Real-time decision support systems in a selected big data environment

by

REGIS FADZI MUCHEMWA

Thesis submitted in partial fulfilment of the requirements for the degree

Master of Technology: Business Information Systems

in the Faculty of Business and Management Sciences

at the

CAPE PENINSULA UNIVERSITY OF TECHNOLOGY

Supervisor: Dr Andre de la Harpe

February 2016

DECLARATION

I, Regis Fadzi Muchemwa, declare that the contents of this thesis represent my own unaided work, and that the thesis 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

The emergence of big data (BD) has rendered existing conventional business intelligence (BI) tools inefficient and ineffective for real-time decision support systems (DSS). The inefficiency and ineffectiveness is perceived when business users need to make decisions based on stale and sometimes, incomplete data sets, which potentially leads to slow and poor decision making. In recent years, industry and academia have invented new technologies to process BD such as Hadoop, Spark, in-memory databases and NOSQL databases. The appearance of these new technologies have escalated to an extent, that organisations are faced with the challenge of determining most suitable technologies that are appropriate for real-time DSS requirements. Due to BD still being a new concept, there are no standard guidelines or frameworks available to assist in the evaluation and comparing of BD technologies. This research aims to explore factors that influence the selection of technologies appropriate for real-time DSSs in a BD environment. In addition, it further proposes evaluation criteria that can be used to compare and select these technologies. To achieve this aim, a literature analysis to understand the concept of BD, real-time DSSs and related technologies is conducted. Qualitative as well as quantitative research techniques are used after interviews are conducted with BI experts who have BD knowledge and experience. Experimental research in a computer laboratory is also conducted. The purpose of the interviews is to ascertain which technologies are being used for BD analytics and in addition, which evaluation criteria organisations use when choosing such a technology. Furthermore, a comparative computer laboratory experiment is conducted to compare three tools which run on Hadoop namely; Hive, Impala and Spark. The purpose of the experiment is to test if system performance is different for the three tools when analysing the same data set and the same computer resources. The imipirical results reveals nine main factors which impact the selection of technologies appropriate for real-time DSS in a BD environment, and ten application independent evaluation criteria. Furthermore, the experiment results indicate that system performance in terms of latency, is significantly different among the three tools compared.

I. Table of contents

II. List of Figures

III. List of Tables

IV. List of key words

Business intelligence, DW, ETL , BD analytics, OLAP, MapReduce, Hadoop, in-memory databases, decision support system, massively parallel processing systems, distributed processing, NoSQL databases, BD, benchmark.

V. List of Abbreviations

Abbreviation	Meaning
ACID	Atomicity, Consistency, Isolation and Durability
BASE	Basically Available, Soft-state, Eventually consistent
BD	Big Data
BI	Business Intelligence
CAP	Consistency, Availability, Partitioning
DSS	Decision Support System
DW	Data Warehouse
ETL	Extract-Transform-Load
IQ	Interview Question
HDFS	Hadoop Distributed Filing System
OLAP	On-Line Analytical Processing
ODS	Operational Data Store
SRQ	Sub Research Question

1. CHAPTER ONE: INTRODUCTION TO THE STUDY

1.1 Introduction

e-Commerce, Internet banking, social media platforms and other technology instruments are producing data characterised by large volume, high speed, variety, and veracity (Chen, Mao & Liu, 2014). Many organisations are aware of the value and opportunities that big data (BD) can offer them, but incorporating BD into business strategies is challenging. The management of BD is complex with the result that many tools are flooding the market in order to assist organisations and at the same time, to generate revenue for them. These tools are designed to support decision support systems (DSS), business intelligence (BI) and real-time data processing in BD environments.

DSS and BI systems are information systems that support organisational decision making activities at various levels of management (Delic, Douilet & Dayal, 2001; Keen & Peter, 1980). According to Dayal, Wilkinson, Castellanos and Alkis (2009), the objective of a DSS is to provide the latest information for a given decision making context. According to Watson and Wixom (2007:1), a BI framework consists of two main activities namely; "getting data in" to the data warehouse (DW) and "getting data out" to business users from the DW. The goal of BI is to collect, integrate, cleanse and mine information to support business users to make decisions, based on data stored in the DW (Chaudhuri, Dayal & Narasayya, 2011; Dayal *et al.*, 2009). Data in the DW should provide decision makers with a single unified view of data (Chaudhuri *et al.*, 2011; Chaudhuri & Dayal, 1997).

The emergence of new business models demand significant changes to the original BI architecture, with one of its requirements being, the need to offer real-time access to data (Pereira & Azevedo, 2012; Delic *et al.*, 2001). This approach, where raw operational data is made available for analytical processing, typically supported by OLAP tools, has long been proven to be key to improving decision making processes in firms (Pereira & Azevedo, 2012; Watson & Wixom, 2007). A real-time BI approach differs from the traditional BI scenario where DW updates are performed periodically e.g. at the end of the day, and outside of business hours (Pereira & Azevedo, 2012).

An increase in volume and speed of both, structured and unstructured data, has been experienced over the years in industries and has brought about challenges to BI infrastructures (Gantz & Reinsel, 2011; Manyika, Chui, Brown, & Bughin, 2011). Companies acquire large volumes of data from a variety of sources and leverage this data through analysis to support effective decision making and provide new products and services. The fundamental requirement of business analytics is the ability to scale to the immense volume of data that need to be extracted, processed and analysed in a real-time mode (Doulkeridis & Nørvåg, 2013). The rise in the volume of large data sets has brought about challenges which are driving the innovation and development of appropriate solutions to these challenges. Chen *et al.* (2014) note that advances in IT, for example mobile applications, have made it easy for data to be generated in very high volume and high speed. "…therefore, we are confronted with the main challenge of collecting and integrating massive data from widely distributed data sources." (Chen *et al.*, 2014:3). Further, Chen *et al.* (2014) find that such data in volume, variety and velocity surpasses the capacity of conventional database and DW technologies, and its real-time requirement stresses any available computing capacity.

The challenges faced by BD lead to inadequate, inaccurate and inconsistent data presented to online analytical processing (OLAP) and reporting systems, resulting in slow and sometimes poor decision making by users. Governments, industry and researchers are exploring and developing more solutions to solve challenges brought about by BD (Chen *et al.*, 2014; Xiong *et al.*, 2013). According to Ghazal, Rabl, Huand Raab (2013), as these new technologies mature, a need appears to develop ways to evaluate and compare these. This research focuses on assessing technologies appropriate for real-time DSSs in a BD environment.

1.2 **Research Problem**

1.2.1 **Background to Research Problem**

The rapid growth of the internet and widespread use of digital technologies and applications such as e-commerce, has resulted in unprecedented high volumes, speed, value and diverse types of data (BD) aiding in the emergence of the concept of BD. It is inefficient and ineffective to use conventional database solutions to process this type of data, as it causes slow and sometimes poor decision making.

This leads to the on-going widespread search, development and adoption of new technologies to solve challenges brought about by BD. According to Yan (2014:1-8), BD needs a new set of technologies and tools capable of processing, storing and managing large datasets, and sometimes in real time. Ghazal *et al.* (2014) note that it is essential for industries to be able to compare and evaluate these new technologies as they mature.

1.2.2 **Research problem statement**

Although it has been accepted that real-time access to BD is fundamental to improve decision-making processes in organisations, Zhong *et al.*, (2013) and Dayal *et al.,* (2009), find BD has brought about challenges to organisations whom still rely on traditional real-time BI technologies. As mentioned before, it is inefficient and ineffective to use traditional BI tools to process BD, where real-time access to data is required eliminating slow and sometimes poor decision making, which is based on insufficient data (Duggal, & Paul 2013; Chaudhuri *et al.*, 2011; Kemper & Neumann, 2011; Tank, Ganatra, Kosta, & Bhensdadia, 2010). The inefficiency and ineffectiveness of existing technologies occurs when using stale and sometimes inconsistent data sets at reporting points, while analysing data (Chardonnens, Cudre-mauroux, Grund & Perroud, 2013; Davenport, Barth & Bean, 2012). Although new technologies have been developed and reported by Hossain (2013), Lee, Kwon, and Farber (2013), Garber (2012) and Kemper and Neumann (2011), it is also noted that industry and researchers are in the process of developing and exploring more BD solutions (Chen & Zhang, 2014; Cuzzocrea , Sacca, & Ullman, 2013; Philip Chen & Zhang, 2014). It is found that a gap still exists in literature on how to compare and select appropriate technologies (Ghazal *et al.*, 2013) for real-time DSS in BD environments. From the above discussion, the following research problem is formulated;

Determining appropriate tools for real-time DSS in a BD environment is difficult and complex resulting in inefficient and ineffective systems leading to poor decision making by users.

1.3 Research Questions and sub-research questions

This research is primarily driven by the research questions and sub-research questions discussed in this section. The questions are asked in order to investigate factors that influence the selection of technologies that are appropriate for real-time DSSs in a BD environment. Each sub-research question is used to formulate interview questions (guide) which in turn, are used to collect data from subject experts (BI and BD experts).

1.3.1 Research questions

The following research questions (RQ) are formulated to satisfy the criteria required to solve the stated research problem.

> RQ 1: What factors influence the selection of technologies appropriate for real-time DSS in a BD environment?
>
> RQ 2: How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?

The objective of each research question, including sub-research questions (SQ) and methodology used to answer the questions, are provided in Table 1.1.

Table 1.1 Research questions, sub-research questions, methodology and objectives of research questions.

No.	Research question	Methodology	Objective
RQ 1	What factors influence the selection of technologies appropriate for real-time DSS in a BD environment?	Semi-structured interviews	To identify factors that influence the selection of technologies appropriate for real-time DSS in a BD environment.
SQ1.1.	What is the relationship between the characteristics of data and selection of data analytic tools in a BD environment?	Semi-structured interviews.	To identify characteristics of BD and its impact on the selection of technologies appropriate for real-time DSS.
SQ 1.2	What existing technologies are appropriate for real-time DSS in a BD environment?	Semi-structured interviews.	Identify existing technologies appropriate for real-time DSS in a BD environment.
RQ 2	How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?	Semi-structured interviews and computer laboratory experiments.	Identify evaluation criteria for technologies appropriate for real-time DSS in a BD environment.

SQ 2.1	What are the existing guidelines, frameworks, criteria, or measures applicable when comparing analytics tools for real-time DSS in data environments?	Semi-structured interviews	Identify evaluation criteria that can be used to select technologies appropriate for real-time DSS in a BD environment.

1.3.2 Research Hypotheses for laboratory experiment

The perception based on results from the qualitative interviews and literature review suggests that the most important evaluation criterion for real-time DSSs is performance in terms of low latency and throughput. The following hypotheses are formulated in order to test this finding in a computer laboratory experiment:

H_0 – The mean query execution times for Impala, Spark and Hive are equivalent.

H_1 – The mean query execution times for Impala, Spark and Hive are not equivalent.

1.4 **Research aim**

The aim of this research is to explore real-time DSS including related technologies to identify factors that influences the selection of technologies appropriate for real-time DSS in a BD environment. A framework including evaluation criteria to compare and select technologies most appropriate for real-time DSS in a BD environment is proposed.

1.5 **Research Methodology**

In this section an overview of the research philosophy, research approach, research strategy, data collection and data analysis techniques used for this research, is presented.

1.5.1 **Research philosophy**

O'Leary (2004) describes research philosophy as a set of assumptions that define an intellectual perception of how the world operates and knowledge is produced. According to Saunders, Lewis and Thornhill (2009), research philosophy relates to the development of knowledge in a particular field and the nature of that knowledge. The authors state that research philosophy adopted by researchers should contain important assumptions about the way in which the researcher views the world. These assumptions are ontological (truth about reality) and epistemological (how to get knowledge) (Holden & Lynch, 2004). In this research, the ontological stance of subjectivism is proposed and the epistemological position of pragmatism followed.

1.5.2 **Research Approach**

According to Teddlie and Tashakkori, (2009), a mixed method approach uses both deductive and inductive logic in a distinctive sequence described as the inductive-deductive research cycle. The authors further posit that any of the two (inductive or deductive) can be used first, followed by the other, depending on where the researcher is in terms of studying the phenomenon of interest. According to Blaikie (2009), inductive research is such that generalisations are derived by induction from data. Blaikie (2009) further explains that with inductive research, the researcher begins with collecting qualitative data followed by searching for patterns or categories in the data and ends up with abstract descriptions of defined categories of the data. The identified categories become generalisations or themes, and a combination of these themes becomes theory (Blaikie, 2009). On the other hand, the deductive approach starts off with theory, which is borrowed, produced or invented in the form of a deductive argument. This argument can be a hypothesis, a prediction, or the regularity that is to be explained (Blaikie, 2009). In some research cases, an existing researcher's theory could be used in its original or modified form or alternatively, the theory could be constructed using elements from findings obtained from previous research. In this research, the inductive approach is used to identify factors that influence the selection of technologies appropriate real-time decision support in a BD environment and to propose a framework of evaluation criteria. One element of the theory generated by the inductive approach is used to drive deductive phase of this study. Finally, in this research, a mixed method approach as defined by by Teddlie and Tashakkori (2009) is used being the most appropriate.

1.5.3 **Research strategy**

As stated in Section 1.5.2, an exploratory sequential mixed methods research design is used for this study. The research commenced with a qualitative study and then followed by a quantitative experiment design. In the qualitative phase of the study, ten BI experts with knowledge of BD in South Africa are identified and interviewed, using a set of semi-structured interview questions. The qualitative study is used to identify factors that influence the selection of technologies appropriate for real-time

DSS in a BD environment. The next stage is to generate a framework containing evaluation criteria. The quantitative study is conducted using a computer-based laboratory experiment to compare Hive, Impala and Spark with the objective, of confirming the top criterion identified in the qualitative phase.

1.5.4 **Data collection**

Refering again to Section 1.5.2, qualitative as well as quantitative data are collected using semi-structured interviews and computer laboratory experiments respectively, in a mixed methods research design. Creswell (2009) as well as Maree (2012), characterise mixed methods research containing elements of both, qualitative and quantitative methods. According to Teddlie and Tashakkori (2009), there are six main types of mixed methods research, but only the exploratory sequential mixed methods design is used in this research. In the exploratory sequential mixed methods design, the result of the qualitative strand feeds the generated theory as input into the quantitative strand. This research, thus begins with a qualitative approach (interview BI experts who have knowledge of BD in South Africa) in order to explore the concept of BD, identify factors that influence the selection of technologies appropriate for real-time DSS to generate evaluation criteria used to select such technologies. This is followed by a quantitative approach, where computer laboratory comparative experiments are conducted. Maree (2009), states that quantitative data collection is used with experimental, quasi-experimental, and non-experimental designs to summarise a large number of observations/cases. According to Maree (2009), the quantitative study is either guided by research questions or hypotheses. In this research the quantitative enquiry is guided by hypotheses as discussed in Section 1.3.2. This approach is found appropriate for this research because BD is still a new concept. The qualitative phase is used to clarify the concept of BD by interviewing BI experts who have knowledge and experience about BD and real-time DSS. The output from this phase is then used in the experiments in order to test the theory generated.

1.5.5 **Data analysis**

The focus of this section is to discuss the techniques used to analyse and interpret the collected data for the purpose of building a theory to communicate in essence what the data reveals. As stated in Section 1.5.4, both qualitative and quantitative data are collected for this study. The following sections provide an overview of the analysis techniques used in this study and are; qualitative data analysis and quantitative data analysis.

1.5.5.1 Qualitative data analysis

The preparation involved organising, arranging, and creating a general sense of the narrative text for transcribed interviews (Saunders *et al*., 2009). Each interview record is reviewed soon after the interview session to ensure that information is recorded correctly. This is done with the aid of a voice recorded interview session. According to Creswell and Clark (2011:206), preparing qualitative data involves organising the document or visual data for review or transcribing text from interviews into files for analysis. Each interview is transcribed and stored as a text document ready for analysis. The qualitative data was analysed using the content analysis technique. The results are discussed in Chapter 4.

1.5.5.2 Quantitative data analysis

This section provides an overview of the quantitative data analysis collected from the computer laboratory experiments. According to Mouton (1996), in quantitative data analysis, the researcher analyses data based on the type of questions or hypotheses and uses appropriate statistical tests to address those questions or hypotheses. In this study, Generalised Linear Models in SPSS were used to analyse data collected from the experiments. The data comprised of variables having the following characteristics:

- Tool used – a categorical variable with values; Hive, Impala and Spark.
- Data size – a numerical independent variable measured in gigabytes.
- Time taken – a dependent continuous variable which represents the amount of time taken by a tool when analysing a given dataset with known data size.

1.6 **Units of Analysis**

For the qualitative data, analysis was conducted on BI experts who have knowledge and experience on BD, while for the quantitative study, data analysis was conducted on technologies used. A total of ten BI experts in South Africa were interviewed. The experiments were carried out using three technologies namely; Hive, Spark and Impala.

1.7 Delineation

In this research, no new real-time system, nor will a new architecture or new technology for DSS in a BD environment be designed. However, the research is designed only to suggest evaluation criteria that can be used to compare and evaluate technologies appropriate for real-time DSS in a BD environment. In the experiment phase, the tools are tested using structured and not unstructured data. The real-time aspect is tested with data at rest and not streaming as streaming would require more time and resources than was available for this research. This research is not designed to propose an end-to end BD benchmark.

1.8 Contribution

The concept of BD is still a new phenomenon which is in the process of being accepted and adopted by many companies especially, in South Africa. The contribution of this research is to generate evaluation criteria that can be used by organisations when adopting BD technologies for real-time DSS. The output of this research can also be used as a template for further research on BD benchmarks.

1.9 Significance of study

In this research, it is acknowledged that BD is still in its infancy and there is limited literature available as research and development of BD products is going on. This research fills the gap in that no significant literature exists around the evaluation and selection of technologies appropriate for BD analytics as noted by Dilpreet and Reddy (2014) and Liu, Lftikhar and Xie (2013). The output of this study will also be useful to enlighten BI experts and academic scholars relating to the concept of BD and its related technologies.

1.10 Research ethics consideration

Various ethical considerations needed to be addressed as highlighted by Gray (2009) and Maree (2012). This research involves human subjects and therefore, ethical issues relating to human participants are adhered to. It is also a requirement to obtain ethics approval from the Cape Peninsula University of Technology (CPUT) ethics committee. Strategies are put in place in order to deal with challenges relating to confidentiality, anonymity, right of privacy, voluntary participation, protection from harm and trust (Maree, 2012; Gray, 2009). Letters of consent and the right to confidentiality were obtained from the participants, also from the management of participants' organisations and CPUT. This is essential as stated by Greener (2011:64), that there are some ethical issues that researchers need to be aware of. The researcher informs participants before they willingly, participated in this research, and are:

- Voluntary participation: Participants voluntarily participate and also have every right to withdraw whenever they want from the research. This is addressed in letters of consent obtained from the participants.
- No harm to participants: This research does not require any form of experiments which can cause harm to participants.
- Anonymity: This research treats all respondents anonymously.
- Confidentiality: This research does not identify any participants in any way.
- Deception: This research is conducted with honesty and truth, participants are informed what the research is meant for, and what it hopes to achieve.

Leedy and Ormrod (2010:101) mention additional ethical issues that need to be addressed in this research, as follows:

- Beneficence: The organisations involved in the study will benefit from this research, by being able to use the outcomes (evaluation criteria) as a guide towards the adoption of technologies appropriate for real-time DSS.
- Justice: There is equal distribution of risk and benefits among the participants, and no discrimination tolerated.
- Informed consent: Participants are informed what the research is all about, and affords them to decide if they want to participate.
- Right to privacy: Participants are offered their right to privacy in this research.

1.11 Summary

This chapter provides an overview of the thesis and introduces the research problem, aim, research questions, research methodology, research approach, research techniques used for this study and ethical considerations. To recapitulate, the aim of this research is to explore real-time DSS, its related technologies and to identify factors that the influence the selection of technologies appropriate for real-time DSS in a BD environment. The primary objective is to propose evaluation criteria that can be used as a guide to assess and determine technologies that are appropriate for real-time DSS in a BD environment. To achieve this aim and objectives, a research problem statement is formulated and two main research questions posed. The problem statement for this research is centred on the complexity and difficulty of determining appropriate tools for real-time DSS in a BD environment which leads to inefficient and ineffective systems resulting in poor decision making by users. The two main research questions driving this research are:

- What factors influence the selection of technologies appropriate for real-time DSS in a BD environment?
- How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?

This chapter is followed by the literature review in Chapter 2. The detailed research methodology is presented in Chapter 3. This is followed by research findings and discussion in Chapter 4 and Chapter 5, respectively. Chapter 6 finalises the thesis with conclusions and recommendations.

The next chapter also provides a discussion of the concept of real-time DSS, BD and related technologies, from a literature review perspective.

2 CHAPTER TWO: LITERATURE REVIEW

In Chapter 1, an introduction, aim and objectives of this research are discussed. The problem statement and research questions, methodology and ethical considerations are also covered. To recapitulate the aim of this research; to explore real-time DSS, its related technologies and to investigate what factors influence the selection of technologies, appropriate for real-time DSS in a BD environment. The main objective as dicussed, is to propose evaluation criteria that can be used as a guide to assess and select appropriate technologies for real-time DSS in a BD environment. The research commences by formulating the research problem statement which forms the focal point for this research. The problem statement is discussed in Section 1.2.2. This is followed by formulating research questions given in Section 1.3.1. The next step is to identify key words from the research problem statement, research questions and the aim of the research. Armed with the identified key words, the next step is to identify existing literature related to the concept of real-time DSS, BD technologies and frameworks. The objective of the literature study is to contexctualise and relate this research to current literature on this topic. Different literature databases were used which include; Google Scholar, CPUT databases such as, Emarald and other scientific journal and conference databases. Furhermore IEEE Xplore Digital Library, ACM Digital Library, SpringerLink and ScienceDirect were accessed via the online library at CPUT.

The literature review process commences with analysing he the concept of DSS and BI in Section 2.1, followed by analysis of literature on the concept of real-time DSS and the impact of BD on real-time DSS. In Section 2.3, the researcher explores the concept of BD in detail by studying the various definitions of BD. In Section 2.4, various characteristics of BD are discussed as described in existing literature. Section 2.5 provides the analysis of different technologies used for DSS in a BD environment with the objective of gaining an understanding of all technologies that can be used in a BD environment. Finally, Section 2.6 contains existing literature in relation to the process of BD technologies with the objective of identifying any gaps.

2.1 **Business Intelligence**

Within a business intelligence setting, the quality of a decision is determined by the volume, timeliness, quality and variety of data available to the decision maker (Mcguire, Manyika & Michael, 2012). Technologies used to process and analyse data also influences the quality of decision making. The advent of the Internet and related digital technologies, accelerates the intensity of competition, increases the volume of data available to decision makers, resulting in short decision making cycles (Delic *et al*., 2001; Tank *et al*., 2010). This exposes users to large inflows of data and exerts pressure on them to make quick and critical business decisions, to maximise profitability, mainly of their core business (Delic *et al.*, 2001). According to Marshall and de la Harpe (2009), it is essential for up-to-date information to be available to decision makers with as little latency as possible. The demand for fresh (up to date) data by enterprises leads to the development and adoption of real-time DSS, making it possible to deliver timely decision support (Watson & Wixom, 2007). Furthermore, Watson and Wixom (2007) state that many enterprises benefit from real-time DSS by making it possible to influence current decision making, operational business processes and customer facing applications as things are happening.

Although the importance of real-time DSS has been highlighted (Chaudhuri *et al.*, 2011; Sahay & Ranjan, 2008; Delic *et al.*, 2001), new technologies to support real-time DSS have been developed according to (Dehne & Zaboli, 2012; Jörg & Dessloch, 2010). BD has rendered conventional database technologies ineffective and inefficient for real-time DSS (Manyika *et al*., 2011; Pavlo, Paulson & Rasin, 2009). The ineffectiveness and inefficiency becomes apparent when systems fail to effectively deliver accurate and complete information timeously to users, which can affect the requirement for quick decision making (Singh & Singh, 2012). According to Ghazal *et al*. (2013), there is an increasing interest in BD by academia, industry and a large user base, which has seen the birth of several technologies designed for the processing of BD.

These new technologies range from in-memory databases, NOSQL databases, Cloud computing solutions to many others such as providing reporting and analytic tools. Large commercial database vendors such as Oracle, IBM and Microsoft have intensified tweaking and transforming their existing data processing and analytics infrastructures, to integrate with BD frameworks such as Hadoop (Mctaggart, 2008) and Spark (Zaharia *et al.*, 2010). The open source world has also released many tools to analyse BD such as Hive (Thusoo *et al*., 2009), Impala (Wanderman-Milne & Li, 2014), Spark (Zaharia *et al*., 2012; Zaharia *et al*., 2010), Pig (ApachePig, 2013) and MapReduce (Mctaggart, 2008). On the other hand, as mentioned earlier, commercial vendors of database and data analytics products are investing in research and development towards the development of hardware and software products to manage and analyse BD (Purcell, 2013; Singh & Singh, 2012). It is noteworthy that these technologies are new and there is limited knowledge and skills in organisations to manage them, let alone to identify appropriate technologies for a BD analytics environment. According to Ghazal *et al.* (2013) and Yan (2013), it is not apparent to determine the value of BD technologies as there are no available guidelines or frameworks for comparing and evaluating BD tools. According to Ghazal *et al*. (2013) and Xiong *et al*. (2013), it is essential for organisations to be able to compare and evaluate these new technologies as they mature. In the following sections, a review of literature on BD and its related technologies is conducted.

2.2 **Real-time DSSs (BI)**

The traditional BI architecture is composed of source systems, operational data store (ODS), extract-transform-load (ETL) process, a data warehouse (DW) or one or more data marts and a set of reporting tools as shown in Figure 2-1.

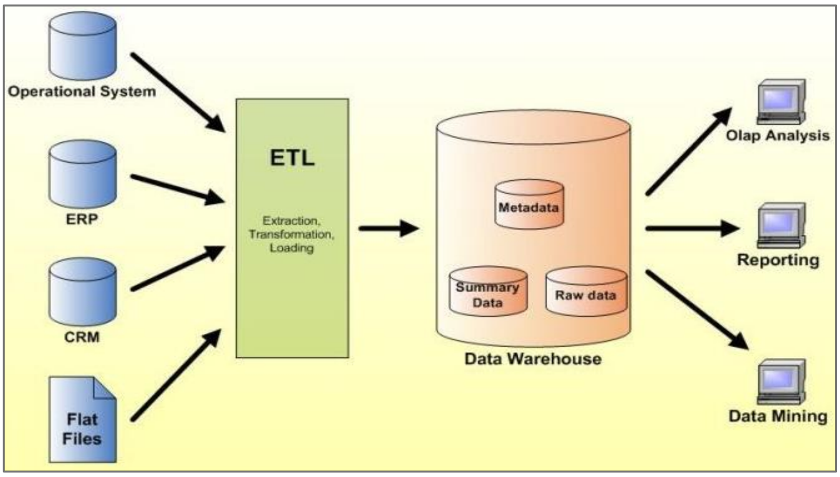


Figure 2-1 Traditional BI architecture (Source: Volts, 2015:2)

DWs are traditionally refreshed in a periodic manner, most often on a daily basis, however, there is some delay between a business transaction and its appearance in the DW (Jörg & Dessloch, 2010; Tank *et al*., 2010). The most recent data remains in the operational source systems where it is unavailable for analysis until the next schedule kicks off. Traditionally, this BI architecture is designed for tactical and strategic decision making where historical data is analysed for reporting and for building analytic models (Dayal *et al*., 2009). The ETL process periodically extracts, cleanses, integrates, transforms and loads data into the DW for query and reporting from multiple source systems. The ETL design and implementation is an intensive activity which takes up a high percentage of the amount of work required to build DW projects. The complexity of the ETL process causes data to be updated in the DW periodically instead of in real-time. The ETL process has been identified as a bottle neck, delivering fresh data to the DW (Tank *et al*., 2010). The trend of BI is shifting towards real-time BI, where BI reports and analytics are required to support operational activities (Farooq, Sarwar & Mansoor, 2010). In real-time BI, data is analysed as soon as it lands in the organisation, that is, as soon as a business event has occurred (Sandu, 2008).

According to Jörg and Dessloch (2010), for timely decision making, business users ask for fresh data and to meet this requirement, some businesses are making use of near real-time data warehousing which shortens the DW refreshment intervals thereby delivering source data to the DW with low latency. Jorg and Dosslech (2010) prove that near real-time data warehousing has anomalies which cause inconsistencies between the DW and its sources. They state that analysis based on inconsistent data will likely lead to wrong decisions being made and this is also supported by Tank *et al*. (2010). Although Jorg and Dosslech (2010: 11) proposed several solutions to prevent what they called "…change data mismatch" and "…change propagation delays", these solutions are not scalable and work only with structured relational data.

Tank *et al.* (2010) also propose high performance joins and high performance aggregations as possible ways to refresh the DW with low latency but this is not adequate in a BD environment where the volume and speed of data is very high and the data is sometimes unstructured. There are several other approaches proposed to deliver data into the DW in real-time (Dayal *et al*., 2009; Golden Gate & Inc, 2009), but according to recent reported trends, the nature of BD has rendered traditional real-time technologies inadequate (Liu *et al*., 2013). In the following sections, the concept of BD is explored and literature on technologies used to process BD, is discussed.

2.3 **BD definition**

In literature, numerous definitions of BD are available. The first definition of BD is reported by Gartner (known as META) in 2001, although the term BD was not used at the time. BD is described in Gartner as *"…*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*"* (Laney, 2001: 1-2).

According to Manyika *et al.* (2011:6), BD refers to "…data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered BD*".* The authors affirm in this definition that to classify data as 'big', is contextual because there is no volume limit or threshold associated although, volume is the single attribute.

Most literature reports on the definition of BD using the "3V" model (volume, velocity and variety) defined by Laney (2001). However, some researchers have added one, and others two more "Vs", for veracity and value. For the purpose of this research, the following definition is adopted from Yan (2014:8) and is used, as it embraces all aspects under investigation such as, technologies and real-time analysis: "…BD refers to large and/or diverse datasets either in motion or at rest, and a set of technologies handling these data sets. BD at rest, is analysed periodically. BD in motion refers to datasets that are processed and analysed in real time or immediately when they are received. BD technologies refer to a union of tools, platforms, and systems, and analytics and applications that allow data scientists to capture, store, govern and analyse large and diverse datasets" Yan (2014:8).

2.4 **BD Characteristics**

There are many factors which influences the development of BD systems which include, the exponential growth of data, changing user behaviour and globalisation (Tekiner & Keane, 2013). The convenience, affordability and accessibility of online applications, e-Business, e-Commerce and electronic devices have led to people spending more time online and using mobile devices, thereby creating large volumes of data. According to Tekiner and Keane (2013), organisations are constantly looking for opportunities to increase their competitive advantage in a competitive global market place, by using efficient and effective data analytical models and all forms of data available. This spans from structured to unstructured data. The authors also note that it is essential for analytic applications to present findings and reports in a clear and concise form. The five main characteristics of BD; volume, velocity,variety,veracity and value, are discussed in Sections 2.4.1 through 2.4.5.

2.4.1 **Volume**

When large pools of data are brought together and analysed, businesses can derive new patterns and make quality decisions which is the basis of competition and growth for individual firms (Letouze, 2012; Singh & Singh, 2012;). According to Singh and Singh (2012), the size of data being generated in organisations is very large in measures of terabytes and petabytes. Liu (2013) overcame the challenge of high volume data requiring technologies that store vast amounts of data in a scalable fashion. The author suggests using technologies that use distributed approaches, for querying and deriving actionable information and insights from the BD, with very low latency.

2.4.2 **Velocity**

Velocity refers to the rate at which new data arrives, or the rate at which changes are made to existing data (Stonebraker, Madden & Dubey, 2013; Letouze, 2012; Agrawal *et al*., 2009). Agrawal *et al.* (2009:10) state that the rate at which data arrives is called ”…acquisition rate challenge”. The rate at which data is updated is called “…timeliness challenge”, and this corresponds to an acceptable time to analyse data and act on it while it is flowing in. The objective is to quickly and consistently maintain a persistent state of data in the database in a real-time fashion (Stonebraker *et al*., 2013).

2.4.3 **Variety**

According to Singh and Singh (2013), BD consists of a variety of data types which are both, structured and unstructured in nature such as, text, videos, pictures and audio files. Liu (2013) states that variety refers to the proliferation of data types from social media, machine to machine, and mobile sources in addition to traditional transactional data and this data no longer fits into neat, easy to consume structures. The author further states that the diversity of BD requires new techniques and approaches to storing, moving, processing and reporting methods in order for businesses to derive deeper insights and new values from BD.

2.4.4 **Veracity**

According to Dong and Srivastava (2013), data sources in a BD environment, even in the same domain, inherently have different quality, with significant differences in the coverage, accuracy and timeliness. This is consistent with observations made by Singh and Singh (2013), who refer to varacity as the trust that an organisation places on the data. Yan (2013:2) describes veracity as the "…integrity of data". The consistency and accuracy in data for decision making is determined by the number of data sources. In BD environments there are multiple data sources and therefore there is a challenge of trust in the data available to decision makers.

2.4.5 **Value**

According to Yan (2013:4), value refers to "…the usefulness of data". The author states that it is the potential value in BD which makes it a "…hot…" topic. According to Manyika *et al*. (2011:14), there are five new kinds of value that might come from BD:

- Creating transparency in organisational activities that can be used to increase efficiency.
- Enabling more thorough analyses of employee and system performance in ways that allow experiments and feedback.
- Segmenting populations in order to customize actions

- Replacing / supporting human decision making with automated algorithms.
- Innovating new business models, products, and services.

