The Forgotten Middle Layer: Metadata-Driven Governance For Semi-Structured Big Data

Uday Kumar Reddy Gangula

ukgangula@gmail.com

***Abstract*:**

Big data technologies enable advanced analytical capabilities, yet they produce numerous project failures because of a fundamental architectural deficiency. The paper reveals a "forgotten middle layer" in contemporary data systems, which arises from the implementation of schema-on-read data lakes in conjunction with semi-structured formats, such as JSON and XML. The complexity of modern data systems exceeds the capabilities of traditional governance models, which results in data becoming both untrustworthy and unusable. The proposed metadata-driven governance framework utilizes active metadata management to deliver data discovery, lineage, and quality, as well as policy enforcement capabilities. The paper demonstrates architectural patterns for this method while showing its implementation in Apache Atlas platforms and developing a step-by-step adoption framework. Metadata governance establishes structure and trust in semi-structured data environments, connecting raw infrastructure with effective analytics.

Index Terms: Big Data, Data Governance, Metadata Management, Semi-structured Data, Data Lake, Apache Atlas, Data Lineage, Schema Inference.

I. INTRODUCTION

The significant data era promised organizations transformative benefits in business intelligence, operational efficiency, and competitive advantage. Organizations started gathering large amounts of data from transactional systems, social media platforms, mobile devices, and Internet of Things (IoT) devices. The rapidly growing quantities, speeds, and types of data have created significant opportunities for analysis. Many initiatives have failed to produce meaningful business results from their significant investments in Hadoop and Spark technology.

The primary reason for this failure emerges from the incompatibility between data governance methods and new architectural frameworks. The shift from traditional schema-on-write data warehouses to schema-on-read data lakes created a governance void [1]. Data lakes excel at inexpensive data ingestion and storage of various datasets; however, they often delay critical operations such as integration, quality assurance, and semantic definition, which creates a gap between raw data infrastructure and analytical applications.

The governance gap highlights a strategic dilemma that arises when organizations opt for technology-focused approaches without proper business alignment or establish unattainable expectations without understanding technical limitations. Research indicates that cultural elements and organizational factors are the primary challenges hindering the success of big data.

The paper suggests that a "forgotten middle layer" exists between data storage/processing platforms (e.g., HDFS, Spark) and consumption tools (e.g., BI platforms, data science notebooks). The middle layer needs to be established as a metadata-driven governance framework that provides data discoverability, trust, and semantic clarity. The paper analyzes governance problems in contemporary architecture, establishing metadata as the foundation for this middle layer and developing an operational framework for its deployment.

II. THE GOVERNANCE GAP IN MODERN DATA ARCHITECTURES

Modern data platforms have become ineffective for traditional data governance because they employ schema-on-read processing and support the ingestion of unstructured data, as well as a diverse range of data types. The powerful architectural changes disrupt the strict control systems that previously maintained data quality, consistency, and usability.

A. Architectural Shift: From Data Warehouses to Data Lakes

Enterprise data architecture has transitioned from schema-on-write data warehouses to schema-on-read data lakes. The requirement for data cleansing and transformation into predefined schemas in traditional warehouses maintains quality standards but restricts both agility and scalability [1].

Data lakes store unprocessed heterogeneous data without needing an initial schema definition. The cost-efficient flexible data lake approach eliminates the quality control mechanisms that ETL processes in data warehouses used to enforce. The delay of critical governance functions, such as integration, standardization, and validation, until the consumption period results in an increasing distance between infrastructure and analytics capabilities. A governance model must be developed to manage unstructured data in rapidly changing environments.

TABLE I- **COMPARISON OF DATA WAREHOUSES AND DATA LAKES**

Characteristic	**Data Warehouse**	**Data Lake**
Data	Structured, Processed	Structured, Semi-structured, Unstructured, Raw
Processing	Schema-on-Write	Schema-on-Read
Schema	Well-defined, pre-existing	No predefined schema
Storage	Expensive for large volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, easily reconfigured
Usage	Well-defined, operational reporting	Future, experimental, exploratory analytics
Data Quality	Clean, trusted, single source of truth	Raw requires end-user integration and validation
Governance	Mature, centralized control	Maturing requires new approaches
Skills	Heavy IT reliance (DBAs, ETL developers)	Self-service, more technical analysts (Data Scientists)

B. Emergence of Semi-Structured Data

Data lakes are increasingly populated by semi-structured formats such as JSON and XML. These formats include internal markers (e.g., keys, tags) that allow partial schema inference at read time, providing both flexibility and minimal structure [2], [3].

