A Conceptual Framework for Big Data Governance

Abdulrahman Albladi and Eisa Alanazi

Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia

Summary

Big Data management has attracted a lot of attention in the past few years. This is driven by a large and complex data continue to be generated daily in different aspects of our lives. Hence, there is a clear need for regulations and policies to govern how to extract, process, and collect big data.

Governing big data is one of the most critical issues that need to be investigated. This paper defines a conceptual framework for big data governance. We outline five basic steps to ease the process of governing big data.

Key words:

Big Data, Data Governance, Framework

1. Introduction

Most of organizations are seeking to adopt technology in their business to facilitate their work. As an impact of this adoption a new field became a trend which is Big Data. Big data is a term that refers to a massive amount of data that cannot be processed or stored by traditional software. People started to notice the increasing amount of information or what known as “information explosion” by 1941. However, the “big data” concept started to arise in 1980, according to the Oxford English dictionary. Some researchers define big data in terms of technologies and analytical methods. However, most of them extend Laney’s 3 V’s theory [13] to define big data. The value of big data depends on the ability to collect, prepare, and analyze it to make data-driven decisions.

Laney’s theory was initially used as data dimensions that enterprises must deal with to manage their data. The three V’s are Volume, Velocity, and Variety. Volume refers to the size of big data in terabytes or the number of records, tables, and files. Therefore, this big size of data needs scalable storage units and advanced processing techniques. Velocity refers to the speed of data growth and the needed analyzing speed: mobile apps, sensors, monitoring cameras, ...etc. Produces a stream of data in a brief period, this fast-growing stream, most of the time, needs a real-time analysis to make immediate and customized decisions [8]. Variety refers to the diverseness of data structures and types. Due to the variety and increase in data sources, they produce various types of data: structured e.g., databases, unstructured e.g., text, pictures, videos, audios, and semi-structured e.g., XML. New innovative technologies needed to prepare and organize these data in order to leverage them. In addition to volume, velocity, and variety, some other dimensions of data were defined to describe big data. Researchers at IBM introduces Veracity as the fourth V [17]; it refers to data noise and uncertainty. Moreover, the level of reliability of big data affects the decision-making process. Through the years, researchers continue adding more dimensions to define big data. De Mauro et al[7]. (2016) analyzed most common researches on big data to propose a formal unified definition, that is “Big Data is the Information asset characterized by such a High Volume, Velocity, and Variety to require specific Technology and Analytical Methods for its transformation into Value.” [7]. Volume, Velocity, and Variety describe the dimensions of big data, while technology and analytical methods refer to the required resources to turn information assets into valuable data.

To ensure the benefits of utilizing data and avoiding any unexpected missuse of data, Data Governance principle showed. Governance as term indicates the method used by the organization to ensure that strategies are developed, monitored and achieved [18]. However, Data Management Association provide a definition of Data Governance as “the exercise of authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets”. Moreover, Data Governance is above data management and more detailed [1]. The authors in 16] defined Data Governance as “a system of decision rights and accountabilities for information-related processes, executed according to agreed upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods”. In addition, the work in [15] proposed a different definition of Data Governance as “specifies the framework for decision rights and accountabilities to encourage desirable behavior in the use of data. To promote desirable behavior, data governance develops and implements corporate-wide data policies, guidelines, and standards that are consistent with the organization’s mission, strategy, values, norms, and culture”. However, the definition of Data Governance is not is not unique. Generally, data governance is not a technical feature. It all about policies, guidelines and standards.

2. Importance of Big Data Governance

As per [11] the reasons of adopting data governance in organizations are data accountability, security improvement,

Manuscript received January 5, 2020
Manuscript revised January 20, 2020

reduction of overall costs, consistency between data and business functions, and provision of quality. Big Data Governance is at the top of organizations concerns the reason of that is losing data governance means losing control of their crucial data [10]. Thus, applying governance in such an organization will allow them to control their data. Another point as per McKinsey, the productivity of some sector in US was improved after utilizing big data which has a direct effect to enhance economy [14].

Moreover, any organization regardless the category if it applies the suitable governance framework in term of big data it will improve their economic position as well as the organization activity. Data Quality is a core for the governance process. Also, a strong governance framework is important for companies [12] to helps the decisions-makers to obtain the values of governance.

