technical know-how, relation and time range interaction with data.

This assisted the researcher in forming an impression of participants and their departments, and based on the impression the actual interview followed with marked relevance and varying difficulty levels as many participants mentioned that "these BD" questions were difficult. The interview process then divided into business and technical interviews (Figure 4.6).

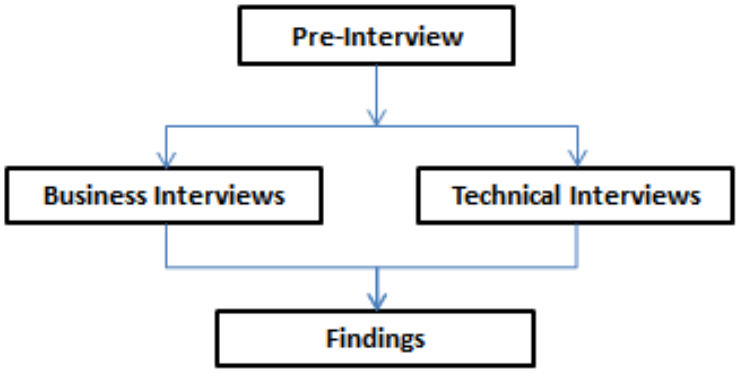


Figure 0.6: Interview diagram

Eighteen (18) participants were interviewed. All participants acknowledged data as being 'super' useful as mentioned by one participant and that data has changed from being another one of business's by-products to becoming an enterprise asset. Every interview participant was asked the pre-interview question to negotiate the succeeding interview questions (Table 4.4). All participants affirmed the importance of data many times.

According to P17, "data is everything; I cannot make decisions without data".

P16 (Annexure A, Table 7.9) stated that:

> …data is everything, and that we have come to understand that you make better decisions consulting data. Data brings information, data helps realise what is working and what is not.

P3 mentioned that:

> …the nice thing about all these is that we are getting better at making decisions with data; being a start-up we need data to even tell us what is going on, this makes data such an important part of the puzzle.

The importance of data is echoed by P4 who asserted that "…data is everything; data is the power house of technology".

Table 4.4 holds the pre-interview question, research questions and sub-questions.

Table 0.4: Research questions

Question	Question statement
Pre-interview question	What is your strategic, tactical and operational involvement with data?
Research question 1	What are the factors affecting business to leverage Big Data for competitive advantage?
Sub-question 1.1	What is business doing to leverage Big Data to gain a competitive advantage?
Sub-question 1.2	What is the business's view of Big Data in terms of competitive advantage?
Sub-question 1.3	What are the policies and strategies for leveraging Big Data?
Sub-question 1.4	What information do businesses want to get from Big Data?
Sub-question 1.5	What kind of data is being curated as part of Big Data?
Research question 2	How can Big Data be leveraged in a media organisation to gain competitive advantage?
Sub-question 2.1	How can Big Data be utilised to gain a competitive advantage?
Sub-question 2.2	How can a business implement Big Data curation?

The pre-interview question assists in substantiating the functions of the various departments, how the departments use data, and also allow participants to be grouped for business or technical questions.

Table 7.1 (Annexure A) summarises the pre-interview findings. This identifies how participants are grouped from an organisational point of view. Based on initial findings, participants are grouped to answer business-oriented or technical interview questions as per the Table 7.1 pre-interview findings.

P10 stated that his department is responsible for designing and building the data structures needed for centralising data as most of the data are in silos, fragmented across departments and other affiliated organisations such as magazines On the Dot and Google Analytics. P10 also mentioned that the organisation is a start-up with large volumes of inaccessible magazine data.

Magazine data is important because the data is huge. The data has been collected over years with distinct attributes enriching the data, which is the dark business data required for BD analytics. But Spree is restrained by the lack of historical data. P10 stated that, "…at the moment we have no historical data in the business".

This view is also shared by P1, P2 and P6 who described the Spree setting as being a start-up, inferring to the lack of significantly colossal data for analytics in support of decision-making. They did not attribute the lack of historical data to the organisation being a start-up. However, P7, P10 and P14 indicated that the lack of historical data for decision-making is a limiting factor, and that this can be attributed to data unavailability for generating insight to affect decision-making and other important business growing opportunities. This, P6 stated as:

> Ideally, there has to be a massive database for drawing insight, so much that your message can be more targeted and segmented. When saying one thing to everybody, you must be vague so not to isolate anybody, but with the right segmented data, you can customise the message to impact better. This is number two on my list to effect (Annexure A, Table 7.13).

P1 and P11 expressed the importance of data through financial planning in the form of budgets for procurements and providing buyers with purchasing figures for their various categories which are done on a Rand value scale. Due to the lack of self-service and availability of raw data provisioned, P1 depends on BI to provide data for the creation of clearance percentages. P1 uses the information to categorise sellers as to being poor, medium or best sellers. Poor sellers are added to a discount tab. The aim is to direct buyers into profitable situations using data. P1 P2 and P6 named data as being a mandatory piece of the puzzle to deliver service. The challenge here is the lack of historical data for better analysis, as mentioned in the preceding paragraph.

According to P4,

> …coming from Amazon, normally we'll look at historical data to predict … what availability and future lead times are going to be, we will predict stocks arrival time, and lead times are going to be (Annexure A, Table 7.9).

P6, as head of Social Media, asserted that influencing communication strategies and building social media content to attract and communicate with customers in the form of marketing, are some of the important goals of her department. She stated that:

> I influence the communication strategies especially regarding social media, at the moment Spree is starting; there isn't any data to work with. It is like, we have the Sarie database but you don't want to base any learning on that because that's a different consumer altogether. But any data we get from there I will just influence the social media or communication (Annexure A, Table 7.13).

Also, messaging and packaging of data to a well-segmented and connected customer group in an ideal situation ranks second on her (P6's) list of most important things to do; hence customer information and system content are paramount.

P6 mentioned that in segmenting customers, it is essential to look at elements and attributes such as:

> …the type of products they purchased, how often they purchased, when they purchased, are they frequent buyers, what is the size of their basket and are they the end user or buying for someone? (Annexure A, Table 7.13).

According to P6, different channels of communication have been identified over time, some of which are the use of social media through Facebook, blogs, Instagram, twitter, email marketing, Pinterest, radio and TV, but the organisation does not have much relevant historical data to ascertain customer behavioural patterns or insight to influence communication as anticipated by business. There is still no profiling or customer segmentation. Messages meant for customers are packaged in a certain way with the right tone based on their affiliation with the organisation to have the right impact. Customers may be new; they may never have bought before or even be existing customers, but what matters most is controlling communication for the brand's sake especially to promote a positive brand sentiment.

According to P18, data is constantly at work in the supply chain. This participant divided customer service into inbound and outbound data, and stated that none of the data can yet be considered as BD because the entirety of the data can be captured in an Excel spreadsheet. P18 further stated that:

> …you have all the SKUs; just to be clear, none of these data can be classified as BD at this point, it is very small amounts of data because you are talking about SKUs in thousands, and it can be contained within an Excel spreadsheet so it's really not BD (Annexure A, Table 7.12).

This does not waive the complexity between data fields. An important challenge in this section (Supply Chain and Customer Care) is dealing with variations—being a config to simple. A config (configurable) allows a seller to manage multiple options of a product. A config is a virtual item that holds many lower items, hierarchically arranged which are real products that appear on the web site. In this section, information from suppliers also play a major role allowing the department to determine costs and related variables, capacity planning, future lead time predictions, delivery date estimation, identification of patterns of demand and labour size planning. Outbound data helps with things such as space planning in the warehouse.

Collating the pre-interview findings from Table 7.1, many keywords emerged which were later reduced into categories and then abstracted to themes. The emerged pre-interview findings can be summarised in the following list:

Pre-IQ Findings

i) Data usage, application, complexity and deployment vary by department.
ii) All participants have experience with data-driven decision-making.
iii) All interview participants acknowledge data is an important enterprise asset.
iv) There are nineteen magazines in Naspers.
v) Nine are affiliated to Spree with more to join soon to form an affiliate business model.
vi) The nine magazine companies have massive data that is inaccessible to Spree.
vii) Data curation from a curator point of view is divided into technical and business curatorial data.
viii) Business Management conceptualises and contextualises business cases to identify the business side of proposed business idea.

The section below focuses on business interviews and technical interviews.

Business interviews entail semi-structured questions that ascertain business-oriented qualitative answers that focus on business processes and Business Management with data as a decision-making tool. Technical interviews, on the other hand, focus on the platform deployed, curation strategies, the technological, infrastructural aspects of the organisation relative to data curation, and the technical side of using data. Tables 4.5 and 4.6 hold information about interviews and related departments. Four (4) of the eighteen (18) interviews were repeat interviews for clarification on the preceding interview, this is indicated in column 4 of Tables 4.5 and 4.6.

Table 0.5: Business interviews

Interview number (INT)	Participant Number	Department	Number of interviews with participant(s)
1	1 , 11	Merchandising	2
2	2	Business Intelligence	1
3	3	Business Management	1
6	6	Social Media and Marketing	1
7	7	Marketing	1
12	8, 9	Marketing	1
13	12	Business Management	2
14	4, 13	Supply Chain and Customer Care	2
15	18	Supply Chain and Customer Care	1

Table 0.6: Technical interviews

Interview number	Participant Number	Department	Number of interviews with participant(s)
2	10	Business Intelligence	1
4	4	Supply Chain and Customer Care	1
5	14	Business Management	1
10	15,16	Technical	2
11	17, 6	Marketing	1

4.4.2 Business interviews

Participants acknowledged a web presence with 24-hour visibility in the form of an e-commerce store. P3 and P14 mentioned the following as preceding the creation of every software system within the organisation as part of creating the core business model and planning the architecture of the product including its look and feel. This helps the organisation to determine the following questions:

- What are we going to actually do?
- What is the business case?
- What is the business side of an idea?

P3 and P14 with their team identified what data to log from a department point of view as users will be in the systems domain. P14 stated that,

> …prior to developing a product, and this starts with planning the architecture of the product. What the product is going to look like. What I and my team do is to identify what data to record (capture) because users will be using this product and will generate data in the process.

This generates diverse data from transactional to behavioural data which includes logs and metadata. Based on the data, the organisation is able to ascertain:

- Are people are using the product?
- Is the organisation making the anticipated revenue?
- Is the organisation getting the right business data to measure metrics?

According to P12, it is important to collect data on where the system is failing to meet stakeholder expectations, for example results of A/B Testing.

There are many different systems involved in this e-commerce solution which form the entire architecture. As mentioned earlier, this includes Magento, Google Analytics, Bookmaster, the website and other affiliates like the magazine companies affiliated to Spree. Details about these data fragments are in Section 4.2.2.6.

4.4.2.1 RQ1: What are the factors affecting business to leverage Big Data for competitive advantage?

a) Sub-Q 1.1: What is business doing to leverage Big Data to gain a competitive advantage?

Interviews with 18 participants brought to light the importance of data to Spree with all participants mentioning that data is everything to Spree. However, not all participants concur that Spree has BD, curating BD or know what BD is, but all participants acknowledged data as being extremely important to Spree. Of the total number of participants, 22 per cent were unaware of BD, 78 per cent were aware but disagreed on Spree having or curating BD. In addition, 50 per cent of the total number of participants disagreed that Spree had BD while 28 per cent were strongly opinionated about Spree having BD.

P4 (Annexure A, Table 7.11) stated that:

> …data is everything, data is super useful, and data is technology power house. It allows us to run the business with insight…

P4 continued by mentioning how data is applied to every facet of business, especially where it concerns decision-making, planning, forecasting, fraud detection, space management and labour planning. She stated that:

> …data is applicable in almost every aspect of business, and besides, every decision should be data driven…

According to P1, budget formulation spurs the department on to have insight so we need to curate data. It is an absolute impossibility without careful evaluation of transactional data for insight into bestsellers, medium and bad sellers. This is necessary to steer the business towards a profitable situation. P1's (Annexure A, Table 7.12) words were:

> Data is everything, because the minute we don't have read on what is happening with sales and all that, it is virtually impossible to improve or push profitability.

P4 again mentioned that "data is very critical for analysis".

According to P7,

> …the essence of data is invaluable to business hence businesses' reliance on data, especially for decision-making.

Reinforcing the true meaning of data to businesses, all of the participants mentioned that data is needed for decision-making across every facet of business.

There are three kinds of decisions for businesses using data:

- Operational decisions
- Tactical decisions
- Strategic decisions

Finding 1: Data is an extremely important enterprise asset.

Finding 2: Data is critical for analysis and decision-making.

Finding 3: Data in Supply Chain and Customer Service falls under inbound or outbound data.

According to P3, a business manager, Spree is curating BD. He stated:

> I will say yes we are collecting BD; we have a good view on data, especially from a business model perspective as the business model shapes out technologies. We are curating BD to a degree but not to the level we should be doing. We are restricted by what tools are provided us. Like Google Analytics, it will store every page the user has been to and we will be able to go in there and say what is the most popular page, how many pages has the user been to (Annexure A, Table 7.11).

P12 indicate that most software systems are built in-house by the technology theme in collaboration with other departments; this gives the organisation the opportunity to conceptualise and identify what data needs to be logged from a department's point of view as to benefit others. P12 again said that:

> …customer data is incredibly important to us. We find marketing is really expensive and once we do that, getting a way to contact customers afterwards is incredibly important. So customer information, email addresses, phone numbers are really important. This we are good at storing though not at the scale we had like (Annexure A, Table 7.11).

However, P4, a former employee of Amazon, who led the global forecasting division of Amazon at Seattle and who is currently head of Customer Service and Supply Chain at Spree stated that Spree does not have BD yet. In her words:

> I will not classify this as BD; no, BD to me by definition, you want to talk about terabytes of data but because Spree is collecting so much clickstream information, clickstream information at some point we will look at and to me that falls close to BD (Annexure A, Table 7.9).

P4 also asserted that, "just to be clear, none of this data can be classified as BD yet; it is very small amounts of data".

P4 again highlighted that the part of the business affiliate model which might have BD is magazines, except this data is far out of reach for Spree. In terms of the proposed business model, magazines should be a direct affiliate, which they are, but access to data remains a secondary matter that has not yet been resolved.

On the other end of the scale, P6 mentioned that she believes Spree is curating BD, except none of it is available to her in terms of what she needs for work. She stated that, "I am sure they are collecting BD at the back-end but none of the data is available to me in terms of what I need, like customer profiling data" (Annexure A, Table 7.13).

P1 came across BD as a former employee of another retail store and drawing insight from data to make decisions is not uncommon. She mentioned that:

> I deal with data upon data…, but it all comes to having the right systems and at Spree now. I need to be super organised to run on these manual systems. According to senior management we will run on that for the next couple of months (Annexure A, Table 7.8).

Finding 4: There are conflicting and opposing opinions about Spree having BD.

Finding 5: Operational data users are uncertain about Spree curating BD.

Finding 6: There is valuable (BD) data that is outside the reach of Spree.

Finding 7: There is a gap in communication between data curators and data users.

Finding 8: Many data operations are still on a manual level, increasing the propensity for errors.

Finding 9: There is no historical data for decision-making as Spree is a start-up.

According to P3, it is essential to log every possible piece of micro data pertaining to customers, customer browsing log data for behaviour analysis, products, transaction and transaction-related data. The reason being, marketing is very expensive and with customers coming to browse the site, it is essential to gain insight about customers and be able to contact them. Furthermore, with the implementation of the Magento system and Google Analytics (GA), customer clickstream data (weblog) are stored by default—first on the web servers and second in Magento and GA. This facilitates varied analytics on customers and system (application) usage. Insight on customer data and customer behaviour on the site is at their disposal to gather further insight.

One of the consequences of customer profiling and segmentation is the effectiveness of marketing that it furnishes to the curating firm so that messages are targeted appropriately to the right group of customers to generate better results. This validates logging customer data and gaining insight into customer behaviours.

According to P12, there is a vast amount of data being stored about customers as this helps the organisation confirm data patterns and realise what works and what does not work. P12 stated that:

> …we are a new company; we need information regularly to ascertain what is working and what is not (Annexure A, Table 7.10).

Finding 10: Business Management identifies what data to log per department's need for information.

Finding 11: Spree logs varied data about business entities using the Magento system, Google Analytics and On the Dot (OTD).

b) Sub-Q 1.2: What is the business's view of Big Data in terms of competitive advantage?

P11 mentioned that there are two sides of buying: customer buying and business purchasing, such as Spree buying stock. Deciding how much of what particular item to order depends on insight into customer buying. P11 said that when customer buying and business purchasing correspond, they become more profitable and competitive. In her words:

> The aim is to get the buying and purchasing parallel as much as possible…, so if you see a big sale on the market then we have not paralleled it enough and we are taking a big knock on margin, a bad season (Annexure A, Table 7.12).

Knowing how much of what particular item to stock up requires insight into customer buying patterns. P1, from Merchandising, indicated that the data collected (which are in the form of transactional data), carry footprints that allow for improved decision-making. She adds that the business cannot go a step further without the data:

> The data I get relates 100 per cent to the entire needs of the business, meaning I am unable to operate without the data; without it, the business will not survive (Annexure A, Table 7.12).

For Merchandising, according to both P11 and P1, budget formulation depends entirely on curated data and market place happenings even though the data comes in the form of high level reports.

Finding 12: Business is profitable and competitive when customer buying patterns and business buying are aligned.

Finding 13: Transaction and transaction-related data are required for analysis into buying.

Finding 14: Insightful business operation demands analysing enriched data.

Finding 15: Merchandising formulates budget; budget formulation is based on market place events and data insight.

Finding 16: Merchandising uses current available information from reports to plan and make decisions.

Finding 17: Business Management evaluates system patronage and the success of a project generates business data and curates data for decisions from logged data.

c) Sub-Q 1.3: What sort of data is being collected as part of data curation?

P12, from a curation point at the inception of building a system, identified what data to collect because users will be using the system and this generates large amounts of data. Collecting such data helps the organisation in many ways, including evaluating patronage and relevance of the system to users. According to P12, it is customer data that is the most important as business needs to find a way to contact customers. He states:

> Customer data remains incredibly important to the whole equation. But, whilst we don't actively curate a lot of our customer data, when we need to make a certain business decision we are fairly good at going in there for what is necessary and curating that data to help us make that business logic (Annexure A, Table 7.11).

Some of the data collected, as stated by participants P1 and other respondents, include transactional data and transaction-related data. P3 mentioned that many varied data are logged besides customer-oriented data, including products such as data in the form of simples and configurables, product categories, payment methods, pricing groups, pricing models, promotion graphics and check-out graphics.

Finding 18: Many varied datasets (especially data needed for operation) are being collected as part of data curation, including transactional and transaction-related data.

Finding 19: In the supply chain there are variations (config and simples) which involve many SKUs with complex interaction fields.

Finding 20: Supply Chain and Customer Care uses historical data to predict product availability, future lead times and stock arrival.

Finding 21: Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data.

d) Sub-Q 1.4: What are the policies and strategies for leveraging Big Data?

Participants that answered this question were unaware of any documented policies and strategies to direct data collection and storage other than to direct anybody needing to know about policies and strategies to senior management.

According to P3, there is a lack of understanding about privacy laws and how these impact their usage of data.

P3 replied by saying that:

> …it is fairly difficult to restrict what can be collected and what cannot. Also regarding the issue of policies, there are no stated data curation policies (Annexure A, Table 7.11).

According to P2, from the BI department, strategies and policies are very important as data helps assess the current state of the organisation. This necessitates data management, governance and auditing of data for a single global view. But, what remains a challenge is the lack of documented strategies and policies surrounding data. She (P3) emphasised this point by initially stating that there are no policies and strategies of which she is aware.

Finding 22: There are no documented policies and strategies for data curation.

e) Sub-Q 1.5: What information does business want to get from data?

According to P2, message packaging for the social media environment and marketing is extremely important. The message has essential attributes such as tone, content, relevance, target audience and these attributes need to be built into the message to impact the target audience for the right response. Lack of this will mean vague messaging which may end up alienating customers. When a message lacks relevance for a particular group of customers, they consider these to be spam and cause a negative impact of reducing brand affinity and lowering brand sentiment.

For instance, P6 stated that:

> …on mother's day we had to send messages out but who are we talking to? Loss will only come from not being able to target a message at the right time, to the right person at the right time, so knowing the consumer through a life-cycle approach is extremely valuable.

P2 mentioned that it pays for the business to be able to separate big buyers from small buyers, and also to identify first time buyers for loyalty programs and maintenance. She believed that with BD, the organisation should be in a better position to administer loyalty for gains. In her words:

> Historical data for information is extremely important, especially data to profile customers for segmentation and to build content (Annexure A, Table 7.12).

On a granular level, some of the information business wants to obtain from data includes data that easily leads to insight, such as campaigns that directly improve sales and site patronage. The main reason for essence of data is to leverage its opportunities to become more profitable and competitive. From a departmental perspective, data is needed for operation. When the departments' processes become based on data, insight to optimise processes emerge leading to process improvements and maturation. The specific data and information needed at departmental level are listed in Table 4.3. From the technical team, the business is doing better every day and soon enough it will become mandatory to build into the system a yield management and recommendation engine. These require customer transactional and transaction-related data.

Finding 23: Business micro and macro data, though hind-sighted, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business.

Business makes many decisions, mainly operational, tactical and strategic. Strategic decisions are long-term decisions towards sustainability and profitability. Tactical decisions are seasonal with a much short timeline. Operational decisions turn to be very short-term, the kind of decisions made by floor managers, supervisors and people in authority on operational level. This forms the basis of data collection for analysis to guide decisions that improve business. P12 from Business Management mentioned that to market effectively, marketing must be directed to the right segment of customers at the right time. This again increases the need to be able to contact customers based on an analysis of customer demographics. In his words:

> It will not make sense for a winner of a promotion to come all the way from Namibia for a R1000 voucher; this can only be avoided by having insight into customer demographics (Annexure A, Table 7.11).

According to the Merchandising department, categorising products into best, medium and poor sellers depends on buying patterns and enriched data from sales. This progresses into determining which products are restocked and the quantities of stocked items. It also forms the basis of budgeting.

Finding 24: Data driven decision-making warrants profitability and productivity.

Finding 25: Business Management creates a budget (top line figure), which is handed down to Merchandising; Merchandising then re-budgets creating budget sub-categories from the top line figures.

Finding 26: With the budget, Merchandising determines the number of items per variations (config and simple) to buy.

P2 from the BI department mentioned the fragmented state of data at Spree as being a major impediment to analysis and gaining insight into data.

P2 further stated that:

> We have data spread across Google, Magento, OTD, and also at the magazine data in support of our business model. At least if we could put the first 3 together, we should be on our way to analysing data and getting insights (Annexure A, Table 7.9).

The statement of P2 is supported by P4, indicating that the current state of data makes it extremely difficult to use for any form of analysis.

> In the context of Spree, if I am limited to that area where we have BD, to go forward is magazine information… but you want to create attributes, except the way data is spread across the different servers, it will be a hellish job to try to bring all the datasets together in a meaningful way (Annexure A, Table 7.9).

P1 stated that the integrity of data may be compromised or jeopardised due to manual data handling.

> Because the data sent to me is manually generated and I must also alter to suit my needs, this may end up jeopardising the data integrity causing a high propensity for errors.

Business is at a point where there is a predominant manual manipulation of data, making the data error prone. All participants cited the organisation as being a start-up and as a result it lacks the necessary data volume for insight generation, meaning historical data to generated trends, patterns and insight is lacking.

Many participants also mentioned that the data available to them can easily fit onto an Excel spreadsheet which accentuates the fact that is it still a start-up and has not matured in terms of data for reliable decision-making.

There is a lack of consensus as to whether Spree is curating BD or not. P4 started off the interview highly opinionated about Spree not having or curating BD, but was quick on two occasions to dismiss this claim, stating areas where Spree might have BD. P2 and P7 similarly agreed, when at the beginning of the interview they mentioned that Spree is not "there yet" with BD curation, but halfway through the interview, these same participants changed their views and stated that Spree has BD and is curating BD.

Finding 27: Data is spread across multiple servers outside the perimeters of the business, making the data inaccessible for analysis.

Finding 28: The way data is stored across the different servers may increase the difficulty in bringing the data together for analysis.

Finding 29: Manual data handling may make data prone to errors, thereby compromising the quality of data which may affect decision-making.

Finding 30: There is a lack of consensus as to whether Spree has BD or is curating BD.

Finding 31: The lack of historical data lessens the generation of insights and patterns.

Finding 32: Lack of customer segmentation may reduce the impact of messages as messages must be vague.

4.4.2.2 RQ 2: How can Big Data be leveraged in a media organisation to gain competitive advantage?

Leveraging data for competitive advantage implies gaining a competitive edge through data to improve and optimise business processes for sustainability and profitability. According to P11, the learning acquired from buying patterns and transactional data spurs smart buying for the department. In her statement she mentioned that:

> …attaining competitiveness is only the beginning; remaining competitive through continual optimisations and data leveraging becomes a necessity (Annexure A, Table 7.11).

The benefits of leveraging data are further accentuated by P6 through the impact of targeted messaging gained from the likes of customer segmentation, profiling and loyalty programs.

According to P6, these are well engaged and effective leveraging of BD.

> …leveraging BD may culminate… Improve in page analytics to get micro and macro data for insight in creating content. Curate to foster brand affinity.

P13 was quick to emphasise the importance of having a clear BD goal prior to even commencing curation as this is an industry norm. She mentioned that many organisations have jumped into BD curation blindly to their own demise. It is absolutely mandatory to have clear data goals which identify opportunities and possible setbacks that may arise.

Finding 33: Customer returns data is a source of rich data to optimise customer care and support.

Finding 34: Leveraging data insight to create content will improve brand affinity and customer awareness.

Finding 35: It is critical to have a clear BD goal.

a) Sub-Q 2.1: How can BD be utilised to gain a competitive advantage?

P6 advised that social media, as a platform to create brand awareness, is a highly reactive environment needing quick actionable data and turnaround time. For example, results of A/B testing could be leveraged when patterns are identified quickly. It is imperative to realise that this cannot be stored for future except when building a knowledge base of identifiable patterns to study trends.

According to P13, the entire back-end depends on market place happenings which are reflected in data in the form of patterns and trends. How to organise the product warehouse, manage labour, control product capacity—these are all insights that can be attained by looking at data patterns.

P2 stated that the cost of marketing is a deterrent but BD and its benefits facilitate effective marketing. A curator may be able to ascertain traffic sources of information, thereby directing marketing expenditure for optimum gains.

Finding 36: Social media is reactive; hence actionable data is needed for fast action.

Finding 37: Marketing is expensive but BD will help the organisation market effectively.

b) Sub-Q 2.2: How can a business implement BD curation?

Prioritising and incentivising customers and potential buyers are important to P11, who believes that putting the customer first is part of building an integrated customer life-cycle. For example, a customer who is called by his name gives them the impression that this company cares and knows me. When a good buyer is rewarded, it improves the customers' affinity to the company.

See Annexure A, Table 7.3, for a summary of the business interview findings.

4.4.2.3 Business interview summary

Business makes three kinds of data-driven decisions based on insight from data to optimise sales, processes and profitability. These decisions are operational, tactical and strategic decisions. As curated data is at the heart of decision-making, the organisation is bound to progress towards set goals, except interviewees are uncertain as to whether the data used in this decision-making is BD. Secondly, participants who are aware, either state that the data at Spree does not qualify to be called BD, primarily due to the data size and other factors which employees could not adequately articulate. Many different perceptions of BD definitions emerged, with some participants disqualifying currently available data as being BD; participants also mentioned that the data can fit onto an Excel spreadsheet. Some participants are aware of BD and its uses, and were strongly opinionated that Spree has BD and is curating BD. Many findings emerged which are summarised in Table 7.3.

Table 7.6 provides a summary of technical findings (Annexure A, Table 7.6).

4.4.3 Technical Interviews

Table 7.5 in Annexure A presents findings of interviews with technical personnel. Participants in this group are predominantly from the technical development and BI team. In the case where technical experts were employed in Supply Chain and Customer Care, Business Management, and Marketing departments, they were also interviewed.

P10 from the BI department said that "Spree is a start-up and not functioning yet within the realm of BD". P10 further mentioned that business is currently at a stage where most of their activities are centred around curatorial and data provisioning processes mainly for scheduled reporting, *ad hoc* reporting and data provisioning. These are accomplished through the use of industry tools such as SQL server 2012, SQL server integrations services (SSIS), SQL server reporting services (SSRS), and data mart and data warehouse.

According P10, most departmental reports are scheduled for delivery daily using SSRS. Most business operational data are one day late. According to P4 and P10, from operational teams and the operation decision maker's perspective, this is still within a reasonable time range for decision-making. P10, P15 and P16 from the BI department pointed out the need for real-time data but current systems can only provide data one day late.

4.4.3.1 RQ1: What are the factors affecting business to leverage Big Data for competitive advantage?

Participants 4, 10, 14, 15, 16 and 17 all mentioned that data is an enterprise asset needed for decision-making. According to P14,

> …at the inception of business, most decisions were made based on gut (Annexure A, Table 7.10 & Table 7.11).

P14 further stated that business has transitioned to acknowledging the core of data. Many steps have been taken to integrate data into the decision-making process, especially long-term decision-making as "we always need to be able to tell that people are using the system and that the organisation is generating the needed revenue from the system" (app). From a technical point of view, it is essential to know the performance of the system during peak hours.

The performance metrics are needed for hardware provisioning, resource allocation and software system improvement. P4 and P10 acknowledged that data are in silos, necessitating steps to bring these data together to be able to draw insight and answer questions such as what the best time is to initiate campaigns for marketing and what gets consumers preferring the offered social media articles, images and adverts? Business needs to know how conversions are improving, if traffic is increasing to the site. And, when traffic to the site increases, does it imply a higher conversion rate? According to P2,

> …analysable data are in silos all over the place, necessitating steps to centralise the data as part of building the data marts and data warehouse (Annexure A, Table 7.10).

P8 said that testing out functionality by analysing site metadata, AB testing and multivariate results helps the organisation to realise the usefulness of parts of the system. He stated that:

> …the business is still a start-up and we have no way of knowing what works and what does not work, so the way around that is to collect and analyse these data.

Participants mentioned data as being everything. Data assists to measure performance of systems, functionality and sustainability. According to P4, data is the technology power house. Data encapsulates the successes and failures of the business, hence a source of

insight for business monitoring. According to P10, data helps communicate findings about business. P4 mentioned that these are justifiable comments to warrant more data management. According to P10, the aim is to provide on demand data to business in the form of reports, aggregated macro format data as in dash boards and raw data.

Finding 38: Spree is a start-up and not functioning within the realm of BD yet.

Finding 39: The need for data for decision-making is mentioned through all departments and by all interview participants.

Finding 40: The business is a start-up so not much historical data exists.

Finding 41: There is a need for real-time data but current systems and implementation only provision data one day late.

Finding 42: Data is an enterprise asset that needs business monitoring and steering.

P4 mentioned that Spree may not be in the realm of BD, but Spree has Internet streaming and clickstream data which by definition is BD.

In her words:

> I will not classify this as BD. Big Data by definition, you want to talk about terabytes of data but because we are collecting so much clickstream information at some point we want to start looking at that; to me that falls closer to BD world and when it comes to BD, it starts as small data. The underlying fact is that I can handle this data within my Excel control…

According to P4, Spree does not have enough sizable data to classify as BD although Spree is collecting BD. Current data is not overwhelming yet. This decision is based on the fact that the available data size is small, data will fit onto Excel spreadsheets and as an organisation, "we are not overwhelmed with data yet, rather we have a lack of data".

According to P4, Spree is not curating BD. In her words:

> The word curating scares me because we are not curating at all. I will say we are managing data (Annexure A, Table 7.9).
>
> But yet again, because Spree is collecting clickstream data and BD is an extension of data, I should say there is BD. I might say we are not curating BD, but think of BD as an extension of data (Annexure A, Table 7.9).

According to a BI interview participant (P10), business acknowledges the importance of data but the business is still a start-up and that business is not operating in the domain of BD yet.

Rather, available data is fragmented so the focus of the department is to centralise data by building and architecting data marts and data warehouses to contain both granular data and aggregated data for business monitoring. However, half way through the interview P10 realised that the Magento platform implemented allowed for the collection of clickstream data. This realisation made the participant immediately change the stance from "we don't have BD and we are not curating BD" to "we have BD except it is not very big yet" (Annexure A, Table 7.10).

According to P17, there is BD. He stated that:

> Big Data is being collected in the form of Instagram feeds, Facebook campaigns, social media and structured data. I think BD is our structured data and unstructured data.

Finding 43: Big Data is an extension of data.

Finding 44: Interview participants who are aware of BD are operating with a limited definition of BD.

a) Sub-Q 1.1: What is business doing to leverage Big Data to gain a competitive edge?

Business is at a monitoring phase, aiming to centralise all segregated data to commence analysis in the near future, hence the deployment of Microsoft data curation tools to facilitate the process (P10). The process of data management is augmented by SQL server integration services and reporting services for data grouping and dissemination. Data collection, especially from Google to gain metric information, daily sales, inventory data, traffic, clickstream data, metadata, transaction-related data and transactional data is done using SSIS as a curation tool.

As part of curation, data marts are being created to service departments. These data marts are being created for departmental use, e.g. sales data mart and customer data mart; these and other data marts combine to form the data warehouse. These are online analytic process implementations (OLAP) optimised for data analysis so data structures, especially tables, are de-normalised for data retrieval.

P17 from the Marketing department indicated that centralising data to ask questions is a requirement as businesses will be able to gain insight in customers and products together and also ascertain what deters customers from purchasing. One other source of insight which P17 proposed as important for her department is customers that abandon shopping carts.

Finding 45: Business Intelligence (BI) is in the process of building and creating data structures to centralise data for analysis to gain insight into curated data.

Finding 46: Traditional marketing is expensive.

Finding 47: Formal data modelling processes aimed at identifying information need is done by Business Management who may not be in possession of the needed expertise or knowledge of other departments' core functionality.

Finding 48: Marketing is at a point of wanting to analyse customer, product and sales data for better insights for better marketing.

Finding 49: Departmental use and need for information mature and evolve at different stages and levels.

b) Sub-Q 1.2: What is the business's view of Big Data in terms of competitive advantage?

The Marketing department initiates campaigns anticipating an increase in traffic and conversions. P17 mentions that understanding the customer base and knowing when to initiate a campaign is strategic to obtaining results.

According to P6, social media is very reactive, need actionable data and the ability to respond quickly as social media cannot be planned ahead of time. Rather it is essential to be in a ready state to tap into the opportunities of the social platform. This differentiates market leaders from followers. She states that, "for me this is being competitive".

According to P4, data, and for that matter BD, is pivotal in service design and service delivery to customers. As an organisation and department, it will be a good idea to aim at Six Sigma goals which is cost reduction and total customer satisfaction. In the Customer Care section, this is reducing cost through insights by identifying where to cut back or reduce wastage, for example, reducing customer returns through a reduction in defects per million metric. Relative to Six Sigma, the aim is striving for quality to the point of near perfection. She mentioned that Spree might not be at the point of implementing Six Sigma yet in terms of customer contacts per week, but the bottom line is gaining insight into customer service and being able to improve this section to the point of tracing the cause of returns easily and destroying it so as not to repeat the mistake.

Customer contacts may imply a customer calling to return an item, a customer receiving an incorrect item, a customer receiving a damaged item, and/or a customer complaining because they are disgruntled with service.

P4 mentioned that labour size planning, capacity management and stock holding layout are important benefits of gaining insight into buying patterns.

Finding 50: Campaigns are initiated to create awareness which then propels more traffic to the site.

Finding 51: Customer base insight is necessary in achieving results from campaigns.

Finding 52: Curating data to improve service delivery is pivotal to achieving the Six Sigma goal.

Finding 53: Next generation products may come from managing customer returns with insight.

Finding 54: Supply Chain and Customer Care may trace errors and sources of errors or defects using historical data.

c) Sub-Q 1.3: What are the policies and strategies for leveraging Big Data?

Not much was said about this as there are no documented policies.

Finding 55: From a technical point of view there are no documented policies, strategies, frameworks and curation models, except a Magento ERD diagram.

d) Sub-Q 1.4: What information does business want to get from Big Data?

Business monitoring and steering is dependent on data; data provides information relative to metrics, statistics, key performance indicators, critical success factors and business initiatives. The aggregation to constitute these elements is obtained from low-level data that is combined and grouped, especially by BI and furnished to Business Management and other departments for varied reasons, which may include directing the business, for example, data that is collected to promote the loyalty program, which include details of first time buyers or buyers who refer others to buy and customer data for business to follow up on customers. Other forms of data also include inbound and outbound data to predict lead time and delivery time, conversion rate and to gain insight into campaign release dates and best sellers data to ascertain which products to buy more of. The organisation also collects data about best buyers to segment customers for customer profiling. On a granular level, some of the data collected include product information, customer information, sales data, and system data.

Finding 56: Business aims to centralise data as a means to better answer key business questions.

e) Sub-Q 1.5: What kind of data is being curated as part of Big Data?

P10 mentioned that granular transactional and transaction-oriented data are being collected which covers customers, products, sales and browsing information. This, P14 broke down into customer-oriented data, including customer name, session information, email address and time, among others. According to P4, clickstream data is being collected for future analysis on customer usage of the system.

Finding 57: Spree is collecting clickstream data and BD is an extension of data.

Merchandising aims to align customer buying to business buying products; the closer these two are aligned the more profitable the organisation will be, implying that all buying decisions by business need an analysis of customer buying patterns (P14). Furthermore, the same department uses product, customer and sales data to work out best sellers, best buyers and best sales days. Insight gathered from here helps the business plan ahead for smart merchandising.

The Marketing department launches campaigns, advertises Spree on web affiliate portals (marketing spend), and measures sales and the aftermath of all marketing efforts, including the results of AB testing. Data proving successful initiatives as in the case of reactivity of social media are actioned quickly. Insights from this data help departments manage future initiatives better. According to P14, marketing is very expensive so it is important to collect the right data so decision-makers are confident about decisions. Some decisions include budg*et* allocation for marketing expenditure and message packaging for a targeted audience.

Business Management evaluates which parts of the system are doing well through looking at site metadata.

Finding 58: Decision-making based on data is applicable to every facet of business.

Participants mentioned many different data challenges ranging from decentralisation to deter analysis, manual data handling which may increase the propensity for errors, poor infrastructure and lack of BD specific infrastructure, and the lack of historical data as some of the immediate data challenges. According to P14,

> There is so much manual data handling which increases the propensity for errors and compromise the quality of data.

Finding 59: Data quality may be compromised due to human intervention or manual data handling, reducing the accuracy of findings.

4.4.3.2 RQ 2: How can BD be utilised to gain a competitive advantage?

a) Sub-Q 2.1: How can a business implement Big Data curation?

According to the Technical department and BI, Spree is not there yet and so this question is slightly out of scope, but they believe with the current centralisation, curation processes are in place; Spree should get there sooner or later.

Finding 60: There are no plans of BD integration as of yet; Spree is not there yet.

b) Sub-Q 2.2: How will BD curation contribute to the growth of Spree?

Big Data curation will give spree a huge advantage over competitors; it is a good opportunity for analysis but this still remains a speculation as Spree is not there yet. It will give us the opportunity to analyse customer behaviour, improve the system and become more competitive.

Finding 61: Big Data curation gives the curator a competitive edge.

Annexure A, Table 7.7, presents interview questions and findings.

4.4.3.1 Technical interview summary

The section on technical interviews sums up the technical aspects of this research more from a department specific standpoint. It focuses on the technical curatorial processes of Spree, use of technology to accomplish diverse tasks as the organisation seeks insight from curated data to monitor and steer itself towards sustainability and profitability. Technologies touched on by participants include Magento, SSRS, SSIS, SS 2012 and other fragmented data sources that the Business Intelligence department seeks to centralise for analysis.

Findings from these interviews are summarised in Table 7.5 (Annexure A), detailing what is relevant with respect to findings in the varied departments. Not all questions were answered in this section as some participants felt some questions were not applicable in the present setting of the research entity.

4.5 Themes development

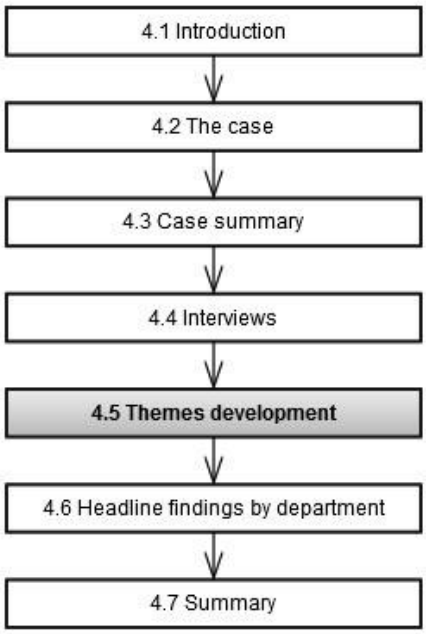


Figure 0.7: Chapter layout – Themes development

Theme identification is fundamental to qualitative research (Ryan & Bernard, 2003:85), yet simultaneously mysterious as the broad sharing of techniques to identify themes seem hindered by epistemological differences, boundaries and disciplinary prerogatives. Ryan and Bernard (2003) mention that this analysis involves four main tasks which include discovering themes, sifting through these themes to a manageable few, building hierarchies of themes, and linking themes into theoretical models. Steps 2 to 3 deploy the process of data reduction, a qualitative data process that reduces themes to a manageable, relevant and connected few. Tables 7.13, 7.14 and 7.15 provide an overview of theme development.

From eighteen interviews covering 6 different departments on leveraging BD for competitive advantage and data curation from a departmental perspective, twelve (12) themes emerged and will be discussed in the next section.

