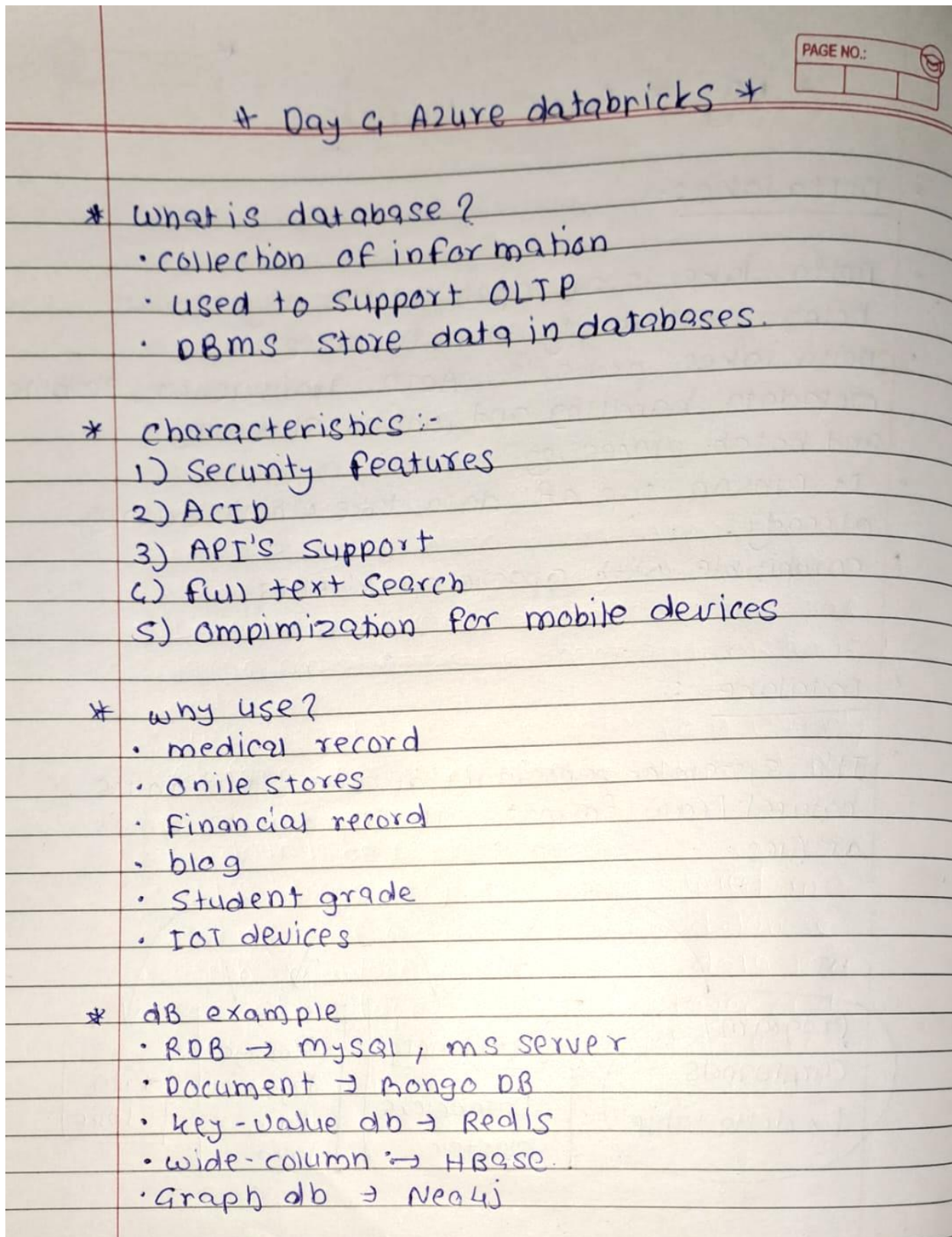


Azure Databricks Assignment 4

Pratik Wani

- Notes



* 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 analyst and data scientist
- power daily operations

* Data warehouse example:-

- Amazon Redshift
- IBM db2 warehouse
- Microsoft Azure Synapse
- 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 data lake is database?

- it's repository for data stored in a variety of ways including database
- with modern 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

* characteristics :-

- stores large amount of structured, semi-structured and unstructured 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 incrementally 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 Analytics.

