Assignment

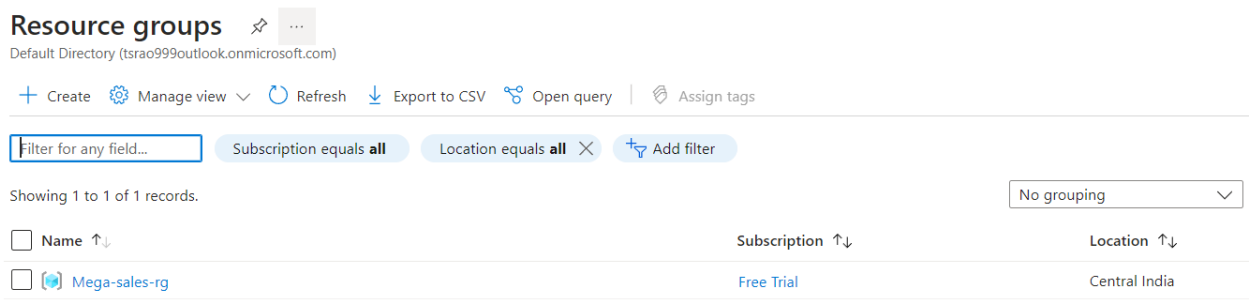
Utilize Azure Data Factory (ADF) to ingest Orders and Customers data, and execute fundamental transformations on the datasets.

Task -1: Ingest orders.csv file from external URL to ADLS Gen2

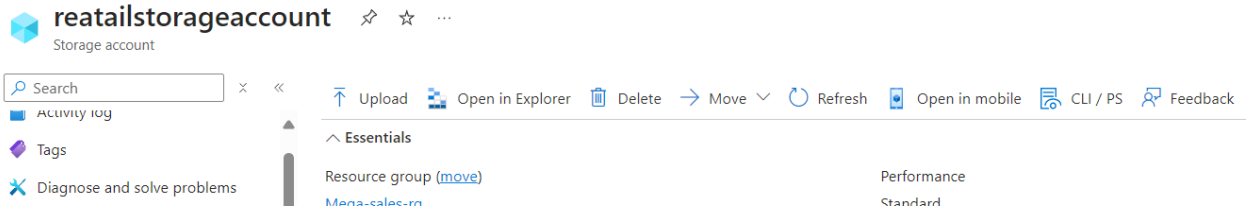
Dataset orders.csv can be downloaded from below link
https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv

Implementation:

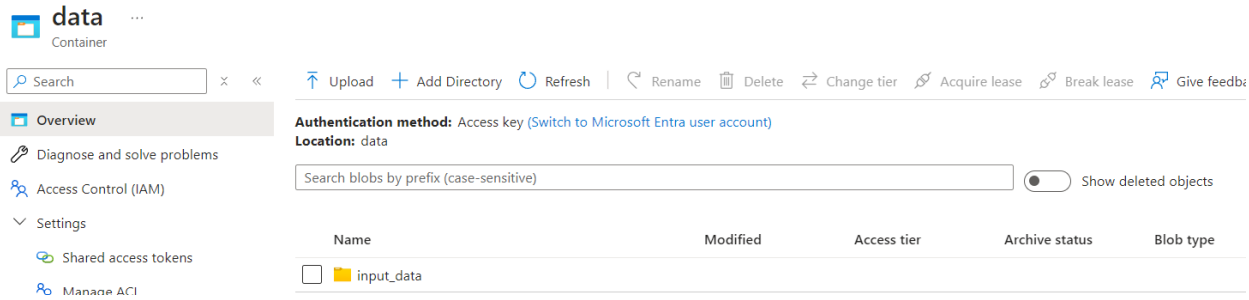
1. I have created resource group.



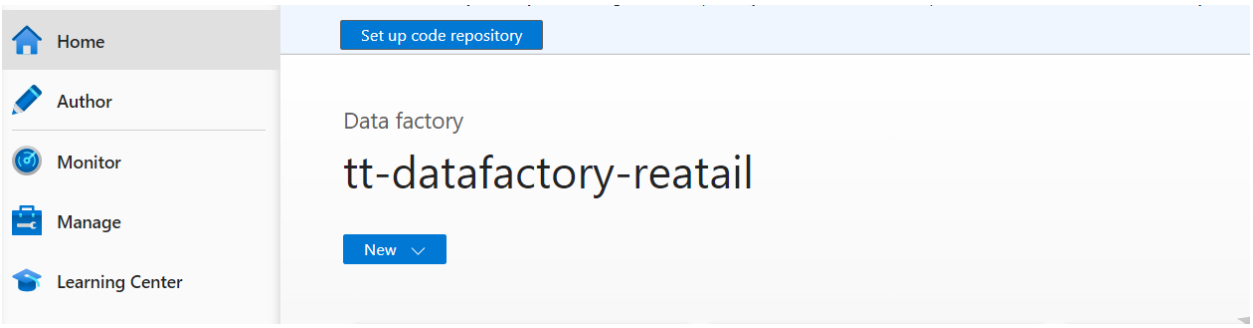
2. I have create a Storage Account (Enable Hierarchical namespace to make it as data lake storage and not just the blob storage)



3. I have created container with directory inside storage account.



4. **I have created a Resource - Azure Data Factory:**



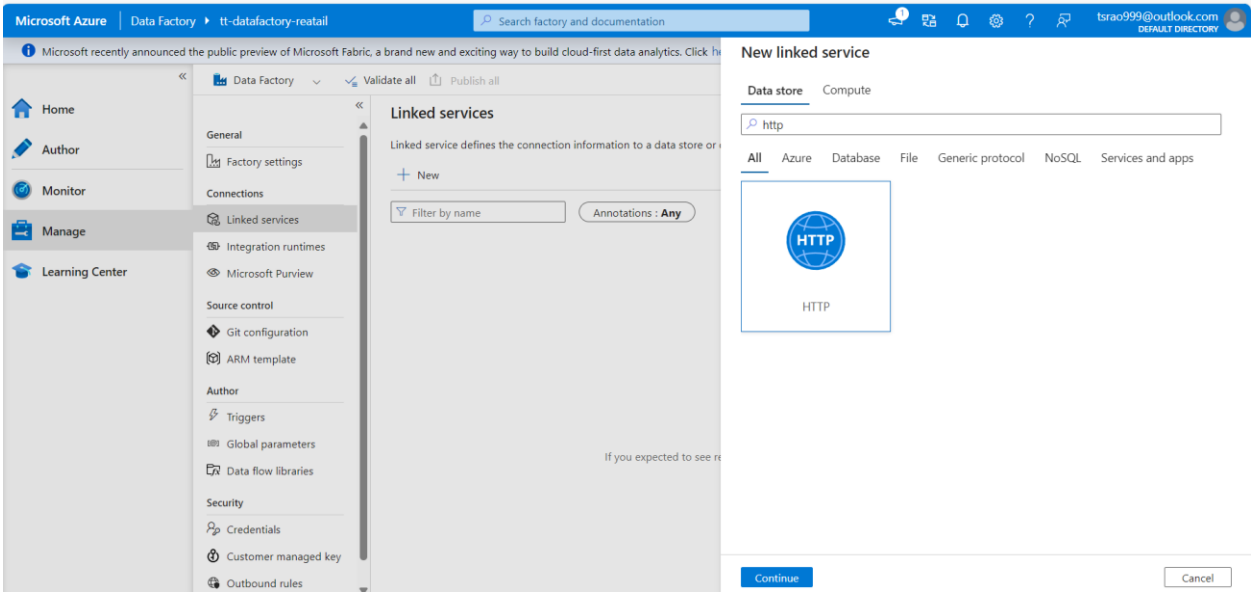
5. **Create a Linked Service for Source (choose HTTP connector):** orders.csv can be downloaded from below link.

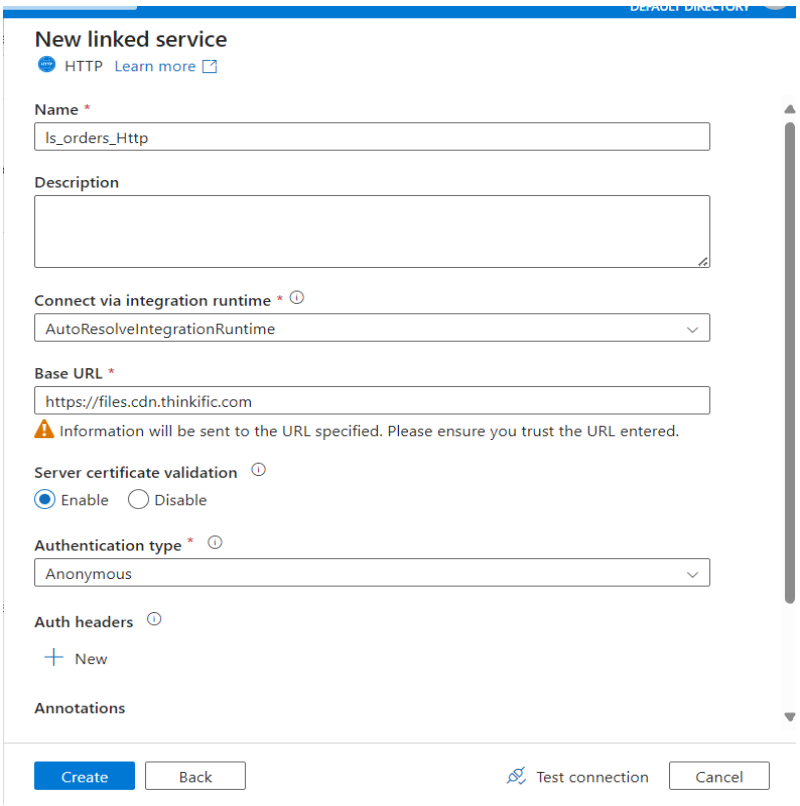
"https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv "

a) **Base URL** - https://files.cdn.thinkific.com
b) **Relative URL** - file_uploads/349536/attachments/c28/5fb/25b/orders.csv

I have launched the **Azure Data Factory** to create the Linked service and datasets and pipeline

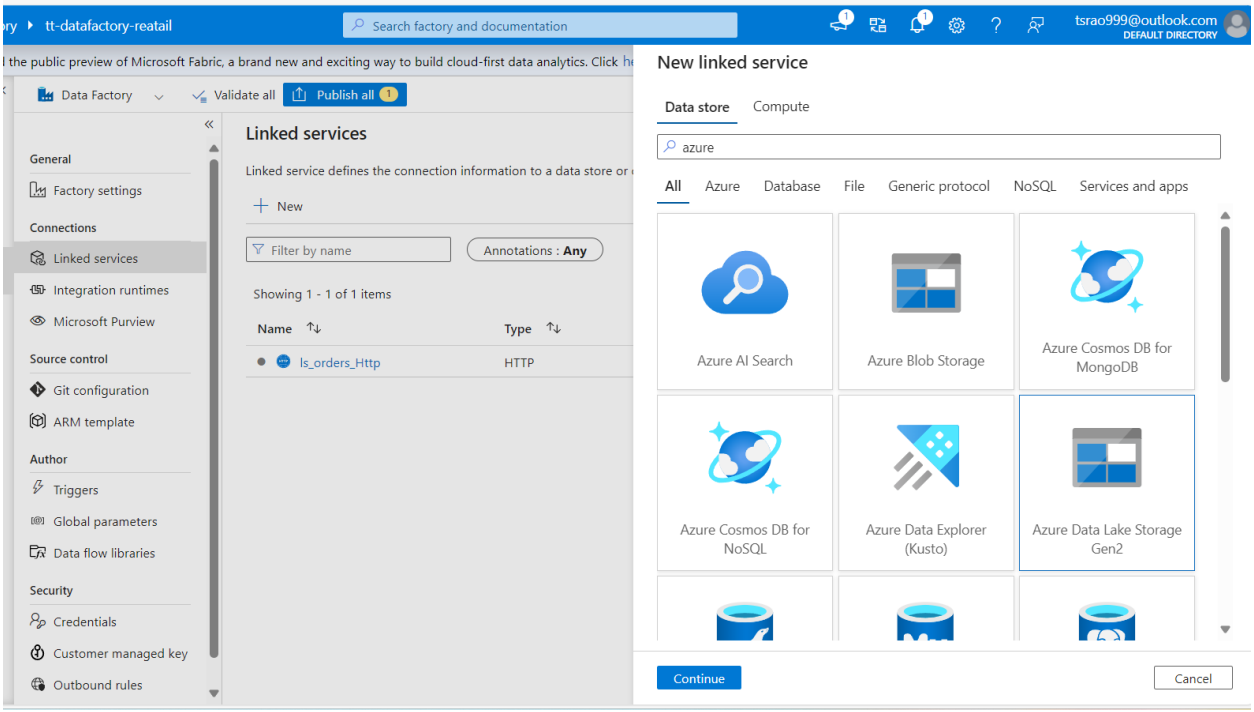
Check below screenshot for creating linked service for our source i.e. Azure Data lake Gen2 Go to **Monitor => Linked Service => New => search Http server and select Azure data Lake Gen2 => Continue**

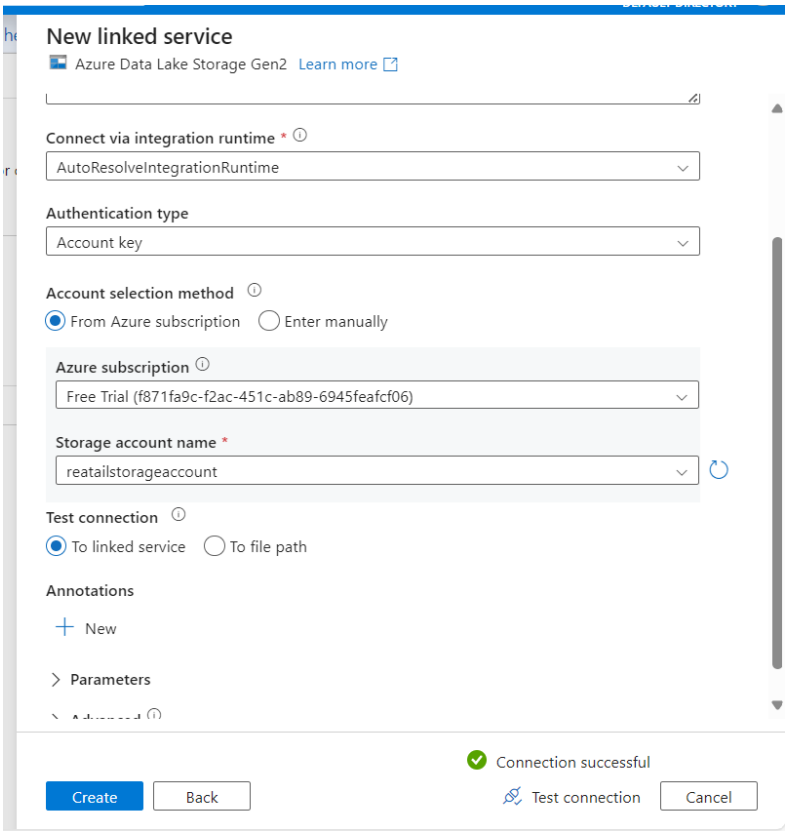




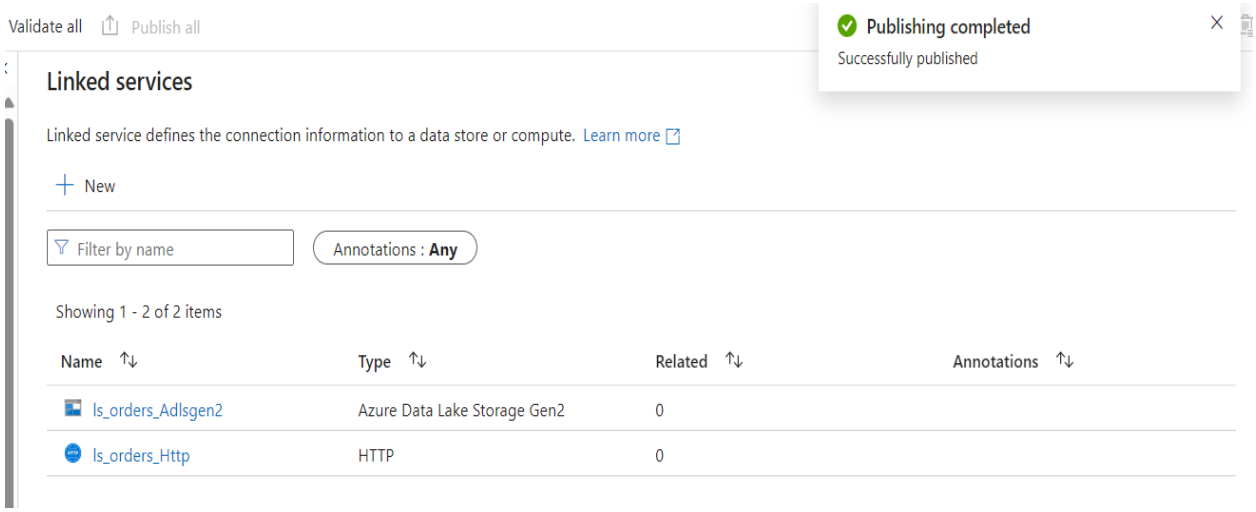
I have create a Linked Service for Sink (choose Data Lake Gen 2 connector):

- Check below screenshot for creating linked service for our sink to ADLS gen storage



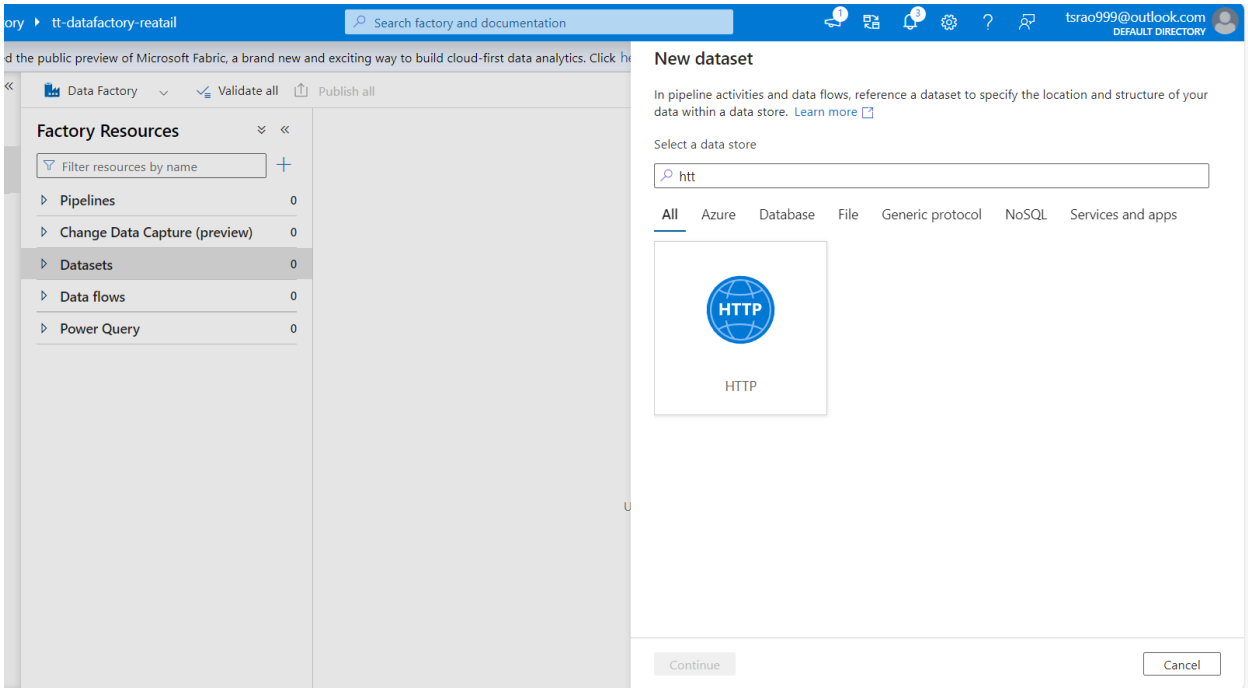


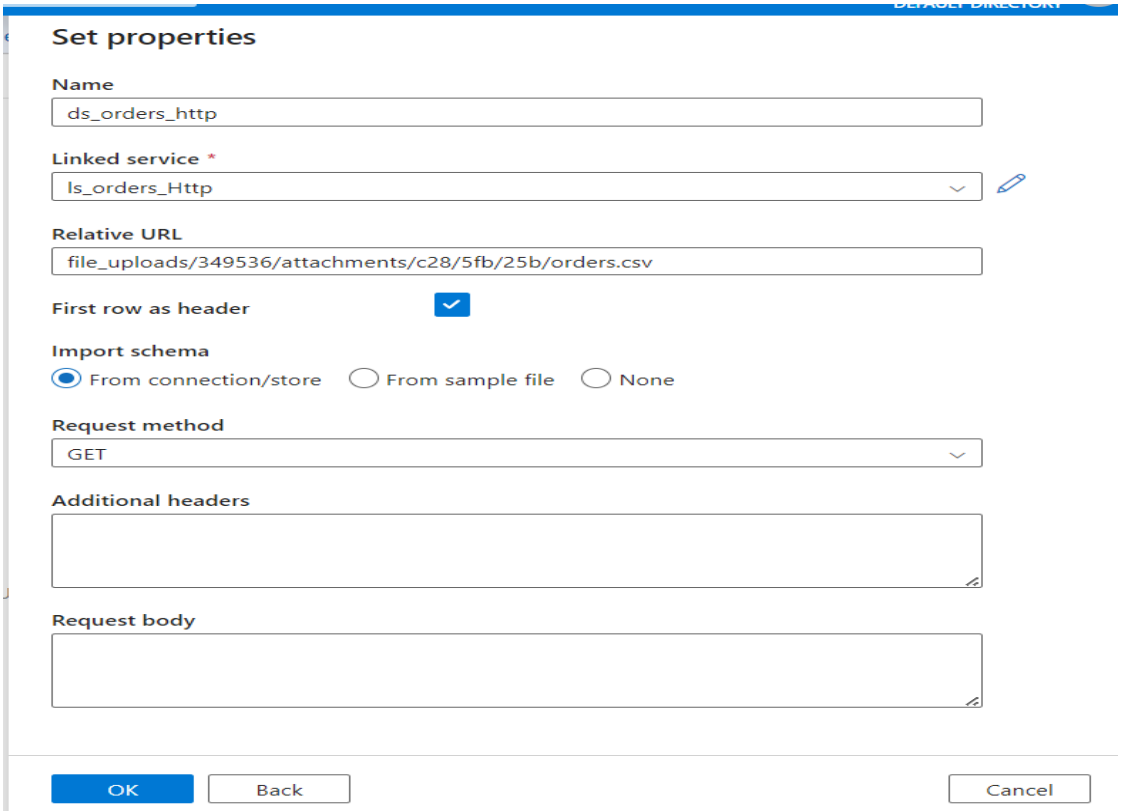
After creating new linked services in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.



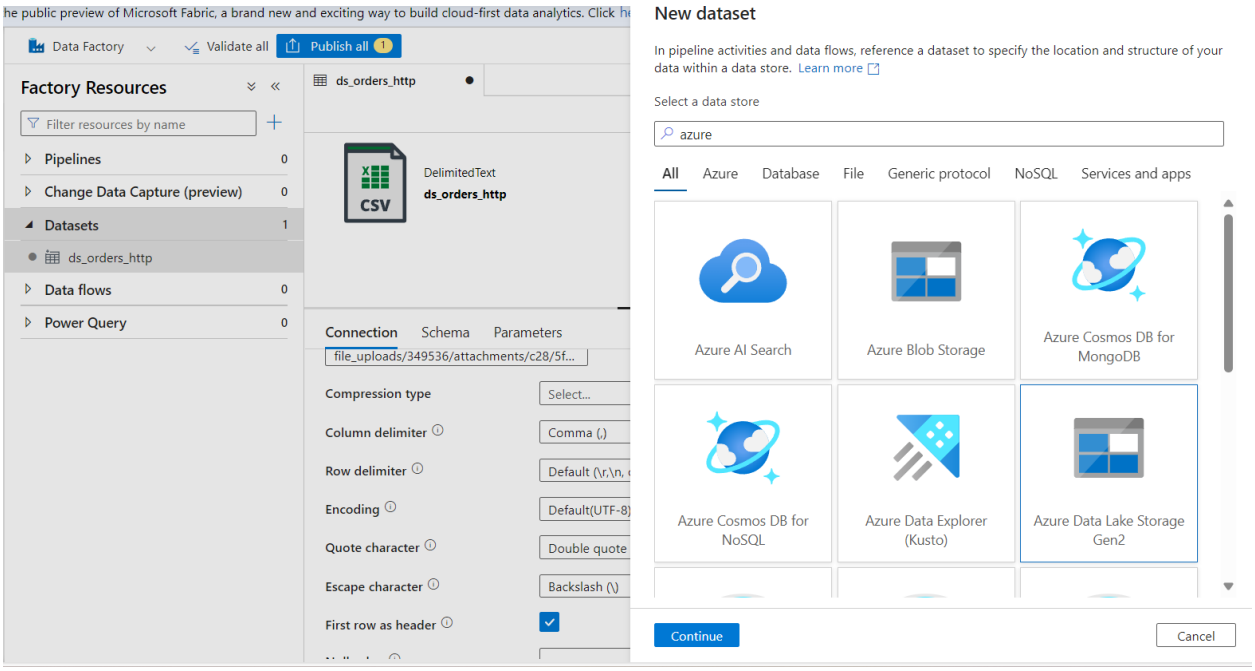
I have created **dataset for Source** (choose CSV format and also provide the relative URL as it is for the HTTP Linked Service)

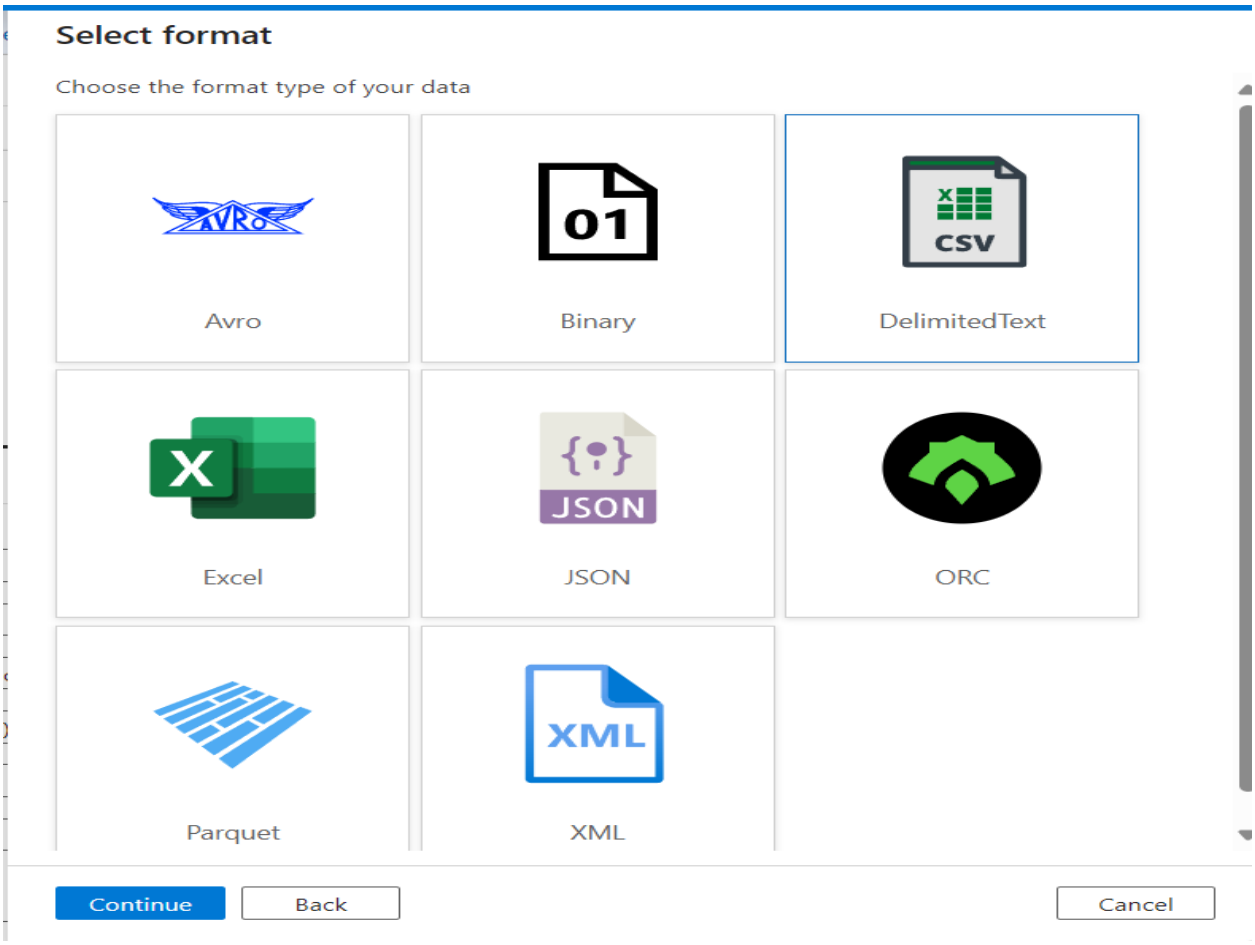
- To create dataset click on Author => Datasets => New datasets

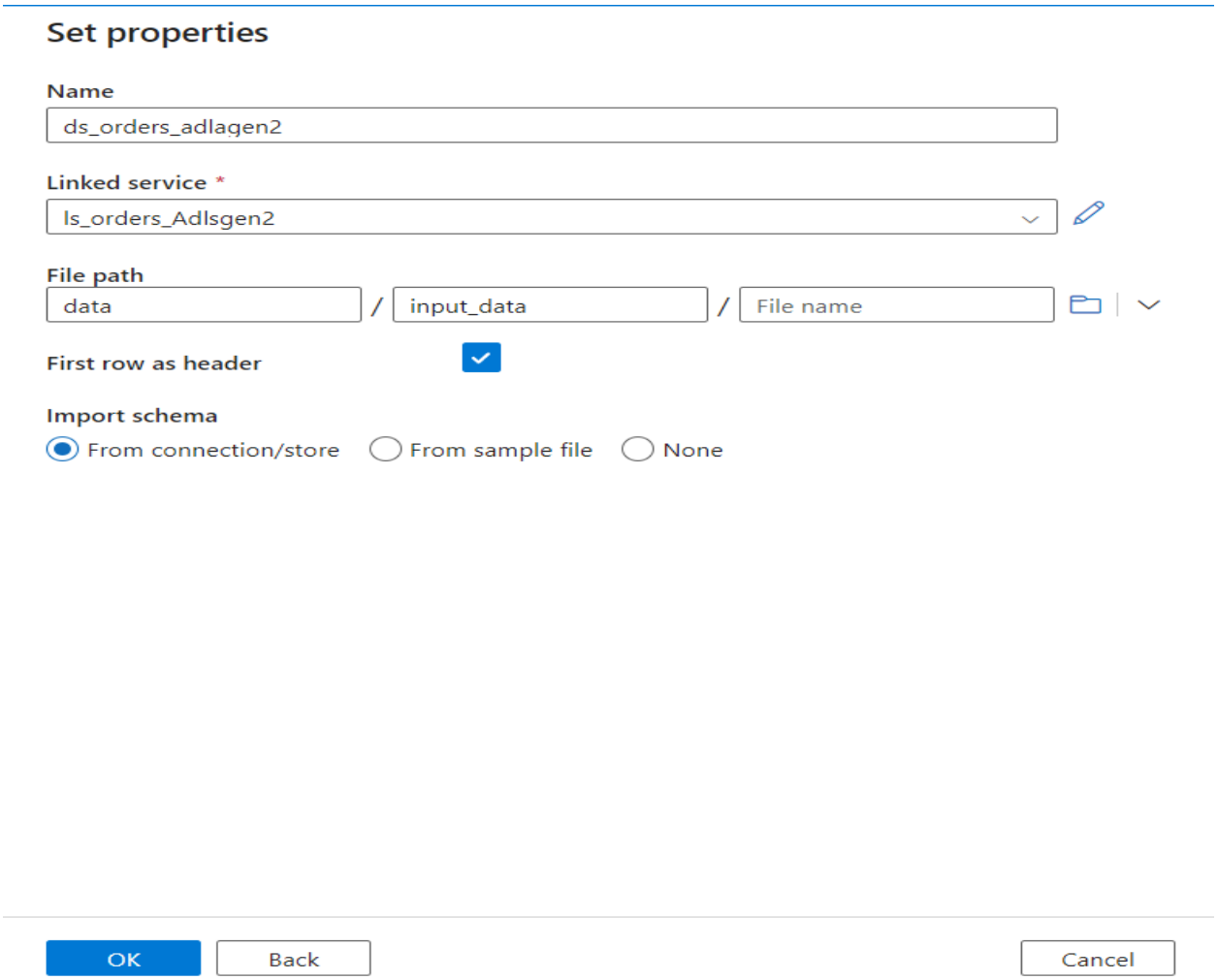




I have created dataset for Sink:



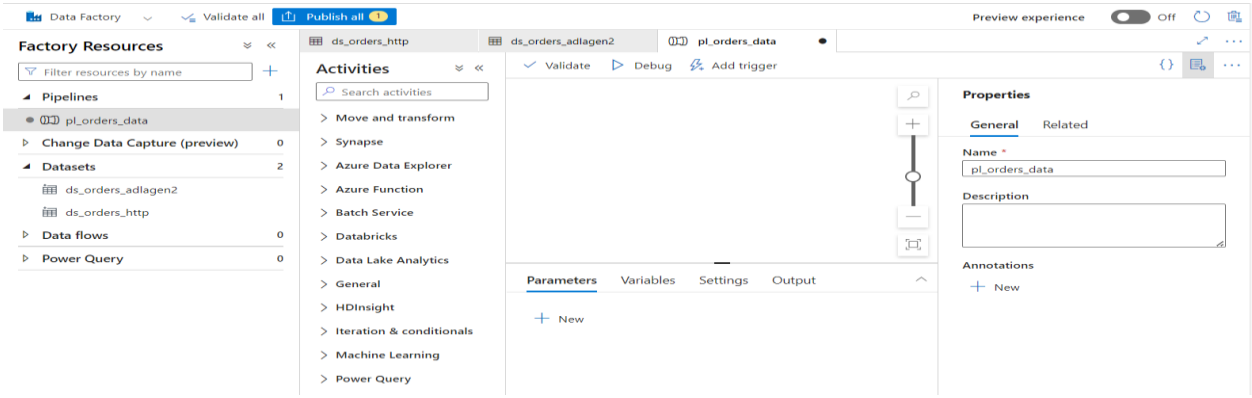


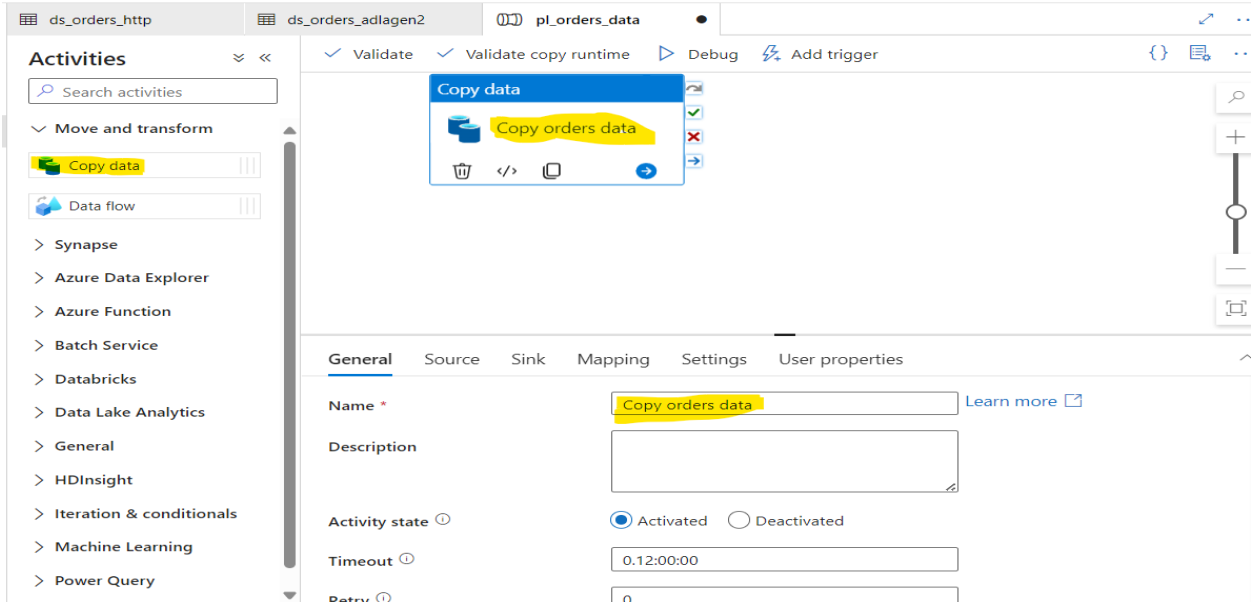


After creating new datasets in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

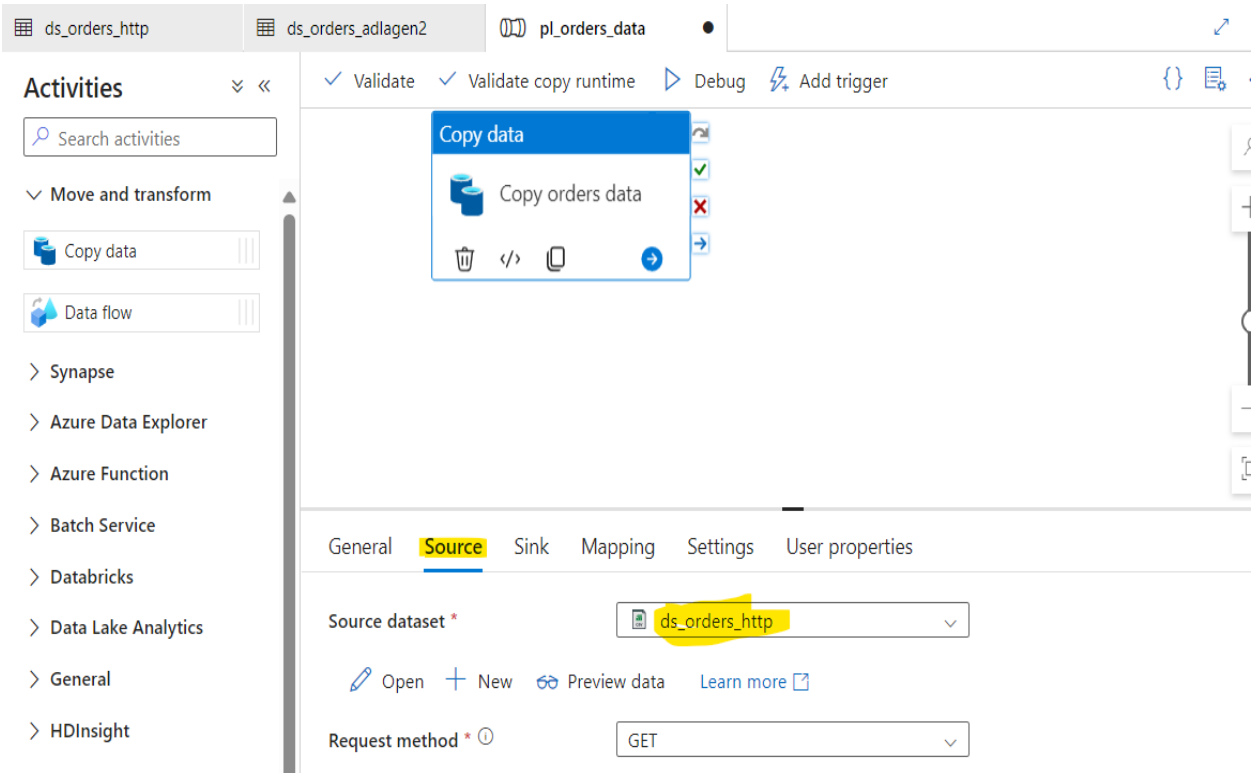
I have created a data Pipeline and Create a Copy activity within the Pipeline:

Now click on “Move and transform” and drag copy activity in the pipeline as shown below. Rename the pipeline copy data activity as shown in screenshot.

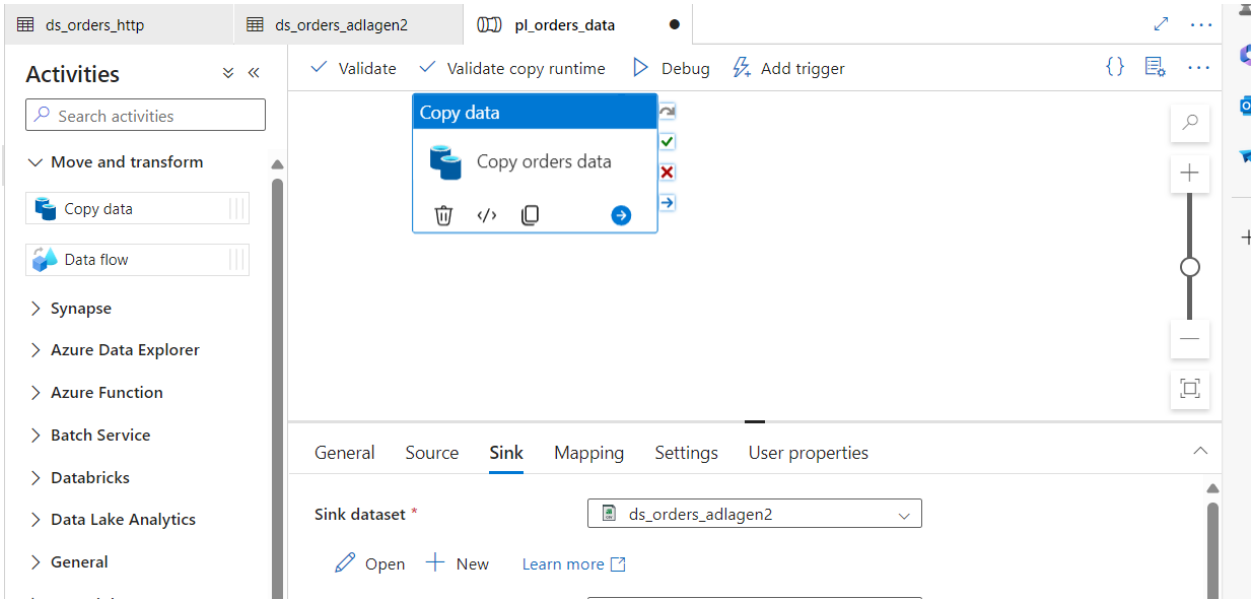




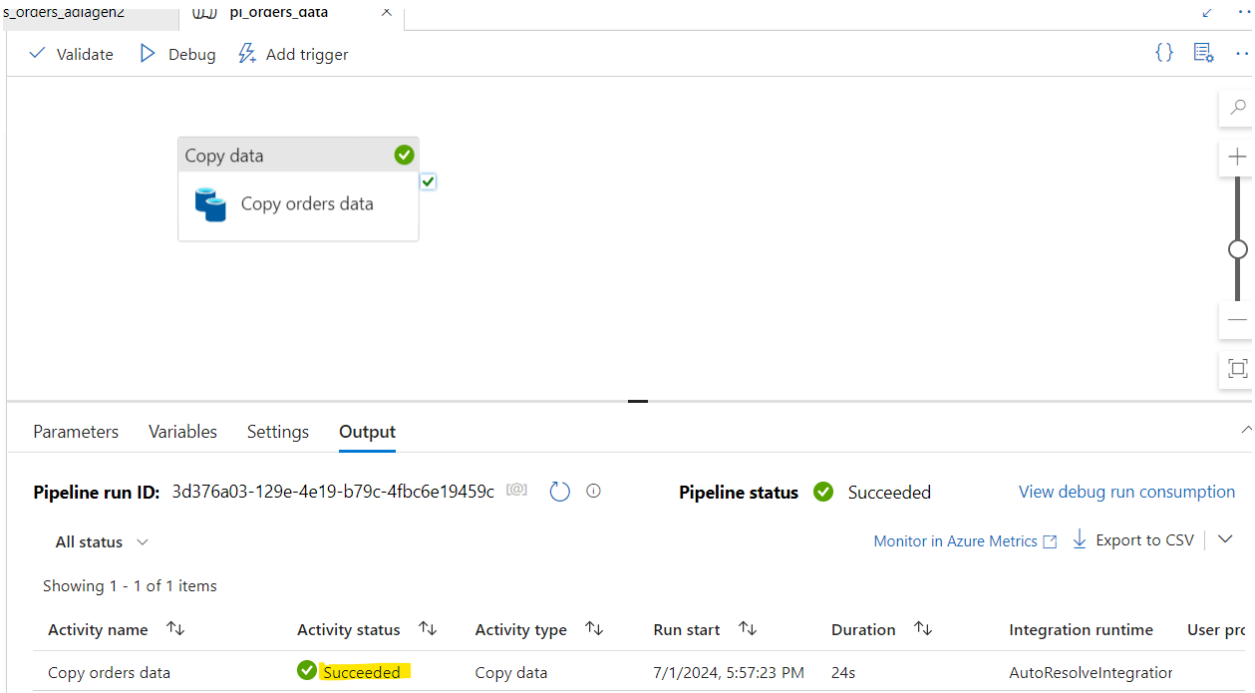
Now in source select dataset here “ds_orders_http”



Now in sink select dataset for sink here “ds_orders_adlsgen2”

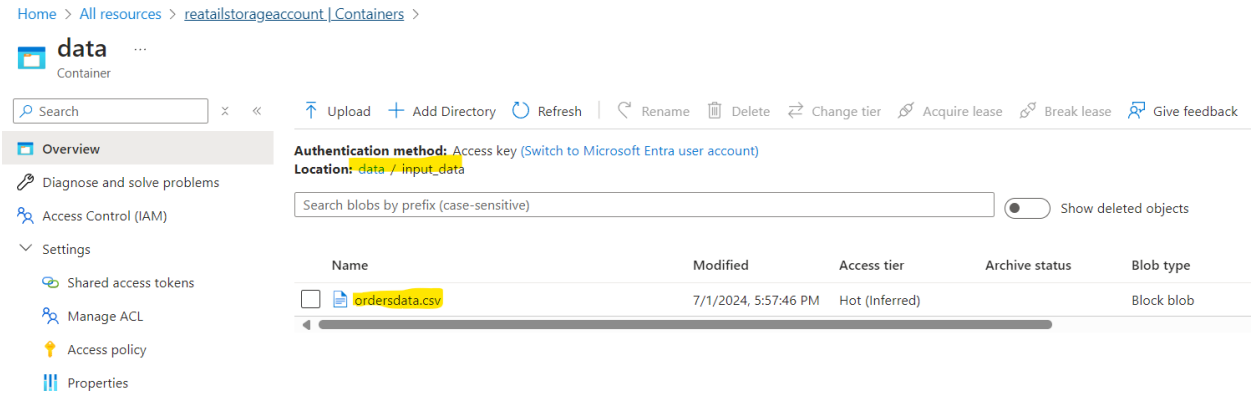


Debug and validate the pipeline, and upon successful validation, proceed to publish the pipeline:



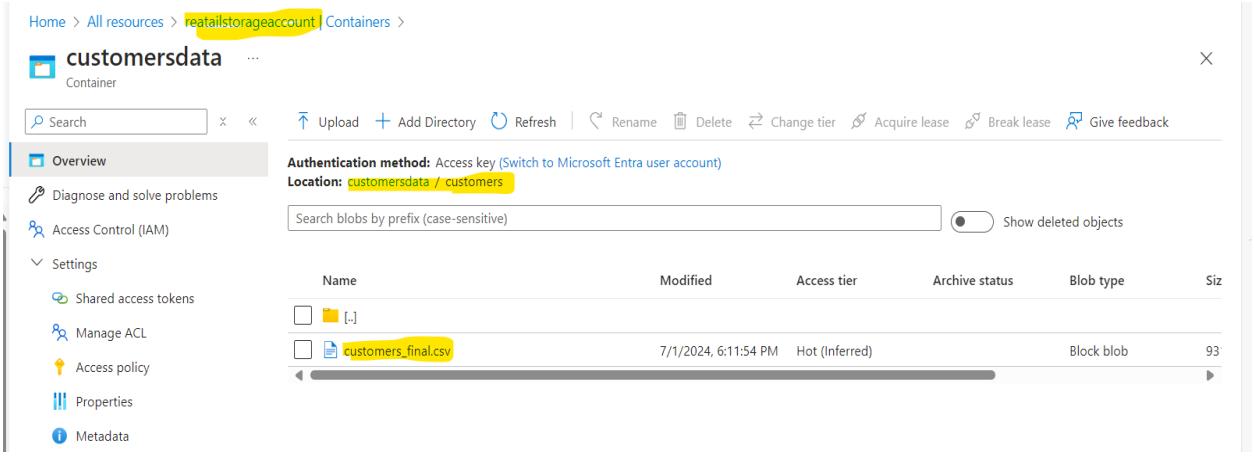
And if this is successful then publish the pipeline.

After creating new pipelines in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

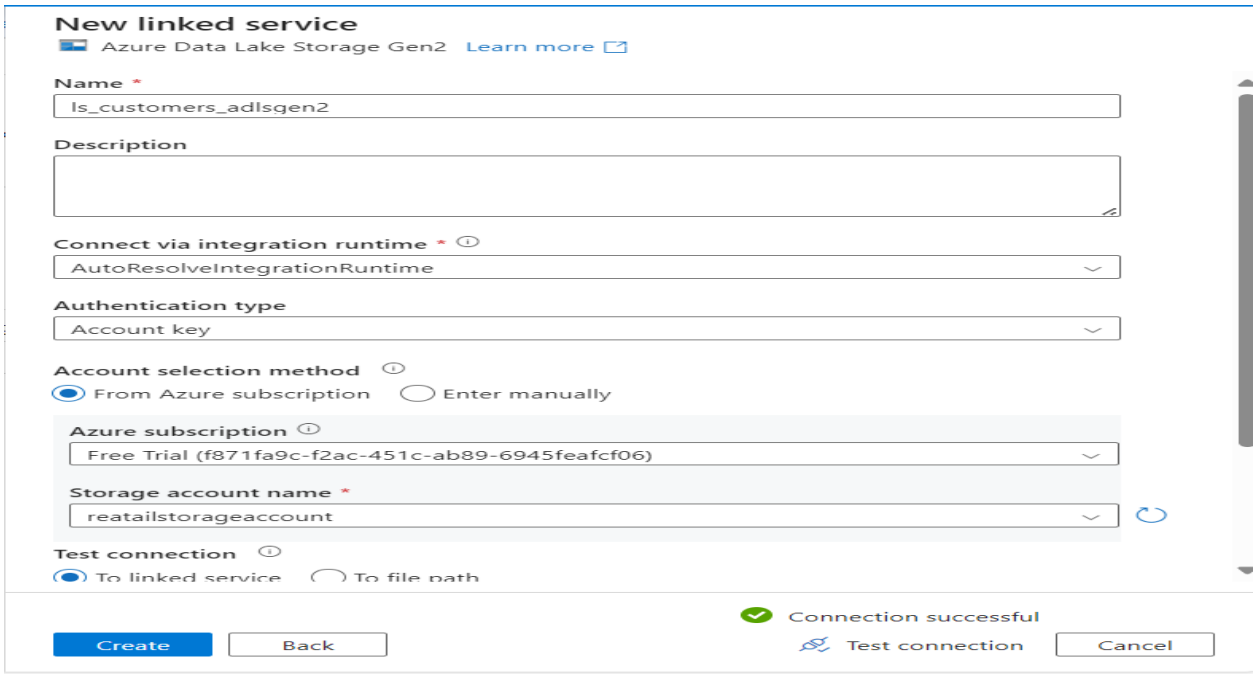


Task -2: **Ingest customers.csv file from ADLS Gen2 to Azure SQL Database.**

1. I have uploaded customer's data in **ADLS Gen2 storage account.**

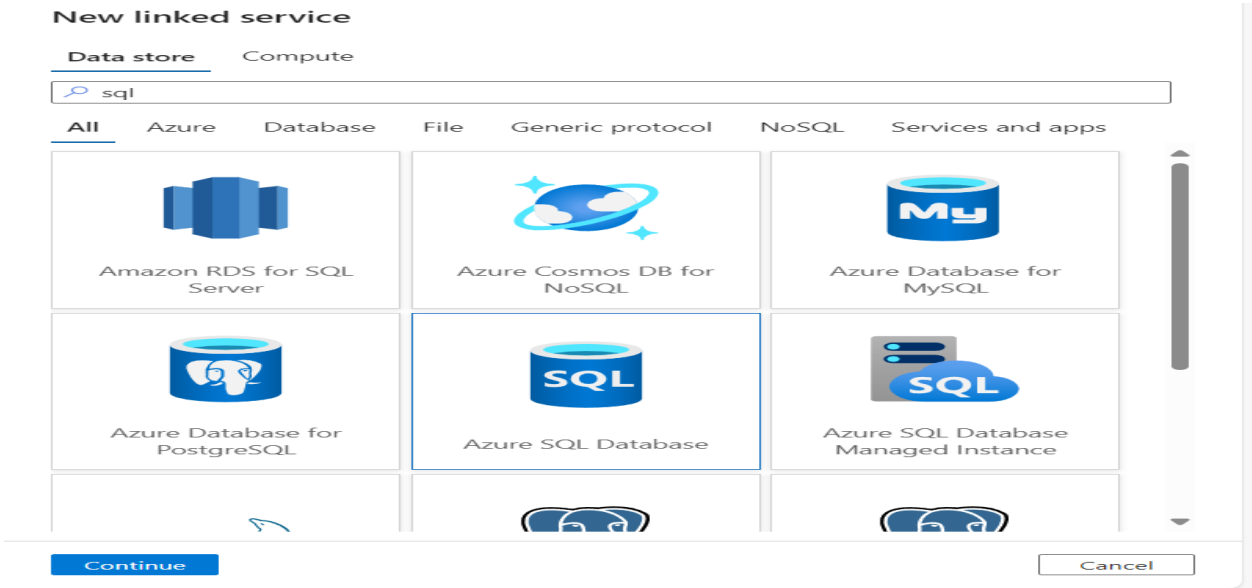


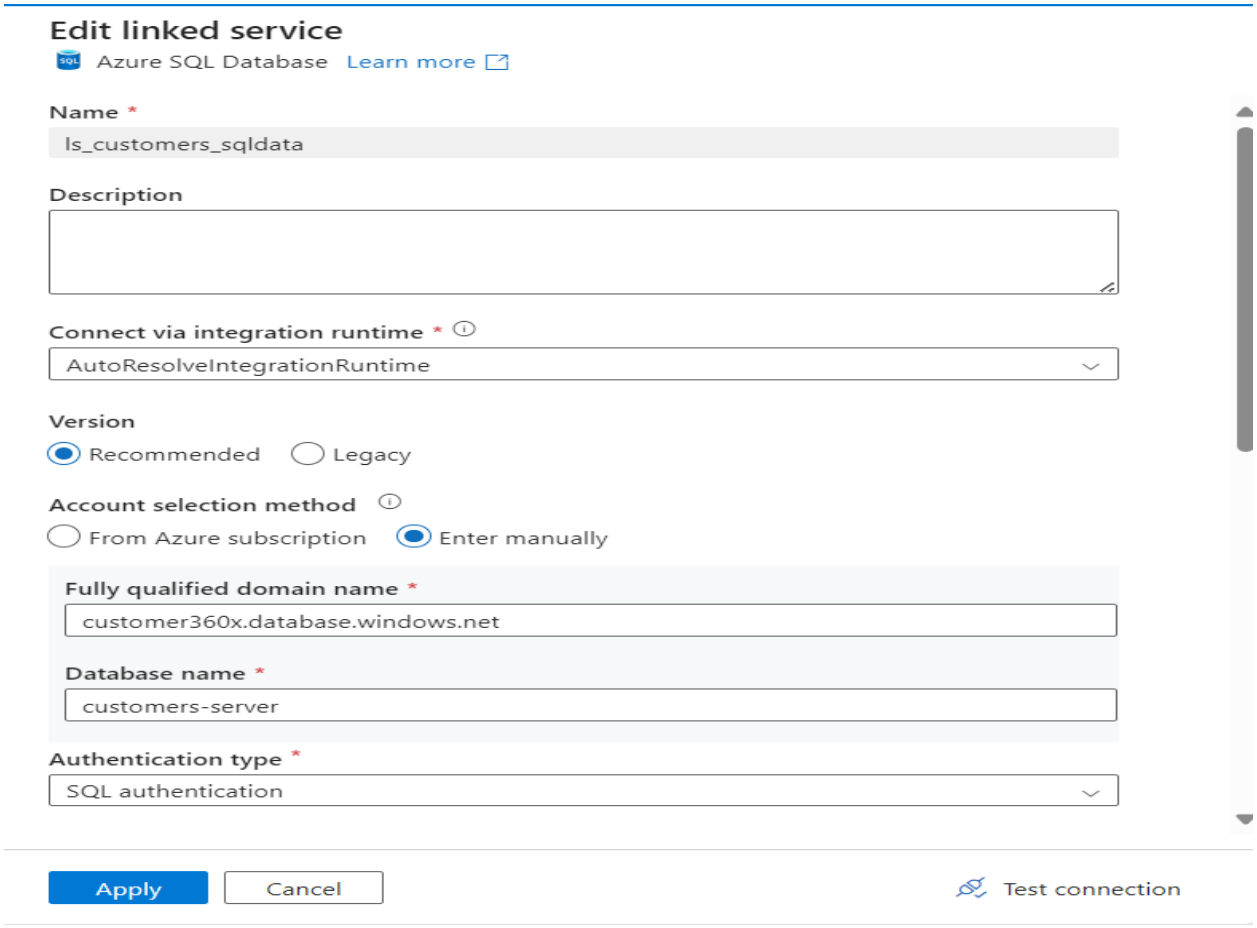
I have Create a Linked Service for Source (Adls gen2):



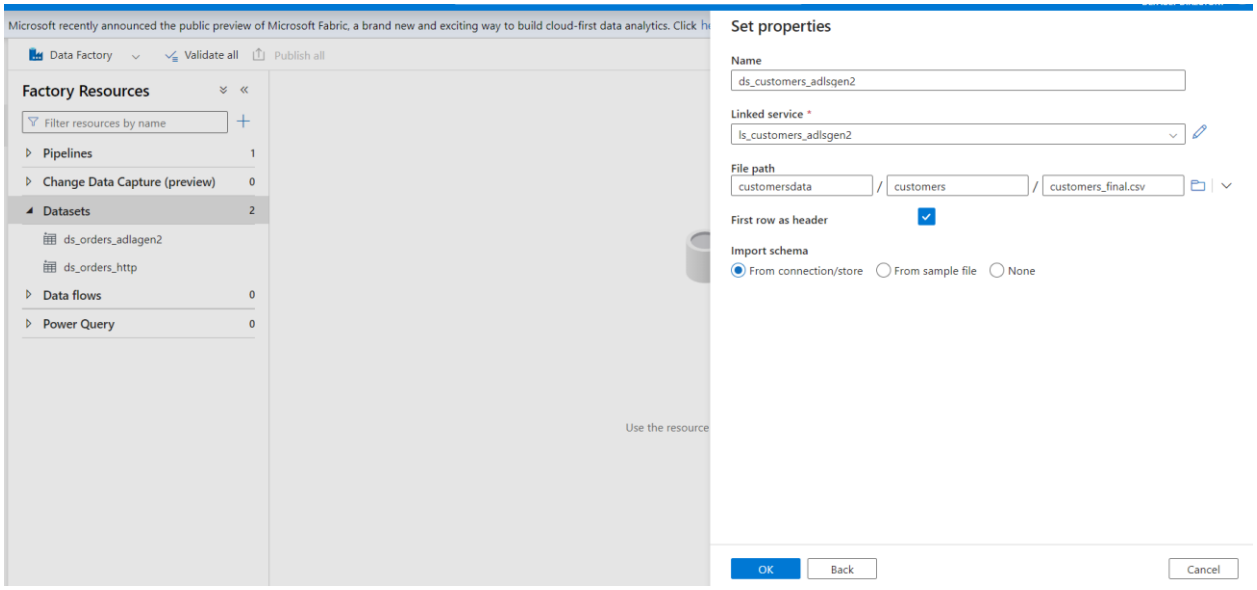
I have created a SQL Database for sink system:

I have Create a Linked Service for Sink (Azure SQL Database):

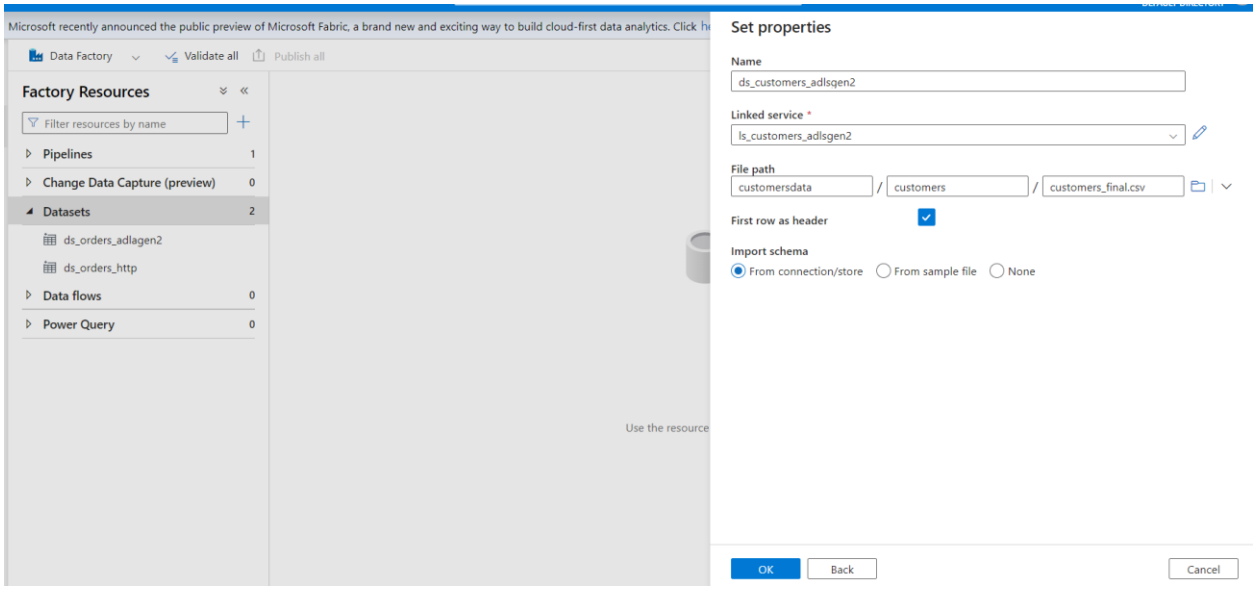


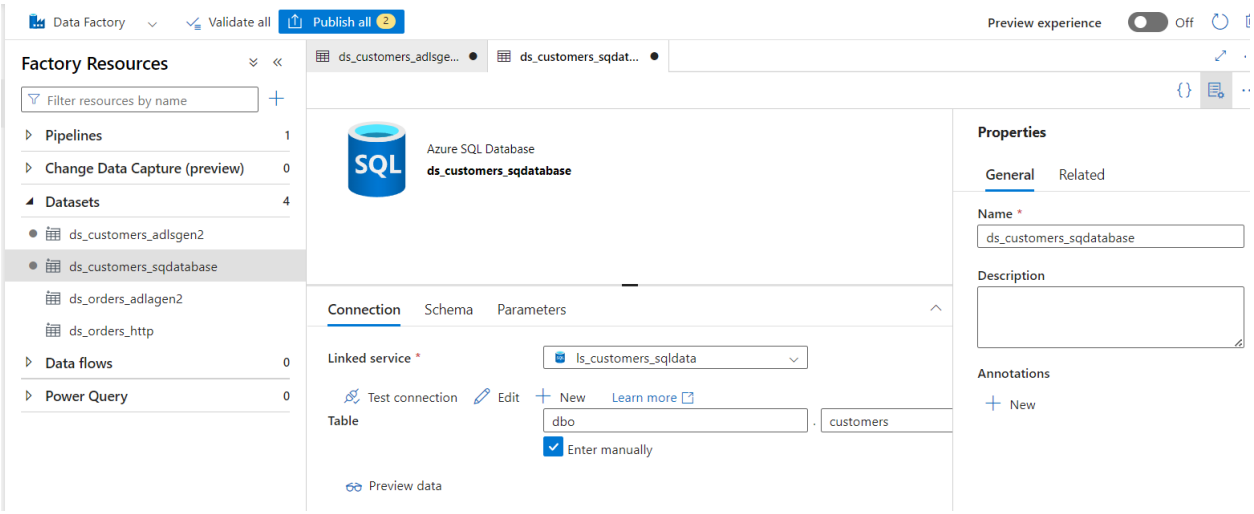


To create dataset click on Author => Datasets => New datasets for source system:



Create Dataset for Sink (SQL database) :

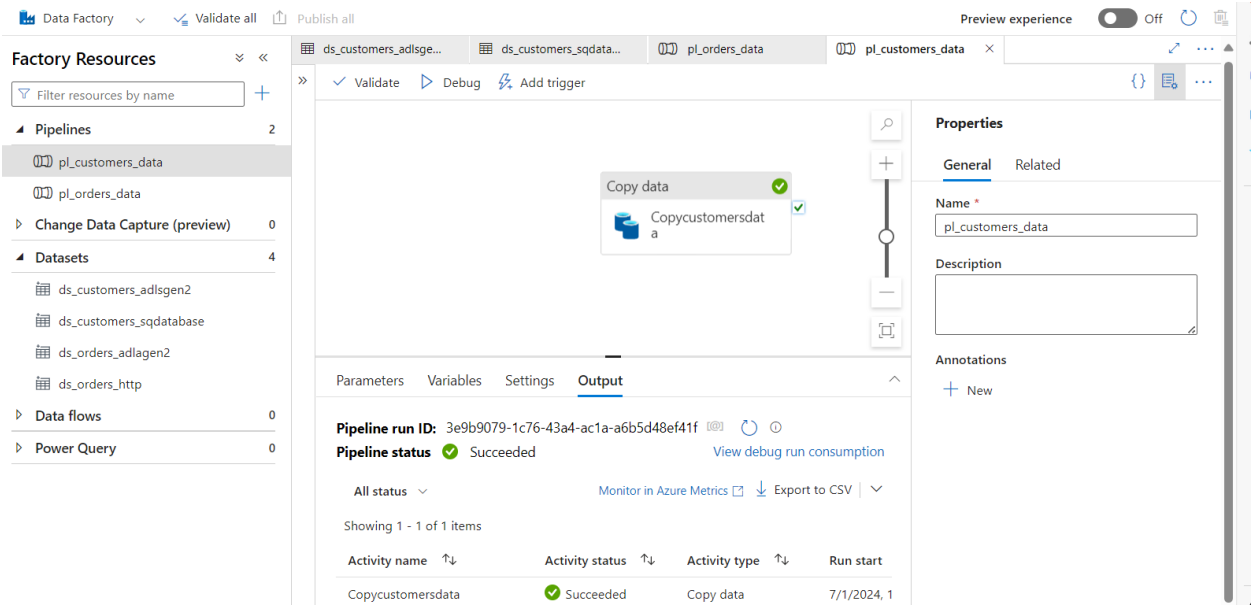




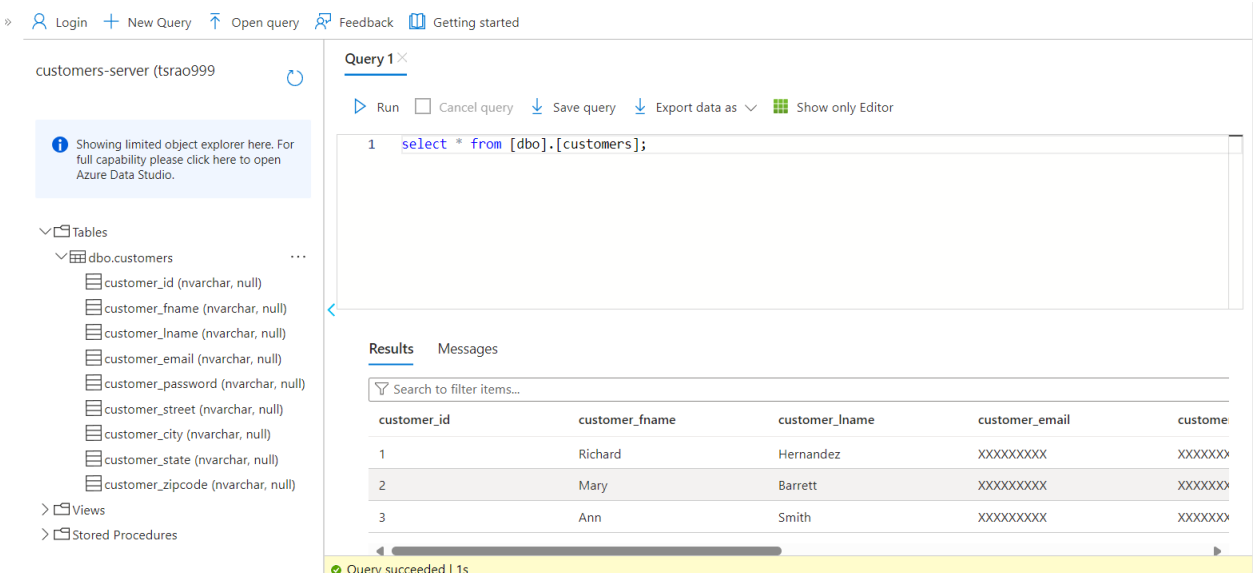
After that publish all the datasets in azure data factory.

I have created a data Pipeline and Create a Copy activity within the Pipeline:

Now click on “Move and transform” and drag copy activity in the pipeline as shown below. Rename the pipeline copy data activity as shown in screenshot.

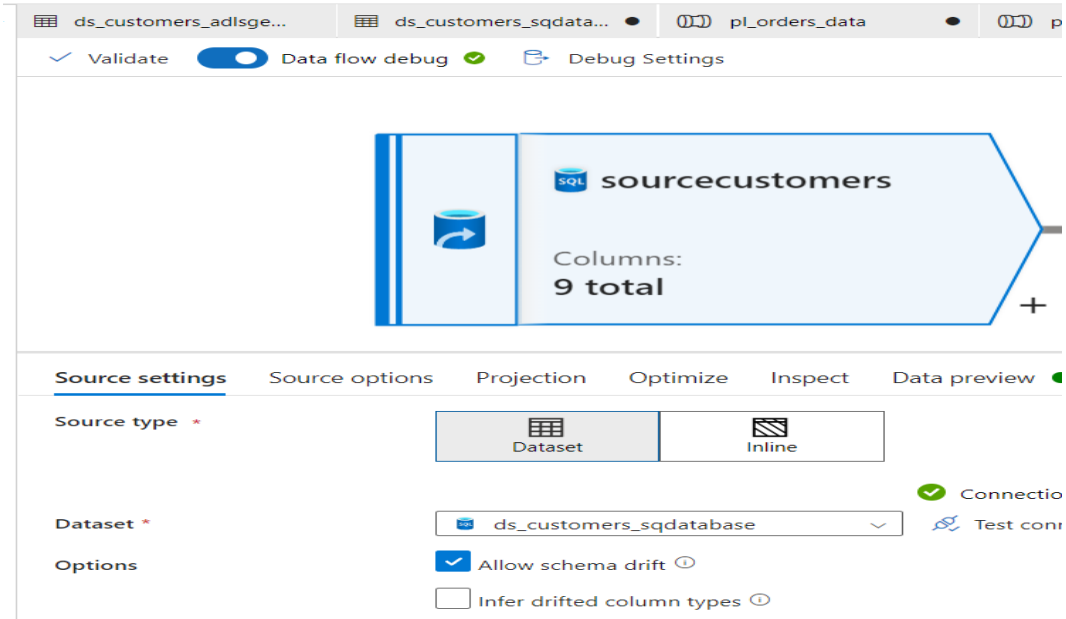


After this activity checking customers table in SQL database and I have attached below screenshot

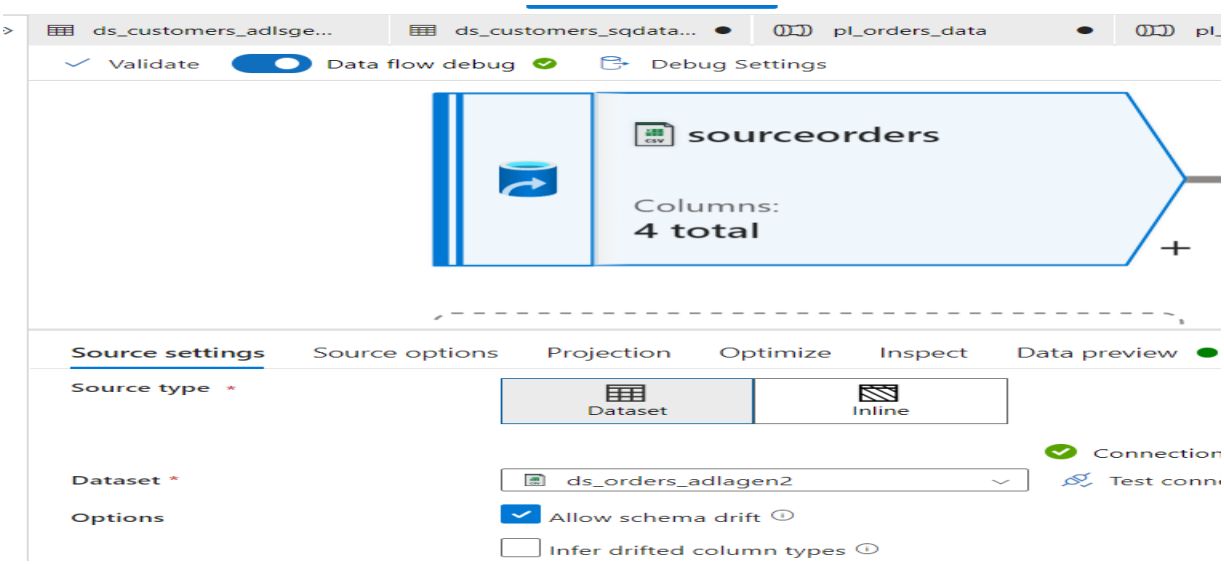


Mapping the datasets using dataflow in 2 sources:

1) Azure SQL database

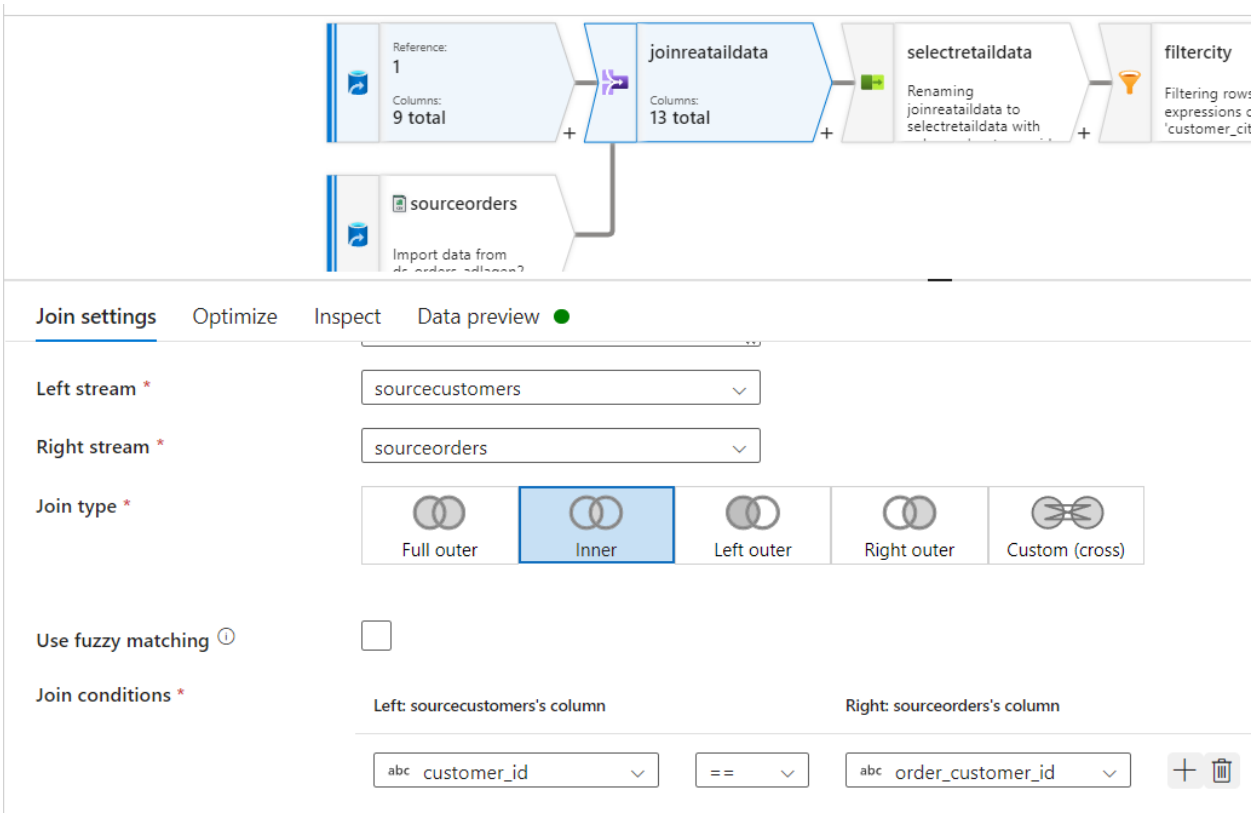


2) ADLS Gen2 storage Account

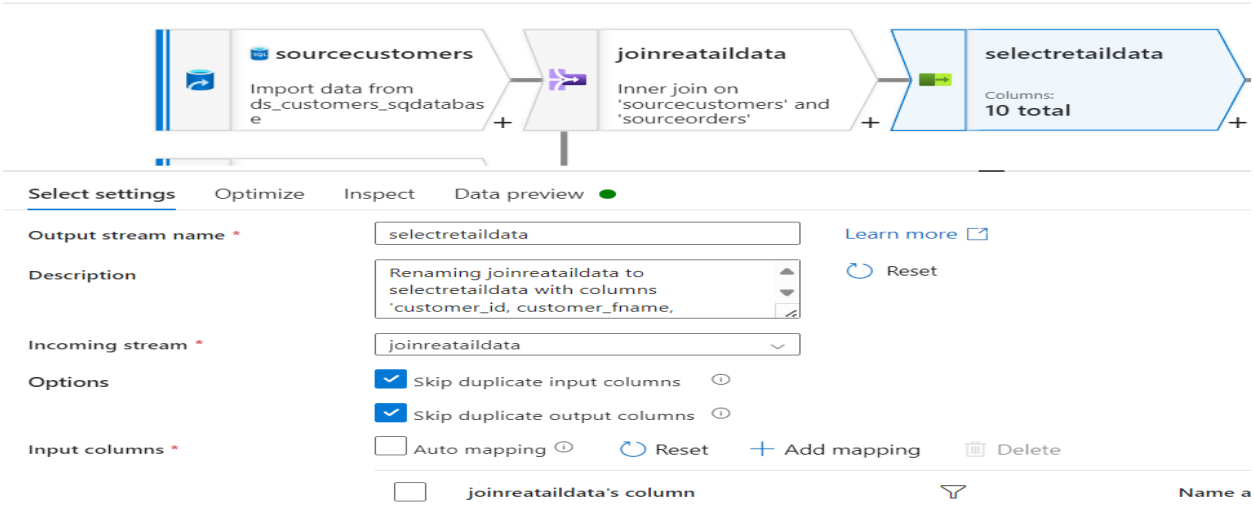


I have mapped two sources to do some transformations in dataflow below.

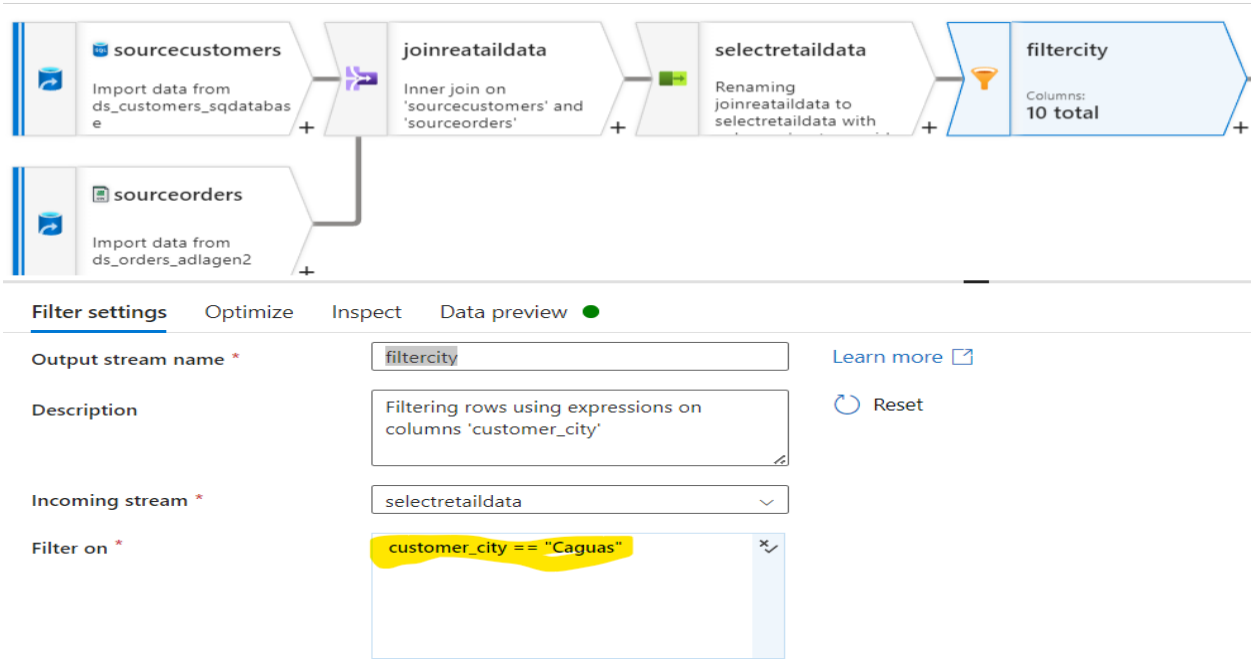
Join Transformation: I am using inner join the two sources based on common field form customer source in **customer_id** and orders source in **order_customer_id** using to create one join transformation.



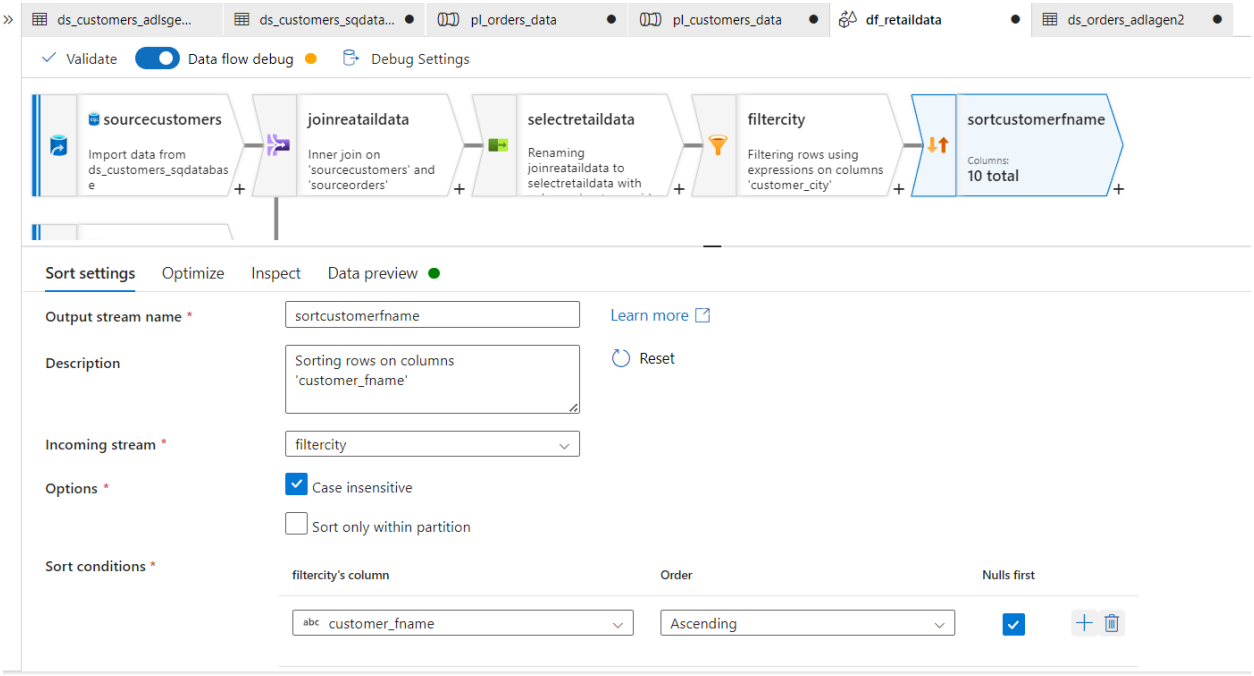
Select Transformation: incoming stream data (joinretaildata) using to select transform activity through remove the 3 fields in selectretaildata.



Filter Transformation: incoming stream data (selectretaildata) using to filter on city is Caguas to transformed output stream (filtercity) and I have attached screen shot below.



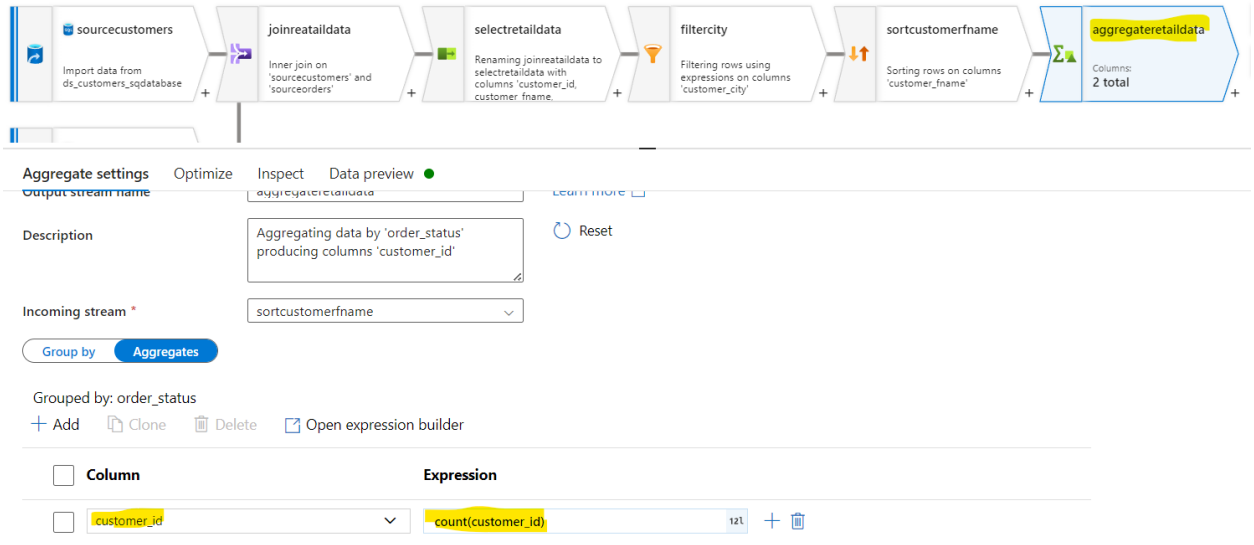
Sort Transformation: **I** have sorted the ascending order the customer fname based on filtercity incoming stream and I have attached screenshot to check it once**.**



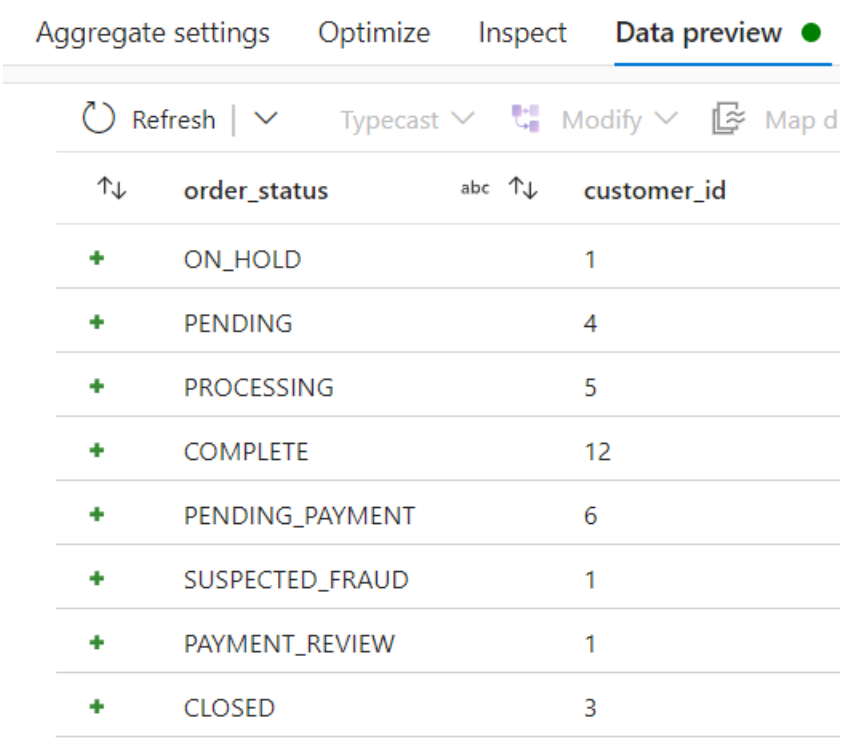
Aggregate Transformation: I am using aggregate transform to found total number of customer id's based on order status.

Now select the group by the "order status" column and Click on "Aggregates" mention order_id column and click on "Open expression builder"

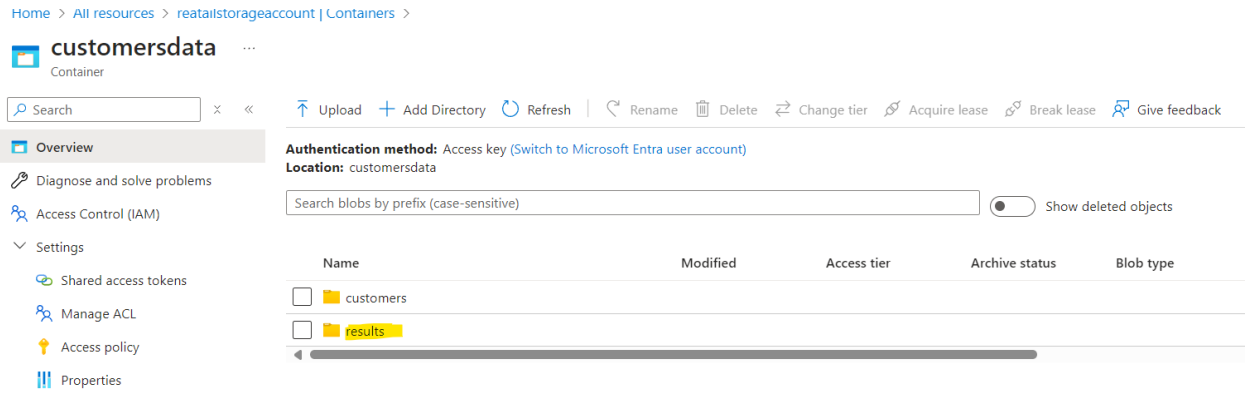
Now mention count (order_id) in "Expression" refer attached screenshot and click "save and finish"



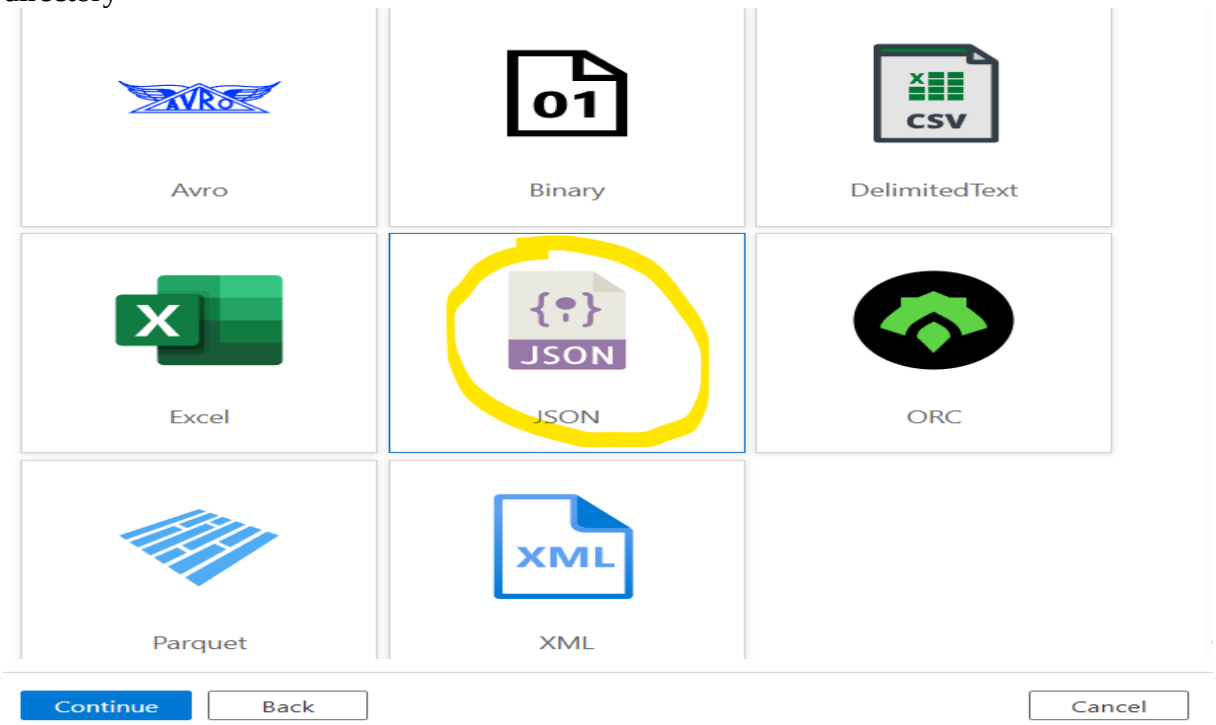
After this click on "Data Preview" and refresh it to see all the change

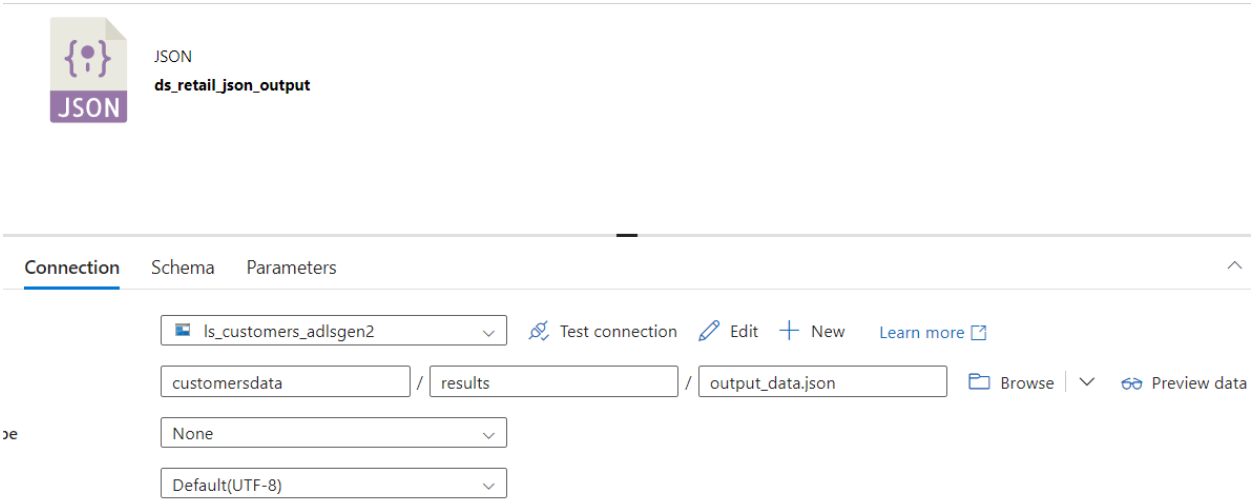


We have to Add the sink but before proceeding ahead we will create the “result” directory in the “data” container in ADLS gen2 storage “reatailstorageaccount” that we have created.

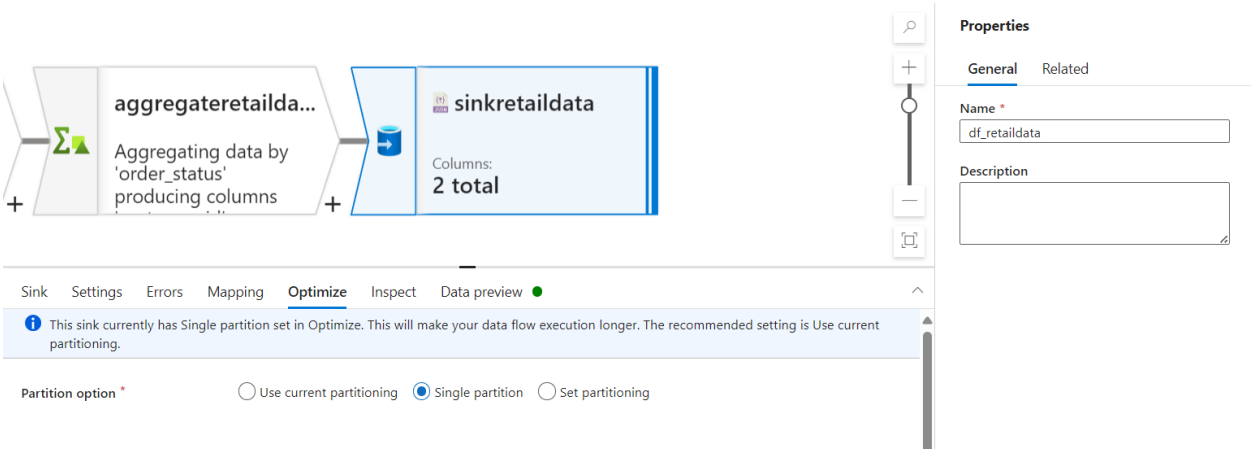


Also create dataset “ds_retail_json_output” for storing “output_data.json” data in this “result“ directory



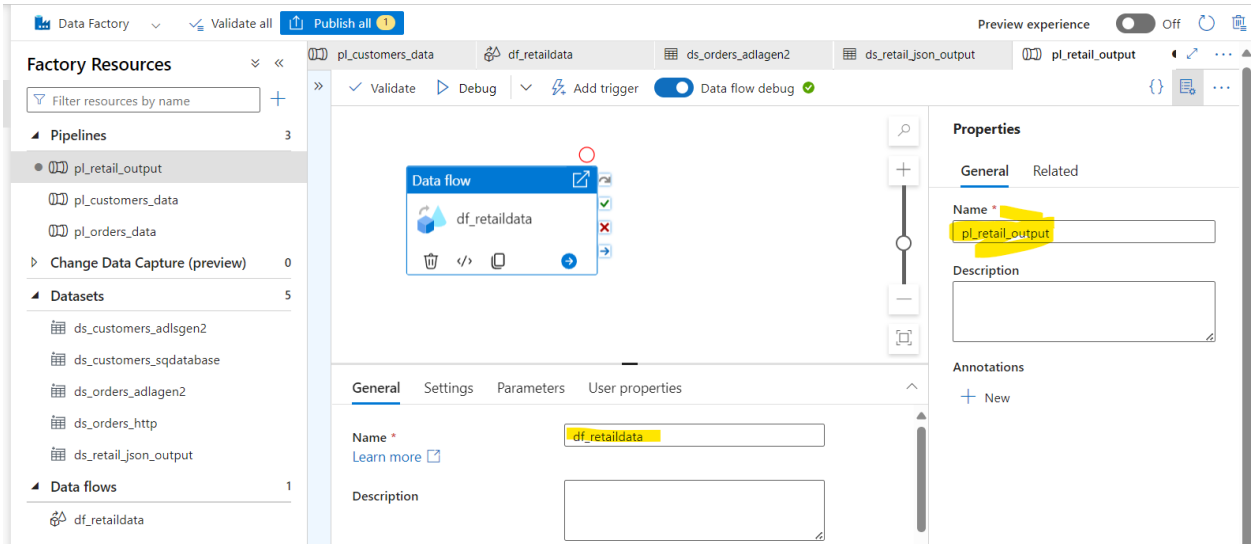


For “df_retaildata” click on “Optimize” option and select “Single partitioning” to store the complete output in a single file.

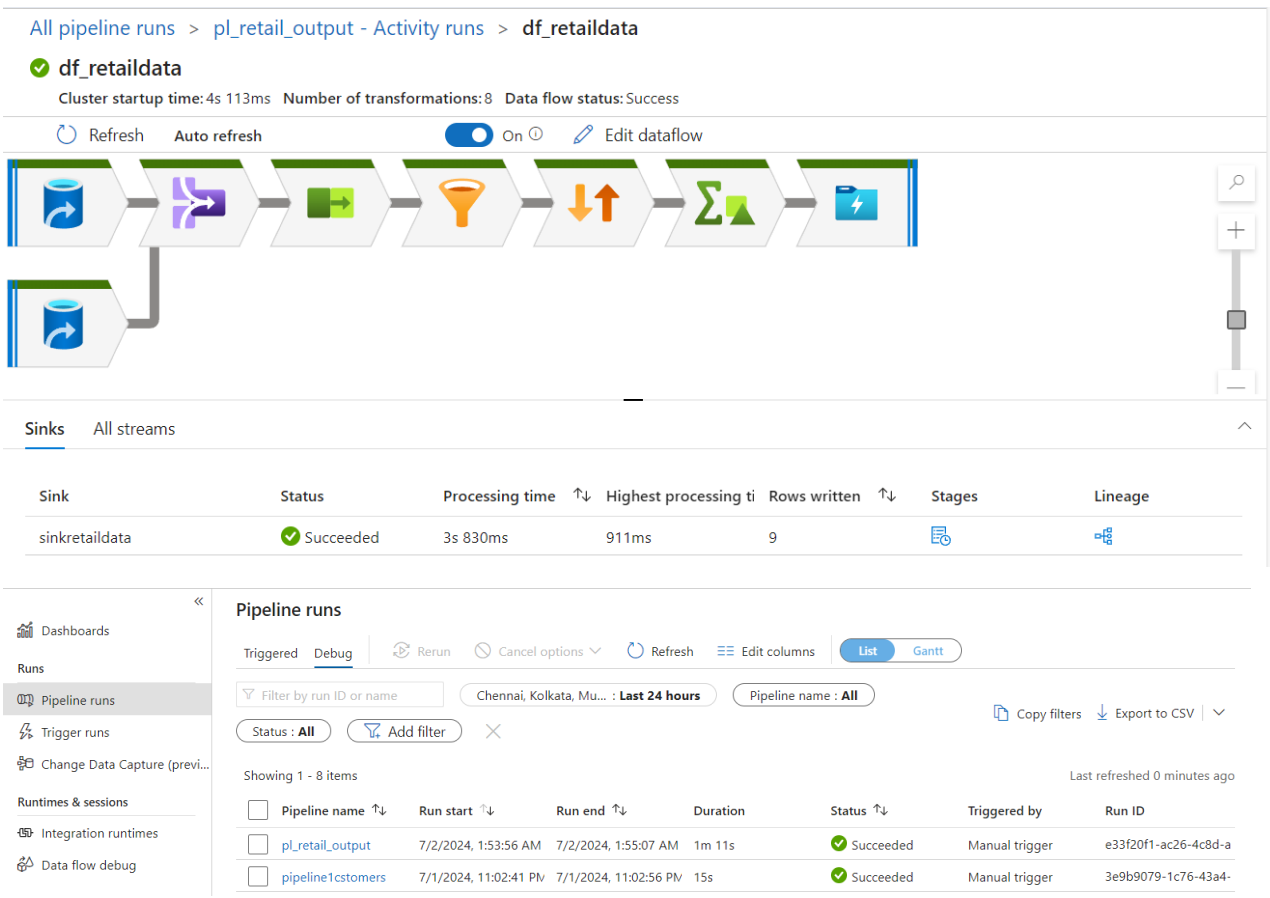


Note: Before proceeding ahead. Be sure to publish these changes to make them active and available for use in my data workflows.

I have created the new pipeline “pl_retail_output” and will drag this dataflow in the pipeline “pl_retail_output”.



Debug the pipeline and validate it. i can see this running pipeline in the monitor tab (Debug) and publish it.



I have opened my storage account and see the output data in results container.