2.5 **BD technologies**

There are several technologies designed for acquiring, storing, processing and analysing BD. The techniques share common characteristics of scalability, elasticity, fault-tolerance and low latency reads (Bakshi, 2012; Marz, Narthan & Warren, 2012). According to Dobre and Xhafa (2013), as the volume of large data sets grew and exceeded the capacity of existing DBMSs, the database industry responded with a number of solutions. In this section, an overview of some of the technologies that have been developed that can be used for real-time DSS in a BD environment is provided. These are; massively parallel databases, in-memory databases,hybrid databases, NOSQL databases and the Hadoop framework in Sections 2.5.1 through 2.5.5.

2.5.1 **Massively Parallel databases**

Massively Parallel Processing (MPP) databases allow database loads to be split among many processors (Dobre & Xhafa, 2013). MPP databases are based on a distributed architecture (Özsu & Valduriez, 2011). The distributed architecture is designed for high scalability and fault tolerance when processing large volumes of data (Dobre & Xhafa, 2013). MPP databases employ a "…shared nothing" architecture, where each node has its own CPU, memory and disk (Bakshi, 2012:2).

2.5.2 **In-memory databases**

In-memory databases are designed to provide quick data analysis while it is in memory (RAM) rather than on disk. This speeds up the data analytics processes, even when the size of data becomes excessively large (Garber, 2012). In-memory databases are significantly faster than traditional disk based databases because they hold all data in memory. Examples of in-memory databases include SAP Hana (Lee *et al.*, 2013) and Oracle TimesTen (Oracle, 2009).

2.5.3 **Hybrid databases**

A development within in-memory databases, is the concept of hybrid databases. Traditionally, the On-Line Transaction Processing (OLTP) database is separated from OLAP DW with the later, being periodically refreshed by an ETL process. Kemper and Neumann (2011) state that this separation has many disadvantages, including data freshness issues due to the delay caused by only periodically initiating the ETL. Furthermore, exasperated by data staging and excessive resource consumption due to maintaining two separate information systems. To bridge such two separate systems, hybrid databases have been developed such as, Hyper (Kemper & Neumann, 2011), ScyPer (Mühlbauer, Rödiger & Reiser, 2013) and SAP Hana (Lee *et al.*, 2013). The concept behind in-memory hybrid databases is for OLTP transactions and OLAP queries to be performed on the same main memory resident database, but without interfering with each other. The goal is to meet real-time requirements currently being demanded by users wanting to process large scale data.

2.5.4 **NOSQL databases**

The origin of the term NOSQL reported, is attributed to Johan Oskarsson, who used it in 2009 to name a conference about "…open-source, distributed, non-relational databases*"* (NOSQL-meetup, 2009). However, Hossain (2013) states that NOSQL acronym was coined as far back as 1998. According to the Apache foundation, NOSQL is a general term meaning that the database is not a relational database management system (RDBMS) which supports SQL, as its primary access language (Apache, 2014). NOSQL databases are suitable when working with large volumes of data or when the nature of data cannot fit into relational databases. Hossain (2013), describes NOSQL systems as distributed, non-relational databases designed for large-scale data storage and for massively-parallel data processing across a large number of commodity servers. They use non-SQL, or not only SQL languages and mechanisms, to interact with data through some feature application programming interfaces (APIs) that convert SQL queries to the system's native query language or tool.

NOSQL database systems are designed to scale to thousands or millions of users executing updates and reads (Pokorny, 2013). Hossain (2013) and Pokorny (2013) state that traditional database technology offers transactional processing characterised by ACID (Atomicity, Consistency, Isolation and Durability) properties. This ensures persistent integrity and consistency of data in all situations of data management. However, it is proven that scaling out of ACID-compliant systems is difficult. This led to the development of a new concept used by NOSQL databases called the CAP (Consistency, Availability, Partitioning) theorem (Pokorny, 2013)

Pokorny (2013) further states that the CAP theorem implies that for any system sharing data, it is impossible to guarantee simultaneously, all the CAP theorem properties. Hossain (2013) also supports this by stating that NOSQL databases have now loosened up the consistency requirement in order to achieve better availability and partitioning. This resulted in systems known as BASE (Basically Available, Soft-state, Eventually consistent) (Neubauer, 2010). Pokorny (2013:5) describes the BASE theorem as*;* "…an application works basically all the time (basically available), does not have to be consistent all the time (soft state) but the storage system guarantees that if no new updates are made to the object eventually all accesses will return the last updated value. Availability in BASE is achieved through supporting partial failures without total system failure".

There are three main categories of NOSQL databases namely; key-value stores, column-family / wide-Column stores and document stores. Although each category works well in specific application scenarios, Hossain (2013) identifies four main primary uses of NOSQL databases. These are: large-scale data processing (parallel processing over distributed systems), embedded IR (basic machine-to-machine information look-up & retrieval), exploratory analytics on semi-structured data (expert level) and large volume data storage which in turn is divided into unstructured, semi-structured and small-packet structured.

2.5.5 **Hadoop framework**

Hadoop, according to Borthakur (2007) and Mctaggart, (2008) is an open source software project that enables scalable distributed processing of large data sets, across clusters of commodity servers. Hadoop has many similarities with existing distributed file systems such as, Google File System (GFS) and others (Ghemawat, Gobioff & Leung, 2003). HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop supports applications under a free license.

There are three Hadoop sub-projects:

- Hadoop Common: common utilities package
- HDFS: Hadoop Distributed File System with high throughput access to application data
- MapReduce: A software framework for distributed processing of large data sets on computer clusters.

2.5.5.1 HDFS – Hadoop Distributed File System

The main concept behind Hadoop is to move processing towards the data in a distributed file system manner, instead of bringing data to processes (Dean & Ghemawat, 2008; Bakshi, 2012). A single file is split into blocks and these blocks are distributed in the Hadoop cluster nodes. The input data in HDFS is treated in a write-once fashion and processed by MapReduce, which is discussed in more detail in section 2.5.5.2. The results are then written back in HDFS. The data in HDFS is protected by a replication mechanism among the nodes. This provides reliability and availability, despite node failures. There are two types of HDFS nodes: DataNode, which stores the data blocks of the files in HDFS; and NameNode, which contains the metadata, with enumeration of blocks of HDFS and a list of DataNodes in the cluster.

2.5.5.2 MapReduce

MapReduce is a programming model and software framework, first developed by Google (Dean & Ghemawat, 2008). It is Intended to facilitate and simplify the processing of vast amounts of data in parallel, on large clusters of commodity hardware in a reliable and fault-tolerant manner (Mctaggart, 2008). In a MapReduce implementation, a map function is specified that processes a key/value pair, to generate a set of intermediate key/value pairs, and to reduce functions that merge all intermediate values associated with the same intermediate key (Dean & Ghemawat, 2008).

According to Bakshi (2012:3), MapReduce provides a framework for programmers to leverage the distributed systems for processing datasets in two distinct phases:

- Map phase – divides the workload into smaller sub-workloads and assigns tasks to Mapper, which processes each unit block of data. The output of Mapper is a sorted list of key/value pairs. This list is passed (also called shuffling) to the next phase.
- Reduce – Analyses and merges the input to produce the final output which is then written to the HDFS in the cluster. Clients read the results from the HDFS.

Although Hadoop was initially designed to process large data sets in batch mode, recent implementations now have the capability to answer user queries in a real-time manner. There are several implementations of Hadoop currently on the market and newer versions are being developed to meet new user requirements. Some vendors are integrating Hadoop with traditional databases (Su & Swart, 2012) to leverage on the strengths of both technologies.

2.5.5.3 Hadoop Implementations

There are three common implementations of Hadoop namely; Apache Hadoop from Hortonworks, Cloudera Hadoop from Cloudera and MapR. The ecosystem for the Hadoop implementation is shown in Figure 2-2. Each Hadoop implementation comes with several components (mainly from Apache) used for storing data, accessing data, processing data and analysing data. In this research, the Hadoop implementation from Cloudera is used.

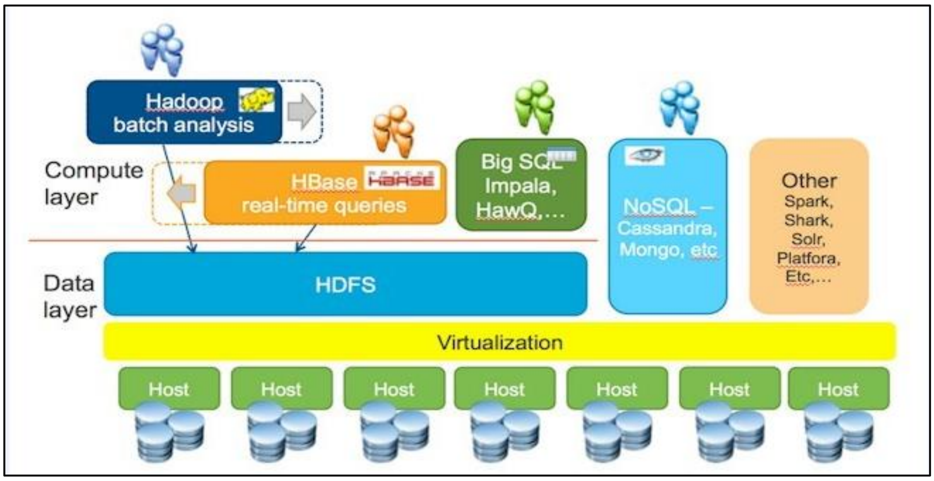


Figure 2-2 Hadoop Ecosystem (Source: Morgan, 2013:2)

2.5.5.4 Impala

Impala is an open-source MPP database built for the Hadoop ecosystem (Wanderman-Milne & Li, 2014). The authors state that Impala is designed to combine the flexibility and scalability that is expected from Hadoop with performance and SQL support offered by commercial MPP databases. Impala currently executes queries 10 to 100 times faster than existing Hadoop solutions and is comparably equal to commercial MPP databases. This allows end users to run interactive exploratory analytics on BD. Impala is designed for analytic workloads, rather than OLTP.

2.5.5.5 HBase

According to Hortonworks (Apache, 2014), HBase is a non-relational (NOSQL) database that runs on HDFS. It is columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. HBase adds transactional capabilities to Hadoop, allowing users to conduct data manipulation operations such update, insert and delete. HBase also provides random and real-time access to BD. Hbase is designed to store large tables with billions of rows and millions of columns. HBase also provides a wide range of benefits which include; fault tolerance, ear real-time lookups, atomic and strongly consistent row-level operations and automatic sharding and load balancing of tables

2.5.5.6 Hive

Hive is an open-source data warehousing solution built on Hadoop (Thusoo *et al.*, 2009). According to the authors, Hive supports queries expressed in a SQL-like declarative language called HiveQL, which is compiled into map-reduce jobs and executed on Hadoop. The language supports a type system with support for tables containing primitive types, collections like arrays and maps, and nested compositions of the same. According to Thusoo *et al.* (2009) the Hive database has a model which is composed of tables, partitions and buckets.

2.5.5.7 Spark

Spark is a cluster computing framework which supports applications with working sets, while providing similar scalability and fault tolerance properties to MapReduce (Zaharia *et al.*, 2010). According to Zaharia *et al.* (2012), the main abstraction in Spark is that of a resilient distributed dataset (RDD). RDD is a read-only array of objects partitioned across nodes that can be rebuilt if a partition is lost. In this framework, users can explicitly cache an RDD in memory across machines and reuse it in multiple MapReduce-like parallel operations. RDDs achieve fault tolerance through a notion of lineage: if a partition of an RDD is lost, the RDD has enough information about how it was derived from other RDDs, to be able to rebuild just the damaged partition. According to Zaharia *et al.* (2010), on Hadoop, each query incurs significant latency (tens of seconds) because it runs as a separate MapReduce job and reads data from disk. Spark can outperform Hadoop by 10 times in iterative machine learning jobs, and can be used to interactively query a 39 GB dataset with sub-second response time (Zaharia *et al.*, 2012). Spark can also be run on a Hadoop cluster. Several other components are built on top of Hadoop ecosystems in addition to the above. The recent trend has seen more technologies both open source and commercial, being built around the Hadoop and Spark frameworks.

2.5.6 **Lambda architecture**

A wide range of technologies to process BD such as NOSQL databases, MapReduce and Spark have evolved. However, there is no single tool that provides a complete solution to BD challenges. Marz *et al.* (2012) proposed a data-processing architecture designed to handle large quantities of data by taking advantage of both batch and stream-processing methods. According to Marz *et al.* (2012), this architecture addresses latency, throughput, and fault-tolerance by using batch processing to provide views of batch data, while at the same time, using real-time processing to provide views of fresh data. This architecture is made up of three layers namely; speed layer, batch layer and serving layer as shown in Figure 2-3.

2.5.6.1 Batch layer

According to Marz *et al.* (2012), the batch layer contains the master copy of the batch dataset and pre-computes views on that master dataset. Indexes are then created on the pre-computed views to provide quick access to data. The idea behind the batch view is to provide answers to queries using pre-computed results sets instead of scanning through all the data at the time of querying. The views are continuously refreshed from scratch.

2.5.6.2 Serving layer

Marz *et al.* (2012) states that the results of the functions executed in the batch layer are uploaded into the serving layer so that it can be efficiently queried. The serving layer is a specialised distributed database that loads in batch views, makes them queryable, and continuously refreshes them as they are re-computed by the batch layer.

2.5.6.3 Speed layer

The speed layer contains views on real-time data. The speed layer does incremental updates of the views as it receives new data.

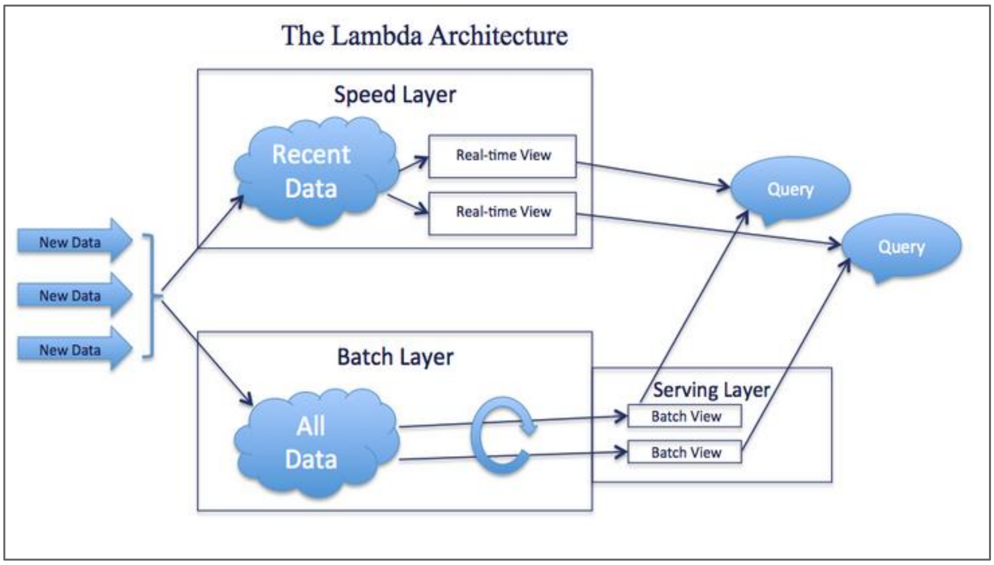


Figure 2-3 Lambda Architecture (Source: Piekos, 2014:4)

2.6 **BD technology evaluation**

As gleaned from Section 2.5, several technologies have been designed and developed to process BD. For example, Ghazal *et al.* (2013) and Dilpreet and Reddy (2014), also made these observations but noted further, that there is a gap in existing literature on how these technologies can be compared and evaluated. It is important to note that the excitement and interest in BD is continuously driving the development of more and more, open source and commercial BD technologies. This is further making it difficult and complex for organisations to identify and determine technologies that are appropriate for real-time DSS requirements in a BD environment. Therefore, there is a need for guidelines, frameworks or end-to-end benchmarks which are easy to use to help enterprises in evaluating and comparing these tools (Dilpreet & Reddy, 2014). A number of BD technology benchmarks have been proposed as noted by Bakshi (2012), Ghazal *et al*. (2013) and Gualtieri (2013), but at the time of this study, no end-to end standard benchmarks could be identified from existing literature. Those that are available do not offer this and are very complex to use (Ghazal *et al*., 2013: Liu *et al*., 2013). What compounds this challenge is the shortage of BD technology expertise especially, in South Africa. Also at the time of this study, there is no educational institution in South Africa offering courses on BD. This study therefore, seeks to propose evaluation criteria that can assist organisations in determining appropriate technologies for BD analytics.

2.7 **Summary**

In this chapter, existing real-time DSSs and BD technologies are reviewed. It is highlighted that the trend in most enterprises is driving BI and DSSs towards real-time, or operational BI for competitive advantage. Furthermore, it is apparent that although industry has developed several technologies for real-time DSSs, it is evident from literature that traditional BI tools are no longer adequate for managing BD. Traditional BI tools are not designed to process high volume, heterogeneous and high velocity data and therefore, enterprises are not leveraging BD. Organisations are experiencing growing volumes of structured, semi-structured and unstructured data especially from new sources such as machines, sensors, logs and social media. In this chapter, it is also revealed that industry and academia is responding positively, to the challenge of BD by designing and developing technologies that can handle the new data sources and types for real-time DSS. Referring to Section 2.6, the challenge within enterprises now is to identify and determine tools that are appropriate in a given BD environment because there are no available guidelines or standard benchmarks that can be used to compare and evaluate these technologies. The next chapter is devoted to the research methodology and techniques used for this research.

3 CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY

3.1 Introduction

This chapter provides a discussion of the formulated research design process, epistemology stance, philosophy and research approach to achieve the objectives and aim stated in Section 1.3 and 1.4, respectively. Furthermore, the researcher discusses the interview data collection technique, the experimental design technique and ethical considerations. In Chapter 1, the researcher established the aim of this research, which is to identify factors that influence the selection of technologies appropriate for real-time DSS in a BD environment and to propose evaluation criteria that can be used to assess and select such technologies. It is established in Chapter 2 that evaluating and selecting technologies appropriate for real-time DSS in a BD environment is complex. It is reported that industry and academia have developed numerous technologies to process BD but there are no guidelines or end-to end standard BD benchmarks available to assist in assessing and selecting appropriate technologies. This research therefore explores the concept of real-time DSS, BD and its related technologies in order to achieve the aim and objectives set for this study.

Creswell (2009) describes research methodology as a perspective that provides a philosophical frame of reference for approaching research. According to O'Leary (2004:85), research methodology is "…the framework associated with a particular set of paradigmatic assumptions used to conduct research." Section 3.2 through to Section 3.9 discusses the research philosophy, research choice, research strategy, data collection, data analysis, delineation and ethics considerations for this research.

3.2 Research philosophy

Research philosophy relates to the development of knowledge in a particular field and the nature of that knowledge (Saunders, Lewis & Thornhill, 2009). The philosophy adopted by a researcher contains assumptions about the way in which the researcher sees the world (Saunders *et al.*, 2009). These assumptions are sometimes known as worldviews, paradigms or beliefs (Denzin & Lincoln, 2008) that the researcher has when conducting research. There are four main paradigms within which a research can be positioned namely; positivism (Henning, 2004), interpretivism (Ritchie & Lewis, 2003; Saunders *et al*., 2009; Teddlie & Tashakkori, 2008), advocacy/participatory (Johnson & Onwuegbuzie, 2012) and pragmatism (Plano Clark & Creswell, 2008). These assumptions are ontological and epistemological (Holden & Lynch, 2004). Epistemology "…is concerned with ways of knowing and learning about the social world" (Ritchie & Lewis 2003:26). In this research, an ontological stance of subjectivism is taken while the epistemological position of pragmatism is assumed.

Pragmatism employs any system of philosophy and and uses both qualitative and quantitative approaches (mixed methods) to collect data (Creswell, 2009). According to Creswell (2009), pragmatic researchers use methods, techniques and procedures of research that best meet their needs and purposes. Pragmatism is found appropriate for this study because the concept of BD is still a new concept which needs the use of multiple methods and techniques.

3.3 **Research Choice**

There are two main research choices that researchers may follow when conducting research, namely; the *"…mono and multiple"* methods (Saunders, Lewis & Thornhill, 2009:152). Multiple methods are divided into a multi-method choice and a mixed-method choice. Mixed methods are further divided into mixed-method and mixed-model research. Mixed method research uses qualitative as well as quantitative data collection techniques but does not combine the two collection techniques. On the other hand, mixed-model research combines qualitative and quantitative data collection techniques by for example taking qualitative data converting it to quantitative data and vice versa. For this study, the sequential exploratory (Teddlie & Tashakkori, 2009; Plano Clark & Creswell, 2008) typology of the mixed method approach was used. In this method, qualitative data is collected by interviewing BI experts and analysing qualitatively. The theory generated from the qualitative study (top evaluation criterion) is used as input in the quantitative study. The quantitative data is collected through experiment design and analysed statistically. Qualitative and quantitative data collection and analysis techniques are discussed in detail in Sections 3.6 and 3.7, respectively.

3.4 **Research strategy**

Saunders *et al.* (2009:600) define research strategy as *"…the general plan of how the reserahcer will go about answering the research question".* There are various research strategies available that can be used by a researcher which include case study, experiment, grounded theory, interviews, action research and ethnography (Saunders *et al*., 2009). This research uses interviews and experiment as research strategies to explore the concept of BD and achieve the aim and objectives stated in Chapter 1.

3.5 **Units of analysis**

Babbie and Mouton (2001:84) describe unit of analysis as *"…the WHAT of your study: what object, phenomenon, entity, process or event you are interested in investigating".* According to Bless, Higson-Smith, Kagee (2006), the unit of analysis can be a person, object, an individual, group of people, an organisation, a time period or a social artefact. from whom or where data is collected by the researcher. These authors state that data from a unit, describes the unit, but when put together with data from similar units, the data then provides an accurate picture of the population. In this research, there are two units of analysis namely BI expert and a DSS. The qualitative enquiry was conducted by gathering data from BI experts who have knowledge and experience about BD and its related technologies. The objective was to find how technologies appropriate for real-time DSS in a BD environment are compared and evaluated. The unit of analysis is therefore the BI expert who can provide insights and perceptions about the selection of technologies appropriate for real-time DSS in a BD environment. The unit of analysis for the quantitative enquiry was the DSS and the units of observation were the three tools used in the experiment namely, Hive, Impala and Spark.

3.6 **Data collection techniques used**

In this section, the techniques used for collecting data for the research are discussed. According to Blaikie (2004), there are three types of data that can be gathered by a researcher namely, primary data, secondary data and tertiary data. The author describes primary data as data generated by the researcher, secondary data as raw data that has been collected by another person and tertiary data, as data that is collected by someone else which is not raw data but is found in analysed format. In this research, primary data are collected using both qualitative techniques and quantitative techniques (Gray, 2009; Teddlie & Tashakkori, 2009; Henning, 2004).

3.6.1 Qualitative data collection techniques

Interviews

In this research, the interview guide used is presented in Appendix A, and comprises of open-ended and closed-ended questions (Teddlie & Tashakkori, 2009; Plano Clark & Creswell, 2008; Denscombe, 2007). The semi-structured interview technique is used to collect data (Teddlie & Tashakkori, 2009) about the nature of information that organisations use for decision making, the technologies used to analyse that information and the factors that drive the selection of such technologies. The research questions are discussed in Section 1.3 and forms the centre of the qualitative enquiry (Horvat, Heron, Agbenyega & Bergey, 2013) for this study. Ten BI experts who have knowledge about BD were identified in South Africa and interview sessions were scheduled directly with the interviewees. Each interview session takes between 15 and 30 minutes. Five participants are interviewed face-to-face, three participants are interviewed over the telephone and two participants are interviewed over Skype. Note-taking is used to record the information obtained from the interviews during the interview as detailed notes were made. The Interviews are conducted in such a way that the interviewees are free to express their own views and feelings about BD technologies, and how BD analytics is being done in their organisations. All the participants are asked the same set of questions and in sequential order. When the first two interviews are conducted, the schedule consisting of 15 questions, the questions are reduced to 11 questions as the subject under study became clearer. Once the interviews are done the interviews are transcribed and offered to the participants to validate the correctness of the transcription.

Sampling

Purposive sampling (Creswell & Plano Clark, 2011; Creswell, 2009; Plano Clark & Creswell, 2008; Henning, 2004; Babbie & Mouton 2001) or to be precise, expert sampling (Kumar, 2011), is used as the sampling strategy for data collection.

According to Teddlie and Tashakkori (2009:173), purposive sampling techniques involve *"…*selecting certain units or cases based on a specific purpose rather than randomly*".* In this research, the participants are selected from a population having the following characteristics:

- Participant must be a BI expert, BD expert or manager of a BD project.
- Must have some experience in BD analytics projects or the participant's organisation must be have been involved with BD technologies proof of concept (POC).
- The participant's organisation must have data which display the characteristics of BD as described in Section 2.4.

3.6.2 **Quantitative data collection techniques**

In this section, a description of quantitative research and quantitative data collection techniques is provided. Quantitative data collection techniques include questionnaires, tests, and some form of structured interview (Teddlie & Tashakkori, 2009), experimental, quasi-experimental, and non-experimental (Maree, 2012). In this research, quantitative data is collected using an experiment design technique. According to Denzin and Lincoln (2008), quantitative study emphasises on the measurement and analysis of causal relationships between variables. The data collected in this part of the research is in numerical form and the following three variables are defined:

- Tool used – the data type for this variable is nominal / categorical with possible values being: Spark, Impala and Hive.

- Volume – the size of data in Gigabytes analysed by each tool. This is a numerical variable.

- Time – the query execution time taken to bring back results in seconds (s) per tool when analysing a given volume of data. This variable is continuous.

Creswell (2009) describes quantitative research as a means for testing theories by examining the relationship between variables. In this research, the qualitative study, in Section 3.6.1, generated theory first, which was then used to generate the hypothesis specified in Section 1.3.2. This approach is supported by Teddlie and Tashakkori (2009), who state that the hypotheses can be based on known theory, results from a previous research, or some other rationale about the relationship among social phenomena. The following section will now discuss quantitative experiments.

3.6.2.1 Experiments

Denscombe (2007:48) describes an experiment as "…an empirical investigation under controlled conditions designed to examine the properties of, or relationship between specific factors*".* According to Saunders *et al*. (2009), experiments aim to be used to study relationships between two or more variables, for example, to establish whether a change in one variable results in a change in another dependent variable. The qualitative research is used to identify attributes of different technologies that can be used to assess and select appropriate technologies for real-time DSS in a BD environment. Based on the perceptions of the interview participants, the most important attributes (variables) are used to in an experiment setting to evaluate purposively selected tools. In this this regard Hive, Impala and Spark, described in Section 2.5.5 are evaluated based on the hypothesis formulated on the identified attributes.

i) Technology selection

In this research, the distribution of Hadoop (Cloudera, 2014) is used. This is purposively selected for the experiment because Cloudera is one of the leading distributions of Hadoop and is pre-packaged with all the tools selected for this study which are Hive, Spark and Impala. In the following sections, the cluster setup is presented, followed by a description of the workload and finally a description of the method of evaluating the performance of the selected technologies. The initial intent is to evaluate and compare five technologies; Spark, Hive, Impala, Pig and Hbase which are described in detail in Chapter 2, Section 2.5.5. However, after review of existing literature, Hbase and Pig were discarded. Hbase was discarded because it is merely a data store while Pig appears to have been superceded by new developments in the BD analytics space, as it is no longer widely used.

ii) Cluster configuration

A multi-node cluster of four computer servers (nodes) is setup in a computer laboratory at CPUT. The most recent version of Cloudera, CDH 5.3.3 (Cloudera, 2014) is installed on the cluster with its default versions of Hive, Impala and Spark. In the cluster, no performance tuning is conducted by either, changing the operating system configuration, or CDH settings. To setup the CDH cluster, the step-by-step configuration procedures adopted from Noll (2014) and Cloudera (2014) are used. The essential packages for the experiment in this study installed on the cluster include Hive, Spark, Impala, YARN (MR2), Zookeeper and HDFS. Each node on the cluster is installed with a 64 bit operating system, Ubuntu 12.04.4 LTS, codenamed Precise. The hardware specifications for the four cluster nodes are shown in Table 3-1. In total, the cluster for this research has 72G RAM and 1050 G (1Tb) of hard disk space.

Table 3-1 Experiment: Hardware specification

Role	Name	Disk Size (G)	RAM (G)	OS Type	Processor	Network Card
NameNode	Master	300	32	64 bit	I3-2100 CPU @ 3.10 GHzx4	1000 Mb/s
Datanode	Hadoop1	250	16	64 bit	I3-2100 CPU @ 3.10 GHzx4	1000 Mb/s
Datanode	Hadoop2	250	16	64 bit	I3-2100 CPU @ 3.10 GHzx4	1000 Mb/s
Datanode	Hadoop3	250	8	64 bit	I3-2100 CPU @ 3.10 GHzx4	1000 Mb/s

iii) Cluster architecture

A Hadoop cluster must have a NameNode and one or more data nodes (Mctaggart, 2008). One of the nodes on the cluster is configured as the NameNode (Resourcemanager) because it has more memory and hard disk space. This is named “master”. All four of the nodes including the “master” node, are also configured as data nodes. Three of the data nodes are named ”hadoop1”, “hadoop2” and “hadoop3”. One data node, “hadoop2”, is configured to host a MySQL database server which hosts the cluster’s meta-store mainly used, to define structures of tables for Hive, Impala and sparkSQL. Figure 3-1 depicts the experiment’s Hadoop cluster architecture design used in this study.

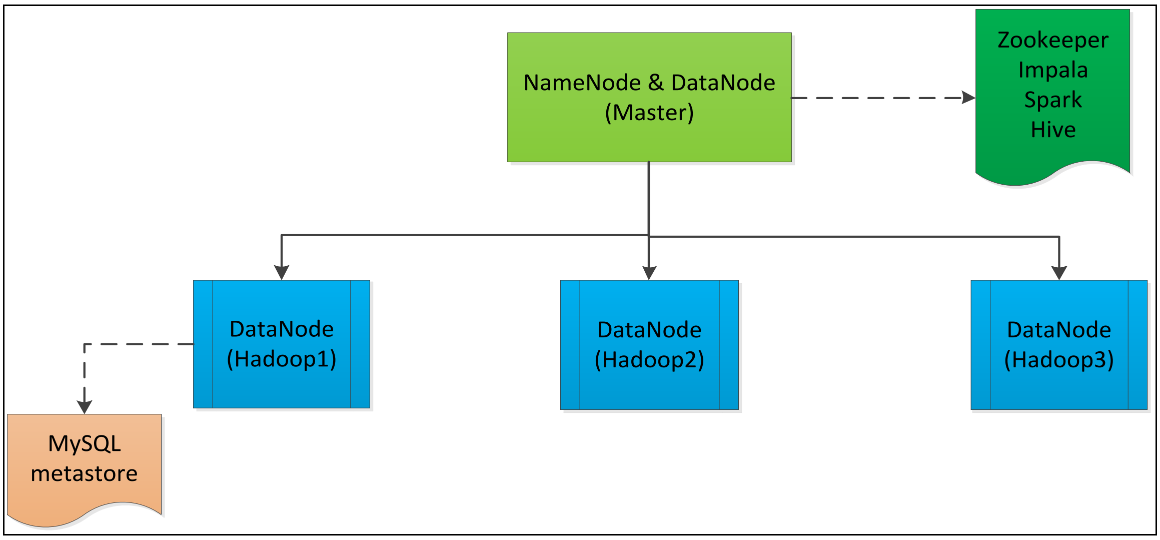


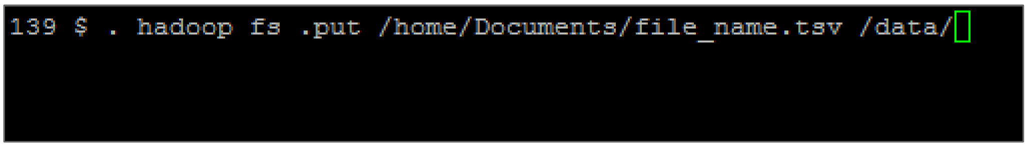
Figure 3-1 CDH Architecture for experiment

3.6.2.2 Workload

In this research, workload is used to describe the data sets used for the experiments. It is noteworthy to mention that sourcing data to use for the experiments was difficult. One of the organisations from where two participants are drawn for this research has to provide two sets of real world log files. Confidential information related to the company and its clients are obfuscated for confidentiality reasons before the data could be made available for the research. One text file contains attributes of publishers for on-line advertising. The file has 55 columns and 12 million records. The second file contains on-line conversions data for the company. The file has 25 columns and 53 million records. Both files are tab (\t) separated. The two file structures are specified in appendix B and C. Initially, the total size for the data provided for the experiments was 60GB.

3.6.2.3 Experiments Execution

Preparing for the experiments, involved copying the files described in Section 3.6.1.2 from the operating system, onto HDFS using the command depicted in Figure 3-2.

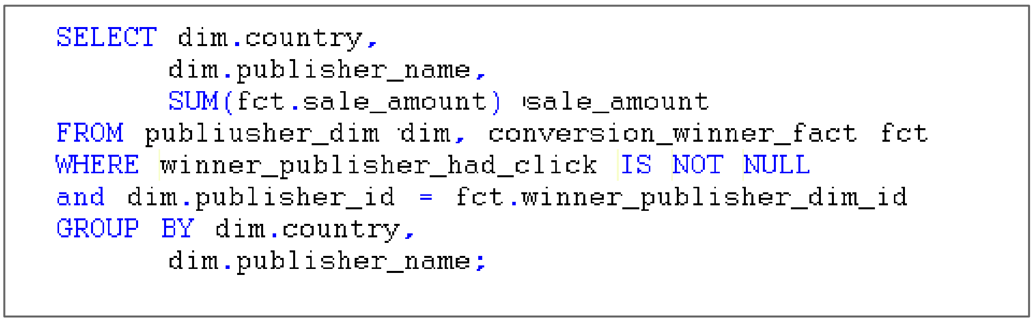


```
139 $ . hadoop fs .put /home/Documents/file_name.tsv /data/
```

Figure 3-2 Command to copy file from OS to HDFS

The goal is to store the data in a central location such that all the tools to be used for the experiments are able to access the data. After the files are stored in HDFS, external tables are created for each file in Hive through the Hive command line. In order to test Hive, Impala and Spark against larger data sets, a parquet file format table (ApacheParquetIncubator, 2014), needed to be created in Hive to store the conversion winner fact file's data. This enabled storage of approximately 10 GB, 50 GB, 100 GB, 250 GB, 500 GB and 1TB, uncompressed data in the parquet table.

