**Graduation Project Report**

Anti-Money Laundering (AML) Detection System

**Prepared By**

Ahmed Osama

Nahla Mohamed

Mona Osama

Rowan Amr

Adham Medhat

Mahmoud Magdy

**Supervised By**

Eng. Ahmed Essam Azab

**Overview**

Our project focuses on developing a predictive analytics system to identify potential money laundering activities and suspects in global financial transactions. By analyzing transaction patterns and behavioral data using some tools such as SQL server for Data warehousing and Power BI for visualization.

**ETL Process for AML Data Warehouse**

**Purpose of Data Warehousing for the AML Project**

The purpose of data warehousing in this anti-money laundering project is to create a centralized repository for storing, managing, and analyzing large volumes of transaction data from various sources. This enables:

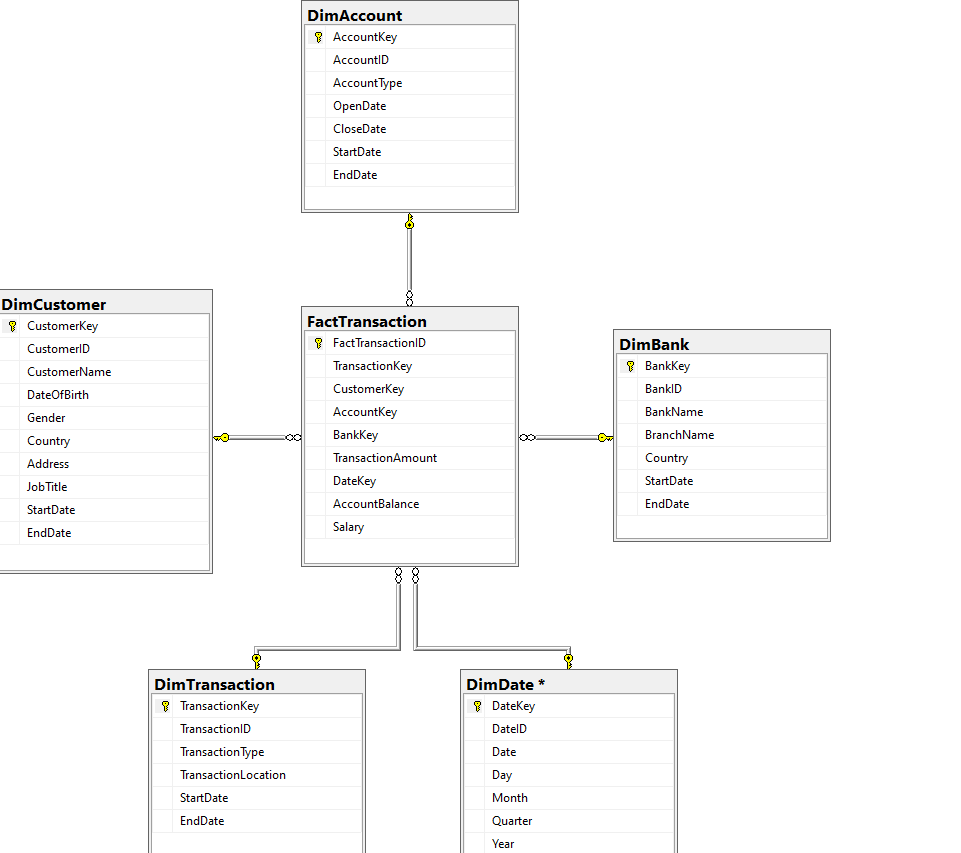
1. **Enhanced Data Integration:** Consolidating data from multiple banks and financial institutions allows for a comprehensive view of client transactions.
2. **Improved Query Performance:** The structured star schema design facilitates faster and more efficient querying, enabling analysts to quickly identify suspicious patterns and trends.
3. **Historical Analysis:** Storing historical transaction data supports longitudinal studies and trend analysis, essential for detecting evolving money laundering tactics.
4. **Data Quality and Consistency:** A data warehouse ensures high-quality, consistent data through rigorous preprocessing, making it reliable for compliance and reporting.
5. **Support for Advanced Analytics:** The warehouse serves as a foundation for implementing machine learning models and advanced analytics, helping to predict and identify potential money laundering activities effectively.

