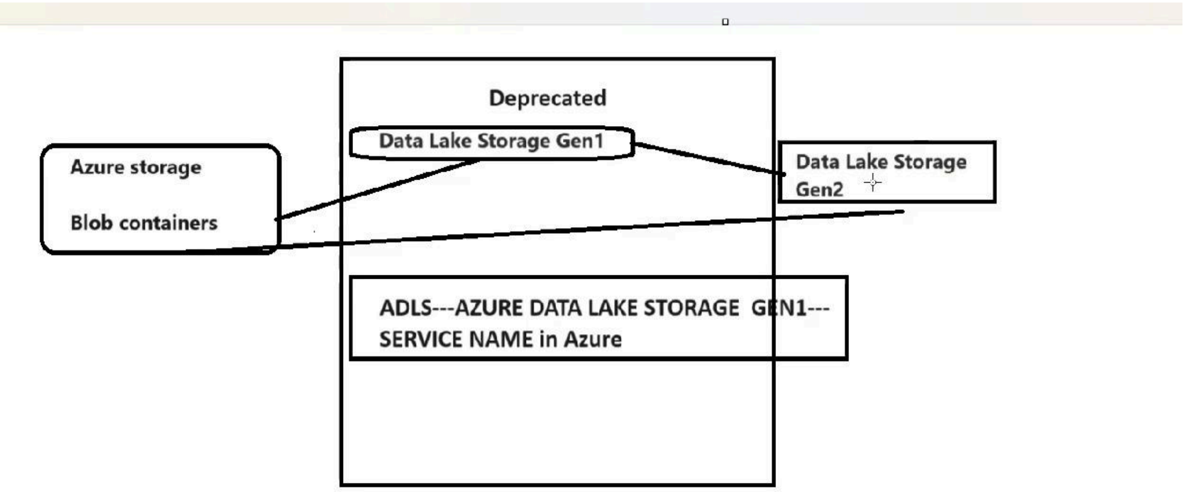
NAME : AKULA SHARATH CHANDRA
BATCH: DATA ENGINEERING
DATE :16-02-2024
TOPICS: ADLS,CREATE A DATA LAKE STORAGE GEN2 ACCOUNT,UNDERSTAND THE STAGES FOR PROCESSING BIG DATA

1)ADLS:

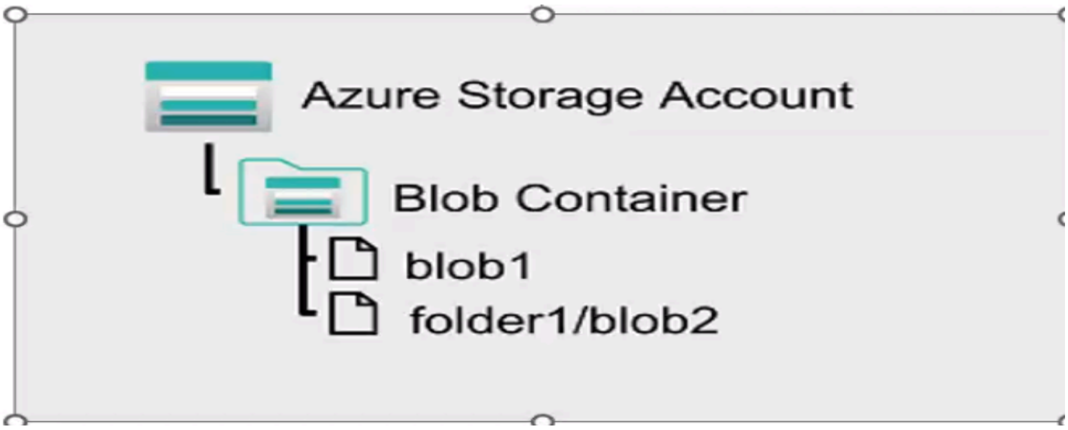
—>In Azure Blob storage, you can store large amounts of unstructured ("object") data in a flat namespace within a blob container. Blob names can include "/" characters to organize blobs into virtual folders", but in terms of blob manageability the blobs are stored as a single-level hierarchy in a flat namespace.

—>Azure Data Lake Storage Gen2 builds on blob storage and optimizes I/O of high- volume data by using a hierarchical namespace that organizes blob data into directories, and stores metadata about each directory and the files within it. This structure allows operations, such as directory renames and deletes, to be performed in a single atomic operation. Flat namespaces, by contrast, require several operations proportionate to the number of objects in the structure. Hierarchical namespaces keep the data organized, which yields better storage and retrieval performance for an analytical use case and lowers the cost of analysis.

