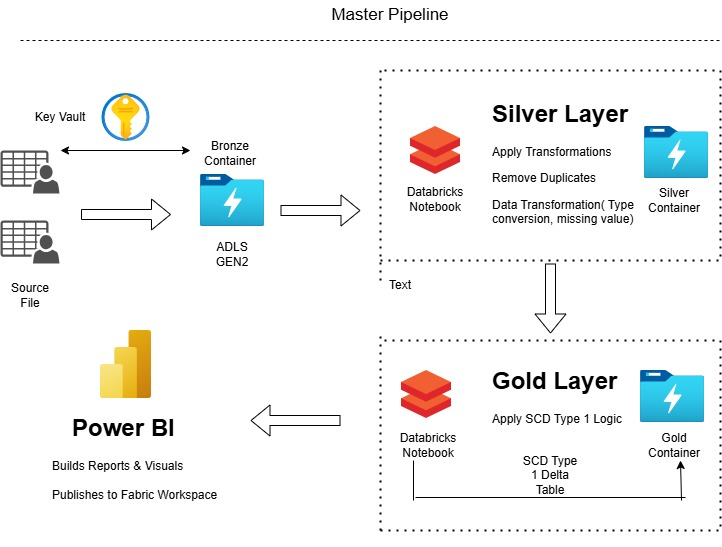
****

Architecture of the Project\_2

**1. Introduction**

This report documents the development and implementation of a robust data pipeline project for managing transactions and loan data of a customer. The pipeline is built using Azure Data Lake Storage (ADLS) Gen2, Azure Key Vault, and Databricks, with a final visualization layer on Power BI.

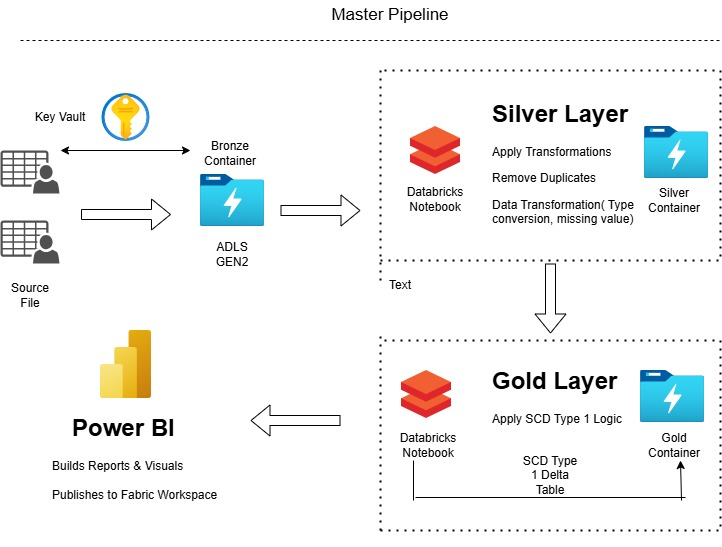
**2. Objective**

To design and implement a scalable and accurate data pipeline that ingests raw financial datasets, processes them into meaningful insights using SCD Type 1 delta format, and visualizes the outcomes via Power BI for better decision-making.

**3. Tools and Technologies Used**

* **Azure Data Lake Storage (ADLS) Gen2**
* **Azure Key Vault**
* **Databricks Notebooks (PySpark/SQL)**
* **Delta Lake**
* **Power BI**
* **Draw.io** (for architecture diagram)

**4. Project Architecture**



Architecture of the Project

**4. Project Architecture**

*An architecture diagram will be included here to show:*

* Bronze Layer (Raw Data Ingestion)
* Silver Layer (Cleaned/Transformed Data)
* Gold Layer (SCD Type 1 Delta Tables)
* Visualization Layer (Power BI)

