A Holistic Framework for Complex Big Data Governance

Rita Uwaeluemuaka Obatolu

A thesis submitted in partial fulfilment of the requirements of the

University of East London for the degree of

Professional Doctorate in Data Science

School of Architecture, Computing & Engineering

September 2021

Abstract

Big data assets are large datasets that can be leveraged by organisations if the capabilities exist, but it also brings considerable challenges. Despite the benefits that can be realised, the lack of proper big data governance is a major barrier, making the processing and control of this data exceptionally difficult to execute correctly. More specifically, organisations reportedly struggle to incorporate big data governance into their existing structures and business models to derive value from big data initiatives.

Big data governance is an emerging research domain, gaining attention from both Information Systems scholars and the practitioner community. Nonetheless, there appears to have been limited scientific research in the area and most existing data governance approaches were limited, given they do not address end-to-end aspects of how big data could be governed. Furthermore, no suitable framework for handling data governance against big data complexities was found to be available. Thus, the main contribution of the work presented in this thesis is to address this requirement; by advancing research in this field and designing a novel holistic big data governance framework capable of supporting global organisations to effectively manage big data as an asset, thereby obtaining value from their big data initiatives.

An extensive systematic literature review was done in order to uncover the published content that reflects the current state of knowledge in big data governance. To facilitate the creation of the proposed framework a grounded theory methodology was used to analyse openly available parliamentary inquiry data sources, with particular focus on identifying the core criteria for big data governance. The resulting novel framework generated provides new knowledge by identifying several big data governance building blocks; which are classified as strategic goals, execution stages, enablers and 22 core big data governance components to ensure an effective big data governance programme. Moreover, thesis findings indicate that big data complexities extend to the ethical side of big data governance and this is taken into consideration in the framework design. An ‘ethics by design’ component is proposed to influence how this can be addressed in a structured way.

Key Words: Big Data Governance, Big Data Governance Framework, Data Governance, Ethical Big Data Governance, Big Data Issues, Big Data challenges, Big Data risk factors, Systematic Review

Contents

List of Tables

List of Figures

List of Accompanying Material

1) Atlas ti Code: *Thesis Project (Rita Obatolu).atlproj9*

Abbreviations

BD: Big Data

DG: Data Governance

ROI: Return on Investment

DQ: Data quality

BDQ: Big Data Quality

MDM: Master Data Management

IS: Information Systems

IT: Information Technology

IEC: Information Economy Council

ROI: Return on Investment

RACI: Responsible Accountable Consulted Informed

DLC: Data Life Cycle

BDLC: Big Data Life Cycle

E2E: End to end

BU: Business Unit

KPI: Key Performance Indicators

PII: Personal Identifiable Information

GDPR: General Data Protection Regulation

IBM: International Business Machines Corporation

Bus obj: Business Objective

ML: Machine learning

TDWI: The Data Warehouse Institute

Acknowledgements

This thesis study is dedicated to my husband Kofo Obatolu, my partner through life. I do not take the unending love and support he has given me through this doctoral journey for granted. Without his love and support, this achievement would have been much harder to attain. It is also dedicated to my children (Karl, Kevin & Jessica), I am hoping that setting my sights on the honour of being called "Dr. Mum" will show them the endless possibilities in life if they strive and work hard at achieving their goals.

I would like to say thank you to my mum, Mrs Christiana Remi Dieyi, for the sacrifices she has made my entire life so that I could achieve my own dreams. Encouraging me and always being there for me at each step along the way. I will always be indebted to her. Similarly, gratitude goes to my late father, Mr Albert Chukwuka Dieyi, for altering the course of our family's history by summoning the courage to leave Lagos to the United Kingdom many years ago, to offer his family some of the opportunities that otherwise might not have been available. Remembering the words he used to say was the encouragement I needed to never give up during this journey – "Those that do it, do not have two heads".

I will remain forever grateful for the support and encouragement of my former Supervisor Professor Allan Brimicombe, whose thought provoking feedback and advice helped ensure that this would become a quality piece of work. My thanks also go to Dr Yang Li for his supervision during the final stages of my doctorate, after Professor Allan Brimicombe retired last year. I am grateful to them both for the guidance and direction they provided from the beginning to completion of the programme.

Most importantly, I give thanks to Almighty God who is and always will be "the author and finisher of my faith". For what God cannot do, does not exist.

1.0 Introduction

1.1 Research Problem and Purpose of Thesis Study

Since its emergence, big data (BD) has continually received attention from all sectors. Existing literature provides evidence that organisations are becoming increasingly serious about the notion of data as an asset (Huff et al, 2020; Abraham et al, 2019; Nielsen et al, 2018; Niemi et al, 2016; Khatri et al, 2010; Otto, 2011a), presenting numerous ways that BD is being exploited by organisations to create value across sectors of the global economy. Analysing these large datasets, has become a basis of competition and decision making within modern companies which underpins a new wave of productivity growth innovation, customer acquisition and retention. There are further revelations within the scientific and practitioner community of significant improvements from using BD for enhanced customer knowledge, better targeted services, economic and social benefit (Aversa et al, 2021; Brinch et al, 2021; Sestino et al, 2020; Vassakis et al, 2018). Arguably, organisations who fail to effectively utilise these datasets and the insights that can be generated from them could potentially fall behind their competitors.

Although the benefits of BD are considerable; global organisations are confronted by the challenges of BD that impacts various industry sectors. There remains a large amount of issues when implementing BD initiatives related to the characteristics of the data itself (e.g. data volume, variety, velocity, veracity, volatility, quality and discovery), as well as process and management challenges e.g. privacy, security, governance and ethical aspects (Sivarajah et al, 2017; Cui, 2019). Drawbacks hampering the ethical and responsible use of BD by both private and public sector impact public trust, whilst issues related to transparency and security can also limit their uptake. Consequently, whilst BD is undoubtedly one of the valuable assets organisations possess, it is reported in the literature to be increasingly onerous to manage and control (Kim et al, 2018; Soares, 2012). Authors also highlighting costly redundant resource hungry projects, data sources that are not rationalised and inflexible systems that do not cater for changing business needs (Tan et al, 2017; Cheong et al, 2007).

Some researchers argue that to effectively utilise or prepare for the transition to BD, global organisations that adhere to an established data governance (DG) framework are better

positioned to achieve greater success in both the short and long term; hence, better prepared to succeed or derive the desired business value (Otto, 2011a; Khatri et al, 2010). Earlier work emphasise that performed correctly, a robust big data governance program can remove the guesswork out of discovering and using the right data to make business decisions – regardless of its source, what data type it is, or its velocity (Soares, 2012). Nonetheless, there appears to be major challenges associated with adopting accessible big data governance approaches to manage the complexities introduced by the use of BD systems, as modern global organisations strive to leverage data as an asset. Big data governance is reported to be a huge challenge for Chief Data Officers (CDOs), Chief Information Officers (CIOs), Data Stewards (DSs) and other BD stakeholders. Hence, organisations are increasingly concerned with this domain, owing to the benefits that can be derived from governance of this data.

According to the literature review conducted, scientific research in the DG domain is limited, fragmented and still lacks a commonly accepted definition. Despite recognition and the call for DG frameworks by researchers (Ho et al, 2020; Konig, 2021; Praseyto, 2015; Wende, 2007; Niemi et al, 2013), this study observes that there is a limited amount of them that address big data governance and this specific area of research appears to be in its infancy. Although there is some research that contributes to the understanding of DG and there is evidence of DG frameworks that have already been proposed, nonetheless all these frameworks appear to have limitations when considering BD complexities. Firstly, none of the DG frameworks appear to take a holistic approach on how to address big data governance within global operating organisations, all tending to concentrate on singular aspects within this domain. Secondly, little attention has been paid to building systematic and structured research that compares the variety of existing DG frameworks against BD complexities and compare their usefulness in this aspect. Review of the literature indicates a considerable degree of ambiguity and a lack of research that specifically studies criteria for governing BD holistically. Additionally, there are increasing ethical tensions regarding some of the uses of BD which appear to be ignored by earlier DG frameworks which is addressed in the thesis study.

Another salient issue is that some organisations reportedly struggle to integrate DG into their structures and business models to create value from BD initiatives or programs. An examination of the literature discovered that most existing DG frameworks do not lay any emphasis on the "how" of implementing DG frameworks to handle BD complexities, which could be a reason for this problem. As organisation endeavours to implement BD initiatives,

it is important they have the appropriate guidance to develop "fit for purpose" big data governance frameworks to support the management of BD systems and guidance on how they can incorporate these into existing organisation structures.

Thus, this doctorate thesis proposes to focus research direction on organisations that are increasingly confronted by the global challenge of accelerated accumulation of BD from disparate channels, specifically where it is used for delivering BD services, precision target marketing or deriving intelligence to inform business decisions. These organisations face increased demands from the added complexity of BD collection, storage, analysis, processing, integration, as well as legislative and regulatory compliance demands. Given the complexities within these operating scenarios, these organisations need to establish robust foundations to govern BD assets effectively and subsequently be better positioned to achieve success from BD initiatives.

1.2 Aim and Objectives of the Study

The overall aim of the doctorate thesis is to design a holistic and accessible big data governance framework that can serve as a toolkit for organisations to govern BD aspects, so they can advance in their BD driven strategies. The objectives are to:

- Perform a systematic literature review to understand the state of the art of big data governance

- Examine existing DG frameworks, considering the limitations of these approaches to govern BD aspects, with the view to producing design recommendations for an improved approach.

- Perform data analysis on parliamentary inquiry submissions, to extract the common BD complexities to be mitigated through effective big data governance.

    - It is deemed necessary to establish a system for introduction and way to recognise BD challenges or risk factors, to aid the understanding of which factors need to be mitigated through effective governance and enhance the overall success of the framework.

- Provide a detailed step-by-step description of the methodology used.

- Design and create a novel conceptual big data governance framework. Including an 'ethics by design' component which forms part of the overall big data governance framework.

1.3 Aim of the Big Data Governance framework

Authors from both the research and practitioner community caution that a robust big data governance framework is crucial to the success of BD initiatives and the right management of big data (Kim et al, 2018; Al-Badi et al, 2018; Li et al, 2019; Soares, 2012). Decision makers require a thorough comprehension of how to reduce risks inherent in BD initiatives and systems to realise the expected outcomes (Kim et al, 2018; Al-Badi et al, 2018). The novel BDG framework proposed in this thesis study seeks not only to address core BD challenges confronting organisations, but to optimise strategic BD opportunities through effective governance and control. There is a requirement to establish big data governance frameworks in organisations for decision-making (Al-Badi et al, 2018; Kim et al, 2018), at both the strategic and operational levels.

During the course of our investigation, it was unveiled that the academic research and practitioner community alike utilised diverse terminology to describe conceptual schematics e.g. frameworks, models etc. We have formed the viewpoint that our proposed framework can be thought of as a conceptual framework consistent with the contentions of earlier authors, that such frameworks outline key constructs and relationships between them. It is a suitable method that introduces a set of handrails to guide organisations, practitioners and researchers to communicate big data governance concepts to either a layman or subject matter expert in a comprehensible way. This is crucial, as from our literature review and data analysis stages it is evident there is a broad range of interconnected BD challenges or risk factors that are influenced by governance, and which fall under the big data governance scope. Big data governance is not just about BD technology in isolation, but also about other business and social related domains such as ethics, finance, management, regulatory compliance, processes, etc. A framework is seen as a specification of big data governance as a concept with interrelated component levels that can strategically be utilised for governing

BD as an organisational asset. This thesis was based on the premise that when studied as a group of components, the representational framework contributes an accessible body of knowledge on big data governance and provide holistic insights derived from the research investigation.

1.4 Research Questions

The following research questions will be addressed to achieve the overall research objective.

What are the design elements of an effective Big Data governance framework, to manage big data as an asset, so as to optimise processes in such a way as to maximize the leveraged value?

In order to answer the main research question, the following sub-research questions are formulated:

1) What are the core big data complexities that organisations need to address?
2) What are the existing data governance frameworks to manage big data as an asset and how useful are they?
3) What are the elements of a big data governance framework?
4) What type of framework can guide big data governance?

1.5 Contribution to Knowledge

Evidence in the research literature and subsequent data analysis supports the opinions that the lack of attention to big data governance, leads to substantial challenges measured in both economic and human terms, either of which creates negative organisational outcomes. Nonetheless, there is limited research focus on the BD challenges and risk factors that organisations need to address through effective governance. In addition, organisations reportedly struggle to incorporate big data governance into their existing structures and business models to derive value from BD initiatives. This thesis research provides an original contribution to knowledge related to this area in several ways.

- The main contribution is the design of a novel big data governance framework to support organisations to execute an effective big data governance programme for improved outcomes with their BD initiatives.

- It uncovers those challenges posed by BD complexities extends to the ethical side of big data governance and increases focus on areas of ethical relevance as a key component of big data governance.

- A novel 'ethics by design' component is proposed within the overall big data governance framework to assist organisations in governing the ethical implications of BD.

- Critical analysis of independent parliamentary inquiry submissions undertaken as part of the thesis study, identify challenges and risk factors that may arise from the adoption of BD, proving a shift from old governance processes and expanding the limited research of the governance implications presented by BD.

  As far as the research conducted reveals, such analysis on this dataset has never been conducted for these research purposes before.

- The thesis provides evidence that a high level of complexity exists in the multi-dimensional aspects that BD challenges provide to DG. Identifying a large number of BD challenges, risks and vulnerabilities, that can be used by stakeholders within organisations as terms of reference in their strategic big data governance decision making. This can be utilised as a knowledge base to aid both professional and non-professional stakeholders to possess some ideas regarding the BD areas that require handling.

These original contributions are expected to be significant for both practitioners and researchers, justified by the limited research and struggling practices in big data governance. It will add new knowledge to IS science regarding big data governance. It will also simultaneously contribute knowledge to another intended audience of practitioners e.g. Chief Information Officers, Chief Data Officers, Business Executives, Chief Risk Officers; Data

Stewards, IT Managers etc., that are responsible for realising the benefits of BD initiatives, by providing an accessible and holistic big data governance framework to improve big data governance capabilities within their organisations. This will support their business goals of managing BD as an asset and aid to realise the desired business value from BD initiatives.

Researchers can benefit by applying the conceptual framework in the conduct of similar research in other organisational settings. Additionally, in the conduct of research extending the framework and examining different aspects of the framework in more specific strategic contexts.

1.6 Structure of the Thesis

The thesis is presented in 6 chapters and organised as follows:

- Chapter 1 introduces the background of the thesis study, emphasising the importance of big data governance approaches in addressing BD complexities, and noting the limited research in this area. The research aims, objectives and research questions are also considered in this chapter. Furthermore, it also highlights the main contributions to knowledge made by this research.

- Chapter 2 examines the relevant literature on big data governance, BD complexities and related areas, which are pertinent to the research problem identified. Based on a systematic review performed, existing DG frameworks are discussed. Crucially, the coverage and weak points of these previous frameworks are discussed to assist in providing context to the reader on the state of the art in research, as well as the research gaps identified.

- Chapter 3 provides an overview of the methodology employed for this research study. Firstly, it describes the research method utilised and also the theoretical basis underpinning the research approach. The methods of data collection and the way the parliamentary inquiry dataset for this research has been analysed are explained. Some limitations of this thesis study are also discussed here in this chapter.

- Chapter 4 discusses the inductive analysis findings and results. Based on these two chapters, the results of the grounded theory process of identifying BD challenges and risk factor concepts are discussed. Moreover, it establishes and discusses the criteria for big data governance which influence how these could be addressed as revealed by our investigation.

- Chapter 5 presents the proposed novel big data governance framework, its components and supporting approach to governance of BD complexities that affect organisations. It includes an ethical governance component in the proposed framework, by integrating concepts from the process.

- Chapter 6 finalises the thesis outlining the objectives achieved by answering the research questions and the core contributions delivered in this work. Lastly, it discusses the limitations of the thesis study and potential directions for future research.

1.7 Summary of Chapter One

Chapter one presents the background into the problem and motivation for this thesis study. It also provides an outline of the aims and objectives, the research questions, and the major contributions to knowledge. Lastly, the chapter provides an overview of the entire thesis structure. The next chapter will present the systematic literature review.

2.0 Literature Review

This chapter presents a review of the existing literature related to big data governance. The up-to-date literature review was conducted to assist in providing a theoretical understanding of the research landscape in big data governance, including its implications on governing big data as an organisational asset. Furthermore, it pinpoints some big data challenges and potential direction to address based on the current research literature. It considers the existing data governance frameworks from both academic and practitioner fields.

Critically, the research objectives drive the literature review methodology. The objectives focused on gaining an in-depth understanding of data governance frameworks and proposals that have been designed to manage data as an asset and examine if they meet the emerging needs of organisations facing BD challenges. The review set out to achieve the objectives of the following research questions in table 1:

Research Questions	**Motivation**
What are the core Big Data complexities that organisations need to address?	The goal is to discover from the literature which challenges, or risk factors related to Big Data need to be addressed by data governance.
What are the existing data governance frameworks to manage Big Data as an asset and how useful are they?	The aim is to identify the existing data governance frameworks, examine if they meet the needs of Big Data initiatives, and to identify the gaps in this area to support the main aim of the study. Additionally, this part of the review assists this study to identify what aspects of big data governance has been researched and what has not been studied.

Table 1: Research questions and motivations

These research questions were used to establish the content and structure, determine the primary studies, critically evaluate these studies and analyse their results.

The first section presents the concept of 'data as an asset' to set the context as to why big data governance is extremely important in terms of organisational value and competitive advantage today. In the next sections, literature examining the emergence of BD and its related challenges is then presented. This is deemed necessary to recognise the BD challenges and risk factors that need to be addressed, to establish a mode of introduction and a way to recognise the depth of the research problem. A systematic review which concentrated on identifying existing DG frameworks is also presented, examining whether they pay attention to addressing BD challenges or propose appropriate big data governance approaches.

2.1 Data as an Asset

Prior to exploring the nature and requirements for big data governance within organisations, it is important to understand the concept of data as an organisational asset. The Hawley Committee developed the concept of data as an asset in a report in 1994, which defined it as "data that is or should be documented and that has value or potential value" (Oppenheim et al, 2003). A straightforward illustration of data as an organisation's asset, includes companies that utilise customer data they store and analyse to grow revenue; by increasing cross sell / upsell rates and enhancing customer retention rates through targeted marketing campaigns and offers based on the profiling of historic buying patterns.

The quantities of data each organisation produces has grown exponentially and in recent years has shown no evidence of decreasing. Analysts have predicted that the large quantity of data generated globally is anticipated to reach 180 ZB in 2025 (Economist, 2017), hence providing BD with a dominant role in shaping the change and growth of markets and enterprises in the digital era of the 21st century (Vassakis et al, 2018; Ghasemaghaei, 2021). Moreover, the emergence of innovative technology such as BD analytics systems, customer relationship management systems and social networking platforms, have led to novel information exchange patterns for companies, customers, vendors as well as competitors. An earlier study had revealed that the market for BD technology grew considerably at about 23% per year, reaching $48 billion in 2019 (Grover et al, 2018). Furthermore, it has increased the amounts of data collection points. This combination of improved data sources with flexible integration, the emergence of business intelligence tools and BD analytics capabilities has

been a major driver in the development of data driven decision making. In 2012, a study highlighted those organisations using the BD generated in decision-making processes reported improved overall productivity by 5-6%, as well as higher levels of profitability, improved financial performance and increased market share (McAfee et al, 2012). Notwithstanding there are barriers to leveraging this data as identified during the literature review.

There is a broad consensus among researchers that managed correctly, data can become an organisation's most valuable assets, helping it to remain competitive, meet customer requirements and ensure costs are optimised (Ghasemaghaei et al, 2021; Brinch et al, 2021; Otto 2011a; Khatri et al, 2010; Soares, 2012). Recognition of data as a corporate asset, imply that some form of DG would be beneficial for effective data management (Abraham et al, 2019; Cheong et al, 2007). The structures within modern organisations are frequently distributed in nature. Making it important to focus on key organisation BD being available, accurate and properly integrated (Janssen et al, 2020). Organisations from various industries are striving to manage data as an enterprise asset, to be shared and reused throughout the company, across business processes and multiple software systems (Mao et al, 2021; Manyika et al, 2011; Panian, 2010). Data is ultimately a reusable asset if managed correctly. However, the value of data is determined by its use and for it to be fit for purpose, it must maintain its quality and meet other criteria. According to Otto, the formal objective of DG is to increase the value of data assets (Otto, 2011a). He argues that DG manages data assets and maximises its value through quality control. The diagram in figure 1 shows Otto's representation of his view on DG and its related concepts.

Others confirm the views of Otto (2011a), but also add that due to its value it needs to be governed and its quality also managed, monitored and secured (Khatri et al, 2010; Were et al 2017; Rifaie et al, 2009). Additionally, Wende states that DG ensures the relevant high-quality information products are supplied to the user (Wende 2007).

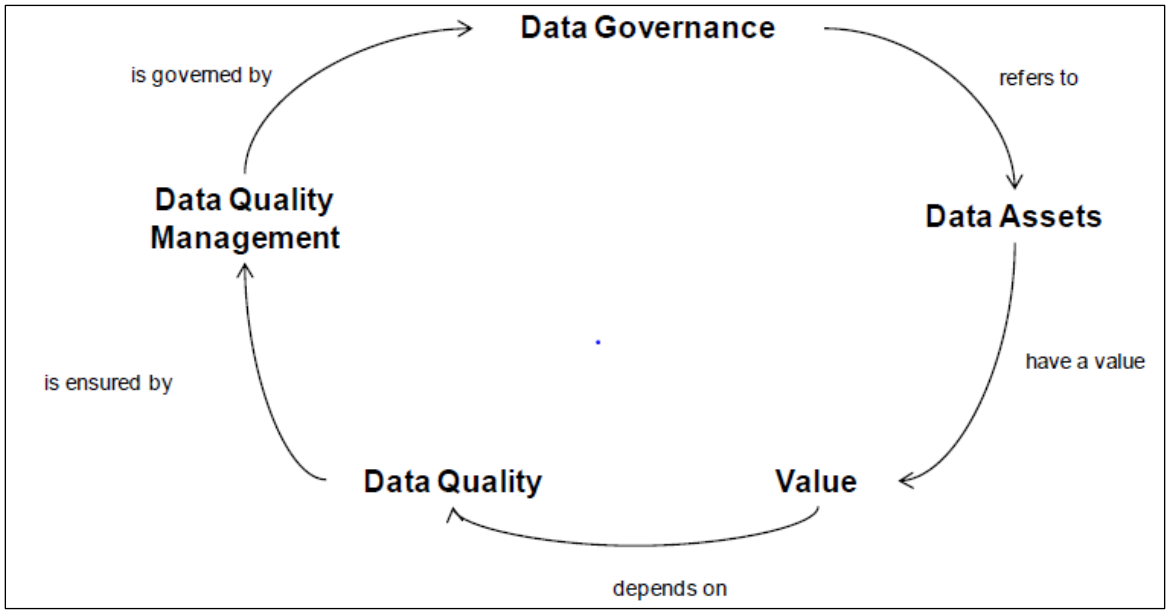

Figure 1: Data Governance and Related concepts (Adapted from Otto, B. 2011a)

It is contended by various studies that DG increases the value of data as an asset by supporting the factors that ensure the use of it increases within the organisation. Data's value is not currently represented on firm's balance sheets; however, research has revealed that it is increasingly considered to be one of the most important assets. The review conducted identified, that to increase the value of data as a business asset and obtain the expected competitive advantage, companies and government organisations need to establish standards, policies, processes for usage, development and management of data to create the right organisational structure and to develop the supporting technology structure. Here, big data governance emerges as a structured approach to handle some of the complexities with BD.

2.2 The Advent of Big Data: A further Imperative for Data Governance

The findings from the systematic literature review unveils evidence that a considerable amount has been written by researchers and practitioners on the BD paradigm. Majority of the research papers reviewed were analytical in nature. However, irrespective of the research approach most papers were aligned in the view that when managed, processed and analysed appropriately, it has the potential to offer colossal opportunities (Brinch et al, 2021; Albadi et al, 2018; Gunasekaran et al, 2017, Philip Chen et al, 2014; Motta et al, 2016). The digital domain has been leveraged by organisations in recent years for its ubiquitous, pervasive,

monitoring, tracking and user information gathering capacity. Since its emergence, companies have been exploiting its large scale, fast moving, complex streams of datasets that possess the potential to transform the way organisations conduct decision-making (Janssen et al, 2020), find otherwise hidden insights and obtain competitive advantage over their competitors (Brinch et al, 2021; Sestino et al, 2020; Raguseo et al, 2018; Janssen et al, 2016). The concept of BD has also become a driver for innovation. Notwithstanding, it was also evident from the review undertaken that the challenges and requirements for managing and processing BD are significant, and these need to be addressed to obtain benefits from its use.

Gartner conceptualises BD as "high-volume, high velocity and high variety data assets that require novel forms of cost-effective processing to enable improved decision making, insight discovery and process optimisation" (Gartner, 2012). BD is closely related to BD analytics, which is needed to create value from data (Brinch et al, 2021; Holsapple et al, 2014). Senior business leaders globally are flooded with data, characterised by its variations and inundated by its sheer volumes which offer these potential insights through BD Analytics (Aversa et al, 2021; Aversa et al, 2018; Riggins et al, 2017). Some authors classify the different types of BD in their work. Firstly, examples of BD consist of web and social media generated data (Soares, 2012; Brous et al, 2016). The second category is classified as machine-to-machine data (Soares, 2012; Brous et al, 2016). Thirdly there is biometric data (Malik, 2013; Soares, 2012. Fourth, there is big transaction data and finally there is human generated data (Soares, 2012; Malik 2013). Irrespective of the type of BD, it is perceived by some researchers to be extremely valuable; aiding to foster productivity growth, enhance the decision-making process and drive new revenue streams across various industry sectors (Maniyika et al, 2011). For example, mobile phones that have been designed to leverage BD to learn the owner's usage practices and preferences, or that contains applications tailored to user requirements, is deemed to be more valuable than a mobile phone that has not been customised to a user's requirements. In addition, enterprises can design products that support customer requirements better, by leveraging BD. BD usage can lead to improved health outcomes, enhance the quality of community engagement with government, or promote reduction in prices due to pricing transparency. Consequentially, customers and citizens are direct and indirect beneficiaries of BD innovation, also benefitting from a large amount of the economic surplus that BD facilitates.

However, whilst some studies focused on highlighting the undeniable operational and strategic opportunities BD provides, a closer examination of the literature during the review revealed that working with BD raises new challenges related to its capture, storage, analysis and visualisation. Moreover' some of these studies contend that if all sectors are to harness the value and opportunities BD provides, there is an accompanying requirement to develop robust data driven strategies and development of the right capabilities to derive the expected value. With BD comes subsequent changes in the way data is generated, collected, maintained and utilised. It comes from a variety of sources and structures including, social network feeds, audio and video links, sensor data, email etc. These data types are in different formats 'structured' or 'unstructured' and are created continuously by people both internal and external to the organisation. Against this backdrop, both Soares (2012) and Kim et al (2018) argue that in order to leverage BD and effectively prepare for the transition to BD, organisations adhering to established DG frameworks are positioned to achieve greater short-term or long-term success (Soares, 2012; Kim et al, 2018). An inadequate DG framework will cause issues such as irregular data standards, insufficient data quality and low management effectiveness. Thus, DG frameworks are becoming more vital in public and private organisations to manage data as an asset (Mao et al, 2022; Katri et al, 2010). Echoing both Kim et al and Soares, other researchers agree with this view and further add that big data governance is a matter of crucial significance for organisation's that rely on BD to drive value (Li et al, 2019; Tallon, 2013; Al-badi et al, 2018).

To support the outcomes of this review, the following section provides a presentation of some of the BD challenges discussed in the existing literature and which big data governance frameworks are needed to address. Crucially some of the problems identified have yet to be solved holistically by any framework and there are outstanding open challenges and implications for this thesis research.

2.3 An Investigation of the Big Data Complexities

Global operating businesses confront increased demands with regards to data capture, processing, integration, storage and analysis for their data platforms and infrastructures when implementing BD. In this study, BD is viewed in the context of leveraging data as an asset

and the subsequent requirement for businesses to derive greater amounts of information from large volumes of data than ever before, from multiple and often new sources, consumable upon creation (Ghasemaghaei, 2021; Arista et al, 2016; Baesens, 2014). Along with today's data deluge "a new scientific paradigm has emerged called data-intensive scientific discovery, also known as BD problems" (Chen et al, 2014). Some researchers have questioned the traditional approaches to managing data as companies have increasingly started to embrace BD. Its high velocity, voluminous and multi-structured characteristics from a proliferation of disparate data sources, in a variety of formats have raised new difficulties on how best to manage and govern the data (Li et al, 2019; Al-badi et al, 2018; Kim et al; 2018; Sivarajah et al, 2017; Soares, 2012). BD can be acquired from both inside and outside the organisation, in the form of structured and unstructured data, and it can create problems for organisations attempting to capture, aggregate, analyse and extract value from it (Raguseo, 2018)

Whilst conducting the search for relevant papers to be included in the review, there appeared to be an increase in the number of studies discussing BD and its challenges from 2012 onwards. In fact, the searches conducted in Scopus using the defined search terms discussed later in chapter 3, significantly did not display any results prior to 2012. This highlights the traction and rise in importance of this focus area among practitioners, governments and the research community globally from this period. The vast majority of papers reviewed outline substantial challenges related to BD. A simplified view of some the examples of BD challenges revealed in the literature are; Increase in data size impacting operational cost and the capabilities of computation (Philip Chen et al, 2016), technical and operational issues in value creation (Sivarajah et al, 2017), the problems establishing appropriate BD analysis and visualisation rules from reliability of source data (Vassakis et al, 2018; Mithas et al, 2013), the complications associated with querying and analysing large volumes of data as well as inadequate data warehousing and infrastructure (Mao et al, 2022; Rifaie, 2009; Kune et al, 2016), fragmented architecture and legacy systems, leading to the lack of BD availability (Nielsen et al, 2018) issues of privacy and compliance (Cui, 2019; Saxby, 2014), ethical issues (Ho et al, 2020, Novak et al, 2020, Cui, 2019; Tallon, 2013), security and integrity issues (Kim et al, 2018; Yallop et al, 2020; Cervone, 2016), problems arising from flaws, inaccuracies and incompleteness of the data itself (Kim et al, 2018; Sukarmar et al, 2015), challenges with data integration (Mao et al, 2021; Sivarajah, 2016), processing barriers to

ensuring the correct BD is available in the correct format when and where required (Sanders, 2016) , skills and capability gaps (Raguseo, , 2019; Vassakis et al, 2018; Whyte et al, 2016), legislative and regulatory compliance challenges (Favaretto et al, 2020; Kemp, 2014).

Interestingly, because traditional computational systems are inadequate for processing and analysing BD, some of the papers reviewed described BD problems in strict relation to the technical challenges that arise. Problems related to the volume and velocity of BD, for instance, were contented to produce several issues to traditional IT structures given that organisations have legacy systems and infrastructure that are unsuitable for collecting, storing and processing the large quantities of heterogeneous data that is created at increasingly higher velocity. Mao et al (2021) stated that with the growing quantities of user data, the scope of businesses continues to extend. Arguing that nonetheless, the traditional front- and back-end systems of some organisations are overloaded, unable to match the operating speeds required (Mao et al, 2021). Some of these scholars argued that given BD technologies are dealing with large amounts of high-speed heterogeneous data from different sources, this increases the probability of the data being inaccurate or untrustworthy, generating problems related to the veracity of BD (Favaretto et al. 2019, Rehman et al, 2021).

To a greater extent, people now reside in a networked, with numerous data sources and constant proliferation of BD e.g. smart mobile phones, social networking sites, video sharing, online messaging, online purchasing, online financial transactions etc. Over the past few years many global operating companies have embarked on medium to large scale, multi-year next generation digital transformation projects to derive value from BD, at great cost to the company and with varying degrees of success (Ghasemaghaei, 2021; Grover et al, 2018; Manyika et al, 2011). According to recent surveys however, 80–87% of the BD projects never generate a sustainable solution (Escobar et al, 2021). Consultancies and third-party vendors are making huge profits providing businesses with BD consultancy services in all areas of the BD domain, which inadvertently drives up project and adoption costs. The interest of numerous organisations on how to utilise BD and new data sources e.g. mobile and social etc., is spurred by opportunities to make expeditious and enhanced decisions via sophisticated analytics. For example, a major aspect of social media for marketing management decision-making is perhaps the access it provides to the opinions, ideas and preferences of either existing or potential customers. Marketers can use social media sites for direct interaction with potential customers via social and mobile media, using ‘engaging’

social media campaigns and 'viral' social media content to gain significant intelligence to inform decisions. Leading some researchers to highlight the negative problems this causes society, including some of the inherent biases built into BD systems algorithms (Micheli et al, 2020; Kitchen, 2017). Some contending that the consequential discriminatory practices of BD mining can for the most part be traced to human bias and shortcomings of the law (Favaretto et al, 2019). Owing to the increased complexities associated with these type of use cases, businesses are now confronting new ethical, legal and regulatory challenges with big data governance. To illustrate, Europe introduced major changes to the General Data Protection Regulation (GDPR) on 25th May 2018. This ushered in tighter controls on the use of customer data albeit with limitations (Delacroix et al, 2019). Figure 2 below presents some of the tighter controls introduced. For instance, penalties for breach of these regulations could result in penalty fines of up to €20 million or 4% of annual turnover, whichever is greater.

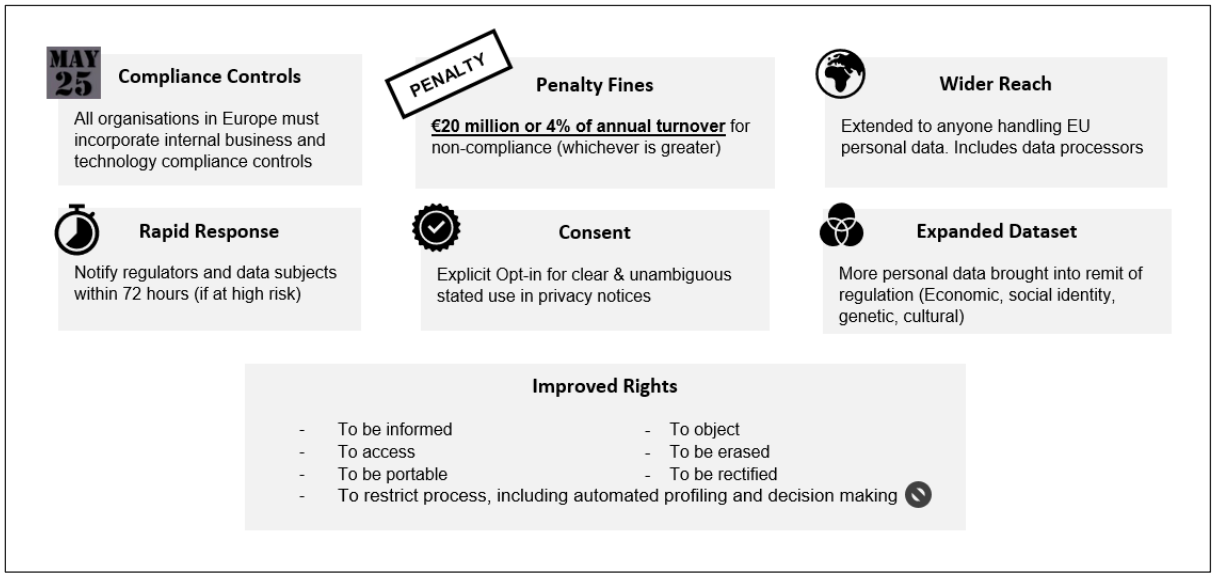

Figure 2: GDPR (General Data Protection Regulation) - Changes to Regulation

Further additional measures were communicated by the European Union in 2020, including a Digital Services Act and a Data Act (European Commission), introducing further considerable compliance challenges to organisations. It is argued that although GDPR introduces tighter regulations around the storage, management and transfer of PII data. Notwithstanding, BD collection and usage could be impacted by data regulations that differ from country to country, adding further complexity to govern the data for both local and

international operating organisations. Furthermore, these regulatory requirements increase pressure on organisations to have a robust handle on what, where and how the BD is being used and stored. A robust big data governance program is critical to the data visibility and categorisation required for the differing compliance rules. Ultimately, big data governance can be leveraged to enable and facilitate the goals of GDPR but has further far-reaching purposes. Including the governance of non-personal BD.

The problems with BD worsen when you investigate the specific areas of data security that come under the umbrella of governance risk and compliance (Mansfield-Devine, 2017). Socio-ethical and privacy issues also emerge with the increasing analysis of social networking data by organisations (Sestino et al, 2020; Cui, 2019; Tallon, 2013; Soares, 2012). These ethical issues are unveiled as one of the core BD problems that must be addressed. Legal rights in relation to BD present a further complex picture (Cui, 2019' Kemp, 2014). In his work, Kemp argues that intellectual property rights, contract and regulation are discrete norms which possess their own technical and often inconsistent rules. With ever increasing data volumes, variety and velocity the challenge to manage and protect the data is becoming progressively more difficult.

In the further context of BD, another significant challenge is the management effort required at strategic and operational level (Kache et al, 2017; Sivarajah et al, 2017; Praseyto, 2015). There are challenges in master data management e.g. How to structure unstructured data and ensure data quality, which is made more complex when handling data from disparate data sources. Literature reviewed during this research, has also revealed organisations are confronting issues in implementing the holistic policies, processes and standards in the collection, storage, processing, searching, transferring and sharing of BD (Jin et al, 2015; Philip Chen et al, 2014, Kim et al; 2018). BD brings new issues to governance since much of the data that organisations would want to leverage comes from external sources outside of their control. Additionally, it is less structured and less understood than traditional transactional data. Problems with integrating BD with traditional company data were also cited in the literature (Malik, 2013; Mao et al, 2022; Sivarajah, 2016), as some companies have evolved from being product centric to customer centric to gain a 360 view of the customer. The insights, products and services that emerge from the analysis & processing of this data is where the value lies and what drives the competitive advantage (Aversa et al, 2021; Raguseo, 2018).

Kim et al (2018) examined some of the BD issues and risks that need to be eliminated to support BD implementation in an NPS case analysis in Korea. These researchers call out the problems of personal information leakage and privacy infringement, concerns about data monopoly as well as data quality and service. Organisations face the problems of ascertaining if it is accurate or misleading. This aligns with the concerns raised that inaccurate information exists on the internet and has been circulated for long periods of time. Establishing the time value of data and determining if the data can be trusted or fit to be utilised for decision making is cited as a huge problem (Bizer, et al 2012; Sukarmar et al, 2015; Malik, 2013).

Additionally, data quality is critical for businesses to meet a variety of business needs such as compliance with legal or regulatory provisions, integrated customer management, efficient reporting, or automated business processes (Otto, 2011c; Wende, 2007; Khatri et al, 2010). Studies suggests that without DG, indicators of poor data quality become apparent e.g. inconsistent data, redundant data, non-availability and less accountability (Otto, 2011b, Cheong et al, 2007; Zhang, 2020). BD in its raw form, holds no value, until it is processed and analysed for a specific goal. This means that in its raw form it cannot be validated until it is processed, which could potentially be too late. Consequentially, BD is not always as trustworthy as traditional data, hence governance is important (Soares, 2012). In the era of BD, Cohn (2014) emphasises that "whilst data is a catalyst for innovation, DG is a catalyst for quality and value is derived from well governed quality data". If companies are to succeed in the quest to, obtain a return on investment from BD initiatives which are geared to exploit data to derive business value and competitive advantage, they will also need to address issues in big data governance at the strategic, tactical and operational levels. According to Davenport, there is a requirement for data strategy or DG to ensure critical data and other resources are managed (Davenport, 2007). Furthermore, it is argued that it requires a holistic process view which encompasses the sourcing, preparation, storage, analysis, access and usage of BD (Phillips-Wren, 2015; Holsapple et al, 2014).

Watson (2004) highlights that DG is integral to any business. The absence of DG is a serious problem that is considered to require careful consideration to minimise the risk of failed software implementation, safeguard information and realise the concept of data as an asset (Khatri et al, 2010). Nevertheless, research reveals that a core challenge for organisations is to develop an appropriate framework for how BD, analytics, frontline tools and people come together to create business value, particularly for BD initiatives (Li et al, 2019; Kim et al,

2018; Al-Badi et al, 2018; Biesdorf et al, 2013). Both the scientific and practitioner community state that the scope, diversity and integration associated with the governance of BD is really challenging.

All of the above BD challenges have wider implications when it comes to governance. DG fundamentally aims at creating a competitive advantage for businesses by fostering a holistic approach to vital company data. One researcher defines it as "a companywide framework for assigning decision-related rights and duties in order to be able to adequately handle data as a company asset" (Otto, 2011b). Researchers cite that companies require enterprise-wide data strategy and governance to achieve competitive advantage or to derive value. The key differentiation between the terms 'Governance' and 'Management' is primarily that governance refers to the decisions that must be taken and who makes these decisions in order to ensure effective management and use of resources, in contrast management involves executing decisions (Khatri el al, 2010). Consequently, management is influenced by governance (Otto, 2011a).

Big data governance mechanisms that are fit for purpose to ensure ethical, regulatory and legal compliance, as well as support effective control of BD is considered crucial.

2.4 State of the Art in Data Governance Research

Research shows that DG research took root in 1980s, it is still a relatively new research area without a commonly accepted definition (Abraham et al, 2019; Otto, 2011c, Wende, 2007); as a result, it is argued in this research thesis that it requires further definition and validation. Some attempts at data governance definitions were discovered in both the practitioner community and scientific research papers; hence to provide some research direction some of the definitions found are presented in table 2 below.

No.	Author(s) (Year)	Data Governance Definition (s)	Domain
1	Cohen, 2006	"The process by which a company manages the quantity, consistency, usability, security and availability of data".	Scientific Research

2	Newman & Logan, 2006	"The compilation of decision rights, processes, standards, policies and technologies required to manage, maintain and exploit information as an enterprise resource".	Practitioner
3	Thomas, 2006	"A strategy that "refers to the organisational bodies, rules, decision rights and accountabilities of people and information systems as they perform information-related processes. Data governance sets the rules of engagement that management will follow as the organisation uses data".	Practitioner
4	IBM Software Group, 2007	"A quality control discipline for adding new rigour and discipline to the process of managing, using, improving and protecting organisational information".	Practitioner
5	Karel, 2007	"A process by which an organisation formalises the judiciary duty for the management of data assets critical to its success".	Practitioner
6	Weber, Otto, & Osterle, 2009	"A framework for decision rights and accountabilities to promote desirable behaviour in the use of data".	Scientific Research
7	Khatri & Brown, 2010	"Data governance refers to who holds the decision rights and is held accountable for an organization's decision making about its data assets".	Scientific Research
8	The Data Governance Institute, 2012	"A system of decision rights and accountabilities for information related processes, executed according to agreed upon models, which describe	Practitioner

		who can take what actions, with what information, and when, under what circumstances, using what methods".	
9	Soares, 2012	"Policies relating to optimisation, privacy and monetisation of BD practice".	Practitioner
10	Abraham, Brocke & Schneider, 2019	"Policies and procedures adopted by organisations to manage data; it is a cross-functional framework for managing data as a strategic enterprise asset. Specifying decision rights and accountabilities for an organisation's decision-making about its data".	Scientific Research

Table 2: Outline of Several Data Governance Definitions

From initial review of the literature, there are differing views amongst researchers on how data governance should be handled and limited research showing how existing data governance frameworks address specific challenging aspects of BD. Wang et al originally prescribed that the information production process must be well defined, moreover it should consist of adequate controls such as quality assurance, delivery time-management and production, arguing that the information production process be handled similarly to producing a physical product for a customer (Wang et al, 1998). Other researchers and practitioners have positioned that the data governance approach should be handled differently. Weber et al (2009) adopted Weills 2004 definition of IT governance, aligning on the theory that it is a framework for organisations to outline decision rights and accountabilities which drive the behaviors required in the use of data. Similarly, Khatri and Brown (2010) aligned with this description as the table above shows. The aspects of assigning responsibilities pertaining to the management of data, is an aspect emphasised in the data governance models by some researchers (Cheong et al, 2007; Wende, 2007; Weber et al, 2009). There are further differing views in the practitioner community. IBM (2007) defines data governance as a quality control discipline and states that its goal is to add rigour and control to the process of managing, utilising, improving and safeguarding organisational information. By contrast, Gartner have a process view of data governance. Given its many influences and the differing definitions of data governance uncovered in the existing literature

– it starts to clarify why confusion proliferates when implementing data governance programs within industry. Further exasperated when BD is put into the mix.

Survey results published by TDWI based on responses from 394 participants, unveiled that over 61% stated that the 'lack of understanding of governance' is also a barrier to data governance in their organisation (Russom, 2008). It is contended in this research thesis that data governance maturity levels are relatively low across a variety of industries, hence the requirement to contribute towards improvement in this important area. According to the data shown from this earlier survey in Figure 3 below, 1.64 out of 5 is the average maturity score across industries in relation to data governance (With 0 being the lowest maturity level score an industry sector could be scored at and 5 being the highest possible score). The graph in figure 3 further reflects that the Insurance industry was the most mature sector in the practice of data governance with a relatively low score of 1.95 out of 5. It also indicates that the Distribution industry was found to be the least mature sector overall with an even lower score of 1.01 out of 5.

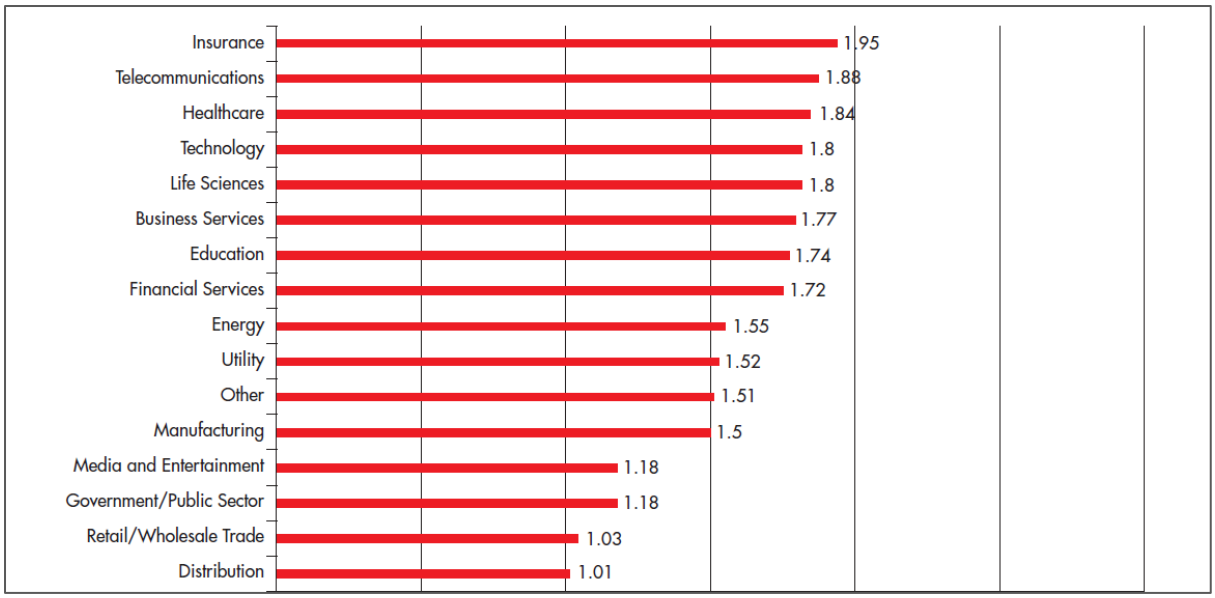

Figure 3: Room for Data Governance improvement across industries (Source: Informatica, 2015)

It is argued that these varying levels of maturity lead to complications in the practice of exploiting data assets (Nielsen et al, 2018). Findings from some researchers highlight that internally many organisations erroneously view data as a technological problem (Manyika et al, 2011, Fisher, 2009). This has led to huge investments in expensive multi-year BD projects

without delivering the promised business value and return on investment. These researchers maintained that this is also an executive problem that needs to be governed strategically to address BD challenges that block business success of the organisation. BD Scientific research about big data governance is limited (Kim et al, 2018) and during the review only a limited number of studies were found which have discussed this increasingly important area. Significantly, the first and only data governance definition which particularly relates to BD was that of Soares (Soares, 2012).

Furthermore, it was uncovered that a great deal of the papers reviewed discussed data governance in general and were descriptive in nature, and only a few papers included data governance frameworks. Following further investigation, it was uncovered that they appear to have limitations as none appear to take a holistic approach in attempting to address the BD challenges or risk factors cited in this thesis research. Moreover, there does not appear to be any overview of existing data governance frameworks and how they compare to address BD complexities that organisations are confronted with. Which this review serves to do.

Subsequently a total of 29 data governance frameworks were uncovered during the literature review and are referenced in the table 3 below.

Scientific publications (17 in total)	Cheong and Change (2007), Wende (2007), Khatri and Brown (2010), Webber and Otto (2007), Panian (2010), Tallon (2013), Otto (2011), Al-Ruithe (2016), Rosenbaum (2010), Prasetyo and Surendro (2015), Kim and Cho (2018), Zhang et al (2017), Al-Badi et al (2018), Abraham et al (2019), Yebenes et al (2019), Mao et al (2021), Li et al (2019)
Practice-oriented publications (12 in total)	Thomas (2006), IBM (2007), Radcliff (2007), TDWI (2010), Soares (2010), Soares et al. (2012), Soares (2013), Loshin (2013), Reeves and Bowen (2013), DAMA International (2014), Karel (2015), Informatica (2017)

Table 3: List of publications selected containing Data Governance Frameworks

The table in Appendix B provides a summarised description of each of the frameworks. More importantly, it accentuates that only 6 out 29 of these frameworks explicitly address big data governance complexity aspects. 5 of these big data governance frameworks are identified from the scientific research publications and 1 from the practice-oriented publications.

The earlier data governance framework of Wang et al (1998) was concentrated towards process management and coordination, mainly useful for the traditional IT function that performed information process management. Additionally, as this was years before the trend in BD research took off, the specific challenges particularly addressing aspects of BD were evidently lacking. After several years came the Data Governance Institute's (DGI) framework (Thomas, 2006), closely followed by Wende's scientific paper, which is one of the earliest scientific efforts to propose a data governance framework (Wende, 2007). The DGI's framework, identifies 10 universal components of data governance programs, split into three separate categories: namely, rules, rules of engagement, people and organisational bodies and processes. The framework's focus areas address policy, standards, strategy, data quality, architecture/integration, management support, data warehouse and business intelligence. In comparison Wende proposed a data governance model, comprising of three components concentrating on data quality management. It adheres to the school of thought that combines technical as well as business driven perspectives to deal with strategic and operational challenges of quality management. The main contribution was to propose a flexible data governance model, with a framework comprising of roles, responsibilities and decision areas. Identifying a set of data quality roles, by proposing a RACI chart. However, in both these approaches all aspects of big data governance have not been addressed. This inadvertently could leave businesses with some gaps and open questions, which leaves certain aspects open to inaccurate interpretation.

