2020

Exploring Mid-Market Strategies for Big Data Governance

Kenneth Stanley Knapton III
Walden University

Walden University

College of Management and Technology

This is to certify that the doctoral study by

Ken Knapton

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Jodine Burchell, Committee Chairperson, Information Technology Faculty
Dr. Steven Case, Committee Member, Information Technology Faculty
Dr. Gary Griffith, University Reviewer, Information Technology Faculty

Chief Academic Officer and Provost
Sue Subocz, Ph.D.

Walden University
2020

Abstract

Exploring Mid-Market Strategies for Big Data Governance

by

Ken Knapton

MS, Walden University, 2018

MBA, Brigham Young University, 2004

BSc, Utah Valley University, 1997

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Information Technology

Walden University

August 2020

Abstract

Many data scientists are struggling to adopt effective data governance practices as they transition from traditional data analysis to big data analytics. Data governance of big data requires new strategies to deal with the volume, variety, and velocity attributes of big data. The purpose of this qualitative multiple case study was to explore big data governance strategies employed by data scientists to provide a holistic perspective of those data for making decisions. The participants were 10 data scientists employed in multiple mid-market companies in the greater Salt Lake City, Utah area who have strategies to govern big data. This study's data collection included semi-structured in-depth individual interviews (n = 10) and analysis of process documentation relating to big data governance in those organizations (n = 4). Through thematic analysis, 4 major themes emerged from the study: ensuring business centricity, striving for simplicity, establishing data source protocols, and designing for security. One key recommendation for data scientists is to minimize the data noise typically associated with big data. Implementing these strategies can help data scientists' transition from traditional to big data analytics, which could help those organizations to be more profitable by gaining competitive advantages. The strategies outlined in this study can lead to positive social change by proactively addressing the ethical use of personally identifiable information in big data. By implementing strategies relating to the segregation of duties, encryption of data, and personal information, data scientists can mitigate contemporary concerns relating to the use of private information in big data analytics.

Exploring Mid-Market Strategies for Big Data Governance

by

Ken Knapton

MS, Walden University, 2018

MBA, Brigham Young University, 2004

BSc, Utah Valley University, 1997

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Information Technology

Walden University

August 2020

Dedication

I would like to dedicate this research study to my wife, Debbie. Without her support, this academic journey would not have been possible.

Acknowledgments

This doctoral study is the culmination of efforts from many individuals. I would like first to thank my wonderful committee chairperson, Dr. Jodine Burchell. I am so grateful for her support, guidance, and encouragement throughout this process. I would also like to thank my other committee members, Dr. Steven Case and Dr. Gary Griffith. I would also like to thank the Program Director Dr. Gail Miles. Finally, I am grateful to all of the participants and gatekeepers in the organizations that participated in this study. I am thankful for their time, effort, and thoughtful responses to my questions.

Table of Contents

List of Tables

List of Figures

Section 1: Foundation of the Study

Background of the Problem

Data generation has increased exponentially in recent years (Bello-Orgaz, Jung, & Camacho, 2016), with no signs of this trend stopping. Bello-Orgaz et al. (2016) estimated that globally, 2.5 exabytes of new data are being generated per day, and the Government Accountability Office (2017) estimated that by 2025, there would be between 25 and 50 billion devices connected to the Internet and generating data. Even with the vast amount of data available to them, organizations effectively use less than 5% of their available data (Zakir, Seymour, & Berg, 2015). This emergence of big data has introduced data management challenges involving processing speed, data interpretation, and data quality for organizations that wish to consume complex information (Lee, 2017). The traditional methods, frameworks, strategies, and tools for data governance and analysis are outdated and no longer adequate for processing the vast amount of data available to organizations today, thus making current strategies ineffective for handling big data (Bello-Orgaz et al., 2016). Big data analytics challenges arise from issues relating to data that are too vast, unstructured, and moving too fast to be managed by traditional means (Zakir et al., 2015).

Companies of all sizes are dealing with this new big data environment and struggling to determine how best to analyze big data as a critical driver of strategic business decisions. Larger companies have more resources to direct toward this problem, but mid-market organizations face similar issues with less available capital to apply to the problem. To remain competitive, however, it is just as critical for them to find ways to address this data deluge that faces companies of all sizes.

Problem Statement

Contemporary outdated data processing systems are unable to handle the data deluge of exponentially increasing amounts of data that we are generating daily (Sivarajah, Kamal, Irani, & Weerakkody, 2017). More than 40% of organizations are currently challenged to attract and retain skilled data scientists, while by 2020, the U.S alone will need more than 190,000 skilled data analysts (Mikalef, Giannakos, Pappas, & Krogstie, 2018). The general IT problem is that there is a lack of knowledge regarding data governance principles for analyzing big data. The specific IT problem is that some data scientists lack big data governance strategies that provide a holistic perspective of those data for making decisions.

Purpose Statement

The purpose of this qualitative multiple case study was to explore the big data governance strategies employed by data scientists to provide a holistic perspective of those data for making decisions. The population for this study was data scientists employed in three mid-market companies in the greater Salt Lake City, Utah area who have strategies to govern big data. The findings from this study could potentially affect social change by creating additional jobs in mid-market companies. Job creation would be accomplished through increased capability of these companies to have a better understanding of meaning of data which they gather and use to make strategic and operational business decisions daily. This increased and more targeted information could lead to increased profitability for these mid-market companies, which in turn could lead to increased jobs available in these communities. These new jobs could elevate the mid-

market organization's community by providing better pay for residents and improving their lives in general.

Nature of the Study

This study used a qualitative method to explore the big data governance strategies employed by data scientists to provide a holistic analysis of those data for making decisions. Qualitative studies are useful when investigating peoples' views or human behavior, or when attempting to find out how well something is performed, but for which quantitative data does not offer a complete picture (Kelly, 2017). Because strategies for data governance involve understanding how well those governance activities are performed, the qualitative method was appropriate. Quantitative studies can demonstrate trends and correlations, but are not as capable of providing explanations or reasons (Kelly, 2017). Quantitative studies are focused on predictions and theories based on cause-and-effect relationships, which are testable and generalizable (Gehman et al., 2018). Since this study was not attempting to prove a hypothesis, a quantitative approach was not appropriate. A mixed methods approach would involve combining quantifiable data with an in-depth understanding of participants' behavior and actions (McCusker & Gunaydin, 2015). Because strategies related to data governance do not involve quantifiable statistical data, a mixed methods approach was not appropriate.

I used a multiple case study methodology for this study. A case study methodology allows a deep examination of an issue or phenomenon in its natural setting (Arseven, 2018). Because this study focused on strategies used within organizations that are using big data for decision-making, this design approach was warranted. An ethnographic approach requires embedding the researcher in a cultural context to observe

the behaviors of a people in a particular cultural context (Watson, 2018). I did not select the ethnographic design because this study was not investigating any specific culture, customs, or social group. Thus, it was not appropriate for this study. The phenomenological design was not appropriate because it involves how participants live through, interpret, or experience the topic under research (Sousa, 2014). Because this study was not investigating perceptions of participants, the phenomenological approach would not have been appropriate.

Qualitative Research Question

What big data governance strategies do data scientists employ to provide a holistic perspective of those data for making decisions?

Interview Questions

Demographic Questions

1. What is your role in data gathering and analysis?
2. Please describe your background concerning big data analysis.
3. How long have you been involved in big data analysis?

Interview Questions

1. How do you ensure that you are analyzing all available data for any particular model?
2. How do you determine appropriate sources of data for any specific model or analysis?
3. What processes are in place to ensure that all applicable data are available to you when needed for analysis?

4. What countermeasures are in place for risk mitigation regarding the validity of the analysis that you perform (i.e., the risk of not having current data, or of incorrect correlation of data from various sources)?
5. What have you found to be most successful in building data models based on big data from various sources?
6. Why do these processes give you confidence that you are analyzing all applicable data?
7. How do these processes ensure the protection of the data?
8. What controls are in place to ensure the privacy of the data?
9. What other issues are there that we should discuss as relating to big data analysis?
10. What other areas are there left to discuss?

Conceptual Framework

I selected the organizational information processing theory (OIPT) as the lens through which I examined data governance strategies used by data scientists to provide a holistic perspective of those data for making decisions. Jay R. Galbraith introduced the OIPT in 1974 (Galbraith, 1974). The central concept of the theory is that organizations need relevant information to make decisions. Three concepts primarily comprise the ability to use that information for decision-making efficiently: the need within an organization to process information, the organization's inherent ability to process information, and the gap between the two (Premkumar, Ramamurthy, & Saunders, 2005).

The OIPT aligns well with this study because of the direct association between the amount of information that is available to an organization and the ability for that organization to process that information to improve the decision making within that

organization. Data scientists are starting to employ novel strategies in terms of governance and analysis of big data to transform large quantities of information into valuable insights for strategic business decision-making processes. By using the lens of the OIPT, I was able to assess strategies used by data scientists to determine whether these strategies are increasing the ability of the organization to process information and thereby reduce slack resources within the organization. By viewing governance of big data through the lens of the OIPT, I focused on strategies of data scientists to properly govern the vast amount of data available to them as a mechanism to reduce the uncertainty of information analysis to enable better decision-making throughout their organizations.

Definition of Terms

Data/Information Governance: An emerging discipline involving managing, processing, and controlling information throughout an organization. It encompasses processes, procedures, and rules involving the management of information throughout an organization (Mullon & Ngoepe, 2019).

IT Governance: The set of processes and oversight within an organization that ensures that investments in IT generate business value, mitigate risks associated with IT initiatives, and operate with accountability and traceability for those who are funding the IT resources (Grossman, 2018).

Assumptions, Limitations, and Delimitations

Assumptions

Walsh (2015) described assumptions as unconscious beliefs that the researcher accepts as accurate which can produce biases in terms of both perceptions and thoughts. I

assumed that participants answered my questions to the best of their abilities. I also assumed that individuals within organizations that I interviewed were subject matter experts within their organization based on job title and work performed. I assumed that documentation these organizations provided to me was official and reflected their current processes and procedures.

Limitations

Teichler (2014) defined limitations of a study as those aspects of the study that are beyond the control of the researcher, despite their best efforts. Limitations of this study are primarily centered around the fact that I performed a qualitative case study, and as such, findings may not be able to be extrapolated beyond organizations selected for the case study. Qualitative studies in general have inherent limitations since qualitative assessments are limited by incomplete understanding of judgments made by the researcher and lack of consensus among researchers regarding the rigor of deriving those judgments (Cook, Kuper, Hatala, & Ginsburg, 2016).

Delimitations

Delimitations are boundaries defined by the researcher that guide the study itself (Kongnso, 2015). This study was geographically limited to the greater Salt Lake City, Utah, area. Additionally, participants for this study were limited to those who were involved directly in data modeling and data analysis strategy and processes within case study companies, as well as those who were involved in the design of data governance processes and practices. Specific roles or job titles for this responsibility may differ between organizations. Finally, the organizations that I selected for the case study were required to have some experience analyzing big data.

Significance of the Study

Contribution to Information Technology Practice

This study may provide strategies to data scientists in mid-market companies who lack strategies for governing big data. This study may provide value to the overall IT practice by uncovering insights about successful data governance practices that are in use within mid-market organizations. Overall contributions to IT practice may potentially include additional insights into best practices that mid-market organizations have adopted and from which they have found success as they struggle to deal with the deluge of information that is available to them for making critical strategic business decisions. Kemp (2014) indicated that nearly 85% of Fortune 500 organizations are unable to use their data effectively, and a structured approach to data governance, policy, and strategy is a critical factor for project success. Mid-market companies have a similar need to use data available to them for strategic business decisions effectively, and they can typically react and implement new strategies more rapidly than their larger competitors.

Implications for Social Change

The potential for social change involves the increased ability of companies to meet the needs of their customer base due to an increased understanding within the company of the meaning of data which they gather and use to make strategic and operational business decisions daily. This improved customer focus could lead to higher profitability, which in turn would enable these organizations to hire more employees and increase wages. These overall improvements in businesses could lead to overall elevation of residents in these communities by allowing them to earn more money and improve their way of life. Also, better overall management could lead to better employment

opportunities for all types of employees, as well as improved customer experiences for customers of mid-market companies. Effective use of big data could benefit customers by providing more specific and targeted solutions to their needs for small companies that are part of their local communities. In sum, the ability for data scientists to effectively use big data in mid-market companies could be an equalizing force that could benefit customers, employees, and owners of all types of mid-market companies.

A Review of the Professional and Academic Literature

For this study, I focused on three basic concepts when researching the professional and academic literature: issues involving the adoption and use of big data as a reliable and viable technology, applicability of the OIPT as a conceptual model for applying big data governance in the enterprise decision-making process, and the use of the case study methodology to study this topic. I searched for academic literature with Google Scholar, Science Direct, IEEE Explore, Research Gate, and the Walden Library as primary sources for academic research. I used the following search terms to initiate my search at these websites: big data governance, big data challenges, big data decisions, big data benefits, big data roles, big data security, big data ethics, small business big data, data governance, big data analytics, data scientist roles, data analysis roles, big data job description expectations, OIPT, organizational information processing theory, Galbraith, enterprise decision, small business decision making. I also employed the suggested search links that appeared in the search results on these various sites to continue my search. As I studied the literature, I frequently used reference lists of cited articles as alternate sources. I focused my research on contemporary publications, with the intent of

using articles published in 2016 or later. I also researched seminal articles and publications when researching the OIPT and case study topics.

The content of the literature included in this review includes history, contemporary discourse, and contrasting views of the OIPT as well as all aspects of big data governance. I selected seminal sources where applicable and available. I focused my review of literature on the data governance aspects of big data, including expectations of big data analytics, quality and volume of data, challenges due to evolving from traditional to big data management, ethical and privacy concerns, organizational and human resource concerns, and IT and data governance in general. I intentionally left the discussion of specific toolsets such as Hadoop and R out of this study as these are analytic technologies and not explicitly related to the overall governance of big data. I reviewed more than 250 sources in total, eventually eliminating more than 110 sources as not relevant, duplicative, or repetitive. In my review, I did not encounter any studies that specifically addressed the governance of big data and its effect on strategic decision-making in the manner that I am studying.

To document this literature review, I start with a discussion of the conceptual framework and then proceed to relevant literature regarding big data. The review of big data literature starts with a definition of big data, leading to a discussion of its benefits and uses. I then discuss some of the challenges associated with big data, which leads to a discussion of related technologies such as the Internet of Things (IoT) and cloud computing. I end the literature review by discussing IT and data governance as relates specifically to big data initiatives.

This literature review consists of 105 articles relating to the OITP and big data and the associated technologies such as the IoT, cloud computing, data governance, and information governance. Eighty-two percent of articles reviewed were published between 2016 and 2020. Ninety-one percent of these articles were peer-reviewed as verified primarily by searching Ulrich's Periodicals Directory and reviewing journal information directly in many cases.

Application to the Applied IT Problem

The purpose of this qualitative multiple case study is to explore big data governance strategies employed by data scientists to provide a holistic perspective of those data for making decisions. In the next sections, I will provide a thorough discussion of the conceptual framework and current literature regarding big data.

Conceptual Framework

The conceptual framework for this study is the OIPT. I will discuss the OIPT in-depth as well as provide an analysis of both supporting and contrasting theories.

OIPT. OIPT was a seminal theory on organizational design, which contributed to Galbraith later becoming known as the father of modern organizational design theories (Galbraith, 2017). It was Galbraith's mentor, James D. Thompson, who first introduced him to the concept of viewing organizations as systems and studying the inherent need for organizations to process information (Galbraith, 2017). Galbraith's theory was also heavily influenced by Herbert Simon, a Nobel Prize-winning economist who, in conjunction with James Marsh, introduced the concept of slack resources within an organization to aid in information processing (Galbraith, 2017). The basis of the OIPT is that organizations are comprised of individuals who must deal with both predictable and

unpredictable information (Galbraith, 2017). Tushman and Nadler (1978) added that the OIPT directly applies to the gathering, interpretation, and synthesis of information within an organization for decision-making. Feurer, Schuhmacher, and Kuester (2019) explained that the fundamental basis for OIPT is centered around the uncertainty and complexity of tasks to be completed in a task workflow and is based on the concept that organizations should be designed to enable decision-making. All of Galbraith's major writings also cite the contributions of Alfred D. Chandler, which demonstrates that Chandler's work influenced Galbraith's connection of an organization's strategy with its structural form (Galbraith, 2017).

Zelt, Recker, Schmiedel, and vom Brocke (2018) articulated that the OIPT describes organizations as information processing systems that collect, process, and distribute information, and Allegrini and Monteduro (2018) added that greater uncertainty is associated with the need to process more information during task execution. Galbraith (1974) explained that the greater the uncertainty of a task, the more information must be processed by decision-makers to execute that task. Gupta, Kumar, Kamboj, Bhushan, and Luo (2019) explained that a lack of information within an organization leads to uncertainty, and Tushman and Nadler (1978) explained that uncertainty was the difference between information processed and required to complete a task. Foerstl, Meinlschmidt, and Busse (2018) theorized that the OIPT evolved in response to organizational design problems stemming from the size-induced complexity of organizations in the 1970s.

Tushman and Nadler (1978) identified management as collecting, processing, and combining relevant information for organizations. Bartnik and Park (2018) added that

management's role is to reduce both uncertainty and equivocality by clarifying mission statements and organizational priorities for organizations. Gao, Liu, Guo, and Li (2018) explained that merely increasing information available does not resolve uncertainty as individuals do not have unlimited processing capacity and will reach information overload.

Zelt, Recker et al. (2018) clarified that the OIPT builds on prior contingency theories by aligning with the concepts that the organizational structure needs to fit variables both inside and outside of the organization. Duvald (2019) added that the OIPT is one of many contingency theories that identify that while there is no one best way to organize, not all ways of organizing are equally valid. DiMaggio and Powell (1983) established institutional theory in 1983, which establishes that processes that can establish practices, policies, and norms. Chrisman, Chua, Le Breton-Miller, Miller, and Steier (2018) indicated that governance in general is usually the domain of institutional theory. The application of institutional theory focuses mainly on coercive, mimetic, and normative pressures that shape behavior within an organization and not on the effect on strategic decision making (Chrisman et al., 2018).

Duvald (2019) explained that equivocality arises from the existence of multiple and often conflicting interpretations of data, which can only be mitigated by additional processing mechanisms rather than simply gathering more data. Zelt, Recker, et al. (2018) added that increasing equivocality leads to a need for systems to interpret and assign meaning to the information being processed.

Galbraith (1974) theorized that as task complexity increases, organizations can deal with increased uncertainty of task completion in only four ways, two of which

address the reduction of the information needing to be processed, and the other two addressing the possibility of increasing the overall capacity to process information (see Figure 1). Premkumar et al. (2005) added that the OIPT has three primary elements: the need within an organization to process information, their inherent ability to process information, and the gap between the two. Galbraith (2012) explained that organizations can deal with this gap in one of two ways: they can either increase the capacity of the organization to process more information, or they can decentralize the interdependence on that information. Cao, Duan, and Cadden (2019) defined the information processing capability of an organization as their capacity to capture, integrate, and analyze data and information and use those insights in the context of organizational decision-making. Wang, Xu, Fujita, and Liu (2016) added that combining heterogeneous sources of data into new information sets is one way to reduce uncertainty. Gao et al. (2018) theorized that using the lens of the OIPT, it becomes clear that increasing access to information will only serve to create information overload rather than benefiting decision-making processes. Ever-increasing amounts of information coupled with the inability of the organization to process that information lead to organizations adopting two typical strategies for coping with the inherent uncertainty that ensues: developing buffers to reduce the effect of uncertainty and implementing structural mechanisms and information- processing capabilities to enhance information flow and thereby reduce uncertainty (Premkumar et al., 2005).

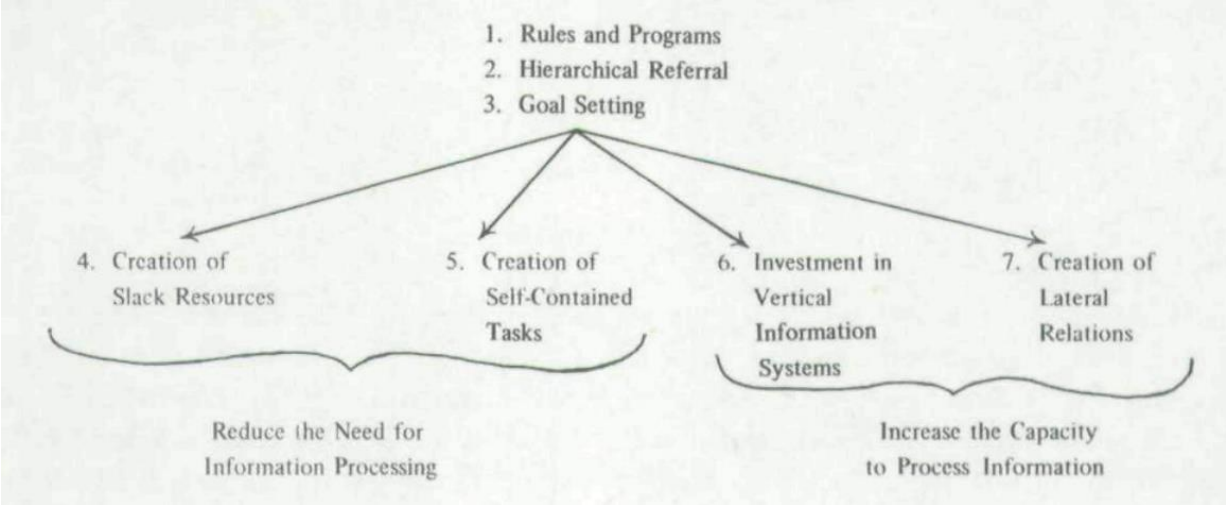

Figure 1. Organizational design strategies. From "Organizational Design, an Information Processing View" by J. R. Galbraith, 1974, *Interfaces*, 4(3), p.30. Reprinted with permission.

While the concept was quite simple, Burton, Obel, and Håkonsson (2015) indicated that Galbraith's subsequent work that built upon this theory became the foundation for how to make matrixed organizations work well. Because of the lack of vertical IT systems to analyze information, the only option was to create matrix organizations to process information and make business decisions. Feurer et al. (2019) explained that one of the general concepts of the OIPT is that as lateral organizational relationships are established, the number of decisions referred upward is reduced, which then leads to more effective decision-making. Obel and Snow (2014) posited that the theory of organizational design allowed many areas of research to be brought together to create a model of organizational design with significant predictive power and that it is even more relevant in today's world of the dramatic increasing availability of information in the form of big data. Intezari and Gressel (2017) added that the ability of

organizations to make strategic decisions amidst uncertainty and ambiguity is predicated their ability to learn and reconfigure their knowledge base continuously.

Additional research regarding the OIPT led Galbraith to investigate matrix organizations as a means to decentralize decision making, which set the foundation for Galbraith's most famous theory on organizational design, named the star model (Galbraith, 2017). Galbraith (2014) demonstrated the potential impact of big data on an organizational design using the star model, as shown in figure 2. The star model included five dimensions of organizational structure: strategy, structure, processes, people, and rewards (Miterev, Turner, & Mancini, 2017). Obel and Snow (2014) indicated that the star model became the most recognized and widely accepted organizational design model. Galbraith based the star model on his belief that organizations are inherently made up of individuals who need to process information to accomplish the work of that organization (Galbraith, 2017).

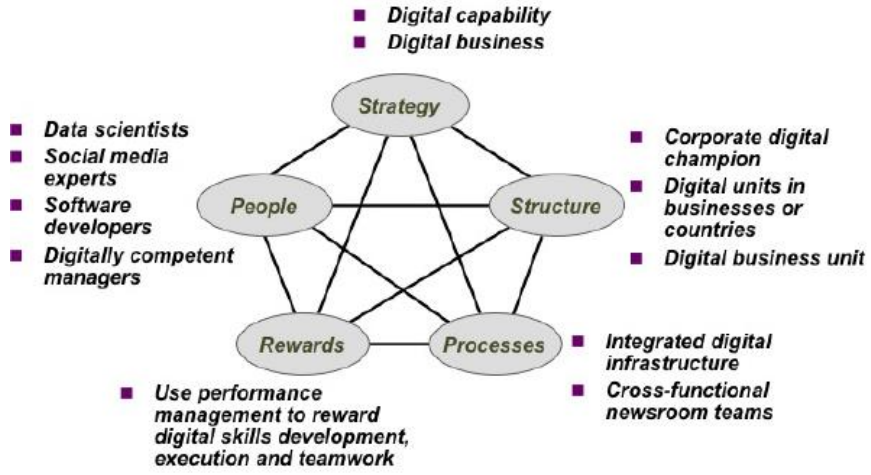

Figure 2. Big data's impact on an organization. From "Organizational design challenges resulting from big data" by J. R. Galbraith, 2014, *Journal of Organizational Design*, *3*(1), p.12. Reprinted with permission.

