



University of Fort Hare
Together in Excellence

**Big Data Use at an Automotive Manufacturer: A Framework
to Address Privacy Concerns in Hadoop Technology**

By

Prenisha Padayachee

201400169

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Commerce

In

Information Systems

In the

Faculty of Management and Commerce

of the

University of Fort Hare

Supervisor: Prof. R Piderit

November 2021

Abstract

An automotive manufacturer can generate big data through accessible data points from internal and external Internet of Things (IoT) data sources connected to the production line. Big data analytics needs to be applied to these large and complex datasets to realise the associated opportunities, such as an improved manufacturing process, optimised supply chain management, competitive advantage and business growth. In order to store, manage and process the data, automotive manufacturers are using Apache Hadoop technology. Apache Hadoop is a cost-effective, scalable, and fault-tolerant technology. However, there has been a concern raised regarding the privacy of big data in Apache Hadoop. A key challenge in Hadoop technology is its ineffective security model, making the data susceptible to unauthorised users.

Consequently, a breach in data privacy results in automotive manufacturers becoming victims of theft of trade secrets and intellectual property via corporate spies. This theft has a negative impact and results in the loss of company reputation, business competitiveness and business growth in the automotive market. This study investigated a solution to ensure big data privacy when using Hadoop technology.

The Selective Organisational Information Privacy and Security Violations Model (SOIPSV) and the Capability Maturity Model (CMM) provided the theoretical base for this study. The researcher undertook a literature analysis and qualitative study to understand and address the identified research problem. The primary data was collected from ten Information Technology (IT) specialists at a local automotive manufacturer. These specialists participated in an interview session, which also included the completion of a questionnaire. All questions were pre-determined and open-ended, and the participants' responses were recorded. Primary data was analysed using the inductive approach by identifying relevant themes and sub-themes. In contrast, the literature analysis included academic journals, conference proceedings, websites, and books, which were critically discussed in this study.

This study's findings indicated various measures to be implemented by the automotive manufacturer to address the research problem. Critical success factors were derived from the identified measures, which addressed significant data privacy issues in using Hadoop technology.

The identified critical success factors included: control of internal and external data sources; monitor the value of big data towards improving the automotive manufacturing process and user behaviour; implementation of user authentication; encryption to secure data; disaster recovery and backup plan; execution of authorisation and Access Control List (ACLS); conduct audits and regular reviews of user access to data; apply data masking to sensitive data and tokenization to secure data; build own infrastructure to store and analyse data; install regular security updates and update passwords regularly. Each factor had a purpose that examined big data management, governance and compliance in detail. The identified factors contributed towards ensuring data privacy in the use of Hadoop technology.

These factors were categorised into contextual and rule and regulatory conditions adopted from the SOIPSV. Identified conditions were then aligned to the five-level CMM. Each condition was expanded upon at various maturity levels to form a framework that addressed the main research problem. The framework's application was described as an independent assessment of each critical success factor and provided a guide through various maturity levels. The framework's purpose was to address and overcome big data privacy concerns in using Hadoop technology at a local automotive manufacturer.

Declaration of Originality

I, Prenisha Padayachee, (ethical clearance number: PID031SPAD01) hereby declare that:

- The work in this treatise is my own work.
- All sources used or referred to have been documented and recognised.
- This treatise has not previously been submitted in full or partial fulfilment of the requirements for an equivalent or higher qualification at any other recognised educational institution.

Signature: Padayachee

Date: 22 November 2021



University of Fort Hare
Together in Excellence

Acknowledgements

I would like to thank the following people who were involved in and who contributed to this research:

- God, for providing me with strength and patience in order to complete this study.
- My supervisor, Professor Piderit, for her continuous advice, guidance, and encouragement throughout this research project.
- The Department of Information Systems at the University of Fort Hare, for their assistance and support, provided throughout my academic career.
- The Govan Mbeki Research and Development Centre at the University of Fort Hare, for the financial support they have provided me throughout this research project.
- Finally, I would like to thank my family and friends for their encouragement, understanding, and support throughout this endeavour.



University of Fort Hare
Together in Excellence

Table of Contents

Abstract	i
Declaration of Originality	iii
Acknowledgements.....	iv
List of Figures	x
List of Tables	xi
Chapter 1	2
The Problem & Its Setting.....	2
1.1 Introduction	2
1.2 Problem statement	4
1.3 Research question and sub-questions	6
1.4 Significance of the study	7
1.5 Theoretical perspective.....	8
1.6 Literature review	17
1.6.1 Introduction	17
1.6.2 Generation and use of big data in the automotive manufacturer	17
1.6.3 Privacy concerns of big data in Hadoop technology in the automotive manufacturer	18
1.6.4 Solutions to address privacy concerns of Hadoop in the automotive manufacturer.....	20
1.7 Research methodology	20
1.7.1 Paradigm and approach	21
1.7.2 Data collection methods	22
1.7.3 Data analysis methods	23
1.7.4 Population and sample.....	25
1.8 Delimitation of the study	26
1.9 Ethical considerations.....	27
1.10 Outline of chapters	27
Chapter 2	30
The Generation and Use of Big Data in the Automotive Manufacturer	30
2.1 Introduction	30
2.2 Overview of Industry 4.0.....	31
2.3 Sources of big data	35
2.3.1 External sources	35
2.3.2 Internal sources.....	37

2.4 Characteristics of big data	40
2.4.1 Volume	43
2.4.2 Velocity	44
2.4.3 Variety	44
2.4.4 Value	45
2.4.5 Veracity	46
2.5 Big data analytics	46
2.5.1 Descriptive analytics	48
2.5.2 Diagnostic analytics	48
2.5.3 Predictive analytics.....	48
2.5.4 Prescriptive analytics.....	49
2.6 Uses cases of big data in the automotive manufacturing industry	49
2.6.1 Optimise supply chain management.....	50
2.6.2 Improve the manufacturing process	51
2.7 Conclusion.....	54
Chapter 3	58
Privacy Concerns of Big Data in Hadoop Technology in the Automotive Manufacturer	58
3.1 Introduction	58
3.2 Overview of Apache Hadoop technology	60
3.2.1 Hadoop in the automotive industry	60
3.2.2 Drawbacks of Hadoop technology	62
3.3 Comparison between privacy and security.....	63
3.4 Privacy challenges of big data in Hadoop technology in the automotive manufacturer	65
3.4.1 Data breaches	66
3.4.2 Fragmented data	68
3.5 Privacy issues related to the selective organisational information privacy and security violations model	69
3.6 Conclusion.....	72
Chapter 4	75
Privacy Solutions for Big Data in Hadoop Technology in the Automotive Manufacturer ...	75
4.1 Introduction	75
4.2 Security measures for the protection of data privacy in Hadoop technology	76
4.2.1 Implementation of user authentication	77
4.2.2 Execute authorisation/Access Control Lists	77

4.2.3 Implement encryption to secure data.....	78
4.2.4 Conduct audits.....	79
4.3 Privacy requirements of big data in Hadoop technology	79
4.4 Data privacy protection methods.....	80
4.4.1 Conduct regular reviews of user access to data.....	81
4.4.2 Apply data masking to sensitive data	81
4.4.3 Implement disaster recovery and backup plan	82
4.4.4 Monitor user behaviour	82
4.4.5 Apply tokenization to secure data	83
4.4.6 Build own infrastructure to store and analyse data.....	84
4.5 Privacy and security recommendations	84
4.6 Privacy measures related to the Selective organisational information privacy and security violations model.....	85
4.7 Conclusion.....	87
Chapter 5	90
Research Design and Methodology.....	90
5.1 Introduction	90
5.2 Philosophical research paradigms	91
5.2.1 Positivism	92
5.2.2 Interpretivism	94
5.2.3 Critical theory.....	96
5.2.4 Design science research.....	97
5.2.5 Selecting an appropriate research paradigm.....	98
5.3 Research Methodology.....	102
5.3.1 Design Science Research.....	102
5.3.2 Capability Maturity Model.....	108
5.4 Research Format.....	109
5.5 Data Collection Methods.....	111
5.5.1 Primary Data Collection Methods	111
5.5.2 Secondary Data Collection Methods	113
5.6 Population and sampling method	113
5.6.1 Sampling method.....	114
5.6.2 The principle of saturation.....	114
5.7 Data analysis	115



University of Fort Hare
Together in Excellence

5.7.1 Pattern matching.....	115
5.7.2 Thematic analysis.....	117
5.8 Research Evaluation.....	121
5.9 Conclusion.....	122
Chapter 6.....	125
Empirical Analysis and Discussion.....	125
6.1 Introduction.....	125
6.2 Interpretation of primary data.....	126
6.3 Theme 1: Generation and use of big data.....	126
6.3.1 Sub-Theme 1.1: Sampled demographics.....	127
6.3.2 Sub-Theme 1.2: The Internet of Things and its effect on the automotive manufacturer.....	130
6.3.3 Sub-Theme 1.3: Big data generation and collection.....	132
6.3.4 Sub-Theme 1.4: Big data storage.....	136
6.3.5 Sub-Theme 1.5: Required tools and techniques.....	137
6.3.6 Sub-Theme 1.6: Creating value from data.....	138
6.3.7 Sub-Theme 1.7: Organisational culture.....	141
6.4 Theme 2: Privacy challenges.....	143
6.4.1 Sub-Theme 2.1: Privacy Awareness.....	143
6.4.2 Sub-Theme 2.2: Access control to the Internet of Things devices.....	144
6.4.3 Sub-Theme 2.3: Associated organisational risks.....	145
6.4.4 Sub-Theme 2.4: Consequences of violating privacy in Apache Hadoop.....	146
6.5 Theme 3: Privacy protection.....	149
6.5.1 Sub-Theme 3.1: Impact on core organisational values.....	149
6.5.2 Sub-Theme 3.2: Methods to protect privacy.....	150
6.6 Conclusion.....	153
Chapter 7.....	158
A Framework for Reducing the Impact of Privacy Concerns of Big Data in Hadoop Technology at the Automotive Manufacturer.....	158
7.1 Introduction.....	158
7.2 Critical success factors.....	159
7.3 Composition of the framework.....	164
7.3.1 Contextual conditions.....	165
7.3.2 Rule and regulatory conditions.....	166
7.4 Proposed framework.....	167

7.5 Framework application.....	170
7.6 Conclusion.....	171
Chapter 8	173
Conclusion and Recommendations.....	173
8.1 Introduction	173
8.2 Theoretical framework	173
8.3 Research methodology	174
8.4 Research objectives	175
8.5 Research contribution.....	180
8.6 Research evaluation.....	181
8.7 Limitations and recommendations for future research	182
8.8 Conclusion.....	182
9. References	184
List of Abbreviations and Acronyms.....	201
Glossary	203
Appendix A: Ethical Clearance.....	204
Appendix B: Overview and informed consent.....	206
Appendix C: Interview and questionnaire.....	208



University of Fort Hare
Together in Excellence

List of Figures

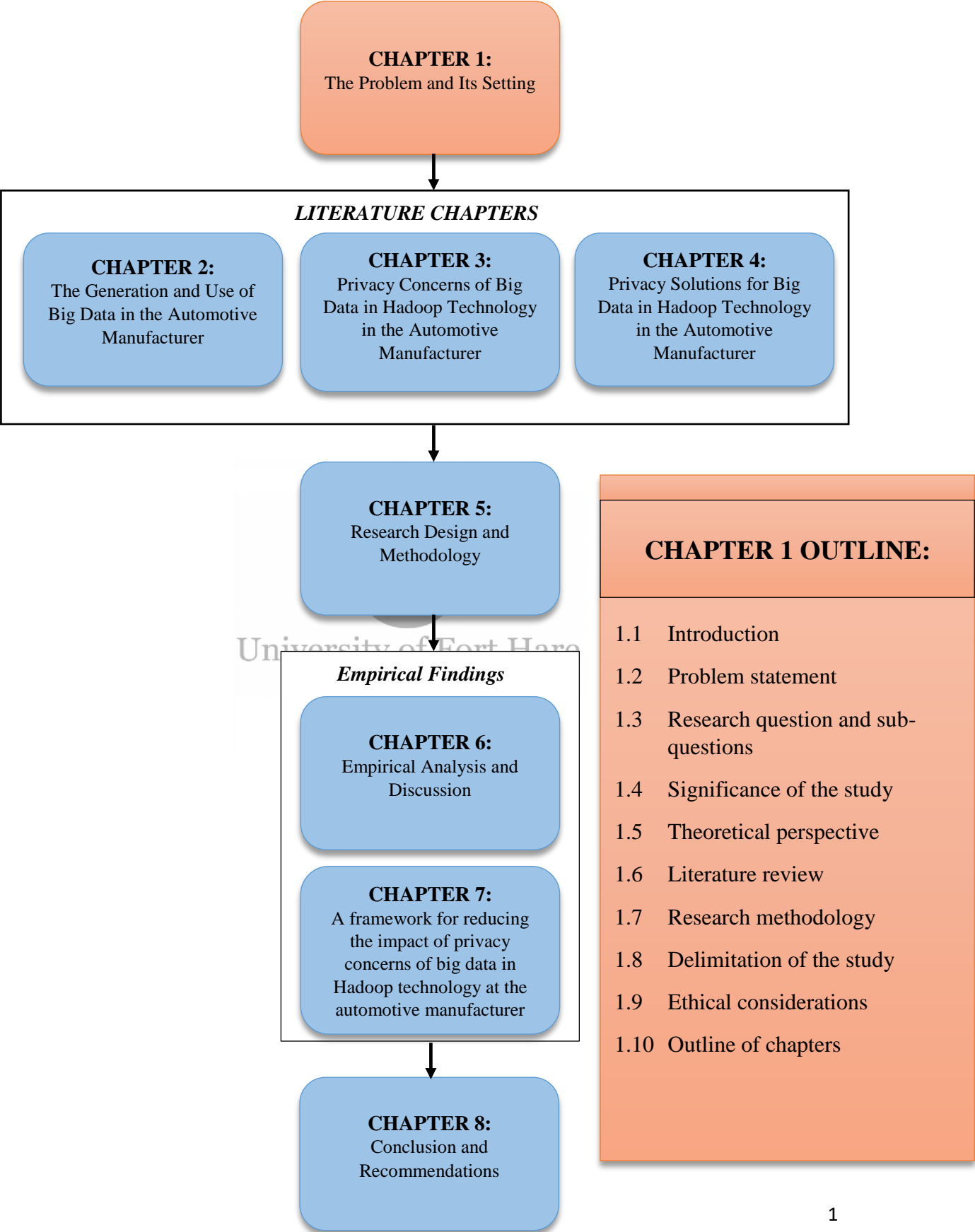
Figure 1.1: Selective organisational information privacy and security violations model (Wall et al., 2015)	10
Figure 1.2: Data collection and analysis process	25
Figure 1.3: Outline of chapters	28
Figure 2.1: Common characteristics of big data	43
Figure 2.2: Summary of process between big data components	55
Figure 3.1: Key features of Apache Hadoop (Huawei, 2021)	61
Figure 3.2: Classification of a data breach (Frankenfield, 2019; Martin, 2019)	67
Figure 3.3: Selective organisational information privacy and security violations model (Wall et al., 2015)	71
Figure 5.1:Continnum of core ontological assumptions (Collis & Hussey, 2009)	101
Figure 5.2: Information Systems Research Framework (Hevner et al., 2004)	104
Figure 5.3: Five stages of the capability maturity model (CMMI Institute 2017)	108
Figure 5.4: The theory of pattern matching (Trochim, 2001)	116
Figure 5.5: Empirical process phases	121
Figure 6.1: Respondents based on IT role	128
Figure 6.2: Responses to changes in organisation structure	142
Figure 6.3: Responses on privacy awareness	144
Figure 6.4: Responses to the negative internal consequences of violating privacy of big data in Hadoop	147
Figure 6.5 : Responses on violation of big data in Hadoop will face serious consequences	148
Figure 6.6: Responses on privacy of big data and impact on organisational core values	150
Figure 6.7: Response to device encryption	151
Figure 6.8: Response on using different passwords on IoT devices	151
Figure 6.9: Responses on installing manufacturers security updates	152
Figure 6.10: Responses on unambiguity on privacy and security standards	153
Figure 7.1: Framework composition	165
Figure 7.2: Factors of contextual conditions	166
Figure 7.3: Factors of rule and regulatory conditions	167

List of Tables

Table 2.1: Elements of Industry 4.0	32
Table 2.2: Examples of external data sources and types in the automotive manufacturing industry (Zaki et al., 2019)	36
Table 2.3: Examples of internal data sources and types (Zaki et al., 2019)	37
Table 2.4: Big data characteristics identified by authors	40
Table 2.5: Types of big data in an automotive manufacturing environment (Wang & Alexander, 2015)	45
Table 2.6: Types of big data analytics (Deloitte, 2019, Riahi & Riahi, 2018)	47
Table 3.1: Differences between privacy and security (Jain et al., 2016)	64
Table 5.1: Philosophical assumptions of four research paradigms (Adebesin, Gelderblom, & Kotzé, 2011)	99
Table 5.2: Features of the positivist and interpretivist paradigm (Collis and Hussey 2009)	100
Table 5.3: Design Science Research Guidelines (Hevner et al., 2004)	105
Table 5.4: Themes and sub-themes of the study	118
Table 5.5: Quality in Positivist and Interpretivist research (Oates., 2006)	121
Table 6.1: Research participants identity	130
Table 6.2: Questionnaire key and description	146
Table 7.1: Critical success factors to address privacy concerns of big data in Hadoop technology at the automotive manufacturing industry	159
Table 7.2: A framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer	168
Table 8.1: Overview of findings	179



University of Fort Hare
Together in Excellence



Chapter 1

The Problem & Its Setting

1.1 Introduction

Traditional automotive manufacturers are undergoing a revolution due to an emerging data source and associated technologies, namely, big data (Nagy, Oláh, Erde, Máté, & Popp, 2018). Gupta, Gupta and Singhal (2014, p.266) define big data as “data that exceeds the processing capacity of traditional databases.” Big data includes information gathered from social media, mobile devices, machine data, video and voice recordings (Lovalekar, 2014). Syafrudin, Alfian, Fitriyani and Rhee (2018) describe the nature of big data as large, unstructured and fast-moving. As a result, Riahi and Riahi (2018) characterised big data into the five V's: volume, velocity, variety, value and veracity.

An automotive manufacturer's facilities need to manage the complexity of the data produced and use the correct analytical techniques to extract useful information from large datasets (Nagy et al., 2018; Wang & Alexander, 2015). When data is collected in digital form, it is analysed to quicken the planning process and disclose patterns that can be used to improve business strategies (Deloitte, 2019 & Tole, 2013). Big data needs to be processed instantly so that the automotive manufacturer can receive real-time information on vehicle or part failure (Deloitte, 2019). Obtaining this information in real-time and applying big data analytics allows the manufacturer to improve the production process and maintain a competitive position in the automotive market (Ajah & Nweke, 2019; Deloitte, 2015).

Apache Hadoop is used in the automotive manufacturing environment to manage all data effectively. This data management includes cost-efficiently storing, processing, analysing and transforming complex data (Educba, 2020 & Tole, 2013). Apache Hadoop is a popular open-source framework that allows large datasets to be distributed across clusters of

computers using simple programming models (Bhagyashree & Koundinya, 2020). Furthermore, the framework stores and analyses big data through various built-in modules (Ishwarappa, 2015 & Tole, 2013). However, this framework has raised information privacy concerns due to its security vulnerabilities (Jain, Gyanchandani, & Khare, 2016; Yadav, Maheshwari, & Chandra, 2019).

Security refers to various practices and processes which can be implemented to ensure that data cannot be used or accessed by unauthorised individuals or parties (Jain et al., 2016). However, if a security vulnerability is evident, the automotive manufacturer's information privacy is breached as sensitive information is explicitly available for illegal use (Moura & Serrão, 2015).

This research study investigated information privacy concerns at a local automotive manufacturer. Information privacy issues resulted in the manufacturer becoming vulnerable to infiltration from unauthorised users, who can access and modify sensitive information (Frankenfield, 2019). If sensitive information becomes compromised, altered, or disclosed in public without the consent of the automotive manufacturer, this causes there to be an information breach and a loss of information privacy (Tawalbeh, Muheidat, Tawalbeh, & Quwaider, 2020; O'Donovan, Leahy, Bruton, & O'Sullivan, 2015). An information breach or the loss of information privacy is a significant concern as brand reputation, business growth and competition can potentially be lost in the automotive market (Ajah & Nweke, 2019; Data Privacy Manager, 2020).

This study was conducted using the interpretivist approach to understand the problem from a subjective perspective and thereafter provide necessary explanations of the research problem (Clough & Nutbrown, 2012). Primary data was collected from a South African automotive manufacturer, where the qualitative data collection approach was applied. The primary data collection was interviews and questionnaires from ten IT (Information Technology) big data specialists.

The interview and questionnaire aimed to gain further insight into big data, privacy concerns, and big data solutions in Hadoop technology at an automotive manufacturer. A

literature analysis was conducted using books, academic journals, conference proceedings and websites.

Furthermore, the principle of saturation was applied to determine that sufficient information had been collected and no new information would be realised. Thematic analysis was used in the empirical analysis to identify relevant themes of this study. The literature review and primary data were analysed using Microsoft Excel and NVivo to answer the main and sub-research questions. This was completed in the form of a framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer. Design science was applied in establishing the framework, which was the output of this study and was in the form of an artefact that was used to address the research problem.

This chapter outlined big data and addressed the problem statement from which the primary and sub-research questions were derived and presented the study's significance. Furthermore, an introduction to the Information Systems (IS) theory: Selective Organisational Information Privacy and Security Model (SOIPSVM), a short literature review, research methodology, scope of the study, and ethical considerations were discussed. Lastly, an outline of the chapters in this study was presented. The following section identified and addressed the problem statement of this research study.

1.2 Problem statement

Big data makes the product lifecycle in the automotive manufacturing process more efficient, decentralised and well-connected (Chhetri, Rashid, Faezi, & Al Faruque, 2018). Consequently, supply chain management and operations within the automotive manufacturing process become optimised and improved (Deloitte, 2019 & Reidy, 2018). However, automotive manufacturers have realised information privacy concerns due to the complex nature of big data stored and processed on the Apache Hadoop technology (Jain, Gyanchandani, & Khare, 2016; Yadav, Maheshwari, & Chandra, 2019). Automotive manufacturers use Hadoop technology to process, manage, store and analyse big data (Educba, 2020 & Yadav et al., 2019). Due to an ineffective security model present in

Hadoop technology, information privacy concerns have been raised (Bhathal & Singh, 2019; Sharma & Navdeti, 2014).

Information privacy issues, such as data breaches, can make sensitive information easily accessible to illegal users (Frankenfield, 2019). If this problem is ignored, the automotive manufacturer will be at risk because it is more likely that sensitive information can be accessed and tampered with. This can affect production technologies and analytical results (Tawalbeh, Muheidat, Tawalbeh, & Quwaider, 2020; O'Donovan, Leahy, Bruton, & O'Sullivan, 2015). Further, the automotive manufacturer's trade secrets can potentially be accessed due to prevalent privacy issues in Hadoop technology (Summersgill & Coviello, 2019). In effect, the opportunities associated with big data will not be achieved, production levels will be negatively impacted and the automobile will not be of high quality (Bhathal & Singh, 2019; Wang & Alexander, 2015). Ultimately, the company's reputation will be negatively impacted; and business competitiveness and growth will be lost in the automotive market (Ajah & Nweke, 2019; Data Privacy Manager, 2020).

Therefore, it is vital to address the source of information privacy concerns in Hadoop technology, as this will ensure that a holistic solution is implemented should potential information attacks occur (Sharma & Navdeti, 2014; Jain et al., 2016). This will enable the automotive manufacturer to gather, process and analyse the massive amounts of data effectively generated from production technologies. As a result, leveraging big data will ensure a transformation in the automotive production processes (Nagy, Oláh, Erde, Máté, & Popp, 2018). Therefore, privacy concerns need to be effectively addressed to maintain or exceed the competition and business growth in the automotive market. Furthermore, it is essential to ensure that critical information is not compromised (Ajah & Nweke, 2019; Data Privacy Manager, 2020).

However, some automotive manufacturers may instead build their own infrastructure according to their specifications to store and process the data generated from the Internet of Things (IoT) devices (Tole, 2013). The downfall in this method is the automotive manufacturer will need to employ specialists to maintain the infrastructure, and it is incredibly costly. Therefore, making Hadoop a more feasible financial option (Lorant,

2016 & Tole, 2013). *The problem investigated was to identify a solution to address big data privacy concerns in Hadoop technology at a local automotive manufacturer.* The following section defined the main and sub-research questions for this study. Furthermore, each research question was analysed through a short literature discussion.

1.3 Research question and sub-questions

The following research question was formulated:

How can the privacy concerns of big data in Hadoop technology at an automotive manufacturer be addressed?

The main research question was split into three relevant sub-questions:

i) How is big data being generated and used at an automotive manufacturer?

Big data can be generated from internal and external data sources (Ismail, Truong, & Kastner, 2019). Internal sources identified included sensors, Radio-Frequency Identification (RFID) and business applications. In contrast, external data sources included mobile devices, social media and point of sale (Zaki, Theodoulidis, Shapira, Neely, & Tepel, 2019). The large datasets generated from the data sources contain different data types, such as sensor and contextual data (Wang & Alexander, 2015). Big data analytics is applied to the generated datasets to track defects, optimise production processes, improve manufacturing uses cases and advance the automotive manufacturing process (Deloitte, 2019 & Klöser, 2019).

ii) What are the types of privacy challenges experienced within the Hadoop environment at an automotive manufacturer?

Apache Hadoop is a popular big data technology used to store and process large datasets (Bhathal & Singh, 2019). However, several concerns have been raised regarding information privacy due to its flawed security model, which makes it vulnerable to information attacks (Jain et al., 2016). Considering the complexity and technicality of big

data and Hadoop's ineffective security model, information privacy challenges are prevalent in Hadoop technology (Bhathal & Singh, 2019).

Challenges such as data breaches and fragmented data can cause an individual, group or software system to gain unauthorised access and retrieve sensitive information resulting in information theft (Bhathal & Singh, 2019 & Jain et al., 2016). The compromised information is used for fraudulent purposes, which negatively impacts the automotive manufacturer's reputation and competitiveness in the market (Bhathal & Singh, 2019 & Frankenfield, 2019).

iii) What measures can an automotive manufacturer take to protect their privacy?

To protect privacy, Tole (2013, p. 35) states that “some companies choose to build their infrastructure for storage and manipulate the data they have”. The downfall of this method is that it is costly, maintenance by trained personnel is required, and more features are necessary for the infrastructure as the organisation develops (Lorant, 2016). Adequate information privacy can be maintained through the implementation of security and governance mechanisms. Identified security measures included authentication, authorisation, Access Control List (ACLS), encryption and audits (Sharma & Navdeti, 2014; Singh, 2014). In contrast, governance measures were identified to protect the privacy of data stored on Hadoop technology. These measures included: conducting regular reviews of user access to data, data masking, monitoring user behaviour, disaster recovery and backup and tokenization (Deloitte, 2017; Mattsson, 2014; Shacklett, 2016; Simon & Ramesh, 2016). The next section discussed the significance of this research study.

1.4 Significance of the study

Automotive manufacturers have acknowledged that big data and analytics is the new source of revenue and innovation, which has led to disruptive product innovation, new business models and digital services (Dremel, 2017). Big data is the next step towards real-time predictive and efficient manufacturing, leading to high business competitiveness and growth within the automotive manufacturing industry (Klöser, 2019 & Nagy et al., 2018). However, automotive manufacturers have been slow to adopt this emerging technology

because of fundamental changes needed for implementation. The technological shift includes changes to work processes, departmental structure and organisational culture (Dremel, 2017).

The successful adoption of big data and analytics has resulted in an information privacy barrier for automotive manufacturers using Hadoop technology. If information privacy issues remain unaddressed, sensitive information can be compromised, affecting production technologies and analytics to extract useful information (Tawalbeh et al., 2020 & O'Donovan et al., 2015). Therefore, the automotive manufacturer will not realise the associated value of big data, resulting in production target levels and the quality of products not being met. Ultimately, business competitiveness, growth and brand reputation in the market area will be lost (Ajah & Nweke, 2019 & Data Privacy Manager, 2020).

This study was significant as it identified and addressed big data privacy concerns in Hadoop technology through critical success factors extended into a framework. Furthermore, this study investigated associated use cases and big data value in the automotive manufacturer. This study aims to create a framework to answer the main research question and add to the existing knowledge base of the information systems domain. The following section discussed the underlying theory used for this study.

1.5 Theoretical perspective

Theories are constructed to explain, predict and understand a phenomenon. This extends into challenging and expanding upon an existing knowledge base from a critical perspective. A theory allows the researcher to connect the study conducted within the current knowledge base that focuses on a specific subject area. This will provide the researcher with an initial structure to conduct the research study (University of Southern California, 2018). Therefore, to study the research problem, Selective Organisational Information Privacy and Security Violations Model (SOIPSVM) and the Capability Maturity Model (CMM) were used as the two underlying theories. The following section discussed the SOIPSVM and CMM.

1.5.1 The selective organisational information privacy and security violations model (SOIPSVM)

A theoretical framework is “a structure that guides research by relying on a formal theory, constructed using an established, coherent explanation of certain phenomena and relationships” (Grant & Osanloo, 2014, p.13). Theoretical frameworks introduce and describe a theory, explaining why the identified research problem exists and how it can be solved (Grant & Osanloo, 2014). The framework established in this research study aimed to address several privacy concerns of big data in the automotive manufacturer. This was done through SOIPSVM and the five levels of the CMM.

This study used SOIPSVM to comprehensively address privacy concerns of big data usage in the automotive manufacturer. Information privacy concerns are prevalent because the information is easily accessible through the flawed security model present in Hadoop technology (Jain et al., 2016 & Wall, Lowry, & Barlow, 2015). Therefore, information privacy is of primary importance within a connected and information-intensive digital age. Consequently, information privacy and security laws have become ubiquitous within many areas of society (Wall et al., 2015).

However, repetitive privacy and security breaches indicate a failure in current privacy regulations that keep important information secure (Krauss, 2014 & Wall et al., 2015). To address this issue, Wall et al. (2015) proposed a theoretical model called the SOIPSVM, as depicted in **Figure 1.1**.

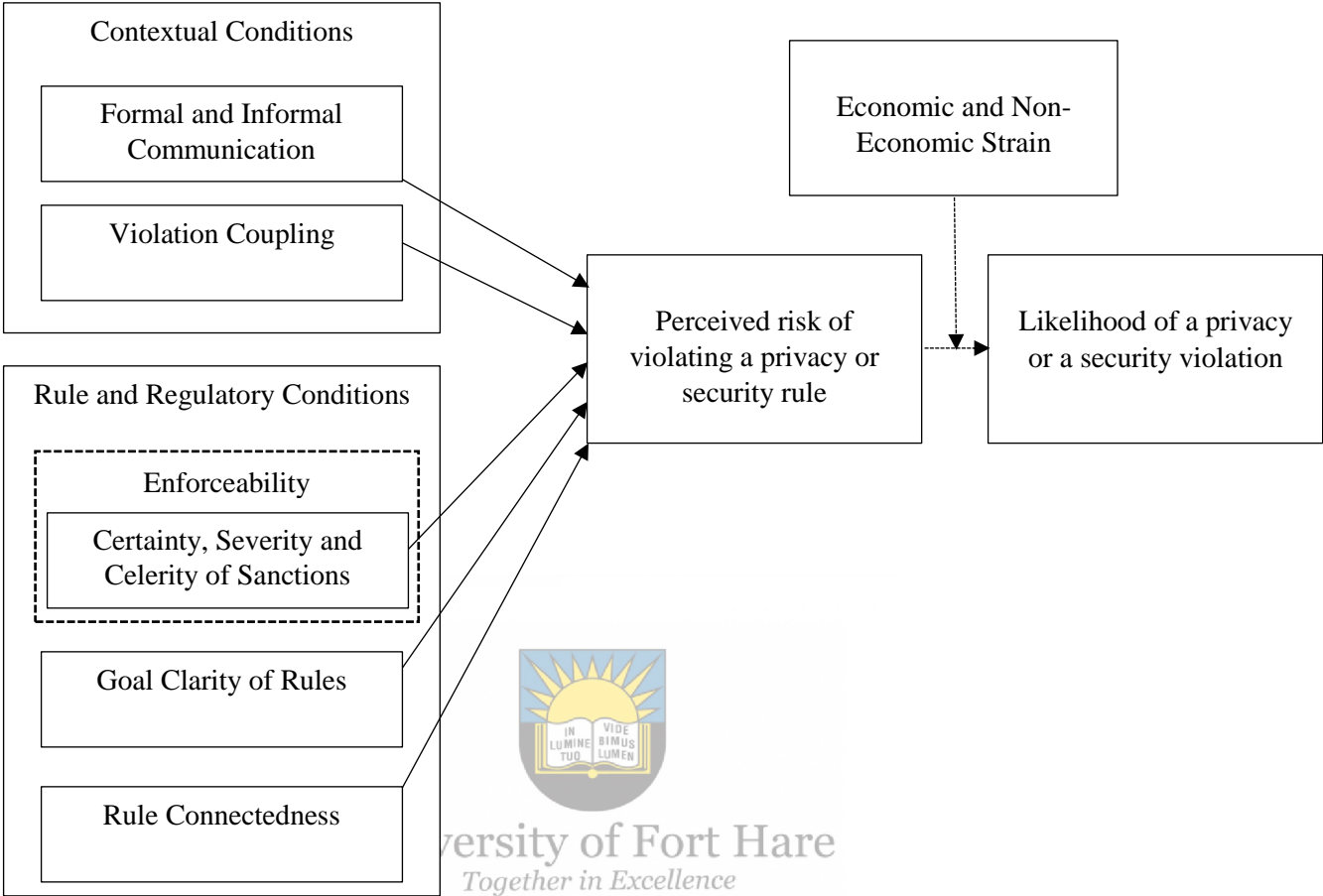


Figure 1.1: Selective organisational information privacy and security violations model (Wall et al., 2015)

The SOIPSVM explains that large organisations such as an automotive manufacturer need to “select an externally governed privacy or security rule for violation in response to organisational strain or slack resources” (Wall et al., 2015, p.1). Externally governed privacy and security rules are concerned with laws, policies and standards relating to privacy and security, compiled by external agencies or collectives. External agencies or collectives enforce these rules through regular monitoring, sanctions and fines (Wall et al., 2015).

Furthermore, according to Wall et al. (2015), it is evident that organisations have succumbed to violating privacy and security rules by intentionally misusing sensitive data

or failing to protect it efficiently. Therefore, SOIPSVM aims to address this issue. Each component of the SOIPSVM was explained briefly per the objective of the theory.

1.5.1.1 Contextual conditions

Contextual conditions consist of formal and informal communication structures and violation coupling (Wall et al., 2015).

1.5.1.1.1 Formal and informal communication structures

Formal and informal communication structures influence the flow of communication and organisational structure, which can either obstruct or maintain a communication flow, impacting the organisation's perception of risk (Wall et al., 2015). According to Wall et al. (2015), formal communication structures comprise bureaucratic structures. Bureaucratic structures consist of fixed procedures, policies and constraints, where an element consists of structural secrecy.

Furthermore, these structures follow a high degree of formality, consisting of centralised and rigid communication flows (Burley, 2018 & Wall et al., 2015). Examples of formal IT communication structures in the automotive manufacturer includes IT policies, upper-level management approval for business transactions and collaboration with other units in the industry, and the extent of decision-making responsibility for capital budgeting and new product introduction (Wall et al., 2015).

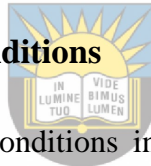
In contrast, informal communication structures are not part of the formal structure of the workplace. Instead, they follow and develop a communicative relationship through a lateral context. (Wall et al., 2015). Examples of informal communication structures include relational networking opportunities with inter-departmental teams. Communication and the organisational structure play an integral role in the success of the automotive manufacturer. It is important for communication channels to be established within and outside the fixed organisational structure (Liebe, Tichy, Knauss, Ljungkrantz, & Stieglbauer, 2018).

1.5.1.1.2 Violation coupling

Lehman and Ramanujam (2009) describe violation coupling as the likelihood that privacy and security violations will lead to a result, which can be “positive such as performance improvement or negative such as regulatory penalties” (Lehman & Ramanujam, 2009, p.648). When the coupling is tight, the connection between the violation and the corresponding outcome can be distinguished, and the automotive manufacturer can predict the result of the breach. Furthermore, when the violation outcome is positive, the automotive manufacturer controls the consequences and is less risky (Wall et al., 2015).

Violations linked to a negative outcome and associated behaviour can result in fines and the automotive manufacturer’s corporate image loss. The negative consequences increase risk perception and decrease the automotive manufacturer’s control over the result (Lehman & Ramanujam, 2009).

1.5.1.2 Rule and regulatory conditions



Elements of rule and regulatory conditions include certainty, severity and celerity of sanctions conceptualised into enforceability. Furthermore, other factors include goal clarity of rules and rule connectedness (Wall et al., 2015). These conditions need to be implemented as rules to regulate the organisation's conduct (Law Insider, 2018).

1.5.1.2.1 Enforceability

In the context of SOIP SVM, enforceability is referred to as how well behaviour, in correlation to detection and sanction, is monitored together with the adverse effects. According to the SOIP SVM, the enforceability element comprises certainty, severity, and celerity of sanctions. External regulatory bodies put these sanctions in force when an organisational privacy or security rule violation occurs (Wall et al., 2015).

The automotive manufacturer must apply enforceability by regularly auditing company policies, including the extremity and response time to violations. Furthermore, regulatory agencies must continuously monitor the automotive manufacturer’s actions and share an interdependent relationship. (Wall et al., 2015).

1.5.1.2.2 Goal clarity of rules

It is important for the automotive manufacturer to clearly state all rules and policies without ambiguity to prevent confusion on rule outcomes. Goal clarity refers to regulations that contain objectives that do not inhibit ambiguity and alternative interpretations. Also, the rules provide more detail on how to attain the objectives. Therefore, when the clarity of a rule's goal is refined, the rule's expected results are defined (Wall et al., 2015).

1.5.1.2.3 Rule connectedness

Information privacy and security rules have a high-level rule connectedness. This means that they are associated with a vast number of rules. An increase in privacy and rule connectedness will increase an organisation's perceived risk (Wall et al., 2015). Therefore, the automotive manufacturer must clearly define all rules which are connected to other rules or policies, as this can determine whether the automotive manufacturer complies with other security rules or whether a violation of a rule results in violation of other rules at the same time (Wall et al., 2015 & Law Insider, 2018).

1.5.1.3 Perceived risk of violating privacy and security

The conditions discussed above have a direct impact on the organisation's perceived risk of a violation occurring. The SOIPSVM indicates that an increase in the automotive manufacturer's perceived risk of violating an externally governed privacy or security rule will decrease the likelihood of a rule violation. However, the automotive manufacturer should establish the organisation's risks should severe sanctions be issued due to a violation (Wall et al., 2015).

1.5.1.4 Economic and non – economic strain

The economic and non-economic strain influences the perceived risk of violating a privacy and security rule and the likelihood of a breach. Economic strain, as per SOIPSVM, relates to strain in the automotive manufacturer's performance. The automotive manufacturer's performance in comparison to competitors provides an operationalised economic strain. As a result, this can affect organisational behaviour (Wall et al., 2015). In contrast, non-

economic strain refers to the failure to achieve desired goals and the incapacity to carry out core values. This results in devised solutions; however, if solutions are unavailable, non-compliant behaviour occurs (Wall et al., 2015).

1.5.1.5 Likelihood of a privacy or security violation

The likelihood of privacy or a security violation refers to the impact of the automotive manufacturer's elements mentioned above. These elements and the regulatory environment may result in the organisation violating externally governed rules to protect its privacy and security of sensitive information. The automotive manufacturer can assess their environment by establishing whether core functions of the business can or cannot be practised should a privacy or security role be followed. (Wall et al., 2015).

1.5.1.6 Summary of the selective organisational information privacy and security violations model in the automotive context

In summary, the SOIPSVM describes how organisational structures and processes, including contextual conditions and characteristics of regulatory rules, can modify the automotive manufacturer's understanding of risk. The elements of contextual conditional include formal and informal communication structures and violation coupling (Wall et al., 2015).

The difference between the two communication structures is the formal communication structure includes IT policies and a rigid communication flow within the automotive organisational structure. In contrast, informal communication structures entail a natural form of communication and interaction amongst employees of other divisions within the automotive manufacturer (Wall et al., 2015 & Burley, 2018).

Violation coupling refers to the probability of a privacy or security violation occurring, causing a positive or negative result. The positive outcome is the automotive manufacturer's ability to speculate and control consequences due to the tight coupling between the violation and result (Lehman & Ramanujam, 2009). Negative outcomes cause the automotive manufacturer to increase risk, which decreases control over the result. The

impact of a negative outcome is severe and can result in the automotive manufacturer obtaining sanctions, which can cause damage to the company's corporate image (Wall et al., 2015).

Rule and regulatory conditions include enforceability, goal clarity of rules and rule connectedness. Enforceability refers to the regularity of the automotive manufacturer auditing company policies and the consequences due to non-adherence. The automotive manufacturer's regulations and policies must not contain ambiguous information, which can cause various interpretations by employees at the automotive manufacturer. The objectives of the policies and regulations must clearly be stated to meet the goal clarity element. Lastly, rule connectedness entails outlining all rules related to other company rules or policies (Wall et al., 2015).

The purpose of defining rule connectedness is to ensure that the automotive manufacturer and employees abide by the regulations and decipher whether a rule violation can result in several related rule violations (Wall et al., 2015). Non-compliance with the mentioned conditions results in a privacy and security rule violation occurring in the automotive manufacturer. This can be influenced by economic or non-economic strain. Economic strain impacts the automotive manufacturer's production volumes which can affect organisational behaviour.

Non-economic strain is concerned with not attaining goals and the inability to meet the core values of the automotive manufacturer. As a result, when the automotive industry's performance levels do not correlate with its target level, the effect is that there is a likelihood of rule violation occurring (Wall et al., 2015). The following section discussed studies that have used the SOIPSV theory and the appropriateness of choosing SOIPSV as the underlying theory for this study.

1.5.1.7 Appropriateness of the selective organisational information privacy and security violations model theory for this study

Several studies have incorporated SOIPSV as the theory used to solve a research problem. In a study by Dorasamy, Haw and Vigian (2017), it was identified that the rise in

the IoT had worsened security and information privacy challenges, increasing cybercrime incidents. The main reason for the increasing cybercrime incidents was due to organisations and individuals violating regulations and rules. The study used the SOIPSVM to understand the organisational attitude and behaviour towards the contextual, rule and regulatory conditions. The perceived risk of violating a privacy or security rule was hypothesized to determine the likelihood of a privacy and cybersecurity rule violation. The authors of the study aimed to eliminate the adverse effect of IoT by building a privacy and cybersecurity model (Dorasamy, Haw, & Vigian, 2017).

Ahmadu, Hussin and Bahari (2021) studied the leakages of classified information at institutions of higher learning due to the lack of security. Existing security measures were ineffective due to poor regulatory enforcement, communication structure, and human behaviour. It was found that information security can be improved upon through the enforceability of rules and regulations. The study used SOIPSVM to develop and validate a security violation prevention model to mitigate the leakage of sensitive information at higher-level institutions (Ahmadu, Hussin, & Bahari, 2021).

Similarly, the SOIPSVM was appropriate for this study as the model aimed to explain the violation of privacy and security rules and the behaviour of organisations (Wall et al., 2015). In the context of the automotive manufacturer, there were vulnerabilities experienced in Hadoop technology which was used to store and process big data. Due to privacy and security violations, sensitive information was easily available to be illegally misused.

Privacy issues of big data in Hadoop technology need to be addressed using effective measures. If this is not done, the automotive manufacturer may experience competitive loss and a decline in business growth (Ajah & Nweke, 2019 & Wang & Alexander, 2015). This study incorporated a qualitative data collection approach, a literature review, and two theoretical perspectives: SOIPSVM and CMM to address the research problem.

Measures to address information privacy concerns were identified as critical success factors. According to contextual and rule and regulatory conditions, factors were

categorised and expanded upon, which formed part of the SOIPSVM model's components. Once factors were categorised, they were aligned to the five levels of the CMM. As a result, a framework was constructed to address the identified factors at different levels of maturity. Assessing the elements at different levels of maturity provided a better solution as the aim was to improve at each level of maturity. The next section discussed the literature of the identified sub-research questions.

1.6 Literature review

1.6.1 Introduction

Delgado (2017) states that IT has played a significant role in the automotive manufacturing revolution. Previously, the production process was manual and too tedious (Delgado, 2017). However, Delgado (2017) and Klöser (2019) state that the IoT and digitalisation have resulted in the automation of processes and various forms of mechanisation, resulting in large datasets generated from internal and external IoT sources. Along with the value that big data provides to the automotive industry, several information privacy issues in the environment need to be addressed (Jain et al., 2016).

This brief literature review considered the generation and use of big data in the automotive manufacturer, followed by a discussion on information privacy concerns of big data in Hadoop technology. Finally, measures to address the information privacy concerns of big data in Hadoop technology at the industry were briefly examined. The following section discussed the generation and use of big data in automotive manufacturers.

1.6.2 Generation and use of big data in the automotive manufacturer

Sources of big data in the automotive environment are either external or internal. External sources focus on the consumer environment and include social media, mobile devices, and sensors (Zaki et al., 2019). In contrast, internal sources focus on the automotive manufacturing environment, including machine-to-machine interaction, smart meters, and business applications (Wang & Alexander, 2015; Zaki et al., 2019).

Automotive manufacturing operations and systems produce continuous data streams such as sensor, event, contextual, test, product failure, product, and process performance data from these sources (Lee, Lapira, Bagheri, & Kao, 2013; Wang & Alexander, 2015). The generated data is complex, and as a result, Riahi and Riahi (2018) have characterised big data into the five V's: volume, velocity, variety, value, and veracity.

Due to the complexity of the datasets generated, automotive manufacturers require next-generation data management, integration, processing systems, and storage, allowing the firm to collect, manage, store, and analyse data quickly, efficiently, and cost-effectively. As a result, the automotive manufacturer uses Apache Hadoop (Bhathal & Singh, 2019). Big data technologies are driven by advanced data analytics used to analyse the datasets and provide insight from the past and the future (Ajah & Nweke, 2019 & Deloitte, 2019). The next section discussed privacy concerns of big data in Hadoop technology in the automotive manufacturer.

1.6.3 Privacy concerns of big data in Hadoop technology in the automotive manufacturer



University of Fort Hare

Hadoop technology is an open-source software framework and an economical method of storing large volumes of various data types compared to traditional approaches. This big data technology contains extensive processing power and can control unlimited concurrent tasks (Educba, 2020; Gutierrez, 2015). Hadoop is scalable, fault-tolerant, flexible, low cost, distributable, and can store and process large and various data types (Wróbel & Wikira, 2019). Hadoop is used as a storage mechanism to store millions or billions of data transactions (Educba, 2020). Hadoop is significant for the automotive manufacturer as it enables data patterns to be effectively monitored to prescribe instructions on managing the data (Gutierrez, 2015).

Previous patterns identified will not match new patterns, as Hadoop is updated continuously with fresh data streams (SAS, 2022). Further, Hadoop technology needs to know what needs to be communicated and act accordingly (Tole, 2013; Gutierrez, 2015). In effect, Hadoop will lead to new opportunities, such as vehicle design and production

improvements and improved quality. However, due to the complexity of big data and the ineffective security model in Hadoop, several concerns are raised regarding information privacy, making it susceptible to many threats in the automotive manufacturer (Bhathal & Singh, 2019; Jain et al., 2016).

Security is a concern in Hadoop technology, because the infrastructure was built with a poor security model (Moura & Serrão, 2015). Security issues arise due to the large volume of data stored, which is not encrypted (Chu K., 2020). This means that Hadoop would be easily susceptible to cyber-attacks, as the infrastructure does not contain encryption at a storage or network level (Abualkishik, 2019). Furthermore, Hadoop utilises Java which is programming language that is largely exploited (Chu K., 2020). Therefore the presence of a poor security model in Hadoop can result in sensitive data being compromised leading to an information privacy violation (Jain et al., 2016).

