Santeri Palomäki

IDENTIFYING AND VALIDATING KEY CHALLENGES OF BIG DATA-BASED DECISION-MAKING: A FRAMEWORK MAPPING OUT CHALLENGES FROM DATA TO DECISIONS

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2020

TIIVISTELMÄ

Palomäki, Santeri
Identifying and validating key challenges of Big Data-based decision-making: A framework mapping out challenges from data to decisions
Jyväskylä: Jyväskylän yliopisto, 2020, 70 s.
Tietojärjestelmätiede, Pro Gradu -tutkielma
Ohjaaja: Kazan, Erol

Big Datan rooli yritysten päätöksenteossa on muuttunut yhä tärkeämmäksi viime vuosikymmenen aikana. Syitä tähän ovat muun muassa huomattava kasvu datan määrässä maailmassa, sekä sen keräämisessä ja prosessoinnissa tehdyt harppaukset. Monet haasteet ovat nostaneet päätään yritysten pyrkiessä niittämään Big Datasta saatavia hyötyjä päätöksenteossaan, ja tämä on vaikeuttanut liiketoimintahyötyjen maksimointia. Nämä haasteet ovat liittyneet esimerkiksi dataan, prosessointiin ja johtamiseen. Big Datan muuttuessa tunnetummaksi ilmiöksi, on siihen kohdistuvan tutkimuksenkin määrä kasvanut sen mukana. Tämä on johtanut hajanaiseen näkemykseen Big Datan määritelmästä alan kirjallisuudessa. Tämän tutkielman tarkoitus on tarjota nykyaikainen ja kattava määritelmä Big Datalle, sekä perusteellinen kartoitus Big Data-pohjaiseen päätöksentekoon liittyvistä haasteista. Kirjallisuuskatsaus toteutettiin näiden tavoitteiden saavuttamiseksi. Kirjallisuuskatsauksen lisäksi järjestettiin teemahaastatteluja alan ammattilaisille vaihtelevilla taustoilla ja työhistorioilla. Haastattelujen pohjalta tunnistettiin 16 teemaa, joiden kautta validoitiin alan kirjallisuudessa löydettyjä haasteita. Tutkimuksen tuloksena on yksityiskohtainen kuvaus kaikista alan kirjallisuudessa merkittäviksi todetuista haasteista, jotka tulee huomioida Big Data-pohjaisessa päätöksenteossa, sekä ajankohtainen määritelmä itse Big Datalle. Lisäksi kehitettiin ja validoitiin uusi viitekehys, jolla visualisoidaan vielä yksityiskohtaisemmin tutkimuksessa tunnistettujen haasteiden välisiä suhteita. Tutkielman tuloksissa esitellään myös haastateltujen alan ammattilaisten näkemys nykypäivän oleellisimmista Big-Data-pohjaisen päätöksenteon haasteista yrityksille, mikä toimii tärkeänä käytännön implikaationa tämän tutkielman osalta.

Avainsanat: big data, big data-analytiikka, päätöksenteko, datapohjainen päätöksenteko

ABSTRACT

Palomäki, Santeri
Identifying and validating key challenges of Big Data-based decision-making: A framework mapping out challenges from data to decisions
Jyväskylä: University of Jyväskylä, 2020, 70 pp.
Information systems science, Master's thesis
Supervisor: Kazan, Erol

Big Data's role in organizational decision-making has become increasingly important during the last decade. This is due to, inter alia, a massive increase in the amount of data in the world, as well as advancements made in gathering and processing techniques for data sets of this size. A plethora of challenges have been noted to present themselves as organizations are trying to reap the benefits of Big Data in decision-making, thus hindering the realized business benefits. These challenges are related to, for example, data, processing, and management. As Big Data has become more relevant as a phenomenon, research of it has also increased. This increased research has created a scattered view of Big Data definition in the literature of the field. This study seeks to provide a current, all-inclusive definition of BD and to comprehensively map out relevant challenges associated with Big Data-based decision-making. To achieve this, a literature review was conducted to identify key Big Data-based decision-making challenges found in the literature of the field. In addition to the literature review, a set of semi-structured interviews was conducted with industry professionals with varied backgrounds and professional experience. Based on the interviews, 16 different themes were identified and further used to validate the challenges found in the literature of the field. The result of this study is a detailed description of all relevant challenges that should be addressed in Big Data-based decision-making accompanied by a definitive explanation of BD itself. A new validated framework is also provided to further visualize the relations between different challenges identified in this study. Additionally, challenges found most relevant by the practitioners of the field are presented in the results of this study, which provides important practical implications for this thesis.

Keywords: big data, big data analytics, decision-making, data-driven decision-making

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

Data available for analysis in the world is massive. By estimate, the amount of data in the world in 2020 will be 40 zettabytes ("Big Data Statistics 2019", 2019). A zettabyte is equal to 1 trillion gigabytes. If one gigabyte were equal to one drop of water, trillion gigabytes would add up to around 50 000 000 liters of water. And the rate of growth is imminent, as 90% of all this data has been generated just over the past two years ("Big Data Statistics 2019", 2019). Due to this massive increase in data volume, Big Data (BD) and Big Data Analytics (BDA) have taken over the world during the last decade. It has been predicted by Press (2017) that the BDA market will surpass $203 billion in worldwide revenue by 2020. Research has also noted that it is very difficult to open a popular publication without running to at least a side note or reference regarding BD or data analytics in general (Agarwal & Dhar, 2014). As this was the case in 2014, it is presumably an even more prevalent trend in the current day world of 2020. Data analytics' competitive capabilities have also been clear for a while due to studies recognizing that best-performing organizations utilize data analytics five times more often than lower performers (LaValle, Lesser, Shockley, Hopking & Kruschwitz, 2011).

BD has become a key part of business processes for many reasons. Economic and social transactions have moved online (Agarwal & Dhar, 2014), and storage costs have decreased in combination with advancements in computer processing power (Moorthy et al., 2015). Further – as described above – exponentially more data has become available for organizations to utilize. All these combined with a certain level of hype around BD and BDA have led to more and more organizations adopting BDA to their business processes to reap the benefits. According to Russom (2013), 75% of organizations manage some sort of BD.

BD poses many possibilities for organizations. It offers the ability to examine and measure micro-level data to address policies and business strategies, provides cost reductions (Balachandran & Prasad, 2017; Thabet & Soomro, 2015), enhances business performance (Moorthy et al., 2015), improves decision-making (Balachandra & Prasad, 2017; Thabet & Soomro, 2015) and improves

existing products and services (Saggi & Jain, 2018; Thabet & Soomro, 2015). Also, BD offers a brand-new research context for academics for qualitative- and quantitative studies, as well as design science (Agarwal & Dhar, 2014). And researches have embraced this, as demonstrated in figure 1, which displays the increase of BD related studies in Scopus, Google Scholar, and ScienceDirect databases during the current decade. The numbers were calculated by searching for papers using "big data" in the article's title, abstract, or keywords as the search term.

What makes BD a unique possibility for organizations is its broad application possibilities. Examples of BD application fields are business insights (Palanimalai & Paramasivam, 2016; Strauß, 2015) e.g. marketing and business strategy building, health-care (Batarseh & Latif, 2016; Kościelniek & Pluto, 2015; Strauß, 2015), operation management (Basha et al., 2019), biotechnology (Kościelniek & Pluto, 2015), IT (Kościelniek & Pluto, 2015), market trend prediction (Kościelniek & Pluto, 2015; Hariri, Frederics & Bowers, 2019; Strauß, 2015), and fraud detection (Strauß, 2015; Balachandra & Prasad, 2017). Due to a wide spectrum of applications, BD has also completely transformed the analytics market. Big Data Analytics (BDA) possess the ability to deliver faster and better decisions, which is a key motivator for BDA adoption (Janssen, van der Voort & Wahyudi, 2017). Thus, accurate, timely, and better decision-making through BD has become a requirement in today's business world (Delen & Demirkan, 2013). De Mauro, Greco, and Grimaldi (2015) even predict that even though there already exist many BD applications, they are expected to grow. All in all, BD might be the most important so-called "tech disruption" since the internet (Agarwal & Dhar, 2014).

However, BD and BDA adoption introduces many challenges for businesses to address:

- *Data challenges* like BDA platform performance and scalability (Garg, Singla & Jangra, 2016; Sivarajah, Kamal, Irani & Weerakkody, 2017; Ali, Gupta, Nayak & Lenka, 2016), and massive magnitude of data and its heterogeneity (Labrinidis & Jagadish, 2012; Bertino, 2013; Zhong et al., 2016).
- *Process challenges* that deal with data processing issues like capturing and analyzing the data (Janssen et al., 2017; Sivarajah et al., 2017; Zicari, 2014).
- *Management challenges* like leadership, talent management, and decision-making (McAfee, Brynjolfsson, Davenport, Patil & Barton, 2012; Shamim, Zeng, Shariq & Khan, 2019).
- *Security and privacy issues* (Garg et al., 2016; Kuner, Care, Millard & Svantesson, 2012; Latif et al., 2019; Balachandran & Prasad, 2017) make it increasingly difficult for businesses to harness the full potential of BD.

Even though many theoretical challenges have been identified in the literature of the field, the key goal of this thesis is validating challenges such as described above on a practical level.

1.1 Motivation

Two key variables serve as motivation for this study: fast evolution of BD as a phenomenon, and a notable increase in research conducted on the subject. Together, they create a scattered understanding of the subject, of which this study seeks to clarify. Current studies often focus on one narrow field of BD application, or a specific perspective to the subject. This has created a very high number of studies on the topic, but with a lot of dispersion regarding definitions and conclusions. Also, as the field has evolved at a fast pace, some of the arguably fresh research – referring to studies published in the last decade – might already be dated. Thus, a status-check is in order.

Sheng, Amankwah-Amoah, and Wang (2019) present additional motivation by stating that "research is needed to advance further understanding and utilization of BDA in managerial applications". This study aims to provide the reader with an overview of BD definitions, applications, challenges, and related frameworks. In addition to the reasons presented above, BD is evolving further as we speak and is a present and relevant trend and interest for many organizations, thus its research is justified.

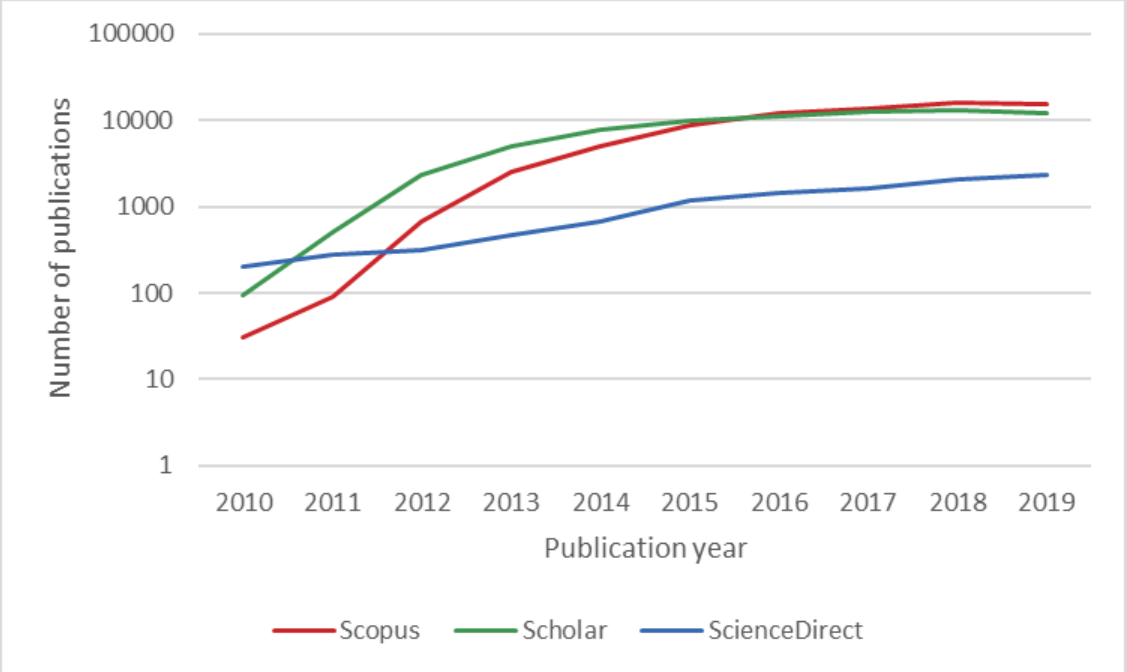

Figure 1: Logarithmic representation of yearly publications related to Big Data in Scopus, Google Scholar, and ScienceDirect databases.

1.2 Research questions

As can be drawn from the motivation, there is a lack of an all-inclusive definitive paradigm of BD-based decision-making and challenges related to it. To address this research gap, the following research questions were formulated:

- RQ1: How can Big Data be defined?
- RQ2: What are the most relevant challenges associated with Big Data-based decision-making identified in the literature?
- RQ3: Which of the challenges of RQ2 are the most relevant to the practitioners of the field?

To answer these research questions, a narrative literature review was carried out followed by a set of semi-structured interviews conducted to practitioners of the BDA field with varied backgrounds and professional experience.

1.3 Structure

The structure for the rest of this thesis is the following: After this introduction, the literature review is presented. This literature review includes utilized research methodologies, defining BD, and presenting key challenges related to BD-based decision-making. After the literature review, we present our methodology for the qualitative research section of the study. The methodology section includes presenting the chosen empirical methodologies, as well as the interview data-analysis methods. Next, the results of this study are presented. In the results section, the 16 themes identified in the semi-structured interviews are examined individually, and a set of challenges presented in the literature are validated. The final section of the thesis is reserved for discussion of the results of this study, and to provide a conclusion by answering our research questions presented earlier.

2 LITERATURE REVIEW

2.1 Methodology

A traditional narrative literature review was selected as a method for building the theoretical background of the thesis due to the multiple strengths of the method presented in the relevant literature. The main purpose of a literature review is provided by Baumeister and Leary (1997), who explain that a literature review's function is to serve as a link between the massive amount of printed knowledge of a given topic and the reader who doesn't have time to analyze all the available literature.

The term "narrative literature review" has been debated as an abstract term. Thus, to clarify, when referring to a narrative literature review in the context of this thesis, we refer to "comprehensive narrative synthesis of previously published information", as defined by Green, Johnson, and Adams (2006) in their highly cited paper of this topic.

To set certain standards to our literature review, we utilize Webster and Watson's (2002) criteria for ideal literature review, which are the following:

- The research topic is motivated
- Key terms are defined
- The research topic is appropriately confined
- The study analyses relevant literature
- Implications drawn from the literature review are justified with theoretical explanations and practical examples
- Useful implications for researchers are presented in the conclusion

The source material for the literature review was gathered by utilizing well known and comprehensive scientific databases of the field. The databases chosen for this thesis were ScienceDirect, Scopus, Web of Science, and Google Scholar. The initial search was conducted with the following query:

big data AND decision-making AND (challenge or threat)

Additional literature was searched using a slightly different query to consider the possibility that Big Data as a term might not be necessarily mentioned if the paper was about data-driven decision-making in general. The secondary query was conducted as follows:

data-driven decision-making AND (challenge OR threat)

To emphasize source material's relevance in the current day world, the results of the searches were limited to only include papers from the past five years (2015-2019). The results were sorted by their citation count, and relevant articles were selected for closer analysis by skimming through the articles' abstracts. ScienceDirect, Scopus, and Web of Science were the main source databases for the papers, whereas Google Scholar was mostly used to check for relevant articles that might have been missed in the search in prior databases mentioned above.