* Database vs Datawarehouse vs Data lake

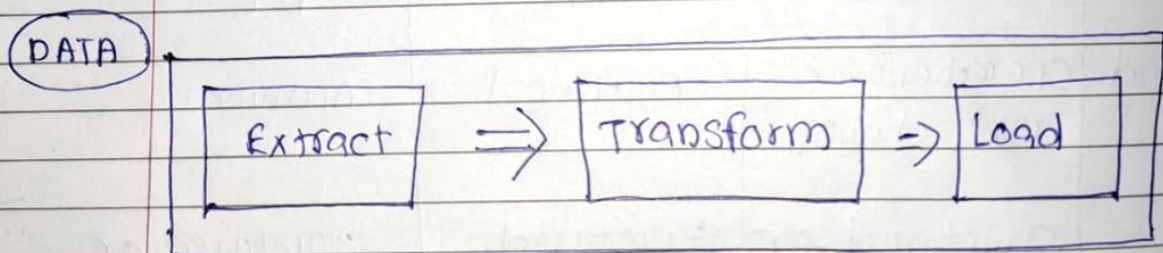
	DB	DW	DL
Workload	operational & Transactional	Analytical	Analytical
Data type	Structured, or semi structured	Structured, semi-structured or unstructured	structured or semi structured un unstructured.

	DB	DW	DL
User	App ⁿ dev	Business Analysts, App ⁿ dev, data scientist	Data Scientist, Business Analyst
pros	fast queries	Business Analysis	fast data storage
cons	limited data analytical capabilities	Difficult to design	Require efforts to organize data

* ETL process

Seq is also imp

data pipeline



Raw data
from the
Source

tools:

Data loading
in DW, DL,
DB

- Storage Account and Container

Microsoft Azure

Search resources, services, and docs (G+)

Home > hexadeb1056_1708020145719 | Overview > hexadeb1056

hexadeb1056 | Containers

Storage account

Search

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Storage Mover

Data storage

Containers

File shares

Queues

Tables

Security + networking

Networking

Front Door and CDN

Container

Change access level

Restore containers

Refresh

Delete

Give feedback

Search containers by prefix

Name	Last modified	Anonymous
<input type="checkbox"/> \$logs	15/02/2024, 23:32:56	Private

New container

Name *

container1

Anonymous access level ⓘ

Blob (anonymous read access for blobs only)

Blobs within the container can be read by anonymous request, but container data is not available. Anonymous clients cannot enumerate the blobs within the container.

Advanced

Create

Give feedback

- Shared Access Signature

Microsoft Azure

Search resources, services, and docs (G+)

Home > hexadeb1056_1708020145719 | Overview > hexadeb1056

hexadeb1056 | Shared access signature

Storage account

Give feedback

A shared access signature (SAS) is a URI that grants restricted access rights to Azure Storage resources. You can provide a shared access signature to clients who should not be trusted with your storage account key but whom you wish to delegate access to certain storage account resources. By distributing a shared access signature URI to these clients, you grant them access to a resource for a specified period of time.

An account-level SAS can delegate access to multiple storage services (i.e. blob, file, queue, table). Note that stored access policies are currently not supported for an account-level SAS.

Learn more about creating an account SAS

Allowed services ⓘ

☒ 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

Allowed blob index permissions ⓘ

☒ Read/Write ☒ Filter

Start and expiry date/time ⓘ

Start 02/15/2024 23:36:13

End 02/16/2024 07:36:13

UTC+05:30L Chennai, Kolkata, Mumbai, New Delhi

- IP Address to Generate SAS and Connection String

Microsoft Azure

Search resources, services, and docs (G+/I)

Home > hexadeb1056_1708020145719 | Overview > hexadeb1056

hexadeb1056 | Shared access signature ☆ ...

Storage account

Search

Give feedback

Enables deletion of versions

Allowed blob index permissions ⓘ

Read/Write Filter

Start and expiry date/time ⓘ

Start 02/15/2024 23:36:13

End 02/16/2024 07:36:13

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

Allowed IP addresses ⓘ

103.183.229.153

Allowed protocols ⓘ

HTTPS only HTTPS and HTTP

Preferred routing tier ⓘ

Basic (default) Microsoft network routing Internet routing

Some routing options are disabled because the endpoints are not published.

Signing key ⓘ

key1

Generate SAS and connection string

- SAS and Connection string

Microsoft Azure

Search resources, services, and docs (G+/I)

Home > hexadeb1056_1708020145719 | Overview > hexadeb1056

hexadeb1056 | Shared access signature ☆ ...

Storage account

Search

Give feedback

Some routing options are disabled because the endpoints are not published.

