

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 Data Engineer 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

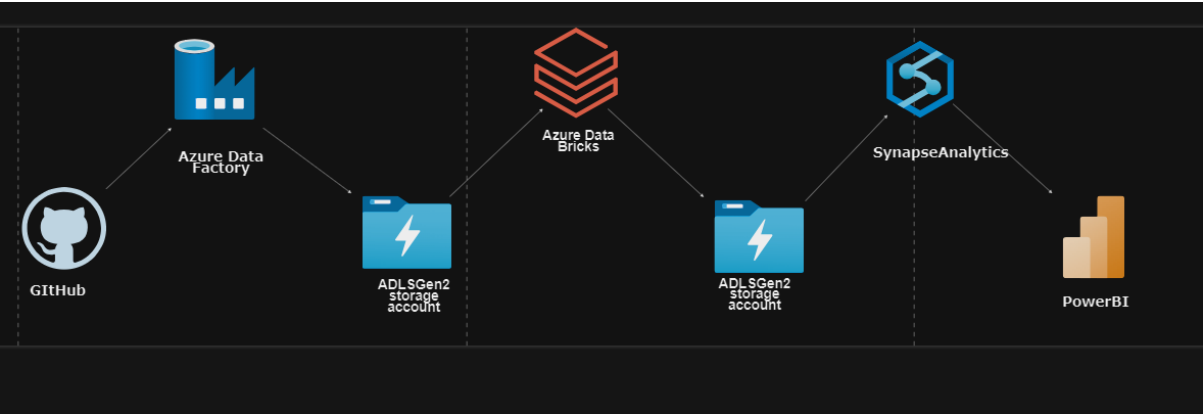
Post Until: Deadline for application submission

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

Home > adlsgen2storage1_1710244043631 | Overview >

adlsgen2storage1 Storage account

Search

Upload Open in Explorer Delete Move Refresh Open in mobile CLI / PS Feedback

Essentials

Resource group (move) [2ndtime12mar2024](#)

Location eastus

Subscription (move) [Pay-As-You-Go](#)

Subscription ID a0a30499-8e4f-491b-b9ff-1a61d1cc7694

Disk state Available

Tags (edit) [Add tags](#)

Performance Standard

Replication Locally-redundant storage (LRS)

Account kind StorageV2 (general purpose v2)

Provisioning state Succeeded

Created 3/12/2024, 5:17:40 PM

Properties Monitoring Capabilities (5) Recommendations (0) Tutorials Tools + SDKs

[JSON View](#)

The dataset is uploaded to a container named github

Home > adlsgen2storage1_1710244043631 | Overview > adlsgen2storage1 | Containers >

github Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

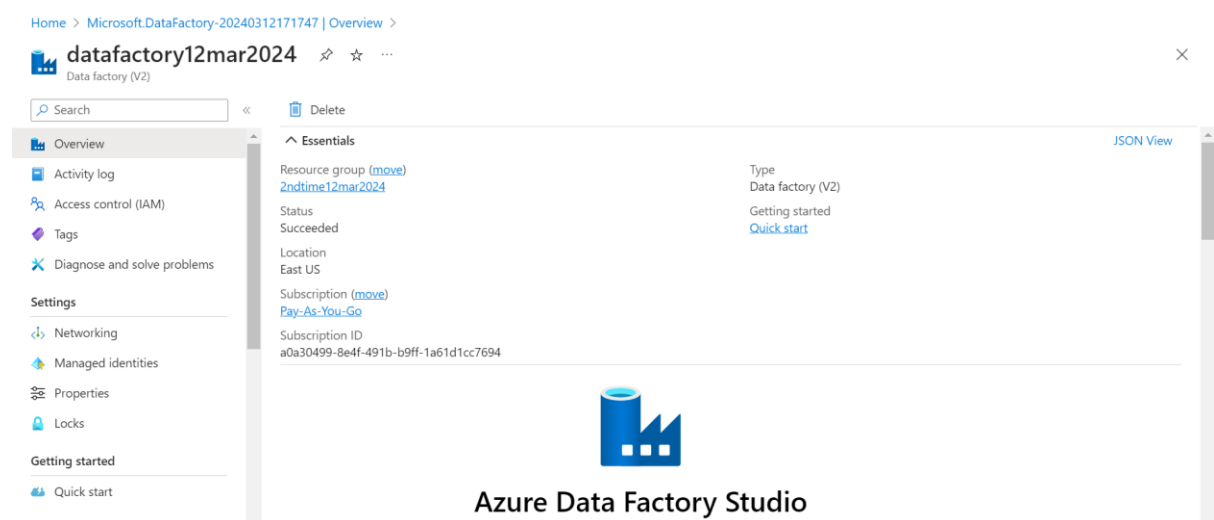
Authentication method: Access key ([Switch to Microsoft Entra user account](#))