Further literature was found by utilizing backward reference searching, which means analyzing the originally selected articles' reference lists. The goal here was to identify possible pioneer studies that were excluded from the initial search due to the limitations set for the publishing year. The result of this source material-gathering method is a combination of articles from the past five years to provide current day knowledge, with a broad set of supporting pioneer studies of the field to confirm the information found from the fresher papers.

2.2 Defining Big Data and Big Data Analytics

Defining Big Data (BD) has always been a troublesome task. Firstly, the rapid evolvement of BD during the last decade makes coming up with a definitive definition challenging, especially as the definition should also stand the test of time. Secondly, because BD is not a single concept. It is rather a combination of multiple approaches that happen to share a name, since BD is such a broad construct. It can be seen from the product-oriented perspective as a complex, diverse, and distributed data sets (N.N.I Initiative, 2012), from the process-oriented perspective as a new tool for process-optimization (Kraska, 2013), from the cognition-based perspective as a concept that exceeds the capabilities of current technologies (Adrian, 2013), or from the social movement perspective as a new revolutionary approach that has the potential to completely change the field of organizational management practices (Ignatius, 2012).

Even though explicitly defining BD can be complicated, researchers have widely agreed on multiple variables to be associated with BD to better understand its attributes and dimensions. This frame of thought has been called the prism of Vs (Jabbar, Akhtar & Dani, 2019) since it has become standard to link words starting with letter V with BD. A set of most frequently used Vs has taken root in the literature, and these are volume, velocity, variety, veracity, value, variability, visualization, and volatility. The usage of Vs has evolved with the

phenomena of BD itself and new Vs find their way into the BD definition as more research is conducted. Table 1 displays the usage frequency of various Vs in the literature. Studies in the table were selected as they all provide their take on which Vs should be associated with BD and further because these studies cover a decent time frame – almost a decade – to easily compare the differences of which Vs have been used during certain time frames.

Table 1: Frequently used Vs for describing Big Data

Authors	Volume	Variety	Velocity	Veracity	Value	Variabil.	Visualiz.	Volatil.
Chen et al., 2012	x	x	x					
Bertino, 2013	x	x	x					
Borne, 2014	x	x	x	x	x	x		
Thabet & Soomro, 2015	x	x	x	x				x
Gandomi & Haider, 2015	x	x	x	x	x	x		
Ali et al., 2016	x	x	x	x	x			
Horita et al., 2017	x	x	x	x				
Sivarajah et al., 2017	x	x	x	x	x	x	x	
Basha et al., 2019	x	x	x	x	x		x	x
Hariri et al., 2019	x	x	x	x	x			

The table demonstrates very well – as it is sorted by publication year with oldest publications being on the top – how more Vs have been introduced to the field as years have passed. However, we can also see that newer Vs have a harder time taking root as a standard, thus they are more scattered across literature. In contrast to this, the initial Vs became an industry standard and have stayed as one. In further sections, we take a closer look and comprehensively define all the Vs mentioned above.

2.2.1 Big Data (BD) and Big Data Analytics (BDA)

Separating BD from BDA is a key construct in the field of data analytics, and critical dichotomy as we move forward in this thesis. As we go further in the defining of BD, we will learn that BD is a broad concept covering a multitude of different attributes and having a wide range of definitions. However, defining

BDA is considerably easier, yet as important. Akter and Wamba (2016) define BDA as a process, which involves the collection, analysis, usage, and interpretation of data, intending to gain insights and create business value, which in the end leads to competitive advantage. We can draw from this definition that BD itself is a mere object or resource and BDA is the tool that is used to turn that object into an advantage. A practical example would be that BD is the oil beneath the Earth's surface and BDA is the oil rig used to access it and the benefits that can be processed from that resource.

A wide variety of techniques are used in BDA and there are multiple outcomes of the usage of BDA. Sivarajah et al. (2017) group these outcomes to descriptive-, inquisitive-, predictive-, prescriptive-, and pre-emptive analysis. Descriptive analysis is used to examine and chart the current state of business (Joseph & Johnson, 2013). The inquisitive analysis uses the data for business case verification (Bihani & Patil, 2013), i.e. charting which business opportunities to chase based on a risk-reward analysis. Predictive analysis aims to forecast future trends and possibilities (Waller & Fawcett, 2013). Prescriptive analysis's purpose is to optimize business processes to, for instance, reduce variable costs (Joseph & Johnson, 2013). To highlight the difference between the latter two, predictive analysis helps organizations by providing decision-makers with possible future scenarios to consider, whereas prescriptive analysis provides concrete steps to achieve the desired outcome. And finally, pre-emptive analysis is used to determine what actions to take to prepare for undesirable future scenarios (Smith, Szongott, Henne & Von Voigt, 2012). Examples of BDA techniques are data mining, predictive modeling, simulation modeling, prescriptive methods, and business intelligence to name a few (Saggi & Jain, 2018). However, this thesis will not dive deeper into BDA methods and technologies, as they are out of the scope of this thesis.

2.2.2 Volume

The volume of big data refers to the massive magnitude, amount, or capacity of the data at hand for enterprises to analyze (Akter et al., 2019; Basha et al., 2019; Hariri, Fredericks & Bowers, 2019; Moorthy et al., 2015; Thabet & Soomro, 2015). The pure volume of data available on the current day world – as described in the introduction – is the attribute that arguably created the term BD. Though there is not a concrete standard of what volume of data counts as BD, Bertino (2013) argues that data sized ranging from terabytes to zettabytes refer to the volume attribute of Big Data. Volume can be seen as the fundamental essence of BD, as the sheer amount of data branches out to the other attributes of BD creating a multitude of other issues.

2.2.3 Variety

Data variety refers to the fact that BD is often captured through multiple different channels, which leads to data being in numerous different formats within a

BD database (Basha et al., 2019; Moorthy et al., 2015). Different data formats are commonly defined as structured-, unstructured-, and semi-structured data (Bertino, 2013; Garg, Singla & Jangra, 2016; Hariri et al., 2019).

Structured data, in this case, refers to data that can be captured, organized, and queried relatively easily (Philips-Wren, Iyer, Kulkarni & Ariyachandra, 2015), and has a clear, defined format (Garg et al., 2016). Examples of structured data are names, dates, addresses, credit card numbers, etc. Semi-structured data, on the other hand, lacks the standardized structure associated with structured data but has features that can be identified and categorized (Philips-Wren et al., 2015) by, for instance, separating data elements with tags (Hariri et al., 2019). Examples of semi-structured data are emails, HTML, and NoSQL databases. Finally, unstructured data is poorly defined and variable data (Akter et al., 2019). Unstructured data cannot be processed with structured data since the data does not fit in pre-defined data models (Casado & Younas, 2014). Data such as audio files, images, videos, metadata ("data about when and where and how the underlying information was generated" (Kuner et al., 2012)), and social media data can be categorized as unstructured. From the categories above, most of the data collected by organizations is unstructured (Bhimani, 2015). For example, Facebook processes 600 TB of data every day, and 80% of all this data is unstructured (Garg et al., 2016).

2.2.4 Velocity

The velocity attribute covers two aspects. Firstly, it refers to the pace at which data is generated, or the rate at which the data grows (Akter et al., 2019; Basha et al., 2019; Moorthy et al., 2015). And secondly, to the organization's capacity and capability to process the generated data with minimal delay (Thabet & Soomro, 2015; Chen, Mao & Liu, 2014). As the data streams today are high in velocity, this results in continuous data streams and makes it critical for enterprises to analyze and act upon this data as fast as possible (Bertino, 2013). Since data, in general, has a short shelf life (Thabet & Soomro, 2015), the faster new data is generated, the faster old data becomes less relevant and possibly flawed.

Garg et al. (2015) state that real-time analysis of data is a requirement for extracting business value out of it. They also argue that the speed at which an organization can analyze data correlates with greater profits for the said organization (Garg et al., 2015). Sivarajah, Kamal, Irani, and Weerakkody (2017) closely associate velocity with variety by explaining that the high rate of data generation is heterogeneous in structure. What this means in practice, is that the faster data is generated the faster more heterogeneous data should be analyzed, which has been deemed challenging.

2.2.5 Veracity

As the volume, variety, and velocity above mostly describe properties or attributes of BD, veracity deals with the underlying nature of the data. It refers to the

uncertainties, unreliability, noise, biases, quality, authenticity, trustworthiness, and possibly missing values in a given data set (Akter et al., 2019; Basha et al., 2019; Moorthy et al., 2015; Thabet & Soomro, 2015). This makes veracity a critically important aspect of BD to consider, as Garg et al. (2016) describe by stating that data should be reliable and clean for it to be useful.

Data veracity is categorized into three categories: good, bad, and unidentified (Hariri et al., 2019). On a general level, good veracity of data means its trustworthiness can and has been verified, bad veracity refers to certainly unreliable, noisy, or biased data, and unidentified veracity means a data set's trustworthiness is yet to be determined. Veracity is a relevant topic in any data analytic context but is greatly highlighted in Big Data Analytics (BDA), as verified by Sivarajah et al. (2017), who explain that veracity is caused by complex data structures and imprecisions in large data sets. Two aspects that are highly present when dealing with BD. For instance, in a practical setting traditional data sets might not have any veracity at all if the data set's size is manageable and it is logically structured throughout. Even if some veracity exists, the verification process in the traditional data set context is not that labor-intensive. In the BD context, the large data sets and complicated data structures are, by definition, present from the start of the process. The data verification process is extremely labor-intensive and to some degree uncertain due to the massive size of BD sets.

2.2.6 Value

Value in the context of BD has two distinct characteristics. On one hand, it refers to the economic business value that can be extracted from processed data and its usefulness for decision-making (Akter et al., 2019; Hariri et al., 2019; Moorthy et al., 2015). On the other hand, the value of BD is the high value of the data itself (Basha et al., 2019). Two examples to clarify this dichotomy: organization can extract value from BD by processing it and transforming it into business insight. In this case, the value of BD refers to the economic value extracted from it. We can compare this to the second kind of value, which would be the case where an organization possesses highly valuable data that it can sell to third parties interested in the data. The second case would represent the high value of data itself.

On a more practical level, we can compare an organization having a business strategy based on business insights gained from BDA to social media giants like Facebook that control massive amounts of user data that are sold to advertisers. The second aspect of BD value – the possession of highly valuable data – is often overlooked in the literature, in which it is often stated that BD value is gained by improving decision-making quality (Janssen, van der Voort & Wahyudi, 2017; Economist Intelligence Unit, 2012). Value is also highly susceptible to human interference, as the analysis of BD is open to human interpretation, thus the analysis generates little to no value if the end-users of the analytics process cannot understand it (Labrinidis & Jagadish, 2012). This is also verified by Thabet and Soomro (2015), who state that analysis has very limited

value if it is not understood by the decision-makers. In practice, no decision-maker can make good decisions by just looking at a set of numbers or a graph on the screen. The context of said numbers or visual representations has to be understood by the decision-maker.

2.2.7 Variability

Variability refers to the fact that data's meaning can change frequently (Sivarajah et al., 2017; Moorthy et al., 2015). The context of the data plays a critical role in the analysis process of data, as it can considerably change the meaning of said data (Sivarajah et al., 2017). In addition to the frequently changing meaning of data, variability also refers to the constantly changing flow of data (Gandomi & Haider, 2015). Critical aspects to consider when dealing with data variability, are how to verify the data context, and how prepared an organization is to data streams with altering velocity. As discussed in the velocity section, the organization's data processing speed should match the data flow velocity to consistently draw business value out of it. Variability in data flow does not only affect the data processing requirements, but also storage requirements. The organization's data storage should be able to handle constantly changing the velocity of data flow.

Data context becomes most relevant when conducting BDA in the context of natural languages. In every language, words do not necessarily have a static meaning. The analysis of word context is critical to draw relevant conclusions out of such data sets. For example, when analyzing natural language and algorithm runs into a homonym (a word that can have two or more different meanings), it has to understand the context to determine the word's meaning correctly. Otherwise, the meaning of the entire sentence, tweet, or message can change, which after many repetitions leads to faulty or noisy data with increased uncertainty.

2.2.8 Visualization

Visualization of BD deals with representing knowledge gained from BDA as effectively as possible, and in an understandable form (Basha et al., 2019; Sivarajah et al., 2017). The desired goal of visualization is to present data in an appropriate format and context to ensure that it is effortless for the target audience to consume it (Garg et al., 2016) and draw conclusions. Kościelniek and Puto (2015) see visualization as an essential function to obtain business benefits from BD.

Common techniques used in visualization are for example tables, histograms, flow charts, timelines, or Venn diagrams (Wang, Wang & Alexander, 2015). By successful visualization, it is possible to remove much of the data interpretation aspect, which can often impede decision-making. There are many BD visualization tools available in the market – each with distinct strengths and weaknesses – and one should be chosen for the data requirements at hand (Ali

Gupta, Nayak & Lenka, 2016) rather than seeking a one-size-fits-all solution. What makes visualization extremely important is that with effective visualization of a data set, managers or decision-makers can make more informed decisions. McAfee et al. (2012) state that "data-driven decisions are better decisions as they are decided based on evidence rather than intuition". Visualization is the aspect that enables decision-makers to make data-driven decisions.

2.2.9 Volatility

Volatility of BD defines how long the data is valid and thus, how long an organization should store it in their databases (Thabet & Soomro, 2015). Determining the volatility of a BD set is to determine a point of data from whereon it is no longer relevant for analysis (Basha et al., 2019). High volatility data's analytical usefulness is rather short, and low volatility data retains its analytical relevance for a longer period. For instance, data related to market trends can be considered highly volatile, as there is a possibility of a sudden shift in the market for example in a situation where a new technology is introduced that has the potential to revolutionize the field. On the other hand, geographical data like location data of tectonic plate borders is low volatility data, because even though the plates' locations are changing, the changes are most of the time slow and predictable. Earthquake prediction would be considerably more difficult if this kind of seismologic data were highly volatile. Table 2 summarizes the definitions of Vs associated with BD discussed above.

2.2.10 Additional definitions

As discussed in the first paragraph of chapter 2.2, the definition of BD can be viewed from multiple different perspectives. This means, that the prism of Vs approach is in no way the only way researchers have attempted to define BD.

De Mauro, Greco, and Grimaldi (2015) aimed to build an all-inclusive yet compact definition for BD. In doing so, they categorized BD definitions into three different categories: First category being describing BD through the prism of Vs discussed earlier. The second category focused on the technological requirements for BD processing, as Dumbill (2012) put it, data is big if it "exceeds the processing capacity of conventional database systems". The final category highlighted BD's impact on the societal level stating it to be a cultural, technological, and also a scholarly phenomenon (Boyd & Crawford, 2012).

By trying to combine aspects and nuances of all these three categories, they came up with the following definition: "Big Data represents the information assets characterized by such as high volume, velocity, and variety to require specific technology and analytical methods for its transformation into value". The catalyst behind this definition was that BD's evolution had been quick and disordered, which lead to a situation that universally accepted formal statement of its meaning did not exist (De Mauro et al., 2015). This is considered

to be the newest as well as the most comprehensive definition of BD extended by Latif et al. (2019), who defined BD as "advanced technology process that enables to store, capture, and process the large and complex data sets generated from various data sources".

