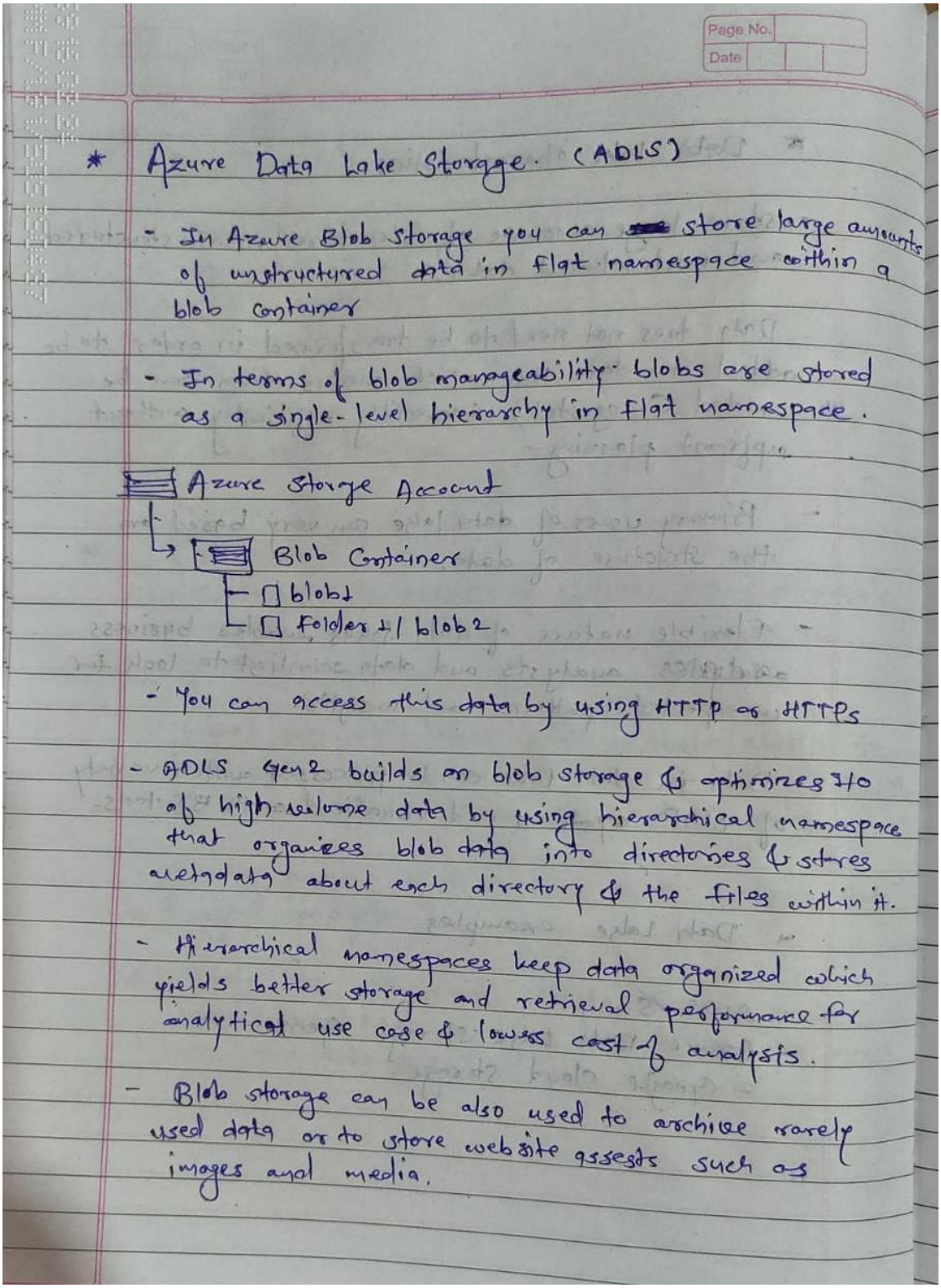
Data Engineering Batch 1 **Day 20:16/02/2024**

Name: Pradip Bochare

Notes



* Azure Data Lake Storage. (ADLS)

- In Azure Blob storage you can store large amounts of unstructured data in flat namespace within a blob container

- In terms of blob manageability. blobs are stored as a single-level hierarchy in flat namespace.

