

# Unified Data Governance Strategy for Enterprises

Ramla Suhra  
H-E-B Digital  
4217 Valley Oaks Dr  
Leander, TX, USA

## ABSTRACT

In the rapidly advancing fields of Data, Artificial Intelligence and Machine Learning, Data governance is essential for ensuring the ethical, effective and secure management of data. Many large data organizations face challenges in establishing effective data governance infrastructures to support advanced data uses, interoperability, robust data security, and high-quality data. This is especially acute when dealing with heterogeneous systems and the need to share data with numerous external organizations. To address these challenges, data organizations must adopt data governance frameworks that prioritize interoperability, data privacy, security, and data quality. This requires establishing standards not only for data structure, storage, and usage within organizations but also for how data is governed and circulated throughout the data ecosystem. This paper argues that all these cannot be achieved without a unified approach to information governance across organizations. Lack of coherence can lead to data inaccessibility, decreased efficiency, poor data quality, increased security risks, and compromised data integrity. To enhance adaptability, flexibility, and efficiency in data usage we propose a unified framework that integrates the principles of information, interoperability, security, and data quality governance. This paper explores the key elements of unified data governance, the challenges organizations face in implementing it, and best practices for success. Furthermore, we explore a proposed architecture for the unified governance framework and the role of different components in supporting this framework, including data cataloging, metadata management, access control management and data security monitoring and alerting. There is also a section which aims to further the discourse on what data governance looks like across the AI/ML data lifecycle.

## General Terms

AI, Data-catalogs, Data-governance, Data-security.

## Keywords

AI, Data-catalogs, Data-governance, Data-security, Unified, Machine Learning.

## 1. INTRODUCTION

As data volumes continue to grow, proper data governance becomes increasingly essential for large enterprises. With the proliferation of data sources, the growing complexity of data landscapes, and the increasing regulatory pressures, organizations must establish robust data governance frameworks to ensure data quality, security, compliance, and overall business value[1][2][9].

This includes managing and controlling all data stores within an organization, even those containing similar datasets across different systems due to multiple applications that generate or utilize them. This is particularly important in organizations that have adopted hybrid cloud as well as on-prem architectures.

To gain better visibility into all datasets, including where they are stored, how they are accessed, their lineage, computation cost and how they might be affected by new regulatory compliance rules, a unified governance approach has become essential.

Enterprise Unified Data Governance (EUDG) provides a comprehensive approach to managing data across an entire enterprise. It involves establishing policies, standards, and processes to govern the collection, storage, use, sharing, and protection of data assets. By centralizing data governance efforts, organizations can improve data quality, enhance decision-making, and mitigate risks.

EUDG offers a vital “bird’s-eye view” perspective of all data in an enterprise which effectively manage data across the disparate systems within an enterprise.

## 2. THE SIGNIFICANCE OF EUDG

Traditionally enforcement of data security governance often falls on IT teams to create solutions on-premises that exist in silos for each database or service. A unified approach to data security and access governance unites the needs of the organization’s data platform, information security, and data analytics teams on a single platform.

Adoption of unified approach offers many advantages as listed below [6].

### 2.1 Data Quality and Consistency

It ensures that data across the organization is accurate, consistent, and reliable, improving decision-making.

This can be achieved through standardizing definitions, building data quality metrics, assigning data ownership and stewardship.

### 2.2 Enhanced Security and Privacy

By implementing strong data governance practices, organizations can protect sensitive information and comply with regulations like GDPR and HIPAA.

This can safeguard and manage sensitive data across infrastructure, apps, and endpoints.

Digital Data Security teams in the data organizations can have a single pane of glass view of all data activities within the organization.

### 2.3 Reduce Redundant Data Management

Unified data catalogs can drastically reduce the data fragments and processes generating redundant data.

### 2.4 Improved Efficiency and Productivity

Unified data governance can streamline data management processes, reducing redundancy and increasing efficiency.