In order to ensure consistency of results, each tool is executed against each of the above datasets for ten times and the time taken to return results in seconds(s) is recorded. The behaviour of CPU utilisation, memory consumption and disk I/O is also observed and recorded for the whole cluster. After each execution of each tool, the whole cluster is restarted to ensure that cache memory is cleared, and that every execution has almost the same resources available for each test run and at the same time, ensuring independence of each execution from influence from the previous tool execution. Figure 3-3 depicts the query that is executed in the experiments to compare the three technologies. This query is adopted from the organisation's common queries used for reporting when they want to know the total amount of sales made per country and per publisher.



```
SELECT dim.country,
       dim.publisher_name,
       SUM(fct.sale_amount) sale_amount
FROM publiusher_dim dim, conversion_winner_fact fct
WHERE winner_publisher_had_click IS NOT NULL
and dim.publisher_id = fct.winner_publisher_dim_id
GROUP BY dim.country,
       dim.publisher_name;
```

Figure 3-3 Query executed

3.7 Data analysis techniques

In this section, a discussion of the techniques used to analyse and interpret collected research data is provided for the purpose of building a theory to communicate the essence of what the data reveals. In Sections 3.7.1 and 3.7.2 the researcher discusseses in more detail the aspects of qualitatative and quantitative data analysis techniques, respectively for this research.

3.7.1 **Qualitative data analysis**

According to De Vos, Strydom, Fouche and Delport (2011), qualitative data analysis involves bringing order, structure and meaning to the mass of data. Content analysis as defined by De Vos *et al*. (2011) as textual analysis, is used to analyse the narrative data collected using interviews. Qualitative data text can be in the form of focus group notes, observations, interviews, written texts, visual images, and any tangible interpretable artefacts, but in this research, it is in the form of interview notes. Qualitative content analysis involves grouping data together into chunks and then assigning them to broader categories of related meaning. In this way, the data is structured into codes and themes which can then be applied to all the text. Patterns embedded in the text can be identified and more categories and sub-categories can be developed. This process is called coding, which enables a researcher to contribute findings to the discipline under study (De Vos *et al*, 2011).

In preparation for qualitative data analysis, the researcher uses the guideline suggested by Creswell (2008). The preparation involves organising, arranging, and having a general sense of the information that is collected. Each interview is recorded correctly with the aid of voice recording interview session. According to Creswell and Plano Clark (2011:206), "…preparing qualitative data involves organising the document or visual data for review or is transcribed and stored as a text document, ready for analysis.

Coding

According to Denscombe (2007), coding is extracting small units (phrases, sentences or paragraphs) from text, assigning them into defined categories, and then grouping the defined categories into themes, identifying relationships among the themes and categories. The author defines codes as tags or labels attached to collected raw data. The coding label can come from the exact words of the participants (i.e., in vivo coding), phrases composed by the researcher, or concepts used in the social or human sciences. The core feature of qualitative data analysis is the coding process. In this research, some categories were identified during literature analysis which was then used to build interview questions. The same categories were used as codes and categories for qualitative data analysis but with additional codes derived from the words of interviewees as stated by Denscombe (2007).

3.7.2 **Quantitative data analysis**

In quantitative data analysis, the researcher analyses data based on the type of questions or hypotheses and uses appropriate statistical tests to address the questions or hypotheses (Mouton, 1996). Teddlie and Tashakkori (2009), state that quantitative data analysis involves the analysis of numerical data using techniques that include:

i) Simply describing the phenomenon of interest.

ii) Looking for significant differences between groups or among variables.

Mouton (1996), states that quantitative data can be analysed statistically or mathematically. There are two main categories of statistics, namely descriptive statistics and inferential statistics (Dunn, 2010; Mouton, 1996). Descriptive statistics is concerned with organising and summarising available data to make it more comprehensible, while inferential statistics is concerned with the kinds of inferences that a researcher can make when generalising data collected (Saunders *et al*. 2009; Blaikie, 2004 & Mouton, 1996). In this research, descriptive statistics is found to be appropriate and is used to show the relationship between the three technologies that are evaluated and compared. The relationship between the size of data and performance of each technology is also presented using descriptive statistics. Denscombe (2007) and Dunn (2010) describe six main categories of quantitative data namely nominal, ordinal, interval, ratio, discrete and continuous. As stated in Section 3.6.2, the quantitative research phase collected three primary variables; execution time measured in seconds, data size measured in gigabytes (GB) and the technologies used to analyse the data. Volume (data size) and execution time are numeric continuous variables while technology is a nominal variable.

3.8 **Delineation**

This research does not propose or design a new system or a new architecture for a real-time decision support system. There is no comparison and evaluation of all layers of a real-time decision support system as the focus is on analytics tools only. Data analysis was restricted to structured data. This research does not propose an end-to-end BD benchmark. Scalability and fault tolerance of technologies is not evaluated in this research. This research does not delve into detail on how Spark, Impala and Hive process queries but merely note the differences or similarities in execution time. It must be noted that due to time and resource limitations, the real-time aspect is tested using data ‘at rest’, that is, data stored in HDFS and not in a streaming environment.

3.9 **Ethical issues**

This section provides an outline of ethical issues around this study and how they were addressed. Gray (2009:68) describes ethics as moral principles adopted by a researcher which are concerned with ”the appropriateness of the researcher’s behaviour in relation to the subjects of the research or those who are affected by it.”

As indicated in Chapter 1, two research techniques are used; a qualitative interview of BI experts with knowledge of BD and BD technologies from organisations in South Africa were interviewed, and an experiment conducted. One interviewee’s organisation provided test data used as a load to carry out computer laboratory experiments. In this research setting, there are various ethical issues that needs to be addressed, as highlighted by Maree (2012), Resnik (2011) and Gray (2009). Ethical issues relating to human participants are addressed and strategies are designed to deal with challenges relating to confidentiality, anonymity, right of privacy, voluntary participation, protection from harm and trust (Maree, 2012; Resnik, 2011 & Gray, 2009). In this regard, data collection, data analysis, interpretation and reporting of research findings are conducted in an ethical manner (Maree, 2012; Gray, 2009). Furthermore, the research process conforms to the code of ethics of scientific research in general, and also the code of ethics of the supporting organizations not to infringe with stated organisational ethics. In this regard, legally binding documents (Non-Disclosure Agreement) stating the rights of the organization, individuals (interviewees) and the CPUT as the sponsors of this research study are signed. Letters of consent and right to confidentiality are obtained from the participants and also from management of the participants’ organisations.

The data collection is conducted in an ethical manner by ensuring validity and trustworthiness as stated by Maree, (2012) in order to avoid deception as stressed by Bless *et al.* (2006) and Gray (2009). It must be noted that data analysis and reporting is conducted in an ethical manner to avoid fabrication and falsifying information (Bless *et al.*, 2006). According to Bless *et al*. (2006:145), ethical responsibility rests with the researcher who "…should report on technical shortcomings, failures, limits of the study, negative findings and methodological constraints".

In summary, the following ethical issues were addressed in relation to this research:

- Autonomy (respect for the person and a notion of human dignity) – In this regard, signed letters of consent from the interviewees were obtained. Before the interviews were conducted, participants were acquainted with the study and furnished with the interview questions before the consent letter was signed. The participants were informed of their right not to participate and to withdraw any time as stated by Gray (2009) and no one was forced, either overtly or covertly to participate (Bless et al, 2006).
- Beneficence (benefit to the research participant) – According to Gray (2009), this covers what the participants will gain from the research. Bless *et al*. (2006), state that it is essential for a research project to also potentially contribute to the interests of others. The organisations from where the participants were drawn will be provided with a report of the study. They were made aware that copies will be retained by CPUT.
- Non-malfeasance (absence of harm to the research participant) – According to Bless *et al*. (2006), the basic principle of research is that participants should not be harmed in the research project either intentionally or non-intentionally. On this regard, the name and logo of the selected organisations remain anonymous in this research for reasons of confidentiality. The interviewees are referenced by their job titles only. The test data obtained as load for experiments was first muddled in order to remove any reference to their name or their

clients' identity for confidentiality purposes. This was to ensure that the organisation's integrity will not be put in jeopardy by this research.

- Justice (notably distributive justice) – equal distribution of risk and benefits between communities. According to Bless *et al*. (2006:142), "the principle of justice is based on the assumption that all people are equal". The authors further posit that in the research project there must be no suggestions of discrimination based on race, gender, disability status, income level or any other attribute or a participant. Although this aspect of ethics is very important, it was not applicable in this research.
- Fidelity – This principle requires faithfulness and adhering to agreements between the researcher and the participants. According to Bless *et al*. (2006), the researcher must not deceive or divulge any confidential material as this is ethically wrong. The authors also emphasize the importance of respecting the participants' rights and dignity by the researcher during the research process.

3.10 **Summary**

In this research the ontological view of subjectivism is taken while the epistemological stance of pragmatism is selected in order to explore the concept of BD and identify factors, that may influence the selection of technologies appropriate for real-time DSS in a BD environment. The exploratory sequential mixed methods design approach is used in this research. The research is conducted in two phases. In the first phase, qualitative data is collected using semi-structured interviews in order to propose a framework of factors that influence the selection of technologies appropriate for real-time DSS in a BD environment. Content analysis is used to analyse data gathered using semi-structured interviews. In the second phase of the research, quantitative data is collected by evaluating and comparing Hive, Impala and Spark. Quantitative data is analysed statistically. The objective of the quantitative study is to test if the three technologies (Impala, Spark and Hive) have different query execution times when analysing data. The results of the quantitative data are used to support the results of the qualitative research. The next chapter provides a discussion of the research findings.

4 CHAPTER FOUR: RESEARCH FINDINGS

4.1 Introduction

According to Babbie and Mouton (2001), all fieldwork culminates in the analysis and interpretation of collected data. As discussed in Chapter 3, this research uses both, qualitative and quantitative data.

The researcher uses a semi-structured interview guide, shown in Appendix A, where the interview questions are formulated, after the literature review in Chapter 2. The researcher ensures that the interview questions are unambiguous as far as possible. During the interviews, all words and concepts deemed unclear, are clarified. The participants' responses are recorded using notes and a mobile device operated as a voice recorder. At the end of each interview session, the collected information is transcribed. The researcher then applies content analysis to analyse the data by focusing on common words, sentences and themes governed by the interview guide. The quantitative data is collected by means of an experiment to evaluate and compare Hive, Impala and Spark. The data collected from the experiments are analysed statistically using SPSS software.

In the following sections both, the qualitative and quantitative results and findings are discussed in detail. Qualitative findings are presented by linking the research problem, research questions and interview questions to the answers of the participants. The quantitative findings are presented after the statistical analysis procedure on the data collected from the experiments.

4.2 Qualitative findings

Interviews

In this research, a total of ten participants are interviewed. The participants are drawn from different companies operating in different industrial sectors in South Africa as depicted in Table 4-1. The findings from the interviews of each research question, are discussed in this section.

Table 4-1 Interview Participants

Participant No.	Designation	Industry
1	Chief Technology Officer	Customer Relationship Management
2	Business development manager	Internet Service Provider
3	Financial Director	Finance
4	Head Business Intelligence	Software Development (BI)
5	BI Analyst	Software Development (BI)
6	Manager Business Intelligence	On-Line media marketing
7	BI Analyst	On-Line media marketing
8	Manager	On-Line media marketing
9	BI Manager	Retail
10	Network Analyst	Networking

RQ1. What are the factors that influence the selection of technologies appropriate for real-time DSSs in a BD environment?

This research question sought to identify factors that drive the selection of technologies appropriate for real-time DSSs in a BD environment. The research question is divided into two sub-research questions, each of which are further used to formulate the interview questions. In the following sections, the researcher discusses the findings of the sub-research questions.

SRQ1.1. What is the relationship between characteristics of data and technologies used for analysing data in a real-time environment?

This sub-research question affords interviewees to describe the characteristics of data used for decision making in their organisations, and the technologies used to analyse the data. It also allows the researcher to establish any relationships that exists between the characteristics of data and the technologies actually used to analyse the data, and therefore, identify factors that drive the selection of technologies appropriate for real-time DSS in a BD environment. Eight interview questions comes from this sub-research question and the responses of these are now presented.

IQ1. What are the sources of data in your organisation used for decision making?

This question is used to identify the sources of data available within organisations, from where the participants are drawn. The question is also used to establish any relationships between data sources and the technologies appropriate for real-time DSSs in a BD environment. As shown in Table 4-2, all the participants interviewed acknowledged the existence of multiple sources of data in their organisations.

Table 4-2 Sources of data

Source of data	No. of Participants	Technologies used
Internet and on-line application systems	10	Hadoop, traditional BI tools, NOSQL databases and In memory databases.
OLTP systems	10	Traditional BI tools
Digital systems (e.g. sensors)	2	Traditional BI tools
Monitoring systems (e.g. CCTV)	1	Data is not analysed
Social media (e.g. Facebook and Twitter)	2	Hadoop and Google analytics
Audio voice recording systems (e.g call centre)	1	Data is not analysed
Mobile network data traffic	1	Analytics tool called Sandvine

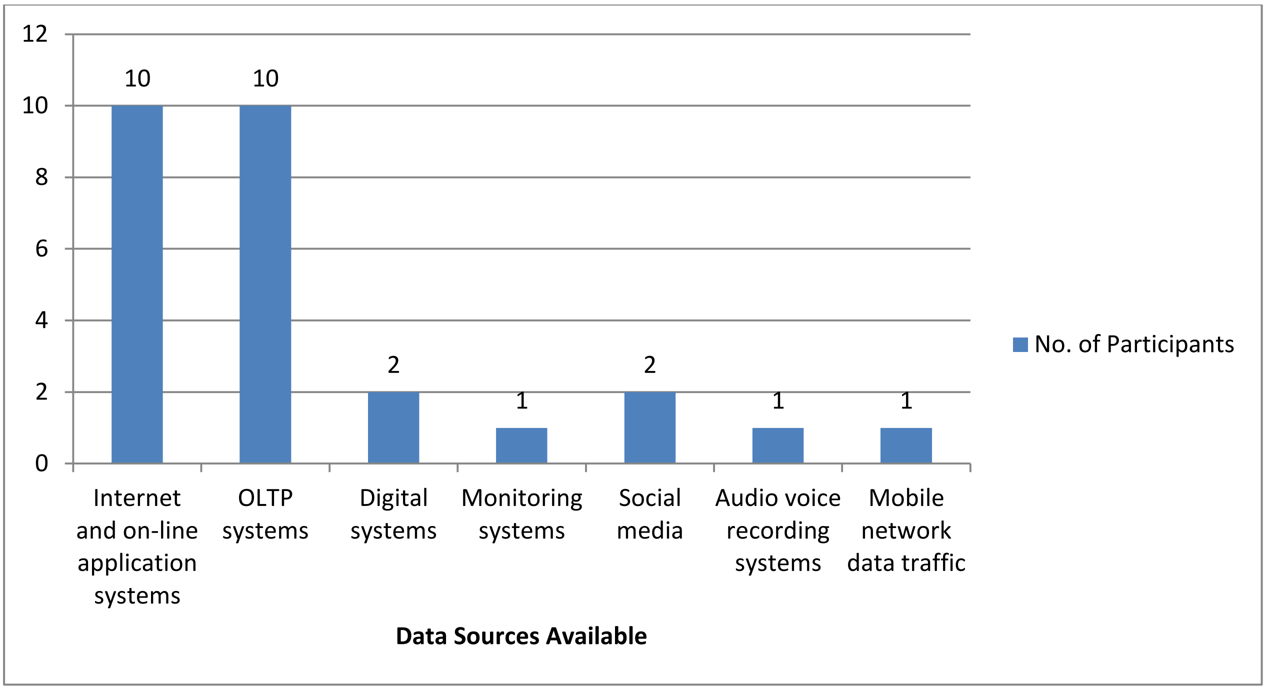


Figure 4-1 Data sources graph

Refering to Table 4-2 and Figure 4-1, all the participants in this research use data from OLTP systems for decision support purposes. Of the ten participants who source data from OLTP systems, Table 4-2 depicts that all of the participants only use traditional BI tools to analyse data.

Furthermore, the data in Table 4-2 and Figure 4-1 reveal that all participants use data for decision support which is generated by the Internet, and on-line application systems such as, on-line stores and websites. The graph in Figure 4-2 depicts the relationship between data sourced from the Internet and on-line applications, as well as the technologies used to analyse that data. The results indicate that no participants use Hadoop only to analyse data, and three participants only use traditional BI tools. None use either in-memory databases or NOSQL databases only and six use a combination of any of these two and traditional BI tools.

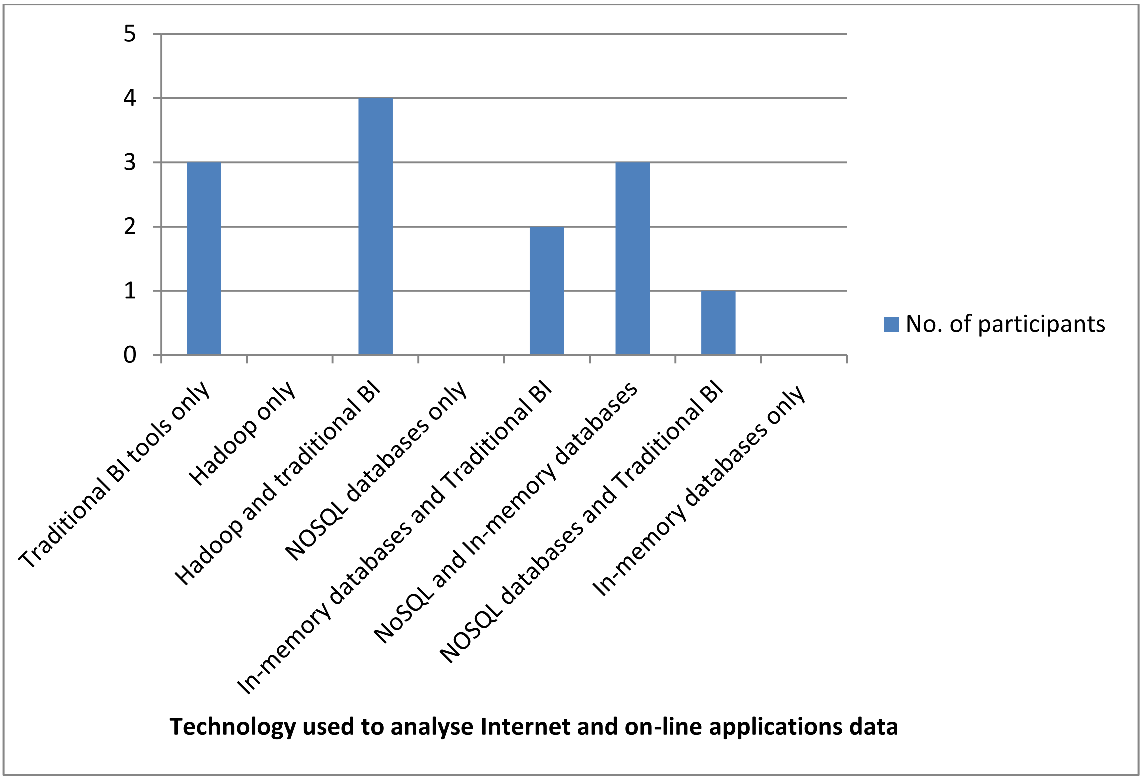


Figure 4-2 Technologies used to analyse Internet and on-line data

The data in Table 4-2 depicts that two participants mentioned that they use source data from digital devices and sensors, but this data is stored in relational databases from where it is integrated into data warehouses through ETL processes. This effectively means that these two participants rely on traditional BI tools to analyse data from this source.

According to participant (P2), his organisation’s main sources of data for decision making are financial and billing OLTP systems. In addition, P2 mentions that his organisation sources data from digital devices and sensors on the network. The participant states: “The company has a whole bunch of core routers and other networking devices. Each of those devices generates usage data which helps us to know how much utilisation we have and how much capacity we have in our network at any given time.” (Appendix P:134).

According to participant P3, in addition to a financial and billing system, another source of data used for decision making is their company website (Internet and On-line system). This data is however, stored in a relational database. Participants P5 states that data used for decision making in his organisation obtained from a call centre, comprises 99% OLTP systems and 1% online survey information. The survey information is stored in a relational database.

Participant P9 states that his company uses network traffic as the main source of data. The company analyses video, audio and other file formats on various networks as users are watching videos online or downloading them. Participant P9 states: “We monitor network traffic and analyse metadata about on-line videos such as Youtube videos as they are being watched across the country or as they are being downloaded. Our goal is to analyse this data in real-time and provide our clients with reports as the activities are happening for corrective action or for improving services.” (Appendix P:134).

All participants acknowledge that their organisations have active accounts on social media platforms such as Facebook and Twitter, but only P1 mentions that his company has started analysing data in near real-time mode using BD technologies. In his own words, P1 states: “Our data includes customer related information collected from Facebook and Twitter. Facebook has a graphing API, and you can get access to feeds and posts through this API. Similarly, Twitter has an API and we are able to integrate easily. We use Text Analytics API to analyse sentiment on this data in real-time. So the main source of data is Internet and on-line applications as well as OLTP systems in our organisation." (Appendix P:134).

P10 mentions that his organisation currently relies on Google analytics to get sense out of data sourced from the company’s Facebook page and states: “Our company has a Facebook page with over a million followers. On this page, we receive comments about our services and products from different parts of South Africa and so we need to know what they are saying about our brands and shops. At present we rely on Google analytics to get information from our Facebook page and we also have someone who is dedicated to monitoring this page for sentiment analysis and this is really difficult. It would have been ideal to have platforms that can assist us to analyse this data possibly in real-time.” (Appendix P:134)

According to participants P6, P7, P8, P9 and P10, OLTP systems, Internet and on-line application systems, constitute their main source of data used for decision making in their organisations. However, the participants acknowledge the presence of other sources of data such as, website logs, video and call centre voice recorded messages and information accumulating on their Facebook and Twitter accounts. None of these participants are leveraging these data sources at present.

Finding 1: OLTP systems constitute the major source of data used for decision making in organisations. Traditional BI tools are the main technology used for DSS.

Finding 2: Organisations that analyse internet and on-line applicaions use BD technologies such as Hadoop, NOSQL databases and In-memory databases. Organisations which analyse Internet and on-line applications data, first store the data in relational databases before analysis is done.

Finding 3: Although organisations are gathering data from social media platforms such as Facebook and Twitter, this data is not fully used in decision making.

Finding 4: Unstructured data sources such as CCTV are not being leveraged by organisations.

IQ2. How would you describe the data available in your organisation used for decision making?

This question aims to identify the attributes of data that exists within organisations from where the participants are drawn. The researcher uses this information to establish if any relationships exists between the attributes of data, and the technologies used to analyse the data.

Table 4-3 Characteristics of data data available in organisations

Characteristic of data	No. of Participants
High volume	9
Quickly changes	9
Comes from different sources	10
Comes from a single source	0
Unstructured (Files, Videos)	3

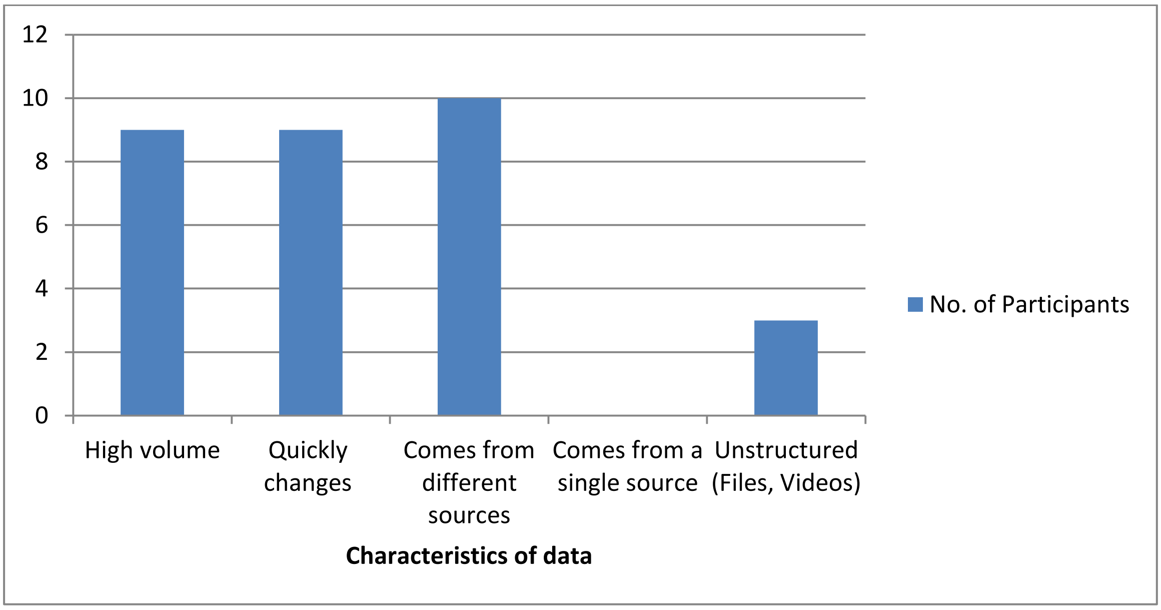


Figure 4-3 Characteristics of data available in organizations

As depicted in Table 4-3 and the graph in Figure 4-3, nine participants describe their data to change quickly and/or arrives quickly into their analytics platforms. For example, P6 states that: “Data is added into the system every second. There are few changes to the data once it has landed into the systems but it comes in very fast. A lot of data comes in very quick and so we do a lot of inserts but few updates.” (Appendix Q:135).

All participants describe their data originating from multiple sources, while none of the participants state that data used in their organisations for decision making, originates from a single source. For example, participant P2 states: “…our data comes from different places” (Appendix Q:135). Nine participants describe their data being high volume, while only one participant states that his organisation analyses low volumes of data, but the data has complex structures to analyse.

From the data collected, it is evident that participants organisations that generate high volumes of data and depend on traditional BI tools, struggle to process data in real-time. All participants describe their data as structured in nature. According to participant P2, the structure and the format of data analysed is known in advance and the participant states: “We always know the format and structure of the data that we analyse in advance. We are an Internet service provider and we have other people’s data moving across our network but we have kind of metadata about that data and we use that for our decision making. We need to know how much data is being moved but we don’t know anything specific to the data moving in our network. Our main challenge with this data is that it comes in very high volume and is difficult for us to analyse in real-time with the technologies we currently have.” (Appendix Q: 135).

Participants P4, P5, P6, P7, P8, P9 and P10 also state that their data for decision making is structured in nature but the volume of that data is continuously increasing. Participant P4 states: “…the other challenge we have with the structured data is that it is massively growing in volume. We need technologies that are scalable as the volume of both structured and unstructured data increases.” (Appendix Q:135).

According to P10, although the data used for decision making in the organisation is structured in nature, the rate at which the structure changes, and the ability for the development team to reflect those changes in reports for business decision making is impacting decision makers. Participant P10 states: “We need technologies which can allow any change in data type or structure to be reflected immediately in our analytics and reporting environment. This is not possible with existing technologies because of a long development cycle of the ETL process.” (Appendix Q:135). According to P6, the organisation analyses structured data stored in a relational database and the participant says: “In our organisation, we track actions. There is a fact table which keeps clicks, impressions and other buying activities between media partners and clients.” (Appendix Q:135). Two participants mention that their organisations analyse unstructured data from social media platforms but one of these relies on Google analytics for reporting purposes. One participant, P9 states that his organisation analyses unstructured data in the form of Youtube videos and other files downloaded from the internet (Appendix Q:135).

Finding 6: There is a relationship between the volume of data generated by an organisation and the selection of technologies used to analyse that data in real-time.

Finding 7: The format (structured or unstructured) of data available in an organisation has an influence on the technology that can be used for DSSs.

Finding 8: The rate at which updates occur in the source systems, and the need to reflect those changes to the reporting and analytics environment, determines the type of technologies that are appropriate for real-time decision support.

IQ3. How would you describe the volume of data available for decision making in your organisation?

The researcher uses this question to determine what the impact of high volume of data is on the choice of technologies appropriate for real-time DSSs in a BD environment. After asking participants to describe the volume of data generated and analysed in their organisations, the participants are requested to state the technologies they currently use to analyse data and what would be ideal for real-time DSSs. Table 4.4 provides scales of data available within organisations for use in decision making.

Table 4-4 Data volume scales available in organisations

Volume description	No. of Participants	Technologies used to analyse data
Very high – Zeta bytes and above	1	Generates zeta bytes but cannot analyse it.
High – Petabytes	1	Uses analytics tool called Sandvine.
Medium - Terabytes	8	All participants use traditional BI tools. Three participants have started using BD technologies such Hadoop, NOSQL and In-memory DB.
Low - Less than Terabytes	1	Uses MongoDB and Redshift.

As depicted in Table 4-4, eight participants stated that their systems receive medium volume of data and one participant states that the data in her organisation is low volume, but with complex structures which are difficult to analyse with traditional BI tools. Nine participants state that the volume of data in their organisations is continuously growing and this is impacting their databases and data warehouse technologies' performance. For instance, P10 says: "We have seen a sharp increase in the volume of data generated when we bought a new IBM campaign management system. Every day, the system sends out millions of marketing email and SMS messages to clients and prospects. The system also receives messages back from the targeted people. We have high volume of data being generated but we are struggling to analyse this data especially in real-time as we have to wait for the ETL process to run over night." (Appendix R:137).

Findings 9: Organisations with high data volumes and using traditional BI technologies are facing scalability challenges.

Finding 10: Technologies that can scale with volume of data and that can allow reporting to be in real-time are ideal for high volume data.

IQ4. How would you describe the rate at which data changes within your systems for analytic purposes?

The objective of this question is for the researcher to ascertain how often data arrives in the system, or how often the data is updated. The participants are asked to disclose what they currently use to analyse this data, and what would be the ideal technologies to analyse it in real-time. As seen in , nine participants state that data changes within their systems per second, while one participant states that data changes per day in the source systems.

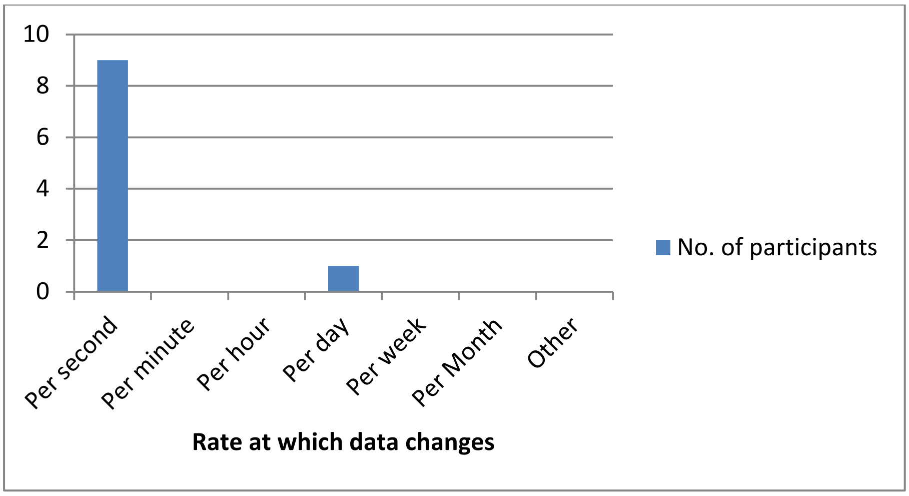


Figure 4-4 Rate at which data changes

According to participant P6, data changes quickly and arrives quickly into the analytics platform but currently, it is not analysed in real-time due to technological limitations.

Participant P2 finds data arrives and changes per second, but is analysed per hour, per day and per month. P2 further states that: “The data arrives into our systems in real-time but we base our analytics per day. We aggregate and look at it per day. Some data is analysed per hour except for system logs, which are obviously per second, but those are just for platform performance monitoring. The data changes quickly but is analysed per day.” (Appendix S:138)

According to participant P1, data within their systems change at a very high rate: “We use BD technology to assist us with managing the velocity of the data coming in. Data arrives per second and there are a lot of changes on data per second.” (Appendix S:138)

Finding 11: Although data arrives and changes quickly within the systems for the organisation of nine participants, none of the organistions interviewed analyses data in real-time due to technology limitations. Two organisations analyse data in near real-time (one hour after data has arrived in their analytics platform) but their desire is to analyse the data in real-time.

Finding 12: The rate at which data arrives and gets updated in a system, influences the technologies used to analyse it in real-time.

IQ5. What type of data formats/structures exist in your company?

This question sought to establish if there is a relationship between the structure or format of data in an organisation and the technologies that are appropriate for real-time decision support. The participants are asked to state the data formats available in their organisations and the technologies they are using to analyse that data. The participants are further requested to state the technologies that would be ideal to analyse that data. According to the data collected, all participants state that they analyse data structured in files or relational databases. Traditional BI tools are used to analyse this data.

Five participants namely P1, P9, P10, P4 and p5 state that they generate and store data in image, video and audio format while all the participants mentioned that they also have data generated from social media platforms such as Facebook and Twitter. Participant P10 states that his organisation has started analysing data from Facebook using Google analytics but not for real-time DSS. Although the five participants state that they store unstructured data in the form of images and videos, P9 states that his organisation actually analyse these data formats for decision making. Participant P9 indicates further, that his organisation analyses Youtube videos on data networks. P9 states: “This includes how long the video was watched; the time it was watched, location in which the video was downloaded and how much bandwidth was consumed.” (Appendix T:139)

Finding 13: The format and structure of data influences the selection of technologies appropriate for real-time DSSs in a BD environment.

