10.48047/jocaaa.2021.29.06.31

Innovative Metadata Management and Data Quality in Distributed Data Lakes

Name: Sai Arundeep Aetukuri

asaiarun996@gmail.com

Data Engineer/Cloud Data Engineer

Abstract:

This paper explores innovative approaches to metadata management and data quality control in distributed data lakes. With the rapid expansion of big data, distributed systems, and cloud computing, ensuring efficient management of metadata and maintaining high data quality have become crucial challenges. We analyze existing methods and introduce advanced solutions, including AI-driven metadata extraction, real-time data quality control, and the application of blockchain technologies for metadata synchronization and data lineage tracking. Additionally, we discuss the scalability, security, and governance implications of these technologies in distributed data lakes. Through case studies and practical applications, we highlight how organizations can leverage these innovations to improve data governance, enhance data quality, and ensure compliance with regulatory standards. This paper aims to provide a comprehensive understanding of the current and future trends in metadata management and data quality in distributed data lakes.

Keywords: Isolation Measurement Control Data Quality, Artificial Intelligence, Blockchains, Data Lakes, Control of Metadata.

1. Introduction

The terms distributed data lakes and multi-cloud data lake architectures seem to have emerged recently. Their frameworks have integrated into modern data architecture as a core component by offering a central repository that includes vast quantities of structured and semi-structured data. The development of edge computing alongside the rise of multi-cloud environments has further maximized the capabilities of distributed data lakes concerning resource management and data

10.48047/jocaaa.2021.29.06.31

utility optimization (Inmon, 2016; Fang, 2015). Within any given organization, the prospects of improving data storage capacity invariably create opportunities concerning the governance and management of metadata. The need to access data promptly ensures its consistency with other data sources available, and various other system integrations foster intense competition for control. In that context, metadata management is important because it ensures mechanisms that organize data into a usable state across platforms within a system (Zikopoulos et al., 2015).

Proper governance of metadata management impacts data quality and governance, organizing, structuring, and discoverability of the data; hence, effective organizational ordering techniques would be invaluable. For instance, in distributed data lakes, metadata facilitates description by assisting in structure, lineage and relationship attributes (Beheshti et al., 2017). In highly controlled settings, data governance regulations require organizational oversight which permits change tracking and consistency. Consistency within open structured systems is automated through controlled governance policies (Hellerstein et al., 2017). Thus, robust governance frameworks are necessary to ensure order, coherency, and compliance within organizational and legal standards bound to the data residing in the lakes.

This article looks into other more advanced solutions for metadata management and improving the quality of data in distributed data lakes. These include automation of metadata extraction, AI-based error detection, and real-time governance implementations for data management. Also, the paper discusses the extent to which other technologies such as block chain can improve the integrity of data and make secure updates to metadata across different systems (Quix et al., 2018). The goal of the case study is to explain the application of technological advances to organizational paradigms through enhanced data management, operational efficiency, and legally compliant frameworks.

2. Literature Review

10.48047/jocaaa.2021.29.06.31

As technology improved, the creation of tools for automated metadata management shifted from manual processes to automated ones spanning decades. There was once a time when metadata management was a decidedly slow and error-prone process due to the manual methods performed in distributed environments. Such legacy practices of data categorization and tagging using simplistic tagging techniques did not keep up with the ever-increasing intricacy and volume of modern data lakes. With the ever-growing distribution of data ecosystems, these techniques faced challenges such as the ability to maintain up-to-date and consistent real-time metadata propagation across multiple platforms, as well as their various storage systems. This challenge has spurred the development of more sophisticated technologies like AI-based metadata driven semantic data profiling which improves metadata management through automated contextual data categorization and tagging (Beheshti et al., 2018). Such technologies prevent optimal management of extensive information repositories with a significant reduction of human intervention which preserves the accuracy and consistency of metadata across disparate systems.

