International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 7, Issue 2, 61-64, 2026
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V7I2P110

Original Article

Modernizing Anti-Money Laundering (AML) Data Pipelines Using Cloud-Native Architectures on Azure and Databricks

Mukesh Kumar Mishra
Individual Contributor.

Received On: 17/02/2026 ***Revised On: 26/03/2026*** ***Accepted On: 04/04/2026*** ***Published On: 11/04/2026***

Abstract: *Financial institutions must analyze large volumes of transactional data while complying with strict regulatory requirements aimed at preventing financial crime. Many existing Anti-Money Laundering (AML) systems rely on legacy infrastructures that use batch-oriented ETL pipelines and static rule-based monitoring, resulting in limited scalability and delayed fraud detection. This paper proposes a cloud-native AML data pipeline architecture implemented using Microsoft Azure and Azure Databricks. The framework integrates real-time data streaming, scalable data lake storage, and machine learning–based analytics to detect suspicious financial behavior. Data governance, lineage tracking, and automated compliance checks are embedded within the pipeline. The proposed approach enhances processing throughput, improves analytical accuracy, and enables financial institutions to respond more effectively to emerging financial crime patterns.*

Keywords: *Anti-Money Laundering (AML), Cloud-Native Architecture, Azure Databricks, Delta Lake, Real-Time Streaming, Financial Compliance, Machine Learning, Data Governance*.

1. Introduction

The rapid digital transformation of the global banking sector has resulted in exponential growth in financial transaction volumes. While digital banking platforms provide convenience and accessibility, they also increase the complexity of monitoring financial activities for potential illicit transactions. Anti-Money Laundering (AML) programs therefore play a critical role in safeguarding the financial ecosystem from misuse by criminal organizations.

Financial institutions are required to maintain monitoring mechanisms that can detect suspicious activities and report them to regulatory authorities. In the United States, regulatory oversight is enforced by multiple agencies including the Financial Crimes Enforcement Network (FinCEN), the Office of Foreign Assets Control (OFAC), the Securities and Exchange Commission (SEC), and the Federal Reserve System. These organizations require financial institutions to maintain transparent audit trails, effective monitoring systems, and robust reporting processes.

Traditional AML infrastructures were primarily designed for batch-based transaction processing environments. As transaction volumes continue to increase and financial crime techniques evolve, these systems face significant operational limitations. Consequently, financial institutions are increasingly exploring cloud-native architectures capable of supporting scalable analytics, real-time data processing, and intelligent risk detection models.

2. Limitations of Traditional AML system

- Dependence on batch-oriented ETL processes executed daily or weekly
- Limited data lineage and traceability capabilities
- Slow alert processing and delayed decision-making workflows
- High occurrence of false-positive alerts generated by static rule engines
- Operational inefficiencies due to manual investigation processes
- Limited scalability of on-premise infrastructure
- Inability to process high-frequency transactional data streams

The proposed architecture addresses these limitations by introducing scalable cloud-native processing, real-time analytics, and machine learning driven detection capabilities. The shift toward cloud-native architectures in anti-money laundering (AML) enables real-time detection, elastic compute scaling, and AI-driven risk scoring. A cloud-based AML framework built on Microsoft Azure and Azure Databricks enables integrated streaming analytics, scalable storage, and advanced machine learning–driven detection capabilities to detect complex financial crimes. The architecture adopts a layered data processing framework commonly referred to as the Medallion architecture, where data is organized into Bronze (raw ingestion), Silver (cleansed datasets), and Gold (analytical datasets) layers to improve data quality, lineage, and governance. The Medallion architecture model widely used in modern data

lake systems organizes datasets into multiple refinement layers.

Key benefits of proposed Architecture

- Enhanced Detection Accuracy
- Real time data streaming and processing
- Cost Efficiency and scalability
- Improved regulatory compliance
- Rapid analysis of large volumes of data

3. Proposed Architecture Overview

As illustrated in Fig. 1, data is collected from multiple heterogeneous sources, including flat files, DB2 databases, and real-time event streams. This data is initially processed by the core banking system to support operational decision-making and transaction processing. Subsequently, the data is ingested into the cloud environment through Azure Event Hub within the Microsoft Azure platform.

In the Azure data platform, the incoming data is first stored in the Bronze layer of Azure Data Lake Storage Gen2 (ADLS Gen2), where it is preserved in its raw format without any transformation. The raw data is then processed and transferred to the Silver layer, where data cleansing, normalization, and structuring are performed to create a standardized and reliable data model. Further transformation, enrichment, and business logic are applied in the Gold layer to generate curated datasets optimized for analytics and downstream applications.