SRQ 1.2 What are the tools available on the market for use in analyzing BD?

IQ6. What do you currently use to analyse your data?

From the data collected, all the participants indicate that they currently use traditional database and DW as the main technologies for analysing data in their organisations. The interview results also reflect that none of the participants uses Hadoop as the only tool for DSSs but some organisations have started moving towards the NOSQL and Hadoop platforms in order to achieve the goal of real-time DSSs and overcome the challenge of BD.

Two participants, P1 and P6, indicate that in addition to the DW, they have started using Hadoop to analyse data. P6 states that his organisation is in a transitional phase of moving their analytics platforms from traditional database and data warehouse to Hadoop environment. Participant P2, states that his organisation has started exploring the use of Hadoop as a platform for data analytics.

According to participant P2, the organisation uses different technologies to analyse different types of data. P2 states: “We currently use traditional database and data warehouse technologies namely Sybase, Oracle and Greenplum. We also use Round Robin Database (RRD) which is a time series database. We have a plan to use Hadoop but we are not there yet. However, we sell Cloudera Hadoop to our clients.” (Appendix U:140). Three participants P1, P2 and P5 in this study use in-memory databases for analysing data. P2 and P5 use Greenplum while participant P1 uses SAP Hana.

P1 and P3 use NOSQL databases to analyse data. P1 states that his organisation uses Hadoop and NOSQL databases to analyse unstructured data and the participant states: “We use NOSQL databases because data like Facebook and Twitter continually change their data structures which are usually in JSON file format. So change in data structure is a big thing in BD analytics. NOSQL databases can handle changes in data structure more easily, as you don't have to define the structure upfront. Structure is implicit in the JSON structure. You can query the items even if the structure changes dramatically. Another problem that NOSQL databases solve is performance with large complex data because you don't have to join tables as you don't have to normalise your data. NOSQL databases are superfast at summarizing and querying the data." (Appendix U:140).

Participant P3 responds: "We use a sharded MongoDB implementation, which we have mapped to Amazon Redshift so that we can analyse the data. We found that MongoDB was terrible for data analysis as queries were complex and slow. Even pretty basic queries would impact on customer experience so we upload deltas to Amazon Redshift every hour in order to improve system performance." (Appendix U:140).

In describing the technologies used in his organisation, P7 states: "We are an open source shop. We have sharded Mysql databases and an ETL tool. Then we have a reporting application called Infobrite but we now want to get out of the Mysql space and out of Infobrite into Hadoop. We want to replace our ETL tools simply because we are running into scalability problems. You cannot linearly add more machines and you cannot scale well as volumes of data increase. You don't expect your performance to double up after doubling up your machines. But on the Hadoop space you can scale with no problems. We have been using sharded databases but it's proving to be more expensive. We are also moving to Impala for reporting which runs off Hadoop. Our star schema is now on the impala platform." (Appendix U:140).

According to P4 and P5, their organisations still rely on traditional database and data warehouse for their DSSs although they have also started looking at Cloudera Hadoop. According to participant P5: "When we get to a point where we don't know what we are expecting, then we will move more towards BD infrastructure. At present, we will continue to use traditional database and data warehouse technologies because we know in advance what we are looking for in our data.“ (Appendix U:140).

Finding 14: Organisations are resorting to sharding databases in order to scale up to the challenge of high volume of data and complexity of data structures.

Finding 15: Scalability drives the selection of technologies appropriate for real-time DSSs in a BD environment.

Finding 16: The need to discover unknown patterns in data, has an influence on technology choice to analyse data.

Finding 17: The ability to obtain answers to questions not known in advance.

IQ7. If you are using Hadoop, which distribution do you use?

The objective of this question is for the researcher to explore different versions of Hadoop available on the market. After asking participants which distribution of Hadoop is used in their organisations, they are further asked to provide the reasons why they selected a particular version of Hadoop. From the data collected, only three organisations involved in this research actually use Hadoop. Two participants P1 and P7, state that their organisations use Cloudera Hadoop while P7 states further that, "Cloudera hadoop plays better with the existing infrastructure. It is a pure open source tool which is easy to maintain." (Appendix V:141). According to participant P4, his organisation "…has recently started using Pivotal distribution of Hadoop but they are still in the exploratory stages" (Appendix V:141). According to participant P1, "Hadoop and NoSQL databases are not yet fully mature" (Appendix V:141). Participant P1 agrees to an extent with P7, who states "…the maturity of analytics on Hadoop is still very low. There is a very small population of users of Hadoop in South Africa at the moment" (Appendix V:141).

Finding 18: The adoption of Hadoop by organisations in South Africa is still at its infancy stage. The few organisations that have Hadoop implementations already installed, are still in exploratory or transitional phases.

Finding 19: Maturity of technology has an influence on the choice of technologies.

Finding 20: The ability to integrate with existing technologies and systems.

IQ8. If you are using Hadoop, what analytic components do you use?

This question was initially designed to establish an understanding of the different analytics and reporting tools available on the market that can be used to report data stored in Hadoop. However, after discovering that most organisations are not actually using Hadoop, the researcher then asks participants to provide any reporting and analytics tools used in their organisations for decision making.

Participants P6, P7 and P8 respond that their organisations use Impala, Hive and Spark for analysing data stored in Hadoop. According to participant P6, the organisation uses an integration of Spark and Kafka for their near real-time ETL process. According to participant P1, his organisation recently started using Cloudera Hadoop and therefore, analyses data stored in Hadoop using Hive.

Participant P1 responds that his organisation uses a number of tools to prepare and analyse data used for decision making. P1 states that: "We use Hive because it uses a SQL variant to get data out of Hadoop. We also use SAS, R, Tibco Spotfire and Zoom Data which is very good at handling large data sets in seconds. With ZoomData one is able to visualize 1 billion rows, with additional 1 million rows per second in real time. In some cases we actually transfer the Hadoop data to MongoDB and I use ZoomData to analyse data. We also use Impala and Hive to analyse some data that is sitting on Hadoop. In addition to that we also use SAS, Tableau and R to analyse data that we query from Redshift. Finally, we also use Microstrategy for reporting and analytics." (Appendix W:142). Although participant P7 states that his organisation is now using Impala, Hive and Spark, structured data is stored in a Mysql database and DW for analytics and reporting purposes. The organisation of participant P7 uses Infobrite for reporting. The data is transferred from a sharded Mysql database to a DW using an ETL tool.

Finding 21: Ease of use of a technology has an influence on the selection of appropriate technologies. For instance, technologies which use a SQL variant to query data, are preferred by users.

Finding 22: Ability to analyse large datasets with low latency (sub-seconds) by a technology has an influence on the selection of technologies that are appropriate for real-time DSS in a BD environment.

RSQ2. How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?

SRQ 2.1 What are the existing guidelines, frameworks, criteria, or measures applicable when evaluating analytic tools for real-time DSS in BD environments?

The objective of this sub-research question is to identify important criteria or attributes to be considered, when selecting technologies that are appropriate for real-time DSSs in a BD environment. The data collected for this research question is now be discussed.

IQ9. When selecting analytic tools for real-time decision making in a BD environment, what criteria did you consider?

The responses to this question from the participants are summarised in Figure 4-5.

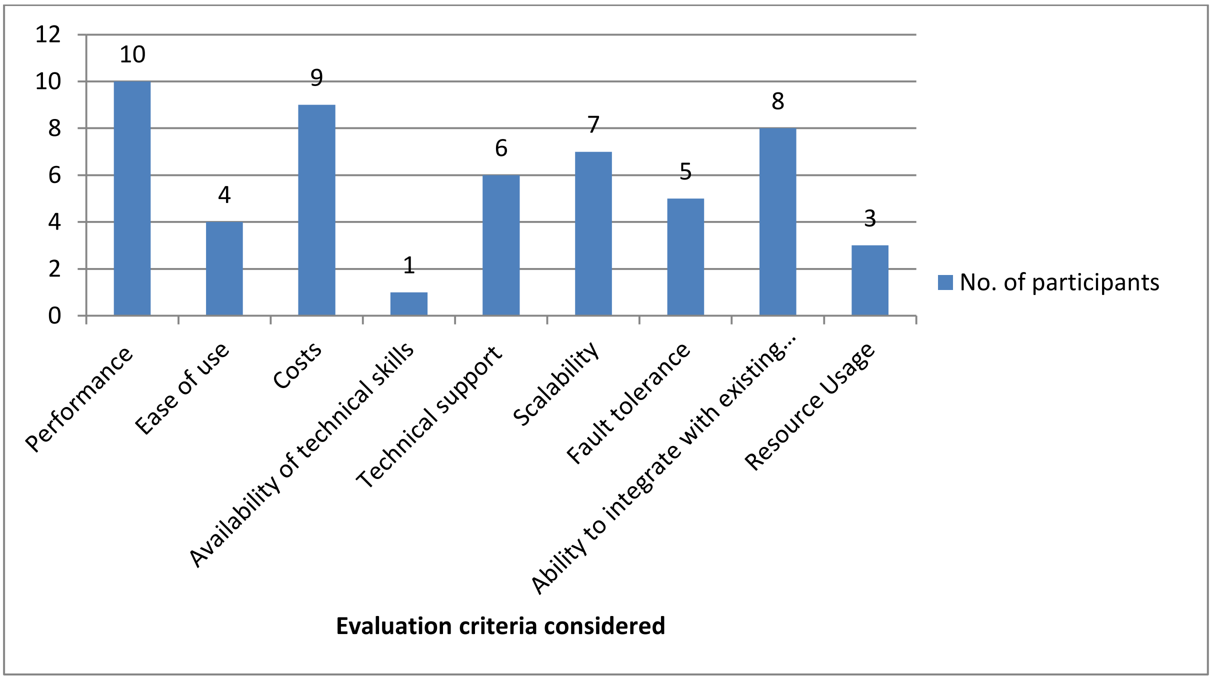


Figure 4-5 Evaluation criteria considered by participants

In Figure 4-5, analysed interview results are depicted showing that all 10 participants use performance of a system to assess technologies for DSS. It can be gleaned from the results that the performance of a technology in terms of throughput and/or latency, is the most important criterion considered when assessing technologies appropriate for real-time DSSs. Participant P4 states that: "…nowadays, memory is very cheap. The hardware doesn't drive requirements but performance influences the choice of technology used for analysing BD in real-time. How fast do we need to load data is what drives requirements. To us scalability and fault tolerance is also critical as our data volumes are continuously increasing, for example, adding RAM to the environment should be easy." (Appendix X:143). The comments made by P4 in this regard also show the importance of assessing technologies by looking at scalability, fault tolerance and the usage of computer resources.

Although four participants, P3, P4, P5 and P9 agree that technologies which are easy to use would be preferred over technologies which are not easy to use, all participants (P1 to P10), acknowledge that most BD technologies are still new in the industry and therefore organisations need to skill up their employees on how to use these technologies. The results indicate that availability of technical skills and ease of use are not important things to look at when selecting appropriate technologies for real-time DSSs in a BD environment. In relation to technical skills availability, participant P7 states that finding skilled people with expertise to manage and maintain BD technologies is difficult, and therefore, ease of use and technical skills availability are not very important attributes to consider when evaluating BD technologies. In support of the perception of P7, participant P3 states that the challenge of technical skills availability in BD analytics "…is not only IT related but it is also difficult to find people who can write statistical models." (Appendix X:143). Participant P5 states that it is essential to consider the availability of technical skills when selecting technologies that are appropriate for real-time DSS in a BD environment.

As depicted in Figure 4-7, eight participants (P1, P2, P4, P5, P6, P7, P8 & P10) are of the opinion that technologies appropriate for real-time DSSs in a BD environment, should be easy to integrate with existing systems. This could imply that existing technologies have an influence on the selection of technologies that are appropriate for real-time DSS in a BD environment. Participant P5 summarises this by saying: "Corporates are still cowed in legacy systems and one must consider how new technology will be integrated with existing technologies and other systems. There is need for continuity and you need to be able to connect to existing data sources with the new technologies." (Appendix X:143).

Participant P9 states that the exisiting data base of his organisations and data warehouse systems, were easy to integrate with the new analytics system that was acquired. Participant P9 states: “…my company had a product from the same supplier which was going to be easy to integrate with the new system for data analytics. So the existing supplier and the existing technology had an influence on the technology that was eventually selected.” (Appendix X:143).

On the aspect of the ability to integrate with existing systems, participant P7 emphasises the need for consistency in reporting when an organisation switches from an existing system to a new technology. Furthermore, participant P7 states: “When switching, make sure that your reporting remains consistent. We must ensure that we don’t have a single point of failure.” (Appendix X:143).

Participants P1 to P10 state that they use costs to assess and determine appropriate technologies for DSS. Participant P2 explains clearly how his organisation considers costs when selecting technologies: “When we consider costs, we look at TCO (total cost of ownership or total cost of the infrastructure over the time period). We look at how much it is going to costs us for hardware, how much software will cost and how much maintenance will cost. We consider all these things and then we compare three or four different technologies. So we look at it over the lifetime of the service and not just one particular individual item. So our decision is based on TCO. Some products are licenced per user, some per node and some per CPU core.” (Appendix X:143).

From the interview results, another important aspect that should be considered when determining technologies appropriate for real-time DSS in a BD environment, is the licensing which in turn, has an impact on costs. Participant P4 raises this aspect and is of the opinion: “In a data system environment, the system can grow in the number of CPU cores, the number of users or number of nodes. The moment you add a new node or a new user, it means costs will go up. One needs to understand the type of licencing mode because some technologies are licenced per user, some per node and some per CPU core.” (Appendix X:143).

Participants P6, P7 and P8 agree that they opt for open source technologies such as Hadoop and Spark because costs associated with building and maintaining such an environment, are lower compared to commercial technologies.

From the interview results, participants P4 and P5 are of the opinion that the availability of technical support for a technology is an important criterion when evaluating and selecting technologies appropriate for real-time DSSs.

The results also reveal that there are three important resources which to consider when determining technologies appropriate for real-time DSS in a BD environment namely; i) CPU utilisation, ii) memory usage and iii) disk input/output.

Nine participants (P1, P3, P4, P5, P6, P7, P8, P9 & P10) state they consider CPU usage when selecting technologies. Participant P2 for example, states: “To us, CPU utilisation is very important in our decision making process because of the way we are charged by the technology vendors. For example, Greenplum is charged per CPU core, so CPU utilisation is important.” (Appendix J:128).

The amount of memory used when processing data is different from one technology to the other, and this can have an impact on the performance of systems and also on the costs incurred when building and maintaining a DSS. Participants P1, P4 and P5 reveal that they prefer implementing technologies which process data while stored in RAM, as this improves the performance of the system. To these participants, system performance is more critical than the costs associated with adding RAM. Participant P4 for instance, states: ”RAM is very cheap now and therefore the cost of memory is no longer critical to us. We would therefore rather have a system that crunches data in memory while answering queries in real-time than a slow system with low memory.” (Appendix J:128).

Participants (P1, P2, P4, P5, P6, P7, P8 & P9) state disk I/O can be used to assess and select appropriate technologies for real-time DSSs. For example, participant P2 states: “Disk I/O is important to us because we generate a lot of data which we need to store quickly without affecting other users of the system.” (Appendix J:128). Furtehermore, participant P4 emphasises the importance of considering the technoilogy disk I/O by stating: “…a data warehouse separated from the analytics server may cause performance issues. Make sure the new technology architecture doesn’t impact on other systems performance.” (Appendix J:128).

IQ10. Which Industry is your company and what sort of insights do you expect from your data?

This interview question is used to ascertain if there is a relationship between business user requirements, and the technologies chosen for real-time DSSs in a BD environment. It is evident from the results that business user requirements influences the choice of technologies that are appropriate for real-time DSSs.

Participant P1 whose company works in the customer data management industry states: “Our core business is customer experience management. Our main focus is to build products for customer experience, and allow real-time monitoring of customer experience. From our data, we expect to identify risky customers, identify areas of weakness and strength in the business and then decide which intervention programmes are necessary. To meet this requirement, it needs real-time access to operational data, just as it arrives into our systems as things are happening. The key to our business is the ability to provide quick access to information for decision making.” (Appendix Y:144).

According to participant P2 who operates in the networking industry, there are two approaches to DSSs namely, to answer questions currently not being answered by existing technologies, and being able to predict something. Participant P9 also operates in the networking industry, and states that user requirements appear to influence the technologies appropriate for real-time decision support. In support of this view, P9 states: “One issue I have been investigating often is people want to know the reason certain subscribers are getting poor network quality. The problem is that you cannot tell as there are so many factors that influence this, for example, some people will upload multiple videos in multiple driver stations. They might be doing something else and at the same time watching a video and still tell you that they are having a bad network experience but you might not really know if they are watching a video. So there are all different kinds of opportunities to measure things you couldn’t measure before BD technologies came on board.” (Appendix Y:144).

Participant P10 provides a user story that appears to imply that business user requirements have an influence on the selection of technologies that are appropriate for real-time DSS in a BD environment. Therefore, participant P10 states: “The company interfaces with its clients through multiple channels such the company website, Facebook, twitter, email, SMS and through telephone calls. These channels generate massive amounts of data but only a small percentage is being leveraged. Ideally, we would want to have technology that can analyse all data that is available within the business as quickly as possible and make quick operational business decisions to our competitive advantage. We would want to respond to customer sentiments but at the moment we are unable to do so as we do not have technologies that can assist us in that.” (Appendix Y:144).

According to P4 and P5, user requirements, and the technologies available on the market, forms the basis of their selection process. Participant P4, states: “We considered requirements from the users and then checked what was available on the market to do the job required. We then compiled a matrix of all important things and weighted them. We then looked for vendors for the top three products and requested them to present their products. Each product had scores and in the end, the product that scored higher than the others was selected.” (Appendix Y:144).

Finding 23: Business user requirements drive the choice of technologies used for real-time DSSs.

Finding 24: Performance in terms of throughput and/or latency is the top criterion used to evaluate technologies appropriate for real-time DSSs.

Finding 25: Ability of a new technology to integrate with existing technologies and systems has an impact on the choice of technologies appropriate for real-time DSS in a BD environment.

Finding 26: The data collected reveales the following additional criteria that are important when determining technologies appropriate real-time DSSs in a BD environment: Costs, licensing models, resource usage, scalability, fault tolerance, security, technical support.

Finding 27: Technical skills availability can be considered as a criterion with which technologies can be compared but it should not be given high priority.

Finding 28: The usage of resources such as CPU, RAM and disk input/output by a technology has an impact on the system's performance.

Finding 29: The number of processors (CPU) required by technologies in order to give high system performance required by real-time DSSs can have an impact on licensing costs.

IQ11. How can an organisation evaluate analytic tools appropriate for real-time DSS in a BD environment?

The objective of this question is to ascertain from participants, what process they follow when selecting appropriate technologies for DSSs. The data collected for this question is depicted in Figure 4-5. It can be observed from Figure 4-5, that four of the participants (P2, P6, P7 & P8) indicate that they use a benchmark designed internally in their organisation to assess and select technologies used for DSS purposes. The internal benchmarks include proof of concept activities (POC) carried out by an organisation. Participant P2 in explaining the process followed to assess and select the database technologies currently being used in his organisation states: "To do the testing, we ran and compared three different database technologies. We took a sample of our dataset, uploaded it into the databases and then looked at the time it took to execute a query. Finally, we compared the amounts of time taken to execute the query by each of the products and based our decision on this. To us, performance was very critical. So we used an internal benchmark and not an existing benchmark." (Appendix Z:146).

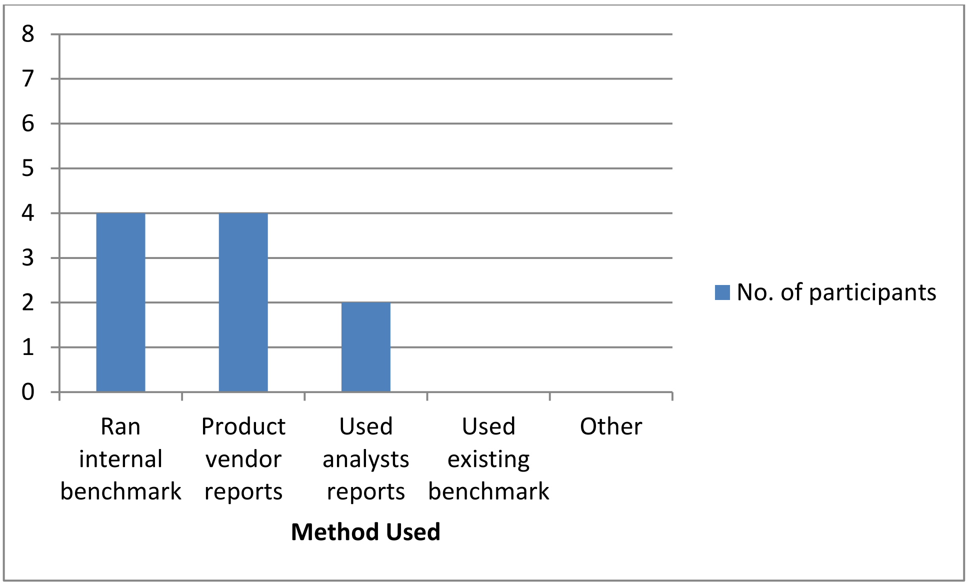


Figure 4-6 Method used to assess technologies by participants

Data collected from the participants show that four of the participants (P1, P3, P4 & P5) rely on product vendor reports in order to determine technologies that are appropriate for DSS. In this study, product vendor reports include presentations made by product vendors during the selection process. Participant P3 reveals that their selection process is influenced by on-line reports produced by analysts such as Gartner. Participant P3 states the opinion with regards to how the organisation selects Tableau and R as an analytics tool: "I was involved with the selection of Tableau and R which we use for analysing data. In the selection process, we used the recent Gartner Magic Quadrant report on Tableau and we also considered recommendations given by data analysts sharing their work on-line through platforms such as GitHub. We seriously consider what these people recommend and therefore much of our decision to go with Tableau (and with R) was based on their recommendations." (Appendix Z:146).

Finding 30: Organisations rely on product vendor reports.

Finding 31: On-line reports produced by analysts such as Gartner have an influence on the selection of technologies for DSS in BD environments.

4.3 **Quantitative findings**

The results from the literature review in Section 2.6 and qualitative interviews, indicate that system performance in terms of either, latency or throughput, is one of the top criterion considered when selecting appropriate technologies for real-time DSSs in a BD environment. This is because business competitiveness today is driven by quick access to data sources and the ability to make quick decisions based on all forms of available data and not just a subset of available data. The objective of the quantitative investigation is therefore, to ascertain if performance of different technologies, which are designed for DSS in a BD environment, is different or not. The results of this analysis can be used to determine if performance has an impact on the selection of technologies appropriate for real-time DSSs in a BD environment. In order achieve this, comparative computer laboratory experiments are conducted as discussed in Section 3.6.2. The results of the analysis of the data collected is now presented.

4.3.1 **Experiment results**

As stated in Section 3.6.2.2, the experiments are conducted by executing a query exhibited in Figure 3-3 using Hive, Spark and Impala against two sets of files with structures shown in appendix B and C. The two files are first loaded into HDFS and then data from the files uploaded into parquet tables, defined in Hive meta-store having the same structure as the files. One table is defined as a dimension table and has a static number of rows, while the other table is used as a fact table and '*manipulated'* by changing the number of rows during the experiments. As stated in Section 3.6.1, each of the tools (Impala, Hive and Spark) is tested with data sets grouped according to the number of transactions (rows) in the fact table. These are; 12 million rows, 52 million rows, 103 million, 255 million, 510 million and one billion two hundred million rows. Although the researcher intends to increase the number of transactions in the fact table to over one billion two hundred rows, there is no adequate disk space on the platform for this. The results of the experiments is now discussed.

4.3.2 **The research hypothesis**

The experiments are carried out in order to test whether the performance of technologies used to analyse data is equivalent or not. For the purpose of the reader, the hypothesis defined in Chapter 3 will be stated again in this Section:

H_0 – The mean query execution times for Impala, Spark and Hive are equivalent.

H_1 – The mean query execution times for Impala, Spark and Hive are not equivalent.

4.3.3 **Data collected**

This section provides a description of the data collected when query one (Q1) was executed using the 3 tools (Impala,Hive and Spark). Table 4.5 through Table 4.8 depicts the execution times taken in seconds (s) when Impala, Hive and Spark are used to execute Q1 against the data sets described in Section 4.3.1.

4.3.3.1 Impala

The time taken to successfully execute Q1 in Impala is indicated in Table 4-5 and in Figure 4-7. The results indicate that as the volume of data (number of rows in the fact table) increases up until 510,000,000 rows, the query execution time also increased. When the number of rows are pushed up to 1 200 000 000, Q1 aborts with error an message: '*MEM LIMIT reached*' and therefore there are no data entries for the last row in Table 4-5.

Table 4-5 Q1 on Impala

	Time taken in seconds									
Rows (GB)	**test1**	**test2**	**test3**	**test4**	**test5**	**test6**	**test7**	**test8**	**test9**	**test10**
12	0.90	0.89	0.87	0.89	0.89	0.89	0.88	0.91	0.93	0.95
52	1.00	1.00	1.04	1.02	1.06	1.03	0.98	1.01	1.05	1.06
103	1.13	1.10	1.10	1.09	1.09	1.11	1.10	1.10	1.10	1.17
255	1.52	1.51	1.61	1.49	1.48	1.51	1.49	1.51	1.53	1.52
510	2.00	2.03	2.00	2.06	2.01	2.10	1.99	1.99	2.00	2.00
1 200										

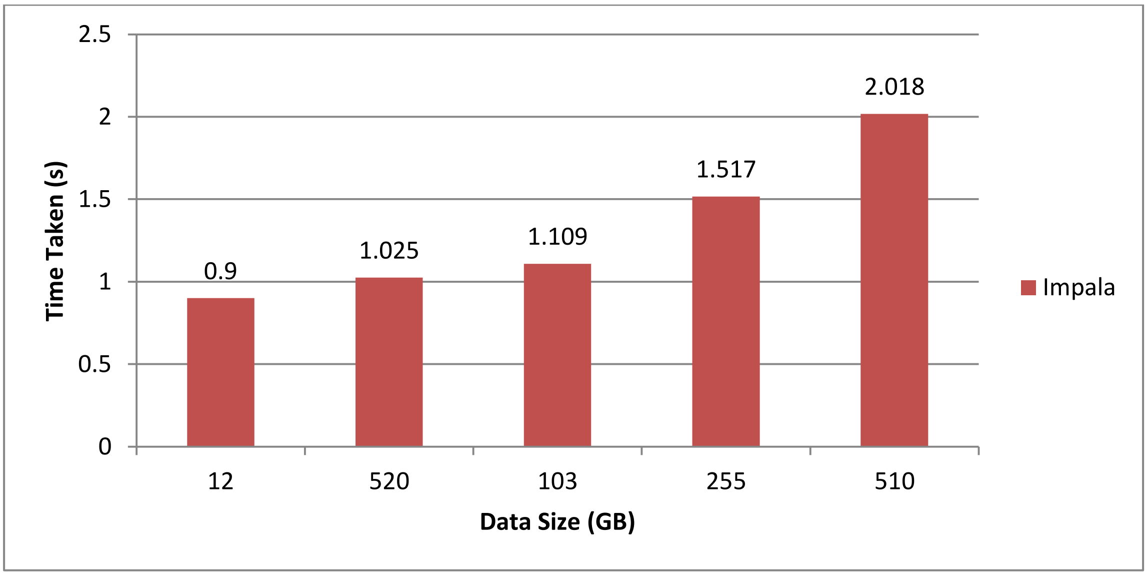


Figure 4-7 Graph: Q1 on Impala

4.3.3.2 Hive

Table 4-6 and Figure 4-8 depicts the amount of time taken to successfully execute Q1 with Hive. The data is in a pattern, that as the number of rows in the fact table increases, the query execution time also increases.

Table 4-6 Q1 on Hive

	Time taken in seconds									
Rows	**test1**	**test2**	**test3**	**test4**	**test5**	**test6**	**test7**	**test8**	**test9**	**test10**
12	33.96	33.99	34.01	34.06	34.05	34.02	34.02	34.00	34.04	33.97
52	44.77	44.67	44.69	44.66	44.70	44.66	44.67	44.69	44.66	44.69
103	59.98	60.00	59.97	60.05	59.94	60.00	59.97	60.06	60.03	60.03
255	101.97	101.99	101.98	102.00	101.98	102.02	101.91	101.95	101.99	102.07
510	181.95	181.98	181.87	181.90	181.98	181.88	181.93	181.95	181.99	182.04
1 200	789.82	789.86	789.83	789.75	789.84	789.89	789.91	789.97	789.83	789.84

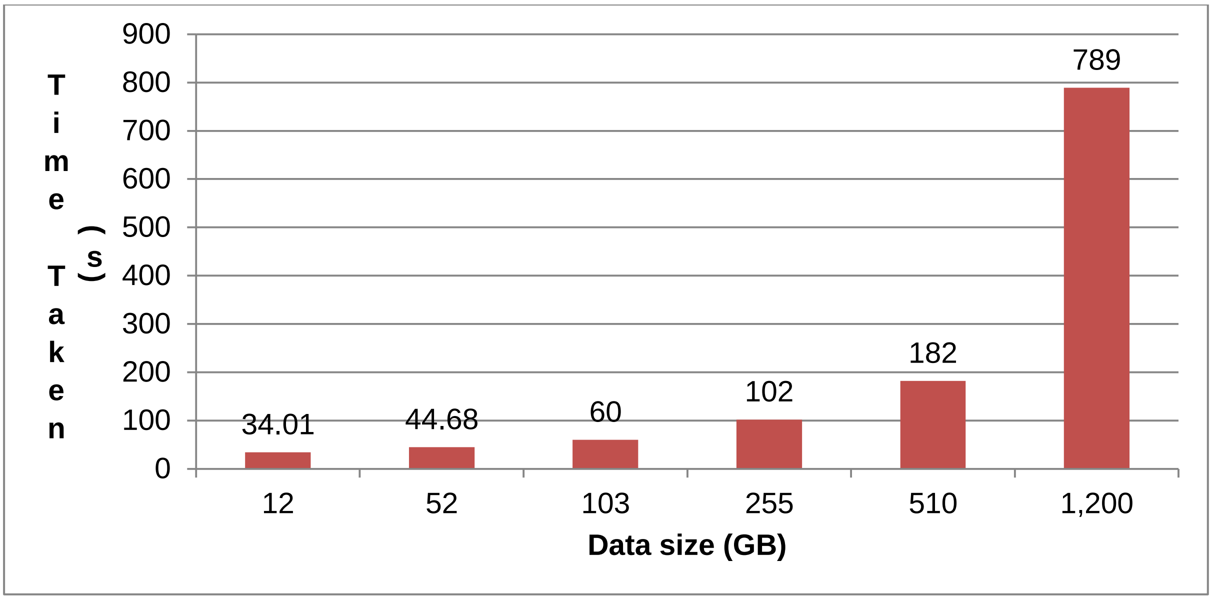


Figure 4-8 Graph: Q1 on Hive

4.3.3.3 Spark

The amount of time taken to execute query Q1 using Spark is indicated in Table 4-7**.**

Table 4-7 Q1 on Spark

	Time taken in seconds									
Size(GB)	**test1**	**test2**	**test3**	**test4**	**test5**	**test6**	**test7**	**test8**	**test9**	**test10**
12	0.54	0.49	0.52	0.56	0.48	0.47	0.51	0.42	0.50	0.53
520	0.53	0.54	0.53	0.52	0.50	0.49	0.48	0.45	0.59	0.60
103	0.62	0.60	0.61	0.59	0.66	0.63	0.58	0.54	0.55	0.57
255	0.80	0.69	0.79	0.77	0.78	0.76	0.79	0.75	0.80	0.74
510	1.00	1.01	1.12	1.03	1.01	1.01	1.04	1.01	1.02	1.03
1 200	2.03	2.08	2.07	2.08	2.03	2.01	2.00	1.99	1.98	1.95

Figure 4-9 depicts that the average execution time for each category of data size for Q1 increases as the number of rows in the fact table increases, and the same pattern detected for Impala, is also applicable for Spark. However, unlike the Impala query which abortes at the 1200M dataset level, Spark executes successfully with the mean execution time of 2.02 seconds

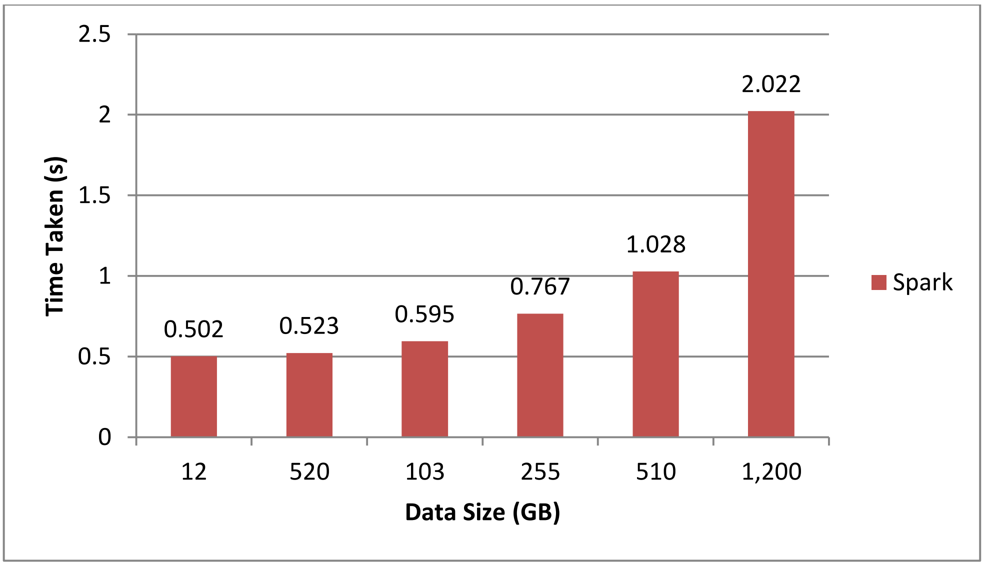


Figure 4-9 Graph: Q1 on Spark

The data collected also indicates that when Q1 is executed, Impala and Spark have lower execution times when compared to Hive. This is depicted in Figure 4-10, however, the pattern will be tested statistically in Section 4.3.4.