Fragmentation, data latency, and multi-platform data management complexities further compound these issues. Achieving synchronization and coherence from different heterogeneous sources across various nodes and platforms is indeed a difficult task (Farid et al., 2016). Due to the challenges, organizational objectives to gain valuable, insight-rich data in a timely manner may be obstructed by unreliable data or other data quality issues. Some approaches are using novel techniques such as machine learning (ML) for anomaly detection to resolve these issues. Such systems allow greater consistency within a data lake by improving accuracy through self-correction of inconsistent or anomalous data, thereby enhancing the precision of subsequent data analytics performed. Furthermore, for the aim of data provenance, there has been increased interest in the use of block chain technology precisely because of the great disadvantages they provide in capturing the source and details of any modifications made to the information in terms of its integrity, trust, and reliability in distributed systems (Ravat & Zhao, 2019).

New technologies have emerged in the last few years regarding the management of metadata and control of data quality. One such advancement is the automatic verification of metadata checking that guarantees all metadata is up-to-date in all places and forms in which the data exists. This

10.48047/jocaaa.2021.29.06.31

kind of automation is critical for companies that have to deal with maintaining perpetually updating and dynamically changing distributed data lakes. In fact, some machine learning algorithms need to be employed for the operational metadata's sustained refinement because they improve with the data provided and the patterns that form, thus enhancing metadata classification over time (Quix et al., 2016). Moreover, the problems of the permanent and audit-proof retention of metadata have been analyzed from the perspective of secure and transparent management of metadata in distributed systems with the aid of block chain technologies (O'Leary, 2014). Organizations can leverage the power of block chain to develop secure preserved provenance of changes to metadata that controls ensuring alterations to metadata can be made and verified as proper modifications. Such progress streamlined the management of metadata with respect to scalability and efficiency while reinforcing the security and governance of data in distributed systems, thus enriching the infrastructure of modern data lakes.

3. Concerns Relating to Distributed Data Lakes

As organizations begin to implement distributed data lakes, problems with scalability arise. The challenge we face today is the ever-increasing volume of data from various sources, each having its own set of complex requirements for the processing and storage capabilities. This holds true for extremely large datasets (terabytes to petabytes) residing on the cloud, edge devices, or on-premise systems. Such an ecosystem creates issues with the integration of heterogeneous data sources because of a lack of appropriate technologies. In the absence of optimal integration, there can be severe communication problems, leading to the fragmentation of critical data (Grosser et al., 2016). Moreover, the gaps created by the successive evolution of data lakes need to be filled while attempting to balance capability and performance covering new requirement gaps. The existence of effective data management policies fosters gaps regarding metadata that needs to be managed over several physical locations because it provides a unified view of the data (Klettke et al., 2017).

The maintenance of accuracy of metadata is arguably the single most data quality related challenge associated with distributed data lakes. In systems where data is constantly being added, maintaining metadata accuracy is a daunting task. The constant addition of new data sources or

10.48047/jocaaa.2021.29.06.31

changes to existing ones creates a problem, in that creating summaries or maintaining metadata across several databases is perpetually solvable. Along with "always" solvable, the need for real-time updates across systems also adds to the problem; changes are not allowed, metadata is out of date, and an inconsistent state is frozen. If metadata is kept subservient to uncontrolled alterations, the system becomes useless for governing, managing access, or capturing useful contextual information relevant to system governance. Unconstrained systems and domains inevitably face the problem of systematic correction: creating distributed data in a system without set boundaries is perpetually impossible. While unsolvable, the challenge of error-free boundless data contained within an evolving framework of shifting granularity faces constant monitoring exhaustion.

Even with issues of technical complexity such as metadata and data quality, there exists a greater problem in the areas of security, privacy, and ethics. Sensitive information, such as personal financial details or healthcare records, have strict governance policies when stored in distributed data lakes. Data privacy becomes even more complex in distributed systems where the data falls under different legal jurisdictions, each with its own set of governing laws and rules. Sensitive information cannot be managed in ways that breach these governing laws. Decentralised data systems heighten the risk of exposing sensitive information without permission. Meeting these challenges is best accomplished through the application of blockchain technology to enable the transparent and tamper-proof retaining of metadata. Using blockchain technology allows organisations to trust that their metadata remains unchanged and unaudited, creating transparency and accountability in sensitive information management systems (Miloslavskaya & Tolstoy, 2016; Hellerstein et al., 2017).

