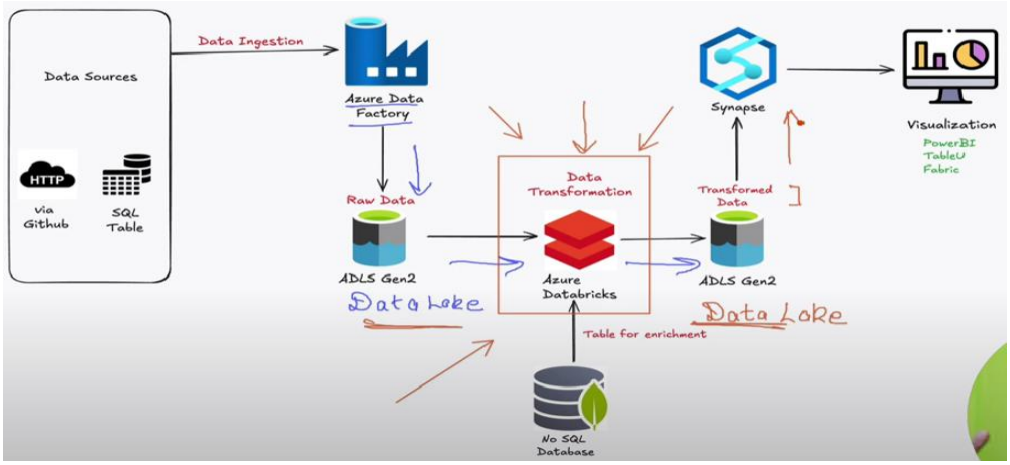
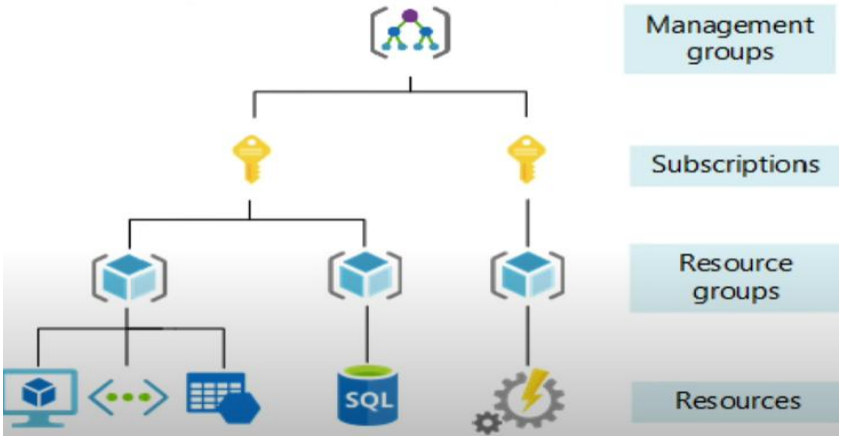
Data Engineering Project : Ecommerce

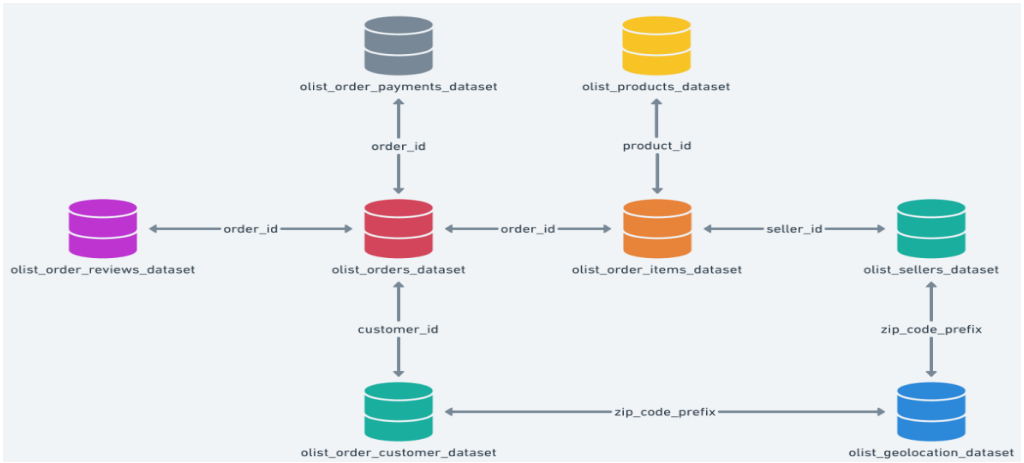
Architecture of the project



Azure account Structure



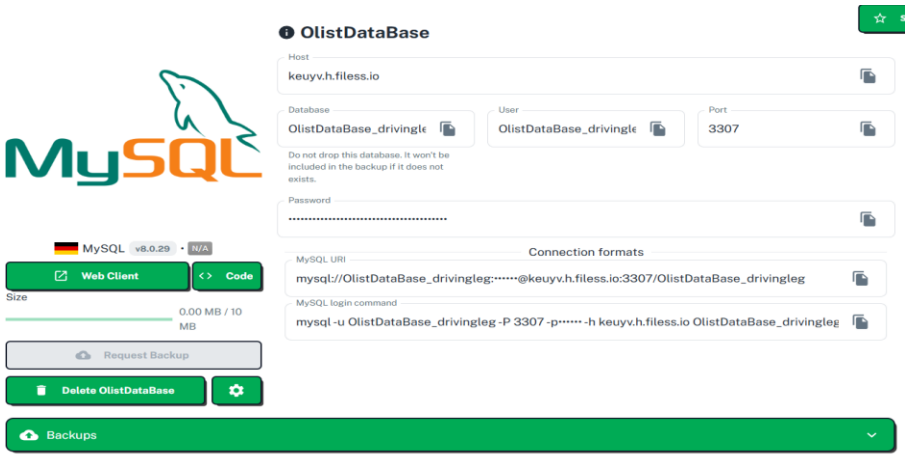
Brazillian Ecommerce Data by Olist [dataset link]

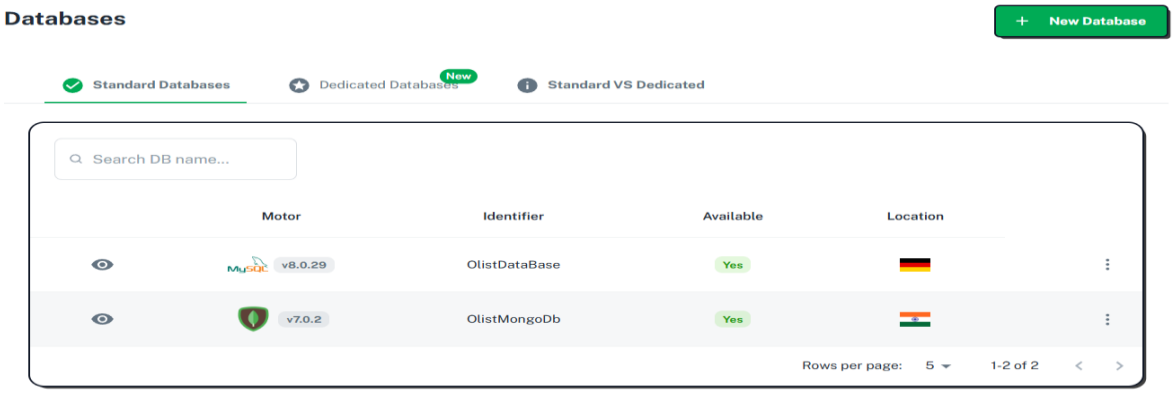


Github : [link]

Filess.io : [link] **:->** this helps us to store our SQL and Nosql database

Data Ingestion: here we have our data in http(github) and Filess.io Database i.e we ingested here





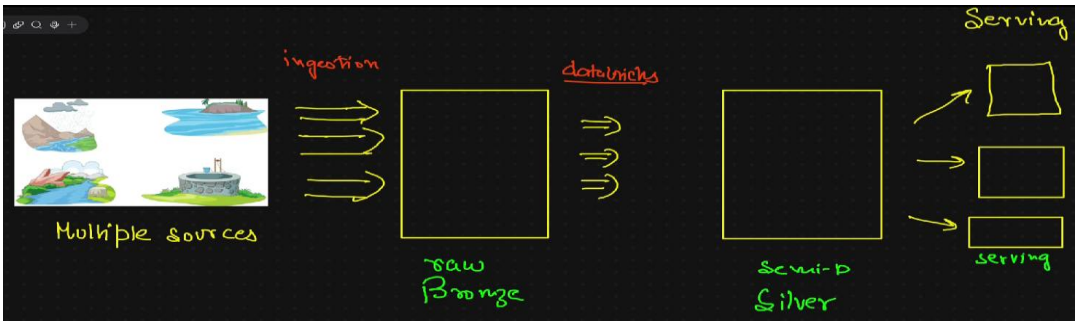
Here in above databases servers to Insert our data now form github and or through the local device

→We just need to write a simple python scripts do insert data with MySQL-connector [link]

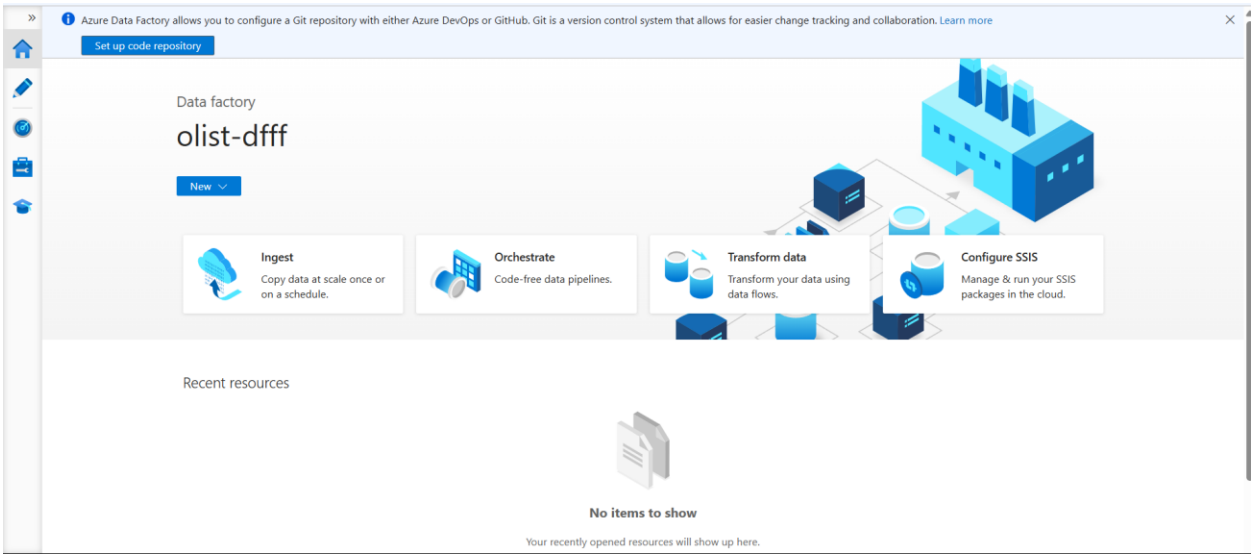
During ingestion if we face any problem of while uplading just adjust the ***batch_size***

Now we have data in github and servers so perform **Ingestion** ingest this data on *Azure Data Factory*

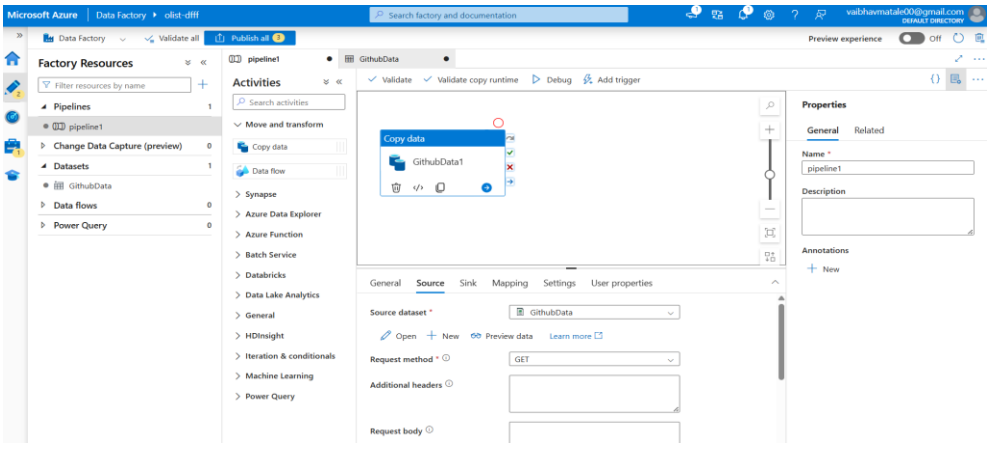
Pipeline



Data factory



Ingesting data fron Github



Create a storage account

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * : Azure for Students

Resource group * : ecomm-live

Create new

Instance details

Storage account name * : olistdatastorageaccount

Region * : (Asia Pacific) Central India

Deploy to an Azure Extended Zone

Primary service : Select a primary service

Performance * : **Standard:** Recommended for most scenarios (general-purpose v2 account)

Premium: Recommended for scenarios that require low latency.

Redundancy * : Locally-redundant storage (LRS)

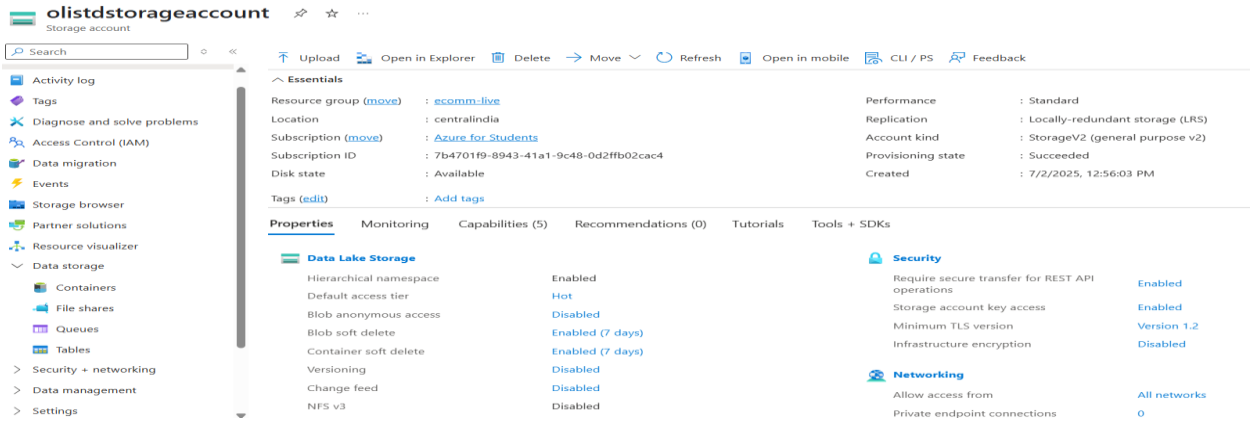
Redundancy is the copies of this data stored in local centre if they failed or server is down you won't acces the data for change it geo or zone but accordingly cost is applied higher

Hierarchical Namespace

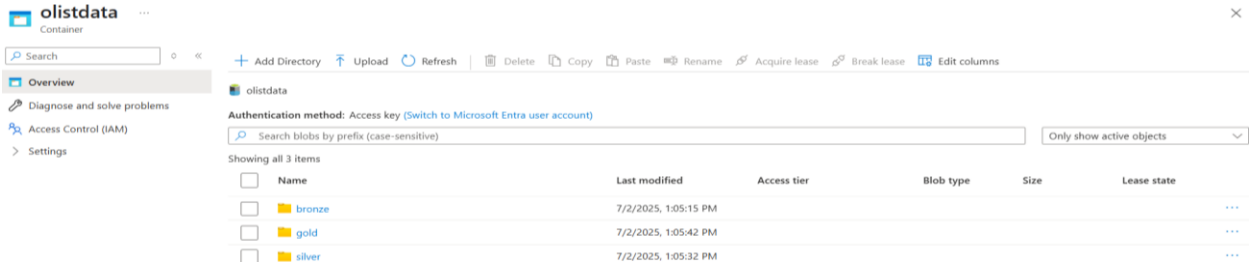
Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) Learn more

Enable hierarchical namespace

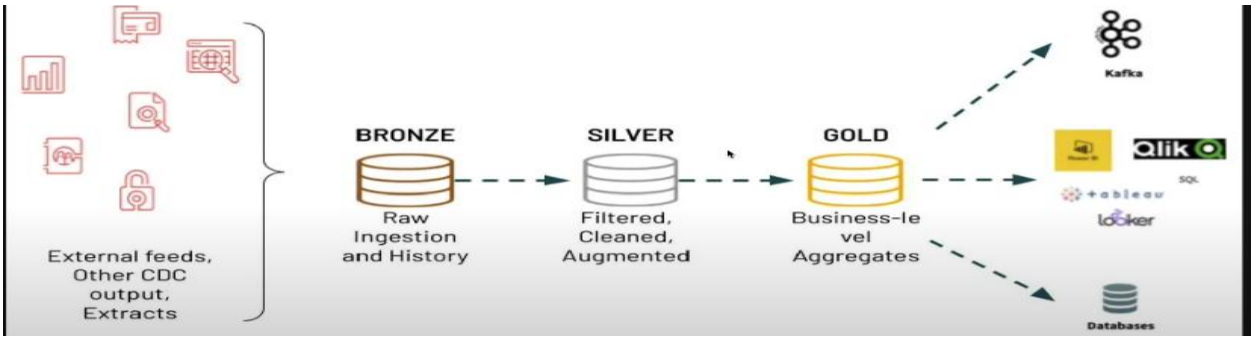
Tern on this to create folders and all in hierarchy



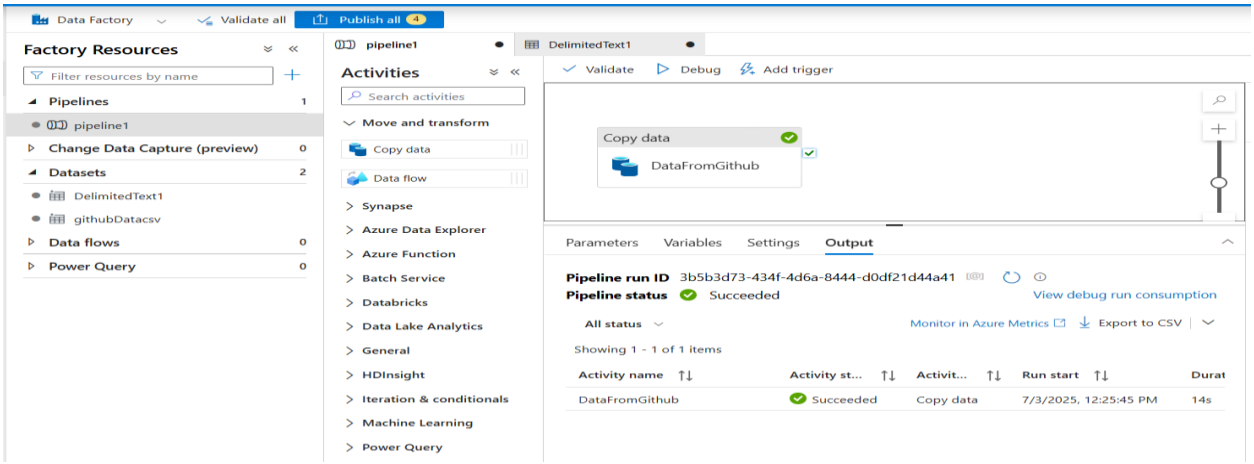
I've just created the container olistdata with directory bronze, silver, gold by medellian acrhitec..



Medallian Architecture : of storing data

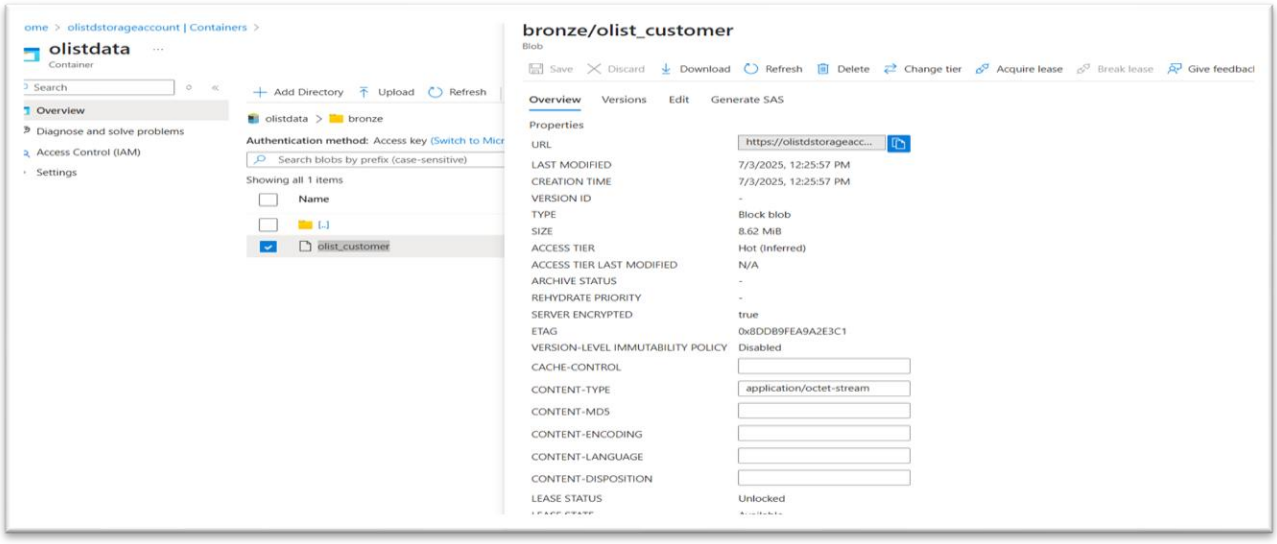


Now we created the container and after that created pipeline in data factory with this storage account



Here above we need to validate and Debug this pipeline after that will get this above page tick

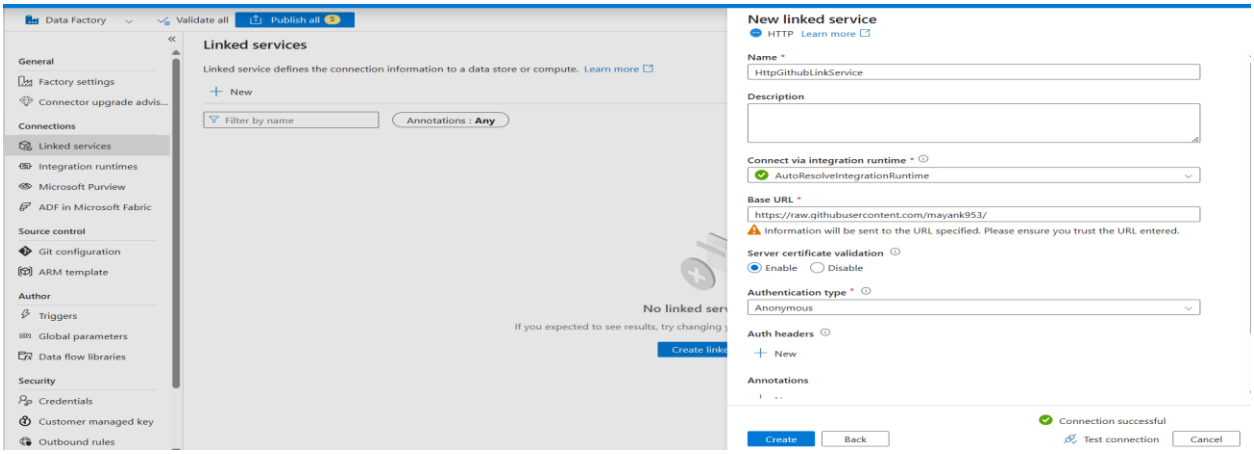
Here above pipeline we take data from github source and sink it to Azure Storage account with ADLS Gen2

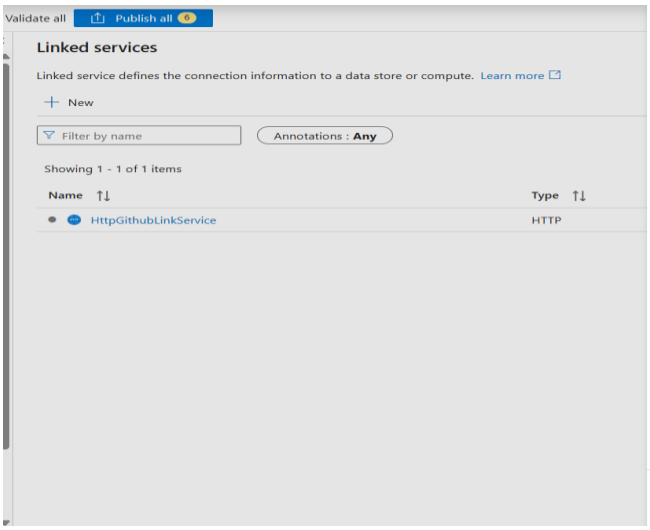


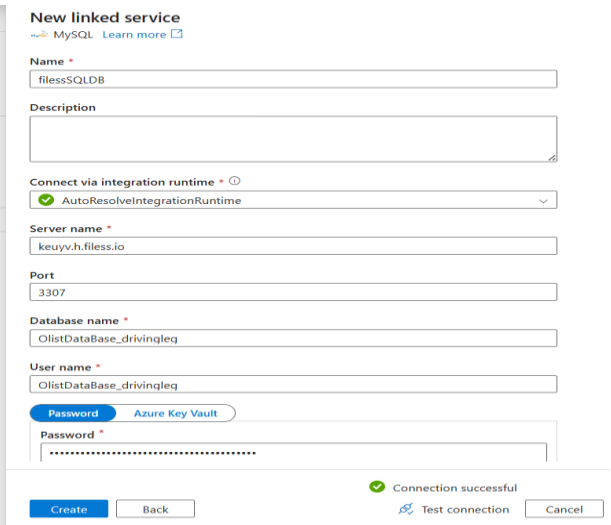
Here we get that olist customer folder in storage account

tier->how frequenly we can use this data ->archive , hot , etc according to cost we use this

Instead of adding single data we can use **Linked service** **->**provide just base url here

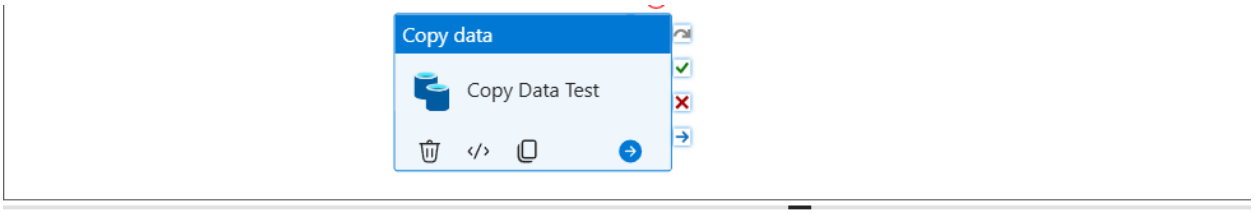


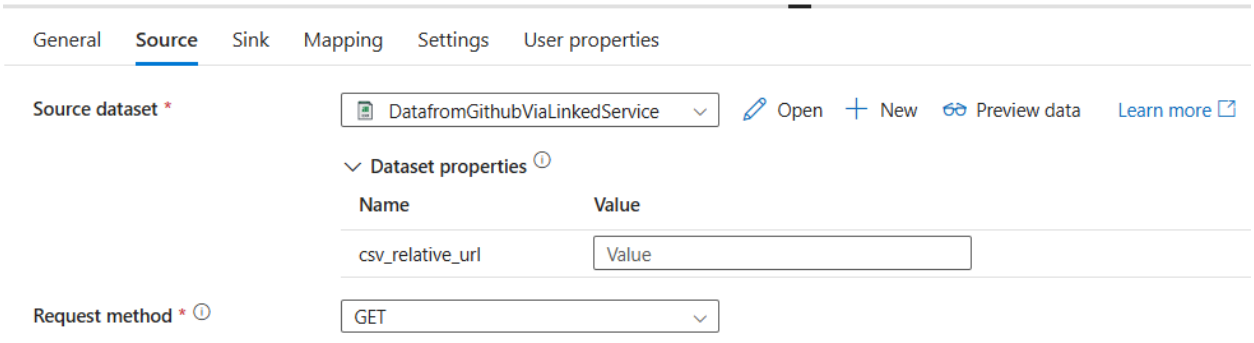


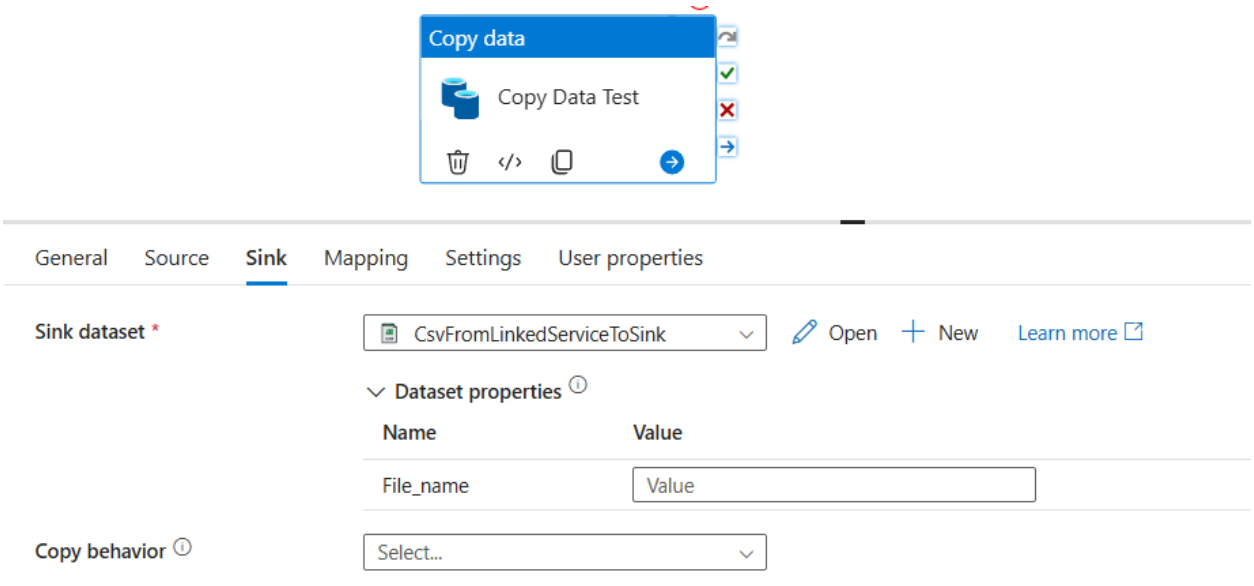


Here we linked to Database server files which mysql data

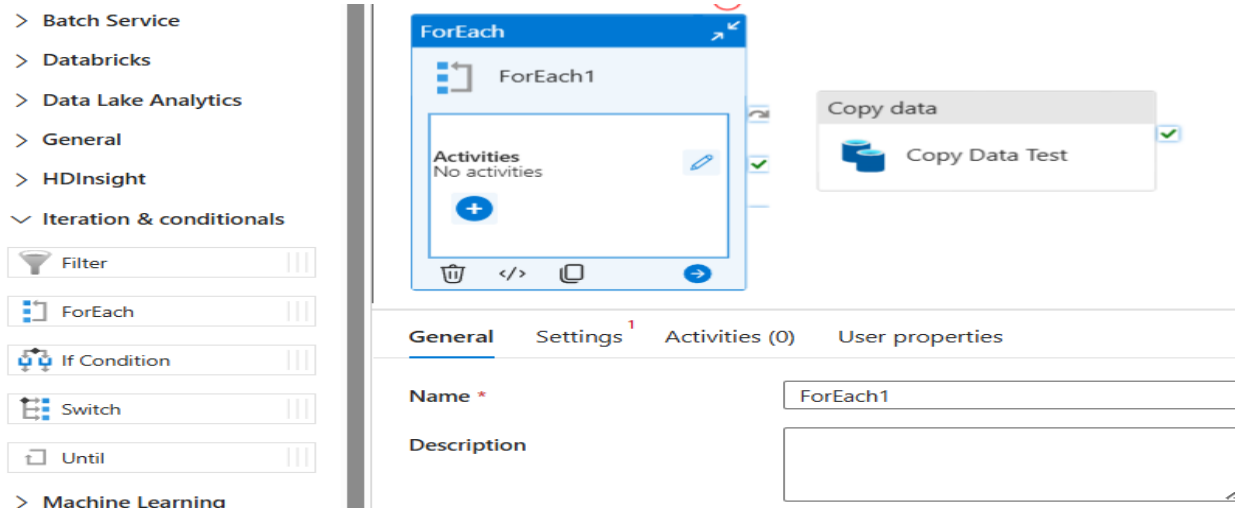
->Will now make a pipeline of this all linked servicres in ***Source*** we have data from sources like *http* and in ***Sink*** we have data from *Azure Data Storage* in *bronze folder*







This is for each loop which helps to get multiple data in loop instead of getting single single data

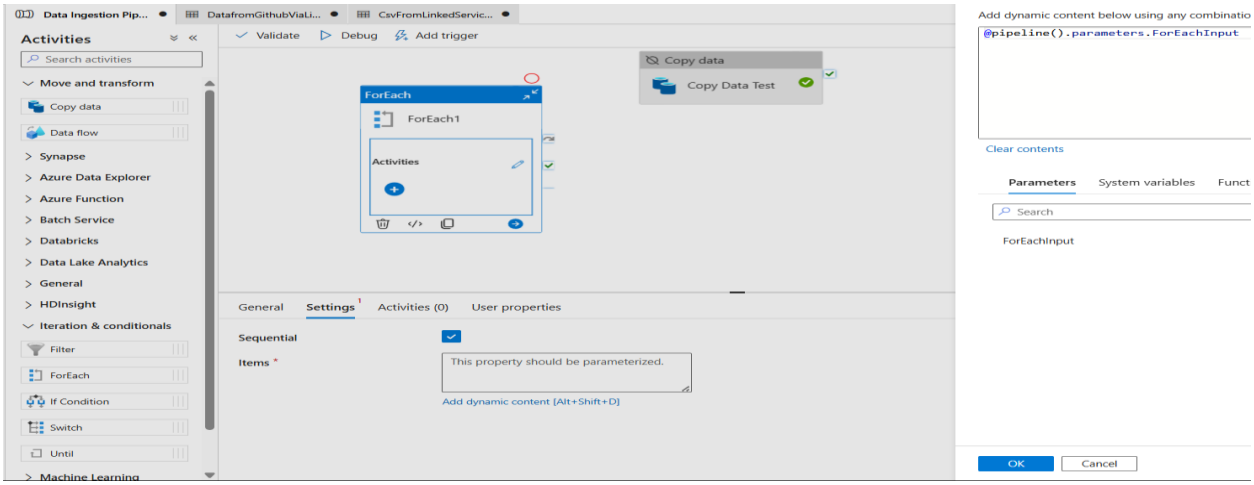


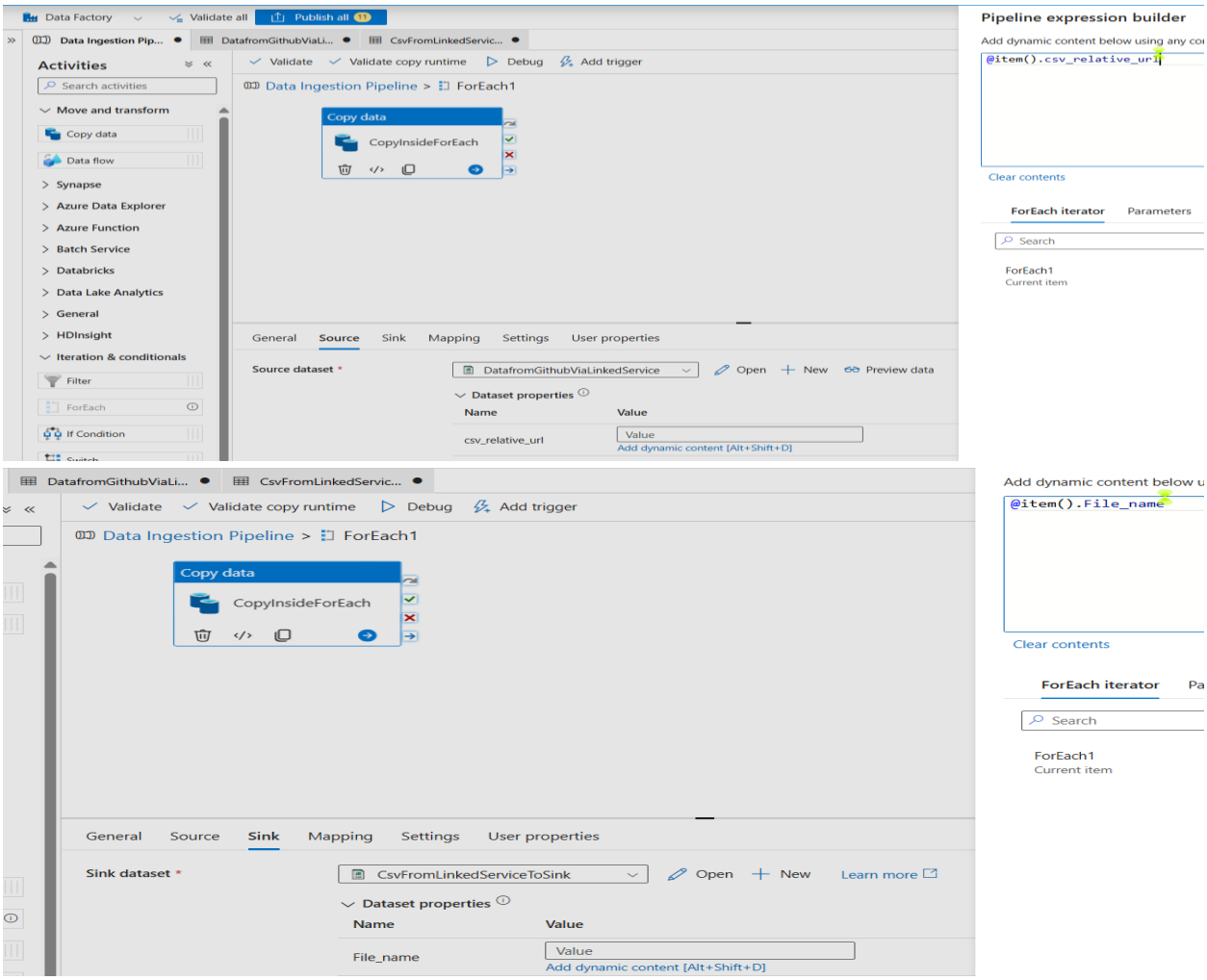
This below base url we pass in for each to as we set our Relative url in main pipeline it will get all the data in a loop with ietrations.

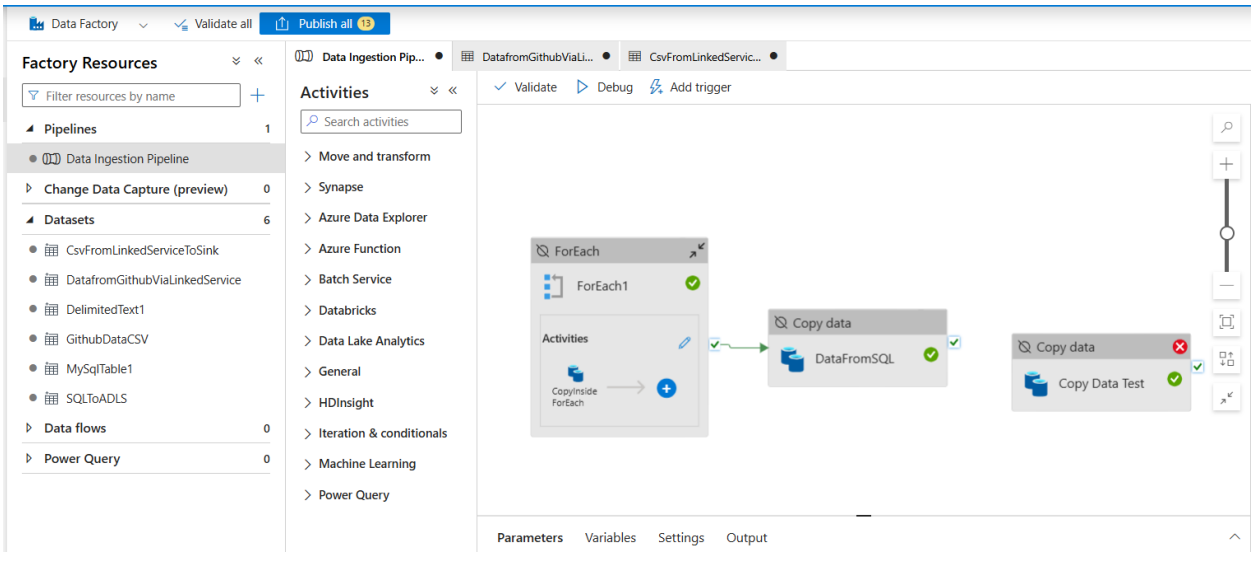

```
[
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_customers_dataset.csv",
  "file_name": "olist_customers_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_geolocation_dataset.csv",
```

```
  "file_name": "olist_geolocation_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_order_items_dataset.csv",
  "file_name": "olist_order_items_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_order_reviews_dataset.csv",
  "file_name": "olist_order_reviews_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_orders_dataset.csv",
  "file_name": "olist_orders_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_products_dataset.csv",
  "file_name": "olist_products_dataset.csv"
 },
 {
  "csv_relative_url": "BigDataProjects/refs/heads/main/Project-Brazillian%20Ecommerce/Data/olist_sellers_dataset.csv",
  "file_name": "olist_sellers_dataset.csv"
 }
]
```

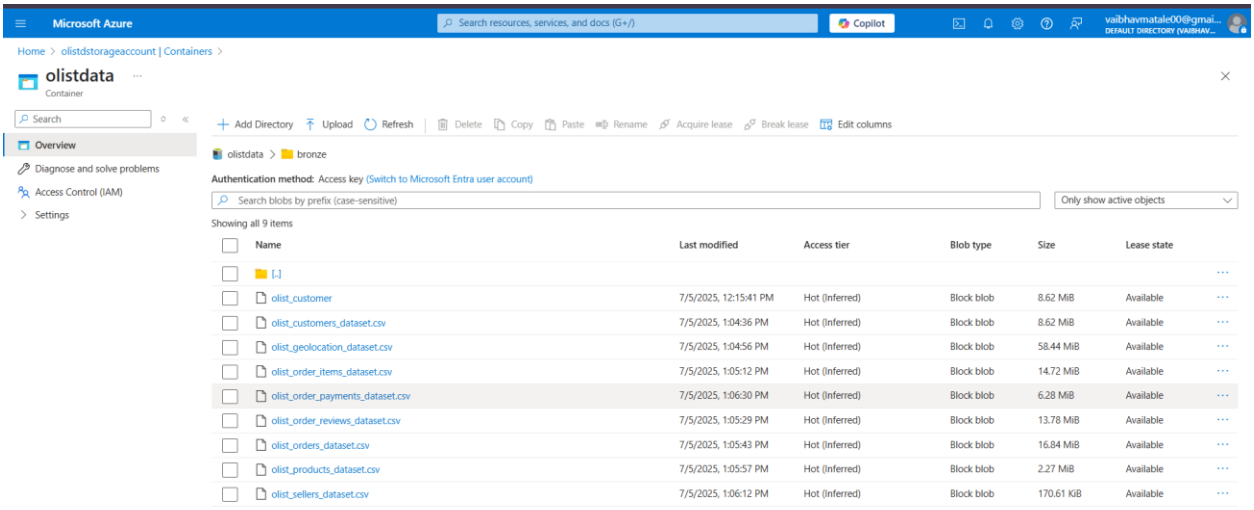

We basically pasete this to our parameter “ForEachInput” which is typed Array



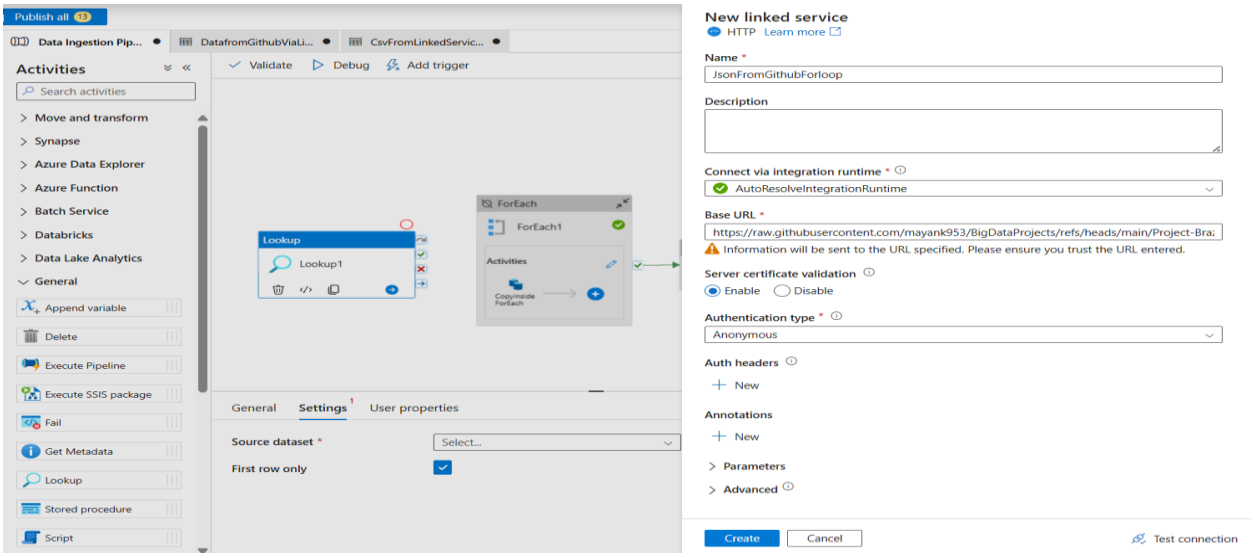

Data Factory
Validate all
Publish all
Activities
Search activities
Move and transform
Copy data
Data flow
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Filter
ForEach
If Condition
Validate
Validate copy runtime
Debug
Add trigger
Data Ingestion Pipeline > ForEach1
Copy data
CopyInsideForEach
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
DatafromGithubViaLinkedService
Open
New
Preview data
Dataset properties
Name
Value
csv_relative_url
Value
Add dynamic content [Alt+Shift+D]
Pipeline expression builder
Add dynamic content below using any co
@item().csv_relative_url
Clear contents
ForEach iterator
Parameters
Search
ForEach1
Current item
DatafromGithubViaLi...
CsvFromLinkedServic...
Validate
Validate copy runtime
Debug
Add trigger
Data Ingestion Pipeline > ForEach1
Copy data
CopyInsideForEach
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
CsvFromLinkedServiceToSink
Open
New
Learn more
Dataset properties
Name
Value
File_name
Value
Add dynamic content [Alt+Shift+D]
Add dynamic content below u
@item().File_name
Clear contents
ForEach iterator
Search
ForEach1
Current item


Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
Data Ingestion Pipeline
Change Data Capture (preview)
Datasets
CsvFromLinkedServiceToSink
DatafromGithubViaLinkedService
DelimitedText1
GithubDataCSV
MySqlTable1
SQLToADLS
Data flows
Power Query
Data Ingestion Pip...
DatafromGithubViaLi...
CsvFromLinkedServic...
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
ForEach
ForEach1
Activities
CopyInside ForEach
Copy data
DataFromSQL
Copy data
Copy Data Test
Parameters
Variables
Settings
Output

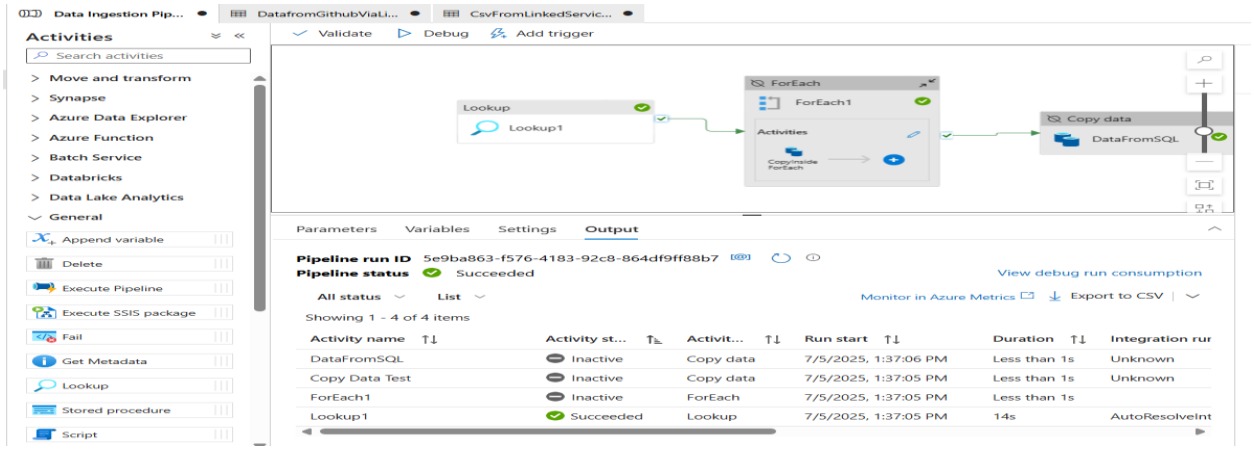
Till we have loaded data from sql server and github in our storage account

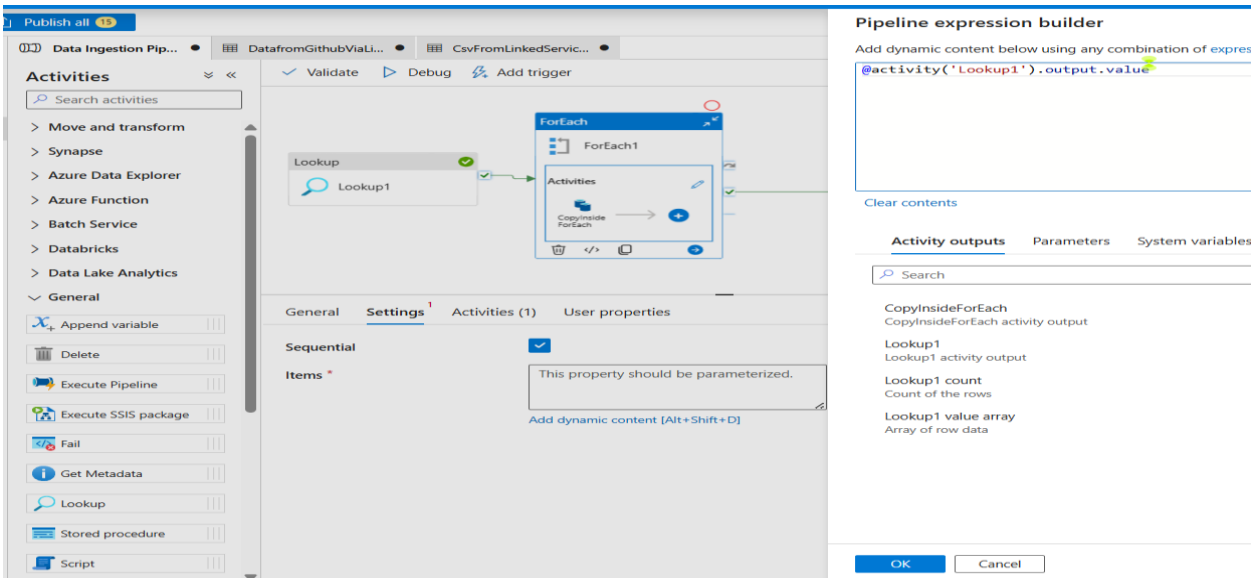


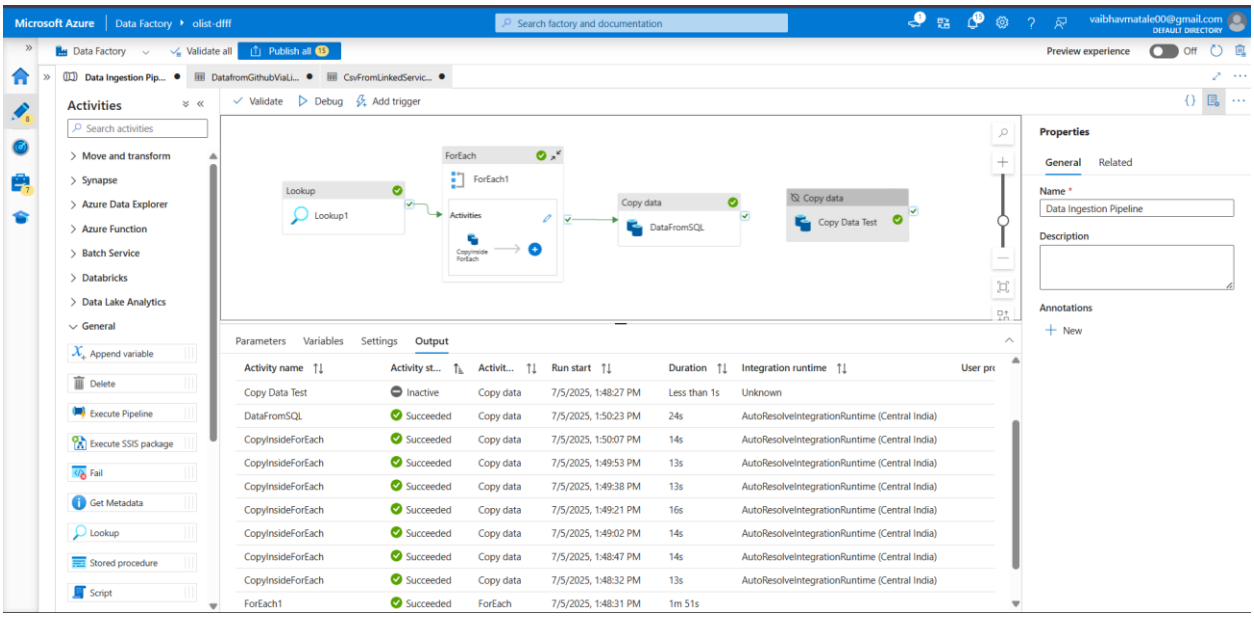
Lookup



Here we pass url of links of data in json from github so we pass raw json link







Here I published my this above pipeline

DataBricks : Databricks is a unified data analytics platform built for big data and machine learning. It is based on Apache Spark and allows you to do everything from data engineering, data science, to machine learning — all in one place.

It offers:

- A collaborative environment for data scientists, data engineers, and analysts
- Support for Python, Scala, SQL, and R

- Built-in ML & DL libraries like MLlib, TensorFlow, PyTorch, scikit-learn
- Auto-scaling and auto-termination for big data clusters
- Integration with many data sources like Azure Data Lake, AWS S3, JDBC, etc.

Azure Databricks?

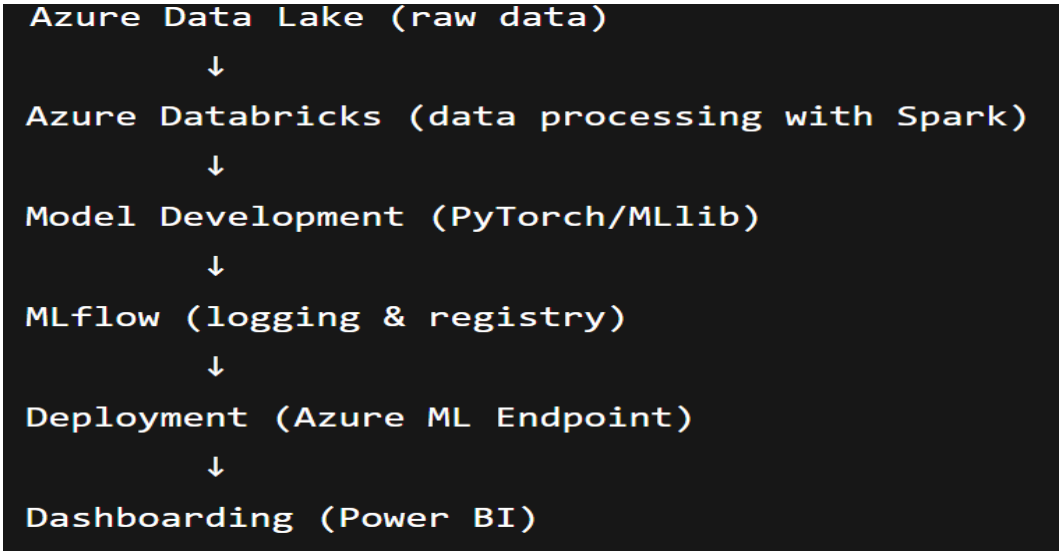
Azure Databricks is a **Microsoft Azure**-based version of Databricks, fully integrated with the Azure ecosystem.

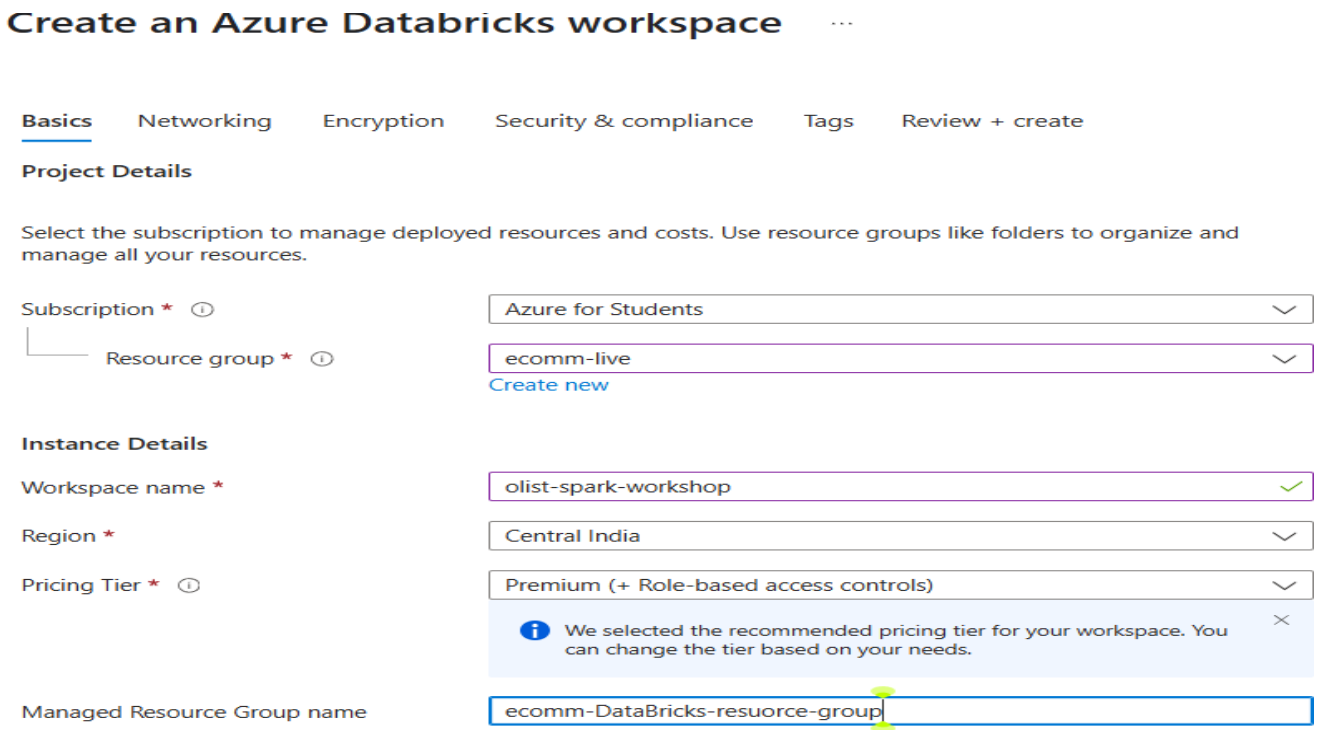
✅ It combines:

- The **Databricks platform**
- With **Azure services** like:
    - **Azure Data Lake Storage**
    - **Azure Synapse Analytics**
    - **Azure Machine Learning**
    - **Azure Key Vault**
    - **Azure DevOps**

✳️ **Features:**

- One-click setup for Spark clusters
- Enterprise-grade security with **AAD (Azure Active Directory)**
- Integrated **workspace with notebooks (like Jupyter)**
- Can connect to Power BI, Azure Blob, SQL DBs, etc

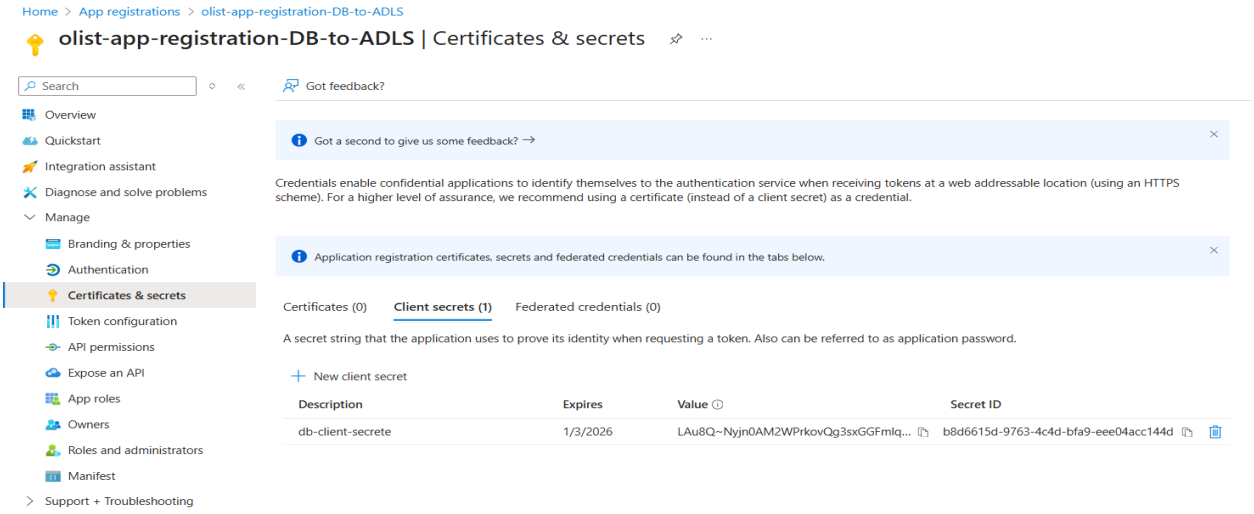




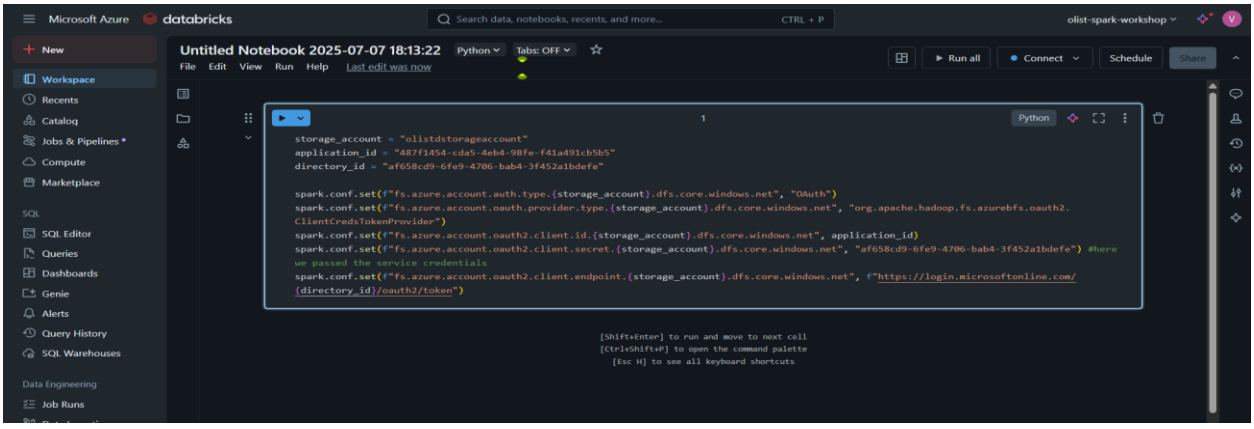
Now We do Azure DataBricks to Azure Lake Gen 2 Connections

Refer this link For Azure Data Lake Connection Tutorial

We have to do a connection for the connection to ADLS any provide app_id and other to there



Here we pass all this in Notebook of Databricks for the connection



Assigning Role to access storage containers in Azure with that we've created registration olist-db-adls