Data Governance as Success Factor for Data Science

Paul Brous, Marijn Janssen, Rutger Krans

▶ **To cite this version:**

Paul Brous, Marijn Janssen, Rutger Krans. Data Governance as Success Factor for Data Science. 19th Conference on e-Business, e-Services and e-Society (I3E), Apr 2020, Skukuza, South Africa. pp.431-442, 10.1007/978-3-030-44999-5_36 . hal-03222837

HAL Id: hal-03222837

https://inria.hal.science/hal-03222837v1

Submitted on 10 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Data Governance as Success Factor for Data Science

Paul Brous[1] [0000-0002-0593-1168], Marijn Janssen[1][0000-0001-6211-8790] and Rutger Krans[2]

[1] Delft University of Technology, Delft, Netherlands
[2] Rijkswaterstaat, Utrecht, Netherlands
P.A.Brous, M.F.W.H.A.Janssen@tudelft.nl, Rutger.Krans@rws.nl

Abstract. More and more, asset management organizations are introducing data science initiatives to support predictive maintenance and anomaly detection. Asset management organizations are by nature data intensive to manage their assets like bridges, dykes, railways and roads. For this, they often implement data lakes using a variety of architectures and technologies to store big data and facilitate data science initiatives. However, the decision-outcomes of data science models are often highly reliant on the quality of the data. The data in the data lake therefore has to be of sufficient quality to develop trust by decision-makers. Not surprisingly, organizations are increasingly adopting data governance as a means to ensure that the quality of data entering the data lake is and remains of sufficient quality, and to ensure the organization remains legally compliant. The objective of the case study is to understand the role of data governance as success factor for data science. For this, a case study regarding the governance of data in a data lake in the asset management domain is analyzed to test three propositions contributing to the success of using data science. The results show that unambiguous ownership of the data, monitoring the quality of the data entering the data lake, and a controlled overview of standard and specific compliance requirements are important factors for maintaining data quality and compliance and building trust in data science products.

Keywords: Data lake, data governance, data quality, big data, digital transformation, data science, asset management.

1 Introduction

More and more, asset management organizations are introducing data science initiatives to support the digital transformation of their business processes [1]. However, in order for data science to be successful, it is vital that asset management organizations are able to trust the integrity of the digital environment [2], [3]. Managers have, in the past, found it difficult to trust data science products as, for example, the data is often found to be lacking the required quality [4]–[7]. Furthermore, as suggested by Wallis et al. [7], data collections are only as valuable as the data they contain, and users need to be able to trust the data based on the integrity of the data systems and the intrinsic quality of the data. Managers need to be able to trust data science products before they are confident enough to use these products to support their business processes to make crucial decisions [6]. Examples of these decisions in the asset management domain are maintaining dykes or replacing a bridge. Decisions in these scenarios have long term implications and wrong

decisions can be expensive and risky. A lack of trust in data science projects can often be attributed to the lack of data quality, and the success of data science projects is often highly reliant on the quality of the data being used [8]–[10]. There is no single factor defining the successful outcomes of a data science project [11], [12], but recently data governance has gained traction by many organizations as being important for ensuring quality and compliance in data science outcomes [11], [13]. However, it remains unclear how data governance contributes to the success of data science outcomes, leading to calls for more research in this area [11], [14], [15].

Data Governance can be defined as "*the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets*" [16] (p. 67), and can provide direct and indirect benefits [17]. For example, Brous et al. [14] showed that adoption of data governance can improve operational efficiency, increase revenue, reduce risk (for example with regards to privacy violations), reduce costs, improve perception of how information initiatives perform, improve acceptance of spending on information management projects, and improve trust in information products.

