Bootcamp Project 1 - Data Pipeline for Customer Account Analysis

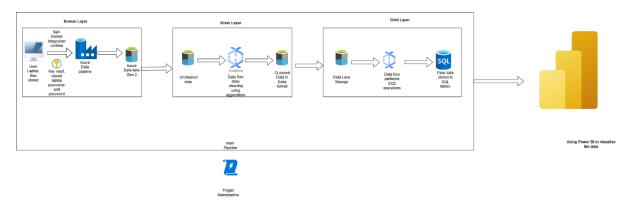
1. Objective:

The project aims to design and implement a robust data pipeline for processing customer account data. This includes copying data from a backend team's storage account, performing necessary transformations using ADF and upserting (inserting or updating) data from a file stored in Azure Data Lake Storage ADLS GOLD Storage into SQL database table. The pipeline aims to ensure efficient, accurate, and scalable data processing to support downstream analytics and reporting needs.

2. Tools Required:

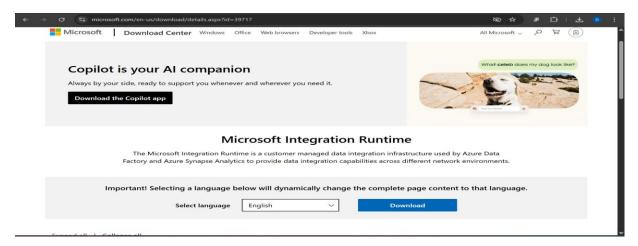
Azure Data Factory Local System Azure SQL data base

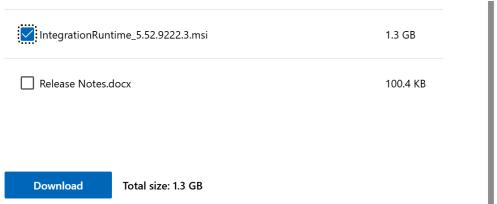
3. Created an Architecture Diagram

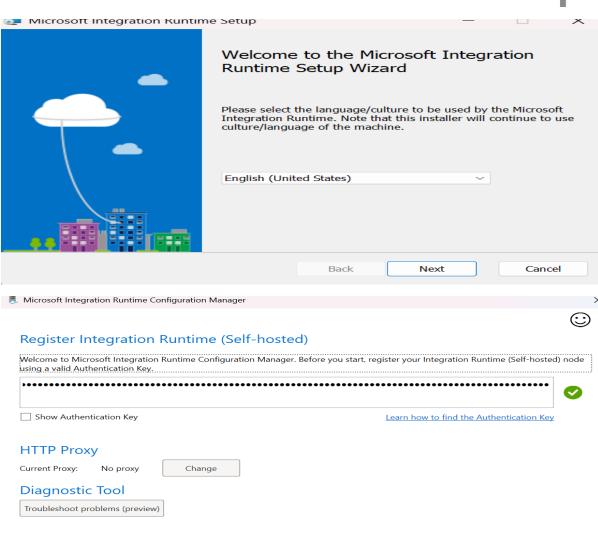


4. Started Creating the Pipeline

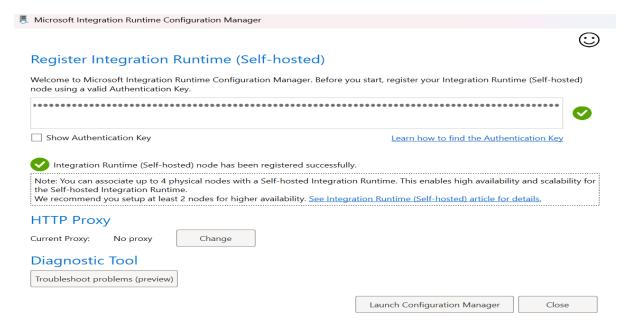
- 1. First Data pipeline, is to copy data from on premise file storage system to ADLS into Bronze folder.
 - First, we have to download selfhosted integration runtime in local system to connect to Azure cloud.



