Bootcamp Project 2 – Financial Data Analysis

# Outline of the Project:

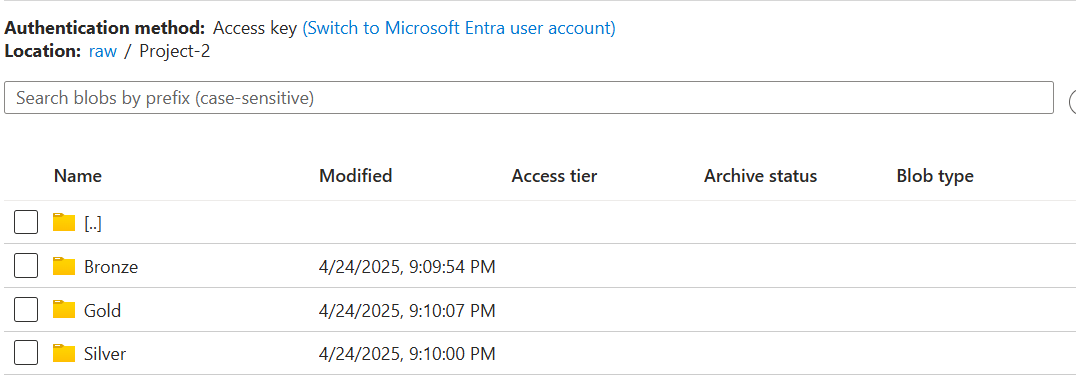
You are given 5 csv files which represent data related to Transactions and Loan acquired by Customers. Your task is to do proper ETL process for the data which includes data cleaning and transformation for storage purpose and store the processed data into SCD Type 1 Dimension Tables. Also, create table which encapsulate combined data created by joining raw tables. And At last, populate the data to Desktop Power BI.

Use only Databricks for all the computation and transformations.

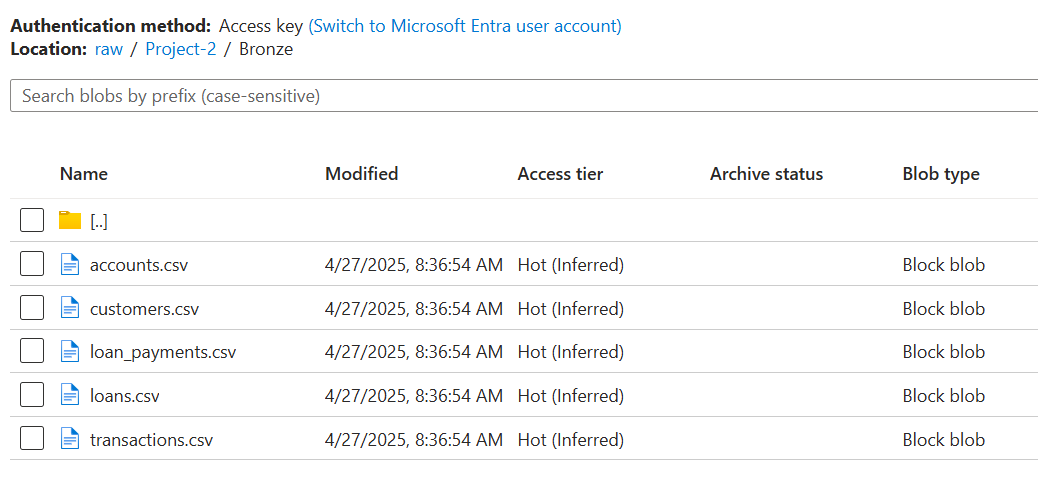
# Work Flow of Project:

*Storing of the CSV files in ADLS G2 (bronze layer)*

* I’ve created 3 directories in ADLS G2 (bronze, silver, Gold) which willbe used in the project to follow medilean structure:

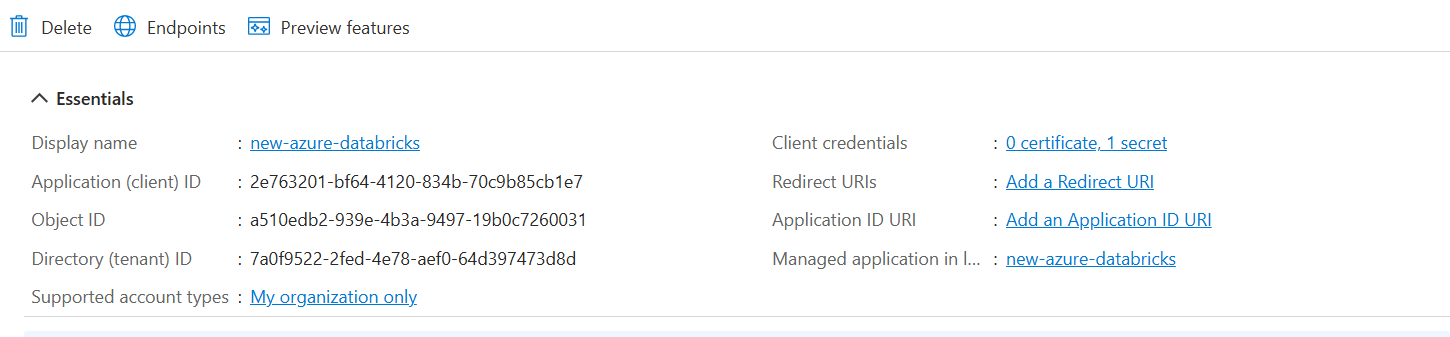


* I’ve stored 5 csv files in bronze directory:



Creating connection between ADLS G2 and databricks using Service principal method:

* To create service principal, we need to do new app registration in Microsoft Entra ID:



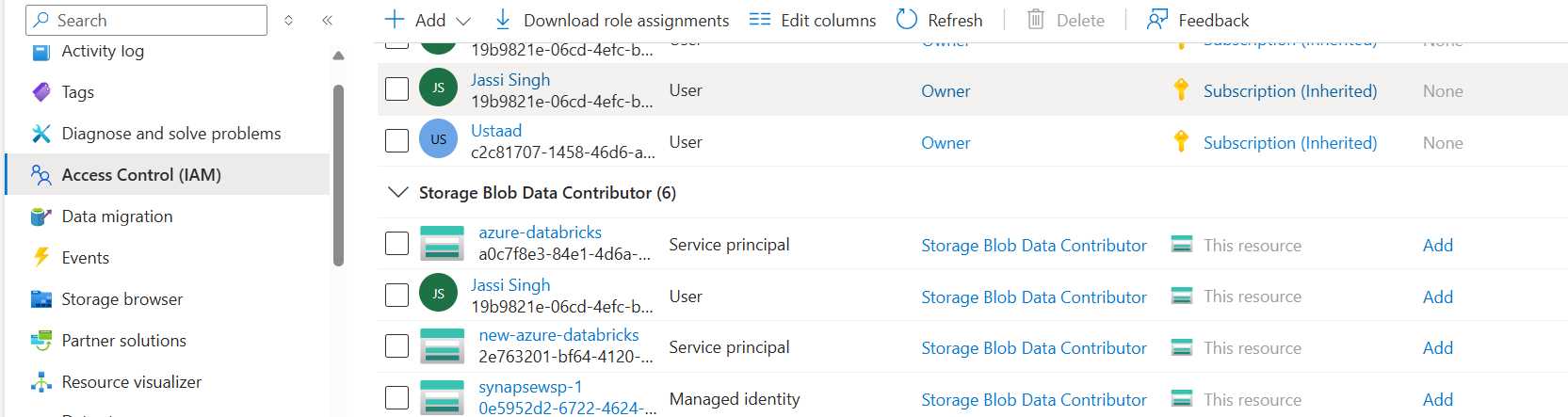
Store application ID and Tenant ID of the app in your system for future use.

Next, we need to create Client secret.

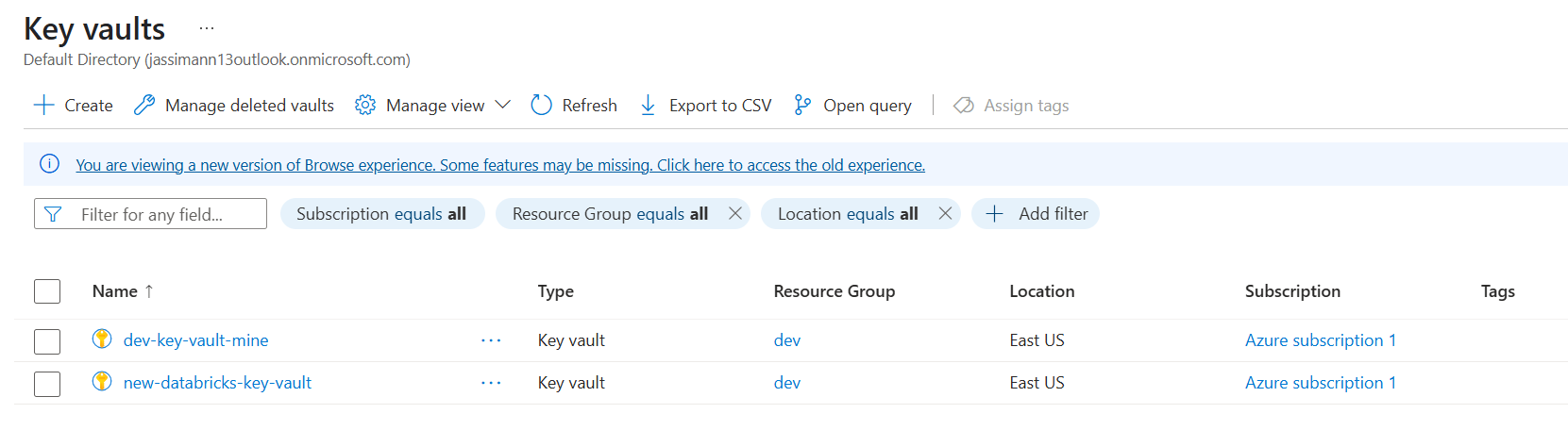
Create the secret and store it’s value with you for future use (we cann’t access this value later)

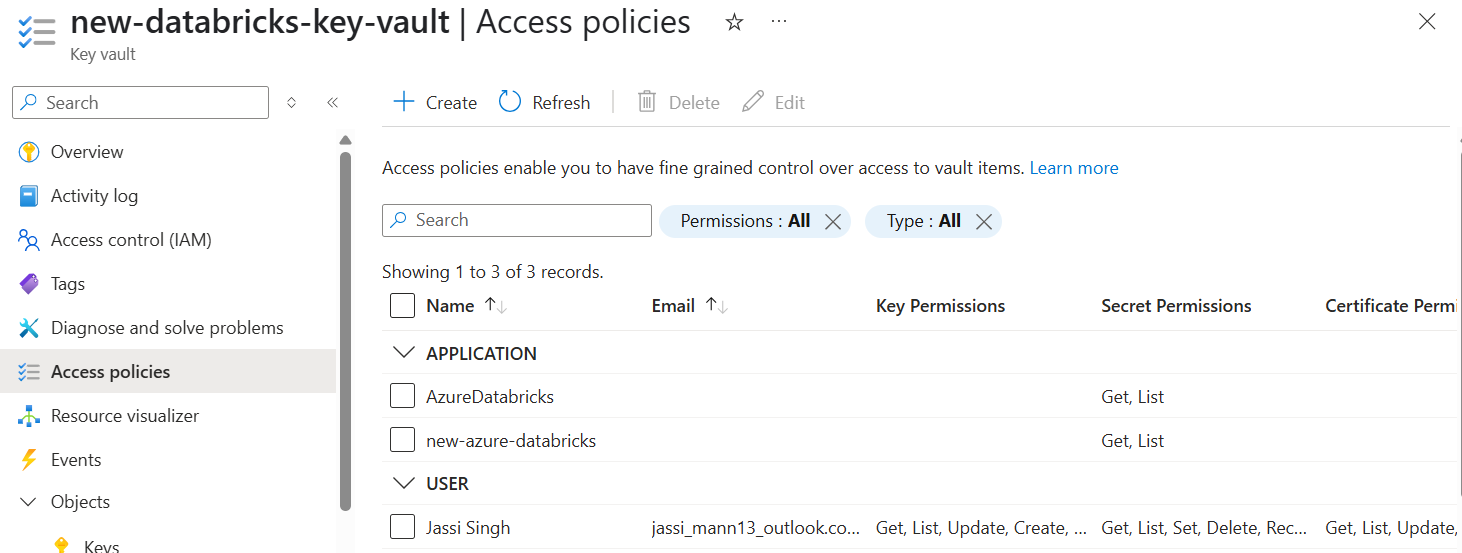
After saving all 3 info in local, we need to give Storage Blob contributor role to this new service principal.