Table 2: Summary of definitions of different Vs linked to Big Data

Attribute	Description	Associated literature
Volume	Pure magnitude of available data ranging from terabytes to zettabytes	Akter et al., 2019; Basha et al., 2019; Hariri, et al., 2019; Moorthy et al., 2015; Thabet & Soomro, 2015; Bertino, 2013
Variety	Data is captured from multiple sources and in multiple formats, specifically structured, unstructured, and semi-structured formats	Basha et al., 2019; Moorthy et al., 2015; Bertino, 2013; Garg, et al., 2016; Hariri et al., 2019; Philips-Wren et al., 2015; Akter et al., 2019; Casado & Yonas, 2014; Bhimani, 2015
Velocity	The speed of which new data is generated. Organizations' data processing speed must match with the generation speed to draw insights from the data	Akter et al., 2019; Basha et al., 2019; Moorthy et al., 2015; Thabet & Soomro, 2015; Chen et al., 2014; Bertino, 2013; Sivarajah et al., 2017; Garg et al., 2015
Veracity	Overall quality of data that manifests through noise, biases, trustworthiness, and missing values in a data set. Veracity is categorized as good, bad, or undefined	Akter et al., 2019; Basha et al., 2019; Thaber & Soomro, 2015; Hariri et al., 2019; Sivarajah et al., 2017
Value	The economic value that can be drawn from processing the data to improve decision-making, or high value of data set itself	Akter et al., 2019; Hariri et al., 2019; Moorthy et al., 2015; Basha et al., 2019; Janssen et al., 2017; Economist Intelligence Unit, 2012; Labrinidis & Jagadish, 2012; Thabet & Soomro, 2015
Variability	Changes in the meaning or context of data, or the data flow	Sivarajah et al., 2017; Moorthy et al., 2015; Gandomi & Haider, 2015
Visualization	Presentation of BD analysis in an effective and understandable format	Basha et al., 2019; Sivarajah et al., 2017; Garg et al., 2016; Kościelniek & Puto, 2015; Wang et al., 2015; Ali et al., 2016
Volatility	Determination of how long data is valid for analytic purposes	Thabet & Soomro, 2015; Basha et al., 2019

De Mauro et al.'s. definition was slightly altered by Moorthy et al. (2015), who state that "Big Data refers to information assets characterized by high volumes, velocity, variety, variability with veracity subjected to a specific technology and analytical methods for deriving value with virtue". They motivate this

definition by adding that volume alone is not capable of defining BD, and the analysis factor is a critical part of the equation (Moorthy et al., 2015).

All in all, even though a wide spectrum of definitions exists for BD, as it is a remarkably broad term, BD definition should be tied to the context in which it is discussed. The prism of Vs combined with the additional definitions presented here offers an adequate understanding of the concept itself but to fully understand the term in a given context, one should be able to apply their knowledge to the situation at hand. For example, if the issue and hand is a lacking technological infrastructure to process BD, it should not be viewed as a sociological construct in that context, but instead, the technological attributes of BD should be the main focus.

2.3 Big data decision-making challenges

To capitalize on the benefits of BD, organizations need to address a variety of challenges introduced by BD. The presence of these challenges can be seen from statistics as well. Around 80% of businesses have failed in the implementation of their BD strategies (Asay, 2017; Gartner, 2015). Also, over 65% of organizations report that they have experienced below-average returns from their investments to BD management (Baldwin, 2015). Ransbotham, Kiron, and Prentice (2016) say that "the percentage of companies that report obtaining a competitive advantage with analytics has declined significantly over the past two years". This implies that as BD has become more popular and available, more companies not capable of addressing BD challenges have yet attempted to adopt it into their business processes.

Challenges vary by type and there are many opinions of which ones are the most essential ones to tackle. Sivarajah et al. (2017) introduce a framework for categorizing BD-related challenges. They sort the challenges related to BD into three groups: data challenges, process challenges, and management challenges. This framework will serve as the foundation of this thesis's method of describing relevant challenges, though this framework will be slightly expanded to express the importance of security and privacy, as they have become increasingly highlighted in recent years due to media attention given to data breaches and insufficient security. In addition to media attention, security is difficult to categorize as being a purely managerial challenge it is described in some studies, as it requires attention throughout the whole process starting from the data itself. Finally, one single entity in the process cannot be named solely responsible for the security, as it is fundamentally more of a mindset that should be held by all included parties (from management to the operational employees) than a concrete function in the process. Additionally, visualization is also highlighted as an independent part of the expanded framework, as it has become a key variable in the studies of more recent years and the results indicate it might be more relevant than thought so far.

2.3.1 Data challenges

Data challenges represent the portion of challenges that are related to the fundamental nature of BD, meaning the V's (Sivarajah et al. 2017) presented in the first chapter of our literature review. In other words, what challenges the pure essence of BD brings in for organizations to consider.

The volume of the datasets is a challenge itself, as well as being a key factor that enables many of the other challenges to exist. The outright size of the data makes retrieving it, processing it, and inferring it challenging (Barnaghi, Sheth & Henson, 2013). Additionally, sheer volumes of data introduce challenges related to scalability and uncertainty (Hariri et al., 2019). Especially uncertainty due to data volume is a significant challenge to consider. Data is often analyzed with statistical methods, and when the volume of the dataset becomes great enough – like when dealing with BD sets – it can lead to weak signal analysis, which means overlooking statistically insignificant possibilities (Raikov, Avdeeva & Ermakov, 2016). These statistically insignificant possibilities, even though highly unlikely, can cause massive consequences if manifested. Strauß (2015) describes this as following: "So-called black swans [or the statistically insignificant possibilities] are exceptionally and highly improbable events, but they can have a particularly high impact". Volume's role as more of an enabling factor for other challenges is described by Bertino (2013), who notes that volume alone might be the least difficult problem to address when organizations are dealing with big data. This is further verified by Janssen et al. (2017), as according to them the other challenges of BD become more prominent due to the volume of the data. They also note that "main challenge found was not dealing with the volume but… dealing with variety, velocity, veracity, and validity of data" (Janssen et al., 2017).

As the datasets are already large, difficulties related to the variety of data are enhanced. As the data is not consistent but is gathered in a multitude of different formats and sources, it becomes very challenging to understand and manage this kind of data (Chen et al., 2012; Chen et al., 2013). Hariri et al. (2019) describe that analyzing unstructured and semi-structured data is challenging because the data comes from heterogeneous sources with many different data types and representations. The key attribute of unstructured data is that it requires major processing to be used in the analysis, which further requires adequate infrastructure to accomplish (Tabesh, Mousavidin & Hasani, 2019). This is expensive and hinders especially small enterprises' capability of utilizing BD. The need for modern infrastructure is highlighted by findings of Thabet and Soomro (2015) that point out that only 20% of data can be processed by traditional systems used for data analysis.

The velocity of the data presents its challenge, especially as the data processing speed of the organization should match the data generation speed. It is challenging to manage data that is generated with high velocity (Chen et al., 2013). The reason for this is that as the data should be processed as close to real-time as possible, only one section of the data is provided and this might give

different implications than when the whole dataset is examined (Janssen et al., 2017). It is also noted by Meredino et al. (2018) that there exists a clear mismatch between BD velocity and the capacity to respond quickly – meaning that currently, organizations are not able to process data in real-time – that further complicates the velocity aspect of BD. Sivarajah et al. (2017) also address this aspect by stating that the growth of data seems to out-speed the advancements made in computing infrastructures.

The veracity of the data accumulates pressure for data analysis accuracy. The biases, uncertainties, imprecision, noise, and general messiness creates a challenge of verifying the data for it to be precise enough to be used in analytics (Vasarhelyi, Kogan & Tuttle, 2015). Quality of data is a significant issue, as stated by Raikov, Avdeeva, and Ermakov (2016) who describe more than 40% of total data as being "dirty". This dirt can be human- or machine inducted (Raikov, Avdeeva & Ermakov, 2016). Human inducted dirt refers to data that was contaminated due to human action, whereas machine inducted dirt refers to data ruined by something else than human action. For instance, falsely tagging items in a dataset is human inducted dirt. A system failure leading to corrupt data, on the other hand, is machine inducted dirt. Janssen et al. (2017) describe the challenge related to noise in data is that the data is incorrectly connected, identities of persons are confused, a wrong place is mentioned, or some data from different periods are connected. According to Hamoudy (2014), some researchers have even stated veracity to be the greatest challenge related to BD. This might be due to the human factor being a key concept when addressing this challenge.

The context of the data becomes a challenge as organizations deal with data variability. As mentioned before, the context of the data can drastically change the meaning of it (Sivarajah et al., 2017), thus creating a challenge to build algorithms able to interpret data contexts. Janssen et al. (2017) state that the context of the collected data is often not known.

The value becomes increasingly more difficult to extract as the data sets grow in volumes. Data contains significant amounts of useless or irrelevant information, which makes it harder to extract the useful, beneficial, valuable, or "golden" information from the data (Zaslavsky, Perera & Gergakopoulos, 2013). Even if managing valuable information is achieved by an organization, it is extremely challenging to do it in a cost-efficient way (Abarwajy, 2015).

Visualization was categorized to be one of the key V-attributes of BD. But as research was conducted, it turned out to contain considerably more challenges than the rest of the Vs, as well as being more exposed to the human factors. Thus, a separate sub-chapter was decided to be created to address visualization challenges.

2.3.2 Data visualization

"Big data visualization method is concerned with the design of a graphical representation in the form of a table, images, diagrams, and spontaneous display

ways to understand the data" (Saggi & Jain, 2018). What makes visualization a relevant challenge aspect to examine individually, is that many challenges presented in this thesis are connected to it, and it provides additional challenges itself. Moreover, visualization is one of the key components associated with effective decision-making and human interpretation offers new dimensions to consider.

Ali et al. (2015) describe that when analyzing BD sets, interesting patterns can be found, but the result of such analysis is usually raw numbers regarding these patterns and thus, are difficult to interpret. They list visualization challenges as being visual noise, information loss, large image perception, high rate of image change, and serious performance requirements. They define the challenges as follows: visual noise describes the relativity of data sets. Different entities of a large data set are often difficult to separate. Information loss is closely connected to data latency, as the latency can be decreased by reducing the visibility in a data set, but this leads to information loss for the interpreter. Visual mechanical output can easily outclass physical perception capabilities, and this is called large image perception. The high rate of image change refers to data velocity, as if the refresh rate in a visualized image is too high, no decision-maker can react to these rapidly updating values. And finally, to represent visualization dynamically – as required in BD context – the performance requirements are considerably higher than in static visualization (Ali et al., 2016).

Visualization methods and technologies should also be designed in a way that the interpreter can interact with the data. This is important due to the frequent changes in the provided information and data sources (Horita et al., 2017). For instance, if a dynamic visualization of data is updated once per minute, the interpreter should be able to interact with the visualization to inspect the changed elements more closely. Otherwise, the benefit of the dynamic visualization is hindered. Ali et al. (2016) agree with this by declaring interactivity as being "the most important feature that visualization must have". Interactivity is not only a requirement that should exist, but the visualization system should also encourage it (Wang et al., 2015).

Chen & Zhang (2014) declare performance requirements as well as scalability and response time as being highly problematic when trying to visualize large data sets. All these aspects are highlighted by not only the volume of BD but by the presence of high amounts of unstructured data as well. Ali et al. (2016) also highlight this by specifying that "Big Data visualization tool must be able to deal with semi-structured and unstructured data".

Data visualization comes hand in hand with data interpretation. After the analysis is conducted and certain insight is extracted from data, this analysis needs to be interpreted by the decision-makers, which can lead to assumptions (Bertino, 2013) that increase uncertainty. "Knowledge is the ability to interpret data and information" (Ekambaram, Sørensen, Bull-Berg & Olsson, 2018). Thus, data should be visualized in a way that leaves little room for interpretation, or the decision-makers should be comfortable enough with data analytics in general that their interpretations are based on previous knowledge rather than as-

sumptions. Even with sufficient knowledge from the decision-makers' side, large, complex, and puzzling nature of BD sets to tackle the mental capacities of humans that make deciphering and interpreting such data increasingly challenging (Sammut & Sartawi, 2012). Strauß (2015) declares the correct interpretation of information provided by BD being a fundamental challenge. He continues by stating that "without interpretation of the data, the only valid fact about data is its existence, but the data itself does not reveal whether it is valid or true in a certain context or not" (Strauß, 2015). Managers play the leading role in the interpretation of data, but data scientists can help with the interpretation process by providing technical findings to the decision-making managers (Tabesh et al., 2019). These technical findings can be for instance information regarding the analysis process or insight of the data gathering methods. Thabet and Soomro (2015) agree with this as they declare that "it's not enough that the decision-makers see the data, they should also understand where the results came from".

Visualization is not only about presenting data efficiently. A key challenge is also to design a system that provides effective tools for data visualization. This kind of component is referred to as system visibility. Visibility measures the support provided for data visualization (Basha et al., 2019). Table 3 summarizes the challenges recognized in this chapter.

2.3.3 Process challenges

Process challenges refer to the challenges regarding the processing of the data, like capturing and analyzing it (Sivarajah et al. 2017; Thabet & Soomro, 2015). Another way to describe the process challenges is to formulate them as "how to" question, like "how to capture, integrate, process, and transform data" (Thabet & Soomro, 2015).

Defining what data an organization is interested in, how to filter out the irrelevant or uninteresting data, and generating and storing metadata has been deemed a challenge (Thabet & Soomro, 2015; Bertino, 2015). To gather information from a data set for analysis, data variety plays a critical role as most of the time the data is not in the format required for processing, thus there needs to exist a process to extract the data and to transform it into a format ready for analysis (Thabet & Soomro, 2015; Bertino, 2015). Additionally, if there exists uncertain data (possibly noisy data with incorrect information) within the data set, it needs to be verified (Bertino, 2015). As Garg et al. (2016) state, "if data is not proper or accurate then it will affect the decision-making capabilities of an organization".

When designing methods for data analysis, multiple requirements have to be considered. Bertino (2015) points out that these methods must be able to address heterogeneous, noisy, and dynamic data, as well as the complex relationship within the data. He also states that these method requirements can only be achieved with scalable data mining algorithms and powerful computational infrastructure (Bertino, 2015).

Wang et al. (2016) present two key strategies for the data analysis process. The scientific strategy "investigates natural phenomena, acquires new knowledge, integrates and/or corrects the existing knowledge and interprets the laws of nature from the obtained multiple sources of data". And the engineering strategy or decision informatics "pays more attention to the requirement of real-time decision-making in the presence of Big Data. It is supported by information technologies and decision science, and underpinned by data fusion/analysis, decision modeling, and systems engineering" (Wang et al. 2016). Two main approaches being present for analytics creates a challenge to choose the appropriate one for each situation.

Process challenges also extend to the platform used for BD processing. Basha et al. (2019) list challenges for the BDA platform: scalability, reliability, fault tolerance, data latency, and analytics. Scalability measures a system's capability to deal with a growing workload. Reliability is "a measure of the user to show the degree of dependency on data". Fault tolerance refers to a system's capability of functioning even if individual components fail. Data latency means delays in the processing of data. Finally, analytics describes the system's support for the decision-making process based on a great volume of data. An efficient BDA platform should cover all these challenges, i.e. be highly scalable, reliable, and fault-tolerant, minimize the latency in the analysis process, and offer a high level of support for decision-making.

2.3.4 Management challenges

Management challenges address the managerial side of BD utilization (Zicari, 2014) and relate directly to BD decision-making quality (Shamim, Zeng, Shariq & Khan, 2019). Management challenges are also called non-technical challenges and defined as "challenges which are arisen by management problems of service suppliers and users, rather than technical challenges related to Big Data processing" (Wang et al., 2016).