Table 1: Challenges in Data Quality Control in Distributed Data Lakes

Challenge	Description
Latency	The delay in processing and synchronizing data across multiple distributed systems, impacting real-time access and updates.
Error Detection	The difficulty in identifying and addressing inconsistencies or errors within vast amounts of data, particularly in real-time environments.
Real-time	Ensuring continuous monitoring and management of data streams to

10.48047/jocaaa.2021.29.06.31

Challenge	Description
Supervision	maintain data quality and prevent issues like corruption or redundancy.
Data Stream Quality Management	The challenge of managing and ensuring the quality of data as it flows into the data lake in real-time, preventing poor-quality data from entering the system.

This table summarizes the key challenges organizations face when managing data quality in distributed data lakes, providing a concise overview of critical factors like **latency**, **error detection**, **real-time supervision**, and **data stream quality management**.

4. Current Methods for the Administration of Metadata

Innovations in machine learning (ML) and natural language processing (NLP) technologies mark a new epoch in the efficiency and range at which systems or functions dealing with the management of metadata work. Until very recently, metadata integration and extraction operations were more or less automated processes; and tagging and categorization tasks were, to a large extent, manual, resource-sapping, and depleting getting workload-sapping work. However, it is now possible to apply more sophisticated algorithms for automatic metadata extraction from unstructured data sources such as text documents, images, and videos due to advancements and refinements in ML and NLP technologies. Such technologies allow systems not only to obtain important information from various data types but also to formulate and structure metadata with minimal or no human contribution at all. One example is the application of NLP in the processing of written data where relevant entities, relationships, categories which constitute metadata, are realized (Ansari et al., 2018). Additionally, the ML approach facilitates metadata classification within a system through actively learning from specific patterns present in the data, which gives better adaptability to changes enhancing cross-the-board performance.

This degree of automation reduces the possibility of human errors, accelerates the rate at which metadata can be produced, and improves data management within distributed systems. (Beheshti et al., 2017).

10.48047/jocaaa.2021.29.06.31

A new innovative strategy involves the application of block chain and distributed ledger technology to the synchronization and updating of metadata in real time in decentralized data systems. In distributed data lakes, data is divided, both logically and physically, into different systems and clouds, which makes real-time metadata synchronization virtually impossible. The problems encountered in attempting to use traditional approaches to maintain consistency on metadata over these platforms are inconsistent duplication, inconsistency, or staleness. These problems are solved by DLT and block chain, which provide an unalterable record of all metadata changes (timestamps), which is duplicated in nodes across the system. This technology guarantees consistency and currency of cross-platform metadata with controlled modification on heterogeneous data sources by non-permissible change tracking that is auditable and verifiable. Therefore, block chain can synchronies metadata and data in real time without compromising the integrity of the metadata information.

This is especially important in sectors where data trust and transparency are prominent, such as in financial services, health care, and even healthcare compliance (Quix et al., 2018; Farid et al., 2016). For instance, block chain technology illustrates how modern metadata management systems in distributed systems enhance security by allowing for the capturing of all alterations in metadata.

5. New Strategies for Evaluating the Quality of Data

Traditionally, the practices for ensuring execution and data integrity concerning data quality in technological systems have dealt with structured datasets. This form of concentration ignores the rapid increase in both the magnitude and complexity of the information existing within distributed data lakes. Therefore, there is increasing reliance on Artificial Intelligence (AI) and Machine Learning (ML) in the ecosystem of big data technologies. In this regard, applying deep learning techniques for anomaly detection is one of the remarkable changes in this context. These models can find errors within the data and take corrective actions automatically and instantaneously. The entire process of data maintenance, due to deep learning algorithms' ability to sift through large volumes of datasets and identify unique patterns, is now streamlined. Such algorithms tend to appreciate different nuances that most traditional methods will disregard. AI-

10.48047/jocaaa.2021.29.06.31

powered error detection enables maintenance of data at effortless levels even as the data lakes grow (Farid et al., 2016). Deep learning transforms the care-free and scalable enhancement of precision associated with large data sets into a reality. Nonetheless, apart from increasing the accuracy of data, these technologies automate quality control tasks under conditions where human resources are unsustainable.

