Big Data Adoption Challenges and Benefits: The Case of a South African Bank

Johannes Chabane Matsimbi

A research report submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business

Johannesburg, 2020

(Version 2019-2020)

ABSTRACT

The origins of Big Data have not been confirmed; however, it is suspected that it originated in the mid-1990s. Since then, the adoption of Big Data has gained momentum in many industries. This study sought to establish how data is collected and highlight the challenges that are experienced during Big Data adoption. The research also sought to establish how the adoption of Big Data can benefit organisations.

A qualitative study was conducted, consisting of a sample of 15 participants from the South Africa bank's Information Technology (IT) Department. Semi-structured interviews were conducted to collect the data. The data was analysed through the qualitative tool NVivo, to assist with the process of coming up with codes and patterns. Then these patterns were used to create themes to assist with analysis.

The study revealed that the bank uses industry-appropriate tools that are also suitable for collecting and curating Big Data. The study further showed that skills scarcity could be a significant challenge when implementing Big Data projects. The architecture of the organisation, governance around how projects are implemented, as well as common understanding regarding how the data should be packaged, can be a challenge when implementing Big Data projects. Despite the challenges presented by Big Data, the study also found that there are benefits for Big Data adoption. The benefits for Big Data include, amongst others, enhancement of customer-centric marketing, data-driven product development and data monetisation, to mention, but a few.

The study added to the body of knowledge (BoK) by demonstrating that the benefits of Big Data adoption outweigh the challenges presented by Big Data. The study suggests that more investments should be put in place for change management when implementing projects of this magnitude. Change management would assist in monitoring the impact of the projects on how people do their jobs and provide clarity on the required changes regarding how systems will work, such as process optimisation. The knowledge gained from this study provides valuable insights for researchers, bank leaders and Big Data practitioners about the importance of adopting Big Data and the benefits derived from doing so.

KEY WORDS: Big Data, Big Data Analytics, Big Data Tools, Challenges of Big Data, Benefits of Big Data, Predictive Data, Prescriptive Data, Data Monetisation, Financial Sector, Banking, South African Environment

DECLARATION

I, Johannes Matsimbi, declare that this research report is my own work except as indicated in the references and acknowledgements. It is submitted in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

Name: Signature:

Signed at ……

On the …….. day of …… 20…..

DEDICATION

I’d like to dedicate this to my dad, William Chabane Matsimbi. Thank you for your continuous guidance and support. I wouldn’t be half the man I am today if it wasn’t for your contribution. Na khensa papa.

ACKNOWLEDGEMENTS

I would like to acknowledge my wife and children for their patience through this process. I also would like to thank my supervisor for the continuous support and constant direction that was provided. Finally, I would like to thank my employer for their support and for providing me with the opportunity to pursue this academic qualification.

TABLE OF CONTENTS

ABSTRACT ..... **ii**

DECLARATION ..... **iv**

DEDICATION ..... **v**

ACKNOWLEDGEMENTS ..... **vi**

TABLE OF CONTENTS ..... **vii**

LIST OF TABLES ..... **x**

LIST OF FIGURES ..... **xi**

LIST OF ACRONYMS AND ABBREVIATIONS ..... **xii**

CHAPTER 1. INTRODUCTION **13**

1.1 PURPOSE OF THE STUDY 13
1.2 CONTEXT OF THE STUDY 13
1.3 RESEARCH PROBLEM BACKGROUND 15
1.4 RESEARCH OBJECTIVES AND QUESTIONS 16
1.4.1 RESEARCH OBJECTIVES: 16
1.4.2 RESEARCH QUESTIONS: 16
1.5 SIGNIFICANCE OF THE STUDY 17
1.6 DELIMITATIONS OF THE STUDY 17
1.7 DEFINITION OF TERMS: 18
1.8 ASSUMPTIONS 19
1.9 CHAPTER OVERVIEW 19

CHAPTER 2. LITERATURE REVIEW **21**

2.1 INTRODUCTION 21
2.2 THEORETICAL REVIEW 21
2.2.1 BIG DATA BACKGROUND 21
2.2.2 DEFINITION OF BIG DATA 22
2.2.3 DEFINITION OF BIG DATA ANALYTICS 23
2.3 PRIOR STUDIES ON BIG DATA COLLECTION TOOLS 24
2.4 PRIOR STUDIES ON BIG DATA CHALLENGES 26
2.4.1 PROPOSITION 1 30
2.5 PRIOR STUDIES ON BIG DATA BUSINESS BENEFITS 30

2.5.1 PROPOSITION 2 31
2.6 CONCLUSION OF THE LITERATURE REVIEW 32

CHAPTER 3. RESEARCH METHODOLOGY 33

3.1 RESEARCH APPROACH 33
3.2 RESEARCH DESIGN 34
3.3 DATA COLLECTION METHODS AND PROCEDURE 34
3.4 POPULATION AND SAMPLE 35
3.4.1 POPULATION 35
3.4.2 SAMPLE AND SAMPLING METHOD 35
3.5 THE RESEARCH INSTRUMENTS 36
3.6 DATA ANALYSIS AND INTERPRETATION 36
3.7 TRANSFERABILITY AND DEPENDABILITY 39
3.7.1 TRANSFERABILITY 39
3.7.2 CREDIBILITY 40
3.7.3 DEPENDABILITY 40
3.8 ETHICAL CONSIDERATIONS 41

CHAPTER 4. PRESENTATION OF FINDINGS 42

4.1 INTRODUCTION 42
4.1.1 ORGANISATION BACKGROUND 42
4.1.2 PROFILE OF RESPONDENTS 43
4.2 RESULTS PERTAINING TO PROPOSITION 1: 50
4.2.1 BECOMING FAMILIAR WITH THE DATA 50
4.2.2 INITIAL CODE GENERATION: 51
4.2.3 SEARCH AND REVIEW THEMES 53
4.2.4 DEFINE THEMES – THEMATIC ANALYSIS 55
4.2.5 REVIEWING OF THEMES 55
4.2.6 THEME WRITE-UP AND PRESENTATION: 57
4.3 RESULTS PERTAINING TO PROPOSITION 2: 60
4.3.1 BECOMING FAMILIAR WITH THE DATA 60
4.3.2 LOADING OF TRANSCRIPTS INTO AN ANALYSIS TOOL 61
4.3.3 INITIAL CODE GENERATION: 61
4.3.4 SEARCH FOR THEMES 61
4.3.5 THEMATIC ANALYSIS 63

CHAPTER 5. INTERPRETATION OF FINDINGS 67

5.1 INTERPRETATION AND DISCUSSION OF RESULTS FOR PROPOSITION 1 67
5.1.1 THEME 1: ACCESS MANAGEMENT AND GOVERNANCE 68
5.1.2 THEME 2: DATA CURATION PROCESS AND PROCESS CHALLENGES 68
5.2 INTERPRETATION AND DISCUSSION OF RESULTS FOR PROPOSITION 2 70
5.2.1 THEME 1: DATA COLLECTION AND DATA CURATION 70
5.2.2 THEME 2: BIG DATA BENEFITS 71

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS 74

6.1 CONCLUSION FOR RESEARCH QUESTION 1: 75
6.2 CONCLUSION FOR RESEARCH QUESTION 2: 76
6.3 CONCLUSION FOR RESEARCH QUESTION 3: 76
6.4 RESEARCH PROBLEM ADDRESSED 77
6.5 RESEARCH OBJECTIVES AND QUESTIONS ANALYSED 78
6.6 RECOMMENDATIONS 78
6.7 SUGGESTION FOR FURTHER RESEARCH 79

REFERENCES ..... Error! Bookmark not defined.

APPENDIX A: COMPANY RESEARCH PERMISSION LETTER ..... Error! Bookmark not defined.

APPENDIX B: INSTRUMENT..... Error! Bookmark not defined.

APPENDIX C: UNIVERSITY APPROVAL OF TITLE ..... Error! Bookmark not defined.

APPENDIX D: EDITOR CERTIFICATE Error! Bookmark not defined.

LIST OF TABLES

Table 1: Profile of respondents by position 36

Table 2: Sample of initial codes 52

Table 3: Category to code mapping 53

Table 4: Merged category 54

Table 5: Themes for Proposition 1 56

Table 6: Codes for Proposition 2 61

Table 7: Categories for Proposition 2 62

Table 8: Research propositions with emerging themes 67

LIST OF FIGURES

Figure 1: A streamlined Codes to Theme Model for a qualitative inquiry adapted from Arcury & Quandt, 1999..... 39

Figure 2: Consolidated view of the codes 49

Figure 3: Thematic Map for Proposition 1 55

Figure 4: Thematic Map for Proposition 2 63

LIST OF ACRONYMS AND ABBREVIATIONS

Term	Description
ADV	Advanced Data Visualisation
BI	Business Intelligence
BoK	Body of Knowledge
CIB	Corporate and Investment Banking
CIO	Chief Information Officer
DBA	Database Administrator
ETL	Extract Transport and Load
FAIS	Financial Advisory and Intermediary Services
IT	Information Technology
PBB	Personal and Business Banking
RR	Research Respondent
STTM	Source to Target Mapping
SVOT	Single Version of Truth
TDWI	The Data Warehousing Institute

CHAPTER 1. INTRODUCTION

1.1 Purpose of the study

This study aims to reduce the confusion surrounding the myth of Big Data and to demystify the benefits of Big Data implementation and adoption. In this study, the Big Data benefits and challenges such as identifying appropriate tools and the method used for data collection are presented. The present study focuses on one of the four major banks in South Africa. Apart from the importance of providing context in the developing world, this study focuses on the financial services industry, an industry that is arguably overwhelmed with Big Data, but under studied, particularly the banking sector.

1.2 Context of the study

South Africa has a very mature financial sector; however, this sector has been predominantly perceived as the least innovative, and most of the products are prescribed to the customer (PWC, 2017). Incorporating Big Data analysis into the industry will enhance the current operational model for the banking industry. This statement is supported by the study conducted by Chen, Chiang, and Storey (2012), which identified Business Intelligence (BI) as a benefit of Big Data analytics. The benefits also include business insights, understanding change and additional details regarding business cost drivers. In addition, banks will be able to identify which products are relevant for their clients and what products they can decommission. McAfee, Brynjolfsson, Davenport, Patil and Barton (2012) argue that an organisation cannot manage what they cannot measure, which suggests the need for banking institutions to have effective analytical tools to manage their operations.

The way people consume services has changed over the years (Fischer-Baum, 2017). This finding was outlined in an article published by the Washington Post in 2017, which aimed to review the change in technology over the last 30 years. The rise in technology has increased accessibility for cell phones and laptops. The prices of these devices have also dropped drastically over the years (PWC,

2017). Internet access has increased in South Africa over the years, and this has contributed to increased access to information.

An article published by World Wide Worx (2017) indicated that internet access in South Africa had reached 21 million people. This trend of expanding information has created a culture of efficiency and expeditiousness as customers expect nothing less from their service providers (PWC, 2017). The trend further influences the operational processes in companies due to the change in consumer behaviour.

The state of being informed changes consumer market behaviour from a demand perspective as consumers become aware of the product options that are available and whether those products meet their needs. The change in demand changes the supply as well, which, in turn, has resulted in a change in the organisation's operations (Clemons, 2008). Companies must measure this change to manage their business, as argued by MacAfee et al. (2012).

The characteristics of Big Data include Volume, Velocity, Variety and Veracity (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). Volume is defined as the amount of data that flows online, or that is stored in databases. Big Data gives companies access to high volumes of data to work from, and this allows access to more data than they had access to before. Velocity is defined as the speed at which the data flows: the faster the information flows, the better, as real-time information gives a more competitive advantage to companies: the data will be in its more recent form. Variety is defined as the different formats that data can be captured in, which varies from text, images, video, GPS coordinates, social network data and information from actuators. Veracity is defined as the validity and accuracy of data that has been collected.

The context of this study was to analyse the benefits of Big Data for the banking sector while looking at the challenges posed. The study also evaluated what the Big Data characteristics (Volume, Velocity, Variety and Veracity) mean to the bank. The evaluation critically analysed if the bank needs to enhance its tools to manage the increasing volumes. This included vetting if the existing tools are well equipped to deal with velocity and lastly if the bank has appropriate tools to deal with variety and veracity to enable the bank to achieve the perceived benefits of

Big Data. The study followed a case study research design conducted in a South African bank to add a South African context to the available literature.

1.3 Research problem background

Big Data is enabling companies to process and analyse information fast, and this enables them to meet customer expectations. According to a study conducted by Russom and Phillip (2011), quick processing of information can assist with quick decision making. Quick decision making and efficiency are critical, as indicated in the study conducted by PWC (2017). MacAfee et al. (2012) conducted a study to show that the support of executives has an impact on Big Data implementations. The study found that data-driven companies perform better as they can conduct informed product mapping and creation. This is only made possible when the implementation of Big Data projects is supported by executives.