Overall, the data warehouse enhances the project's ability to detect and mitigate financial crime risks while ensuring regulatory compliance.

**Star Schema**

In a star schema designed for an Anti-Money Laundering (AML) project, the central fact table is connected to several dimension tables, facilitating efficient data analysis. Here’s a breakdown of the potential structure:

* Fact Transactions (FactTransactionsID , CustomerID, TransactionAmount, ..)
* Dimensions\_Tables (DimCustomer, DimTransaction, DimBank, ..) as shown



This star schema structure allows for efficient querying and reporting on transaction data, providing insights into potential money laundering activities. Analysts can easily join fact and dimension tables to explore relationships, identify suspicious patterns, and perform comprehensive analyses across different dimensions, such as time, client demographics, and transaction types.

**Create AML Warehouse with Fact and Dimensions tables:**

* **Customer Dimension**

**A screenshot of a computer code

Description automatically generated**

* **Transaction Dimension**

**A screen shot of a computer code

Description automatically generated**

* **Account Dimension**

**A screen shot of a computer code

Description automatically generated**

* **Bank Dimension**

**A screenshot of a computer code

Description automatically generated**

* **Date Dimension**

**A white screen with black text

Description automatically generated**

* **Fact Table**

A computer screen shot of a computer code

Description automatically generated

### **ETL Process Overview for the AML Project:**

The ETL (Extract, Transform, Load) process is crucial for preparing data for analysis in the Anti-Money Laundering (AML) project. Here's an overview of each step:

1. **Extract**

* **Data Generated:**
  + **Bank Transactions:** Comprehensive records of all transactions.
  + **Transaction Amount:** The value associated with each transaction.
  + **Account ID:** Unique identifier for each bank account involved in the transactions.
  + **Country of Transaction:** Geographic location where the transaction occurred.
  + **Address:** Specific address associated with the transaction (if available).
  + **Date of Transaction:** Timestamp indicating when each transaction took place.
  + **Transaction Type:** Classification of the transaction (e.g., Deposit, Withdrawal, Transfer).
* **Data Generation Process:**
* We used Python to write scripts that generated synthetic data, which mimicked real-world transaction patterns.
* Libraries we used: Pandas for creating and manipulating structured datasets efficiently, Faker for generating realistic data such as (names, addresses, … etc), allowing us to simulate customer profiles and their activities.
* **Sample of Code:**

A computer screen shot of text

Description automatically generated

**Sample of Data:**



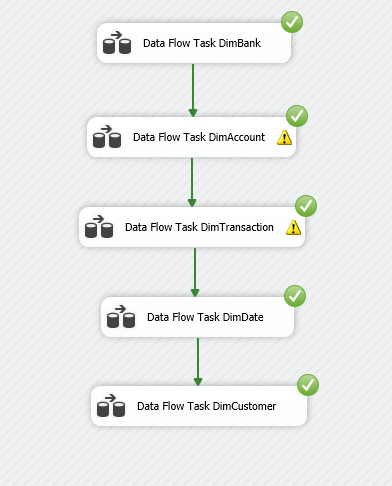
**2. Transform**

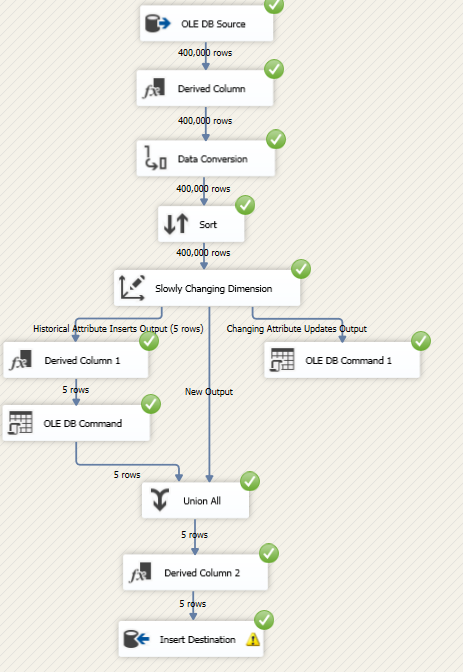
**Objective:** Cleanse and prepare the extracted data for integration into the data warehouse.