Cheong et al introduced a data governance framework in 2007 (Cheong et al, 2007), building on the earlier studies of Thomas in 2006, where it states the lack of authority and clear roles, and responsibilities are contributors to failure via stewardship (Thomas, 2006). They take their findings from a case study of a large utility company in Australia and propose a data governance structure and framework that focuses on data management issues e.g. reactive data management, data assets not easily accessible, problems aligning on data standards and data improvement projects that are overrun and over budget - emphasising the needs for collaboration between the business and IT departments to support organisations. Their research highlights the importance of data governance structures together with policies and procedures for managing data effectively but provides no suggestions on how this can be done. However, they also go on to identify some critical success factors for data governance. Wende and Cheong et al, refer to business intelligence (BI) as one of the major operation

business domains that Data Governance can have great impact providing quality data to users, but apart from that no further information or specific guidelines are provided on how to better support BI or BD analytic systems.

It is explicitly called out in the literature that Gartner's seven building blocks framework was utilised by practitioners for years, nonetheless its core focus is on master data management (Radcliffe, 2007). According to the Gartner, most challenges to do with Master Data Management (MDM) programs are not technical, but governance related and cites that the creation of an MDM Program is important for MDM program success within an organisation. Hence, Gartner introduced MDM governance as one of the building blocks of the overall MDM framework.

The IBM Software Group developed the data governance maturity model, designed with input from 55 organisations (IBM software group, 2007). Its objective was to build consistency and quality control in governance. It positioned that effective data governance can enhance data quality, availability and integrity of company's data, by directly impacting four factors it found that organisations were concerned about i.e. increasing revenue, lowering cost, reducing risk and increasing data governance. The model measures data governance competencies of companies based on 11 domains considered to be crucial to data governance maturity. This model is identified as a maturity framework and a set of assessment questions with answers constructed to indicate maturity levels. However, in this model, how the organisation adopts these tools depend on various factors and it literally focuses on the 'What?' and 'Who?' compared to the 'How?' It argues that each company is unique, possessing its own goals and methods, based on different business scenarios, plans and requirements. However, to avoid challenges that could arise with adoption, it will be argued within the thesis research that all frameworks should provide supportive recommendations or guidance on how data governance can be linked and made accessible regardless of maturity levels. Furthermore, it is apparent that IBM's perspective is that of a software vendor, as its company establishes that its suggested data governance frameworks would require software support (IBM, 2007; Soares, 2010; Soares et al, 2012). Additionally, IBMs model reveals that a company's requirements should make sure that both business and IT issues are clearly outlined. However, it is argued that in turn there is a requirement that data governance frameworks must clearly highlight problems they aid to address. Specifically, with regards to BD.

Weber et al and Khatri et al also released frameworks that were important contributions in 2009 and 2010 respectively (Weber et al, 2009; Khatri et al, 2010). Weber et al, extend the previous data governance research by Wende by suggesting a framework that addresses accountability and decision rights for corporate wide data quality management. Nonetheless, they build upon this and argue that organisation structures are not a one size fits all and promote a contingency approach to data governance. Each company requires a specific data governance configuration that fits a set of contingencies. The research cites these contingencies as firm size, structure and decision-making style and argues that these parameters could impact the model. The contingency approach they developed, proposed that each business should have a specific data governance configuration that fits a set of context factors (Weber et al, 2009). Utilising the Weill & Ross IT governance framework as a starting point for their data governance framework, Khatri et al (2010), presented a framework that proposes five interrelated data decision domains and guidelines highlighting what governance is required for each. The five decision domains are data principles, data quality, access, meta-data, and data lifecycle. Their research emphasises that businesses should establish data principles to understand what policies, standards and guidelines are appropriate, arguing that by first realising data as an asset, the data becomes subject to company controls, standards and policies. What is found useful with this framework, is the consideration given to the necessity for data life cycle decision points. This is deemed by this research to be a vital component for the success of a big data governance and is justified in later sections of the thesis study. Both Weber et al and Khatri et al's research on data governance propose the assumption that actualisation of the governance design options is contingent on context (Khatri et al, 2010; Weber et al 2009), nonetheless none of the frameworks satisfactorily address the BD context holistically.

Another major contributor to the studies on data governance is Otto (Otto, 2011c). In 2011, this researcher created a more comprehensive framework and argued that data governance consists of goals and structures. The goals and structures focus on control, organisational form and roles and committees. Otto asserted in his paper that organisation of data governance programmes had not considered all the aspects required for these to be considered complete. A paper he published aimed to develop a morphology of data governance organisation to highlight this (Otto, 2011b). Several mini-case studies were used to effectively assess the morphology of the organisation of data governance. Niemi (2013)

also observed that all existing sources for data governance from researchers and practitioners at the time his paper was published, focused on placement of decision- making authority for data quality management and organisational structuring (Niemi, 2013). Both Otto and Niemi effectively aligning on the argument that all data governance frameworks designed up to 2011, were incomplete.

The industry association DAMA International published the Data Management framework DM-BOK in 2010, with a second edition DM-BOK2 published in 2014. In summary, it presents a standard industry view of data management knowledge areas. It attempts to outline some of the 'who', 'what' and 'why' of data management, but not the 'how to' methods. It defines 11 core knowledge areas for performing data management, with data governance being one of them (see figure 4). It also identifies strategies for data management maturity analysis. Within its framework it defines data governance "as the planning, oversight and control over management of data and the use of data and data related resources" (Cupoli et al, 2014). DAMA-DMBOK2 demonstrates a technical direction and business components appear lacking in the design.

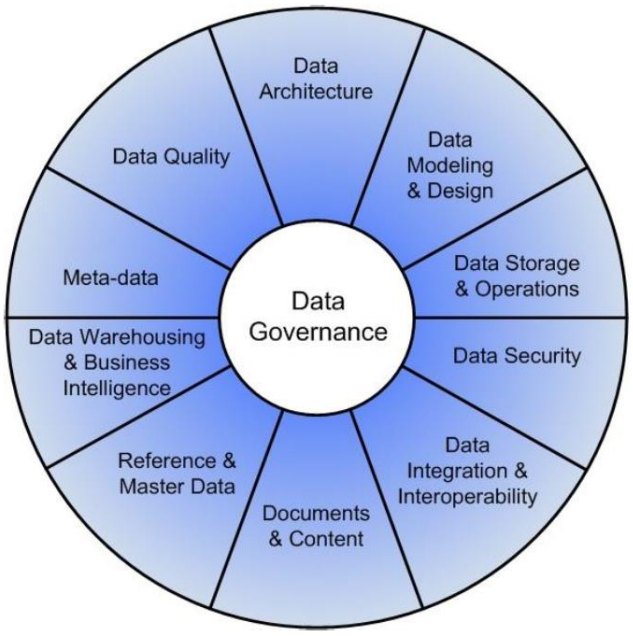

Figure 4: The Data Management Knowledge Area Wheel (DAMA-DMBOK2 Framework, 2014)

Panian's work identifying data as a source of risk, suggests six core data attributes that are usually associated with digital challenges those modern organisations experience i.e.

accessibility, availability, quality, consistency, auditability and security. However, both Panian and DAMA-DMBOK 2 demonstrate a more technical orientation and business value components are not adequately represented in the framework design. According to Prasetyo and Surendro (2015), like IT governance, data governance needs to align with an organisations business strategy (Prasetyo et al, 2015).

Upon further examination of the data governance models that have emerged from the practitioners community, a couple of observations were made. The DAMA International's model designed by industry associations demands the completion of all artifacts in a matrix that could be perceived to contain too many elements that must be built in data governance in an organisation. Whilst the data governance models developed by the DGI consulting organization and TDWI necessitates the development of data governance is formed of several related items (including programs, stages, decision domain, universal objects, and components) that require significant effort to learn and use. IBM's different models for data governance are built from a software provider perspective, so to establish data governance will ultimately require software support. Most of the practitioner models demonstrate a high-level nature with abstract connections between data governance components. Another foreseen challenge is that the exact content of each component, potentially makes it harder for data stakeholders in an organisation to identify their objectives due to different interpretations. This could prove to make the models less accessible to a lot of companies, requiring extensive budgets to pay for consultancy. In the case of this research, it is argued that organisations need a model that can be easily adopted in accordance with the requirements and resources of the organisation.

As stated, six existing publications related to big data governance frameworks were identified (Tallon, 2013; Kim et al, 2018; Zhang et al, 2017; Al-Badi et al, 2018; Soares, 2012; Li et al, 2019). Although the frameworks in these publications contribute to knowledge, nonetheless there are some limitations. The frameworks either address big data governance concentrating on particular dimensions of big data complexity aspects e.g. BD Quality, BD Security, Privacy infringement and managing cost (Kim et al, 2018; Tallon, 2013) or comprise smaller industry sector views more concerned with activities associated to BD use in the sector (Li et al, 2019). Tallon (2013) proposes a range of structural, procedural and relational practices, that are aimed at maintaining a balance between value creation and the risk exposure for

companies that leverage BD. The paper asserted that policies and structures should protect data from errant factors that could destroy or limit its value. Also outlining several inhibitors and enablers that can impact an organisation's ability to design and execute the data governance practices managing BD risk, cost and value. Moreover, the study provided an outline of positive and negative consequences of data governance on company performance. Notwithstanding, the author does not provide a method to establish and assess data governance, albeit the author does supply some understanding into the precursors for, constitution of and significance of big data governance. Zhang et al (2017) conduct research which produces an actor network of big data governance using the translation process of Actor-Network Theory and scrutinising the interconnection between the human factors and non-human factors (Zhang et al, 2017). Notwithstanding, these authors and others do not describe the antecedents or consequences of BD challenges & risks, which is a necessity to comprehend the factors that instigate the adoption of big data governance frameworks and the effects of those frameworks (Zhang et al, 2017; Li et al, 2019). Only Kim et al appear to specifically go into any detail highlighting the BD issues and risks to be addressed by big data governance frameworks, however in our opinion they do not investigate nor validate this area extensively (Kim et al, 2018). They use a framework developed by Panian (2010) as a basis to adapt a framework for The National Pension Service (NPS) of South Korea. Although they use the NPS as a case study, again they do not elaborate in detail or provide a full description of the methodology used to develop the framework further. After the commencement of this thesis study, Al-badi et al (2018) also made attempts at developing a big data governance framework that consists of 8 components. Upon evaluation of the paper there is limited explanation on the methodology utilised. Following a literature review where they uncover some articles discussing big data governance frameworks, they analyse and interpret these, but do not discuss in detail how the criteria for the framework was ascertained. They also provide insufficient evidence of detailed coverage of the various critical elements in the big data governance domain, even though their research clearly evidences the necessity to focus on emerging requirements exposed by the BD domain. Additionally, the resource omits to supply a clear framework for decision-making, leaving the large number of problems involved liable to overwhelm users. Karel (2007), discusses the use of a data governance framework to enable BD. However, this framework is not classified in this thesis study as a big data governance framework as it proposes using traditional data governance processes to enable big data initiatives, which is contrary to the position taken in this thesis study as discussed in section 2.6. Likewise, Abraham et al (2019) recognises and

contextualises some of the different implications related to DG when it comes to BD e.g. DG with a lense on master data quality could be distinct from DG looking at privacy in the context of BD. It also recognises that DG should address BD risk without hampering innovation. However, there conceptual framework strongly seems to focus their concepts on data governance in general, albeit it makes a distinction between traditional data and big data in the framework by splitting the content dimensions of Tiwana et al (2014).

It is worth highlighting that a high number of papers examined during the literature review, did not focus any attention on addressing the ever-increasing ethical tensions posed by some BD types or their associated analytical practices e.g. social media data. To overcome this major deficiency, we proffer a solid proposal. Finally, a significant number of the papers uncovered do not outline the methodology in which these frameworks were constructed. Only 1 out of the 12 practitioner publications described their methods for establishing their frameworks as reflected in the table in appendix B. Soares out of all the practitioners has done the most to build a case for big data governance, however his methodology for developing his recommended approaches were never revealed. In the scientific research studies although methodologies were specified, only 6 out of the 19 went into full detail describing their varied research methods.

2.5 Big Data Governance in Organisations

If the standpoint of several researchers and practitioners is to be accepted, that key assets of today's organisations lie in its data. Then we must also recognise the viewpoint of others who contend that harnessing these BD assets requires a relentless focus on the data being fit for purpose and having adequate big data governance that supports this. According to the outcomes of the literature review, there is an explicit need for organisations to take a holistic approach to big data governance. Millions have been invested by organisations on enterprise systems in the push to become more data driven. With shrinking budgets and increased global competition, it becomes imperative to obtain a return on investment.

Earlier studies have asserted that a framework can play an important role in assisting modern organisation to effectively plan and allocate resources to enhance data driven strategies (Gorry et al, 1971). Sprague suggests that it can assist an organisation to identify components and relationships between parts to understand an otherwise complex system

(Sprague, 1980). Researchers in the data governance research community contend that a data governance model helps organisations to structure and document the accountability for data quality (Wende; 2007; Otto; 2011c). Also stating that an effective data governance framework can help organisations to create a clear mission, achieve clarity, increase confidence in using organisational data, establish accountability, maintain scope and focus, and define measurable success criteria (Abraham et al, 2019; Janssen et al, 2020; Fu et al, 2011). Moreover, a good data governance framework supports compliance and legal efforts over the long term, as well as build better relationships with customers and partners which enhance opportunities (Panian, 2009).

Three major schools of thought uncovered in the literature provide perspective relevant to this thesis research. One evangelises that a core data governance framework, processes and best practices are agnostic and should in reality support any BD initiative or BD technology; hence discourages terms such as cloud data governance, big data governance etc, as this would imply that they are distinct disciplines separate from traditional data governance (Karel, 2007; Cohn, 2015). Arguing that traditional data governance processes should meet the needs and complexities inherent in BD use cases. The other school of thought focuses on the distinct characteristics of BD and assert that the sheer velocity, volume and variety of BD, is invariably introducing complexity in the way organisations capture, store, process and generate it – thereby advocating the importance of a big data governance framework in its own right (Kim et al, 2018; Al-Badi et al; 2018). Somewhere in the middle of these two perspectives, Zhang et al state that "because BD is still data, so big data governance is also part of data governance" but appear to go further with the angle that it should "be a new product or new stage in the development process of data governance, which not only contains more contents, but also involves more factors". They elaborate that BD has three attributes of data (data source), technology (data processing) and society (data application), this determines that big data governance is complex activities consisting of environment, social and technology, and pays more attention to data security, privacy protection, service innovation, organisational decision-making, and other aspects in the application of BD.

Whether the different perspectives are compatible is a matter for debate. The thesis study is not agreeing with the view that data governance does not need to extend to accommodate the distinct characteristics of BD and other complex requirements. The literature review outcomes underscored the relevance and significance of having a set of core principles and a distinct framework for governing BD to get the expected value from it. It seems that all

approaches evaluated have in common that they do not focus on end-to-end aspects of big data governance. In the new BD landscape, this appears to leave an open gap which inadvertently leads to project failure and isolated solutions which in turn could hamper BD strategies aimed at achieving competitive advantage over business rivals or impede them from obtaining value from the BD they are attempting to exploit. Furthermore, big data governance could have several conflicting requirements due to the uniqueness of an environment or industry sector, hence a big data governance framework should cater for handling distinct organisation requirements and priorities, thereby targeting to make them more accessible for global operating companies which will in turn aid to embed these frameworks into existing structures and business models.

To provide clarity on what a big data governance framework is, we examined the attempts at data governance definitions discovered both in the scientific research papers and practitioner community. Given there was no consensus or consistency in the definitions uncovered, it is deemed vital to produce a definition for a big data governance framework to provide clarity within this thesis. This thesis study's novel definition is **a framework which formalises the decision rights, accountabilities, policies and procedures within which an organisation's BD is governed and enabled as a valuable asset. It sets out the core strategy, guidelines, standards and principles that lay the foundation for a sustainable big data governance structure to address the specific BD challenges facing an organisation through the life cycle stages to optimise strategic BD opportunities.** This novel definition above in bold text was produced as an outcome of this thesis study but conforms to the depiction of big data governance in the critiqued literature.

Subsequently, this leads to the research direction of this doctorate thesis which aims to fill the research gap identified.

2.6 Summary of Chapter Two

The above sections of this chapter aimed to pinpoint the research gaps that are present in the current literature regarding big data governance. Big data challenges and risk factors in the literature have been investigated using a systematic literature review protocol that will be discussed further in section 3.3. The objective was to develop knowledge on the evident

complexities that BD presents to organisations that warrant the high-level requirements for a robust big data governance framework. A review of the data governance frameworks uncovered in both the scientific research and practitioner community is also discussed in this chapter, presenting argumentation of the different frameworks regarding their focus on addressing BD issues and risk factors.

The outcome of the systematic literature review contributes to knowledge by revealing an important gap in the literature, in that the linkage between data governance frameworks and the emerging big data complexities that they should address has only been minimally examined to date. Furthermore, the big data governance frameworks that were identified in the literature were found to have shortcomings. Thus, forming the motivation for this thesis study.

3.0 Research Methodology

This chapter describes the methodology utilised to answer the sub-questions and conduct the research for this thesis. The first section describes the research strategy and the underlying research philosophy underpinning the selected methodology. Furthermore, it discusses the theoretical framework within which the research was drawn. The next sections present the research design, followed by detailed discussions on the data collection procedures, software used for investigation and data analysis mechanisms. The chapter culminates by describing the limitations of the methodology.

3.1 Research Strategy & Theoretical Basis of Research

To answer the sub-questions and achieve the objectives of this study, the study approached the research problem by following a qualitative research methodology. Developing a conceptual framework from existing literature and practice evidence is a process of theorisation, which uses grounded theory methodology (Strauss et al, 1990). For several reasons, it is considered a suitable approach considering the motivations and nature of the research thesis.

Grounded theory was first introduced by Glaser and Strauss (1967), as an inductive form of qualitative research. Burney (2008) highlighted the differences between the inductive and deductive approach to research development – also referred to in research as inductive and deductive reasoning. The deductive form of qualitative research is a top-down approach, conducted from the perspective that the research has a distinct theoretical stance prior to starting the data collection process (Saunders et al, 2006). This theoretical stance is established from existing proposition or hypothesis. However, the inductive form of research development is a bottom-up approach, which means the researcher infers the implications of the findings and a theory is the outcome of the approach (Burney, 2008). Theory is developed from the data, progressing from specific to broader generalisation and conclusions based on facts identified in research findings. According to Sanders et al (2007) the purpose of this is to discern what is going on, so as to comprehend an entity or concept in a much better way (Saunders et al, 2007).

Orlikowski (1993) cites a benefit of using grounded theory. The writer aligned with the argument of Eisenhardt (1989) contending that the continual comparative analysis of grounded theory is less likely to result in researcher prejudice. Highlighting the distinction between 1) statistical generalisation that generalises from a population sample and 2) analytic generation which is the generalisation of theoretical concepts and patterns (Orlikowski,1993). Glaser and Strauss (1967) purport other advantages for the method:

- The grounded theory enables the development of an explanation of a phenomenon, whilst simultaneously grounding that account in empirical observations or data. This inductive method is particularly valuable where limited theory exists, or none is adequate.

- The grounded theory provides a method for dealing with the complexity of a social connected domain such as the big data governance paradigm and making it meaningful.

- The grounded theory approach is inherently process orientated, so aids to deal with change and the consequence of change.

However, there are several acknowledged problem areas, which is also experienced during the research.

- Producing a grounded theory can be time consuming.

- Producing a grounded theory is difficult involving simultaneous collection and analysis of a dataset, sampling based on emerging theory and the generation and integration of abstract concepts.

- Grounded theory is readily understandable to the layman, and this could be said to be one measure of a successful theory. However, a criticism is that it is challenging to readers, mainly because it is dense, and its boundaries are difficult to establish. To the extent that practical steps were undertaken to minimise the effect of this limitation.

Moreover, following the goals of this thesis which is to develop a conceptual framework to support big data governance within organisations, it is salient to base the research design on this type of inductive analysis. A conceptual framework is proposed here in order to ensure developed governance systems will adhere to the requirements to address BD complexities i.e. the requirements of ethics, regulations, management, external and internal policies related to the BD context etc. The notion of a framework is chosen as the optimal approach through which to address the identified problem. A conceptual framework is defined as a network or a 'plane' of linked concepts, which also present an integrated way of inspecting a problem under investigation. Moreover, it is a layered structure consisting of a set of subsystems or components, each performing part of the entire intended process and interrelating components through the output of other components (Paradkar, 2011). Grounded theory method offers a procedure of theorisation for building conceptual frameworks based on our inductive analysis approach.

Furthermore, one of the widely used methods within the IS discipline for the analysing qualitative data is coding (Abraham et al, 2019; Mao et al, 2021; Tallon et al, 2013). Specifically, Strauss & Corbin developed open coding techniques as part of the grounded theory method. They state that coding represents "the process of breaking down, examining, comparing and conceptualising data" (Strauss et al, 1990). It is established that one of the benefits of using these tried and tested grounded theory techniques is that it provides systematic guidelines for gathering, synthesing, analysing and conceptualising data for the purpose of theory construction (Strauss et al, 1998). It is designed to be useable with a wide range of research questions and in the context of diverse metatheoretical procedures. For this research it enabled comprehensive grounding by methodical and intensive analysis of the parliamentary inquiry dataset, by meticulously examining wording as well as phrasing, sentence by sentence. Clarity was gained on the emerging theories through the coding process, which necessitated the need to be highly observant of the dataset utilised during the detailed coding and transcribing stage of this research study.

Crucially, grounded theory is a methodology that can be utilised in most research paradigms and has also been used broadly in an interpretive context (Urquart, 2001). The body of literature exemplifies the notion that BD, as well as its associated opportunities and risks is a

social phenomenon in our contemporary times (Raguseo, 2018; Sestino et al, 2020; Vassakis et al, 2018; Kho, 2018). Most social phenomenon are complex and linked to multiple bodies of knowledge that belong to different disciplines as uncovered i.e. business, ethical, technology, legal etc. In addition, grounded theory has proved particularly useful for extrapolating the relation of actions to settings. It can be used for establishing typologies of relevant phenomenon and it can help identify patterns and relationships to complex systems. As such using this qualitative method is considered as an adequate tool for investigating this complex phenomenon in relation to big data governance frameworks. BD is a paradigm that involves complex relationships and systems to govern. Additionally, using qualitative methods for building the conceptual framework for a phenomenon that is linked to multidisciplinary bodies of knowledge is considered appropriate.

Other research papers looking at data governance frameworks within organisational settings have also found this approach suitable to critically investigate topics related to data governance activities, domains and frameworks (Otto 2011, Abraham et al, 2019; Alhassan, 2016). With this pragmatic stance as a backdrop, this inductive approach is used to derive concepts, constructs, principles and relationships that must be addressed by the big data governance framework in relation to governing BD. It is worth noting here again that the grounded theory approach is aids to deal with change and the consequence of change. Building on the history of why data governance was originally utilised for traditional data, the framework of Khatri et al (2010) is used as a basis for the proposed framework and significantly expand this based on the analysis findings. The next section discusses the research design for this study.

3.2 Research Design

The objective of this thesis study is to contribute to knowledge, by recommending a novel big data governance framework that will support global operating organisations to drive their BD strategies. The novel toolkit aims to contribute to improvements in the accessibility of governance processes to manage BD as an asset supporting organisations to maximise its value. To address the goals achieved in the thesis, the research adopted a 5-stage methodology approach as outlined below:

Stage 1: Literature review

The first stage of the thesis research was an extensive two phased systematic literature review to examine the current state of the art of big data governance concepts, existing data governance frameworks and business models. Crucially, the gaps in this chosen research area are identified at this stage to support the core objectives of the research. In addition, the outcome of this systematic and structured review was critically analysed to deduce key elements and dimensions, which could potentially be applied to develop a suitable framework to govern BD. An overview of the two-phased literature review protocol utilised is described later in this chapter.

Stage 2: Data analysis & requirements collection (secondary data sources)

The inductive analysis stage was designed to unearth the distinct core concepts big data governance should address. The literature study buttressed the view that there is limited research on big data governance. Additionally, an in-depth analysis of the interactions governing BD were identified to be missing. There are a broad set of factors that impact BD decision-making within an organisation, notably in scenarios where multiple stakeholders are involved. Hence, as contended in stage 1, a deeper comprehension of the BD complexities and risks is necessary for this study to identify governance areas that can influence a BD framework. Due to the nature of BD and its influence within organisations at deriving value and driving the decision-making quality, it was deemed important to comprehend the risk factors as considered in organisational environments and in practice. Therefore, following the initial literature review, the second stage of this study was to utilise Atlas.ti (qualitative data analysis software) to identify, appraise and synthesise multiple artefacts from chosen data sources.

To increase the credibility of the research, systematic analysis was conducted on a set of openly available UK government data sources which include artefacts such as written evidence, transcripts of oral evidence and detailed reports. The aim of this stage was to obtain a greater impression and explore reported concerns pertaining to BD practices, critical success factors and big data governance requirements to address the research questions. Criteria and requirements unveiled were critically analysed to assess those that affect the variation of the data governance approaches with respect to big data governance. Observations of general patterns, regularities and gaps were induced from the assessment,

which resulted in identification of criteria to address big data governance and design recommendations for a holistic big data governance framework. The data sources utilised are presented in section 3.4.

Stage 3: Design of big data governance framework

The data analysis output and inductive grounded theory process was used to establish the theoretical constructs that underpin the development of the conceptual framework developed. This fourth stage built out the arguments and proceeds to design the framework against BD aspects at different levels.

Stage 4: Evaluation

This stage evaluated some big data governance frameworks identified in stage 2 against the conceptual framework proposed in stage 3. The characteristics of the big data governance frameworks were analysed and compared to identify the similarities and differences between them that affect the variation of the data governance approaches with respect to big data governance. Observations of general patterns, regularities and gaps were induced from the comparison, which resulted in further arguments to validate the design recommendations for improving existing data governance frameworks to address big data governance.

Stage 5: Interpretation, Recommendations and reporting

Interpretation of data was carried out to discuss the results of the analysis. Following interpretation, the outcomes of the research were reported, and design recommendations for the new holistic big data governance were documented based on the findings.

The research process did not follow a strictly sequential approach and at several steps, backward steps were performed. For example, to receive feedback from research supervisor, refine research questions, review emerging literature and implement changes in the design.

Research Stage	Discussion
1. Literature Review	Chapter 2, Chapter 3 section 3.3, 3.3.1, 3.3.2, 3.3.3
2. Data Analysis and requirements collection	Chapter 3 section 3.4, 3.4.1, 3.4.2, 3.5 Chapter 4
3. Design	Section 3.5, 3.6 Chapter 4
4. Evaluation	Section 2.4, 5.4
5. Interpretation & Discussion	Chapter 3, 4, 5
6. Conclusion	Chapter 6

Table 4: Sections that the stages of the research process are discussed

3.3 Literature Review Process Description

The goal of the literature review was to collate and assess the existing studies related to the thesis research objectives, with the aim of achieving impartial and auditable results which can also be replicated.

This part of the research adopted the formal systematic literature review framework proposed in 2007, by Brereton, et al (Brereton et al, 2007). There were reasons for choosing to conduct a systematic review using this approach. “It recommends a frame of reference to identify current research direction, determines research gaps and highlights areas for future research and can also provide current evidence around a technology” (Brereton et al, 2007). Hence, a formal systematic review process was used to identify, appraise and synthesise multiple research papers that met pre-defined eligibility criteria to answer the research questions. By following the approach, it enabled observance of a plan for the systematic review and displays transparent steps in terms of the method followed.

Figure 5 below shows the approach adopted for the systematic review conducted.

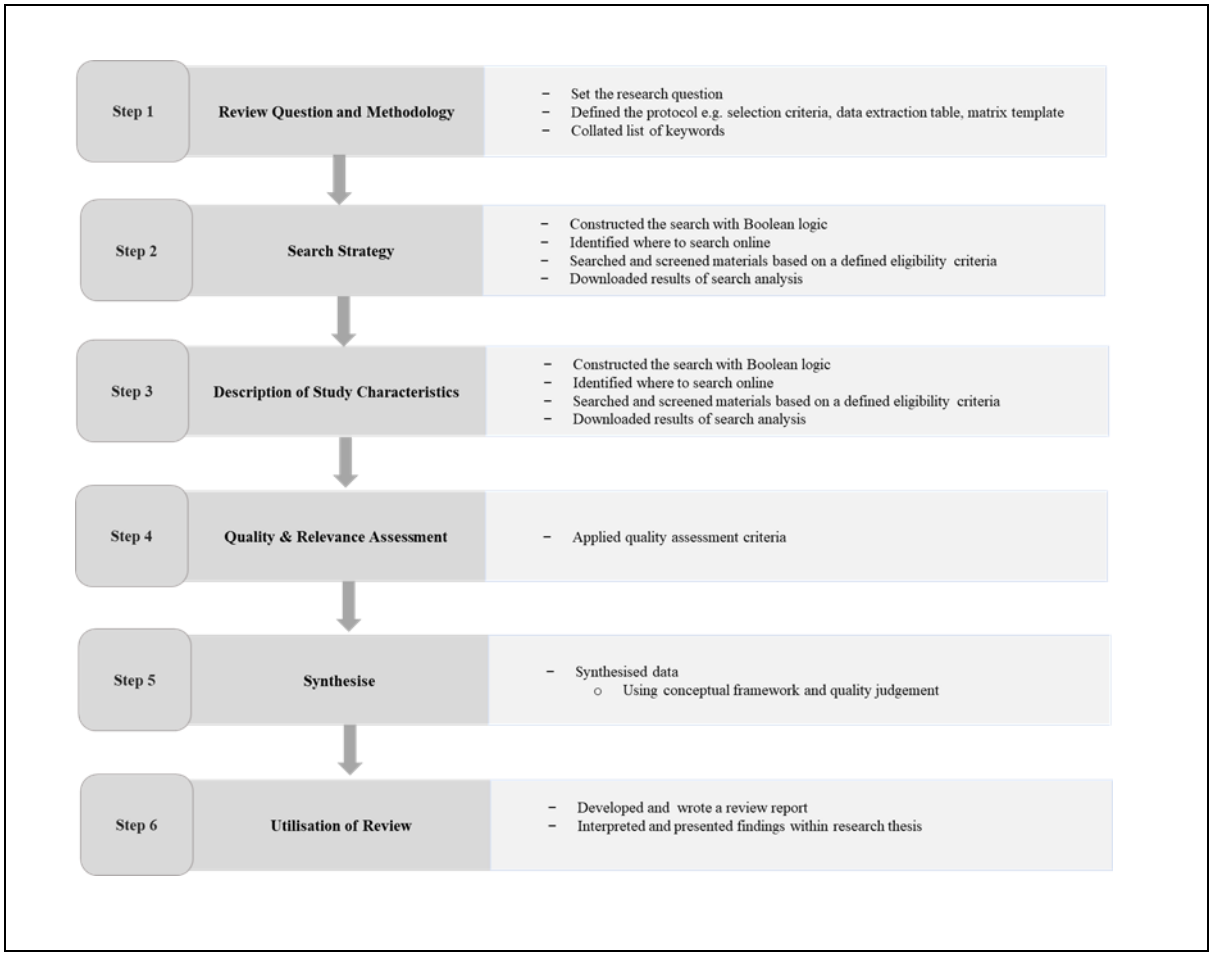

Figure 5: Overview of Systematic Review Approach (Gough et al, 2012)

The following subsections elaborate further on the protocol used to conduct the review.

3.3.1 The Search Strategy

At the onset of the review, a scoping exercise was done to establish where the key journals likely to contain information related to the research could be located. Following on from this exercise, the review was conducted searching Elsevier's scopus database. There were several obvious reasons for choosing the scopus database solely to conduct the online review. Firstly, it is deemed to "cover approximately 36,377 plus titles and numerous international papers, amongst which there is coverage of 34,346 peer-reviewed journals in different areas" (Elsevier website, 2019). Additionally, it integrates with various reputable online research journal databases; therefore, it was decided that it would be a good source where a significant

proportion of published journals and other material would exist. The searches were limited to the abstract, title, and key words fields.

To identify the relevant material, a strategy was found to build the search terms and list out key words as follows:

- Key words derived from the research questions.

- Identification of any synonyms or other terms that could be used for the various key words were identified. For example, synonyms identified for the key word "Data Governance frameworks" for research question 2 were *"Data Governance approach" OR "Data Governance activities".*

  Similarly, examples of synonyms used for the key word search term *"Big Data Governance frameworks"* were *"Big Data Governance approach" OR "Big Data Governance organisation".*

- Use of Boolean operators *AND & OR* to construct the research search string to incorporate synonyms and link major terms.

3.3.2 Study Selection Criteria

The search selection criteria were defined to ensure a pre-defined profiling approach to select existing research papers we downloaded for further evaluation. In a systematic review, there is the necessity for the researcher to explicitly specify the search boundaries to guarantee the quality of the review and focus the scope (Brereton, 2007). The various criteria for selection of the research papers included in the study, are referred to as the inclusion and exclusion criteria and are presented below.

Inclusion Criteria

- To enhance the quality of the review, only published peer-reviewed journals and articles were initially selected. However, this was later extended to include key conference papers after manually reviewing the references and citations of selected papers.

- Only Scopus document types ‘article’ and ‘review’ and “conference” were selected.

- Materials published between the periods ‘All years to present’ were searched to ensure a complete historical understanding of the state of the art.

- Duplicates found in searches were removed.

- If several studies had reported similar research, only the most complete paper was selected for further assessment.

- Papers that were either empirical or conceptual were included, as some researchers argue that it is also important to identify conceptual and theorised research developments (Sivarajah et al, 2017).

- The paper had to be available at no cost.

Exclusion Criteria

Studies that met the listed exclusion criteria below were omitted from the review conducted:

- For research question 2 (table 1, section 2.0), studies that did not specifically discuss data governance approaches, data governance frameworks or data governance models were excluded.

- For research question 1 (table 1, section 2.0), we excluded studies that did not cover Big Data in relation to some of the additional complexities it produces.

- Papers not in English were excluded.

3.3.3 Search Results and Selection

Earlier in chapter 2, we presented the systematic review of past research papers dealing with data governance frameworks and any solutions in the papers highlighted to address big data challenges.

For research question 2 outlined in the previous chapter (i.e., table 1 section 2.0), the keyword search criterion of *"Data Governance" OR "Data Governance frameworks" OR "Data Governance framework" OR "Data Governance model" OR "Data Governance Models" OR "Data Governance approach" OR "Data Governance organisation" or "Data Governance Activities"* was used, yielding 225 articles and review papers published from the year 2005 onwards. Of which 219 were written in English. A further search which used the keyword search criterion *"Big Data Governance" OR "Big Data Governance frameworks" OR "Big Data Governance framework" OR "Big Data Governance Model" OR "Big data Governance Models" OR "Big Data Governance Approach" OR "Big Data Governance Organisation"* was attempted resulting in a further 8 articles and review papers published from 2012 onwards. After reading through the abstracts and introductions of these papers, 36 studies were short-listed for further review.

Following a further manual review of the references and citations of the scientific publications that were shortlisted from the initial review, the search was extended to include conference papers. 5 additional key papers were discovered by extending the search to include conference papers.

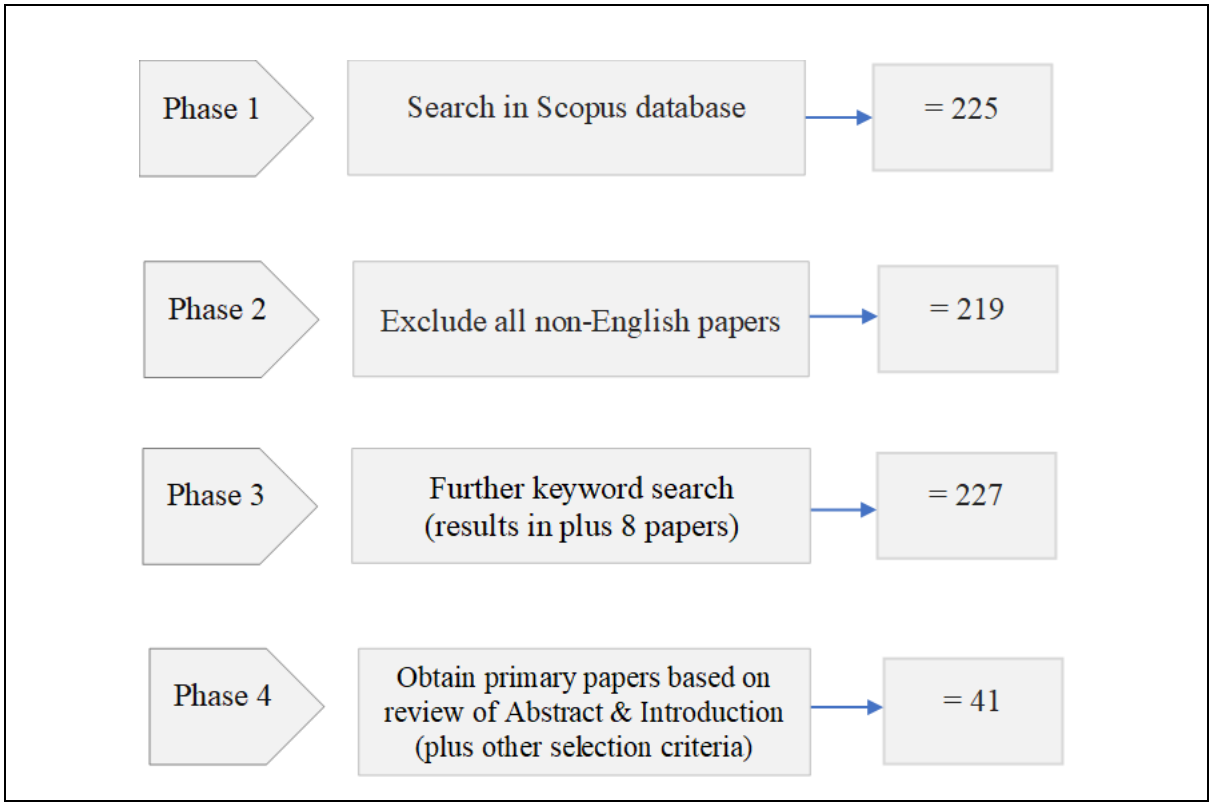

Figure 6: Research question 2: Outline of Scopus search results

Researchers agree there is only limited research on data governance (Otto, 2011b; Wende et al, 2007) and an elaborate analysis of its interactions governing aspects of Big Data were found to be missing. It was also deemed critical to comprehend how scientific concepts are applied in actual business environments and obtain relevant understanding of how data governance is executed or utilised in practice within business organisations. Subsequently, for this thesis research, the results were later combined with data governance sources and contributions from practitioners on data governance. These were mainly; white papers by reputable consulting companies or major software vendors e.g. IBM & Informatica, reports by Analysts such as Gartner, and publications by industry associations such as Data Governance Institute (DGI) and DAMA International. The decision on which practitioner publications to select was taken after the manual review of references and citations of the shortlisted scientific publication and based on consideration of whether the practitioners were deemed to be key players in the industry. We then reviewed these publications selected again for a more detailed analysis, to identify publications that explicitly covered data governance frameworks, as well as to classify them in terms of science or practice-oriented publications. As discussed earlier in chapter 2 a total of 29 data governance frameworks were uncovered during this stage and are referenced in the table 1, section 2.0.

For research question 1, which objectives were to delve further into the Big Data complexities data governance is needed to address, we performed a search with the keyword search criterion "Big Data Issue" OR "Big Data Issues" OR "Big Data Challenge" OR "Big Data Challenges" OR "Big Data problem" OR "Big Data Problems" OR "Big Data Risks" OR "Big Data Complexities". This search yielded 636 papers in English within journals. A further search using the keyword search criterion of "Big Data" AND Issues OR Issues OR Challenge OR Challenges OR problem OR Problems OR Risks OR Complexities AND "Data Governance" yielded 31 additional papers. After reading through the abstracts and introductions of these papers, 93 studies were short-listed for further review.

The overall search, including the manual review of citations and references resulted in a total of 62 publications being included in the full review.

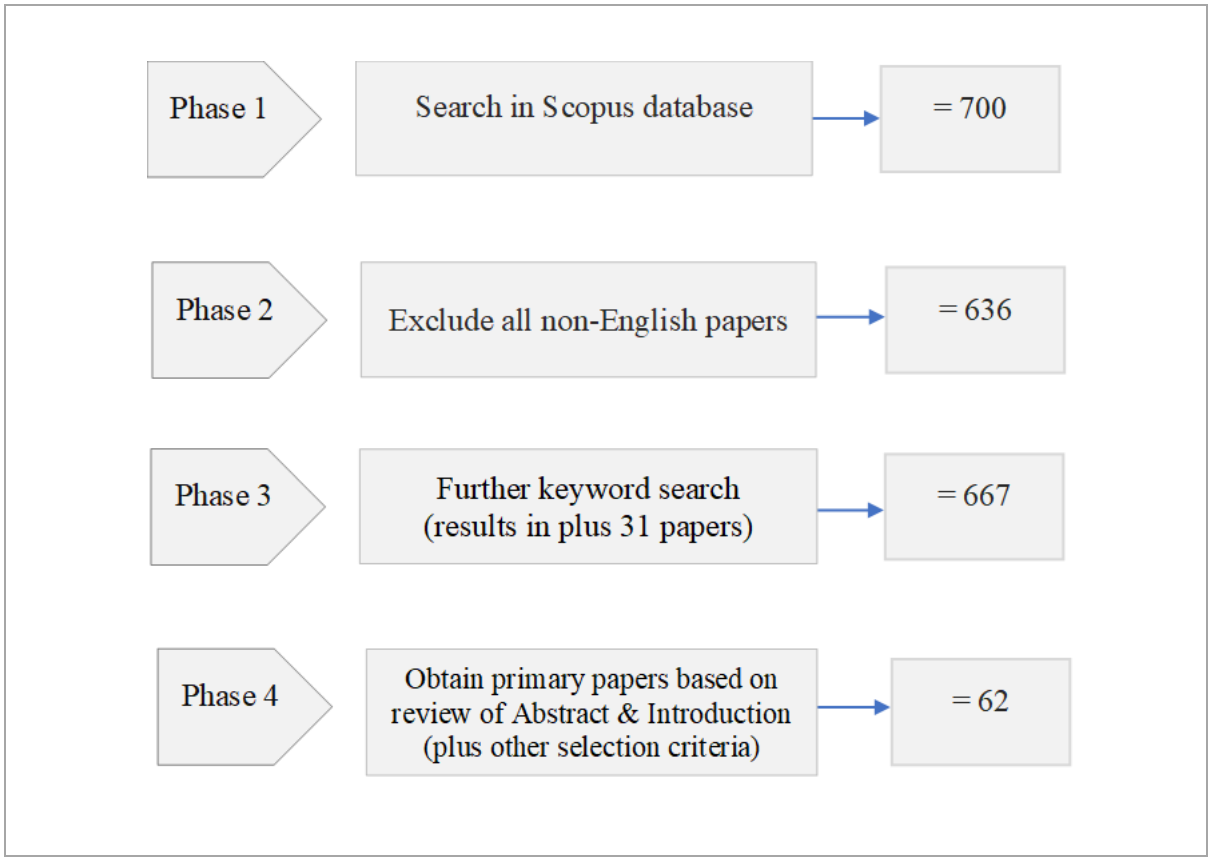

Figure 7: Research question 1: Outline of Scopus paper selection process

The next sections discuss details of the inductive analysis methodology undertaken on secondary data sources in this thesis study.

3.4 Methods for Data Collection and Inductive Analysis

The BD challenges, risk factors and key relationships criteria pinpointed during the research analysis were operationalised through the critical examination of UK parliamentary submissions.

The dataset was compiled in response to inquiries led by the UK Parliament, House of Commons Science and Technology Committee. The datasets used for analysis are outlined below:

- One of the datasets is compiled in response to the 'The BD Dilemma inquiry' in 2015 led by the Science and Technology Committee (UK Parliament, House of Commons). The dataset contains written and oral evidence submitted by various organisations and stakeholders, as well as inquiry responses and reports.

- The second dataset was compiled in response to the "Social media data and real time analytics inquiry", led by the commons select committee. The dataset contains written evidence from various internationally operating organisations and stakeholder e.g. Digital Catapult, LinkedIn Corporation, Twitter, Google, Facebook, Office for National Statistics etc.

The dataset was chosen for this study due to the availability of written and oral evidence submitted on BD challenges or risks by various organisations, independent bodies and stakeholders. As this thesis is orientated towards knowledge contribution in terms of BD governance in an organisational setting, its potential application is expected to be realised by industry practitioners and government stakeholders. Therefore, it was deemed salient that the basis of the conceptual frameworks design initially examines reactions from a sample of practitioners, thereby establishing validity into the problems it proports to address. The inquiry responses and reports submitted were discovered to include some of the UK's foremost experts and focus groups drawn from a broad range of organisational areas, who regularly utilise BD within their organisations or research areas. Therefore, they were considered well placed to provide input into the committee's inquiry. An expert is deemed to be an individual (or group of individuals) who is considered to have built up expertise in a domain through research or a profession. Furthermore, BD by its ubiquitous nature crosses geographical boundaries, hence although the inquiry was run by the UK Parliament, it is salient to note that the dataset was considered suitable as responses were also submitted by global operating companies. More importantly, it was ensured the selection of this dataset was suitable due to the richness of BD issues and implications to its governance covered. Although the datasets utilised for analysis were collected in 2015 and 2014 respectively, the primary literature review conducted buttresses that most of the information highlighted by participants is still prevalent, therefore this is judged to be a suitable secondary data source.

The main objective was to analyse the data and delve further into the issues, gaining a full understanding of the interaction practices and general ecology of this topic area within

organisations. Furthermore, to glean further insights to support the answering of research sub-question 1 & 3, BD risk factors, identify final themes and concepts. Utilising datasets from focused inquiries such as these yielded extensive and useful information related to complex behaviour and explore attitudes towards this complex topic i.e. BD. Additionally, inductive approaches were utilised to extract key requirements or criteria for a governance framework that can address BD challenges and risk factors.

Due to the time available for the research, the study is carried out on two inquiries, with 119 responses from different organisations. Given the suitability and relevance of the topic areas, combined with output from the previous extensive literature review, this was considered sufficient for this thesis research.

3.4.1 Evidence Papers Administration and Participant Characteristics

The first datasets examined was evidence requested by the Science & Technology committee titled 'The BD Dilemma inquiry'. Invitations to participate were published online, with a deadline for receipt of response submissions by 3rd September 2015. Crucially, the datasets are extremely relevant to this research study as the inquiry's primary objective was to investigate BD risk factors as reported by participants practicing in the field. The second batch of datasets were responses to an invitation for written submissions on the topic 'Social Media data and Realtime Analytics'. One of the key objectives of this parliamentary inquiry was to examine the barriers to implementing real-time analysis of this BD type i.e. Social Media data. The deadline published for receipt of response submissions for this specific inquiry was the 31st of March 2014. In both cases the online invitations identified the objectives of the inquiry and assured participants that the evidence supplied would be processed in accordance with the provisions of the Data Protection act 1998. Taking both datasets together, a combined total of 119 responses were analysed. Participants were represented by both local and international organisations from disparate industry sectors.

The inclusion of subsequent themes or critical factors in the conceptual framework is based on how participants from the inquiry report various critical factors or interpretations in relation to BD challenges that are ascertained big data governance needs to address. Due to the substantial number of factors unveiled in the inquiry, it was decided to categorise them into core constructs/variables.

3.4.2 Atlas.ti for Analysis

The following overview of the research method provides a background context to its specific application, before discussing detailed analysis results and findings in the next Chapter. Based on our earlier systematic literature review of big data governance, the results indicated that limited sources have addressed data governance widely and big data governance in particular. The results also showed that there are a lack of empirical studies specifically addressing the challenges or risk factors that big data governance must address. Based on these earlier findings, this research investigates the BD challenges and risks factors negatively influencing the implementation of BD initiatives. This important exercise is performed to extract the common barriers encountered, and requirements to be addressed by the proposed conceptual framework.

The starting point in our research analysis was to import the datasets into Atlas.ti to conduct a systematic data analysis. This was to aid the initial classification of factors in the domain of our proposed framework through coding. Atlas.ti is a qualitative data analysis software and the most recent version (i.e. v8.4.24) at the time of this research was used. It was later upgraded to version 9.1.3. As coding is central for aspects of grounded theory analysis, computer-assisted qualitative data analysis (CAQDAS) is considered well suited to support the analytical approach for this research. There are various studies where Atlas.ti has successfully been used for grounded theory analysis.

Following Strauss et al (1990) inductive approach of examination, components for the proposed framework are generated by analysing the dataset. For the conceptual framework consideration is given to the parameters of BD challenges or risk factors, the characteristics of each complexity and the requirements for big data governance in each case. In recent years, a few writers have approached BD challenges as an emerging requirement of data governance, therefore this justifies the vital step for BD requirements identification in our research techniques (Kim et al, 2018; Al-Badi et al, 2018; Soares, 2012; Zhang et al, 2017; Tallon, 2013).

The key considerations focused on during the investigation to interrogate and code the data are:

I. What are the distinct BD challenges or risk factors reported by respondents of the inquiry?

II. What is the frequency that participants of the inquiry identified these BD challenges? What is the subsequent impact of these BD challenges or risk factors within the organisation if reported?

III. Is there a hierarchy of importance that emerges for each element identified?

IV. Are there any common themes, patterns or linkages between these BD challenges that demand greater attention related to big data governance within organisations?

V. Is there any relevance to address one or more of these BD elements, in the context of long-term inter-organisational and big data governance relationships?

Three very complex rounds of coding and data analysis was completed utilising Atlas.ti. The Atlas.ti coding process was inductive following the thesis methodology selected and described. The research process was also iterative, "requiring a steady movement between concept and data. As well as comparative, requiring a consistent comparison across types of evidence to control the conceptual level and scope of emerging theory" (Orlikowski, 1993). Myers (2006) cites that this technique is required by grounded theory, as it supports a continuous interplay between data collection (Myers, 2009). The use of Atlas.ti for coding enabled this cyclical and iterative approach to our data analysis that would have been difficult to accomplish manually. Secondly, it strengthens the credibility of our research by ensuring processes utilised for data analysis are transparent and replicable.

