Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 21-02-24

CODING ASSESSMENT – AZURE DATABRICKS

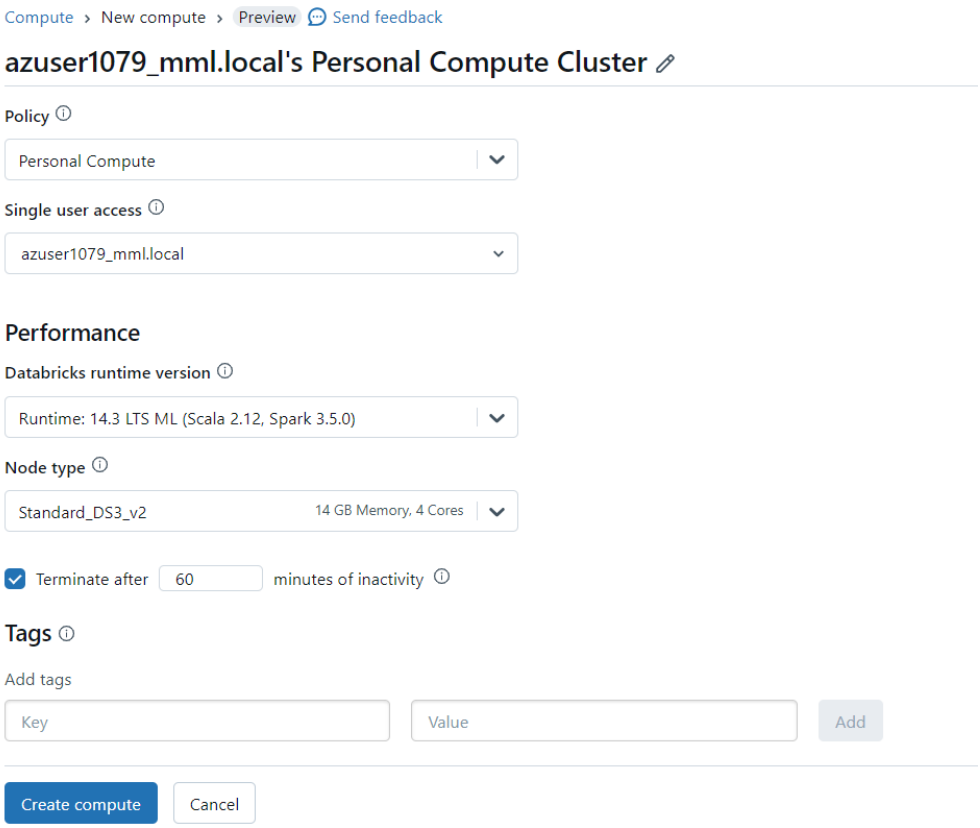
TASK 1

Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks

We intend to visualize the data using charts or graphs. we should be able to see the list of charts and graphs supported by Azure Databricks as shown below. Some of the most used charts are very well supported here like Bar chart, Scatter chart, Maps, Line chart, Area chart, Pie chart etc.

Data profiling is the process of collecting statistics and summaries of data to assess its quality and other characteristics.

- **Create a personal compute cluster in azure databricks**



- **Read csv file and display dataframe**

```
sparkDF = spark.read.csv("/databricks-datasets/bikeSharing/data-001/day.csv", header="true", inferSchema="true")
display(sparkDF)
```

(3) Spark Jobs

sparkDF: pyspark.sql.dataframe.DataFrame = [instant: integer, dteday: date ... 14 more fields]

Table + New result table: OFF

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hu
1	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.8
2	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.6
3	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.4
4	4	2011-01-04	1	0	1	0	2	1	1	0.2	0.212122	0.5
5	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.22927	0.4
6	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.5
7	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.4

731 rows | 16.65 seconds runtime Refreshed now

Command took 16.65 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/21/2024, 10:14:27 AM on azuser1079_mml.local's Personal Compute Cluster

- **Using scatter plot and choosing its axis**

Visualization Editor

Visualization type

Scatter

General | X axis | Y axis | Series | Colors | Data labels

X column

dteday | None

Y columns

casual

Group by

workingday

Error column

Choose column...

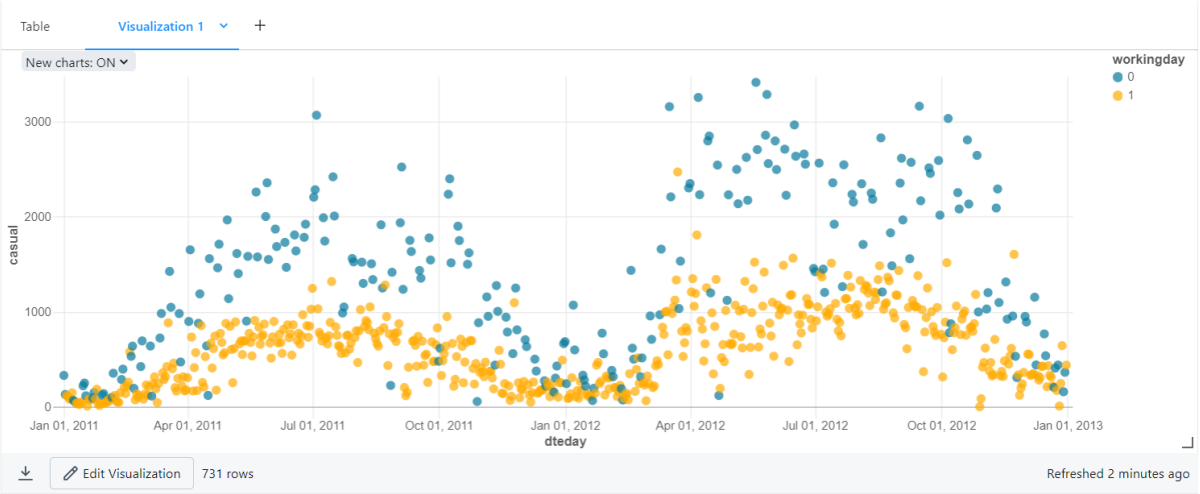
Legend placement

Automatic (Flexible)

