International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 218-221
https://doi.org/10.70917/ijcisim-2026-4985
Article

Scalable Data Lineage And Auditability Models For AML Compliance In Snowflake Environments

Mukesh Kumar Mishra

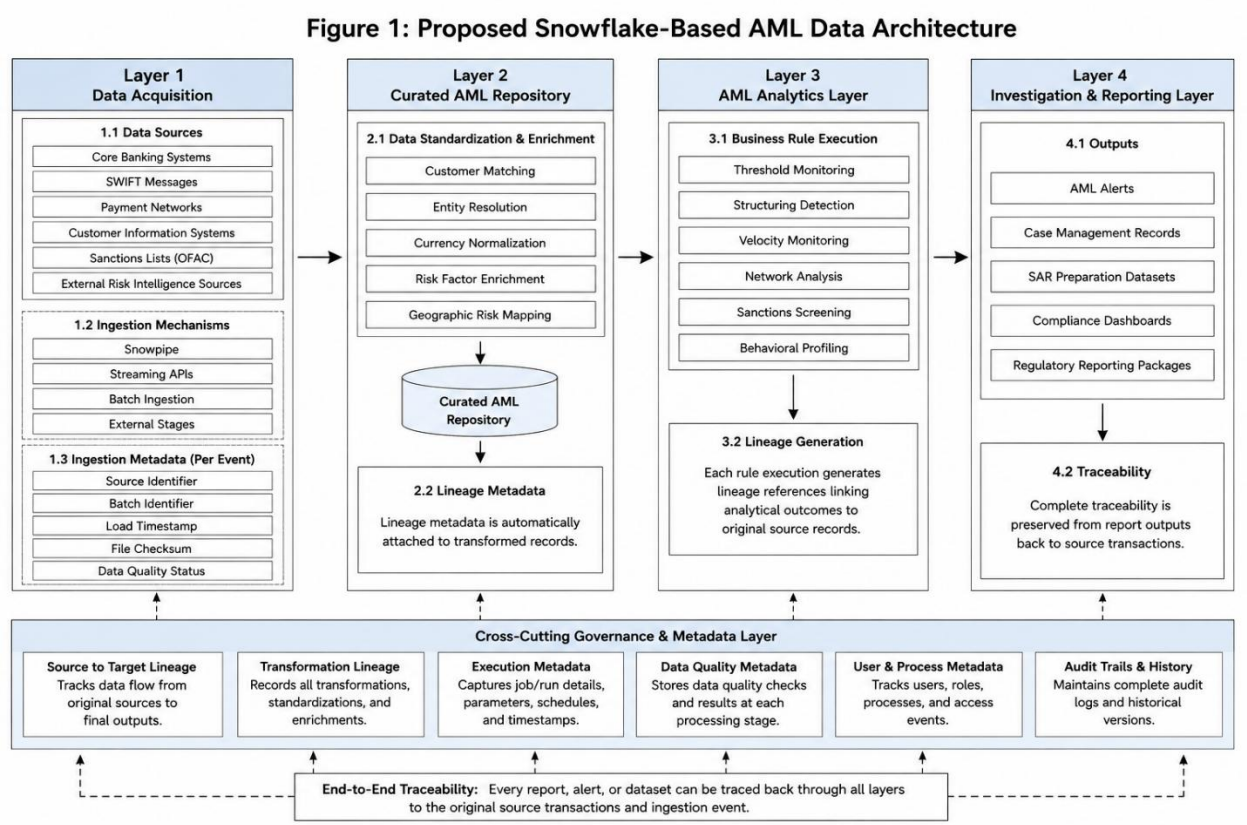
Financial Technology & Data Engineering, India
Email: MISHRAMUKESH2711@GMAIL.COM

Abstract: In today's modern world, Anti-Money Laundering (AML) is very crucial and important. It depends primarily on the integrity, traceability, and explain ability of financial data as the data journey goes through the complex native architectures. Regulators are continuously updating the guidelines of AML and making the process stricter. Regulators mandate financial institutions like bank to demonstrate not only the accurate suspicious activity monitoring systems but also the complete data lifecycle that influences risk score calculations, alerts, sanctions, investigations and regulatory filings. In traditional approach, it is a struggle to find the data chain of custody due to limited visibility of data multiple sources, data paths and inconsistent audit controls across various analytical platforms. This paper presents a scalable data lineage and auditability framework designed for AML compliance on Snowflake based data platforms. The framework introduces multi-layer lineage architecture which gives end to end data visibility across AML pipelines , it includes ingestion traceability, business rule transparency and regulatory evidences. To achieve the framework objective, this architecture leverages snowflake services, object dependencies, access history, task execution and governance records. This paper proposes an architecture to integrate data quality controls, metadata governance, data lifecycle and compliance dashboards which can improve data transparency, reduces fraud and suspicious activities investigation effort, audit reviews and strengthen regulatory readiness and enhances trust in AML monitoring systems.
Keywords: Anti-Money Laundering, Data Lineage, Auditability, Snowflake, Metadata Governance, Financial Compliance, Data Engineering, Regulatory Reporting