Location: github

Search blobs by prefix (case-sensitive) ☐ Show deleted objects

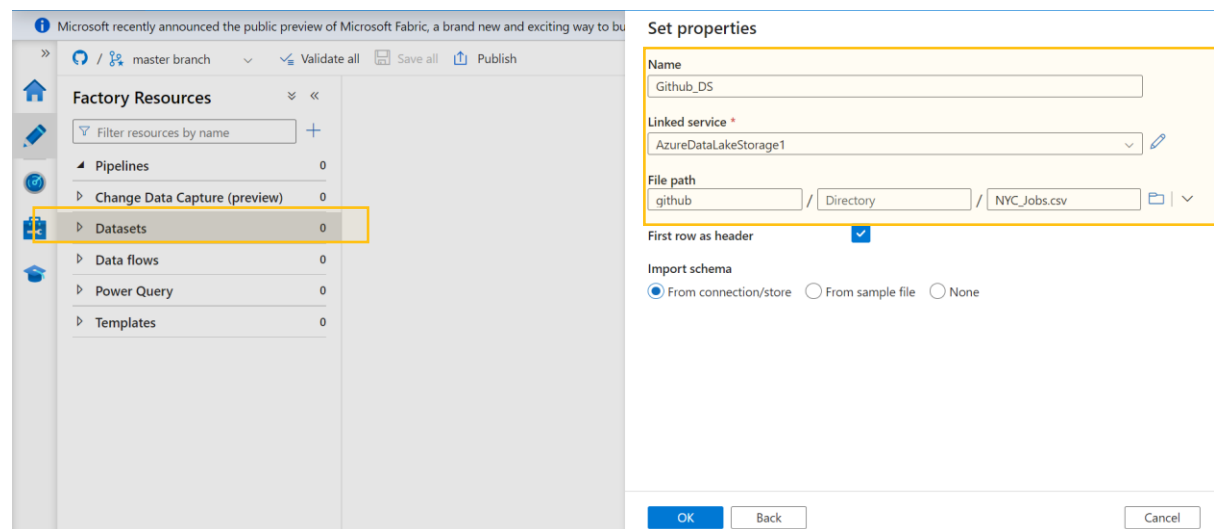
Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> NYC_Jobs.csv	3/12/2024, 5:38:46 PM	Hot (Inferred)		Block blob	41.4