The main objective of the paper is to understand the role of data governance as a factor for successful data science outcomes. Our main research question therefore asks how does data governance contribute to more successful data science outcomes? This paper analyses a case study in the asset management domain with specific regard for the role of data governance as success factor for data science outcomes. The case under study is managed by Rijkswaterstaat in the Netherlands. Rijkswaterstaat is part of the Dutch Ministry of Infrastructure and Water Management and is responsible for the design, construction, management and maintenance of the main infrastructure facilities in the Netherlands. The paper reads as follows. Section 2 presents the background of literature regarding the relationship between data governance, trust and the digital environment. In Section 3 the methodology of the research is described. Section 4 describes the findings of the case study. Section 5 discusses the findings of the case study and section 6 presents the conclusions.

2 Literature Background

Although more attention has been paid to data governance in the literature in recent years, there have been several calls within the scientific community for more systematic research into data governance and its impact on the business capabilities of organizations [18]–[20]. Little evidence has been produced so far indicating what actually has to be organized by data governance and what data governance processes may entail [20], [21], and many organizations find data governance difficult to implement [22], [23]. There appears to be no "one-size-fits-all" approach to data governance [24] and the nuances attached to various domains and organizational types have not yet been extensively described [25], [26]. Furthermore, evidence is scant as to the role data governance plays in ensuring the successful outcomes of data science initiatives [18], [19].

Recent years have witnessed more and more asset management organizations adopting data science initiatives in order to support the digital transformation of their business processes [27], [28], and Van der Aalst [29] go so far as to suggest that organizations without a data science capability may not survive. According to Provost and Fawcett [1] (p.52), data science is "*a set of fundamental principles that support and guide the principled extraction of information and knowledge from data*". From this perspective, data science encompasses a broad range of

knowledge and capabilities such as data-mining and machine learning, which are designed to extract knowledge from data and are important for creating value and moderating risk in data science initiatives. As such, data governance can help organizations make use of data as a competitive asset [21], [33]. Data governance aims at maximizing the value of data assets in enterprises [1], [37]. For example, capturing electric- and gas-usage data every few minutes benefits the consumer as well as the provider of energy. With active governance of big data, isolation of faults and quick fixing of issues can prevent systemic energy grid collapse [38].

Data science can improve asset management decision-making which is needed to facilitate more efficient and secure asset management operations, as well the need for better situational awareness about network disturbances [10], [27]. Data science initiatives such as predictive maintenance modelling generally require big data [10], [30], [31]. Asset management organizations often choose to implement data lakes using a variety of architectures and technologies to store big data and to make this data available for use. A data lake is "*a central repository system for storage, processing, and analysis of raw data, in which the data is kept in its original format and is processed to be queried only when needed*" [32] (p. 456). Data lakes are different to traditional data warehouses which often have their own native formats and structures as data is stored in its original, raw, format [33], [34]. Often, the data processing systems which are required to allow the data to be ingested without compromising the data structure are also included in the definition [32], [34]. The data in the data lake is generally immediately accessible, allowing users to utilize dynamic analytical applications [34], [35]. This immediate accessibility, as well as the retaining of data in its original format presents a number of challenges regarding management of the data lake, including data quality management, data security and access control [33], [36], as well as in maintaining compliance with regards to privacy [21], [36]. As such, data governance has increasingly gained popularity as a means of ensuring data quality and maintaining compliance.

Managing data quality is considered by many researchers to be an important reason for adopting data governance (e.g. [24], [37], [39]). However, big data can provide asset management organizations with complex challenges in the management of data quality. According to Saha & Srivastava [40], the massive volumes, high velocity and large variety of automatically generated data can lead to serious data quality management issues which can be difficult to manage in a timely manner [41]. For example, IoT sensors calibrated to measure the salinity of water may, over time, begin to provide incorrect values due to biofouling. Data science information products often rely on near real-time data to provide timely alerts, and, as such, problems may arise if these data quality issues are not timely detected and corrected.