Some examples are consolidating or removing data silos,

automating data tasks, overall reducing cost.

## 2.5 AI Governance

As AI continues to advance rapidly, unified data governance becomes increasingly crucial.

Data driven ethical AI requires guidelines for access, sharing and availability of consistent standards within organization.

## 2.6 Better Business Insights

High-quality, accessible data enables organizations to gain valuable insights and make data-driven decisions.

## 2.7 Risk Mitigation

By establishing clear data policies and standards, organizations can mitigate risks associated with data breaches, misuse, and non-compliance.

## 2.8 Increased Trust

A well-governed data environment fosters trust among stakeholders, both internal and external.

## 2.9 Overall Cost Optimization

A well-governed data environment fosters trust among stakeholders, both internal and external.

## 2.10 Robust Activity Monitoring And Alerting

Proactive monitoring on data activities to uncover suspicious behaviors.

Helps monitor relational databases, data warehouses, big data and mainframes to generate real-time alerts on policy violations.

## 3. GOVERNANCE APPROACHES

Federated and unified governance are two distinct approaches to managing data within an organization. While they share the goal of ensuring data quality, consistency, and compliance, they differ in their implementation strategies. Upon comparing federated and unified data governance approaches, we have identified the advantages and drawbacks of each.

### 3.1 Federated Governance

In Federated data governance, each business unit or department within an organization has its own data governance policies and procedures. Given below are its advantages.

- *Highly customizable and quicker:* Adapts to the unique needs of different business units with quick turnaround related to policy and decision making.
- *Decentralized Decision-Making:* Empowers business units to make data-related decisions without much wait.

It also has the following disadvantages.

- *Inconsistent Policies:* May lead to inconsistencies in data definitions, quality standards, and security measures.
- *Data Silos:* Can create data silos and hinder data sharing across the organization.
- *Complexity:* Managing multiple governance frameworks can be complex and time-consuming.

Organizations at various maturity stages can adopt federated data governance. However, it is generally more suitable for organizations that:

- *Have a decentralized structure:* Federated governance aligns well with organizations that have a high degree of autonomy within their business units.

- *Are comfortable with risk:* Federated governance can introduce risks such as data inconsistencies and security breaches if not managed effectively.
- *Have established data management practices:* While federated governance can be adopted at different maturity levels, organizations with established data management practices may have a stronger foundation for success.

With Federated approach, organization risks data security, inaccurate insights, and delayed growth.

### 3.2 Unified Governance

In Unified data governance, a central data governance team establishes and enforces consistent policies and procedures across the entire organization. By unifying the data security governance, it promotes data access, privacy, and security governance across your entire organization. You enable your teams to create policies by using a common UI and enforce and execute them across your entire data estate. It has the following advantages.

- *Central view with Consistency:* Ensures consistent data definitions, quality standards, and security measures.
- *Data Sharing:* Facilitates data sharing and collaboration across different units within the organization.
- *Management Efficiency:* Streamlines data management processes and reduces redundancy.

Given below are the disadvantages of this approach.

- *Centralized Administration:* May limit the flexibility to address the specific needs of different business units.
- *Slower implementation:* Central administration can slow down decision-making.
- *Resistance to Change:* May face resistance from business units that are accustomed to having autonomy over their data.

## 4. IMPLEMENTATION CHALLENGES

Implementing unified data governance can be a complex undertaking, fraught with various challenges. Here are some of the key obstacle's organizations often face:

### 4.1 Data Silos[1] and Heterogeneity:

- *Diverse Systems:* Organizations often have a multitude of data systems, each with its own structure, format, and governance practices.
- *Data Isolation:* Data silos [1] can make it difficult to consolidate and manage data across different systems.

### 4.2 Lack of Data Ownership and Accountability:

- *Unclear Responsibilities:* Without clearly defined data ownership, it's challenging to enforce data quality standards and ensure compliance.
- *Accountability Gaps:* A lack of accountability can lead to inconsistencies and errors in data management.

### 4.3 Technical Challenges:

- *Data Integration:* Integrating data from various sources can be technically complex, especially when dealing with large datasets.
- *Data Quality:* Ensuring data accuracy, completeness, and consistency across different systems can be challenging.

#### 4.4 Organizational Resistance:

- *Cultural Barriers:* Implementing unified data governance may require a significant shift in organizational culture and mindset[2].
- It becomes impossible or hard to gain support of the stakeholders.

#### 4.5 Complex Regulatory Landscape:

- *Evolving Regulations:* Compliance requirements can be complex and change frequently, making it difficult to stay up to date.
- *Data Privacy Laws:* Regulations like GDPR and CCPA impose stringent data protection requirements.

Hence after reviewing all the above points, we have concluded potential benefits of unified data governance often outweigh some of the implementation challenges making it the preferred choice for data management.

### 5. STAKEHOLDERS AND REQUIREMENTS

The first step was to gather the set of requirements for data governance and interoperability based on the challenges faced by and the data needs of key stakeholders of data platform in enterprises. Major stakeholders are Data engineers, Data Scientists and Analysts, Business Users, IT and Data Security Department, Data Platform Administrators and Data Governance Team. This is shown in Table 1.

**Table 1. Stakeholder Responsibilities**

Type	Stakeholder	Requirements
Internal	Data engineers	Standardization: Seek clear standards for data modeling, metadata management, and data quality.
		Automation: Desire tools and processes to automate routine data management tasks.
		Integration: Require a framework that supports the integration of diverse data sources.
	Data Scientists and Analysts	Data Discoverability and Accessibility: Ability to discover datasets and easily request for access.
		Data Quality: Require reliable, accurate data for analysis and usage.
		Data Lineage: Desire information about the source system and impacted systems of the data set.
	Data Security Department	Security and Compliance adherence
		Desire a framework that helps mitigate risks associated with data breaches and misuse.
	Data Platform Administrators	Scalability, Easy of operation, Interoperability
	Data Governance Team	Clarity: Need well-defined roles, responsibilities, and decision-making processes.
Centralization: Needs a centralized system to define policies, review and govern data requests, usage, and fulfil all governance responsibilities.		
Enterprise Leaders	Usability: Need data presented in a user-friendly format that is easily understandable	
	Cost efficiency	
External	Customers	Privacy and Transparency: Demand that this model help their personal data to be handled securely and ethically across the enterprise
	External Partnering Orgs	Interoperability: Require compatible data standards and protocols for seamless data exchange.
	Investors	Desire a framework that helps mitigate risks associated with data breaches and misuse.

### 6. SYSTEM DESIGN

While building a unified enterprise platform for data governance here are some key design considerations to keep in mind.

#### 6.1 User Friendly UI

- The platform should be intuitive to use for all stakeholders (both technical and non-technical users).
- Search and discovery tools should be available for users to be able to look up data and controls quickly.

#### 6.2 Central Metadata Management [3] and Discovery

- Ensure the tool can capture metadata from various sources (e.g., databases, data warehouses, cloud platforms, data lakes etc.).
- Even when each of the systems can maintain their metadata separately, a central governance platform can help users collaborate by providing a central repository for metadata.

#### 6.3 Data Governance and Collaboration

- All teams should be able to collaborate using this tool to get access to the data sets that they are interested in an efficient and controlled manner.

- The tool therefore should have a workflow feature for reviewing and approving data access requests.
- Implement mechanisms to enforce data governance policies and standards.

### 6.4 Data Auditability

- Audit logs and dashboards are another key component in this platform which are crucial for investigation in the event of failures or data compromise.

### 6.5 Integration with Existing Tools

- Interoperability[4]: Design the tool to integrate seamlessly

with existing data management tools (e.g., ETL, data warehousing, BI).

- API Support: Provide well-documented APIs to enable programmatic access and integration with other systems.

