Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 7s (2025)

Federated Data Lakehouse Architecture for Multi-Cloud Governance, AI and Analytics

Saurabh Mishra

Pt. Ravishankar Shukla University, Raipur, India

saurabhkumar.mishra@gmail.com

Article History:

Received:10-06-2025

Revised:27-07-2025

Accepted:10-08-2025

Abstract:
Enterprise data architecture requirements have enhanced the pace at which scalable enterprise data architecture is being unified in a manner that is secure, intelligent, and entailed across diverse cloud environments. The federated data lakehouse architecture is a paradigm that integrates the flexibility of data lakes with the structure and management capabilities of data warehouses, while being designed for use across multi-cloud platforms. This review explains how federated lakehouses can facilitate data governance, enable distributed collaboration with AI, and support real-time analytics, while also addressing challenges related to data residency, compliance, and interoperability. The paper provides a comprehensive view of architectural components within cloud computing, data engineering, and artificial intelligence through the convergence of these domains using a federated perspective, coordinated operations, and the advantages inherent in this model. Drawing on recent literature, the article reveals the transformational potential of federated lakehouses in re-engineering enterprise information systems to align with the emerging era of decentralization and intelligent automation.

Keywords- Federated Lakehouse, Multi-Cloud Governance, Artificial Intelligence, Data Engineering

1. Introduction

The rise of cloud computing has provoked dynamism in enterprise data architecture, and the increasing volume of data has further accelerated the pace of change. Traditional monolithic data platforms have proven ineffective in handling the velocity, volume, and variety of data currently encountered, especially within complex multi-cloud environments. Federated data lakehouse designs have emerged to address these complexities, offering a synthesized paradigm that combines the flexibility of data lakes with the performance and structure of data warehouses. These architectures are particularly well-suited for deployment in multi-cloud settings, where governance, artificial intelligence (AI), and analytics must be effectively harmonized. This paper reviews how federated data lakehouse architectures have evolved and scaled, particularly in delivering robust data governance, AI integration, and analytics capabilities across multi-cloud systems.

2. Evolution from Data Lakes to Federated Lakehouses

Businesses embraced data lakes at the onset of the big data management era since they enabled the storage of vast quantities of both unstructured and structured data. However, data lakes were not imposed with predefined schemas and strong governance, which often led to the emergence of data swamps. The concept of lakehouses addressed these limitations by combining the governance and reliability of data warehouses with the flexibility of data lakes.

Lakehouses facilitate ACID transactions, schema enforcement, and fine-grained access control, and are therefore more appropriate for enterprise workloads.

Federated data lakehouses extend this model across multiple clouds, allowing organizations to pool disparate datasets stored in AWS, Azure, Google Cloud, and on-premises environments. This is especially applicable in a landscape where data residency requirements, performance optimization, and cost control compel organizations to decentralize their data resources. Federated architectures enable each data lakehouse instance to remain semi-autonomous, while adhering to shared governance and metadata frameworks to ensure consistency and interoperability [1] [17] .

This evolution is supported by the broader shift toward data fabric and data mesh architectures, which emphasize decentralized ownership, semantic integration, and domain-oriented data products. Federated data lakehouses therefore serve as a structural foundation for realizing these architectural paradigms, enabling organizations to manage data as a strategic asset within increasingly heterogeneous environments [1].

3. Multi-Cloud Integration and Big Data Workflow Management

"Multi-cloud strategies are now essential for organizations seeking to avoid vendor lock-in, leverage best-in-class services from various providers, and comply with diverse jurisdictional regulations. Such strategies align well with the federated data lakehouse model, which supports data processing and integration across heterogeneous cloud infrastructures. One of the key requirements in this context is the coordination of big data processes across multiple cloud environments.

These processes often involve complex Extract, Transform, Load (ETL) operations, real-time data ingestion, and cross-cloud data synchronization. Federated architectures enable the creation of virtualized data views that span cloud boundaries, helping to minimize data replication and reduce latency issues. Additionally, they support the execution of containerized data processing pipelines using technologies such as Kubernetes and Apache Spark, which can be orchestrated through multi-cloud data flow technologies [2] [19].