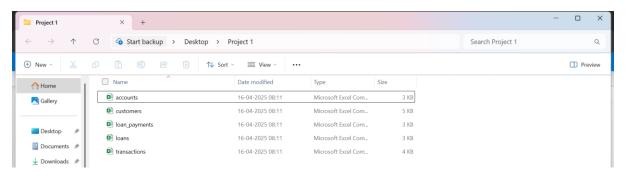




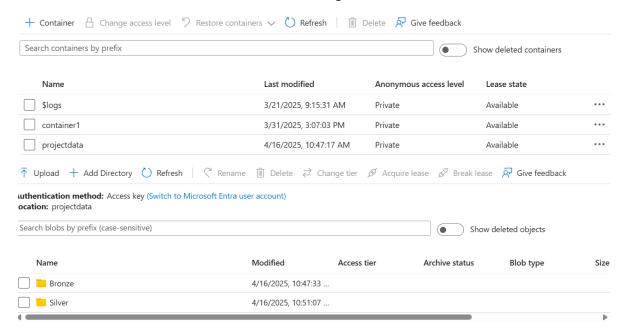
Register Cancel



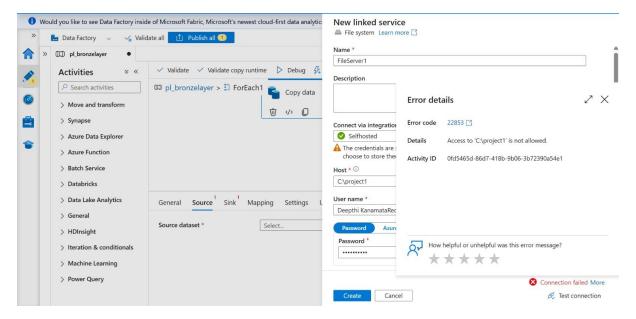
- Selfhosted integration runtime is successfully installed and registered in local system
- This is where files in local system are located.



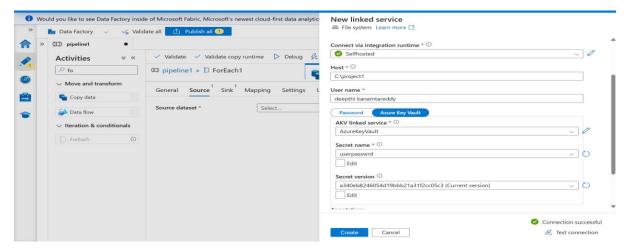
• Now created a container and folders in ADLS gen 2.



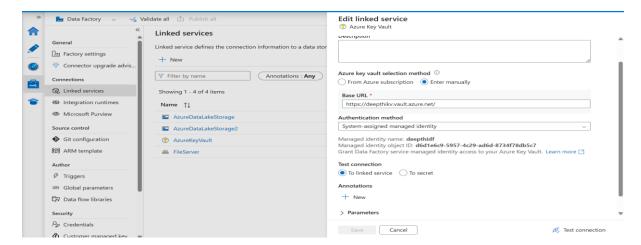
• Now try to create a linked service between local system and Azure cloud.



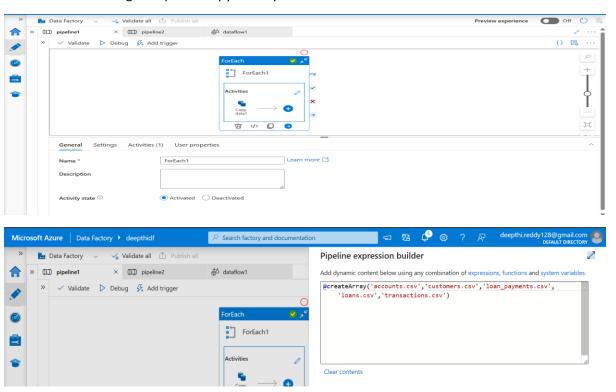
- Faced a connection error while trying to connect to local system, resolved it by using below commands in window power shell.
- First changed directory in power shell using command cd "C:\Program Files\Microsoft Integration Runtime\5.0\Shared".
- Once directory is changed ran command .\dmgcmd.exe -DisableLocalFolderPathValidation, which gave access for selfhosted run time to access local system files.



