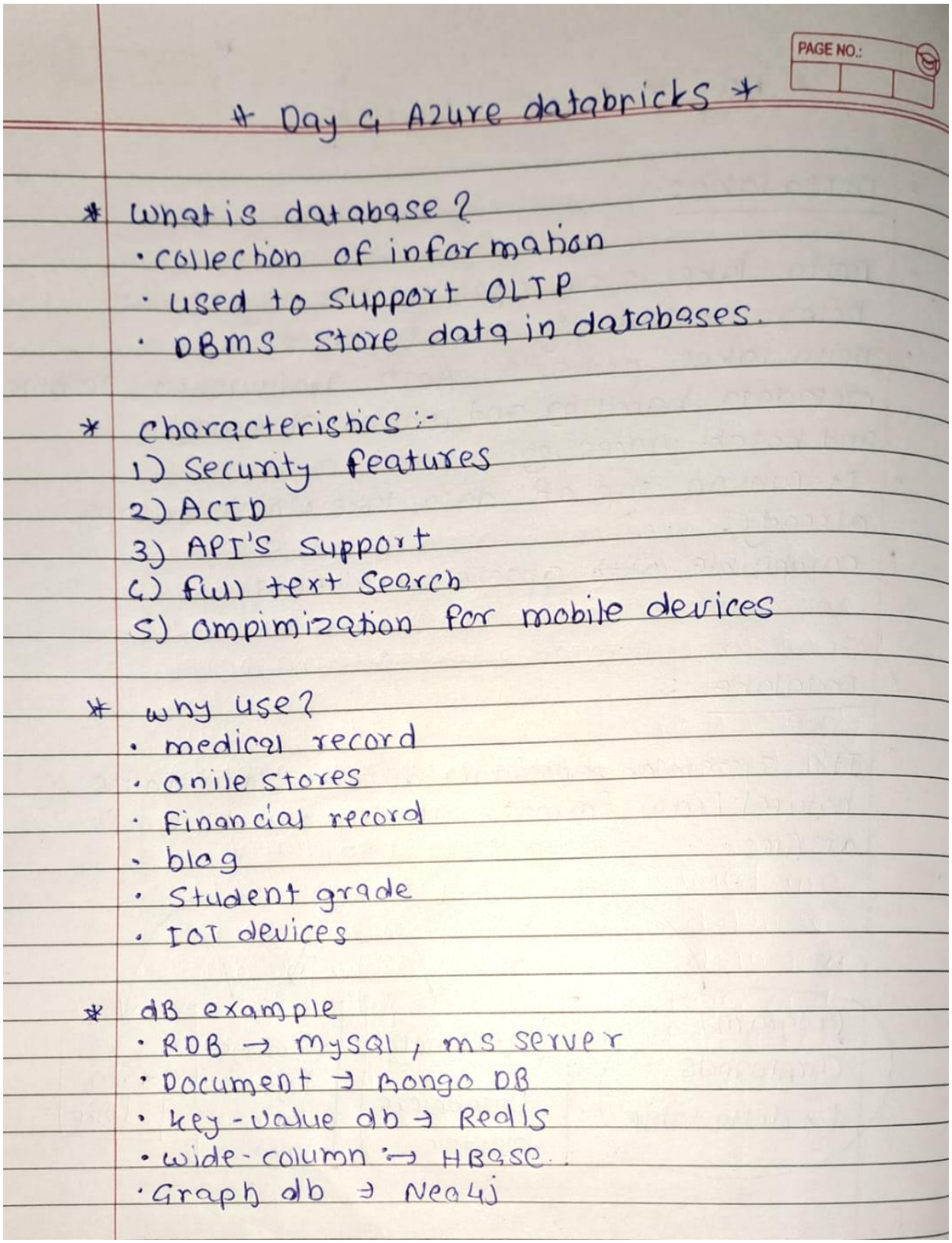
Azure Databricks Assignment 4

Pratik Wani

- Notes



* Day 4 Azure databricks *

* What is database ?
- collection of information
- used to support OLTP
- DBMS store data in databases.

* characteristics :-
1) Security features
2) ACID
3) API's support
4) full text search
5) Ompimization for mobile devices

* why use?
- medical record
- onile stores
- Financial record
- blog
- Student grade
- IOT devices

* dB example
- ROB → mySQl , ms server
- Document → Mongo DB
- key-value db → Redis
- wide-column → HBase.
- Graph db → Neo4j

PAGE NO.:

* OLAP & data warehouse and data lake

- Both warehouse and Lake are meant to support OLAP
- collect data from various sources
- use for analytical purpose

* Data warehouse?
- Store highly structured information from various sources
- stores current and historical data
- Giant db that optimized for analytics

* characteristics :-
- stores large amount of historical data
- ETL process moves data from original source to data warehouse
- predefined and fixed relational schema
- work well in structured data.

