International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

Enhancing Federated Cloud with Explainable Generative Artificial Intelligence for Enterprise Data Modernization and Governance

Maheshwari Muthusamy

Team Lead, Infosys, Jalisco, Mexico

ABSTRACT: Modern enterprises are increasingly adopting federated multi-cloud environments to prevent vendor lock-in, maintain localized data sovereignty, and leverage specialized infrastructure across geographic regions. However, managing distributed data estates across heterogeneous cloud architectures introduces severe operational bottlenecks in data discovery, schema standardization, compliance enforcement, and cross-silo lineage tracking. Traditional centralized data governance models introduce performance friction and struggle to scale alongside high-velocity, unstructured data assets. This paper presents a novel framework integrating Explainable Generative Artificial Intelligence (xGenAI) into federated cloud architectures to automate enterprise data modernization and continuous governance. By embedding localized retrieval-augmented generative models and automated ontology mappers at distributed cloud nodes, the framework dynamically generates unified semantic metadata, maps cross-cloud lineage, and enforces fine-grained access policies. To ensure transparency and regulatory compliance, the system incorporates Shapley Additive exPlanations (SHAP) and attention-map attribution mechanisms, offering human-interpretable reasoning for automated classification, policy decisions, and synthetic data generation. Empirical evaluation across a simulated multi-cloud ecosystem demonstrates significant improvements in metadata auto-cataloging throughput, drastic reductions in policy audit latency, and enhanced explainability metrics over conventional governance toolchains.

KEYWORDS: Federated Cloud, Explainable Generative AI, Enterprise Data Modernization, Data Governance, Retrieval-Augmented Generation, Feature Attribution, Cross-Cloud Lineage.

I. INTRODUCTION

Enterprise digital transformation has accelerated the shift away from centralized data warehouses toward federated, multi-cloud data estates. Organizations deploy microservices, analytical engines, and operational databases across diverse cloud platforms—such as AWS, Azure, Google Cloud, and localized on-premises infrastructure—to capitalize on platform-specific optimizations and adhere to regional data residency regulations (e.g., GDPR, CCPA). While this decentralized approach provides operational flexibility and high availability, it fragments the enterprise data landscape. Data assets become isolated in multi-cloud silos, leading to inconsistent metadata definitions, duplicated storage, untracked lineage, and heightened security vulnerabilities. Modernizing these legacy and fragmented systems requires establishing a unified semantic layer without forcing physically centralized migration, which is often legally or economically infeasible.

Conventional enterprise data governance platforms rely heavily on static, rule-based heuristics and manual curation by domain-specific data stewards. These legacy systems require hand-crafted regex patterns, fixed classification rules, and periodic manual audits to detect sensitive records, map schema relationships, and enforce policy guardrails. As enterprises generate terabytes of semi-structured and unstructured data daily, manual stewardship becomes an operational bottleneck. Traditional tools fail to contextualize complex relationships across distributed environments, frequently resulting in false positives, uncataloged "dark data," and uncoordinated schema drift. Consequently, organizations face compliance penalties, delayed business intelligence, and increased friction when exposing data to downstream artificial intelligence applications.

The emergence of Generative Artificial Intelligence (GenAI) introduces powerful capabilities for understanding contextual semantics, auto-generating schema mappings, translating natural language queries, and synthesizing complex metadata. By deploying generative models across federated cloud nodes, organizations can automate the discovery, cataloging, and harmonization of unstructured and semi-structured data without transferring raw underlying datasets across network boundaries. However, deploying standard "black-box" generative models within enterprise

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

governance introduces critical trust and regulatory risks. Foundation models are prone to hallucinating data lineage, producing non-deterministic classification decisions, and generating opaque transformations. Without verifiable explainability mechanisms, compliance officers and data stewards cannot audit AI-driven governance decisions, rendering pure GenAI solutions unacceptable for highly regulated industries like finance and healthcare.

To bridge this operational gap, this paper introduces an integrated architecture combining Explainable Generative AI (xGenAI) with federated cloud orchestration for enterprise data modernization and continuous governance. The proposed system embeds localized light-weight generative models equipped with Retrieval-Augmented Generation (RAG) and post-hoc attribution frameworks (such as Integrated Gradients and SHAP) across multi-cloud environments. This design enables autonomous data classification, dynamic policy generation, and real-time cross-cloud lineage synthesis while producing human-auditable rationale trails for every automated governance action. The primary objective of this research is to establish a transparent, privacy-preserving, and scalable framework that modernizes legacy data silos into a trustable, unified data fabric.

