Exploring Mid-Market Strategies for Big Data Governance

Ken Knapton (ken@knaptonfamily.net)
Walden University

Research

Keywords: Big Data, Governance, Privacy, Organizational Information Processing Theory

Posted Date: December 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-114534/v1

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. 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.

Introduction

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.

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.

Methods

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. I 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. I followed a semi-structured interview approach by asking a set of predetermined questions and then interacting with participants by asking for more elaboration when responses involved new data related to the research 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

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).

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, Duan, and Cadden, (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.

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).

Results

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? The following four major themes emerged from the analysis: ensuring business centricity, striving for simplicity, establishing data source protocols, and 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 (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, Aasheim, Rutner, and Williams (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, Greco, Grimaldi, and Ritala (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.

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, van der Voort, & Wahyudi, 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, Brooks, Nepelski, Ali, & Moro, 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, Hackney, Tassabehji, & Castelli, 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, Schmiedel, & vom Brocke, 2018), and since that lack of information within an organization leads to the uncertainty to which Galbraith refers (Gupta, Kumar, Kamboj, Bhushan, & Luo, 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, Aithal, & Bhuimali, 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, Fouweather, Gregory, & Vernon, 2017). Alternatively, proper data governance addresses the standardization, accuracy, and availability of those data (Wang, Kung, & Byrd, 2018). Big data governance is critical to ensure that data

are accurate, shared appropriately, and protected per company policies (Cheng, Li, Gao, & Liu, 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, Liu, Guo, & Li, 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, Wah, Cheng, & Wang, 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, Schuhmacher, & Kuester, 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 the Internet of Things (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, Kashif, Talib, Sarwar, & Hussain, 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.

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.

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, Huang, Li, Liu, & Hu, 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, Chua, Le Breton-Miller, Miller, & Steier, 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.

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, Nassar, & Almsafir, 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.

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, Whittington, & Chang, 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.

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.

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.

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, Hendrickx, & de Montjoye, , 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).

Discussion

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.

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.

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.

Conclusion

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.

Abbreviations

OIPT: Organizational Information Processing Theory

IoT: Internet of Things

SLA: Service Level Agreement

PII: Personally Identifiable Information

Declarations

Ethics approval and consent to participate

I obtained approval from the Walden University 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. Walden University's approval number for this study is 03-13-20-0666175 and it expires on March 12th, 2021.

Consent for publication

Not applicable

Availability of data and materials

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 did not use any publicly available datasets. Transcripts of the interviews and documentation received from participating organizations will be preserved for 5 years from the date of the study. The datasets generated and/or analyzed during the current study are not publicly available due to preserving the anonymity of participants and organizations but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

Participants were not compensated in any way.

Authors' contributions

I am the sole author of this study.

Acknowledgments

Dr. Jodine Burchell was the committee chair for this doctoral study. Dr. Steven Case and Dr. Gary Griffith participated as the other members of my committee. Dr. Gail Miles was the Program Director at Walden University when this study received final approval.

References

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

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

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. *InformationFusion*, *28*, 45-59.doi:10.1016/j.inffus.2015.08.005

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Figures

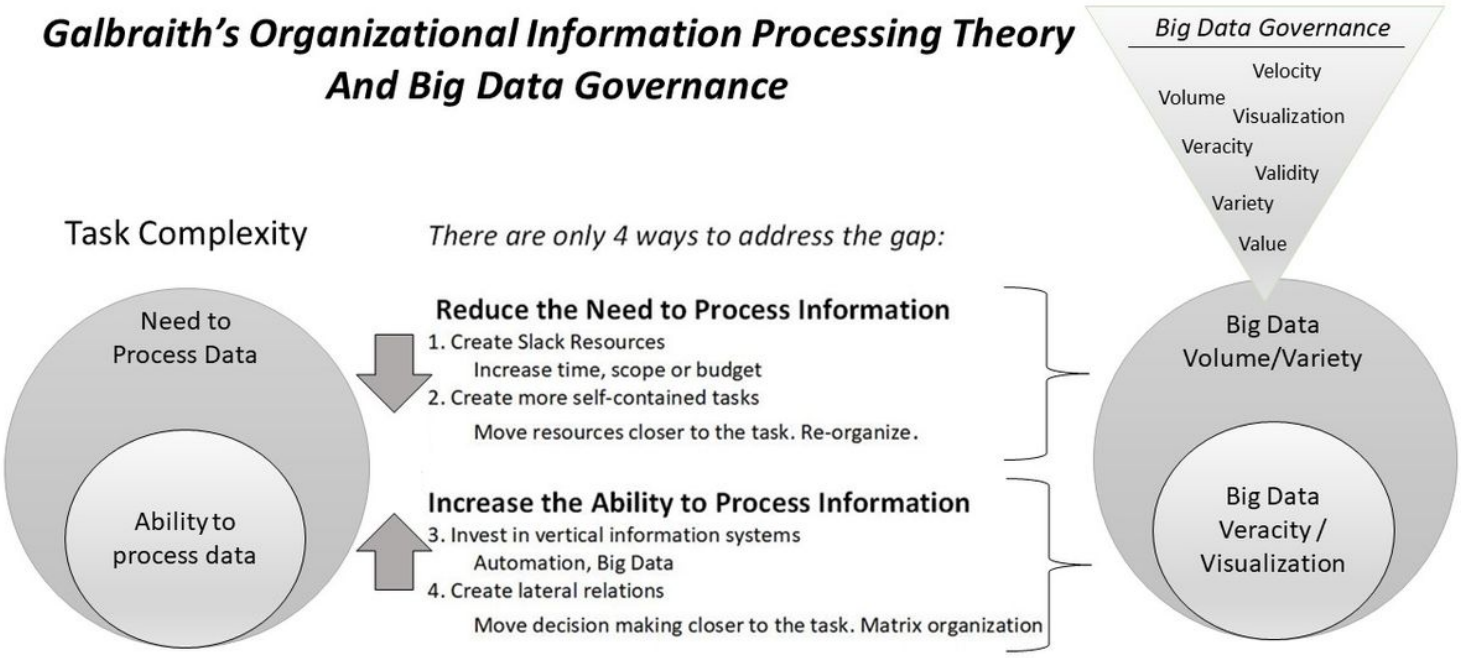

Figure 1

OITP key concepts.