Azure Storage Account
  Blob Container
    blob1
    Folder1/blob2

- You can access this data by using HTTP or HTTPS

- ADLS Gen2 builds on blob storage & optimizes I/O of high volume data by using hierarchical namespace that organizes blob data into directories & stores metadata about each directory & the files within it.

- Hierarchical namespaces keep data organized which yields better storage and retrieval performance for analytical use case & lowers cost of analysis.

- Blob storage can be also used to archive rarely used data or to store website assests such as images and media.

Page No.
Date

* Stages for processing big data

Data lakes have fundamental role in a wide range of big data architectures. These arechitecture can involve creation of:
- enterprise data warehouse
- Advanced analytics against big data
- Real-time analytical solution.

* four stages for processing big data.

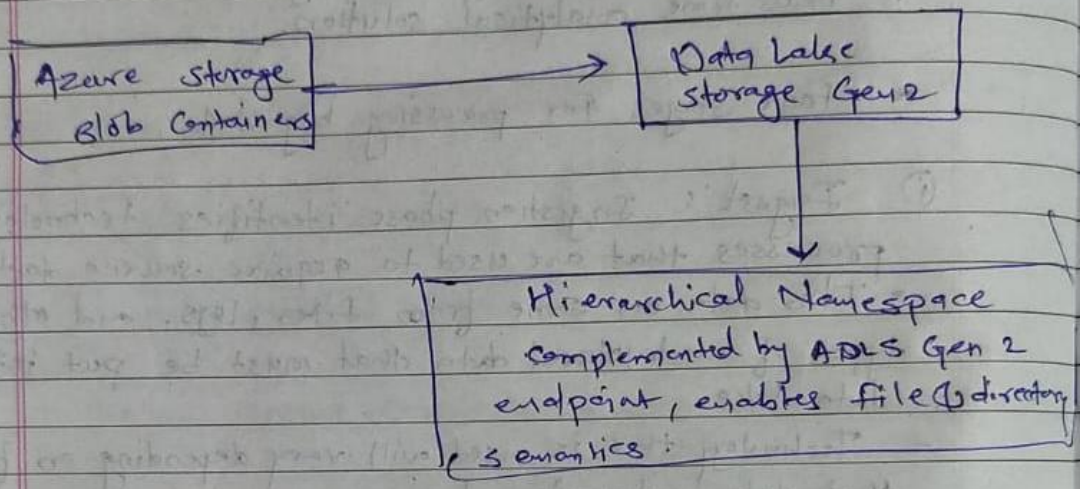
① Ingest: Ingestion phase identifies technology and processes that are used to acquire source data.
- This data can come from files, logs, and other types of unstructured data that must be put into data lake
- Technology that is used will vary depending on frequency that data is transferred.

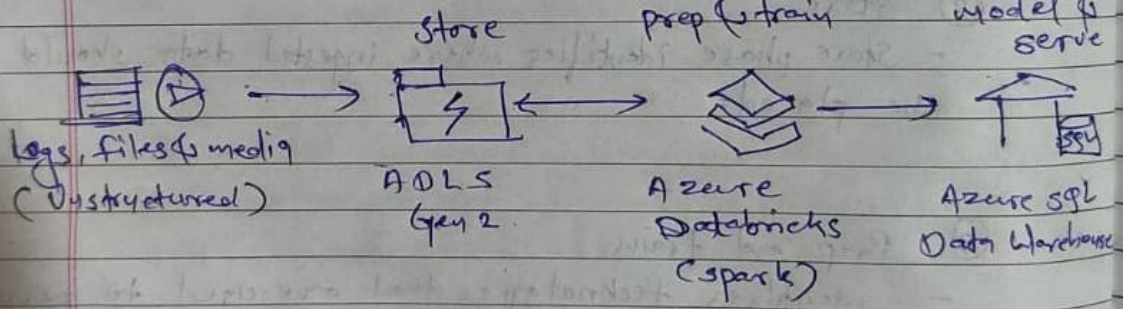
② store
- store phase identifies where ingested data should be placed.