### 6.6 Cost effectiveness and manageability

- The platform's implementation should not lead to a spike in infrastructure costs.
- It should be manageable using infra as code with less manual maintenance overheads.

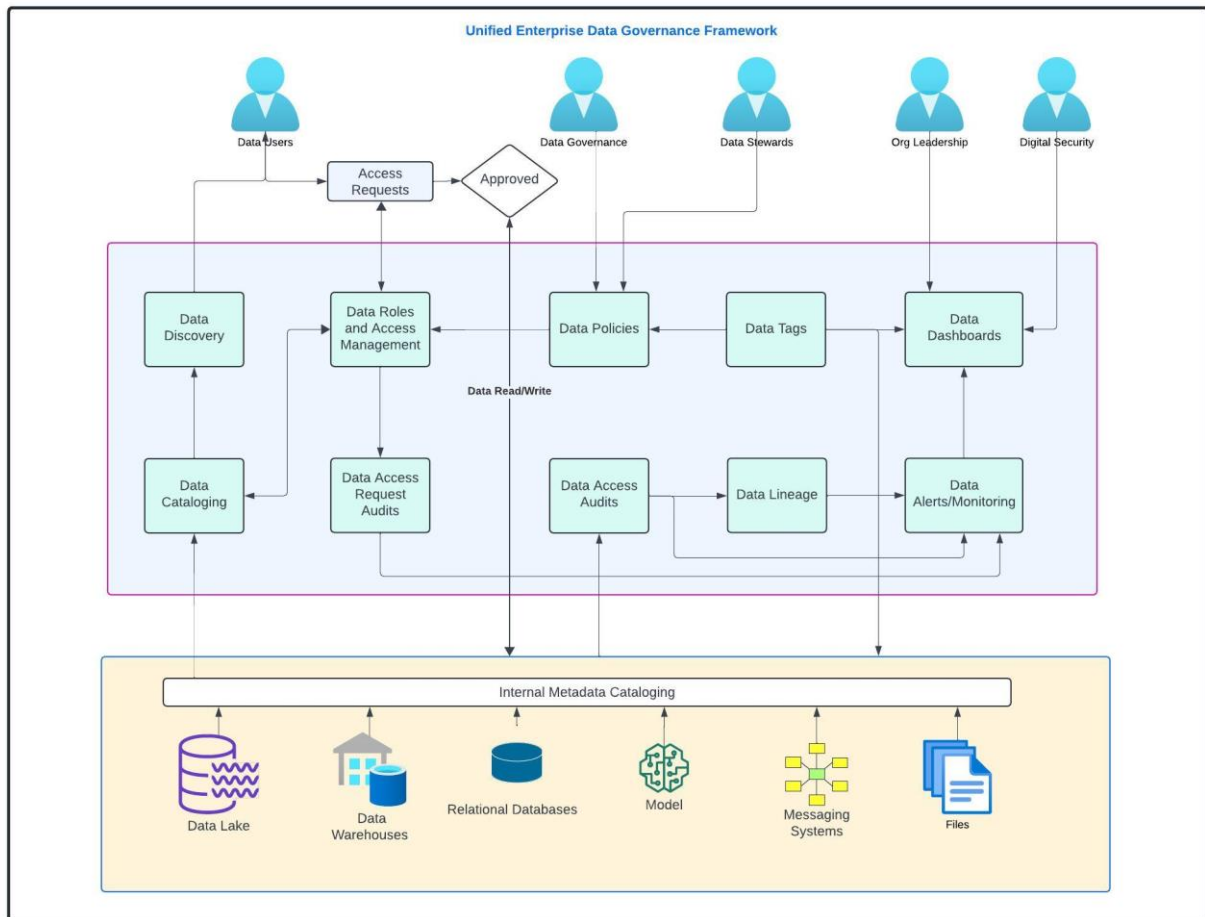


Fig 1: Platform Architecture Diagram

## 7. PROPOSED ARCHITECTURE

The architecture shown in Figure 1 focuses on the key components to be included in the governance framework while trying to integrate with disparate systems within an organization.

