Project: Azure Data Pipeline

Problem Statement: You are working as a Data Engineer at a dynamic and forward-thinking company that leverages Azure as its cloud platform of choice. This company recognizes the immense value locked within data and has embarked on a mission to harness the power of its data assets for informed decision-making and business optimization. In pursuit of this goal, the company has entrusted you with the responsibility of designing and implementing a comprehensive data engineering solution using Azure services and tools.

Objective:

The primary objective of this project is to streamline the end-to-end data lifecycle, from data ingestion to visualization, with a focus on efficiency, scalability, and data quality. To accomplish this, you are required to create a robust data pipeline that begins with the acquisition of external data from GitHub via HTTPS requests, facilitated by Azure Data Factory (ADF). Once acquired, the raw data is to be stored in Azure Data Lake Storage (ADLS), providing a centralized repository for easy access and management.

Subsequently, the project mandates the utilization of Azure Databricks, a powerful data processing platform, to transform and enrich the raw data. Apache Spark and Python will be employed within Databricks to perform the necessary data cleansing, enrichment, and transformation tasks. The transformed data is then to be persistently stored in ADLS as a refined dataset, primed for further analysis.

As the project progresses, the focus will shift towards Azure Synapse Analytics, where the transformed data will serve as the foundation for extracting valuable insights and generating meaningful inferences. Azure Synapse Analytics offers the capacity to perform complex analytical operations, enabling data-driven decision-making at an enterprise level.

Finally, to democratize access to these insights, the project concludes with the integration of Azure Synapse Analytics with Power BI. Power BI will serve as the visualization layer, empowering stakeholders across the organization to interact with and derive actionable insights from the data.

In summary, this project encapsulates the end-to-end journey of data, from acquisition to visualization, within an Azure environment. Your role as a DataEngineer is pivotal in ensuring the successful execution of this project, resulting in a transformative impact on the company's decision-making processes and its ability to gain a competitive advantage in the market.

Dataset Description: The Dataset to be used here is inside the New York City Job Dataset.zip file. Extract the zip file to get all the datasets.

The zip file contains the following dataset files: - NYC_Jobs.csv

The New York City Job Dataset presents a rich tapestry of employment opportunities within the city's public sector. It encapsulates an array of roles, qualifications, and application processes, providing a comprehensive view of the diverse careers available to job seekers in the bustling metropolis of New York City.

There are a total of 30 columns in the dataset and they are listed below:

Job ID: Unique identifier for each job

posting Agency: The government department or organization offering the job

Posting Type: Indicates whether the job posting is internal or external

Of Positions: Number of positions available for the specific job Business Title: Job title used within the organization

Civil Service Title: Official civil service designation for the role.

Title Classification: Classification level of the job title Title Code No: Numeric code associated with the job title

Level: Job level within the organization Job Category: Category of work the job falls under

Full-Time/Part-Time indicator: Specifies if the job is full-time or part-time Career Level: Experience level required for the position

Salary Range From: Minimum salary offered for the job

Salary Range To: Maximum salary offered for the job

Salary Frequency: Indicates if the salary is provided on an annual basis

Work Location: Address where the job is based

Division/Work Unit: Division/Work Unit adds an additional layer of context to employment opportunities within the city's public sector

Job Description: Detailed description of the job responsibilities

Minimum Qual Requirements: Minimum qualifications required for the jobPreferred

Skills: Skills preferred but not mandatory for the job

Additional Information: Any additional information about the job To Apply: Instructions on how to apply for the position

Hours/Shift: Working hours and shift details

Work Location 1: Additional work location information if applicable Recruitment

Contact: Contact person for recruitment-related inquiries

Residency Requirement: Information about residency requirements Posting Date: Date when the job was posted

Post Until: Deadline for application submission

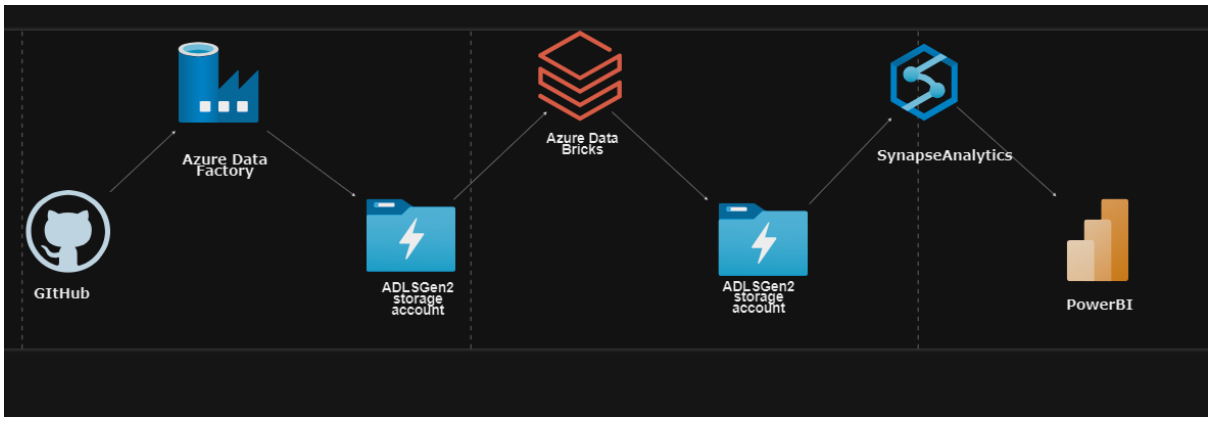
Posting Updated: Date when the job posting was last updated Process Date: Date when the hiring process was initiated

Tasks to be Completed:

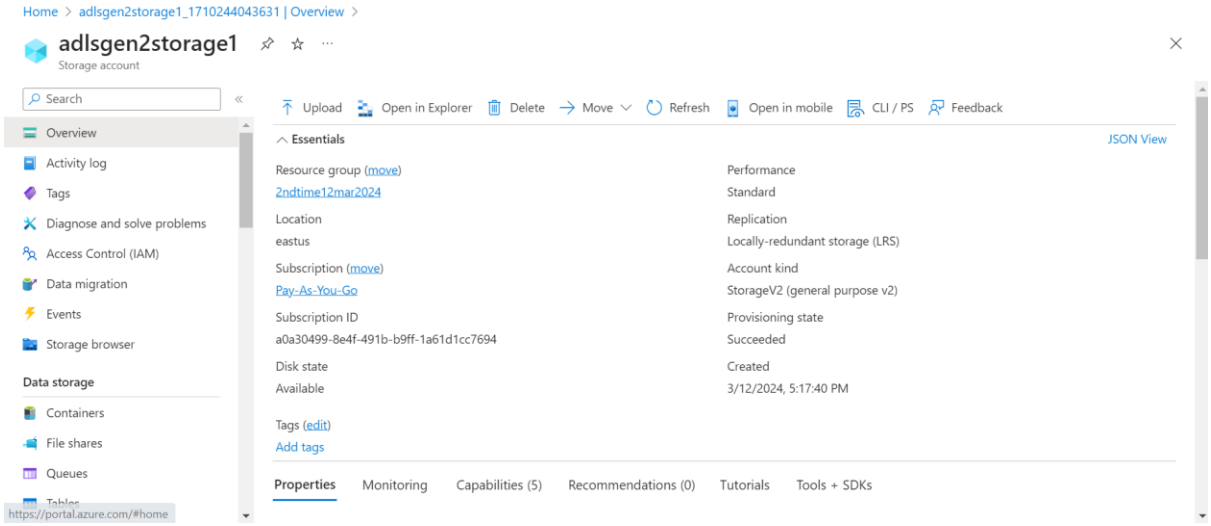
1. Create a diagram illustrating the complete data pipeline you would design and implement for this project. It should showcase the flow of data from its source on GitHub, through Azure Data Factory, Azure Data Lake Storage, Azure Databricks, Azure Synapse Analytics, and finally to Power BI for visualization. Please include the major components and their interactions in your diagram to provide a clear overview of the end-to-end data journey.