McAfee et al. (2012) highlight management challenges in their paper. They agree that the technical challenges related to utilizing BD are real, but the managerial challenges outshine them. Leadership, talent management, decision-making, technology, and company/organizational cultures are the five managerial challenges mentioned in the paper (McAfee et al., 2012).

As McAfee et al. (2012) put it, leadership in the context of BD utilization means companies have a management team that sets clear goals, defines what success looks like, and asks the right question. They also highlight the fact that BD's power does not erase the need for human insight (McAfee et al., 2012). This point also ties into the talent management challenge. Shamim et al. (2019) agree with McAfee et al. (2012) by noting that assigned leadership should possess a clear vision and set goals. Managers should also adapt their leadership style based on the work environment and desired outcomes (House, 1971), which can prove to be challenging.

Talent management refers to the challenge of finding competent personnel with adequate knowledge, skills, and BD capabilities. Janssen et al. (2017) define these BDA capabilities as skills and processes used to transform data inputs into outputs of greater value. The traits required for competent personnel can be categorized into two groups, technological and methodological (Shamim et al., 2019). Technical traits represent the practical know-how of hired staff to transform data into business insight (Shamim et al., 2019), whereas methodological traits mean the ability to transform those business insights into organizational value (De Mauro et al., 2018). McAfee et al. (2012) point out that in addition to the fact that the personnel should be comfortable working with large data quantities, they should also speak the business language to participate in the decision-making process. Tabesh et al. (2019) state acquisition of BD know-how as being a significant challenge for organizations. This can also be recognized from the statistics of Boulton (2015) that point out that 66% of organizations are unable to successfully fill their data scientist positions with qualified applicants.

Decision-making refers to issues hindering an efficient and effective decision-making process. Raikov, Avdeeva, and Ermakov write that the mental image of a decision-maker is full of convictions, perceptual features, cost and practical rules, and individual features that affect his problem resolution. Lack of unified vision in decision-making or strategy also blocks the effective implementation of BD insights (Rogers & Meegan, 2007). This is due to the fact described by LaValle et al. (2011) that decision-makers often lack adequate understanding of BDA and its benefits or applications in business processes. Fundamental knowledge of management or decision-makers is essential for effective implementation of BD insights to business strategy (Ethiraj, Kale, Krishnan & Singh, 2005). Thus, the decision-maker must learn the basics of data analytics to be able to integrate BDA into decision-making (Tabesh et al., 2019). The additional challenge related to the process leading to decision-making is that decision-makers are often provided with useless or irrelevant information that still requires adequate knowledge for further processing (Horita et al., 2017), which hinders decision-making.

Technological challenge refers to the organization's technological competence to process and act upon BD. This technological competency in a fundamental aspect of utilizing BD for analytical purposes (Lawson et al., 2013). McAfee et al. (2012) describe technological challenges as following: "Big data decision-making requires the use of the most effective and cutting-edge technologies to collect, store, analyze and visualize data". These kinds of effective, cutting-edge technologies are often very expensive, and organizations might realize it to be difficult to find available competent personnel for the implementation process. Technological challenges can be seen as the most resource-heavy challenges to address, as they combine the need for money, personnel, and expertise.

Organizational culture is the "set of norms, values, attitudes, and patterns of behavior that defines the core organizational identity" (Denison, 1984). What

makes organizational culture one of the main challenges in BD management is that if something is not part of set organizational norms, employees will not regularly do so (McAfee et al., 2012). If an organization promotes cultural aspects of BD like knowledge exchange and data analytics being high on the list of executive interests, the organization's BD decision-making capabilities are enhanced (Shamim et al., 2019). Promoting this kind of organizational culture can be referred to as a data-driven culture. Gupta and George (2016) define data-driven culture as "the extent to which organizational members (including top-level executives, middle managers, and lower-level employees) make decisions based on the insights extracted from data". Tabesh et al. (2019) express that lack of data-driven culture defined above is one of the leading causes of failure in BD projects.

Shamim et al. (2019) studied managerial challenges' association with BD's decision-making capabilities and found out, that organizational culture has the strongest association, followed by talent management, leadership, and technology, respectively. They conclude by emphasizing the importance of addressing BD management challenges by stating that "firms cannot be successful just because they have access to good data, but they need leadership with clear vision, suitable talent management practices, and most importantly an organizational culture that facilitates the use of big data" (Shamim et al., 2019).

Finally, the governance of data continues to be a key challenge in managing data. Data governance essentially means the protocols and actions taken to ensure data security (Thabet & Soomro, 2015). BD is commonly filled with sensitive or personal information, which makes its governance a matter of significant importance (Thabet & Soomro, 2015). In practice, data governance is used to define who can access what information and when and from where. This is often referred to as access- or identity control. Data governance processes are also responsible for ensuring data quality (Janssen et al., 2017), which ties into the data- and managerial challenges discussed earlier. Creating effective data governance protocols is not the only challenge as Russom (2013) notes that a complete lack of governance is a common – and unarguably more critical – issue as well.

2.3.5 Security and privacy issues

"The need for security, privacy, and accuracy of data is felt more strongly than ever" (Latif et al., 2019). Data security generally covers two aspects: security, and privacy. Security and privacy are sometimes used interchangeably in the literature. However, they can and should be separated as terms. Herold (2002) provides an exhaustive dichotomy of the two by describing them as following: "One must implement security to ensure privacy and difference between security and privacy is that one must use security to obtain privacy". Latif et al. (2019) expand this by describing that security is a process that leads to a certain result, in this case, privacy. Security is the strategy enforced by the organization, and privacy is the end-result of said strategy (Latif et al., 2019).

Thabet & Soomro (2015) highlight the importance of security by describing that BD warehouses possess a chance for huge profits for criminals if attacked. This is also emphasized by the fact that BD is often geographically distributed, which makes it more vulnerable to attacks. They state that the security of BD does not fundamentally differ from general data security. However, the Vs of BD augment all the requirements for effective security management (Thabet & Soomro, 2015). Additionally, BD security is a subject of higher risk as all the sensitive information it contains makes it often not suitable for simple data transmission (Wang et al., 2016). Security also raises scalability issues, as the administrative tools used for security administration should be able to scale to BD magnitudes (Bertino, 2013). The volume also plays a critical role in the security aspect as described by Garg et al. (2016): "sensitive information poses a major threat… to secure all sensitive data because of this huge volume of data available".

Privacy does not only become a challenge as a result of inadequate security. It can also function as an individual challenge because of independent legislation and policies enforced throughout the world. Buhl, Röglinger, Moser & Heidemann (2013) describe privacy's role as BD challenge as following: "we additionally see a myriad of different legal privacy restrictions in different countries turning into one of Big Data's most serious challenges". Narayanan and Shmatikov (2008) explain that to ensure an individual's privacy, data sets should be effectively anonymized. However, they continue by expressing that due to the massive volume of BD, even effective anonymization does not guarantee the unidentifiability of an individual, as the massive amount of data enables complicated reverse processes in the form of deanonymization (Narayanan & Shmatikov, 2008).

2.3.6 Typology of BD decision-making challenges

A framework was created to further illustrate different challenges' roles during the life cycle of the BD-based decision-making process. This typology was built based on the researcher's vision of the practical relationships of the challenges presented in this thesis. The typology aims to offer managers and analysts a visual representation to determine which challenges most likely manifest in different sections of the decision-making process. The framework is presented in figure 2.

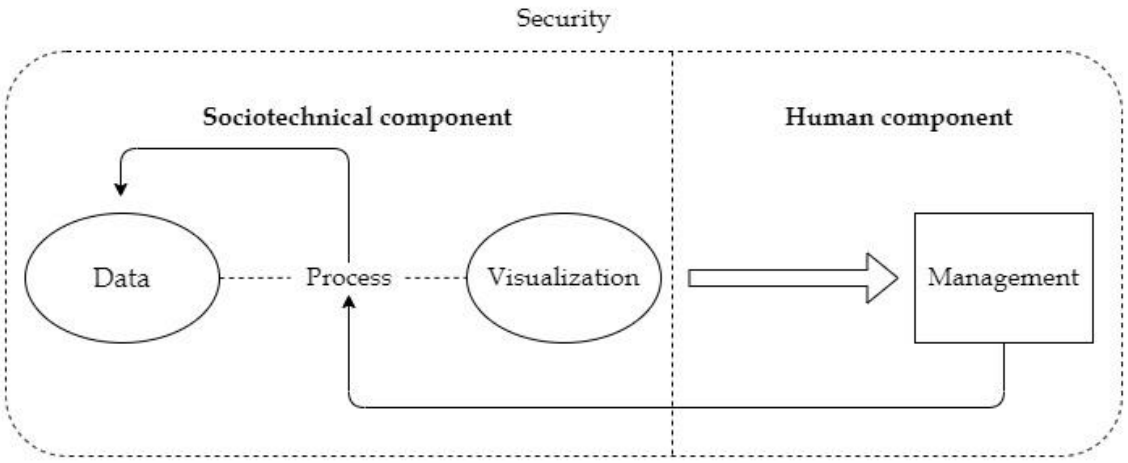

Figure 2: Typology of BD-based decision-making challenges

The main points of the framework (data, process, visualization, management, and security) are drawn from the challenge categories presented earlier and summarized in table 3. The model is divided into two main sections: sociotechnical- and human components. These components reflect the fundamental nature of the process within each sub-section. In other words, the sociotechnical component consists of process stages that utilize modern technologies supported by human expertise. Data, process, and visualization create the sociotechnical component. The human component, on the other hand, switches the focus from technology to human judgment. The human component contains the managerial section of the decision-making process, i.e. the decision-making itself. To differentiate the two, the sociotechnical component consists of clearly pre-defined methods and practices used by an organization to transform raw data into insight. The human component, on the other hand, is highly susceptible to human interpretation and thus different decision-makers might draw different conclusions from the same provided data set. What should be further noted is that in the sociotechnical component the technology is the enabler of the process whereas in the human component the technology is just a supporting function.

The decision-making lifecycle begins with raw data. Even though the process has only begun, all the data challenges are already present. Data must be processed before any further actions can be taken based on it. The goal of the processing is to transform the data into an understandable format, which in this case will be a visualization for the decision-makers. In this section, the process challenges present themselves. The last part of the sociotechnical component is visualization. It is the end-state of data processing. The data should now be in a visualized format, meaning that all relevant data is visualized interactively.

When the visualized data is transferred to the decision-makers, it has transitioned to the human component of the decision-making progress. Managerial challenges become increasingly relevant in this part of the process. Increasingly important, because as we can see from the figure, managerial challenges also branch out to the processing part of the lifecycle too. Talent management, or-

ganizational culture, and leadership playing the key roles here. The abovementioned branching out also applies to the process challenges. By definition, the process begins with the collection of data. This is acknowledged in the typology by expanding the process challenges to the data itself.

Security challenges are an aspect surrounding all the stages of the decision-making lifecycle. What this means in practice, is that rather than addressing security challenges in a specific part of the lifecycle, they should be considered throughout the whole process. This is because even if data is gathered according to all the highest security standards, all that effort can be undermined if the data processing is performed with lacking security. Thus, data security is more of a mindset than concrete action taken towards solving specific security challenges.

The common theme in the typology is that if the challenges are not acknowledged in the respective section of the lifecycle, they might manifest later making efficient execution of the rest of the process difficult. For example, if data veracity is not addressed in the data- and processing sections, it is very difficult to create an accurate visualization to draw insight from. Or an example from another perspective: even if data- and processing challenges are addressed and properly tackled, poor visualization can easily cripple all that work and lead to uneducated decisions made by managers. Even though mistakes in earlier stages might hurt performing the later stages, the typology also works the other way around, meaning that to a certain degree, mistakes in earlier phases can be fixed with success in the later stages. For instance, lacking visualization can be compensated with educated decision-makers, or poor data set to begin with can be undertaken with sophisticated processing and skilled personnel.

Table 3: Summary of Big Data decision-making challenges

Type	Sub-challenge	Challenge description	Literature
Data	Volume	A massive amount of data makes the processing challenging. Introduces scalability and uncertainty issues. Enhances the rest of the challenges.	Barnaghi et al., 2013; Hariri et al., 2019; Raikov et al., 2016; Strauß, 2015; Janssen et al., 2017
	Variety	The multitude of sources and formats make understanding and managing data difficult. Expensive infrastructure requirements.	Chen et al., 2012; Chen et al., 2013; Hariri et al., 2019; Tabesh et al., 2019; Thabet & Soomro; 2015
	Velocity	Organizations' ability to analyze data real-time is lacking. Growth of data outspeeds infrastructure.	Chen et al., 2013; Janssen et al., 2017; Meredino et al., 2018; Sivarajah et al., 2017
	Veracity	Data must be verified. Data quality is hard to ensure.	Vasarhelyi et al., 2015; Raikov et al., 2016; Janssen et al., 2017; Hamoudy, 2014
	Variability	Data context can drastically change its meaning and is often not known.	Sivarajah et al., 2017; Janssen et al., 2017
	Value	Significant amount of useless information. Valuable information is hard to extract.	Zaslavsky et al., 2013; Abarwajy, 2015

Process	Filtering	Defining interesting data. Generating metadata. Transforming data to processing-ready format.	Thabet et al., 2015;
	Processing	Designing adequate analysis methods. Scalability and infrastructure issues. Complex data relationships.	Bertino, 2015
	Uncertainty	Verifying uncertain data and ensuring data accuracy.	Bertino, 2015; Garg et al., 2016
	Process strategy	Multiple processing strategies available for organizations.	Wang et al., 2016
	Scalability	Measure of how capable the system is functioning under a growing workload.	Basha et al., 2019
	Reliability	The measure of how well the system is able to point out data dependencies.	Basha et al., 2019
	Fault tolerance	The system should continue to be operational even in a situation of some individual components failing.	Basha et al., 2019
	Latency	The system must minimize delays in data processing.	Basha et al., 2019
	Analysis	The measure of how well the system is able to provide support for decision-making.	Basha et al., 2019
Management	Leadership	Setting clear goals, defining success. Utilizing human insight. Adapting leadership style.	McAfee et al., 2012; Shamim et al., 2019; House, 1971
	Talent management	Finding competent personnel to deal with BD.	Janssen et al., 2017; Shamim et al., 2019; De Mauro et al., 2019; McAfee et al., 2012; Tabesh et al., 2019; Boulton, 2015
	Decision-making	Decision-makers' mental image affect problem resolution. Lack of unified decision-making vision. Lack of decision-makers' understanding of data analytics. Useless information reaching decision-makers.	Roger & Meegan, 2007; LaValle et al., 2011, Ethiraj et al., 2005; Tabesh et al., 2019; Horita et al., 2017
	Technology	Organization's technological competency is often lacking to utilize BD effectively.	McAfee et al., 2012
	Culture	Employees will not act on something that is not part of organizational norms. Promoting data-driven culture.	Denison, 1984; McAfee et al., 2012; Shamim et al., 2019; Gupta & George, 2016; Tabesh et al., 2019
	Governance	Creating sufficient security protocols. Inadequate or complete lack of governance.	Thabet & Soomro, 2015; Janssen et al., 2017; Russom, 2013
Security	Security	BD possesses huge profits for criminals if attacked. Securing sensitive information. Scalability and administrative issues.	Thabet & Soomro, 2015; Wang et al., 2016; Bertino, 2013; Garg et al., 2016
	Privacy	Addressing different legislations and policies. Deanonymization of sensitive information.	Buhl et al., 2013; Narayanan & Shmatikov, 2008

Visualization	Visual noise	Data sets are relative to each other and thus, hard to differentiate.	Ali et al., 2016
	Information loss	Reducing latency leads to information loss.	Ali et al., 2016
	Image perception	Mechanical output surpasses physical perception.	Ali et al., 2016
	Image change rate	High image change rate makes it difficult to react to changes.	Ali et al., 2016
	Performance and scalability	Dynamical visualization comes with high-performance requirements. Unstructured data is problematic	Ali et al., 2016; Chen & Zhang, 2014
	Interpretation	Decision-makers' assumptions. Humans' mental capabilities. Data origin must be understood.	Bertino, 2013; Ekambaram et al., 2018; Sammut & Sartawi, 2012; Strauß, 2015; Thabet & Soomro, 2015
	Visibility	The measure of how well the system is able to provide support for visualization.	Basha et al., 2019

3 METHODOLOGY

A qualitative methodology was selected as the research approach for this study. More specifically, a set of semi-structured interviews was performed with the interviewees being industry professionals of various backgrounds. The goal of the interviews was to gain practical insights regarding the research questions determined in the beginning of the thesis. The qualitative research methodology is well suited for studies such as this one, as described by Mason (2010), who states that qualitative interviews focus on the meaning of the interviews rather than creating [or testing] hypotheses.

