Name : Parth Nandedkar
Date : 21 Feb 2024
Topics : Azure Databricks
Batch : Data Engineering Batch-1

Q3.Execute & explain, Azure datafactory and its copy activity.

Azure Data Factory is a cloud-based data integration service that allows you to create, schedule, and orchestrate data workflows at scale. It enables you to **collect, transform, and move data** from various sources to different destinations, both within Azure and across on-premises and other cloud environments. Here's an overview of Azure Data Factory and its key components:

Pipelines:

Pipelines are the core construct in Azure Data Factory. They represent a series of data processing steps that perform operations on data. These steps can include data movement, transformation, data loading, and data orchestration activities.
Pipelines are composed of activities, which are the building blocks for performing tasks like copying data from a source to a destination, transforming data using Azure services like HDInsight or Azure Databricks, executing SQL scripts, etc.

Activities:

Activities are the individual processing steps within a pipeline. They represent the actions that are performed on data, such as copying data from a source to a destination, running a Hive query, executing a stored procedure, etc.
Azure Data Factory provides a wide range of built-in activities for common data integration tasks, and you can also create custom activities using Azure Functions or Azure Batch if needed.

Datasets:

Datasets represent the structure of the data being processed within Azure Data Factory. They define the schema and location of the data, whether it's in files, databases, or other data stores.
Azure Data Factory supports various types of datasets, including Azure Blob Storage, Azure Data Lake Storage, Azure SQL Database, Azure Synapse Analytics, and many others.

Linked Services:

Linked Services are connections to external data sources or destinations used by Azure Data Factory activities. They contain the information needed to connect to the data stores, such as connection strings, credentials, and authentication methods.
Linked Services are defined separately from datasets and can be reused across pipelines.

Triggers:

Triggers are used to schedule the execution of pipelines in Azure Data Factory. There are different types of triggers available, including schedule-based triggers, event-based triggers, and tumbling window triggers. Schedule-based triggers allow you to specify a recurrence pattern for running pipelines at regular intervals, while event-based triggers respond to events such as the arrival of new data or the completion of a data processing task.

Integration Runtimes:

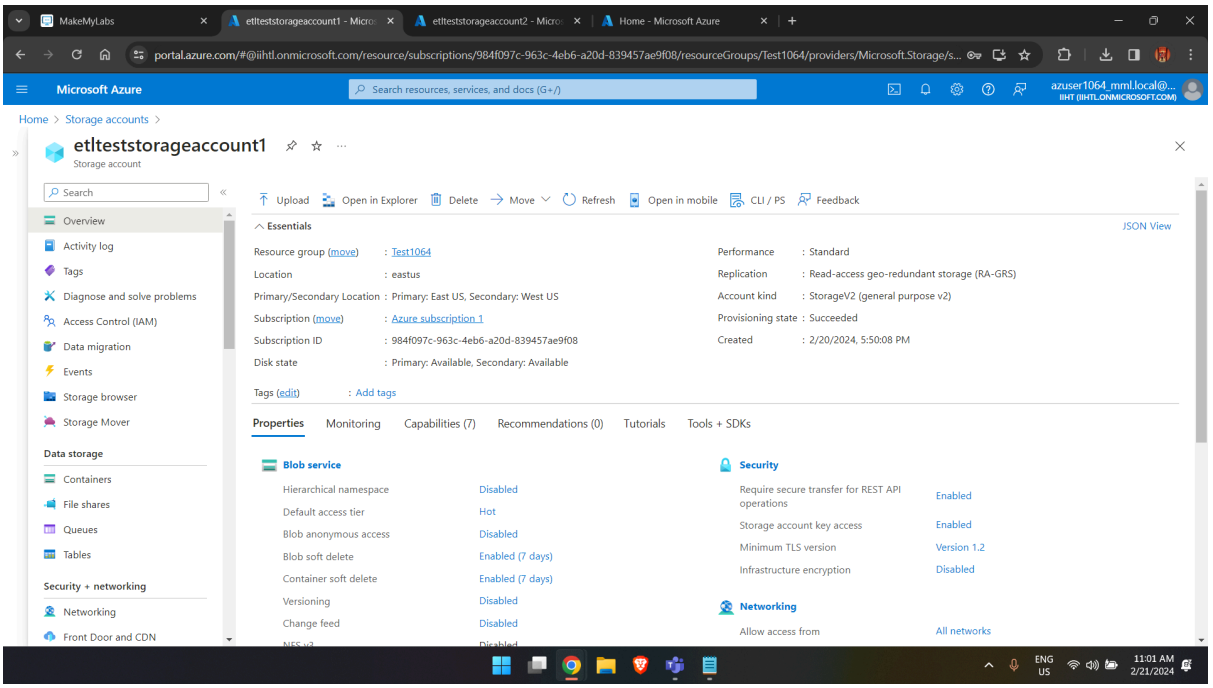
Integration Runtimes are compute environments used by Azure Data Factory to execute data integration tasks. They provide the resources and capabilities needed to connect to data sources, perform data transformations, and move data between different environments.
Azure Data Factory supports three types of integration runtimes: Azure, Self-hosted, and Azure-SSIS (SQL Server Integration Services).

Azure Data Factory simplifies the process of building, deploying, and managing data integration workflows in the cloud, enabling organizations to efficiently collect, process, and analyze data from diverse sources to gain valuable insights and drive business decisions.

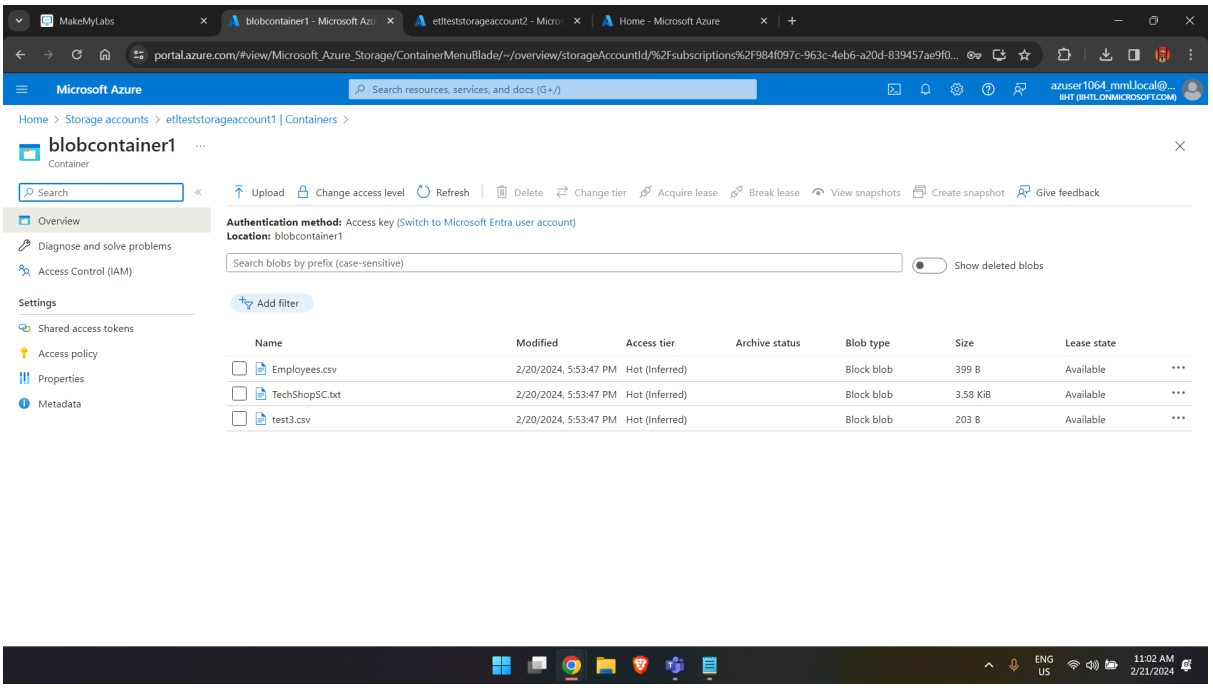
Executing Copy Activity :

Blob to Blob Copy :

Creating two blob storage accounts :

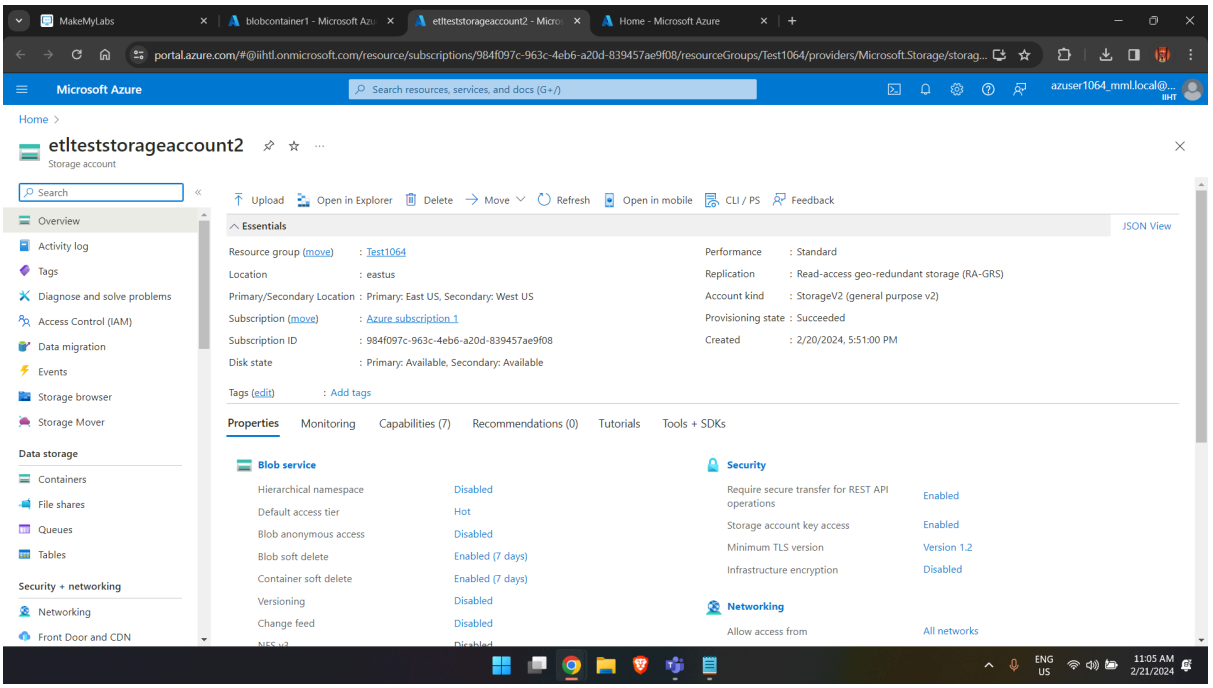


This account is for copying purposes so we will add sample data files in it.

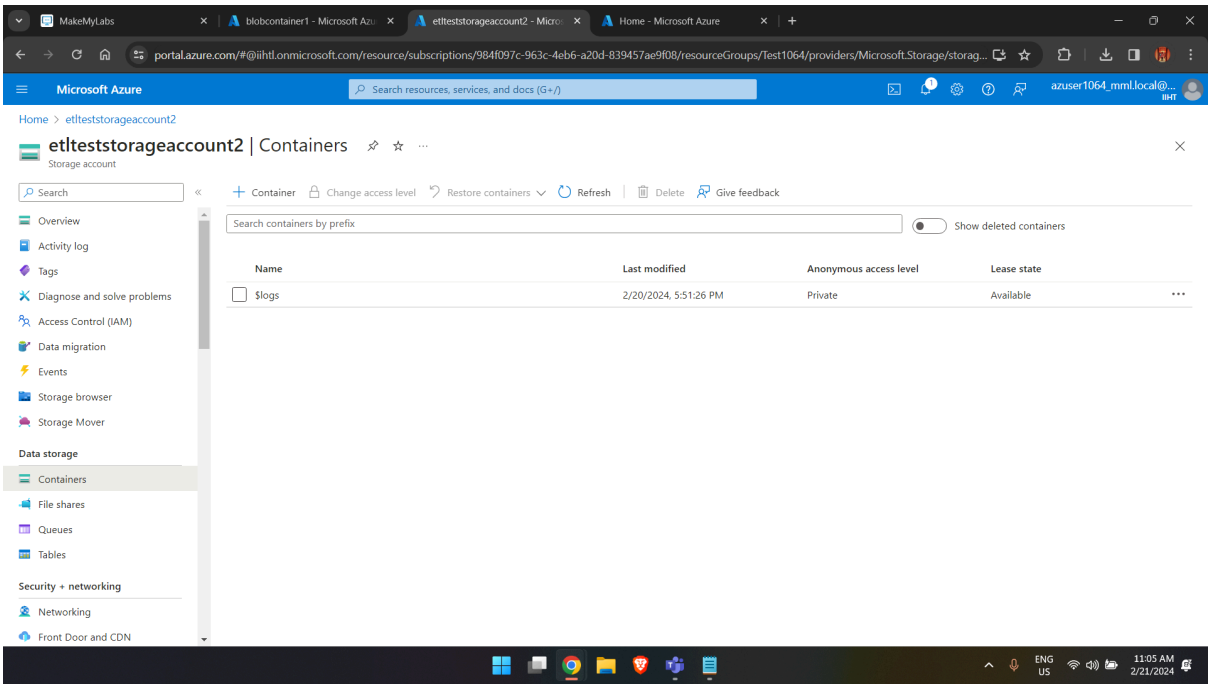


Added the sample data in blobcontainer1. Three files are added as we can see.

Creating a second blob storage account for copying data to it.

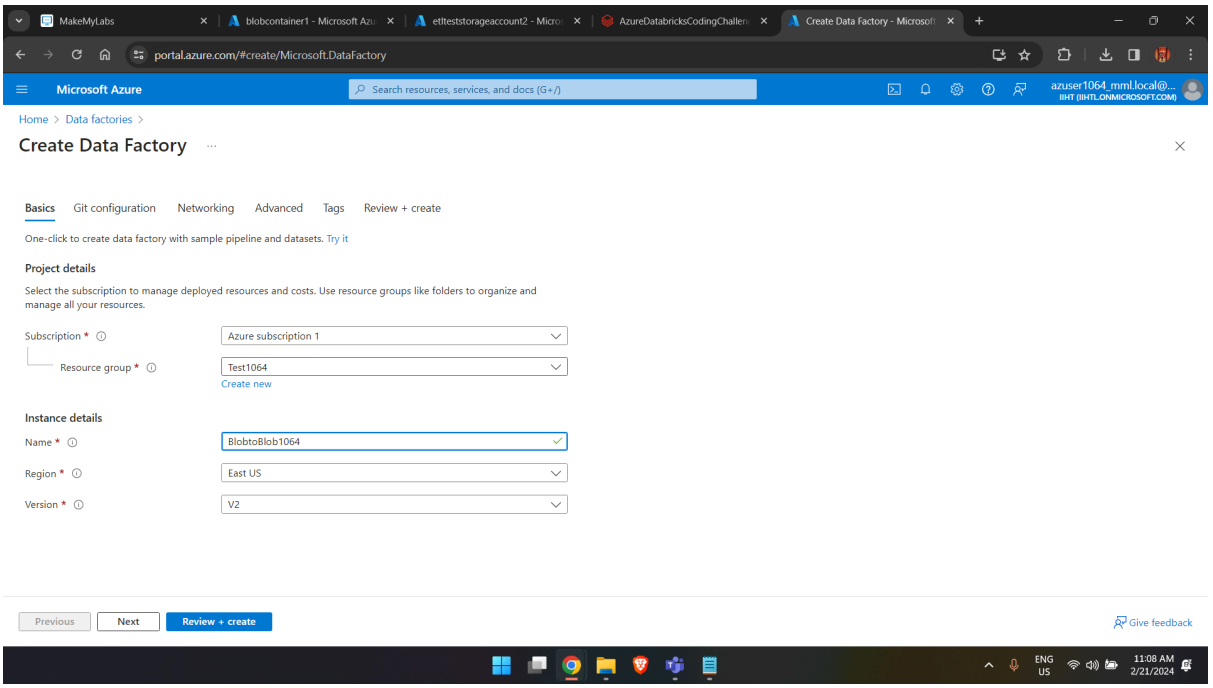


Storage account2 is created.

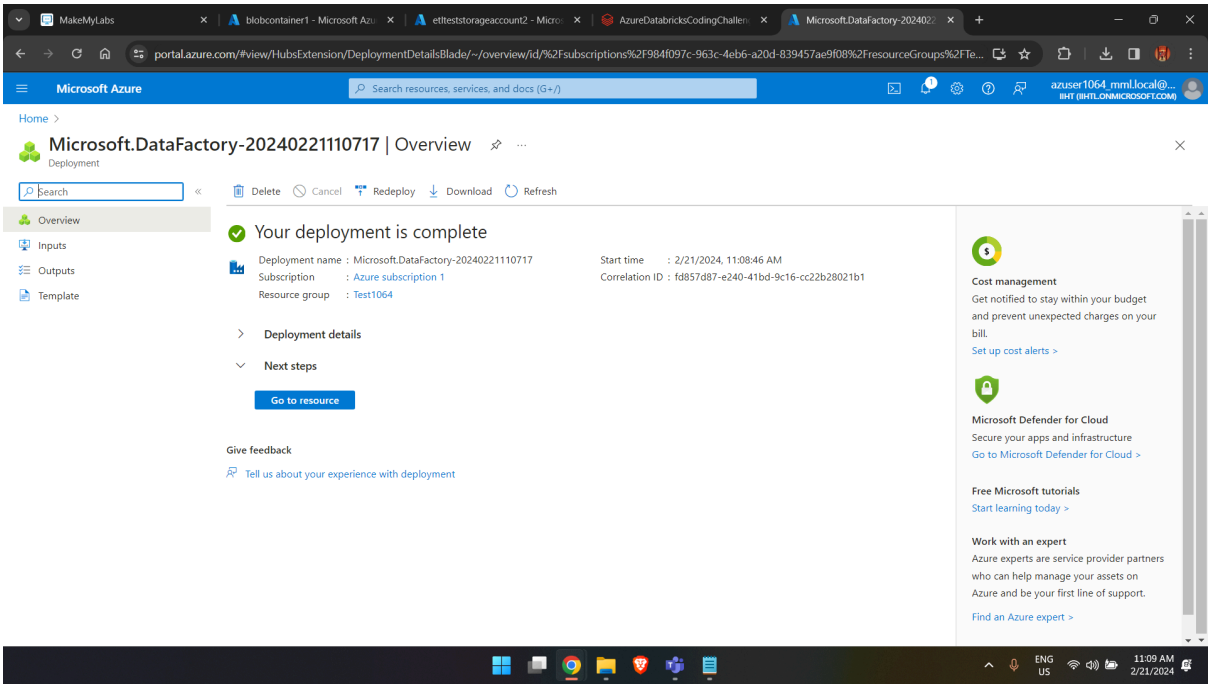


Here there is no container.

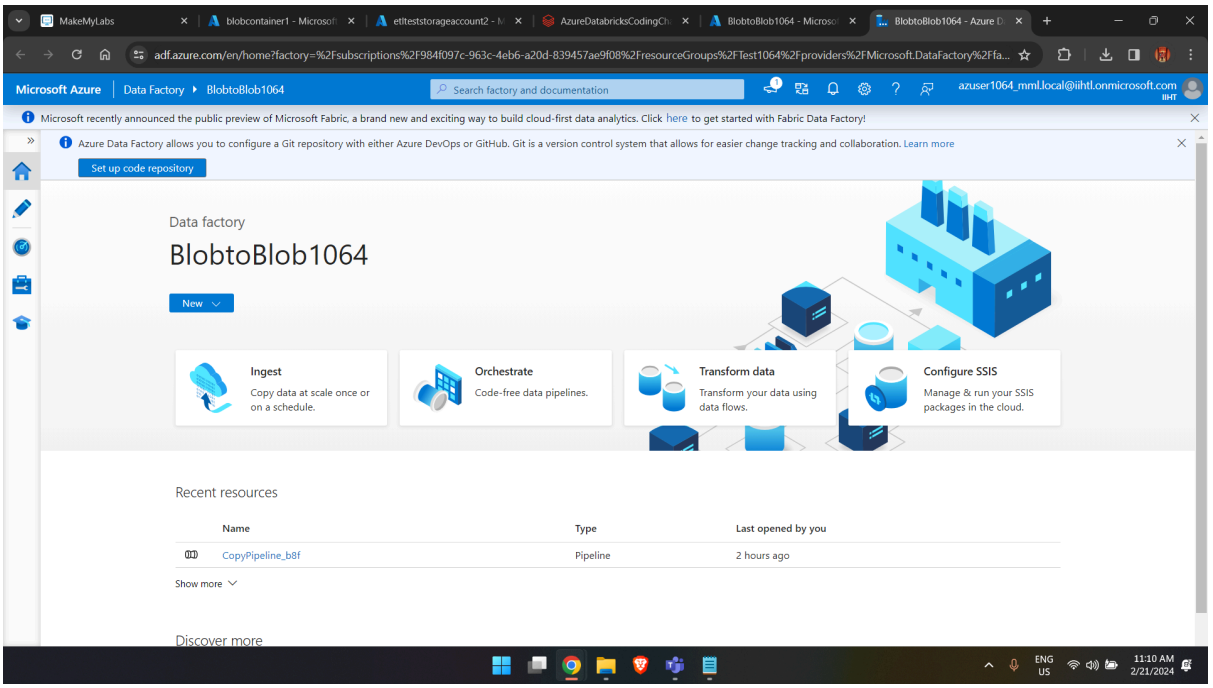
Creating new Data Factory :



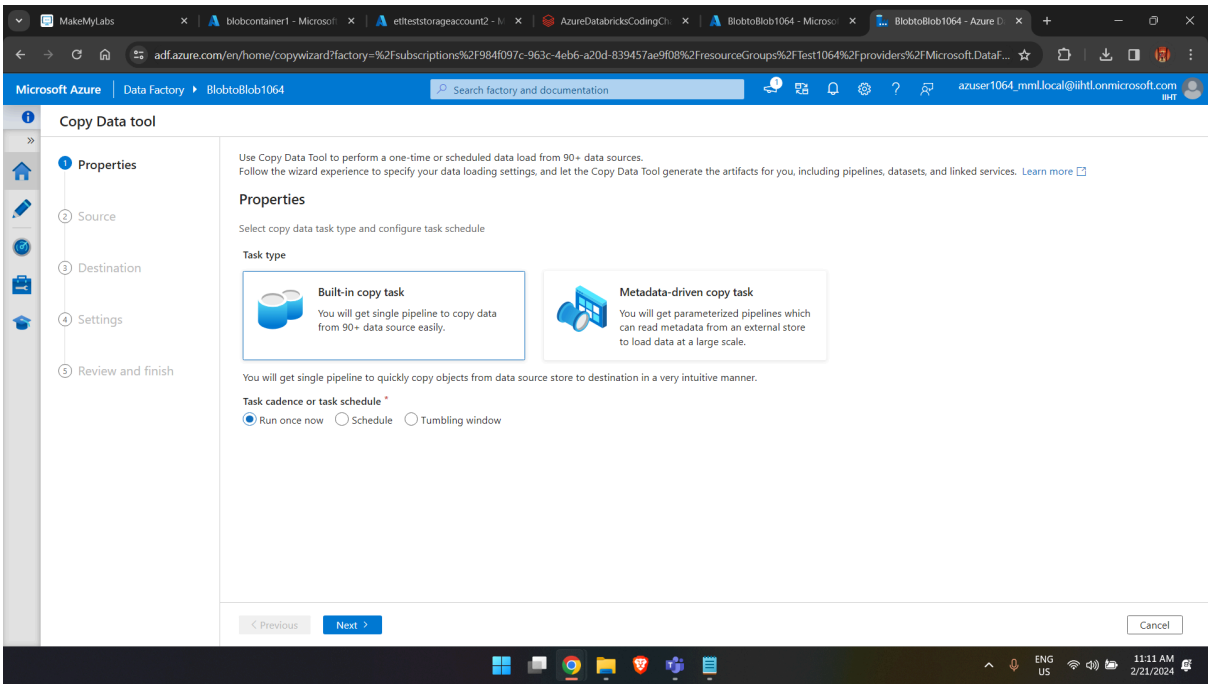
Created a datafactory.



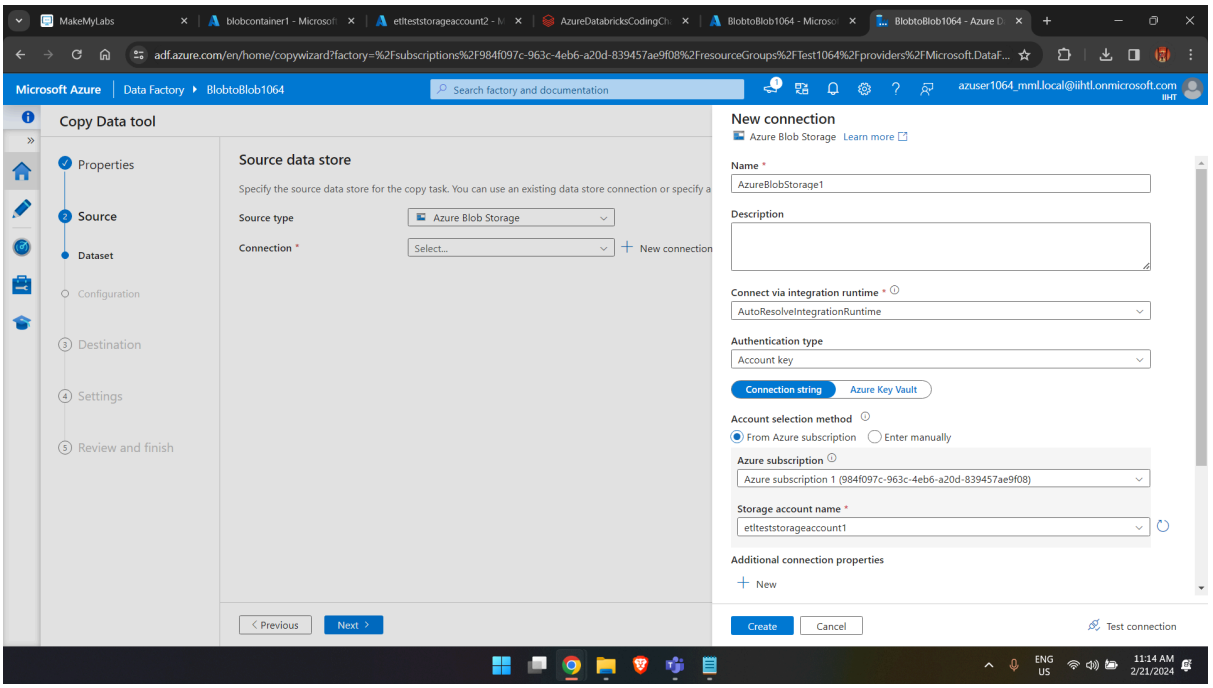
After launching the studio. UI of studio will be generated .



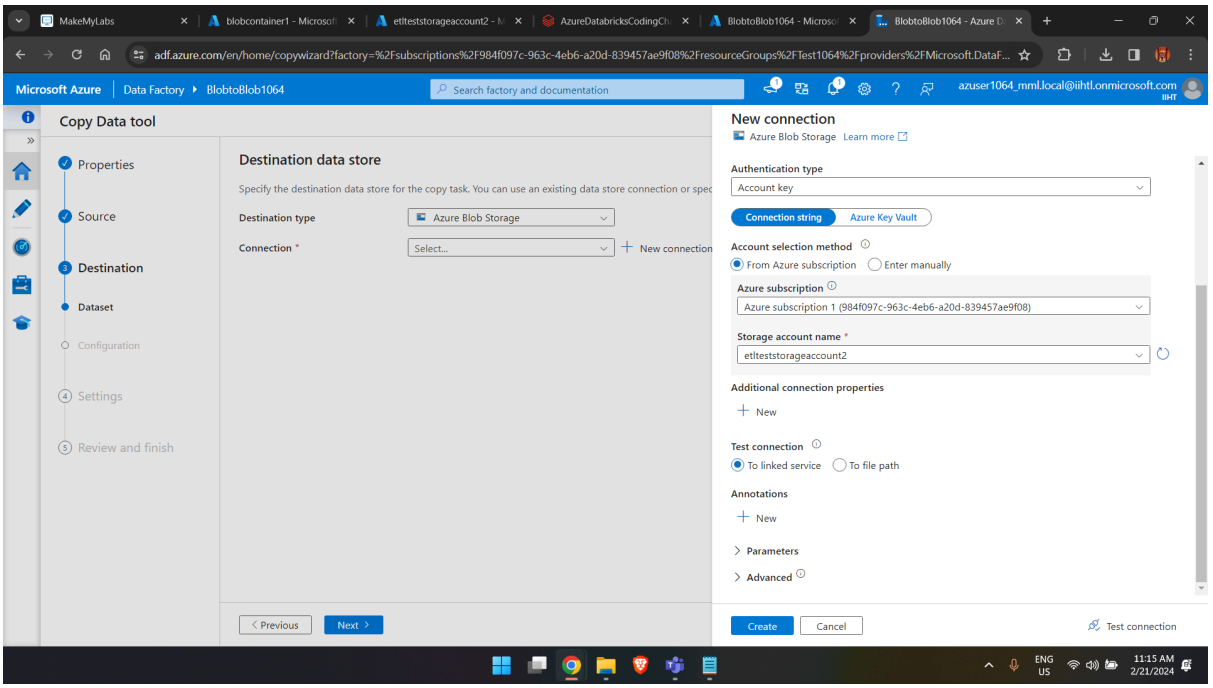
Performing Copy Task:



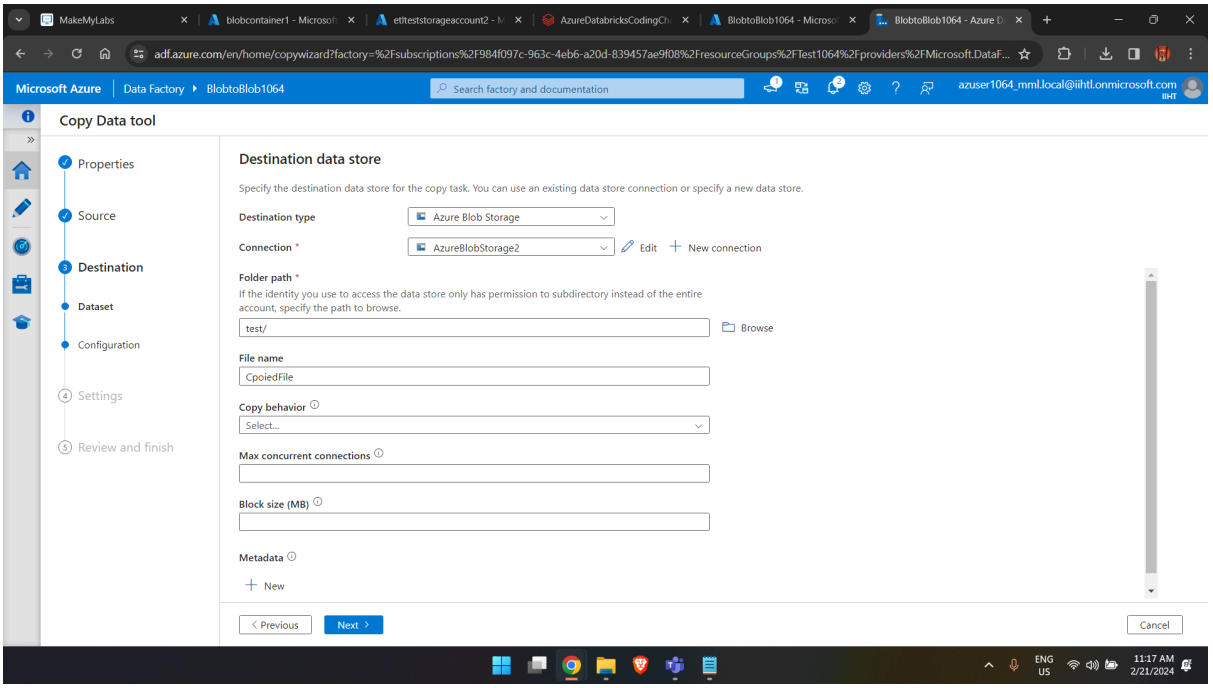
Connecting the source file :



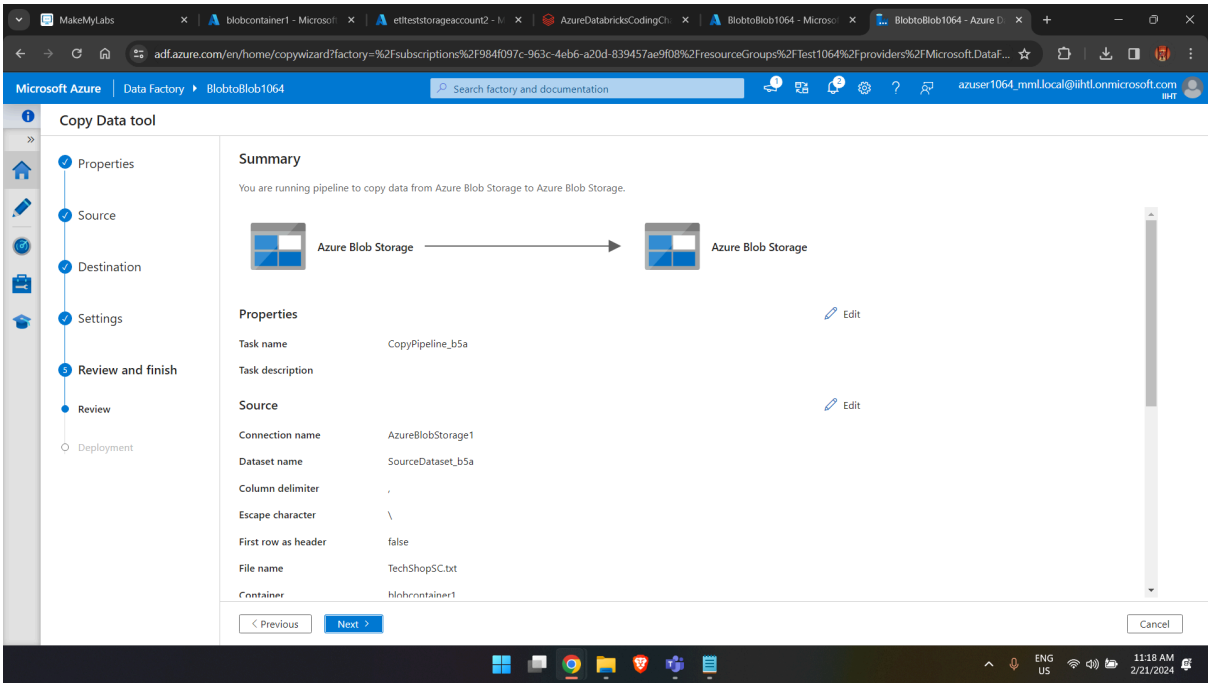
Connecting Destination Storage Account :



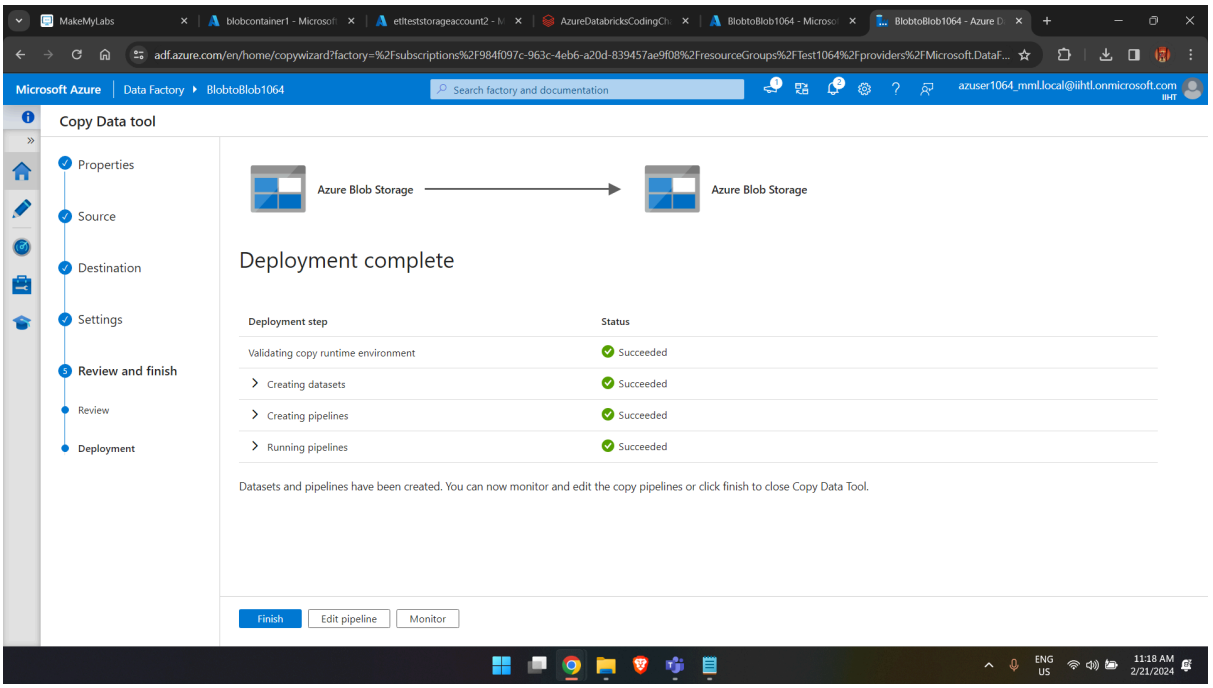
Adding Destination Path :

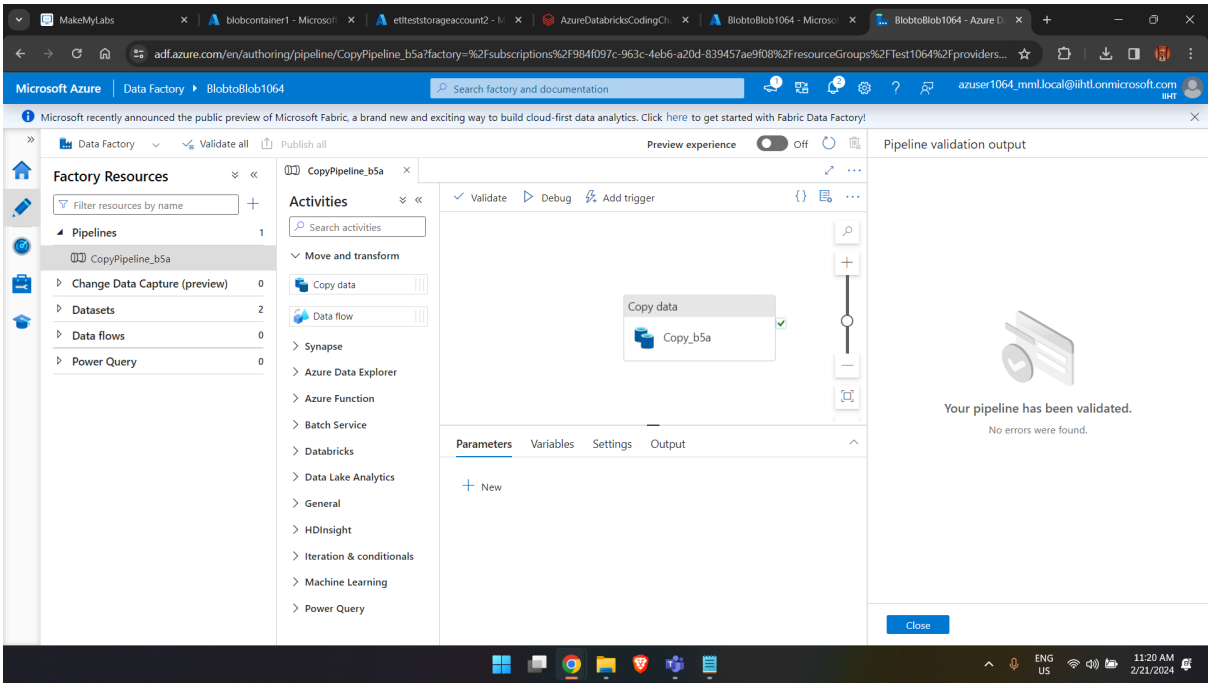


Reviewing the process :



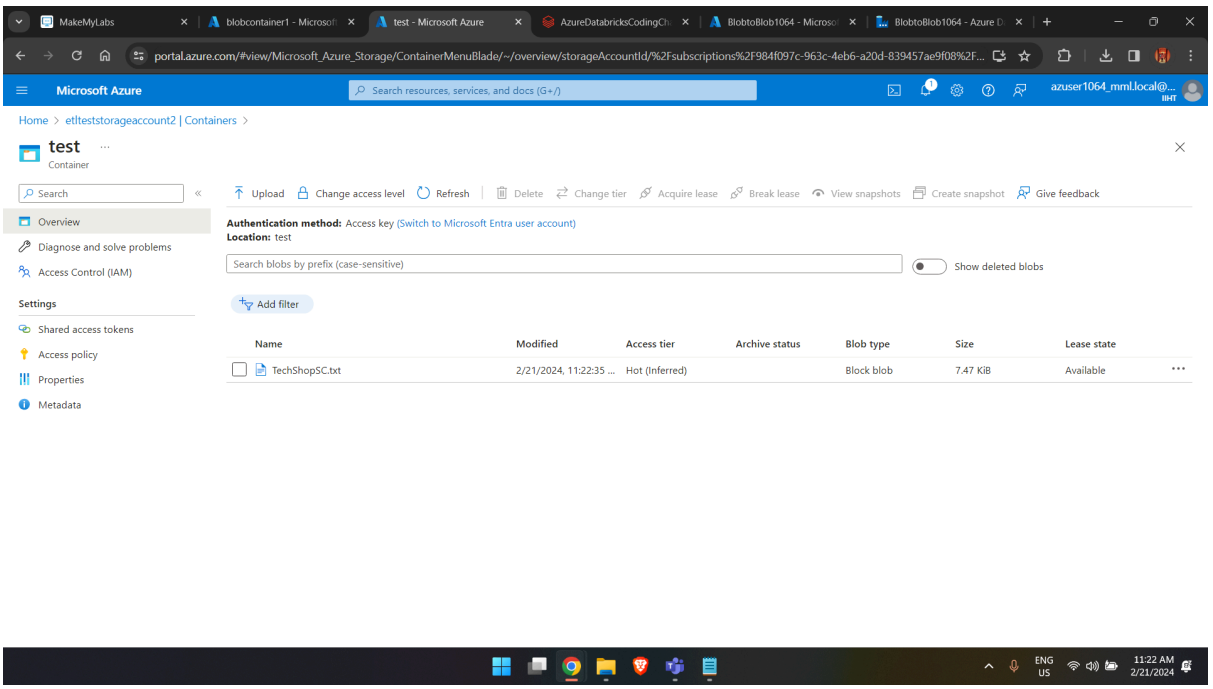
Creating pipeline :





Above Screenshot tells that the pipeline worked successfully.

Checking whether changes are reflected in storage account2 or not.



We can see the file is copied successfully from one blob storage account to another blob storage account.

From source to destination the data copied successfully by using azure data factory like that we can use this service for adls-adls, adls-blob, blob-adls.