2. Execute each phase of the pipeline, capturing comprehensive screenshots at every step, and compile a comprehensive report that encapsulates the inferences and achievements of the project.

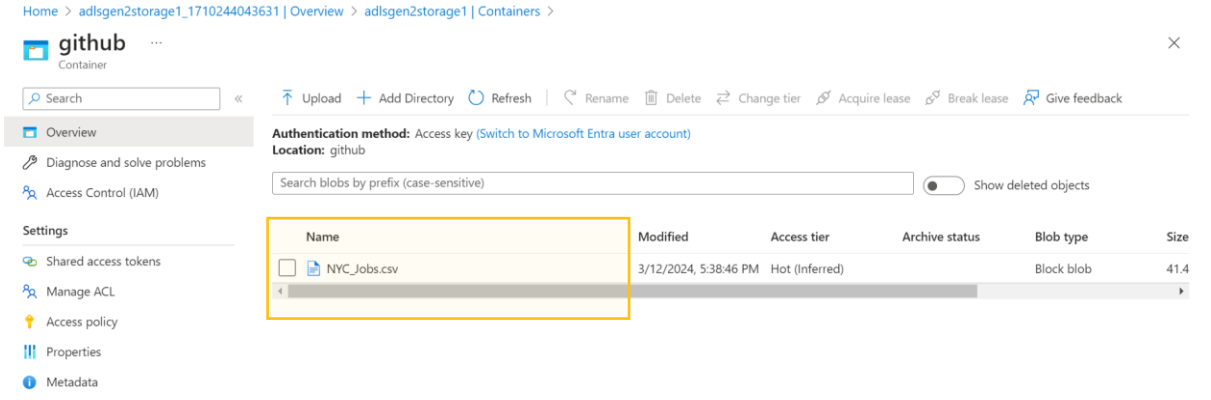
Diagram illustrating the complete data pipeline