Legend items order

Normal

- **Visualization of the scatter plot**

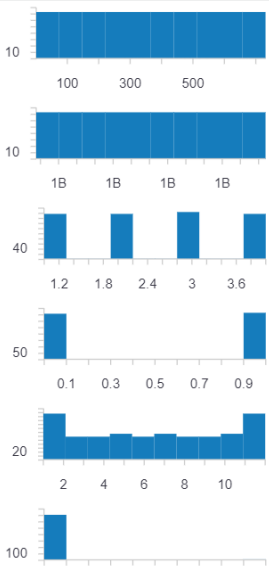


Command took 16.65 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/21/2024, 10:14:27 AM on azuser1079_mml.local's Personal Compute Cluster

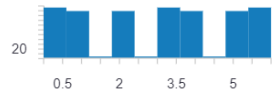
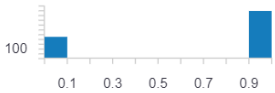
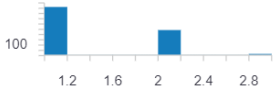
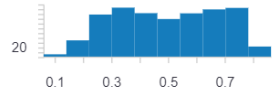
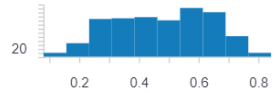
- **Getting Data profile of the data frame**

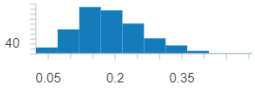
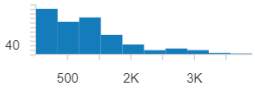
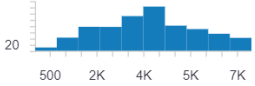
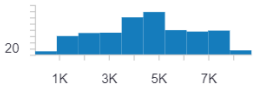
Standard
log expand

	count	missing	mean	std dev	zeros	min	median	max	custom
instant	731	0%	366	211.17	0%	1	366	731	data type: int
dteday	731	0%	1.33B	18.2M	0%	1.29B	1.33B	1.36B	data type: date min: 2011-01-01 max: 2012-12-31
season	731	0%	2.5	1.11	0%	1	3	4	data type: int
yr	731	0%	0.5	0.5	49.93%	0	1	1	data type: int
mnth	731	0%	6.52	3.45	0%	1	7	12	data type: int
holiday	731	0%	0.03	0.17	97.13%	0	0	1	data type: int



- **All the attribute of the table is displayed as data profile**

weekday	731	0%	3	2	14.36%	0	3	6	data type: int	
workingday	731	0%	0.68	0.47	31.6%	0	1	1	data type: int	
weathersit	731	0%	1.4	0.54	0%	1	1	3	data type: int	
temp	731	0%	0.5	0.18	0%	0.06	0.5	0.86	data type: double	
atemp	731	0%	0.47	0.16	0%	0.08	0.49	0.84	data type: double	
hum	731	0%	0.63	0.14	0.14%	0	0.63	0.97	data type: double	

windspeed	731	0%	0.19	0.08	0%	0.02	0.18	0.51	data type: double	
casual	731	0%	848.18	686.62	0%	2	713	3,410	data type: int	
registered	731	0%	3,656.17	1,560.26	0%	20	3,662	6,946	data type: int	
cnt	731	0%	4,504.35	1,937.21	0%	22	4,548	8,714	data type: int	

7.41 seconds runtime

Command took 16.65 seconds -- by azuser1079_mml.local@iihtl.onmicrosoft.com at 2/21/2024, 10:14:27 AM on azuser1079_mml.local's Personal Compute Cluster

TASK 2

Explain Overview of 3 level namespace and creating Unity Catalog objects

The Unity Catalog in Azure Databricks organizes data and AI assets using a three-level namespace, which includes catalogs, schemas (databases), and objects (tables, views, volumes, models).

Metastore is the top-level container for metadata. Each metastore exposes a three-level namespace (catalog.schema.table) that organizes your data.

Catalogs

- Catalogs serve as the top-level containers in the Unity Catalog namespace.
- They are used to organize data assets and provide a logical grouping for schemas and objects.
- Users can see all catalogs on which they have been assigned the USE CATALOG permission.
- Admins can assign default permissions on automatically provisioned catalogs or create new catalogs manually.
- To create a catalog, you can use the Databricks UI or run SQL commands in the notebook with appropriate permissions.

Schemas (Databases)

- Schemas are the second level in the Unity Catalog namespace.
- They organize tables and views within catalogs.
- Users can see all schemas on which they have been assigned the USE SCHEMA permission.
- Admins can assign default permissions on schemas or create new schemas manually within catalogs.

- Unity Catalog includes a default schema named "default" in each catalog, which is accessible to all users in the workspace.
- To create a schema, you can use the Databricks UI or run SQL commands in the notebook with appropriate permissions.

Objects (Tables, Views, Volumes, Models)

- Objects reside within schemas and represent various data and AI assets.
- **Tables:** Store rows of data and can be managed or external. Managed tables are managed by Unity Catalog, while external tables are not.
- **Views:** Read-only objects created from one or more tables or views and reside in the third layer of the namespace.
- **Volumes:** Provide non-tabular access to data stored in any format, containing directories and files. Managed and external volumes are supported.
- **Models**: Machine learning models registered in the MLflow Model Registry.
- Each object must adhere to specific permissions for creation, access, and manipulation.
- To create objects, you can use the Databricks UI, write SQL commands, or execute code in notebooks, ensuring that the user has appropriate permissions

Creating a unity catalog object

To create table object

- To create a table, users must have CREATE and USE SCHEMA permissions on the schema, and they must have the USE CATALOG permission on its parent catalog.
- To query a table, users must have the SELECT permission on the table, the USE SCHEMA permission on its parent schema, and the USE CATALOG permission on its parent catalog.

To create view object

- A view can be created from tables and other views in multiple schemas and catalogs. You can create dynamic views to enable row- and column-level permissions.