The steps adopted in the coding procedure were proposed by Finney & Corbett (2007) and were also drawn from prior data governance and BD challenges research studies in which validity had already been determined to varying degrees (Otto, 2011b; Alhassan, 2016). The steps are summarised below in table 5:

Step	Description
1	Decide the levels of analysis: Written evidence and reports are initially read and re-read to discover terms and concepts of interest. Crucially, this step was conducted to induce or interpret the developments of concepts. Start to identify key metaphors, themes and concepts, which also involved deciding what level of analysis would be conducted during research.
2	Decide how many concepts to code for
3	Decide whether to code for existence or frequency of a concept
4	Decide how to distinguish between concepts
5	Develop rules for coding the text or oral evidence
6	Decide what to do with irrelevant information
7	Coding the text
8	Analysing the results

Table 5: Adopted steps for coding procedure (Finney et al, 2007)

- **Atlas.ti first stage coding:**

Open coding was used for first stage data analysis, as it enabled segments of data to be evaluated in detail and be compared with one another. Due to 1) the limited number of papers that explicitly contribute to big data governance and 2) Limited researched or practitioners outlining their methodology, it was decided to adopt the approach of Alhassan et al (2016) and code inductively for theoretical concepts that could be interpreted as challenges or requirements that needed to be considered for DG activities to address BD related risk factors. It utilised clustering based on connected characteristics and underlying constructs to develop categories from concepts. The organisation of concepts into categories and categories into thematic categories into meta-categories is clear evidence of hierarchical clustering. At its apex is the core category and all other categories are related to this by a series of cause-and-effect relationships. Subsequently in this part of the initial stage, all 119 documents included in the dataset were first read in their entirety to gain familiarity and understanding of the information. Interestingly, assessment of participant comments initially reviewed matched the view that there was still a lot to be done in relation to BD challenges of

which big data governance would be helpful. Consequentially, the assessment proved beneficial in validating the core concept of the conceptual framework and the factors under consideration and raised the confidence level of the thesis study. The dataset was then evaluated again, this time coding each phrase or concept of interest line by line.

This initial stage resulted in indexing and making an initial attempt to systematically collate segments and information of interest to the thesis research. Several conceptual themes were derived using the overall methodology.

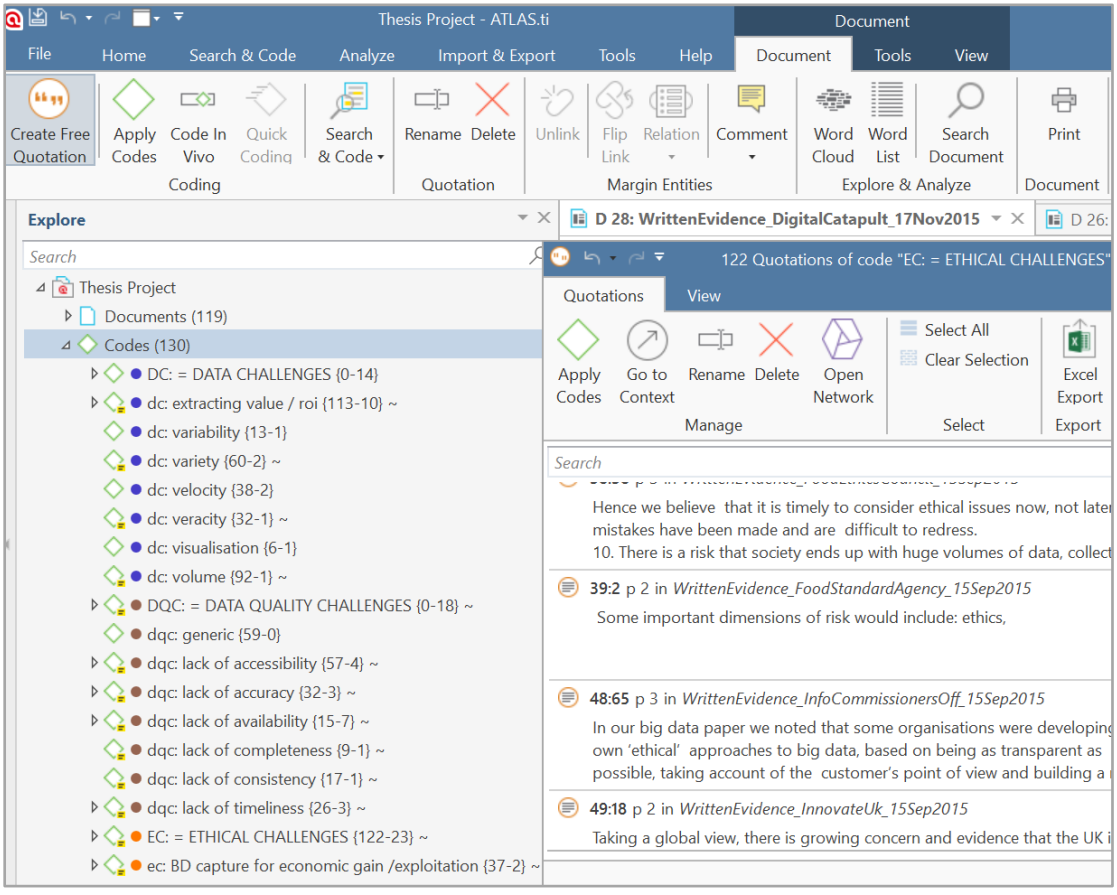

Figure 8: Snapshot of code list from Atlas.ti thesis project

Figure 8 above shows a snapshot of the Atlas.ti thesis project created during this research and some of the coded categories derived through analysis, which reflect specific conceptual themes. The definition and development of each of these constructs are addressed separately in chapter 4. It is worth noting that these elements discussed in the next chapter are by no

means exclusive; nonetheless, they endeavour to give an extensive measure of the combination of various factors for big data governance e.g. determinants / Intensity, antecedents or significance.

Finney and Corbet (2007), advocated that before starting this stage, a decision should be taken whether to code for frequency or existence (Finney et al, 2007). For this study, the decision was made to code for frequency to make feasible a discourse of concepts that are considered prominent or of paramount importance. The code list developed for each category uncovered during data analysis is also shown in the relevant sections within chapters 4.

- **Atlas.ti second stage coding:**

  In the second stage of data analysis, similar or related codes were narrowed down by merging codes into higher order categories and reiteratively compared. This stage helped with streamlining the significant codes. Codes with limited usage or which were considered irrelevant were removed at this stage. Moreover, coded concepts with dissimilar names, but essentially representing the same concept became apparent in the code manager (i.e. "grounded" in the code manager indicated frequency). This step necessitated even closer interrogation of the data to ensure we verified and saturated the categories conceptualised to answer the research questions. At this juncture we are closely examining the conceptual constructs for big data governance that emerged in the initial round of coding, to further decide the most relevant requirements or criteria for the domain under investigation. This further assisted to expose flaws in any original coding or arguments made, revealed previously undetected relationships or underlying constructs, and enabled further expansion of initial analysis. The coding procedure undertaken on all 119 inquiry submissions resulted in 130 code concepts (excluding success criteria codes) which were further grouped into 9 categories.

- **Atlas.ti Stage 3 coding:**

During the third stage of the complex and extensive data analysis conducted, the dataset was again revisited, analysing for patterns and relationships. Considering a conceptual

framework is a layered system, comprising of a group of sub-components which each complete part of an overall process, the aim was to derive a linking or sequence of multiple codes to create networks. The resulting network diagrams represent qualitative visual depictions of the data and themes that emerged from the analysis. At this stage of the research, we identified the patterns in the data and interrelated components, by mapping reoccurring codes. When this stage was completed, lines of argument for the thesis research could be expressed for the overarching concepts, themes and patterns discovered during data analysis to answer the research questions. The methodology and analysis findings in the Atlas.ti third stage coding is discussed further in the following section.

3.5 Interrelationships of Big Data challenges & Risk Factors

The potential interrelationships of the BD challenges and risk factors is also examined, as an outcome of the analysis. Whilst a major goal was on the identification of BD factors in relation to criteria for big data governance, it is imperative to recognise the impact of how these elements may interface with each other. This enables improved comprehension of the various impact of each BD challenge or risk factor within the organisation. Moreover, the interrelationships express the effect of the existence or non-existence of each BD challenge on one another. Furthermore, examining related actions / interactions is a key element for it is "the manner in which any phenomenon is expressed" (Strauss et al, 1990). The related actions and interactions we subsequently unveil are embedded in a set of conditions (causal, contextual or intervening), that to a greater or lesser extent, determine what must be addressed by our proposed big data governance framework. Interactions between the themes uncovered have certain consequences; thus, these might also become part of the conditions for a future set of actions and interactions.

One of the most challenging aspects of this thesis study lies in the interplay of analysing and producing inductions from the large number of lengthy participant submissions, as well as recognising those relationships or interactions to the different themes and categories uncovered. There are numerous contexts for identifying such relationships. However, the research study utilised the inductive analysis method, also supported by inbuilt code co-occurrence queries in Atlas.ti. Using the inbuilt code co-occurrence queries to support

analysis helped to highlight the frequency of co-occurrences between various codes to discover the strength of relationships between them. Closer inspection of the co-occurrences influenced further inspection of the context in which participant referred to particular BD challenges or risk factors. Plus, associations between some of themes were revealed. What is clear from findings is the potential complexity of even the simplest causal relationship due to multiple conditions in operation at any point in time in the BD domain. For example, in table 6 below the code cooccurrence table cell for ***mc: data protection - limitations of consent*** code, reflected that it the most closely related code is ***ec: Informed consent lacking/incomplete*** with a total cooccurrence of 35. The number indicated the overall numbers of cooccurrences.

Code Groups	● lc: data protection & privacy infringement	● mc: data protection - limitations of consent	● mc: data protection - limitations of de-identification	● mc: inappropriate BD usage / sharing
● dc: extracting value / roi	2	2	0	1
● EC: = ETHICAL CHALLENGES	5	8	13	5
● ec: Informed consent lacking/incomplete	4	35	1	4

Table 6: Sample Code Cooccurrence table 1

Code groups	● mc: privacy concerns	● mc: security of BD / BD systems	● oc: bdg - establishing and adopting policies, processes, guidelines	● pc: BD mining & ML barriers
● ec: discrimination/bias/profiling	4	1	2	19
● mc: data protection - limitations of consent	10	4	5	3

Table 7: Sample Code Cooccurrence table 2

In table 7 above, it can be observed that the code cooccurrence table cell for ***pc: BD mining & ML barriers*** code cooccurrence with ***ec: discrimination/bias/profiling*** is much higher than all other codes (=19). This infers that BD mining and machine learning has some causal or linked relationship with ethical challenges that can cause discrimination or bias. This gave some insights which led to further examination. Investigating whether this could be a "associated" or "results in" or "cause of" relationship. If the network diagram is checked in

Atlas ti code, it will be seen that as part of the research study all network nodes (concepts) have been linked accordingly and any observed relationships were labelled.

The inductive analysis approach and sample screening of some code cooccurrences, although time consuming, revealed various permutations of the relationships between the codes. Consequentially, the outcome was the creation of a concept map (network diagram) in the third stage of coding, which assists in representation of identified BD challenges and big data governance, as well as their interconnectedness. This details the structure and relationships of the coded concepts that emerged from the inductive analysis and improves metacognition in the thinking about the knowledge uncovered (Friese, 2014). Furthermore, linking the codes together with named relationships formed a sequence of logic for the developing conceptual framework, assisting to ascertain the hierarchy and importance of each BD risk factor or challenge identified during analysis. It also aided to draw out the linkage to the BD requirements and vital big data governance activities necessary to address them. Hence the importance of this aspect of the research. In practice the process was more complex than that conveyed by this simplified outline. Figure 9 below shows the network diagram which outlines the results for this part of the research. This network diagram has also been submitted in electronic format separately as part of the Atlas ti code that forms part of the work of this thesis.

As depicted in the network diagram, a substantial amount of the BD challenges uncovered are interlinked and pose risks to the quality, availability, accessibility, auditability, security and integrity of BD, which in turn could lead to a trust deficit. For example, our visual representation shows that according to our analysis BD quality challenges are driven by the lack of accessibility through integration, we can seek to address this by considering BD standards, BD technical architecture, metadata etc. Management challenges could also be triggered by a lack of policies and procedures to govern BD as an asset. Lack of defined role and responsibilities within the organisation, skilled workforce gaps or a combination of these factors could also cause interconnected issues. Sometimes a categorised code could legitimately belong to more than one aspect of the BD complexities uncovered within the dataset. For example, we can say that the lack of BD standards could cause or lead to a specific context e.g. a lack of standards leads to BD integration or interoperability issues. In

addition, we can also say that BD standards are an intervening condition i.e. they come between action/interaction and consequences e.g. adoption of BD architecture standards could go some way to improving BD platforms integration and interoperability. We examined if this category should be identified as a cause, an intervening condition or both. On these specific occasions the transcripts were re-read to contextualise the category further. Sometimes it proved difficult to follow the relationship in the dataset, so the creation of a network diagram assisted to make sense of this. Finally, the results indicate that all stakeholders that participated in the inquiry were facing interrelated challenges controlling their BD assets. Schematic representations such as this, serve as an extremely effective system for presenting multiple dimensions and interdependencies in these relationships. The relative immaturity of big data governance research also strengthens usage of diagrammatic representation to articulate concepts. This schematic starts to lay out the key factors or constructs and outlines relationships between them. Analysing the code relationships in this way also aided comprehension of the evolving meanings in relation to the thesis research questions

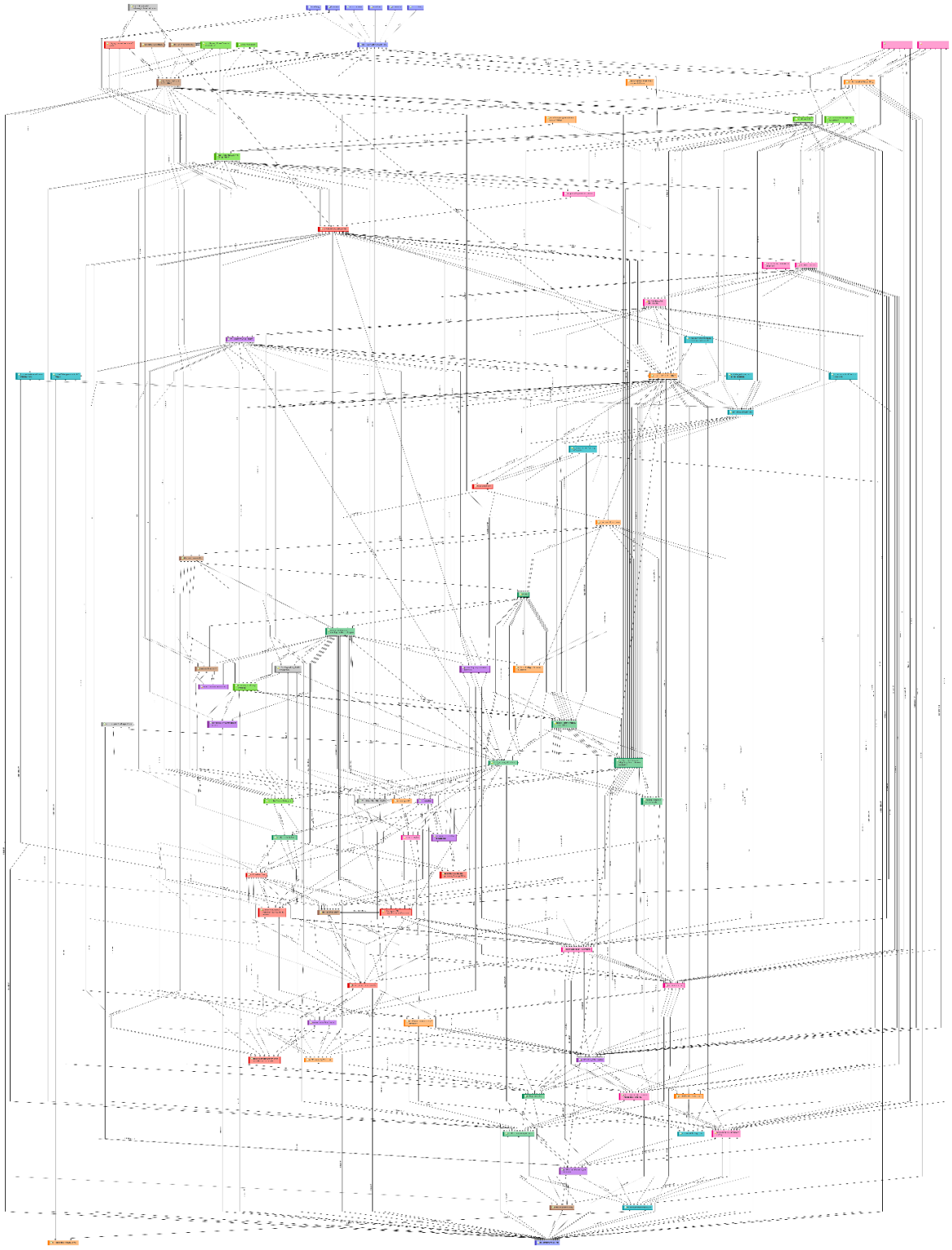

Figure 9: Atlas ti Network Diagram Reflecting BD Challenges Interrelationships

3.6 Big Data challenges & Risk Identification

Before moving to the next chapter, it is necessary to explain the importance of risk identification and its subsequent management within organisations. Research has suggested that globally there is a tremendous wave of innovation, productivity, and growth, including new forms of competition and ways of capturing value - all driven by BD as customers, organisations, and economic sectors exploit its possibilities (Manyika et al, 2011). Innovation is a dominant driver of competitive advantage and business growth for companies. In today's world organisations must innovate in order to survive, sometimes in a climate of huge uncertainty (Taplin et al, 2005). When innovating within a paradigm such as BD, risk is intrinsic (Zhao, 2005). Additionally, risks are inherent to change of legislation or law, market condition changes, labour availability, execution strategy or other external changes. Furthermore, the velocity of BD technology change, competitive operational environments and challenging customer demands, results in the associated innovation intrinsic with BD being very complex as discussed earlier. This leads to significantly less certainty, to possible outcomes from BD initiatives. From an entrepreneur's point of view, BD innovation could have a certain degree of risk, as the innovative activities targeted at the future usually encounter some degree of uncertainty.

The viewpoint of this thesis study is that implementing BD initiatives under conditions of uncertainty, result in tangible risks to organisations e.g. organisations could either gain or suffer losses to profit, loss of organisational reputation, reduction in competitive advantage etc (Berglund, 2007). We argue that BD challenges or risk identification is the paramount first step to development of our proposed conceptual big data governance framework. The identification of real or assumed risk, is a foundational step in decreasing the overall risk of strategic BD initiatives or BD innovation. Moreover, organisations must adopt a strategy for early risk or challenge diagnosis and govern these in order to implement successful BD initiatives. Furthermore, we consider the capability to identify, reduce and monitor these BD challenges or risks through governance to be of vital importance from initiation. Other researchers highlighted this to be significant during the earlier literature review (Tallon, 2013; Al-badi et al, 2018; Kim et al, 2017).

There does not appear to be a sole definition of the word 'risk' that is universally utilised. However, Edwards et al (2005) state the standard definition as "the combination of the frequency or probability of occurrence and the consequence of a specified hazardous event (Edwards et al, 2005). In other literature, the word 'risk' is claimed to be "an uncertain event, or condition that occurs, that has a positive or negative outcome on project objectives (Zhao, 2005). According to Wharton (1992), risk is utilised to "indicate a measurement of the chance of an outcome, the potential size of the outcome or combination of both" (Wharton, 1992).

It is contended in this thesis study that BD innovation, BD technologies and BD analytical practices inevitably comprise a high level of risk activity. Consequentially, organisations need to consider BD risk factors or complexities within their governance approach in order to drive successful BD initiatives. Any challenge that could potentially alter BD performance or BD initiative outcomes, should be governed and controlled. By the very nature of BD in the digital era, risk avoidance is not a suitable option for most organisations who adopt it for innovation or strategic purposes. Therefore, we argue that proactive risk identification is required, as there will be the opportunity to influence the route of events through our proposed governance approach, from the onset and at key decision points. The thesis research design, enables in-depth inductive analysis and assessment of BD risk complexities, based on powerful contributions of parliamentary participants. This is not a risk management study, however it worth stating that there is no formal method for quick recognition of all BD risk factors or challenges. Although the BD risks or challenges we identify during our analysis is not exhaustive or proport to cover every possible scenario. The rigorous analysis we have conducted, plus subsequent identification of BD complexities, uniquely serve as a reference point or checklist for other organisations to positively influence the governance of these BD risk factors. These reference points that this thesis study contributes, is useful to organisations with no prior or additional knowledge of the BD situations that could affect their organisations. Or organisations that have lower levels of maturity with regards to BD governance.

The contribution of these reference points or checklists, do not need to remain fixed or static. The static nature of checklists is a prevalent criticism, but we argue that they should remain

dynamic. It can be used as a starting point with the various BD complexities we identified as part of this research; but is open to additions of new or additional complexities prior or during big data governance implementation.

3.7 Summary of Chapter Three

In this chapter, an overview of the methodologies used in this thesis research is provided. The research method supplies the structure for learning and development of the proposed conceptual big data governance framework. Grounded theory gives the structure for the analysis of qualitative data through inductive coding and constant comparison in the analysis phase utilising an iterative cycle. The parliamentary inquiry submissions supply considerable rich open-source data, covering the topics relevant to this research.

The research which took place over an extended period, undertook to learn about BD complexities and its current constraints in practice. This enabled a better understanding of the requirements of different sectors for big data governance. Through an analysis of current BD challenges and risk factors unveiled by the inquiry participants, we validated there was a vital need for big data governance. Subsequent learning provided the impetus to develop a conceptual framework to aid big data governance in organisational settings.

4.0 Inductive Analysis of Big Data Challenges & Governance Dimensions

4.1 Introduction

This chapter starts to discuss the data analysis results and findings of this thesis study, which form the background to the building blocks of our proposed BDG framework. Moreover, it provides further insight into the grounded foundation of this research.

Previous discussions in this research study provided information on several types of DG frameworks and the limitations of these approaches in handling BD, along with the limitations of some of their methodologies. To construct a big data governance framework which assists decision makers with a suitable approach for governing BD, our study recommends one should first evaluate the BD challenges and risk factors that constitute the core requirements of the framework. Hence, to address the research questions we took a closer examination at key dimensions and priorities reported in participants inquiry submissions. The goal of this examination was to elicit which of these dimensions should be given priority in our novel BDG framework. The choice of this assessment sequence is grounded in the outcomes of the earlier literature review, that highlighted organisations and their decision makers are confronted with simultaneous considerations of both qualitative and quantitative factors impacting their ability to manage BD as an asset. As a consequence, our framework therefore recognises the BD challenges & risk factors to be considered as part of its construction. This stands in contrast to some of the earlier frameworks identified in the academic literature which appear not to address the tangible or intangible factors in this complex area holistically.

Following the identification of BD challenges, BD risk factors and subsequent BDG requirements derived from investigation of the dataset; a conceptual BDG framework is established to accommodate unique BD features which pose as organisational challenges, with the goal to address research questions 1 and 3: -

- What are the core Big Data complexities that organisations need to address?

- What are the elements of a Big Data governance framework?

The fundamental importance of comprehending the BD complexities that DG should address (Kim et al, 2018; Al-Badi et al, 2018; Soares, 2012; Zhang et al, 2017; Tallon, 2013) and the limited research of such relationships in the DG context, accentuates the need to investigate and acquire conclusions from other contributors or practitioners in this field. To the best of this writer's knowledge, there is no empirical studies in the academic literature that address governance for BD holistically in this manner.

The inductive analysis process undertaken reveals several well-defined challenges and risk factors that should be considered as part of a big data governance framework. Moreover, some of the concepts related to the potential success criteria and requirements for big data governance were identified, as well as the interrelationships between these domains. This enabled the study to work towards the design of a conceptual big data governance framework and subsequently, a cross industry strategy for implementing big data governance programmes. We start to discuss these in the next sections.

4.2 Big Data Governance Criteria Development

Our discrete and sequential approach to coding, impacted the big data governance framework design strategy. All the BD complexities (hence core requirements) identified during analysis, were gradually refined through an iterative process until a final criteria list was finalised. These were subsequently included as key components of the conceptual big data governance framework to be addressed dependent on organisational context. 130 independent codes that related to BD dimensions that big data governance should address were discovered, which were later classified into nine major categories. These concepts represent assessment of the BD challenges/risk factors organisations are confronted with, that align to our big data governance framework's core requirements. They are classified as follows:

1) Data Characteristics

2) Data Quality
3) Management
4) Organisational
5) Processing
6) Legal
7) Ethical
8) Technical
9) Financial

The codes that have been grouped under each major concept, have been mentioned in the same thematic expression, but in varying context or meaning.

Irrespective of the industry sector of the participants, the results indicated consensus that the BD deluge and its complexity continues to pose issues which antiquated data governance cannot resolve. The findings in relation to this are presented in both chapter 4 and chapter 5 which we start to discuss under different subsections below.

4.2.1 Big Data Characteristic Challenges

The Information Commissioner's Office provides the definition of BD as "a way of analysing data that typically uses massive datasets, brings together data from different sources and can analyse the data from different sources, and can analyse the data in real-time" (HC648). According to (BIG0057), another participant of the inquiry, what accounts for BD could differ between different kinds of disciplines and industry sectors. This participant maintains BD is more than the generation of large and complex datasets. Contending that it could involve smaller scale datasets characterised with a high-level of variability and complexity i.e. from the capturing and processing of immense amounts of fast-moving data e.g. IoT and environmental monitoring etc. Moreover, they argue that the handling of this data could offer increased complexities where this is personal data or personal identifiable information.

The data characteristics challenges group uncovered during inductive analysis represent the challenges associated with the characteristics of BD. These coded concepts are illustrated in table 8 and discussed further below. The content in the column titled **category** contains the

naming provided to this distinct BD challenge group. In analysing the responses of the inquiry participants, a large number referred to seven distinct characteristics of BD that constitute some of the challenges confronted. Hence during Atlas.ti coding, sub-group codes that represent distinct concepts and nuances within the overarching category were developed. Namely: value (dc: value) → c = 113, volume (dc: volume) → c = 92, variety (dc: variety) → c = 60, velocity (dc: velocity) → c = 38, variability (dc: variability) → c = 13, veracity (dc: veracity) → c = 32, visualisation (dc: visualisation) → c = 6.

Category	Code	Frequency (Grounded)
Data Characteristics Challenges	• dc: extracting value / roi	113
	• dc: variability	13
	• dc: variety	60
	• dc: velocity	38
	• dc: veracity	32
	• dc: visualisation	6
	• dc: volume	92

Table 8: Data Characteristics challenges frequency table

In the frequency column, the frequency count is indicated. Each section in this chapter also contains tables designed to display the frequency of codes (i.e. themes or concepts) that were evaluated during investigation. This aids to reflect the importance given to these challenges by participants as determined by the frequency they are highlighted. Emphasising how grounded the code is, thus indicating some initial view of priorisation that could be given to each concept.

- **Extracting Value**

Many of the participants identified value as a crucial feature, arguing that the requirement to extract business value is at the heart of BD itself. They cite that BD is of no use unless it generates a benefit or provides timely and applicable insights to customer requirements or solves a market or client need. The participants reported on the numerous opportunities

afforded by BD. For example, the British Bankers Association pinpointed that BD was estimated at the time to create 56,000 job opportunities by 2020, which would account for a growth of 160% since 2013. Furthermore, it was estimated that it would contribute a total of £216 billion to the UK economy from 2012 to 2017 (BIG0078). Notwithstanding the obvious advantages identified, many of the participants argued that extracting value from BD in an effective manner was complex. The size of the problem is emphasised by the high code frequency count; value (dc: value) → c = 113. The data analysis conducted also highlighted patterns in the data, which showed that other dimensions used to characterise BD challenges were also linked with challenges extracting value (see figure 9 – network diagram). Related to this discovery, (BIG0060) discussed that 85% of available BD is generated from every device, upload, consumer transaction, tweet, sensor etc. However, the storing or accessing of this data could prove costly, hence organisations need to make decisions on what to store and how long to retain it, insights which we contend is pertinent to consider for the proposed framework. Other participants shared the same perspective, arguing that even when the correct and relevant data is available, extracting value from BD could lead to costly and lengthy analytics or staff being engaged for months on large BD initiatives or data warehousing projects (BIG0075, BIG0065). Additionally, we induced that the speed of changing business needs often could result in these large and complex projects becoming obsolete before completion. This infers that robust governance is required to control cost, resources, implementation and other challenging BD dimensions to ensure that an organisation can extract the intended value from BD. Moreover, to overcome this challenge a key strand of our research output was to detect the importance of considering the strategic nature of this risk factor. Related to this there is a requirement for the design of the BD governance approach to support the internal capabilities and decision processes which negate the levels of organisational risk in this area at both the strategic and tactical level.

Apart from the management, process, technology and other challenges mentioned, participants argue that the introduction of BD must be associated with a return on investment. Failing to obtain a return on investment (ROI) could potentially have financial implications to an organisations balance sheet or hamper their objectives to attain a competitive advantage (BIG0050; BIG0027; BIG0052; BIG0060). EY research was presented during the inquiry, drawing attention to figures that showed 50% of organisation's thought that measuring ROI on BD projects was a top BD challenge along with BD quality (BIG0055). BD projects could be halted by senior leadership teams if there are major differences between the forecasted

value of BD solutions against actual results. Business cases supporting the introduction of BD solutions or initiatives could be challenged by an organisation's decision makers as it is imperative to scrutinise if any BD investment can sustain its competitiveness or create novel business opportunities. For consumers BD must demonstrate value even though the analysis shows that there could be associated cost connected to loss of privacy to contend with. This major point associated to the loss of privacy or privacy infringement, we identified to link directly to other high-level BD risk factors that are identified to hamper organisations deriving value if not governed appropriately. Given the increase of BD acquisition from different platforms and sources, privacy related problems connected to the usage of BD as an asset uncovered a multiplicity of interrelated issues such as evolving ethical, legal and security implications which need to be governed. These interconnected problems will be discussed further in later sections.

- **Volume**

Much was written in the literature regarding the massive volumes of data being generated in today's digital landscape. Next to 'Value' with a frequency count of 113, analysis results depicted the voluminous characteristics of BD to be the next area of challenge in this group with a frequency count of 91. The ubiquitous nature and sheer amounts of the data generated by the Internet of Things (IOT), as well as the proliferation of social media was reported by participants to out speed the advancement of legacy system, infrastructure and their computational methods. On both a business and technology level, the inquiry dataset analysed implied that organisations must tackle core challenges to do with the data deluge. Statistics presented by the government taken from digital insights about the public's daily engagement with its social media platform even as far back as 2013 highlighted the scale of the issue in the table 9 below. This shows that 400 million tweets were posted daily on the Twitter platform, 350 million photos uploaded daily to facebook and over 1 billion videos viewed. Big Brother Watch also raised concerns that the volume of personal information collected and stored by organisations is set to increase due to the continual development of the IOT, so the challenges must be addressed (BIG0006). As a result of the colossal scale of this data, participants reported the problems they face with the acquisition, processing, storage, synchronisation and integration of these large datasets which need to be addressed.

PLATFORM	ITEM	ENGAGEMENT STATS (Per day)
Twitter	Tweet	400m tweets posted
Facebook	Photo	350m photos uploaded
YouTube	Video	1b+ videos viewed

Table 9: Example representation big data volumes

Participants overall supported the standpoint that novel analysis, interpretation and visualisation techniques and tools are required to address challenges caused by its scale. However, examination of the dataset suggests that this goal can really only be achieved where an integrated governance approach to ensure the right people, processes, skillset, financial support exists - and a culture of sustainable capability development exists to support this.

- **Variety**

The third major data group challenge unveiled during analysis was the 'variety' of BD with a frequency count of 60. Variety refers to the various formats of BD organisations have to process, making it problematic to analyse and utilise to make decisions. BD produces diverse forms of structured, semi-structured and unstructured data from different sources e.g. image, video, sensor, biometric data, text and audio. Several participants reported this BD heterogeneity is a huge challenge to comprehend and manage within their organisations (BIG0078; BIG0062; BIG0064; BIG0049). As with the BD feature 'volume' and subsequent features of BD discussed related to this theme, participants raise those enhanced tools, methods and techniques for storing, analysing and interpreting the BD is required. A challenge arises in attempting to evolve BD methods that involve the analysis of the entire complexity of the data, rather than by individual data types. The increase of these different BD format types is reported to have led to insufficient data availability. To support this

argument, it was cited that 80% of data created and subsequently used by companies is unstructured data, increasing at double the rate of structured data (BIG0078). This stage of the study analysis highlighted the relationship between the variety of data and the associated lack of availability, proving problematic for some firms. It indicated a connection between this issue and the criteria to build or improve analytical tools, techniques and methods as shown in the network diagram. However, there are dependencies between management structures, processes and policies that influence these variables. More importantly, analysis repeatedly identified the links between reported BD skill gaps and the lack of capabilities to develop most of the BD tools, techniques or methods to deal with this challenge and others risks.

- **Velocity**

Another major problem uncovered is to do with the velocity of BD. This challenge is due to the high speed of the data and the requirement to manage the rate of heterogeneous data generated from complex networks. These networks include BD produced by the proliferation of digital or mobile devices, which are ever present in the current landscape and are noted to drive the need for real-time analytics and real-time processing of BD. Some of the participating organisations doing business with BD, raised this requirement as an ongoing challenge within their industries. Even though they generally aligned with the view that drawing value from BD is aided by being able to analyse the data in an iterative and timely manner. However, the methods for managing this demand for faster processing of BD analysis platforms, real-time processing for time sensitive and faster analytics results and contextualising its meaning for quick decision making was identified as being extremely complex or in some cases costly (BIG0046; BIG0053).

- **Variability**

Similarly, the inquiry participants recognised that BD issues go beyond variety alone. Observations were made that BD challenges could be amplified by its variability. Not to be confused with variety, variability is to do with the inconsistent data flow and constantly changing nature of BD. Although lower in the hierarchy of other data challenges according to the participants, this feature was identified to be particularly challenging due to the

increasing utilisation of digital media for BD analytics. A relevant example provided during the inquiry was BD obtained from social media platforms like twitter, which generates different types of data which is subsequently mined to conduct sentiment analysis (SMD0023; SMD022). The increased challenges come when mining and attempting to interpret the diverse data types, when simultaneously the meaning of the data is continually and rapidly changing. This leads to additional complexity for BD algorithms and computational processing, to ensure accurate contextualisation of a words meaning.

It was observed that organisations, whether in the private or public sector, must not lack the expertise to govern these variables or to operate effectively with the BD paradigm which has an impact within their operations. Earlier in the study we have unveiled that researcher have propagated the notion that DG attributes to the risk reduction.

- **Veracity**

Considerable attention was given in the inquiry responses to the challenges associated with the veracity of BD. Various participants pinpointed some of the integral discrepancies, inaccuracies, inconsistencies, ambiguity and latency that could be present in BD collected. With respect to the truthfulness, accuracy and quality - the common problem we identified was the challenge of dealing with fabrications, discrepancies and bias inherent from social media networks and web sources, which leads to untrustworthiness of the data and risks when using BD for decision making. One of the participants provides further context to this, stating the availability of BD does not guarantee its fitness for purpose or indeed its authenticity (BIG0053). Within this context concerning the variety of BD Experian stated that "*organisations' BD management capabilities are being put to task for ensuring the data is reliable and accessible for its given purposes – and many are creaking"* (BIG0022). Others raised that appropriate BD tools and analytical methods, need to be developed for the management, processing and mining of data (BIG0001; BIG0004; BIG0008; BIG0022). Within the context of this theme, if BD is to provide value as an organisational asset, the analysis points to the requirement to ensure that the BD is an accurate version of the truth. Participants emphasised the specific requirement to ensure proper governance of this challenge, dealing with the minimum set of processes, policies and procedures to govern the veracity of BD within the organisation. This is also closely connected with the reported

challenge associated with ensuring BD quality linked to accuracy, discussed later in sub-section 4.2.2.

- **Visualisation**

Only five participants raised the BD issues related to visualisation (BIG0047; BIG0048; BIG0057; SMD0004; BIG0075). Notwithstanding its low hierarchy within the data characteristics challenge theme, we include it in our findings due to its close interrelationships with other codes within this category. (SMD0004) argued that a range of BD applications have challenges performing BD visualisation; specifically, due to suboptimal technical performance in response times, functionality and scalability. Speed, scalability and dynamics were all revealed to be major challenges of BD visualisation. Furthermore, visualisation of BD with diversity and heterogeneity is also seen to be a huge problem by these participants. The fundamental goal of BD visualisation discussed in some of the submissions, is to present and convey information concealed in large scale and complex datasets adequately and intuitively. Along with the need to display meaningful results there are of course technical issues to be resolved. Building on this observation, another participant implied that BD visualisation approaches, are more difficult than traditionally smaller datasets. For example, with the new ways of harvesting and ingesting multimedia and multilingual unstructured datasets, BD poses additional issues developing appropriate visualisation and interpretive techniques (BIG0057). This was also said to be linked closely with the big data governance requirements for organisations to ensure the right skillset is available, as well as ensuring appropriate training is in place to develop the right skillset levels of their staff. Significantly, our investigation also revealed patterns and interrelationships related to the legal and ethical issues surrounding BD visualisation, related to new approaches to predictive and sentiment analysis (see figure 9 network diagram).

Figure 10 below indicates the data characteristic challenge themes which were discussed earlier and factors that have an influence on specific areas of BD governance based on the participant's opinions and data analysis results.

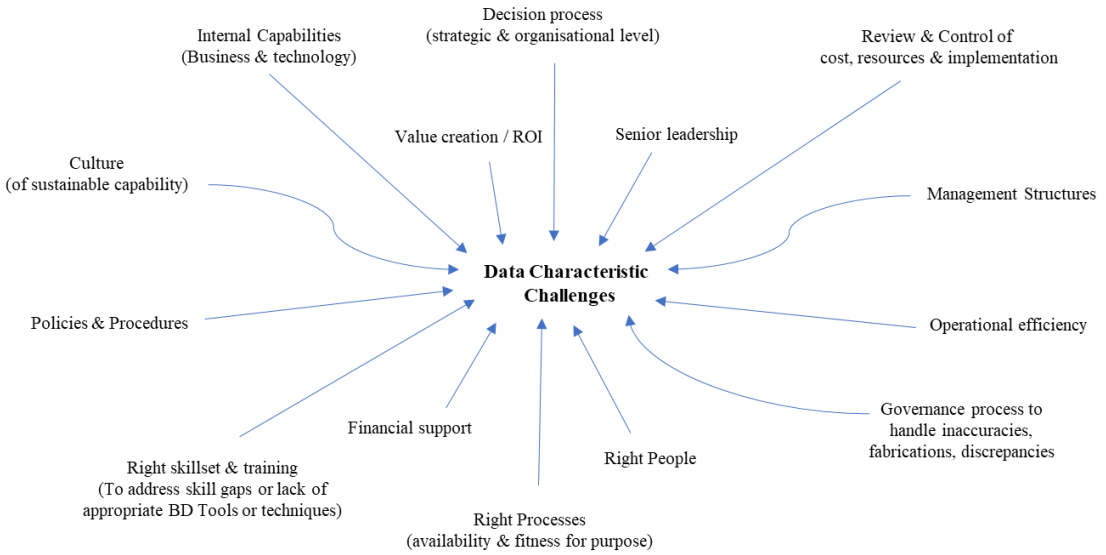

Figure 10: Factors that influence big data governance for Data challenge concept

4.2.2 Big Data Quality Challenges

It was positioned earlier in chapter two that BD is viewed as an organisational asset, but the business value that organisations can exploit from their acquired information pool can be impacted by the levels of data quality. BD must be fit for purpose, or it could lead to poor decision making or failed BD initiatives. Additionally, DQ is vital for compliance, reporting, policy or business processes. Traditional functional goals of DG also refer to DQ controlling (Khatri et al, 2010; IBM, 2007, Otto, 2011b). Given the focus on this area within earlier research, we deemed it imperative to look very closely at participants views in this area to uncover any further information regarding increased BD complexities.

Participating organisations reported about problems with BDQ, which contributed to evidence that big data governance optimisation is needed in this area. Table 10 shows the results of our examination of the dataset and how the distinct themes were coded. It indicates many participants referred to the following core problems that constitute the main areas of concern: BDQ challenges that covered general concerns in this core area but did not specify

particular areas of BDQ concerns (dqc: generic) → c = 59. Gaps in an organisation's accessibility to quality BD (dqc: lack of accessibility) → c = 57, complexities impacting how accurate the BD collected and stored is (dqc: lack of accuracy) → c = 32. Problems concerning bottlenecks which impact the availability, completeness or consistency of BD are (dqc: lack of availability) → c = 15, (dqc: lack of completeness) → c = 9, (dqc: lack of consistency) → c = 17. Finally, the problems related to the lack of timeliness of BD, which is said to be an important concern ensuring BD is available and accessible in a timely manner to make business decisions or provide real-time insights that are fit for purpose (dqc: lack of timeliness) → c = 26.

Category	Code	Frequency (Grounded)
Big Data Quality Challenges	• dqc: generic	59
	• dqc: lack of accessibility	57
	• dqc: lack of accuracy	32
	• dqc: lack of availability	15
	• dqc: lack of completeness	9
	• dqc: lack of consistency	17
	• dqc: lack of timeliness	26

Table 10: Big Data Quality challenges frequency table

Participants revealed that BD quality remains a concern (BIG0063; BIG0023; BIG0006; BIG0048). Ernest & Young argued that research had uncovered that 50% of organisations cited 'data quality' as an issue identifying it is the highest placed issue together with 'return on investment' (ROI) on BD projects (BIG0055). Additionally, analysis results identified that 'value' and 'trust' in BD is often undermined as a result of substandard quality of data at point of recording or collection. This finding highlights an important issue that needs to be addressed by any organisation, as substandard quality of BD will negatively influence an organisations decision making quality.

Findings confirmed that the challenges related to BD quality were multidimensional concepts sometimes arising from the complex features of the data e.g. accuracy, consistency, timeliness, relevance and it being fit for use. Building on this, one participant draws the focus

to four criteria which if not governed effectively could lead to BD quality problems (BIG0006). The content table 11 below outlines the criteria identified.

Criteria	Description
Complete	All applicable data is linked
Accuracy	Common data issues are cleansed (e.g. Typos, misspellings etc)
Available	Required or essential data is accessible on demand. Additionally, users are not required to manually search for data.
Timely	Up to date information is available

Table 11: Big Data quality multi-dimensional aspects. Adapted from Big Brother Watch written evidence (Ref: Crawford et al, 2014. Big Data & Due Process: Toward a Framework to Redress Predictive Privacy Harms)

It was noted that organisations and other stakeholders collect and process BD from multiple sources of varying data quality, resulting in a BD chain. Problems related to BD quality could also lead to additional complexities related to legal and regulatory compliance (HC468b). As shown in the visual concept map in section 3.4 closer examination of the datasets revealed complex patterns and relationships between BD quality and the group of BD challenges discussed earlier i.e. volume, velocity, variety, variability, veracity and visualisation. These complex relationships identified during inductive analysis, emphasised the requisite for big data governance mechanisms to ensure BD quality and to enable contextualisation of BD.

Based on some participants opinion, BD quality risk factors must be one of the proposed big data governance goals. In the main context of big data governance, we drew some foundational inductions from the inquiry output that governance can be one of the main enablers to support an organisations effort to ensure BDQ, by putting in place the right organisational structure for acquisition, storage, processing and exploitation of BD. Essentially, analysis findings determine that big data governance has a direct effect on the concept of BDQ challenges, as participants noted the relationship between the following key governance dimensions to enhance and maintain BDQ: -

- Appropriate strategies to guarantee BDQ
- Identification of roles & responsibilities assigned with a primary goal of ensuring BD is fit for purpose
- Standards and rules
- Policies and procedures
- Technical architecture
- Quality control processes

For example, a few participants mentioned the importance of ensuring the right organisational structure, with the right roles to enable ownership and stewardship responsibilities throughout the BD life cycle. Khatri et al (2010), argued that being cognizant that all data progresses through life-cycle stages is vital to designing data governance. It can be said that although Khatri et al were referring to traditional data, this applies more so to BD given its additional complexities and risks from end to end of the BD life cycle. Therefore, we extend the design of the proposed big data governance framework, incorporating this facet to explicitly reduce end to end risk.

Although big data governance research is in its infancy, some of the earlier approaches for governing data quality also recognised the crucial requirement to enhance, maintain and control the quality of their data assets in order to derive value from it. In order for organisations to derive value and a return on investment in their BD, the data must be fit for organisational use. Otto (2011a) argued that data quality is a critical success factor for organisations to meet various business requirements. Also stating that data quality is pivotal to meet requirements related to integrated customer management, automated processes,

accurate and efficient reporting, compliance etc (Otto, 2011a). This infers the need for monitoring (BIG0052) and measurements (BIG0041; SMD0008) and KPIs for BDQ, with some participants highlighting the need to have the right skills and resource capability in place to: -

1) Develop appropriate metrics to assess BDQ

2) Undertake any appropriate remediation actions when issues are identified.

We argue that organisations must accomplish effective governance processes in order to attain competitive positions through BD. The efficiency of these processes is measurable through defined metrics and key performance indicators (KPIs). KPIs for BD are vital to organisation leader's decision-making and should be relied upon at different levels of an organisation to measure whether successful outcomes have been achieved. An organisation needs to be able to identify its high-value high-risk BD quality issues that support each KPI. Thus, our framework supports ongoing monitoring and audits, as well as the mapping of BDQ dimension to KPIs to provide additional confidence to decision makers.

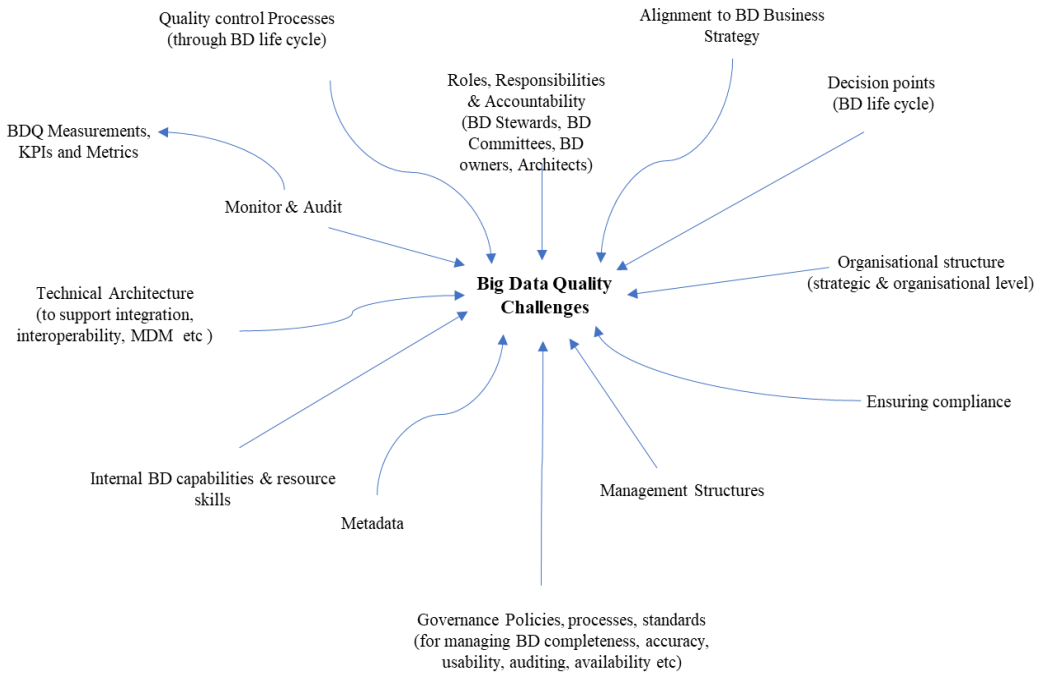

Figure 11: Factors that influence big data governance for the Big Data Quality challenge concept

Figure 11 above indicates the BDQ challenges category which was discussed previously and the factors that have an influence on specific areas of BD governance based on participant's opinions and analysis findings. Moreover, based on earlier research foundations and our subsequent inductive analysis findings - we align to the view that it is vital for organisations to include as part of their BD strategy, a core objective of maximising BD quality. Consequentially, it is a high-level priority within our novel BD governance framework.

4.2.3 Management Challenges

A thematic analysis of the datasets exposed another core group of BD challenges, classified as management challenges. These challenges are connected to a group of issues and risk factors confronted whilst acquiring, accessing or managing BD, and which participants identify as requiring robust BD governance to address. To a great extent most organisations from both the public and private sector that participated in the inquiry considered BD challenges from this category as being highly significant. Table 12 below reflects that amongst the themes uncovered, this conceptual group was within the top three problems unveiled. As the frequency count in the table indicates, majority of the participants asserted that the rise in the scale and diverse channels of BD have inevitably led to BD management challenges around privacy, data protection and security. Whereas some participants placed considerable emphasis on problems associated with inappropriate BD usage and sharing practices. Additionally, some organisations called out the management challenges associated with poorly implemented projects leading to white elephant BD initiatives tying up resources and budget. Participants supported the view that addressing these areas through governance are critical to deriving actionable insights from BD.

Category	Code	Frequency (Grounded)
Management Challenges	• MC: = MANAGEMENT CHALLENGES	
	• mc: generic	20
	• mc: access rights	38
	• mc: data protection	117
	• mc: data protection - limitations of consent	136
	• mc: data protection - limitations of de-identification	77
	• mc: data protection - personal info leakage / breaches	55
	• mc: inappropriate data usage / sharing	64
	• mc: minimising unnecessary collection	13
	• mc: poorly implemented projects	5
	• mc: privacy concerns	233
	• mc: security of BD / BD systems	152

Table 12: Management challenges frequency table

In the next sections, attention is centred on the top three BD management challenges identified by inquiry participants i.e. related to privacy, data protection and security.

- **Privacy Concerns**

One of the predominant findings from the datasets established that challenges related to privacy were an overwhelming concern for most of the participants of the inquiry. The frequency count for this concept was high on the hierarchy of concerns (mc: privacy concerns) → c = 233. The themes and relationships examined painted a rather broad picture, revealing the challenges associated with privacy were multifaceted and multi-dimensional; including issues that were behavioural, technical, ethical and legal in nature. It appears that the research analysis validates debates and assumptions about the need to enhance designs of existing data governance to cater for BD holistically.

Participants from both the private and public sector organisations extensively discussed this concept, reporting that amongst the highest profile risks of developing and implementing BD

technologies was the risks to personal privacy from the processing and analysing of personal data. Furthermore, there was substantial evidence that this challenge indeed has multifaceted dimensions which are technical, ethical, behavioural and legal in dimension. A finding that was evident during examination of several of the core concepts which are proposed as part of the framework.