3. Related Work

To stay in global markets Companies are required to enhance their business processes. The architecture of processes and IT strategy of an organizations are the bases of customers' needs and service offers. The authors propose a model for flexible Data Governance as well as a model to document the company-specific decision-making framework to improve Data Quality Management[19][6]. Data Quality Management is pushed in the area between business and IT, it is responsible to provide stakeholders with accurate data however this data is not necessarily met their requirements, therefore companies will face different issues in applying Data Quality Management. Academic Data Governance is not yet covering all organizations aspects.

Data Governance supposed to address decision rights and accountabilities to encourage desirable behavior in the use of data however the current structure is making all organizations had the same approach which may lead to lack of decision-making. IT governance is more flexible than Data Governance, the IT governance researches proposes a model with three elements: roles, major decisions areas and assignment of accountabilities. The authors in [19] adopt this model to help organizations to build a flexible Data Governance by using IT governance model. However, Data Governance is not a full subset of IT governance, therefore IT and business professionals should work closely to follow corporate governance principles which is a combination between Data Governance and IT governance. The work in [19] comprised of Data Quality Management roles, decision areas and main activities, and responsibilities, and arranged these three in a matrix to define the relations between roles and decision areas. Authors also analyzed studies and reports to make roles and committees with full description for each of them as in . Data Governance model is consisting only the fundamental decision areas and main activities in improving and maintaining corporate data quality. The author in [19] review the existing Data Quality Management approaches and wrote the most consensus decision areas. To specify responsibilities, he came with A Responsibility Assignment Matrix that identify the participants in making decisions and how they make them. Finally this model could helps organization to build a Data Governance model and making decisions more effective and accountable, however the are some limitations in this model refers to build it on IT governance model as well as this model does not cover both the effectiveness aspect and the efficiency aspect at all levels of the organization.

Recently data became very explosive after interchange of information technology with economic society which make data as resource for countries. This explosion of data affects different dimension of economy. After Studied multiple theories and they came to the fact that data governance is used by some industries even though it's not widely used [6][4]. Each of these theories is taking data governance from different perspectives. Authors in [6] conclude that all these theories are not compatible with the current market situation. Therefore, they developed a model for cloud data governance. As a start they defined cloud data governance as: targeting cloud data as a main objective for governance by making policies, strategies, management plans, operations, security privacy and architecture for data to reach the perfect governance model. Data can be stored in two different types: local storage which is the main storage and not using internet and the other type is cloud storage which become more used recently due to widespread applications and user demand over Internet. The cloud data contain many car stones that needs to be evaluate, plan, build and optimize to achieve the full benefits of data as well as applying data governance model. In this framework of cloud data governance, the author suggested a model consist of six core categories and twenty-three processes area, moreover, these six card categories are cloud data strategy, cloud data management, cloud data quality, cloud data operation, cloud data architecture and security privacy. Each one of them has multiple processes, furthermore, each process has proper definition and function. In order to connect these categories, the authors in [6] apply Plan, Do, Act and Check model. PDAC model is knwon as cycle for continual improvement [9]. In addition, to guarantee a mature model, a classification for capacity of cloud data governance divided to: performance, management, definition, measurement and optimization. Each of these levels has his own description and characteristics. Finally to assess the maturity, the authors in [6] created a method which can supervise, evaluate and improve data governance capacity. This method has two main criteria, the first is organizational self-evaluation which is made internally within the organization to improve data governance system.

On the other is hand the third party assessment Which is deepest and the most helpful criteria to improve data governance in the organization.

Big data is a new generation of technology, it can achieve a huge value for organizations and companies. Big data governance reorganized big data by applying roles and regulation for it. The main goal of information governance is to help decision-making, moreover, achieving this goal required sharing, reliability and high-quality of information. Akoka et al [2], studied for current big data organizations came with results that there are many challenges facing these organizations to adopt Big Data Governance. However, to evaluate the governance, authors defined an approach that consider governance as a system has different dimensions, these dimensions must be evaluated by appropriate criteria. Their approach is to evaluate big data governance as a system considering multi-criteria needed for this evaluation. Authors also developed a generic multi criteria hierarchy as well as evaluation procedure to meet their[2] approach. Big data governance has many elements to be considered as a system. One of these elements is the purpose of big data governance which is decision-making. The other one is achieving organizations objectives by implementing big data governance.