To create volume object

- To create a volume, users must have CREATE VOLUME and USE SCHEMA permissions on the schema, and they must have the USE CATALOG permission on its parent catalog.
- To read files and directories stored inside a volume, users must have the READ VOLUME permission, the USE SCHEMA permission on its parent schema, and the USE CATALOG permission on its parent catalog.

To create model object

- To create a model in Unity Catalog, users must have the CREATE MODEL privilege for the catalog or schema. The user must also have the USE CATALOG privilege on the parent catalog and USE SCHEMA on the parent schema.

TASK 3

Execute & explain, Azure DataFactory and its copy activity

Azure Data Factory, a cloud data integration service that orchestrates and automates movement and transformation of data.

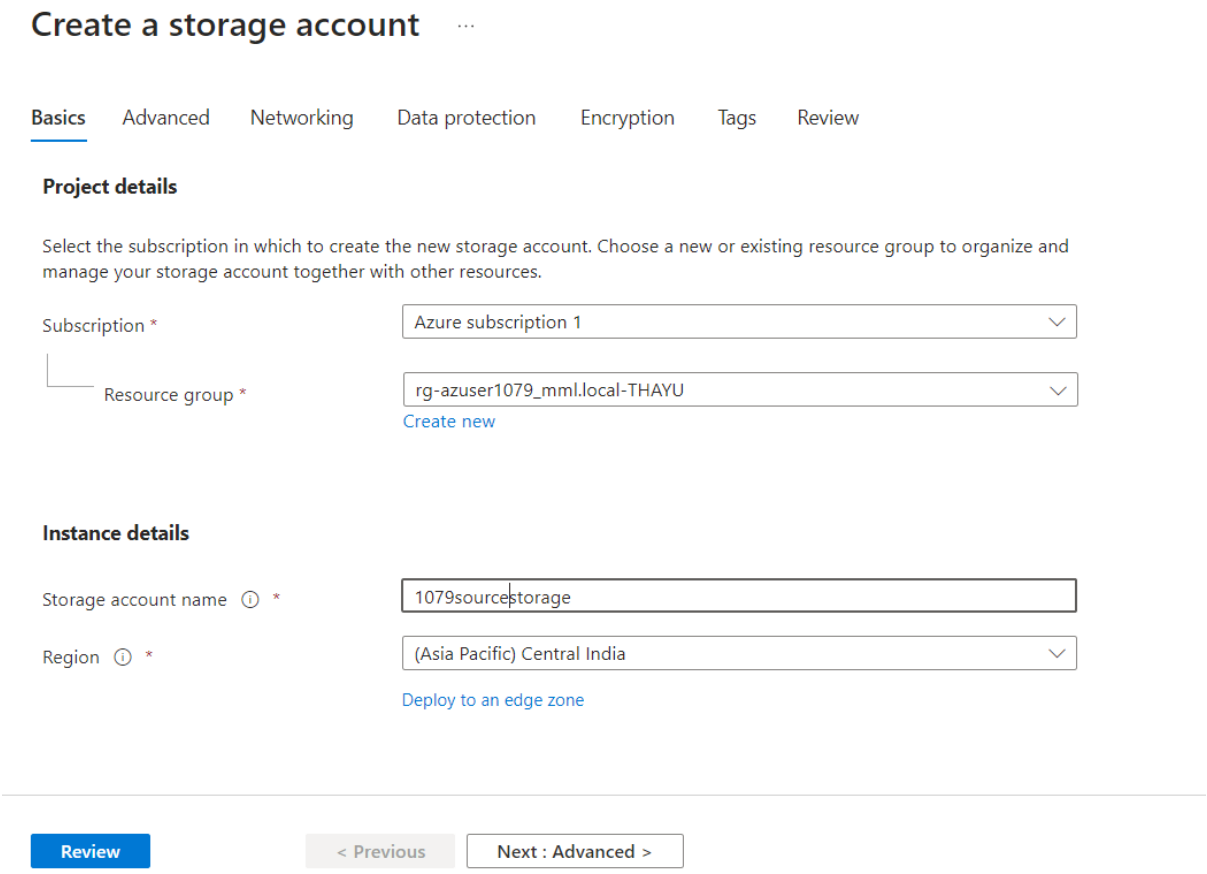
Pipeline

A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task.

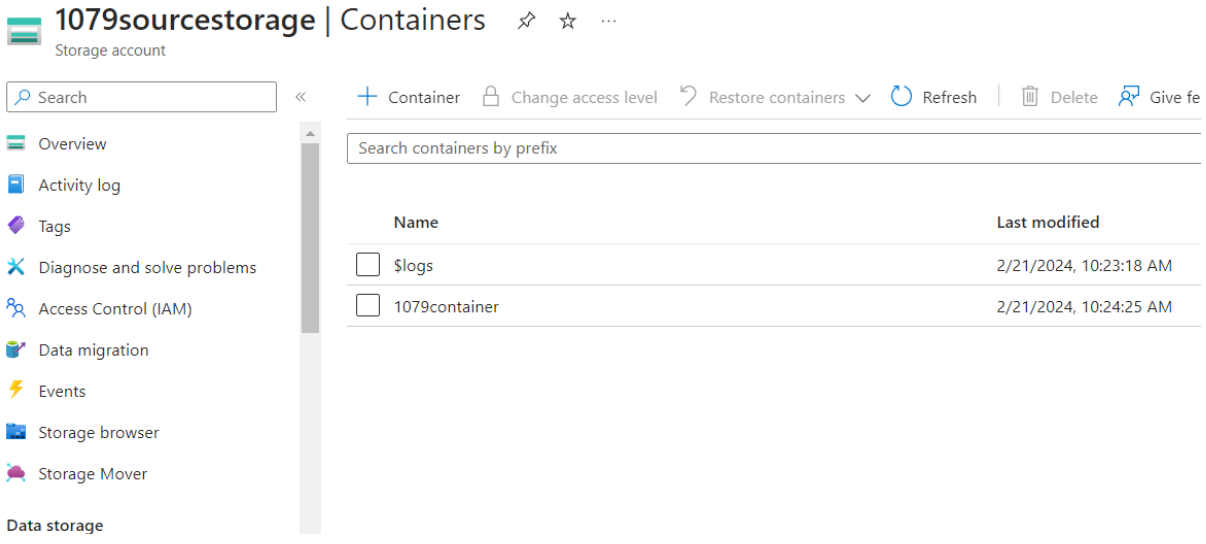
Activity

Activities represent a processing step in a pipeline. Use a copy activity to copy data from one data store to another data store.

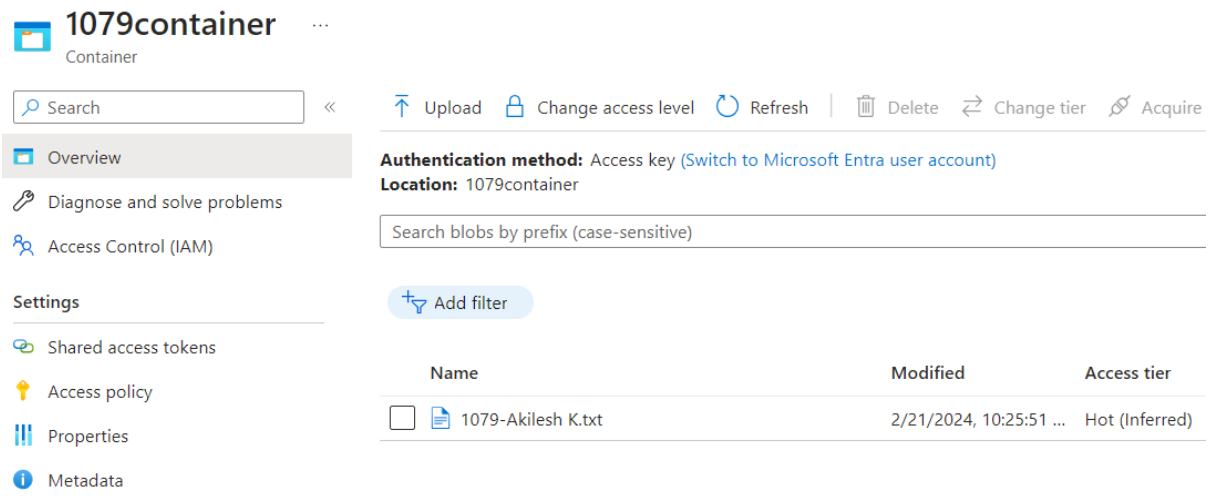
- **Creating a source storage account**



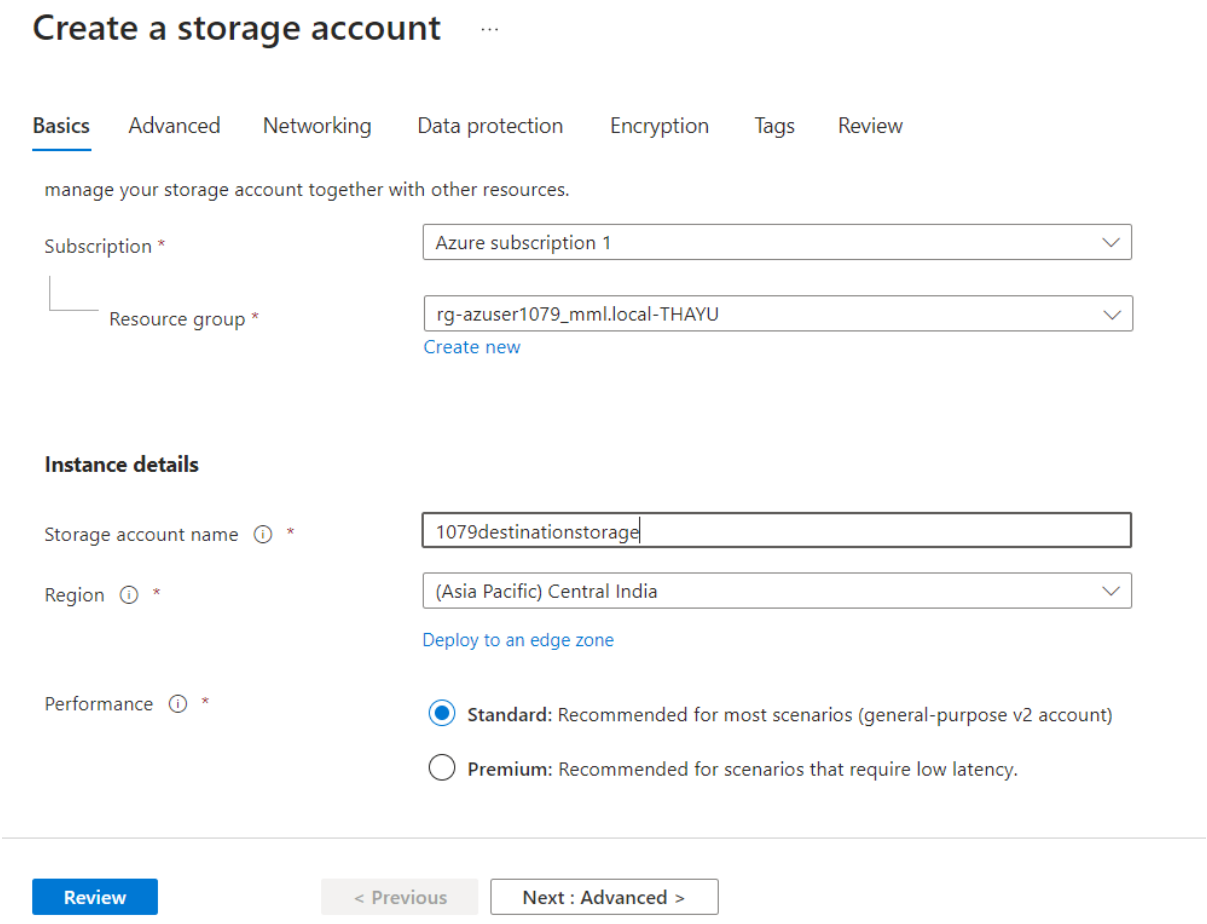
- **Create a container file**



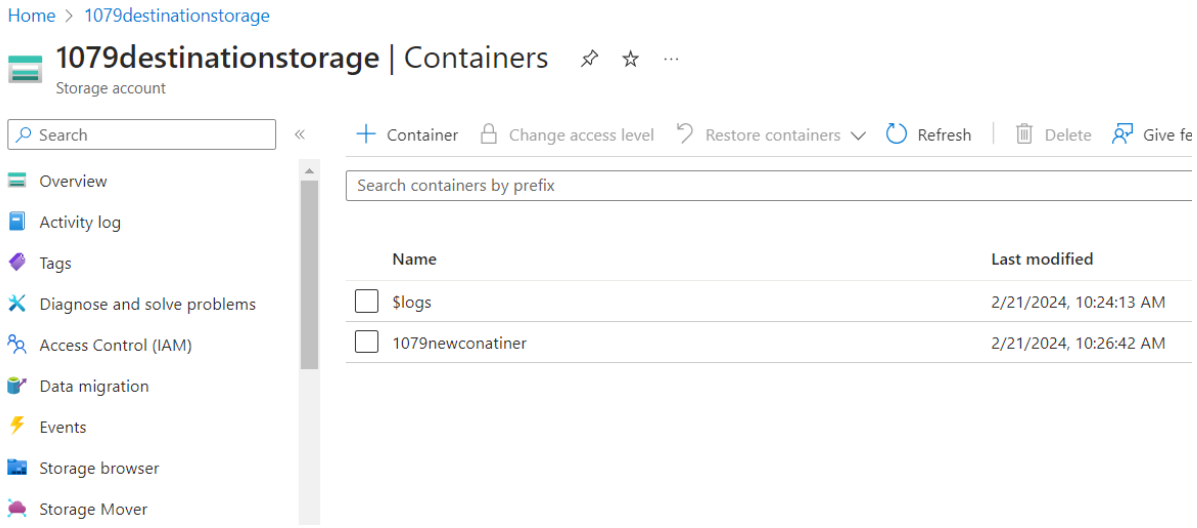
- **Upload a sample file in the container**



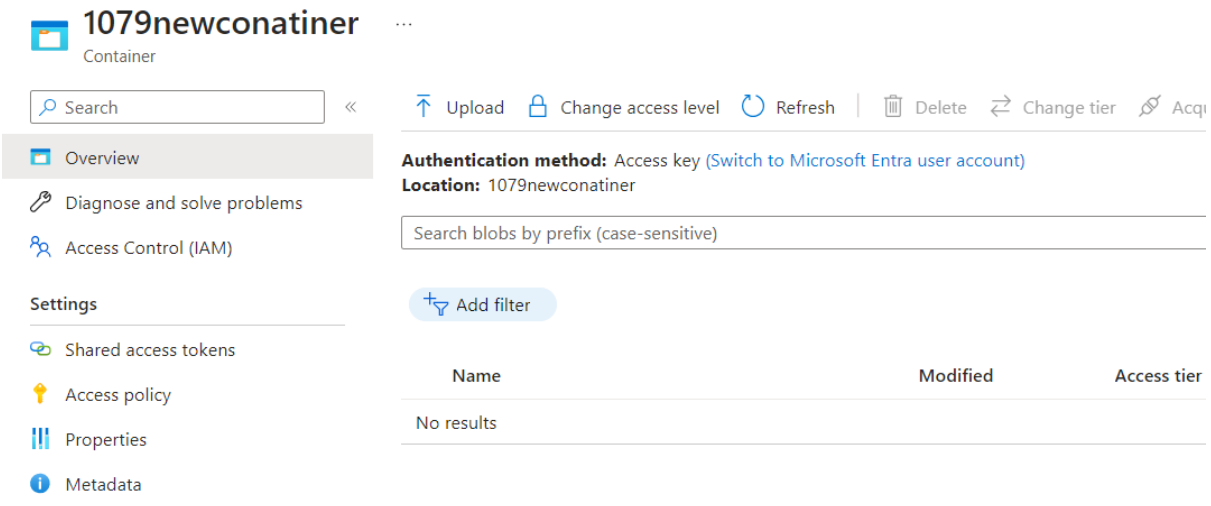
- **Create a destination storage account**