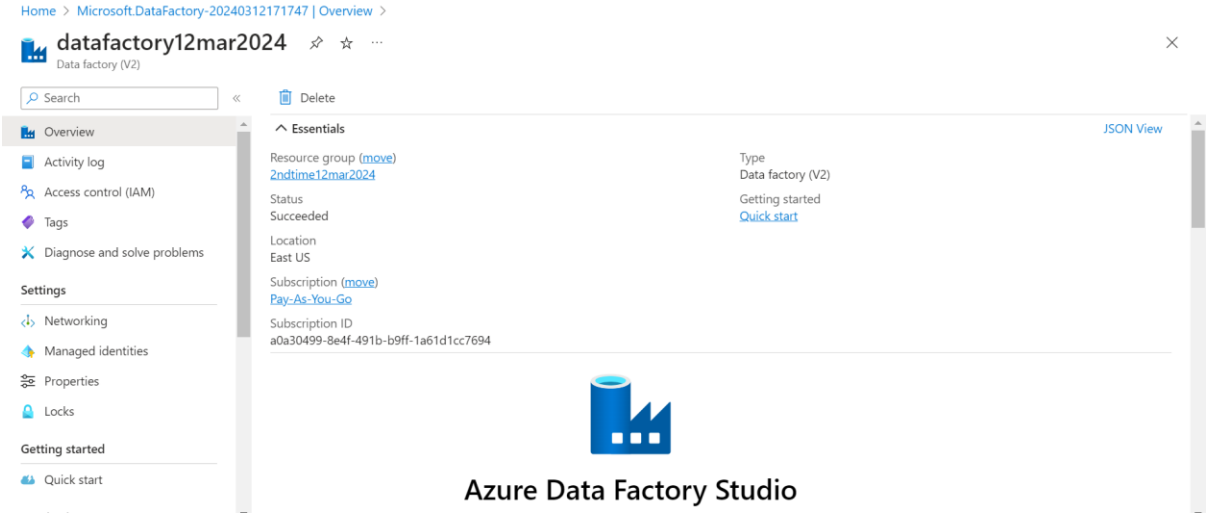
A Gen2 storage account is created



The dataset is uploaded to a container named github

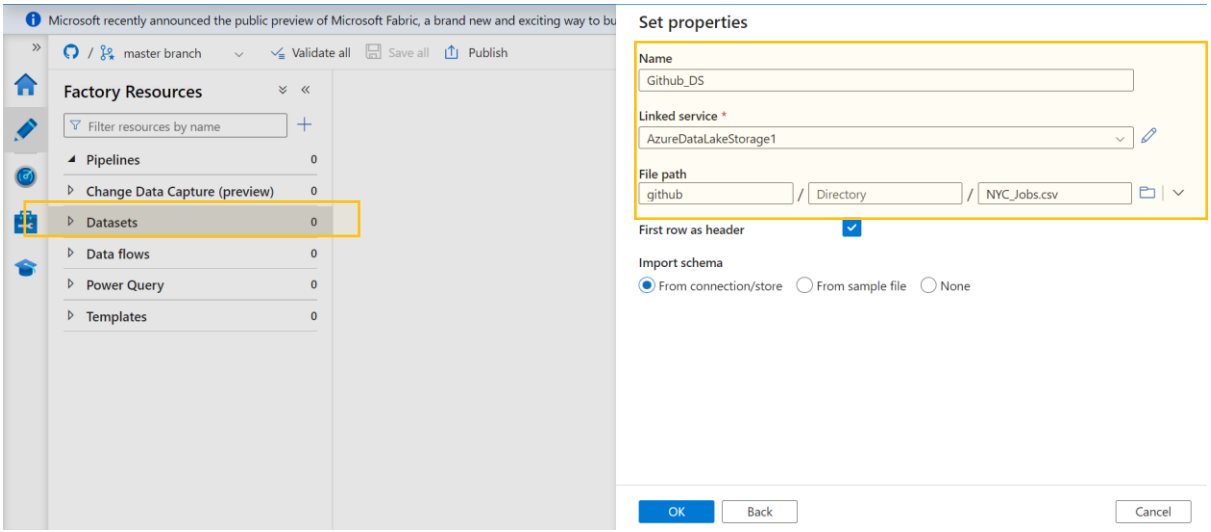


A data factory resource is created

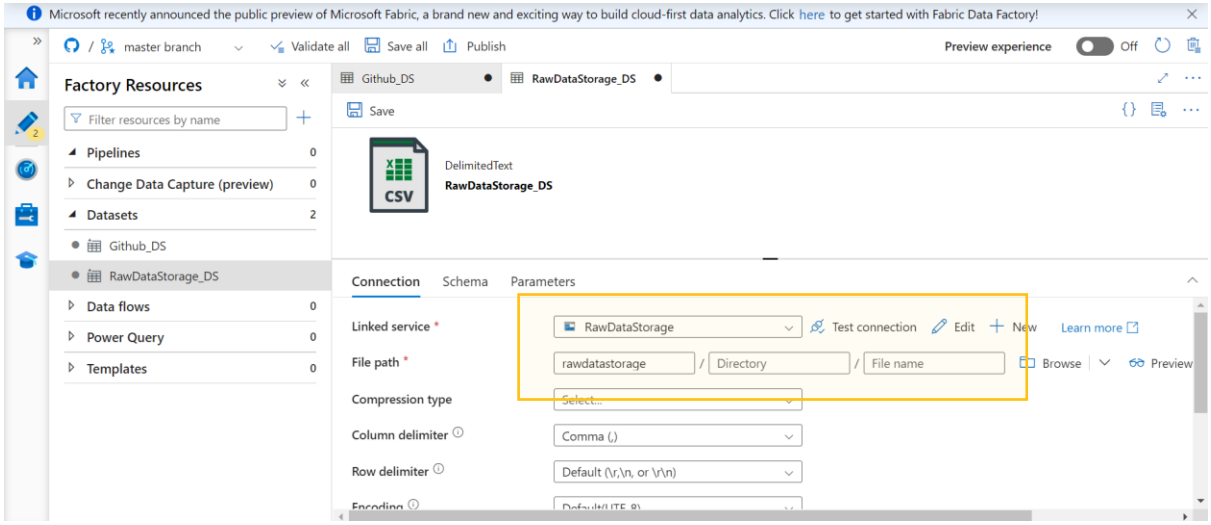


Then linked services are created to the gen2 storage account

After that dataset is creted for the github container

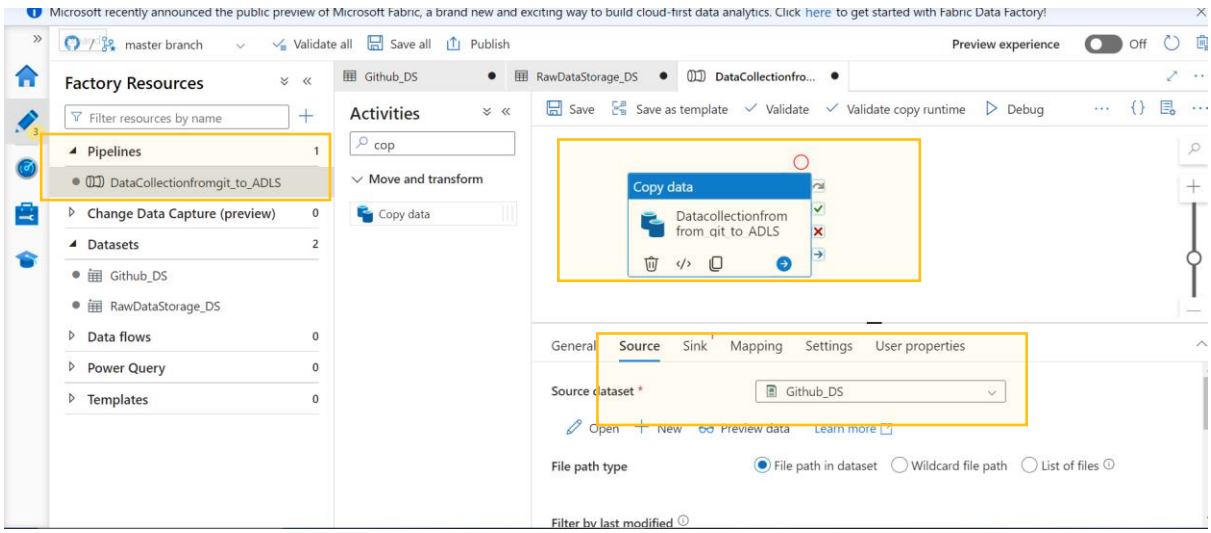


Similarly another dataset is created to the the Raw storage container