Based on business requirements, the curated data is processed using Azure Databricks for large-scale data transformation and feature engineering. The processed datasets are then forwarded to machine learning–based detection channels, where analytical models evaluate transaction patterns to identify anomalies or potentially fraudulent activities.

A real-time risk scoring engine computes a risk score for each transaction or event. If the calculated risk score exceeds a predefined threshold value, the transaction is flagged as a potential fraud case. Consequently, a case is generated and forwarded to the case management system for further investigation. Reporting and visualization tools such as Power BI are used to generate Suspicious Activity Reports (SARs) and analytical dashboards, which are shared with relevant stakeholders for monitoring, compliance, and further decision-making.

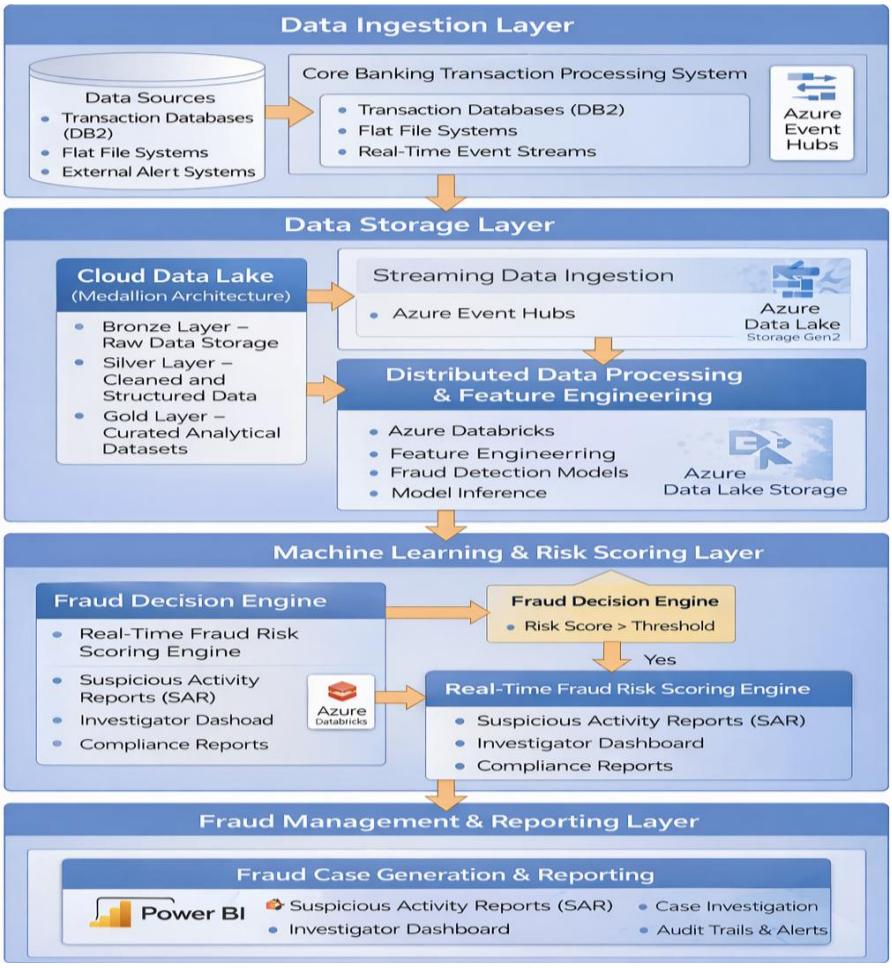

Fig 1: High Level System Architecture – End to End Cloud Native AML Structure

4. Medallion Data Architecture

Table 1: Raw data Structure (Bronze) → Structured data Model (Silver) → Curated data Model (Gold)

Layer	Description	Purpose
Bronze	Raw ingested transactions	Preserves original source data for auditing
Silver	Cleansed & validated	Applies schema validation and data quality rules
Gold	AML analytical datasets	Transformation logic applied - Risk scoring & reporting

5. Machine Learning – Fraud Detection

Conventional AML monitoring platforms rely heavily on predefined rules such as transaction thresholds or geographic risk indicators. Although rule-based detection methods are effective for identifying known fraud scenarios, they often struggle to detect complex laundering patterns that evolve over time.

Machine learning techniques enable the identification of hidden patterns within large financial datasets. Unsupervised anomaly detection models can identify deviations from normal transaction behavior without requiring predefined fraud rules. These approaches allow institutions to detect suspicious activity patterns that may otherwise remain undetected.

Network-based analytical techniques further enhance detection by analyzing relationships among accounts, customers, and counterparties. Graph-based models can reveal hidden transaction networks that may indicate coordinated financial crime operations.

