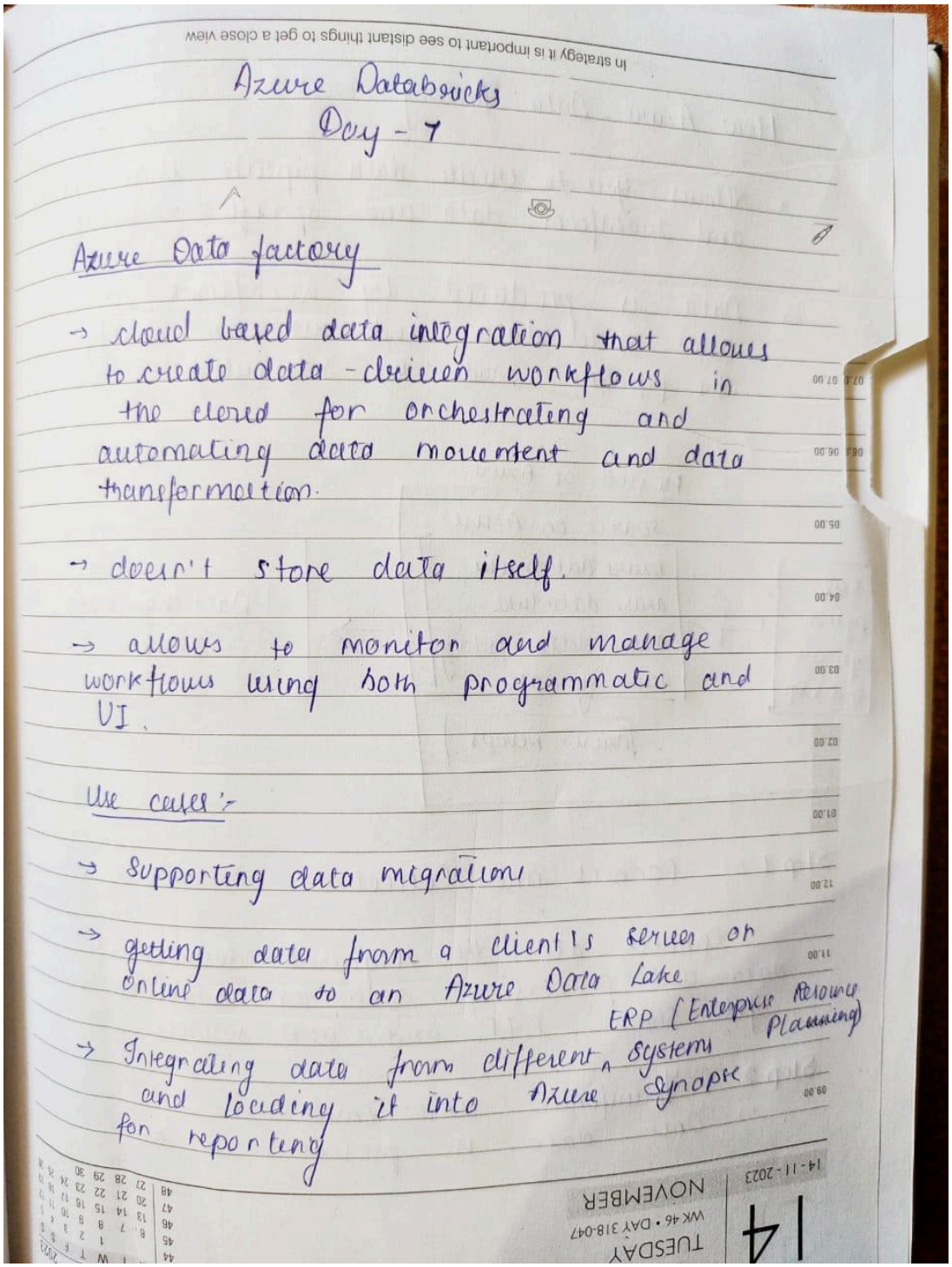
Azure Databricks Assignment - 7

Mitushi Vishwakarma

NOTES :



In strategy it is important to see distant things to get a close view

Azure Databricks
Day - 7

Azure Data factory

→ cloud based data integration that allows to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.

→ doesn't store data itself.

→ allows to monitor and manage workflows using both programmatic and UI.

Use cases :-

→ Supporting data migration

→ getting data from a client's server or online data to an Azure Data Lake

→ Integrating data from different systems ERP (Enterprise Resource Planning) and loading it into Azure Synapse for reporting

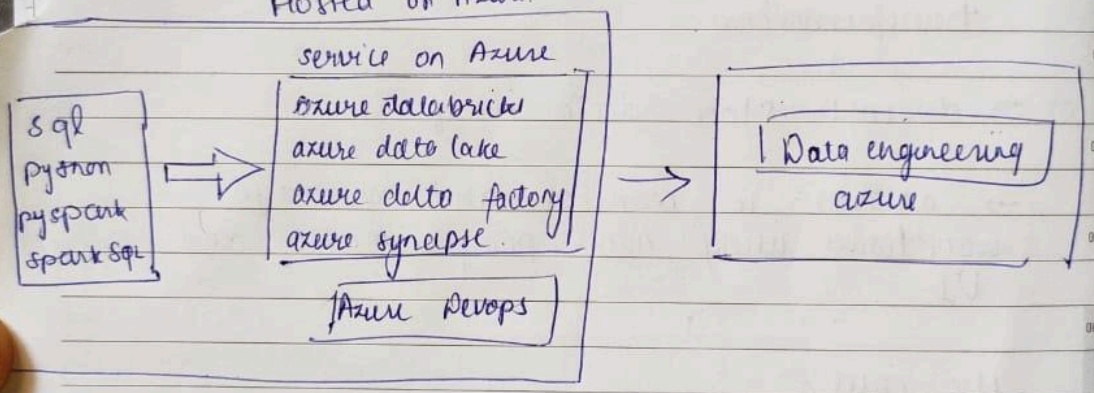
14-11-2023
TUESDAY
NOVEMBER
WK 46 • DAY 318-047
14

When work, commitment and pleasure all become one, nothing is impossible

How Azure Data Factory work?

→ Allows you to create data pipeline that move and transform data and specified schedule.

→ Data is produced by workflows in time sliced data and we can specify the pipeline made as scheduled.



Step 1: Connect and Collect:

→ Connect all the required sources of data and processing such as SaaS services file shares, FTP and web services

Step 2: Transform and Enrich

→ Once data is present in a

Form a mental vision of your goal and cling to it through thick and thin

Centralized data store in the cloud. It is transformed using compute services such as HDInsight Hadoop, Spark, Azure data lake Analytics.

12 SUNDAY

Step 3: Publish :-

→ Cloud → on premise sources like SQL server.

Key components

Four components that work together to define input and output data, processing events.

1. Datasets represents data structure within the data store:

Input dataset :- Input for an activity in pipeline

Output dataset → output for the activity.

11-11-2023

2023
SATURDAY
WK 45 • DAY 315-050
NOVEMBER

WK	M	T	W	T	F	S	S
44			1	2	3	4	5
45	6	7	8	9	10	11	12
46	13	14	15	16	17	18	19
47	20	21	22	23	24	25	26
48	27	28	29	30			

Let our advance worrying become advance thinking and planning

2. A pipeline is a group of activities.

→ Group activities into a unit that together performs a task.

→ may have one or more pipelines.

3. Activities defines the actions to perform on your data.
   → data movement
   → data transformation

4. Linked Services define the information needed for Azure data factory to connect the external resources.

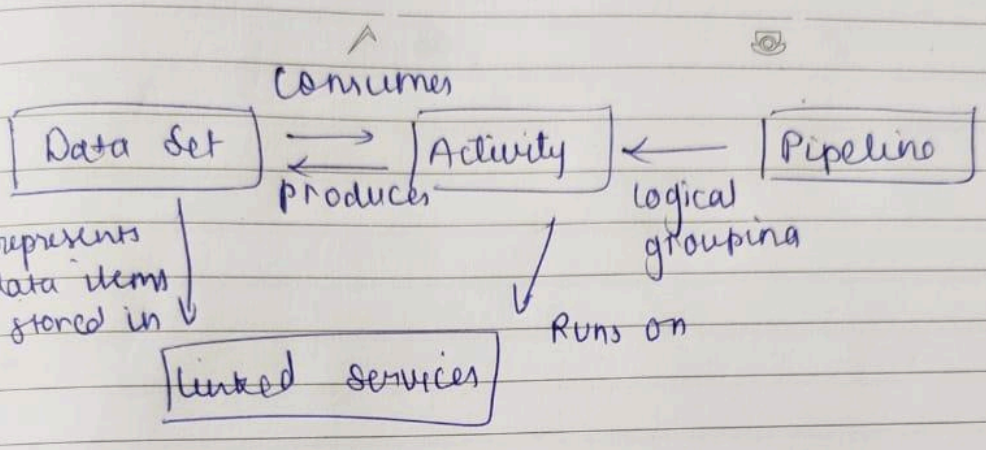
Copy Activity

Copy data from data source (source) to sink data store.

following tools or APIs can be used to create data pipeline in ADF.

Azure portal
Visual Studio
powershell
.NET API

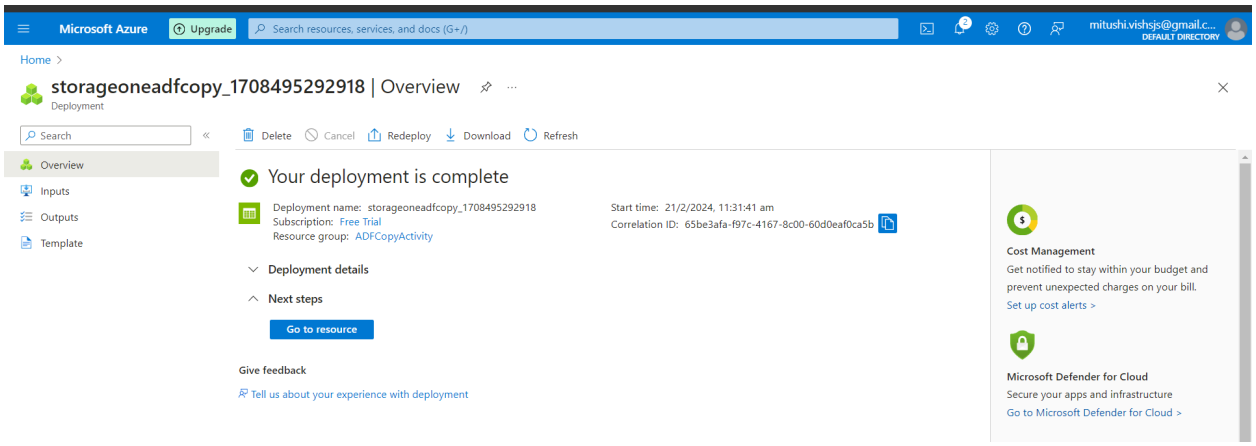
The system of domination is founded on depriving nations of their true identity



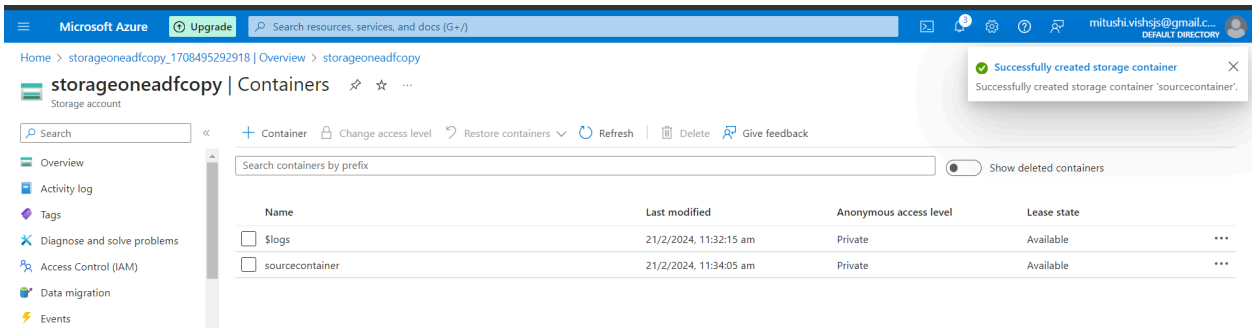
NOVEMBER
2023

Copy Activity : Creating two storage accounts “storageoneadfcopy” and “storagesecondadfcopy” and performing copy activity in data factory.

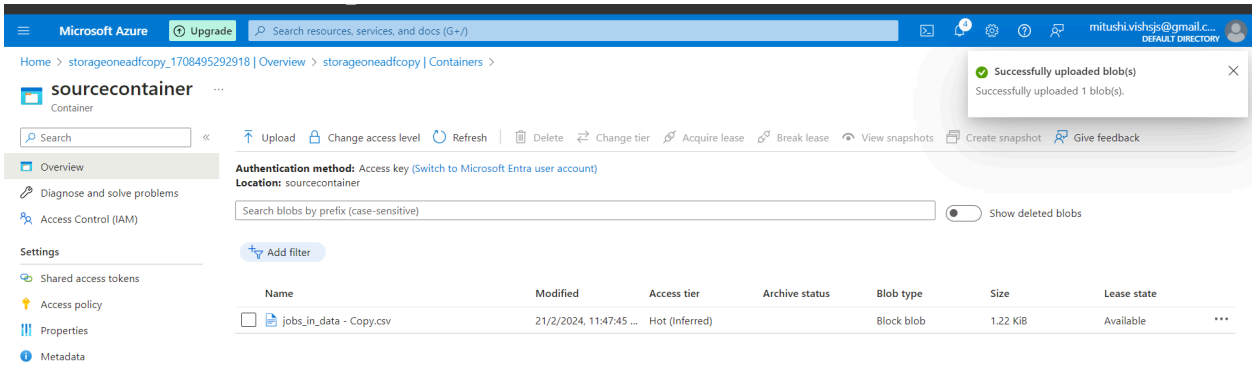
Created first blob storage account and container that will be the source store :



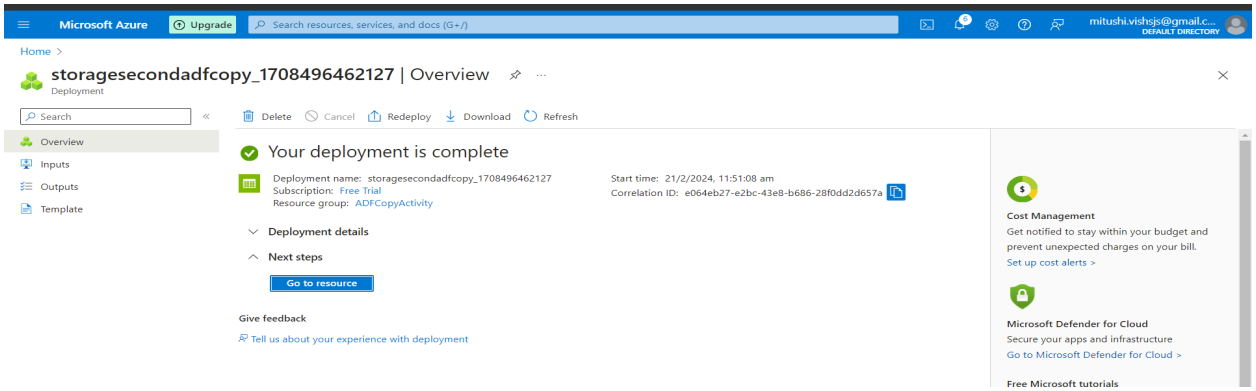
Created container in storage account :



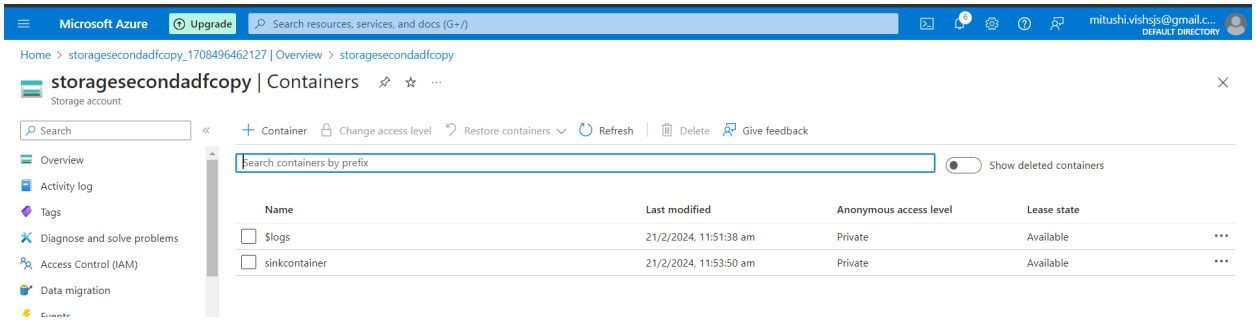
Uploaded file in container :



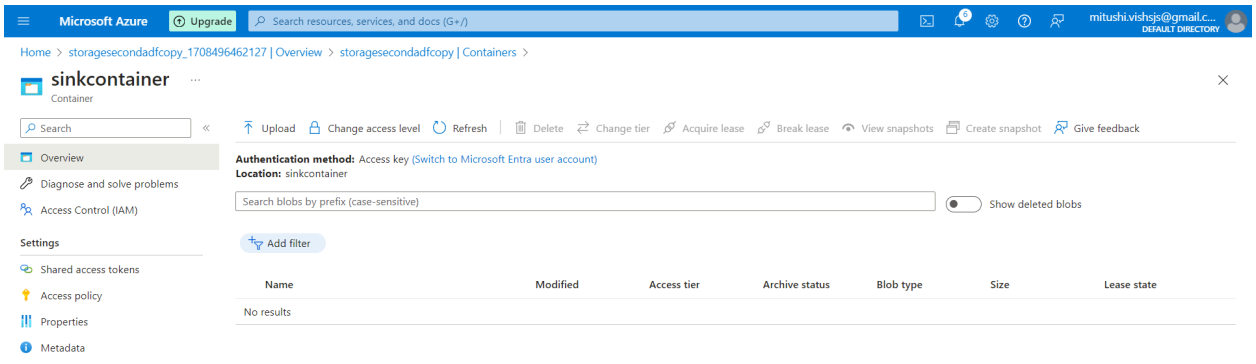
Created second blob storage account and container that will be the sink store :



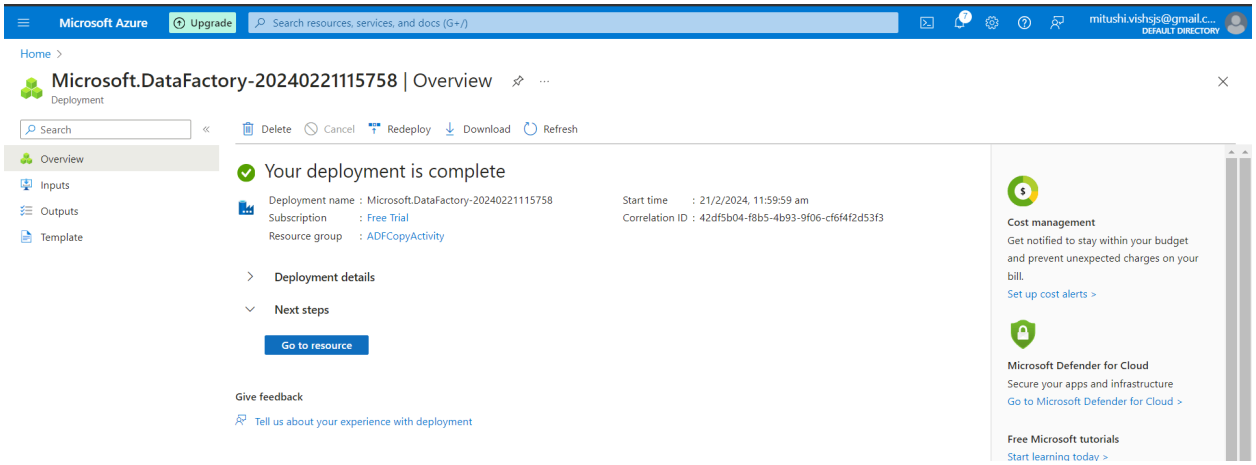
Created container in storage account :



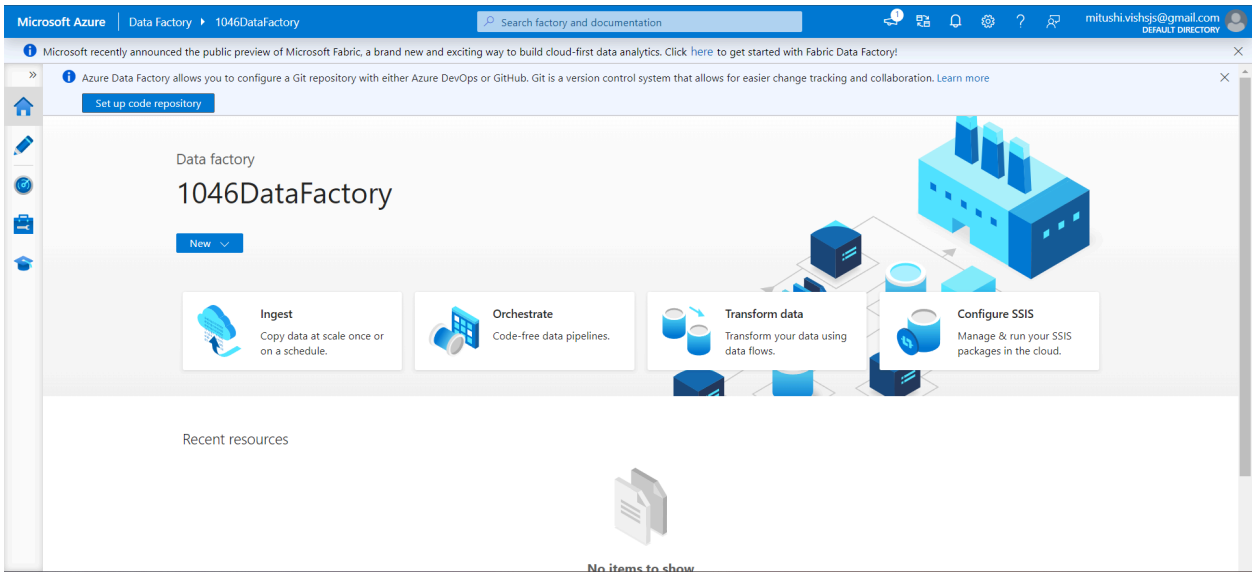
Empty Container :

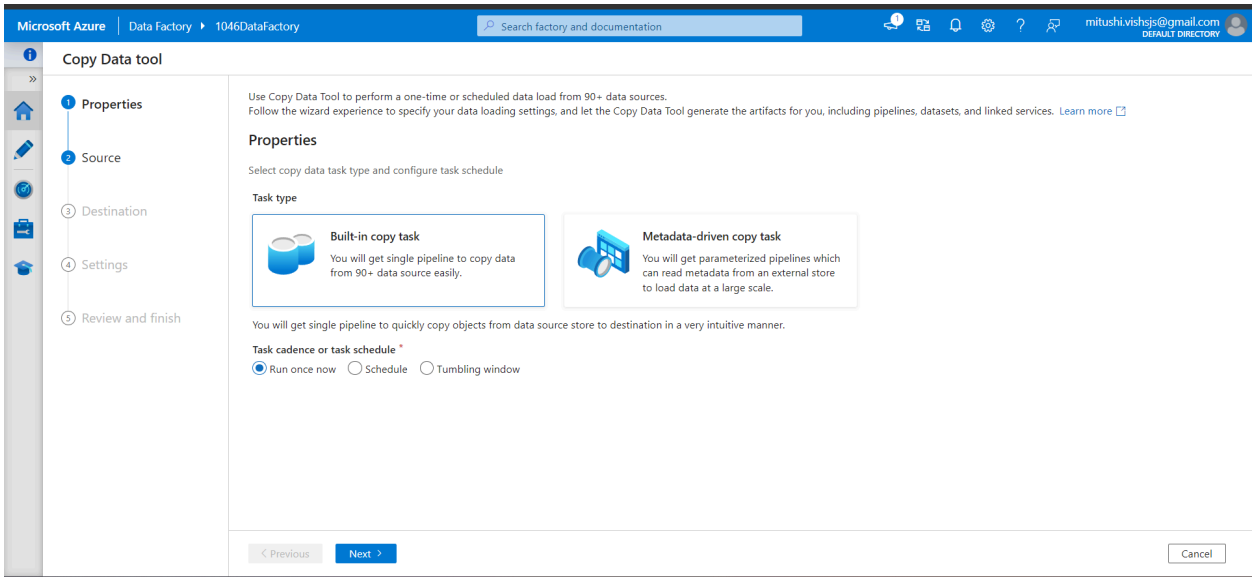


Created Data Factory named “1046DataFactory “ :

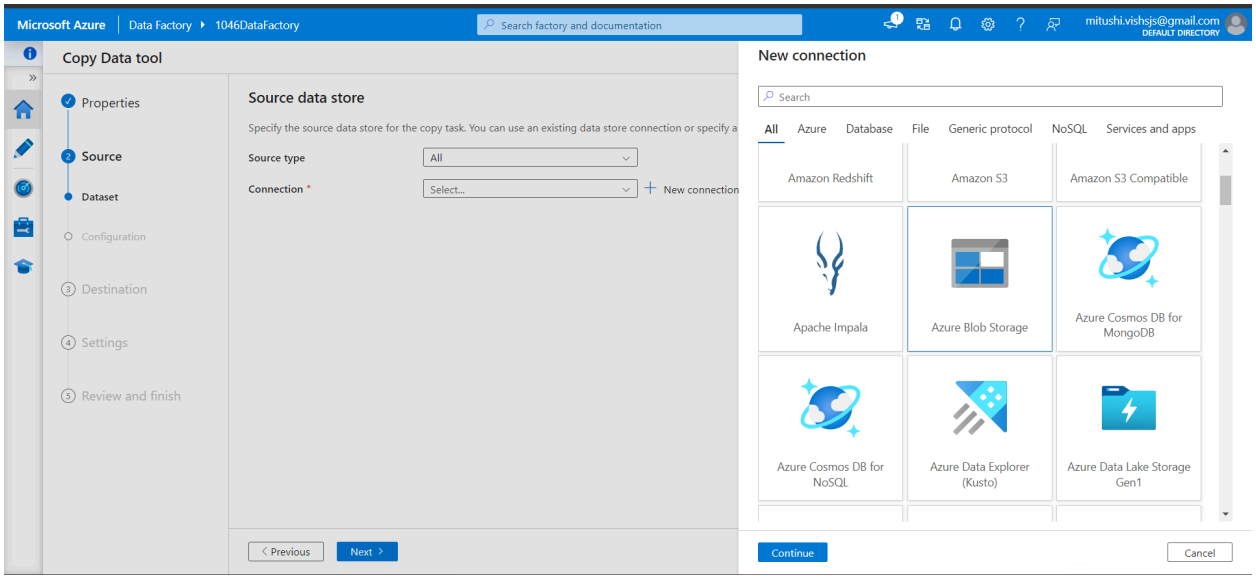


Selecting Ingest to copy data :

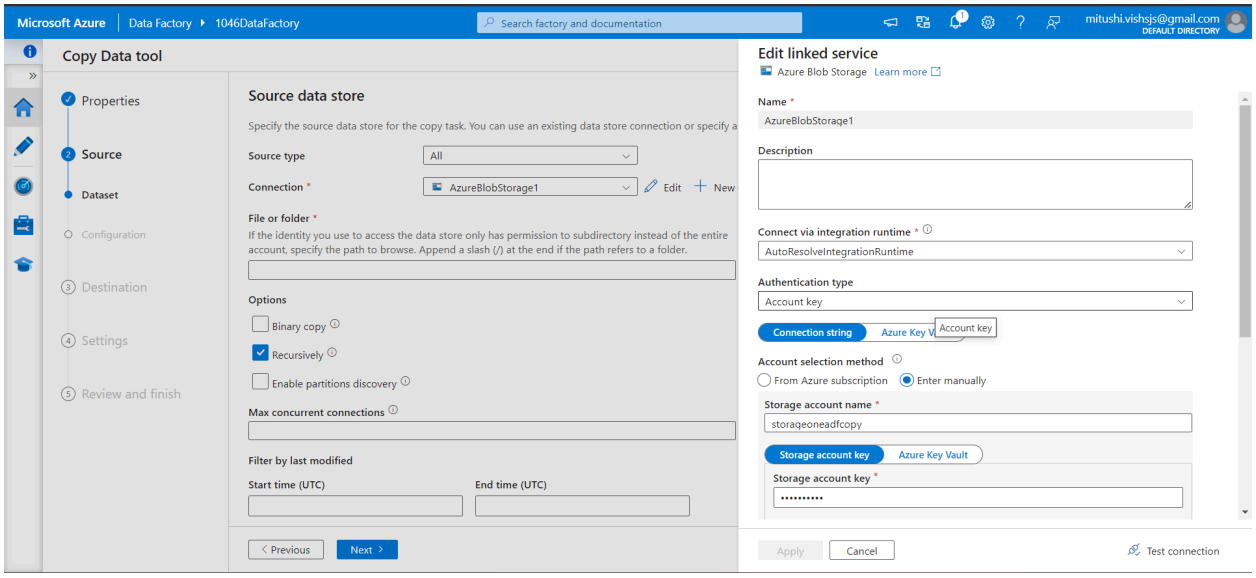




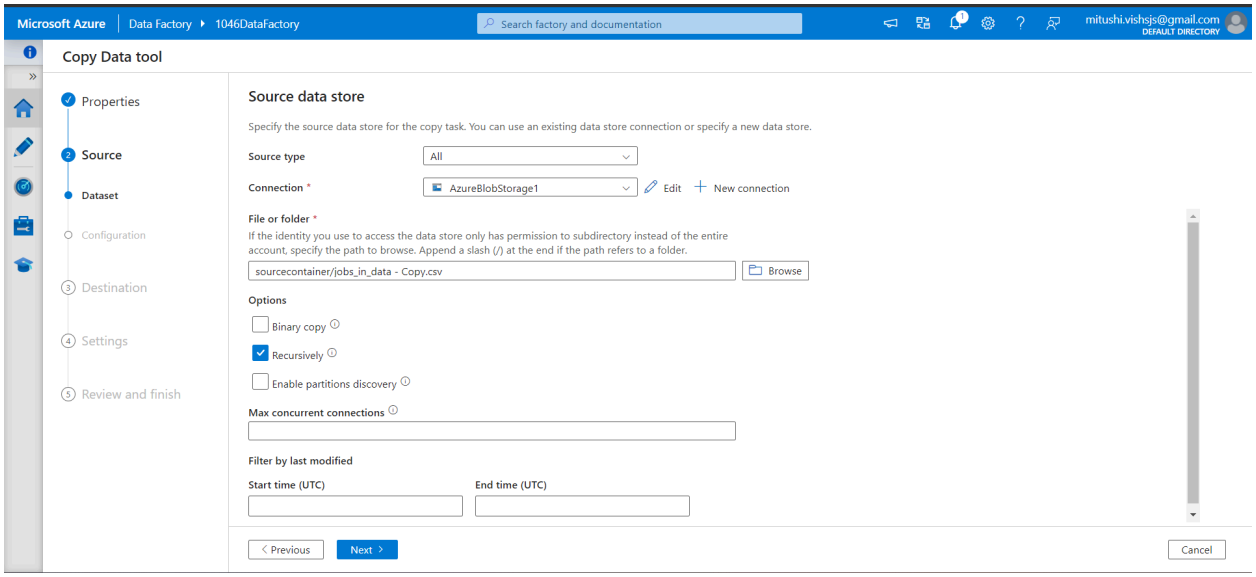
Selecting Source data store :



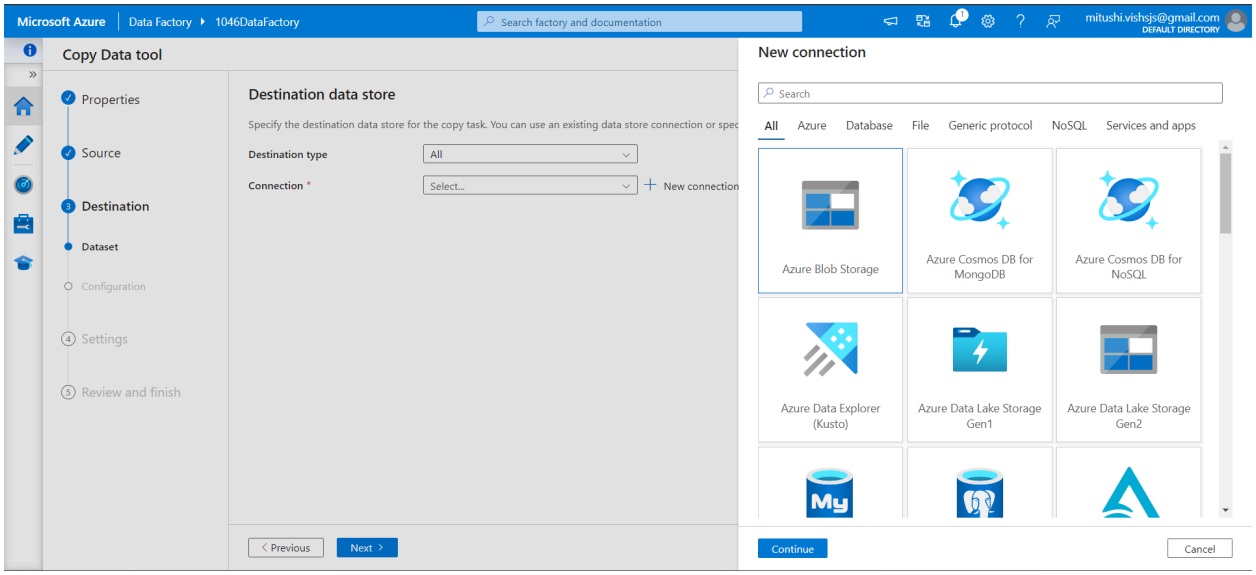
Linked first storage account :



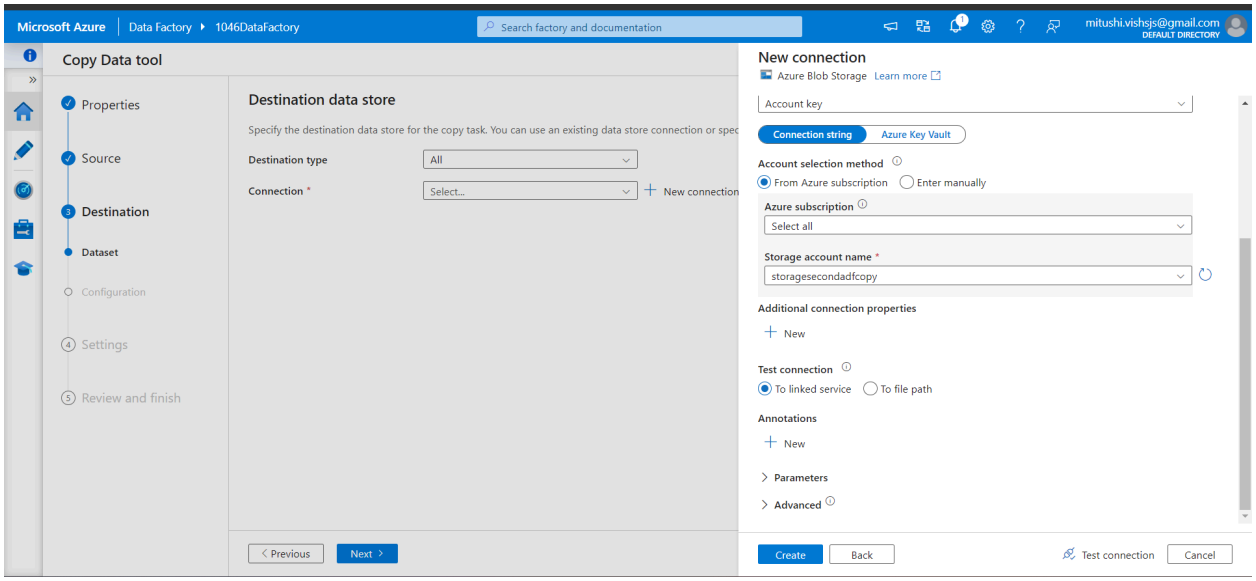
Provided the path for source data store files :



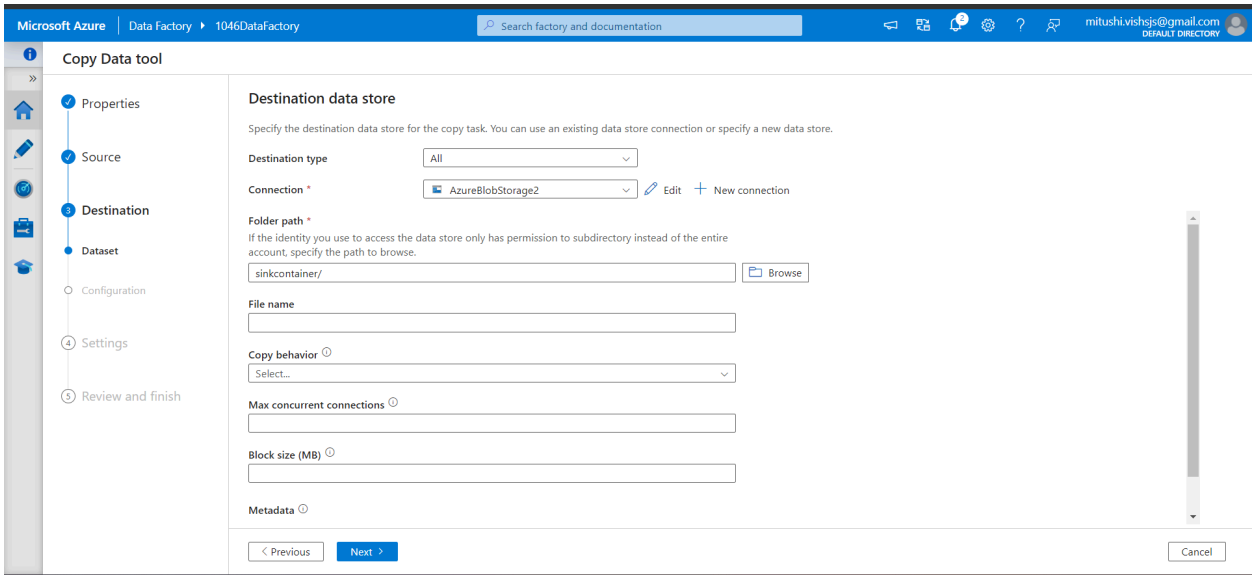
Creating Destination data store connection :



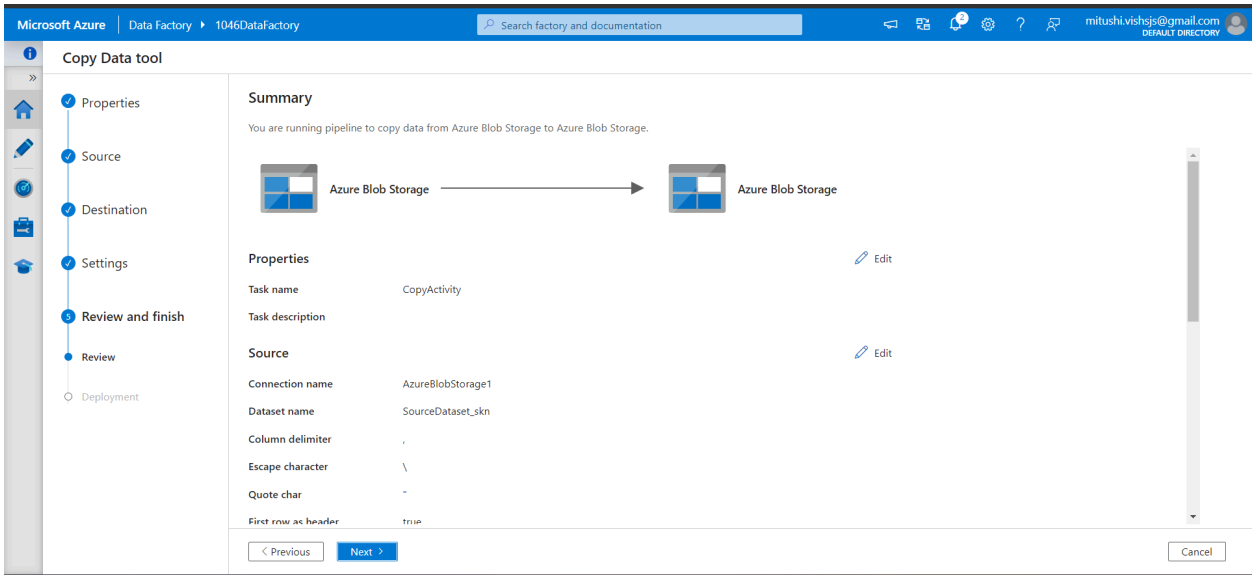
Linked second storage account in the destination connection :



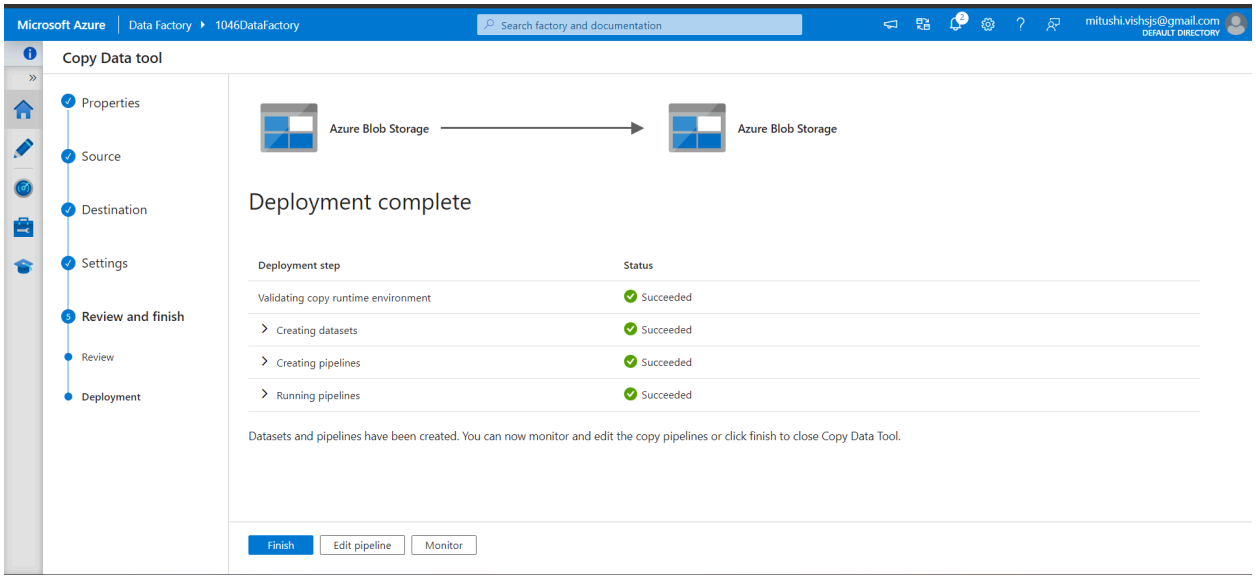
Provided the path where the files will be copied :



Review and submit :



Data Copied successfully :



Checking Copied data file :