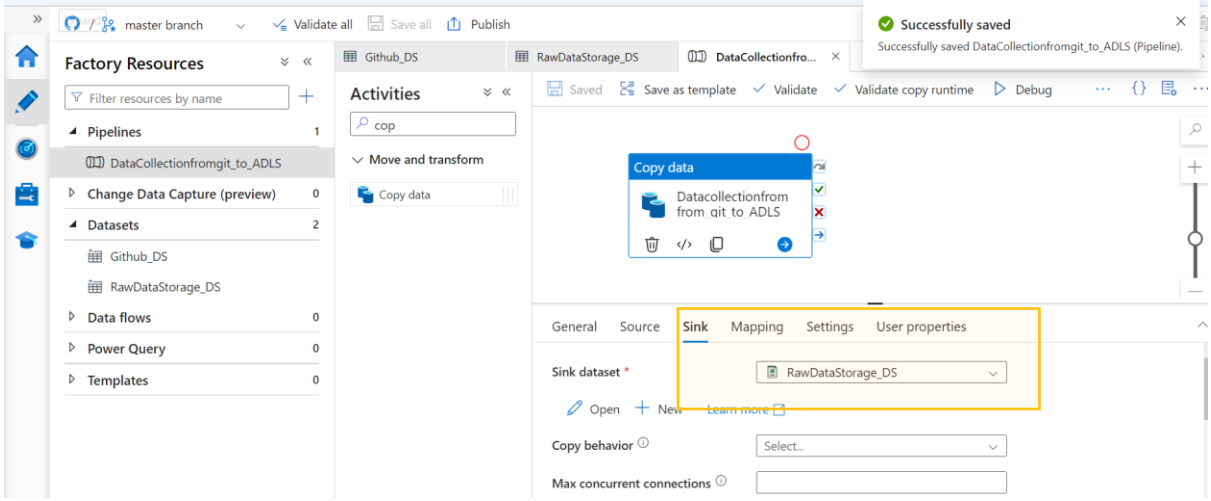


Then a pieline is created with copy data activity

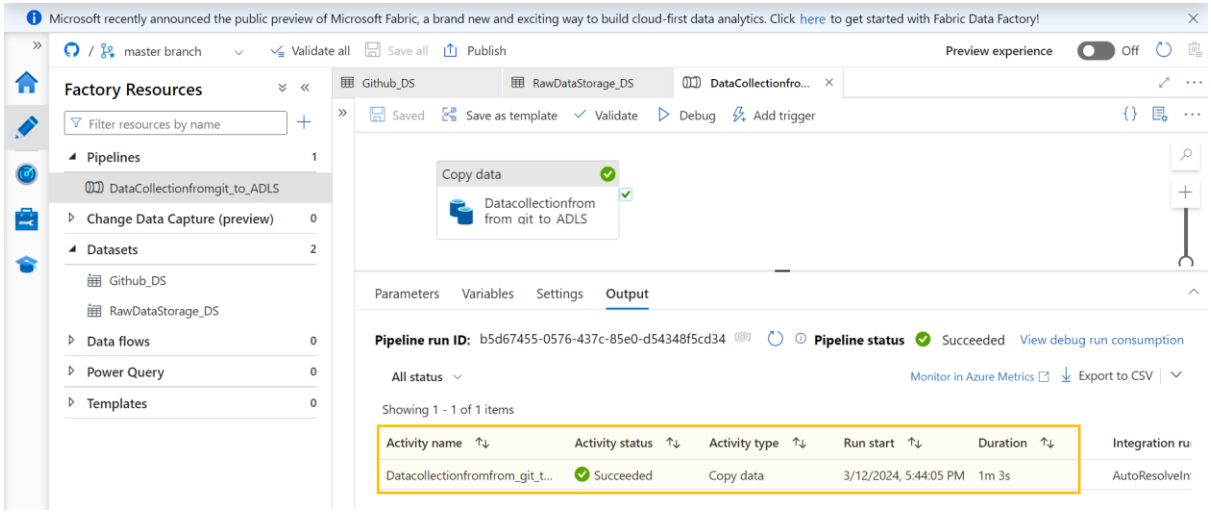
Source is provided as github_dataset



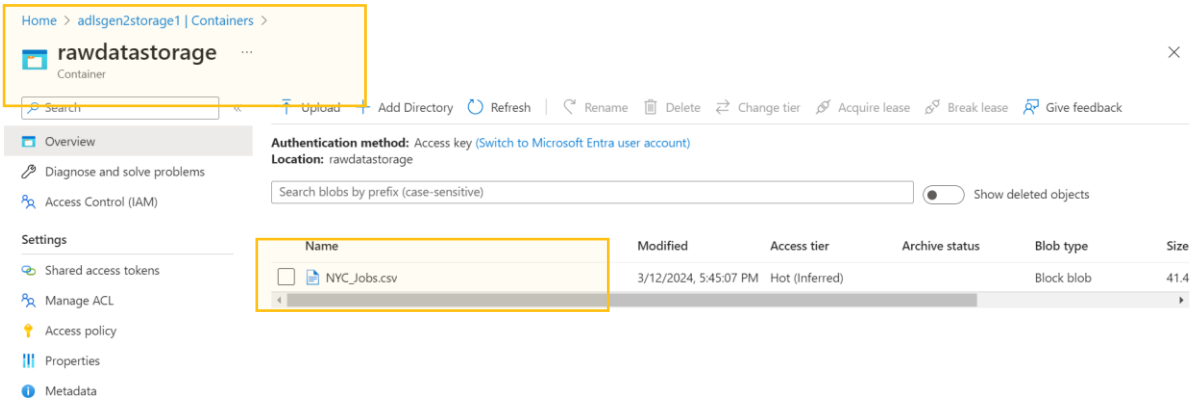
Sink is provided as Rawstorage dataset



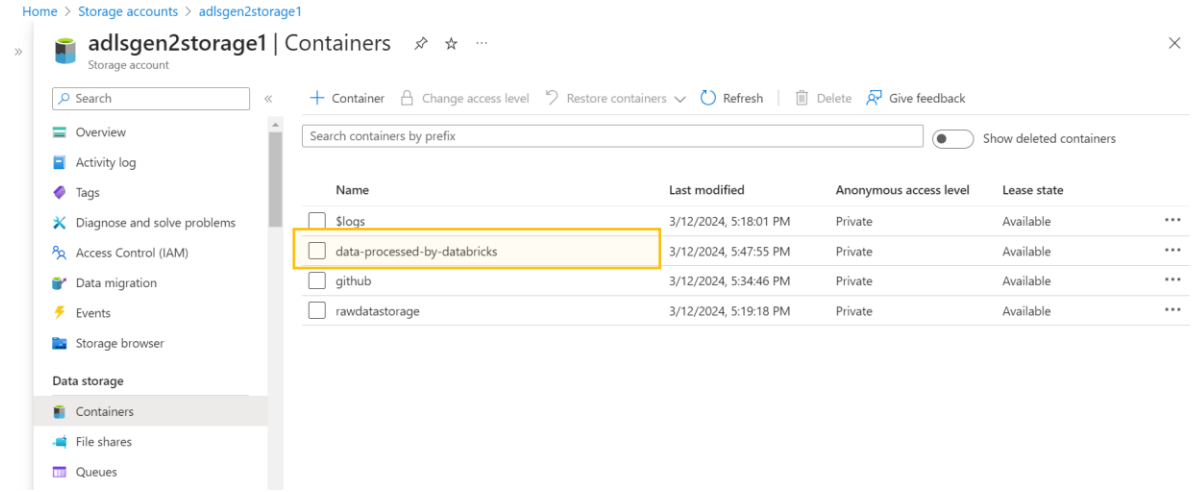
Then the changes are published and pipeline is executed



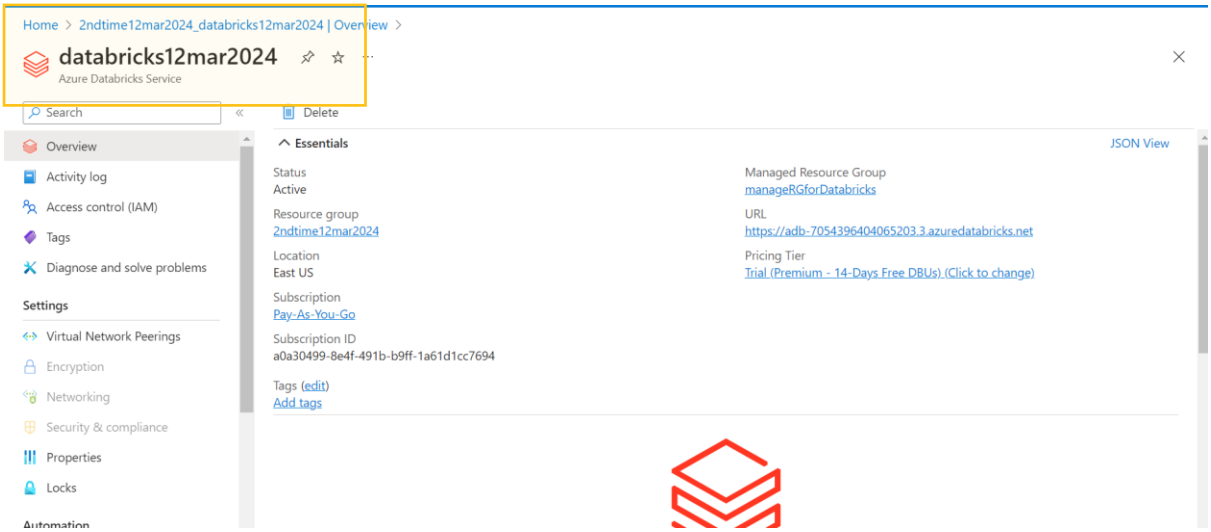
The dataset is successfully copied to the ADLS account