Key attributes of semi-structured data include:

- **Flexible Schema:** Fields and objects can evolve without redefining the data model [4].

- **Self-Describing:** Embedded schema markers enable machine parsing and human interpretation [3].
- **Hierarchical Structure:** Supports nesting for complex relationships beyond relational tables.

Such properties make JSON and XML the default formats for modern sources, such as APIs, IoT feeds, and logs—the core drivers of the big data surge [5]. However, their structural ambiguity complicates governance.

C. Incompatibility of Traditional Governance Models

Conventional governance frameworks operate best in structured, controlled environments; however, they struggle to handle the distributed, schema-agnostic nature of data lakes. Key limitations include:

1) **Loss of Central Control:** Traditional governance systems depend on ETL processes that operate from centralized locations to enforce quality standards and compliance [1]. The schema-on-read model bypasses these control points, allowing raw, unvetted data to accumulate.
2) **Manual Scalability Breakdown:** The rapid growth of big data makes manual governance practices impossible to maintain at scale. The tracking of lineage and classification activities usually relies on error-prone spreadsheets as their default method [6].
3) **Data Swamps:** Data lakes transform into “data swamps” when proper governance is absent because they store undocumented, unreliable data. Users lack semantic clarity in understanding field meanings because they cannot determine whether an ID field represents a customer, product, or transaction. The lack of clarity in the field means that across billions of records creates trust issues and inconsistent analytics results.

The current governance models have proven insufficient because they require a new system to handle semi-structured data at scale and rebuild trust in big data systems.

III. METADATA: THE FOUNDATION OF THE MIDDLE LAYER

Modern data architectures require metadata to establish connections between their governance systems. Organizations should focus their efforts on managing the descriptive metadata that represents their data, rather than attempting to govern raw data directly. A well-designed metadata strategy creates trust while making data discoverable and usable throughout the entire data ecosystem.

TABLE II- KEY FUNCTIONS OF A METADATA-DRIVEN GOVERNANCE PLATFORM 16

Function	Description	Business Value
Data Cataloging & Discovery	Create a searchable inventory of all data assets, enriched with technical, business, and operational metadata.	Reduces the time data analysts spend finding relevant data and prevents redundant data engineering work.
Data Lineage & Provenance	Automatically track the origin, movement, and transformation of data across all systems, from source to consumption.	Builds trust in data; enables impact analysis for changes; simplifies regulatory auditing and root cause analysis.
Business Glossary & Semantics	Provide a centralized repository for business terms, definitions, and rules, linking them to physical data assets.	Ensures a consistent understanding and use of data across the organization, eliminating ambiguity in reporting.

Data Stewardship & Curation	Assign ownership and stewardship responsibilities for data assets; facilitate collaborative annotation and curation.	Improves data quality and accountability; captures tribal knowledge and makes it accessible to all users.
Policy Management & Security	Define and manage data access, security, and privacy policies based on metadata tags and classifications.	Enables consistent policy enforcement across heterogeneous systems, helping to manage compliance with regulations such as GDPR.

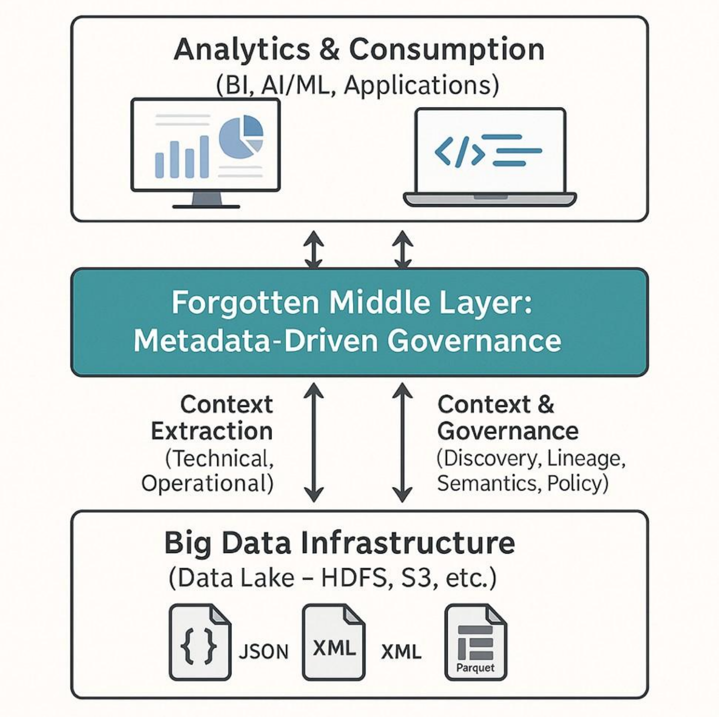

Fig. 1. Three-tier architecture highlighting the often-missed metadata-driven governance layer connecting raw data infrastructure to analytics and consumption