4.5.1 General themes discussion

i) Data as asset

All participants concur with the importance of data to business. Presently, a common trend in business seen in industry and paralleled at Spree is that data has evolved from being another by product of business into becoming an extremely important enterprise asset, with quantifiable metrics such as size, rate of capture, accuracy and diversity.

Many organisations desire to analyse these data especially real-time data for competitive advantage as the benefits posited are countless; some direct benefits may include data to support loyalty programs in marketing, direct marketing, recommendation engine data, dynamic pricing, yield management and improving customer engagement.

All interview participants referred to data as being essential for many diverse reasons with common departmental course. Among many factors, the leading factors spurring Spree to curate data as emerged from interviews and findings include customer profiling for segmentation and targeted marketing, smart buying for merchandising, cross-product sales insights for market basket analysis, decision-making (as these factors affect sales) and marketing. Participants from Business Management also accentuated the essence of integrating data into the business model as the basis for infrastructure architecture and design for value proposition to business customers hinges directly on the group's business model. Participants indicated BD to be important but its status at Spree still remains unclear as many participants are unaware of BD and its curation. Some participants are aware of BD but these participants operate with a limited definition of BD, and participants who are seemingly aware acknowledged BD curation practices as incomplete, inadequate and not on a massive scale to see its direct benefits.

ii) Customer

The Spree customer is an important entity with many variables collected about the customer. This is essentially the customer's footprint on the site and serves as one major data source for many benefits. The customer, in the case of Spree, is an individual or a party that receives a product from the organisation. Marketing and Business Management participants mention that contacting the Spree customer is an essential part of marketing. Placing the customer within a segment and categorising them in terms of buying power fosters management of loyalty programs and customer base management for future integration into the customer's life-cycle. Customer data provides insight into customer behaviour, transactions, returns, buying patterns and order of browsing—these are insightful customer-oriented elements to monitor. As a major source of insight for operation, Business Management participants mentioned the many advantages of collecting customer data from every possible data source. Some of the advantages include insight to improve user experience and better service delivery. Participants frequently mentioned the development of next generation products based on customers' likes and trends as the curated customer data contains the customers' behavioural patterns.

iii) Products

Products in the case of Spree are goods sold to a customer. Products vary from clothing, shoes to accessories for men, women and children. Some products recently introduced include home décor products, gift items and toys. Marketing aims to advertise Spree products to a targeted segment of customers in the right tone and at the right time to gain better results in the form of improved sales. Marketing mentions targeted marketing as an essential activity to gain the needed impact and possible reflection in sales. Merchandising buys products from suppliers who look at customer buying patterns, when customers buy and how business buying aligns with customer buying. This infers profitability and improved business opportunities**.**

iv) Planning

The Supply Chain and Customer Care department manages processing of sales and delivery of items. Managing the warehouse requires planning through insight into buying patterns, especially what items sell the quickest. Based on this insight, labour can be managed and allocated to specific times of data based on traffic (demand on items). Fast selling items are placed closer so to be reached quickly. This is the result of decision-making based on insight from data. From a marketing perspective, traffic expenditure, campaigns and advertising are planned based on insight and best results based on historical data and AB testing results.

v) Decision-making

Decision-making spans the entire business. There are three kinds of decisions made by business—strategic, operational and tactical. The quality of decisions depends on accessible data and the accuracy of the data. The Merchandising department aligns the two different kinds of buying (customer buying and Merchandising department buying to re-sell) to improve and predict profitability. The demise of an organisation becomes quickly eminent with bad decisions. Optimising business success requires making insightful decisions, which is driven by quality data.

vi) Competitive advantage

Competitive advantage infers superiority of an organisation over its competitors in an industry setting. In the case of Spree, being competitive may mean more sustained sales and increased profitability. According to Merchandising, accomplishing the Six Sigma goal based on insight into customer service and smart merchandising will improves brand affinity and help form a better customer impression of the organisation. Participants exerted that

competitive advantage can be likened to information advantage as insight will yield sustainability and profitability.

vii) Marketing

Marketing refers to all the activities the organisation may undertake to promote sales of products and awareness. At Spree, this includes advertising through links on affiliate sites, banners, online advertising, TV commercials and radio, and also through sister company sites. Business managers mentioned that marketing is vital and expensive.

viii) Service

Service is the sum of acts and elements that allow consumers to receive what they need. Customer service such as Spree is vitally important as a differentiator due to the competitiveness of the industry. According to interview participants from Customer Care, Spree as an organisation aims to reach the Six Sigma goal of 3.4 defects per million opportunities. This implies better quality of service and customer care.

ix) Sales

A sale refers to the act of selling products or services in return for money. Sales generate tangible revenue for business. According to Merchandising, when customer buying aligns closely to Spree purchasing on a continuum, that implies profitability. Merchandising uses sales data, customer demand patterns and best seller information to group and categorise products to make smart buying decisions.

x) Analytics

Transforming data into actionable insights is a benefit stemming from creating insight through data analysis. Analytics, for example customer behavioural analysis, provides business with insightful findings for steering and monitoring. Combining detail customer transactions such as sales, returns, consumer comments and weblogs with social media data is a game changer that furnishes the seller with actionable insight to obtain information advantage.

xi) Strategy

The successful implementation of strong and robust strategies gives any organisation a significant competitive edge as mentioned in Section 2.6. Well-formulated strategies lead to superior performance for organisations when successfully implemented. Strategy furnishes business with action plans to gain a desired outcome. Strategy from a business perspective permeates all facets of business but the implementation of a successful business strategy

demands leveraging insight from data. This may require identification of all KPIs and metrics against which success can be measured and the supporting data and processes that may be used.

xii) Business model

A business model describes the rationale of how an organisation creates, delivers and captures value, in economic, social, cultural or other contexts. It represents the plan implemented by a company to generate revenue and make a profit from operations; the Spree business model is described to be one of its kind as Spree as an entity has many affiliate organisations, all operating under one umbrella company—Naspers. The Spree magazine company affiliation allows redirection of Internet traffic for which the magazine company receives revenue when the directed traffic leads to a sale.

4.5.2 Themes

Themes interaction at department level

Interviewed participants came from the mentioned departments; participants had common uses for data and reflected on data and its uses within their departments in a similar way. Sections 4.5.2.1 to Section 4.5.2.6 below reviewed these themes and their interaction at departmental level.

4.5.2.1 Themes: Business Management

A competitive edge is a state of business brought about by the possession of a resource that may bring an organisation into a leading position in a given industry as a customer's needs are met and relevance of the business to customers steadily grows. This is a business attribute that will see especially businesses possessing the right attributes to outperform competitors. This state is promoted by data, insight from data, the business model and the strategy for accomplishing business critical success factors (CSF) and analysis. The use of Big Data has become a crucial way for leading companies to outperform their peers, as BD can unlock significant value by making information transparent so to improve the accuracy of predictions with increasing curated data, and segmenting customers and products based analysis at a granular level. This data may lead to developing next generation products; optimise business through insights and knowledge from data.

Both macro and micro data forming the basis of BD in an enterprise is an extremely important enterprise asset needed for every aspect of business, especially in decision-making, customer segmentation and profiling, forecasting, planning, budgeting, marketing,

reducing and intercepting fraudulent transactions and improving service delivery. Business transfiguration and survivability depends on leveraging and optimising deployment of insight from curated data.

Business makes decisions that fall under operational, tactical and strategic decisions. These decisions affect the growth, survivability and sustainability of the business. Risk leading to the demise of the business is reduced by decisions made based on the availability of enriched data.

4.5.2.2 Themes: Marketing and Social Media

Business Management is aware of curated data and more curatable data that could be of significance to other departments (especially Marketing and Social Media department), but this information is not documented and not shared, so a department like Marketing remains in the dark. The Marketing department it not able to use available data immediately, and as such, this results in a communication gap.

This communication gap infers an information gap between technical data curators and operational data curators for decision-making at Spree. Operational data curators are unaware of what data is being curated at the back-end, while technical curators are fully aware of what data Spree has. This potentially creates a gap which could be bridged through documentation, communication and sharing. This is further exacerbated by unawareness of BD and its opportunities to business for a competitive edge. While curators at inception know and log data needed for analysis, operational data users are unaware of what data is being logged, the granularity, the kind of data, the disparity and the scope. This unawareness creates a data gap as decision-makers cannot leverage curated data for decision-making due to unawareness. Curated data is the basis of business monitoring leading to business insights. Generated insights eventually lead the business into optimisation and further up the ladder of maturation. For instance, inefficient merchandising efforts and products that are not selling could be identified easily with curated data making the business more efficient or smart at Merchandising.

4.5.2.3 Themes: Supply Chain and Customer Service

The entire back-end and supply chain depend on data. Better insight into both macro and micro data fosters efficient and effective running of the back-end. This affects space management, service delivery, fault reduction, fraud detection, labour planning and decision-making. An organisation's ability to meet customers' needs and wants through the five mentioned pillars of service—tangibles, reliability, responsiveness, assurance and

empathy—improves customer service. Improving customer service through attainment of the Six Sigma goals indicates better quality service. In the case of quality of service, the customer still remains the sole judge, though the organisation creates and follows best practices.

4.5.2.4 Themes: Technology

The technology team focuses on development, maintenance and smooth running of the system; this is enhanced mostly by tests and metadata collected to measure performance. Business, at regular intervals, collects logged data to evaluate and monitor the success of sections of the system. This data is provided to business and stakeholders in the form of reports to provide feedback on performance or patronage. Business Management makes decisions as to whether to continue portions of the system or cut back. Improving system functionality and usability are two important aspects of monitor-logged browsing data and metadata. For instance, abandoned shopping carts may supply reasons related to how customers struggle when checking out.

4.5.2.5 Themes: Merchandising

Buying has two forms according to Spree—customer buying and product purchase according to the budget set out for the department. Confirmed demand patterns and trends facilitate aligning customer buying and products purchased; this has its basis in the VALUE OF PERFECT INFORMATION (see Section 2.13). The closer the two are in parallel, the better it is for the business, implying better profitability. Business next generation products emerge quickly from BD and pattern analysis. Effective product management strategies will result in increased revenues, lower costs, improved profitability, enhanced levels of customer service and better returns management.

Relating customer identified demand patterns and trends to buying is a necessity as mentioned by the Merchandising department, when the two parallel with minimal differences then a business has got it right and this promotes competitiveness. Customer click-through, browsing patterns, conversions and struggles are insightful customer data that enhances customer service.

4.5.2.6 Themes: Business Intelligence

The interviewed participants had mixed feelings; these participants were unsure about BD curation at Spree. Some considered Spree as curation BD; BD by some authors is defined as transactional and behavioural data collected or logged as users surf the system. This definition does not factor in size. P4, coming from a background of BD (Amazon) where data

is in large volumes and readily available, considered the availability of data or what is being curated as not being qualified as BD. What is eminent is the fact that Spree is a start-up which was mentioned by all participants, hence the lack of historical data and BD. Many BD participants are also operating with a limited definition of the term BD as they perceive data to be BD only if it is in the realm of the terabytes.

4.6 Headline findings by department

Figure 0.8: Chapter layout – Headline findings by department

4.6.1 Findings: Business Management department

The Business Management department decides on the platform for systems from conceptualisation through inception to completion and maintenance. Since most of the systems used are in-house, building it makes it possible for the organisation to collect as much data on a consumer as possible. Participants from this section are strongly opinionated that Spree is curating BD though the organisation might not be doing it on a larger scale or to the degree needed to garner the full potential of BD.

Customer tracking is important, so all customer or consumer activities are logged, including data needed for tracking consumers. This data is in the form of transactional data and user behavioural data (weblog or clickstream). Most of this data are logged by default on Magento, Google Analytics and the web servers.

4.6.2 Important findings

- Spree is curating BD but not on a large scale
- Systems are built in-house from ground up so there is the potential to collect every piece data that is available
- The data collected allows Spree to track customers

4.6.3 Findings: Technical development team

The technical team works hand-in-hand with BI and Management to identify and implement the collection of data. The department's primary focus is to create, develop and ensure the e-commerce system is running with minimal or no down time. Its functionality is more client-facing as the UI and the functionality of the site falls under its list of responsibilities. This department thinks of BD as Hadoop data clusters and the like, not having Hadoop playing a role in the systems implemented implies no BD. The amount of data being generated is so small it may fit onto an Excel spreadsheet and as a result there is no BD.

- There are no Hadoop clusters so there is no BD
- Spree is using Magento as a persistence platform; it is MySQL, so no BD
- Spree uses Google Analytics to track and store most of its data
- Spree is still a start-up
- There is very little historical data and available data is not qualified to be called BD

4.6.4 Findings: Merchandising department

Participants from the Merchandising department seem partially aware of BD—not relating to Spree but from a previous work environment in retail, except at Spree there is no implementation that connotes or correlates to BD curation. Spree is still a start-up dwindling in small data, with much manual data handling. This increases the propensity for errors and compromises the quality and integrity of data for decision-making.

- There is no BD
- Manual handling of data makes data prone to errors and compromises data quality and integrity
- The level of automation only touches on a few reports but what is ideal is having the entire Spree run on pre-set systems with checks and balances without any human intervention

- Spree is profitable when customer buying parallels spree product buying; this necessitates understanding product demand patterns reflected in customer buying insights

4.6.5 Findings: Marketing department

The Marketing department communicates with customers through emails, online advertising, blogs and social media. What is salient is the lack of integration of a products life-cycle, customer life-cycle and the needed segmentation to allow for targeted messaging; hence messaging has to be vague so as not to miss customers except the impact of the message is reduced, yielding far less results. Furthermore, Spree insists on traditional marketing as it is justifiable in terms of expenditure. Growth hacking, a lean start-up approach, is not quantifiable, hence not an acceptable marketing strategy.

No targeted messaging as of yet, messages need to be vague not to miss customers.

The vagueness of messages reduces the impact of messages.

Contact with customers has to be strategic, not spam or ward off customers but rather attract them and retain them through a well-managed system of a customer life-cycle and messaging.

- There is no product life-cycle management yet
- There is no customer life-cycle yet
- There is no customer segmentation and customer profiling to allow for targeted marketing
- Growth hacking is not accepted as a marketing strategy (cannot be measured)
- Traditional marketing is preferred and can be measured
- Targeted messaging is ‘super’ important but there is no provision yet

4.6.6 Findings: Social Media and Marketing department

Participants work with data and are fully aware of the essentials of data but barely aware of BD. Yet in this department, participants acknowledge that social media is reactive and needs a quick turnaround in data management. This department uses blogs and consumer impersonation to create content as there is not much content data and historical data, though it tries by every means possible not to be viewed as aggressive to the consumers. This, the department believes, will alienate consumers and create a negative customer sentiment and weaken brand affinity.

Data, according to this department, will increase the scope for creative thinking. According to the Social Media and Marketing department, there is no BD, and even if the organisation does have BD, none of is available to them.

- The current Spree consumer is different from the incubation consumer
- There is a bit of historical data from the incubation project but not suitable for the Spree consumer
- Message packaging is important due to tone, content and relevance to the consumer to have the right impact
- A massive amount of data is needed for this kind operation to generate insight for decision-making
- The organisation may incur loss by not being able to target consumers at the right time due to inadequate data
- Data is needed to shape customer expectations and profile customers
- Brand awareness through paid advertising, Google advertisements and web banners need data to evaluate their effectiveness
- Visual communication may improve and boost marketing
- Enriched data with all relevant fields may lead to quick unique selling points
- Some of the important data needed for this department are spend, traffic, pages accessed by customers and analysing viable information
- Understanding that the customer service variable will contribute to growth
- Better insight into data means less room for errors
- Customer profiling helps segment customers for targeted communication and gives insight about the customer life-cycle

4.6.7 Findings: Business Intelligence department

The BI department creates data structures to curate data and work towards centralising data. This is side-lined by other responsibilities like provisioning data for use by other departments in the form of raw data (for self-service), reports, metrics that reflect high level figures of performance and results of tests. While this department is in the centre of data curation, findings from participants in the department reflect that participants may have heard of BD but are unaware of BD and its opportunities to the organisation. Some participants are completely unaware of BD, and furthermore, the department does not consider what data it has or is currently collecting as BD. At the start of the interview, the participant started off highly opinionated that Spree does not have BD and is not curating BD. Halfway through the interview as participants became aware of clickstream data collected in the Magento

repository, participants changed their minds to indicate that Spree is collecting BD but not analysing as of yet.

Findings can be summarised as follows:

- Spree is not curating BD and is not in the realm of BD yet
- The amount of data available or generated thus far does not qualify for BD
- Spree does not have the necessary technology for BD
- Until the data has been centralised from the three silos, there cannot be a needs analysis with present available data
- Data sources are fragmented, decentralised and different
- Data quality is low due to wrong coding and manual data handling
- There is a lot of inaccessible data with magazines
- Department is not overwhelmed with data influx yet
- Spree is a start-up
- There is no historical data
- Spree is currently building data marts and a data warehouse
- Spree is collecting clickstream data so there may be BD

4.6.8 Findings: Supply Chain and Customer Support department

Data, according to the Supply Chain and Customer Support department, is paramount in capacity and labour planning. There might not be a massive historical data as Spree is a start-up unlike Amazon, but what is available is a good candidate to drive decision-making and lead time prediction. It is important to realise that there is always an end goal for using BD as stated by participants. For example, fraud bid data can be used to make sure the department identifies fraud in time and blocks that order. The aim of BD is to produce some analytical information into the system for personalising or for the following:

- Spree is not curating Big Data
- There has to be a clear data goal when curating Big Data
- Big Data is the basis of systems personalisation and recommendations
- There must always be an end goal to using Big Data
- Big Data can distract you from your goals
- There might be BD in the form of clickstream data
- Big Data is an extension of data

4.7 Summary

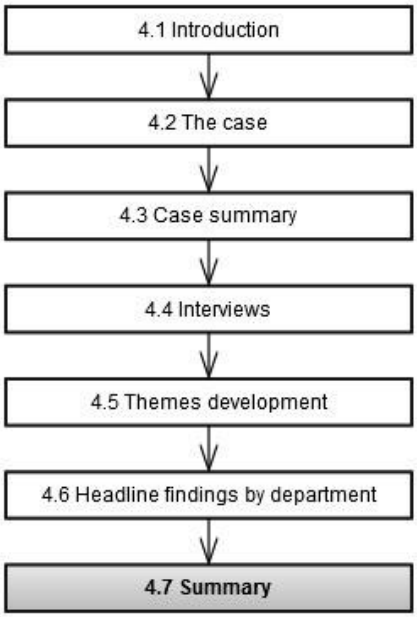


Figure 0.9: Chapter layout - Summary

Spree is an online e-commerce store which operates on the affiliate business model with nine magazine affiliates, and traffic redirected from affiliates that lead to successful sale results in a commission for the referrer.

From business interviews, it was quickly obvious as all of the participants mentioned Spree as being a start-up. This, according to participants, prevented Spree from having the needed colossal data to function in the realm of BD when compared to Amazon and other already established organisations. The lack of huge data implies there is no historical data and the absence of historical data implies it is impossible to study the past to predict the future, which is the whole concept of leveraging BD.

Furthermore, participants mentioned that real-time data for predictive analytics is needed, except where current curation practices are anchored by traditional curational processes. Reports and accessibility to data is one day late; this is the closest to real-time data currently available to Spree. For now it is not a problem, but it will become a problem soon due to competition and economic pressures in the market space.

Most of the customer weblog data is logged for customer behaviour analysis. Data in the environment of Spree is in silos and not concurrent as data in GA seem not to correlate to what is in Magento. These silos make it difficult to perform any meaningful analysis of data. As a result, the business has initiated a centralisation process in the hopes that the data will answer all its questions.

Many participants are unaware of the BD phenomenon. Though every participant acknowledges data as important, some participants suggested that BD is an extension of data and that since Spree is collecting weblog data in the form of clickstream, it implies BD exists. Yet again, some participants mentioned that the lack of huge data size with the present data disqualifies it from being BD. In terms of policies and strategies for curation, Spree has no formally documented policy documents or strategies that standardise or formalise curation. Participants also acknowledge that curating BD furnishes the curator with a competitive edge to promote competition.

5 CHAPTER FIVE: DISCUSSION

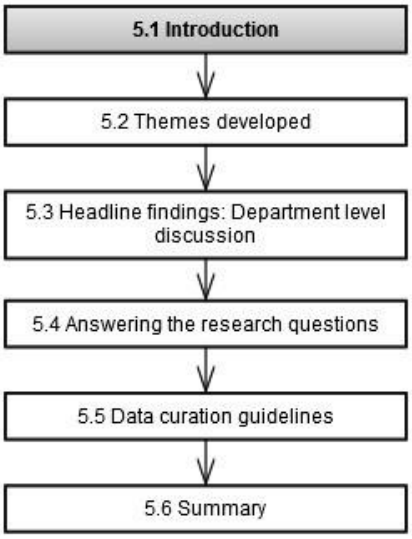


Figure 0.1: Chapter Five layout - Discussion

5.1 Introduction

Chapter Five presents a discussion of the research findings. The discussion is divided into five parts, namely introduction (Section 5.1), themes developed (Section 5.2), headline findings at departmental level (Section 5.3), answering the research questions (Section 5.4), data curation guidelines (Section 5.5) and summary (Section 5.6). Twelve high level themes emerged from 61 findings. The themes include data as an enterprise asset, profiling the customer, developing and satisfying customer needs through planning and decision-making, competitive advantage from information advantage as a product of guided data curation, marketing optimisation, service delivery, sales management, real-time data analytics, building strategies and BD integration into a business model as a business re-engineering process.

This chapter brings to light the concept of BD as an enterprise resource which may enable forecasting the future from real-time and historical data. The discourse covers Solution Engineering, Decision Theory and a BD Business Model Maturation Index as necessary concepts that shape and help create a successful curation process. Chapter Five also adopts the Business Model Maturation Index of Schmarzo (2013:6) to access the level of integration of BD in the case business model. The discussion draws from the findings to propose guidelines and recommendations. Chapter Five is a discussion in the context of the research problem (Section 1.3), namely that companies find it difficult to leverage the opportunities BD offers them in terms of monetising insights or curated data.

5.2 Themes developed

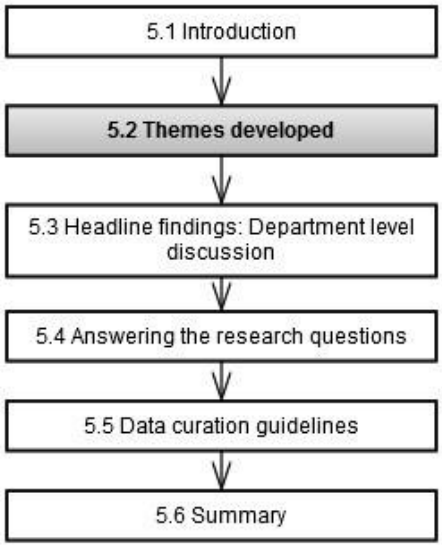


Figure 0.2: Chapter layout – Themes development

Twelve themes emerged from the data analysis (Annexure A, Table 7.11). Most of these themes are discussed in the various sections within the literature and will be further discussed in Sections 5.2.1 through to Sections 5.2.12. The themes are discussed below.

5.2.1 Data as an asset

The essence of curated data is the value it holds in the form of insight or knowledge (data abstracted and contextualised). Sections 2.3, 2.6 and 2.7 address the essentials of data, especially to the merchant from the stance of value proposition. Value proposition as mentioned in Chapter Two is an integral part of the business model. The business model encapsulates all features that are supposed to attract prospective customers to the proposed service(s), which is an important part of the business model. In the transaction and service delivery process, generated data assists the business in learning about its customer's engagement with the business, a process enlisted as part of the business model and plan. The business model (Sections 1.8.2 and 2.6) puts forward strategic choices which require data, as a measuring tool to articulate metrics that lead to business knowledge. This can be leveraged to re-engineer the business model. The value network component of the business model in Section 2.6 highlights the information flow and relates customer, product and supplier information.

Customer data and transaction data are collected at customer touch points for learning and re-introduction into the business processes for optimisation. These data sources hold the potential to re-engineer the organisations value creation process.

Spree (the case) acknowledges the potential that data holds to transform the way it enacts business. The case also shares the view of Schmarzo (2013) in that the insights generated from data sources can be used to optimise customer engagement. From a high level the business model, as an implicit business definition, dictates the information and technology assets required for business to capture value. This is the value captured from the business model perspective as differentiating factors such as cost and profit.

It is imperative to identify the data assets that will drive business transformation forward. This is done by identifying the organisation and department's information requirements and data needs through the use of a data model and enterprise data model as mentioned in the DMBOK guide (Mosley *et al*., 2009). Although the case recognises data as an important asset, the interviewees were divided on the availability of BD and data curation or lack thereof within the organisation's domain. This seems to arise from a lack of understanding of BD, the definition of BD and the functional benefits to the curator. It is not a new phenomenon as Khatri and Brown (2010) report similar findings in that many organisations do not know what data they have, how critical that data is, the sources that exist for critical data or the redundancy degree of their data assets. Furthermore, in a research study which interviewed executives of 51 companies in order to understand how organisations generate value from data, Ross, Beath and Quaadgras (2013) mention that organisations do a poor job of knowing what data they have.

Participants reiterated the importance of data to their departments but have contrasting views of BD and its curation as a dynamic asset needed by the organisation. When participants were asked if Spree is curating BD, participants' had contrasting viewpoints: they were either unaware or aware but flagged the curation processes as insufficient. Participants stated many other reasons why the data at Spree may not qualify as BD. One reason being, curation processes are not affected by programming models such as Hadoop and as such, relatively cheaper commodity technologies are used, for example Google, Yahoo and Amazon. Predominantly, participants mentioned data size as the number one reason why data should not be classified as BD.

However, some participants were strongly opinionated about BD in the case domain. They asserted that Spree has BD and is curating BD. This group of participants mentioned that the technologies are only relevant to help manage the data and surmount the challenges associated with the curation as asserted by Manyika *et al*. (2011) and discussed in Section 1.8. According to Gualtieri (2012), as indicated in Section 1.8, Big Data is about using all the data within the merchants' domain to optimise service delivery to the customer.

Interview participants acknowledged that leveraging BD for insight remains a challenge at Spree because there is no proper data curatorship and responsibility (informed task segregation) for data. This is due to the lack of a BD strategy (Annexure A, Table 7.7 and Findings 38, 60, 4 and 40) as the case is still in an early development stage of its business life-cycle. Spree being a start-up creates the opportunity to build its production system from the ground resulting in a unique opportunity to decide what data to collect. They can decide to collect, for example, data on customer behaviour so that departments can draw specific insights required for specific actions and results. The aim is to supply business with actionable data to re-engineer the business model to be more relevant and remain competitive in the market space through learning and re-introduction of lessons learned back into the pool of input data.

To collect and use data for the different departments in a relevant way, it is imperative to define the information needs (data to empower departments and individuals to make decisions). The information needs of an organisation form the primary contextual composition of an enterprise data model. The enterprise model provides a way of capturing and defining the information needs and data requirements that can be restated as the information and technology assets. For the data requirements, a data model is required that accurately expresses and effectively communicates the data requirements. This resonates with a function of data models, which is a bridge to understanding data between people with different levels and types of experiences and expertise.

The enterprise data model (see Section 2.6 for details) is an integrated subject-oriented data model defining data used across an entire organisation. The data model, as depicted in Section 2.6, is critical to capturing value about specific business elements such as the customer, in the domain. The purpose of the data model at depicted by Mosley *et al*. (2009), makes the data model a necessary input to all future systems development projects and the baseline for additional data requirement analyses and modelling efforts. Spree does not have an enterprise data model (see Annexure A, Table 7.7 and Finding 55). The lack of a data model creates a communication gap. Khatri and Brown (2010) state that in order to manage the inventory data as well as related sources for leveraged benefit, data curators (or stewards as indicated in the DMBOK guide) need to develop an understanding of the types and sources of data, the storage requirements, growth trends and integration. Data curators need to be fully aware of what data they have and how to put this into action for the needed benefits. Ross, Beath and Quaadgras (2013) support Khatri and Brown (2010) by mentioning that many organisations do not know what data they have, the sources that exist or the redundancy degree of data assets. The implication here then remains that an important aspect to launching a BD journey is to know what data already exists.

Rele (2012) states that BD is about understanding how to use all data to meet customer needs in order to grow their business (as discussed in Chapter Two). The understanding of how to use all data is a valid input to knowledge and negates reliance on technology to define the true meaning of BD because BD means many different things to different people. According to Gualtieri (2012:10), BD is "the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers." The difference in views on the availability of BD and its curability implies the need for a discussion that focuses on a reasonable judgement of current data. According to Manyika *et al*. (2011), BD is not about size as the word "big" but depends on the context in the sense that business situations vary, and that there is no volume limit or threshold that makes volume or size the sole validating criteria. This notion is shared by Zikopoulos *et al*. (2013:3) as they define BD as "the adoption of new business processes and analytics approaches that take advantage of that data without any mention to size". According to Manyika *et al*. (2011), the rate of technology advancements and increase in data size warrants the question, "how big is Big Data?" This implies that defining BD in terms of size will not be appropriate as there is no limit on data size. Spree participants aware of BD viewed BD only as huge volumes of data and managing the data requires adopting a programming model such as Hadoop (see Section 5.3.6 and Table 5.6).

The different viewpoints surrounding the availability of BD in Spree is partially caused by participants having misguided notions about BD. These misguided notions imply a specific definition and applicability of BD in the case domain. Participants who were fully aware of BD and acknowledged its associated benefits mentioned that Spree is not curating at the best level because the company is still a start-up and its technology requirements are not fully defined and developed yet. Spree has data spread across Google analytics, Magento and On the Dot. Spree is affiliated with nine magazines and will soon add many more as part of its business model (see Section 2.3.3).

The magazine companies have been in business for many years, some for decades, referencing business data. Magazines have big volumes of data but the data were not curated with analytics in mind. According to interview participants with knowledge of magazines, the way in which data is stored makes the process of leveraging data for insight virtually impossible as the magazine environments are dominated by print and paper technology. This may be as a result of the fact that the data (dark business data) is not curated with analysis in mind and presents a new challenge of having to manipulate the data for the new intended use. Furthermore, except for magazines, all other companies regardless of the affiliation to Spree operate as independent entities whose data are not accessible to the case environment.

It will be valid and beneficial to accept the fact that Spree may not have BD (only huge sizes of data) for now, but the context in which Spree is operating in and the need for detailed information of customers, BD should become part of the Spree strategic plans to place the company among information advantaged organisations (see Section 2.6.1).

Sathi (2012), Schmarzo (2013), Manyika *et al*. (2011) and Nerney (2013) acknowledge that BD is a disruptive technology. It is a multi-year multi-phase journey that mandates a strategic vision which is aligned with industry. However, it is equally essential to have short-term projects with measurable outcomes. Short-term BD projects can be captured and directed with a strategy document whilst a business data model may direct curation for long-term projects and global views of data.

5.2.2 Profiling the customer

It is important to track customers that browse and purchase on the Spree web site. This enables Spree to intelligently contact customers, as mentioned by interview participants. Spree builds its system in-house, thus creating the opportunity to collect all available customer data on-site. Through literature it became apparent that weblogs or clickstream data holds valuable customer insight in the form of preferences significant to knowing customer likes and dislikes. This forms the basis of customer segmentation and profiling as well as relating to the viewed or purchased products. According to Hurtgen, Natarajan, Spittaels, Vetvik and Ying (2012), a highly detailed and segmented view of the customer base enables improved customer service and personalised recommendations for target market efforts, cross-sell and up-sell campaigns. The lack of customer profiling and segmentation implies rifeness of vague messages to customers for an all-inclusive effect. This may reduce the impact and persuasiveness of messages and marketing to the customers. The essence of marketing is the message. The message is the value proposition which is only effective if it has a high impact. However, despite having a high impact, this might be weakened when directed to many different customers as relevance to the prospective customer is key to gaining the correct results.

The value proposition component of the business model aims to attract the prospective customer; the impact is accentuated when it is directed towards a segment of customers with well-defined interests. Spree as a business has a clearly defined target audience in the market space, which is a female mother at home who is connected to the Internet and has the resources to purchase. The market can further be broadened to include young connected professionals. Within this framework there are many smaller segments that can be targeted through the profiling of the individual customer for direct marketing. The more robust the customer profile, the more impact the analytical models will have. Profiling customers to

identify who the organisation's best customers are is important for customer management and sales improvement. Schmarzo (2013:162) states that:

> Combining detailed dark customer transactional data with social and mobile data and advanced analytics may uncover insights to optimise customer engagement life-cycle processes such as targeting, segmenting, acquisition, activation, and retention.

5.2.3 Developing and satisfying customer needs

Capitalising on the power of BD for retail requires combining customer and products insights. The organisation needs to be able to tailor offers and fine-tune each of the customer channels to maximise the appeal of products and drive more sales. With the advent of the Internet, customers are exposed to the ability to shop, compare and evaluate products from many different sites prior to making a decision; participants referred to this as smart buying. This is also applicable in the context of sellers buying to satisfy customer trends and behavioural patterns identified from data. Smart buying is a direct benefit of sales data and market basket analysis according to the Spree merchandising team. The merchandising team purchases products from suppliers after evaluating sales, customer preferences, user behaviour, market basket analysis, product information, fast selling items and traffic analysis to ascertain what items customers may be looking at. This promotes ordering products to foster cross-selling and up-selling which is a fundamental part of market basket analysis. According to the Merchandising department, identifying and categorising products as to best, medium and poor sellers assists the department in buying smart and meeting customer needs. Furthermore, when the two kinds of buying (customer and merchandise) are paralleled, the organisation becomes more profitable.

5.2.4 Planning

Pea (1982:6) defines planning as "a complex form of symbolic action that consists of consciously preconceiving a sequence of actions that will be sufficient for achieving goals". This may be inferred to as the action phase of Solution Engineering according to Nikov (2012). The process of solving a problem or creating steps to solve a problem requires knowledge of the structure in which the problem is embedded. A successful transition to the desired state hinges on following the right steps which make up the plan required to accomplish the solution. This process is comprised of formulating acts that may lead to accomplishing goals.

According to Spree's Supply Chain and Customer Care interviewed participants, decisions and plans regarding the entire back-end rest upon data through which results for seamless management of the back (warehouse) will be attained, and without which the back-end will

be a total mess. Insights from data depend on the planning warehouse storage; capacity and labour depend on insights from data. The current data *status quo* does not promote much of planning with data to a fine level due to the lack of BD sizes for analysis and conclusive findings. Over time and with the accumulation of data, it will become more profitable to hinge on fine grain data to predict lead time and even delivery times. Planning directed by insights from data allows for the creation of a management strategy that promotes better management of the back-end and a high reliability of forecasts.

5.2.5 Decision-making

New and emerging business processing now makes possible what was previously impossible, i.e. operational business intelligence that hinged on analytics to improve business agility which enables automated real-time action and operational decision-making (White, 2011:2). Supporting Big Data to satisfy organisations' business agility implies combining different technologies as part of unified solutions architecture. This is important in today's fast paced business environment of needed fast decision-making (Taylor, 2013). Not all business decisions have to be made in real-time, but some intra-organisational tasks such as processing a credit card purchase may require a quick decision which depends on business agility and the capability not to make poor decisions. Poor decision-making leads to the demise of an organisation, as mentioned in Chapter Two and seconded by interview participants. Contrary to poor decision-making, smart decisions backed by insight from data and analytics will foster business growth and process optimisation, which implies monetisation. According to Davenport (2013:10), BD brings to the table attributes such as the creation of intelligent software, scale out infrastructure and large datasets. These attributes do not only speed up decision-making and data processing but also fosters better internal or customer focused decisions which also ties in with the benefits of the 360 degree view (see Section 2.7) of the customer as a benefit of customer profiling. According to the Business Management department at Spree, business managers at inception used to make decisions based on gut feeling, but management has transitioned into becoming more adept at using data to make decisions. This transition has necessitated a more strategised approach to leveraging data for every needed decision. Identified problems needing redress within the domain environment can be applied to the concepts in Section 2.13 on Solution Engineering and Decision Theory to fabricate a solution state guided by curated data. The section mentioned the value of perfect information as being the quantitative difference between known conditions of certainty minus expected value under conditions of uncertainty. This quantitative state can only be established through curated data and insight from analytics. Knowing the numerical benefit associated may help narrow options pertinent to decision-making as the fewer the presented alternatives in a dilemma, the easier it is to make a necessary decision.

Agrawal *et al*. (2012) and Sathi (2012) acknowledge that the benefits of working with data are many. Manyika *et al*. (2011) concur by mentioning that the benefits are manifold. Business managers understand numbers, facts and figures. With the data influx and the benefits of obtaining actionable insights from BD for decision-making, organisations stand a better chance to monetise curated data. Big Data is a common language that can be used to improve products, services, business models and strategic decisions, unify departments and improve collaboration.

5.2.6 Competitive advantage

There is considerable evidence that decisions based on analytics are more likely to be correct than those based on intuition (Davenport & Prusak, 2005). This in evident in Whites comment that leveraging the value of Big Data helps reduce costs, increase revenues and improve competitiveness. This is the reason why business intelligence and analytics remain top priority on the list of CIOs (White, 2011). The benefits of leveraging BD help reduce the time to value generation by fostering rapid decision-making. Participants aware of BD reveal that BD holds the key to competitive advantage due to service footprints in historical data and the ability to make predictions from the data. Spree may have Big Data, as mentioned by some participants. Its superimposition on the maturation index pins it at business monitoring (see Section 2.13.1) mainly due to the technologies deployed in data curation, which means that data curation at Spree is based on sustainable technologies, the likes of MS SQL server and other relational toolsets. It is important to note that as data collection continues and data sizes grow the organisation is bound to seek alternatives, though this might not be at the fore front of strategy now. Present curation processes are not optimised to surmount scalability challenges; furthermore the reliability of generated data is questionable due to manual data handling. A data strategy geared for BD must have cleansing applied at capture points to retain the confidence in the data coming in. Confidence in data quality will lead to confidence in analytic results and a high integrity of generated actionable data.

Being competitive requires the ability to act upon actionable data quickly to gain the benefits, while the current lack of real-time data implies inability to gain or take advantage of actionable data immediately. Progressing to the next level of the BD journey from monitoring requires leveraging insights from data to optimise processes. This requires advanced analytics, real-time data and dark transactional data to materialise gaining the benefits thereof. Spree has customer data comprising of customer information, sales data and transactional data.

Besides these types of data, the magazine has enormous customer and demographics data which may be deployed for analytics. Google and Facebook marketing provides platforms for

leveraging Big Data. The data from the Google and Facebook platforms combined with external data sources for marketing purposes may furnish Spree partly with the needed competitive edge. Literature and primary sources of data concur that BD is a differentiator to furnish insights and unlock the power of data. The insights with actionable data, preferably real-time data, will position an organisation to outperform competitors.

5.2.7 Marketing optimisation

As part of attaining set organisational objectives partly through increasing visibility, the Spree Marketing department takes traditional as well as an Internet based approaches to marketing, which include creating impressions on TV, radio and the Internet. Free Internet marketing approaches such as viral marketing and growth hacking techniques which are a complete break away from traditional and other quantifiable methods, are not considered by Spree because the outcome or impact is not quantifiable. Marketing is a process responsible for identifying, anticipating and satisfying customer requirements profitably. Londre (2009) expounds further on this by stating that marketing involves the process of planning and executing the conception, pricing promotion and distribution of ideas, goods and services to create exchanges that satisfy individual and organisational objectives.

The business monitoring phase, according to Schmarzo (2013), Schroeck *et al*. (2012) and Mohanty *et al*. (2013), is a starting point to integrated BD in order to gain insight for marketing as the organisation has already gone through the process of identifying key business processes and capturing KPIs, dimensions, metrics, reports and dashboards that support the business processes. What then becomes eminent and next is a transition to the optimisation phase that reflects effort (marketing trends and spends) in relation to returns. The aim is to improve marketing returns in the form of sales as quantifiable benefits from marketing efforts. According to participants at Spree, targeted marketing and segmentation of customers are vital to improve sales and conversion as marketing becomes more targeted, with the impact on the recipient being better and directed.

5.2.8 Service delivery

Knowing the organisation’s customers and profiling them is the start of gaining better insight into customer behaviour and driving sales and repeat sales (Rijmenam, 2014; Davenport, 2013). Customer profiling allows the organisation to engage customers in ways directed by customer insight. According to Rijmenam (2014), Amazon, a market leader in retail, has perfected the art of analysing their customer data to gain insights mainly to drive sales. This is evidential in the functioning of the recommendations engine on their website.

Smaller companies as well as offline companies should leverage BD to drive sales to the next level (Manyika *et al*., 2011). Providing an optimised service empowered by data comes from knowing what data has been collected, what data is required and how to combine the different datasets to gain insights required to increasing and improving sales. Findings 38, 11, 43, 30 and 5 indicate that participants were not aware of exactly what data is being collected and what is available to them, thus depriving them of the richness and possibility of a data-centred service design that engages a customer based on insight.

The quality and level of service to customers indicates accomplishment relative to the Six Sigma goals (García-Alcaraz, Maldonado-Macías & Cortes-Robles, 2014). Six Sigma according to (ibid) is a business process that allows organisations to drastically improve their bottom line by designing and monitoring everyday business activities in ways that cut down waste and resources while increasing customer satisfaction. In the domain of Spree, participants mentioned a Six Sigma goal necessary to reduce errors to a bare minimum. Error reduction leads to less disgruntled customers, a reduction in revenue loss and better service to the customer. It is important and mandatory to minimise errors made per million opportunities as smarter customers do not only look for lower prices but also for a competitive service that parallels the price. García-Alcaraz, Maldonado-Macías and Cortes-Robles (2014) mention that to compete, companies have to provide superior service and quality in order to capture a reasonable market share and increase customer satisfaction. At the Supply Chain and Customer Care departments of Spree, one well-circulated goal is to resolve a sale in about three seconds, as mentioned by an interviewee. While noting that, most fraudulent transactions are only detected by chance; BD and analytics improve the organisations potential to detect fraud efficiently while minimising the negative impact on operational productivity. Fraudulent transactions, as mentioned in Section 2.3 (Paragraph 11), impact the bottom line, therefore identifying it in time helps minimise associated loses.

