Received 24-01-2023; Accepted 14-03-2023; Published 20-03-2023

Exploring Mid-Market Strategies for Big Data Governance

Ken Knapton Ken@KnaptonFamily.net
DIT, Walden University, Minneapolis
MN, United States

Corresponding Author: Ken Knapton

Abstract

Many data scientists struggle to adopt effective data governance practices as they transition from traditional data analysis to big data analytics. This qualitative multiple case study explored big data governance strategies used by data scientists employed in 3 mid-market companies in the greater Salt Lake City, Utah area who have strategies to govern big data. Data were collected via 10 semi-structured, in-depth, individual interviews and analysis of 4 organizational process documents. Four major themes emerged from the study: ensuring business centricity, striving for simplicity, establishing data source protocols, and designing for security. One key recommendation from the findings 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 be more profitable by gaining competitive advantages. By implementing strategies relating to the segregation of duties, encryption of data, and personal information, data scientists can mitigate contemporary concerns relating to using private information in big data analytics.

Keywords: Big data, Governance, Data analytics

1. INTRODUCTION

Data generation has increased exponentially in recent years, with no signs of this trend stopping. Bello-Orgaz et al. [1], estimated that globally, 2.5 exabytes of new data are being generated per day, and the Government Accountability Office [2] 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, organizations effectively use less than 5% of their available data [3]. 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. The traditional methods, frameworks, strategies, and tools for data governance and analysis are outdated and no longer adequate for processing the vast amount of data available to organizations today, thus making current strategies ineffective for handling big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. 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.

Citation: Ken Knapton Exploring Mid-Market Strategies for Big Data Governance. Advances in Artificial Intelligence and Machine Learning. 2023;3(1):51.

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. However, to remain competitive, it is just as critical for them to find ways to address this data deluge that faces companies of all sizes.

Contemporary outdated data processing systems cannot handle the exponentially increasing amounts of data that we are generating daily. 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 [4]. The problem addressed in this study is that some data scientists lack big data governance strategies that provide a holistic perspective of data for making effective decisions.

2. SUMMARY OF THE LITERATURE

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 following 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.

2.1 Organizational Information Processing Theory (OIPT)

The organizational information processing theory (OIPT) was the lens through which data governance strategies used by data scientists were examined to provide a holistic perspective of those data for making decisions. Jay R. Galbraith introduced the OIPT in 1974 [5]. 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 [6].

Obel and Snow [7], posited that Dr. Galbraith's theory of organizational design is even more relevant today than when Galbraith developed it in 1974 because of the dramatically increased availability of information in the form of big data. Park et al.[8], highlighted that the ability to process information for organizational decision-making had been studied extensively. Still, little focus has been placed on the details within the IT component of that decision-making process. Cao et al. [9], articulated that Galbraith's work was later adopted to address organizational decision-making. Jia et al.[10], 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 gathering, processing, and acting on information from within their environment. Hwang et al. [11], added that through the OIPT, Galbraith highlights the essential function of an organizational structure to facilitate the collection, analysis, and distribution of information to reduce uncertainty within the organization.

2.2 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 [12]. Some defined it by the processes required to leverage big data [13]. Wang et al. [14], categorized the various definitions of big data as oriented around the product, process, cognitive, or social aspects. As a case in point, Wamba et al. [15], indicated that big data is actionable and delivers sustained value, measured performance, and provides a competitive advantage, while Sivarajah et al. [16], defined big data by the 7 V's including high volume, high variety, low veracity, high velocity, high variability, effective visualization, and high value. Lee [17], 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 [18], also referred to the value of the data as a critical element to the definition of big data. Herschel and Miori [19], stated that big data refers to capturing, storing, analyzing, and acting upon information 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 those data attributes. 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 [20].

Challenges with big data. While there were varying definitions of what constitutes big data, the literature covered the challenges of gathering and analyzing big data in detail. De Mauro et al. [21], identified that the combination of structured, unstructured, and semi-structured data is challenging to handle with traditional systems. Yang, Huang, Li, Liu, and Hu [22], asserted that due to big data's heterogeneous and unstructured nature, it is challenging for traditional systems to manage and analyze those data effectively. Grover et al. [23], estimated that unstructured data make up 95% of big data. Siddiqa et al. [24], 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 [17], also noted that managing data at velocity strains contemporary data center architectures that are not sufficient to sustain the volume and velocity of heterogeneous data elements of personal and corporate data. Due to the velocity attribute associated with big data, Lee and Kang [25], noted that contemporary analyzing methods such as storing data elements in a data warehouse for later analysis are no longer sufficient. Lee [17], also asserted 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 of attributes of big data make it very difficult to determine the veracity of those data [26].