Workflow governance within decentralized cloud environments remains a significant challenge. Each cloud provider offers native tools for data security, auditing, and lineage tracking. Harmonizing these disparate systems into a unified federated governance layer requires standardized metadata models, cross-cloud identity federation, and centralized policy management systems. Federated data lakehouses support these needs by offering centralized metadata catalogs and unified policy enforcement points that operate across diverse cloud platforms [2].

Figure 1 below illustrates a typical federated data lakehouse architecture deployed across multiple clouds, highlighting the central metadata layer, AI integration modules, and analytic engines operating across federated domains.

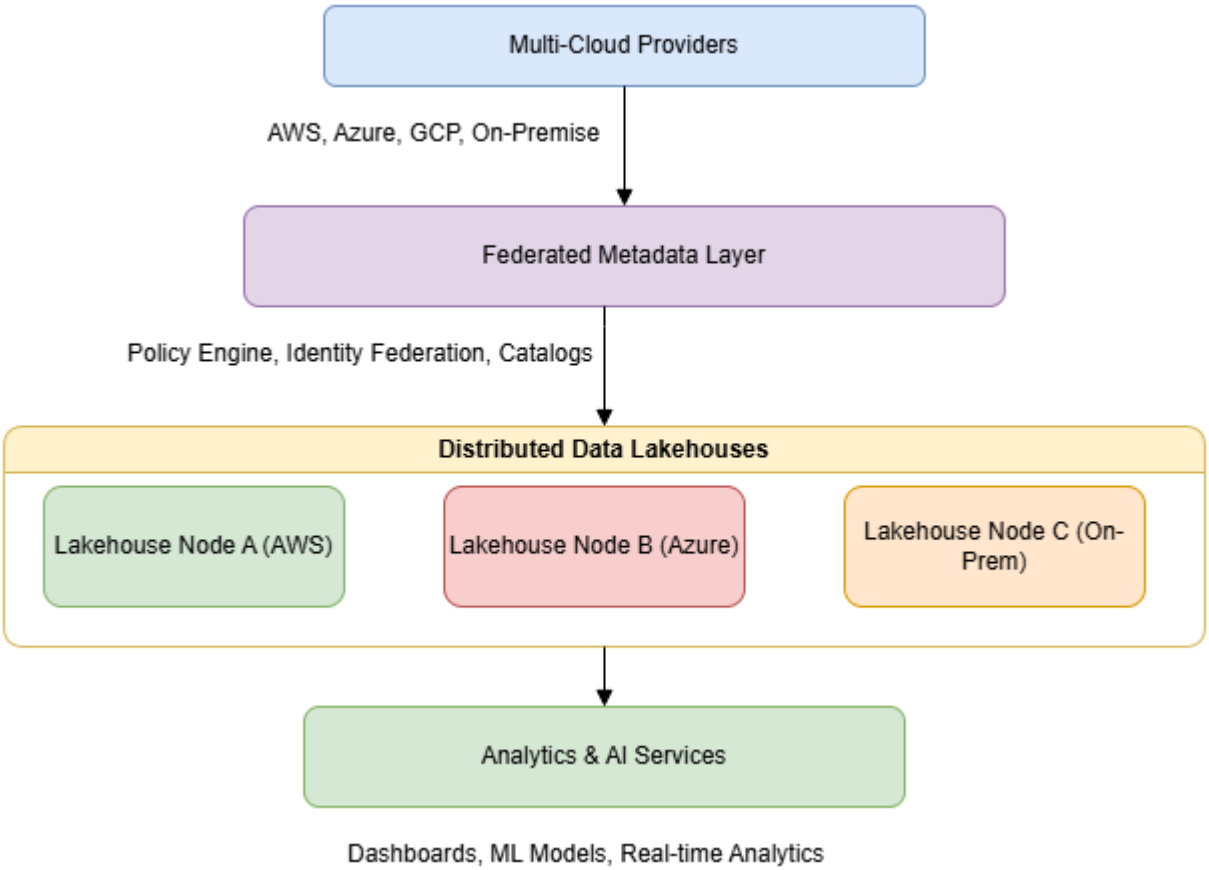

Figure 1: Federated Data Lakehouse Architecture Across Multi-Cloud Platforms

Source: Adapted and conceptualized based on information from sources [1], [2]

4. Governance, Security, and Compliance in Federated Architectures

One of the most important pillars of any modern data architecture is data governance. Governance in multi-cloud environments poses additional challenges due to jurisdictional constraints, non-homogeneous security models, and disparate compliance frameworks. Federated data lakehouse designs address these challenges by separating data governance policies from the underlying physical infrastructure. This enables centralized governance to be defined globally while being enforced locally within each cloud provider environment.