③ Prep and train
- identifies technologies that are used to perform data preparation and model training and scoring for machine learning solutions

Page No.
Date

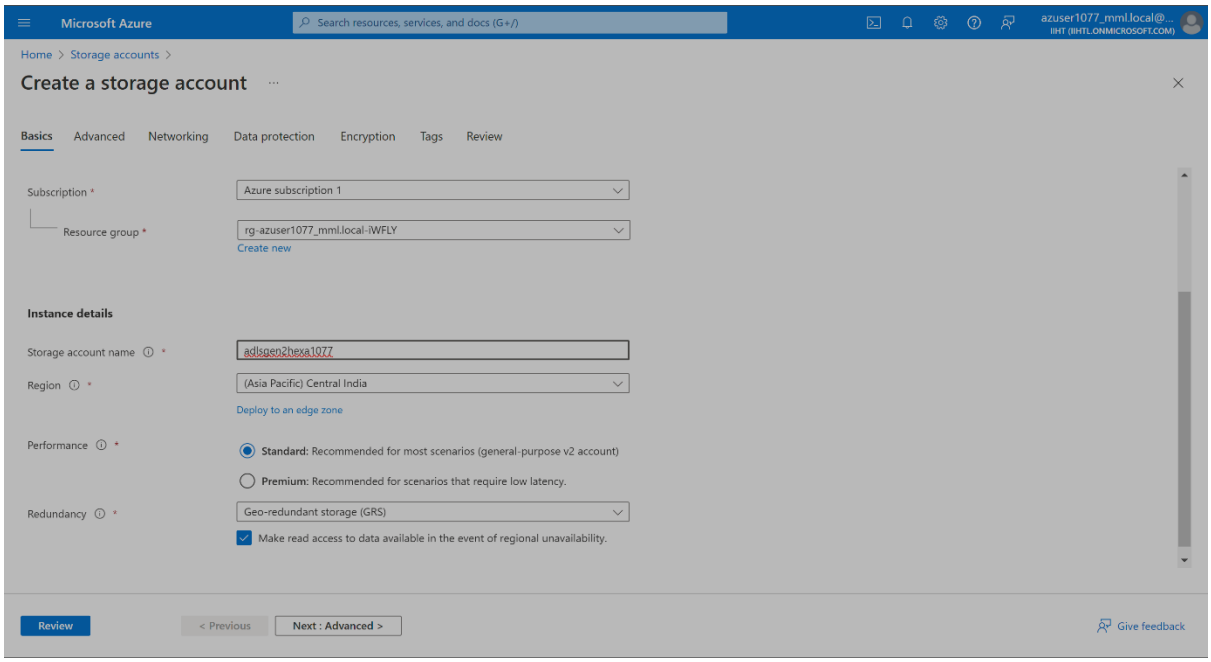
④ Model and serve
- This phase involves technologies that will present data to users.
- This technologies can include visualization tools such as Power BI, or analytical data stores such as Azure Synapse Analytics.