The PWC (2017) study indicates that the South African banking sector has not yet advanced to the level of fully incorporating Big Data analytics, even though they have started incorporating customer-centricity in their strategy. McAfee et al. (2012) argue that the correct transformation of implementing Big Data projects will give businesses a competitive advantage. Organisations will be able to leverage insights into their decision making from customers, which suggests the need for banks to implement Big Data projects effectively to gain competitive advantage. The findings of a more recent literature review conducted by Pejić Bach, Krstić, Seljan, and Turulja (2019) indicate that many companies have large volumes of data that can be exploited and transformed into knowledge to make informed decisions.

Although Big Data has been known for some time, there is a good deal of uncertainty amongst organisations about the correct use of Big Data (Kitchin, 2013). There is a lack of skills, a lack of understanding of the tools to use and a lack of understanding of precisely what benefits will result. The benefits need to be linked to the strategic intent of the company (McAfee et al., 2012). Financial institutions have not included Big Data Scientist as part of their normal operational model. Because of this, the adoption of Big Data has been slow. The lack of these

skills has an impact on the growth and efficiency of these financial institutions as indicated by market analysis. The Big Data adoption approach can provide the necessary support for the growth and efficiency of the organisation (Trelewicz, 2017).

Given these findings, there appears to be a gap in assessing whether the same challenges and perceived benefits could be applied in banking. As far as could be established in this research, there is no literature available that analyses Big Data adoption in South Africa primarily focusing on the tools used to collect Big Data for banks and the challenges and benefits of Big Data adoption.

1.4 Research objectives and questions

The primary purpose of the study was to critically analyse the adoption of Big Data focusing on the tools used for the collection and processing of the data and the challenges and benefits of Big Data in South African banking.

1.4.1 *Research objectives:*

The objectives of the study were to:

- Assess how the bank in the case study collects and processes Big Data and the tools they use for this.
- Analyse the challenges associated with the collection and processing of the Big Data.
- Assess Organisation A's intended benefit of Big Data collection and ascertain if this links with their data monetisation activities on the ground.

1.4.2 *Research questions:*

The research questions, following the objectives of the study, were as follows:

- Which tools can be used for Big Data in the bank in this study?
- What challenges are experienced for Big Data adoption for the bank in the case study?

- What are the business benefits of adopting Big Data for the bank in the case study?

1.5 Significance of the study

This paper analysed the role and adoption of Big Data by conducting a case study in a South African bank. Moreover, it critically reviewed the available tools while identifying the associated challenges and perceived benefits. The study then recommended solutions that were identified in the literature review relating to Big Data adoption. The study added to the existing BoK in the field of Big Data adoption by providing insight into the South African context and the banking sector.

The study provided bank leaders and Big Data practitioners with valuable insights on the challenges presented by Big Data and the importance of adopting Big Data, including the benefits derived from doing so.

1.6 Delimitations of the study

The delimitations of the study were that the study was limited to the banking industry. The interviews were conducted in the IT Division in the bank. The scope of the interviews included direct clients of the IT Division, such as the BI team and Data managers. The study only focused on the Corporate and Investment Banking and Wealth and Investment IT Divisions and did not include personal and business banking.

The delimitations of the study are:

a) The characteristics of Big Data.
b) How Big Data is collected and processed, and the collection tools used.
c) Challenges in collecting and processing the data.
d) Challenges for business application.
e) Perceived benefits and business value linking the bank's strategy.

The abovementioned were investigated in a South African context in the data space of the banking sector.

1.7 Definition of terms:

Term	Description
ADV	Advance Data Visualisation.
Business Intelligence	Berson and Smith (1997) define BI as the technologies, applications and practices for collection, integration and analysis and presentation of business information.
Chief Information Officer	The original definition of a CIO is “a senior executive responsible for establishing corporate information policy, standards and management control over all corporate information resources” (Peppard, Edwards, & Lambert, 2011).
Source to Target Mapping	The invention relates to a method for mapping at least one data column from a database source to at least one data column of a data target, the method comprising defining at least one reference column of the data target and at least one database source column; performing a comparison of data contained in the data column(s) with the reference column(s); and determining mapping candidates between the data column(s) and the reference column(s) (Lingenfelder, Raspl, & Saillet, 2012).

Term	Description
Subject Matter Expert:	An individual with a deep understanding of a process, function or domain.

1.8 Assumptions

The assumptions are as follows:

a) All respondents reflected standard perspectives and experience.
b) No respondents will be biased towards the organisation they work for, and they will be industry-specific.
c) There is already a basic understanding of what Big Data is and its relevance to the population to be interviewed.

1.9 Chapter Overview

Chapter Outline

Chapter	Content
Chapter 1 – Introduction	In this chapter, the researcher introduces the nature of the study by discussing the purpose of the study, the context, background of the research problem, the objective of the study, the significance of the study and the scope of the study.
Chapter 2 – Literature Review	In this chapter, the researcher provides an in-depth literature review of the study. The definitions and characteristics of Big Data are discussed, the theory of Big Data, the tools for

	collecting Big Data, the challenges for Big Data as well as the benefits for Big Data.
Chapter 3 – Research Methodology	In this chapter, the researcher introduces the research methodology used in the study. That is, research approach, research design, data collection methods, the population and sample for the research, research instruments, procedure for data collection, data analysis and interpretation, limitation of the study, transferability, credibility and dependability of the study.
Chapter 4 – Presentation of Findings	The statement of findings was presented in this chapter.
Chapter 5 – Interpretation of Findings	In this chapter, the findings were interpreted and contrasted to the literature review.
Chapter 6 – Conclusions and Recommendations	In this chapter, the study was concluded, and recommendations were made on how to implement and adopt Big Data projects using tools that are linked to the planned outcome. The recommendations also included how to mitigate challenges based on the findings and literature review.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

This chapter assesses previous literature on the relevance of Big Data, the collection methods and the tools used to collect Big Data. The chapter will also evaluate the challenges and benefits thereof. The first section will be a theoretical review on Big Data, and the second section will evaluate prior studies addressing the application of Big Data.

2.2 Theoretical review

2.2.1 *Big Data background*

Boellstorff (2013) conducted a study to demonstrate that theory is necessary for the compilation of Big Data. The study found four conceptual interventions for the Big Data theory. The first concept is dated theory, which states that data is a temporal phenomenon. The second concept is metadata making, which argues that theory is needed when metadata is formed to give it meaning. The third concept argues that theory is needed in the dialectic surveillance and recognition of Big Data. The last concept compares raw data, cooked data and rotted data. This concept explains data before it is given meaning after it has been given meaning and decaying data that is losing meaning. Meaning can only be achieved when theory is included.

Bowker (2014) conducted a qualitative study to investigate the view that Big Data alone is adequate to model outcomes without the use of theory. The study found that the exaggeration of Big Data while excluding theory leads to withering away of interpretation of the data. The study further argues that everything that we create embodies both theory and data.

A study conducted by Wise and Shaffer (2015) found that there are fundamental problems with the analysis of large-scale data if theory is not involved. They maintain that the researcher conducting the analysis needs to assign variables to the data in order to interpret a multitude of micro-results and make them

actionable. The study argues that theories such as pattern-matching theory and the correlation theory do not imply causality of what is being researched. These theories, however, make it possible to operationalise concepts as the operationalisation depends on the theory that has been assigned to them. The purpose of the study was to indicate why theory matters more than ever in the age of Big Data.

2.2.2 *Definition of Big Data*

The origins of Big Data are still unconfirmed, but studies suggest that it originated in Silicon Valley (Gandomi & Haider, 2015). Big Data is defined as the automated collection and analysis of large volumes of data (Fox & Do, 2013). Big Data is also defined by Fox and Do (2013) as a collection of large data sets that cannot be processed using traditional computing techniques; it involves data that is processed by different devices and applications which produce analytics. Provost and Fawcett (2013) argued that when defining Big Data, the concepts of data science and data mining should also be assessed. Data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. It unifies data analytics, complicated algorithms and machine learning in the process of data mining and data interpretation. Data mining is the process of extracting and examining data to create meaningful information from it.

Katal, Wazid, and Goudar (2013) defined Big Data as a large amount of data that requires new technologies and architectures so that it is possible to extrapolate value from it by the process of capturing and analytics. Kitchin (2013) argued that there is no agreed definition of Big Data; however, a survey of emerging literature denotes that Big Data is massive in volume. It has high velocity that is being created close to real-time. It is diverse in variety as it is unstructured and structured. It is exhaustive in scope due to its attribute of capturing the whole population or system. Big Data is also defined as fine-grained in resolution by aiming to be as detailed as possible and flexible and rational in nature.

Chen, Mao, and Liu (2014) referred to Big Data as the amount of data involving a large scale of mainstream software tools. The tools are used in achieving the

goal of collection, storage and management of the data. The tools are also used in the processing of data to bring more active decision information to assist enterprise management within a reasonable time.

Sharma (2015) defined Big Data as a collection of data sets that are too large and too complex to analyse using simple databases and traditional analytical tools. Recent definitions of Big Data incorporate its characteristics. For instance, Gandomi and Haider (2015) defined Big Data as high volume, high velocity and a wide variety of information – assets that require innovative forms of information processing for enriched insights and decision making. Their study further describes the fourth characteristic identified at a later stage as veracity. Veracity explains the unavoidable unreliability that is inbuilt in the data due to the foundation of where it is collected. An example of this would be a human interpretation of the emotions on a social media post.

Gil and Song (2016) defined Big Data as huge volume, complex and rapid growing data sets with numerous autonomous and independent sources. Big Data is also described as a large amount of structured, semi-structured and unstructured data that is collected at an unusually high velocity (Behera, Sahoo, & Pradhan, 2017). Turner and Atkinson (2018) defined Big Data as large volumes of data that inundates businesses during their day-to-day operations. This data is processed through complex algorisms and then analysed for meaning.

2.2.3 *Definition of Big Data analytics*

Russom (2011) defined Big Data analytics as the application of advanced analytic techniques to an extensive data set. In a study that was conducted to expedite user understanding for Big Data analytics and tools, Russom (2011) further describes Big Data analytics as a combination of Big Data and advanced analytics. This is intended to give the user and business sponsors the BI they need for their business.

MacAfee et al. (2012) argued that Big Data analytics seeks to glean intelligence from data to translate it into business advantage. Chen et al. (2012) defined Big Data analytics as the BI and analytics technologies that are focused mostly on

data mining statistical analysis. The study further argues that Big Data analytics requires trial and error and experimentation.

Gandomi and Haider (2015) argued that Big Data is useless in a vacuum; its potential is only realised when it is leveraged to drive business value, and that is achieved through analytics. The study classifies Big Data analytics into five categories. The first category is text analytics, which refers to techniques that extract information from textual data. The second category is audio analytics, which analyses and extracts information from unstructured audio data. The third category is video analytics, also known as video content analysis, involving a diversity of techniques used to analyse, monitor and extract meaningful information from video streams. The fourth category is social media analytics, which refers to the analysis of both structured and unstructured data from social media channels. The fifth category is predictive analytics, comprising a variety of techniques that predict future outcomes based on historic data and current data.

Jha, Jha, and O'Brien (2016) argued that Big Data analytics focuses on transforming vast quantities of data into useable information. The qualitative study also found that intelligent business processes need Big Data analytics to provide a competitive advantage to the business. Saura, Herráez, and Reyes-Menendez (2019) defined Big Data analytics as the process of analysing large amounts of information organised in a structured and unstructured way. The study also argues that the goal for Big Data analytics is to find trends, correlations, patterns, insights or customer preferences to improve the decision-making process.

2.3 Prior studies on Big Data collection tools

Russom (2011) conducted a study to expedite user understanding of the new data analytics tools in recent years. The study was conducted using a sample of 325 respondents from the Data Warehousing Institute (TDWI) – a data management professional’s database. The results of this study found that Hadoop MapReduce is not only suitable for data volumes, but it is also suitable for data variety. Data variety includes structured data, unstructured data and semi-structured data. It was also found that the Enterprise Data Warehouse is

the preferred tool for data analytics. The finding also indicated that from the tools available in the industry, the tools that indicated growth and favour from companies are advanced data visualisation (ADV), advance analytics and predictive analytics tools.

Gandomi and Haider (2015) conducted a qualitative study to analyse the fundamental concepts for Big Data and the tools used. The study was also aimed at combining the previously disjointed debate on the components of Big Data and the standards that define the size and other features of Big Data. The study found that predictive analytics tools, which deal mostly with structured data, overshadow other forms of analytics tools applied to unstructured data, which constitutes 95% of Big Data. This study also argues for the need for new statistical methods for Big Data to address the anomalies that differentiate Big Data from smaller data sets.