A. Defining Modern Metadata

Modern architecture uses metadata as an operational architectural element that extends its definition beyond "data about data" [7]. Raw data requires this context to achieve effective interpretation and utilization. The three main categories of metadata exist as follows:

- **Technical Metadata**: The structural properties of schema definitions, data types, file formats, and physical locations are described by this metadata type. The ingestion process of semi-structured environments reveals only this type of extractable metadata.
- **Operational Metadata**: The system tracks data lifecycle dynamics through lineage and transformation logic, job logs, access patterns, and performance metrics. The system tracks data production processes and data movement paths [8].
- **Business Metadata**: The business context of data includes glossary terms and definitions, ownership and stewardship roles, quality indicators, and regulatory classifications (e.g., PII) [8].

B. Pillars of Metadata-Driven Governance

The integrated metadata repository provides governance functions that support schema-on-read architecture.

The middle layer depends on the following pillars to function as a metadata-driven system:

- **Data Discovery**: The system allows users to find data assets by technical attributes, business terms, or usage patterns, which speeds up insights and decreases tribal knowledge dependency.
- **Data Lineage**: The system provides complete end-to-end data flow tracking, which enables audit trail impact analysis and regulatory compliance [7].
- **Data Quality and Semantics**: The system links technical raw fields to business definitions to maintain consistency between reports and models. The analytics output increases through metadata support of quality scoring and semantic disambiguation.

These functions collectively address the governance challenges of flexible, large-scale data environments. The governance platform requires the essential capabilities, as presented in Table II.

IV. ARCHITECTURAL PATTERNS FOR METADATA-DRIVEN GOVERNANCE

The conceptual "middle layer" has transformed into specific architectural patterns that both open-source and commercial platforms use. The industry has reached an agreement on the fundamental components that support scalable, metadata-driven governance systems.

A. Central Metadata Repository Pattern

The main architectural pattern relies on a centralized metadata repository, which collects and unifies metadata from Hadoop clusters, relational databases, streaming platforms, and BI tools. The unified metadata repository addresses data silo issues by providing a single interface for discovery, lineage, and governance functions.

B. Core Architectural Components

Modern metadata platforms share several architectural features:

- **Ingestion Framework**: The system collects metadata from different sources through real-time hooks and batch crawlers for data in motion [9].
- **Flexible Type System**: The system enables users to create custom data models that handle different assets (Hive tables, Kafka topics, BI dashboards) and their connections.
- **Graph Storage Backend**: The system uses graph databases (e.g., Janus Graph) to store metadata entities, which enables efficient relationship and lineage query capabilities [8].
- **Search Index**: The system uses full-text indexing technologies (e.g., Solr, Elasticsearch) to provide quick metadata discovery that operates independently from the graph layer [10].
- **API and Messaging Layer**: The system enables metadata interaction through RESTful APIs, which support both user interfaces and tools. The Apache Kafka event-driven architecture separates producers from consumers to provide real-time metadata updates, which support policy enforcement [11].

C. Automated Schema Inference

The inference of schema becomes crucial for both JSON and Parquet semi-structured formats. The process of deriving structural metadata occurs through dataset analysis, without requiring predefined schemas [12]. Advanced platforms use machine learning to improve this process by handling ambiguous and variable data more effectively. The automation process plays an essential role in enabling scalable governance operations for schema-on-read environments.

D. Early Metadata Platforms

Early metadata governance tools demonstrate the adoption of these patterns through their implementation.

- **Apache Atlas**: The open-source reference architecture combines flexible type systems with storage capabilities from Janus Graph and indexing from Solr and notification features from Kafka. The system provides exceptional column-level lineage tracking for Hive, Spark, and Storm environments.
- **LinkedIn Wherehouse**: The pioneering metadata portal established a unified platform to combine metadata information with Hadoop and Teradata systems. The system focused on web-scale discovery and lineage functions, which served as a foundation for developing advanced platforms [13].

- **Cloudera Navigator**: A commercial governance suite supporting metadata management, auditing, and visualization. The system created automated audit trails and operated through a central metadata server, which handled both structured and custom metadata [14].

These platforms show how metadata-centric governance architectures are evolving to meet the needs of contemporary data ecosystems.