II. LITERATURE REVIEW

The evolution of enterprise data architecture from centralized data lakes to federated paradigms, such as Data Mesh and Data Fabric, has been extensively documented in recent literature. Dehghani et al. introduced the foundational concepts of Data Mesh, arguing that domain-driven decentralized data ownership avoids the architectural bottlenecks inherent in monolithic data platforms. Subsequent studies by Saraswat et al. expanded on this by evaluating federated cloud environments, demonstrating that distributing data workloads across heterogeneous clouds enhances regional latency and satisfies data sovereignty laws. However, researchers such as Alhassan et al. highlighted that decentralized data estates create severe governance fragmentation. Without centralized visibility, maintaining unified access control policies, verifying data quality, and tracking end-to-end data provenance across cloud providers becomes difficult, leading to elevated security risks and compliance vulnerabilities.

To automate governance across distributed estates, early research explored machine learning techniques for automated metadata extraction and policy enforcement. Supervised classification models and natural language processing (NLP) pipelines, such as BERT-based named entity recognition (NER), were implemented to scan databases and flag sensitive Personal Identifiable Information (PII). While these approaches improved upon static rule engines, studies by Zhang et al. demonstrated their limitations when handling domain-specific jargon, complex nested structures, and polysemic data attributes across legacy enterprise schemas. Furthermore, traditional ML models operate independently within localized silos, failing to infer global schema dependencies or construct holistic semantic ontologies across multi-cloud environments without centralized data aggregation.

The incorporation of Large Language Models (LLMs) and Generative AI into data management has marked a significant shift in enterprise data engineering. Recent work by Chen et al. demonstrated that LLMs fine-tuned on SQL and data management artifacts can automatically generate complex ETL transformations, synthesize documentation, and perform zero-shot schema alignment across disparate sources. Additionally, Retrieval-Augmented Generation (RAG) architectures have been leveraged to ground generative outputs in organizational policy documentation, improving contextual alignment. Despite these advances, scholars like Amershi et al. emphasize that the black-box nature of generative models poses a challenge to automated data governance. Unexplainable AI outputs can cause automated misclassification of restricted data, incorrect policy assignments, and unverified data transformations, which undermine enterprise trust and violate strict regulatory mandates demanding algorithmic accountability.

To address model opacity, Explainable AI (XAI) frameworks have emerged as a critical requirement in automated decision-making platforms. Post-hoc explainability methods, such as SHAP (Shapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention attribution mechanics, have been validated in clinical and financial domains to expose model reasoning. However, existing XAI literature focuses primarily on predictive ML models, with limited research exploring its application to generative architectures operating within federated cloud governance frameworks. Current research lacks a unified approach that integrates explainable generative AI directly into federated cloud data planes to simultaneously drive data modernization (e.g., schema harmonization, code translation) and enforce continuous, auditable data governance. This paper addresses this gap by presenting an explainable, generative-driven framework tailored for federated enterprise data ecosystems.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

III. RESEARCH METHODOLOGY

Federated Multi-Cloud Infrastructure and Data Fabric Setup: A multi-cloud enterprise testbed is constructed using Kubernetes clusters distributed across three cloud environments (AWS EKS, Azure AKS, and Google Cloud GKE) alongside an on-premises PostgreSQL and MinIO environment. A federated control plane (built using Open Policy Agent and Apache Atlas) is deployed to manage global policy definitions and metadata indexing without aggregating raw tenant data. Synthetic and real-world benchmark enterprise datasets—containing structured SQL tables, semi-structured JSON logs, and unstructured PDF policy documents with embedded PII and regulatory attributes—are distributed across the nodes.

Localization of Generative AI Agents and RAG Module Deployment: Lightweight open-weights generative models (e.g., Llama 3 8B, Mistral 7B) are containerized and deployed directly onto localized cloud nodes as governance sidecars. A localized Retrieval-Augmented Generation (RAG) pipeline is implemented at each node using a vector store (Qdrant). The RAG pipeline indexes localized enterprise data dictionaries, historical schema mappings, and regional compliance mandates (GDPR, HIPAA), allowing localized LLM agents to process metadata, generate canonical schema representations, and infer data sensitivity without transferring underlying raw data out of the hosting region.

