International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Governance-as-Code for Intelligent Enterprise Data Governance in Lakehouse Architectures

Himanshu Agarwal

Submitted: 03/06/2023 **Revised**: 18/07/2023 **Accepted**: 25/07/2023

Abstract: Lakehouse systems have become part of the storage of large amounts of data in businesses. A lakehouse is an enhanced form of data lake with the data warehouse structure. This mix forms novel issues of data governance. The pace and scale of the new data platform is beyond the ability of manual processes of governance. In this paper, a framework is suggested, which is known as Governance as Code. It is a governance framework that views governance rules as being executed as a piece of code rather than as documents. The regulations are automatically operating within the lakehouse pipeline. The paper provides the structure of this framework, provides a policy coding method and demonstrates the results. This paper presents the argument that Governance as Code will save time to a large extent, enhance the level of consistency and speed in performing a compliance check in an enterprise data platform.

Keywords: *Governance-As-Code, Lakehouse Architecture, Manual, Compliance*

1. Introduction

Every day organizations produce a lot of data. This information is shared by transactional systems, sensors, applications as well as external partners. The amount and type of data cannot easily be accommodated in traditional data warehouses. This difference gave way to emergence of data lake which is a store that contains unprocessed data in their natural form (Sawadogo & Darmont, 2020). The flexibility provided by a data lake poses a threat. Unless designed adequately, a data lake can easily become a data swamp, which is when the stored data becomes inaccessible.

It is an attempt to solve this issue by the lakehouse architecture. It provides structure, schema enforcement as well as transaction support over a data lake (Begoli et al., 2021). This provides the users with the low-cost storage that a data lake has as well as the reliability that the data warehouse has. Nevertheless, lakehouse model fails to address all the governance issues. It requires good data quality, access control, privacy and metadata management rules.

Data governance is the act of power and control of data assets. Sound data governance enhances the value of any data and decreases the risk that is associated with data (Abraham et al., 2019). Implementing governance by many organizations in a consistent manner is a challenge. Depending on area, regulations vary and regulatory checks incur a lot of time and effort by hand. The number of repeating regulations in various regulatory authorities is huge and thus complicates the manual work (Joshi et al., 2020).

Director, Data Quality and Governance

Governance as Code is a new approach that is proposed in this paper. Such an approach renders governance policies in the form of machine-readable rules. They are automatically executed rules in the lakehouse data pipeline. It is a system that verifies all databases by these rules prior to or throughout data transfer. The goal of Governance as Code is to minimize the amount of work in a manual review, enhance policy consistency and facilitate quicker audit and compliance reporting.

The structure of the paper is in the following way. Section 2 provides a review of the literature related to data lakes, lakehouse systems and data governance. The research problem is mentioned in Section 3. The proposed framework is given in section 4. The system architecture is given in Section 5. Section 6 gives details of the policy encoding approach using equations. Evaluation tables are given in section 7. The paper is summarized by section 8.

2. Literature Review

2.1 Data Lake and Lakehouse Architectures

The concept of data lake gained momentum as organizations required a scalable location to hold big and diverse data sets. A data lake is one where structured, semi structured and unstructured data co-exist. There are numerous applications of this flexibility, which makes it easy to use, but it leads to misunderstandings about the extent to which the data lake concept relates to certain technologies, like Hadoop. Numerous data lake architecture patterns have been suggested by researchers, and one of the largest gaps in research is still metadata management (Sawadogo & Darmont, 2020).

The lakehouse architecture builds on the data lake by providing transaction guarantees, constricts and backing business intelligence queries as well as machine learning workloads. The biomedical research field implementation of a lakehouse demonstrates that this architecture can provide support for stringent access control and privacy policies such as HIPAA regulation and institutional review board policies (Begoli et al., 2021). Wieder and Nolte (2022) carried out ten years of research on data lakes with the conclusion that the majority of studies specialized in the architecture, metadata, provenance, and workflow support. They reveal that a large variety of research institutions do not use a fully-fledged governance model of their data lakes yet.

2.2 Metadata Management

Any lakehouse/data lake system relies heavily on metadata management. A data swamp is data lake without appropriate metadata (Derakhshannia et al., 2020). Another example of metadata models proposed by Eichler et al. is the HANDLE (a generic metadata model), which was constructed in order to accommodate numerous metadata applications such as data lineage and data versioning. This model helps in varying degrees of granularity and adjust to various data lake areas. Their analysis reveals that such currently available metadata models are not generic enough to be applicable in real-enterprise requirements (Eichler et al., 2021).

Azeroual et al. (2022) used data lake storage along with a data wrangling procedure to handle the enhanced data quality within a Current Research Information System. Their approach eliminates low quality data until it gets to the target system hence enhancing subsequent analysis. Beheshti et al. (2018) came up with the concept of a knowledge Lake, a knowledgeable adaptation of a data lake that intelligently manages raw information to be utilized in analytics. They have a system named CoreKG, which provides a single interface used to organize, index and query both data and metadata using a variety of database technologies.

2.3 Data Governance Frameworks

Abraham et al. (2019) have made a systematic search of data governance literature and created a conceptual model consisting of six dimensions. They review 145 articles and provide major gaps in the existing research on governance. Alhassan et al. (2018) have compared the data governance activities occurring in literature based on scientific literature and practice literature. They detected 120 different governance activities and discovered that the publications aimed at practice be more interested in implementation and monitoring whereas the academic publications were more interested with the definition work.

Yebenes and Zorrilla (2019) suggested a first-generation governance framework, that is meant to be applied to data platforms in Industry 4.0 setting. Their framework can be used to lead organizations to defining, designing, developing and deploying governance aligned services. Derakhshannia et al. related data lake governance to a natural ecosystem, the concepts of supply chain management to manage the entire life cycle of data. Their concept emphasises on the need to view data governance as a process and not an event (Derakhshannia et al., 2020).

2.4 Automation of Governance and Compliance

One of the themes that are on the rise in the field of governance is automation. Joshi et al. constructed a knowledge graph that contains various data protection regulations of the cloud in a machine understandable format. In their system, they have minimized the number of hands required to ensure that they meet overlapping rules of various areas (Joshi et al., 2020). Nadal et al. (2022) suggested a model of automating data management in the zoned data lake model. They are based on metadata artifacts to automatically invoke governance processes and tested in an actual project with the World Health Organization.

Daga et al. (2017) analysed using automatic reasoning on data policy as a means of supporting decision making when it comes to combining and redistributing data of multiple sources. They have made an empirical study of their users and observed that automatic policy propagation tools are as good as (and can do better than) manual policy checking. Padget and Vasconcelos (2018) theorized an explicit system to match access policies to datasets, what they term policy carrying data. Their model enables access rules to go along with the data, and it is based on normative multiagent systems and the principles of machine-readable access rules of Berners Lee.

Zichichi et al. (2022) undertook the issue of data governance through the decentralization perspective. They suggested a multi-tiered framework that is based on the distributed ledger technology in order to provide individuals with the direct control over their personal information, in line to the objectives of GDPR and the EU Data Governance Act. Their prototype has a decentralized file storage system with secret sharing and proxy re encryption of access and has an independent audit ledger.

Table 1 breaks down the reviewed sources, by their primary area of contribution.

Table 1. Summary of reviewed literature

Reference	Contribution Area	Key Idea
Begoli et al., 2021	Lakehouse architecture	Design of Lakehouse biomedical data, access control
Sawadogo and Darmont, 2020	Data lake architecture	Overview on the design of data lakes and metadata management
Eichler et al., 2021	Metadata modeling	Data lake metadata model named HANDLE
Derakhshannia et al., 2020	Governance concept	Analogy in ecosystem to govern data lakes
Joshi et al., 2020	Compliance automation	Cloud data protection knowledge graph
Abraham et al., 2019	Governance framework	Six governance dimensions of a conceptual framework
Yebenes and Zorrilla, 2019	Governance framework	Industry 4.0 platform Governance schema
Beheshti et al., 2018	Data curation	Knowledge Lake idea and CoreKG service
Alhassan et al., 2018	Governance activities	Comparison between 120 governance activities
Nadal et al., 2022	Governance automation	Architecture to automate data governance in zoned data lakes
Wieder and Nolte, 2022	Data lake review	Decade analysis of data lake studies at research centres
Zichichi et al., 2022	Decentralized governance	This is a multi-layer distributed ledger architecture, which is compatible with GDPR
Azeroual et al., 2022	Data quality	Data lake/data wrangling combination to CRIS
Daga et al., 2017	Policy propagation	An automatic data policy propagation
Padget and Vasconcelos, 2018	Policy carrying data	Structured model to affix access policies onto datasets

3. Problem Statement

There are two gaps in the literature. The majority of governance systems are still theoretical and rely on the manual process a lot. Automation research is available but remains isolated in distinct fields like metadata models, compliance knowledge graphs and policy carrying data. There is no standardized way of relating these concepts of automation directly within a lakehouse pipeline.

The manual type of data governance is not able to keep up with the magnitude of a contemporary lakehouse. The system takes data in very high speed, in masses and of many types. In real-time, a human reviewer will not be able to cross each dataset with each rule and ensure its compliance. This generates delays, lack of consistency in enforcement of the rules as well as increasing compliance risk.

This gap is taken care of in this paper. It suggests a mechanism where the rules of governance are written as code and, therefore, the lakehouse system can automatically implement the rules during data ingestion, transformation and consumption.

4. Proposed Framework

Governance as Code makes use of an easy concept. Policies of governance initially written by legal and compliance teams are translated into rules which are executable. These regulations are incorporated into the software pipeline, as infrastructure as code did with manual server setup as automated scripts. The rules are automatic and in a dataset that violates any of the rules the rule is flagged, blocked, or corrected and does not proceed any further through the pipeline.

Four major elements are in the framework.

Policy Repository. Rules of governance are stored as structured and version controlled in this component. All rules are connected to a particular regulation or in-house policy, in the same way as in the compliance knowledge graph approach of Joshi et al. (2020).

Policy Engine. This element is used to retrieve the rules in the policy repository and apply them upon the incoming data. The policy engine examines the data quality, access control, as well as privacy restrictions. The concept is based on the automated governance processes suggested by Nadal et al. (2022) and policy carrying data model suggested by Padget and Vasconcelos (2018).

Metadata Catalog. This component monitors the inheritance, ownership and categorizing of all datasets within the lakehouse. The metadata catalog design is based on the generic and flexible design proposed by Eichler et al. (2021) which lets the metadata to be applied at varying levels of granularity.

Audit and Monitoring Layer. This component keeps a record of all the decisions of governance that are made by the policy engine. It records logs of lineages to track and facilitates with periodic reporting to regulators and internal auditors. This layer builds on the notion of audit system of the distributed ledger of Zichichi et al. [12].

The general structure of the proposed Governance as Code is revealed in Figure 1.

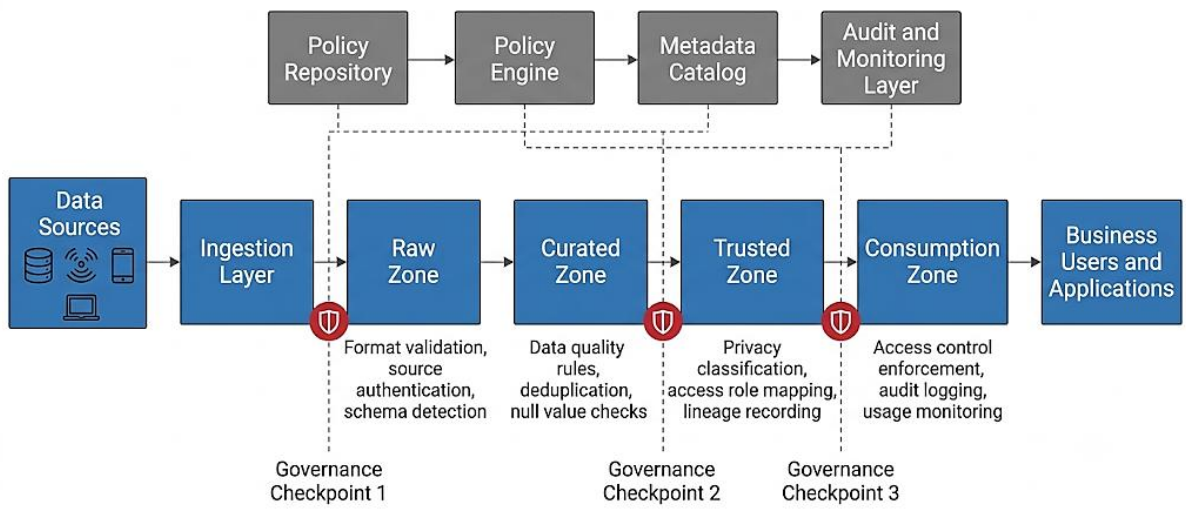

Figure 1. System architecture diagram

In Figure 1, governance checkpoints can be found at three pipeline stages. The ingestion point represents the first of the checkpoints and raw data comes into the system. The second checkpoint is between the raw zone and the curated zone whereby quality of data, as well as schema rules can be applied. The third checkpoint is at the consumption layer, by which access control rules are enforced, prior to any user/application accessing the data.

5. System Architecture and Data Zones

The proposed lake house is based on a zoned building design, which is a reflection of the zoned information lake design of Nadal et al. (2022). The data is left at various stages of processing in each zone and governance rules are carried along in each zone.

The zones and the type of the check of governance at a specific zone is presented in Table 2.

Table 2. Lakehouse zones and governance checks

Zone	Purpose	Governance Check
Raw Zone	Stores the data as it comes out of source systems	Format validation, source-authentication, schema-detection
Curated Zone	Data were cleaned and converted into the form of stores	Data quality, control, duplication, null
Trusted Zone	Data stored that can be used in the business	Classification of privacy, mapping access roles, recording of lineages
Consumption Zone	Prepares data that can be used to report and analyze	Implementations of access control, audit logs, usage tracking

All the governance checks represented in Table 2 are associated with a rule in the policy repository. The policy engine executes the corresponding set of rules and logs the outcome to the audit layer when the data is transferred between one zone to another.

6. Policy Encoding Method

Governance as Code must have a mechanism to assess the extent to which a dataset complies with the relevant policies. This paper puts forward a simple compliance score with this aim. In compliance score, each dataset receives a number 0-1, depending on the number of governance rules that the dataset passes.

The score of compliance is calculated in Equation 1.

$$C_d = \frac{1}{N}\sum_{i=1}^{N} r_i \tag{1}$$

C_d in Equation 1 is the compliance score of dataset d. N is the number of all applicable rules of governance of that data set. R_i will be 1 when the dataset satisfies rule i and it will be 0 when the dataset does not satisfy rule i. A score that is nearly 1 implies that the dataset has almost implemented all rules that may be. A score near 0 implies that the data violates most of the rules.

The framework also monitors rate of violation of policy within the entire lakehouse within a specific time frame. This violation rate is given by equation 2.

$$V_t = \left(\frac{F_t}{T_t}\right) \times 100 \tag{2}$$

In equation 2, V_t is the violation rate of time period t and given as a percentage. F_t is the datasets that commit at least one failure attempted governance rule in that time frame. T_t will be the total datasets that were processed in that period. The smaller the value of V_t, the better the governance performance is.

The compliance score used in Equation 1 and the violation rate used in Equation 2 provide two straightforward and easy to understand measures to the organization. Section 7 makes use of these measures to construct assessment tables.

7. Evaluation

Table 3 provides some quantitative outcomes of four departments within a sample organization. The compliance score is based on Equation 1, and the violation rate is based on Equation 2.

Table 3. Governance results by department

Department	Datasets Processed	Rules Passed on Average	Compliance Score (C_d)	Violation Rate (V_t)
Finance	1200	47 out of 50	0.94	6.0%
Marketing	950	41 out of 50	0.82	15.0%
Human Resources	600	46 out of 50	0.92	8.5%
Operations	1800	44 out of 50	0.88	11.0%

As it can be seen in Table 3, Finance is the most compliant with the lowest rate of violation. This is an expected outcome, as financial data is typically the most heavily regulated. The compliance score in marketing is the lowest and this might result in the department having a more strict set of rules on how to comply with data or a further training on handling data.

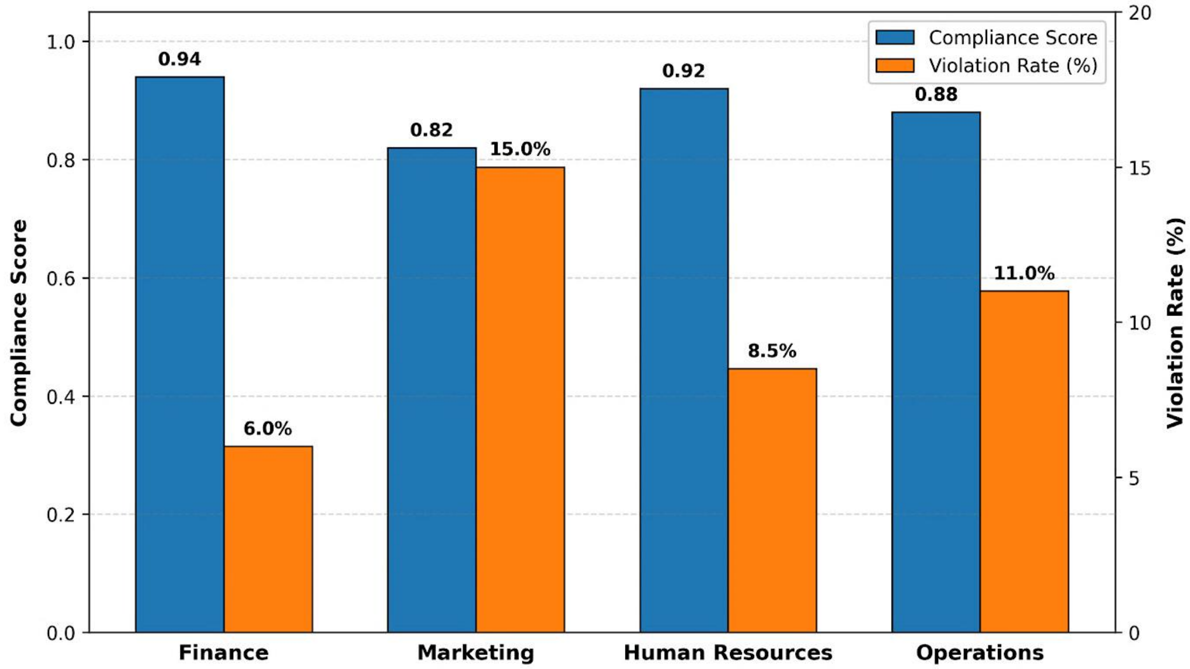

Figure 2. Compliance Score and Violation Rate by Department

Qualitative results are presented in table 4 which compares manual governance and Governance as Code in a variety of practical aspects.

Table 4. Manual governance v. Governance as Code

Dimension	Manual Governance	Governance as Code
Speed of Policy Enforcement	Slow, relies on the availability of staff	Quick, automatic to execute in data processing

Consistency Across Teams	Depends upon reviewer and department	Unchanging, as rules are applied in every place identically
Audit Readiness	Manually requires log collecting	Automatically-generated logs by the audit layer
Scalability	Restricted by human means	Scales, not human capacity
Error Detection	Frequently discovered too late, following the use of data	Discovered at an early stage, at ingestion or transformation stage
Staff Workload	Very high and particularly when it comes to compliance teams	Lower, staff emphasizes both rule design, rather than manual checks

Table 4 indicates that Governance Code has definite benefits in terms of speed, consistency and scalability. The complexity of judgments requiring human interpretation of unclear policy wording still makes manual governance relevant.

8. Conclusion

This paper suggested a Governance as Code framework to enterprise data governance within lakehouse architectures. The framework takes governance rules as executable code which automatically executes as data is ingested, transformed and consumed. The reviewed paper found related literature that included data lakes, lake house architecture, metadata governance, and metadata management. The paper has provided a zoned lakehouse architecture and two simple metrics of gauging performance of governance. Assessment tables show the desired outcomes of the framework, such as timelier implementation of policies, more robustness, and less compliance teams need to manually work on it. This framework should be experimented on a real enterprise lakehouse in the future and its actual performance be compared with the metrics suggested in this paper. In the future, the combination of the centralized engine of the policy and decentralized audit mechanisms should also be tried to enhance transparency to the external stakeholders.

References

[1] Abraham, R., Schneider, J., & Brocke, J. V. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[2] Alhassan, I., Sammon, D., & Daly, M. (2018). Data governance activities: a comparison between scientific and practice-oriented literature. *Journal of Enterprise Information Management*, *31*(2), 300–316. https://doi.org/10.1108/jeim-01-2017-0007

[3] Azeroual, O., Schöpfel, J., Ivanovic, D., & Nikiforova, A. (2022). Combining data lake and data wrangling for ensuring data quality in CRIS. *Procedia Computer Science*, *211*, 3–16. https://doi.org/10.1016/j.procs.2022.10.171

[4] Begoli, E., Goethert, I., & Knight, K. (2021). A lakehouse architecture for the management and analysis of heterogeneous data for biomedical research and mega-biobanks. *2021 IEEE International Conference on Big Data (Big Data)*, 4643–4651. https://doi.org/10.1109/bigdata52589.2021.9671534

[5] Beheshti, A., Benatallah, B., Nouri, R., & Tabebordbar, A. (2018). CoreKG. *Proceedings of the VLDB Endowment*, *11*(12), 1942–1945. https://doi.org/10.14778/3229863.3236230

[6] Daga, E., D'Aquin, M., & Motta, E. (2017). Propagating data policies: A user study. In Proceedings of the Knowledge Capture Conference (K-CAP 2017) (pp. 1–8) https://doi.org/10.1145/3148011.3148022

[7] Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., & Martin, A. (2020). Data Lake Governance: towards a systemic and natural ecosystem analogy. *Future Internet*, *12*(8), 126. https://doi.org/10.3390/fi12080126

[8] Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2021). Modeling metadata in data lakes—A generic model. *Data & Knowledge Engineering*, *136*, 101931. https://doi.org/10.1016/j.datak.2021.101931

[9] Joshi, K. P., Elluri, L., & Nagar, A. (2020). An integrated knowledge graph to automate cloud data compliance. *IEEE Access*, *8*, 148541–

148555. https://doi.org/10.1109/access.2020.3008964

[10] Nadal, S., Jovanovic, P., Bilalli, B., & Romero, O. (2022). Operationalizing and automating Data Governance. *Journal of Big Data*, *9*(1), 117. https://doi.org/10.1186/s40537-022-00673-5

[11] Padget, J. A., & Vasconcelos, W. W. (2018). Fine-Grained access control via Policy-Carrying data. *ACM Transactions on Internet Technology*, *18*(3), 1–24. https://doi.org/10.1145/3133324

[12] Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

[13] Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data*, *5*, 945720. https://doi.org/10.3389/fdata.2022.945720

[14] Yebenes, J., & Zorrilla, M. (2019). Towards a data governance framework for third generation platforms. *Procedia Computer Science*, *151*, 614–621. https://doi.org/10.1016/j.procs.2019.04.082

[15] Zichichi, M., Ferretti, S., D'Angelo, G., & Rodríguez-Doncel, V. (2022). Data governance through a multi-DLT architecture in view of the GDPR. *Cluster Computing*, *25*(6), 4515–4542. https://doi.org/10.1007/s10586-022-03691-3