ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/girish-ganachari/228

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 1

Research Article

Data Governance for Enterprise Data Lakes

Girish Ganachari*

Citation: Ganachari G. Data Governance for Enterprise Data Lakes. *J Artif Intell Mach Learn & Data Sci 2022*, 1(1), 958-961. DOI: doi.org/10.51219/JAIMLD/girish-ganachari/228

Received: 02 July, 2022; **Accepted:** 18 July, 2022; **Published:** 20 July, 2022

***Corresponding author:** Girish Ganachari, USA, E-mail: girish.gie@gmail.com

ABSTRACT

Data lakes contain vast amounts of organised, semi-structured, and unstructured data in scalable and adaptive ways, making them crucial to corporate data management. This flexibility raises data governance difficulties. To guarantee data usability and integrity, this article covers metadata management, data quality, security, and compliance in organisational data lakes. This study examines data governance and data lake enhancements using secondary data from several academic sources.

Keywords: Data governance, Data lakes, Metadata management, Data quality, Data security, Compliance, Enterprise data management

1. Introduction

Data lakes can analyse and store large volumes of diverse data in scalable and adaptive repositories; therefore, enterprises are embracing them. Data lakes can store raw data and perform numerous analytical tasks, unlike data warehouses. However, adaption complicates data governance. Data must be reliable; thus data governance is essential. This framework encompasses metadata, data quality, legislation, and security. Real-time data quality solutions ensure data accuracy, while metadata management keeps data lakes clear of useless data. Maintaining customer trust and protecting sensitive data needs solid security policies and CCPA and GDPR compliance.

2. Metadata Management

2.1. Challenges in metadata management

Discoverability, quality, and governance need metadata management in data lakes. Traditional data management systems characterise data assets with metadata, simplifying discovery, interpretation, and usage. Because data lakes are unstructured and varied, metadata management strategies often fail[1]. Data lakes include text, multimedia, JSON, XML, and database data. Traditional metadata management methods for structured data struggle with many data contexts. Lack of metadata may quickly transform a data lake into a "data swamp," where low-quality and disorganised material collects junk[2]. Preventing this requires metadata. Understanding and using data requires metadata. Companies struggle to discover, understand, and trust repository data without it[3]. Insufficient metadata inhibits data governance, lifecycle management, and quality. It also makes it hard for data scientists and analysts to access and analyse information, which may cause mistakes and inefficiencies.

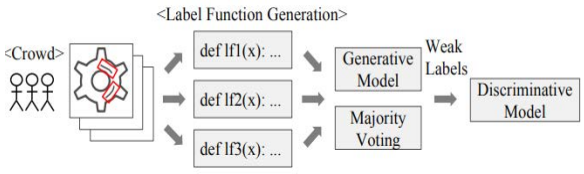

Figure 1: Workflow of data programming for a smart factory application.

2.2. Solutions and Approaches

These problems are solved by systems for managing metadata that are automated and based on artificial intelligence. Automatic metadata discovery solutions like Google's GOODS keep

data fresh and available without flooding the data lake[4]. These systems use machine learning to create and update metadata for new lake data[5]. This strategy increases data discovery and ensures data governance rules for all data assets[6]. As data is ingested, automatic metadata management systems may identify significant attributes, label it with relevant information, and organise it in a searchable catalogue. It enhances metadata input accuracy and minimises information management. These solutions may integrate with data governance systems to ensure metadata fulfils business needs. They handle data efficiently, comply with regulations, and protect data.

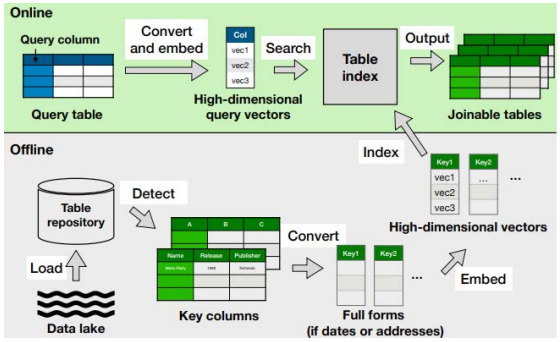

Figure 2: Joinable table discovery framework.

2.3. Case Study: CLAMS

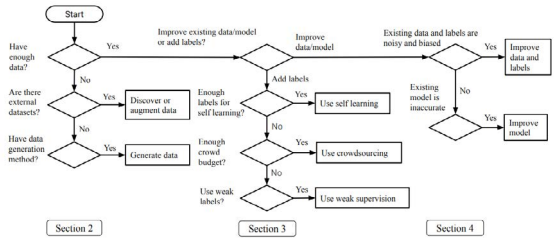

Figure 3: A decision flow chart for data collection.