As well as establishing data management processes which manage data quality, data governance should also ensure that the organization's data management processes are compliant with laws, directives, policies and procedures [42]. For example, Panian [43] states that establishing and enforcing policies and processes around the management of data should be the foundation of effective data governance practice as using big data for data science often raises ethical concerns. Automatic data collection may cause privacy infringements [44], [45] such as cameras used to track traffic on highways which often record personally identifiable data such as number plates or faces of persons in the vehicles. Data governance processes should ensure that these personally identifiable features are removed before data is shared or used for purposes other than legally allowed. Data governance should therefore establish what specific data privacy policies are appropriate [39] and applicable across the organization [38]. For example, Tallon [46] states that

organizations have a social and legal responsibility to safeguard personal data, whilst Power & Trope [47] suggest that risks and threats to data and privacy require diligent attention from organizations.

In summary, asset management organizations often choose to implement data science initiatives such as predictive maintenance and anomaly detection, using methods such as data-mining and machine learning, in order to support the digital transformation of their business processes. Many modern data science methods require big data which is often stored and made available through data lakes. However, asset management organizations are increasingly being faced with challenges which impact the success of data science outcomes, often related to: 1. a lack of trust in the quality of data [40], [41], 2. whether or not the data is being used in an ethical way [46], and 3. whether or not the management and use of the data is compliant with relevant legislation and internal policies [47]. In order to tackle these challenges, data governance assigns responsibilities for decision-making [24], defines processes for monitoring an managing data quality [41], and defines policies for monitoring and maintaining compliance with relevant legislation [47].

The propositions of the research are based on the results of the background literature review as well as on existing theory regarding the principles of data governance in asset management organizations and the reasons why asset management organizations choose to implement data governance [13], [14], [48]. The propositions of the research therefore read as follows:

1. Defining clear roles and responsibilities for data management will result in easier generation of business value from data science efforts.
2. Monitoring and managing data quality will result in more useful outcomes from data science efforts.
3. Compliance monitoring and control is a required condition for data science.

As discussed above, the literature shows that many organizations have implemented data governance in an attempt to improve trust in data science efforts through the improved management of data quality and compliance to relevant legislation.

3 Methodology

This paper describes a single case study using a multi-method approach to investigate the role of data governance as success factor for data science. Case study is a widely adopted method for examining contemporary phenomenon such as the adoption of data governance [49], [50]. In this research we analyze a single case, following the design of an explanatory case study research proposed by Yin [51], including the research question, the propositions for research, the unit of analysis, and the logic linking the data to the propositions. Single case study was selected as being appropriate for this research as there is a need to investigate data governance as success factor for data science in greater detail. In this regard, single case studies may be more appropriate than multiple case studies, as a single case study provides the opportunity to have a deeper understanding of data governance in a specific context [51], [52], in this case, data science efforts in the asset management domain. As suggested by Eisenhardt [50], the research was contextualized by a review of background literature, identifying the generally accepted roles of data governance in a data science context. The literature background reveals data science initiatives often face a number of challenges, and not all efforts lead to successful outcomes [15], [48], [53]. Facing these challenges has led many organizations to adopt data governance as a means of

improving the outcomes of data science efforts [13]. However, data governance remains a poorly understood concept [22], [36] and its contribution to the success of data science has not been widely researched [36]. As discussed above, our main research question therefore asks *how does data governance contribute to more successful data science outcomes*?

Following Ketokivi & Choi [54], deduction type reasoning augmented by contextual considerations provided the basic logic for the propositions to be tested in a particular context, namely data science in an asset management domain. The data analysis in this research utilizes "within case analysis" [55]. Within case analysis helped us to examine the impact of data governance on the success of data science in a single context. In this case, the unit of analysis was a single data science project in the asset management domain. The case selected was managed and implemented by Rijkswaterstaat, often abbreviated to RWS and referred to as such in this paper. RWS is the Directorate-General for Public Works and Water Management and an operational agency of the Ministry of Infrastructure and Water Management of the Netherlands. RWS is charged with the management and maintenance of the major highways, waterways and shipping lanes in the Netherlands. In order to prepare the organization for the case study research project, RWS was provided with information material outlining the objectives of the project.