Semi-structured interviews possess many strengths and possibilities that help to unveil the experiences and opinions of the interviewees. The mains ones being:

I. The interviewer can ask questions outside the interview guide
II. The interviewer can change the original questions in the interview guide
III. The interviewer can probe to a new path if an unexpected theme emerges during the interview

Semi-structured interviews were deemed the most suitable qualitative method for this study, as the topic at hand is somewhat abstract and thus, additional insight can be achieved if the interviewer is allowed to deviate from the original interview guide during the interview.

Determining the proper number of interviews to carry out can be problematic. For example, Galvin (2015) notes that no finite number is enough for interviews. This is due to underlying statistical mathematics that makes it impossible to ever reach 100% confidentiality. In this thesis, the question of how many interviews to conduct was approached through the saturation perspective. On the most basic level, reaching saturation represents the act determining the point after which additional interviews do not offer any further insight regarding the topic. At this point, the research is considered to be saturated. For example, if a researcher plans to perform 10 interviews regarding the perceived health benefits of fruits, and five out of the first six interviewees state that they feel more energetic the more fresh fruit they consume, the researcher might determine the research as being saturated as no additional themes seem to arise.

To determine the point in which saturation is reached beforehand, we turn to relevant studies of the subject. As this thesis mostly seeks to uncover the emerging themes regarding the challenges of BD-based decision-making rather than testing a concrete existing hypothesis, we aimed to reach a saturation level of around 70%. At this saturation level, we have covered most of the arising relevant themes of the subject and can draw implications to practice.

Many highly cited articles have tackled the issue of reaching adequate saturation levels with qualitative interviews. Morgan, Fischhoff, Bostrom, and Atman (2002) state that 5-6 qualitative interviews should be carried out to uncover most of the concepts. According to Guest, Bruce, and Johnson (2006), a 70% saturation level is reached with 6 qualitative interviews. Francis et al. (2010) also agree with this: their study shows that most themes can be identified in the 5-6 interview range. Finally, Namey, Guest, McKenna, and Chen (2016) reached 80% maturity with 8 interviews. Even though the final example might not be as highly cited as the others, it provides a similar frame of reference when aiming to determine the number of interviews to be performed.

To complement the goal of reaching an adequate saturation level, we should ensure we carry out enough interviews to gain relevant insight from practice. Galvin (2015) has created a formula to calculate how many qualitative interviews one should perform for them to gain insight into relevant themes of the whole population. The formula is the following:

$$R = 1 - \sqrt[n]{1 - P}$$

In the formula, R represents the probability that a theme present in proportion R of the whole population is also present in an interviewee. Moreover, n is the number of interviews performed, and P is the probability that a theme will emerge in the n number of interviews, i.e. the confidentiality level of our interviews. If we seek a confidentiality level of 95% and conduct 6 interviews, which is the amount noted to produce an adequate saturation level in the paragraph above, we will get an R-value of 0.39303. What this means in practice is that by conducting 6 qualitative interviews we can be 95% certain that themes that are present in ~40%+ of the population will come up. This is a very much acceptable level for us, as mentioned earlier, we seek to gain knowledge of the emerging themes regarding the challenges of BD-based decision making. Based on this, as by performing 6 interviews, we can be 95% certain we have uncovered themes shared by at least 40% of the population, the research goals are reached.

The interview preparation consisted of two phases, creating the interview guide, and selecting the interviewees. The interview guide was created to serve as the backbone of the interviews containing the initial questions to be asked from the interviewees. The detailed interview guide can be found in appendix 1. The interview guide was also created in Finnish as a portion of the interviewees

spoke Finnish as their first language. The Finnish interview guide can be found in appendix 2. The interviewees were selected with the following goals in mind:

I. The interviewees should represent a broad age range
II. The interviewees should represent multiple organizational hierarchy levels
III. The interviewees should come from more than one organization or department
IV. The interviewees should be competent in the field of data analytics
V. The interviewees should be familiar with data-driven decision making

The fulfillment of criteria I-III was observed by the author/interviewer and criteria IV-V was verified based on interviewees' subjective impression before the interview.

The results of the interviews were analyzed using thematic analysis. In practice, according to Braun and Clarke (2012) thematic analysis seeks to identify patterns in data set and turning those patterns into meaning. They explain the main benefits of thematic analysis to be accessibility and flexibility. In their work, a six-step approach to thematic analysis was introduced, and that approach will be applied to this research as well. The six steps are:

1. Familiarizing yourself with the data
2. Generating initial codes
3. Searching for themes
4. Reviewing potential themes
5. Defining and naming themes
6. Producing the report (Braun & Clarke, 2012)

This approach was followed in the practical coding process of the acquired interview material. The initial codes of step 2 were color-codes used to categorize various elements in the interview material: definitions, validations, exclusions, and quotes. The themes were selected by comparing the color-coded interview material to the decision-making challenges identified in the literature review. The interview results matched well with themes arisen in the literature review and thus, the potential themes were confirmed. The naming of the themes was conducted following the names of challenges in table 3 to create the most logical and easy-to-follow overall structure. The identified themes function as the crossheading in the results chapter.

Unfortunately, the ongoing coronavirus pandemic made arranging the interviews substantially more difficult than expected. In the end, there were five interviews conducted for industry professionals with varied ages, professional backgrounds, and organizational hierarchy levels. However, the author is confident that taking the circumstances into account, we reached an adequate saturation level with these five interviews. All interviewees currently work in an international multidisciplinary organization in various data analysis tasks. Details of the selected interviewees can be found in table 4.

Table 4: Summary of the qualitative study interviewees

Interviewee code	Current job title	Professional experience (y)	BD experience (y)
R1	Senior data analyst	5	2
R2	Data analytics specialist	7	7
R3	Data scientist	12	N/A
R4	Senior manager	22	2
R5	Data scientist	8	N/A

N/A coding in the BD experience column's two cells represents that the interviewee would not classify their experience actual BD experience, but rather experience working with generally large data masses. The interviews' average length was around 45 minutes, which means that for the results there was a little under four hours of interview material to analyze. In the next chapter, we will focus solely on the results of the interviews.

4 RESULTS

A total of 16 different themes were identified from the semi-structured interviews. These themes were BD definition, BDA definition, BD strengths, BD weaknesses, BD opportunities, BD threats, BD utilization in decision-making, utilization challenges, BD integration to decision-making, integration challenges, data challenges, process challenges, visualization challenges, management challenges, security challenges, and typology validation. The themes largely followed the topics discussed in the interviews. In this chapter, we go through the interview results arisen related to each theme individually.

4.1 Holistic view of the identified challenges

There was a clear deviation between the respondents regarding the current day's biggest BD challenge. Data, -availability, and -quality were the answers that arose in most interviews. Data the organizations already possess was deemed as often being flawed, data availability was seen as an issue if the organization does not possess it already, and data quality was noted to be difficult to verify. Though it was also noted that these are the current challenges because organizations are in the earlier phases of adopting BD into decision-making. The rest of the challenges would present themselves later. As one respondent described:

> *"At this point, definitely acquiring and understanding the data. The next part is how to process it. The following phase would be how to pass the information to the management through visualization, but I think that's more in the future for many organizations." -R1*

However, as there was a deviation, a few respondents did not see data or its availability to be an issue at all. The biggest challenge was viewed to be using the data and getting it modeled the right way. The challenges were noted to be highlighted in the human aspect of the process because the rest of the challeng-

es can be tackled simply with enough capital. Sub-chapters 4.2-4.17 present each identified theme separately.

4.2 Big Data definition

The respondents unanimously approached the definition of BD from a very practical point of view. The key approaches to defining BD were through used tools, number of lines, data sources, and data structure. The respondents agreed that the data size requirements for the data to be big is easier to define through tools than a concrete number of lines. For instance, multiple respondents described that it is hard to say if BD should be millions or tens of millions of lines, but in practice, it comes down to the fact that it cannot be opened using a single machine. The need for specific technologies designed to handle BD came up. Data processing perspective was also discussed by the respondents. BD was described to be difficult to handle due to its vast size, and the analytics were said to take a lot of time.

Regarding the data sources, some respondents stated that for the data to be BD, it should be gathered from various sources. Additionally, one respondent adduced that not only should the data be from multiple sources, it should also be created by multiple different actors.

The structure of the data was discussed by most of the respondents. The key findings being that the biggest difference between BD and traditional data is that with BD the data is no longer structured. The lack of data structure also adds requirements and challenges to the processing of the data. However, even though data structure was an aspect mentioned by multiple respondents, it was also noted that data structure challenges are not a requirement for the data to be big. Instead, sometimes there might be just a massive amount of very structured data that is considerably easy to process, but it is still BD. As one respondent put it:

> *"Sometimes the data might be big in size but very uniformed and structured and then it's so easy to move it around and processing it doesn't take much time at all." -R5*

4.3 Big Data Analytics definition

In line with the definition of BD itself, the definition of BDA was also approached from a practical point of view. The definitions were split to process- and technology-centered points of view. By the process centered view, BDA was described as refining and making BD understandable – creating something from BD. Additionally, it was said that the refining of BD is done to find connections within it. One respondent summed BDA up as being all the operations

and manipulations performed using BD. On the other hand, the respondent adopting the technological point of view portrayed BDA as being something that cannot be performed in a reasonable amount of time utilizing a single computer and requires some kind of parallel processing to be effective.

> *"However, when it comes to BDA it starts to require cloud-based parallel processing - like [Apache] Spark or equivalent - so that the calculation can be distributed, and you can no longer do that with one computer given that you would like to it in a reasonable amount of time." -R1*

4.4 Big Data strengths

Identified strengths of BD came largely down to the vast size of the data set. BD was deemed to contain a lot more information compared to traditional data. The increased amount of information was said to also complement to more diverse information. With increasing size and diversity of information, respondents agreed that it is possible to draw more accurate conclusions from BD thus, making BD more preferable compared to traditional data.

The ability to find patterns more easily from BD was identified to be a basis for more accurate conclusions. Interviews pointed out that with BD's large sampling, everything is closer to a normal distribution, which makes it easier to identify patterns. Also, one strength of BD that came up in more than one interview, was that BD enables wider opportunities to explore the data. This is because BD was noted to contain a lot of data to spare compared to traditional data, where you have to utilize all the data available for analysis. This gives an opportunity to play around with the dimensions of the data, explore things more profoundly, and in the end find the most important aspects of everything within the data set. One respondent's description of the dimensions of BD:

> *"Getting high accuracy measures is easier with Big Data, because you have many things to explore, many dimensions to explore." -R3*

4.5 Big Data weaknesses

The key weaknesses of BD identified in the interviews can be associated with data size, quality, variety, transparency, and warehousing. Data size was deemed to increase the time and computational requirements to process the data, as well as demand a lot of infrastructural know-how. Expertise and processing power were identified to be the basic requirements to get any relevant insight out of the data. Data quality was described to be an underlying weakness, as it can sometimes be difficult to verify. Additionally, data variety was

noted to add considerably more stress to the process, as it takes significantly more time to get varied data to processing ready format.

Transparency of the data was identified to be a side effect of the massive size of the data. When analyzing BD, it was said to be easy to miss smaller patterns that might not be relevant for the analysis on their own but can be a key aspect to consider regarding specific cases. Also, there were deemed to be business opportunities hidden in the data that are missed, because the focus is diverted to the big patterns. The transparency issues were also identified to consider the analysis process, as it can be difficult to track what data is being used, which poses a significant risk.

Warehousing was described to become an increasing issue in the future containing a multitude of aspects to consider. Where to put the data, the server, the database, geographical allocation, access control, and processing tools were identified as relevant issues to consider in data warehousing. One interviewee summarized BD weaknesses in the context of traditional data weaknesses:

> *"It might be that when the data is bigger, all your problems are bigger too." -R5*

4.6 Big Data opportunities

Many future opportunities for BD were discussed in the interviews. The well-known metaphor that data is the new oil came up. The more you have it, the more material you have to explore and gain value from. BD was said to enable organizations to better understand their activities, as well as their clients, leading to business insights concerning what the client wants. The real-time analysis aspect was discussed, and it was deemed to produce more accurate results in the future, as the conclusions are based on real-time information rather than investigating what has happened in the past.

One respondent mentioned that there already exists uncapitalized opportunities with BD, as huge amounts of data are gathered but never used. The data might be gathered and not even meant to be used in the analysis, but it still contains a lot of information. Opportunities were also identified regarding the amount of gathered data, which was said to increase in the future covering all aspects of life. One respondent described BD's opportunities as following:

> *"There are possibilities that you are not even aware right know and they will present themselves in the future as we explore the data and see the opportunities that we can identify." -R3*

4.7 Big Data threats

The threats of BD discussed in the interviews focused on privacy, security, and ethical issues for the most part. With data sizes continuously increasing, questions of ethics and ownership of the data emerge. Misuses of data were identified as an increasing threat in the future. Privacy issues were said to be difficult to avoid as the gathered data is often about humans, their activities, and preferences. Also, in a globally functioning business field, continuous transfers of data were identified to possess information security risks.

In addition to the security issues, threats concerning data analysis and the BD market were discussed. Blindly trusting analysis results was identified to be a threat in the future, as the increasing sizes of data sets make it more and more difficult to backtrack the analysis pipeline and identify possible errors in the process. Also, one respondent described the over-analysis of data to be a clear threat, meaning that analysts try to forcefully find things in the data that are not there. One respondent described the BD market to be very saturated possessing a risk. They explained that small companies get acquired by larger companies making the data in the world clustered to a handful of large technology organizations. The respondent described saturated BD market as following:

> *"BD usually belongs to big companies. They can take over the whole business if they collect more data. And this is exactly what they are doing right now." -R3*

4.8 Big Data utilization in decision-making

When discussing BD's utilization in organizational decision-making, it was agreed upon that using data in the decision-making at all usually improves the quality of the information where the decision is based on, as the context for the decision is better understood if it's backed up with data. Multiple concrete actions on how to use BD in decision-making were identified. For example, tracking people's clicking patterns of the organization's website was said to be a method that gives more accurate information on users' activities. Advanced analytics was another identified method, where the goal is to predict different trends, price development, market growth, market changes, and to compare it to the increasing amount of market data to be able to make predictions and adapt the business in a changing market environment.

