Mallu Sravana Sandhya
Assessment-27
05/01/24

Data Lakehouse:
It is a combination or integration of Delta lake which is an open source storage that gather all the information stored in the cloud storage ,process it and Data Warehouse which contains huge amount of databases.

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.

Key features of Lakehouse:

- It provides transactional updates on data updation, deletion and transfer of data which indicates the ACID properties
- Data lakehouse can contain data fthat  belongs to different version and maintains history of data which helps in analysis of version control
- Scalabitlity and performance: it allows any oraganization to process, maintain and analyze huge amount of data.it uses various distributed computing techniques to process and analysis
- Schema enforcement: it provides strong governance and auditing mechanisms by maintaing data quality and consistency
- It works on wide range of data sources and formats irrespective of their structured forms
- Lake house provide a unified way to manage metadata across multiple data sources, making it easier to discover, access, and query data.

**Architectural component of Lakehouse:**

Delta Lake —

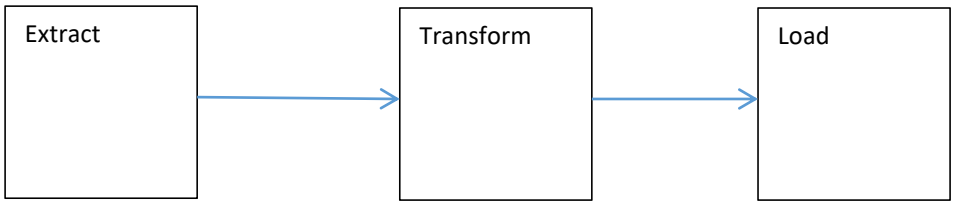
Delta Lake provides a powerful set of features for managing data in a lakehouse architecture. It is a optimized storage layer that allows users to store and manage data in various file formats.Delta Lake also provides a transaction log, which enables ACID transactions and allows for data versioning and rollbacks.

Unity Catalog —

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

- Centralized governance for data and AI
- Provide enhanced query performance and boosts productivity
- Transparency of creating automated lineage for all workloads
- Secure data sharing across organizations

Databricks Delta Live Tables provide one of the key solution to build and manage, reliable and robust data engineering pipelines that can load the Streaming and batch data and deliver high-quality data on the Lakehouse Platform.




In data ingestion there are three layers they are

Bronze layer-it provides the raw data which is collected from data sources .it maintains all the history of organizations and provide data with no filters

Silver layer- it provides filter data it maintains data in semistructured form.It handles the missing data and also does the data type conversation if required. It converts the nested objects into the flat structures .It is responsible for cleaning and filtering bronze data

Gold layer-It contains the well structured, aggregated and modeled data which helps for the business queries

Complexity of Data Delivery

The following problems can be faced while maintaining data quality and consistency with large scale solution data pipelines:

- Difficult to maintain table dependencies and to switch between streaming and batch load.
- Error handling and recovery is time consuming
- Pipeline failure may impact the downstream system and the team relying on it.

Data engineers have to focus on tools instead of doing the development because operational complexity dominates.

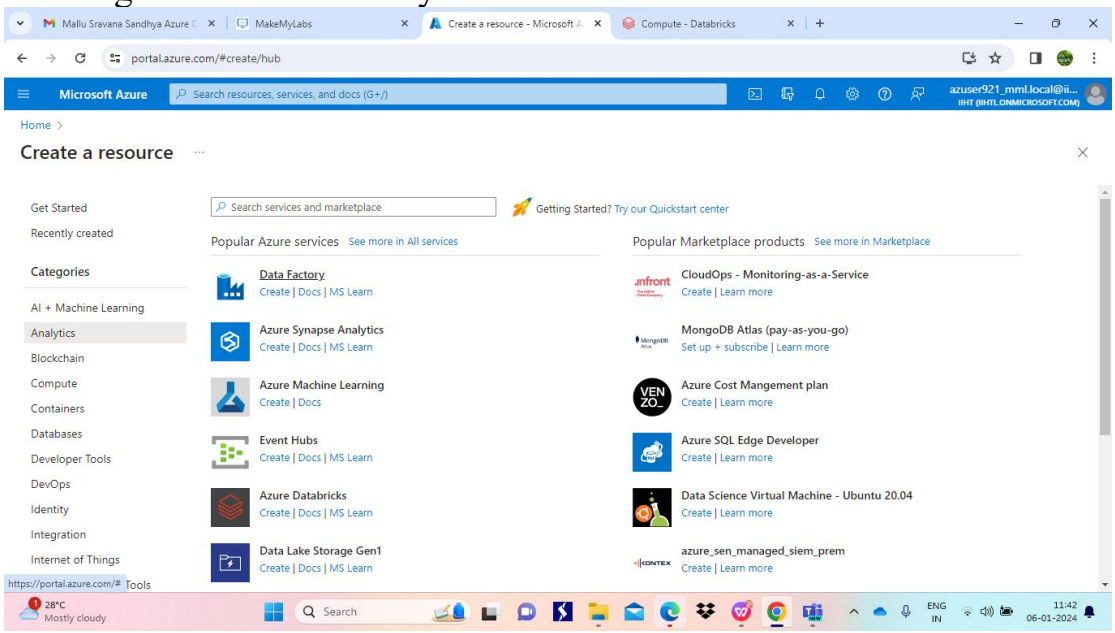
To overcome this data delivery complexity we use DLT Delta Live Tables which helps to solve the complexity issues