Following the suggestions of Yin [51], the case study was conducted using a multi-method approach and multiple data sources were used. Methods used are document analysis and face-to-face interviews. The interviews were conducted during 2019 taking the form of one-on-one, face-to-face interviews. The interviewees were mainly selected from RWS staff members directly involved in the data science project in various roles, but also included other staff members involved in the governance and management of the data and the monitoring of the data in order to ensure saturation. Secondary data sources included relevant internal documentation, including project reports, data governance workshop reports, and data and information technology strategy documents. Company websites which included relevant data governance information and reports on the data science case were also included. Triangulation of aspects of data governance which contribute to the successful outcome of the data science case was made by listing aspects of data governance found in internal documentation and testing these in the one-on-one interviews. In the interviews the interviewees were asked as to the contribution of these aspects of data governance towards the successful outcome of the project. In the interviews the interviewees were also asked to name other aspects of data governance that may have had a significant contribution to the successful outcome of the data science project but which may have been overlooked.

4 Case Study Description

RWS is tasked with the management and maintenance of the national public infrastructure including the construction and maintenance of shipping lanes, major waterways (including flood prevention) and national roads and highways. RWS has a spend of approximately €200 million per annum on asphalt maintenance, with operational parameters traditionally focused on traffic safety. In the past this has led to increasing overspend due either to premature maintenance, or to expensive emergency repairs. The prediction of asphalt lifetime based on traditional parameters has been shown to be correct one third of the time. RWS is seeking to reduce these costs by extending the lifespan of asphalt where possible whilst reducing the number of emergency repairs made by adopting data science techniques for the purpose of predictive, "just-in-time"

maintenance. Using available big data in a more detailed manner, such as raveling data collected by a Laser Crack Measurement System combined with Weigh-in-Motion data has doubled the prediction consistency. According to RWS officials, improving the accuracy of asphalt lifetime prediction has enabled better maintenance planning which has significantly reduced premature maintenance, improving road safety and cost savings, and reducing the environmental impact due to reduced traffic congestion and a reduction in CO^2 emissions. The data science model uses data related to traditional inspections, historical data generated during the laying of the asphalt, road attribute data and planning data, as well automatically generated, streaming data such as weather data, traffic data, and IoT sensor data. The current model takes about 400 parameters into consideration. According to an RWS official, "this number will only grow, as the (*project partners*) continue to supply new data". According to RWS, the ultimate goal is a model that can accurately predict the lifespan of a highway.

With regards to defining *roles and responsibilities* RWS has asked the data managers of each of the datasets used in the data science project to each appoint an executive sponsor or data owner. The data owner is a business sponsor. Once ownership is established, the current and desired future situations are assessed in terms of production and delivery. A roadmap is then established which was translated into concrete actions and a delivery agreement is reached. RWS also uses "open" data from external sources. Due to its many open data partnerships, RWS has implemented a policy of providing knowledge, tools and a government-wide contact network in which best practices are shared with other government organizations. These best practices refer to organization of data management, data exchange with third parties, data processing methods and individual training. According to staff members, RWS has implemented data governance for their big data in order to remain "future-proof, agile and to improve digital interaction with citizens and partners". According to an RWS executive manager, "RWS wants to be careful, open and transparent about the way in which it handles big and open data and how it organizes itself". Furthermore, RWS has introduced the policy of assessing and publishing the monetary cost of data assets in order to raise awareness of the importance of data quality management. This means that every RWS process and every RWS organizational unit is encouraged to be aware of its data needs and the incurred costs.

With regards to *data quality*, RWS has implemented a data quality framework to improve their control of data quality. RWS staff believe that "the return (of the investment) stands or falls with the quality of data and information". As such, according to RWS staff, the underlying quality of the data and information is of great importance to work in an information-driven way. RWS staff members have suggested that, in the past, a significant amount of production time has often been lost due to inadequate data quality. The RWS data quality management process follows an eight step process which begins by identifying: 1. the data to be produced, 2. the value of the data for the RWS primary processes, and 3. a data owner. RWS has developed an automatic auditing tool (AAT) in combination with a Manual Auditing Tool (MAT) to monitor the quality of the data as a product in order to further improve its grip on data quality. According to RWS staff, the AAT and the MAT ensured that quality measurements were mutually comparable, provided tools for more focused management, and caused a change in the conscious use of data as a strategic asset. Alongside with the AAT, the MAT is considered important as it is not yet possible to automate the monitoring of all data quality dimensions. Data quality measuring is centralized at RWS, the goal being to ensure a standardized working method. However, RWS maintained the policy that every data owner is responsible for improvements to the data management process

and the data itself. The RWS data quality framework was based on fitness for use and data quality measurement was maintained according to 8 main dimensions and 47 subdimensions.