5.2.9 Sales management

Rijmenam (2014) states that analysing website statistics and social media can help the organisation identify products that are regularly viewed. The viewed items may not always convert into sales, but it is important to understand the sales dynamics of the viewed products and the customer base for insights. According to Davenport (2013:10), the creation of new products and services for customers is a benefit of curating BD that cannot be overlooked; these benefits have been aggressively pursued by the likes of Google, LinkedIn, and Facebook. The benefits help to gaining insight into sales barriers such as low conversion rates; this may assist the organisation in optimisation as they leverage the benefits of data curation and not losing sales. This is data that traditional analytics already provide, for example Google Analytics.

Combining sales data with social media data may uncover relevant insight into customer purchase patterns, how to implement sales dynamics, and on what bases to implement them. These may include dynamic pricing, recommendations and even yield management.

According to the Marketing department, it may be essential to at all times have insight into external data, especially competitor prices through crawlers that provide real-time data. The rifeness of data on the e-commerce market space has empowered customers and business alike to buy smart; this makes it imperative for retail companies to be alert and keep abreast with competition and contender prices so as to remaining competitive.

5.2.10 Analytics of data

BD analytics, according to Sathi (2012), is becoming integrated with processes and traditional analytics to provide major outcomes such as customising, personalising and changing products based on customer feedback. New products based on insights from curation combine components in response to customer's needs. Spree participants mentioned the lack of data analysis due to data fragmentation, disparity, inaccessibility to affiliate data and the lack of BD-oriented technologies. For an organisation to make use of BD it is essential to know and trust the integrity of data, but the sheer volume, complexity and related difficulties imply insufficiency of traditional and manual methods of discovering, governing and correcting data. Curating BD for analytics becomes worthwhile when integration and governance are implemented with BD applications. Participants mentioned unawareness as to what data exist or form part of curated data. This is asserted by Khatri and Brown (2010) in that many organisations do not know what data they have.

Social Media department participants mentioned influencing communication as the organisation tries to create an impression without alienating customers or spamming customers; segmented and profiled customers from analytics will provide grounds for targeted communication. Participants mentioned again that insight from social media is "reactive requiring", thus implying a need to quickly act upon actionable data. With current curated insights coming through to the department one day late, the impact of social media actionable data is lost. Furthermore, what data is curated 'at the back' remains something to be discovered by the department. This creates a gap that precludes the Marketing and Social Media department from leveraging data opportunities. Data received from BI is a day late, implying the lack of real-time data for analytics. Influencing and shaping the consumers impression of Spree depends partly on how the department engages customers on the social platforms; executing the underlying activities without proper analytics and insight about engaged customers means running blind which implies being ineffective.

5.2.11 Building strategies

Competition is at the core of the success or failure of firms; it determines the appropriateness of the firms' activities such as innovation, cohesive culture and implementation, which contribute to performance (Porter, 1998). As mentioned in Section 1.2, integrating BD into corporate information architectures provides deeper insight into what customers are thinking and how business operates. Some of the tangible payoffs of integrating Big Data into the organisation's strategy include better sales with lower returns rates and improved customer retention. According to Wang (2014), strategy sits at the heart of a competitive approach, which he defines as an elaborate and systematic plan of action to be taken by an organisation with measurable decisions about the direction to be taken. Competitive strategy design hinges on two main questions, which are (i) the attractiveness of the industry for long-term profitability, and (ii) the determining factors, as strategy seeks to establish a profitable and sustainable position against market forces driving the competition.

Spree aims to answer business questions by bringing all fragmented data together; this is a strategy the organisation believes will enable it to gain more insight from all the data collected about customers, products and other business entities. The aim is to become more competitive by looking at the data previously curated in order to improve on internal processes—a direct benefit attained from the Six Sigma goals. The Spree business model proposes to provide service to segmented customers (see Annexure A, Table 7.3, Finding 31). For example, the urban middle-aged, stay-at-home mother taking care of the children with limited time for shopping may be drawn to the Spree brand due to the ease of purchasing online, fully trusting that she will receive the expected products and quality service. This business expectation requires a service design that meets customer needs, using insight from customer sentiments. This may mean incentivising loyal customers or building a strong loyalty program through innovative and optimised service delivery and engaging customers on a social platform for discussions about products and experience. To support this level of expectation and commitment, knowledge and management of the data becomes a critical success factor.

Most of the participants stated that Spree does not have a defined data strategy. BD puts business elements at the heart of corporate strategy (Lumpkin, 2013; Mohanty, *et al*., 2013; Stubbs, 2014). The strategy has to be concise, clear in defining the current and future state relative to what is to be accomplished, and also relevant to business stakeholders. This needs to be done by focusing the processes on supporting the organisation's overall business strategy in alignment with the business initiatives.

5.2.12 Transforming the business model

According to Teece (2010), business innovation will fail to deliver or capture value without a well-developed business model. Engineering the business model, an organisation process goes through various levels of maturity with insights from BD (Mohanty *et al*., 2013).The last two phase of the BD Business Model Maturation Index are data monetisation and business metamorphosis. Data monetisation is that point in a business life-cycle where the organisation is looking to leverage BD for new revenue opportunities. These revenue opportunities may be in the form of leveraging actionable insights that can be deployed or even sold to other organisations. These may also be in the form of recommendations based on customer behavioural patterns to rethink, restructure and optimise customer experiences, re-packaging products, and insights based on trends and patterns identified from market analysis for sale to other organisations and manufactures. Business metamorphosis takes it a step further where an entire ecosystem is created that empowers third party organisations to generate revenue from an analytics-enabled platform from customer usage patterns, products performance patterns and market trends. Business model transformation implies transforming business insights such as customer and product insights to move the business into new markets.

It is mandatory for an organisation to broaden its charter to provide a more comprehensive solution to its customers. This requires understanding the broader picture of what customers are trying to accomplish against what they are doing. Changing the business operational strategy which affects business processes for optimisation implies a need for business model transformation as the business model is only “business process unified”. The Spree business model, as depicted in Section 4.2, creates a conceptual platform that affords Spree discoverability and improved visibility through its affiliate model, except the organisation lacks the ability to integrate its dispersed data into its business model to improve data usability. A similar system of affiliation is seen with Amazon, and also reported by interview participants, except in the case the Spree affiliates operate under one umbrella company, Naspers Media24, placing data within reach although data may not yet be readily accessible. The proxy locations and affiliations create a platform through which collaboration services can be shared.

5.3 Headline findings: Department level discussion

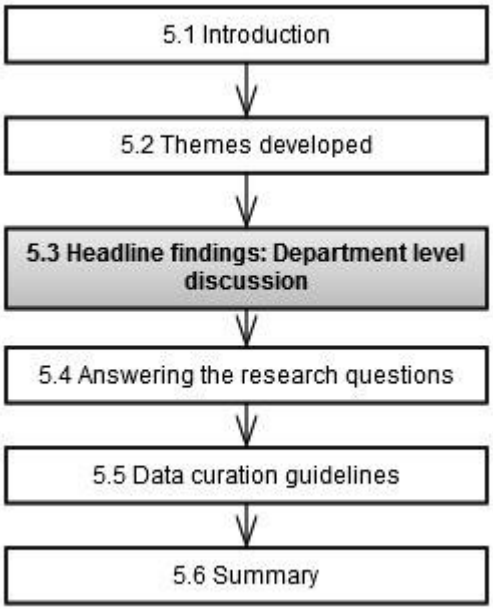


Figure 0.3: Chapter layout – Themes discussion

This section presents a summary discussion of findings pertinent to departments and how the different departments perceive the current data. The findings are in no particular order with respect to Findings in Chapter 4. These findings are from a high-level perspective and relative to departments. Findings reached saturation, hence recurring findings across departments are mentioned prudently to minimise reiteration.

5.3.1 Headline findings: Business Management

The state of the business is directly inscribed in the data elements, which warrants the need for reporting. Data has transitioned into becoming one of businesses most dependent assets. Data provides a way to communicate through status changes reflected in the state of data. This resonates well with Mosley *et al*. (2009), Schmarzo (2013) and Davenport (2014) who affirm that data holds the potential to transform businesses.

According to the BM participants, one initial activity carried out is the identification of what data assets are needed by the departments and the organisation as a whole (Pre-IQ Finding 8 and Findings 10, 17 and 47). But what is not stated is the involvement of stakeholders from the related departments. The identification of the data required is carried out by a selected few individuals from the Business Management department who may not possess the core requisite knowledge for a particular department. This is the reason for the continuous *ad hoc* reports needed from the BI department, as mentioned by BI participants.

Participants from the Social Media and Marketing department mentioned that it might be true that the organisation is collecting BD, except none of the data are available to them. This is due to the lack of a widely communicated business data model documenting what is being curated, what needs to be curated and what is available for use (decision-making and analysis). This constitutes a communication gap. Other interview participants such as Supply Chain and Customer Care stated that Spree does not have BD and the data available is not qualified to be branded as BD. Table 5.1 highlights a summary of the Business Management findings.

Table 0.1: Summary of Business Management headline findings

Headline Finding	Description
43	Big Data is an extension of data.
10	Business Management identifies what data to log per department's need for information.
5	Operational data users are uncertain about Spree curating BD.
7	There is a gap in communication between data curators and data users.
44	Interview participants who are aware of BD are operating with a limited definition of BD.

5.3.2 Headline findings: Merchandising

The impact of merchandising is reflected in sales through sale and demand patterns. Product sales, as mentioned in Section 2.6, depend in part on high-quality merchandising and an appealing store environment to attract consumers (Parker, 2013). According to Henke, Coulbourne and Kadochnikov (2011) the merchandising team acts as a primary educator with the initial task of being able to convince the retailer of the value of merchandising by pointing out the lost sales and profit resulting from sub-optimal stock turns caused by poor merchandising. Merchandising accentuates the importance of data to every business activity. Most importantly, merchandising data leads the buying team to profitable situations through products and seller categorisations, as this may foster smart buying except participants also mentioned the lack of BD. The reason being that Spree is a start-up company and there is not much curation in the form of historical data.

Also, in comparison with previous work environments, the data footprints and deployments are less and underrepresented as to how data can be used. Currently much of the data deployed to the department, including scheduled reports which are automated but still require human intervention, is data prone to errors and affects data integrity. Low data integrity impacts on decision-making as the integrity of findings is questioned due to inadequate representation. The Merchandising department also added that there is a need for real-time data as current data deployments provides actionable data one day late. This implies a loss of opportunity, especially with decision-making regarding social media

marketing as these environments are reactive, thus needing quick turn-around with data and other resources. The Merchandising department interviewees also mentioned that for business to be profitable, it is imperative to match customer buying trends to company buying patterns. When these two align well enough, it implies business is profitable. There are many factors that underlie this but the alignment is mostly based on evaluating demand patterns and forecasting what customers might be looking for. Table 5.2 presents summary headline findings pertinent to the Merchandising department.

Table 0.2: Summary of Merchandising headline findings

Headline Finding	Description
40	The business is a start-up so not much historical data exists.
38	Spree is a start-up and not functioning within the realm of Big Data yet.
41	There is a need for real-time data but current systems and implementation only provide data one day late.
8	Many data operations are still on a manual level, increasing the propensity for errors.
12	Business is profitable and competitive when customer buying patterns matches business buying.

5.3.3 Headline findings: Social Media and Marketing

Spree as an online e-commerce store benefits from extended visibility and reach. With the diverse marketing approaches available, it is able to reach many different connected and unconnected customers as it leverages both traditional and non-traditional methods of marketing. Data business management curators mention that marketing is extremely expensive. Using the Internet is an important platform through which business strategises to reach more customers, attract them and retain them. Business needs to find ways and means to reach customers with information about products and services that are relevant and capable of addressing needs. This only becomes possible when the organisation has curated data significant enough to expose customer needs in the form of trends and patterns that can meet or satisfy these customer needs. This implies smart merchandising efforts based on curated data such as insight from market basket analysis and demand patterns of products. The curated data will assist the organisation research and ascertain what prices customers will be willing to pay. Ultimately the organisation uses information to curb external competition data. Some of the sources of data needed as mentioned by participants include social media data, traffic information and general customer behavioural pattern data. Porter (1998) mentions that an organisation's competitive strategy may hinge on cost leadership, product differentiation or possibly remain in the middle, whatever an organisation decides. The strategy must enable the business to stand out in order to attain the value of competing in the market space.

Marketing at Spree is done mostly on a traditional level and is augmented with Internet marketing. Participants mentioned that the organisation's preferred way of marketing is through traditional means as returns are quantifiable with respect to spend. Free, viral marketing and growth hacking techniques deployed by lean start-ups are not favoured; the reason being the return on investment are not measureable, a lean start-up is a method for developing businesses and products (see glossary). As mentioned earlier, marketing is expensive and investing to create brand awareness and forcing brand impression through marketing must be done based on a strategy that is insightful and well thought-through. Participants indicated that there is no product or customer profiling which, as a result, reduces the impact of messages (messages that are not targeted or unsolicited may imply spamming in the case of emails). Profiling customers forms the basis of a customer life-cycle. A well-segmented customer base will facilitate the creation of a customer value proposition, which will reduce the need for campaigning that may not yield revenue. A customer buys from an organisation expecting a certain value based on what is on offer by virtue of the value proposition; meeting this value proposition rather quickly satisfies the customer and creates a happy customer. Integrating this information into the customer life-cycle or product life-cycle furnishes the organisation with richer insights into customers and products.

According to the marketing interview participants, messages to customers are generalised so as to not exclude potential customers; however, this may constitute spamming and therefore causing the organisation to lose customers, especially high profile customers. Profiling may enable business to generate more revenue as messages are focused and well-defined. Business will then be able to deploy a better yield of techniques in terms of dynamic pricing, recommendations and management of the loyalty program. Currently there is little analysis as data is structured in silos and unmatched across the silos. Table 5.3 presents summary headline findings from the Social Media and Marketing department.

Table 0.3: Summary of Social Media and Marketing headline findings

Headline Finding	Description
11	Spree logs varied data about business entities using the Magento framework, Google Analytics and On the Dot.
2	Data is critical for analysis and decision-making.
6	There is valuable (BD) data currently outside the reach of Spree.

5.3.4 Headline findings: Supply Chain and Customer Support department

Supply Chain and Customer Care mentioned that data is a technology power house; data powers everything within the department including labour management, warehouse capacity management and arrangement of products in the warehouse.

But while companies have an expectation from Big Data analytics in their supply chain, many struggle with the integration or adoption of BD as part of an enterprise-wide process. Insight from data allows for seamless operations within Supply Chain and Customer Care, without which the department will struggle to strategise its daily operational management. Supply Chain on its own collects data about products in the form of inventory, but most of the data within the department is divided into inbound and outbound data. Seeing that the department has to do forecasting of product arrival times both at the customer end and at the warehouse, it is important to build a repository of historical information for insight and prediction. Sometimes the factors that determine the state of a situation might lie outside the control of the organisation but with adequate data and use cases, the department is still able to forecast.

The Supply Chain and Customer Care department is adamant and vehement that Spree is not curating BD yet. According to the interview participants, Spree is not in the realm of BD yet and furthermore, the data the department uses normally fits onto Excel spreadsheets which disqualifies the current data as being classified as BD. A participant with about thirteen years of operational experience from Amazon stated that magazines have BD while Spree is collecting weblogs (Finding 57) in the form of clickstream data—this is the closest Spree is to BD. This impression is partly true considering some specific definitions of BD, but considering the definition by Rele (2012) who place the emphasis on the data and its uses to meet customer needs, this notion is discarded. Participants' view on Big Data reflect a specific definition of BD as the inference here to clickstream data being collected actually implies that Spree has BD. Furthermore, Manyika *et al*. (2011) state that limiting BD to size of data is unfounded as there is no upper limit to data. According to Schmarzo (2013), the process of moving from a monitoring phase to understanding using generated insights, needs BD in the form of BD drivers which include transactional data, unstructured data, real-time data feeds and the integration of predictive analytics. Spree has BD and is continuing to collect BD (Finding 30). Finding 43 indicates BD as an extension of data; there has been a steady growth in the number of unique visits and subscriptions on the site and to the Spree newsletter and customer footprints in the database and weblogs are continuously growing which implies a changing start that might need new approaches to handling data.

Participants again mentioned that magazine companies (Annexure A, Table 7.7 and Pre-IQ 6) have massive data which are not available to Spree to access, stating a few related problems with the way the data is stored and the complete inaccessibility of the data (Finding 6 and 27). Technically, magazines are a part of the business model which was an initially defined benefit of the establishment of Spree—the magazines are joined through the affiliate business model to redirect customers from their web site to Spree.

Insight into what the customer interests are in the magazine site domain might provide insight into the customers' activities on the Spree site. Big Data constitutes all the data within an organisation (Rele, 2012) and not only a selected dataset for a particular purpose. This includes the structured and unstructured data, macro and micro transactions and transaction oriented-data as stated in Chapter One on BD definitions. This conclusively indicates that Spree has Big Data and that participants are operating with a specific definition of BD and the organisation as a whole, thus not taking full advantage of available data assets through integration. Table 5.1 present summary headline findings from the Business Management department.

Participants mentioned that BD is disruptive as supported by literature (Section 2.8). Sathi (2012) indicates BD as a disruptive force as curating BD without proper insight into processes could lead to a loss of investment. Sathi (2012) also states that it is vital to curate data with a clear business goal. Table 5.4 presents summary headline findings from the Supply Chain and Customer Support department.

Table 0.4: Summary of Supply Chain and Customer Support headline findings

Headline Finding	Description
3	Data in Supply Chain and Customer Service falls under inbound or outbound data.
57	Spree is collecting clickstream data and BD is an extension of data.
21	Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data.
54	Supply chain and customer care may trace errors and sources of errors or defects using historical data.
35	It is critical to have a clear BD goal.

5.3.5 Headline findings: Business Intelligence

The current state of data at Spree makes it difficult to furnish business with analysable data. This is as a result of data siloes and data spread across multiple sites. Data is spread across Magento, GA and OTD, not mentioning the inaccessible magazine data. The effect of fragmented organisational data as mentioned in Section 1.8.3 predominantly in traditional organisations where many distinctly functioning departments exist, each of which is focused on a specific function with fixed operational boundaries. As a result the BI department is tasked with having to centralise all data (Annexure A, Table 7.7 and Finding 45).

The state of the data across the different servers makes it difficult for the department to leverage the data (Section 4.3.2.1, participants 2 and 4). There is a similar situation with magazine data, the data is inaccessible and the nature of the storage according to participants will pose a challenge even trying to bring the data together. Data at Spree is currently available in the form of curated transaction data and transaction oriented data are

optimised for business monitoring according to BI department. BI also mentioned that with Spree being a start-up there are a lot of things that the organisation is still learning about the market and that data curation for example information from AB testing helps them ascertain if a particular thing is working or not. Table 5.5 present summary headline findings from the business intelligence department.

Table 0.5: Summary of Business Intelligence headline findings

Headline Finding	Description
45	Business intelligence (BI) is in the process of building and creating data structures to centralise data for analysis to gain insight into curated data.
28	The way data is stored across the different servers may heighten the difficult in bringing the data together for analysis.
27	Data is spread across multiple servers outside the perimeters of the business making the data inaccessible for analysis.
22	Business aims to answer business monitoring questions from centralisation as these data are pre provisioned and in curation.
23	Business micro and macro data, though hind sighted, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business.

5.3.6 Headline findings: Technical development

The Technical department creates and maintains the Spree system through sprint work. Findings from this department also reveal that participants are unaware of BD as many mention that Hadoop clusters and data sizes are too small, but the organisation is well able to take advantage of cloud offerings which many small businesses and start-ups are doing already. As mentioned in Section 2.9 (Zikopoulos, *et al*. 2013, Angeles (2014), “Even a small company can have BD”. Table 5.6 presents summary headline findings from the Technical department.

Table 0.6: Summary of Technical department headline findings

Finding	Description
38	Spree is a start-up and not functioning within the realm of BD yet.

5.4 Answering the research questions

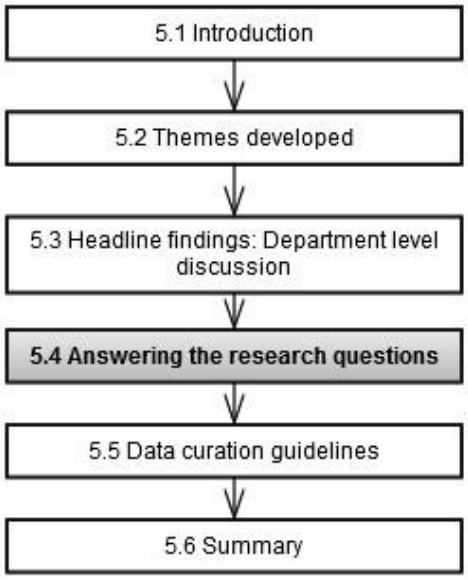


Figure 0.4: Chapter layout – Answering the research questions

5.4.1 Answering Research Question One

Research question 1: What are the factors affecting business to leverage BD for competitive advantage?

To answer this question, five sub-questions were asked. These sub-questions are summarised in Table 5.7. The table provides a high level overview of sub-questions and summarised answers as per findings from various interviews. The aim of question one is to establish the essence of data curation by identifying factors that spur the business to curate data.

Table 0.7: Summary of answers to Research Question One sub-questions

No.	Sub-question	Summary on a high level
1.1	What is business doing to leverage Big Data to gain a competitive advantage?	The main strategy besides data collection and management is to centralise data for analysis to provide answers and a unified view of data for querying.
1.2	What is the business's view of Big Data in terms of competitive advantage?	Data is an enterprise wide asset that has taken centre stage for decision-making. Data provides the keys to business transformation, optimisation and a source of insights. This illustrates the high level of importance Spree gives to BD's contribution to a competitive advantage for the business.
1.3	What are the policies and strategies for leveraging Big Data?	Not much is said about this as there are no documented strategies and policies.
1.4	What information does business want to get from Big Data?	Data to measure business performance in the form of critical success factors underlying business initiatives including KPIs and key metrics.

No.	Sub-question	Summary on a high level
1.5	What kind of data is being curated as part of Big Data?	Macro and micro data in the form of transactional and transaction-related data. Structured and unstructured data.

A review of the answers to research sub-questions one through five brought to light the fact that data curation in the Spree domain is spurred by factors both internal and external to the organisation, some of which the organisation literally had no control over except to strategise its operations and decisions to meet the imposed challenges. These factors include customer, products, planning, decision-making, competitive advantage, marketing, service, sales, analytics and strategy, which all impact on the implicit or explicit business model.

Research sub-question 1.1: What is business doing to leverage Big Data to gain a competitive advantage?

Participants mentioned that the organisation is not in the realm of BD yet as data sizes are still quite small and data is fragmented across Google Analytics, Magento and On the Dot. Magazine data is completely out of reach due to the method of storage and persistence possibly being another insurmountable task to achieve for analysis. According to BI interviewees, business has started centralising all fragmented data sources as it is believed that questions can be answer better with centralised data.

Chandramouly and Stinson (2013) and O'Shea and Shah (2014) are of the notion that centralisation is the key, but contrary to that, Casey, Harbitter, Leary and Martin (2008) believe that centralisation has to be contextualised and validated against factors such as cost, governance, data ownership, performance, function, scalability, security and privacy to evaluate feasibility and beneficence, especially for information-sharing and retrieval systems. LaValle, Lesser, Shockley, Hopkins and Kruschwitz (2011) are of a different opinion; they indicate that data curation should be started in the middle, which is recommendation two of four proposed recommendations they proposed. Recommendation two suggests that, for each opportunity, start with questions and not data. LaValle *et al.* (2011) further mention that organisations are traditionally tempted to gather all available data before beginning the analysis which has often lead to an all-encompassing focus on data management, collecting, cleansing and converting data, leaving little time, energy or resources to understand potential uses. They believe that centralisation of data has failed many organisations and therefore do not recommend this strategy as the starting point of curation. From an organisational perspective, the data challenges faced by Spree are many, including siloes of data, manual data handling, lack of real-time data, lack of customer profiling, segmentation, lack of historical data, poor data quality and incoherent reports, among others.

Eckerson (2011), on the other hand, suggests implementing hybrid architecture, i.e. finding a balance between centralisation and decentralisation; centralisation provides a high degree of rigor, while decentralisation offers a high degree of flexibility. According to BI participants, the organisation's data are spread across Google Analytics, Magento and On the Dot. Magazine data might be colossal and therefore justifies BD availability, but it is not accessible to Spree (which does not seem to be a problem to the participants).

As part of data curation and gaining insight from data, the BI department has initiated a data centralisation process aimed at bringing all the data together. Analysing data, according to the BI department, will lead to insightful findings which relate directly to business entities that deal directly with business value drivers such as customers, products, market analysis, sales, what the best decisions are, engaging with customers through sales, analytics, and competitive strategising for gains.

LaValle *et al.* (2013) suggest that organisations seeking to leverage the opportunities of BD curation should start in what might seem like the middle of the process, implementing the benefits of data curation by first defining the insights and questions (needs list) needed to meet the big business objective and only then proceed with identifying the datasets required for answers.

Research sub-question 1.2: What is the business's view of Big Data in terms of competitive advantage?

Business and all participants acknowledge data as a source of insightful findings to drive business processes, and most importantly business transformation, through business initiatives as critical success factors are met. From a departmental level, Merchandising formulates budgets and decides on quantities based on sales information. Market basket analysis and shopping cart information, including abandoned carts, is a major source of insight. Findings and generated insights enable the merchandising team to liaise with product manufacturers or suppliers and directing them based on these insights.

The Walmart case study is a typical example of how an organisation's view of data may change the entire facet of value creation at supply chain and merchandising levels. In the case study as in Section 2.1, data has the potential to transform both companies and industries. In this sample case study, Walmart invested in software that could track consumer behaviour in real-time for bar codes read at Walmart's checkout counters. This information was intelligently ported to significant stakeholders (manufacturers) to improve products and become more efficient. With this contribution, Walmart eventually ended up dictating to retailers how much product they could stock, at what price, and which promotions to use.

With gained information from point-of-sale systems, retailers were in the position to determine and know much more about consumer behaviour patterns.

Some of the ways in which data and advanced analytics may transform organisational processes include smart procurement, next generation products development, marketing and campaigns, dynamic pricing and yield management, store operations, warehouse operations and returns management. Strategising marketing campaigns is another case of data coupled with analytics. Growth hacking methods are more of a lean start-up approach, hence not necessarily a preferred option. Insight gained from campaigns will help decide where to cut back and where to increase spending for directing more traffic to the site. This is made possible and more effective with BD as access to fine granular data becomes readily available. According to Schmarzo (2013), data equates insight while BD is foresight, paving the way to better predictions.

Decision-making, according to all participants, is one of the most important applications of data insight within the organisation. Strategic decision-makers decide the fate of business initiatives by analysing data. This ramifies further to what decisions will be carried out at departmental levels by operational managers. Decisions are categorised as operational, tactical and strategic. According to interview participants in the Merchandising department, budgetary allocations in terms of figures are provided to Merchandising. They then create a shopping list from data insights, indicating what items are considered 'best sellable'. The department is thus able to buy smart. Decisions are based on sales data, transaction data, customer behaviour and market basket data. It is clear that a lack of substantial historical data, according to interview participants and the fact that Spree is a start-up, seem to hamper business initiatives as participants may take a lukewarm approach to using what is available.

Research sub-question 1.3: What are the policies and strategies for leveraging Big Data?

Interview participants mention that there are no documented policies or strategies for BD curation.

Research sub-question 1.4: What information does business want from data?

Participants mentioned that Spree just evolved from being an incubation project into a live store, which means the current historical data available is not yet sufficient for decision-making. Business Management on the other hand, mentions that since systems are built in house, Spree has the opportunity to gather any data collectable. As a result, business collects data to enable the organisation to contact customers and to market more effectively. As stated by a participant, some of the data the organisation is collecting may help to

address the departmental needs list, KPI-oriented data, competition germane data, enriched data with relevant interconnecting fields, and data leading to unique selling points. This data can be categorised as business micro and macro data, which, though hind-sighted, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business.

Research sub-question 1.5: What kind of data is being curated as part of data?

Through customers visiting the Spree site, it becomes possible to collect all data needed for analysis on transactional and transaction-related data. There is a divide among the Spree participants with regard to curating BD. The participants all agree that data forms the basis of measuring metrics and KPIs relative to business processes. Data collected include sales granular data, transaction data, customer data, products information and session information. Data curated include subscriptions, total visits to site, unique visits, first time visits and repeat visits, among others.

Supply Chain and Customer Care use historical data to predict lead times and evaluate patterns of demand for the different categories. The level and granularity of analysis depends on the level of curation in terms of granularity.

5.4.2 Answering Research Question Two

Research question 2: How can BD be leveraged in a media organisation to gain competitive advantage?

To answer research question 2, two sub-questions were asked. These sub-questions are indicated in Table 5.8. The table provides a high-level overview of sub-questions and summarised answers of findings from various interviews. The aim of question two is to establish how the organisation may utilise data, and more specifically BD, to its advantage as justification for the never-ending data gathering trends. Furthermore, how can this curation be implemented in a way which is beneficial and appropriate in promotion of a particular business model?

Table 0.8: Summary of answers to Research Question Two sub-questions

Number	Sub-question	Summary of findings
2.1	How can BD be utilised to gain a competitive advantage?	Strategise curation to meet data needs both for short-term projects and long-term organisational data needs.
2.2	How can a business implement BD curation?	Plan forward using industry best standards and view what other organisations are doing and have in place. Research from universities could also provide a rich source of information.

A review of the answers to research sub-questions one and two and literature brought to light the fact that traditional data curation approaches are inadequate to meet today's data processing and curation needs. This has prompted organisations to look beyond traditional methods.

With the majority of dynamic organisations deploying NoSQL databases, some have gone to the extent of Polyglot persistence or hybrid persistence to leverage the full benefits of the different systems brought together to synergise.

Research sub-question 2.1: How can BD be utilised to gain a competitive advantage?

All participants appreciate BD as an important enterprise asset. This is reflected in the gradual shift from gut-based decision-making to data-driven decision-making. The shift in mind-set and mandated reliance on data by especially Business Management, the Technical department and other decision makers, has created a common neutral entity for all data curators. The organisation has come to realise the significance of data as an enterprise asset among all other assets, except participants' contribution or view of BD seem clouded by the limited definition and unawareness of BD in its totality and ability to transform the business. According to literature, BD means different things to different people based on interest, but to Spree, the lack of understanding of BD is more detrimental and more of a loss to the organisation than can be comprehended.

Research sub-question 2.2: How can a business implement BD curation?

From a technical and implementation perspective, participants were of the notion that Spree may be collecting data, but current data size does not warrant data being called BD. One participant mentioned that there are no terabytes of data and also no Hadoop clusters in their environments—this implies that participants might be confusing actual data with tools. Participants mentioned that currently there are no documented plans of BD integration into the business model.

Participants did not dwell on BD implementation; they rather focused on the benefits of curating BD as per previous work environments and what knowledge they have gathered in the process relating to BD. They indicated that Spree is a start-up and does not yet have or curate BD, but that BD is beneficial to an organisation. From literature analysis it was inferred that many dynamic organisations realising the potential of BD, moved on to canvas a host of tools to enable them to deal with the data deluge. Some of these organisations are Google, Yahoo, Diggs, Facebook and Amazon. Almost all implementation started off with some form

of relational database to store persist data and then extending to NoSQL implementation of data storage. There are different ways of implementation but most current systems deploy a combination of relational and non-relational approaches.

5.5 Data curation guidelines

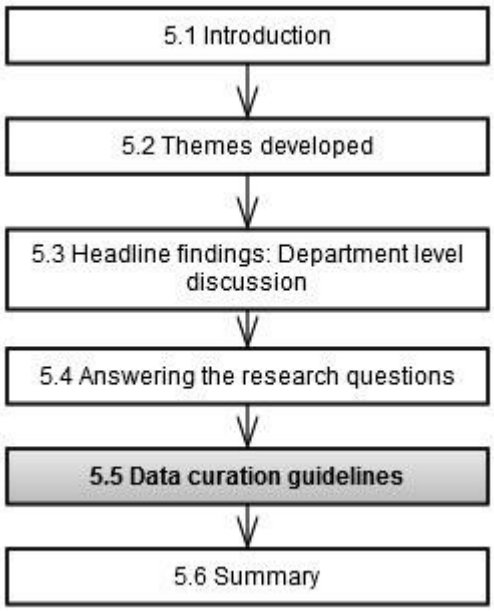


Figure 0.5: Chapter layout – Data curation guidelines

Section 2.3 brings across the fact that BD means different things to different people, especially as analysis of large and ever-increasing datasets become routine, which is necessitated and validated by an organisation's need for information to meet different transformation needs. The definition of BD will continuously shift (see Section 2.3), but what remains constant is that success at the leading edge and competitiveness will be achieved by curators and developers who look beyond off-the-shelf techniques for total algorithms, models and hardware resources available to drive the transformation needs. Data has always been the fuel that powers insightful business thinking (See Sections 2.3 and 5.1), as leading organisations leverage the opportunities of data and insight to surpass competitors in the market space using insights from data trends to uncover user needs. This is essential as business transformation necessitates a data curation journey which is a systematic process that spans many phases from business monitoring to business metamorphosis (see Section 2.13.1 for phases of BD Business Model Maturation Index).

The curatorial guidelines will facilitate start-up companies and establish organisations in line with e-commerce to launch their BD journey. Figure 5.7 depicts a representation of the proposed business curation guidelines.

This research uses the BD Business Model Maturation Index as a guide to ascertain where an organisation may be at in the data curation journey.

The model assumes that the majority of organisations seeking to embark on this journey have their BI traditional processes in place. This may also imply that they may have started collecting transaction data for future analysis. More or less 95 per cent of companies are at the business monitoring phase as they gain insight from data mainly using relational tools to run day-to-day business activities. Data curation processes at the business monitoring phase are based on well-known BI processes (Schmarzo, 2013) such as scheduled and *ad hoc* reporting as sources of insight for decision-making.

The BD Business Model Maturity Index proposed by Schmarzo (2013) is aimed at giving the seller the idea to take advantage of Big Data and advanced analytics to power value creation. The BD Maturity Model details what phases an organisation may pass through in the BD curation journey. The BD Maturity Model stands out as one of only a few approaches to BD integration, but the model does not provide a concise curatorial guide for initiating the BD curation journey. The BD Maturity Model does not touch much on the business model and the fact that identified or tested strategies for improving business processes as way of monetising insights should be integrated into the business model as input and to promote iteration as a way of re-engineering the business model. One of the advantages of curating BD and drawing insights from BD is the identification of trends to integrate user needs to improve customer engagement. The user needs emanate from data interconnections, patterns and identified overlaps which improve data integrity and trust.

The proposed guidelines in Table 6.3, also represented in Figure 5.7, may furnish the curator with process knowledge and provide a knowledge base as direction for data curation. The BD journey requires an initial analysis of the organisation's current status of data, which in part requires an exhaustive review of the business model as this defines the business processes to ascertain how the organisation creates value or seeks to create value. The business model from a business perspective is essential to ensure that the business operates with insight as it forms the basis of value proposition to the customers and validates why the business is the single one entity that delivers that differentiated service. The business model canvas (Figure 5.7 and Section 2.6) helps to collect relevant information about the business as it describes the business model from a theoretical standpoint, dividing the model into four components as mentioned in Section 2.6. This section, for practical

purposes and ease of comprehending the model, focuses on two primary components: customers and value proposition. The customer and value proposition components provide and encompass the essence of the model as business is about delivery service to the prospective customer at a cost worthy of the value proposed and supported by all other business elements. The model helps the seller create and articulate the contents of the individual components of the business model which serve first as input for the business plan and secondly as an input for other business documents.

Figure 0.6: Business model canvas

(Source: Canning, 2014)

A business needs data to monitor itself and direct business initiatives. The data needed to make informed decisions comes in different forms, as mentioned in Section 2.7. The data forms include structured and unstructured data. To capture data that drives business transformation requires a good knowledge of business elements (themes) as depicted in Figure 5.7. The diagram elucidates the creation of a business model, leading to the creation of a data curation strategy from an enterprise culture perspective where all employees and departments adopt the process as culture-based and inter-disciplined. It starts with the members of the organisation being fully aware of what the organisation represents, how it generates revenue and delivers value to the prospective customer. These are inputs that feed directly into the business model canvas (Figure 5.6), which is a subset of the process.

At the heart of the proposed data curation guidelines (Figure 5.7 and Table 6.3) is a brief statement as to who or what the organisation represents in terms of value proposition to the customers and stakeholders. The statement can further be expounded using the business model canvas. The information feeds directly into the business model as value propositions upon which the organisation is hinged. It is important to define and document value proposition as part of the business model and business plan. The process is followed by step 3, which is to create a list of data challenges that serve as the needs list of the organisation.

Examples of needs include questions such as “what technology assets are required to leverage BD?”, or “what technology assets are required to gain insight from customer data as part of analytics and data curation?” Step 3 (Table 6.3) and preceding steps provide data and insight to develop and implement the business model, which determines the service and/or what technology and information assets are required to deliver services.

Completion of the initial analysis, which includes a statement of all data challenges forming the organisation’s data needs list and the establishment of different department-level data models which identify all needed data, is followed by a unified business model which identifies on a global level all data needed for the organisation. The organisation may then continue to evaluate whether to outsource this curation process or engineer curation in-house. Arriving at this decision depends on management and data skills available in-house. Figure 5.7 delineates designing a curation strategy as part of the process of data curation. It is worth mentioning that these processes are repeatable as a data curation process—it is a lifelong process, generating insights into the needs of the business and re-directing information back into the system for progressive maturation until business maturation or metamorphosis has been reached, as indicated in Figure 5.9.

It is essential to note that the curation guidelines could work well with Schmarzo’s (2013) Maturation Index to identify where an organisation is currently positioned in the curation journey. For brevity, the stipulated guidelines do not integrate possible curation tools and technologies that could be deployed for the data curation process. As proposed by Schmarzo (2013), a business tool such as the business model canvas could aid in simplifying laying out the business model (see Figure 5.7). It provides boxes for listing key activities, partners, value propositions, customers and all other components of the business model.

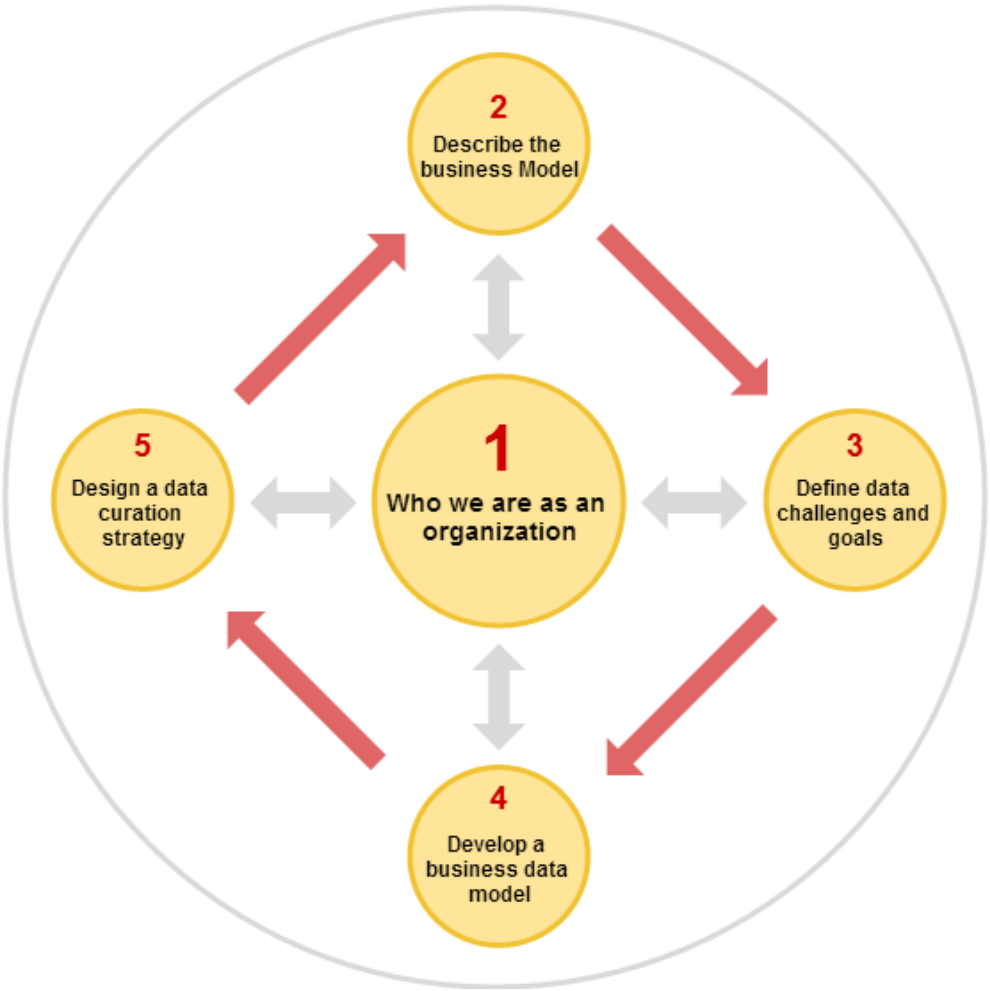


Figure 0.7: Enterprise-wide data curation guidelines

On a more detailed level, the process of brainstorming and designing a curation platform requires an organisation to be able to establish the organisation's need for data (data needs list). It is also essential to know how the data will be utilised, and in which format, for ease of use. Other factors that may be considered include the rate of data progression (increasing volumes), rate of change of business elements, rate of data ingestion, what decisions to be made, points of contact to collect data, and data diversity from a granular departmental level using data models. Insight into these factors will culminate in business model re-engineering through a process of unification. This is referred to as bottom-up approach, as depicted in Figure 5.8. The aim is to ultimately create a value proposition based on insight into business elements that feed into the business model. The business elements represent the identified themes which form the business elements that spur the organisation to curate data for insight (i.e. analysis). These elements are divided into primary business elements and supporting business elements. The technology assets, which inform the entire technology-oriented curation platform, depend on initial investments and technologies the organisation already has in place. The challenge is how to integrate augmenting tools to help capture and manage relevant data that was previously impossible to leverage.