An advanced data lake metadata management system is what the CLAMS system is all about. CLAMS identifies and enforces expressive integrity requirements to improve metadata management and data quality[7]. Quality criteria organise unorganised and partially organised facts into a model[8]. This arranges data for analysis and access independent of presentation. CLAMS finds data problems, recommends fixes, and updates metadata to fit the data lake's state using machine learning and other advanced technologies. Metadata management is needed to keep the data lake from becoming a swamp and preserve its decision-making capacity[9]. Applying past changes to fresh inputs enhances data quality in CLAMS. CLAMS demonstrates how advanced metadata management systems may handle data lake issues. Ensuring metadata is comprehensive, correct, and updated helps organisations maintain data quality, increase data governance, and maximise data asset value [10]. Data lake governance involves sophisticated metadata management to ensure data validity, organisation, and accessibility.

3. Data Quality

3.1. Ensuring data quality

Big data lakes with diverse data make data quality management challenging. These challenges are greater than for typical data warehouses. Data lakes may hold organised, semi-structured, and unstructured data. In contrast, data warehouses manage structured data using schemas[1]. This diversity requires extensive data quality control to ensure correctness and usefulness. Farid et al. (2016) recommend CLAMS (Cleaning, Labelling, and Managing Systems) to enforce quality standards and unify data[2]. CLAMS automated DQM systems increase lake data quality. Data quality management systems identify and fix issues in real time, ensuring the lake includes high-quality data[3]. Advancements in data accuracy and reliability aid analytics and decision-making[4].

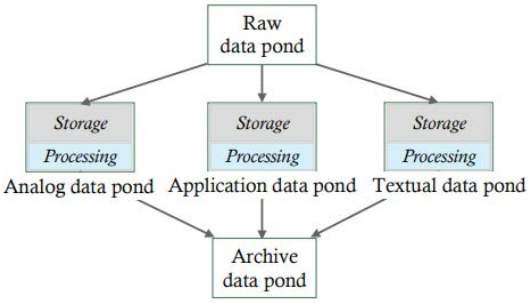

Figure 4: Data flow in a pond architecture.

3.2. Real-world application: Azure data lake store

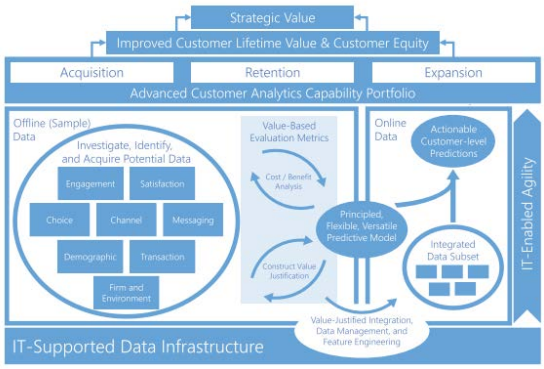

Figure 5: Advanced customer analytics framework.

Good data governance may help data lakes, as Azure Data Lake Store illustrates. This technology securely and reliably handles huge data sets. Data storage is enormous, and security and compliance are excellent[5]. Data governance in Azure Data Lake Store supports automated consistency and integrity checks. Data lake consistency is essential for diverse data[6]. Azure Data Lake Store automates data lake standards and quality control. Complete data governance guidelines show this. Enterprises get precise data and advanced analytics[7].

4. Security and Compliance

4.1. Importance of security and compliance

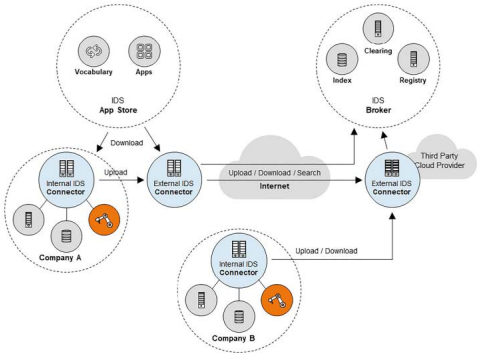

Figure 6: Initial IDS architecture design.

Big data lakes with sensitive data require control and security. Data lakes undermine data security and privacy due to constant

data creation and collection[1]. Addressing these issues requires secure data governance. These include robust access restrictions, encryption to protect data during transmission and storage, and audits to track sensitive data changes and access[2]. It is important to take heed of both the CCPA and the GDPR. Data must be managed securely to protect privacy[3]. Large, dynamic data lakes need constant security monitoring and adaptability. Changes are needed to address new threats and vulnerabilities. Continuous monitoring of illegal access and data breaches decreases loss[4].

4.2. Blockchain for data management

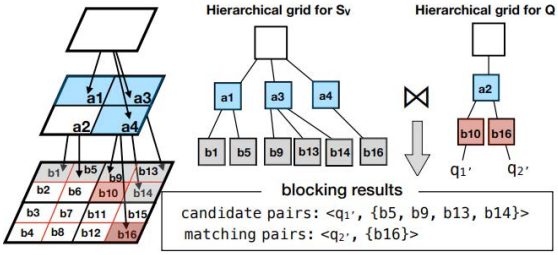

Figure 7: Hierarchical grids for target vectors and query vectors, and the blocking results of matching pairs and candidate pairs.

Blockchain enables the administration of data in a secure and transparent manner. Privacy and control in diabetic healthcare data management using blockchain[5]. Decentralised blockchains may increase data security by eliminating single points of failure. Prevent unauthorised access and manipulation. Companies securely share healthcare data using blockchain, satisfying CCPA and GDPR criteria. Hospitals must share data for complete treatment[7]. Blockchain's openness and immutability provide a reliable, impermeable transaction record for legal compliance. The blockchain meticulously records each transaction and access event, creating an immutable record that can be reviewed for business and regulatory compliance[8].

5. Opportunities in Data Governance

5.1. Automated metadata discovery

AI and machine learning automatically create and update metadata after importing new data. Maintaining data structure and accessibility avoids the lake from becoming a data swamp[24]. Automation that recognises and labels key information may facilitate searching and retrieving[25].

5.2. Real-Time data quality management

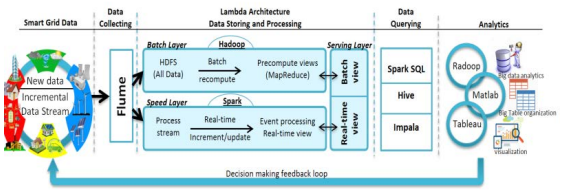

Figure 8: The smart grid big data eco-system to deal with the smart grid big data from data collecting to data analytics, with visualization and feedback loop capabilities.

Integrity and excellence need real-time data quality control. These approaches immediately find irregularities and inconsistencies to ensure processing-quality data[16]. Continuous data quality monitoring should enable firms to resolve issues rapidly to secure data[17].

5.3. Enhanced security protocols

Security measures are required in order to protect sensitive data that is stored in data lakes. Tight access restrictions, multi-factor authentication, and advanced encryption must secure data. Data protection must be stringent under the CCPA and GDPR[18]. These solutions protect data privacy and prevent unauthorised access[19].

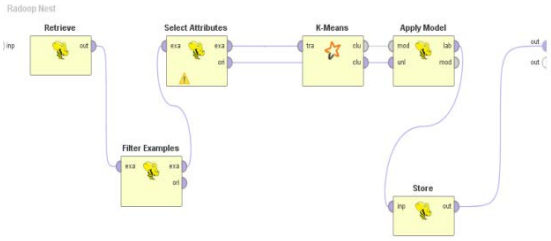

Figure 9: Radoop RapidMiner nest process to preprocess the data, apply the data mining clustering K-means algorithm and store the results into the HDFS repository

6. Case Studies and Applications

6.1. Efficient joinable table discovery

Dong et al. (2021)[10] discover joinable tables in data lakes using high-dimensional similarity. This boosts data lake value by making relevant datasets simpler to find and use[11]. Effective data integration helps companies understand their data[12].

6.2. Open data integration

Miller (2018) outlines the merits and downsides of integrating open data into data lakes for organising and leveraging public information[13]. Data integration efficiency may increase company data lake open data consumption and accessibility[14]. Organisations may use open data to enhance research[25].

6.3. Advanced customer analytics

Advanced consumer analytics combines massive data sets from several sources to provide strategic insights into client preferences and behaviour suggest relationship-oriented big data integration for consumer analytics[16]. Data lakes can store and analyse massive volumes of consumer data, improving customer satisfaction and marketing[17].

6.4. Geospatial data management

Data lakes should use geographic data management, according[22]. Geodata is saved, processed, and evaluated[25]. Data lakes manage vast volumes of geographical data with scale and flexibility for more accurate and thorough spatial analysis[19]. This method may enhance environmental monitoring, disaster management, and urban planning[18].

6.5. Big data in healthcare

Big data is helping to improve hospital operations as well as the care that patients get. Data lakes may handle healthcare data including real-time patient monitoring systems and electronic health records, according to[15]. Unifying healthcare data into a data lake may enhance patient outcomes and clinical decision-making[12].

6.6. Organizational performance and big data analytics

Aljumah et al. (2021)[24] say big data analytics influences organisational performance. To maximise big data analytics advantages, the paper emphasises data governance[23]. Data governance guarantees trustworthy, secure, and compliant data for data-driven decision-making that strengthens organisations[20].

7. Conclusion

Business data lakes need governance. Data, regulations, and quality are protected by good governance. Data lakes are scalable and adaptable; therefore, organisations will always use them, making data governance systems necessary. Data lakes will be crucial to business data management strategy as real-time data quality management, automated metadata discovery, and security improve. Organisations can optimise their data lakes and compete in the data-driven economy by addressing data governance issues and capturing opportunities.

8. References

1. Sawadogo P, Darmont J. On data lake architectures and metadata management. J Intelligent Inform Systems 2021;56: 97-120.
2. Farid M, Roatis A, Ilyas IF, Hoffmann HF, Chu X. CLAMS: bringing quality to data lakes. In Proceedings of the 2016 International Conference on Management of Data 2016; 2089-2092.
3. Nargesian F, Zhu E, Miller RJ, Pu KQ, Arocena PC. Data lake management: challenges and opportunities. Proceedings of the VLDB Endowment 2019;12: 1986-1989.
4. Munshi AA, Mohamed YARI. Data lake lambda architecture for smart grids big data analytics. IEEE Access 2018;6: 40463-40471.
5. Micheli, M., Ponti M, Craglia M, Berti Suman A. Emerging models of data governance in the age of datafication. Big Data & Society 2020;7.
6. Mehmood H, Gilman E, Cortes M, et al. Implementing big data lake for heterogeneous data sources. In 2019 IEEE35th international conference on data engineering workshops (icdew) 2019; 37-44.
7. Ramakrishnan R, Sridharan B, Douceur JR, et al. Azure data lake store: A hyperscale distributed file service for big data analytics. Proceedings of the 2017 ACM International Conference on Management of Data 2017; 51-63.
8. Bogatu A, Fernandes AA, Paton NW, Konstantinou N. Dataset discovery in data lakes. 2020 IEEE 36th international conference on data engineering 2020; 709-720.
9. Lake RW. Big Data, urban governance, and the ontological politics of hyperindividualism. Big Data Society 2017;4.
10. Dalle Mule L, Davenport TH. What's your data strategy. Harvard business review 2017;95: 112-121.
11. Dong Y, Takeoka K, Xiao C. Oyamada M. Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. 2021 IEEE 37th International Conference on Data Engineering 2021; 456-467.
12. Gomes VCF, Queiroz GR, Ferreira KR. An overview of platforms for big earth observation data management and analysis. Remote Sensing 2020;12: 253.
13. Cichosz SL, Stausholm MN, Kronborg T, Vestergaard P, Hejlesen O. How to use blockchain for diabetes health care data and access management: An operational concept. J diabetes science technology 2019;13: 248-253.
14. Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data-AI integration perspective. IEEE Transactions on Knowledge and Data Engineering 2019;33: 1328-1347.
15. Stair RM, Reynolds GW. Fundamentals of information systems. Cengage Learning 2018.
16. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: An overview. Briefings in bioinformatics 2021;22: 812-822.
17. Miller RJ. Open data integration. Proceedings of the VLDB Endowment 2018;11: 2130-2139.
18. Otto B, Jarke M. Designing a multi-sided data platform: findings from the International Data Spaces case. Electronic Markets 2019;29: 561-580.
19. Nocker M, Sena V. Big data and human resources management: The rise of talent analytics. Social Sciences 2019;8: 273.
20. Hellerstein JM, Sreekanti V, Gonzalez JE, et al. Ground: A data context service. CIDR 2017.
21. Kitchens B, Dobolyi D, Li J, Abbasi A. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. J Management Information Systems 2018;35: 540-574.
22. Breunig M, Bradley PE, Jahn M, et al. Geospatial data management research: Progress and future directions. ISPRS Int J Geo-Information 2020;9: 95.
23. Munawar HS, Qayyum S, Ullah F, Sepasgozar S. Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis. Big Data Cognitive Computing 2020;4: 4.
24. Aljumah AI, Nuseir MT, Alam MM. Organizational performance and capabilities to analyze big data: do the ambidexterity and business value of big data analytics matter? Business Process Management J 2021;27: 1088-1107.
25. Rialti R, Zollo L, Ferraris A, Alon I. Big data analytics capabilities and performance: Evidence from a moderated multi-mediation model. Technological Forecasting and Social Change 2019;149: 119781