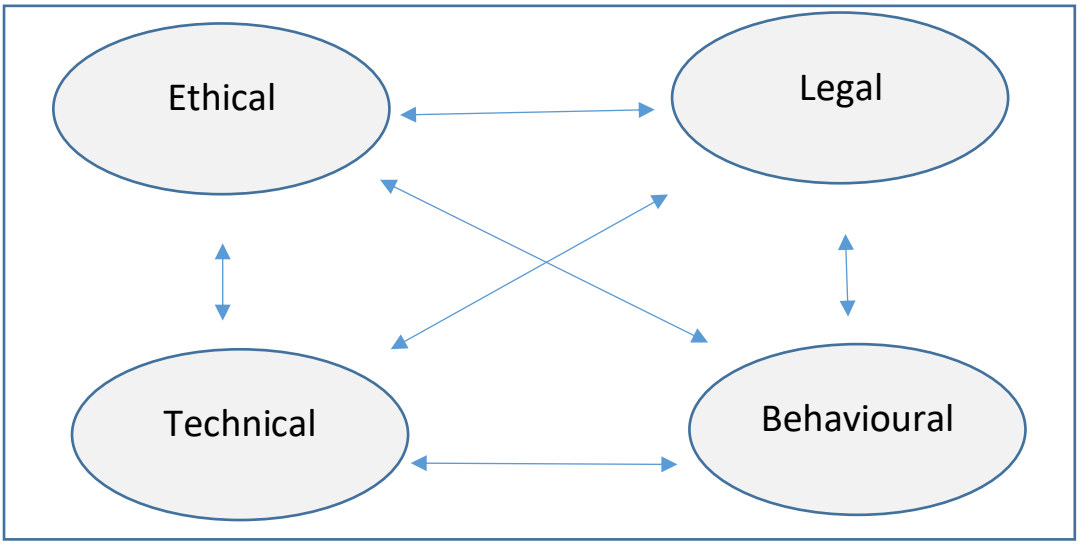

Figure 12: Aspects of Multifaceted Dimensions of BD Challenges as reflected by participants

Various scenarios were discussed regarding the negative ways BD technologies infringe on people's privacy. Respondents expressed concern about the consequences of potential breaches of privacy, citing some examples of past breaches (BIG0006). One of the participants raised that the ubiquitous use of BD platforms such Facebook, Twitter etc, means that individuals disclose personal information such as photos, videos etc irrespective of the risk of the data being utilised maliciously or negatively (BIG0033). For instance, BD can be created without the consent of the concerned person, this includes travel route information, data created whilst searching the internet etc. Due to the proliferation if social networking platforms, digital devices and smart phones, various participants argued that personal data is becoming increasingly available, leading to significantly increased risk of personal information leakage and privacy infringement due to collection of BD. Additionally, several highlighted public concerns about their personal data in these domains being used by businesses for BD analytics purposes such as advertisement targeting, or by national security

agencies for surveillance purposes or to predict a user's intentions. (SMD0001) highlighted that there is evidence of common privacy concerns across demographics. According to Renata Sampson her organisation (Big Brother Watch) conducted a poll in the same year as the inquiry and discovered that 79% of individuals that took part in the poll were concerned regarding online privacy. Moreover 74% of these individuals wanted control of their personal data and only 25% were not concerned regarding online advertisement being tailored to them (HC468b). Two participants discussed a related concern of function creep, noting that some organisations introduce the gradual widening of the use of BD beyond the originally intended purpose of collection (SMD0024; SMD0001). Elaborating that this leads to potential invasion of privacy. Another participant argued that some organisations use of BD is conducted in the context of a "Privacy paradox", which refers to individuals willingly participating in sharing their personal data over online social networks, whilst disagreeing to government bodies or enforcement agencies requesting, compiling or analysing their personal data (SMD0003). Suggesting that a significant number of the public would prefer to have greater control of their personal data. This suggests the need for improved opt-in strategies given they are uncomfortable about their personal data being mined and tracked through BD analytic practices without their knowledge.

Crucially, during analysis we induced that the probability of each of these privacy risk-factors and their potential impact should be investigated. This leans to a requirement for a systematic big data governance approach to embed privacy through risk assessment and privacy impact. The organisation should subsequently prioritise mitigation actions around these in their governance approach. It is induced from participants that both the business units and IT departments should consider risks individually, ensuring that there is formal BD governance with the appropriate privacy access controls, standards, policies, data protection and ethical safeguards put in place, as well as enforced. Similarly, results uncover that for BD management challenges, there are two factors that create risk: internal and external. Participants revalidated the necessity for organisations to have specific governance frameworks to address these BD dimensions, particularly as there is evidence that both organisations and individuals are unsure where the line is drawn when privacy is opened up online. Further conclusions drawn from the thematic analysis is that there are sizeable, interrelated challenges enforcing compliance, minimising unnecessary BD collection, as well as adhering to legal and regulatory rules. This was emphasised by the themes, patterns and

linkages that emerged interrogating the data. A major argument of this thesis study is that ethical governance must also play an increasing part within the proposed big data governance framework to address some of the privacy concerns the participants highlighted. Subsequently ethical governance to support appropriate BD collection, usage and sharing is established as part of the proposed big data governance framework. Figure 13 shows the 'Privacy challenge' subgroup within the Management challenges category and the factors that have an influence on specific areas of BD governance as identified during analysis.

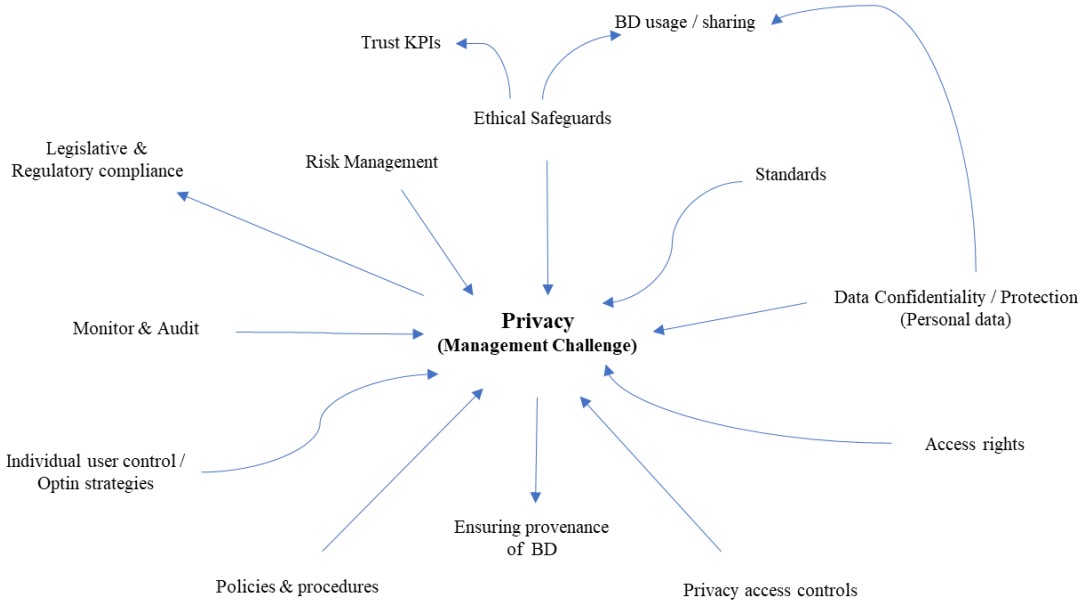

Figure 13: Factors that influence big data governance for the Privacy subgroup of the Management challenge concept

- **Data Protection**

Another major revelation from the results of the analysis was connected to Data Protection challenges. We coded this theme into subgroups as the findings found that participants discussed various distinct challenges within the context of the overall theme. The frequency counts for this concept and its subgroups shown in table 12 were again very high on the hierarchy of challenges outlined (mc: data protection) → c = 117, (mc: data protection – limitations of consent) → c = 136, (mc: data protection – limitations of de-identification) → c = 77, (mc: data protection – personal info linkage / breaches) → c = 55. These were

discussed extensively by participants in written and oral evidence, with organisations establishing that public and consumer privacy, increased the demand to have robust and trusted governance safeguards in place.

Significantly, deeper analysis of the dataset uncovered data protection principles link to numerous problems associated with BD quality. For instance, issues identifying the provenance and ownership of data from BD sources (BIG0077; SMD0008; BIG0041; SMD0022), handling of sensitivity and confidentiality of personal data (BIG0088; BIG022; BIG0045), the access rights (SMD0015; SMD0013) and permissions that come with it and how long to retain it for (BIG0006; SMD0004). These are interesting observations given that regulation for data protection already exists e.g. GDPR. However, research conducted by one of the organisation participants Ernest and Young, which was presented during the inquiry unveiled that 17% of companies were concerned regarding data protection regulation complexities and the potential risk of non-compliance, possibly resulting in huge fines or corporate damage that impairs value or corporate reputation (BIG0055). Further interrogation of the dataset highlights those participants identified various reasons why BD protection is more complex to comply with e.g. the added difficulties complying with cross-jurisdiction regulations during to some of the features of BD etc. The line that is hard to control is where BD discovery, collection and usage is being opened up online. The proposed big data governance framework takes this requirement and others into account in its conceptual design, including the requirement stated by participants for strategies and policies to facilitate data protection regulatory compliance and curation through the BD life cycle.

A further BD risk factor unveiled during the investigation is associated with the identification of personal identifiable information, either through aggregation or data association. Professor Montgomery of the Nuffield Council of Ethics argued that there were various ways that traditional safeguards are less effective (HC468b). For example, the suggestion that anonymisation of data is straight forward is no longer the case with BD. Several respondents raised the risk of re-identification, which is the ability to link several datasets together, so that data which was unidentifiable in possession of the original producer, becomes identifiable once combined with other BD sources or publicly available information (BIG0067; BIG0089; BIG0015).

Another major challenge divulged by participants was the issue of informed consent and permission management. Obtaining genuinely informed consent at each data point was said to raise problems. For example, the question of what happens to informed consent if the data is linked with other datasets to produce new datasets which was not agreed to when shared, or if those utilising the data were not involved in the process of initial BD collection. Various participants contended that individuals providing BD, are unable to foresee the potential consequences of how the BD would be made available for linkage or reuse, suggesting there needs to be greater transparency of motive for BD usage and sharing by organisations. According to (BIG0057), *"informed consent has been a cornerstone of ethical research involving people",* stating that *"gaining genuinely informed consent from individuals relating to each individual BD point is difficult if not impossible"* as there are logistical challenges connected to contacting individuals each time data is utilised. Notwithstanding, some participants made it clear that traditional consent models are no longer fit for purpose when dealing with BD.

GDPR has since been introduced in Europe in May 2018, however big data governance must extend more holistically within organisations to integrate data protection with other evidently interrelated big data governance requirements (BIG0046; BIG0045). Moreover, GDPR is specific to Europe and global operating organisations must have governance in place to comply with data protection for several jurisdictions. Participants from all sectors recognised the critical need to have the right legislative and regulatory frameworks in place. However, they raised concerns about interpretation of the legislative boundaries on how BD can be used, particularly outside the EU. Interestingly, a number of businesses identified that the approach to BD sensitivity must also strike a balance between mitigating risk and encouraging innovation. In addition, the general principles of regulations introducing an increased level of burden on firms was a common problem raised. Organisations would need to also investigate as part of their BD strategy what the logistics, financial and time implications are when dealing with BD protection.

GDPR in Europe and other regions disparate data protection regulation plays its part in combating some of these issues in terms of legislation and regulatory aspects; however, our detailed interrogation of the dataset, builds the case for enhanced processes, procedures,

standards and safeguards to support the mitigation of BD management challenges within our conceptual governance framework. Neither data protection methods such as obtaining consent or de-identification methods are sufficient in isolation to protect a person's privacy or personal interests. Neither are legal and security methods, if not addressed holistically. Thus, with the IOT, organisations must address their organisation's BD governance controls to assist with dealing with increasing challenge of dealing with the cross-border phenomenon. Figure 14 reflects the data protection challenges subgroup and the factors identified as having influence on specific areas of BD governance based on participant's opinions and analysis findings.

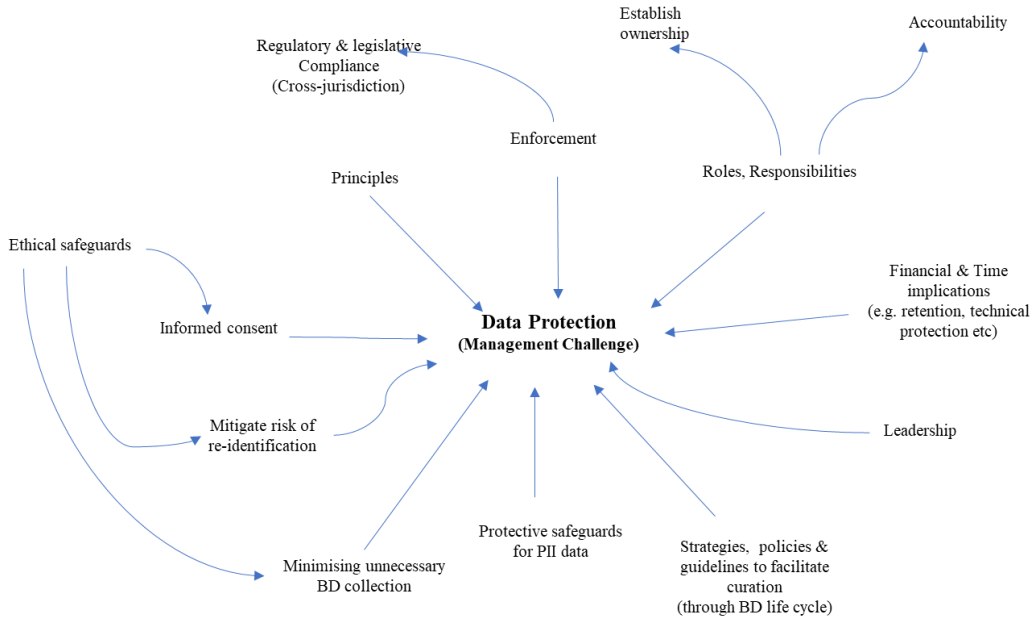

Figure 14: Factors that influence big data governance for the Data protection subgroup of the Management challenge concept

Finally, during our close examination of the datasets, we obtained confirmation that the management challenge theme related to BD security differed from BD data protection in two distinct ways. Whilst both are related to the safeguarding of BD, data protection correlated extensively with personal data. Secondly, public concerns about privacy and the sensitivity of personal data, imposed additional requirements beyond that of data security. As a result, our proposed conceptual framework recognise that these are two distinct components to be governed. BD security is discussed separately in the section below.

- **Big Data Security Challenges**

Findings determined BD security as a core issue for many of the participants (mc: security of BD / BD systems) → c = 152. There was general alignment that if organisations do not address security challenges within the BD domain, then it would negatively impact public trust in an organisation's BD systems, people and processes. Consequentially, impacting the ability to extract value from BD as an asset. Participants reported implementing environments for storage, aggregation and analysis of BD. Subsequently, BD warehouses and systems have risen in number, and with this a rise in security concerns.

Challenges of traditional data are not extremely distinct from BD. However, a high number of the participants contend that there are some distinct challenges with BD which are fundamentally different from traditional data (BIG0022, BIG0055). These challenges are observed to be incremental, rather than fundamentally distinct differences between BD and traditional data security challenges. For example, the variety, velocity and volume features of BD heighten security challenges. Furthermore, the BD landscape is a large-scale distributed system and extremely vulnerable to attack hence it is contended that traditional forms of DG must be optimised to factor in aspects of BD that require heightened security. According to (BIG055), 44% reported that BD elevates data security risks. Stating that by executing organisation wide governance, companies can add notable improvements in each core BD problem area.

Cybersecurity is also reported widely as a significant challenge, with participants discussing the risks posed to individual's privacy, personal security, as well as physical, national and intellectual property. Various respondents reported that it was a continual battle to secure people's identities, or safeguard against illicit transactions, malware attacks and hacking, or even guard against national security attacks from foreign governments. There were reports that criminals had targeted BD warehouses for huge returns and data breaches could cause companies enormous losses and reputational damage (BIG0003, SMD0023). Renate Samson (2015) discussed that in the same year as the inquiry, 90% of UK companies experienced some form of attacking attempt against their systems. Given the distributed characteristics of BD, this suggests that inadequately secure BD technologies and infrastructure can magnify problems related to availability, integrity and accountability. Participants point to the vital requirement for big data governance to ensure access through safe settings and security

controls to ensure organisation systems are resilient to alterations. In addition, having adequate skills, funding, guidelines, policies, training, as well as defined roles and responsibilities in place are also identified as important BD governance success factors to prevent security problems. (BIG0039) contends that the full opportunities of BD can only be achieved if consumers are confident to share their data. Therefore, ensuring BD is secure and personal data is controlled appropriately and transparently is key to maintaining consumer trust and therefore the future of BD.

(BIG0032) stated that trust in organisations to handle the data of individuals in the correct way is at an all-time low. If BD is kept in facilities that are insecure or subject to breaches at the hands of people with questionable agendas - public and consumer trust in the organisation holding the BD will diminish. Maintaining governance practices which enforce and maintain security of the BD both technically and in terms of human access is essential for public trust. The thesis study sets out guiding principles for transparent usage of BD, as well as the requirement for processes to empower individuals to manage, share and control their BD. Robust big data governance that supports cybersecurity, prevents data leakage and keeps BD secure is an important aspect of the proposed conceptual framework based on the findings. Figure 15 below reflects the BD security challenges subgroup and the factors identified as having influence on specific areas of BD governance based on our inductive analysis results.

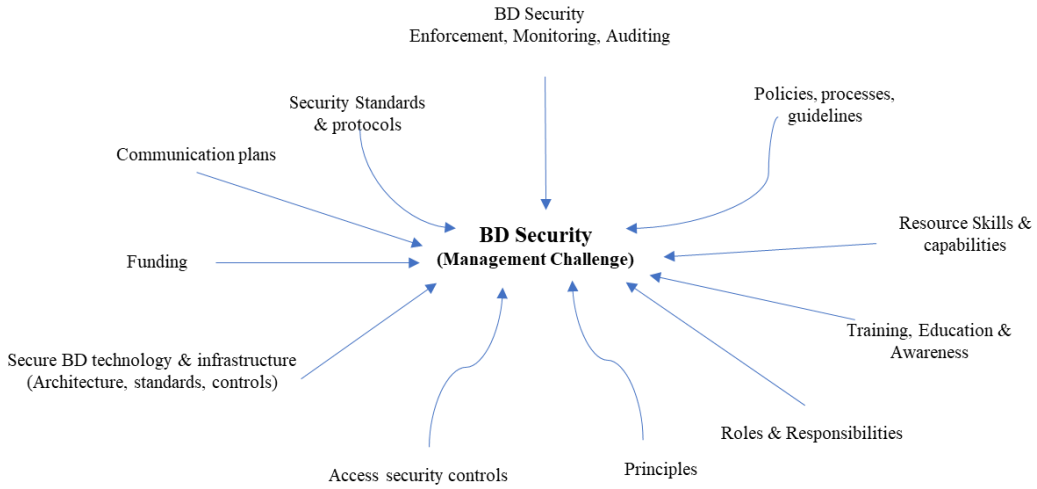

Figure 15: Factors that influence big data governance for the BD security subgroup of the Management challenge concept

4.2.4 Legal challenges

The group of legal challenges with respect to BD paints a very complex picture. Table 13 below presents the coding of this group of challenges that were considered worth pointing out separately and the frequency count of these specific themes. As shown where the dataset referred to legal challenges, without contextualising the problem, this was combined in the overall group frequency count (LC: = Legal Challenges) → c = 60. Other legal challenges identified during analysis were related to items such as: Interpreting and enforcing compliance to data protection, as well as legislation protecting against privacy infringement (lc: data protection & privacy infringement) → c = 93; complexities related to BD ownership and intellectual property (lc: data ownership / intellectual property rights → c = 71; Complexities contracting for BD and developing transparent terms & conditions for BD (lc: contracting for BD & data regulation) → c = 32; Abiding to and enforcing legislation related to BD which crosses country jurisdiction (lc: abiding/enforcing cross border legislation) → c = 33; civil liberties and human rights problems e.g. for example, where BD is mined for profiling or surveillance purposes etc (lc: human rights legislation) → c = 19; Complexities associated with complying to multiple regulations (lc: legal regulatory framework issues) → c = 27; Finally, the insufficient penalties for BD misuse and reported violation of legislation which is also seen to be an issue (lc: insufficient penalties for BD misuse) → c = 29.

Category	Code		Frequency (Grounded)
Legal Challenges	•	LC: = LEGAL CHALLENGES	60
	•	lc: abiding/enforcing cross border legislation	33
	•	lc: contracting for BD & data regulations	32
	•	lc: BD ownership/intellectual property rights	71
	•	lc: data protection & privacy infringement	93
	•	lc: human rights legislation	20
	•	lc: insufficient penalties for BD misuse	29
	•	lc: regulation lagging behind technology development	14
	•	lc: regulatory framework issues	27

Table 13: Legal challenges group frequency table

The inquiry participants cited burgeoning legal issues connected with balancing privacy, anonymisation, security and public benefit. Explaining that the BD explosion has led to the emergence of various sources of legal regulation that regulators and organisations are finding difficult to keep pace with, given the evolving BD environment. Findings reveal BD technologies and tools producing BD were intensifying the pressure on legislation, and there is evidence that current legislation is lagging behind the rapidly evolving BD technology. In most cases, the rapid evolution of BD technology is outside the control of the country in which the organisation operates. For instance, modern developments such as cloud infrastructure, add further complications which might lead to BD being stored outside local or national jurisdiction. At the same time, there is not a one stop shop legal mechanism where BD processing or violations can be dealt with across borders. One participant goes further to state that there is confusion between UK, European and global data protection legislative regulation (BIG0077). Another participant raised that it is extremely concerning when personal data is being transferred to a county or territory outside of its jurisdiction (SMD002), suggesting that governments should think about creating embassy data centres or special economic zones for BD projects. It can be said that these issues cause additional pressure on global operating companies dealing increasingly with a BD cross-border phenomenon. It poses important BD governance considerations, raising questions regarding whom has access rights, about retention and how to ensure the organisation is complying to the different legislations.

The BD ecosystem is based on a complex ecosystem of numerous BD sources, BD users, BD suppliers & vendors etc. Consequentially, contracting for this domain is a challenge. For example, contract structures that licence, moderate and assign risk around BD use constitutes problems for most organisations in a BD market that is said to have inappropriate legislation (BIG0027; HC 245). Moreover, meeting different BD service requirements constitutes problems, as some businesses are not configured to accommodate the various types of clients with versatile, tiered levels of BD service level agreements (BIG0057). A participant contended that there is also distrust in unreliable and incomprehensible terms and conditions, which also cause issues for reputable companies (BIG077). Furthermore, a significant number of participants claim that existing legal regulatory frameworks are sometimes seen as complex or poorly understood. This means that there are compelling requirements to get the legal framework for BD right. Whilst legislation is seemingly lagging behind technology, we

support the argument of a large number of participants that the role of BD governance becomes even more vital. Although participating organisations struggle to develop comprehensive contracts, terms & conditions, it is still important that the end user is informed about how their BD is collected, stored or shared. Hence BD governance should facilitate cross-functional alignment on how these aspects are controlled within the organisation.

The findings revealed a strong link between legal and ethical challenges. New approaches to sentiment analysis, multimedia exploitation and new BD business models, led various participants to call for a framework that curbs data monopoly and government agencies from exploiting BD. Nonetheless, we identified differing perspectives related to this area. Some participants contended that the introduction of strict legislation would cause a large overhead and introduce complex processes for companies that could potentially hinder accessibility to BD. These participants said that care should be taken introducing draconian legislation, as it could be argued that there is a necessity to strike a balance between the rights of the individual versus the public benefit to be gained from BD. This explains why as a criterion, BD governance must enforce transparency and compliance without hindering the ability to exploit BD and attempt not to introduce increased overhead to the organisation (BIG0039).

The more complex area where respondents had significant points to raise was related to BD ownership and intellectual property right. This came top of the hierarchy of legal concerns cited by participants, with results unveiling that BD ownership is a confusing concept for a variety of reasons. There were strong views around the issue of who owns an individual's data. Dame Fiona Caldicott stated in her submission that medical records for instance were not owned by individual UK citizens, but by the Secretary of State (HC468b). Conversely, Chris Combemale the CEO of Direct Marketing Association stated that research conducted by his company, revealed that 82% of customers were of the opinion that the data belonged to them – including their shopping habits and patterns (HC468b). Stella Creasy provided an example that individuals cannot go to Sainsbury's or Tesco's to request their data in order to make like for like comparisons. One participant stressed that as commercial value of BD rises, the questions on who owns it becomes contentious and open to debate (BIG0049). We recognise the fragmentation of BD collection, ownership and control adds to legal ownership issues. For example, the large datasets collected by Nectar card data is owned by Aimia, thus

should be attained with approval from the firm, as well as other firms that collected the data e.g. Homebase, Sainsburys, Expedia etc (BIG0013). Additionally, for social media platforms more questions can be raised based on the analysis findings e.g. who owns the data the citizen, firm or state? Likewise, for BD types such as social media, people co-create pieces of content through a chain of metadata, posts and tags, as well as appearance on group photos or videos posted on the platforms. This raises another need for radical and novel governance practices for managing intellectual property rights in this new manifestation. We argue that organisations must implement governance approaches in response to growing use of personal data in the BD paradigm and protect subsequent negative impacts of this practice to lives of citizens. Participants highlighted individual's concerns about their inability to control BD usage or sharing legally (BIG0080; BIG0065). This emphasises the requirement to introduce governance procedures that empower the individual, not just introduce controls that are practiced internally.

Another grey area is the ownership of data after a user's death, raising ethical concerns in the use of personal data for BD activities. There appears to be a lack of clarity about how the data of a deceased person could be appropriately protected and maintained in a transparent and accessible way. This situation seems to mirror a lack of clarity in legislation both in the UK and globally. Although legislation is starting to respond to the concept of digital death, these responses were said to vary globally, hence requiring robust BD governance approaches to address. Article 17 in the EU Data Protection Directive includes the "Right to Erasure", which gives individuals the right to have content deleted at any time during their life or after upon request, if it identifies them directly or indirectly (SMD0019). There are variations of this legislation in the US States. In Virginia, legislation is specific to the death of a minor. A representative has the authority to take over the deceased minor's online account. However, in Oklahoma, the executor of an estate can take over or terminate the "social media and certain other digital accounts of the deceased, regardless of their age (SMD0019). Such variations in legislation add to this already complex situation, where the content associated with an individual could also be stored in disparate countries globally and controlled by different ISPs, in various legal jurisdictions.

Crucially for participants from all sectors, results reinforce the criticality for BD governance to assure that organisations are complying with that myriad of legislations and address areas where current legislation is inappropriate. Policies and guidelines must be in place to address any potential gaps that could harm the reputation of an organisation and consider how they will legally collect, store, publish, share and analyse BD. Crucially, respondents highlighted that consideration needs to be given within organisations to the legal basis of collection, storage, publishing, sharing and analysis of BD (SMD0014; BIG0065). Participants also identified the requisite for organisations to apply the right cross-functional organisational structure, processes, rules and guidelines to comply with applicable BD legislation and regulation, as well as enforce penalties and sanctions to resources that do not comply. It is a vital requirement to get the right organisational structures, with the right roles and responsibilities and the right departments working together. There was consensus that governance of BD is not just a technology department problem (BIG0048; HC245). Business units across the business must play an important part in BD governance, including the legal department. The locus of control must extend beyond the technology department with cross departmental BD councils and committees.

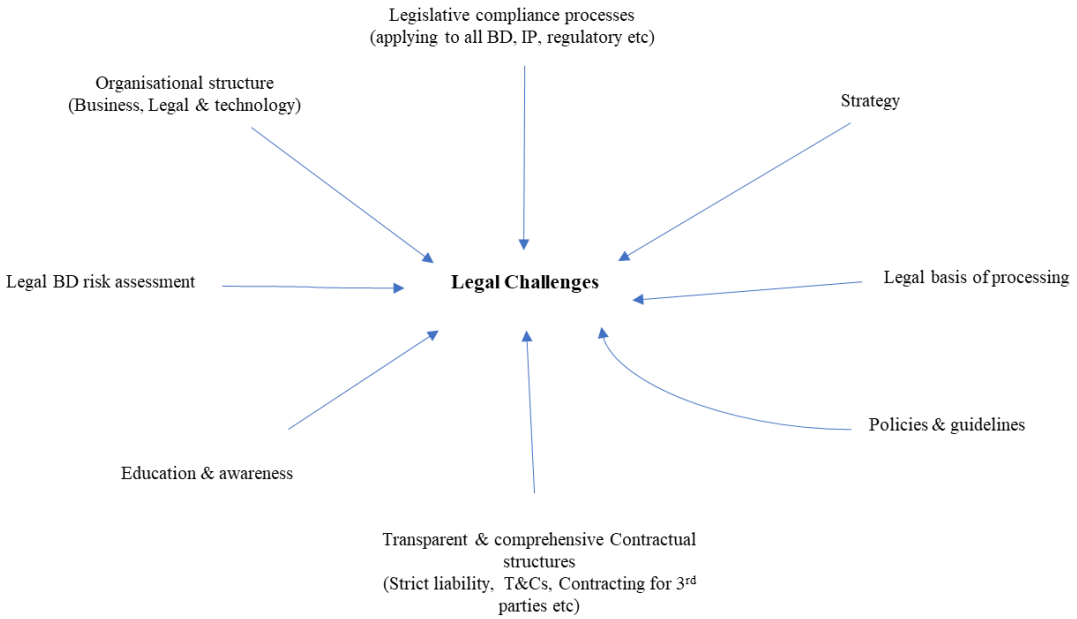

Figure 16: Some Factors that influence big data governance for the Legal challenge concept

4.2.5 Process Challenges

In this research study the process challenge group is classified as the problems confronted whilst processing and analysing BD from the point of acquisition to presenting the output to the end users. There was evidence of clear consensus by participants that the huge volumes, variety, velocity and variability of BD pose significant process challenges. Table 14 below shows the themes revealed during dataset analysis:

Category	Code		Frequency (Grounded)
Process Challenges	•	PC: = PROCESS CHALLENGES	14
	•	pc: aggregation & integration	42
	•	pc: analysis & modelling complexities	63
	•	pc: BD acquisition, storage & retention	79
	•	pc: BD mining & machine learning barriers	62
	•	pc: ensuring the provenance of refined BD	29
	•	pc: interpretation complexities	23
	•	pc: metadata	30
	•	pc: visualisation & provision of results	8

Table 14: Process challenges group frequency table

Table 14 specifies the coded themes produced from the thesis findings. The diversity of BD sources brings abundant BD types, complex data structures and increases the difficulties related to aggregation and integration (pc: aggregation & integration) → c =42. The need to understand its meaning makes it complex to routinise the work and embed the use of BD in operational processes (pc: interpretation complexities) → c = 23. In addition, BD analytics and modelling techniques brings about its own processing challenges (pc: analysis & modelling complexities) → c = 63, along with challenges related to the collection, warehousing & retention of BD (pc: BD acquisition, storage & retention) → c = 79. Organisations increasingly experience barriers utilising datamining and machine learning techniques (pc: BD mining & machine learning barriers) → c = 62, problems ensuring the provenance of refined BD (pc: ensuring the provenance of refined BD) → c = 29, problems processing BD due to the lack of accurate metadata (pc: metatdata) → c = 30, processing challenges related to visualisation and provision of results (pc: visualisation & provision of results) → c = 8.

We delve further into these processing challenges unveiled by participants in the sections below.

- **Aggregation and Integration**

A common barrier to the process of using and sharing BD that was unveiled is related to complexities with aggregation and integration (pc: aggregation & integration) → c =42. There were a number of scenarios that were given as examples. Participants reported that in various organisations BD is siloed, making it problematic to share complete data. It was noted that organisations are grappling with issues transforming BD into agile and actionable insights due to this problem, adding that few organisations have robust capabilities when it comes to utilising multichannel, as well as cross device BD for real-time personalisation and decision making. BD analysis is an increasingly sophisticated process; with some analysis results indicating that the decisioning process and the provenance of BD are also issues for aggregation and integration. The complexity of interactions between different channels that BD is acquired and analysed from are increasing and it is determined that participants are struggling to navigate this challenge. It was observed that it poses interrelated complexities to automating integration and aggregation. From BD application to BD databases there is sometimes the absence of any binding information, and aggregation of this BD is beyond the capabilities of traditional data systems (BIG0018). Frequently, there are no unique identifiers that assists the linking of an individual's information at the correct level (SMD007; BIG0016). This explains why one participant suggests that for BD to maximise its impact, incoming BD must be annotated, validated and integrated. Similarly, there is the requirement to not only be able to implement systems and infrastructure that deal with BD in an interoperable and flexible way, but these capabilities need to also be integrated into the decision-making processes and services. Big data governance addresses decision making related to how organisations govern and control data as an asset, so our analysis infers that this is imperative to be addressed as part of any big data governance framework. Significantly, one of the participants said that BD must be championed from the Boardroom. Arguing that failure to put in place the correct organisational structures, processes and governance frameworks, leads to a siloed approach to BD deployment that limits an organisations capability to discover, measure, create and protect BD value across different operational areas complained about by other participants (BIG0055).

- **Metadata**

Due to the heterogeneous nature of BD and the diverse types of representation and meaning, we unveiled it is not just enough to capture and store BD. According to several participants adequate metadata should be used to resolve processing issues related to the contrast in hosting BD record structure or experimental details. These participants were aligned that where there is insufficient or incorrect metadata, this adversely affects integration or the interpretation of BD (pc: metadata) → c =30. Metadata is an important decision domain identified in earlier literature which enables a method for accurate and dependable description of the data being collected, thus assisting in interpretation of data (Abraham et al, 2019; Khatri et al, 2010; Mao et al, 2021). This suggests that this is an even higher-level criteria for big data governance than participants identified. To be able to capitalise on BD so that it can be utilised for driving business decisions, relies on organisations being able to effectively categorise and organise BD. A primary issue observed is that large amounts of BD is unstructured, and a huge proportion of BD assets can be stored in company silo's that can be held across multiple business units. To add to this complexity both unstructured and structured BD is collected in varied internal and external BD systems, applications, networks and smart-connected devices. Participants recognised that an organisation needs to have the right processes to grow metadata so it can be worked out what the BD is telling you, but metadata hardly comes up as high priority for organisations as induced during analysis. Khatri et al (2010) attempted to enforce meta-data of different kinds e.g. physical, domain-independent & domain-specific, and user metadata have a vital role for discovery, recovery, collation and analysis of BD. Significant for addressing the goals of this thesis, findings als suggest that defining procedures and codes of practice for BD, including the requirement for adequate, explanatory metadata to also protect against the intentional misuse of BD.

- **Ensuring the Provenance of Big Data**

Closely linked to standardisation, but worth calling out separately in this section, is the process challenge associated with ensuring the provenance of BD highlighted by some participants (pc: ensuring provenance of refined BD) → c =29. The frequency count is moderate in relation to other categories; however, interrogation of the dataset revealed a high level of interconnectedness with other BD challenges identified. A primary challenge with using BD analysis results comes from large discrepancies inherent in BD acquisition, as well

as the lack of BD provenance. (BIG0071) claims it is vital to establish BD provenance and ownership, with (BIG0045) also stating that organisations must be able to depend on the accuracy and provenance of BD if it is to be public facing. (BIG0041) supported the view of both these participants, adding that to ensure the quality of BD systems any measurement made by network node should not be ignored as all measurements done by a network node is worthless without being accompanied by a unit of measure and estimate of measurement uncertainty, as well as sufficient metadata to comprehend the conditions of measurement. In other words, being able to track BD right back to its origin is seen as key to generating trust based on a number of participants experience (BIG022; SMD0015; BIG0043). To minimise the risk posed by this challenge; big data governance processes and procedures that track changes that are made to BD, where BD originates and moves to, and who makes changes to it over time are argued to be important big data governance considerations. This historical record of information can then be trusted for BD validation and audit purposes. Moreover, sufficient analytical algorithms and methods are cited as required to process the streams of BD and understand to understand its provenance. We argue that there is an interconnected requirement for adopting big data governance standards and architecture, as well ensuring big data governance policies and procedures around this to mitigate against organisations ignoring these criteria. This is discussed further in section 6.4 of the thesis.

- **Acquisition and Storage of Big Data**

The participants recognised that individuals had concerns, related to how BD is acquired, utilised and stored (pc: BD acquisition, storage & retention) → c = 79. The acquisition of BD from a myriad of sources e.g. ubiquitous mobile devices, remote sensors, cameras, wireless sensor network, tweets etc was reported to add to these concerns, with the BCS arguing that it has been established there is a difference around how BD is stored, used and deleted (BIG0038). Particularly in relation to the processing and storage of personal data for BD initiatives (HC468a). As soon as BD is gathered it must be stored. However, there are deemed to be interrelated risks related to 1) whether it is encrypted 2) whether it is stored securely in servers within the country or facilities external to the country BD was collected 3) How access and permission management will be handled.

Furthermore, analysis findings emphasised that there is additional processing complexity that originates with this kind of data. The massive amounts of BD generated by IoT and BD types

such as social media, out-speeds the advancement of legacy infrastructure and its computational methods (BIG0015; BIG0040; BIG0043). It is contended that this is because traditional data storage systems were not designed for real-time BD analysis (SMD0007). Applying filters or on-line analytical algorithms to process streaming of BD and reduce BD prior to storage is required due to its complexity. This is known to sometimes moderate the velocity and resolution at which BD is gathered and subsequently stored. As uncovered in prior categories discussed, this last example also points to the interrelated processing complexities associated with the BD analytics practices, including inconsistencies, incompleteness, scalability, timeliness and security. Various participants insist that the lack of experienced staff is a vital factor that must be considered when examining risks connected to this processing challenge group category. The skillset required to design intelligent BD tools or to design effective BD processes are also needed to appropriately extract the value of these big datasets. Skilled labour plays a vital role in the operation and development of appropriate processes and new BD technology to address the concerns raised. There were recommendations that leadership need to include within their governance strategies to improve training and skills development for staff working with complex BD processing challenges. This relies on the assumption that there is finance allocated by the organisation, as there are also funding implications related to the problems identified. Not only from a technical capability perspective, but also an ongoing skills development and training perspective.

- **Big Data Mining and Machine Learning Barriers**

Issues exposed during inductive analysis were not just related to the collection or storage of BD. A lot was discussed about problems with BD algorithms and assumptions in big datasets, or just processing to make sense of it. For instance, it is simple to reach inaccurate inferences from BD. Techniques and methods to query or mine BD are significantly different from analysis of traditional datasets (pc: BD mining & ML barriers) → c = 62. BD is dynamic, heterogenous, unstructured, noisy and often unreliable, consequentially old methods of analysis and organisation of the data were agreed to be no longer sufficient by a considerable number of participants. Additionally, the collection and analysis of BD, will be of limited value if it cannot be comprehended by end users (pc: interpretation complexities) → c = 23. Various participants reported complexities with how BD is processed and interpreted. rrors

can emerge from BD feeds, but also erroneous assumptions can be built in models which compounds this issue. Additionally, when using BD to represent the world, care needs to be taken to prevent over-simplification. This means that it is crucial to ensure that the correct questions are being asked of the BD collected, and that interpretation incorporates an understanding of context. This is made even more difficult due to the ubiquity and dynamic features of BD. (BIG0009) mentioned the metaphorical term stove piping, in the context of intelligence, to describe various approaches in which raw data might be presented without appropriate context. To address some of these challenges, (BIG0076) suggests that organisations adopt process integration and standardisation. Arguing that the capability to integrate processes and standardise tasks, will not only aid enhancing the trust in BD results, but will enhance the BD chain. According to (BIG0056), both categorisation and interpretation of BD must be carefully governed and standardised to insure the maintenance of BD quality.

Another participant believes that a core issue connected to bridging the gap between deep BD comprehension to context specific action and subsequent impact, is technological barriers to BD implementation; namely curation and visual language analysis (SMD0004). Additionally, new tools were said to be required to interpret and visualise results of what is largely unstructured BD (SMD0009). A challenge we see with this is defining how BD tools and technological solutions can continually evolve to interpret and visualise this type of BD. Critically BD systems must convey result in a comprehensive way to assist end users with BD interpretation. One would question how big data governance can assist to resolve this, however one of the big data governance focal points here is the criteria that some of participants highlight. They identify a skillset gap pertaining to individuals with the needed capabilities to address these aspects. Skillsets need to be developed in this area to address the reported skills gap by a number of industry sectors. To minimise the risks posed by these gaps, organisations should prioritise strategies that underly the capability to take advantage of BD opportunities and mitigate.

The participants also identified a number of problematic factors of BD mining and ML which they related to ethical challenges for organisations. Due to the significance of this research finding, we elaborate on this area separately in section 4.3.

4.2.6 Financial Challenges

Another distinct category was discovered during detailed examination. This is classified as the financial challenges concept and covers implications related to funding the required activities, processes and technologies for an organisation to succeed in BD initiative, as well as derive the expected return on investment. Table 15 below presents the frequency count for the distinct themes we identified: -

Category	Code	Frequency (Grounded)
Financial Challenges	• FC: = FINANCIAL CHALLENGES	6
	• fc: cost / operational expenditure	26
	• fc: investment in BD technologies & insfrastructure	32
	• fc: Staffing, training & skill development	31

Table 15: Financial challenges group frequency table

Although results show this to be a lower-level criteria compared to others presented in earlier sections, the financial concepts that participants perceived as concrete were included in the criteria list. Analysis revealed that funding for certain aspects of BD is vital for most organisations. General financial challenges that were referred to, but not specifically described (fc: Financial challenges) → c = 6, BD challenges related to cost control and operational expenditure implications (fc: cost/operational expenditure) → c = 26, ensuring appropriate investment in BD technologies and Infrastructure is budgeted for (fc: investment in BD technologies & Infrastructure) → c =32, financial implications of Staffing, Training & skill development (fc: staffing, training & skill development) → c = 31

In the context of the above, a pattern emerged which points to the need for governance which guarantees Leadership teams within organisations ensure funding models are strengthened and ROI is monitored. BD analytic tools, data warehouses, repositories, infrastructure, application development, high performance computing capabilities etc all have an associated cost. Participant's evidence analysed indicate that these BD technologies currently evolve at a rapid rate or could have large upfront costs. Both (BIG0013) and (BIG0048) stated that there was limited resources and a lack of focus on funding as an organisational priority. This

is a risk for BD initiatives and its governance. Hardware does not have an indefinite lifespan, usually having an end-of-life span with 3-5 years and therefore requires continual investment and opex budgets to refresh (BIG0069). With a considerable number of participants revealing that there are financial implications brought about by BD related to the volumes of data generated, related storage costs and BD processing. Although storage costs were said to be cheap, (BIG0021) contends that costs associated with legal requirements compliance and data centre maintenance leads to BD investment cost ineffectiveness for smaller organisations. Other participants raised the impact of higher spending on security, the fines for non-compliance of regulations and damage to organisation's reputations that could also have its own financial consequences. Sharing of BD was also said to be an expensive exercise for most organisations, also requiring skilled staff that have the capability to identify, extract, clean and prepare BD in formats sufficient for usage and sharing. However, there are claims that organisations rarely provide sufficient funding for BD curation. It was observed that the healthcare industry representatives discussed challenges with BD initiatives, linked to high investment costs and intermittent cost cutting exercises within their sector (BIG0066). Whilst they cite significant benefits of BD usage to enhancing patient care, they also argued there was significant problems being experienced relating to how to govern the growing costs of curation and maintenance of these big datasets. According to what was disclosed, it is a costly and time-consuming activity to retrospectively assess quality or conduct quality assurance activity on big datasets. Adding that there were shortcuts taken in BD initiatives due to avoidance of high investment or cost cutting exercises.

Here again, there is a clear recognition by participants of the gap in BD skills and capabilities, hence many of the participants pinpointed the requirement for additional investment in skill development, training and education to satisfy current and anticipated demand at all levels. (BIG0085) contended that BD gathering could be burdensome, so must deliver benefits that can be evaluated against its tangible and intangible costs. Big data governance is a success factor for BD initiatives, and generally should have a positive impact on how BD initiatives perform, and operational expenditure is controlled. It is argued that one of its main purposes is to increase the value of BD and minimise unnecessary BD related costs. Decisions affecting both capital and revenue costs related to BD, provide a direct interdependency to calculating the ROI which in turn impact the organisational outcome of

deriving value from BD assets. A secondary risk highlighted by one participant is that BD initiatives are not adequately costed, driven, or controlled effectively. Thus, becoming a drain on organisation's finances without realising the overall set of tangible benefits. Governance must take into consideration funding for this kind of BD associated costs. BD operation expenditure needs to be scoped and planned into yearly financial budgets and subsequently controlled and monitored. Findings reveal that it is a major big data governance criterion to ensure financial implications are considered and mitigated. As with any BD project, participants views are that the risks are reduced by having realistic finance or investment plan to have appropriate BD infrastructure, technology, processes, resources and training in place. As recommended by TechUK, a long-term funding base for BD innovation should be supported with set budget cycles to ensure development of long-term capital investment plans to provide continued support of BD developments and SME skill development (BIG0039). This implies the concept of a distinct financial big data governance criteria to control these risks aspects strategically. Figure 17 below shows the factors which are induced to influence big data governance for financial related BD challenges based on the thesis analysis findings.

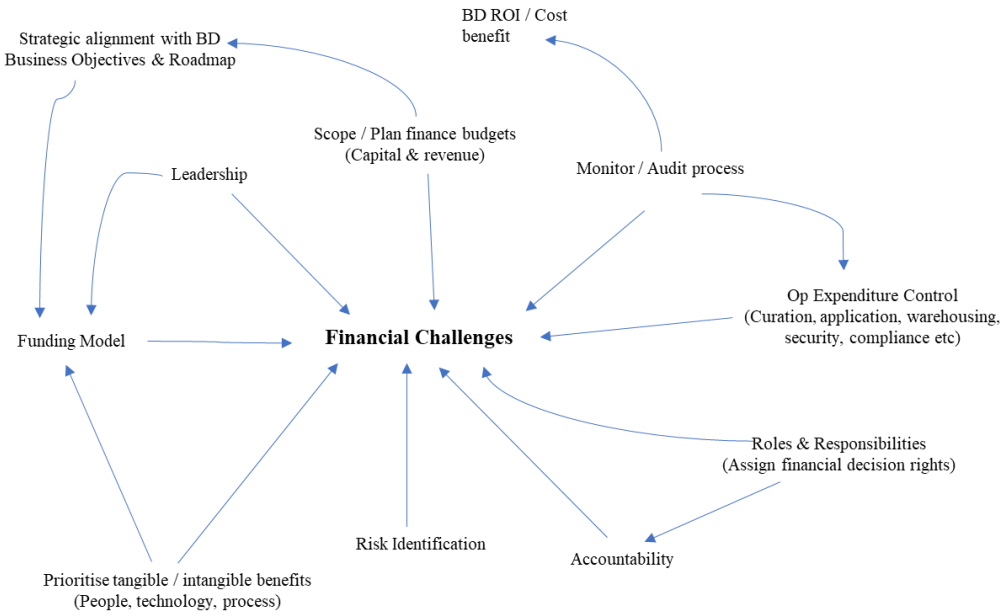

Figure 17: Factors that influence big data governance for the BD of the Financial challenge concept

4.2.7 Technical Challenges

Increasing reliance on BD driven systems is also unveiled to be accompanied with a high of technical risks. This set of technical challenges are attributed to the technological problems that effect big data governance decisions taken regarding BD implementation in organisations, as well as BD technology tools adoption and standards.

Category		Code	Frequency (Grounded)
Technical Challenges	•	TC: = TECHNICAL CHALLENGES	24
	•	tc: BD architectures	13
	•	tc: data protection	43
	•	tc: high perf computing/network capacity / int connectivity	26
	•	tc: Inadequate BD datawarehousing / infrastructure	44
	•	tc: increasingly complex processing & manipulation	41
	•	tc: insufficient models, techniques and tools for BD analysis	38
	•	tc: integration & interopability	54
	•	tc: legacy data insfrastructure & systems	11

Table 16: Technical challenges group frequency table

Based on the research findings, this theme was categorised this theme into eight sub-groups as presented in table 16 above. As shown in the table, where technical problems were highlighted by participants but did not fit into a distinct theme, these were left aggregated in the generic technical challenges theme (TC: = Technical Challenges) → c = 24.

The other sub-groups are outlined below:

- Inadequate BD architecture (tc: BD architectures) → c = 13
- Complexities linked to BD protection (tc: data protection) → c = 43
- Increased requirement for high performance computing, network capacity and internet connectivity (tc: high perf computing/network capacity /int connectivity) → c = 26

Inadequate data warehousing and infrastructure for BD (tc: inadequate BD warehousing / infrastructure) → c = 44

- Burden of legacy data infrastructure & systems, requiring considerable upgrades to meet the needs of BD features (tc: legacy data infrastructure & systems) → c = 11

- Increasingly complex computational methods required for complex BD processing and manipulation (tc: increasingly complex processing & manipulation) → c = 41

- Inadequate models, techniques and tools for BD processing and analysis (tc: insufficient models, techniques and tools for BD analysis) → c = 38

- Complexities linked to integration and interoperability both internal and external to the organisation (tc: integration & interoperability) → c = 54

In seeking an explanation for big data governance success, it is induced that all the technical challenges that participants identified as factors that impact BD services include performance, quality, BD protection, privacy, availability, integrity, processing, availability, reliability, security, standardisation, compatibility and ownership. This is significant, as these factors impact the decisions related to big data governance implementation. Figure x below presents the influences of big data governance in relations to the technical challenges as analysed and called out by the participants.

The increasing reliance on BD analytics and implementation of BD driven systems comes with major technical risks within the context called out above. A number of participants stressed that it is critical that issues that impact these aspects are addressed to prevent the creation of vulnerable BD systems that fail in cascade, leading to widespread practical and economic damage. Or fail to meet the requirements of the use case meant to be solved. We discuss some of the challenges raised by participants below.

- **Scalable and Flexible Big Data Technology & Infrastructure**

A number of participants highlighted that the proportion of effort essential to handle and process BD, is determined by having in place scalable and flexible BD technology and infrastructure. Explaining their practical experience, some participants reported having particular challenges in terms of legacy systems and infrastructures historically used for traditional data types (BIG0002; BIG0071), arguing that these systems need to be upgraded to deal with the inherent characteristics of BD on a regular cycle of several years (BIG0040). Not only does this link with a technology problem, but several participants also reveal this is closely linked with problems related to the additional investment that is required when most organisations are working with tight budgets and ongoing cost cutting exercises. It was further revealed that the inability to upgrade legacy infrastructure could lead to high risk factors. Put simply, access to disparate BD sources and the ingestion of huge volumes of data could lead to increased amounts of "noise" in data and reduce the levels of BD quality. Proprietary software also poses issues for support and backward compatibility for legacy systems, that notably increases the risks of BD loss or loss of access to the data (BIG0066). Several participants emphasised they do not have mature organisation ready capabilities around implementation, competencies, support, training, etc.