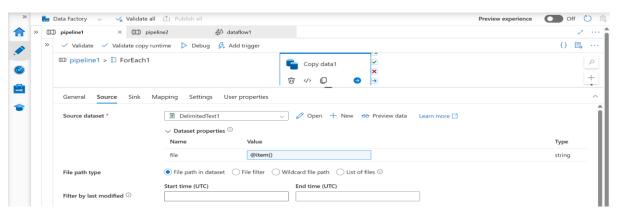
- Connection created successfully.
- Created a linked service for key vault where I stored system password as a secret.



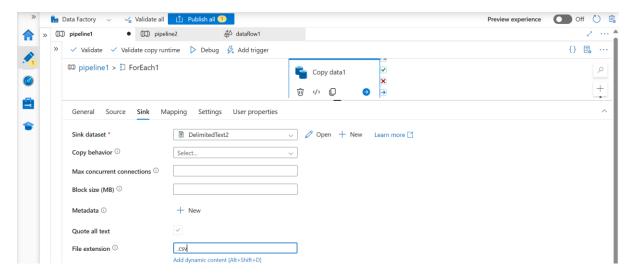
 Created a pipeline with foreach activity, which will loop through an array of file names and will give input to copy activity.



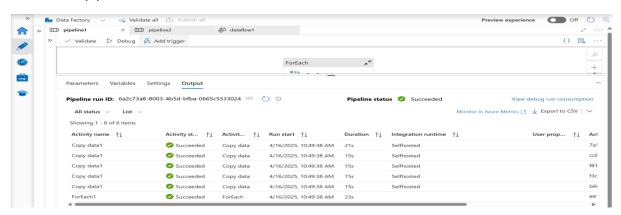
Copy activity source.



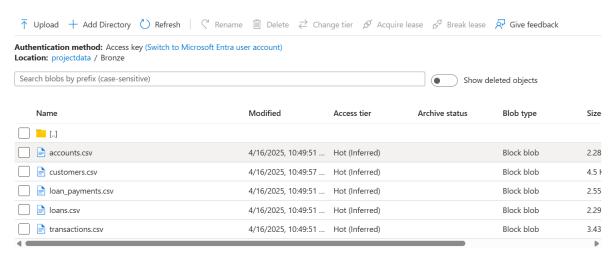
Copy activity sink.



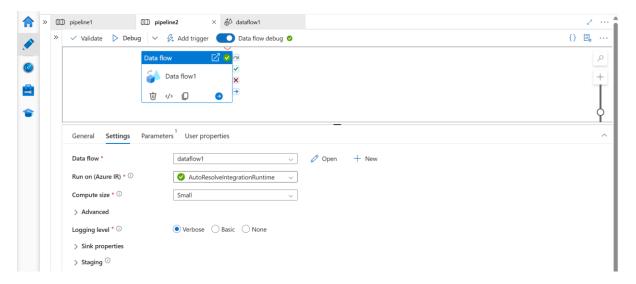
• Ran pipeline.



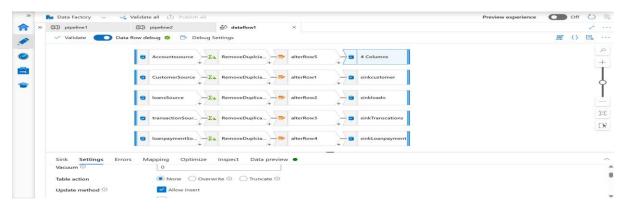
• Pipeline ran successfully and data from on premise is copied to Bronze layer in ADLS.