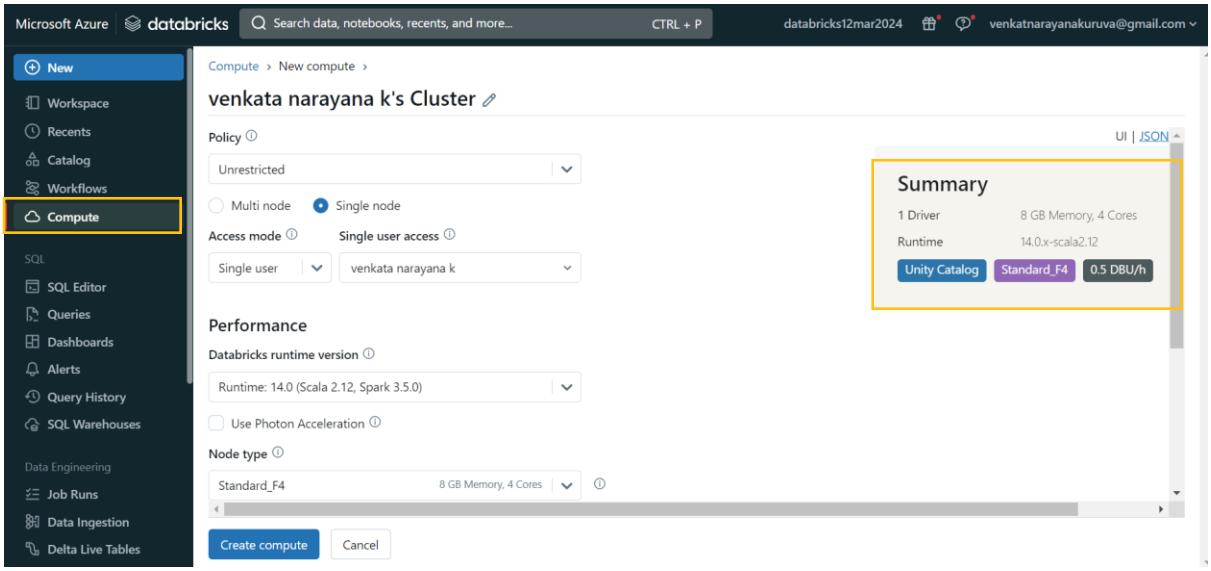
Then another container is created “data-processed-by-databricks” for databtricks output files



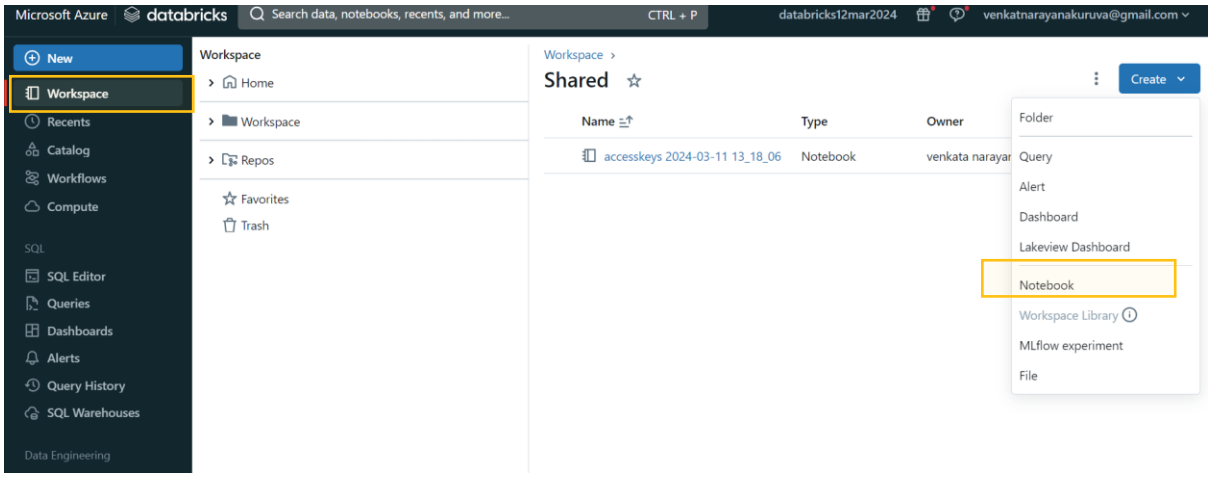
Databricks resoutrce is created



Databricks resource is launched and compute cluster is created



Then navigated to the workspace and created a notebook



The container having the dataset is mounted to databricks

```
06:00 PM (14s)                                                    1
dbutils.fs.mount(
    source = "wasbs://rawdatastorage@adlsgen2storage1.blob.core.windows.net",
    mount_point = "/mnt/rawdatastorage",
    extra_configs = {"fs.azure.account.key.adlsgen2storage1.blob.core.windows.
    net":"smokFhI2kPy6PbQ0R2OruePSIOJhDGbuDMrMDXz5RIIQcSiIDFLdhzmUyQDekRpVK0dHlE2/zRNO+AStRRf/FQ=="}
)
```

```
True
```

```
4 minutes ago (1s)                                                2
dbutils.fs.ls("/mnt/rawdatastorage")
```

```
[FileInfo(path='dbfs:/mnt/rawdatastorage/NYC_Jobs.csv', name='NYC_Jobs.csv', size=43419510, modificationTime=171024570700
0)]
```

Then the container created for the outputs of the databricks is mounted to data bricks

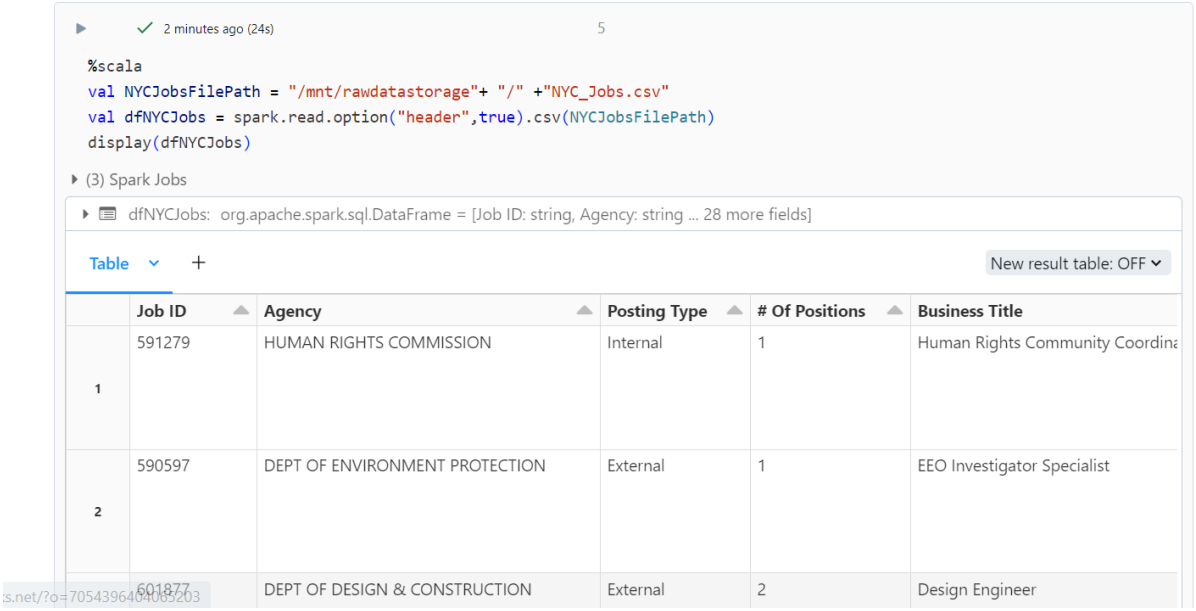
```
4 minutes ago (11s)                                               3
dbutils.fs.mount(
    source = "wasbs://data-processed-by-databricks@adlsgen2storage1.blob.core.windows.net",
    mount_point = "/mnt/data-processed-by-databricks",
    extra_configs = {"fs.azure.account.key.adlsgen2storage1.blob.core.windows.
    net":"smokFhI2kPy6PbQ0R2OruePSIOJhDGbuDMrMDXz5RIIQcSiIDFLdhzmUyQDekRpVK0dHlE2/zRNO+AStRRf/FQ=="}
)
```

```
True
```

```
4 minutes ago (<1s)                                               4
dbutils.fs.ls("/mnt/data-processed-by-databricks")
```

```
[]
```