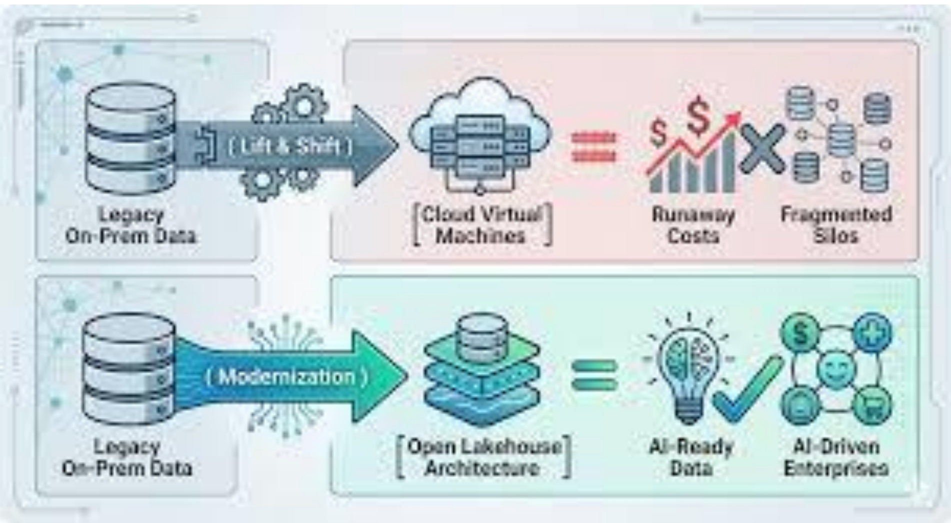

FIG1: Enhancing Federated Cloud with Explainable Generative Artificial Intelligence

Explainability Engine Integration and Feature Attribution Mechanics: To ensure model outputs are interpretable and auditable, an Explainability Module is integrated into the generative pipeline. When an xGenAI agent assigns a data sensitivity classification, executes a schema transformation, or generates a data access rule, a companion attribution engine processes the prompt-token weights using Integrated Gradients and SHAP token-level scoring. The engine outputs a structured, human-readable JSON "Explanation Manifest" that explicitly details:

- Key input tokens and metadata fields that triggered the classification.
- Confidence scores and underlying regulatory policy clauses fetched by RAG.
- Graph-based lineage attributions showing source-to-target mapping logic.

Empirical Evaluation, Stress Testing, and Benchmarking: The integrated framework is subjected to automated modernization and governance scenarios, comparing performance against conventional governance toolchains. Quantitative metrics are evaluated across key dimensions:

- **Auto-Cataloging Precision & Recall:** Accuracy of PII and business attribute identification across unstructured/structured sources.
- **Modernization Throughput:** Speed of legacy schema translation and automated ETL generation.
- **Policy Audit Latency:** Time required to evaluate, explain, and approve dynamic data access requests across federated nodes.
- **Explainability Audit Score:** Evaluation by human compliance experts on the clarity, correctness, and auditability of generated explanation manifests.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

Advantages

- **Automated Semantic Harmonization:** Generative AI agents rapidly analyze and unify disparate, legacy schemas into standardized enterprise data products without requiring manual, time-consuming metadata tagging.
- **Verifiable and Auditable Compliance:** Explainability mechanics provide clear attribution trails for every automated governance decision, enabling compliance officers to audit AI-generated classifications and satisfy strict regulatory requirements.
- **Privacy-Preserving Governance:** Localized federated execution ensures that sensitive data stays within its regional cloud boundary, fulfilling strict data sovereignty laws while maintaining global visibility.
- **Accelerated Data Modernization:** Automated generation of ETL pipelines, data documentation, and SQL queries speeds up the transformation of legacy data silos into analytics-ready cloud assets.

Disadvantages

- **High Computational Infrastructure Costs:** Deploying local generative models and vector databases across multiple cloud nodes increases compute, memory, and specialized hardware (GPU/TPU) requirements.
- **Increased System Latency Overhead:** Generating real-time feature attributions (SHAP/Integrated Gradients) alongside LLM inference adds computational latency compared to basic rule-based scanners.
- **Cold-Start Knowledge Base Dependency:** The quality of RAG-driven classification depends heavily on the accuracy and completeness of the enterprise's initial policy documentation and vector store indexing.
- **Complex Cross-Cloud Attribution Tuning:** Calibrating attribution thresholds and managing prompt drift across distinct generative model sidecars in heterogeneous cloud environments requires specialized operational expertise.