Contemporary view. Contemporary thought regarding the OIPT shows the theory has remained relatively consistent since its inception with incremental additions and developments. Kroh, Luetjen, Globocnik, and Schultz (2018) explained that the evolution of IT systems was evidence that the original theory accurately described how better processing of information within an organization could benefit task completion, and the increased ability of employees to process information using technology provides a mechanism for improving decision-making within an organization due to the improved ability to process more copious amounts of information farther down within the organization.

Obel and Snow (2014) posited that Dr. Galbraith's theory of organizational design is even more relevant today than it was when Galbraith developed it in 1974 because of the dramatically increasing availability of information in the form of big data. Park, Sawy, and Fiss (2017) highlighted that the ability to process information for organizational decision making had been studied extensively, but little focus has been placed on the details within the IT component of that decision-making process. Cao et al. (2019) articulated that Galbraith's work was later adopted to address decision-making within an organization. Jia, Blome, Sun, Yang, and Zhi (2020) clarified that researchers Tushman and Nadler built upon Galbraith's work by interpreting that said organizations are inherently information- processing systems that are intrinsically- programmed to manage uncertainty by gathering, processing, and acting on information from within their environment. Hwang, Kim, Hur, and Schoenherr (2019) added that through the OIPT Galbraith highlights the essential function of an organizational structure, which is to facilitate the collection, analysis, and distribution of information to reduce uncertainty

within the organization.

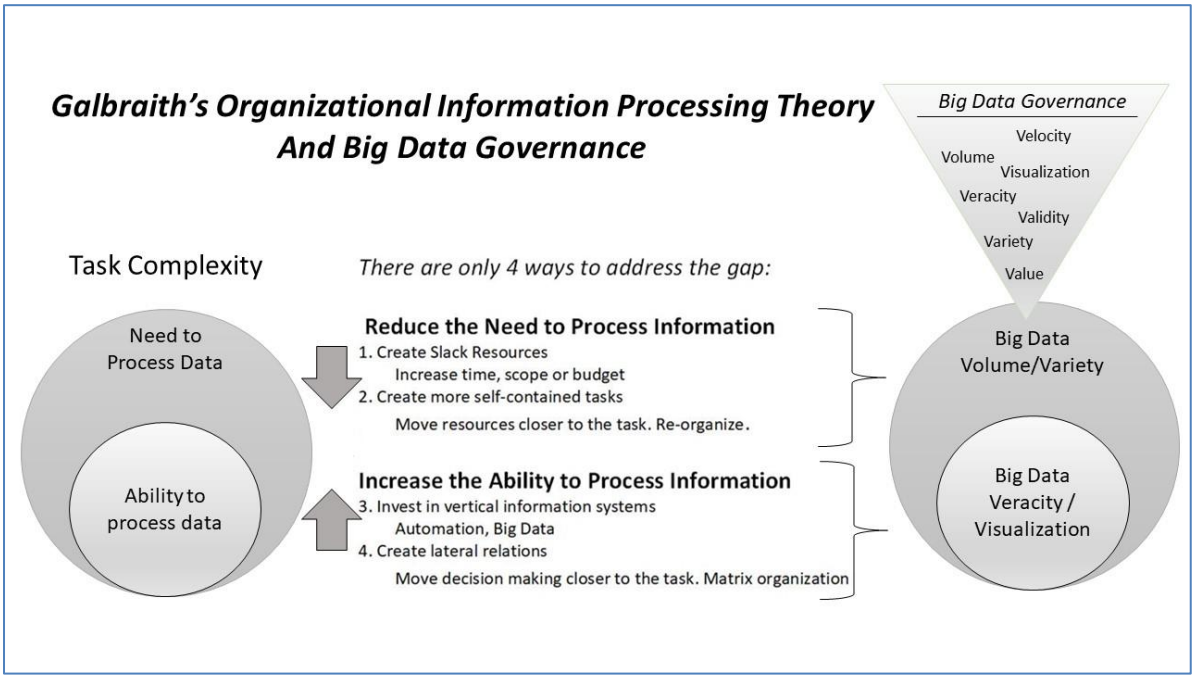

Figure 3. OITP key concepts.

Galbraith (2012) hypothesized that big data would add another dimension to the organizational information processing needs within organizations due to the constant interplay of increasing complexity and interdependence of information and systems within that organization. Zelt, Schmiedel, and vom Brocke (2018) articulated that the application of OIPT leads to an understanding of the factors influencing an organization's information processing capacity. Galbraith (2012) predicted that big data would become that technology that would finally allow for the increased information processing alternative to becoming an effective alternative. Galbraith became fascinated by the way companies were using big data to gain valuable insights into both their organizational decision-making capability and to learn more about their customers (Galbraith, 2017). He believed that big data would become the next significant dimension in organizational

design by providing a digital function similar in power and importance to organizations existing organizational structure (Galbraith, 2014).

The literature identified some limitations of the OIPT. Haußmann, Dwivedi, Venkitachalam, and Williams (2012) identified that Galbraith focused mainly on technical aspects of information processing within an overall organizational design and did not address the individual and interpersonal effects of information processing thoroughly. Kroh et al. (2018) articulated that the OIPT emphasizes structural mechanisms when addressing information flows and task uncertainty. Haußmann et al. (2012) further clarified that the reduction of uncertainty is not the only factor to consider when assessing IT adaptation for improvement of decision making within an organization. Kroh et al. (2018) also stated that information processing capabilities enable organizations to receive external input and translate that into usable information within the organization. Haußmann et al. (2012) clarified that Galbraith mentions the external perspective but does not include this as essential to the overall OIPT concepts.

Despite the limitations, Haußmann et al. (2012) pointed out that the OIPT is considered more than a theoretical model to serve academia and remains a theory with wide adaptation and practical implications within the IS community. The OIPT is very applicable to big data, as shown in figure 3. Before his death, Galbraith was working on another book that was to focus on how big data would affect decision making within organizations (Galbraith, 2017). Obel and Snow (2014) highlighted that his previous articles on big data provided critical insights about improving the scale and speed of dissemination of information within an organization. Galbraith believed that there were three main hurdles to overcome for big data to truly demonstrate valuable insights for an

organization: the first hurdle was the power struggle and friction caused by big data within the current organizational structure. The second hurdle was the creation of a digital information and decision process. The final hurdle was the ability of an organization to invest in digital resources required to develop analytic capabilities across the entire organization (Galbraith, 2017).

Analysis of Supporting Theories

Absorptive capacity theory (ACT). The ACT was introduced in 1990 by Cohen and Levinthal as a firm's ability to identify, consume, transform, and apply valuable external information (Cohen & Levinthal, 1990). The capacity of an organization to assimilate such information is positively related to improved performance (Cohen & Levinthal, 1990). Conceptually, the ACT is similar to the information processing theory (IPT), but the ACT is focused on the organization level while the IPT is focused on the individual level. Butler and Ferlie (2019) identified that Cohen and Levinthal have been criticized for narrowly defining the ACT on R&D only and that subsequent definitions have expanded it to a broader definition centered around the acquisition of knowledge, in general, increasing a firm's overall performance. This lack of clear definition has led to an ambiguous characterization of the ACT in contemporary terms with some scholars focusing on organizational learning, others on innovation, and yet others on resource-based knowledge views of the theory (Chaudhary & Batra, 2018). While the ACT provides a lens for organizational learning from external sources, the OIPT is a more appropriate lens to apply to this study because of the clearly understood definition of the OIPT and the direct connection to decision making.

Organizational information theory. Organizational information theory (OIT) was first published in 1979 by Karl Weick and updated in the second edition of his book published in 2015 (Weick, 2015). One of the central tenets of the OIT was the premise that organizations that operate in an information-rich environment have a great need to make sense of equivocal information (Weick, 2015). Weick defined a process called retrospective sensemaking, which entails seeking to understand what is currently happening by identifying what was done in the past, understanding why those actions were taken, and determining if those actions were adequate for the institution's goals (Weick, Sutcliffe, & Obstfeld, 2005). Weick introduced the concepts of ambiguity and uncertainty and posits that for an organization to progress, they must reduce or eliminate ambiguity and uncertainty (Chadwick & Pawlowski, 2007). This concept of reducing or removing uncertainty is like the central tenet of the OIPT but without the clarity of how to reduce that uncertainty. The OIT is focused more on the concept of sensemaking regarding past decisions made, while the OIPT is more concerned with understanding how to reduce uncertainty by gathering and utilizing more information.

Analysis of Contrasting Theories

Information processing theory (IPT). The IPT was established in 1956 and attempted to identify and explain how information is processed, captured, and retrieved by individuals (Mitchell, Mitchell, & Mitchell, 2009). The IPT attempted to address the way information is acquired, stored, and retrieved from human memory (Mitchell et al., 2009). The central concept of the IPT was that as the volume of information continuously increases, it becomes more difficult for individuals to process due to their limited cognitive processing ability, and they experience information overload (Gao et al., 2018).

IPT focused more on the cognitive capabilities of the human brain when processing information, whereas the OIPT focused more on the impact of information processing within an organization. The IPT may be an appropriate lens to view the creation and dissemination of information from big data on an individual level, but the OIPT is a much more appropriate lens for viewing governance of big data at an organizational level.

Organizational knowledge creation (OKC). Nonaka, Byosiere, Borucki, and Konno (1994) introduced the OKC theory to define how knowledge passes throughout an organization. The OKC is a theory describing knowledge creation that extends both up and down the organizational structure from middle management based on both explicit and tacit communications (Nonaka, Hirose, & Takeda, 2016). Milosevic, Bass, and Combs (2018) explained that the OKC highlights the interrelated phases of knowledge emergence and knowledge formalization within an organization. Kemp (2014) addressed this concept concerning big data when he explained that there is a gap between the amount of data that organizations can gather and the ability of that organization to leverage that information in meaningful ways. Akbar, Baruch, and Tzokas (2017) noted that the OKC focused on the logic of appropriateness, which is a situation-driven response that evolves through socialization and discovery. While applicable to big data, the OKC focused more on the creation of knowledge within an organization, while the OIPT is more focused on the gap that exists between the ability to obtain information and the needed information to complete tasks within the organization.

Big Data Governance

Definition of big data. The term big data is not yet formally defined and was used inconsistently in the literature. The actual definition was still evolving with some

people defining big data by what it is, while others tried to define it by what it does (Gandomi & Haider, 2015), and some defined it by the processes required to leverage big data (Mikalef, Pappas, Krogstie, & Giannakos, 2017). Wang et al. (2016) categorized the various definitions of big data as being oriented around the product, process, cognitive, or social aspects of big data. As a case in point, Wamba et al. (2017) indicated that big data is actionable and delivers sustained value, measured performance and provides a competitive advantage, while Sivarajah et al. (2017) defined big data by the 7 V's of high volume, high variety, low veracity, high velocity, high variability, effective visualization, and high value. Lee (2017) referred to additional dimensions of big data as including complexity and decay, meaning that often, data elements of big data require immediate processing to be of value. Kemp (2014) also referred to the value of the data as a critical element to the definition of big data. Herschel and Miori (2017) stated that big data refers to those activities around capturing, storing, analyzing, and acting upon information that is gathered by both humans and devices, where networks transmit that information.

The most common definition of big data had less to do with the actual amount of data, and more to do with the attributes of those data. Most researchers agreed that this includes the following elements, often referred to as the 5 V's: huge volume, high velocity, wide variety, low veracity, and high value (Jin, Wah, Cheng, & Wang, 2015). Bello-Orgaz et al. (2016) stated that these attributes were first identified in 2001 as the 3 V's, namely: high volume, high velocity, and a wide variety. Gandomi and Haider (2015) added that other attributes of veracity, variability, and value were added later. Song and Zhu (2016) proposed that while variety and veracity are challenging to deal with, the value was by far the most critical dimension of big data. Gartner added a qualifier to this

definition in 2012 that indicated that big data also required new forms of processing to gain insights from those data (Bello-Orgaz et al., 2016). Sivarajah et al. (2017) indicated that the most recent definition now has 7 V's: Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value.

Big data definitions in literature were also commonly used in conjunction with either the size of the data set or with the technology used to enable its utilization. Patel, Roy, Bhattacharyya, and Kim (2017) indicated that big data are anything over gigabytes of data, while others such as Bello-Orgaz et al. (2016) indicated that big data refers to terabytes or petabytes in size. Lee (2017) stated that the threshold for Big Data is a minimum of 1TB but added the qualifying sentiment that the minimum threshold would evolve as technology is developed to handle more copious amounts of data. De Mauro, Greco, and Grimaldi (2015) pointed out that the term big data is also often used in conjunction with the technology, such as Hadoop, that is used to enable the use of big data. Patel et al. (2017) defined big data simply as data sets that are too large to be handled by old traditional methods and technologies, while Bello-Orgaz et al. (2016) defined big data as data sets that are so massive as to extend beyond the ability for traditional database and software tools to manage and analyze them effectively.

A proper definition for big data must extend beyond the size or attributes of the data and should also include the concept of the usefulness of the raw data. Bello-Orgaz et al. (2016) indicated that the definition of big data should include the ability to make sense of those data and exploit their value. De Mauro et al. (2015) articulated that one of the fundamental reasons for big data existing is the ability that it provides to make informed decisions from both structured and unstructured data and failing to include that aspect in

the definition would not provide a full understanding of what big data is intended to be. De Mauro et al. (2015) proposed the following formal definition of big data: “Big data is the information asset characterized by such a high volume, velocity, and variety to require specific technology and analytical methods for its transformational value” (p. 103). Alternatively, Kemp (2014) offered the following definition of big data: “Big data is shorthand for the aggregation, analysis and increasing value of vast exploitable datasets of unstructured and structured digital information” (p. 483). Additionally, Wang et al. (2016) offered the following definition: “Big data start with large-volume, heterogeneous, autonomous sources with distributed and decentralized control and seek to explore complex and evolving relationships among data” (p. 750). Since I will be focusing on the governance aspect of big data, rather than the unique tools required for the analysis of big data, I will use the Wang definition for this study.

Challenges with big data. While there were varying definitions of precisely what constitutes big data, the literature covered the challenges around gathering and analyzing big data in great detail. De Mauro et al. (2015) identified that the combination of structured, unstructured, and semi-structured data is challenging to handle with traditional systems. Yang, Huang, Li, Liu, and Hu (2017) pointed out that due to the heterogeneous and unstructured nature of big data, it is challenging for traditional systems to manage and analyze those data effectively. Grover, Chiang, Liang, and Zhang (2018) estimated that unstructured data make up 95% of big data. Siddiqa et al. (2016) clarified that traditional infrastructures are not designed to handle the distributed computing functionality required to analyze the large quantity and variety of elements found in big data. Lee (2017) also pointed out that managing data at velocity strain contemporary data

center architectures which are not sufficient to sustain the volume and velocity of heterogeneous data elements of both personal and corporate data. Due to the velocity attribute associated with big data, Lee and Kang (2015) stated that contemporary analyzing methods such as storing data elements in a data warehouse for later analysis is no longer sufficient. Lee (2017) also pointed out that big data often must be processed immediately, or the value will be lost, such as in the case of patient monitoring or environmental safety-related systems. The volume and variety attributes of big data make it very difficult to determine the veracity of those data (Matthias, Fouweather, Gregory, & Vernon, 2017).

Complexity and velocity. Identifying and categorizing big data at both velocity and volume poses problems for contemporary systems. Kemp (2014) articulated that identifying the big data engine holistically across the entire enterprise for input, processing, and output is significantly different than what traditional data systems have required. Yang et al. (2017) pointed out that metadata can provide significant benefits for normalizing unstructured data elements for traditional structured data systems, but this brings about another challenge regarding how to automate the generation of metadata for such data elements. The high velocity poses an additional problem of lack of storage capabilities for the transient data as it flows through the analyzing systems. Yang et al. (2017) further described that maintaining traditional storage systems with existing redundancy technologies, such as a redundant array of independent disks (RAID), breaks down at the scale needed for big data.

The complex nature of structured and unstructured heterogeneous data elements is adding strain to existing analytic systems. Jin et al. (2015) articulated that data scientists

currently lack an understanding of the relationship between data complexity and computational complexity as relates to big data processing. Yang et al. (2017) further noted that traditional database management systems lack the scalability for managing and storing the unstructured elements of big data. Jin et al. (2015) pointed out that among industry experts, there is not a good understanding of how to deal with the complexity that comes from complex data structures, complex data types, and complex data patterns inherent with big data. When considering the velocity associated with big data, Yang et al. (2017) pointed out that it is clear that high-dimensional data cannot be processed efficiently within the time constraints required to process big data effectively. Jin et al. (2015) articulated that data complexity is one of the significant challenges with big data as compared to traditional data. Yang et al. (2017) argued that traditional algorithms require structured homogenous data to be effective, and they tend to break down when applied to the heterogeneous environment of big data.

Data quality is also quickly becoming an issue with analyzing big data. Data quality is especially problematic because, as Herschel and Miori (2017) pointed out, organizations tend to have a false sense of confidence in those data simply because of the hype around big data. Yang et al. (2017) stated that due to the heterogeneity and complexity of big data, the accuracy, completeness, redundancy, and consistency of those data become very difficult to manage, thus reducing data quality. Hashem et al. (2015) added that the multiple sources of big data are less verifiable than was the case with traditional data, which calls into question the quality of some of the data. Umer, Kashif, Talib, Sarwar, and Hussain (2017) stated that determining the origin and transformation of data elements contained in big data is one of the primary challenges of big data

currently, and Cervone (2016) added that external data sources are less likely to have internal data stewards or data governance processes established to govern its use within the organization.

Poor data quality can lead to ineffective decisions within an organization. Begg and Caira (2012) argued that poor data quality is disruptive and dangerous to a business. As both the variety and volume of data increase with big data from IoT devices, Lee (2017) pointed out that the quality of those data tends to decrease. High velocity and real-time analysis can create a timing issue, since not all the needed data may be available simultaneously for a specific analysis algorithm. Janssen, van der Voort, and Wahyudi (2017) argued that incomplete data can result in partial analysis, which can paint an inaccurate overall picture. Yang et al. (2017) pointed out that the veracity attribute of big data calls for preprocessing to improve the quality of the data to be processed. This need for real-time processing, in turn, creates an additional challenge regarding the current processing power of existing systems.

Volume and security. The resource requirements associated with the analysis of big data are stretching traditional systems to their computing limits. Yang et al. (2017) articulated that the processing needs for big data currently exceed the capabilities of traditional systems, and Lee and Kang (2015) concluded that the limitation of processing power poses a significant challenge for analyzing big data. Intezari and Gressel (2017) added that to extract information from big data, new techniques, and advanced tools are required. Jin et al. (2015) identified that this is a problem at the architectural level, as the system architecture of traditional systems lacks the operational efficiency to analyze big data effectively. Lee and Kang (2015) added that this problem extends to the software

layer as contemporary event processing engines do not support the parallel computing needs required for managing complex event processing associated with big data. While cloud computing may offer some relief, Yang et al. (2017) pointed out that it also adds another set of challenges, such as networking bandwidth needed for processing such large amounts of data.

Security and privacy are also challenging for big data analytics. Lee (2017) pointed out that big data can include many elements of personal information, which Sivarajah et al. (2017) indicated can raise grave concerns for individuals, companies, and governments, such as location-based information, which is found within many forms of big data. Gao et al. (2018) noted that individuals disclose very personal information on social networks often unconsciously, and Flyverbom, Deibert, and Matten (2019) added that this information now allows corporations to encroach on the private lives of individuals at an unprecedented scale. Metcalf and Crawford (2016) added that there are so many publicly available data sets about individuals that they become highly identifiable when correlated together.

Utilizing big data is potentially damaging to overall security and privacy. Herschel and Miori (2017) found that data sources that had not previously had an issue with privacy may end up threatening privacy once combined with other data sources in the big data construct. Yang et al. (2017) demonstrated that in the cloud computing environment, data owners have significantly less control over data confidentiality and integrity due to the virtualized environment of cloud systems, and Ahmed et al. (2017) added that security concerns are a significant detractor to faster adoption of IoT devices. Carbonell (2016) added that there is a big data divide between people and the information

collected about them through IoT, and pointed out that individuals are rarely granted access to their data.

Big data require changes to traditional security controls. Lee (2017) argued that weak security controls around big data can lead to financial loss and reputational damage. Duncan, Whittington, and Chang (2017) identified that because big data originates from various potentially insecure sources, it is inherently insecure. Yang et al. (2017) stated that traditional methods of encryption break down in the high-velocity and high-volume environment of big data analysis, and performance and scalability start to become an issue when applying contemporary encryption algorithms to big data. Sivarajah et al. (2017) identified that ubiquitous data sources associated with big data magnify the lack of advanced security controls within traditional infrastructure systems.

Privacy issues relating to big data are significant because of the predictive analytical power that is associated with big data. Herschel and Miori (2017) indicated that combining data mining and analytics with big data creates new opportunities for organizations to generate new data with significant predictive accuracy relating to specific consumers. This concept was first publicly documented in a now-famous article in which Target used a teenage girl's purchase history to predict that she was pregnant before she informed anyone (Mai, 2016). Herschel and Miori (2017) pointed out that the Target incident demonstrates the new reality that organizations may be privy to information that consumers never intended to share with them. Baruh and Popescu (2017) argued that the lack of control over information has the potential to create more harm to consumers than the benefits that the consumer might obtain from predictive analytics, such as targeted marketing to meet their current needs. With the proliferation of IoT

sensors, social media, and other big data sources, people today are revealing very personal information daily, often without their knowledge or consent (Mai, 2016), an activity that Jain, Gyanchandani, and Khare (2016) defined as passive data generation. Mehmood, Natgunanathan, Xiang, Hua, and Guo (2016) added that passive data may easily be combined with other publicly available sources to generate new information about the individual, which that person may not have intended to disclose, nor had permitted the organization to obtain or retain.

Ethical issues. Passive data generation leads to significant ethical issues for organizations that analyze big data. Herschel and Miori (2017) posited that big data changes the nature of ethical debates at a fundamental level by redefining the power of information and the extent to which free will can guide users' actions. Carbonell (2016) identified three distinct classes of individuals involved in big data: those who generate the data, those who can collect it, and those who have the expertise to analyze it. Mai (2016) added that the primary ethical issue is not whether to collect the information, but how and when it is ethically responsible for analyzing that information. Carbonell (2016) posited that companies gathering big data are gaining a privileged position that provides them with unique insights into individuals activities of which individuals themselves may not be aware.

The use of big data is also raising new consumer privacy concerns. Herschel and Miori (2017) identified that predictive analysis of passive data can not only generate new insights, but it can also violate the privacy of consumers by re-identification of data sets that had previously been de-identified. Metcalf and Crawford (2016) pointed out that publicly available data can pose significant and severe risks to individuals and

communities when combined with other data sets. One of the first instances of this occurred in 2006 as Mehmood et al. (2016) described when AOL release 20 million search queries that had been de-identified, only to have researcher re-identify those data within just a few days. Metcalf and Crawford (2016) described a similar case where public information regarding New York taxicab rides was combined with other sources to identify private information about specific cab drivers, their religious practices, and even to track celebrities that used the cabs in question.

Proactive efforts to obfuscate data become less effective because of big data. Herschel and Miori (2017) pointed out that re-identification of previously obfuscated data has been documented in such cases as anonymous health information shared between providers, and anonymous dating profiles were combined with public Facebook profiles and facial recognition. In both cases, the consumers thought their data were anonymous, but big data analytics re-identified those data revealing private information that the organization did not have permission to obtain. Rocher, Hendrickx, and de Montjoye (2019) found that 99.98% of Americans could be re-identified with any data set with as few as 15 demographic data elements. Mai (2016) stated that this concept calls into question the perception that has been applied to consumer data privacy since the 1970's that consumers provide informed, rational consent for their data to be used by the organization that they are providing access to their data. Herschel and Miori (2017) explained that organizations analyzing big data have an ethical responsibility to understand how to use the data while protecting the privacy and confidentiality of those data.