after that a dataframe is created out of the dataset



2 minutes ago (24s) 5

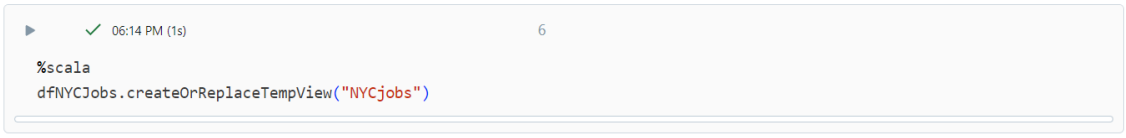
```
%scala
val NYCJobsFilePath = "/mnt/rawdatastorage"+ "/" +"NYC_Jobs.csv"
val dfNYCJobs = spark.read.option("header",true).csv(NYCJobsFilePath)
display(dfNYCJobs)
```

(3) Spark Jobs

dfNYCJobs: org.apache.spark.sql.DataFrame = [Job ID: string, Agency: string ... 28 more fields]

Table + New result table: OFF

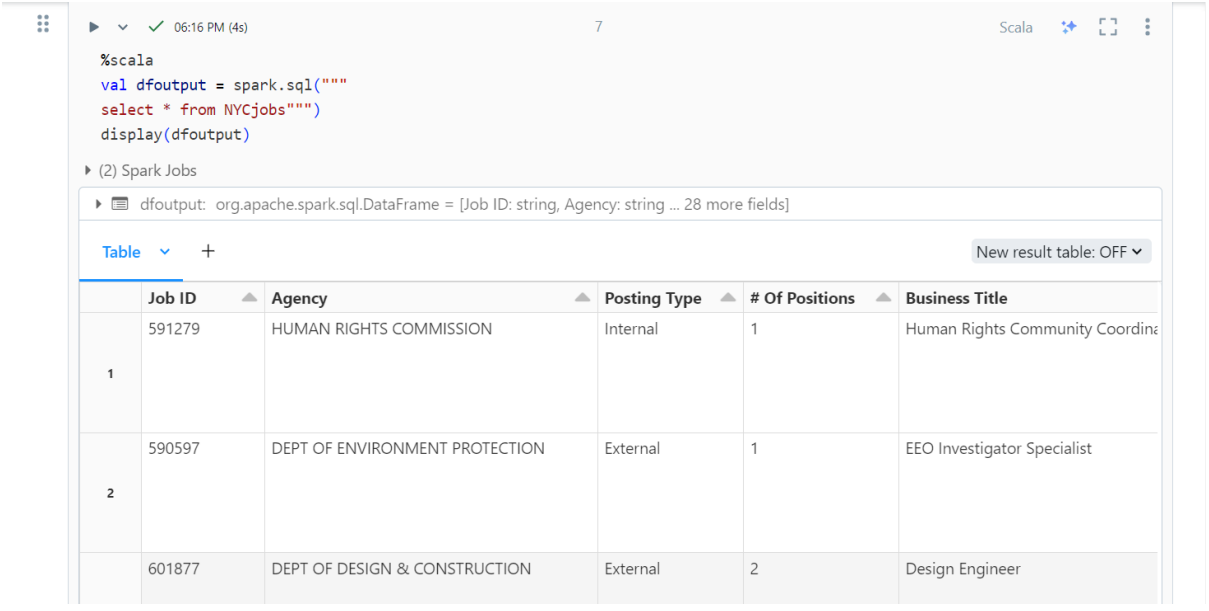
	Job ID	Agency	Posting Type	# Of Positions	Business Title
1	591279	HUMAN RIGHTS COMMISSION	Internal	1	Human Rights Community Coordina
2	590597	DEPT OF ENVIRONMENT PROTECTION	External	1	EEO Investigator Specialist
	601877	DEPT OF DESIGN & CONSTRUCTION	External	2	Design Engineer



06:14 PM (1s) 6

```
%scala
dfNYCJobs.createOrReplaceTempView("NYCjobs")
```

Then a varible named output is created with a processed dataset



06:16 PM (4s) 7 Scala

```
%scala
val dfoutput = spark.sql("""
select * from NYCjobs""")
display(dfoutput)
```

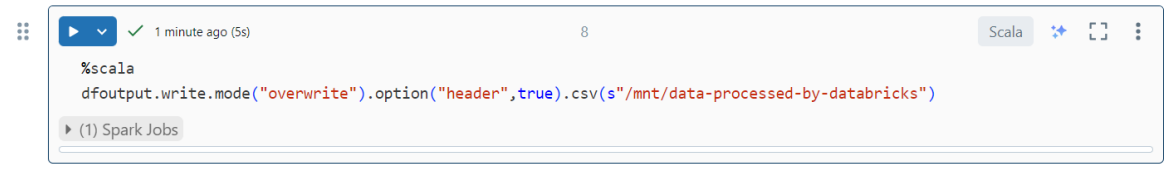
(2) Spark Jobs

dfoutput: org.apache.spark.sql.DataFrame = [Job ID: string, Agency: string ... 28 more fields]

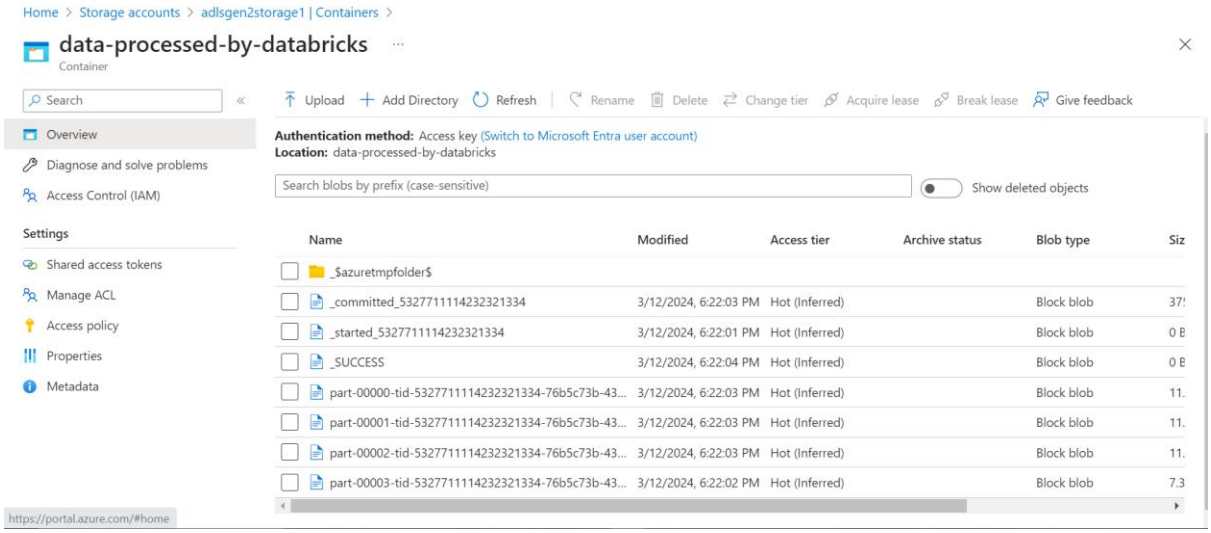
Table + New result table: OFF

	Job ID	Agency	Posting Type	# Of Positions	Business Title
1	591279	HUMAN RIGHTS COMMISSION	Internal	1	Human Rights Community Coordina
2	590597	DEPT OF ENVIRONMENT PROTECTION	External	1	EEO Investigator Specialist
	601877	DEPT OF DESIGN & CONSTRUCTION	External	2	Design Engineer

