Azure Data Engineer: Batch 6

(October 2023)

**Submitted by: Roopmathi Gunna**

# Assignment 8 b: Advanced Tough questions

## Advanced Questions

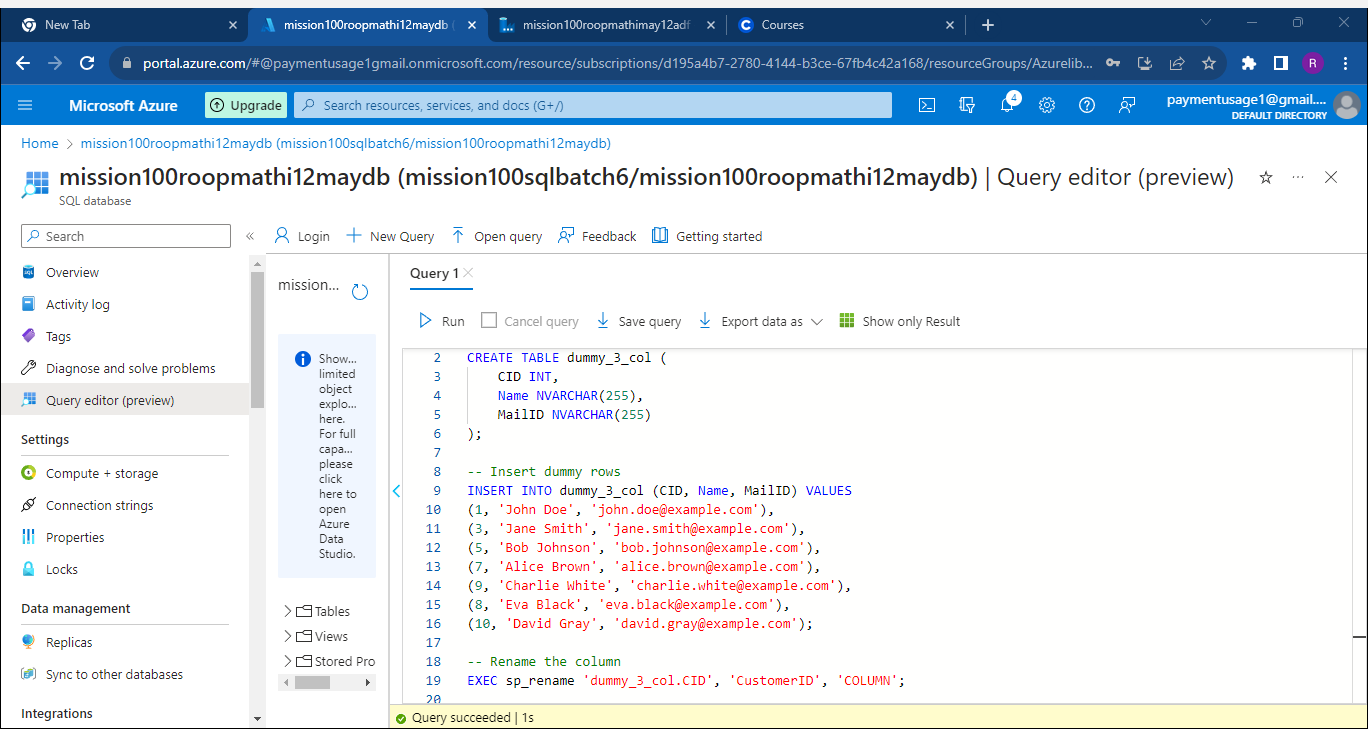
## 1. Create a pipeline to copy the customer data from the CSV file to the SQL table with only 3 columns. While copying the data, ensure that you copy all the rows where CustomerID is an odd number. Also, check if CustomerID exists in the table then update it.