1. INTRODUCTION

Banking and other financial institutions process massive volume of transactions across different payment systems, customer onboarding platforms, applications screening sanctions, different banking channels, digital banking services like crypto platforms. As per the regulatory norms, financial institutions are required to maintain transparency to how transactional data is sourced, transformed, analyzed and ultimately utilized in decision making. In traditional methods , data used to be in traditional data warehouses and it was consumed and analyzed as per the user requirement but now after technology advancement data migrated to cloud native environments like Snowflake, AWS S3 (Amazon web Services – Simple Storage Service) which made data lineage more difficult. Figure-1 shows the proposed snowflake architecture

Figure-1- Proposed approach



Modern AML ecosystems have data feeds from multiple sources, multiple data ingestion mechanisms, data streaming, artificial Intelligence, machine learning models and reporting systems which introduces additional complexity and can complicate the data lifecycle from origin to target until compliances. This complexity and data tracking difficulties can result in increased regulatory scrutiny, delayed investigations and reduced confidence in suspicious activity detection processes. To solve the above problems, this paper proposes a practical lineage architecture that aligns Snowflake-native capabilities with AML compliance objectives

2. AML Challenges

As per the regulatory compliance requirements, AML compliance functions must demonstrate data traceability from source->target->report, data transformation transparency, alert generation, fraud activities details, historical data, access accountability and audit documents. In simple terms, regulators provide mandate to firms to explain how a specific transaction contributed to an alert, how the alert evolved through investigations, and which transformations influenced the final compliance outcome. Financial institutions commonly experience common data traceability

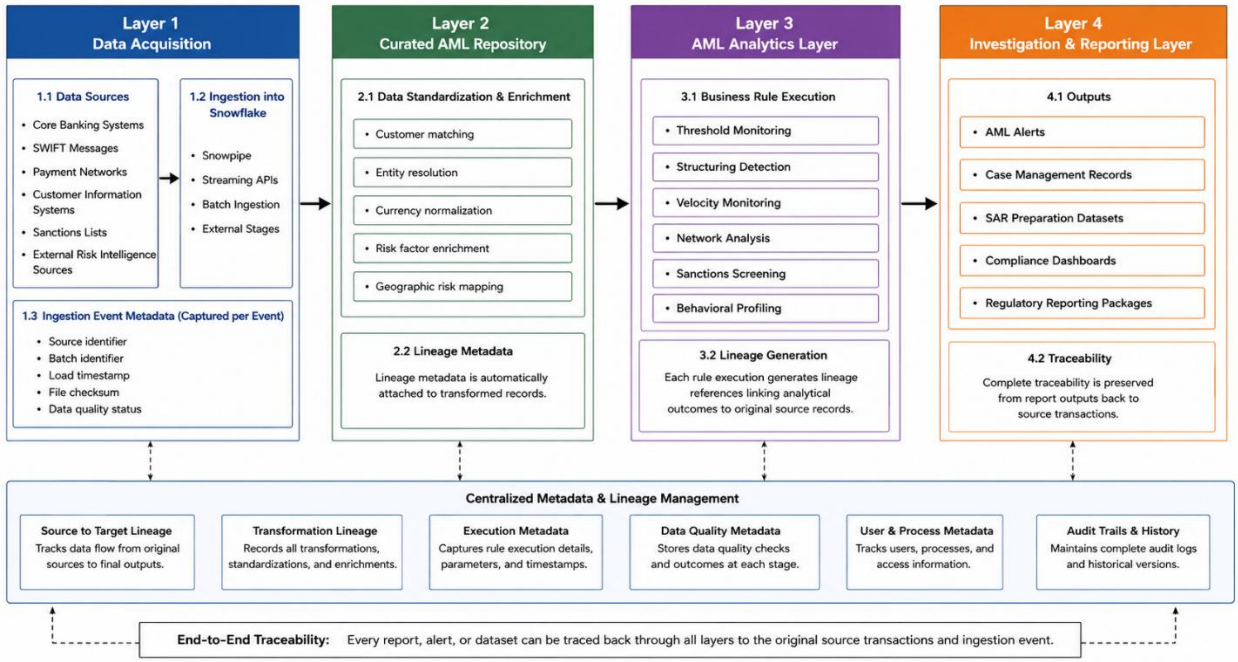
gaps like incomplete metadata capture, undocumented transformation logic, fragmented audit trails, inconsistent lineage across environments and limited visibility into workflows which gets amplified more as the volume of transaction data increases.