Complexity and velocity. Identifying and categorizing big data at both velocity and volume poses problems for contemporary systems. Kemp [18], 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. [22], noted that metadata can provide significant benefits for normalizing unstructured data elements for traditional structured data systems, but this brings about another challenge regarding automating the generation of metadata for such data elements. The high velocity poses an additional problem of lack of storage capabilities for the transient data flows through the analyzing systems. Yang et al. [22], 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 adds strain to existing analytic systems. Jin et al. [20], articulated that data scientists currently lack an understanding of the relationship between data complexity and computational complexity related to big data processing. Yang et al. [22], further noted that traditional database management systems lack the scalability for managing and storing the unstructured elements of big data. Jin et al. [20], asserted 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. [22], noted that high-dimensional data cannot be processed efficiently within the time constraints required to process big data effectively. Jin et al. [20], also discussed that data complexity is one of the significant challenges with big data as compared to traditional data. Yang et al. [22], 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.

Volume and security. The resource requirements associated with big data analysis are stretching traditional systems to their computing limits. Yang et al. [22], articulated that the processing needs for big data currently exceed the capabilities of traditional systems, and Lee and Kang [25], concluded that the limitation of processing power poses a significant challenge for analyzing big data. Intezari and Gressel [27], added that new techniques and advanced tools are required to extract information from big data. Jin et al. [20], 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 [25], added that this problem extends to the software layer as current 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. [22], asserted 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 [17], noted that big data can include many elements of personal information, which Sivarajah et al. [16], 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. [28], noted that individuals disclose very personal information on social networks often unconsciously, and Flyverbom et al. [29], added that this information now allows corporations to encroach on the private lives of individuals at an unprecedented scale. Metcalf and Crawford [30], added that there are so many publicly available data sets about individuals that they become highly identifiable when correlated together.

Ethical issues. Passive data generation leads to significant ethical issues for organizations analyzing big data. Herschel and Miori [19], posited that big data changes the nature of ethical debates fundamentally by redefining the power of information and the extent to which free will can guide users' actions. Mai [31], added that the primary ethical issue is not collecting the information but how and when it is ethically responsible for analyzing that information. Carbonell [32], 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.

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 [33], explained that organizations must be willing to address the managerial and organizational changes required to leverage big data initiatives effectively. Kemp [18], 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 [17], asserted 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 [18], 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 [18], added that a top-down approach to big data governance 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. Carbonell [32], 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. Lee [17], 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. [20], added that contemporary data analysts lack a solid understanding of dealing with the complexity found with complex big data structures. In a study of 430 firms with advanced analytics capabilities, Lee [17], identified that approximately 66% indicated that they cannot employ enough qualified analysts to extract the value from the data they have gathered. Lee [17], added that the McKinsey Global Institute estimated that an additional 140,000 – 190,000 big data analysts would 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 and 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. [34], argued that escalating information availability and visibility improves decision making, and Mahdi et al [35], asserted that knowledge dissemination increases teamwork. Kemp [18], 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. [22], noted out that due to the heterogeneous attribute of big data, the visualization of those data in a human-understandable manner poses a significant challenge.

2.3 Opportunities for Decision-making

One of the more important themes found in the literature involved big data and its effect on 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 [17], asserted that in 2016, 5.5 million new devices were connected each day to the Internet to share, process, and analyze information. That number was expected to grow to 20.8 billion by 2020. Ahmed et al. [36], stated that the number of human beings globally has now been surpassed by the number of connected devices. The overall number of connected devices was expected to reach 50 billion by 2020. Herschel and Miori [19], added that the number of network connections was expected to reach 18.9 billion, approximately 2.5 connections for every person on earth. The Organization 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 et al. [37], 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 merging the physical and virtual worlds. Because of the amount of data generated by IoT devices, Paul et al. [38], explained that some data scientists claimed that big data provides a direct link between cyber and physical systems. Dourish and Cruz [39], utilized the term datafication to describe the transformation of physical, social action into quantifiable digital data, enabling real-time predictive analytics. Flyverbom et al. [29], 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 et al. [40], 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 [19], added that the analysis of this minutia leads organizations to generate new insights into individual consumers' behaviors and activities, often without the consumer's consent.

Analytics of vast data sets can have significant predictive power. The Government Accountability Office (2017) identified that new big data analytics technologies would extract hidden patterns and correlations from these extensive data sets. Ahmed et al. [36], stated that combining IoT and big data analytics promises to significantly improve decision-making within businesses. Carbonell [32], 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 [41], added that the advantage of big data over traditional data analysis is better accuracy, which leads to higher confidence in data-driven decisions.

Impact on decision-making. Harnessing predictive analytics with big data provides significant competitive advantages. Intezari and Gressel [27], 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 et al. [42], predicted that the economic value of organizations could be elevated by effectively utilizing the deeper insights available from big data. Cao et al. [9], found that by improving information processing capabilities, the decision-making effectiveness within the firm increased and created a competitive advantage, especially when those capabilities were rare, valuable, and inimitable. Grover et al. [23], 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. [43], added that the distribution of big data analytics within organizations had a direct effect on improved business decisions within the organization.