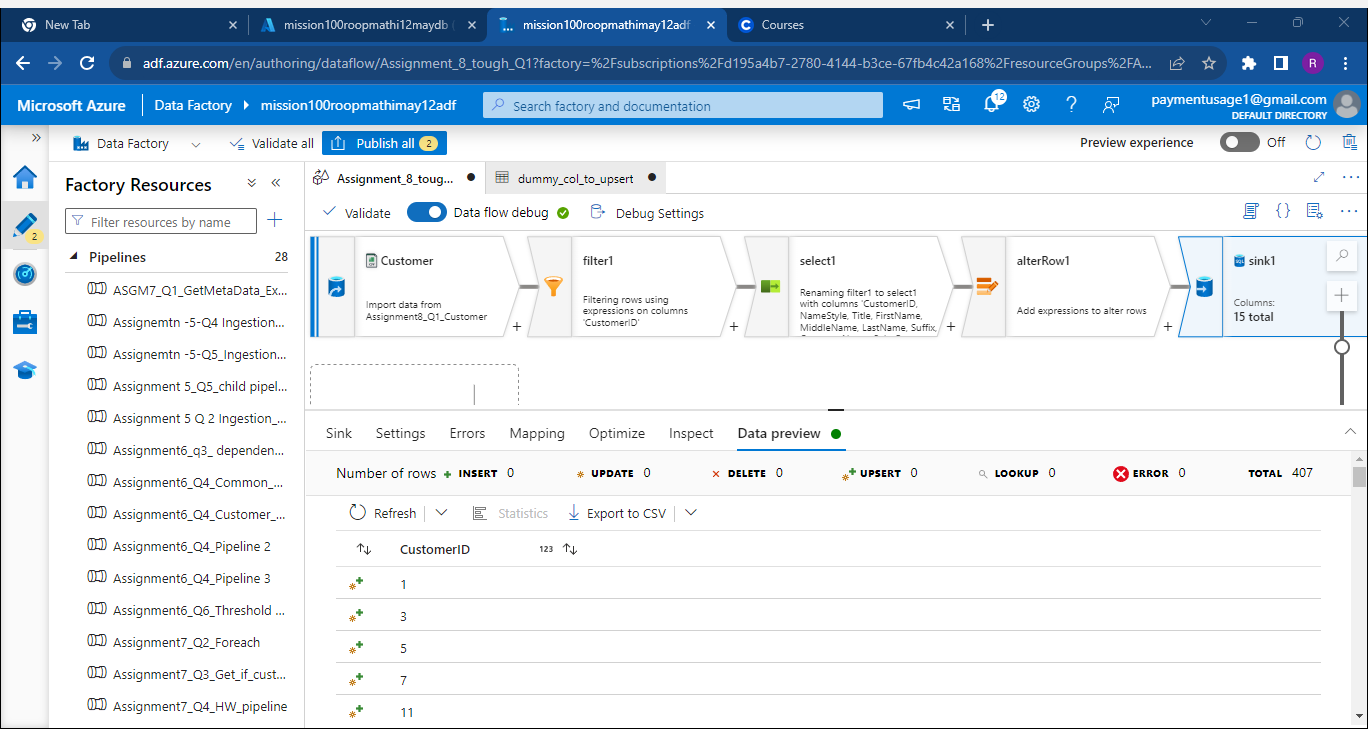
## The file attached with the question.