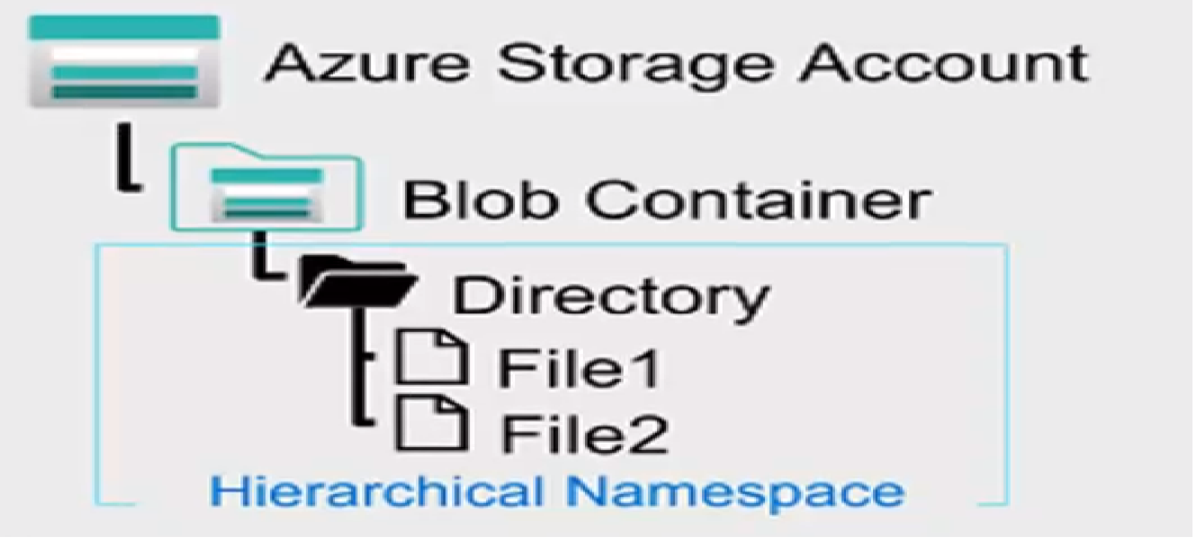
ARCHITECTURE OF ADLS:



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.



2)CREATE A DATA LAKE STORAGE GEN2 ACCOUNT:

—>Firstly sign in to the Microsoft Azure account.

—> Then we need to create a storage account by adding a resource group and adding hierarchical namespace in the advanced settings. This is because if we want to create a storage gen 2 account we need to enable that .

—>Then after creating the storage account we need to create the container.

—>Then after Creating the container we need to go to the shared access signature .

—>Thereby we need to allow the resource types(service,container,object).

—>After allowing that we need to click on the generate SAS and connecting string (this is to connect to the storage explorer).

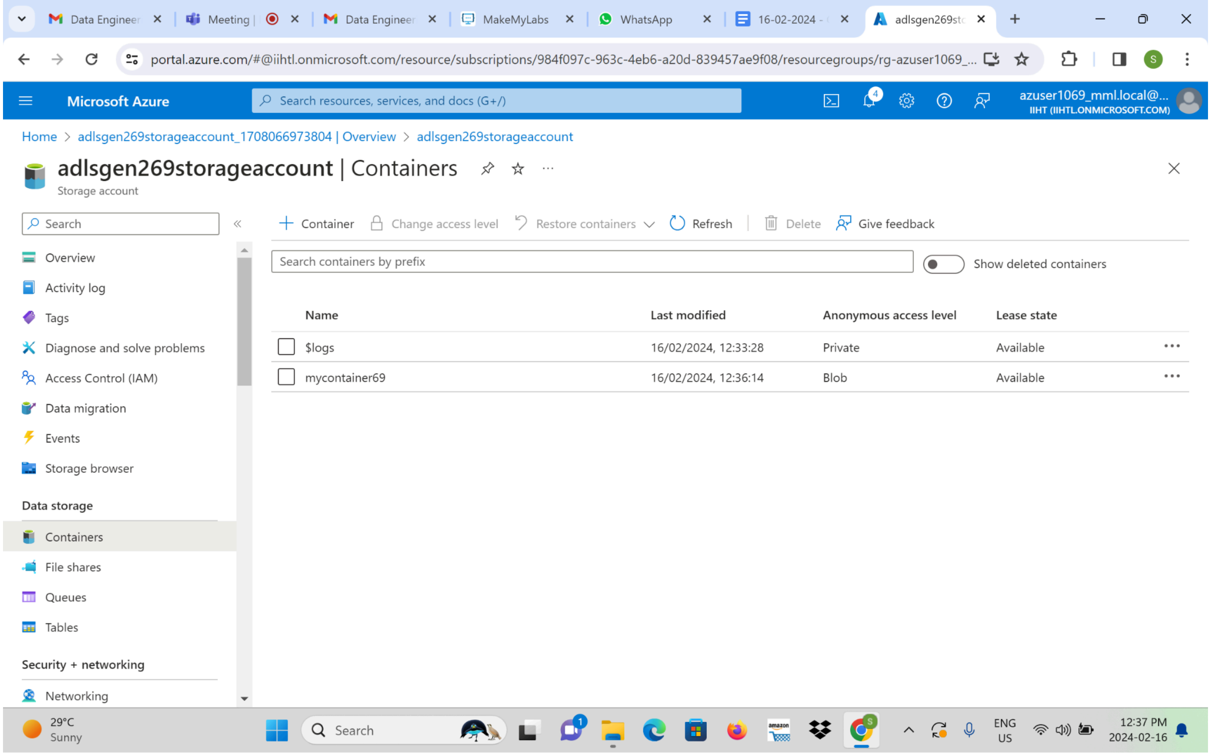
—>After that we need to open the storage explorer and click on the connect to azure resource.