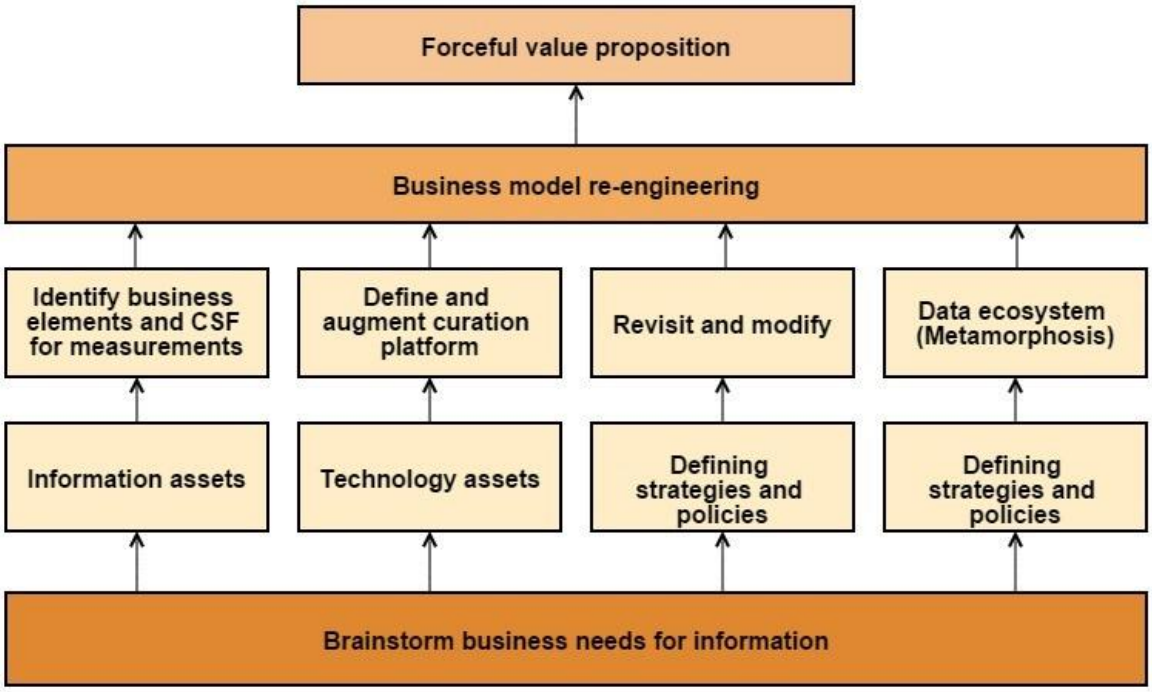


Figure 0.8: Forceful value proposition

The alternative is a top-down approach that commences by stating organisational goals and branching into subsequent departments to ascertain and identify the information needs, data assets, curation strategies and possible alternative data revenues that may originate from data curation and data management. These form the basis of the business model and compelling value proposition or development phase of the business life-cycle based on the needs of customers detectable in data patterns and trends. This is important when answering business questions as part of compiling the organisational data needs list. The curation design process requires a fair knowledge of Decision Theory, Solution Engineering and the business model to deal with enterprise data challenges through a well-defined BD strategy that launches a guided data curation journey as described in Sections 2.13 and 2.13.1. This will ensure the implementer has adequate knowledge and has done enough groundwork on the data curation process.

This research seeks to bring across the significance of BD curation and integration into the business model for information advantage. It is important to note that the technology aspect of BD curation is not fully covered in this research though the researcher gave an overview of polyglot persistence and the different underlying components of a hybrid persistence platform currently deployed by cutting edge organisations such as Google and Amazon. The essence of BD curation is insight generation from BD for differentiation, in other words, using diverse data sources to support service delivery, for example using a 360 view of the customer to generate insight to guide value proposition.

The principle is to collect data or generate insight from streaming or collected data and actioning the insight. It is also essential to measure outcomes through data metrics to ensure that the curation process is meeting set goals or the process is yielding data that can be used as a source of insight for questions.

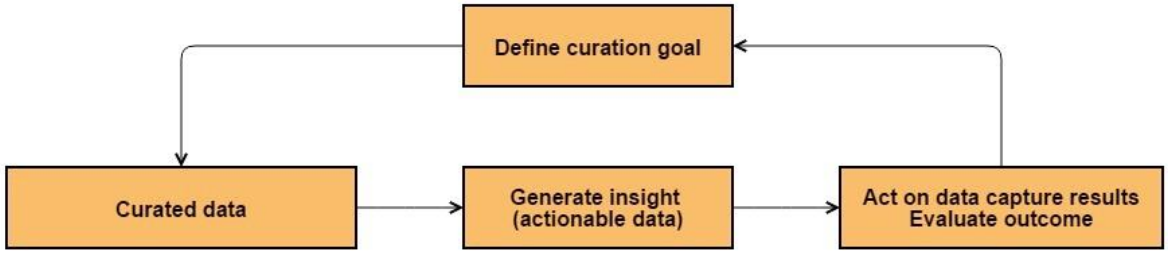


Figure 0.9: Ingest and evaluate data against goals

The BD journey should commence with brainstorm sessions that aid data curators to identify game changing data forms needed and sections of the organisation that data may play a role in attaining the needed data advantage. The business model canvas provides the opportunity to witness, identify, strategise, design, curate, direct implementation and measure success based on identified metrics. The essence of the business canvas is to establish desirability for the organisation's proposed value items in the form of products, services or experiences. With data analysis comes the generation of insights that lead to patterns identification, interconnections and overlaps; these culminate in the identification of trends that could also expose user needs.

The two major components of Figure 5.10 represent the business elements and the technology assets.

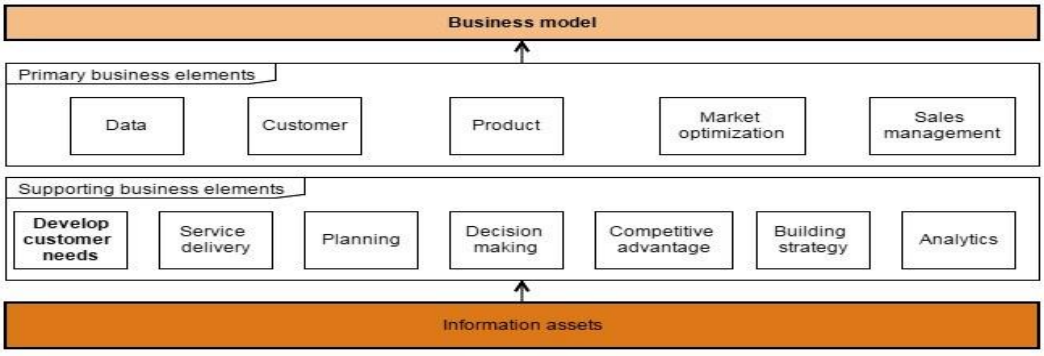


Figure 0.10: Information assets of the organisation

The business elements, which include data, customer, product, market optimisation and sales management, are the primary business drivers that spur the organisation into gaining knowledge insights. These elements are managed and promoted by the supporting elements which eventually feed into the business model. The existing relational technologies and disruptive non-relational curation technologies represent the complete curation platform, significant to collect and manage BD as indicated in Figure 5.11.

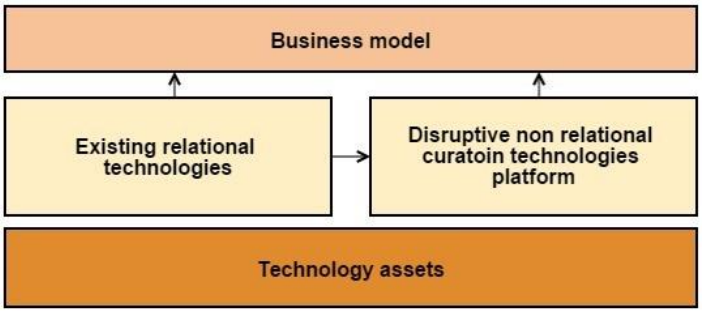


Figure 0.11: Technology assets of the organisation

Defined strategies and policies are not discussed because interview participants indicated that they are unaware of any documented policies or strategies.

5.6 Summary

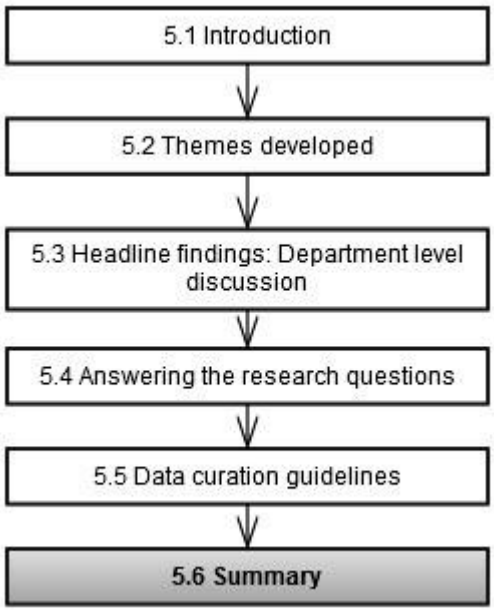


Figure 0.12: Chapter layout – Summary overview of discussion

Every organisation needs differentiating factors to transform as a business, remain relevant and be profitable. These needs extend to information in the form of curated data with footprints of the current business. From this information (contextualised data), the business may be steered into profitable situations and optimised processes using insights drawn from (analysing) data. Basch (2012) states that Google as an organisation realised search engines are not just about finding search results with links to click on, but rather a planned intention on the side of an organisation to satisfy the customer's need for information. According to Basch (2012), this was an obvious concept to Inktomi engineers back in the year 2000 during the declining phases of Inktomi, except selling this idea to executive management was a challenging task. Inktomi was the number one search engine then in the world, contracted by Yahoo to deliver search results.

Interview participants from the Social Media and Marketing departments are unaware of Spree curating BD and do not know what BD is. The Business Management department is strongly opinionated about collecting many varied forms of data such as customer data, transaction data, transaction-related data, web stream data in the form of metadata, and customer clickstream. This forms the basis of BD as asserted by Rele (2012). The participants' unawareness implies the lack of a well-documented, widely communicated data model and business model to communicate on a departmental and organisational level the data needs of each department in alignment with the entire organisation's data needs.

The lack of a data model within Spree creates data or communication gaps as readily available data that may serve other departments, remains unknown to curators who need data for decision-making. This leads to the loss of worthwhile revenue as the organisation is unable to take advantage of the benefits of curated data. This is affirmed by Khatri and Brown (2010:4), stating that "many organisations do not know what data they have". All participants acknowledged data as important. Many participants mentioned that data is a technological power house. Data is everything; it powers business for effective decision-making, segmentation and customer targeting, among others. Participants however asserted that Spree's business is not functioning in the realm of BD yet. This is based on a specific (narrow) definition of BD.

Participants noted that the data available to Spree's decision-makers is one day late, thus implying the insufficiency of data implementation using current BI tools for alerts and reporting. Curating BD is about having an information advantage, which is the cultivation of skills, mind-set, processes and technologies to use information to operate more efficiently, increase customer loyalty, grow market share and create business opportunities that were not previously possible (Adduci *et al*., 2011). For analytics and a reactive approach to dealing with actionable data and implementation of identified insights, it is crucial to deploy real-time

data to speed up or slow down time in order to gain value from data, which in turn depends on measures taken to avail real-time data.

Participants were uncertain about the data status of BD. To ascertain their standpoints, participants were asked if Spree was curating BD, that is, does Spree have BD and is Spree curating large volumes of data? The responses as shown in Annexure A, Table 7.9, indicate that participants were uncertain about the organisation's BD status. This lack of consensus is a major concern. The same pattern is also reflected in many other organisations. As mentioned by Ross, Beath and Quaadgras (2013), the main reason for investments failing to pay off, is that companies do not effectively process the data they already have. They do not know how to manage or analyse the data in ways that enhance their understanding. This holds back the organisation from leveraging the benefits of BD. When participants were asked about the curation of BD at Spree, many different responses were given, with about 28 per cent confirming that Spree does have data. Many were unaware of BD, while the others disqualified current curated data as BD based on the premise that the data is not colossal. Some participants mentioned that BD requires special tools, and that the data currently available at Spree fits on Excel spreadsheets, which implies that extra tools are not needed.

Chapter Five presents a discussion and analysis of the research findings. The discussion is divided into four parts and elaborated on accordingly. Many findings emerged which where categorised and eventually resolved into themes. The themes expound on entities and factors on a high level to ascertain interaction at departmental level and how Spree could be spurred into curating BD for a competitive advantage. Themes include data as an asset, profiling the customer, developing and satisfying customer needs, planning, decision-making, competitive advantage, marketing optimisation, service delivery, sales management, analytics of data, building of strategies and business model re-engineering. Insights from literature analysis and various definitions in practice conclude that Spree is curating BD albeit not on a big scale as stated by participants. Furthermore, as visits to the Spree web site increase, data sizes are bound to grow. This proves to be confirmation of participants who commented that BD is an extension of data. The emerged themes are elaborated on at a general and departmental level with respect to how the interaction takes place.

The chapter further discusses the research questions to evaluate the tangibility of contribution to knowledge. The next chapter discusses the conclusion and recommendations of the study.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

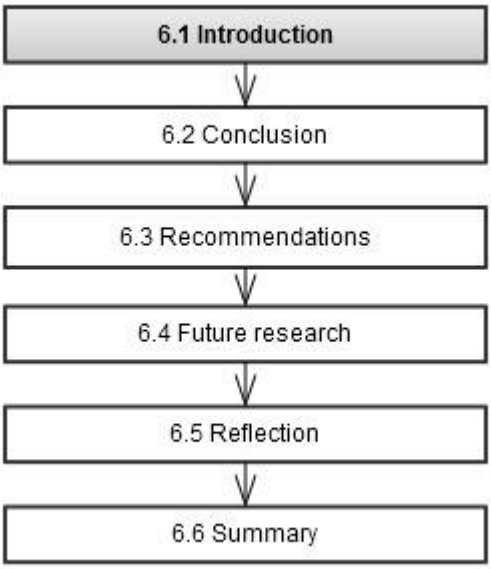


Figure 6.1: Chapter Six layout – Conclusion and recommendations

6.1 Introduction

Big Data has become an imperative business asset taking centre stage in organisations that seek to stand out as an industry strong competitor with data as a driver. With current data trends of data influx and continuous data collection (as described in Section 1.3) prompted by advancing technology and organisational need for information, there are key factors affecting the data that businesses collect. These factors include volume, velocity, variability and accuracy (Sathi, 2012; Schmarzo, 2013; Zikopoulos, *et al*., 2013; Mohanty *et al*., 2013). Both structured and unstructured data is collected. The unstructured data varies based on the type of business and the business model implemented. Organisations have curated structured data using relational data tools, though over time the problem of scalability and performance became eminent as data sizes increased. Furthermore, as business realised the potential of unstructured data as a source of insight, this same lack of structure made it difficult for any practical use. As a result, business had to rely on technologies that broke away from the mundane relational tools.

The scope of data needed by organisations to steer into profitability needs new data management approaches. Some organisations have addressed the need for new curational approaches through an ecosystem of tools called Hadoop and its cluster of tools, but the implementation vary in terms of design of the business model from organisation to organisation.

While this has been accomplished by some leading data companies, many still remain in the dark, seeking ways and means to learn and implement what these leading technology companies know and use that give them the needed advantage to continuously lead in their industry. This advantage helps to continuously improve their business models.

This research addressed the following **problem statement**:

Companies find it difficult to leverage the opportunities Big Data offers them in terms of monetising the content of curated data.

To address the above research problem, two primary **research questions** were asked:

i) What are the factors affecting business to leverage Big Data for competitive advantage?
ii) How can Big Data be leveraged in a media organisation for business to gain a competitive advantage?

The research questions have sub-questions to help drill deeper into canvassing the true and implied meanings in order to address the ontology and epistemological demands of this research, thereby creating a new form of knowledge usable by both start-up organisations and established organisations as they seek to launch the organisational lifelong BD journey.

6.2 Conclusion

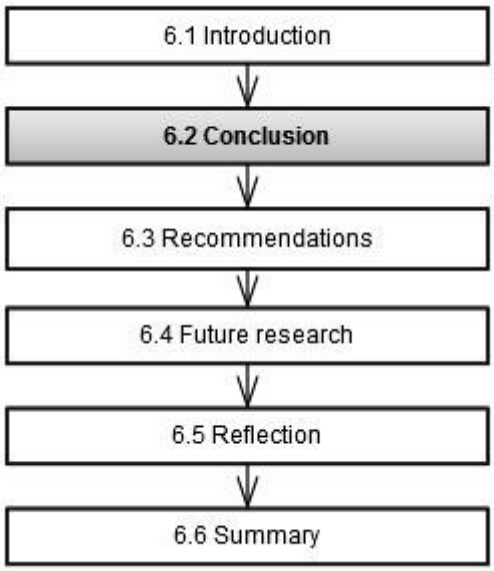


Figure 6.2: Chapter layout - Conclusion

The aim of the research was to explore how BD can be used to utilise opportunities such as leveraging insights to gain information advantage, offered to the organisation. A further aim was to explore how BD can create competitive advantages such as monetising the content of curated data.

The high level objective of this research is to propose BD curation guidelines (see Section 5.5), adoptable by a digital multimedia organisation. The created guidelines propose curation procedures aimed at simplifying, directing and improving the curation process, ultimately for better gains where competitive advantage is a concern. Malm (2013) mentions that curating BD presents the data curator with benefits that include a competitive edge, presenting business transformation data to the Board, and driving innovative products especially for start-ups and also establish organisations. This notion is shared by Manyika *et al.* (2011) as stated in Section 1.2. In carrying out the research, it is worth noting that a *technology platform independent* approach has been used.

Through literature analysis (secondary data), organisational source data and interviews (primary data), twelve business differentiating factors emerged which satisfied research question one. These are the factors that affect business to leverage BD to gain a competitive advantage in the market space and market place. Factors include data as an asset, profiling the customer, developing and satisfying customer needs, planning, decision-making, competitive advantage, marketing optimisation, service delivery, sales management, analytics of data, building of strategies and business model re-engineering. Sections 5.2.1 to 5.2.12 deliberate further on these concepts. Other contributing concepts worth mentioning include Decision Theory and Solution Engineering.

Though some of the business entities, for example decision-making and competitive advantage, might not appear directly in an entity model as influencers to collect data, they play equally significant roles. They have all been identified as being salient to the enterprise data user in different ways. In some cases, these business entities form the core elements to be deliberated on when considering why and how things are to be done especially relative to Solution Engineering and decision-making through the key elements of Decision Theory.

There are sub-research questions for research question one which focuses on business efforts to curate data, business's impression of data, strategies and policies, how the data curated is to be used, and what kinds of data the organisation is seeking to curate to obtain a competitive edge over competitors. For brevity and ease of assimilation, this is tabled in Table 6.1.

Table 6.1: Research Question One sub-questions and findings

Pre-interview and sub-question	Finding
Pre-interview	All interview participants acknowledged that data is an important enterprise asset. Nine magazines are affiliated to Spree with more to join soon. The nine magazine companies have massive data that is inaccessible to Spree. There is valuable (BD) data that is outside the reach of Spree. Data curation from a curator standpoint divides into technical (IT) and business curatorial processes.
Sub-question 1.1	Spree is a start-up and not yet functioning within the realm of BD. There is a need for real-time data (low latency data) but current systems and implementations only provision data one day late. Operational data users are uncertain about Spree curating BD. There is a gap in communication imposed by the lack of a well-documented and widely shared data model between data curators and data users. Participants acknowledged that BD is an extension of data. Data curators are operating with a limited definition of BD.
Sub-question 1.2	Business Management identifies what data to log per department's need for information. Spree has most of its data in silos. Spree logs varied data about business entities using the Magento system, Google Analytics and On the Dot (OTD). Business intelligence (BI) is in the process of building and architecting data structures to centralise data for analysis to gain insight into curated data.
Sub-question 1.3	Not much is said about this as there are no documented strategies and policies.
Sub-question 1.4	A variety of data (especially data needed for operations) is being collected as part of data curation which includes transactional and transaction-related data. Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data. There are no documented policies, strategies, frameworks or a curation model except a Magento ERD diagram. Business aims to obtain all necessary business monitoring data from centralisation as these data are pre-provisioned and in curation to answer relevant business questions.
Sub-question 1.5	Business data, though in hind sight, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business. Spree is collecting weblogs (clickstream data) and BD is an extension of data. Data is spread across multiple servers outside the perimeters of the business, making the data inaccessible for analysis. The way data is stored across the different servers may increase the difficulty in bringing the data together for analysis. Manual data handling may make data prone to errors, thereby compromising the quality of data for decision-making. There is a lack of consensus as to whether Spree has BD and is curating BD.

Addressing research question two took on a more technology-oriented approach by asking the question: "How can Big Data be leveraged in a media organisation for business to gain a competitive advantage?" This research question sought to uncover how the research entity is currently curating data or BD (if the two are thought of as different data forms).

The organisation's data process exposed a curation process that hinged firmly and only on relational tools. Schmarzo's (2013) BD maturation index model depicted this as being part of the business monitoring phase of the BD curation journey. This does not imply that Spree has taken a wrong turn. Rather, it is building data structures necessary to launch a successful BD journey. The model is extensively discussed in Chapter 2. Steps to reach business metamorphosis are indicated in the maturation index and further clarified from a ground-up level in the BD curation guidelines described in Section 5.5.

Participants were unaware of BD and how it exactly affected the business model. Those who were aware circumvented the topic and thought of its discussion as being a little too ambitious, reason being that the data sizes currently at Spree are not as colossal as depicted by some organisations leveraging BD, i.e. Facebook, Yahoo and Amazon. Yet business models vary, and application of BD to effect change within the enterprise vary. Though underlying curational processes are similar, there are many justifying definitions that imply that Spree has BD and is curating BD; furthermore the participants defined BD very limited making them lose out on curated data that could already have been leveraged for differentiation. In Section 4.2.3.6, the Magento framework is stated as being the persistence framework used by the Spree e-commerce store. There are many Magento extensions available to a merchant using the Magento Framework—Windsor Circle, Bronto and Springbot are available extensions for segmentation and targeting. A/B testing can be done with Optimise while Lexity and RJ Metrics are optimised for analytics. This implies that Spree could already be differentiating with already curated data to generate insight. Responses to research sub-question two are tabled below in Table 6.2.

Table 6.2: Research Question Two sub-questions and findings

Sub-question	Findings
Sub-question 2.1	There are no plans for BD integration as of yet; Spree is not there yet.
Sub-question 2.2	Curate BD using technologies that hinge on promoting competitiveness through new technologies for a cutting edge advantage.

A further aim of this research was to address leveraging BD for competitive advantage first by ascertaining which factors affect business to leverage BD, and second how BD can be leveraged in a media organisation to gain a competitive edge.

Research question one was addressed by identifying the factors that prompt Spree to curate data, which include data as an asset, profiling the customer, developing and satisfying customer needs, planning, decision-making, competitive advantage, marketing optimisation, service delivery, sales management, analytics of data, building of strategies and business model re-engineering.

Research question two was addressed by establishing the insufficiency of the relational data warehouse approach to current curation processes, and addressing this through the new polyglot persistence approach that deploys the strengths of the different NoSQL databases in addition to the use of relational databases in the form of a hybrid persistence layer.

The answer to question two was furthermore augmented by dividing data curation into short-term and long-term approaches using Solution Engineering and the BD strategy document to describe the problem and solutions state. This was done to have a clear understanding of the expected value of perfect information. Data curation as a problem (opportunity) requires knowledge of the structure of the domain to create the necessary steps to reach the solution state (planning and strategy) and eventual integration into the business model upon success. These all form part of proposing BD curation guidelines and recommendations that could be applied systematically through Solution Engineering and Decision Theory in a data curation environment to harness the opportunities BD offers in a more scientific way.

6.3 Recommendations

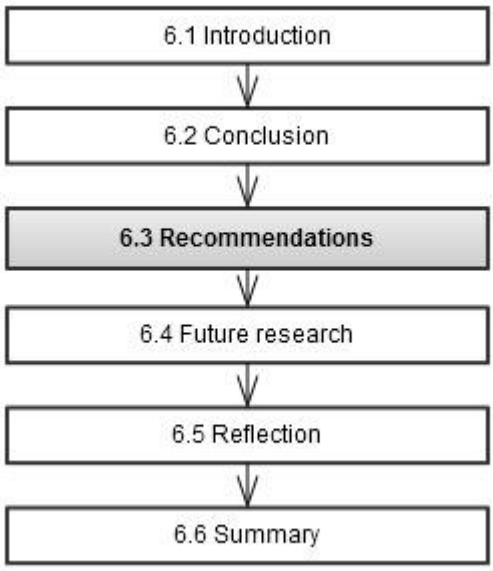


Figure 6.3: Chapter layout - Recommendations

This section focuses on a summary of BD curation guidelines and recommendations as described in Section 5.5. Table 6.3 presents the guidelines.

Table 6.3: Big Data curation guidelines

Step	Activity	Description	Objective
1	Define concisely what the organisation does as a value proposition to a prospective customer.	Who are we as an organisation?	Bring out the value points by developing a business model.
2	Scheme out a fully-fledged business model with emphasis on data curation using Figure 5.4 in Chapter 5.	A clear business model.	Ascertain what factors will affect the business to curate BD.
3	Identify all data challenges and how data will be used. Based on the model of the business, is the organisation system phasing clients and having more insertions (data input), or rather more retrieval, or both? This is crucial for what technology architectures to deploy.	Define data challenges and data goals in modules by functions or department.	Identify at department level what data to curation to a fine grained level.
4	Create department or function level data models focusing on the need for information and unify into a business data model (bottom-up approach); the alternative being a top-down approach, could also be used. Identify all data users and model use cases for users and how data will be accessed, saved and used to generated value from departmental level up to global organisational level with the help or supervision of a data office or data architect.	Department level data model, business data model.	Map out data models.
5	Create a data curation strategy. Using Solution Engineering for both short-term and long-term curation projects.	Create data architecture which factors in all Information collected from previous steps. Note that this is repeatable. At this stage the organisation may decide to outsource or build in-house.	Decision-making and technology dependant phase; outsource curation or develop in-house.

The following section details recommendations for the curation of BD and initiation of the BD journey. Initiating a BD curation journey that may lead possibly to insight monetisation and big business metamorphosis requires a well thought out curation strategy with brainstorming sessions that internalise data and predict where the organisation is headed.

a) Internalise curated data prior to BD journey

Companies with a culture of evidence-based decision-making from guided data curation ensure that all decision makers have performance data readily available. Knowledge of what data is based on being fully aware of which data exists, how to gain access to the data, and

how to leverage its opportunities. Using real-time data and analytics for decision-making will promote a culture of evidence-based decision-making. This will improve the time to value creation, leading to actionable data that may be used. The knowledge of internalised data has to be communicated widely across the organisation and possibly incentivised where necessary to promote the use and dependence of the generated insight as an organisational culture.

b) Create a data culture

Creating and fostering a culture of data-centred decision-making will improve the organisation's chances of success as employees are encouraged to see data as assets such as human capital, infrastructure or skills. The benefits of curating data become even clearer as employees become mandated to participant largely due to executive support. The integration of data to drive institutionalisation will drive a cultural shift to promote data as a decision-making tool.

c) Enterprise data model

Data curation should be aligned to organisational goals and business initiatives, and have their success measured with delineated critical success factors. Curation should be sub-divided into short-term or long-term projects based on intensity, depth and perceived duration of the project. Curational processes should bear marked and hierarchical significance from departmental level to organisational level using a data model (department level) and an enterprise data model (global) as these capture the information needs, tools, process and practices at a given level. An enterprise data model should be developed and maintained jointly by data stakeholders working together in data curatorship teams per subject area and significance, and coordinated by the enterprise data architect. This should be revised at predetermined intervals and changes or extensions reviewed, approved and adopted.

d) Strategise data curation with a pre-determined goal as a guide

Curating BD can be overwhelming when data sizes increase. Having a data goal prior to any curation process will save the organisation much trouble. Strategising curation to be a short-term or long-term project with concisely defined goals and curation steps (Section 2.13.2) will foster better use of data. A short-term curation project could be a small defined data curation task aimed at collecting data for a particular defined goal. For example, data to improve customer experience on site as a way of improving customer engagement (see Section 2.3.1 - content curation to improve customer experience). Short-term curation could be initiated for projects to study or analyse market baskets, shopping carts and cross selling, among others.

Outcomes of these curational projects could be fed into models for other projects to furnish the organisation with the needed insights.

e) Act on actionable data immediately

Actionable data for business, and especially social media, must be acted upon immediately when patterns and trends emerge. As an enterprise it is important to be willing to put to use insights gained from curation as the process becomes futile when collected data does not empower the business to distinguish itself as advantaged.

f) Improve curatorial processes by identifying data source contact points

Identifying data sources requires brainstorming business use cases to identify all possible customer contact points. Every effort must be made to collect data about business elements such as customers and customer behaviour to improve customer experience and engage them to improve user experience and sales.

g) Polyglot persistence for optimal provision

Business systems should use optimised features of tools and databases in a mixed way to leverage the strengths of different tools. Relational databases are great for simple data curation needs, however where demanded, scalability and performance along other curation factors become a challenge, data curators or proponents should quickly look beyond normal curation processes for the benefits thereof.

h) Invest in Big Data personnel

The organisation should identify, develop retain and invest in BD personnel and train current staff to gain the needed BD skills for data curation.

I) Design the data curation process

Architecting a curation technology is a sure way to gain the benefits of data curation as the process encapsulates the business model and draws on the strength of many different systems to meet the data needs of the organisation. See Annexure A Table 7.5 for a sample polyglot persistence architecture that draws on the strength of a relational database and NoSQL database tool (Hadoop) to process data that may significantly furnish the organisation with cost savings, performance gains and better processing.

j) Alter business model

As an organisation, initiating the BD journey implies that the organisation may discover many different insights which require changing the business model. The business model in some context is conceptually referred to as business processes, implying that as business processes are optimised, the competitive strategy is changing and these changes should reflect in business documents and the business model.

k) Document and communication data models across organisation

Documenting business data models and communicating the model across the entire department helps the organisation create a culture of dependence on data. Dependence on data will promote data interaction which will end up with abridged high-level versions of data models in a non-technical language comprehensible especially by non-technical employees. Needing to use data, this addresses the communication gap identified among findings in chapter 4.

l) Prioritise business initiatives and build strategy

A start-up organisation which develops curation processes and chooses technologies with respect to its business model has the advantage of being able to weigh the many different options available. The organisation is at a point where it can try out alternatives, that is, make mistakes, learning and quickly moving on. As the organisation matures, it is forced to choose among a growing number of analytics tools while being faced with having to deal with the lack of adequately skilled personnel in the industry.

Spree is confused with the state of BD largely due to unawareness, lack of knowledge as to applicability, and how limited or incomplete BD is defined. This is exacerbated by the lack of a Spree BD strategic document with enterprise-wide and department-wide data models. Spree should take a position that they may not have large data sizes but they are collecting data continuously according to Gualtieri's (2012) BD definition as in Section 5.2.1. Furthermore, the Magento framework allows merchants to start leveraging BD from inception. This is affirmed by the many BD analysis tools (extensions) available to sellers on the platform relative to their business model. Accepting that available data at Spree may not be colossal, but the context in which Spree is operating in and the need for detailed information of customers, products and customer behaviour in a 360 degrees perspective, BD should become part of the Spree strategic plans to place Spree among information advantaged organisations (see Section 2.6.1).

6.4 Future research

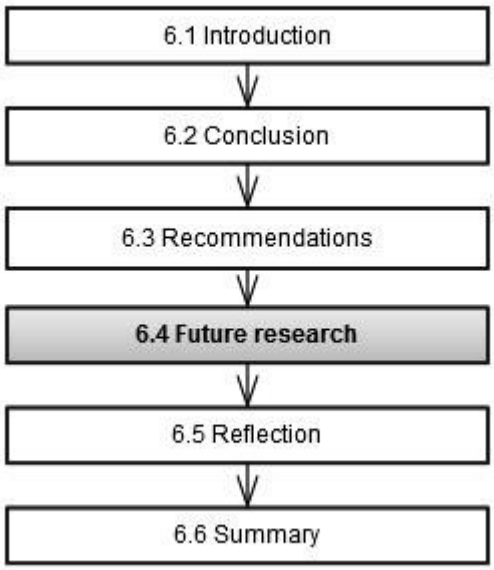


Figure 6.4: Chapter layout – Future research

6.4.1 Polyglot persistence and Business Intelligence

NoSQL databases are schema-less curation platforms for BD using a programing model (Hadoop and MapReduce) to leverage the opportunities of BD as a parallel process. Its advantage over current relational approaches to data curation makes it an industry-preferred approach to data curation, yet its initial setup and implementation could be a challenge to many start-ups and upcoming business. The benefits span deployment over commodity hardware, fault tolerance and the absence of the need for a predefined schema as in relational databases. For businesses that are already on the relational platform, the integration of Hadoop and MapReduce could yield many benefits with many organisations reporting huge savings and operational process improvements as the data curator has the opportunity to hone in on the strengths of different databases to synergise output for their production environments. It is important to research the use of multiple different databases to furnish organisations, especially start-ups, with the needed information required to materialise polyglot persistence platforms.

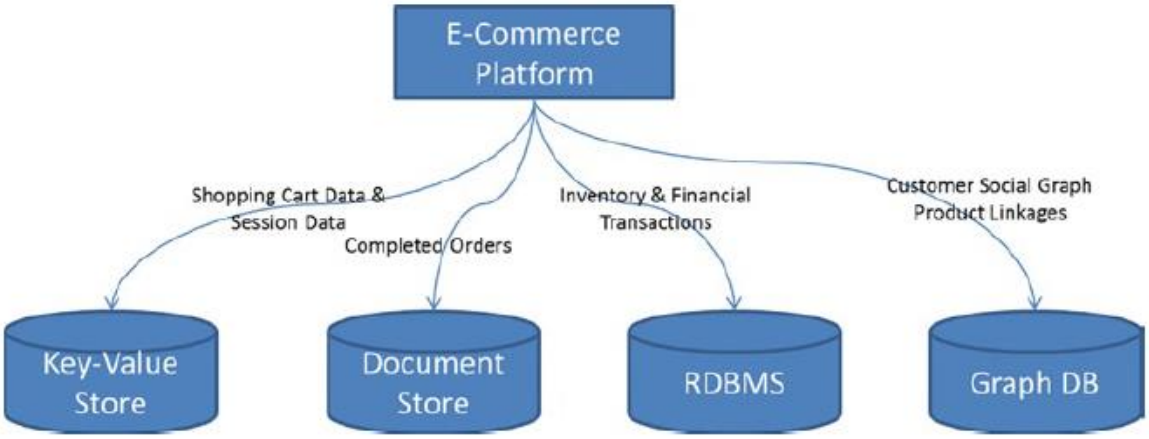


Figure 6.5: Conceptual view of a polyglot persistence layer

(Source: Mohanty et al., 2013:103)

6.4.2 Big Data and fraud

Participants mentioned BD as being an added benefit for fraud analysis in e-commerce. Improving the quality of service, enhancing security and not inconveniencing customers requires business process prudence based on insights from data and analytics. It is easy to inconvenience legitimate customers over fraud alert. While fraud detection and prevention are vital, mitigating inconveniencing a legitimate customer is even more vital. What remains a challenge is creating a balance to allow the use of BD in an effective way to detect and eliminate fraudulent transactions using many different data sources while improving the customer experience. What then are the different data sources that can provide a rich data source to identify and forecast fraud prior to its occurrence based on streaming data?

6.4.3 Increase site traffic does not warrant sales improvement

Interview participants mentioned that after launching campaigns, traffic to the site improves to huge numbers but converting to sales remains a challenge. What could be the reason for customers flogging to the site yet do not buy? Could it be that these customers do not have credit cards? Or do they want items to be cheaper? This still remains a mystery and requires scientific research to unearth the reasons behind the trend.

6.4.4 New scalable software architectures for challenges

The aim of many organisations curating Big Data is a channelled path from its present data status quo to business metamorphosis where it may have created an extensive data platform for third party system users or organisations to deploy data (a form of data commoditisation) as part of service delivered by the organisation.

This kind of platform seems to be required due to the benefits associated with Big Data curation and the growth of data in volumes, variety, velocity and accuracy. Such implementations, due to demand, mandate large scale system architectures which imply significant system architecture challenges especially with distributed systems. Some foreseeable problems that may arise include data consistency, temporary failures, concurrent processing and data replication.

BD curation depends mainly on two primary tools MapReduce and NoSQL due to the large BD jobs processed. However many queries for information are small, dynamic and interactive. MapReduce and NoSQL implementations seem to defeat the purpose when deployed for functions accomplishable by traditional implementations yet with increasing data sizes, traditional systems perform poorly in delivery due to speed and diversity of data forms. How then is the feasibility of a combined, new systems architecture that leverages both a relational storage approach and new data curation paradigms such as NoSQL to yield better delivery in querying and retrieval?

6.4.5 Evaluating the effectiveness of growth hacking techniques in marketing

There are paid and unpaid advertising. The highest return on investment is on unpaid advertising. The challenge then is how merchants and start-ups could leverage unpaid advertising to promote goods and services with quantifiable benefits. Using growth hacking is a well-known technique used by lean start-ups with minimal financial input.

6.4.6 Big Data monetisation as part of data supply chain

Creating new business revenue sources from data is an opportunity that many businesses hope to attain, especially mobile and telecommunication companies. The business models and value delivered to the customer culminates in the discovery, capture, storage, analysis dissemination and use of the data. The data, due to continuous generation, grows colossal in size with hidden yet relevant patterns and trends that may result in the insights that can be monetised by the curating organisation. Data monetisation forms part of a data supply chain with ethical and regulatory vectors which have not been fully explored, hence the lack of a much needed knowledge base for the interested organisation. Furthermore, there is a lack of a framework as a guide for organisations seeking to commence with the BD monetisation journey as part of a BD curation process. This framework will help identify value in the data supply chain as sellers and buyers sit at opposite ends of the continuum with processes in place to optimise the value generation to all stakeholders.

6.4.7 Big Data service refinery

Big Data service refinery refers to increasing corporate data storage efficiency by breaking down silos across data stores and source as a way to optimise data warehousing. Predominantly, the majority of organisations are curating data using traditional data tools which have proven inadequate with velocity, volume and variety of current structured and unstructured data. The integration of disruptive technologies into the business space, according to many practitioners, comes with huge cost savings, which is the reason these technologies are considered disruptive.

There is a lack of knowledge as to how this service refinery cost savings can be quantified to give practitioners an idea to warrant pursuing this course for guided data management. Improving corporate data storage efficiency may also imply the application of new and improved software algorithms for data streams and storage such as data classification for optimised data retrieval. The challenge then is how to leverage identified and efficient techniques to improve storage and data retrieval times for a more efficient design.

6.4.8 Business model, dig data engineering and framework

An organisation's business model reflects a conceptual representation of how the organisation obtains and delivers value to prospective customers. That is, the business model reflects how an organisation operates to accomplish its goals. The business model dictates which information and technology assets are needed for operation. As organisations plan to undertake their Big Data curation journeys, being aware of the business model may enable a seller to strategise and collect insightful data to affect how business may be conducted in a sustainable and data-oriented way while positioning the curating firm in a state of 'information advantage' such as better and faster decisions from enhanced information. For example, data monetisation could be a secondary source of revenue for an organisation that knows its data and what possible challenges the data could solve both inside and outside the organisation.

It is apparent that traditional curation systems that use online analytical processing (OLAP) technologies are moderately easy to implement especially from a framework standpoint considering the many offerings, but transitioning to fully-fledged Big Data analytical implementations or combining both is a challenge. Furthermore, there are not many, if at all, proven and tested Big Data frameworks that support such optimised implementations. The basis of a BD framework is to better understand large scale Big Data tenets such as partitioning and distribution across table to enhance parallelism and adopt the intricacies of

replication, indexing and data management techniques to improve data throughput as the basis of a data-centred organisational **Big Data Framework**.