3. AML Data Architecture - Snowflake

The proposed architecture consists of four logical layers – Data Acquisition, curated AML repository, data Analytics layer, Reporting layer. Data acquisition includes the data sources or input feeds to AML like core banking systems, swift messages, payment networks, customer Information systems, sanctions list from OFAC (Office of Foreign Assets Control) and other external sources. Data is then loaded into Snowflake with either configured data pipelines / Streaming APIs /Batch Ingestion and each ingestion event is identified with source identifier, batch identifier, load timestamp, file checksum and status. Second layer is AML Curated repository. In this layer, data is

enriched and standardized through customer matching, entity resolution, currency normalization, risk factor enrichment and geographic risk mapping especially sanctioned countries. Third layer is Analytics layer. It consists of business rule executions like threshold monitoring, structuring detection, velocity monitoring ,sanctions screening and behavioral profiling which help to generate data tracing from source to analytics layer. Fourth layer is the reporting layer whose output includes AML alerts, case management records, SAR (Suspicious activity report), compliance dashboards, regulatory reporting packages which can be audited by regulators. Figure-2 shows the same data flow. Complete traceability is preserved from report outputs back to source transactions.

Figure-2



4. Data Framework

The proposed framework consists of four lineage dimensions – Technical , Business, Operational and Regulatory. Technical lineage captures information like schema details, task execution history and data journey paths which provides traceability paths. Business lineage captures AML rules, Risk score calculations, Investigation and compliance decisions which explains any data changes. Operational lineage captures Pipeline execution history, Job failures (if any),processing durations and data quality outcomes which supports data governance. Regulatory lineage captures alert generation evidence, investigation decisions, regulatory reporting lineage and audit documentations.

5. AML Risk Management Model

The proposed auditability model contains five control domains – source of funds tracking, data transformation/curation transparency, access accountability, financial history reconstruction and regulatory evidences. Source of funds tracking links to source system, data acquisition event, data processing pipeline and the business owner. Data transformation/curation transparency maintains all historical version changes and links to deployment history, approval records and change requests. All historical data and version are maintained through CDC (Change data capture) logic. Access accountability enables monitoring of user activities, sensitive data access and privilege access. For financial history reconstruction investigators can recreate historical alerts , risk score calculations and data snapshots to capture the data changes. Regulatory evidences package includes source records, data lineage flow chart or diagrams, validation results and investigation documents (vary on case-to-case basis).

6. Benefits for AML Compliance Programs

Using Snowflake in proposed architecture helps in implementation of scalable data flow by supporting access history, object dependencies, data streams, tasks, tags/classification, secured schemas, data sharing info. Proposed framework can help organizations to achieve faster regulatory examinations, enhanced data trustworthiness, reduced audit preparation effort, improved investigation efficiency, increased transparency of AML decisions and enhanced governance controls

7. CONCLUSION

The scalable framework presented in this paper combines technical, business, operational, and regulatory lineage dimensions within a Snowflake-based architecture. By integrating metadata governance, audit controls, and lineage automation, organizations can create a sustainable compliance foundation capable of supporting growing transaction volumes and increasingly complex regulatory expectations. Data dependency tracking and auditability have evolved from technical governance concerns into critical compliance capabilities within modern AML programs. As financial institutions continue migrating to various cloud server or serverless architectures which are specifically built, deployed, and managed to take full advantage of cloud computing models, maintaining end-to-end traceability becomes essential for regulatory confidence, operational effectiveness, and investigative transparency. Future research may explore integration with AI-driven AML monitoring systems, automated lineage validation mechanisms, and real-time compliance observability platforms. Furthermore, with the advancements of Quantum computing, it will further enhance data lineage traceability and fraud investigations.

REFERENCES

1. FATF, “International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation,” Financial Action Task Force, Paris, France, 2024.
2. Basel Committee on Banking Supervision, “Sound Management of Risks Related to Money Laundering and Financing of Terrorism,” Bank for International Settlements, Switzerland, 2020.
3. U.S. Department of the Treasury, “Bank Secrecy Act Anti-Money Laundering Examination Manual,” FFIEC, Washington, DC, USA, 2023.
4. M. Jans, M. Alles, and M. Vasarhelyi, “A Field Study on the Use of Process Mining of Event Logs as an Analytical Procedure in Auditing,” The Accounting Review, vol. 89, no. 5, pp. 1751–1773, 2014.
5. P. K. Chan and S. J. Stolfo, “Toward Scalable Learning with Non-Uniform Class and Cost Distributions,” Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 164–168, 1998.
6. Microsoft, “Azure Databricks Documentation,” Microsoft Learn, 2025.
7. Databricks, “Medallion Architecture Overview,” Databricks Documentation, 2025.
8. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.
9. FFIEC. Bank Secrecy Act / Anti Money Laundering Examination Manual. U.S. Treasury, 2023.
10. Fiore, U., et al. “Using GANs for Improving AML Detection.” Engineering Applications of AI, 2019.
11. Giebler, C., et al. “Big Data Quality: A Survey.” Journal of Big Data, 2020.
12. Glavic, B. “Data Provenance: A Categorization of Existing Approaches.” IEEE Data Engineering Bulletin, 2014.
13. Simmhan, Y. L., Plale, B., & Gannon, D. “A Survey of Data Provenance in e Science.” ACM SIGMOD Record, 2005.