Moreover, authors in [2] combined results of existing researches to make generic multi-criteria hierarchy. This generic multi criteria hierarchy depends on five dimensions: goal, environment, structure, activity and evolution. Based on existing process called Analytic Hierarchy Process, [2] created the evaluation procedure. Depends on the dimensions and some associated criteria form a hierarchical tree of nodes: (i) non-terminal nodes for the dimensions and sub dimensions. (ii) terminal nodes for elementary domains. To evaluate dimensions, the evaluator gives a weight for each domain in this particular dimension. This research proposes a new approach to evaluate big data governance which can be useful and applicable. Arguably we can say that this approach is discovering new areas of governance that was not reached before. However due to the research date this approach has a validity issue because it is not yet applied. Data Governance is the new concept to manage Big Data platforms by enabling road map for data transactions. Cloud computing is a very promising technology to enhance data utilization, however, adopting this technology is not widespread because it needs more rules and Policies to manage data.

Organizations should consider Data Governance strategy before they apply cloud computing to face the major concern of cloud users which is losing control on their data, by applying Data Governance organizations will gain security, privacy and quality of data. After reviewing the current designs that made by cloud computing companies, Al-Ruithe et al[4], found that designing Data Governance for cloud computing is complex, thus they proposed a design for a framework which focusing on five key processes that can save both of provides and users in could. The difficulty of applying Data Governance on cloud comes from making rules for cloud data rather than traditional IT. Making these rules for Data Governance in cloud is shared between IT, business and legal departments and they should identify these rules carefully to avoid any issue could happen in term of data management and privacy.

Due to the lack of papers that address applying of Data Governance in cloud computing, the authors in [4] proposed five effective steps to be as new framework for any organization will transfer to cloud computing environment and covering the data transactions side with a strong rules agreed by all concerned parties. These steps started by setting the structure of Data Governance to clarify roles and responsibilities between related teams, then have a look at the current Data Governance rules and evaluate it. After that making the functions and activities that should be considered while implementation cloud computing services. Then it's an important initiate negotiation contract to evaluate cloud service providers and finally after choosing a provider they should create Service level agreement that include all policies and guidelines [4]. As we can see in [4], there is a flexible framework can be applied in many government departments, semi-governments or private organizations. This framework is combining both of business and IT to design and implement data governance.

Governance in the cloud needs to understand, reasonable and modify the relationships between distinctive cloud actors or stakeholders in terms of roles and responsibilities [5]. However, there are some limitations founded as that this framework was not applied and measured to check the usability of it as well as the steps need to be rearranged assessing the currant governance should become first before building a structure, however, the language of the work in [4] is simple and easy to comprehend by the public. Comparing frameworks above, we can see that [3] framework was mainly designed for decision makers in organizations. This framework able to be applied in different types of organizations public or private regardless the size of these organizations (i.e., small, medium or large). The complexity of interactions for data access, processing and/or updates between the cloud consumer and cloud provider specially in public sector led the authors in [3] to design this conceptual framework that came from the Analytic Theory. The proposed conceptual framework has been developed based on five phases and each phase includes important components. However, this framework is the first one who designed to apply a cloud data governance in terms of decision-making which make it unique and able to be developed more in future.

4. Big Data Governance Framework

- Specify Organization Type

In this step among many organizations specialty work with big data should be specified. The reason of determine the field is brining all the needed related stakeholders in order to highlight number of factors such as: roles and responsibilities between data governance teams. As well as define the guideline for establishing governance committees and meetings.

- Evaluate the Current Governance

This step is important to review and evaluate the current situation of data governance. Moreover, to check the ability of organization to do major and minor changes in their process to implement perfect governance. This step should be made by all related committees in order to ensure the quality and maturity of the evaluation and document all evaluation steps.

- Design Governance Framework

This step will be about functionality of governance. Functions are the activities of data that should be considered by committees during the designing step. The components of data governance include policies, procedures, roles and responsibilities, management plan and communications hierarchy.

- Review Framework and Agreement

After making the design all related committees should review the proposed framework and assess all governance components in order to sign it. The following signature on the framework documents will be made by managements to ensure this framework is commutable with mission and vision of their organization as well as securing sensitive data form being manipulated or analyzed by unauthorized person.

- Lunching Framework and Monitoring Performance

Last step of this framework is go live to set all roles and policies to big data. Management should assign teams to monitoring the performance of this framework and audits all processes related to data to ensure the efficiency of the framework. These teams should report frequently to management and this report must be reviewed by designing committees to enhance the framework and updates policies and procedures.