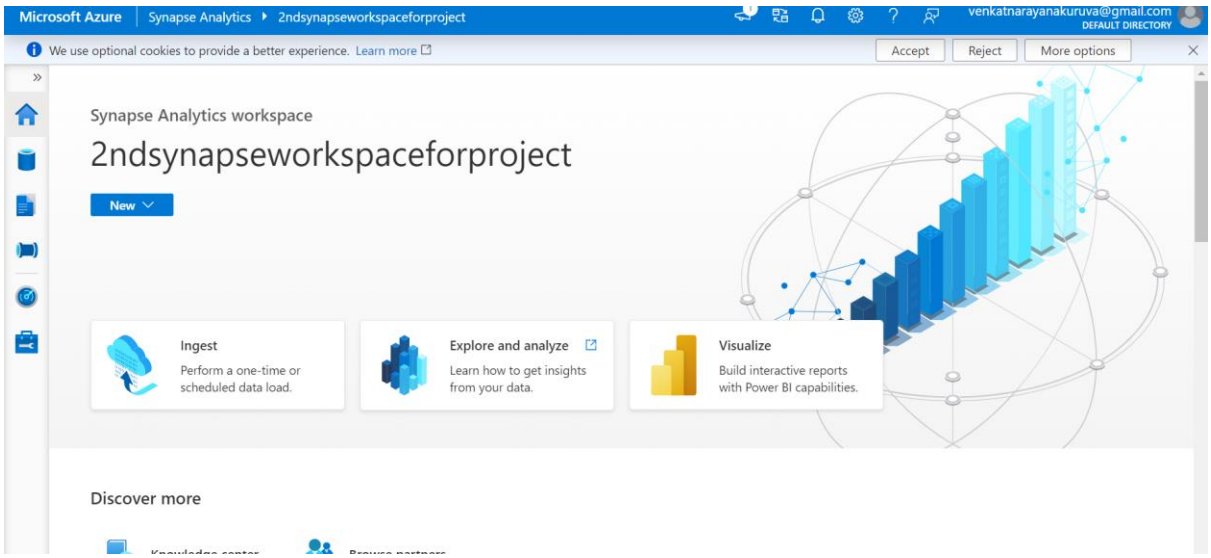
Finally the processed dataset is stored to the container “data-processed-by-databricks”



We can check the output files stored in the container by databricks

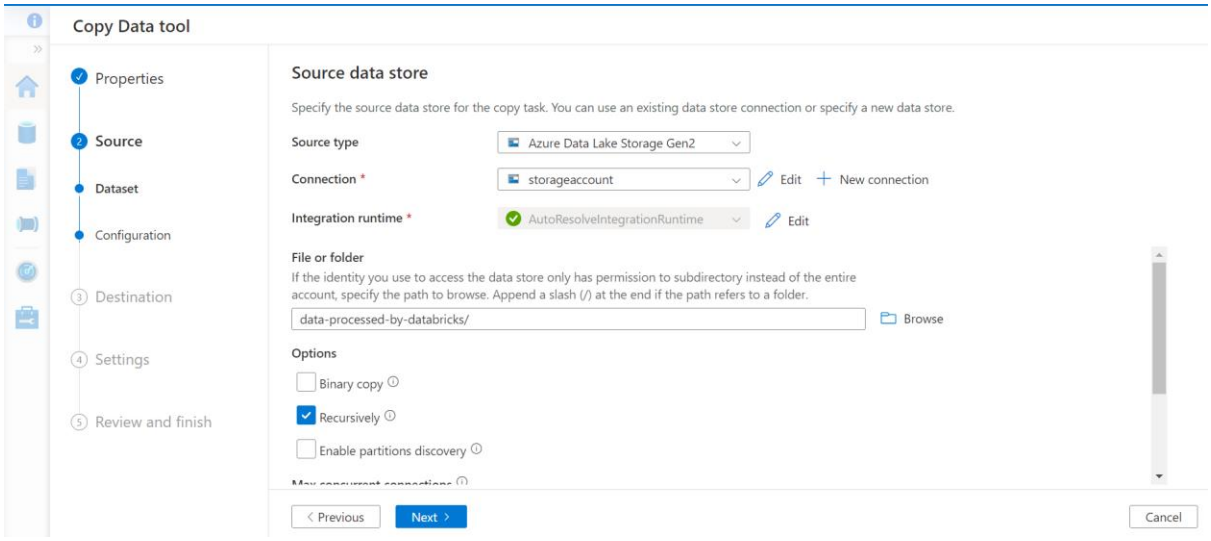


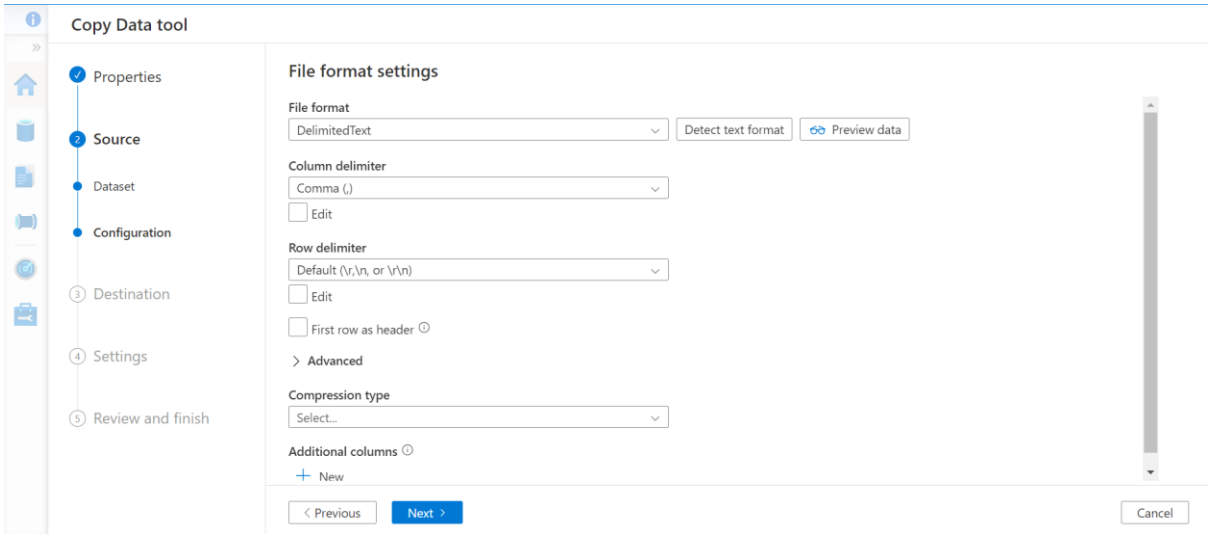
Then a synapse analytics resource is created