* data warehouse uses :-
- when we have to store large amount of data.
- use for indepth analysis
- use for bussiness analysist and data scientist
- power daily operations

PAGE NO.:

* Data warehouse example:-
  - Amazon Redshift
  - IBM db2 warehouse
  - Microsoft Azure Snyapse
  - Snowflake
  - Teradata Vantage

* Data lake:-
  - Repository of data from disparate sources that is stored in its original, raw format
  - Like data warehouse, data lakes store large amounts of current and historical data.
  - has ability to store data in a variety of formats like JSON, BSON, CSV, Avro

* is datalake is database?
  - it's repository for data stored in a variety of ways including database
  - with mordern technology, a data lake can also form the storage layer of database.
  - Tools like Dremio, Atlas data lake can give a db-like view in data stored in your data lake

PAGE NO.:

* characteristics :-
- stores large amount of structured, semi-structured and unstructred data.
- can contain everything from relational data to json documents to PDF to audio files.
- Data need not to be stored in order to be added to the data lake, which means data can added incredibly without planning.
- can be processed with a variety of OLAP systems and visualized with BI tools.

* Data lake example :-
- AWS S3
- Azure data lake
- Google cloud storage
- AWS Athena.
- DB SQL Anlytics.

* Database vs Datawarehouse vs Data lake

	DB	DW	DL
Workload	operational & Transactional	Analytical	Analytical
Data type	Structured, or semi structured	Structured, semi-structured or unstructured	structured or semistructured ~~or~~ unstructured.

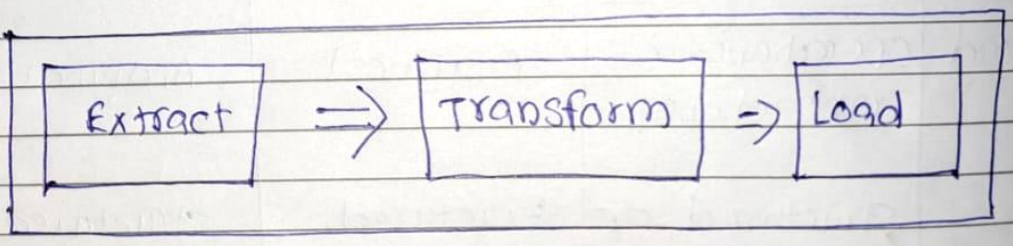
	DB	DW	DL
User	App^n dev	Bussiness Analysis, App^n dev, data scientist	Data Scientist, Bussiness Anq.
pros	fast queries	Bussiness Analysis	Easy data storage
cons	limited data analytical capabilities	Difficult to design.	Require efforts to organize data