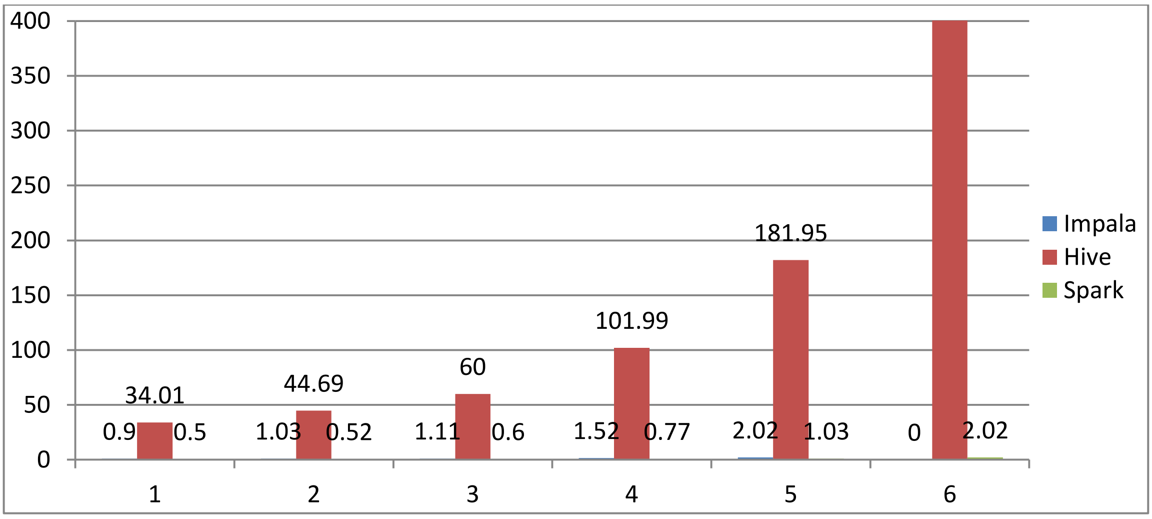


Figure 4-10 Execution time taken per tool

4.3.4 **Statistical analysis and interpretation of collected data**

The data collected from the experiments and described in Section 4.3.3 is analysed using Generalized Linear Models in SPSS. The dependent variable does not portray normally distributed data and therefore, non-parametric tests are used to analyse the data. Table 4-8 depicts that there are 170 records (observations) included in the analysis. This represents the total number of repeated executions conducted for the three tools and the different datasets used in the experiments. As stated in, Q1, it is executed ten times per tool and for each of the 6 datasets. This implies that each of the tools expected to yield 60 records. However, due to Impala aborting at 1 200M datasets, 10 records are excluded from the statistical analysis.

Table 4-8 Case Processing Summary

	N	Percent
Included	170	94.4%
Excluded	10	5.6%
Total	180	100.0%

The tool ratio (%) used of the records analysed, is depicted in Figure 4-11.

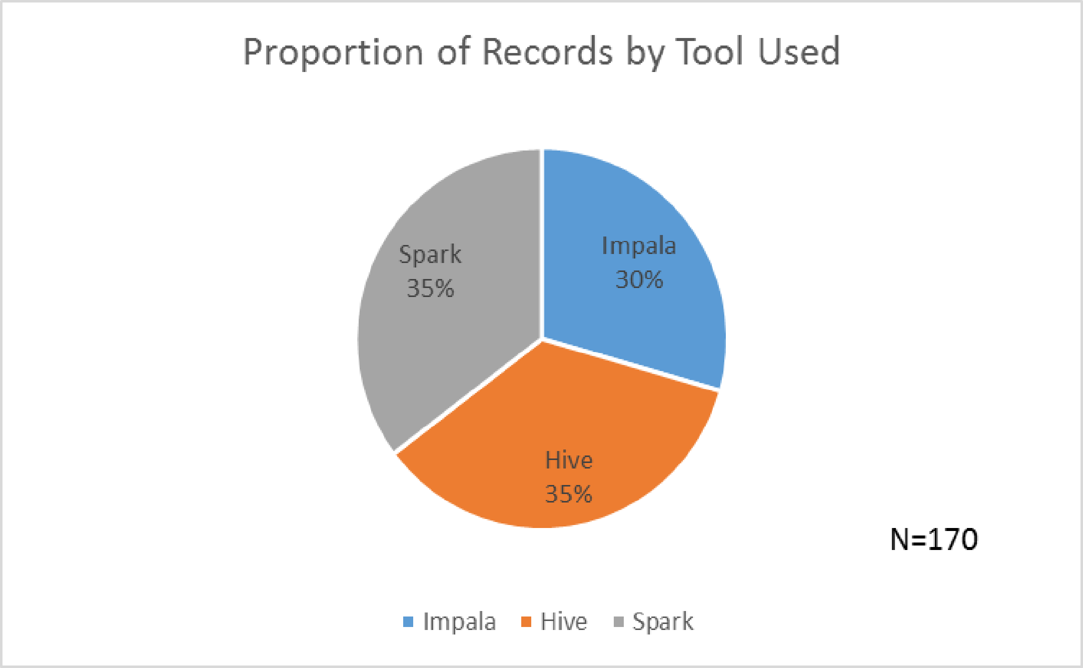


Figure 4-11: Ratio of records by tool used

Table 4-9 depicts that there are 30 records per data set, except for 1200M, which has 20 records.

Table 4-9 Categorical Variable Information

			N	Percent
Factor	**Tool Used**	Impala	50	29.4%
		Hive	60	35.3%
		Spark	60	35.3%
		Total	170	100.0%
	Number of rows	12M	30	17.6%
		52M	30	17.6%
		103M	30	17.6%
		255M	30	17.6%
		510M	30	17.6%
		1200M	20	11.8%
		Total	170	100.0%

Table 4-10 depicts the continuous dependent variable information. The results indicate that time taken in seconds, varied from 0.42 seconds to 789.97 seconds. The mean for the distribution is 72.03, with a standard deviation of 186.202.

Table 4-10 Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Time in seconds	170	.42	789.97	72.0291	186.20250

4.3.4.1 Goodness of Fit

As depicted in Table 4-11, the Goodness of fit test shows the query execution time between Hive, Spark and Impala. The Deviance value (0.232, see Table 4-11) has a chi-square distribution and tests the goodness-of-fit of the model. The resulting p-value from the Chi-square distribution is 1.0, showing that this model is a good fit[1]. This suggests that the null hypothesis can be rejected with the conclusion that the mean query execution time for Hive, Spark and Impala are not equivalent.

Table 4-11 Goodness of Fit

Statistic	Value	Df	Value/df
Deviance	.232	153	.002
Scaled Deviance	170.000	153	
Pearson Chi-Square	.232	153	.002
Scaled Pearson Chi-Square	170.000	153	
Log Likelihood[b]	319.524		
Akaike's Information Criterion (AIC)	-603.049		
Finite Sample Corrected AIC (AICC)	-598.519		
Bayesian Information Criterion (BIC)	-546.605		
Consistent AIC (CAIC)	-528.605		
Dependent Variable: Time in seconds **Model: (Intercept), Tools Used, RowsCode, ToolsUsed * RowsCode**			

4.3.4.2 Tests of Model Effects

The tests of model effects are contained in Table 4-12 and depict that execution time for Impala, Spark and Hive is statistically significantly different (p<0.001). This implies that given datasets used for the experiments, the time taken to execute Q1 using Spark, Impala and Hive is different. The results also revealthat there is a statistically significant difference (p<0.001) in the amount of time taken by the same tool when executed on different datasets. Lastly, the results further revealthat there is a significant interaction between the tools used, and the number of rows used (p<0.001).

Table 4-12 Tests of Model Effects

Source	**Type III**		
	Wald Chi-Square	Df	Sig.
(Intercept)	816779906.784	1	.000
ToolsUsed	931478679.545	2	.000
RowsCode	1561249878.289	5	.000
ToolsUsed * RowsCode	1583636813.538	9	.000
Dependent Variable: Time in seconds **Model: (Intercept), ToolsUsed, RowsCode, ToolsUsed * RowsCode**			

4.3.4.3 Estimated Marginal Means 1: ToolsUsed

Table 4-13 depicts the estimated marginal means for the three tools being different, confirming the results shown by the goodness of fit and test of model effects results. The mean execution times for the Impala, Hive and Spark are 1.31s, 202.08s and 0.91s, respectively. Although the mean execution time taken by Spark to execute Q1 is close to that taken by Impala, Spark has lower execution time which is good for real-time DSS.

Table 4-13 Estimated Marginal Means 1: ToolsUsed

ToolsUsed	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
Impala	1.3138	.00522	1.3036	1.3240
Hive	202.0813	.00477	202.0720	202.0907
Spark	.9062	.00477	.8968	.9155

4.3.4.4 Pairwise comparisons of tools used

The results discussed above merely indicate that the execution time for Impala, Hive and Spark is different, but does not show which individual tools are different from each other. Pairwise comparison statistics indicate the individual tools that are different from each other. Appendix D contains the pairwise comparisons of estimated marginal means based on the original scale of dependent variable times, in seconds. Appendix D depicts the results being statistically significant (p<0.001) indicating that Impala has lower execution times than Hive, with a mean difference of -200.76 seconds. Impala has a higher execution time than Spark with a mean difference of 0.41 seconds. The execution time of the two tools is statistically significantly different (p<0.001). Furthermore, the results indictae that the execution time for Spark is statistically significantly different from that of Impala (p<0.001). Spark produced lower execution times than Impala with a mean difference of -0.41. On the other hand, the execution time for Spark is statistically lower than that of Hive, with a mean difference of 201.02. Overall, the results reveal that the mean difference is significant at 0.05.

4.3.4.5 Overall Test Results: tools used

The Wald chi-square tests statistic indicated in Table 4-14, depicts the effect of tools used on the execution time when executing Q1. This test is based on the linearly independent pairwise comparisons among the estimated marginal means. As seen in Table 4-14, the data collected show that the three tools tested in this study are statistically different in terms of time taken when executing the same amount of data using the same query Q1.

Table 4-14 Overall Test Results: tools used

Wald Chi-Square	Df	Sig.
1149424545.082	2	.000

4.3.4.6 Estimated Marginal Means 2: datasets

The graph in Figure 4-12 and the data in Table 4-15 contain data of the relationship between the data size (number of rows) and the mean time taken to execute Q1. The results indicate that the mean time in seconds increases, as the number of rows increases. Of note, the standard error for each of these means, is the same except for the 1200M dataset, which is attributed to no data for Impala at this dataset.

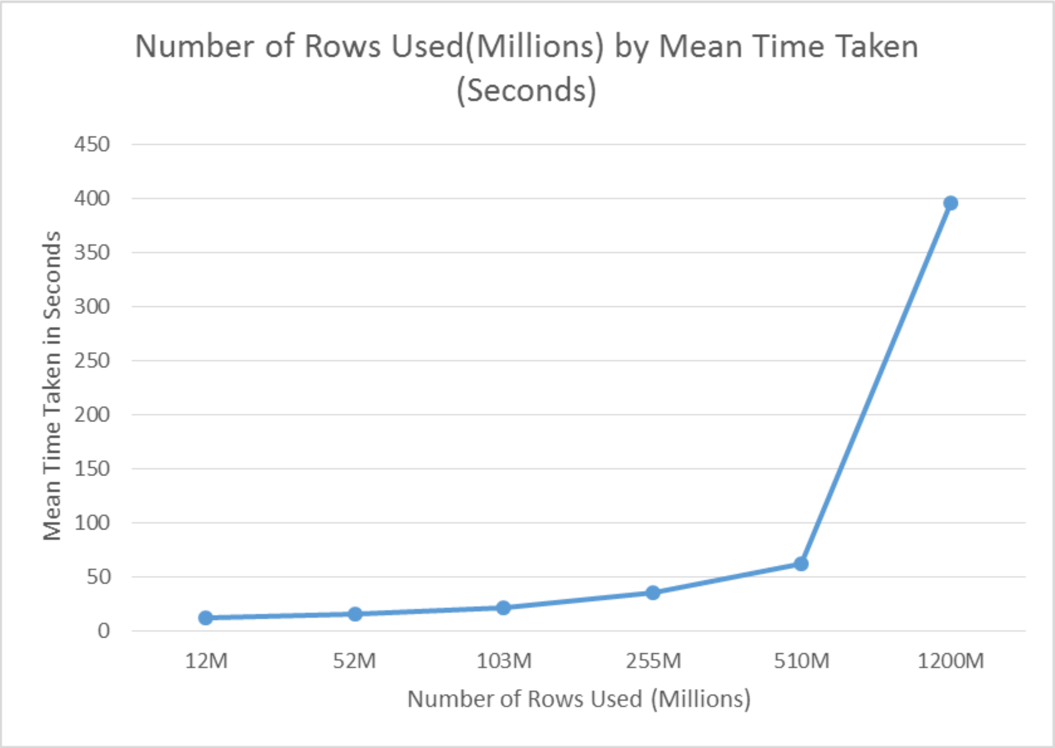


Figure 4-12 Number of rows and mean time taken

Table 4-15 Estimated marginal means: datasets

RowsCode	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
12M	11.8047	.00674	11.7914	11.8179
52M	15.4113	.00674	15.3981	15.4246
103M	20.5690	.00674	20.5558	20.5822
255M	34.7567	.00674	34.7434	34.7699
510M	61.6643	.00674	61.6511	61.6776
1200M	395.9380	.00826	395.9218	395.9542

4.3.4.7 Pairwise Comparisons: datasets

Appendix E contains the pairwise comparisons of estimated marginal means based on the original scale of dependent variable time in seconds. This is when Q1 is executed against different datasets. Further in Appendix E, depicting the data, for the overall mean time taken in seconds to execute Q1 by all the tools against the different datasets. This transpires that the mean time taken at each dataset, for example 12M, is compared to the mean time taken by the rest of the datasets (52M, 103M, 255M, 510M and 1200M). All the pairwise comparisons depict that there is a statistically significant difference (p<0.001) between the datasets. This suggests that there is an association between the number of rows used, and the time taken in seconds. In particular, as the number of rows increases, the time taken in seconds also increases. The mean difference is significant at the .05 level.

4.3.4.8 Overall Test Results: datasets

The Wald chi-square test is used to reflect on the effect of datasets size (row count) on the dependent variable time in seconds. This test is based on the linearly independent pairwise comparisons among the estimated marginal means. The overall test result is depicted in Table 4-16 that reveals that there is a statistically significant difference among the tools used (p<0.001), when executing Q1 against different datasets.

Table 4-16 Overall Test Results: datasets

Wald Chi-Square	df	Sig.
1779195743.825	5	.000

4.3.4.9 Estimated Marginal Means 3: ToolsUsed* RowsCode

The data in Table 4-17 can be gleaned to be the mean time taken at the intercept of each tool used, and the number of rows (dataset) used. The results indicate that for all the tools used, the amount of time taken increases as the number of rows increases. This result confirms the trend observed in the earlier analysis, conducted above in Section 4.3.4.6.

Table 4-17 Estimated Marginal Means 3: ToolsUsed* RowsCode

ToolsUsed	RowsCode	Mean	Std. Error	95% Wald Confidence Interval	
				Lower	Upper
Impala	12M	.9000	.01168	.8771	.9229
	52M	1.0250	.01168	1.0021	1.0479
	103M	1.1090	.01168	1.0861	1.1319
	255M	1.5170	.01168	1.4941	1.5399
	510M	2.0180	.01168	1.9951	2.0409
	1200M				
Hive	12M	34.0120	.01168	33.9891	34.0349
	52M	44.6860	.01168	44.6631	44.7089
	103M	60.0030	.01168	59.9801	60.0259
	255M	101.9860	.01168	101.9631	102.0089
	510M	181.9470	.01168	181.9241	181.9699
	1200M	789.8540	.01168	789.8311	789.8769
Spark	12M	.5020	.01168	.4791	.5249

	52M	.5230	.01168	.5001	.5459
	103M	.5950	.01168	.5721	.6179
	255M	.7670	.01168	.7441	.7899
	510M	1.0280	.01168	1.0051	1.0509
	1200M	2.0220	.01168	1.9991	2.0449

4.3.4.10 Pairwise comparisons at tool used and dataset intercept

The data in Appendix F depicts the pairwise comparison of the estimated marginal means, based on the original scale of the dependent variable time in seconds. The data reveals that at each dataset, the time taken using Impala, Hive and Spark to execute Q1, is statistically significantly different (p<0.001). The same pattern can be detected in Section 4.3.4.10 and is also revealed in this section, where regardless of the volume of data being analysed by the tools, the execution time is different for each of the tools.

4.3.5 **Findings from experiments**

The data collected and analysed, indicate that there is a relationship between the volume of data and the time taken to execute a query. As the volume of data increases, the time taken directly increases for all the tools. The data analyses also indicates that at all datasets prepared for the experiments, the three tools behaved differently with regards to the speed (time taken to return the results). Considering the statistical analysis in Section 4.3.4, Spark produced results in less amount of time compared to Impala and Hive, while Hive took the longest to produce results. Therefore, the conclusion is derived that the mean query execution times for Impala, Spark and Hive are not equivalent, thereby rejecting H_0. This result implies that execution time can be used to compare and evaluate technologies that are appropriate for real-time DSS in a BD environment. Furthermore, from the experiment data, it is evident that the 3 tools portrayed different patterns in terms of memory consumption, CPU utilisation and disk I/O. This is manifested by Impala appearing to use more memory resources when processing data. This is clearly revealed when Impala aborts processing at a data size of 1200M. The results also indictae that Spark uses more memory when processing data than Hive. Hive therefore, uses the least of the three tools of memory when processing data. Another significant finding is that Impala and Spark portraye low usage of disk I/O, compared to Hive at all datasets used.

4.3.6 **Summary**

In this research, data is collected using interview techniques and computer laboratory experiments. Interview data is analysed using the qualitative technique of content analysis. A total of 31 findings are produced from the qualitative analysis of interview data. The perception based qualitative findings indicate that system performance in terms of speed of producing reports, is the most important attribute of a real-time DSS. Based on this finding, a set of hypothesis is formulated in order to test this finding in a computer laboratory experiment setting. The objective is to ascertain if the performance of different technologies is equivalent or not, when processing the same amount of data under the same conditions. The experiment results proved statistically, that the query execution time for the three tools is different. Based on these results, it implies that performance of different technologies, when processing data, can be used to assess technologies that are appropriate for real-time DSS in a BD environment. In Chapter 5, the details of the findings from this research are discussed and also related to existing literature.

5 CHAPTER FIVE: DISCUSSION

5.1 Introduction

This chapter provides a discourse on the research findings presented in Chapter 4. The chapter consists of five sections; introduction, themes developed, answering research questions, BD technologies evaluation criteria and summary.

Nine main themes emerged from the findings in this research. The themes are: i) multiple data sources, ii) Volume of data, iii) high velocity data, iv) Structured and unstructured data, v) business user requirements, vi) integration with traditional database and DW technologies, vii) scalability, viii) system performance and viiii) technical skills availability. This discussion brings to light the concept of real-time DSS in a BD environment as an enterprise resource that can give organisations a competitive advantage. The discussion also addresses BD technologies and frameworks that have recently been developed by industry and academia to solve challenges of BD currently faced by organisations. Finally, the chapter draws from the findings in Chapter 4, to propose BD technologies evaluation criteria and recommendations that can be used to decide appropriate technologies for real-time DSS in a BD setting.

5.2 Themes developed

As shown in Appendix H, nine main themes materialised from the research findings, and will now be discussed in detail in Sections 5.2.1 through 5.2.9.

5.2.1 Theme 1: Multiple data sources

The pattern observed in the interview data collected indicates that organisations with multiple sources of data, and are using traditional BI tools are not using all data sources available, for decision making. The data trends also show that organisations using BD technologies such as Hadoop, NOSQL databases and Spark have multiple data sources. Furthermore, the perceptions of participants suggest that organisations that are planning to implement BD technologies are motivated by; the requirement to integrate and leverage data generated by new data sources such as Facebook and Twitter. This implies as data sources increase, organisations want to adopt technologies which can acquire, store and analyse data from these sources.

The prevalent use of the internet, the increased adoption of e-commerce solutions, and the increased use of mobile and on-line applications, results in multiple systems generating data available for decision making (Chen *et al*., 2014; Tank *et al*., 2010;). This is consistent with the pattern observed in the interview collected data research where all the participants acknowledge the presence of multiple sources of data in their organisations. This is depicted in Table 4-2 and Table 4-3, and in addition, to Appendixes L and M. According to Dong and Srivasta (2013:2), "…the number of data sources for BD is much higher than for traditional data sources, the data sources are extremely heterogeneous, and many of the data sources are very dynamic." This implies that the number of data sources in an organisation has an impact on the selection of technologies appropriate for real-time DSS in a BD environment.

5.2.2 **Theme 2: Volume of data**

Volume is one of the key characteristics of BD that emerged in Chapter 2 covering the literature. This research reveals there is a relationship between the volume of data generated and analysed by an organisation, and the selection of technologies appropriate for real-time DSS in a BD environment. To reach this conclusion, the interview participants are asked to describe the volume of data available in their organisations and the technologies currently used to analyse such data. Participants are further asked to state ideal technologies in cases where existing technologies are not ideal. As depicted in Tables 4-3 and 4-4, and Appendixes P and Q, although all participants in this research still use relational database and data warehouse technologies, some have started using BD technologies such as Hadoop and NOSQL databases. This is in an attempt to solve the volume challenge. For example, P6 states that his organisation has started implementing Hadoop and Spark in order to overcome the challenge of high volume data. As depicted in Table 4-4, organisations where data sizes are in the Terabyte region and above, use BD technologies such as Hadoop, or stated that they would ideally use BD technologies. Noted by Bakshi (2012), growth in data brings the challenge of storage, analysis and extraction of insights from it. In the business world today, it is not uncommon for even small to medium enterprises, to generate very high volumes of data. Liu *et al.* (2013) state that overcoming the volume challenge requires both, technologies that store vast amounts of data in a scalable fashion, and technologies that use distributed approaches to querying and deriving actionable information and insights from the BD. This is also supported by Hu *et al.* (2014) who assert that a BD analytics system must be able to support very large data sets created now and in the future.

5.2.3 **Theme 3: High Velocity data**

As revealed in Chapter 2.4.2, velocity is one of the key attributes of BD. According to Liu (2013), velocity refers to the frequency, or rate with which data is generated, received or shared. The data collected using interviews after analysis, suggests that participants organisations that have high velocity data are either, exploring to migrate their DSS onto BD technology platforms such as Spark and/or Hadoop, or they have already started implementing these technologies. The findings further indicate that that participant organisations that have low velocity data, are satisfied using traditional BI tools for real-time DSS. This indicates that velocity has a significant impact on the selection of technologies appropriate for real-time DSS. This observation is consistent with findings by Liu (2013) who states: "…the velocity of large data streams from a vast range of devices and click streams not only creates requirements for greater real-time use-cases, but also power the ability to parse text, detect sentiment, and identify new patterns. Real-time analytics require fast matching and immediate feedback loops based on alignment with geo location data, social media, user history and current sentiment."

5.2.4 **Theme 4: Structured and Unstructured data**

The results from the narrative interviews indicate that all the participants' organisations have some form of structured and unstructured data generated by various systems. Structured data is stored in relational databases while unstructured data is stored in file systems and in NOSQL databases. The interview results indicate that a small portion of unstructured data stored in NOSQL databases are analysed, while unstructured data not stored in NOSQL databases are not analysed for decision making. The results also indicate that organisations using traditional database and data warehouse technologies, only analyse structured data. The conclusion can be made that the structure of data generated by an organisation has an impact on the choice of technologies appropriate for real-time DSS in BD environment.

5.2.5 **Theme 5: Business user requirements**

This research collected participant perceptions to determine if business user requirements have an impact on the selection of technologies appropriate for real-time DSS in a BD environment. To do this, the participants are first requested to list the technologies currently being used in their organisations for real-time DSS. Participants are then further asked to state the business user requirements that drive them to choose the technologies currently being used, or that would be ideal for real-time DSS in their organisations. Although the participants have different business user requirements from each other, the general perception is that business user requirements significantly influence the choice of technologies appropriate for real-time DSS in a BD environment. This is consistent with observations made by Singh & Singh (2012) and also Kimball (2011). According to Kimball (2011), business user requirements drive the selection technologies appropriate for BD analytics.

5.2.6 **Theme 6: Integration with traditional database systems**

The findings from the interviews indicate that the ability to integrate new technologies with existing systems has an impact on the choice of technologies appropriate for real-time DSS in a BD environment. It appears the reason for this is that many organisations invested substantially in resources using traditional database and data warehouse systems.

5.2.7 **Theme 7: Scalability**

This section explores the concept of scalability and how this influences the selection of technologies appropriate for real-time DSS in a BD environment. According to Dilpreet and Reddy (2014), scaling is the ability of systems to adapt to increased demands in terms of data processing as the volume of data increases. The concept of scalability is also illustrated by Marz *et al*. (2012) when describing the challenges of BD. This illustration is synonymous with user stories collected during the interviews conducted in this research. For example, P6 explains how his organisation resorts to sharding the MySQL database and data warehouse platforms, in an effort to adapt to the deluge of data landing in their DSS environment. There are two types of scaling namely, horizontal scaling and vertical scaling. According to Dilpreet and Reddy (2014), horizontal scaling is also known as "scale out" which involves distributing the workload across many servers. In this setup, multiple independent computers (running individual instances of the operating system) are integrated to boost the processing. On the other hand, vertical scaling also known as “scale up”, involves a single server implementation with more processors, more memory and faster hardware running a single instance of the operating system. Researchers have developed numerous technology platforms which can either scale up or scale out, as depicted in Chapter 2.5. Based on the findings from the interviews and literature review, scalability has an influence on the selection of technologies that are appropriate for real-time DSS in BD environment.

5.2.8 **Theme 8: System performance**

The perception-based results derive from this research indicates that system performance in terms of speed or latency and throughput has an influence on the choice of technologies appropriate for real-time DSS in a BD environment. According to Dilpreet and Reddy (2014:15), performance can be measured in terms of speed or throughput where “…speed refers to the ability of the platform to process data in real-time whereas throughput refers to the amount of data that the system is capable of handling and processing simultaneously.”. All the participants, including those drawn from organisations still relying on traditional database technologies are of the opinion that performance is important when selecting technologies appropriate for real-time DSS in a BD environment. This observation appears to be in line with observations reported by Marz *et al.* (2012) who state that the drive for businesses to make quick decisions on large volumes of data drawn from various sources drives the requirement for high performing systems with low latency. Dilpreet and Reddy (2014) also state that business users need to be clear about whether the goal of the technology is to optimize the system for speed or throughput.

5.2.9 **Theme 9: Technical skills availability**

The results from this research indicate that none of the participants selected technologies for real-time DSS, based on the availability of technical skills. This implies that technical skills availability has no impact on the choice of technologies appropriate for real-time DSS in a BD environment. Existing literature on BD technologies appears to be silent about technical skills. Dilpreet and Reddy (2014) actually mention that BD skills are very scarce. The perception based results from the participants indicate that BD technologies are still new in industry and if organisations have to make use of these technologies, they have to up skill their employees.

5.3 **Evaluation criteria framework**

The research investigates the processes followed to evaluate and select technologies used for DSS in their organisations. The participants are also asked to state the features or properties of technologies that they assessed in order to compare and evaluate for appropriateness in real-time DSS in a BD environment. The perception-based results indicate that there are no standard end-to-end benchmarks, guidelines or frameworks available for comparing and evaluating technologies appropriate for real-time DSS in a BD environment. This finding is consistent with observations made by Dilpreet and Reddy (2014) and Liu *et al.* (2013). According to Ghazal *et al*. (2013), there are no end-to-end standard benchmarks available for BD technologies, which makes it very difficult for organisations to evaluate, compare and select appropriate technologies. This resonates well with findings from this research as some participants profess ignorance of any benchmarks that can be used to compare technologies. Dilpreet and Reddy (2014) also support this perception by noting that existing benchmarks that they found in their survey were designed for specific products and are difficult to use. Furthermore, the findings from the interview data, experiments and literature review indicate that there are 9 evaluation criteria that organisations can use to assess and select technologies appropriate for real-time DSS in a BD environment. The identified criteria are: i) performance, ii) ability to scale, iii) ability to process fast changing data, iv) ability to process structured and unstructured data, v) ability to integrate data from multiple sources, vi) ability to seamlessly adapt to changes in data structure, vii) fault tolerance capability, viii) ability to integrate with existing technologies and data analytics platforms and viiii) costs. These are now discussed based on the perceptions of the participants and existing literature on BD section 5.3.1 to section 5.3.9.

5.3.1 **Performance**

All the participants use system performance as a metric to compare and evaluate technologies appropriate for real-time DSS as depicted in Appendix X. Performance can be measured using throughput or latency. The results indicate that some participants use latency, while others considered throughput and yet others, use both, throughput and latency to compare and evaluate technologies appropriate for DSS. The results of the comparative experiments confirm the findings from the interview about performance. As seen in Section 4.2, the statistical analysis results indicate the three technologies tested have significantly different query execution time when analysing the same amount of data. This implies that technologies appropriate for real-time DSS can be compared and evaluated based on performance. Kimball (2011) as well as Reddy and Dilpreet (2014) appear to support this finding by including performance in their lists of features to consider, when choosing BD technologies in general.

5.3.2 **Ability to scale**

Organisations are experiencing some form of continuous growth in the volume of data generated by its disparate systems. According to P6, in 2014 the organisation’s average number of transactions per month was one billion two hundred, but by June 2015, the number of transactions doubled per month. The impact is that the organisation finds it difficult to maintain the required system performance by scaling the system by sharding the database and data warehouse platforms. The main reason is that this process is costly to the organisation. To overcome this challenge, the organisation has started moving their analytics system to a Hadoop platform, which has so far, proved to be highly scalable and fault tolerant. According to Hu *et al.* (2014:13), BD analytics systems must be able to support very large data sets created now and in future and all the system’s components must be capable of scaling to address the ever growing size of data. The perceptions from the interviewees and the existing literature suggest that scalability is an important consideration when evaluating appropriate technologies for real-time DSS in a BD environment. As observed in a survey by Hu *et al.* (2014), different technologies apply different scaling strategies and some are more expensive than others.

5.3.3 **Ability to process fast changing data**

The results of the interviews conducted in this research show that all participant organisations have a certain percentage of data as high velocity attributes. Either data is captured into the source systems per second, or updated per second. The results also indicate that none of the participants actually analyse this data in real-time. Only participant P7, mentions that his organisation analyses data in near real-time mode. P2 describes his organisation's data as high velocity, which means it arrives into the database in real-time but analytics and reporting on this data is done per day because they use traditional BI tools.

5.3.4 **Ability to process structured and unstructured data**

Generally, organisations are generating both, structured and unstructured data types from various systems. Examples of systems and applications generating unstructured data include social media platforms, digital sensors, emails, campaign management systems, websites and document management systems. The results in this research show that more than 90% of data currently being analysed for decision making by organisations is structured in nature while a small percentage is unstructured. This is despite the fact that organisations are generating very high volumes of unstructured data in the form of documents, emails, videos, website logs and other text data from social media platforms. The observations made in this research show that relational databases are used to analyse structured data only while organisations that are actually analysing unstructured data, use new technologies such as Hadoop and NOSQL databases. The results from the interview data also suggest that more and more systems are generating high volumes of unstructured data and it is the perceptions of the participants that organisations would gain more value by integrating both structured and unstructured data for decision making. The ability of a technology to process both structured and unstructured data types should form the basis of comparing and evaluating technologies that appropriate for real-time DSS in a BD environment.

5.3.5 **Ability to integrate data from multiple sources**

The results from the research indicate that 100% of the organisations in this research generate data from multiple sources. All the organisations have a wide range of OLTP systems, e-commerce sites and of late accounts in the social media space. The discouraging pattern seen from the interview data is that none of the participant organisations analyse all the data from the multiple sources in real-time despite the benefits of real-time DSS as seen in Chapter 2. Furthermore, the results concur that organisations that are analysing data from multiple sources in near real-time, use BD technologies such as Hadoop, NOSQL databases and Spark. On the other hand, organisations who rely on traditional BI tools for decision support systems, are not analysing data from multiple sources in real-time because they depend on batch ETL processes to extract, transform and load data into the data warehouse. Based on these findings, it suggests that the ability of a technology to integrate data from distributed sources can be used to compare and evaluate technologies that are appropriate for real-time decision support systems in a BD environment.

5.3.6 **Ability to seamlessly adapt to changes in data structure**

In a traditional BI setting, the goal is to make decisions based on attributes of data that are known in advance. This concept entails that the data structure is known in advance and database structures and reports are built before data is collected for decision making. The results from the interview data indicate that organisations who use RDBMS technologies find it difficult and time consuming to adapt reporting and analytics systems each time a change in data structure is made at source systems. As discussed in Section 2.2, this approach is not appropriate from decision support systems in a BD environment, where the data structures change very quickly and data structures are complex. The interview data collected clearly indicates that the organisations that have started discovering insights and patterns from data that changes in structure, quickly come to rely on Hadoop based solutions and Spark to analyse data. Although these organisations are not yet analysing the data in real-time, it appears Hadoop based solutions and Spark offers the ability to adapt to data structure changes with ease because the data is stored in its raw format on file systems. According to Marz *et al.* (2012), Hadoop based solutions allow businesses to run arbitrary functions on arbitrary data sets.