Signing key ⓘ

key1

Generate SAS and connection string

Connection string

BlobEndpoint=https://hexadeb1056.blob.core.windows.net/QueueEndpoint=https://hexadeb1056.queue.core.windows.net/FileEndpoint=https://hexadeb1056.file.core.windows.net/TableEndpoint=https://hexadeb1056.table.core.windows.net

SAS token ⓘ

?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytf&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWwN7Ga8bXQUmatl%2Bt5xJfemTjz0AJVDRp2...

Blob service SAS URL

https://hexadeb1056.blob.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytf&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWw...

File service SAS URL

https://hexadeb1056.file.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytf&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWw...

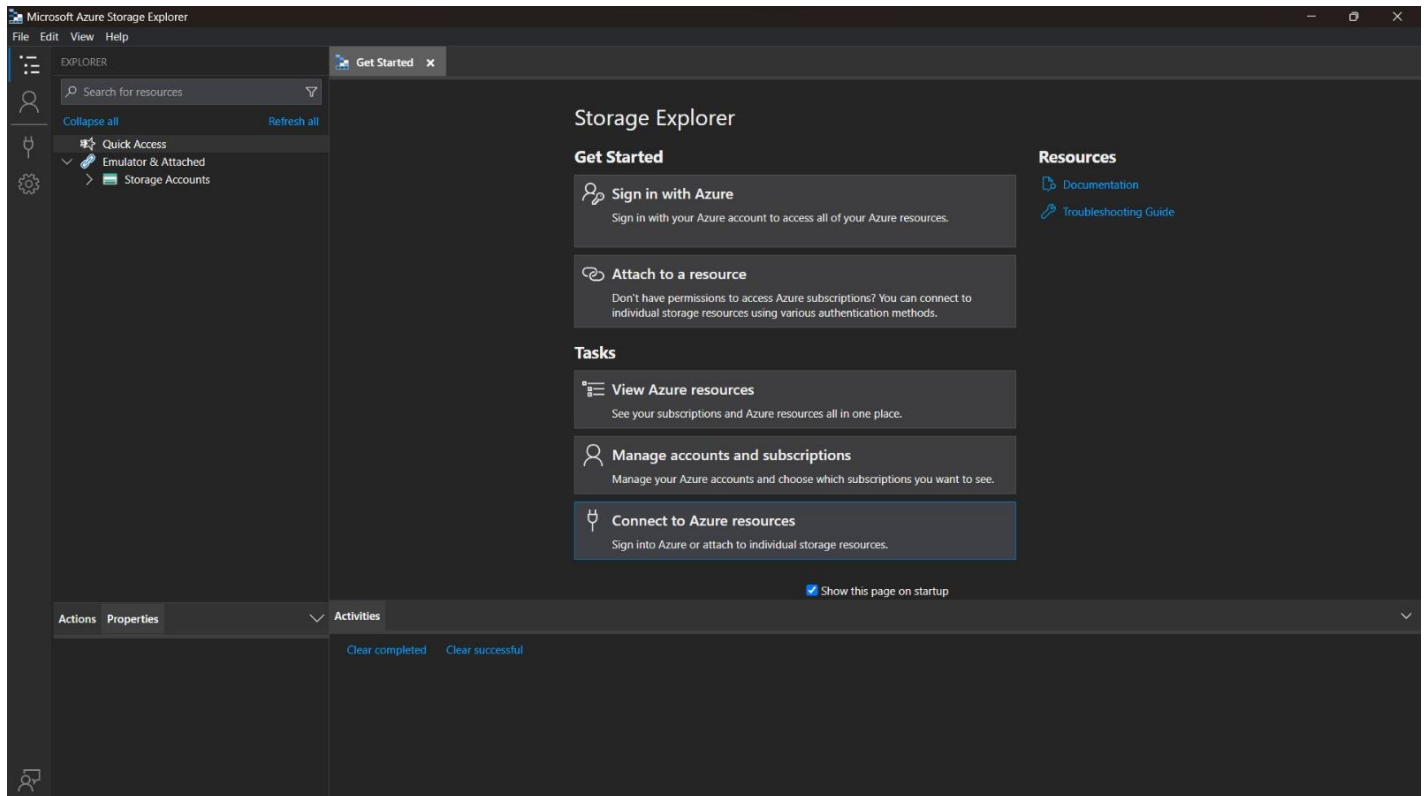
Queue service SAS URL

https://hexadeb1056.queue.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytf&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWw...