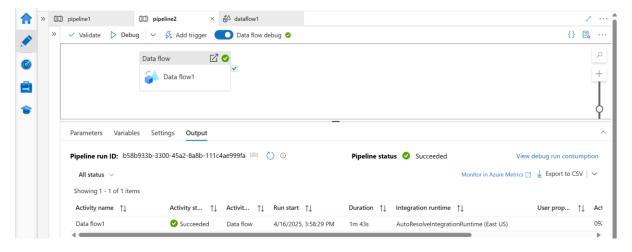
- 2. Create second pipeline for silver layer, where we have to clean the data, which involves removing duplicates.
 - Create a pipeline, with a dataflow



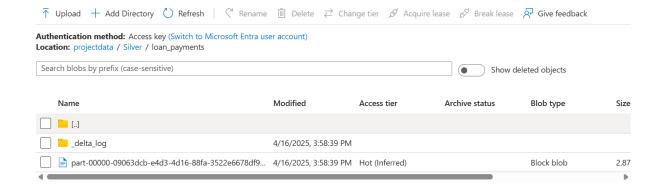
- Inside dataflow, we will have source, Aggregate transformation, alter row transformation and sink for all 5 files.
- Here we are using groupBy(), to remove duplicates by grouping them with specific ID's
 and then aggregating all other columns by using first() which return first not null values.
- Alter row transformation is used to perform upsert operation.



• Run pipeline successfully and data is loaded in silver folder in ADLS.



• Data is loaded to specific folders in Silver Directory for each file.

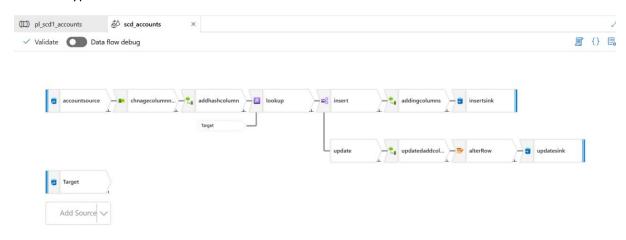


3. Now we have to create 5 different pipelines to perform SCD type 1 and 2 on 5 different tables.

Here we are performing SCD 1 on accounts file, transactions file, customer file.

SCD type 2 on loans and loan_payment file.

SCD Type 1 on accounts file.



Here we are using a dataflow to perform SCD type 1 on accounts file,

1. Source Transformation

- Data is ingested from the Silver Layer.
- Acts as the initial source for the pipeline.

2. Select Transformation

• Used to **rename columns** as per the desired naming convention or schema requirements.

3. Derived Column Transformation (Hash Column)

- A new column is added using the crc32 hash function.
- The hash column helps in identifying data changes for comparison and update logic.

4. Target Setup – Azure SQL Database

An Accounts table is created in the target SQL database.

• The table includes all columns from the source data, plus additional metadata columns (created_by, created_date, updated_by, updated_date).

5. Lookup Transformation

- Performs a **Left Join** between the **source data** and the **target SQL table**.
- Used to check if a record already exists in the target.

6. Conditional Split Transformation

- Segregates records into **Insert** and **Update** branches:
 - o **Insert**: Records not found in the target.
 - Update: Records found with differences (based on hash comparison or existing logic).

7. Derived Column Transformation (Metadata Columns)

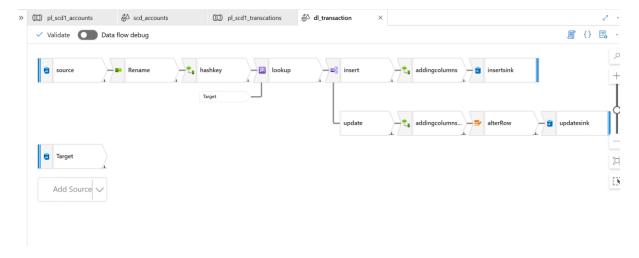
- Insert Path:
 - o Adds values to created_by, created_date, updated_by, and updated_date.
- Update Path:
 - Updates only the updated_by and updated_date fields.

8. Alter Row Transformation (for Update Path)

- Adds update conditions:
 - Sets rows to update if 1==1 (always true) as condition has already been logically filtered in the Conditional Split.

9. Sink Transformation

- Final data write operation:
 - o Insert Path Sink writes new records to the Azure SQL Accounts table.
 - Update Path Sink updates existing records based on the Alter Row configuration.
- SCD type 1 on transaction table.



