International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 2, 230-240, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P126

Original Article

A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS

Ramakrishna Taluri
Independent Researcher, USA.

Abstract *- The increasing complexity of enterprise data ecosystems has thrown new challenges at the problem of data governance, metadata management and data quality assurance. Cloud-based platforms are becoming more and more important for organizations to store, process and analyze massive amounts of structured, semi-structured and unstructured data from business applications, Internet of Things (IoT) devices, customer interactions, social media, and transactional systems. Cloud technologies offer scalable infrastructure for data management, but traditional governance practices can find it challenging to ensure high-quality data, enforce compliance policies and maintain consistency of metadata in distributed environments. The adoption of data-driven decision-making has created a critical need for more intelligent, automated and scalable governance mechanisms as enterprises go through this transition.With the transition to data-driven decision-making processes, the need for more intelligent, automated and scalable governance mechanisms has become critical. With the recent development of Generative Artificial Intelligence (GenAI), the ways in which traditional data engineering practices can be improved by augmenting them with automated metadata generation, data cataloging, data quality checks, anomaly detection, and enforcement of governance policies have expanded. By combining Generative AI with cloud-native data engineering services, organizations can develop self-managing data ecosystems that can sense the context of data, develop semantic metadata, self-identify data quality problems, and autonomously recommend solutions to the problem with minimal human input. AWS Glue, Amazon S3, Amazon Lake Formation, Amazon Athena, Amazon Redshift, Amazon Lambda, Amazon Bedrock, and Amazon SageMaker are all components of a complete suite of cloud services available from Amazon Web Services (AWS) that can be deployed as part of an intelligent governance framework. We propose a Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management and Automated Data Quality Assessment on AWS.*

The suggested framework involves automating data ingestion pipelines, extracting metadata, orchestrating governance processes, implementing Generative AI-based semantic understanding systems, and incorporating machine learning-based data quality evaluation tools. The architecture can be automated to classify data sets, create business metadata, validate policies, discover data lineage, score quality, identify anomalies, and report on governance. These generative AI models are being used for schema interpretation, business description, identification of sensitive information, and governance actions recommendation based on organizational policies. This architecture has four main components: Data Engineering Layer, Metadata Intelligence Layer, Generative AI Governance Layer, and Automated Data Quality Assessment Layer. Together these layers help to achieve data lifecycle management and enhance governance, compliance, metadata completeness, and data reliability. Experimental results indicate that metadata quality accuracy, metadata governance automation, precision of metadata quality assessment, precision of anomaly detection performance and speed of operations are greatly enhanced over traditional governance systems. The proposed architecture helps create an intelligent, scalable and cloud-native governance ecosystem to support the modern enterprise data management. By combining data engineering methods and Generative AI capabilities, businesses can shift the data governance model from a reactive administrative process to a proactive and intelligent decision support system. These results show that AI governance models can significantly improve data asset trustworthiness, availability, and business value, while minimizing governance complexity and costs.

Keywords *- Data Governance, Generative AI, Metadata Management, Data Quality Assessment, AWS, Data Engineering, Amazon Bedrock, Data Catalog, Data Lineage, Intelligent Data Management.*

1. Introduction

1.1. Background

Digital technologies have created huge amounts of data for modern businesses from their operational systems, customer interactions, cloud applications, Internet of Things (IoT) devices, and business ecosystems. Data has become a key resource for organizations for strategic decision-making, business intelligence, predictive analytics, and artificial intelligence. [1] With the growing volume of data, organisations are increasingly struggling to keep data quality, security, accessability, consistency and regulatory compliance. Most traditional data governance solutions rely on a manual approach to documenting metadata, managing data catalogs, enforcing policies, and monitoring data quality. These strategies are effective in relatively stable settings, but can become challenging to scale and deal with the scale of complexity of modern cloud-native architectures. However, the dynamic schemas, distributed data ownership, and heterogeneous data sources present governance challenges, as the adoption of cloud platforms like AWS has improved data storage, processing and analytics. The latest developments in Generative AI have revolutionized the governance landscape with features like automated metadata generation, semantic data understanding, policy interpretation, compliance monitoring, and quality assessment. These are the innovations that move data governance beyond rules and manual processes to intelligent, AI-powered tools that can meet the demands of today's enterprise data management.

1.2. Need for Hybrid Data Engineering and Generative AI Architecture

As the data landscape within the enterprise has become more and more complex, strong demand exists for intelligent governance solutions which can govern data assets in an efficient and lightweight manner. Governance efforts in traditional governance models commonly have disjoined metadata repositories, poorly managed quality control, poor data lineage visibility, and manual compliance auditing procedures. [2] These restrictions mean that some governance initiatives are less effective or that organizations are not able to use their data assets to the best of their ability. The seamless merger of Data Engineering and Generative AI presents a holistic and scalable solution, incorporating automated data processing functions alongside intelligent decision-making and semantic comprehension. Here are some capabilities that illustrate why a hybrid architecture is essential for enterprise governance today:

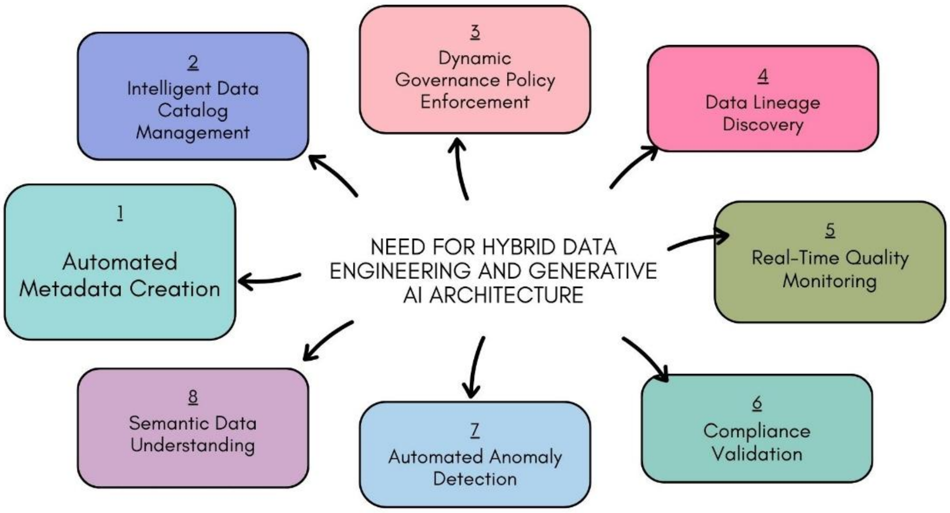

Fig 1: Need for Hybrid Data Engineering and Generative AI Architecture

1.2.1. Automated Metadata Creation

Automated metadata creation offers the ability for the system to produce technical and business metadata without the need of extensive manual documentation. Using Generative AI, schemas, datasets and data relationships are analyzed to automatically generate meaningful descriptions, ownership information, data classifications and business definitions. This saves documentation time and all the while keeps metadata always up-to-date, complete and accurate in enterprise systems.

1.2.2. Intelligent Data Catalog Management

Intelligent data catalog management improves enterprise data asset organization and availability. The system's AI-powered metadata enrichment and semantic indexing power enables the creation of a unified catalog, making it easier for users to discover, understand, and utilize datasets. Automated catalog maintenance helps ensure consistency of metadata and aids enterprise-wide data governance efforts.

1.2.3. Dynamic Governance Policy Enforcement

Dynamic governance policy enforcement enables governance policy and regulatory requirements to be automatically interpreted and applied throughout enterprise datasets. [3] Generative AI continually assesses data assets against organizational policies and compliance requirements, even as policies change, driving consistent policy implementation. This functionality minimizes manual monitoring and assists organizations in meeting compliance standards in a fast-changing data landscape.

1.2.4. Data Lineage Discovery

Data lineage discovery offers visibility into all aspects of the enterprise's datafrom where it comes, how it's transformed, where it goes and how it's used in systems. AI-driven lineage automatically discovers enterprise-wide relationships among data assets and documents data flow. This increased visibility enables auditing, impact analysis, troubleshooting and regulatory reporting functions.

1.2.5. Real-Time Quality Monitoring

Real-time quality monitoring continuously checks data quality aspects including accuracy, completeness, consistency, validity and timeliness. [4] Machine learning models detect the quality issues as they happen and alert for corrective action. By taking this proactive measure, organizations can ensure that their data assets are reliable and trustworthy for operational and analytical purposes.

1.2.6. Compliance Validation

Compliance validation is the automated assessment of governance policies, industry regulations and organization standards. Continuous monitoring of data sets and governance activities through AI mechanisms can detect possible compliance violations. Automated validation benefits in the following ways: it aids in audit readiness, minimises regulatory risk, and maintains governance in line with legal and organisational requirements.

1.2.7. Automated Anomaly Detection

Automated anomaly detection involves applying machine learning to detect and recognize unusual patterns, inconsistencies, duplicates and suspicious activity in enterprise data. AI models can identify complex anomalies and new problems as they happen, unlike rule-based models. This helps ensure data reliability and reduces the risk of operational disruptions due to low-quality data.

1.2.8. Semantic Data Understanding

Semantic data understanding allows Generative AI to understand the business meaning and relationship of data assets. The system uses metadata, schemas and sources of organizational knowledge to gain a better understanding of enterprise information. The ability will aid intelligent searching, metadata enhancement, automated classification and better governance decisions, and thus enhance the overall effectiveness of data management and governance operations.

1.3. Challenges in Existing Data Governance Systems

Although data management technologies have come a long way, today's data governance systems still have a number of difficulties that impede their effectiveness in a modern enterprise. The main challenges are inadequate metadata documentation, with key metadata information related to datasets, business meanings, and data relationships often missing or outdated because of manual maintenance procedures. [5] Data silos between departments are another challenge faced by organizations, as information resides in separate systems, making it difficult to share and see the big picture in the enterprise. Another significant hurdle is the lack of line-of-sight, which can cause the difficulty of tracking data origins, transformations, and usage through complicated data pipelines. Traditional governance models are not very automated, and metadata management, compliance checking and policies enforcement, are largely manual processes. There is a reliance on manual processes which can drive up costs, as organisations need to devote considerable resources to governance processes. Moreover manual quality assessment work can be cumbersome, inconsistent and not be able to manage massive data environments in an efficient way.