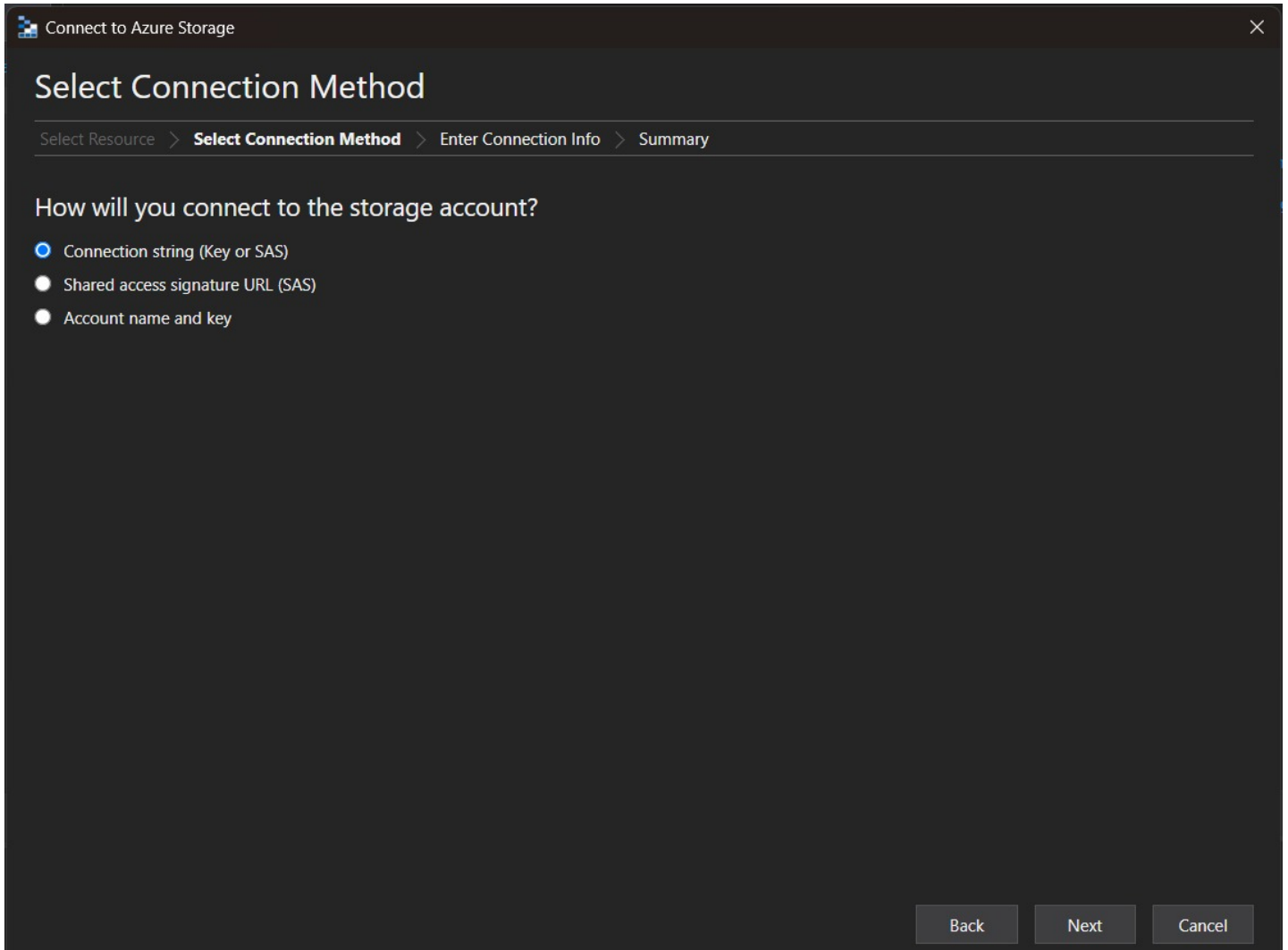
Table service SAS URL

https://hexadeb1056.table.core.windows.net/?sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytf&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWw...

- Storage Explorer



- Connect to Azure



The screenshot shows a dark-themed window titled "Connect to Azure Storage" with a close button in the top right corner. The main heading is "Select Connection Method". Below it is a progress bar with four steps: "Select Resource", "Select Connection Method" (which is highlighted), "Enter Connection Info", and "Summary". The question "How will you connect to the storage account?" is followed by three radio button options: "Connection string (Key or SAS)" (selected), "Shared access signature URL (SAS)", and "Account name and key". At the bottom right, there are three buttons: "Back", "Next", and "Cancel".