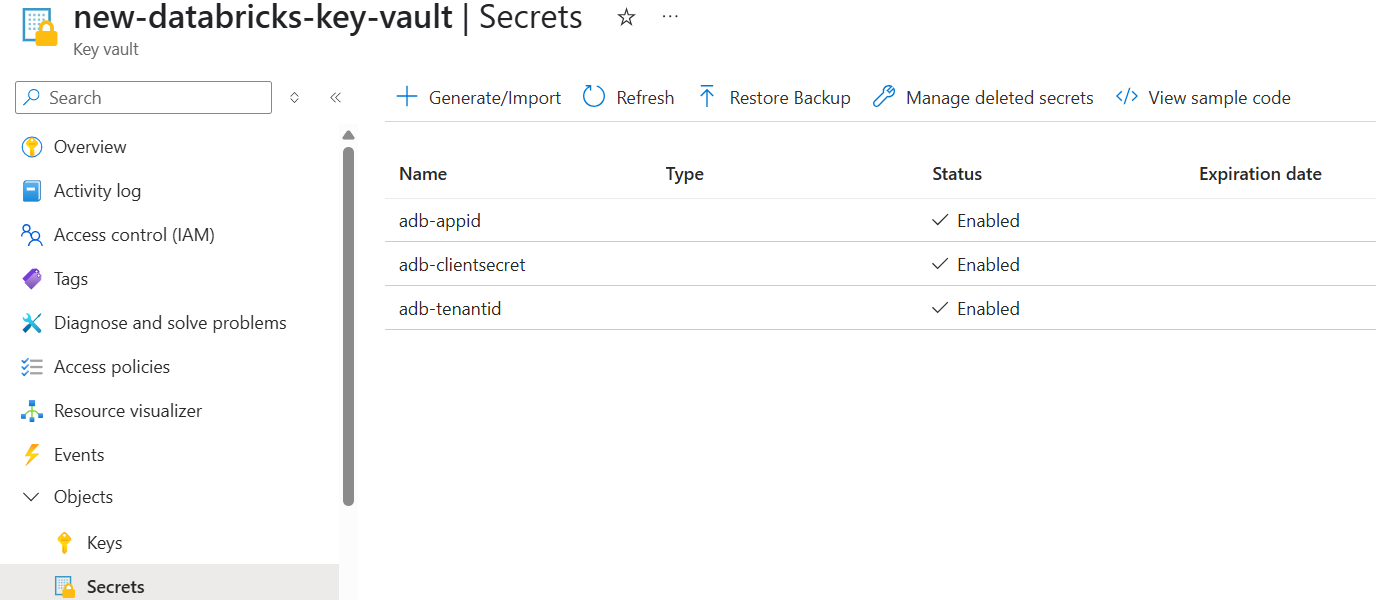
To do that, go to ADLS G2 contianer> IAM access> add new role>select Blob Storage contributor role> assign to new service principal name.



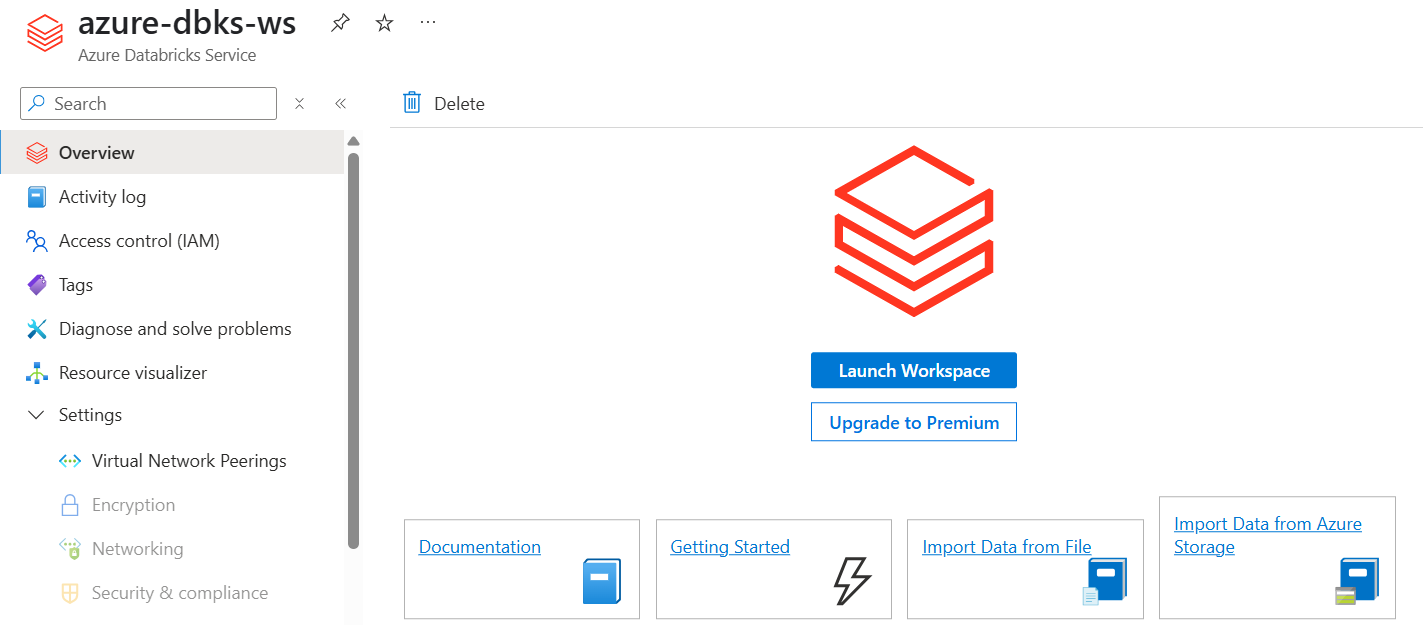
* Creation of key vault and add 3 secrets in key vault and give permissions of reading and listing secrets to New service principal:







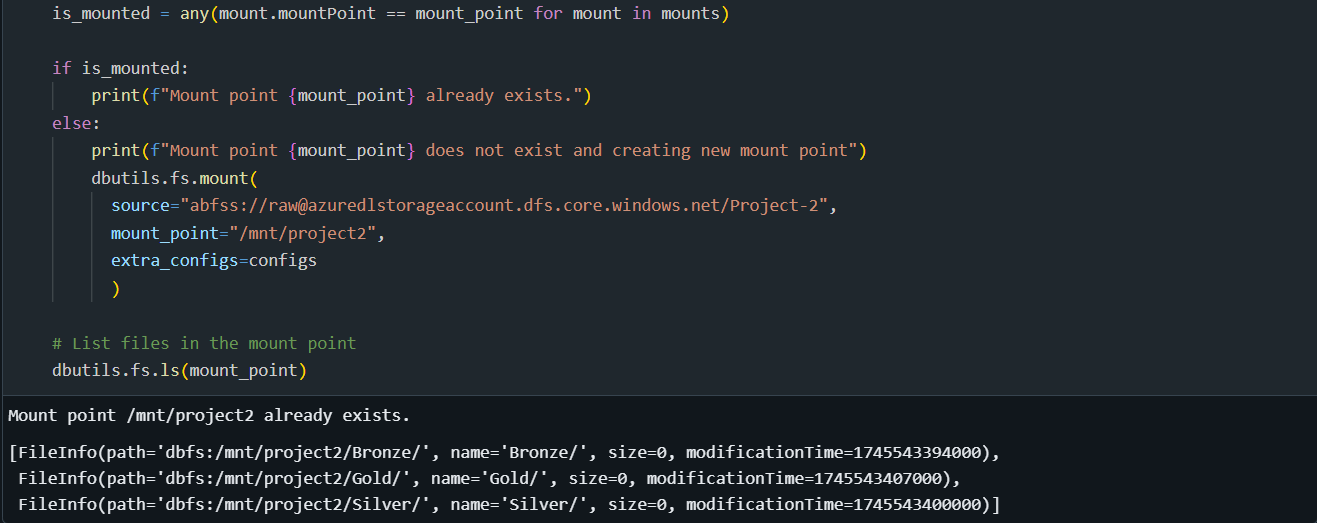
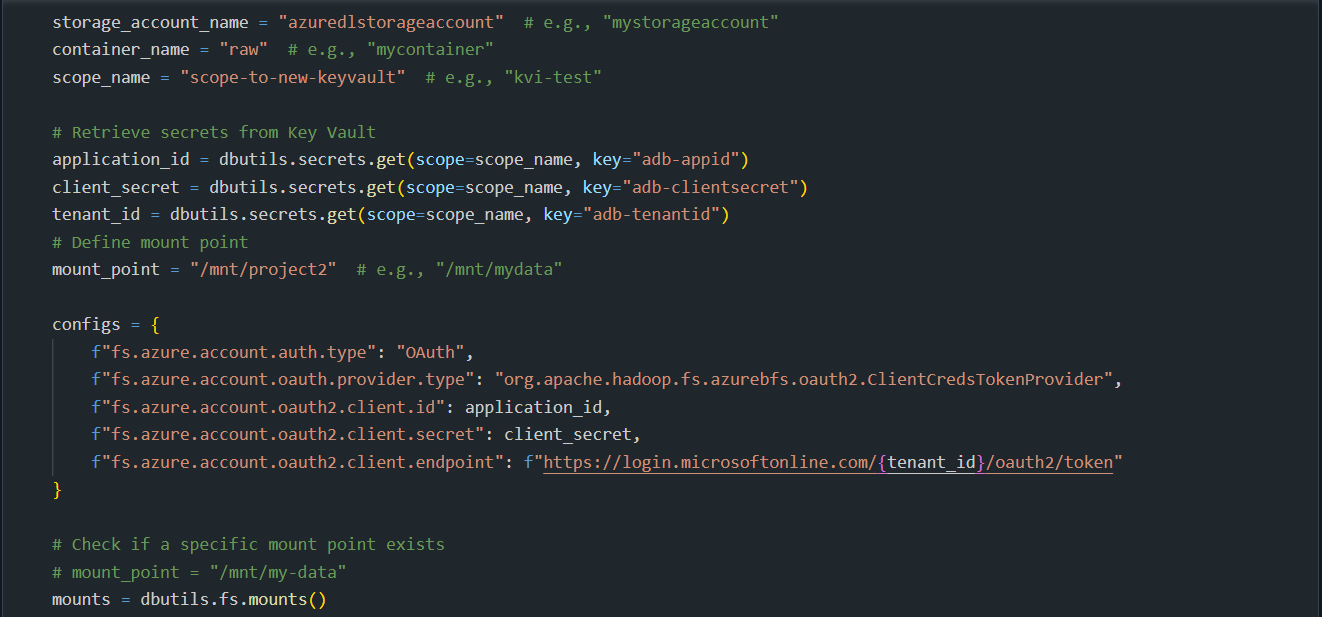
* Creation of Azure Databricks Workspace



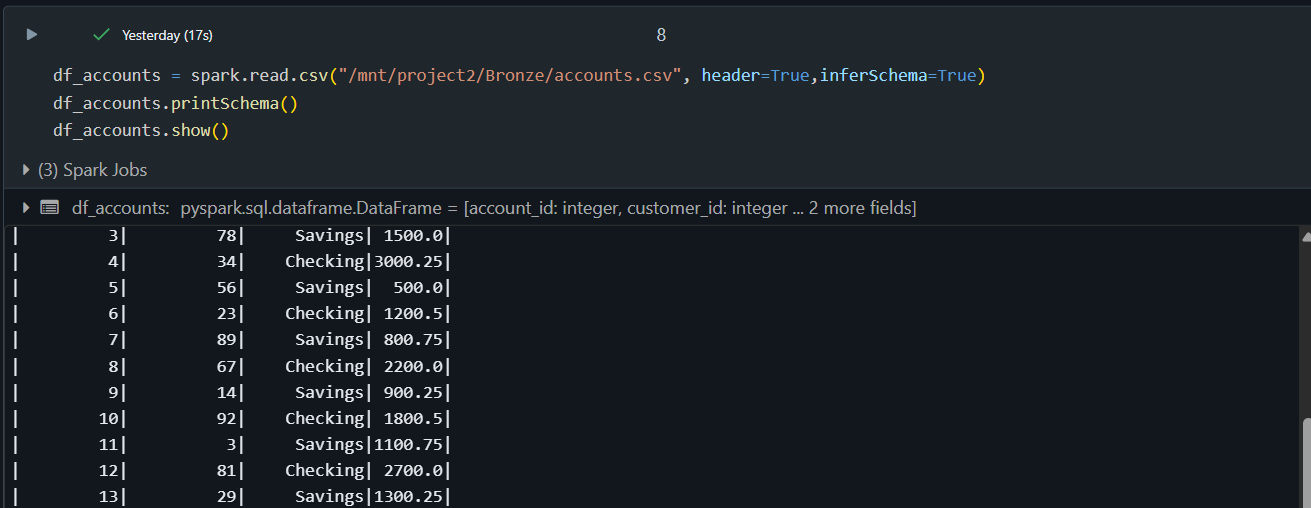
* To access ADLS G2 from Databricks notebook, we need to create scope in notebook so that, we can access key vault secrets we have created. For that, we need connection string/URI of Key vault and resource ID of it too.
* Go to <-------databricks url---->/#secrets/createScope
* Then you can access the key vault secrets by using dbutils.secrets commands