The fundamental concept of individual control upon which privacy is based is called into question by the existence of big data. Mai (2016) pointed out that when combining and analyzing individual data points from multiple sources reveals new information about individuals, and those consumers are neither able to restrict or control access to their data. Jain et al. (2016) defined privacy as the ability for a person or group to stop information about themselves from becoming known to individuals or organizations that they have not provided specific authorization to possess that information. Baruh and Popescu (2017) referred to this as privacy as an individual good where users are free to negotiate their acceptable levels of privacy. Herschel and Miori (2017) stated that when secondary data sources can be combined to reverse-engineer private information about consumers, as happened with both Target and AOL, the current concept of privacy control as dictated by consumers is not enough to truly protect individual privacy.

Organizational challenges. In addition to the technical challenges around big data, some companies are also experiencing organizational challenges when introducing big data analytics. Wang and Hajli (2017) explained that to leverage big data initiatives effectively, organizations must be willing to address the managerial and organizational changes required. Kemp (2014) explained that there is a gap between the amount of data that organizations can gather and the ability of that organization to leverage that information in meaningful ways. Lee (2017) pointed out that many big data projects fail to achieve their designed outcomes, thus degrading management confidence in big data and potentially stifling further investment in big data initiatives. Kemp (2014) warned that demand for the rich information that arises from big data analysis is fueling a

bottom-up approach to data governance, which may lack sufficient legal, compliance, and regulatory controls. Kemp (2014) added that a top-down approach to data governance of big data may result in a lack of responsiveness and flexibility.

The demand for big data analytics combined with the new skills needed to analyze big data properly is creating another challenge around skilled talent within the industry. Lee (2017) identified that because of the complexity of analyzing the unstructured and heterogeneous attributes of big data, there is a shortage of qualified data scientists to accomplish this task. Jin et al. (2015) added that contemporary data analysts lack a solid understanding of how to deal with the complexity found with complex big data structures. In a study of 430 firms with advanced analytics capabilities, Lee (2017) identified that approximately 66% indicated that they are not able to employ enough qualified analysts to extract the value from the data that they have gathered. Lee (2017) added that the McKinsey Global Institute estimated that an additional 140,000 – 190,000 big data analysts will be needed just in the U.S. to meet the demand for big data analytics within corporations.

The velocity, complexity, and volume of big data, as well as the decay of the usefulness of those data, pose challenges for disseminating the analysis of big data in ways that can help meet the business need. Gupta et al. (2019) argued that escalating information availability and visibility improves decision making, and Mahdi, Nassar, and Almsafir (2019) asserted that knowledge dissemination increases teamwork. Kemp (2014) warned that getting that information from the data analysis to the people within the organization who need it in a timely and efficient manner remains a significant challenge. Further, Yang et al. (2017) pointed out that due to the heterogeneous attribute

of big data, the visualization of those data in a human-understandable manner poses a significant challenge.

The value of big data is hampered by the lack of big data expertise and governance within the organization. Matthias et al. (2017) identified that effectively utilizing big data to make decisions poses a considerable challenge for companies. Paul, Aithal, and Bhuimali (2018) stated that big data analytics rely on fuzzy logic and inductive statistics, and Matthias et al. (2017) explained that improper or incomplete analysis of big data can lead to misinterpretation of those data, and lead to incorrect business decisions. Herschel and Miori (2017) added that because big data is by definition obtained from multiple sources, and those sources are not always known or validated, those data sets are often incomplete or inaccurate. To be able to access the benefits from the extreme amounts of data available to them, Ahmed et al. (2017) articulated that companies will need to develop and implement new platforms for analyzing the extreme volumes of data generated by connected devices.

Opportunities for decision-making. One of the more important themes in the literature involved how big data can affect decision making. These decision-making opportunities include the IoT, the impact of big data on decision making, and big data analytics. Each of these are described in the following sections.

IoT and big data. One of the primary contributors in the context of big data is the IoT. Lee (2017) pointed out that in 2016, 5.5 million new devices were connected each day to the Internet to share, process, and analyze information, and that number was expected to grow to 20.8 billion by 2020. Ahmed et al. (2017) stated that the number of human beings globally has now been surpassed by the number of connected devices, and

the overall number of connected devices was expected to reach 50 billion by 2020, and Herschel and Miori (2017) added that the number of network connections was expected to reach 18.9 billion, which amounts to approximately 2.5 connections for every person on earth. The Organisation for Economic Cooperation and Development estimated that by 2020, the average family of four would have 50 internet-connected devices (Government Accountability Office, 2017). Bilal, Khalid, Erbad, and Khan (2018) warned that IoT data analysis is driving new challenging requirements for traditional systems due to the high degree of mobility, real-time data analytics, and extreme latency sensitivity.

One of the results of the proliferation of IoT devices is the phenomenon of merging the physical and virtual worlds. Because of the amount of data generated by IoT devices, Paul et al. (2018) explained that some data scientists claimed that big data provides a direct link between cyber and physical systems. Dourish and Cruz (2018) utilized the term datafication to describe the transformation of physical, social action into quantifiable digital data, which enables real-time predictive analytics. Flyverbom et al. (2019) added that datafication implies that social lives have a digital counterpart in the form of digital traces left by merely participating in social networking. Galliers, Newell, Shanks, and Topi (2017) articulated that one of the critical issues with this collision of physical and virtual worlds is the fact that the minutia of everyday life suddenly becomes part of big data that is then algorithmically assessed without further human interaction, often without the individual's knowledge or permission. Herschel and Miori (2017) added that the analysis of this minutia leads to organizations generating new insights into individual consumer's behaviors and activities, often without the consumer's consent.

The amount of new information generated daily continues to increase. Lee and Kang (2015) explained that extreme amounts of data were already being generated even without considering the IoT, as evidenced by Google which was generating about 25 petabytes of data daily, and Sivarajah et al. (2017) highlighted that estimates are that globally 2.5 quintillion bytes of data are generated per day. Bughin (2016) pointed out that various reports predict that data creation would continue to increase by between 40% and 60% per year. Herschel and Miori (2017) stated that projections indicated that the total volume of data would increase to 40 zettabytes by 2020.

Analytics of those vast data sets can have significant predictive power. The Government Accountability Office (2017) identified that new big data analytics technologies would be capable of extracting hidden patterns and correlations from these extensive data sets. Ahmed et al. (2017) stated that combining IoT and big data analytics promises to improve decision making within businesses significantly. Carbonell (2016) described the power shift that arises as organizations start utilizing big data to predict future events and even intervene before those events are set in motion. Chauhan and Sangwan (2017) added that the advantage of big data over traditional data analysis is better accuracy, which leads to higher confidence in data-driven decisions. This confidence comes because, as Wang and Hajli (2017) identified, big data analytic systems provide insights into hidden correlations between data elements that were previously impossible to determine with traditional data analytics tools. The Government Accountability Office (2017) predicted that businesses that could analyze data from IoT devices could build competitive advantages over their competition. In the manufacturing industry, for example, Popovič, Hackney, Tassabehji, and Castelli (2018) identified

multiple case studies that revealed that big data analytics has already improved manufacturing operations due to the insights gained from that analysis. Baruh and Popescu (2017) highlighted one example of this change is the fact that automobile insurance companies have leveraged big data from GPS sensors to shift from a risk-based model to a habit-based model using real-time monitoring of individual consumers.

Impact on decision-making. Harnessing predictive analytics with big data provides significant competitive advantages. Intezari and Gressel (2017) articulated that the success of an organization is primarily affected by the competitive ability of managers to make strategic decisions in the face of uncertainty and ambiguity. Braganza, Brooks, Nepelski, Ali, and Moro (2017) predicted that the economic value of organizations could be elevated by effectively utilizing the deeper insights available from big data. Cao et al. (2019) found that by improving information processing capabilities, the decision-making effectiveness within the firm increased and created competitive advantage, especially when those capabilities were rare, valuable, and inimitable. Grover et al. (2018) agreed and clarified that big data analytics were likely inimitable because they are so closely tied to the decision-making culture and leadership within the firm. The Government Accountability Office (2017) added that decision making could be enhanced by analyzing aggregated data from IoT devices. Popovič et al. (2018) added that the distribution of big data analytics within organizations had a direct effect on improved business decisions within the organization.

Competitive advantage from predictive analytics derives from the ability of organizations to quickly adjust to trends uncovered by those analytics. Ahmed et al. (2017) explained that companies can use big data analytics to drive consumer behavior

by watching for real-time trends, such as an empty parking lot, and reacting in real-time to promote specials that bring customers into the store. The Government Accountability Office (2017) suggested that advanced data analytics would be able to draw conclusions automatically from various discreet data sources what used to require human intervention, such as suggesting alternative routes in real-time due to traffic patterns, weather, and other conditions. Lee and Kang (2015) further added that the availability of personal location information from big data, such as real-time commuter information, was projected to save $600 billion annually by 2020.

Big data analytics. Big data analytics and effective decision making are closely correlated. Braganza et al. (2017) articulated that effective analysis of big data enabled more evidence-based decisions, while Wamba et al. (2017) added that big data enabled data-driven decision making. Paul et al. (2018) further clarified that analyzing those data had become a critical need for every business sector. Janssen et al. (2017) stated that effectively analyzing big data had the potential to change the way that organizations make decisions fundamentally. Intezari and Gressel (2017) articulated that due to the uncertainty and ambiguity related to strategic decision-making organizations needed to reconfigure and reassess their knowledge base continuously. Zakir et al. (2015) observed that companies that oriented their decision making around fact-based data analytics outperformed their counterparts in the market. According to Jin et al. (2015), big data has had a significant impact on almost every business sector and significant industry already, and to remain competitive, Lee (2017) stated that all companies would need to build big data expertise. Günther, Rezazade Mehrizi, Huysman, and Feldberg (2017) pointed out that there is quite a bit of evidence currently that big data can have a significant impact

on an organization's business models. For organizations to extract value from big data, Günther et al. (2017) stated that it is critical for them to continually realign work practices, organizational designs, and stakeholder interests.

Organizations are still struggling to understand how to effectively utilize big data analytics to guide real-time decisions. According to Patel et al. (2017), big data was the most trending contemporary topic in the computer science field. Wang et al. (2016) pointed out that between the years 2000 and 2016, there were a total of 7,121 publications relating to big data, including 2,924 articles published in over 1,000 different journals. Matthias et al. (2017) discussed multiple case studies of organizations utilizing big data analytics to provide evidence that significant operational insights are now obtainable from big data that were previously unavailable from traditional analytics. Janssen et al. (2017) posited that the real value of big data is found in its potential to improve decision quality. Gandomi and Haider (2015) identified that one of the most prolific uses of big data is predictive analytics, where data analysts use a variety of techniques to attempt to predict future outcomes based on patterns found within big data. Janssen et al. (2017) warned that care must be given, however, since decision quality may be degraded if the decision-maker does not understand the relationships between the various data elements used as part of the analysis.

Organizational and employment impact of big data. It is not yet clear how to most effectively organize to best utilize big data within the organization. Günther et al. (2017) discussed the fact that scholars were actively debating what the most appropriate organizational model would be to most effectively take advantage of big data analytics. According to Janssen et al. (2017), the collection and processing of big data were quite

often accomplished across departments and even across various companies, requiring collaboration and partnerships that do not exist with traditional data. Braganza et al. (2017) added that this lack of role clarity across the organization was inhibiting the ability to use required big data resources efficiently.

The cross-departmental collaboration required for big data to be successful causes challenges for contemporary firms. Because of the value and direct relationship between big data analytics and improved quality of decisions within the organization, Popovič et al. (2018) identified that power shifts had been documented within organizations that are effectively utilizing big data analytics to make fact-based and real-time decisions. Braganza et al. (2017) added that big data processes required new roles and organizations to interact, which had not previously needed to work together. While there was no consensus regarding the impact of big data analytics from IoT on employment, the Government Accountability Office (2017) predicted that it was likely there would be a combination of both new job creation and job loss.

The need for new organizational behavior also brings the need for new governance processes. Ahmed et al. (2017) articulated that integrating and analyzing big data were revolutionizing business processes by turning big data into valuable insights not available with traditional technology. Janssen et al. (2017) added that the quality of decisions from big data was not influenced solely by the amount of data, but more importantly, by how those data were collected and processed. Matthias et al. (2017) pointed out that without appropriate analysis, it did not matter that large amounts of data were collected. De Mauro, Greco, Grimaldi, and Ritala (2018) agreed that companies needed to acquire the right expertise to deal with these technological advances. As stated

by Ahmed et al. (2017), value was extracted from big data when information was transparent and usable to the organization.

Cloud computing. The adoption of cloud computing capabilities enhances the adoption of big data within organizations. Hashem et al. (2015) articulated that big data and cloud computing are conjoined technologies. Because IoT devices are generating much of the data, Marjan et al. (2017) highlighted that cloud computing is expected to provide lightning-fast analysis of fast-moving data. Nadjaran Toosi, Sinnott, and Buyya (2018) added that cloud computing systems are the preferred architecture for hosting these data-intensive applications. Although big data systems are lagging behind IoT development, Marjan et al. (2017) clarified that the two should be developed simultaneously due to their inter-dependence. Ahmed et al. (2017) added that combining IoT with big data analytics allows for the processing of data closer to the source of that data.

Cloud computing architecture provides the foundation for big data analytics. According to Zakir et al. (2015), cloud storage systems offered capabilities that were not available in their on-premises counterparts, such as the pay-as-you-go payment model, the elasticity of capacity, and risk mitigation through business continuity and long-term retention. Varghese and Buyya (2018) pointed out that data centers house most systems that encompass cloud computing, yet Stergiou, Psannis, Kim and Gupta (2018) highlighted that the intensive computations required to analyze big data, and the mass storage needed to support it, were often inefficient even in contemporary cloud systems. According to Marjan et al. (2017), big data analytics required the same or faster-

processing speed of traditional analytic systems at similar or less cost and operating on much more volatile data sets.

Contemporary cloud systems are still lagging behind the requirements for big data analytics. Nadjaran Toosi et al. (2018) identified that one of the most challenging scaling issues relating to contemporary data analysis solutions is the location of the data relative to the computing resources, with transfer latency of data often exceeding the computational analysis time. Bilal et al. (2018) added that the IoT is creating new challenges for cloud architectures when dealing with data systems due to the latency associated with these large data sets. Stergiou et al. (2018) suggested that significant changes to the hardware and software architecture will be needed in support of a new generation of cloud services to support the real-time analytics of data from the IoT. Varghese and Buyya (2018) suggested that these changes include the decentralization of data center cloud services by incorporating edge devices in cloud computing architecture. Most contemporary big data cloud solutions assumed a centralized data center cloud services model, but Varghese and Buyya (2018) suggested that soon, these architectures would need to include the use of edge computing such as cloudlets, fog computing, and mobile cloud computing.

IT governance. Without proper governance practices, technological solutions to big data problems will not be enough to solve the current problems. Watson and McGivern (2016) noted that business intelligence initiatives had a low chance of success without governance practices that provide access to high-quality, secure, and well-modeled data. Braganza et al. (2017) added that extracting value from big data requires not only the appropriate technological solutions but also requires that big data analysis

becomes part of the fabric of an organization. According to Alreemy, Chang, Walters, and Wills (2016), IT Governance played a critical role in creating this fabric by establishing a link between the business and IT that has the potential to provide a competitive advantage by minimizing expenditures and making better use of time.

Repeatable and efficient processes are required to extract value from any data initiative. Wilkin, Couchman, Sohal, and Zutshi (2016) articulated that IT governance is the set of systems that direct and control IT operations to meet stakeholder expectations regarding financial and environmental operations. Alreemy et al. (2016) defined IT Governance as the process that is established to control investment in IT and to guide those investments toward achieving business objectives. Wu, Straub, and Liang (2015) noted that efficient utilization of IT within an organization requires effective IT governance. Braganza et al. (2017) stated that the repeatability of big data processes requires reliable and sustainable business processes.

Effective IT governance is critical for IT organizations to effectively meet stakeholder expectations in all areas of information systems. Wu et al. (2015) stated that the single most critical predictor of value from an IT organization was an effective IT governance process. Braganza et al. (2017) suggested that effective business processes could release value from big data initiatives that have, as of yet, been challenging to reproduce consistently. Wu et al. (2015) identified that IT Governance was critical due to the significant investment that is generally associated with IT within an organization. Mahdi et al. (2019) added that the processes within an organization that create, share, and use knowledge were highly critical strategic capabilities that create a competitive advantage to the organization.

IT organizations need to implement formal IT governance practices to align outcomes with stakeholder expectations. According to Alreemy et al. (2016), the most important critical success factor for IT governance is the alignment between the business and IT. Wilkin et al. (2016) noted that alignment between IT and business initiatives was the most widely realized benefit of IT governance processes. Alreemy et al. (2016) articulated that appropriately implemented IT governance practices could mitigate business risks. Wilkin et al. (2016) added that business alignment and risk mitigation were the primary driving forces for implementing formal IT governance practices.

Smaller organizations are less likely to formalize their governance practices than their larger competitors. Although both smaller and larger organizations understand the strategic need for IT governance, Wilkin et al. (2016) noted that larger organizations were more likely to implement formal IT governance practices than smaller organizations. Wilkin et al. (2016) added that smaller organizations faced significant challenges in managing technology assets so that they could sustain a competitive advantage. Begg and Caira (2012) also added those smaller companies tended to hold the belief that they must first implement better overall IT governance before they can begin implementing data governance.

Leadership involvement with IT governance of big data. Contemporary IT leaders lack strategies for managing big data initiatives. Braganza et al. (2017) noted that there did not appear to be evidence for the assumption within the literature that processes for managing big data exist and that resources are well managed. Mikalef et al. (2017) added that research had focused on big data technologies for data analysis, storage, and visualization, but there has been a gap of research regarding how organizations would

need to adjust to embrace these technological innovations. Braganza et al. (2017) stated that senior leadership needed processes that they could implement to extract strategic value from big data. When addressing the challenges associated with big data analysis, Begg and Caira (2012) noted that processes need to be very flexible and comprised of various roles, decision areas, and assignments of specific responsibilities, such as technical data steward, business data steward, executive sponsor, and other such roles. Braganza et al. (2017) articulated that the value of big data would continue to be limited until businesses could deliver repeated benefits from within the same organization over time and further clarified that Big data processes needed to be flexible enough to change due to variations in both internal and external forces.

Data governance. In addition to IT governance, organizations also require specific data governance practices to effectively manage their big data initiatives. To be able to extract value from raw data, Song and Zhu (2016) noted that organizations require a well-defined, systematic approach to governing their data. Mikalef et al. (2017) clarified that Data Governance was the term used to describe these activities, which included roles, structures, and decision mechanisms around IT resources. However, Begg and Caira (2012) articulated that there was no evidence to date of a standard implementation framework for data governance. Hashem et al. (2015) noted that policies, principles, and frameworks around how big data are used and leveraged within an organization were posing enormous challenges for those organizations. Posavec and Krajnovic (2016) recognized that the lack of formal data governance processes was one of the primary reasons that organizations failed to utilize their data effectively.

The governance of big data is critical to the success of any big data initiative. Cervone (2016) articulated that while data governance was crucial in most environments, it was critical for big data, and organizations could not add it as an afterthought. Wang, Kung, and Byrd (2018) identified that the governance of big data refers to the *how-to* aspects of harnessing the data within the organization. Cheng, Li, Gao, and Liu (2017) added that data governance referred to the series of policies that an organization used to define cloud data strategy, management, operations, and optimization. Chrisman et al. (2018) further clarified that governance referred to both formal and informal rules, practices, and processes that were implemented to direct and control behavior in a consistent manner throughout the organization. Wang et al. (2016) described information fusion as the merging of information from various heterogeneous sources toward a consistently accurate and useful representation. They also clarified that a clear strategy was required for this process to be successful.

Effective data governance can lead to improved business outcomes. When organizations implement governance well, Wang and Hajli (2017) identified that big data initiatives had the potential to provide insights into hidden data correlations for meaningful decision making. Wang et al. (2018) added that proper data governance addressed the standardization, accuracy, and availability of those data. When the organization implemented governance poorly, Wang and Hajli (2017) noted that organizations experienced substantial financial costs for little value. By effectively applying governance principles to data governance initiatives, Wang et al. (2018) found that organizations could better meet organizational goals with lower costs and shorter

timelines. Mahdi et al. (2019) added that to be successful at all, organizations must methodologically exploit their knowledge assets.

Effective big data governance practices are not easy to implement but can have a significant impact on the overall business. Hashem et al. (2015) noted that adopting big data governance practices that provide a balance between value creation and risk mitigation was imperative to organizations. Mahdi et al. (2019) found that knowledge management practices within an organization were significantly and positively related to sustainable competitive advantage. Cheng et al. (2017) added that big data governance was critical to ensure that data are accurate, shared appropriately, and protected per company policies. Al-Ruithe, Benkhelifa, and Hameed (2016) further added that data governance was also vital to demonstrate the ability to manage and monitor data in the cloud environment where the organization does not control the systems and services. Due to the high volume and velocity of big data, Siddiqa et al. (2016) stated that key governance elements such as integrity, auditability, accountability, stewardship, and change management become extremely challenging.

Effective data governance within the IT organization can help elevate the IT organization to more strategic alignment with the rest of the business. As more organizations rely on data as a critical differentiating asset, Watson and McGivern (2016) identified that governance around the data becomes more critical to the success of the company, which in turn shifts IT and business intelligence into a more strategic alignment with business initiatives. Shao (2019) added that alignment between overall information systems strategy and business strategy is a critical precursor to overall profitability and competitive advantage. Data governance refers to the people, processes

and technologies that are used to leverage data to create value for the organization by managing, protecting and using data (Watson & McGivern, 2016), as well as the control and authority over data-related rules and accountability of individuals and systems accessing and utilizing those data (Hashem, et al., 2015).

Small companies are slower to recognize the strategic value of data as compared to their larger competitors. Begg and Caira (2012) found that small companies did not appear to have the same understanding of the inherent value of data as to their larger counterparts. Wilkin et al. (2016) added resource constraints in smaller organizations tended to limit the organization's ability to be competitive in areas such as IT competencies, knowledge management, and innovation. Begg and Caira (2012) found that many small companies still viewed data as a means to an end, as opposed to an asset with inherent value, and they were hesitant to undertake data governance practices out of concern for inappropriate data retirement or other possibilities for disruption of live data needed for operations.

Small businesses face larger hurdles to effectively utilizing big data than larger organizations. Begg and Caira (2012) predicted that data governance would become an increasingly critical issue for small businesses as vendors and providers forced them into a cloud business model. When considering the cost of making decisions from bad data, Watson and McGivern (2016) stated that the cost of data governance was becoming a foundational cost of doing business in today’s world. However, according to Wilkin et al. (2016) IT governance competencies were limited in smaller organizations because of their reduced capacity to identify, assess, and acquire IT talent, knowledge, and skill set.

To remain competitive, Begg and Caira (2012) added that small businesses would soon be forced to adopt data governance initiatives.

Roles in big data governance. The lack of skilled and knowledgeable personnel is of even greater concern than the technical challenges surrounding the adoption of big data within an organization. Technology is actively being developed to fill the gaps identified in big data analysis. However, Song and Zhu (2016) pointed out that there was a recognized shortage of people who can think critically about big data and who knew how to use these new technologies to extract value from those data. Zakir et al. (2015) added that this required a skill set that was new to most IT organizations. One such new role in the industry was the Data Scientist, which Song and Zhu (2016) identified as an emerging role where people were tackling these big data challenges and was considered one of the top 25 jobs in America in 2016. However, Hu, Luo, Wen, Ong, and Zhang (2018) identified that there was no agreed-upon definition of data scientist across different organizations and Gardiner, Aasheim, Rutner, and Williams (2018) warned that it was challenging to categorize job functions across different companies based only on job title due to the very loose interpretation of that role at each company. De Mauro et al. (2018) pointed out that organizations were still defining the job description of Data Scientist, and this role currently had unrealistically high expectations for managing big data. De Mauro et al. (2018) added that extracting value from big data required more than a single role working together across multiple disciplines.

Because big data initiatives intersect with so many disciplines it is a challenge to find all of the needed skills in a single person or role. Gardiner et al. (2018) found that data scientists were required to possess a wider variety of skills than traditional

technologists due to the multi-faceted nature of this new role. De Mauro et al. (2018) stipulated that the main focus of the role was around the data itself, and the associated analytical methods for transforming those data into usable insights. However, in their analysis of contemporary, big data job descriptions, Börner et al. (2018) indicated that soft skills such as collaboration, communication, and writing and problem-solving were all required for the role of a data scientist. Chen and Zhang (2017) also found skills relating to social sciences among the most common qualifications listed for big data roles. Gardiner et al. (2018) identified that contemporary data scientists required both analytical and soft skills.