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

Back Next Cancel

- Connection Information

 Connect to Azure Storage ✕

Enter Connection Info

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

Display name:

storage-account-1

Connection string:

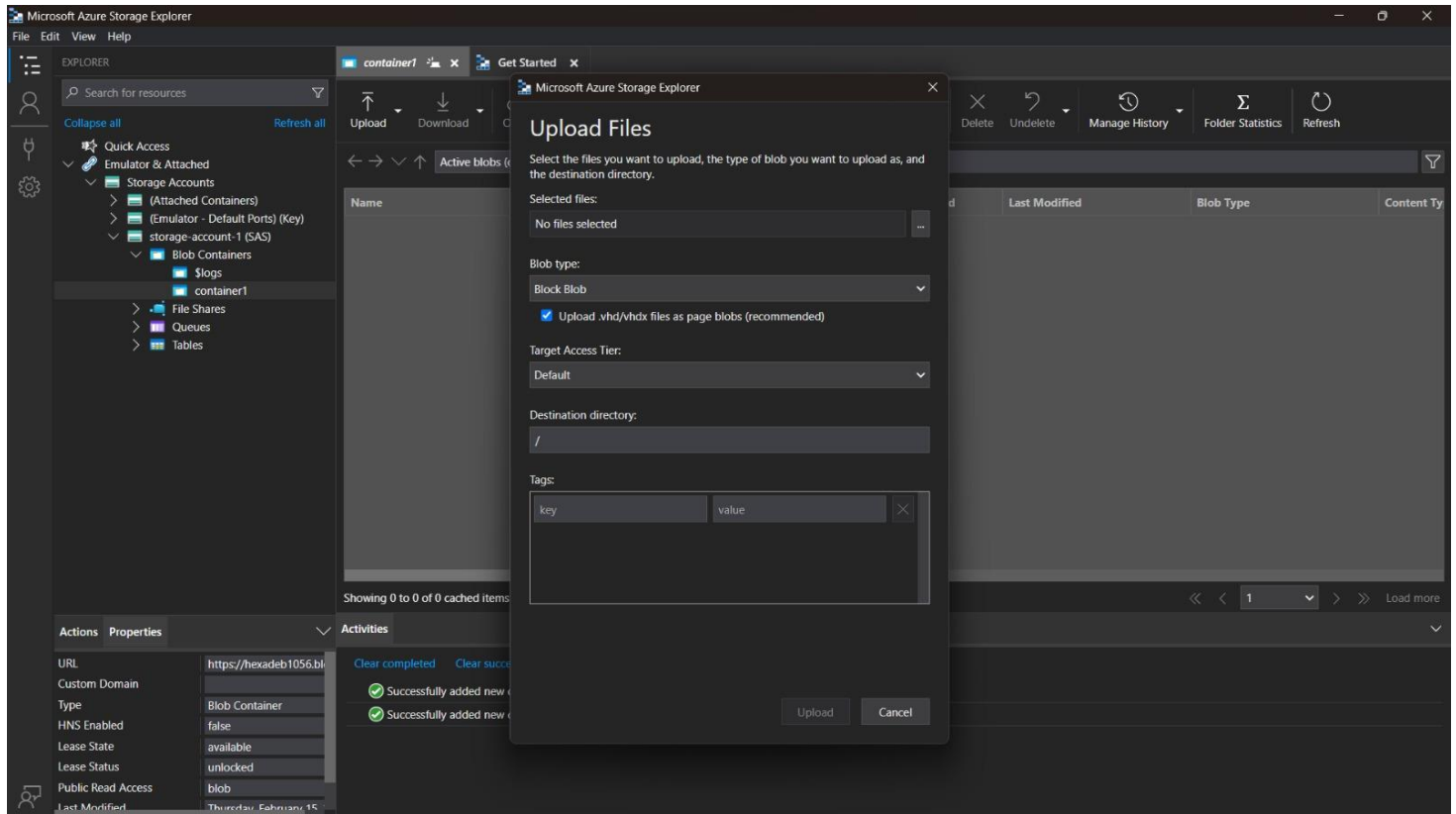
```
BlobEndpoint=https://hexadeb1056.blob.core.windows.net/;QueueEndpoint=https://hexadeb1056.queue.core.windows.net/;FileEndpoint=https://hexadeb1056.file.core.windows.net/;TableEndpoint=https://hexadeb1056.table.core.windows.net/;SharedAccessSignature=sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupiytfx&se=2024-02-16T02:06:13Z&st=2024-02-15T18:06:13Z&sp=103.183.229.153&spr=https&sig=iQ2rMJWwN7Ga8bXQUmatl%2Bt5xjFemTjz0AJVDRp2%2FLI%3D
```