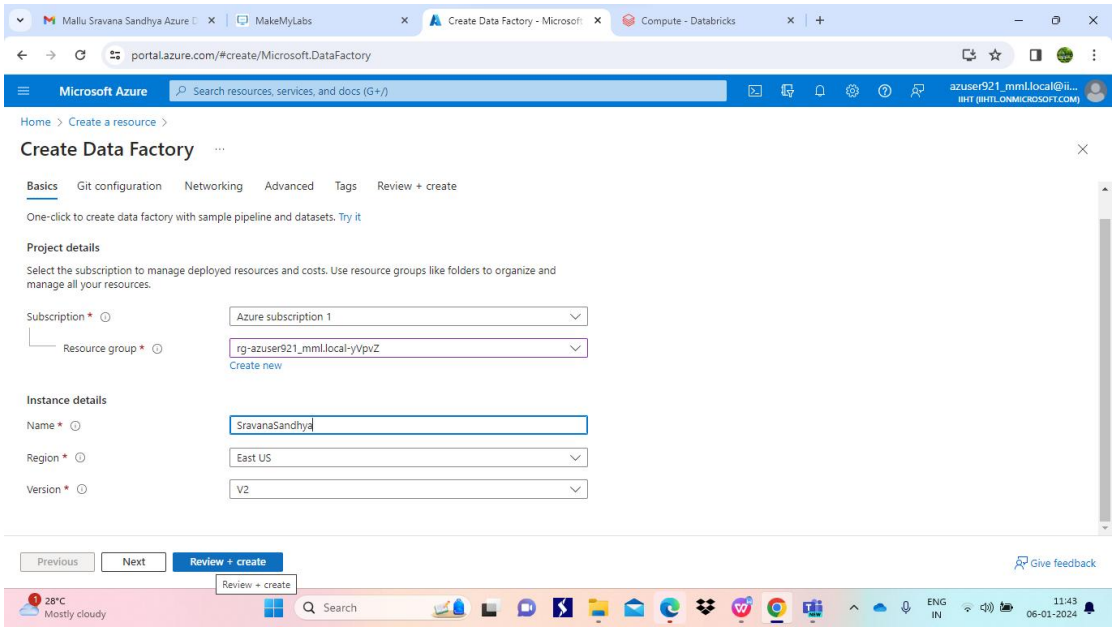
Delta Live Tables or DLT is one of the best way to do ETL process in Lakehouse.It is solely responsible for performing data transformation and managing the task orchestration, cluster management, monitoring, data quality, and error handling.

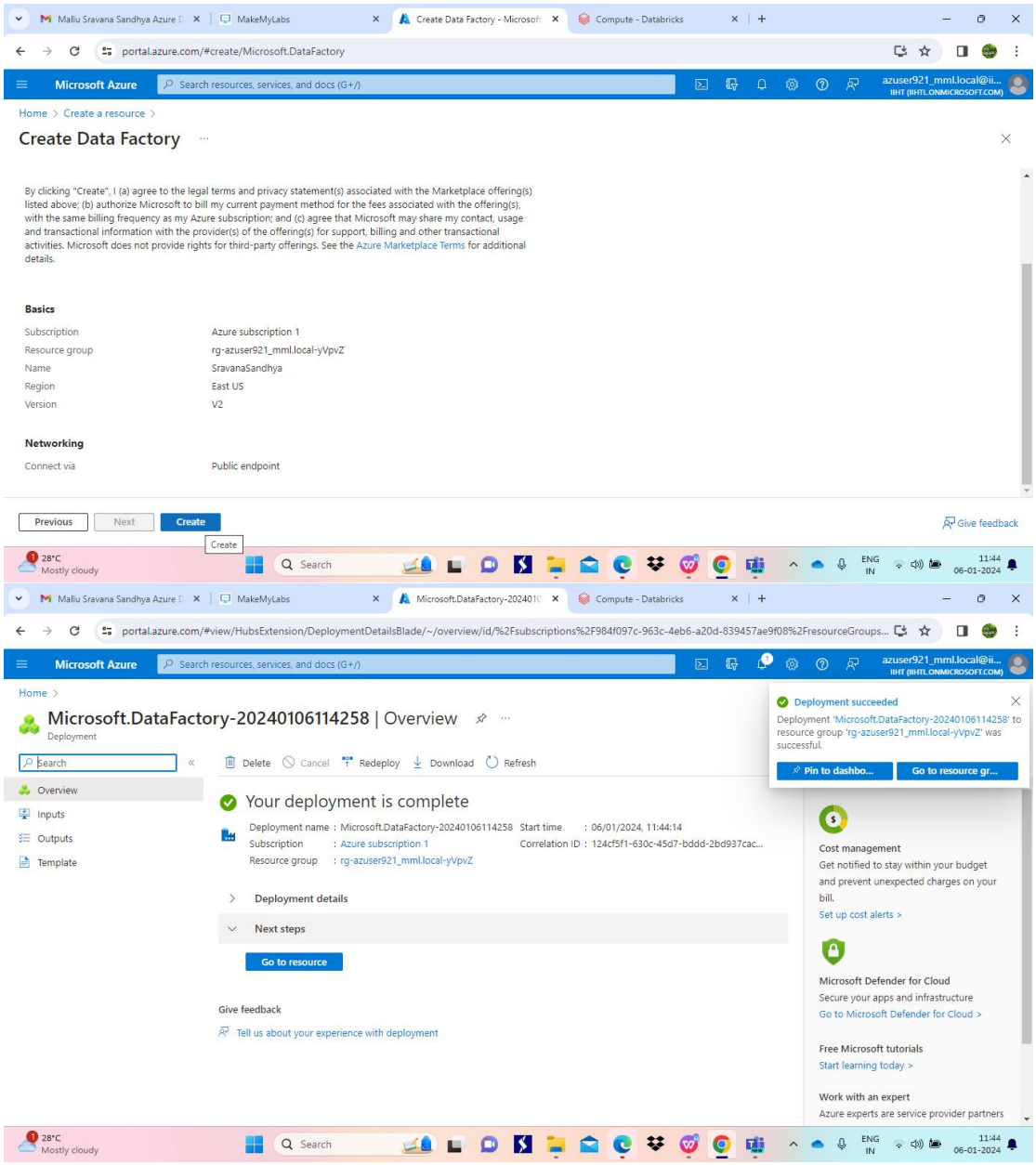
Key benefits of working with Delta Live tables are:

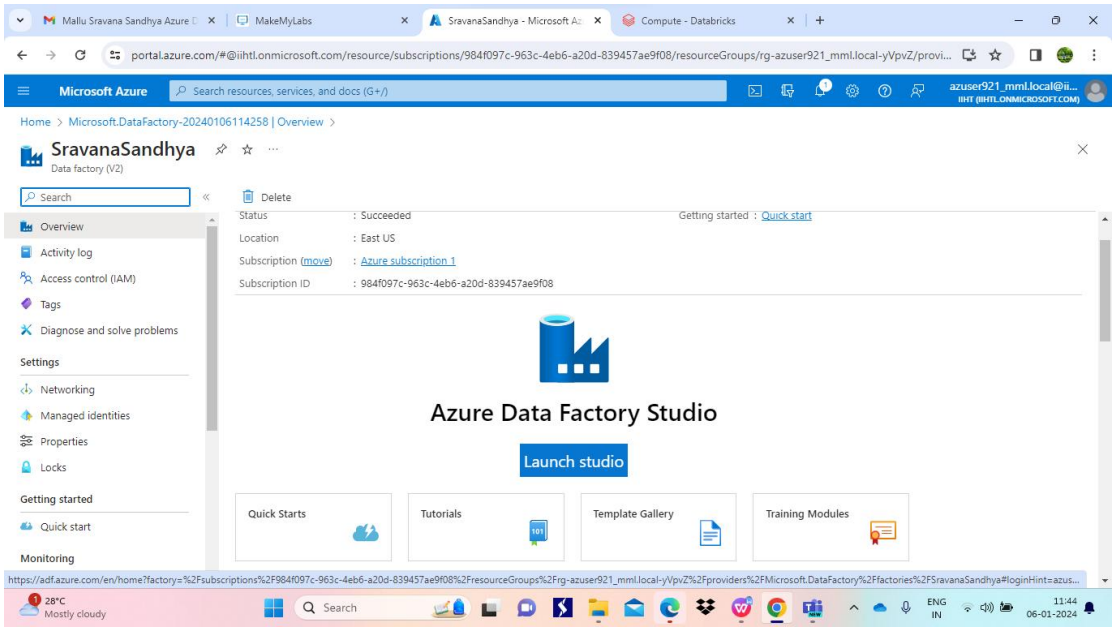
- Accelerates the ETL Process
- Automatically manage your infrastructure
- Ensure high data quality
- Unify batch and streaming