Organizations are currently scrambling to identify individuals with this unique confluence of skills. Börner et al. (2018) identified 69,405 jobs posted for data scientists or data engineers within the six years between 2010 and 2016. Gardiner et al. (2018) stated that McKinsey previously projected that by 2018, there would be a 50% - 60% gap between the supply and demand for big data analytics talent. As De Mauro et al. (2018) pointed out, the deep analytical expertise required of most data scientists was not requisite for maintaining the real competitive advantage expected from big data. Gardiner et al. (2018) added that the skills required to do so include such diversity as mathematics, probability, statistics, programming, data management, data visualization, data mapping, to name just a few.

Relationship of the Study with Previous Research

Existing literature on the topic of big data primarily focuses on two overarching topics: the analytical value of big data and the challenges associated with utilizing big data as a source of information. These two high-level themes categorize the various

studies and other literature described in the previous sections. The literature regarding data governance does not apply solely to big data, but rather the available studies focus on the benefits of data governance in general terms and on the benefits of data governance in general. This study adds to the body of research by focusing specifically on the implementation of traditional data governance principles when utilizing big data as the data source for analytics for decision making within an organization.

Transition and Summary

In this section, I defined my research by introducing both the problem and purpose statements and introduced my research question and the supporting interview questions that I used in the interview process. I defined different terms that will be referenced in this study and explained the limitations, delimitations, and assumptions that apply to this study. I further demonstrated that big data analytics brings with it both high expectations for a positive impact on business decisions as well as significant challenges in the overall governance of those data. In the review of the professional and academic literature, I described how Galbraith's work on the OIPT set a solid foundation for the study of big data in general. I introduced the OIPT and explained how this theory was utilized to guide this study. The literature review provided an overview of big data in general, as well as some of the critical challenges that arise from big data analysis. I discussed the topics of security, the IoT, cloud computing, organizational impacts of big data analysis, and IT governance. The literature review culminated with a discussion of data governance and its implementation within mid-size organizations. The following section identifies the method for performing the study, including the role of the

researcher, ethical considerations, and other vital elements and details around the application of the study itself.

Section 2: The Project

This section of the paper contains information about the participants, sample, population, and research methodology. I provide justifications for my decisions regarding the design of the study and address ethical considerations. I also describe my approach for data collection and analysis.

Purpose Statement

The purpose of this qualitative multiple case study was to explore big data governance strategies employed by data scientists to provide a holistic perspective of those data for making decisions. The population for this study was data scientists employed in three mid-market companies in the greater Salt Lake City, Utah area who have strategies to govern big data. The findings from this study could potentially affect social change by creating additional jobs in mid-market companies and enabling small and medium-size companies to better cater to the needs of their specific customers. Job creation would be accomplished through increased capability of these companies to have a better understanding of data which they gather and use to make strategic and operational business decisions daily. This increased and more targeted information could lead to increased profitability for these mid-market companies, which in turn could lead to increased jobs. Customers would benefit because local small and medium-sized businesses would now be able to customize their offerings to specifically meet the needs of their customers in a similar way that large companies with much more extensive research & development and marketing budgets do.

Role of the Researcher

Qualitative research dictates that the researcher is the principal data collection instrument, and as such, the researcher maintains as essential a role in the research as the participants themselves (Draper, 2015). As the primary instrument for data collection, the researcher is not only an assessor or observer but also the primary analyst for interpretation and synthesis of data (Cook et al., 2016). My role as the sole researcher in this study was to perform interviews, gather data, and analyze and present those data in an unbiased manner. Kelly (2017) explained that in this role, it is crucial to identify experiences, history, and knowledge to the reader of the study to uncover and disclose any biases. This is essential so the reader can better comprehend my biases, as well as where I can apply my expertise within the context of the study. Since 1988, I have worked in various roles within IT, with much of my early career as a developer and enterprise architect. In the latter part of my career, I have worked as Chief Information Officer in various regulated industries, and have focused on process improvement and interpersonal management of technology employees in various roles within IT. My experience has not been with big data or data analytics specifically, although I have managed teams responsible for data analysis. I consider myself unbiased concerning the governance of big data. I have spent my entire professional career in the same general geographical area where I conducted this study. As data from multiple sources were analyzed, it was my responsibility to search for similarities and differences among those data while making sure to avoid any bias.

To mitigate bias, the researcher must strive to be reflexive throughout the study, continuously asking if the results are consistent with the research question and whether

the researcher has conducted the analysis fairly (Kelly, 2017). The researcher must keep records and transcriptions of interviews, as well as documentation of their feelings and impressions from interview sessions (Kelly, 2017). While performing interviews, I kept this in mind and continually sought fairness as I conducted the interviews and listened to responses. I thoroughly documented my feelings and impressions as well as respondents' answers. I also employed member checking to ensure that I accurately interpreted participants' responses. I watched the body language of participants to identify any questions that may have made them uncomfortable and to attempt to identify any areas of potential participant bias.

Studies involving human participants require ethical approval (Kelly, 2017). Ethical considerations require that no harm will come to participants as a result of their participation in this research (Kelly, 2017). The Belmont Report outlines ethical principles and guidelines for protecting human subjects while conducting research, explicitly identifying beneficence, justice, and respect for all persons in the study (U.S. Department of Health & Human Services, 1979). I achieved beneficence, justice, and respect for all participants by treating all participants fairly and with respect and both clarifying and protecting information they provide to me as part of this study. I adhered to the process outlined in the ethical research section of this study.

Mitigating bias during data collection includes carefully planning interactions with participants. An interview protocol can assist in the development of an inquiry-based conversation (Castillo-Montoya, 2016). The researcher should make the process of data collection as explicit as possible by establishing a formal interview setting (Ponelis, 2015). Semi-structured interviews provide a means for the researcher to ask specific

questions to elicit open-ended responses that allow participants to guide the dialogue in ways that the researcher could not have anticipated a priori (Brown & Danaher, 2019). Interview protocol should be structured to ask multiple participants the same starting questions as a means to maintain a steady target for data saturation and mitigation of bias (Fusch & Ness, 2015). Because interactions between the researcher and participants add value to the research, the researcher can strategically arrange the interview setting so participants feel invited or authorized to share interpretive knowledge that would not result otherwise (Malli & Sackl-Sharif, 2015). I worked to ensure that any potential biases were mitigated by following a planned interview protocol (see Appendix B) and watching for participant reactions throughout the semi-structured interview process.

Semi-structured interviews provide a setting for thoughtful reflexivity regarding the overall research question (Brown & Danaher, 2019). The research design must link the research question and overall purpose to processes for data collection and analysis to draw conclusions from those data (Ponelis, 2015). The data collection protocols include the types of people who will be interviewed, documents that must be analyzed, and observations that the researcher must make (Yin, 2017). While the researcher should make the data collection process as explicit as possible, a semi-structured approach is extremely valuable to allow for the exploration of new information that will arise during the interview process (Ponelis, 2015).The interviewer needs to ask well-structured questions while listening to and interpreting answers to find a balance between passively listening and overdirecting the interview (Ponelis, 2015). I followed a semi-structured interview approach by asking questions as outlined in Appendix B and interacted with participants by asking for more elaboration when responses involved new data related to

the research question (see Appendix B). I maintained field notes, including observations during the interview process.

Participants

Participants of qualitative studies are typically selected based on their involvement with the topic being studied based on the research question (Fink, 2000) and their ability to provide detailed descriptions regarding the study topic (Draper, 2015). Well-informed interviewees can provide insights into history and decision-making that influence organizations to end up with models that are currently in use within their organizations (Yin, 2017). Participants were individuals who were employees or contractors working within the partner organizations and had intimate knowledge of data management and analytical strategies of the organizations, had been in a data scientist type of role for a minimum of 5 years, were data analytic solutions architects at these companies, and had no recurring working relationship with me.

A gatekeeper can help the researcher by providing access to key individuals within an organization (Fusch & Ness, 2015). I used a gatekeeper within each organization to assist in identifying participants and selecting appropriately private locations for interviews. I met with these individuals to discuss the research question and overall strategy for completing this study, and we reviewed criteria for participation together. I used these interactions to learn about the organization as well as help clarify participant criteria, and I selected participants that met the criteria for this study based on this information. I then directly contacted participants to obtain their informed consent and schedule interviews with them.

It is critical for the researcher to both build trust and to establish a rapport with participants when using a semi-structured interview process in a qualitative method (Brown & Danaher, 2019). Rapport comes as the researcher and participant move in and out of empathetic moments during the interview process (Prior, 2017). Response bias can arise from participants when rapport leads to behaviors that support the researcher's perceived beliefs (Latkin et al., 2016). I did not select participants with whom I had a prior relationship to mitigate response bias, but rather, I established rapport with participants by learning about the culture of their organization before meeting with them and adapting to their dress and behavioral standards. I used video conferencing technology to allow the participants to choose a location for the interviews that simultaneously allowed the participant to be in a comfortable location while still maintaining their privacy. After obtaining their agreement with the informed consent form, I engaged with the participants and briefly explained to them my background, history, and experience with data governance and big data. I did not go into too much detail to mitigate any potential bias. Finally, I ensured that all participants understood that my role was to gather information and not to judge the validity of their architectural decisions.

Research Method and Design

Research Method

A qualitative method was most appropriate for this study. The central defining capability of the qualitative approach is to answer how and why questions (Ponelis, 2015). These studies are particularly applicable to applied disciplines since the findings are more easily applied to improve specific processes, programs, or problems in practice

(Ponelis, 2015). Qualitative studies require the researcher to possess both emotional maturity and solid interpersonal skills to recount the experiences of the participants accurately in their own words (Collins & Cooper, 2014). Qualitative studies are most useful when investigating how well people perform something (Kelly, 2017). Because this study required an in-depth exploration of the strategies and methods employed by data scientists, a qualitative approach was most appropriate.

Quantitative studies are focused more on correlations and trends than on providing explanations or reasons (Kelly, 2017). Quantitative studies are used when the study intends to predict outcomes or explain theories which are based on cause-and-effect relationships and to determine the correlation between variables (Gehman et al., 2018). Quantitative studies are not as capable of providing explanations for processes (Kelly, 2017). Since there was no hypothesis associated with this study, a quantitative approach was not the most appropriate methodology. If this study had been investigating correlations between big data analysis and specific outcomes, then a quantitative approach may have been more appropriate. I considered using a quantitative approach but deemed a qualitative approach to be more appropriate due to the research question for this study.

Mixed method studies are completed by using both qualitative and quantitative data in the same study (Halcomb & Hickman, 2015). A researcher could employ a mixed-method design by performing the qualitative and quantitative methods of research in either a sequential or concurrent process (Imran & Yusoff, 2015). A researcher should ask themselves whether the mixed-method approach will add more value than using a single research method (McKim, 2017). Adding a quantitative design element to my

study would not have yielded any additional insights into the research question since there are no statistical or quantitative elements to be measured and no hypothesis to be tested. Since a quantitative approach alone was not appropriate for this topic then adding a quantitative element by employing a mixed-method approach was also not inappropriate for this study.

Research Design

I used a multiple case study methodology for this study. A case study methodology allows for a deep examination of an issue or phenomenon in its natural setting (Arseven, 2018). Because this study involved an exploration of strategies used within organizations that are using big data for decision making, this design approach was warranted. This design methodology is most useful when situations or phenomena are investigated for exploration to gain specific insights or knowledge (Elman, Gerring, & Mahoney, 2016). The case study methodology also allows the researcher to observe in person the operation of the governing body to witness the various attributes of the successful data governance program in the mid-market organization. This opportunity for personal observation was critical to this study because considerable research has been conducted regarding information technology governance programs but there is no agreement on what specific attributes constitute an effective information technology governance program (Alreemy et al., 2016). It was important to use the case study design to personally observe and document the strategies used by data scientists using big data. A case study design is appropriate when the study is intended to be inductive and attempts to make broad generalizations (Kelly, 2017). A multiple case study tends to be more generalizable than a single case study would be (Gehman et al., 2018). Because this

study was investigating different strategies used by data scientists a multiple case study was more appropriate for uncovering strategies that could be generalized to other organizations.

An ethnographic design would require embedding the researcher in a cultural context to observe specific behaviors of a group of people in a specific cultural context (Mannay & Morgan, 2015; Watson, 2018). The researcher becomes part of the participant group in ethnographic design (Kassan et al., 2018). An ethnographic approach would have been appropriate if this study intended to identify the specific culture, norms, or customs of data scientists. Since I studied the strategies used by data scientists and not their norms or customs an ethnographic design was not appropriate.

The phenomenological design focuses on how the participants live through, interpret, or experience the topic under research (Adams & van Manen, 2017; Sousa, 2014). A phenomenological design is intended to explore how the participants make sense of a phenomenon (Adams & van Manen, 2017). Since this study was not investigating the perceptions of participants, the phenomenological approach was not appropriate. In this study, I investigated the strategies employed by data scientists and I was not looking specifically at their reaction to a specific phenomenon. The phenomenological design was not appropriate for this study.

Data saturation is achieved when no new information is being uncovered, when there is enough information to replicate the study, and when further coding is not feasible (Fusch & Ness, 2015; Lowe, Norris, Farris, & Babbage, 2018). Data saturation means justifying that the amount of data obtained is not more or less than the data needed for research (Saunders et al., 2018). To achieve data saturation, I needed to interview

participants until no new information was being revealed. Data saturation was determined by assessing the responses as I received them and adjusting my questions in real-time to identify any possible new information. Data saturation was achieved by interviewing all willing participants within the multiple organizations that were involved in the process of data analysis concerning big data.

Population and Sampling

This qualitative case study's population consisted of data scientists that utilize big data analytics to analyze data for decision making within their organization. I selected organizations that are either based in the greater Salt Lake City, Utah area, or that have a technical presence in that geographical area. It can be challenging to gain access to SME's unless the researcher has a network to leverage (Ponelis, 2015). To mitigate this challenge, the population for the study was drawn from companies that are known to the researcher, either by the association in the Utah Chapter of the Association for Information Management or through membership in the Big Data Meetup group. Participant case study organizations were those that were currently implementing or using big data to make business decisions. The population for this study had the knowledge and expertise in the area of big data analytics and the associated governance around those data. The population for the in-person interviews was made up of data scientists within the organization that were directly involved in the gathering, analysis, and modeling of big data elements.

Fink (2000) indicated that the number of participants in a qualitative study should be as many as needed to uncover the information needed to answer the research question. Fusch and Ness (2015) added that when performing a case study, the researcher cannot

identify a specific number of interviews required to reach saturation; however, the researcher will take what they can get from the case study organization. For this study, I used purposeful sampling to select the participants from the population. The sample should be determined by identifying individuals within each organization that have a high level of interaction with that population (Whittingham, Barnes, & Dawson, 2016). For this study, I determined the sample by working with the gatekeeper to identify individuals within the organizations who had a high level of interaction with big data. I planned to interview between five and 10 participants from two organizations. The final number of participants was dependent on achieving data saturation. I completed the study with 10 participants from three different organizations. While different studies use distinct inclusion criteria, it is acceptable to select participants based on their presence in a specific context (Fink, 2000). Once the population was defined by the selection of the case study organizations, the number of participants was be determined relative to the number of data scientists involved in the data modeling and analysis process within those organizations and when data saturation is achieved.

Semi-structured interviews can take a variety of forms that can include face-to-face, telephone, webchat, or email (Brown & Danaher, 2019). Typically case study interviews are conducted in an interactive setting where the participants can feel at ease responding to the questions (Malli & Sackl-Sharif, 2015) and where the interviewer can establish rapport with the participants (Brown & Danaher, 2019). Interview settings should be carefully considered to ensure the comfort of the participant while simultaneously reducing distractions and background noise (Dikko, 2016). I conducted all of the interviews using video conferencing, thus allowing the participants to interview

in a location that is familiar to them. The interviews typically lasted for less than 60-minutes.

Ethical Research

Ethical research practice includes acquiring informed consent and voluntary participation of participants (Eisenhauer, Tait, Rieh, & Arslanian-Engoren, 2019). The researcher must protect the privacy and confidentiality of participants (Yip, Han, & Sng, 2016). All participants should enter research voluntarily and with enough information for them to decide whether to participate (U.S. Department of Health & Human Services, 1979). I explained the interview process to each participant before providing them with the informed consent form for their signature. In the email that I sent them with the informed consent document, I outlined the purpose of the study along with the benefits, risks, and nature of the study (see appendix C). I also informed them of my plans to record the interview, and I also asked for their consent for this recording at the start of the interview. Before acceptance into the study, all participants were required to respond to the email with their consent. This study was conducted ethically, adhering to all ethical guidelines as outlined by the university as well as best practices within academic research.

The Institutional Review Board is responsible for overseeing human studies research, ensuring that the researcher conducts the study according to generally accepted ethical standards and that it adheres to federal and state regulations as well as with university policies and procedures (Caldamone & Cooper, 2017; Cseko & Tremaine, 2013; Qiao, 2018). I obtained approval from the Walden Institutional Review Board (IRB) before contacting any participants or collecting any data. By the informed consent

process, all participants signed a form that informed them of their rights as a study participant, including the method for withdrawing from the study. Participants have anonymity in this process, which was enhanced by the private setting in which the interviews took place. Participants were also asked to maintain the confidentiality of the study by refraining from discussing the names of those involved in the individual interviews or discussing the content of the interview itself with anyone else while the study was underway. All efforts were made to protect all participants from any harm as a result of their participation in this study.

Names of the organizations, as well as the names of the individual participants, are masked in the study documentation. I maintained a spreadsheet with the names of the organizations and participants in which I assigned generic names to each one such as organization I, participant I, participant II, for as many organizations and participants I have included in the study. This spreadsheet is the only place where the participant's names are associated with the masked names. This spreadsheet is stored with the other study documentation in an encrypted OneDrive folder, and I did not include them in the final study documentation. This process provides an audit trail and a simplified method of referring to each participant in an unidentifiable manner throughout the study documentation. Participants were allowed to decline to participate in the process or withdraw at any time. The process for withdrawal from the study was to inform the researcher that they are not willing to participate in the interview process. Upon receiving their notice to withdraw, I removed them from the interview process. There were no repercussions to the participant for their request to withdraw.

Following the Walden University process, data from this study will be maintained for five years after the study, and I will keep them in an encrypted OneDrive folder during that period. After five years, I will destroy all the data from this study. Participants will also have no monetary or non-monetary incentives offered for their participation in this study.

I interviewed participants in a one-on-one setting, with member checking utilized to ensure that I had documented their responses in alignment with their intent. I did not use the participant's names in the study and only referred to them by a number. I asked that each participant maintain confidentiality and that not discuss our interactions with any other employees at their company.

Data Collection

Data Collection Instruments

In qualitative research, the researcher is the primary data collection instrument (Clark & Vealé, 2018). Researchers should present their findings in a way that readers can understand both their approach and their findings (Sutton & Austin, 2015). Qualitative studies commonly use these four methods of data collection: interviews, focus groups, document reviews, and observation (Gill, Stewart, Treasure, & Chadwick, 2008). As the researcher of this qualitative study, I was the primary data collection instrument, and I used interviews and operational document analysis as the other data collection mechanisms. I followed the defined interview protocol (see Appendix B) when conducting the interviews. Methodological triangulation was used to correlate these multiple sources of information, and from which I generated the study findings. Fusch and Ness (2015) stated that triangulation goes a long way to achieving data saturation.

Process documentation and transcripts from the interviews were coded and correlated to uncover trends and standard practices.

As the primary data collection mechanism, the influence of the research cannot be removed entirely from the investigative equation in a qualitative study (Draper, 2015). Data collection in qualitative case studies may take several forms, including interviews, documentation review, on-site observations, and illustrative materials (Yin, 2017). Case study data collection should follow a formal protocol, but the specific information that may become relevant is not always known upfront (Yin, 2017). Establishing a rapport with the participants can set them at ease, which will also increase the effectiveness of the semi-structured interview process and generate mutually beneficial outcomes (Brown & Danaher, 2019). Interviews in a qualitative study can provide rich and informative data that expand the topic (Sutton & Austin, 2015). During the interview process, it was also essential to ask multiple participants the same questions to achieve data saturation and to avoid trying to hit a moving target (Fusch & Ness, 2015). Semi-structured interviews help guide the conversation by allowing the interviewer to ask additional questions based on responses from the participant that enable reciprocation between the interviewer and the participant (McIntosh & Morse, 2015). Semi-structured interviews provide a mechanism for participants to be open to revealing additional relevant information that may not otherwise be identified (McIntosh & Morse, 2015). I used a semi-structured approach to the interview process. I did not interrupt responses and allowed the participants to share as much information as they felt was needed. I asked clarifying questions as needed throughout the interview.

Document analysis can be used by qualitative researchers to gather pertinent facts about the organization (Dunne, Pettigrew, & Robinson, 2016). Document analysis provides a mechanism for reviewing and evaluating documents from the organization being studied (O'Connell, Mc Carthy, & Savage, 2018). Sometimes the document analysis may reveal relevant information about the topic that the researcher would not otherwise discover (Dunne et al., 2016). Analysis of documentation can provide the researcher with insights into how the organization presents itself as a business (Lawson, 2018). Document analysis can also be used to corroborate information obtained from interviews to further validate the study (Dunne et al., 2016). The use of triangulation during data analysis significantly increases the reliability of the results by using multiple data sources to reach conclusions about the data (Fusch & Ness, 2015). Methodological triangulation is the specific type of triangulation that is utilized to correlate information gathered from multiple data collection sources (Fusch & Ness, 2015). Triangulation not only enhances reliability, but also helps to ensure data saturation, and is the method in which a researcher explores different levels and perspectives of the same phenomenon (Fusch & Ness, 2015). I used operational document analysis in addition to interviews in this study. By working with the gatekeeper, I gained access to the process and standards documentation that is used by the organization to govern their big data procedures. I evaluated the documentation for applicability to this study. I used methodological triangulation and member checking as validation mechanisms.

Data Collection Techniques

Data collection in qualitative research occurs primarily through unstructured methods of verbal interaction and observations (Morse, 2015). Thorough qualitative

analysis requires that data be gathered from multiple sources to enable triangulation (Twining, Heller, Nussbaum, & Tsai, 2017). Using multiple data sources will allow for the methodological triangulation of those data points to enhance the validity of the study (Morse, 2015). For this study, I gathered data from two distinct sources, including semi-structured, face-to-face participant interviews, and a review of procedural documentation. I conducted in-depth interviews of individual participants using the questions in the Interview Questions section and recorded the session using GoTo Meeting software on a laptop computer. I also took notes where I recorded my impressions of the responses as well as observations about the participant's body language and other environmental details. The GoTo Meeting recordings of the interviews were transcribed into Word documents using a transcription service. I had a confidentiality agreement signed with the transcription service before the use of that service.

I gained access to participants and documentation through a gatekeeper within the organization. I ensured that each participant understood and agreed to the informed consent process, and I conducted the interviews in a private and comfortable setting for the participants. I acquired IRB approval before initiating contact with any participants.

Once I had their consent to participating in the study, I worked with the participants to arrange a time for the video conference interview, which I enabled a remote face-to-face setting within the greater Salt Lake City, Utah area. I worked with the participants to identify a date and time that fit within their schedule. I ensured that the interview took place in a location that is free from distractions and with which the participants are comfortable and familiar, and which protects their privacy. These details are essential because the location of the interview can influence the content and data

collection process itself, and distractions should be minimized as much as possible (Dikko, 2016).

I informed the participant that the interview was being recorded for accuracy and later transcription. I also informed them that they can request that I turn off the recording at any time. I also informed the participant that I will be taking notes during our interview and that I would be recording both their responses and my impressions of the interview. I ensured that they were comfortable with the process before proceeding.

One of the dangers of qualitative research is when data collection is superficial, such as taking participant's comments at face value (Twining et al., 2017). Participant or member checking is used in mitigating this danger by allowing the participant to clarify intent and comment on my interpretations of their input (Twining et al., 2017). A form of member checking can also be used during the process of the interview by asking for clarification and more depth of specific responses by the participants during the interview process (Morse, 2015). Qualitative researchers need to be both gathering and analyzing data concurrently in an iterative process during the interview (Twining et al., 2017). During the data analysis phase, my summary of the interview was provided to the participant for their review, and edits were made accordingly. This process continued until the participant was satisfied that the transcript reflected their responses accurately. No online or web-based software was utilized in this process. All digitized data relating to this study is stored on my encrypted OneDrive and will be maintained there for the required five-year period. The digital data will then be deleted.