1. Source Transformation

- Data is ingested from the Silver Layer.
- Acts as the initial source for the pipeline.

2. Select Transformation

• Used to **rename columns** as per the desired naming convention or schema requirements.

3. Derived Column Transformation (Hash Column)

- A new column is added using the crc32 hash function.
- The hash column helps in identifying data changes for comparison and update logic.

4. Target Setup – Azure SQL Database

- An **Transaction table** is created in the target SQL database.
- The table includes all columns from the source data, plus additional metadata columns (created_by, created_date, updated_by, updated_date).

5. Lookup Transformation

- Performs a Left Join between the source data and the target SQL table.
- Used to check if a record already exists in the target.

6. Conditional Split Transformation

- Segregates records into **Insert** and **Update** branches:
 - o **Insert**: Records not found in the target.
 - Update: Records found with differences (based on hash comparison or existing logic).

7. Derived Column Transformation (Metadata Columns)

- Insert Path:
 - Adds values to created_by, created_date, updated_by, and updated_date.
- Update Path:
 - Updates only the updated_by and updated_date fields.

8. Alter Row Transformation (for Update Path)

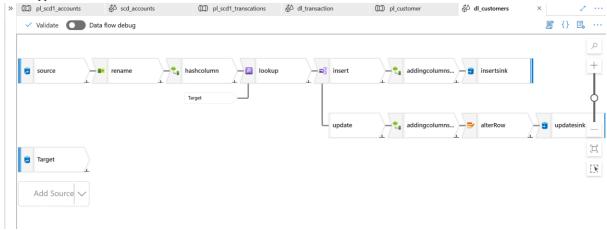
- Adds update conditions:
 - Sets rows to update if 1==1 (always true) as condition has already been logically filtered in the Conditional Split.

9. Sink Transformation

- Final data write operation:
 - o **Insert Path Sink** writes new records to the Azure SQL Accounts table.

o **Update Path Sink** updates existing records based on the Alter Row configuration.

SCD type 1 on Customers table.



1. Source Transformation

- Data is ingested from the **Silver Layer**.
- Acts as the initial source for the pipeline.

2. Select Transformation

• Used to **rename columns** as per the desired naming convention or schema requirements.

3. Derived Column Transformation (Hash Column)

- A new column is added using the crc32 hash function.
- The hash column helps in identifying data changes for comparison and update logic.

4. Target Setup - Azure SQL Database

- A **Customer table** is created in the target SQL database.
- The table includes all columns from the source data, plus additional metadata columns (created_by, created_date, updated_by, updated_date).

5. Lookup Transformation

- Performs a Left Join between the source data and the target SQL table.
- Used to check if a record already exists in the target.

6. Conditional Split Transformation

- Segregates records into **Insert** and **Update** branches:
 - Insert: Records not found in the target.
 - Update: Records found with differences (based on hash comparison or existing logic).

7. Derived Column Transformation (Metadata Columns)

Insert Path:

Adds values to created_by, created_date, updated_by, and updated_date.

Update Path:

Updates only the updated_by and updated_date fields.

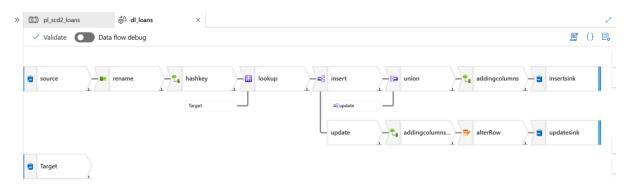
8. Alter Row Transformation (for Update Path)

- Adds update conditions:
 - Sets rows to update if 1==1 (always true) as condition has already been logically filtered in the Conditional Split.

9. Sink Transformation

- Final data write operation:
 - o Insert Path Sink writes new records to the Azure SQL Accounts table.
 - o **Update Path Sink** updates existing records based on the Alter Row configuration.

• SCD type 2 on Loans File



1. Source Transformation

- **Source Layer:** Silver Layer (Azure Data Lake or similar).
- Action: Ingest source data into the data flow pipeline.