**5. Step-by-Step Implementation**

**Step 1: Data Ingestion**

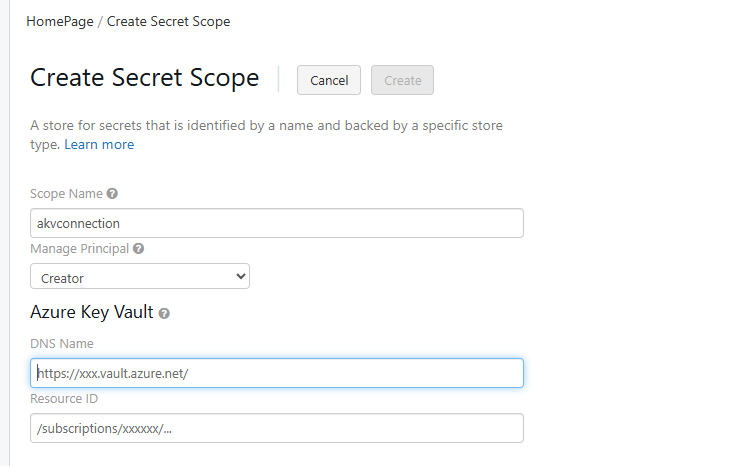
* Source: Backend team storage account containing files:
  + accounts.csv
  + customers.csv
  + loan\_payments.csv
  + loans.csv
  + transactions.csv
* Sink: ADLS Gen2 raw (Bronze) container.



List of the files

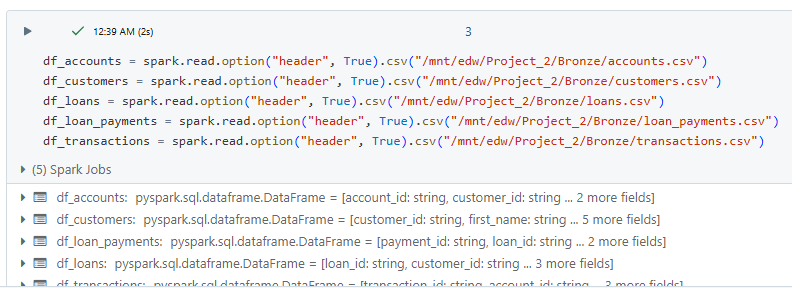
**Step 2: Data Cleaning and Transformation**

* Environment: Databricks Notebooks
* Mount ADLS Gen2 using scope from Azure Key Vault.

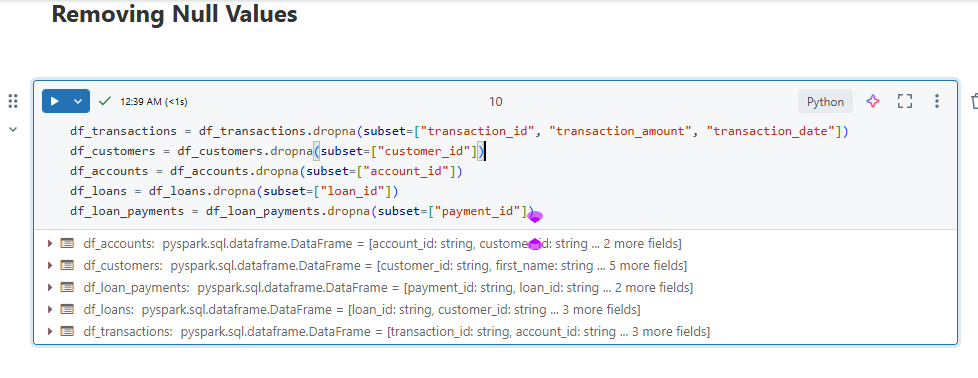


Created the Scope and stored the Connecting Details

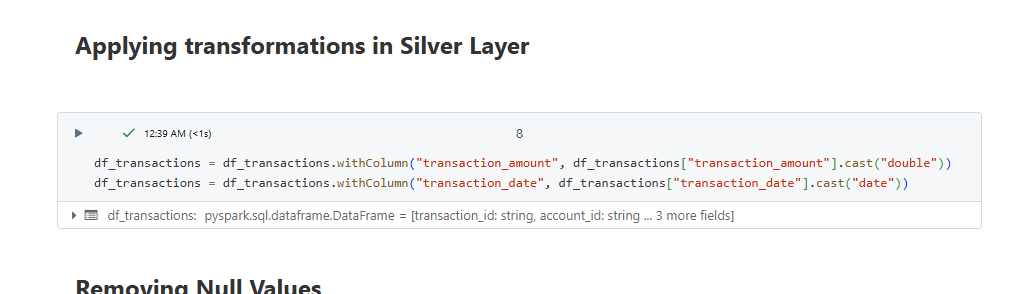
* Read data using Spark with explicit schemas (avoid inferSchema).