# Creating Azure Data Factory

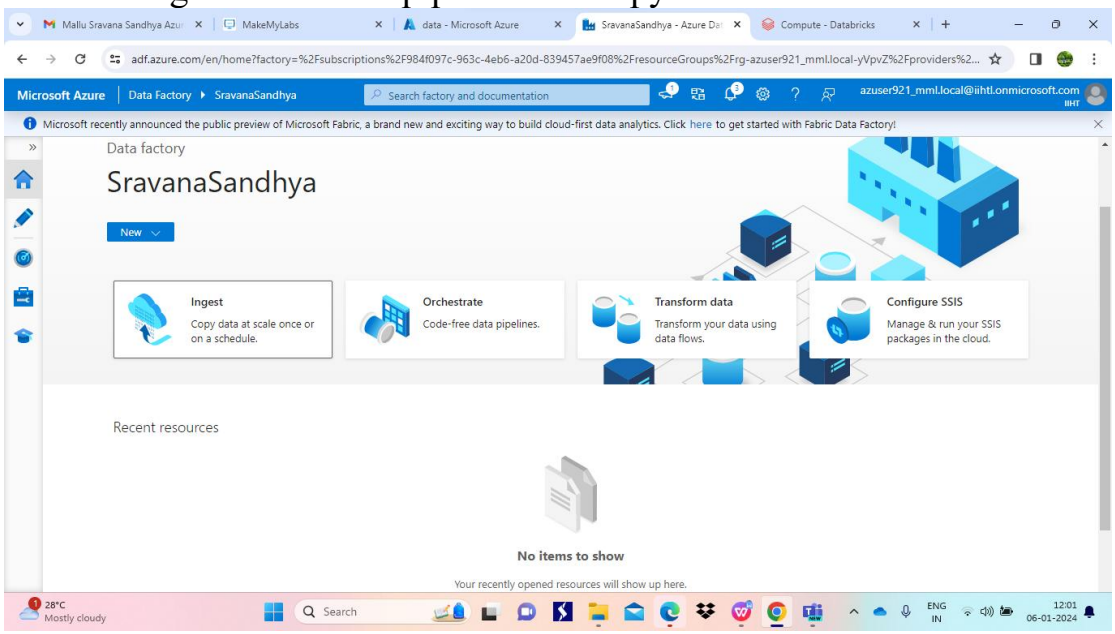


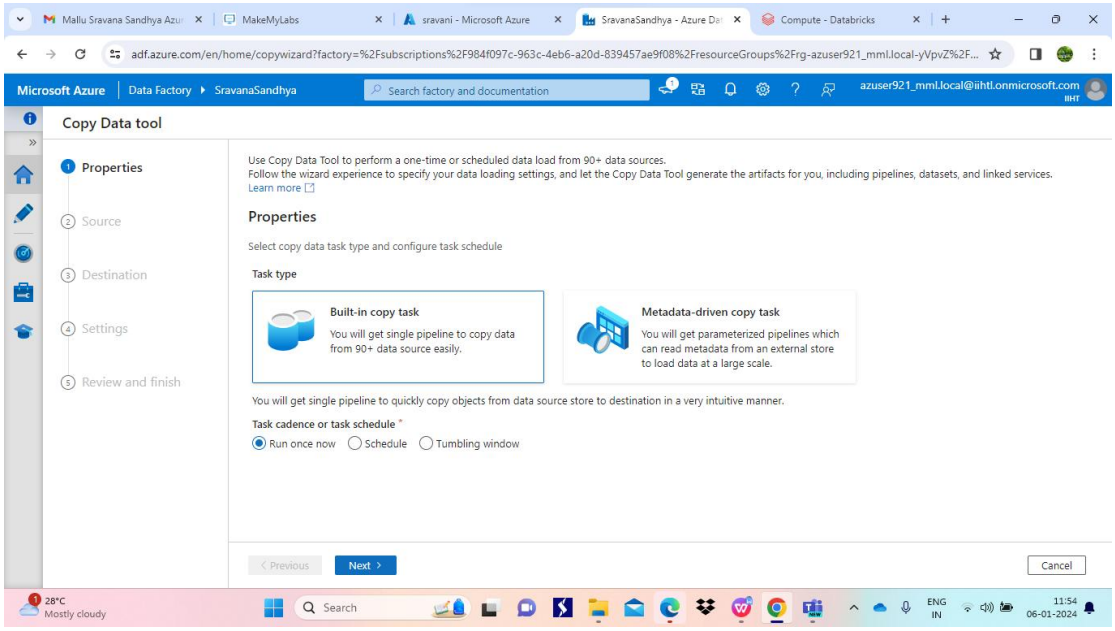

Mallu Sravana Sandhya Azure D
MakeMyLabs
Create Data Factory - Microsoft
Compute - Databricks
portal.azure.com/#create/Microsoft.DataFactory
Microsoft Azure
Search resources, services, and docs (G+/)
azuser921_mml.local@ii...
IIHT (IIHTL.ONMICROSOFT.COM)
Home > Create a resource >
Create Data Factory
By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Terms for additional details.
Basics
Subscription
Azure subscription 1
Resource group
rg-azuser921_mml.local-yVpvZ
Name
SravanaSandhya
Region
East US
Version
V2
Networking
Connect via
Public endpoint
Previous
Next
Create
Give feedback
28°C
Mostly cloudy
Search
ENG
IN
11:44
06-01-2024
portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2F984f097c-963c-4eb6-a20d-839457ae9f08%2FresourceGroups...
Microsoft.DataFactory-20240106114258 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name : Microsoft.DataFactory-20240106114258
Start time : 06/01/2024, 11:44:14
Subscription : Azure subscription 1
Correlation ID : 124cf5f1-630c-45d7-bddd-2bd937cac...
Resource group : rg-azuser921_mml.local-yVpvZ
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Deployment succeeded
Deployment 'Microsoft.DataFactory-20240106114258' to resource group 'rg-azuser921_mml.local-yVpvZ' was successful.
Pin to dashbo...
Go to resource gr...
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