Big data analytics. Big data analytics and effective decision-making are closely correlated. Braganza et al. [42], articulated that practical big data analysis enabled more evidence-based decisions, while Wamba et al. [15], added that big data-enabled data-driven decision making. Paul et al. [38], further clarified that analyzing those data had become a critical need for every business sector. Janssen et al.[44], stated that effectively analyzing big data had the potential to change how organizations make decisions fundamentally. Intezari and Gressel [27], 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. [3], observed that companies that oriented their decision-making around fact-based data analytics outperformed their counterparts in the market. According to Jin et al. [20], big data has had a significant impact on almost every business sector and industry already. Lee [16], stated that all companies would need to build big data expertise to remain competitive. Günther et al. [45], noted that there is quite a bit of evidence currently that big data can significantly impact an organization's business models. For organizations to extract value from big data, Günther et al. [45], stated that it is critical to continually realign work practices, organizational designs, and stakeholder interests. Organizations are still struggling to understand how to utilize big data analytics to guide real-time decisions effectively.

Organizational and employment impact of big data. It is not yet clear how to effectively organize to best utilize big data within the organization. Günther et al. [45], discussed that scholars were actively debating the most appropriate organizational model to take advantage of big data analytics most effectively. According to Janssen et al. [44], 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. [42], added that this lack of role clarity across the organization inhibited the ability to use required big data resources efficiently.

The cross-departmental collaboration required for big data to succeed 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. [43], identified that power shifts had been documented within organizations that effectively utilize big data analytics to make fact-based and real-time decisions. Braganza et al. [42], 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 requires new governance processes.

IT Governance. Without proper governance practices, technological solutions to big data problems will not solve the current challenges. Watson and McGivern [46], noted that business intelligence initiatives had a low chance of success without governance practices providing high-quality, secure, and well-modeled data. Braganza et al. [42], added that extracting value from big data requires the appropriate technological solutions and requires that big data analysis becomes part of the fabric of an organization. According to Alreemy et al. [47], IT Governance played a critical role in creating this fabric by establishing a link between the business and IT that can provide a competitive advantage by minimizing expenditures and making better use of time.

Effective IT governance is critical for IT organizations to meet stakeholder expectations in all areas of information systems effectively. Wu et al. [48], stated that the single most critical predictor of value from an IT organization was an effective IT governance process. Braganza et al. [42], suggested that effective business processes could release value from big data initiatives that have, as of yet, been challenging to reproduce consistently. Mahdi et al. [35], added that the processes within an organization that create, share, and use knowledge were highly critical strategic capabilities that create a competitive advantage.

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. [49], noted that larger organizations were more likely to implement formal IT governance practices than smaller organizations. Wilkin et al. [49], added that smaller organizations faced significant challenges in managing technology assets to sustain a competitive advantage. Begg and Caira [50], also added that smaller companies tended to believe 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. [42], 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. [13], added that research had focused on big data technologies for data analysis, storage, and visualization. Still, there has been a research gap regarding how organizations would need to adjust to embrace these technological innovations. Braganza et al. [42], stated that senior leadership processes required that they could implement to extract strategic value from big data. When addressing the challenges associated with big data analysis, Begg and Caira [50], 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. [42], 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 manage their big data initiatives effectively. To be able to extract value from raw data, Song and Zhu [51], noted that organizations require a well-defined, systematic approach to governing their data. Mikalef et al. [13], clarified that Data Governance was the term used to describe these activities, which included roles, structures, and decision mechanisms around IT resources. Posavec and Krajnovic [52], recognized that the lack of formal data governance processes was one of the primary reasons organizations failed to utilize their data effectively.

The Governance of big data is critical to the success of any big data initiative. Cervone [53], 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 et al. [54], identified that the Governance of big data refers to the *how-to* aspects of harnessing the data within the organization. Cheng et al. [55], added that data governance refers to the series of policies that an organization used to define cloud data strategy, management, operations, and optimization. Chrisman et al. [56], further clarified that Governance referred to both formal and informal rules, practices, and processes implemented to direct and control behavior consistently throughout the organization.

Effective data governance can lead to improved business outcomes. When organizations implement governance well, Wang and Hajli [33], identified that big data initiatives could provide insights into hidden data correlations for meaningful decision making. Wang et al. [54], added that proper data governance addressed the standardization, accuracy, and availability of those data. When the organization implemented governance poorly, Wang and Hajli [33], noted that organizations experienced substantial financial costs for little value. By effectively applying governance principles to data governance initiatives, Wang et al. [54], found that organizations could better meet organizational goals with lower costs and shorter timelines.

Effective big data governance practices are not easy to implement but can significantly impact the overall business. Hashem et al. [57], noted that adopting big data governance practices that balance value creation and risk mitigation was imperative to organizations. Mahdi et al. [35], found that knowledge management practices within an organization were significantly and positively related to sustainable competitive advantage.