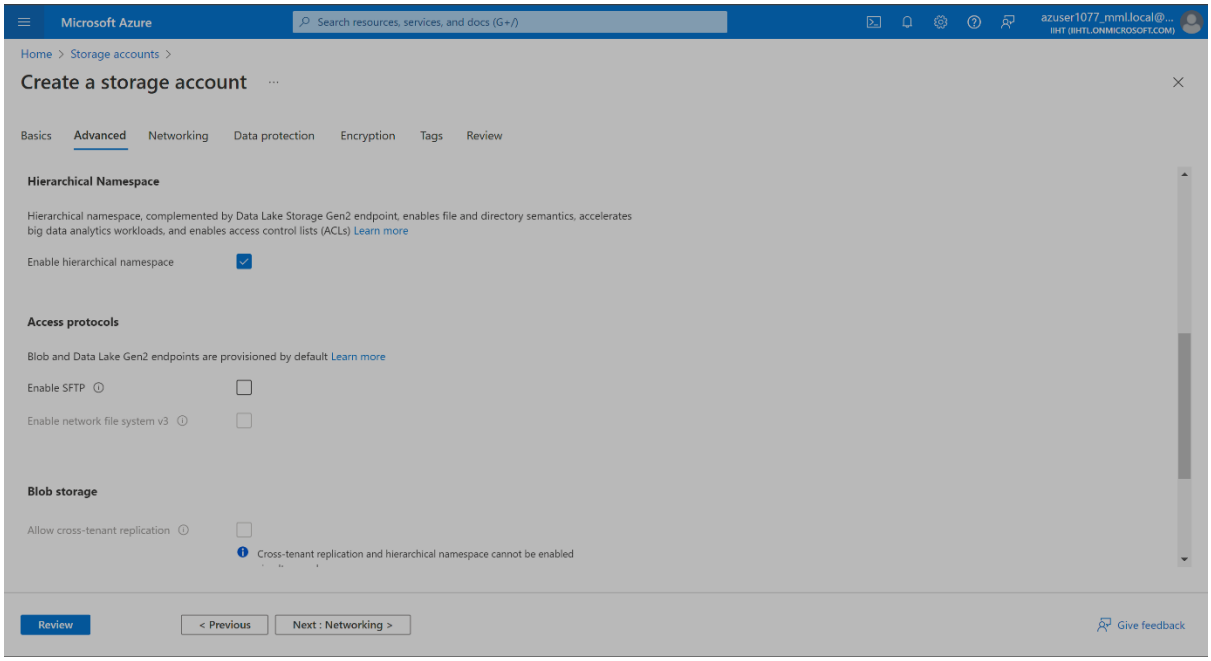


Making Azure Data Lake Storage Gen2

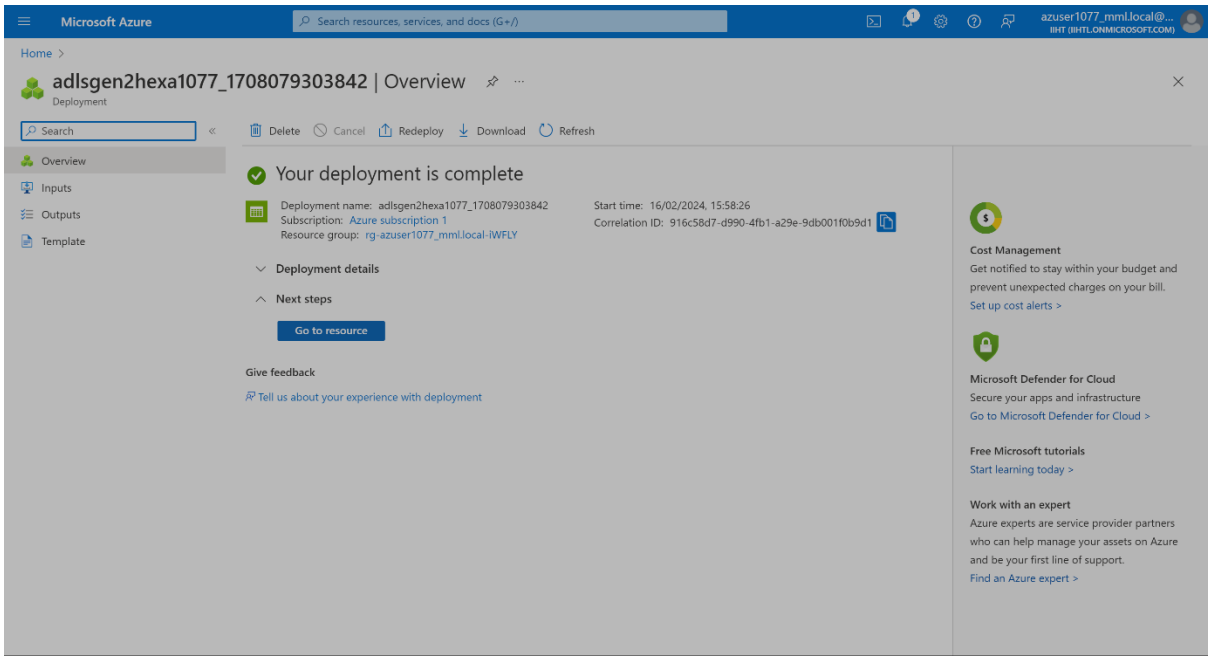
- Selecting Resource group and other parameters for ADLS



