NAME : AKULA SHARATH CHANDRA
BATCH: DATA ENGINEERING
DATE :21-02-2024

AZURE CODING CHALLENGE

1)Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks.
2)Explain Overview of 3 level namespace and creating Unity Catalog objects.
3)Execute & explain Azure datafactory and its copy activity.

ANSWERS:

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

EDA:

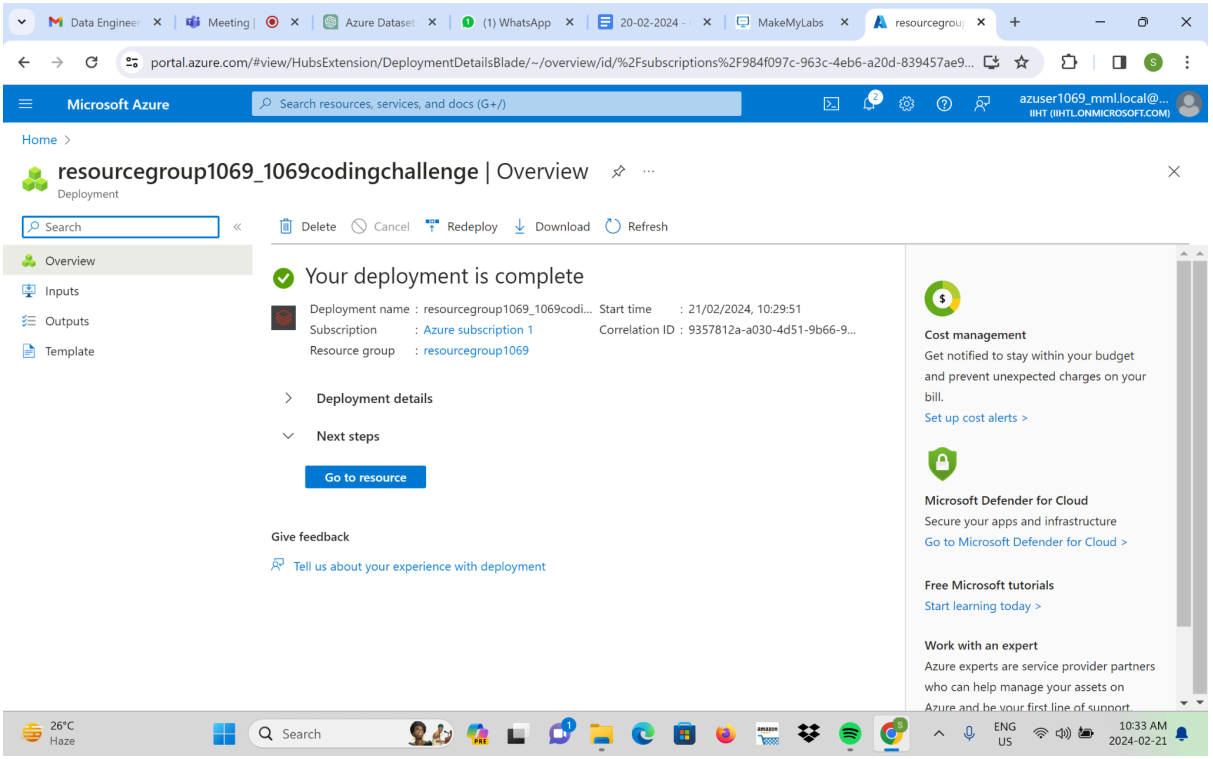
Exploratory Data Analysis is a data analysis approach that involves summarizing, visualizing, and understanding the main characteristics of a dataset to uncover patterns, trends, relationships.

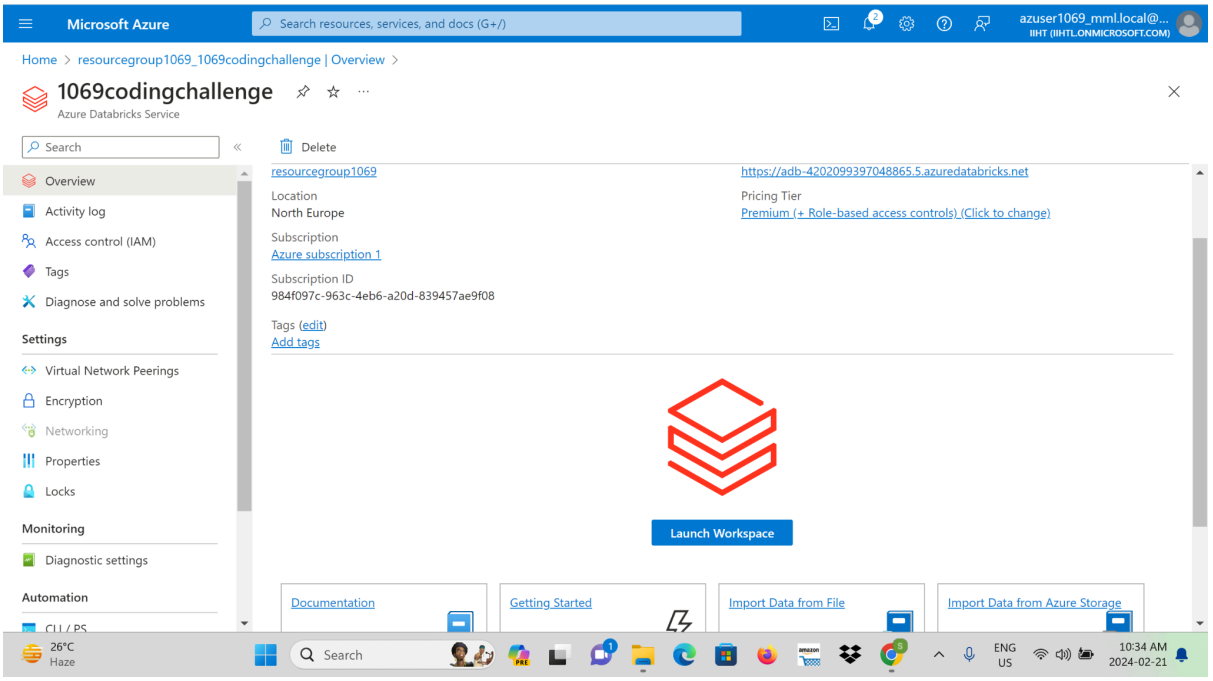
Here are some common steps involved in EDA using Databricks,

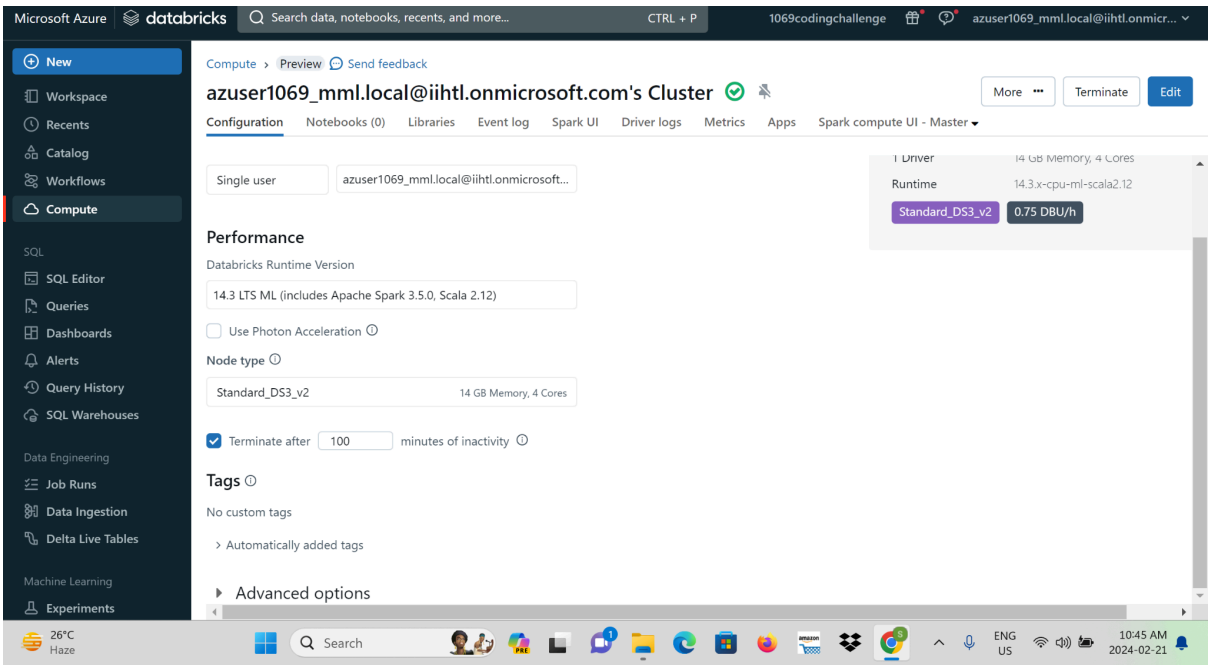
Data Loading: Databricks allows users to load data from various sources such as databases, data lakes, or cloud storage into Spark DataFrames.

Data Profiling: Users can generate summary statistics and descriptive analytics to understand the structure and distribution of data. Databricks offers tools for generating summary statistics, histograms, box plots, and other visualizations.

Data Visualization: Databricks provides integration with popular visualization libraries like Matplotlib, Seaborn, and Plotly, allowing users to create insightful plots and charts to visualize patterns and relationships within the data.
Data Cleaning: EDA often involves identifying and handling missing values, outliers, and inconsistencies in the data. Databricks offers powerful data manipulation and transformation capabilities using Spark SQL and DataFrame APIs.
Feature Engineering: Databricks enables users to create new features or derive insights from existing features to improve model performance. This may involve tasks such as feature scaling, encoding categorical variables, or creating new variables through transformations.
Statistical Analysis: Databricks allows users to perform statistical tests and analyses to validate assumptions, test hypotheses, and uncover relationships between variables.
Interactive Exploration: Databricks notebooks provide an interactive environment where users can iterate on their analysis, visualize results, and collaborate with team members.
Scalability: Databricks leverages the distributed computing capabilities of Apache Spark, enabling EDA on large-scale datasets that may not fit into memory on a single machine.


Microsoft Azure
Search resources, services, and docs (G+/)
azuser1069_mml.local@...
IIHT (IIHTL.ONMICROSOFT.COM)
Home
resourcegroup1069_1069codingchallenge | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : resourcegroup1069_1069codi...
Subscription : Azure subscription 1
Resource group : resourcegroup1069
Start time : 21/02/2024, 10:29:51
Correlation ID : 9357812a-a030-4d51-9b66-9...
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Start learning today >
Work with an expert
Azure experts are service provider partners who can help manage your assets on
26°C
Haze
Search
ENG
US
10:33 AM
2024-02-21


Microsoft Azure
Search resources, services, and docs (G+/)
azuser1069_mml.local@...
IIHT (IIHTL.ONMICROSOFT.COM)
Home
resourcegroup1069_1069codingchallenge | Overview
1069codingchallenge
Azure Databricks Service
Search
Delete
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Virtual Network Peerings
Encryption
Networking
Properties
Locks
Monitoring
Diagnostic settings
Automation
resourcegroup1069
Location
North Europe
Subscription
Azure subscription 1
Subscription ID
984f097c-963c-4eb6-a20d-839457ae9f08
Tags (edit)
Add tags
https://adb-4202099397048865.5.azuredatabricks.net
Pricing Tier
Premium (+ Role-based access controls) (Click to change)
Launch Workspace
Documentation
Getting Started
Import Data from File
Import Data from Azure Storage
26°C
Haze
Search
ENG
US
10:34 AM
2024-02-21


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
1069codingchallenge
azuser1069_mml.local@iihtl.onmicr...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
Compute › Preview Send feedback
azuser1069_mml.local@iihtl.onmicrosoft.com's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark compute UI - Master
Single user
azuser1069_mml.local@iihtl.onmicrosoft...
Runtime
14.3.x-cpu-ml-scala2.12
Standard_DS3_v2
0.75 DBU/h
Performance
Databricks Runtime Version
14.3 LTS ML (includes Apache Spark 3.5.0, Scala 2.12)
Use Photon Acceleration
Node type
Standard_DS3_v2
14 GB Memory, 4 Cores
Terminate after
100
minutes of inactivity
Tags
No custom tags
Automatically added tags
Advanced options
26°C
Haze
Search
ENG
US
10:45 AM
2024-02-21

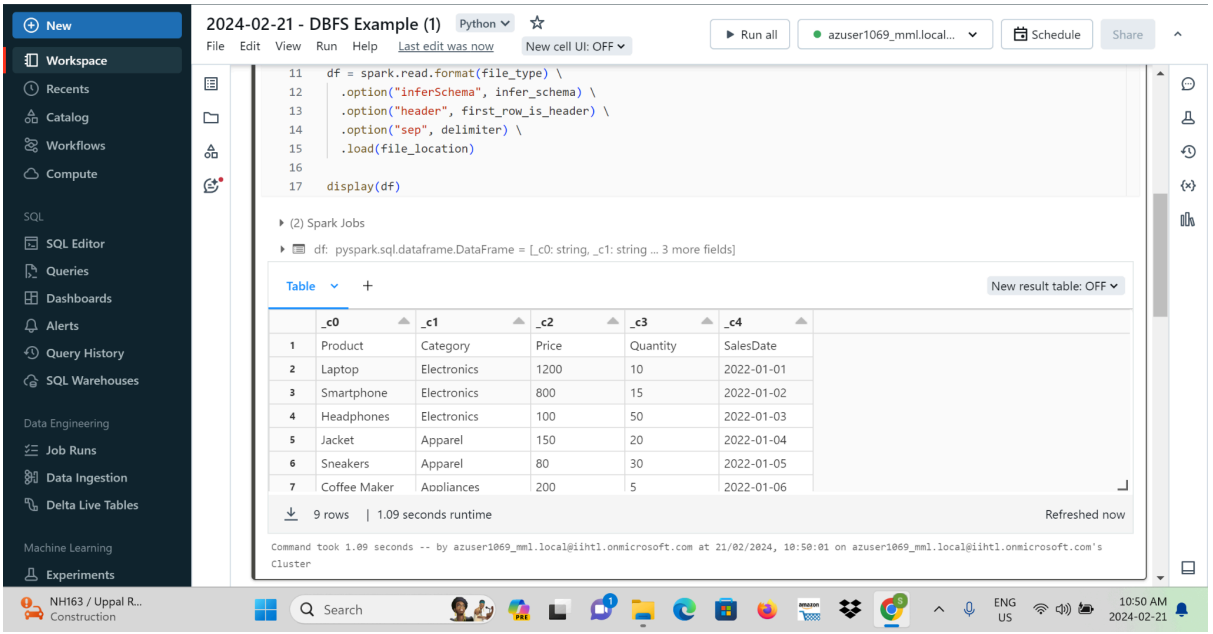
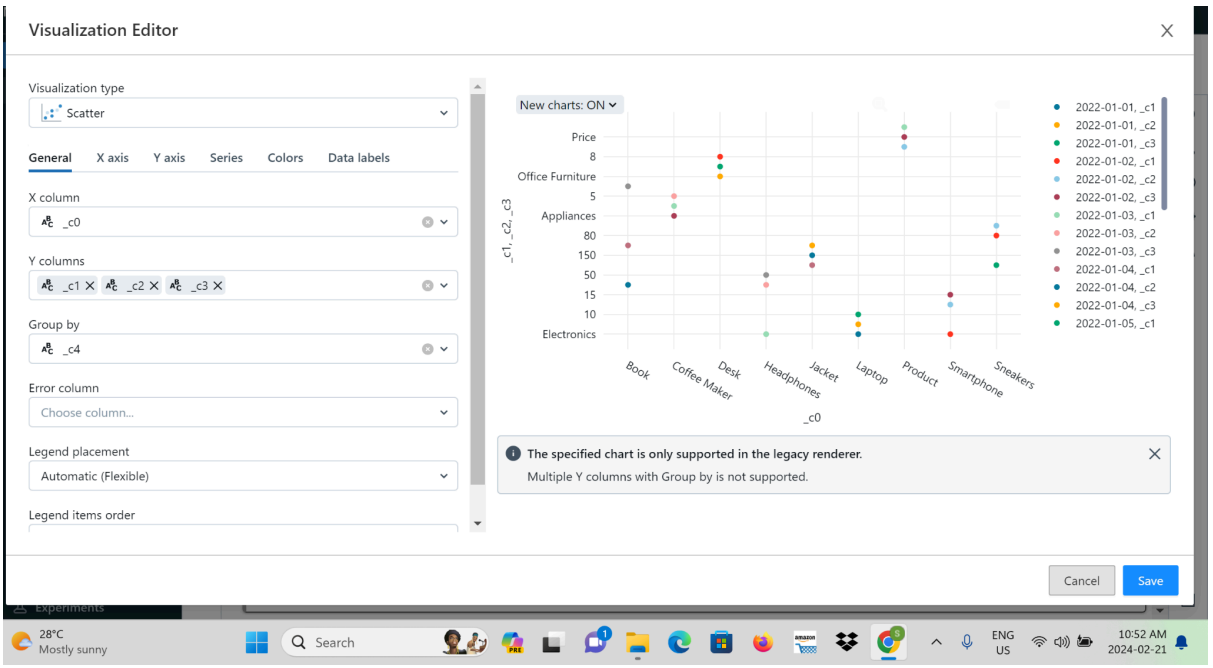
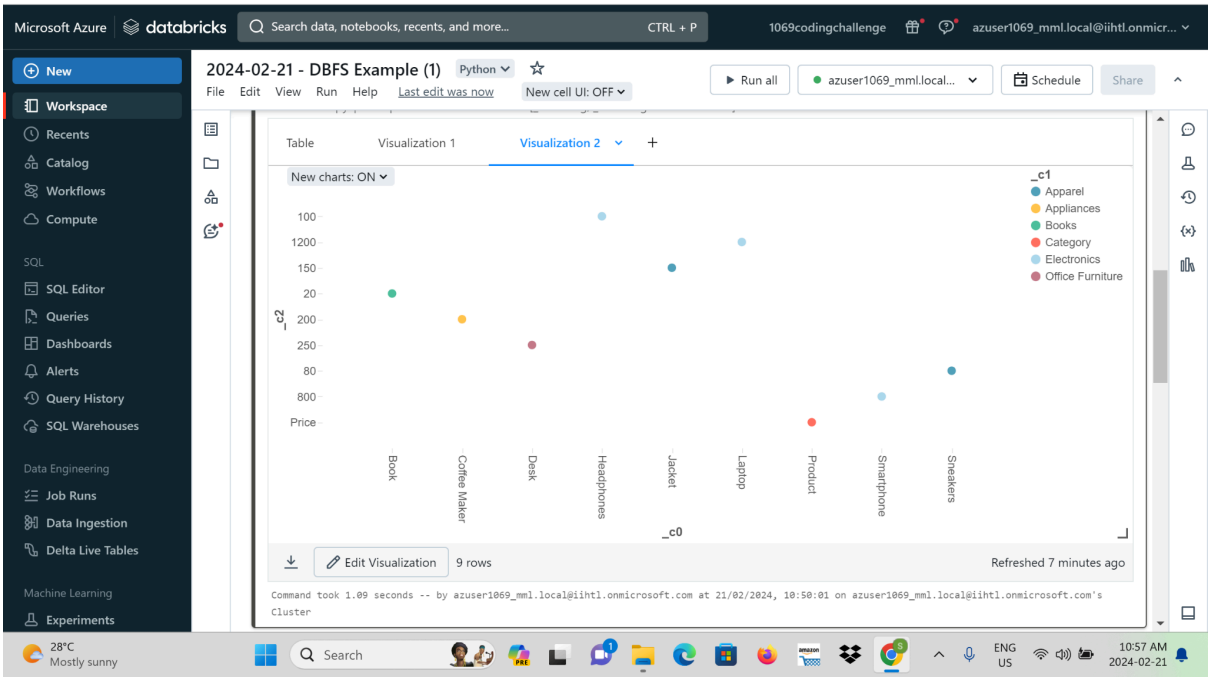

New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
2024-02-21 - DBFS Example (1)
Python
File
Edit
View
Run
Help
Last edit was now
New cell UI: OFF
Run all
azuser1069_mml.local...
Schedule
Share
df = spark.read.format(file_type) \
.option("inferSchema", infer_schema) \
.option("header", first_row_is_header) \
.option("sep", delimiter) \
.load(file_location)
display(df)
(2) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 3 more fields]
Table
New result table: OFF
_c0 _c1 _c2 _c3 _c4
1 Product Category Price Quantity SalesDate
2 Laptop Electronics 1200 10 2022-01-01
3 Smartphone Electronics 800 15 2022-01-02
4 Headphones Electronics 100 50 2022-01-03
5 Jacket Apparel 150 20 2022-01-04
6 Sneakers Apparel 80 30 2022-01-05
7 Coffee Maker Appliances 200 5 2022-01-06
9 rows | 1.09 seconds runtime
Refreshed now
Command took 1.09 seconds -- by azuser1069_mml.local@iihtl.onmicrosoft.com at 21/02/2024, 10:50:01 on azuser1069_mml.local@iihtl.onmicrosoft.com's Cluster
NH163 / Uppal R...
Construction
Search
ENG
US
10:50 AM
2024-02-21


Table
Visualization
Data Profile
_c0
_c2
_c3
_c4
1
Price
Quantity
SalesDate


Visualization Editor
Visualization type
Scatter
General
X axis
Y axis
Series
Colors
Data labels
X column
_c0
Y columns
_c1
_c2
_c3
Group by
_c4
Error column
Choose column...
Legend placement
Automatic (Flexible)
Legend items order
New charts: ON
Price
8
Office Furniture
5
Appliances
80
150
50
15
10
Electronics
Book
Coffee Maker
Desk
Headphones
Jacket
Laptop
Product
Smartphone
Sneakers
_c0
_c1, _c2, _c3
2022-01-01, _c1
2022-01-01, _c2
2022-01-01, _c3
2022-01-02, _c1
2022-01-02, _c2
2022-01-02, _c3
2022-01-03, _c1
2022-01-03, _c2
2022-01-03, _c3
2022-01-04, _c1
2022-01-04, _c2
2022-01-04, _c3
2022-01-05, _c1
The specified chart is only supported in the legacy renderer.
Multiple Y columns with Group by is not supported.
Cancel
Save
28°C
Mostly sunny
Search
ENG
US
10:52 AM
2024-02-21


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
1069codingchallenge
azuser1069_mml.local@iihtl.onmicr...
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Experiments
2024-02-21 - DBFS Example (1)
Python
File
Edit
View
Run
Help
Last edit was now
New cell UI: OFF
Run all
azuser1069_mml.local...
Schedule
Share
Table
Visualization 1
Visualization 2
New charts: ON
100
1200
150
20
200
250
80
800
Price
_c2
Book
Coffee Maker
Desk
Headphones
Jacket
Laptop
Product
Smartphone
Sneakers
_c0
_c1
Apparel
Appliances
Books
Category
Electronics
Office Furniture
Edit Visualization
9 rows
Refreshed 7 minutes ago
Command took 1.09 seconds -- by azuser1069_mml.local@iihtl.onmicrosoft.com at 21/02/2024, 10:50:01 on azuser1069_mml.local@iihtl.onmicrosoft.com's Cluster
28°C
Mostly sunny
Search
ENG
US
10:57 AM
2024-02-21

3)Execute & explain Azure datafactory and its copy activity?

A)<u>DATA FACTORY</u>: Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation. ADF does not store any data itself. It allows you to create data-driven workflows to orchestrate the movement of data between supported data stores and then process the data using compute services in other regions or in an on-premise environment. It also allows you to monitor and manage workflows using both programmatic and UI mechanisms.

—> It works in 3 steps

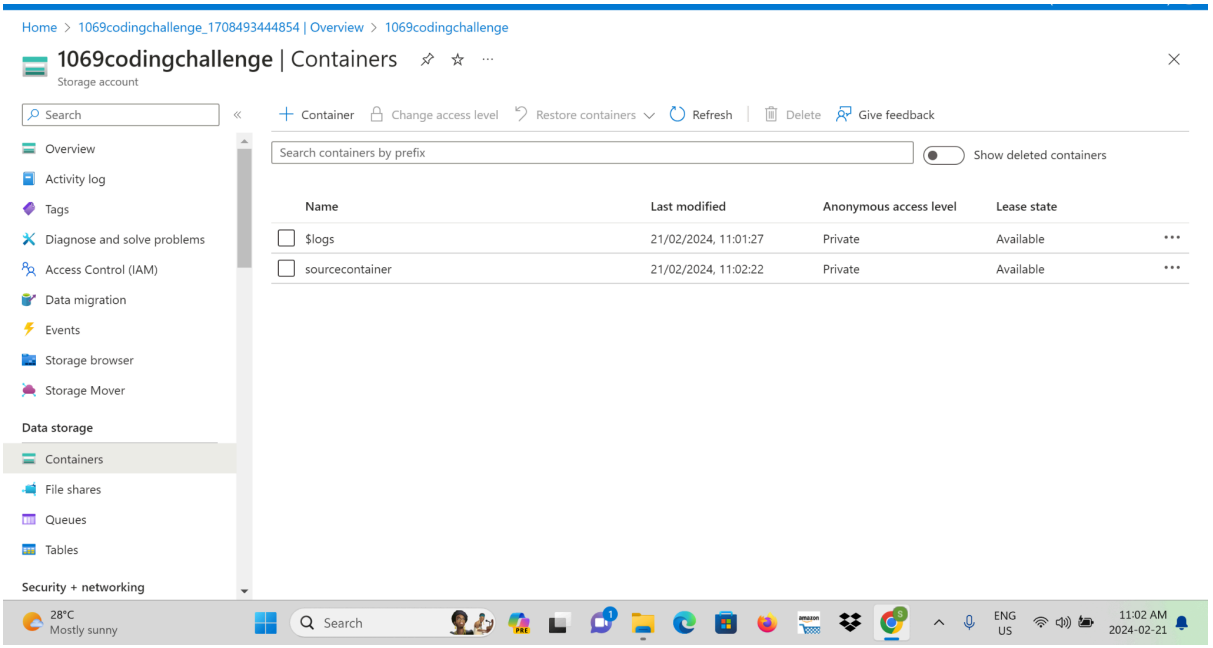
i) Connect and collect

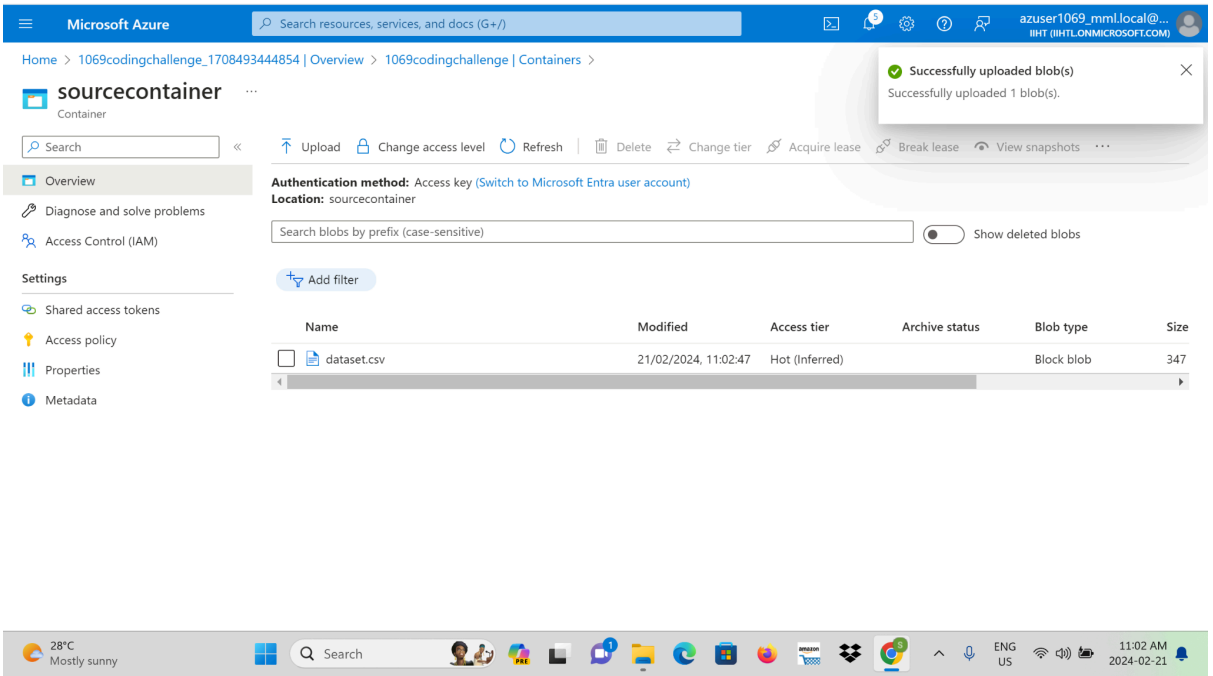
ii) Transform and enrich

iii) Publish

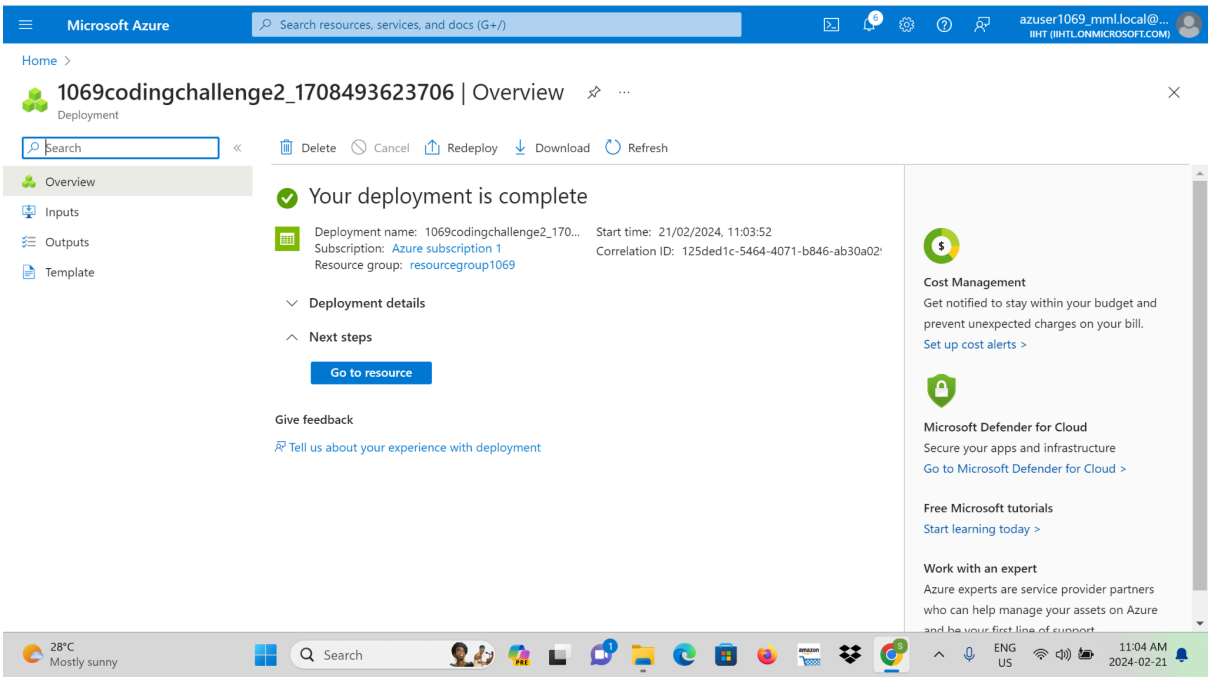
COPY ACTIVITY: Copy Activity in Azure Data Factory copies data from a source data store to a sink data store.

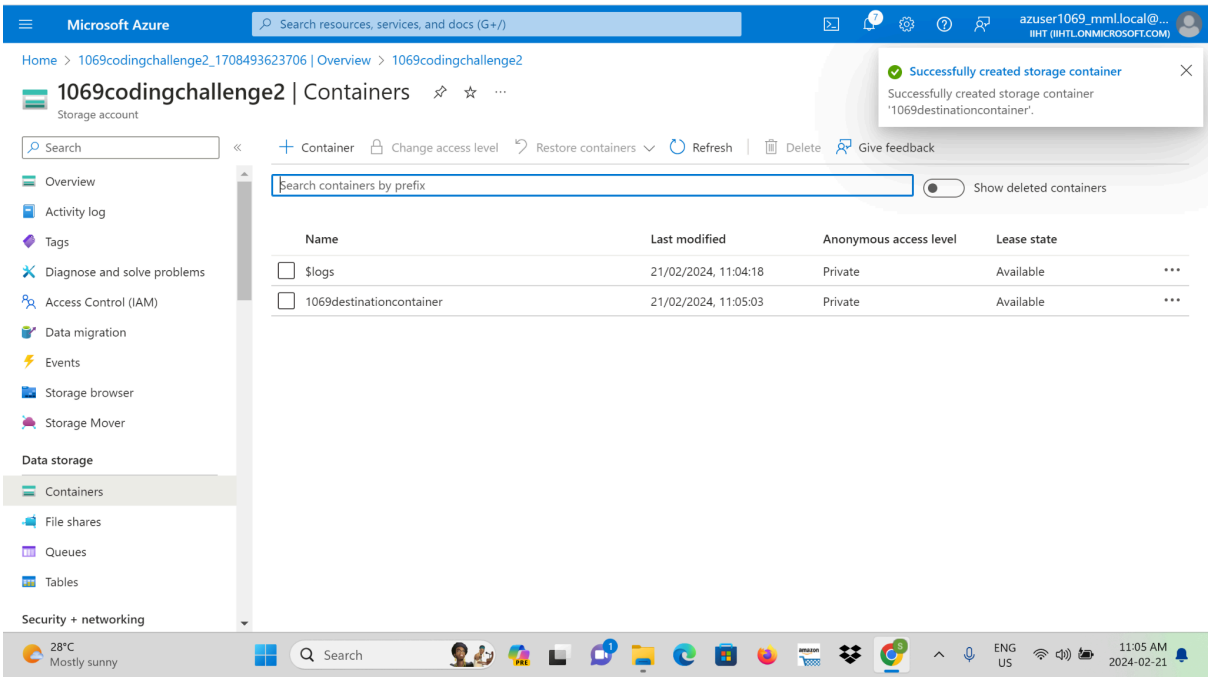
—> to create a data factory account firstly we need to create two storage accounts one is for source group and one is for destination group.
—> I have created a source storage account and then created a container and uploaded a file in it .

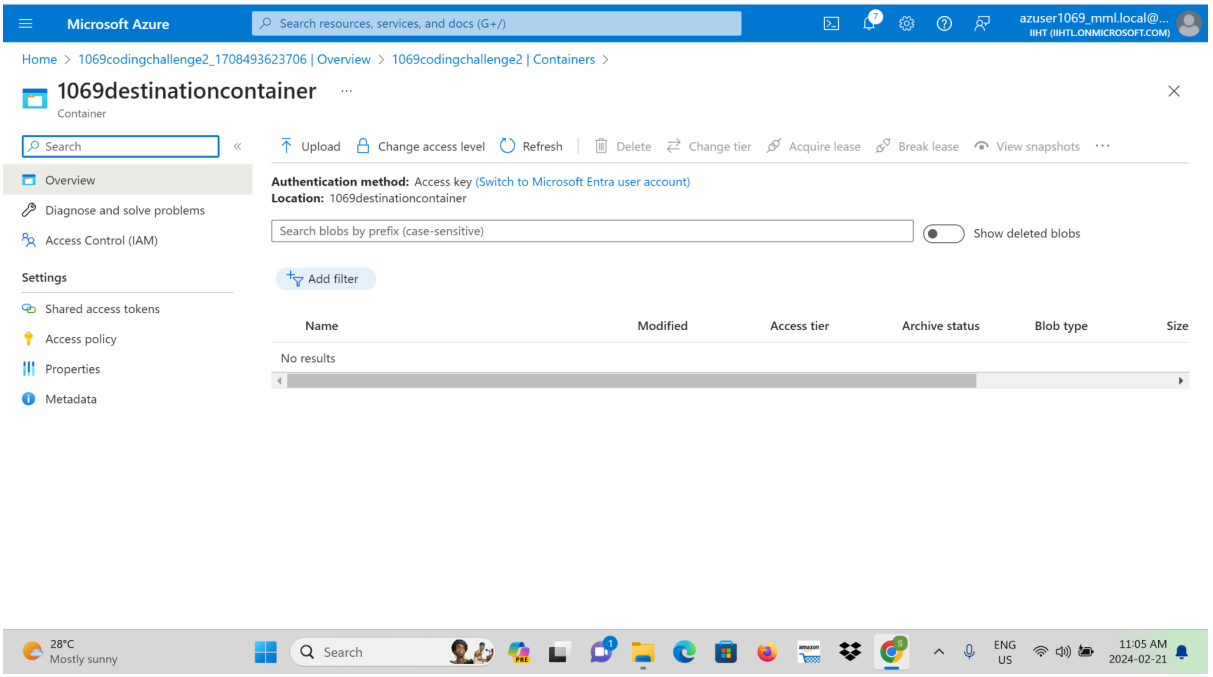




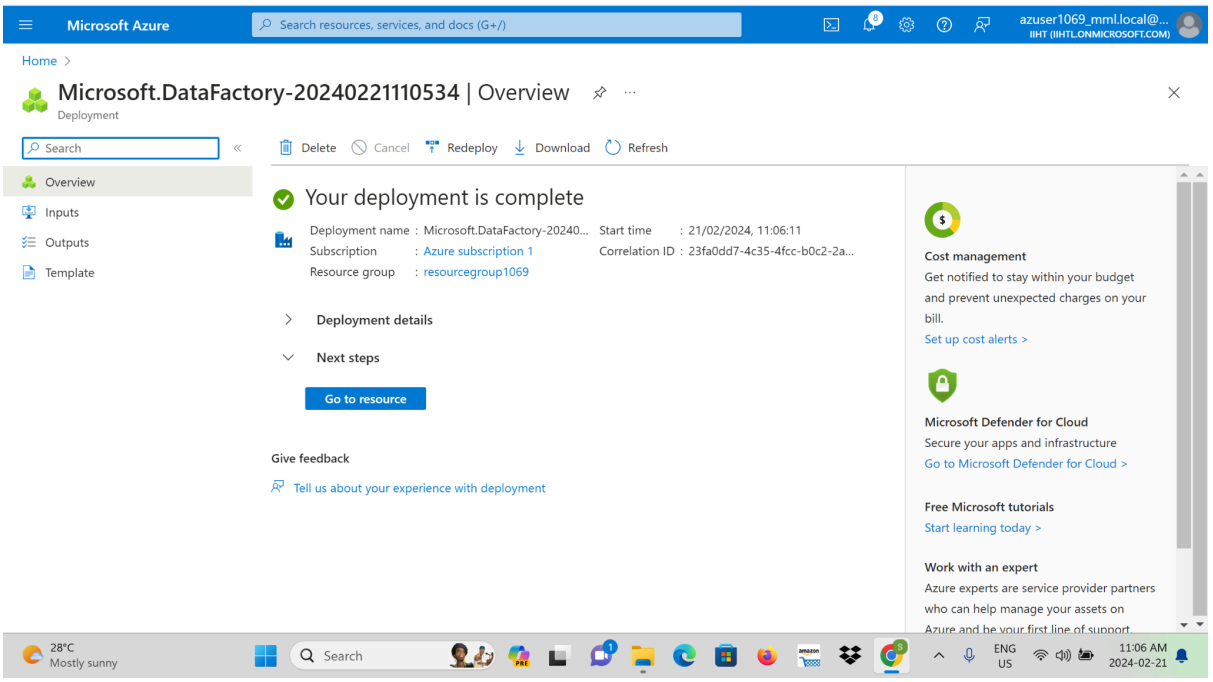
→I have created another storage account for the destination group and then I have created a container without upload a file in it.

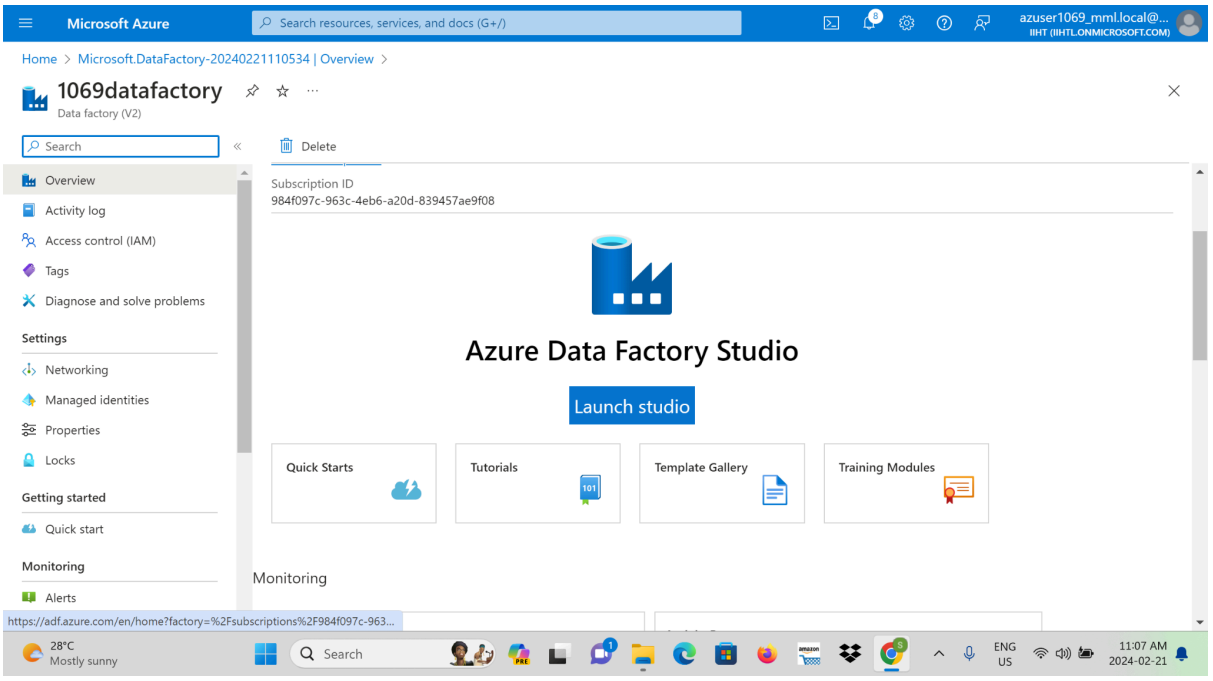




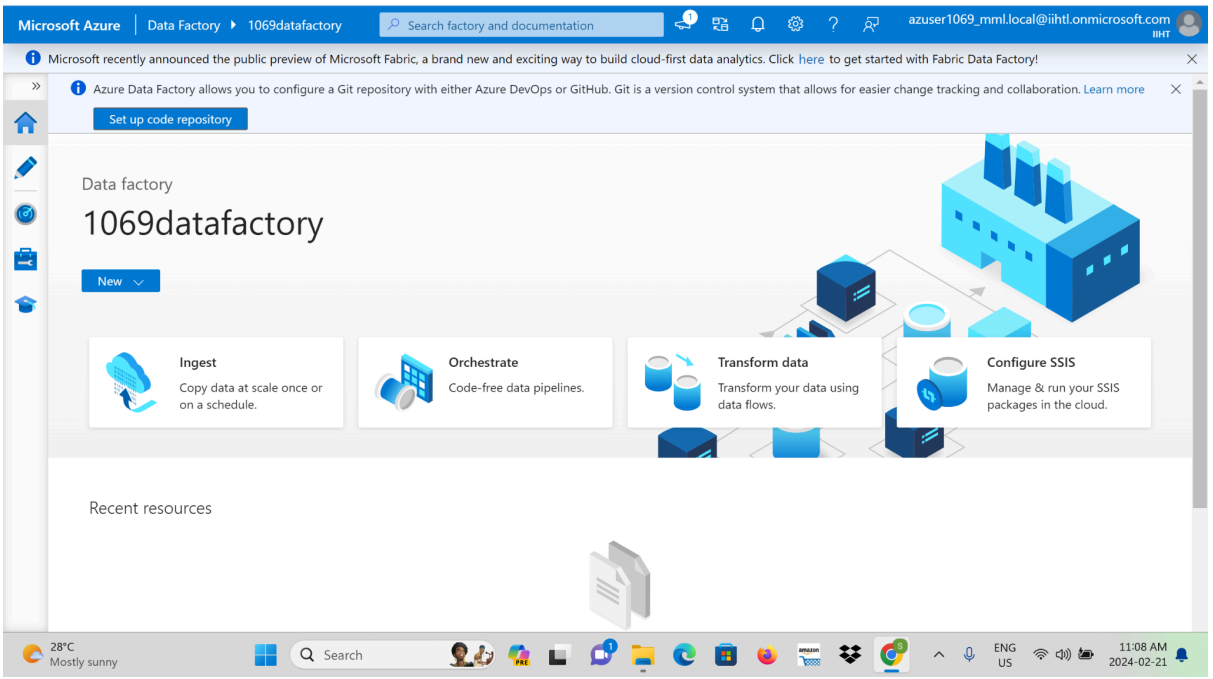

Microsoft Azure
Search resources, services, and docs (G+/)
azuser1069_mml.local@...
IIHT (IIHTLONMICROSOFT.COM)
Home > 1069codingchallenge2_1708493623706 | Overview > 1069codingchallenge2 | Containers >
1069destinationcontainer
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Upload
Change access level
Refresh
Delete
Change tier
Acquire lease
Break lease
View snapshots
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: 1069destinationcontainer
Search blobs by prefix (case-sensitive)
Show deleted blobs
Add filter
Name
Modified
Access tier
Archive status
Blob type
Size
No results
28°C
Mostly sunny
Search
ENG
US
11:05 AM
2024-02-21

—>Then I created a data factory account with the same resource group and clicked on launch studio.

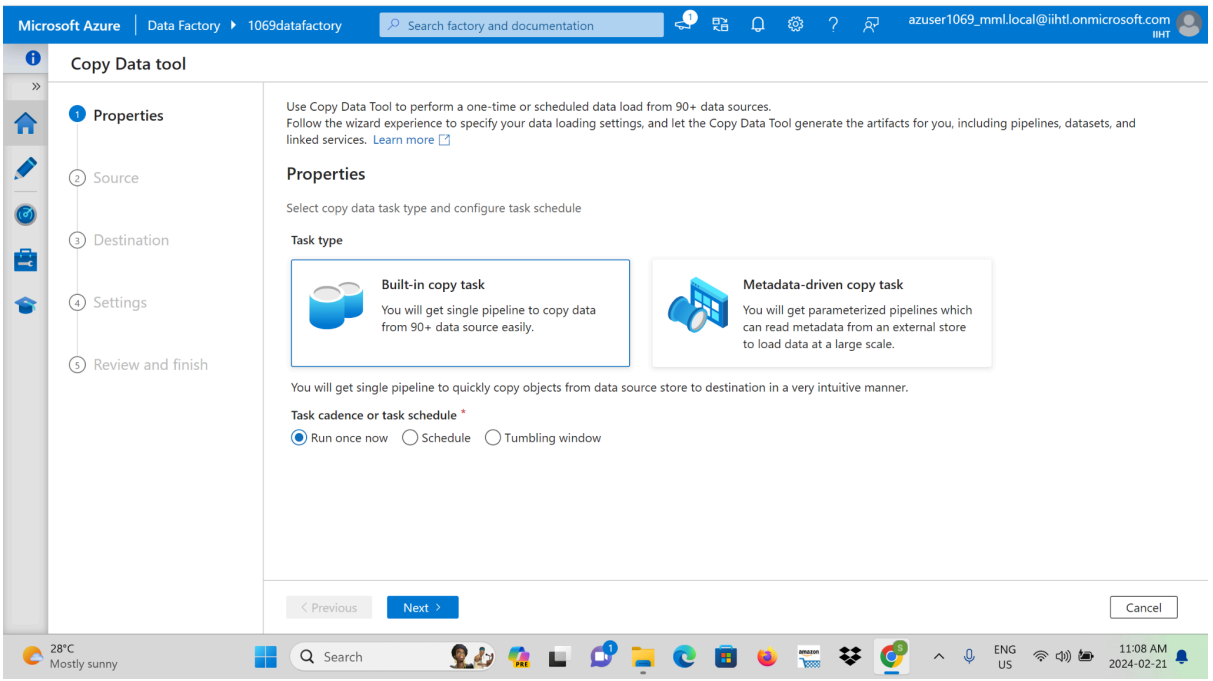




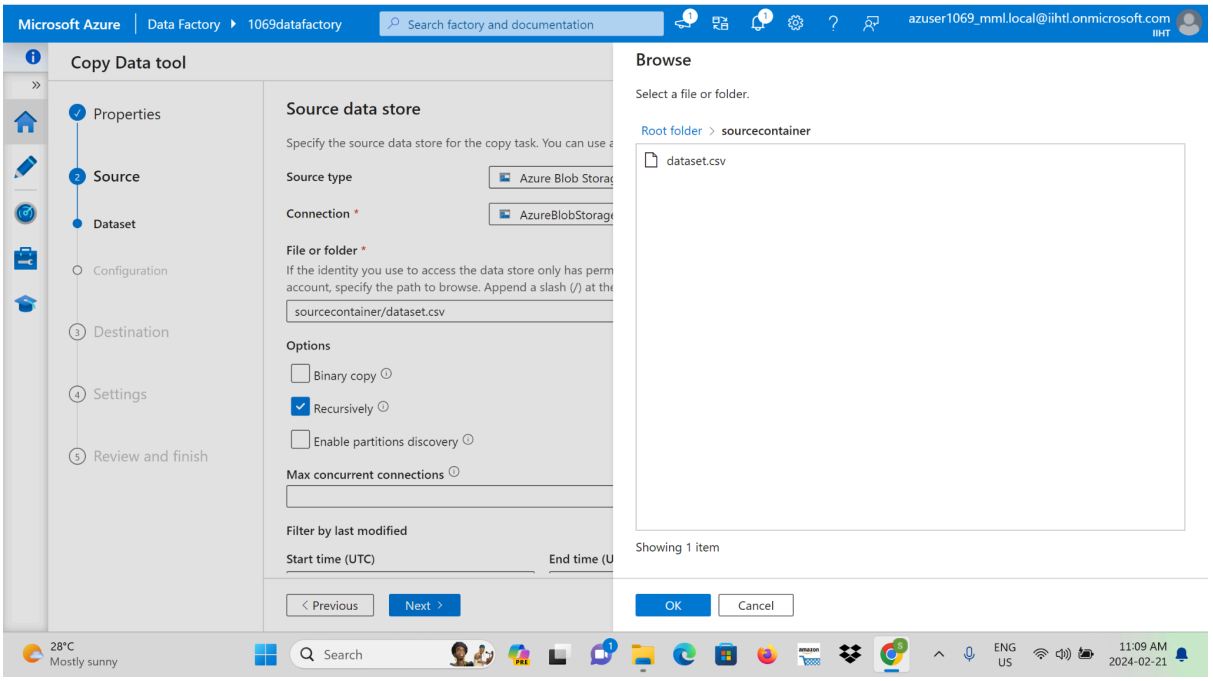
—>Then I have selected an option called ingestion to set the connection .

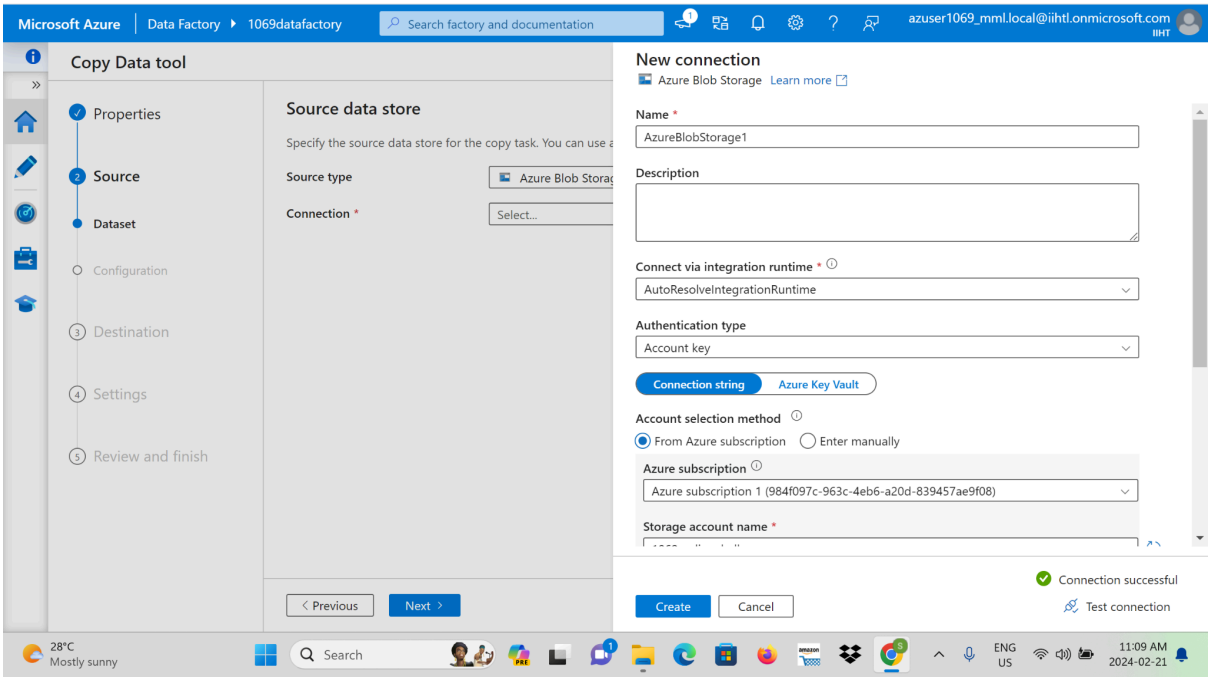


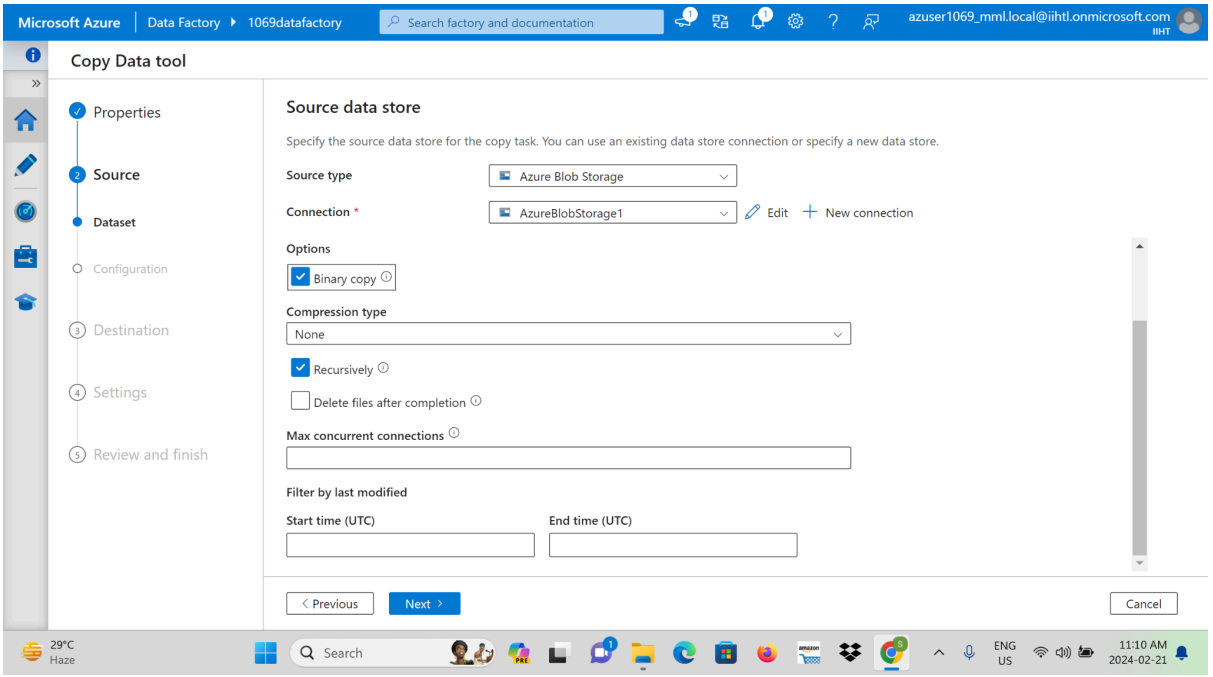
—>Then I have given all the properties and clicked on next .



—>Then I have created a source data source and then I have selected azure blob storage as source type and created a connection by giving a source storage name and source container in it.

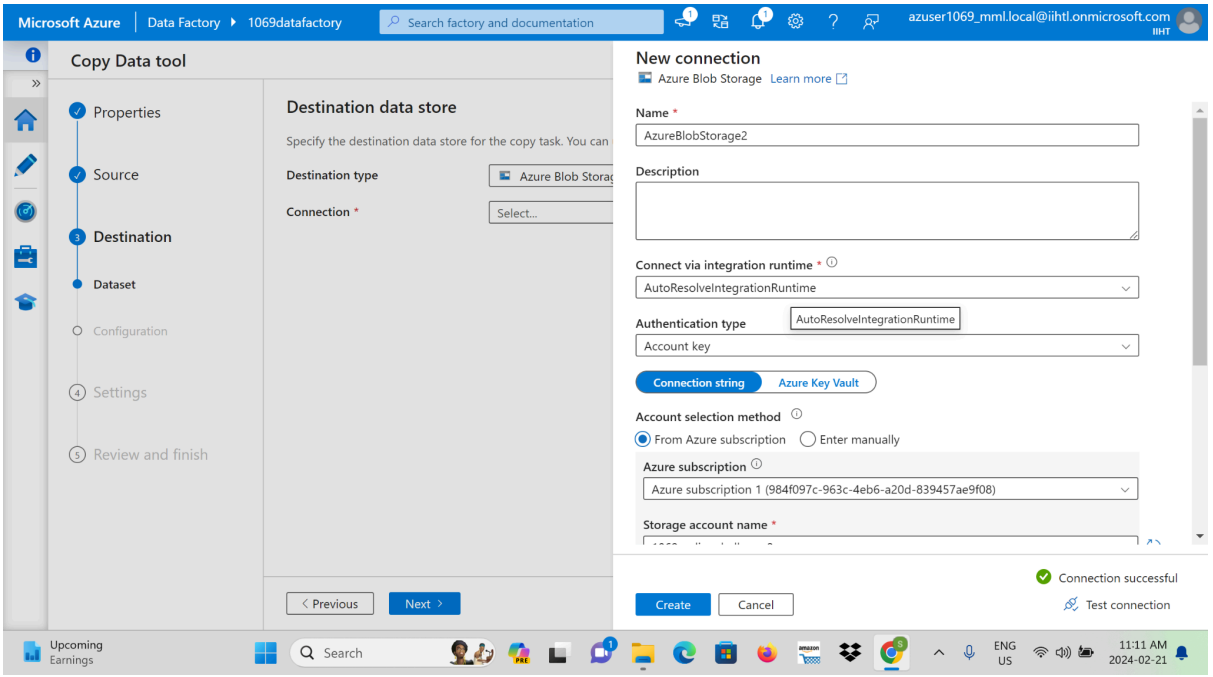


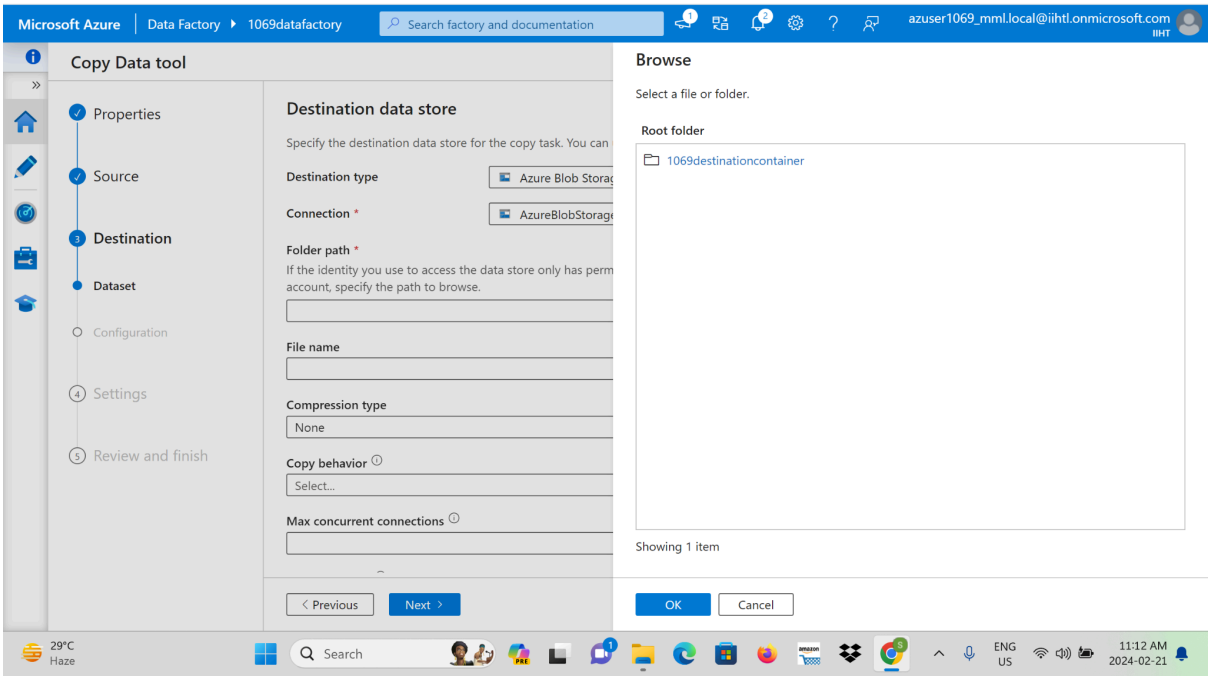


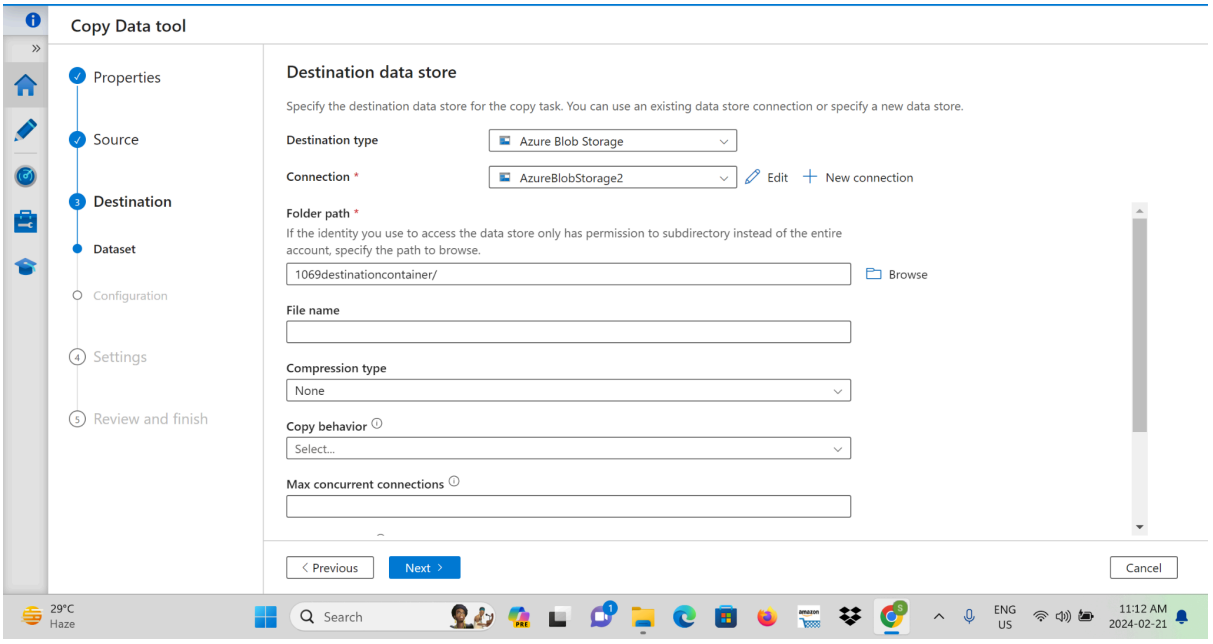


—>Later I have created a destination data source and then I have selected azure blob storage as destination type and created a connection by giving a destination storage name and destination container in it.

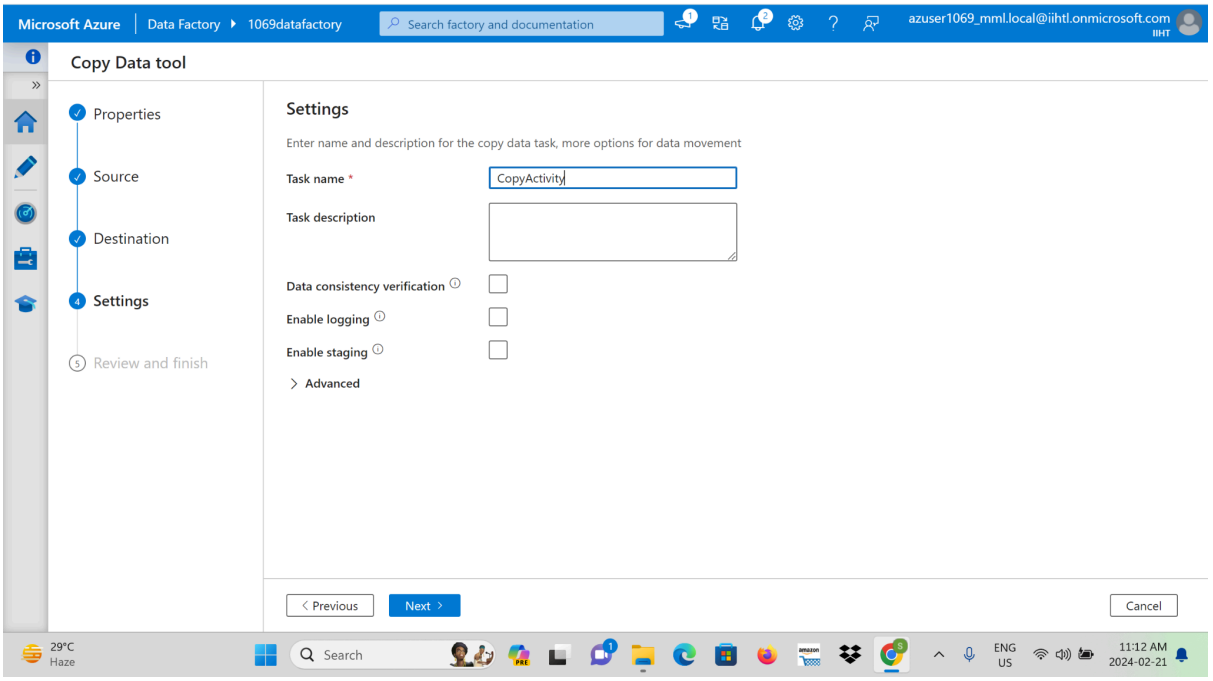
—>In the destination container there is no file present because we are copying source to destination type .So our motive is to copy the data from source and review at destination group . For that we need to give an empty destination container.



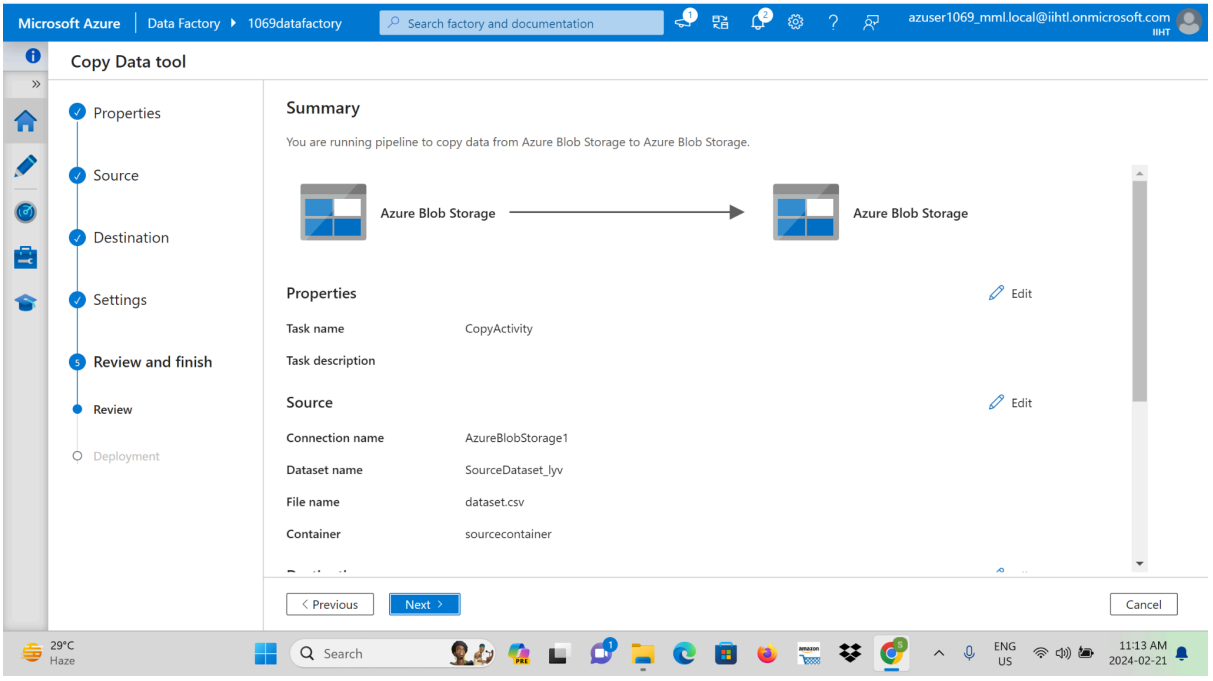

Microsoft Azure
Data Factory
1069datafactory
Search factory and documentation
azuser1069_mml.local@iihtl.onmicrosoft.com
IIHT
Copy Data tool
Properties
Source
Destination
Dataset
Configuration
Settings
Review and finish
Destination data store
Specify the destination data store for the copy task. You can
Destination type
Azure Blob Storag
Connection *
AzureBlobStorage
Folder path *
If the identity you use to access the data store only has perm
account, specify the path to browse.
File name
Compression type
None
Copy behavior
Select...
Max concurrent connections
Previous
Next
Browse
Select a file or folder.
Root folder
1069destinationcontainer
Showing 1 item
OK
Cancel
29°C
Haze
Search
ENG
US
11:12 AM
2024-02-21

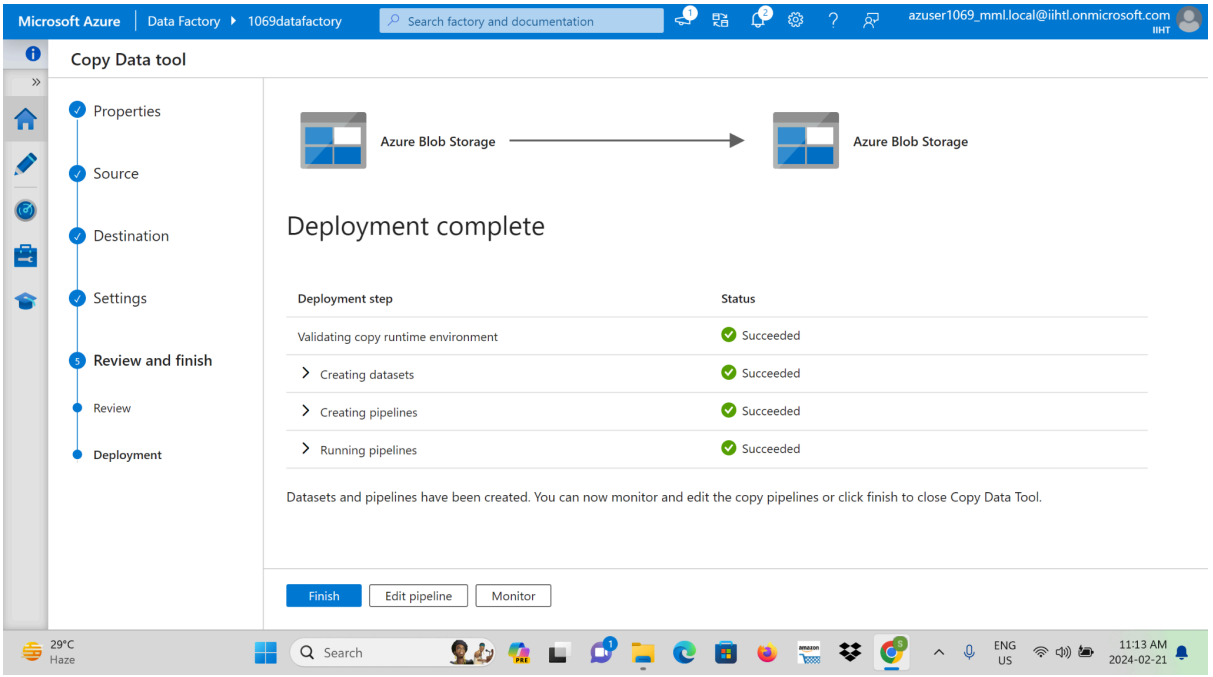

Copy Data tool
Properties
Source
Destination
Dataset
Configuration
Settings
Review and finish
Destination data store
Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.
Destination type
Azure Blob Storage
Connection *
AzureBlobStorage2
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.
1069destinationcontainer/
Browse
File name
Compression type
None
Copy behavior
Select...
Max concurrent connections
Previous
Next
Cancel
29°C
Haze
Search
ENG
US
11:12 AM
2024-02-21

—>Then we need to give the task name as “CopyActivity” for the above process and click on the next .

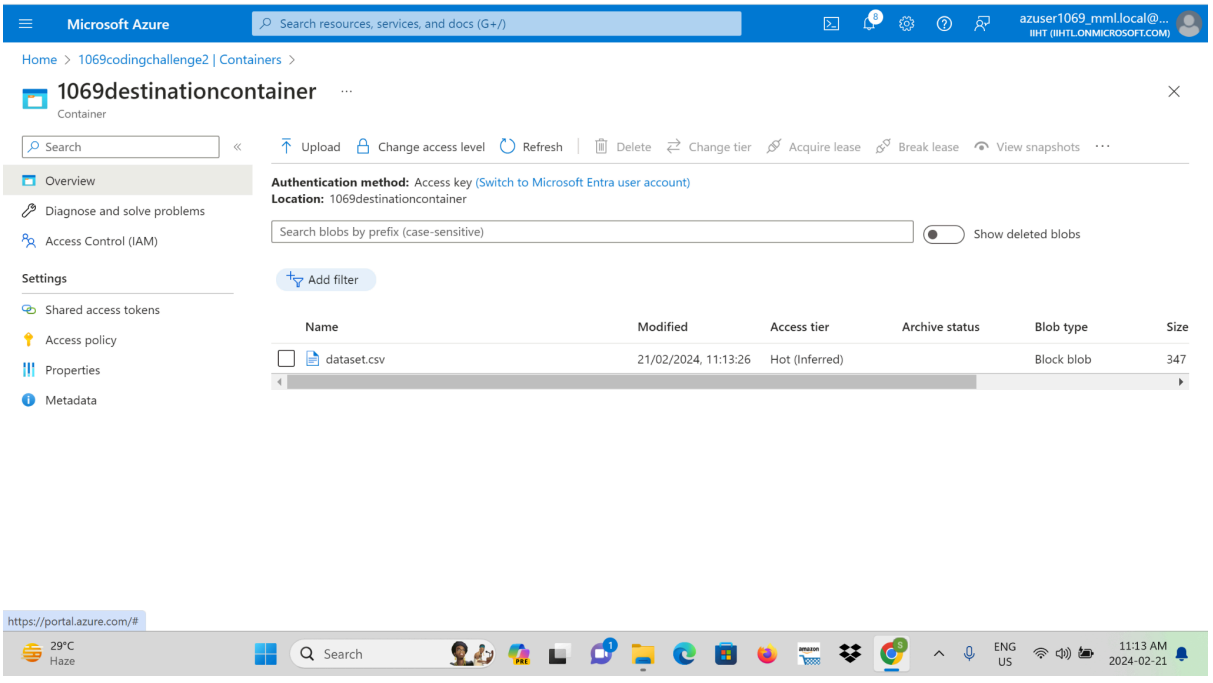


—>After giving the task name we need to review all the settings and click on finish .





—>Then we can see that datasets and pipelines are successfully created and running successfully.
—>Thereby we can see that in the destination container the file has been successfully copied from the source account .



2)Explain Overview of 3 level namespace and creating Unity Catalog objects?

A)

In Unity Catalog, the hierarchy of primary data objects flows from metastore to table or volume:

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

Catalog: The first layer of the object hierarchy, used to organize your data assets.

Schema: Also known as databases, schemas are the second layer of the object hierarchy and contain tables and views.

Tables, views, and volumes: At the lowest level in the data object hierarchy are tables, views, and volumes. Volumes provide governance for non-tabular data.

Models: Although they are not, strictly speaking, data assets, registered models can also be managed in Unity Catalog and reside at the lowest level in the object hierarchy.