Ansh Ranjan
Databricks Case Study
EXERCISE 4 – Data Storage and Retrieval

TASK 1: save the transformed data into Azure blob

1. Mounting gold layer storage container

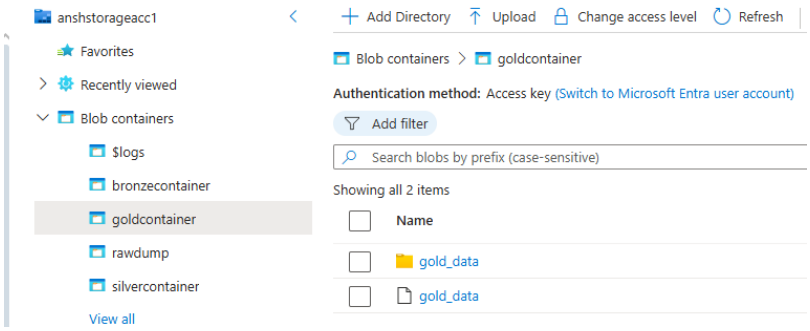
```
dbutils.fs.mount(
    source="wasbs://goldcontainer@anshstorageacc1.blob.core.windows.net/",
    mount_point="/mnt/goldlayer",
    extra_configs={f"fs.azure.account.key.anshstorageacc1.blob.core.windows.net":"VpBXbgXijajAUZHJS3xwP5fj2NWFzKMjaYykkEbt7c7gD28KMwH4PRFsluc2icGJw49Lb6x6KVMD+AStrARmBA=="}
)
```

```
True
```

2. Writing the dataframe in form of Delta Tables in Blob container

```
df_bronze_with_duration.write.format("delta").mode("overwrite").save("dbfs:/mnt/goldlayer/gold_data")
```

3. Checking data in gold container



TASK 2: Read the saved data back into df

1. Reading the delta table back as a new df

```
read_df = spark.read.format("delta").load("dbfs:/mnt/goldlayer/gold_data")
read_df.show(5)
```

```
▸ (1) Spark Jobs
▸ read_df: pyspark.sql.dataframe.DataFrame = [id: string, vendor_id: integer ... 12 more fields]
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+------------------+------------
-------------------+-----------+-----------+
|       id|vendor_id|    pickup_datetime|   dropoff_datetime|passenger_count|   pickup_longitude|    pickup_latitude| dropoff_longitude|   dropoff_latitude|store_and_fwd_flag|trip_duration|t
rip_duration_minutes|day_of_week|hour_of_day|
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+------------------+------------+-
-------------------+-----------+-----------+
|id3380832|        1|2016-04-09 09:56:27|2016-04-09 10:11:12|              1|-73.97080993652344| 40.75213623046875|-73.99332427978516|  40.7461051940918|                 N|         885|
14.75|          7|          9|
|id0033606|        1|2016-04-22 13:59:39|2016-04-22 14:02:47|              1| -73.9861831665039| 40.76225662231445|-73.97857666015625|40.759071350097656|                 N|         188|
```

TASK 3: Explore Databricks Storage Options

Azure Databricks Storage Options

1. **DBFS (Databricks File System)**
    - Managed storage layer built into Databricks.
    - Mounts your cloud storage as /dbfs/.
    - Easy for quick data loading and temp storage.

```
%fs ls
```

Table +

	path	name	size	modificationTime
1	dbfs:/Volume/	Volume/	0	0
2	dbfs:/Volumes/	Volumes/	0	0
3	dbfs:/databricks-dataset...	databricks-dataset...	0	0
4	dbfs:/databricks-results/	databricks-results/	0	0
5	dbfs:/mnt/	mnt/	0	1744452793000
6	dbfs:/volume/	volume/	0	0
7	dbfs:/volumes/	volumes/	0	0

2. **Mounting Azure Storage (Blob or ADLS)**
    - Mount external storage (Blob or ADLS Gen2) to /mnt/your-mount-name.
    - Allows persistent storage with access to raw, bronze, silver, gold layers.
    - Uses dbutils.fs.mount() with access keys or service principals.

```
dbutils.fs.mount(
    source="wasbs://goldcontainer@anshstorageacc1.blob.core.windows.net/",
    mount_point="/mnt/goldlayer",
    extra_configs={f"fs.azure.account.key.anshstorageacc1.blob.core.windows
)
```

```
True
```

3. **Direct Access with ABFS or WASBS URLs**
    - No mount required.
    - Example: "abfss://container@storage.dfs.core.windows.net/"
    - Best for secure, scalable access with Unity Catalog.
4. **External Tables in Data Lake (Delta)**
    - Store Delta tables in ADLS or Blob and register them in Hive or Unity Catalog.
    - Enables scalable data lake architecture (bronze/silver/gold).