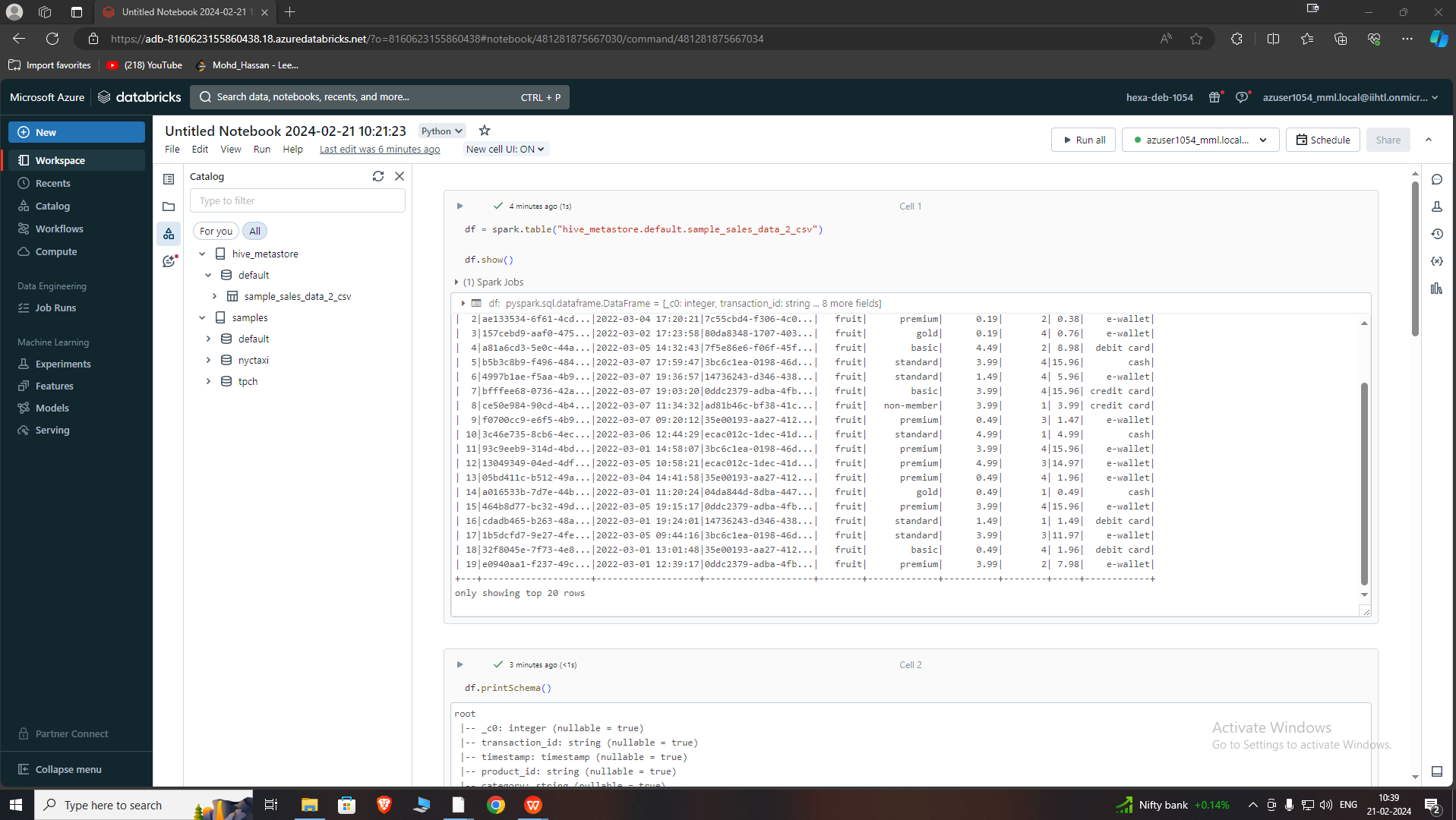
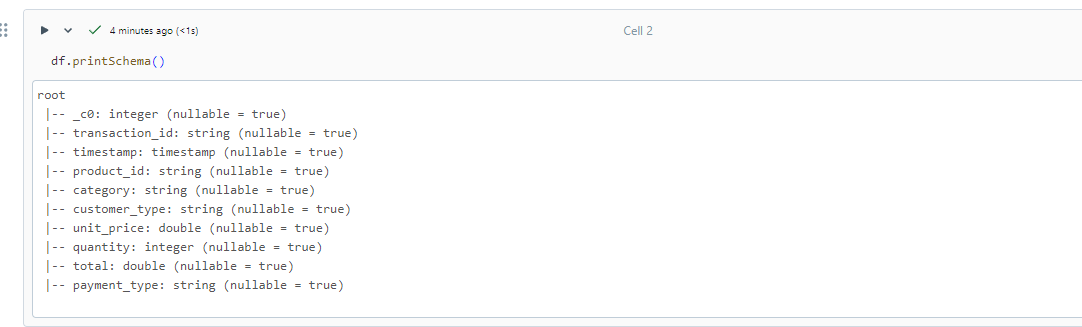
1. **Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks**

Exploratory data analysis (EDA) includes methods for exploring data sets to summarize their main characteristics and identify any problems with the data. Using statistical methods and visualizations, we can learn about a data set to determine its readiness for analysis and inform what techniques to apply for data preparation.

Reding the csv file



“printSchema” is a method used in Apache Spark's PySpark API to display the structure of a DataFrame or Dataset. It presents the column names, data types, and nullability status in a human-readable format, helping users understand the data's layout and make informed decisions during data processing.



Getting the count of customer by customer type

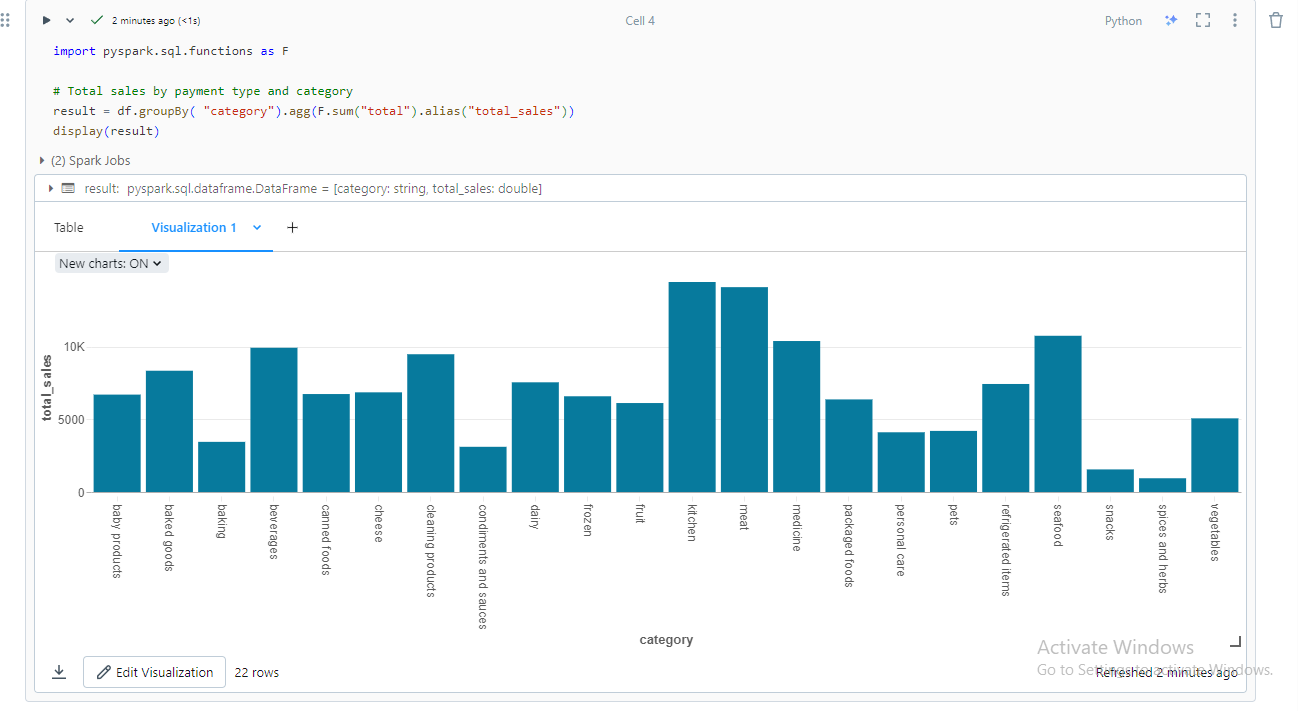


**Data Visualization**

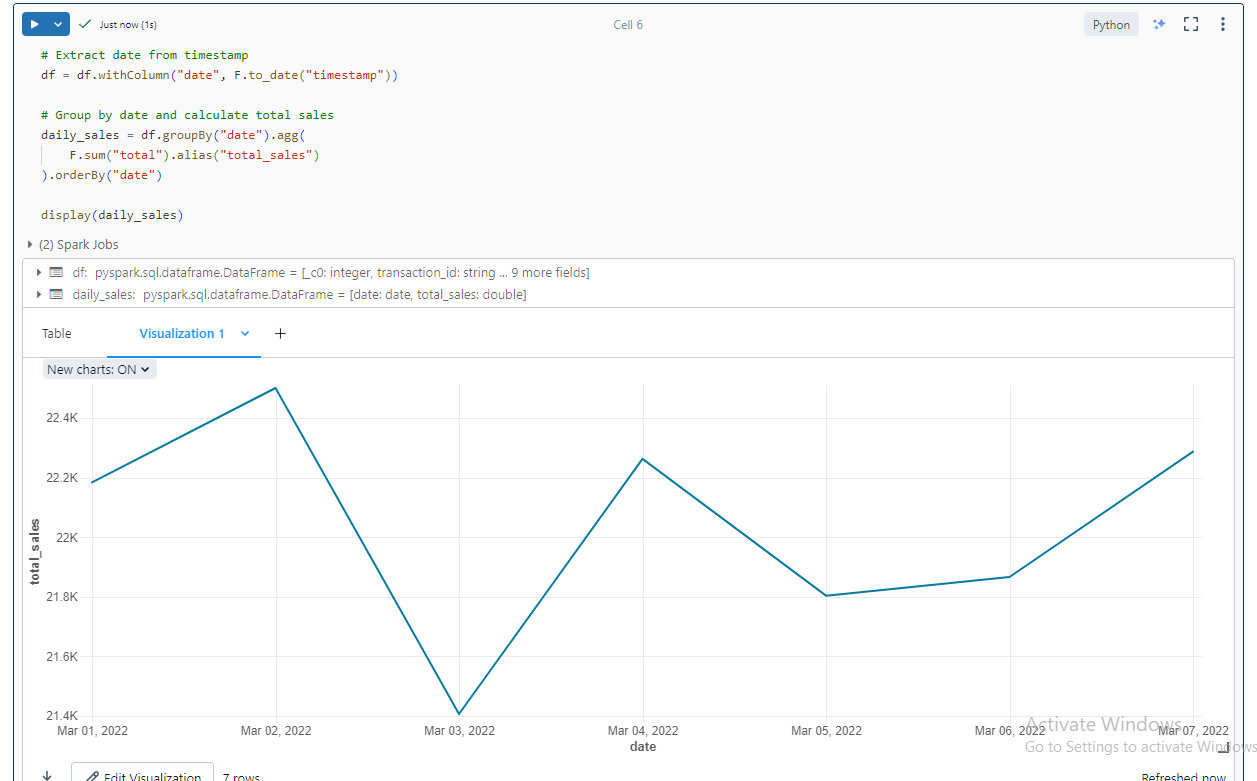
Azure Databricks has built-in support for charts and visualizations in both Databricks SQL and in notebooks.To create a visualization, click + above a result and select Visualization. The visualization editor appears.

Total Sales by Category:

We can calculate the total sales amount for each combination of category.



This code snippet extracts the date from a timestamp column in a DataFrame (df). Then, it groups the DataFrame by date, calculates the total sales for each date, orders the results by date, and displays the total sales per date.



1. Explain Overview of 3 level namespace and creating Unity Catalog objects.

## Unity Catalog Overview:

Unity Catalog is a unified governance solution for data and AI assets within the Databricks lakehouse. It provides centralized access control, auditing, lineage, and data discovery capabilities across Databricks workspaces.

Key features of Unity Catalog include:

* Define once, secure everywhere: Unity Catalog allows administrators to manage data access policies that apply consistently across all workspaces.
* Standards-compliant security model: Based on standard ANSI SQL, administrators can grant permissions at the level of catalogs, databases (schemas), tables, and views.
* Built-in auditing and lineage: Unity Catalog captures user-level audit logs and tracks how data assets are created and used across different languages.

## Three-Level Namespace:

The hierarchy of primary data objects in Unity Catalog flows from metastore to table or volume:

1. **Metastore:** The top-level container for metadata. Each metastore exposes a three-level namespace: catalog.schema.table.
2. **Catalog:** The first layer of the object hierarchy, used to organize data assets.
3. **Schema:** Also known as databases, schemas form the second layer and contain tables and views123.

to create Unity Catalog objects, specifically focusing on models within the three-level namespace:

Model Creation:

* A model resides in the third layer of the three-level namespace: catalog.schema.table.
* In the context of Unity Catalog, a “model” refers to a machine learning model registered in the MLflow Model Registry.
* To create a model in Unity Catalog, users must have the CREATE MODEL privilege for the catalog or schema.

3.Execute & explain, Azure datafactory and its copy activity.

Azure Data Factory is a cloud-based service designed for creating data-driven workflows that streamline data movement and transformation tasks. Unlike storing data internally, it focuses on orchestrating and automating these processes across various data stores. With Azure Data Factory, you can seamlessly manage the movement of data between supported data sources and utilize compute services for data processing, whether it's in other regions or within an on-premise setup. Additionally, the service offers comprehensive monitoring and management capabilities, accessible through both programmatic interfaces and user-friendly UI tools.

Azure Data Factory use cases . ADF can be used for:

* Supporting data migrations
* Getting data from a client’s server or online data to an Azure Data Lake
* Carrying out various data integration processes
* Integrating data from different ERP systems and loading it into Azure Synapse for reporting

**Executing Datacopy activity on Azure Data Factory**

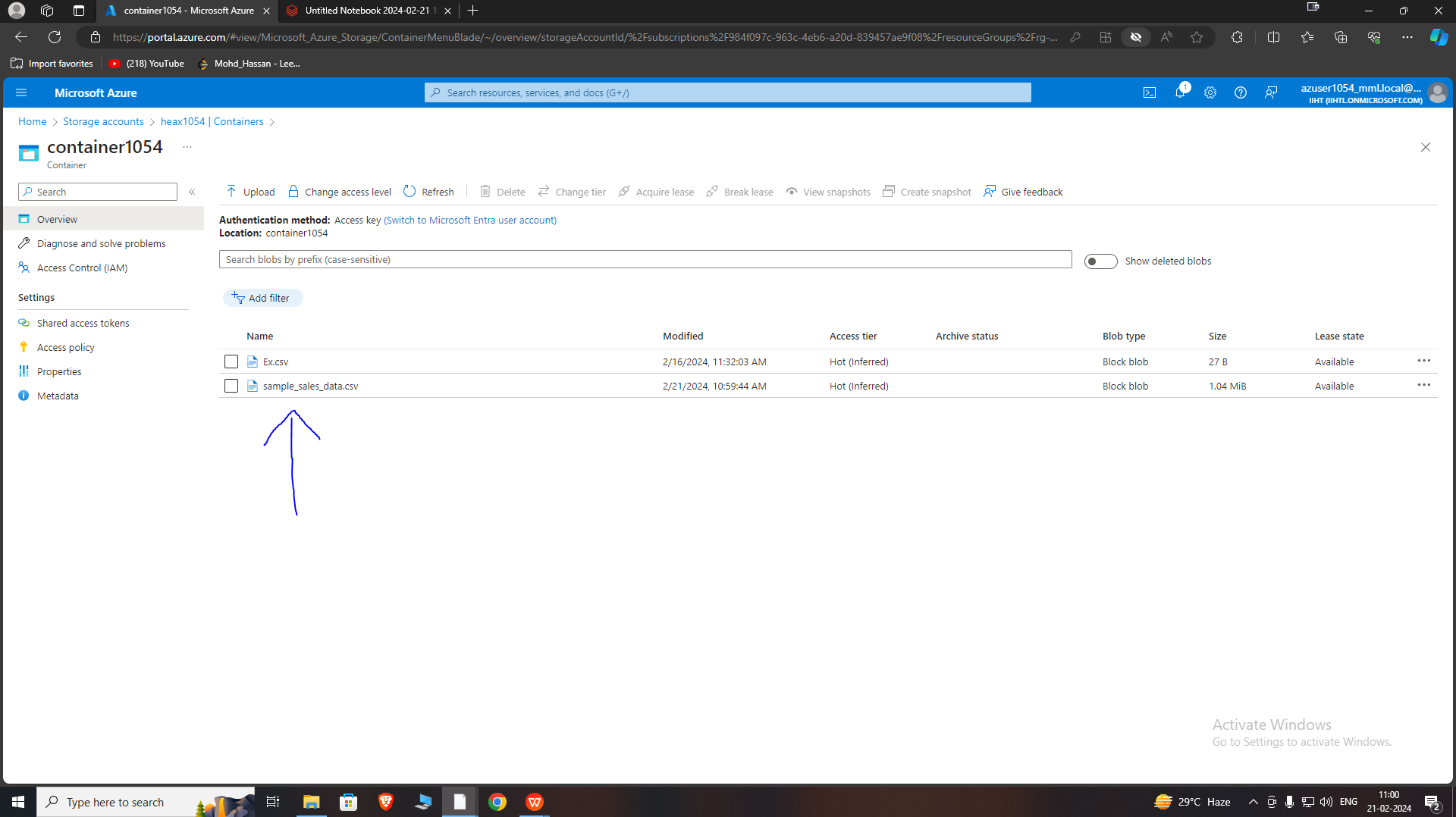
To start migrating the data on Blob storage to Azure SQL, the most simple way is to use Data Copy Wizard, which is currently in preview. It allows you to quickly create a data pipeline that copies data from a supported source data store to a supported destination data store

Copying the file ‘samples\_sales\_data.csv’ from azure blob storage account named “heax1054” to azure datalake gen 2 storage account named “datalake1054”

Azure blob storage ------> Azure data lake gen2

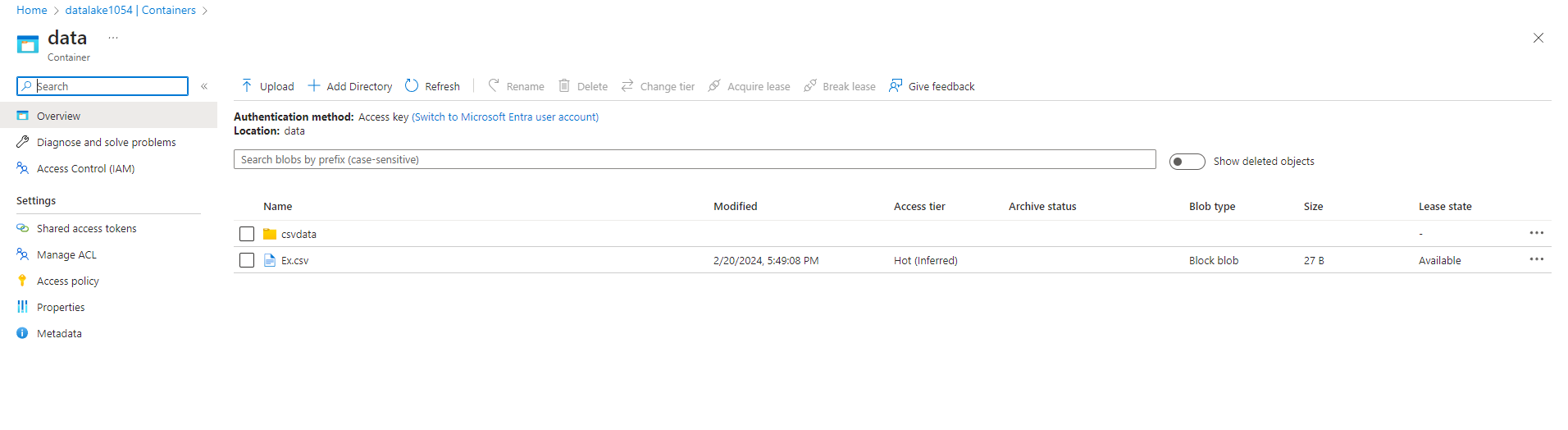
heax1054 -------------------> datalake1054

**Azure Blob Storage**



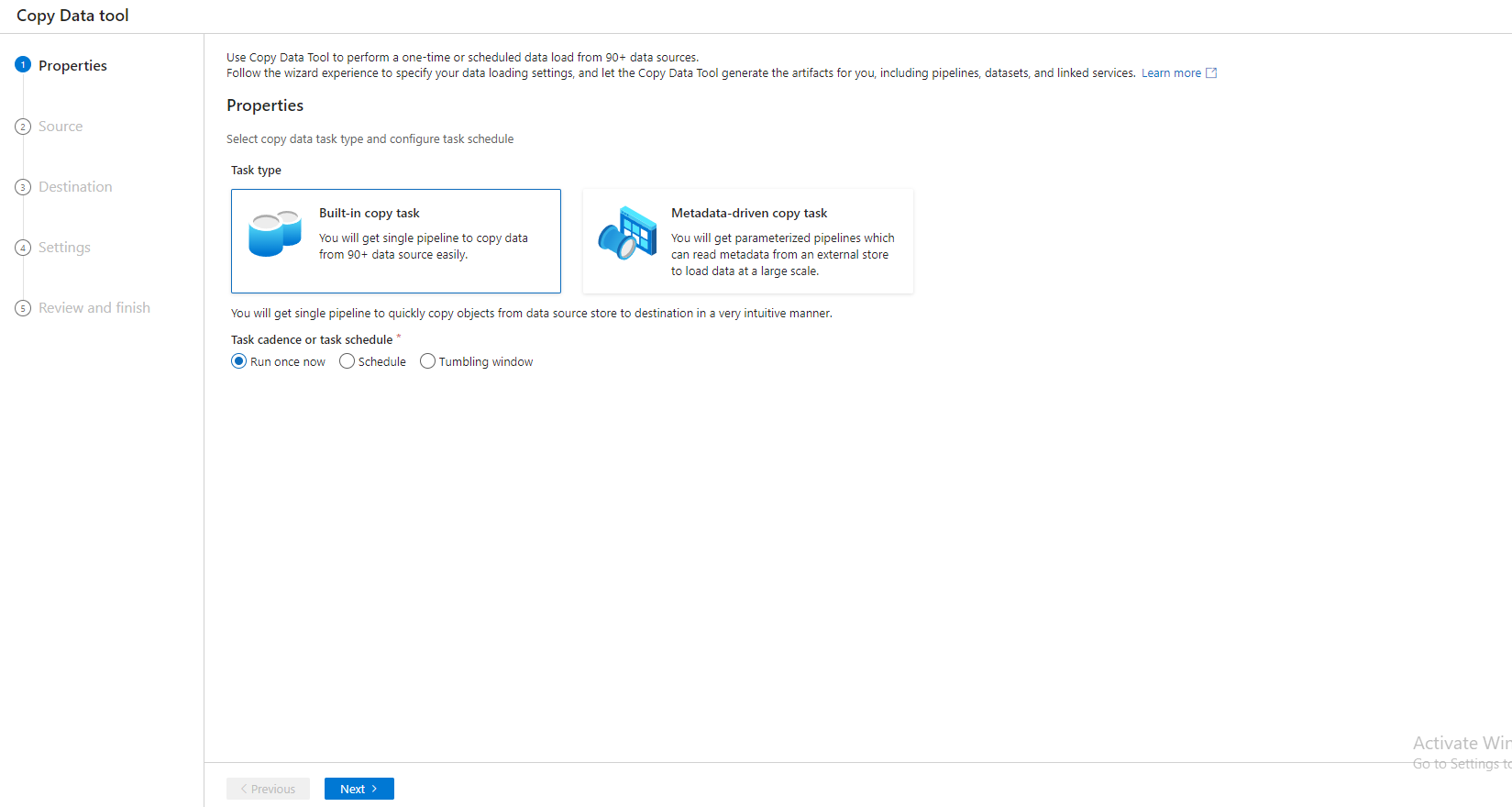
Here , we can see the ‘samples\_sales\_data.csv’ in the container 1054 which is inside the azure blob storage account named heax1054

## Datalake

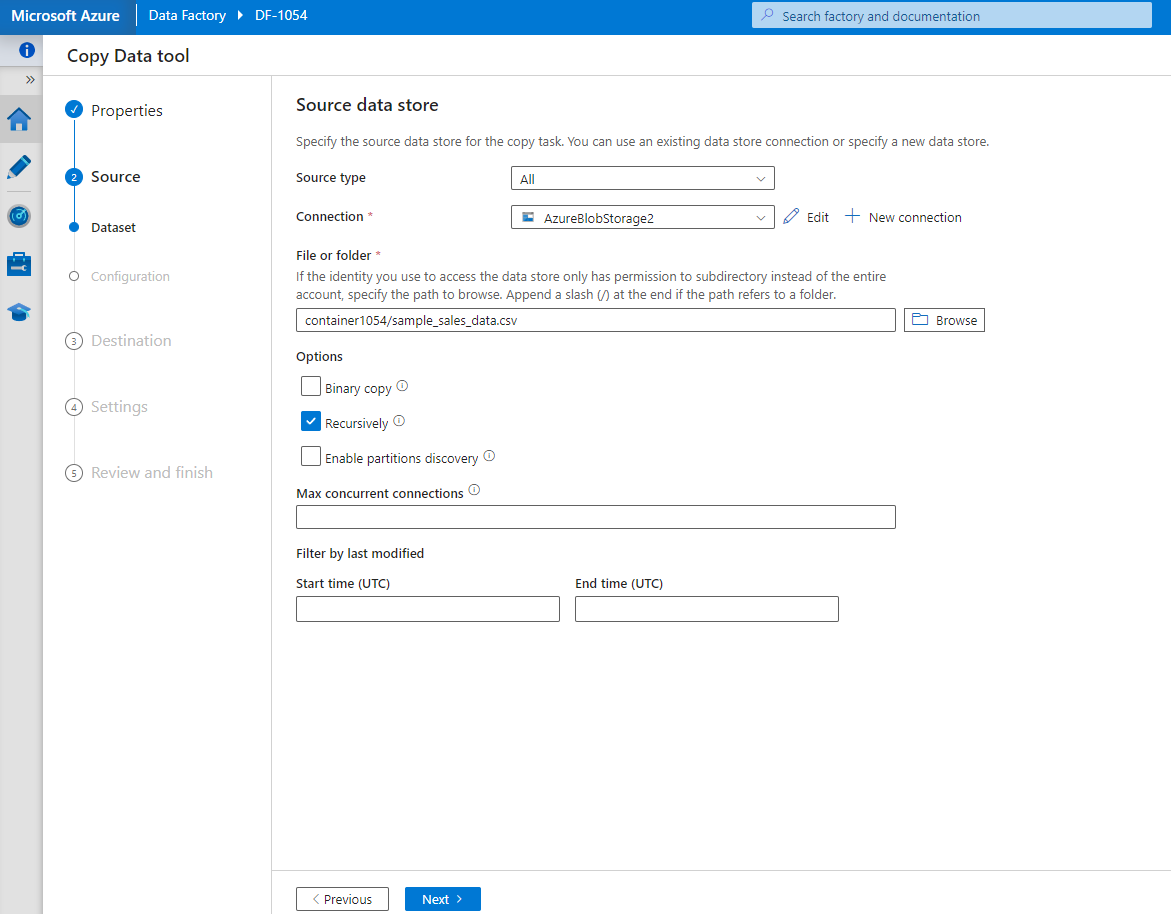


Here , we can see the target location in the data which is inside the azure datalake gen2 storage account named “datalake1054”

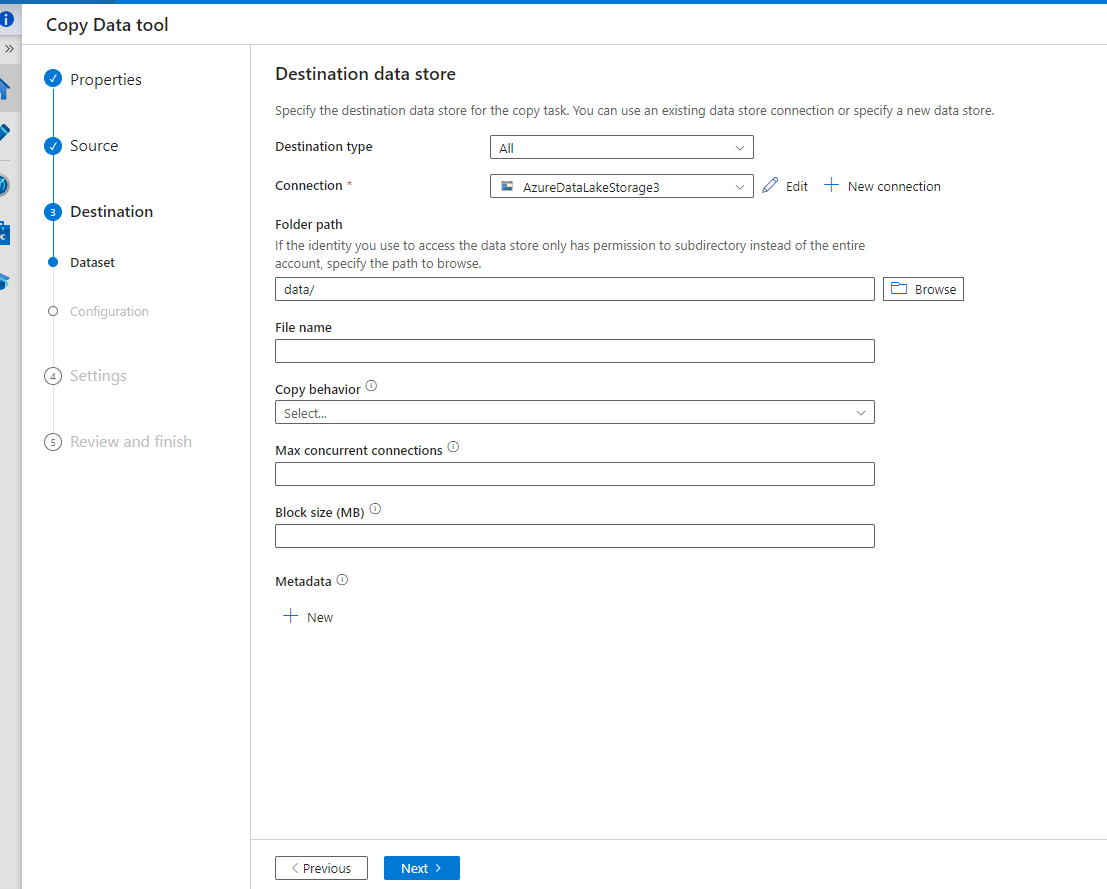
**Process of Ingesting**

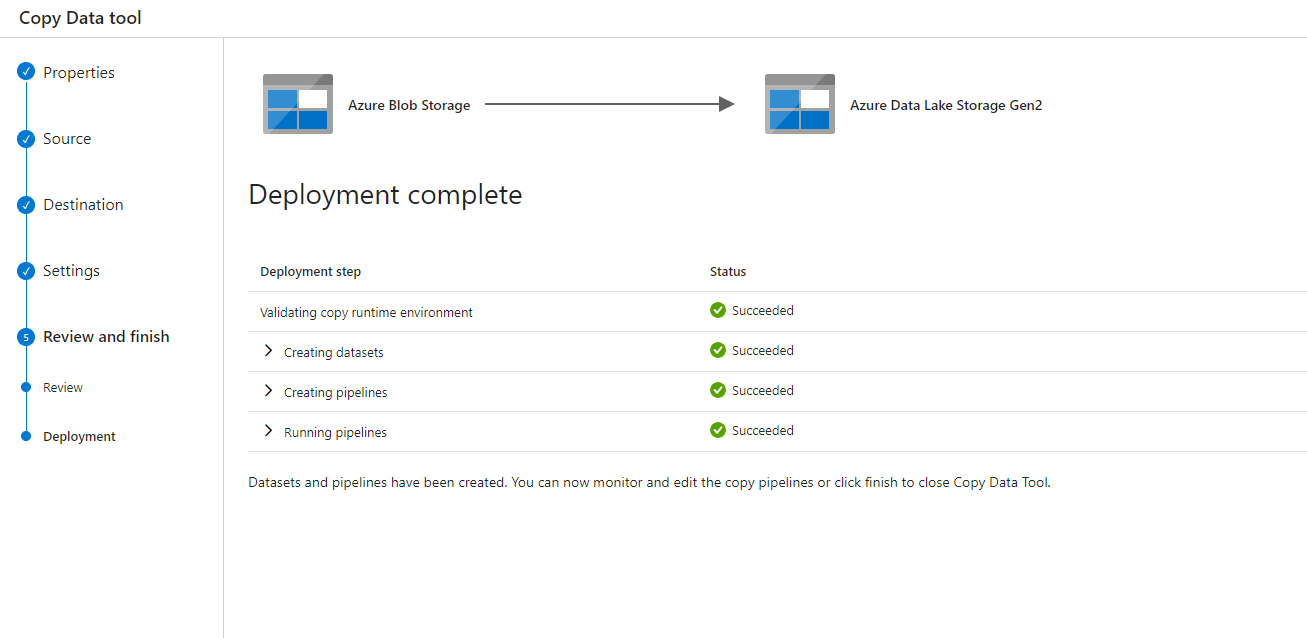


Making connection with the blob storage and selecting the file

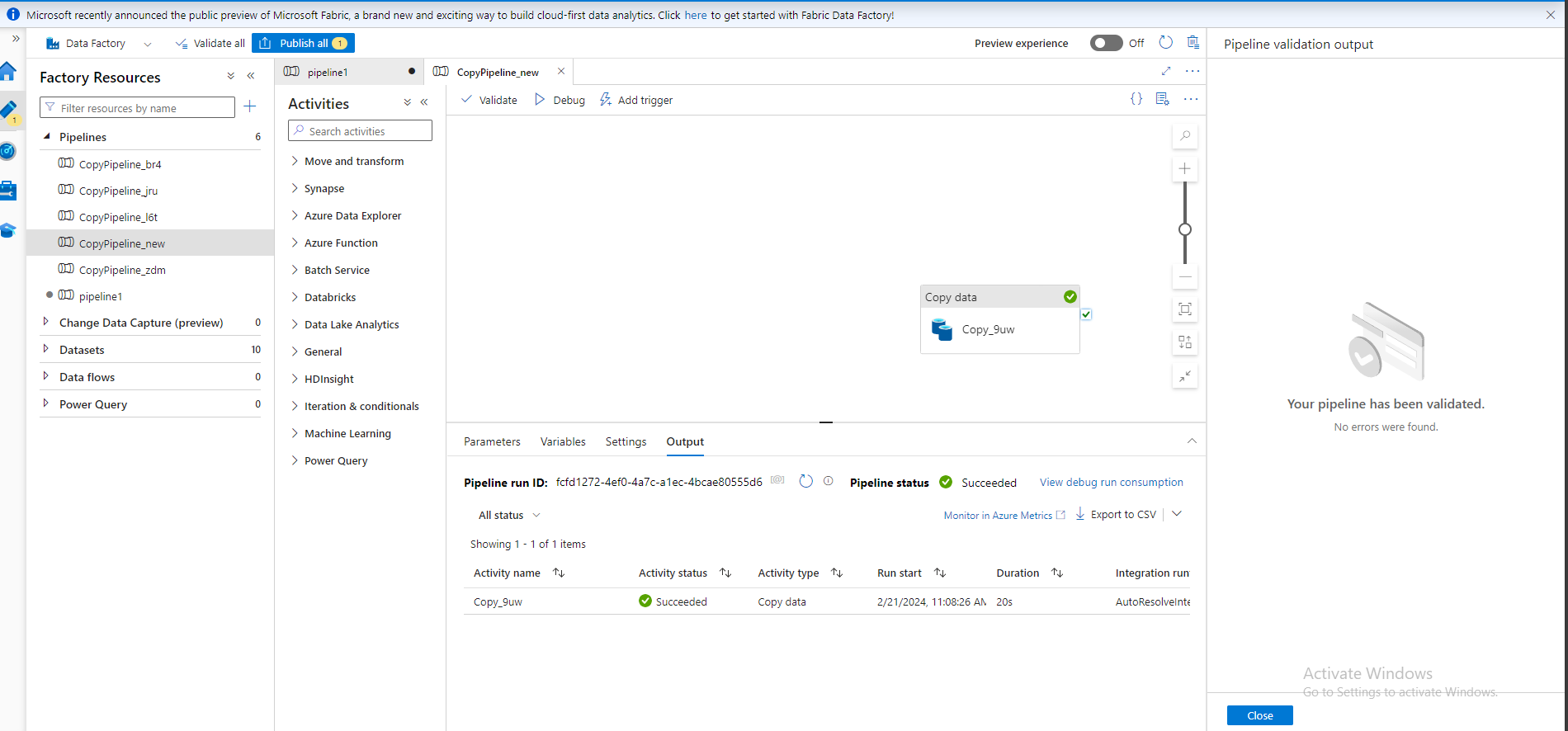


Making connection with the Datalake storage and specifying the target location





**Validating and debugging the pipeline**



**copy activity successfully completed the file was copied at the target location in the datalake1054 storage account**