5.3.7 **Fault tolerance**

Although the interview questions used in this research do not have a question designed specifically to investigate fault tolerance, this concept features frequently in existing literature on BD and is also mentioned repeatedly, by the participants being an important consideration when comparing and evaluating technologies appropriate for real-time DSS in a BD environment. This concept is well illustrated by Marz *et al.* (2012) and refers to the ability of a system to continue processing data when one component of the system fails. The results show that the participants who use RDBMS for DSS never mentioned fault tolerance which seems to imply that traditional database technologies do not have fault tolerance capabilities. On the other hand, fault tolerance is cited by the participants who are already using Hadoop based solutions and Spark, as one of the strengths of these technologies. This is also supported according to the founders of Hadoop. Spark founders also boast about fault tolerance in Spark in literature. According to Marz *et al. (2012),* fault tolerance is achieved by using multiple nodes which are managed centrally. Each data block is sent to multiple nodes. Based on the perceptions of participants and also the findings from existing literature, this research concludes that fault tolerance can be used as criterion applicable when comparing and evaluating technologies appropriate for real-time DSS in a BD environment.

5.3.8 **Ability to integrate with existing technologies and data analytics platforms**

The results of the interview data indicate that all the participants indicate that in an organisation, it is important for new technologies to be able to integrate with existing systems. According to participants P4 and P5, many organisations are still cowed in traditional database and data warehouse technologies and huge financial investments have gone into building these systems. This seems to imply that organisations will continue for more years running data warehouse solutions. Furthermore, BD is still a new concept and as seen in Section 4.2. A small percentage of organisations interviewed indicated that they have started implementing BD technologies. Based on the interview data, it suggests that a technology's ability to integrate with existing data technologies and systems, is key when selecting appropriate technologies for real-time DSS in a BD environment.

5.3.9 **Costs**

According to the findings from the interview data collected in this research, 100% of the participants mention that cost is an important aspect when comparing and evaluating technologies appropriate for real-time DSS in a BD environment. According to participant P2, costs are considered as total cost of ownership (TCO) or total cost of the infrastructure over a time period. Generally, TCO is comprised of hardware and software costs, maintenance, training and licencing costs. Furthermore, the data collected reveals that different technologies have different licensing models. Some products are licenced per user, some per server node and some per CPU core. The licencing model has an impact on TCO.

5.4 **Answering research questions**

As stated in Chapter 3, this research is driven by three research questions. This section provides a summary of the answers to the sub-research questions used to guide this research.

Research question 1: What factors influence the selection of tools for real-time DSSs in a BD environment?

The objective of this question is to explore the concept of BD, its related technologies and identify factors that influence the selection of technologies appropriate for real-time DSS in a BD environment. To achieve this, two sub-research questions are asked and participants' summarised answers are depicted in Table 5-1.

Table 5-1 Research question 1: summary of answers

No.	Sub-research question	Summary of participants' responses
1.1	What is the relationship between the characteristics of data and selection of data analytics tools in a BD environment?	The selection of technologies appropriate for real-time DSS in a BD environment is influenced by the emergence of multiple data sources, rise in data volume, high velocity data, the business user requirements to analyse and report on unstructured and diverse data types.
1.2	What existing technologies are appropriate for real-time DSS in a BD environment?	A wide range of technologies have been developed which can be used to process BD. These include; NOSQL databases, In-memory databases, Spark, Streaming technologies, Hadoop, MPP databases, GreenPlum, PivotalImpala and commercial appliances.

The responses to this question received from participants, suggest several factors that influence the selection of technologies that are appropriate for real-time DSS in a BD environment. These factors include the characteristics of BD, business user requirements, the characteristics of technologies such as resource usage, performance and pricing of technologies.

Sub-research question 1.1: What is the relationship between the characteristics of data and selection of data analytics tools in a BD environment?

This question is designed in order for the participants to describe the data that is available in their organisations and used for decision making. The participants are asked to state the technologies used for analysing this data, and in cases where the technologies being used were not ideal, the participants are asked to state what would be ideal. The objective is to identify any relationship between the characteristics of data and the technologies used to analyse that data. The answers to this question and findings from existing literature suggest that the characteristics of data in an organisation have an impact on the selection of technologies appropriate for real-time DSS in a BD environment. The interview results revealed that data which is characterised by high volume, high velocity, multiple and diverse data sources, structured and unstructured data types, complex data structures require new set of technologies to analyse it. Traditional BI tools are mainly used to analyse structured data but as the data volume increases, indications are that organisations adopt BD technologies such as Hadoop, Spark and In-memory databases. The same pattern is also seen with high velocity data where indications are that when a lot of changes happen quickly, organisations find it difficult to manage data using traditional BI tools. This leads to the adoption of new technologies such as Hadoop and Spark. High volume data require technologies which are scalable. Furthermore, it was seen that as the number of data sources increase, organisations find it difficult to integrate and analyse data in real-time using traditional BI tools. The data integration process takes a long time which is not ideal for real-time requirements. The small percentage of organisations which are analysing both structured and unstructured data types use NOSQL databases and Hadoop or a combination of these new technologies with relational databases. Organisations which rely on traditional BI tools are not leveraging unstructured data types. This implies that the data type has an impact on the selection of technologies appropriate for real-time DSS in a BD environment.

Sub-research question 1.2: What existing technologies are appropriate for real-time DSS in a BD environment?

This question is designed to explore technologies appropriate for real-time DSS in a BD environment. The answers from the participants and findings from the literature indicate that traditional BI tools are not appropriate for real-time DSS in a BD environment. This is seen in organisations which use traditional BI tools but are not leveraging BD. Furthermore, it is seen that organisations which are analysing BD either in real-time, or near-real time, use BD technologies such as Hadoop, Spark, Impala and NOSQL databases.

Research question 2: How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?

The objective of this question is to explore the process of evaluating technologies appropriate for real-time DSS in a BD environment and to propose evaluation criteria applicable for comparing and evaluating these technologies. Table 5-2 depicts the summarised answers given by the interview participants in response to this question.

Table 5-2 Research question 2: summary of answers

No.	Sub-research question	Summary of participants' responses
2.1	What are the existing guidelines, frameworks, criteria, or measures applicable when comparing analytics tools for real-time DSS in data environments?	There are no standard BD benchmarks that can be used by organisations to compare and evaluate technologies appropriate for real-time DSS in a BD environment. No standard framework or guideline could be identified that can be used to compare and evaluate technologies appropriate for real-time DSS. Organisations rely on product vendor reports, on-line analysts' reports such as GitHub and Gartner magic quadrant reports. Evaluation criteria were identified that can be used to compare and evaluate technologies that are appropriate for real-time DSS in a BD environment.

Sub-research question 2.1: What are the existing guidelines, frameworks, criteria, or measures applicable when comparing analytics tools for real-time DSS in data environments?

The perception based results from the interviews appear to indicate that there are no standard guidelines, frameworks or benchmarks currently designed to assist organisations in evaluating and selecting technologies which are appropriate for real-time DSS in BD environments. Indications from the interviews are that organisations rely on the following methods when selecting these technologies:

- Individual intuition by a few IT experts who have knowledge about BD.
- Micro benchmarks which are designed for specific products.
- Product vendor reports
- Internal benchmarks which are run as a proof of concept.
- On-line analysts report from GitHub and other analysts such as Gartner.

Furthermore, based on the findings from the interviews and experiments, 12 evaluation criteria are identified that can be used to evaluate and select technologies that are appropriate for real-time DSS in a BD environment. These are show in Figure 5-1

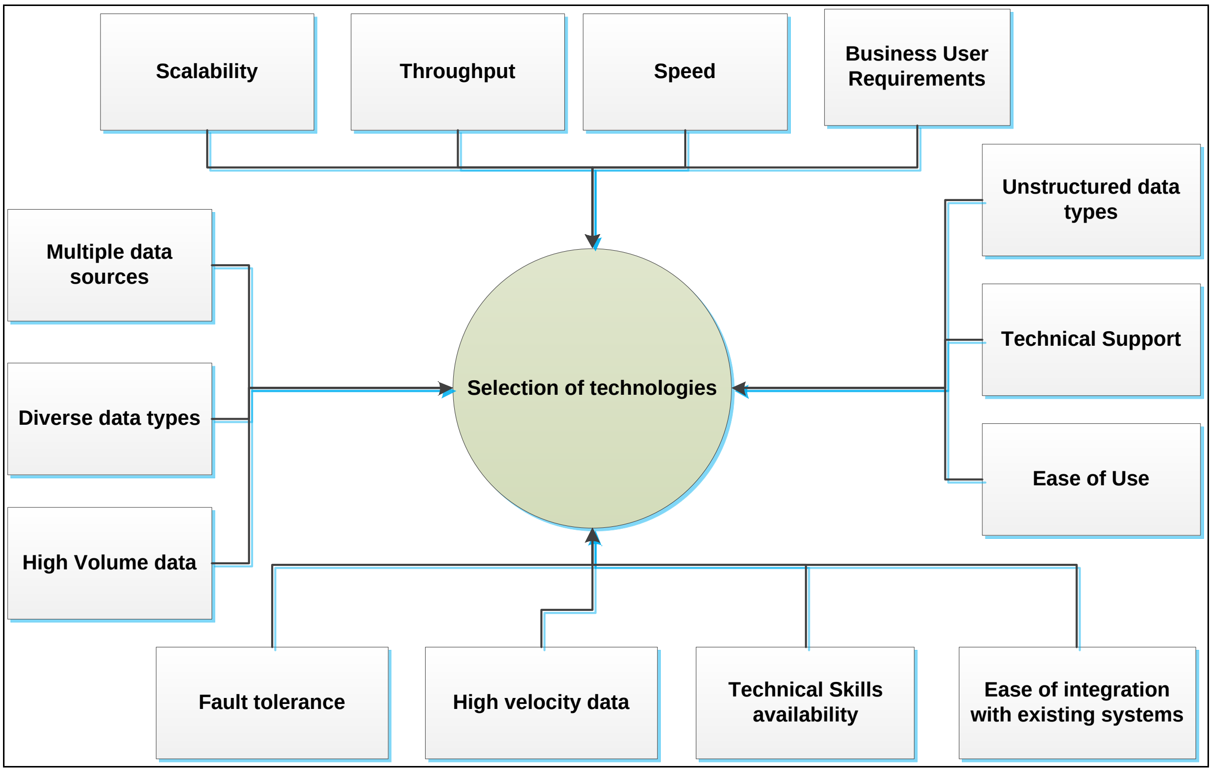


Figure 5-1 Evaluation Criteria framework

5.5 Summary

This chapter presents a discussion of the qualitative and quantitative research findings. The discussion is made up of two main parts relating to the two main objectives of this research. This is; to identify factors that influence the selection of technologies appropriate for real-time DSS, and to propose evaluation criteria framework that can be used to assess and select such technologies. The findings from the qualitative data analysis are placed into categories and after extensive analyses, nine themes are derived. The themes developed from this research include: multiple data sources, volume of data, high velocity data, structured and unstructured data, business user requirements, integration with traditional database and DW technologies, scalability, system performance and technical skills availability. Furthermore, nine main evaluation criteria are derived from the analysis of the data collected and depicted Section 5.3. The next chapter presents the conclusion, recommendations for further research and a reflection on this research process.

6 CHAPTER SIX: CONCLUSION

6.1 Introduction

The aim of this research is to explore factors that influence the selection of appropriate technologies for real-time DSS in a BD environment and to find evaluation criteria that can be used by organisations to determine these technologies. To achieve this aim, the research adopted a sequential exploratory mixed methods approach by using semi-structured interviews and computer laboratory experiments. The interview participants provide insights on the concept of BD and real-time DSSs. The preceding chapters are made up of an introduction to the research, reviewed literature, research methodology, findings and discussion of the findings. The research problem, aim and objectives of this research as well as the research questions which guided the research, are detailed in Sections 1.3 and 1.4. Chapter 2 delves into existing literature by various authors in order to explore the concept of BD, real-time DSS and BD technologies. The comprehensive literature review is guided by the research questions posed in Chapter 1.3. The research philosophy, research method, research approach, design and techniques applied in this research are explained in Chapter 3. Chapter 4 provides a presentation of the findings obtained from the semi-structured interviews and the computer laboratory experiments.

All these chapters have assisted in identifying factors which influence the selection of technologies appropriate for real-time DSS in a BD environment and to propose evaluation criteria that can be used to evaluate and determine these technologies.

As seen in Chapter 2.5, the emergence of BD has seen the development of numerous technologies and computing frameworks designed to alleviate BD related challenges. The main challenge faced by organisations is how to assess and select the best technology for their real-time DSS needs. Ghazal *et al.* (2013) state that there is immense interest in BD by both, academia and industry, which has driven both commercial and open source technology providers to develop a wide variety of products to store and process BD. The authors assert that as these products mature, there is a need to evaluate and compare these systems. In existing literature as at the time of this research, the concept of comparing and evaluating technologies for BD seems to be common to performance alone.

In this research, it is observed that organisations are driving more towards real-time DSSs or operational BI in order to gain competitive advantage. According to Farooq *et al.* (2010), business users now require latest or real-time data for the purpose of analysis and decision making. This requirement has seen the development of a variety of technologies designed for real-time data integration, analytics and reporting. However, the emergence of BD has rendered existing traditional BI tools inefficient and ineffective when delivering data for analytics. Several surveys on big data technologies (Dilpreet & Reddy, 2014; Chen & Zhang, 2014; Doulkeridis & Nørvåg, 2013; Liu *et al.*, 2013; Begoli, 2012), have been conducted which reveals that industry and academia have invented new approaches and technologies to process BD. However, the challenge is that none of the newly developed technologies is a one-size-fits-all solution. The new technologies are so numerous that organisations are faced with the challenge of determining what is appropriate for their requirements because big data is still a new concept and there are no standard guidelines or frameworks available to assist in evaluating and comparing big data technologies. From the literature review conducted in Chapter 2.6, it is revealed that there are several BD benchmarks proposed (Ghazal *et al*., 2013; Xiong *et al.*, 2013) but as noted by Dilpreet and Reddy (2014), these benchmarks are designed for specific products and are not standard. In this research, several factors which influence the selection of technologies appropriate for real-time DSS are identified through literature analysis (secondary data), interviews (primary data) and experiments (primary data). As discussed in Section 5.2, these factors include, multiple data sources, volume of data, velocity of data, variety of data structures (structured and unstructured), business user requirements, integration with existing systems and platforms, scalability and system performance.

Although some of the organisations involved in this research are not yet using BD technologies for real-time DSS requirements, answers to the research questions provided by participants from these organisations play a key role in the research as some of organisations have already started seeing the impact of BD on traditional BI tools.

Research question 2 and its sub-research questions are used to identify evaluation criteria that can be used to assess and select technologies that are appropriate for real-time DSS in a BD environment. The criteria identified in this research include system performance, ability to scale, ability to process fast changing data, ability to integrate both structured and unstructured data types, ability to integrate data from multiple sources, ability to seamlessly adapt to changes in data structures, ability to integrate with existing systems, costs associated with the technology and computer resource consumption by the new technology.

It is noted that all the participants in this research mentioned that system performance is the most important of all the identified criteria. According to the perception based results from the interviews and findings from existing literature, this is largely because real-time DSS is concerned with delivering fresh information quickly to decision makers. To test if performance can actually be used to compare and evaluate technologies, experiments were conducted as explained in chapter 3. The experiment results indicate that query execution time for Impala, Hive and Spark are statistically significantly different. This appears to imply that the performance of technologies can be used to compare and evaluate big data analytics technologies. The evaluation criteria identified are discussed in detail in section 5.3 and the assessment of each criterion depends on the organisation’s priorities.

6.2 **Contributions**

The concept of BD is still a new phenomenon which is in the process of being accepted and adopted by many companies especially, in South Africa. The aim of this research is to explore the real-time DSS in a BD environment and propose evaluation criteria that can be used to assess and select BD technologies that are appropriate for real-time DSS. The main contribution from this research is therefore, the proposed evaluation criteria. The proposed criteria can assist researchers and practitioners, making well informed decisions about the right choice of technologies appropriate for real-time DSS in a BD environment. Furthermore, this research can be used as a starting point for further research by academics or researchers. The research output can also help readers who are new to the topic of BD to gain an understanding of the concepts on BD and its related technologies for the purposes of real-time DSS.

6.3 Recommendations for further research

This research has highlighted a number of areas on which further research would be beneficial.

- The future research and work can potentially involve evaluating different technologies within a specific environment by applying ratings or weights to the evaluation criteria proposed in this research. In this case, the aim will be to choose the right technology for a particular application. This approach could provide a first step to analyse the effectiveness of each of the technologies when handling real-world applications.
- This research can be furthered by widening the geographic area for the interviews to include other countries especially, those that have accepted BD and are already using BD technologies at a wider scale than in South Africa. By having such a large sample survey, valuable perceptions and insights can be obtained which will be useful in many practical and research activities.
- The experimental evaluation conducted in this research can be further improved by widening the tools, increasing the data loads, including unstructured data and using a more complex algorithm such as k-means. This will require an upgraded platform for the experiments by adding more disk space, RAM, CPU and network infrastructure.
- Another area of interest that can be considered as a follow-up to this research, is evaluating technologies based streaming data instead of data at rest.
- Another aspect that can potentially be investigated is comparing different technologies based on ease of use. In this case, the objective is to consider technologies designed to do the same function, setup environments in different places or companies and investigate how quick the different companies get to use the technologies.

6.4 Reflection

The semi-structured interviews are composed of both, open-ended and close-ended questions, in order to guide the participants when answering questions. The open ended questions allow participants to express fully their understanding and experience of the concept under investigation. The qualitative research phase of the research is restricted to companies based in South Africa. However, this results in a small sample size of 10 participants. As mentioned in further research, it may be advisable to include other participants from other countries in order to proceed with a large sample survey. In this research, there are no pre-interview sessions. Pre-interviews may have assisted in refining the interview questions before the actual data collection. Some concepts ended up being corrected during the interview sessions and this resulted in trimming down the number of questions from 15 to 11.

For the empirical evaluation of Spark, Impala and Hive, only structured datasets were used and the aspect of unstructured and semi-structured data analysis was not tested.

6.5 **Summary**

Chapter 6 focuses on the conclusions, the contributions and recommendations for future research. The research questions posed in Chapter 1 and the hypothesis are answered supported in-depth analysis in Chapter 4, with the aid of semi-structured interviews and experimental evaluations, according to the delimitation and scope of this research.

REFERENCES

Agrawal, D., Bernstein , Philip Bertino, E., Davidson, Susan Dayal, Umeshwar Franklin, Michael Gehrke, J., Haas, L., Halevy, A. & Han, J. 2009. *Challenges and Opportunities with Big Data Challenges and Opportunities with Big Data*. http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf. [14 May 2014].

Apache. 2014. *The Apache HBase^{TM} Reference Guide.* http://hbase.apache.org/book.html#arch.overview.nosql. [13 Jun 2014].

ApachePig. 2013. *Welcome to Apache Pig!* http://pig.apache.org/. [13 Jun 2014].

Bakshi, K. 2012*. Considerations for Big Data : Architecture and Approach.* : 1–7. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6187357. [10 May 2014].

Blaikie, N. 2009. *Designing Social Research.* Cambridge: Polity Press.

Begoli, E. 2012. *A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data*. *Proceedings of the WICSA/ECSA 2012 Companion Volume on - WICSA/ECSA '12*: 177. http://dl.acm.org/citation.cfm?doid=2361999.2362039. [20 May 2014]

Borthakur, D. 2007. *The hadoop distributed file system: Architecture and design*. *Hadoop Project Website*: 1–14. https://svn.eu.apache.org/repos/asf/hadoop/common/tags/release-0.16.3/docs/hdfs_design.pdf. [11 May 2014].

Chardonnens, T., Cudre-mauroux, P., Grund, M. & Perroud, B. 2013. *Big Data Analytics on High Velocity Streams : A Case Study.* : 784–787. [13 May 2014].

Chaudhuri, S. & Dayal, U. 1997. *An overview of data warehousing and OLAP technology*. *ACM SIGMOD Record*, 26(1):65–74. [13 May 2014].

Chaudhuri, S., Dayal, U. & Narasayya, V. 2011. *An overview of business intelligence technology*. *Communications of the ACM*, 54(8):88. http://portal.acm.org/citation.cfm?doid=1978542.1978562. [19 March 2014].

Chen, M., Mao, S. & Liu, Y. 2014. Big Data: A Survey. *Mobile Networks and Applications*, (January): 171–209. http://link.springer.com/10.1007/s11036-013-0489-0. [23 March 2014].

Clare Bless, Craig Higson-Smith, Ashraf Kagee, 2006. *Fundamentals of social research methods an African perspective*. Cape Town: Juta

Cloudera. 2014. Download 5.3.3. http://www.cloudera.com/downloads/cdh/5-3-3.html [15 Aug 2014].

Cuzzocrea, A., Saccà, D. & Ullman, J. 2013. *Big data: a research agenda*. *Proceedings of the 17th International…*: 198–203. http://dl.acm.org/citation.cfm?id=2527071. [20 March 2014].

Dana S. Dunn, 2010. *The practical researcher a student guide to conducting psychological research.* New York: Wiley-Blackwell.

Davenport, T.H., Barth, P. & Bean, R. 2012. *How " Big Data " is Different.* , 54(1). http://www.stevens.edu/howe/sites/default/files/MIT-SMR How Big Data is Different.pdf. [20 August 2015].

David B. Resnik, JD., Ph.D. 2011. *What is Ethics & Why is it important.* New York: Oxford University Press.

De Vos, AS., Delport, CSL., Fouche, CB., Strydom, H. 2011. *Research at grass roots: For the social sciences and human services professions.* Pretoria: Van Schaik.

Dayal, U., Wilkinson, K., Castellanos, M. & Alkis, S. 2009. *Data Integration Flows for Business Intelligence*. : 1–11. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.366.6196&rep=rep1&type=pdf. [25 May 2014]

Dean, J. & Ghemawat, S. 2008. *MapReduce: simplified data processing on large clusters. Communications of the ACM*: 1–13. http://dl.acm.org/citation.cfm?id=1327492. [11 May 2014].

Dehne, F. & Zaboli, H. 2012. *Parallel Real-Time OLAP on Multi-core Processors. 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*: 588–594. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6217470. [12 April 2014].

Delic, K.A., Douillet, L. & Dayal, U. 2001. *Towards an Architecture for Real-Time Decision Support Systems* : 303–311. http://ieeexplore.ieee.org.ezproxy.cput.ac.za/stamp/stamp.jsp?tp=&arnumber=938098. [14 August 2014].

Dilpreet, S. & Reddy, C.K. 2014. *A survey on platforms for big data analytics.* , (iv). http://dmkd.cs.wayne.edu/Papers/JBD14.pdf. [29 March 2015].

Dobre, C. & Xhafa, F. 2013. *Parallel Programming Paradigms and Frameworks in Big Data Era*. *International Journal of Parallel Programming*. http://link.springer.com/10.1007/s10766-013-0272-7 [14 April 2014].

Dong, X.L. & Srivastava, D. 2013. *Big Data Integration.* : 1188–1189. http://www.vldb.org/pvldb/vol6/p1188-srivastava.pdf. [19 June 2014].

Doulkeridis, C. & Nørvåg, K. 2013. *A survey of large-scale analytical query processing in MapReduce. The VLDB Journal*, (123). http://link.springer.com/10.1007/s00778-013-0319-9 [2 April 2014].

Duggal, P.S.& P.S. 2013. Big Data Analysis : Challenges and Solutions. : 269–276.

Earl Babbie and Johann Mouton, *The practice of social research*, 2001, Cape Town: Cambridge University Press.

Elizabeth Henning, Wilhelm van Rensburg and Brigitte Smit, 2004. *Finding your way in qualitative research*. Pretoria: Van Schaik.

Erin Horvat, Mary Lou Heron, Emily Tancredi-Brice Agbenyega and Bradley W. Bergey, 2013. *The beginner's guide to doing qualitative research*, New York: Teachers College Press.

Farooq, F. & Sarwar, Mansoor, S. 2010. *Real-Time Data Warehousing For Business Intelligence*: 10.1145. http://dl.acm.org/citation.cfm?id=1943666. [12 May 2014].

Gantz, B.J. & Reinsel, D. 2011. *Extracting Value from Chaos State of the Universe : An Executive Summary.* , (June): 1–12. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. [10 May 2014].

Garber, L. 2012. *Using in-memory analytics to quickly crunch big data. Computer*: 16–18. http://www.computer.org/csdl/mags/co/2012/10/mco2012100016.pdf. [11 March 2014].

Ghazal, A., Rabl, T., Hu, M. & Raab, F. 2013. *Bigbench: Towards an industry standard benchmark for big data analytics*. *… on Management of data*: 1197–1208. http://dl.acm.org/citation.cfm?id=2463712. [9 May 2014].

Ghemawat, S., Gobioff, H. & Leung, S. 2003. *The Google file system. ACM SIGOPS Operating Systems …*, 37(5): 29. http://portal.acm.org/citation.cfm?doid=1165389.945450 [4 June 2014].

Golden Gate, S. & Inc. 2009. *Going Real-Time for Data Warehousing and Operational BI Enabling Real-Time Data Integration.* , (Cdc). http://datasolutions.searchdatamanagement.com/document;5132934/datamgmt-abstract.htm. [19 May 2014].

Gray E.D, 2009. *Doing Research in the real World, 2nd ed.* London: Sage.

Greener, I. 2011. *Designing Social Research: A Guide For The Bewildered.* 1st ed. London: Sage.

Gualtieri, M. 2013. *Evaluating Big Data Predictive Analytics Solutions.* New York: Forester.

Holden, M.T. & Lynch, P. 2004. *Choosing the Appropriate Methodology : Understanding Research Philosophy.* , (2002): 397–409. http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=c75b9408-1d88-4f0b-ab56-6f360effae75%40sessionmgr4004&vid=1&hid=4201. [20 June 2014].

Hossain, S.A. 2013. *NoSQL Database: New Era of Databases for Big data Analytics-Classification, Characteristics and Comparison*. *… Journal of Database …*, 6(4): 1–14. http://www.earticle.net/Article.aspx?sn=207903. [23 March 2014].

Jane Ritchie, Jane Lewis, 2003. *Qualitative research practice: A guide for social science students*, Los Angeles: Sage.

John W. Creswell, 2009. *Research Design, Qualitative, Quantitative, and Mixed Methods Approaches.* Los Angeles: Sage.

Johann Mouton, 1996. *Understanding social research*. *1st ed*. Pretoria: J.L. van Schaik.

Johnson, R.B. & Onwuegbuzie, A.J. 2012. *Mixed Methods Research : A Research Paradigm Whose Time Has Come.* , 33(7): 14–26.

John W. Creswell and Vicki L. Plano Clark, 2011. *Designing and conducting Mixed Methods Research.* Lincoln: Sage.

Jörg, T. & Dessloch, S. 2010. *Near real-time data warehousing using state-of-the-art ETL tools. Enabling Real-Time Business Intelligence*. http://link.springer.com/chapter/10.1007/978-3-642-14559-9_7. [19 April 2014].

Keen, G. & Peter, W. 1980. *Decision support systems: a research perspective. 54*. http://18.7.29.232/handle/1721.1/47172 [17 April 2014].

Kemper, A. & Neumann, T. 2011. *HyPer: A hybrid OLTP&OLAP main memory database system based on virtual memory snapshots. Data Engineering (ICDE), 2011 IEEE …*: 195–206. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5767867. [11 March 2014].

Laney, D. 2001. *3D data management: Controlling data volume, velocity and variety. META Group Research Note*, (February 2001). http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:3D+Data+Management:+Controlling+data+volume,+velocity+and+variety#0. [23 April 2014].

Lee, J., Kwon, Y. & Farber, F. 2013. *SAP HANA distributed in-memory database system: Transaction, session, and metadata management. … (ICDE), 2013 IEEE …*: 1165–1173. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6544906 [23 March 2014].

Leedy, P.D. & Ormrod, J.E. 2010. *Practical Research: Planning and Design*. 9th edition. New Jersey: Pearson Educations Inc.

Letouze, E. 2012. *Big Data for Development : Challenges & Opportunities.* , (May). http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf. [12 May 2014].

Liu, X., Iftikhar, N. & Xie, X. 2013. *Survey of Real-time Processing Systems for Big Data.* In pp. 356–361. http://delivery.acm.org/10.1145/2630000/2628251/p356-liu.pdf?ip=198.54.223.106&id=2628251&acc=ACTIVE SERVICE&key=646D7B17E601A2A5.BB5645D894760FF1.4D4702B0C3E38B35.4D4702B0C3E38B35&CFID=534429282&CFTOKEN=71354785&__acm__=1439015785_4e0751d17a444f32d1d4f29. [15 June 2014].

Manyika, J., Chui, M., Brown, B. & Bughin, J. 2011. *Big data: The next frontier for innovation, competition, and productivity.* , (May). http://www.citeulike.org/group/18242/article/9341321. [18 April 2014].

Maree K. 2012. *Complete your thessis or disserttion successfully: Practical guidelines.* Cape Town: Juta & company Ltd.

Mark Saunders, Philip Lewis and Adrian Thornhill, 2009. *Research Methods for business students, fifth edition*. New York: Prentice Hall.

Martyn Denscombe, 2007. *The Good Research Guide for small-scale social research projects*. Finland: WS Bookwell.

Marshall, L. & Harpe, R.D. la. 2009. *Decision making in the context of business intelligence and data quality*. *South African Journal of …*, 11(June): 1–15. http://reference.sabinet.co.za/sa_epublication_article/info_v11_n2_a2. [19 April 2014].

Marz, Narthan; Warren, J. 2012. Big Data Principles and best practices for scalable realtime data systems. Greenwich: Manning Publications.

Mcguire, T., Manyika, J. & Michael, C. 2012. Why Big Data is the new competitive advantage. http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.U25XUYGSyZc. [29 May 2014].

Mctaggart, C. 2008. *Hadoop / MapReduce*. http://www.cs.colorado.edu/~kena/classes/5448/s11/presentations/hadoop.pdf. [21 May 2014].

Michael G. Noll. 2014. *Applied Research. Big Data. Distributed Systems. Open Source. Running a multi-node storm cluster.* http://www.michael-noll.com/tutorials/running-multi-node-storm-cluster/. [16 Aug 2014].

Morgan, T. 2013. *VMware teaches Serengeti big-data virt new Hadoop tricks*. http://www.theregister.co.uk/2013/04/02/vmware_serengeti_hadoop_update/. [29 July 2014].

Mühlbauer, T., Rödiger, W. & Reiser, A. 2013. *ScyPer: A Hybrid OLTP&OLAP Distributed Main Memory Database System for Scalable Real-Time Analytics*. *BTW*: 499–502. http://db.in.tum.de/people/sites/roediger/papers/muehlbauer2012scyper.pdf. [17 April 2014].

Neubauer, P. 2010. *Graph databases, NOSQL and Neo4j.* http://www.infoq.com/articles/graph-nosql-neo4j. [11 May 2014].

Norman Blaikie, 2004. *Analyzing quantitative data*. London: Sage.

Norman K. Denzin and Yvonna S. Lincoln, 2008. *The landscape of qualitative research*. New York: Sage.

NOSQL-meetup. 2009. *NOSQL meetup*. http://www.eventbrite.com/e/nosql-meetup-tickets-341739151 [20 June 2014].

Oracle. 2009. *Orcale TimeTen In-Memory Database Architectural Overview*. , (6). http://download.oracle.com/otn_hosted_doc/timesten/603/TimesTen-Documentation/arch.pdf. [13 July 2014].

Özsu, M. & Valduriez, P. 2011. *Principles of distributed database systems*. http://books.google.com/books?hl=en&lr=&id=TOBaLQMuNV4C&oi=fnd&pg=PR7&dq=Principles+of+distributed+database+systems&ots=LpGo9D_S3f&sig=opMcuXh-V1k69A3Qegid-p5gbZk. [19 April 2014].

Pavlo, A., Paulson, E. & Rasin, A. 2009. *A comparison of approaches to large-scale data analysis*. *… on Management of data*. http://dl.acm.org/citation.cfm?id=1559865. [15 March 2014].

Pereira, D. & Azevedo, L. 2012. *Real time data loading and OLAP queries: Living together in next generation BI environments*. *… of Information and Data …*, 3(2): 110–119. http://seer.lcc.ufmg.br/index.php/jidm/article/view/183. [7 March 2014].

Philip Chen, C.L. & Zhang, C.-Y. 2014. *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences*, (January). http://linkinghub.elsevier.com/retrieve/pii/S0020025514000346. [30 January 2014].

Pokorny, J. 2013. *NoSQL databases: a step to database scalability in web environment. International Journal of Web Information Systems*, 9(1): 69–82. http://www.emeraldinsight.com/10.1108/17440081311316398 [11 April 2014].

Purcell, B. 2013. *The emergence of " big data " technology and analytics*. : 1–7.

Ranjit Kumar, 2011. *Research methodology, a step by step guide for beginners*. London: Sage.

Sahay, B.S. & Ranjan, J. 2008. Real time business intelligence in supply chain analytics. *Information Management & Computer Security*, 16(1): 28–48. http://www.emeraldinsight.com/10.1108/09685220810862733. [20 February 2014].

Sandu, D.I. 2008. *Operational and real-time Business Intelligence.* , 3(3): 33–36. http://revistaie.ase.ro/content/47/06Sandu.pdf. [29 March 2015].

Singh, S. & Singh, N. 2012. Big Data analytics. *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*: 1–4. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6398180. [14 Jul 2015].

Stonebraker, M., Madden, S. & Dubey, P. 2013. *Intel " Big Data " Science and Technology Center Vision and Execution Plan.* , 42(1). http://dl.acm.org/citation.cfm?id=2481537. [15 June 2014]

Su, X. & Swart, G. 2012. *Oracle in-database Hadoop: when MapReduce meets RDBMS*. *Proceedings of the 2012 ACM SIGMOD International …*: 779–789. http://dl.acm.org/citation.cfm?id=2213955. [25 March 2014].

Tank, D.M., Ganatra, A., Kosta, Y.P. & Bhensdadia, C.K. 2010. *Speeding ETL Processing in Data Warehouses Using High-Performance Joins for Changed Data Capture (CDC)*. *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, (Cdc): 365–368. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5656810 [2 March 2014].