Unified Governance Framework will help standardize and centrally govern data assets and the related key functionalities like management of roles, policies, metadata definitions or tags, data ownership assignment/delegation. Data policies once defined can be applied across all datasets conforming to those rules. These policies help define what quality data looks like for the organization.

Central metadata definitions and cataloging helps with better data organization thereby uncovering and eventually reducing data redundancy and silo issues. Another critical component in

this framework is the data access control module with data access request and approval workflows embedded in it.

To ensure effective data governance, it is essential to implement various data access techniques tailored to different user needs and roles. Role-based access controls (RBAC) is a widely used method in which access permissions are assigned based on roles within an organization. Each role is assigned specific permissions through which users only access data that they are authorized to. For instance, data stewards and data engineers might have access to raw data for preprocessing and anonymization, while data scientists, ML engineers and analysts access production data tagged or labelled for their use cases like analytics, model training and development.

On the other hand, attribute-based access controls (ABAC) allow access based on a user's attributes, such as their role, project title, or data sensitivity levels. ABAC provides more

granular control than RBAC. An example is restricting access to certain datasets based on the sensitivity of its data, thereby enhancing security and compliance.

RBAC and ABAC can be achieved by a combination of standards like tagging or labelling data assets, setting up data role and policies. Centralizing standards for naming, labeling, and documenting datasets are hence key to this framework, and they can also significantly reduce redundancy and inconsistencies.

Data lineage capture can help us understand downstream impact of a data set alteration. As data changes, it can have a significant trickle-down effect. It provides a clear picture of data usage, which helps analyze duplicate reports, deprecated (old) data, and other processes so that computing and people resources can be used more efficiently.

Effective data quality governance practices during the downstream stages are critical for the development of reliable AI models. Ensuring data integrity involves continuous monitoring and validation of the data used for training and evaluation. This includes implementing robust data validation frameworks that can detect and correct anomalies in real time. Organizations can leverage this central platform for managing data spread across different centers in various formats like relational databases, hybrid cloud, flat files and streaming sources[4][7].

To top this all, a unified approach can act as one stop solution

for all data access monitoring/auditing and alerting. Data dashboards can be easily presented to leadership or any other stakeholders through EUDG framework.

## 8. SIGNIFICANCE OF EUDG

AI data governance[5][6][9] is a specialized subset of data governance that focuses on the ethical, legal, and technical aspects of managing data used in artificial intelligence (AI) systems. It ensures that AI systems are developed and deployed in a responsible and transparent manner, aligning with organizational values and ethical standards. Artificial Intelligence technologies leverage huge data sets to generate predictions through Machine learning such as Large Language Models (LLM), Deep Learning, Supervised and Unsupervised learning.

While these are great transformations which will help make better data driven decisions, their results, performance and ethical build rely heavily on the governance of the underlying data. Since AI or ML Models often with highly sensitive data, a more centralized approach might be preferable to ensure strong security and compliance[5][6][9].

Approaching AI governance[5][6][9] as a federated model might give more flexibility to teams, however it can be challenging to maintain consistency across different data governance practices on a longer run. Thus, it is better to embrace unified governance for AI data governance. Unified Data Governance can address the challenges (shown in Table 2) with respect to AI data governance [5][6][9].

**Table 2. Platform Challenges**

Challenges	Details	Opportunity
Data Quality Issues	Data from diverse sources if not controlled with often cause consistency issues within AI models.	Data standardization from Unified approach can help mitigate such issues.
	Data evolution over a period can cause Model behavior drift.	This can reduce the risk of inconsistencies and errors in AI models.
Model bias and fairness	Defining and measuring fairness in AI is complex and often requires domain-specific knowledge.	Unified standardization of data definitions can centralize the domain knowledge and reduce these challenges.
Data Security Issues	Since AI deals with data at all sensitivity levels, it can lead to unauthorized data access and breaches.	Data Access control component solves this concern.
Model scalability Issues and Central metadata cataloging for AI models	With out proper knowledge of data access patterns, it becomes difficult to tune AI models to perform efficiently.	Unified standardization of data definitions can centralize the domain knowledge and reduce these challenges.
	Just like data, models also become unmanageable over a course of time.	Same as above
Accountability	Establishing accountability for the actions and outcomes of AI systems.	Managing ownership and permissions on models can solve this

## 9. FUTURE ENHANCEMENTS

To improve the effectiveness of the model, we can incorporate some of the following future state enhancements:

- Leverage machine learning to detect and alert on abnormal and potentially risky activity based on user access patterns captured via logs.
- Automated description generation for datasets based on the naming standards and columns standards and labels.

- Automated data discovery to identify secure data sets.
- Automated labelling or tagging and automatic push down of tags to its lineage datasets.
- Regular data quality assessments and auditing procedures should be incorporated into formal governance processes. As problems emerge, they can be addressed at the root cause through updated policies and standards.

- Set up an ongoing responsibility to monitor compliance, analyze where quality is breaking down and recommend improvements.

## 10. CONCLUSION

In conclusion, by implementing Unified Data Governance framework organizations can achieve greater operational efficiency, improved decision-making and regulatory compliance, while positioning their business for sustained growth and success in an increasingly data-driven world.

It also emphasizes the crucial role of a holistic data governance approach in AI. By ensuring that data governance frameworks are robust and comprehensive, we can develop AI/ML systems that are not only technologically advanced but also ethically sound and aligned with societal values.

## 11. REFERENCES

- [1] C. Foster, "What is a Data Silo and Why is It Bad for Your Organization?," Plixer, Jul. 27, 2018. <https://www.plixer.com/blog/data-silo-what-is-it-why-is-it-bad/> (accessed Oct. 16, 2024).
- [2] "Understand Data Governance Trends & Strategies | Gartner," Gartner, 2024. <https://www.gartner.com/en/data-analytics/topics/data-governance> (accessed Oct. 16, 2024).
- [3] N. Pedersen, "Future of Master Data Management: Trends in 2023-2025," Stibosystems.com, Sep. 2024. <https://www.stibosystems.com/blog/the-next-frontier-of-master-data-management-and-the-trends-that-are-driving-it> (accessed Oct. 15, 2024).
- [4] Tumulak, J. Tin, and K. Keshavjee, "Towards a unified framework for information and interoperability governance," *Studies in Health Technology and Informatics*, Feb. 2024, doi: 10.3233/shti231310.
- [5] Open Data Institute, T. Carey-Wilson, E. Simperl, A. Das, and I. Reklós, "Understanding data governance in AI: Mapping governance," Open Data Institute, Jun. 2024. [Online]. Available: [https://theodi.cdn.ngo/media/documents/Understanding\\_data\\_governance\\_in\\_AI\\_Mapping\\_governance.pdf](https://theodi.cdn.ngo/media/documents/Understanding_data_governance_in_AI_Mapping_governance.pdf)
- [6] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li, "Data Management for Machine Learning: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, Feb. 2022, doi: <https://doi.org/10.1109/tkde.2022.3148237>.
- [7] Rahimi, Farhoud Jafari Kaleibar, Fatemehzahra Feizi, Amirhossein Hossein Nia, and Houman Kashfi, "Navigating Data Governance in the Telecom Industry," Nov. 2023, doi: <https://doi.org/10.1109/icaea60387.2023.10414472>.
- [8] "DGI Data Governance Framework - The Data Governance Institute," The Data Governance Institute, May 15, 2024. <https://datagovernance.com/the-dgi-data-governance-framework/> (accessed Oct. 15, 2024).
- [9] Y. Jernite et al., "Data governance in the age of Large-Scale Data-Driven Language Technology," *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2022, doi: 10.1145/3531146.3534637