REFERENCES

1. Sudhan, S. K. H. H., & Kumar, S. S. (2016). Gallant Use of Cloud by a Novel Framework of Encrypted Biometric Authentication and Multi Level Data Protection. *Indian Journal of Science and Technology*, *9*, 44.
2. Gujarathi, M. (2024). Monolith-to-microservices migration in enterprise operations platforms: A workflow-oriented architecture. *International Journal of Research Publications in Engineering, Technology and Management, 7*(1), 10018–10029.
3. Raja, G. V. (2023). AI Driven Secure Intelligent Framework for Fraud Detection Cybersecurity and Cloud Based Enterprise Systems. *International Journal of Advanced Research in Computer Science & Technology (IJARCST), 6*(5), 9068-9076.
4. Potdar, A., Gottipalli, D., Ashirova, A., Kodela, V., Donkina, S., & Begaliev, A. (2025, July). MFO-AIChain: An Intelligent Optimization and Blockchain-Backed Architecture for Resilient and Real-Time Healthcare IoT Communication. In *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)* (pp. 1-6). IEEE.
5. Challa, R. (2023). Engineering AI Factories: Scalable HPC Design Patterns for Next-Generation Foundation Model Infrastructure. *International Journal of Computer Technology and Electronics Communication, 6*(4), 7342-7351.
6. Venkiteela, P. (2024). Strategic API modernization using Apigee X for enterprise transformation. *Journal of Information Systems Engineering and Management, 9*(4s), 14. https://jisem-journal.com/index.php/journal/article/view/13168
7. Billah, M. M., Nath, A. D., Das, D., Mahmud, T., & Rahman, R. (2023). Skin Cancer Classification using NasNet. *World J. Adv. Res. Rev, 19*, 1652-1658.
8. Soundappan, S. J. (2022). Blockchain Enabled Zero Trust Security Architecture for Cloud Native Distributed Applications. *International Journal of Emerging Trends in Engineering and Management Research, 7*(4), 12167.
9. Meesala, A. (2024). Machine Learning Enabled Governance Framework for Autonomous Enterprise Platforms and Intelligent Data Ecosystems. *International Journal of Computer Technology and Electronics Communication, 7*(6), 9899-9909.
10. Nanagowda, B., & Kumar, S. (2025). Publishing Of Reports Via Camunda Workflow Orchestration for A Financial Institute. *Advances in Consumer Research, 2*(4).
11. Sudhan, S. K. H. H., & Kumar, S. S. (2015). An innovative proposal for secure cloud authentication using encrypted biometric authentication scheme. *Indian Journal of Science and Technology, 8*(35), 1-5.
12. Narayanan, S. (2023). Cloud-native generative artificial intelligence for autonomous third-party risk intelligence: A zero-trust supply chain assurance framework. *International Journal of Computer Engineering and Technology, 14*(1), 283–297. https://philarchive.org/archive/NARCGA
13. Konakalla, K. (2024). Exploring the functionality of omnichannel in Salesforce Service Cloud to improve agent efficiency. *International Journal of Future Multidisciplinary Research (IJFMR), 6*(1).