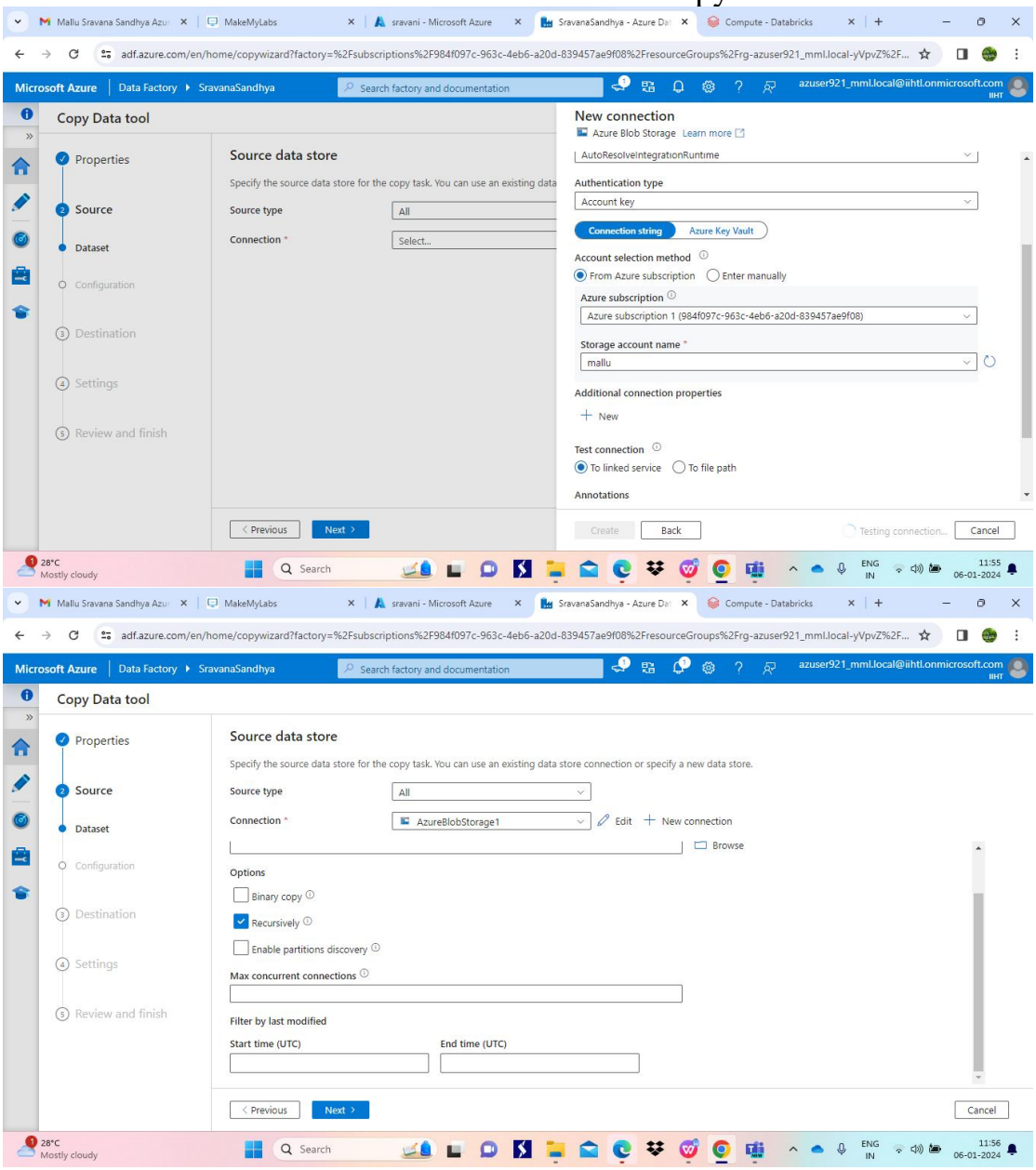

Click on ingest to create a pipeline and copy the file



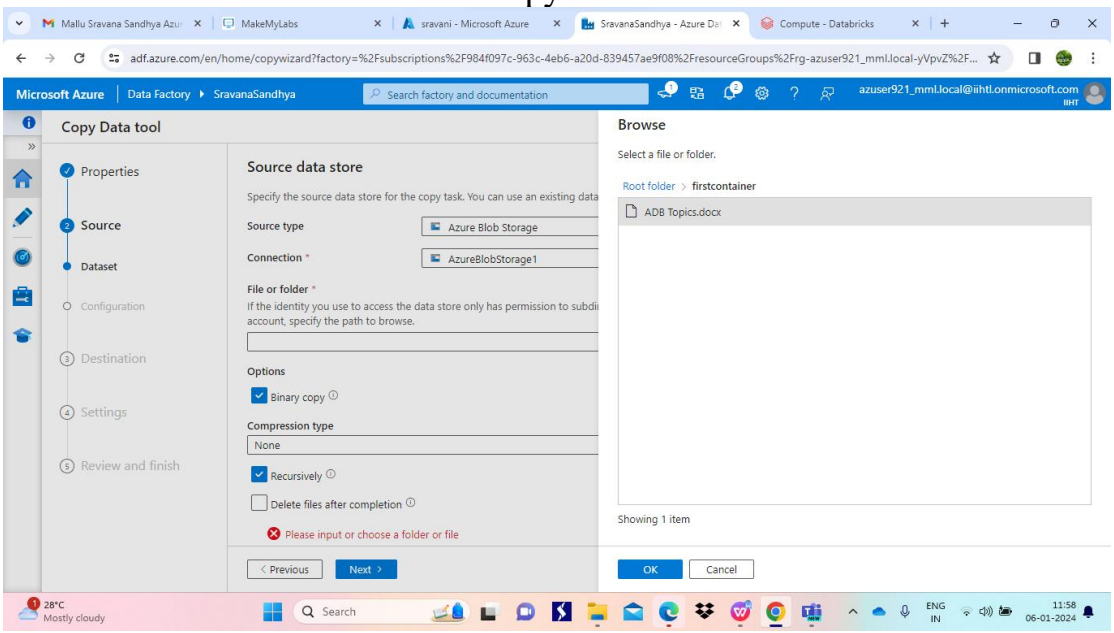
Click on next



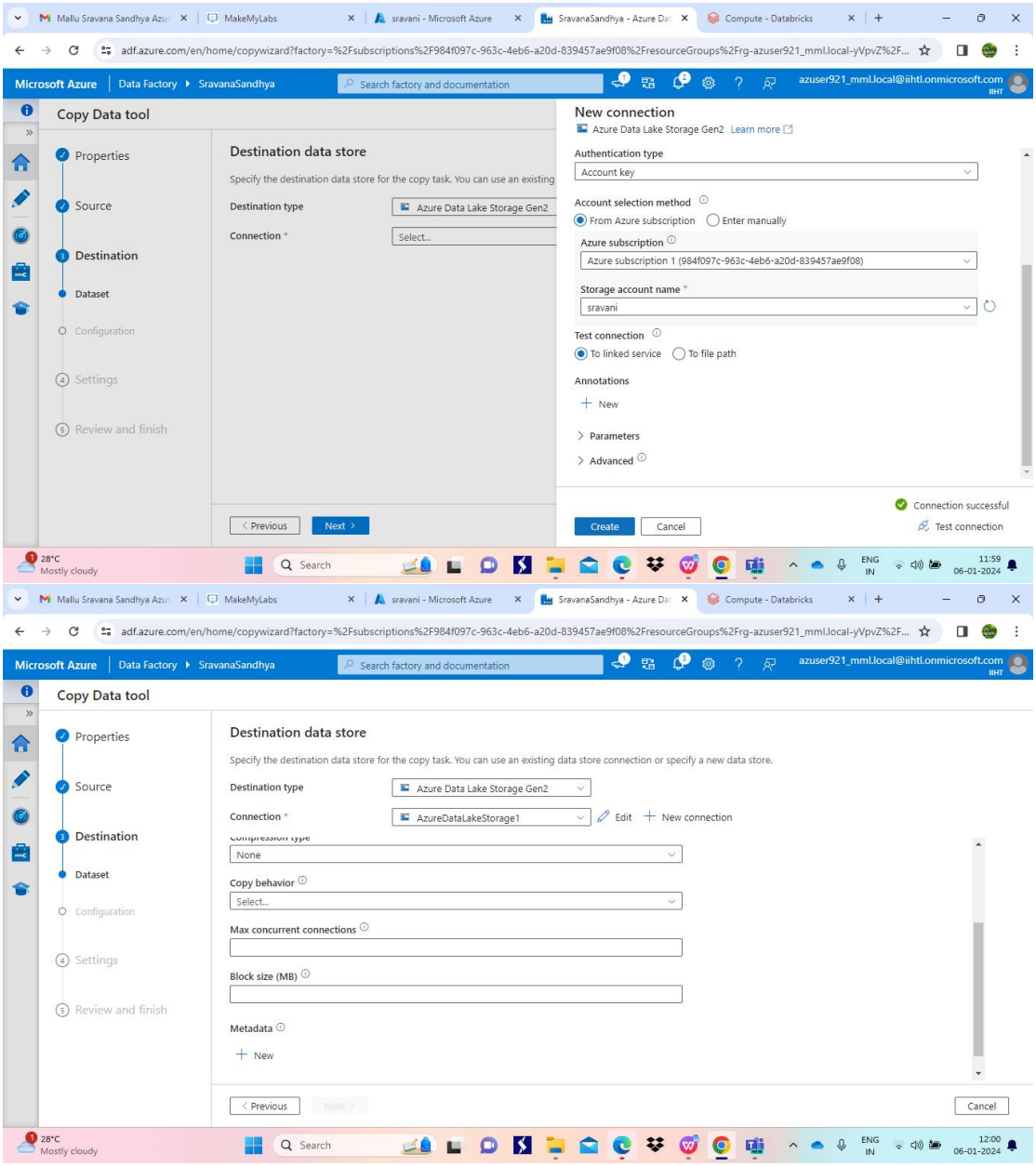
Fill the details in souce from where we have to copy file