Data Organization

The protection of the anonymity and confidentiality of participants is a critical aspect of ethical research (Castillo-Montoya, 2016). It is also critical to quality research to ensure that the data are transparent (Bengtsson, 2016). The use of a spreadsheet for this purpose makes the data readily accessible (Shapiro & Oystrick, 2018). I used a spreadsheet to categorize and organize the data and findings from this study. This spreadsheet is stored on an encrypted OneDrive for both security and recovery purposes. I used participant numbers to identify the participants both within the spreadsheet as well as in the physical folders to ensure that the participant names are not associated with their data in any way. Before being imported into the computer-assisted qualitative data analysis software (CAQDAS), the documentation for each participant was stored in the folder on the OneDrive that is associated with their organization. I used the folder to store interview notes, emails, consent forms and other documentation from the participant or that was provided to me by that participant.

Due to the remote nature of the interview process, there was no physical paperwork collected during this study. All electronic data, including a data backup of the CAQDAS application, are stored in the same OneDrive folder location. This process ensures accessibility to the data while protecting both the security of the data and the ability to restore in the case of accidental deletion. Participants were not identified in any way other than by their assigned numbers to protect anonymity. After five years, the information on the OneDrive will be deleted.

Data Analysis Techniques

After collecting the data, it must be analyzed to find meaning within the data that is relevant to the research questions and the overall purpose of the study (Bengtsson, 2016). Data analysis is the process by which the researcher reviews the collected data repeatedly to uncover either the correlation or contrast that provides meaning for the study (Fusch & Ness, 2015). Grouping data into common topics and themes allows for better data analysis, as it focuses the large data sets into smaller and more isolated elements (Maguire & Delahunt, 2017). My review focused on the strategies, processes, and procedures identified within the organization for overseeing the governance of big data. I reviewed the documentation, interview notes, transcripts, and other data sources multiple times, searching for common themes and descriptions with each review. I reviewed and analyzed the operational documents as a mechanism for validating the descriptions from the in-depth interviews. I identified major themes from the various sources and used color coding of the data to identify and correlate significant themes.

I reviewed the data gathered repeatedly to uncover meaningful information to inform the research. I specifically searched for data that answered my central research question of what big data governance strategies do data scientists employ to provide a holistic perspective of those data for making decisions? Identified themes were highlighted in various colors on the digital documentation in the CAQDAS.

Software packages can be used by qualitative researchers to assist in the analysis of qualitative data to look for dimensions, categories, and subcategories of analysis (Freitas, Ribeiro, Brandão, de Almeida, & de Souza, 2019). Software packages such as this enable the researcher to relinquish the merely technical tasks relating to the

organization of the findings that do not require intellectual effort (Freitas et al., 2019). A Computer-Assisted Qualitative Data Analysis Software (CAQDAS) tool is a computer-based software tool that can automate the qualitative data analysis process (Moylan, Derr, & Lindhorst, 2015). Some researchers argue that using a CAQDAS tool for qualitative data analysis provides better results than manual analysis would (Moylan et al., 2015).

For this study, I utilized a CAQDAS tool called QDA Miner. This software tool was used for the text-based analysis of interviews and open-ended questions. QDA Miner was used for coding of the various data sources from my data collection. I entered the information from the various data sources into QDA Miner and used the coding feature of the software to help me to analyze the information and categorize the data into themes, groups, and categories. This software helped me to discover relationships among the data elements. I was looking for reoccurring themes and strategies from the various sources of data gathered during my data collection phase.

Reliability and Validity

Validity in a qualitative study is the extent to which the instrument of measurement has measured the concept that it was intended to measure (Dikko, 2016). Reliability is defined as the ability to consistently and without bias, achieve similar results (Dikko, 2016). Trustworthiness in a qualitative study is established through four criteria known as credibility, dependability, confirmability, and transferability (Forero et al., 2018). Credibility, dependability, and confirmability in qualitative studies, replace internal validity, reliability, and objectivity (respectively) in quantitative studies (Morse, 2015). Credibility and trustworthiness in qualitative studies are enhanced using analytic processes around these four criteria (Twining et al., 2017).

Taken together, these criteria attempt to speak to the overall rigor of the qualitative study, but these four criteria have never been tested or questioned since their introduction in 1981 (Morse, 2015). While qualitative rigor is the overall goal, validity may be supported inherently by reliability, so researchers should be most concerned with overall trustworthiness, and lean more on traditional reliability and validity concepts as defined in general research terms (Morse, 2015). This rigor can be accomplished by implementing qualitative comparative analysis, which consists of clarifying the question of external validity, ensuring internal validity, and explicitly adopting a specific mode of reasoning (Thomann & Maggetti, 2017).

Credibility

Credibility establishes confidence that the results are accurate, credible, and believable from the perspective of the participants (Forero et al., 2018). Korstjens and Moser (2018) explain that credibility demonstrates that the viewpoints expressed by the participants and aligns with the researcher's representation of those viewpoints. Ensuring that the viewpoint of participants is accurately reflected in the study findings aids in producing research that is both ethical and reliable (Cypress, 2017). The researcher achieved credibility by prolonged and varied engagements with each participant, and by documenting the interview process and protocol (Forero et al., 2018).

To achieve credibility, I ensured that I had enough time with each participant to gather the needed information, and I thoroughly documented the interview process and protocol. Korstjens and Moser (2018) identify triangulation as one of the critical methods of achieving credibility. Qualitative researchers often use triangulation to increase the probability that their study findings will be found credible (Nowell, Norris, White, &

Moules, 2017). I utilized both methodological and data triangulation in this study to enhance credibility. I also employed member checking to enhance the credibility of the study. Korstjens and Moser (2018) explain that the use of member checking to gather feedback regarding the interpretations, conclusions, and analytical categories will enhance credibility. This process builds confidence in the accuracy of the responses. Ensuring the accuracy of responses helps to ensure the validity of the overall study and will help determine the reliability of the content.

Dependability

Dependability ensures the findings are repeatable if the study were repeated with the same set of participants, coders, and context (Forero et al., 2018). Korstjens and Moser (2018) define dependability as the stability of the findings over time. It is achieved when the researcher explains the logical and traceable research process and when the process is documented (Nowell et al., 2017). Dependability is also achieved by providing a detailed description of study methods and ensuring that a stepwise replication of the data is documented (Forero et al., 2018).

To achieve dependability, I provide a detailed description of the study methods used. I clearly described the process that I followed to both collect and analyze data, and I ensured that I documented a stepwise replication of the data. Providing an audit trail of the research process enhances dependability (Korstjens & Moser, 2018; Nowell et al., 2017) by providing steps for other researchers to follow if they attempt to replicate the study. I provide an audit trail within my findings that will help the reader to understand not only how the data were collected but also to clearly understand my thought process

during analysis, my interpretations and impressions during the interview process, and the logic that I employed to reach conclusions from the data.

Confirmability

Confirmability provides confidence that the results would be confirmed or corroborated by other researchers (Forero et al., 2018). This confidence is primarily achieved by data analytics processes such as methodological triangulation (Abdalla, Oliveira, Azevedo, & Gonzalez, 2018; Forero et al., 2018). Nowell et al. (2017) state that confirmability is achieved when credibility, transferability, and dependability are all reached (Nowell et al., 2017).

To achieve confirmability, I ensured that credibility, transferability, and dependability were all achieved. I utilized both data and methodological triangulation to gather and analyze the data for this study. I also clearly described my thought process and the logic that I used to make decisions regarding this study. I documented the interview process, including thoughts, impressions, choices, and decisions made during the data gathering and analysis phase of the study.

Transferability

Transferability provides confidence that the study results could be generalized or transferred to other contexts or settings (Forero et al., 2018; Nowell et al., 2017). Korstjens and Moser (2018) explain that transferability refers to the ability of the reader to assess whether the findings apply to their environment or situation, also known as the transferability judgment. The clear implication is that the reader determines transferability rather than the researcher (Korstjens & Moser, 2018).

As the researcher, I cannot know how individual readers will want to transfer the findings to their environment. The only mechanism that a researcher can use to enhance transferability is to provide detailed and thick descriptions of the participants and research process so the reader can determine applicability to their setting (Nowell et al., 2017). Transferability is primarily achieved through the analytic process of participant selection and ensuring data saturation (Forero et al., 2018). I documented the selection of participants and provided detailed descriptions of their selection and participation criteria for this study to provide as much clarity to the reader as possible.

External validity increases when findings from the population can be generalized beyond the individual case studied (Thomann & Maggetti, 2017). Care must be taken in the selection of the population for the case study. Organizations must be selected that will provide in-depth knowledge that is relevant, critical, and feasible for answering the research question (Thomann & Maggetti, 2017). Additionally, Welch and Piekkari (2017) point out that case studies generalize to a conceptual theory, rather than to a population. By generalizing to the OIPT, this study gains external validity by a selection of a population that is utilizing big data analytics, and the findings of the case study could be generalized based on the OIPT.

Obtaining feedback on the interview protocol can enhance its reliability and trustworthiness as a research instrument (Castillo-Montoya, 2016). The interview protocols defined in Appendix B were reviewed by my committee to ensure the reliability of the protocol and enhance the trustworthiness of my research. Pre-defined interview protocols enhance reliability and increase the quality of data obtained from research interviews (Castillo-Montoya, 2016).

Transition and Summary

This section restated the purpose of the study and provided information regarding the methodology, participants, and decisions relating to the design of the study. Also discussed were the ethical considerations, researcher role, and approaches for data collection and analysis. Finally, I discussed validity and reliability. This section details how I conducted my study relative to my selected methodology and design.

The next section will include a presentation of the findings of my study, as well as the application to practice and future research. Social change will be discussed, as well as personal reflections on this overall study and process.

Section 3: Application to Professional Practice and Implications for Change

The focus of this study was data governance strategies that data scientists are using in practice for big data. I will present the findings from my data collection and analysis and explain how these findings may contribute to the academic literature. I will also explain how findings may affect social change and professional practice as big data analytics matures as an industry practice. Finally, I will make suggestions for future study and present my reflections on the overall study.

Overview of Study

The purpose of this qualitative multiple case study was to explore the big data governance strategies employed by data scientists to provide a holistic perspective of those data for making decisions. The data for this study came from semi-structured interviews that I conducted with data scientists currently working on big data solutions in the greater Salt Lake City, Utah area. The documents analyzed for this study are 10 participant interview transcripts and four process documents from three organizations. This section begins with a brief overview of the study methodology, and then I present the findings from the study.

Presentation of Findings

I began this study with the intent of answering the following research question: What big data governance strategies do data scientists employ to provide a holistic perspective of those data for making decisions? In this section, I will present the findings from my data gathering and analysis relating to this question. The study consisted of 10 participants from three different organizations with an average of 9.4 years of experience working in big data analytics. I used methodological triangulation by obtaining process

documentation from each organization in addition to the interviews held with the participants. The following four major themes emerged from the analysis: ensuring business centricity, striving for simplicity, establishing data source protocols, and designing for security. I will present each of these themes in this section.

Theme 1: Ensuring Business Centricity

Data scientists cannot provide answers to business questions without first having a solid understanding of the business question they are answering. The better the business question is understood by the data scientist, the better they can design an analytic model to answer that question. This theme includes four common subthemes: understanding business needs, partnering with businesses, maintaining contextual awareness, and minimizing data noise (see Table 1).

Table 1

Subthemes for Ensuring Business Centricity

	Interviews		*Documents*	
Subtheme	Count	References	Count	References
Understanding Business Needs	10	26	3	8
Partnering with Businesses	10	43	3	8
Maintaining Contextual Awareness	7	21	4	6
Minimizing Data Noise	10	34	2	4

Subtheme: Understanding business needs. Having a clear understanding of business needs was a universal theme that came up in every participant discussion. Before a data scientist can provide answers to business questions, they must understand the needs that they are addressing. Understanding business needs is an important theme

from both analytic and data governance perspectives. Understanding business needs helps data scientists identify appropriate data sources and elements that will be needed to appropriately and completely address the business question. Each participant discussed the importance of understanding the question they were being asked to answer before they started assessing data sources and elements for analysis. There were a variety of different ways through which they engaged with their business partners. Participants 1, 3, 4, 5, 6, and 7 discussed interviewing their business partners before initiating any analysis. P5 said: "I do the fact-finding into their requirements and also the composition of what their data looks like...Once I get adequate information … I make a determination on how to construct and load the model." P6 said: "Before developing a model, we interview or have discussions with the business owners that have expertise in that space to tell us the hypothesis or reasoning …that might have an influence on the decisions that [we need to make]."

P2 and P8 referred to the long tenure of the data scientists in their organization and indicated that this provides them with significant business context. Because of their tenure in the business, they understand the business need immediately when presented with a new question to be answered. P8 explained that when they receive a request for a new model, "generally speaking, we know what we are looking for" because of their experience in their specific business. All 10 participants discussed the importance of having a clear definition of the business need as a critical step to deciding which data sources they needed to connect with to get their data. P7 indicated that "the more definition that we have upfront from the end-user of what they want to get out of it, the

better the process, the smoother it goes, the faster we can deliver that datamart to answer their questions."

Gardiner et al. (2018) explained that big data scientists are required to possess a wider variety of skills than traditional technologists due to the multifaceted nature of this new role. Alignment between overall information systems strategy and business strategy is a critical precursor to overall profitability and competitive advantage (Shao, 2019). De Mauro et al. (2018) found that while the main focus of the role involves the data itself, associated analytical methods for transforming those data into usable insights are critical to success because of the complex nature of big data. Data scientists cannot bridge this gap without having a clear understanding of business needs to which they are responding when creating their analysis or models from big data.

The concepts of the OIPT identify that pushing decisions down in the organization yields better decisions by those closer to the tasks. The OIPT directly applies to the gathering, interpretation, and synthesis of information within an organization for decision-making (Tushman & Nadler, 1978). Data scientists require a clear understanding of business needs to be able to interpret and synthesize big data for strategic decision-making purposes. The increased ability within an organization to process information using technology provides a mechanism for improving decision-making within that organization due to the improved ability for employees to process more copious amounts of information farther down within the organization (Kroh et al., 2018). It is only by understanding business needs within the organization that data scientists can build useful models for processing that information and disseminating it lower in the organization.

Subtheme: Partnering with the business. Maintaining a partnership with the business was also a universal theme that all 10 participants mentioned as a critical success factor. This subtheme is closely aligned with the subtheme of understanding business needs. By creating partnerships between various technical and business roles, data scientists can learn more about business needs. Developing a partnership between business roles and data scientists establishes a level of trust and communication that enables data scientists to better leverage data to help answer strategic business questions. P4 said:

> A few years ago...people weren't as educated about what data we had and what kind of questions to ask. [In prior years], we used to come up with the questions. They'd ask me some questions to see if we can get some data to answer that. And then usually we will try to understand what they're going to do with the data. And most of the time that really didn't lead to anything… But more lately, I think people are a lot more aware of what data we have and what kind of questions to ask.

P9 similarly discussed the process of educating their business partners regarding the type of information that was needed and available to answer their business strategy questions. Partnerships between business roles and data scientists helps to communicate more clearly about what data elements are available for analysis to answer business questions.

P8 discussed using a very iterative approach with their business partners as they build their models, continually having conversations about what needs to be changed, adjusted, or clarified as they work to answer the questions that are being asked. They described their partnership in this way:

> So if you ask for something today and I deliver it to you tomorrow, the likelihood is it's not going to fulfill exactly what you wanted, even though you may have attempted to communicate to me and I attempted to receive what you were saying. At the end of the day, it's going to take us [several] exchanges back and forth. 'Well, that's close, but I really wanted to do this.' 'Okay, let me go back and do that...', 'That's a lot closer, but now that I'm using it, I see that it needs this other element.' So … what I am looking at is kind of bringing people in and working with people to make sure we get it. You know, we get what has been requested. We get that right.

All 10 participants stated the importance of maintaining an ongoing partnership with their business partners as a critical strategy to defining what data to collect and from which data sources to collect those data.

Collecting and processing of big data require cross-departmental collaboration and partnerships that do not exist with traditional data (Janssen et al., 2017). Successful implementation of big data analytics requires new roles and organizations around the business to interact, which had not previously needed to work together (Braganza et al., 2017). Because of these new partnerships, some power shifts have been documented within organizations that are effectively utilizing big data analytics to make fact-based and real-time decisions (Popovič et al., 2018). Effectively utilizing big data to make decisions within organizations is requiring new partnerships between data scientists and business roles that were not as clearly required previously.

Uncertainty is defined as a key driver of information sharing within the OIPT. Partnerships must be cultivated with the intent of sharing and utilizing information to

reduce uncertainty around business decisions. The success of any organization is primarily affected by the competitive ability of managers to make strategic decisions in the face of uncertainty and ambiguity (Intezari & Gressel, 2017). One of the key tenets of the OIPT is that the greater the uncertainty of a task, the more information must be processed by the decision-makers to execute that task (Galbraith, 1974). Since organizations in the OIPT are simply information processing systems that collect, process and distribute information (Zelt, Recker et al., 2018), and since that lack of information within an organization leads to the uncertainty to which Galbraith refers (Gupta et al., 2019), it follows that the greater the ability to partner across the various departments to share information within an organization the greater will be their ability to reduce uncertainty.

Subtheme: Maintaining contextual awareness. Contextual awareness of the data elements is critical to providing a valid analysis. Seven of the 10 participants mentioned maintaining contextual awareness as a key strategy for managing information from big data. For the analysis to be useful to business decision-makers, data scientists must ensure they understand the context of the data from the various data sources before building a model from those data. Participants 1, 4, 5, 7, 8, and 9 each discussed maintaining contextual awareness while gathering, preparing, and analyzing big data. P1 stated that "we record context around the data…there's a tremendous amount of context" and clarified that "with additional context, you have a much, much richer fingerprint that is harder to accidentally duplicate…you have a greater chance of uniqueness when retaining that context and metadata."

The specific strategies for how to maintain this context vary by organization. P9 articulated that their organization maintains the data lineage throughout the analysis process to provide contextual validity to the analysis. They explained that by maintaining data lineage, they could determine that the "source provided additional information" and that they could then check "to see if [they] can gather additional information from that source." P9 explained further that context can help them to know "if a particular column changes, what reports downstream of that data are going to be affected and need to be adjusted, rather than deal with the outcomes after the fact." They explained that context helps them to determine the scope of the analysis. Using context, they can determine whether the report will meet with executive scrutiny as the executives attempt to correlate information from various reports. Context is important to help those executives accurately interpret the results of the analysis.

P1 and P5 indicated that metadata can be used to maintain context. P1, P4, and P9 indicated that the data source itself could provide some important context to the data. P7 suggested that correlating contextual data from various sources is "when the analytics [are] a lot more powerful for us." P8 mentioned that it is important to maintain contextual awareness when automating results so that the context is not lost because of the automated manipulation.

Big data analytics rely on fuzzy logic and inductive statistics (Paul et al., 2018). Considering that some big data sources may be unknown or unvalidated some big data analytics may be incomplete or inaccurate (Herschel & Miori, 2017). Without contextual awareness, big data solutions can quickly become a liability rather than an asset. Improper or incomplete analysis of big data can lead to misinterpretation of those data,

and lead to incorrect business decisions (Matthias et al., 2017). Alternatively, proper data governance addresses the standardization, accuracy, and availability of those data (Wang et al., 2018). Big data governance is critical to ensure that data are accurate, shared appropriately, and protected per company policies (Cheng et al., 2017). This governance includes maintaining contextual awareness of both the data elements and the data sources.

One of the central tenets of the OIPT is the need to close the gap of uncertainty by processing more information within the organization. Duvald (2019) indicated that the OIPT includes the mitigation of equivocality in addition to uncertainty. Duvald (2019) also explained that equivocality arises from the existence of multiple and often conflicting interpretations of data. The information dispersed throughout the organization to close the gap identified by the OIPT must be accurate, or the analysis will not hold up to the scrutiny of the business partnership. Maintaining the original context of the data elements lends credibility to the analysis and reduces both uncertainty and equivocality of that analysis.

Subtheme: Minimizing data noise. More data does not always mean better analysis. Eight of the 10 participants discussed the importance of minimizing the noise created from big data. P3 said:

> A lot of the data is really just noise. You need to be sure to pick the things that actually matter for the analysis... With the amount of data that we collect, you will never be able to use it all…but a high percentage of the data that we are collecting is pretty useless ... Just because you're collecting it doesn't mean that you can use it for anything.

P1 stated something very similar when discussing the amount of information that their organization gathers. P8 stated that reducing the noise is “a matter of narrowing the broader data set down to what we're specifically trying to get to.”

P1, P2, P3, P4, and P9 stated that they do not attempt to analyze all of the data for any particular analysis because there is simply too much of it. P2 said: “I don't think it has been necessary to look at all the available data.” P3 added: “I would say that you don't need to analyze all the data necessarily… I don't feel like you need to or should actually use all the data.” P4 similarly stated: “I don't think we ever are analyzing all the available or all the applicable data to answer any question. Practically, I don't think it's possible.”

These participants agree that analyzing all of the data is not an achievable goal and stated that it is more important to narrow the data sets to the most applicable data for the specific business question before attempting an analysis. P3 stated: “I think that right now there are so many more questions that can be answered with smaller data.” Simply adding more data into the analysis does not result in better analytics or better decisions.

One of the key responsibilities of data scientists is to reduce the noise and simplify the data analysis process. This theme is confirmed in the literature, as several studies have highlighted the idea that leveraging big data does not simply imply that obtaining more raw data will lead to better information. The quality of decisions from big data is not influenced solely by the amount of data collected but more so by how those data are collected and processed (Janssen et al., 2017). Matthias et al. (2017) pointed out that without appropriate analysis, it does not matter that large amounts of data are

collected. Value is extracted from big data when information is transparent and usable to the organization (Ahmed et al., 2017).

The reduction of uncertainty by processing more information is a key concept of the OIPT. Uncertainty in the context of the OIPT is defined as the difference between the information processed and the information required to complete a task (Tushman & Nadler, 1978). Merely increasing the information available does not resolve that uncertainty (Gao et al., 2018). The success of an organization is primarily affected by the competitive ability of managers to make strategic decisions in the face of uncertainty and ambiguity (Intezari & Gressel, 2017). When data scientists have a clear understanding of the uncertainty that the managers face, they can then design models with minimal data noise to assist these organizational managers in resolving that ambiguity, thus resulting in better decisions deeper within the organization. The ability for organizations to make strategic decisions amid uncertainty and ambiguity relies on their ability to continuously learn and reconfigure the organization's knowledge base (Intezari & Gressel, 2017). Data scientists can help their organizations face uncertainty when they understand the business need, partner with the business to ask and answer the right questions, maintain contextual awareness of the data elements, and minimize the overall data noise.

Theme 2: Striving for Simplicity

Because of the volume, variety, and velocity attributes of big data the analytics can become complex very quickly. Complexity is one of the significant challenges with big data as compared to traditional data (Jin et al., 2015). All 10 participants mentioned the importance of simplicity in their responses in some manner. Process documentation also reminds data scientists to reduce duplication and complexity in their design. The

fundamental basis for the OIPT is also centered around the uncertainty and complexity of task completion and is based on the concept that organizations should be designed to reduce that uncertainty and to enable decision-making (Feurer et al., 2019). Striving for simplicity is a key skill that will help data scientists to deal with the inherent complexities of big data while simultaneously offering solutions that can be maintained over time. This theme consists of four subthemes: minimizing the data sources, using new tools, simplicity of design, and using automation (see Table 2).

Table 2

Subthemes for Striving for Simplicity

	Interviews		*Documents*	
Subtheme	Count	References	Count	References
Minimizing the Data Sources	9	23	2	4
Using new tools	8	23	0	0
Simplicity of Design	10	28	3	9
Using Automation	8	15	4	11

Subtheme: Minimizing the data sources. Strategically limiting the sources of data was a common concept that was raised among the various participants. This theme is counter-intuitive when referring to big data. It is not uncommon for big data to infer many data sources, but high volume and velocity can also occur with a small number of data sources. This is particularly true of IoT where a single data source could have a significant number of endpoints. Big data does not necessarily have to mean copious data sources but could simply indicate many endpoints generating data. This theme is related to the number of data sources and should not be confused with the number of data-

generating endpoints. The reduction of the number of data sources was a concept outlined as a strategy for maintaining the simplicity of design. This subtheme is also closely related to the subtheme of minimizing data noise. One potential source of noise is having too many data sources providing irrelevant or duplicative data elements.