It was also stated that the method itself is not as relevant as the goal: to find new information that is not traditionally available and getting that information to the decision-makers. And internal goal for BD usage was also identified: how the organization functions with the data and how different matters influence each other within the organization.

Some limitations for the utilization were also discussed. To gain benefits from BD analytics, the process should be thought of as a part of the organization's strategy and the organization should know what they want to analyze. Also, the organization should not necessarily base their decisions on data alone. As one respondent described it:

> *"I don't see decisions being fully made based on data during the next few years, as it's hard to trust in and people usually want a human to run an organization rather than a bot." -R2*

4.9 Utilization challenges

Utilization of BD in decision-making was identified to contain a wide spectrum of challenges. The ones that came up most often were communication with decision-makers, understanding the data and its usages, organizational competence, and privacy issues.

Communication with decision-makers was identified as a key utilization challenge. The respondents deemed it difficult to, first of all, get the information from the analysis to the decision-makers, and to get them to listen to it. It was highlighted in the interviews that the decision-makers might often be used to simpler forms of data analysis, for example, Excel reports, and therefore might not be aware of something that is uncovered from BD analysis. Thus, it is difficult to make the data seem trustworthy to the decision-makers. BD analysis often contains new information for the target organization, so demonstrating the value of that new information and providing transparency regarding the analytics process was identified as a challenge.

It was noted to already be difficult to understand what data from a single source tells and why that it. When combining data from multiple sources in BDA, it was said to become even more challenging to define correct key performance indicators (KPI) to understand how the data works, and how it can be transformed into decisions. Additionally, when combining data from multiple different sources, they are not designed to work together, which increases the stress for the analytics process. It was also stated regarding data usage and understanding that many organizations possess a lot of data, but do not know what to do with it. Determining what to do with the data and forming the questions the organization wants answered was described as a utilization challenge. This is due to the current day situation, where data is easy to collect but hard to apply to business.

Data usage and -understanding challenges were identified to be closely linked to organizational competence. It was said that gathering, understanding, and applying it to the decision-making context requires a lot of professional competence which many organizations do not have. Organizations were also said to lack competence in the ability to identify the data already within the organization. Data is not identified as an asset because proper information map-

pings are not conducted. Privacy issues were described to be an underlying aspect complicating the utilization process further due to multiple laws, regulations, and perspectives it brings to the table. One respondent narrated current-day organizations' issues as following:

> *"Many organizations don't consider data as an asset, which leads to lacking resources allocated towards managing it. Data should be identified as an asset, and after that, a management template should be created for it."-R4*

4.10 Big Data integration to decision-making

BD integration to decision-making was deemed to be an issue at a very practical level. It was said that we are still very far from utilizing BD full-time, especially in the context of larger companies. The situation is a little bit better in more flexible start-ups. The key reason for this is the challenges in the integration process. A very practical challenge is to assign IT infrastructure and tools in a way that it is possible to process the number of items required daily. This kind of infrastructure was noted to be expensive and the integration to require very specific competence. Additionally, the integration process was said to easily take years before the first application of the methodology or process is ready. One respondent described the integration process as following:

> *"It [the integration] changes the organizational structure, the IT-structure, it changes how people are working and used to working, it changes the decision-making processes, and the results can be contradictory to the experiences of the people." -R3*

4.11 Integration challenges

The integration challenges identified in the interviews were change management, technological challenges, organizational challenges, integration process design, and change management.

The technological challenges consisted of IT infrastructure challenges, and challenges concerning the integration of multiple data sources and systems together and were identified to be the easiest challenge to resolve. This is because organizations only have to get their hands to more equipment and people who know how to work with IT. However, resolving this challenge does require a notable amount of capital.

Organizational challenges were said to mean challenges regarding the personnel in the process and resolving them was deemed to be significantly more difficult than in the case of the technological challenges. What makes resolving organizational challenges more demanding is that it takes just one

manager who does not believe that implementing a new decision-making method is necessary.

The integration process design was said to be challenging, because it contains multiple variables to be considered, and all the concrete mechanisms for the integration should be carefully planned to achieve desired results. The change management challenges were described to consist of resistance to change and communication issues. Regarding the communication, communicating the actual benefits of tracking something that has not been tracked before was identified as a challenge. Concerning the change management, one respondent described it as following:

> *"It's really difficult for people working in that company to just accept that they have to change their way of working or make a decision where they don't fully understand why that decision is made." -R3*

4.12 Data challenges

The key data challenges identified in the interviews were data availability, quality, and relevance. These issues were summarized as to where to get the information, how high-quality it is, are there any mistakes, how much information is missing, and how much you have to patch up the data yourself. It was noted that data quality has a direct correlation to the data's usability. Aspects like noise, data gaps, and incorrect data were described to be the most prevalent issues affecting data quality. Especially when discussing the gaps in the data, it came up that bad parts, or the gaps, in the data have to be filled or removed, which means you are altering the outcomes of the data either by adding assumptions or removing data. It was however also noted that the data does not have to be all-around perfect for it to be usable in analysis. The issue with data relevance was described to be associated with the organization's needs: the gathered data has to be relevant to the organization's specific needs at hand and identifying these needs might be complicated. One respondent described the practical data challenges as following:

> *"Sometimes we get the wrong data, sometimes we get too little data, sometimes we believe we have the right data but later realize it's not, we might realize some pattern and need another set of data from another component of their system." -R5*

4.13 Process challenges

Key challenges identified with data processing were cooperation, business-IT alignment, manual operations, and transparency. Cooperation was noted to be

a key enabler for the final product to be correct. Building the process to support and encourage cooperation was deemed challenging. Building the process was said to also cover the infrastructure side, i.e. how to get the infrastructure to support the process so the data can be used.

Business-IT alignment means converting business needs and technical implementation together. It requires proper models and algorithms that were noted to be challenging to design. Also, forming the questions correctly to take both business and technical sides into account was discussed as a noteworthy issue. Business-IT alignment also requires a certain type of translation between the two sides, which requires specific competence from the personnel in the analytics process.

Manual processes were said to be a process challenge, because there are a lot of them in the analytics pipeline, which is time consuming even with modern tools. As the manual processes take time, data latency was discussed to become an issue as the data should be processed as close to real-time as possible. Additionally, multiple manual phases significantly increase the risk of mistakes in the process. Transparency was linked to the risk of mistakes: if a mistake is made in the analytics process, it might be hard to pinpoint where the mistake occurred, why it occurred, and what are the results, let alone just noticing that it has happened. One respondent illustrated how one mistake can undermine the whole data analytics process:

> *"If someone messes up the average and the median, that can change everything. Just because someone mistakenly used average instead of median. If I want to summarize." -R3*

4.14 Visualization challenges

Visualization challenges identified were visualization scope, choosing metrics, visualization design, and visualization interaction. Visualization scope issue was described a determining the essential aspect to be visualized: whether to visualize the analysis itself or the whole data mass. There is a clear tradeoff as visualizing only the analysis results can leave important information about the context out of the visualization scope, but on the other hand, visualizing the whole data mass is difficult to perform effectively.

The interaction with the visualization was another challenge brought up in multiple interviews. It was said that the user should be able to dig into the data and interact with it. However, visualizing large data masses in a way that it can be effectively interacted with was noted to be difficult.

Choosing correct metrics was identified to be closely connected to the visualization design process. The visualization should be designed in a way that it answers exactly the question the user is looking to answer. Not only should the answer be found in the visualization, it was discussed that the visualization should be in a format that the decision-maker can make decisions based on it.

Choosing the correct metrics was noted to play a key role in the visualization design process, as incorrect metrics can make it difficult for the user to identify relevant information. A respondent described the determination process of correct metrics as being an issue:

> *"It is possible to measure basically anything - you name it - but what is important and relevant and actually supports the decision-making is challenging to assess." -R2*

4.15 Management challenges

The most prevalent management challenges arisen in the interviews were communication, management attitudes, determining the analysis questions, and interpretation of the data.

Communication was deemed as a challenge present throughout the process. It was said that many decisions are often made in other parts of the process than actual decision-making, and effectively communicating them are important validation for data- and process quality and transparency.

Management attitudes that were discussed in the interview refer to a certain level of curiosity that decision-makers need to possess to effectively make decisions based on data. On a practical level, this meant that the decision-makers must have a personal interest in the data – where it is from and how it is processed – to be able to draw credible insight out of the data. Otherwise, some relevant information can be lost. This was also closely associated with another point of view that came up in the interviews: it is challenging to communicate the value of the data in decision-making and how it is created.

Determining correct questions was seen as a key component in the BD management, because if the management wants answers to the wrong questions, everything else will be done incorrectly. One respondent described the process of asking the wrong questions the following way:

> *"If they are asking something that is totally irrelevant to their processes or they are asking something beyond their capabilities or just plainly something that can't be answered at all." -R3*

It was said that humans and their experience and expertise cannot be fully replaced in the data and play a key role in the management of the data.

Finally, the interpretation of the data was deemed an important aspect to consider, because there is not always a right or wrong answer in the data, but instead, someone has to interpret it. It was noted that a lot of interpretation also exists in the data processing phase, and any interpretations made should be communicated to the decision-makers so that they are aware of the context of the produced visualization.

4.16 Security challenges

The security challenges identified can be divided into two categories: regulatory challenges and human challenges. Concerning regulatory challenges, there was noted to exist a lot of restrictions that make BDA increasingly challenging. A key example that arose in multiple interviews was the General Data Protection Regulation (GDRP). GDPR introduces multiple challenges to the data usage. Organizations might possess data but are unable to use it due to GDPR. GDPR also adds many restrictions on how to acquire the data, how to process it, and when the organization has to delete it. Of course, GDPR is not the only regulation affecting the data security, but it was the example most often used during the interviews.

Human challenges discussed by respondents were process design, access management, analytics team composition, and common practices. It was said that the analytics process is always designed by humans. And the process should be designed carefully as all parts of the analytics pipeline require security. In addition to current day security, the process should also consider the future: how to prepare for future security breaches. As the process is designed by humans, there is a risk of design flaws or that some aspects are not considered.

Access management was described as an essential security control for BDA. Access management was defined as managing who has the access – and in what scope – to the data and visualizations in different phases of the process. It was said that there should exist controls that prevent manual changes to the data after a certain process milestone to prevent misuse.

> *"The data should be protected in a way that the data, process, or visualization can't be changed manually before the decision-making. Identity- and access management are the key" -R4*

Regarding analytics team composition, a risk was identified that BD is often hosted in cloud services, but the analytics team more often than not does not include a security expert or cloud service specialist. This was deemed to lead to a situation where the analytics team utilizes cloud-based tools without fully understanding the security risks.

Common practices were said to include a challenge, as to how analysts use the data varies depending on the individual. Personal data can be shared just by speaking of it, or inadequate reporting might expose personal information. What this means that the challenge is to educate all members of the analytics team to be sufficiently competent with security-related matters that the risk of security mistakes is minimized.

Table 5 below displays the challenges identified in the literature review sections and connects them with quotes acquired from the semi-structured interviews to show which challenges' practical relevance was validated.

Table 5: Validation quotes for identified challenges

Type	Sub-challenge	Validation quote	Synthesis
Data	Volume	You are missing insignificant – or smaller – patterns that are not visible through the whole data set but might be significant for specific cases.	As BDA is focused on finding patterns in large data sets, statistically insignificant aspects often get overlooked, even though they might contain important information.
	Variety	The variety of the data [is probably the greatest weakness]. It requires a considerable amount of time to normalize and get to an analysis-ready format.	Data that comes from multiple different sources in multiple different formats forces a lot of manual labor to the analytics process.
	Velocity	[The goal is] to be able to conduct analysis based on real-time, rather than what has happened in the past.	More reliable and relevant analysis results are reached by analyzing data in real-time but the practical requirements for it are tough.
	Veracity	One [challenge] is obviously the size but it is not only that. It also matters whether the data is structured or unstructured. It makes a great difference in how to process the data.	Whether the data is structured or unstructured can drastically change the analytics process. Large data masses' analysis might be fairly straightforward if the data is in a structured format, and unstructured data on the other hand can make smaller data sets far more complex.
	Variability	In addition to what the data shows, it has to be understood if there are quality issues.	The context of the data has to be documented and understood to fully grasp the meaning of the data
	Value	We have to strongly communicate the actual benefits of tracking something that has not been tracked before and why it should be tracked now.	In addition to the valuable data having to be identified from the data set, its value also has to be demonstrated to achieve transparency and justification for the analysis.
Process	Filtering	The bad parts of the data – incorrect data, missing data, etc. – you have to remove them or fill them in, which means you are altering the outcomes of the data either by adding assumptions or removing data.	At the same time, data filtering is a required step in the analysis pipeline, and an action that directly modifies the outcome of the process. Thus, filtering of the data should be carefully documented to later assess the analysis outcome based on the earlier filtering.

	Processing	Even with modern tools, the processing starts to take a while.	The technologies for BDA are constantly evolving but at this point an efficient analytics process is time-consuming.
	Uncertainty	Where the mistake occurred, why it occurred, and what are the end-results, let alone just noticing that it has happened.	Backtracking in a complex process introduces challenges to consider. The importance of documentation during the process is highlighted to ease identifying and tracking mistakes in the process.
	Process strategy	The business strategy should be tied to the data and map out how the strategy can be executed utilizing the data.	Business-IT alignment is necessary to achieve the desired business results from the BDA process.
	Scalability	Not validated through the interviews.	-
	Reliability	Not validated through the interviews.	-
	Fault tolerance	Not validated through the interviews.	-
	Latency	Manual operations wipe the real-time aspect, i.e. data latency becomes an issue.	Multiple manual phases in the BDA process introduce data latency, which directly interferes with the real-time data analysis goal.
	Analysis	Not validated through the interviews.	-
Management	Leadership	The person who uses the visualization has to possess a certain level of curiosity towards the data to be able to fully realize the information in it.	Managers' personal traits such as curiosity towards the subject play a key role in transforming the data results into desired and justified decisions.
	Talent management	It [BD utilization] requires a lot of professional competence which many organizations do not have.	Organizations have to invest in talent management to acquire necessary competence into the organization to fully take advantage of BDA.
	Decision-making	It is another issue to get that information to decision-makers and get them to actually listen to it.	In addition to producing the analysis, measures have to be taken to ensure the analysis results are heard on the management floor.
	Technology	With BD, there are issues from the beginning. Where to put it, the server, the database, geographical allocation, access control, processing	Many and varied technological aspects to consider require sufficient competence in the organization as well as capital to tackle the technology challenges in practice.

		tools, etc.	
	Culture	When it comes to individuals it is a bit more difficult for the people that are not working with data to accept that an algorithm can give a suggestion that is more useful than their own opinion.	Organization's culture should encourage data-driven decision-making to ensure minimal resistance from the employees when basing business decisions on data.
	Governance	Not validated through the interviews.	-
Security	Security	You also have to be aware of the possible breaches and possible restrictions that will come in the future so you are prepared to answer all the possible audit questions to ensure you use the data in a proper way.	Security aspects are a continuous matter to consider. It is not enough to fulfill today's security standards, but instead, organizations should also be ready for possible future regulations and security threats.
	Privacy	Are people's information processed and anonymized correctly or are actual people's information presented to management?	Data anonymization process should be carefully designed and controlled to ensure individuals' privacy and to minimize the possibility of deanonymization.
Visualization	Visual noise	How to get big data masses visible to the decision-makers so they can properly dig into it and absorb it [is challenging].	Big Data sets' massive size introduces lots of noise that can transfer to the visualization as well. Minimal visual noise to create an effective visualization is challenging to ensure.
	Information loss	Images and alike are difficult to bring forth in the visualization so it can become a little superficial what can be shown to the user.	Data sets can contain data that is difficult to present in the final visualization due to structural- or confidentiality issues. This directly leads to information loss and should be documented so that the context is clear to the decision-makers.
	Image perception	Whether to visualize the analysis or try to somehow visualize the whole data mass.	Visualizing only the results of the analysis leads to lacking context but trying to visualize the whole data mass might introduce issues regarding image perception, as such a large amount of information is presented.
	Image change rate	Not validated through the interviews.	-

Performance and scalability	Not validated through the interviews.	-
Interpretation	There always is not right and wrong answer in data, but instead, someone has to interpret it.	Interpretation of the data is a key variable that can fundamentally change the implications drawn from the data. Interpretation is also an abstract matter to consider, as it might be impossible to know which interpretation is correct in a given situation beforehand.
Visibility	Not validated through the interviews.	-

4.17 Typology validation

The typology presented in sub-chapter 2.3.6 was validated through the interviews. Most of the respondents agreed that on a high-level the typology describes the challenges of the BDA decision-making process very well. The critique received was about data's weight in the typology, people's roles in the pipeline, and the iterative nature of BDA processes.