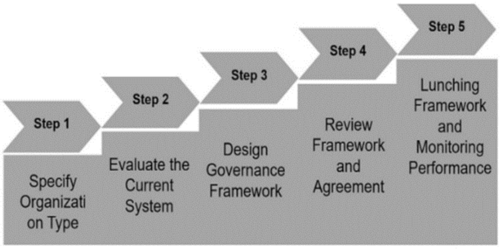
Step 1
Step 2
Step 3
Step 4
Step 5
Specify Organizati on Type
Evaluate the Current System
Design Governance Framework
Review Framework and Agreement
Lunching Framework and Monitoring Performance

Fig. 1 Major Steps in Big Data Governance

5. Conclusion and Future Work

Big data governance has been a trend topic in the last years, a lot of research focused on this topic and presented different results related to different issues of big data. However, due to variety of big data, there is a need for a framework that governs different activities. The framework proposed in this paper is focus in the organization type by allowing each category to establish their own governance model far from arbitrary picking any frameowk, which may affect their business wrongly. The future work will be applying this framework in some enterprises and companies in Saudi Arabia and observe the outcome to improve it.

References

[1] B. Sunil. Performance Measurement Metrics for IT Governance. ISACA Journal. 2016;6:21-7.

[2] J. Akoka and I. Wattiau. Evaluation of big data governance-combining a multi-criteria approach and systems theory. In 2019 IEEE World Congress on Services (SERVICES), volume 2642, pages 398–399. IEEE, 2019.

[3] M. Al-Ruithe and E. Benkhelifa. A conceptual framework for cloud data governance-driven decision making. In 2017 International Conference on the Frontiers and Advances in Data Science (FADS), pages 1–6. IEEE, 2017.

[4] M. Al-Ruithe, E. Benkhelifa, and K. Hameed. A conceptual framework for designing data governance for cloud computing. Procedia Computer Science, 94:160–167, 2016.

[5] M. Al-Ruithe, E. Benkhelifa, and K. Hameed. Data governance taxonomy: Cloud versus non-cloud. Sustainability, 10(1):95, 2018.

[6] G. Cheng, Y. Li, Z. Gao, and X. Liu. Cloud data governance maturity model. In 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pages 517–520. IEEE, 2017.

[7] A. Mauro, M. Greco, and M. Grimaldi. A formal definition of big data based on its essential features. Library Review, 65(3):122–135, 2016.

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

[9] C. N Johnson. The benefits fo pdca. Quality Progress, 35(5):120, 2002.

[10] R. KL Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang, and B. S. Lee. Trustcloud: A framework for accountability and trust in cloud computing. In 2011 IEEE World Congress on Services, pages 584–588. IEEE, 2011.

[11] S Kumar. Data governance: An approach to effective data management. White paper, Satyam Computer Services, Ltd, 2008.

[12] T. T. Lajara and A. C. Gastaud Mac ̧ada. Information governance framework: The defense manufacturing case study. 2013.

[13] D. Laney. 3d data management: Controlling data volume, velocity and variety. META group research note, 6(70):1, 2001.

[14] James Manyika. Big data: The next frontier for innovation, competition, and productivity. http://www. mckinsey.

com/Insights/MGI/Research/Technology and Innovation/Big data The next frontier for innovation, 2011.

[15] E. Niemi. Designing a data governance framework. In Proceedings of the IRIS Conference, At Oslo, Norway, volume 14, 2011.

[16] B. Otto. A morphology of the organisation of data governance. In ECIS, volume 20, page 1, 2011.

[17] M. Schroeck, R. Shockley, J. S., D. R. Morales, and P. Tufano. Analytics: The real-world use of big data. IBM Global Business Services, 12(2012):1–20, 2012.

[18] K. Weber,Boris Otto, and Hubert O. A contingency approach to data governance. Journal of Data and Information Quality (JDIQ), 1(1):4, 2009.

[19] K. Wende. A model for data governance-organising accountabilities for data quality management. ACIS 2007

Abdulrahman Albladi received his Bachelor degree from Computer Science Department at Umm Al-Qura University in 2012. He is currently a master student at the Computer Science Department. His research interests include healthcare systems and data science.

Eisa Alanazi received his B.Sc. degree in Information Systems from King Saud University, in 2007, and the MSc and PhD degree from the University of Regina, Canada in 2011 and 2017 respectively. He is currently an Assistant Professor at the Department of Computer Science, College of Computers and Information Systems in Umm Al-Qura University in Saudi Arabia. His research interests include preference learning and reasoning.