Additionally, Natural Language Processing methods can be applied to analyze unstructured investigation notes or payment descriptions, providing additional contextual information for investigators.

6. Performance - Throughput

Traditional AML platforms built on batch-processing infrastructures often struggle to support high-volume transaction processing. Cloud-native distributed computing platforms significantly improve throughput by enabling parallel processing across scalable computing clusters. Cloud-native data platforms using distributed processing frameworks can achieve throughput exceeding 5,000 transactions per second, representing a tenfold increase in processing capacity.

This scalability enables financial institutions to process large volumes of transactional data in near real time, significantly improving fraud detection responsiveness.

7. Risk Score Calculation – Traditional Vs. Modern

As per the traditional method, when any alert gets generated by any external system, alert is processed by decision channels. Impact of first party, second party is evaluated related to specific alert(s). Then, multi-levels of processing are done and then a Risk score is assigned. If the risk score > threshold limit, it is considered as a fraud and if risk score < threshold limit, it is not a fraud. This rule-based workflow often introduces operational latency due to manual investigation and multi-stage alert evaluation and lot of manual evaluation and calculation. By following the Modern model-based architecture all above issues can be avoided.

8. Security Based Architecture

For login into Azure AD, many security components like Role Based Access Control (RBAC), Multi factor authentication (MFA), Single Sign On (SSO), Key Vault Secret Management, private endpoints are in place. Security and regulatory compliance remain critical requirements for financial systems. The proposed architecture integrates multiple security controls available within the Azure ecosystem. Security layer architecture leverage several Azure security components from user to Azure Data Lake Storage Generation2 (ADLS Gen 2) via Azure Active Directory (Azure AD), credential pass-through or managed identities to access the data without hardcoding secrets in the notebook and Azure Databricks. Azure Data pipelines in Azure Data factory follow the Azure AD security components like RBAC, MFA etc.

9. Implementation Challenges in Financial Institutions

Implementing cloud-native AML architectures requires careful integration with existing banking systems and regulatory reporting frameworks. Many financial institutions operate legacy core banking platforms that generate transactional data in heterogeneous formats. Data standardization, schema enforcement, and real-time validation mechanisms are therefore essential to ensure the reliability of AML monitoring processes. Additionally, institutions must address governance requirements such as auditability, regulatory traceability, and secure data access when deploying cloud-based analytics platforms.

Table 2: General Analysis – Legacy Vs ML Cloud

Dimension	Legacy AML	Modern Cloud AML
Processing Mode	Batch	Real-Time
Model Capability	Rule-Based	ML + Graph
Scalability	Fixed	Elastic
Compliance Readiness	Reactive	Proactive
DevOps	Manual	Continuous Integration /Continuous Deployment (CI/CD) Enabled

10. Conclusion

The modernization of AML data pipelines using cloud-native technologies offers significant advantages in scalability, processing speed, and analytical capability. By integrating real-time data streaming, distributed processing frameworks, and machine learning models, financial

institutions can significantly improve their ability to detect suspicious financial activities. The proposed architecture using Microsoft Azure and Azure Databricks provides a scalable platform capable of handling large volumes of transactional data while maintaining regulatory compliance. As financial crime continues to evolve, cloud-based AML systems offer the flexibility and analytical power necessary to support proactive fraud detection and regulatory reporting.

References

1. Financial Action Task Force, 'International Standards on Combating Money Laundering,' 2023.
2. FinCEN, 'Bank Secrecy Act AML Examination Manual,' 2023.
3. Microsoft Azure Architecture Center, Big Data Analytics Guide, 2024.
4. Databricks, 'Delta Lake Architecture Overview,' 2024.
5. J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2011.
6. I. Goodfellow, Deep Learning, MIT Press, 2016.
7. S. Singh, 'Real-Time AML Analytics Using Big Data,' 2024.
8. T. White, Hadoop: The Definitive Guide, O'Reilly, 2015.
9. A. Gandomi, 'Big Data Analytics Methods,' 2024.
10. OECD Financial Crime Report, 2023.
11. IEEE Big Data Conference Proceedings, AML Analytics Paper, 2024.
12. World Bank Financial Integrity Report, 2023.
13. Gartner Report on Cloud Data Platforms, 2024.
14. McKinsey Global Banking Fraud Study, 2023.
15. NIST AI Risk Management Framework, 2024.
16. M. Weber et al., "Anti-Money Laundering in the Era of Big Data," IEEE Transactions on Big Data, 2022.
17. N. M. Alharbi, "Machine Learning Approaches for Financial Fraud Detection," IEEE Access, 2021.