Due to the flawed security model present in Hadoop technology, automotive industries that use the technology are at high risk (Bhathal & Singh, 2019; Sharma & Navdeti, 2014). Furthermore, big data clusters contain fluid data which can be replicated across numerous servers (Bhathal & Singh, 2019). This results in the data becoming available for fragmentation, which is considered a data privacy risk associated with Hadoop technology (Bhathal & Singh, 2019; Lovalekar, 2014).

Due to rapid technological advancements such as big data and Hadoop technology, the automotive manufacturer has become more susceptible to information theft concerning products produced, especially high-priced or high-demand items (Tasmin, Rahman, Jaafar, Hamid, & Ngadiman, 2020). Therefore automotive manufacturers have become victims of theft of trade secrets and intellectual property via corporate spies (Summersgill & Coviello, 2019). Furthermore, the industry can lose its brand reputation and experience a loss of competition in the automotive market (Ajah & Nweke, 2019; Data Privacy Manager, 2020). The next section discussed solutions that could be implemented to address big data privacy concerns in Hadoop technology.

1.6.4 Solutions to address privacy concerns of Hadoop in the automotive manufacturer

Sharma and Navdeti (2014) suggest that there needs to be a holistic solution to solve big data privacy challenges in Hadoop technology, due to the technology having a poor security model (Moura & Serrão, 2015). Security measures can be implemented to protect the privacy of big data in Hadoop technology. Identified measures included authentication, authorisation, ACLS encryption and audits (Sharma & Navdeti, 2014; Singh, 2014).

The most common solution is to encrypt all data. Encrypting the data with a personal key identifier allows the data to be indecipherable for unauthorised persons (Yadav et al., 2019). However, if required to read or analyse the data, one must use the same software to encrypt it (Tole, 2013). Further measures which were identified to protect the privacy of Hadoop technology included: conducting regular reviews of user access to data, application of data masking to sensitive data, monitoring user behaviour, application of tokenization to secure data and implementation of disaster recovery and backup plan.

Alternatively, for automotive manufacturers to gain information privacy, Lorant (2016) and Tole (2013) have suggested that some automotive industries build their own big data infrastructure. However, the big data platform is extremely complex and evolving continuously. Therefore all the required knowledge for big data and privacy in Hadoop technology must be obtained and be kept up to date (Abualkishik, 2019). The next section discussed the research methodology applied in this study.

1.7 Research methodology

Marczyk, DeMatteo and Festinger (2005) describe research design as a blueprint of a research study and its application to the research project. Furthermore, the strategy contains a comprehensive description of the objectives to answer and solve the research questions identified (Saunders, Lewis, & Thornhill, 2007). In contrast, a methodology is the steps to investigate the research problem and apply techniques to identify, select, process and critically analyse information to understand the identified research problem. Furthermore,

how the data is collected and analysed must be considered (University of Southern California, 2018).

In summary, the research design strategy and methodology must be chosen coherently to support the research problem's outcome and solution. This helps strengthen academic work and writing (Marczyk et al., 2005; Saunders et al., 2007). The research design used in this study is design science. Design science is applicable, as this study's output is an artefact in the form of the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer (Marczyk et al., 2005).

1.7.1 Paradigm and approach

Kuhn (1962) first introduced the research paradigm to establish a research culture based on a basic set of beliefs, assumptions, and values that researchers have in common regarding nature and how research is conducted. Furthermore, Saunders et al. (2007) state that a paradigm is a pattern, structure and framework of academic ideas, values and assumptions for theory and research. It contains assumptions, representations and techniques to answer and produce good quality research (Saunders et al., 2007). According to Clough and Nutbrown (2012) and Saunders et al. (2007), research can be categorised into polarisations. These include qualitative, quantitative, positivism or interpretivism.

This research study was conducted using the interpretivist approach. According to Clough and Nutbrown (2012), interpretivism emphasises investigating phenomena to gain a perspective and understand and interpret events, experiences, and social structures. Furthermore, this paradigm allows the researcher to analyse experiences and establish constructive meaning from them. This enables knowledge and theories to be developed from observations and interpretations to be made on a phenomenon (Saunders et al., 2007). In summary, interpretivism allows researchers to understand subjective reality and then provide meaningful explanations (Clough & Nutbrown, 2012; Saunders et al., 2007). This study incorporated a qualitative data collection method. The next section discussed data collection methods.

1.7.2 Data collection methods

Data collection refers to collecting information from sources to discover an answer to the research problem and evaluate it (Salkind, 2010). Data collection involves two types of data: primary and secondary data (Saunders et al., 2007).

Primary data is described as an original data source, whereby the data is gathered first-hand by the researcher to fulfil a research study (Saunders et al., 2007). Furthermore, primary data can be collected through quantitative, qualitative or mixed methods. However, it is considered more costly and time-consuming than secondary data (Clough & Nutbrown, 2012).

The quantitative approach to data collection is concerned with numerical data analysis through polls, questionnaires, and surveys. Furthermore, computational techniques can manipulate statistical data (Saunders et al., 2007; University of Southern California, 2018). Secondly, the qualitative data collection method provides a platform for exploratory research to understand reasons and opinions, achieve insights and help construct ideas (Clough & Nutbrown, 2012). Lastly, according to Clough and Nutbrown (2012) and Saunders et al. (2007), the mixed methods approach is used in research by collecting, analysing and combining quantitative and qualitative research.

This research study incorporated the qualitative approach. The researcher aimed to get an in-depth understanding of perceptions and motivations to develop the a framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer, in response to the research problem. The primary data collection method employed by this study was interviews. The interviews included IT specialists' perspectives on the big data phenomenon, privacy concerns and big data solutions in Hadoop technology at the automotive manufacturer. Furthermore, respondents were requested to complete a questionnaire. The questionnaire contained pre-determined statements, where each respondent had to provide a judgement based on a rating.

In contrast, Johnston (2014) explains that secondary data is a type of data that has been collected by someone other than the user. Furthermore, secondary data allows a re-analysis

and re-interpretation of existing research that has been conducted (Johnston, 2014). This enables researchers to test new theories and frameworks. Therefore, this data is utilised by an investigator for a research purpose. Crossman (2016) explains that using secondary data provides a feasible option for researchers with limited time and resources to conduct their research. Furthermore, current literature was discussed in the form of secondary data (Saunders et al., 2007). This study critically reviewed primary data and secondary data, which enabled the researcher to understand the topic in-depth and propose a solution to the research problem.

The collection and use of primary and secondary data resulted in developing a theoretical framework in this study. Primary data was collected through interviews and questionnaires. Secondary data used included printed media, such as books, academic journals, and conference proceedings. Additionally, online media such as electronic journals, industry white papers and reports were examined. Using these forms of primary and secondary data helped answer the identified research question and sub-questions. The next section discussed the data analysis methods used in this study.

1.7.3 Data analysis methods

Clough and Nutbrown (2012) and the University of Southern California (2018) advocate that data analysis methods assess data using analytical and logical techniques to investigate constituents of the data obtained. The analysis phase occurs after the data has been gathered and reviewed from the relevant source to determine the research problem (Clough and Nutbrown 2012, Saunders et al., 2007). An analysis was conducted from the literature review of the underlying study. The literature content was used in conjunction with the primary data that had been collected. This study entailed a qualitative approach. Therefore inductive reasoning was used to develop a logical conclusion (Clough & Nutbrown, 2012).

Consequently, primary and secondary data were collected and analysed using the inductive reasoning technique. Data collected from the interviews and questionnaires with relevant experts were analysed using Microsoft Excel and NVivo 11 software. These two software tools were used to present a graphical representation of the primary data analysis. **Figure**

1.2 illustrates a summary of the data collection and analysis process conducted in this study.

Pattern matching was used in this study as an analysis method. According to Sinkovics (2018), pattern matching is the most advised technique for analysing qualitative data. Web Centre for Social Research Methods (2006) indicates that pattern matching involves comparing an observed empirical-based pattern to a predicted theoretical pattern. Once compared and the patterns co-exist, the result will strengthen the internal validity of the research project. Due to the concept of pattern matching, Sinkovics (2018) argues that the application of pattern matching results in a meticulous and structured research process, as it incorporates systematic planning and conceptualisation. Furthermore, thematic analysis was applied following the six phased methods defined by Braun and Clarke (2006). The six phases of thematic analysis include: familiarising yourself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report.



University of Fort Hare
Together in Excellence

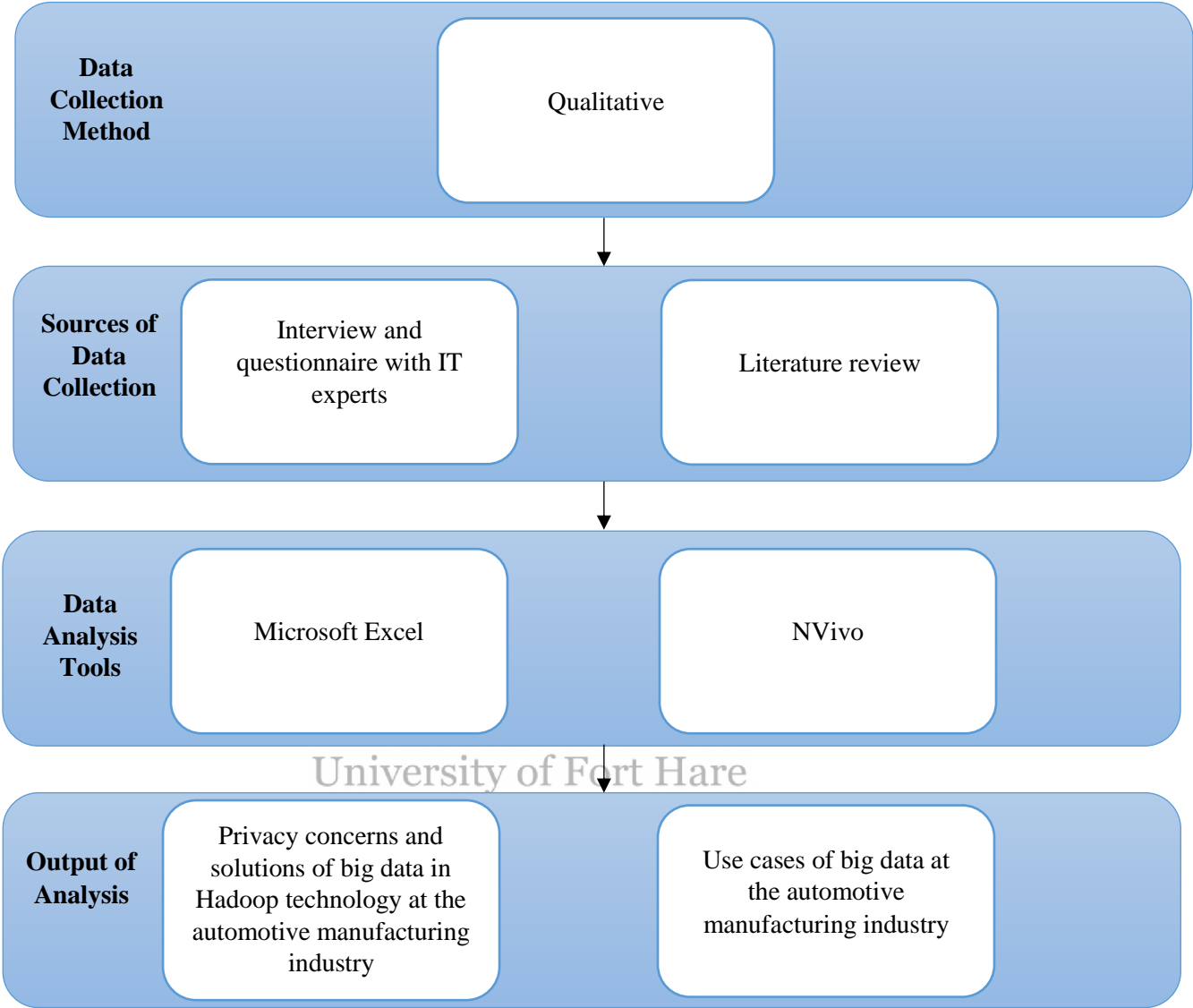


Figure 1.2: Data collection and analysis process

1.7.4 Population and sample

The collection of data from interviews and questionnaires was conducted in a South African based automotive manufacturer. Interviews were conducted within the IT division. The interview population consisted of 67 members from the IT Department. However, ten individuals were part of the sample size. Respondents specialising in big data and Hadoop roles such as the system administrator, business analyst, graduate students, data scientist,

data engineer, IT manager, software developer, and solutions architect were part of the primary data collection process.

The principle of saturation was adopted when conducting the interviews. According to Saunders, Sim, Kingstone, Baker, Waterfield, Bartlam, Burroughs, and Jinks (2018), the principle of saturation is a criterion used to show that sufficient information has been collected to duplicate the research study, and no additional new information has been obtained. This means that once data collection and analysis are conducted, additional data collection and analysis can be discontinued. To achieve the methodological principle, it is essential to devise the number of interviews to be undertaken to reach the principle of data saturation (Saunders, et al., 2018). However, Fusch and Ness (2015) indicate that failure to obtain data saturation impacts the research study's quality and validity. After the ten interviews and questionnaire were conducted, it was clear that no new information had been obtained, and therefore the principle of saturation was achieved. Thus, the principle of saturation was a significant factor in the underlying research study. The following section examined the delimitations of this research study.

1.8 Delimitation of the study

This research study focused on identifying and addressing big data privacy concerns within Hadoop technology in a locally based automotive manufacturer. The study focused on the impact big data has on the automotive manufacturing process. The privacy concerns relate to production (vehicle) data generated from IoT devices connected to the automotive production line and does not concern employees, customers or sales data. Therefore, the Protection of Personal Information Act (POPIA) is not considered, as personal information was not in the scope of this study. The targeted users of this study included the IT department and management at the company. The following section considered ethical considerations that needed to be accounted for in completing this research study.

1.9 Ethical considerations

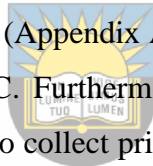
Ethics is regarded as applying morals and standards to collect, analyse, report, and publish a research subject (Saunders et al., 2007). This is essential is due to the involvement of different people from various disciplines. Furthermore, sensitivity to differences needs to be respected (Saunders, et al., 2018). Ethics such as trust, respect and accountability are included in morals and standards (University of Southern California, 2018). Ethics specific to this research study included:

- All data collected remained anonymous.
- The experts interviewed remained confidential.
- Experts had the right to volunteer to participate and withdraw from the research study.

Ethical clearance from the University of Fort Hare's Research Ethics Committee (UREC) was obtained before collecting data (Appendix A). This research study complied with the ethical regulations stated by UREC. Furthermore, management approval was obtained from the automotive manufacturer to collect primary data. The following section outlines the chapters of this research study.

1.10 Outline of chapters

The outline of this research study is depicted in **Figure 1.3**.



University of Fort Hare
Together in Excellence

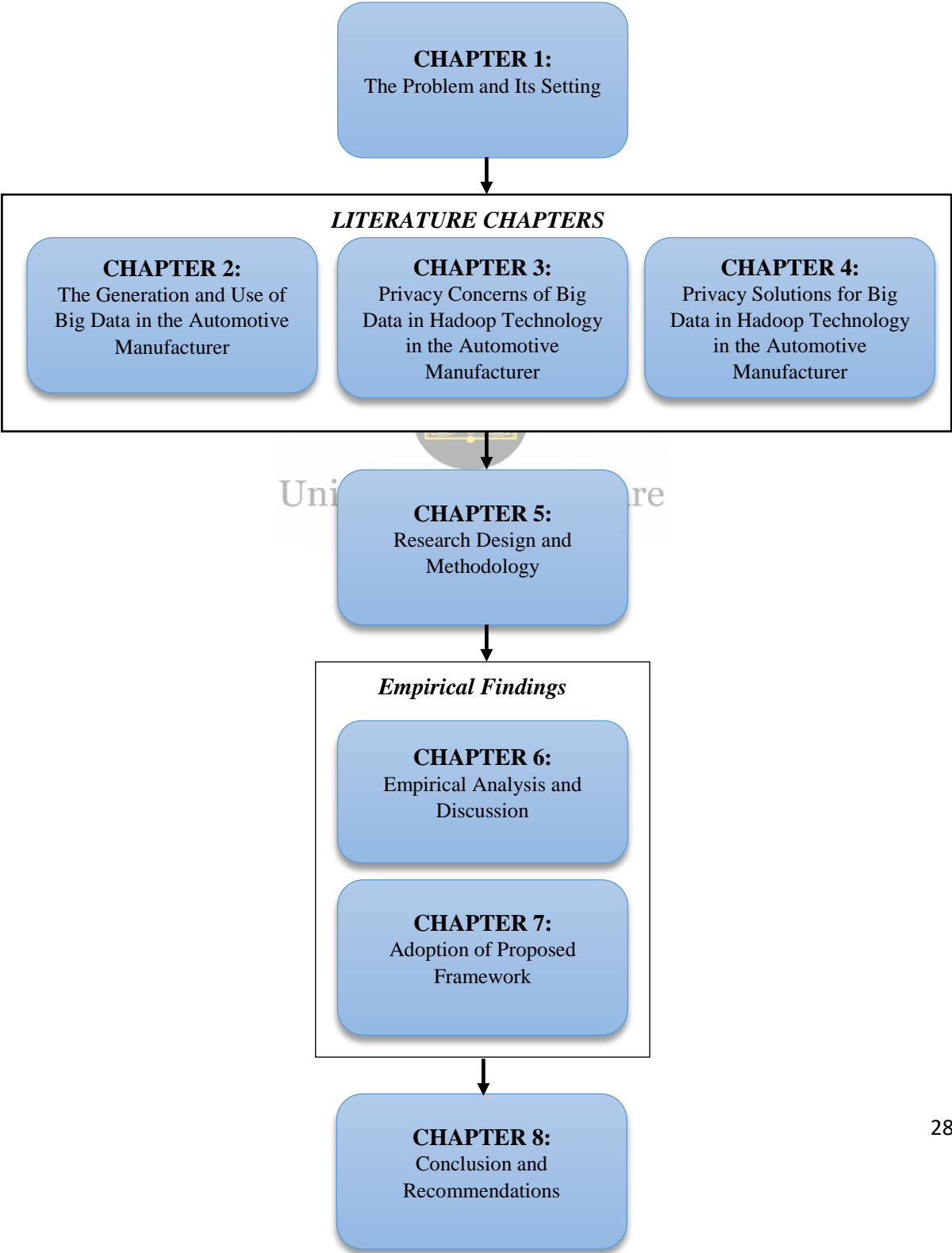
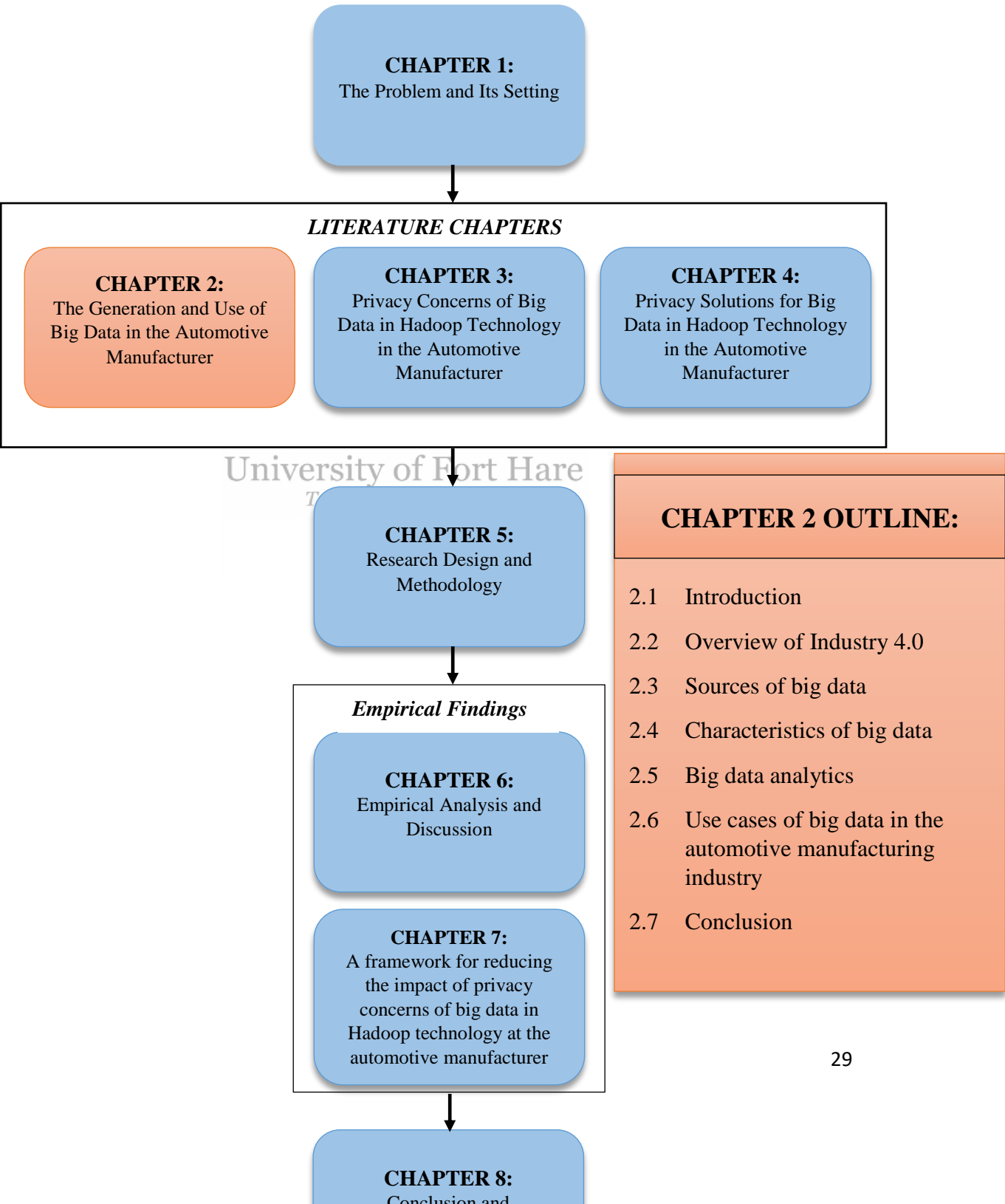


Figure 1.3: Outline of chapters



Chapter 2

The Generation and Use of Big Data in the Automotive Manufacturer

2.1 Introduction

The automotive manufacturing industry previously involved slow and tedious production processes, which produced a few units at a time (Delgado, 2017). Furthermore, according to Ploner (2013) and Müller and Fay and vom Brocke (2018), obtaining data was time-consuming and costly because it was collected from the meter readings of production processes as a handwritten record. However, the development of the assembly line and technological change initiated the manufacturing revolution, which resulted in the automotive industry undergoing a profound technological transformation, otherwise known as the fourth industrial revolution (Oracle, 2015; Delgado, 2017).

Modern automotive manufacturing industries observe significant influences from big data and analytics, resulting in improved operational efficiency, automation, effectiveness, and profitability for the sector (Infor, 2015; Nagy et al., 2018). The automation of processes and numerous forms of mechanisation have resulted in large volumes of data generated from data points in machines and various sources (Delgado, 2017). Klöser (2019) and Tole (2013) indicate that when data is collected in digital format, it will develop manufacturing industries to a new level by optimising operations. The large quantities of data generated need to be examined through big data analytics, including tools to sort and analyse the data. Once the data has been analysed, patterns can be discovered to improve the organisation's strategies (Deloitte, 2019).

Hadoop technology is an open-source framework used to process, manage, store, and analyse big data (Educba, 2020; Yadav, Maheshwari, & Chandra, 2019). Although big data provides opportunities and a competitive advantage in the automotive manufacturer,

privacy concerns have been raised regarding Hadoop technology. Privacy concerns have been raised within the technology due to its flawed security model (Bhathal & Singh, 2019; Sharma & Navdeti, 2014). Privacy issues in Hadoop technology result in compromised essential information by unauthorised users, company reputation damage and possible loss of competitive advantage (Ajah & Nweke, 2019; Data Privacy Manager, 2020; Jain, Gyanchandani, & Khare, 2016).

The purpose of this study was to develop a framework that addressed the underlying research problem. Therefore, to address the main research problem, it was important to understand the types of data sources, the complexity in the nature of the data, and realise the associated use cases that benefit the industry. As a result, this chapter provided an overview of Industry 4.0. Furthermore, it discussed how big data is generated from various Internet of Things (IoT) sources, including the characteristics that make up the data, the use and need of advanced data analytics to extract meaningful information from datasets and uses cases of big data in the automotive manufacturer.

2.2 Overview of Industry 4.0




University of Fort Hare

Automotive manufacturing production systems are currently undergoing a digital transformation to incorporate various enabling technologies. These technologies allow for an intelligent, well-connected, decentralised, and flexible production lifecycle (Chhetri, Rashid, Faezi, & Al Faruque, 2018; Machado, Winrotha, & Ribeiro da Silva, 2019). This era of technological advancement is referred to as the fourth industrial revolution or Industry 4.0 (Machado et al., 2019).


Industry 4.0 functions through the use of advanced technologies (Machado et al., 2019). As a result, Tay, Chaun, Aziati and Ahmed (2018) identified enabling technologies as elements of Industry 4.0. These elements and their use in the automotive manufacturing industry are represented in **Table 2.1**.

Table 2.1: Elements of Industry 4.0

(Engineering Simulation and Scientific Software, 2021; Toshiba, 2020; Deloitte, 2019; Machado et al., 2019; Beamler Additive Manufacturing, 2018; Deloitte, 2017; Wasmund, 2017; Tay et al., 2018 & Deloitte, 2015)

Elements of Industry 4.0	Use in the automotive manufacturer
Autonomous Robots	
<ul style="list-style-type: none"> Robots will interact with each other and collaborate with humans. Grasp complicated tasks, learn from mistakes and improve performance. 	<p>Autonomous robots can optimise the automotive supply chain process by efficiently transporting vehicle parts between the warehouse and production floor. This enables the employees to focus on value-added tasks, reduces error, and optimises vehicle parts' picking and sorting process in the warehouses.</p>
Simulation	
<ul style="list-style-type: none"> Operators can simulate a machine setting in a virtual model before implementation into the real world. Decreases machine setup times, improves quality and plant operations. 	<p>Simulations assist the automotive manufacturing industry with producing vehicles to meet stringent requirements.</p> <p>Automotive application simulations enable engineers to deploy innovations quicker while ensuring safety and reliability through digital prototyping and virtual testing.</p>
Internet of Services	
<ul style="list-style-type: none"> Production activities are triggered through data flows to make daily mobility safe, easier, and efficient. Acts as a service vendor to provide services through the internet. 	<p>Automotive manufacturers are producing vehicles with hardware and software components that can be upgraded. The automobile is sensor ready, and the upgrades provide extra intelligence via the Internet of Services.</p>

Internet of Things (IoT)	
<ul style="list-style-type: none"> • Devices connected to the internet. • Enables interaction, collection and exchanging of data over the internet in real-time. 	IoT devices such as network sensors and intelligent devices are used on the automotive manufacturing floor to collect artificial intelligence and analytics data.
Cyber Physical System (CPS)	
<ul style="list-style-type: none"> • Automated systems provide integration of the natural world with cyberspace. 	The CPS is used in the automotive manufacturing industry to collect data and analyse it using artificial intelligence technologies within the cyber or digital environment. The findings are applied back into the physical world to create value add. This solution enables automotive manufacturers to develop a virtual vehicle prototype, which allows them to perform validations of the complex automotive systems to improve quality and productivity.
Cloud Computing	
<ul style="list-style-type: none"> • Machine data will be stored in the cloud storage system. • Enables production systems to be more data-driven. 	Automotive manufacturers are using cloud computing for efficient data processing and storage. Furthermore, risks associated with data loss are reduced.
Additive Manufacturing (AM)	
<ul style="list-style-type: none"> • Enables for small batch production of customised and lighter products. • Reduces logistic costs and stock. 	Additive manufacturing enables automotive industries to customise vehicle assembly tools by improving functionality, productivity, reducing

	weight and cost compared to traditional tools.
Augmented Reality (AR)	
<ul style="list-style-type: none"> • Used as predictive maintenance. • Promote virtual training. 	The automotive manufacturing industry use augmented reality to develop a virtual testing environment that can be extended for more development and innovation. This reduces costs associated with building physical models and operational costs.
Big Data	
<ul style="list-style-type: none"> • Collection and analysis of large and complex datasets from data sources. • Enables real-time decisions. 	Large volumes of data are generated from various machines connected to the automotive production line, known as big data. Big data must be analysed through analytics to discover patterns and improve operations.

University of Fort Hare
Together in Excellence

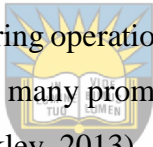
Industries are investing in the characteristics of Industry 4.0 with anticipation to enable processes, machines, employees and products to be integrated into a single network for data collection, data analysis and assessing industry development and performance improvement. Industries that fail to adopt the technology of Industry 4.0 will cause the organisation to lag due to operations not being digitised to match competitors (Nagy et al., 2018).

In the context of the automotive manufacturer, Industry 4.0 is fundamentally changing the developments of automotive production and the value chain. Industry 4.0 opens new opportunities, ensuring the fulfilment of consumer needs, optimising flexibility, and the quality of the automotive production process (Machado et al., 2019). However, to advance in the Industry 4.0 era, the automotive manufacturing industry will need to harness numerous advanced technologies (Nagy et al., 2018).

Digitalisation across the automotive values chain and life cycle generate a plethora of data in real-time. Automotive manufacturers that successfully capture, process, analyse, and store the data can realise exponential growth in opportunities to accelerate along the innovation curve (Dutt, Natarajan, Wilson, & Robinson, 2020). This research study was based on big data in Apache Hadoop and investigated a solution to address information privacy concerns in the Hadoop framework. The next section discussed sources of big data in the automotive manufacturer.

2.3 Sources of big data

The automotive industry has been acknowledged to have a high demand in production rates with a high-efficiency level. This ensures customer satisfaction and competitive costs in the market (Tasmin, Rahman, Jaafar, Hamid, & Ngadiman, 2020). Therefore, automotive manufacturers are poised to utilise big data technologies to capitalise on operational and business data to optimise manufacturing operations and address business needs. As a result, the big data phenomenon represents many promising opportunities and challenges for the automotive industry (Kurtz & Shockley, 2013).



University of Fort Hare

Current automotive manufacturers are equipped with the relevant resources for capturing and tracking data (Infor, 2015). Data can be generated from external and internal sources in the automotive manufacturer. The advancement of technology and digitalisation of the production process has increased data generated from the sources (Ismail, Truong, & Kastner, 2019).

The data generated can provide the automotive manufacturer with the necessary information regarding their customers, products, processes, people, production and equipment (Infor, 2015; Wang & Alexander, 2015). The following section discussed external and internal data sources at the automotive manufacturer in detail.

2.3.1 External sources

External sources focus on the consumer environment and have become more accessible with the advancement of technology (Infor, 2015). The need for satisfying customer

demands, growth and improvements creates instances of re-thinking manufacturing operations (Zaki, Theodoulidis, Shapira, Neely, & Tepel, 2019). This can be done through the generation and analysis of data produced from external data sources. Examples of external data sources and types are shown in **Table 2.2** below.

Table 2.2: Examples of external data sources and types in the automotive manufacturing industry (Zaki et al., 2019)

External (Consumer environment)	
Data source	<ul style="list-style-type: none">• Sensors• Social media• Mobile devices• Public web• Point of sale
Data type	<ul style="list-style-type: none">• Product usage• Customer feedback• Market data

External data sources are used to build customer data. However, anonymity generates higher response rates for data generation. The data generated is used to construct profiles of customers and specific prospects. This includes parameters such as colour and design preferences, common buying triggers or evaluation criteria’s (Infor, 2015). Furthermore, one of the most robust and popular external sources is social media (Infor, 2015). Automotive manufacturers can better understand what current and potential customers prefer in their products (Infor, 2015).

However, Schatsky, Camhi, and Muraskin (2019) discuss external data sources focusing on consumers and material suppliers. Managing big data in the procurement process ensures better manufacturing operations, specifically supply chain management and production planning (Russo, Confente, & Borghesi , 2015). Ismail et al. (2019) add that other external sources include government (incentive programs), strategic partners and distribution channels.

2.3.2 Internal sources

In contrast to external sources, internal sources focus on the automotive enterprise (Zaki et al., 2019). This means that internal data is generated by sources within the automotive production process and includes manufacturing technologies, automated systems, Manufacturing Execution System (MES) and Enterprise Resource Planning System (ERP). Automotive manufacturers can use their systems to capture and analyse the data that has been generated (Infor, 2015). For example, the ERP system can generate data on products, processes, and people in all functional areas of the organisation (Wang & Alexander, 2015). Other data sources and types are represented in **Table 2.3** below.

Table 2.3: Examples of internal data sources and types (Zaki et al., 2019)

Internal (Business environment)	
Data source	<ul style="list-style-type: none">• Sensors• Smart meters• Radio-Frequency Identification (RFID)• Business applications
Data type	Distribute production aims for flexibility, agility, and greater customer orientation, as well as mass customisation.


Lee, Lapira, Bagheri and Kao (2013) state that the development of IoT-based smart sensors has resulted in data generation, and the collection process is straightforward. The automotive manufacturer collects, stores and processes all of the data that the machines generate and utilises smart sensors to increase the range of data (Fox, 2017). The use of smart sensors and the IoT enables data to be collected instantly from machines and the various equipment used in the automotive manufacturing industry (Oracle, 2015).

Smart sensors can effectively be used and integrated with monitoring systems in the automotive industry (Syafudin, Alfian, Fitriyani, and Rhee, 2018). The sensor data is generated in real-time, large amounts, fast velocity, unstructured, and complex (Syafudin

et al., 2018 & Tole, 2013). The data is large in variety and includes test data, product failure data, and product and process data that needs to be efficiently stored, managed and analysed (Wang & Alexander, 2015).

The smart sensors are built into the automotive manufacturing machines and can distinguish various conditions, such as location, weight, temperature, balance, vibration and humidity levels. These conditions can be examined to identify and project any performance complications that may require service, repair, or machine replacement. This allows the automotive manufacturer to be cautioned in advance as several issues may be detected, and intervention can occur before operations are interrupted (Infor, 2015; Wang & Alexander, 2015).

Furthermore, Syafrudin et al. (2018) discussed that IoT based sensors are recognised as a solution towards providing more efficient monitoring of the production process and therefore identified the following benefits:

- 
- *Improved working conditions:* The IoT based sensors can be used for predictive maintenance to monitor workplace equipment. Advanced notifications are provided for any structural failures in any connected devices, enabling the automotive manufacturer to take proactive action in response to the failure (Syafrudin et al., 2018).
 - *Fault diagnosis:* Automotive production operations that use IoT sensors can result in unprecedented downtime. Big data analytics can be applied to the data generated to identify the root cause and solution to the problem (Deloitte, 2019).
 - *Quality prediction:* Data generated from IoT sensors and devices can be used for advanced and real-time identification of vehicle defects before it leaves the production line (Syafrudin et al., 2018; Wang & Alexander, 2015).
 - *Better decision making:* The large amounts of data generated from IoT sensors can be analysed to support effective decision-making in the automotive industry. The analysed data enables management to provide a strategic outlook on the organisation and make effective decisions when needed (Ismail et al., 2019).

However, Lee et al. (2013) add that connecting smart sensors to machines does not allow users to make more informed decisions. Therefore, it is vital that the data generated from sources can be analysed effectively to produce the correct information at the right time. Traditionally, the data generated was used for monitoring alerts when operating conditions were met (Infor, 2015). However, the automotive manufacturer is now using predictive technologies to mitigate and prevent failures from occurring (Machado et al., 2019).

IoT applications rely heavily on machine-to-machine interaction, where communication between the devices is limited or without human intervention. Communication allows for automating tasks, sending commands, and distributing information (Hanada, Hsiao, & Levis, 2018). This means that the IoT can collect data from the production machines and send it to ERP systems and cloud systems to provide the automotive manufacturer with new opportunities for operational excellence. This machine data can offer insight into how equipment functions on the factory floor.

A detailed product lifecycle analysis can help engineers construct future design improvements and performance enhancements (Wang & Alexander, 2015). This data also gives the automotive manufacturer the capability to estimate opportunities for the future, which helps with sales forecasting and inventory management in anticipation of the changing demands of the business world (Infor, 2015; Wang & Alexander, 2015).

However, special attention must be paid to processing big data from IoT devices to cloud-based ERP systems. Due to the large amounts of data generated from IoT devices, a scalable platform is needed to handle the complexity of the data. Proper handling of voluminous data is a sensitive topic to maintain the integrity of data. Currently, the shortfall is that there is no effective data management solution for the cloud ERP system to handle the data (Tavana, Hajipour, & Oveisi, 2020)

Once data has been generated from internal and external sources, the data's complex nature, including the characteristics, needs to be considered (Tole, 2013; Syafrudin et al., 2018). The following section examines the characteristics of big data.

2.4 Characteristics of big data

According to Riahi and Riahi (2018), big data can be classified into the five V's, representing the characteristics of the data generated. The 3V's - volume, velocity, and variety - signify the primary characteristics of big data (Rizk, Bergvall-Kåreborn, & Elragal, 2017). Furthermore, according to Rizk et al. (2017) and Riahi and Riahi (2018), two dimensions have been defined to describe big data. This includes value and veracity. However, Hariri, Frederickson, and Bowers (2019) identified many other characteristics such as variability, viscosity, validity, and viability, which has defined big data. Furthermore, Hussein (2020) studied big data's characteristics, which have now been extended into 56 V's. **Table 2.4** was compiled to compare the different characteristics which various authors had identified. This study focused on the 5 V's, which is commonly known. The 5 V's and identified characteristics are shown in **Figure 2.1**, followed by a description of each characteristic.

Table 2.4: Big data characteristics identified by authors

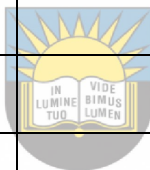
Big data characteristic	Riahi and Riahi (2018)	Rizk, Bergvall-Kåreborn and Elragal (2017)	Hariri, Frederickson, and Bowers (2019)	Hussein (2020)
Volume	X	X	X	X
Velocity	X	X	X	X
Variety	X	X	X	X
Value	X	X	X	X
Veracity	X	X	X	X
Variability			X	X
Viscosity			X	X
Validity			X	X
Viability			X	X

Venue				X
Vocabulary				X
Vagueness				X
Vulnerability				X
Volatility				X
Visualisation				X
Virality				X
Virtual				X
Valences				X
Virility				X
Vendible				X
Vanity				X
Voracity				X
Visible				X
Vitality				X
Vincularity				X
Verification				X
Valour				X
Verbosity				X
Versality				X
Varnish				X
Vogue				X



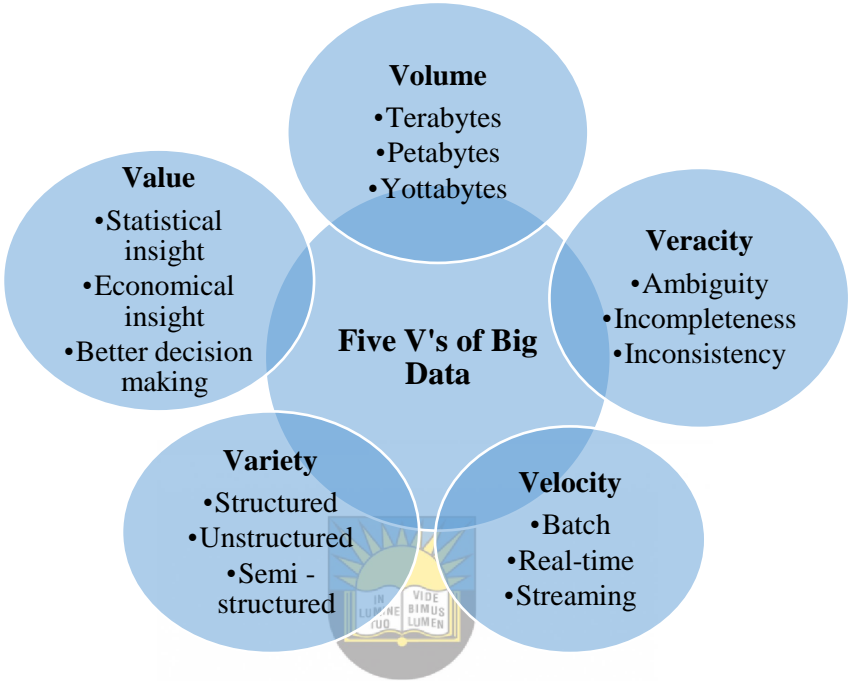
University of Fort Hare
Together in Excellence

Vault				X
Veil				X
Vulpine				X
Verdict				X
Vet				X
Vane				X
Vanilla				X
Victual				X
Vantage				X
Varmint				X
Vivify				X
Vastness				X
Voice				X
Vaticination				X
Veer				X
Voyage				X
Varifocal				X
Version control				X
Vexed				X
Vibrant				X
Virtuosity				X



University of Fort Hare
Together in Excellence

Veritable				X
Violable				X



University of Fort Hare

Figure 2.1: Common characteristics of big data

(Delgado, 2017; Hariri et al., 2019; Mikalef et al., 2018 & Wang & Alexander, 2015)

2.4.1 Volume

The volume of big data refers to the significant size or the large scale of the dataset generated, which needs to be gathered and analysed to extract meaningful information (Riahi & Riahi, 2018; Younas, 2019). Gandomi and Haider (2015) and Hariri et al. (2019) describe the volume of big data as relative and determined by factors such as time and the type of data, making it impractical to define the threshold of its volume.

Hussein (2020) states the size of the data generated just from one data point can be measured in terabytes, petabytes or even extended to yottabytes. The voluminous nature of big data is portrayed as a significant challenge to manipulate and analyse the data because it requires utilising many resources to present the required results (Gandomi & Haider,

2015; Hariri et al., 2019). However, Hussein (2020) indicates an advantage in analysing massive amounts of big data, providing better results and efficiency. Currently, automotive manufacturers generate an exponential amount of data from IoT devices connected to the production line. However, the volume of data generated will continue to increase rapidly due to IoT's ongoing development and expansion on the production line (Stephens, 2020; Deloitte, 2015).

2.4.2 Velocity

The velocity of big data is concerned with the frequency at which the data is generated (Riahi & Riahi, 2018). Gandomi and Haider (2015) further add that the speed at which the data is analysed and acted upon forms part of defining the velocity of big data. However, according to Wang and Alexander (2015), the velocity of big data is also concerned with whether the data is generated in batch, real-time or streams. Therefore, the speed at which the data is processed must meet the rate at which the data is generated (Hariri et al., 2019).

For automotive manufacturers, the main challenge is to keep up with the relatively high-end user requests for streamed data over multiple manufacturing devices. Data transfers are performed less than the capacity of the systems. As a result, transfer rates are said to be restricted, but user requests are unlimited. This makes the streaming of data in real-time a challenge (Tole, 2013).

2.4.3 Variety

Variety is the third characteristic of big data and refers to the various types of data stored, analysed, and used by the automotive industry to achieve results (Younas, 2019). The types of data include structured, semi-structured and unstructured data (Hariri et al., 2019).

Structured data is well-organised and can easily be stored (Hariri et al., 2019). Furthermore, structured data is advantageous because it can be easily entered, stored, analysed and queried (Wang & Alexander, 2015). Examples of automotive manufacturing data include the data stored on relational databases, spreadsheets and enterprise systems. Unstructured data in its raw format makes it difficult to analyse (Hariri et al., 2019).

As a result, unstructured data requires decoding before data values can be elicited. Examples include log files and operator shift reports. However, compared to structured and unstructured data, semi-structured data varies by containing tags or markers that separate data elements and reveal hierarchies of records (Wang & Alexander, 2015). The three types of big data and examples of automotive manufacturing data are shown in **Table 2.5**.

Table 2.5: Types of big data in an automotive manufacturing environment
(Wang & Alexander, 2015)

Structured Data	Unstructured Data	Real-time, semi-structured data
Spreadsheets	Operator shift reports	RFID tags
Relational databases	Machine logs, error logs	Machine builder standards (sensors, vibration, pressure)
Enterprise data warehouse	Images	XML
Files stored on the manufacturing PC's	Manufacturing social platforms	Manufacturing historians (time-series data structures)

2.4.4 Value

Data contains no meaning if the industry does not obtain value from the datasets (Hussein, 2020). This characteristic is concerned with the context, usefulness for decision making, added value and economic insights realised from processing and analysing big data generated (Mikalef, Pappas, Krogstie, & Giannakos, 2018; Hariri et al., 2019). In addition to this, the value of data is determined by the integrity of the data (Tole, 2013). The valuable information needs to be extracted from the datasets to be used further for decision making (Delgado, 2017).

The value of big data is regarded as a significant factor for the automotive manufacturer to have the ability to outperform competitors, create new growth opportunities and realise benefits. For example, automotive manufacturer has access to pools of valuable data generated and analysed by their IoT devices. Networks are connected to these technologies, enabling employees to report their production data in real-time (Machado et al., 2019).

2.4.5 Veracity

Veracity is the fifth characteristic of big data, mainly concerned with the data's consistency, quality, and trust (Younas, 2019). Furthermore, it refers to the origin of the data, the accuracy of the IoT data source and its significance.

It is important to define whether the data generated is from a reliable source. Some data can be inconsistent, ambiguous, or incomplete resulting in veracity categorised as good, bad, and undefined (Hariri et al., 2019). This means that there may be unreliability inherent in some datasets. For example, automotive consumer opinions on social media entail uncertain ideas. However, they do contain valuable information that can be used further.

Therefore, analysing uncertain data through tools and analytics also forms part of the process (Gandomi & Haider, 2015). Considering the complexity of big data and its distinctive characteristics, automotive manufacturers need to utilise tools and techniques to analyse and extract meaningful patterns from large datasets (Riahi & Riahi, 2018).

As a result, advanced data analytics enables the automotive industry to achieve the required results (Tole, 2013). The following section provided an overview of big data analytics and examined the various types of analytics used in automotive manufacturers.

2.5 Big data analytics

For automotive manufacturers to remain competitive in the market, they are re-examining their current business models and identifying new strategies for growth, long term investment in advanced technologies and efficient operations. Due to this change, automotive manufacturers rely on advanced analytics, such as big data analytics (Deloitte, 2019).

Traditional data analytics cannot be used to analyse big data as it loses effectiveness due to the five V's concept discussed in the previous section (Hariri et al., 2019; Hussein, 2020). As a result, advanced analytics is utilised to perform semi-autonomous or autonomous analysis of large scale and complex datasets, discover patterns, market trends,

and additional valuable information through advanced techniques and tools (Deloitte, 2019; Riahi & Riahi, 2018). Big data analytics requires careful attention and an in-depth understanding of the importance of the data. In a study by Ogbuke, Yusuf, Dharma and Mercangoz (2022), it was discussed that laws and regulations need to provide sufficient protection of the industries data, when big data analytics is concerned. Security breaches and information violations can potentially occur when complex and the enormous amounts of data is involved. Therefore, it was identified that it is necessary to extract the value of the data without any violation to ethical, security and privacy violations of big data in the automotive environment (Ogbuke et al., 2022). This is the concept where many automotive manufacturers fall short (Infor, 2015).

Analysing the data in real-time assists automotive manufacturers with perspectives from the past and the future. This means big data analytics provides insight into knowing what occurred (descriptive), understanding the root cause (diagnostic), identifying what might happen (predictive) and determining how to control future occurrences (prescriptive) (Ajah & Nweke, 2019). **Table 2.6** indicates questions that can be associated with the four types of big data analytics identified. Each type of big data analytics identified was briefly discussed in the sections that follow.