Role-based access control, data masking, and encryption of data at rest and in transit are security mechanisms that can be applied uniformly across the federated system. Identity management tools that help establish a unified security posture include federated single sign-on and cross-cloud identity mapping, which support the implementation of advanced identity and access management systems. In addition, federated architectures enable real-time data lineage tracking and audit logging, which are essential for regulatory compliance in the financial, healthcare, and government sectors [3].

The architecture also supports policy-based data residency controls, allowing organizations to specify where data may be stored and processed. This capability is critical for compliance with regulations such as GDPR, HIPAA, and CCPA. Furthermore, federated lakehouses incorporate

AI-driven governance tools, including automated anomaly detection, sensitive data classification, and data quality checks, which reduce the operational burden on data stewards [12].

The built-in availability of these tools within federated lakehouse platforms ensures that data is not only accessible and analyzable but also secured and governed in accordance with an organization's risk management and compliance strategies [3].

5. Federated Machine Learning and Collaborative AI in Multi-Cloud

The growing adoption of artificial intelligence (AI) and machine learning (ML) is driving exponentially increasing demands for access to high-quality data of large volume and variety. These requirements can result in redundant data transfers and heightened security concerns when using traditional centralized architectures. Such challenges are addressed by federated data lakehouses through federated machine learning (FML), which leverages the ability to train models on decentralized data without requiring that data to be transferred from its source.

In federated ML, data stored within autonomous cloud environments is used to train local models. These models are then aggregated to form a global model, preserving data privacy while harnessing the collective intelligence of distributed data sources. This approach is particularly valuable in sectors that prioritize data sovereignty, including healthcare, finance, and government [4] [14] .

Federated architectures provide the computational infrastructure and coordination mechanisms required for FML, including support for secure multi-party computation and differential privacy. The underlying layers of the lakehouse ensure the availability of high-quality training data that is well governed and consistent across nodes. Artificial intelligence accelerators, such as GPUs and TPUs, can be adaptively deployed across clouds depending on model complexity and training requirements.

Such collaborative AI models enable predictive analytics, fraud detection, and personalized recommendations without violating data locality requirements. Federated AI also promotes inter-organizational collaboration, whereby multiple organizations can jointly develop and benefit from shared AI models while retaining control over their proprietary datasets [4].

The federated lakehouse model therefore serves as a key enabler of ethical, scalable, and collaborative AI in multi-cloud environments due to its decentralized yet coordinated approach to AI development.

TABLE 1: Key Features Comparison – Traditional Data Warehousing vs. Federated Lakehouse in Multi-Cloud Environments

Data Storage Format	Structured only	Structured and unstructured
Schema Enforcement	Rigid	Flexible with schema evolution
Scalability	Vertical	Horizontal, cloud-native

Data Governance	Centralized	Federated with centralized policies
Cloud Compatibility	Limited	Multi-cloud and hybrid support
Support for Machine Learning	Minimal	Built-in, federated learning support
Real-time Analytics	Limited	Fully supported
Cost Efficiency	High (licensing, storage)	Optimized through cloud-native tools
Data Sovereignty & Compliance	Complex to enforce	Natively supported across regions
AI Integration	External tools needed	Seamless with built-in accelerators

Source: Compiled and interpreted from references [1] to [4]

6. Redefining ETL and Data Engineering in Hybrid Cloud Setups

Federated lakehouse architectures are also redefining traditional ETL processes, which have long been central to data warehousing. Traditional ETL pipelines involve transforming data midstream before loading it into a centralized data warehouse. However, this model is no longer viable in multi-cloud environments, where data is geographically and logically distributed.

In contrast, the federated lakehouse enables ELT (Extract, Load, Transform) pipelines, where raw data is first loaded into lakehouse layers across different cloud platforms and then transformed. Distributed compute resources are executed directly within these lakehouse layers. This significantly reduces data movement and improves operational efficiency. It also allows data engineers to develop modular and reusable transformation logic using tools such as Apache Airflow, dbt, and native cloud services [5] [20].