With regards to *compliance*, RWS has translated their data policies and principles into a data agenda in which the opportunities, risks and dilemmas of their data policies and ambitions are identified in advance and are made measurable and practicable. Terms and definitions have been coordinated with the Dutch legal framework related to the environment to ensure compliance. Responsibilities relating to compliance to privacy laws are centralized and RWS has assigned privacy officers to this role. The CIO has the final responsibility for ensuring that privacy and security are managed and maintained, however, business data owners are held accountable for ensuring compliance to dataset specific policy and regulations.

5 Discussion

Case study methodology was used in this research to identify the role that data governance plays as success factor for data science. The choice for an in-depth, single case study was based on the contemporary nature of both data science and data governance and the need to study data governance as success factor for data science in greater depth. The study was conducted as a single case study and the results should be regarded in this light. Single case study has been criticized in the past due to the difficulty of providing a generalizing conclusion [51], [56]. In order to overcome this, the data collection made use of multiple sources including reports, presentations and face-to-face interviews. More research is recommended in this area to test the applicability of the propositions in other domains and organizational types. The study was conducted in the asset management domain as asset management organizations by nature are often data rich due to the need to monitor the state of the infrastructure assets. This may limit the applicability of the study for domains which are less data intensive, however the essence of generating value from data is likely to be the same in other domains.

5.1 Proposition 1: Defining clear roles and responsibilities for data management will result in easier generation of business value from data science efforts

Proposition 1 proposes that data science is likely to generate more business value if responsibilities for data management are clearly defined. RWS has many various open data partners, as well as a large variety of sources from which the data is collected. As a result RWS has experienced difficulties in managing responsibilities for data quality and data management processes. RWS has therefore assumed a leadership role in maintaining a government-wide contact network in which knowledge, tooling and best practices with regards to data management and data sharing are shared with other government organizations. Internally, RWS has assigned business sponsors to assume ownership of datasets so that roles and responsibilities of data management are clearly defined. In order to ensure that sufficient resources are made available for data quality management, RWS has also defined a “price” for each dataset so that business owners are aware of the value of each dataset. This allows the organization to treat the data as a business asset, promoting the need to maintain the expected quality of each dataset.

5.2 Proposition 2: Monitoring and managing data quality will result in more useful outcomes from data science efforts

Proposition 2 proposes that data science is more likely to result in useful outcomes if data quality is monitored and controlled. RWS actively monitors their data inputs by means of an "automatic audit tool". RWS has assembled a library of business rules which form the input for the calculation of the data quality. The results of the calculations are displayed in the form of a dashboard which indicates whether the calculated values fall within acceptable limits or not. The acceptable limits are described in the RWS data quality framework which has standardized the calculation and description of data quality throughout RWS. The results of the data quality monitor are used to define which interventions need to be taken in order to achieve the desired levels of quality and also to monitor the effects of the interventions on the data quality. Traditionally, data quality projects at RWS were based on "hearsay" from staff whereby the general feeling was that the quality was below requirements. The AAT has allowed RWS to be more data driven with regards to their data management processes. According to RWS staff, the active monitoring of data quality has led to "identification of gaps in data governance, harmonization of processes across organizational departments, increased awareness and cost savings".

