SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Hadoop in Enterprise Data Governance: Ensuring Compliance and Data Integrity

Govindaiah Simuni[1*] and Atla Amarnathreddy[2]

[1]Technology Manager, Bank of America, Charlotte, USA. E-mail: simunig@gmail.com
[2]Senior Technology Manager, Bank of America, Charlotte, USA. E-mail: atla.amarnathreddy@gmail.com

Article Info

Volume 4, Issue 2, November 2024
Received : 19 June 2024
Accepted : 07 October 2024
Published : 05 November 2024
doi: 10.51483/IJDSBDA.4.2.2024.71-78

Abstract

This paper examines how Hadoop can improve the data governance processes in enterprises with reference specifically to compliance and data consistency in the present climate of data-led business. Due to the ever-improving organizational technologies and processes that gather a lot of information, it has been deemed necessary for organizations to have structures that can follow when it comes to governance. To that end, the research focuses on Hadoop as one of the potent frameworks for analyzing big data and its readiness to facilitate imitable strategies in data management. Challenges for big data One Data Indonesia identified in this study able to provide insight and future work direction through two case analyses: the Presidential Decree No. 39/2019 on One Data Indonesia and analyzing the big data governance in cybersecurity; the common challenges include complexity in integration, non-compliance with the rule of law, and employee resistance to change. In addition, it also indicates ideas for future development, like automation, more intensive analytical tools utilization, and collaboration between the departments. The study demonstrated that by implementing Hadoop solutions and applying extensive governance approaches, any organization can manage the challenges of data management issues and blend them with legal requirements and perimeters to unlock the values and escalate organizational decision-making and performance.

Keywords: *Hadoop, Data governance, Compliance, Data integrity, Big data*

1. Introduction

In the current and ongoing technological advancements where enterprises deal with large amounts of data, both benefits and drawbacks are associated with the exponential growth of enterprise data. Companies and other organizations are creating new data, acquiring and processing data in quantities that in the past were

[*] Corresponding author: Govindaiah Simuni, Technology Manager, Bank of America, Charlotte, USA. E-mail: simunig@gmail.com

unimaginable, and that is why data governance is not only desirable but also necessary (Misran *et al.*, 2022). EDG is the practice, processes, policies, and tools used in managing enterprise information assets to ensure compliance with set policies and that the data is accurate, secure, and easily accessible. The implications of not being able to effectively manage data are dire and include things like data leaks, fines, and lost reputation (Yang *et al.*, 2019).

Hadoop is the leading sophisticated free software that has influenced enormously as the key tool in the field of big data management. It undergoes immense data processing, storage, and management and thus has become an indispensable component of many enterprises' technology platforms. In terms of data manipulation, organizations can increase data processing efficiency, check for data accuracy, and address the various compliances dictated by Hadoop's distributed computing environment (Padmavathi and Sudha, 2014). The Hadoop tool continues to become more valuable to organizations as regulatory activity expands worldwide (Alves-Araújo and Alves, 2013). This research aims to discuss Hadoop in the context of enterprise data governance with an emphasis on its relevance in compliance and data quality within today's diverse business environment.

1.1. Overview

It can be defined as an open-source distributed computational environment whose main objective is to manage the processing of large amounts of data across clusters of computers (Girija *et al.*, 2022). It is comprised of components such as HDFS for storage and MapReduce for computation that enable the handling of large amounts of structured and unstructured data. The architecture of Hadoop allows parallel application data processing, which allows it to solve the problem of the growth of data volumes generated in modern enterprises (Mahanti, 2021). Figure 1 explains the Data Governance Framework which includes Ownership, Accessibility, Knowledge, Quality, and Security.

