|  |  |
| --- | --- |
| NAME | KAILASH S |
| DATE | 05-01-24 |
| ASS NO | 7 |
| TOPIC | Delta lakehouse, unity catalog, copying of files using azure data factory |

**DELTA LAKE**:

Delta Lake is the optimized storage layer that provides the foundation for storing data and tables in the Databricks lakehouse. Delta Lake is [open source software](https://delta.io/) that extends Parquet data files with a file-based transaction log for [ACID transactions](https://docs.databricks.com/en/lakehouse/acid.html) and scalable metadata handling.

**LAKEHOUSE:**

The Lakehouse platform is built on top of Delta Lake, which is an open-source storage layer that adds reliability, scalability, and performance to data lakes. It provides ACID transactions, schema enforcement, and indexing to data lakes, making them more suitable for data warehousing workloads.

**Unity Catalog:**

Unity Catalog is a Unified Governance for All Data & AI Assets including files, tables, machine learning models and dashboards in your lakehouse on any cloud. It is designed to help organizations manage and organize their data in a scalable and efficient manner.

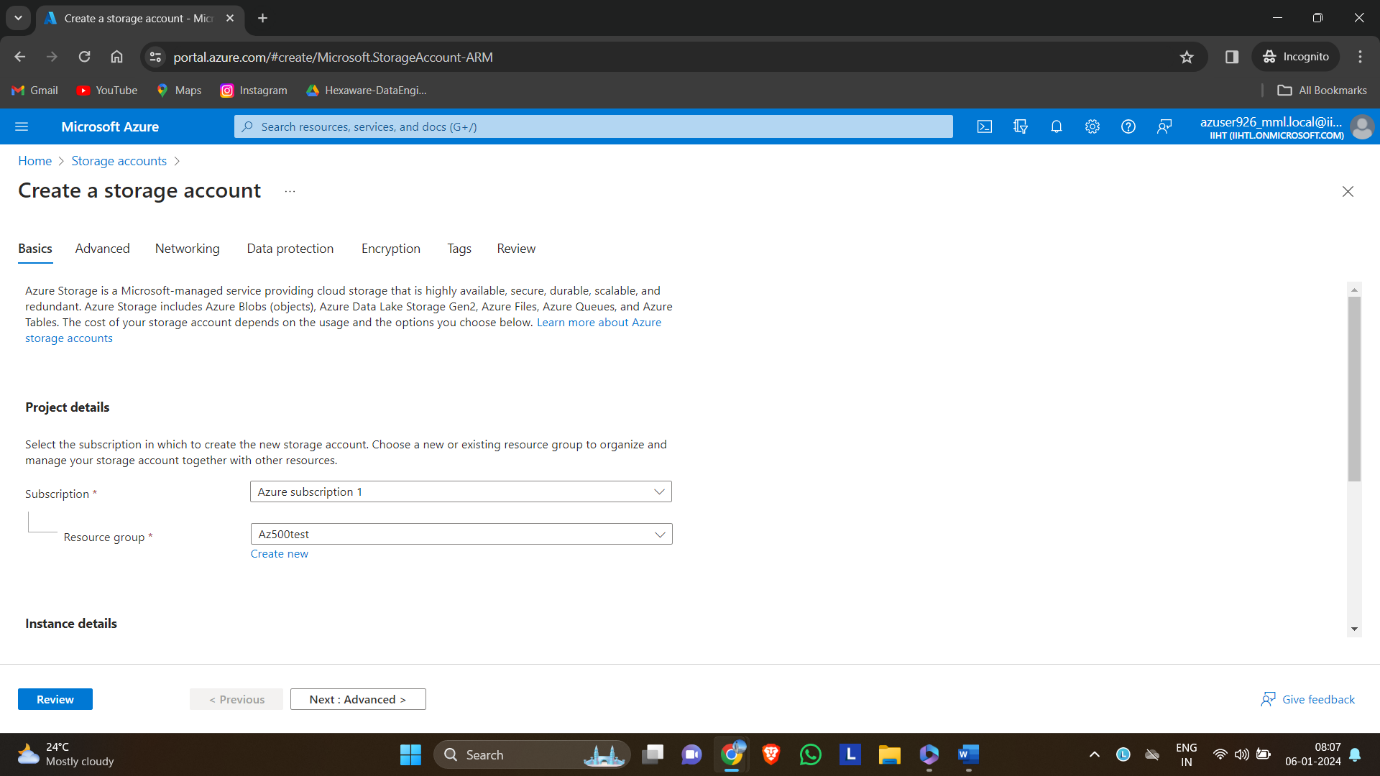
**MEDALLION ARCHITECTURE:**

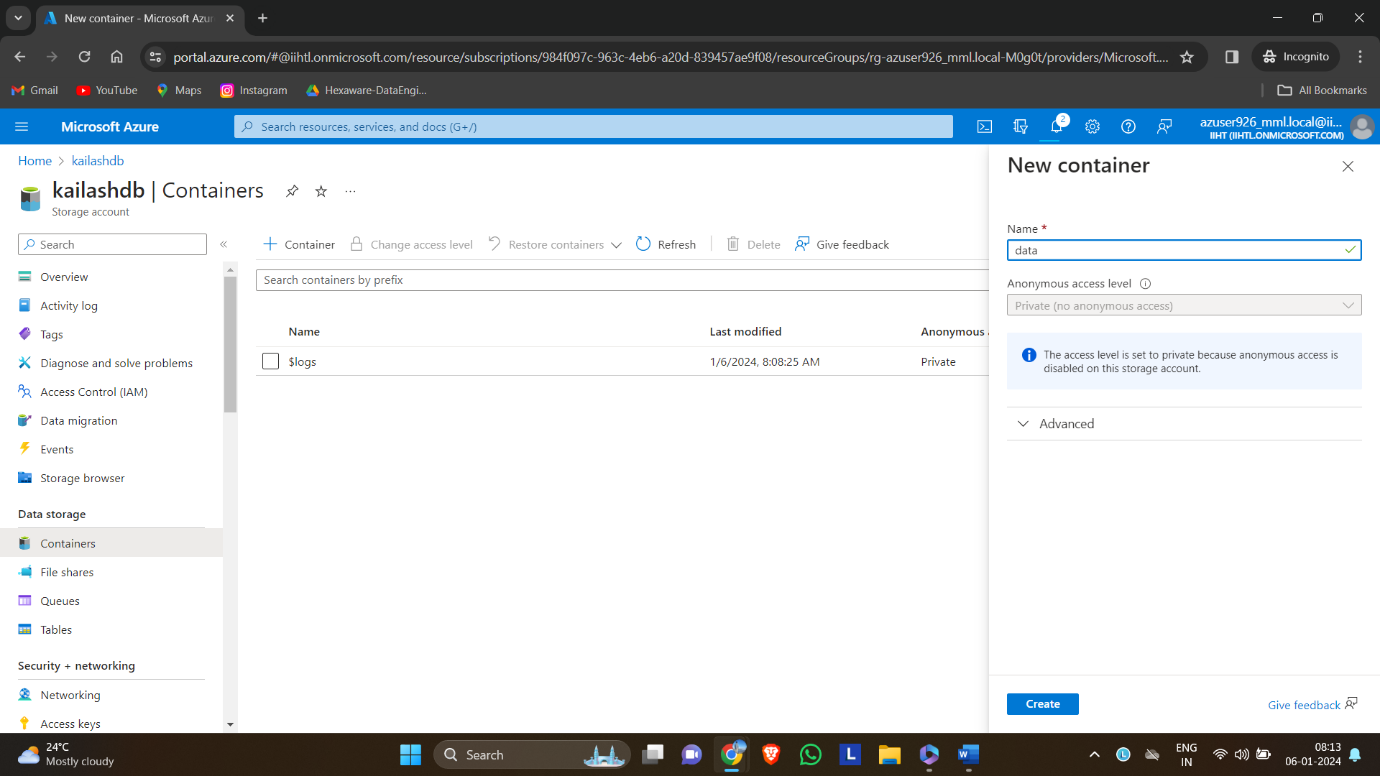
**BRONZE LAYER**- contains raw data as original format

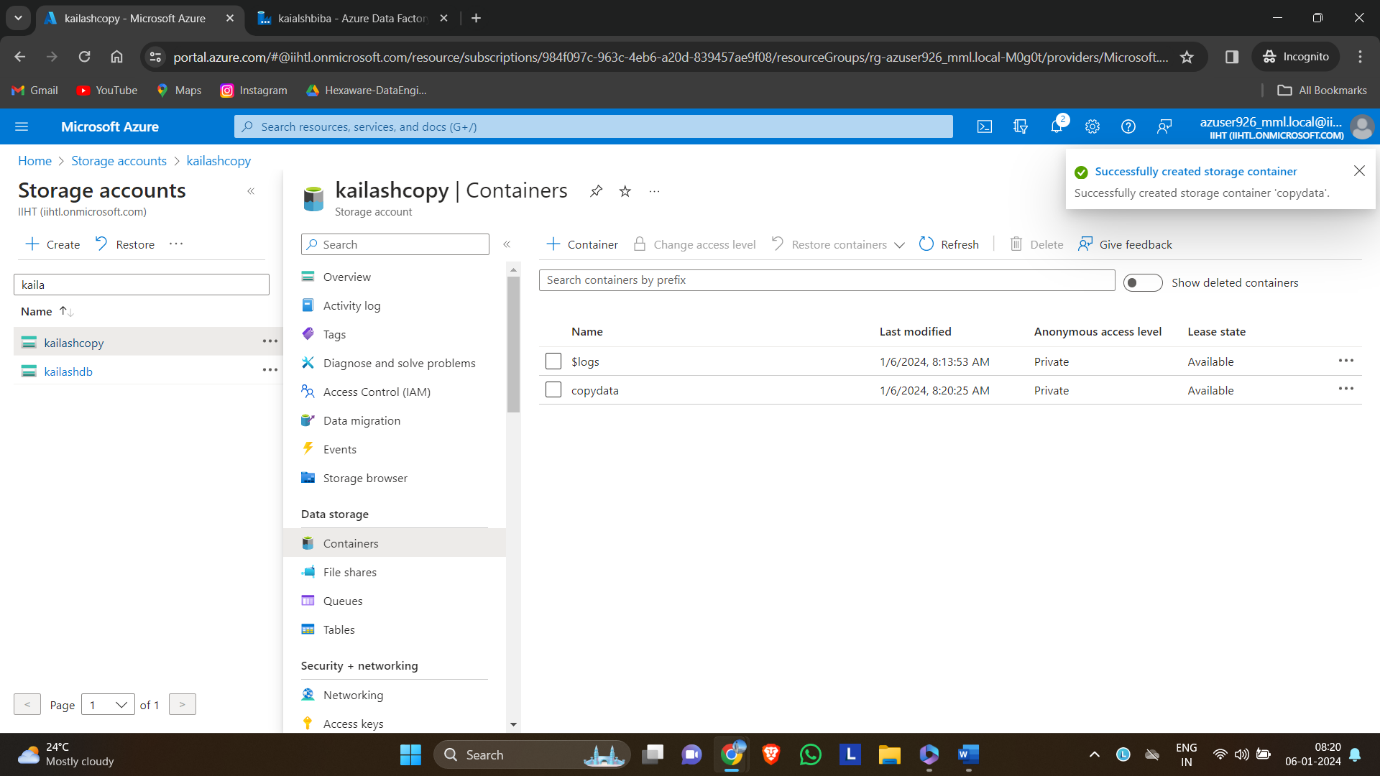
**SILVER LAYER**- Silver layer is responsible for cleaning and filtering the Bronze data.

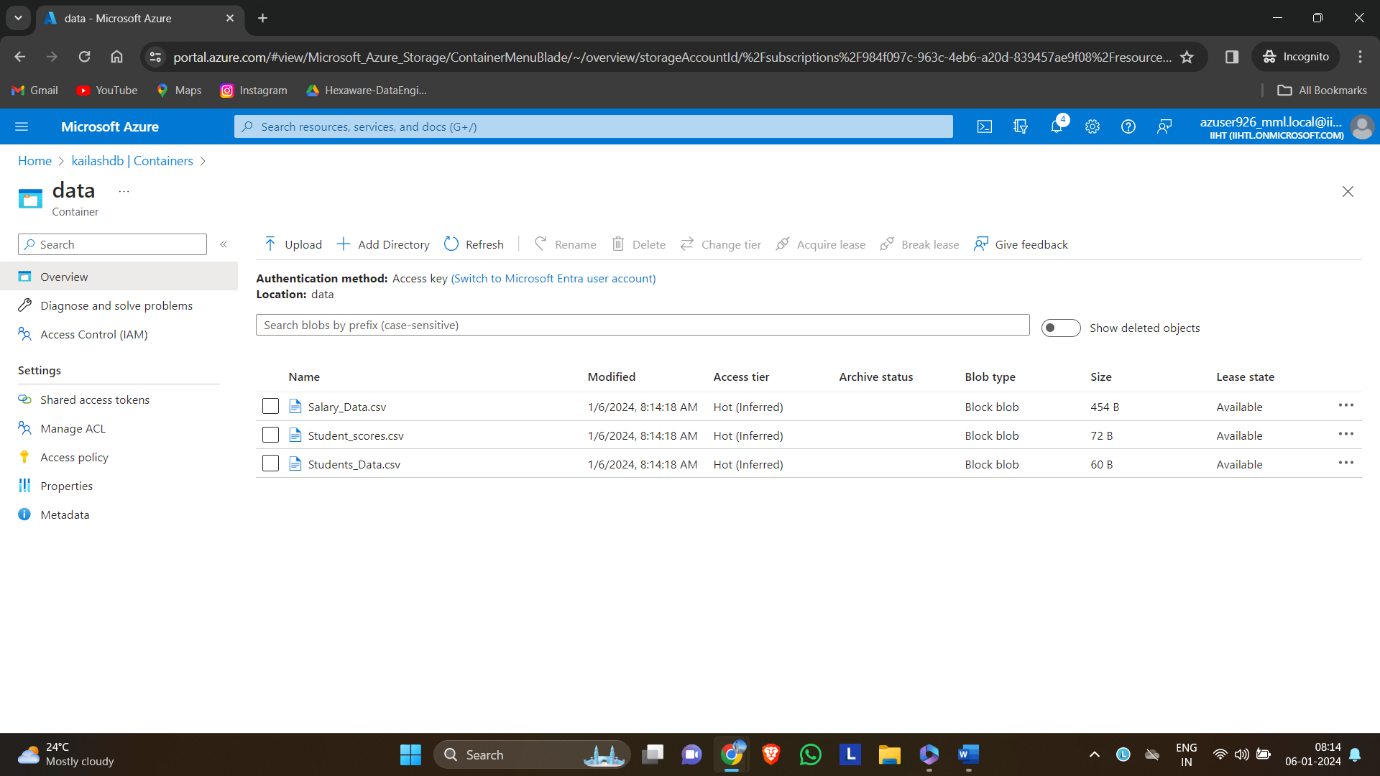
**GOLD LAYER**- This layer contains the transformed, aggregated and modeled data

**Copy data using azure data factory:**

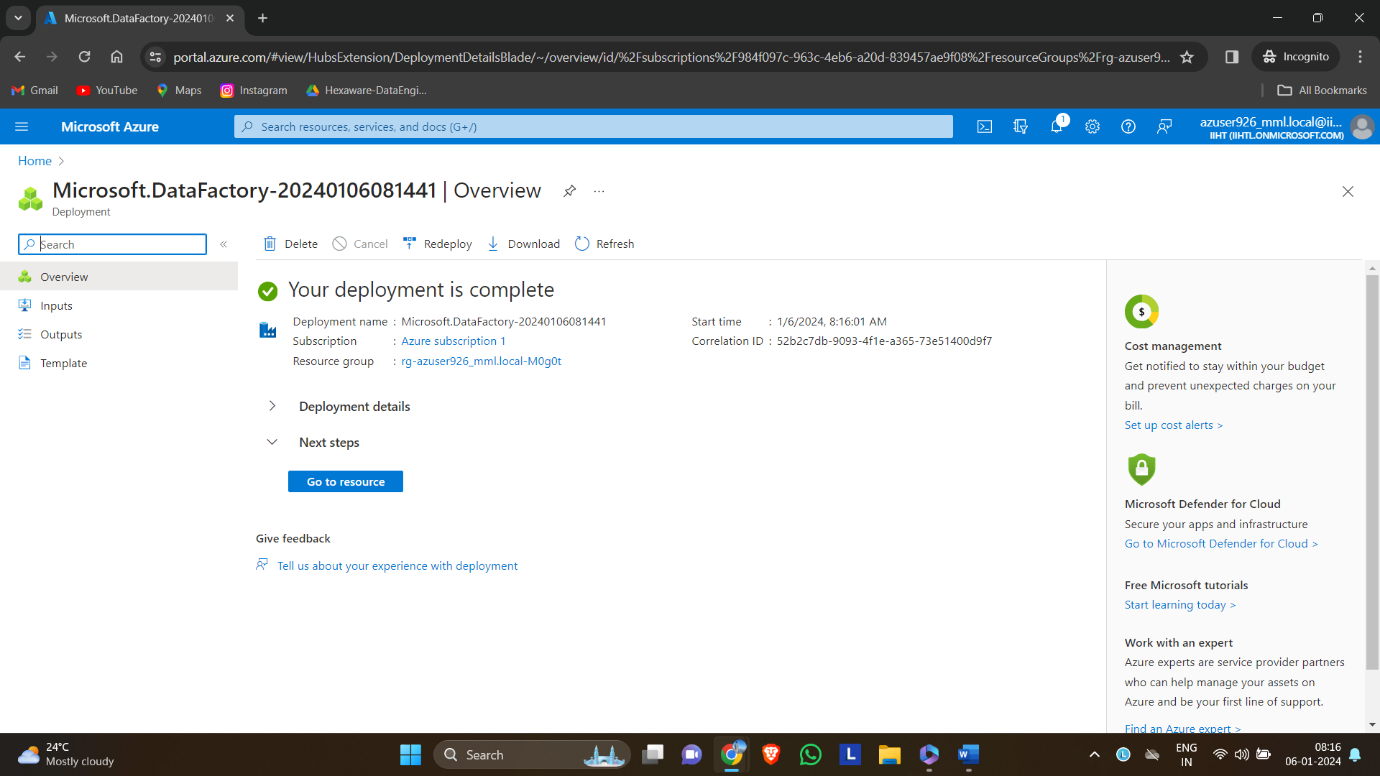
**Step-1:** creating two storage accounts to copy the data and adding the sample data in one account.

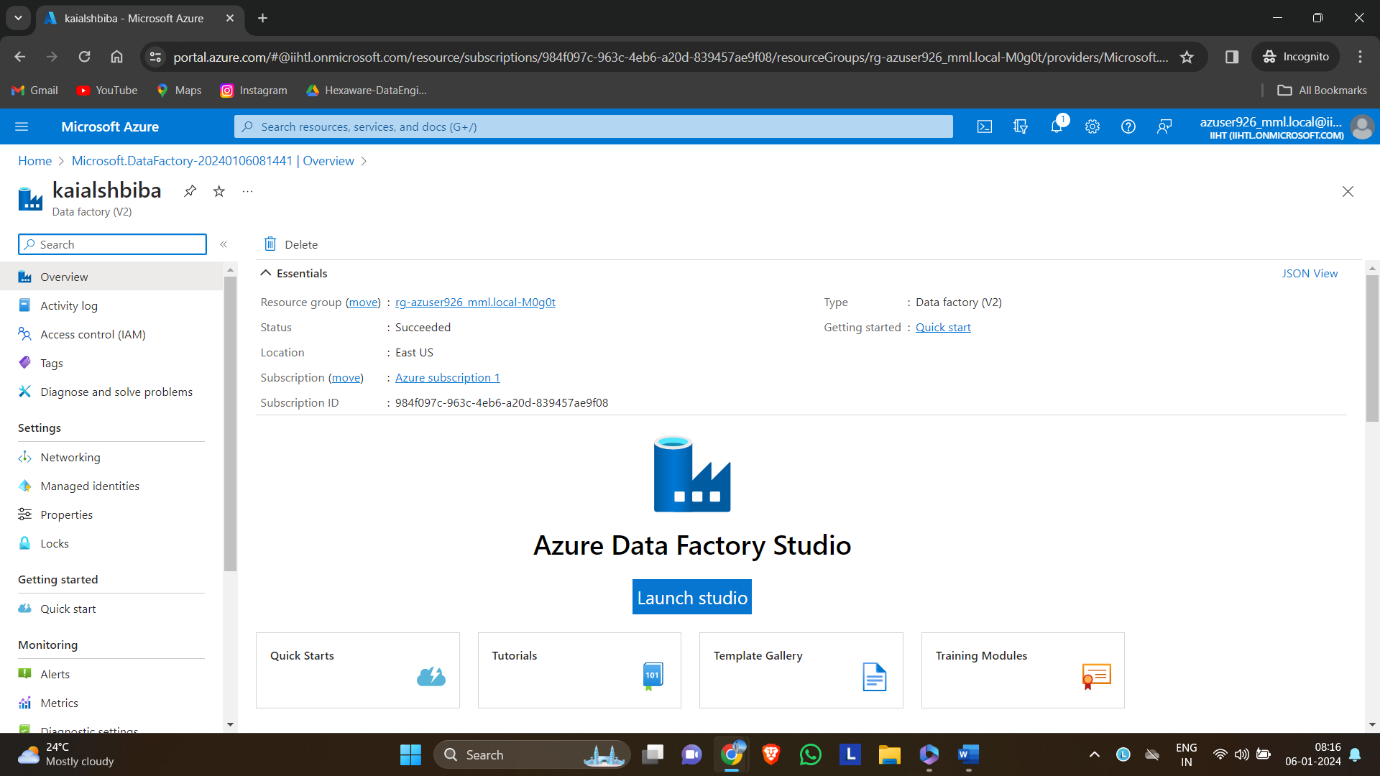




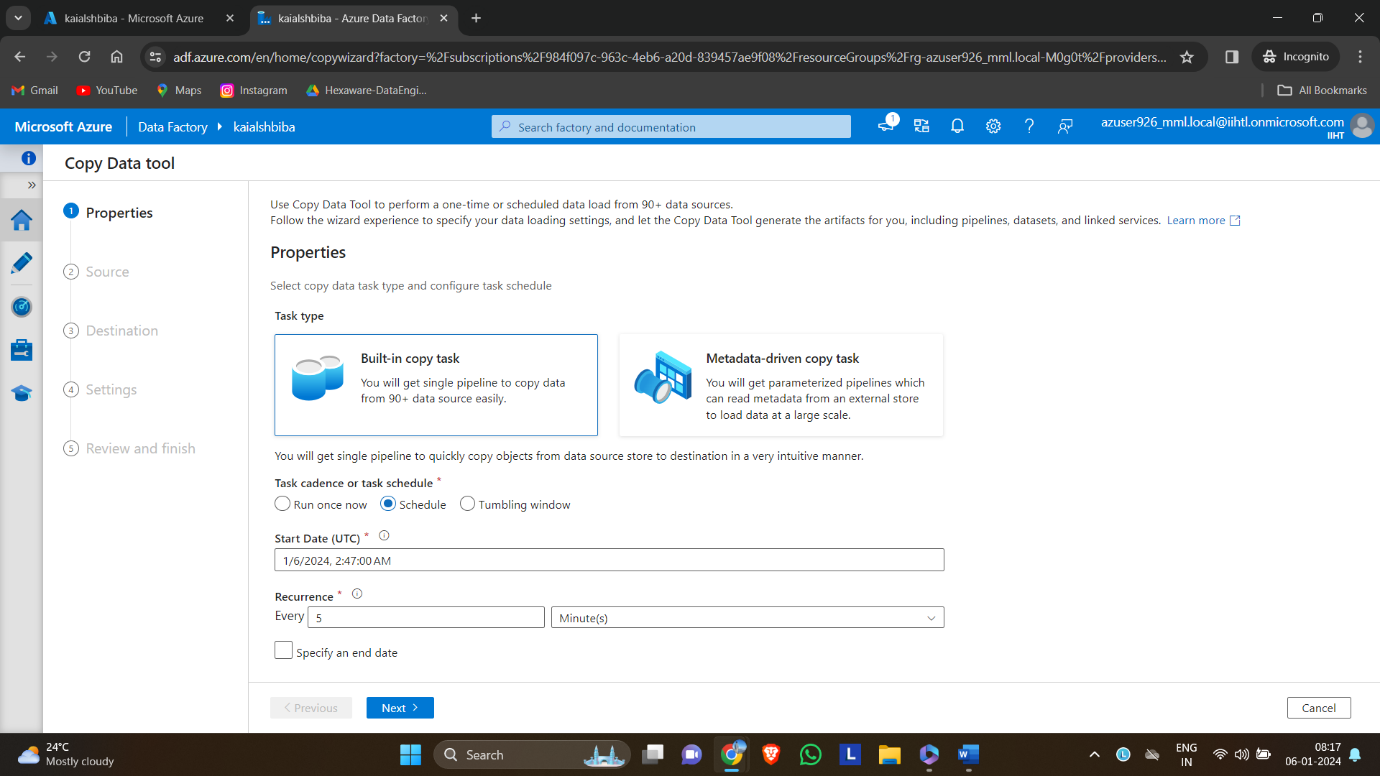


**Step2:** Deploying the azure datafactory.





After creating the data factory click ingestion in the studio and select built in copy task.

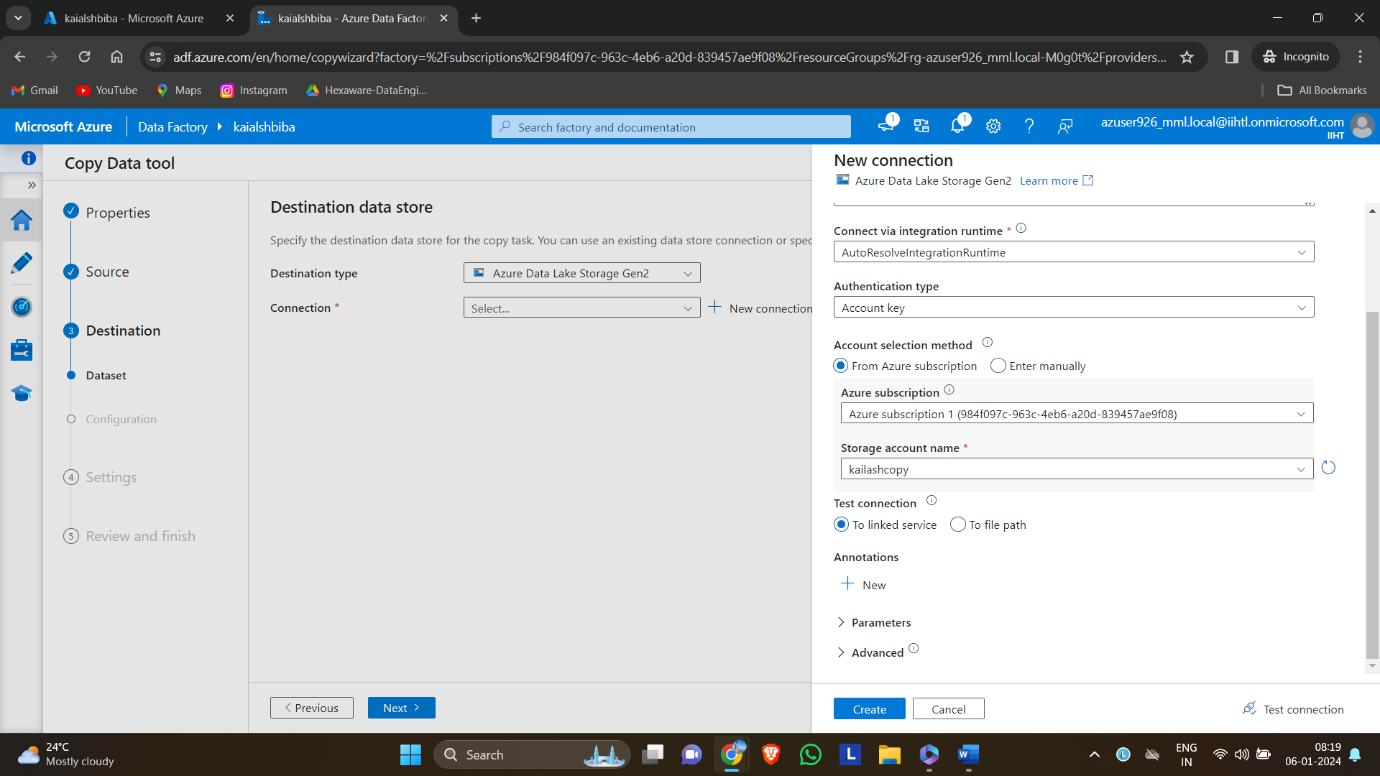


Select the source storage account to be copied.

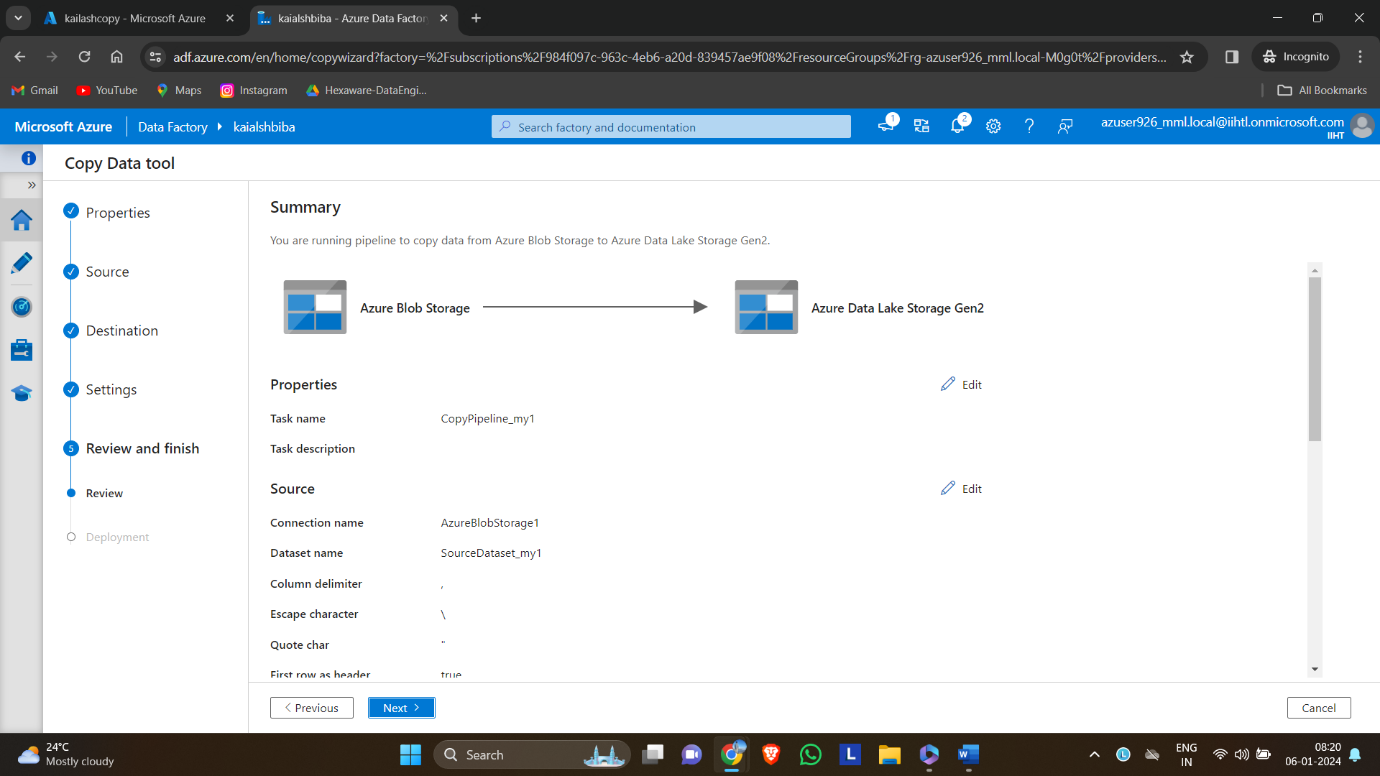
A screenshot of a computer

Description automatically generated

Selecting the destination storage account



After reviewing the process we can deploy the pipeline in the data factory studio.



The pipeline will run automatically and copies data according to the scheduled time.