* ETL process

seq is also imp

data pipeline

DATA

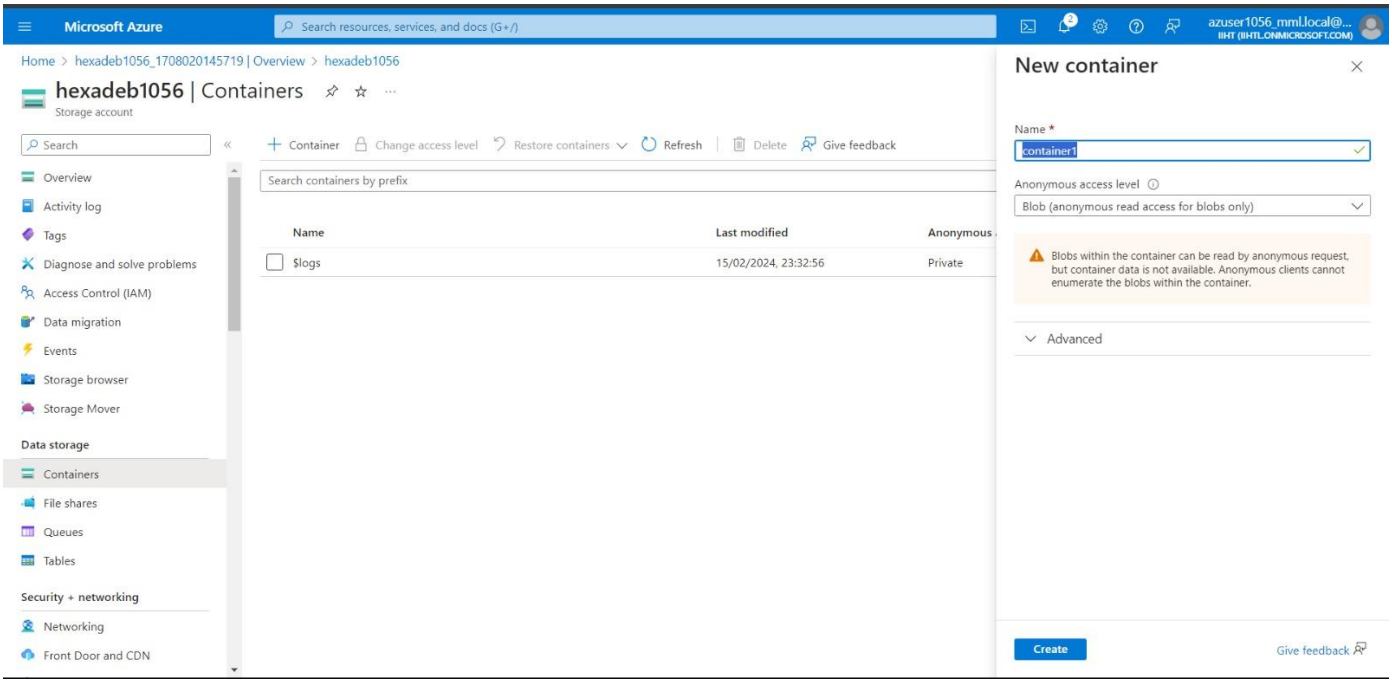


Raw data from the Source

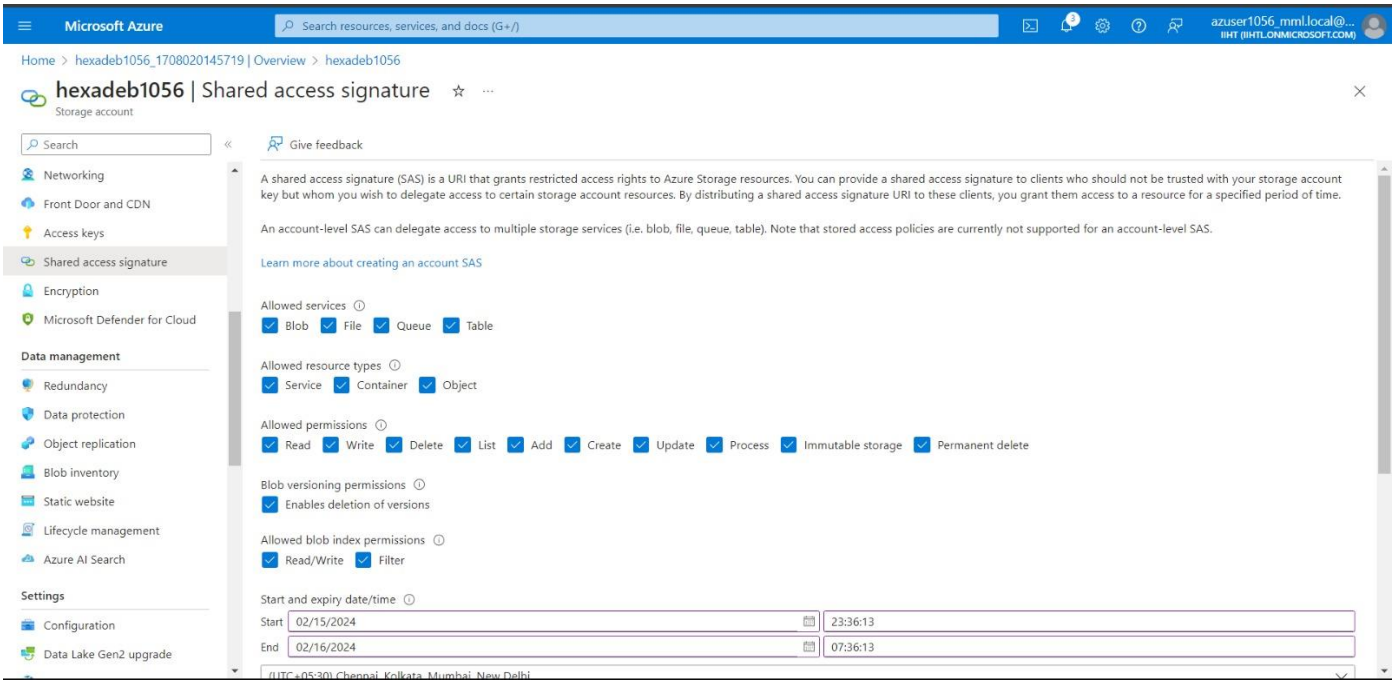
tools:

Data loading in DW, DL, DB

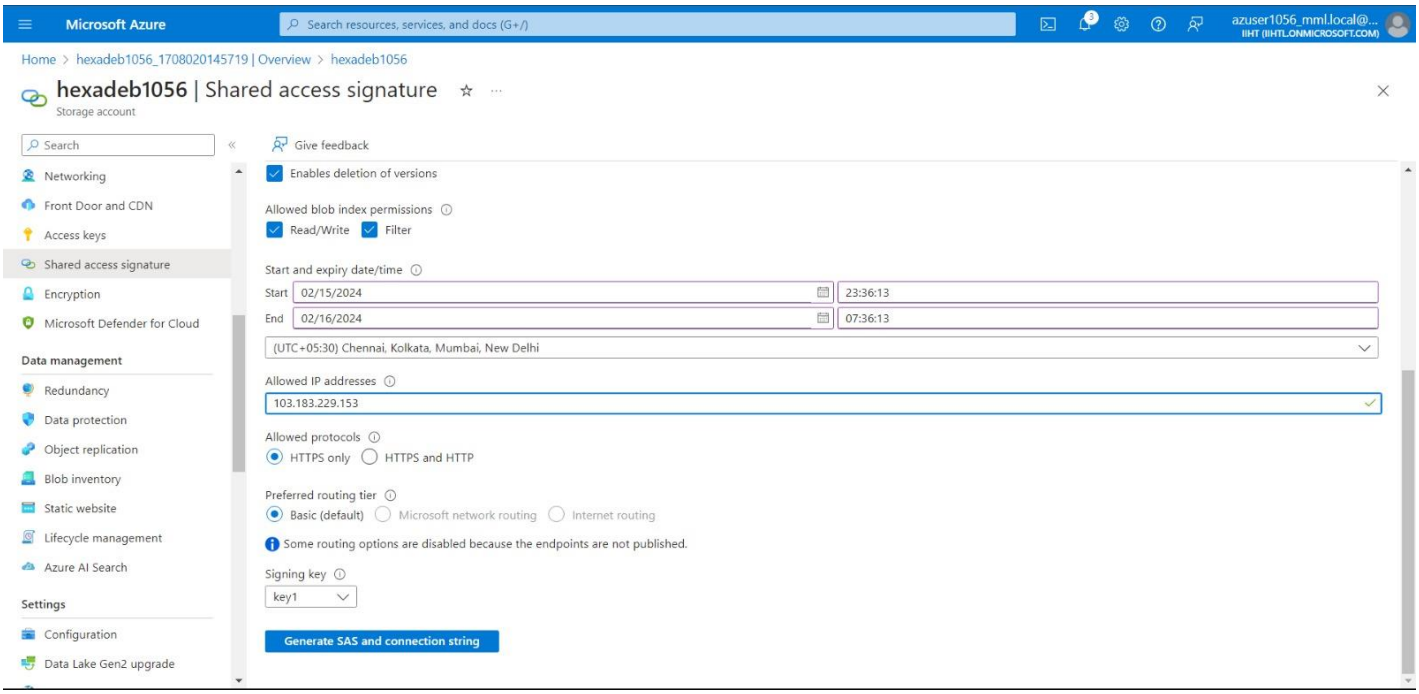
- Storage Account and Container



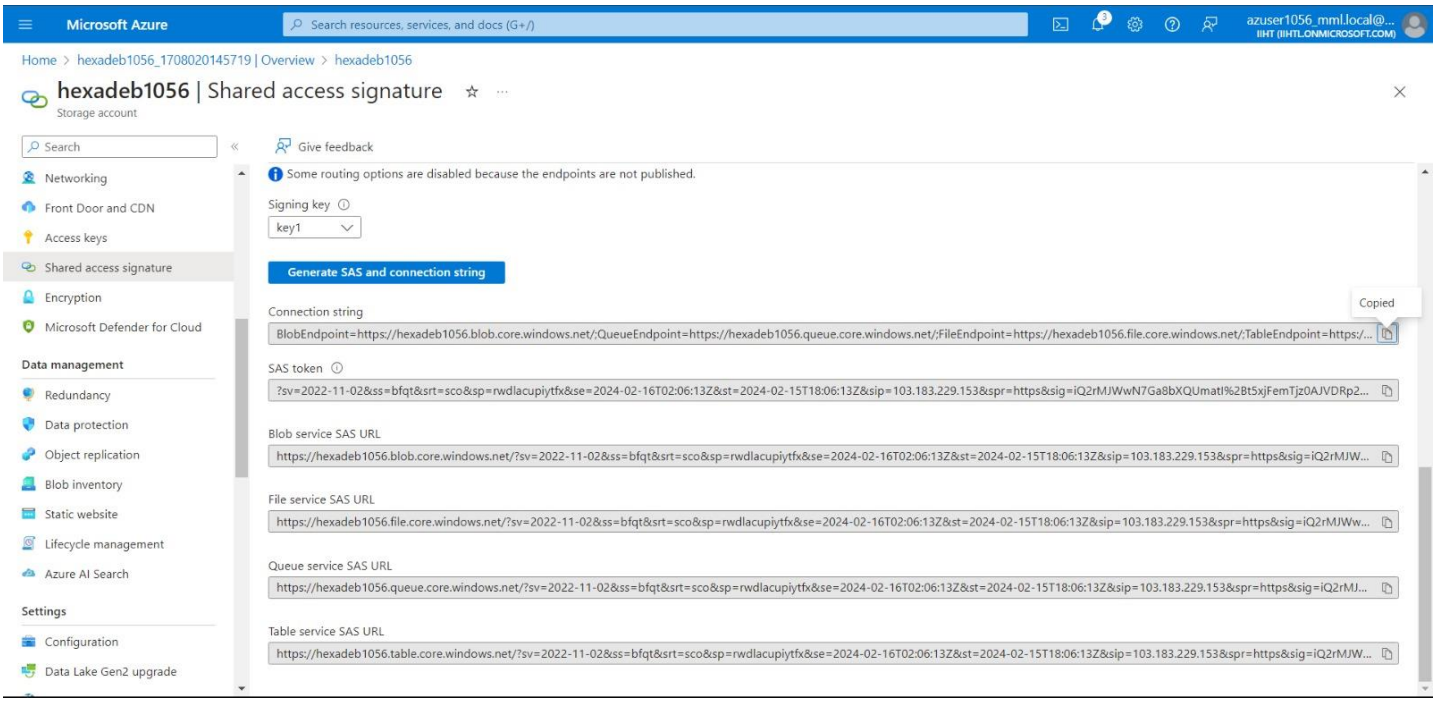