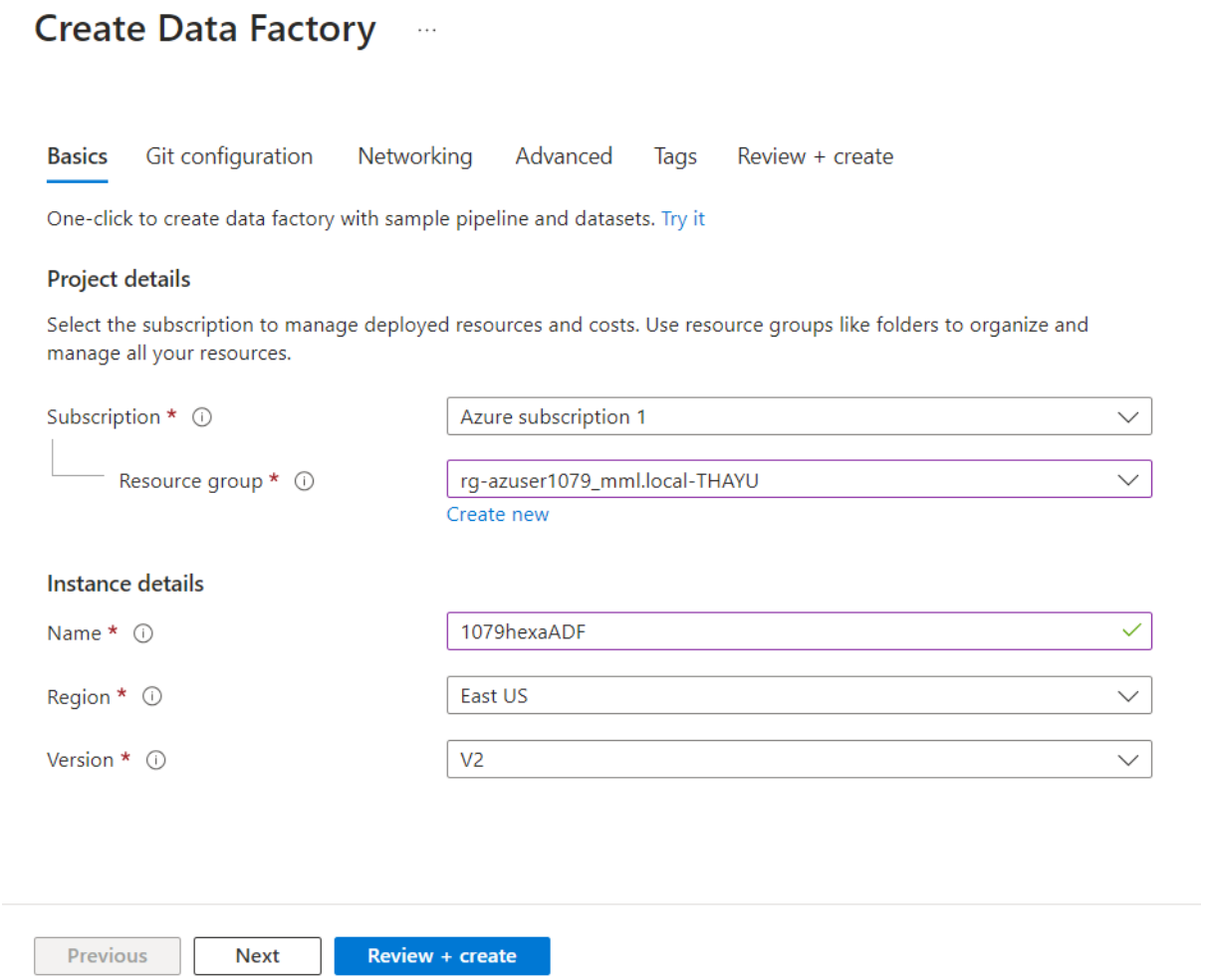
- **Create a destination container file**