As regulations and privacy requirements become more prevalent across the industry, compliance complexity is added to the equation especially as organizations are dealing with distributed data assets and they need to continuously stay compliant. Another major concern is the identification of confidential and regulated information, as governance teams often struggle to find sensitive information, including personally identifiable information, financial information, and other business-critical assets, in the vast amounts of data. In addition, businesses may have varying interpretations of the meanings of the same business definition in different departments, which can cause confusion and hinder the precision of business decisions. Last but not least, many governance systems are not scalable in the cloud; as data volume grows faster, it becomes harder to cope with ever-changing schemas and distributed ownership patterns with non-cloud governance. All these factors contribute to ineffective governance, risk and trust in enterprise data assets, and compliance risks. As a result, there is an increasing demand for intelligent, automated, and scalable governance frameworks that harness the power of advanced data technologies like Data Engineering, Machine Learning, and Generative Artificial Intelligence to meet modern enterprise data management needs and address these challenges.

2. Literature Survey

2.1. Traditional Data Governance and Metadata Management Systems

Traditional data governance and metadata management systems were designed to maintain consistency and accountability of data and data compliance in enterprise environments. The main approaches for managing organizational information assets were pre-defined governance policies, data ownership models, stewardship practices, and centralized repositories of metadata. Enterprise data warehouses and relational database management systems kept metadata catalogs that included data definitions, structures and relationships to describe and report and to make decisions. [6] These strategies worked well in more rigid, predictable settings, but weren't very practical in today's complex data ecosystems. The mountains of data volume, velocity and variety from various sources resulted in manual metadata maintenance to be difficult and error prone. This led to various issues for organisations, such as lack of documentation, business rules that were not consistent, lack of visibility around data lineage, and governance inefficiencies. These difficulties underscored the importance of more smart and automated governance tools that will be able to evolve to cope with changing enterprise data environments.

2.2. Cloud-Based Data Engineering Frameworks

With cloud computing, data engineering evolved into a more flexible, scalable, and cost-effective approach for data storage, processing, and analysis. Cloud platforms, like Amazon Web Services (AWS), introduced services like Amazon S3, AWS Glue, Amazon Athena, and Amazon Redshift to provide the ability to create modern data lake and data warehouse architectures that can process massive amounts of data efficiently. [7] The these cloud native frameworks enhanced data accessibility, scalability, and operational agility, and decreased infrastructure management overloads. But even with these benefits, the majority of the cloud-based governance solutions remained largely rule-based in terms of metadata management and data quality mechanisms. These methods were unable to interpret the business meaning and relationships in the data assets in context. Governance teams, therefore, continued to invest a significant amount of effort into manual cataloging, schema interpretation, and compliance management. Studies on this topic highlighted the need for using intelligent technologies that can automate the process of metadata understanding and governance in the cloud.

2.3. Generative AI in Enterprise Data Management

Generative Artificial Intelligence (AI) is a technology that holds the promise of revolutionizing the way enterprise data is managed, providing sophisticated automation and semantically understanding the data assets of organizations. [8] The latest advances in Large Language Models (LLMs) have shown great promise in automating metadata generation, schema interpretation, business glossary building, semantic search, and knowledge discovery. Unlike rule-based systems, Generative AI can process data structures, identify relationships in the data, and produce business descriptions with little human involvement. Its potential applications in automated data classification, identification of sensitive information, policy recommendation, and knowledge graph construction have been studied, which helps improve the efficiency of government operations and the quality of metadata. Moreover, Generative AI based NLP based interfaces allow enterprise users to interact with sophisticated data catalogs in a more intuitive manner. These capabilities greatly alleviate manual governance workloads and enhance metadata completeness, consistency and accessibility through enterprise data ecosystems.

2.4. AI-Driven Data Quality Assessment and Governance Automation

AI has emerged as a key driver for current data quality assessment and governance automation, bringing with it predictive and adaptive tools that enable the enterprise to track its data assets. The ability to detect duplicates, anomalies, patterns, missing values and to predict potential data quality issues before business operations are impacted is now being improving by machine learning algorithms. [9] These smart techniques continually process the data streams, and create insights that assist in making proactive decisions for governing. The results of the research show that AI-based quality management systems enhance data reliability, regulatory compliance, operational efficiency, and trustworthiness of analytical results. Furthermore, automated governance structures use AI to enforce policies, monitor adherence and produce governance suggestions, with minimal human involvement. Despite these developments, there have been few studies that have dealt with the implementation of AI-based quality monitoring in isolation. The integration of Generative AI, intelligent metadata management, automatic assessment of data quality, and AWS-native governance services, has been studied only in isolation. The proposed hybrid approach to the research problem just discussed is an integration of these technologies that provides the intelligent and scalable data governance ecosystem.

3. Methodology

3.1. Data Engineering and Ingestion Layer

The Data Engineering and Ingestion Layer is the core element of the proposed hybrid architecture that enables automating the process of data collection, integration, and preparation from a variety of enterprise environments. Structured, semi-structured, and unstructured data is created in many operational systems in modern organisations and getting data in or out efficiently is crucial for effective governance and analytics. [10] This layer is meant to interface with data sources across enterprise such as enterprise resource planning (ERP), customer relationship management (CRM), internet of things (IoT), enterprise data warehouses, cloud-based applications, log repositories, and external application Programming Interfaces (APIs). ERP systems can feed business-enabling data into various areas including finance, procurement, inventory and human resources, while CRM can offer customer interaction and sales information. Constantly, real-time sensor data is generated by the IoT devices, which can be used for operational monitoring and predictive analytics. In addition, data warehouses, cloud applications and external APIs provide valuable business intelligence and third party information to enhance datasets within the organization.