Nine participants mentioned some form of limiting the data sources in their responses. When explaining how they select appropriate data sources for any model or analysis P1 explained that “there’s always going to be more data…we don’t have hundreds of sources. We have a handful of core primary sources of data”. P2 identified a handful of core sources of data that can provide answers to most of the business questions asked at their organization. P3 reiterated similarly that “there are a finite set of sources…we know what's available to us and what's easily accessible”. P6 articulated that they strategically limit their data sources to only those that are valuable for the specific business question being analyzed. All of the participants mentioned that there could be more data available to them from other sources, but that minimizing the sources of data is one of the key strategies that they follow to deal with the ever-increasing complexity of various big data sources.

Determining the origin and transformation of various data elements obtained from the myriad big data sources is currently one of the primary challenges of big data (Umer et al., 2017). Because of the difficulty of verifying the various data sources associated with big data the validity of analysis can be questionable (Hashem et al., 2015). Additionally, external data sources are less likely to have internal data stewards or data governance processes established to govern its use within the organization (Cervone, 2016). Strategically determining a minimal set of sources for big data analytics within the

organization can provide the data scientist with confidence in both the validity of the data as well as confidence in the analytics.

The OIPT defines two mechanisms for closing the gap between the need to process information and the ability to process information. Organizations can either reduce the need to process information or increase the ability to process information (Galbraith, 1974). Reducing the data sources in any analysis accomplishes both, and thus reduces the gap from both sides simultaneously. The primary goal of closing the gap is to reduce uncertainty. Tushman and Nadler (1978) defined uncertainty as the difference between the information processed and the information required to complete a task. This does not imply that simply gathering more information from myriad sources will result in the reduction of uncertainty. Merely increasing the information available does not resolve the uncertainty as individuals do not have unlimited processing capacity and will reach information overload (Gao et al., 2018). By focusing on a limited set of data sources, each of which can provide a high volume of data elements, data scientists can build confidence in their analytics while both reducing the need to process more information from other sources and increasing the ability to process more information by using big data analytics.

Subtheme: Using new tools. Many new tools have been developed in recent years to simplify the analysis of big data. These new and advanced tools are required to extract information from big data (Intezari & Gressel, 2017). These new tools are also important from a governance perspective as they help data scientists to better interact with data in ways that traditional tools do not. Eight of the 10 participants discussed the use of these new big data tools as a strategy for coping with the complexity of big data.

The adoption of these new tools is still an ongoing process. None of the participants indicated that they are currently using all of the various tools available to them. All eight of the participants that discussed tools indicated a desire or plan to start using more of these specialized tools soon to further refine and enhance their data analysis.

P2 and P5 indicated that they believe the new tools will help them become more efficient at interacting with big data. P2 said:

> I think there's a skill gap that we're trying to address to understand what's changed in…data engineering over the last five to 10 years and how do we adapt and take advantage of that? We've primarily been stuck in kind of one technology stack for a long time and we've become very efficient and good at it…so what's compelling them to want to change and look at other alternatives and options?

P2, P3, P4, and P7 discussed their plans to leverage more cloud technology soon, and they all expressed confidence that this would bring new capabilities to their ability to effectively govern big data. P2 indicated that they "eventually want to put our data [in the] cloud…right now, it is … on-prem. And so how can we leverage some of the new paradigms and tools to get the flow of data a little bit quicker?" P9 said:

> The volume that they talked about were…challenges [for companies] like Google and Facebook and Twitter [because of] the volume of data they were processing. We had to create tools to meet those challenges and we were then able to use those tools. When they're working with a thousand times more data than we have, and they've built a tool that can solve that solution, it usually handles all our smaller problems.

As larger organizations adopt and utilize these new tools small and medium size businesses can benefit from the use of these same tools to solve their big data challenges.

Traditional analytic systems lack operational efficiency to analyze big data effectively (Jin et al., 2015). Traditional tools for data governance and analysis are outdated and are no longer adequate for processing the vast amount of data available today, making current strategies ineffective for handling big data (Bello-Orgaz et al., 2016). Intezari and Gressel (2017) added that to extract information from big data, new techniques, and advanced tools are required. Many organizations are starting to analyze big data with traditional tools but are quickly establishing plans to implement new and specialized tools.

As organizations learn to process larger amounts of data new tools will play a key role in that process. Application of the OIPT leads to an understanding of the factors influencing an organization's information processing capacity (Zelt, Schmiedel, & vom Brocke, 2018), which include the tools used within the organization to process information. Galbraith (2012) predicted that big data would become that technology that would finally allow for the increased information processing alternative to becoming an effective solution. Galbraith (2012) envisioned the implementation of big data solutions would be driven by the constant interplay of increasing complexity and interdependence of information and systems within that organization.

Subtheme: Simplicity of design. Process documentation from one of the organizations sets a standard of the simplicity of design for data scientists. The documentation urges data scientists to “make things as simple as possible, but no simpler”. Another document instructed data scientists to “reduce complexity for greater

flexibility and lower cost". This concept was universally reiterated by all 10 participants and was discussed in three of the four process documents. Designing for simplicity is not easy and should be considered a key skill to develop for any data scientist. P1 expressed that big data analytics is not being done correctly today particularly because some data scientists design overly complex solutions. P5 explained:

> The way we process [big data] simplifies it and makes it easier to review…I think the number one thing at play would be to use a simple a process as possible, frankly…I have so much confidence in our processes... because it's not complex. It's not overly complex. It's not muddied. There's not stuff all over the place. We keep it clean, lean, and to the point. Straight forward. The processes and the model itself, it's doing the work. The data doesn't have to do the work.

P6 discussed the importance of keeping solutions simple yet not shying away from complexity when it is required. Finding this balance is a key skillset for any data scientist to effectively govern big data for analytic solutions.

The complexity of big data sources creates a strain on traditional analytic systems because of the nature of structured and unstructured heterogeneous data elements. Complexity in big data comes from the complex data structures, complex data types, and complex data patterns that are inherent with big data (Jin et al., 2015). Heterogeneity adds to the complexity of big data making big data solutions inherently difficult to manage (Yang et al., 2017). Data complexity also adds to the overall complexity of big data as compared to traditional data (Jin et al., 2015). Because of the inherent complexity of big data, the solutions built from multiple big data sources can quickly become very complex. Maintaining simplicity in the constructs of the data governance will assist in

developing more useful analytic systems. Jin et al. (2015) pointed out that among industry experts, there is not a good understanding of how to deal with the complexity that comes from complex data structures, complex data types, and complex data patterns inherent with big data. Data scientists who can simplify these constructs form advantage over those who are still struggling to understand the relationship between data complexity and computational complexity as relates to big data processing

Through the OIPT Galbraith posits that organizations are inherently information processing systems intrinsically programmed to manage uncertainty by gathering, processing, and acting on information from within their environment (Jia et al., 2020). The OIPT builds on contingency theories (Zelt, Recker, et al., 2018) which focus mainly on the attributes within an organization that shape behavior (Chrisman et al., 2018). The OIPT specifically focuses on the essential function of an organizational structure, which is to facilitate the collection, analysis, and distribution of information to reduce uncertainty within the organization (Hwang et al., 2019). As solutions are designed for simplicity they will be more readily adopted within that organization. According to the constructs of the OIPT complex information needs to be made available lower in the organization to reduce the uncertainty of task completion (Feurer et al., 2019). The OIPT helps shape behavior within organizations to take complex information systems and simplify them so that information can be used farther down in the organization by individuals closer to the task. Simplicity is a key attribute of this overall solution.

Subtheme: Using automation. The use of automation in the overall design can add significant value to big data governance solutions. Because of the low veracity attribute of big data, it becomes important to perform preprocessing to improve data

quality (Yang et al., 2017). With the rapid decay of the value of big data the need for automation becomes clear. Eight of the 10 participants mentioned the use of automation to assist in the movement, validation, and correlation of data from big data sources. P4 included automation as part of their job description when discussing their role in the organization. P6 indicated that they "are responsible for developing automated models to drive our pricing …strategies." P7 and P10 stated that they rely very heavily on automated notifications and alerts from their various automated systems that gather data from the big data sources daily. P2 stated that they "have automated validation tests…we've gotten more sophisticated over the past couple of years." Process documentation from one of the organizations reinforces the concept of automation by discussing the need for systems to handle fluctuations of batch processing automatically and without interruption of the overall data flow. P10 indicated that they have "automated processes" and stated that they "highly rely" on them. Without the heavy use of automation, it would be extremely difficult to keep up with the high volume, large velocity, and low veracity of big data and it would be impossible to maintain the validity of the analytic solutions.

The value of some big data elements decays much more rapidly than traditional data. To extract value from many big data solutions the data elements must be processed immediately (Lee, 2017). This cannot be accomplished without automation. The volume and variety attributes of big data also tend to make it very difficult to determine the veracity of those data (Matthias et al., 2017). The processing of big data and the validation of information from big data sources must be automated to provide value to the

organization. Attempting to accomplish big data solutions without automation would be a futile effort.

The central tenet of the OIPT is the need for any organization to close the gap between their need to process information and their inherent ability to process information (Premkumar et al., 2005). Kemp (2014) identified a similar gap between the amount of data that organizations can gather and the ability of that organization to leverage that information in meaningful ways. Galbraith (2012) clarified that organizations can deal with this gap in one of two ways: either they can increase the capacity of the organization to process more information, or they can decentralize the interdependence on that information. Combining heterogeneous sources of data into new information sets is one way to reduce the uncertainty identified in the OIPT (Wang et al., 2016). When considering the temporary value of raw big data and the need to reduce uncertainty within an organization by processing more information it becomes clear that automated solutions are required to process that information quickly while it still has value.

Theme 3: Establishing Data Source Protocols

Clearly defined protocols for gathering and validating data from various sources are critical strategies to respond to the volume, velocity, and variety attributes of big data. In the absence of standards and well-defined processes for data validation, a data scientist could not have confidence in their analysis. Incomplete data can result in partial analysis, which can paint an inaccurate overall picture (Janssen et al., 2017). This theme consists of three subthemes: following defined standards, establishing validation processes, and reducing duplication (see Table 3).

Table 3

Subthemes for Establishing Data Source Protocols

	Interviews		*Documents*	
Subtheme	Count	References	Count	References
Following defined standards	8	18	3	17
Establishing validation processes	10	43	3	4
Reducing duplication	5	16	3	6

Subtheme: Following defined standards. Eight of the 10 participants and three of the four process documents referenced following defined standards. In this context, standards refer to both industry standards and internal organizationally defined standards. P2, P6, and P7 discussed the standard practice of establishing service level agreements (SLAs) with their data source vendors for both timeliness and quality of data. P2 stated that they have "strict SLAs around…external data", and P7 added that they "set up SLAs with our vendors of when that data needs to be available for us to pull in." Establishing SLAs for the timely access and gathering of data from the data sources assures that those data will be available when needed for analysis.

P4, P5, P6, and P9 discussed various internal standards that have been defined within their organization for both quality and timeliness of departmental deliverables. These standards include such practices as peer review of solutions, daily review of exception reports and alerts, and time commitments of deliverables between departments. Process documentation from three of the case organizations referenced the flexibility, economies of scale, and support of heterogeneous environments as benefits of following established open standards. One of those documents explains that the "use of standards

provides the ability to leverage the knowledge and efforts of others. Risk is reduced, proven solutions are implemented, and needless diversity and duplication are prevented." Another document explains that "standardization helps achieve economies of scale, reduces complexity, and improves flexibility." Defining and documenting standards within the organization provides the data scientist with consistency and reliability in their big data governance solutions. Adhering to published industry standards around security and protection ensures the protection of data within big data solutions.

To be successful in the current environment organizations must methodologically exploit their knowledge assets (Mahdi et al., 2019). This implies the establishment of formal processes and procedures around their data assets. Organizations experience substantial financial costs for little value when the organization implements data governance poorly (Wang & Hajli, 2017). Knowledge management practices within an organization are significantly and positively related to sustainable competitive advantage (Mahdi et al., 2019). Establishing and following defined internal standards and adhering to published standards provide the data scientist with a reliable and secure foundation for their big data solutions.

The essential function of any organization as highlighted by the OIPT is to facilitate the collection, analysis, and distribution of information to reduce uncertainly within the organization (Hwang et al., 2019). The fundamental basis for the OIPT is centered around the uncertainty and complexity of tasks to be completed in workflow and based on the concept that organizations should be designed to enable decision-making (Feurer et al., 2019). Because of the vast amount of data involved with complex big data

solutions organizations cannot provide reliable mechanisms for reducing uncertainty without formal standards controlling those solutions.

Subtheme: Establishing validation processes. Validation of data sources is a key protocol that participants identified for multiple data sources due to the low veracity attribute of big data. The volume and variety attributes of big data make it very difficult to determine the veracity of those data (Matthias et al., 2017). This subtheme also aligns well with the subthemes of minimizing data noise and minimizing data sources. This is because many big data sources are often insecure which leads to potentially inaccurate analysis (Duncan et al., 2017). Developing processes for validating data sources is a key governance practice for data scientists.

All 10 of the participants discussed processes that they have in place to validate the various data sources in their analysis. P1 and P5 discussed applying a scoring mechanism to each data source. P1 explained it as follows: “We score the data sources, and we'll pick the [specific data elements] from the most accurate, the most available, and the most trusted data source and that is adjudicated in real-time”. This scoring algorithm provides them with confidence in the data source which can help them to determine which data source to leverage as new questions are posed. The score is updated and maintained as new information arises about the validity of the information from that data source.

P2, P3, P4, P5, P6, P7, and P8 all mentioned performing validation on the data themselves to ensure validity. They each do this in different ways with some performing a line-by-line review personally and others performing spot checks and validating their findings with peers and stakeholders. P6 said:

> We also go through a process of validating data using multiple sources…we are getting data from the competitor, but we also review what that data is telling about the rates that we put in the marketplace…so we have the ability to compare the results …[through] multiple rounds of validation.

P3 said:

> Everything that we do is validated by us personally. We have data coming from our website, our call center, from our stores, from Salesforce... hundreds of different sources and anytime that we use this in our analysis, we validate every column that we're using against some other source that we know is true to make sure that the data [are] correct.

P9 articulated that they do "peer reviews to have multiple eyes take a look at an area …We'll also do two sets of analysis; have two different groups do the analysis and compare the numbers." P1 also described an automated validation process when they said, "We go through a signing process to make sure the data is not being tampered with." Regardless of the specific strategy used every participant was concerned about utilizing some defined, documented, and replicable process for validation of their big data analytics. P5 said:

> In the validation process, if there's anything that needs to be changed, that's where you should start. You get that part sorted first. You get that approved first as quickly as possible. If the data simply can't be validated or can't be approved for whatever reason, then you can deal with it, number one.

Establishing a validation process for data sources and specific data elements is a critical step in the big data governance process.

Organizations tend to have a false sense of confidence because of the hype around big data (Herschel & Miori, 2017). Validating the origin and transformation of data elements contained in big data is one of the primary challenges of big data (Umer et al., 2017). The validity of big data analytics can be called into question due to the multiple sources of big data which are less verifiable than was the case with traditional data (Hashem et al., 2015). Data scientists must establish validation processes for their data sources and for their analytics that provide confidence to the organization in the accuracy of their big data systems. Big data governance is critical to ensure that data are accurate, shared appropriately, and protected (Cheng et al., 2017)

The OIPT includes the mitigation of both uncertainty and equivocality (Duvald, 2019). Management's role in any organization is to reduce both uncertainty and equivocality for their organization (Bartnik & Park, 2018). Ensuring that the analysis of information for task completion and decision making is valid, vetted, and accurate is of critical importance for good decision making. Knowledge management practices within an organization are significantly and positively related to sustainable competitive advantage (Mahdi et al., 2019). If the consumers of the big data analysis are not confident in the results, they will not be reducing uncertainty but creating more uncertainty. The main goal of OIPT is to reduce uncertainty, not increase it.

Subtheme: Reducing duplication. Another strategy that arose from the study was that of reducing the duplication of data. This subtheme aligns well with the subthemes of minimizing data sources, minimizing data noise, and striving for simplicity. Copying data from one location to another is time-consuming and introduces risk to the validity of the data itself. P1 and P5 discussed using metadata to refer to the primary

source of the data elements as a key strategy to reduce the duplication of data. P1 described the value of not duplicating data in this way: “We try and not lose contact with the source data…once you've made a copy of it and moved it, it could have been from anywhere and anything. And we keep our finger on the pulse at the data source”. P10 articulated a similar concept when they explained that they “make sure that I do not duplicate the data…the same information should not appear in different tables in the form of different dimensions and different facts. That is the main thing I would look at. No duplication.”

Two process documents from two different organizations also mentioned the importance of leveraging metadata as a key strategy for big data governance. One document urges data scientists to “minimize redundancy and reduce duplication” because it “helps reduce complexity and promotes greater efficiency”. Utilizing metadata reduces risk by leaving the source data untouched and maintaining information about where the source data reside. P5 reinforces this concept when they explain that “I think that number one…would have to be metadata and the critical importance of that. It has to be clean, crisp, consistent information and you need to get as much detail as possible about that dataset”. By refraining from copying the data and analyzing the data where it resides data scientists can avoid these traditional challenges.

Challenges arise with big data analytics from the issues relating to data that are too vast, unstructured, and moving too fast to be managed by traditional means (Zakir et al., 2015). Using traditional strategies of copying data for analytics purposes creates significant challenges due to the rapid decay of many big data elements. Leveraging metadata is one way to normalize unstructured big data elements for analytic purposes

(Yang et al., 2017). Traditional data copy and retention methods break down when applied to big data because of the attribute of high velocity. Storage of big data becomes challenging as transient data flows through the analyzing systems (Yang et al., 2017). Minimizing the copying and duplication of data is a key governance strategy for data scientists as they transition from traditional analytic methods to big data analytics.

The OIPT directly applies to the gathering, interpretation, and synthesis of information within an organization for decision making (Tushman & Nadler, 1978). The gap identified by the OIPT directly relates to the ability of an organization to gather and analyze data. There is a limit to the amount of data that an organization can gather and store, but by analyzing the data where it resides organizations increase their ability to analyze more data. By designing systems that analyze data without requiring that those data be copied onto local storage devices, organizations increase their ability to process larger amounts of data than what was previously available to them due to the physical limitations of local storage. Ensuring data source protocols by following defined standards, establishing validation processes, and reducing duplication of data can increase the validity of big data analytics by data scientists.

Theme 4: Designing for Security

Protection of security and privacy are critical data governance concepts for any solution involving personal data. Data scientists are among the first line of defense when it comes to protecting both the raw data and the individual privacy of the consumers who are represented in those data. With the ever-increasing threat of a data breach, any organization must ensure the protection of the information within its control. This theme

consists of three subthemes: segregation of duties, using encryption, and protecting private information (see Table 4).

Table 4

Subthemes for Designing for Security

	Interviews		*Documents*	
Subtheme	Count	References	Count	References
Segregation of duties	9	35	3	17
Using encryption	3	9	4	10
Protecting private information	10	15	4	10

Subtheme: Segregation of duties. Security and privacy are best addressed in the data governance process before the data scientists have access to those data for analysis. When asked about the protection of the data nine out of 10 participants discussed the segregation of duties concerning the data lineage within their organization. Additionally, three of the four process documents referenced segregation of duties. In every organization studied there were either data owners or data stewards who were responsible for protecting access to the information. P2 described that data custodians in their organization prepare the various data environments for the data scientists to access and stated that "when the environment is ready for [data scientists] to work in, they wouldn't see [private] data". If a data scientist needed additional information to perform their analysis, they would need to obtain permission from the data owner or data steward before they were able to access the information. P3 articulated this process as follows: "if there's something that we want that isn't available, we have to go through steps to get that from the BI team…[they would] be the ones that need to give us access to that". Security

controls on the various systems of record were active to prevent such access without appropriate permission.

P3, P7, P8, P9, and P10 also discussed having role-based access controls within their systems that restrict access to private data based on role. P3 indicated that the "BI team…would be considered more of the gatherers, and then on the data science side we do more of the analysis". P8 highlighted a role distinction between the data architect and reporting writing roles as follows: "I'm not a report writer, I'm not an architect... I kind of come before the report, I pull together the requirements". Process documentation from one of the organizations also specifies that "effective [segregation of duties] ensures access to company…data is restricted to only authorized and appropriate personnel". They also discussed the segregation of duties among the IT team with database administrators controlling access to the data while data scientists, business analysts, and business intelligence roles all accessing only the information that they require to accomplish their job.

Maintaining appropriate segregation of duties between those who are responsible for data access and those who analyze information is critical to the protection of the data. P6 spoke of the inherent conflict that exists between data owners who want to protect data and restrict access and data scientists who want to access and review as much data as possible. P6 said:

> There is some conflict between data science and data warehouse on what data is available to the data science team. The data science team wants access to …more data, more raw data, more extensive data. There is some unwillingness to provide

> access to all of the data that is needed…there are concerns about privacy and other aspects.

This is a healthy conflict that forces data scientists to justify all of the access to data in terms of a business question that they are working to answer. This highlights the alignment between this subtheme and the subtheme of understanding the business need.

The primary ethical issue is not whether to collect the information, but how and when it is ethically responsible to analyze that information (Mai, 2016). Herschel and Miori (2017) stated that data sources that had not previously had an issue with privacy may end up threatening privacy once combined with other data sources in the big data construct. Big data analytics must be comprised of various roles and assignments of specific responsibilities, such as technical data steward and business data steward (Begg & Caira, 2012). Data owners need to stay focused on their role of protecting individual privacy and data security of the information that they gather. Maintaining segregation of duties between those who are responsible for gathering the data, those who are responsible for analyzing the data, and those who are responsible for protecting the data creates natural boundaries around data governance within the organization that lead toward better security overall.

Segregation of duties is not specifically addressed in OIPT. However, the concepts related to pushing decision making down in the organization are relevant to this theme. The premise of OIPT is to minimize uncertainty within an organization by closing the gap between the amount of information needed for a task and the ability to process information around the task (Premkumar et al., 2005). This means that according to the constructs of OIPT those data need to be made available lower in the organization (Feurer

et al., 2019). Making information available lower in the organization implies having a larger audience for the information which in turn implies a greater need for security and privacy controls. By granting information to the individuals who need it when they need it, role-based access to that information becomes an important security control to limit access to private information. The net effect of pushing the decision making down within the organization aligns with the concept of segregating duties and limiting access to private and protected information to only those individuals who need to see those private data as part of their role within the organization.

Subtheme: Using encryption. Encryption of data at rest and data in transit has long been a foundational security control of data governance. Many regulated industries require that encryption is implemented as one of many security controls to protect data against unauthorized access. When asked about the protection of the data three of the 10 participants specifically mentioned encryption. P5 indicated that "the data is encrypted. So, it is not readable or hackable in any way. We keep that consistency and the data is always protected." P8 said:

> Files don't go out the door without some form of protection. …our file transfers now happen via SFTP. We also wrap those files in a PGP or GPG encryption…we also want to make sure… that wherever that data is sitting and wherever it eventually lands that it is encrypted at rest. So, it is encrypted as it sits, it is basically double encrypted as it goes. And then it's still encrypted as it sits at its destination.

Use of encryption is a common practice to protect personal information for both data at rest and data in transit.

The remaining seven participants stated that their security team has controls in place for the protection of the data which does not rule out encryption being used as part of that protection. In addition, three of the five process documents referred to encryption of the data as a key security control. One of those documents indicated that “it is important that we comply with security requirements, laws, and regulations as we design our systems.” P5 also discussed the idea that consumers are concerned about sharing their data with companies who are not protective of those data and stated that encryption helps provide a level of comfort for their customers.

Because of the volume and velocity of big data traditional encryption at rest is inefficient when applied to big data solutions (Yang et al., 2017). Encryption in transit is also questionable when applied to big data as many data sources associated with big data lack advanced security controls (Sivarajah et al., 2017). It is because of these challenges that data scientists must take additional care within their environments to proactively enforce encryption of data both at rest and in transit to enhance the security of big data solutions. Encryption technology itself must also improve to handle the speed requirements that businesses are demanding from big data analytics.

Similar to the theme of segregation of duties the specific concept of encrypting data is not directly addressed in the OIPT. However, the safeguarding of information is critical to maintaining confidence in any information-sharing solution within an organization. The OIPT defined the essential function of an organizational structure to facilitate the collection, analysis, and distribution of information (Hwang et al., 2019). Alignment between overall information systems strategy and business strategy is a critical precursor to overall profitability and competitive advantage (Shao, 2019). Data

governance of that information refers to the control and authority over data-related rules and accountability of individuals and systems accessing and utilizing those data (Hashem, et al., 2015). Ensuring the protection of those data through industry-standard security controls such as encryption is crucial to the ability of an organization to share information widely throughout the various organizational levels of the organization. This is critical to the success of the concepts outlined in OIPT.