—>Then we need to click on the storage account or service.

—>Thereby we need to select a connection method to shared access signature URL(SAS).

—> Then we need to paste the blob service SAS URL in it and we need to click on the next.

—> Thereby the connection will be added successfully.



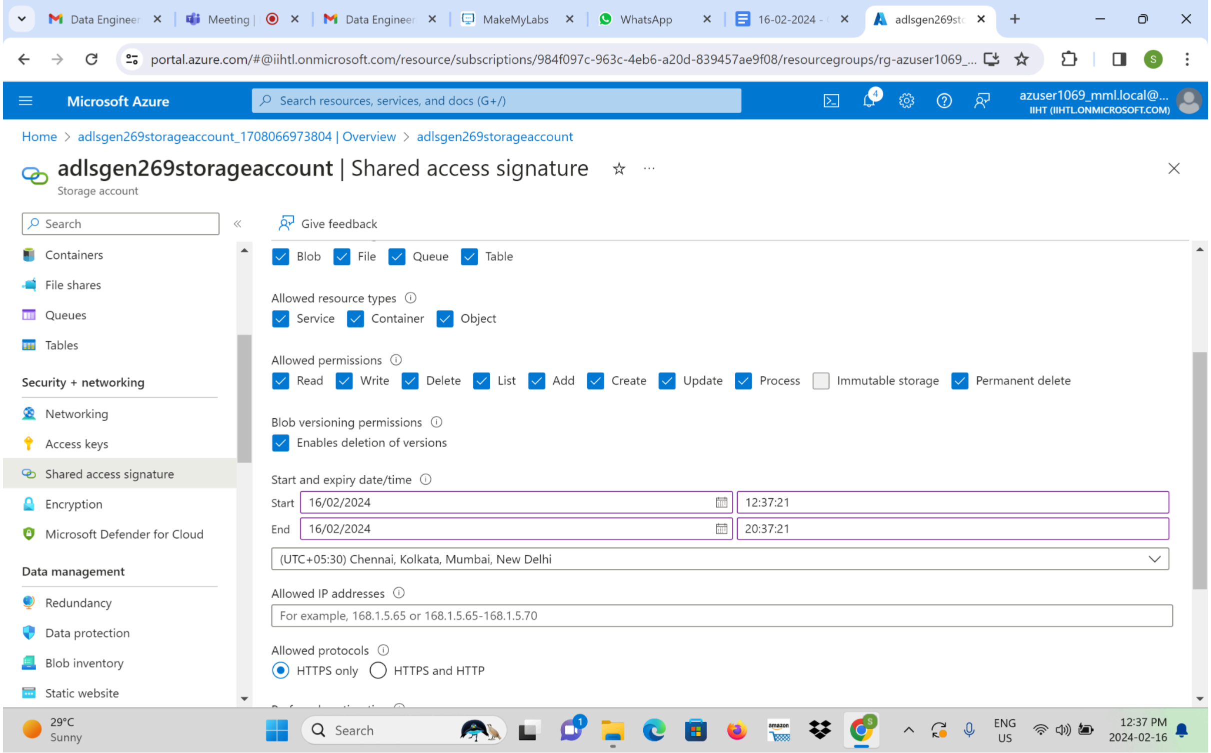
portal.azure.com/#@iihtl.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourcegroups/rg-azuser1069_...

Home > adlsgen269storageaccount_1708066973804 | Overview > adlsgen269storageaccount

adlsgen269storageaccount | Containers

Storage account

	Name	Last modified	Anonymous access level	Lease state
□	$logs	16/02/2024, 12:33:28	Private	Available
□	mycontainer69	16/02/2024, 12:36:14	Blob	Available



portal.azure.com/#@iihtl.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourcegroups/rg-azuser1069_...