A data factory resource is created

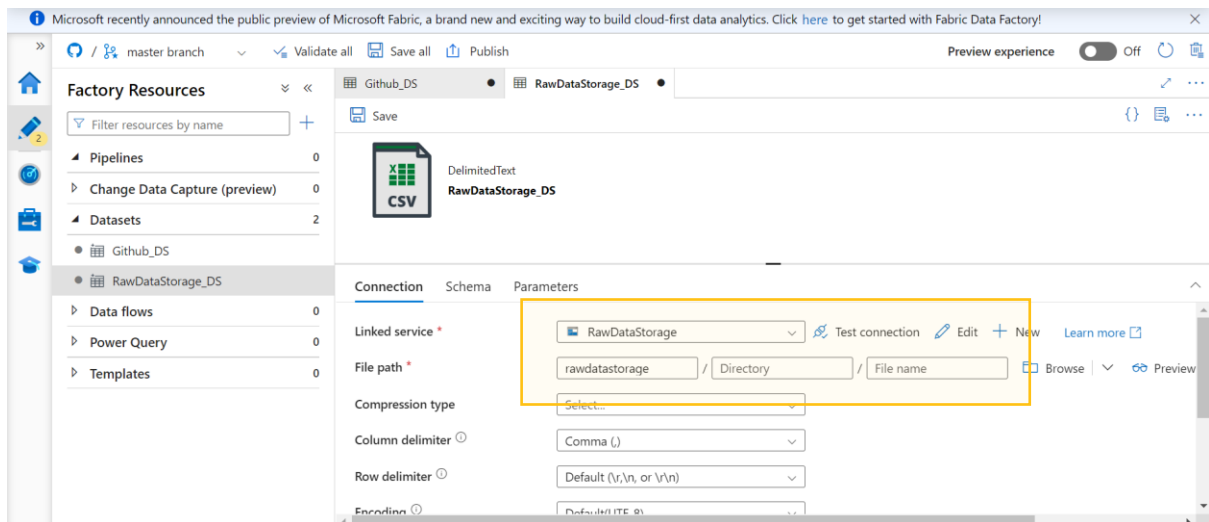


Then linked services are created to the gen2 storage account

After that dataset is created for the github container

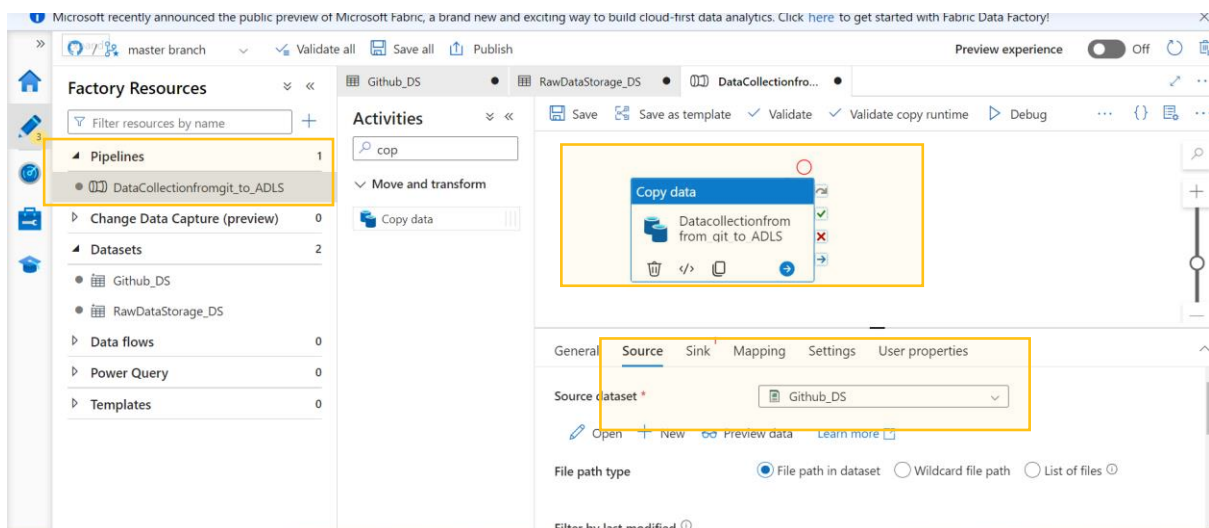


Similarly another dataset is created to the the Raw storage container

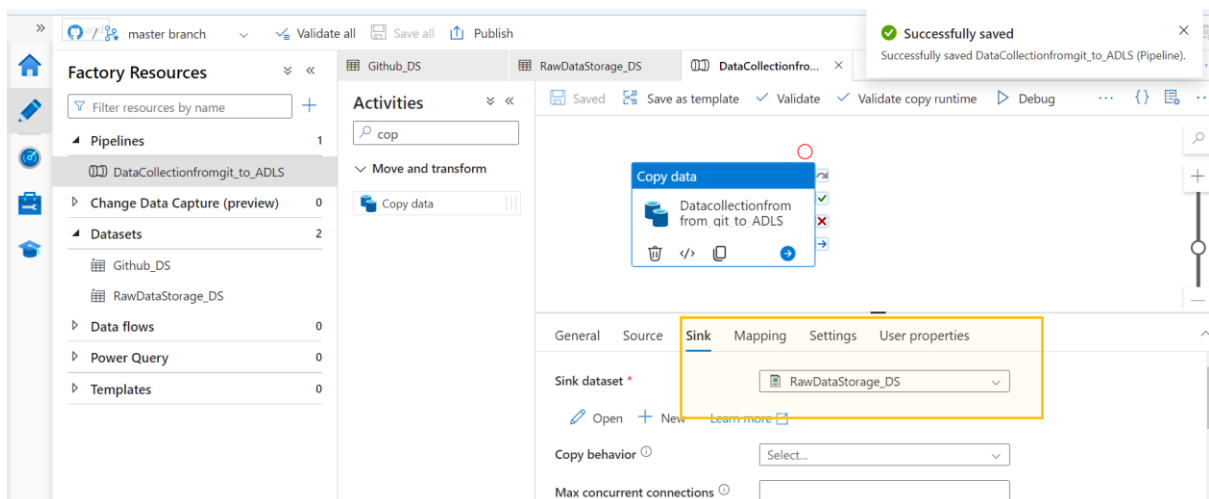


Then a pipeline is created with copy data activity

Source is provided as github_dataset



Sink is provided as Rawstorage dataset



Then the changes are published and pipeline is executed

The screenshot shows the Microsoft Fabric Data Factory interface. On the left, the 'Factory Resources' pane lists 'Pipelines' with 'DataCollectionfromgit_to_ADLS' selected. The main area displays the 'Copy data' activity, 'Datacollectionfrom from git