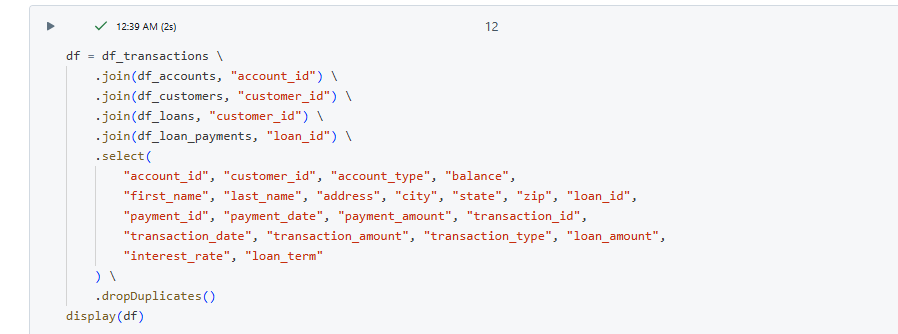
* Cleaning: Remove irrelevant records, handle missing values.



Removed the Null Values from the all the 5 Source files.

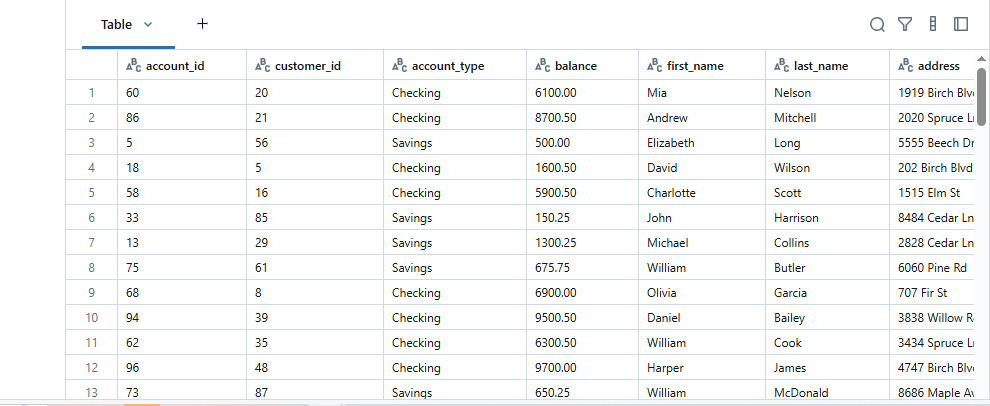


Changed the Data types in the few columns in the file like here we changed the transaction\_amount to double and changed the datatype of transaction\_data to data

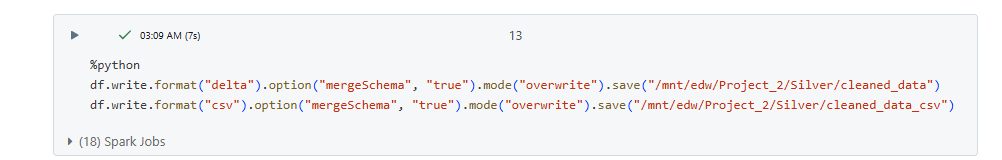
* Transformation:
  + Join all five datasets based on Account ID, Transaction ID, Customer ID, Loan ID, Payment ID.