The very presence of the data flow increases various governance issues. Data stewardship refers to the mechanism of regulating how data is managed, guarded from different risks, and thereby used in line with legal and regulatory requirements. With the GDPR, HIPAA, CCPA, and many other rules and standards in place that require much more control over data privacy and security, enterprises face the challenge not only of

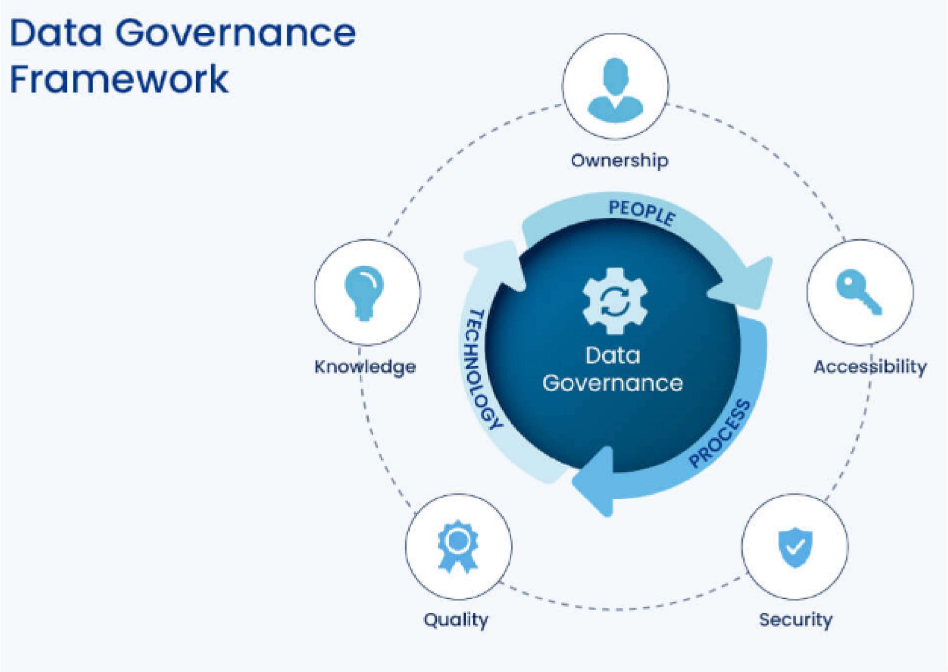

Figure 1: Data Governance Framework

Source: www.acceldata.io. https://www.acceldata.io/article/benefits-of-data-governance handling huge amounts of data but also of handling it in the right way and compliant with the law (Hariharan, 2022).

Hadoop is highly regarded for its EDG implementation because of its versatility and capacity. Adopting Hadoop into the enterprise data governance frameworks will enable compliance with these standards, data accuracy and integrity, and security (Wibowo and Sandikapura, 2019). This makes Hadoop an invaluable tool for enterprises that wish to solve two related problems—how to manage large volumes of data and how to store data that has become subject to ever-more-stringent regulations.

1.2. Importance

The role of Hadoop in enterprise data governance is mainly centered on the fact that you need a platform that can handle large volumes of structured and unstructured data. Data governance allows an organization to protect, manage, and utilize data effectively to maintain conformity to existing rules and regulations like GDPR or HIPAA. Hadoop offers the scalability and flexibility required to meet these governance challenges, improve data lineage, enhance data accuracy, and secure private information. Moreover, Hadoop's real-time processing capacity accelerates enterprises' capability to observe compliance and identify other aspects of governance rapidly. Figure 2 explains Data Governance Key Elements.