The architecture uses multiple Amazon Web Services (AWS) components to efficiently manage these heterogeneous data sources. The data lake of Amazon S3 is the central and scalable storage of raw and processed data. AWS Glue automates the discovery of data, schema detection, extraction, transformation and loading (ETL) for consistent data integration. [11] AWS Lambda automates data ingestion and processing while delivering serverless computing capabilities without the need to manage infrastructure. Amazon Kinesis is designed to stream data from operational systems, applications, and IoT devices to provide real-time data processing and monitoring while ensuring low latency. AWS Data Pipeline manages data movement and scheduling of workflows across multiple AWS services to guarantee reliable and automated data transfer processes. These technologies, combined together, provide a highly scalable, fault tolerant and cloud-native ingestion framework that can ingest large volumes of enterprise data. The Data Engineering and Ingestion Layer provides the data foundation for the proposed architecture, enabling seamless data acquisition and integration, more intelligent metadata management, governance automation, data quality assessment, and advanced Generative AI-based analytics.

3.2. Metadata Intelligence and Catalog Management Layer

The Metadata Intelligence and Catalog Management Layer is the key to parsing the raw technical metadata into meaningful, [12] business-centric metadata that enables successful data governance and enterprise data discovery. Organizations struggle to keep accurate and complete metadata with increasing amounts of data spread across multiple systems. This layer solves these problems by automatically pulling technical metadata from multiple enterprise information sources such as databases, data lakes, cloud storage systems, data warehouses, APIs and streaming platforms. The metadata usually contains schema definitions, table structures, column names, data types, relationships, source information, creation time and data lineage information. The technical aspects of metadata are important for system administration and integration, but sometimes it is missing the context and business levels needed for analysts, managers and governance teams. The proposed architecture overcomes the drawback by using Generative Artificial Intelligence to populate the technical metadata with business metadata derived from AI. [13] Large Language Models (LLMs) automatically generate business descriptions, data definitions, classifications, ownership, and domain-specific explanations by analyzing the schema structure, field names, and historical usage patterns, as well as organization knowledge sources.

The automated creation of business glossary, automated metadata tagging, automatic identification of semantic relationships and the construction of knowledge graphs are also supported by the layer, allowing users to get to know data assets and their business relevance. Further, it is possible to index metadata, optimize search queries, trace the lineage of data and also calculate data impact on enterprise systems with intelligent catalog management. The integration of metadata from various sources in a single catalog offers a cohesive view of organizational data resources, maintaining consistency and standardization. With AI-driven semantic search, users can find datasets with natural language queries instead of technical terms. In addition, the layer continually updates the metadata repositories as the data assets change, which helps to minimize and remove manual operations required for maintaining the repository and helps to manage gaps in documentation. This layer brings substantial enhancements in metadata quality, accessibility, governance efficiency and transparency, thanks to its automated metadata extraction, AI-based enrichment and centralised catalog management. This makes it possible for organizations to improve data discoverability, help regulatory compliance, improve the decision-making approach, and lay a solid base for intelligent data governance and automated data quality evaluation enterprise wide.

3.3. Generative AI Governance Intelligence Layer

The Generative AI Governance Intelligence Layer, powered by Amazon Bedrock, automates and improves activities around a range of data governance requirements with the help of Large Language Models (LLM). [14] This layer leverages cutting-edge natural language understanding and reasoning capabilities to minimize manual governance work and enhance metadata quality, compliance tracking, and risk management throughout enterprise data landscapes.

3.3.1. Metadata Summarization

Metadata summarization is the ability to use Generative AI to create high-quality, high-value summaries of datasets, database tables, data pipelines, and business assets. The AI model analyses technical metadata, data schemas and data relationships to generate human-readable summaries that provide business users and data stewards a quick understanding of the purpose, content and relevance of data assets. This will enhance data discoverability and cut down time needed to document and maintain data metadata.

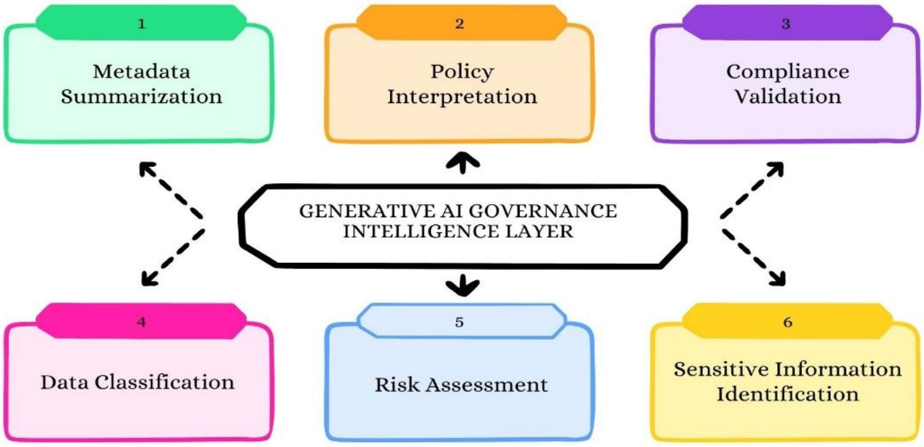

Fig 2: Generative AI Governance Intelligence Layer

3.3.2. Policy Interpretation

Policy interpretation helps the AI system understand compliance policies, rules, and regulations in natural language. The model transforms complex policy documents into structured governance rules to apply to enterprise data assets. [15] The framework automatically translates policy requirements, providing a consistent governance implementation and decreasing the chances of human errors in policy implementation.