Joined all 5 Source files and stored the unique columns

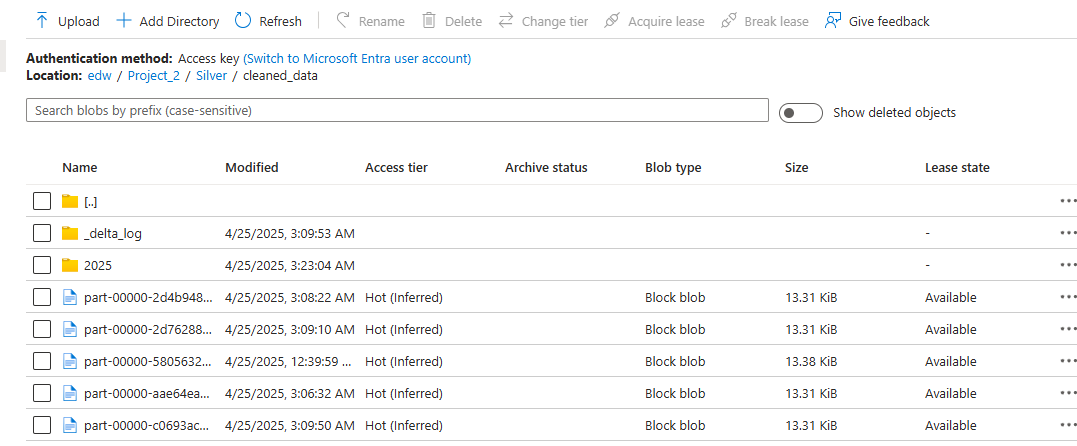
* + Final output file includes: Account ID, Transaction ID, Customer ID, Loan ID, Payment ID, Amount, Dates.

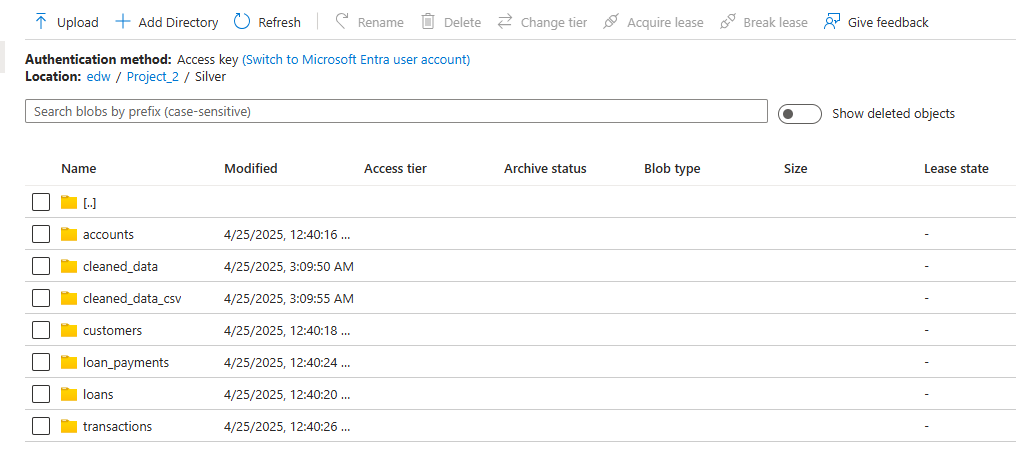
This is the final output of the final joined file.

* + Write cleaned file in Delta format.



Write the combined files in the edw

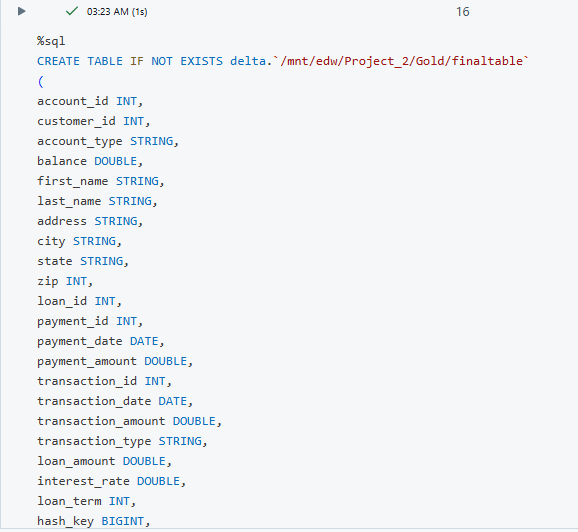


* + Other processed datasets saved in Parquet format.

These are the files Stored in the form of Parquet form.