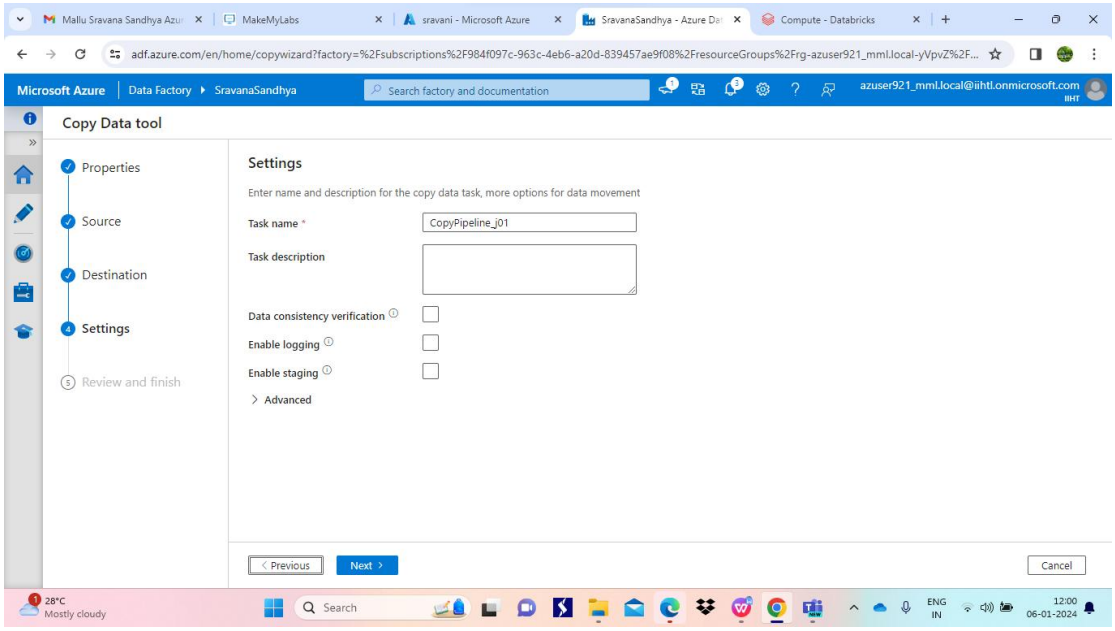
Select the file which we have to copy



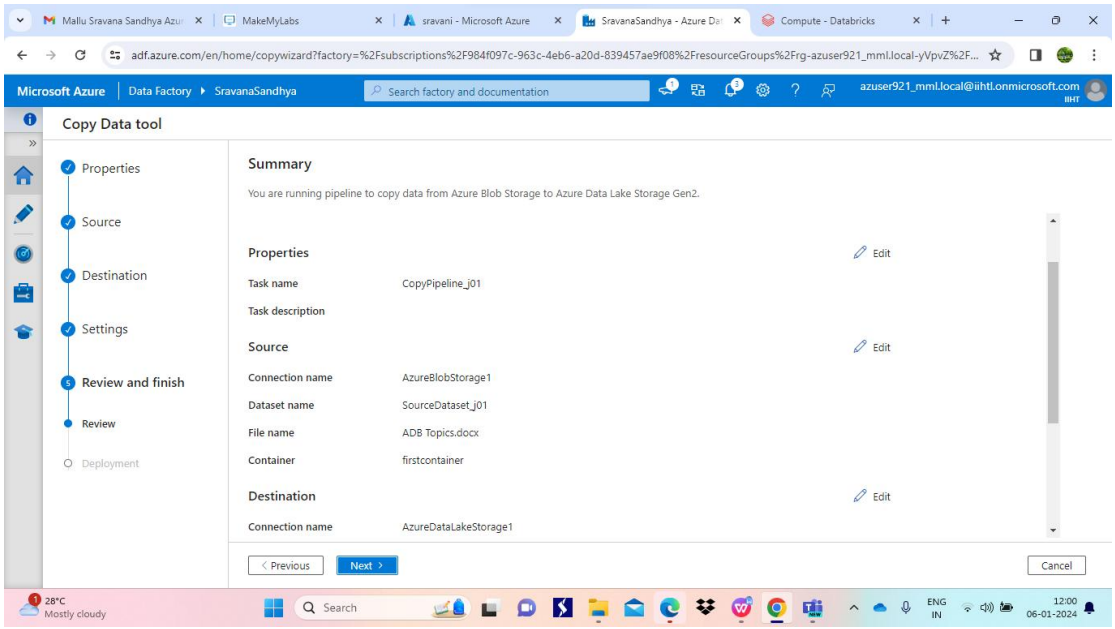
Give the destination location



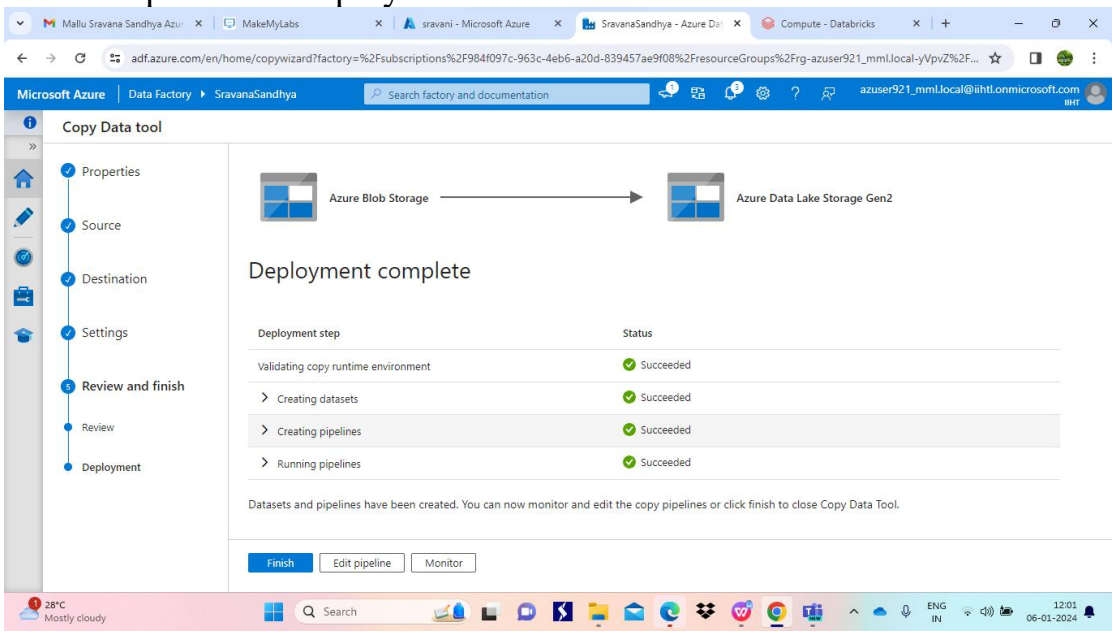