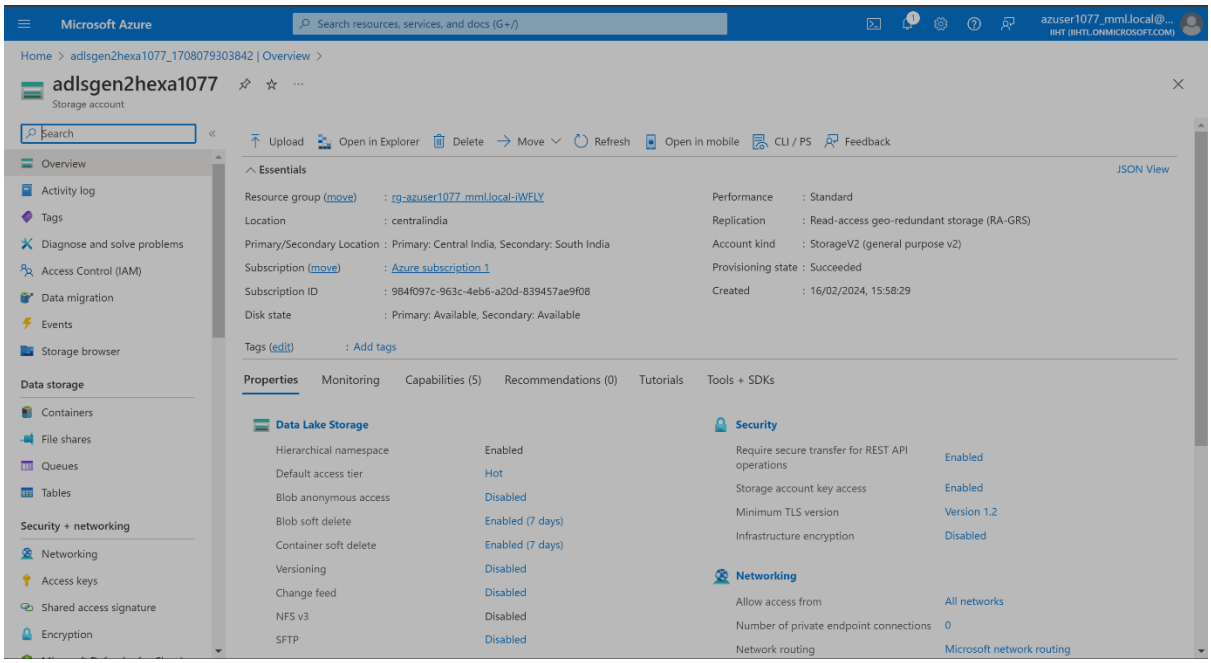
- Enabling Hierarchical Namespace