to ADLS'. Below, the 'Output' tab shows the 'Pipeline run ID: b5d67455-0576-437c-85e0-d54348f5cd34' with a status of 'Succeeded'. A table lists the activity details:

Activity name	Activity status	Activity type	Run start	Duration	Integration ru
Datacollectionfromfrom_git_t...	Succeeded	Copy data	3/12/2024, 5:44:05 PM	1m 3s	AutoResolveIn

The dataset is successfully copied to the ADLS account

The screenshot shows the Azure Storage Explorer interface for the 'adlsgen2storage1' container. The 'rawdatastorage' container is selected, and the file 'NYC_Jobs.csv' is highlighted in the list. The table below shows the file details:

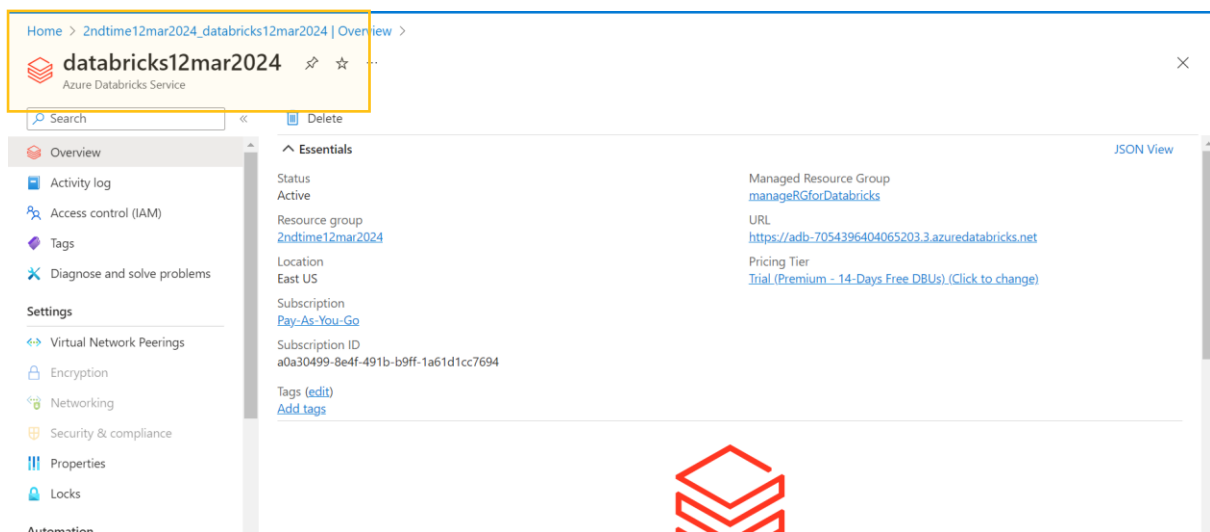
Name	Modified	Access tier	Archive status	Blob type	Size
NYC_Jobs.csv	3/12/2024, 5:45:07 PM	Hot (Inferred)		Block blob	41.4