One participant provided an example related to the company Talk Talk, which had taken a well-publicised hit due to the lack of investment to secure BD legacy systems (BIG0088). (BIG002) spoke about the requisite to future-proof IT infrastructure to be prepared for integration with existing and future BD technologies. As discussed in earlier sections, this implies that organisations need a strategy to factor in funding required to scale and upgrade BD systems. Additionally, where BD is used to drive physical systems, specific technical challenges need to be considered around the complexity, uncertainty and resilience of the systems that are created, by integrating previously separate networks of sensors and devices.

- **High performance computing, network capacity and internet connectivity**

Another challenge raised, is one that has historically existed in computer architectures for decades i.e. CPU-heavy but i/o poor. Various participants argued that this system imbalance is amplified with BD, hampering the progress of discovery from BD. One of the participants

went further to state that a potential technical barrier to realisation of BD ambitions is the lack of access to adequate computing capacity (BIG0027). Claiming that organisations that need to transfer BD between organisations in large scale are being hampered by legacy hardware and software as highlighted above. BD features also means that adequate network capacity and internet connectivity is challenging. Bottlenecks were reported with network bandwidth capacity, particularly where the magnitude of communication is considerable (BIG0075; BIG0027). Although there was no evidence found whether this is related to cloud, non-cloud or distributed systems.

- **Technical Data Protection & Security Controls**

It is observed that BD transmissions also demand for increased integrity checking due to the increased security vulnerability. Oral evidence provided by one participant recommends deep technical work to connect and bring about the best privacy settings & safeguards (HC468b), suggesting that there be stronger penalties for organisations that do not put in place appropriate safeguards. There are technical physical and procedural requirements to ensure BD security & privacy. However, several participants said establishing user identities and permission management within siloed structures are challenges which need enhanced governance to monitor and control. Claiming there are a number of ways in which previous safeguards are ineffective (HC468a; BIG0006; BIG0049). For example, whether BD is stored on secure servers or within cloud facilities internal or external to the country the organisation physically operates. Capabilities to diagnose or meaningfully respond to an infiltration in real-time is lacking, and most say they rely on technologies that cannot be used for this purpose at all (BIG0049). Additionally, there are limited technical solutions for BD products that require to proactively get informed consent (BIG0027). Broadly, an organisation that lacks the ability to tackle these BD challenges through efficient big data governance practices put themselves at huge risk in gaining a competitive advantage, risk reputational damage or at risk from regulators. The perspective most commonly mentioned by participants was that every BD project requires deep technical understanding in the best privacy settings to utilise, as well as need to investigate what the BD purpose for collection is and what safeguards are core requirements. Analysis findings also uncover those organisations need to have as part of big data governance the ambition to be able to attain the policies, capabilities, knowledge and skillset to resolve given these broad problems.

- **Integration and Interoperability**

Problems with integration and interoperability were found to be the highest reported technical challenge. Crucially, our inductive analysis shows that new BD methods, architecture, volume, veracity and variety impose additional risks of lack of control and governance over BD, and this necessitates extra organisational focus in this area.

Architecture to support integration with different BD sources needs to be interoperable (SMD0009; SMD0007). The overwhelming volume and velocity of incoming BD, plus the multiple channels accessible for decision making purposes produce consequential issues for technical integration (BIG0022; BIG0065). It was discussed earlier in the literature review that the multichannel environments, creates the need to connect via a single consumer view to drive BD analytics of consumer profiles, glean business intelligence and support decision making from BD. BD analytical packages which conduct various kinds of processing such as mining and statistical analysis from diverse sources, means that integration and interoperability are major criteria for efficiency. These additional complexities are recognised by some participants as an ongoing problem for some businesses to navigate. (BIG0046) states that few companies have strong capabilities when it comes to utilising integrated multichannel or multidevice BD for real-time decisioning and personalisation. Significantly, public sector participants in particular highlight a huge problem with this, making recommendations for government to invest in BD infrastructure in a way that ensures it is interoperable and flexible (HC468b). Moreover, participants suggest not just to address this issue physically, but also to introduce processes that support these aspects as a matter of course. For instance, (SMD0023) says procurement processes used by organisations could result in BD systems that are incompatible and hard to integrate. This results in an interconnected problem, where there disparate BD collection is built up, subsequently leading to difficulty in users obtaining accurate analysis and insights (SMD0021). For BD analytical packages which conduct various kinds of processing such as mining and statistical analysis, integration and interoperability are major criteria. Additionally, BD ecosystem has multiple data warehouses and systems across a wide array of suppliers, making it complex for some organisations to comply with legislation due to the lack of integration e.g., when receiving a request to erase customer records. Another salient issue raised by participants is the lack of coordination between database suppliers, hosting providers and other stakeholders in the BD

landscape. According to (SMD0023) procurement processes used by the UK government may lead to incompatible systems that make it difficult to integrate.

We discuss the problems identified regarding inadequate BD standards and the overarching BD issues related to this in chapter 5. However, we note here that the lack of technical standardisation is also interconnected to what will be discuss later. According to (BIG0076), technical standardisations is required in the following areas to ensure BD integration and interoperability:

- Query languages including non-relational queries to support diverse data types (XML, PDF, JSON, multimedia etc) and BD operations e.g., matrix operations.
- Domain-specific languages
- Semantics of eventual consistency
- Advanced network protocols
- For efficient data transfer: General and domain specific ontologies and taxonomies for describing BD semantics including interoperation between ontologies.
- BD security and privacy access controls
- Remote distributed and federated analytics (taking the analytics to BD) including data and processing resource discovery and data mining.
- BD sharing and exchange.
- BD quality and veracity description and management.

- **BD Architecture**

Closely interconnected to the problems participants identified above is the complexities establishing robust BD architecture to support integration and interoperability. Participants raised that 1) design principles that can be used to make BD capture and reuse clearer, 2) protocols that allow BD to link across repositories and 3) architecture designs capable of scalability to record BD from globally distributed disparate BD sources will enhance BD integration and interoperability (BIG0012; HC245; SM0007) even though they contend this is complex. (SMD0009) argues that whichever standards or protocols are introduced in this regard would need to allow maximum flexibility.

- **Insufficient tools and techniques for Big Data Analysis**

According to (BIG0048) the technical barriers to analysing BD sets are significant, particularly where expertise and access to high performance computing facilities are concerned. Additionally, capabilities to handle the analysis of BD volumes and intricacies of this type of information in a timely manner was said to be complex. It is worth noting that timeliness is of utmost technical priority for real-time BD applications for most sectors represented in the inquiry e.g., finance, biomedicine, transport, social media companies etc. However, there is still technical challenges reported for stream processing involved in BD. BD was not only reported to be producing changed directions in the development of hardware, but also software architectures and the synchronisation of these large datasets.

Big data governance is not just a technology problem, it is also a business problem. However, getting the right mix of BD skills and technical capabilities to resolve some of these challenges are vital. There is an important role for IT as some participants rightly pinpointed, as a number of these challenges require the technical competence to address knock-on impacts to BD security protection, personal privacy protection, performance, integration etc. Big data governance would also support the need to build partnerships with other areas of the business to effectively govern and control risks - whether that be investment, procurement, training or other related dimensions. The idea that organisations will get to a steady technical state relatively quickly does not feel likely based on the research analysis. However, participants discourse show evidence of a consensus that these technical barriers must be effectively governed over time, with clear guidelines, policies, processes and procedures, as well the right roles and responsibilities identified. One participant goes as far to say that the ambition should not be to try to address all these areas from the centre. The ambition should be to spread the capability, knowledge and skillset very broadly across the organisation, with big data governance capabilities to address these integrated within the decision- making processes.

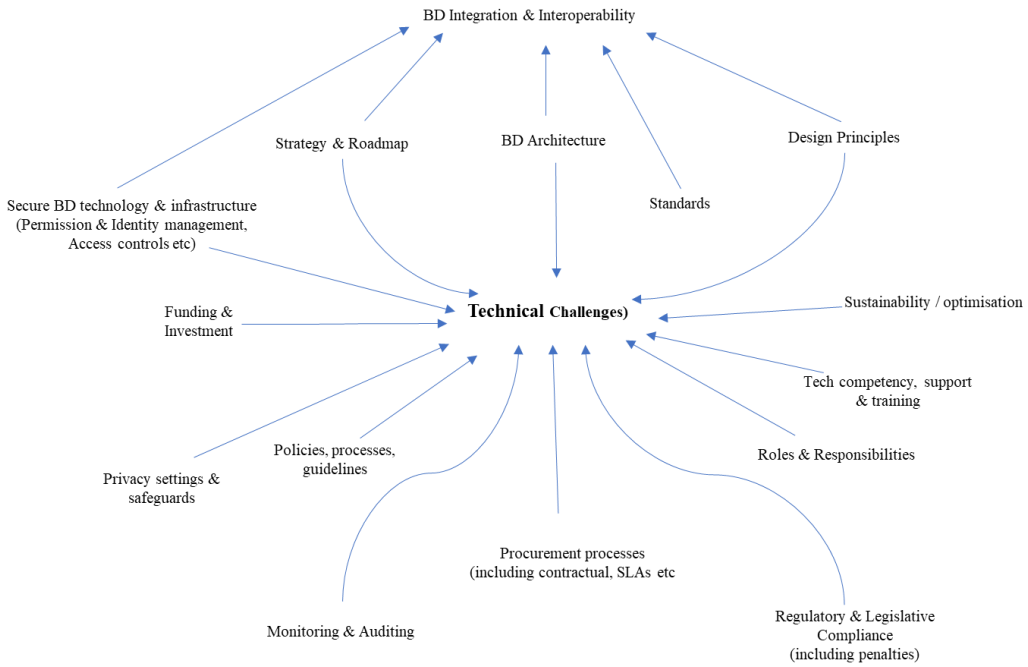

Figure 18: Factors that influence big data governance of the Technical challenge concept

4.3 Ethical Concerns Associated with Big Data.

This section presents further core findings from the data analysis conducted. Critical to this thesis research, in-depth investigation identifies another unique category of BD challenges which have novel ethical implications for big data governance. The participants from both the private and public sectors disclose a number of distinct ethical concerns regarding the collection and usage of BD. This was a continued theme in a large number of inquiry submissions and high on the hierarchy of requirements for our proposed big data governance framework. Due to the major significance of this finding to the design of our novel big data governance framework, it is expanded upon within a separate section from the other BD challenges and risk factors exposed earlier in chapter 4. This does not reduce the significance of other findings discussed earlier; nonetheless, to achieve some of the research outcomes this concept requires more expressiveness.

Throughout the process of coding, grouping and examining the dataset, significant insights were derived from how the participants described various aspects of distinct ethical concerns. The grounding of the findings shows that they are part of the top 5 BD challenges discussed

by participants. Table 17 outlines the ethical codes and how grounded they were in the datasets analysed for this thesis study.

Category		Code	Frequency (Grounded)
Ethical Challenges	●	EC: = ETHICAL CHALLENGES	122
	●	ec: BD ownership/IP obscure	15
	●	ec: civil liberties / surveilance	59
	●	ec: BD capture for economic gain /exploitation	37
	●	ec: discrimination/bias/profiling	64
	●	ec: Functional creep /concerns with secondary use	29
	●	ec: imbalance of power/control	15
	●	ec: Informed consent lacking/incomplete	42
	●	ec: privacy infringements	34
	●	ec: questionable use /misuse of BD/BD analytical methods	21
	●	ec: replacing/supporting human decision making	12
	●	ec: safeguards prevent BD sharing	7
	●	ec: security of BD	31

Table 17: Ethical challenges code frequency count

To avoid confusion and provide clarity, the term ethics in the dictionary pertains to dealing with the principles of morality or being in accordance with the rules and standards pertaining to right or wrong practices or conduct. These moral principles of conduct govern an individual's or organisation's behaviour or the handling of an activity. More importantly ethics influence people or organisation's decisions-making, their rights and responsibilities or how they lead their lives. It is worth noting here that ethical is not the same as being lawful (BIG0015). This was buttressed by a participant of the inquiry, highlighting organisations should avoid relying on compliance with the law to guarantee that BD usage is morally appropriate. This research study's stance is that legal compliance by organisations, should never be deemed sufficient to safeguard morally reasonable utilisation of BD. This viewpoint is also supported by several of the participants.

In terms of scope, observations made during analysis found that participants appeared to collectively utilise the terminology 'dig Data' to refer to the data itself, as well as the analytical methods used to process it. This aligns with the viewpoint of this thesis study; as from the starting point, it is contended that the entire BD ecosystem connected to people,

processes and systems is of concern when considering BD challenges to be addressed in the proposed big data governance framework.

The emphasis on communicating wider ethical implications were expressed by a number of participants in various context. According to (BIG0008), Ethicists are only just becoming aware of the complex issues within the existing privacy paradigms to do with BD issues. Participants articulated there is a growing call for ethical behaviour which is not limited to the users of BD services. Professor McAuley of the Horizon Digital Economy Research Institute pointed out that various small businesses were concerned regarding how to work with the ethical issues surrounding BD e.g., social media data, whilst still utilising the BD to inform their business requirements. Claiming this was leading to "tension between innovation and regulation" (HC 245). Sureyya Cansoy of TechUK, also reported that an "increasing number of companies were concerned about ethics questions", with Professor McAuley further arguing that "many small companies, even large ones, want to be seen to be behaving ethically and are getting somewhat annoyed at some of the unethical behaviours of others" (HC 245). These ethical challenges some of which reflect conflicting perspectives between BD innovation and the call to have this regulated, should be addressed with a practical, equitable and collaborative approach when designing big data governance given the conflicting angles.

As reflected in table 17, the ethical related themes and patterns grounded in the dataset ranged from the following summarised below:

- Infringement of people's privacy due to integration of BD technologies and systems.
- Whether consent methodologies some organisations utilise are appropriate.
- Concerns related to functional creep of BD, as well its secondary use by organisations
- Problem connected to the purpose of BD acquisition
- The potential for discrimination and bias being perpetuated by profiling of individuals irrespective of BD anonymisation.
- Increased surveillance and intrusion of civil liberties

- Abuse and misuse of the BD and its related BD analytical practices
- Security of Big Data - connected to unauthorised access, unauthorised use etc
- The replacement of human resources to support decision making.
- Unscrupulous BD capture for economic gain and the impact of BD monopoly by a group of organisations
- Limitless life span retaining BD.

As with several other BD challenges already discussed in this chapter, it was observed a number of these ethical concerns were interlinked, hence this is reflected in how codes were merged during the coding of this category. For example, due to the myriad of interrelated patterns a number of these ethical challenges were merged into the generic code EC: = Ethical Challenges, with a frequency count of 122. Where participants discuss in-depth about a particular ethical challenge, but the frequency count is not high these unique problems were combined into this overall generic code. However, there were some instances that although the frequency of some of the coded concepts was not high, the ethical challenges they posed were significant enough to be separated into distinct codes in this thesis study.

In a view similar to ours, (BIG0057) and some other participants highlight the importance of understanding the level of BD ethical risk, particularly concerning privacy, security, trust, and develop the appropriate responses such as proportionate ethical approaches that support big data governance. Hence as a step towards a more extensive appreciation of the evolving ethical concerns relevant to BD, the research study delved deeper into this particular area and present the perspectives reported by the participants in section 4.3.1 below. This further contributes to the novelty of this research, as to our knowledge previous big data governance frameworks have not addressed this aspect in-depth.

4.3.1 Data Analysis: Participants Ethical Concerns Associated with Big Data.

The comprehension of ethical concerns discussed by participants can be improved by examining them within the BD lifecycle. Factors which lead to the myriad of interconnected ethical challenges, were observed to occur at disparate stages of the life cycle, which implies the requirement for distinct big data governance procedures. Khatri (2010) identifies the data

life cycle as a key decision domain that shapes how data should be handled. Aligned with the viewpoint of this author, this research study considers the end-to-end big data life cycle (BDLC) of interest when analysing the points at which ethical challenges occur in the BD life cycle. Crucially, Khatri (2010) argues that realising all data moves through life cycle stages is central to designing data governance, contending that the DLC decisions include how the data is defined, produced, retained and retired etc. Consequentially, the ethical concerns revealed during interrogation of the dataset were mapped to these BDLC stages to assist in the design of the ethical component of our proposed big data governance framework.

To provide some clarity, we define the BDLC as the order of stages that BD passes through from its initial capture to its distribution and reuse, to the archiving or deletion of BD at its end-of-life stage. The stages are outlined below:

- BD Capture (collection)
- BD processing
- BD synthesising
- BD usage
- BD publication
- BD discovery
- BD analysis
- BD archiving
- BD deletion

The table in Appendix C shows some examples of the ethical implications of BD we uncovered during our analysis, which we mapped to the BDLC stage they might occur. Hence, deemed to be a further valuable contribution of this thesis research. Although it has been attempted here to supply a comprehensive list of all ethical challenges which emerged from this research analysis, this study could not discuss in detail the context of every conceivable ethical concern due to the word count limit. Therefore, we select a few of the issues for expanded discussion in the following subsections. The selection criteria for those discussed is based on the higher frequency count, or their close interrelationships to some categories discussed in earlier sections of chapter 4.

- **Trust Deficit**

The initial strand to be explored seeks to establish the importance of the link between ethical standards and public trust to the decision-making process for BD initiatives at both the strategic and tactical level. Results revealed close links in the dataset that infer that the unethical use of BD inevitably undermines public trust in both private and public sector organisations. There is also a clear link between privacy infringement and a consequential consumer trust deficit. The data analysis suggests that where organisations acquire BD to innovate, maintain competitive advantage or improve operational efficiency, any BD misuse could lead to a lack of public trust or pose a huge risk to an organisation's reputations. Analysis of the dataset unveiled that breaching that trust leads to considerable backlash and undermines an organisation's BD efforts. Honesty and trust are currency for organisations to do business with their customers.

According to one participant:

> *"An essential precursor to realising the massive opportunities of big health data to save and improve lives (and benefit the economy) will be ensuring the public is well-informed about how their data are collected, stored, linked and shared and ensuring their legitimate concerns about privacy are taken seriously and addressed effectively. Patients, the public and healthcare professionals must understand and trust the system. Building that trust is fundamental"* (BIG0030).

(BIG0038) argued that misuse of BD has impaired confidence that some citizens have on digital interactions with organisations, to the extent 30% of the population can be classified as non-sharers. Similarly, attempts by organisations to shift norms or impose new norms without engagement with the wider public, also risks undermining trust and therefore the objectives of the BD initiative. These views expressed have severe implications for an organisation, given underlying requirements to extract value from BD initiatives.

The thesis study identified that if organisations from the different sectors are to play a role in the delivery of public BD services or products, this must be engaged within reasonable ethical expectations of the public and devising ethical behaviours. As a step towards this, we argue that organisations need to take additional care to determine situational ethical risks of BD initiatives. Making sure to examine upfront how BD initiatives could disproportionately impact different people or groups and explore how individuals might value the possible BD benefits and BD hazards. Taking this upfront step, enables organisations to explore what forms of ethical BD governance is needed within their enterprise.

- **Privacy Infringements**

One of the principal risks related to BD and its associated practices uncovered, is the difficulties for those organisations using it to meet ethical and privacy standards. Broadly speaking, the ethical challenges subgroup called privacy infringement (ec: privacy infringements) → c = 34 refers to the unwelcome exposure, inference, tracking or leakage of confidential or private information. It is determined inductively that there is a high risk that BD could be unethically exploited by individuals, private and public actors, resulting in privacy infringements in the wider context of living in the digital era in which BD and personal data is routinely captured by organisations. From social networking platforms to financial transactions to shopping habits etc participants highlighted BD is being routinely captured, analysed and integrated, increasing susceptibility to privacy infringements. The frequency count of this specific issue was moderately high in the hierarchy of ethical issues uncovered. This mirrored the high level of privacy related concerns indicated within the management challenges group discussed earlier in chapter 4.

For a considerable number of participants there appeared to be acceptance that BD analytics, including the use of social media BD types has an ethically challenging privacy dimension. To illustrate, (SMD0004) argued that BD has the capability to power analysis from the macro level of international opinions and trends to the individual psychology at a private level. Stating:

"The BD collected go beyond demographics to include people's private life via social media. Combining this information with other sources such as credit card data, supermarket loyalty card data and various others, allows the linkage of private life with consumer behaviour at a deep level" (SMD0004).

Likewise, (BIG0006) elaborated on a further example of privacy infringement through the narrative of another popular BD analytics example connected to the US retail chain, Target. The participant recounted:

> "*Famously they sent a teenage girl marketing material related to pregnancy which were based on seemingly disconnected purchases she has made, including vitamins, soap and cotton wool. By correlating the data, Target predicted the girl was pregnant. They knew of her pregnancy before her father"* (BIG0006)

In a thought-provoking contribution, (BIG0019) offered insights into a practical way of looking at this BD related ethical problem in terms of paradoxes – the transparency paradox, identity paradox, and power paradox. Referring to the identity paradox which speaks to challenges such as privacy infringement, the participant contends:

> "*Though BD evangelists talk in terms of miraculous outcomes, this rhetoric ignores the fact that BD seeks to identify at the expense of individual and collective identity"* (BIG0019).

This contribution also implies that BD usage should not be aimed at causing resulting harm. The participant contends that recognising the paradoxes of BD, which convey its risks along with its potential, will help us to better understand its evolution

Despite the acknowledgement by a high number of participants regarding various ethical concerns connected to privacy violations, Baroness Shields the Minister for Internet and Safety, tempered this point by pointing out conflicts in individual's attitudes towards BD. This participant's viewpoint is that there is a discrepancy in terms of what individuals think concerning trusting BD. Commenting:

> *"If you talk to teenagers, they do not care, they have given up privacy and decided that they are happy to share absolutely everything in their lives and have it catalogued"* (HC 468).

Notwithstanding this contribution, (BIG0085) expressed that BD containing the personal data of young people should not be seen as a commodity for unethical exploitation unless people are willing to consider their exploitation in the digital world as more acceptable than in the real world. Arguing that organisation cannot continue to separate the exploitation of BD, from the exploitation of the person, particularly when it relates to young people's personal data (BIG0085). Moreover, the earlier rhetoric ignores the fact that analysis shows evidence that BD seeks to identify at the expense of the individual and collective identity. Based on the evidence, we induce that BD usage should avoid causing resulting harm, and organisations have an ethical responsibility to look out for individuals interests through appropriate governance of BD initiatives.

During the course of research analysis, it was found the reported ethical implications associated to privacy infringement possessed complex and intertwined relationships. There are two key interconnected features of ethical practice reported that are examined here: 1)

Privacy and anonymisation and 2) Ownership and consent. Due to the complex interrelationships of these ethical problems with other ethical themes uncovered, they are examined further with other interrelated issues where applicable.

- **Lack of Informed Consent**

The ubiquitous nature of BD usually results in the consent of BD being sort at the time of BD acquisition. Some participants alluded to two obvious ethical issues that arise when organisations make further use of the BD. Questions such as 1) Would the original consent obtained still reflect the preferences of the person who provided it? 2) Does the new proposed usage fall within the uses that the person who provided it originally intended? (BIG0015) emphasised that an individual providing data about themselves cannot fully comprehend the potential consequences of how their BD will be available for reuse or linking, arguing that BD protection methods such as anonymisation cannot solely protect all their interests. Whilst some responses analysed reveal general agreement that consent acknowledges a person's right to decide against certain uses of BD. However, as both participants (SMD0021) and (SMD0004) suggest, it does not prevent damage occurring to the individuals when the consequences of BD use maybe misunderstood or unforeseen. It therefore needs to be supported with big data governance procedures that ethically obtain informed consent.

As discussed in one of the submissions, informed consent has been considered a cornerstone of ethical research which involves individuals (BIG0057). Yet 21 participants admitted that BD has brought about increased complexities to obtaining informed consent relating to each person at every specific BD life stage. They also admitted that traditional models of consent are not fit for purpose in an age where BD analysis etc are expanding the risks for potential privacy invasion, misuse of BD and revealing the limitations of consent to use and share BD. One participant commented:

> *"There is however a rising tide of concern about BD analytics and the use of personal information and the unintended consequences of its use and misuse. Key to this is the issue of consent and permission management, the act of making an informed choice about how data is used"* (BIG0032)

(BIG0070) also placed further emphasis on this ethical problem, stating:

> "*BD often rely on the collection of personal data and, to varying extents, private data. The biggest challenge lies with the notion of informed consent, that the person has sufficient information and understanding to make a rational decision" (BIG0070).*

Given the relationships and links to other BD challenges, some of the big data governance criteria to address would include strategy, defined roles and responsibilities, as well as policies determining the organisations ethical and permissible use of BD without violating a clients or consumers privacy rights.

Earlier in section 4.2.3, 56 participants highlighted interrelated issues reflected in the management challenges code group, also pinpointing management issues related to the limitations of some current consent methodologies and terms & conditions. Findings also show that there are both ethical and management problems noted with company's terms and conditions e.g. in some cases they change without their customers knowledge etc. Additionally, from an ethical perspective, there is also purported to be evidence of impractical consent models which have been adopted by companies for convenience, which fail to bring any tangible benefits to the individual and are not transparent. We argue that organisations are required to provide clear and understandable consent statements and models that customers or the public opt-in to, based on the data analysis results.

- **Ownership and Intellectual Property**

A subgroup of participants delved into the ethical problem surrounding ownership of BD. It is observed from the inductive analysis that this is a complicated area. Organisations carry out BD collection for innovation or for the intention of doing business, however BD does not emerge from nowhere. Several participants made the key point that it is disputable that companies own the BD they receive from individuals.

This theme was expanded by participant (SMD0015), contending that commercial companies such as Facebook and Twitter profess legal ownership of BD published on their platforms. However, a company like Twitter claims tweets published are owned by the people who write them, the tweets are posted in public, but collectively packages them as a sellable product (SMD0015). Delving into aspects of this topic; there is a valid argument that owning something does not give the owner the go ahead to use it as they choose. An example this which one participant supplies is the ownership of medical records. Commenting:

> "*Owning something does not give the owner carte blanche to use it as they choose. Ownership of medical records does not permit a GP to sell those records to insurance companies"* (SMD0004).

Furthermore, this participant disputed if companies own the BD collected from individuals:

> *"If I share an original idea with you, it is still my idea"* (SMD0004).

The participant contends that this stance is widely recognised in professional ethics, academic practice, and intellectual property and copyright law, hence the strong perspective that this should also apply cross industry for BD applications or BD services. (BIG0018) commented that clear ethical frameworks relating to intellectual property rights of BD were required. Other participants also elaborated on the ethical problems to do with ownership and consent. It is interpreted from the analysis of the dataset that it is paramount that organisations fully appraise and govern the ownership context in which BD initiatives will be executed. It is also noted that potential harm due to BD misuse must be taken into consideration, both on the part of those to whom the BD pertains and of those who may be affected by the data. BD collected and stored by organisations is normally about data subjects. The analysis findings show that Ethical problems emerge where BD involves dealing with personal data.

An ethical issue concerning ownership of data after a user's death was also cited as a grey area, raising ethical concerns about the use of personal data in BD initiatives. (SMD0019) claimed that the data capability strategy of the UK government at the time, did not really address this query, claiming that there was a lack of clarity on how the deceased data can be suitably protected or transparently or accessibly maintained. This participant goes further to argue that this lack of clarity exists both in the UK and globally. This ownership scenario also drew attention to further questions on the timespan limits for which BD can be ethically held.

Even if existing legislation was straightforward and coherent internationally, the thesis study's viewpoint is that ethical concerns affecting ownership remains. For example, participants raised that creation of content on social media platforms causes a deluge of BD e.g. comments, tags, metadata, photographs and videos. This thesis research supports the notion that organisations must give extensive thought relating to the rights they provide data subjects, affording them some sense of ownership and control over their data, which forms part of their identity. Where there are gaps in legislation, analysis results suggest that organisations should safeguard that BD handling is morally or ethically acceptable. Hence,

the requirement to provide some big data governance guidelines to address this among other areas reported in the research findings.

- **Functional Creep / Secondary Big Data Use**

Another key finding is the ethical concerns of several participants regarding the collection and use of BD, as it relates to the eventual secondary use of the BD, or the functional creep of the BD to ends other than those originally envisaged. Although the intention behind the collection of BD by specific organisations might be authentic. Due diligence should be taken to safeguard that the BD is not subject to function creep if the collected BD appears to have advantages for other entities or other purposes than those initially envisaged.

Several of the cases utilised by participants to illustrate this position stemmed from types of BD such as social media. For example, (SMD0015) considered the ethical dimensions of using personal social media data. describing the case of individuals using the social media platform, Twitter. This participant highlighted that, even though "*there is a process of consent as part of creating a Twitter account*" notwithstanding, "*it may not be entirely clear to the account holder how their Tweets might be used for secondary purposes, including research*" (SMD0015).

Some stakeholders argued that as BD moves across boundaries the risks of function creep occurs and data can be abused by those it is entrusted with (SMD0004, SMD0001). Similarly, whilst BD collected by organisations may be utilised for the aims for which it was initially collected. Research findings highlighted that it could be utilised for other purposes. For example, medical BD, may be stored for the advantage of serving the individual and alternately for the benefit of public health. Additionally, a further use of BD could be to aid insurance companies in determining premiums on an individual or group membership basis. This is further usage of the BD collected by organisations is referred to as function creep. (SMD0001) thus argues that:

> *"Further uses of data are not necessarily wrong. However, if informed consent has not been given to these uses and the subjects are not liable to be subjected to surveillance, then these further uses are not ethically legitimate. New consent must be sought, or grounds for liability established"* (SMD0001).

Additionally, whilst the insights and the competitive advantage derived from BD was never denied by participants, the practices of sharing BD with other parties for secondary use other than which it was originally intended raised calls for this to be controlled and governed. A participant claimed:

> *"There needs to be better clarity on which types of organisation will be able to access the data and for what purposes. 'Red lines' need to be clear and there should be no surprises emerging in future years"* (BIG0030).

Further investigation uncovered that ethical good practice guidelines for the secondary use of BD remains underdeveloped. As part of their written submissions one participant contends that some resources have been produced, however good practice lags behind considerably compared to that in place for other types of data (SMD0015). Critically, it is essential that the reality and likelihood of these distinct ethical risks are recognised within the proposed big data governance Framework.

- **Ethical Security Barriers**

The security barriers highlighted by some participants refer to ethical dimensions of an organisation's self-interest, that have implications for corporate social responsibility and protection against reputational harm. Ethical and responsible use of BD by private and public sector organisations has the potential to provide a range of benefits to the public such as more personalised and better targeted services, and simplified transactions. However, observations show that issues around security and trust can limit their uptake; for example, where BD analytics link and combine multiple anonymised datasets which can lead to an individual being identified. The fourth report from the Inquiry committee noted:

> *"The bringing together of different datasets which have been de-anonymised can also make it possible to re-identify individuals. We need to establish a balance between security and privacy, and the trade-offs made between the two should be made much more explicit"* (HC 468).

Furthermore, because verification of some BD types potentially involves identifying its source, this approach increases ethical concerns associated not only to informed consent and anonymisation, but also concerns related to accessing this type of BD.

> *"Overriding or ignoring the requirement for informed consent involves harms amounting to stealing a person's data, demonstrates a lack of respect, and requires justification"* (SMD0004)

Moreover, the damage to an organisation's reputation caused by data breaches or violations, unauthorised disclosure or access of BD, unauthorised retention of BD, or perceived BD misuse could have a knock-on impact (BIG0015; BIG0028; SMD0005). Subsequently, we induced the need to proactively take ethical governance steps to minimise these interconnected risks within the proposed governance framework.

- **Unfairness**

Some participant's contributions provoked this research to think about ethical implications related to potential BD misuse in ways that increase unfairness. Various inquiry participants present details around the controversies surrounding BD analytics usage. The integrated BD landscape is reported to enable BD analytics and algorithm-based systems to be deployed using machine learning or other forms of artificial intelligence. These increasing number of algorithm-based BD systems were reported by these participants to have given rise to the capabilities of organisations to conduct datamining, sentiment analysis and profiling of people's data to draw inferences from them. Some raised concerns about BD profiling and analytics practices increasing the risk of surveillance or prosecution (BIG0006, SMD0003; BIG0015), leading to the impact or denial of insurance claims (BIG0052; BIG0053), compromising creditworthiness (BIG0078), social sorting (SMD0001) etc. These participants make the key point that these types of examples could have far reaching negative impacts, highlighting the opportunities for organisations to discriminate against clients, customers or citizens need to be controlled.

The analysis of BD and subsequent usage of the types of inferences discussed above, raised common questions in the inquiry submissions about the fairness of BD processing. There is proof from the analysis conducted that some BD practices pose threats to how organisations treat individuals or groups fairly. Whilst providing oral evidence Christopher Graham elaborates:

> *"The more problematic it can be, particularly where inferences are being drawn about individual's behaviours to their disadvantage: for example, differential pricing.*

If I get a worse deal than you online because of all sorts of inferences drawn about my purchasing behaviour, that is not fair" (HC 468: Oral Evidence)

Participants pinpointed the need for safeguards which prohibits profiling of individuals according to characteristics which would normally be discriminatory of their personal information within BD initiatives. This is reflected in the relatively high frequency count for this subgroup of ethical challenges (*ec: discrimination/bias/profiling)* → c = *64.*

The findings drew attention to another interrelated problem, where BD analytic programmes or stored big datasets could come with substantial shortcomings and bias. These types of instances were revealed to include scenarios where inherent bias is programmed into BD software programmes e.g. poorly defined variables, excluded phenomena, missing data etc. It was claimed these scenarios could lead to incorrect or misinterpreted BD contributing to disparities (BIG0011). (BIG0053) argues that BD collection and analysis is not necessarily neutral. The assumptions built into the meanings derived from BD and the analytic processes for drawing conclusions could affect the person in many ways or reinforce bias or discrimination. Furthermore, investigations show that considerations relating to the fairness of monitoring practices occurred frequently, with participants having different opinions about this. (SMD0003) also noted that people in the UK are increasingly uncomfortable about the way BD is mined or tracked. We induced that these different perspectives appear to result from differing opinions of whether a monitoring practice is fair or not. A fair practice is one that is consistent with a user's reasonable expectations concerning the collection, use and dissemination of BD. For example, different attitudes to BD mining result from different assessments of practices and not from different conceptions of fairness. Due to these types of reasons, various participants assert that strict ethical safeguards should be introduced and observed by organisations. (SMD00011) reported on a study which found that a concern for fairness is common amongst users, which adds weight to participants calls for ethical safeguards.

- **Power Control Imbalance**

Questions regarding the collection and expanding use of BD capabilities to learn about wider patterns and trends have been discussed briefly in earlier sections. We exposed during our data analysis that this trend also has ethical implications concerning power control imbalance

and monopoly (ec: imbalance of power / control) → c = 15. This is also closely interrelated with the code group (ec: BD capture for economic gain / exploitation) → c = 37. (BIG0006) revealed 46% of individuals shared concerns that they are being harmed by larger firms collecting their BD. The utilisation of BD outweighing the privacy and interests of the individual was reported as a frequently re-occurring trade-off. (SMD0003) argued that increasingly people are no longer just worried about government breaches of their privacy through the utilisation of BD, but also about the exploitation of large corporations using BD systems to collect online customer data. Advocating:

> *"In this context, the government's plan to use social media data for 'studying social processes as they unfold at the level of populations' is bound to create controversy, which may become a barrier for implementation. From an ethical perspective of good governance, it is imperative that the concerns of the UK public regarding data sharing are known, understood and solved"* (SMD0003)

The BD being generated and acquired was also reported to be increasingly locked up and in control of a small group of organisations, presented as easy to use consumer services provided for free use to the consumer. In return for the use of these free BD services, the individuals BD or the performance analysis of it is used by these organisations for business intelligence or data mining. (SMD0013) claimed that the public appear to perceive data mining of BD for commercial purposes as much less acceptable than when utilised to provide free beneficial government services. However, some participants purported that the public were largely unaware of the practice of compiling this BD for secondary use. The BD is in essence outsourced to the individuals it is about, without them really knowing they are practically doing the work related to these services and reaping minimal benefit or reward which raises ethical questions. Participants considered how these practices lead to an imbalanced concentration of power which are not methodologically removed from human design or bias, leading to disproportionate power to the organisations who own these, not the recipients or users of the BD services.

4.3.2 Discussion of Key Findings

The thesis findings build a case that robust ethical governance controls are a major requirement and should be integrated into big data governance frameworks. The examples used to illustrate the inquiry participants position, aided this study to draw out some of the

ethical implications addressed in our proposed big data governance framework that have previously been ignored. The dataset was thoroughly queried to discover distinct ethical concerns and identify patterns connected across this BD risk category, to determine problems experienced in practice and underpin these with views associated to practitioners in both the private and public sectors. Whilst it might be desirable to expand on why associations exist within the dataset and if they might be meaningful, we deem this information irrelevant to meet our goals. The objective here is to comprehend BD challenges in practice and predict potential risk factors within this domain based on evidence submitted.

We draw conclusions from our analysis that there is a huge dilemma on how BD is ethically used, stored and accessed by organisations. Participant's discourse connected to BD ethical challenges are formed by concerns related to the consequences of widespread adoption of BD analytical and mining practices, in which public harm, discrimination and unfairness can be perpetuated. Significant problems with far reaching consequences for organisations are intrinsic, including ethical implications for coordinating transparency, potential for BD misuse or secondary use, verification of BD protection and review of future ethical BD innovation trajectories. Moreover, organisations using BD could infringe on people's privacy, discriminate, evade accountability, use BD for surveillance, manipulate or misinform public opinion etc. If allowed to extend without proper oversight, accountability or governance, these extensive powers could be harmful. Results also showed evidence for a strong ethical criterion in big data governance to enable the ethical capturing, utilisation and storage of BD, in methods that are transparent, fair and democratic. This requirement needs to be addressed at an organisation level, as findings inferred that ethically correct behaviour in BD application enhanced efficiency and improved effectiveness. As we previously highlighted under other section of this thesis, our results indicate that the interrelated ethical implications connected to trust, transparency, funding and training must also be governed.

The thesis study did not analyse in-depth variation in perspectives by participant group nor consider challenges that might be unique by BD type, indicating an area of future research. However, it was noted that conflicting views emerged from the detailed examination of the datasets. Participants from organisations within the public and research sector argued in most cases that it is necessary to consider ethical disclosure of public data and its scope for the

sake of public interest. On the other hand, some private organisations commercialisation of BD pulled them in the other direction of the argument. Participants noted the disposition of BD to push against the boundaries of data protection. It is a tangible issue with various BD applications that some participants stated it has meant an intrusion to private life, heightened risk of discrimination through profiling and of individuals being subject to decisions on the basis of processing that are not transparent to them. It was observed that different risks arise from various BD scenarios. Some advocates argued that compliance solutions can be used proportionately and matched to these different risks. Nonetheless, we induced that some participants recognised current methods such as anonymisation, privacy by design and some other safeguards used for BD protection, to be insufficient for ethical big data governance governance and address this within the proposed big data governance framework. We share a similar viewpoint to some participants that the changing context for potential BD misuse means that compliance with the law is not sufficient to ensure BD initiatives are ethically appropriate. The Medical sector appear to have long recognised that what is lawful may not necessarily be ethical. Which is why they have a code of ethics that is broader than the requirements of the law. In the same way, the dataset analysis we conducted revealed that certain BD usage may not necessarily be ethical. Although morality and law may overlap to some degree, nonetheless from our research, scarcely any organisation concerned with insuring ethical BD practices would agree that regulatory compliance or compliance to the law equates to being ethical (BIG0015; BIG0008; HC245). Given the changing context and potential for BD misuse or re-use, this means that compliance with the law is not sufficient to ensure a BD initiative is ethically appropriate.

There is evidence from our research findings that those that manage BD initiatives, have a continuous role to promote and protect the legitimate rights and interests of the people who provide their BD irrespective of the consent given. Similarly, there were suggestions by participants for people to adopt a leading role in the thinking around ethics for BD at an organisational level. Highlighting that development and maintenance of ethical approaches are factors that would address users and organisations moral and ethical concerns, as well as help the success of BD initiatives. We align with this viewpoint and propose a robust ethical component within big data governance that ensures the right type of BD is being gathered, utilised and stored, in methods that are fair, transparent and democratic as supported by the analysis findings.

Importantly, we established through our analysis that addressing ethical concerns is a vital requirement to ensuring trust between organisations and the public. From the standpoint of addressing the goals of our research, it was imperative we recognised these challenges identified and designed BD governance approaches which build trust into BD systems and processes, ensuring that BD technologies and services are not only reliable, robust and secure, but usage is done in ethical ways. The inquiry participants depiction of the ethical problems is not to be taken as an exhaustive list; but are intended to portray the breadth and complexity of the problems that are related to the big data governance domain we propose to solve.

In the next section, we present the inquiry participants ideas and suggested recommendations that can influence the ethical oversight and governance of BD.

4.3.3 Participants Strategies to Address Ethical Concerns Associated with Big Data

This section presents some participant's recommendations extracted from the research findings. From the inductive analysis method employed in the thesis study, we are able to portray some of the key suggestions of participants to breakdown complex ethical issues related to BD. The advantage of examining these here, is to elicit potential suggestion of significance which might contribute to the interception of unethical based conditions within organisations, which could be formalised as part of the big data governance framework. It is not the goal to give a general conclusive scheme about the significance of each suggestion mentioned or a hierarchical order between them. However, it is believed that the process of examining some of the participants suggestions is something more than simply applying ready-made rules to established facts. There is a notion generally accepted in grounded theory that the value of all other methods remains relative, and each might still contribute to the interception of factors based on conditions (Strauss et al, 1990; Orlikowski, 1993). Therefore, the central method of reasoning selected for establishing the framework is inductive analysis, which also takes advantage of some criteria and suggestions pinpointed by participants below.

- **Strategies to Establish Fairness**

Analysis results support the case that organisation's use of BD and any associated practices should not cause resulting harm, discrimination or unfairness to individuals or groups. It lays emphasise on the need for strategies, principles and processes that do not allow BD initiatives within organisations to widen inequity. The investigation pointed to some key findings on what organisations could do to facilitate fair BD practices e.g. 1) Outline principles within the organisation that establish fairness of BD initiatives, 2) Establish off limit BD usage and BD methods, 3) Establish off limit topics, 4) Train personnel on all ethical guidelines established within the organisation. Similarly, another key finding is the idea to create internal processes to contemplate on and prevent uncertainties, bias and limitations in BD. However, to achieve this goal means that organisations would require processes and funding to ensure the right skillsets are in place.

In-depth investigation determines that there is a vital requirement for approaches to address the ethical dimensions of BD analytics. Apart from being concerned with compliance to data protection and privacy regulations, there is indications that these approaches should consider assessing the ethical basis for BD analytics as well. Notably these strategies should consider the pertinent ethical question that ensure organisations look beyond what they are legally permitted to do with the BD, but causes organisations to ask, "should we be doing this activity with the data?" It is implied that if organisations adopt ethical approaches to control BD analytics, it could go some way to address basic criteria that personal data use within BD initiatives processing was fair. Some participants were of the view that prospectively identifying relevant stakeholder values and interests in any BD initiative is essential. Key suggestions were for organisations to have some process for stakeholder mapping and contemplation on this aspect as an important initial step to discern what informs these interests and where they are located e.g. These interests may be private, economic or political etc. The aim of this crucial reflections is to explicate moral content, which might get the organisation to uncover prior strategic imperatives or value commitments that need to be reviewed or brought into question. Explicating this information may shed light on what is morally acceptable to the stakeholder. Subsequent crucial reflection could highlight to the organisation areas of 'hard constraints' or 'strategic imperatives' that ought to be evaluated.

One participant called for organisations to take extra care to establish those interests that could be particularly at risk, or that could arise from diverse values when executing BD initiatives. Arguing that the following two actions listed below are necessary for organisations to scrutinise which forms of respect for people's freedoms (i.e. consent) and which forms of governance may be required (BIG0015).

1) Identifying why the results of a specific BD initiative might disproportionately impact specific people or groups i.e. the identification of situational vulnerabilities

2) Gaining awareness of how different individuals value the potential benefits or risks of BD initiatives.

- **Wider Public Engagement**

Participants had ideas for wider public engagement, proposing that mechanisms be set up to discuss ethical concerns with members of the public and incorporate their point of view in decision making. Earlier discussions also highlighted that the BD paradigm necessitates increased involvement from the public in how their BD is being used. It seems the overall objective of this recommendation is to identify methods to promote public trust and buy-in of BD initiatives. To this end, these advocates highlighted that the BD paradigm should necessitate increased involvement in how their data is being used. (BIG0080) argues that at a practical level, any system that promotes the flow of BD should grant transparency and control to the consumer, to obtain their trust and engagement. Some participants standpoint is that stakeholder views should influence the design and execution of BD initiatives (BIG0080; BIG0069; BIG0033). They are also of the view that BD stewards could assist with this public engagement to elicit public views and ensure appropriate oversight. Plus, introducing communication plans increase awareness, educate and involve the public on the benefits of BD initiatives. With implications that this could increase buy-in to BD initiatives and foster wider BD sharing.

Additionally, they had ideas for organisations to contemplate BD use, by establishing what existing privacy norms are engaged. Stating that these might include norms sharing data within social groups, norms of professional confidentiality or broader acceptance of BD use.

They argue that key findings from any consultations or research of public opinion would be informative at the initial stages of a BD initiative. Caution was recommended when relying on former understanding, mainly because values, interests and circumstances for individual BD initiatives may differ. Highlighting to us that there is the requirement for ongoing monitoring and engagement.

- **Policies, Guidelines and Processes for Ethical Governance**

A considerable number of participants overwhelmingly endorsed the introduction of a series of voluntary ethical policies, guidelines and processes for BD Use and sharing. Opinions mostly reflected aspirations to balance ethical concerns against the capability to utilise BD in a meaningful way to innovate or gain competitive advantage. Nonetheless, policies, procedures and terms of reference which offer clarity and direction for addressing these issues where widely suggested by participants. Particularly where compliance with the law did not fully reflect moral norms or perspectives. Where there are gaps in ethical regulation, voluntary processes for governance were suggested.

- **Assignment of Roles and |Responsibilities for Ethical Oversight**

Some participants underscored the roles and responsibilities of organisations that participate in producing and administering BD. Recommending establishing a council of BD ethics, BD ethics stewards, ethics review panels and other organisational roles as a means of addressing ethical complexities related to balancing privacy, anonymisation, security and public benefit. Results uncovered that BD stewardship could be organised in a way to proactively safeguard ethical BD use and control, whilst also promoting transparency, fairness and high ethical standards. Another important set of responsibilities suggested for BD stewards or similar is to elicit public views, concerns and expectations in BD, and make efforts to ensure that BD usage align with public expectation and values. (BIG0015) asserting that this should include a pre-requisite of some process of stakeholder mapping and subsequent reflection after opinions are garnered. Suggestions were that BD stewards liaise with external parties such as Community Advisory Boards to provide an avenue for the public to have input into how BD regarding them is collected, stored, distributed, and interpreted.

(SMD0012) recommended that the government should ensure that research involving BD containing personal data is overseen by "ethics review panels that operate under published guidelines, as happens in the academic field".

There was a powerful set of ideas unveiled, with implications that all users of BD should be held to account for their handling and use of the BD. Big data governance guidelines need to be made clear to personnel within the organisation and adherence monitored effectively. With (BIG0030) recommending that "Sanctions need to have real teeth and be applied transparently to those who have broken the rules". Recommendations highlighted participants recognition on how these defined big data governance roles and responsibilities could ensure that an organisation mitigate ethical risks.

- **Focus on BD Ethical Governance**

As reported earlier, participants generally recognised the premise that organisations should not rely simply on compliance with the law to ensure that BD activities is morally appropriate or to address ethical dimensions. Findings determine that robust big data governance approaches offer means to respond the ethical concerns and complexities associated with BD. Including ethical components into big data governance can influence how organisations can engage in ethical and transparent BD activities, find and counter any ethical deficiencies and other engrained limitations, and appropriately frame big data governance guidelines or policy implications. The Information Commissioner's office drew attention to the BD paper where they discussed that some organisations were developing their own ethical approaches to BD, based on the philosophy of transparency, considering customer viewpoints. The goals were highlighted as aiming to forge a relationship of trust with the public.

Several participants also recommended the development of robust and future-proof ethical frameworks to limit risks from BD disclosure or public trust (BIG0016; BIG0057; SMD0009; HC245). We argue that existing big data governance frameworks are limited and need to be overhauled to include mandatory ethical components if organisations are to cope with the

scale, veracity and velocity of the constantly evolving BD environment and the associated BD analytical practices.

The next section 4.4 discusses some overarching organisational challenges that have explicit significance in establishing core big data governance criteria, as determined from the dataset analysis.

4.4 Overarching Organisational challenges

The organisational challenges are revealed to be overarching problems, that apply to organisational areas of effectiveness which could form barriers to big data governance implementation. From our observations made during analysis, we discovered that these could impact organisations at the strategic, tactical and operational level.

Category	Code	Frequency (Grounded)
Organisational Challenges	OC: = ORGANISATIONAL CHALLENGES	0
	oc: bdg	72
	oc: bdg - compliance	45
	oc: bdg - defined roles & responsibilities	60
	oc: bdg - enforcing accountability	27
	oc: bdg - establishing & adopting standards, principles	105
	oc: bdg - establishing and adopting policies, processes, guidelines	149
	oc: bdg - leadership & Sponsorship	21
	oc: bdg - maturity levels	5
	oc: bdg - strategy & roadmap	41
	oc: bdg - transparency & control	117
	oc: culture & mindset	47
	oc: bd monopolies	16
	oc: bd trust deficit	105
	oc: requirement for improved data literacy	35
	oc: skills - recruitment/retention of skilled staff	32
	oc: skillset & capabilities gaps	150
	oc: training / skill dev / awareness	65

Table 18: Organisational challenges group frequency table

As shown in frequency table 18 above, this study categorises organisational challenges into several sub-groups that must be considered for big data governance from the perspective of participants. Not adopting appropriate governance and control of BD usage and its associated

practices is a problem that is referred to by a large number of participants. This big data governance coding group consists of distinct nuances, hence we classified them further into several subgroups: -

- The general lack of focus on the governance of BD (oc: bdg) → c = 72
- Increased effort to ensure and monitor compliance, to complex regulation and legislation in the BD environment (oc: bdg – compliance) → c = 45
- Lack of defined big data governance cross-functional roles and responsibilities within the organisation (oc: bdg – defined roles & responsibilities) → c = 60
- Inadequate organisational structure to enforce accountability (oc: bdg – enforcing accountability) → c = 27.
- Insufficient policies, processes & guidelines established, and the problems related to adopting these within the organisation (oc: bdg – establishing and adopting policies, processes, guidelines) → c = 149.
- Lack of focus on the development and adoption of standards and principles both internal and external to the organisation (oc: bdg – establishing and adopting standards, principles) → c = 105.
- Lack of senior level leadership and sponsorship to support big data governance implementation (oc: bdg – leadership & sponsorship) → c = 21.
- Inadequate establishment and adoption of processes, procedures and guideline to govern BD through the life cycle (oc: bdg – processes, procedures & guidelines) → c = 83.
- Lack of focus on the importance of setting the strategy and roadmap for BD as a strategic asset (oc: bdg – strategy & roadmap) → c = 41.
- Insufficient transparency and control related to BD usage and sharing, due to inadequate focus on big data governance (oc: bdg – transparency & control) → c = 117.