2. Select Transformation

- **Purpose:** Rename columns from the source as needed.
- Action: Standardize column names to align with the target schema.

3. Derived Column Transformation (Hash Column)

- Purpose: Add a new column to help with change detection.
- Action:

- o Add a hash column using the crc32() hash function.
- This will act as a unique fingerprint for each record.

4. Target Configuration

- Target: Azure SQL Database.
- **Table:** accounts (contains all columns from the source).
- **Purpose:** Compare incoming records with existing ones.

5. Lookup Transformation

- Join Type: Left Join.
- Purpose: Compare source records with the target table (loans) based on business key(s).
- Output: Helps identify whether a record is new or has changed.

6. Conditional Split Transformation

- **Purpose:** Classify records into:
 - o **Insert:** New records not found in the target.
 - o **Update:** Records with changed hash value.

7. Update Path

a. Derived Column Transformation (for Update)

- Add Columns:
 - updated_by
 - updated_date
 - o isActive (set to **0**, marking the old record as inactive).

b. Alter Row Transformation

• Action: Use UpsertIf(1==1) to handle updates.

c. Sink Transformation

- Target Table: loans (Azure SQL Database).
- Action: Load the updated records.

8. Insert Path

a. Union Transformation

- Action: Combine outputs of:
 - Split@Insert
 - Split@Update
 (We do this because updated records will be inserted as new rows with updated values.)

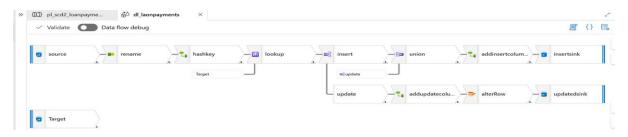
b. Derived Column Transformation (for Insert)

Add Columns:

- created_by
- o created_date
- updated_by
- updated_date
- o isActive (set to 1 for active records)

c. Sink Transformation

- Target Table: loans (Azure SQL Database).
- Action: Load all new/inserted records.
- SCD Type for loans_payment file.



1. Source Transformation

- Source Layer: Silver Layer (Azure Data Lake or similar).
- Action: Ingest source data into the data flow pipeline.

2. Select Transformation

- **Purpose:** Rename columns from the source as needed.
- Action: Standardize column names to align with the target schema.

3. Derived Column Transformation (Hash Column)

• Purpose: Add a new column to help with change detection.

Action:

- o Add a hash column using the crc32() hash function.
- o This will act as a unique fingerprint for each record.

4. Target Configuration

- Target: Azure SQL Database.
- Table: accounts (contains all columns from the source).
- Purpose: Compare incoming records with existing ones.

5. Lookup Transformation

- Join Type: Left Join.
- **Purpose:** Compare source records with the target table (loans_payment) based on business key(s).
- Output: Helps identify whether a record is new or has changed.

6. Conditional Split Transformation

- **Purpose:** Classify records into:
 - o **Insert:** New records not found in the target.
 - o **Update:** Records with changed hash value.

7. Update Path

a. Derived Column Transformation (for Update)

- Add Columns:
 - o updated_by
 - updated_date
 - o isActive (set to **0**, marking the old record as inactive).

b. Alter Row Transformation

• Action: Use UpsertIf(1==1) to handle updates.

c. Sink Transformation

- Target Table: loans_payment (Azure SQL Database).
- Action: Load the updated records.

8. Insert Path

a. Union Transformation

- Action: Combine outputs of:
 - Split@Insert
 - Split@Update
 (We do this because updated records will be inserted as new rows with updated values.)

b. Derived Column Transformation (for Insert)