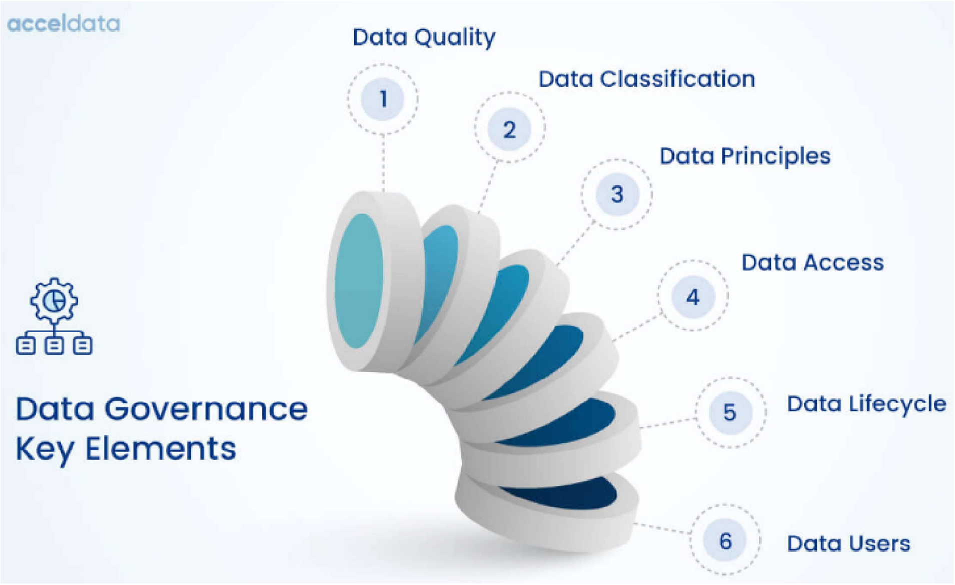

Figure 2: Data Governance Key Elements

Source: www.acceldata.io. https://www.acceldata.io/article/benefits-of-data-governance

1.3. Research Objectives and Scope

The purpose of this research is to understand Hadoop's role in defining enterprise data governance strategies by enforcing compliance with regulations and standards and by protecting data from any form of alteration.

1.3.1. Objectives

1. In order to understand how Hadoop facilitates data integrity in enterprise data management.
2. To review the barriers and constraints of Hadoop solutions for compliance initiatives.
3. Subsequently, the best practices for adopting Hadoop in enterprises' current data governance structures will be reviewed.
4. To know how Hadoop improves security and data privacy in extended enterprises.
5. This format will enable you to meet the necessities of the research topic and make it straightforward and definite.

1.4. Research Questions

6. How does Hadoop ensure data integrity in enterprise data governance?
7. What challenges arise when using Hadoop to comply with data governance?

2. Literature Review

Enterprise data governance has become more important over the recent past because of the technological growth and expansion of data across organizations. Data governance includes practices regarding the planning, collection, organization, usage, sharing, protection, and disposal of data (Suman, 2020). With organizations struggling to manage big data, frameworks such as Hadoop have become important structures that enable organizations to practice good data management. This literature review is organized around applying Hadoop in enterprise data governance and its ability to provide compliance and data control. Figure 3 explains ecosystem of hadoop.

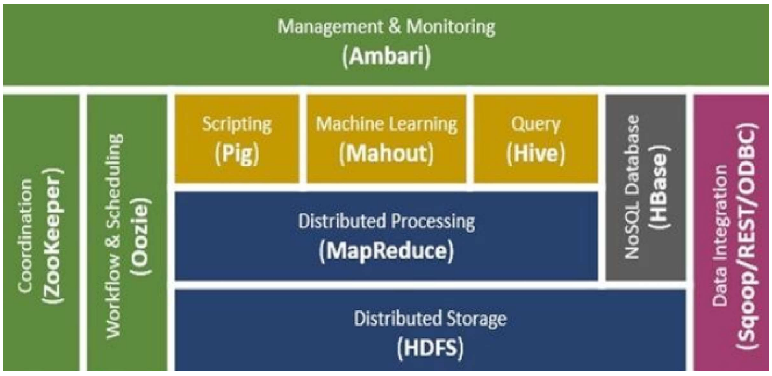

Figure 3: Ecosystem of Hadoop

Source: Researchgate.net (2024)

2.1. The Role of Hadoop in Data Governance

Hadoop is now a standard for processing and storing big data in different sectors across the globe. Its distributed structure makes it easy for organizations to store and process large amounts of data. By incorporating Hadoop within data information management frameworks, organizations can easily implement efficient means of complying with a given organization's regulations (Essakimuthu *et al.*, 2021).

2.2. Compliance and Regulatory Challenges

As the regulatory authorities pay more attention to the companies, the latter must strengthen the data governance framework to correspond to the requirements. Laws like GDPR and HIPAA demand that companies control the accessibility and utilization of information. Hadoop has tools that enable organizations to track data lineage and the ability to enforce regulation policies for compliance with such regulations.

The evolution in civil and industrial data regulations requires organizations to implement preventive measures for compliance issues. New legislation like GDPR and HIPAA means that data must be processed securely, and there must be proof that this is the case. Identifying and tackling these issues go hand in hand with Hadoop's data governance abilities since this platform helps organizations improve end-to-end data security and accountability (www.acceldata.io). From a Hadoop standpoint, security administrators can use Apache Ranger to define precise user access rules, while data stewards can use Apache Atlas to define data lineage and source. Such a level of visibility is important since it ensures compliance with policies, which are vital in responding to audits and inquiries from regulatory agencies in a manner that complies with the legal and regulatory requirements for data management. As such, Hadoop is a perfect solution for enterprises to adapt to the challenging context of data governance and compliance, including protective measures against potential legal consequences and ensuring the prevalence of accountability practices (Researchgate.net, 2024).

2.3. Ensuring Data Integrity

Accomplishing accurate, complete, and timely data is one of the critical components of data management since data quality should be proper in its use. Hadoop's architecture also shields data from data integrity issues by providing replication and other consistent methods to ensure quality data in an organization. Data integrity is critical as enterprises use Hadoop for an ever-growing range of big data processing activities.

Hadoop offers compelling benefits to organizations for data management, but it comes with issues that organizations encounter (Hariharan, 2022). This makes it hard to deploy Hadoop environments in a manner that adheres to well-defined corporate policies. Furthermore, corporates need to look for ways and means to fill the skill gap within the organization to unlock the full potential of Hadoop and, more importantly, align with governance frameworks already in place.

2.4. Best Practices for Hadoop Integration

Some of the best practices for organizations to get the maximum benefits of Hadoop in enterprise data governance include using automation tools in the Hadoop ecosystem. Apache Ranger and Apache Atlas should be integrated, which will help simplify the enforcement of governance policies, improve data lineage, and improve auditing. Traditional organizations can, therefore, gain from conceptualizations of big data management by offering a structured framework through which big data challenges and risks can be managed effectively.

3. Methodology

This research, therefore, employs a case study research method to assess the applicability of Hadoop and big data governance in increasing compliance and data quality in the enterprise. The strength of this approach is that two very different case studies are examined, and the practical lessons learned from applying the data governance framework and the issues encountered in the process are elucidated.

3.1. Case Study Analysis

Two distinct case studies are analyzed:

- **Case Study 1** (Misran *et al.*, 2022) **– The Presidential Decree on One Data Indonesia:** This paper explores the institutional context for information management envisioned by the Presidential Decree on One Data Indonesia to promote governance in both the center and the regions. A major focus of the study is to show how blockchain can improve data transmission and ownership as one of the solutions to managing large volumes of big data in city planning and taxation in regions.
- **Case Study 2** (Yang *et al.*, 2019) – **Big Data Governance in Cybersecurity:** The second case analysis focuses on applying a big data governance framework in organizations with a focus on cybersecurity. It shows that big data should be managed to provide quality access for machine learning while adhering to legal and ethical requirements. Therefore, this research outlines an organization's framework to make relevant decisions and ensure that data security, privacy, and trustworthiness are achieved.

The research will analyze the earlier cases to discover the positive experiences and problems associated with using Hadoop and big data technologies for data governance.

4. Findings and Discussion

The findings from the case study analysis reveal essential insights into the effectiveness of Hadoop and big data governance in enhancing compliance and ensuring data integrity across various organizational contexts.

4.1. Analysis of Case Study Results

- **Case Study 1** (Misran *et al.*, 2022) **– The Presidential Decree on One Data Indonesia:** This case-based analysis also highlights how 'rules' that fall under the broad category of regulations influence data regulation. The enforcement of the Presidential Decree on One Data Indonesia that the government adopted has forced organizations to have a proper way of managing data, hence improving governance. Blockchain technology has become a revolutionary factor, and integrating the technology has led to improved data security and the communicability of data. Blockchain enhances data lineage to improve efficiencies; therefore, meets the transparency and accountability needed to conform to the laws. However, the analysis also reveals some limitations, including privacy infringement and data governance, which must be dynamic to capture changing legislation. The process of managing big data has proven to be fruitful for urban planning and tax collection, and it shows how crucial data is for public sector activities.

- **Case Study 2** (Yang *et al.*, 2019) **– Big Data Governance in Cybersecurity:** The second case study focuses on the importance of big data guardianship frameworks in security situations. Companies that have implemented effective governance structures state they have experienced positive changes like data availability and decision-making. They show that good governance practices improve the ability of organizations to collect and manage data, which remains a critical enabler of successful machine learning and, by extension, advanced analytics. Also, using a stated framework, respect for the legal and ethical requirements is underlined, especially in protecting personal data. However, as with any change management, there are barriers that organizations experience when adopting such frameworks, such as resistance to change, issues related to incorporating new technologies in the organization, and the constant need to train the staff. The research evidence indicates clear benefits of a proactive approach to data governance, which supports greater information sharing and collaboration and increases an organization's security posture.

All in all, the two presented case studies also demonstrate how Hadoop and big data governance can improve compliance difficulties and data veracity. That is why, today, there is a need to manage data more effectively, and the broad adoption of new technologies like blockchain or the establishment of extensive governance systems is critical for organizations attempting to adapt to these new conditions. These case studies help create benchmarks that need to be followed to overcome common issues faced by organizations related to data governance and meet the regulatory requirements of different countries.

5. Challenges and Future Directions

There are some issues related to Hadoop deployment and adoption of big data governance strategies: integration is not an easy process; compliance standards may change over time; data privacy can be an issue; and people may resist such change. It means that organizations have to overcome these challenges in order to adequately handle and secure information. Future developments in data management must concentrate on the possibilities of using definition application tools, advanced analytical tools, cross-functional teamwork, strict adherence to decentralization, moral principles, and training sessions. With these issues, it is possible to identify new ways of effective data management, following organizational regulations, and maintaining data quality in a highly regulated environment.

5.1. Challenges in Implementing Data Governance Frameworks

- **Complexity of Integration:** Integrating Hadoop and big data governance involves compatibility issues and requires skilled personnel.
- **Regulatory Compliance:** Maintaining evolving regulations like GDPR and HIPAA demands ongoing updates and training.
- **Data Privacy Concerns:** Balancing data utilization with protecting sensitive information presents significant challenges.
- **Resistance to Change:** Cultural barriers may hinder the adoption of new technologies and governance practices.

5.2. Future Directions in Data Governance

- **Emphasis on Automation:** Utilizing automated tools for monitoring compliance and enforcing policies can enhance data management efficiency.
- **Adoption of Advanced Analytics:** Integrating machine learning can provide insights into data usage patterns and identify compliance risks.

- **Enhanced Collaboration Across Departments:** Fostering cross-functional teams can create a unified strategy for data governance.
- **Focus on Ethical Data Use:** Developing frameworks that address ethical concerns will ensure transparency and accountability.
- **Investing in Training and Awareness:** Providing training programs will equip employees with the knowledge to navigate new governance practices effectively.

Organizations can improve their data governance capabilities by addressing these challenges and pursuing future directions, ensuring compliance and maintaining data integrity in a complex digital landscape.

6. Conclusion

With the increase in the volume of generated data and new stiff legal regulations, data governance becomes critically important for organizations that must adhere to compliance requirements and simultaneously guarantee data quality. Hadoop and big data governance frameworks present a favorable platform to address issues emerging from the enhanced data complexity environment. The case studies discussed in this paper prove that integrating modern technologies, including blockchain, can increase data protection, improve data readability, and meet the requirements of various regulations, e.g., GDPR or HIPAA.

Nevertheless, organisations have some issues to solve, such as the difficulty of system integration, problems associated with constant changes in legislation, and staff resistance to change. This implies that solving these problems should be achieved with the help of active tools and methods based on automation and analytics, as well as cooperation between departments. Additionally, any organization that seeks to employ ethical data in its operations must be very committed to presenting a culture of ethical conformity by investing in its employees.