## Table: Create dummy table with column (CID, Name, mailid)

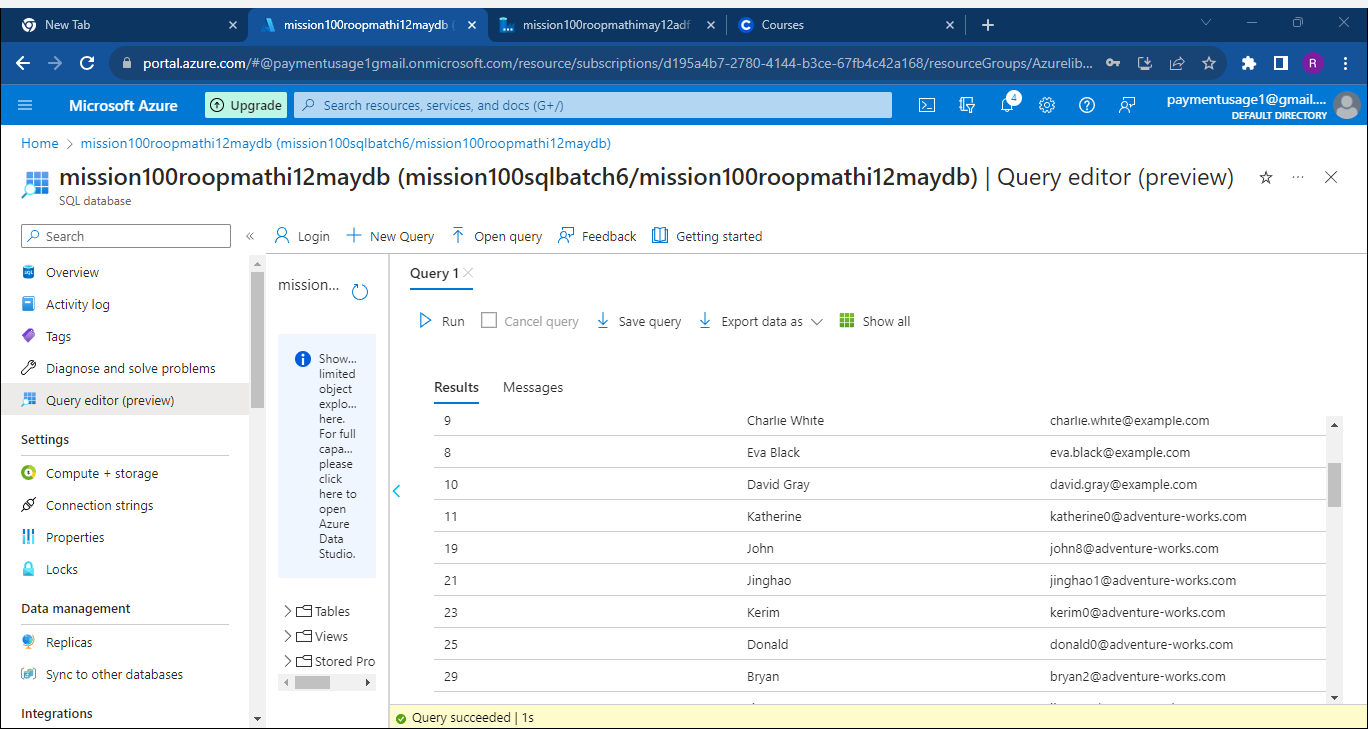
## Insert dummy row with id 1,3,5,7,9,8,10

First create a dummy table with three columns like this:



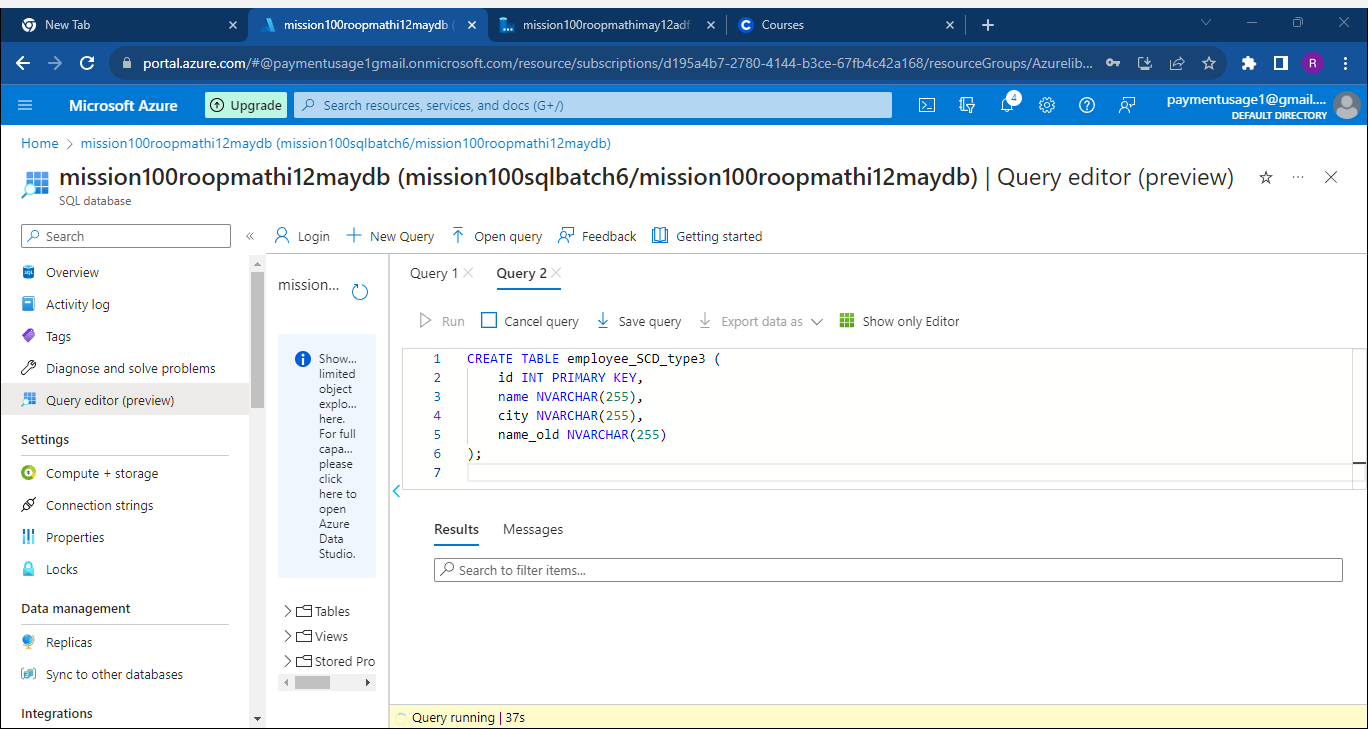


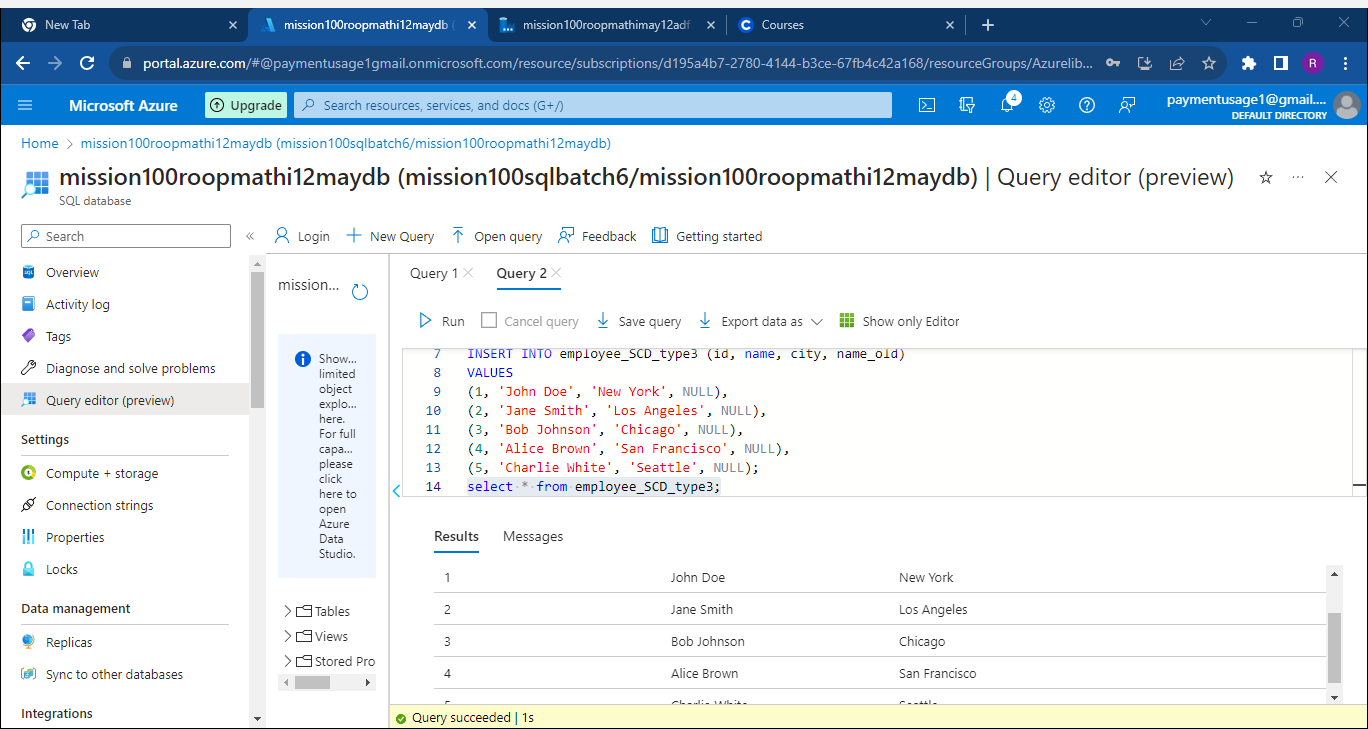
Here you can see that rows with Customerid 11 onwards have been upserted into the dummy\_col\_table we created earlier.