6.5 Reflection

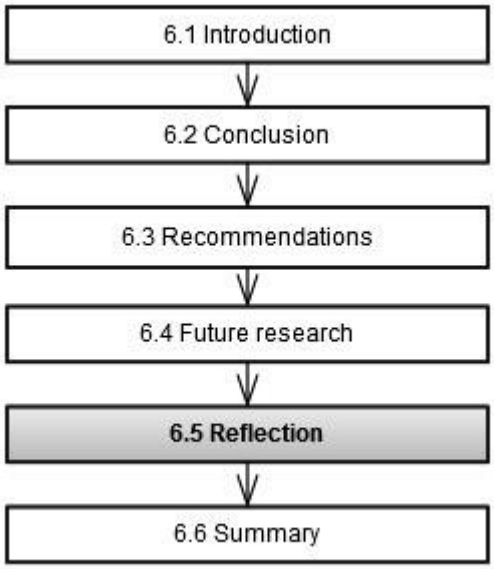


Figure 6.6: Chapter layout – Reflection

Pre-interviewing participants to divide subsequent interviews into business and technical interviews, was an idea to made interviews more focused. This approached allowed interviews to proceed with marked relevance as not all participants had knowledge of BD. Interviews were semi structured and open ended to allow the interview process to take new turns based on how participants answered questions. Some participants requested interview questions prior to interview so answers may have been thought out which may have mitigated the richness of answers ascertained. This research is a single case study of the organisation, to validate findings and results obtained from the research further investigation will be required for similar institutions (multi case study) to ascertain differences and possible similarities so inferences can be drawn where necessary.

Curation guidelines and recommendations were not tested in the case environment due to time constraints and work scope, testing to ascertain the applicability and where necessary improvements can be made of the proposed guidelines and recommendations will enable further researcher based on results.

6.6 Summary

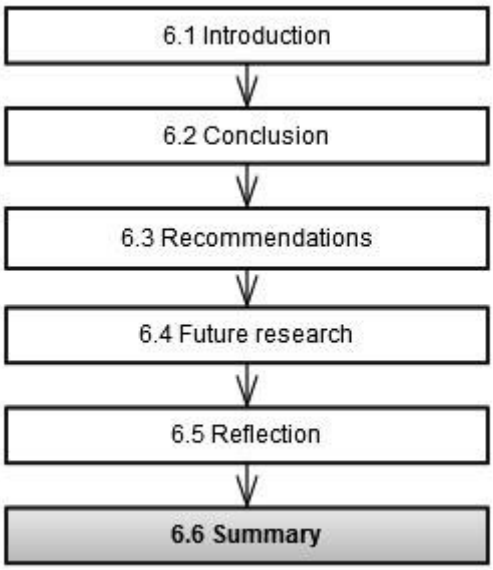


Figure 6.7: Chapter layout – Summary

Chapter 6 focuses on conclusions and recommendations of the research. Research questions one and two were answered with sub-questions to help canvass more insights into answering the main research questions. As stated in Section 6.3, Spree should take the position that they do have BD at their disposal, and that they are collecting BD. Accepting that available data at Spree may not be colossal, but given the context in which Spree is operating and the need for detailed information about business elements, BD should become part of the Spree strategic plans to place itself among information advantaged organisations (see Section 2.6.1).

REFERENCES

Adam, I.O. 2014. *The ontological, epistemological and methodological debates in Information systems research: a partial review*. [Online]. Available from: http://ssrn.com/abstract=2411620 orhttp://dx.doi.org/10.2139/ssrn.2411620. [Accessed: September 9, 2014].

Adamopoulos, A. 2013. *Data classification and storage optimisation.* [Online]. Available from: http://www.snia.org/sites/default/files/Sales_Qual_Data_Strategies-ILM.pdf. [Accessed: October 2, 2014].

Adduci, R., Blue, D., Chiarello, G., Chickering, J., Mavroyiannis, D., Michandani, S., Schleier-Smith, J., Solimando *et al.* 2011. *Big Data: big opportunities to create business value*. [Online]. Available from: http://www.emc.com/microsites/cio/articles/big-data-big-opportunities/LCIA-BigData-Opportunities-Value.pdf. [Accessed: November 1, 2014].

Adriaenssens, C. 2013. *Big Data & opportunities for research*. [Online]. Available from: http://www.ipsos.com/content/big-data-opportunities-research. [Accessed: November 1, 2014].

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L. *et al*. 2012. *Challenges and opportunities with Big Data*. [Online]. Available from: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1000&context=cctech. [Accessed: September 20, 2014].

Agrawal, D., Das, S. & Abbadi, E.l.A. 2011. *Big Data and cloud computing: current state and future opportunities*. [Online]. Available from: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf. [Accessed: September 3, 2014].

Agrawal, R., Ailamaki, A., Bernstein, P.A., Brewer, E.A., Carey, M.J., Chaudhuri, S., Doan, A., Florescu, D. *et al*. 2008. *The Claremont report on database research*. [Online]. Available from: http://db.cs.berkeley.edu/claremont/claremontreport08.pdf. [Accessed: April, 2014].

Alt, R. & Zimmermann, H.D. 2001. Introduction to special section-business models. *Electronic markets -The International Journal*, 11(1):1019-6781, October.

An, A. 2012. *Big Data–big deal*. [Online]. Available from: http://khamreang.msu.ac.th/miwai13/doc/keynote-talk.pdf. [Accessed: September 19, 2014].

Angeles, S. 2014. *Big Data: what small businesses don't understand.* [Online]. Available from: http://www.businessnewsdaily.com/6868-5-things-about-big-data-small-businesses-don-t-understand.html. [Accessed: November 1, 2014].

Austin, T. & Mitcham, J. 2007. *Preservation and management strategies for exceptionally large data formats: Big Data*. [Online]. Available from: http://www.casparpreserves.eu/Members/ArchaeologyDataService/Papers/preservation-and-management-strategies-for-exceptionally-large-data-formats-big-data-1/. [Accessed: May 20, 2013].

Babbie, E. & Mouton, J. 2009. The practice of social research. Cape Town: Oxford University Press.

Babbie, Earl. 2001. *The practice of social research.* 9th ed. Belmont: Wadsworth.

Baird, G.M. 2013. *The future of water infrastructure asset management, part 3: breaking down organisational silos as barriers to cost savings*. [Online]. Available from: http://www.mwhglobal.com/wp-content/uploads/2013/08/AWWA-Aug-2013-The-Future-of-Asset-Management-Part-3-Breaking-down-Organisational-Silos-as-Barriers-to-Cost-Savings_1.pdf. [Accessed: November 1, 2014].

Bandaranayake, T. 2012. *Understanding research philosophies and approaches*. [Online]. Available from: http://www.slideshare.net/thusharabandaranayake/understanding-research-philosophies. [Accessed: March 20, 2014].

Banerjee, S., Bolze, J., Mcnamara, J. & O'Reilly, K. 2011. *How Big Data can fuel bigger growth*. [Online]. Available from: http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Outlook-How-Big-Data-can-fuel-bigger-growth-Strategy.pdf. [Accessed: September 30, 2013].

Basch, D. 2012. *A relevant tale: how Google killed Inkomi*. [Online]. Available from: http://diegobasch.com/a-relevant-tale-how-google-killed-inktomi. [Accessed: January 21, 2013].

Becla, J. & Wang, D.L. 2005. *Lessons learned from managing a petabyte*. [Online]. Available from: http://www.cidrdb.org/cidr2005/papers/P06.pdf. [Accessed: June 20, 2013].

Ben-Gan, I., Sarka, D. & Talmage, R. 2012. *Querying Microsoft SQL server 2012: training kit.* California: O'Reilly Media.

Ben-Shabat, H. & Gada, K. 2012. *Beauty and the E-commerce beast*. [Online]. Available from: http://www.atkearney.com/documents/10192/642824/Beauty+and+the+E-Commerce+Beast.pdf/cb1c6e4b-7bfb-4caa-b463-9fbc2302f9db. [Accessed: November 18, 2014].

Bendassolli, P.F. 2013. Theory building in qualitative research: reconsidering the problem of induction. *Forum Qualitative Social Research*, 14(1), January.

Bi Insider. 2014. *Column and row base database storage.* [Online]. Available from: http://bi-insider.com/business-intelligence/column-and-row-based-database-storage/. [Accessed: November 9, 2014].

Biesdorf, S., Court, S. & Willmott, P. 2013. *Big Data: what's your plan?* [Online]. Available from: www.mckinsey.com/insights/business_technology/big_data_whats_your_plan. [Accessed: June 21, 2014].

Bizer, C., Boncz, P., Brodie, M.L. & Erling, O. 2012. The meaningful use of Big Data: four perspectives—four challenges. *ACM SIGMOD Record*, 40(4):56-60, December.

Bohé, A., Hong, M., Macdonald, C. & Paice, N. 2013. *Data monetization in the age of Big Data*. Accenture. [Online]. Available from: at: http://www.accenture.com/us-en/Pages/insight-data-monetization-big-data-mobile-operators.aspx. [Accessed: September 25, 2013].

Boyd, D. & Crawford, K. 2012. Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society,* 15(5):662-679, May.

Brockman, M.S. & Russell, S.T. 2009. *Decision-making/reasoning skills*. Building Partnership for Youth: National 4-H Council and the University of Arizona.

Bruell, A. 2013. *Ad holding companies look to spread data wealth beyond media agencies: Omnicom, Publicis, WPP create cross-agency groups to leverage data.* Adage. [Online]. Available from: http://adage.com/article/agency-news/ad-holding-companies-spread-data-wealth-media-agencies/240373/. [Accessed: March 19, 2013].

Buhl, H.U., Röglinger, M., Moser, D.K.F. & Heidemann, J. 2013. Big Data. *Wirtschaftsinformatik*, *55*(2):63-68, April.

Burrows, M. 2006. *The Chubby lock service for loosely-coupled distributed systems*. [Online]. Available from: http://static.googleusercontent.com/media/research.google.com/en//archive/chubby-osdi06.pdf. [Accessed: June 10, 2014].

Canning, L. 2014. *Business model you*. [Online]. Available from: http://blog.entrepreneurthearts.com/2012/01/13/business-model-you/. [Accessed October 21, 2014].

Casey, T., Harbitter, A., Leary, M. & Martin, I. 2008. Secure information sharing for the US Government. White paper. *Nortel Technical Journal*.

Castrejon-Castillo, J., Vargas-Solar, G., Collet, C. & Lozano, R. 2014. *Model-driven cloud data storage*. [Online]. Available from: https://hal.inria.fr/hal-00922888/document. [Accessed: December 2, 2014].

Chandramouly, A. & Stinson, K. 2013. *Enabling Big Data solution with centralised data management*. [Online]. Available from: http://www.google.co.za/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CCYQFjAC&url=http%3A%2F%2Fwww.intel.co.za%2Fcontent%2Fdam%2Fwww%2Fpublic%2Fus%2Fen%2Fdocuments%2Fwhite-papers%2Fenabling-big-data-management-solutions-with-centralized-data-management.pdf&ei=56uaVPTsFcjsUt2igaAL&usg=AFQjCNHpJ3M4Eo4iivqgBbDiBIqZqnuSVw&sig2=3HbnDR9hmQk5XKulj327bQ&bvm=bv.82001339,d.d24. [Accessed: November 10, 2014].

Chaudhuri, S. 2012. *What next? A half-dozen data management research goals for Big Data and the cloud*. [Online]. Available from: http://www.forrester.com/pimages/rws/reprints/document/112461/oid/1-PBE69P. [Accessed: October 6, 2014].

Chenail, R.J. 2012. *Conducting qualitative data analysis: managing dynamic tensions within.* [Online]. Available from: http://www.nova.edu/ssss/QR/QR17/chenail-tensions.pdf. [Accessed: July 15, 2014].

Chiu, H-C., Hsieh, Y-C., Kao, Y-H. & Lee, M. 2007. The determinants of email receivers' disseminating behaviours on the Internet. *Journal of Advertising Research*, 47(4):524-534.

Codd, E.F. 1970. A relational model of data for large shared data banks. *ACM,* 13(6):377-387.

Crawford, K. & Schultz, J. 2014. *Big Data and due process: toward a framework to redress predictive privacy harms.* [Online]. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325784. [Accessed: May 15, 2014].

Creswell, J.W. 2013. *Research design: qualitative, quantitative, and mixed-methods approaches*. CA: Sage.

Creswell, J. & Clark, P.V. 2007. *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cuzzocrea, A., Song, l. & Davids, K.C. 2011. Analytics over large-scale multidimensional data: the Big Data revolution. DOLAP '11 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP. *ACM*:101-104. October.

Das, S. 2012. *Harnessing the power of Big Data in real time through in-memory technology and analytics.* [Online]. Available from: http://www3.weforum.org/docs/GITR/2012/GITR_Chapter1.7_2012.pdf. [Accessed: June 26, 2012].

Davenport, T.H. 2013. *At the Big Data crossroads: turning towards a smarter travel experience*. Amadeus IT Group. Madrid: Rivington Press.

Davenport, T.H. 2014. *Big Data at work: dispelling the myths, uncovering the opportunities*. [Online]. Available from: http://hbr.org/product/big-data-at-work-dispelling-the-myths-uncovering-the-opportunities/an/16574-HBK-ENG?referral=01240?referral=01240. [Accessed: April 3, 2014].

Davenport, T.H. & Prusak, L. 2005. *Working knowledge: how organisations manage what they know.* [Online]. Available from: http://www.acm.org/ubiquity/book/t_davenport_1.html. [Accessed: June 2, 2014].

Davey, I. 2013. *Consumers, Big Data, and online tracking in the retail industry: a case study of Walmart.* [Online]. Available from: http://www.scribd.com/doc/187338572/Walmart-Privacy-ReportConsumers-Big-Data-and-Online-Tracking-in-the-Retail-Industry-A-Case-Study-of-Walmart#force_seo. [Accessed: October 26, 2014].

Davis, K. & Patterson, D. 2012. *Ethics of Big Data*. Sebastopol: O'Reilly Media.

Devlin, B., Rogers, S. & Myers, J. 2012. *Big Data comes of age.* [Online]. Available from: http://www-03.ibm.com/systems/hu/resources/big_data_comes_of_age.pdf. [Accessed: August 01, 2014].

Dijcks, J. 2013. *Oracle: Big Data for enterprise.* Redwood Shores: Oracle.

Duhigg, C. 2013. *How companies learn your secrets.* [Online]. Available from: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0. [Accessed: March 12, 2013].

Dumbill, E. 2012. *Planning for Big Data.* [Online]. Available from: http://radar.oreilly.com/edd/. [Accessed: June 24, 2012].

Easterby-Smith, M., Thorpe, R. & Jackson, P. 2008. *Management research*. 3rd ed. London: Sage.

Eckler, P. & Rodgers, S. 2010. *Viral marketing on the Internet*. Wiley International Encyclopedia of Marketing 4.

Eckerson, W. 2011. *Enterprise data strategy.* [Online]. Available from: http://docs.media.bitpipe.com/io_10x/io_100166/item_417254/Creating%20an%20Enterprise%20Data%20Strategy_final.pdf. [Accessed: September 2, 2014].

Elo, S. & Kynga, S.H. 2008. The qualitative content analysis process. *Journal of Advanced Nursing,* 62(1):107-115.

Engström, A. & Salehi-Sangari, E. 2007. *Assessment of business-to-business e-marketplaces' performance.* [Online]. Available from: http://epubl.ltu.se/1402-1544/2007/22/LTU-DT-0722-SE.pdf. [Accessed: November 10, 2014].

Eriksson, P. & Kovalainen, A. 2008. *Qualitative Methods in Business Research*. London: Sage Publications.

Everest, T. 2014. Resolving the qualitative-quantitative debate in healthcare research. *Academic Journals,* 5(1):6-15, February.

Flick, U. 2013. *The Sage handbook of qualitative data analysis*. London: Sage.

Flowers, P. 2009. *Research Philosophies—importance and relevance.* Cranfield School of Management, UK. [Online]. Available from: http://www.networkedcranfield.com/cell/Assigment%20Submissions/research%20philosophy%20-%20issue%201%20-%20final.pdf. [Accessed: July 20, 2012].

Floyer, D. 2012. *Enterprise Big-data.* [Online]. Available from: http://wikibon.org/wiki/v/Enterprise_Big-data#Big-data_Definition1. [Accessed: April 27, 2012].

Forsyth, C. 2012. *For Big Data analytics there's no such thing as too big.* [Online]. Available from: http://www.cisco.com/en/US/solutions/ns340/ns517/ns224/big_data_wp.pdf. [Accessed: March 13, 2013].

Fryersolution. 2013. *Accelerating the value of Big Data for global clients.* [Online]. Available from: http://freyrsolutions.com/big-data/. [Accessed: September 21, 2014].

Ganis, M. 2013. *Social media Big Data analytics.* [Online]. Available from: csis.pace.edu/ctappert/dps/d861-14/ganis.ppt. [Accessed: November 7, 2013].

Gantz, J. & Reinsel, D. 2012. *The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the Far East*. [Online]. Available from: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. [Accessed: January 4, 2015].

García-Alcaraz, J.L., Maldonado-Macías, A.A. & Cortes-Robles, G. 2014*. Lean manufacturing in the developing world: methodology, case studies and trends from Latin America.* New York: Springer.

Gelber, R. 2012. *Wharton professor pokes hole in Big Data balloon*. [Online]. Available from: http://www.datanami.com/datanami/2012-05-03/wharton_professor_pokes_hole_in_big_data_balloon.html/. [Accessed: January 1, 2014].

Grant, R.M. 2010. *Contemporary strategy analysis and cases: text and cases*. Indiana: John Wiley & Sons.

Gray, D.E. 2013. *Doing research in the real world*. London: Sage.

Grace, L.K.J., Maheswari, V. & Nagamalai, D. 2011. Analysis of weblogs and web user in web mining. *International Journal of Network Security & Its Applications (IJNSA),* 3(1), January.

Gualtieri, M. 2012. *The pragmatic definition of Big Data.* [Online]. Available from: http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data. [Accessed: November 2, 2014].

Hadjigeorgiou, C. 2013. *RDBMS vs. NoSQL*. [Online]. Available from: http://www.epcc.ed.ac.uk/sites/default/files/Dissertations/2012-2013/RDBMS%20vs%20NoSQL%20-%20Performance%20and%20Scaling%20Comparison.pdf. [Accessed: June 23, 2014].

Halpin, T. & Morgan, T. 2010. *Information modeling and relational databases*. Burlington: Morgan Kaufmann.

Hardesty, L. 2012. *MIT, Intel unveil new initiatives addressing Big Data.* [Online]. Available from: http://web.mit.edu/newsoffice/2012/big-data-csail-intel-center-0531.html. [Accessed: June 18, 2012].

Hatch, M.J. & Cunliffe, A.L. 2006. *Organization theory: modern, symbolic, and postmodern perspectives.* 2nd ed. New York: Oxford University Press.

Hebner S. 2012. *Middleware software: IBM Software group at pulse general session 2012, Day 1*. [Online]. Available from: http://www.youtube.com/watch?v=j1p2C4cbnuI. [Accessed: July 20, 2012].

Hitzler, P. & Janowicz, K. 2013. Linked data, Big Data, and the 4th paradigm. *Semantic Web*, *4*(3):233-235.

Henke, C., Coulbourne, C. & Kadochnikov, N. 2011. *The value of smarter merchandising, how adopting a customer centric merchandising can increase revenue and margins.* [Online]. Available from: https://www.ibm.com/smarterplanet/global/files/us__en_us__retail__rew03010-usen-00_hr.pdf. [Accessed: November 20, 2014].

Hockenson, L. 2013. *Leveraging Big Data in the world of enterprise*. [Online]. Available from: http://thenextweb.com/insider/2013/03/20/big-data-for-business/. [Accessed: September 30, 2013].

Hoskins, C.N. & Mariano, C. 2004. *Research in nursing and health: understanding and using quantitative and qualitative methods.* 2nd ed. New York: Springer.

Howe, B. 2013. *Introduction to Data Science.* [Online]. Available from: https://class.coursera.org/datasci-001/lecture/21. [Accessed: June 20, 2014].

Hritzuk, N., Esquero, I., Jones, K. & Burke, E. 2013. *The consumer decision journey: retail understanding consumer decision-making along the retail path to purchase*. Microsoft advertising consumer insights.

Hsu, K. 2013. *Fraud detection, identify potential fraud.* [Online]. Available from: http://www.datameer.com/solutions/usecases/fraud-detection.html. [Accessed: November 1, 2014].

Hutchby, I. & Wooffitt, R. 2008. *Conversation analysis*. London: Polity.

Hurtgen, H., Natarajan, S., Spittaels, S., Vetvik, O.J. & Ying, S. 2012. *Crunch time: using BD to boost telco marketing capabilities.* McKinsey & Company, Inc.

Hyett, N., Kenny, A. & Dickson-Swift, V. 2014. Methodology or methods? A critical review of qualitative case study reports. *International Journal of Qualitative Studies on Health and Well-being, 9:23606.* [Online]. Available from: http://dx.doi.org/10.3402/qhw.v9.23606. [Accessed: October 03, 2014].

Ivarsson, T. 2010. *NoSQL for dummies.* [Online]. Available from: http://www.slideshare.net/thobe/nosql-for-dummies. [Accessed: October 6, 2013].

Jacobs, A. 2009. The pathologies of Big Data. ACM. *Commun ACM*, 52(8):36-44, August.

Jackson, J. 2011. *IBM Watson vanquishes human jeopardy foes*. [Online]. Available from: http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html. [Accessed: April 1, 2012].

Jahnke, L., Asher, A. & Keralis, S.D.C. 2012. *The problem of data.* Washington, DC: Council on library and information resources. [Online]. Available from: http://www.clir.org/pubs/reports/pub154/pub154.pdf. [Accessed: November 8, 2014].

Jansen, H. 2010. *The logic of qualitative survey research and its position in the field of social research methods*. [Online]. Available from: http://www.qualitative-research.net/index.php/fqs/article/view/1450/2946. [Accessed: November 5, 2014].

Jonker, J. & Pennink, B. 2010. *The essence of research methodology: a concise guide for Master and PhD students in Management Science*. Heidelberg: Springer.

Kadam, A., Shaikh, R. & Parab, P. 2013. *Data collection: primary & secondary.* [Online]. Available from: http://www.slideshare.net/parabprathamesh/primary-sec. [Accessed: May 23, 2014].

Kaminski, J. 2011. Theory applied to informatics—Lewin's Change theory. *CJNI: Canadian Journal of Nursing Informatics*, 6(1):1-4, March.

Kamakura, W.A. 2007. Cross-selling: offering the right product to the right customer at the right time. *Journal of Relationship Marketing*, 1(6):3-4, February.

Kaplan, B. & Maxwell, J.A. 2005. Qualitative research methods for evaluating computer information systems. In *Evaluating the organizational impact of healthcare information systems.* New York: Springer: 30-55.

Keller, B. 2011. *Marriott Hotel website—study report under the authority of Institut für Software-Ergonomie und Usability.* [Online]. Available from: http://miratech.fr/v5bis/wp-content/themes/miratech/blog/IUTP_Marriott_Study_Report. [Accessed: December 1, 2014].

Kelly, J. 2012. *There's no excuse for not leveraging Big Data: services angle.* [Online]. Available from: http://servicesangle.com/blog/2012/07/19/theres-no-excuse-for-not-leveraging-big-data/. [Accessed: March 18, 2013].

Kelton Research. 2010. *Global survey: the business impact of Big Data*. [Online]. Available from: http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEOv.pdf. [Accessed: June 20, 2014].

Kimball, R. & Ross, M. 2013. *The data warehouse toolkit, the definitive guide to dimensional modeling.* Indiana: Wiley.

Kinsella, E.A. 2006. *Hermeneutics and critical hermeneutics: exploring possibilities within the art of interpretation*. [Online]. Available from: http://www.qualitative-research.net. [Accessed: September 21, 2013].

Khatri, V. & Brown, C.V. 2010. Designing data governance. *ACM*, 53(1):148-152, January.

Kolluru, R. 2012. *Hands on turn text into information: regular expression makes business sense of unstructured data.* [Online]. Available from: http://www.teradatamagazine.com/v12n01/Tech2Tech/Turn-Text-Into-Information/. [Accessed: June 3, 2013].

Kohlbacher, F. 2005. *The use of qualitative content analysis in case study research.* [Online]. Available from: http://nbn-resolving.de/urn:nbn:de:0114-fqs0601211. [Accessed: September 24, 2013].

Laningham, S. 2010. *Tackling Big Data with Hadoop and IBM big sheets*. [Online]. Available from: http://www.ibm.com/developerworks/podcast/dwi/cm-int082410-bigsheets.html. [Accessed: June 10, 2014].

LaValle, S., Hopkins, M.S., Lesser, E., Shockley, R. & Kruschwitz, N. 2011. *Analytics: the new path to value. How the smartest organizations are embedding analytics to transform insights into action.* [Online]. Available from: http://public.dhe.ibm.com/common/ssi/ecm/gb/en/gbe03371usen/GBE03371USEN.PDF. [Accessed: September 1, 2014].

LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. & Kruschwitz, N. 2013. Big Data, analytics and the path from insights to value. *MIT Sloan Management Review*, *21*.

Leblanc, R. 2011. *Why Big Data? Why now? IBM Corporation.* [Online]. Available from: http://www.slideshare.net/MauricioSWG/robert-leblanc-why-big-data-why-now. [Accessed: June 25, 2012].

Leblanc, R. 2012. *Pulse general session 2012, Day 1*. [Online]. Available from: http://www.youtube.com/watch?v=j1p2C4cbnuI. [Accessed: July 20, 2012].

Laudon, K.C. & Traver, C.G. 2002*. E-Commerce: business, technology, society*. Boston: Addison Wesley.

Laudon K.C. & Traver C.G. 2009. *E-Commerce 2014*. New York: Prentice Hall.

Leidwinger, S. 2013. *Enhanced 360 degree view of the customer.* [Online]. Available from: http://www.ibmbigdatahub.com/video/big-data-use-case-2-enhanced-360-degree-view-customer. [Accessed: October 2, 2014].

Li, S., Sun, B., Alan, L. & Montgomery, A. 2011. Cross-selling the right product to the right customer at the right time. *Journal of Marketing Research,* 48(4):683-700, August.

Lin, S.H. 2013. *An introduction to Decision Theory*. Cambridge: MCGraw Hill.

Linder, J. & Cantrell, S. 2000. *Changing business model: surveying the landscape*. Cambridge: Accenture.

Londre, L.S. 2009. Marketing, the marketing mix (4p's), and the nine p's. [Online]. Available from: http://www.londremarketing.com/documents/Nineps10232008.pdf. [Accessed: November 20, 2014].

Louwers, J. 2013. *Adding a V to the four Big Data thinking.* [Online]. Available from: http://johanlouwers.blogspot.com/2013/07/adding-v-to-four-v-big-data-thinking.html. [Accessed: September 7, 2014].

Lopez, S.C. 2009. *A qualitative study of an e-commerce organization in transition*. New York: ProQuest.

Lumpkin, G. 2013. *Multidimensional data: integrating Big Data into corporate information architectures gives companies new insight.* [Online]. Available from: http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf. [Accessed: November 8, 2014].

Malm, P. 2013. *Three reasons why Big Data is awesome*. [Online]. Available from: https://econsultancy.com/blog/63365-three-reasons-why-big-data-is awesome#i.14g9r89113ie5i. [Accessed: June 2, 2013].

Malicke, D. 2012. *Airing out your data with a status board.* Launch and learn: University of Michigan.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A.H. 2011. *Big Data: the next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Marciano, R. 2012. *Big Data curation: innovative approaches and techniques.* [Online]. Available from: http://www.lis.illinois.edu/events/2012/02/03/richard-marciano-big-data-curation-innovative-approaches-and-techniques. [Accessed: June 1, 2012].

Mayer-Schönberger, V. & Cukier, K. 2013. *Big Data: a revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Maas, A.V. 2013. *Mastering strategy and implementation.* [Online]. Available from: http://www.slideshare.net/ArnoudvanderMaas1/article-mastering-strategy-implementation. [Accessed: September 10, 2013].

Mayring, P. 2014. Qualitative content analysis: theoretical foundation, basic procedures and software solution. Forum Qualitative Social Research (submitted).

McLeod, S.A. 2008. *Case study method*. [Online]. Available from: http://www.simplypsychology.org/case-study.html. [Accessed: September 1, 2014].

McCallum, E.Q. 2012. *Bad data: mapping the world of data problems*. Cambridge: O'Reilly.

Meijer, E. & Bierman, G. 2011. A co-relational model of data for large shared data banks. *Communications of the ACM*, *54*(4):49-58, April.

Merson, P.F. 2009. *Data model as an architectural view.* [Online]. Available from: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1280&context=sei. [Accessed: February 20, 2014].

Merriam, S. 2009*. Qualitative research:* a *guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Mohanty, S., Jagadeesh, M. & Srivatsa, H. 2013. *Big Data imperatives: enterprise Big Data warehouse, BI implementations and analytics*. Berkeley, CA: Apress.

Mosley, M., Brackett, M., Earley, S. & Henderson. 2009. *The DAMA guide to the data management body of knowledge (DAMA-DMBOK Guide)*. Bradley Beach: Technics Publications, LLC.

Modell, M.E. 2007. *Data analysis, data modelling and classification.* New York: McGraw-Hill.

Myers, D.M. 2009. *Qualitative research in business & management.* London: Sage.

Myers, M.D. & Newman, M. 2007. *The qualitative interview in IS research: examining the craft.* [Online]. Available from: http://www.carlosmello.unifei.edu.br/Disciplinas/Mestrado/PCM-10/Textos%20para%20Leitura/Texto_Leitura_Atividade_2.pdf. [Accessed: October 21, 2014].

Naspers Media24. 2014. *Naspers: our group.* [Online]. Available from: http://www.naspers.com/page.html?pageID=3. [Accessed: July 3, 2014].

Nayak, A., Poriya, A. & Poojary, D. 2013. Types of NoSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems* (IJAIS). ISSN: 2249-0868. Foundation of Computer Science FCS, New York, USA.

Neuman, W.L. 2011. *Social research methods: qualitative and quantitative approaches*. Cape Town: Pearson.

Nerney, C. 2013*. McKinsey report: Big Data at centre of disruptive technologies.* [Online]. Available from: http://data-informed.com/mckinsey-report-big-data-at-center-of-disruptive-technologies/. [Accessed: November 1, 2013].

Nikov, A. 2012. *E-commerce business models and concepts*. [Online]. Available from: http://www2.sta.uwi.edu/~anikov/INFO3435/lectures/02-EC-lect-business-models-concepts.pdf. [Accessed: September 07, 2014].

Nickols, F. 2012. *Solution Engineering: ten tips for beefing up your problem solving toolbox*. [Online]. Available from: http://www.nickols.us/Solution_Engineering_Tutorial.pdf. [Accessed: July 20, 2014].

Nkwi, P., Nyamongo, I. & Ryan, G. 2001. *Field research into socio-cultural issues: methodological guidelines*. Yaounde, Cameroon, Africa: International Center for Applied Social Sciences, Research and Training/UNFPA.

Orlikowski, W.J. & Baroudi, J.J. 1991. Studying information technology in organizations: research approaches and assumptions. *Information Systems Research,* 2(1):1-28, March.

Oosthuizen, M.J. & Phil, D.L. 2012. *The portrayal of nursing in South African newspapers: a qualitative content analysis.* [Online]. Available from: http://uir.unisa.ac.za/bitstream/handle/10500/8897/ajnm_v14_n1_a6.pdf?sequence=1, Africa Journal of Nursing and Midwifery. [Accessed: November 4, 20014].

O'Shea, V. & Shah, R. 2014. *Big Data in capital markets: at the start of the journey*, [Online]. Available from: http://share.thomsonreuters.com/general/PR/Big%20Data%20IB_White%20Paper_Aug2014.pdf. [Accessed: November 1, 2014].

Parise, S. 2012. *Four strategies to capture and create value from Big Data*. [Online]. Available from: http://www.iveybusinessjournal.com/topics/strategy/four-strategies-to-capture-and-create-value-from-big-data#.UV7t0KJcwfN. [Accessed: March 2, 2013].

Parker, M.G. 2013. *Annual report and notice of annual meeting, Nike, INC.* [Online]. Available from: http://investors.nikeinc.com/files/nike2013form10K.pdf. [Accessed: November 1, 2014].

Parkinson, G. & Drislane, R. 2011. *Qualitative research.* Online Dictionary of the Social Sciences. [Online]. Available from: http://bitbucket.icaap.org/dict.pl. [Accessed: July 17, 2014].

Pea, R.D. 1982. *What is planning development the development of?* [Online]. Available from: http://web.stanford.edu/~roypea/RoyPDF%20folder/A11_Pea_82d.pdf. [Accessed: December 1, 2014].

Piateski, G. & Frawley, W. 1991. *Knowledge discovery in databases*. Massachusetts: MIT.

Pickard, A.J. 2013. *Research methods in information*. London: Facet Publishing.

Pederson, S. 2012. *From data to decision: delivering value from 'Big Data.* [Online]. Available from: http://bigdata.brightplanet.com/Portals/179268/docs/BrightPlanet_Creating%20Intelligence%20from%20Big%20Data.pdf. [Accessed: March 7, 2013].

Pederson, S. 2013. *Exploiting Big Data from the deep web: the new frontier for creating intelligence*. [Online]. Available from: http://bigdata.brightplanet.com/Portals/179268/docs/BrightPlanet_Creating%20Intelligence%20from%20Big%20Data.pdf. [Accessed: November 9, 2014].

Peräkylä, A. 2008*. Conversation analysis.* The Blackwell Encyclopedia of Sociology Online. [Online]. Available from: http://blogs.helsinki.fi/perakyla/files/2008/10/conversationanalysis_08l1.pdf. [Accessed: May 30, 2014].

Perdue, T. 2012. *NoSQL: An overview of NoSQL databases.* [Online]. Available from: http://newtech.about.com/od/databasemanagement/a/Nosql.htm. [Accessed: June 22, 2012].

Pereira, F. 2007. *MPEG multimedia standards: evolution and future*. [Online]. Available from: http://lsdis.cs.uga.edu/GlobalInfoSys/Structured-and-Unstructured-for-EIPs.pdf. [Accessed: June 22, 2012].

Peterson, J. 2014. *Philosophical paradigms, data collection, and analysis design of a green technology education mixed methods research*. [Online]. Available from: https://www.linkedin.com/today/post/article/20140714153011-31971126-philosophical-paradigms-data-collection-and-analysis-design-of-a-green-technology-education-mixed-methods-research. [Accessed: October 20, 2014].

Pokorny, J. 2011. NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, 9(1):69-82.

Pokorny, J. 2013. NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, 9(1):69-82.

Porter, M.E. 1998. *The competitive advantage: creating and sustaining superior performance.* NY: Free Press.

Posey, M.B. 2010. *Intelligent health care, strategies for getting the most out of your business intelligence systems.* [Online]. Available from: http://docs.media.bitpipe.com/io_10x/io_101432/item_447818/SHealthIT_Bizanalytics_Intelligent_Health_final.pdf. [Accessed: September 10, 2014].

Prajapati, V. 2013. *Big Data analytics with R and Hadoop.* Birmingham-Mumbai: Packt.

Primesberger, C. 2011. Big ideas about Big Data. *eWeek*, 28(13):34-37, October.

Ramanathan, R. & Raja, K. (eds.). 2013. *Service-driven approaches to architecture and enterprise integration*. IGI Global.

Rappa, M. 2001. *Business models on the web: managing the digital enterprise.* [Online]. Available from: digitalenterprise.org/models/models.html. [Accessed: December 8, 2014].

Rees, R. 2010. *NoSQL, no problem: an intro to NoSQL databases.* [Online] Available from: http://www.thoughtworks.com/insights/blog/nosql-no-problem-intro-nosql-databases. [Accessed: May 18, 2014].

Rele, A. 2012. *Principal product manager, Magento Big Data and analytics*. [Online]. Available from: http://www.magentocommerce.com/blog/comments/making-big-data-useful-for-smbs/#sthash.I6i2nZP1.dpuf. [Accessed: January 9, 2013].

Ries, E. 2011. *The lean startup: how today's entrepreneurs use continuous innovation to create radically successful businesses*. New York: Random House LLC.

Rijmenam, M.V. 2014. *Ways Big Data can drive your sales to the next level.* [Online]. Available from: http://smartdatacollective.com/bigdatastartups/199006/four-ways-big-data-can-drive-your-sales-next-level. [Accessed: May 10, 2014].

Rho, J.J., Moon, B., Kim, Y. & Yang, D. 2004. Internet customer segmentation using web log data. *Journal of Business & Economics,* 2(11), November.

Rogge, E. 2011. *Googlizing BI with search-based applications: the data warehousing Institute*. [Online]. Available from: http://tdwi.org/articles/2011/06/08/googlizing-bi-with-search-based-applications.aspx. [Accessed: September 30, 2013].

Rosenbush, S. & Totty, M. 2013. How Big Data is changing the whole equation for business. *Wall Street Journal*, *11*.

Ross, J.W., Beath, C.M. & Quaadgras, A. 2013. *You may not need Big Data after all.* [Online]. Available from: https://hbr.org/2013/12/you-may-not-need-big-data-after-all. [Accessed: August 9, 2014].

Ryan, G.W. & Bernard, H.R. 2003. Techniques to identify themes. *Field methods*, 15(1):85-109, June.

Sathi, A. 2012. *Big Data analytics: disruptive technologies for changing the game*. Van Nuys: Mc Press.

Saunders, M., Lewis, P. & Thornhill, A. 2007. *Research methods for business students.* 4th ed. Harlow: Prentice Hall Financial Times.

Saunders, M., Lewis, P. & Thornhill, A. 2009. *Research methods for business students*. London: Pearson Education.

Schmarzo, B. 2013. *Understanding how data powers big business*. New Jersey: Wiley.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. & Tufano, P. 2012*. Analytics: the real-world use of Big Data. How innovative enterprises extract value from uncertain data*. Executive Report. [Online]. Available from: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=XB&htmlfid=GBE03519USEN#loaded. [Accessed: November 21, 2014].

Schubert, P. & Koch, M. 2002. *The power of personalisation: customer collaboration and virtual communities.* [Online]. Available from: http://bas.uni-koblenz.de/bas/publications_light.nsf/9419ff0c508bbae8c1257219004fef63/dbbc9c1dbe38ba89c1257359004908d4/$FILE/schubert_koch.pdf. [Accessed: May 20, 2014].

Scotlant, J. 2012. Exploring the philosophical underpinnings of research: relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. Qatar University. *Canadian Center of Science and Education*, 5(9):1916-4750.

Seeger, M.S. 2009. *Key-Value stores: a practical overview.* [Online]. Available from: http://blog.marc-seeger.de/assets/papers/Ultra_Large_Sites_SS09-Seeger_Key_Value_Stores.pdf. [Accessed: January 10, 2014].

Sen, S. 2012. *Big Data and BI with SQL server and Apache Hadoop.* [Online]. Available from: https://www.youtube.com/watch?v=HM0YX7mpplk. [Accessed: October 21, 2014].

Sensmeier, L. 2013. *How Big Data is revolutionizing fraud detection in financial services*. [Online]. Available from: http://hortonworks.com/blog/how-big-data-is-revolutionizing-fraud-detection-in-financial-services/. [Accessed: September 1, 2014].

Setzer, V.W. 2006. *Data, information, knowledge and competence.* [Online]. Available from: http://www.ime.usp.br/~vwsetzer/data-info.html. [Accessed: June 15, 2014].

Shafer, S.M., Smith, H.J. & Linder, J.C. 2005. The power of business models*. Business Horizons*, 48(3):199-207.

Stubbs, E. 2014. *Big Data big innovation: enabling competitive differentiation through business analysis*. New Jersey: Wiley.

Șaovă, G. & Raduteanu, M. 2013. Optimizing Ecommerce sites through the use heat map. *European International Journal of Science and Technology,* 2(4).

Steiner, J. 2009. *Managing unstructured with Oracle, database 11g*. [Online]. Available from: http://www.oracle.com/us/products/database/options/spatial/039950.pdf. [Accessed: November 13, 2014].

Stephens, R. 2008. *Beginning database design solutions.* Indianapolis: Wiley.

Stemler, S. 2001. *An overview of content analysis.* [Online]. Available from: http://pareonline.net/getvn.asp?v=7&n=17. [Accessed: October 5, 2014].

Stonebraker, M. 2010. SQL v NOSQL databases. *MIT Commun*, 53(4).

Swardt, M. 2008. *Factors influencing the choice to shop online: a psychological study in a South African context.* [Online]. Available from: http://upetd.up.ac.za/thesis/available/etd-11252008-120107/unrestricted/dissertation.pdf. [Accessed: November 9, 2014].

Tawari, S. 2011. *Professional NoSQL*. Indiana: Wiley.

Taylor, J. 2013. *Delivering customer value faster with Big Data analytics*. [Online]. Available from: http://www.fico.com/en/wp-content/secure_upload/DeliveringCustomerValueFasterWithBigDataAnalytics.pdf. [Accessed: November 2, 2014].

Taylor-Powell, E. & Renner, M. 2003. *Analysing qualitative data.* University of Wisconsin-Extension. Corporate extension Madison, Wisconsin. [Online]. Available from: http://learningstore.uwex.edu/assets/pdfs/g3658-12.pdf. [Accessed: August 21, 2013].

Teece, D.J. 2010. *Business models, business strategy and innovation.* [Online]. Available from: http://www.elsevier.com/locate/lrp. [Accessed: September 21, 2014].

Tharakan, K. 2006. *Methodology of social sciences: positivism, anti-positivism and the phenomenological mediation*. [Online]. http://philpapers.org/archive/THAMOS-2.pdf. [Accessed: November 1, 2014].

Thomas, D.R. 2003. *A general inductive approach for qualitative data analysis*. Auckland. New Zealand.

Tonytam. 2010. *What has technology done for words lately?* [Online]. Available from: http://blog.wordnik.com/what-has-technology-done-for-words-lately. [Accessed: January 21, 2013].