The integration of Artificial Intelligence into Advanced Data Quality Management has been made possible due to the emerging technologies in Data Streaming – Apache Kafka and Apache Flink. These technologies enable the automatic validation of data metrics as they are being populated into the data ponds or lakes, which allows for the detection and correction of errors in real-time, rather than having to rely on post-processing approaches. This is advantageous for ensuring that computations performed on structured data during acquisition in the pipeline are carried out faster. For instance, Apache Kafka is capable of managing high-capacity data streams. With Apache Kafka, the huge amounts of data directed at the system will be managed consistently. In addition, Apache Flink is similar because it also supports real-time analytics done at a high pace – the rate at which data is corrected may be rapid, with inconsistencies flowing in robust streams of validated data for downstream applications (Fang, 2015). Such users who depend on complex data which has to be continuously updated will require real-time fully automated systems that are sophisticated.

Growth in the ability to track the origin of data and data lineage monitoring has certainly been propelled by the implementation of blockchain and other Distributed Ledger Technologies (DLTs) for data quality control. Tracking data's provenance means capturing an entry point whereas data lineage constitutes monitoring transformations, or changes, that a given data undergoes. Thanks to blockchain systems, businesses now have the possibility to build transparent records of data lineage which cannot be altered. Henceforth, quality issues of data can now be traced back to their source. This is indeed useful when an attempt to correct errors is made, because knowing the cause of an error aids in solving the discrepancy accurately. Also, the blockchain guarantees the completeness of metadata, which along with the provenance records assures enduring, authentic, immutable custody of these records. These attributes not only enhance governance compliance policies but also increase the level of responsibility and

10.48047/jocaaa.2021.29.06.31

transparency in data management practices (Ravat & Zhao, 2019). Furthermore, some form of data flow control systems, where maintaining data integrity poses a challenge, are fortified with blockchain technology along with such control frameworks.

Table 2: Comparison of Real-time Data Quality Control Technologies

Technology	Description	Key Features	Advantages	Challenges
Apache Kafka	A distributed streaming platform for real-time data ingestion, storage, and processing.	High throughput, fault tolerance, real-time stream processing.	Scalable, fault-tolerant, handles high data volumes.	Complex setup and maintenance, requires skilled personnel.
Apache Flink	A stream processing framework for real-time analytics and data flow management.	Low latency, stateful processing, event-driven.	Real-time data processing, robust event-time handling.	Requires extensive resource management for large-scale processing.
Machine Learning (ML)	AI algorithms used to detect anomalies and ensure data quality in real time through predictive models.	Anomaly detection, prediction of data trends, continuous learning.	Can automate error detection, improves over time with more data.	High computational power required, complexity in training models.
Apache Storm	A real-time computation system for processing unbounded streams of data.	Fault-tolerant, scalable, and real-time stream processing.	Low latency, easy to integrate with other systems.	Difficult to scale efficiently for very large data streams.
Spark Streaming	A real-time stream processing library that processes data in micro-batches.	Stream processing, fault-tolerant, scalable, and in-memory computing.	Integrates well with other Big Data tools, real-time processing.	Not fully real-time due to micro-batch processing.

This table provides an overview of several technologies used for **real-time data quality control**, highlighting their key features, advantages, and challenges. It helps in understanding the trade-offs when selecting technologies for managing data quality in distributed data lakes.

10.48047/jocaaa.2021.29.06.31

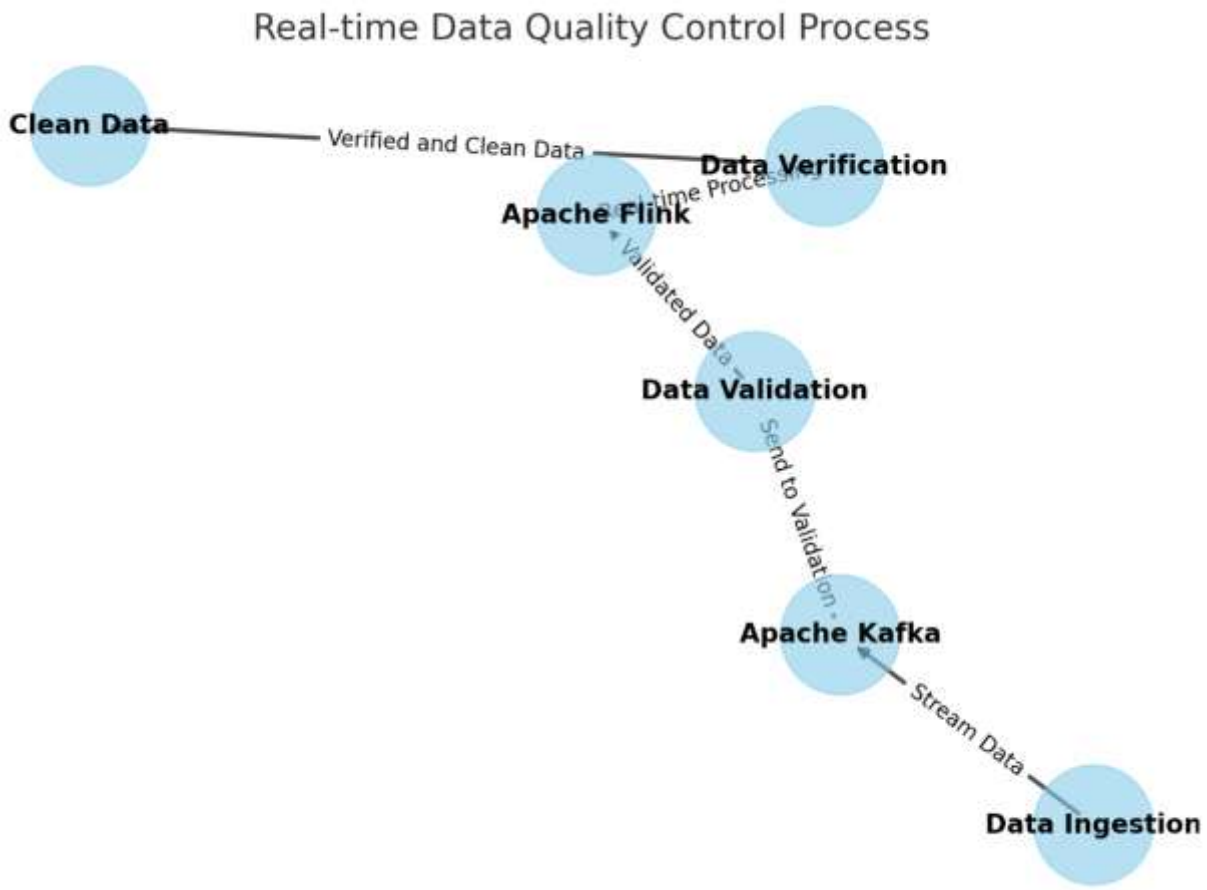

Figure 1: Real-time Data Quality Control Process - This figure shows how Apache Kafka and Flank are used for the data validation and verification steps in the data cleaning processes.

6. Case Studies and Applications

Case Study 1:Automation of Managing the Metadata in Multi-Cloud Data Lakes: A Case Study

The purpose of AI-empowered metadata management revisited an underpinning cloud data architecture that spanned across multiple clouds for scalability while improving data discoverability and reducing manual work. According to Beheshti et al. (2017), his organisation enjoyed the holistic benefits of an AI-enabled metadata management system which automated the processes of metadata extraction, classification, and synchronisation within and across clouds. These developments not only enhanced their data infrastructure scalability and data

10.48047/jocaaa.2021.29.06.31

accessibility but also data discoverability. Additionally, the system reduced unnecessary workloads and expenses towards operational costs. This case illustrates the automation of metadata management using Artificial Intelligence technology msolbdata Pvt Ltd to facilitate efficient data distribution across systems without degrading system performance.