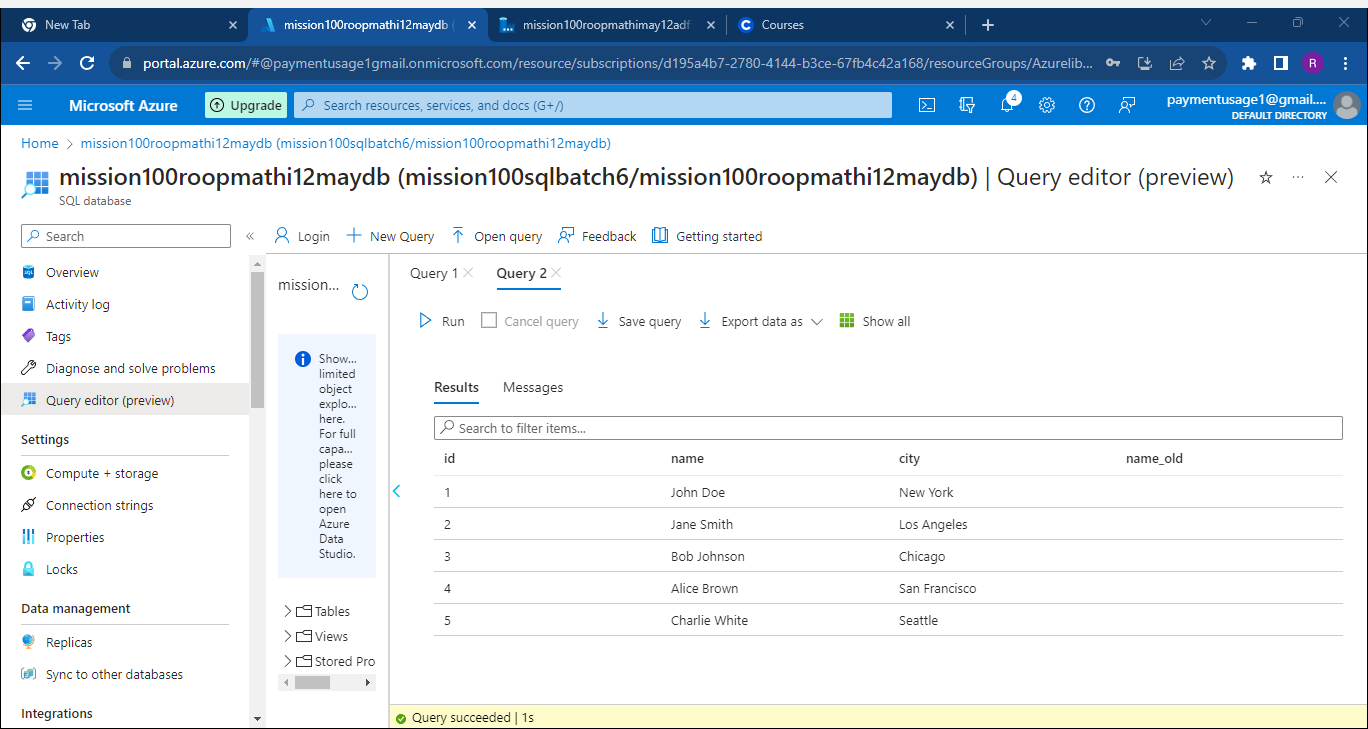
## 2. Create a pipeline to implement the SCD Type 3 example using the Dataflow.