V. IMPLEMENTING METADATA-DRIVEN GOVERNANCE: A FRAMEWORK

A metadata-driven governance layer implementation represents a strategic initiative that extends beyond the scope of a single project. The path to success requires a staged maturity model, along with specific roles and performance-based metrics.

A. Phased Maturity Model

Organizations develop their governance capabilities through four distinct stages.

- Level 1 – Ad-Hoc/Chaotic: No centralized governance. Metadata is missing or siloed. The data lake resembles a data swamp, where discovery depends on informal knowledge and trust is low.
- Level 2 – Foundational: A central data catalog is introduced. Technical metadata is ingested using crawlers and schema inference. Basic search and discovery become possible.
- Level 3 – Managed: The organization begins to focus on business metadata. Stewards define glossaries and quality rules. The automation of data lineage occurs for essential pipelines and governance policies to receive proper documentation.
- Level 4 – Optimized: The governance system operates proactively through automated processes. The system utilizes metadata to ensure security through Apache Ranger, while also monitoring quality and optimizing costs. The governance framework supports organizational agility instead of creating obstacles to it.

B. Key Roles and Responsibilities

The success of effective governance requires teams to work together functionally while maintaining clear roles and responsibilities.

- **Data Steward:** The business domain expert maintains full responsibility for data quality definition and curation and data usage management [6].
- **Data Curator:** The technical position focuses on maintaining accurate metadata, proper tagging, and catalog structure.
- **Data Governance Council:** The leadership body establishes organizational policies while handling ownership disputes and maintaining strategic alignment throughout the organization.

C. Measuring Success

The evaluation of governance outcomes should focus on business results instead of measuring metadata quantities. Key metrics include:

- Efficiency Metrics:
  - The reduction of time needed to discover and understand data stands as one of the key performance indicators.
  - The system provides automated capabilities for compliance and audit reporting.
- Quality & Trust Metrics:
  - The number of errors in reports that stem from data issues decreases.
  - The data trust score increases because of lineage tracking, checks, and endorsements.
- Value Metrics:
  - The number of new use cases that become possible because of previously inaccessible data.
  - The valuation of information assets has become quantifiable through an emerging CFO-driven governance practice [15].

VI. CHALLENGES AND FUTURE DIRECTIONS

The development of metadata-driven middle-layer architecture did not eliminate existing challenges, which led to the identification of future directions for data governance.

A. Remaining Technical Challenges

- **Scalability:** Exabyte-scale data lakes create performance challenges for graph database queries and search engine indexing, which demand high-availability architectural solutions.
- **Schema Inference Accuracy:** The process of inferring schemas from complex nested JSON or XML data remains imprecise. The research community actively investigates methods to achieve optimal precision-recall tradeoffs when processing large JSON datasets [12].
- **Real-time Lineage:** The development of batch lineage tracking in Hive and Spark has progressed, but real-time systems such as Kafka and Storm still lack effective column-level lineage tracking capabilities.

B. Organizational and Cultural Hurdles

- **Skills Gap:** The insufficient number of professionals with expertise in both technical elements and business operations is still a barrier to effective governance initiatives.
- **Cultural Resistance**: Most organizations view governance as a limiting factor. Executive sponsorship, combined with a cultural transformation that emphasizes governance as a trust-building mechanism and agile data sharing practices, will help organizations overcome this challenge.

C. The Future: AI-Driven and Standardized Governance

1) **Augmented Data Management:** AI/ML were expected to transform governance by automating metadata discovery, anomaly detection, and more-reducing IT dependency and enabling proactive governance [15].
2) **Standardization Efforts:** Standards bodies, including IEEE, started working on metadata interoperability, like Schema.org's success in structuring web data [4]. These efforts aim to unify governance across hybrid and multi-cloud environments.

Ultimately, the middle layer is evolving from a passive metadata repository to an intelligent, autonomous system capable of inferring context, detecting issues, and enforcing policies, realizing its role as the self-governing brain of the data ecosystem.

VII.CONCLUSION

The ongoing high failure rates of big data projects stemmed from architectural and strategic issues rather than technological problems. The transition from data warehouses with controlled schema-on-write architecture to data lakes with flexible schema-on-read capabilities created a vital governance gap, known as the "forgotten middle layer," that separated raw data infrastructure from value-generating analytics. The adoption of semi-structured data formats, such as JSON and XML, exacerbated the control gap because their flexible nature undermined traditional centralized management systems, creating "data swamps" that were both untrustworthy and unusable.