Back

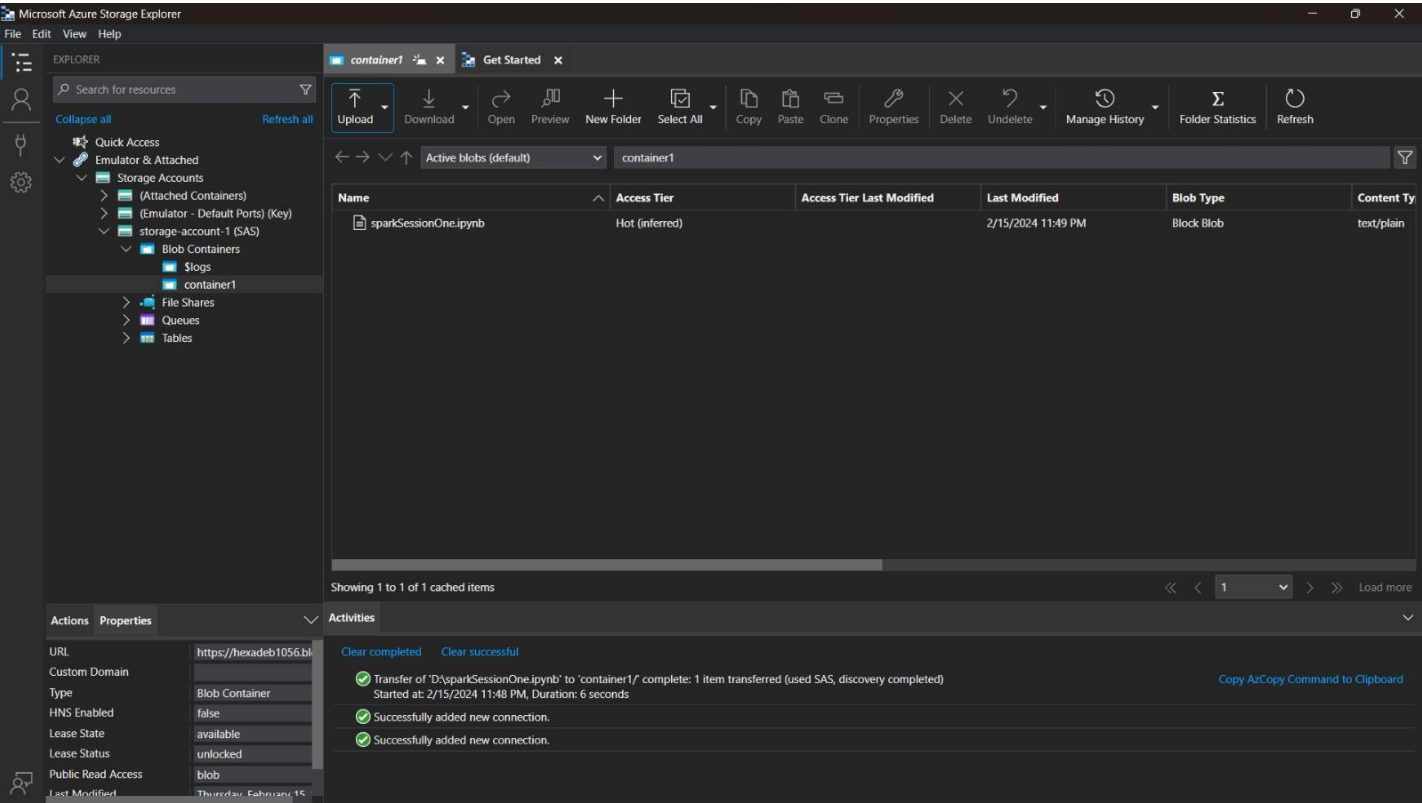
Next

Cancel

- Blob Containers



- File Upload



- Also Visible in Azure

Microsoft Azure

Search resources, services, and docs (G+)

azuser1056_mml.local@...
IH1T (IH1TLODMICROSOFT.COM)

Home > hexadeb1056_1708020145719 | Overview > hexadeb1056 | Containers >

container1

Container

Search

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Create snapshot

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: container1

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> sparkSessionOne.ipynb	15/02/2024, 23:49:06	Hot (Inferred)		Block blob	4.34 KiB	Available	...