Effective data governance within the IT organization can help elevate it to more strategic alignment with the rest of the business. As more organizations rely on data as a critical differentiating asset, Watson and McGivern [46], identified that Governance around the data becomes more essential for the company's success, which in turn shifts IT and business intelligence into a more strategic alignment with business initiatives. Shao [58], added that alignment between overall information systems strategy and business strategy is a critical precursor to overall profitability and competitive advantage.

Small companies are slower to recognize the strategic value of data compared to their larger competitors. Begg and Caira [50], 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. [49], added that resource constraints in smaller organizations tended to limit the organization's ability to be competitive in IT competencies, knowledge management, and innovation. Begg and Caira [50], 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.

3. METHODS

Based on the background and literature review, 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 data for making decisions. The target population for this study was data scientists employed in mid-market companies in the greater Salt Lake City, Utah area. Specifically, the sample for this study was 10 data scientists employed in three target companies from the study area who have strategies to govern big data.

4. RESEARCH QUESTION

RQ1: What big data governance strategies do data scientists employ to provide a holistic perspective of data for making decisions?

5. DATA COLLECTION

As is common in a case study design, access to participants and documentation was accomplished through the use of a gatekeeper within each organization. Open-ended interview questions and organizational documents were used as the data collection methods.

Interviews. Interview questions included the following:

1. How do you ensure that you analyze all available data for any particular model?
2. How do you determine appropriate data sources 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 you perform (i.e., the risk of not having current data or 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?

Document review. We used operational document analysis in addition to interviews in this study. By working with the gatekeeper, we gained access to the process and standards documentation that is used by the organization to govern their big data procedures. We then evaluated the documentation for applicability to this study.

6. DATA ANALYSIS

A thematic analysis was used to provide an analysis of 10 transcripts and 4 organizational documents to explore the big data governance strategies employed by data scientists to provide a holistic analysis of data for making decisions. The interviews were manually transcribed and formatted into a text file. The data were loaded into QDA Miner, a qualitative and mixed methods data analysis tool. The responses that emerged from participants during the interview process were coded and then analyzed. Themes were developed by way of the analysis and coding process performed.

7. RESULTS

Participants were interviewed to answer the overarching research question: What big data governance strategies do data scientists employ to provide a holistic perspective of those data for making decisions? The final sample consisted of ten participants from three different organizations with an average of 9.4 years of experience working in big data analytics.

8. MAJOR THEMES

The following themes were developed from the data collected and will be discussed in more detail below:

Theme 1: Ensuring Business Centricity

Theme 2: Striving for Simplicity

Theme 3: Establishing Data Source Protocols

Theme 4: Designing for Security

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.

Understanding business needs. Having a clear understanding of business needs was a universal theme in every participant's discussion. Before data scientists can provide answers to business questions, they must understand the needs they are addressing. Understanding business needs is an essential theme from analytic and data governance perspectives. Understanding business needs helps data scientists identify appropriate data sources and elements that will be needed to address the business question. Each participant discussed the importance of understanding the question they were being asked to answer before assessing data sources and elements for analysis. There were various 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. 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. All 10 participants discussed the importance of clearly defining the business need as a critical step to deciding which data sources they needed to connect with to get their data. Data scientists cannot bridge this gap without clearly understanding the business needs to which they are responding when creating their analysis or models from big data.

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 understanding business needs. Data scientists can learn more about business needs by creating partnerships between various technical and business roles. 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. P9 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 help 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. 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. 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.

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 critical strategy for managing information from big data. For the analysis to be useful to business decision-makers, data scientists must 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 discussed maintaining contextual awareness while gathering, preparing, and analyzing big data. 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. P9 explained 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 decide if the report meets with scrutiny as the executives attempt to correlate information from various reports. Context is important to help those executives accurately interpret the analysis results. 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 vital context to the data. P8 mentioned that it is crucial to maintain contextual awareness when automating results so that the context is not lost because of the automated manipulation. This Governance includes maintaining contextual awareness of both the data elements and the data sources.

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. P1 discussed 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 the data for any analysis because there is simply too much of it. These participants agree that analyzing all 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. One of the key responsibilities of data scientists is to reduce the noise and simplify the data analysis process.

Theme 2: Striving for Simplicity

Because of big data’s volume, variety, and velocity attributes, analytics can quickly become complex. Complexity is one of the significant challenges with big data compared to traditional data [20]. 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. OIPT is based on the concept that organizations should be designed to reduce that uncertainty and enable decision-making [59]. Striving for simplicity is a critical skill that will help data scientists deal with big data’s inherent complexities 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.

Minimizing the data sources. Strategically limiting the data sources was a common concept 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 few data sources. This idea 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 noise source 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. P2 identified a handful of core data sources that can provide answers to most of the business questions asked at their organization. P6 articulated that they strategically limit their data sources to only those valuable for the specific business question being analyzed. All participants mentioned that there could be more data available to them from other sources but that minimizing the data sources is one of the key strategies that they follow in dealing with the ever-increasing complexity of various big data sources. 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 and confidence in the analytics. 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.