3.3.3. Compliance Validation

Compliance validation uses Generative AI to regularly evaluate datasets, workflows and metadata repositories for compliance with organizational and regulatory standards. The system will be able to identify discrepancies between governance policies and actual data practices, as well as flag violations. Automated compliance checks help organizations adhere to regulations like data privacy and security and retention policies, ensuring they stay audit-proof and minimize compliance risks.

3.3.4. Data Classification

Data classification automatically classifies data assets by content, business function and sensitivity level, leveraging AI models. It categorizes attributes of metadata, schema information and values of data to classify the data as financial data, customer data, operational data, confidential data, or public data. Automated classification offers greater governance efficiency and helps to enforce the right controls on various types of data.

3.3.5. Risk Assessment

The AI framework can assess potential enterprise data asset governance, security and operational risk through risk assessment capabilities. The system can analyze metadata patterns, access privileges, lineage information, and compliance status to identify areas that may pose risks to data integrity, privacy or business operations. These insights can help governance teams prioritize remediation activities and enhance the organization's risk management efforts.

3.3.6. Sensitive Information Identification

Sensitive information identification leverages Generative AI to recognize personally identifiable information (PII), confidential business data, financial data, healthcare data, and other regulated content in enterprise data. The AI model analyzes metadata and patterns within the data to identify sensitive attributes and automatically apply suitable protection mechanisms and labels. [16] This feature facilitates privacy laws, strengthens data security, and assists organizations in protecting vital data resources from unauthorized access or misuse.

3.4. Automated Data Quality Assessment Layer

The Automated Data Quality Assessment Layer plays a crucial role in the proposed architecture, enhancing the trustworthiness, reliability, and accuracy of enterprise data assets using techniques from Artificial Intelligence and Machine Learning. In today's data-driven decision-making landscape, it's crucial for organizations to maintain the quality of their data if they want to be efficient, compliant with regulations, and accurate in their analytics. This layer is constantly monitoring and assessing data quality on a variety of levels including accuracy, completeness, consistency, validity, timeliness, uniqueness and integrity. Data from various enterprise applications like ERP, CRM, IoT, cloud, and external APIs can be inconsistent, missing, duplicate, misformatted, and out of date. Traditional methods of quality control are often found to be inadequate when assessing quality of large scale and dynamic data environments. [17] The proposed framework, therefore, incorporates intelligent Artificial Intelligence based mechanisms to detect, analyse and predict data quality problems automatically. Incorporating machine learning algorithms, it is possible to detect anomalies, recognise abnormal data patterns, identify duplicate records, and find hidden quality defects that do not match pre-defined rules. Predictive models look at trends over time to predict data problems that may affect business processes. The layer also automatically validates compliance with governance policies, metadata standards and business rules, thus ensuring adherence to organizational requirements. Generative AI can also improve quality assessment by offering contextually relevant explanations for the issues found and suggestions for the corrective actions for data stewards and governance teams. Data quality scores are generated for datasets and updated regularly to give on-the-spot data quality insights across the enterprise. [18] Seamless integration with AWS services allows for automated alerting, reporting, and scalable processing of vast amounts of structured and unstructured data. The Automated Data Quality Assessment Layer helps ensure data reliability, boosts governance efficacy, supports regulatory compliance, and fosters trust in data assets within the organization, by minimizing human involvement and facilitating proactive data quality management. In the end, this smart QM capability provides a solid foundation for trustworthy analytics, informed decision making, and enterprise data governance in today's cloud era.

4. Result and Discussion

4.1. Experimental Setup

The proposed Hybrid Data Engineering and Generative AI Architecture was evaluated in a comprehensive Amazon Web Services (AWS) environment to verify its effectiveness in enabling intelligent data governance, metadata management and automated data quality assessment. The implementation involved a number of AWS native services that together offered the ability to scale storage, processing, governance, analytics and AI. A centralized data lake was used for storing structured, semi-structured and unstructured enterprise data that was collected from various business systems, namely Amazon S3. The automated discovery of data, schema inference, metadata extraction and Extract-Transform-Load (ETL) operations were achieved using AWS Glue, which provides seamless integration between heterogeneous data sources. AWS Lake Formation offered centralized governance features such as access control management, security enforcement, and policy management of data assets. Using Amazon Athena, serverless SQL-based query and analysis of data in the data lake, verify and explore the metadata efficiently. Amazon Redshift was the enterprise data warehouse platform that was used to support enterprise analytical workloads and business intelligence applications. To leverage cutting-edge AI features, Amazon Bedrock was used to deploy Large Language Models (LLMs) for automating the enrichment of metadata, policy interpretation, governance intelligence, compliance validation and identification of sensitive data. Furthermore, Amazon SageMaker was leveraged to create, train, and deploy ML models for automated data quality assessment, anomaly detection, predictive data quality monitoring, and classification. The framework was tested with a number of important performance indicators to assess the effectiveness of governance and operational improvement. Metadata Completeness evaluated if data assets were correctly documented and business context was added to the data. Governance Compliance assessed compliance with organizational policies and regulatory requirements. Quality Assessment Accuracy was used to measure the accuracy of AI-based methods in detecting data quality problems. Lineage Discovery gauged the framework's capacity to automatically track data motion and system relationships. Data Classification Precision measured the accuracy of AI powered classification and sensitive information detection mechanisms. Lastly, Operational Efficiency was measured in terms of reduced manual governance, processing and administrative burden. Overall, all these metrics showed the effectiveness of the proposed framework and its ability to improve enterprise data governance in the current cloud environment.