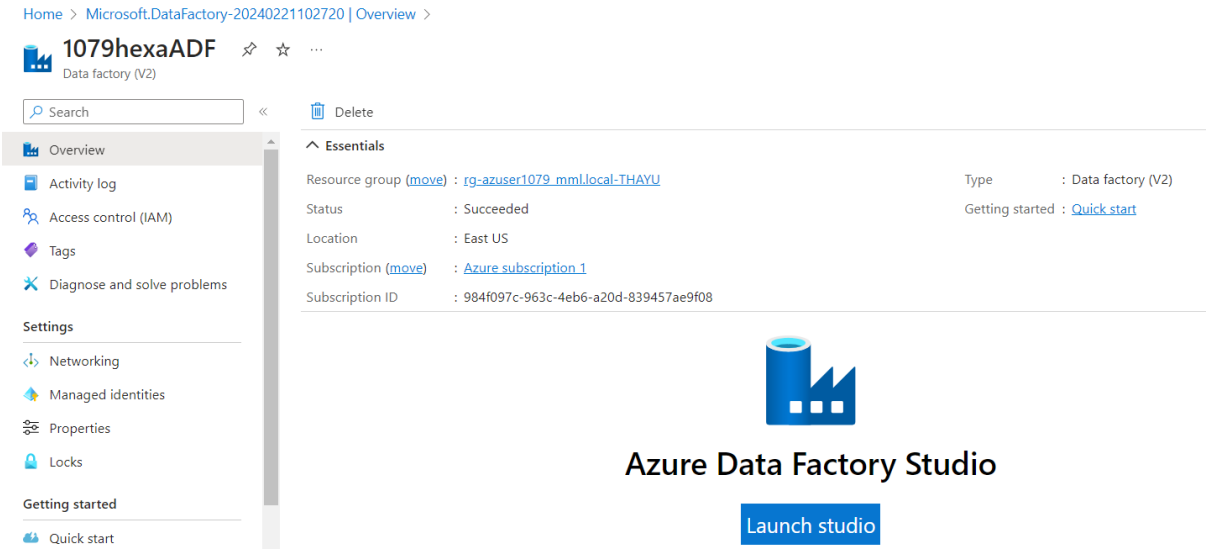
- **Keep the container empty**



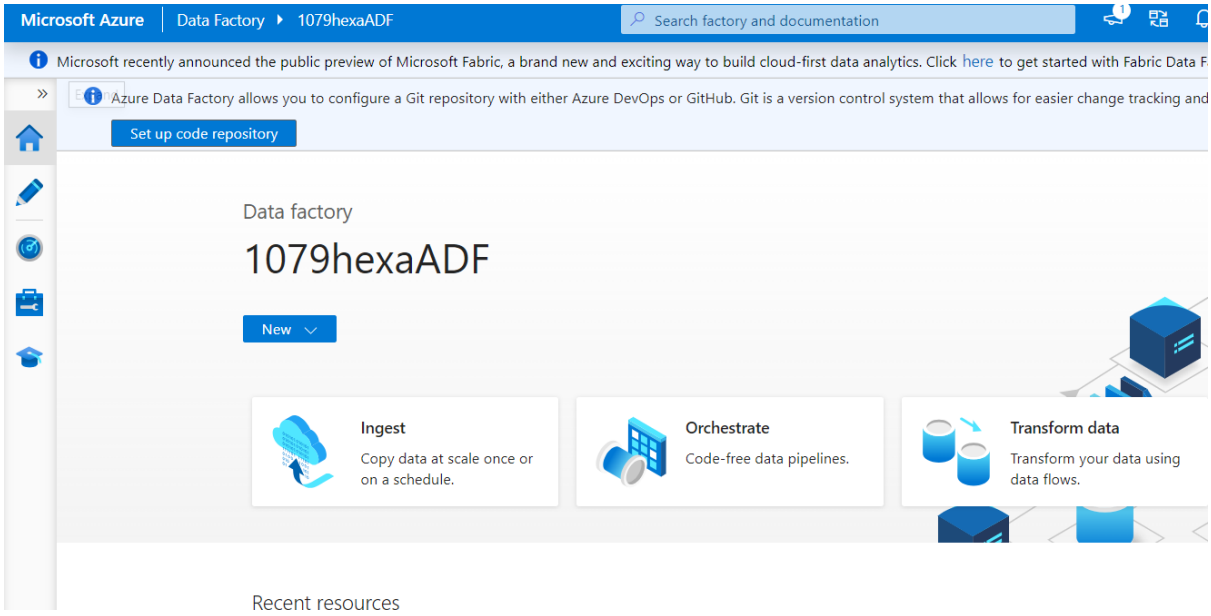
- **Create a azure datafactory**



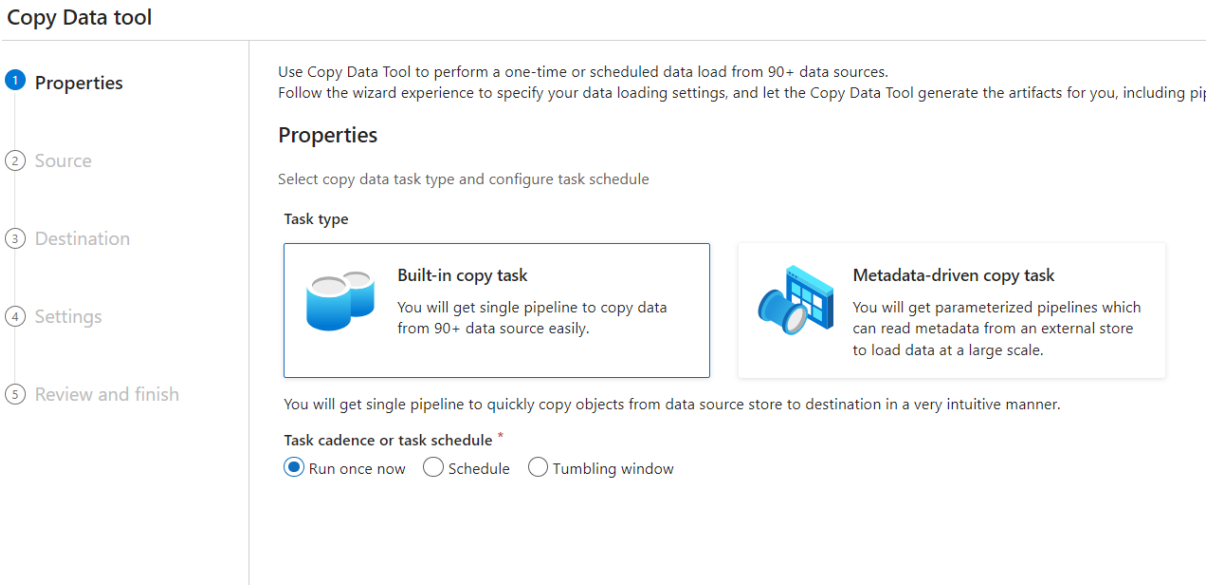
- **Launch the Data Factory studio**



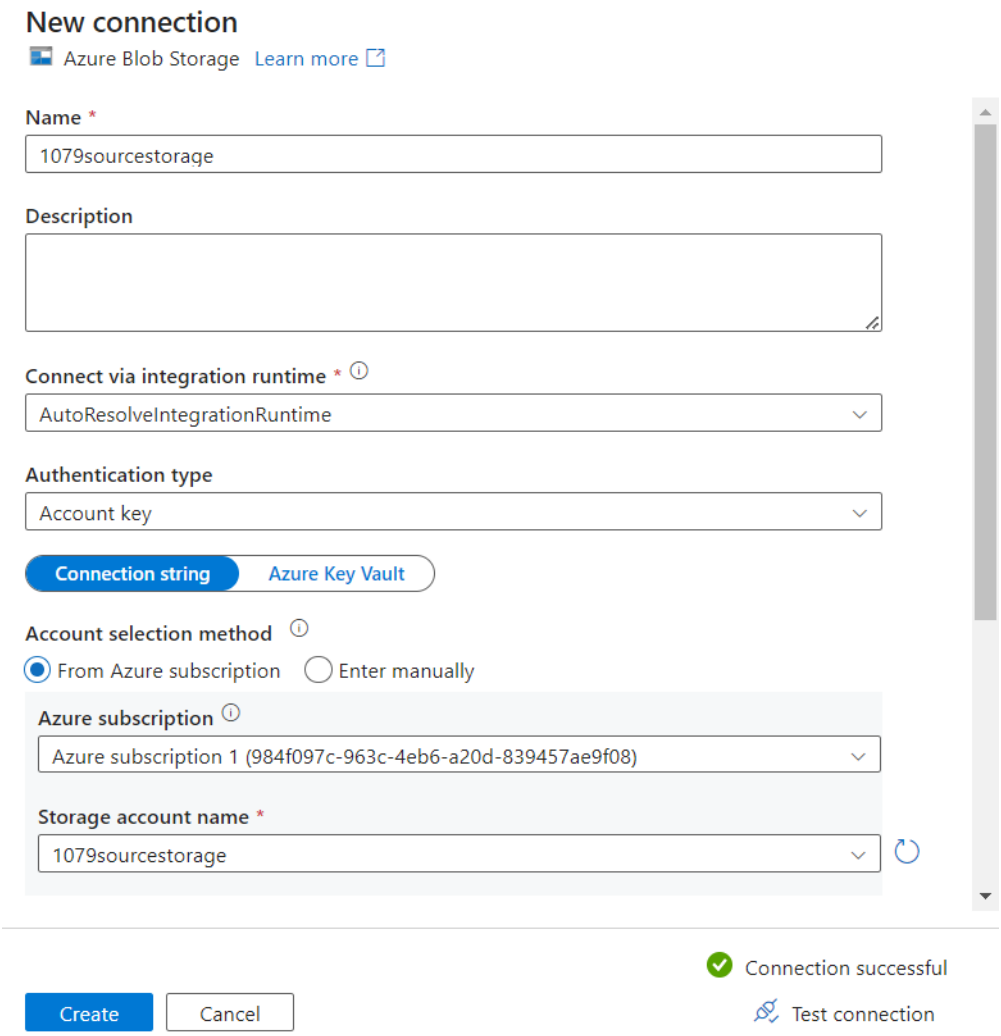
- **click ingest data**



- **Click built in copy task**



- **Choose azure storage and create a new connection by mentioning the storage name**



- **Specify the source container path**

Source data store

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

Source type Azure Blob Storage

Connection * 1079sourcestorage Edit New connection

File or folder *

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. Append a slash (/) at the end if the path refers to a folder.