Click on next



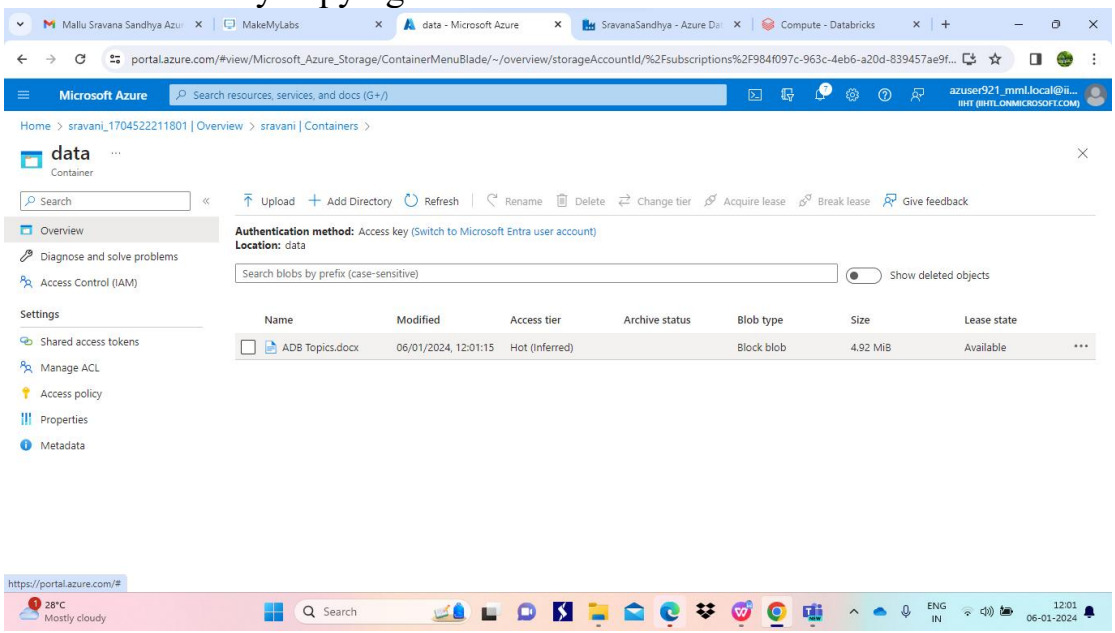

Review the souce and destination and click next



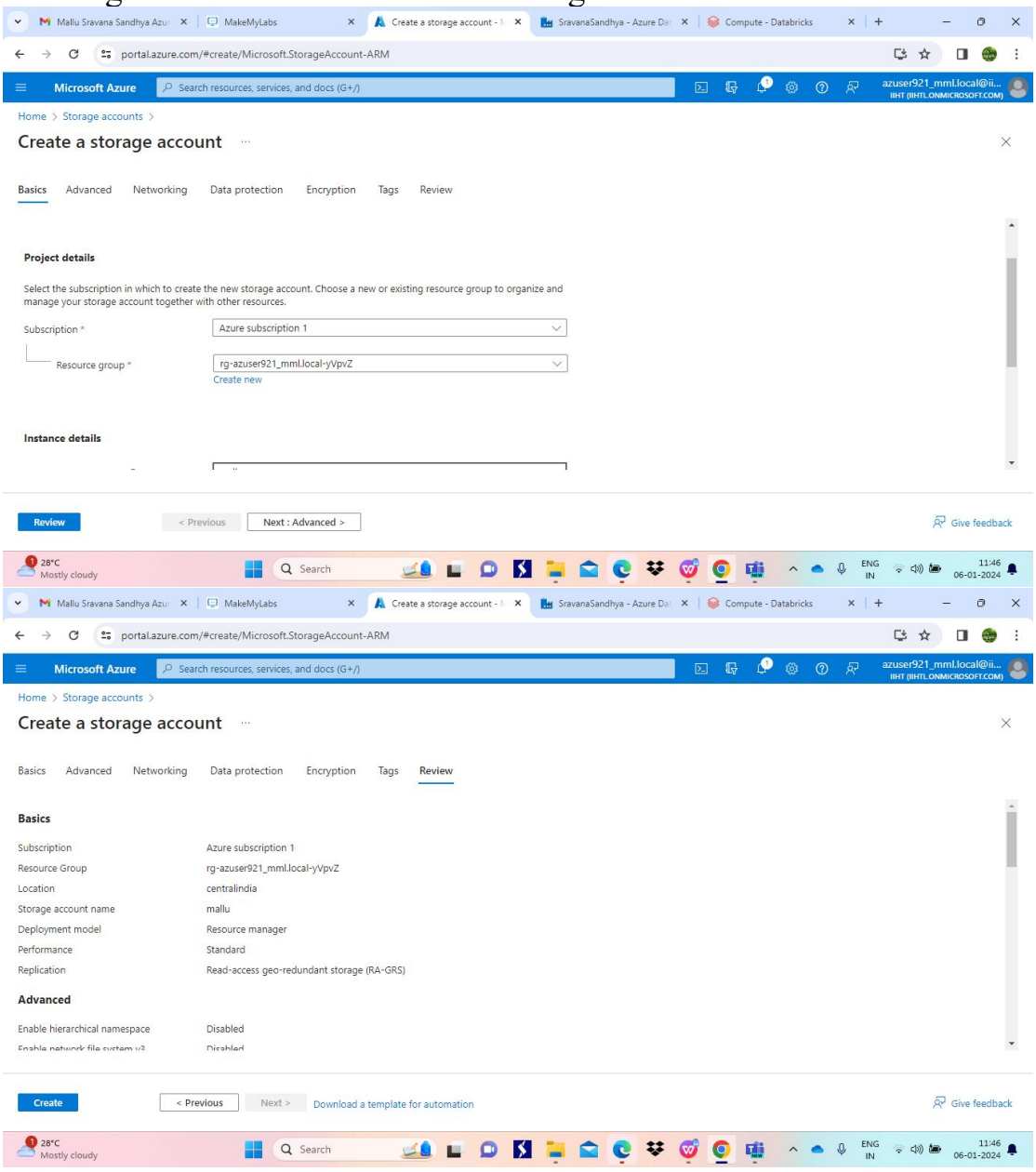
after completion of deployment click on finish

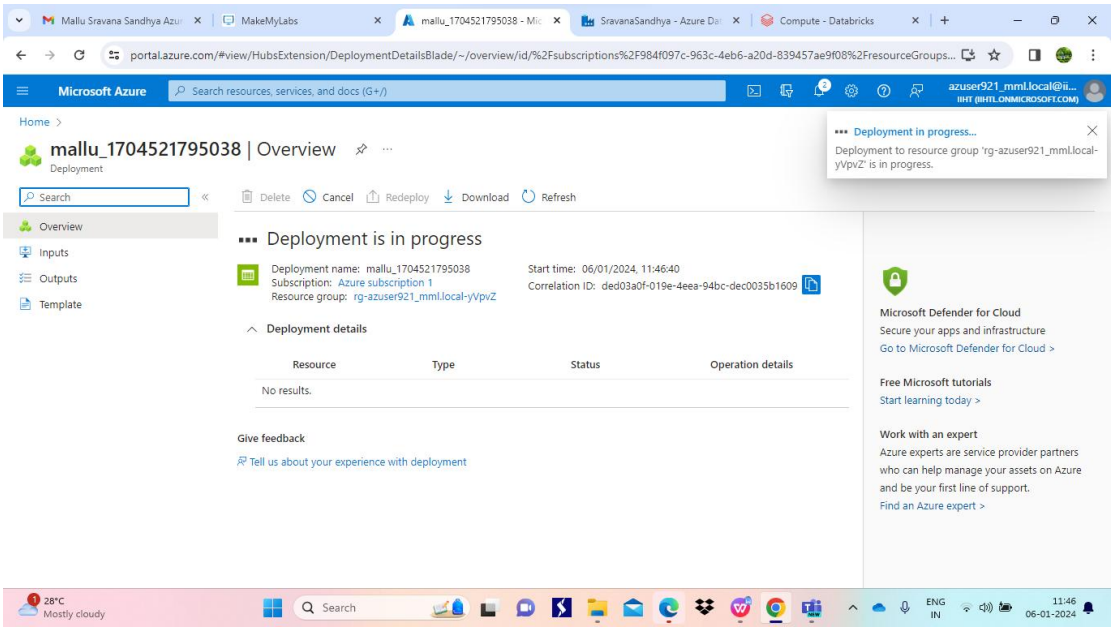



After successfully copying of file we can see the file in destination folder

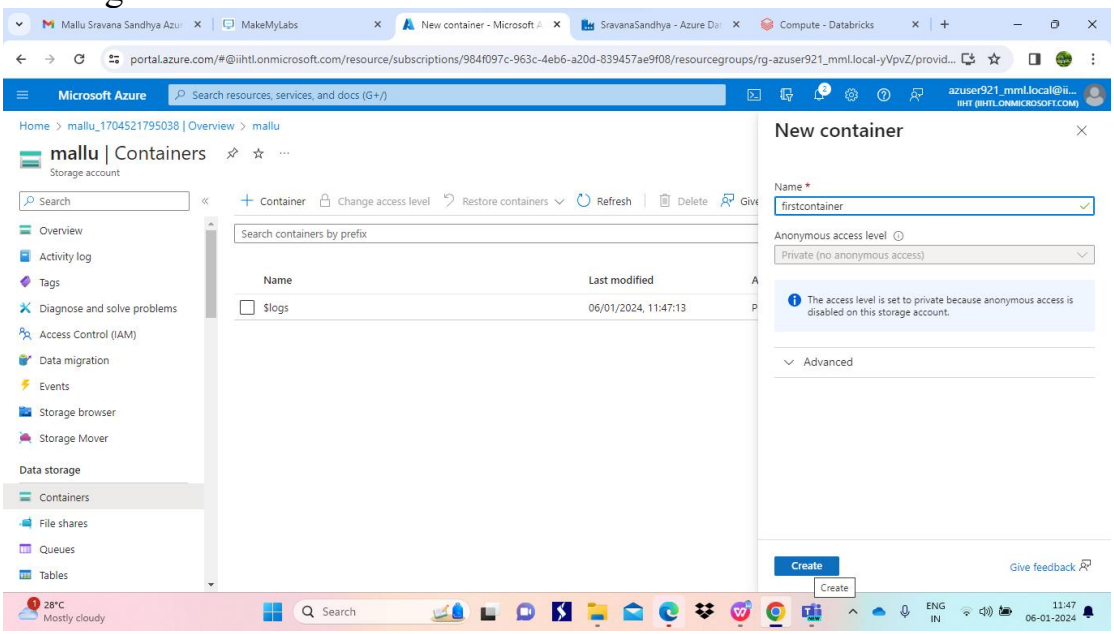


# Creating source location which is a storage account

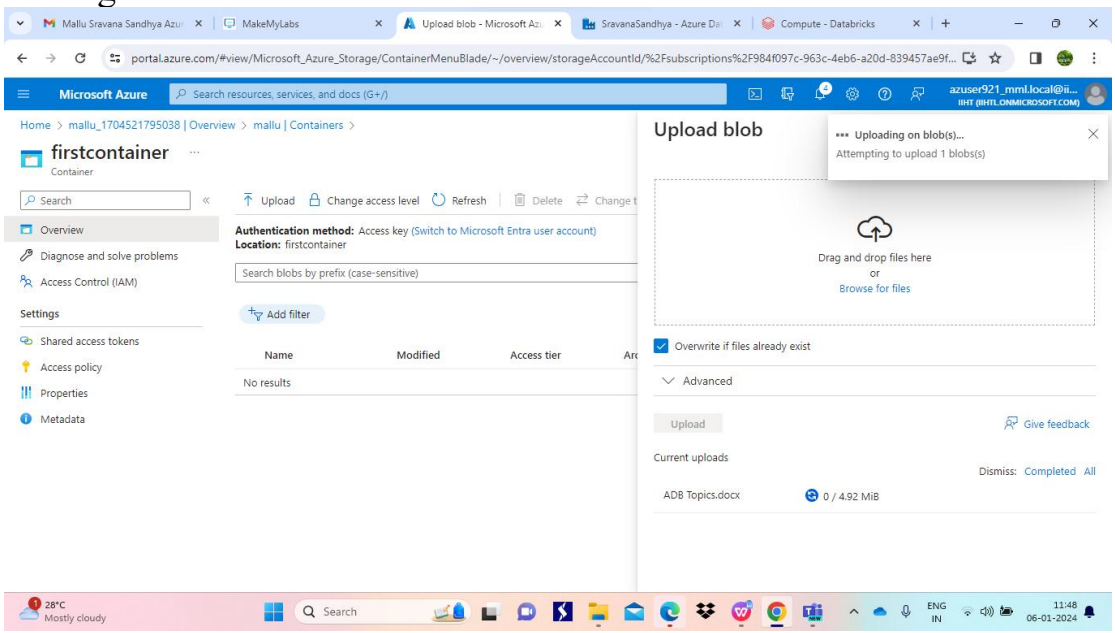





creating a folder in the source location



Adding file in the source location




Similarly create a  destination storage account and the file is copied successfully