In another study, Sharma (2015) assessed the rise of Big Data, the tools used for the collections and processing and the challenges that come with it. This study also critiqued the mining of Big Data, its application, challenges and the tools that are available for the analysis of Big Data. The study found that Hadoop MapReduce is becoming prominent in the field of Big Data. The study also found that Big Data is a rising pioneer because of its relevance and application in almost all the major sectors. The study, however, further suggests that it is necessary for data scientists to do more research to investigate this area to combat challenges for collection tools, processing and applications that could arise in the future. This subject is still going to be more diverse, which will increase its complexity.

Vanka and Sudha (2017) conducted a qualitative study to review recent technologies created for Big Data. The study was also intended to assist with the election and acceptance of the right combination of Big Data technologies, technological needs and requirements that detail the specific application of the data. The study found that the Hadoop Ecosystem is the most suitable for Big Data.

A quantitative study was conducted by Dhaenens and Jourdan (2019) to explain how data mining problems can be considered as combinatorial optimisation problems, and how metaheuristics can be used to address them. The study found that knowledge discovery in database processes can be used as a tool in recognising effective, novel, valuable, and understandable patterns from large datasets during the process of data mining or data collection.

Another study was conducted by Jabbar, Akhtar, and Dani (2019) to evaluate which Big Data tools would be appropriate for professions such as marketing and if these tools can access real-time data for better and faster accurate decision making. The study found that Apache Storm and Hadoop are the appropriate tools when fast access to data is required. Achieving secure Big Data collection based on trust evaluation and true data discovery makes the data collection process innocuous and effective.

Big Data collection roots from different sources and application. It is, therefore, imperative to have a mechanism to ascertain if the data from these sources can be trusted and if the quality of data from these sources is adequate. The best tool to use to guarantee trust from these sources is the implementation of the trust evaluation model. This was a finding from a study conducted by Lv and Zhu (2020), to evaluate the specific measures of data security collection from the two aspects: trust evaluation of data nodes and true data discovery.

2.4 Prior studies on Big Data challenges

In the qualitative study conducted by Russom (2011) to expedite user understanding of the new data analytics challenges in the recent years, it was found that it is necessary for companies to address the challenge of skills. This is because data analytics requires specialised skills for data mining and data warehousing. The study also found that companies need to ensure that they have the correct business case for the projects that incorporate Big Data to avoid loss of money. Problems with skills, sponsors and software are the leading barriers for data analytics.

MacAfee et al. (2012) conducted a qualitative study to demonstrate that support from executives has an impact on the success of Big Data implementations and

resolution of Big Data challenges. The findings of the study indicate that in addition to the correct talent management and correct technology use, there should be the right level of engaged leadership to avoid the challenges of implementation and adoption. The decision making of the company must be aligned with what the data indicates. The study further advocates that the culture of the company should support data-driven decisions and not executive experience instincts. These contribute to the challenges in Big Data adoption. Big Data power, however, does not eliminate the need for a company to have vision or human insight.

Fox and Do (2013) conducted a study to analyse critically the Big Data hype and the tools used. He used a realist study methodology informed by references to predictive theory and action research. This critical analysis revealed that the hype about Big Data underplays the many challenges in achieving information and transformational effects. Practical implications of Big Data initiatives are not just about the technology and tools used; these initiatives include many complicated inter-related factors.

A study was conducted by Fox and Do (2013) to analyse Big Data and its challenges. It was found that with the collection of data come new complexities. Companies need to know precisely what it is they want to do with the data and the best ways to process the data once they have collected it to minimise implementation challenges. This is not usually an easy task for most companies, as this is still vague for them. The study further explains this to be a pre-existing complexity that has been a challenge for some time.

Katal et al. (2013) conducted a study to analyse Big Data challenges and ethical practices. The study found that there are two main challenges for IT professionals handling Big Data. The first challenge is designing a system that handles Big Data and the second challenge is to filter the correct data from all the data collected that will add value to the business.

There is a need for new methods of handling and analysing data sets that consists of billions of observations generated on a dynamic basis in a variety of forms. Kitchin (2013) identified this as a challenge in a study to analyse Big Data opportunities, challenges and risks.

Kaisler, Armour, Espinosa, and Money (2013) conducted a qualitative study aimed at critically analysing the issues and challenges of Big Data. The study also aimed at assessing methodologies for Big Data analysis and design. The study found that IT researchers and practitioners of Big Data are challenged to design suitable systems that can handle data effectively, and that can analyse data to give it relevant meaning. This challenge is due to the rapid growth rate of Big Data that is exceeding the growth in skills.

Fan, Han, and Liu (2014) undertook research to assess the features and challenges of Big Data and how these features impact the paradigm shift on statistical, computational methods as well as computational architecture. The study found that Big Data is characterised by high dimensionality and large sample sizes that pose three challenges.

The first challenge was that dimensionality brings noise accumulation, distorted correlation and incidental consistency and the second challenge was that high dimensionality combined with large sample size brings high computational costs and algorithm complexity and instability. The last challenge was that the large samples of Big Data are aggregated from multiple sources collected at different times using different technologies.

Sharma (2015) conducted a qualitative study to assess the rise of Big Data and the accompanying challenges. The study also evaluated the mining of Big Data, its application, challenges and the tools that are available for the analysis of Big Data. The findings of the study consolidate the Big Data challenges into three categories: analysis of data, privacy and security of data and representation of data. Analysis of Big Data describes the difficulty associated with analysing unstructured data and ensuring that the correct data that is being retrieved. The second challenge appraises the privacy and security in Big Data applications, and the last challenge addresses the representation of data to ensure its meaningfulness.

Bello-Orgaz, Jung, and Camacho (2016) conducted a qualitative study to analyse the recent achievements of and challenges for social Big Data. The study found that the recent social Big Data challenges are in areas such as data storage, data

processing and data representation and how data can be used for pattern mining, analysing user behaviour and visualising and tracking data.

Using a qualitative study, Behera, Sahoo, and Pradhan (2017) emphasised the challenge of the representation of Big Data; however, the study focused on real-time processing of the data. The finding of this study illustrated that it is challenging to retrieve stored data efficiently, to process it and to visualise it in real-time. Social data demonstrates the same difficulty. Other challenges identified in this study include data storage due to the volume of Big Data, handling and representation of the data.

Cloud-hosted data poses a different challenge. In a case study conducted by Feng and Zhao (2017), it was found that risk would need to be managed differently for forensic investigations that are reliant on cloud services for data hosting. The purpose of the study was to assess trends for digital forensics and to assess the Internet of Things – devices that collect data stored in a cloud computing environment. When risk is identified, the user needs to disconnect from the cloud to ensure that no damage is done.

A study was conducted by Sivarajah, Kamal, Irani, and Weerakkody (2017) to analyse Big Data challenges and analytical methods. The study found that data challenges are the group of challenges associated with the characteristics of the data itself. Process challenges are the group of challenges encountered while processing and analysing the data, that is, from capturing the data to interpreting and presenting the results. Finally, the management challenges for Big Data are the group of challenges encountered while accessing, managing and governing the data

Consequently, the challenges create an unfortunate view for some companies where they question the actual value of Big Data. There are still discussions regarding the relevance of Big Data and the differentiation between Big Data analysis and regular analysis. This was one of the findings of a qualitative study done by Turner and Atkinson (2018) aiming to analyse the emerging effects and challenges due to Big Data in different business functions, as well as how the insights gathered are affecting business strategy. The view of Turner and

Atkinson (2018) is not popular amongst scholars, but this is still a view that should be considered.

2.4.1 *Proposition 1*

Proposition 1*: Big Data collection requires planning and tactics.*

2.5 Prior Studies on Big Data business benefits

According to a study conducted by Russom (2011), intending to expedite user understanding of Big Data benefits, it was found that Big Data can benefit organisations with customer relations, including segmentation, improved targeting and changed customer behaviour. The second business benefit that was found was BI, which includes business insights, understanding change and additional details regarding cost drivers. The other benefits relate to analytic applications, such as fraud detection, risk quantification and market sentiment analysis. Big Data also has the potential to assist with process automation in future.

Abbasi, Albrecht, Vance, and Hansen (2012) conducted a qualitative study that critically analysed the BI aspect of Big Data to increase business benefit. The study also aimed to create a framework to combat data fraud. The framework incorporates three areas of focus; meta-fraud, meta-learning and enhanced financial fraud detection. Experiments conducted in these areas revealed that each component of the framework makes an important contribution towards the overall efficacy of the framework for combating data fraud. Meta-fraud frameworks increase confidence in predictions for data fraud to benefit the business.

A study was conducted by McAfee et al. (2012) using public research from web search data that was compared with a model of the National Association of Realtors. The study aimed to indicate how the new flow of information due to Big Data can change the company's performance. The study found that Big Data benefits businesses by improving decision making and predictive analysis. The findings also highlighted the importance of an organisation's culture in business decision making. If the executives or leadership are not willing to follow what the

data is saying as opposed to their intuition, they will not get maximum value from the data.

In the same school of thought, H. Chen, R. H. Chiang, and V. C. Storey (2012) conducted a study focusing on BI by reusing the IBM Trend-Tech Report (2011) and the Bloomberg Business Week Survey (2011). The purpose of the study was to show the benefit of Big Data for BI and analytics. The findings indicated that, in addition to being data-driven, the application of BI could leverage opportunities presented by the data and domain-specific analytics needed in many essential and high-impact application areas that impact the business positively.

Fan et al. (2014) conducted a qualitative study probing the Chinese systems that measure fiscal budget performance. The study found that the appropriate combination of Big Data tools and the analytics thereof provides an accurate evaluation of the fiscal budget performance, which was a great business benefit for the performance of the Chinese budget.

In e-commerce, controlled experiments and the inclusion of Big Data has become the new foundational scientific tool for decision making (Gandomi & Haider, 2015). The purpose of the study was to test the design, measurement, financial constraints, analysis and decision making for e-commerce. The study was conducted using controlled experiments of Walmart data. The finding of the study also indicated that Big Data analytics has a major impact on decision making for the benefit of the business.

The study conducted by Jha et al. (2016) critically analysed the impact of Big Data analytics on business processes. The findings of the study indicated that the actual real impact is dependent on the architecture of the organisation. It is not only Big Data that matters, but the whole business process of the organisation, including the correct use and storing of paper-based information that will have a realised impact that translates to profit.

2.5.1 *Proposition 2*

Proposition 2: *Effective Big Data planning and tactics has a positive influence on business outcomes.*

2.6 Conclusion of the literature review

As far as the researcher could establish, data is a temporal phenomenon, and for it to make sense, meaning needs to be given to it. The analysis of data becomes meaningful when the data has been collected from a variety of sources, in different states and in high volumes. Hadoop seems to be the most preferred tool for data collection in the Big Data field. There is also consensus from the scholars that when Big Data is implemented and adopted correctly, business benefits are realised; however, this does not mean there will be no implementation and adoption challenges in the process.

In addition to what the researcher could establish, the literature that was reviewed indicated the following gaps:

- There is no study available that focuses on a Big Data implementation and adoption case study in the banking industry from a South African perspective.
- There is a need to research Big Data benefits and challenges for the banking sector and how data is practically collected and processed in this sector to create business value.

CHAPTER 3. RESEARCH METHODOLOGY

The previous chapter covered the literature review; this chapter focuses on the research methodology used to conduct the study of Big Data application in the banking sector in South Africa. The chapter also describes and categorises the target population, the research approach, research design, population and sample, the research instruments, procedures for data collection, data analysis and interpretation and limitations of the study.

The chapter outlines the logic behind the transferability, dependability, credibility and confirmability of the study as well as the demographic profile of respondents and ethical considerations for the study.

A qualitative research methodology was followed in this paper. Semi-structured interviews were conducted within a specific population to collect data. The data was then correlated with the literature review to outline similarities and discrepancies.

3.1 Research approach

The research approach was qualitative in nature. The data collection method was in the form of an empirical study with feedback from respondents in semi-structured interviews. A deductive research approach was followed, and this included constructing propositions and using available literature and other data collection tools such as interviews and documentation of observed practice. This approach is suitable when the researcher has time constraints, which is the case with this research. The researcher will also use an inductive approach by contributing new data to the BoK regarding whether the existing Big Data tools evaluation models are suitable for banking or if they need an extension.

This research aims to investigate a real-world setting, and qualitative research is suitable for that (Golafshani, 2003): this is why a qualitative research method was chosen. Good qualitative research can help us understand a situation that could prove to be incomprehensibly confusing (Flinders & Eisner, 1994).

3.2 Research design

The research design was in the form of an empirical case study conducted at a South African bank to understand the subject in sufficient detail. Thematic analysis was used to classify the collections process, challenges and benefits into themes and groupings. The classification assisted in defining classes of the collection process, the challenges and the benefits.

The empirical qualitative case study allowed the researcher to uncover deeper meaning during the interview process; it also allowed the researcher to understand the setting where the research is conducted (Golafshani, 2003). The qualitative research case study further allowed the researcher to have access to data on the ground, and the researcher was exposed to observing how this process is executed by industry professionals at the bank.