1079container/ Browse

Options

- [ ] Binary copy
- [x] Recursively
- [ ] Enable partitions discovery

Max concurrent connections

- **Select the new storage account for destination datastore**

New connection

Azure Blob Storage Learn more

Name *

1079destinationstorage

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication type

Account key

Connection string | Azure Key Vault

Account selection method

(•) From Azure subscription ( ) Enter manually

Azure subscription

Azure subscription 1 (984f097c-963c-4eb6-a20d-839457ae9f08)

Storage account name *

1079destinationstorage

Connection successful

Test connection

Create Cancel

- **Mention the destination container path**

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 Blob Storage

Connection * 1079destinationstorage Edit New connection

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.

1079newconatiner/ Browse

File name

- **Create a pipeline name**

Settings

Enter name and description for the copy data task, more options for data movement

Task name * copyActivity-1079

Task description

Data consistency verification

Fault tolerance

Enable logging

Enable staging

Advanced

- **Summary of the pipeline is provided**

Summary

You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.

Properties

Edit

Task name	copyActivity-1079
Task description	

Source

Edit

Connection name	1079sourcestorage
Dataset name	SourceDataset_opx
Column delimiter	
Escape character	\
Quote char	"
First row as header	false

- **The pipeline gets executed**

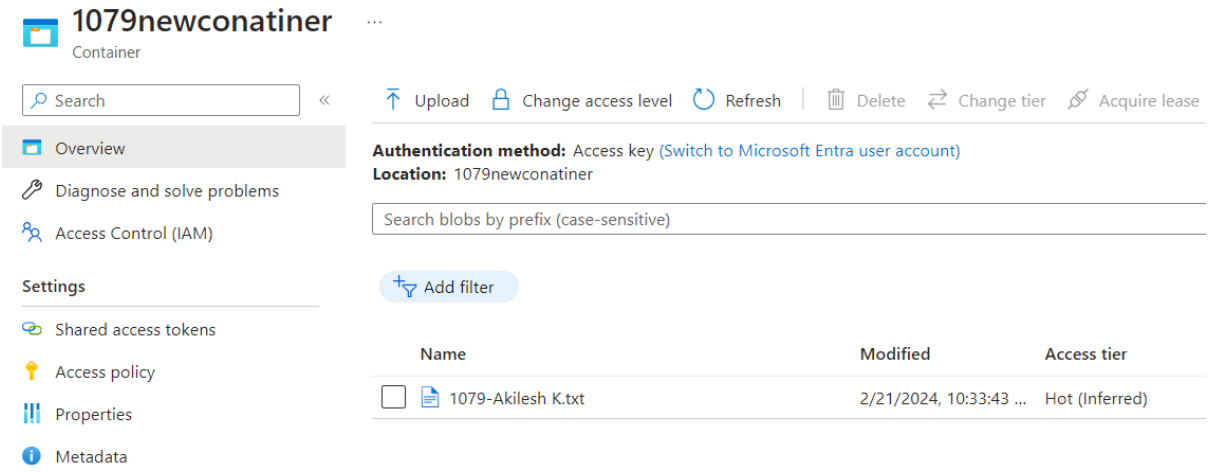
Deployment complete

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

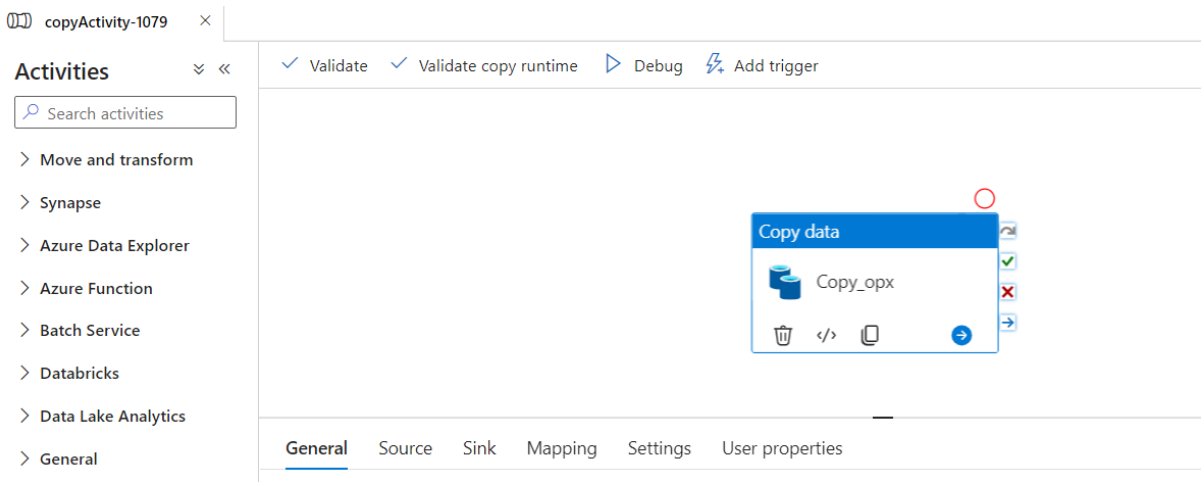
Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

Finish | Edit pipeline | Monitor

- **The data gets transferred to destination blob storage**



- **The pipeline is executed successfully**



A pipeline run is an instance of the pipeline execution. Pipeline runs are typically instantiated by passing the arguments to the parameters that are defined in pipelines. The arguments can be passed manually or within the trigger definition.