Building a Financial Transaction ETL Pipeline

\*1. Introduction\*

\*Objective:\*

To design and implement an ETL (Extract, Transform, Load) pipeline that processes raw transaction data from the Synthetic Financial Dataset.

\*Goals:\*

- Cleanse and validate the dataset.

- Perform transformations such as currency conversion and timestamp standardization.

- Aggregate data at different intervals (daily, weekly, monthly).

- Implement data quality checks.

- Generate summary statistics and profiling reports.

\*2. Dataset Overview\*

The \*Synthetic Financial Dataset\* contains transactional data with fields such as:

- step: Simulation time step (hourly).

- type: Transaction type (e.g., CASH-IN, PAYMENT, etc.).

- amount: Transaction amount.

- nameOrig, nameDest: Account IDs for origin and destination.

- isFraud, isFlaggedFraud: Fraud detection indicators.

\*3. Technical Approach\*

The ETL pipeline is designed in the following stages:

\*3.1 Extract\*

- Load the dataset using pandas.

\*3.2 Cleanse and Validate\*

- \*Remove duplicates\* to eliminate redundancy.

- \*Drop missing values\* to ensure data integrity.

- Validate transaction types against predefined valid types.

- Ensure transaction amounts are positive.

\*3.3 Transform\*

- Convert transaction amounts using a currency conversion rate (e.g., USD to EUR).

- Standardize the step field into human-readable timestamps.

\*3.4 Aggregate\*

- Summarize the transaction data at \*daily, \*\*weekly, and \*\*monthly\* intervals.

\*3.5 Data Quality Checks\*

- Identify issues such as negative amounts, invalid transaction types, and missing values.

\*3.6 Reporting and Profiling\*

- Generate summary statistics for data distribution.

- Visualize transaction amount distributions using histograms.

\*4. Implementation\*

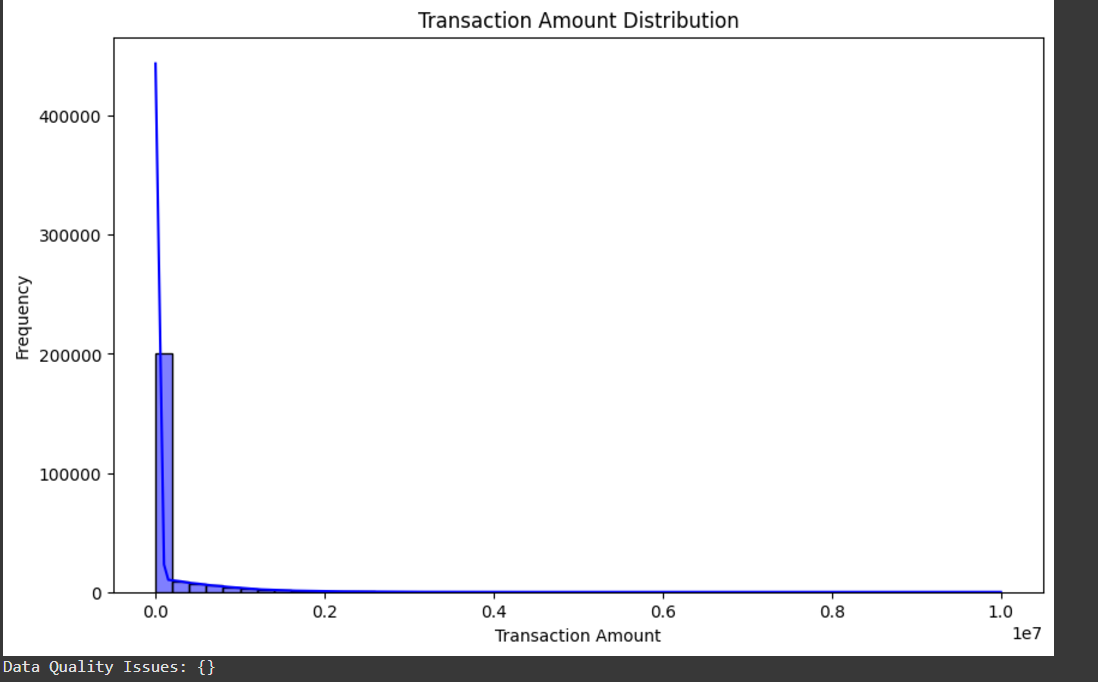
The pipeline is implemented using \*Python\*. Below is a detailed breakdown of the code components:

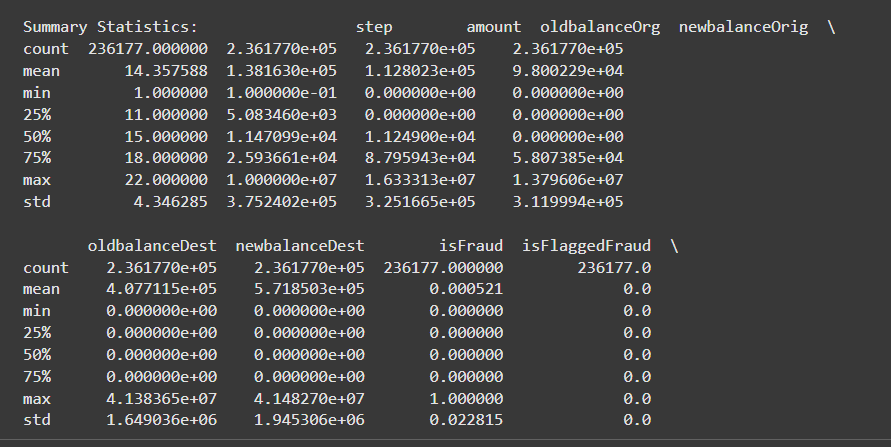
\*4.1 Code Implementation\*

Python

TD1.pynb

\*5. Output and Results\*





\*Key Outputs:\*

1. \*Data Quality Issues Report:\*

- Identifies anomalies such as negative amounts and invalid transaction types.

2. \*Summary Statistics:\*

- Provides mean, median, min, max, and other statistics for transaction data.

3. \*Aggregated Data:\*

- Summarized metrics at daily, weekly, and monthly intervals.

4. \*Visualizations:\*

- Histogram depicting the distribution of transaction amounts.

\*5. Conclusion\*

The ETL pipeline successfully processes transaction data, ensuring data quality and generating actionable insights. It provides a scalable and reusable framework for similar financial datasets.