5.3 Proposition 3: Compliance monitoring and control is a required condition for data science

Proposition 3 proposes that compliance with relevant legislation is a necessary and required condition for data science. RWS has had a central, IT-centered approach to data privacy to ensure that legal requirements and guidelines regarding the European General Data Protection Regulation (GDPR) are standardized and consistent throughout the organization. RWS has published a transparent list of systems in which personal data is collected, and has published detailed instructions as to how personal data may be viewed and, where necessary, deleted. RWS has appointed privacy and compliance officers to assume this responsibility and has appointed the CIO has the responsible executive sponsor. The monitoring of other compliance related activities is done using the AAT or the MAT. Responsibility for the actions flowing from the results of the AAT or the MAT lies with the data managers and ownership lies with the data sponsor. This hybrid approach allows RWS to standardize compliance processes where possible, whilst also being able to tailor customized solutions for particular data issues. Currently the feasibility of a nationwide data platform for asphalt pavement data is being explored in which easy data accessibility, authorization, storage, scalability, architecture, plateau planning, solution directions and cost estimations are addressed.

6 Conclusions

In this research paper we analyzed a case study regarding the governance of data in a data lake in the asset management domain to identify factors contributing to the success of using data science. The objective of the case study is to understand the role of data governance as success factor for data science. The case under study is a data science project which predicts the maintenance requirements of asphalt on national highways over time. Three propositions were defined on the basis of existing theory on data governance, namely: 1. defining clear roles and

responsibilities for data management will result in easier generation of business value from data science efforts, 2. monitoring and managing data quality will result in more useful outcomes from data science efforts, and 3. compliance monitoring and control is a required condition for data science. These propositions were derived from the literature and confirmed in the case study, suggesting that data governance should be regarded as an important success factor for data science outcomes. The results show that clearly defined ownership of the data, monitoring the quality of the data entering the data lake, and a controlled overview of compliance requirements are important factors for successful data science outcomes. The results also show that efficient management of compliance may be performed by developing centrally managed, standardized solutions for privacy and security requirements. However, system-specific compliance requirements need to be developed by data managers and these requirements should be owned by a business sponsor who assumes responsibility for these requirements. As such, the results show the data governance is an important success factor for data science outcomes as it ensures that data quality and compliance are effectively managed.

7 References

[1] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," Big Data, vol. 1, no. 1, pp. 51–59, Feb. 2013.

[2] Council on Library and Information Resources, Ed., Authenticity in a digital environment. Washington, D.C: Council on Library and Information Resources, 2000.

[3] R. Randall, D. Peppers, and M. Rogers, "Extreme trust: the new competitive advantage," Strategy & Leadership, Nov. 2013.

[4] S. Lin, J. Gao, and A. Koronios, "The need for a data quality framework in asset management," presented at the Australian Workshop on Information Quality, Adelaide, Australia, 2006, vol. 1.

[5] J. Symons and R. Alvarado, "Can we trust Big Data? Applying philosophy of science to software," Big Data & Society, vol. 3, no. 2, p. 205395171666474, Dec. 2016.

[6] S. Passi and S. J. Jackson, "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects," Proc. ACM Hum.-Comput. Interact., vol. 2, no. CSCW, pp. 1–28, Nov. 2018.

[7] J. C. Wallis, C. L. Borgman, M. S. Mayernik, A. Pepe, N. Ramanathan, and M. Hansen, "Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries," in Research and Advanced Technology for Digital Libraries, vol. 4675, L. Kovács, N. Fuhr, and C. Meghini, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 380–391.

[8] G. Manco et al., "Fault detection and explanation through big data analysis on sensor streams," Expert Systems with Applications, vol. 87, pp. 141–156, Nov. 2017.

[9] D. Lee and R. Pan, "Predictive maintenance of complex system with multi-level reliability structure," International Journal of Production Research, vol. 55, no. 16, pp. 4785–4801, 2017.

[10] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in Bulk Power System Dynamics and Control - IX

Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium, 2013, pp. 1–9.

[11] J. S. Saltz and I. Shamshurin, “Big data team process methodologies: A literature review and the identification of key factors for a project’s success,” in 2016 IEEE International Conference on Big Data (Big Data), Washington DC,USA, 2016, pp. 2872–2879.