The solution to this problem requires explicit design and implementation of metadata-driven governance frameworks to establish the middle layer. Organizations can recover essential data discovery and lineage functions, along with quality management and semantic consistency, by creating a central repository that integrates technical, operational, and business metadata. Open-source projects, such as Apache Atlas, along with early industry platforms like LinkedIn's WhereHows, and commercial solutions like Cloudera Navigator, demonstrate that this approach represents a necessary and convergent design pattern for managing big data.

The implementation of this middle layer requires a phased maturity model, along with new roles for data stewards and curators, and the continuous delivery of business value. The path toward an automated and intelligent governance layer continues to advance through machine learning technology, while industry standards provide foundational support despite existing technical and cultural obstacles. Data-driven organizations must recognize that mastering metadata is a crucial strategic requirement for their success. The modern data landscape requires strategic mastery of metadata as its central imperative to control complexity and extract the substantial yet elusive potential of big data.

REFERENCES:

[1] P. Tyagi and H. Demirkan, "The biggest big data challenges." Nov. 07, 2016. doi: 10.1287/lytx.2016.06.05.

[2] A. A. Frozza, R. D. S. Mello, and F. De Souza Da Costa, *An Approach for Schema Extraction of JSON and Extended JSON Document Collections*. 2018, pp. 356–363. doi: 10.1109/iri.2018.00060.

[3] J. L. C. Izquierdo and J. Cabot, "JSONDiscoverer: Visualizing the schema lurking behind JSON documents," *Knowledge-Based Systems*, vol. 103, pp. 52–55, Jul. 2016, doi: 10.1016/j.knosys.2016.03.020.

[4] R. V. Guha, D. Brickley, and S. Macbeth, "Schema.org," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, Jan. 2016, doi: 10.1145/2844544.

[5] A. Oguntimilehin and O. Ademola, "A review of big data management, benefits and challenges," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 5, no. 6, pp. 433–438, Jun. 2014, [Online]. Available: https://www.researchgate.net/publication/280933768_A_Review_of_Big_Data_Management_Benefits_and_Challenges

[6] R. Abraham, J. Schneider, and J. V. Brocke, "Data governance: A conceptual framework, structured review, and research agenda," *International Journal of Information Management*, vol. 49, pp. 424–438, Dec. 2019, doi: 10.1016/j.ijinfomgt.2019.07.008.

[7] W. Eckerson, "Ten characteristics of a Modern Data Architecture," *Datalere*, Nov. 25, 2018. https://datalere.com/articles/ten-characteristics-of-a-modern-data-architecture

[8] "Apache Atlas – Architecture," Jun. 14, 2018. https://atlas.apache.org/1.0.0/Architecture.html

[9] A. Reeve, "Managing Data in Motion: data integration best practice techniques and technologies," Elsevier, 2013. [Online]. Available: https://d117h1jjiq768j.cloudfront.net/docs/default-source/default-document-library/progress/documents/book-club/managing-data-in-motion.pdf?sfvrsn=ae6e268f_0

[10] "Apache Atlas – Building & Installing Apache Atlas," Jun. 28, 2019. https://atlas.apache.org/2.0.0/InstallationSteps.html

[11] J. Kreps, "The Log: What every software engineer should know about real-time data's unifying abstraction," *LinkedIn*, Dec. 16, 2013. https://engineering.linkedin.com/distributed-systems/log-what-every-software-engineer-should-know-about-real-time-datas-unifying

[12] M. A. Baazizi, H. B. Lahmar, D. Colazzo, G. Ghelli, and C. Sartiani, "Schema inference for massive JSON datasets," in *HAL (Le Centre Pour La Communication Scientifique Directe)*, Venice, Italy. [Online]. Available: http://dx.doi.org/10.5441/002/edbt.2017.21

[13] A. Kalam, "MANAGEMENT PERSPECTIVES OF DATA-DRIVEN, ECOSYSTEM-BASED BUSINESS TRANSFORMATION," thesis, 2019. [Online]. Available: https://repositum.tuwien.at/bitstream/20.500.12708/8506/2/Kalam%20Alin%20-%202019%20-%20Management%20perspectives%20of%20Data-driven%20ecosystem-based...pdf

[14] M. Small, "Big Data Analytics – security and compliance challenges in 2019," 80072, Apr. 2019. [Online]. Available: https://www.comforte.com/fileadmin/Collateral/WP_KC_Security_and_Compliance_Challenges_in_2019.pdf

[15] "Gartner's top data and analytics predictions for 2019 - Health Data Management," *Health Data Management*, Jan. 18, 2019. https://www.healthdatamanagement.com/articles/gartners-top-data-and-analytics-predictions-for-2019