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

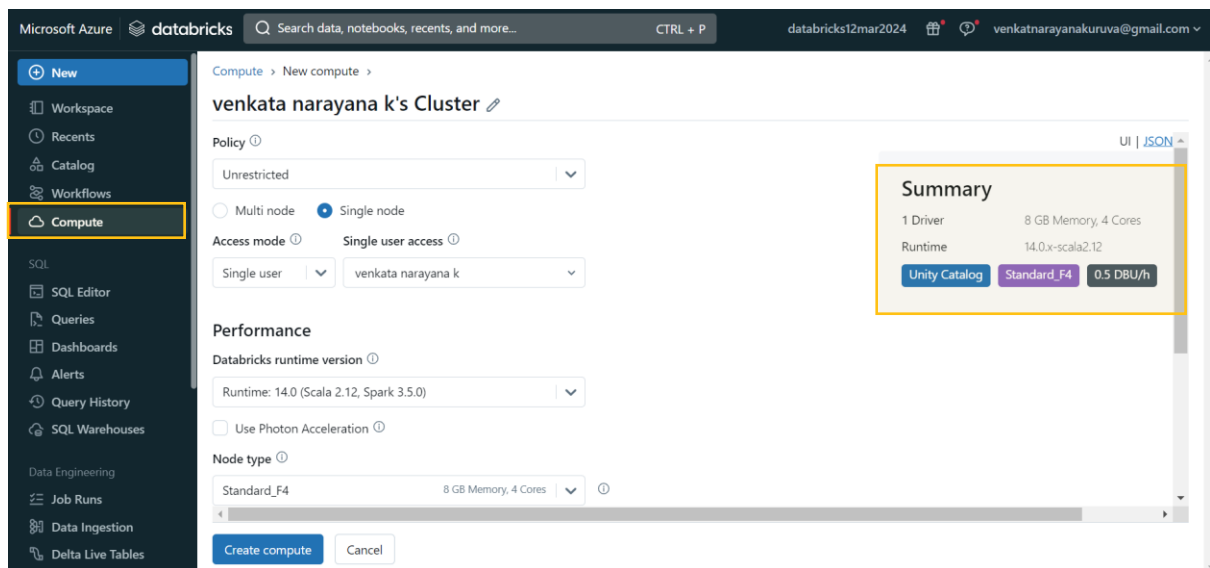
The screenshot shows the Azure Storage Explorer interface for the 'adlsgen2storage1' container. The 'data-processed-by-databricks' container is highlighted in the list. The table below shows the container details:

Name	Last modified	Anonymous access level	Lease state
\$logs	3/12/2024, 5:18:01 PM	Private	Available
data-processed-by-databricks	3/12/2024, 5:47:55 PM	Private	Available
github	3/12/2024, 5:34:46 PM	Private	Available
rawdatastorage	3/12/2024, 5:19:18 PM	Private	Available