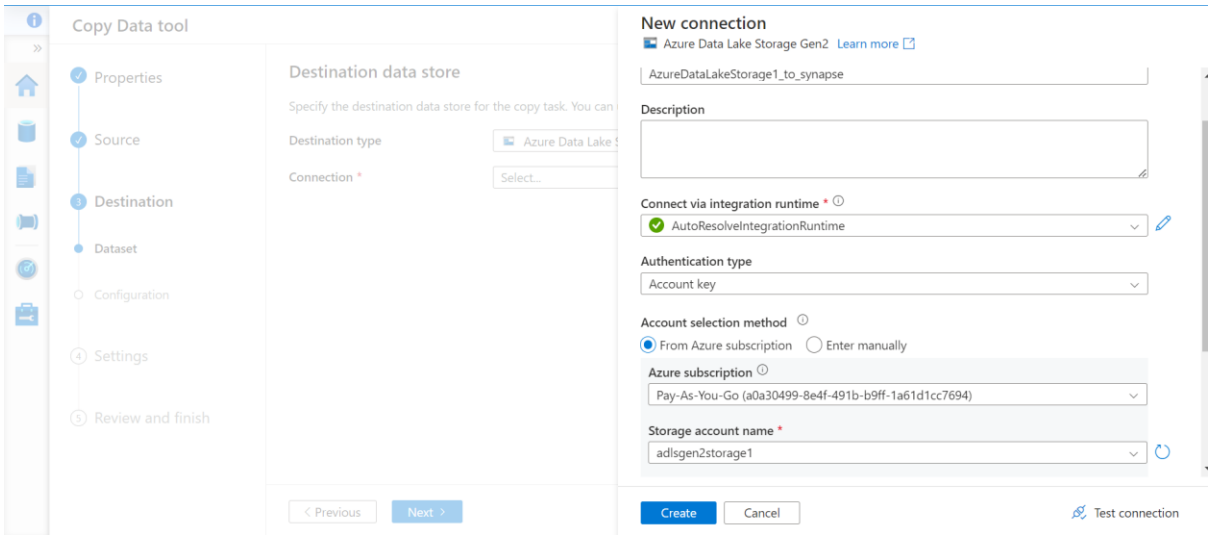
The data produced by the databricks is ingested into the synapse analytics

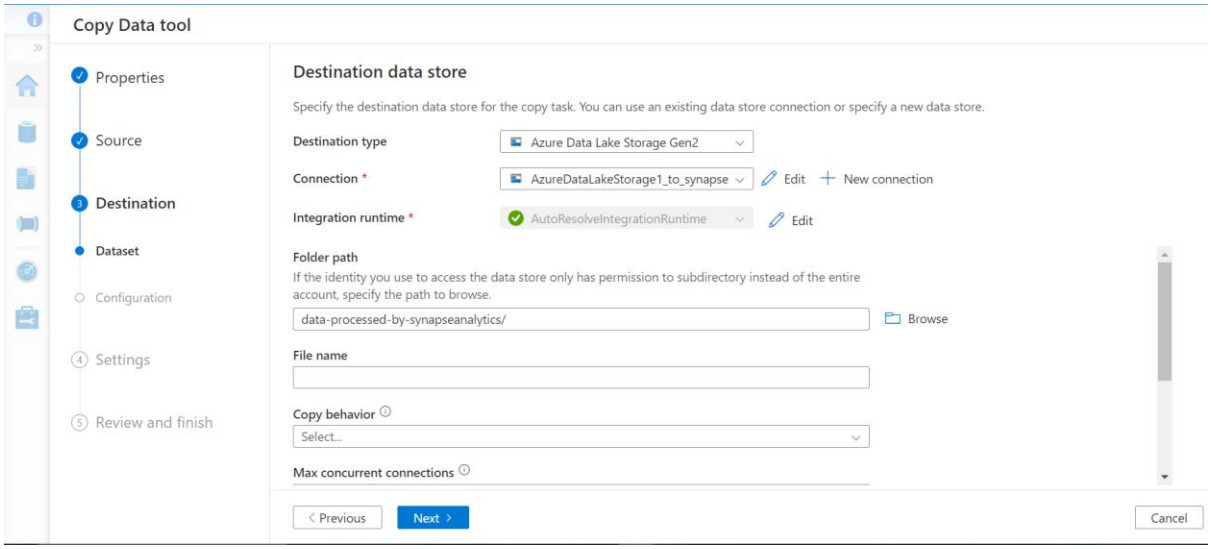
Source is output of the databricks



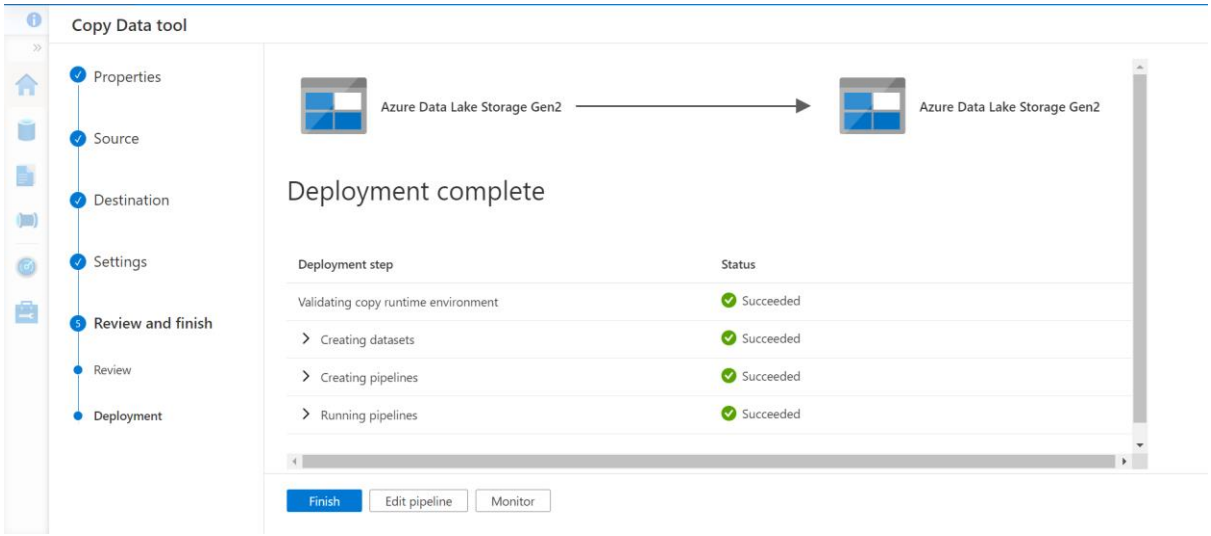


Destination is synapse's defualt storage





The data is successfully ingested into the synapse analytics for the further use



At last processed insights are visualised through the PowerBI