For years, data governance remained a baseline concept for many organizations; however, as organizations start focusing on the future, the following strategies should be prioritized to address the challenges ahead on the path that leads to success. Thus, enterprises satisfy compliance requirements and successfully implement the potential of data assets to create new values and solve various management and operational problems thanks to applying Hadoop and strict governance measures.

References

Alves-Araújo, A. and Alves, M. (2013). Checklist of Sapotaceae in Northeastern Brazil. *Check List*, 9(1), 59, February. doi: https://doi.org/10.15560/9.1.59

Essakimuthu, A., Karthik Ganesh, R., Santhana Krishnan, R. and Harold Robinson, Y. (2021). Enhanced Hadoop Distribution File System for Providing Solution to Big Data Challenges. *Intelligent Systems Reference Library*, 71-83. doi: https://doi.org/10.1007/978-3-030-57835-0_7

Girija Periyasamy, Rangaswamy, E. and Nawaz, N. (2022). Big Data Systems Architecture and Data Security Fundamentals—Case Study Approach for a Hospital in Singapore. *Lecture Notes in Networks and Systems*, 277-287, November. doi: https://doi.org/10.1007/978-3-031-17746-0_23

Hariharan Pappil Kothandapani, (2022). Optimizing Financial Data Governance for Improved Risk Management and Regulatory Reporting in Data Lakes. *International Journal of Applied Machine Learning and Computational Intelligence*, 12(4), 41-63, Accessed: Sep. 07, 2024. [Online]. Available: http://neuralslate.com/index.php/Machine-Learning-Computational-I/article/view/137

Mahanti, R. (2021). Data Governance and Data Management Functions and Initiatives. *Data Governance and Data Management*, 83-143. doi: https://doi.org/10.1007/978-981-16-3583-0_3

Misran, M. Syaifuddin, Nurmandi, A. and Khadafi, R. (2022). A Meta-Analysis of Big Data Security: Using Blockchain for One Data Governance, Case Study of Local Tax Big Data in Indonesia. *www.atlantis-press.com*, March 01. https://www.atlantis-press.com/proceedings/iconpo-21/125970961

Researchgate.net, (2024). https://www.researchgate.net/profile/Dharminder-Yadav/publication/341215709_IJETTCS-2017-07-14-17/links/5eb42eff299bf152d6a38b6e/IJETTCS-2017-07-14-17

Suman Shekhar, (2020). An In-Depth Analysis of Intelligent Data Migration Strategies from Oracle Relational Databases to Hadoop Ecosystems: Opportunities and Challenges. *International Journal of Applied Machine Learning and Computational Intelligence*, 10(2), 1-24, Accessed: Aug. 22, 2024. [Online]. Available: http://neuralslate.com/index.php/Machine-Learning-Computational-I/article/view/133

Vanka, Padmavathi and Sudha, T. (2014). Big Data Technologies: A Case Study. *Research Journal of Science and Technology*, 9(4), 639-642, Accessed: Oct. 01, 2024. [Online]. Available: https://www.indianjournals.com/ijor.aspx?target=ijor:rjst&volume=9&issue=4&article=025

Wibowo, S. and Sandikapura, T. (2019). Improving Data Security, Interoperability, and Veracity Using Blockchain for One Data Governance, Case Study of Local Tax Big Data. *International Conference on ICT for Smart Society (ICISS)*, November. doi: https://doi.org/10.1109/iciss48059.2019.8969805

www.acceldata.io. https://www.acceldata.io/article/benefits-of-data-governance

Yang, L., Li, J., Elisa, N., Prickett, T. and Chao, F. (2019). Towards Big Data Governance in Cybersecurity. *Data-Enabled Discovery and Applications*, 3(1), December. doi: https://doi.org/10.1007/s41688-019-0034-9

Cite this article as: Govindaiah Simuni and Atla Amarnathreddy (2024). Hadoop in Enterprise Data Governance: Ensuring Compliance and Data Integrity. *International Journal of Data Science and Big Data Analytics,* 4(2), 71-78. doi: 10.51483/IJDSBDA.4.2.2024.71-78.