[12] P. Cato, P. Golzer, and W. Demmelhuber, “An investigation into the implementation factors affecting the success of big data systems,” in 2015 11th International Conference on Innovations in Information Technology (IIT), Dubai, United Arab Emirates, 2015, pp. 134–139.

[13] P. Brous, P. Herder, and M. Janssen, “Governing Asset Management Data Infrastructures,” Procedia Computer Science, vol. 95, pp. 303–310, 2016.

[14] P. Brous, M. Janssen, and R. Vilminko-Heikkinen, “Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles,” in International Conference on Electronic Government and the Information Systems Perspective, 2016, pp. 115–125.

[15] A. Yoon, “Data reusers’ trust development,” Journal of the Association for Information Science and Technology, vol. 68, no. 4, pp. 946–956, 2017.

[16] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge. Technics Publications, 2017.

[17] J. Ladley, Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program. Newnes, 2012.

[18] J. Fruehauf, F. Al-Khalifa, J. Coniker, and L. L. P. Grant Thornton, “Using The Bolman And Deal’s Four Frames In Developing A Data Governance Strategy,” Issues in Information Systems, vol. 16, no. 2, 2015.

[19] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, “The rise of ‘big data’ on cloud computing: Review and open research issues,” Information Systems, vol. 47, pp. 98–115, 2015.

[20] B. Otto, “A morphology of the organisation of data governance.,” in ECIS, 2011, vol. 20, p. 1.

[21] V. Morabito, “Big Data Governance,” in Big Data and Analytics, Springer, 2015, pp. 83–104.

[22] C. A. Mathes, “Big data has unique needs for information governance and data quality,” Journal of Management Science and Business Intelligence, vol. 1, no. 1, pp. 12–20, 2016.

[23] N. Thompson, R. Ravindran, and S. Nicosia, “Government data does not mean data governance: Lessons learned from a public sector application audit,” Government Information Quarterly, vol. 32, no. 3, pp. 316–322, Jul. 2015.

[24] K. Wende and B. Otto, “A Contingency Approach to Data Governance,” presented at the International Conference on Information Quality, Cambridge, USA, 2007.

[25] C.-S. Wang, S.-L. Lin, T.-H. Chou, and B.-Y. Li, “An integrated data analytics process to optimize data governance of non-profit organization,” Computers in Human Behavior, vol. 101, pp. 495–505, Dec. 2019.

[26] R. Abraham, J. Schneider, and J. vom Brocke, “Data governance: A conceptual framework, structured review, and research agenda,” International Journal of Information Management, vol. 49, pp. 424–438, Dec. 2019.