We cannot discuss in-depth all these codes and the entire findings uncovered due to the word count limits of this thesis. However, we try to elaborate on some of the analytical results. The analysis results show that many participants emphasised the pressing requirement to fully consider the appropriate governance surrounding these challenges identified. One example of

this given is that of the National Pupils database, with one participant reporting that the current levels of governance for this does not meet what is required due to the sensitivity and volume of its BD held (BIG0085). Contending that the organisation responsible must start independent oversight and increased governance of BD handling and release. Another participant's view is that apprehensions over privacy, security and ownership of BD should not be utilised to obstruct BD innovation, arguing that robust governance approaches can address these concerns (BIG0046). (BIG0055) supports this view, emphasising that without the right organisational structure, processes and governance framework established within the organisation, the collection and analysis of BD across the organisation will be practically impossible. Arguing that it would result in a siloed approach to BD implementation that hinders an organisation's ability to discover, measure, create & protect BD value across diverse functional areas within the business. As reported by one participant, in their view "*effective governance of the use of BD is indispensable"* (BIG044).

Several actions were discovered during the analysis for addressing these challenges. Although big data governance of an organisation could be at different levels of maturity, any organisation executing BD initiatives is said to require strong big data governance processes, guidelines and policies in place to avoid misuse and misinterpretation of BD. Transparent processes and procedures to govern BD within the end-to-end life cycle, supports the development of trust in BD quality as people can readily see evidence of clear affirmative action on how BD is processed. Or any subsequent trade-offs. Establishing and adopting common policies were also be considered as part of big data governance as is evident in the coding. Participants suggested that organisations must address this in the context of their big data governance policies on regulatory & legislative compliance, privacy, security, BD sharing & processing, IP & liability, BD retention and deletion. Moreover, in the wider context of policies on staff training, staff recruitment & staff retention to address earlier flagged concerns necessary for BD skills development and retention. This is said to increasingly require focus to address highlighted BD skill shortages and support a policy strategy to boost the BD talent pipeline.

Participants called for the establishment of organisation structure of accountability, well defined roles and responsibilities as a means of addressing the emerging BD related

challenges (HC992). Several participants argued that it is crucial to assign people for BD associated activities, establishing accountability for BD requirements and BD processes. Our analysis reveals that this is a crucial success factor for big data governance, with some participants pointing out roles that should be assigned duties for accomplishing effective big data governance cross-functionally. A sample of the roles suggested by participants are outlined in section 6.4.

One of the conditions identified for success is that those assigned play leading roles in clarifying or providing guidance on management, processing, ethical, legal, data protection and other BD related issues. In addition, roles should be assigned with the responsibility to drive the enablement of BD sharing through the breakdown of organisational silo's, ensure the auditing of BDQ and help ensure good big data governance practice generally. Various participants argued that these actions discussed would enable organisation to control and monitor BD, plus give the organisation an opportunity to gain a competitive advantage or obtain value from BD initiatives. (BIG0015) suggests it is likely that the consequences of BD misuse are intrinsically difficult to identify and significantly under-reported; thus, *"those who govern BD initiatives must have a continuing duty to promote and protect the interest and rights of people, irrespective of the terms of any consent given."*

It is inferred by some participants, that those currently charged within organisation are not effective in their roles (HC992). Therefore, dealing with this focus area also means that for every defined role and responsibility, these should be supported with appropriate terms of reference in the shape of guidelines, policies, standards, principles and consistent job descriptions. This ensures that staff assigned to big data governance roles or working with BD are aware of and understand their responsibilities. From the analysis findings, we recommend in the proposed framework that these roles are cross-functional as outlined in chapter 6, as well as cover responsibilities at the strategic, tactical & operational level. We determined that unless this cross-dimensional focus is placed on big data governance, gaps will remain unresolved holistically. Furthermore, the analysis also reveals the need for senior leadership and sponsorship as a factor for BD initiatives if big data governance is to be effective. For example, one participant claimed that "*it was indispensable to gaining board and management level buy-in and funding approval"* (BIG0055). Plus, it was argued that it

was crucial to build strong working relationships between senior level stakeholders and operational teams for the successful deployment of BD projects.

Organisations possessing proper BD strategies and roadmap were identified by a significant number of participants as playing a critical role to improve success of big data governance goals. This consists of various elements. (BIG0075) refers to the commitment on a coordinated and long- term roadmap to sustain strategies for BD collection, BD use cases, as well as long term skills, training and retention strategies etc. (BIG0067) argues that producing a BD strategy emphasises the importance of BD as an asset and prepares the organisation to take full advantage of the BD ecosystem through governance. (BIG0065) flagged concerns about handling BD initiatives as a series of piecemeal projects and initiatives, saying its treatment in general is fragmented and lacks clarity of governance that is important for the establishment of long-term value. The results of our analysis determined that without defining a strategic context for the exploitation of BD, including around how it can be utilised in combination with other internal and external BD sources, the potential will remain largely unrealised. The analysis results underscore the requirements for strategies and policies to facilitate curation throughout the BDLC stages. This is not just a technology problem, but includes strategically governing the kind of BD available, its protection and retention, the collection handling and linking BD to data subjects, as well as the implications of evolving BD processing capabilities.

Interrogating the dataset, several participants argued about the correlation between leadership being vital to set the strategy, roadmap and culture of an organisation. Drawing a linkage to ensuring defined roles and responsibilities, with leaders across functions keeping in touch with BD users externally to focus on BD agendas (HC468a; HC468c; BIG006). In addition, strategically addressing the development training and retention of BD skills is a risk that if organisations do not address, will impact their ability to grow in BD sector or possess the required competitive advantage (HC992). (BIG0039) and some other participants pinpointed the strategic requirement to boost BD talent and address the skills shortage through development of policies and schemes geared towards staff retention. Similarly, strategies that support a clear BD funding commitment was recommended. Investment and funding are important criteria for BD to be a success, but a few participants pinpointed that oftentimes

executive sponsorship and strategic BD roadmaps are lacking. This inevitably appears to lead to insufficient funding, as it is established that initiatives have not been suitably linked to the organisations business objectives.

There is a common theme about the need to develop standards to ensure that BD is useable and fit for purpose both within and across industry sector. (BIG0076) indicated that there were various gaps in BD related standardisation tasks 1) Big Data use cases, definitions, vocabulary and reference architectures e.g., system, platforms, online/offline etc, 2) Metadata specifications and standards including those related to BD provenance, 3) Standards related to BD application models e.g. batch streaming etc. The participant claims organisations across industry sectors are developing distinct approaches to BD management, storage and analysis in silos which adds to the challenge. Similarly, there was said to be the requirement for new BD security standards. Plus, there are also a lack of standard ethical approaches for obtaining consent etc. The lack of co-ordinated approach across sector results in the increased risk of inconsistency, incompatibility & interoperability of BD solutions developed. Several participants recommend the need for coordinated BD standards for new BD systems (BIG0076; BIG0018; SM0008; HC245). Outcomes from analysis show it is an important success criterion to have robust ethical standards developed and applied to enable discoverability, interpretability, and subsequent reuse of BD. A participant claimed that without robust standards, organisations will not be able to build sustainable markets in BD (BIG0076). There is also the risk to accessibility to high quality BD and trusting the veracity of BD from source systems, which in turn impedes the sharing of BD between agencies and its broader use in the BD chain. There appears to be evidence of work going on to attempt to improve standards (HC468a; BIG0076). Notwithstanding, participants discusses the positive the role that the adoption of voluntary standards can play in the area of ethical BD usage. Standards are designed to set out coherent and unambiguous provisions and objectives. They are separate from legal and regulatory systems, which organisations could adopt voluntarily to compliment current legislation. Organisations recommended to champion a set of common standards to ensure BD is captured and shared efficiently across distinct platforms. Perhaps participants views are not surprising. According to Panian (2010) and Wende (2007) standards are a major domain for DG, important to establish DQM through the organisation and assuring compliance with organisation strategy and legislation governing data. Khatri et al (2010) also states that it is vital to determine the roles with the decision rights for

determining standards for DQ. Based on the thesis findings, we recommend that activities to drive the establishment and adoption of ethical standards are embedded into the big data governance process, ensuring the standardisation process addresses the end-to-end BDLC. Before concluding the discussions on the oc: bdg subgroups, it is important to raise the participants major concerns raised regarding insufficient transparency and control of BD usage and sharing (oc: bdg – transparency & control) → c = 117. There is a huge concern revealed about how BD is used, stored and accessed, which is said to have led to an interrelated trust deficit by the public (oc: bd trust deficit) → c = 105. Misuse of BD can damage public confidence in interactions with organisations, with (BIG0038) presenting statistics which indicate that due to high profile misuse, 30% of the UK population can be categorised as non-sharers. This participant argues that fixing relationships of trust between organisations and the public is not only ethical, but mandatory if BD opportunities available are to be realised. From our analysis, another participant states that it is essential for big data governance to build trust, recommending that organisations establish a Trust board to enforce these goals and protect BD, as well as ensure organisations possess the technical capability to be trustworthy (BIG0074). Speaking from the perspective of the Healthcare industry, this participant further argues that *"Driving lawful and trusted collaborative BD uses on a global scale is essential for advancing the research and innovation needed for future healthcare and wellbeing".*

Another participant concluded that, "*regardless of whether citizens feel threatened by BD, they feel they have lost control over their personal information"* (HC 245). In a similar vein, the ICO claimed that *"two thirds of people feel they have lost control of their BD"* (HC468b). Related to this challenge, Chris Combemale revealed outcomes of a study which showed *that "90% of people want more control, so while 72% were comfortable, they absolutely want more control; they want more openness and transparency, and they will work only with companies they trust" (HC468b).* However, putting forward some defence, (SMD0008) stated that for some BD types such as social media, BD spreads fast between BD applications, and although its trajectories might in theory be possible to reconstruct, contending it is not often practical to trace where each BD element goes, stating "*BD changes as it flows through systems e.g. undergoing various annotations, revisions, translations and transformations, making it difficult to comprehend if the BD is the same".* Consequentially, this participant argues that for this and other technical reasons it is difficult to revert BD introduced to BD ecosystems, and difficult to attribute responsibility for so doing.

Examining the frequency count; it is apparent that a considerable number of participants perceive that individual's trust is negatively impacted by a lack of transparency in BD environments and practices. For example, there were concerns raised that, individuals are not often aware of the exact nature of BD processing actions, algorithm logic or decision-making process applied to BD analytics. However, it is suggested that having the right type of skilled leadership within the organisation, who are accountable for big data governance and who are formerly assigned to oversee the safeguarding of personal data used for BD initiatives is vital to reduce this mistrust (BIG0046). Even though legal clarification might be needed, this specific scenario might also require new forms of consent or permission management to be created, whilst good practice guides might also be of use. It is also a recommendation to promote guidance and advice on rights and responsibilities of BD users (SMD0007). Individuals must be made aware, but also be provided with relevant guidance on what can happen with BD and be allowed to provide or withhold their informed consent, irrespective if some BD projects seem to be involved in the concept of implied consent. Big data governance is deemed to be an avenue to ensure this transparency, control, trust and compliance. Ensuring that appropriate organisation structures, processes, standards & policies are in place to facilitate privacy & security, as well as proper ethical, technical and legal practices.

From the observation made, we also found other organisational challenges subgroups which closely linked to human risks factors. These subgroups are closely interconnected and refer to problems that influence the successful implementation & adoption of big data governance in organisations. One particular participant stresses that these human factors require more attention than technology factors in enabling the continued evolution of BD provision and usage (BIG0048). These challenges are categorised into the following subgroups: -

- Bottleneck in changing organisation cultures and adapting big data governance to a new way of effectively deriving value from BD (oc: culture & mindset) → c = 47.

- Lack of people with the right skills and capabilities related to implementing BD initiatives and associated big data governance to drive success in the organisation (oc: skillset & capabilities gap) → c = 150.

- Inadequate strategies and processes to support staff training, internal and external awareness and skills development (oc: training/skills dev/awareness) → c = 65.

- Issues impacting the recruitment and retention of skilled and competent staff to drive the responsible exploitation of BD (oc: recruitment/retention of skilled staff) → c = 32.

We discuss some of these subgroups further below: -

- **Culture & Mindset**

Participants contend that there are cultural risks to the success of BD within organisations (BIG0063; UKBIG0047), stressing that these cultural problems need to be overcome. One participant argued that driving a BD culture is important, as BD is the new competitive advantage for business (BIG0060). IDC explored the impact of BD on business by surveying more than two thousand firms and discovered that there are two types of organisations: Those which have taken a leadership position when it comes to BD; and those that are not making best use of BD. Discovering that the leaders i.e. businesses that embraced BD, derived significantly more dividends from those investments in BD e.g. increased revenue, improved productivity, and reduced costs. Another participant revealed that *"data–driven companies are over 10% more productive than 'dataphobes' firms that don't exploit their BD"* (BIG0047). Claiming they estimate that if all such dataphobes were to make good use of BD in driving their business decisions, they would see an improvement in competitive advantage. Examples that wrong culture and mindset were impacting negatively on the organisations were provided by some participants. For example, reflecting on the review, one participant discussed disappointment stating *"I was very disappointed when we revisited the new 'Caldicott principles' … to find that the culture in the NHS of sharing information had not moved in the way we hoped. Today, the benefits of sharing patient BD have yet to be realised"* (HC468c).

The implications are that making a step change in this direction is not a technical activity, the bottleneck is in changing cultures and adapting governance to a new way of attributing ethical practices. The analysis results identified a few activities that are recommended to address this risk. In common with many change projects, senior managers need to ensure BD initiatives are not undermined by employee resistance to change. This means that

implementing a change management plan as part of the big data governance programme is an important element for changing mindsets of staff (BIG0045). One participant argues that this needs to be a board level issue and chief executives need to be responsible for putting in place the right culture across their companies. In addition, principles that could act as a road map for putting the right ethical culture into the organisation, is seen to be as important as details of the rules that compliance officers might look at. What is significant from all the insights provided by participants is important criteria for organisations to adopt BD-centric ethical cultures to support big data governance decision making.

- **Big Data Skillset & Capability Gaps**

Participants of different sectors involved in the inquiry were unanimous in their concern about the lack of multi-skilled personnel in the context of BD and BD analytics. The skills gap and the shortage of staff with the right BD competencies was said to be a major concern, which if not overcome would impede an organisation's ability to be leaders in BD development, adoption and exploitation. Furthermore, this risk is said to impede an organisation from deriving value from BD as a strategic asset. Participants argue that for BD to succeed the appropriate technical capability to capture, store, process, secure, publish and maintain BD needs to exist. Similarly, business knowledge to define and articulate the desired business outcomes and successfully sponsor initiatives also need to exist. Furthermore, business insights to ensure resources are effectively deployed are also needed to ensure a successful strategy. To buttress this point (BIG0053) stated that Mckinsey had predicted that *"in 2018 the US alone faced a shortfall of 140,000 to 190,000 individuals with the required expertise and a shortfall of another 1.5 million managers and analysts with the skills to understand and make decisions"* connected to BD. The recent literature review conducted as part of this thesis study also reveals this as a problem. It was argued by some that the problem related to the trends in BD skills demand would not abate in the future. A high number of participants argued that skills development is of critical importance at all levels, hence organisations providing opportunities to train, up-skill their staff and embed new skills is essential. This requirement in itself is a challenge, given that BD permeates all areas of an organisation operations and inevitably the requirements are diverse. Repeated themes and links were revealed between the reported BD skill gaps and the shortage of capabilities to develop most of the BD tools, techniques and methods to handle the various

challenges identified. This points to the need for an organisation's big data governance framework to put in place the strategies, policies and processes for recruitment, training and skill development. Partnering with the government and other parties where applicable. Notably some participants also recommended supporting career paths and training standards to aid skills development and retention of the multidisciplinary skills required. Also focusing on the right levels of competence or awareness of big data governance processes and procedures. Another idea was that Leadership should address these challenges in their strategies and strengthen the funding models to ensure adequate funding to support the success of BD skills development and training. We infer from the analysis that leadership would need to strengthen funding for other BD aspects and maintain control over cost/operational expenditure. We suggest establishing a big data governance strategy where organisations commit to a policy that supports them to access, develop and retain BD talent. As with other BD challenge categories, we also contend that the risks are lessened by having a realistic finance plan to support this.

Focusing strongly on the inquiry participant's different descriptions of BD challenges and risk factors, it emerged from participants of the inquiry that there are considerable BD issues constituting a dilemma to organisations, where robust governance is required to control cost, resources and other BD dimensions in order to extract the expected value and return on investments. What is also evident is a core requirement to consider the end-to-end BDLC, in which driving value through BD should be considered in a well-defined structured manner throughout the governance process. As a result of this analytical observations, a core property of our big data governance framework contains guidelines to assess risk factors that can occur through the BDLC. Even during early stage of this assessment, BD risk identification through the BDLC is vital to ensure the right decision points are determined. Figure 19 below starts to illustrate examples of some of the constructs uncovered:

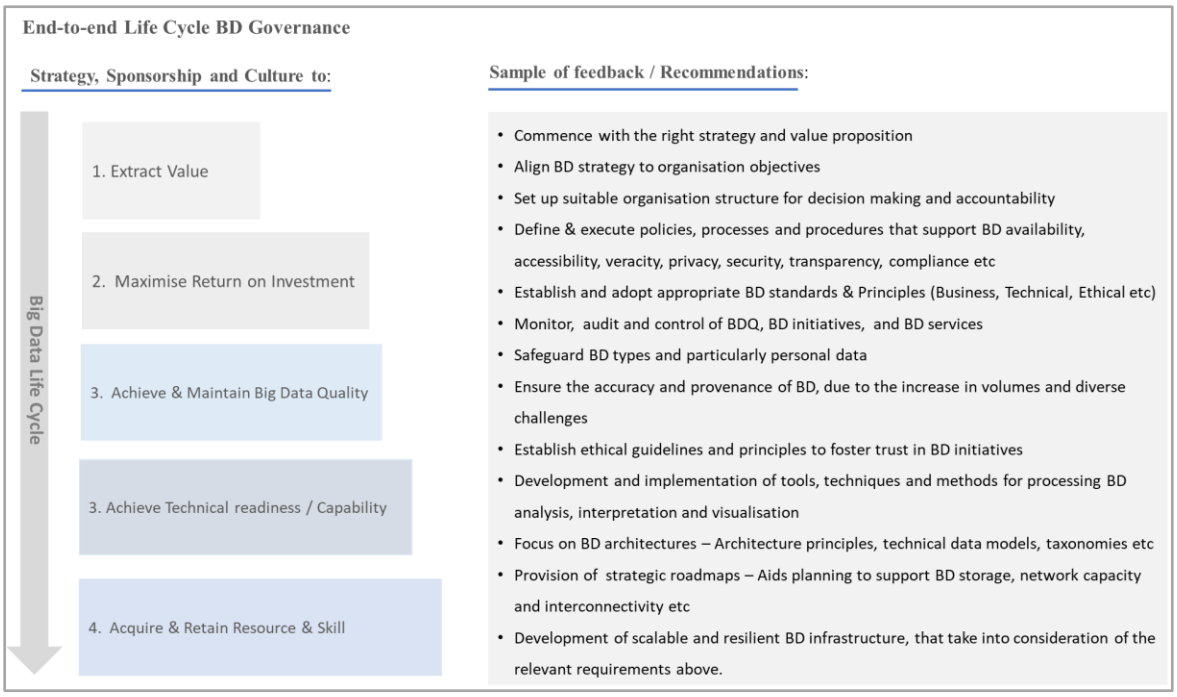

Figure 19: A cross section of participant feedback & recommendations

4.5 Summary of Chapter 4

In the preceding chapter it is explained how the analysis results unveiled specific BD challenges and risk factors, disclosed through investigation of the inquiry participant's submissions from both the private and public domain. The knowledge and feedback of these participants aided the study to ascertain some of the relevant BD issues to be addressed by the proposed big data governance framework. Moreover, the findings determined that participants within different sectors are fully conscious of the magnitude and multiplicity of the BD related problems which is consistent with one of the motivations of our research.

Informed by the analysis, we also revealed key ethical concerns relating to BD. Participant's perceptions of different ideas to address these ethical concerns were also examined. By examining the ethical concerns and subsequent suggestions to resolve them, useful insights were derived about how organisations see various aspects of these ethical issues related to BD.

There was enough evidence from the research findings which point to 1) a major requirement for organisations to refocus their approach regarding big data governance and 2) distinct

concepts that influence intervention by big data governance. Consequentially, these elements form the constructs of the proposed conceptual big data governance frameworks which we present in the next chapter.

5.0 Building Blocks of a Big Data Governance Framework

5.1 Introduction

In previous chapters a clear gap is identified that there is no suitable big data governance framework to effectively govern the additional BD challenges and risk factors that impact BD initiatives in organisations. Earlier DG frameworks concentrated on data quality and the governance of data in traditional IT settings, as outlined in chapter 2. The only three big data governance frameworks that were earlier identified were either limited in scope or did not explain the methodology adopted to derive the framework. Thus, this chapter introduces a novel big data governance framework, proposed as part of the thesis to improve big data governance within organisations. Hence, answering the overall research question and fulfilling the main objective of this thesis research.

The framework is created based on the grounded theory methodology employed and the identification of BD challenges, as well as influential big data governance requirement concepts. Conclusions are drawn from the inductive analysis that taking a holistic framework-based approach will address some of the key contentious BD challenges identified in the inquiry datasets examined. These key BD challenges identified as part of this thesis research, are discussed in previous chapters. In addition, building on past history of why DG was originally utilised for traditional data, we significantly extend the framework of Khatri et al (2010) as a basis for our framework, based on our inductive analysis findings. Furthermore, we provide here a novel component for the proposed big data governance framework, which contributes to support and guide organisation's big data governance programmes to address ethical concerns uncovered. This proposed ethical component is represented by two distinct parts. The first relates to the different principles that assist to set the direction for ethical decisions and establish some boundaries for the ethical usage of BD within organisations. The second part aids to elicit the design and execution requirements for ethical decision points through the BD life cycle Now this chapter describes and justifies the components of the framework, to design, execute & maintain effective big data governance within an organisation.

5.2 The Proposed Big Data Governance Framework

Khatri et al's (2010) earlier DG framework design focuses on five decision domains: Data quality, data principles, data lifecycle, data access and metadata. This framework provides a simplified, but one-dimensional perspective on DG which centre's their framework on some core data decision domains (see figure 20).

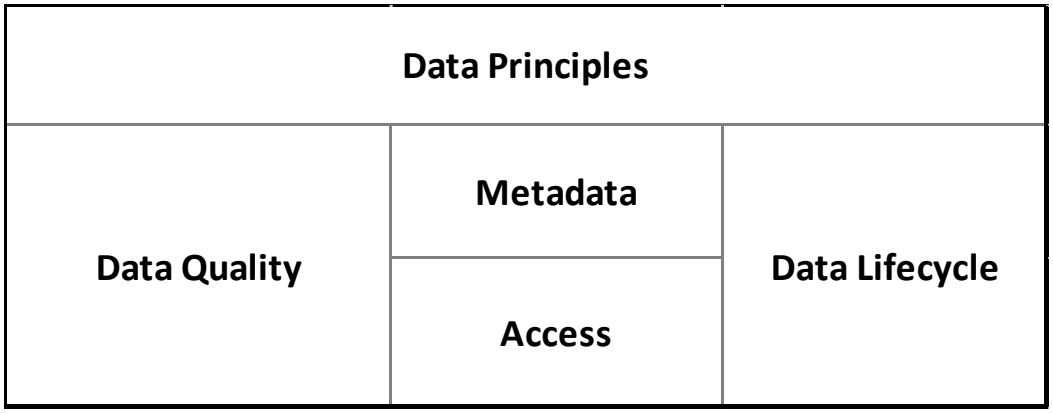

Figure 20: Decision Domains for Data Governance (Adapted from Khatri et al, 2010).

Evidence from the inductive analysis reveals it has limited depth in the interconnectedness with the organisation & governance processes related to BD challenges or risk aspects. By contrast, the design of the proposed big data governance framework consists of 4 strategic goals, 4 enablers, implementation stages and 22 key components that influence how organisations can address BD challenges or risk factors as shown in figure 23. The proposed strategic goals should remain at the centre of any big data governance programme initiated and are namely 1) the oversight of BD, 2) BD risk reduction, 3) obtaining the expected value from BD and 4) ensuring the transparency and control of BD. These strategic goals are the main objectives and results an organisation should aim to achieve over time. The four crucial enablers support the big data governance programme's effectiveness, as well as help the 22 key components of the big data governance programme to successfully influence how an organisation addresses the BD challenges and risk factors that confront them. The four core enablers are 1) risk assessment, 2) communication and awareness plan, 3) change management plan and 4) competencies development and training. All components of the proposed big data governance framework are depicted based on this foundational framework,

and the extensive findings from inductive analysis undertaken. Rather than purely utilising the literature, the inductive analysis research findings assisted to present real-life necessities as reported by the practitioners to support designing the novel big data governance framework and significantly expand on Khatri et al's earlier work. We explain the stages of the big data governance framework next in section 5.2.1

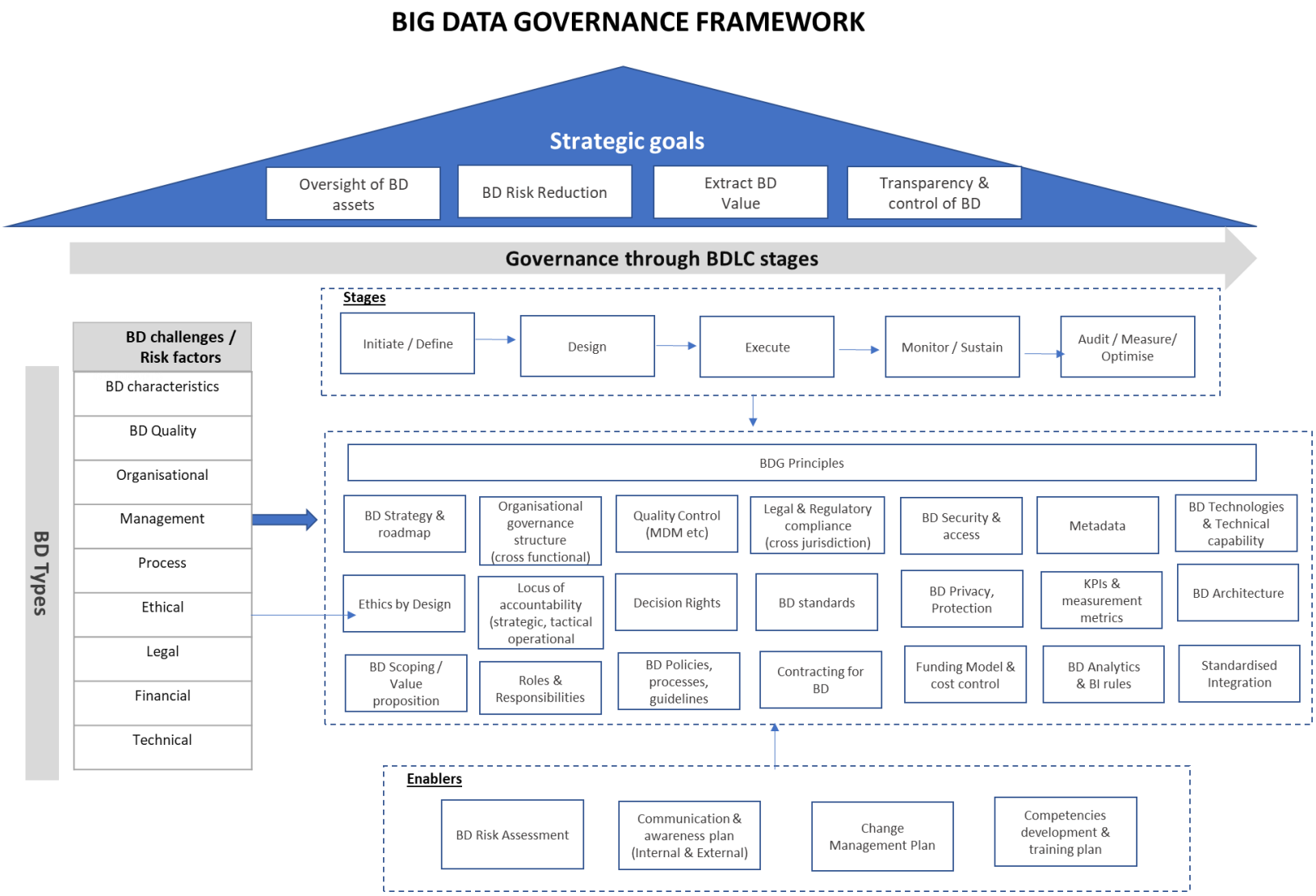

Figure 21: Dimension of Proposed Big Data Governance Framework

5.2.1 Stages of the Big Data Governance Framework

Considering the main goal of this thesis study which is to apply big data governance to alleviate the additional problems placed on organisations involved in BD initiatives and considering there is a key requirement to start the governance process from the very early stages of planning, it is necessary to outline the stages that components can be mapped to in order to apply them. It ensures that a clear and concise outline of all stages needed to execute the big data governance approach is understood. These stages are shown in figure 24 below. The first stage is the very beginning that covers activities that defines the BD scope & value proposition, ensure Senior Leadership sponsorship and buy in, carry out BD challenges and risk identification, understand big data governance requirements in order to mitigate against unsuccessful initiatives, and agree big data governance organisation structure. Furthermore, the design of big data governance organisation structure should consider the proper cross-functional distribution of roles and responsibilities for BD related decision rights. Looking at this from the strategic, tactical and operational level. It is important also to mention that a vital task at this stage is to ensure that key success criteria, metrics and key performance indicators are defined for big data governance as these will be very important for measuring the success of big data governance programmes implemented.

The second stage is the design stage which must come after a circle of BD risk identification to fully understand what mitigations are necessary. This stage will involve the design and creation of the necessary big data governance policies, procedures, processes, standards, principles an organisation will utilise to drive desired behaviours in BD usage, BD quality management, ensure compliance with regulations, protect BD assets from unauthorised usage or access etc. These are specifically important, as all stakeholders must have the appropriate terms of reference to guarantee that big data governance stands a chance of succeeding. Additionally in this stage the communication and change management plans are designed.

After these initial stages, comes stage 3. This is the stage that an organisation can execute their big data governance programme and commence to govern their BD assets utilising a formalised and controlled mechanism. The expectation is that all big data governance stakeholders would have been trained to perform their expected responsibilities and an

ongoing training program exists to ensure core competencies are maintained, irrespective of changing demands. Stage 4 is the stage where the big data governance programme is monitored & sustained ongoing. The stage is required to ensure that monitoring and checks are put in place that is successfully being adopted. Moreover, is necessary to establish the culture and mindset of continuous improvement for the big data governance programme. Stage 4 is closely linked to stage 5, to audit, measure and optimise big data governance. At this stage, a big data governance review is carried out on the progress of the big data governance programme, measured against initially defined KPIs and metrics. Following this review the necessary adjustment to optimise areas of big data governance are carried out where necessary. The advantage of this stage is to enhance the organisations big data governance maturity levels and consequentially improve the outcomes of BD initiatives and related practices.

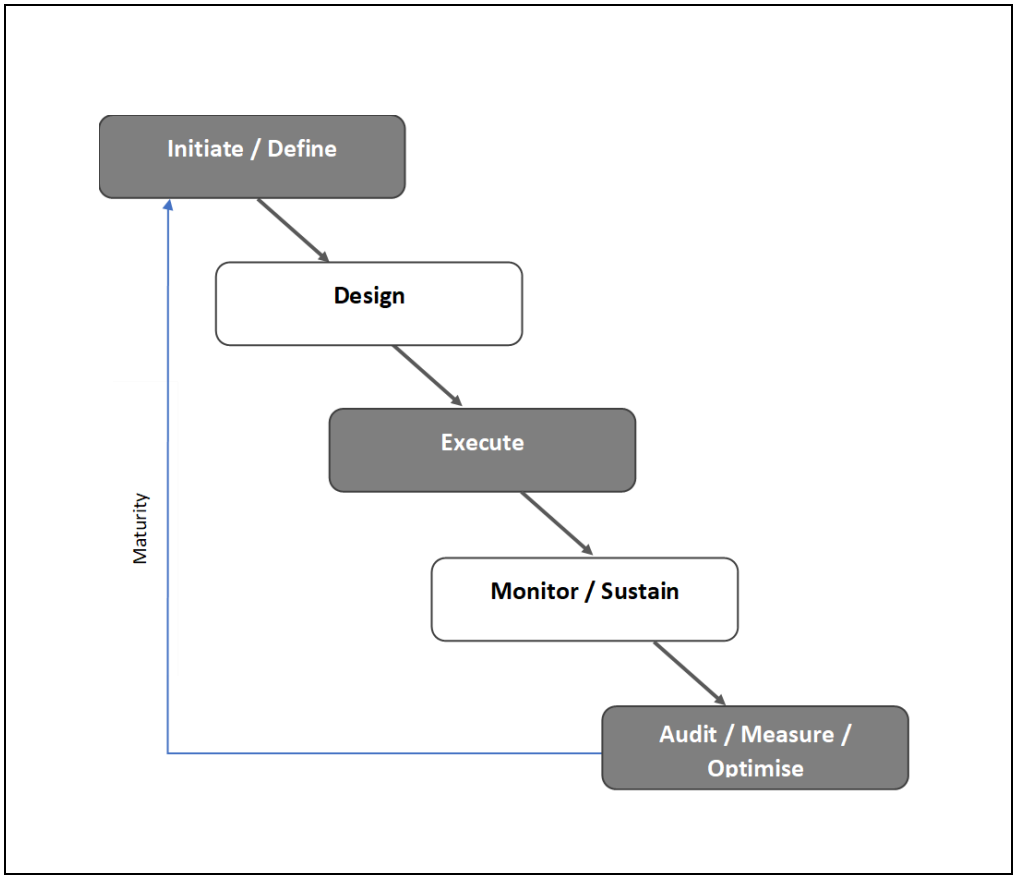

Figure 22: Big Data governance framework stages

5.2.2 Key Conceptual Big Data Governance Framework Components

It is important to focus on defining, analysing, measuring and optimising big data governance. The major principles of our big data governance building blocks are hinged around BDLC decision domains and governance controls consisting of the following components discussed below: -

- **BD Strategy & Roadmap**

The strategy and roadmap elements of the big data governance framework accelerates the orientation and alignment to organisational strategy and supports organisational goals to achieve a return on investment. Investigations uncovered that successful BD initiatives need a notable coupling of strategic direction, objectives and vision that big data governance must mobilise. Consequentially, by executing the big data governance strategy, BD initiatives & use cases will align with core organisation objectives, ensuring business value can be derived by aligning to business growth and operational efficiency goals. This also provides benefit by enabling the organisation to strategically determine business as usual people and processes for BD operations. Subsequently, this big data governance component enhances an organisations capacity to allocate required resources, allocate funding and manage priorities. Moreover, introducing this strategy driven focus enables big data governance to mature rapidly.

A roadmap is a short to long term plan for BD initiatives related to the BD technology, process and people involved (Soares, 2012). Soares (2012) has also shared our viewpoint and those of participants, that the big data governance program should have a roadmap that increases the visibility and importance of BD assets to senior leadership and provides of potential strategic or tactical requirements; for instance, funding, resources, decommissioning of legacy system and infrastructure etc. We recommend defining a clear roadmap, that assists the workstream that will assess people implications e.g. organisation structure design, culture, recruitment, training, skills development and support. In terms of technology, a robust strategy and clear roadmap aids the organisation to strategically develop a best practice BD environment which is future-proofed, highly scalable and can be leveraged across the

organisation. More importantly, it prevents ad hoc and reactive BD problem resolution and non-standardised BD processes and environment.

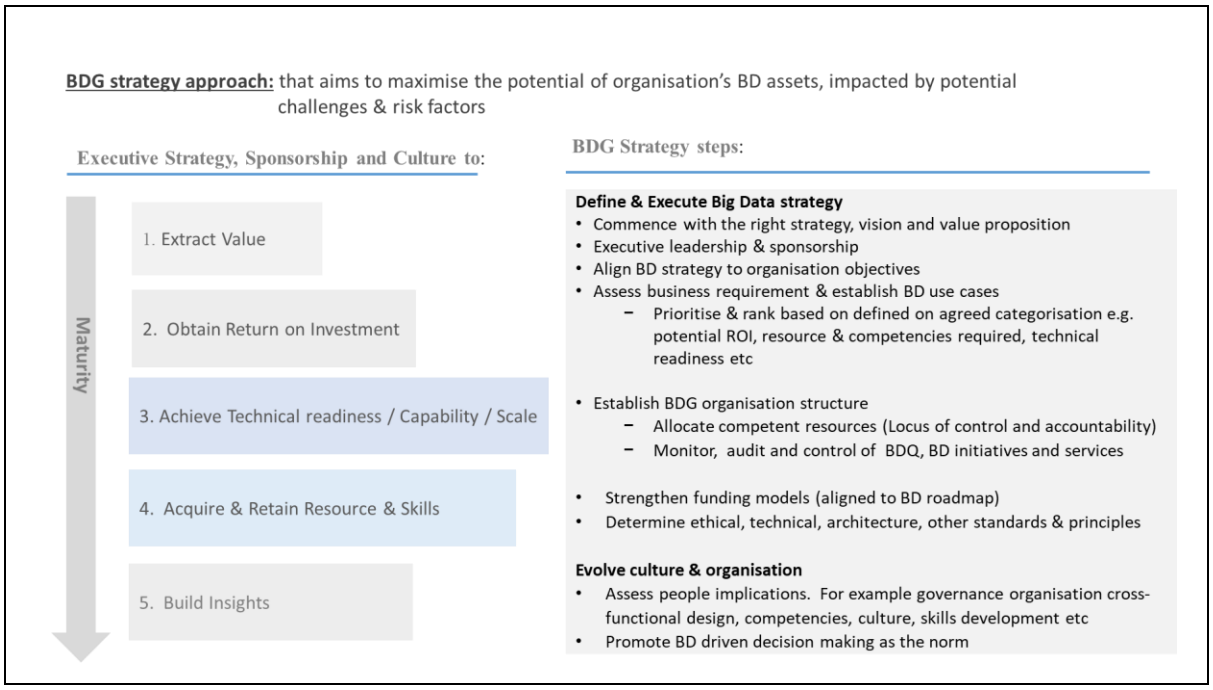

Figure 23: Strategic Context for Big Data governance

- **Big Data Scope and Use Cases**

Big data governance produces the required structure that should guide organisations to obtain value of BD as an organisation's asset. Not all BD use cases will be of value for every organisation, nor would all BD immediately benefit from being collected by the organisation e.g. the overall cost might not justify the rewards etc. Hence, as a vital step of the big data governance framework, we propose the organisation should first scope their BD use cases in the first step. Then prioritise and rank these BD use cases based on cost benefit, technical readiness, resource skillset capability etc. Participants relayed that sometimes there is confusion in the area of scoping appropriate BD use cases. To solve this aligning BD initiatives with the organisation's core business strategy and objectives will provide clarity for this process. Objectives define the goals and strategies of an organisation which will inform priorities. Furthermore, aligning to the organisation's business objectives will ensure

the focus is on the right BD value proposition. Key BD item prioritisation can be driven by the following principles:

- Principle 1 – Business goals (delivers value, builds towards vision)
- Principle 2 – Feasibility (capability, within current organisational constraints)
- Principle 3 – Quick wins, (optimised timeline & immediate impact)
- Principle 4 – Compliant to architecture principles and targeted future state

This was also depicted in the investigation outcomes; Organisations should commission all BD initiatives and related value propositions with senior leadership sponsorship.

- **Big Data Risk Assessment**

As a core step in the initiate stage all relevant BD challenges or risks related to the specific BD use cases are identified. This stage informs what potential or future risks related to the BD initiative needs mitigation. Furthermore, at what level of risk they could impact (e.g. high, medium or low impact levels), and what priority should be given to addressing this risk based on resource, budget, time available etc. Traditionally, the task of risk identification is to gather these from stakeholders e.g. SMEs, BD users, customers etc. For large BD initiatives there could be a large number of risks identified coming from numerous stakeholders. Such early risk identification aims to ensure that the organisation is unambiguous about their BD use cases and fully reflect on both current challenges and future risk that needs to be avoided. Another main reason for suggesting this approach for the proposed big data governance framework is the fact that this component allows the potential BD challenges e.g. Ethical, Technical, Legal, Privacy infringement etc. to be systematically analysed within the governance process. In subsequent decision making, any additional big data governance accountabilities, activities, guidelines, training or soft goals could be identified for relevant stakeholders both internal & external to the organisation. In order to have the context, this closer inspection makes the mapping process to organisation dependencies or to locus of accountabilities easier. Dependencies are also on other third parties in order to achieve big data governance goals, perform big data governance activities and execute required mitigating processes. One of the main contributions of this thesis study is that the robust thesis findings in chapter 4 can be utilised as terms of reference for organisations review potential BD challenges and risks that they could be confronted with. This reduces the need

to start this exercise from scratch. Specifically, it serves as an aid for organisations at differing levels of big data governance maturity.

- **Organisation Structure**

Findings produced valuable insights on the criteria for the organisational aspects of the framework. In an effort to clarify this big data governance element, many participants pinpointed various roles and responsibilities that would form a locus of big data governance accountability in organisations, clarifying that it is vital to shape the big data governance organisation structure that is responsible for defining, executing, monitoring and sustaining big data governance, as well as performing big data governance related decision-making activities. Indeed, some of the earlier works in DG have included the identification of roles, responsibilities, accountabilities as core dimensions of their DG approaches (Khatri et al, 2010; Weber, 2007; Otto, 2011b; Alruthie et al, 2016), mapping these to organisational placements and a hierarchy of decision-making authority. In our big data governance framework we are initially doing the same, but since we are including a unique ontological method of BD risk identification to formalise our framework, thus we can uniquely and automatically map these dimensions together. By contrast we are taking design patterns which specifically focus on and include the locus of accountabilities connected to BD dimensions. Develop locus of accountabilities & responsibility assignment matrix to clarify which big data governance stakeholders should be responsible, accountable, consulted and informed - ensuring this takes into consideration the existing structure of the organisation. Each responsibility assignment should be mapped to the right big data governance processes. Where applicable include these big data governance responsibilities to job descriptions and annual appraisal objectives to drive the right behaviours. Table 19 presents a group of diverse roles and responsibilities which form the suggested big data governance roles and organisational accountabilities assigned to these roles

Earlier research shows that DG involves organisation, roles and decision-making to ensure the control of data value. Khatri et al (2010) support this standpoint. In their view DG establishes organisational decision-making responsibilities related to data assets. Our findings also provide evidence of this, thus elements of our big data governance comprises the set-up of the right organisation structure. Participants identified that it is crucial to have appropriate

leadership, as well as big data governance roles & responsibilities that drive a BD centric culture and promote the desirable behaviours in BD usage.

	Roles
Strategic	Executive & Senior Leaders Executive Sponsors Chief Data Officers BD Architects
Manage	BD Stewards Architects BD protection officer Compliance officers Data controllers Legal personnel
Coordinating & Approval Committees / Boards	BD Steering groups (internal & external) Ethics BD Steering Committee Council of BD ethics BD Trust Board Community awareness group
Tactical	BD owners
Operations	SME's (Business & IT) BD legal experts BD Security experts Procurement BD Tech support

Table 19: Example of big data governance roles identified by participants

What is distinct about the components of the proposed framework is we have been able to identify new obligations and responsibilities to be considered within big data governance programmes. Apply these in big data governance specific context and align them to roles and responsibilities uniquely identified to address BD complexity aspects e.g. BD ethics councils, Community Awareness Boards, BD Architects, BD owners, Trust Councils etc (see examples in table 19). The right organisation and leadership structure is a paramount component of big

data governance. For example, the big data governance organisation should be constituted of a hierarchy of Executive leadership, BD stewards, BD architects etc with the responsibility for defining and setting direction on BD principles, BD standards, BD guidelines, BD taxonomy and rules for BD usage or sharing. Additionally, clear roles accountable for the quality, availability, reliability, transparency and trustworthiness of BD should be covered within the structure.

Finally, the design of our big data governance organisation structure is cross functional, complimentary with the existing structure of the organisation at strategic, tactical and operational levels. This is necessary as the organisational structure gives a mechanism to ascertain priorities, establish cooperation, and enable communication between stakeholders. A relevant finding uncovered during analysis, is that the roles and responsibilities identified by participants should have forums to interact with each other in various methods. However, it is important to define the locus of accountability (Khatri, 2010; Otto 2011a, Cheong et al, 2007; Wende, 2007). Thus, although forums for interaction should exist, we argue it critical to identify the roles in big data governance, but also the boundaries of responsibility. Thus, we advocate this as part of the framework.

- **Leadership & Executive Sponsorship**

Senior Leadership is crucial for the success of big data governance. A committed leadership team is an essential component to set the strategic direction, foster a BD centric culture and mindset that promotes the desired behaviours, as well as the highest level of BD quality and compliance. Their sponsorship of the big data governance programme accelerates the strategic direction and sets the appropriate priorities organisation wide. Without strong leadership backing, the prioritisation of funding, training plans, BD quality oversight, BD skillset development, technical capabilities, and all other areas which the organisation should be cognisant of will lack focus and big data governance could fail.

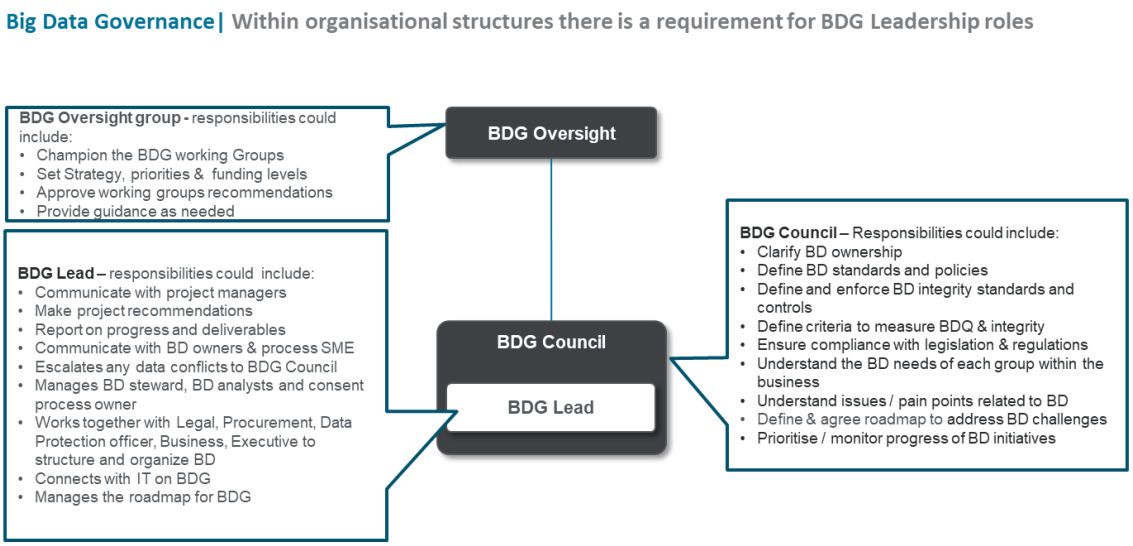

Figure 24: Example big data governance leadership roles and responsibilities

- **Communication Plan**

A coherent, accessible overarching big data governance for the quality control, availability, transparency and trustworthiness of BD is required. Given the complexity of existing governance processes hinder their interpretation, as well as the capability to communicate them efficiently, we advocate a coordinated communication strategy as a key component of the proposed framework. This communication strategy should be designed to reach the widest number of big data governance stakeholders. As discussed earlier, the lack of awareness related to big data governance was contended by participants to be a major stumbling block in advancing big data governance. Thus, the communications strategy component of the framework is responsible for advancing enterprise-wide awareness of BD asset governance. The goal of the component is to clearly communicate big data governance policies, guidelines, standards, principles etc to ensure that all stakeholders are aware of these and any associated big data governance activities. Raising awareness with the public also provides transparency to the organisational controls in place, which in turn develop trust in organisations BD systems, processes and people. Moreover, an effective communication strategy forms a clear framework for discourse, resolution and driving awareness, consequentially aiding problem solving.

- **Big Data Governance Policies, Processes & Guidelines**

Establishing and adopting clear and transparent policies, guidelines and processes is a core pillar of our big data governance framework. As explained, we were able to apply the results of our research analysis to have identified these as key components of the framework. The big data governance obligations and activities will not be successfully implemented without clearly defined processes, guidelines and policies. These offer appropriate terms of reference, clarity and direction which aids the organisation to resolve some of the BD related challenges. Policies are required to facilitate curation throughout the lifetime of BD assets. Policies for the distinct aspects we identified in earlier chapters should not be ignored e.g. Policies to cover:

  - appropriate handling and reporting of BD usage, sharing & processing to guarantee compliance with regulations or legislation
  - Privacy & Data protection
  - Security management etc

In this context clear and communicated policies and processes will encourages the transparency of big data governance, allows organisations to fulfil obligations to show publicly what they are doing. In turn this enables discussion regarding them and shows proof that there are big data governance procedures in place to hold the organisation to account. Here we propose big data governance processes that include those that

  - Minimise risk of privacy infringement
  - Improves BDQ management
  - Reduce risk around misuse and misinterpretation of BD etc.
  - Preserve public confidence and trust etc

This not only allows focus on outputs of defined processes etc, but to continually monitor and measure whether the big data governance achieves the required results. Another advantage of these components is that they enable personnel to have suitable direction and enhances confidence across the organisation that personnel are supported to apply big data governance appropriately in a manner that assists at driving value from BD assets.

Organisations should classify & implement the procedures to meet the BD safeguarding needs identified by the BD risk identification process in earlier stages. Be careful to consider

evolving regulations and customs when establishing policies regarding the acceptable use of BD types e.g. Social Media data, PII data etc. The senior leadership team of the organisation involved in the strategy should work with those involved with the day-to-day processing of BD to establish the right policies & transparent processes of accountability across the organisation that can be adopted and understood by all stakeholders. We argue based on the evidence of our research, that without these an organisation cannot be trusted to collect or use BD of customers or citizens responsibly.

- **Principles and Standards**

We consider BD principles and standards to be primary components in the proposed big data governance framework. Principles were derived from some existing principles recommended by Khatri et al (2010) and from inductions made from analysis results. The framework focuses on the execution of principles to be implemented as part of the big data governance programme, to be used to guide the design, development and implementation of policies and controls related to big data governance risk. Setting principles for organisation decision making related to BD, establishes the linkage with the business and determines the extent to which BD is an organisation wide asset (Khatri et al, 2010) and thus what particular policies, standards and guidelines are appropriate for successful usage and sharing of BD. In addition, they are utilised as a basis for monitoring, benchmarking and auditing facets of the organisations BD initiatives or practices.