Modern data engineering now emphasizes data observability, pipeline resiliency, and real-time processing. Federated architectures offer built-in support for schema evolution, versioning, and rollback capabilities, enabling pipelines to adapt flexibly to schema changes or processing failures. Moreover, they support parallel data processing across nodes, thereby improving time-to-insight [5].

These capabilities are especially critical for AI and analytics workloads, which require clean, current, and well-modeled data. The convergence of data engineering and federated architectures underscores the need for scalable, efficient data pipelines that are fully integrated into the governance and AI layers of the enterprise data platform [5].

7. Convergence of Cloud Computing and Data Warehousing

The need for scalable, intelligent, and cost-effective data solutions has driven the integration of cloud computing and data warehousing, marking a significant architectural shift in enterprise technology. Traditional data warehouses were typically hard-coded and lacked scalability; however, with the advent of cloud computing, data platforms have become dynamic and

capable of scaling to petabyte sizes with minimal overhead. This convergence extends to the federated lakehouse model, which incorporates enhanced cloud-native services into the core data stack.

Within this model, each federated node of a multi-cloud lakehouse can operate independently, utilizing cloud-native serverless query engines, self-scaling compute clusters, and controlled storage solutions. This simplifies infrastructure provisioning and maintenance, and is highly cost-effective. It also enables data practitioners to work closer to the data source and benefit from in-region processing to meet data residency and latency requirements [6] [21] .

Moreover, this convergence enables a tighter integration of DevOps and DataOps pipelines, supporting continuous integration and continuous deployment (CI/CD) of data artifacts. Data transformations, models, and machine learning pipelines can be tested, monitored, and deployed at scale. The federated lakehouse architecture supports this through centralized observability, lineage tracing, and dependency mapping across cloud environments.

These capabilities are essential for ensuring agility in data-driven decision-making. The dynamic, predictive, and intelligent nature of modern data warehousing can be transformed into a reactive system component by leveraging the compute elasticity and service diversity of the cloud [6].

8. AI-Driven Insights and Intelligent Modernization

One of the main components of enterprise modernization is the federated lakehouse, and artificial intelligence has increasingly become a central element of these strategies. By enabling seamless access to distributed, high-quality data across geographies and business units, federated architectures support real-time analytics, natural language processing, computer vision, and predictive modeling. Cloud-native AI frameworks such as TensorFlow, PyTorch, and Scikit-learn can be trained and deployed within lakehouse environments and configured to support auto-scaling and distributed training [7] [13] .

The close integration between AI workloads and federated data lakehouses also enables the continuous retraining and deployment of models, which is essential for keeping models up to date in dynamic business environments. Real-time feedback loops can be established, whereby model predictions influence downstream processes and the resulting outcomes are fed back into training datasets. This allows enterprises to deploy real-time decision intelligence systems that evolve in response to changes in market conditions, operations, or customer behavior.

Another important advantage is the ability to use AI to optimize the operation of the lakehouse architecture itself. AI-powered optimization engines can automatically allocate compute resources based on workload patterns, identify data quality issues, and recommend schema optimizations. These intelligent capabilities reduce the need for manual intervention while improving system stability and performance [7].

Notably, federated AI models trained within these architectures are inherently more scalable and privacy-aware, addressing the growing demand for responsible AI. Responsible AI

frameworks can be integrated to ensure model outputs are explainable, biases are mitigated, and decision-making processes remain transparent across the organization [7].

9. Institutional Data Management with Data Lakehouses

Data repositories in institutions, particularly academic and research institutions, are increasingly adopting data lakehouse architectures to enhance scalable, compliant, and accessible data management. These organizations generate vast amounts of research, experimentation, and collaborative work across departments, often spanning national borders. A federated lakehouse model enables these institutions to achieve decentralized data ownership while maintaining centralized compliance and access control.

Structured metadata management, version control, and persistent identifiers are features provided by lakehouses and are indispensable to institutional repositories. Lakehouses support a more holistic approach to data stewardship by accommodating both structured and unstructured data types, including spreadsheets, genomic sequences, and multimedia files [8] [11] .