- Shared Access Signature



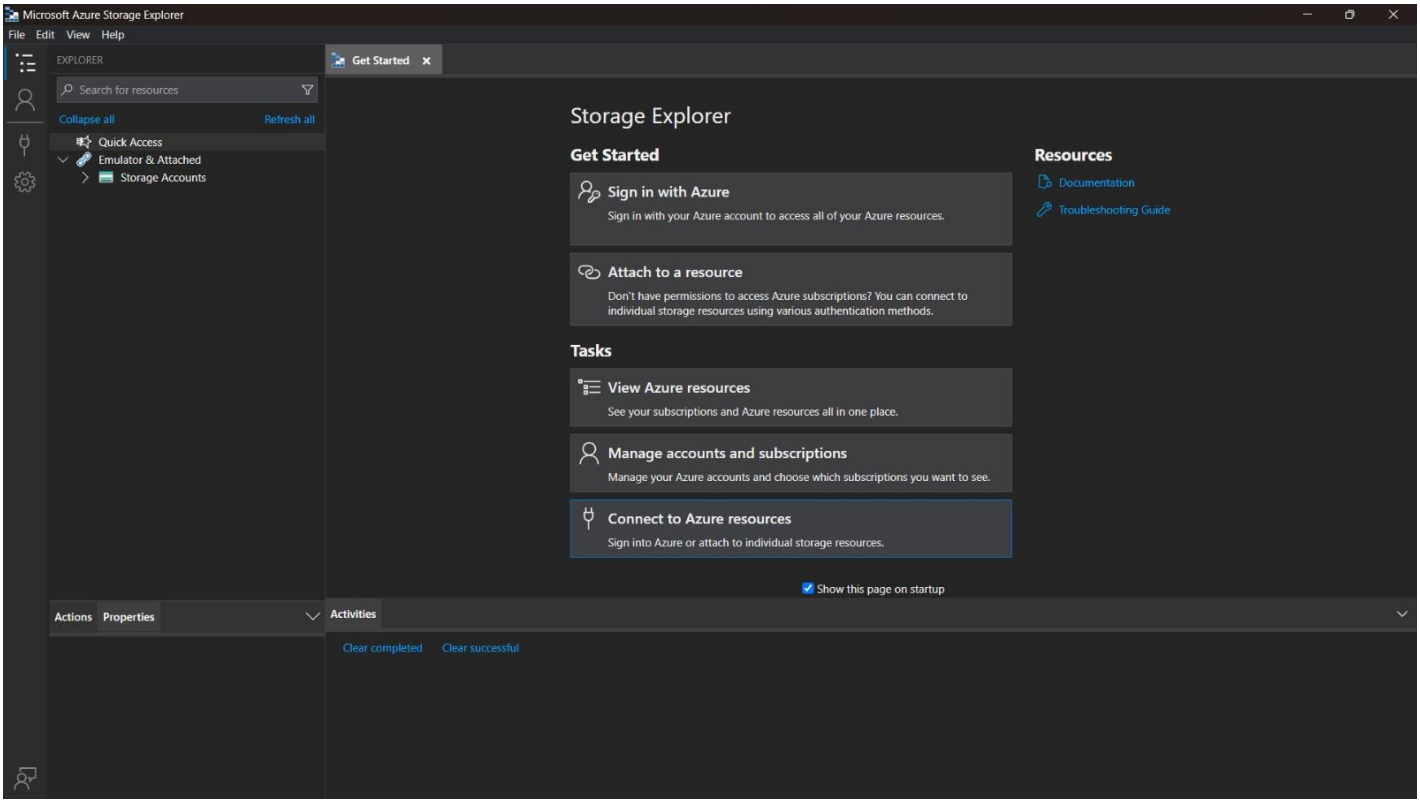
- IP Address to Generate SAS and Connection String