Teddlie, C. & Tashakkori, A., 2009. *Foundations of Mixed Methods Research.* Thousand Oaks, CA: Sage.

Tekiner, F. & Keane, J.A. 2013. *Big Data Framework.* http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6722011. [22 May 2014].

Thusoo, A., Sarma, J. & Jain, N. 2009. *Hive: a warehousing solution over a map-reduce framework. Proceedings of the ….* http://dl.acm.org/citation.cfm?id=1687609 [25 March 2014].

Volts, G. 2015. *Data Warehousing and mining*. http://alchetron.com/Data-warehousing-and-mining-739-W. [21 July 2015].

Wanderman-Milne, S. & Li, N. 2014. *Runtime code generation in Cloudera Impala. IEEE Data Engineering Bulletin*: 31–37. ftp://131.107.65.22/pub/debull/A14mar/p31.pdf [12 May 2014].

Watson, H.J. & Wixom, B.H. 2007. *The Current State of Business*. Intelligence.http://ieeexplore.ieee.org.ezproxy.cput.ac.za/stamp/stamp.jsp?tp=&arnumber=4302625. [20 May 2015].

Vicki L. Plano Clark and John W. Creswell, 2008. *The mixed methods reader. Lincoln*: Sage.

Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y., Bai, X., Li, Y. & Xu, C. 2013. A characterization of big data benchmarks. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, (1): 118–125.

Yan, J. 2013. Big Data , *Bigger Opportunities collaborate in the era of big data.* http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf. [16 June 2014].

Zaharia, M., Chowdhury, M., Das, T. & Dave, A. 2012. *Fast and interactive analytics over Hadoop data with Spark.* : 45–51. https://www.usenix.org/system/files/login/articles/zaharia.pdf. [12 May 2014].

Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S. & Stoica, I. 2010. Spark : Cluster Computing with Working Sets. http://www.cs.berkeley.edu/~matei/papers/2010/hotcloud_spark.pdf. [17 August 2014].

Zhang, H., Chen, G., Ooi, B.C., Tan, K.-L. & Zhang, M. 2015. *In-Memory Big Data Management and Processing: A Survey. IEEE Transactions on Knowledge and Data Engineering*, 27(7): 1–1. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7097722. [12 July 2015].

Zhong, T., Doshi, K., Tang, X., Lou, T. & Li, Z. 2013. *On Mixing High-Speed Updates and In-Memory Queries. prof.ict.ac.cn*: 102–109. http://prof.ict.ac.cn/bpoe2013/downloads/papers/S7212_5873.pdf. [23 March 2014].

Zina O'Leary, 2004**.** *The Essential Guide to Doing Research*. London: SAGE**.**

APPENDICES

Appendix A: Research interview questions

Number	Interview question
1	What are the sources of data in your organization used for decision making? *You can describe how data is generated in your organisation for example*: - Internet and On-Line Applications - Transactional database applications. - Digital Systems(e.g. sensors) - CCTV - Other(s)
2	How would you describe the data available in your organization used for decision making? *You can use examples below but feel free to describe what is available.* - Voluminous - Quickly changes - Comes from different sources - Comes from a single source - Unstructured in nature (files, video). - Other(s)
3	How would you describe the volume of data available for decision making in your organization? - Very High – Zeta bytes and above - High – Petabytes - Medium – Terabytes - Low – Less than Terabytes
4	How would you describe the rate at which data arrives in your system or the rate at which data changes within your systems for analytic purposes? Example: - Per second - Per minute - Per hour - Per day Tell us more about how the data is generated and analysed.
5	What type of data formats/structures exist in your company? Are all these data types being analysed? Please feel free to add more and describe. - Text files - Pictures

	- Video - Voice
6	What do you currently use to analyze your data? Example: - Traditional database and data warehouse (OLAP) - Hadoop - Hybrid (Combination of Hadoop and Traditional databases) - In-Memory databases (e.g. SAP-Hana) - NOSQL databases (e.g. MongoDB). - Other (Please Specify) Are the technologies being used capable of analysing all data available?
7	If you are using Hadoop, which distribution do you use? - Apache - Cloudera - Amazon - Other
8	If you are using Hadoop, what analytic components do you use? - Impala - Hive - Pig - Spark - Other (Please Specify)
9	When selecting technologies for analysing big data, what features of the technology and other factors did you consider? Tell us more about what was considered. For example: - Performance - latency - Technology resource utilisation such as memory, CPU and disk I/O - Ease of use - Costs - Availability of technical skills - Technical Support
10	How did you evaluate and compare the different technologies when you selected what you are currently using to analyse data?

Appendix B: Dimensional file/table structure

```
CREATE EXTERNAL TABLE publisher_dim(
  id INT,
  version INT,
  date_from TIMESTAMP,
  date_to TIMESTAMP,
  publisher_id INT,
  network_id INT,
  rating INT,
  status STRING,
  publisher_name STRING,
  activation_date TIMESTAMP,
  deactivation_date TIMESTAMP,
  sales_owner_id INT,
  sales_owner_name STRING,
  account_owner_id INT,
  account_owner_name STRING,
  publisher_type STRING,
  classification STRING,
  persona STRING,
  media_source_type STRING,
  media_source_type_displayname STRING,
  media_source_subtype STRING,
  media_source_ircm_campaign_id INT,
  media_source_conversion_credit_rule STRING,
  hide_in_directory INT,
  platform_approval_state STRING,
  branded_signup_account_id INT,
  branded_signup_campaign_id INT,
  source_type STRING,
  source_account_id INT,
  risk_level STRING,
  donot_pay INT,
  donot_pay_reason STRING,
  vetted_to_pay INT,
  country STRING,
  state STRING,
  url STRING,
  prom_mthd_online_email INT,
  prom_mthd_online_content INT,
  prom_mthd_online_shopping INT,
  prom_mthd_online_coupon INT,
  prom_mthd_online_loyalty INT,
  prom_mthd_online_sem INT,
  prom_mthd_cashback_site INT,
  prom_mthd_subaffiliates_cpanetwork INT,
  prom_mthd_incentivized_consumer INT,
  prom_mthd_mobile INT,
  prom_mthd_offline_tv INT,
  prom_mthd_offline_radio INT,
  prom_mthd_offline_print INT,
  prom_mthd_offline_billboard INT,
  prom_mthd_offline_outdoor INT,
  prom_mthd_offline_directmail INT,
  prom_mthd_offline_other INT,

  prom_mthd_other INT,

  dlu TIMESTAMP

 )  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