Federated deployment of lakehouses also enhances the ability of institutions to collaborate globally without violating data sovereignty or intellectual property agreements. Each institution or department maintains its own data node while participating in a broader data-sharing consortium. A centralized governance policy ensures that data usage aligns with ethical guidelines and legal requirements, including those established by research ethics boards and funding bodies [18].

Additionally, lakehouses can be integrated with institutional authentication mechanisms such as LDAP and SAML to allow secure, role-based access to datasets. This integration ensures that only the data relevant to a researcher, administrator, or external partner is accessible, thereby mitigating the risk of data leakage or misuse.

This architectural approach has transformed research data management by aligning with the scalability, governance, and accessibility needs of knowledge institutions, offering a forward-looking solution equipped to meet future demands [8].

10. Scalable Data Engineering and Cloud ETL Pipelines

Set to become more complex in multi-cloud environments, federated lakehouse architecture offers a simplified and extensible approach to managing ETL and ELT pipelines through cloud-native data engineering. It is imperative to develop stable and efficient pipelines capable of ingesting, processing, and serving data within distributed systems. The use of abstraction and automation in orchestrating pipelines across clouds facilitates scalability in this context [9].

Modern ETL pipelines increasingly utilize advanced tools such as Apache Beam, Google Cloud Dataflow, and AWS Glue to manage complex tasks. These tools are natively integrated with schema evolution, schema error handling, and parallel processing in lakehouse storage layers. Federated architectures support these workflows by enabling seamless data movement and transformation across geographically dispersed storage nodes, while ensuring consistency through synchronized metadata [15].

Data pipeline observability is also enhanced through centralized dashboards that provide insights into job performance, error rates, and data freshness. These capabilities allow data teams to proactively detect and respond to anomalies, thereby ensuring the health of both the pipelines and the data. Furthermore, lakehouses enable dynamic resource allocation based on workload demands, balancing cost and performance effectively.

Another key advantage is integration with streaming systems such as Apache Kafka and Amazon Kinesis, which are essential for real-time analytics. Federated lakehouses support both batch and stream processing, eliminating the traditional architectural divide between OLAP and OLTP systems, and providing a unified approach to data engineering and analytics [9].

The use of Infrastructure as Code (IaC) and containerization technologies like Docker and Kubernetes further enhances reproducibility and cross-environment portability—capabilities that are increasingly demanded by enterprises operating in regulated and dynamic industries.

11. Advancements in Data Stack Architectures and BI Integration

The success of federated lakehouse architectures has been closely tied to the evolution of the modern data stack. Unlike classical monolithic data platforms, the modern data stack consists of interoperable, modular components that can be deployed in cloud environments. These components include data ingestion tools (e.g., Fivetran, Talend), storage solutions (e.g., Delta Lake, Apache Iceberg), transformation tools (e.g., dbt), orchestration frameworks (e.g., Airflow), and business intelligence platforms (e.g., Tableau, Power BI, Looker).

Federated lakehouses integrate these tools within a unified architecture that supports end-to-end data operations in the cloud. This modularity allows businesses to adopt a plug-and-play model, selecting the most suitable tool for each function while ensuring compatibility through standardized APIs and metadata schemas [10].

One of the major advantages of this architecture is the ability to deliver business intelligence and data visualization at scale. BI tools can present information to decision-makers accurately and in real time, thanks to unified access to clean, well-modeled data across cloud environments. Data models structured within lakehouses help maintain semantic consistency and reusability, ensuring that business metrics carry the same meaning across departments and geographic regions.

Self-service analytics portals—where non-technical users can query data, build reports, and browse dashboards without assistance from data engineers—also support the democratization of data access. This approach leads to faster time-to-insight and empowers domain experts to make informed decisions using real-time data [16].

Figure 2 below illustrates the adoption trend of federated lakehouse architectures over time, indicating increased uptake across sectors such as healthcare, finance, retail, and public services.