Using new tools. Many new tools have been developed in recent years to simplify big data analysis. These new and advanced tools are required to extract information from big data. These new tools are also necessary from a governance perspective as they help data scientists better interact with data in ways that traditional tools do not. Eight of the 10 participants discussed using 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 the various tools available. All eight participants who discussed tools indicated a desire or plan to start using more of these specialized tools soon to further refine 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, 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 govern big data effectively. As larger organizations adopt and utilize these new tools, small and medium-sized businesses can benefit from using these same tools to solve their big data challenges. 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.

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 big data analytics is not being done correctly today, mainly because some data scientists design overly complex solutions. P6 discussed the importance of keeping solutions simple yet not shying away from complexity when it is required. Finding this balance is a key skill set for any data scientist to govern big data for analytic solutions effectively. The complexity of big data sources strains traditional analytic systems because of the nature of structured and unstructured heterogeneous data elements. Data scientists who can simplify these constructs form an advantage over those who are still struggling to understand the relationship between data complexity and computational complexity as relates to big data processing.

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 vital to perform preprocessing to improve data quality. With the rapid decay of the value of big data, the need for automation becomes clear. Eight of the 10 participants mentioned 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. 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. 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. The value of some big data elements decays much more rapidly than traditional data. The processing of big data and validating 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. When considering the temporary value of raw big data and the need to reduce uncertainty within an organization by processing more information, 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 big data's volume, velocity, and variety attributes. In the absence of standards and well-defined processes for data validation, data scientists cannot have confidence in their analysis. Incomplete data can result in partial analysis, which can paint an inaccurate overall picture [44]. This theme consists of three subthemes: following defined standards, establishing validation processes, and reducing duplication.

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 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. 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 peer review of solutions, daily review of exception reports and alerts, and time commitments of deliverables between departments. Process documentation from three 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 data protection within big data solutions. 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. 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.

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 [26]. This subtheme also aligns well with the subthemes of minimizing data noise and minimizing data sources. This alignment is because many big data sources are often insecure, leading to potentially inaccurate analysis [60]. Developing processes for validating data sources is a key governance practice for data scientists. All 10 participants discussed having processes in place to validate the various data sources in their analysis. P1 and P5 discussed applying a scoring mechanism to each data source. A 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 data validation 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. For example, P1 described an automated validation process as 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 to validate their big data analytics. Establishing a validation process for data sources and specific data elements is critical in big data governance.

Organizations tend to have a false sense of confidence because of the hype around big data [19]. Validating the origin and transformation of data elements contained in big data is one of the primary challenges of big data [61]. The validity of big data analytics can be questioned due to the multiple sources of big data, which are less verifiable than was the case with traditional data [57]. 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. If the big data analysis consumers are not confident in the results, they will not be reducing uncertainty but creating more uncertainty.

Reducing duplication. Another strategy that arose from the study was 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 risks 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 data duplication. P10 articulated a similar concept when they explained that they ensure that they do not duplicate the data because the same information should not appear in different tables in the form of different dimensions and different facts.

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 scientist 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. By refraining from copying the data and analyzing the data where it resides, data scientists can avoid these traditional challenges. Minimizing the copying and duplication of data is a crucial governance strategy for data scientists as they transition from traditional analytic methods to big data analytics.

There is a limit to the amount of data that an organization can gather and store. Still, organizations increase their ability to analyze more data by analyzing the data where it resides. By designing systems that analyze data without requiring that those data be copied onto local storage devices, organizations can process more significant 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 data duplication can increase the validity of big data analytics by data scientists.

Theme 4: Designing for Security

Security and privacy protection 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 represented in those data. With the ever-increasing threat of a data breach, any organization must protect the information within its control. This theme consists of three subthemes: segregation of duties, using encryption, and protecting private information.

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 data protection, 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 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. If data scientists needed additional information to perform their analysis, they would need to obtain permission from the data owner or data steward before they could access the information. 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 private data based on role. Specifically, P8 highlighted a role distinction between the data architect and report writing roles. 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 access only the information they require to accomplish their job. Maintaining appropriate segregation of duties between those responsible for data access and those who analyze information is critical to protecting the data. P6 spoke of the inherent conflict between data owners who want to restrict access to protect data and data scientists who wish to access and review as much data as possible. There is a healthy conflict that forces data scientists to justify all access to data in terms of a business question they are working to answer. This highlights the alignment between this subtheme and understanding the business need. 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 responsible for gathering the data, those responsible for analyzing the data, and those responsible for protecting the data creates natural boundaries around data governance within the organization that lead toward better security overall.