**Step 3: SCD Type 1 Implementation**

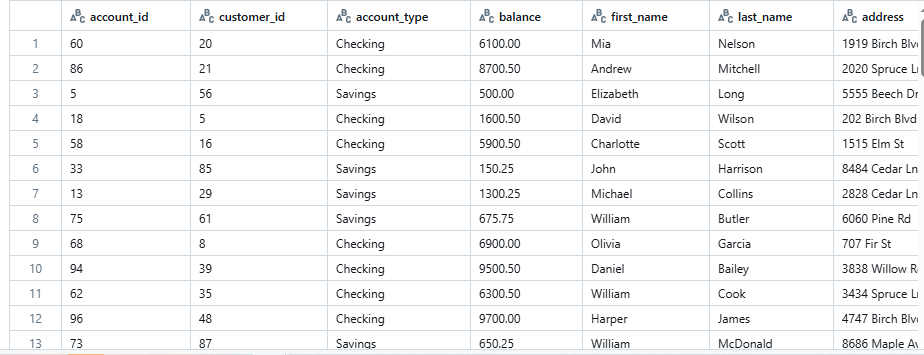
* Create Delta tables using Databricks Notebooks.
* Store final curated tables in the Gold layer (ADLS Gen2).
* Implement upserts (overwrite) logic to maintain SCD Type 1.

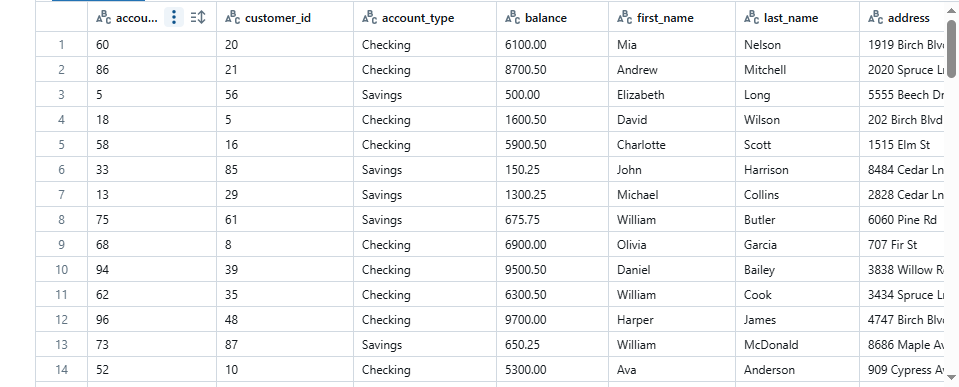


This creates a Target destination where we can see the updated data.

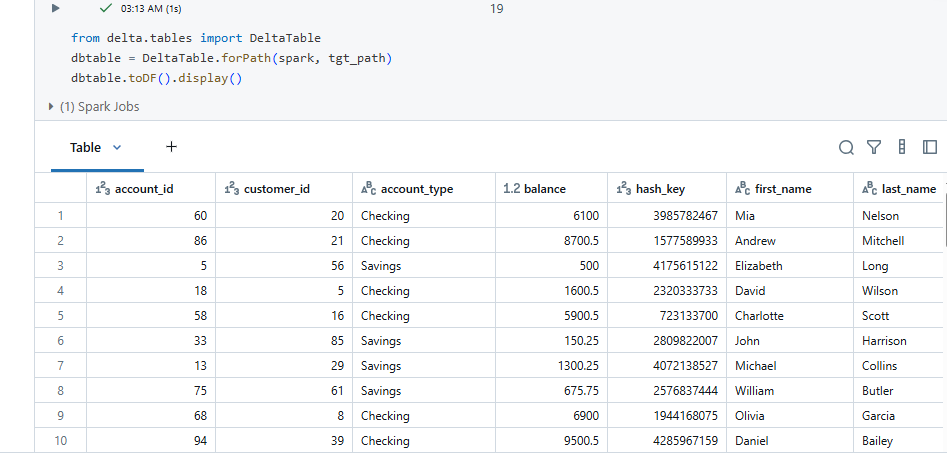


With the Help of we can create the Parameters in the PySpark in this we taking the file date as the input. 