3.3 Data collection methods and procedure

Semi-structured interviews were used for the data collection method. Observations for the empirical study were also documented. Semi-structured interviews are suitable data collection methods because they allow the flexibility of open-ended questions with some structured content as well, which will lead to more valid, well-grounded construction of realities (Golafshani, 2003). The data collection procedure was conducted in an informal setting that ensured an environment that allows for open engagement.

The collected data consisted of 15 transcripts from 15 respondents according to the sample target specified. The below procedure was followed:

- An email was sent to participants to schedule interviews with the interview questions attached.
- Reminder emails were sent to the respondents as the time approached.
- Notepad was used to capture additional information.
- A recording device was used to transcribe at a later stage.

3.4 Population and sample

3.4.1 *Population*

Saunders and Lewis (2012) define the population as the complete set of research contributors that are accessible to the researcher. The population for this study was the bank's Group IT Department. The bank's strategy strongly encourages continuous improvement, simplification, and finding new ways of serving their customers better by being more client-centric. Big Data is an area of interest for the bank due to the perceived benefits to assist with this. The findings from the study are beneficial to the organisation, and this contributes to the reason why the population is suitable for the study.

3.4.2 *Sample and sampling method*

Saunders and Lewis (2012) define a sample as a subgroup of the population. The sampling method that was used was a purposive sample. Saunders and Lewis (2012) define this as "a type of non-probability sampling in which the researcher's judgment is used to select the sample member based on a range of reasons and premises". The researcher used his judgment to choose appropriate subject matter expects that would have the industry knowledge and information regarding the subject. The researcher conducted 15 interviews comprising data engineers, solution architects, a data analyst, platform leads and product owners.

Table 1: Profile of respondents by position

Job Title	Number To Be Sampled
Senior Data Engineers	4
Intermediate Data Engineer	1
Senior Solution Architect	1
Intermediate Business Product Owners/BI specialist	4
Senior Application Platform Heads	4
Senior Data Analyst	1
TOTAL number of respondents	15

3.5 The research instruments

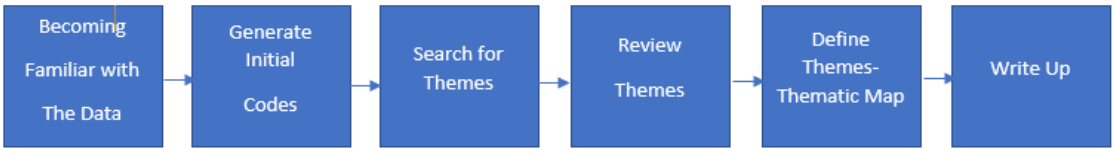
The instruments used were a notepad to note additional information, a recording device for the interviews, Otter to assist with transcribing and NVivo to assist with analysis. The interviews were semiformal in nature to allow rich, detailed information. According to Saunders and Lewis (2012), researchers using semi-structured interviews usually have a list of questions to ask. The researcher had a list of questions; however, since these were face-to-face interviews, additional questions were asked where necessary. Observations were documented in the process as well.

3.6 Data analysis and interpretation

Thematic analysis was used to develop sequential, systematic methods for the collection of data. Thematic analysis was also used to assist the researcher with the systematic process of handling the analytic phases during the inquiry period, as well as to help with streamlining and the integration of the data collection and

analysis process (Charmaz & Belgrave, 2007). The process included the details of how raw data was collected through interviews during the study. The interview questions were open-ended to allow expression and respondents could provide detailed opinion and information on the subject.

This process is depicted below:



Source: Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars (Maguire & Delahunt, 2017).

The steps below were followed:

Step 1: Becoming familiar with the data

A recording device was used to record the interviews. This process was followed by vetting to check that the audio was clear and in a satisfactory quality to be transcribed. The recorded interviews were then loaded on the Otter tool to assist with transcribing. The transcribing process then commenced for all the recordings with the aid of the Otter computer-based tool. The generated text transcripts had errors that required additional effort from the researcher to listen to the recording and fix the errors on the generated transcripts. An initial impression on the information collection was written.

Step 2: Generate initial codes

After this process, the next step was loading the raw interview data into NVivo, which is a computer-aided analysis tool. The purpose of the tool was to assist with coding and analysis of the data. Initial codes were created using NVivo.

Step 3: Search for themes

After the process of code generation, the step that followed was to search for themes. The study was concerned with addressing specific research questions, and as a result, a theoretical thematic analysis was followed. Given this, the coding of each segment focused on data that was relevant or interesting to the research question. Categories were then created by grouping the codes into logical patterns.

Step 4: Review themes

The codes were listed, and those that were familiar were merged to create categories. After a thorough examination of the codes, those that were found to fit together were combined to form a theme. For illustration, a few codes that were relating to the activities required for Big Data collection, such as tools involved, different data formats, the data touch points involved, data quality management and data steward's guidance, were merged. These codes assisted in creating a Collection Process Theme.

Step 5: Define themes – Thematic Map

In this step, the code classes and groupings assisted in identifying taxonomies. This was the final enhancement of the themes. This process was to capture the core meaning of each theme. The relationship between these was depicted with a Thematic Map.

Step 6: Write-up

In this step, the themes were analysed according to the research objectives, research propositions and literature review presented in the paper. A detailed write-up of this process was produced.

Figure 1 depicts the coding, categorisation and theme generation process.

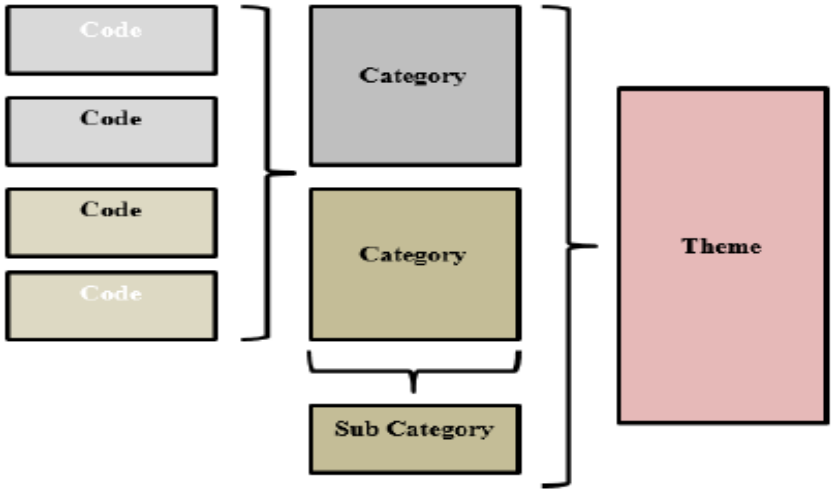


Figure 1: A streamlined Codes to Theme Model for a qualitative inquiry adapted from Arcury & Quandt, 1999.

Limitations of the study

Limitations are:

- Time – Research had to be handed in within five months.
- Data sensitivity – The information that will be shared could be considered as part of the strategic competitive advantage and will need to be managed carefully.
- Researcher bias – This is as a result of qualitative research being subjective, and it could reflect the perspective of the researcher (Saunders & Lewis, 2012).

3.7 Transferability and dependability

3.7.1 *Transferability*

This refers to the data collection and analysis methods used to enhance the transferability of the study. This process ensured that a systematic data collection process is followed so that the validity of the study is not compromised. The process legitimised qualitative research as a scientific study (Charmaz & Belgrave, 2007). Thus, this assists with transferability for future use. The data

was collected in a case study during working hours, and the study was supported by the organisation studied. The respondents were in their natural state of mind, and the work surroundings were normal, making it easier for them to be open about their experiences during the data collection process. A case study is not meant to be transferable, and this study did not aim to be transferable.

3.7.2 *Credibility*

The use of appropriate software for data analysis and the assistance of an expert (supervisor) to interpret the data enhanced the credibility of the study. The software assists with giving a base of appropriate required support because of its capabilities. The supervisor will provide input and direction regularly to ensure that the information provided is valid and the correct scientific processes are followed.

The supervisor will assist in validating if:

- The correct methodology was followed as planned
- The study data collection process is adequate
- The sample size and responses are adequate
- Does the study measure what it claims to measure?

The process mentioned above will help in ensuring the credibility of the study.

3.7.3 *Dependability*

The background for this study and the description of the theory in Big Data technology, Big Data collection process, Big Data challenges, Big Data analytics and Big Data benefits is supported by a systematic approach that enhanced the dependability of the study. This process was supported by an external audit from the supervisor to evaluate the data collection process. The evaluation included an assessment to check whether the findings, interpretation and recommendations are all supported by the data collected in the study.

3.8 Ethical considerations

Wealth South Africa IT Department CIO provided the sign-off to allow the study to be conducted.

In addition to this, the following was considered:

- Consent forms were signed by all respondents who were included in the study.
- The researcher signed the plagiarism form, as stated in the university guidelines.
- The research permission communication process from the university was followed.
- Ethical clearance from the university was obtained.
- Privacy - Respondents were not mentioned by name.
- Data sensitivity - The information that was shared could be considered as part of the strategic competitive advantage and will need to be managed carefully.

CHAPTER 4. PRESENTATION OF FINDINGS

4.1 Introduction

In this chapter, the background of the organisation is provided, followed by the presentation of the findings of the study. The analysis process is presented, including the coding process, analysis of codes and the steps in Thematic Analysis that were taken in the study. The findings of the analysis process are presented concerning Proposition 1. This is followed by the presentations of the findings in relation to Proposition 2.

4.1.1 *Organisation background*

Bank A (referred to as 'Organisation A' or 'the organisation') is a South African bank listed on the Johannesburg Stock exchange and is one of the top leading banks in the South African Market. Bank A believes in focusing on African investments, and this is reflected in their presence in over 20 countries in the African market. Bank A is a financial institution that offers the following:

- Transactional banking
- Lending
- Saving
- Borrowing
- Risk management
- Investment
- Insurance
- Wealth management

The abovementioned services are categorised under the following business areas:

- Personal and Business Banking
- Corporate and Investment Banking
- Wealth Management

This study only focused on the IT participants from the Corporate and Investment Banking and Wealth Management Departments. The Personal and Business Banking Departments were not included in the scope.

4.1.2 *Profile of respondents*

The research respondents' (RR) profiles are described in this section. The profiles included the titles, job levels, experience in the bank, and experience in the field of study.

In the context of this study, the participants had to have experience in and knowledge of dealing with data. Arcury and Quandt (1999) stated that how you select your respondents is crucial in a qualitative study. It is essential to select participants that have experience and knowledge of the subject matter.

Profiles provided below:

RR1 Profile
RR1 is a White, male Senior Data Engineer who has been working for the organisation for five years. RR1 has 21 years in the data field and currently operates in the capacity of a senior manager in the organisation. RR1 has the responsibility for setting up standards of operation for the team to operate under; he ensures that work is produced upholding those standards for the Wealth IT Business Unit. **RR1 is very knowledgeable in the data field and has been operating with various data collection, curation and storage tools.**

RR2 Profile
RR2 is a Coloured male Senior Data Engineer who has been working for the organisation for four years. RR2 has 19 years in the data field of study and currently operates in the capacity of a senior manager in the organisation. **RR2 is responsible as a team leader, and he is responsible for the quality of the work the team produces in the Insurance Business Unit. RR2 is also actively involved in the day-to-day operations of the team.**

RR3 Profile
RR3 is an Indian, male Senior Data Engineer in Corporate and Investment Banking (CIB). RR3 has been working for the organisation for eight years, and he has been in the data field for 12 years. **RR3 is responsible for the quality of delivery of the CIB business unit.**

RR4 Profile
RR4 is an Indian, male Executive Head Data Platforms, Data Services and Artificial Intelligence (AI) and Machine Learning for the organisation. RR4 is operating at group level and is responsible for the management of the data platforms, the service provided by the data team and building AI capability for the organisation that will assist with machine learning and advanced analytics. **RR4 is responsible for the technical execution of the group strategy to achieve a single view of client. This is done in partnership with the Executive Head of Data from the business. RR4 has been with the organisation for 24 years and has 12 years' experience in the data space.**

RR5 Profile
RR5, an Indian male, is the Head of Information Architecture for Personal and Business Banking (PBB) responsible for the architectural flow of the information in the organisation. The **PBB business area is where the customer information is kept.** **RR5 is currently managing the technical implementation for a single view of client as in the PBB business unit, and he is putting the groundwork in place for the organisation, as PBB is where the customer data resides.** **RR5 has been with the organisation for** ten **years and has been in the data field for** ten **years.**

RR6 Profile
RR6 is a Black, male employee of the organisation and he operates as the Head of Data Quality and Data Profiling Management for the Platform Team. RR6 is responsible for the data quality of the organisation's data and customer profiles before the data is included in the reservoir. **RR6 has been with the organisation for 13 years, and he has been in the data field for 12 years.**

RR7 Profile
RR7 is the Head of AI platforms for the Data Services team. He is responsible for building AI capabilities for the data that is collected in the Data Lake. **He is male and Indian. RR7 is responsible for building the machine learning capability for the organisation and introducing models that should assist the organisation with understanding their customers better, as well as enhancing the predictive and prescriptive analytics.** **RR7 has been with the organisation for four months; however, he has been in the data field for seven years.**

RR8 Profile
RR8 is a Black, male BI Manager. RR8 has been with the organisation for 12 years and has been working in the data field for 12 years. **RR8 is responsible for developing BI reports for the Wealth Business area and he is also responsible for implementing the business rules for producing meaningful and valuable business insights.**

RR9 Profile
RR9 is a White, female BI Manager for the Insurance Business area. RR9 is responsible for creating BI IT engineering capabilities that will add value to producing business insights for the Insurance Business area. **RR9 has been with the organisation for five years; however, she has been in the data space for over 25 years.**

RR10 Profile
RR10 is a male, Chinese Senior Manager in the Platform team. He is responsible for the reporting and visualisation delivery that is produced from the data that has gone through the quality process. **RR10 has been with the organisation for 14 years, and he has been in the data field for 14 years.**

RR11 Profile
RR11 is a Black, male Junior Business Intelligence Manager in the Insurance Business space responsible for building the BI reports. RR11 is also responsible for actioning the strategy for his area of work by implementing the deliverables indicated on the outlined roadmap. **RR11 has been working for the organisation for seven years**, **and** he **has been in the data field for 12 years.**

RR12 Profile
RR12 is a male, Indian Senior Enterprise Architect who has been assigned to the Wealth Business area. RR12 is responsible for assisting the Wealth Business area with contributing to the single view of client but moving all their data to the Data Reservoir and contributing towards enabling the group in collecting meaningful data. **RR12 has been working in the organisation for ten years, and he has been in the data field for ten years**.