Subtheme: Protecting private information. Protecting the privacy of individuals is critical to organizations that are analyzing big data. Individuals are revealing very personal information on social networks and through IoT devices often unconsciously (Gao et al., 2018) and without their knowledge or consent (Mai, 2016). When implementing data governance over big data it is critical for the data scientists to be aware of personally identifiable information (PII) within their system and to proactively put controls in place to protect those data.

When asked about privacy 10 out of 10 participants indicated that they are cognizant of privacy and that it is important to their organization. P4 explained that “we don't use customer identifying information for any analysis, but …there's one process that uses that and we hash [the PII to] protect the data.” Additionally, one of the process documents describes that the system in use in that organization will automatically determine whether the information being requested is PII and if so, will return obfuscated data rather than the raw data. P8 discusses the importance of ensuring that appropriate data sharing agreements are in place before sharing any PII. P1 explained that they obtain consumers' permission prior to accessing their information: “if I want to look at Patient X's x-rays, Patient X has to authorize that”. P2 explained that they scrub the PII from the

system on-demand: “we've invested in some scrubbing technology. So, when we, even on the application development side, if an engineer has to pull data down to a local environment…the PII’s scrubbed.”

Regardless of the specific manner in which the personal data are protected, it is critical for organizations to understand what personal information they have access to and to put processes and controls in place to protect that information. Big data can include many elements of personal information (Lee, 2017). 99.98% of Americans could be reidentified with any data set with as few as 15 demographic data elements (Rocher et al., 2019). Data scientists analyzing big data can no longer assume that consumers have provided informed, rational consent for their data to be used by the organization (Mai, 2016). This means that organizations analyzing big data have an ethical responsibility to understand how to use the data while protecting the privacy and confidentiality of those data (Herschel & Miori, 2017).

Tushman and Nadler (1978) articulated that the OIPT directly applies to the gathering, interpretation, and synthesis of information within an organization for decision making. If individuals have not provided their consent for such use the organization could be at risk. One of the general concepts of the OIPT is that as lateral relationships are established, the number of decisions referred upward is reduced, which then leads to more effective decision-making (Feurer et al., 2019). This means that according to the constructs of the OIPT those data need to be made available lower in the organization (Feurer et al., 2019). Hwang et al. (2019) added that OIPT highlights the essential function of an organizational structure, which is to facilitate the collection, analysis, and distribution of information to reduce uncertainty within the organization. Considering the

privacy requirements enacted in recent legislation in the form of the California Consumer Protection Act the organization must focus on privacy before providing access to such information for decision making within the organization. By implementing the strategies of segregation of duties, use of encryption, and protection of privacy data scientists can successfully design their solutions for security.

Applications to Professional Practice

By implementing the concepts outlined in these four themes, data scientists can implement governance practices that will assist them in working with big data. The four themes described in this study guide data scientists to provide a foundation for big data analytics that will ease the transition from traditional data analytics to the use of big data for decision making within the organization. These practices can help provide better information and insights to those individuals who are closer to the work within the organization as theorized in the OIPT. The implementation of the practices associated with these four themes will allow individuals within the organization to work more autonomously and to close the gap between information needed and information available for the front-line employees. As described in the key tenant of the OIPT this application to professional practice should result in better efficiencies throughout the organization.

Data scientists who use the strategies described in these findings could improve their effectiveness as agents of change for their organization. Adopting big data governance practices that provide a balance between value creation and risk mitigation is imperative to organizations (Hashem et al., 2015). Learning and implementing these practices could make data scientists more valuable to their organization which could

result in an increase in both efficiencies of the organization and improved value of the data scientist. By implementing the governance practices outlined in this study the data scientist could improve their position within the organization by bringing them closer to the business itself. This increased focus on the business will benefit both the role of the data scientist as well as the efficiency of the solutions that the data scientist can provide to the organization.

Implications for Social Change

If the implementation of these practices results in the efficiencies theorized in the OIPT, the result would be better decisions being made by the organization. Better decision making within the organization would occur because the information would be made available lower in the organization and closer to those who need the information to do their job. Making better information available lower in the organization would in turn result in increased responsiveness and time to market for the organization, which could provide the overall organization with a significant competitive advantage. Competitive advantage could lead to increased profitability for the organization. The cumulative effect of these practices could lead to more employment opportunities for the communities where these mid-market companies are based.

Implementation of these practices could also result in more efficient use of data within the organization, resulting in increased profitability for the overall organization. Increasing the efficiency of data use means that the potential for social change is the increased ability for companies to meet the needs of their customer base because they now have an increased understanding of the meaning of the data that they have. These business improvements could benefit the local communities where the company is

located by allowing the employees to earn more money and improve their way of life. Improved business performance could also lead to better employment opportunities and improved customer experiences. Effective use of the big data themes in this study could benefit customers by providing them more specific and targeted solutions to their needs by the small companies that are part of their local communities. In sum, the ability to effectively utilize big data in mid-market companies could be an equalizing force that could benefit customers, employees, and owners of all types of mid-market companies.

A social change could also result from data scientists enacting the security strategies outlined in the findings from this study. Weak security controls around big data can lead to financial loss and reputational damage (Lee, 2017). Implementing the security controls outlined in the findings can help ensure better security of big data solutions and thus protect the reputation of the institution and mitigate financial losses. By implementing the strategies relating to the segregation of duties, encryption of data, and protection of personal information, data scientists can mitigate the contemporary concerns relating to the use of private information in big data analytics.

The use of multiple data sources that may themselves be insecure leads to potentially insecure solutions (Duncan et al., 2017). During the completion of this study, Simon Weckart tricked Google Maps into reporting that 100 cell phones in a wagon were a traffic jam demonstrating the ease with which some big data systems can be fooled by injection of IoT devices (Torres, 2020). Validation of data sources can mitigate this concern and enhance the overall security of big data analytic solutions, which could result in less possibility of injection of malicious and falsified data streams into AI-driven analytic systems. By protecting the privacy of individuals and ensuring the validity of the

data sources, positive social change can be accomplished in the embodiment of trusted solutions for autonomous artificial intelligence systems.

Recommendations for Action

Policies, principles, and frameworks around how big data are used and leveraged within an organization are posing enormous challenges for those organizations (Hashem et al., 2015). The results of this study validate this concept from literature. Data governance is critical for big data, and data scientists cannot add it as an afterthought (Cervone, 2016). By following the findings of this study, data scientists can thoughtfully and proactively implement data governance into their big data solutions from the start. To be able to extract value from raw data, organizations require a well-defined, systematic approach to governing their data (Song & Zhu, 2016). Data scientists can leverage the data governance strategies outlined in the findings of this study to design appropriate governance models for their big data initiatives.

Data scientists should review the findings of this study to determine how best to apply these practices into their big data solution architectures. The lack of formal data governance processes was one of the primary reasons that organizations fail to utilize their data effectively (Posavec & Krajnovic, 2016). Implementation of these themes can guide data scientists who are trying to move from traditional data analytics to big data analytics. To be able to extract value from raw data, Song and Zhu (2016) noted that organizations require a well-defined, systematic approach to governing their data. The themes outlined in this study provided a foundation for such a systematic approach to big data governance. Begg and Caira (2012) argued that there is no evidence to date of a standard implementation framework for data governance. This study helps to set a

foundation for that standard implementation framework. As data scientists implement these themes, they will also be establishing a set of best practices within the industry. Data scientists should review the themes identified in this study and should look for ways to implement these concepts as they implement big data analytics within their organizations.

Recommendations for Further Study

This study focused on the general data governance practices of data scientists who are utilizing big data to help make decisions within their organization. Data governance of big data was a broad topic about a subject that is in its infancy within current practice. Big data by itself is of little use when considering the application to an organization. The power of big data comes when combined with both IoT and Machine Learning. I would recommend that future studies focus on the combination of these three technologies. IoT is creating additional data sources for big data, and machine learning is starting to provide faster analysis of those data. As future studies focus on the powerful combination of big data governance combined with IoT and machine learning, more insights will emerge regarding the benefits of big data to the decision-making process within an organization.

Security, privacy, and ethical use of big data are also critical aspects of data governance when implementing big data solutions. Re-identification of previously obfuscated and anonymous data has been documented in multiple cases (Herschel & Miori, 2017). Additional study is warranted in these areas to determine the effectiveness of the security controls and the extent to which the strategies outlined here are effective in reducing the ability to correlate big data elements into data sets that contain private

information not previously permitted by the data owner. The security strategies outlined in this study should be further studied to determine their effectiveness.

Reflections

As I reflect on this process, my thoughts turn first to the current state of the practice of big data governance. Documentation regarding data governance processes lacks in all the organizations that I studied. Each organization had documentation regarding their security processes, and one had design documentation regarding their overall approach to big data analysis. Still, very little of the documentation was focused on the data governance aspects of how to handle data within their organization. Most of the documentation was focused on overall IT practices and only lightly touched on the data specifically. While design specifications are important, it is also important to document decisions around the governance process. It is important to have documentation stating why certain governance decisions were made and how to implement those decisions in practice.

It was also clear to me that the participants validated the lack of clarity described in literature around the definition of big data. Some of the participants did not think they were working with big data until I discussed the definition of big data with them. There are also still many different perspectives on how to handle big data in the future. One participant indicated that they should not collect any more data until they can effectively utilize all of the data collected already, and another participant opined that they should collect much more data and simply store it until they can determine the best use for it. While there was very good alignment around the four themes that emerged from this

study, there were also many conflicting ideas about some of the other strategies in use today.

Another concept that was identified throughout this process is the fact that the data scientists that I spoke with take their responsibilities very seriously. Each one was very conscious of the fact that they do not own the data that they are accessing, and the need to maintain privacy and security is paramount to them. As I reviewed my notes from the interviews, I noticed that there are several occasions where I noted from their body language that the participant took their role as data steward seriously and that they clearly understood that they were interacting with other people's data. The stewardship that these participants feel became clear early in the process when some potential participants declined to participate because they were concerned about sharing too much information about the data that they gather. The concept of data stewardship versus data ownership is well understood by the individuals that I interacted with for this study.

As I reflect on the process, my thoughts also turn to the environment in which I conducted the study. I entered the data gathering phase of this study during the COVID-19 pandemic of 2020. The environmental changes that came as a result of the pandemic caused some added challenges to the data gathering process that I had not previously anticipated. Some of the partner organizations had to delay scheduling interviews because their employees were busy responding to the challenges of providing support to their company while employees transitioned to working from home. Others had to delay their participation because their department had gone through a layoff, and the remaining employees were dealing with the emotional stress of having lost some co-workers while having to take on the added work that was left by these departing employees. These

various effects of the pandemic caused the data gathering phase to take longer than I had initially anticipated.

I was surprised at the difficulty of finding partner organizations willing to participate in this study. During the proposal phase of my study, I had identified two organizations that I was planning to use as partner organizations. I spoke with the leadership of both organizations, and they both seemed excited to participate. Once I obtained the approval of the proposal and approached these two organizations to obtain their Letter of Cooperation, both backed out. In one case, the leadership was uncomfortable with the fact that they would not be able to review and approve the individual interview content before inclusion in my data analysis. They were concerned that something might be revealed that might reflect poorly on their company, even though I assured them that confidentiality would be paramount. As I searched for other organizations to participate, it was challenging to find organizations mature enough in their big data processes to participate in this study. In one case, the potential partner company had been gathering quite a bit of data from IoT devices, but they were not yet doing any analysis of those data. The difficulty of locating companies that are currently using big data to make decisions is a clear indication that we are still incredibly early in the adoption of big data analytics as a true strategic capability within mid-market organizations.

After I found organizations that met the criteria, I then met with the challenge of obtaining permission from the potential participants to agree to the interview process. Several potential participants declined to participate because they were uncomfortable sharing details of their data governance practices outside of their organization. As I

discussed the process with potential participants, it became clear to me that many of the individuals involved in the data gathering and analysis process take their role in protecting information very seriously. Thus, they are more reluctant to discuss details of their process with others. While not completely surprising, this reluctance to share details regarding their big data governance processes did pose an additional challenge to the overall study process.

Summary and Conclusions

The focus of this study was to explore big data governance strategies that data scientists are currently using in practice. I presented four major themes that provide insights into those strategies: ensuring business centricity, striving for simplicity, establishing data source protocols, and designing for security. Implementation of these strategies can assist mid-market organizations in making the transition from traditional data analytics to big data analytics, which could, in turn, help those organizations to be more profitable by gaining competitive advantages. I explained the possibility of social change in the way that individuals' private information is gathered as part of a big data strategy. Following the strategies outlined in the four themes of this study for big data governance can lead to the improved overall protection of individual privacy.

References

Abdalla, M. M., Oliveira, L. G. L., Azevedo, C. E. F., & Gonzalez, R. K. (2018). Quality in qualitative organizational research: Types of triangulation as a methodological alternative. *Administration: Teaching and Research*, *19*(1), 66-98. doi:10.13058/raep.2018.v19n1.578

Adams, C., & van Manen, M. A. (2017). Teaching phenomenological research and writing. *Qualitative Health Research*, *27*(6), 780-791. doi:10.1177/1049732317698960

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. *Computer Networks*, *129*, 459-471. doi:10.1016/j.comnet.2017.06.013

Akbar, H., Baruch, Y., & Tzokas, N. (2017). Feedback loops as dynamic processes of organizational knowledge creation in the context of the innovations' front-end. *British Academy of Management*, *29*(3), 445-463. doi:10.1111/1467-8551.12251

Allegrini, V., & Monteduro, F. (2018). The role of uncertainty in performance information disclosure. *International Journal of Public Sector Management*, *31*(5), 583-598. doi:10.1108/IJPSM-08-2017-0216

Alreemy, Z., Chang, V., Walters, R., & Wills, G. (2016). Critical success factors (CSFs) for information technology governance (ITG). *International Journal of Information Management*, *36*(6, Part A), 907-916. doi:10.1016/j.ijinfomgt.2016.05.017

Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2016). A conceptual framework for designing data governance for cloud computing. *Procedia Computer Science*, *94*, 160-167. doi:10.1016/j.procs.2016.08.025

Arseven, I. (2018). The use of qualitative case studies as an experiential teaching method in the training of pre-service teachers. *International Journal of Higher Education*, *7*(1), 111-125. doi:10.5430/ijhe.v7n1p111

Bartnik, R., & Park, Y. (2018). Technological change, information processing, and supply chain integration: A conceptual model. *Benchmarking: An International Journal*, *25*(5), 1279-1301. doi:10.1108/BIJ-03-2016-0039

Baruh, L., & Popescu, M. (2017). Big data analytics and the limits of privacy self-management. *New Media, & Society*, *19*(4), 579-596. doi:10.1177/1461444815614001

Begg, C., & Caira, T. (2012). Exploring the SME quandary: Data governance in practise in the small to medium-sized enterprise sector. *Electronic Journal of Information Systems Evaluation, 15*(1), 11. Retrieved from http://www.ejise.com/

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59.doi:10.1016/j.inffus.2015.08.005

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open, 2*, 8-14. doi:10.1016/j.npls.2016.01.001

Bilal, K., Khalid, O., Erbad, A., & Khan, S. U. (2018). Potentials, trends, and prospects in edge technologies: Fog, cloudlet, mobile edge, and micro data centers. *Computer Networks*, *130*, 94-120. doi:10.1016/j.comnet.2017.10.002

Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., … Evans, J. A. (2018). Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, *115*(50), 12630-12637. doi:10.1073/pnas.1804247115

Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, *70*, 328-337. doi:10.1016/j.jbusres.2016.08.006

Brown, A., & Danaher, P. A. (2019). CHE Principles: facilitating authentic and dialogical semi-structured interviews in educational research. *International Journal of Research & Method in Education*, *42*(1), 76-90. doi:10.1080/1743727X.2017.1379987

Bughin, J. (2016). Big data, big bang? *Journal of Big Data*, *3*(1), 2. doi:10.1186/s40537-015-0014-3

Burton, R. M., Obel, B., & Håkonsson, D. D. (2015). How to get the matrix organization to work. *Journal of Organization Design*, *4*(3), 37. doi:10.7146/jod.22549

Butler, M. J. R., & Ferlie, E. (2019). Developing absorptive capacity theory for public service organizations: Emerging UK empirical evidence. *British Journal of Management*. doi:10.1111/1467-8551.12342

Caldamone, A. A., & Cooper, C. S. (2017). The institutional review board. *Journal of Pediatric Urology*, *13*(6), 557-558. doi:10.1016/j.jpurol.2017.10.006

Cao, G., Duan, Y., & Cadden, T. (2019). The link between information processing capability and competitive advantage mediated through decision-making

effectiveness. *International Journal of Information Management, 44,* 121-131. doi.org:10.1016/j.ijinfomgt.2018.10.003

Carbonell, I. M. (2016). The ethics of big data in big agriculture. *Internet Policy Review, 5*(1), 1-13. doi:10.14763/2016.1.405

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *The Qualitative Report*, *21*(5), 811-830. Retrieved from http://nsuworks.nova.edu/tqr

Cervone, H. F. (2016). Organizational considerations initiating a big data and analytics implementation. *Digital Library Perspectives*, *32*(3), 137-141. doi:10.1108/DLP-05-2016-0013

Chadwick, S. A., & Pawlowski, D. R. (2007). Assessing institutional support for service-learning: A case study of organizational sensemaking. *Michigan Journal of Community Service Learning*, *9*. Retrieved from https://eric.ed.gov

Chaudhary, S., & Batra, S. (2018). Absorptive capacity and small family firm performance: Exploring the mediation processes. *Journal of Knowledge Management*, *22*(6), 1201-1216. doi:10.1108/JKM-01-2017-0047

Chauhan, S. K., & Sangwan, S. (2017). Big data analytics. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(4), 4. Retrieved from http://www.ijritcc.org/

Chen, H., & Zhang, Y. (2017). Educating data management professionals: A content analysis of job descriptions. *The Journal of Academic Librarianship*, *43*(1), 18-24. doi:10.1016/j.acalib.2016.11.002

Cheng, G., Li, Y., Gao, Z., & Liu, X. (2017). Cloud data governance maturity model. *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 4*. doi:10.1109/ICSESS.2017.8342968

Chrisman, J. J., Chua, J. H., Le Breton-Miller, I., Miller, D., & Steier, L. P. (2018). Governance mechanisms and family firms. *Entrepreneurship Theory and Practice, 42*(2), 171-186. doi:10.1177/1042258717748650

Clark, K. R., & Vealé, B. L. (2018). Strategies to enhance data collection and analysis in qualitative research. *Radiologic Technology*, *89*(5), 5. Retrieved from http://www.radiologictechnology.org/

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128. doi:10.2307/2393553

Collins, C. S., & Cooper, J. E. (2014). Emotional intelligence and the qualitative researcher. *International Journal of Qualitative Methods*, *13*(1), 88-103. doi:10.1177/160940691401300134

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine, 91*(10), 1359-1369. doi:10.1097/ACM.0000000000001175

Cseko, G. C., & Tremaine, W. J. (2013). The role of the institutional review board in the oversight of the ethical aspects of human studies research. *Nutrition in Clinical Practice*. doi:10.1177/0884533612474042

Cypress, B. S. (2017). Rigor or reliability and validity in qualitative research: perspectives, strategies, reconceptualization, and recommendations. *Dimensions*

of Critical Care Nursing, *36*(4), 253-263. doi:10.1097/DCC.0000000000000253

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics (pp. 97-104). doi:10.1063/1.4907823

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. doi:10.1016/j.ipm.2017.05.004

Dikko, M. (2016). Establishing construct validity and reliability: Pilot testing of a qualitative interview for research in Takaful (Islamic Insurance). *The Qualitative Report*, *21*(3), 10. Retrieved from https://nsuworks.nova.edu/tqr

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organization fields. *American Sociological Review*, *48*(2), 147-160. doi:10.2307/2095101

Dourish, P., & Cruz, G. E. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data, & Society*, *5*(2), 205395171878408. doi:10.1177/2053951718784083

Draper, J. (2015). Ethnography: Principles, practice, and potential. *Nursing Standard*, 29(36), 36-41. doi:10.7748/ns.29.36.36.e8937

Duncan, B., Whittington, M., & Chang, V. (2017). Enterprise security and privacy: Why adding IoT and big data makes it so much more difficult. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-7). Antalya: IEEE. doi:10.1109/ICEngTechnol.2017.8308189

Dunne, B., Pettigrew, J., & Robinson, K. (2016). Using historical documentary methods to explore the history of occupational therapy. *British Journal of Occupational Therapy*, *79*(6), 376-384. doi:10.1177/0308022615608639

Duvald, I. (2019). Exploring reasons for the weekend effect in a hospital emergency department: An information processing perspective. *Journal of Organization Design*, *8*(1). doi:10.1186/s41469-019-0042-0

Eisenhauer, E. R., Tait, A. R., Rieh, S. Y., & Arslanian-Engoren, C. M. (2019). Participants' understanding of informed consent for biobanking: A systematic review. *Clinical Nursing Research*, *28*(1), 30-51. doi:10.1177/1054773817722690

Elman, C., Gerring, J., & Mahoney, J. (2016). Case study research: Putting the quant into the qual. *Sociological Methods, & Research, 45,* 375–391. doi:10.1177/0049124116644273

Feurer, S., Schuhmacher, M. C., & Kuester, S. (2019). How pricing teams develop effective pricing strategies for new products: Pricing teams and new product pricing strategies. *Journal of Product Innovation Management*, *36*(1), 66-86. doi:10.1111/jpim.12444

Fink, A. S. (2000). The role of the researcher in the qualitative research process. A potential barrier to archiving qualitative data. *Forum: Qualitative Social Research*, *1*(3), 16. doi:10.17169/fqs-1.3.1021

Flyverbom, M., Deibert, R., & Matten, D. (2019). The governance of digital technology, big data, and the internet: New roles and responsibilities for business. *Business, & Society*, *58*(1), 3-19. doi:10.1177/0007650317727540

Foerstl, K., Meinlschmidt, J., & Busse, C. (2018). It's a match! Choosing information processing mechanisms to address sustainability-related uncertainty in sustainable supply management. *Journal of Purchasing and Supply Management*, *24*(3), 204-217. doi:10.1016/j.pursup.2018.02.002

Forero, R., Nahidi, S., De Costa, J., Mohsin, M., Fitzgerald, G., Gibson, N., … & Aboagye-Sarfo, P. (2018). Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine. *BMC Health Services Research*, *18*(1). doi:10.1186/s12913-018-2915-2

Freitas, F., Ribeiro, J., Brandão, C., de Almeida, C. A., & de Souza, F. N. (2019). How do we like to learn qualitative data analysis software? The Qualitative Report, *24*(13), 21. Retrieved from https://nsuworks.nova.edu/tqr/vol24/iss13/8/

Fusch, P. I., & Ness, L. R. (2015). Are we there yet?: Data saturation in qualitative research. *The Qualitative Report, 20,* 1408–1416. Retrieved from http://nsuworks.nova.edu/tqr

Galbraith, J. (2014). Organizational design challenges resulting from big data. Journal of Organization Design, *3*(1), 2-13.