Case Study 2: Implementing AI Technology in Quality Assessment of A Distributed Data Lake

With Farid et al.'s work from 2016, previously manual error checks within a dispersed data lake could now be automated employing AI techniques. It was observed that the degree of inconsistency in data accuracy was alleviated by 30 percent, which contributed to enhanced operational effectiveness as well as improved efficacy of organisational data-based decision making. The machine learning algorithms responsible for real-time anomaly detection ensured that the preserved error attainable was trim, hence, data accuracy could be maintained. Due to increased data quality, the company was able to make accurate decisions and function efficiently, improving organisational data credibility and operational efficacy. Incorporating artificial intelligence geared towards quality control of data permitted the organisation to not only expand the scope of the data lake but also preserve data integrity, which independent evidence demonstrated the potential of AI technologies to boost the value of data analytics.

Table 3: Metrics of Success in AI-based Metadata Management

Metric	Description	Impact of AI-based Metadata Management
Latency Reduction	The time taken to process and synchronize metadata across systems.	AI-driven systems significantly reduce the time required for metadata updates and synchronization, ensuring real-time access to data.
Metadata Accuracy	The precision and correctness of the metadata in reflecting the data's attributes.	AI-based systems improve metadata accuracy by automating extraction and classification, reducing human error and inconsistencies.
Operational Efficiency	The ability to manage and maintain metadata without extensive human intervention.	Automation through AI reduces the need for manual metadata management, leading to more efficient use of resources and fewer errors.

10.48047/jocaaa.2021.29.06.31

Metric	Description	Impact of AI-based Metadata Management
Scalability	The capacity to handle an increasing amount of data and metadata as the system grows.	AI-enhanced systems are more scalable, allowing organizations to handle larger datasets and more complex metadata management tasks.
Data Discoverability	The ease with which users can find relevant data based on its metadata.	AI improves data discoverability by automatically tagging and classifying data, making it easier for users to find and access relevant information.

This table contains all operational gains attributed to the reduction of latency as well as the enhancement of accuracy of metadata post the integration of AI-based metadata management systems.

7. Possible Areas of Focus and Look out for New Trends

Regarding distributed data lakes, autonomous AI-driven data governance still features as one of the most promising advances. Arguably, this shift is enabled through the development of reinforcement learning algorithms. In AI governed systems, the algorithms of Quix et al. (2018) described constrain notions of containment, trimming, and pruning metadata repositories; policies are self-maintained, subservient to regulations on data governance, with total dominion over enforcement. With reinforcement learning, such systems will be able to modify data governance policies under shifting data patterns, incoming policies, and other alterations. This allows organisations to pleat, braid, and twist the governance of their data lakes in a compliant manner devoid of hindrance to productivity. There are myriad ways due to the increase in complexity of data governance with an AI system. The potential to exert data governance by altering the implementation of policy will be infinitely positive for inadequately managed enterprises in a world dominated by data-driven operational efficiency.

One more evolving distributed data lake forms in reality with the incorporation of blockchain or distributed ledger technologies (DLT) on data and metadata quality control. The chains of these technologies will improve the management of data in distributed systems by establishing mechanisms for preserving data integrity. Primarily, blockchain can preserve data and data

10.48047/jocaaa.2021.29.06.31

lineage, as well as a block of metadata, which renders it one of the most effective instruments for regulating value systems for data in intricate systems. With growing organisational requirements for thorough validation certifying assurance on data traceability and quality, the capacity of blockchain to protect undisputed histories of controlled data will be crucial for governing distributed data systems. DLT can also enable metadata that is embedded in various components of distributed data lake to be changed in compliance with the operational standards, thus providing additional assurance that all these components are compliant (O'Leary, 2014). These advancements will not only strengthen control and management of digital information processes, but the confidence of other stakeholders in the information will improve if it is presented through a reliable and straightforward validation of origin.