There was particular attention related to establishing standards by practitioners participating in the inquiry, as it related to governance controls on integration, improvement of master data management and BD quality. From the technology perspective the organisation should focus big data governance on developing standards for BD architecture, BD infrastructure, BD technology; BD analytical applications etc; recognising the importance of considering these in the big data governance programme to address gaps in this area as discussed earlier in chapter 5. BD standards should facilitate BD quality, BD integration and interoperability through standardised BD protocols, common BD taxonomies, standardised language and syntax, metadata etc. According to Khatri et al (2010), different kinds of metadata perform a role in the discovery, analysis, retrieval and collation of BD. The main benefit of the component for establishing and adopting standards is to

1) Link diverse cultures within an enterprise

2) Bridge different BD languages, definitions or semantics

3) Clarify disparate BD sources

4) Reduce inconsistencies that affect the availability, veracity, accuracy, interpretability and accessibility of BD.

- **Change Management**

Effective big data governance should be able to adapt to the changing BD landscape, disparate BD ecosystems, changing regulations and rapid BD innovation climate. Being able to quickly adjust to this changing environment is crucial, hence the change management component proposed in our big data governance framework is a necessity for sustaining success, irrespective of both the internal and external changes that might cause problems. The organisations big data governance change management procedure should enable the organisation to make agile changes as required to address BD challenges. Change is continual; thus, organisations must be flexible to adjust processes, policies, guidelines etc or the organisation could end up with stagnated big data governance procedures that will inevitably affect outcomes.

- **Focus on BD Security**

The security big data governance component defines and applies end to end lifecycle security controls both at the business and technical levels. For big data governance it is evident that security dimensions are core criteria for organisations, considering the weight given to this by participants. The component provides coverage to activities, including applying technical approaches such as identity, access and permission management. Or defining security policies and procedures which support organisational principles related to security. The organisational security principles should promote policies and processes which help prevent the existence of BD security risk concepts, and meet the discovered BD requirements for confidentiality, integrity, availability, authentication and authorisation. BD assets should be held in a secure location, whilst in turn being accessible when needed to authorised people with the approved access.

- **BD Architecture**

In the context of executing big data governance from a BD technology perspective, the proposed BD architecture component involves cultivating architecture standards and best practices which decrease some of the inherent risks in BD environments and BD technology implementations. Managing the BD architecture is a crucial competency without which any big data governance organisation will struggle. The BD architecture component should facilitate the futureproofing of the BD environment & infrastructure, so it is extensible and scalable to be redeployed across multiple BD platforms and BD technologies. Ensuring that the organisation can be flexible to accommodate the different BD types and BD velocity, whilst aiding accuracy and reliability. Moreover, BD architecture standards guides the forward-looking design of the BD landscape to foster BD quality, integration, interoperability, and usability. An organisation cannot effectively govern what it is unknown to it. For big data governance purposes this part of the component specifies the requirement for documentation that describes the BD resource and illustrates how the components fit together. One of the major challenges identified with BD was the complexity related to ensuring the provenance of BD amongst other issues. It was induced from the research that the minimum architecture documentation includes:

- A model that provides descriptions of the different BD domains
- Logical models that describe business views of the BD
- Entity definitions
- Implementation maps illustrating the physical location of the data entity
- Documents showing the of business unit interactions with BD entities i.e. creation, reporting, updating, deleting.

The architect also responsible for documents that show BD flow, data lineage, and attribute/element definitions. In changing environments, “current state” and “desired state” models are valuable (TDWI, 2010), thus this is imperative for the ever-changing BD environments.

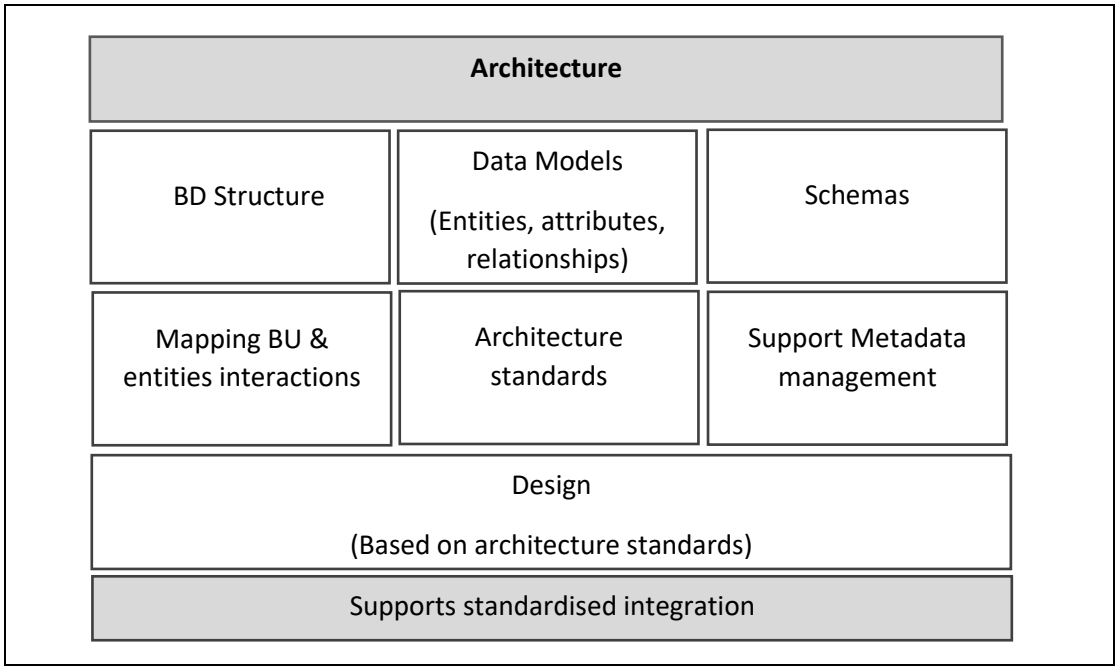

Figure 25: BD Architecture frame of reference

- **Oversight of Big Data Quality**

The main assumption for this thesis study is that affirmative action must be taken to positively reduce the risk of poor BDQ. As discussed by earlier researchers, poor BDQ can affect an organisation at both strategic and operational levels (otto, 2011; Soares, 2012; Khatri et al, 2010). Without BDQ the data will not be fit for purpose and there will be no way to obtain the expected value from BD assets.

The conceptual framework component was thus situated in a way to support the BDQ criteria as defined above. The eight BDQ quadrants identified in our inductive analysis are shown on the right and the two organisation benefits depicted on the left (see figure 26). Taken together the strategic relationships are the focus of this part of the framework. Governance components that influence the mitigation of BDQ problems was touched upon in earlier section 4.2.2, so it is not the goal to discuss these again here. With respect to our big data governance framework, BD quality is positioned as a primary component with BDQ decision domains closely interconnected with all other components as a mechanism to insure BDQ aspects.

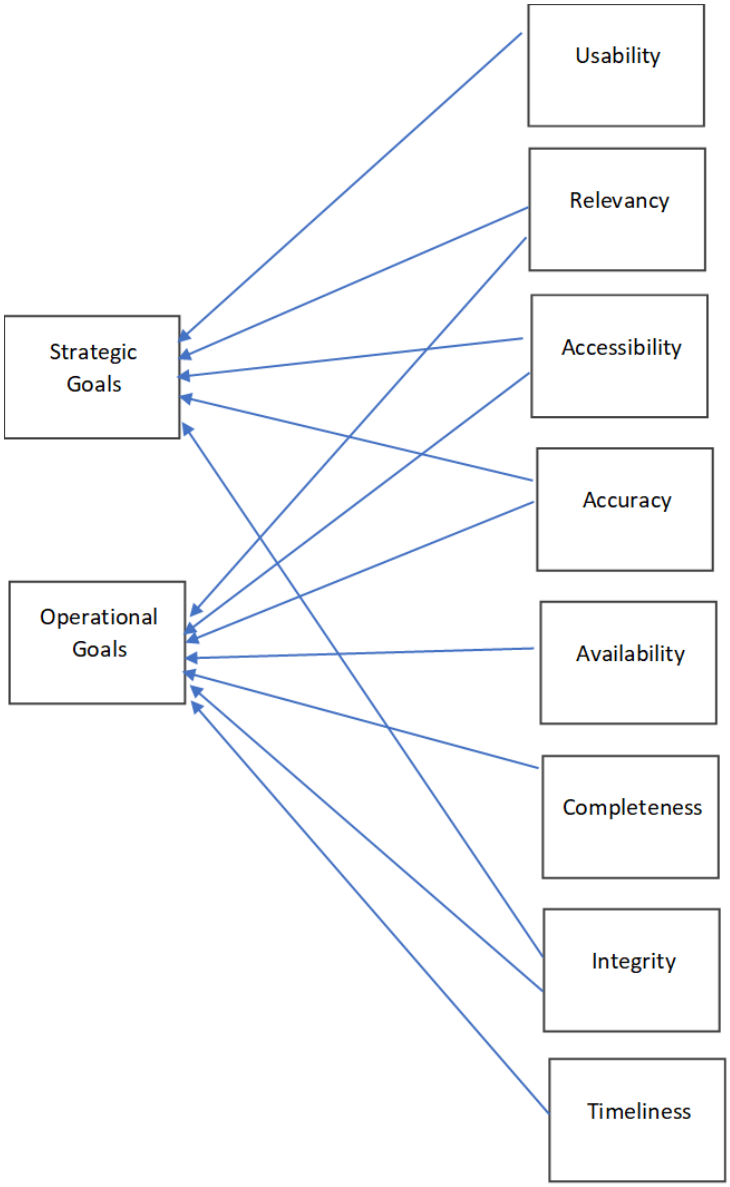

Figure 26: Strategic and Operational Goals for BDQ

Most big data governance roles and responsibilities identified as part of the framework should be pivotal for safeguarding BDQ through the BD lifecycle. Focus must be on master data management and that BD is fit for purpose, ensuring business rules and locus of accountabilities to support this.

- BDQ management should be built into the big data governance programme deliverables.

- BDQ standards and targets should be agreed upfront in conjunction with key big data governance stakeholders notably Chief Data Officer and big data governance Council members.
    - All big data governance council members should be identified
    - BD stewards & BD quality specialist (business & IT) should be salaried staff or funded within approved business cases.
    - Define & enforce governance through standards, policies and controls to govern BDQ (management and use)
- Identify areas of BDQ risk within the organisation.
- Define and execute the BDQ management approach, including how data cleansing could be addressed. As part of the overall approach the following should be performed at a minimum, but not limited to: -
    - Evaluate the key BD items required, assessing the completeness and ability to access and use (in context of business objectives)
    - Conducted BD sources prioritisation
    - Recognise the need to delete/archive data. Retention schedule should be created for all BD stored.
    - Examine how BD platforms would integrate to other data hubs or multiples sources etc.

The big data governance programme should target to continually optimise the BDQ capability, as various BD is collected to BD platforms. BDQ measurements and metrics should be defined to enable successful monitoring and reporting, ensuring appropriate remediation is taken and funding is in place to execute remediation**.**

- **<u>BD Technical Governance</u>**

As it was uncovered that various issues related to BD also occur due to deficient BD technologies and technical processes to analyse or process it efficiently; in our technology component the BD standards, BD architecture, BD policies, BD procedures corresponding to the collection, storage, retention, security, privacy, analysis, processing, data protection, legislative compliance etc need to be identified.

Although big data governance is not just about technology; technology could aid in the automation and scaling of enforcement of big data governance processes, policies, and standards. Particularly standardised integration related processes can be automated. Similarly, BD technologies could supply built-in capabilities to access, security, monitoring or cleansing. Given the context of the complicated BD ecosystem, it imperative to establish and maintain data lineage and the provenance of BD. Participants indicated they were struggling to work round the BD technical skills gap. Reporting problems recruiting resources with the right skills or building internal capabilities for handling BD as most of the technologies and methods require high levels of expertise. As part of the technical component, organisations need to focus on strategies to recruit and retain the right competencies.

Organisations are advised to invest in tools that help make certain BD is accurate, available, timely and fit for purpose at all times. Legacy systems and tight budgets are a problem here for some organisations; however, they should decipher how to integrate what already exists where possible. Findings indicate that focusing strongly on architectural standards could also support to facilitate this. Alternatively postpone certain BD initiatives until there is technical readiness. As part of the framework guidelines the following is also expected, but not limited to:

- Classify all BD within the organisation (as per approved standards and taxonomies)
- Assess BD security risk and define controls
- Design and communicate security policies
- Set up physical & logical access security controls to block unauthorised access to private, sensitive or valuable BD.
- Define and execute processes to keep up to date on legislative and data protection regulation, whilst also employing BDLC best practice safeguards.

In the BD landscape the appropriate big data governance safeguards must also be in place to enable organisations work with trusted third parties i.e. contractual agreements, non-disclosure agreements etc**.** Lastly, we have taken on board participant views to improve BD funding models related to BD technologies, technical training etc. We discuss this related component later.

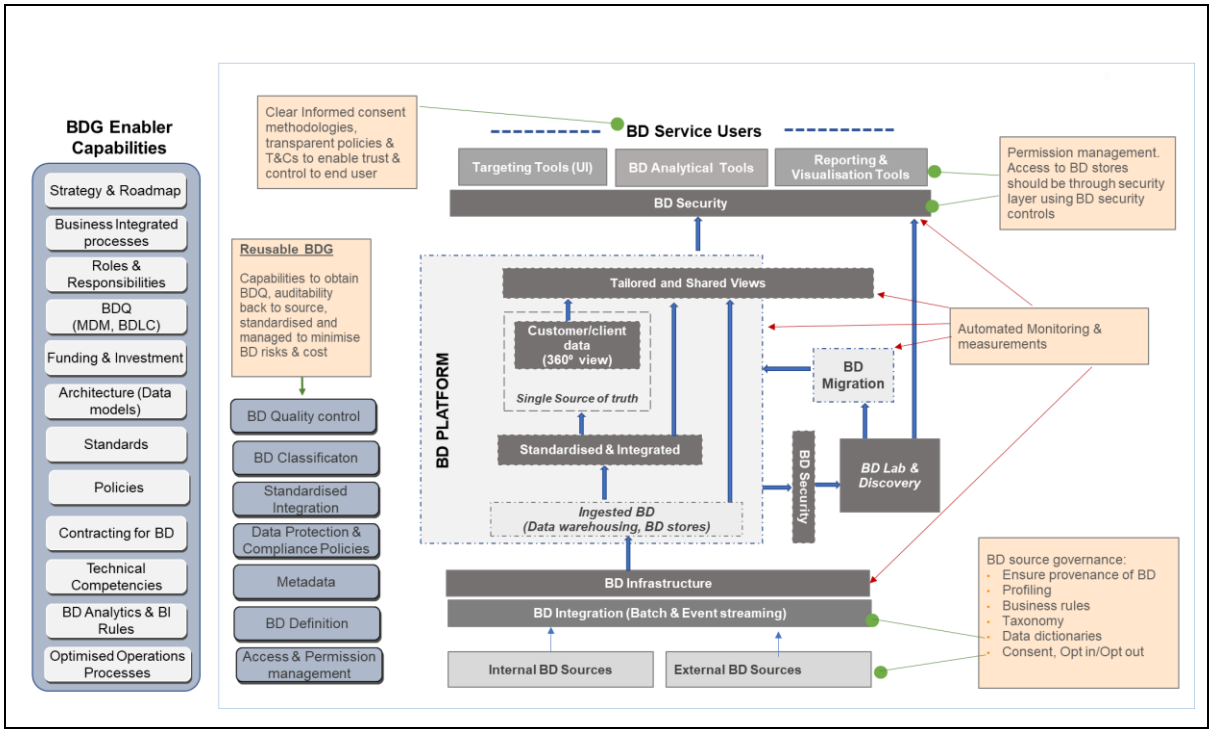

Figure 27: Enabler Capabilities & Proposed Big Data Governance Technology Blueprint – BD and functional flows

- **Training and Skills Development plans**

Closely linked to the strategy component, is the part of our framework which reflects the creation and execution of training and skills development plans as crucial big data governance activities. These are vital to maintain, optimise and drive adoption of the BD competencies that fosters the success of BD initiatives. Moreover, they are identified as important activities to ensure organisations meet the big data governance goals for the education, awareness and competency development of big data governance stakeholders and BD users.

- **Funding model**

As part of the proposed big data governance an appropriate funding model should be made available to support the success of BD initiatives. One of the first steps in this part of the framework is to assess the value proposition of the BD use cases identified during scoping at the start of any initiative, based on the cost-benefit. Inductive analysis results indicate that

BD initiatives could increase the cost associated with 1) sophisticated BD technology & infrastructure, 2) BD security, 3) recruiting and retaining the highly skilled resources required etc. Recognising these areas involve decisions affecting both an organisation's capital expenditure and revenue, we were able to establish a direct link for the criteria to evaluate the return on investment. Incorporation of this separate component that addresses the goal of assessing cost related risks, ROI and ensuring appropriate funding is allocated. It guarantees the organisation is focused on governing and controlling costs associated with BD delivery to ensure they are manageable e.g. Associated technology and infrastructure costs, ongoing base fees for BD applications, upfront and ongoing integration costs, 3rd party licence fees are reasonably priced etc. As mentioned before, the big data governance gives the directive to set-up a cross-functional organisation structure with the right locus of accountabilities. It is the proposed that there is representation from Finance and Procurement divisions of the organisation on big data governance committees actively consulting for this area, even though executive leaders should be accountable.

- **Monitoring, Measurements and KPIs**

It is paramount to prioritise the monitoring and measuring of the big data governance programme, at the end of an elected timeframe or at intermediate milestones. Results should be audited and measured against success metrics and key performance indicators defined by the organisation (Thomas, 2006; Al-Badi et al, 2018; Soares, 2012). Hence, this component creates the metrics and KPIs to measure big data governance effectiveness, thus establishing the benchmarks that can be used for monitoring and identifying areas for optimisation during the monitoring & audit stages of the framework. It is hard to foresee every risk, or every exact action required from the onset of any big data governance programme. Organisations must actively monitor and audit their BD operating environments and assess performance of the big data governance programmes against these defined metrics in orders to assess where changes or adjustments are needed. Additionally, progression being made can be reviewed against key business requirements & objectives, cost-benefit, technical aspects, operational efficiency, and the effect of the big data governance programme on BD value creation.

High priority should be given to monitoring and measurement of BDQ, Compliance with legislative, adoption of ethical guidelines etc. It is worth noting that all processes, guidelines, standards, policies etc should have been specified and the controls for big data governance

implemented, before monitoring and auditing is undertaken. Monitoring and measuring these results means that outcomes or benefits can be disseminated to relevant stakeholder's organisation-wide, thereby fostering ongoing support, finance etc to the big data governance programme. Along with the change management component of the framework, this component emphasises the framework is vested in supporting a BD ecosystem that puts tools and processes in place for the organisation to adjust to changing BD environments, as well as make sufficient course corrections required to address BD challenges.

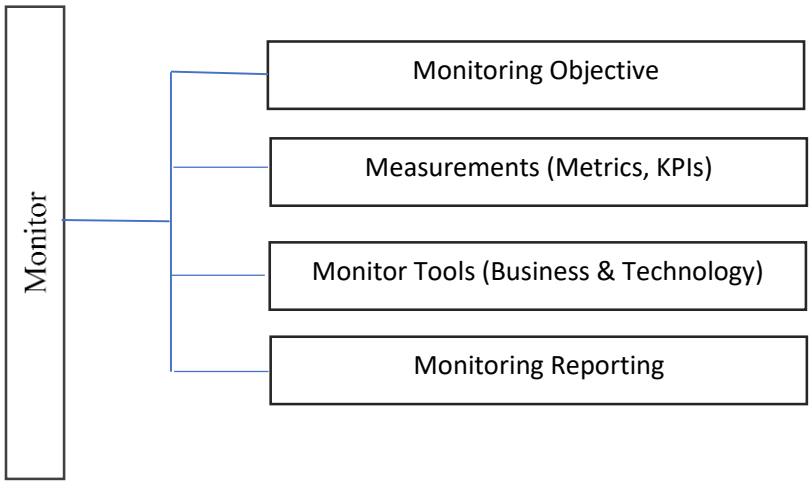

Figure 28: Monitor frame of reference for Big Data Governance

- **Legislative and Regulatory Compliance**

As explained, organisations should be able to apply the correct rules to every BD use case, which they are able to analyse upfront in our big data governance framework at the right points of BD system context. Where there is clarity required in the application of any legal rules or obligations, the organisation structure should ensure that the right roles have been assigned to clear up areas of ambiguity or clarify general language of the law. The purpose here is compliance and we are looking for a systematic and static simulating component for our framework, which is also based on local and international compliance demands. Tools covering regulatory rules and guidelines is selected here to aid comprehension of regulations

related to BD and to aid organisation compliance. Additionally, definitions of the law and applicable legislative compliance resources pertinent to the relevant BD laws should be developed, including guidelines, policies, contractual agreements, terms and conditions etc. Legal ownership & IP rights in relation to BD is also prioritised in our framework e.g. in relation to copyright, confidentiality, patent rights to BD innovations and software applications.

Khatri et al (2010) also corroborate that there are strong connections between DG and compliance, notwithstanding these aspects were not really focused on within their framework. By contrast, we address compliance as part of the big data governance process using an ontology-based approach for representing the compliance processes and policies, as well as checking the organisations ongoing compliance with them. This is based on core parts: one to identify the BD compliance risks and the other to capture the compliance requirements. Guidelines from government, independent authorities, international or professional bodies are also to be collated and made available to big data governance stakeholders. Processes should also exist to enforce policies at the system level. The big data governance framework also recognises the necessity for global operating organisations to comply with cross-jurisdiction legislation and regulation. This is represented in the design of our big data governance framework. It is clear the European Data Protection directives (GDPR) is available as a reference for regulation within Europe. Notwithstanding, global operating organisations are faced with numerous international data protection regulations, competition laws, local legislation etc, that the big data governance framework highlights should be addressed. For this reason, our framework includes Compliance Officers, Data protection Officers and Legal Counsels in the locus of accountability to work cross-functionally with other big data governance stakeholders to define and clarify legal and regulatory compliance rules, as well as monitor and enforce compliance. Crucially all BD users and big data governance stakeholders should be provided with ongoing training to educate them on the organisation's and their local or international regulatory responsibilities. Based on the organisation's operating environment requirements; training plans that provide awareness of the relevant data protection & legislative regulations should be designed. Case law interpretations must also be created and shared with stakeholders.

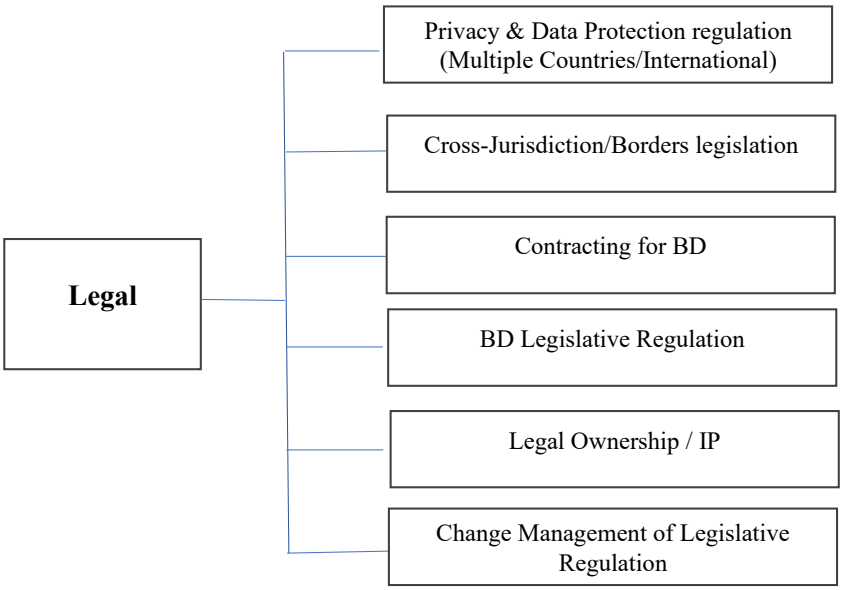

Figure 29: Legal frame of reference for Big Data Governance

- **Privacy Component**

Closely interconnected to the legislative and regulatory compliance components discussed is the privacy component. A few of the earlier DG frameworks concentrate on recommending guidelines for data protection to protect PII data and advocate compliance with regulation to protect privacy (IBM, 2007; Thomas, 2006; Khatri et al, 2010). Some speak to regulation compliance, pointing to policies drawn out from legal content (Panian, 2010). As part of data protection regulation such as GDPR, privacy policies are cultivated by technical controls and guidelines to protect data subjects. Most data protection regulations already concentrate on privacy risk assessment through the technical system life cycle as a solution to document privacy requirements and provide some technical approaches such as identity management, consent & preference management, anonymisation etc. as means to safeguard individual's privacy. This was also evident from the analysis undertaken as part of the research. Applying these existing safeguards still must be undertaken, but by contrast the proposed big data governance framework concentrates on the implementation of principles related to safeguarding the privacy of BD platforms and the privacy of individuals, taking into account the spectrum of privacy required for BD and the apparent requirement for improved consent

methodologies. These privacy principles govern the identification, design, development and implementation of privacy controls and policies that specifically address the BD complexities uncovered during research analysis, as well as guide the relevant decision domains. During the BD risk identification and mitigation assessment process, organisations should focus on identifying BD privacy control and policies that can be implemented to fulfil high priority privacy protection requirements related to BD risk. High on the agenda here for leadership teams should be to consider whether (or not) an individual can be considered identifiable during BD processing. Appropriate methodology for informed consent should be designed and implemented by organisations to ensure the public have full awareness of the BD services they are opted in to and that any changes in the use of their BD is transparent to them. The gaps in the current consent methodologies that were previously used for traditional data was a major finding of the research, hence the framework stipulates organisations create and execute informed consent methodologies as part of the privacy component.

It is observed that data controllers within organisations operating in Europe have an obligation under the GDPR data protection principles, to enact suitable organisational and technical processes to protect against unauthorised and illegal processing that impact an individual's privacy. Hence this component links closely with the legislative and regulatory compliance component of our framework. The framework advocates defining big data governance metrics / success measures for monitoring and auditability of privacy processes and policy adherence i.e. for both technical and business. If under any circumstances there is a breach owing to malicious or accidental activities, clear processes and lines of communication should be put in place to report this and defined remediation taken.

A small organisation might not have the luxury of many personnel to choose from, hence one role might take on multiple responsibilities as long as the competence and bandwidth enables this.

5.2.3 Focus on BD analytical and Mining governance practices

Finally, although not defined as a separate component we point out that the big data governance framework considers one of the primary triggers for big data governance within

many organisations; that is to leverage BD for analysis and draw insights from BD held in disparate BD systems, as well as to leverage new BD technological advancements for analytics. Hence to effectively deploy BD initiatives our framework proposes that components should facilitate the effective governance to achieve these aims and minimise the associated risk. There could be various ways within big data governance to help achieve these goals and we do not proport to have come up with exhaustive big data governance guidelines or principles for the components identified to facilitate this. Nonetheless based on our previous findings. The target is facilitating the rationalisation of BD architecture and BD quality improvements to support the outcomes of BD initiatives. Moreover, ensuring that governance decision-making targets the improvement of BD availability and accessibility, as well as promote consistency with multiple BD tools and practices. Thus, the BD architecture and BDQ control component facilitates the governance to drive effectiveness of BD analytical goals.

Similarly, some core big data governance guiding principles for BD and BD analytics platforms that might assist organisations response to the constant rate of business change in the arena is a necessity given previous findings. Given the socio-ethical issues related to BD data mining and machine learning practices that were uncovered in chapter 4, it is advised that data mining and profiling guidelines are defined. The ethical big data governance component is discussed in the next section.

5.3 Ethical Big Data Governance by Design

This section discusses the main features of the ethical component proposed for integration into the design of our big data governance framework. It introduces a new ethics-by-design approach for addressing ethical challenges organisations could confront. Figure 30 presents the ethical component dimensions designed following the inductive analysis findings, indicating how it would be arranged within the BDG framework, along with incorporating ethical BD principles to future-proof and minimise ethical risks.

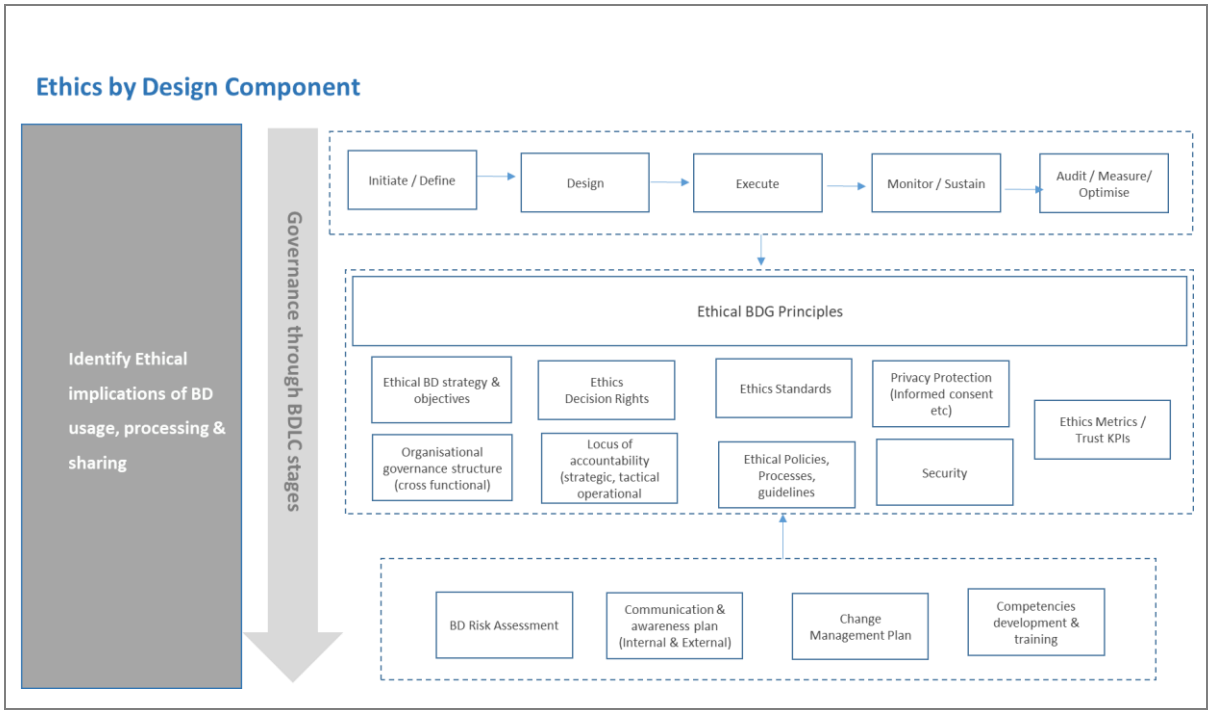

Figure 30: Novel 'Ethics by Design' component of the Big Data Governance framework

Solving an issue by progressively breaking it down into smaller and more manageable pieces is a generic design method. Weber et al (2009) contend that "one size, does not fit all" when considering governance approaches (Weber et al, 2009). This suggests that when considering the design of a framework, flexibility should be built in so that components can be selected dependent on the business use case or maturity level of the organisation. A big data governance strategy that discourages flexibility could hinder adoption, depending on the 1) organisational culture, 2) personality of the implementors, and 3) resources available. An innovative and successful strategic framework is one that is developed to focus on solving key problems, but also ensures there is flexibility and accessibility for its users.

From the research findings there is evidence that there is a high level of complexity in the multidimensional ethical aspects that must be considered by big data governance. Consequentially, in our big data governance framework we purport that governance should be structured in a method that integrates an ethics component which primary objective is to enforce high ethical standards. As many of the participants highlighted that 'one size does not fit all', it should be flexible enough to be adopted specific to the organisation's needs and based on the BD use case to be addressed. The process of ethical governance supported in

the big data governance framework is a 'ethics by design' ontology consisting of decision points, locus of accountability and principles. We introduce the phrase 'ethics by design', which we term to mean the scoping, design and integration of ethical governance principles into big data governance structures, processes and guidelines. This defines a set of ethical ground rules which sit above and beyond other regulations e.g. data protection regulations etc, to direct the organisation when planning BD activities and support them in considering ethical practices related to BD. Plus, mitigate against unintended consequences. This ethical governance component should be relevant at all levels of the organisation and assist to provide clear and consistent ethical tolerances to all stakeholders cross functionally and cross jurisdiction.

It is decided that a reasonable starting point for the proposed ethical component of the big data governance framework was for oganisation's to first establish where in the BDLC the ethical concern or risks occurs. This action should be performed for a number of reasons. As highlighted previously Khatri's et al's earlier DG framework highlights 5 decision domains, one of which is the data life cycle (Khatri et al, 2010). A view of ethical problems could be improved by considering these problems initially against the BDLC. Secondly, organisations decisions made at all stages of the BDLC or BD analytic process will have consequences for both the BD users or BD subjects (i.e. customer, client etc) through the life cycle. An organisation needs to consider where in the life cycle any decisions to address BD issues are likely to occur, as the decisions points offer a means of integrating risk management. Finally, it also serves as an aid to ensuring there is some thought to ethics by design in our governance framework. Thus, structuring BD initiative planning actions to take into consideration decisions points that apply to the BDLC, with associated consequences for organisational decisions and activities related to big data governance. It is worth flagging here that we show that these are not just technically relevant, but also have social importance. Moreover, the approach adopted is suitably flexible enough to be adjusted based on transparent stakeholder feedback and future risk factors that might emerge in the rapidly changing BD landscape.

Each initial stage consists of an information gathering activity e.g., examining potential risks through each BDLC stage (see appendix C), followed by a decision targeted at mitigation. As

it is seen in table in the appendix C, we were able to decompose ethical implications at each BDLC stage, of which further description for each BD ethical challenge could be included by the organisation. This makes sure that consideration is given to the inception, tracking and movement of BD through the ecosystem. Although the numbers of ethical problems are huge, nevertheless it is less complicated, more understandable to review each problem gathered and subsequently prioritise.

Information gathering should involve relevant assessments or checks, preparation of paperwork based on various assessment activities and resources to enable decisions for each decision point (with the scope to incorporate a wide range of challenges and potential risks). One of the vital questions to ask at this stage is, can we improve the way which BD is ethically collected and analysed so that people can trust the organisations BD initiative and then identify better actions. According to Khatri et al (2010), governance approaches specify what decisions should be made to guarantee effective management and use of the decision domain (BD) and who has the locus of accountability for decision making. The specific focus on decision points provides a means for incorporating ethical actions into the BD innovation model. Each decision has a risk factor and requires information that provides forms of risk management against unethical practices. Therefore, the decisions offer a means of integrating ethical governance focused risk management concepts into BD initiative planning or implementation model by design. In other words, this governance component could be viewed as a risk management mechanism; filtering out potential BD ethical risks that are likely to occur at different BDLC stages and offering a process for the organisation to reflect on the perceived ethical risk. In terms of the decision points, instead of taking 'Yes' or 'No' decisions, this component identifies the ethical risks, and the BD initiative can proceed, but the organisation may need to investigate the ethical issues identified thoroughly given it be a major source of risk - the organisation can decide whether to proceed with the initiative as planned, consider mitigation actions or abandon the BD initiative completely. By considering the ethical implications of BD initiatives at each BD life cycle stage within the proposed framework, with feedback loops and decision points allowing for the consideration of different issues, it can be claimed that the process is interactive. The locus of accountability should be cross-functional, and this stage could involve stakeholders from various different departments.

Participant ideas already summarised in chapter 4 provide a list of some of the recommendations revealed during investigation, we have singled out the significant considerations given to setting out suggested ethical principles for BD sharing and governance of BD initiatives. (BIG0015) examines the ethical issues heightened by BD and sets out suggested ethical principles for governance of BD. Although this participant focuses particularly within the context of Healthcare and Biomedical research, we examined if some of the suggestions could be considered useful for wider general usage in other fields. Their stance infers that a 'one size fits all' solution to ethical governance of BD initiatives is not possible, however it was induced that when an organisation starts to develop an initiative, a set of principles can be followed for improving ethical BD practices. For instance

- Any BD initiatives should take both private and public interest into account.

- Any BD initiative must respect the fundamental rights of the individual e.g. the right not to have their privacy infringed upon etc. Including, limits to which the government or others have specific power to interfere with citizens privacy in the public interest.

- An organisation's decision makers should actively seek participation, as well as take active steps to identify how individuals expect their BD to be used and ensure they act on those expectations.

- Procedures and policies for formal accountability should be included for BD initiatives to ensure the organisation adopts ethical BD practices.

Secondly, (BIG0080) also flag the work on ethical data sharing principles led by UK cabinet Office and Information Economy Council. Findings indicate that these could be a foundation for developing trust if the ethical principles are integrated to some degree as a premise for BD sharing. We cross referenced these principles with some of the ethical issues discussed by participants with a view to establishing if there is any potential for reuse.

Although the suggestions discussed by both participants here are in the context of Healthcare and Biomedical research (BIG0015) and principles applied in data sharing (BIG0080), the same techniques could also be applied in the ethical component of the proposed big data governance framework since it is thought the nature of ethical analysis is common between

them and our ethical argument. Unlike these participants, however, we apply to big data governance that help support a higher level of ethical standards for BD initiatives. Embedding ethical principles at the centre of big data governance for collection, processing, sharing and publishing of BD could assist in developing a foundation of customer or public trust for BD initiatives according to findings. For example, incorporating the ethical principles to ensure public transparency and control on which participants have reached consensus as being vital, fosters trust in BD systems, processes and people. The inquiry results indicated that the notion of trust, fairness and transparency are increasingly crucial for consumers engaging with organisations using their BD. It is important to ensure ethical approaches to address concerns uncovered within the big data governance framework to ensure value is obtained and minimisation of harms are considered.

Khatri et al (2010) argues that data principles determine the connection with the business e.g. an organisations decision to standardise ethical BD business processes signifies that there must be a defined business owner of BD assets (data principle). By outlining the business uses of BD, a data principle confirms the magnitude of how BD is seen as an organisation wide asset, and therefore what specific ethical guidelines, policies or processes are appropriate to mitigate ethical concerns. Khatri et al (2010) go further to suggest that each principle is supported by a rationale and a set of implications, making sure to consider external data usage. Moreover, data principles define the desirable behaviours for cross functional personnel within the organisation i.e. across both business and IT functions. With implications that business owners have a vital role in data quality management, as well as its governance through the life cycle, interpretability and access.

We endeavoured to scrutinise principles from the IEC Ethical Data sharing principles presented by the (BIG0015) and principles proposed by (BIG0080) and others to simulate ethical principles for the big data governance framework's ethical component. To accomplish this, we utilised the same inductive reasoning method to set apart key elements into our results. These elements were prepared in sequence and were examined for more meaning and refinement of ethical recommendations extracted from the dataset. We also revealed in the previous sections how some of the uncovered ethical BD implications impact decisions through the BDLC. Here we take the concepts and obligations and map together with some

further concepts we included in the novel ethics component of the big data governance framework. Therefore, requirements and observations from our overall research findings, were mapped and integrated to form the ethical component.

To perform the stage of ethics component creation, we needed to extract key ethical questions to ask from our thesis findings. An important step to identify and distinguish these pertinent questions was to refer to ethical concerns and implications revealed by participants. Each ethical problem extracted in the earlier analysis discussed in chapter 4, was a pre-step to examining the right ethical principal and standards areas that need to be responded to. Outcomes inform the the 'ethical by design' steps of the ethical big data governance component, and these are summarised below:

1) Leadership should define a strategy for ethical big data governance which links with the organisation's business objectives

2) Establish specific roles and locus of accountability within the organisation with the responsibilities for ensuring the defining, management, transparency, and enforcement of ethical big data governance controls and standards (both internal and external). For example, establish Ethical Review Boards, Ethic Councils/Committees and stewardship roles, community awareness boards etc. Some of these roles should drive key ethical big data governance domain decisions as follows:

Ethical Big Data Governance Domain Decisions
- What are the ethical principles & standards in respect to the organisations BD practices? - What are the organisations ethical usage and sharing boundaries for BD? - What are the preferable ethical behaviours for employing BD as enterprise assets?

- How do gaps in the regulatory environment influence the voluntary uses of ethical big data governance that the organisation will adopt? - How will ethical risk assessment be undertaken on an ongoing basis? - What is the big data governance approach for documenting the ethical processes & policies required to mitigate ethical malpractice? - What are the methods of raising awareness of the organisation's ethical uses and policies of BD i.e., both internally & externally? - How does the organisation engage the public? - How will ethical big data governance be monitored and measured e.g., Trust KPIs etc?

Table 20: Ethical Big Data governance domain decisions

3) Establish processes and forums to report and handle cases of ethical malpractice that contravenes the organisation's policies and standards.

4) Create and execute training and awareness plans for ethical big data governance guidelines, processes, policies, and standards

   The organisation should ensure all guidelines, procedures, policies and decision points are understood and relevant to all levels and/or departments of the organisation. It is particularly important that ongoing training and funding is considered and put in place to ensure ethical guidelines, procedures, policies, decision points and code of practice is fully understood by stakeholders at all levels.

5) Create communication plans for ongoing internal and external stakeholder awareness.

6) Open channels for public participation
   a. Inform and engage the public

b. Establish external community boards / committees with the responsibility to drive ethical practices of BD
c. Involve the public in the design and governance of BD initiatives, where there benefit, and values can be expressed and considered
d. Cultivate customer and client trust.

7) Create measurements and KPIs to monitor effectiveness of ethical big data governance. Then execute ongoing monitoring, auditing and optimisation based on measurement results.
   - Establish BD trust KPIs which consistently monitor impact of ethical safeguards put in place and monitor levels of trust in the organisation e.g. number of right to erasure requests. Have they been actioned? How long did they taken to be actioned?

Table 21 summarises some ethical big data governance principles which also builds on the Human Right legislation work as far back as the 1950s discussed in the written evidence.

Ethical BDG Principles	
Access & Ownership:	BD shared by the public is owned by them, and the public can replicate or withdraw permission to it at at will
Control:	Informed consent must be given for all BD types usage and sharing
Retention:	Collection of BD without an intended purpose should be minimised. Solely the BD required for the particular processing operation should be utilised
Transparency:	The customer or client must have knowledge regarding the utilisation of their BD
Correctness:	The BD source must be accurate and easily updated
Fairness	BD should not be processed or shared in a way that would bring harm to individuals or groups
Simplicity:	Services must be designed to be understandable to the members of the public e.g. clear optin to provide informed consent
Compliance:	The public desire to be assured that BD services meet the principles above

Table 21: Subset of potential ethical big data governance principles

Organisations can add or adjust ethical principles according to the decisions made about them and the strategy to apply. As has been shown, any of the compositing elements of the component are providing an obligation which their non-existence could cause unethical BD practices. For example, if the terms and conditions to utilise a BD service is presented to the data subject in an intelligible form or has not used plain and clear language, the BD controller is not complying to its ethical obligation to the data subject. Therefore, we need ethical BD principles to understand the meaning of clear and plain language in detail and guidelines that serve as handrails for staff, in order to apply it in the form to the data subject which is ethical.

5.4 Evaluation

The purpose of this section is to evaluate the big data governance framework in order to examine whether it fulfils the goals of this thesis research. A subset of earlier identified DG frameworks discussed in chapter 2 are evaluated and compared against our novel big data governance framework to assess their appropriateness for addressing BD challenges and risk factors. We focus the evaluation solely on those frameworks identified from the academic literature and considered to be big data governance frameworks as indicated in the systematic literature review matrix (see appendix B). To compare them, verification was conducted by all the evidence collected from the literature review and inductive analysis findings of the parliamentary inquiry dataset. Silverman considers that the ability of the researcher to relate their findings to an existing body of knowledge is a key criterion for evaluating works of qualitative inquiry (Silverman, 2000). In addition, criteria for evaluation are also extracted from the ISO 8000 standard to assess how well they fill the gaps that form the motivation for this research. A previous author had used a similar approach for evaluation of their earlier big data governance framework (Al-Badi, 2018). It is deemed suitable to utilise a similar approach for evaluation given the similar subject matter and given that ISO 8000 also specifies standards for maintaining quality controls. Crucially, because BD needs to be governed with definitive reference to the BD operational or BD analytic requirements for application (Soares, 2012), the evaluation is also again performed against three scenarios. The sample scenarios are adapted from Soares (2012) publication given a case study methodology has not been used in this thesis study.

Scenario 1:

Hospital staff of a healthcare provider could use biometrics, so they do not have to utilise user IDs and passwords. Instead, they can utilise biometric scanning to log into the system quickly as they arrive at a workstation. Hospitals must work with legal personnel, notwithstanding, to do a privacy impact assessment regarding the usage and retention of biometric data about their staff.

Scenario 2:

A telecommunications provider builds detailed models on the percentage of customers that stop using their company's product or service during certain time frames. The models include big transaction data e.g. call records and social media data. The overall practicality of the model is dependent on traditional attributes of the master data related to the customer. For instance, name, date of birth, income and location. This large company also wants to execute a predictive BD analytics strategy connected to predicting this churn. They also decide they want to outsource these analytics activities to a third-party vendor overseas, which means that customer records must be sent to the vendor daily. There is significant concern connected to safeguarding the security and privacy of the large quantities of customer data.

Scenario 3:

An information technology department of a large organisation is using BD for analysis of application logs to gain insights that could lead to system performance improvements. The application vendor log files are in various formats; hence the must be standardised before they can be of use. It has come to light that their marketing business unit also want to use twitter feeds for sentiments analysis to investigate customers views about their products. Integrating sentiment analysis with customer profiles could be challenging and might result in privacy issues.

Table 22 below presents the results of the evaluation. The result of the comparison is shown by the concepts included or not included in each framework under evaluation; where 'included' = √, 'not included' = X, 'partial inclusion' = P.

The results revealed from this evaluation procedure, depicted that none of the existing frameworks are suitable to address big data governance comprehensively and that the proposed novel big data governance framework contributes in various ways to achieve this goal. The evaluation was against the type and number of governance concepts utilised in each framework.

Based on the assessment the big data governance framework contained the most concepts compared to the other framework when looking at three separate cases. It demonstrates that endeavours to address end to end aspects of big data governance in the design was successful. Some of the other framework's strength are apparent in specific singular dimensions where they go into detail being prescriptive e.g. Weber (2007) gives a description of responsibility assignment matrix comparable to a RACI chart. However, we can address this in future work.

The outcomes of our evaluation process illustrates that each BD complexity is considered by a mix of big data governance countermeasures in our designed approach. We have also considered a process for ongoing monitoring, sustainability and optimisation of the big data governance programme.

Frameworks	Key Dimensions																														
	BD Governance	Methodology to derive provided	BD Risk Assessment	Strategy & data centric culture	Organisation Structure & Decision-making	Alignment BD initiatives with Bus obj.	Scoping / BD use case identification	Value Proposition / ROI	Decision rights	Clear Roles & Responsibilities,	Stakeholder Identification	Process Identification	Principles	Policies & Guidelines	Standards	Quality Management/Control	Implementation	Monitoring & Measuring	Technical	Standardised integration (interoperability)	Architecture	Metadata	Privacy protection	Ethical	Financial Model \ Cost control	Legislative & Regulation	Compliance	Security & Access	E2E Life Cycle	Communication Plan	Change Management	Training & Awareness	BD Skills & competency dev (Bus & IT)
*BDG Frame work	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	P	√	√	P	√	P	√	√	P	√	√	√	√	√
Kim & Cho (2018)	√	x	P	√	P	x	√	x	x	P	P	√	x	P	√	P	P	P	P	x	x	x	√	x	x	x	P	P	x	x	x	x	x
Al-Badi et al (2018)	√	√	x	x	√	x	√	x	x	x	√	P	x	P	P	x	x	P	P	x	x	x	P	x	x	P	P	P	x	√	x	x	x
Zhang et al (2017)	√	√	x	P	P	x	x	x	P	P	√	P	x	P	√	P	x	x	x	x	√	P	P	x	x	x	x	P	P	x	x	x	x
Li et al (2019)	√	√	x	P	P	P	√	x	x	P	P	P	x	√	P	P	x	x	P	x	P	x	P	x	x	P	P	P	P	x	x	x	P

Tallon (2013)	√	x	P	x	X	x	x	P	x	P	P	P	x	P	x	x	x	P	x	x	x	x	x	x	P	x	x	P	x	x	x	P	x

Table 22: Frameworks comparison against dimensions of Big Data Governance

5.5 Summary of Chapter 5

This chapter has presented the design of the novel big data governance framework for addressing the additional BD complexities that organisations are confronted with. The proposed approach in this research is an attempt to answer to the demand for big data governance. The chapter outlines the proposed big data governance component to address the ethical implications of BD which is a major contribution of the thesis study. Moreover, the big data governance framework will be a useful reference framework for organisations to guide them on their big data governance journey. Lastly, evaluation results which illustrated how the proposed big data governance framework compares to some other earlier big data governance frameworks was also presented in this important chapter of the thesis.

6.0 Conclusion, Limitations and Future works

This chapter concludes the thesis by summarising how the study has fulfilled the research goals answering the research questions, describes the limitations and proposes some areas for future research.

6.1 Summary of the Thesis Study

BD is increasingly regarded as a valuable asset by organisations globally. Notwithstanding plausible evidence of the limited attention to big data governance leading to significant problems for organisations executing BD initiatives; there is a lack of focus to this point on big data governance frameworks that specifically address the BD challenges and risk factors holistically. There are few studies on big data governance and those that exist either do not address end to end BD complexity aspects, or do not discuss the methodology used to design them. This thesis study filled this gap, by designing a novel big data governance framework, utilising a rigorous grounded theory methodology. Based on a systematic review and the inductive analysis of parliamentary inquiry submissions, major BD challenges impacting both private and public organisations, as well as the core criteria for big data governance were identified. A framework with four strategic goals, 22 core components and four enablers has been proposed, which will serve as a toolkit for organisations to address BD challenges that they may be confronted with when implementing BD initiatives. A novel 'ethics by design' component has also been created with the purpose of handling the ethical tensions raised by BD. The framework also makes it possible to perform risk identification and assessment on proposed BD initiatives. This is to ensure that the BD initiatives are always accompanied by a risk assessment methodology which allows BD challenges and risk factors to be identified, so that mitigation can be put in place through effective governance. Additionally, the extensive BD challenges, risk factors and vulnerabilities we unveiled as part of this thesis study will serve as valuable references for both public and private organisations, providing guidance to areas that require focus and prioritisation.

The thesis study answered the following research questions: -

- **Research Question 1**

"What are the core Big Data complexities that organisations need to address?"