Home > adlsgen269storageaccount_1708066973804 | Overview > adlsgen269storageaccount

adlsgen269storageaccount | Shared access signature

Storage account

☑ Blob ☑ File ☑ Queue ☑ Table

Allowed resource types

☑ Service ☑ Container ☑ Object

Allowed permissions

☑ Read ☑ Write ☑ Delete ☑ List ☑ Add ☑ Create ☑ Update ☑ Process □ Immutable storage ☑ Permanent delete

Blob versioning permissions

☑ Enables deletion of versions

Start and expiry date/time

Start 16/02/2024 12:37:21

End 16/02/2024 20:37:21

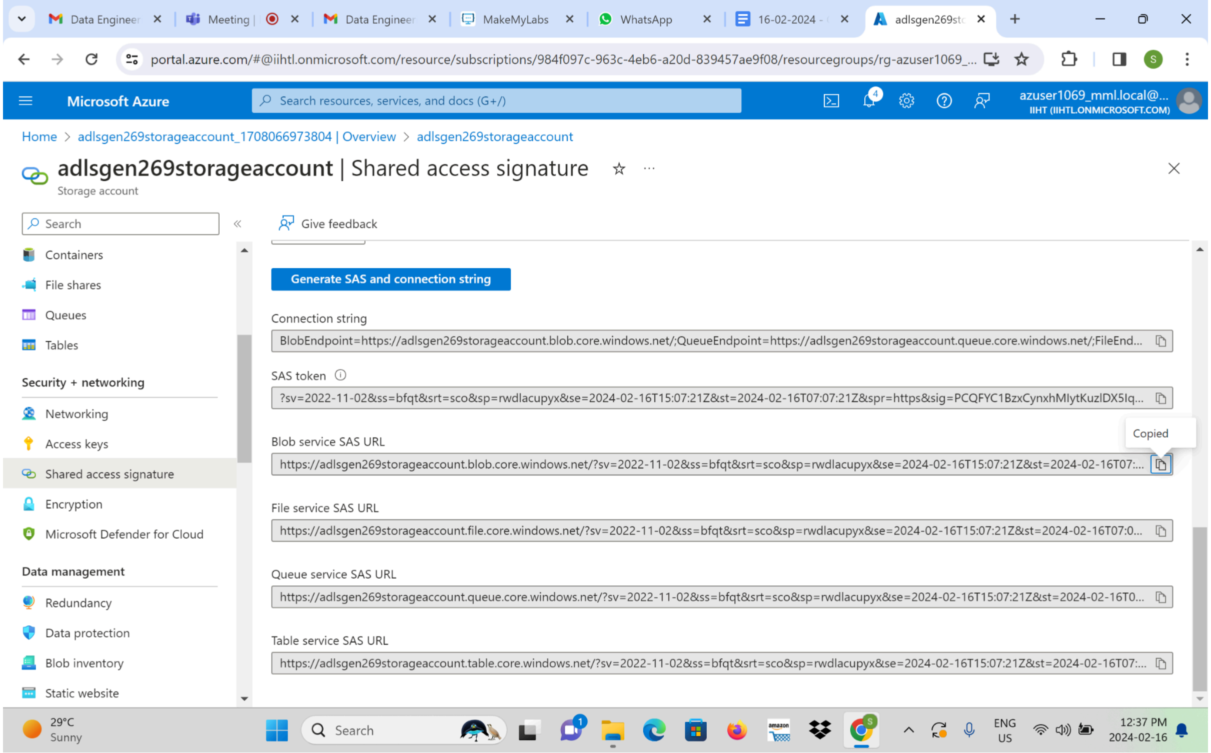
(UTC+05:30) Chennai, Kolkata, Mumbai, New Delhi

Allowed IP addresses

For example, 168.1.5.65 or 168.1.5.65-168.1.5.70

Allowed protocols

● HTTPS only ○ HTTPS and HTTP



portal.azure.com/#@iihtl.onmicrosoft.com/resource/subscriptions/984f097c-963c-4eb6-a20d-839457ae9f08/resourcegroups/rg-azuser1069_...

Home > adlsgen269storageaccount_1708066973804 | Overview > adlsgen269storageaccount

adlsgen269storageaccount | Shared access signature

Storage account

Generate SAS and connection string

Connection string

BlobEndpoint=https://adlsgen269storageaccount.blob.core.windows.net/;QueueEndpoint=https://adlsgen269storageaccount.queue.core.windows.net/;FileEnd...

SAS token

?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-02-16T15:07:21Z&st=2024-02-16T07:07:21Z&spr=https&sig=PCQFYC1BzxCynxhMlytKuzlDX5Iq...

Blob service SAS URL

https://adlsgen269storageaccount.blob.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-02-16T15:07:21Z&st=2024-02-16T07:...

Copied

File service SAS URL

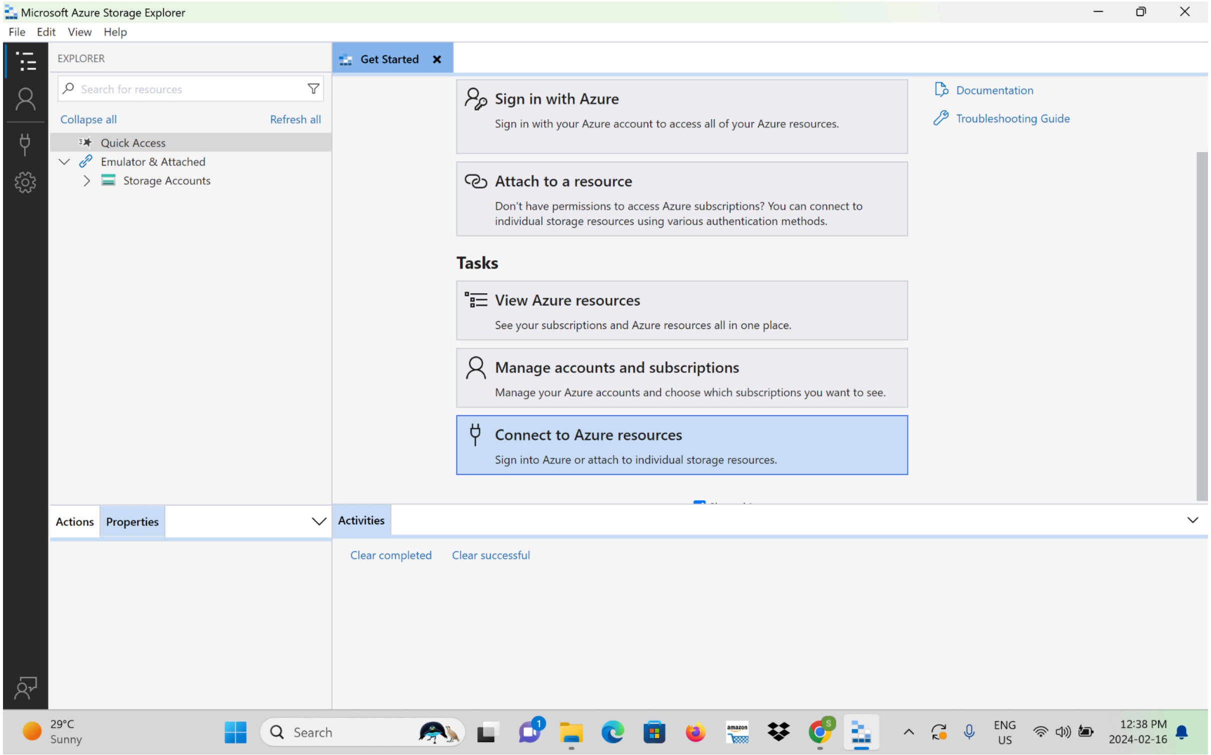
https://adlsgen269storageaccount.file.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-02-16T15:07:21Z&st=2024-02-16T07:0...

Queue service SAS URL

https://adlsgen269storageaccount.queue.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-02-16T15:07:21Z&st=2024-02-16T0...

Table service SAS URL

https://adlsgen269storageaccount.table.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-02-16T15:07:21Z&st=2024-02-16T07:...



Microsoft Azure Storage Explorer

Sign in with Azure

Sign in with your Azure account to access all of your Azure resources.

Attach to a resource

Don't have permissions to access Azure subscriptions? You can connect to individual storage resources using various authentication methods.

Tasks

View Azure resources

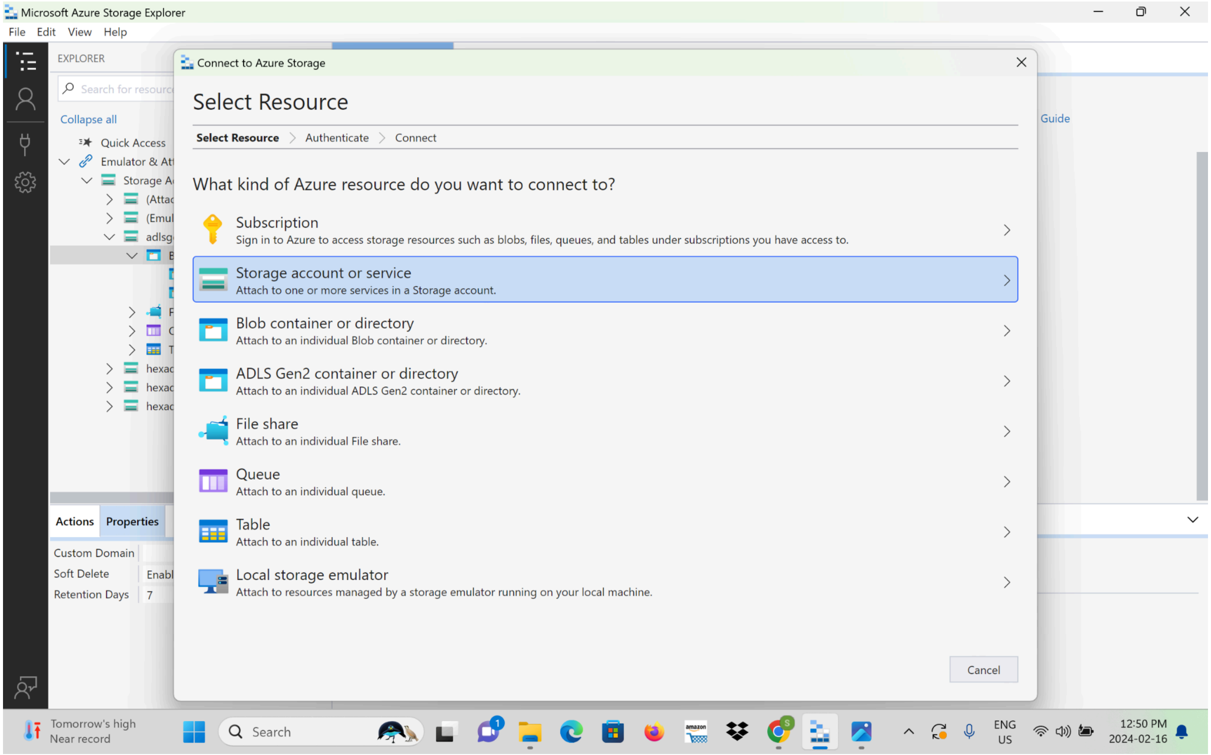
See your subscriptions and Azure resources all in one place.

Manage accounts and subscriptions

Manage your Azure accounts and choose which subscriptions you want to see.

Connect to Azure resources

Sign into Azure or attach to individual storage resources.



Connect to Azure Storage

Select Resource

Select Resource > Authenticate > Connect

What kind of Azure resource do you want to connect to?

Subscription
Sign in to Azure to access storage resources such as blobs, files, queues, and tables under subscriptions you have access to.

Storage account or service
Attach to one or more services in a Storage account.

Blob container or directory
Attach to an individual Blob container or directory.

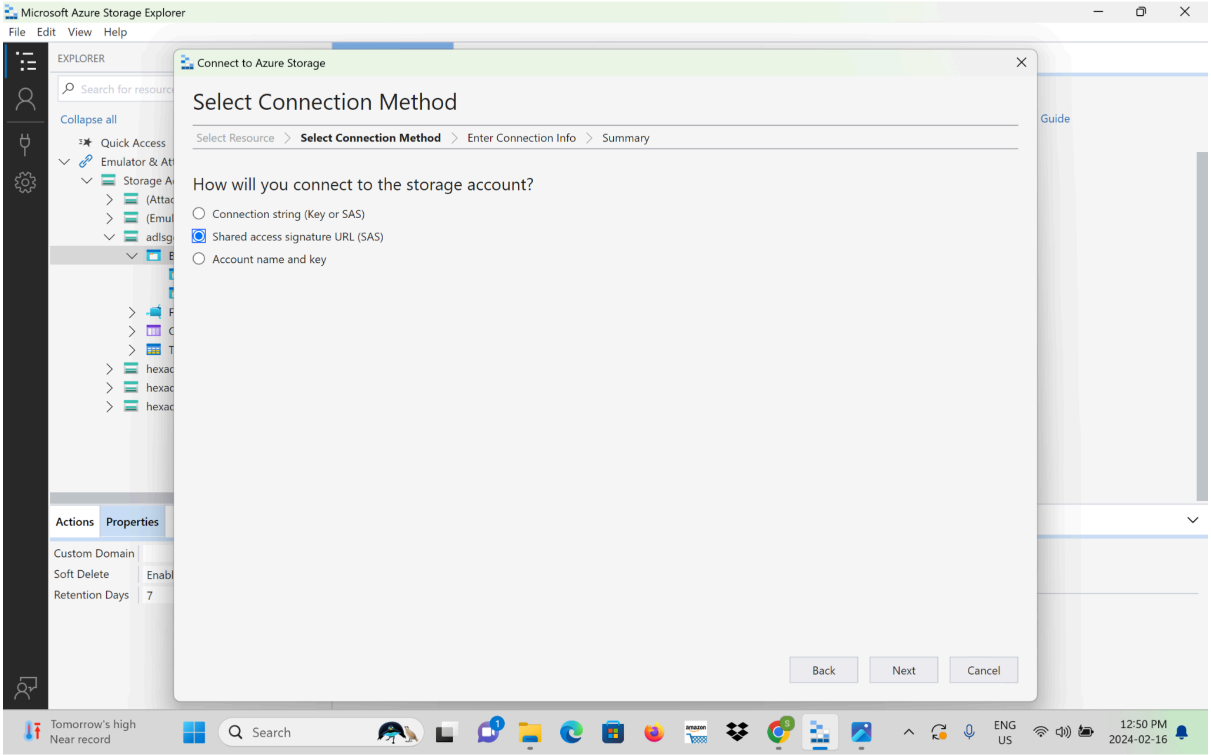
ADLS Gen2 container or directory
Attach to an individual ADLS Gen2 container or directory.

File share
Attach to an individual File share.

Queue
Attach to an individual queue.

Table
Attach to an individual table.

Local storage emulator
Attach to resources managed by a storage emulator running on your local machine.



Connect to Azure Storage

Select Connection Method

Select Resource > **Select Connection Method** > Enter Connection Info > Summary

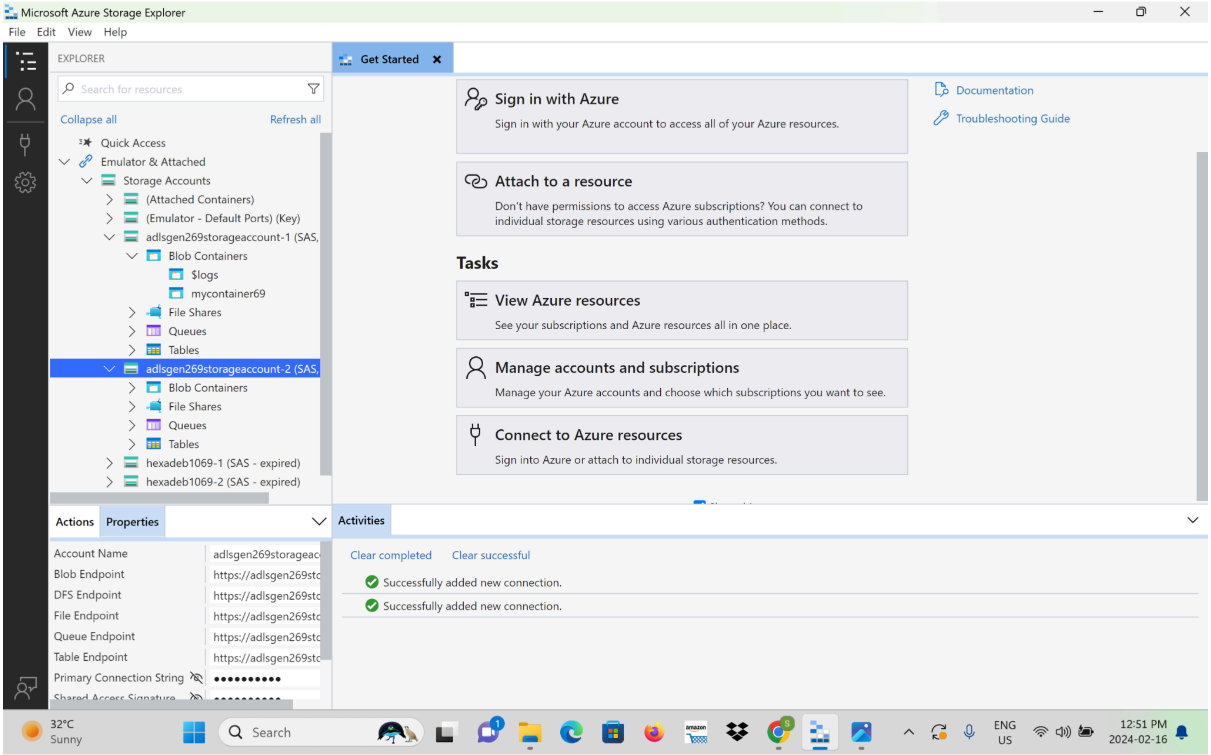
How will you connect to the storage account?

○ Connection string (Key or SAS)

● Shared access signature URL (SAS)

○ Account name and key

Custom Domain	
Soft Delete	Enabl
Retention Days	7



- Quick Access
- Emulator & Attached
  - Storage Accounts
    - (Attached Containers)
    - (Emulator - Default Ports) (Key)
    - adlsgen269storageaccount-1 (SAS,
      - Blob Containers
        - $logs
        - mycontainer69
      - File Shares
      - Queues
      - Tables
    - adlsgen269storageaccount-2 (SAS,
      - Blob Containers
      - File Shares
      - Queues
      - Tables
    - hexadeb1069-1 (SAS - expired)
    - hexadeb1069-2 (SAS - expired)

Sign in with Azure

Sign in with your Azure account to access all of your Azure resources.

Attach to a resource

Don't have permissions to access Azure subscriptions? You can connect to individual storage resources using various authentication methods.

Tasks

View Azure resources

See your subscriptions and Azure resources all in one place.

Manage accounts and subscriptions

Manage your Azure accounts and choose which subscriptions you want to see.

Connect to Azure resources

Sign into Azure or attach to individual storage resources.

Actions | Properties

Account Name	adlsgen269storageac
Blob Endpoint	https://adlsgen269stc
DFS Endpoint	https://adlsgen269stc
File Endpoint	https://adlsgen269stc
Queue Endpoint	https://adlsgen269stc
Table Endpoint	https://adlsgen269stc
Primary Connection String	••••••••••
Shared Access Signature	••••••••••

Activities

Clear completed   Clear successful

✓ Successfully added new connection.

✓ Successfully added new connection.

—> we can add the connection from another method too

—>For that we need to detach the storage account from the storage explorer .

—>After detaching we need to go the access key and there we need to copy the storage account and connection string

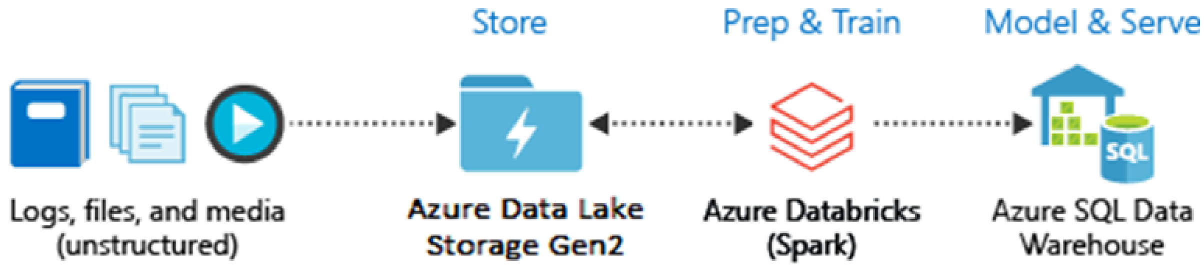
From that we need to paste it on the storage explorer.
—>For that, first we need to connect to an azure resource and then we need to select the ADLS gen2 container or directory resource.
—>From there we need to select the connection string for the connection method .
—>Thereby we need to paste the link and the connection will be established successfully thereby we can see the hierarchical structure of our data.

3)UNDERSTAND THE STAGES FOR PROCESSING BIG DATA:

Data lakes have a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:

—>An enterprise data warehouse.
—>Advanced analytics against big data.
—>A real-time analytical solution.

There are four stages for processing big data solutions that are common to all architectures:



i)**INGEST:**The Ingest phase in big data processing involves gathering data from various sources like files or logs and storing it securely in a data lake. The choice of tools depends on how frequently and quickly data needs to be collected. For periodic batch movements, tools like Azure Synapse Analytics or Azure Data Factory are suitable, acting as organized pipelines. In contrast, for real-time data ingestion, options like Apache Kafka for HDInsight or Stream Analytics ensure immediate processing. In simple terms, the Ingest phase focuses on efficiently collecting and storing data, adapting tools based on the speed and frequency of data arrival.

ii)STORE: In the Store phase of big data processing, we decide where to keep the collected data securely. Azure Data Lake Storage Gen2 is a safe and scalable storage option that works well with popular big data processing tools. Essentially, this step is about choosing a reliable place to store our gathered data, and Azure Data Lake Storage Gen2 offers a secure and scalable solution that plays nicely with various big data processing technologies.

iii)PREP AND TRAIN: In the Prep and Train phase, we use specific technologies to get our data ready and train models for machine learning solutions. This involves preparing the data and teaching models to make predictions. Common tools for this phase include Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Machine Learning. So, it's like getting our data in shape and teaching our systems to make smart predictions using technologies that make the process easier.

iv)MODEL AND SERVE: In the Model and Serve phase, we use technologies to show the data to users. This could involve using tools that help visualize the data, like Microsoft Power BI. Additionally, we might use analytical data stores such as Azure Synapse Analytics to store and manage the data. Usually, we use a mix of different technologies based on what the business needs. So, it's about using tools to present the data in a way that's easy for people to understand and use.