Bootcamp Project 2 - Transactions and Loan Data for a Customer

**Objective:**

The project aims to design and implement a robust data pipeline for processing customer account data. This includes copying data from ADLS GEN2 (Bronze layer) and transforming the data in the Silver layer using Data bricks Notebooks and GOLD Storage into the SCDType 1 Delta Table in ADLS GEN2. The pipeline aims to ensure efficient, accurate, and scalable data processing to support downstream analytics and reporting needs.

# Project Steps:

**Step 1: Data Ingestion (Raw(Bronze) Container to ADLS GEN2 (Silver)**

1. Source: Specify the backend team's storage account details and the location of the data to be copied.
   1. accounts.csv
   2. customers.csv
   3. loan\_payments.csv
   4. loans.csv
   5. transactions.csv
2. Sink: Define your data lake storage account and the raw(bronze) container where the copied data will be placed.

Reference: This step is similar to copying a dataset from Kaggle: ht[tps://www.kaggle.com/datasets/varunkumari/ai-bank-dataset](http://www.kaggle.com/datasets/varunkumari/ai-bank-dataset)

# Step 2: Use Databricks Notebooks to remove the duplicates

1. Use Databricks Notebooks (use aggregate transformation/windows/assert)
   1. Read Data: Read data from five different sources within the raw(bronze) container.
   2. Data Cleaning: Implement logic to identify and remove any hanging or irrelevant data from the sources.
   3. Data Transformation: Apply necessary transformations to prepare the data for further processing. This might involve schema changes(avoid inferSchema), data type conversions, or handling missing values (parquet or delta)
   4. Create a scope using key-vault
   5. Use Mount point for connecting to ADLS Gen 2
   6. Pass the files
   7. Join the 5 files and generate a single file based on columns below:
   8. Account ID, Transaction ID, Customer ID, Loan ID, Payment ID, Amount, Dates (put in one separate file)
   9. Remove the duplicates for the above combination
   10. The file should be generated in the DELTA file format
   11. Other files can be in Parquet format, and then do SCD Type 1

# Step 3: Use Notebooks using SCD Type technique (SCD 1)

1. Use Notebooks in the pipeline
   1. Place data into ADLS GEN2 in DELTA format (GOLD Layer)
   2. Schedule the Pipelines, in the data bricks itself

# Step 4: Use Power BI for Data Visualization

* Use ADLSGEN2 as a source to PowerBI Reports and build visuals on top of it
* Publish the developed report into the Fabric Workspace

Repeat the Steps using the automated triggers.

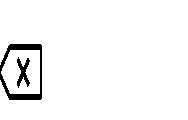
# Additional Considerations

* **Dynamic Parameters:** Incorporate dynamic parameters within your data pipelines to enable configuration changes without modifying the pipeline code itself. This enhances flexibility and simplifies pipeline management.

* **Key Vault:** Utilize Azure Key Vault to securely store sensitive information used in your pipelines, such as connection strings or credentials.

# 

# Deliverables

1. Documentation: Create a well-structured document outlining the steps involved in the data pipeline, including screenshots and detailed explanations. [Include Screenshots]
2. Code Snippets: Share relevant code snippets from your Databricks notebooks (SQL and/or PySpark) and pipeline activity configurations (JSON) in a dedicated repository (e.g., GitHub).
3. Follow the Naming convention while saving your project documentation

***“Your Name\_Project no.\_Project Nameˮ*** Example file name: Sruthi\_BC001\_Dataset and share in your individual group

1. Draw an architecture diagram (Use Draw.io), and use connections between them