[27] M. A. Waller and S. E. Fawcett, “Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management,” Journal of Business Logistics, vol. 34, no. 2, pp. 77–84, 2013.
[28] S. J. Berman, “Digital transformation: opportunities to create new business models,” Strategy & Leadership, vol. 40, no. 2, pp. 16–24, 2012.
[29] W. Van Der Aalst, “Data science in action,” in Process Mining, Springer, 2016, pp. 3–23.
[30] H. Chen, R. H. Chiang, and V. C. Storey, “Business intelligence and analytics: from big data to big impact.(Special Issue: Business Intelligence Research)(Essay),” MIS Quarterly, 2012.
[31] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” International Journal of Production Economics, vol. 165, pp. 234–246, 2015.
[32] J. Couto, O. Borges, D. Ruiz, S. Marczak, and R. Prikladnicki, “A mapping study about data lakes: An improved definition and possible architectures,” presented at the Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE, 2019, vol. 2019-July, pp. 453–458.
[33] C. Madera and A. Laurent, “The next information architecture evolution: The data lake wave,” presented at the 8th International Conference on Management of Digital EcoSystems, MEDES 2016, 2016, pp. 174–180.
[34] N. Miloslavskaya and A. Tolstoy, “Big Data, Fast Data and Data Lake Concepts,” Procedia Computer Science, vol. 88, pp. 300–305, Jan. 2016.
[35] S. Ullah, M. D. Awan, and M. Sikander Hayat Khiyal, “Big Data in Cloud Computing: A Resource Management Perspective,” Scientific Programming, 2018. [Online]. Available: https://www.hindawi.com/journals/sp/2018/5418679/. [Accessed: 18-Oct-2019].
[36] J. A. Kroll, “Data Science Data Governance [AI Ethics],” IEEE Security & Privacy, vol. 16, no. 6, pp. 61–70, 2018.
[37] D. B. Otto, “Data Governance,” Bus Inf Syst Eng, vol. 3, no. 4, pp. 241–244, Jun. 2011.
[38] P. Malik, “Governing Big Data: Principles and practices,” IBM J. Res. Dev., vol. 57, no. 3–4, p. 1, Jul. 2013.
[39] V. Khatri and C. V. Brown, “Designing data governance,” Commun. ACM, vol. 53, no. 1, pp. 148–152, Jan. 2010.
[40] B. Saha and D. Srivastava, “Data quality: The other face of big data,” in 2014 IEEE 30th International Conference on Data Engineering, 2014, pp. 1294–1297.
[41] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, “Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications,” International Journal of Production Economics, vol. 154, pp. 72–80, 2014.
[42] D. Wilbanks and K. Lehman, “Data governance for SoS,” International Journal of System of Systems Engineering, vol. 3, no. 3–4, pp. 337–346, 2012.
[43] Z. Panian, “Some practical experiences in data governance,” World Academy of Science, Engineering and Technology, vol. 38, pp. 150–157, 2010.
[44] G. Cecere, F. Le Guel, and N. Soulié, “Perceived Internet privacy concerns on social networks in Europe,” Technological Forecasting and Social Change, vol. 96, pp. 277–287, 2015.

[45] T. van den Broek and A. F. van Veenstra, “Governance of big data collaborations: How to balance regulatory compliance and disruptive innovation,” Technological Forecasting and Social Change, vol. 129, pp. 330–338, 2018.
[46] P. P. Tallon, “Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost,” Computer, vol. 46, no. 6, pp. 32–38, Jun. 2013.
[47] Power and Trope, “The 2006 survey of legal developments in data management, privacy, and information security: The continuing evolution of data governance,” Bus. Lawyer, vol. 62, no. 1, pp. 251–294, 2006.
[48] P. Brous, M. Janssen, D. Schraven, J. Spiegeler, and B. C. Duzgun, “Factors Influencing Adoption of IoT for Data-driven Decision Making in Asset Management Organizations,” presented at the 2nd International Conference on Internet of Things, Big Data and Security, 2017, pp. 70–79.
[49] J. Choudrie and Y. K. Dwivedi, “Investigating the research approaches for examining technology adoption issues,” Journal of Research Practice, vol. 1, no. 1, p. 1, 2005.
[50] K. M. Eisenhardt, “Building Theories from Case Study Research,” The Academy of Management Review, vol. 14, no. 4, pp. 532–550, Oct. 1989.
[51] R. K. Yin, Case Study Research: Design and Methods. SAGE, 2009.
[52] J. Gustafsson, “Single case studies vs. multiple case studies: A comparative study,” Engineering and Science, Halmstad University, Halmstad, Sweden, pp. 1-15, 2017.
[53] Q. H. Cao, I. Khan, R. Farahbakhsh, G. Madhusudan, G. M. Lee, and N. Crespi, “A Trust Model for Data Sharing in Smart Cities,” presented at the IEEE International Conference on Communications 2016 (ICC 2016), 2016.
[54] M. Ketokivi and T. Choi, “Renaissance of case research as a scientific method,” Journal of Operations Management, vol. 32, no. 5, pp. 232–240, Jul. 2014.
[55] M. B. Miles and A. M. Huberman, Qualitative Data Analysis: An Expanded Sourcebook. SAGE, 1994.
[56] Z. Zainal, “Case study as a research method,” Jurnal Kemanusiaan, vol. 5, no. 1, 2017.