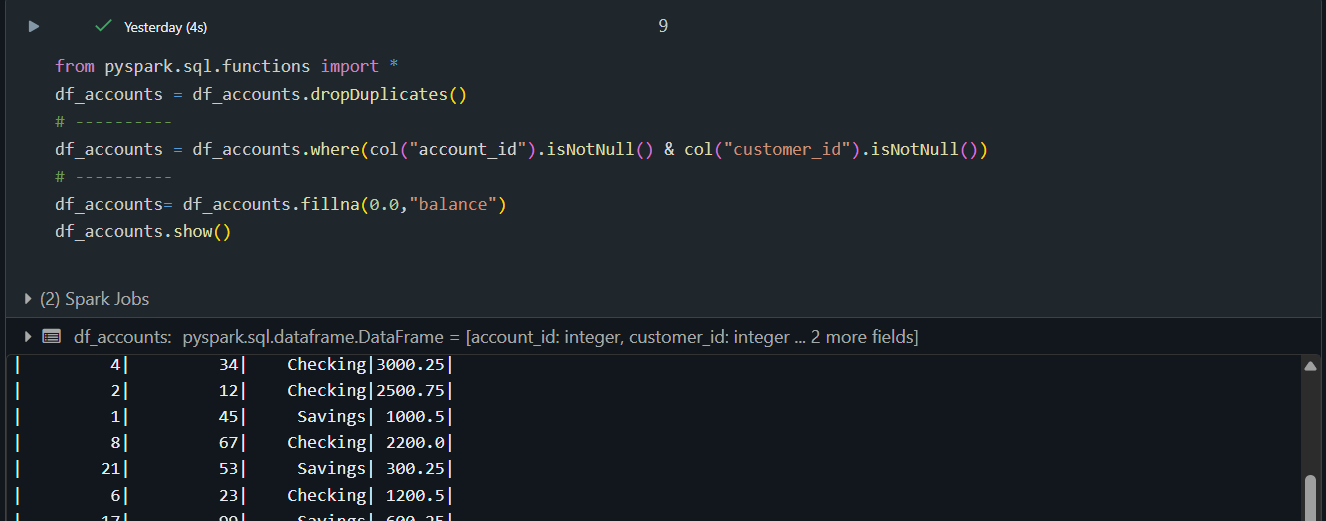
* Connect with ADLS G2 using service principal method by using following Code:



* To Read CSV file from ADLS G2 directory :