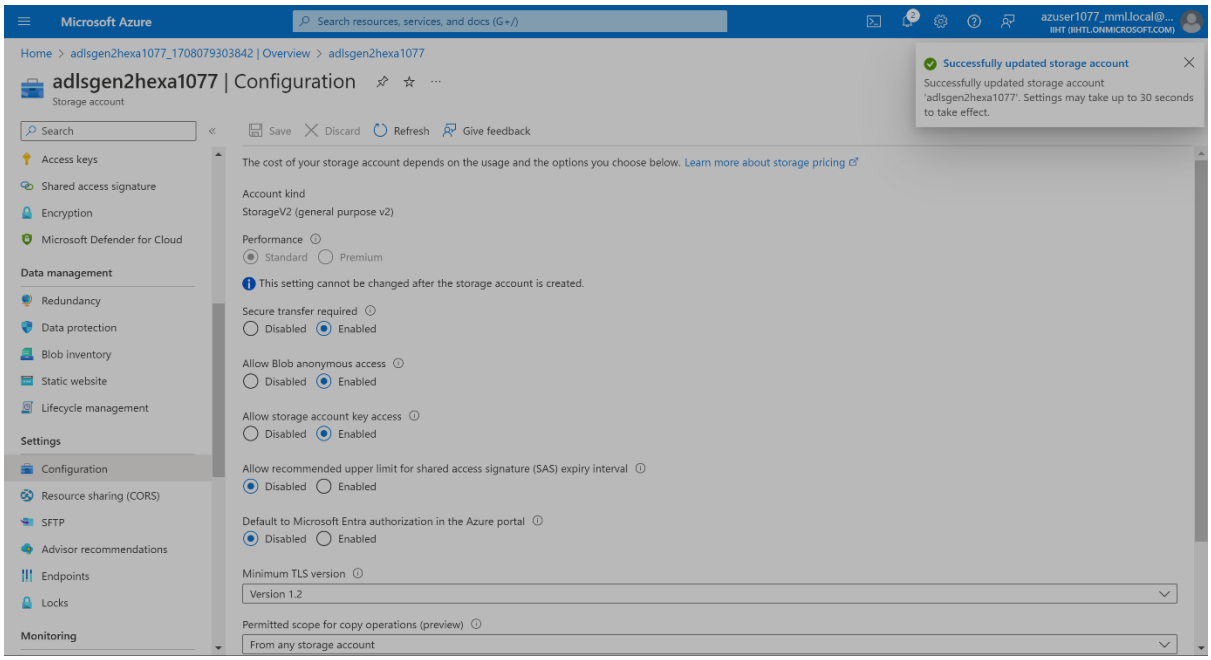
- ADLS is deployed



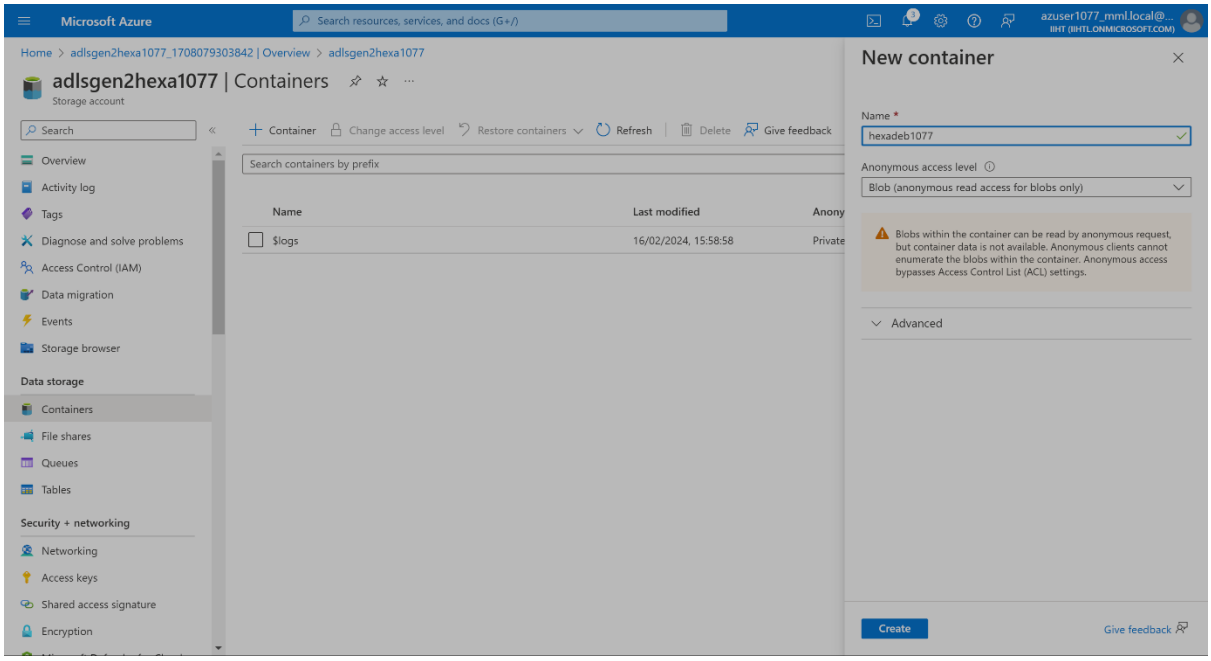
- Overview of ADLS



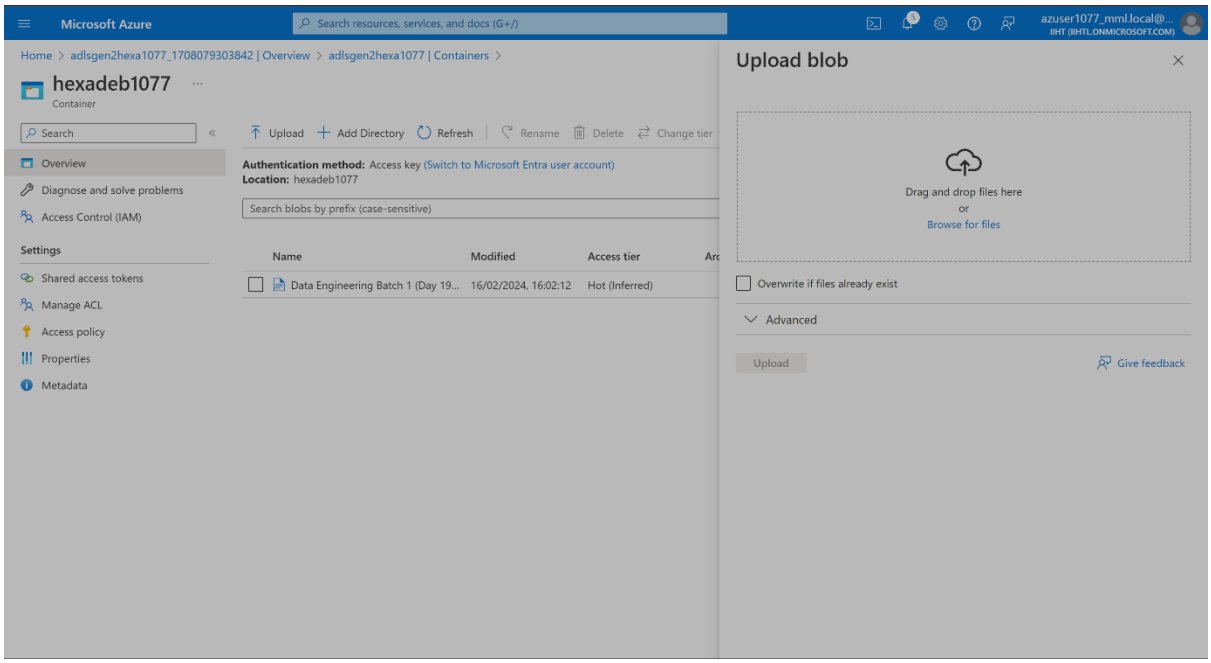
- Changing Configuration to allow Blob access