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 be implemented to protect data against unauthorized access. Three of the 10 participants specifically mentioned encryption when asked about data protection. The use of encryption is a common practice to protect personal information for data at rest and data in transit. The remaining seven participants stated that their security team has controls in place to protect 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 data encryption as a key security control. One of those documents indicated that it is important to comply with security requirements, laws, and regulations as we design our systems. Because of these challenges, 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 must also improve to handle the speed requirements that businesses demand from big data analytics.

Protecting private information. Protecting the privacy of individuals is critical to organizations that are analyzing big data. Individuals reveal very personal information on social networks and IoT devices, often unconsciously and without their knowledge or consent. When implementing data governance over big data, data scientists must be aware of personally identifiable information (PII) within their system and proactively put controls in place to protect those data. When asked about privacy, 10 out of 10 participants indicated that they are mindful of privacy, and it is crucial to their organization. P4 explained that they don't use customer identifying information for any analysis, but "there's one process that uses that, and we hash to protect the data". Additionally, one of the process documents describes that the system in use in that organization will automatically determine whether the requested information is PII and 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 before accessing their information. Regardless of the specific way personal data is protected, organizations must understand what personal information they have access to and put processes and controls in place to protect that information.

9. SUMMARY

As data scientists strive to adjust to the unique demands that arise when transitioning from traditional data analysis to analyzing big data, they necessarily must leverage new tools and improved processes. The unique challenges resulting from the big data characteristics of volume, velocity and veracity require a new approach for data analysis. By implementing the four themes outlined in this study data scientists can improve their ability to make that transition. As they ensure business centricity, strive for simplicity, establish data source protocols, and design for security, data scientists can implement big data analytics processes that will allow their business to make data driven decisions from big data that could uncover new sources of profitability and efficiency within their organizations. Implementing these study findings will also help them to balance the need for deeper data analytics while maintaining the privacy of the individuals represented in those data.

10. DISCUSSION

10.1 Implications for 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 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. Implementing the practices associated with these four themes will allow individuals within the organization to work more autonomously and close the gap between information needed and information available for the front-line employees.

Data scientists who use the strategies described in these findings could improve their effectiveness as change agents for their organization. Adopting big data governance practices that balance value creation and risk mitigation is imperative to organizations. Learning and implementing these practices could make data scientists more valuable to their organization, resulting 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, data scientists 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 and the efficiency of the solutions that the data scientist can provide to the organization.

10.2 Future Research

This study focused on data scientists' general data governance practices who utilize big data to help make decisions within their organizations. 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. We recommend that future studies focus on combining these three technologies. IoT is creating additional data sources for big data, and machine learning provides 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.

When implementing big data solutions, security, privacy, and ethical use of big data are also critical aspects of data governance. Re-identification of previously obfuscated and anonymous data has been documented in multiple cases. Additional study is warranted in these areas to determine the effectiveness of the security controls and the extent to which the strategies outlined here effectively reduce 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.

10.3 Recommendations for Action

Policies, principles, and frameworks around how big data are used and leveraged within an organization pose enormous challenges for those organizations. The results of this study validate this concept from the literature. Data governance is critical for big data, and data scientists cannot add it as an afterthought. By implementing the findings of this study, data scientists can thoughtfully and proactively implement data governance into their big data solutions from the start. To extract value from raw data, organizations require a well-defined, systematic approach to governing their data. 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.

The lack of formal data governance processes was one of the primary reasons organizations failed to utilize their data effectively. Data scientists should review the findings of this study to determine how best to apply these practices to their big data solution architectures. Implementation of these themes can guide data scientists trying to move from traditional data analytics to big data analytics. To extract value from raw data, 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 [50], 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 look for ways to implement these concepts as they implement big data analytics within their organizations.

11. CONCLUSION

The focus of this study was to explore big data governance strategies that data scientists are currently using in practice. We 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. We explained the possibility of social change in how individuals' private information is gathered as part of a big data strategy. Following the strategy outlined in the four themes of this study for big data, governance can lead to the improved overall protection of individual privacy.

References

[1] Bello-Orgaz G, Jung JJ, Camacho D. Social Big Data: Recent Achievements and New Challenges. Inf Fusion. 2016;28:45-59.

[2] https://www.gao.gov/products/GAO-17-75

[3] Zakir J, Seymour T, Berg K. Big Data Analytics. Issues Inf Syst. 2015;16:81.

[4] Mikalef P, Giannakos MN, Pappas IO, Krogstie J. The Human Side of Big Data: Understanding the Skills of the Data Scientist in Education and Industry. IEEE Global Engineering Education Conference (EDUCON). 2018; 2018: 503-512.

[5] Galbraith JR. Organization Design: An Information Processing View. Interfaces. 1974;4:28-36.

[6] Premkumar G, Ramamurthy K, Saunders CS. Information Processing View of Organizations: An Exploratory Examination of Fit in the Context of Interorganizational Relationships. J Manag Inf Syst. 2005;22:257-294

[7] Obel B, Snow CC. Jay R. Galbraith Memorial Project. J Organ Des. 2014;3.