RR13 Profile
RR13 is a male, Indian Product Owner and Data Steward for the Insurance business. RR13 is responsible for representing business products that should be delivered by the IT team. RR13 is also responsible for data for the insurance space from a business perspective. **RR13 has been working for the organisation for 20 years, and he has 12 years' experience in the data field.**

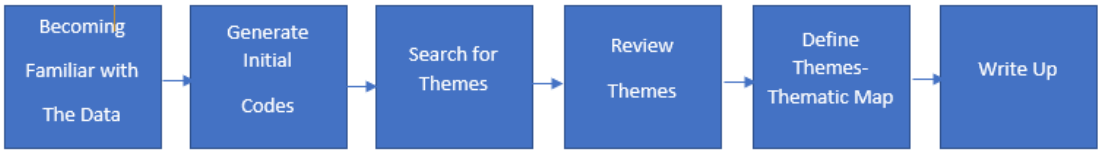
RR14 Profile
RR14 male, Black Data Engineer in the Wealth IT area. R14 is responsible for implementing the data road map for the Wealth Business area. **RR14 has been working for the organisation for five years, and he has eight years in the data field.**

RR15 Profile
RR15 is a Black female Data Analyst and Design Lead in the Wealth IT area. RR**15 is responsible for depicting the data road map and guiding the team with analysis capabilities in implementing the data road map for the Wealth IT area.** **RR15 has been working for the organisation for four years; she has seven years' experience in the data field.**

The research questions were considered during the process as follows:

1. Which tools can be used for Big Data in the banking sector?
2. What challenges are experienced for Big Data adoption in the banking sector?
3. What are the business benefits of adopting Big Data in the banking sector?

Thematic Analysis Process



Source: Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars (Maguire & Delahunt, 2017).

The above analysis process was followed, which resulted in a consolidated view of the codes that were collected. This process was followed by a breakdown of the results per proposition.

Access Attributes	Governance Attributes	Data Attibutes			
person access	data governance	mask data	current data challenges	data lake	data security
role-based access	governance issues	advanced data	customer data management	data landing	data set
unauthorized access	**Issues Attributes**	analysing data	data acquisition pattern	data monetization framework	data standards
user access	skill issues	data encryption	data aggregation	data output	data stewards
Business Attributes	dependency issues	data sources	data agreement	data owner	data storage
business benefits	**Landing Attributes**	big data	data analytics	data platform	data strategy
business level access	landing zones	big data adoption	data architect	data processing	data touch points
business rules	landing areas	big data governance	data catalogue	data products	data visualization
business space	**Tools Attributes**	big data perspective	data cleaning process	data quality analysis	data volumes
business users	adequate tools	big data process	data committee	data quality management team	data warehouse
Collection Attributes	big data collection tools	brokers data	data dictionary	data quality tools	head data quality
big data collection tools	Fit for purpose tools	businesses data	data engineering	data reservoir	historical data
collection point	data quality tools	data types	data environment	data safety	internal data
Case Attributes	Hadoop main tool	collecting data	data field	data science	large data
case study	main visualization tool	commercializing data	data formats	data science model	mainframe data
use case	open source tools	correct data	data governance	data science workbench	periodic data quality
business case	profiling tools	different data assets	financial data	replicating data	transporting data
User Attributes	real time tools	different data attributes	group data services	right data	unstructured data
specific user roles	reporting tools	different data elements	predominant data collection tools	sensitive data	data lake
use case journeys	vendor tool	different data formats	profile data asset	small data	
user guides	visualization tool	enterprise data warehouse	pure data	specific data marts	
	analytical reporting tools	existing data initiatives	quality data	technical data catalogues	
		exploratory data analysis	raw data	teradata data sites	
		external data	Data manual	transactional information data	

Figure 2: Consolidated view of the codes

4.2 Results pertaining to Proposition 1:

The results relating to Proposition 1 are presented in this section.

Proposition 1*: Big Data collection requires planning and tactics.*

4.2.1 *Becoming familiar with the data*

A recording device was used to record the interviews. This process was followed by a vetting to check if the audio was clear and that the quality was acceptable for transcription. The process of transcribing for all the recordings commenced. This was done using Otter, a computer-based tool.

Initial impression: *Big Data collection tools require planning and tactics.*

It seems that the organisation has a challenge with lengthy processes when allowing stakeholder access for the data platforms. Access Management for the end-to-end process for Big Data adoption seems cumbersome. As much as the organisation does not have an issue with the availability of tools, the data collected seems to suggest that there is a challenge in the shortage of skills associated with the end-to-end collection and curation process.

The Hadoop suite seems to be the preferred technology stack when working with the Data Reservoir and the Data Lake. However, the CIO teams that are involved with the implementations do not have adequate skills and training to operationalise the work that needs to be done in their respective areas. This seems to have an impact on the time it takes to reach the planned strategy of a single view of client. The single view of client is the strategic initiative of the organisation to move all the customer data and system data linking to the customer to a single repository to have a single source of the truth when it comes to their client. Data that is being moved to the Data Reservoir needs to be of good quality. There seems to be a process in place to deal with data quality; however, the challenge seems to be that there is still significant reliance on the Data Source to provide clarity.

The rigorous compliance process appears to be frustrating for the data engineers and other role-players. The experience for the people interviewed suggest that the process is time-consuming and tedious, and the teams seem to believe that it should be optimised. In as much as the data catalogue assists with interpretations and mapping during the processing and curation process, what was noticed is, it does not replace domain knowledge.

There is still a need for the technical team to work with the data stewards or domain experts. The data catalogue is deemed as too technical, and this is impacting the adoption of usage by the business areas. Last, data compatibility seems to be a challenge. This is due to legacy systems that were procured in the early stages of the organisation; now these systems seem to be outdated, and they seem to need additional effort for the extraction and extrapolation of meaning from the data.

Loading of transcripts into an analysis tool

Interviews were recorded using a recording device and loaded on the Otter tool to assist with transcription. Having completed this process, the step that followed was loading the raw interview data into NVivo, a computer-aided tool. The purpose of the tool was to assist with coding and analysis of the data.

4.2.2 *Initial Code generation:*

Once the interview data was loaded in the tool, the researcher proceeded with the process of code generation. The study was concerned with addressing specific research questions, and as a result, theoretical thematic analysis was followed. Given this, the researcher coded each segment of data that was relevant or interesting aligning to the research question. Categories were then created subsequent to this, grouping the codes together into logical patterns

Table 2: Sample of initial codes

Description	Code
Access codes	access
	person access
	role-based access
	unauthorised access
	user access
Business codes	business
	business benefits
	business-level access
	business rules
	business space
	business users
	data collection
	Big Data collection tools
	collection point
	business data
	current data
	data analytics
	data dictionary
	data governance
	data stage
	data touchpoints
	enterprise data warehouse
	transporting data
Governance codes	governance
	data governance
	governance issues
Challenges codes	issues
	skill issues
	dependency issues
Landing area codes	landing
	landing zones
	landing areas
Tools codes	tools
	adequate tools
	Big Data collection tools
	fit-for-purpose tools
User codes	user
	business users

The sample above is a subset of the consolidated view of the codes

4.2.3 *Search and review themes*

After a thorough examination of the codes, there were codes that were found to be fitting together to form a theme. The codes were then organised into categories that shared similar attributes (Rogers, 2018). As an illustration, the table below shows a representation of the output of the categorisation of the codes.

Table 3: Category to code mapping

Category	Codes
Access	person access
	user access
	role-based access
	unauthorised access
Business	business-level access
	business users
Data Sources	brokers data
	businesses data
	external data
	financial data
	transactional information data
	unstructured data
	raw data
	sensitive data
Data Curation Process	business rules
	customer data management
	data acquisition pattern
	data aggregation
	collection point
	mask data
	analysing data
	data encryption
	data cleaning process
	data conversions
	data engineering
	data science model
	data science workbench
Data Governance	role-based access
	user access data governance
	business-level access
	business users

Category	Codes
	access management
	unauthorised access
	governed access
	landing zones
	Big Data process
	Big Data governance
	data standards
	data committee
Challenges	governance issues
	skill issues
	dependency issues
	governance delay issues
	dependency issues
	lack case studies
	use case understanding
	business case understanding
	current data quality challenges

Subsequent to this, further analysis indicated the need for two categories. Access and Business were combined to form a new category: Access Management. This process is depicted below:

Table 4: Merged category

Merged Category	Category	Codes
Access Management	Access	person access
		user access
		role-based access
		unauthorised access
	Business	business-level access
		business users

4.2.4 *Define Themes – Thematic analysis*

To ensure that the organisation of the study could identify the challenges associated with the adoption of Big Data, the relationships that link the categories associated with the challenges were investigated. This is depicted in a Thematic Map below;

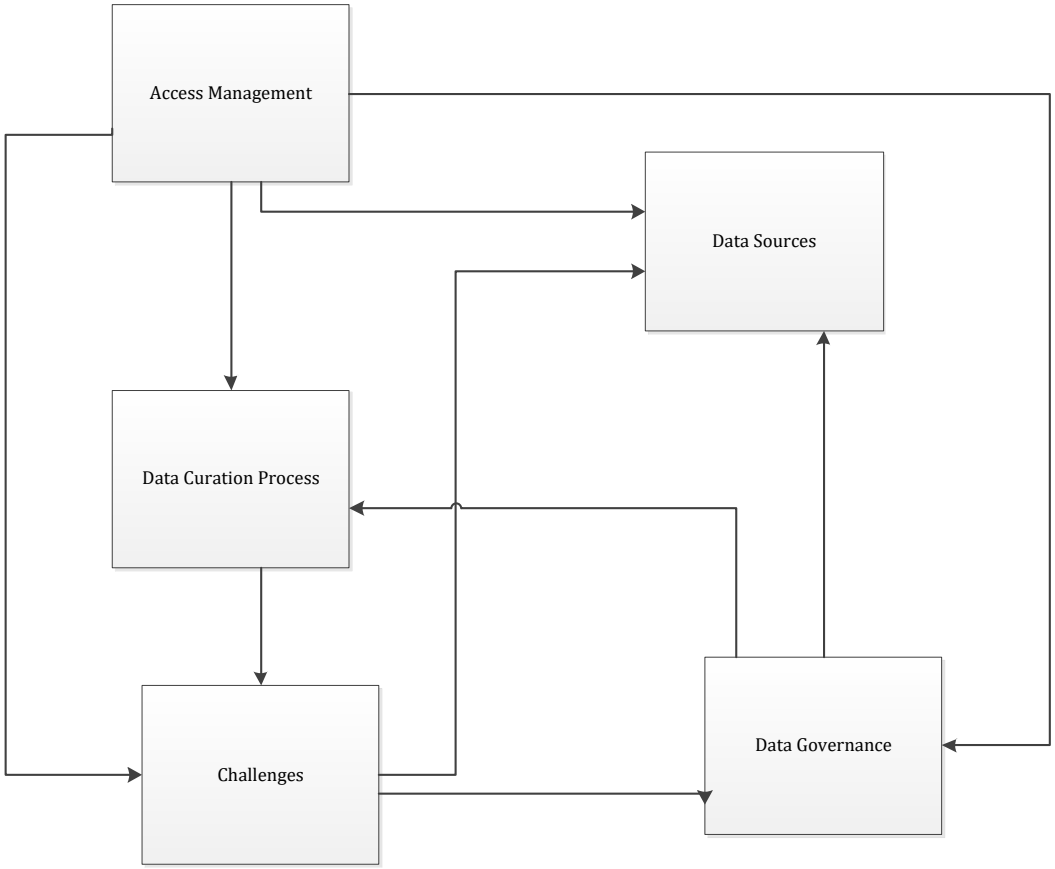


Figure 3: Thematic Map for Proposition 1

From the identified relationships in the categories, the following themes emerged: Challenges, Data Governance, Data Curation, Data Source and Access Management.

4.2.5 *Reviewing of themes*

The researcher analysed the identified categories and grouped the categories that had similar attributes and could fit with each other. During this process, there appeared a need to rearrange the codes further. Some of the codes in the

categories Business, Data Sources and Data Curation were joined together to be part of a Data Governance theme as these codes were more fitting as part of that theme.

Similarly, some of the codes from Access and Data Curation were moved and grouped to be part of the Data Governance theme. The categories Challenges, Access Management and Data Curation, were converted to themes because they linked to the research questions, as they seemed suitable as themes. They indicated that they had great significance for the area in the study being conducted. The final themes for Proposition 1 were: Access Management, Data Curation Process, Data Governance and Challenges.

The following table depicts the reviewed themes for Proposition 1:

Table 5: Themes for Proposition 1

Theme	Category
Access Management	Access Management
	Data Sources
Data Curation Process	Data Curation Process
Data Governance	Data Governance
Challenges	Challenges

4.2.6 *Theme write-up and presentation:*

Theme 1: Access Management

Access Management emerged as a theme that was well understood by the RRs. There was an understanding as to why it is necessary to be adhered to; however, the governance process required before access is granted seems to be unwieldy. The operational process involved in managing access seems to be time-consuming. These delays do not appear to be caused by limitations in technology but rather by lack of process optimisation.

RR2*: "**We have things like role-based access, and that gets monitored very closely with the DBAs. And it can get a bit heavy because there is a prerequisite document that needs to be filled up as to why you would need this."*

The same sentiment is highlighted by RR3 below, also detailing the impact on the internal operational delays;

RR3: "*On top of this, many of these, like teams, give frameworks and practices that we have to adhere to otherwise we won't get access to the data. To add to that, this is the reason why projects take so long to get delivered. It takes about three to six months to go through this process before we can even start because it's in the bank, everything is highly regulated*."

RR13: "*Yes, we do. I think the bank has a little bit of red tape with regards to the data and how it protects its data*."

Theme 2: Data Curation process

Data curation is the process of cleaning and organising the data so that it will be in a consumable state for the data owner and data users. The rules that are used to clean the data are usually based on the intended use outlined by the data owners. This is handled via the Extract Transport and Load (ETL) process, the Source to Target Mapping process (STTM) and the Data Quality Remediation process. The end-to-end process is seen to be delayed by dependencies on other teams and governance around the process.

RR1: *“We're using the normal ETL principles, translations as per the STTMs set up by business.”*

RR13: *“So we use different processes and methods or applications as well. So we compare, we look at the different data, we compare to the industry standards that we have. We work a lot with the data quality management team within the bank as well, to find out what processes and what, and how they actually apply data cleaning in there. But yeah, we do a lot of comparisons against the source data. To check the data there, if there's data conformity shift, we have to conform to different standards as well. So, there's different rules and standards in place that have been put, if the data does not conform to that. We actually have to get it cleaned according to those standards. And if it's not, we try to put business rules in place into the system as well, to make sure that when at the point of capture the certain data conformity and standards that are applied as well.”*

RR14: *“In my case, I would say, because we work hand-in-hand with, with business. So mainly, we've got this STTM or Source to Target Mapping, where there would be logic that, that we need to apply.”*

There is also a detailed process that has been bedded down to ensure data quality that was explained well by RR5 below:

RR5: “*So we [do] manual remediation and the automated remediation or systematic remediation. The manual remediation job is specifically to eyeball the information [in] the cases where we need to, and then you have automated. We also use a third-party source to enrich our data sources, and we have Hive-Jobs application to remove the redundancy.”*

Theme 3: Data Governance

Data governance in the context of the study seems to be the governance process around who should have access to the data collected, the platforms and tools for the collections process and limitations according to business areas according to regulation or data agreements. The data curation in itself does not pose a challenge to the collection process; however, the issues are more enhanced by the detailed Data Governance process it has to follow. This process seems to have a lot of red tape, and that can be time-consuming.

RR13: "*I think the bank has a little bit of red tape in regard to the data and how it protects its data, we actually mask data into the databases, we use different methods of secure transformation of data."*

RR7*: "…so we're not allowed to pull data into our Big Data platform unless it's been authorised by the data owner. So we only land data there if the data owner has given us approval."*

Theme 4: Challenges

The challenges that have been highlighted in the data that has been collected speak to operational dependencies on other teams, skills and common understanding amongst stakeholders due to the technology being new in the organisation as well. These challenges are reflected in detail by the data collected below:

RR1: "…*dependency issues. So, we've got a lot of dependencies between source systems. Another thing is understanding the data by both the developers and the users, what's needed to get the desired result from what the business expects."*

RR2: "…*my biggest challenge I have at the moment is dependency, dependency on another team."*

RR3: "*So we, we are a centralised capability for CIB. Within use cases. So again, I'm heavily dependent on the teams that have given me the data."*

RR4: "*I think the bigger challenge is the proficiency of the user community."*

RR7: "…*understanding the metadata associated with the tables or the columns or whatever it is, if that's not available, then interpretation of the data could be a challenge."*

RR15: "*So I think when we made the decision to embrace EIM 3.2 I think all the checks and balances and contracts and you know, things should have been made, and patents should have already been bedded down. And the whole thing about going to platforms and saying, okay, so I hear about the fact that you are saying we can stream some data using a Kafka topic. But then you told, we don't yet have a specialist."*

4.3 Results pertaining to Proposition 2:

In this section, the results relating to Proposition 2 were presented.

Proposition 2*: Effective Big Data planning and tactics has a positive impact on business outcomes.*

4.3.1 *Becoming familiar with the data*

The researcher worked on Otter to review the transcripts, focusing on themes that might emerge relating to the research questions mentioned in Section 4.2.1.

Initial impression: the positive impact of Big Data adoption

Based on the data collected, the organisation of the study seems to have a good technical foundation to handle the end–to-end collection and processing of the data. The Hadoop suite seems to be the preferred technology stack for the Data Reservoir and Data Lake. The Data Reservoir is the strategic initiative by Organisation A to enable them to store data in a single repository. There seems to be a process in place, such as Hive-Jobs for checking and dealing with data redundancy before the data is stored. There also appears to be a periodic data quality profiling exercise to deal with the quality and correctness of data from source. An Infosphere module called Data Stage seems to be the preferred tool for cleaning up the data, as well as checking for data conformity. This seems to suggest that the organisation invested in fit-for-purpose tools to avoid technology challenges.

In as much as there seems to be a rigorous compliance process in the organisation, there is also a general understanding that the organisation deals with sensitive information that needs to be protected. There is an understanding as well regarding the benefits of the rich analytics that will be available due to the Big Data collection. There seems to be a strategic intent by the organisation to monetise the data that is being collected by building suitable data products as well as client-centric data-driven solutions as well as following initiatives that are indicated by the collected data.

4.3.2 *Loading of transcripts into an analysis tool*

NVivo was used as the tool of choice for analysis of the transcripts, as mentioned in Section 4.2.1. The purpose of the tool was to assist with coding and analysis of the data that was collected.

4.3.3 *Initial Code generation:*

In this section, the concern was addressing specific research questions, and as a result, theoretical thematic analysis was used, focusing on Big Data benefits. Given this, the coding of each segment of data was done in accordance with what was relevant or interesting in relation to the research question. This was followed by the creation of categories, grouping the codes into logical patterns

Table 6: Codes for Proposition 2

Description	Code
Collection Codes	
	Big Data collection tools
	collection point
Data Codes	
	current data
	data analytics
	data dictionary
	data governance
	data stage
	enterprise data warehouse
	transporting data
Tools Codes	
	adequate tools
	Big Data collection tools
	fit-for-purpose tools
Big Data Benefits Codes	
	adoption benefits
	business benefits

The above sample is a subset of the consolidated view of the codes

4.3.4 *Search for themes*

After a thorough examination of the codes, the researcher found codes that were fitting together to form themes. The codes were then organised into categories

that share similar attributes (Rogers, 2018). As an illustration, the table below shows a representation of the output of the categorisation of the codes. The following categories were identified: Data Collection Process, Data Curation Process and Benefits.

Table 7: Categories for Proposition 2

Category	Codes
Data Collection Process	enterprise data warehouse
	data stage
	transporting data
	fit-for-purpose tools
	landing areas
	customer data management
	data acquisition pattern
	data quality management
	data quality tools
	data reservoir
	data quality tools
	Hadoop main tool
	main visualisation tool
	open source tools
	visualisation tool
	data lake
Data Curation Process	business rules
	customer data management
	data acquisition pattern
	data aggregation
	collection point
	mask data
	analysing data
	data encryption
	data cleaning process
	data conversions
	data engineering
	data science model
	data science workbench
Big Data Benefits	data analytics
	commercialising data
	data products
	data quality analysis
	data monetisation framework
	data visualisation

4.3.5 *Thematic analysis*

To ensure that the organisation could identify the benefits that are associated with the adoption of Big Data, it was necessary to look at the relationships that linked the categories that were associated with Big Data benefits. The collection of data affords the business a volume and variety of data that enables them to make informed decisions from the insights retrieved from analytics. The choice of tools used also enables the business to collect and store this data coming in at a high velocity. The tools also empower business to curated data.

Figure 4 depicts a relational model of the identified patterns for Big Data benefits.

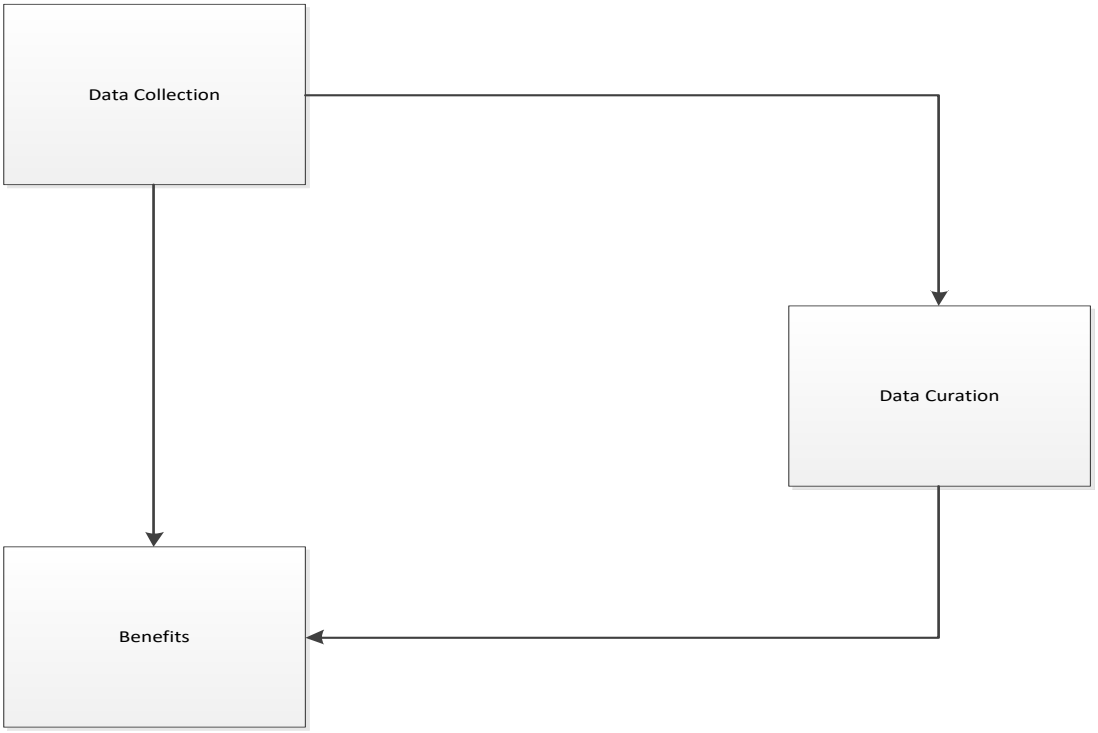


Figure 4: Thematic Map for Proposition 2

Theme Discussion:

Theme 1: Data Collection

The data collection process describes the process and what is entailed to collect the data. This is achieved through fit-for-purpose tools that can cater for the high velocity of data and the translation of the data. The process also describes the variety of the data and the tools associated with the varied data. Lastly, the high volumes of the data, as well as the technology that is required for storing and

integrating this data, is outlined. This is all conducted in a secure and stable environment.

RR4: "*I think the, the attraction of a Big Data environment using Hadoop as the underlying technology layer, is the fact that you can use cheap commodity hardware. So, you don't need to have expensive high-end appliances. It is built for failure. It's built for redundancy. So, if you lose one node, these many others and data is replicated multiple times in that process. So that's the attraction of the environment."*

RR4: "*I think in the data processing, there's quite a rich ecosystem of tools, supporting Big Data technologies. So there's no shortage of tooling around it. So whether it is Spark type technologies, or Hive, which gives you traditional sequel type, capabilities on the Big Data environment, or more advanced frameworks like TensorFlow and PySpark for advanced machine learning.*"

RR5: "*So what are your Big Data collection tools, so collection, and so there's [a] difference between collection versus storage. So collection happens at different instances. For example, we have tool for ... pilot, which is for images, we just we have too, we have triple, Avia Triple C, which is basically the voice recording. And then you have. So we talked about the images, we talked about the voice, we don't have videos. And we have external data sources, that's it, and biometric*."

Theme 2: Data Curation

Data curation relating to Big Data benefits speaks to the value that is brought by the curated data views that are consumable by the business or data owners. This is the process where business rules have been applied, and data has been cleaned and masked to a suitable level. This is highlighted by the data that was collected below.

RR2: "*So from my experience, what would happen is business would request access to a table, the access would be granted to them if they meet the prerequisites. And based on that the data would then be analysed so that the business rule can cater for the ETL process in order to clean up the data the way the business users would want to see. I think that's why the bank went with Data Stage [a]s you've got a lot of stages within the actual tool to do clean-ups like*

removing duplicates, doing lookups, doing filtering. So, you know, there's a lot of features of data stage that would assist the business at the end of the day to see the end result of what they need."

RR11: "…*so basically, it's extracting data, huge volumes of data, and being able to store that data in a central repository. So that obviously we can be able to give our business users high quality of data. So it's mainly storing that huge, huge volumes of data."*

RR4: "*We have a large mainframe environment, so we use Ab Initio for processing mainframe data and often to do data conversions, we have an in-house built Spark framework which we use for processing. We use Kafka as another capability and as well as NiFi. So those are the predominant data collection tools. I mentioned use Ab Initio because we have mainframe data, which is not always in first normal form, so it needs to be translated into consumable format. Spark is really great for distributed processing. The nature of Big Data environments is you have a distributed cluster with different storage endpoints and different compute endpoints. So Spark is really good at processing around those environments. Kafka we use for real-time event-based architectures and NiFi for something similar."*

Theme 3: Big Data Benefits

Big Data benefits are the actualised benefits that are achieved through the collection of Big Data. This is achieved through the data analytics, the creation of data products and the actual monetisation of the data through these created products and the competitive advantage created by the collected customer insights. This is reflected in the data collected from the respondents below:

RR4: "*I think what the, the key benefit is the ability to bring large amounts of data together, apply a host of techniques and processes that we know, that typically wouldn't have been possible to conduct that sort of analysis, you know, in in the past, and actually, you know, use really advanced capabilities like machine learning to, to, to create a to create a far higher return on your data asset. You know, we've seen that, you know, there's almost been a plateau around the value we go out of data for point in time, because the techniques that were available*

didn't allow you to fully exploit that data. Actual return that you're getting out of working with the data, it is obviously predicated on starting with the right problem to begin with."

RR5*:* "*So, real-time monitoring you can view them on a quality of it, you can view the business outcome of it, you can validate the business outcome, you can validate the volumes, you can track the performance, and obviously there are many more things, insights you can get, you can, you can also predict what will business want So, you can apply a level of predictive analysis says and then yeah, so there are multiple use cases."*

RR15: "*I think in keeping a good shape of your market, because if you know what customers want, and are always addressed in servicing, whatever the needs are, then there won't be a need for you to lose, you know, a customer base to your competitors."*

CHAPTER 5. INTERPRETATION OF FINDINGS

This chapter discusses the interpretation of the findings. Interpretation involves a discussion of the emerging themes and comparing and contrasting these themes to the literature review. The categorisation of the themes and their link to the research questions is depicted below;

Table 8: Research propositions with emerging themes

Research Proposition	Themes
There are challenges with Big Data collection and application in business.	Access Management
	Data Curation Process
	Governance Process
	Process Challenges
Big Data has a positive impact on the business.	Data collection
	Data Curation
	Big Data Benefits

5.1 Interpretation and discussion of results for Proposition 1

In this section, the interpretation of the results pertaining to Proposition 1 was discussed. The discussion entailed a process of contrasting the results with the literature review. In order to ensure a logical flow in the discussion, the themes that emerged from Chapter 4 were combined where relevant.

5.1.1 *Theme 1: Access management and governance*

Data that is heavily regulated, such as in the organisation, requires that there should be controls in place to ensure that the data is protected. The process of Access Management ensures that correct access is granted to the user for the appropriate business need. Organisation A also has Data Chinese Walls as part of their governance process, ensuring that only the business unit that is authorised to see or use certain data can see only that data and nothing else. Hadoop being the technology of choice for the organisation is suitable for Access Management as it has an access layer that manages file distribution. This layer is capable of constructing role-based configurations that support adequate role-based access.

This finding from the research is supported by the study conducted by Sharma (2015) that confirmed that Hadoop is the tool most suitable for Big Data. The purpose of this study was to assess the rise of Big Data, giving focus to the tolls associated with the collection and processing of Big Data. A similar study conducted to assist with the selection and adoption of the right combination of Big Data technologies, validates Sharma's finding further and confirms that the Hadoop Ecosystem is the most appropriate tool for Big Data due to its makeup of the Processing Layer, Querying Layer and the Management Layer with the ability to handle Access Management. This study was carried out by Vanka and Sudha (2017)

The correlation indicated that Organisation A has the correct tools and processes in place for Access Management. In as much as the process may be seen as time-consuming, it is not a technology issue but a process issue that can be improved.

5.1.2 *Theme 2: Data Curation Process and Process challenges*

Subsequent to further analysis, data curation was found to be the process of organising data in a suitable and consumable form for the data owners and the users. This process was found to be very important as it assists in transforming the data to be meaningful for the end-user. In the context of challenges that are experienced during data curation, Organisation A found this process time-

consuming and challenging. The process of interpreting the data by using a data catalogue is not always a seamless exercise, but Organisation A understands the value add that is obtained from the curated views of the data. In as much as this process is complicated, it is a necessary process as it contributes to the curation of the data.

A study conducted by Sivarajah et al. (2017) indicates this to be a common challenge in practice as well. The study found that data challenges are a group of challenges associated with the characteristics of the data itself. Process challenges are the group of challenges encountered while processing and analysing the data, that is, from capturing the data to interpreting and presenting the final results. The management challenges for Big Data are the group of challenges encountered while accessing, managing and governing the data.

Technology in use to collect the data does not seem to be an issue for the organisation: the issue seems to be the skills that are required to operate the tasks to reach the required final state. The same finding was made in the study conducted by Russom and Phillip (2011). The study indicated that, although the technology in use for Big Data implementation might be obtainable, there is still an element of skill shortage that needs to be addressed. These skills are still new to most organisations.

The data collected in the study shows that business understanding is crucial to ensuring that there are positive results with Big Data projects. The organisation needs to ensure first that the data is collected to resolve the right problem. This indicates that there should be an understanding from the organisation regarding the use case they would like to resolve.

This finding is confirmed by a study conducted by McAfee et al. (2012). The study indicated that knowing which problem to resolve is essential and can be achieved by partnering with domain experts. McAfee et al. (2012) further state that it is crucial for companies to set clear goals, ask business areas the right questions and ensure that that they understand what success should look like to solve the right problems. This would guarantee the success of Big Data projects, not just a collection of enormous amounts of available data, but a clear plan on what needs to be achieved.

The data collected from the respondents also indicated that data standards are a challenge during Big Data collection and processing due to the data being collected from different source systems and from a variety of data types. The resolution of this can be achieved by employing a quality standards management process that should also be partnered with domain experts.

The study conducted by Fan et al. (2014) indicates that large samples of Big Data aggregated from multiple sources collected at different times, using different technologies, pose significant challenges during the collection process. This observation is further supported by the findings from the research participants, which indicate that when data is sourced from different systems and processed or stored by different systems, this creates dependency issues in the end-to-end process. These dependencies slow down the curation and storage process because of misaligned priorities and long periods waiting for approval from the different teams.

5.2 Interpretation and discussion of results for Proposition 2

In this section, the interpretation of results for Proposition 2 was provided, and these were discussed in detail and contrasted with the literature review. Some of the identified themes were combined for discussion as it seemed logical to group these when providing a detailed interpretation.

5.2.1 *Theme 1: Data collection and Data Curation*

Data collection should be done through fit-for-purpose tools to process data velocity, variety and volumes. The research findings indicate that the organisation is aware of this, and this is shown by their large toolbox for the different data characteristics. The organisation uses Ab Initio for processing mainframe data. There is an inbuilt Spark framework for processing data as part of the collection and curation process. Kafka and NiFi are also part of the tools available in the organisation's toolbox. For other unstructured data, such as images and voice recording, Organisation A is using Avia Triple C. There are external data sources that can handle biometric information and there are social media analysis tools available to Organisation A.

The strategic intention for Organisation A to have a single view of a client is being actioned by creating a single repository for the data they have as an organisation. The tool used for this is the Hadoop Ecosystem. The same tool is used for the Data Lake, with the strategic intent of using the raw data stored there for experimental and exploratory purposes. Hadoop supports other tools so that the intended use can be realised.

To provide an example, Organisation A uses Hive as a supporting tool for traditional sequel requirements and TensorFlow and PySpark for advanced machine learning. This demonstrates that Organisation A caters for data volumes and velocity through Hadoop. Variety and veracity are catered for through ETL tools and other fit-for-purpose tools that are available for use for the organisation.

The Hadoop suite has proven to be the most suitable Big Data tool, as indicated in the study done by Vanka and Sudha (2017). This study also indicated that it is necessary to understand the purpose of these available tools. This will assist with understanding which tools are applicable and available for a specific use. Organisation A demonstrated this understanding through the use of tools such as Ab Initio for mainframe data collection and conversions and processing, Spark Framework for processing, and Kafka and NiFi for collection. All these tools have the capability of plugging into Hadoop. These tools can handle the customer management process as well as the data acquisition patterns that are prescribed by the organisation. The study by Vanka and Sudha (2017) supports the notion that the Hadoop Ecosystem is the most suitable for Big Data.

5.2.2 *Theme 2: Big Data Benefits*

Big Data benefits can be divided into organisational benefits and data monetisation benefits. This was a finding from the data provided in the study of Organisation A. Organisational benefits include the rich data that is available to provide insights to the business to make better and more informed decisions. These insights are achievable through data analytics, made possible by tools such as Power BI and ClickView. These tools do not only assist with BI reporting but also with predictive and prescriptive analytics used to provide solutions and forecasts to the customer and the organisation.

Data monetisation is the opportunity provided by the collected data to create revenue for the organisation. This opportunity is created through the knowledge and insights gained from the curated views as well as exploratory and experimental projects conducted on the data. Organisation A has built capabilities and a platform to benefit holistically, operationally and economically. The data collected from the study also indicates that there should be a strategic intention for the collection of the data, and this should have active backing from the leadership of the organisation. A constant is that a review of processes is necessary to ensure the organisation is performing its tasks optimally and efficiently.

The study conducted by Russom (2011) aligns with the view of Organisation A, as indicated in the findings that state the benefits of Big Data to be BI, including business insights, understanding change and additional details regarding cost drivers. This addresses the need to use the gained insights to adjust operations and plan for revenue growth as the data indicates. A study conducted by H. Chen et al. (2012) states that the benefit of Big Data is products or outcome that is produced from the data. The BI applications can also leverage opportunities presented by the data and domain-specific analytics. These analytic outputs are needed in many essential and high-impact application areas. These domain-specific analytics also impact business positively by contributing insights for data monetisation.

Organisation A is creating data products from the data that is collected. The process of data-driven product development is not only for client products, but it is also process-enhancing software and models from machine learning and experimental processes. The benefit is made possible by ensuring from the start that there is a problem that can be resolved and that the correct questions and methods are applied to resolve the problem. This is all supported by the relevant technology, such as ClickView and Power BI, to reflect what the data indicates.