It was said that data's weight in the pipeline should be increased, because it plays a much bigger role in the practical process and serves as a most basic building block for everything else. Additionally, it was mentioned that the data's multiple types should be somehow modeled. People's role in the pipeline was not deemed clear enough, and it should be somehow modeled that the analytics pipeline is created by humans – that there exists a process owner. And finally, it was noted that the typology does not quite reflect the iterative nature of the BDA process, and thus, it should be better represented in the typology. Based on this feedback, the created model for the typology was improved:

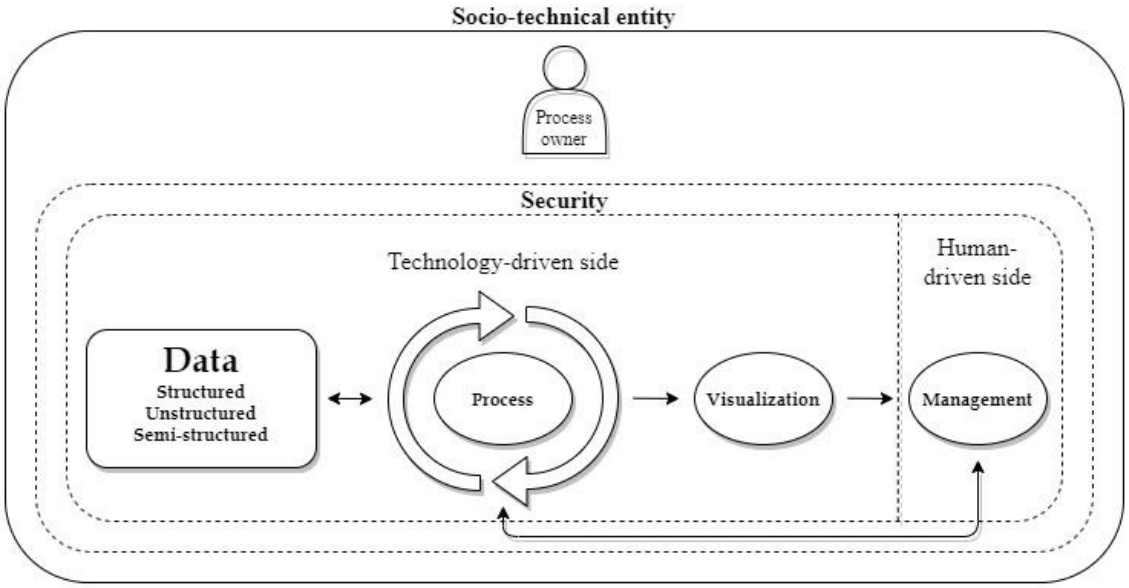

Figure 3: Revised typology of BD based-decision-making challenges

All the functions explained in sub-chapter 2.3.6 apply to the revised typology as well. However, changes made based on the interview validation better illustrate the typology in a practical scenario. Data challenges' weight in the typology was highlighted, and different data types were added to the data challenges section for clarification. Process challenges' iterative nature is considered in the revised version, as well as required two-way communication between the management- and process sections. A process owner was added to represent the human challenges attached to the building of the pipeline. Additionally, socio-technical- and human components were re-iterated as technology- and human-driven sides to better differentiate the two and to avoid unnecessary confusion. Finally, the whole pipeline was defined as a socio-technical entity.

5 DISCUSSION AND CONCLUSIONS

In this thesis, a narrative literature review and a set of five qualitative, semi-structured interviews were conducted about Big Data (BD), Big Data Analytics (BDA), and challenges related to decision-making based on them. Three research questions were formulated to reach the research goals set for this thesis:

- RQ1: How can Big Data be defined?
- RQ2: What are the most relevant challenges associated with Big Data-based decision-making recognized in the literature?
- RQ3: Which of the challenges of RQ2 are the most relevant to the practitioners of the field?

To answer these research questions, the literature review was divided into two parts, the first focusing on comprehensively defining BD. The second part of the literature review was dedicated to identifying the key challenges related to BD-based decision-making. In addition to the literature review, five semi-structured interviews were carried out to industry professionals of varied backgrounds and professional experience.

5.1 Discussion

We introduced a wide variety of different definitions associated with BD. The most popular way in the literature is to define BD through multiple Vs associated with it, mainly volume, variety, velocity, veracity, variability, value, visualization, and volatility. Multiple alternative definitions were also provided.

To answer the first research question, we conclude that BD is a broad and fickle term strongly associated with the context it is used in. Thus, to choose the appropriate definition, one should be familiar with the given context. The Vs are an adequate way of describing BD's technical attributes, but where it falls behind is fully capturing BD as a socio-technical – as requiring a combination of people and technology – and cultural concept – as in changing the way data analysis is conducted in the future. On the other hand, some other definitions

describe BD well on a higher level to the average user but fail to comprehensively describe the technical aspects, requirements, and possibilities of BD.
As for the second research question, we identified a wide variety of different challenges related to BD and BDA. These challenges were categorized as data-, process-, management-, security-, and visualization challenges. What is worth noting is that as decision-making is the final step in the BDA process before implementation, all challenges – from data to visualization – are connected to the decision-making. Thus, to enhance the decision-making capabilities of an organization, it must not ignore data- and process challenges to better focus on management challenges and expect desired results. In other words, challenges in the BDA process should not be examined as silos – as in data processing being one silo and decision-making being another – but instead as a continuous flow of activities. In this activity flow, challenges ignored in previous activities reflect the following activities. To further emphasize this, a framework was created, validated, and revised to display BD decision-making challenges as a linear process rather than silos. To find a summary of all identified theoretical challenges one should refer to table 3, where we have summarized all of our findings from the literature.

Semi-structured interviews were chosen as the qualitative research method for this study. Five interviews were performed to industry professionals of varied backgrounds and professional experience. English and Finnish interview guides can be found in appendixes 1 and 2. A total of 16 different themes were identified based on the interviews. The themes were BD definition, BDA definition, BD strengths, BD weaknesses, BD opportunities, BD threats, BD utilization in decision-making, utilization challenges, BD integration to decision-making, integration challenges, data challenges, process challenges, visualization challenges, management challenges, security challenges, and typology validation.

BD and BDA definitions were approached from a very practical perspective by the respondents. Key findings were that according to the respondents, BD cannot be stored on a single machine and BDA cannot be performed with a single computer in a reasonable amount of time. BD was described to be data from multiple sources and produced by multiple actors. BDA was noted to be refining BD and creating something from it with a goal of finding connections or patterns.

Strengths of BD were identified to be its vast size and how much information it contains. Finding patterns from BD was seen as easier and more credible, as there is more data to spare and everything is mathematically closer to normal distribution. Key weaknesses of BD were said to be data size, -quality, -variety, transparency, and warehousing. Because of its massive size, BD requires specific expertise and infrastructure requirements. Data quality is often difficult to verify, and data variety makes its processing time-consuming. Transparency issues made it difficult to backtrack possible errors in the analytics process. Warehousing issues were identified problematic as there are many aspects to consider.

Future opportunities for BD usage were vague. BD was described as the future's oil that enables organizations to better understand themselves and their clients. Also, it was discussed that many of the future opportunities are challenging to imagine yet. Respondents identified BD threats revolving around security, privacy, and ethical issues. Also, blindly trusting analysis and over-analyzing was deemed a threat.

BD's usage in organizational decision-making was said to improve decision-making quality. Respondents identified practical actions on how BD can be used in decision-making. The goal of the usage was deemed to be more important than the method: to find new information not traditionally available and getting that information to decision-makers. Utilization was also noted to contain many challenges. Communication with decision-makers, demonstrating value, transparency, combining multiple data sources, data usage, and organizational competence were the key challenges identified.

Integrating BD to decision-making was said to come with many practical challenges, and that in today's world large organizations are far away from fully taking the benefit of BD. Technological challenges identified in the integration process were said to come from strict infrastructure requirements, but also seen as the easiest challenge to resolve with enough capital. Organizational challenges were deemed more problematic, as in how to create an organizational culture that encourages data-driven decision-making. The integration process itself was also deemed challenging, as there are many variables involved and everything has to be carefully planned.

Key data challenges arisen in the interviews were data availability, -quality, and -relevance. Data availability refers to issues acquiring the data, and data quality refers to noise, gaps, and incorrectness in the data that need to be addressed, thus altering the outcome. Data relevance was said to do with organizations' ability to identify the data relevant to their specific needs.

Cooperation, business-IT alignment, manual operations, and transparency were process challenges identified in the interviews. Cooperation throughout the process was seen as a key enabler for success but building a process to encourage it was found challenging. Business-IT alignment in the process was said to require very specific competence and know-how, thus requiring talent management. Processing of BD contains a lot of manual functions that were said to take time, thus increasing data latency, and increasing the risk of human errors. Transparency issues arose from the manual phases: backtracking mistakes in the process was noted to be difficult.

Challenges associated with data visualization had to do with determining visualization scope and metrics and designing the visualization and interaction possibilities. Determining scope and metrics was said to be an issue, because without succeeding, the visualization might answer the wrong questions. Designing visualization and interaction possibilities was noted challenging, as there exists a tradeoff between information loss and technical capabilities.

Management challenges identified were communication, management attitudes, determining analysis questions, and interpreting the data. Communica-

tion was seen as an issue, because many decisions are already made in the processing of the data, and without effective communication of those decisions, the management might make uninformed decisions as the context is not clear. Management attitudes covered the fact that decision-makers should possess personal curiosity towards the data and its context to make fully informed decisions based on it. Determining analysis questions was seen as a key building block for the whole process, because if management asks the wrong questions, the whole analytics process might provide wrong, or insufficient information. Interpretation of data was seen challenging, because there is not always right or wrong answers in the data. Also, interpretation is conducted before the decision-making changing the context of the analysis.

Regulatory- and human challenges were identified regarding BD security. Organizations must be aware of all regulations and act respectively. Human challenges regarding security start with the design of the process: the process is built by humans and some information security aspects might be overlooked. Analysis teams were also said to often lack security experts, leaving the team with a limited understanding of security aspects. Access management was identified as a key control to improve process security.

To answer our third research question, data, data availability, and -quality were the most popular answers as being the most relevant BD decision-making challenges to the practitioners. It also came up that this is due to organizations being in the earlier phases of BD decision-making adoption, and the rest of the phases would become bigger issues in the future. Data usage and modeling were also brought up as relevant challenges, as well as its management. The typology presented in sub-chapter 2.3.6 was validated in the interviews, and the revised version of it is presented in sub-chapter 4.17.

5.2 Theoretical and practical implications

For academics, by synthesizing a broad range of BD studies, the results of this study provide an accurate, inclusive, current-day explanation on defining BD. Different ways of defining BD were discussed, and it was concluded that BD definition is highly dependent on the context it is used in. Additionally, this thesis offers a widescale analysis of all challenges that relate to BD-based decision-making. The challenges were categorized based on their respective position in the BD-based decision-making pipeline.

For practitioners, the thesis offers empirical evidence on which challenges of BD-based decision-making are the most relevant for the organizations today. In addition to this, through the validation process performed in this thesis, the practitioners of the field get a comprehensive practical view on which challenges might present themselves in their BD-based decision-making pipeline. This enables the organizations to prepare and create controls to tackle these challenges before they manifest in practice. Through the empirical results, practitioners can also get valuable information on various challenges' attributes and

roles in practical BD-based decision-making. Finally, for academics and practitioners alike, a framework was created to better illustrate the BD-based decision-making pipeline and different challenges' roles and positions in the pipeline.

5.3 Conclusion

This thesis provides contributions to both, the academic community, and practitioners alike. The first chapter of the literature review presents current knowledge of defining BD, examining multiple studies to provide a comprehensive summary of the topic. The second chapter first of all describes challenges often found scattered in the literature of the field. Secondly, it expands current frameworks by categorizing security and visualization as their main challenge groups and provides a completely new validated typology of addressing BD decision-making challenges as a linear process. Thirdly, it uncovers the industry professionals' views regarding the topic, thus providing needed empirical validation for the challenges identified in the literature. Table 6 presents the challenges found in the literature of the field originally summarized in table 3, and compares them to the interviews to summarize, which challenges were validated through the semi-structured interviews.

5.4 Limitations

The main limitations of the study are focused on the time frame of the source material. As the goal with the source literature was to use as many recent studies as possible, some relevant studies were possibly missed during the backward reference searching of the selected articles. Additionally, the ongoing COVID-19 pandemic made it substantially more difficult to get professionals to participate in interviews, as many understandably had their priorities elsewhere. Thus, the goal set for the number of interviewees was off by one and not reached. Finally, as the interviews were performed for members of one large international multi-disciplinary organization, the results are difficult to generalize to the whole industry, including small to mid-size actors.

5.5 Future research agenda

Further research leaves plenty of room for both qualitative and quantitative studies. Quantitative studies should be conducted on a broader industry level, to determine whether different kinds of organizations recognize different challenges. There is also justification for longitudinal study in an organization

adapting BDA as a part of their decision-making process to determine which challenges prove most critical for the implementation process of BD based decision-making. Additional research should also be conducted to explore the practical business consequences of different BD-based decision-making challenges manifesting in practice.

Table 6: Challenges validated through semi-structured interviews

Type	**Sub-challenge**	**Validated through the interviews (Y/N)**
Data	Volume	Y
	Variety	Y
	Velocity	Y
	Veracity	Y
	Variability	Y
	Value	Y
Process	Filtering	Y
	Processing	Y
	Uncertainty	Y
	Process strategy	Y
	Scalability	N
	Reliability	N
	Fault tolerance	N
	Latency	Y
	Analysis	N
Management	Leadership	Y
	Talent management	Y
	Decision-making	Y
	Technology	Y
	Culture	Y
	Governance	N
Security	Security	Y
	Privacy	Y
Visualization	Visual noise	Y
	Information loss	Y
	Image perception	Y
	Image change rate	N
	Performance and scalability	N
	Interpretation	Y
	Visibility	N

5.6 Acknowledgments

I want to thank my thesis supervisor, Ph.D. Erol Kazan for the continuous support throughout the process of writing this thesis, and for actively challenging my way of thinking for improved results.

REFERENCES

Abawajy, J. (2015). Comprehensive analysis of big data variety landscape. *International journal of parallel, emergent, and distributed systems, 30*(1), 5-14.

Adrian, M. (2013). Big data. *Teradata Magazine*, *1*(11).

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research 25*(3), 352-363.

Akter, S., Bandara, R., Hani, U., Wamba, S. F., Foropon, C., & Papadopoulos, T. (2019). Analytics-based decision-making for service systems: A qualitative study and agenda for future research. *International Journal of Information Management, 48*, 85-95.

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets, 26*(2), 173-194.

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016, December). Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 656-660). IEEE.

Asay, M. (2017). 85% of big data projects fail, but your developers can help yours succeed. TechRepublic. Retrieved 5.12.2019, from https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

Balachandran, B. M., & Prasad, S. (2017). Challenges and benefits of deploying big data analytics in the cloud for business intelligence. *Procedia Computer Science, 112*, 1112-1122.

Baldwin, H. (2015). When big data projects go wrong. Forbes. Retrieved 5.12.2019, from https://www.forbes.com/sites/howardbaldwin/2015/01/22/when-big-data-projects-go-wrong/#5e82ab086427

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: big data challenges in the web of things. *IEEE Intelligent Systems*, (6), 6-11.

Basha, S. M., Rajput, D. S., Bhushan, S. B., Poluru, R. K., Patan, R., Manikandan, R., & Kumar, A. (2019). Roadmap. *International Journal on Emerging Technologies, 10*(2), 50-59.

Batarseh, F. A., & Latif, E. A. (2016). Assessing the quality of service using big data analytics: with application to healthcare. *Big Data Research, 4*, 13-24.

Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of general psychology, 1*(3), 311-320.

Bhimani, A. (2015). Exploring big data's strategic consequences. *Journal of Information Technology, 30*(1), 66-69.

Bertino, E. (2013). Big Data--Opportunities and Challenges Panel Position Paper. In 2013 *IEEE 37th Annual Computer Software and Applications Conference* (pp. 479-480). IEEE.

Big Data Statistics 2019. Retrieved 14.11.2019, from (https://techjury.net/stats-about/big-data-statistics/#gref)

Bihani, P., & Patil, S. T. (2014). A comparative study of data analysis techniques. *International journal of emerging trends & technology in computer science, 3*(2), 95-101.

Borne, K. (2014). Top 10 Big Data challenges—A serious look at 10 Big Data V's. Retrieved 4.12.2019, from https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Braun, V., & Clarke, V. (2012). Thematic analysis.

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big data. *Business & Information Systems Engineering, 5*, 65-69

Casado, R., & Younas, M. (2015). Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience, 27*(8), 2078-2091.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications, 19*(2), 171-209.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data. *Information sciences, 275*, 314-347.

Delen, D., & Demirkan, H. (2013). Data, information, and analytics as services. *Decision Support Systems, 55,* 359-363.

De Mauro, A., Greco, M., & Grimaldi, M. (2015, February). What is big data? A consensual definition and a review of key research topics. *In AIP conference proceedings* (Vol. 1644, No. 1, pp. 97-104). AIP.

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management, 54*(5), 807-817.

Denison, D. R. (1984). Bringing corporate culture to the bottom line. *Organizational Dynamics, 13*(2), 5-22.

Dumbill, E. (2013). Making sense of big data. *Big Data, 1*(1), 1-2.

Economist Intelligence Unit. (2012). The deciding factor: Big data & decision-making. In C.G. reports (Ed.), (pp. 24): Economist Intelligence Unit.

Ekambaram, A., Sørensen, A. Ø., Bull-Berg, H., & Olsson, N. O. (2018). The role of big data and knowledge management in improving projects and project-based organizations. *Procedia computer science, 138*, 851-858.

Ethiraj, S. K., Kale, P., Krishnan, M. S., & Singh, J. V. (2005). Where do capabilities come from and how do they matter? A study in the software services industry. *Strategic management journal*, *26*(1), 25-45.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, *25*(10), 1229-1245.

Galvin, R. (2015). How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge?. *Journal of Building Engineering*, *1*, 2-12.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management, 35*(2), 137-144.

Garg, N., Singla, S., & Jangra, S. (2016). Challenges and techniques for testing of big data. *Procedia Computer Science, 85*, 940-948.

Gartner. (2015). Gartner says business intelligence and analytics leaders must focus on mindsets and culture to kick start advanced analytics. Retrieved 5.12.2019, from https://www.gartner.com/en/newsroom/press-releases/2015-09-15-gartner-says-business-intelligence-and-analytics-leaders-must-focus-on-mindsets-and-culture-to-kick-start-advanced-analytics

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of chiropractic medicine*, *5*(3), 101-117.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods*, *18*(1), 59-82.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, *53*(8), 1049-1064.

Hamoudy, E. (2014). Analyzing 6Vs of Big Data using System Dynamics. *The 2nd Scientific Conference of the College of Science*, 75-83.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data, 6*(1), 44.

Herold R (2002) What is the difference between security and privacy. InformationShield.

Horita, F. E., de Albuquerque, J. P., Marchezini, V., & Mendiondo, E. M. (2017). Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems, 97*, 12-22.

Ignatius, A. (2012). From the editor: Big data for sceptics. *Harvard Business Review 90*, 12-12.

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

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research, 70*, 338-345.

Joseph, R. C., & Johnson, N. A. (2013). Big data and transformational government. *IT Professional, 15*(6), 43-48.

Kraska, T. (2013). Finding the needle in the big data systems haystack. *IEEE Internet Computing*, *17*(1), 84-86.

Kuner, C., Cate, F. H., Millard, C., & Svantesson, D. J. B. (2012). The challenge of 'big data' for data protection. *International Data Privacy Law, 2*(2), 47-49.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5*(12), 2032-2033.

Latif, Z., Lei, W., Latif, S., Pathan, Z. H., Ullah, R., & Jianqiu, Z. (2019). Big data challenges: Prioritizing by decision-making process using Analytic Network Process technique. *Multimedia Tools and Applications, 78*(19), 27127-27153.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics, and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21-32.

Lawson, R. A., Blocher, E. J., Brewer, P. C., Cokins, G., Sorensen, J. E., Stout, D. E., ... & Wouters, M. J. (2013). Focusing accounting curricula on students' long-run careers: Recommendations for an integrated competency-based framework for accounting education. *Issues in Accounting Education, 29*(2), 295-317.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. In *Forum Qualitative Sozialforschung/Forum: qualitative social research* 11(3), Art. 8.

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

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big data: prospects and challenges. *Vikalpa, 40*(1), 74-96.

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2002). *Risk communication: A mental models approach*. Cambridge University Press.

Namey, E., Guest, G., McKenna, K., & Chen, M. (2016). Evaluating bang for the buck: a cost-effectiveness comparison between individual interviews and focus groups based on thematic saturation levels. *American Journal of Evaluation*, *37*(3), 425-440.

N.N.I Initiative, Core techniques and technologies for advancing big data science and engineering (BIGDATA). (2012). Retrieved 5.12.2019, from https://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

Palanimalai, S., & Paramasivam, I. (2016). Big data analytics bring new insights and higher business value - an experiment carried out to divulge sales forecasting solutions. *International Journal of Advanced Intelligence Paradigms,* 8(2), 207-218.

Phillips-Wren, G. E., Iyer, L. S., Kulkarni, U. R., & Ariyachandra, T. (2015). Business Analytics in the Context of Big Data: A Roadmap for Research. *CAIS, 37*, 23.

Press, G. (2017). 6 predictions for the $203 billion big data analytics market. Forbes. Retrieved 5.12.2019, from https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#5bd4f2de2083

Raikov, A. N., Avdeeva, Z., & Ermakov, A. (2016). Big data refining on the base of cognitive modeling. *IFAC-PapersOnLine*, *49*(32), 147-152.

Ransbotham, S., Kiron, D., & Prentice, P. K. (2016). Beyond the hype: the hard work behind analytics success. *MIT Sloan Management Review, 57*(3), 3-16.

Rogers, P., & Meehan, P. (2007). Building a winning culture. *Business Strategy Series*, *8*(4), 254-261.

Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management*, *54*(5), 758-790.

Sammut, G., & Sartawi, M. (2012). Perspective-taking and the attribution of ignorance. *Journal for the Theory of Social Behaviour*, *42*(2), 181-200.

Shamim, S., Zeng, J., Shariq, S. M., & Khan, Z. (2019). Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: A dynamic capabilities view. *Information & Management, 56*(6), 103135.

Sheng, J., Amankwah-Amoah, J., & Wang, X. (2019). Technology in the 21st century: new challenges and opportunities. *Technological Forecasting and Social Change, 143*, 321-335.

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.

Smith, M., Szongott, C., Henne, B., & Von Voigt, G. (2012, June). Big data privacy issues in public social media. *In 2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)* (pp. 1-6). IEEE.

Strauß, S. (2015). Datafication and the seductive power of uncertainty—A critical exploration of big data enthusiasm. *Information, 6*(4), 836-847.

Tabesh, P., Mousavidin, E., & Hasani, S. (2019). Implementing big data strategies: A managerial perspective. *Business Horizons, 62*(3), 347-358.

Thabet, N., & Soomro, T. R. (2015). Big Data Challenges. *Journal of Computer Engineering & Information Technology, 4*(3).

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big Data in accounting: An overview. *Accounting Horizons, 29*(2), 381-396.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77-84.

Wang, L., Wang, G., & Alexander, C. A. (2015). Big data and visualization: methods, challenges, and technology progress. *Digital Technologies, 1*(1), 33-38.

Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision-making: An overview on challenges and trends of Big Data. *Information Sciences*, *367*, 747-765.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. *International Conference on Advances in Cloud Computing*. (ACC-2012), Bangalore, India, July 2012, pp 21-29.

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering*, *101*, 572-591.

Zicari, R. V. (2014). Big data: Challenges and opportunities. *Big data computing, 1*, 103-128.

APPENDIX 1: INTERVIEW GUIDE

1. *Introduction*
   1.1. Introduce yourself
   1.2. Introduce the study
   1.3. Inform interviewee of confidentiality
   1.4. Inform interviewee of the right not to answer a question if they do not wish to
   1.5. Inform interviewee of right to stop the interview if they wish
   1.6. Get consent for audio recording

2. *Questions about the interviewee*
   2.1. What is your professional background?
      2.1.1. What is your current job title?
      2.1.2. How long is your professional background?
   2.2. Describe how does your work relate to Big Data
      2.2.1. How long have you worked with Big Data?

3. *Big Data questions*
   3.1. Please define Big Data
   3.2. Please define Big Data Analytics
   3.3. What are the greatest strengths of Big Data?
   3.4. What are the greatest weaknesses of Big Data?
   3.5. What are the greatest opportunities of Big Data?
   3.6. What are the greatest threats of Big Data?

4. *Questions about Big Data-based decision-making challenges*
   4.1. In general, how can Big Data be utilized in organizational decision-making processes?
      4.1.1. How is Big Data currently utilized in your organization's decision-making processes?
   4.2. In general, is it challenging to integrate Big Data into organizational decision-making processes?
      4.2.1. Has it been challenging to integrate Big Data into your organization's decision-making process?
      4.2.2. What makes the integration challenging?
   4.3. In general, is it challenging to utilize Big Data in organizational decision-making processes?
      4.3.1. Has it been challenging to utilize Big Data in your organization's decision-making process?
      4.3.2. What makes the utilization challenging?

5. *Questions about the typology for BD-based decision-making*

5.1. Present and explain the typology to interviewee
5.2. What makes the data aspect of Big Data challenging?
5.3. What makes the processing of Big Data challenging?
5.4. What makes the visualization of Big Data challenging?
5.5. What makes the management of Big Data challenging?
5.6. What makes the security of Big Data challenging?
5.7. Does this typology relate to a practical Big Data-based decision-making process?
5.8. Which section of the typology is the most challenging for organizations to tackle?
5.9. Would you like to add/point out/delete something regarding this discussion?

6. *Conclusion*
6.1. Conclude the interview
6.2. Thank the interviewee
6.3. Inform the interviewee how the study will proceed

APPENDIX 2: HAASTATTELURUNKO

1. *Johdanto*
   - 1.1. Esittele itsesi
   - 1.2. Esittele tutkimus
   - 1.3. Informoi haastateltavaa haastattelun luottamuksellisuudesta
   - 1.4. Informoi haastateltavaa mahdollisuudesta olla vastaamatta kysymykseen
   - 1.5. Informoi haastateltavaa mahdollisuudesta lopettaa haastattelu, mikäli he haluavat
   - 1.6. Pyydä suostumus haastattelun nauhoitukseen (ääni)

2. *Kysymykset haastateltavasta*
   - 2.1. Millainen on ammatillinen taustanne?
     - 2.1.1. Mikä on nykyinen työnimikkeenne?
     - 2.1.2. Kuinka pitkä ammatillinen taustanne on?
   - 2.2. Kuvailisitteko, miten työnne liittyy Big Dataan?
     - 2.2.1. Kauanko olette työskennelleet Big Datan parissa?

3. *Kysymykset Big Datasta*
   - 3.1. Määrittelisittekö Big Datan?
   - 3.2. Määrittelisittekö Big Data -analytiikan?
   - 3.3. Mitkä ovat Big Datan suurimpia vahvuuksia?
   - 3.4. Mitkä ovat Big Datan suurimpia heikkouksia?
   - 3.5. Mitkä ovat Big Datan suurimpia mahdollisuuksia?
   - 3.6. Mitkä ovat Big Datan suurimpia uhkia?

4. *Kysymykset Big Data-pohjaisen päätöksenteon haasteista*
   - 4.1. Yleisellä tasolla, miten Big Dataa voidaan hyödyntää organisaation päätöksentekoprosesseissa?
     - 4.1.1. Miten Big Dataa tällä hetkellä hyödynnetään teidän organisaationne päätöksentekoprosesseissa?
   - 4.2. Yleisellä tasolla, onko Big Datan integroiminen osaksi organisaatioiden päätöksentekoprosesseja haastavaa?
     - 4.2.1. Onko Big Datan integroiminen teidän organisaationne päätöksentekoprosesseihin ollut haastavaa?
     - 4.2.2. Mikä tekee integroimisesta haastavaa?
   - 4.3. Yleisellä tasolla, onko Big Datan hyödyntäminen organisaation päätöksentekoprosesseissa haastavaa?
     - 4.3.1. Onko Big datan hyödyntäminen teidän organisaationne päätöksentekoprosesseissa ollut haastavaa?
     - 4.3.2. Mikä tekee hyödyntämisestä haastavaa?

5. *Kysymykset Big Data-pohjaisen päätöksenteon typologiasta*
   5.1. Esittele ja selitä typologia haastateltavalle
   5.2. Mikä tekee Big Datan datapuolesta haastavaa?
   5.3. Mikä tekee Big Datan prosessoimisesta haastavaa?
   5.4. Mikä tekee Big Datan visualisoimisesta haastavaa?
   5.5. Mikä tekee Big Datan johtamisesta haastavaa?
   5.6. Mikä tekee Big Datan tietoturvasta haastavaa?
   5.7. Liittyykö tämä typologia mielestänne käytännön Big Data-pohjaiseen päätöksentekoprosessiin?
      5.7.1. Miksi/miksi ei?
   5.8. Mikä typologian osio on organisaatioille haastavin?
   5.9. Haluaisitteko lisätä/korjata/poistaa jotain liittyen tähän keskusteluun?

6. *Yhteenveto*
   6.1. Vedä haastattelu yhteen
   6.2. Kiitä haastateltavaa
   6.3. Informoi haastateltavaa, miten tutkimus etenee tästä