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

14. Gopinathan, V. R. (2023). Intelligent Cloud Security through Continuous Threat Detection and Risk Assessment. *International Research Journal of Innovative Engineering*, *7*(6), 13571-13581.
15. Thota, S. K., & Anumula, S. K. (2024). Quantum-Enabled Drones for Battlefield Information Dominance: Integrating Sensing, Computing, and Secure Communications. *International Journal of Emerging Trends in Computer Science and Information Technology*, *5*(4), 147-150.
16. Sojol, J. I., Alam, M. S., Hossain, N., & Motahar, T. (2019, February). Smart School Bus: Ensuring Safety On The Road For School Going Children. In *2019 21st International Conference on Advanced Communication Technology (ICACT)* (pp. 734-739). IEEE.
17. Kale, P. (2024). AI-Augmented DevSecOps Pipelines for Secure and Efficient Software Delivery in Cloud-Native Platforms. *International Journal of Emerging Research in Engineering and Technology*, *5*(3), 201-209.
18. Kanji, M. R. K. (2022). A Unified Data Warehouse Architecture for Multi-Source Forest Inventory Integration and Automated Remote Sensing Analysis. *Journal Of Engineering And Computer Sciences*, *1*(5), 10-16.
19. Anand, L. (2025). Modernizing Enterprise Systems through Generative AI Autonomous Operations and Cloud-Native Engineering. *International Journal of Humanities and Information Technology*, *7*(02), 54-69.
20. Mohammed, S. (2021). Hybrid cloud architecture strategy for global infrastructure operations. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *3*(6), 4078-4081.
21. Sugumar, R. (2023, September). A Novel Approach to Diabetes Risk Assessment Using Advanced Deep Neural Networks and LSTM Networks. In *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)* (pp. 1-7). IEEE.
22. Juvvadi, R. R. (2024). Embedding ESG and carbon accounting into the financial ledger: A sub-ledger architecture for CSRD-aligned reporting. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *6*(1), 7531–7535.
23. Dasari, H. P., & Kesarpu, S. (2025, August). Kafka event sourcing for real-time risk analysis. *International Journal of Computational and Experimental Science and Engineering*, *11*(3), 6012–6018. https://doi.org/10.22399/ijcesen.3715
24. Meesala, L. K. (2024). Converging Infrastructure and Security: A Maturity-Based Approach to Cloud-Native Data Protection, SIEM Optimization, and Compliance Automation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *5*(1), 283-289.
25. Chaba, A. (2025). Agent Orchestration: A New Paradigm for Autonomous and Scalable MarTech Ecosystems. *ISCSITR-International Journal of Computer Science and Engineering (ISCSITR-IJCSE)*, *6*(4), 63–76.
26. Akter, K. S., Onik, T. A., Yasin, M., Nabil, M. A., Hossain, I., & Akter, S. (2024). AI-Driven Early Detection of Chronic Kidney Disease: A Comparative Benchmark of Ensemble Learning Models with Clinical Explainability. *Vascular and Endovascular Review*, *7*(2), 480-493.
27. Seetala, S. R. (2023). Real-Time Data Monitoring Using Cloud Observability Tools: Architectures, Techniques and Emerging Practices. *J Artif Intell Mach Learn & Data Sci*, *6*(4), 3367-3374.
28. Hossain, M. S., Ali, M., & Haider, P. S. (2022). Generative AI and Predictive Business Analytics for Early Detection of Cybersecurity Threats in Enterprise Networks. *American Journal of Economics and Business Management*, *5*(12).
29. Prasanna Kumar Natta. (2022). AI-driven inventory intelligence for large-scale retail operations: A framework for real-time store-level stock accuracy. *International Journal of Advanced Engineering Science and Information Technology*, *5*(2), 8740–8751. https://doi.org/10.15662/IJAESIT.2022.0502002
30. Mudusu, S. K. (2025). Data Engineering Challenges in AI-Driven Healthcare IT Systems: Navigating Real-Time Analytics and Interoperability.
31. Rohit Wadhwa. (2024). Designing Event-Driven Enterprise Systems with Distributed Data Sharding and Partitioning Strategies. *ISCSITR- International Journal of Computer Applications (ISCSITR-IJCA)*, *5*(1), 22–35.
32. Vollem, S. (2017). An architectural and strategic analysis of enterprise-scale re-engineering approaches for modernizing legacy financial systems through Java-centric software paradigms and intelligent cloud automation frameworks. *International Journal of Scientific Research in Science, Engineering and Technology*, *3*(3), 878-896.
33. Vasa, M. R. (2025). Cloud-Oriented Deep Learning Models for Smart Healthcare Automation and Predictive Risk Analytics. *International Journal of Future Innovative Science and Technology (IJFIST)*, *8*(4), 15306.
34. Sikarwar, V. (2025). AI-Augmented in Enterprise Domain Modeling and its impact on Data Modernization projects. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *7*(3), 9944-9952.
35. Alex Mathew. (2023). Threat defense through cyber fusion. *International Journal of Computer Science and Mobile Computing*, *12*(1), 24–27. https://doi.org/10.47760/ijcsmc.2022.v12i01.003
36. Gunda, S. R. (2025). Optimizing Cloud Computing Resource Utilization Through Intelligent Allocation and Containerization Strategies. *Journal of Computer Science and Technology Studies*, *7*(8), 238-244.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805036

37. Singh, I. K. (2024). Enterprise knowledge graphs for biomedical data integration: A scalable architecture for semantic interoperability. *International Journal of Engineering & Extended Technologies Research (IJEETR), 6*(4), 8165–8176.
38. Jayaraman, S., Rajendran, S., & P, S. P. (2019). Fuzzy c-means clustering and elliptic curve cryptography using privacy preserving in cloud. *International Journal of Business Intelligence and Data Mining, 15*(3), 273-287.