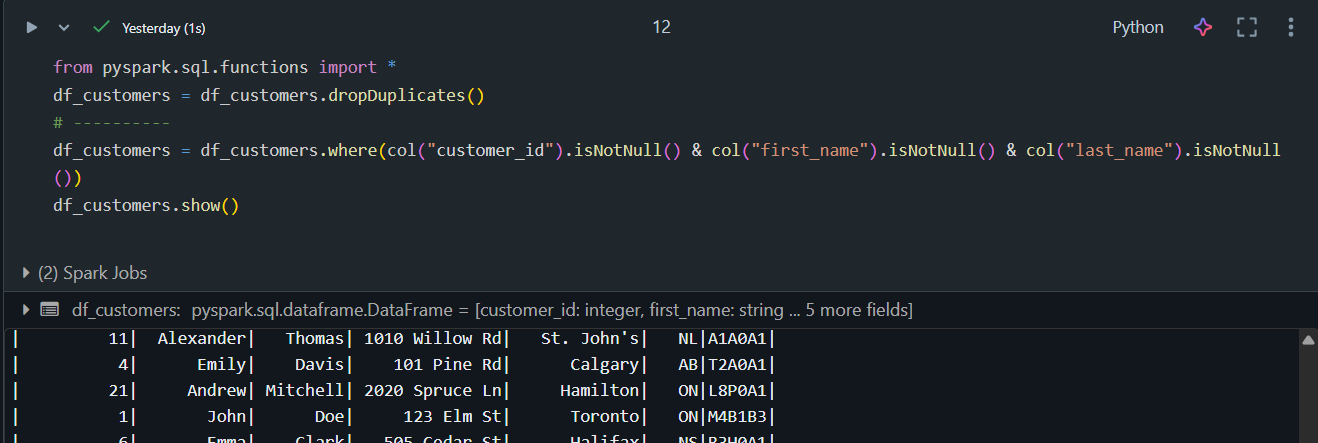


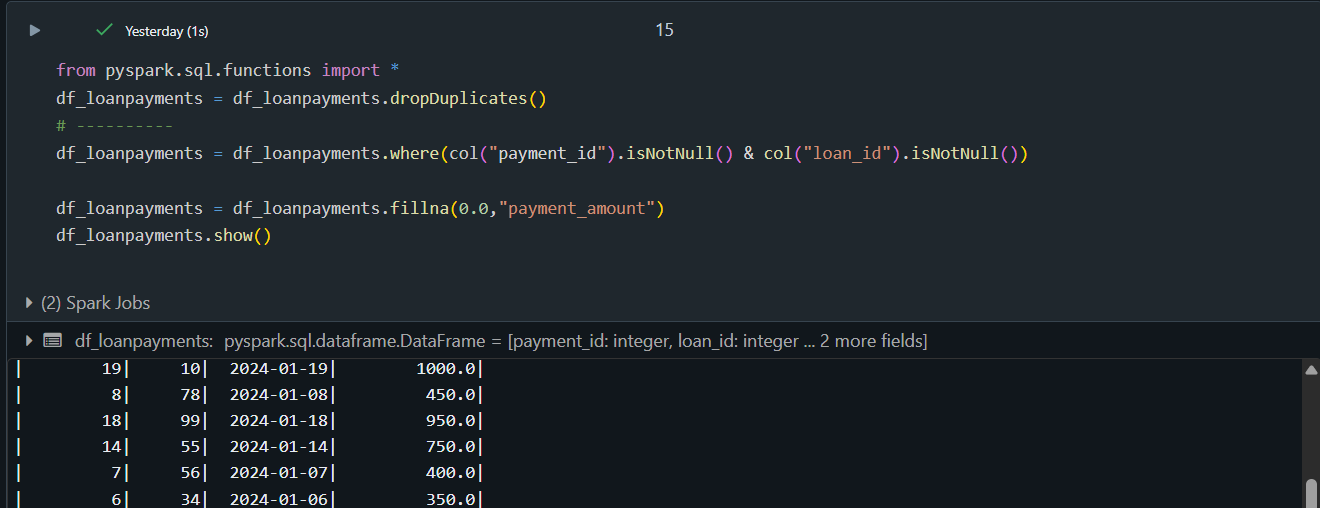
* *Data Transformation I did were :*

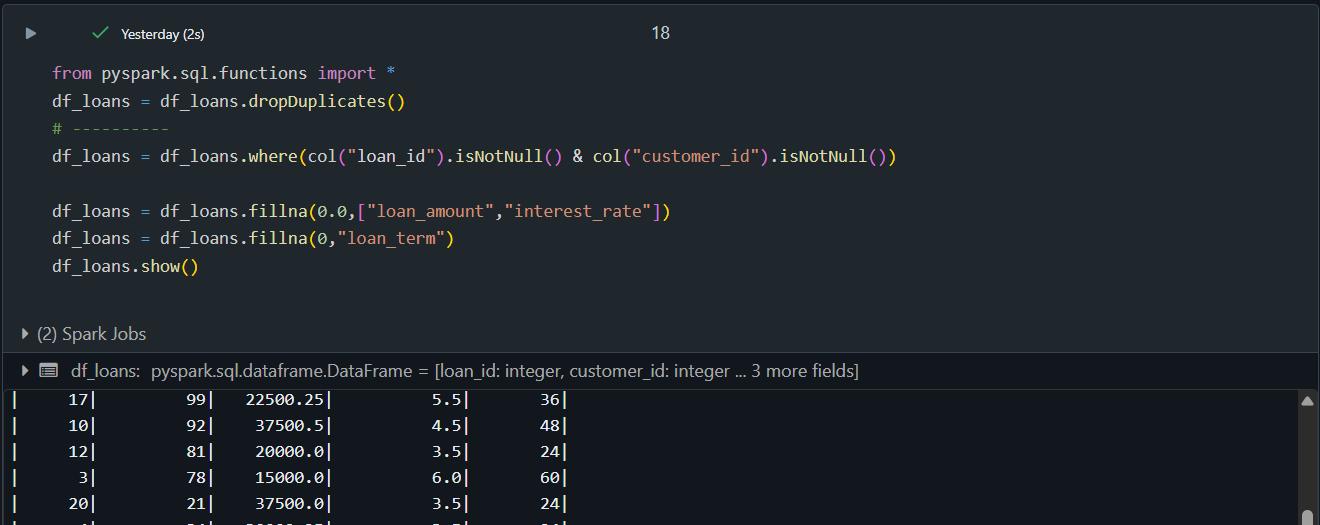


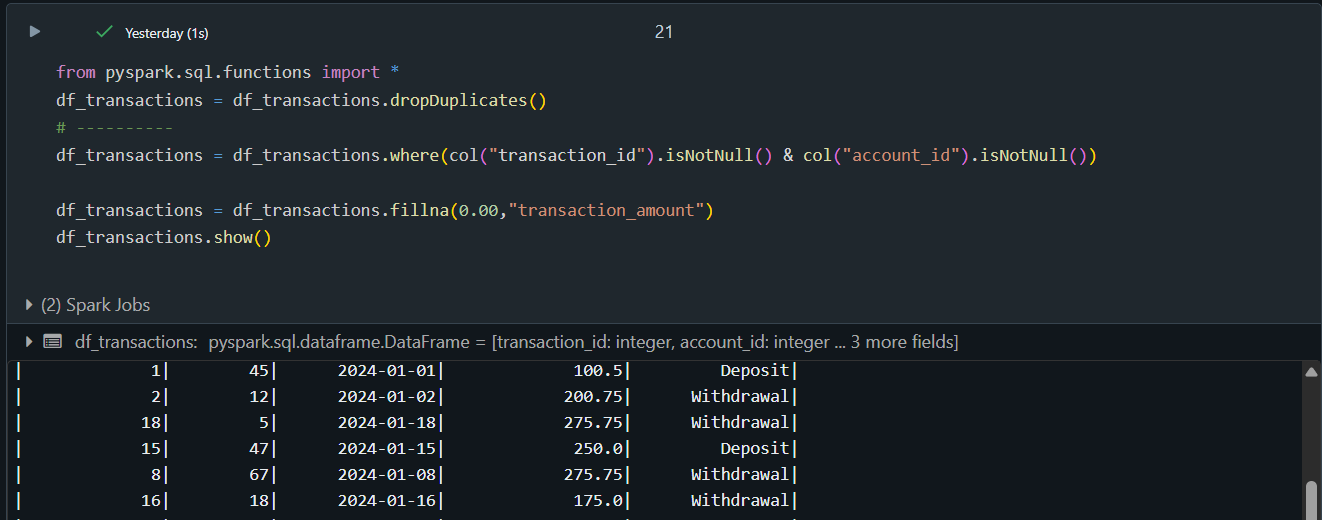
1. Dropping duplicates records
2. Filtering out records, which have non-null values of particular columns
3. Setting up default values to those column values which will be NULL.

* I’ve applied these transformations to all 5 Dataframes:

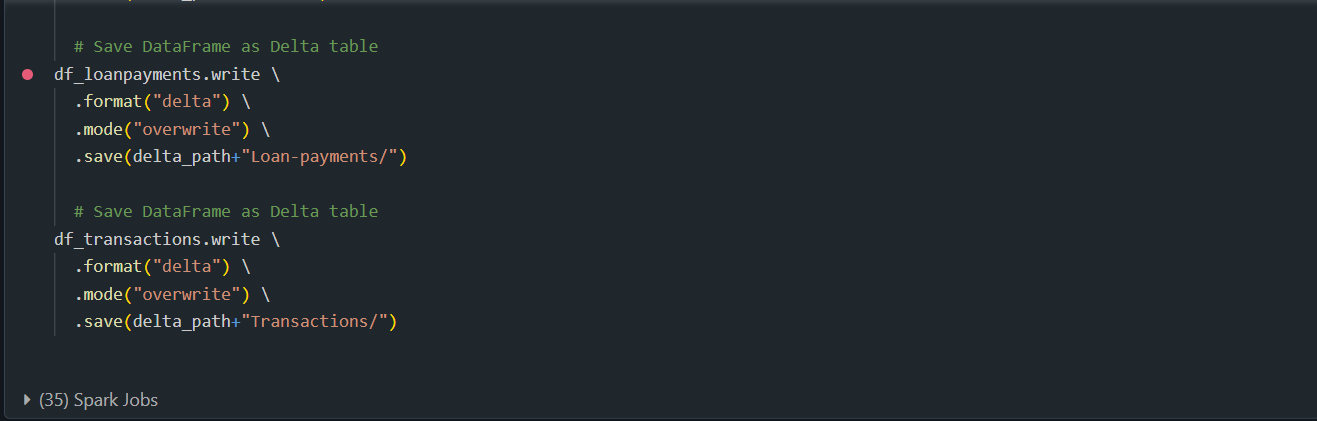
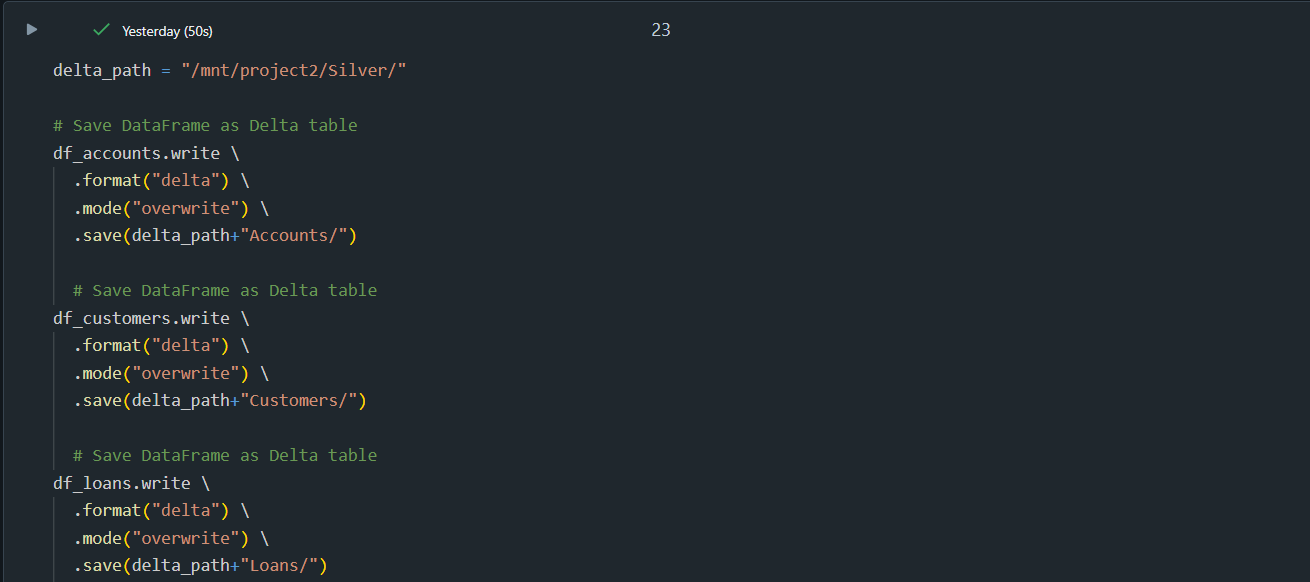






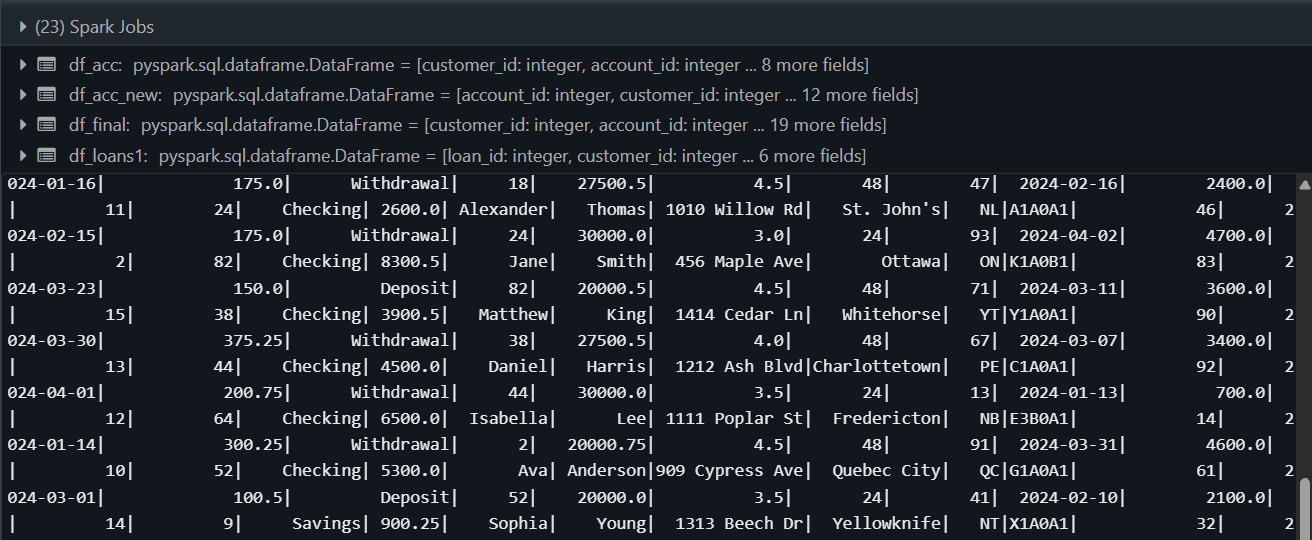


* Saving all transformed Data frames into delta format in silver layer – ADLS G2

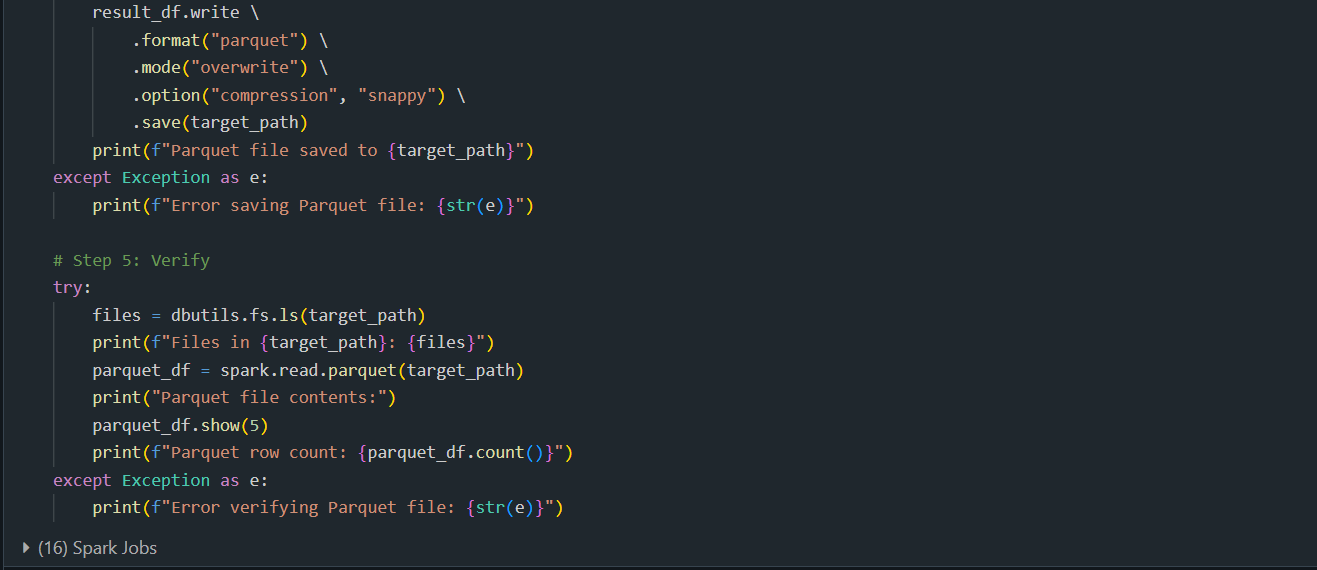
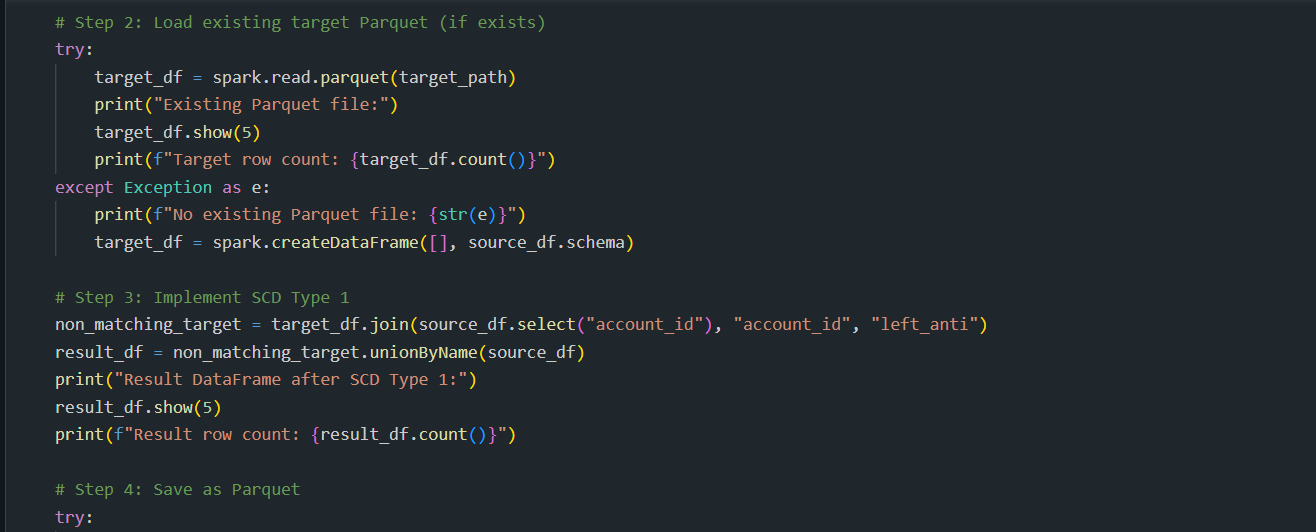
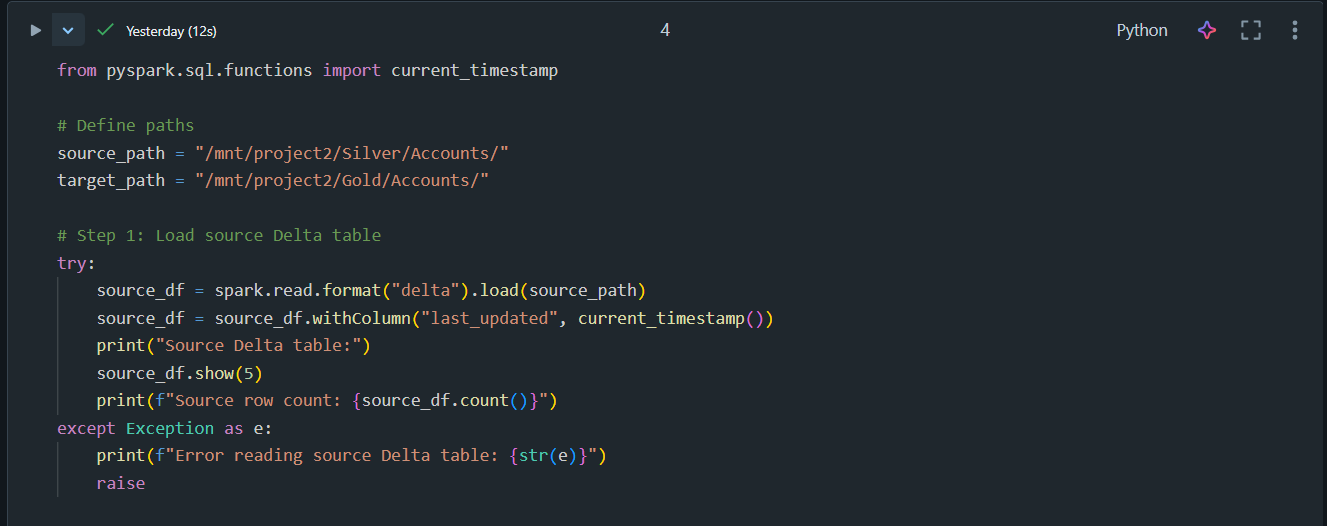


* To combine all Data frames into one common DF by using their relationships with each other, I’ve created a new Data frame and store it into delta table in ADLS Silver layer.

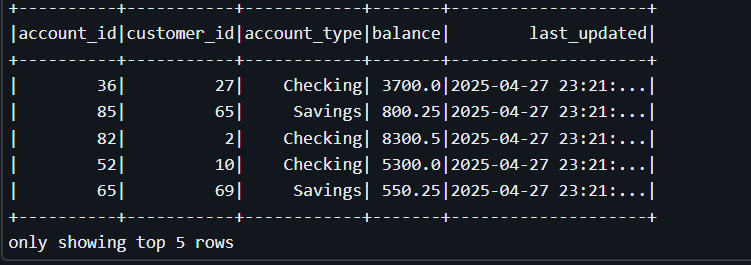




# CONVERTING DATA INTO SCD TYPE 1 DIMENSION TABLES BY USING DATABRICKS:



* New Data frame created :



# Populating Data on Power BI Dashboard:

