Akilesh K

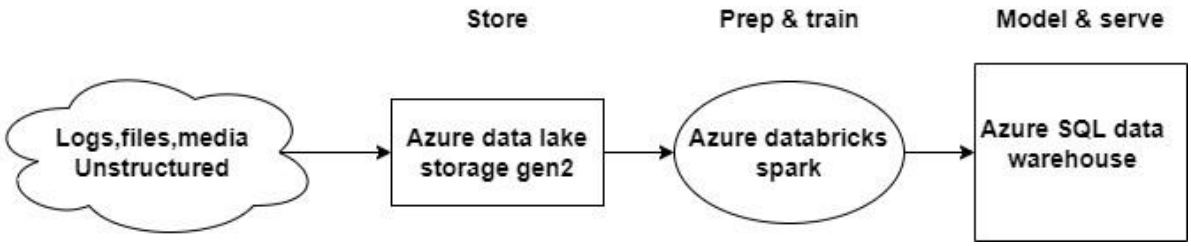
k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 16-02-24

DAY 20 – Azure -Data Lake Storage

ETL Model



The data lifecycle typically involves several key phases:

- **Ingestion:** This phase involves acquiring source data from various sources such as files, logs, and other unstructured data. Technologies like Azure Synapse Analytics, Azure Data Factory, Apache Kafka, or Stream Analytics may be used for batch or real-time data movement.
- **Storage:** In the storage phase, ingested data is placed into a secure and scalable storage solution. Azure Data Lake Storage Gen2 is commonly used for this purpose due to its compatibility with big data processing technologies.
- **Preparation and Training:** In this phase, technologies such as Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Machine Learning are used for data preparation and model training in machine learning solutions.
- **Model and Serving:** Finally, the model and serve phase involves presenting the data to users. This may involve visualization tools like Microsoft Power BI or analytical data stores such as Azure Synapse Analytics. Often, a combination of multiple technologies is used based on business requirements.

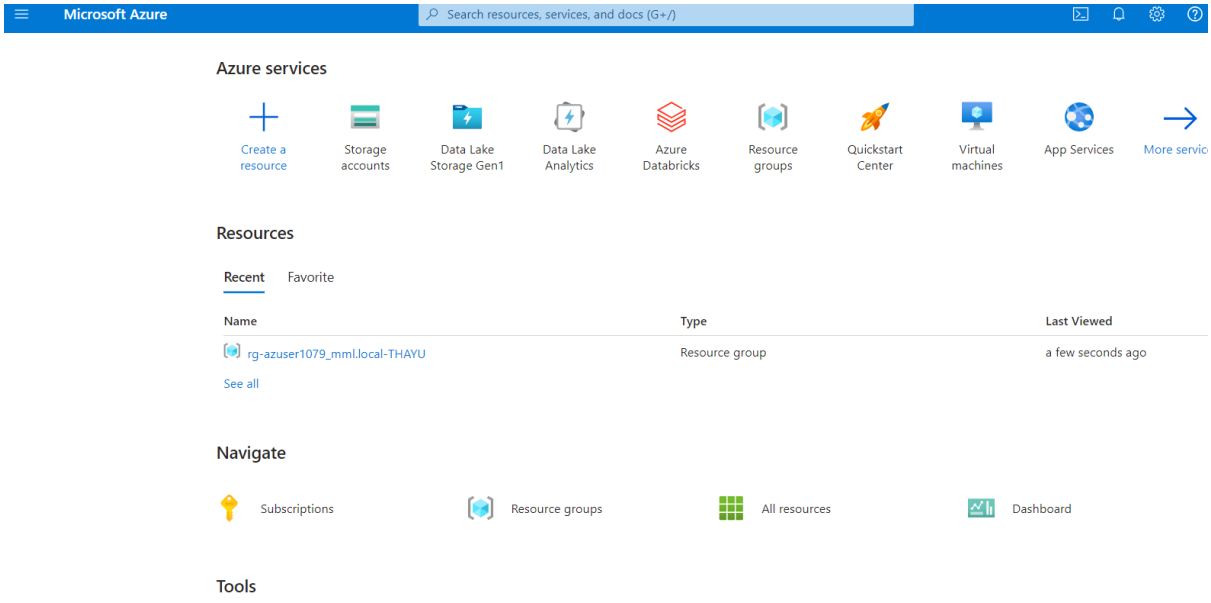
Overall, the data lifecycle encompasses processes for acquiring, storing, preparing, analyzing, and presenting data, with various technologies utilized at each stage to enable efficient data management and decision-making.

Data lake storage

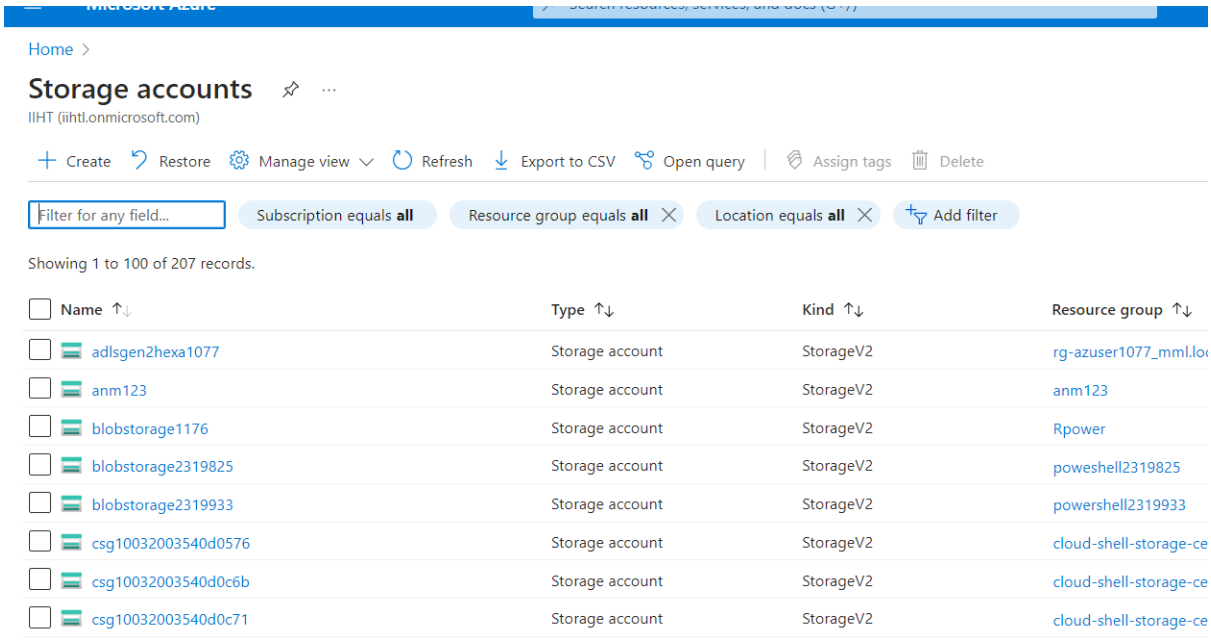
- If you want to store data without performing analysis on the data, set the Hierarchical Namespace option to Disabled to set up the storage account as an Azure Blob storage account. You can also use blob storage to archive rarely used data or to store website assets such as images and media.

- If you are performing analytics on the data, set up the storage account as an Azure Data Lake Storage Gen2 account by setting the Hierarchical Namespace option to Enabled. Because Azure Data Lake Storage Gen2 is integrated into the Azure Storage platform, applications can use either the Blob APIs or the Azure Data Lake Storage Gen2 file system APIs to access data.

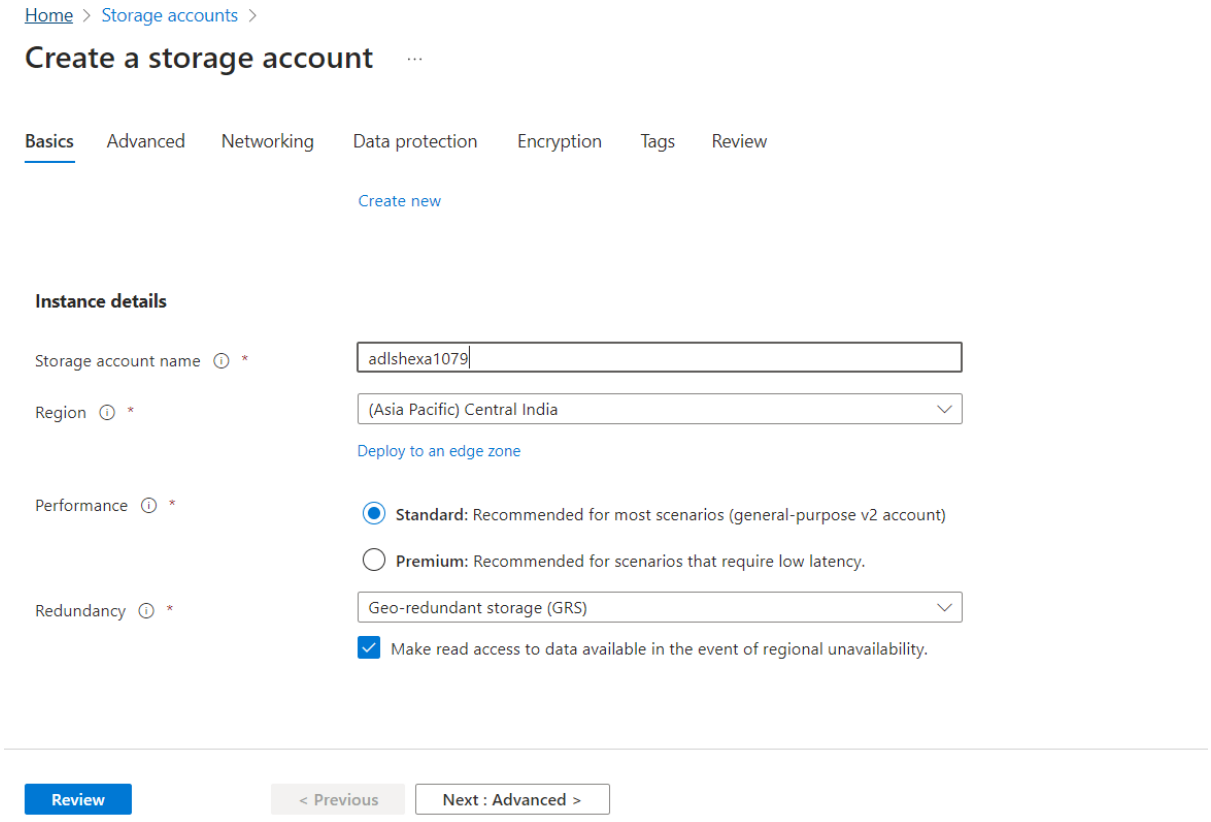
Login into Azure portal



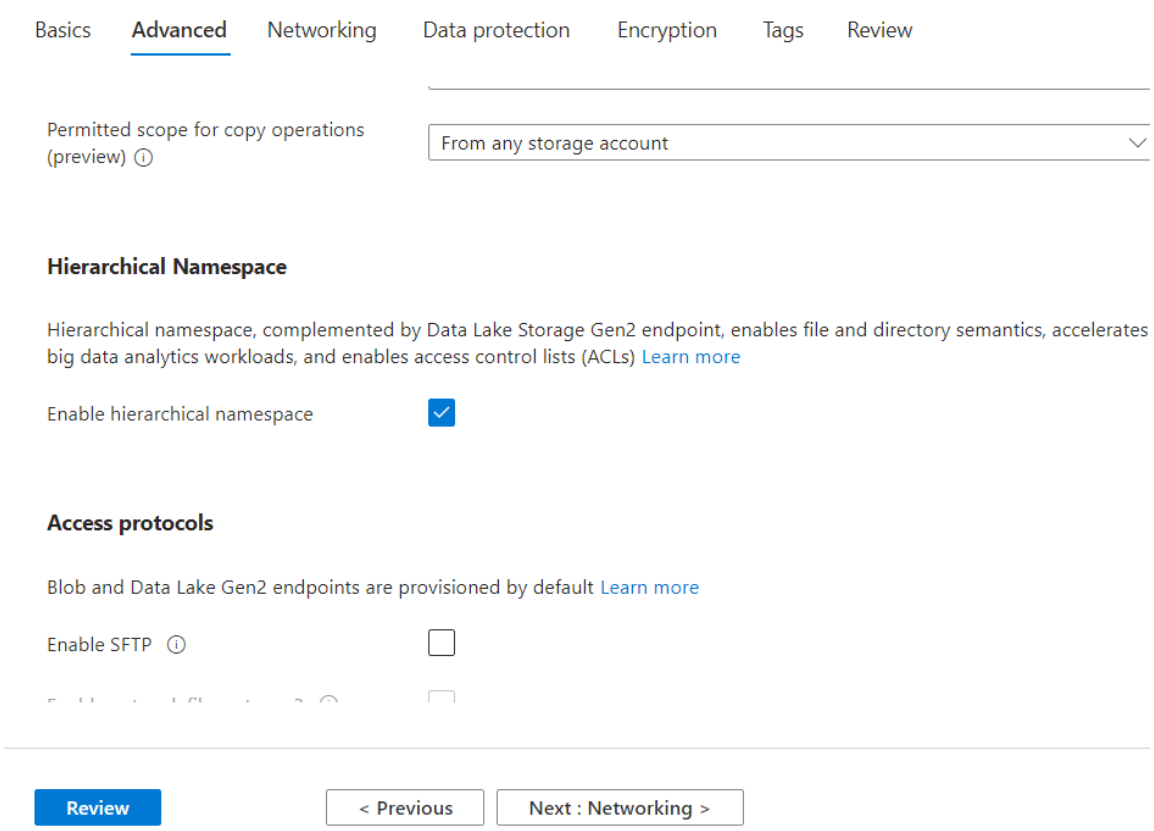
Go to a storage account



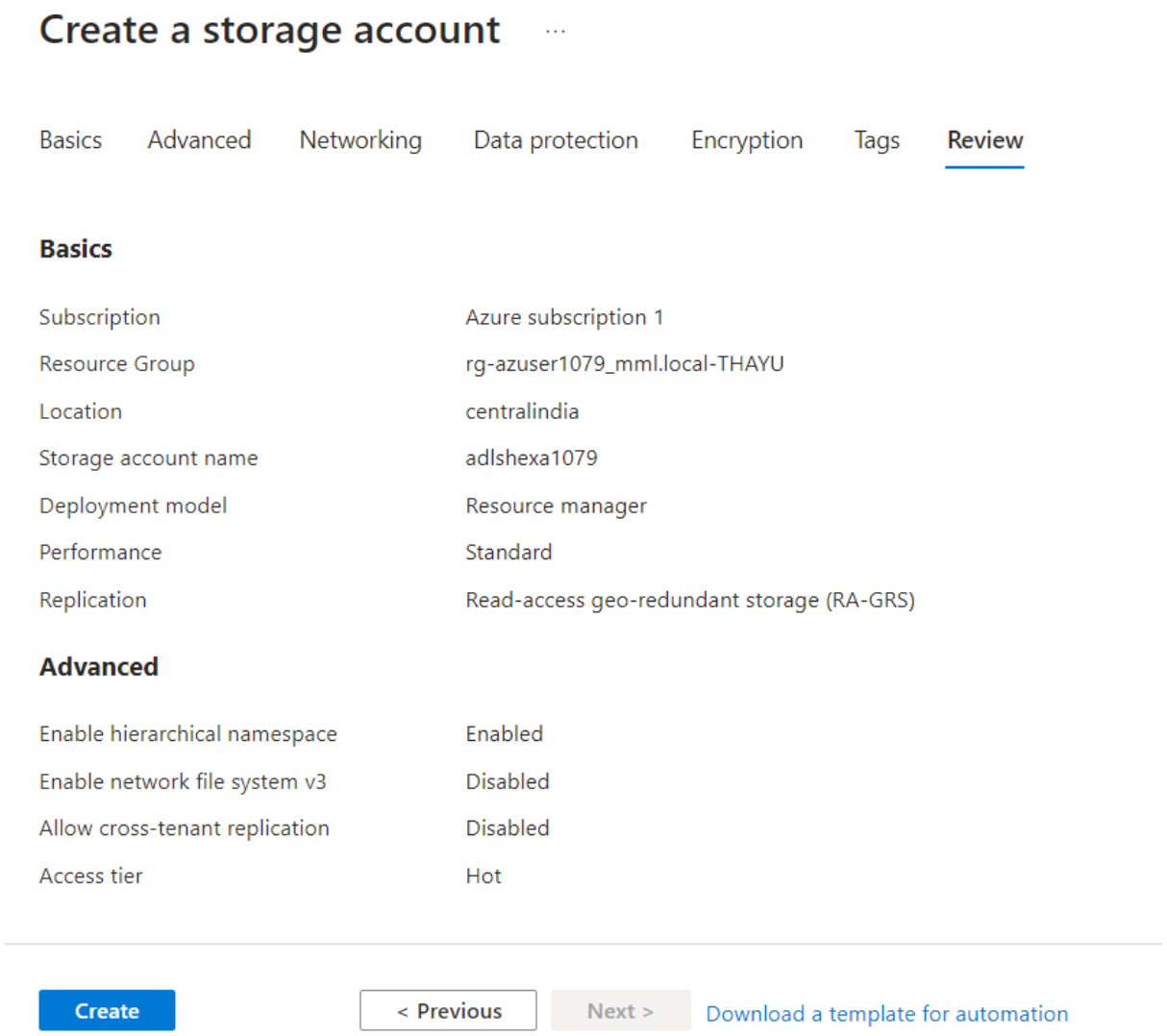
Create a storage account



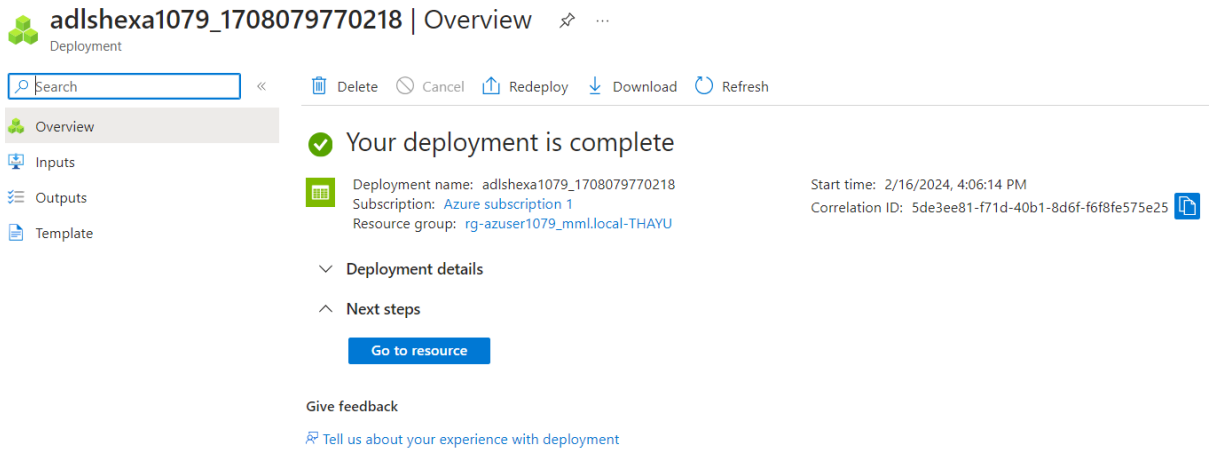
In the Advanced option ,check enable hierachial Namespace



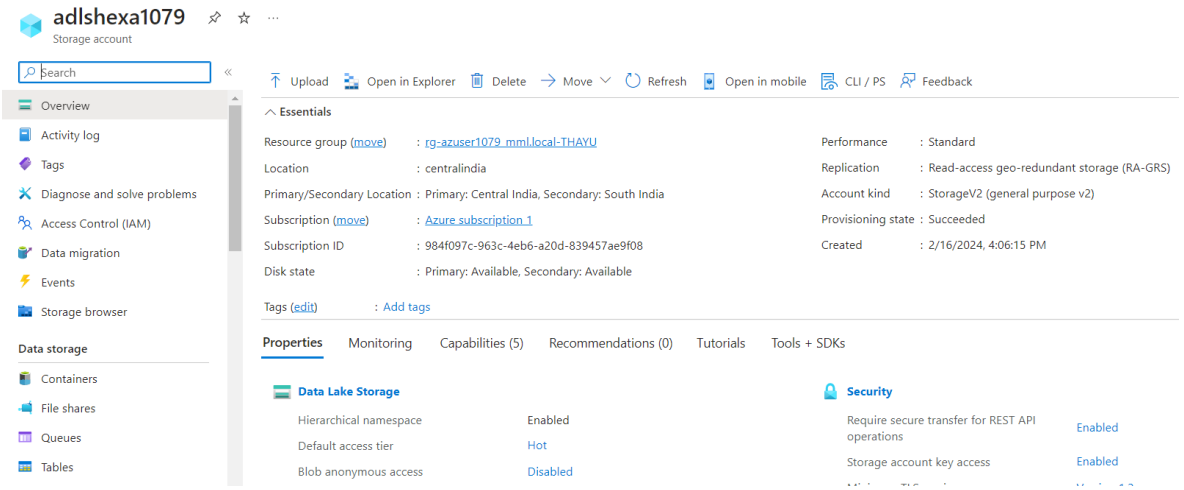
Review the storage account



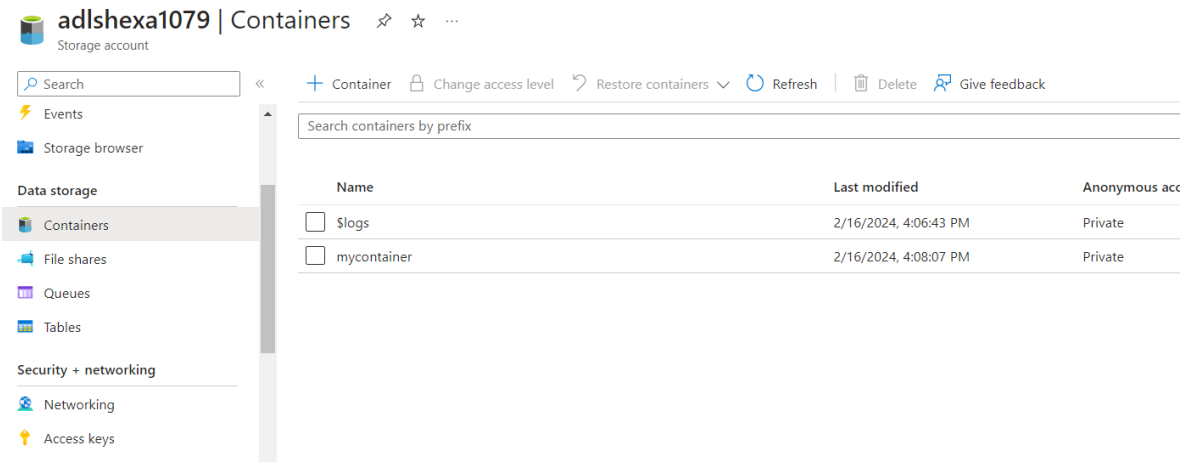
Storage account is deployed



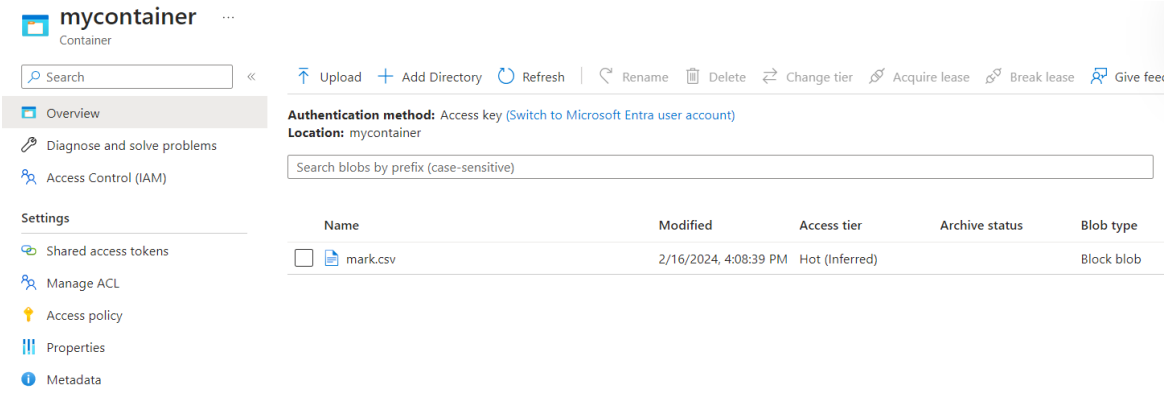
Open the storage account



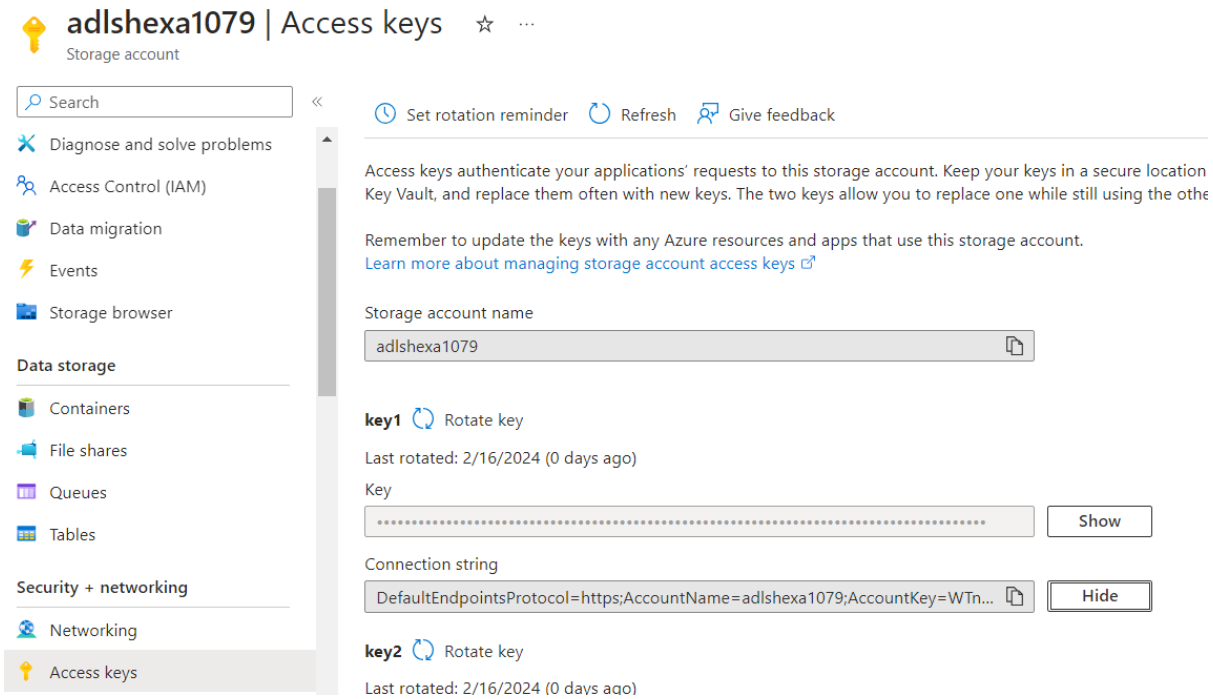
Create a container



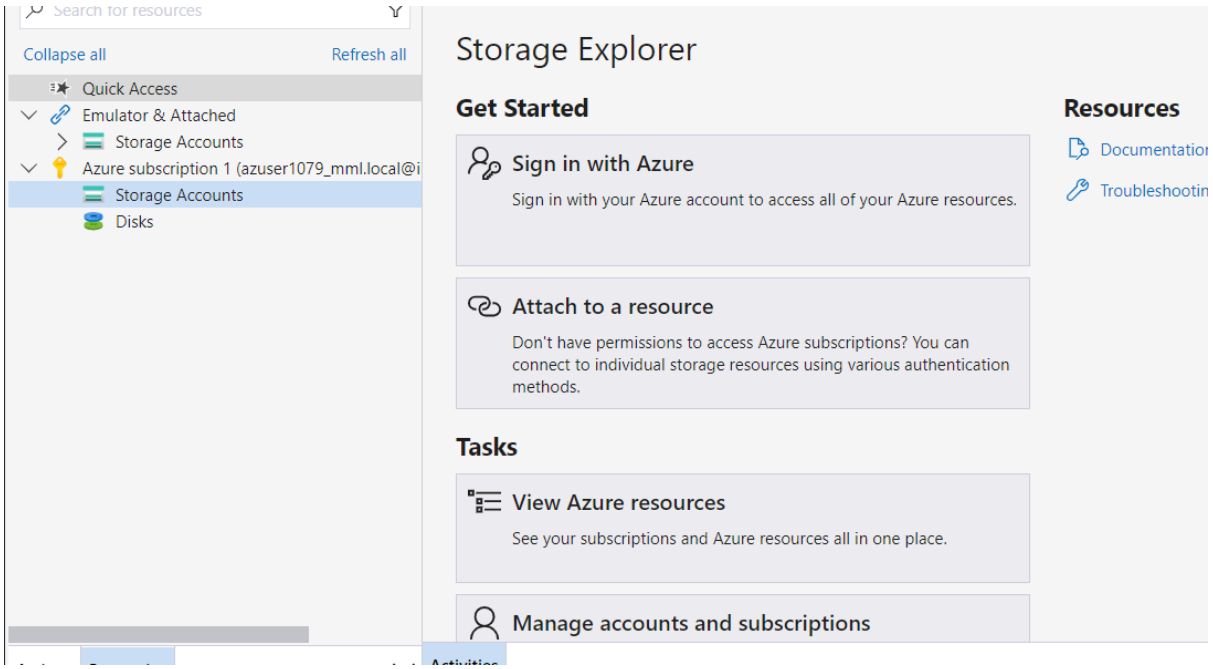
Add sample files



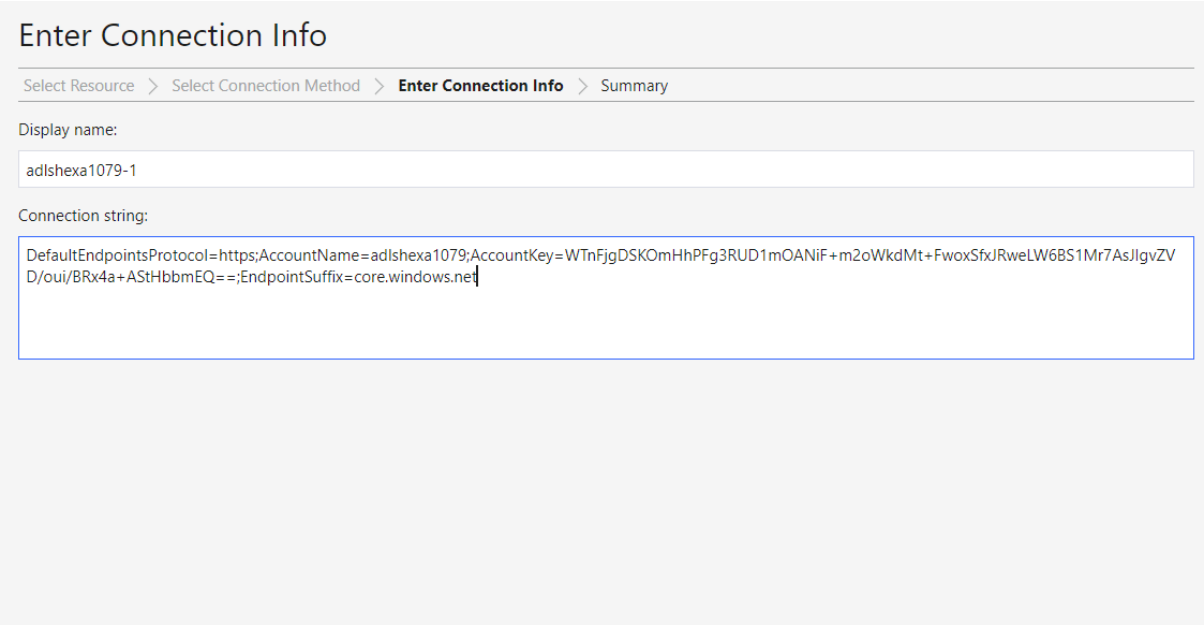
Copy the access key-connection string



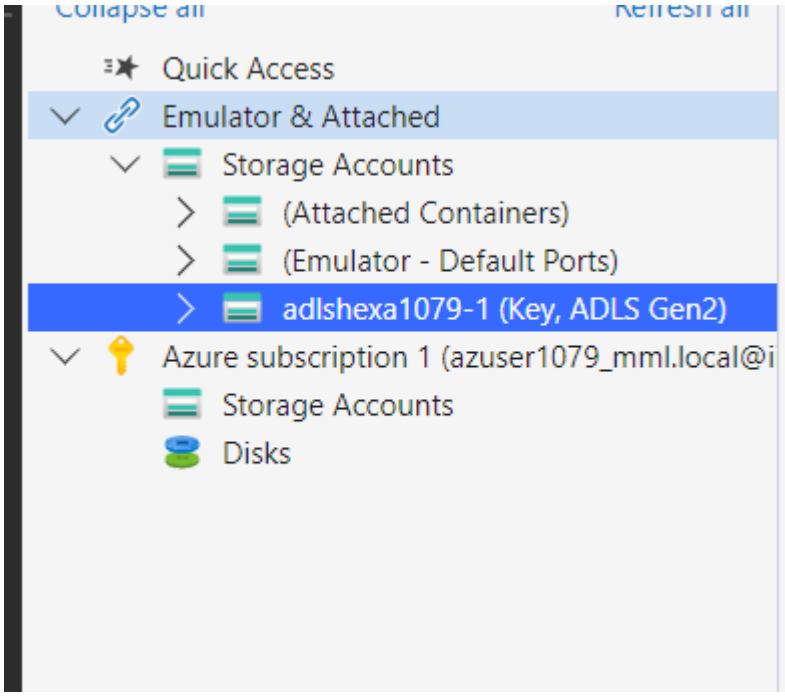
Open azure storage explorer



Connect it by pasting the connection string



ADLS Gen2 is added to the file explorer



Sample file is present here in the dashboard