Let’s create an employee\_SCD table first:

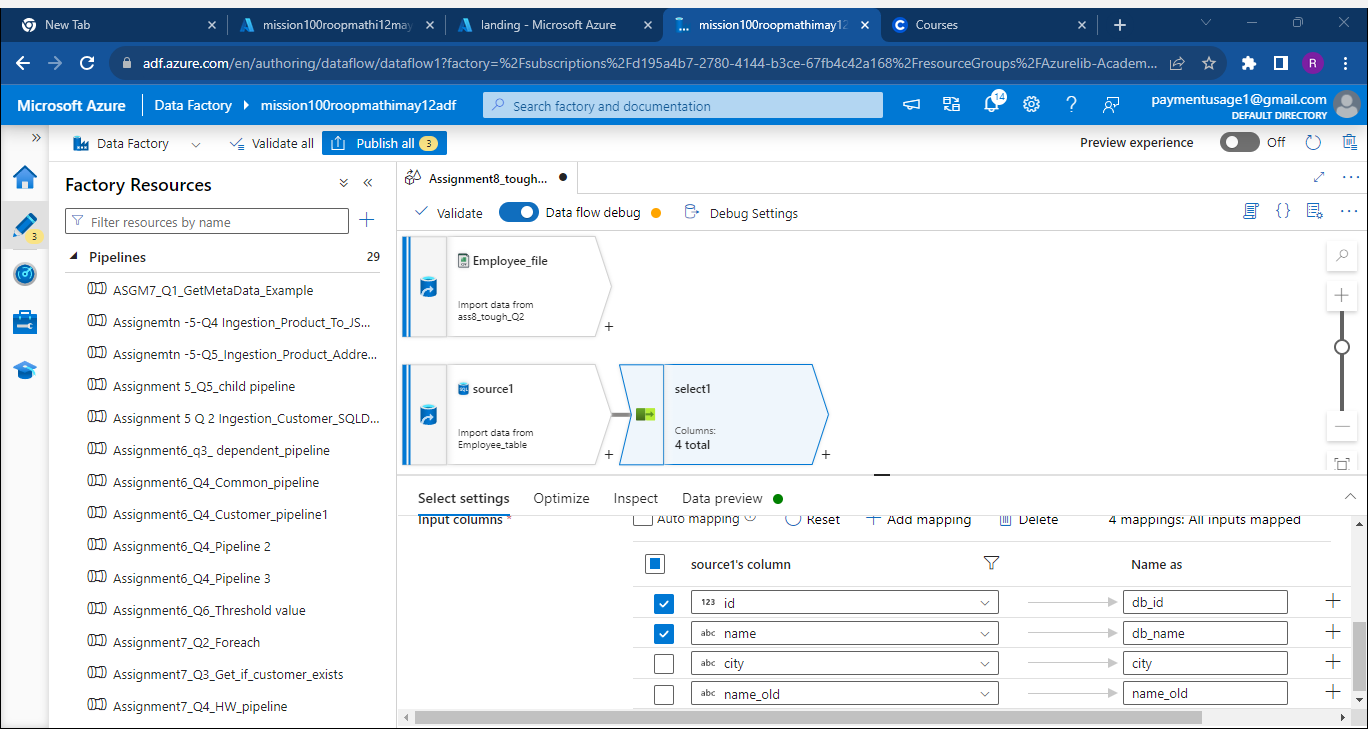




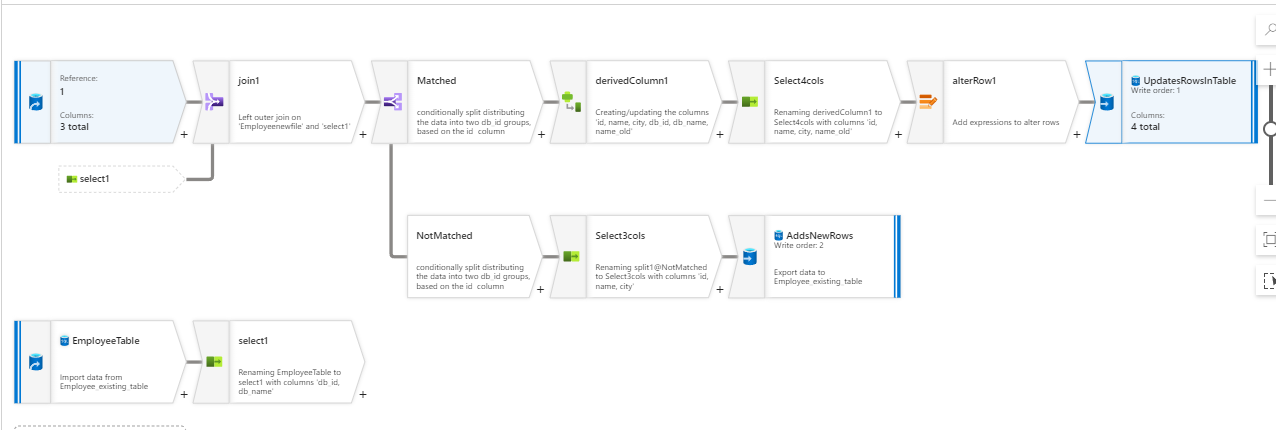
Before the Dataflow transformation



Now let's create the Data Flow : Source 1 is Employee.txt file on our ADLS folder and Source 2 is the Employee table in Azure SQL DB , lets do the transformation.



## This data flow validates records in Source 1 (Employee.Txt) against our employee table. If a record exists, the city name in the Employee table is updated with the new information from Employee.Txt. If a record doesn't exist, new rows are added to the Employee table.



Now let’s execute this dataflow in a Data pipeline, we see it executed successfully.

A screenshot of a computer

Description automatically generated

## Here is the SCD\_Type3 Employee Table after this dataflow was executed.

## A screenshot of a computer Description automatically generated

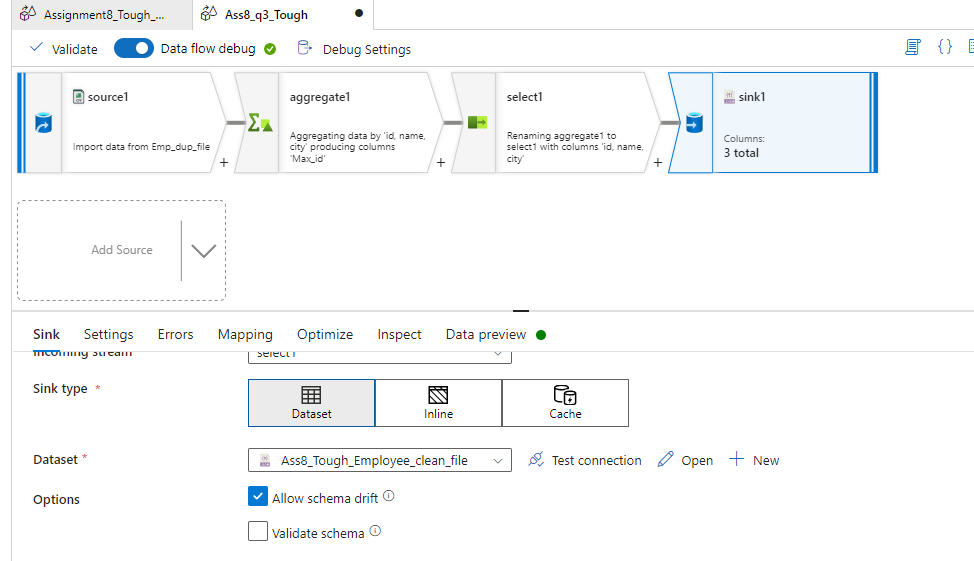
## A screenshot of a computer Description automatically generated

Jane Smith, Alice Brown and Charlie White have been updated to Jane Johnson, Alice black and Charlie Brown respectively.

Three new rows with IDs 6, 7, and 8 have been added.

## 3. Create a pipeline to remove the duplicate record in the CSV file and save it as JSON.

Let's first create the dataflow to execute this logic:



This is what data in source1 looks like:

A screenshot of a computer

Description automatically generated

This is what the data moving into the Sink looks like after the set of transformations in our data flow:

A screenshot of a computer

Description automatically generated

Let’s finally build and execute a Data Pipeline that will execute this data Flow:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated