Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 05 Issue: 06 | June-2025

Optimizing Multicloud Data Lakes with Delta Lake and Unity Catalog for Regulated Industries

Shivam Jawla[1], Rohan Patel[2], Santosh Patel[3], Sadhana Dubey[4]

[1]*ITC Infotech , Bengaluru,India, shivam.7744@gmail.com*
[2]*Accenture, Tokyo,Japan, Rohan.patel@teknikoz.com*
[3]*IBM, Shenzhen,China, Santy.patel@gmail.com*
[4]*HSBC,Guangzhou, China, Dubeysadhana@gmail.com*

Abstract -The transition of highly regulated industries towards a multi-cloud approach for enhanced resilience, scalability, and vendor-neutral data architectures highlights the critical importance of implementing a unified, compliant, and high-performance data fabric. This paper explores Delta Lake and Unity Catalog on Databricks as methods for creating scalable, secure, and compliance-ready data lakes on AWS and Azure. We examine architectural forms, governance systems, and performance enhancement strategies suitable for finance, healthcare, and governmental entities.

Key Words: Multi-cloud, compliance, Delta Lake, Unity Catalogue, Governance.

1.INTRODUCTION

This document shows the suggested format and appearance of a manuscript prepared for SPIE journals. Accepted papers will be professionally typeset. This template is intended to be a tool to improve manuscript clarity for the reviewers. The final layout of the typeset paper will not match this template layout.

1.1 Multicloud Takes Off in Regulated Industries

Regulated industries present unique data management challenges, such as compliance (e.g., HIPAA, GDPR, SOX), data residency restrictions, and high availability. Multicloud strategies offer flexibility and risk mitigation but introduce governance and interoperability complexities.

1.2 Objectives

This paper aims to demonstrate-

- how Delta Lake ensures ACID compliance and performance across multicloud environments.
- Showcase governance across AWS and Azure via Unity Catalog.
- Propose a framework tailored for regulated sectors.

2. CORE TECHNOLOGIES

2.1 Delta Lake

Delta Lake enhances data lakes with ACID transactions, scalable metadata, and unified batch-stream processing. Its key features are as follows-

- ACID transactions
- Time travel
- Schema enforcement and evolution
- Scalable metadata

Fig 1: Key features of Delta Lake

2.2 Unity Catalog

Unity Catalog provides fine-grained access control, data lineage tracking, and audit capabilities across Databricks environments.Key Features of Unity catalog are-

- Controlled metadata
- RBAC and ABAC support
- Multicloud compatibility
- Identity provider integration (e.g., Azure AD, AWS IAM)

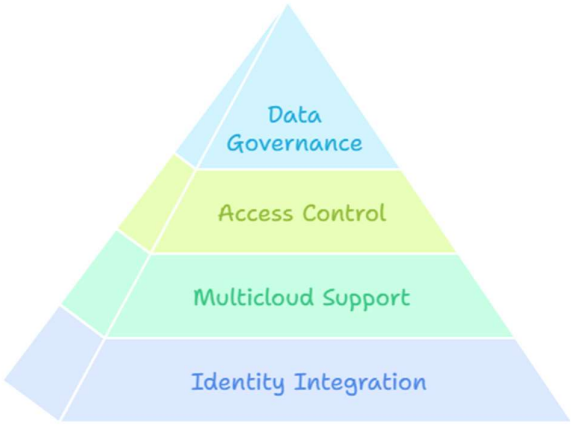

Fig 2: Unity Catalog Hierarchy

3. OVERVIEW OF MULTICLOUD ARCHITECTURE

3.1 Principles of Design

Data Sovereignty: Maintain jurisdictional data boundaries.

Interoperability: Use open formats (Parquet, Delta) and APIs.

Security by Design: Encrypt data in transit and at rest; enforce least privilege access.

Scalability: Use autoscaling clusters and serverless compute.

3.2 Reference Architecture

The reference architecture is designed to support scalable, secure, and compliant data lake operations across AWS and Azure using Databricks as the unified analytics platform. It integrates core components for storage, compute, governance, and data movement.

🔷3.2.1 Storage Layer

Amazon S3 (AWS) and Azure Data Lake Storage Gen2 (ADLS Gen2) serve as the foundational storage systems.

These are optimized for storing large volumes of structured and unstructured data in open formats like Parquet and Delta.

Delta Lake adds transactional consistency and schema enforcement on top of these storage systems.

🔷3.2.2 Compute Layer

Databricks Workspaces on AWS and Azure provide distributed compute environments for data engineering, analytics, and machine learning.

These workspaces are connected to the storage layers and support serverless and autoscaling clusters for efficient resource utilization.

🔷 3.2.3 Governance Layer

Unity Catalog acts as the centralized governance and metadata management system.

It enables fine-grained access control, data lineage tracking, and audit logging across both cloud platforms.

Unity Catalog supports RBAC and ABAC, and integrates with identity providers like Azure AD and AWS IAM.

🔷 3.2.4 Data Movement and Sharing

Delta Sharing facilitates secure, real-time data sharing across cloud boundaries without data replication.

Databricks Connect allows developers to interact with Databricks from local IDEs.

Cloud-native pipelines (e.g., Azure Data Factory, AWS Glue) are used for ingestion, transformation, and orchestration.

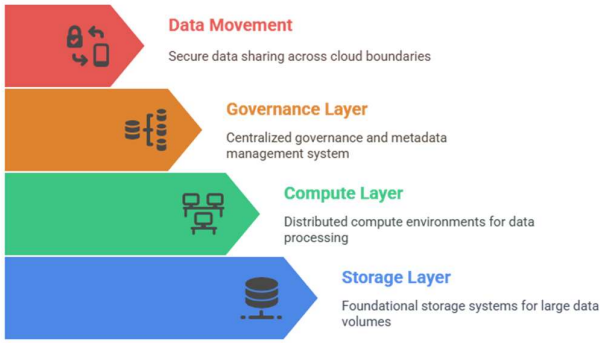

Figure 3: Multicloud Reference Architecture

4. COMPLIANCE AND GOVERNANCE

4.1 Regulatory Requirements

HIPAA: Audit logging, data encryption, access control

GDPR: Data minimization, right to erasure, data lineage

SOX: Change tracking, audit trails

4.2 Unity Catalog for Compliance

This auditability, lineage, and dynamic access—creates a governance trifecta that tightly aligns with regulatory expectations while supporting a scalable and flexible multicloud data lake architecture

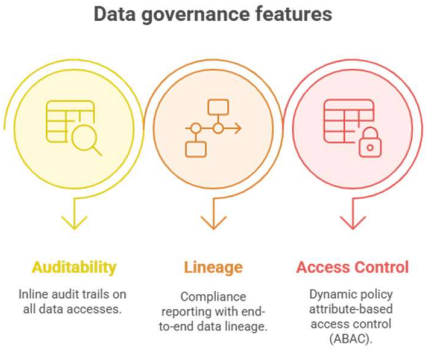

Figure 4: Compliance Mapping with Unity Catalog

Auditability

Unity Catalog features inline audit trails that document all access requests and data interactions as they occur in real time. This indicates that all inquiries, alterations, and data access by a user are documented in real-time. The centralized nature of these logs, being write-only, facilitates the tracing of user activity, the identification of abnormalities, and the demonstration of compliance with regulations like SOX or HIPAA, enabling security teams and auditors to perform these tasks with ease. The visibility helps deter unauthorized individuals and facilitates thorough forensic investigations if necessary.

Lineage

Unity Catalog features built-in end-to-end data lineage that automatically monitors the source of the data, the transformations it undergoes, and its final destination. This metadata mapping encompasses data sources, transformations, and outputs across notebooks, dashboards, and pipelines. In contexts where traceability holds significant value, like finance and healthcare, this form of lineage reporting indicates that, when integrated with the collection and dissemination of personal data, one can demonstrate adherence to data governance regulations such as GDPR.

Access Control (ABAC)

Unity Catalog featuring Dynamic Attribute-Based Access Control (ABAC) allows for the creation of policies that adapt based on user attributes (like department or location), data tags (including sensitive or PII), and contextual conditions (such as time of access, geography, etc.). This approach enhances the conventional role-based access control (RBAC) method by offering organizations detailed, immediate oversight regarding access to view or utilize specific data.

Unity Catalog supports:

Inline audit trails

End-to-end data lineage

Dynamic ABAC policies

5. PERFORMANCE OPTIMIZATION

5.1 Delta Lake Optimizations

Z-Ordering: Improves query performance on frequently filtered columns

Data Skipping: Reduces I/O by ignoring irrelevant files

Optimize and Vacuum: Periodic compaction and cleanup

5.2 Cross-Cloud Performance

Delta Sharing: Enables secure, real-time data sharing across clouds

Caching: Delta Caching and Photon Engine reduce latency

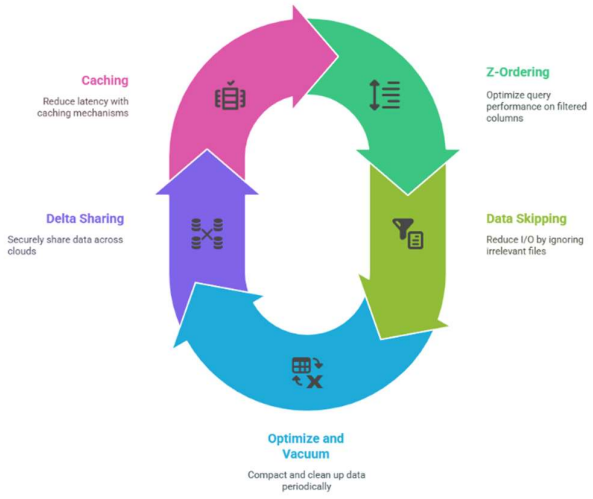

Figure 5: Data Lake Performance Optimization cycle

6. CASE STUDY: FINANCIAL SERVICES

A global bank deployed a multicloud data lake using Databricks on AWS and Azure.

Outcomes:

50× reduction in data access latency

Unified administration across 12 regions via Unity Catalog

Full GDPR and SOX compliance through audit logging and lineage

7. CHALLENGES AND FUTURE DIRECTIONS

7.1 Challenges

Cross-cloud latency and egress charges

- Identity federation complexity
- Evolving regulatory landscape

7.2 Future Directions

Integration of AI/ML governance frameworks Expansion of Unity Catalog to new data modalities Enhanced hybrid cloud support

8. CONCLUSION

Delta Lake and Unity Catalog provide a robust foundation for building scalable, compliant, and performant multicloud data lakes. For regulated industries, this architecture supports innovation while maintaining strict adherence to compliance standards.

REFERENCES

1. Databricks. (2024). Delta Live Tables and Unity Catalog Integration. https://databricks.com
2. Databricks. (2024). Predictive Optimization in Unity Catalog. https://databricks.com
3. Databricks. (2024). Enterprise-Wide Governance with Unity Catalog. https://databricks.com
4. Correction to: An Improved Multiexposure Image Fusion Technique by Nazish, et al. Big Data 2023;11(3):215–224;
5. Bhadresh Shiyal. "Beginning Azure Synapse Analytics", Springer Science and Business Media LLC, 2021.