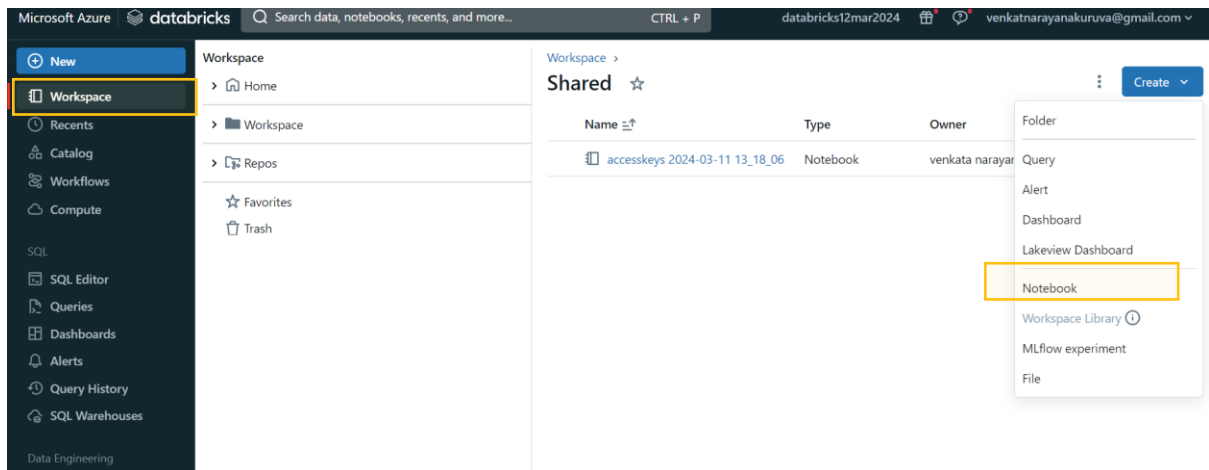
Databricks resource is created



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": "smokFhI2kPy6PbQ0R2OruePSIOJhDGBuDMrMDXz5RIIQcSiIDFLdhzmUyQDekRpVK0dH1E2/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=1710245707000)]
```

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": "smokFhI2kPy6PbQ0R2OruePSIOJhDGBuDMrMDXz5RIIQcSiIDFLdhzmUyQDekRpVK0dH1E2/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 Coordin
2	590597	DEPT OF ENVIRONMENT PROTECTION	External	1	EEO Investigator Specialist
	601877	DEPT OF DESIGN & CONSTRUCTION	External	2	Design Engineer

s.net/?o=7054396404065203

06:14 PM (1s) 6

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

Then a variable 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 Coordin
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”

1 minute ago (5s)

8

Scala