The thesis study examined the BD challenges and risk factors that impact organisational outcomes of BD initiatives. By conducting a systematic literature review the initial research gap was identified. The initial review also revealed emerging BD complexities that organisations confront that have become additional challenges as identified by the scientific field. As part of our transparent grounded theory methodology, inductive analysis approaches where utilised to analyses the selected Parliamentary Inquiry dataset. Rather than purely utilising just the literature, the inductive analysis and thorough coding process used to examine the inquiry papers, aid to present real-life necessities as reported by practitioners to operationalise our BDG framework approach. The reiterative inductive approach we used permitted our research to study this complex BDG paradigm, providing a detailed comprehension and insights from real-world context to answer the research question.

Based on feedback and arguments from several participating organisations from both the private and public sector, the core BD challenges and risk factors that need to be addressed through appropriate governance were determined. Therefore, identifying a full set of BD complexities that occur from specific organisational environments and BD ecosystems.

- **Research Question 2**

"What are the existing data governance frameworks to manage Big Data as an asset and how useful are they?"

During the two phased systematic literature review current DG frameworks were identified and investigated. Due to this research being in its infancy, current DG frameworks from the practitioner fields were also critiqued. These DG frameworks are presented in section 2.4 and were initially critically analysed for their usefulness in fulfilling the BDG gaps. Furthermore, evaluation undertaken of the existing frameworks also concluded that none of the earlier works satisfied the full set of BDG criteria. Even though some DG frameworks have strengths in particular areas, all of them did not meet the criteria to fulfil end to end aspects of BDG.

- **Research Question 3**

"What are the elements of a Big Data governance framework?"

An iterative inductive analysis process on inquiry participants dataset established the governance criteria which influence organisations capability to govern BD, as well as mitigate against potential BD challenges (Chapter 4). This formed the foundation for the design of BDG components that formed the building blocks of our proposed BDG framework discussed in chapter 5. In the same chapter, the way in which the distinct stages of the framework are meant to operate was discussed. We also describe BDG framework components of each stage of the framework and justification was given for these.

- **Research Question 4**

"What type of framework can guide Big Data governance?

A novel conceptual BDG framework is synthesised and designed. It is created by extending an earlier framework that had been designed to fill the governance gaps to do with traditional data in the literature. This earlier framework consists of five key decision domains: quality, principles, metadata, access & data lifecycle. We integrate this with output derived from our research analysis findings, forming the main construct of the proposed BDG framework. A significant contribution is that the framework makes it possible to perform a BD risk identification assessment as part of BDG. This secures that governance is always accompanies by a risk assessment procedure that takes BD challenges and risk factors into account, so that the organisation can actively address them. As discussed, the stages and components of our framework are presented in chapter 5, providing a comprehensive structure for BDG. Another core dimension of its structure is a novel "ethics by design" component as part of BDG to address the ethical tensions connected to BD usage and its associated practices. This is significant contribution of this thesis research, as emerging ethical implications for BDG was high on the hierarchy of challenges identified during analysis. And existing frameworks do not address this high-level concern uncovered.

Crucially the research work is significant, as it provides an original contribution to existing knowledge of big data governance, which is in its infancy, to support the management of big data as an asset. The discrete application of each component of the framework can lead to big data governance improvements and can highly influence the control of BD assets within

organisation's operating environments. It will add new knowledge to IS science regarding big data governance. It will also simultaneously contribute knowledge to another intended audience of practitioners e.g. Chief Information Officers, Chief Data Officers, Business Executives, IT Managers etc., that are responsible for realising the benefits of Big Data initiatives, by providing an accessible and holistic big data governance framework to improve big data governance capabilities within their organisations. This will support their business goals of managing BD as an asset and aid to realise the desired business value from BD initiatives.

This next section discusses some limitations of the research study.

6.2 Limitations

This thesis identifies findings that represent overall viewpoints from the Parliamentary inquiry dataset, employing a quantitative method. Although it has been attempted here to provide a comprehensive list of challenges and risk factors from BD collection or usage, this research could not cover the implications to every conceivable BD complexity. In addition, our research was based upon the assumption that the participant responses accurately reflected the knowledge and perceptions of those participants with respect to BD challenges, BD risk factors and related organisational outcomes. As such, the findings of this study may not necessarily reflect the generalised views of a wider selection of organisations or of organisations outside the industries represented in our dataset sample. Notwithstanding, examining data for evidence and gleaning information from experts are good ways to gather knowledge, as we did in this study. Whilst it might be desirable to expand on why associations exist within the dataset and if they might be meaningful, we do not deem this information relevant to meet our goals. The objective here is to comprehend BD challenges in practice and predict potential risk factors within this domain based on evidence submitted.

The timeframe available for this thesis study added some constraints. Thus, although the case study or action research methodologies could have been alternative methods to conduct this research, we excluded them as options as the timeframe for investigation was limited. Obtaining the required organisational approval to conduct a case study proved difficult and

would have caused delays due to external factors. Although the grounded theory technique used in this research has also been sharply criticised as being time consuming, nonetheless it also has the advantage of reducing the risk of external delays outside of our control. Future works can be planned, to execute a multi-case study or action research to confirm the big data governance framework within organisational settings.

Additionally, due to the timeframe available, our study considered a representative sample of DG frameworks identified from scopus during our systematic review of the literature. Nevertheless, it cannot be considered completely exhaustive, as frameworks from other sources could also exist. Albeit it is considered that the examples that have been addressed are key DG frameworks and provide a thorough representative picture of the current state of the art and the gaps that exist in big data governance.

6.3 Future works

Various limitations of this thesis study and a suggestion for future work were discussed above. Nonetheless, some suggestions for future work remain and are discussed here. Researchers are encouraged to seek further validation of this big data governance framework in an organisational setting. Such research can produce valuable validation of the research finding and can provide recommendations to enrich its capabilities. This is a research field that is in its infancy (Otto, 2011b; Wende et al, 2007) that needs to keep on developing. Previous authors have pointed out that it is an onerous task to study big data governance based on actual use cases (Kim et al, 2018) and this can be seen as one of the contributions of this thesis study. It would be useful to see how BD challenges and risk factors to be addressed by big data governance would be shaped by stakeholders from different industry sectors. Future research could investigate this by developing a methodology to determine priorities within specific organisation settings. The thesis has discussed and justified the big data governance components that form the building blocks of the proposed framework but have not been prescriptive in the resources or artifacts suggested. This is an area that is to be considered in future, as it might prove useful for organisations at lower big data governance maturity levels.

Finally, this thesis study has gone some way in discussing a big data governance approach to govern the ethical usage and sharing of BD by organisations. As a recommendation for future work the investigation of mandatory international ethical governance standards for BD would be helpful.

References

ABRAHAM, R., SCHNEIDER, J. & VOM BROCKE, J. 2019. Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management,* 49**,** 424-438.

ADDO-TENKORANG, R. & HELO, P. T. 2016. Big data applications in operations/supply-chain management: A literature review. *Computers and Industrial Engineering,* 101**,** 528-543.

AIKEN, P. 2016. Experience: Succeeding at data management-BigCo attempts to leverage data. *Journal of Data and Information Quality,* 7.

AL-BADI, A., TARHINI, A. & KHAN, A. I. Exploring big data governance frameworks. Procedia Computer Science, 2018. 271-277.

AL-RUITHE, M., BENKHELIFA, E. & HAMEED, K. A Conceptual Framework for Designing Data Governance for Cloud Computing. 2016. 160-167.

AL-RUITHE, M., BENKHELIFA, E. & HAMEED, K. 2018. Data governance taxonomy: Cloud versus non-cloud. *Sustainability (Switzerland),* 10.

AL-RUITHE, M., BENKHELIFA, E. & HAMEED, K. 2019. A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing,* 23**,** 839-859.

ALHASSAN, I., SAMMON, D. & DALY, M. 2018. Data governance activities: a comparison between scientific and practice-oriented literature. *Journal of Enterprise Information Management,* 31**,** 300-316.

AMIRI, P. A. D. & GAVGANI, M. R. 2016. A review on relationship and challenges of cloud computing and big data: Methods of analysis and data transfer. *Asian Journal of Information Technology,* 15**,** 2516-2525.

ANSELL, J. & F., W. 1992. *Risk-Analysis, Assessment and Management,* UK, Wiley.

ARISTA, S., ARUN GNANARAJ, A. & GNANA JAYANTHI, J. 2016. A comprehensive study on big-data analytics-tools, techniques, technologies and applications. *International Journal of Control Theory and Applications,* 9**,** 479-488.

AVERSA, J., DOHERTY, S. & HERNANDEZ, T. 2018. Big Data Analytics: The New Boundaries of Retail Location Decision Making. *Papers in Applied Geography,* 4**,** 390-408.

AVERSA, J., HERNANDEZ, T. & DOHERTY, S. 2021. Incorporating big data within retail organizations: A case study approach. *Journal of Retailing and Consumer Services,* 60.

BAESENS, B. 2014. *Analytics in a Big Data World*, Wiley Publishin.

BIESDORF, S., COURT, D. & WILLMOTT, P. 2013. Big data: What's your plan? *McKinsey Quarterly***,** 40-51.

BIZER, C., BONCZ, P., BRODIE, M. L. & ERLING, O. 2011. The meaningful use of big data: Four perspectives - Four challenges. *SIGMOD Record,* 40**,** 56-60.

BRERETON, P., KITCHENHAM, B. A., BUDGEN, D., TURNER, M. & KHALIL, M. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software,* 80**,** 571-583.

BRINCH, M. 2018. Understanding the value of big data in supply chain management and its business processes: Towards a conceptual framework. *International Journal of Operations and Production Management,* 38**,** 1589-1614.

BRINCH, M., GUNASEKARAN, A. & FOSSO WAMBA, S. 2021. Firm-level capabilities towards big data value creation. *Journal of Business Research,* 131**,** 539-548.

BROUS, P., JANSSEN, M. & KRANS, R. 2020. Data Governance as Success Factor for Data Science.

BROUS, P., JANSSEN, M. & VILMINKO-HEIKKINEN, R. 2016. Coordinating decision-making in data management activities: A systematic review of data governance principles.

C., E. 2001. *Coping with Strategic Risk,* UK, Prentice Hall.

CERVONE, H. F. 2016. Organizational considerations initiating a big data and analytics implementation. *Digital Library Perspectives,* 32**,** 137-141.

CHEN, H., CHIANG, R. H. L. & STOREY, V. C. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems,* 36**,** 1165-1188.

CHEONG, L. K. & CHANG, V. The need for data governance: A case study. ACIS 2007 Proceedings - 18th Australasian Conference on Information Systems, 2007. 999-1008.

COHN, B. L. 2015. Data Governance: A Quality Imperative in the Era of Big Data, Open Data and Beyond *ISJLP,* 10**,** 812-824.

CUI, S. Privacy Revolution in the Big Data Age: From Privacy Protection to Personal Data Governance. 2019.

CUPOLI, P., EARLEY, S., AND HENDERSON, D. 2014. *DAMA-DMBOK2 Framework*, The Data Management Association.

DAVENPORT, T. H. A. H., J.G 2007. *Competing on Analytics: The New Science of Winning,* Boston, M.A, Harvard Business School Press.

EDWARDS, P. & BOWEN, P. 2005. Risk management in project organisations. *Civil Engineers Australia,* 77**,** 33.

ESCOBAR, C. A., MCGOVERN, M. E. & MORALES-MENENDEZ, R. 2021. Quality 4.0: a review of big data challenges in manufacturing. *Journal of Intelligent Manufacturing,* 32**,** 2319-2334.

FAVARETTO, M., DE CLERCQ, E. & ELGER, B. S. 2019. Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data,* 6.

FISHER, T. 2009. *The Data Asset – How Smart Companies Govern their Data for Business Success*, Wiley and SAS business series.

FU, X., WOJAK, A., NEAGU, D., RIDLEY, M. & KIM, T. 2011. Data governance in predictive toxicology: A review. *Journal of Cheminformatics,* 3.

GAO, L. Exploring data governance in big data age. Systems and Computer Technology - Proceedings of the 2014 International Symposium on Systems and Computer Technology, ISSCT 2014, 2015. 179-184.

GARTNER. 2012. *Big data. IT Glossary* [Online]. Gartner. Available: http://www.gartner.com/it-glossary/big-data/ [Accessed 16 October 2018].

GHASEMAGHAEI, M. 2021. Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics. *International Journal of Information Management,* 57.

GORRY, G. A. S. M., M. 1971. *A Framework for Management Information Systems* [Online]. Sloan Management Review. Available: http://cpe.njit.edu/dlnotes/MIS645/Frame_Management_Info.pdf [Accessed 2 February 2018].

GOUGH, D., OLIVER, S. & THOMAS, J. 2012. *An Introduction to Systematic Review,* London, SAGE.

GROUP, I. S. 2007. The IBM Data Governance Council Maturity Model: Building a Roadmap for Effective Data Governance.

GROVER, V., CHIANG, R. H. L., LIANG, T. P. & ZHANG, D. 2018. Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems,* 35**,** 388-423.

GUNASEKARAN, A., PAPADOPOULOS, T., DUBEY, R., WAMBA, S. F., CHILDE, S. J., HAZEN, B. & AKTER, S. 2017. Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research,* 70**,** 308-317.

HO, C. W. L., ALI, J. & CAALS, K. 2020. Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bulletin of the World Health Organization,* 98**,** 263-269.

HOLSAPPLE, C., LEE-POST, A. & PAKATH, R. 2014. A unified foundation for business analytics. *Decision Support Systems,* 64**,** 130-141.

INFORMATICA 2017. Holistic Data Governance: A Framework for Competitive Advantage.

JANSSEN, M., BROUS, P., ESTEVEZ, E., BARBOSA, L. S. & JANOWSKI, T. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly,* 37.

JANSSEN, M., VAN DER VOORT, H. & WAHYUDI, A. 2017. Factors influencing big data decision-making quality. *Journal of Business Research,* 70**,** 338-345.

JIN, X., WAH, B. W., CHENG, X. & WANG, Y. 2015. Significance and Challenges of Big Data Research. *Big Data Research,* 2**,** 59-64.

KACHE, F. & SEURING, S. 2017. Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. *International Journal of Operations and Production Management,* 37**,** 10-36.

KAMIOKA, T., LUO, X. & TAPANAINEN, T. An empirical investigation of data governance: The role of accountabilities. Pacific Asia Conference on Information Systems, PACIS 2016 - Proceedings, 2016.

KAREL, R. 2007. Enabling Data Governance in a Big Data World. Available from: https://blogs.informatica.com/2015/11/11/enabling-data-governance-big-data-world/#fbid=xpH26_YNMgz.

KEMP, R. 2014. Legal aspects of managing Big Data. *Computer Law and Security Review,* 30**,** 482-491.

KHATRI, V. & BROWN, C. V. 2010. Designing data governance. *Communications of the ACM,* 53**,** 148-152.

KHO, N. D. 2016. The state of big data. *EContent,* 39.

KIM, H. Y. & CHO, J. S. 2018. Data governance framework for big data implementation with NPS Case Analysis in Korea. *Journal of Business and Retail Management Research,* 12**,** 36-46.

KITCHIN, R. 2017. Thinking critically about and researching algorithms. *Information Communication and Society,* 20**,** 14-29.

KÖNIG, P. D. 2021. Citizen-centered data governance in the smart city: From ethics to accountability. *Sustainable Cities and Society,* 75.

KUNE, R., KONUGURTHI, P. K., AGARWAL, A., CHILLARIGE, R. R. & BUYYA, R. 2016. The anatomy of big data computing. *Software - Practice and Experience,* 46**,** 79-105.

LI, Q., LAN, L., ZENG, N., YOU, L., YIN, J., ZHOU, X. & MENG, Q. 2019. A Framework for Big Data Governance to Advance RHINs: A Case Study of China. *IEEE Access,* 7**,** 50330-50338.

LOSHIN, D. 2013. Data Governance for Master Data Management and Beyond.

MALIK, P. 2013. Governing Big Data: Principles and Practices. *IBM Journal of Research and Development,* 57**,** 1-13.

MANSFIELD-DEVINE, S. 2017. Data Governance: Going Beyond Compliance. *Computer Fraud and Security,* 2017**,** 12-15.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C., AND BYERS, A.H. 2011. Big data: The next frontier for innovation, competition, and productivity.

MAO, Z., WU, J., QIAO, Y. & YAO, H. 2022. Government data governance framework based on a data middle platform. *Aslib Journal of Information Management,* 74**,** 289-310.

MARELLI, L., TESTA, G. & HOYWEGHEN, I. V. 2021. Big Tech platforms in health research: Re-purposing big data governance in light of the General Data Protection Regulation's research exemption. *Big Data and Society,* 8.

MCAFEE, A. & BRYNJOLFSSON, E. 2012. Big data: the management revolution. *Harvard business review,* 90**,** 60-"66, 68, 128".

MICHELI, M., PONTI, M., CRAGLIA, M. & BERTI SUMAN, A. 2020. Emerging models of data governance in the age of datafication. *Big Data and Society,* 7.

MITHAS, S., LEE, M. R., EARLEY, S., MURUGESAN, S. & DJAVANSHIR, R. 2013. Leveraging big data and business analytics. *IT Professional,* 15**,** 18-20.

MORABITO, V. 2015. *Big data and analytics: Strategic and organizational impacts*.

MOTTA, G., PUCCINELLI, R., REGGIANI, L. & SACCONE, M. 2016. Extracting value from grey literature: Processes and technologies for aggregating and analyzing the hidden "Big Data" treasure of organizations. *Grey Journal,* 12**,** 15-25.

NEWMAN, D. L., D. 2006. Governance Is an Essential Building Block for Enterprise Information Management. Stamford, CT: Gartner Research.

NIELSEN, O. B., PERSSON, J. S. & MADSEN, S. 2019. Why governing data is difficult: Findings from Danish local government.

NIEMI, E. 2013. Working Paper: Designing a Data Governance Framework. *IRIS.* Oslo, Norway.

NOVAK, R. & PAVLICEK, A. 2021. Data experts as the balancing power of big data ethics. *Information (Switzerland),* 12**,** 1-14.

NOVÁK, R. & PAVLÍČEK, A. Big data ethics and specific differences from general data ethics. 2020. 223-230.

NUNN, S. 2009. Driving compliance through data governance. *Journal of the American Health Information Management Association,* 80**,** 50-51.

OPPENHEIM, C., STENSON, J. & WILSON, R. M. S. 2003. Studies on information as an asset I: Definitions. *Journal of Information Science,* 29**,** 159-166.

OTTO, B. 2011a. Data governance. *Business and Information Systems Engineering,* 3**,** 241-244.

OTTO, B. A morphology of the organisation of data governance. 19th European Conference on Information Systems, ECIS 2011, 2011b.

OTTO, B. 2011c. Organizing Data Governance: Findings from the telecommunications industry and consequences for large service providers. *Communications of the Association for Information Systems,* 29**,** 45-66.

PANIAN, Z. 2009. Recent advances in data management. *WSEAS Transactions on Computers,* 8**,** 1061-1071.

PANIAN, Z. 2010. Some practical experiences in data governance. *World Academy of Science, Engineering and Technology,* 62**,** 946-953.

PHILIP CHEN, C. L. & ZHANG, C. Y. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences,* 275**,** 314-347.

PHILLIPS-WREN, G., IYER, L. S., KULKARNI, U. & ARIYACHANDRA, T. 2015. Business analytics in the context of big data: A roadmap for research. *Communications of the Association for Information Systems,* 37**,** 448-472.

PIERCE, E., DISMUTE, W.S. AND YONKE, C.L. 2008. Industry Report: The State of Information and Data Governance: Understanding How Organizations Govern their

Information and Data Assets. Baltimore, MD.: International Association for Information and Data Quality,.

PRASETYO, H. N. & SURENDRO, K. 2015. Designing a data governance model based on soft system methodology (SSM) in organization. *Journal of Theoretical and Applied Information Technology,* 78**,** 46-52.

RADCLIFFE, J. 2007. The Seven Building Blocks of MDM: A Framework for Success.

RAGUSEO, E. 2018. Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management,* 38**,** 187-195.

RAGUSEO, E. & VITARI, C. 2018. Investments in big data analytics and firm performance: an empirical investigation of direct and mediating effects. *International Journal of Production Research,* 56**,** 5206-5221.

REEVES, M. G. & BOWEN, R. 2013. Developing a data governance model in health care. *Healthcare financial management : journal of the Healthcare Financial Management Association,* 67**,** 82-86.

REHMAN, A., NAZ, S. & RAZZAK, I. 2021. Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*.

RIFAIE, M., ALHAJJ, R. & RIDLEY, M. Data governance strategy: A key issue in building enterprise data warehouse. iiWAS2009 - The 11th International Conference on Information Integration and Web-based Applications and Services, 2009. 587-591.

RIGGINS, F. J. & KLAMM, B. K. 2017. Data governance case at KrauseMcMahon LLP in an era of self-service BI and Big Data. *Journal of Accounting Education,* 38**,** 23-36.

ROSENBAUM, S. 2010. Data governance and stewardship: Designing data stewardship entities and advancing data access. *Health Services Research,* 45**,** 1442-1455.

RUSSOM, P. 2006. Taking Data Quality to the Enterprise through Data Governance.

RUSSOM, P. 2008. Data Governance Strategies: Helping Your Organization Comply, Transform, and Integrate. Renton, WA.: TDWI.

SAXBY, S. 2014. The 2013 CLSR-LSPI seminar on electronic identity: The global challenge-Presented at the 8th International Conference on Legal, Security and Privacy issues in IT Law (LSPI) November 11-15, 2013, Tilleke & Gibbins International Ltd., Bangkok, Thailand. *Computer Law and Security Review,* 30**,** 112-125.

SESTINO, A., PRETE, M. I., PIPER, L. & GUIDO, G. 2020. Internet of Things and Big Data as enablers for business digitalization strategies. *Technovation,* 98.

SHAW-TAYLOR, Y. 2014. Making quality improvement programs more effective. *International Journal of Health Care Quality Assurance,* 27**,** 264-270.

SIVARAJAH, U., KAMAL, M. M., IRANI, Z. & WEERAKKODY, V. 2017. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* 70**,** 263-286.

SOARES, S. 2010. *The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices,* US, MC Press

SOARES, S. 2012. *Big Data Governance: An Emerging Imperative*, Mc Press.

SOUZA, R. & PATIL, C. 2016. Big data integration, privacy, and security. *The Human Element of Big Data: Issues, Analytics, and Performance.*

SPRAGUE JR, R. H. 1980. A framework for the development of decision support systems. *MIS Quarterly: Management Information Systems,* 4**,** 1-26.

STRAUSS, A. L. & CORBIN, J. 1990. *Basics of Qualitative Research,* Thousand Oaks, CA, Sage.

STRAUSS, A. L. & CORBIN, J. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory,* London, Sage.

SUKUMAR, S. R., NATARAJAN, R. & FERRELL, R. K. 2015. Quality of Big Data in health care. *International Journal of Health Care Quality Assurance,* 28**,** 621-634.

TALLON, P. 2013. Corporate governance of big data: Perspectives on value, risk, and cost. *Computer,* 46**,** 32-38.

TALLON, P. P., RAMIREZ, R. V. & SHORT, J. E. 2013. The information artifact in IT governance: Toward a theory of information governance. *Journal of Management Information Systems,* 30**,** 141-178.

TAN, K. H., JI, G., LIM, C. P. & TSENG, M. L. 2017. Using big data to make better decisions in the digital economy. *International Journal of Production Research,* 55**,** 4998-5000.

TAPLIN, R. & SCHYMYCK, N. 2005. An interdisciplinary and cross-cultural approach. *Risk Management and Innovation in Japan, Britain and the United States.* UK: Routledge.

THALER, R. H. & TUCKER, W. 2013. Smarter information, smarter consumers. *Harvard Business Review,* 91.

THOMAS, G. 2006. The DGI Data Governance Framework.

URQUHART, C. & FERNÁNDEZ, W. 2013. Using grounded theory method in information systems: The researcher as blank slate and other myths. *Journal of Information Technology,* 28**,** 224-236.

VASSAKIS, K., PETRAKIS, E. & KOPANAKIS, I. 2018. Big data analytics: Applications, prospects and challenges. *Lecture Notes on Data Engineering and Communications Technologies.*

WANG, R. Y. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems,* 12**,** 5-34.

WANG, R. Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM,* 41**,** 58-65.

WEBER, K., OTTO, B. & ÖSTERLE, H. 2009. Ones size does not fit all -A contingency approach to data governance. *Journal of Data and Information Quality,* 1.

WENDE, K. A model for data governance - Organising accountabilities for data quality management. ACIS 2007 Proceedings - 18th Australasian Conference on Information Systems, 2007. 416-425.

WENDE, K. & OTTO, B. A contingency approach to data governance. Proceedings of the 2007 International Conference on Information Quality, ICIQ 2007, 2007.

WERE, V. & MOTURI, C. 2017. Toward a data governance model for the Kenya health professional regulatory authorities. *TQM Journal,* 29**,** 579-589.

WHYTE, J., STASIS, A. & LINDKVIST, C. 2016. Managing change in the delivery of complex projects: Configuration management, asset information and 'big data'. *International Journal of Project Management,* 34**,** 339-351.

WINTER, J. S. & DAVIDSON, E. 2019. Big data governance of personal health information and challenges to contextual integrity. *Information Society,* 35**,** 36-51.

YALLOP, A. & SERAPHIN, H. 2020. Big data and analytics in tourism and hospitality: opportunities and risks. *Journal of Tourism Futures,* 6**,** 257-262.

YEBENES, J. & ZORRILLA, M. Towards a data governance framework for third generation platforms. 2019. 614-621.

ZHANG, G. A data traceability method to improve data quality in a big data environment. 2020. 290-294.

ZHANG, S., GAO, H., YANG, L. & SONG, J. Research on big data governance based on actor-network theory and Petri nets. Proceedings of the 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design, CSCWD 2017, 2017. 372-377.

ZHAO, J. G. Marrying risk register with project trending. 2005. 1-6.

Appendices

Appendix A Parliamentary Inquiry Dataset: Participants List and Corresponding References

Parliamentary Inquiry: The Big Data Dilemma			
Atlas ti #	**References**	**Participant(s) List**	**Industries / Representation**
D1	HC 992	House of Commons Science & Technology Committee. Government Response to the Committee's Fourth Report of Session 2015–16	Other
D2	HC 468c	House of Commons Science & Technology Committee. Fourth Report of Session 2015–16	Written Evidence (Other)
D3	HC 468b	Oral Evidence from Royal Statistical Society, Department of Business Innovation & Skills, and the Department for Culture, Media & Sport.	Oral Evidence (Other)
D4	HC 468b	Oral Evidence from Direct Marketing Association, Nuffield Council of Bioethics, Big Brother Watch & Information Commissioners Office	Oral Evidence (Other)
D5	BIG0079	Professor Sir John Tooke, President of the Academy of Medical Sciences	Medical/Health Care

D6	BIG0063	Alzheimer's Research UK	Research & Development
D7	BIG0042	ANS Group	Technology
D8	BIG0030	The Association of Medical Research Charities	Medical/Health Care
D9	BIG0062	Barclays Bank	Banking/Financial
D10	BIG0038	BCS, The Chartered Institute for IT	Professional Body
D11	BIG0006	Big Brother Watch	Campaign group
D12	BIG0005	Mediatrac	Information Technology
D14	BIG0078	British Bankers Association	Banking/Financial
D15	BIG0076	British Standards Institution	National standards body
D16	BIG0064	The British Academy	Education
D17	BIG0023	The British Heart Foundation	Non-Profit
D18	BIG0051	The British Library	Library
D19	BIG0009	The British Pharmacological Society	Medical/Health Care
D20	BIG0049	The British Vehicle Rental and Leasing Association	Transportation
D21	BIG004	Brunella Longo	Other
D22	BIG0056	British Veterinary Association (BVA)	Medical/Health Care
D23	BIG0007	CompanyNet	Information Technology
D24	BIG0013	Consumer Data Research Centre – University College London	Research & Development
D25	BIG002	The Department for Environment Food & Rural Affairs (Defra)	Public sector
D26	BIG0069	The Department for Business, Innovation and Skills, and the Department for Culture Media & Sport	Public sector
D27	BIG0089	Mr Edward Vaizey MP, Minister for Digital Economy	Public sector
D28	BIG0080	Digital Catapult	Information Technology

D29	BIG0048	Digital Science	Information Technology
D30	BIG0014	The Direct Marketing Association (UK) Limited	Trade association
D31	BIG003	Dr Kevin McNish	Education
D32	BIG0040	The E-infrastructure Leadership Council (ELC)	Information Technology
D33	BIG0046	EMC	Information Technology
D34	BIG0055	Ernst & Young LLP	Consulting
D35	BIG0021	The Economic and Social Research Council (ESRC) Big Data Network, Phase 2	Research & Development
D36	BIG0022	Experian	Information services
D37	BIG0074	The Farr Institute of Health Informatics Research	Research & Development
D38	BIG0019	The Food Ethics Council	Non-Profit
D39	BIG0037	The Food Standards Agency	Public sector
D40	BIG0045	Fujitsu Services Ltd	Information Technology
D41	BIG0081	Funding Circle	Other
D42	BIG0020	GARNet	Other
D43	BIG0077	GBG	Information Technology
D44	BIG0034	Genetic Alliance UK	Non-Profit
D45	BIG0067	The Greater London Authority and Transport for London	Transportation
D46	BIG0001	The Institute of Chartered Accountants in England and Wales (ICAEW)	Accounting body
D47	BIG0065	The Institution of Engineering and Technology (IET) and the Royal Academy of Engineering	Research & Development
D48	BIG0033	The Information Commissioner's Office	Public sector
D49	BIG0026	Innovate UK	Public sector

D50	BIG0071	The Institute of Cancer Research	Research & Development
D51	BIG0052	The Insurance Law Research Group, University of Southampton	Research & Development
D52	BIG0054	Internet Advertising Bureau UK	Advertising
D53	BIG0010	Involve	Non-Profit
D54	BIG0085	Jen Perssen	Public sector
D55	BIG0027	Jisc	Information Technology
D56	BIG0061	John Innes Centre	Research & Development
D57	BIG0090	Leaderboarded Ltd	Information Technology
D58	BIG0068	Lloyds Register Foundation	Non-Profit
D59	BIG0088	Medconfidential	Medical/Health Care
D60	BIG0036	Medical Schools Council	Medical/Health Care
D61	BIG0060	Microsoft	Information Technology
D62	BIG0017	Ms Susan Hall	Legal
D63	BIG0032	Mydex CIC	Data services
D64	BIG0012	National Oceanography Centre	Non-Profit
D65	BIG0041	National Physical Laboratory	Metrology
D66	BIG0047	Nesta	Innovation agency
D67	BIG0015	Nuffield Council of Bioethics	Medical/Health Care
D68	BIG0084	Open Data Institute	Data services
D69	BIG0070	The Open University	Education
D70	BIG0073	PHG Foundation	Medical/Health Care
D71	BIG0082	Professor Lorna Woods	Other
D72	BIG0050	Pure Storage	Information Technology
D73	BIG0057	Research Council UK	Research & Development
D74	BIG0029	Royal College of Physicians	Medical/Health Care
D75	BIG0044	Royal Statistical Society	Professional Body
D76	BIG0059	SAP	Information Technology
D77	BIG0075	Science and Engineering South Consortium	Science & Engineering
D78	BIG0024	Smart Energy GB	Non-Profit

D79	BIG0018	South London and Maudsley NHS Foundation Trust	Medical/Health Care
D80	BIG0028	Swanbarton Ltd	Energy
D81	BIG0086	TechUK (Sue Daley)	Non-Profit
D82	BIG0039	TechUK	Non-Profit
D83	BIG0035	The Royal Society	Research & Development
D84	BIG0011	UK Computing Research Committee	Engineering & Technology
D85	BIG0008	UK Data Archive	Data services
D86	BIG0016	University of Cambridge Big Data Strategic Research Initiative	Education
D87	BIG0043	University of Edinburgh	Education
D88	BIG0053	University of Essex	Education
D89	BIG0058	Wellcome Trust	Non-Profit
D90	BIG0066	Wellcome Trust Sanger Institute	Research & Development
Parliamentary Inquiry: Social media data and real time analytics			
Atlas ti #	**References**	**Participant(s) List**	**Industries / Representation**
D91	HC245	Science and Technology Committee	Government
D92	SMD0001	Dr Kevin Macnish	Other
D93	SMD0002	EMC	Information Technology
D94	SMD0003	Professor Liesbet van Zoonen	Other
D95	SMD0004	University of Leeds	Education
D96	SMD0005	University of Essex & University of East London	Education
D97	SMD0006	Aidan McCormack, Department of Work and Pensions	Public sector
D98	SMD0007	UKCRC	Research & Development
D99	SMD0008	Alliance for Useful Evidence	Research & Development
D100	SMD0009	Web Science Trust and the EPSRC project SOCIAM	Non-Profit

D101	SMD0010	Digital Science	Information Technology
D102	SMD0011	Institute of Communications Studies, University of Leeds	Education
D103	SMD0012	Horizon Digital Economy Research Institute	Research & Development
D104	SMD0013	Dr Ella McPherson and Dr Anne Alexander (University of Cambridge)	Education
D105	SMD0014	The Open University	Education
D106	SMD0015	Dept of Social Statistics, University of Manchester	Education
D107	SMD0018	Information Commissioner's Office	Public sector
D108	SMD0019	FRSA	Other
D109	SMD0020	Department for Business, Innovation and Skills, the Department for Culture, Media and Sport, and the Ministry of Justice	Public sector
D110	SMD0021	Big Brother Watch	Campaign group
D111	SMD0022	Economic & Social Research Council (on behalf of RCUK)	Research & Development
D112	SMD0023	TechUK	Non-Profit
D113	SMD0024	Daniel Trottier, PhD	Other
D114	SMD0025	Office for National Statistics	Other
D115	SMD0026	Facebook	Information Technology
D116	SMD0027	LinkedIn	Information Technology
D117	SMD0028	Yahoo	Information Technology
D118	SMD0029	Twitter	Information Technology
D119	SMD0030	TechUK (GDPR Working Group)	Non-Profit
*D120	SMD0031	Google (Joint submission with TechUK, hence not coded separately)	Information Technology

Appendix B Summarised Review Matrix of Data Governance Frameworks

Author(s) (Year) Title	DG Approach / Framework identified	Methodology	BD Governance framework / Approach specified
Cheong, L.K. and Chang, V. 2007. The Need for Data Governance: A Case Study	Data governance framework which is organised in three dimensions: Organisational bodies and policies, standards and processes and data governance technology, distributing the internal components that correspond to DG capabilities and activities along these dimensions. It proposed three different levels of engagement between IT and business departments i.e. strategic, tactical, and operational. Additionally, the paper identifies 9 DG Critical Success factors:	Case Study	No.
Khatri, V. and Brown, C.V. 2010. Designing Data Governance	Data governance framework proposing five interrelated decision domains: • Data principles • Data quality • Metadata • Data access • Data life cycle.	Field Research	No

	A one-page design matrix is proposed for communicating a company's data governance approach.		
Wende, K. 2007. A Model for Data Governance - Organising Accountabilities for Data Quality Management.	Data governance model comprising three components i.e. data quality roles, decision areas and responsibilities.	Not specified	No
Weber, K. and Otto, B. 2007. A Contingency Approach to Data Governance.	Data governance model that provides a matrix comparable to a RACI covering data quality roles, decision areas and responsibilities. The model documents the data quality roles and their type of interaction with data quality management activities that affect data governance, highlighting that a data governance configuration is unique to a given organisation. It proposes a first set of contingencies and aimed to demonstrate their impact on the data governance model proposed. Two design parameters facilitate understanding the way	Community action research project	No

	contingencies affect the individual design of a company's data governance model. They range from two opposed value pairs centralised/decentralised and hierarchical/cooperative.		
Weber, k., Otto, B., and Osterle, H. 2009. One Size Does Not Fill All – A Contingency Approach to Data Governance.	Data governance model which documents data quality roles and their type of interaction with data quality management activities, decision areas and responsibilities. In addition, the approach defines a data contingency model for data governance.	Community action research project consisting of subject matter experts from six international companies spanning various industries.	No
Panian, Z. 2010. Some Practical Experiences in Data Governance.	Data governance framework with four components that aid to manage and develop six key enterprise data attributes: accessibility, availability, quality, consistency, auditability, and security. The four key components of the data governance framework proposed include standards, policies and processes, organisation structure and technology.	Not specified	No

Tallon, P. 2013. Corporate Governance of Big Data: Perspectives on Value, Risk and Cost.	Data governance practices which comprise three separate categories: Firstly, structural practices that cover 1) user involvement in policy setting and evaluation, 2) steering committees to assess data value and costs, 3) creation of data ownership and stewardship roles and responsibilities. Secondly, operational practices that span all stages of the information life cycle. Finally, relational practices such as 1) The education of users and non-IT Managers on storage utilisation and cost. 2) Development of communications regarding policy effectiveness and user needs.	Research conducted at Loyola University, Maryland.	Yes.
Otto, B. 2011. Organising Data Governance: Findings from the Telecommunicatio ns Industry and Consequences for Large Service Providers.	A conceptual data governance framework, which shows data governance as an organisational design task comprising of three dimensions i.e. organisational goals, organisational structure, and organisation transformation. The framework also specifies empirical indicators related to the organisational dimensions.	Case study research	No.

Al-Ruithe, M., Benkhelifa, E., and Hameed, K. 2016. Data Governance Taxonomy: Cloud versus Non-Cloud. .	Overall taxonomies of data governance for cloud versus non-cloud computing networks. It presents its taxonomy of data governance development requirements in the organisational, technical, and legal context.	Systematic literature review	No.
Rosenbaum, S. 2010. Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access.	Data stewardship as a mechanism for Data Governance	Insights from stakeholder and commenters views	No
Prasetyo, H.N., and Surendro, K. 2015. Designing a Data Governance Model based on Soft System Methodology (SSM) in Organisation.	Data governance model based on a Soft System Methodology approach.	Literature Review	No
Kim, H.Y., and Cho, J. 2018. Designing a Data Governance Framework for Big Data	Big data governance framework to facilitate implementation of big data services. Components include organisation, standards and guidelines, processes, and policies.	Case Analysis	Yes.

Implementation with NPS Case Analysis in Korea.			
Zhang, S., Gao, H., and Song, J. 2017. Research on Big Data Governance Based on Actor-Network Theory and Petri nets.	Big data governance model based on actor-network theory and Petrinets.	Actor-network theory	Yes
Al-Badi, A., Tarhini, A. and Khan, A.I. 2018. Exploring Big Data Governance Frameworks.	Conceptual big data governance framework, which consists of eight components	Literature Review	Yes
Thomas, G. 2006. The DGI Data Governance Framework.	Data Governance Institute (DGI) data governance framework with 10 universal components that put emphasis on three areas i.e. Rules and Rules of engagement, People, and organisational bodies and finally, processes.	Not specified	No
IBM Software group. 2007. The IBM Data Governance Council Maturity Model: Building a Roadmap for	IBM data governance maturity model with 11 domains. Each of the 11 categories of the model contain many sub-categories grouped into five levels of maturity, which include:	Not specified The model was professed to of been designed with	No

Effective Data Governance.	Level 1 – Initial Level 2 – Managing Level 3 – Defined Level 4 – Quantitatively managed Level 5 – Optimising	input from a collaborative council of 55 leading companies. The exact methodology utilised was not described in detail.	
Radcliff, J. 2007. The Seven Building Blocks of MDM: A Framework for Success.	Governance is defined as one of the seven building blocks of Gartner's master data management framework. The seven building blocks of the MDM framework are MDM Vision, MDM Strategy, MDM Governance, MDM Organisation, MDM Processes, MDM Technology Infrastructure, MDM Metrics.	Not specified	No
The Data Warehousing Institute. 2010. TDWI Data Governance Fundamentals.	A model that outlines the data governance organisation, some decision areas and suggested processes to govern organisation data. It also notes that there are different types of Data Steward roles i.e. by subject area, business unit, project, and enterprise. Furthermore, suggests some key processes to govern data and	Not specified	No

	guidelines for driving a data governance program.		
Soares, S. 2010. The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices	Data Governance processes, involving 14 steps and almost 100 sub-steps to perform integrated data governance.	Not specified	No
Soares, S., Deutsch, T., Hanna, S., and Malik, P. 2012. Big Data Governance: A Framework to Assess Maturity	Maturity Assessment Model for big data governance.	Not specified	Yes
Soares, S. 2012. Big Data Governance: An Emerging imperative	Big data governance framework which provides a guide to implementing data governance within a company. In summary, it recommends: • A maturity assessment • A business case justification for implementing big data governance initiatives. • A data governance roadmap to guide implementation.	Not specified.	Yes

	The book also provides specification of some anonymised industry use cases.		
Loshin, D. 2013. Data Governance for Master Data Management and Beyond.	Framework for data governance management, which defines roles and accountabilities for the sponsorship, oversight, coordination and stewardship of data within an organisation. It proports that there are 3 major aspects of data governance for master data management. 1. Managing critical data elements 2. Setting information policies and data rules 3. Enforcing accountability.	Not specified	No
Reeves, M.G., and Bowen, R. 2013. Developing a Data Governance in Health Care.	Data governance model to reduce risk and enhance compliance. It recommends a metric to measure the value of a data governance program – confidence in data – dependent assumptions (CIDDA). CIDDA is measured as follows: CIDDA = G x M x TS G = Confidence that data are good enough for their intended purposes	Case Analysis	No

	M= Confidence that data mean what you think they do TS= Confidence that you know where the data comes from and trust their source. Although there are no industry benchmarks for CIDDA, each organisation can assign its own values and goals to the metric.		
Cupoli, P., Earley, S., and Henderson, D. 2014. The Data Management Association DAMA-DMBOK2 Framework	Data management framework, consisting of 11 knowledge areas, with data governance positioned as a central component. Each knowledge area has a context diagram that delineates and structures the scope of the area.	Not Specified	No
Karel, R. 2015. Enabling Data Governance in a Big Data World.	Data governance model with process stages to enable big data. (This framework proposes traditional data governance processes to enable big data initiatives). This practitioner reports that there are over twenty distinct processes segmented into four core process stages – all of which are iterative	Not specified	No.

	and encompass numerous parallel activities depending on the stage of data governance maturity a company has reached.		
Informatica, 2017. Holistic Data Governance: A Framework for Competitive Advantage.	Data governance framework with ten complimentary facets	None specified	No
Abraham, R., Brocke, J.V., and Schneider, J. 2019. Data Governance: A conceptual framework structured review, and research agenda	A conceptual data governance framework broken down against six dimensions. The key dimensions are related to the governance mechanisms i.e. structural, procedural, and relational mechanism. This framework also specifies the antecedents that impact adoption of DG and consequences of the effects of DG. This framework is not classified as a big data governance framework. However, it recognises the context of distinct BD types and the requirement to handle BD risk.	Literature Review	No

Li, Q et al (2019) A Framework for Big Data Governance to Advance RHINs: A Case Study of China	A framework consisting of 3 domains and 12 guidelines for healthcare big data governance based on Chinese practice. Centred on addressing the issues to the rise of regional health information networks.	Case Analysis	Yes
Mao, Z., Wu, J., Qiao, Y and Yao, H (2021) Government Data Governance Framework based on a data middle platform	A government data governance framework with the data middle platform as its core. The paper exposed the key requirements and issues of government data governance, and address the main problem related to facilitation of the datafication of administrative processes.	Literature review	No
Yebenes, J and Zorrilla, M (2019). Government Data Governance Framework based on a data middle platform	Proposes a conceptual schema that can guide an organisations data governance programme for third generation platform to achieve industry 4.0 objectives. ceptual tool which guides organizations to define, design, develop and deploy services aligned with its vision and	Literature review	No

Appendix C Sample Representation of Participant's Ethical concerns mapped to the BDLC stages

Big Data Lifecycle stage	**Breakdown of Ethical concerns related to BD and its associated practices**	**Reference – A few representative examples of participant's submissions**
All stages	Complexity and scale of BD and related BD analytics impedes transparency or trust	BIG002, BIG0038, BIG0033, HC 468 (oral Evid), BIG0085, BG001, BIG0058
All stages	Inappropriate governance / oversight for dealing with BD ethical challenges	HC 992, BIG0069, BIG0019, BIG0037, BIG0015, SMD0001, HC 245, SMD0022, SMD0003, BIG0018
All stages	BD misuse in ways that increase unfairness	BIG0038, HC 468, BIG0069, BIG0019, BIG0015, SMD0004, SMD0011, BIG0052
All stages	Complexity of BD hinders auditability/verification within BD ecosystem	HC 468, SMD0013, BIG0043, BIG002, SMD0018
All stages	Risk of privacy infringement by organisations utilising BD	BIG002, HC 468, BIG0069, BIG0046, BIG0021, BIG0019, SMD0013, SMD0015, BIG0011, HC 245, SMD0005, SMD0021
All stages	Procedures / rules for BD protection not utilised	HC 468, BIG0037, BIG0015, SMD0012, SMD0018
All stages	Shortcomings leading to control/power disparities e.g. • no control over how personal data in BD is used/reused/shared. • public engagement not sought (or not possible) even when citizens are the data subjects of BD initiatives.	BIG0069, BIG0038, BIG0085, HC 245, SMD0012, SMD0013, SMD0021, BIG0070, SMD0015, SMD0024

	• deficient BD analysis techniques etc.	
All stages	BD ownership challenges i.e., the question of ownership once data of a subject is captured by an organisation is a complicated ethical issue e.g. data ownership after a subject's death is a grey area.	BIG0006, BIG0018, SMD0021, SMD0004, SM0015, SMD0030
All stages	Lack of organisation policies, procedures or good practice guidelines to prevent ethical problems	BIG0019, HC 992, BIG0015, HC 468, SMD0008, SMD0009, SMD0013, SMD0003, SM0015
All stages	Inadequate or missing standards to prevent ethical problems connected to BD collection, analysis, usage or sharing	HC 468, BIG0019, HC 245
All stages	Unscrupulous BD capture for economic gain / monopoly	HC 468 (oral Evid), BIG0080,
All stages	Lack of accountability, roles and responsibilities to address ethical handling of BD	HC 992, HC 468, HC 468 (oral Evid), SMD0003, SMD0009, BIG0006, BIG0037, BIG0015, SMD0015
All stages	Insufficient ethical emphasis on the security of BD	BIG002, BIG0069, SMD0015, SMD0005, SMD0004
BD Capture	Data minimisation i.e. Not ensuring the bare amount of BD is collected	BIG0019, BIG0037, BIG0033
BD Capture	Lack of Informed consent i.e. • are there protocols in place to ensure a subject's consent for collection is truly informed? • does original consent acquired still reflect the preferences of the individual who originally provided it?	BIG003, BIG0037, BIG015, HC 245, SMD0001, SMD0013, SM0015, BIG0032,
BD Capture	Ethical concerns with the purpose of BD collection	BIG0021, SMD0001, SMD0004, SMD0003, BIG0055, BIG0037, HC 245

BD Capture	Bias introduced when factors are selected, excluded or developed into BD algorithms or BD software programs.	BIG008, BIG0064, BIG0053
BD Capture, sharing	Data subjects choose not to participate in BD initiatives due to a trust deficit related to perceived unethical practices	BG0011, SMD0003, BIG0038, SMD0018
BD Capture	Inappropriate ethical regulatory framework hindering innovation or research	BG0011, HC 245, SMD0028, BIG0074
BD processing	Failure to accurately integrate/aggregate BD	BIG002, BIG0055, SMD0004, SMD0007, BIG0085
BD processing	Inadequate BD governance/code of practice/frameworks for ethical BD use and related BD analytics	HC 468, BIG0019. BIG0037, BIG0015, BG0011, SMD0013, SM0015, BIG0018, HC 245
BD processing/capture	Lack of safeguards for BD use or re-use for unexpected purposes or purposes not disclosed	BIG0015, BG011, BIG0017, BIG0006, BIG0032, BIG0074
BD processing/capture	Concerns with secondary use of BD	BIG003, BIG0015, SMD0004, SMD0015, SMD0021, BIG0055, SMD0018, SMD0024
BD processing/curation	Personal data within BD systems insufficiently anonymised or re-identified	BIG002, BIG0037, BG0011, SMD0004, BIG0069, BIG0015, SMD0013, SMD0015, SMD0023
BD processing/curation	Inadequate BD infrastructure to block ethical issues	BIG0051, SMD0009, BIG0070
BD processing/application	Linking and combining of multiple anonymised datasets which can lead to re-identification/questionable BD analysis etc.	BIG002, BIG0037, BIG015, HC 468, BIG0006, BIG0069, BIG0085, BIG0043, BIG0058, SMD0023
BD processing/application	Biased analytical methodology or biased utilisation of BD collected	BIG0008, SMD0012, BIG0070, BIG0011, BIG0016, SMD0025

BD processing/application	bias in BD analytic methods	BIG0085, SMD0021, BIG0033, BIG0052, BIG0015, SMD0011
BD processing/application	Decision-making by automated techniques replacing human decision making	BIG0078, BIG0040, BIG0045, BIG0052, HC 468 (Oral Evid), BIG0022
BD processing/application	Functional creep or utilisation of BD analytics for unintended purposes	BIG002, BIG0027, SMD0001, SMD0013, SMD0021, SMD0015, BIG0055, BIG0006, BIG0033
BD processing/application	Lack of differentiation between the use of BD for commercial or research purposes	BIG0071, SMD0007
BD curation	Limitless life span of BD: • Concerns with BD being retained for longer than is needed. • Organisations archiving or failing to destroy BD	BIG0015, SMD0018, BIG0033,
BD curation	Security of BD e.g. unauthorised disclosure or access to BD	BIG003, HC 468, BIG0015, SMD0001, SMD0004, SMD0013, SMD0018, SMD0015
BD sharing	Failure to make BD public	SMD0024, SMD0015
BD sharing	Inappropriate BD sharing practices	BIG0038, HC 468, BIG0074, SMD0013, BIG0090, SMD0024
BD sharing	Lack of appropriate ethical safeguards preventing BD sharing or limiting uptake of BD	HC 468, BIG0015, SMD0009, BIG0038
BD application	Broad based monitoring / surveillance beyond ethical boundaries	BIG0019, HC 245, SMD0011, SMD0013, SM0018, SMD0021, SMD0024
BD application	Unfair predictions or targeting by BD analytical tools or techniques	BIG0054, SMD0001, SMD0011, SMD0021,

		SMD0018, BIG0057, BIG0006, BIG0033
BD application	Individual or group-level harm from BD profiling	BIG0054, BIG0015, BIG008, SMD0001, SMD0004, SMD0012, SMD0021, SMD0024, BIG0052, BIG0057
BD application	BD ownership or Intellectual property obscured or compromised	SMD0001, SMD0004, SMD0013, SMD0019, SMD0015, SMD0018, SMD0024
BD deletion	Premature or unintentional BD destruction	HC468a