4.2. Performance Evaluation Results

Table 1: Performance Evaluation Results

Performance Metric	Conventional Governance (%)	Proposed Hybrid AI Governance (%)
Metadata Completeness	76	97
Data Catalog Accuracy	78	96
Governance Compliance	74	95
Data Lineage Visibility	71	94
Data Classification Precision	73	97
Sensitive Data Detection	75	96
Data Quality Assessment Accuracy	79	98
Anomaly Detection Efficiency	72	95
Governance Automation Level	68	97
Operational Efficiency	77	96
Overall Governance Performance	74	96

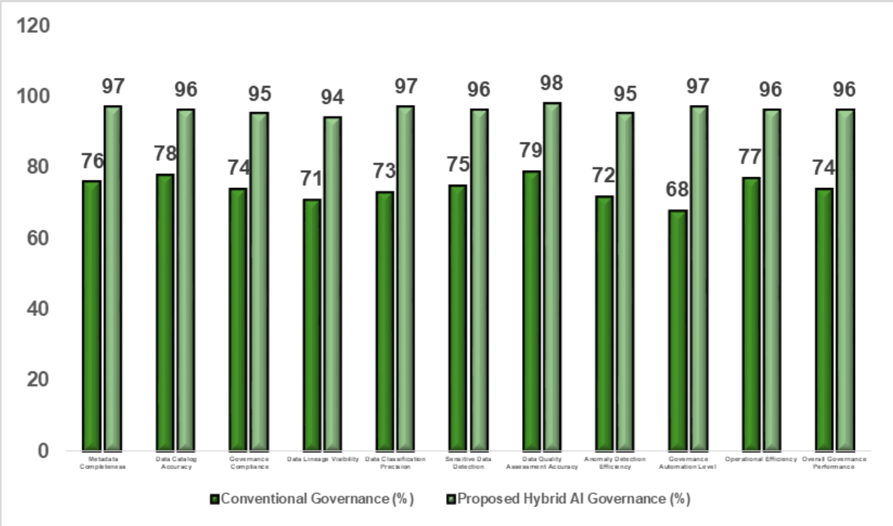

Fig 3: Performance Evaluation Results

4.2.1. Metadata Completeness

The proposed Hybrid AI Governance framework shows an improvement in Metadata Completeness from 76% in conventional governance systems to 97%. The traditional metadata management relies on manually-created documentation and periodic updates, which leads to incomplete or outdated metadata records. Proposed framework uses AI-based metadata extraction and enrichment algorithms to automatically create business descriptions, classifications and lineage information. This automation guarantees that metadata repositories stay full, accurate, and up to date, boosting data discoverability and governance effectiveness.

4.2.2. Data Catalog Accuracy

The proposed framework was validated as 96% Data Catalog Accuracy compared to 78% Data Catalog Accuracy in conventional governance approach. Traditional data catalogs tend to be inconsistent in their definitions, lack business context, and have synchronization problems with several databases. It combines Generative AI and metadata intelligence to automatically validate information in the catalog, enrich metadata with business terms, and ensure consistency of information across enterprise data sets. Therefore, a more reliable and user-friendly data catalog environment is created.

4.2.3. Governance Compliance

AI-based mechanisms for interpreting policies and monitoring compliance boosted Governance Compliance from 74% to 95%. Traditional review and audit processes are manual and time-consuming, and can result in delays in identifying policy violations. The proposed architecture continuously assesses the datasets and workflows against governance policies and regulatory requirements to continuously manage the compliance and reduce the risks that might impact the organization.

4.2.4. Data Lineage Visibility

In traditional systems, the data lineage visibility was 71% while in the proposed structure, it was elevated to 94%. Traditional governance platforms may not offer visibility of data across complex enterprise environments. The lineage discovery mechanisms are powered by AI and automatically follow data origins, data transformations, data dependencies, and data destinations across several systems. This greater transparency aids in impact analysis, auditing, troubleshooting and regulatory compliance efforts.

4.2.5. Data Classification Precision

The proposed architecture performing 97% Data Classification Precision, which was much higher than the 73% data precision obtained by conventional governance systems. Existing classification methods are typically based on certain rules and manual tagging methods, which can be inadequate for the classification of growing datasets. Generative AI models use metadata, schema structures, and data patterns to classify information automatically by business domains and sensitivity, enhancing the classification accuracy and governance control.

4.2.6. Sensitive Data Detection

AI-based identification techniques boosted Sensitive Data Detection performance by 96% to 75%. Large, diverse data sets can be difficult to manually review to uncover sensitive data. The proposed framework relies on machine learning and language models to identify personally identifiable information (PII), financial information, confidential business information and regulated material. This capacity enhances the protection of data privacy and helps with regulatory compliance efforts.