```
%scala
dfoutput.write.mode("overwrite").option("header",true).csv(s"/mnt/data-processed-by-databricks")
```

(1) Spark Jobs

[Shift+Enter] to run and move to next cell

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

Home > Storage accounts > adlsgen2storage1 | Containers >

data-processed-by-databricks

Container

Search

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: data-processed-by-databricks

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> _\$azuretmpfolder\$					
<input type="checkbox"/> _committed_5327711114232321334	3/12/2024, 6:22:03 PM	Hot (Inferred)		Block blob	37...
<input type="checkbox"/> _started_5327711114232321334	3/12/2024, 6:22:01 PM	Hot (Inferred)		Block blob	0 B
<input type="checkbox"/> _SUCCESS	3/12/2024, 6:22:04 PM	Hot (Inferred)		Block blob	0 B
<input type="checkbox"/> part-0000-tid-5327711114232321334-76b5c73b-43...	3/12/2024, 6:22:03 PM	Hot (Inferred)		Block blob	11...
<input type="checkbox"/> part-0001-tid-5327711114232321334-76b5c73b-43...	3/12/2024, 6:22:03 PM	Hot (Inferred)		Block blob	11...
<input type="checkbox"/> part-0002-tid-5327711114232321334-76b5c73b-43...	3/12/2024, 6:22:03 PM	Hot (Inferred)		Block blob	11...
<input type="checkbox"/> part-0003-tid-5327711114232321334-76b5c73b-43...	3/12/2024, 6:22:02 PM	Hot (Inferred)		Block blob	7.3...

https://portal.azure.com/#home

Then a synapse analytics resource is created

Microsoft Azure

Synapse Analytics

2ndsynapseworkspaceforproject

venkatnaranakuruva@gmail.com

DEFAULT DIRECTORY

We use optional cookies to provide a better experience. Learn more

Accept

Reject

More options

Synapse Analytics workspace

2ndsynapseworkspaceforproject

New

Ingest

Perform a one-time or scheduled data load.

Explore and analyze

Learn how to get insights from your data.

Visualize

Build interactive reports with Power BI capabilities.

Discover more

Knowledge center

Support center

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

Source is output of the databricks

The screenshot shows the 'Copy Data tool' interface with the 'Source' step selected in the left-hand navigation pane. The main panel is titled 'Source data store' and contains the following fields:

- Source type:** A dropdown menu set to 'Azure Data Lake Storage Gen2'.
- Connection:** A dropdown menu set to 'storageaccount', with an 'Edit' link and a '+ New connection' button.
- Integration runtime:** A dropdown menu set to 'AutoResolveIntegrationRuntime', with an 'Edit' link.
- File or folder:** A text input field containing 'data-processed-by-databricks/' and a 'Browse' button.
- Options:** A group of checkboxes: 'Binary copy' (unchecked), 'Recursively' (checked), and 'Enable partitions discovery' (unchecked).

At the bottom, there are navigation buttons: '< Previous', 'Next >', and 'Cancel'.

The screenshot shows the 'Copy Data tool' interface with the 'File format settings' step selected in the left-hand navigation pane. The main panel contains the following fields:

- File format:** A dropdown menu set to 'DelimitedText', with 'Detect text format' and 'Preview data' buttons.
- Column delimiter:** A dropdown menu set to 'Comma (,)', with an 'Edit' link.
- Row delimiter:** A dropdown menu set to 'Default (\r,\n, or \r\n)', with an 'Edit' link.
- First row as header:** An unchecked checkbox.
- Advanced:** A section header with a right-pointing arrow.
- Compression type:** A dropdown menu set to 'Select...'.
- Additional columns:** A section with a '+ New' button.

At the bottom, there are navigation buttons: '< Previous', 'Next >', and 'Cancel'.

Destination is synapse's default storage

The screenshot shows the 'Copy Data tool' interface with the 'Destination' step selected in the left-hand navigation pane. The main panel is titled 'Destination data store' and contains the following fields:

- Destination type:** A dropdown menu set to 'Azure Data Lake Storage Gen2'.
- Connection:** A dropdown menu set to 'Select...'.

On the right side, there is a 'New connection' panel with the following fields:

- Connection name:** A text input field containing 'AzureDataLakeStorage1_to_synapse'.
- Description:** A text input field.
- Connect via integration runtime:** A dropdown menu set to 'AutoResolveIntegrationRuntime'.
- Authentication type:** A dropdown menu set to 'Account key'.
- Account selection method:** Radio buttons for 'From Azure subscription' (selected) and 'Enter manually'.
- Azure subscription:** A dropdown menu set to 'Pay-As-You-Go (a0a30499-8e4f-491b-b9ff-1a61d1cc7694)'.
- Storage account name:** A dropdown menu set to 'adlsgen2storage1'.

At the bottom, there are navigation buttons: '< Previous', 'Next >', 'Create', 'Cancel', and a 'Test connection' button.

Copy Data tool

1 Properties

2 Source

3 Destination

4 Dataset

5 Configuration

6 Settings

7 Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type Azure Data Lake Storage Gen2

Connection * AzureDataLakeStorage1_to_synapse Edit + New connection

Integration runtime * AutoResolveIntegrationRuntime Edit

Folder path

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

data-processed-by-synapseanalytics/ Browse

File name

Copy behavior Select...

Max concurrent connections

< Previous Next > Cancel

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

Copy Data tool

1 Properties

2 Source

3 Destination

4 Settings

5 Review and finish

6 Review

7 Deployment

Azure Data Lake Storage Gen2 → Azure Data Lake Storage Gen2

Deployment complete

Deployment step	Status
Validating copy runtime environment	✓ Succeeded
> Creating datasets	✓ Succeeded
> Creating pipelines	✓ Succeeded
> Running pipelines	✓ Succeeded

Finish Edit pipeline Monitor

At last processed insights are visualised through the PowerBI