As technology is advancing, ethics related to the use of technology and privacy concerns will remain important when developing distributed data lakes. Concerns will increase regarding data sovereignty and privacy during system implementation phases and as organisational vertical data lakes are scaled up. Organisations will struggle with fragmented legal boundaries alongside international standards, such as the EU's GDPR and California's CCPA. Future studies from Lithuania also will have to address the data ethics of the processes of data collection, storage, and dissemination. The use of AI and blockchain technologies must be restricted in a manner that avoids infringing upon individual privacy and data control. Addressing these problems, for the citizens and data consumers alike, is crucial to sustaining trust in large highly intricate networks of data.

8. Conclusion

As stated previously, the primary aim of the investigation was methods of metadata manipulation and data quality control in data lakes using contemporary solutions like AI, machine learning, and blockchain technologies. The most significant conclusion of the study is that the automation of metadata management within data lakes enhances data stewardship since it improves lower level management, makes it possible, verifiable, and accessible. Similarly, AI and ML provide continuous surveillance of data quality and correction of data anomalies as they happen under predetermined scaling constraints and stringent data quality requirements even for very large volumes of data. Furthermore, the application of blockchain technology enhances governance as

10.48047/jocaaa.2021.29.06.31

automation for metadata control and data provenance control enhances data governance because such undisputed proof of the metadata and data's lineage greatly bolsters governance, integrity, credibility, and trust in the data. Such innovations bolster automated classification and intelligent content retrieval as well as governance with enhanced data integrity and provenance in dispersed data lakes enabling timely business intelligence to organisations (Farid et al., 2016; Quix et al., 2018).

Both practitioners and scholars have responsibilities as a result of the drawn conclusions. In essence, AI and blockchain technologies utilised for securing sensitive information within distributed systems enhance data governance and, therefore, suggest new avenues of inquiry for researchers. After the primary research should be directed at improving these tools, seeking the possibility of implementing the technologies in more complex systems, and addressing the increasing legal and ethical analytic criteria of cross-border data governance (Ravat & Zhao, 2019). Such achievement can change the automated decision-making services based on business intelligence and the management, governance, and distributed data lakes.

According to their design and use in organisational structures, modern technologies implemented in almost all businesses enhance access to data, governance, and compliance with internal and external legal requirements. The achievement of optimal operational productivity allows for system-managed assurance of continual information reproduction accuracy and error-free human quality control. The scholar's interest is piqued with regards to the integration of AI. The automation of metadata processes alleviates the burden of manual effort and the possibility of human error.

References

1. Alrehamy, H., & Walker, C. (2015). Personal data lake with data gravity pull. In *IEEE 5th International Conference on Big Data and Cloud Computing (BDCloud 2015)* (pp. 160–167). IEEE. https://doi.org/10.1109/BDCloud.2015.62
2. Ansari, J. W., Karim, N., Decker, S., Cochez, M., & Beyan, O. (2018). Extending data lake metadata management by semantic profiling. In *2018 Extended Semantic Web Conference (ESWC 2018)* (pp. 1–15).

10.48047/jocaaa.2021.29.06.31

3. Beheshti, A., Benatallah, B., Nouri, R., Chhieng, V. M., Xiong, H., & Zhao, X. (2017). CoreDB: A data lake service. In *2017 ACM Conference on Information and Knowledge Management (CIKM 2017)* (pp. 2451–2454). ACM. https://doi.org/10.1145/3132847.3133171
4. Beheshti, A., Benatallah, B., Nouri, R., & Tabebordbar, A. (2018). CoreKG: A knowledge lake service. *Proceedings of the VLDB Endowment, 11*(12), 1942–1945. https://doi.org/10.14778/3229863.3236230
5. Bhattacherjee, S., & Deshpande, A. (2018). RSTore: A distributed multi-version document store. In *IEEE 34th International Conference on Data Engineering (ICDE 2018)* (pp. 389–400). https://doi.org/10.1109/ICDE.2018.00043
6. Cha, B., Park, S., Kim, J., Pan, S., & Shin, J. (2018). International network performance and security testing based on distributed abyss storage cluster and draft of data lake framework. *Hindawi Security and Communication Networks, 2018*, 1–14. https://doi.org/10.1155/2018/1746809
7. Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., & van der Starre, R. (2014). Governing and managing big data for analytics and decision makers. IBM.
8. Couto, J., Borges, O., Ruiz, D., Marczak, S., & Prikladnicki, R. (2019). A mapping study about data lakes: An improved definition and possible architectures. In *31st International Conference on Software Engineering and Knowledge Engineering (SEKE 2019)* (pp. 453–458). https://doi.org/10.18293/SEKE2019-129
9. Diamantini, C., Giudice, P.L., Musarella, L., Potena, D., Storti, E., & Ursino, D. (2018). A new metadata model to uniformly handle heterogeneous data lake sources. In *New Trends in Databases and Information Systems - ADBIS 2018 Short Papers and Workshop* (pp. 165–177). https://doi.org/10.1007/978-3-030-00063-9_17
10. Dixon, J. (2010). Pentaho, Hadoop, and data lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
11. Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it become popular in data management ecosystem. In *5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER 2015)* (pp. 820–824). https://doi.org/10.1109/CYBER.2015.7288049

