Syed Babar Ali
School of Science and Engineering

CS5304 Big Data Services & MLOps — Spring 2025 (Quiz # 3) — March 5th, 2025

1. Databricks Workspace

a. What is a dbc file? What is it used for? (2)

- dbc file is a package that can contain a folder of notebooks or a single notebook. A Databricks archive is a JAR file with extra metadata and has the extension .dbc.
- To allow you to easily distribute Azure Databricks notebooks and other objects

b. How can you run one notebook from another notebook? (1

Ans: Using %run magic command

c. What is the magic command to interact with shell from databricks notebook? (1)

Ans: Using %sh magic command

d. How can we run a notebook or JAR either immediately or on a scheduled basis? (1)

Ans: Using databricks jobs

e. Provide the command to set spark configuration parameters from your notebook. (2)

Ans: `spark.conf.set(“key”, “value”)`

2. Databricks Secrets

a. Regarding databricks secret management, what is the purpose of creating a secret scope? (2):

Ans: Scope are used to group secrets per application

3. Databricks CLI

a. Briefly provide the steps required to configure **databricks cli** with your databricks account. (2)

Ans: a) Execute the command: databricks configure, b) When prompted, provide the workspace URL and personal access token

b. Provide the command to list all files in a user’s workspace. (2)

Ans: databricks workspace list

4. Databricks File System & Unity Catalog (metastore)

a. What is the difference between managed vs external tables? (3)

- **Managed tables** are fully managed by Unity Catalog, which means that Unity Catalog manages both the governance and the underlying data files for each managed table.
- **External tables** are tables whose access from Azure Databricks is managed by Unity Catalog, but whose data lifecycle and file layout are managed using your cloud provider and other data platforms.

b. What is the purpose of a service principal when configuring external storage in Azure databricks? (3)

Ans: Service principal is the identity that databricks assumes when connecting with your external storage (Azure ADLS Gen 2) where that identity is assigned the proper access rights (Azure blog storage contributor).

Service principal is essentially an identity that is used by one azure service to connect to another azure service in which that identity was given access perissmissions.

c. What problem unity catalog solves in an azure databricks environment? Provide at least 3 features that unity catalog provides. (3)

Ans: 1) centralized governance for data and AI including user access and permission management, 2) Built-In Search and Discovery, 3) Secure data sharing and collaboration

d. Briefly explain the 3-level object hierarchy in the unity metastore. (1)

Ans: Under the unity metastore, a catalog is created. Then the chema(database) is created in the catalogs, which then contains the table objects besides the other objects. Tables are then accessed using the 3 level object hierarchy such as: catalog.database.table

e. What is the difference in object namespace in hive metastore vs unity catalog? (2)

Ans: Since there is no unity metastore and no catalogs, database objects are accessed directly using hive_metastore object name: hive_metastore.database.table, whereas unity metastore objects are accessed through catalogs (mentioned in the previous answer)

f. Creating an external table in unity catalog requires combination of which two previously configured unity catalog objects? (2)

Ans: a) credential, b)location

g. Managed tables reside under which object in unity metastore? (2)

Ans: schema (database)

h. What is the purpose of delta sharing? What is required to configure delta sharing in databricks? What is required from the delta sharing client? (3)

Ans: Delta Sharing is a secure data-sharing platform that lets you share data and AI assets with users outside your organization, whether or not those users use Databricks.

Delta Sharing is configured through unity catalog and required remote workspace ID if sharing with a databricks user. Otherwise, client is system independent, and is just required to read parquet files.

- Client authenticates to Sharing Server
- Client requests a table (including filters)
- Server checks access permissions
- Server generates and returns pre-signed short-lived URLs
- Client uses URLs to directly read files from object storage

i. What is the purpose of bronze, silver and gold layers based on databricks guidance? Provide one possible way to map these layers to unity metastore objects. (3)

Ans: Bronze tables have raw data (JSON, CSV, Events/IoT, RDBMS data, etc.).
Silver tables will give a more refined view of our data. We can join fields from various bronze tables to improve streaming records or update account statuses based on recent activity.
Gold tables give business-level aggregates often used for dashboarding and reporting.
b) name the metastore based on the environment (dev, test, prod), then use three catalogs under that metastore with the name: raw (for bronze), ods (for silver) [ods: operational datastore], and dm (for gold) [dm: data mart]