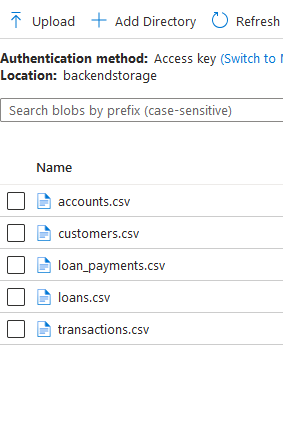
**Data Pipeline Documentation: Customer Account Analysis**

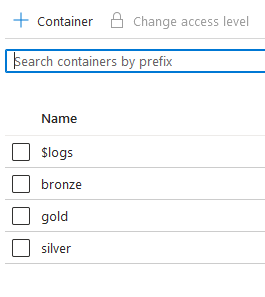
This document outlines a data pipeline for processing customer account data, including copying data from a backend team's storage account, performing transformations in Databricks, and exposing data for analysis in Azure Synapse Analytics.

**Step 1: Data Ingestion (Backend Storage to Raw (Bronze Container)**

Data is available in the backend storage account.

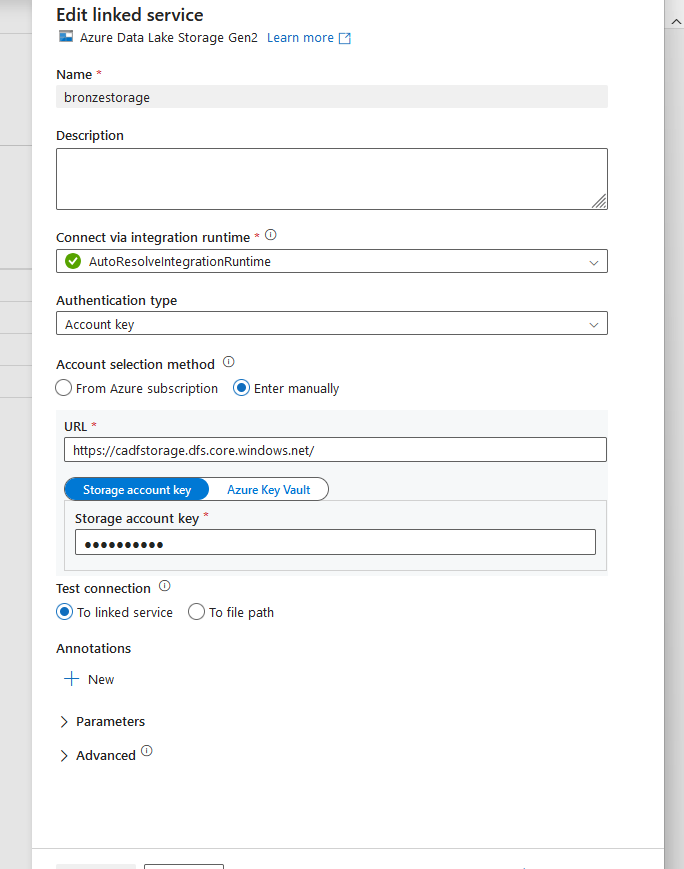


Create another storage account with bronze, silver and gold containers.



Create Data Factory resource > Launch workspace > Manage > Linked Service

* Create a linked service for backend storage container and another for bronze container.



* Then go to Author > Pipelines > New Pipeline > Move and Transform > Copy Data
  + Under Copy data – In source > Create new data set for data lake storage and choose the backend storage container for the path. Choose Wildcard file path to move multiple files at a time.
  + In sink > Create new data set for data lake storage and choose bronze container in new storage account for the path.
  + Then publish all and debug.
  + The files are transformed to the bronze container.



**Step 2: Databricks Activity (Incremental/Delta Processing)**

* Create Databricks Workspace> Create a cluster with access mode – no isolation shared.
* Go to Wrokspace> create new folder – Project pipeline > Create new notebook.
* Read the data and clean the accounts file by matching the customer id with the customer\_id column in customers table.
* Code snippet –

**# Select only the customer\_id column from the customer DataFrame**

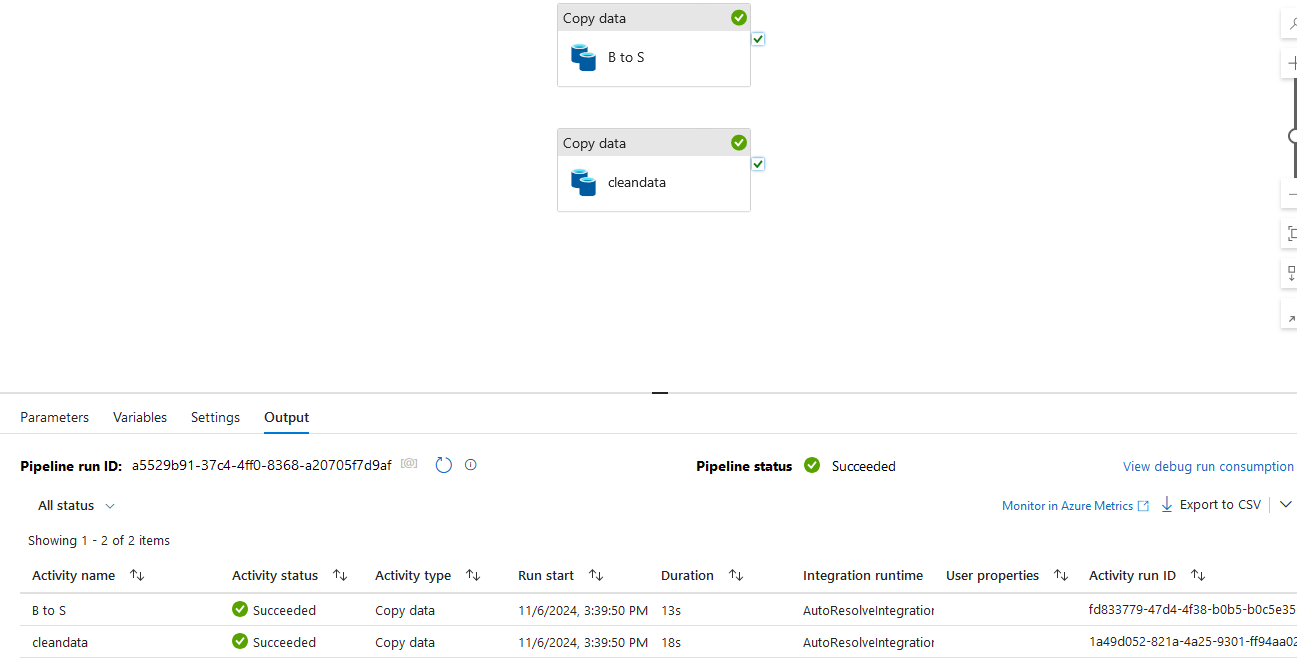
valid\_customers = customers\_df.select("customer\_id").distinct()

**# Join accounts with valid\_customers on customer\_id**

cleaned\_accounts\_df = accounts\_df.join(valid\_customers, on="customer\_id", how="inner")

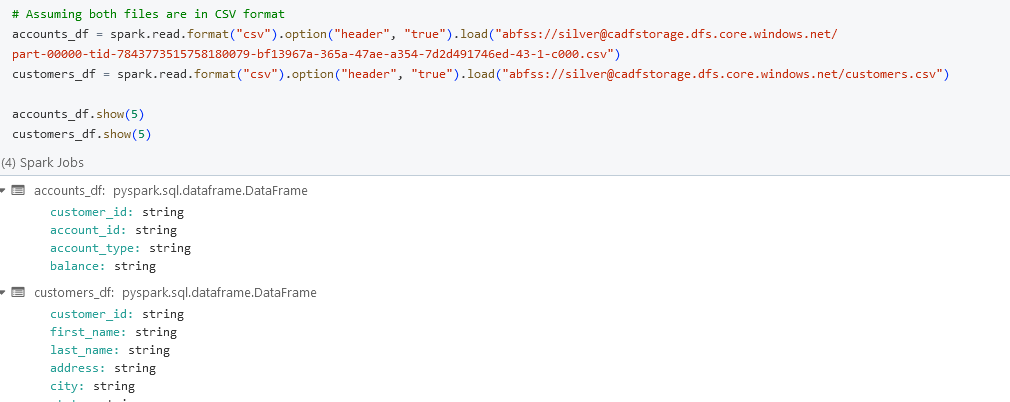
cleaned\_accounts\_df.show(20)

* Then finally write this file in the bronze container.
* Now using the data factory pipeline, move the data in bronze container to silver container.
* Create two data sets – one for the cleaned data file and one for rest of the files.
* Create a pipeline and execute them.



**Step 3: Databricks Activity (ETL Processing)**

* Read the accounts and customers tables from silver container.



* Then transform the data by aggregating the balance column



* Finally write this file to gold container.



**Step 4: Azure Synapse Analytics**

* Create synapse analytics and open the studio.
* Go to > Manage > Linked Service > New > Create linked service for silver container and gold container.
* Go to Data > linked> select the respective container, the files in the container appear on the right window.
* Right click on the respective file > sql > create external table - a window emerges – create a database and click create script.



Customers ext table



Tot balance per customer



transactions