Table 2.6: Types of big data analytics (Deloitte, 2019, Riahi & Riahi, 2018)

Types of Big Data Analytics	
Descriptive Analytics	
•	What is happening?
Diagnostic Analytics	
•	Why did it happen?
Predictive Analytics	
•	What is likely to happen?

Prescriptive Analytics

- | |
|--|
| <ul style="list-style-type: none"> • What should be done? |
|--|

2.5.1 Descriptive analytics

Descriptive analytics analyses historical data to identify and provide insight into the cause of certain situations in the production process (Ajah & Nweke, 2019; Lepenioti, Bousdekis, Apostolou, & Mentzas, 2020). Descriptive analytics can be closely associated with diagnostic analytics by identifying why certain events occurred through the application of data mining and statistical techniques. Once the methods have been applied, patterns from the collected data can be discovered (Ajah & Nweke, 2019; Riahi & Riahi, 2018). For example, the automotive manufacturer can use descriptive analytics as a vehicle design trend by identifying opportunities to differentiate between next-generation car models (Deloitte, 2019).



2.5.2 Diagnostic analytics

Diagnostic analytics analyses historical data to identify the root cause of certain events and understand the reason and behaviours (Riahi & Riahi, 2018). Therefore, diagnostic analytics is integrated with descriptive analytics to account for why certain circumstances occurred using patterns identified in the collected data (Ajah & Nweke, 2019). For example, the automotive manufacturer uses diagnostic analytics to determine the reason behind a line stop in the production process (Wang & Alexander, 2015).

2.5.3 Predictive analytics

Predictive analytics is based on assumptions on past data trends that account for and predict the future (Deloitte, 2019). Advanced statistics, information software or operations research methods are applied to identify variables, build predictive models, and forecast simulations for the future (Ajah & Nweke, 2019; Riahi & Riahi, 2018). The result is for the automotive manufacturer to predict opportunities for the organisation to improve their products and identify production failures in advance (Ajah & Nweke, 2019). For example,

predictive analytics enhances the quality of a vehicle by using artificial intelligence to predict quality concerns, their root cause and measures to address the problem (Deloitte, 2019).

2.5.4 Prescriptive analytics

Prescriptive analytics is of research interest, as it is not as mature as the other types of big data analytics identified (Lepenioti et al., 2020). It examines data to provide simulations, recommendations and determines what should be done ahead of a problem in the production process (Deloitte, 2019). Descriptive analytics uses historical data, and predictive analytics assists with forecasting production issues that may occur. Therefore, prescriptive analytics uses these parameters to identify a viable solution (Ajah & Nweke, 2019). For example, prescriptive analytics assists automotive manufacturers with reducing time to take action and resolve a problem by offering advanced analysis techniques before an issue occurs (Deloitte, 2019).



As a result of effectively applying big data analytics, the automotive manufacturer can utilise the results to identify and improve big data use cases within the production process. The following section will identify and discuss use cases of big data in the automotive industry.

2.6 Uses cases of big data in the automotive manufacturing industry

Leveraging big data is vital for automotive manufacturers to transform the production process and remain competitive in the industry (Kurtz & Shockley, 2013). The large and complex data is extracted from the IoT devices, and analytics is applied to identify, analyse, and maintain opportunities associated with the use cases of big data in the production process (Syafudin et al., 2018). The following sub-sections discussed various use cases of big data in automotive manufacturers.

2.6.1 Optimise supply chain management

Supply chain management consists of planning and managing all sourcing materials, procurement, and logistics activities in the production process. This includes coordinating and collaborating with suppliers and mitigating risks in supply. Supply and demand are essential factors in supply chain management (CSCMP, 2020).

Automotive suppliers provide over twenty thousand parts for each vehicle produced. These parts can either be locally produced or imported from suppliers across the world. Therefore, making the supply chain process complex (Reidy, 2018). The supply chain process aims to capitalize on big data generated from data sources and predictive analytics to obtain a competitive advantage in procurement, inventory management, and logistics processes in supply chain management (Russo et al., 2015). This research study focused on the automotive manufacturer as an actor in the process.

2.6.1.1 The procurement process



Procurement is the process of purchasing materials from suppliers, which the automotive manufacturer requires to produce vehicle units (Procurement Academy, 2019). The application of big data analytics provides visibility to the procurement process between manufacturer, supplier, carrier, and external factors such as weather or road conditions (Russo et al., 2015).

Data visibility and real-time predictive analysis provide insight into the material in stock and availability. Furthermore, it allows for managing pricing from various suppliers, demand, and external conditions, which may delay materials delivery. As a result, buyers can share data to understand which materials are necessary, when needed, and assess the suppliers' services (Russo et al., 2015; Wang & Alexander, 2015).

2.6.1.2 Inventory management

Inventory management is concerned with storing materials in a storage location within the manufacturing plant and managing materials used within the production process (Russo et al., 2015). Predictive analytics assists automotive manufacturers in obtaining real-time information on the capacity of various storage locations and production lines within the plant. Automotive manufacturers can follow best practices to reduce inventory levels, which also reduces operational costs. This optimises inventory levels and ensures compliance for transporting hazardous materials (Reidy, 2018; Russo et al., 2015).

2.6.1.3 Logistics management

Logistics management is concerned with planning, implementing, and controlling efficient forward and reverse flows of materials between the point of origin and the significance of consumption on the production line (CSCMP, 2020). Effective data analytics can be applied in logistics management to improve the delivery time notifications of materials in transit (Russo et al., 2015).



This means real-time alerts can be triggered if a shipment is delayed, and supply chain professionals can take alternatives. As a result, logistics and production processes can be optimised and just in time deliveries can be managed more efficiently (Reidy, 2018).

2.6.2 Improve the manufacturing process

Big data and the application of advanced data analytics enable automotive manufacturers to realise associated benefits and improve their operations in the automotive manufacturing process (Deloitte, 2019). Improved sub-processes include quality assurance and error design prevention, fault diagnosis, better decision-making, and machine utilisation. These sub-components within the manufacturing process were discussed in the following sub-section.

2.6.2.1 Quality assurance and error design prevention

Traditional quality checks involved manually inspecting the product after the manufacturing process was complete. However, big data and advanced analytics prove vital, ensuring that quality control is more effective and efficient in the production process (Davenport, Patil, & Snaidauf, 2018).

Customers can manufacture and purchase faulty vehicles without quality control, causing brand name damage, high costs, and dangerous consequences. Therefore, quality control is a crucial part of the production process in the automotive manufacturing industry (Matthews, 2018).

The detection of defect tracking and quality is a functional form of predictive analytics used to identify the product or part failure in the production process (Ajah & Nweke, 2019; Infor, 2015). The product data generated from IoT devices on the production line assists the automotive manufacturing specialists to identify design flaws and weak product elements (Syafudin et al., 2018; Wang & Alexander, 2015). These flaws and elements can be detected early and in real-time, ensuring better detection of product defects. Furthermore, supplier attributes are also relevant if the use case improves quality (Machado et al., 2019). This means the detail of every part that constitutes the product is tracked, enabling automotive manufacturers to determine the suppliers and sub-contractors that meet or exceed expectations and remove those sub-contractors who are performing poorly (Infor, 2015). Ultimately, the product lifecycle is improved, making the product of high quality (Wang & Alexander, 2015).

2.6.2.2 Fault diagnosis

Automotive manufacturing operations involve complex processes and can result in production downtime due to breakdowns, faults and other holdups in the process. Therefore, automotive manufacturers need to identify the root cause of the problem (Deloitte, 2019). The root causes analysis assists automotive manufacturers in determining the cause-and-effect relationship. Once the root cause has been identified, mitigation strategies are distinguished and applied through diagnostic analytics (Ismail et al., 2019).

The efficiency of this analysis can be improved by applying both diagnostic and predictive analytics. Diagnostic analytics can be used to identify the driver behind production downtime and defects. Therefore, it provides support in understanding the problems causing production downtime. Predictive analytics enables the automotive manufacturer to predict downtime so that solutions can be implemented proactively (Deloitte, 2019).

2.6.2.3 Improved decision making

Automotive manufacturing decisions require a lot of time and directly impact the production process and performance of the organisation. Therefore, the industry needs to ensure that findings support the strategic outlook of the enterprise (Ismail et al., 2019).

The large amounts of data generated from data sources need to be analysed by managers to assist them with decision making (Syafudin et al., 2018). Descriptive, diagnostic, predictive and prescriptive analytics can be utilised to make more informed decisions. Big data and its analytics enable the manufacturer to increase the scope of their data analysis and improve their decision making to support business strategies (Deloitte, 2019; Riahi & Riahi, 2018).



2.6.2.4 Machine utilisation

Predictive analytics in automotive manufacturing enable the enterprise to use machine loss (Lee et al., 2013). Automating the analysis of data generated from sensors and machines' operation enables manufacturers to identify when devices need to be brought online or shut down to prevent any fall-backs.

Furthermore, data captured inside the machines can trigger alerts to signal that an issue has been found to avoid preventative maintenance. This is a critical aspect in ensuring that all machines operate at their optimal level (Robinson, 2016).

2.6.3 Improved Customer Services

Data generated from external sources goes under predictive analysis to discover the manufacturing parameters which are most important to customers.

After that, the automotive manufacturer can develop a new product or improve the current product to satisfy the customer's needs (Wang & Alexander, 2015).

Furthermore, due to the advancement of technology, automotive manufacturers can use data generated using sets of big data generated from digital footprints. This data can be used to learn more about customer preferences, improve customer service, sales, optimise marketing, and be a leader amongst automotive competitors (Deloitte, 2019).

2.7 Conclusion

The automotive manufacturing industry is currently undergoing a digital transformation due to the large amounts of data produced from IoT devices on the production line. Big data is produced from various internal and external data sources. External sources are concerned with the customer and supplier environment. However, internal sources are within the production process in the business environment. The data generated from various sources is highly complex, consisting of several characteristics that define big data from traditional data. Characteristics include volume, velocity, variety, value, and veracity.

For automotive manufacturers to understand the value and use cases behind the data, they need to adopt advanced analytics to extract essential and relevant information from the datasets. The types of advanced analytics which can be applied include descriptive, diagnostic, prescriptive and predictive. Effective analytics can optimise supply chain management processes, improve the manufacturing process and customer services. **Figure 2.2** summarises the data sources, types of analytics applied, and the types of use cases discussed in this chapter.

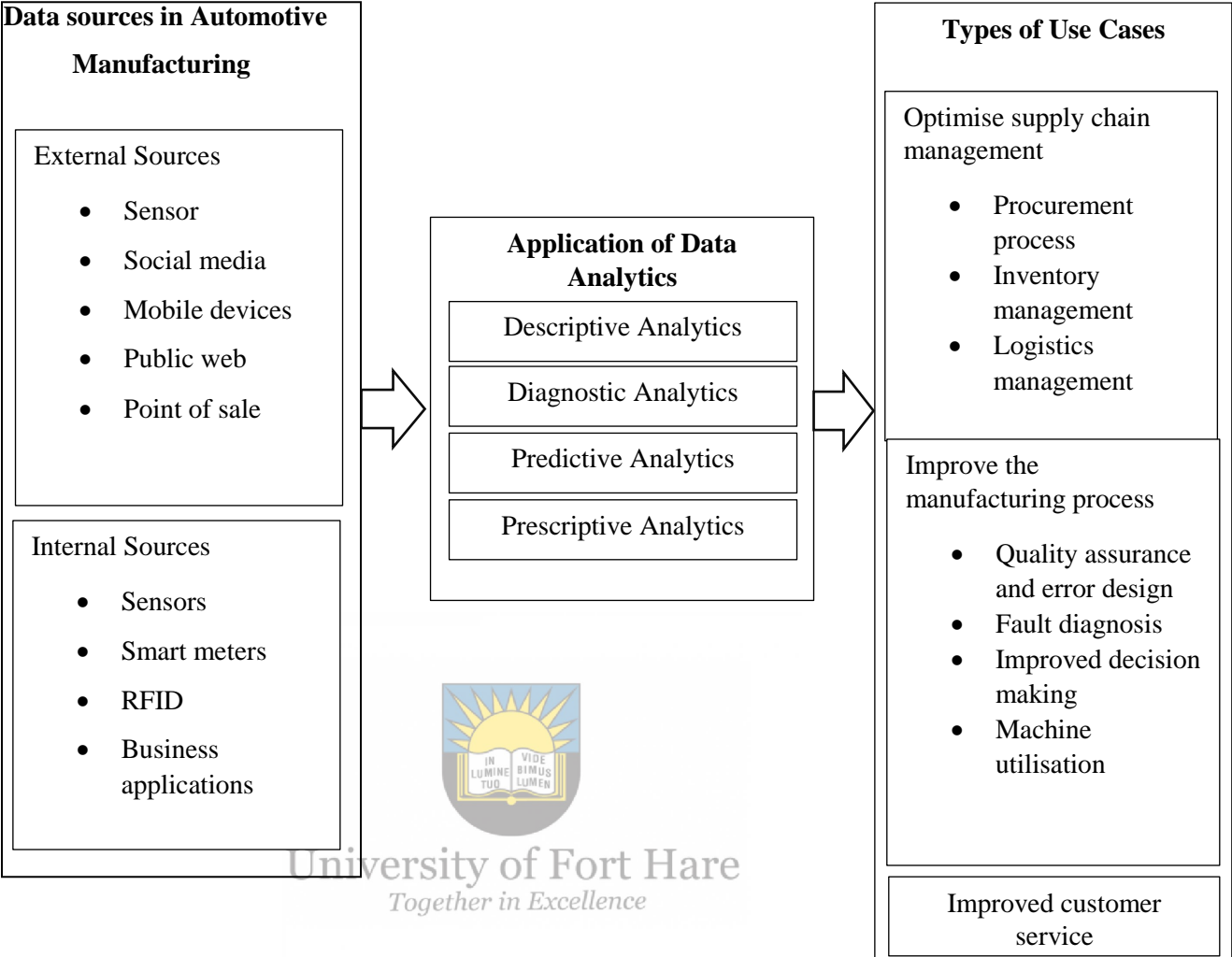


Figure 2.2: Summary of process between big data components

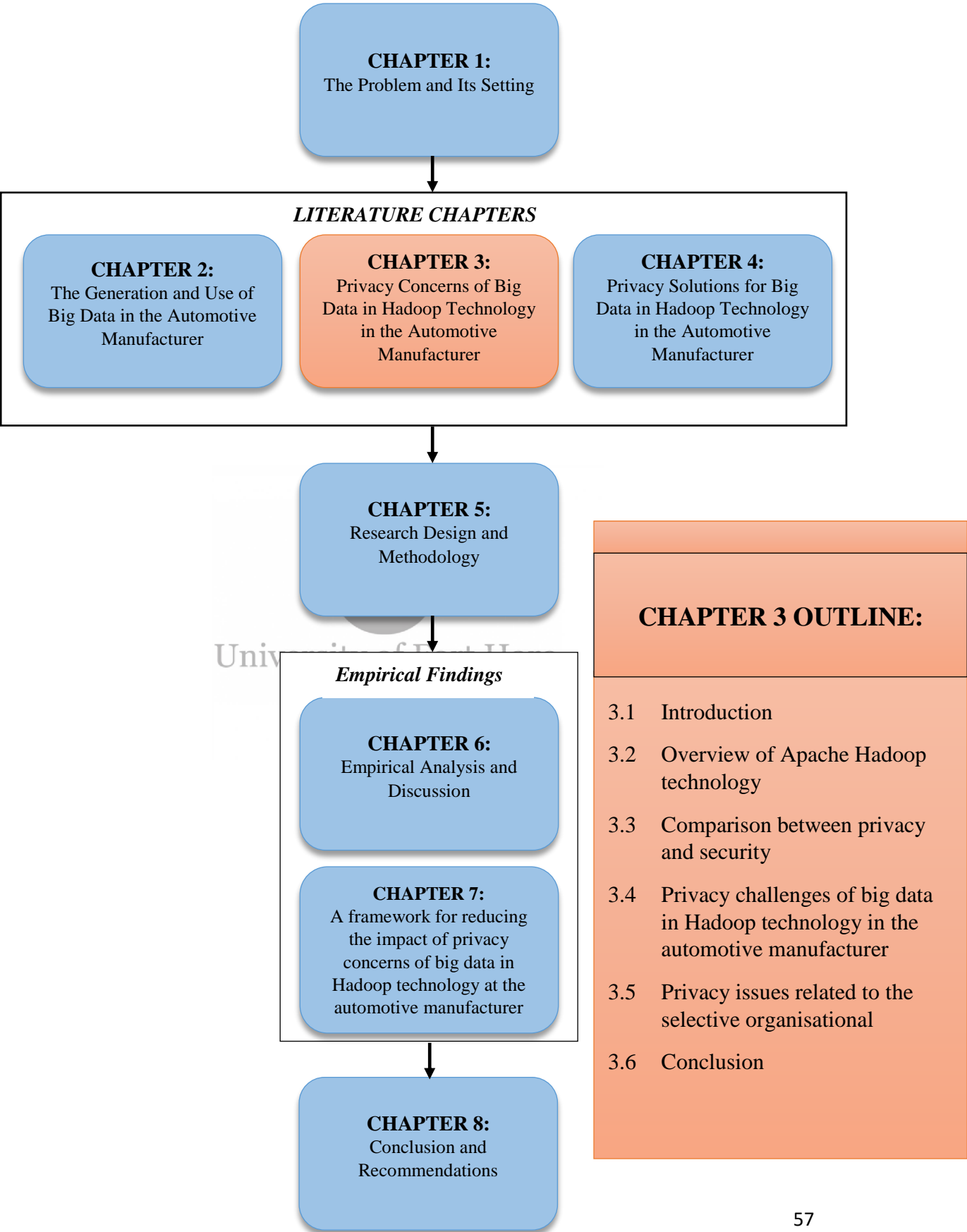
The automotive manufacturer that adopts the latest technologies, such as big data, into their production process can be identified as a key competitive differentiator (Tasmin et al., 2020). This will benefit shortcomings that may be experienced in the current market by optimising supply chain management and improving the manufacturing process and customer services.

Big data is considered an important advanced technology that the automotive manufacturer must utilise to remain competitive and ensure business growth in the industry. However, to realise the associated opportunities and use cases of big data, the industry must address the privacy issues prevalent in Hadoop technology. Hadoop technology is used to manage, process, and store big data. Therefore, it is imperative to ensure that unauthorised people or groups cannot compromise the data.

This chapter covered the fourth industrial revolution, identified and discussed the IoT sources of big data generation and the four types of big data analytics which can be applied to large and complex datasets. This provided the base of the chapter to understand the associated benefits big data has on the automotive production process. The next chapter discussed privacy concerns of big data in Hadoop technology in the automotive manufacturer.



University of Fort Hare
Together in Excellence



Chapter 3

Privacy Concerns of Big Data in Hadoop Technology in the Automotive Manufacturer

3.1 Introduction

The volume of data generated from various data sources in the automotive manufacturing industry has grown exponentially, resulting in the big data phenomenon (Ajah & Nweke, 2019). However, due to the complex nature of big data, there are difficulties in storing, analysing, and applying further procedures to realise the associated benefits (Jain, Gyanchandani, & Khare, 2016). In addition, traditional data warehouse tools and technologies cannot be used due to the complexity of the data (Hussein, 2020). As a result, big data analytics is applied, allowing the automotive manufacturer to discover patterns from the datasets and extract valuable insight required for innovation, decision-making, and business growth (Infor, 2015; Riahi & Riahi, 2018). Therefore, big data technologies, such as Apache Hadoop, must store, manage the process effectively and analyse the large sets of heterogeneous data generated from various internal and external data sources in the automotive manufacturing environment (Apache Hadoop, 2020; Agrawal, 2016).

The automotive manufacturer has adopted big data and analytics in production operations to optimise supply chain management, improve the manufacturing process, and make decisions and customer services (Deloitte, 2019; Riahi & Riahi, 2018 & Russo, Confente & Borghesi, 2015). Automotive manufacturers predominantly use Apache Hadoop for cost and process efficiency (Bhathal & Singh, 2019).

However, information privacy vulnerabilities exist in Hadoop technology, where sensitive information can become compromised (Tawalbeh, Muheidat, Tawalbeh, & Quwaider, 2020; O'Donovan, Leahy, Bruton, & O'Sullivan, 2015). If the automotive manufacturer's information is compromised, the results of big data analytics will be inaccurate, the benefits will not be realised, and competitors can take advantage of data breaches (Ajah & Nweke, 2019; Data Privacy Manager, 2020).

This study focused on Apache Hadoop, used in various industries, including the automotive manufacturer (IBM, 2015). Apache Hadoop is an open-source framework and one of the most economical methods for storing and processing big data. However, due to the flexibility of the framework, some vulnerabilities are realised (Bhathal & Singh, 2019). This research study focused on addressing information privacy concerns of big data in Hadoop technology.

Privacy is concerned with the appropriate use and sufficient control to access the industry's data (Li, Li, Niu, & Chen, 2019). If privacy issues are left unaddressed, the organisation's data is threatened, and potential information attacks can occur. This negatively affects the automotive industry's reputation, business growth, and competition (Bhathal & Singh, 2019; Data Privacy Manager, 2020). As a result, big data's opportunities and use cases cannot be realised (Bhathal & Singh, 2019; Kurtz & Shockley, 2013).

This chapter discussed the open-source framework used by the automotive manufacturer for big data storage and processing: Apache Hadoop. An overview of Hadoop was discussed, including its use in the automotive industry and its drawbacks. In addition, the concept of privacy was discussed. Privacy issues associated with the use of Hadoop were also identified and elaborated upon. Furthermore, the Selective Organisational Information Privacy and Security Violations Model (SOIPSVM) model was incorporated into the discussion of this chapter.

Identification and examining privacy concerns of big data in Hadoop technology assisted with understanding the source of the problem, which was addressed through the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer, in Chapter Seven. The next section provided an overview of Apache Hadoop technology.

3.2 Overview of Apache Hadoop technology

Traditionally, only files that had the disk capacity of a standard PC could be saved. Furthermore, to process the files, they needed to be sent to a local device. However, big data technologies such as Hadoop enable users to save and process file data much larger than the common disk capacity (Wróbel & Wikira, 2019).

Apache Hadoop is open-source software that can analyse and store large datasets within a reasonable amount of time (Bhathal & Singh, 2019). Instead of depending on expensive proprietary hardware, the advanced technology contains a distributed and parallel computing platform that is scalable and fault-tolerant (Gutierrez, 2015 & Kumar, 2015). This means that Hadoop technology has been built in a way that considers the preparation of massive datasets using simple programming models (Bhagyashree & Koundinya, 2020)

Hadoop can scale up from a single server to many and divides data across multiple system infrastructures for processing (Educba, 2020). Furthermore, it can create a map of the scattered data contents to be easily found and accessed (Tole, 2013). Wróbel and Wikira (2019) mention that Hadoop technology can create clusters of data. If any clusters fail, then the mechanism continues to operate without losing any data. In contrast to traditional platforms, Kumar (2015) identified that Hadoop could store any data type in its original format. The use of Hadoop in the context of the automotive manufacturer and its drawbacks must be understood. These were discussed in the sections that follow.

3.2.1 Hadoop in the automotive industry

The automotive manufacturer uses Hadoop to control challenges such as storage, mining, and analysing complex data (Tole, 2013). It is considered useful for automotive

manufacturers because it is cost-efficient to store and process data (Educba, 2020). Furthermore, it is used to perform various analyses and transformations (Tole, 2013). As depicted in **Figure 3.1**, the automotive manufacturer can take advantage of the key aspects of Apache Hadoop, namely: computing power, fault tolerance, flexibility, scalability, low cost, ability to store and process huge amounts of any data type (Huawei, 2021).

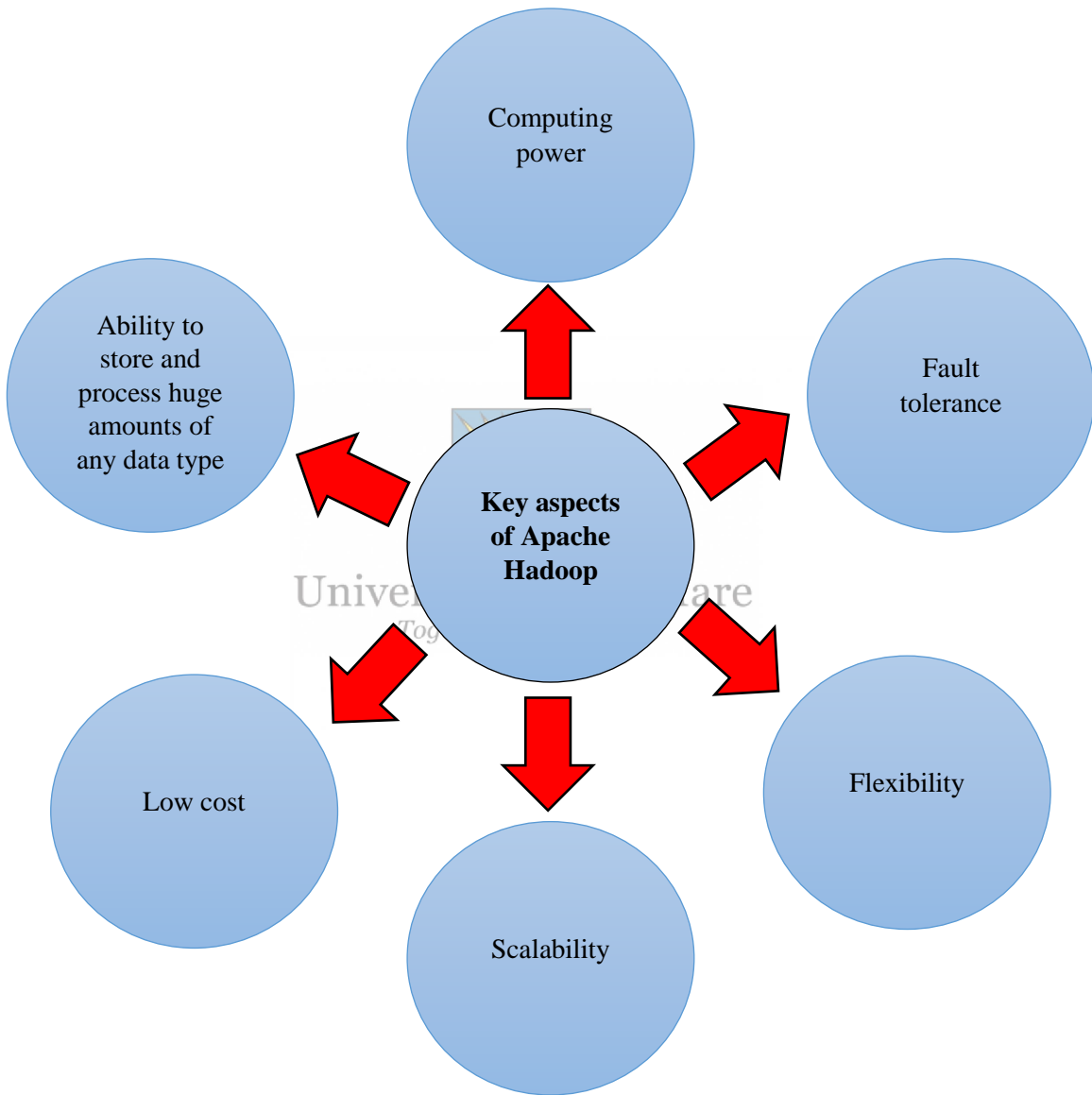


Figure 3.1: Key features of Apache Hadoop (Huawei, 2021)

- *Computing power* – Hadoop has a distributed computing infrastructure that enables fast big data processing. The more computing nodes used in the infrastructure result in more powerful processing (Huawei, 2021).
- *Fault tolerance* – Data and application processing are protected from distributed computing downtime. This means that Hadoop can continue to function without data losses if any node fails (Wróbel & Wikira, 2019).
- *Flexibility* – The large volumes and variety of data generated from the Internet of Things (IoT) devices can be stored and processed on Hadoop (Huawei, 2021).
- *Low cost* – Apache Hadoop is an open-source, free, open-source framework that uses commodity hardware to store data (Huawei, 2021; Bhathal & Singh, 2019).
- *Scalability* – Adding more nodes to the system with assist with handling the voluminous amount of data (Huawei, 2021).

Therefore, deploying Hadoop technology allows automotive manufacturers to manage all their data effectively, leading to new opportunities being realised. As identified in Chapter 2, opportunities included improving the manufacturing process and optimising supply chain management (Deloitte, 2019; Russo et al., 2015).

Bhathal and Singh (2019) state that Apache Hadoop is the leading big data storage and processing technology. However, Yadav, Maheshwari and Chandra (2019) argue that although several opportunities may be realised from the effective use of Hadoop technology, there have also been several concerns raised. The next section discussed the drawbacks of Hadoop technology.

3.2.2 Drawbacks of Hadoop technology

New technologies always encompass new challenges. Yadav et al. (2019) indicate that the most challenging and emerging concern for industries is the insurance of privacy and security of big data. Section 3.3 of this chapter discussed privacy and security to get an understanding of the differences. Furthermore, privacy concerns of big data in Hadoop technology were identified and elaborated upon.

Traditional Relational Database Management System (RDBMS) security has undergone many security evaluations to improve its capacity (Sharma & Navdeti, 2014). In contrast, Hadoop technology was released with minimal security support and has not undergone the same evaluation level, resulting in little assurance of the level of security implemented (Bhathal & Singh, 2019; Sirisha, Kiran, & Karthik, 2018).

Originally, Hadoop was poorly developed in terms of security (Moura & Serrão, 2015). There was no security model, no authentication of users and services, and no information privacy (Sirisha et al., 2018). This meant that anybody could insert random code which could be executed. Although auditing and authorisation controls were used in early distributions, access controls could be easily avoided because any user could impersonate another. This resulted in ineffective security controls and models (Bhathal & Singh, 2019; Sharma & Navdeti, 2014).

Later, authorisation and authentication were applied, but they had several drawbacks attached to them. Users could make the mistake of deleting large amounts of data within seconds with a distributed delete. Additionally, all users had the same level of access to data in the cluster, meaning any user could read any cluster dataset (Sharma & Navdeti, 2014).

Furthermore, according to Perwej (2019), firewalls do not sufficiently provide security in Hadoop Technology. This is because all firewalls represent a single layer of defence around a soft interior. Once the firewall has been breached, the data clusters are open for attack from unauthorised users (Sharma & Navdeti, 2014). Privacy and security regarding the big data environment is critical issue that needs to be addressed. The absence of an effective security model can result in data being undermined and easily compromised (Jain et al., 2016). The next section differentiated between privacy and security and identified big data privacy concerns resulting from an ineffective security model in Hadoop technology.

3.3 Comparison between privacy and security

Security is concerned with the practice of protecting information and information assets. The protection of information and information assets is done through the effective use of

technological mechanisms, processes, and training to prevent unauthorised usage, disclosure, disruption, modification, inspection, recording and destruction of data (Fruhlinger, 2020). The automotive manufacturer's technological mechanisms and tools collect real-time data about asset inventory vulnerabilities and threat detections. Security professionals use the tools and data to identify irregularities should an asset vulnerability be exploited (Automotive World, 2021).

Privacy is the advantage of having control over how sensitive information is collected and used. It is concerned with the proper handling of confidential or sensitive data, meeting regulatory requirements, protecting the confidentiality and integrity of the data, and preventing it from being exploited (Jain et al., 2016). For example, a Vehicle Identification Number (VIN) identifies a particular vehicle being produced in the automotive production process. This unique number is linked to a customer. Therefore, a user privacy issue is identifying customer information during internet transmission (Bowman, 2020; Jain et al., 2016).

In summary, information privacy focuses on the use and governance of sensitive data to use it appropriately. In contrast, security protects data from malicious attacks and improper use of stolen data to profit. The differences between privacy and security have been summarised in **Table 3.1**.


Table 3.1: Differences between privacy and security (Jain et al., 2016)

Privacy	Security
Appropriate use of sensitive information.	Confidentiality, integrity, and availability of data.
Capability to decide the destination of the individuals or organisations sensitive information.	Capability to be confident that decisions on data are respected.
Possible to have poor privacy and good security practices.	Difficult to have good privacy practices without a good data security model.

Data is a critical asset for any organisation, including the automotive manufacturing manufacturer (Agrawal, 2016). Effective use of data is crucial for business continuity and

for making effective decisions based on information extracted from the datasets (Yadav et al., 2019). However, a security breach and data misuse can seriously affect the automotive manufacturer, both legally and financially (Krauss, 2014; Wall et al., 2015). For example, the impact can be attributed to ransomware attacks, production downtime, lost production and scrapping of vehicles, theft of intellectual property and damage to the automotive manufacturer's reputation (Automotive World, 2021).

Ransomware is malware that holds the victim's device hostage until a ransom has been paid. The hackers use malware to lock and encrypt files on the device, making the data inaccessible. When the ransom is paid, the hackers may or may not provide the victim with a decryption key to regain access (van der Kleut, 2021). In 2020, Honda was a victim of ransomware which had impacted its operations across the world. The attack affected access to servers, emails, and internal production systems, which halted production operations (Tidy, 2020).



Data theft increases every day as several breaches are available for attackers to access automotive manufacturers sensitive information. The greater the volume of the data, the greater the risk associated as cyber criminals try to hack, steal, and sell the data using various tools and mechanisms available (Simon & Ramesh, 2016). This research study only focused on addressing information privacy concerns. The following section identified and discussed privacy issues and challenges of big data in the Hadoop environment.

3.4 Privacy challenges of big data in Hadoop technology in the automotive manufacturer

Big data is generated from IoT devices in various forms and voluminous amounts (Syafudin, Alfian, Fitriyani, & Rhee, 2018). Apache Hadoop provides the capability to process the data at a fast speed and a minimum cost (Educba, 2020). However, security issues arise due to the large volume of data stored in the database, which is not regular or encrypted. Hadoop's tools and technologies to handle the datasets do not have proper security, enabling hackers to steal data and copy it to any storage device (Chu K., 2020). In a recent study, Abualkishik (2019) discussed that Hadoop is vulnerable to security and

privacy breaches as it was developed by Java, which contains flaws that are often exploited by cybercriminals. This means that sensitive data stored on Hadoop would become an appealing target for exfiltration, corruption, unauthorised access and data modification.

The security flaw in Apache Hadoop causes it to become vulnerable to data breaches or fragmentation which can be intentionally exploited or accidentally triggered, causing the loss of information privacy due to unauthorised data access, data theft and unwanted disclosure of information (Bhathal & Singh, 2019 & Chu, 2020).

As a result of an unreliable security model in Hadoop technology, data breaches and data fragmentation can compromise automotive manufacturers' sensitive information (Jain et al., 2016). The following section discussed data breaches and data fragmentation in more detail.

3.4.1 Data breaches



Frankenfield (2019) defined a data breach as the unauthorised access and retrieval of sensitive information by an individual, group or software system. This is due to an ineffective security model present, which results in data being compromised. The compromised data can be used for identity theft and fraudulent purposes, which affects the reputation and competitiveness of the company (Martin, 2019). As mentioned in the previous section, Hadoop technology lacks a reliable security model. Therefore there is a high risk of a data breach occurring. Furthermore, Bhathal and Singh (2019) have indicated that the investigation into data breaches in Hadoop technology has become a neglected issue, which has resulted in several threats to sensitive data.

Trend Micro (2018) identified the following steps in a data breach operation:

- *Research* – The cybercriminal investigates and identifies potential weaknesses in an organisations security. This includes people, systems and the network.
- *Attack* – The cybercriminal executes initial contact using the network or a social attack.

- *Network/Social attack* – A network attack occurs when a cybercriminal takes advantage of infrastructure and system application weaknesses to exploit an organisations network. Social attacks include tricking or baiting employees into providing access to the organisation's network. For example, an employee may receive a malicious attachment and be asked to enter their credentials to access the attachment.
- *Exfiltration* – Once cybercriminals can access a computer, they can attack the organisations network and access sensitive information.

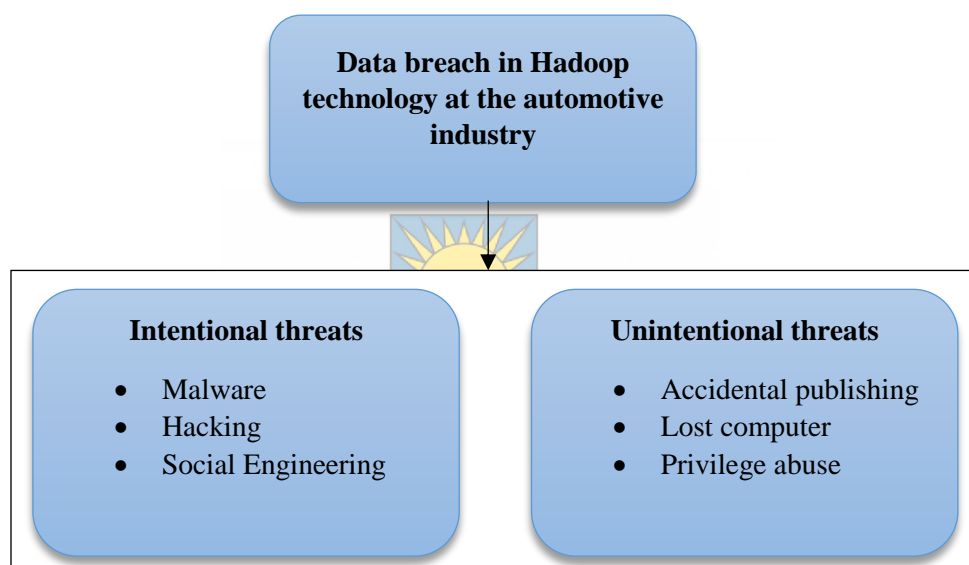


Figure 3.2: Classification of a data breach (Frankenfield, 2019; Martin, 2019)

However, Frankenfield (2019) further identified that data breaches occur intentionally or unintentionally. As shown in **Figure 3.2**, an intentional data breach occurs when a cybercriminal hacks into the company or individual's system for accessing sensitive information using the steps mentioned above. In 2021, a group of hackers attacked Kia Motors America and subjected the automotive manufacturer to a ransomware attack to decrypt information. The automotive manufacturer experienced IT system outages where malware encrypted live data and the company's backups (Stumpf, 2021). However, after

the IT outage, Kia Motors America denied reports that it was a victim of a ransomware attack (Kovacs, 2021).

In contrast, an unintentional data breach occurs when an information custodian negligently uses corporate tools. For example, an employee may access insecure websites, lose a laptop or cellphone, which can cause a data breach (Frankenfield, 2019; Martin, 2019).

In 2021 Volkswagen experienced an unintentional data breach, where an unauthorised third party gained access to personal information about customers and interested buyers used for digital sales and marketing. The exploited data included personal information such as phone numbers, email addresses, date of birth, account numbers, and information about the vehicle purchased, leased or inquired about. The cause of the data breach was a security vulnerability present in an electronic file that contained the information (Shepardson, 2021).

3.4.2 Fragmented data



Data fragmentation is considered a data privacy risk associated with Hadoop technology (Bhathal & Singh, 2019). Big data clusters contain fluid data, allowing multiple copies to move to and from the various nodes, resulting in redundancy (Lovalekar, 2014). The data becomes available for fragmentation and can be shared across many servers. This results in more complexity being added due to the fragmentation, meaning that a privacy risk arises due to the absence of an effective security model in Hadoop technology (Lovalekar, 2014; Sharma & Navdeti, 2014; Sirisha et al., 2018).

Due to security flaws in Hadoop technology, information privacy issues will arise within the big data environment (Jain et al., 2016). The automotive manufacturer is susceptible to information theft depending on the automobile models being produced and their popularity. This information includes how units are built for high-priced or high-demand items (Tasmin et al., 2020).

The SOIPSVI indicates that organisations violate privacy and security rules by exploiting or not properly protecting data. This means that contextual and rule and regulatory

conditions are deviating from the standard norm and need to be reviewed to prevent further attacks and data breaches in the organisation (Wall et al., 2015).

The global automotive market and business competition are continuing to grow (Cision, 2020). Therefore, automotive manufacturers need to be alert as they could potentially target theft of trade secrets and intellectual property via corporate espionage.

Therefore, for information to be considered a trade secret, the automotive manufacturer must take the required measures to protect their data. If the automotive manufacturer maintains a trade secret, the industry may need to explain how it has kept the information confidential over the period (Krauss, 2014).

The automotive manufacturer is more likely to face security threats such as cyber espionage, denial of service and web application attacks (Moura & Serrão, 2015; Khan, 2020). Krauss (2014) concluded in his study that in most manufacturing industries, the sources of security incidents usually come from current employees. Former employees, competitors and hackers follow. The automotive manufacturer can become affected by having employee records, personal information about customers and suppliers compromised. Also, the loss or modification of records and theft of intellectual property such as processes and institutional knowledge can have harmful consequences to the automotive manufacturing industry (Jain et al., 2016).

Therefore, the automotive manufacturer needs to address the poor security model present in Hadoop technology to protect the privacy of sensitive information. This will enable the industry to use big data technologies and data more efficiently to be more competitive, ensure business growth, maintain company reputation, and meet regulatory compliance requirements. The following section discussed privacy issues related to the SOIPSVM.

3.5 Privacy issues related to the selective organisational information privacy and security violations model

The SOIPSVM was the underlying theory used in this study to address big data privacy concerns in Hadoop technology at a local automotive manufacturer. The automotive

manufacturer uses Hadoop technology to store and process big data. However, hackers can take advantage of the security vulnerabilities within the framework. Cyber-attacks can occur, and sensitive information can be exploited (Jain et al., 2016; Wall et al., 2015). The SOIPSVM was chosen for this study as it aimed to explain the violation of privacy and security regulations and organisational behaviour (Wall et al., 2015).

The SOIPSVM model depicted in **Figure 3.3** explains how the organisational structures and processes, including characteristics of regulatory rules, can change the perception of risk when an organisation's performance does not meet the desired levels of its competitors. This affects the likelihood of rule violations (Wall et al., 2015).

The SOIPSVM consists of various elements which were briefly discussed. The elements of contextual conditional include formal and informal communication structures and violation coupling. Contextual conditions are represented by formal and informal communication structures and violation coupling. Formal communication structures include policies, fixed procedures, and rigid communication channels within the organisational structure in the automotive manufacturer. Informal communication structures are a colloquial form of interacting amongst employees in the automotive manufacturer. The impact of formal and informal communication flows in the automotive organisational structure can influence the organisation's risk perception. Violation coupling refers to the probability of a privacy and security violation leading to a positive or negative outcome (Lehman & Ramanujam, 2009).

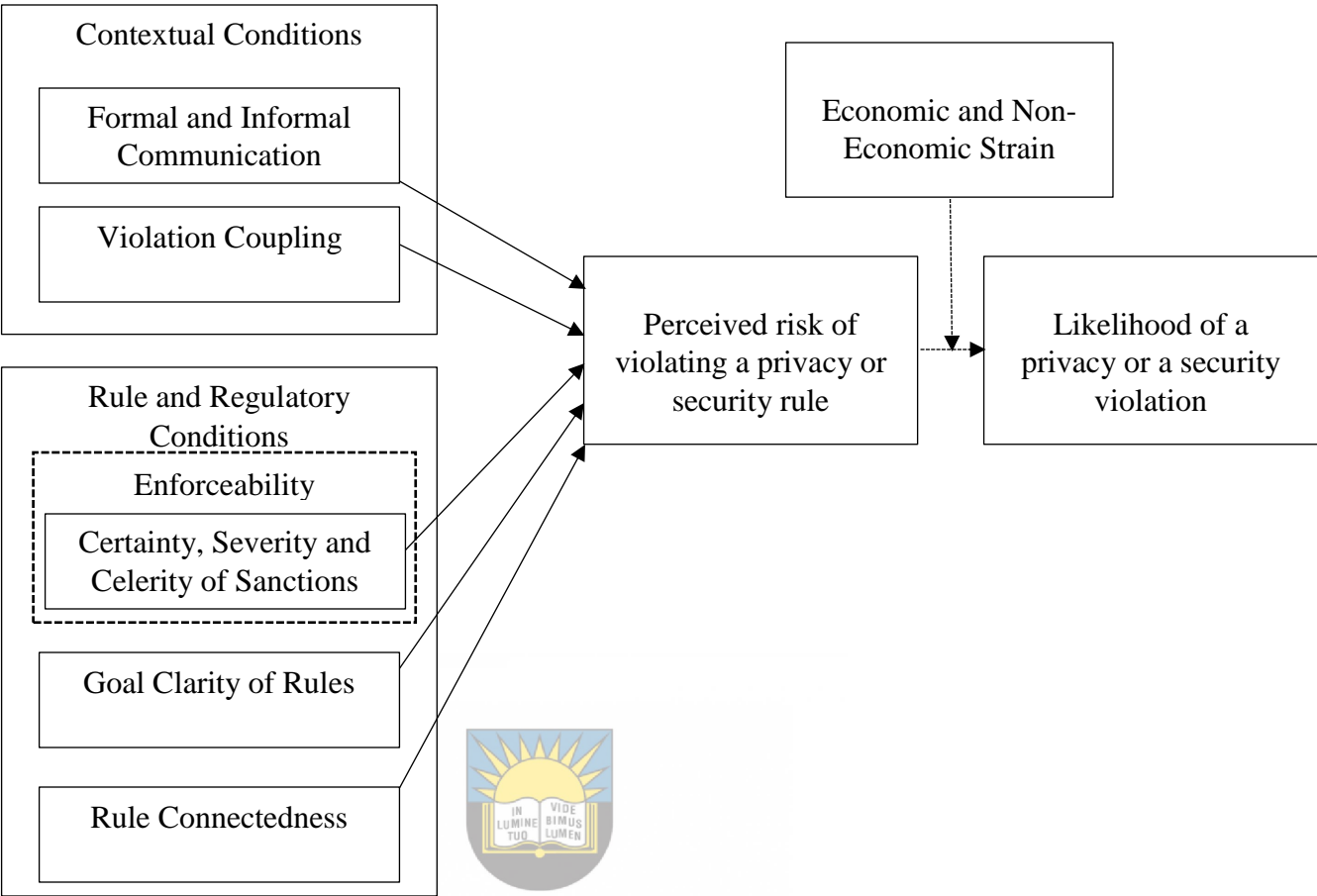


Figure 3.3: Selective organisational information privacy and security violations model (Wall et al., 2015)

When the outcome is positive, the automotive manufacturer has authority, and there is less risk perceived. However, if the automotive manufacturer experiences a loss in corporate image, this results in a negative outcome. A negative outcome is also associated with the automotive manufacturer not controlling the outcomes and increasing perceived risk. (Lehman & Ramanujam, 2009; Wall et al., 2015).

Rule and regulatory conditions consist of three elements: enforceability, goal clarity of rules and rule connectedness. Enforceability is concerned with the consistency of auditing policies and procedures which has been adapted in the company. Goal clarity refers to specific policies and regulations which clearly state an outcome.

Rule connectedness is concerned with defining all rules and regulations concerning other policies and rules in the automotive manufacturer to interpret whether a rule violation will give rise to more rule violations (Wall et al., 2015).

Rule connectedness and goal clarity of a rule increase risk perceptions (Wall et al., 2015). When these two conditions are not enforced properly, it can result in a perceived risk of violating a privacy or security rule. This means that due to the contextual and regulatory conditions not being met, data breaches or fragmentation can occur, resulting in a privacy violation (Frankenfield, 2019; Wall et al., 2015)

The perceived risk of violating a security or privacy rule is influenced by economic and non-economic strain. Economic strain is concerned with an organisation's performance compared to a competitor. In the context of the automotive manufacturer, economic strain is impacted by automotive production volumes, which influences organisational behaviour. Non-economic strain is related to regulatory pressure, consumer expectations and not the incompetence in delivering core values of the automotive manufacturer (Wall et al., 2015). In contrast, the non-economic strain results from a conflict between privacy and security rules and organisational values. When the privacy and security rules conflict with corporate values, the organisation does not abide by regulations to maintain its core values. This applies even when the perceptions of risk are high.

The likelihood of a privacy or security violation is influenced by economic and non-economic strain. If the strain is seen as successful in the organisation, it can result in privacy and security violations, resulting in fines and loss of corporate image (Wall et al., 2015). Therefore, the automotive manufacturer must implement a holistic and effective solution to address the identified privacy concerns in Hadoop technology (Sharma & Navdetti, 2014). Implementing a viable solution will enable the automotive manufacturer to realise the opportunities associated with big data in Hadoop technology (Hussein, 2020).

3.6 Conclusion

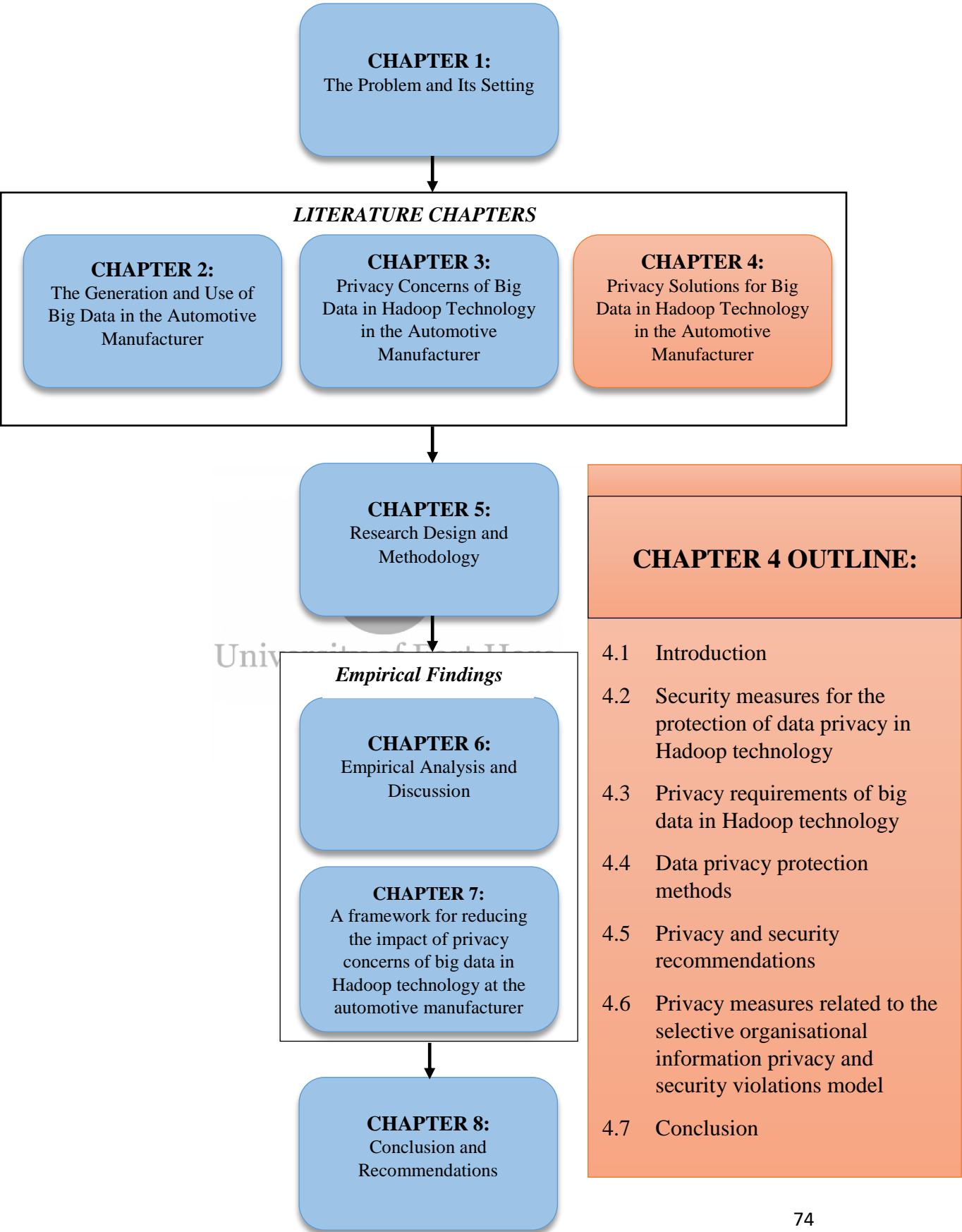
Automotive manufacturers use Hadoop technology to store, mine and analyse complex datasets generated from IoT devices. However, it is an insecure technology as it has a poor

security model associated with it. Due to the ineffective security model, this chapter identified potential privacy challenges of big data in the Hadoop environment in the automotive industry. Privacy challenges included data breaches and fragmented data.

Due to these privacy challenges, it has become easy for unauthorised users to access sensitive information generated from IoT data points. Once unauthorised groups or individuals have gained access to sensitive information, the automotive manufacturer's privacy is violated, and information becomes compromised. As a result, corporate secrets, information regarding processes, intellectual property theft, and the modification or deletion of important information can result. It can cause the organisation to lose its business competitiveness and business growth due to tampered information. To address the identified privacy concerns in Hadoop technology, chapter four identified security and governance measures required to ensure effective data privacy in Hadoop technology to address the problem.



University of Fort Hare
Together in Excellence

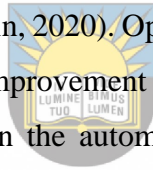


Chapter 4

Privacy Solutions for Big Data in Hadoop Technology in the Automotive Manufacturer

4.1 Introduction

Automotive manufacturing industries are currently generating big data from the Internet of Things (IoT) devices connected to the production line (Delgado, 2017). As a result, advanced analytics is applied to the datasets to realise associated opportunities for sustainability in the industry (Hussein, 2020). Options include optimisation of supply chain management processes and an improvement in the automotive production process. Associated use cases of big data in the automotive industry were discussed further in chapter two.



University of Fort Hare
Together in Excellence

Big data has taken the automotive industry to a new level by optimising all operations (Deloitte, 2019). Automotive industries use Apache Hadoop to manage, store and process big data within their environment (Educba, 2020; Yadav et al., 2019). However, Hadoop has been built with an ineffective security model, making it vulnerable to information attacks (Bhathal & Singh, 2019). As a result, data breaches and fragmented data can occur (Bhathal & Singh, 2019; Frankenfield, 2019). If the organisation's data is compromised, it results in reputational damage, loss of business growth and competitive loss within the automotive market (Ajah & Nweke, 2019; Wang & Alexander, 2015).

To address the privacy challenges mentioned in chapter three, several solutions were proposed in this chapter. According to Sharma and Navdeti (2014), there needs to be a holistic solution to address the big data privacy challenges. Traditional solutions can only protect small amounts of static data and are unsuitable for handling the complexity of big

data (Moura & Serrão, 2015). Therefore, it is important to understand how the collection of large datasets consisting of unstructured, structured and semi-structured data can be protected through effective mechanisms. The most common solution is to encrypt all data in the Hadoop environment (Yadav, Maheshwari, & Chandra, 2019).

There may be cases where automotive manufacturers do not solely have assurance with the use of Hadoop technology, even with the implementation of required data privacy mechanisms (Lorant, 2016; Tole, 2013). As a result, some automotive manufacturers may choose to build their infrastructure for processing and storing the data to ensure that data is secure based on company regulatory and privacy requirements (Lorant, 2016; Pratt, 2021). However, it is extremely costly, and regular infrastructure maintenance is needed (Lorant, 2016).

This chapter identified and discussed security measures to protect big data privacy in Hadoop in an automotive manufacturer. Furthermore, privacy requirements and solutions of big data in Hadoop were identified and discussed. Recommendations for effective privacy protection was presented, and the Selective Organisational Information Privacy and Security Violations Model (SOIPSVM) was incorporated into the discussion of this chapter. The identified solutions from this chapter formed the foundation for the framework presented in chapter seven. The next section identified security measures to protect data privacy in Hadoop technology in the automotive manufacturer.

4.2 Security measures for the protection of data privacy in Hadoop technology

After analysing the study done by Bhathal and Singh (2019), Kapil, Agrawal, Attallah, Algarni, Kumar & Khan (2020) and Sharma and Navdeti (2014), it is understood that Hadoop security is not strong and ineffective. However, security mechanisms can be implemented to provide effective data privacy in Hadoop technology. The four pillars of security solutions proposed in a study done by Sharma and Navdeti (2014) and Singh (2014) were discussed in the following sections.

These pillars included authentication, authorisation and Access Control List (ACLS), encryption and audits (Sharma & Navdeti, 2014; Singh, 2014). The identified four pillars ensure Hadoop clusters are more secure and assist in effective data privacy to protect the automotive manufacturer's sensitive information. The following section discussed user authentication implementation.

4.2.1 Implementation of user authentication

According to Abdullah, Håkansson and Moradian (2017), authentication allows for the verification of users when accessing technology. Hadoop in the automotive industry needs to be integrated with an existing authentication system such as a corporate directory (Deloitte, 2017). Kerberos is an authentication protocol that can be implemented with Hadoop clusters. It is a pluggable authentication for Hyper-Text Transfer Protocol (HTTP) web consoles, allowing web applications and console implementors to configure their authentication mechanism. The main purpose of Kerberos is to verify the identity of a user or host (Guo, Wang, Wu, & Al-Nabhan, 2020; Sharma & Navdeti, 2014).

However, Benjelloun and Lahcen (2015) argue that authentication mechanisms are complex and difficult to handle across distributed clusters and datasets. Furthermore, Benjelloun and Lahcen (2015) proposed that the Single Sign-On (SSO) concept simplifies user authentication, allowing access to various clusters within the same unique account. Bazaz and Khaliq (2016) defined SSO as an access control technique that requests users to log on only once. This allows them to access the resource and services after the successful login without asking them to log on multiple times. As a result, SSO enables users to authenticate only once to access resources. The following section discussed authorisation and ACLS.

4.2.2 Execute authorisation/Access Control Lists

Authorisation refers to providing users with specific permissions to perform actions (Deloitte, 2017). If the data owner thinks the data stored and processed on Hadoop may exploit sensitive information which should not be shared, then an authorisation technique can be implemented to restrict access to the information (Jain, Gyanchandani, & Khare,

2016). Hadoop technology uses file-based permissions to control the access of users. The access control to files incorporates name nodes through the file permissions of user groups (Sharma & Navdeti, 2014).

Sharma and Navdeti (2014) state although Hadoop can be set up to perform access control via user or group permissions and ACLS, this may not be sufficient. However, Apache Hadoop (2020) argues that access control lists can implement permission requirements for specific users. Maayan (2020) further states a robust user control policy must have a foundation of automated role-based settings and policies. The policy-driven access control protects Hadoop from threats through the automation of complex user control management. The next section discussed encryption implementation to secure data.

4.2.3 Implement encryption to secure data

Jain et al. (2016) indicate that information privacy protection in the data storage phase is based on encryption techniques. Encryption is concerned with ensuring privacy in sensitive information and securing sensitive data stored and processed on Hadoop technology (Matthews, 2019 & Navdeti, 2014). Hadoop is a distributed system that runs on distinguished machines. Therefore, data is transferred over the network regularly. As a result, there is a demand to translate sensitive information into the Hadoop ecosystem to provide value (Apache Hadoop, 2020 & Zettaset, 2014).

Sensitive data within the data clusters need protection which should be secured at rest and in motion. This means that the data must be protected during transmission to and from the Hadoop system (Kapil, et al., 2020). According to Kapil et al. (2020) and Tole (2013), data can be encrypted using a personal key. The personal key makes the data unreadable for unauthorised individuals. However, there is a disadvantage to the encryption method.

The same software used to encrypt the data needs to be used to read and analyse it. The process becomes more complicated if the automotive manufacturer wants to make it available for every software used. The first step would entail encrypting the data using special encryption software. After the data has been used for manipulation and analysis, it must be decrypted each time it is used. After the process has been finished, it will need to

be encrypted again (Tole, 2013; Sharma & Navdeti, 2014). The next section discussed audits as a security measure for data privacy protection in Hadoop.

4.2.4 Conduct audits

Security breaches can be intentionally exposed or triggered unknowingly (Chu K., 2020). Therefore, performing audits are critical to comply with data compliance requirements because it ensures that security controls are working effectively and identifies ways to bypass them (Garcia, 2016). An Apache Hadoop system audit would entail periodically verifying logs in the entire Hadoop ecosystem (Chu K., 2020). Audit logs are used as a common method to keep evidence actions performed in the Hadoop infrastructure. The logs allow administrators to review historical actions performed in Hadoop (Garcia, 2016).

Certain components of Hadoop contain an audit log. However, audit log monitoring tools need to be implemented into the components which do not have built-in logs (Sharma & Navdeti, 2014). After the logs in Hadoop are implemented, it is important to proactively monitor the data clusters for security breaches and any suspicious activities (Garcia, 2016). The next section discussed the privacy requirements of big data in Hadoop technology.

4.3 Privacy requirements of big data in Hadoop technology

Big data analytics and its associated opportunities have interested various industries (Klöser, 2019). The automotive manufacturer continuously generates and collects large amounts of data from IoT devices connected to the production line. Hadoop can store and process sensitive data generated from sensors, machines, and customer data. The generated data undoubtedly creates numerous opportunities for the sector (Kurtz & Shockley, 2013). The automotive manufacturer uses Hadoop technology to keep sensitive data in Hadoop clusters (Bhathal & Singh, 2019). Hadoop technology is a distributed system that enables the automotive manufacturer to store massive amounts of complex data and process the data in parallel (Educba, 2020).

However, the automotive manufacturer does not utilise the technologies because of the absence of an effective security and privacy model in the Hadoop environment (Jain et al.,

2016). This means that the danger of data privacy breaches occurring is present (Frankenfield, 2019). Therefore, the automotive manufacturer needs to comply with privacy terms and regulations (Jain et al., 2016).

After the poor security model associated with Hadoop technology was identified, several solutions were provided, including a technological perspective of these solutions. The four pillars of security, namely: authentication, authorisation and ACLS, encryption and audits, need to be implemented according to the SOIP SVM model (Sharma & Navdeti, 2014; (Wall, Lowry, & Barlow, 2015). This means that contextual and rule and regulatory conditions form part of the automotive manufacturer's measures. These conditions are intended to reduce the risk of there being a perception of a privacy violation. This perception should not be heavily influenced by economic and non-economic strain. Therefore there will be a decrease in the likelihood of a privacy violation occurring (Wall et al. 2015). To address the challenges experienced within Hadoop technology, the next section identified and discussed measures to protect data privacy in Hadoop technology in the automotive manufacturer.



4.4 Data privacy protection methods

The automotive manufacturer redefines their Information Technology (IT) processes to manage the semi-structured, unstructured, and structured data that form part of big data generated from IoT data points (Agrawal, 2016). Controls may be implemented to manage the in-flow of data generated from data points. Still, there has been concern regarding the data stored on various repositories from where it can be accessed. Therefore, it is important for appropriate governance measures to be taken using Hadoop technology in the automotive manufacturer (Bhathal & Singh, 2019; Mattsson, 2014).

Big data is constantly growing, and its processing is becoming quicker. Therefore, it is important to view big data privacy from a holistic approach (Jain et al., 2016). Several governance measures can be taken to protect the data, including conducting regular reviews of user access to data, data masking, monitoring user behaviour, disaster recovery and backup and tokenization (Deloitte, 2017; Mattsson, 2014; Shacklett, 2016; Simon &

Ramesh, 2016). There may be cases where automotive manufacturers may choose to build their infrastructure to store and access their data (Lorant, 2016; Tole, 2013). These measures are described in the sections that follow. The following section discussed reviewing users access to data.

4.4.1 Conduct regular reviews of user access to data

A user access review is a control mechanism used periodically to identify and verify that only authorised users can access applications and infrastructure (Ramaseshan, 2019). If employees access rights are unknown and not reviewed, it increases the risk of data theft and system vulnerability. This means that hackers can exploit old user accounts and access permissions to access the automotive manufacturers' system environment (Ramaseshan, 2019). According to Shacklett (2016), IT should conduct meetings with corporate stakeholders who access the data repositories and review data access permissions for authorised personnel. These meetings should take place on a semi-annual or annual basis.

Upon discussion, access permissions can be influenced by either adjusting it upwards or downwards based on employee responsibilities (Shacklett, 2016). It is important to note that when employees and contractors are no longer employed in the organisation, they should instantly be removed from access to the repositories. If access is not removed, the vulnerability can be exploited and cause reputation loss to the automotive manufacturer (Ramaseshan, 2019). The next data privacy protection method discussed is the application of data masking.

4.4.2 Apply data masking to sensitive data

Data masking is also referred to as data obfuscation, data anonymisation, or pseudonymization (Dilmegani, 2021). Data masking is used to edit sensitive data, not share it with others external to the automotive manufacturer. The masking preserves the type and length of the structured data, replacing it with distorted and worthless value (Kapil et al., 2020; Shacklett, 2016). Masked data looks and can act like the original, but only authorised users and processors can read it. Dilmegani (2021) states the need for data masking has increased due to the following reasons:

- The need to copy production data for application testing and business analytics modelling.
- Employees unintentionally threaten the organisation's data privacy policy. A study indicated that 79% of Chief Information Officers (CIOs) believed employees put company data at risk accidentally.
- Data protection acts in countries force organisations to strengthen their data protection. Otherwise, fines will be issued.

Therefore, data masking should be considered if the automotive manufacturer sells big data to third parties (Mattsson, 2014; Shacklett, 2016). However, Dilmegani (2021) argues that even if an organisation applies complex and comprehensive data masking techniques, there is a chance that an unauthorised user will be able to identify the trends in masked data. Therefore, there is a risk of sensitive information being released to third parties. The following section discussed disaster recovery and backup plans.

4.4.3 Implement disaster recovery and backup plan

A disaster recovery enables the automotive manufacturer to continue with operations should the data centres fail. The disaster recovery plan is supported through backups, replication, and mirrors (Awasthi, 2020). A backup is like the production system, but it is not available immediately. A replication is a close resemblance of the production system, where data is replicated at a scheduled interval. Lastly, a mirror is an exact copy of the production system and is immediately available (Deloitte, 2017).

The automotive manufacturer must use the lessons learned from previous years and apply them to their updated disaster recovery and backup plans. The reviewing of policies will assist the organisation with determining those which are outdated and others that are still valid (Pittaluga, 2021). The next section discussed the monitoring of user behaviour.

4.4.4 Monitor user behaviour

Monitoring user behaviour is concerned with the continuous monitoring of the access routine of users and developing an outlook on the behavioural aspect of the user accessing

the data (Raguvir & Babu, 2020). Anomaly detection identifies the items, events or observations which do not comply with an expected pattern. This results in a problem such as fraudulent activities and structural defects (Reghunath, 2017). An alert should be issued if the usage pattern does not match how the user normally interacts with the data or if anomalies are found (Raguvir & Babu, 2020).

Logs from Apache Hadoop and the runtime environment can also be used to detect and diagnose irregular behaviour (Chu K., 2020). However, monitoring and getting insight from log data is a challenge as developers need to make decisions that impact the quality and usefulness of log data. This includes the logs statements and which information should be included in the log messages (Cândido, Aniche, & van Deursen, 2021). The following section discussed the application of tokenization for data privacy.

4.4.5 Apply tokenization to secure data

Tokenisation is concerned with replacing sensitive data with a random value while retaining all critical information about the data. The substituted values and strings of characters that replace sensitive data are called tokens and have no value if reached (Close, 2019). This means that tokenisation securely stores the original value outside the original environment (Mixson, 2021). Tokenisation is a non-destructive form of obfuscation where the data becomes recoverable through a unique security identifier.

Data becomes secured at rest, in use, transit, and analytics (Batchelor, 2019; Mattsson, 2014). However, Turner (2020) discussed the tokenisation adds more complexity to the IT infrastructure. Furthermore, it was identified that tokenisation removes vulnerabilities should third parties be involved. As a result, the automotive manufacturer needs to ensure that suppliers have appropriate and secure systems to protect sensitive production information (Turner, 2020). The last data privacy protection method identified is a self-built infrastructure discussed in the next section.

4.4.6 Build own infrastructure to store and analyse data

The storing and analysis of large datasets is integral for the automotive manufacturer to realise big data opportunities (Agrawal, 2016). This requires a complex hardware infrastructure, and if more data is stored, more hardware systems will be needed. According to Moura and Serrão (2015), such hardware systems can only be relied upon over a certain period.

The automotive manufacturer may not be comfortable using Hadoop technology due to its several flaws. Production faults may occur when the hardware has been intensively used, resulting in a system malfunction. The organisation cannot afford to lose the data they have attained over the years (Tole, 2013). As a result, the automotive manufacturer may take the initiative to build their data repositories to store and analyse the data. However, this solution will depend on the size of the industry and its financial means. This method to protect privacy is extremely costly, and constant maintenance is required (Lorant, 2016; Tole, 2013). The following section provided recommendations for data privacy and security.



University of Fort Hare

4.5 Privacy and security recommendations

Considering the four pillars of security, methods to protect big data from unauthorised users are important to identify different data sources, the origin, and who is entitled to access the data (Ramaseshan, 2019). It is recommended by Moura and Serrão (2015) that different security procedures should be closer to data sources and the data itself. This is so that security can be provided at the origin of where the data is generated, and mechanisms of control and prevention can work simultaneously.

It is important to conduct an accurate classification to identify critical data and align it with the organisation's information security policy. This allows for the enforcement of access controls policies. Big data security and privacy enforcement ensure that data is trustworthy throughout the data lifecycle, including collecting data to its usage (Colombo & Ferrari,

2019; Moura & Serrão, 2015). The next section discussed privacy measures in association with SOIPSVM.

4.6 Privacy measures related to the Selective organisational information privacy and security violations model

This study addressed the research problem by using SOIPSVM as the underlying theory. An automotive manufacturer has raised information privacy issues due to insufficient security mechanisms in Hadoop technology used to store and process big data generated from IoT devices in the production line. Insufficient security in Hadoop creates a high risk for the automotive manufacturer's sensitive data to become compromised (Jain et al., 2016; Wall et al., 2015). Considering Hadoop security and privacy solutions in the big data environment, these factors must comply with the SOIPSVM model (Wall et al., 2015). This section discussed the privacy and security measures in conjunction with SOIPSVM.

Wall et al. (2015) proposed the SOIPSVM due to large organisations' consistent privacy and security breaches. This meant that current privacy regulations were inadequate. The SOIPSVM identified that large organisations must choose an externally governed privacy or security rule, such as laws and policies collated by external agencies. Externally governed privacy or security rules are chosen for violation in retort to organisational strain. Furthermore, organisations have violated privacy and security rules by deliberately exploiting sensitive data or not having effective measures to protect sensitive data (Wall et al., 2015). The SOIPSVM consists of the following factors, which were briefly discussed.

Contextual conditions are factors that contribute to an organisation's operations. It consists of formal and informal communication structures and violation coupling. Formal and informal communication structures impact the automotive organisational structure and communication flow. Formal communication structures are defined by fixed procedures, policies, and a rigid communication flow (Burley, 2018; Wall et al., 2015). Informal communication structures do not follow the formal structure and methods in the organisation.

Violation coupling represents the probability that privacy and security violations will lead to a positive or a negative result (Lehman and Ramanujam, 2009). When the result is positive, the automotive manufacturer has authority over the outcome and low risk. However, if the result is negative, the automotive manufacturer can experience fines and a loss in the corporate brand (Wall et al., 2015). In the context of this literature review in this study, the researcher identified the following contextual conditions:

- Control of external and internal sources
- Implementation of disaster recovery and backup plan
- Monitor the value of big data towards improving the manufacturing process

The purpose of the rule and regulatory conditions is to control the automotive manufacturer's conduct and consists of enforceability, goal clarity of rules and rule connectedness (Law Insider, 2018; Wall et al., 2015). Enforceability is concerned with the adoption of the condition is tracked together with the unfavourable effects in the automotive manufacturer. The automotive manufacturer applies enforceability by implementing efficient and streamlined audits and monitoring of response times to privacy and security violations. Goal clarity of rules refers to the automotive manufacturer's regulations and policies, containing specific goals and providing sufficient information on how the objectives can be achieved. Rule connectedness is concerned with privacy and security rules which are inter-connected. This means that an increase in privacy and connectedness will increase the automotive manufacturer's perceived risk occurrence (Wall et al., 2015). Automotive manufacturers must consider rules which are interdependent on rules or policies (Wall et al., 2015; Law Insider, 2018). The following rule and regulatory conditions were identified in the literature review of this research study:

- Implementation of user authentication
- Execution of authorisation and ACLS
- Implementation of encryption to secure data
- Conduct of audits
- Conduct regular reviews of user access to data
- Application of data masking to sensitive data

Contextual and rule and regulatory conditions have been implemented to address the repetitive occurrences of privacy breaches, making the automotive manufacturer vulnerable to its information becoming insecure, resulting in a privacy violation (Frankenfield, 2019; Wall et al., 2015).

When these conditions are not met, and a violation has occurred, there is a perceived risk that the likelihood of violation of a privacy rule has occurred. The SOIPSVM explains that an increment in the automotive manufacturer's perceived risk of contravening an externally governed privacy or security rule will lessen the likelihood of a rule violation. However, when these conditions are in place, the perceived risk of a privacy violation is low, in conjunction with economic and non-economic strain. Consequently, this results in a lower likelihood of privacy or security violation occurring as effective privacy mechanisms are implemented. The likelihood of a privacy or security violation is influenced by economic and non-economic strain (Wall et al., 2015).

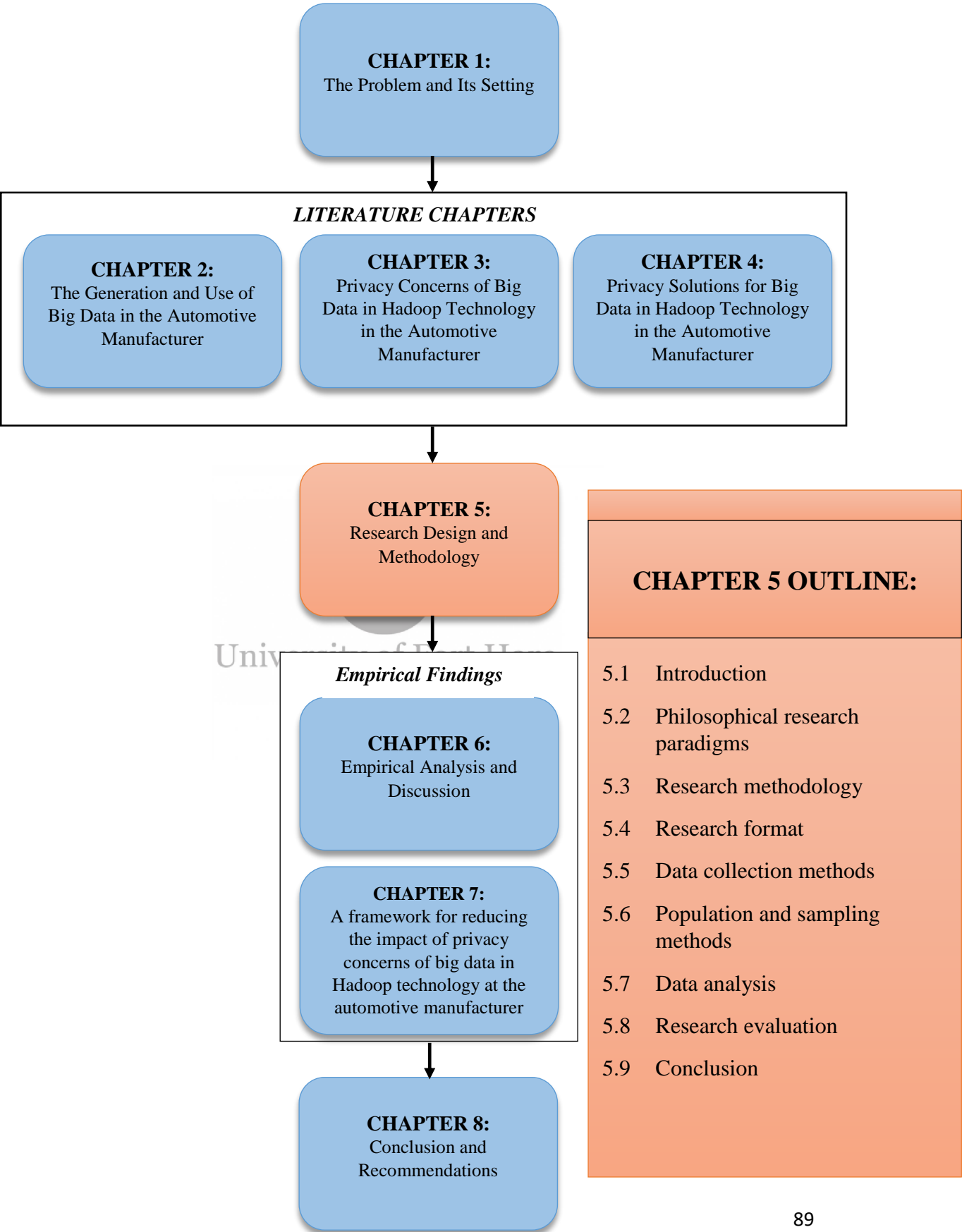
Economic and non-economic strain is dependent on the perceived risk of a security or privacy violation. Economic strain is associated with the automotive manufacturer's performance in automotive production compared to competitors in the market, which can impact organisational behaviour. In contrast, non-economic strain refers to the automotive manufacturer's incompetence to achieve organisational goals and core operations (Wall et al., 2015). The SOIPSVM explains the likelihood of a privacy or security violation due to contextual and rule and regulatory conditions not being met, increases the perceived risk of a violation occurring and economic or non-economic strain at the automotive manufacturer. Therefore, the automotive manufacturer must devise effective measures and practices for data privacy in Hadoop technology.

4.7 Conclusion

To address the privacy issues identified in chapter three, the automotive manufacturer can implement various mechanisms. The source of privacy issues is due to Hadoop technology having an inadequate security model. Therefore, it is easy for the organisation's data to be compromised by unauthorised users and third parties.

A four-pillar authentication mechanism, authorisation, encryption and audits were discussed to address the underlying security model. These four pillars ensure Hadoop clusters are more secure in the repository, therefore establishing information privacy. Furthermore, measures to protect data privacy in Hadoop technology in the automotive manufacturer were identified and discussed. Measures included conducting regular reviews of user access to data, applying data masking to sensitive data, implementing disaster recovery and backup plans, monitoring user behaviour, and applying tokenisation to secure data. The automotive manufacturer can also build and implement their infrastructure, depending on their financial means. According to the SOIP SVM model, the identified violation factors is categorised into conditions. Each condition results in there is a perceived risk of privacy and security rule violation. Economic and non-economic strain can influence this, resulting in a distinguished likelihood of a violation occurring.

Along with the security and privacy measures in place, several recommendations have also been made. It is of utmost importance for automotive manufacturers to protect their data from unauthorised users to realise their opportunities and have a competitive advantage in the market area. As a result, of the literature analysis discussed in chapters two, three and four, the next chapter discusses the research design and methodology used in this study to obtain and analyse the data collected.



Chapter 5

Research Design and Methodology

5.1 Introduction

The automotive manufacturing industry has realised big data opportunities in the current competitive environment (Deloitte, 2017). Big data generated from the Internet of Things (IoT) devices and the application of advanced data analytics enables the industry to leverage the benefits of having a connected, effective, and efficient production line (Deloitte, 2019; Reidy, 2018). Use cases of big data in the automotive manufacturer were identified and examined in chapter two of this study. However, with big data and related technology, Apache Hadoop has raised information privacy concerns due to its poor security model (Bhathal & Singh, 2019). The automotive manufacturer cannot address information privacy concerns of big data in Hadoop technology, resulting in the industry becoming vulnerable to information breaches and fragmentation (Bhathal & Singh, 2019; Frankenfield, 2019). Solutions to ensure data privacy were identified and examined in chapter four.

The literature discussed in chapters two, three and four set the foundation to understand the main research problem. In summary, the literature content discussed the benefits and use cases of big data in automotive manufacturers. Furthermore, privacy concerns of big data in Hadoop technology in the automotive manufacturer were identified and addressed.

Kivunja and Kuyini (2017) state that research needs to be conducted by applying a research paradigm. Furthermore, the researcher must understand the content of the paradigm and use it within the context and objectives of the study. Saunders, Lewis and Thornhill (2007) argue that the researcher must understand the assumptions of the paradigm chosen, which will form the base of the research strategy. This will improve the academic work and writing of the study, enabling the reader to have a clear perspective of the research outline

and how the solution to the research problem will be addressed (Kivunja & Kuyini, 2017; Saunders et al., 2007).

In agreement with the research problem, this research study aims to formulate a framework to assist the automotive manufacturer in addressing big data privacy concerns in Apache Hadoop. If the framework is implemented, the organisation will be more knowledgeable of use cases, limitations, and key big data requirements in the automotive manufacturer. These requirements are needed to secure and protect the privacy of data generated from IoT devices.

The objective of this chapter was to elaborate on the methodology applied to this research study. It provided a clear direction on how primary data was collected and analysed. Furthermore, this chapter provided more detail into how the research study was conducted, and the objectives addressed. Research paradigms and methodologies were defined and examined in the context of this research project. This chapter's discussion included the research format, data collection methods, population, sample size, data analysis methods, and research evaluation. The next section discussed the philosophical research paradigms which can be applied to research studies.



5.2 Philosophical research paradigms

As discussed in chapter one, the objective of a research paradigm is to guide how a research project will be conducted. Various paradigms exist in research that can be differentiated based on their respective philosophical assumptions. These philosophical assumptions support the research strategy and methodology chosen for the research study (Saunders et al., 2007). This section examined research paradigms in detail. After which, the types of research paradigms were critically discussed. Lastly, the paradigm for this study was elaborated upon.

American philosopher Thomas Kuhn proposed and introduced the concept of paradigms. The paradigm is based on patterns and a philosophical way of thinking (Kuhn, 1962). Hirschheim and Klein (1989) argue that paradigms must be applied to research and guide

the development and implementation of systems in Information Systems (IS) and Information Technology (IT).

In the context of educational research, a paradigm can be defined as a technique of investigating a social phenomenon, from which specific understandings can be obtained and explained (Kivunja & Kuyini, 2017). Comparably, Oates (2006) explained that research paradigms are the underlying philosophical perspectives of individuals concerned with the world they live in and the research performed. In addition, a paradigm is based on philosophical assumptions about the nature of reality (ontology), the relationship between knowledge and reality (epistemology) and the tools used to investigate reality (methodology), which researchers use to create the latest understandings of real-life problems. (Collis & Hussey, 2009; Saunders et al., 2007).

Therefore, the paradigm is an important element in research. This provides the underlying belief of how research should be conducted and how the results should be interpreted. As a result, the choices made in the research process had a significant impact on the interpretation of data gathered and the outcome of this research study.

Collis and Hussey (2009) recognise the positivist and interpretivist research paradigms. However, Kivunja and Kuyini (2017) identified four philosophical paradigms in research, namely: positivism, interpretivism, critical theory and lastly, the pragmatic paradigm. Additionally, design science research is an emerging research paradigm in computing (Naidoo, Gerber, & van der Merwe, 2012)

The following sections compare the identified research paradigms: positivism, interpretivism, critical, pragmatism and the design science research paradigm. Furthermore, a motivation for the selected paradigm in this study was provided.

5.2.1 Positivism

Kivunja and Kuyini (2017) state that the positivist paradigm was first proposed by French philosopher Auguste Comte (1798 – 1857). Oates (2006) states that the positivist paradigm

shares an origin of the natural science approach to research and is based on the following two assumptions:

- The world is ordered and regular, not random.
- The world can be investigated through an objective approach.

This scientific method involves constructing hypotheses and conducting experiments to examine observations and answer questions (Kivunja & Kuyini, 2017). Therefore, positivistic researchers believe that knowledge can be acquired through observation and experiment (Rahi, 2017). Oates (2006) states that deductive reasoning provides explanatory theories through fixed, pre-determined research design and objective measures to understand a social phenomenon. Furthermore, Sekaran and Bougie (2013) indicate that positivists view the world by the laws of cause and effect through a scientific approach towards research. This paradigm's support is dependent on experiments after testing the cause-and-effect relationships through manipulation and observation (Sekaran & Bougie, 2013; Oates, 2006). As a result, positivists develop hypotheses to test and use the findings from the experiments to confirm or reject the hypotheses.

Positivistic studies share the following six characteristics (Oates, 2006):

- *The world exists independently of humans:* A physical and social world exists to be studied, captured, and measured.
- *Measurement and modelling:* This world is recognised through observations, measurements and producing models in the form of hypotheses or theories.
- *Objectivity:* The researcher is an impartial observer, and facts are independent of the researcher's values and beliefs.
- *Hypothesis testing:* The research is based on empirical testing, which either confirms or refutes the hypotheses.

- *Quantitative data analysis*: Mathematical modelling and statistical analysis produce a logical and objective means of analysing observations and results.
- *Universal laws*: Researchers aim to produce generalizations.

Kivunja and Kuyini (2017) state that the positivist paradigm is the preferred worldview for research that interprets observations through facts or objectively measured. Arguably, positivism was constructed to study the natural world, making it less suitable for studying the social world (Oates, 2006). In contrast to the positivist paradigm is the interpretivist paradigm. Positivists criticise the interpretivist paradigm due to it being non-scientific. Furthermore, the interpretivist paradigm is less established compared to the positivist paradigm (Oates, 2006). This paradigm is subjective and dependent on the researcher's social context. The next section discussed interpretivism in detail.

5.2.2 Interpretivism



Interpretivism was constructed in response to the criticisms of the positivist paradigm (Collis & Hussey, 2009). According to De Vos, Strydom, Fouche and Delport (2011), the interpretivist paradigm can be traced to German socialist Max Weber and German philosopher Wilhelm Dilthey. This paradigm aims to understand the subjective world of human experience and highlight the nature of individuals characters and engagement in social and cultural life (Chowdhury, 2014; Kivunja & Kuyini, 2017). Creswall (2014) indicates that interpretivism uses the inductive process to provide an interpretive understanding of a social phenomenon compared to positivism. However, Oates (2006) suggests that the interpretivist paradigm understands computing as a social process developed and interpreted by people depending on the social setting. As a result, interpretivism aims to understand the social context of IT.

Oates (2006) argues that interpretivism does not prove or reject a hypothesis compared to positivism. Furthermore, Oates (2006) explains that interpretive research identifies, explores and explains how factors in a social phenomenon are related and interdependent. In the context of this research study, the social setting was privacy concerns of big data in Hadoop technology in the automotive manufacturer. Therefore, effective measures to

address the privacy concerns of Hadoop technology in the automotive manufacturer were identified and studied in this research project. Ultimately, the objective of an interpretivist study is to understand unique social phenomena, such as the automotive manufacturer.

Interpretive studies share the following six characteristics (Oates, 2006):

- *Multiple subjective realities*: Each person or group of people can perceive the world differently. Therefore, there is no single version of the truth.
- *Dynamic, socially constructed meaning*: Social constructions such as language, shared meanings and understandings are used to access and transmit the understanding of reality.
- *Researcher reflexivity*: Researchers must be reflexive or self-reflective to acknowledge that the research study's assumptions, beliefs, values, and actions will impact the research study. Furthermore, recognising interactions with others can result in meanings, understandings and practices being deferred.
- *Study of people in their natural social settings*: People are studied in their natural environment instead of an artificial world. Furthermore, the researcher's prior understandings or expectations must not inflict upon the participants' study.
- *Qualitative data analysis*: The interpretivist paradigm generates and analyses qualitative data
- *Multiple interpretations*: Researchers will arrive at more than one explanation in their study, but the most relevant one is discussed and explained further.

Supporters of the interpretivist paradigm strongly believe that research is subjective. This means that they develop subjective explanations from their experiences on a social phenomenon (Rahi, 2017). Furthermore, from an interpretivist perspective, research can influence the researcher's beliefs and values (De Vos et al., 2011).

In contrast to the interpretivist paradigm, the critical theory relies on social justice issues such as political, social, and economic influences (Kivunja & Kuyini, 2017). The next section examined the critical theory paradigm.

5.2.3 Critical theory

Critical theory is based within a social context and challenges prevailing political, cultural, and power relations (De Vos et al., 2011). This paradigm is seen as a social critique and seeks to control or improve the society from circumstances that constrain them (Myers, 1997).

According to Ritchie, Lewis, Nicholls and Ormston (2013), critical theory is influenced by neo-Marxism, feminism, and critical race theory, which have social and cultural factors influencing people's lives. Furthermore, Oates (2006) states that people create the social reality, whereby economic, political and cultural influences impact and shape reality. Therefore, Rahi (2017) identified that critical theory enables researchers to focus on political and social issues in the modern world, with a plan to address and eliminate the causes of social problems, such as inequality, oppression, domination and alienation.

The critical theory paradigm shares the following five characteristics (Oates, 2006):

- *Emancipation*: Critical researchers aim to liberate people and not only understand and explain the social phenomenon.
- *Critique of tradition*: Critical researchers do not merely accept the status quo but instead confront and challenge the existing patterns of power and taken-for-granted assumptions
- *Non-performative intent*: Critical researchers aim to focus on meeting managers' maximized profits and strengthening their power and control
- *Critique of technological determinism*: Critical researchers oppose the motive that people and societies need to adapt to technology. However, they argue that they should shape the technology that we develop.

- *Reflexivity*: Similarly to the interpretivist paradigm, critical researchers recognise that societal and organisational factors and history influence them. Furthermore, the influence on their methods, values and actions impact the research.

Oates (2006) argues that supporters of the critical theory paradigm condemn the interpretive paradigm for failing to analyse patterns of power and control, which regulate the world's reality. Design science research is a recent paradigm in computing, which claims to strengthen IT artefacts' use and the practical relevance of IS research (Naidoo et al., 2012). Design science research is discussed in the following section.

5.2.4 Design science research

Design science research is also known as the socio-technologist paradigm (Gregor & Hevner, 2013 & Vaishnavi & Kuechler, 2008). According to Weber (2010), design science research originated from the architectural and engineering discipline. Furthermore, this paradigm is extensively used within engineering but is considered an emerging paradigm, seeking acceptance within the computing discipline (Naidoo et al., 2012). Hevner, March, Park, and Ram (2004) discuss two paradigms most applicable to research in the IS discipline: behavioural and design science.

The behavioural science paradigm aims to develop and prove theories that explain and predict human or organisational behaviour. In contrast, the objective of the design science paradigm is to extend human and organisational capabilities via the creation of useful IT artefacts. Therefore, the behavioural and design science paradigm is significant to the IS discipline, which involves the convergence of people, organisations and technology (Hevner, March, Park, & Ram, 2004).

Design science enables researchers to develop knowledge in their specialised fields and design solutions to the problems identified in their relevant discipline (Naidoo et al., 2012). Therefore, Hevner et al. (2004) and Drechsler and Hevner (2016) classified design science as a problem-solving paradigm to ensure that knowledge and understanding of a problem are obtained by constructing and applying an IT artefact.

Weber (2010) criticises design science because it is not an accepted research approach or paradigm. Furthermore, the study indicated that design science should instead be used as a research approach in the three aforementioned paradigms, with an exception made to its origin. Goecks, de Souza, Librelato and Trento (2021) added the practicality of design science research needs to be expanded upon. This means the theory should be translated into practice. Lastly, Elragal and Haddara (2019) indicated that it is difficult to find guidelines on an appropriate evaluation strategy for design science artefacts.

Possible research paradigms were identified and critically discussed. The following section justified the selection of the interpretivist and design science research paradigm for this study.

5.2.5 Selecting an appropriate research paradigm

According to Oates (2006), positivism, interpretivism and the critical theory paradigm are mostly used in research studies within the IT and IS discipline. Furthermore, Oates (2006) states that positivism was the standard paradigm for research in the past. However, the interpretive and critical theory paradigms are often criticised for their characteristics. Over the years, researchers have adopted the interpretivist approach, where design science research is also becoming a widely used paradigm for research studies (Hevner et al., 2004; Oates, 2006). However, critical theory is less known and rarely used within research studies as compared to the paradigms mentioned above. According to Kivunja and Kuyini (2017) and Saunders et al. (2007), a paradigm is composed of four elements, namely:

- *Ontology*: This assumption is based on the nature of reality.
- *Epistemology*: This assumption concerns the nature of knowledge and what constitutes acceptable and valid knowledge.
- *Methodology*: This refers to the research design, methods, approach, and procedures used to investigate the identified research problem.

- *Axiology*: This assumption studies the role of values and ethics within the research process.

Therefore, it is important to understand the elements because they consist of the basic assumptions, beliefs, norms, and values that form each paradigm's foundation. **Table 5.1** depicts the four philosophical assumptions and characteristics to consider when choosing the research paradigms discussed: positivism, interpretivism, critical theory, and design science research.

Table 5.1: Philosophical assumptions of four research paradigms
(Adebesin, Gelderblom, & Kotzé, 2011)

<u>Research Paradigm</u>	<u>Philosophical Assumptions</u>			
	Ontology	Epistemology	Methodology	Axiology
Positivist	<ul style="list-style-type: none"> • A single reality • Law like 	<ul style="list-style-type: none"> • Objective • Detached observer 	<ul style="list-style-type: none"> • Experimental • Quantitative • Hypothesis testing 	<ul style="list-style-type: none"> • Truth • Prediction
Interpretivist	<ul style="list-style-type: none"> • Multiple realities • Socially constructed 	<ul style="list-style-type: none"> • Empathetic • Observer subjectivity 	<ul style="list-style-type: none"> • Participation • Qualitative • Interpretation 	<ul style="list-style-type: none"> • Contextual understanding
Critical theory	<ul style="list-style-type: none"> • Socially constructed reality • Discourse • Power 	<ul style="list-style-type: none"> • Suspicious • Political • Observer constructing versions 	<ul style="list-style-type: none"> • Deconstruction • Textual analysis • Discourse analysis 	<ul style="list-style-type: none"> • Inquiry is value bound • Contextual understanding • The researcher's values affect the study
Design Science	<ul style="list-style-type: none"> • Multiple contextually situated realities 	<ul style="list-style-type: none"> • Knowing through making 	<ul style="list-style-type: none"> • Developmental • Impact analysis of artefact on 	<ul style="list-style-type: none"> • Control • Creation • Understanding

		<ul style="list-style-type: none"> Context-based construction 	the composite system	
--	--	--	----------------------	--

The researcher needs to examine the philosophical assumptions (ontology, epistemology, methodology and axiology) for each paradigm. This will assist the researcher in comparing and selecting the underlying paradigm for the research study concerned (Collis & Hussey, 2009; Kivunja & Kuyini, 2017).

The paradigm adopted influenced the methodology chosen for the research study. As a result, the study was influenced by the most dominant paradigm within the research area and the problem to be investigated (Oates, 2006; Collis & Hussey, 2009). To differentiate between the positivist and interpretivist paradigms, Collis and Hussey (2009) identified the features. **Table 5.2** depicts the differences.

Table 5.2: Features of the positivist and interpretivist paradigm

(Collis and Hussey 2009)

<u>Factor</u>	<u>Positivist Paradigm</u>	<u>Interpretivist Paradigm</u>
Qualitative and quantitative data	Produces quantitative data.	Produces qualitative data.
Sample size	Uses large samples.	Uses small samples.
Theories and hypotheses	Concerned with hypothesis testing	Concerned with generating theories.
Types of data	Highly specific and precise data.	Rich and subjective data.
Location	Location is artificial.	Location is natural.
Reliability	Produces results with high reliability.	Produces results with low reliability.
Validity	Produces results with low validity.	Produces results with high validity.
Generalisability	Results generalised from the sample to the population.	Results are generalised from one setting to another.

In **Figure 5.1**, Collis and Hussey (2009) illustrated the positivist and interpretivist paradigms. Both paradigms oppose each other in the form of a continuum. The continuum of these assumptions is identified with six identifiable stages. The far left depicts the positivist (objective end), and the far-right depicts the interpretivist (subjective end).

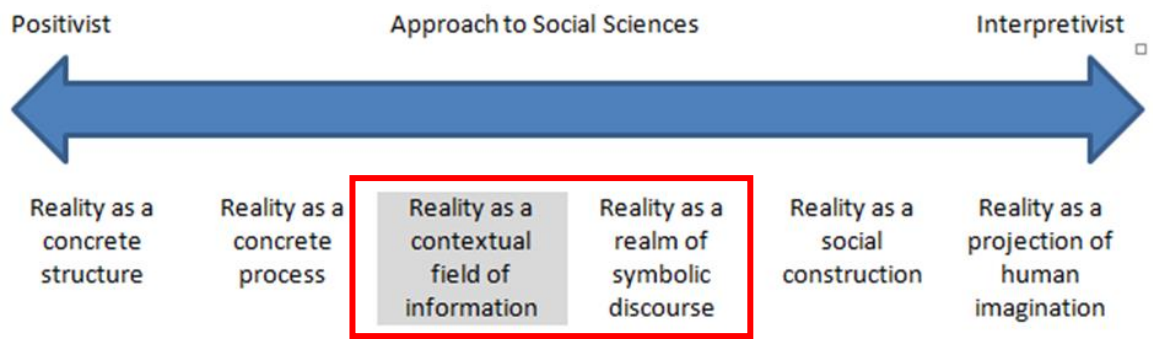


Figure 5.1: Continuum of core ontological assumptions
(Collis & Hussey, 2009)

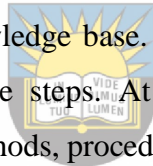
Collis and Hussey (2009) indicate that the assumptions of the one paradigm are slowly being replaced by those of the other paradigm along with the movement of the continuum. Furthermore, it is explained that only a few researchers operate exclusively in one paradigm. Using a combination of the paradigms (indicated between the objective and subjective end above) allows the researcher to achieve a broader and complementary view of the identified research problem (Collis & Hussey, 2009).

This research study focused on addressing privacy concerns of Hadoop technology in the automotive manufacturer. The paradigm used in this study is the interpretivist paradigm, which according to Sekaran and Bougie (2013), is commonly paired with the qualitative approach. This research study was positioned towards the centre of the extremes of the positivist and interpretivist paradigms. Through the literature review process, the research was focused more on the interpretivist extremes. However, the empirical process positioned the research to both positivistic and interpretive extremes.

Therefore, this research study followed an interpretivism approach but had positivistic aspects. In addition, the design science paradigm was applied to this study. Hevner et al. (2004) explain that the design science research paradigm focuses on providing guidelines for the iterative assessment of research artefacts to improve the performance of the artefact. The paradigm was applicable as the objective of this study was to develop an effective framework that assisted with addressing privacy concerns of big in Hadoop technology at the automotive manufacturer. The following section discussed and substantiated the research methodology which was implemented in this study.

5.3 Research Methodology

Research methodology is defined as procedures, principles, data collection and analysis techniques applied to the research process (Collis & Hussey, 2009). Kumar (2011) states the purpose of applying research methodology is to understand the researcher's profession and advance the professional knowledge base. This includes finding the answers to the research questions through flexible steps. At each step in the research process, the researcher must choose various methods, procedures and models of research methodology, which assists in achieving research objectives (Kumar, 2011).



University of Port Harcourt
Together in Excellence

The following section discussed the research methodology selected for this study, namely: design science research. Furthermore, the Capability Maturity Model (CMM) was discussed, which was used together with the Selective Organisational Information Privacy and Security Violations Model (SOIPSV) to compose a framework to meet the objectives of this study.

5.3.1 Design Science Research

The research methodology selected for this study is design science. Design science research is concerned with the development of a socio-technical artifact which is used for IS evaluation (Gregor & Hevner, 2013). Hevner et al. (2004) further explained that design science research aims to create and assess IT artefacts to solve identified organisational problems. This research project aimed to develop and produce a framework that can assist

automotive manufacturers with addressing privacy concerns of big data in Hadoop technology.

Creating an artefact enabled the researcher to address and understand the identified problem and the feasibility. The artefact produced from this study was a framework. Hevner et al. (2004) and Gregor and Hevner (2013) categorised the four types of IT artefacts as follows:

- *Constructs*: Vocabulary or symbols used to define and communicate problems and solutions. The constructs are clearly defined with descriptions at an abstract level.
- *Models*: Representation of real-world situations which assists in understanding the problem and solution
- *Methods*: Define processes and provide aid in solving problems
- *Instantiations*: Indicate that constructs, models, or methods can be applied to an existing system



In **Figure 5.2** below, Hevner et al. (2004) depict the IS framework by integrating the design science and behavioural science paradigms. This illustration aims to assist researchers with understanding, executing and evaluating IS research. Design science is a suitable methodology for this study as it provides the technique required to produce and evaluate a research artefact while including relevant stakeholders in the study context.

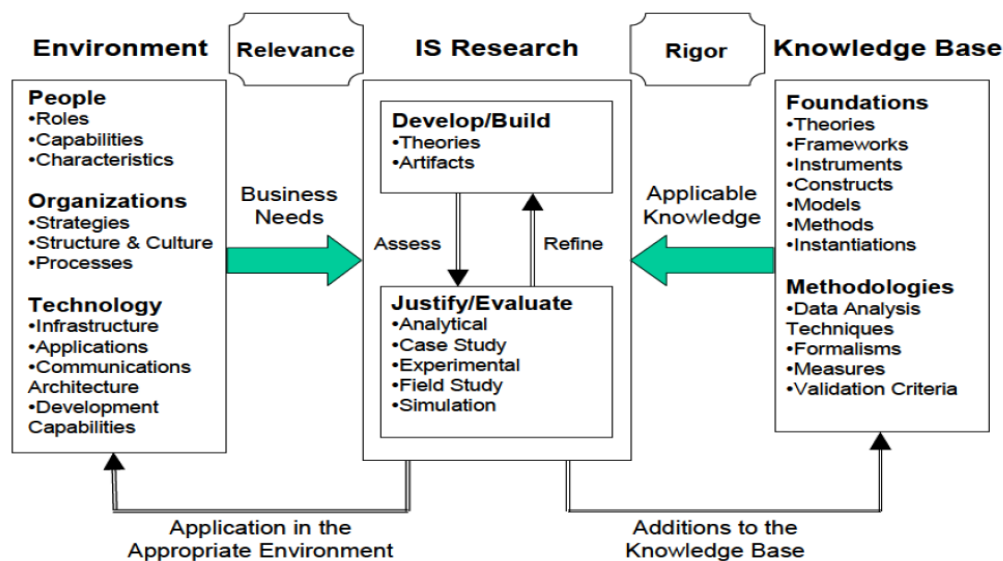


Figure 5.2: Information Systems Research Framework (Hevner et al., 2004)

In the illustration above, Hevner et al. (2004) identified the impact of the environment and existing knowledge base in IS research. The environment included people, organisations and technology, which define the business problem and form the context of the research study (shown as a business need). For example, this research study was in the context of Hadoop technology used within the automotive manufacturer.

The knowledge base consists of existing theories and methodologies which are used to construct a research artefact. After reviewing the business need (environment) and the current theories and methodologies (knowledge base), it was distinguished that IS research is investigated in two stages: build and evaluate. This process indicated that the researcher would build an artefact, evaluate it against primary data collection methods, and refine it until the artefact addresses the identified business need. Until the researcher has subsequently refined the artefact, it will then be applied to the environment, and the artefact will become an addition to the knowledge base (Hevner et al., 2004). This means the evaluation of the artefact should be a rigorous contributing factor to the knowledge base. Evaluation would indicate whether the artefact works, does not work and how it is associated with experiences (Drechsler & Hevner, 2016). If the artefact is evaluated to be viable and can reveal its worth with evidence, it needs to address a criteria, namely: validity, utility, quality and efficacy (Gregor & Hevner, 2013).

In this study, the literature review process-initiated relevance by assessing the people, organisation, and technology in the environment. Within the environment, relevance was commended through the interview and questionnaire process. Furthermore, the context within the big data environment in the automotive manufacturer was identified in the literature content of this research study.

Design science is suitable to studies that produce an artefact, such as the framework in this research project. The framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer, was constructed as an artefact, where the rigour and relevance were applied and evaluated against. The framework was proposed and refined by the application of empirical findings.

Research rigour is established in this study through the foundational theory, which was reviewed in the literature review. The SOIPSVM and the CMM, and the design science research guidelines provided rigour to the research process. The research design established rigour through the seven design science guidelines, as shown in **Table 5.3**. The framework was applied to the concerned environment, while it contributed to the knowledge base of IS.

University of Fort Hare

Table 5.3: Design Science Research Guidelines (Hevner et al., 2004)

<u>Guideline</u>	<u>Description</u>	<u>Application to Study</u>	<u>Chapter Concerned</u>
1. Design as an Artefact	Design science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation.	This study generated a framework to address privacy concerns of big data in Hadoop technology in the automotive manufacturer.	Establishment of the framework (Chapter Seven)
2. Problem Relevance	The objective of design science research is to develop technology-based solutions to	In this study, the identified problem is prevalent privacy concerns of big data in	Literature review (Chapter Two, Three and Four)

	address important and relevant business problems.	Hadoop technology in the automotive manufacturer. This contributes to the automotive industry not recognising big data's associated benefits, causing the industry to lose business competitiveness. A solution is established by IT within the context of the research problem.	
3. Design Evaluation	A design artefact's utility, quality, and efficacy must be rigorously demonstrated via well-executed evaluation methods.	The proposed research framework is evaluated through applicable primary and secondary data gathering and analysis techniques.	Literature review and data collection and analysis (Chapter Two, Three, Four and Six)
4. Research Contributions	Effective design science research must provide clear and verifiable contributions in the design artefact, design foundation, and/or design methodologies.	The contribution of this study is the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer. This was considered a foundation	A framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer. (Chapter Seven)

		contribution as it contributes to the existing IS knowledge base.	
5. Research Rigor	Design science research relies upon applying rigorous methods in both the construction and evaluation of the design artefact.	This research study will utilise applicable data gathering and analysis techniques. The framework will be evaluated against interviews.	Research evaluation (Chapter Six and Seven)
6. Design as a Search Process	The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.	This guideline will be applied by conducting interviews to ensure applicability to the problem domain.	The empirical process was conducted in the context of the research process. (Chapter Six)
7. Communication of Research	Design science research must be presented effectively both to technology-oriented as well as management-oriented audiences.	This guideline will be fulfilled by writing the research paper, outlining the contribution, and summarising the research project.	The research output is the framework which indicates the findings of the research project. (Chapter Seven)

The research design used in this study is design science. **Table 5.3** depicts the seven design science research guidelines applied respectively to this study's output. However, there are limitations with the application of design science research to this research study, as the 7 steps could also indicate false-positive or false-negative results. A false-positive would mean the findings indicate that the artefact works, but infact it does not work in the environment. The false-negative would indicate in the findings that the artefact does not work, when infact it does work in the environment (Gregor & Hevner, 2013). Furthermore,

another limitation is not being able to apply the 7 steps of design science research from an empirical perspective. A framework was constructed to address the identified research problem. The following section described the additional theory, which was used in this study, namely: CMM.

5.3.2 Capability Maturity Model

The CMM signifies five levels of increasing organised and systematically mature processes (CMMI Institute, 2017). This enables continuous improvements for organisations to achieve high-performance operations (CMMI Institute, 2017). The CMM was used in this study by comparing the identified privacy solutions against the CMM stages at different operational levels. The aim was to determine a solution for improvement at every stage until optimisation. **Figure 5.3** illustrates the five stages of the maturity model. Each stage is explained briefly in the sections that follow.



Figure 5.3: Five stages of the capability maturity model (CMMI Institute 2017)

5.3.2.1 Initial

At an initial level, processes are undocumented and in dynamic change. Processes are only performed for a specific purpose, making the processes unpredictable and reactive (CMMI Institute, 2017). The organisation does not enforce a stable environment, resulting in its success being dependent upon the capabilities of people and not the usage of proven processes.

This level signifies that organisations overcommit, abandon processes in an urgent situation, and cannot repeat their past success (CMMI Institute, 2017 & International Software Testing Qualifications Board, 2016).

5.3.2.2 Repeatable

This level of maturity signifies that some processes are repeatable to ensure consistent results. Processes are unlikely to be rigorous, but if rigorous processes persist, it ensures that existing processes are maintained in an urgent situation (CMMI Institute, 2017; International Software Testing Qualifications Board, 2016).

5.3.2.3 Defined

The organisation implements a standard set of processes and controls which are defined and documented. The standards act as measures of guidance and are subject to improvement over time. This maturity level is seen as proactive instead of reactive (International Software Testing Qualifications Board, 2016).

5.3.2.4 Managed

At this level, process metrics are used to achieve process objectives across various operational conditions. The process is quantitatively managed according to an agreed set of metrics and aligned to meet the requirements of internal and external stakeholders. Therefore, at this level, process capability is established (CMMI Institute, 2017).

5.3.2.5 Optimising

The last level of maturity is seen as being stable and flexible (CMMI Institute, 2017). Processes at this level focus on continuous improvement through incremental and innovative technological changes. Furthermore, the organisation can respond to opportunities and change (International Software Testing Qualifications Board, 2016). The next section discussed the research format of this study.

5.4 Research Format

Collis and Hussey (2009) identified and classified research according to the purpose of the relevant study, namely: exploratory, descriptive, analytical and predictive research. The following contrasts are associated with the various formats.

- *Exploratory*: Exploratory research is conducted in new areas of research, where there is very little literature on the research problem. This type of research aims to look for patterns, ideas, or hypotheses that enable the researcher to understand the research problem (Collis & Hussey, 2009).
- *Descriptive*: Center for Innovation in Research and Teaching (2019) indicates that descriptive research provides a detailed analysis of a situation, subject, behaviour or phenomenon. It is used to identify and acquire information on a specific problem which will assist the researcher with examining the identified problem furthermore (Collis & Hussey, 2009).
- *Analytical*: Analytical is also known as explanatory research, which is a continuation of descriptive research. This type of research aims to understand the phenomena and why outcomes have occurred by discovering and measuring causal relations (Collis & Hussey, 2009).
- *Predictive*: Predictive research aims to generalise by predicting the foundation of hypothesized and general relationships. Therefore, the solution to a research problem in a study can be generalised to other studies (Collis & Hussey, 2009).

The research format applicable to this study was descriptive research. Research studies can either follow inductive or deductive reasoning. Deductive reasoning involves developing a theory tested through empirical observations (Collis & Hussey, 2009). This type of reasoning results in particular instances being deduced from general conclusions. In contrast, inductive reasoning develops a theoretical structure from empirical observations, where a generalised conclusion is induced from instances (Collis & Hussey, 2009).

This research study was based on the inductive reasoning approach. Therefore, the researcher began with general observations by formulating a main research question and sub-questions, where patterns were identified. The general conclusions were recognised through recommendations based on a framework to address privacy concerns of Hadoop technology in the automotive manufacturer. The above discussion incorporated the

research purpose and logic of this study. The data collection methods used to obtain and analyse primary and secondary data in this study were discussed next.

5.5 Data Collection Methods

The data collection of this study incorporated both primary and secondary resources. These resources assisted the researcher with answering the associated research questions and met the objectives of this study. According to Myers (1997), primary data sources are referred to as unpublished data and gathered from the participants within the organisation concerned. In contrast, secondary data is referred to any previously published materials, such as books and articles (Myers, 1997). The primary and secondary data collection methods used in this study are discussed in the following section.

5.5.1 Primary Data Collection Methods

All respondents were sent the ethical clearance (Appendix A), informed consent form and an overview of the study (Appendix B). The primary data collection methods which were used in this research project were interviews and questionnaires. Interview questions and the questionnaire (Appendix C) were presented in a Portable Document Format (PDF) and sent to all respondents via email before the data collection session. The questions compiled for the interview and the questionnaire were derived from existing studies and categorised based on the sub-research questions of this study and the SOIPVSM model.

Each session was separated into two parts: the interview and the questionnaire. The results were available instantly as the respondents completed the questionnaire within the duration of the session. The two methods are explained below.

5.5.1.1 Interview

Saunders et al. (2007) define an interview as a discussion between two or more people with a purpose or objective. Furthermore, interviews assist the researcher in obtaining valid and reliable data which is relevant to identified research questions and objectives (Collis & Hussey, 2009)

The concept of interviews can be associated with both the positivist and interpretivist paradigms. Furthermore, the positivistic approach indicates that structured, closed-ended questions should be used in the interview. At the same time, the interpretivist approach suggests unstructured questions, which are not prepared in advance (Collis & Hussey, 2009).

Interviews were conducted with the big data specialists in IT. The interviews were approximately 20 minutes long, where most questions were open-ended, following the interpretivist approach. The interview questions were derived from a previous study by Andersson and Axelsson (2018). The interview questions were grouped together into the below sections based on the literature of this study.

- Generation and use of big data
- Privacy challenges in Hadoop technology
- Measures to protect privacy

Answers to the questions were written by the researcher and recorded on a mobile device, where it was transcribed to Excel and NVivo for analysis. The results were analysed and discussed in Chapter six of this study. The next section discussed the questionnaire process used in this study.

5.5.1.2 Questionnaire

A questionnaire is a pre-defined set of questions, where each person is requested to respond to the set of questions in a pre-determined order (Saunders et al., 2007). The questionnaire enables the researcher to elicit the responses, analyse and interpret the data to address the identified research problem (Collis & Hussey, 2009).

For this study, the interview was first to be conducted. Afterwards, the respondents were requested to complete a short questionnaire, resulting in a positivistic element in this study. The questionnaires consisted of close-ended questions, where respondents were requested to provide a rating to further provide the researcher with empirical data for analysis and interpretation. The questionnaire was derived from the SOIPSVM study which focused on

the consequences of violating privacy and measures to protect privacy of big data in Hadoop technology at the automotive manufacturer (Wall, Lowry, & Barlow, 2016).

5.5.2 Secondary Data Collection Methods

This research project incorporated secondary data, which was collected from various sources. The information obtained was used throughout the research process in this study. Secondary sources included the following:

- Conference proceedings
- Academic websites and articles journals from academic journals
- Various internet sources
- Theories and frameworks
- Past research projects

All secondary data has been referenced, ensuring that credit has been given to the original author. Furthermore, all effort has been made to ensure that the content of this study is current and relevant to the research project.



University of Fort Hare
Together in Excellence

5.6 Population and sampling method

The population in a research study refers to the total number of members who conform to a certain criterion (Chaudhury & Banerjee, 2010). However, it is not a feasible option to recruit the entire population into a research study. Researchers should recruit a sample size based on the population (Majid, 2018). In this study, the population consisted of 67 members from the IT Department in the automotive manufacturer. The sample size was 10 IT specialists with specialised knowledge related to big data. Primary data in interviews and questionnaires were collected from the IT department at the automotive manufacturer concerned.

Further, ethical clearance from the University of Fort Hare's Research Ethics Committee (UREC) was obtained before the data collection was conducted, which complied with the ethical regulations stated by UREC (Appendix A). Approval for the use of data was

obtained from the IT management of the company. The next section discussed the sampling method used in this study.

5.6.1 Sampling method

Majid (2018) refers to a sampling method selecting a statistical representative of individuals based on the population size. Two sampling methods can be adopted into a research study probability and non-probability sampling (Taherdoost, 2016).

The probability sampling method applies to quantitative studies, while the non-probability technique is associated with qualitative research (Taherdoost, 2016). Each sampling method has different techniques associated, where the researcher can choose one most applicable technique. This study was qualitative, and the sampling method most suitable was the non-probability strategy. Furthermore, the purposive sampling technique was applied to the study.

According to Etikan, Musa and Alkassim (2016), the purposive sampling technique is used to choose participants based on the individual's quality. This includes individuals who are well informed about specific phenomena. This sampling method enabled the researcher to concentrate on the respondents who would constructively assist the relevant research study. As a result, the sample size in this study included ten specialists in big data. The job role of the ten respondents included the system administrator, business analyst, graduate students, data scientist, data engineer, IT manager, software developer and solutions architects. As a result of selecting the purposive sampling technique for the qualitative study, the principle of saturation was also applied. The following section discussed the principle of saturation.

5.6.2 The principle of saturation

Saunders et al. (2018) define the principle of saturation as the point when obtained data is no longer providing new information about the research problem being investigated. The study's researcher found that using ten participants, which comprised employees working in IT, resulted in saturation. Thus, no further interviews and questionnaires were conducted.

5.7 Data analysis

Data analysis refers to processing data into meaningful insights (Kumar, 2011; Collis & Hussey, 2009). This study used two approaches for data analysis, namely: pattern matching and thematic analysis. The following tools were used to analyse the primary data gathered from the interview and questionnaire.

- *Microsoft Excel*: Excel was used to sort the data and generate visuals.
- *NVivo 11*: NVivo was used to analyse the text compiled from the interview and questionnaire.

5.7.1 Pattern matching

Pattern matching is a data analysis method used in qualitative research. It involves the comparison of a predicted theoretical pattern with an observed empirical-based pattern. The purpose of the pattern matching techniques is to strengthen the rigour of the study. If the empirical-based pattern matches the predicted ones, the findings contribute towards the study's internal validity (Sinkovics, 2018). The theory of pattern matching is depicted in **Figure 5.4**.


 University of Fort Hare
 Together in Excellence

The top part of **Figure 5.4** depicts the theoretical realm. Trochim (2001) states the theory may originate from a formal tradition of theorizing, ideas of the researcher, or a combination of both. The conceptualisation task involves the translation of the ideas into a specific theoretical pattern. The bottom section of **Figure 5.4** is the observation realm. The observation realm includes direct observations in the form of field notes and formal objective measures. The collection of operationalisations relevant to the theoretical realm is called the observational pattern. The deductive task involves the link between the two patterns. Based on the extent that the patterns match, the research can conclude that the theory used may predict the same observed pattern (Trochim, 2001).

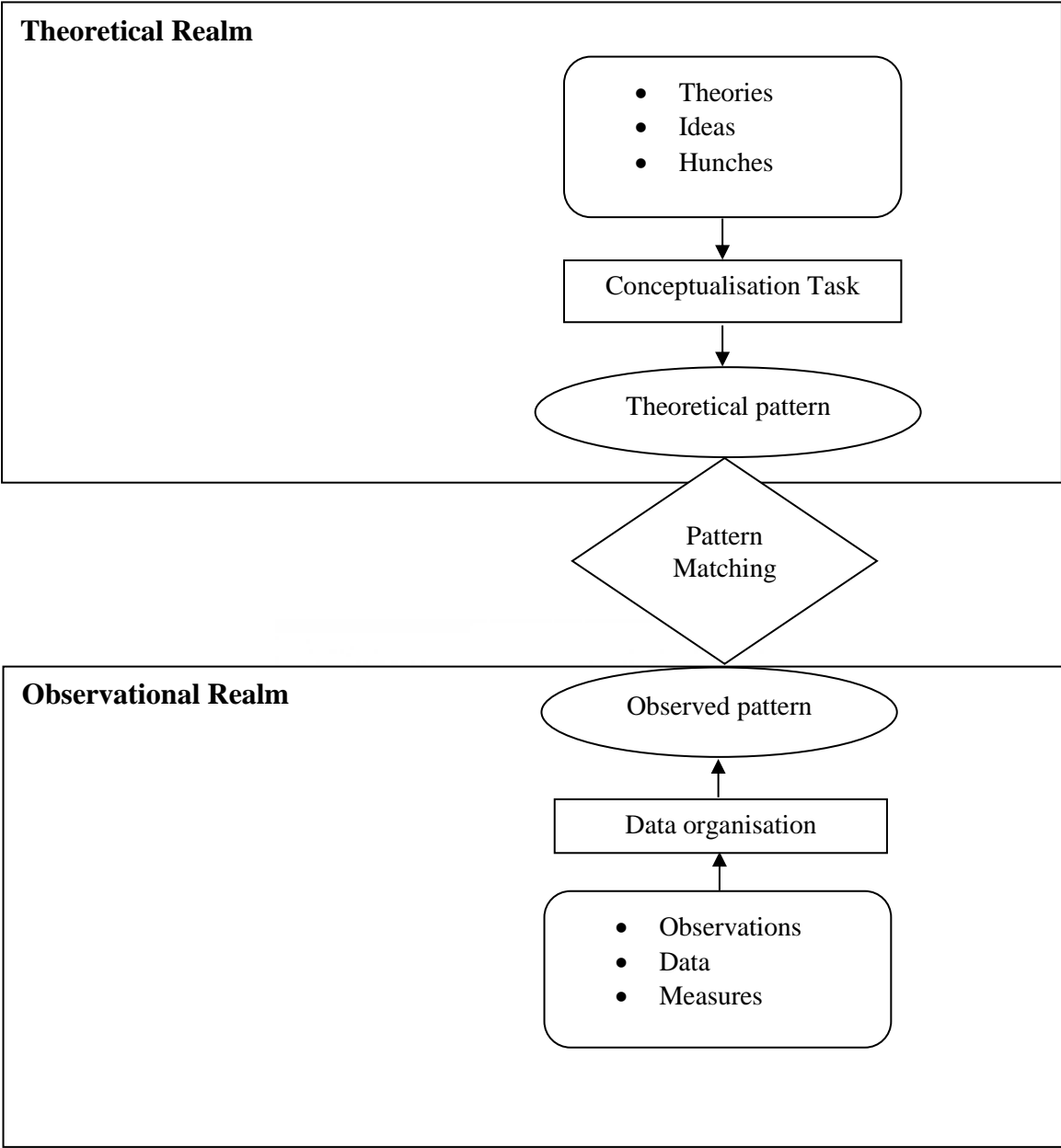


Figure 5.0.4: The theory of pattern matching (*Trochim, 2001*)

This research study used pattern matching to link the theoretical pattern with the observed realm (Sinkovics, 2018). The theoretical pattern used in this study included the SOIP SVM and existing studies discussed in the literature review. The data collected in this study in the interviews and questionnaire constituted as part of the observational realm. The theoretical and observed realms were analysed to identify any similar patterns. The next step for analysis was to identify relevant themes in this study through thematic analysis.

5.7.2 Thematic analysis

Braun and Clarke (2006) define thematic analysis as a method for identifying, analysing and reporting themes within data. This research study adopted Braun and Clarke's six phased approaches designed for thematic analysis. Furthermore, the NVivo research analysis software tool and Microsoft Excel was used in this study. The six phases of thematic analysis are: familiarising yourself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report (Braun & Clarke, 2006). Each phase of thematic analysis was discussed in relation to this study.

5.7.2.1 Familiarising yourself with the data

The researcher acquaints themselves with the data collected to ensure an in-depth and sufficient understanding of the data contents (Braun & Clarke, 2006). The researcher of this study conducted the interviews and the questionnaire. Therefore the process of understanding the data content was easier.

The interview and questionnaire were structured into a format which was categorised according to the identified sub-research questions, namely:

- *How is big data being generated and used in the automotive manufacturer?*
- *What are the types of privacy challenges experienced within the Hadoop environment in the automotive manufacturer?*
- *What measures can the automotive manufacturer take to protect their privacy?*

The interviews and questionnaire were transcribed into three main themes: generation and use of big data, privacy challenges, and privacy protection. All data collected from the interview and questionnaire were transcribed into an Excel spreadsheet, which formed part of the cleaning process and getting familiar with the data. The next phase was to generate the initial codes.

5.7.2.2 Generating initial codes

After obtaining an in-depth understanding of the datasets collected, the researcher was able to make a thorough comparison of important phrases that impacted the main research

question (Braun & Clarke, 2006). Saldaña (2013, p.3) indicates that coding in qualitative research is represented by a “word or short phrase that symbolically assigns a summative, salient, essence-capturing, and or evocative attribute for a portion of language-based or visual data.”

The themes and sub-themes depicted in **Table 5.4** were identified and extracted from the interview and questionnaire. Each sub-theme was aligned to the primary data collection methods to understand the impact on one or both primary data collection methods used in this study.

Table 5.4: Themes and sub-themes of the study

Main themes	Sub-themes	Interview	Questionnaire
Theme 1: Generation and use of big data	1.1 Sampled demographics	X	
	1.2 IoT and effect on the automotive manufacturer	X	
	1.3 Data generation and collection	X	
	1.4 Big data storage	X	
	1.5 Required tools and techniques	X	
	1.6 Creating value from data	X	
	1.7 Organisational culture	X	
Theme 2: Privacy challenges	2.1 Privacy awareness	X	X
	2.2 Access control to IoT devices	X	
	2.3 Associated organisational risks	X	X
	2.4 Consequences of violating privacy in Apache Hadoop		X
Theme 3: Privacy protection	3.1 Impact on core organisational values		X
	3.2 Methods to protect privacy	X	

Identifying the themes and sub-themes after the six-phased thematic analysis was important as it referred to the empirical and analysis chapter (Chapter 7) of this study. Upon completing the coding phase, the next phase entailed searching for themes discussed in the next section.

5.7.2.3 Searching for themes

Nowell, Norris, White and Moules (2017, p.8) define the concept of a theme as “an abstract entity that brings meaning and identity to a recurrent experience and its variant manifestations. As such, a theme captures and unifies the nature or basis of the experience into a meaningful whole.”

This phase begins after all data has been collated and the researcher has established a list of codes identified from the dataset. The third phase re-focuses on the analysis and entails sorting the codes into potential themes and collating all coded data extracts within the themes. After that, the relationship between the codes, themes and the different levels of themes can be established (Braun & Clarke, 2006).

This research study constructed the three main themes from the sub-research questions and the literature review. The coded data was integrated into the relevant themes, which are illustrated in **Table 5.4**. All themes were reviewed to evaluate the degree of rigour and relevance.



5.7.2.4 Reviewing themes

During this phase, the researcher refines the identified themes. This entails reviewing the coded data extracts for each theme to determine if it forms a coherent pattern. The validity of each theme is considered to establish it reflects the context in the datasets. Inefficiencies in the initial coding and themes can be identified in this phase, and changes can be made (Nowell et al., 2017).

In the context of this research study, the researcher re-checked the themes to identify rigour and key elements to address the research problem. The themes and sub-themes were reviewed against the objectives of this study. After completing the review of identified themes in this study, the themes were defined and named.

5.7.2.5 Defining and naming themes

For each theme, the researcher is required to conduct a detailed analysis. The detailed analysis entails identifying the story behind each theme and how the theme fits into the

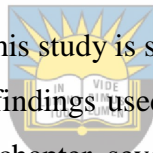
overall story. This process ensures that the data in relation to the research questions do not contain overlapping themes. The researcher will identify whether sub-themes exist within the theme (Braun & Clarke, 2006).

The three objectives of this study were used to define and name the main and sub-themes. The final themes and sub-themes for this study are illustrated in **Table 5.4**. The last step in the thematic analysis is producing the report.

5.7.2.6 Producing the report

Producing the report in the form of a write-up of the communication of findings is the last step of thematic analysis. This involves merging the analytical narrative and the coded data extracts to provide the reader with a coherent story about the data in context to existing studies. The report should focus on key aspects which address the research problem (Braun & Clarke, 2006).

The empirical process followed in this study is summarised into four sequential phases, as illustrated in **Figure 5.5**. The key findings used to develop a framework, to address the research problem is discussed in chapter seven of this study. As a result, when the interpretivist approach is used in a study, it needs to be evaluated for quality. The next section discussed the characteristics for the evaluation to be successful.



University of Fort Hare
Together in Excellence

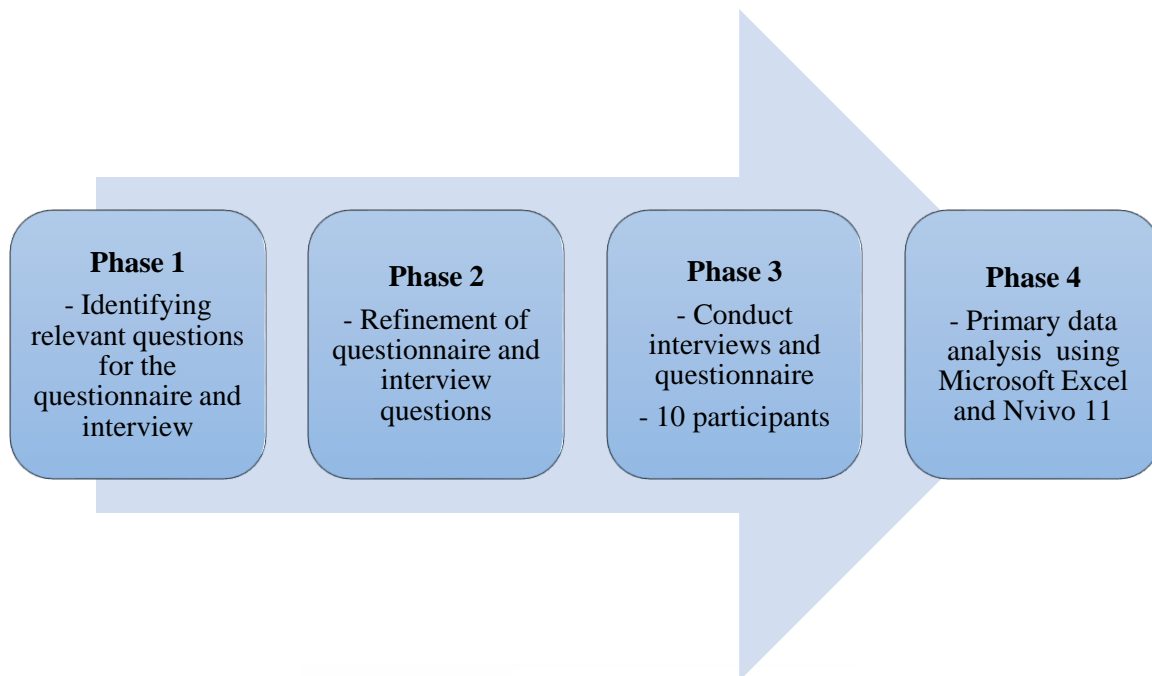


Figure 5.5: Empirical process phases

5.8 Research Evaluation



Oates (2006) provides a set of measurement parameters that can evaluate the quality of the research for both positivistic and interpretivist research. These characteristics must be met for the evaluation to be successful. **Table 5.5** indicates the characteristics.

Table 5.5: Quality in Positivist and Interpretivist research (Oates., 2006)

Positivism	Interpretivism
Validity	Trustworthiness
Objectivity	Confirmability
Reliability	Dependability
Internal validity	Credibility
External validity	Transferability

The interpretivist paradigm was used to evaluate this research study in chapter eight. The criteria for this paradigm are described below (Oates, 2006).

- *Trustworthiness*: The amount of trust which can be placed on this research study.
- *Confirmability*: This is concerned with whether sufficient information has been gathered and analysed to determine whether the findings flow from the data and experience of the setting.
- *Dependability*: The accuracy of the research process and documentation.
- *Credibility*: The research problem was accurately identified and described so that research findings are creditable.
- *Transferability*: The findings of this study can be transferred to other research areas.

5.9 Conclusion

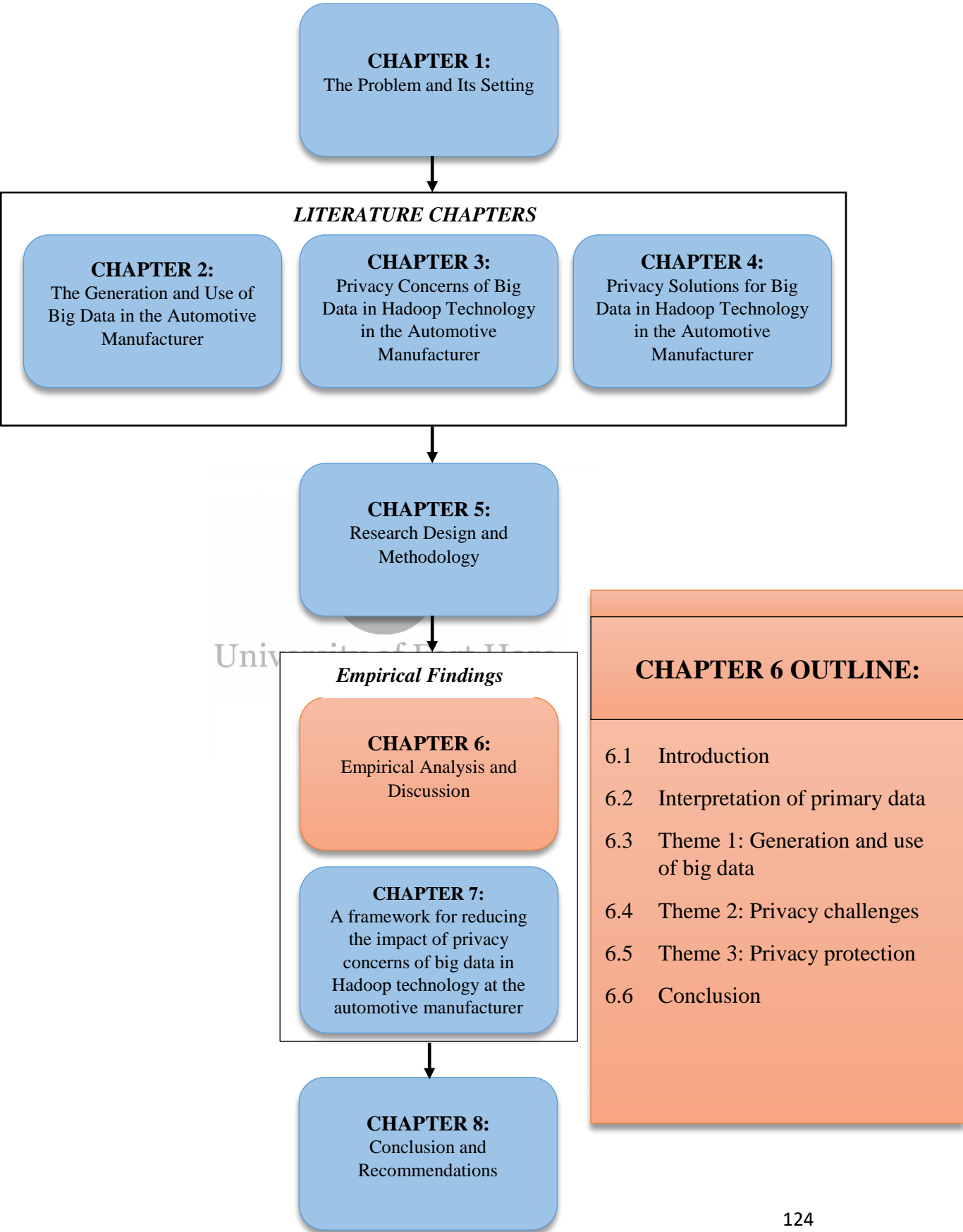
IoT devices connected to the automotive production line generates large amounts of data. The generated data is analysed using big data analytics to optimise supply chain management and improve the automotive manufacturing process. The automotive manufacturer utilises Hadoop technology to manage, process and store complex data. Although the data generated and analytics provides an optimised automotive production process, there are privacy concerns in Hadoop technology that cannot be ignored. Examples of information privacy concerns identified included: data breaches and fragmented data.

The root cause of information privacy concerns indicated that Hadoop technology has an ineffective and failed security model. If these concerns were left unaddressed, the automotive manufacturer would be at risk due to the security vulnerability. This means sensitive information would be available for access and can potentially be tampered with by unauthorised users. In effect, company trade secrets can become public, use cases of big data in the automotive manufacturer will not be realised, automotive production levels and quality will not be met, which is detrimental to the automotive manufacturer's reputation. This research study aimed to address the information privacy concerns in Hadoop technology through critical success factors that were extended into a framework.

The philosophical research paradigm of interpretivism was applied to this study because of the subjective nature and complexity of information privacy of big data in Hadoop technology. The research format of this study was descriptive and incorporated inductive reasoning. Furthermore, the design science methodology was discussed and evaluated to be suitable due to the artefact produced in this study. The artefact was in the form of a framework that addressed the research problem.

The primary data collection method used in this study is the qualitative data collection approach. Primary data was collected in interviews and questionnaires from 10 big data IT specialists in the automotive manufacturer. The principle of saturation was applied after no new information was obtained in the primary data collection process. Secondary data was also considered to understand the context of the research problem and create the foundation of the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer. The data analysis of this study was conducted using two applications, namely: Microsoft Excel and NVivo 11.

Furthermore, pattern matching and thematic analysis were incorporated into the data analysis section of this study. Lastly, this chapter covered the research evaluation used to assess the quality of an interpretivist research study. The next chapter presented the empirical analysis of this study.



Chapter 6

Empirical Analysis and Discussion

6.1 Introduction

The previous chapter discussed the empirical research process used in this study to collect and analyse the primary data. This study used the interpretivist paradigm as a guide in the research process. Furthermore, the qualitative approach in an interview and questionnaire was used in this study's primary data collection process. Chapters two, three and four reviewed the literature content, which provided the foundation in answering the main research question. The objective was to address the main research problem discussed in chapter one, which was stated as follows: *How can the privacy concerns of big data in Hadoop technology at an automotive manufacturer be addressed?*

This chapter presented and discussed the findings from the empirical research conducted. The empirical process applied to this study is illustrated in **Figure 5.5**. The purpose of the empirical analysis was to explore further the concepts discussed in the literature review of this research study. The concepts were put into the research context for Information Technology (IT) professionals to provide practical opinions and explanations.

The interview and questionnaire (Annexure C) were compiled by selecting and refining relevant questions from the Selective Organisational Information Privacy and Security Violations Model (SOIPSVM) and existing studies. These questions were categorised into different themes. Each question against the applicable theme was analysed and categorised into a sub-theme respectively. The identified themes and sub-themes are illustrated in **Table 5.4**.

The interview and questionnaire were conducted with ten employees at the IT department in the automotive manufacturer. The respondents represented a distribution of various IT positions, responsibilities, and levels of experience within the organisation. Furthermore, all respondents were in IT and linked to the industry's big data and analytics team. Therefore, the response rate has been deemed a representation of the IT department in the organisation and was considered acceptable. The results were analysed using Microsoft Excel and NVivo 11. Furthermore, the findings incorporated the principle of saturation and pattern matching to compile a framework that addressed the identified research problem. The following section discussed the analysis and interpretation of the primary data collected.

6.2 Interpretation of primary data

The findings of each theme, literature content, participants responses and the researcher's conclusions were included in the discussion to provide a rich and concise perspective on the privacy concerns of big data in Hadoop technology in the automotive manufacturer. Direct quotes from the respondents were provided in italics in the analysis discussion. Each theme and its applicable sub-theme were key in formulating the conclusion, in answering the main research question. The first theme discussed in the next section was the generation and use of big data in automotive manufacturers.

6.3 Theme 1: Generation and use of big data

The production line in the automotive manufacturer contains advanced technologies, such as the Internet of Things (IoT) which results in large amounts of data being generated from various internal and external data sources (Ismail, Truong, & Kastner, 2019). This data is analysed for realising use cases and opportunities such as optimising supply chain management, improving the automotive manufacturing process and customer service, which was discussed in detail in Chapter two of this study (Deloitte, 2019; Russo, Confente, & Borghesi, 2015).

However, Ajah and Nweke (2019) state that effective data analytics must be conducted to realise the associated benefits and use cases in the production process. As a result, this theme sought to gain insight on IoT devices, data output and practical uses from IT experts in the automotive manufacturer.

The aim was to obtain clarity on the process from data generation to realising current use cases in the automotive manufacturer. The sub-themes that explore this concept were: sampled demographics, the effectiveness of IoT devices in the automotive manufacturer, big data generation and collection, big data storage, required tools and techniques, creating value from data and organisational culture. Additionally, the generation and use of the big data theme focused on the first research sub-question: *How is big data being generated and used in the automotive manufacturer?* The sub-themes were discussed in the following section.

6.3.1 Sub-Theme 1.1: Sampled demographics

Distinct roles have been established for all highly skilled professionals who work with big data. All team members work differently but have the same objective in realising the use and benefits of using data to improve processes (McAllister, 2018). Therefore, it was important to understand how various IT specialists contribute to the big data phenomenon in the automotive manufacturer.

This sub-theme was intended to establish and provide detail on the IT role of each respondent from the big data team in the automotive manufacturer. The interview data analysed the IT role of each respondent. This enabled the answers to be assessed within each role category, as shown in **Figure 6.1**. Based on the feedback provided by each respondent, each role was elaborated upon. It was important to note that one respondent shared both the system administrator and data engineer roles.

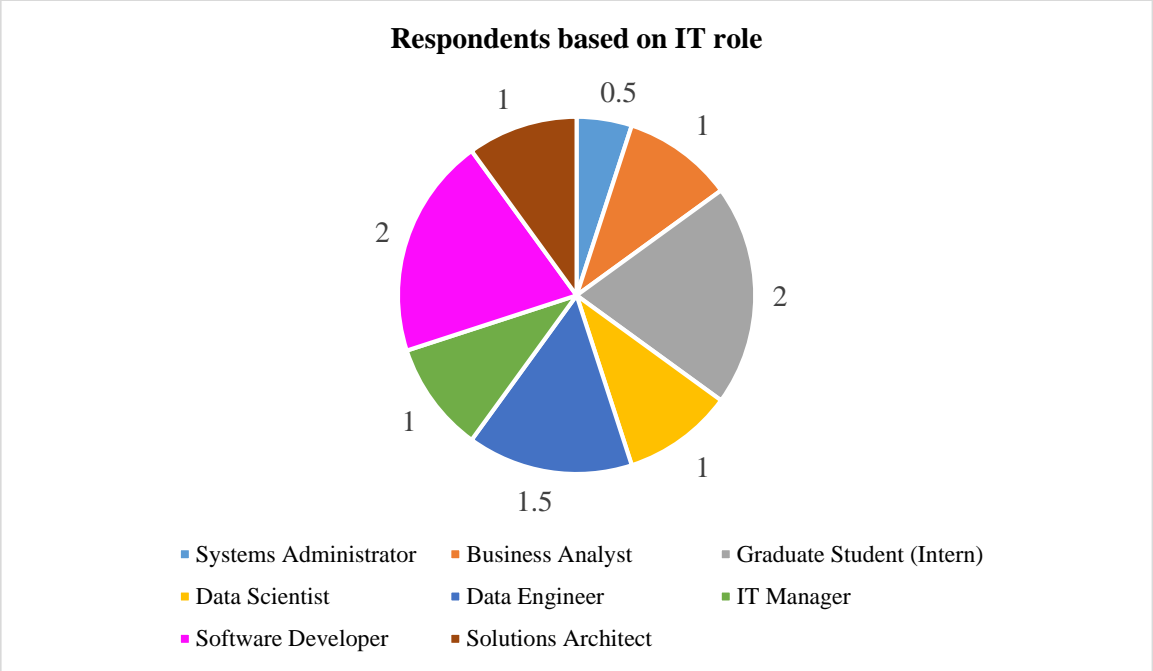


Figure 6.1: Respondents based on IT role

Each respondent was asked to describe their job function briefly. The purpose of this was to understand each IT role in a practical context to big data in the automotive manufacturer. Each position was described as follows.

- System Administrator

“Maintaining files.”

- Business Analyst

“I am a Business Analyst within the production space in IT looking after quality systems, big data topics and second level support for engineering applications. In terms of big data, I look at getting use cases and check how they can be implemented in the plant and how artificial intelligence and machine learning can be used to bring value to the plant.”

- Graduate Student

“My job entails software development. Learning, understanding technologies and skills required.”

- Data Scientist

“Involves collecting and analysing data for improving business processes.”

- Data Engineer

“I work on the data collection platform. I am involved with the development, operations, and data engineering, moving data from one place to another.”

- IT Manager

“Responsible for production and shopfloor systems. It's all about coordinating the teams and making sure the required service level agreements have been fulfilled.”

- Software Developer

“Develop software that solves business use cases.”

- Solutions Architect

“Design and build big data, artificial intelligence and IoT solutions.”

Discussion of findings



The overall view was that various IT specialists were key members of big data in the automotive manufacturer. Each member interacted differently with the data; however, they would collaborate to realise the value obtained from big data. The key finding deduced from this sub-theme was the automotive manufacturer needed a central team of IT specialists, who used the data in various ways based on their function.

To understand the interview and questionnaire perspectives, each specialist's role was given an identifier used throughout the analysis and discussion of this study. This enabled the researcher to identify the responses of each respondent based on their role. **Table 6.1** represented the identifiers that had been allocated against each IT specialist role from **Figure 6.1**. The next sub-theme was related to the perception of IoT devices and their effect on automotive manufacturers.

Table 6.1: Research participants identity

IT Specialist Role	Number of Respondents	Identifier
System Administrator	0.5 -Shared role	SADE1
Business Analyst	1	BA1
Graduate Student	2	GS1, GS2
Data Scientist	1	DS1
Data Engineer	1,5 – Shared Role	DE1, SADE1
IT Manager	1	ITM1
Software Developer	2	SD1, SD2
Solutions Architect	1	SA1
Total Respondents	10	

6.3.2 Sub-Theme 1.2: The Internet of Things and its effect on the automotive manufacturer



Industry 4.0 is the new period for merging IoT technology and the automotive manufacturer (Subrahmanyam & Aruna, 2017; Tay, Chaun, Aziati, & Ahmed, 2018). Based on the interview data collected, this sub-theme highlighted the extent of IT professionals' understanding of IoT technology and its impact on the automotive manufacturer. The IT specialists were asked to describe the IoT and provide their perspective on its effect on the automotive manufacturer.

IT specialist DS1 described the IoT as follows:

“IoT is a collection of things connected by the internet. Many devices connected that talk to achieve a certain goal, gain insight, things working together.”

Furthermore, IT specialist ITM1 pointed out that IoT had transitioned in connecting devices compared to the past.

“In the past small and big devices could be connected to provide data, but this is about connecting things which couldn't be connected before.”

IT Specialist SD1 expressed the challenges of connecting IoT devices in the automotive environment.

“Directly, devices which are providing us with more data and measuring points. How these devices are plugged into the environment is a challenge on its own.”

Based on the responses from the IT specialists, it was evident that IoT is a network of devices connected to provide the automotive manufacturer with massive amounts of big data. However, connecting devices on the production line can be complicated. Furthermore, all respondents acknowledged that IoT was positively affecting the automotive industry.

IT professional ITM1 believed that,

“From an automotive perspective, there will be lots to learn, and it will help in adding value to what our colleagues do.”

Similarly, IT professional SADE1 provided an example of how IoT devices can be used to add value to business processes:



“Predictive maintenance for part failure. Also, identify the criticalities and identify which stations on the production line require more people.”

The responses from the IT professionals indicated that IoT was assisting the industry in having a more connected production line. IoT devices connected to the production line enabled the visibility of assets, processes, resources, and products. This supports more efficient operations and improved productivity (Gelmato, 2020).

Discussion of findings

Industry 4.0 has resulted in the digitalised manufacturing sector being established, where IoT devices such as sensors are built into all manufacturing equipment, components, and products (Tay et al., 2018). The establishment of IoT devices has created a platform for digital transformation, which significantly impact automotive manufacturers (Mallon, 2018). This study found that IT professionals had understood the advancement of technology, resulting in IoT devices being used in the organisation.

Furthermore, the associated benefits of IoT devices have been recognised. However, since IoT and big data is new digitalisation topic, there was a constant learning culture amongst the IT professionals to realise the associated use cases and how to plug these devices into the environment. Therefore, the automotive manufacturer must create a workplace of collaboration, learning and developing the professional's competencies (Machado, Winrotha, & Ribeiro da Silva, 2019).

To this end, the key finding deduced from this sub-theme was that IoT devices had a positive effect in automotive manufacturers. The next sub-theme discussed data generation and collection.

6.3.3 Sub-Theme 1.3: Big data generation and collection

Automotive production is a complex process that involves a high logistical effort. The production line is split into four main facilities: press, body, paint, and assembly (Bysko, Krystek, & Bysko, 2020).



University of Port Harcourt
Together in Excellence

- *Press shop*: Raw materials supplied by external suppliers such as steel, aluminium and plastic are transformed, shaped, and moulded to form panels, such as the side frames, doors, bonnets, and roofs (Giampieri, Ling-Chin, Smallbone, & Roskilly, 2020).
- *Body shop*: The panels are then mounted together to form the vehicle shell, also referred to as the body in white. This process involves welding operations, joining, and forming, which is done automatically by robots or manually. The body in white is then transferred to the paint shop facility (Giampieri et al., 2020).
- *Paint shop*: This facility entails transferring the body in white to various dip tanks for cleaning, protecting corrosion, and preparing the metal for painting (Bysko et al., 2020). The painted vehicle body is then transferred to the assembly facility through conveyors (Giampieri et al., 2020).
- *Assembly*: This is the last step in the vehicle production process, where the assembly sub-components are mounted onto the painted vehicle body. Furthermore,

vehicle quality, inspection and testing checks are performed (Giampieri et al., 2020).

IoT devices and the production line development have increased big data generated from various internal and external data sources in the automotive manufacturer (Syafurudin, Alfian, Fitriyani, & Rhee, 2018). Examples of big data collected include sensors, Radio-Frequency Identification (RFID), robotic and application data. IoT technologies relevant to the automotive production line use a Manufacturing Service Bus (MSB), a flexible integration concept for intelligently and digitally networking the automotive manufacturer.

The MSB is a global interface used for connecting and integrating various old and new technologies to be used simultaneously. The MSB controls the large and complex amounts of big data generated (Fraunhofer IPA, 2016). The MSB contains Message Queue Telemetry Transport (MQTT) protocols that collect and transport big data into relevant storage architectures. Furthermore, MQTT is an open-source protocol or a message broker responsible for receiving messages, filtering them, deciding who is interested in them, and sending the message to the subscribed clients (Cherradi, Boulmakoul, & El Bouziri, 2016). After that, data analytics is applied to monitor the production process (Syafurudin et al., 2018).

This sub-theme focused on analysing interview data and understanding how big data can be generated and collected within the automotive production process. Furthermore, it explored the types of big data generated, areas that can be managed, and data collected within the automotive production area. IT professionals were asked to explain how big data is generated and collected in the automotive manufacturer.

IT Professional DS1 described the process as follows:

“Most of the data comes from robots. Robots have sensors regarding motion, position, temperature, torque, current, voltage, power, and metrics supplied to the industry by the manufacturers of the machines. Bosch’s Programmable Logic Controller (PLC) capture data in 2 milliseconds and expose it to IT. Collected via MSB.”

DE1 further elaborated that:

“All devices on the production line which are capable of recording data do so. For example, robots know their current position and send that position to the person subscribed to the information. Welding guns measure their welding time and current and send that information. All information is sent to a message broker called an MQTT software, which listens to all information and makes it available to all who need to subscribe to it.”

IT professional DS1 explained how IoT devices could be built into technologies where data cannot be generated:

“IoT sensors are embedded in technologies where we can’t get data. Systems transmit data to us. Technology experts or technology manufacturers are heavily involved and supply IoT sensors.”

In other words, the automotive production line contains machines that are integrated with IoT devices, such as sensors. Furthermore, IoT devices can be built into technologies where data cannot be generated. These IoT devices assist with generating data based on variants such as motion, temperature, and power. All the information is collected via the MSB and transmitted to subscribed clients through MQTT software. Respondents were asked to describe the types of data collected from the various areas of the production line.

BA1 described the various forms of data which can be collected on the assembly and body shop facilities:

“Data in the form of images. Data in the form and numbers (readings from different technologies on the shop floor). Data can be collected from the production line, mainly assembly and body shop.”

Furthermore, ITM elaborated that all process data is collected:

“Process data - anything to do with machine movement, health status, temperatures, positions. Data can be collected mainly in the body shop and some parts of the assembly line.”

This meant that all data generated from IoT devices could be collected in various forms, based on the technologies used on the production line. The main facilities from which data can be generated and collected are the body shop and assembly facilities. The respondents were further asked to describe what is done with the data which is collected.

GS1 described that the data is used for analysis purposes to identify any faults in machinery: *“Analyse to check factors such as faults.”*

Furthermore, SA1 described that the data is analysed to send information back to the suppliers of the technologies used on the production line. This is to ensure that suppliers provide efficient services to the automotive manufacturer.

“Send data back to suppliers in real-time on technologies. Suppliers can plan better, do diagnostics on systems to ensure better service. Plan better for production, plan better on production capacity. Based on technology and version.”

The data is also used to identify any abnormalities in the production process in advance. ITM stated:

“Collecting data from spot stud, AI model to do predictive weld maintenance and give a predictive analysis of when welds go out of tolerance to enable the responsible maintenance and production employee to take measures to ensure it does not go out of parameter and quality defects.”

The data which is collected is used for analytical purposes to identify any trends and real-time abnormalities. This data assists the automotive manufacturer in improving the manufacturing process by providing insight into technologies, predicting faults, and determining how prescriptive controls can be used to prevent occurrences from happening in the future.

Discussion of findings

IoT devices and automotive manufacturers directly influence each other. An important aspect of collecting data is ensuring that a communication channel exists between the IoT

device and the storage infrastructure (Cherradi et al., 2016 & Fraunhofer IPA, 2016). The data generated can be analysed for various uses to improve the production process and the quality of the product (Deloitte, 2019).

The key finding deduced from this sub-theme is that MQTT and MSB technologies were used as a communication channel from the IoT devices to the storage device. All data generated from IoT devices were collected. This data exists in various forms such as image, text, and robotic data. Once the data is collected, the data needs to be stored to perform analytics. The following section discussed big data storage in detail.

6.3.4 Sub-Theme 1.4: Big data storage

Big data storage refers to the scalable storage and management of complex, large, and unlimited datasets. This ensures that applications can access the data for reading and writing purposes (Cavanillas, Curry, & Wahlster, 2016).

All organisations have a data retention period for as long as it is required. A retention period refers to the period in which business data and records are stored for compliance purposes. When the retention period ends, the data is removed to mitigate risk in keeping out of date information and ensure unnecessary storage space is used, resulting in further costs (Sage, 2020).

Based on the interview data, this sub-theme aimed to understand the types of data stored and not stored. Furthermore, the reasoning behind why this data is stored or not stored. IT professionals were asked to elaborate on what data is currently stored and the reason for storage.

ITM1 indicated that all collected data is stored:

“The only data we store is what we have collected. The more data we store, the better. There's nothing that we are channelling to throw away.”

DE1 mentioned that retention periods are used to store data. If the data must be stored for a long period, then separate storage must be done.

“We try to store as much as possible, for a short period time, about 15 days. Long term storage must be done separately.”

Furthermore, DS1 indicated that each data stream has a retention period. Data storage is where Hadoop technology is utilised in the automotive manufacturer.

“Retention policy is set per data stream. All data is stored; this is where Hadoop plays a role.”

IT participants were further asked to describe the types of data that is not stored. However, due to all data being stored, this question was deemed irrelevant.

Discussion of findings

The key finding is that data collected from IoT devices is stored on big data storage mechanisms for a certain retention period. The retention period policy is applied to each data stream. However, if data needs to be stored for a long period, this is done on a separate storage device. Hadoop technology is a factor in big data storage. Sub-themes 1.5 to 1.7 focused on the usage of big data in automotive manufacturers. The next section discussed the required tools and techniques to analyse data.

6.3.5 Sub-Theme 1.5: Required tools and techniques

Big data tools and techniques comprise analysis tools such as charts, diagrams designed to collect, interpret, and present data for decision-making purposes (Tague, 2005). This sub-theme focused on tools and techniques used to analyse the datasets. In the interview, IT respondents were asked to explain how the data is analysed in-house.

DE1 mentioned that queries could be written to the datasets.

“Can do direct queries with the data.”

ITM1 mentioned that algorithms could be written to applications such as Kibana, Power Bi and customised products.

“Kibana can be used. Easy to use, the business also uses it. Can write complicated algorithms and analytics. Using BI, use Power BI to do analytics of data. We have our customised products as well.”

SA1 further mentioned that free and open-source technologies could be used.

“Free, open-source technologies are used to build components to analyse data.”

Discussion of findings

The purpose of big data analysis tools and techniques is to analyse and visualise data to identify trends, correlations, and patterns within the datasets (Tague, 2005).

DE1 mentioned that queries could be written to the data. A query is a simple data request from one or multiple database tables (Herawan, 2020). As mentioned by ITM and SA1, open-source applications such as Kibana and Power Bi can be used for advanced analysis purposes. Complex algorithms are written into the data on the analytical applications to show all relevant data. An algorithm is a procedure or a formula used to perform a task or solve a problem (Kosek, 2015).

Therefore, the key finding of this sub-theme was that queries are used for simple analysing purposes such as reading data from database tables. However, applications and algorithms can be used for more intense analysing. Organisations need to identify which applications suit the environment.

6.3.6 Sub-Theme 1.6: Creating value from data

In the past, the product development process in the automotive manufacturer consisted of manual processes which produced a limited number of units (Delgado, 2017). Furthermore, data were not widely available as it was recorded manually by operators through handwritten records (Ploner, 2013). However, the significant advancement of technology has resulted in large and complex datasets being generated from IoT technologies. The automotive manufacturer must realise the associated value and importance of efficient data analysis to optimise operational activity (Machado et al., 2019). This sub-theme

highlighted the advancement of technologies and the value associated with big data generation in automotive manufacturers. During the interview, the respondents were asked to elaborate on the current product development process.

BA1 underlined the impact of the fourth industrial revolution on the automotive manufacturer.

“Newer technologies are catering for the fourth industrial revolution. They allow for devices to be connected to and have data streamed or extracted. We perform an analysis on the data.”

Furthermore, ITM1 provided an example of spot welds done manually by operators, prone to human error.

“A lot more equipment. In the past, there were more manual spot welds. With people, there was an error. Now, we are technologically advanced, and it is done by robots ensuring precision.” However, SA1 expressed that there is reluctance in the adoption of newer technologies.

“Changes in technology, however, the problem is the adoption of change as lack of trust.”

Respondents were further asked to explain how value can be created from the data collected in the production development process.

DS1 described that variables would be downsized:

“Cutting costs, time, labour.”

Therefore, technological advancement has resulted in the proficiency of operations. In the past, manual operations resulted in human error, causing production line stops, the reworking of car units and scrapping. Further, the data generated and collected is used to make efficient decision making. However, the adoption of technological change and trust remains a concern. In an analysis of technology adoption, Thiesse and Fliesch (2007) identified that technological compatibility, perceived complexity and business process re-engineering are important technological factors to consider when adopting new

technologies. Top management support, organisational strategy, competitive pressure and environmental impact are further the main drivers in diffusing newer technologies in the production line. Several factors are required for the industry to create value from the data which is generated and collected. IT Professionals were asked to mention the necessary elements.

SA1 stipulate that “*Skills*” would be an important factor.

Furthermore, BA1 described that identifying the correct use cases are important.

“Need to see where the problems are. Get use cases or statistics from stakeholders. Analyse where losing money, and where brand reputation is being damaged.”

The respondents were asked to describe what possible use cases are envisaged in the automotive manufacturer.

DE1 mentioned various use cases from collected data in the automotive manufacturer:

“Unnecessary part ordering or variants of parts. Identify and eliminate the parts and save on cost. Optimise any process. Optimise current processes. Follow processes in sequence. Use the data for planning future lines. Improve next production line.”

The automotive production process is heavily reliant on data to make efficient decisions. Therefore, the industry needs skilled personnel who can extract and provide insight into the raw data. A business strategy enables the organisation to identify objectives, allowing the industry to use data to achieve business goals. This is in the form of a use case (Syafudin et al., 2018).

Based on the interview, the main use case identified was to optimise the current production line process. Lastly, the respondents were asked to describe how data provides value internally and externally for car manufacturers.

GS1 expressed the value behind planning and logistics.

“Internally: Use cases and insights from data. Data-driven insight. Externally: Rate supplier.”

Discussion of findings

Currently, the production development process is technology-driven. The advancement of technology has resulted in data being the driver for the automotive manufacturer to realise value. The key findings where skills are considered is an important factor for the workforce in the automotive manufacturer.

This ensures that data is generated, collected, and analysed with accuracy. Furthermore, internal and external use cases are the drivers behind data-driven decision making.

6.3.7 Sub-Theme 1.7: Organisational culture

Organisational culture refers to shared values and beliefs concerned with how people should work and behave. Furthermore, it incorporates the approach taken for decision making and working activities (McLaughlin & James, 2013).

This sub-theme aimed to understand the requirement from business partners to realise big data value. Furthermore, it explored the impact of organisational structure in translating big data insights into action.



Interview data were analysed where IT respondents were asked to describe the requirements for the automotive manufacturer to capture value, as value has now become co-created by various stakeholders.

BA1 described that the use cases should not be IT-dependent.

“Need to have someone from production driving the use case. Someone from business who can verify and be hands-on with use case. Should not be reliant on a developer.”

Furthermore, SA1 mentioned that there needs to be a collaboration with stakeholders.

“Collaboration and reach consensus on how we get the data and share data with stakeholders.”

Lastly, IT respondents were asked whether they think the organisational structure needs to be changed to understand data translation into action better. **Figure 6.2** summarises the responses from participants of this study. Seven respondents agreed that the organisational structure needs to change. Three respondents disagreed, and one respondent was unsure. ITM1 further mentioned that the workforce must adapt to the changes and understand the value of data in the production process.

“People’s attitude towards data must change. Buy into using the data into using efficiently.”

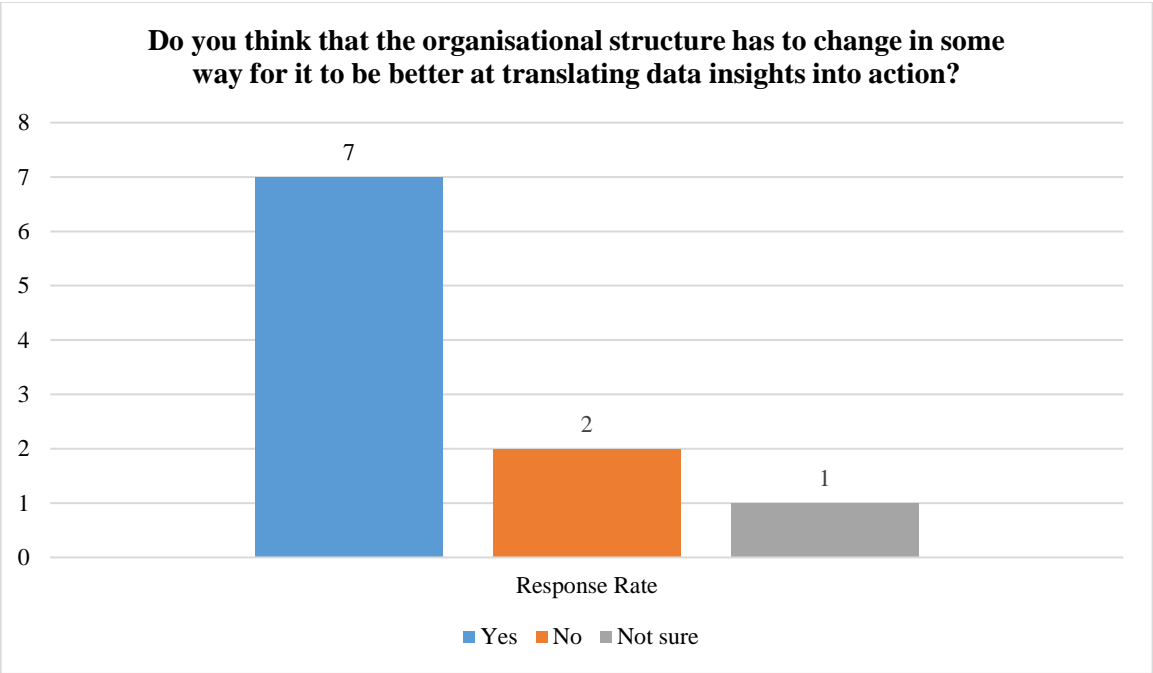


Figure 6.2: Responses to changes in organisation structure

Discussion of findings

This sub-theme identified that cooperation and collaboration with stakeholders are important to realise value. For example, collaboration with suppliers will assist with the supply chain management process. Furthermore, the organisational structure provides a setback, as people are not adaptive to change and realising data value.

6.4 Theme 2: Privacy challenges

Automotive manufacturers invest in Apache Hadoop technology to store and process data (IBM, 2015). However, due to the flexibility of the framework, privacy vulnerabilities have been realised (Bhathal & Singh, 2019). As a result, this theme aimed to gain an IT professional insight on privacy challenges in Hadoop technology in the automotive manufacturer. The aim was to acquire clarity from the automotive manufacturer regarding potential privacy risks associated with Hadoop technology.

The identified sub-themes which explored this notion were privacy awareness, access control to IoT devices, associated organisational risks and the consequences of violating privacy in Apache Hadoop. Additionally, the privacy challenges theme focused on the second research sub-question of this research project: *What types of privacy challenges are experienced within the Hadoop environment at the automotive manufacturer?* The first sub-theme discussed was privacy awareness.

6.4.1 Sub-Theme 2.1: Privacy Awareness

This sub-theme explored privacy awareness amongst IT professionals. During the interview, respondents were asked whether privacy concerned them when using devices connected to the internet. **Figure 6.3** provides a visual representation of the responses.

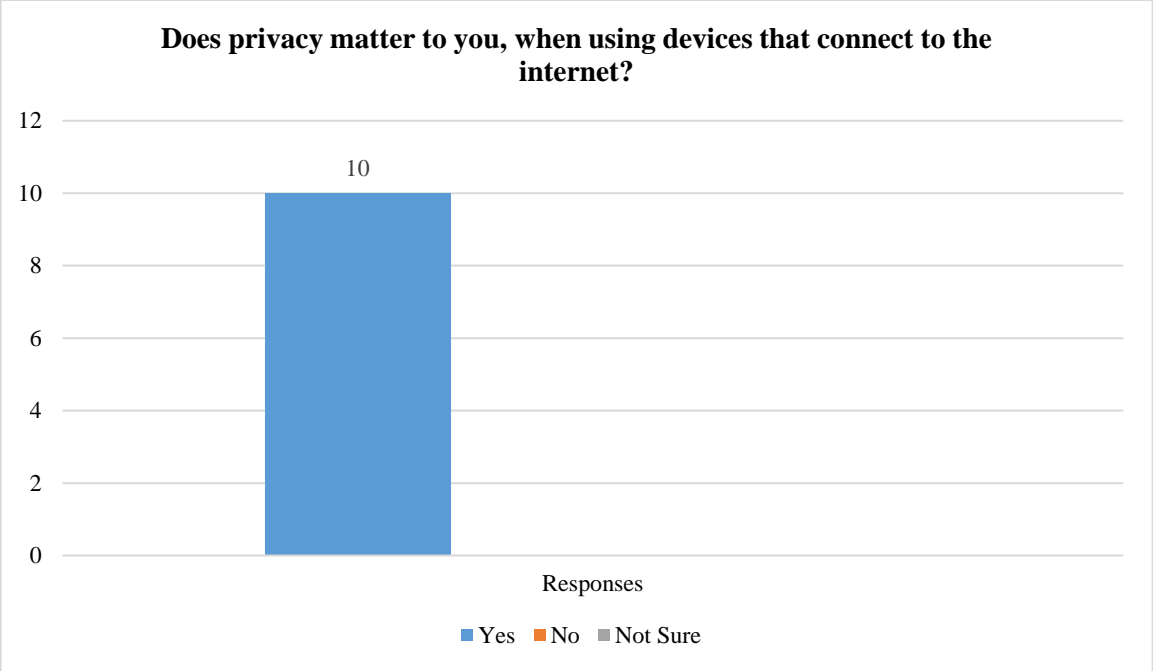


Figure 6.3: Responses on privacy awareness

Discussion of findings

This sub-theme identified that privacy awareness is important when using devices connected over the internet. The next sub-theme focused on access control to IoT devices.

6.4.2 Sub-Theme 2.2: Access control to the Internet of Things devices

Access control is implemented to control access to devices to ensure the confidentiality of business-sensitive information (Andaloussi, Ouadghiri, Maurel, Bonnin, & Chaoui, 2018). The respondents were asked to elaborate on who has access to IoT devices in the automotive manufacturer, which formed part of the interview data obtained.

DE1 indicated that

“Access is restricted.”

BA1 further stated

‘People may get temporary access should it be required.’

Discussion of findings

Based on the findings, it is evident that not all stakeholders have access to IoT devices. Sufficient access control has been implemented. However, implementation of only access control does not mean that the devices are protected from unauthorised users. Therefore, additional measures need to be considered. The next theme explored associated organisational risks.

6.4.3 Sub-Theme 2.3: Associated organisational risks

This sub-theme explored the risks associated with the violation of big data privacy through the interview data obtained. IT respondents were asked to describe what they believe is the risk for the industry due to it being issued severe sanctions for violations of big data privacy in Hadoop technology.

DS1 indicated that



“Hadoop is a data store. The risk is there.”

Furthermore, BA1 mentioned that the automotive manufacturer’s reputation could get damaged.

“If big data is compromised, the brand image can get compromised. Someone externally can use the data for their benefit and end up damaging the company reputation.”

Discussion of findings

Violation of privacy in Hadoop technology can have a serious impact on the industry. Data and technologies need to be secured to prevent any compromises from unauthorised users. The end impact can cause permanent damage to the automotive brand image, which will result in a loss in the competition.

6.4.4 Sub-Theme 2.4: Consequences of violating privacy in Apache Hadoop

If companies violate privacy in Hadoop technology, there are several consequences that associated perpetrators can face. These consequences are based on company policies and rules. This theme aimed to understand the severity of the consequences and whether these consequences are known to the organisation.

Respondents were asked to complete a questionnaire and provide a rating on each statement. The ratings had a key and a description, as indicated in **Table 6.2**.

Table 6.2: Questionnaire key and description

Key	Description
1	Strongly disagree
2	Disagree
3	Agree
4	Strongly Agree

The response from IT professionals was presented in charts. **Figure 6.4** illustrated whether key decision-makers within the automotive manufacturer were aware of the internal negative impact, should there be a breach of big data privacy in Hadoop technology. One respondent strongly disagreed, three respondents disagreed, two respondents agreed, three strongly agreed, and one response is unknown. Therefore, it was identified that the latter of the responses lies within the strongly agree or agree option.

This further indicated that 40% disagree with the statement. Therefore, it was important for the automotive manufacturer to ensure that privacy and security policies are clearly defined and that employees abide by the rulings of the company.

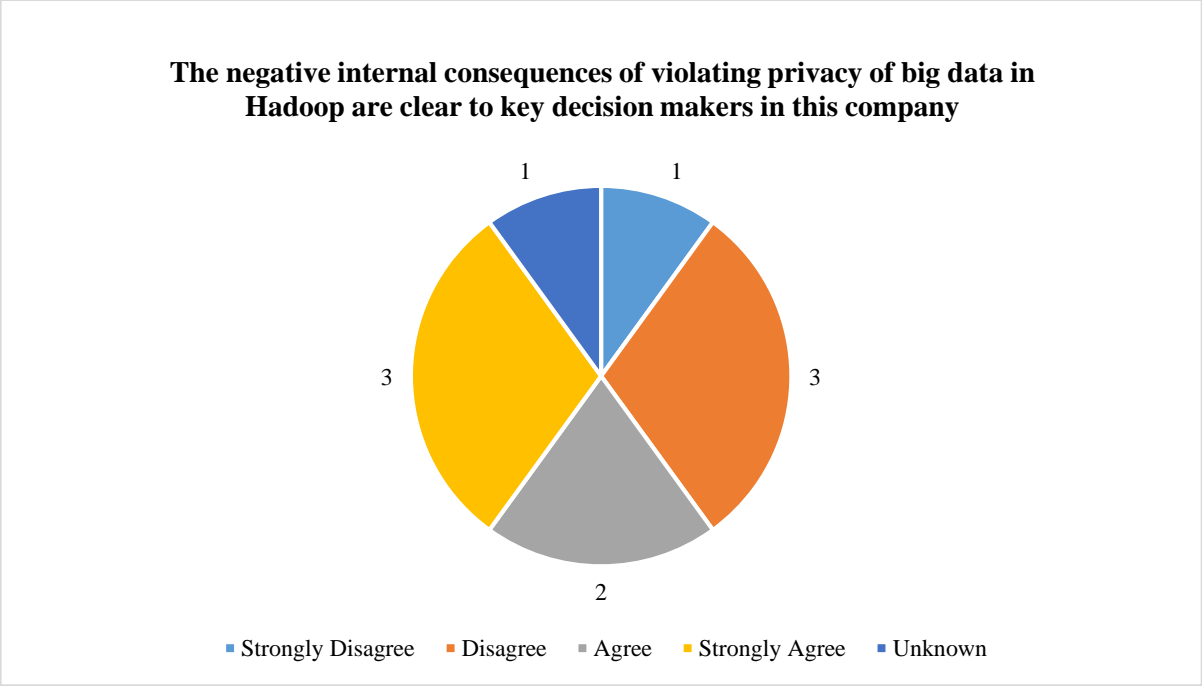


Figure 6.4: Responses to the negative internal consequences of violating privacy of big data in Hadoop

Figure 6.5 identified responses based on consequences organisations should if they face violate the privacy of big data in Hadoop technology. Three respondents strongly agreed, five agreed, two strongly disagreed, and zero disagreed that serious consequences would be faced. Based on most of the responses, there are consequences stated in privacy and security policies should the automotive manufacturer violate big data privacy.

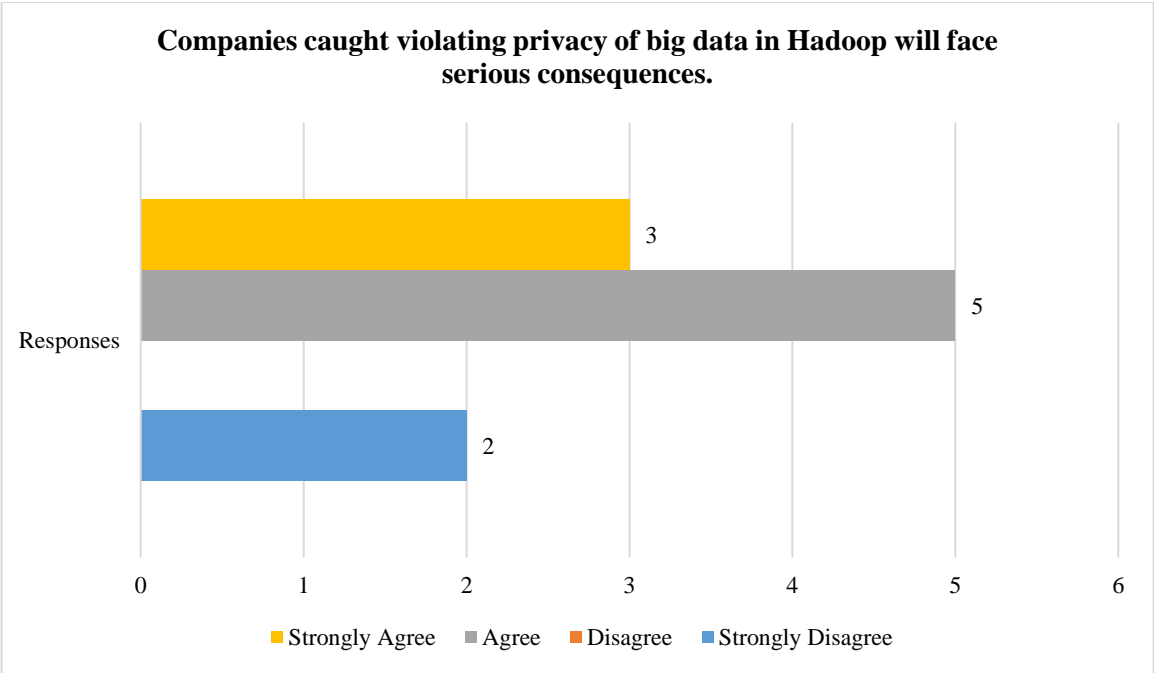


Figure 6.5 : Responses on violation of big data in Hadoop will face serious consequences

Discussion of findings

This theme aimed to understand whether the internal consequences of violating privacy in Hadoop are clear to decision-makers in the automotive manufacturer. Furthermore, the objective of this theme was to understand whether the automotive manufacturer would face the consequences should they get caught violating the privacy of big data in Hadoop technology.

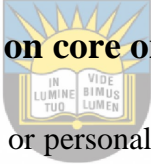
Based on the findings, it was understood that the decision-makers are aware of the negative impact on the automotive manufacturer. Furthermore, if faced with a violation, then consequences will be faced. This means that the automotive manufacturer does have IT privacy and security policies that the organisation must abide by. The next theme discussed privacy measures to address privacy concerns of big data in Hadoop technology in the automotive manufacturer.

6.5 Theme 3: Privacy protection

As discussed in Chapter two and section 6.3.4 of this chapter, the automotive manufacturer utilises Apache Hadoop technology for big data processing and storage (IBM, 2015). However, measures for privacy protections need to be implemented to protect corporate information (Jain, Gyanchandani, & Khare, 2016).

This theme focused on the last research sub-question of the research project, which is stated as: *What are the types of privacy challenges experienced within the Hadoop environment at the automotive manufacturer?* The first sub-theme focused on privacy measures and the effect on core organisational values. Additionally, the last sub-theme in this section identified and examined methods used to protect the automotive manufacturer's sensitive information from being accessed by unauthorised users. The first sub-theme discussed in the next section was the impact on core organisational values.

6.5.1 Sub-Theme 3.1: Impact on core organisational values



Organisational values refer to social or personal beliefs which employees share within an organisation. Ultimately, organisational values guide important aspects that impact decisions and actions within the company (Towerstone Leadership Centre, 2016). This sub-theme explored whether organisational values are sacrificed when big data privacy in Hadoop technology is followed.

Based on the interview data, **Figure 6.6** indicated one respondent strongly agreed, 0 agreed, four disagreed, and five strongly disagreed regarding the automotive manufacturer's values being sacrificed due to following the privacy of big data in Hadoop technology. This meant that by following privacy rulings, the core business operations would not be negatively impacted.

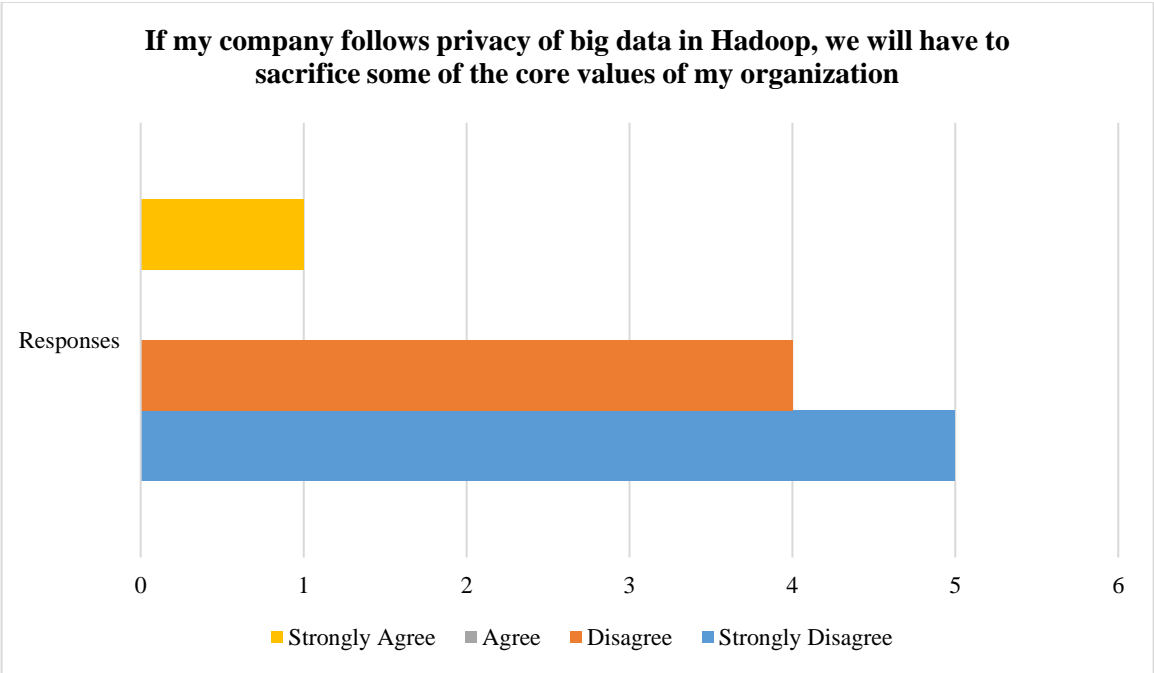


Figure 6.6: Responses on privacy of big data and impact on organisational core values

Discussion of findings

This sub-theme explored the impact on organisational core values based on whether the automotive manufacturer follows big data privacy in Hadoop. Based on the findings, there is no negative impact on the core values of the automotive manufacturer.

6.5.2 Sub-Theme 3.2: Methods to protect privacy

Encryption involves encoding data on devices so that it is hidden from unauthorised users. The importance of encryption is to protect confidential information, sensitive data and furthermore enhance security between the application and server (Matthews, 2019).

In the questionnaire, IT respondents were asked whether IoT devices were encrypted. The objective was to understand whether encryption is currently being used on devices in the automotive manufacturer. As shown in **Figure 6.7**, 80% of respondents agreed that IoT devices are encrypted. The next section discussed password technology.

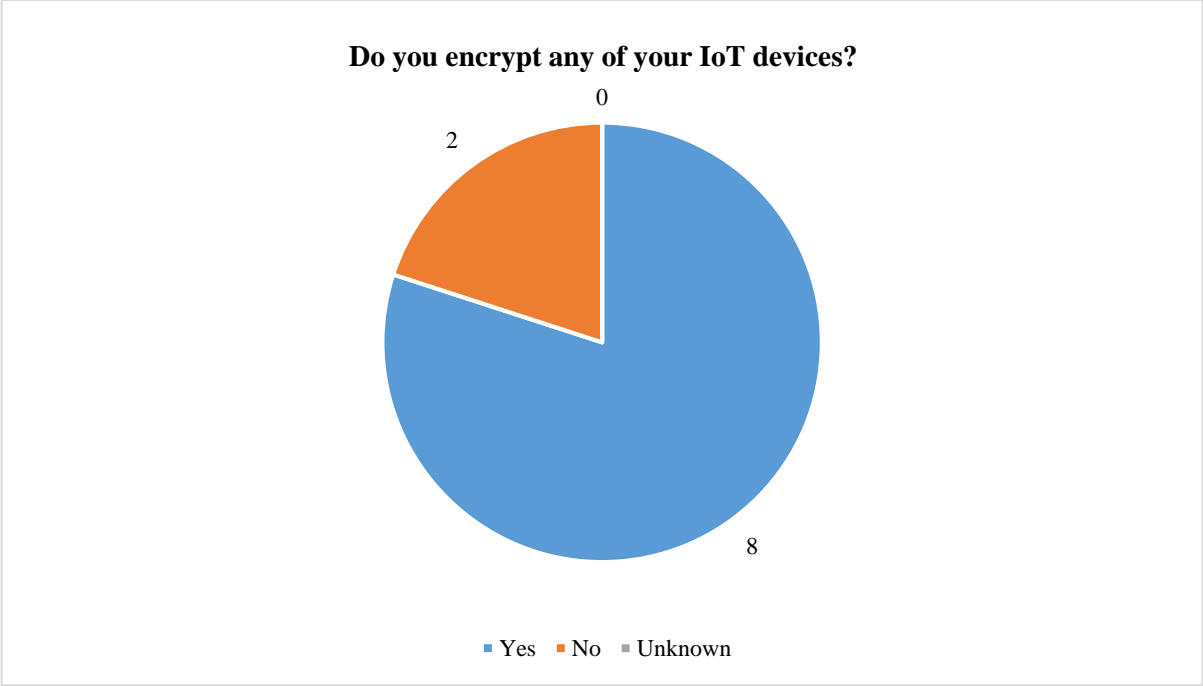


Figure 6.7: Response to device encryption

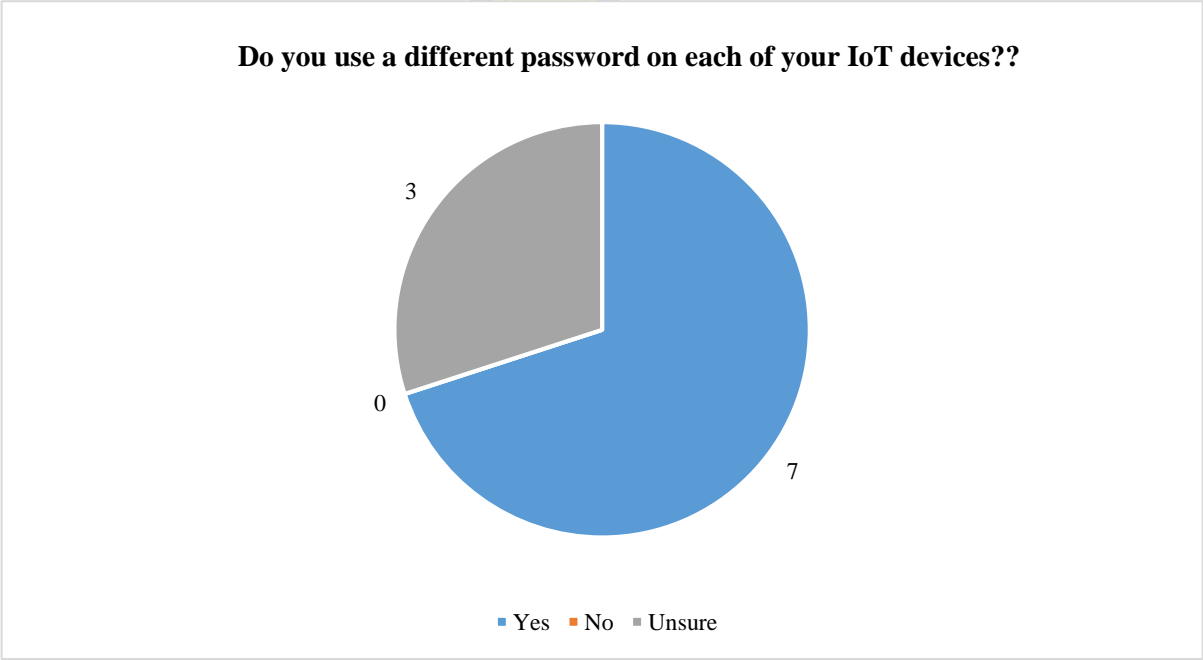


Figure 6.8: Response on using different passwords on IoT devices

The objective of passwords is to enable the first line of defence from unauthorised access to any device and information (Martinelli, 2018). IT respondents were asked if they use a different password on each of their devices. 70% of the respondents agreed to use different passwords on their IoT devices. The responses are illustrated in **Figure 6.8**.

A software vulnerability is a security flaw found in a program or operating system. Therefore, the purpose of security updates is to assist in patching security flaws and protecting data (Microsoft, 2019). Based on **Figure 6.9**, seven out of the ten respondents installed security updates on IoT devices.

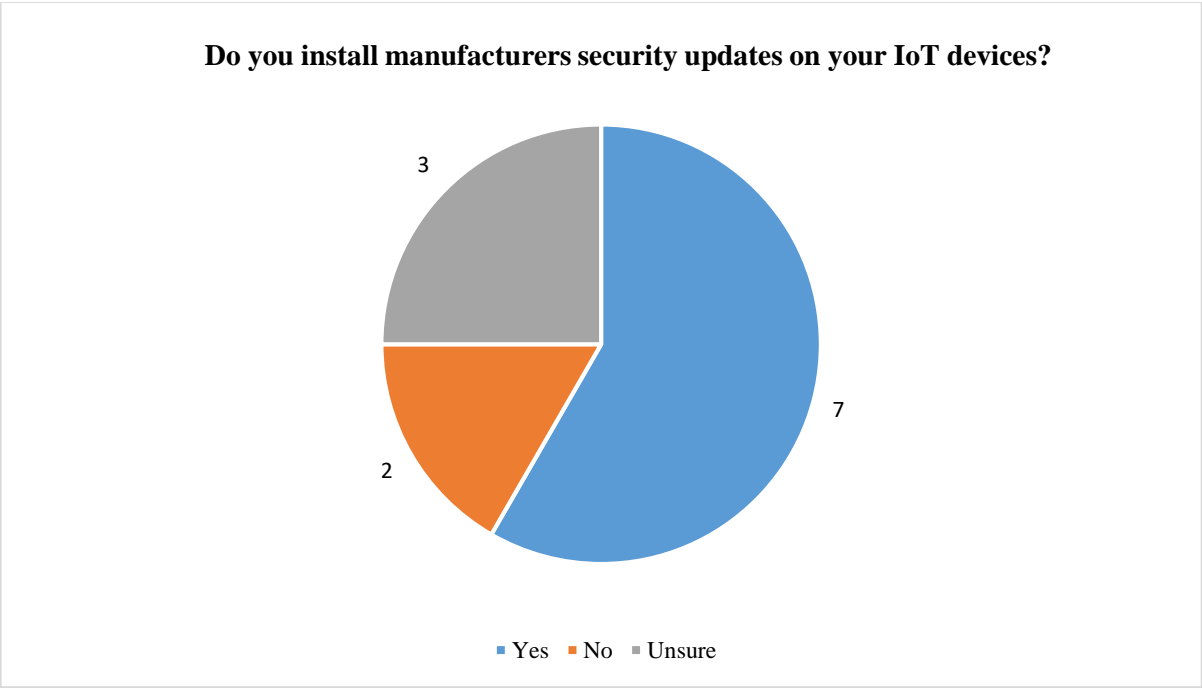


Figure 6.9: Responses on installing manufacturers security updates

Information Systems Security Officers (ISSO) are responsible for establishing and implementing information security and privacy policies and standards to protect information and prevent unauthorised access (Chron, 2020). Based on **Figure 6.10**, 50% of the responses agreed that the privacy and security standards are unambiguous in terms of privacy in Hadoop technology.

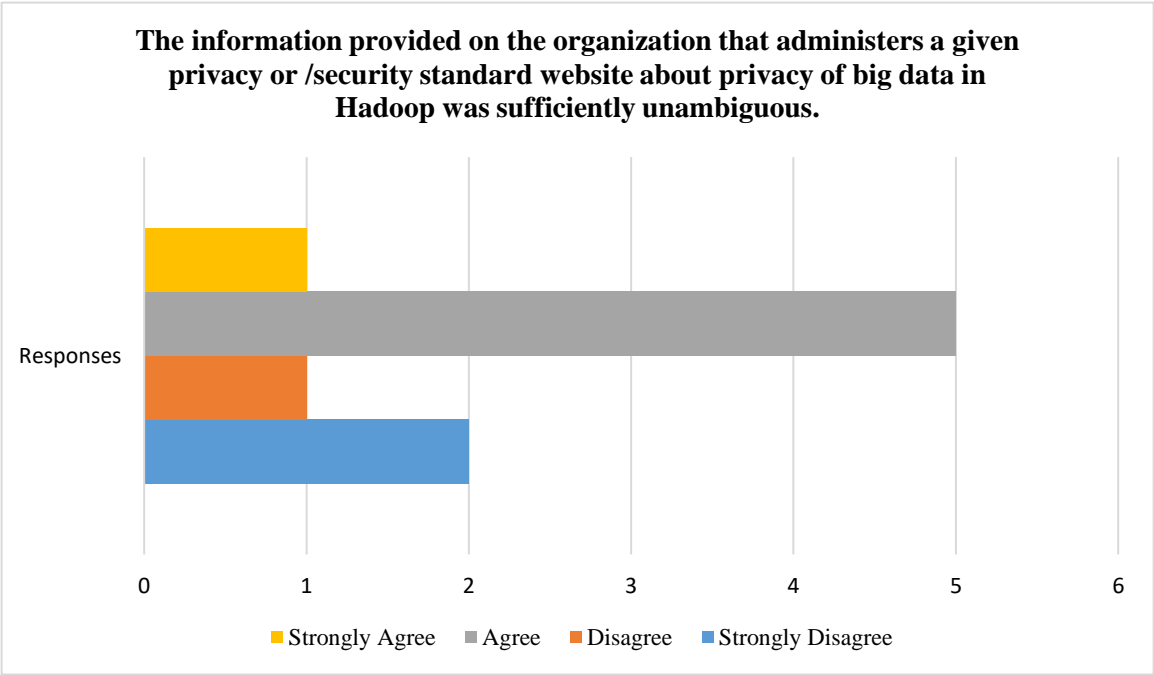


Figure 6.10: Responses on unambiguity on privacy and security standards

Discussion of findings

This theme focused on ensuring the privacy of big data in Hadoop technology in the automotive manufacturer. It was identified that most automotive manufacturers' IoT devices have encryption and use different passwords. Furthermore, security updates would be rolled out regularly to these devices to ensure that any security flaws are patched. Lastly, ISSO specialists provide sufficient and unambiguous information when implementing security and privacy policies in the organisation.