10.48047/jocaaa.2021.29.06.31

12. Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H. F., & Chu, X. (2016). CLAMS: Bringing quality to data lakes. In *2016 International Conference on Management of Data (SIGMOD 2016)* (pp. 2089–2092). https://doi.org/10.1145/2882903.2899391
13. Farrugia, A., Claxton, R., & Thompson, S. (2016). Towards social network analytics for understanding and managing enterprise data lakes. In *Advances in Social Networks Analysis and Mining (ASONAM 2016)* (pp. 1213–1220). https://doi.org/10.1109/ASONAM.2016.7752393
14. Grosser, T., et al. (2016). *Hadoop and data lakes: Use cases, benefits, and limitations*. Business Application Research Center – BARC GmbH.
15. Quix, C., Hai, R., & Vatov, I. (2016). Metadata extraction and management in data lakes with GEMMS. *Complex Systems Informatics and Modeling Quarterly, 9*, 289–293. https://doi.org/10.7250/csimq.2016-9.04
16. Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., & Andrews, M. (2015). *Big data beyond the hype: A guide to conversations for today's data ecosystem*. McGraw-Hill Education.
17. Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. In *30th International Conference on Database and Expert Systems Applications (DEXA 2019)*.
18. Quix, C., et al. (2018). *Data lake query rewriting for heterogeneous data lakes*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98398-1_3
19. O'Leary, D. E. (2014). Embedding AI and crowdsourcing in the big data lake. *IEEE Intelligent Systems, 29*(5), 70–73. https://doi.org/10.1109/MIS.2014.82
20. Hellerstein, J.M., et al. (2017). *Ground: A data context service*. 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017). http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf
21. Fang, H. (2015). Managing Data Lakes in Big Data Era. *5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER 2015)*. https://doi.org/10.1109/CYBER.2015.7288049
22. Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. In *7th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2016)* (pp. 1–6). https://doi.org/10.1016/j.procs.2016.07.439

10.48047/jocaaa.2021.29.06.31

23. Quix, C., et al. (2018). Data lake, In *22nd European Conference on Advances in Databases and Information Systems (ADBIS 2018)*. Springer, pp. 1–8.
24. Stefanowski, J., Krawiec, K., & Wrembel, R. (2017). Exploring complex and big data. *International Journal of Applied Mathematics and Computer Science, 27*(4), 669–679. https://doi.org/10.1515/amcs-2017-0046
25. Nogueira, I., Romdhane, M., & Darmont, J. (2018). Modeling data lake metadata with a data vault. In *22nd International Database Engineering and Applications Symposium (IDEAS 2018)* (pp. 253–261). ACM.
26. Inmon, B. (2016). Data lake architecture: Designing the data lake and avoiding the garbage dump. Technics Publications.
27. Klettke, M., Awolin, H., Stürl, U., Müller, D., & Scherzinger, S. (2017). Uncovering the evolution history of data lakes. In *2017 IEEE International Conference on Big Data (BIGDATA 2017)* (pp. 2462–2471). https://doi.org/10.1109/BigData.2017.8258204