Turban, E. & King, D. 2003. *Introduction to e-Commerce*. Upper Saddle River (NJ): Prentice Hall.

Vakali, A. & Terzi, E. 2002. Video data storage policies: an access frequency based approach. *Computers & Electrical Engineering*, 28(6):447-464.

Vaish, G. 2013. *Getting started with NoSQL: your guide to the world and technology of NoSQL*. Birmingham-Mumbai: Packt Publishing.

Villars, R.L. & Vesset, D. 2014. *Building a datacenter infrastructure to support your Big Data plans*. [Online]. Available from: http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/big-data/245209January.pdf. [Accessed: June 20, 2014].

Vogt, P. 2013. *Big Data at the weather company.* Interviewed by Chris Smith. [Online]. Available from: http://www.guardian.co.uk/media-network/media-network-blog/2013/mar/20/weather-company-big-data. [Accessed: June 26, 2014].

Wahyuni, D. 2012. The research design maze: understanding paradigms, cases, methods and methodologies. *Journal of Applied Management Accounting Research*, 10(1):69-80, June.

Walker, T.C. 2010. The perils of paradigm mentalities: revisiting Kuhn, Lakatos and Popper. *Perspectives on Politics,* 8(02):433-451.

Wang, H. 2014. *Theories of competitive advantage.* [Online]. Available from: http://eurekaconnection.files.wordpress.com/2014/02/p-33-43-theoriesof-competitive-advantage-theori-ebook_finaljan2014-v3.pdf. [Accessed: December 1, 2014].

Weinberger, A. 2012. *How to leverage Big Data to monetize customer experiences*. [Online]. Available from: http://www.marketingpower.com/ResourceLibrary/Documents/Whitepapers/Autonomy%20Whitepaper%20Final%202.28.2012.pdf. [Accessed: September 30, 2013].

Welman, C., Kruger, F. & Mitchell, B. 2006. *Research methodology*. 2nd ed. Cape Town: Oxford University press.

White, C. 2011. *Using Big Data for smarted decision making.* [Online]. Available from: ftp://ftp.boulder.ibm.com/software/tw/Using_Big_Data_for_Smarter_Decision-Making_v.pdf. [Accessed: December 2, 2014].

Whitehead, E.J. 2002*. Uniform comparison of data models using containment modeling*. [Online]. Available from: http://xml.coverpages.org/WhiteheadHT02.pdf. [Accessed: January 2, 2013].

Yin, R.K. 2003. Case study research: design and methods. 3rd ed. *Applied Social Research Methods:*5.

Yin, R.K. 2010. *Qualitative research from start to finish*. New York: Guilford Press.

Yin, R.K. 2014. *Case study research: design and methods*. London: Sage.

Zarrella, D. 2010. *The Social media marketing book.* [Online]. Available from: http://hbr.org/2013/12/you-may-not-need-big-data-after-all/ar/1. [Accessed: September 1, 2014].

Zikopoulos, P. 2013. *Big Data for small and medium-sized businesses*. [Online]. Available from: http://www.ibm.com/midmarket/us/en/att/pdf/FV_August_Big_Data_v2.pdf?ca=fv1308&me=feature2&re=usartpdf?. [Accessed: November 2, 2013].

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsh, T., Corrigan, D. & Giles J. 2013. *Harness the power of Big Data: IBM Big Data platform*. New York: McGraw-Hill.

Zikmund, W.G., Babin B.J., Carr, J.C. & Griffin M. 2013. *Business research methods.* Mason, USA: Cengage Learning.

ANNEXURES

ANNEXURE A: Interview Findings

Table 7.1: Pre-interview findings

Pre-interview question	P1, P11	P2, P10	P3, 14	P6	P7, P8, P9, P14	P4, P13, P18	P15, P16	Findings
What is your strategic and tactical involvement with data?	Budget formulation, Planning, decision-making, bestseller categorisation or identification of best, medium and poor sellers, macro data, micro data.	Building and architecture data structures focusing on core e-commerce persistence relative to BI, centralised fragmented data, automation of data curation and serving, reporting, information or data provision, code core functionality for data curation, provision data, insight.	Conceptualise and contextualise business, identify business opportunities, business cases, visualise the business side of ideas, identity types of data to store (curation), define data curation process, define KPIs/KPMs, competitive edge/ advantage, strategies and policies.	Influence communication strategy, identify communication channels, marketing, message packaging, identify consumer (clientele), customer segmentation, recommendations, competition.	Message packaging, get and retain customers, marketing, customer segmentation, integration of customer product life-cycle, decision-making, competitive pricing, influence communicatio n strategy.	Supply chain data (inbound and outbound), customer service data, variations, actual capacity planning, fault finding, reduce defects (defects per million opportunities), customer service affiliations or referrals, customer service contacts, historical data, Inbound/ outbound data, delivery time promise/ prediction, lead times prediction, historical data, forecasting.	Creation of site, configurations, dynamic pricing, recommendation engine, suggestion engine, provide data to BI (provision data), make data available (core), code core functionality for data curation.	Determine and formulate budget, planning, decision-making, bestseller identification, build and architect data structures, reporting, conceptualise and contextualise business opportunities, identify apposite data for curation, influence communication strategy, identify communication channels, micro data, macro data, message packaging, client segmentation, customer acquisition and retention, integration of customer product life-cycle, dynamic pricing, recommendations and use of recommendation engines, referrals or affiliates, supply chain data, fault finding, defects reduction, storage capacity and labour planning, sales forecasting, lead time predictions, delivery time promise, analytics, marketing and advertising, historical data the basis of BD, start-up, BD, competition, pattern identification.

Table 7.2: Findings of business interviews

RQ 1 & 2 and sub-questions	P1, P11	P6	P3, 14	P4, P13, P18	P2, P10	Findings
RQ 1 What are the factors affecting business to leverage Big Data for competitive advantage?	Data as an asset, push profitability, decision-making, forecasting, identify patterns and trends, insight, monitor business.	Marketing, influence communication, messaging, packaging, insight, customer profiling, creation of content, start-up.	Consumer tracking, data is everything, product improvement (software system), decision-making, insight, forward thinking, product life-cycle.	Data super useful, planning (capacity, labour, space management), forecasting, prediction of lead time, outliers detection, improve service delivery, gain insight, customer life-cycle and products life-cycle, fraud detection, personalisation, clear goal for curation, faster service delivery time, decision-making, warehouse planning.	The need for data as a decision resource, decentralisation of data and a lot of unstructured data, data in silos, company is a start-up and BI is leading to test out new ideas, this helps us to immediately see what is working and what is not, better decision with analysing curated data, quick decision-making, better to analyse data as a whole rather than decentralised, there are users on the site and their behaviour, origin, target needs to be known .	Enterprise asset, decision-making, forecasting, patterns and trends (insight), segmented clientele, customer profiling, targeted messaging, content management, product improvement, next generation products (product improvement), customer life-cycle, product life-cycle, the need for data as a decision resource.
SQ 1.1 What is business doing to leverage Big Data to gain a competitive advantage?	Bring in automated systems, improve data integrity by reducing manual data processing.	Impersonating customers, driving traffic to site, customer profiling, customer segmentation, targeted messaging, emails, get more Facebook likes, create better content.	Log all relevant data, consumer data, data to answer questions, data to improve product (system), track customer usage of system across pages, store transactional data and consumer data to GA and Magento.	Create needs list, log clickstream data, have clear data goal, but not curating BD now.	Build and architecture data structures for data centralisation, collecting data to analyse user behaviour on the site, integrating the three data sources.	Centralise data, provisioning data, data self-service, investing in data personnel, create and modify needs list.
SQ 1.2 What is the business's view of Big Data in terms of competitive advantage?	Data drives everything, important for insights, product segmentation, direct merchandising into profitable situations-	Marketing spend, traffic, retail cycle, pages accessed by customers. What do they spend? Are they happy to come again? Competitive	Asset for decision-making, an incredibly important resource to answer questions, Enterprise's most important asset. Yes.	Data is everything, resource for planning, predictions, provide insight to reduce errors or defects, provide better service, trace defects,	Data use not to be so important but now it is at the centre stage (the core focus), data allows us to see what is working and what does not work, there is a big advantage to curating data,	Data is super important as an asset across the board, every decision made should be based on data, like what to stock, based on what customers like and dislike, insight into customer

RQ 1 & 2 and sub-questions	P1, P11	P6	P3, 14	P4, P13, P18	P2, P10	Findings
	that is everything, data helps meet customer needs and push profitability. No BD.	advantage, unique selling point, having insight means less room for errors, high propensity to sell once you track life-cycles, self-service data super essential, social media is now, so making the data readily available for self service is better, that is an enterprise asset, it drives decision-making and generation of insight. Uncertain but yes and none if it is available to me.		the entire back-end depends on what gets ordered the most, the arrangement of stock depends on the movement of stock, place them at the front or far back, capacity planning, how much labour is needed depends on data from the site, clicks, orders that come, and what time they come in, based on that I plan how much labour is needed at a particular time, technology power house, very important asset. No, later answered yes.	being a start-up we are at ground zero, after centralisation we then move to Big Data and competitive advantage, data is everything, especially an asset for monitoring business No, later answered yes.	struggles, conversion data, customer behaviour, click through, data is source of information for capacity planning; data is everything, technology power house, enterprise asset. Yes and No.
SQ 1.3 What kind of data is being curated as part of Big Data?	Metrics-oriented data.	Customer and Social Media data.	System logs, all kinds of data.	AB test results, data to show what works and what does not.	Data to provide insight on marketing, campaigns, statistics, user behaviour.	Data to run business, get and remain competitive, data to confirm and identify patterns in the form of consumer transaction data and consumer behaviour data, AB test results.
SQ 1.4 What are the policies and strategies for leveraging Big Data?	None at all.	Nothing	Lack of knowledge of privacy laws, no particular data curation policies, no curation framework.	No Big Data, no policy.	None at all, no data curation framework.	No defined policies and strategies.

RQ 1 & 2 and sub-questions	P1, P11	P6	P3, 14	P4, P13, P18	P2, P10	Findings
SQ 1.5 What information does business want to get from data?	Data to evaluate marketing effectiveness, KPIs (targets for 3 to 5years, visited pages, views, unique visits, number of orders, very much sales based data, cost per clicks, conversion rate, and FB likes, marketing targets, traffic sources, GMV), Budget formulation, best seller categorisation, products selection, compromised data quality, manual data handling.	Enriched data with relevant interconnecting fields, data leading to unique selling points (CP), business monitoring decisions like spend on traffic and campaigns and advertising but fragmented data, inconsistent and inaccurate data.	Consumer tracking information, information to answer questions, customer information, whether to continue with sections of the system or not based on results and usage patterns, not curating data to the extent we should, there is more room for improvement.	Browsing information, consumer struggles, conversion data, click through patterns, likes and dislikes, whether to flag an order as fraud or not, labour allocation, space management, the entire back-end depends on what gets ordered the most, the arrangement of stock depends on the movement of stock—place at front or far back, capacity planning, how much labour is needed depends on data from the site, clicks, orders that come in, what time do they come in? Based on that I plan how much labour is needed at a particular time, lack of access to data.	User behaviour data to customise marketing, data to analyse users and what they are doing, where have they come from? What exposure do they have to marketing prior to buying? The entire running of the business depends on data, data fragmented across many different servers.	Answers to needs list, KPI - oriented data, competition germane data, budget formulation, best seller categorisation, products selection, business monitoring decisions such as spend on traffic and campaigns and advertising, system performance, back-end monitoring.
RQ 2 How can BD be leveraged in a media organisation to gain competitive advantage?	Get data to align with buying and purchasing to push up profitability, buying determines profitability, ensure the two variables are close.	Integrate customer & product life-cycle with targeted messaging, improve in-page analytics to get micro & macro data for insight in creating content, curate to foster brand affinity.	Glenn data to transform product returns into a profit centre, collate all customer communications and make them readily available at all times to on-duty members.	Critical to have a clear data goal, identify other opportunities to use Big Data, centralise data.		Clear data goal, needs list aligned to data curation, AB testing, create a forward plan to meet targets, centralise data.

RQ 1 & 2 and sub-questions	P1, P11	P6	P3, 14	P4, P13, P18	P2, P10	Findings
SQ 2.1 How can BD be utilised to gain a competitive advantage?	Use BD to direct next generation products of business.	Use AB testing results to direct curation, visual communication is super essential for marketing, social media is reactive—act fast.	Improve communication with customers by defining standards as directed by insight from data.	The entire back-end depends on what gets ordered the most, the arrangement of stock depends on the movement of stock—place at front or far back, capacity planning, how much labour is needed depends on data from the site, clicks, orders that come, what time do they come in? planning the back-end.	Marketing is expensive, Big Data will help us market effectively as a start-up, we will know when to run campaigns, and where to spend money and when not to, better communication with customer, marketing tells you where your traffic is coming from, information on visual experience.	Act quickly on data.
SQ 2.2 How can a business implement Big Data curation?	Develop predictive analytics solutions to identify and prioritise potential customers, best sellers, and opportunities which will aid in developing next generation products.	Do not use data to generate hypothesis, rather use to confirm patterns and trends.	Invest in data personnel.	Identify what data is being generated and how this niche could be exploited to the benefit of the business.		Investing in Big Data strategies, Big Data systems and up-skill personnel to get Big Data aware, develop an organisational culture that promotes Big Data.

Table 7.3: Summary of business interview findings

Question No.	Question	Findings
RQ 1	**What are the factors affecting business to leverage Big Data for competitive advantage?**	
SQ 1.1	What is business doing to leverage Big Data to gain a competitive advantage?	**Finding 10:** Business Management identifies what data to log per department's need for information. **Finding 11:** Spree logs varied data about business entities using the Magento system, Google Analytics and On the Dot (OTD).
SQ 1.2	What is the business's view of Big Data in terms of competitive advantage?	**Finding 1:** Data is an extremely important enterprise asset. **Finding 2:** Data is critical for analysis and decision-making. **Finding 3:** Data in Supply Chain and Customer Service falls under inbound or outbound data. **Finding 4:** There are conflicting and opposing opinions about Spree having BD. **Finding 5:** Operational data users are uncertain about Spree curating BD. **Finding 6:** There is valuable (BD) data that is outside the reach of Spree. **Finding 7:** There is a gap in communication between data curators and data users. **Finding 8:** Many data operations are still on a manual level increasing the propensity for errors. **Finding 9:** There is no historical data for decision-making as Spree is a start-up. **Finding 12:** Business is profitable and competitive when customer buying patterns and business buying are aligned. **Finding 13:** Transaction and transaction-related data are required for analysis into buying. **Finding 14:** Insightful business operation demands analysing enriched data. **Finding 15:** Merchandising formulates budget; budget formulation is based on market place events and data insight. **Finding 16:** Merchandising uses current available information from reports to plan and make decisions. **Finding 17:** Business Management evaluates system patronage and the success of project generates business data and curates data for decisions from logged data.

Question No.	Question	Findings
SQ 1.3	What kind of data is being curated as part of Big Data?	**Finding 18:** Many varied datasets (especially data needed for operation) are being collected as part of data curation, including transactional and transaction-related data. **Finding 19:** In the supply chain there are variations (config. and simples) which involve many SKUs with complex interaction fields. **Finding 20:** Supply Chain and Customer Care uses historical data to predict product availability, future lead times and stock arrival. **Finding 21:** Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data.
SQ 1.4	What are the policies and strategies for leveraging Big Data?	**Finding 22:** There are no documented policies and strategies for data curation.
SQ 1.5	What information does business want to get from data?	**Finding 23:** Business micro and macro data, though hind sighted, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business. **Finding 24:** Data driven decision-making warrants profitability and productivity. **Finding 25:** Business Management creates a budget (top line figure), which is handed down to Merchandising; Merchandising then re-budgets creating budget sub-categories from the top line figures. **Finding 26:** With the budget, Merchandising determines the number of items per variations (config. and simple) to buy. **Finding 27:** Data is spread across multiple servers outside the perimeters of the business, making the data inaccessible for analysis. **Finding 28:** The way data is stored across the different servers may increase the difficulty in bringing the data together for analysis. **Finding 29:** Manual data handling may make data prone to errors, thereby compromising the quality of data for decision-making. **Finding 30:** There is a lack of consensus as to whether Spree has BD or is curating BD. **Finding 31:** The lack of historical data lessens the generation of insights and patterns. **Finding 32:** Lack of customer segmentation may reduce the impact of messages as messages must be vague.
RQ 2	**How can BD be leveraged in a media organisation to gain competitive advantage?**	
SQ 2	How can BD be utilised to gain a competitive advantage?	**Finding 33:** Customer returns data is a source of rich data to optimise customer care and support. **Finding 34:** Leveraging data insight to create content will improve brand affinity and customer awareness. **Finding 35:** It is critical to have a clear BD goal. **Finding 37:** Marketing is expensive but BD will help the organisation market effectively. **Finding 36:** Social media is reactive hence actionable data is needed for fast action.
SQ 2.2	How can business implement BD curation?	**Finding 45:** Business Intelligence (BI) is in the process of building and creating data structures to centralise data for analysis to gain insight into curated data.

Question No.	Question	Findings
		Finding 55: From a technical point of view there are no documented policies, strategies, frameworks and curation models, except a Magento ERD diagram.

Table 7.4: Technical participant response indicating unclear status of Big Data

	P10	P14	P17	P15, P16	P4	Findings
SQ 1.2 Is Spree curating Big Data?	No	Yes	Yes	Yes/No	No	Big Data status unclear.

Table 7.5: Findings from Technical interviews

RQ 1 & 2 and sub-questions	P10	P 14	P17	P15, P16	P4	Findings
Q1 What are the factors affecting business to leverage BD for competitive advantage?	Data is an asset, start-up company, decision-making, decentralise data, support business critical success factors.	Tracking customer and know if our site is doing well by testing out new functionality, AB testing, multivariate test.	Decision-making			Data is an asset, start-up company, decision-making, decentralise data, support business critical success factors, tracking customer and know if our site is doing well, decision-making.
SQ 1.1 What is business doing to leverage Big Data to gain a competitive advantage?	Not yet there, but centralising data sources.	At inception documenting all needed data and implementing logging.	Collecting data to ask questions of data, an increase in traffic to site does not mean convert everything.	Currently business has Magento and Google Analytics, great platforms for data curation, the information source.	Centralising data sources into a unified one.	Not yet there but centralising data sources, collecting data to ask questions, at inception document all needed data and implement logging, ask relevant questions for insight, everything, facts.
SQ 1.2 What is the business's view of Big Data in terms of competitive advantage?	Data is an asset, provides insight for when to run marketing campaigns, source to understand the customer base knowing where they come from, data is everything, data that	Big source of differentiation is customer data insight and following up right with the customer, Asset , data will help push profitability and meet customer needs,	Provide insight as to how to spend money on marketing, yes there is BD.	Data helps the organisation to stand out, yes/no.	Enriched data insight provides opportunity to optimise process and accomplish CSF, data is facts we as a business need for diverse reasons, No BD.	Data is an asset, provides insight into when to run marketing campaigns, source to understand the customer base, knowing where they come from, big source of differentiation is the customer and communicating with the

RQ 1 & 2 and sub-questions	P10	P 14	P17	P15, P16	P4	Findings
	help measure targets, and communicate profitability, No BD.	yes there is BD.				customer, provide insight as to how to spend data on marketing, data that helps measure targets, asset, everything, yes and no BD.
SQ 1.3 What are the policies and strategies for leveraging Big Data?	No documented policies and strategies yet.	None	None	None	None	No documented policies.
SQ 1.4 What information does business want to get from Big Data?	Metrics, stats, KPI-oriented data (targets for 3 to 5 years; pages, views, unique visits, number of orders, very much sales based, cost per clicks, conversion rate, and the likes, marketing targets, traffic sources, GMV).	Data to track customers for marketing and campaigns.	Data to determine how to spend on marketing.	Data relating to all entities in the system, logs, metadata, system logs.	Traffic, returns data, identification of source of error to improve, this is for Six Sigma and quality of service.	Metrics, stats, KPI-oriented data, data to track customers for marketing and campaigns, data to determine how to spend on marketing.
SQ 1.5 What kind of data is being curated as part of Big Data?	Granular transactional and transaction-oriented data, marketing spend decisions, buying, decentralised and fragmented data.	Customer data, data about transactions, all decisions from buying, marketing, manual data handling increasing propensity for errors.	All stats including business relevant data, determine which parts of the site are doing well and how to optimise, clickstream data, decentralised data, data not unified and not in one single view.	We have access to web stream data, stock arrangement depends on stock movement, planning.	Micro and macro e-commerce retail data, almost every business decision has to consult data first, lack of historical data.	Granular transaction and transaction-oriented data, customer data, all stats including business relevant data, marketing spend decisions, buying, all decisions from buying, marketing, determine which parts of the site are doing well, decentralised and fragmented data, manual data handling increasing propensity for errors, data not accurate.

RQ 1 & 2 and sub-questions	P10	P 14	P17	P15, P16	P4	Findings
RQ 2 **How can Big Data be utilised to gain a competitive advantage?**	We are not there yet.	Drawing insight to make decisions.	We are not there, data is very small.	Improve design of systems.	Always have clear goals, use BD to promote automation of every kind, identify fraud and intercept transaction, use BD to personalise, improve delivery time.	We are not there yet, drawing insight to make decisions, we are not there, data is very small.
SQ 2.1 **How can a business implement BD curation?**	We have started by centralising out data stores, with BD we are at ground zero.	Getting the right systems and automating.	Getting infrastructure and knowing what data to collect.	Align the business model to data curation and system architecture, document and publicise available data.	Enforce the need for BD, make data an enterprise-wide responsibility.	We have started by centralising out data stores, with BD we are at ground zero, getting the right systems and automating, getting infrastructure and knowing what data to collect.
SQ 2.2 **How will BD curation contribute to the growth of Spree?**	Gives Spree a huge advantage over competitors, good opportunity for analysis but we are not there yet, opportunity to analyse users and their behaviour on the site, are customers exposed to other sites and prices before making a purchase?	Make us competitive.	Improve the system.	Establish the essence of data across the enterprise, show or celebrate every successful decision, successful using data.	Data on the back-end controls everything from capacity planning to warehouse and product arrangement.	Gives Spree a huge advantage over competitors, good opportunity for analysis but we are not there yet, opportunity to analyse users and their behaviour on the site, are customers exposed to other sites and prices before making a purchase?, make us competitive, Improve the system.

Table 7.6: Summary of technical interview findings

Question No.	Question	Findings
RQ 1	**What are the factors affecting business to leverage BD for competitive advantage?**	
SQ 1.1	What is business doing to leverage Big Data to gain a competitive advantage?	**Finding 38:** Spree is a start-up and not functioning within the realm of BD yet. **Finding 39:** The need for data for decision-making is mentioned through all departments and by all interview participants. **Finding 40:** The business is a start-up so not much historical data exists. **Finding 41:** There is a need for real-time data but current systems and implementation only provision data one day late. **Finding 42:** Data is an enterprise asset needed for business monitoring and steering. **Finding 43:** Big Data is an extension of data. **Finding 44:** Interview participants who are aware of BD but are operating with a limited definition of BD. **Finding 45:** Business Intelligence (BI) is in the process of building and creating data structures to centralise data for analysis to gain insight into curated data. **Finding 48:** Marketing is at a point of wanting to analyse customer, product and sales data for better insights for better marketing. **Finding 49:** Departmental use and need for information mature and evolve at different stages and levels.
SQ 1.2	What is the business's view of Big Data in terms of competitive advantage?	**Finding 50:** Campaigns are initiated to create awareness which then propels more traffic to the site. **Finding 51:** Customer base insight is necessary in achieving results from campaigns. **Finding 52:** Curating data to improve service delivery is pivotal to achieving the Six Sigma goal. **Finding 53:** Next generation products may come from managing customer returns with insight. **Finding 54:** Supply Chain and Customer Care may trace errors and sources of errors or defects using historical data.
SQ 1.3	What kind of data is being curated as part of Big Data?	**Finding 55:** From a technical point of view there are no documented policies, strategies, frameworks or curation models except a Magento ERD diagram.
SQ 1.4	What are the policies and strategies for leveraging Big Data?	**Finding 56:** Business aims to centralise data as a means to better answer key business questions.

Question No.	Question	Findings
SQ 1.5	What information does business want to get from data?	**Finding 57:** Spree is collecting clickstream data and BD is an extension of data. **Finding 58:** Decision-making based on data is applicable to every facet of business. **Finding 59:** Data quality may be compromised due to human intervention or manual data handling, reducing the accuracy of findings.
RQ 2	How can BD be leveraged in a media organisation to gain competitive advantage?	
SQ 2.1	How can a business implement BD curation?	**Finding 60:** There are no plans of BD integration as of yet; Spree is not there yet.
SQ 2.2	How will BD curation contribute to the growth of Spree?	**Finding 61:** Big Data curation gives the curator a competitive edge.

Table 7.7: Research questions and related findings

Question	Finding	Theme	Category
Pre-interview question	**Pre-IQ Finding 1:** Data usage, application, complexity and deployment vary by department. **Pre-IQ Finding 2:** All participants have experience with data-driven decision-making. **Pre-IQ Finding 3:** All interview participants acknowledge data is an important enterprise asset. **Pre-IQ Finding 4:** There are nineteen magazines in Naspers. **Pre-IQ Finding 5:** Nine are affiliated to Spree with more to join soon to form an affiliate business model. **Pre-IQ Finding 6:** The nine magazine companies have massive data that is inaccessible to Spree. **Pre-IQ Finding 7:** Data curation from a curator point of view is divided into technical and business curatorial processes. **Pre-IQ Finding 8:** Business Management conceptualises and contextualises business cases to identify the business side of proposed business idea.	Data Customer Business model	
RQ 1	**What is business doing to leverage Big Data to gain a competitive advantage?**		
SQ 1.1	**Finding 7:** There is a gap in communication between data curators and data users. **Finding 8:** Many data operations are still on a manual level increasing the propensity for errors. **Finding 10:** Business Management identifies what data to log per department's need for information. **Finding 11:** Spree logs varied data about business entities using the Magento system, Google Analytics and On the Dot (OTD). **Finding 18:** Many varied datasets (especially data needed for operation) are being collected as part of data curation, including transactional and transaction-related data. **Finding 27:** Data is spread across multiple servers outside the perimeters of the business, making the data inaccessible for analysis. **Finding 28:** The way data is stored across the different servers may increase the difficulty in bringing the data together for analysis. **Finding 29:** Manual data handling may make data prone to errors, thereby compromising the quality of data which may affect decision-making. **Finding 38:** Spree is a start-up and not functioning within the realm of BD yet.	Data Analysis Decision-making Competitive advantage Strategy Business model Marketing sales	

Question	Finding	Theme	Category
	Finding 42: Data is an enterprise asset that needs business monitoring and steering. **Finding 43:** Big Data is an extension of data. **Finding 45:** Business Intelligence (BI) is in the process of building and creating data structures to centralise data for analysis to gain insight into curated data. **Finding 47:** Formal data modelling processes aimed at identifying information need is done by Business Management who may not be in possession of the needed expertise or knowledge of other departments' core functionality. **Finding 50:** Campaigns are initiated to create awareness which then propels more traffic to the site. **Finding 56:** Business aims to centralise data as a means to better answer key business questions. **Finding 57:** Spree is collecting clickstream data and BD is an extension of data. **Finding 60:** There are no plans of BD integration as of yet; Spree is not there yet.		
SQ 1.2	**Finding 1:** Data is an extremely important enterprise asset. **Finding 2:** Data is critical for analysis and decision-making. **Finding 3:** Data in Supply Chain and Customer Service falls under inbound or outbound data. **Finding 4:** There are conflicting and opposing opinions about Spree having BD. **Finding 5:** Operational data users are uncertainty about Spree curating BD. **Finding 6:** There is valuable (BD) data that is outside the reach of Spree. **Finding 30:** There is a lack of consensus as to whether Spree has BD or is curating BD. **Finding 7:** There is a gap in communication between data curators and data users. **Finding 8:** Many data operations are still on a manual level increasing the propensity for errors. **Finding 39:** The need for data for decision-making is mentioned through all departments and by all interview participants. **Finding 40:** The business is a start-up so not much historical data exist. **Finding 44:** Participants aware of BD operate with a limited definition of the phenomenon. **Finding 43:** Big Data is an extension of data.	Data Analysis Competitive advantage	
SQ 1.3	**Finding 12:** Business is profitable and competitive when customer buying patterns and business buying are aligned. **Finding 13:** Transaction and transaction-related data are required for analysis into buying patterns. **Finding 14:** Insightful business operation demands analysing enriched data.	Sales Strategy Analysis	

Question	Finding	Theme	Category
	Finding 15: Merchandising formulates budget; budget formulation is based on market place events and data insight. **Finding 16:** Merchandising uses current available information from reports to plan and make decisions. **Finding 17:** Business Management evaluates system patronage and the success of a project generates business data and curates data for decisions from logged data. **Finding 52:** Curating data to improve service delivery is pivotal to archiving the Six Sigma goal. **Finding 53:** Next generation products insight comes from managing customer returns with insight. **Finding 50:** Campaigns are initiated to create awareness which then propels more traffic to the site.	Decision-making Planning Business model Service Products	
SQ 1.4	**Finding 18:** Many varied datasets (especially data needed for operation) are being collected as part of data curation which includes transactional and transaction-related data. **Finding 20:** Supply Chain and Customer Care uses historical data to predict product availability, future lead times and stock arrival. **Finding 21:** Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data. **Finding 22:** There are no documented policies and strategies for data curation. **Finding 55:** From a technical point of view there are no documented policies, strategies, frameworks, curation model except a Magento ERD diagram.	Data Strategy Analysis Competitive advantage	
SQ 1.5	**Finding 16:** Merchandising uses current available information from reports to plan and make decisions. **Finding 17:** Business Management evaluates system patronage and the success of project generates business data and curates data for decisions from logged data. **Finding 19:** In the supply chain there are variations (config and simples) which involve many SKUs with complex interaction fields. **Finding 22:** There are no documented policies and strategies for data curation. **Finding 23:** Business micro and macro data, though hind sighted, incomplete, inaccurate and disjointed, provides the basis for steering and monitoring business. **Finding 27:** Data is spread across multiple servers outside the perimeters of the business, making the data inaccessible for analysis. **Finding 28:** The way data is stored across the different servers may increase the difficult in bringing the data together for analysis. **Finding 29:** Manual data handling may make data prone to errors, thereby compromising the quality of data for decision-making.	Strategy Analysis Data Decision-making Competitive advantage Products Strategy	

Question	Finding	Theme	Category
	Finding 30: There is a lack of consensus as to whether Spree has BD or is curating BD. **Finding 32:** Lack of customer segmentation may reduce the impact of messages as messages must be vague. **Finding 46:** Traditional marketing is expensive. **Finding 49:** Departmental use and need for information mature and evolve at different stages and levels. **Finding 51:** Customer base insight is necessary in achieving results from campaigns. **Finding 53:** Next generation products may come from managing customer returns with insight. **Finding 54:** Supply Chain and Customer Care may trace errors and sources of errors or defects using historical data. **Finding 57:** Spree is collecting weblog data (clickstream data) and BD is an extension of data. **Finding 58:** Decision-making based on data is applicable to every facet of business.		
RQ 2			
SQ 2.1	**Finding 33:** Customer returns data is a source of rich data to optimise customer care and support. **Finding 35:** It is critical to have a clear BD goal. **Finding 36:** Social media is reactive hence actionable data is needed for fast action.	Data Customer Data Products Planning Decision-making	
SQ 2.2	**Finding 9:** There is no historical data for decision-making as Spree is a start-up. **Finding 12:** Business is profitable and competitive when customer buying patterns and business buying are aligned. **Finding 13:** Transaction and transaction-related data are required for analysis into buying. **Finding 14:** Insightful business operation demands analysing enriched data. **Finding 20:** Supply Chain and Customer Care uses historical data to predict product availability, future lead times and stock arrival. **Finding 21:** Supply Chain and Customer Care may evaluate patterns of demand for the different categories using historical data. **Finding 24:** Data driven decision-making warrants profitability and productivity.	Data analytics Competitive advantage	

Question	Finding	Theme	Category
	Finding 25: Business Management creates a budget (top line figure), which is handed down to Merchandising; Merchandising then re-budgets creating budget sub-categories from the top line figures. **Finding 26:** With the budget, Merchandising determines number of items per variations (config and simple) to buy. **Finding 31:** The lack of historical data lessens the generation of insights and patterns. **Finding 32:** Lack of customer segmentation may reduce the impact of messages as messages must be vague. **Finding 33**: Customer returns data is a source of rich data to optimise customer care and support. **Finding 34:** Leveraging data insight to create content will improve brand affinity and customer awareness. **Finding 37:** Marketing is expensive but BD will help the organisation market effectively. **Finding 41:** There is a need for real-time data but current systems and implementation only provision data one day late. **Finding 48:** Marketing is at a point of wanting to analyse customers, products and sales data for better insights for better marketing. **Finding 52:** Curating data to improve service delivery is pivotal to achieving the Six Sigma goal. **Finding 59:** Data quality may be compromised due to human intervention or manual data handling, reducing the accuracy of findings. **Finding 61:** Big Data curation gives the curator a competitive edge.		

Table 7.8: Diggs polyglot persistence architecture terms and descriptions

Term	Description
Facebook connect	Allows user of Facebook and Digg to connect their accounts; this allows the added benefit of sharing web content across.
Digg dialog	Allows a dialog among users.
Digg Bar	Allows access to diverse functionalities without having to leave page.
Digg API	Software developers can write code to interface with Digg through this API.
Digg APP	Follows close integration with other sites; allows users to browse Digg's content.

Table 7.9: Sample Interview Supply Chain and Customer Service division

Interviewer	Cecil Nartey
Interview number	Interview number 4
Interviewer	Lekha Bhargavi
Moderator	Dr Andre de la Harpe
Date and Time	14 June 2013. 12:00pm
Venue	Black and white room, 19th floor, Naspers building
Employment capacity	Head of Supply Chain and Customer Service division
Company and Department	Spree, Media24
Consent form signed	Yes
Follow up message sent	Yes
Interviewee background and profile	**Interjection: What is your role in the organisation?** I don't work with Touchlab, I work with Spree. Touchlab was the innovation hub, we pulled out Spree from there and Spree is in its own world. I joined Spree. I basically head the Supply Chain and Customer Service division. I have been here for about 6 weeks now. I joined Spree at the beginning of May. I was at Amazon up until 2 months ago prior to Spree. I have been with Amazon for the past 13 years, that is, 2000 to 2013. I moved to India for the last 3 years and wanted to move back to the states before this. came up.

	Interjection: What did you do at Amazon? At Amazon I did a whole lot of things, but largely data involved. In Amazon I was in Supply Chain as well. I joined Amazon as an analyst there 13 years ago. When I left I was leading the Forecasting division, when I left Amazon Seattle. So when I left I was leading the Global Forecasting world. So what Global Forecasting does is, we use BD easily works. So what global forecasting does, we are forecasting for 2 million products on a daily basis, using historical data. The amount of Items that we are producing forecast for is, when I say 2 million in the catalogue. No all of them are actively selling, but there is still a million that we are actively producing xyz in time forecast and the kind of products vary across 70 to 90 different product lines. So we are talking forecasting for books, which is the first product line that everybody associates Amazon with. We also sell things like tractors and tools and aircrafts supplies in a parallel/pile. A parallel is the most challenging when it comes to forecasting accurately. So through my time there I have done different things, Supply Chain, Procurement and Fraud. I have done My last three years I have spent in Fraud. In Amazon you find that data is huge and across the board and the definition of BD consist of the amount of clickstream data that we get. Fraud uses all the technologies that apply in the BD world. Largely machine learning techniques and you can use real-time on the systems that's applied across different groups. So Forecasting used machine learning, Fraud uses statistical analysis as well as machine learning techniques. And the system that we build in India, which is part of Fraud. So I was part of Fraud called Abuse, which was my domain, unlike the forecasting before. Prior to that of course, a lot of contribution on Finance.
Interviewee request form	Signed
Question 1	**What is your tactical and strategic involvement with data?** At Spree now, today the data that we use within Spree is used largely on my end is on Supply Chain and Customer Service end. I will separate it as Inbound data, which is basically from the suppliers world, you have all the SKUs. Just to be clear, none of this can be classified as BD at this point. It is very small amounts of data. Because you are actually talking about SKUs in thousands and it can be contained within an Excel spreadsheet so it's really not BD. That does not mean that the interactions between the data fields are not complex. So you have everything going on from when a purchase order is placed and you have thousands of titles being created, so all those come to us. So one of the biggest challenges could be the variation as I call them, which is config to simple, which is one of the biggest challenges in the apparel world where you have one virtual product and associated to that you have multiple children and those are the once that the customer sees. So that's just a catalogue as I call it. The next set of information is related to suppliers, that is purchase order data, so that will be everything related to how much suppliers will charge us for it, invoices, purchase order details about when deliveries will come. The reason I am saying this fields are important is because this will lead to the next step, which is through its currently not done in Spree yet. But we want to get to a point where we can plan for capacity, products that are coming in. plan in a way to tell the customer that yes we expect this product at this time.

	This is a regular time and coming from Amazon what we use to do, we will look at historical data, that's when the BD stuffs come in, we apply models based on historical data to predict based on what you have available and what the future lead times are going to be. And based on that we make promises to the customer, you have delivery date promises even when the product is not in stock. So that's when the when date will come in. So you have catalog, purchase orders, the next is you move on to products coming into the warehouse and that's when the receipt date comes in which is sort of a related that's when the inbound world comes in. on the outbound, SKUs become the link between inbound and outbound is where you are shipping customer orders to the customer, that's also supply chain thing that that also leads to use figuring out this is the pattern of demand, for these set of categories, for these set of catalog and based on that you plan a lot of things. You plan for what you want to buy in the future, for products you also plan for do we have sufficient space in the warehouse for these kinds of products because this is what we are expecting to come in, these are the customers that ordering, obviously we are bringing in more categories. So for instance in Spree's world today we only do cloths but we have plans to do toys in a couple of months from now. So we have to know that toys in terms of data the dimensions are very different. Right, so it does not fit into the regular space. So again data we collect from the suppliers we build it up, we the data stream, so again that's the data that we get from the outbound world. Customer service contacts come, so metric around customer service so, so like I said this is not an old business. This is the difference between Amazon and Spree right, we have about 30 to 100 customer contacts or rather 200 contacts in a week, by contacts I mean email, phone calls etc. At Amazon we're talking millions of contacts in a week. We have customer service associates all over the world, thousands and thousands of them. So the metric used there literally defects per million opportunities. The goal was a Six Sigma goal. The Six Sigma states that the quality you're trying to achieve is less than or equal to 3.4 defects per million opportunities, that's the goal. So for every million orders you send by orders I mean transactions with customer, you are allowed 3.4 defects. That's the goal but it could be 90 but that's significantly better than 1 million for every 80. I don't know right now we are at 50, when I classify something as a defect. It could be a: • customer calling for a return, • customer receiving an incorrect item, • customer receiving a damaged good, • customer calling and saying hey I was displeased with your item because your courier did not deliver there or here. So a defect can be classified in different ways. That's the other data I collect from the customer service point of view and my goal will be to look at the data and drive it down. And that's super important goal. If I have to work backwards from the customer I start there and say ok this is the customer did not like so and we can go all the way to the first pipe line as a catalogue right, a customer receive and incorrect item could mean the way we set up our catalog is incorrect. So when you ask me about. What data streams we use I use all of them.