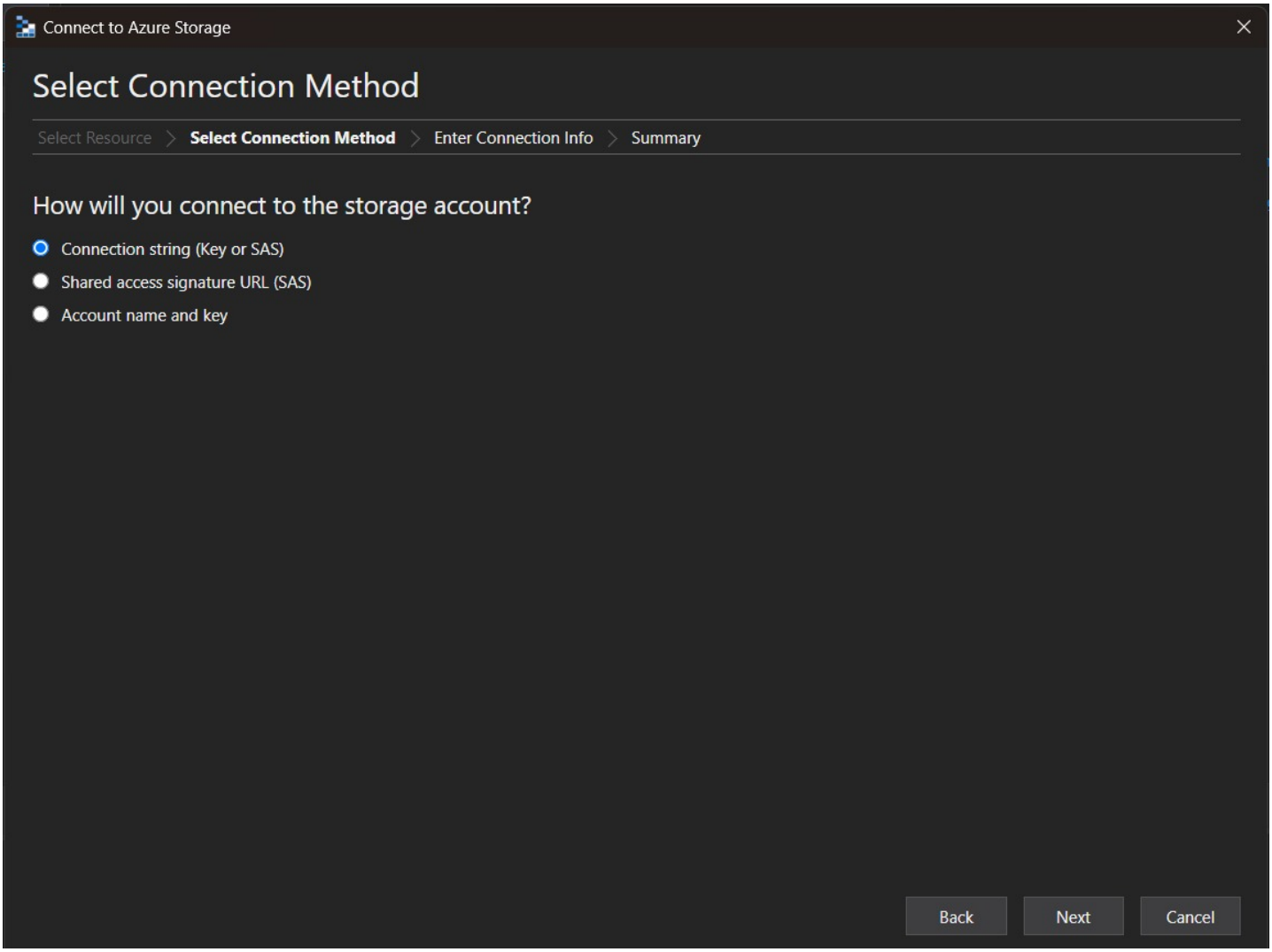
- SAS and Connection string



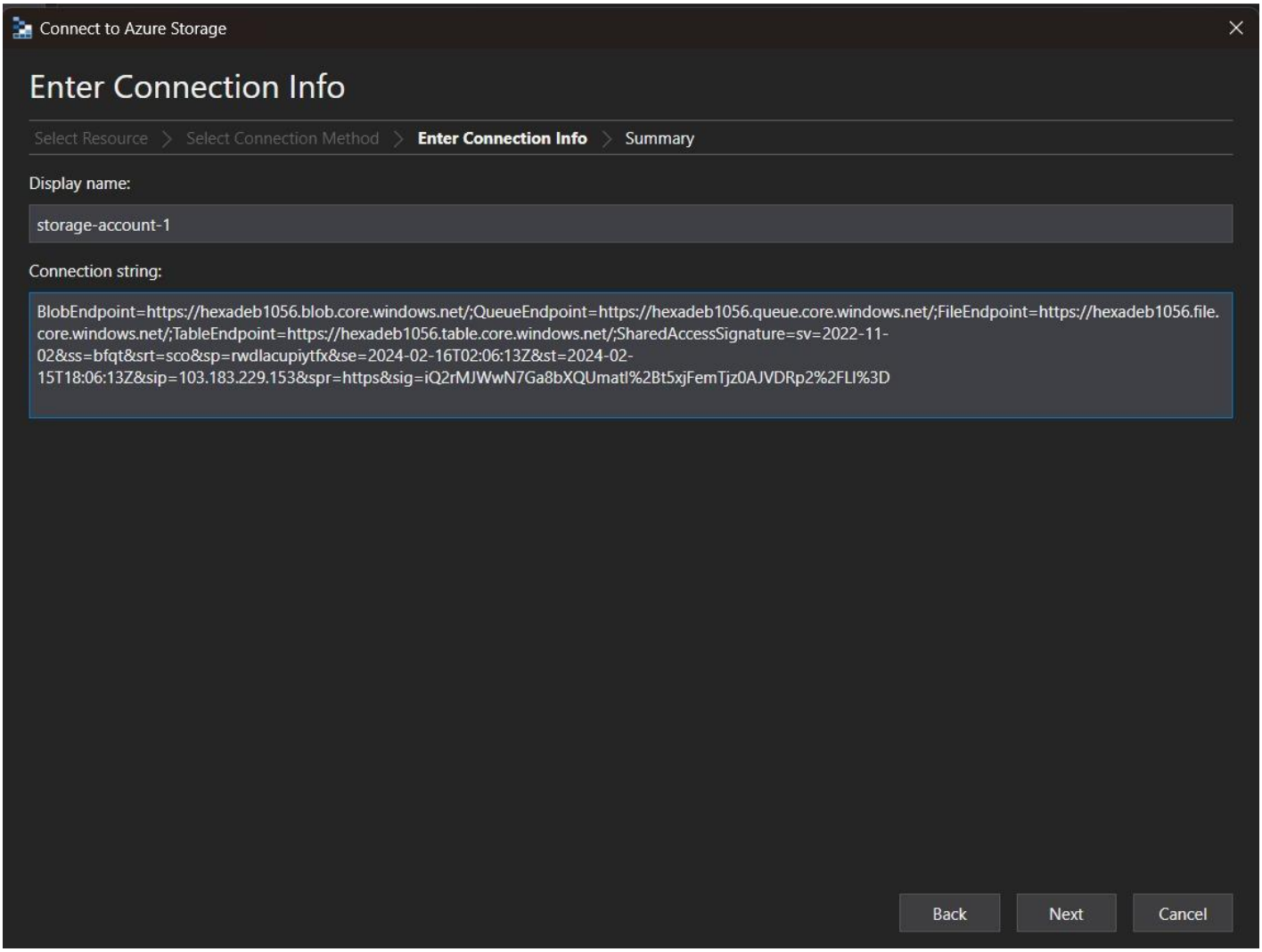
- Storage Explorer



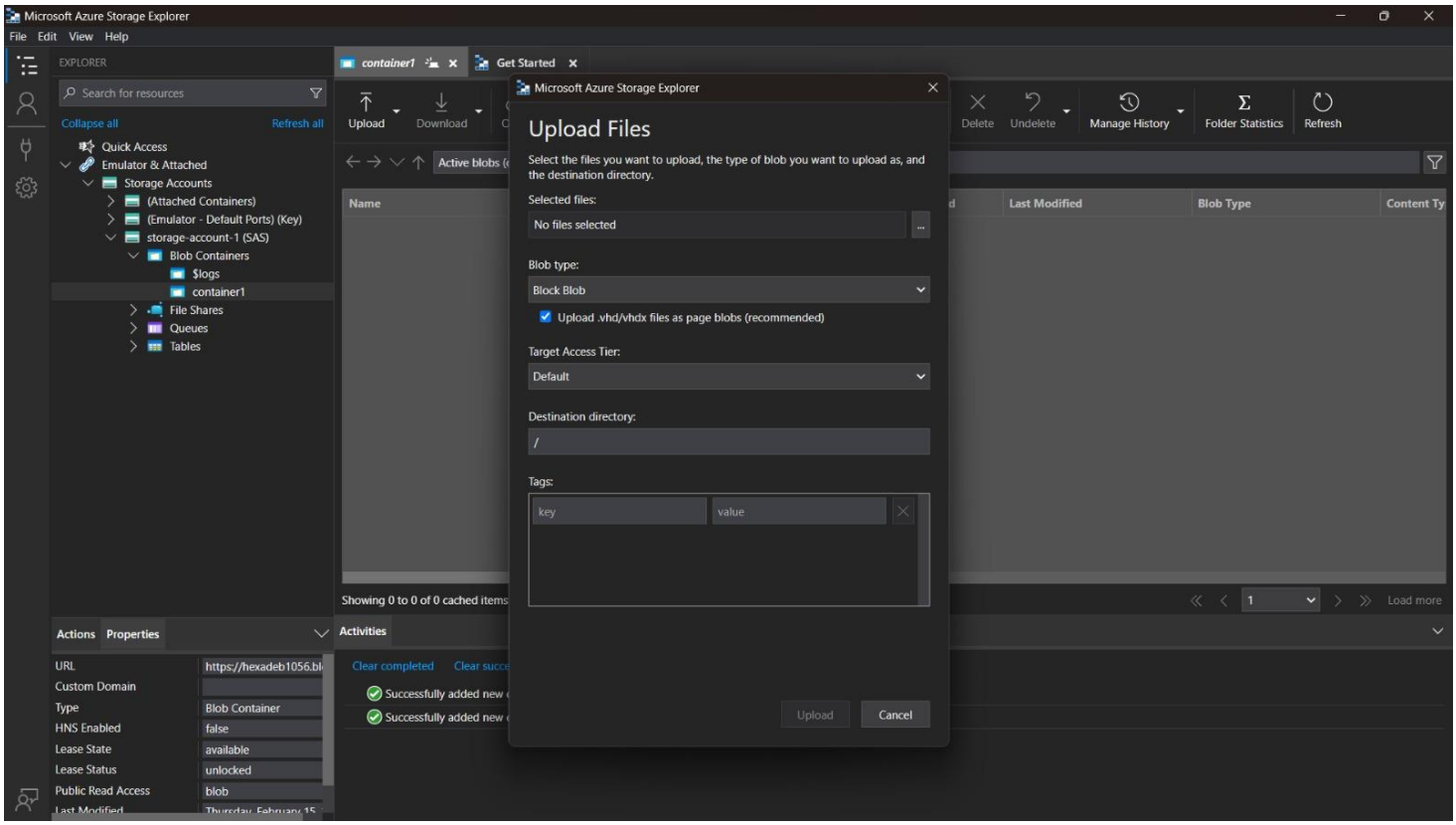
- Connect to Azure



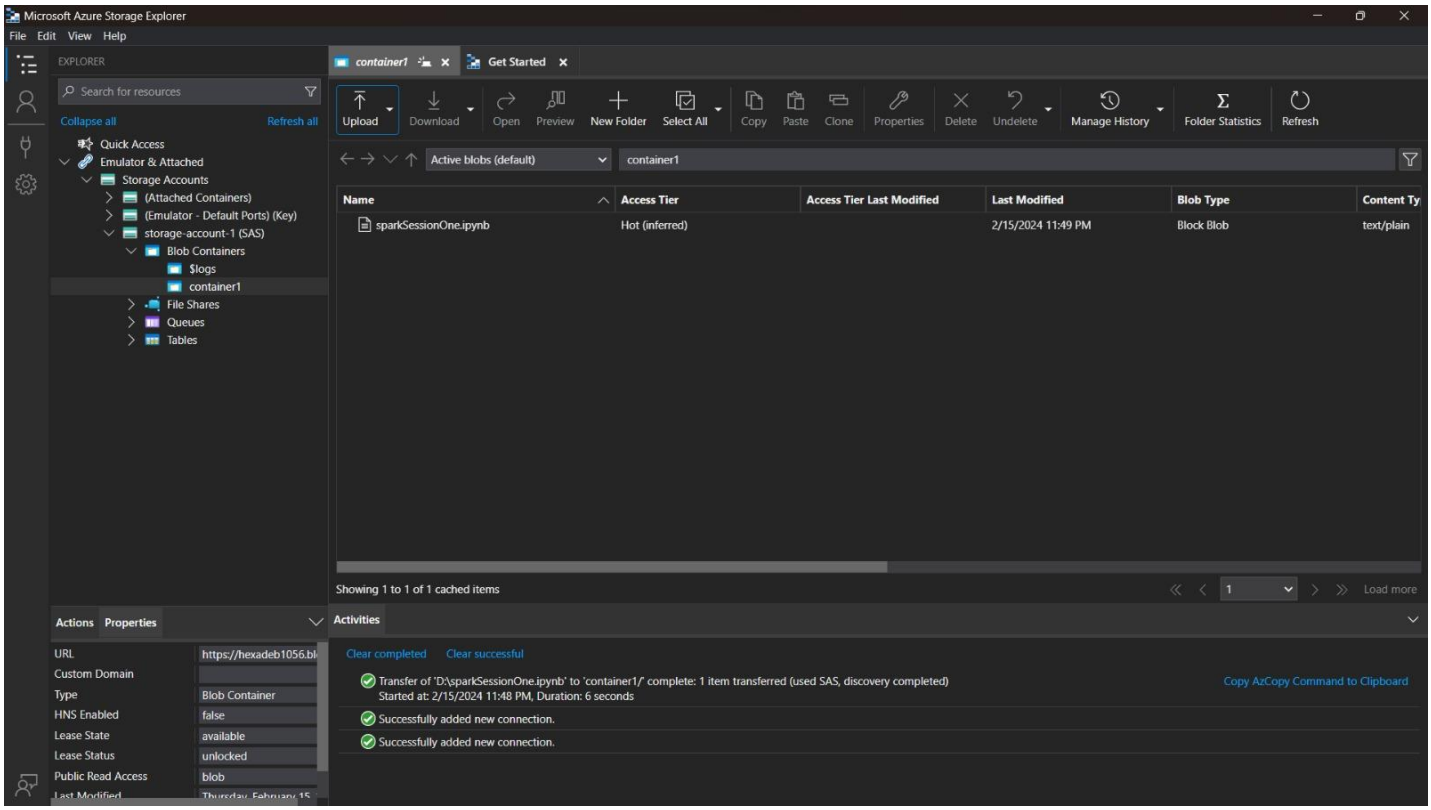
- Connection Information



- Blob Containers



- File Upload


Microsoft Azure Storage Explorer
File Edit View Help
EXPLORER
Search for resources
Collapse all
Refresh all
Quick Access
Emulator & Attached
Storage Accounts
(Attached Containers)
(Emulator - Default Ports) (Key)
storage-account-1 (SAS)
Blob Containers
$logs
container1
File Shares
Queues
Tables
container1
Get Started
Upload
Download
Open
Preview
New Folder
Select All
Copy
Paste
Clone
Properties
Delete
Undelete
Manage History
Folder Statistics
Refresh
Active blobs (default)
container1
Name
Access Tier
Access Tier Last Modified
Last Modified
Blob Type
Content Ty
sparkSessionOne.ipynb
Hot (inferred)
2/15/2024 11:49 PM
Block Blob
text/plain
Showing 1 to 1 of 1 cached items
Load more
Actions
Properties
URL
Custom Domain
Type
HNS Enabled
Lease State
Lease Status
Public Read Access
Blob Container
false
available
unlocked
blob
Activities
Clear completed
Clear successful
Transfer of 'D:\sparkSessionOne.ipynb' to 'container1/' complete: 1 item transferred (used SAS, discovery completed)
Started at: 2/15/2024 11:48 PM, Duration: 6 seconds
Copy AzCopy Command to Clipboard
Successfully added new connection.
Successfully added new connection.

- Also Visible in Azure