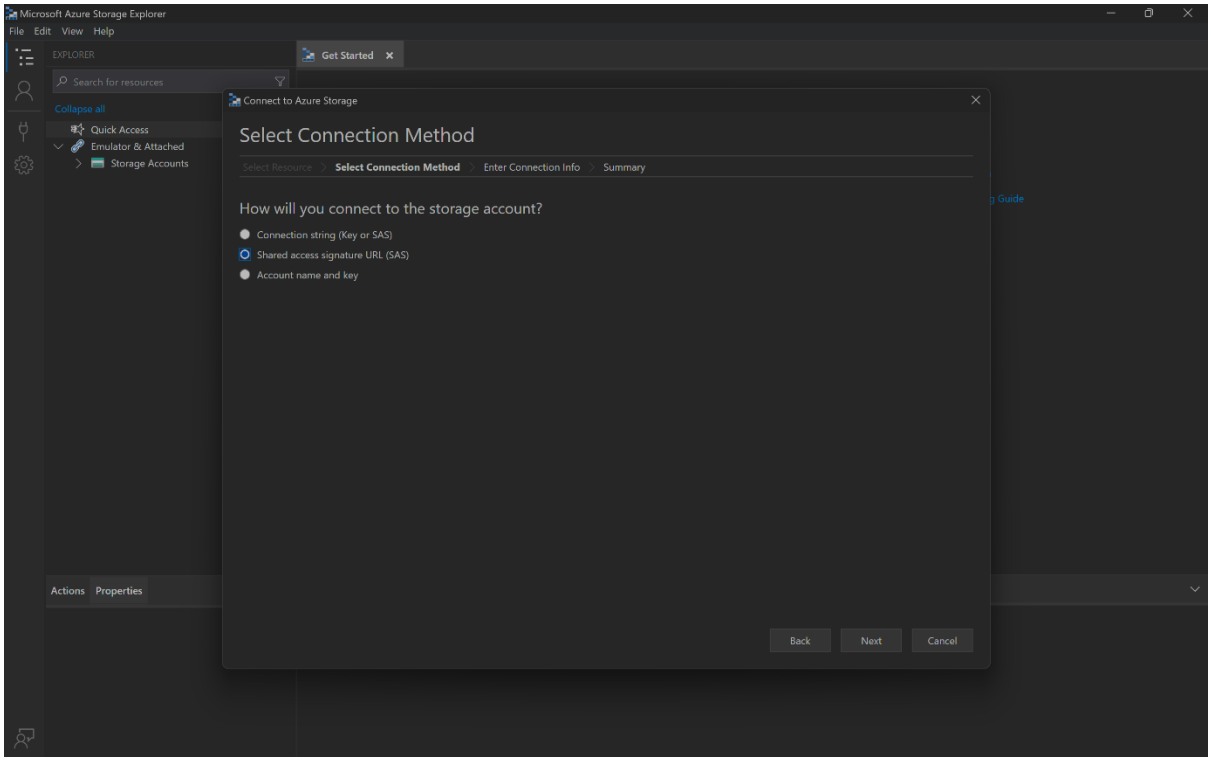
- Creating Container in ADLS



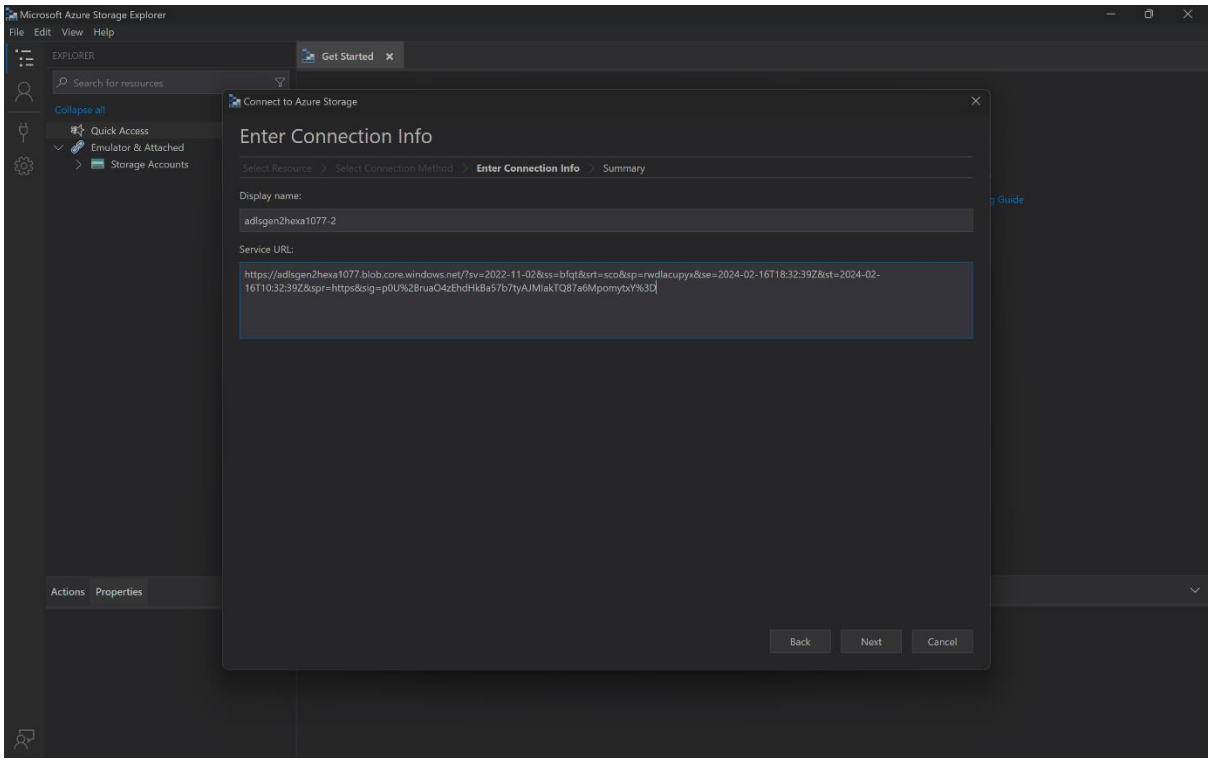
- Uploading files in ADLS



- Connecting ADLS to Microsoft Azure Storage Explorer



- Giving Service URL to connect



- ADLS is connected