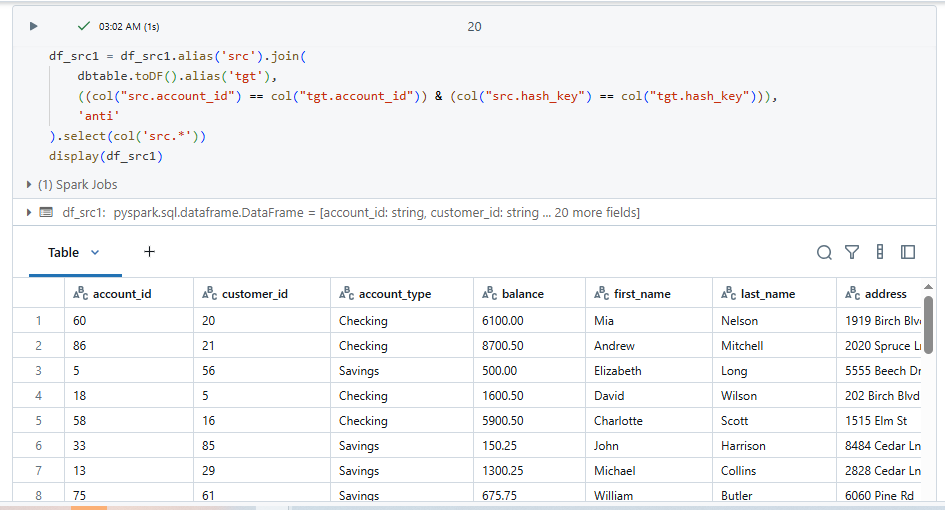
In this step we will be taking the source and the target path of the file and displaying it below.



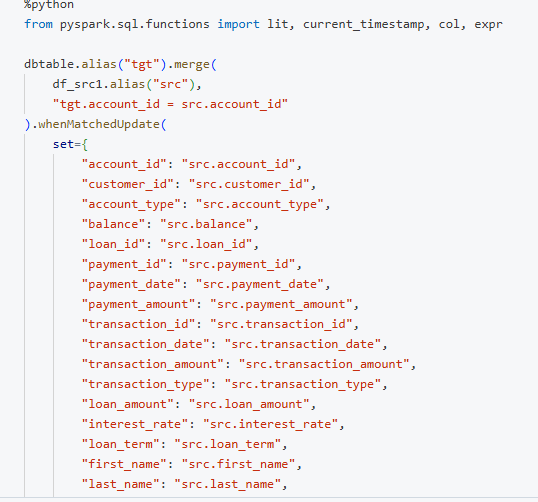
These are the source and the target file.