[8] Park Y, Sawy O, Fiss P. The Role of Business Intelligence and Communication Technologies in Organizational Agility: A Configurational Approach. J Assoc Inf Syst. 2017;18:648-686.

[9] Cao G, Duan Y, Cadden T. The Link Between Information Processing Capability and Competitive Advantage Mediated Through Decision-Making Effectiveness. Int J Inf Manag. 2019;44:121-131.

[10] Jia F, Blome C, Sun H, Yang Y, Zhi B. Towards an Integrated Conceptual Framework of Supply Chain Finance: An Information Processing Perspective. Int J Prod Econ. 2020;219:18-30.

[11] Hwang S, Kim H, Hur D, Schoenherr T. Interorganizational Information Processing and the Contingency Effects of Buyer-Incurred Uncertainty in a Supplier's Component Development Project. Int J Prod Econ. 2019;210:169-183.

[12] Gandomi A, Haider M. Beyond the Hype: Big Data Concepts, Methods, and Analytics. Int J Inf Manag. 2015;35:137-144.

[13] Mikalef P, Pappas IO, Krogstie J, Giannakos M. Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda. Inf Syst e-Business Manag. 2018;16:547-578.

[14] Wang H, Xu Z, Fujita H, Liu S. Towards Felicitous Decision Making: An Overview on Challenges and Trends of Big Data. Inf Sci. 2016;367-368:747-765.

[15] Wamba SF, Gunasekaran A, Akter S, Ren SJ, Dubey R, et al. Big Data Analytics and Firm Performance: Effects of Dynamic Capabilities. J Bus Res. 2017;70:356-365.

[16] Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical Analysis of Big Data Challenges and Analytical Methods. J Bus Res. 2017;70;Suppl C:263-86.

[17] Lee I. Big Data: Dimensions, Evolution, Impacts, and Challenges. Bus Horiz. 2017;60:293-303.

[18] Kemp R. Legal Aspects of Managing Big Data. Comput Law Sec Rev. 2014;30:482-491.

[19] Herschel R, Miori VM. Ethics, Big Data. Technol Soc. 2017;49:31-36.

[20] Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. Big Data Res. 2015;2:59-64.

[21] De Mauro A, Greco M, Grimaldi M. What Is Big Data? A Consensual Definition and a Review of Key Research Topics. AIP Conference Proceedings. 2015;1644: 97-104.

[22] Yang C, Huang Q, Li Z, Liu K, Hu F. Big Data and Cloud Computing: Innovation Opportunities and Challenges. Int J Digit Earth. 2017;10:13-53.

[23] Grover V, Chiang RHL, Liang TP, Zhang D. Creating Strategic Business Value From Big Data Analytics: A Research Framework. J Manag Inf Syst. 2018;35:388-423.

[24] Siddiqa A, Hashem IAT, Yaqoob I, Marjani M, Shamshirband S, et al. A Survey of Big Data Management: Taxonomy and State-Of-The-Art. J Netw Comput Appl. 2016;71:151-66.

[25] Lee JG, Kang M. Geospatial Big Data: Challenges and Opportunities. Big Data Res. 2015;2:74-81.

[26] Matthias O, Fouweather I, Gregory I, Vernon A. Making Sense of Big Data – Can It Transform Operations Management? Int J Oper Prod Manag. 2017;37:37-55.

[27] Intezari A, Gressel S. Information and Reformation in KM Systems: Big Data and Strategic Decision-Making. J Knowl Manag. 2017;21:71-91.

[28] Gao W, Liu Z, Guo Q, Li X. The Dark Side of Ubiquitous Connectivity in Smartphone-Based SNS: An Integrated Model From Information Perspective. Comput Hum Behav. 2018;84:185-193.

[29] Flyverbom M, Deibert R, Matten D. The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business. Bus Soc. 2019;58:3-19.

[30] Metcalf J, Crawford K. Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide. Big Data Soc. 2016;3.

[31] Mai JE. Big Data Privacy: The Datafication of Personal Information. Inf Soc. 2016;32:192-199.

[32] Carbonell IM. The Ethics of Big Data in Big Agriculture. Internet Policy Rev. 2016;5:1-13.

[33] Wang Y, Hajli N. Exploring the Path to Big Data Analytics Success in Healthcare. J Bus Res. 2017;70:287-299.

[34] Gupta S, Kumar S, Kamboj S, Bhushan B, Luo Z. Impact of Is Agility and HR Systems on Job Satisfaction: An Organizational Information Processing Theory Perspective. J Knowl Manag. 2019;23:1782-1805.

[35] Mahdi OR, Nassar IA, Almsafir MK. Knowledge Management Processes and Sustainable Competitive Advantage: An Empirical Examination in Private Universities. J Bus Res. 2019;94:320-34.

[36] Ahmed E, Yaqoob I, Hashem IAT, Khan I, Ahmed AIA, et al. The Role of Big Data Analytics in Internet of Things. Comput Netw. 2017;129:459-471.

[37] Bilal K, Khalid O, Erbad A, Khan SU. Potentials, Trends, and Prospects in Edge Technologies: Fog, Cloudlet, Mobile Edge, and Micro Data Centers. Comput Netw. 2018;130:94-120.

[38] Paul PK, Aithal PS, Bhuimali A. 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). 2018;6:21-27.

[39] Dourish P, Gómez Cruz E. Datafication and Data Fiction: Narrating Data and Narrating With Data. Big Data Soc. 2018;5.

[40] Galliers RD, Newell S, Shanks G, Topi H. Datification and Its Human, Organizational and Societal Effects: The Strategic Opportunities and Challenges of Algorithmic Decision-Making. J Strateg Inf Syst. 2017;26:185-190.

[41] Chauhan SK, Sangwan S. Big Data Analytics. International Journal on Recent and Innovation Trends in Computing and Communication. 2017;5:4.

[42] Braganza A, Brooks L, Nepelski D, Ali M, Moro R. Resource Management in Big Data Initiatives: Processes and Dynamic Capabilities. J Bus Res. 2017;70:328-337.

[43] Popovič A, Hackney R, Tassabehji R, Castelli M. The Impact of Big Data Analytics on Firms' High Value Business Performance. Inf Syst Front. 2018;20:209-222.

[44] Janssen M, van der Voort H, Wahyudi A. Factors Influencing Big Data Decision Making Quality. J Bus Res. 2017;70: 338-45.

[45] Günther WA, Rezazade Mehrizi MH, Huysman M, Feldberg F. Debating Big Data: A Literature Review on Realizing Value From Big Data. J Strateg Inf Syst. 2017;26:191-209.

[46] Watson HJ, McGivern M. Getting Started With Business-Driven Data Governance. Bus. Intell. J.. 2016;21(1):4-7.

[47] Alreemy Z, Chang V, Walters R, Wills G. Critical Success Factors (CSFs) For Information Technology Governance (ITG). Int J Inf Manag. 2016;36:907-916.

[48] Wu SP-J, Straub DW, Liang T-P. How Information Technology Governance Mechanisms and Strategic Alignment Influence Organizational Performance: Insights From a Matched Survey of Business and IT Managers. MIS Q. 2015;39:497-518.

[49] Wilkin CL, Couchman PK, Sohal A, Zutshi A. Exploring Differences Between Smaller and Large Organizations' Corporate Governance of Information Technology. Int J Acc Inf Syst. 2016;22:6-25.

[50] Begg C, Caira T. Exploring the SME Quandary: Data Governance in Practice in the Small to Medium-Sized Enterprise Sector. Electron J Inf Syst Eval. 2012;15:11.

[51] Song IY, Zhu Y. Big Data and Data Science: What Should We Teach? Expert Syst. 2016;33:364-373.

[52] Posavec AB, Krajnovic S. Challenges in Adopting Big Data Strategies and Plans in Organizations. In: 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Vol. 2016. Opatija, Croatia: IEEE Publications; 2016: 1229-1234.

[53] Cervone HF. Organizational Considerations Initiating a Big Data and Analytics Implementation. Digit Libr Perspect. 2016;32:137-1341.

[54] Wang Y, Kung L, Byrd TA. Big Data Analytics: Understanding Its Capabilities and Potential Benefits for Healthcare Organizations. Technol Forecasting Soc Change. 2018;126:3-13.

[55] Cheng G, Li Y, Gao Z, Liu X. Cloud Data Governance Maturity Model. 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). 2017;4.

[56] Chrisman JJ, Chua JH, Le Breton-Miller I, Miller D, Steier LP. Governance Mechanisms and Family Firms. Entrep Theor Pract. 2018;42:171-186.

[57] Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, et al. The Rise of ”Big Data” on Cloud Computing: Review and Open Research Issues. Inf Syst. 2015;47:98-115.

[58] Shao Z. Interaction Effect of Strategic Leadership Behaviors and Organizational Culture on Is-Business Strategic Alignment and Enterprise Systems Assimilation. Int J Inf Manag. 2019;44:96-108.

[59] Feurer S, Schuhmacher MC, Kuester S. How Pricing Teams Develop Effective Pricing Strategies for New Products: Pricing Teams and New Product Pricing Strategies. J Prod Innov Manag. 2019;36:66-86.

[60] Duncan B, Whittington M, Chang V. Enterprise Security and Privacy: Why Adding IoT and Big Data Makes It So Much More Difficult. In: International Conference on Engineering and Technology (ICE T). Vol. 2017. Antalya: IEEE Publications; 2017:1-7.

[61] Umer M, Kashif M, Talib R, Sarwar B, Hussain W. Data Provenance for Cloud Computing Using Watermark. Int J Adv Comput Sci Appl. 2017;8.