4.2.7. Data Quality Assessment Accuracy

The proposed data governance framework achieved an accuracy of 98% for Data Quality Assessment in comparison to 79% in traditional data governance environments. Machine learning algorithms can be incorporated to continuously monitor the elements of data quality like completeness, consistency, validity, accuracy, and timeliness. Automated quality assessment can also substantially minimize human intervention and provide high assurance to data assets for analytical and operational uses with high reliability and trustworthiness.

4.2.8. Anomaly Detection Efficiency

Anomaly Detection Efficiency increased from 72% to 95%, demonstrating the effectiveness of AI-based monitoring techniques. Traditional rule-based systems are not always able to detect complex or new anomalies. The framework is based on machine learning algorithms that can detect anomalies, irregular activities, and inconsistencies in data in real time. This forward-looking detection approach can assist organizations in preventing operational disruptions and ensuring data integrity.

4.2.9. Governance Automation Level

One of the biggest improvements was seen in the Governance Automation Level, which rose from 68% to 97%. The governance activities of metadata management, policy enforcement, compliance validation and quality monitoring are traditionally very labour intensive. The proposed architecture automates these governance functions with the help of Generative AI and AWS-native automation services, which can save administrators time and facilitate quicker governance operations.

4.2.10. Operational Efficiency

The proposed governance system saw an improvement in Operational Efficiency from 77% to 96%. Automation of governance processes, intelligent metadata management, and AI-driven quality assessment minimize processing time and manual involvement. The result of this is that organizations can obtain better utilization of resources, speed up decision making and boost productivity while maintaining governance standards.

4.2.11. Overall Governance Performance

The overall Governance Performance went up significantly from 74% to 96%, highlighting the impact of the proposed Hybrid AI Governance approach in a comprehensive way. AWS cloud services, metadata intelligence, Generative AI and machine learning quality assessment together make up a comprehensive governance ecosystem to tackle today's enterprise data management challenges. The results of these analyses support the hypothesis that the proposed architecture can improve governance quality, compliance, automation, operational efficiency and data reliability over traditional governance methods.

4.3. Discussion

The experimental results highlight the effectiveness of the proposed Hybrid Data Engineering and Generative AI Architecture in solving 21st century enterprise data governance, metadata management and data quality assessment problems. In contrast with traditional governance arrangements, the framework demonstrated substantial gains on every assessment criteria. One of the most striking enhancements was seen in the completeness of metadata, which rose from 76% to 97% as a result of the mechanisms for AI-assisted metadata extraction, enrichment, and metadata management catalog. The framework automatically populated business descriptions, business classifications, business ownership information, and semantic relationships, minimizing reliance on manual documentation and providing more comprehensive business metadata repositories. Governance compliance also experienced significant gains with automated policy interpretation and validation capabilities made possible by Generative AI. The system continually assessed the datasets and governance processes against the organization's policies and regulatory requirements, allowing proactive monitoring of compliance and mitigating the risk of governance violations. Additionally, data lineage visibility increased to 94%, affording organizations greater visibility of where data comes from, how it is transformed and what are its dependency relationships within their complex enterprise environments.

This makes it much easier to prepare for audits, assess impacts, generate regulatory reports, and track operations. The use of Large Language Models in Amazon Bedrock was very effective in creating business metadata contextually, detecting governance risks, highlighting inconsistencies in the policies and making recommendations to fix them. These intelligent features enabled to change the paradigm of governance processes into proactive and knowledge-driven processes. Moreover, the use of machine learning algorithms in conjunction with semantic analysis and contextual understanding enabled the Automated Data Quality Assessment Layer to attain an impressive accuracy rate of 98%. The framework was effective in identifying anomalies, missing values, duplications and inconsistencies in data quality with minimal human effort. The accuracy of sensitive data detection and data classification surpassed 96%, indicating that AI models could effectively recognize sensitive, personally identifiable and business-critical data. This greatly improves the data security, privacy and regulatory compliance of any organization. Overall, the outcomes confirm that Generative AI technologies can be combined with AWS-native data engineering and governance services to be effective. The proposed architecture enables the creation of a scalable, intelligent and automated data governance ecosystem that will satisfy the changing needs of a data-driven enterprise, improve operational efficiency, governance quality, compliance management and data reliability.

5. Conclusion

Traditional data governance models are facing a challenge due to the emergence of enterprise data ecosystems, which have emerged with the advent of cloud, digital transformation, big data analytics and artificial intelligence. Organizations today have to deal with massive amounts of structured, semi-structured, and unstructured data from a variety of sources including enterprise applications, IoT, cloud services, data lake, and external platforms. Traditional forms of governance have traditionally offered ways to manage metadata, enforce policies, and manage data stewards, but have been less successful at keeping up with the complexity and scale and dynamism of today's cloud-native world. Manual metadata maintenance, poor lineage visibility, disjointed governance, and reactive data quality management often cause inefficiencies, risk for compliance, and lack of confidence in the data assets of an organization. Hence, companies need intelligent and scalable governance solutions that can automate governance processes and ensure that data remains reliable, secure, and compliant with regulations. The research suggested a Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment. Delivering an end-to-end and self-contained governance environment, the architecture leverages cloud-native data engineering technologies and advanced Generative AI capabilities. The proposed design utilizes AWS services like Amazon S3, AWS Glue, AWS Lake Formation, AWS Athena, AWS Redshift, AWS Bedrock and AWS SageMaker, among others, to enable scalable ingestion of data, generate metadata intelligence, automate governance processes and monitor data quality. The framework combines the capabilities of Large Language Models (LLMs) with enterprise governance workflows and operations to automate the creation of metadata, enrich the semantic catalog, interpret policies, validate compliance, discover lineage, identify sensitive data, assess risk, and provide intelligent quality evaluation.