6.6 Conclusion

Effective use of big data in the automotive manufacturer has benefited the organisation, such as supply chain optimisation and improved production. The automotive manufacturer uses Hadoop technology due to its effectiveness to control and to process the big data generated from IoT devices connected to the production line.

However, Hadoop technology has been established with security vulnerabilities, which has raised privacy concerns in the automotive manufacturer. Unauthorised users and third parties can take advantage of the security vulnerabilities by accessing or tampering with company sensitive information. As a result, business use cases can become available to the public and competitors. Furthermore, the automotive manufacturer can establish a poor reputation and brand due to low production levels and bad quality vehicles. Therefore, this research study aimed to establish a framework, which addressed the privacy concerns of Hadoop technology in the automotive manufacturer.

The interpretivist paradigm was applied to this research study. The primary data collection method incorporated conducting an interview and questionnaire session with ten IT specialists at a locally based automotive manufacturer. The interview and questionnaire were split into three themes relevant to this study's main research and sub-questions. The identified themes, sub-themes and findings were summarised in the next section.

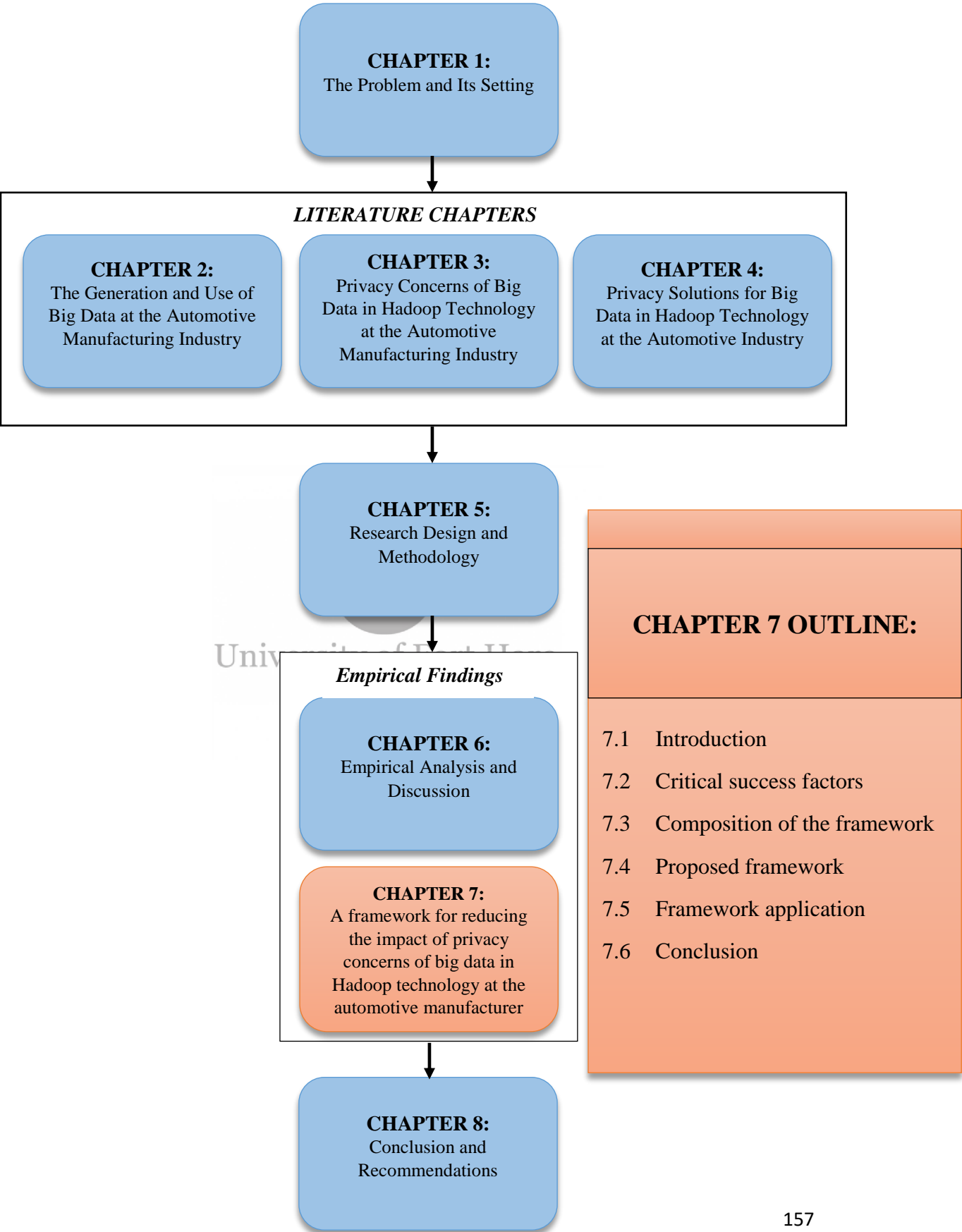
- *Theme 1: Generation and use of big data*
 - 1.1 Sampled demographics
 - This sub-theme identified the distinct roles for each respondent who worked with big data in the automotive manufacturer. This enabled the researcher to understand how their role impacted the phenomenon. Distinct roles of the ten respondents included: system administrator, business analyst, graduate students, data scientist, data engineer, IT manager, software developers and solutions architect. Therefore, the automotive manufacturer required a central IT team that collaborates findings to realise big data opportunities.
 - 1.2 IoT and effect on the automotive manufacturer
 - IoT and the fourth industrial revolution has resulted in the digital transformation of the production process in the automotive manufacturer. It was found that IoT devices had a positive effect and brought about the value to the organisation.
 - 1.3. Data generation and collection

- A communication channel needs to exist between the IoT device and the storage infrastructure such as Hadoop. The analysis found that MQTT and MSB technologies are used as communication channels. Once the data is collected from the IoT devices, big data analytics is applied to the dataset to realise the value and opportunities of big data.
 - 1.4 Big data storage
 - Data is collected from IoT devices and stored on storage infrastructures for a retention period per data stream. If data is stored for a long period, then the data is stored on a separate device.
 - 1.5 Required tools and techniques
 - Big data analysis tools and techniques are used to analyse and visualise data patterns and trends. Open-source applications such as Kibano and Power Bi can be used for analysis. Further, queries and algorithms can be written to the datasets for analysis purposes.
 - 1.6 Creating value from data
 - The production process is currently technology-driven, resulting in data being the main factor in realising its associated opportunities. Skills are an important requirement for the workforce to ensure that data is generated, collected, and analysed accurately.
 - 1.7 Organisational culture
 - Co-operation and collaboration are important to realise value. It was further identified that the organisational structure is not adaptive to the technological change in the fourth industrial revolution.
- *Theme 2: Privacy challenges*
 - 2.1 Privacy awareness
 - Privacy awareness is an important factor when using devices that are connected to the internet.
 - 2.2 Access control to IoT devices
 - Access control has been implemented as not all stakeholders have access to the IoT devices. However, additional measures also

needed to be considered to protect big data privacy in Hadoop technology.

- 2.3 Associated organisational risks
 - Privacy violation in Hadoop can result in critical data being compromised. This can negatively impact the brand image and result in a loss in competition with other automotive manufacturers.
- 2.4 Consequences of violating privacy in Apache Hadoop
 - The automotive manufacturer had IT privacy and security policies implemented if an employee violated a rule.
- *Theme 3: Privacy protection*
 - 3.1 Impact on core organisational values
 - Following privacy techniques and methods did not negatively impact any of the automotive manufacturer's core values.
 - 3.2 Methods to protect privacy
 - IoT devices use encryption and different passwords. Further, security updates are rolled out regularly to ensure the security flaws are patched. ISSO specialists provide sufficient information when implementing relevant policies to ensure privacy and security at the organisation.

The literature and data analysis chapter were used to construct critical success factors, which were extended into a framework in the next chapter.



Chapter 7

A Framework for Reducing the Impact of Privacy Concerns of Big Data in Hadoop Technology at the Automotive Manufacturer

7.1 Introduction

The fourth industrial revolution has resulted in the digital transformation of automotive production systems (Machado, Winrotha, & Ribeiro da Silva, 2019). Big data generated from the Internet of Things (IoT) on the production line has equipped the automotive manufacturer to improve supply chain management, the production process, and the automobile's quality (Deloitte, 2019; Russo, Confente, & Borghesi, 2015).

The automotive manufacturer uses Hadoop technology to manage challenges for storing, managing and analysing the complex and large amounts of data generated from IoT devices connected to the production line. Furthermore, Hadoop technology is cost-efficient to store and process the big data generated (Educba, 2020 & Tole, 2013). However, Hadoop technology was built with a poor security model. This means the automotive manufacturer's sensitive information can easily become compromised by third parties and unauthorised users (Tawalbeh, Muheidat, Tawalbeh, & Quwaider, 2020; O'Donovan, Leahy, Bruton, & O'Sullivan, 2015). Ajah and Nweke (2019) discussed if the automotive manufacturer's sensitive information is compromised, the results of big data analytics will be erroneous, the use cases of big data will not be realised, and competitors can exploit the automotive manufacturer, causing a competitive loss.

This study incorporated two theories: The selective Organisational Information Privacy and Security Violations Model (SOIPSVM) and the Capability Maturity Model (CMM).

The SOIPSVM theory includes conditions that reduce the likelihood of a privacy violation from occurring. The CMM aims to improve systematic and organised processes through a five-level maturity model. The literature analysis discussed how big data could be generated and used from internal and external data sources. The effective usage of big data had resulted in privacy issues, which was identified. These findings were then addressed by several solutions that can assist solve the privacy issues associated with big data in the automotive manufacturer. However, these solutions failed to address privacy based on how organisational structures and processes use conditions to understand and address the raised concerns.

Critical success factors were first constructed, and the identified privacy solutions were incorporated based on the literature and data analysis solutions. These factors were then extended into a framework that provided a refined solution towards addressing privacy concerns using the conditions associated with the SOIPSVM model and the five-stage process of the CMM.



7.2 Critical success factors

University of Fort Hare

Critical success factors were constructed from the literature in chapters two, three, four and the empirical findings from Chapter six. Each critical success factor and its purpose was explained in **Table 7.1**.

Table 7.1: Critical success factors to address privacy concerns of big data in Hadoop technology at the automotive manufacturing industry

<u>Critical Success Factors</u>	<u>Purpose</u>	<u>Reference: Literature Review</u>	<u>Reference: Empirical Data</u>
Control of external and internal sources	Effectively store, manage, and analyse data generated from internal and external sources. External sources focus on the consumer environment and include sensors and social media.	Chapter 2: Section 2.3	Chapter Six: Section 6.3.3 and 6.4.2

	<p>Internal sources in the automotive manufacturer's production environment include smart meters and business applications.</p> <p>IoT devices have a direct relationship with an automotive manufacturer. Therefore, a communication channel needs to be set up between the IoT devices and the storage infrastructure. The empirical findings identified that a Message Queue Telemetry Transport (MQTT) and Manufacturing Service Bus (MSB) is used as the communication channel from the IoT to the storage technology. The data collected can be in numerous forms, such as image, text, and robotic data. Once collected, big data analytics must be applied to the datasets for the automotive manufacturer to realise the associated value.</p>		
Monitor the value of big data towards improving the automotive manufacturing process	<p>Perform big data analytics (descriptive, diagnostic, predictive & prescriptive) to realise future benefits of big data in the automotive manufacturer. Benefits include optimised supply chain management and an improved manufacturing process.</p> <p>The empirical analysis discussed</p>	Chapter 2: Section 2.6.1 & 2.6.2	Chapter Six: Section 6.3.6

	<p>how technology and data drive the automotive production process.</p> <p>Furthermore, data enables the automotive manufacturer to realise the value of data. Skills is an important requirement for the workforce to ensure that big data is generated, collected, analysed, and managed effectively.</p>		
Implementation of user authentication	Authentication allows for users to be verified when accessing the technology. Hadoop needs to be integrated with an existing authentication technique to ensure user authentication.	Chapter Four: Section 4.2.1	
Execute authorisation and Access Control Lists (ACLS)	Hadoop technology uses file-based permissions and ACLS to manage the access of users. Hadoop can be built with access control implemented through user or group-based permissions, which may be insufficient. Furthermore, an effective user control policy must contain automated role-based settings and policies.	Chapter Four: Section 4.2.2	
Implement encryption to secure data	Data can be encrypted with the use of a personal key. The personal key makes the data unreadable for unauthorised individuals. However, the same software used to encrypt the data must read and analyse the data.	Chapter Four: Section 4.2.3	Chapter Six: Section 6.5.2

	Encryption aims to protect sensitive data and optimise security between the application and server. 80% of respondents agreed to encrypt IoT devices in the automotive manufacturer.		
Conduct audits	Audits are performed in Hadoop to meet compliance requirements. This is done periodically. Actions performed in Hadoop are logged, enabling administrators to review historical logs and actions performed in Hadoop.	Chapter Four: Section 4.2.4	
Conduct regular reviews of user access to data	Conduct meetings with corporate stakeholders who access the data repositories and review data access permissions for authorised personnel. Access permissions must be adjusted based on changes to employee responsibilities.	Chapter Four: Section 4.4.1	
Apply data masking to sensitive data	Edit sensitive data so that it is not shared externally with the organisation concerned. This means masking will preserve the type and length of the structured data, replacing it with worthless value.	Chapter Four: Section 4.4.2	
Implement disaster recovery and backup plan	Ensure continuation of business operations should the data centres fail. The disaster recovery plan consists of backup, replicated or mirrored systems. The disaster recovery and backup plans must be updated with lessons learned.	Chapter Four: Section 4.4.3	

Monitor user behaviour	This entails continuous monitoring of the access routine of users and developing an outlook on the behavioural aspect of the user accessing data. If an anomaly is detected, this means there could be an observation or event which does not correlate to the expected pattern and immediately needs to be alerted for potentially fraudulent activities.	Chapter Four: Section 4.4.4	
Apply tokenisation to secure data	Tokenisation entails replacing sensitive data with irregular values, which keeps all critical aspects of the data intact. Clear text is replaced with a random value that is the same data type and length. Tokens have no worth if breached and secure the original value of the data outside the original environment.	Chapter Four: Section 4.4.5	
Build own infrastructure to store and analyse data	Depending on the size and financial means of the automotive manufacturer, it can take the initiative to build its data repositories to store and analyse the data.	Chapter Four: Section 4.4.6	
Install regular security updates	Security updates assist in patching security flaws and vulnerabilities found in a program or an operating system. Regular security updates also assist with protecting data. 70% of respondents install security		Chapter Six: Section 6.5.2

	updates on IoT devices in the automotive manufacturer.		
Update passwords regularly	Updating passwords regularly enables the first line of defence from unauthorised access to any device and information. It was identified that 70% of IT respondents use different passwords on each of their IoT devices.		Chapter Six: Section 6.5.2

7.3 Composition of the framework

After constructing the critical success factors, these factors were further extended into a framework. Swanson (2013) identified that a framework is derived from theories explaining, predicting, and understanding a phenomenon. The framework aims to support a theory of a research study and understand an objective, meaning that it can be adjusted accordingly (Swanson, 2013).

The framework was composed of critical success factors that were identified in **Table 7.1**. These factors were categorised into the two conditions that form the components of the SOIPSVM. The two components of conditions are contextual conditions and rules and regulatory conditions. These conditions were then aligned and expanded in correlation to the five stages of the CMM. The composition of the framework is illustrated in **Figure 7.1**.

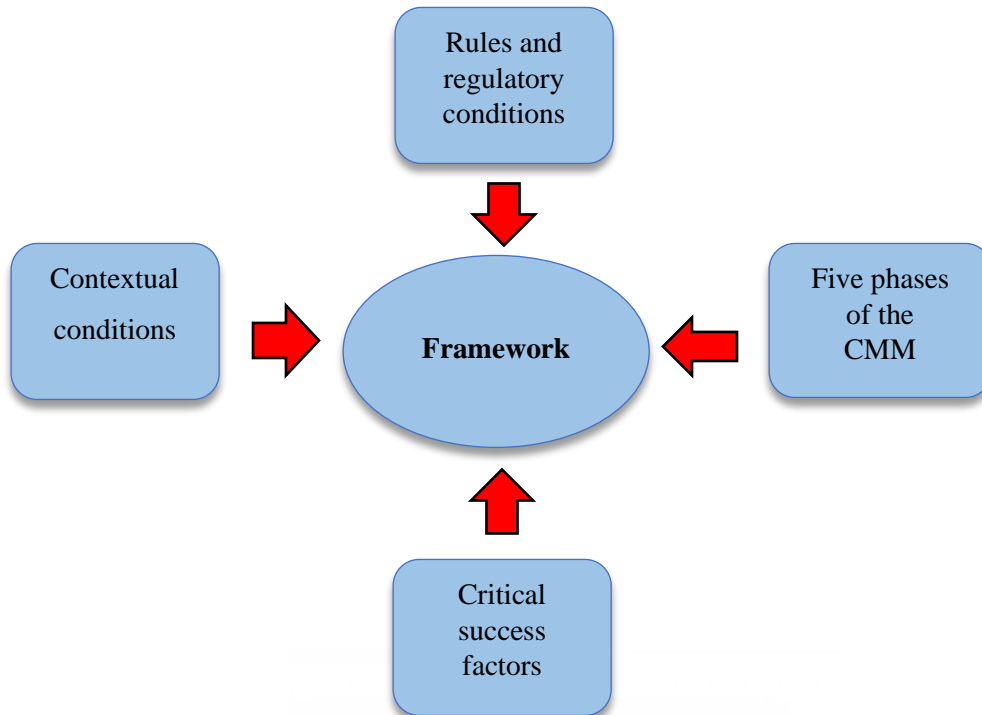


Figure 7.1: Framework composition

7.3.1 Contextual conditions

Contextual conditions contribute to an organisation's functioning (Wall, Lowry, & Barlow, 2015). The contextual conditions identified in the literature and data analysis include controlling external and internal sources, monitoring the value of big data in improving the manufacturing process, and implementing disaster recovery and a backup plan, as illustrated in **Figure 7.2**. Big data is generated from internal and external sources. Therefore, these sources need to be controlled effectively to produce valuable data.

Once this is done, big data's value towards improving the automotive manufacturing process can be monitored to provide accurate results. Furthermore, implementing disaster recovery and a backup plan ensures that production can continue as usual should the automotive manufacturer's data centres fail due to discrepancies.

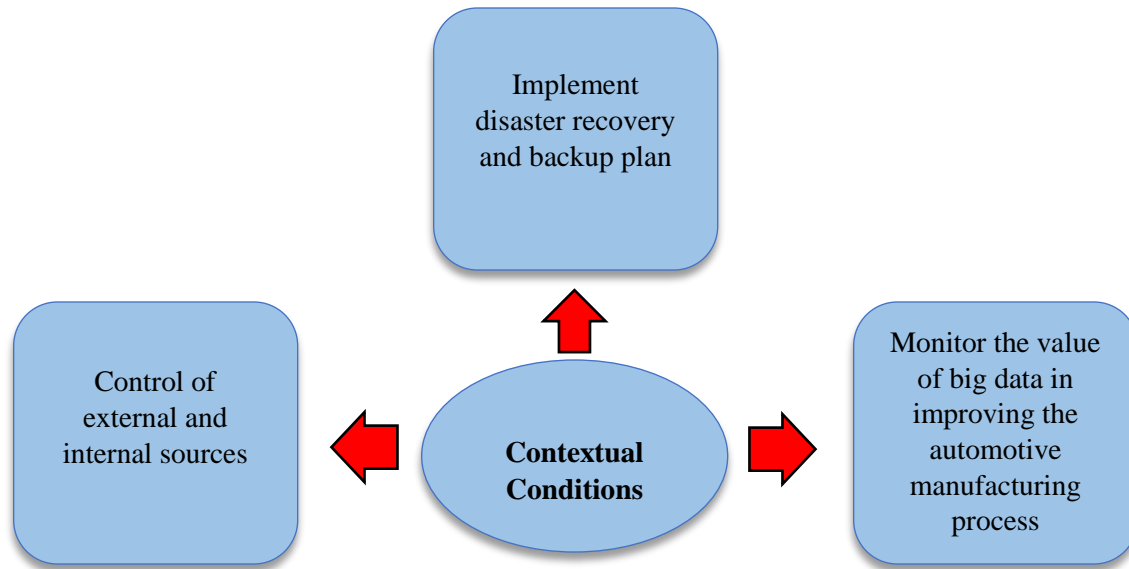


Figure 7.2: Factors of contextual conditions

7.3.2 Rule and regulatory conditions

Rules and regulatory conditions are viewed as being constraints that are external and formal, like laws. These conditions must be low in ambiguity (Wall et al., 2015). Rule and regulatory conditions were identified as the automotive manufacturer's measures to protect their big data privacy in Hadoop technology. This included implementation of user authentication, encryption to secure data, execute authorisation and ACLS, conduct audits, regular reviews of user access to data, apply data masking to sensitive data, tokenization to secure data, monitoring user behaviour, installation of regular security updates, update passwords regularly and build own infrastructure to store and analyse data. These conditions are illustrated in **Figure 7.3**.

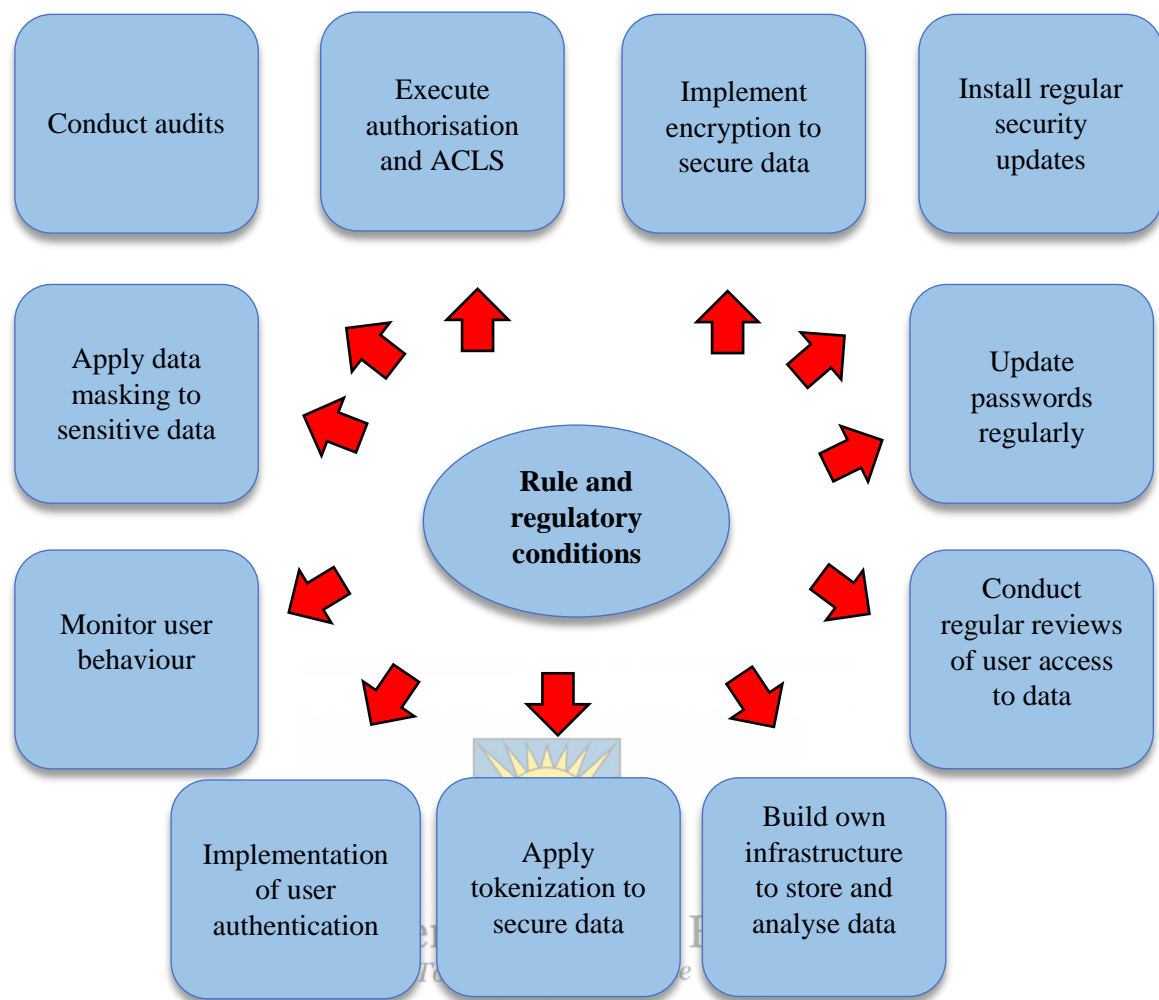


Figure 7.3: Factors of rule and regulatory conditions

7.4 Proposed framework

The framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer aims to solve the prevalent privacy issues of big data in Hadoop technology. Various factors were identified in terms of contextual, rule, and regulatory conditions, as shown in **Figure 7.2** and **Figure 7.3**. These conditions were established from the SOIPSVM model. Once these conditions were identified, they were compared at various levels of the CMM. The CMM was used in this study to depict continuous improvement and systematic processes of each critical success factor (CMMI Institute, 2017).

The purpose of comparing conditions at different levels was to determine the transition of the conditions at different levels and for the automotive manufacturer to identify how they can improve factors identified. Therefore, the maturity levels would gradually progress until the condition identified had been optimised. The framework is illustrated in **Table 7.2**.

Table 7.2: A framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer

	<u>Initial</u>	<u>Repeatable</u>	<u>Defined</u>	<u>Managed</u>	<u>Optimising</u>
<u>Contextual Conditions</u>					
Control of external and internal sources	No control over the external and internal sources.	A policy is drafted for the control of external and internal sources.	The policy is implemented.	Review the policy to ensure adherence to the policy.	Review best practices and update the policy regularly.
Implement disaster recovery and backup plan	No disaster recovery and backup plan.	Disaster recovery and data backup plans are drafted.	Disaster recovery and a backup plan are implemented.	Review the disaster recovery and backup plan.	Conduct regular reviews and update the disaster recovery and backup plan.
Monitor the value of big data towards improving the automotive manufacturing process	Data is generated in voluminous amounts.	Data analytics is used to extract value from large datasets.	Predictive analysis is conducted.	Evaluate predictions to ensure accuracy in the manufacturing process.	Monitor and control the automotive manufacturing process to meet or exceed predictions
<u>Rule and Regulatory Conditions</u>					
Implementation of user authentication	No authentication or poor authentication has been implemented for client or server applications.	Choose and examine appropriate authentication methods.	User authentication is effectively implemented.	Review the effectiveness of user authentication methods and ensure required standards are met.	Monitor and control user authentication.

Execute authorisation and ACLS	Access control is lacking, or it has not been implemented.	Suitable access control methods are identified and chosen.	Access control is implemented.	Determine whether access control is effective to ensure privacy and security standards are met.	Monitor and update access control.
Implement encryption to secure data	Data is not protected during the transmission to and from the Hadoop system.	The encryption method is identified and chosen.	Encryption is implemented.	Evaluate whether data clusters are secure at rest and in motion.	Monitor and control encryption.
Conduct audits	Unauthorised or inappropriate access by authorised users.	An audit is permitted on the Hadoop system.	Conduct an audit on the Hadoop system periodically.	Assess the Hadoop system to ensure compliance with security requirements.	Review audits conducted.
Conduct regular reviews of user access to data	No control on who can access data repositories.	Prepare meetings with stakeholders who access data repositories on a semi-annual or annual basis.	Implement control access and conduct user access reviews.	Review data access permissions for authorised personnel.	Manage and update access control to data repositories.
Apply data masking to sensitive data	No preservation of sensitive data sold to third parties.	Data masking is chosen as a preservation technique.	Data masking is implemented.	Determine whether data masking has preserved the sensitive data from being shared with external organisations.	Monitor the masked sensitive data.
Monitor user behaviour	Abnormality in user behaviour is detected.	The access routine of the user is monitored.	Control access is implemented.	Evaluate whether access control is effective to prohibit	Monitor and control user behaviour through effective access controls.

				abnormal behaviour.	
Apply tokenization to secure data	Data is insecure at rest, at use, in transit and analytics.	Tokenization is chosen as being effective for securing data at various levels.	Implement tokenization to text.	Review the flexibility in levels of data security privileges.	Monitor and control tokenization.
Install regular security updates	Security updates are not installed regularly.	Compile a plan to identify required security updates and rollout period.	Implement the plan to ensure that all security flaws are patched with the update.	Review the rollout period of the security updates.	Monitor and regularly review the plan to ensure effectiveness.
Update passwords regularly	Passwords are not updated regularly.	Compile a password policy to ensure that passwords are updated regularly.	Implement the password policy.	Review the password policy to ensure adherence to the policy.	Review best practices and update the password policy regularly.
Build own infrastructure to store and analyse data	Numerous flaws inhibited Hadoop technology.	Develop complex hardware infrastructures to store and analyse data.	Implement new infrastructure and data repositories.	Review the effectiveness of building your infrastructure.	Constantly maintain infrastructure data repositories.

7.5 Framework application

The method of application of the framework would be an independent assessment for each condition identified in **Table 7.2**. For each condition being considered for the application into Hadoop technology, the framework would guide the process by using different maturity levels. To perform the assessment, the following steps would need to be conducted:

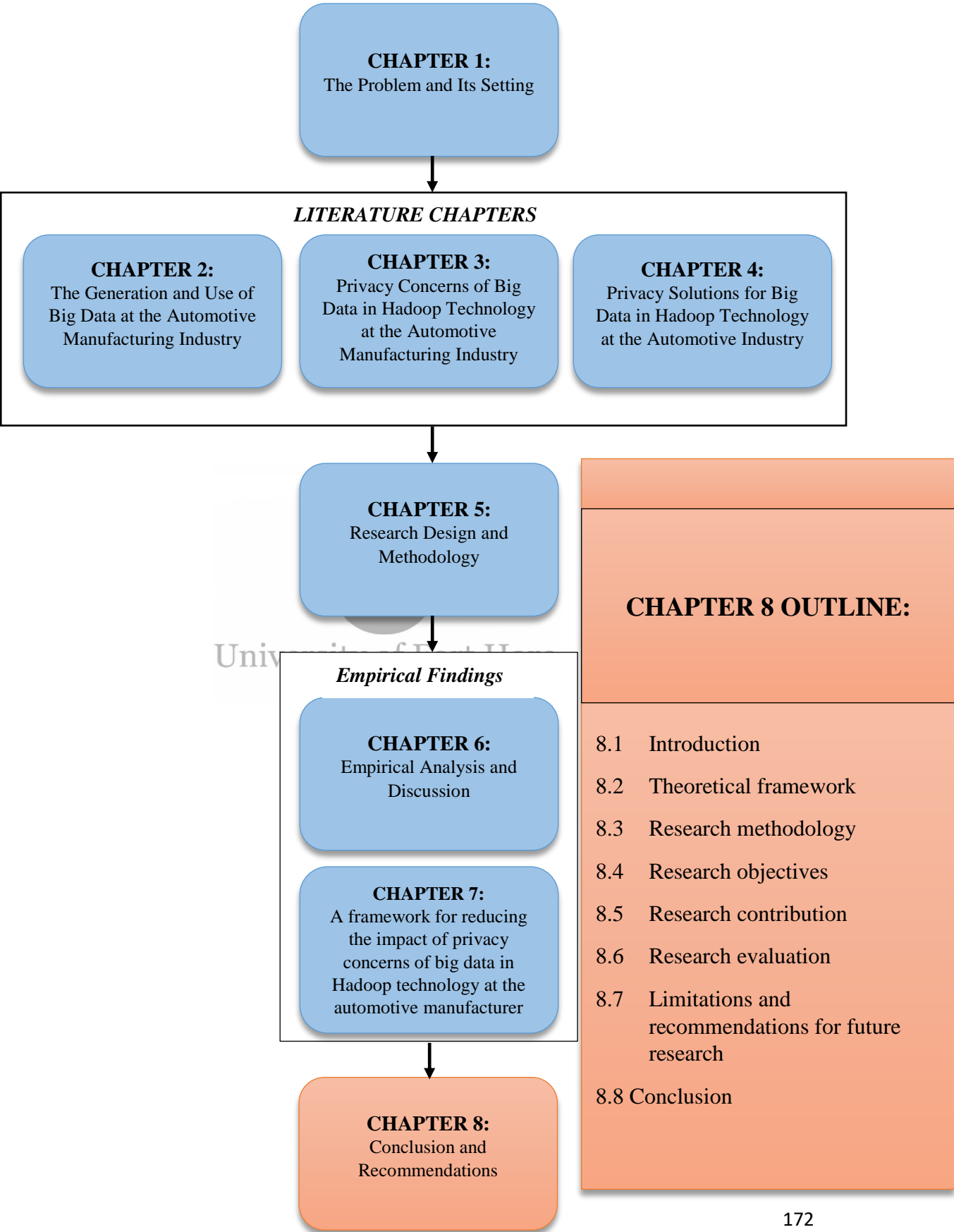
- *Identify relevant sub-conditions:* Identify the relevant critical success factors (sub-conditions) and whether it is categorised as contextual or rule and regulatory condition.
- *Assess the condition:* Evaluate the current scenario of the sub-conditions identified.

- *Establish a maturity level for each condition:* Identify the maturity level on the continuum after evaluating the current sub-conditions identified.
- *Implement steps to improve maturity where levels are lower:* If the identified maturity level is not optimised, implement the required steps per the continuum to achieve an optimised maturity level.
- *Re-assess on a period basis:* Re-assess the sub-conditions periodically to identify if the optimised level has been achieved or whether the maturity has decreased on the continuum. If the maturity level has not been achieved, repeat the steps until the optimised maturity level has been attained.

7.6 Conclusion

Critical success was first constructed and then extended into a framework. The framework titled as a framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer, was formulated using components from existing models. These components were identified in the literature analysis. Contextual and rule and regulatory conditions were adapted from SOIPSVM, and the five phases of maturity were adapted from the CMM. A combination of these components resulted in the framework in **Table 7.2**.

This framework answered the research question: *How can the privacy concerns of big data in Hadoop technology at the automotive manufacturer be addressed?* Privacy solutions were identified and categorised into contextual and rule and regulatory conditions of the SOIPSVM. These conditions were then extended and explained at various stages of the maturity model. The prime motive was to reach the last level, which is the optimised level. The factors were utilised to illustrate the gradual progress of how the critical success factors underwent a transition through the various stages of the CMM and how big data privacy in Hadoop technology can be addressed at multiple levels. The final chapter concluded this research study, and recommendations were made on this research study.



Chapter 8

Conclusion and Recommendations

8.1 Introduction

The preceding chapters have contributed towards identifying a solution to the research problem and objective of the study. Chapter one provided the introduction, background and discussion of the research problem and objectives. Chapter two focused on the generation and use of big data in automotive manufacturers. Chapter three identified and discussed the barriers of Hadoop technology and included the privacy challenges. Chapter four outlined and examined security and data privacy protection measures implemented in the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer.



Chapter five provided the research design and methodology adopted in this study. The empirical analysis and discussion were examined in Chapter six. Furthermore, Chapter seven discussed the measures to address the research problem through critical success factors, extended into a framework.

Chapter eight examined the theoretical background and contribution of the study. This was followed by a summary of the research study and a discussion on each research question. Limitations and future research were outlined. This chapter concluded with an evaluation of the study based on the overall findings.

8.2 Theoretical framework

The underlying theory used in this study was the Selective Organisational Information Privacy and Security Violations Model (SOIPSVM). This framework was chosen to provide a comprehensive foundation towards understanding and addressing the privacy concerns of big data in Hadoop technology in the automotive manufacturer.

Hadoop technology contains a poor security model. Therefore, information is easily available to unauthorised individuals. If an ineffective security model exists, then the organisation's data privacy can be compromised by information attacks (Jain, Gyanchandani, & Khare, 2016; Wall, Lowry, & Barlow, 2015).

Therefore, information privacy is of utmost importance, especially in the current digital age. As a result, information privacy and security laws have become mandatory across various industries (Wall et al., 2015). This theory's contextual and rule and regulatory conditions were used to categorise the identified critical success factors, contributing to this research study's framework.

In addition, the Capability Maturity Model (CMM) model was used to construct the framework of this study. The importance of the CMM model was that it signified five maturity levels of processes, where each level allowed for continuous improvement to reach a higher performance of operations (CMMI Institute, 2017). Each sub-condition under the contextual and rule and regulatory were compared against the five levels of the CMM. The SOIP SVM and CMM were incorporated into the research study framework to address privacy issues of big data in Hadoop technology in the automotive manufacturer. The following section discussed the research methodology of this study.

8.3 Research methodology

This study was conducted using the interpretivist paradigm approach. Further, the qualitative data collection method was used. This study followed the design science methodology, which focused on creating an artefact to solve the research problem. This study adopted the seven guidelines of design science research.

- *Design as an artefact:* The final output of this study was a framework. The framework's purpose was to address privacy concerns of big data in Hadoop technology in the automotive manufacturer.
- *Problem relevance:* For this study, the problem investigated was the privacy concerns of big data in Hadoop technology in the automotive manufacturer. A

solution was proposed in the form of critical success factors, which was extended into a framework.

- *Design evaluation*: The framework was to be evaluated through an independent assessment (Section 7.4).
- *Research contributions*: The contribution of this study was the framework, which ensured big data privacy in Hadoop technology in the automotive manufacturer.
- *Research rigour*: In terms of rigour, this research project incorporated the qualitative method for data collection. The CMM assessed the factors of the framework at different levels of maturity.
- *Design as a search process*: This guideline was achieved by collecting primary data and secondary data.
- *Communication of research*: This guideline was met through thesis and journal articles written on this research study.

This research study incorporated an interview and questionnaire for primary data collection. Secondary data included the literature review, critically analysed journals, books, websites, and conference proceedings. The literature review was examined to form the base of the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer. The literature review and the empirical findings were used to define the critical success factors of this study. Critical success factors were aligned with the SOIP SVM and the five levels of the CMM, to produce the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer.

8.4 Research objectives

The primary objective of this research study was to construct a holistic framework to assist automotive manufacturers in overcoming the privacy issues associated with big data in Hadoop technology. This framework was presented and described in Chapter 7

Section 1.3 of Chapter One defined the main research question, and three sub-questions were derived. The purpose of the research questions was to direct the researcher in this

study. The research questions were assessed to conclude and evaluate the study. The primary research question was defined as follows:

How can the privacy concerns of big data in Hadoop technology at an automotive manufacturer be addressed?

The construction of the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer, enabled the Information Technology (IT) department to become better informed of the solutions, which can be implemented to ensure data privacy in Hadoop technology. This would enable the automotive manufacturer to realise the benefits associated with big data, achieve competitive advantage and business growth. To achieve the primary objective, three sub-research questions were derived from the main research question.

i) *How is big data being generated and used at the automotive manufacturer?*

The literature content was discussed in Chapter two and was analysed to answer this sub-question. This chapter discussed the advancement of the production line due to Industry 4.0. Internal and external data sources were defined, and examples were provided. Internal data sources focused on the business environment and included business applications and Radio-Frequency Identification (RFID). In contrast, external sources included the consumer and supplier environment. Examples of internal sources included sensors and social media.

Due to the complexity of big data, various characteristics were identified and defined. Characteristics of big data included volume, velocity, variety, value, and veracity. This chapter highlighted the importance of applying big data analytics to realise big data's benefits and associated use cases in automotive manufacturers. The types of data analytics which can be applied include descriptive diagnostic (knowing what occurred), predictive (what may happen), diagnostic (root cause) and prescriptive analytics (future occurrences). If effective big data analytics is applied, the automotive manufacturer can experience an improvement in the manufacturing process and optimise supply chain management.

Chapter six, Theme One, focused on answering this sub-question from an empirical perspective. It was found that the advancement of technology had resulted in data being a strong driver for the automotive manufacturer to realise the opportunities associated with big data. Further, Message Queue Telemetry Transport (MQTT) and the Manufacturing Service Bus (MSB) technologies are used as a communication channel between the Internet of Things (IoT) and the infrastructure such as Hadoop.

As a result, data collected from the IoT devices can only be stored onto the storage device for a specific retention period. It was further identified that tools and techniques are required to apply effective big data analytics to the data. Organisational culture impacted the automotive manufacturer in realising the opportunities that big data could achieve for the company. Therefore, co-operation and collaboration with stakeholders was important factor to consider.

- ii) *What are the types of privacy challenges experienced within the Hadoop the environment at the automotive manufacturer?*

Chapter three addressed the types of privacy challenges prevalent in Hadoop technology. This chapter provided an overview of Hadoop in the context of the automotive manufacturer. It was found that Hadoop technology was used in the automotive manufacturer to control storage, mining, and the analysis of complex data. Further, Hadoop technology was a cost-efficient infrastructure that could be used. It was identified there are drawbacks to the technology. The disadvantage of Hadoop is that it has an ineffective security model present, resulting in data privacy concerns occurring. Therefore, a comparison between privacy and security was conducted to get an understanding of the two concepts. It was identified that data breaches and fragmented data are the data privacy issues associated with Hadoop in the automotive manufacturer. The identified privacy issues were then discussed in the context of the SOIPSVM.

Furthermore, Chapter Six, Theme Two, focused on privacy challenges from the empirical findings. It was identified that privacy awareness is crucial when using devices connected to the internet. Based on the results, it was found that not all stakeholders have access to IoT devices connected to the production line.

Further, sufficient control had been implemented, but that did not mean that the devices were protected from information attacks. Therefore, additional measures had to be considered. If a violation of privacy occurred in Hadoop technology, it could seriously impact the industry. For example, the automotive manufacturer could experience a loss in the competition. It was also understood that IT privacy and security policies had been implemented in the industry.

iii) What measures can the automotive manufacturer take to protect their privacy?

The literature of this sub-question was discussed in Chapter four of this study. Chapter four identified security measures that can be implemented to address the insecure security model in Hadoop technology. Security measures included implementing user authorisation and Access Control List (ACLS), encryption to secure data, and to conduct audits. It further explored data privacy protection techniques that can be applied in the form of governance methods. These methods included conducting regular reviews of user access to data, using data masking to sensitive data, implementing disaster recovery and backup plans, monitoring user behaviour, applying tokenization to secure data, and building of own infrastructure to store and analyse data.

University of Fort Hare
Together in Excellence

Chapter Six, Theme Three, focused on privacy protection in the organisation. It explored the methods to protect the privacy of big data in Hadoop in the automotive manufacturer. It was found that the automotive manufacturer does have encryption methods in place, and different passwords were used to access the devices. Furthermore, security updates were regularly rolled out to the devices to ensure that all security flaws were patched. Additionally, it explained that no industry's core functions would be exploited should effective and efficient measures are implemented to address the privacy issue prevalent in Hadoop technology. **Table 8.1** provides an overview of the findings of this study.

Table 8.1: Overview of findings

Main Research Question: How can the privacy concerns of big data in Hadoop technology at an automotive manufacturer be addressed?				
Research Objective	Research Question	Literature review, questionnaire, or interview	Results	Conclusion
To identify the sources of big data in the automotive production process and associated use cases.	How is big data being generated and used at the automotive manufacturer?	<ul style="list-style-type: none"> • Literature review - Chapter Two. • Interview and Questionnaire - Chapter Six: Section 6.4. 	Internal (business environment) and external (consumer or supplier environment) sources exist. Use cases included optimisation of supply chain management and improvement in the manufacturing process.	Effective use of big data enables a well-connected production line, competitive advantage, and business growth.
To uncover the source of privacy challenges that impact the automotive manufacturer from realising big the benefits and opportunities of big data.	What are the types of privacy challenges experienced within the Hadoop environment at the automotive manufacturer?	<ul style="list-style-type: none"> • Literature review - Chapter Three. • Questionnaire and Interview - Chapter Six: Section 6.4. 	Hadoop technology has a poor security model established, therefore impacting data privacy. Identified privacy challenges included data breaches and data fragmentation.	To obtain a clear understanding of the impact and link between the poor security model and data privacy in Hadoop.
To identify the measures which the automotive manufacturer can	What measures can the automotive manufacturer take to	<ul style="list-style-type: none"> • Literature review – Chapter Four 	Measures that can be implemented include user authentication, authorisation/ACLS, encryption, audits, user access reviews,	To establish measures that can be used to address the research problem. The

implement to address data privacy issues in Hadoop	protect their privacy?	<ul style="list-style-type: none"> Questionnaire – Chapter Six: Section 6.5. 	data masking, user behaviour monitoring, tokenization, development of own infrastructure, regular security updates, and disaster recovery and backup plans.	identified measures formed the base for critical success factors, which was extended into a framework.
<p style="text-align: center;">Final Product:</p> <p style="text-align: center;">A framework to assist the automotive manufacturing industry in addressing big data privacy concerns in Hadoop technology.</p>				

8.5 Research contribution

Current literature content of big data privacy concerns in Hadoop technology at the automotive manufacturer is limited (Bhathal & Singh, 2019). Therefore, this research study is a valuable contribution to the Information Systems (IS) research domain. This study proposed critical success factors to address big data privacy concerns in Hadoop technology at a local automotive manufacturer. Furthermore, the critical success factors were extended into a framework, which assessed the factors at different maturity levels.

The critical success factors and framework was proposed and discussed in Chapter Seven. This was the primary contribution of this research project. The CMM model was an important factor in assessing the critical success factors as it signified five levels of maturity. Each level enabled continuous improvement to reach the optimised level. The optimised level signified the highest performance of operations (CMMI Institute, 2017). The critical success factors were constructed considering the measures identified in Chapter Four and Chapter Six (Section 6.6) of this study.

The automotive manufacturer and academia can use this research study to understand the privacy challenges faced when using Hadoop technology. The contribution of this study will assist with identifying measures which can be implemented iteratively to minimize

information privacy vulnerabilities and further enable the industry to understand the value that big data has on the automotive manufacturer.

8.6 Research evaluation

The interpretivist research approach differs from the positivist approach. Oates (2006) identified the criteria for interpretivist research and positivist research. These criteria are mainly used as a research evaluation to ensure the credibility of the research project. These are shown in **Table 5.5**.

The interpretivist paradigm was used in this study. The criteria for this research study against the paradigm was described below:

- *Trustworthiness*: The researcher was provided with an ethical clearance form, a declaration of originality had been completed, the plagiarism report (Appendix D) has been attached to this study, which met the requirements as stated by the Information Systems (IS) department of the University of Fort Hare. The respondents to the primary data collection had completed an overview and informed consent form before the primary data collection process.
- *Confirmability*: This research study adopted the principle of saturation, where the data gathered did not provide any new information about the research problem. As a result, the study had gathered adequate information.
- *Dependability*: The contribution of this study incorporated two models, namely: SOIPVSM and CMM. Furthermore, the researcher engaged in notetaking and recording during the interview process to accurately summarise the findings from the interview.
- *Credibility*: Ten big data IT specialists were part of the primary data collection in this study. The specialists were able to provide an accurate representation of their daily job, which was related to the context of this research study. The literature was critically discussed using recent studies.
- *Transferability* – This research study applied a methodology to construct a suitable framework to address the research problem. Transferability was considered

appropriate where the framework could be adopted outside of the automotive production process and within the business sectors of the organisation.

This study adopted the interpretivism criteria, making it a credible research study.

8.7 Limitations and recommendations for future research

This study addressed privacy concerns of big data in Hadoop technology at the automotive manufacturer and only focused on the manufacturing process. A recommendation was to empirically test the framework for reducing the impact of privacy concerns of big data in Hadoop technology at the automotive manufacturer.

This study was restricted in context, as it was conducted in South Africa at an automotive manufacturing plant. More research could be undertaken in other industries and also automotive manufacturers outside of South Africa. Future research could be investigated regarding the data privacy of cars connected to the internet. Furthermore, more critical research focused on personnel will address organisational culture challenges when adopting advanced technologies such as big data and Hadoop in the automotive manufacturer. Lastly, the design science research could be applied rigourously to ensure and the artefacts works as it was intended to, and that

8.8 Conclusion

There are numerous benefits the automotive manufacturer can experience with the effective use of IoT devices and big data analytics. This includes supply chain optimization, improved production process, business growth and competitive advantage. Hadoop technology is used to store, manage, and process big data in the automotive manufacturer's production environment. This infrastructure is used in the automotive manufacturer due to its fault-tolerance, cost efficiency and scalability.

However, Hadoop technology has numerous security vulnerabilities, making it susceptible to information privacy threats from third parties and unauthorised users. In effect, the business will lose competition, produce low-quality vehicles, and establish reputational

damage. This research project presented a study addressing big data privacy concerns in Hadoop technology at the automotive manufacturer.

The outcome of this study was the framework developed to assist the organisation with identifying solutions to ensure data privacy in Hadoop technology. The significance of this study was that the automotive manufacturer would realise the associated benefits of incorporating big data and Hadoop into the organisation. Adopting the technologies and the identified solutions will improve the automotive manufacturer's operations, business growth, and competitive advantage.



University of Fort Hare
Together in Excellence

9. References

- Abdullah, N., Håkansson, A., & Moradian, E. (2017). Blockchain based Approach to Enhance Big Data Authentication in Distributed Environment . *Ninth International Conference on Ubiquitous and Future Networks (ICUFN)* (pp. 887 - 892). Milan: Institute of Electrical and Electronics Engineers.
- Abualkishik, A. Z. (2019). Hadoop and Big Data Challenges. *Journal of Theoretical and Applied Information Technology*, 97(12), 3489-3500.
- Adebesin, F., Gelderblom, H., & Kotzé, P. (2011). Design research as a framework to evaluate the usability and accessibility of the Digital Doorway. *Proceedings of the 2011 Design, Development and Research Conference*. Cape Town.
- Agrawal, V. (2016). *The Impact of Big Data and Analytics on Manufacturing [Infographic]*. Retrieved from Tech Co: <http://tech.co/big-data-analytics-manufacturing-2016-12>
- Ahmadu, B., Hussin, A. C., & Bahari, M. (2021). Development and Validation of a Classified Information Assurance Scale for Institutions of Higher Learning. In F. Saeed, F. Mohammed, & A. Al-Nahari (Eds.), *Innovative Systems for Intelligent Health Informatics* (pp. 857-868). Cham: Springer.
- Ajah, I. A., & Nweke, H. F. (2019). Big Data and Business Analytics: Trends, Platforms, Success Factors and Application. *Big data and Cognitive Computing*, 3(32), 1 - 30.
- Andaloussi, Y., Ouadghiri, M. E., Maurel, Y., Bonnin, J. M., & Chaoui, H. (2018). Access control in IoT environments: Feasible scenario. *The 8th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS-2018)* (pp. 1031–1036). Porto: Elsevier B.V.
- Andersson, A., & Axelsson, M. (2018). *Creating and Appropriating Value from Connected Vehicle Data*. Gothenburg, Sweden: Chalmers University of Technology.
- Apache Hadoop. (2020). *HDFS Permissions Guide*. Retrieved from Apache Hadoop: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsPermissionsGuide.html>
- Automotive World. (2021). *Cyber-security in manufacturing demands Defense in Depth*. Retrieved from Automotive World: <https://www.automotiveworld.com/articles/cyber-security-in-manufacturing-demands-defense-in-depth/>

- Awasthi, A. (2020). Disaster Recovery - Foundation Pillars. *International Journal of Science and Research (IJSR)*, 9(2), 1360-1362.
- Batchelor, K. (2019). *What Is Tokenization, and Why Is It So Important?* Retrieved from Forbes: <https://www.forbes.com/sites/forbestechcouncil/2019/08/13/what-is-tokenization-and-why-is-it-so-important/>
- Bazaz, T., & Khalique, A. (2016). A Review on Single Sign on Enabling Technologies and Protocols. *International Journal of Computer Applications*, 151(11), 18-25.
- Beamlar Additive Manufacturing. (2018). *Business cases: 3D printing in the automotive industry*. Retrieved from Beamlar Additive Manufacturing: <https://www.beamlar.com/3d-printing-in-the-automotive-industry/>
- Benjelloun, F.-Z., & Lahcen, A. A. (2015). Big Data Security: Challenges, Recommendations and Solutions. In K. Munir, M. Al-Mutairi, & L. Mohammed, *Handbook of Research on Security Considerations in Cloud Computing* (pp. 301-313). Pennsylvania: IGI Global.
- Bhagyashree, A., & Koundinya, A. K. (2020). Convergent Analytical Tools for Big Data Applications in Hadoop Environment. *International Journal of Engineering Applied Sciences and Technology (IJEAST)*, 4(9), 283-285.
- Bhathal, G. S., & Singh, A. (2019). Big data: Hadoop framework vulnerabilities, security issues and attacks. *Array (1-2)*, 1-8.
- Bowman, C. P. (2020). *What is a vehicle identification number (VIN)?* Retrieved from Coverage : <https://www.coverage.com/insurance/auto/what-is-a-vin/>
- Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Burley, K. (2018). *What is a bureaucratic organisation?* Retrieved from Chron: <http://smallbusiness.chron.com/bureaucratic-organization-20379.html>
- Bysko, S., Krystek, J., & Bysko, S. (2020). Automotive Paint Shop 4.0. *Computers and Industrial Engineering*, 139(1), 105546 - 105558.
- Cândido, J., Aniche, M., & van Deursen, A. (2021). *Log-based software monitoring: a systematic mapping study*. PeerJ Computer Science.
- Cavanillas, M. J., Curry, E., & Wahlster, W. (2016). *New Horizons for a Data-Driven Economy*. Switzerland: Springer International Publishing.
- Center for Innovation in Research and Teaching. (2019). *Overview of Descriptive Research*. Retrieved from Center for Innovation in Research and Teaching: https://cirt.gcu.edu/research/developmentresources/research_ready/descriptive/overview

- Chaudhury, S., & Banerjee, A. (2010). Statistics without tears: Populations and samples. *Industrial Psychiatry Journal*, 19(1), 60–65.
- Cherradi, G., Boulmakoul, A., & El Bouziri, A. (2016). Smart Data Collection Based IoT Protocols. *Big data & Applications 12th edition of the Conference on Advances of Decisional Systems*. Marrakech: ASD.
- Chhetri, S. R., Rashid, N., Faezi, S., & Al Faruque, M. A. (2018). Security Trends and Advances in Manufacturing Systems in the Era of Industry 4.0. *Journal of Hardware and Systems Security*, 2(1), 51-68.
- Chowdhury, M. F. (2014). Interpretivism in Aiding Our Understanding of the Contemporary Social World. *Open Journal of Philosophy*, 4(1), 432-438.
- Chron. (2020, May 18). *Responsibilities of an Information System Security Officer*. Retrieved from Chron: <https://work.chron.com/responsibilities-information-system-security-officer-15533.html>
- Chu, K. (2020). *Apache Hadoop: A Review on Security Issues and Solutions for HDFS*. Retrieved from Towards data science: <https://towardsdatascience.com/apache-hadoop-a-review-on-security-issues-and-solutions-for-hdfs-5ba06861b7cd?gi=84ec814e2a1b>
- Cision. (2020). *The global automotive motors market size is projected to grow from USD 20,321 million in 2020 to USD 25,719 million by 2025, at a CAGR of 4.8%*. Retrieved from Cision PR Newswire: <https://www.prnewswire.com/news-releases/the-global-automotive-motors-market-size-is-projected-to-grow-from-usd-20-321-million-in-2020-to-usd-25-719-million-by-2025--at-a-cagr-of-4-8-301113089.html>
- Close, D. (2019). *Tokenization: Your Secret Weapon for Data Security?* Retrieved from ISACA: <https://www.isaca.org/resources/news-and-trends/industry-news/2019/tokenization-your-secret-weapon-for-data-security>
- Clough, P., & Nutbrown, C. (2012). *A Student's Guide to Methodology* (3rd ed.). London: Sage Publications.
- CMMI Institute. (2017). *What Is Capability Maturity Model Integration (CMMI)?* Retrieved from CMMI Institute: <http://cmmiinstitute.com/capability-maturity-model-integration>
- Collis, J., & Hussey, R. (2009). *Business research: A practical guide for undergraduate and postgraduate students*. London: Palgrave Macmillan.
- Colombo, P., & Ferrari, E. (2019). Access control technologies for Big Data management systems: literature review and future trend. *Cybersecurity*, 2(3), 1-13. doi:10.1186/s42400-018-0020-9

- Colombo, P., & Ferrari, E. (2019). Access Control Technologies for Big Data Management Systems: Literature Review and Future Trends. *Cybersecurity*, 2(3), 1-13.
- Creswall, J. (2014). *Research Design: Qualitative, Quantitative and mixed method approaches* (4th ed.). California: Sage Publications, Inc.
- Crossman , A. (2016). *Understanding Secondary Data Analysis*. Retrieved from ThoughtCo: <https://www.thoughtco.com/secondary-data-analysis-3026536>
- CSCMP. (2020). *CSCMP Supply Chain Management Definitions and Glossary*. Retrieved February 15, 2020, from Council of Supply Chain Management Professionals (CSCMP): https://cscmp.org/CSCMP/Educate/SCM_Definitions_and_Glossary_of_Terms.aspx
- Data Privacy Manager. (2020). *Data Breach and Reputation Management*. Retrieved from Data Privacy Manager: <https://dataprivacymanager.net/data-breach-and-reputation-management/>
- Davenport, T., Patil, A., & Snidauf, D. (2018). *A revolution in data-driven quality improvement*. Retrieved from Deloitte: <https://www2.deloitte.com/us/en/insights/topics/analytics/a-revolution-in-data-driven-quality-improvement.html>
- De Vos, A., Strydom, H., Fouche, C., & Delport, C. (2005). *Research at Grass Roots for the Social Sciences and Human Service Professions* (Third ed.). Pretoria: Van Schaik.
- Delgado, R. (2017). *Big Data's Transformation of the Manufacturing Industry*. Retrieved from Data Informed: <http://data-informed.com/big-datas-transformation-of-the-manufacturing-industry/>
- Deloitte. (2015). *Big Data and Analytics in the Automotive Industry - Automotive Analytics Thought Piece*. Retrieved from <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/manufacturing/deloitte-uk-automotive-analytics.pdf>
- Deloitte. (2017). *Five key principles to secure the enterprise Big Data platform*. Retrieved from <https://www2.deloitte.com/content/dam/Deloitte/de/Documents/technology/TECH-Hadoop-Best-Practices-Security-Big-data-platform-large-enterprises-2017.pdf>
- Deloitte. (2017). *Using autonomous robots to drive supply chain innovation*. Retrieved from <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/us-manufacturing-autonomous-robots-supply-chain-innovation.pdf>

- Deloitte. (2019). *Implementing advanced analytics - Disruption in the automotive industry*. Retrieved from Deloitte: <https://www2.deloitte.com/uk/en/insights/industry/automotive/auto-analytics.html>
- Dilmegani, C. (2021). *Data Masking: What it is, how it works, types & best practices*. Retrieved from AI Multiple: <https://research.aimultiple.com/data-masking/>
- Dorasamy, M., Haw, S.-C., & Vigian, T. (2017). Cyber Security Violation in IOT-Enabled Bright Society: A Proposed Framework. *Twenty First Pacific Asia Conference on Information Systems* (pp. 244-249). Langkawi: AIS Electronic Library (AISeL).
- Drechsler, A., & Hevner, A. (2016). A Four-Cycle Model of IS Design Science Research: Capturing the Dynamic Nature of IS Artifact Design. *11th International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, (pp. 1-8). St John.
- Dremel, C. (2017). Barriers to the Adoption of Big Data Analytics. *Twenty-third Americas Conference on Information Systems*, (pp. 1-10). Boston.
- Dutt, D., Natarajan, V., Wilson, A., & Robinson, R. (2020). *Steering into Industry 4.0 in the automotive sector*. Retrieved from Deloitte Insights: <https://www2.deloitte.com/us/en/insights/industry/automotive/industry-4-0-future-of-automotive-industry.html>
- Educba. (2020). *Uses of Hadoop*. Retrieved from Educba: <https://www.educba.com/uses-of-hadoop/>
- Elragal, A., & Haddara, M. (2019). Design Science Research: Evaluation in the Lens of Big Data Analytics. *Systems*, 7(27), 1-8.
- Engineering Simulation and Scientific Software. (2021). *Automotive - The ideal solutions for vehicle engineering*. Retrieved from Engineering Simulation and Scientific Software (ESS): <https://www.esss.co/en/automotive-industry/>
- Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1-4.
- Fox, C. (2017). *Big Data in Manufacturing*. Retrieved October 4, 2018, from Triangle Information Management: <https://triangleinformationmanagement.com/big-data-manufacturing/>
- Frankenfield, J. (2019). *Data Breach*. Retrieved from Investopedia: <https://www.investopedia.com/terms/d/data-breach.asp>
- Fraunhofer IPA. (2016). *Manufacturing service bus - the flexible solution for digitizing production processes*. Stuttgart: Fraunhofer Institute for Manufacturing

- Engineering and Automation IPA. Retrieved from <https://www.fraunhofer.de/en.html>
- Fruhlinger, J. (2020). *What is information security? Definition, principles, and jobs*. Retrieved from CSO: <https://www.csoonline.com/article/3513899/what-is-information-security-definition-principles-and-jobs.html>
- Fusch, P. I., & Ness, L. R. (2015). Are We There Yet? Data Saturation in Qualitative Research. *The Qualitative Report*, 20(9), 1408-1416.
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts , methods and analytics. *International Journal of Information Management*, 35(1), 137-144.
- Garcia , E. (2016). *Hadoop audit and logging "back in time"*. Retrieved from Security Week: <https://www.securityweek.com/hadoop-audit-and-logging-back-time>
- Gelmato. (2020). *Smart manufacturing and the IoT is driving the next industrial revolution*. Retrieved from Gelmato: <https://www.gemalto.com/iot/inspired/smart-manufacturing>
- Giampieri, A., Ling-Chin, J., Smallbone, A., & Roskilly, A. P. (2020). A review of the current automotive manufacturing practice from an energy perspective. *Applied Energy*, 261, 114074.
- Goecks, L. S., de Souza, M., Librelato, T. P., & Trento, L. R. (2021). Design Science Research in practice: review of applications in industrial engineering. *Gestão & Produção*, 28(4), 1-19.
- Grant, C., & Osanloo, A. (2014). Understanding, selecting, and integrating a theoretical framework in dissertation research: creating the blueprint for your “house”. *Administrative Issues Journal: Connecting Education, Practice and Research*, 4(2), 12 - 26.
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact . *MIS Quarterly*, 337-355.
- Guo, L., Wang, J., Wu, H., & Al-Nabhan, N. (2020). XML Security Protection Scheme Based on Kerberos Authentication and Polynomials Authorization. *Mathematical Biosciences and Engineering*, 17(5), 4609–4630.
- Gupta, R., Gupta, S., & Singhal, A. (2014). Big Data : Overview. *International Journal of Computer Trends and Technology (IJCTT)*, 9(5), 266 - 268.
- Gutierrez, D. (2015). *Big data technology for manufacturing*. Retrieved from Inside big data: <http://insidebigdata.com/2015/04/16/big-data-technology-for-manufacturing/>
- Hanada, Y., Hsiao, L., & Levis, P. (2018). Smart Contracts for Machine-to-Machine Communication: Possibilities and Limitations. *Conference: 2018 IEEE*

- International Conference on Internet of Things and Intelligence System (IOTAIS)* (pp. 130-136). Stanford University.
- Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, 6(44), 1 - 16.
- Herawan. (2020). *What is a Query? Database Query Explained*. Retrieved from Hostinger: <https://www.hostinger.com/tutorials/what-is-a-query>
- Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), pp. 75-105.
- Hirschheim, R., & Klein, H. K. (1989). Four Paradigms of Information Systems Development. *Communications of the ACM*, 32(10), 1199 - 1216.
- Huawei. (2021). *Some important features of Apache Hadoop*. Retrieved from Huawei Enterprise Support Community : <https://forum.huawei.com/enterprise/en/some-important-features-of-apache-hadoop/thread/735321-893>
- Hussein, A. A. (2020). How Many Old and New Big Data V's Characteristics, Processing Technology, and Applications (BD1). *International Journal of Application or Innovation in Engineering & Management (IJAIEEM)*, 15-27.
- IBM. (2015). *Making the case for big data and Hadoop in the enterprise*. New York: IBM Corporation.
- IBM. (2016). *Apache Knox gateway overview*. Retrieved from IBM: https://www.ibm.com/support/knowledgecenter/en/SSPT3X_4.1.0/com.ibm.swg.im.infosphere.biginsights.admin.doc/doc/knox_overview.html
- Infor. (2015). Big Data in manufacturing: A compass for growth. 1-8. Retrieved from www.infor.com/content/industry-perspectives/big-data-in-manufacturing.pdf/
- International Software Testing Qualifications Board. (2016). *What is Capability Maturity Model (CMM)? What are CMM Levels?* Retrieved from ISTQB Exam Certification: <http://istqbexamcertification.com/what-is-cmm-capability-maturity-model-what-are-cmm-levels/>
- Ishwarappa, A. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science*, 48, 319 - 324. doi:10.1016/j.procs.2015.04.188
- Ismail, A., Truong, H.-L., & Kastner, W. (2019). Manufacturing process data analysis pipelines: a requirements analysis and survey. *Journal of Big Data*, 6(1), 1 - 26.
- Jain, P., Gyanchandani, M., & Khare, N. (2016). Big Data Privacy: A Technological Perspective and Review. *Journal of Big Data*, 3(25), 1-25.

- Johnston, M. P. (2014). Secondary Data Analysis: A Method of which the Time Has Come. *Qualitative and Quantitative Methods in Libraries (QQML)*, 3(1), 619 – 626.
- Kapil, G., Agrawal, A., Attaallah, A., Algarni, A., Kumar, R., & Khan, R. A. (2020). Attribute Based Honey Encryption Algorithm for Securing Big Data: Hadoop Distributed File System Perspective. *PeerJ Computer Science*. doi:10.7717/peerj-cs.259
- Khan, Y. I. (2020). *Automotive Cyber Security Challenges: A Beginner's Guide*. Independently Published.
- Kivunja, P., & Kuyini, P. B. (2017). Understanding and Applying Research Paradigms in Educational Contexts. *International Journal of Higher Education*, 6(5), 26 - 41.
- Klöser, S. (2019). *Big data analytics in manufacturing: How do we leverage existing data?* Retrieved from Thrive: <https://thrive.dxc.technology/eur/2019/01/14/big-data-analytics-in-manufacturing-how-do-we-leverage-existing-data/>
- Knauf Industries Automotive . (2020). *What does the Fourth Industrial Revolution look like in the automotive industry?* Retrieved from Knauf Industries Automotive: <https://knaufautomotive.com/what-does-the-fourth-industrial-revolution-look-like-in-the-automotive-industry/>
- Kosek , P. (2015). *What is an Algorithm? - Definition & Examples*. Retrieved from Study.com: <https://study.com/academy/lesson/what-is-an-algorithm-definition-examples.html>
- Kovacs, E. (2021). *After IT Outage, Carmakers Kia and Hyundai Say No Evidence of Ransomware Attack*. Retrieved from Security Week: <https://www.securityweek.com/carmakers-kia-and-hyundai-say-no-evidence-ransomware-attack>
- Krauss, R. (2014). *Manufacturing: Big Industry, Big Security Challenges*. Retrieved from Business Insights: <https://businessinsights.bitdefender.com/manufacturing-big-industry-big-security-challenges>
- Kuhn, T. S. (1962). The Structure of Scientific Revolutions. *International Encyclopedia of Unified Science*, 2(2).
- Kumar , G. B. (2015). An Encyclopedic Overview of ‘Big Data’ Analytics. *International Journal of Applied Engineering Research*, 10(3), 5681-5705.
- Kumar, R. (2011). *Research Methodology: A step by step guide for beginners* (3rd ed.). London: SAGE Publications Ltd.
- Kurtz, J., & Shockley, R. (2013). *Analytics: The real-world use of big data in manufacturing*. New York City: IBM Global Services. Retrieved from <https://www.ibm.com/downloads/cas/ONBGKB82>

- Lafuente, G. (2014). *Big Data Security - Challenges & Solutions*. Retrieved March 03, 2017, from MWR Infosecurity: <https://www.mwrinfosecurity.com/our-thinking/big-data-security-challenges-and-solutions/>
- Law Insider. (2018). *Definition of Regulatory Conditions*. Retrieved from Law Insider: <https://www.lawinsider.com/dictionary/regulatory-conditions>
- Lee, J., Lapira, E., Bagheri, B., & Kao, H.-a. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, 1(1), 38 - 41.
- Lehman, D., & Ramanujam, R. (2009). Selectivity in organisational rule violations. *The Academy of Management Review*, 34(4), 643-657.
- Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive Analytics: Literature Review and Research Challenges. *International Journal of Information Management*, 50, 57 - 70. doi:10.1016/j.ijinfomgt.2019.04.003
- Li, F., Li, H., Niu, B., & Chen, J. (2019). Privacy Computing: Concept, Computing Framework, and Future Development Trends. *Engineering*, 5(6), 1179–1192.
- Liebe, G., Tichy, M., Knauss, E., Ljungkrantz, O., & Stieglbauer, G. (2018). Organisation and communication problems in automotive requirements engineering. *Requirements Eng*, 23(1), 145–167.
- Lorant, A. (2016). *DIY vs. fully integrated Hadoop – What’s best for your organization?* Retrieved from Networkworld: <https://www.networkworld.com/article/3153186/diy-vs-fully-integrated-hadoop-what-s-best-for-your-organization.html>
- Lovalekar, S. (2014). Big Data: An Emerging Trend In Future. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(1), 538-541.
- Maayan, G. D. (2020). *Big Data Security: Challenges and Solutions*. Retrieved from Data Versity: <https://www.dataversity.net/big-data-security-challenges-and-solutions/>
- Machado, C. G., Winrotha, M. P., & Ribeiro da Silva, E. H. (2019). Sustainable manufacturing in Industry 4.0: an emerging research agenda. *International Journal of Production Research*.
- Majid, U. (2018). Research Fundamentals: Study Design, Population, and Sample Size. *Undergraduate Research in Natural and Clinical Science and Technology (URNCSST) Journal*, 2(1), 1-7.
- Mallon, S. (2018). *IoT Is The Most Important Development Of The 21st Century*. Retrieved from Smart Data Collective: <https://www.smartdatacollective.com/iot-most-important-development-of-21st-century/>

- Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of Research Design and Methodology*. New Jersey: John Wiley & Sons.
- Martin, N. (2019). *What is a data breach?* Retrieved from Forbes:
<https://www.forbes.com/sites/nicolemartin1/2019/02/25/what-is-a-data-breach/#3db6d25114bb>
- Martinelli, K. (2018, April 6). *Password Security Guidance*. Retrieved from High Speed Training : <https://www.highspeedtraining.co.uk/hub/password-security-guidance/>
- Matthews, K. (2018). *How big data is improving quality control and testing*. Retrieved from The Innovation Enterprise Channels:
<https://channels.theinnovationenterprise.com/articles/how-big-data-is-improving-quality-control-and-testing>
- Matthews, K. (2019). *7 Advantages of Using Encryption Technology for Data Protection*. Retrieved from Smart Data Collective: <https://www.smartdatacollective.com/5-advantages-using-encryption-technology-data-protection/>
- Mattsson, U. (2014). Bridging the Gap Between Access and Security in Big Data. *ISACA Journal*, 6(1), 1-5.
- Maxwell, J. (2005). *Qualitative Research Design: An Interactive Approach*. California: SAGE Publications.
- McAllister, J. (2018). *7 Key Members of Every Big Data Team*. Retrieved from Inside Big Data: <https://insidebigdata.com/2018/02/16/7-key-members-every-big-data-team/>
- McLaughlin, J., & James, G. (2013). *What is Organizational Culture? - Definition & Characteristics*. Retrieved from Study.com:
<https://study.com/academy/lesson/what-is-organizational-culture-definition-characteristics.html>
- Microsoft. (2019, March 11). *Description of the standard terminology that is used to describe Microsoft software updates*. Retrieved from Microsoft Support:
<https://support.microsoft.com/en-za/help/824684/description-of-the-standard-terminology-that-is-used-to-describe-micro>
- Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Journal of Information Systems and e-Business Management*, 16(3), 547-578.
- Mixson , E. (2021). *An introduction to tokenisation*. Retrieved from Cyber Security Hub (CSHUB): <https://www.cshub.com/executive-decisions/articles/an-introduction-to-tokenizationnbs>

- Moura, J., & Serrão, C. (2015). Security and Privacy Issues of Big Data. In J. A. Moura, & C. Serrao, *Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence*. IGI Global. doi:10.4018/978-1-4666-8505-5.ch002
- Müller, O., Fay, M., & vom Brocke, J. (2018). The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. *Journal of Management Information Systems*, 35(2). doi:10.1080/07421222.2018.1451955
- Myers, M. D. (1997). Qualitative Research in Information Systems. *MIS Quarterly*, 21(2), 241-242.
- Nagy, J., Oláh, J., Erde, E., Máté, D., & Popp, J. (2018). The Role and Impact of Industry 4.0 and the Internet of Things on the Business Strategy of the Value Chain—The Case of Hungary. *Sustainability*, 10(10). doi:10.3390/su10103491
- Naidoo, R., Gerber, A., & van der Merwe, A. (2012). An Exploratory Survey of Design Science Research amongst South African Computing Scholars. *Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT)*. Pretoria: ACM International Conference Proceeding Series. doi:10.1145/2389836.2389876
- Normandeau, K. (2013). *Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity*. Retrieved from Inside Big Data: <https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, 16(1), 1-13.
- O'Donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. (2015). Big data in manufacturing: a systematic mapping study. *Journal of Big Data*, 2(20), 1 - 22.
- Oates, B. (2006). *Researching Information Systems and Computing*. London: Sage Publishers.
- Ogbuke, N. J., Yusuf, Y. Y., Dharma, K., & Mercangoz, B. A. (2022). Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society. *Production Planning & Control*, 33(2-3), 123-137.
- Olivier, M. (2004). *Information Technology Research: A practical guide for Computer Science and Informatics*. Pretoria: Van Schaik Publishers.
- Oracle. (2015). Improving Manufacturing Performance with Big Data. 1 - 20. Retrieved from www.oracle.com/us/technologies/big-data/big-data-manufacturing-2511058.pdf

- Perwej , Y. (2019). The Hadoop Security in Big Data: A Technological Viewpoint and Analysis. *International Journal of Scientific Research in Computer Science and Engineering*, 7(13), 1-13.
- Pittaluga, I. (2021). *5 tips for better business continuity and disaster recovery plans*. Retrieved from GCN - The technology that drives government IT: <https://gcn.com/articles/2021/06/23/disaster-recovery-plans.aspx>
- Ploner, L. (2013). *Industry Insiders Report: Big Data in manufacturing - Part 1*. Retrieved from Renesas: <https://www.renesas.com/en-in/about/web-magazine/edge/global/13-big-data.html>
- Pratt, M. (2021). *Building a big data architecture: Core components, best practices*. Retrieved from Tech Target: <https://searchdatamanagement.techtarget.com/feature/Building-a-big-data-architecture-Core-components-best-practices>
- Procurement Academy. (2019). *Procurement Strategy Development*. Retrieved from Procurement Academy: <https://www.procurement-academy.com/procurement-competences/procurement-strategy-development/>
- Raguvir, S., & Babu, S. (2020). Detecting Anomalies in Users – An UEBA Approach. *Proceedings of the International Conference on Industrial Engineering and Operations Management* (pp. 863-876). Dubai: IEOM Society International.
- Rahi, S. (2017). Research Design and Methods: A Systematic Review of Research Paradigms, Sampling Issues and Instruments Development. *International Journal of Economics & Management Sciences*, 6(2), 1 - 5.
- Ramaseshan, S. (2019). *Effective User Access Reviews*. Retrieved from ISACA: <https://www.isaca.org/resources/isaca-journal/issues/2019/volume-4/effective-user-access-reviews>
- Reghunath, K. (2017). Real-time intrusion detection system for big data. *International Journal of Peer to Peer Networks (IJP2P)*, 8(1), 1-20.
- Reidy, S. (2018). *Have you done enough to optimize your automotive supply chain?* Retrieved February 15, 2020, from Arviem: <https://arviem.com/optimize-your-automotive-supply-chain/>
- Riahi, Y., & Riahi, S. (2018). Big Data and Big Data Analytics: Concepts, Types and Technologies. *International Journal of Research and Engineering*, 5(9), 524-528.
- Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (2013). *Qualitative Research Practice: A guide for social science students and researchers*. London: Sage Publishers.

- Rizk, A., Bergvall-Kåreborn, B., & Elragal, A. (2017). Digital Service Innovation Enabled by Big Data Analytics - A Review and the Way Forward. *Proceedings of the 50th Hawaii International Conference on System Sciences*, (pp. 1247 - 1256).
- Robinson, A. (2016). *4 Big Benefits for the Use & Implementation of Predictive Analytics In Manufacturing*. Retrieved from Cerasis:
<http://cerasis.com/2016/03/16/predictive-analytics-in-manufacturing/>
- Russo, I., Confente, I., & Borghesi, A. (2015). Using big data in the supply chain context: opportunities and challenges. *The 16th European Conference on Knowledge Management ECKM 2015* (pp. 649 - 656). Udine: Academic Conferences and Publishing International Limited.
- Sage. (2020). *GDPR - Manage your business data retention period*. Retrieved from Sage Business Cloud Accounting Start : <https://help.accounting.sage.com/en-gb/start/settings/setup-gdpr-retention-period.html>
- Saldaña, J. (2013). *The Coding Manual for Qualitative Researchers* (2nd ed.). Los Angeles: SAGE Publications Ltd.
- Salkind, J. N. (2010). *Encyclopedia of research design*. London: Sage Publications.
- SAS. (2022). *Hadoop - What it is and why it matters*. Retrieved from SAS:
https://www.sas.com/en_zh/insights/big-data/hadoop.html
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., . . . Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, 52(4), 1893–1907.
- Saunders, M., Lewis, P., & Thornhill, A. (2007). *Research Methods for Business Students* (4 ed.). Essex: Pearson Education Limited.
- Schatsky, D., Camhi, J., & Muraskin, C. (2019). *Data ecosystems - How third-party information can enhance data analytics*. Retrieved from Deloitte:
<https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html>
- Sekaran, U., & Bougie, R. (2013). *Research Methods for Business* (6th ed.). West Sussex: John Wiley & Sons Ltd.
- Shacklett, M. (2016). *4 security measures that strengthen big data governance*. Retrieved from Tech Republic: <http://www.techrepublic.com/article/4-security-measures-that-strengthen-big-data-governance/>
- Sharma, P. P., & Navdeti, C. P. (2014). Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(2), 2126-2131.

- Shepardson, D. (2021). *VW says data breach at vendor impacted 3.3 million people in North America*. Retrieved from Auto Economic Times:
<https://auto.economictimes.indiatimes.com/news/passenger-vehicle/cars/vw-says-data-breach-at-vendor-impacted-3-3-million-people-in-north-america/83452059>
- Simon, L., & Ramesh, S. (2016). Big Data - Understanding the Security Issues. *International Journal for Scientific Research & Development - IJSRD*, 4(5), 370 - 372.
- Singh , A. J. (2014). Security Issues in Big Data: In Context with Hadoop. *International Journal of Innovative and Emerging Research in Engineering*, 2(3), 127-130.
- Sinkovics, N. (2018). *The SAGE handbook of qualitative business management research methods*. Thousand Oaks, US: Sage Publications Inc.
- Sirisha, N., Kiran, K., & Karthik, R. (2018). Hadoop Security Challenges and Its Solution Using KNOX. *Indonesian Journal of Electrical Engineering and Computer Science*, 12(1), 107-116.
- Smeda, J. (2015). Benefits, business considerations and risks of big data. *Master's thesis, Stellenbosch University*. Retrieved from
<http://scholar.sun.ac.za/handle/10019.1/96684>
- Stephens, D. (2020). *Why the Car is the Secret to a Connected World*. Retrieved from IoT Evolution World: <https://www.iotevolutionworld.com/smart-transport/articles/444885-why-car-the-secret-a-connected-world.htm>
- Stumpf, R. (2021). *The Apparent Hackers Behind Kia's Ransomware Attack Are Demanding Millions in Bitcoin*. Retrieved from The Drive:
<https://www.thedrive.com/tech/39309/the-apparent-hackers-behind-kias-ransomware-attack-are-demanding-millions-in-bitcoin>
- Subrahmanyam, V., & Aruna, K. (2017). Future Automobile an Introduction of IoT. *International Journal of Trend in Research and Development (IJTRD)*, 88-90.
- Summersgill, M., & Coviello, A. (2019). *Insight: The Future of Automotive Trade Secret Litigation*. Retrieved from Bloomberg Law: <https://news.bloomberglaw.com/ip-law/insight-the-future-of-automotive-trade-secret-litigation>
- Swanson, R. (2013). *Theory Building in Applied Disciplines*. San Francisco: Berrett-Koehler Publishers.
- Syafrudin, M., Alfian, G., Fitriyani, N. L., & Rhee, J. (2018). Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing. *Sensors*, 18(9), 2946 - 2519.
- Tague, N. R. (2005). *The Quality Toolbox, Second Edition*. Milwaukee, USA: Quality Press.

- Taherdoost, H. (2016). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *International Journal of Academic Research in Management (IJARM)*, 5(2), 18-27.
- Tasmin, R., Rahman, N. S., Jaafar, I., Hamid, N. A., & Ngadiman, Y. (2020). The Readiness of Automotive Manufacturing Company on Industrial 4.0 Towards Quality Performance. *The International Journal of Integrated Engineering*, 12(7), 160-172.
- Tavana, M., Hajipour, V., & Oveisi, S. (2020). IoT-based enterprise resource planning: Challenges, open issues, applications, architecture, and future research directions. *Internet of Things*, 11(1), 1-28.
- Tawalbeh, L., Muheidat, F., Tawalbeh, M., & Quwaider, M. (2020). IoT Privacy and Security: Challenges and Solutions. *Applied Sciences*, 10(12), 4102-4119.
- Tay, S. I., Chaun, L. T., Aziati, A. N., & Ahmed, A. N. (2018). An Overview of Industry 4.0: Definition, Components and Government Initiatives. *Journal of Advanced Research in Dynamical & Control Systems*, 10(14), 1379 - 1387.
- Thiesse, F., & Fliesch, E. (2007). *Adoption and Diffusion of RFID Technology in the Automotive Industry*. Zurich: Auto-ID Labs.
- Tidy, J. (2020). *Honda's global operations hit by cyber-attack*. Retrieved from BBC: <https://www.bbc.com/news/technology-52982427>
- Tole, A. A. (2013). Big Data Challenges. *Database Systems Journal*, 5(3), 31 - 40.
- Toshiba. (2020). *Cyber-Physical Approach in the Automotive Industry: Stepping into the Future of Model-Based Development*. Retrieved from Toshiba: <https://www.toshiba-clip.com/en/detail/p=485>
- Towerstone Leadership Centre. (2016, May 1). *What are organisational values and why are they important*. Retrieved from Towerstone Leadership Centre: <https://www.towerstone-global.com/what-are-organisational-values-and-why-are-they-important/>
- Trend Micro. (2018). *Data Breaches 101: How They Happen, What Gets Stolen, and Where It All Goes*. Retrieved from Trend Micro: <https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/data-breach-101>
- Trochim, W. (2001). *The Research Methods Knowledge Base* (2nd ed.). Chicago: Thomson Learning.

- Turner, C. (2020). *Tokenization VS. Encryption: Pros and Cons*. Retrieved from eSecurityPlanet: <https://www.esecurityplanet.com/threats/tokenization-vs-encryption/#tokenization>
- University of Southern California. (2018). *Research Guides*. Retrieved from USC Libraries: <http://libguides.usc.edu/writingguide/theoreticalframework>
- Vaishnavi, V., & Kuechler, W. (2008). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. Florida: Auerbach Publications.
- van der Kleut, J. (2021). *What is ransomware and how to help prevent ransomware attacks*. Retrieved from Norton: <https://us.norton.com/internetsecurity-malware-ransomware-5-dos-and-donts.html>
- Wall, J. D., Lowry, P. B., & Barlow, J. (2016). Organisational violations of externally governed privacy and security rules: Explaining and Predicting Selective Violations Under Conditions of Strain and Excess. *Journal of the Association for Information Systems (JAIS)*, 17(1), 39 - 76.
- Wang, L., & Alexander, C. A. (2015). Big Data in Design and Manufacturing Engineering. *American Journal of Engineering and Applied Sciences*, 8(2), 223 - 232.
- Wasmund, R. (2017). *The internet of services in Industrie 4.0*. Retrieved from Concept Systems Inc: <https://conceptsyste.msinc.com/the-internet-of-services-in-industrie-4-0/?v=a284e24d5f46>
- Web Centre for Social Research Methods. (2006). *Pattern Matching for Construct Validity*. Retrieved from Web Centre for Social Research Methods: <http://www.socialresearchmethods.net/kb/pmconval.php>
- Weber, S. (2010). Design Science Research: Paradigm or Approach? *America's Conference on Information Systems (AMCIS)*. Peru: AIS Electronic Library .
- Winton, N. (2021). *In 2021 With Europe Strongest, Barring Anything Unexpected*. Retrieved from Forbes.
- Woolledge, S. (2014). *7 Ways Hadoop Can Optimize Manufacturing Performance*. Retrieved from mbtmag: <https://www.mbtmag.com/article/2014/10/7-ways-hadoop-can-optimize-manufacturing-performance>
- Wróbel , G., & Wikira, M. D. (2019). Overview of Big Data platforms. *Journal Computer Sciences Institute (JCSI)*, 13(1), 283-287.
- Yadav, D., Maheshwari, H., & Chandra, U. (2019). Big Data Hadoop: Security and Privacy. *2nd International Conference on Advanced Computing and Software Engineering (ICACSE)* (pp. 358-365). Sultanpur: SSRN Electronic Journal.

- Younas, M. (2019). Research Challenges of Big Data. *Service Oriented Computing and Applications*, 13, 105 - 107.
- Zaki, M., Theodoulidis, B., Shapira, P., Neely, A., & Tepel, M. F. (2019). Redistributed Manufacturing and the Impact of Big Data: A Consumer Goods Perspective. *Production Planning & Control*, 30(7), 568-581.
- Zettaset. (2014). *The Big Data Security Gap: Protecting the Hadoop Cluster*. Retrieved from Zettaset.



University of Fort Hare
Together in Excellence

List of Abbreviations and Acronyms

ACLS – Access Control List

AM - Additive Manufacturing

AR - Augmented Reality

CIO - Chief Information Officer

CMM – Capability Maturity Model

CPS - Cyber Physical System

ERP – Enterprise Resource Planning

HTTP - Hyper-Text Transfer Protocol

IoT – Internet of Things

IS – Information Systems



ISSO - Information Systems Security Officer

University of Fort Hare
Together in Excellence

IT – Information Technology

IoT – Internet of Things

MES - Manufacturing Execution System

MSB - Manufacturing Service Bus

MQTT - Message Queue Telemetry Transport

PDF - Portable Document Format

PLC - Programmable Logic Controller

RDBMS – Relational Database Management System

RFID - Radio-Frequency Identification

SOIPSVM – Selective Organizational Information Privacy and Security Violations Model

SSO - Single Sign-On

UREC - University of Fort Hare's Research Ethics Committee

VIN - Vehicle Identification Number



Glossary

Apache Hadoop Technology - An open-source software platform that processes big data. It contains a Java-based programming framework that supports processing and storing large datasets in a distributed computing environment.

Big data - Data of a large volume, velocity, variety, value, and veracity that exceeds the processing capacity of traditional databases.

Capability Maturity Model (CMM) - The CMM signifies five maturity levels of increasing organized and systematically mature processes

Big Data analytics - Processes used to extract useful information from large datasets. Data is extracted and categorised to identify and analyse patterns and techniques for business competitiveness.

Privacy - Privacy is defined as the appropriate use of data.

Security - Security is defined as all the practices and processes to ensure data is not being used or accessed by unauthorised individuals or parties.

Selective Organizational Information Privacy and Security Violations Model SOIPSVM - The acronym for selective organisational information privacy and security violations model explains how organisational structures and processes, including characteristics of regulatory rules, can modify an organisation's understanding of risk.



University of Fort Hare
Together in Excellence

Appendix A: Ethical Clearance



University of Fort Hare
Together in Excellence

ETHICS CLEARANCE REC-270710-028-RA Level 01

Project Number:	PID031SPAD01
Project title:	Big Data in the Automotive Industry: A Framework to address privacy concerns in Hadoop Technology.
Qualification:	Master of Commerce in Information Systems
Principal Researcher:	Prenisha Padayachee
Supervisor:	Prof R. Piderit
Co-supervisor:	N/A

On behalf of the University of Fort Hare's Research Ethics Committee (UREC) I hereby grant ethics approval for PID031SPAD01. This approval is valid for 12 months from the date of approval. Renewal of approval must be applied for BEFORE termination of this approval period. Renewal is subject to receipt of a satisfactory progress report. The approval covers the undertakings contained in the above-mentioned project and research instrument(s). The research may commence as from the 22/08/19, using the reference number indicated above.

Note that should any other instruments be required or amendments become necessary, these require separate authorisation.

Please note that the UREC must be informed immediately of

- Any material changes in the conditions or undertakings mentioned in the document;

- Any material breaches of ethical undertakings or events that impact upon the ethical conduct of the research.

The Principal Researcher must report to the UREC in the prescribed format, where applicable, annually, and at the end of the project, in respect of ethical compliance.

The UREC retains the right to

- Withdraw or amend this approval if
 - Any unethical principal or practices are revealed or suspected;
 - Relevant information has been withheld or misrepresented;
 - Regulatory changes of whatsoever nature so require;
 - The conditions contained in the Certificate have not been adhered to.
- Request access to any information or data at any time during the course or after completion of the project.

Your compliance with DoH 2015 guidelines and other regulatory instruments and with UREC ethics requirements as contained in the UREC terms of reference and standard operating procedures, is implied.

The UREC wishes you well in your research.

Yours sincerely



Professor Pumla Dineo Gqola
Acting UREC-Chairperson
22 August 2019

Appendix B: Overview and informed consent

Overview



Prenisha Padayachee

University of Fort Hare

Big Data in the Automotive Industry: A Framework to address privacy concerns in Hadoop Technology

Overview and voluntary participation:

Dear participant

Thank you for taking the time to participate in this interview and questionnaire. I am a student in the Information Systems Department at the University of Fort Hare (East London Campus), currently conducting research for my M.Com Information Systems degree under the supervision of Prof. Roxanne Piderit. This research study focuses on creating a framework to address privacy concerns of big data in Hadoop technology at an automotive industry. This interview and questionnaire should take approximately 20 minutes to complete. This is an independent research study and participation is voluntary. This research study, interview and questionnaire adheres to the code of ethics stipulated in the researchers' ethical clearance certificate (No: PID031SPAD01).

If you have any questions regarding the interview, questionnaire or this research project in general, please contact Prenisha Padayachee via email (prenisha007@gmail.com) or Prof Roxanne Piderit at (043) 704 7094. Please find an informed consent below indicating your agreement to participate in this study.

We look forward to your response.

Yours sincerely,

Prenisha Padayachee

Informed consent and overview

Informed consent form: I hereby agree to participate in research regarding a framework to address privacy concerns of big data in Hadoop technology at an automotive industry. I understand that I am participating freely and without being forced in any way to do so. I also understand that I can stop at any point, should I not want to continue and that this decision will not in any way affect me negatively. I understand that this is a research project whose purpose is not necessarily to benefit me personally. I have received the telephone number of a person to contact should I need to speak about any issues which may arise in this interview and questionnaire. Your responses will be recorded, treated as strictly confidential and the anonymity of respondents is assured. Therefore, no person or firms will have access to your interview and questionnaire.

Sign: _____

Date: _____



University of Fort Hare
Together in Excellence

Appendix C: Interview and questionnaire

This section presented the template used for the interview and questionnaire.


Big Data in the Automotive Industry: A Framework to address privacy concerns in Hadoop Technology

General Questions

- Can you tell me a little bit about what your job entails?
- What is the Internet of Things, and how is it affecting the automotive industry?

1. How is big data being generated and used in the automotive manufacturing industry?

Generation of big data


- 
- How is the data generated and collected?
 - What data can you collect today? In which areas can data be collected?
 - What do you do with the data collected today?
 - What data is stored today? And why?
 - What data is not stored? And why?
 - How much does it cost to store data?

Use of big data

- How can data be analysed in-house? With which tools/techniques?
- What kind of different business models for monetizing data from connected things do you know of?
- What do you think is needed to create value from data collected by cars?
- What possibilities do you see with collected data in the automotive industry? What use cases?
- How does car data create value for car manufacturers? (e.g., internally/externally)

- How do you think the vehicle manufacturers' role in value creation will change with the automotive industry moving towards data-enabled services?
- As value becomes increasingly co-created by different actors/partners, what do you think is important for vehicle manufacturers to capture the value?
- How does the product development process look today?
- How can you create value from data in your product development process?
Cutting costs/improving product?
- Do you think that the organizational structure must change in some way to be better at translating data insights into action?

2. What types of privacy challenges are experienced in the automotive manufacturing industry in the big data environment, specifically Hadoop technology?

- 
- Does privacy matter to you when using devices that connect to the internet?
 - Who has access to your IoT devices?
 - What do you believe is the risk for your organization due to the possibility that: My organization could be issued severe sanctions for violations of [privacy of big data in Hadoop]?

Ratings & Comments (1 – Strongly disagree, 2- Disagree, 3 – Agree, 4 – Strongly Agree)

- The positive internal consequences of violating [privacy of big data in Hadoop] are clear to key decision-makers in this organization.
- The positive internal consequences of violating [privacy of big data in Hadoop] are clear to key decision-makers in this organization
- The negative internal consequences of violating [privacy of big data in Hadoop] are clear to key decision-makers in this organization.
- Organizations caught violating [privacy of big data in Hadoop] will be severely punished.

- Organizations caught violating [privacy of big data in Hadoop] will be reprimanded.
- Organizations caught violating [privacy of big data in Hadoop] will face serious consequences.
- For our organization, actions against violating [privacy of big data in Hadoop] are instantaneous.
- For our organization, actions against violating [privacy of big data in Hadoop] are timely
- For our organization, actions against violating [privacy of big data in Hadoop] are immediate.
- If my organization follows [privacy of big data in Hadoop], we will have to sacrifice some of the core values of my organization
- If you were an employee at this organization, what is the likelihood that you would have violated [privacy of big data in Hadoop]?

3. What measures can manufacturers take to protect their privacy?

- Do you encrypt any of your devices?
- Do you use a different password on each of your IoT devices?
- Do you install manufacturers security updates on your IoT devices?

Source: (Andersson & Axelsson, 2018)

Ratings/Questionnaire Ratings & Comments (1 – Strongly disagree, 2- Disagree, 3 – Agree, 4 – Strongly Agree)

- Our business transactions with other units should be approved by upper management
- Upper management has the ultimate power to decide whether we collaborate with other units in the organization.
- The information provided on the [organization that administers a given privacy or /security standard] website about [privacy of big data in Hadoop] was sufficiently unambiguous.

- A violation of [privacy of big data in Hadoop would be problematic because it would entail a violation of other rules at the same time.



University of Fort Hare
Together in Excellence