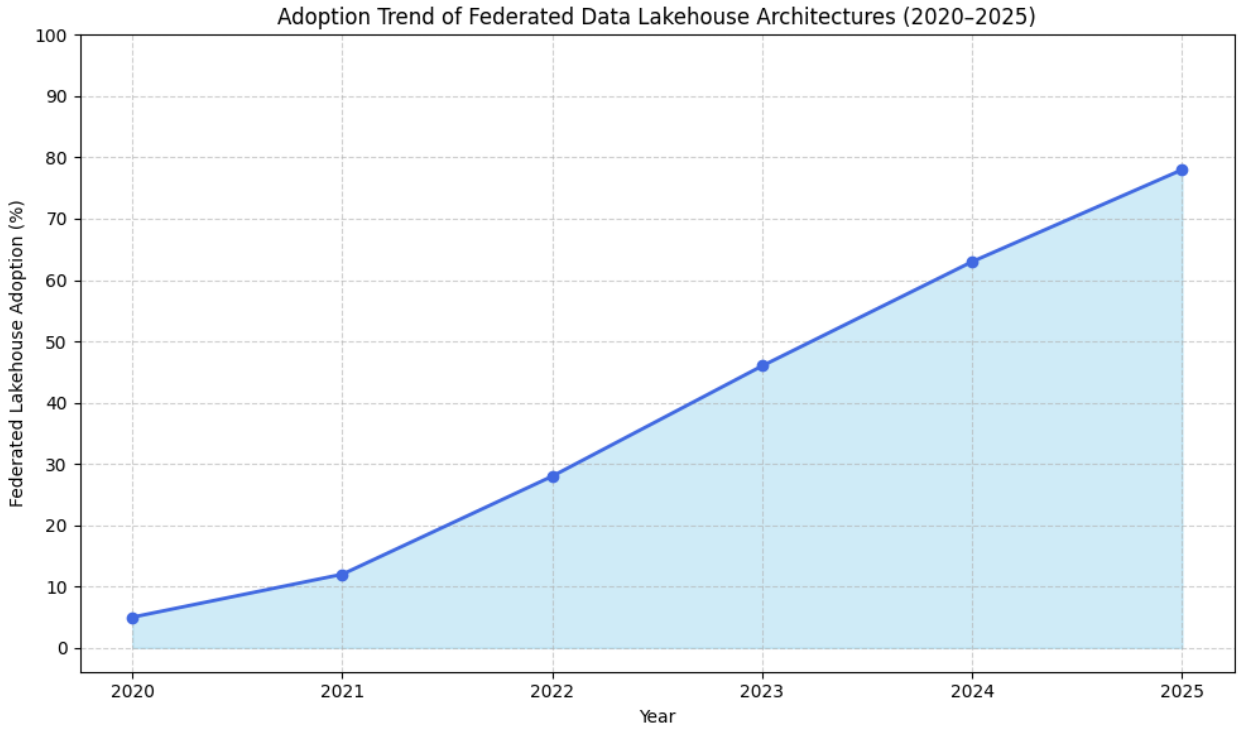

Figure 2: Adoption Trend of Federated Data Lakehouse Architectures (2020–2025)

Source: Synthesized from data trends and market analysis based on [10]

12. Conclusion

The federated data lakehouse is an emerging paradigm in enterprise data architecture, developed to address the challenges of multi-cloud governance, AI enablement, and data-driven decision-making. Federated lakehouses assist organizations in operating their data assets more effectively by combining the scalability of data lakes with the structure of data warehouses and extending these capabilities across diverse cloud environments. They enable strong governance controls, efficient data engineering, scalable AI workloads, and real-time analytics within a unified yet decentralized architecture. As organizations continue to navigate digital transformation, the federated data lakehouse represents a forward-looking solution for building intelligent, compliant, and agile data environments.

References

[1] Eeti, E. S. (2024). Architectural patterns for big data analytics in multi-cloud environments. The International Journal of Engineering Research, 8 (3), 16-25.[TIJER](tijer tijer/viewpaperforall. php? paper= TIJER2103003).

[2] Peter, H. (2024). Multi-Cloud Data Lake Architecture for Scalable AI/ML Model Deployment.

[3] Polisetty, S. M. (2024). CLOUD-NATIVE LAKEHOUSES: MULTI-CLOUD STRATEGIES FOR BUSINESS INTELLIGENCE AND DATA ANALYTICS. Technology (IJRCAIT), 7(1).

[4] Merseedi, K. J., & Zeebaree, S. R. (2024). The cloud architectures for distributed multi-cloud computing: a review of hybrid and federated cloud environment. The Indonesian Journal of Computer Science, 13(2).

[5] Levandoski, J., Casto, G., Deng, M., Desai, R., Edara, P., Hottelier, T., ... & Volobuev, Y. (2024, June). BigLake: BigQuery's evolution toward a multi-cloud lakehouse. In Companion of the 2024 International Conference on Management of Data (pp. 334-346).

[6] Perugu, P. K. (2024). AI-Driven Solutions for Data Governance in Multi-Cloud Ecosystems. Available at SSRN 5119378.

[7] Bhat, J. (2024). Designing Enterprise Data Architecture for AI-First Government and Higher Education Institutions. International Journal of Emerging Research in Engineering and Technology, 5(3), 106-117.

[8] Sundar, D. (2024). Enterprise Data Mesh Architectures for Scalable and Distributed Analytics. American International Journal of Computer Science and Technology, 6(3), 24-35.

[9] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2024). A Federated Zero-Trust Security Framework for Multi-Cloud Environments Using Predictive Analytics and AI-Driven Access Control Models. International Journal of Emerging Research in Engineering and Technology, 5(2), 95-107.

[10] Dolhopolov, A., Castelltort, A., & Laurent, A. (2024). Implementing federated governance in data mesh architecture. Future Internet, 16(4), 115.

[11] Christopher, A. J. (2024). Unified Apache-Based AI Cloud Analytics for Financial Intelligence across Smart Waste Management and Healthcare. International Journal of Research and Applied Innovations, 7(5), 11410-11418.

[12] Rajasekar, M. (2024). Secure Digital Banking with Federated AI: An AWS Cloud-Based Predictive Analytics Architecture for Financial Risk Intelligence. International Journal of Research and Applied Innovations, 7(3), 10735-10740.

[13] Adelusi, B. S., Ojika, F. U., & Uzoka, A. C. (2024). Advances in Scalable, Maintainable Data Mart Architecture for Multi-Tenant SaaS and Enterprise Applications.

[14] Jiang, C., Liu, S., & Ma, P. (2024, December). Innovative Architecture and Key Technology of Database for Multiple Scenarios. In 2024 International Conference on Ubiquitous Computing and Communications (IUCC) (pp. 286-291). IEEE.

[15] Deswandikar, A. (2024). Engineering Data Mesh in Azure Cloud: Implement Data Mesh Using Microsoft Azure's Cloud Adoption Framework. Packt Publishing Ltd.

[16] Schiller, R. J., & Larochelle, D. (2024). Data Engineering Best Practices: Architect robust and cost-effective data solutions in the cloud era. Packt Publishing Ltd.

[17] Cross-Platform Analytics Harmonization in Multi-Tenant Retail Environments Using Adobe and Tealium [Eshita Gupta. (2025). Cross-Platform Analytics Harmonization in

Multi-Tenant Retail Environments Using Adobe and Tealium. International Journal of Computational and Experimental Science and Engineering, 11(4). https://doi.org/10.22399/ijcesen.4122]

[18] Designing Scalable Multivariate Testing Frameworks for High-Traffic Ecommerce Platforms [Gupta, E. (2025). Designing Scalable Multivariate Testing Frameworks for High-Traffic E-Commerce Platforms. International Journal of Basic and Applied Sciences, 14(8), 167-173. https://doi.org/10.14419/47mq5944]

[19] Enabling Analytics Governance in Agile Product Teams: A Scalable Tagging and QA Framework [Gupta, E. (2025). ENABLING ANALYTICS GOVERNANCE IN AGILE PRODUCT TEAMS: A SCALABLE TAGGING AND QA FRAMEWORK. International Journal of Applied Mathematics, 38(7s), 1161-1172.]

[20] Torres, D. J. G. (2024). AI-Enhanced DevOps Pipeline for Real-Time Patient Monitoring: Leveraging Databricks Data Intelligence and SAP-Integrated Cloud Workloads. International Journal of Computer Technology and Electronics Communication, 7(6), 9770-9774.

[21] Bukhari, T. T., Oladimeji, O., Etim, E. D., & Ajayi, J. O. (2024). Cloud-native business intelligence transformation: Migrating legacy systems to modern analytics stacks for scalable decision-making. International Journal of Scientific Research in Humanities and Social Sciences, 1(2), 744-762.