* **Data Preparation Using SSIS:**
  + **Derived Columns:**
    - **ID Transformation:** Modify account IDs by removing prefixes (e.g., change "ACC1102" to "1102") for standardization.
    - **Null Value Removal:** Eliminate rows with null values in critical fields to ensure data integrity.
    - **Date Formatting:** Convert date fields to a standardized format that includes day, month, and year, …etc.
  + **Data Conversion:**
    - Ensure all data types are consistent and appropriate for analysis
  + **Slowly Changing Dimensions (SCD):**
    - Implement SCD to track changes in dimension data over time, ensuring historical accuracy.
  + **Merge:**
    - Combine data from various sources into a unified dataset, linking transaction records to client and account details.
  + **Sort:**
    - Organize the data by key attributes (e.g., transaction date or account ID) to optimize loading and querying in the data warehouse.
* **Utilizing SSIS:**
  + Leverage SQL Server Integration Services (SSIS) to automate and streamline these transformation processes, ensuring efficiency and accuracy.
* **Slowly changing Dimension**

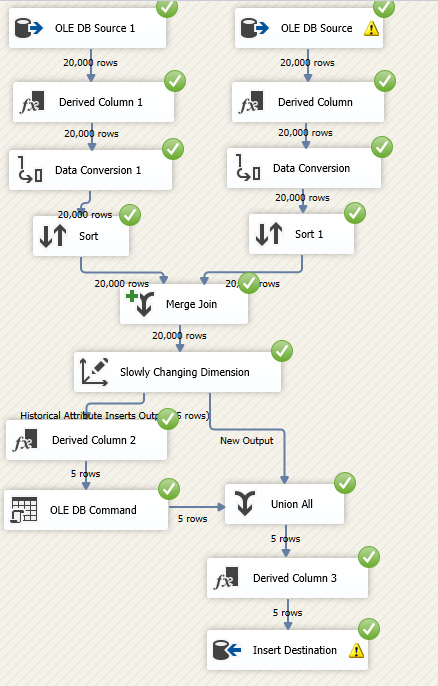
A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse. It is considered and implemented as one of the most critical ETL tasks in tracking the history of dimension records.

* Create Data Flow Task for Each Dimension:

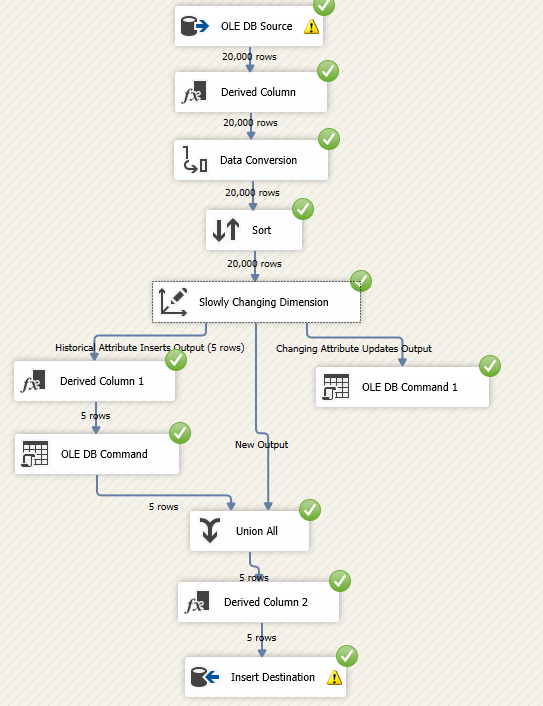


A. Create Data Flow Task for Transaction Dimension

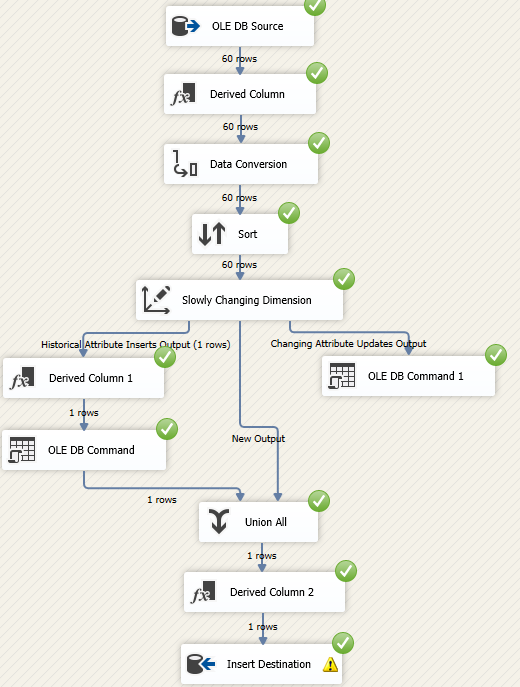
1. B. Create Data Flow Task for Account Dimension



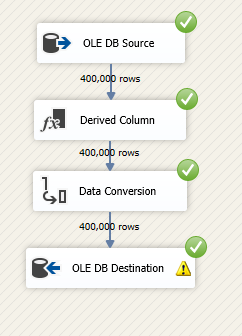
C. Create Data Flow Task for Customer Dimension

1. 

D. Create Data Flow Task for Bank Dimension

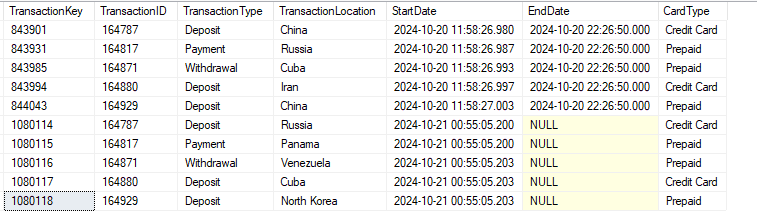


E. Create Data Flow Task for Date Dimension



**Results:**

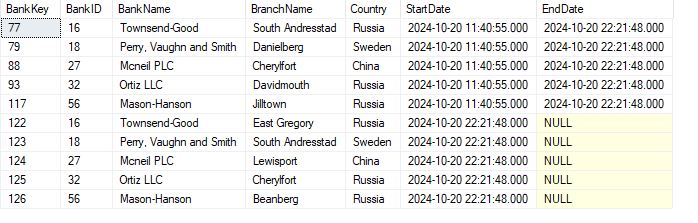
* For Transaction Dimension:



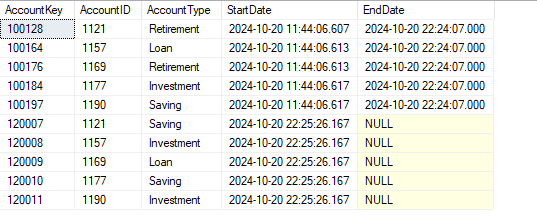
* For Customer Dimension:



* For Bank Dimension:



* For Account Dimension:



* Create Data Flow Task for Fact Transaction:

A screenshot of a computer

Description automatically generated

**Result:**

**A table of numbers and letters

Description automatically generated**

This transformation phase ensures the data is clean, standardized, and ready for effective analysis in the data warehouse.

**3. Load**

**Objective:** Store the transformed data in the data warehouse.

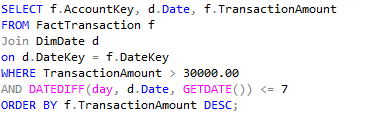
* Load data into the structured star schema designed for the AML project, populating fact and dimension tables.