```
LOCATION '/data/publisher_dim';
```

Appendix C: Fact file/table structure

```
CREATE EXTERNAL TABLE conversion_winner_fact3

(

   process_num  INT,

   oid  STRING,

   conversion_type INT,

   action_id STRING,

   promo_code STRING,

   ref_type STRING,

   conversion_datetime  TIMESTAMP,

   first_click_referral_datetime TIMESTAMP,

   last_click_referral_datetime TIMESTAMP,

   campaign_id INT,

   campaign_dim_id  INT,

   action_tracker_dim_id INT,

   winner_publisher_dim_id INT,

   first_click_publisher_dim_id  INT,

   last_click_publisher_dim_id INT,

   network_dim_id INT,

   conversion_date_dim_id INT,

   sale_amount FLOAT,

   payout FLOAT,
   number_of_unique_participants INT,

   number_of_unique_clicks  INT,

   unique_participants STRING,

   winner_publisher_had_click INT,

   doe TIMESTAMP,

   landing_page_url STRING

) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

LOCATION '/data/conversion_winner_fact3';
```

Appendix D: Pairwise comparisons: tools used

(I) ToolsUsed	(J) ToolsUsed	Mean Difference (I-J)	Std. Error	Df	Sig.	95% Wald Confidence Interval for Difference	
						Lower	Upper
Impala	Impala						
	Hive	-200.7675[a]	.00707	1	.000	-200.7814	-200.7537
	Spark	.4076[a]	.00707	1	.000	.3938	.4215
Hive	Impala	200.7675[a]	.00707	1	.000	200.7537	200.7814
	Hive						
	Spark	201.1752[a]	.00674	1	.000	201.1619	201.1884
Spark	Impala	-.4076[a]	.00707	1	.000	-.4215	-.3938
	Hive	-201.1752[a]	.00674	1	.000	-201.1884	-201.1619
	Spark						

Appendix E: Pairwise comparisons:datasets

(I) RowsCode	(J) RowsCode	Mean Difference (I-J)	Std. Error	Df	Sig.	95% Wald Confidence Interval for Difference	
						Lower	Upper
12M	12M						
	52M	-3.6067[a]	.00954	1	.000	-3.6254	-3.5880
	103M	-8.7643[a]	.00954	1	.000	-8.7830	-8.7456
	255M	-22.9520[a]	.00954	1	.000	-22.9707	-22.9333
	510M	-49.8597[a]	.00954	1	.000	-49.8784	-49.8410
	1200M	-384.1333[a]	.01066	1	.000	-384.1542	-384.1124
52M	12M	3.6067[a]	.00954	1	.000	3.5880	3.6254
	52M						
	103M	-5.1577[a]	.00954	1	.000	-5.1764	-5.1390
	255M	-19.3453[a]	.00954	1	.000	-19.3640	-19.3266
	510M	-46.2530[a]	.00954	1	.000	-46.2717	-46.2343
	1200M	-380.5267[a]	.01066	1	.000	-380.5476	-380.5058
103M	12M	8.7643[a]	.00954	1	.000	8.7456	8.7830
	52M	5.1577[a]	.00954	1	.000	5.1390	5.1764
	103M						
	255M	-14.1877[a]	.00954	1	.000	-14.2064	-14.1690
	510M	-41.0953[a]	.00954	1	.000	-41.1140	-41.0766
	1200M	-375.3690[a]	.01066	1	.000	-375.3899	-375.3481
255M	12M	22.9520[a]	.00954	1	.000	22.9333	22.9707
	52M	19.3453[a]	.00954	1	.000	19.3266	19.3640
	103M	14.1877[a]	.00954	1	.000	14.1690	14.2064

	255M						
	510M	-26.9077[a]	.00954	1	.000	-26.9264	-26.8890
	1200M	-361.1813[a]	.01066	1	.000	-361.2022	-361.1604
510M	12M	49.8597[a]	.00954	1	.000	49.8410	49.8784
	52M	46.2530[a]	.00954	1	.000	46.2343	46.2717
	103M	41.0953[a]	.00954	1	.000	41.0766	41.1140
	255M	26.9077[a]	.00954	1	.000	26.8890	26.9264
	510M						
	1200M	-334.2737[a]	.01066	1	.000	-334.2946	-334.2528
1200M	12M	384.1333[a]	.01066	1	.000	384.1124	384.1542
	52M	380.5267[a]	.01066	1	.000	380.5058	380.5476
	103M	375.3690[a]	.01066	1	.000	375.3481	375.3899
	255M	361.1813[a]	.01066	1	.000	361.1604	361.2022
	510M	334.2737[a]	.01066	1	.000	334.2528	334.2946
	1200M						

Appendix F: Pairwise comparisons: tools used per dataset (row count)

RowsCode	(I) ToolsUsed	(J) ToolsUsed	Mean Difference (I-J)	Std. Error	Df	Sig.	95% Wald Confidence Interval for Difference	
							Lower	Upper
12M	Impala	Impala						
		Hive	-33.1120[a]	.01652	1	.000	-33.1444	-33.0796
		Spark	.3980[a]	.01652	1	.000	.3656	.4304
	Hive	Impala	33.1120[a]	.01652	1	.000	33.0796	33.1444
		Hive						
		Spark	33.5100[a]	.01652	1	.000	33.4776	33.5424
	Spark	Impala	-.3980[a]	.01652	1	.000	-.4304	-.3656
		Hive	-33.5100[a]	.01652	1	.000	-33.5424	-33.4776
		Spark						
52M	Impala	Impala						
		Hive	-43.6610[a]	.01652	1	.000	-43.6934	-43.6286
		Spark	.5020[a]	.01652	1	.000	.4696	.5344
	Hive	Impala	43.6610[a]	.01652	1	.000	43.6286	43.6934
		Hive						
		Spark	44.1630[a]	.01652	1	.000	44.1306	44.1954
	Spark	Impala	-.5020[a]	.01652	1	.000	-.5344	-.4696
		Hive	-44.1630[a]	.01652	1	.000	-44.1954	-44.1306
		Spark						
103M	Impala	Impala						
		Hive	-58.8940[a]	.01652	1	.000	-58.9264	-58.8616

		Spark	.5140[a]	.01652	1	.000	.4816	.5464
	Hive	Impala	58.8940[a]	.01652	1	.000	58.8616	58.9264
		Hive						
		Spark	59.4080[a]	.01652	1	.000	59.3756	59.4404
	Spark	Impala	-.5140[a]	.01652	1	.000	-.5464	-.4816
		Hive	-59.4080[a]	.01652	1	.000	-59.4404	-59.3756
		Spark						
255M	Impala	Impala						
		Hive	-100.4690[a]	.01652	1	.000	-100.5014	-100.4366
		Spark	.7500[a]	.01652	1	.000	.7176	.7824
	Hive	Impala	100.4690[a]	.01652	1	.000	100.4366	100.5014
		Hive						
		Spark	101.2190[a]	.01652	1	.000	101.1866	101.2514
	Spark	Impala	-.7500[a]	.01652	1	.000	-.7824	-.7176
		Hive	-101.2190[a]	.01652	1	.000	-101.2514	-101.1866
		Spark						
510M	Impala	Impala						
		Hive	-179.9290[a]	.01652	1	.000	-179.9614	-179.8966
		Spark	.9900[a]	.01652	1	.000	.9576	1.0224
	Hive	Impala	179.9290[a]	.01652	1	.000	179.8966	179.9614
		Hive						
		Spark	180.9190[a]	.01652	1	.000	180.8866	180.9514
	Spark	Impala	-.9900[a]	.01652	1	.000	-1.0224	-.9576
		Hive	-180.9190[a]	.01652	1	.000	-180.9514	-180.8866
		Spark						

1200M	Impala	Impala						
		Hive						
		Spark						
	Hive	Impala						
		Hive						
		Spark	787.8320[a]	.01652	1	.000	787.7996	787.8644
	Spark	Impala						
		Hive	-787.8320[a]	.01652	1	.000	-787.8644	-787.7996
		Spark						

Appendix G: Findings of interviews

RQ 1 & 2 and sub-research questions	P4,P5,P6,P7,P9	P10	P7,P8,P2,P1	Findings
RQ 1. What are the factors that influence the selection of technologies appropriate for real-time DSSs in a BD environment?	System Performance (Latency and throughput), existence of multiple sources of data, frequency of data changes.	Unstructured data sources, requirement to gain insight on social media data, continuous growth in data, Complexity of data structures, Integration with traditional BI tools.	Dynamic data structures, Scalability.	Discover unknown patterns, insights, reporting, and predictive analytics on all data available. Integration of data from multiple sources. Centralized reporting. Availability of technical support, Technical skills availability, Ease of use, Network usage, integration with existing technologies systems and architectures.
SRQ1.1. What is the relationship between	Increase in data sources, internet and on-line provision of	Monitoring operational digital gadgets. Real-	Integrate and leverage all data sources,	As data sources increase it becomes difficult to use traditional

characteristics of data and technologies used for analysing data in a real-time environment?	reports to customers in real-time, continuous growth of structured data, requirements to discover patterns and insights from social media data and other unstructured data sources such as videos and audio, Real-time monitoring of customer experience, identify risky customers, identify areas of weakness, detect fraudulent transactions in real-time and handle this in large quantities of data.	time monitoring of network activity, Monitoring customer sentiments on social media platforms, take corrective or proactive action for competitive advantage, To prevent loss of customers, loss of opportunities and loss of revenue. Track actions, clicks, and buying activities on-line and suggest offers in real-time. Real-time access to operational data, File formats.	Data structure is not known in advance, Scalable technologies, Frequency of data structure changes and ability to reflect changes quickly in reporting platform.	BI tools. As data sources increase, it needs the use of BD technologies. As complexity of data structure increase, use BD technologies. High data volume cannot be handled in real-time by traditional BI tools.
SRQ2. What are the tools available on the market for use in analyzing BD?	Traditional data warehouse, Tableau, Hybrid databases, ETL tools, Google analytics.	In-memory databases NOSQL databases, GreenPlum, Impala, Hive, Shark, Kafka, MPP databases, distributed databases.	Hadoop, Spark, Time series database (Round Robin database), Cloudera Hadoop, Hortonworks, Mapreduce, MapR Amazon redshift R.	Hadoop based solutions, distributed database solutions, data warehousing solutions, hybrid of data warehouse and Hadoop solutions, In-memory databases, NOSQL databases,MPP databases integrated with Hadoop. NOSQL databases combined with Hadoop or Spark, Streaming technologies

				such as Kafka.
RSQ2. How can an organization evaluate technologies appropriate for real-time DSS in a BD environment?	Use of existing BD benchmarks, use of existing BI benchmarks, utilization of product vendor reports, Use of on-line reports by analysts, Use of existing evaluation framework.	Executing internal proof of concept in the form of experiments to compare and evaluate technologies.		Use of micro benchmarks for big data technologies, use of known BD evaluation frameworks, reliance on product vendor reports.
SRQ 2.1. What are the existing guidelines, frameworks, criteria, or measures applicable when evaluating analytic tools for real-time DSS in BD environments?	Cost of ownership, system performance, speed of loading data, speed of producing reports, Ease of use, availability of technical skills, skilled resources, throughput.	Resource usage, hardware, ability to integrate with existing technologies, legacy systems, need for continuity, ability to connect to existing data sources in addition to new sources, Consistency in reporting after implementing new technology, CPU utilization, memory consumption , CPU utilization, impact of new technology on existing network and system performance.	Scalability, Fault tolerance, Licensing models, license per data node, license per CPU or license per user. Open source technologies because of costs. Commercial products because of technical support availability.	Costs of technologies, system performance, ability to scale, fault tolerance, ability to integrate data from multiple sources, ability for technology to integrate with existing BI systems and technologies, Ease of use, ability to process high data volume, ability to process both structured and unstructured data types, ability to process high velocity changing data, Computer resource usage such as CPU utilization, memory consumption and of data processing on network performance.
SRQ 2.2. How can an organization	Product vendor reports, used existing	Used internal benchmark.	On-line reports, analysts' reports	Develop list of evaluation criteria of that

evaluate analytic tools appropriate for real-time DSS in a BD environment?	big data benchmark.		e.g. Gartner magic quadrant report. Carried out a POC to compare and evaluate technologies before making a decision.	can be used to compare and evaluate technologies, develop BD benchmarks.

Appendix H: Summary of interview findings

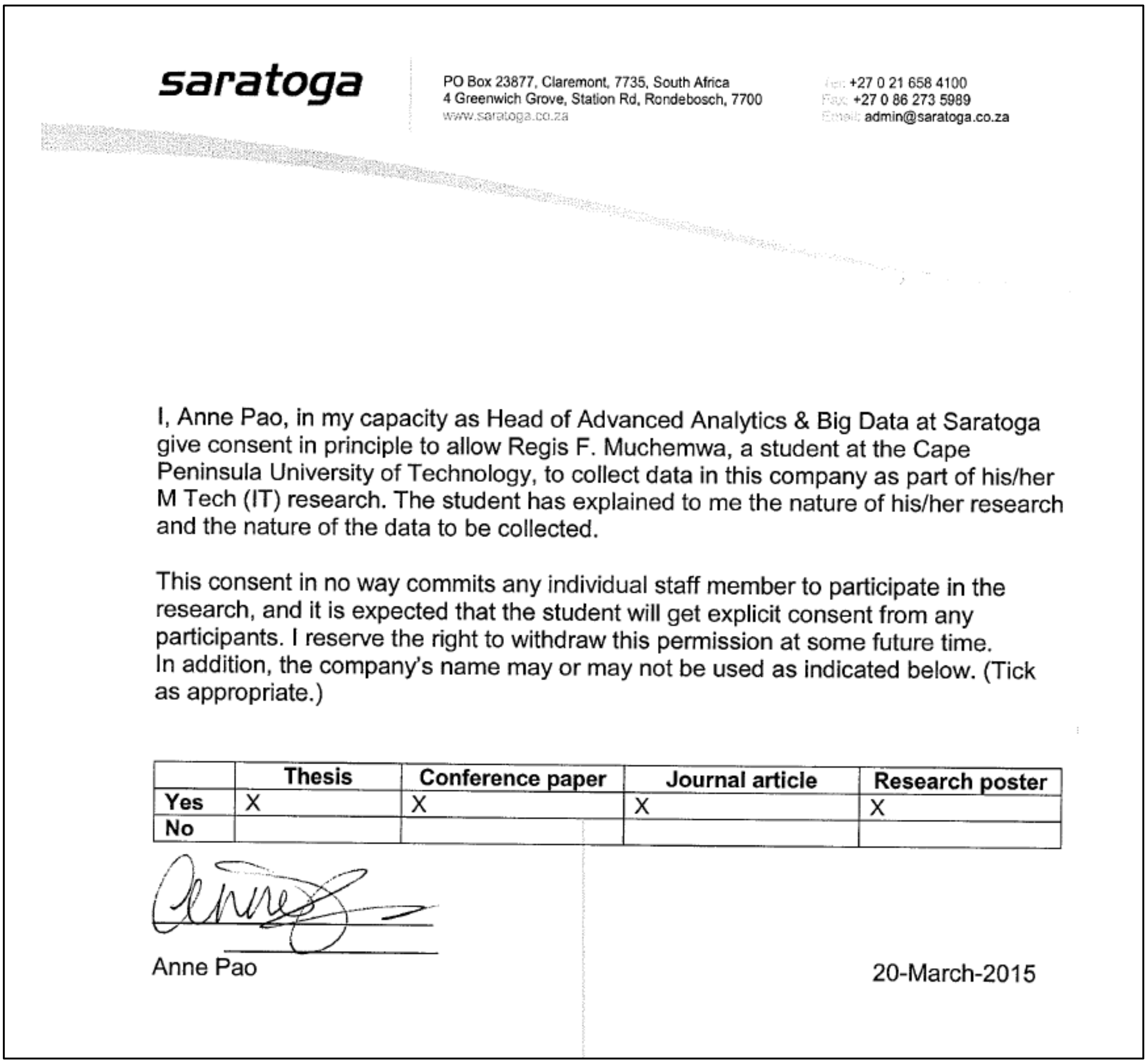
Question No.	Question	**Findings**
RSQ 1	***What are the factors that influence the selection of technologies appropriate for real-time DSSs in a BD environment?***	
SRQ 1.1	What is the relationship between characteristics of data and technologies used for analyzing data in a real-time environment?	**Finding 1**. Internet and on-line applications seem to influence organizations to use BD technologies such as Hadoop, NOSQL databases and In-memory databases. **Finding 2.** Although organizations are gathering data from social media platforms such as Facebook and Twitter, this data is not fully leveraged. **Finding 6:** There is a relationship between the volume of data generated by an organization and the selection of technologies used to analyse that data in real-time **Finding 7:** The format (structured or unstructured) of data available in an organisation has an influence on the technology that can be used for DSSs. **Finding 8:** The rate at which updates occur in the source systems and the need to reflect those changes to the reporting and analytics environment determines the type of technologies that are appropriate for real-time decision support. **Findings 9:** Organisations with high data volumes and using traditional BI technologies are facing scalability challenges.

		Finding 10: Technologies that can scale with volume of data and that can allow reporting to be in real-time are ideal for high volume data. **Finding 11:** Although data arrives and changes quickly within the systems for the participants’ organisations, none of the companies interviewed analyses data in real-time due to technology limitations. Two organisations analyse data in near real-time (one hour after data has arrived in their analytics platform) but their desire is to analyse the data in real-time. **Finding 12:** The rate at which data arrives and gets updated in a system influences the technologies used to analyse it in real-time. **Finding 13**: The format and structure of data influences the selection of technologies appropriate for real-time DSSs in a BD environment.
SRQ 1.2	What are the tools available on the market for use in analyzing BD?	**Finding 14:** Organisations are resorting to sharding databases in order to scale to the challenge of high volume of data and complexity of data structures. **Finding 15:** Scalability drives the selection of technologies appropriate for real-time DSSs in a BD environment. **Finding 16:** The need to discover unknown patterns in data has an influence on technologies used to analyse data. **Finding 17:** Ability to get answers to questions that are known in advance. **Finding 18**: The adoption of Hadoop by organisations in South Africa is still at its infancy stage. The few organisations that have Hadoop implementations are still in exploratory or transitional phases. **Finding 19:** Maturity of technology has an influence on the choice of technologies. **Finding 20:** Ability to integrate with existing technologies and systems.

		Finding 21: Ease of use of a technology has an influence on the selection of technologies appropriate. For instance, technologies which use a SQL variant to query data are preferred by users. **Finding 22:** Ability to handle large datasets in seconds by a technology has an influence on the technologies that are appropriate for real-time DSS in a BD environment.
RQ 2	How can an organisation evaluate technologies appropriate for real-time DSS in a BD environment?	
SRQ 2.1	What are the existing guidelines, frameworks, criteria, or measures applicable when evaluating analytic tools for real-time DSS in BD environments?	**Finding 23**: Business user requirements drive the choice of technologies used for real-time DSSs. **Finding 24:** Performance in terms of throughput and/or latency is the top criterion used to evaluate technologies appropriate for real-time DSSs. **Finding 25:** Ability of a new technology to integrate with existing technologies. **Finding 26**: The data collected revealed the following additional criteria that are important when determining technologies appropriate real-time DSSs in a BD environment. **Finding 27**: Costs of ownership. **Finding 28:** Licensing models. **Finding 29**: Resource network and hardware resource usage. **Finding 30**: Ability to scale with increase in data volumes **Finding 31:** Ability to handle system failure **Finding 32**: Availability of technical support.

		Finding 33: Technical skills availability can be considered as a criterion with which technologies can be compared but it should not be given high priority. **Finding 34:** The usage of resources such as CPU, RAM and disk input/output by a technology has an impact on the system's performance. **Finding 36:** The number of processors (CPU) required by technologies in order to give high system performance required by real-time DSSs can have an impact on licensing costs.
SRQ 2.2	How can an organization evaluate analytic tools appropriate for real-time DSS in a BD environment?	**Finding 37:** Organisations are relying on product vendor. **Finding 38:** Business user requirements drive the selection of technologies.

Appendix I: Permission Letter from company to carry out interviews



saratoga

PO Box 23877, Claremont, 7735, South Africa
4 Greenwich Grove, Station Rd, Rondebosch, 7700
www.saratoga.co.za

Tel: +27 0 21 658 4100
Fax: +27 0 86 273 5989
Email: admin@saratoga.co.za

I, Anne Pao, in my capacity as Head of Advanced Analytics & Big Data at Saratoga give consent in principle to allow Regis F. Muchemwa, a student at the Cape Peninsula University of Technology, to collect data in this company as part of his/her M Tech (IT) research. The student has explained to me the nature of his/her research and the nature of the data to be collected.

This consent in no way commits any individual staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.
In addition, the company's name may or may not be used as indicated below. (Tick as appropriate.)

	Thesis	Conference paper	Journal article	Research poster
Yes	X	X	X	X
No				

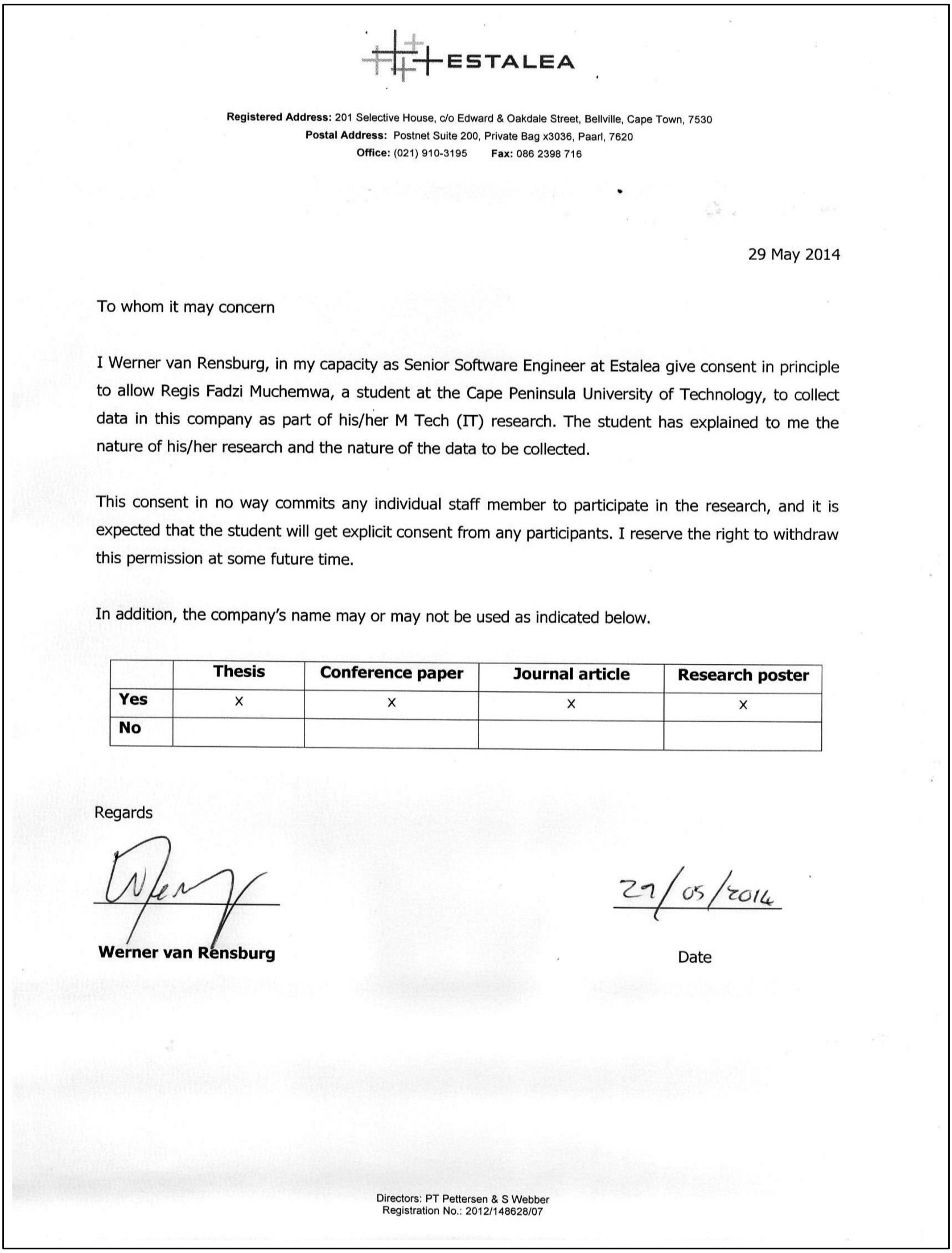
Anne Pao

20-March-2015

Appendix J: Other evaluation criteria

	Comments
Participants	
P1	Performance, technical support and costs, ability to process large volumes of data which changes quickly.
P2	To us, CPU utilisation is very important in our decision making because of the way we are charged by the technology vendors. For example, Greenplum is charged per CPU core, so CPU utilisation is important.
P3	Performance and ease of use and ability to handle complex data structures with ease.
P4	RAM is very cheap now and therefore the cost of memory is no longer critical to us. We would therefore rather have a system that crunches data in memory while answering queries in real-time than a slow system with low memory. A data warehouse separated from the analytics server may cause performance issues. Make sure the new technology architecture doesn't impact on other systems performance.
P5	Costs, resource usage, performance, Scalability and fault tolerance.
P6	Performance, technical support, costs, scalability, resource usage, and ease of use.
P7	Performance, costs, scalability, fault tolerance, resource usage, business user requirements,
P8	Performance, costs, scalability, fault tolerance, resource usage, business user requirements,
P9	Performance, ease of use and costs
P10	Performance, scalability. Costs and technical support

Appendix K: Consent letter from company to do interviews



ESTALEA

Registered Address: 201 Selective House, c/o Edward & Oakdale Street, Bellville, Cape Town, 7530
Postal Address: Postnet Suite 200, Private Bag x3036, Paarl, 7620
Office: (021) 910-3195 **Fax:** 086 2398 716

29 May 2014

To whom it may concern

I Werner van Rensburg, in my capacity as Senior Software Engineer at Estalea give consent in principle to allow Regis Fadzi Muchemwa, a student at the Cape Peninsula University of Technology, to collect data in this company as part of his/her M Tech (IT) research. The student has explained to me the nature of his/her research and the nature of the data to be collected.

This consent in no way commits any individual staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.

In addition, the company's name may or may not be used as indicated below.

	Thesis	**Conference paper**	**Journal article**	**Research poster**
Yes	x	x	x	x
No				

Regards

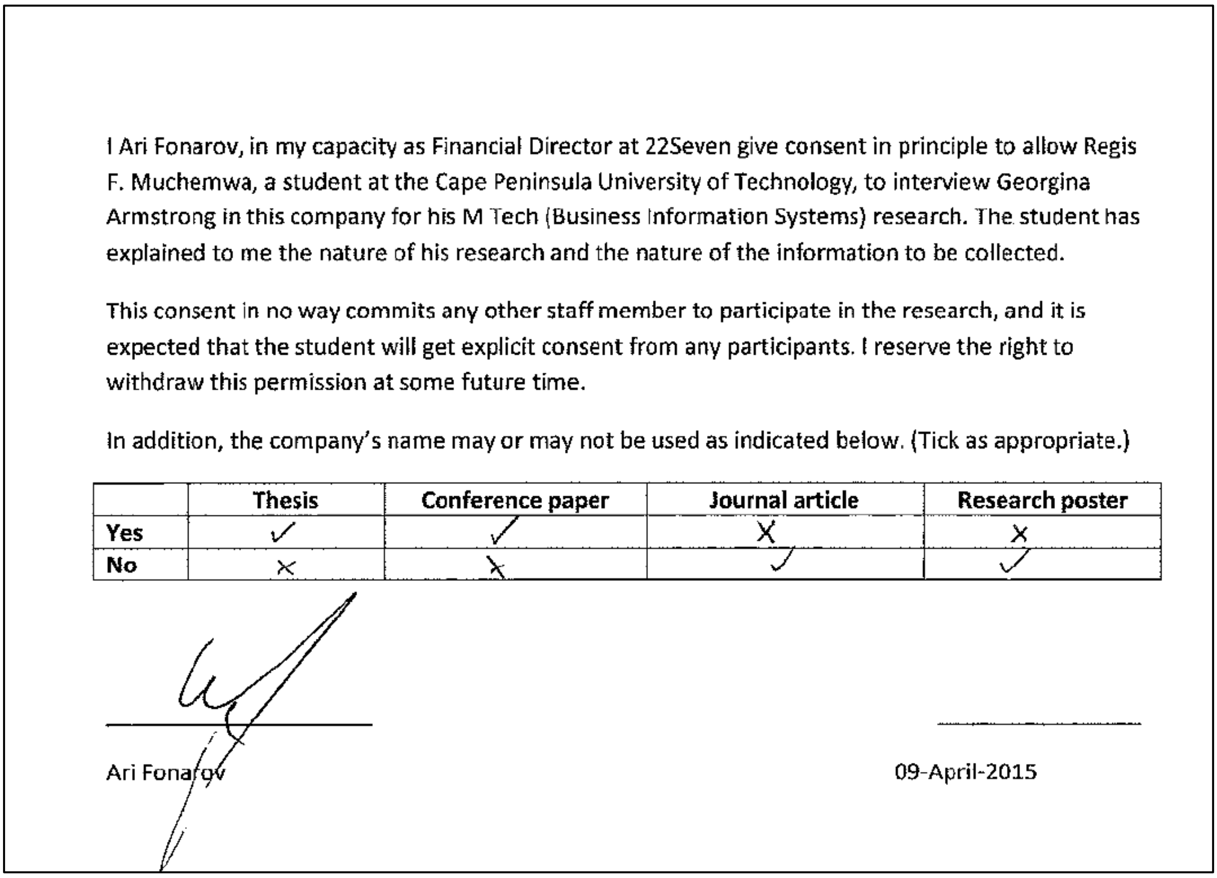
Werner van Rensburg

29/05/2014

Date

Directors: PT Pettersen & S Webber
Registration No.: 2012/148628/07

Appendix L: Permission Letter from company to carry out interviews



I Ari Fonarov, in my capacity as Financial Director at 22Seven give consent in principle to allow Regis F. Muchemwa, a student at the Cape Peninsula University of Technology, to interview Georgina Armstrong in this company for his M Tech (Business Information Systems) research. The student has explained to me the nature of his research and the nature of the information to be collected.

This consent in no way commits any other staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.

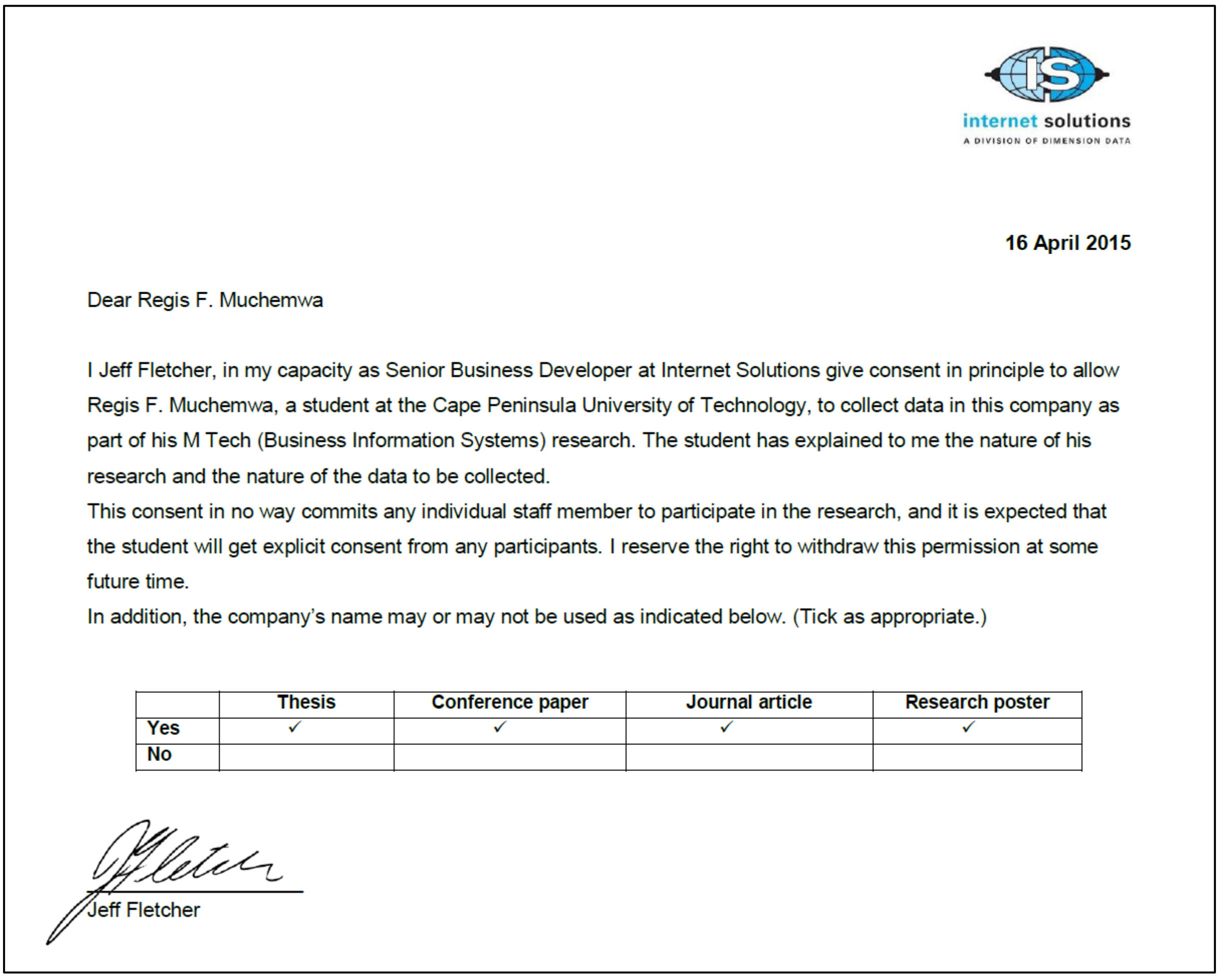
In addition, the company's name may or may not be used as indicated below. (Tick as appropriate.)

	Thesis	Conference paper	Journal article	Research poster
Yes	✓	✓	X	X
No	×	X	✓	✓

Ari Fonarov

09-April-2015

Appendix M: Permission Letter from company to carry out interviews



internet solutions
A DIVISION OF DIMENSION DATA

16 April 2015

Dear Regis F. Muchemwa

I Jeff Fletcher, in my capacity as Senior Business Developer at Internet Solutions give consent in principle to allow Regis F. Muchemwa, a student at the Cape Peninsula University of Technology, to collect data in this company as part of his M Tech (Business Information Systems) research. The student has explained to me the nature of his research and the nature of the data to be collected.

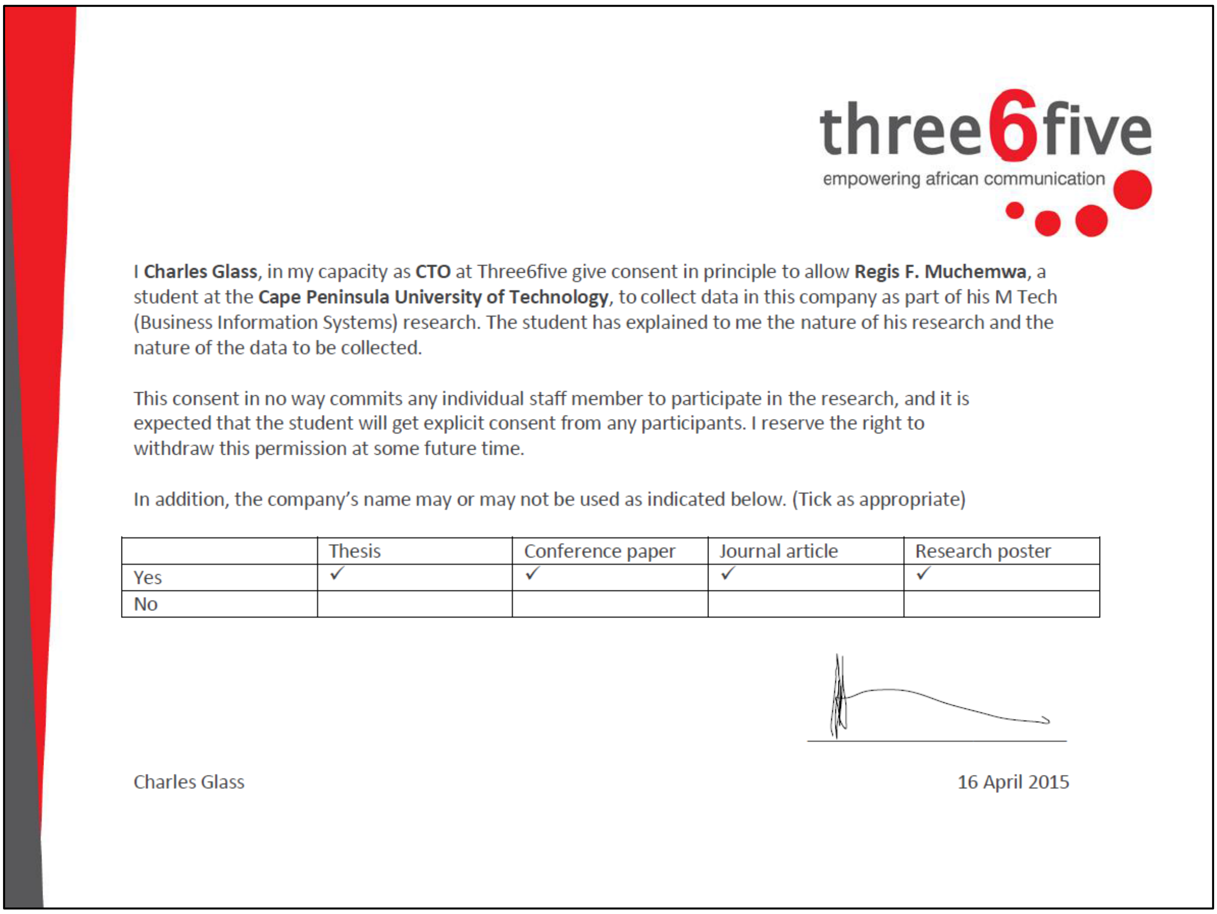
This consent in no way commits any individual staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.

In addition, the company's name may or may not be used as indicated below. (Tick as appropriate.)

	Thesis	**Conference paper**	**Journal article**	**Research poster**
Yes	✓	✓	✓	✓
No				

Jeff Fletcher

Appendix N: Permission Letter from company to carry out interviews



three6five
empowering african communication

I **Charles Glass**, in my capacity as **CTO** at Three6five give consent in principle to allow **Regis F. Muchemwa**, a student at the **Cape Peninsula University of Technology**, to collect data in this company as part of his M Tech (Business Information Systems) research. The student has explained to me the nature of his research and the nature of the data to be collected.

This consent in no way commits any individual staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.

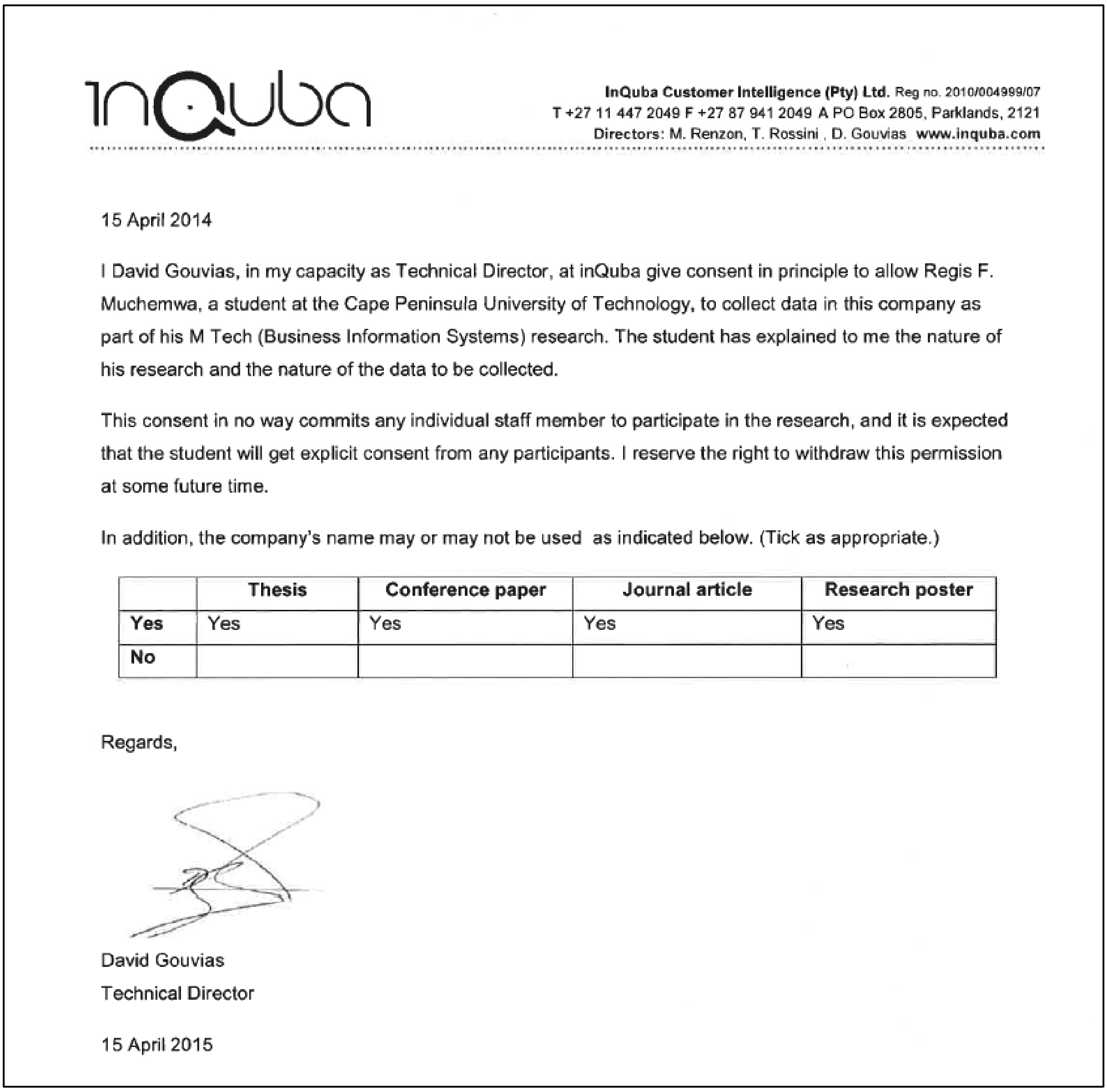
In addition, the company's name may or may not be used as indicated below. (Tick as appropriate)

	Thesis	Conference paper	Journal article	Research poster
Yes	✓	✓	✓	✓
No				

Charles Glass

16 April 2015

Appendix O: Permission Letter from company to carry out interviews



inQuba

InQuba Customer Intelligence (Pty) Ltd. Reg no. 2010/004999/07
T +27 11 447 2049 F +27 87 941 2049 A PO Box 2805, Parklands, 2121
Directors: M. Renzon, T. Rossini , D. Gouvias **www.inquba.com**

15 April 2014

I David Gouvias, in my capacity as Technical Director, at inQuba give consent in principle to allow Regis F. Muchemwa, a student at the Cape Peninsula University of Technology, to collect data in this company as part of his M Tech (Business Information Systems) research. The student has explained to me the nature of his research and the nature of the data to be collected.

This consent in no way commits any individual staff member to participate in the research, and it is expected that the student will get explicit consent from any participants. I reserve the right to withdraw this permission at some future time.

In addition, the company's name may or may not be used as indicated below. (Tick as appropriate.)

	Thesis	Conference paper	Journal article	Research poster
Yes	Yes	Yes	Yes	Yes
No				

Regards,

David Gouvias
Technical Director

15 April 2015

Appendix P: Answers to interview question 1

What are the sources of data in your organization used for decision making?

Participant	Answers	Comments
P1	Internet and On-line application systems, OLTP systems and Social media platforms e.g. Facebook	Our data includes customer related information collected from Facebook and Twitter. We use Text Analytics API to analyse sentiment on this data in real-time. So the main source of data is Internet and on-line applications as well as OLTP systems in our organisation.
P2	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	The company has a whole bunch of core routers and other networking devices. Each of those devices generates usage data which help us to know how much utilisation we have and how much capacity we have in our network at any given time.
P3	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	Data source are financial systems and Internet and On-Line application systems.
P4	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	99% combination of online and transactional and 1 % Online surveys information, obtained through call centre agents
P5	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	99% combination of online and transactional and 1 % Online surveys information, obtained through call centre agents
P6	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	Transactional data generated by on-line application systems. Internal platform that generates very large amounts of data which becomes the primary input into the data analytics environment
P7	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	Internal platform that generates very large amounts of data which becomes the primary input into the data analytics environment
P8	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors).	Internal platform that generates very large amounts of data which becomes the primary input into the data analytics environment
P9	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors), mobile network data traffic	We monitor network traffic and analyse metadata about on-line videos such as Youtube videos as they are being watched across the country or as they are being downloaded. Our goal is to analyse this data in real-time and provide our clients with reports as the activities are happening for corrective action or for improving services.

P10	OLTP systems, Internet and On-line application systems, digital systems (e.g. sensors), social media platforms (e.g. Facebook), digital systems, audio voice recording systems and CCTV	Mainly Transaction database systems. There is also on-line and internet data sources which are not being analysed. There is also data from the call centres in the form of voice recorded messages. In the company's stores, there are massive amounts of CCTV data which are not being analysed to identify trends or patterns that can help

Appendix Q: Answers to Interview Question 2

How would you describe the data available in your organization used for decision making?

Participants	Answers	Comments
P1	High Volume, quickly changes, different sources, both structured and unstructured in nature.	
P2	High volume, quickly changes, structured and unstructured, comes from different sources	Our data comes from different places. We always know the format and structure of the data that we analyse in advance. We are an Internet service provider and we have other people's data moving across our network but we have kind of metadata about that data and we use that for our decision making. We need to know how much data is being moved but we don't know anything specific to the data moving in our network. Our main challenge with this data is that it comes in very high volume and is difficult for us to analyse in real-time with the technologies we currently have. Data is High volume, quickly changes, comes from different data sources and is generated in structured and unstructured form but only structured data is analysed.
P3	Quickly changes and originates from two sources. Data is low volume.	
P4	High volume, quickly changes, multiple sources, both structured and unstructured.	High volume, quickly arrives but less updates, generates data in multiple places but only a single source is analysed. Also have unstructured data which is never analysed by the system. The other challenge we have with the structured data is that it is massively growing in volume. We need technologies that are scalable as the volume of both structured and unstructured data increases.

P5	High volume, quickly arrives but less updates, multiple sources, unstructured data and structured	High volume, quickly arrives but less updates, generates data in multiple places but only a single source is analysed. Also have unstructured data which is never analysed by the system. The other challenge we have with the structured data is that it is massively growing in volume. We need technologies that are scalable as the volume of both structured and unstructured data increases.
P6	Data quickly arrives, but less changes, high volume, comes from different sources, both structured and unstructured	Data is added into the system every second. There are few changes to the data once it has landed into the systems but it comes in very fast. A lot of data comes in very quick and so we do a lot of inserts but few updates. Data is characterised by high volume, quick insertions and comes from different sources. Data is both structured and unstructured but analysis is restricted to structured data only.
P7	High volume, comes from different sources, structured and unstructured, quickly changes.	Data is characterised by high volume, quick insertions and comes from different sources. Data is both structured and unstructured but analysis is restricted to structured data only.
P8	High volume, quickly changes, both structured and unstructured, analysis restricted to structured data	Data is characterised by high volume, quick insertions and comes from different sources. Data is both structured and unstructured but analysis is restricted to structured data only.
P9	High volume, high velocity, both structured and unstructured comes from different sources	Data is characterised by high volume, high velocity and is both structured and unstructured. Meta data about videos and other file formats are analysed,
P10	Quickly changes in structure, high volume, structured and unstructured, comes from different sources.	We need technologies which can allow any change in data type or structure to be reflected immediately in our analytics and reporting environment. This is not possible with existing technologies because of a long development cycle for the ETL process.

Appendix R: Answers to Interview Question 3

How would you describe the volume of data available for decision making in your organisation?

	Answers and Comments
Participant	
P1	Medium – terabytes
P2	Medium in terms of what we keep. A lot of the information we don't keep, we drop we generate zeta bytes of data but we don't use that for decision making.
P3	Low – less than 500Gb but data structure is complex in terms of what we analyse. We don't analyse unstructured data that we generate.
P4	Volume – Between low and medium. 80% low and 20% medium. Telecoms. Network traffic high volume and quickly changes.
P5	Volume – Between low and medium. 80% low and 20% medium. Telecoms. Network traffic high volume and quickly changes.
P6	We not yet in the petabytes region but we are expecting to go there very soon. We are looking at about 2,6 billion transactions per month. We are expecting this to grow and that's why we are looking to move to a different platform. In 6 months we have moved from 1.2billion transactions per month and that's why we want to cater for this growth.
P7	We not yet in the petabytes region but we are expecting to go there very soon. We are looking at about 2.6 billion transactions per month. We are expecting this to grow and that's why we are looking to move to a different platform. In 6 months we have moved from 1.2 billion transactions per month and that's why we want to cater for this growth.
P8	We not yet in the petabytes region but we are expecting to go there very soon. We are looking at about 2,6 billion transactions per month. We are expecting this to grow and that's why we are looking to move to a different platform. In 6 months we have moved from 1.2billion transactions per month and that's why we want to cater for this growth.
P9	Data is in the petabytes region.
P10	We have seen a sharp increase in the volume of data generated when we bought a new IBM campaign management system. Every day, the system sends out millions of marketing email and SMS messages to clients and prospects. The system also receives messages back from the targeted people. We have high volume of data being generated but we are struggling to analyse this data especially in real-time as we have to wait for the ETL process to run over night. Data being analysed is within the terabytes range but video (CCTV) and voice data (call centre)

Appendix S: Answers to Interview Question 4

How would you describe the rate at which data changes within your systems for analytic purposes?

	Comments
Participants	
P1	Data changes per second. We use big data technologies to assist us with managing the velocity of the data coming in. and the real-time management of that data
P2	The data arrives into our systems in real-time but we base our analytics per day. We aggregate and look at it per day. Some data is analysed per hour except for system logs, which are obviously per second, but those are just for platform performance monitoring. The data changes quickly but is analysed per day.
P3	Data is analysed Per hour - except our system logs, which are obviously per second, but those are just for platform performance monitoring.
P4	The data changes quickly but is analysed per day. Data warehouse.
P5	The data changes quickly but is analysed per day. Data warehouse.
P6	We receive second new data arrives into the environment through clicks on the on-line adverts.
P7	Data arrives by second or less but we do not yet have the capability to analyse it in real-time. Our goal the coming year is to produce user reports within a few hours and eventually in real-time but at the moment we are only doing so through nightly batch process.
P8	The data arrives into the analytics environment at a very high rate.
P9	There data is changing on a per second basis.
P10	All data is currently being analysed per day but the frequency of changing includes per hour, per day and per month.

Appendix T: Answers to Interview Question 5

What type of data formats/structures exist in your company?

	Comments
Participants	
P1	Flat files, Videos and Audio files and relational databases
P2	All text. Log files. We know the format in advance. Some of the data is stored in relational databases. Videos and emails we don’t analyse. Have Facebook page but we don’t analyse,
P3	Text files but very complicated
P4	Flat files, Videos and Audio files but analyse only flat files
P5	Text files, videos, email and audio but analyse text only.
P6	The files are generated by the application but analytics is happening on the warehouse
P7	The files are generated by the application but analytics is happening on the warehouse
P8	The files are generated by the application but analytics is happening on the warehouse
P9	This includes how long the video was watched; the time it was watched, location in which the video was downloaded and how much bandwidth was consumed.
P10	Mainly structured flat files and relational databases.

Appendix U: Answers to Interview Question 6

What do you currently use to analyse your data?

	Comments
Participants	
P1	We use NOSQL databases because data like Facebook and Twitter continually change their data structures which are usually in JSON file format. So change in data structure is a big thing in BD analytics. NOSQL databases can handle changes in data structure more easily, as you don't have to define the structure upfront. Structure is implicit in the JSON structure. You can query the items even if the structure changes dramatically. Another problem that NOSQL databases solve is performance with large complex data because you don't have to join tables as you don’t have to normalise your data. NOSQL databases are superfast at summarizing and querying the data.
P2	We currently use traditional database and data warehouse technologies namely Sybase, Oracle and Greenplum. We also use Round Robin Database (RRD) which is a time series database. We have a plan to use Hadoop but we are not there yet. However, we sell Cloudera Hadoop to our clients.
P3	We use a sharded MongoDB implementation, which we have mapped to Amazon Redshift so that we can analyse the data. We found that MongoDB was terrible for data analysis as queries were complex and slow. Even pretty basic queries would impact on customer experience so we upload deltas to Amazon Redshift every hour in order to improve system performance.
P4	Traditional databases, In-memory databases. Have started testing Cloudera Hadoop and using Hadoop based solution Pivotal.
P5	When we get to a point where we don’t know what we are expecting, then we will move more towards BD infrastructure. At present, we will continue to use traditional database and data warehouse technologies because we know in advance what we are looking for in our data. We use GreenPlum and Pivotal HDS.
P6	Traditional database and DW solutions. Have also started using Cloudera Hadoop and migrating our DSS onto this new platform.

P7	We are an open source shop. We have sharded Mysql databases and an ETL tool. Then we have a reporting application called Infobrite but we now want to get out of the Mysql space and out of Infobrite into Hadoop. We want to replace our ETL tools simply because we are running into scalability problems. You cannot linearly add more machines and you cannot scale well as volumes of data increase. You don't expect your performance to double up after doubling up your machines. But on the Hadoop space you can scale with no problems. We have been using sharded databases but it's proving to be more expensive. We are also moving to Impala for reporting which runs off Hadoop. Our star schema is now on the impala platform.
P8	
P9	Network Analytics - from Sandvine. They had a product from the same company so it was by choice to get a product from the same company. The existing supplier had an influence on the selection.
P10	We have started talking about big data technologies to help with analytics of both structured and unstructured data. At present we make if SAS, an ETL tool called Datastage and Oracle databases.

Appendix V: Answers to Interview Question 7

If you are using Hadoop, which distribution do you use?

	Comments
Participants	
P1	Hadoop and NOSQL databases are not yet fully mature. Very few organisations are using Hadoop especially in South Africa. Cloudera is popular and very easy to setup and configure.
P2	Cloudera Hadoop
P3	No Hadoop
P4	Pivotal HDS and Greenplum which is like relational with a backend of Hadoop. Maturity of analytics on Hadoop is still very low. There is a very small population of users of Hadoop in South Africa at the moment.

P5	Pivotal HDS and Greenplum which is like relational with a backend of Hadoop. Maturity of analytics on Hadoop is still very low. There is a very small population of users of Hadoop in South Africa at the moment.
P6	Cloudera Hadoop
P7	Cloudera Hadoop
P8	Cloudera Hadoop
P9	No Hadoop
P10	No Hadoop

Appendix W: Answers to Interview Question 8

If you are using Hadoop, what analytic components do you use?

	Comments
Participants	
P1	We use Hive because it uses a SQL variant to get data out of Hadoop. We also use SAS, R, Tibco, Spotfire and Zoom Data which is very good at handling large data sets in seconds. With ZoomData one is able to visualize 1 billion rows, with additional 1m per second in real time. In some cases we actually transfer the Hadoop data to MongoDB and I use ZoomData to analyse data. We also use Impala and Hive to analyse some data that is sitting on Hadoop. In addition to that we also use SAS, Tableau and R to analyse data that we query from Redshift. Finally, we also use Microstrategy for reporting and analytics.
P2	We use Impala and Hive because they are the default on Cloudera and they are starting points for most people.
P3	We use Tableau and R to analyse data that we query from Redshift.
P4	Microstrategy – reporting and analytics. We have also started using Hive
P5	Microstrategy – reporting and analytics.
P6	
P7	Infobrite, Impala and Hive

P8	Infobrite, Impala and Hive
P9	Sandvine
P10	We don’t use Hadoop but for reporting we use Oracle OBIEE.

Appendix X: Answers to Interview Question 9

When selecting analytic tools for real-time decision making in a BD environment what criteria did you consider?

	Comments
Participants	
P1	Performance, availability of technical support, costs, scalability, resource usage and fault tolerance.
P2	When we consider costs, we look at TCO (total cost of ownership or total cost of the infrastructure over the time period). We look at how much it is going to costs us for hardware, how much software will cost and how much maintenance will cost. We consider all these things and then we compare three or four different technologies. So we look at it over the lifetime of the service and not just one particular individual item. So our decision is based on TCO. Some products are licenced per user, some per node and some per CPU core.
P3	Performance, resource usage, scalability, ability to process high volume data, ability to handle quick changing data and costs.
P4	Nowadays, memory is very cheap. The hardware doesn’t drive requirements but performance influences the choice of technology used for analysing BD in real-time. How fast do we need to load data is what drives requirements. To us scalability and fault tolerance is also critical as our data volumes are continuously increasing, for example, adding RAM to the environment should be easy. In a data system environment, the system can grow in the number of CPU cores, the number of users or number of nodes. The moment you add a new node or a new user, it means costs will go up. One needs to understand the type of licencing mode because some technologies are licenced per user, some per node and some per CPU core.

P5	Corporates are still cowed in legacy systems and one must consider how new technology will be integrated with existing technologies and other systems. There is need for continuity and you need to be able to connect existing data sources with the new technologies.
P6	
P7	When switching, make sure that your reporting remains consistent. We must ensure that we don’t have a single point of failure.
P8	
P9	My company had a product from the same supplier which was going to be easy to integrate with the new system for data analytics. So the existing supplier and technology had an influence on the technology that was eventually selected.
P10	

Appendix Y: Answers to Interview Question 10

Which Industry is your company and what sort of insights do you expect from your data?

	Comments
Participants	
P1	Our core business is customer experience management. Our main focus is to build products for customer experience, and allow real-time monitoring of customer experience. From our data, we expect to identify risky customers, identify areas of weakness and strength in the business and then decide which intervention programmes are necessary. To meet this requirement, it needs real-time access to operational data, just as it arrives into our systems as things are happening.

P2	That’s question we are trying to answer. The BD infrastructure doesn’t automatically create insight but the infrastructure actually creates costs. So from our perspective, we are actually trying to answer that question. We were replacing some old equipment, so we have stuff that provide us with insights, something that we have always known like how long does it take us to close a call and what is the utilisation of different parts of our network. To start looking at insights, there are two things that someone has to consider. Are they trying to answer a specific question that is not currently being answered. Your BI should answer certain questions but is there a business question not being answered. The second part is trying to predict something. That’s where most people are looking at big data. Given that we have this huge amounts of data, can we take it and predict something about the future of our business to say, are we losing out to our competitor or is one of our products no longer going to be bought or could we predict that a customer will leave us. Those I think are the insights that people are expecting to get out of big data but the complexity of building a predictive algorithm is quite high.
P3	
P4	We considered requirements from the users and then checked what was available on the market to do the job required. We then compiled a matrix of all important things and weighted them. We then looked for vendors for the top three products and requested them to present their products. Each product had scores and in the end, the product that scored higher than the others was selected.
P5	
P6	
P7	
P8	
P9	One issue I have been investigating often is people want to know the reason certain subscribers are getting poor network quality. The problem is that you cannot tell as there are so many factors that influence this, for example, some people will upload multiple videos in multiple driver stations. They might be doing something else and at the same time, watching a video and still will tell you they are having a bad network experience but you might not really know if they are watching a video. So there are all different kinds of opportunities to measure things you couldn’t measure before BD technologies came on board.

P10	The company interfaces with its clients through multiple channels such the company website, Facebook, twitter, email, SMS and through telephone calls. These channels generate massive amounts of data but only a small percentage is being leveraged. Ideally, we would want to have technology that can analyse all data that is available within the business as quickly as possible and make quick operational business decisions to our competitive advantage. We would want to respond to customer sentiments but at the moment we are unable to do so as we do not have technologies that can assist us in that.

Appendix Z: Answers to Interview Question 11

How can an organization evaluate analytic tools appropriate for real-time DSS in a BD environment?

	Comments
Participants	
P1	
P2	To do the evaluation we ran and compared three different database technologies. We took a sample of our dataset, uploaded it into the databases and then looked at the time it took to execute a query. Finally, we compared the amounts of time taken to execute the query by each of the products and based our decision on this. To us, performance was very critical. So we used an internal benchmark and not an existing benchmark.
P3	I was involved with the selection of Tableau and R which we use for analysing data. In the selection process, we used the recent Gartner Magic Quadrant report on Tableau and we also considered recommendations given by data analysts sharing their work on-line through platforms such as GitHub. We seriously consider what these people recommend and therefore much of our decision to go with Tableau (and with R) was based on their recommendations.

P4	We considered requirements from the users and then checked what was available on the market to do the job required. We then compiled a matrix of all important things and weighted them. We then looked for vendors of the top three products and requested them to present their products. Each product had scores and in the end, the product that scored higher than the others was selected.
P5	
P6	we ran internal benchmarks to compare different products
P7	
P8	
P9	
P10	Conducted proof of concept before we made a choice.,