Question 2	**So would you say Spree is not curating Big Data?** I will not classify this as BD. No BD, to me by definition, you want to talk about terabytes of data but because we are collecting so much clickstream information right, clickstream information at some point we want to start looking at that, to me that will fall closer to BD world so BD sort of starts from small data. Because I can handle within my Excel control I can still apply the same techniques as I use in BD. But it's not worth it. You need that complex techniques that you use on BD for the amount of data we handle today like today's world of Spree but that does not mean we should not, we still use some of the basic rules like data quality issues. In normal statistical technique, I rem first at Amazon, I rem outliers detection we check out the habits data client, when we got to BD when we started on machine techniques there was not a single data field that was ignored. So everything had some information in it. So outliers gave us as much information as the average. So you didn't really chuck out data in the world of machine learning, you just threw everything into the same engine like random forest engine or decision support system (DSS), we use these 2 models big time but there were teams that used neural networks, carts which where regression engines, so there we engines that you could just throw the data, get the data back review the output and review that these does not belong here, the point was you glen the information before you chuck the data out you did not eliminate the data upfront. You did it afterwards, those kinds of things are easily those techniques are done in Spree. I can do it now; I don't need to throw it into a random forest engine. To get the data. I can do it in Excel right so you can still apply some of the techniques but we are not in the realm of BD right now I think. Let's say we get to 5000 orders per day. Hopefully we will; then we start slowly moving into that realm. The area that I will say that BD does exist in Spree is magazines, so magazines is our close partner but my world does not directly interact with that, I don't care where the orders come from. When the orders come in I jump in but the front end has a lot to do with magazines; its all over the place—that's the place where we can lavish BD because the customer base is huge, the world has been around for a long time, there is a whole lot of data in the database to work with. The demographic data exist, the article data exist, what people like, don't like exist so you can actually build a whole catalogue based on likes and dislikes on customers looking at print data what I don't know is how that data is stored. That is very critical for analysis. The biggest problem that I have seen that the data is not stored right. So you want to be able to access the data. If you not able to access the data, its kind of pointless. So for magazines you should speak to Krishna about it.
Question 3	**With your experience and expertise, what will you say are some of the factors that affect Spree in leveraging BD?** In the context of Spree if I am limited to that the area where we have BD to go forward is magazine information. We not there yet but these two are the areas that I think we can leverage. On the other hand, the existing business is for instance put all the data on servers which is what I see right now. Like a Spree server not that easy to bring that information out so you need to be able to build, figure out build features, out of that dataset, first of all by features, you are creating attributes right now the way its spread the way its stored on servers across the place like a Sarie server is very different from other servers, so if you want to get similar datasets across magazines, it's going to be a hellish job. So that could be the first factor impeding BD usage because if you g*et* all

	the data together, imagine the amount of data you have. You have like 17 magazines in the world of Naspers. You just pull all the data. You have want to play around with right there. Even if you get one year of BD across this 17 magazines. Ok let's say start with 7. It's still big, so the first thing I think is being able to get access to that data—simply access putting in an area or format that it can be accessed. Even if it's like in the form of flat files. They are easy to handle, like sheets everywhere which is being stored today. On the flip side, if the Spree data itself, the clickstream data, that will become usable faster because of the way we are string it behind the scenes, like it's still within the dev world, parsing the data might be hard but you know that's possible because it's in simple data bases and flat files behind the scene so you can actually get to the data faster. So since you built the Bi dashboard for magazines, you should have some idea of where the data is today. Clearly its more accessible do you feel? **Interjection:** The data was everywhere, dispersed. We need to centralise now we have the data mart up and running with SSIS. We pull the data into the data mart, so is in a usable format as BD concept.
Question 4	**In your experience with Amazon and with BD, are the problems you experienced at Amazon different from what you experience her now?** Amazon is slightly different, BD as a concept has been there, but it is only coming to walk in the last 5 to 6 years so the problems are kind of similar but because Amazon is a technology company at heart, most or everything was built by engineers behind the scenes. So actually it was friendlier for engineers, so engineers could get to the data much faster, the usage of it starts there. So in that regard alone it was slightly simple but the same problem with different data, building data marts, building Hadoop, basically so when you have large amounts of data using Hadoop to spread the data, parallel the data, data computation but its slightly different in the sense that you are actually dealing with BD problems. You are dealing with: "How do I compute this large amount of data at the same time? The problem is not where the data is from, the data exist but now I need to operate on it. We are at the point where we need the data at one place but by the time we start using the data, the data already exist, we got the next problem. How do I manage the data? Today I feel like we still are not at the point where we have a data mart built out completely yet at Spree, so once a data mart is built out then I will say ok, managing becomes a problem.
Question 5	**What about the quality of data?** The quality is a problem everywhere, so it's a problem at Amazon, it's a problem at Spree and how that is resolved is again relentlessly. In the beginning there were a lot of quality issues, so the idea was so its automation at the end of the day that reduces or resolves the problem. The entire performance dashboards that were built to cater for quality issues in the data stream but in the beginning we were talking this same exact problems as here.

	So cleansing the data is a big problem everywhere, so we use the same techniques that you use to cleanse data here. I remember my first mentor at Amazon about 13 years ago, he used Excel. At some point it was Excel, we were content with Excel, we used forecast engines in Excel and all that and he spent 40 per cent of his time just looking at the data, cleaning the data up. That was a significant amount of time then but over the course of years, we made sure that, that part was the cleansing happened, but happened automatically. We moved on to detecting or to systems that will call out these are the issues, move them out so we clean the data as we go along the pipeline, so the manual effect of cleansing the data reduced significantly. And the keywords are to make sure we are capturing it upfront. Like I said, do not lose the information contained in the data, so should not mean that you're losing the information. You are putting it into the feed buckets, you still looking at the data but not ignoring it to say you trimming. Trimming is a classic example you just trim the top *x* per cent. You're losing information there. It means let the entire data go through and take the top *x* per cent. **Interjection:** I think that's an important point you making there. Instinctively we think cleansing means you're throwing things away. With BD it becomes more important. First of all, like you said, what I think the quality of data is like garbage-in-garbage-out syndrome that not what I mean, doesn't matter whether its huge amounts of data, I mean, when it's key that important is as tight as possible. So know what you are letting into the system, you know when there is an issuer as it comes out; that's why manual entry, you have lots of checks and balances in place so by the time it gets to the point of analysis you letting it through. In Fraud it becomes a huge thing, because in Fraud you are dealing with needles in haystacks if you are looking at millions of orders and only a small percentage of it is fraud, so you are literally hunting for those outliers in a mountain. Those outliers mean something; in fact, they are the only ones that mean something to us the way you operate on data in that world is very different and you can't let go of any data stream that tells you a story. So this is why it's super important that, if you are not confident of what you are letting into the system, then you waste a lot of time when you analyse the data, so all we did is that the imports are controlled. This order came in, it came in at this time blah… so you know exactly what caused that difference to happen you will not allow manual entries into the system.
Question 6	**What were some of the policies that were in place for leveraging Big Data?** **Comment: Explain curating? Data management through its life-cycle of beneficence.** Ok curate scares me, I will use manage because we are not curating at all. The best way to explain that will be examples. Data is used like I said was used across the board at Amazon, the few that I was familiar with was the fraud world, the affiliate world, at Spree as well when magazines help function as an affiliate Amazon has a huge associate model. There are hundreds of thousands of associates who can allow put Amazon link and when they click through they get a sale, so again we get to be information-based on the click they get a share, and based on what they are paying, we make change to our model because it's a hit to our final bottom line so all this are areas where BD is stored. So I am going to try and pick examples and see if I answer your question there. So let's say fraud, customer orders, mostly there are places in distributed databases for managing it in a way again I will follow your question, if you say BD.

	I am assuming you're talking about eh world of analytics, that's can be used in multiple ways. The idea is that you are using that to produce some analytical information back into the system for personalising or for throwing up a fraud flag. At eh time and at this space. There is always an end goal of using BD, so in the world of fraud it was to make sure we are able to identity fraud in in time and block that order at the time and in place so the BD analytics was actually done with that goal in mind. Similarly, that's personalisation, that's a huge area where BD was used and then it was used to make sure that the customers get recommendations based on what they have bought and their friends have bought. But we are using huge amounts of data but to analyse it. But the output to the start. So as you are working with BD, that goal needs to be very clear because you can get lost in the world of BD and lose sight of what the goal is and sometimes its four different teams working on the same data with 4 different goals in mind. That's ok but you can't expect that you are just going to have this petabytes of data thrown at you and you can come up with a list of things to do at the same time. At the end of the day its still data common sense, you still have to make sure that the data does not distract you from your goal. So for that reason at a high level the data will be stored in distributed systems, and distributed systems because there is always redundancy built in there were teams like fraud teams, there were besides the actual database that they use they also had flat files because a lot of the BD techniques, a lot of machine learning techniques that are built are using flat files and this works much faster. Hadoop was used, cloud was used, but the data started getting stored in the cloud a lot more. So there is an entire and cloud storage offering… basically we use the cloud a lot to access the data and that's when Hadoop came in bug time, I am not sure if you know Hadoop. So it was used for parallel data and then at the very we also had to use this for tagging the data, to the order there is a whole lot of redundancy built in, there is also something that the customer sees, web services, that's where thousands of servers, 100s of thousands of servers are viewing the data at the same time. So ever order that's coming in has to be literally resolved in 2 milliseconds because that's the order that behind the scenes is using BD but the time it gets to the front end and responds back. You literally had 2 milliseconds to respond back, so that's ….. the we operated on. The BD was used to come up with the results but the results were stored, in the system framework was built a way that, it would respond back very quickly, so a lot of the technologies that they used, technologies that were specifically built to address the speed of computation, so that happened a lot. You operated with parallel data so your results were very quick so what will take days should take minutes, so every thousands order you get feeding into your casting set your validation set and its actually running its running as many times as possible during the day in some cases almost real-time. So the random forest engine for instance was built was real-time, is an engine built behind the scenes that that use petabytes of data, but what was used are was just the final output, let's say if there are four cycles, it will use cycle four for real-time computation. The analytics team will work on the remaining set and feed the new orders coming in through that engine and calculate the end results. On a daily basis, typically it was on a daily basis, in effect curating or leveraging BD gives you the opportunity to deal with many problems at the same time but with different goals in mind. Meaning that it makes you competitive. Data is power. **Correction: you just said dealing with the same problem allows you to deal with multiple problems using the same set of data? That is try, but not necessarily the goal.** I could just use order data for instance, each order let's say has 80 elements. Maybe I just need one tag, something as fraud or not, so I am using just one goal, the idea is I am using 100 per cent of the data, to be as accurate as possible, so BD allows me to be very more accurate much faster.

Question 7	**What are some of the decisions that we need to make using BD?** Good question, we should think of BD as an extension of BD; every decision we make should be data driven. Some decisions include what to stock. It starts from there, which I said depends on the customer likes and dislikes, their browsing patterns, think of that as another decisions because the data element is so small, there is tones of information we can get from how the customer struggles on the site, conversion data, how long does the customer stay on the site, and what did they click through to. This is just an example, I mean everything can be driven by data, so starting with what to store where to put it on the site how to market it is a BD driven industry. The entire back-end actually depending on what products gets ordered the most, this is just an example I will perhaps move stuff from the warehouse to let's say a forward location that takes longer or faster if this is what's going to come on a daily basis. Capacity planning, how much labour I plan for, that depends again on data that comes from studying the website. Let's say I am looking at the website clicks and most of the orders are places between lunch time and two, I will have more labour planned between two and four, so I can get more out so if your think about it. Nearly every single thing needs data, nothing can you do without data. Just to expand, where are the opportunities to use BD like at Spree? Web stream data is the biggest that we have that is something we can call as direct applications, but in my mind there is nothing you can't do without data. Big Data is the extension of data and this the bottom line, people think BD is sitting free at the north pole, but it's just here and that's what's valuable. Big Data to me implies technology power house, in the world that I am used to we were users of BD, we also built, I was there when Hadoop was launched, I know what happened before and the day that people started using it at Amazon, other people used other technologies but at Amazon, that became. We built in-house tools to manage BD. What could be useful to you, I think, is for people to recognise that the data that we have is super useful. There is a framework that is built in in the world that makes it easy for none technology savvy people like none engineers to begin using it. For instance, like what's available in the web services world, Hadoop world makes it easy for people to with very basic knowledge on how to pull in the framework. Say I have all this data, it can be just flat files sitting somewhere or I can take all the Excel flat files sitting somewhere— now I can throw in all the BD technologies into it because the framework for anybody to use, its free, its open source and they can just start applying it and making sense of it. But the critical thing is to have a goal. At the end of the day you must have an idea what you want to use. The data for the answer is simple; you can use it for everything. So anything you want to do, you can apply it. I have a lot of friends that kick off start-ups—they left Amazon; they all started with their knowledge because they come from Amazon. And they are comfortable with dealing with data. Like my friends who is the pay data which is huge, foreign engines and payment engine, the people who use advertising data to be able to promote the right advertisement. To the customer for any website, I remember this is actually for media clients that one of my friends built, because media, take for instance Naspers, they have to know which space and what space will give them the most click return, and that BD that that they are talking about can change that for you. What is most important is speed and performance.

	You can run days and days of analysis, but into hours of analysis, you want to be able to get BD. Data is not everything; it was always there for centuries. It's just the way we are dealing with the data. We know the technology to analyse and the results in seconds if you chose to. That's what BD means to me—BD became powerful; people realised that we have a framework and technology in place that allows us to use it. Amazon has had this data for 15 to 20 years but only started using it like the last 6 to 7 years because that's when the technology got built; it was a result we started looking at it. So that's what makes BD useful. Because think about the world, the world always had these data, but it makes it so much easier. **Recommendation:** why don't you think about eh way BD can be utilised because you should just think that it's there and don't listen to me when I say Spree does not have, so it's that right, it's just and extension. It's important.

Table 7.10: Sample Interview Business Intelligence

Interviewer	Cecil Nartey
Interview number	Interview number 4
Interviewers	Sheldon Redman and Allan Mcluckie
Date and Time	02 April 2013, 11:00am
Venue	Black and white room, 19th floor, Naspers building
Employment capacity	BI manager Spree (Allan), Data analyst (Sheldon)
Company and Department	Media24, Spree, Marketing
Consent form signed	Yes
Follow up message sent	Yes
Interviewee background and profile	Allan came from Mocality, Media24, and Sheldon was with Sarie as a data analyst from the magazine's inception.
Question 1	**What is your tactical and strategic involvement with data?** Allan: I think we work with a lot of the tactical and strategic data on a daily basis, and my involvement will be building and architecting a data store to centralise and store the organisation's data. Sheldon: We are all directly involved with the data. In terms of what type of data we work with, our data is very decentralised and unstructured, and that's why we are trying to centralise it.
Question 2	**What types of data do you work with?** Allan: mostly structured data and very decentralised.
Question 3	**How will improving the timely availability of data improve decision-making?** Sheldon: We are pretty much a start-up still and we are constantly testing new ideas, and we immediately get to see how they are doing. That's why we are almost very much looking for real-time information. The best we can get now is one days delay. We would like it more real than that to make decision right away. And change anything that we are testing.

	Allan: obviously the more time you have with data, the more you can analyse it, the better decisions you can make so obviously the quicker we can get data, the quicker we can make decisions.
Question 4	**What data challenges do you have?** Allan: Most of the data is decentralised, so this makes it difficult to analyse. It will be better to analyse data as a whole instead of different, little pieces of data.
Question 5	**What is data to Touchlab or how important is data to Touchlab?** **Interjection: And why is that?** Because we have come to understand that any decision you make, you will make a better decision if it's backed with data. Data brings transformation; it helps you realise what's working and what's not working.
Question 6	**Is there anything like bad data, dirty or inaccurate data?** Sheldon: Specifically for us at Touchlab yes there is, we get data from multiple sources and these data don't correlate. And because of that, you spend your time trying to find out why this data is inaccurate rather than spending your time analysing the data. As to data being bad, I think its more the format it comes in, is it in a format you can work with. And then talking about data, the biggest thing for me is reliability and accuracy. I mean, we are getting data from three different sources and they are given use three different values so what do you choose? Allan: I don't think there is such a thing as bad data. I think it's how you interpret that and how you are looking at that data, that you are making your decisions. Like, you have two reports coming from the same system and both are given different sales results. I don't think it's the data, I think it's the way the reports are being written. They are inaccurate, they not comparing apples with apples.
Question 7	**What are the companies KPIs?** Our KPI's are our targets, like three to five year plans. Currently we are on a 1one to two year plan. Our KPIs, which are in the form of page views, unique visits, number of orders, very much sales based. Cost per clicks, conversion rate and the likes. Marketing targets. Note: look at the daily report for the targets. Traffic source, GMV

Question 8	**How will data curation contribute to the growth of this company?** Sheldon: this relates to the question asked earlier, what do we use data for? It's like right now we are a start-up and our data is the only thing that tells us what works and what did not work. Should we continue with this or should we discontinue?.
Question 9	**What is the businesses view of Big Data in terms of competitive advantage?** As a business we do realise that there is a huge advantage, a great opportunity in analysing Big Data, which we are not there yet, looking that we are a start-up. **Interject: if we are not there yet? What is the business doing to get us there?** Well, currently we are working on centralising our data stores. Currently we are working on our operational reports, talking about Big Data we are ground zero, where we are still working on trying to get there. And then secondly, we will look at Big Data and competitive advantage as a second priority.
Question 10	**Assuming we should have a Big Data curation platform implemented, what are some of the policies you think will be in place.** We are not there yet, so we don't have an answer to that. **Interjection: Ok, assuming we should eliminate the word big and only work with data, what are some of the policies that you foresee being in place?** **Further explanation to justify why this question by interviewer:** With Magento we are able to track clickstream on our web site and that is considered part of Big Data, especially with the implementation of the Magento database. Allan: I think we have Big Data, we are recording Big Data, and Big Data is our Instagram feeds, Facebook campaigns, a lot of social media, and a lot of structured data. I think Big Data is our structured data and unstructured data. We are recording all that information; we are just not at the stage where we can be analysing it. I think its record now and then analyse at a later stage. Sheldon: I think if we are to think ahead, a big part of it will come to granular level so it comes to specific users, analysing specific users and what their behaviors is. All of that stuff. Where did they come from, what are they doing on our site, are they exposed to multiple marketing sources before making a purchase, do they ever make a purchase, its very much user behaviour.

Question 11	**What data curation modules do we have in place?** I guess based on our current environment, our data curation modules I will say is Magento, and also Google Analytics, and Bookmaster. Those three systems are our core systems. And our challenge will be integrating those three systems.
Question 12	**Does the organisation have a data curation framework as a guideline for data curation?** Allan: No I don't think.
Question 13	**How often does the organisation do an audit of data so the organisation has a single view of data?** Sheldon: We are a start-up; we have not gone through a full financial year yet. A month ago we had one person working on reporting; now we have three.
Question 14	**One a scale of 1 to 10 how would you say the data available to you is compliant to the 4 C's (Compliant, Consistent, Current, Clean)** Sheldon: putting a number to it, it's definitely going to be a low number. Our data is current; *consistency* is the biggest issue for us; *clean* will be quite low as well. As to the data, we don't have multiple copies of multiple things. The thing with consistent, if we are actually looking at the data or the reports—that we get. I think the data is not inaccurate but rather the report that is created. The reports don't pull the data correctly. **Interjection: But isn't the report just pulling the data that it finds there?** It is, but it's not pulling the right information, it's not coded up correctly. I mean, if I compare sales in Bookmaster and sales in Magento, we will have discrepancies. I think the information is there for us to pull the information and work with it but at the moment the way we are working with it, it's not consistent. Given consistent 5 is actually high. If we are talking about the reports I will give it a 1. Because they are not consistent. It is something that Louna also brought up. Consistency within the Magento reporting, it is something that we are aware of. Particularly being a start-up, even when Winston was doing the reports. She was also worried about the consistency of data.

Question 15	**Some data curators subscribe to right time than real-time.** I think there is a lot of time to reading data, the fact that you have, does not mean that you are looking at it. Even with Krishna, certain of his data, he has access to it on a daily basis but he does not look at the monthly data on a daily basis. He looks at that once or twice a month. So I think there is a right time to the data. Sheldon: Well something I have been thinking of is we have access to a lot of data. We could send all the information daily. It will become overloading information, people don't need it. Deciding what data we have and how to break it up on a daily basis and monthly basis, which needs it on different time frame. What type of information, top line reporting and more granular stuff, like marketing so we have to break it up between those factors. What's needed, who needs it and what's needed? And data that's needed might not be needed in six months' time. So it's what is needed, who needs it and why.
Question 16	**What kind of decisions need to be made using Big Data to foster being competitive and remaining competitive?** Allan: It's everything from marketing to when to run campaigns and statistics, its crucial in our start-up environment. It helps us know where to spend money and when to spend money. You know we are buying a lot of traffic out there, mainly just to find the most cost effective way of getting our name out there building our brand. Sheldon: I think the decisions around communicating with the customer—we need to understand the customer and then around that communicating with the customer. Because marketing tells you where your traffic is coming from, how much you're paying for it and you can relate the cost to revenue. Is it effective or not? But if you look at streams that you're not doing well, how do you understand why you're not doing well? Do you send out surveys to understand your customer profile? Especially for us where a big part differentiation is customer service. What needs have we met? Are people like the visual experience? I think we will like to do an analysis on what is selling; we will like to know what is selling, are people buying too many shirts or shoes, is belts our biggest seller? Then we need to buy more belts. Those are all decisions that need to be driven by the data, in order to make those decisions. So it affects every part of our business. I think particularly with the spending of money.
Question 17	**Analysis paralysis?** Allan: I think my understanding of that is, when you have so much information, you battle to make decisions. You are inundated with information. I think we don't have that problem; we definitely don't have too much information at this stage to make our decision-making too much difficult. We have the opposite where we don't have enough information available. **Interject: How about a situation where you know a decision needs to be made but you don't have the kind of information you need in order to make that decision?**

	Sheldon: I think when it comes to that, we have loads of data, because the setup we have, it's the timing and the way we put the data that creates problems for us. It's a very manual process. The problems that we have stem from that.
Question 18	**In your capacity as a data analyst, how will you advice we preclude analysis paralysis?** Sheldon: I think it's just what we are doing, centralise out data, the biggest part of that, is knowing we are working with accurate data, you don't have to confirm it. You can just go on and analyse it. As long as you're analysing and you feel it's accurate to base decisions on it. If you have the information and you are not going to base decisions on it, you won't have analysis paralysis. Allan: I think also it's there is consistency with your data, like if you're looking at one version of your data like sales and everyone knows that sales is sales across the board and we are not getting two or three different views of sales from different sources, it will make decision-making very easier. I don't think when Krishna sits there and makes his decisions he has too much problem with too much data. The problem is us getting it from all this different sources and trying to get it and give it to him. So there are two different ways of looking at it. We've got enough data but when the actual guys are making the decisions, they are not overwhelmed with too much data on their hands.
Question 19	**Relative to destination thinking, how do you relate data and results—statistical results and identifiable results?** Sheldon: I am not sure how to answer that. But like we have been saying, we don't get to being able to analyse our data. Talking about statistical relationships, that all comes with analysing your data. In my mind it's all about seeing trends. Having the data relating to statistical relationships, having the data in the right format to identify trend and relating information from different sections to each other. Across marketing, sales, across merchandising. And that's how you have to relate the information. In the long-term, once you have more data and history then you could start looking at statistical relationships. Especially when you want to do model building, I don't think you can do that now when you have a very short history of data and you are a start-up. It's difficult to relate data and identify patterns. Like when you are busy testing one thing, there are ten different things going on at the same time.
Question 20	**What are your preferences with object and data modeling?** **Is there a preference to model data or object separately or together?** Allan: We don't have a preference at this stage (not very sure).

Question 21	**How important is a data model to communicate ideas?** Sheldon: I think it's quite important because any ideas you get needs to be supported with a data, and the data comes from the data model. It makes it easy to paint the picture. Like we are the ones working with the data, we picking up the trend. And we believe something is working. You not just going to say that it's working, you will say here, its working and this are the data to back it up. And that kind of thing especially to support which idea that you have.
Question 22	**How do you relate information to the needs of the organisation?** Sheldon: I think that's answered, that is, testing to see what's working and what is not, relating it; it's about taking that information and turning it into decisions. Allan: I think I have heard people say that don't even bother analysing the data if it's not going to affect the decisions you make. If you are going to carry on making the same decisions you have always made, then don't even bother spending money on this data centralisation strategy.
Question 23	**What happens to old data?** Allan: we definitely don't have any old data, but there might be some data that is archived. Like transaction log. Transaction data, we might reach a stage where we will have to archive it.

Table 7.11: Sample Interview Business Management

Interviewer	Cecil Nartey
Interview number	Interview number 4
Interviewer	Mr Nick Smith
Moderator	Dr Andre de la Harpe
Date and Time	02 May 2013, 11:00am
Venue	Black and white room, 19th floor, Naspers building
Employment capacity	Business manager and subject matter expert, Senior data curator
Company and Department	Spree, Media24
Consent form signed	Yes
Follow up message sent	Yes
Interviewee background and profile	
Interviewee request	From my side what I really want to know or see what we as a company are able to learn from it. I think it's amasing that we have someone living right in the heart of our data and is actually doing academic work. So we as a company will like to get insight that.
Question 1	**What is your tactical and strategic involvement with data?** I work for Spree, Media24 as a business manager. My tactical and strategic involvement with data is to identify as a business (i) what are we going to actually do, (ii), what is the business case, and (iii) what is the business side of an idea? This is all prior to developing a product, and this starts with planning the architecture of the product. What the product is going to look like. What I then do is identify what data to record (capture) because users will be using this product and will generate data in the process. So what bits of that data do we need to store? This decision is made based on being able to do a couple of things, which are to determine: (i) Are people using our product? (ii) Are we making money? (iii) Are we getting the right business data out of this? (iv) Most importantly, collect data on where we are going wrong, especially where users are not using the product the way we expected?

	So my involvement with data strategically and tactically in short is to determine what to log. That is mainly, conceptualisation of application.
Question 2	**Technically, exactly what type of data do you curate?** **Comments: Can you give example? Is it customer data? What exactly? How does the data relate to Touchlab?** When you say data, how exactly, you don't mean the low level string, customer, or intent or something. Customer data is incredibly important to us. We find marketing is really an expensive thing to do. To track customers to our products is really expensive, and once we do that, getting a way of contacting them afterwards is incredibly important. So customer information, things like email address, phone numbers, are really important. So customer information is one of our biggest ones. So second to that, a lot of the curation is to do with data that will ultimately allow us to do final analysis, answer certain questions about whether something is working or not, whether when we added something or change something, it will likely help us make more money or improving a product. So typically in the app and web environment, that is what we will call stream data, what is the user doing? Where are they coming from and where are they going? From which page to what page? **Input (by moderator: So you keep that data?)** Answer: Yes we follow all that. **Input (by moderator: And transactional data?)** Answer: For any business, transactions happen. All of that gets stored. There is transaction data and transaction associated data. This is comprised of customer data, transaction associated data, values, what was bought, where did we attract, acquire that customer from. Did they receive an email, have they shopped. All these get stored in a transactional fact table. **Input (by moderator: So that's really a 360 view?)** Answer: very much so. On that, about what data we store, I think we are very good at storing a lot of data and the reason for that is, they are in our space… Unlike, I suppose, things that are offline, there is a lot of plug and play analytics that you can use. The big one is goggle analytics, which logs absolutely everything about a user being on our website. Even the custom things that we are doing, because we code a lot of our things from the ground up we have the ability to choose what we log and we turn to log a lot. So the type of data is a massive variety. And we build up a lot and store all that data.

Question 3	**Is Touchlab curating Big Data?** **Comments (Curation within the confines of the tool)** The way I understand that question is, we have this massive variety of raw data, are we taking it and abstracting it to a level higher than the raw format and making sense out of it? Or taking it to a higher level? I will say Yes we are. We have a pretty good view on data. We are doing it to a degree but not to the level we should be doing. The level of curation is limited to specific sets of data, what our tools provide us with, what our tools allow us to. Like Google Analytics, it will store every page of our site that the user has been to and we will be able to go in there and say what is the most popular page. How many pages have the user been to? So I will say that's curation but more than that tool; we do very little beyond what that tool provides us. So I will say yes and no. So going out and doing more than that, looking beyond what that tool has to offer, we do a very poor job of it. I say poor in terms of what is possible. I say poor in terms of what other people are doing. I don't think other people are doing a good job of that.
Question 4	**How will data curation contribute to the growth of Touchlab** I think that what we are getting better at doing is using data to make decisions. While we don't actively curate a lot of our custom data. When we need to make a certain business decision we are fairly good at going in there for what is necessary and curating that data to help us make that business logic. I am trying to get a good example for you. On one of the products we are working on. We believe that some section of our website is not useful at all. Our high level view does not give us much of that. So what we have had to do is delve deeper into the raw log data, aggregate all that data and then make that decision based on that. I think that our decisions are very data based. I think the culture here at the managerial level where I have seen in the past, our decisions are made based on gut. We have transitioned to knowing that our decisions must be made based on data. At that we are out to get that data. I don't think the getting of that data is done particularly well. We make a few mistakes in it because we don't understand some basics and some statistical problems and the biases that come with that. So we will go out and curate data but we will do it incorrectly because we don't understand some basics in curating data in a certain way. But typically decisions will be made on data.

Question 5	**Are there any policies or strategies outlined for the curation of Big Data?** Policies, as a media company or particularly any company in South Africa, especially a media company that often interact with customers and often have customer information, we are often worried about privacy laws that are coming in. there is on that side a lack of understanding of the privacy laws and how they impact our ways of using data. And the company as a whole understands that concern that they are bringing about policies to say if you're attracting customers like this you can't market directly to them blah blah. So there are those kinds of policies coming in. But I will say there is no policies like these are how you will structure your data; this is what data you'll collect. I think part of that is because what we do is fairly different. Its fairly difficult to give a restrictive policy because it changes from application to application. But part of it is not having thought about it or putting it into use. **Input (by moderator: There is no framework that you use, as it happens, you move?)** Answer: Yes. **Digression:** This is the boss's toy (interviewee shows interviewer and moderator an app designed for senior management to record or input their Key metrics or Key performance indicators (KPIs)). Probably about year or two ago, Koos said, we have all this companies within Naspers generating massive amounts of data. But we as a business don't have a way of seeing that quickly and getting a finger on the pulse for a business. So they started this dashboard team to build this iPad app. With this, Naspers requested every company within the stable to report their key metrics into the app. So all managers and CEOs report there key metrics into this app. The next review is tomorrow. So at the review meeting you go and meet Koos and report back on your business. How is your business doing? And this is all driven solely from this app. So how this happens is, you've defined your key metrics. It is deliberated upon, normally this will be critiqued, why are you not measuring this? Why are you not measuring that? From there, you'll go out and find the data and import it into this app. So whenever he wants he can see how you're doing. Based on that when you're having a review with him, you have all that data that corresponds to your key metrics in front of you. The next one is a travel app, indicating what's budget and where we are. Our trading profit and that sort of thing. And then we go into the more custom things like how many daily downloads are we getting. What are our active users doing? How many users do we have? How many of our users are returning three or four times in a month and where do we rank in the app store? I can show you newspapers. **Input (moderator: Is this real-time or data gets fed in at the end of the month?)** Answer: Typically how it has worked, data gets fed in at the end of the month. They will export an Excel file and someone will have to copy that data across. What it's moved towards is data automation. Like this is my Google Analytics profile, I am reporting on page views, etc., pull that across making it real-time. It's up to what Google Analytics can provide.

	So that was the first step in automation. The next step was, is to allow a company a file with your information on a file and have a crawler at regular intervals go and crawl the data to dynamically update the app. Personally I have that running to my travel company. I collect data from various sources and aggregate them there and it is consumed from that central location. With that my data is pretty much up to date. I think that's where we are heading. A more real-time view of things. I thought that will be really important. Naspers and Koos Bekker were really forwarding in thinking that Naspers, the people needing to make decisions, may have this data at their disposal. And further, people will come into meetings and when quizzed what are your page views with events happening verses six months ago and they have no idea. But now with this they have the information in front on them. So Naspers at a group level does an incredible job of curating data if you think of Naspers as a company with holdings. It curates the data that it needs incredibly well. That's one part of the story, the other part is what sits in each company and what data is generated there and I think that's what we have not cracked yet. The very high level data is great, but the very low level data we have not cracked yet.
Question 6	**What kinds of decisions to you need to make to be competitive and remain competitive?** **Comment: Remind me of the question again? He's great thinker. Thinking very deep.** A couple of thoughts on this. Sometimes we collect these massive amounts of data and we think we can go into it; that can give us a massive advantage. So we go, we get the data, we build this as part of our product or say people are not converting on our store and we draw right deep into the data we think we have found the reason. We go and make those changes based on what we have seen and it has no impact. It seem like this is really an issue for us. That is being able to find the right actionable things to do using that data. We often think that we can get a competitive edge from looking at data and realise that we have done something that didn't work. So being able to identify what that thing is from the data. The data will give u various options so making the right one is what we don't always get right. In terms of competitive advantage, what I have found amasing with the accessibility of data today is that very frequently you don't just have access to your own information. You also have access to your competitors' information because from a website information is publicly accessible so you can go out and crawl your competitor's information. In the past if you were pick n pay you wanted to see what Woolworths was doing, you will have to walk into their store, have a look around and if you wanted to get any idea of stock or the variety of clothing, you will have to go count each individual thing. So you didn't have particularly good insight into your competitor's data. Whereas now it takes a developer an hour to programmatically go out and find out how many different sample, and styles you have on your competitors website and so in terms of competition that is really an amasing thing and the power of programing and business is being aligned. We get that insight.

	In terms of our own data to create competitive advantage, it really relates to us being able to ask some questions about our products and very quickly get a responds to that so as I said we will say we believe building the same thing for a web site or killing or removing a certain page will allow us to get better and we can very quickly get into that data and very quickly determine if that decision is the right one. **Input (Moderator: But you also have the sense that there is this massive amount of data there and somewhere in there, there must be a competitive advantage that's not popping out, so to investigate what this is, is problematic.)** Answer: You are 100 per cent correct. I think that if we know what we are looking for it is fairly easy to get it out. If we don't and we just looking for something, its very difficult. Potentially there is a disconnect between the people who can go out and look into the data and the people that can identify a competitive advantage. So I will say all of our managers, none of them have the ability to go and query that data. They themselves can't go and check the raw data and put into something that's curated. And then I will say on the other side our programmers who interact with the data all the time and are incredibly familiar with the data have no idea what they should be looking for from the data. So a BI team goes a long way towards doing that and being able to instil in that team what we are looking for business-wise and being able to instil in that team what business in looking for and them having the skill to mine that is incredibly important. For example, a little while back we had no BI at all, so our managers didn't know how to get into the data and our programmers had no idea what our managers were looking for which accentuated the disconnection. So in building a BI team what brought on Sheldon, who had all the know how in statistical science, we got him exposed to the raw data with the help of the programmers, that was great but Sheldon didn't have the IT skills to get in there programmatically to bring the data, meant that he was relying on someone to bring him the data and that's when we found the second person with the programing ability to go out and push that data out. That really pushed us ahead. So now we had the analytical ability and the programing ability combined. Personally I was lucky enough to be able to go out and programmatically get data and I was running the business-side of my product but that worked for me and my little part of what I managed, but no one else was getting that benefit so we have a massive advantage being able to match up the business side and the data side. And I still don't think we are doing a good enough job of it. I think to really do it well, our BI guys should sit close to the programing guys, especially when they are actually developing a product and should be able to assist them by telling them what to log or a format or framework of what to log, because a manager does not understand the data side of it well. They think that they know what they want out of it. They see very high-level stuff. Whereas the BI guy knows I am going to ask some questions, I don't know what they are, I will have to ask a range of questions of the data, a data raw enough for me to be able to get insight out of it, but it's got to be in a way that I can use it easy.
Question 7	**How sensitive do you think the business is towards Big Data curation and getting or extracting knowledge?** So, I think the decision makes this environment know that they can go and get that data, so they go to the BI team or go to the programmer and say I need to know this and right often there is some loss of translation where the person going to do the querying doesn't quite understand what the decision-maker is asking.

	So you get problems from day-to-day but say goes through alright, you can get that insight but the problem with that is not instant. So you can't get it immediately. This on a lot of decisions making it's got to be immediate and or real-time. And second problem is because the manager decision-maker doesn't understand what the data looks like; very often they don't know what the can ask the BI team. E.g. I know we log certain types of data because I sat at product level and I know that you can get for example the screen resolution of the browsers or our websites, so the decision-maker down the line might go, we really don't know how big to make our pictures, but they don't realise that we keep a log of the screen size. Or another example is, we want to know thing about gender, are there are more males on our site than females. And at some point we have taken down the ID numbers. A data person knows that the ID number will give you the gender of some but the person making decisions very often doesn't know that so because they don't know (i) the depth of the data stored, and (ii) what you can often get out of it to know the way you aggregate certain things. That you could do a regression test on something, they don't ask the right questions from the BI person, so two-fold. They can't get back quickly because they don't understand how to query it and they done know what is possible with the data. So many times I have absolutely blown people's minds by just going, oh you have the data sitting right there why don't you go and ask something of it? And then they realise that they have the opportunity to do that. I think to be able to do that you have to understand what makes up the bottom bits of that data, what is getting stored. What types of queries are possible but do you think it's possible for business people like hard core business people who sell and do things to understand and do that? I think with enough exposure, yes. I don't think we can come to them with Big Data tools saying here is a curated set of data, you can get anything you want from this data, query it. I don't think we can expect them to be able to know what to be able to ask. They might be able to tell you a high-level business decision but then, like you said, you'll be missing the nuggets sitting in there. I think the only way that you can really sort that problem out is to make your decision-makers more familiar with the type e of data that is being stored and is successful.
Question 8	**Do you have those people involved in your strategic and operational sessions when you discuss data?** I will say here we typically do more; more frequently we will have that because there is quite a bit of closeness between the person making the decision on the business level and who is determining what the product looks like. But we are very unique in that way and very often I see here when a team gets over six people, the business guy becomes the head of the business and the product manager comes below that and he is very much development focused and not necessarily making the decisions. So a business guy says we need a page that does this and someone below him goes and decides what kind of data gets stored. So typically your decision makes are abstracted from that data strategy and potentially when you building a product to have that decision makes aware of you data strategy. I will say it doesn't happen tightly enough; doing that requires that that manager has some good understanding of data and what can be stored and what is sorted and how it works.

	Question: How accurate do you perceive the data you work with? That is a problem, in that data is often misunderstood and I think the one that bites me the most is from a statistical point of view. I am trying to think of common biases that are introduced here, the biggest one that I see happening is people not understanding the sample size correctly. So they will observe something and then this is how it is because I have seen it like this but then the sample size is so small that there is a misinterpretation of the big picture and so. They don't realise they need to go out and get the bigger sample and so that's what this Big Data should be able to give us is a big sample of what's going on. So, frequently I see people who have no technical understanding making common statistical mistakes from advising perspectives from not having a big enough sample. The other one that I see is correlation and causation. Because we can see these two variables correlate we believe or think that this thing is the cause. So they will go out and help in sorting out the cause but actually because two things correlate does not mean that one is the cause. So just a basic statistical understanding there will allow them to realise, I should regress this against other variables. It's not necessarily a problem with the data itself. Its mistakes with the way it's interpreted. The actual storing of data, I think we do a good job when we store and we have to for many reasons. When you get someone's email address wrong, it causes us big time. Occasionally we make mistakes in the way we pull data and we don't have enough checks in place to say we are counting the number of downloads—are we sure that we are getting the right numbers of downloads? Perhaps there is something in the code that is doubling it. Like, we had a mistake the other day where all of our flight bookings people were making, were being doubled and so our numbers were great but actually it was rather half of that. If someone book for two passengers instead of doubling the price, we quadrupled it, so there is some mistakes at the level. Typically they get found quickly and we store that data without too big a mistake. One last thing, potentially sometimes we don't clean data well enough when we store it. The best example is telephone numbers, for example people putting 082 or 27 or people forgetting the area code or people doing postal code with three numerals instead of four. Typically I find though that we can fix a lot of that by just having some rules in place. Telephone number knows if it starts with a zero, put a 27 and drop the zero. We have a couple of issues around sanitisation but nothing like the deeds office.
Question 9	**How often does the organisation do an audit of data to ascertain that the organisation has a single view of data, i.e. the truth?** I will say we don't often in terms of a lot of the data that we store; we don't go back and question it. We take it as fact. I think we can do that for a lot of it because it is very technical thing where someone is viewed a page and we have logged that page. So unless we have a fundamental problem where there is double page views, counts, or something like that which we will generally pick up because …… *spike*

	Input (You look at exceptional reporting) However we do have a number of checks in place for transactional data so we will have a front-end system that keeps log of stock and how much stock we do have in store so that we can sell it on line. And we do have another system that keeps information about the stock in the warehouse. Then in the warehouse when deliveries take place we have a count there of stock and then after all of that we have a count that we have paid to suppliers that we get invoiced for a certain amount. So those four or so systems have to hook up. That is some kind of effective auditing that happens there. That is like a check on each one when they don't add up. We then can go and query but it is very much exception-based.
Question 10	**How do you know if some gives you his ID number, that it is indeed the right person?** ID numbers are very infrequently needed on online-based offerings. There is no reason to it. I know fore magazines, ABCs which is circulations, ID numbers are required and they often ask for it so they can get the age and gender from it. But I don't think that there is any checking on that from what I understand. They only way to check that is to ask for a copy of the ID and we never do that. But I don't even know if the copy is legit.
Question 11	**And in terms of addresses, do you really care?** If he is getting something delivered our assumption is that if you have some delivered, you will do your best to get the address correct. Other times we will get information from users; we will just have to rely on them in telling the right thing. The way that I turn to try and do it, is ask the user for as little information as possible and then determine the information from the signal. So for example, we don't ask the user where they live but there province is useful to us because we can ask them relevant information based on what province they live in. we just test their cell phone GPS, get a location and then assume that's the persons province. So where possible, if I can take something that is fact because of a signal that is untampered. From a GPS or a network like Vodacom, I will always go for one of those instead of asking the user. Two reasons for that is a mission to have the user to give data I don't want to have to ask, and they don't like giving information. And I know that it is actually correct. I take it from a source other that the user.

Table 7.12: Sample Interview Merchandising

Interviewer	Cecil Nartey
Interview number	Interview number 3
Interviewers	Sherene Conrad
Date and Time	02 April 2013, 2:00pm
Venue	Black and white room, Absa building, 13th floor
Employment capacity	Merchandising planner, subject matter expect
Company and Department	Media24, Spree, Marketing
Consent form signed	Yes
Follow up message sent	Yes
Interviewee background and profile	
Question 1	**What is your tactical and strategic involvement with data?** My role with data especially is to determine budgets, those budgets are decided top line, and it is then handed to me. Then what I do with that which is a purchasing figure, I then break it down into categories and sub categories. That will be on a rand value scale. At a cost excluding vat. It further gets broken down into units required per item, the number of config that you need as well as our average RSPs, costing's. And those are then given to the buyers and then they will go and execute buying towards these budgets. So it's like a shopping list that I formulate. So at the moment we have no history in the business. So now that we are starting Spree, Spree is actually a new entity so it will slowly over time create its own history, plus what the buyers bring as new trends and what is happening in their market place. To formulate their budget based on what is available. And this goes down to every config that you see on the website. So if you see a dress in red blue and green, we have planned to have the three colours. Not necessarily the red blue and green. But we have planned to have three of those dresses. **Interject: that takes care of the tactical side of things. How does this affect the strategic?** Now that budget then becomes an actual, once the buyers attach a product to it, it becomes an actual, it starts feeding through the system. It goes through Magento, Bookmaster. It is then stored there in terms of sales figure. So then what I require of BI is to pull that information, give it to me, and then I analyse that information into our clearance percentages. Using a clearance percentage, I then decide what is a good seller, medium seller and poor seller. What happens then is to top-up our best sellers or base sellers.

	We then put the poor sellers into our discount tab or our daily tab to clear out. So what I do with what you guys send me, is I manipulate it into a report and normalise it …..into a tool to push buyers into a more profitable situation. It's like BI, but more like buying orientated.
Question 2	**Is data available to you in a timely manner for decision-making?** Look, there is a weekly report that gets sent out to me from BI, and then I manipulate it into something like a report. So I can put it into a proper formal report and then on Tuesday. I will present it to the buyers for buying; it comes to me at the beginning of the week. So I will say that data gets to me in a timely manner.
Question 3	**How accurate do you perceive the data that you work with?** I will say it's like a five to six at the moment, purely because of system errors and the data is pulled manually; it's pulled from two sites and it's a very manual process. It's pulled from Magento and Bookmaster; it's then collated and I chop and change it. So a high propensity for errors. In previous work environments, I have that information in pre-set format so the data integrity there is higher because I don't have to manipulate the information manually. So the aim is to get to a point where a system pushes out the data to me and I don't have to manipulate it but rather just make decision. So it's now five to six because its very manual. And there are problems that I sift through as I go through.
Question 4	**How will improving the quality of data improve decision-making?** It will make the decision-making far more confident, like if we see red pampers are the number one seller, there will be no doubt in our mind that it is in fact a good seller. Now, as we sit and sift through data, we ask, is this reasonable, and in a lot of cases in October last year, we had a lot of problems with reports not coming through properly. Purely because it was a manual exercise but at least we have reached a point where we can rely on it to about 90 per cent, but it takes a lot of work to get it to that 90 per cent, so initially it's a five to six and I have to sift it to get it to 95per cent. So what will help me improve the accuracy is if it is a system-generated report. That is the level of manual operation degrades the integrity of the report.
Question 5	**What data challenges do you face?** Because the data that is sent to me is a manually generated data, there is a high propensity for errors. So the integrity of the data is compromised or jeopardise a little bit. That is, there is so much manual data handling which increases the propensity for errors and compromise the quality of data.

Question 6	**What is data to this organisation?** Data is everything, because the minute we don't have read on what is happening with sales and all that, it is virtually impossible to improve or push profitability. Because if we know red dresses will sell we need to buy more red dresses because that's what the customer wants.
Question 7	**In your own words, what do you think business could do to leverage the kind of data that you will need to stand out?** I think it all leads to when are we getting a setup system? And that is a very huge factor. I mean, you can work with the data but a preconfigured system where the bugs are minimal will be the best thing that ever happened to us. It will optimise performance and delivery, especially with the rate that we are projected to grow. The business is not a small business; it will grow into a huge online business.
Question 8	**Have you come across the concept of Big Data?** Yes I mean in all retailers you dealing with Big Data; this is not an uncommon thing for me to deal with because I deal with data upon data upon data, this is what I do. So it all comes to getting a system, so like right now, you need to be so super organised to run on the manual system. And to my understanding, we will run on that for the next couple of months from the numbers side.
	Competition germane data of a business should comply with the four C's, which include Complete, Clean, Current and Consistent. On a scale of 1 to 10 how compliant will you say the data available to you is? 10 being the highest. I will give it five to six, but eventually it should improve like get to nine to ten. As it stands, when I get data, I have to clean it, edit it before it get to a more usable state.
Question 9	**Besides buying and forecasting, what other decisions do you think you will need to make with data in order to make us more competitive?** All relates to buying, there is two sides to buying, it's the purchasing and the customer buying. And we use the reports to match to what our purchasing and what the customer is buying. The closer the match, the more profitable we are, the further the match the less profitable we are, because meaning that the customer does not prefer our goods. That is, they are buying certain things and staying away from others. So if we can get it so close that there is a 10 per cent gap, then we will be super profitable. So if we have bought 20 per cent black dresses and customers are only buying 10 per cent black dresses, then we should rather reduce black dresses.

	So all the decisions we have to make pertain to buying. And that is essentially the business. And the closer we get our purchasing closer to our buying, the more competitive we will be and remain. Because the minute you become unprofitable then that means you're not profitable enough. The whole idea is to get the buying and the purchasing parallel as much as possible. I mean, in a perfect world, the idea is to get it parallel so getting it close as possible is the aim right now. So if you see a massive sale in the market place then that means we have not paralleled it enough and we are taking a known on margining. We have not had a good season. The fact is the core of the business is retail; it does not deviate from the fact that we sell to make a profit, putting it on the website means we are selling. So our buying is driving everything at the moment.
Question 10	**How do you relate data to the needs of the organisation?** Look, the data that I get relates 100 per cent to the entire needs of the business, meaning you cannot operate this business without data; the business will not survive. We will not know if we are profitable or not except from an accounting level, even accountants need this information. I regularly have accountants come to me for this information.

Table 7.13: Sample Interview Social Media and Marketing

Interviewer	Cecil Nartey
Interview number	Interview number 4
Interviewer	Ms Sheri-Lee Carver-Brown
Moderator	Dr Andre de la Harpe
Date and Time	02 may 2013, 11:00am
Venue	Black and white room, 19th floor, Naspers building
Employment capacity	Brand manager and subject matter expert
Company and Department	Media24, Spree, Marketing
Consent form signed	Yes
Follow up message sent	Yes
Interviewee background and profile	Spree is a new online fashion store. We are very close (affiliates) with some of media24 magazines like • Sarie • Fair Lady • Grazia • You • Huisgenoot • Drum and basically the fashion editors select items that our buyers have bought. So if you go on the site, you can say I am a You magazine reader, click on the You magazine and you will see products from the Spree collection that's for You specifically. It's like selecting what we think you will like so I do that social media for that.
Moderator explains his role	My responsibility is to guide him and ensure he is on the right path, so sometimes I will interrupt him rudely. This is to say that this is also a training session. Note: Before the interruption, she said something; I don't know if you picked it. She said finds it complicated. So that's a tip for you to realise that you need to explain more than you had to do with some of your previous interviews like with Nick Smith. Else she might give you exceptions of it and then we may pick up errors.

Question 1	**What is your tactical and strategic involvement with data?** Influence the communication strategies. Specifically regarding social media. At the moment Spree is starting, there isn't any data to work with. It's like we have the Sarie database but you don't want to base any learning on that because that's a different consumer altogether. But basically any data inside that we do get. I will just influence the social media or communication strategy. **Input: In which way?** It will determine which channels we use to communicate, how we package the message to our consumer, what is it that you going to tell them and why it is even relevant to them. It's really being the consumer and packaging the message to them in a different way. At the moment, most the people here were with Sarie so they still have the same customer in mind and that influences their design, how they wrap things in terms of communication, etc., but it's not really relevant to a Spree customer. So I just need to influence that and direct the communication differently. **Input: Who is your typical client?** Female, aged 25 to 39 if not older, its high lsm, living in your metro arrears, having a lot of money to spend and having Internet access.
Question 2	**In an ideal situation where data is readily available to you, what kind of information will you be dealing or working with?** In an ideal world, if I look at the company that I worked before, the Foschini group, which is for Markham (men's wear)—its massive, you had a crazy database available to you. So the insight available to you is better. So your message can be so much more targeted and segmented, instead of saying one thing. If you're saying something to everybody, you have to be so vague to try not to isolate anybody, but if you have insights into the database, your message is a lot more targeted, you know exactly how you can get that message to that person. You can like customise that message to affect them stronger and your results will be so much better. So things like the types of products that purchase for example, how often they purchase, are they frequent purchasers with small baskets size or are they big purchasers? Are they big once off purchasers or twice a year? Are they the end-user or are they purchasing for other people? It's all that type of thing. For example on Mother's day we don't know if we speaking to mothers or people buying for their mothers. So, it's that that of thing. It all about the message. They only way we will lose out, is how to target a message to the right person at the right time at the right tone of voice.

	Input: This is segmentation in an accurate way? **Input: So you don't need data out of the customer profile, like stock data, or image data, like clothing, jewellery?** In terms of content creating we do need that, for example, everything that's on the site that's on the what's new area, we pull through into Facebook, so we will say this has just been unpacked. In the store, Spree store, we definitely need that because that forms the message in terms of stock levels that gets through to the digital marketing lady—the one that does the email sends. So they won't send an email to the database unless they are guaranteed that the product is in the store. So only when they have ticked the box that it gets sent through to me. So it's important, but the customer profile is more important. It does not help if product is available and I don't know what product is relevant to the consumer. Yet again, you don't want to promise what you don't have. **Input: where do you get that data?** From my colleague, and they get it from the database capturing team. I currently pull directly from the data capturing team. In my assumption, I am acting like the consumer. So if it's on the site, I am putting it there, if it's not on the site, it shouldn't be there. Because its now social media, you can't plan for it weeks in advance. It's useless. Social media is very reactive, and you're posting two, three times a day. So an email is more planned and there is one to three a week and the rest you can kind of plan in advance, speak to the data team.
Question 3	**Will you say Spree is currently curating Big Data?** I am sure they are on the back-end but non that is available to me in terms of what I need like customer profiling. For example, in an ideal world, when we spoke to the guys from fab.com, that New York online site, it's almost like, in an ideal world you have this database and you should be able to see what stage of the customers product life-cycle they are in and you will communicate to that person in a different way than you will communicate to a new person; than someone who is been buying from you for years verses a big spender, to someone that likes to window shop. For me, at the moment we don't even know who that person is, for me that's an ideal world, that's what I'd like to see happening. Obviously this is a start-up. **Input: Do you understand the concepts of curation of Big Data?** I will assume that, the curation of Big Data is the collation thereof. **Note: Interviewer explains Big Data by saying that Big Data is data sizes that force curators to look beyond normal data curation processes…**

	Interviewer interjection 1: Is Big Data like traffic sources—that type of thing? **Interviewer interjection 2: Give me another example that's not marking related? Is it stock turnover?** Answer: Let's say for example, Home Affairs. There is 70 million IDs in that database and everyday its fed ID numbers, images, finger printing, pictures and all that, and the management of that is huge Big Data stuff, especially the quality around that which quite sucks, because about two to three years ago there was absolutely no control over that data. Actually, until today it's the same situation. In the case of Kalahari, it is a good example of stock. It initially started off without stock, you need to go and look at what the customers are buying and decide what amounts of stock do we have to carry. **Interview interjection:** Customer call logs, all fall within the brackets of Big Data. Another example of Big Data, rather more towards the beneficence aspect, collecting diverse prices from competitors to enable your organisation get more competitive and proceed with dynamic pricing to land a sale… At the moment in terms of what you're saying about page visits, what is more interesting for me is Google Analytics to see traffic sources for example. It's …, not direct click through or some of the processes are; its more brand awareness. **Interjection by moderator:** Throwing more light on Big Data. Where your Big Data advantage will lie is to get, for example, data from e, DSTV, entertainment, who is watching what at what time in terms of your target market. For example, with Kalahari, who are the people on fashion books, Sarie, and all the other magazines? If they do it right, you should be able to get that data. Responds to interjection: that should be fantastic. It depends on how you ask it, but I think it could be available to you.
Question 4	**Who do you ask to get these things (information)?** Well, in the past the Foschini group had a whole building called Infotech, which sits behind the fashion building, and you literally send an email. It usually takes about 5 days so you build that into your time. So you send the mail to say I am looking for males within this age group that have spent R12 000 in the past, and specify the criteria. I need 50 000 names because I am sending out an SMS. And then five days comes and they will send me this database. So that's what I am saying, like I don't know who exactly we ask from, so like I am saying, that will be fantastic! And that will be with account holders. So if we are able to branch it out to people who are not even branched out yet, that's really awesome. Especially a lot of people look only for fashion, especially like Spree where we are saying it's for the average woman that doesn't always know about fashion. Like, Sal36 is very niche, is very high; Zando is like everything and you have to make your mind up. We are saying you can buy anything on the site and trust that it's ok.

	So it's like for your mum of two kids as well, doesn't have time as well. She is watching Cartoons Network with the kids; its also saying what are they reading, food and what's the latest for them. At the moment the only way we getting information like that is through gorilla tactics like speaking to bloggers and influences and that type of thing and consuming media that our consumer will consume. But its all assumptions and you make your own mind up based on what you think it is. It's not hard fact and you don't know who those people are.
Question 5	**What kinds of decision will Spree have to make to be competitive and remain competitive?** I will say targeted message. Obviously we are online and the big idea is you have to push products. It's just spam. I have unsubscribed from everything. It's irritating, the amount that you get from them, and you can't select what type of communication you get from them. For example, if there was data available to them or to us for example I could say here is her birthday, she is not very fit so don't advertise active wear to her; she works in corporate, fantastic. So if we start putting these tribes of consumers together, you then do targeted marketing directly to them which is guaranteed for them to buy into it. **Interjection: So for marketing, the most important thing is customer segmentation?** Answer: Yes, 100 per cent! It's easy to get the product info, it is what is on the site, for me at least, because I don't want to drive; my core focus is to drive people to the site. And how else do I drive them there for things that aren't there or give them a very different view of what's there or not and my work is too active to go thru the whole process of social media? It will be fatal. They will only go once. So all in all, it's the consumer side of things that matters.
Question 6	**How do you relate information to the needs of the organisation?** Like I said, it's all through segmentation and targeting; it's the ultimate fundamental frontline with regards to that. For me, especially my role, there is so many things that are going to drive sales. Like (i) your paid advertising, (ii) your Google ads, and (iii) your web banners. But brands that make you want to buy things even when your product sucks, there is always customer issues, logistics issues that you don't have control over, any that's going to be your saving grace; is the brand and making people fall in love with the brand, it's the affinity and the brand equity and the only way that I can do that is by knowing our customer and telling them what they want to hear. So it's like your best friend. It's almost like your best friend, personifying it.
Question 7	**How do you get customers to come to your Facebook?** What we do at the moment is a very soft approach and obviously the business is used to the hard approach. Because I feel like its being obnoxious, it makes the brand look desperate. So its brand damaging like, get this now, buy that now. So what we do is make sure that content is king. So we make sure what we have on the site is content credible.

	Interject: So where do you get the content? From our website and from our fashion team. Where do they get it from? It's their opinion. They will look at the London fashion weeks in terms of high level trends, what is trending internationally. The trend for winter's baroque, which is like a shoe, golden, and we have to make sure we pushing that on our platform, so that we are sure to be fashion credible. So as a brand, our equity has increased, so there is a bit of content management. 100 per cent. Without that you become like a Zando where you just post a product and nothing else with it. It's a big part of it. And then what we have done at the moment, we have launched our social media platforms, just Facebook and twitter because there is not use launching platforms and they are not populated until we have launched. Because we need to do photo shoots and all that, so what we did is that fashion week, we launched with the two platforms. And all the trendiest people that we spotted, we gave them a business card that said Spree at the back. It just said "spotted" and our website and we posted all of their photos onto our website. So we spoke to that need of wanting to get spotted. That naughty badge of you stylish and you acknowledge that, that's how on Facebook we have just crossed over 600 in two weeks, and on twitter. We almost at the 200 mark, no paid advertising, its literally getting someone there, letting them tag themselves and seeing this is actually cool and letting them share the content with their friends and referrals and that. So once we have gotten to that point where we have data available to us, then we will be able to do the paid version. So that's how we are building our audience—it's purely through content. ***Additional information from interviewee*** A lot of times, its seems the marketers are having to drive what we need from data, for example, when we were asked what tools, what paid advertising do we do so they can measure the traffic from the source, to determine our ROI, when you give them the platforms like tumbler and blogs, they say but it's not paid so it's not going to be added in terms of measuring traffic, but my point is, it needs to be. It's your highest ROI because you're not paying anything. It's guaranteed, why don't you put your focus on there? If there does, a great understand from the IT and technical guys on the value behind those free marketing tools, the value behind understanding those consumers and not just all other 'nitty gritties'. Sometimes they think you need to pay a rand in order to get a two rand. But you don't have to pay anything. If we can just get the proper data, if we just have the proper insights, we can do so much more. We don't know what's available, we might say we want name, age, address, but like you said, we can tell you who is shopping at Zando and you like flip, you know what we can do with that, so its more working closely and an understanding what's available and you understand what is needed.
Question 8	**Is data available to you in a timely manner?** Most of the time, simply because of Google Analytics and the like, because in the past and I don't know how it will work. If we had this Big Data, what process will in terms of us briefing you what we need or it will be a report but I think with social media, it's like now. So if you give it to us, it will be better than us having to ask it all the time—-self-service.

	Interject: Does it mean much has not gone into conceptualising what needs to be provided to you to deliver to business? The difficulty is that a lot of people don't understand marketing; a lot of marketers don't even understand the different ways of marketing they are used to the conventional TV, full page, half page, and black and white. What does that do for you? You can't even measure it. If people understand what you can pull from it. And we spoke to the guys from fab, like I said that, New York based international shop don't need something, then we don't need it either. When I asked them about the product life-cycle and they laughed and they said they are actually talking about launching it now, incorporating it into their database, it's a Big Database, but you don't know anything about that person and the person has invested so much time and money in you, and you know deadly squat, the customer becomes nothing. You know how it is when you go to a shop and they greet you by your name, you feel like wow. They know me. Whereas they just looked at your card, they know your surname. Brands have lost that and a brand is nothing without that. So I agree, I think its understanding,
Question 9	**Accuracy, I have always given 10 out of 10.** **Do you really need accurate data? Let's say the guy is 39 or 40, does it really matter?** Small things like that, it does not really matter but when it comes to things such as uhm, for example, we use to do new store launches or something that was geographically specific, like you will hate to be in Cape Town and get a competition that is Johannesburg specific. ……..Facebook, is it that important that you cut down to absolute size? I am not talking about email address fashion is such a liquid kind of thing. We are definitely not saying if you 50 years old, we don't want you. That's not what we're saying, but certain information has to be accurate. It just depends on what part of the data it is…….accuracy. Interject: That's why I am asking, because many times, you will from an IT perspective, and you talk accurate, it means 100 per cent, but when you talk business accuracy it does not mean 100 per cent—it's more trends, it's like the average, so it's not necessarily 100 per cent to the decimal.
Question 10	**How will improving the accuracy of data improve your decision-making?** Increase scope for creative thinking. But yet again, when I said accurate, it means not different from what you mean inaccurate, but its important because real accurate data kills innovation but sometimes your aim is to be innovative and its part from the marketers' fault as well. Because a lot of the time marketers to the detriment they use research, using Big Data to prove something, but it shouldn't be. It should just be used to confirm something, so it should not tell them anything new per say. It does not make sense so it's almost like if they say we are targeting white women and they see but oh, majority of SA women are actually black, then they freak and go oh, but we need to target black people because if you're objective.

Question 11	**How will improving the timely availability of data improve your decision-making?** It will make us faster to market—Expedia is available—the faster we can be. This will increase the brand perception and customer relationship. So they can see for example, if we are the first people to wish them a happy birthday, the day before their birthday everyone else jumped on-board. It will seem like we really do care more and make us more competitive. **Interjection: But your timeliness isn't brain surgeon that needs the information now. You know the guy's birthday is next month so you want it so you can plan time, yet you want to be relevant.** I think when it comes to the product side of things, it's like real-time and at least a day or two give or take. But in terms of the other consumer information, a month is fine because it does not change that drastically. The think it does not help if we get the brand, the product and the logistics 100 per cent, but the product is wrong because you don't know what the consumer want to buy and nobody wants to shop from you because its pep version of online retail. So that's why I am saying, a lot of time the marketing side of what is needed is undervalued or underestimated. **Interject: So your look and feel must be good? You can't afford a pep sort of look, that's visual communication. If it's a pep version we just have to make sure we are talking to the pep customer and that's where it kind of gets disjointed.**
Question 12	**What data challenges do you face?** Lack of resource in which to act, we don't have the availability of Big Data, we don't know how to access it. And sometimes we when we do have a lot of it there is no resources in order to do something with it, so its two parts. **Question: What business challenges germane to Big Data do you face?** I said, lack of access with regard to social media and is intangible. So uhm, its almost like what I am trying to access is totally intangible. So explaining to the guy sitting on the business side because I can't tell you how much the audience needs to be, there aren't like set targets in terms of rand, people you can't measure things like that in terms of social media so that's the biggest challenge in terms of explanation.
Question 13	**What is data to this organisation?** Spend, traffic, retail cycle, so when someone goes to the site how long do they spend there, what pages do they access what do they spend, are they happy to come again?

	Competitive advantage? Unique selling point. Is there anything like bad, dirty data? Answer: Yes I assume that will be inconsistent, so I am not sure. In addition, its data that's enriched, verified, making sure that all the fields are available in terms of what's relevant to me. What are the businesses KPI's? Turn over stock, expenses, that type of thing. For me the most important thing is customer services, the product, the people, understanding what exactly they want they will contribute to growth.
Question 14	**How does data relate to competitive advantage?** Knowing the insights means less room for error. There is a high propensity to sell once you track life-cycle, you start picking up trends and for example, now we see we are getting a lot of traffic to the site from social media but not a lot of people are buying. So my feeling tells me that they may be uncomfortable with buying, so they are kind of just window shopping. Why is the increased traffic from social media e-commerce website not culminating in buying? They may be uncomfortable with buying or not have credit cards. And if they are uncomfortable buying, then I need to create content that will make them buy so I tell them the things we are doing well, your credit card details are safe, etc., so for example when we see that people are buying low cost low risk then we can send them communication relating to shoes, but we don't have that yet.
Question 15	**What are the precursors to failure?** My number one is my objective, so if we don't know what we are going to use the data for does not help that its there right, we don't need to do that as well. And also data to confirm hypothesis, not to create one.
Question 16	**Don't you think that if you have this two zillion of data sitting there with all the nuggets, don't you think that looking at that can pop important things to you?** If it ties back to the objective, but what I am always nervous about and what I have seen, rem with marketing the only data I ever used, separate to consumer data, was AMPS or 80 20. It's an online thing, it collects amps and all that starts SA stuff and you can basically find out how many South Africans are aged between 18 to 24 females drink Windhoek or whatever. That's the type of data I have had access to in the past and I found that if you don't have the objectives, you end up knee jerking and you don't want to knee jerk. There are things that will pop up if you can say find this awesome trend through analysing these data by showing that consumers are actually this, then I must say cool, as long as it confirms or supports the objectives that were initially set out.

	Where I feel a lot of the time, a lot of people don't know what those objectives are and they try to look to Big Data for answers and that's where things get. **Interject: if you want to force data to support, you can.** Answer: And that's the scary part; you can just massage it till it works for you and that's what I am not comfortable with, and a lot of agencies do that. So that will come and say this TV channel is launching a new magazine, we want you to buy adverts; this is a fantastic thing about our mags, it reaches this consumers and I am thinking anybody can do that. They charge you quite a bit so that's why I feel strongly about confirming specific tools being used. I just heard about Magento and the only thing that stands out for me is traffic and spend.
Question 17	**Being competitive?** The most important thing to me is customer acquisition verses retention; if we had data in support that, that will be fantastic because its two completely different strategies.
Question 20	**How will that be?** It's almost like all the consumers who have not bought but have been to the site; or consumers like that you get from Zando, who have shopped by you, have been shopping by you, you do loyalty programs and stuff with those verses acquisition where you just trying to get them to sell to build the brand. Whereas the other guys know the brand is awesome, that's why they have bought. So at the moment you trying to communicate to different markets the same message because we don't have a choice, but it will be great if we got to the point where they are separate.

Table 7.14: Emerged research themes

Theme	Context
Data as an asset	An important enterprise asset taking centre stage
Customer	A service beneficiary
Product	Item to be sold
Planning	The act or process of drawing up plans or layouts for some enterprise
Decision-making	The cognitive process of reaching a decision
Competitive advantage	A circumstance that puts an organisation in a superior business position
marketing	The commercial process involved in promoting and selling and distributing a product or service
Service	Work done by the organisation to benefit the customer
Sales	The business of selling goods or services
Analytics	Systematic computational analysis of data
strategy	An elaborate and systematic plan of action

Table 7.15: Summary of response to Spree having Big Data

Participant responds	**Number**	**Frequency in Percentage**
YES	5	28%
NO	9	50%
Uncertain (unaware)	4	22%

ANNEXURE B: Data Analysis

Table 8.1: Pre-interview question and summarised responses by department

Pre-interview question	Interview grouping							
	Interview 1	**Interview 2**	**Interview 3**	**Interview 4**	**Interview 5**	**Interview 6**	**Interview 7**	**Summarising**
Question	**Buying and Merchandising**	**Business Intelligence**	**Business Management**	**Social Media Marketing**	**Marketing and advertising**	**Warehousing and Customer Services**	**Technology**	
What is your strategic and tactical involvement with data?	Determine budget (top line figure), budget formulation, planning, decision-making, bestseller categorisation or identification of best, medium and poor sellers.	Building and architecting data structures, core e-commerce services, centralisation of fragmented data, automation of data curation, reporting, Information or data provision.	Conceptualise and contextualise business, identify business opportunities, business cases, visualise the business side of ideas, identity types of data to store (curation), define data curation process, define KPIs/KPMs, competitive edge/ advantage.	Influence communication strategy, identify communication channels, message packaging, Identify consumer (clientele), customer segmentation, recommendations	Message packaging, get and retain customers, customer segmentation, integration of customer product life-cycle, decision-making, competitive pricing, influence communication strategy.	Supply chain data (inbound and outbound), customer service data, variations, actual , capacity planning, fault finding, reduce defects (defects per million opportunities), customer service affiliations or referrals, customer service contacts, historical data, inbound/outbound data, delivery time promise/ prediction, lead times prediction, historical data. Forecasting.	Dynamic pricing, recommendation ending, suggestion engine, provide data to BI (provision data), make data available (core), code core, functionality for data curation.	Determine budget, planning, decision-making, bestseller identification, build and architect data structures, reporting, conceptualise and contextualise business opportunities, identify apposite data for curation, influence communication strategy, identify communication channels, message packaging, client segmentation, customer retention, integration of customer product life-cycle, dynamic pricing, recommendation engines, referrals, supply chain data, fault finding, defects reduction, capacity planning, sales forecasting, predictions.

Table 8.2: Themes development

Code	Category	Subcaterogy level interaction
DSM	Decision Making	WSM
BAD	Build and Architect data structures/ data or Information Provision	CFD, ADC, CES, DTC,CSY,DSI, DCF, DTP/INP /REP, DCF, REP, RTI
CCB	conceptualize and contextualize business	BUC, VBS,BDO, BUO, DCF
COM	Communication	ICS, ICC,MSP, PKM
CST	Customer / Customer service	CUS, GRC,NCG,KNC/VSS, CBI/CBG,CUI,CSD, CSA, CSE, CNT, PGP
REC	Recommendations	RCE,
PLC	Product life cycle/Products	CPL, VAR, ACT
FOR	Forecasting /Predictions	LTP, DTP,
CMP	Competitive Advantage/Edge	DNP, KPI/KPM, PIP, PFO, BSD, INS, BRA,WTC, KBT, LSM, CNI, AFM
STC	Start-up Company	
BPL	Business Planning	LSP, SSM/SSP, CAP
PTI/PTC	Pattern Identification/Confirmation	POD
MKT	Marketing	ICM, CNM, CCI, MSG, KNC/VSS, PGP, PKM, NCG, BDO, AND, AFM,CUI, CNT, BRA, EMF, CNB, RCI, GRC, CSD
DTA	DATA as an Asset	TYD, ODT, HSD, IOD
COM	Communication	MSG, MKT, ICM, INC, CNM, CCI
QNA	Questions and Answers/ Analysis	PCH/CUT, CTR, NPS, NRT, NDD, CDG, NAQ,
BDC	Big data collected	MDE, RLD, BDD, NBD, DTC, CDD, MVD, SDC, DDG, RAF, CBR, CUT/CBR, IOD, NDD
DTM	Data management	PCH/CUT, CTR, NPS, NRT, NDD, CDG, NAQ,
SER	Service and service delivery (optimization) ****, Customer service	DFR, FAF, CSE, CNT, PGP
FRA	FRAUD	FRD, OUD
	Fragmented data, data silos	
ANA	Analytics	
	Business continuity	
STP	Strategies and policies, Ethical considertaions	
	Return on investment	
	Recommendations engine	

Table 8.3: Establish themes frequency

Row Labels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Occurrence frequency	Count of Interviewees
Competitive edge/advantage	1	2				3		1			3			4	1	1	1		**17**	**9**
Pattern identification	3		1	2		1	1		1	1		4		1	3			3	**21**	**11**
Data as an asset	6	5		1	2	2				6			1			6			**29**	**8**
Data fragmented/isolated/silos		4							3						5			2	**14**	**4**
start-up	3	6	3	1		2		4				3			2				**24**	**8**
analytics			2			1		2			4					5			**14**	**5**
Predictions				3														1	**4**	**2**
Business Planning	4		1	4			7		1	1	1		1						**20**	**8**
Service	1			3		2												3	**9**	**4**
customer	2	2	3	11		9					1	3		3		4			**38**	**9**
Communication		4		3		8			1	5		1	4		1			2	**29**	**9**
Data management or curation				1		1			1	1			1			1		4	**10**	**7**
Decision making	12	9	5	6		1	1	1	1	3		1	1			1			**42**	**12**
Product life cycle						6		1	1		1			1					**10**	**5**
Customer life cycle																				
Marketing and advertizing		2				5	1		1	1			1		1				**12**	**7**
Strategies and policies			1	1					1		1			1					**5**	**5**
Product				1		4		6		1	4	3		3		1	1		**24**	**9**
Segmentation (customer, product, message)			6		4							3				1		5	**19**	**5**
Data Provision		4		5		2			1						7				**19**	**5**
Personel mentoring	1		1		1			2		1				2	1		1	2	**12**	**9**
life cycle(system, product, customer)		1	1		2		2		3				4		1		3		**17**	**8**
Return on Investment						1			1				1				1		**4**	**4**
Z Grand Total	**33**	**41**	**27**	**46**	**14**	**54**	**19**	**17**	**16**	**20**	**15**	**18**	**14**	**15**	**22**	**20**	**7**	**22**	**420**	

Table 8.4: Abridged themes frequency

super categories	Frequency
Competitive edge/advantage	17
Pattern identification	21
Data as an asset	29
Data fragmented/isolated/silos	14
start-up	24
analytics	14
Predictions	4
Business Planning	20
Service	9
customer	38
Communication	29
Data management or curation	10
Decision making	42
Product life cycle	10
Customer life cycle	
Marketing and advertizing	12
Strategies and policies	5
Product	24
Segmentation (customer, product, message)	19
Data Provision	19
Personel mentoring	12
life cycle(system, product, customer)	17
Return on Investment	4

Table 8.5: Data analysis of emerged themes

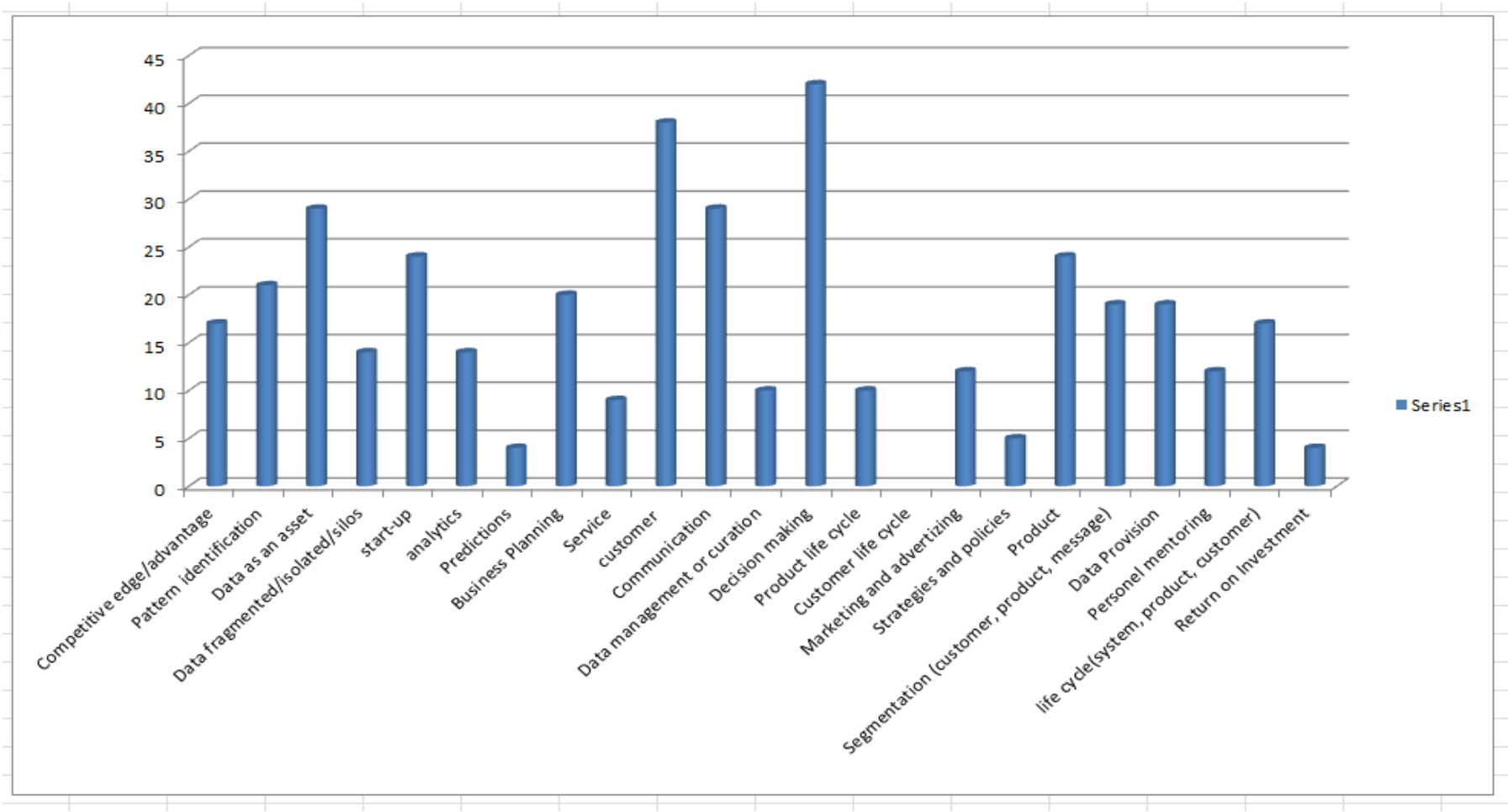
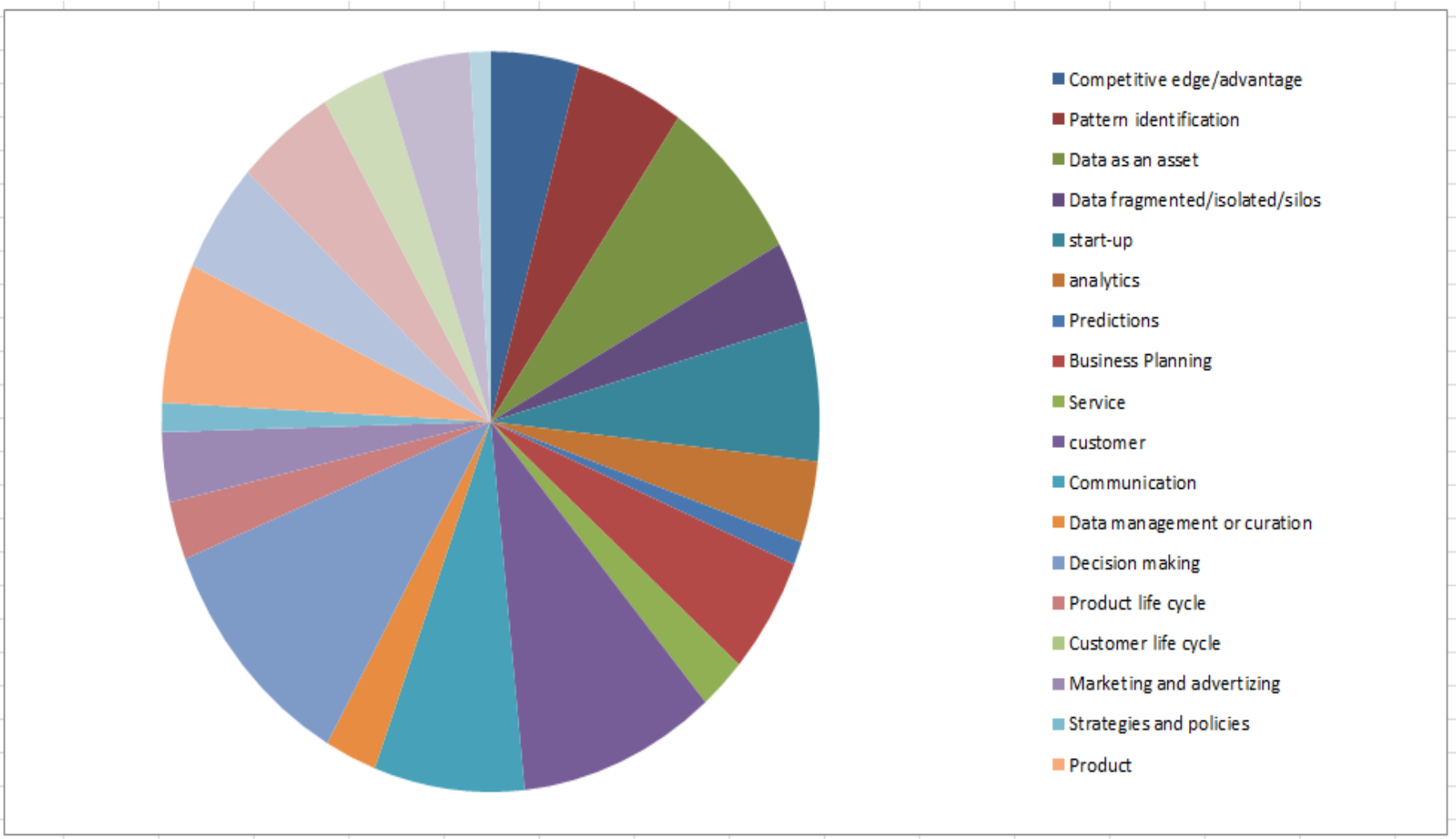


Table 8.6: Pie chart grouping of emerged themes



ANNEXURE C: E-commerce Retail Concepts

Table 9.1: E-commerce retail concepts

Terminology	Description
Cross promotion	The promotion of a website through other traditional forms of advertising which may include newspapers, TV, radio, billboards and magazines.
Distribution channel	A method by which products and services are sold as part of value proposition to customers such as Internet website and retailers.
Offline transaction processing	Capture of order and credit card information for later authorisation and transaction processing through card swipe terminal.
Settlement	Upon shipment of customer purchased goods, sellers are required to key in transaction for settlement at which time the customer's credit card is charged for the transaction and the proceeds deposited into the seller account.
Shipping confirmation	A message notifying customer of shipment in the form of an electronic test or email notification.
Affiliate marketing	A marketing system where other sites and e-commerce stores can sign up to re-sell products of other sites.
Content management	A system of tools and interfaces for management of website content.
Customer service management	Management of customer relationships to facilitate acquiring and retaining through customer information analysis.
Email marketing services	Mass sending of emails to customers as a means of marketing to a seller's customer base.
Inventory management	The management of inventories including ordering, quantities, release-dates.
Mobile commerce system	Systems that can offer sales and promotions on mobile devices.
Payment processing	A system of payment in real-time transactions that enable sellers to accept credit card and electronic check payments.
Search engine marketing	Internet marketing strategies designed to promote visibility.
Search tools	Website specific search systems (interfaces).
Shipping rates	Systems that provide shipping rate information on selected products.
Web analytics	The analysis of user activity on a website to ascertain user behavioural patterns for systems optimisation and promotion of business.
Website performance monitoring	A set of systems that check and ensure that a website is up and working.

ANNEXURE D: Letter of Consent for Interview

Department of Information Technology

Letter of informed consent

Data represents the lowest level of abstraction from which information and knowledge can be derived (Malicke, 2011:15). It forms the building block of information within an organisation allowing for knowledge and facts to be obtained. Drawing from its epistemological essence, data represents the object of knowledge as presented to the mind.

The leveraging of new value streams from Big Data, though initially a source of challenge for many organisations has become the source of competitive advantage in enterprises since the 1980's (Forsyth, 2012; Jacobs, 2009). This is especially true with the confluence of enterprise Information technology (IT), cloud computing, social media, eScience and media in combination with mobility and emerging social trends which are re-shaping the technology industry (Floyer, 2012).

Your participation in the research will be appreciated, as we request your time for participation in a research interview session. The interview will be in a semi-structured format where a set of questions we will be administered in an interview fashion by the researcher. This interview should span about an hour.

Purpose of this study (Objective)

The study is focused on the possible competitive advantage a media organisation stands to benefit curating Big Data, through

a) The proposition of guidelines for the curation of Big Data
b) And recommendations for Big Data curation

The participation in this study is voluntary; respondents are free to withdraw at any time, the interview sessions will be audio recorded, which is subject to the consent of respondents. The audio recording and contents of the interview captured will be strictly used for research purposes only. There will be no risk of personal/emotional/physical/mental/ harm of any kind inflicted by this study, and discussions on sensitive topics will be avoided. The identities and other personal information of the participants will not be disclosed, and no information collected will be accessible beyond the immediate researchers involved.

By signing this letter, the participant acknowledges his or her informed consent as related to the study.

On the completion of the research study, feedback on the findings will be made available to the participants to increase their knowledge of the subject matter and help in future decisions making.

Participant	Organisation	Signature/Date
Researcher	Institution	Signature/Date
Supervisor	Institution	Signature/Date