The study by McAfee et al. (2012) found that Big Data benefits businesses by improving decision making and predictive analysis. The findings also highlighted the importance of an organisation’s culture in business decision making. If the executives or leadership are not willing to follow what the data indicates (as opposed to their intuition), they will not get maximum value from the data. This is

aligned with Organisation A's view on leadership involvement for Big Data adoption, and it indicates that Organisation A is on the recommended path for achieving their goal. The study conducted by Jha et al. (2016) indicated that the actual real impact is dependent on the architecture of the organisation. This is supporting the notion of Organisation A of creating a platform and building capabilities to achieve the intended benefits through the architecture of the organisation.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

The purpose of this study was to establish the challenges and benefits of Big Data adoption for a bank in the South African context. The study's primary aim was to establish whether a bank would experience similar challenges and benefits when implementing and adopting Big Data projects or if new theories would emerge when focusing in a developing economy in a South African context. The following research propositions were formulated as part of the study.

Proposition 1: *Big Data collection requires planning and tactics.*

Proposition 2: *Effective Big Data planning and tactics has a positive impact on business outcomes.*

The study indicated that Big Data collection requires planning and tactics. The tactics required are mostly associated with a shortage of skills. This is because Big Data applications are new to most organisations. The other challenges that require tactics are due to a common understanding of how the data is translated from different sources. This process revealed that the data dictionary is not always adequate. There is still a strong reliance on the domain experts. Finally, the dependencies that are created because of the way the organisation has structured its operations also exacerbates the slow pace of the implementation process. This has also been identified as an area that requires tactics and further planning.

The study indicated that there is a positive impact and a benefit for business in adopting Big Data. These benefits emerge when good planning and good tactics have been put in place. The benefits are also achieved through the insights gained from data analytics. These contribute towards how the business understands its customer, which then leads to the business' competitive advantage. The information also assists the business with data-driven product development. These factors enable client-centricity, predictive and prescriptive analytics, as well as data monetisation opportunities.

The propositions were addressed by answering the following research questions:

1. Which tools can be used for Big Data in the bank in the case study?
2. What challenges are experienced for Big Data adoption for the bank in the case study?
3. What are the business benefits of adopting Big Data for the bank in the case study?

6.1 Conclusion for Research Question 1:

Which tools can be used for Big Data in the banking sector?

The literature review has shown that Hadoop is the preferred tool for Big Data collection. The Hadoop Ecosystem (also known as a suite) is compiled in such a way that it consists of the Processing layer for batch and real-time processing and the Querying layer for running queries against the data sets. The ecosystem also consists of an Access layer for the collection and aggregation of data that is used for analytics and a Management Layer for the workflow jobs of the different systems interacting on the Hadoop platform. This ecosystem makes it possible for supporting technology to be plugged into Hadoop to enhance the functionality of the layers.

Similar to the literature review, the findings from the data collected indicate that the organisation also prefers Hadoop as their Big Data tool of choice. They use other fit-for-purpose tools, such as Ab Initio for mainframe environment, Kafka and NiFi for collection of data, Spark framework and Data Stage for processing and Infosphere for data quality checks, all plugging into Hadoop to achieve their intended purpose.

The literature review and the data collected from the research participants are aligned, showing that Hadoop is a suitable Big Data tool. The organisation seems to be on a level with industry standards in terms of using the appreciate technology for Big Data adoption.

6.2 Conclusion for Research Question 2:

What challenges are experienced for Big Data adoption in the banking sector?

The literature review indicates that one of the challenges experienced during Big Data adoptions is the need for specialised skills for data mining. The business also needs to have the correct business case to start with to avoid losing money. Organisations should also ensure that the technology they employ is used correctly and, last, there should be executive buy-in so that the organisation can follow data-driven decision making.

A similar pattern is shown by the data collected in the research. The organisation also seems to have an issue with finding the appropriate skills required for their end-to-end process for Big Data adoption. What comes across as a new challenge from the data collected is the dependency issues between teams. This is due to the team construct in the organisation: as a result, there are delays added to the process.

The data also indicates an issue with common understanding amongst the affected stakeholders of the data catalogue (also known as the data dictionary). The data collected also highlights further delays that are introduced by the governance process during the curation and processing phases. The outcome of the study indicates that the organisation in the case study is not facing unique issues that have not been tackled by other companies implementing Big Data projects. The organisation could take direction from other industry players in resolving these challenges, but this also presents an opportunity for innovation to present solutions that will set them apart from the rest.

6.3 Conclusion for Research Question 3:

What are the business benefits of adopting Big Data in the banking sector?

The literature review demonstrates Big Data benefits to include the ability to use target marketing for customers. Big Data also allows for monitoring customer

behaviour, which then leads to improved BI due to the better insights derived from the data analytics. The monitoring provides details for cost drivers. The literature further indicates that the organisation will make better decisions because of the predictive analysis enabled by Big Data. The literature additionally states that in order for an organisation to be successful with Big Data projects, there has to be a level of executive alignment, and this needs to show in the architecture of the organisation.

The collected data indicates that Big Data enables data analytics, which allows the business to gain more insights into their customer base. This aligns with the findings from the literature review.

Similarly, the analytics also allow for behaviour monitoring for the organisation to be aware when something changes in their customer patterns. The organisation is busy with a strategic project of a single view of client by moving data into a reservoir. As reflected in the literature review, this shows commitment from the executives by building the foundation needed for their strategic intent. This strategic intent will enable the organisation to have a single source of the truth for client data and improve on client-centricity and data-driven marketing. This single view of client will further enable the data monetisation that will be achieved through the development of data-driven products. Organisation A seems to be well advanced to benefit from adopting Big Data.

6.4 Research problem addressed

The research problem sought to establish how the organisation can tackle the challenges that are associated with the adoption of Big Data for banks. The research also sought to establish if the same benefits experienced by other organisations for implementing Big Data projects would hold for a bank, thus contributing to the BoK in a South African context.

From the literature review it has been established that it is common to experience challenges when embarking on a Big Data adoption exercise, but ensuring that there is leadership alignment on how success should look contributes to the success of the implementation. This is also enabled by the correct selection and use of the supporting technology and tools to achieve the desired outcome. The

findings of the study indicate that the organisation has the right support from the executives and has invested in the correct technology for Big Data adoption. This demonstrates that they have the correct support to tackle the challenges they could face, and they also have the right technological backing to achieve their objective.

6.5 Research objectives and questions analysed

The study findings highlight the tools that are used for Big Data in Organisation A. The data collection tools vary depending on the kind of data that needs to be collected; however, these were found to be fit-for-purpose tools. Hadoop was found to the most suitable tool for Big Data. The study findings were also able to highlight the challenges associated with Big Data adoption for Organisation A.

Finally, the study managed to note the benefits and perceived benefits for Big Data adoption and its intended primary use, which is a single repository of the data to have a single view of client. The single view of client will then enhance how Organisation A creates value for the client based on the insights collected, eventually leading to their competitive advantage. All the findings were found in this case study conducted at a bank to provide a South African context.

6.6 Recommendations

The recommendations below were provided to Organisation A:

- Invest in process optimisation to fast-track their governance process during the approval phase. This will accelerate end-to-end execution allowing the organisation to reach their milestones faster.
- Invest in the training required by the team and ensure that talent management is thorough to ensure that these skills are not lost to competitors.
- Ensure that change management for this journey is measurable to ensure that the correct support is given to the team throughout the process of Big Data adoption.

- Organisation A needs to socialise the need for business to collect clean, high-quality data. The business front line is the main source for the data collection process. If clean, high-quality data is collected, it will reduce work for downstream systems.
- Partner with FinTech companies to assist with skills in areas where the organisation lacks the skills.
- Track data monetisation more aggressively and report on the value to the business to encourage the continuation of investment in Big Data decision making.

6.7 Suggestion for further research

This study was qualitative owing to time constraints that limited the sample size to only 15 participants. It is recommended for future studies to increase this sample size and change the study to a quantitative framework to allow for attributes to be measurable. Actioning research that assesses the impact and effectiveness of the chosen Big Data collection and processing tools is important. This future research should also track the rate of Big Data adoption, the increase and decrease of the challenges as well as constructing a model to measure data monetarisation.

A more extensive study will provide additional insight into the research problem, further adding to the BoK. It is recommended that, in a wider study, more time should be taken to allow for a greater reach of data engineers, data stewards and heads of department in the different platform areas, as well more business owners and business users.

As this research was based on information collected from a case study in one single major bank, it might appear that Big Data adoption is more beneficial to big banks as opposed to smaller banks. It is important for future research to include other banks of different composition to the one studied.

6.8 References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: A meta-learning framework for detecting financial fraud. *Mis Quarterly, 36*(4).

Arcury, T. A., & Quandt, S. A. (1999). Participant recruitment for qualitative research: A site-based approach to community research in complex societies. *Human Organization*, 128-133.

Behera, R. K., Sahoo, A. K., & Pradhan, C. (2017). *Big Data Analytics in Real Time-Technical Challenges and its Solutions.* Paper presented at the 2017 International Conference on Information Technology (ICIT).

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*

28, 45-59.

Berson, A., & Smith, S. J. (1997). *Data warehousing, data mining, and OLAP*: McGraw-Hill, Inc.

Boellstorff, T. (2013). Making big data, in theory. *First Monday, 18*(10).

Bowker, G. C. (2014). Big data, big questions| the theory/data thing. *International Journal of Communication, 8*, 5.

Charmaz, K., & Belgrave, L. L. (2007). Grounded theory. *The Blackwell encyclopedia of sociology*

Chen, Chiang, & Storey. (2012). Business intelligence and analytics : From big data to big impact. *Mis Quarterly, 36*(4).

Chen, Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and Applications*

19(2), 171-209.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly, 36*(4).

Clemons, E. K. (2008). How information changes consumer behavior and how consumer behavior determines corporate strategy. *Journal of Management Information Systems*

25(2), 13-40.

Dhaenens, C., & Jourdan, L. (2019). Metaheuristics for data mining. *17*(2), 115-139.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review, 1*(2), 293-314.

Feng, X., & Zhao, Y. (2017). *Digital forensics challenges to big data in the cloud.* Paper presented at the 2017 IEEE International Conference on Internet of Things (iThings)

Fischer-Baum, R. (2017, 26 November 2017). Tech generation Retrieved from https://www.washingtonpost.com/graphics/2017/entertainment/tech-generations/

Flinders, D. J., & Eisner, E. W. (1994). Educational criticism as a form of qualitative inquiry. *Research in the Teaching of English*, 341-357.

Fox, S., & Do, T. (2013). Getting real about Big Data: applying critical realism to analyse Big Data hype. *International Journal of Managing Projects in Business, 6*(4), 739-760.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*

35(2), 137-144.

Gil, D., & Song, I. Y. (2016). Modeling and management of big data: challenges and opportunities. In: Elsevier.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report, 8*(4), 597-606.

Jabbar, A., Akhtar, P., & Dani, S. (2019). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*.

Jha, M., Jha, S., & O'Brien, L. (2016). *Combining big data analytics with business process using reengineering.* Paper presented at the 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS).

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). *Big data: Issues and challenges moving forward.* Paper presented at the 2013 46th Hawaii International Conference on System Sciences.

Katal, A., Wazid, M., & Goudar, R. (2013). *Big data: issues, challenges, tools and good practices.* Paper presented at the 2013 Sixth international conference on contemporary computing (IC3).

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography, 3*(3), 262-267.

Lingenfelder, C., Raspl, S., & Saillet, Y. (2012). Method for mapping a data source to a data target. In: Google Patents.

Lv, D., & Zhu, S. (2020). Achieving secure big data collection based on trust evaluation and true data discovery. *Computer Security*

101937.

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *The All Ireland Journal of Teaching, 9*(3).

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard business review, 90*(10), 60-68.

Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *11*(5), 1277.

Peppard, J., Edwards, C., & Lambert, R. (2011). Clarifying the ambiguous role of the CIO. *MIS Quarterly Executive, 10*(1), 31-44.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data, 1*(1), 51-59.

PWC. (2017). The Future of Banking : A South African Perpective

Rogers, R. H. (2018). Coding and Writing Analytic Memos on Qualitative Data: A Review of Johnny Saldaña's The Coding Manual for Qualitative Researchers. *The qualitative report, 23*(4), 889-892.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19*(4), 1-34.

Saunders, M. N., & Lewis, P. (2012). *Doing research in business & management: An essential guide to planning your project*: Pearson.

Saura, J. R., Herráez, B. R., & Reyes-Menendez, A. (2019). Comparing a traditional approach for financial Brand Communication Analysis with a Big Data Analytics technique. *IEEE Access, 7*, 37100-37108.

Sharma, S. (2015). *Rise of Big Data and related issues.* Paper presented at the 2015 Annual IEEE India Conference (INDICON).

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263-286.

Trelewicz, J. Q. (2017). Big data and big money: The role of data in the financial sector. *IT Professional, 19*(3), 8-10.

Turner, H. C., & Atkinson, D. (2018). Strategic decision making: The effects of big data. *Economic Social Development: Book of Proceedings*, 126-136.

Vanka, P., & Sudha, T. (2017). Big Data Technologies: A Case Study. *Research Journal of Science Technology*

9(4), 639-642.

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics, 2*(2), 5-13.