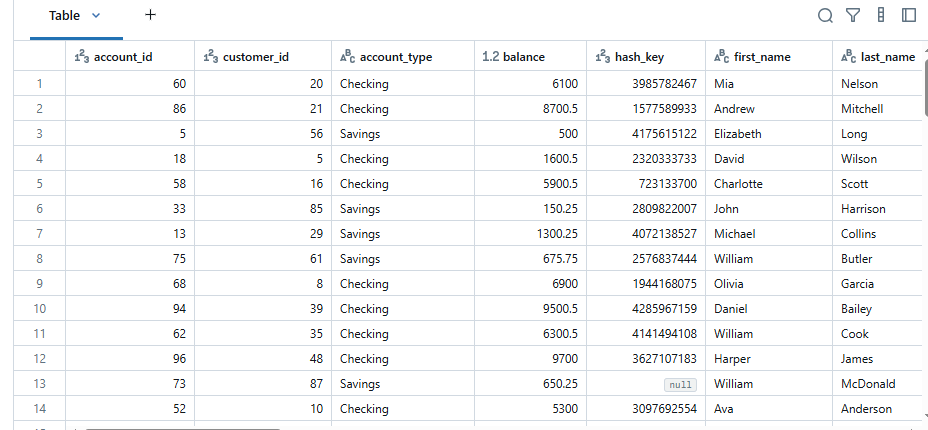


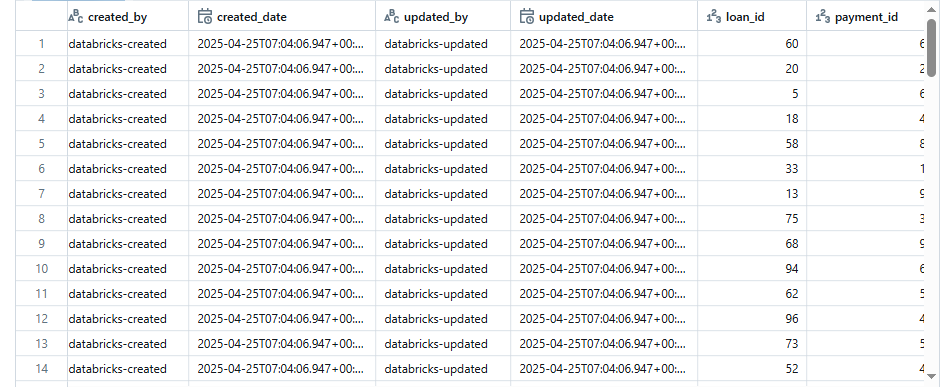
Here we converting the table into the dataframe inorder to perform scd type 1 on the table.



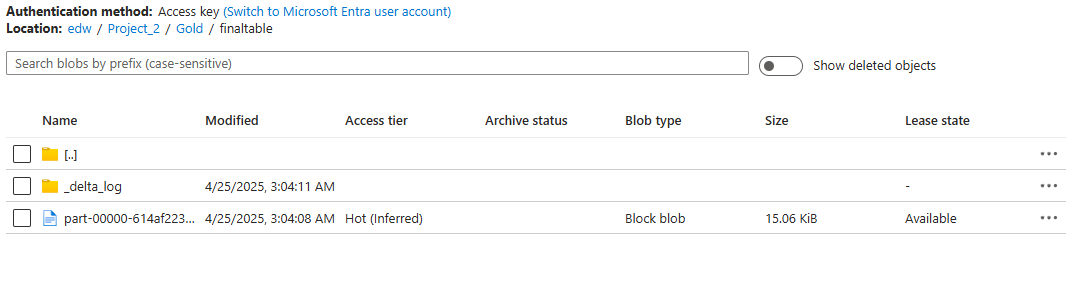
In this step we are joining the source file and the Target file with the help of account\_id column and the hash\_key

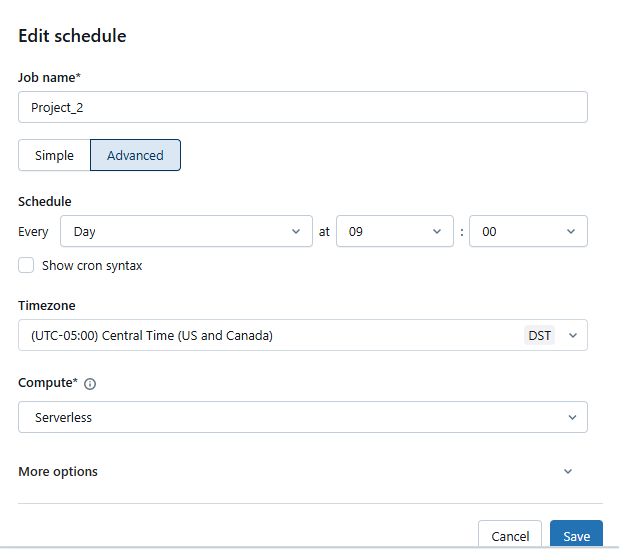




This is the final table after performing SCD Type – 1



Schduled the Workflow daily at 9 AM.



**Step 4: Data Visualization with Power BI**

* Connect Power BI to ADLS Gen2 (Gold Layer).
* Develop visuals such as:
  + Loan disbursement trends
  + Payment defaults
  + Customer demographics
  + Transaction volume over time
* Publish dashboards to Fabric Workspace.

A screenshot of a computer screen

AI-generated content may be incorrect.

Accounts Report

A screenshot of a computer screen

AI-generated content may be incorrect.

Customer Report

A screenshot of a computer

AI-generated content may be incorrect.

Loans Payments Report

A screenshot of a graph

AI-generated content may be incorrect.

Loan Report

A screenshot of a graph

AI-generated content may be incorrect.

Transaction Report