Galbraith, J. R. (1974). Organization Design: An information processing view. *Interfaces*, *4*(3), 28-36. doi:10.1287/inte.4.3.28

Galbraith, J. R. (2012). The future of organization design. *Journal of Organization Design, 1*(1), 3-6. doi:10.7146/jod.2012.1.2

Galbraith, S. (2017). Jay R. Galbraith. In D. B. Szabla, W. A. Pasmore, M. A. Barnes, & A. N. Gipson (Eds.), *The Palgrave Handbook of Organizational Change Thinkers*

(pp. 1-20). Cham: Springer International Publishing. doi:10.1007/978-3-319-49820-1_39-1

Galliers, R. D., Newell, S., Shanks, G., & Topi, H. (2017). Datification and its human, organizational and societal effects: The strategic opportunities and challenges of algorithmic decision-making. *The Journal of Strategic Information Systems*, *26*(3), 185-190. doi:10.1016/j.jsis.2017.08.002

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007

Gao, W., Liu, Z., Guo, Q., & Li, X. (2018). The dark side of ubiquitous connectivity in smartphone-based SNS: An integrated model from information perspective. *Computers in Human Behavior, 84*, 185-193. doi:10.1016/j.chb.2018.02.023

Gardiner, A., Aasheim, C., Rutner, P., & Williams, S. (2018). Skill requirements in big data: a content analysis of job advertisements. Journal of Computer Information Systems, *58*(4), 374-384. doi:10.1080/08874417.2017.1289354

Gehman, J., Glaser, V. L., Eisenhardt, K. M., Gioia, D., Langley, A., & Corley, K. G. (2018). Finding theory–method fit: A comparison of three qualitative approaches to theory building. *Journal of Management Inquiry*, *27*(3), 284-300. doi:10.1177/1056492617706029

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: Interviews and focus groups. *British Dental Journal*, *204*(6), 291-295. doi:10.1038/bdj.2008.192

Government Accountability Office. (2017). Internet of things, Status and implications of an increasingly connected world. (GAO Publication No. 17-75). Washington, D.C.: U.S. Government Printing Office. Available at https://www.gao.gov/products/GAO-17-75

Grossman, R. L. (2018). A framework for evaluating the analytic maturity of an organization. *International Journal of Information Management*, *38*(1), 45-51. doi:10.1016/j.ijinfomgt.2017.08.005

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems*, *35*(2), 388-423. doi:10.1080/07421222.2018.1451951

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, *26*(3), 191-209. doi:10.1016/j.jsis.2017.07.003

Gupta, S., Kumar, S., Kamboj, S., Bhushan, B., & Luo, Z. (2019). Impact of IS agility and HR systems on job satisfaction: An organizational information processing theory perspective. *Journal of Knowledge Management*. doi:10.1108/JKM-07-2018-0466

Halcomb, E., & Hickman, L. (2015). Mixed methods research. *Nursing Standard: promoting excellence in nursing care*, *29* (32*)*, 41-47. doi:10.7748/ns.29.32.41.e8858

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, *47*, 98-115. doi:10.1016/j.is.2014.07.006

Haußmann, C., Dwivedi, Y. K., Venkitachalam, K., & Williams, M. D. (2012). A summary and review of Galbraith's organizational information processing theory. *Information Systems Theory (Vol. 29)*. Springer, New York, NY. doi:10.1007/978-1-4419-9707-4_5

Herschel, R., & Miori, V. M. (2017). Ethics, & big data. *Technology in Society*, *49*, 31-36. doi:10.1016/j.techsoc.2017.03.003

Hu, H., Luo, Y., Wen, Y., Ong, Y.-S., & Zhang, X. (2018). How to find a perfect data scientist: A distance-metric learning approach. *IEEE Access*, *6*, 60380-60395. https://doi.org/10.1109/ACCESS.2018.2870535

Hwang, S., Kim, H., Hur, D., & Schoenherr, T. (2019). Interorganizational information processing and the contingency effects of buyer-incurred uncertainty in a supplier's component development project. *International Journal of Production Economics*, *210*, 169-183. doi:10.1016/j.ijpe.2019.01.019

Imran, A., & Yusoff, R. M. (2015). Empirical validation of qualitative data: A mixed method approach. *International Journal of Economics and Financial Issues, 5*(1). Retrieved from www.econjournals.com/index.php/ijefi

Intezari, A., & Gressel, S. (2017). Information and reformation in KM systems: Big data and strategic decision-making. *Journal of Knowledge Management*, *21*(1), 71-91. doi:10.1108/JKM-07-2015-0293

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1). doi:10.1186/s40537-016-0059-y

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research, 70*, 338-345. doi:10.1016/j.jbusres.2016.08.007

Jia, F., Blome, C., Sun, H., Yang, Y., & Zhi, B. (2020). Towards an integrated conceptual framework of supply chain finance: An information processing perspective. *International Journal of Production Economics, 219*, 18-30. doi:10.1016/j.ijpe.2019.05.013

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Visions on Big Data, 2*(2), 59-64. doi:10.1016/j.bdr.2015.01.006

Kassan, A., Goopy, S., Green, A., Arthur, N., Nutter, S., Russell-Mayhew, S., … Silversides, H. (2018). Becoming new together: Making meaning with newcomers through an arts-based ethnographic research design. *Qualitative Research in Psychology*, 1-18. doi:10.1080/14780887.2018.1442769

Kelly, K. (2017). A different type of lighting research – A qualitative methodology. *Lighting Research, & Technology*, *49*(8), 933-942. doi:10.1177/1477153516659901

Kemp, R. (2014). Legal aspects of managing big data. *Computer Law, & Security Review, 30*(5), 482-491. doi:10.1016/j.clsr.2014.07.006

Kongnso, F. (2015). Best practices to minimize data security breaches for increased business performance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3739769)

Korstjens, I., & Moser, A. (2018). Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. *European Journal of General Practice, 24*(1), 120-124. doi:10.1080/13814788.2017.1375092

Kroh, J., Luetjen, H., Globocnik, D., & Schultz, C. (2018). Use and efficacy of information technology in innovation processes: The specific role of servitization. *Journal of Product Innovation Management, 35*(5), 720-741. doi:10.1111/jpim.12445

Latkin, C. A., Mai, N. V. T., Ha, T. V., Sripaipan, T., Zelaya, C., Le Minh, N., … Go, V. F. (2016). Social desirability response bias and other factors that may influence self-reports of substance use and HIV risk behaviors: A qualitative study of drug users in Vietnam. *AIDS Education and Prevention: Official Publication of the International Society for AIDS Education, 28*(5), 417-425. doi:10.1521/aeap.2016.28.5.417

Lawson, L. V. (2018). Documentary analysis as an assessment tool. *Public Health Nursing*, *35*(6), 563-567. doi:10.1111/phn.12520

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, *60*, 293-303. doi:10.1016/j.bushor.2017.01.004

Lee, J. G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research, 2*(2), 74-81. doi:10.1016/j.bdr.2015.01.003

Lowe, A., Norris, A. C., Farris, A. J., & Babbage, D. R. (2018). Quantifying thematic saturation in qualitative data analysis. *Field Methods*, *30*(3), 191-207. doi:10.1177/1525822X17749386

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education, 8*(3), 14. Retrieved from http://ojs.aishe.org/index.php/aishe-j/article/view/335/553

Mahdi, O. R., Nassar, I. A., & Almsafir, M. K. (2019). Knowledge management processes and sustainable competitive advantage: An empirical examination in private universities. *Journal of Business Research*, *94*, 320-334. doi:10.1016/j.jbusres.2018.02.013

Mai, J.-E. (2016). Big data privacy: The datafication of personal information. *The Information Society*, *32*(3), 192-199. doi:10.1080/01972243.2016.1153010

Malli, G., & Sackl-Sharif, S. (2015). Researching one's own field. Interaction Dynamics and Methodological Challenges in the Context of Higher Education Research, 16. doi:10.17169/fqs-16.1.2225

Mannay, D., & Morgan, M. (2015). Doing ethnography or applying a qualitative technique? Reflections from the 'waiting field.' *Qualitative Research*, *15*(2), 166-182. doi:10.1177/1468794113517391

Marjan, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: Architecture, opportunities, and open research challenges. IEEE Access, 5, 5247-5261. doi:10.1109/ACCESS.2017.2689040

Matthias, O., Fouweather, I., Gregory, I., & Vernon, A. (2017). Making sense of big data – can it transform operations management? *International Journal of Operations, & Production Management*, *37*(1), 37-55. doi:10.1108/IJOPM-02-2015-0084

McCusker, K., & Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion*, *30*(7), 537-542. doi:10.1177/0267659114559116

McIntosh, M. J., & Morse, J. M. (2015). Situating and constructing diversity in semi-structured interviews. *Global Qualitative Nursing Research, 2*, 1-12. doi: 10.1177/2333393615597674

McKim, C. A. (2017). The value of mixed methods research: A mixed methods study. *Journal of Mixed Methods Research, 11*(2), 202-222. doi:10.1177/1558689815607096

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. *IEEE Access*, *4*, 1821-1834. doi:10.1109/ACCESS.2016.2558446

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data, & Society*, *3*(1), 205395171665021. doi:10.1177/2053951716650211

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 503-512. doi:10.1109/EDUCON.2018.8363273

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and E-Business Management*. doi:10.1007/s10257-017-0362-y

Milosevic, I., Bass, A. E., & Combs, G. M. (2018). The paradox of knowledge creation in a high-reliability organization: A case study. *Journal of Management*, *44*(3), 1174-1201. doi:10.1177/0149206315599215

Mitchell, R. K., Mitchell, B. T., & Mitchell, J. R. (2009). Entrepreneurial scripts and entrepreneurial expertise: The information processing perspective. In A. L. Carsrud,

& M. Brännback (Eds.), *Understanding the Entrepreneurial Mind* (pp. 97-137). doi:10.1007/978-1-4419-0443-0_6

Miterev, M., Turner, J. R., & Mancini, M. (2017). The organization design perspective on the project-based organization: a structured review. *International Journal of Managing Projects in Business, 10*(3), 527-549. doi:10.1108/IJMPB-06-2016-0048

Morse, J. M. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research, 25*(9), 1212–1222. doi:10.1177/1049732315588501

Moylan, C. A., Derr, A. S., & Lindhorst, T. (2015). Increasingly mobile: How new technologies can enhance qualitative research. *Qualitative Social Work, 14*(1), 36-47. doi:10.1177/1473325013516988

Mullon, P. A., & Ngoepe, M. (2019). An integrated framework to elevate information governance to a national level in South Africa. *Records Management Journal, 29*(1/2), 103-116. doi:10.1108/RMJ-09-2018-0030

Nadjaran Toosi, A., Sinnott, R. O., & Buyya, R. (2018). Resource provisioning for data-intensive applications with deadline constraints on hybrid clouds using Aneka. *Future Generation Computer Systems, 79*, 765-775. doi:10.1016/j.future.2017.05.042

Nonaka, I., Byosiere, P., Borucki, C. C., & Konno, N. (1994). Organizational knowledge creation theory: a first comprehensive test. *International Business Review, 3*(4), 337-351. doi:10.1016/0969-5931(94)90027-2

Nonaka, I., Hirose, A., & Takeda, Y. (2016). 'Meso'-foundations of dynamic capabilities: Team-level synthesis and distributed leadership as the source of dynamic creativity. *Global Strategy Journal*, *6*, 168-182. doi:10.1002/gsj.1125

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1-13. doi:10.1177/1609406917733847

Obel, B., & Snow, C. C. (2014). Jay R. Galbraith Memorial Project. Journal of Organization Design, *3*(2). doi:10.7146/jod.17953

O'Connell, S., Mc Carthy, V. J. C., & Savage, E. (2018). Frameworks for self-management support for chronic disease: A cross-country comparative document analysis. *BMC Health Services Research*, *18*(583), 1-10. doi:10.1186/s12913-018-3387-0

Park, Y., Sawy, O., & Fiss, P. (2017). The role of business intelligence and communication technologies in organizational agility: A configurational approach. *Journal of the Association for Information Systems*, *18*(9), 648-686. doi:10.17705/1jais.00001

Patel, B., Roy, S., Bhattacharyya, D., & Kim, T.-H. (2017). Necessity of big data and analytics for good e-governance. *International Journal of Grid and Distributed Computing*, *10*(8), 11-20. doi:10.14257/ijgdc.2017.10.8.02

Paul, P. K., Aithal, P. S., & Bhuimali, A. (2018). Business informatics: with special reference to big data as an emerging area: a basic review. *International Journal on Recent Researches in Science, Engineering, & Technology (IJRRSET)*, *6*(4), 21-27. Retrieved from http://www.jrrset.com/2018/volume6issue4/paper4.pdf

Ponelis, S. (2015). Using interpretive qualitative case studies for exploratory research in doctoral studies: A case of information systems research in small and medium enterprises. *International Journal of Doctoral Studies*, *10*, 535-550. doi:10.28945/2339

Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2018). The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers*, *20*(2), 209-222. doi:10.1007/s10796-016-9720-4

Posavec, A. B., & Krajnovic, S. (2016). Challenges in adopting big data strategies and plans in organizations. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1229-1234). Opatija, Croatia: IEEE. doi:10.1109/MIPRO.2016.7522327

Premkumar, G., Ramamurthy, K., & Saunders, C. S. (2005). Information processing view of organizations: An exploratory examination of fit in the context of interorganizational relationships. *Journal of Management Information Systems*, *22*(1), 257-294. doi:10.1080/07421222.2003.11045841

Prior, M. (2017). Accomplishing "rapport" in qualitative research interviews: Empathic moments in interaction. Special Issue: The social life of methods, Guest Editors: Gabriele Kasper and Steven J. Ross. *Applied Linguistics Review*, *9*(4), pp. 487-511. doi:10.1515/applirev-2017-0029

Qiao, H. (2018). A brief introduction to institutional review boards in the United States. *Pediatric Investigation*, *2*(1), 46-51. doi:10.1002/ped4.12023

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 10*(1). doi:10.1038/s41467-019-10933-3

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., … Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality, & Quantity, 52*(4), 1893-1907. doi.org:10.1007/s11135-017-0574-8

Shao, Z. (2019). Interaction effect of strategic leadership behaviors and organizational culture on IS-Business strategic alignment and enterprise systems assimilation. *International Journal of Information Management, 44*, 96-108. doi:10.1016/j.ijinfomgt.2018.09.010

Shapiro, S. J., & Oystrick, V. (2018). Three steps towards sustainability: Spreadsheets as a data collection analysis system for non-profit organizations. *Canadian Journal of Program Evaluation*, *33*(2), 247-257. doi:10.3138/cjpe.31157

Siddiqa, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications, 71*, 151-166. doi:10.1016/j.jnca.2016.04.008

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*(Supplement C), 263-286. doi:10.1016/j.jbusres.2016.08.001

Song, I.-Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364-373. doi:10.1111/exsy.12130

Sousa, D. (2014). Validation in qualitative research: General aspects and specificities of the descriptive phenomenological method. *Qualitative Research in Psychology*, *11*(2), 211-227. doi:10.1080/14780887.2013.853855

Stergiou, C., Psannis, K. E., Kim, B.-G., & Gupta, B. (2018). Secure integration of IoT and cloud computing. *Future Generation Computer Systems*, *78,* 964-975. doi:10.1016/j.future.2016.11.031

Sutton, J., & Austin, Z. (2015). Qualitative research: Data collection, analysis, and management. *The Canadian Journal of Hospital Pharmacy*, *68*(3), 226-231. doi:10.4212/cjhp.v68i3.1456

Teichler, U. (2014). Opportunities and problems of comparative higher education research: The daily life of research. *Higher Education*, *67*, 393-408. doi:10.1007/s10734-013-9682-0

Thomann, E., & Maggetti, M. (2017). Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools. *Sociological Methods, & Research*, 004912411772970. doi:10.1177/0049124117729700

Torres, E. (2020). *Artist tricks Google Maps into recording traffic jam with 99 cellphones and a wagon*. ABC News. https://abcnews.go.com/International/artist-tricks-google-maps-recording-traffic-jam-99/story?id=68754956

Tushman, M. L., & Nadler, D. A. (1978). Information processing as an integrating concept in organizational design. *Academy of management review*, *3*(3), 613-624. doi:10.5465/amr.1978.4305791

Twining, P., Heller, R. S., Nussbaum, M., & Tsai, C.-C. (2017). Some guidance on conducting and reporting qualitative studies. *Computers, & Education*, *106*, A1-A9. doi:10.1016/j.compedu.2016.12.002

Umer, M., Kashif, M., Talib, R., Sarwar, B., & Hussain, W. (2017). Data provenance for cloud computing using watermark. *International Journal of Advanced Computer Science and Applications*, *8*(6). doi:10.14569/IJACSA.2017.080654

U.S. Department of Health & Human Services. (1979). The belmont report. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Varghese, B., & Buyya, R. (2018). Next generation cloud computing: New trends and research directions. *Future Generation Computer Systems*, *79*, 849-861. doi:10.1016/j.future.2017.09.020

Walsh, R. (2015). Wise ways of seeing: Wisdom and perspectives. *Integral Review*, *11*(2), 156-174. Retrieved from integralreview.org/issues/vol_11_no_2_walsh_wise_ways_of_seeing.pdf

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research, 70*, 356-365. doi:10.1016/j.jbusres.2016.08.009

Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of big data. *Information Sciences*, *367-368*, 747–765. doi:10.1016/j.ins.2016.07.007

Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research, 70*, 287-299. doi:10.1016/j.jbusres.2016.08.002

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3-13. doi:10.1016/j.techfore.2015.12.019

Watson, H., & McGivern, M. (2016). Business Intelligence Journal, Vol. 21 No. 1. *Business Intelligence Journal*, 21(1), 5. Available from https://tdwi.org/research/2016/03/business-intelligence-journal-vol-21-no-1.aspx

Watson, T. (2018). Ethnography and the management of organisations. In: Ciesielska M., Jemielniak D. (eds) Qualitative Methodologies in Organization Studies. Palgrave Macmillan, Cham. doi:10.1007/978-3-319-65217-7_6

Weick, K. (2015). Karl E. WEICK (1979), The social psychology of organizing, Second Edition: Paperback: 294 pages Publisher: McGraw-Hill (1979) Language: English ISBN: 978-0075548089. M@n@gement, vol. *18*(2), 189-193. doi:10.3917/mana.182.0189.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409-421. doi:10.4337/9781849807630.00024

Welch, C., & Piekkari, R. (2017). How should we (not) judge the 'quality' of qualitative research? A re-assessment of current evaluative criteria in International Business. *Journal of World Business*, *52*(5), 714-725. doi:10.1016/j.jwb.2017.05.007

Whittingham, K., Barnes, S., & Dawson, J. (2016). Capturing the carer's experience: a researcher's reflections. *Nurse Researcher*, *23*(5), 31-35. doi:0.7748/nr.23.5.31.s7

Wilkin, C. L., Couchman, P. K., Sohal, A., & Zutshi, A. (2016). Exploring differences between smaller and large organizations' corporate governance of information

technology. *International Journal of Accounting Information Systems*, *22*, 6-25. doi:10.1016/j.accinf.2016.07.002

Wu, S. P.-J., Straub, D. W., & Liang, T.-P. (2015). How information technology governance mechanisms and strategic alignment influence organizational performance: Insights from a matched survey of business and IT managers. *MIS Quarterly*, *39*(2), 497-518. doi:10.25300/MISQ/2015/39.2.10

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, *10*(1), 13-53. doi:10.1080/17538947.2016.1239771

Yin, R. K. (2017). *Case Study Research and Applications: Design and Methods*. SAGE Publications. Retrieved from https://books.google.com/books

Yip, C., Han, N. R., & Sng, B. L. (2016). Legal and ethical issues in research. *Indian Journal of Anaesthesia*, *60*(9), 684-688. doi:10.4103/0019-5049.190627

Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. *Issues in Information Systems*, *16*(2), 81. Retrieved from http://www.iacis.org/

Zelt, S., Recker, J., Schmiedel, T., & vom Brocke, J. (2018). A theory of contingent business process management. *Business Process Management Journal*. doi:10.1108/BPMJ-05-2018-0129

Zelt, S., Schmiedel, T., & vom Brocke, J. (2018). Understanding the nature of processes: An information-processing perspective. *Business Process Management Journal*, *24*(1), 67-88. doi:10.1108/BPMJ-05-2016-0102

Appendix A: Copyright Permission for Galbraith Images

Ken Knapton

From: Sasha Galbraith <[redacted]@jaygalbraith.com>
Sent: Wednesday, August 29, 2018 5:36 PM
To: Ken@KnaptonFamily.net
Subject: Re: Sasha Galbraith Contact: Permission to use copyrighted images

Hi Ken,

Thanks for writing — and for asking — about the two images you'd like to reproduce. You may have permission to reproduce the two you cited below for your study, provided you reference them properly.

Jay was fascinated by Big Data and would probably have written a book about the organization design implications for companies that embrace data analytics. If you think of it, I would love to see what you come up with in your study when you've finished it.

Kind regards,
Sasha

Sasha Galbraith, Ph.D.
Galbraith Management Consultants
303 S. Broadway, Ste. 200-403
Denver, Colorado 80209

Appendix B: Interview Protocol

Interview Protocol for Individual Interviews

1. Introduce myself to the participant and describe my role as a doctoral student and researcher
2. Ensure the Informed Consent Form has been signed
3. Remind participants of their voluntary participation in the study and their right to withdraw at any time
4. Briefly discuss the concept of big data and data governance with the interviewee and explain that we are here to discuss the governance mechanisms in place that assist in the analysis of big data with a focus on its effect on strategic decision making within the organization
5. Remind the participant that all answers should be based upon their current experience and current organization
6. Remind the participant that the interview will be recorded and let the participant know that the audio recording is about to begin
7. Begin the recording. State again that the interview is being recorded, and have them provide approval for the recording
8. Ask the defined questions in a semi-structured fashion, allowing for re-ordering of the questions as the interview progresses

Demographic Questions

9. What is your role in data gathering and analysis?
10. Please describe your background concerning big data analysis.

11. How long have you been involved in big data analysis?

Interview Questions

12. How do you ensure that you are analyzing all the available data for any particular model?
13. How do you determine the appropriate sources of data for any specific model or analysis?
14. What processes are in place to ensure that all applicable data are available to you when needed for analysis?
15. What countermeasures are in place for risk mitigation regarding the validity of the analysis that you perform (i.e., the risk of not having current data, or of incorrect correlation of data from various sources)?
16. What have you found to be most successful in building data models based on big data from various sources?
17. Why do these processes give you confidence that you are analyzing all applicable data?
18. How do these processes ensure the protection of the data?
19. What controls are in place to ensure the privacy of the data?
20. What other issues are there that we should discuss as relating to big data analysis?
21. What other areas are there left to discuss?
22. Thank them for their participation, and remind them of the member checking process that
23. End the recording

Appendix C: Participant Invitation Email

Subject: Invitation to participate in a Big Data research study

Dear [Recipient],

My name is Ken Knapton, and I have been given your name as someone who is knowledgeable about big data in your company. I am a doctoral student at Walden University, and I am conducting a doctoral study regarding the governance of big data. The purpose of this study is to identify big data governance strategies used by data scientists.

I would like to request your participation in the study. Participation is voluntary, and you may withdraw from participation at any time. Please see the attached consent form for a detailed description of the research study. Please reply and let me know if you are interested in participating in this study.

Thank you for your consideration,

Ken Knapton, MBA, MIT
DIT Student
College of Management, & Technology
Walden University

Appendix D: Letter of Cooperation from a Research Partner

[Community Research Partner Name]

[Contact Information]

[Date]

Dear Ken Knapton,

Based on my review of your research proposal, I give permission for you to conduct the study entitled "Exploring Mid-Market Strategies for Big Data Governance" within the [Insert Name of Community Partner].

As part of this study, I authorize you to:

Contact the data scientists that we identify as potential participants, either via email or by phone

Obtain their informed consent from each potential participant

Conduct an initial interview and a follow-up member checking interview with them of approximately 60 minutes each, and each using GoTo Meeting or a private location where the interviewee is comfortable

Collect documentation related to the big data governance process within our organization

Individuals' participation will be voluntary and at their own discretion.

We understand that our organization's responsibilities include:

Identify potential participants by using the following criteria, as outlined in your proposal:

Data Scientists (not necessarily using this formal job title) who are actively participating in the analysis of big data within our organization

Individuals that are employees or contractors working within the organization and who have intimate knowledge of the data management and analytical strategies of the case study organizations

Individuals that have been in a data scientist type of role for a minimum of 5 years

Individuals that have architected the data analytic solutions at these companies

Individuals with whom you do not have a recurring working relationship

Provide you with contact information for all the individuals that meet these criteria

Work with you to identify the initial set of individuals from this population to interview

Provide any documentation regarding the big data analysis process, policy and procedure within our organization

Help to ensure the privacy of each participant by not discussing this study in any meetings or other open forums

Provide authorization by emailing this form to both Ken.Knapton@WaldenU.edu and IRB@mail.waldenu.edu

We reserve the right to withdraw from the study at any time if our circumstances change.

I understand that you will not be naming our organization in the doctoral project report that is published in Proquest.

I confirm that I am authorized to approve research in this setting and that this plan complies with the organization's policies.

I understand that the data collected will remain entirely confidential and may not be provided to anyone outside of the student's supervising faculty/staff without permission from the Walden University IRB.

Sincerely,

[Authorization Official]

[Contact Information]

Walden University policy on electronic signatures: An electronic signature is just as valid as a written signature as long as both parties have agreed to conduct the transaction electronically. Electronic signatures are regulated by the Uniform Electronic Transactions Act. Electronic signatures are only valid when the signer is either (a) the sender of the email, or (b) copied on the email containing the signed document. Legally an "electronic signature" can be the person's typed name, their email address, or any other identifying marker. Walden University staff verify any electronic signatures that do not originate from a password-protected source (i.e., an email address officially on file with Walden).