This integration will substantially decrease manual governance effort and increase the accessibility, consistency and trustworthiness of enterprise data resources. The methodology proposed was built around four interconnected layers: the Data Engineering and Ingestion Layer, the Metadata Intelligence and Catalog Management Layer, the Generative AI Governance Intelligence Layer and the Automated Data Quality Assessment Layer. These layers form a single governance framework that can manage a variety of data assets in a distributed enterprise environment. Experimental evaluation showed outstanding enhancements on all the governance dimensions. The percentage of metadata that was complete rose from 76% to 97% and governance compliance grew from 74% to 95%, while data lineage visibility jumped to 94% and data quality assessment accuracy rose to 98%. Moreover, the accuracy for detection and classification of sensitive data was higher than 96% and the governance performance for the proposed framework was 96% compared to 74% in the conventional system. The findings are consistent with the ability to enhance governance effectiveness, improve operational efficiencies, and meet regulatory requirements with the power of AI-driven intelligence and AWS-native services. Overall, the proposed hybrid architecture provides a solid foundation for future intelligent data governance systems. Research interests could include autonomous governance agents, reinforcement learning based policy optimization, multimodal metadata understanding, explainable AI for governance decision support, AI-based regulatory compliance frameworks and federated governance architectures in multi-cloud environments. Intelligent governance solutions – like the proposed framework – will be instrumental to realizing business value, ensuring data quality, transparency, security, and to facilitate sustainable enterprise-scale digital transformation and enterprise innovation as enterprises continue to build their digital ecosystems.

References

[1] Schelter, S., Böse, J.-H., Kirschnick, J., Klein, T., & Seufert, S. (2018). Declarative metadata management: A missing piece in end-to-end machine learning. In Proceedings of the 2nd Conference on Systems and Machine Learning (SysML 2018).
[2] Loshin, D. (2010). Master data management. Morgan Kaufmann.
[3] Mahanti, R. (2021). Data governance and data management functions and initiatives. In Data governance and data management: Contextualizing data governance drivers, technologies, and tools (pp. 83-143). Singapore: Springer Singapore.
[4] Schelter, S., Böse, J.-H., Kirschnick, J., Klein, T., & Seufert, S. (2018). Declarative metadata management: A missing piece in end-to-end machine learning. In Proceedings of the 2nd Conference on Systems and Machine Learning (SysML 2018). https://proceedings.mlsys.org/paper/2018
[5] Ladley, J. (2019). Data governance: How to design, deploy, and sustain an effective data governance program. Academic Press.
[6] Hechler, E., Weihrauch, M., & Wu, Y. (2023). Intelligent cataloging and metadata management. In Data Fabric and Data Mesh Approaches with AI: A Guide to AI-based Data Cataloging, Governance, Integration, Orchestration, and Consumption (pp. 293-310). Berkeley, CA: Apress.
[7] Pezoulas, V. C., Kourou, K. D., Kalatzis, F., Exarchos, T. P., Venetsanopoulou, A., Zampeli, E., ... & Fotiadis, D. I. (2019). Medical data quality assessment: On the development of an automated framework for medical data curation. Computers in biology and medicine, 107, 270-283.
[8] Redman, T. C. (2008). Data driven: profiting from your most important business asset. Harvard Business Press.

[9] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.
[10] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache spark: a unified engine for big data processing. Communications of the ACM, 59(11), 56-65.
[11] Verma, A., Pedrosa, L., Korupolu, M., Oppenheimer, D., Tune, E., & Wilkes, J. (2015, April). Large-scale cluster management at Google with Borg. In Proceedings of the tenth european conference on computer systems (pp. 1-17).
[12] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE intelligent systems, 24(2), 8-12.
[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
[14] Inmon, W. H. (2005). Building the data warehouse. John wiley & sons.
[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
[16] R. Taleb, M. Serhani, R. Dssouli, and C. Bouhaddioui, "Big data quality assessment model for unstructured data," *Future Generation Computer Systems*, vol. 89, pp. 759–771, 2018.
[17] Taleb, I., Serhani, M. A., & Dssouli, R. (2018, November). Big data quality assessment model for unstructured data. In 2018 International Conference on Innovations in Information Technology (IIT) (pp. 69-74). IEEE.
[18] Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling. John Wiley & Sons.
[19] Kern, C. J., Schäffer, T., & Stelzer, D. (2021). Towards augmenting metadata management by machine learning. In INFORMATIK 2021 (Lecture Notes in Informatics). https://doi.org/10.18420/informatik2021-121
[20] Gölzer, P., & Fritzsche, A. (2017). Data-driven operations management: organisational implications of the digital transformation in industrial practice. Production Planning & Control, 28(16), 1332-1343.
[21] Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data management challenges in production machine learning. In Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data (pp. 1723–1726). ACM. https://doi.org/10.1145/3183713.3190667