Project: Daily Batch Extraction Application for On-demand Queries With AWS

- Dataset Info: Money laundering remains a significant global issue, driving the need for improved transaction monitoring methods. Current anti-money laundering (AML) procedures are inefficient, and access to data is difficult/restricted by legal and privacy issues. Moreover, existing data often lacks diversity and true labels. This study introduces a novel AML transaction generator, creating the SAML-D dataset with enhanced features and typologies, aiming to aid researchers in evaluating their models and developing more advanced monitoring methods.

The dataset incorporates 12 features and 28 typologies (split between 11 normal and 17 suspicious). These were selected based on existing datasets, the academic literature, and interviews with AML specialists. The dataset comprises 9,504,852 transactions, of which 0.1039% are suspicious. It also includes 15 graphical network structures to represent the transaction flow within these typologies. The structures, while sometimes shared among typologies, vary significantly in parameters to increase complexities and challenge detection efforts. More details about these typologies are available in this paper (B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan and G. Aksu, "Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset," 2023 IEEE International Conference on e-Business Engineering (ICEBE), Sydney, Australia, 2023, pp. 47-54, doi: 10.1109/ICEBE59045.2023.00028. https://ieeexplore.ieee.org/document/10356193). The dataset used in this project is a reduces version (8% randomly sampled) of the source dataset containing 76039 transactions.

Features of the SAML-D dataset:

• Time and Date: Essential for tracking transaction chronology.

• Sender and Receiver Account Details: Helps uncover behavioural patterns and complex banking connections.

• Amount: Indicates transaction values to identify suspicious activities.

• Payment Type: Includes various methods like credit card, debit card, cash, ACH transfers, cross-border, and cheque.

• Sender and Receiver Bank Location: Pinpoints high-risk regions including Mexico, Turkey, Morocco, and the UAE.

• Payment and Receiver Currency: Align with location features, adding complexity when mismatched.

• 'Is Suspicious' Feature: Binary indicator differentiating normal from suspicious transactions.

• Type: Classifies typologies, offering deeper insights.

- Context: At the start of a customer relationship, institutions gather personal and financial data to verify the customer’s identity through Know Your Customer (KYC) procedures. These records form the basis for monitoring future transactions.

Financial institutions use automated software to monitor transactions in real-time, flagging those that exceed predefined thresholds or show unusual patterns (e.g., large cash deposits, rapid transfers between accounts). Suspicious Activity Reports (SARs) are generated if a transaction is deemed questionable.

AML regulations require institutions to keep records of transactions for a specified period (typically five years or more). This includes details of transactions (dates, amounts, recipients), customer profiles, and the SARs filed.

When a transaction or series of transactions trigger suspicion (such as structuring, large cash deposits, or transfers to high-risk countries), institutions are required to file SARs with regulators. These reports often include transaction details, customer information, and reasons for suspicion.

- Project scenario: A Financial institution generates thousands of international transaction entries daily. To ensure regulatory compliance, Financial institutions are required to periodically review their AML programs and transaction records. Regulators may also conduct audits to ensure that institutions are properly tracking and reporting suspicious transactions. To reduce the risk of non-compliance, the organization reinforces its governance frameworks by occasionally querying its AML transaction generator output. Therefore, to enable AML analysts to run ad-hoc queries without overloading their OLTP database, we will implement a data pipeline using AWS Glue to extract, transform, and load data from an Amazon RDS for Aurora MySQL cluster to an Amazon S3 bucket. Then utilize Amazon Athena for efficient on-demand query execution, ensuring seamless access to customer transaction data for an enhanced AML tracking program.

- Steps: This solution uses glue to extract transform and load data from RDS to S3 bucket. It then uses Athena to run on demand queries.

1. Create a RDS for Aurora MySQL cluster

2. Create a S3 bucket

3. Create an AWS Glue crawler to create databases and tables from the Aurora MySQL cluster using a JDBC connection (An AWS glue crawler is used to extract raw data from RDS. The data schema is

inferred and imported into the Glue Data Catalog).

4. Use AWS Glue Studio to visually create jobs to ingest customer table data from the AWS Glue Data Catalog (The data is transformed by an AWS glue job, removing personally identifiable information (PII) in this case, the date of birth, before archival storage in an S3 bucket.).

5. Users run queries, as needed, on the data by using the Amazon Athena query editor, which retrieves schema info from the Data Catalog.

6. Processed query data is saved to a different S3 bucket for later on-demand reporting use.