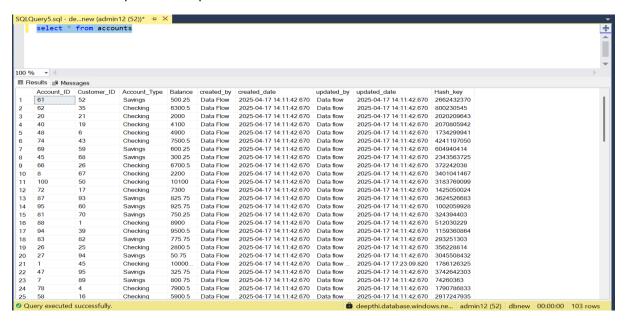
- Add Columns:
 - created_by
 - created_date
 - updated_by
 - o updated_date
 - isActive (set to 1 for active records)

c. Sink Transformation

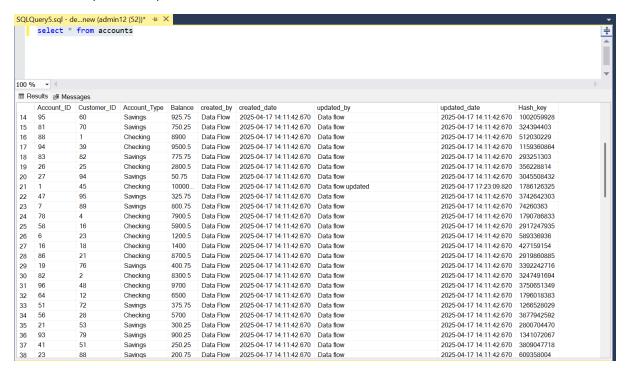
- Target Table: loans (Azure SQL Database).
- Action: Load all new/inserted records.

Tables output:

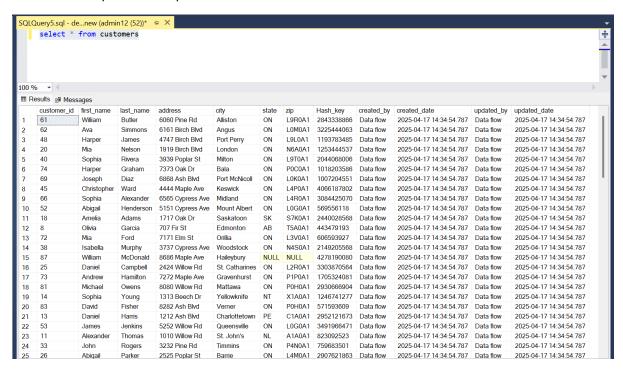
Accounts table output before update



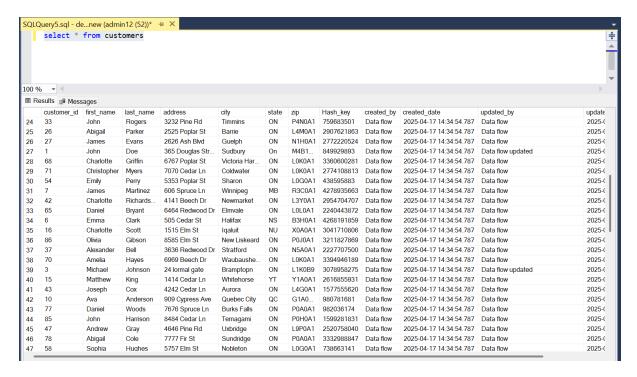
Accounts output after update



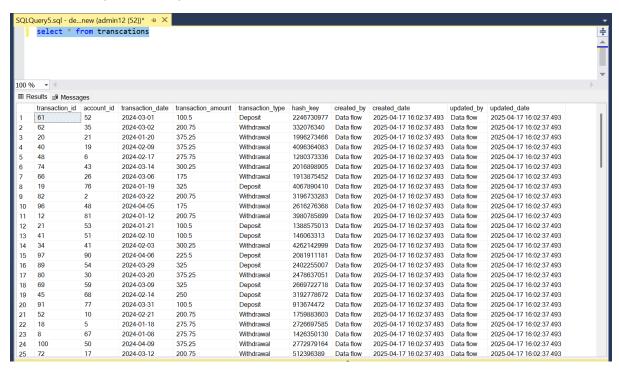
Customers Output before update



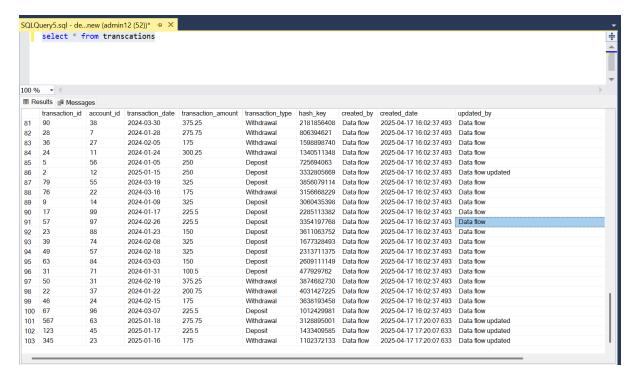
After update



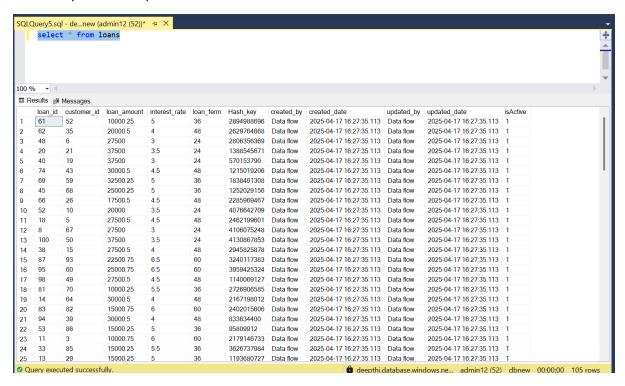
Transaction output before update



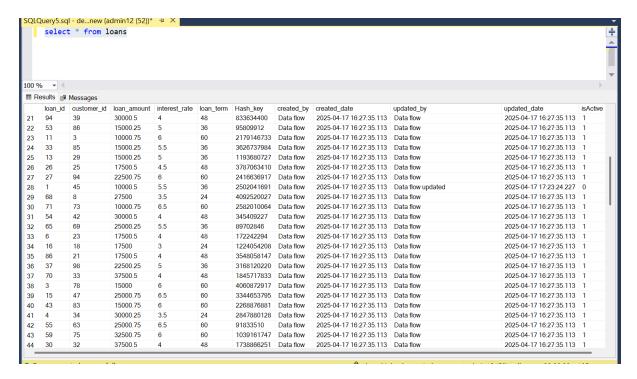
After update



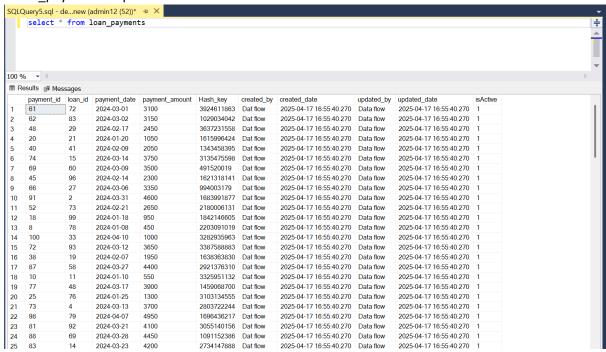
Loans output before update



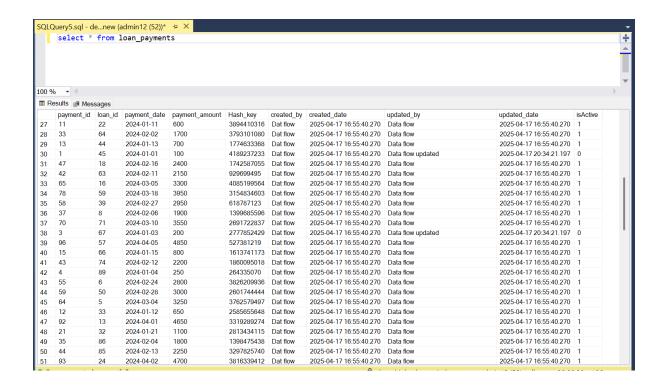
After update



Loans_payment output



After update



Created a trigger to run pipeline at 8:30 every day