**Summary:**

The ETL process ensures that the data used in the AML project is accurate, consistent, and ready for analysis. By systematically extracting, transforming, and loading data, the project can leverage advanced analytics to identify and mitigate money laundering risks effectively.

**SQL Scripts and Queries:**

This query retrieves account transactions that have an amount greater than 30,000, and have occurred within the 7 days:

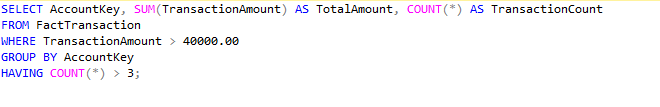


**Result of Query:**

**A screenshot of a data

Description automatically generated**

This query retrieves the total sum of transaction amounts greater than 40,000, and the count of those transactions. It groups the results by **AccountKey** and only includes accounts that have more than 3 such transactions.

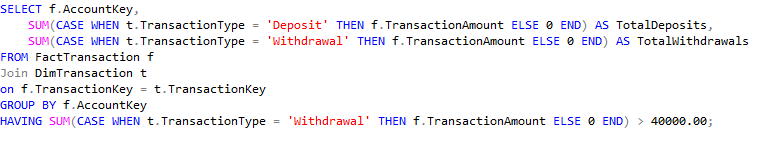


**Result of Query:**

**A screenshot of a computer

Description automatically generated**

**This query retrieves the AccountKey, the total amount of deposits, and the total amount of withdrawals for each account by summing the TransactionAmount based on the transaction type (either "Deposit" or "Withdrawal"). It filters to include only accounts where the total amount of withdrawals exceeds 40,000.**



**Result of Query:**

**A screenshot of a computer

Description automatically generated**

**This SQL query identifies suspicious transactions by grouping transactions based on AccountKey, summing the TransactionAmount, and counting the number of transactions for each account. It filters for transactions above 40,000 and shows only accounts where the total sum of transactions exceeds 500,000.**

A white background with black text

Description automatically generated

**Result of Query:**

**A screenshot of a computer

Description automatically generated**

This query retrieves transaction details. It selects TransactionAmount for transactions that occurred between 2 days to track the accounts that applied suspicious transactions.

A screen shot of a number

Description automatically generated

**Result of Query:**

A screenshot of a data

Description automatically generated

This query identifies repeated transactions for each account within a 24-hour period.

A screenshot of a computer

Description automatically generated

**Result of Query:**

**A screenshot of a phone number

Description automatically generated**

This query retrieves customer transactions from specific high-risk locations ('Iran', 'Syria', 'Russia', 'North Korea', 'China'). It counts the number of transactions and sums the total transaction amount for each customer in those locations, only including customers who have more than one transaction in these locations.

A screenshot of a computer code

Description automatically generated

**Result of Query:**

**A screenshot of a computer

Description automatically generated**

**Visualization**

**A close-up of a graph

Description automatically generated**

The demonstrated dashboard presents key insights from our case study on Anti-Money Laundering (AML) system. The data selected through our queries highlights the most suspicious transactions:

-The total value of transactions above 40,000 amounts to 71.92 billion.

-Syria has the highest value and number of transactions, exceeding 8 billion.

-North Korea, while one of the most suspicious countries, ranks lowest with over 7 billion in transactions.

-In terms of transaction durations, July has the longest average time between transactions, while February has the shortest.

-When comparing transaction types by number, withdrawals have a higher total value than deposits.

**Conclusion**

This project successfully implemented a framework for detecting potential money laundering activities using synthetic transaction data. By simulating various transactions, including those linked to sanctioned countries and unusual amounts, we analyzed patterns that indicate suspicious activities. The structured data and SQL queries provided valuable insights for identifying potential risks. This approach offers a foundation for improving AML detection and can be expanded with machine learning and real-time monitoring for enhanced accuracy in real-world applications.