Project Documentation: FinTech Data Migration Pipeline (Synapse)

---

1. Project Overview

The **FinTech Data Migration Pipeline** is designed to process and transform data across three stages: **Bronze**, **Silver**, and **Gold**. This pipeline leverages **Azure Synapse Analytics** to orchestrate data migration, transformation, and loading, using data stored in **Azure Data Lake Storage (ADLS)** and **Azure SQL Database**. The pipeline includes dynamic processing and uses **PySpark notebooks** for transformations between layers.

The pipeline processes five key tables in the **Fintech schema**:

- Account
- Customer
- Loan
- Payment
- Transaction

---

2. Pipeline Design Overview

2.1 FinTech Container Structure

- **ADLS Structure**:
    - The **fintech container** in **ADLS** contains three main folders: **Bronze**, **Silver**, and **Gold**.
    - Data moves from the **Bronze layer** (raw, untransformed data) to the **Silver layer** (cleansed and enriched data) and finally to the **Gold layer** (analytical data).

2.2 Azure SQL Database:

- Stores metadata about the tables in the fintech schema.
- The pipeline dynamically queries the database to identify the tables that need to be processed.

---

3. Synapse Pipeline Components

3.1 Linked Services

- **ADLS Linked Service**:

    - Connects to the **fintech container** in **ADLS** containing the **Bronze**, **Silver**, and **Gold** folders.
- **SQL Database Linked Service**:
    - Connects to the **Azure SQL Database** containing metadata on the tables that need processing.

3.2 Integrated Datasets

- **BronzeLayer Dataset**: Points to the **fintech container** in **ADLS** and provides access to the data in the **Bronze layer**.
- **SQLDatabase Dataset**: Provides access to the **Azure SQL Database**, enabling dynamic querying for the required tables.

3.3 Pipeline: fintech_data_migration

The **fintech_data_migration pipeline** orchestrates the flow of data from the **Bronze** layer, through the **Silver** layer, and into the **Gold** layer. The pipeline includes the following activities:

1. **Lookup Activity (SQL Query)**: Queries the **SQL Database** to fetch a list of tables to process.
2. **ForEach Activity (Loop)**: Iterates over the list of tables and processes each dynamically.
3. **Copy Data Activity (Within ForEach Loop)**:
    - Source: Dynamically queries the tables (Account, Customer, Loan, Payment, and Transaction) from the **SQL Database**.
    - Sink: Writes the data to the **BronzeLayer** in **ADLS**, stored in a dynamic path.
4. **Notebook Activity (Bronze to Silver Transformation)**: Uses **PySpark notebook** (**Bronzetosilverprocess**) to clean and enrich the data from the **Bronze** layer to the **Silver** layer.
5. **Notebook Activity (Silver to Gold Transformation)**: Uses another **PySpark notebook** (**Silvertogoldprocess**) to apply business rules and aggregations, transforming data into the **Gold** layer.
6. **Web Activity (Success Notification)**: Sends a success email notification via a **Logic App** once the **Silvertogoldprocess** notebook completes successfully.
7. **Web Activity (Failure Notification)**: Sends a failure email notification via a **Logic App** if the **Silvertogoldprocess** notebook fails.

---

4. PySpark Notebooks for Transformation

4.1 Bronze to Silver Transformation (Bronzetosilverprocess Notebook)

The **Bronzetosilverprocess** notebook processes raw, unstructured data from the **Bronze** layer, performing data cleansing and enrichment. **Key Operations**:

- **Data Cleansing**: Handles incomplete, invalid, or missing data.
- **Enrichment**: Adds calculated fields or merges data with external sources.
- **Format Transformation**: Converts raw data into structured formats.
- **Deduplication**: Removes duplicate records for quality assurance.

Example Transformations:

- Filtering incomplete records.
- Changing data types for compatibility.
- Joining data (e.g., merging customer information with account details).

After these transformations, data is saved in the **Silver** layer, ready for deeper analysis.

4.2 Silver to Gold Transformation (Silvertogoldprocess Notebook)

The **Silvertogoldprocess** notebook takes the cleansed and enriched data from the **Silver** layer and applies business logic and aggregations to prepare the data for final use. **Key Operations**:

- **Business Logic Application**: Applies complex rules to prepare data for decision-making.
- **Data Aggregation**: Summarizes data to create useful metrics.
- **Final Data Formatting**: Transforms data into the required format for reporting.

Example Transformations:

- Aggregating transaction data for total spending per customer.
- Summing loan payments and balances.

The output data is stored in the **Gold** layer, optimized for reporting and analytics.

---

5. Global Parameters

- **To**: Recipient email address (e.g., ozairshaikh164@gmail.com)

- **SuccessSubject**: Subject for success email ("Pipeline Executed successfully").
- **SuccessContent**: Content for success email ("Hey ozair, Your Pipeline Executed successfully").
- **FailedSubject**: Subject for failure email ("Pipeline has not Executed successfully").
- **FailedContent**: Content for failure email ("Hey ozair, Your Pipeline has not Executed successfully").

These parameters are used to customize the email notifications for the pipeline's success or failure.

---

6. Conclusion

The **FinTech Data Migration Pipeline** processes data in three stages: **Bronze**, **Silver**, and **Gold**. It performs data migration, cleansing, enrichment, and aggregation using **PySpark notebooks**. The **Bronzetosilverprocess** notebook handles basic transformations and data enrichment, while the **Silvertogoldprocess** notebook applies advanced business rules and aggregates data for analytics.

The pipeline ensures real-time communication of its status via **email notifications** for both successful and failed executions. With this architecture, stakeholders can rely on accurate, timely, and actionable insights from the pipeline's output. The **FinTech Data Migration Pipeline** represents a flexible and automated solution for data transformation in financial services.