INDIAN DATA CLUB

CODE BASICS

databricks

14 DAYS

AI CHALLENGE

DAY 05

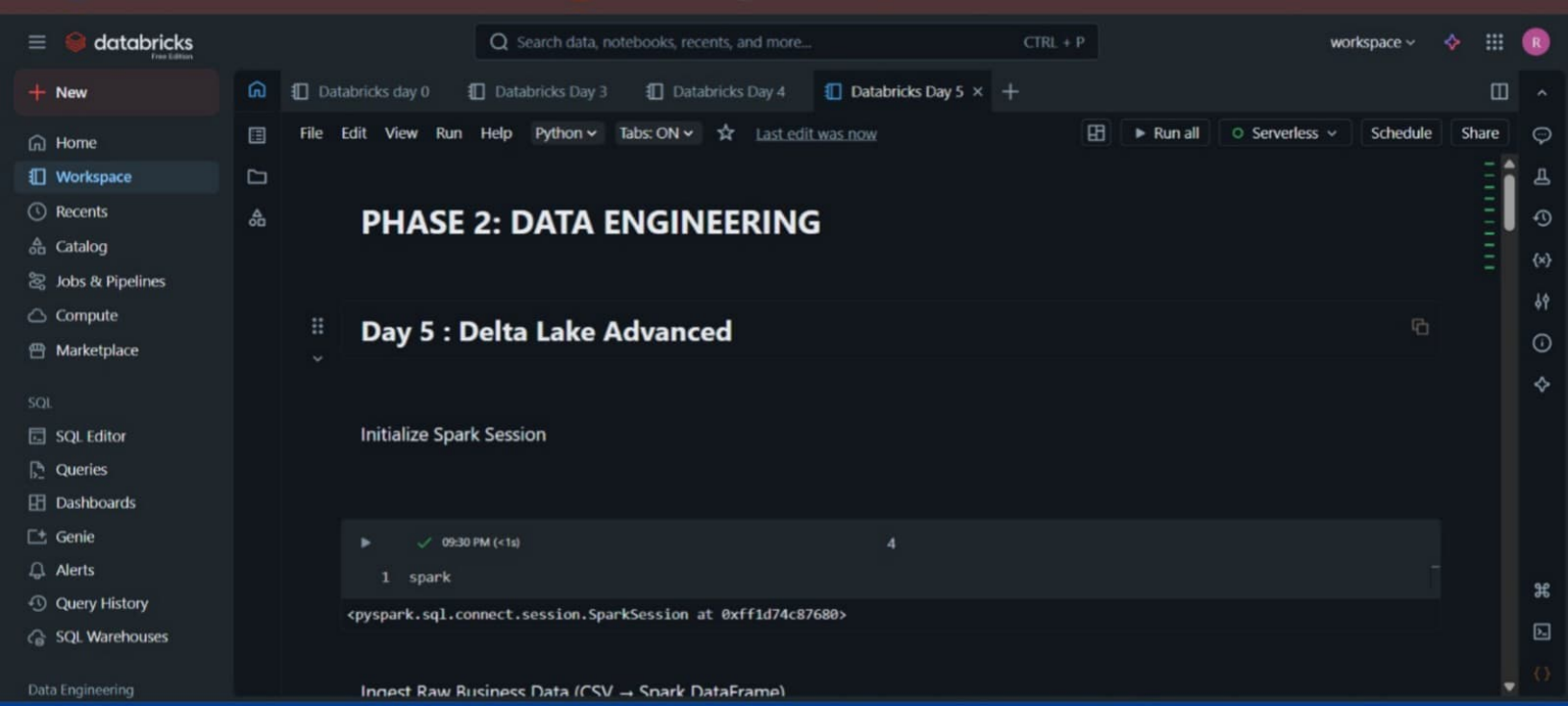
Topic:

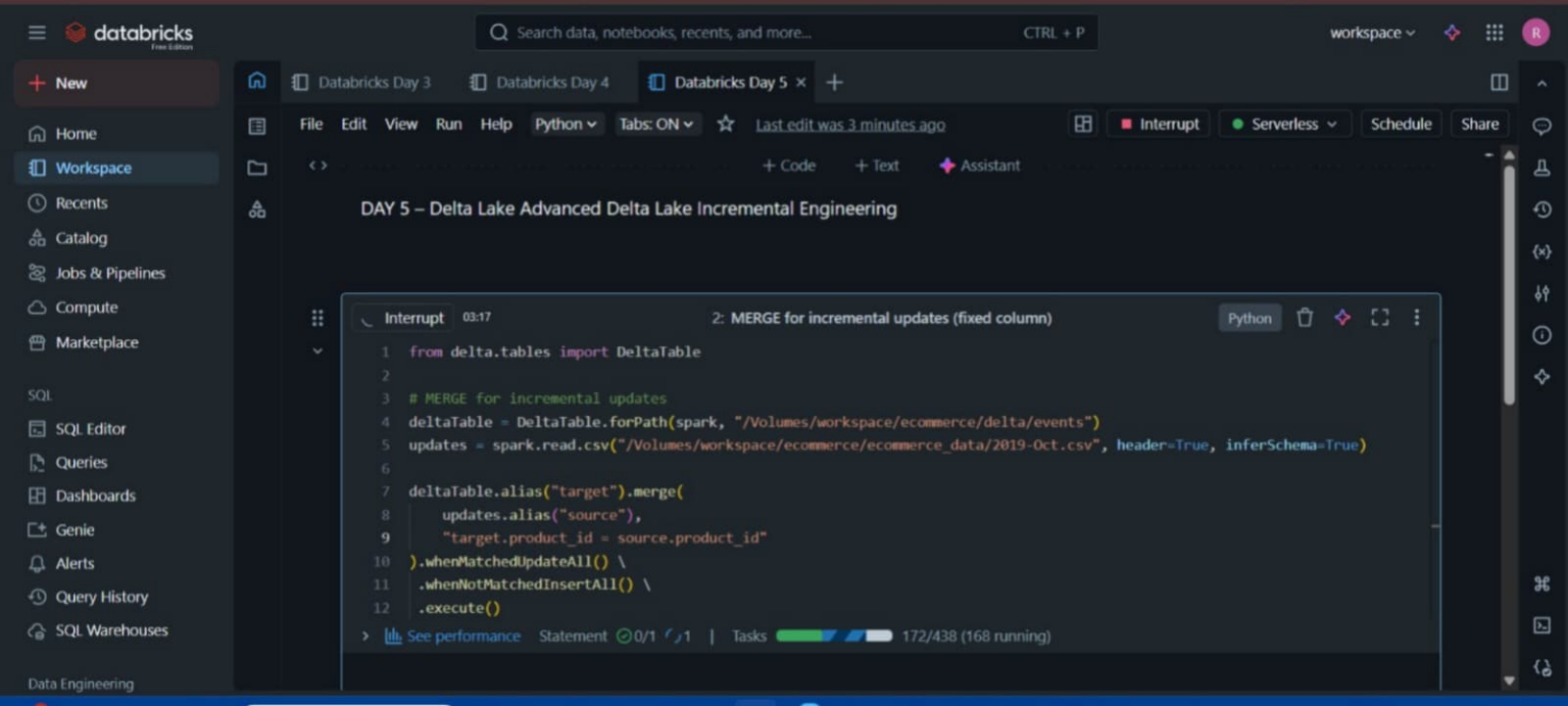
Delta Lake Advanced

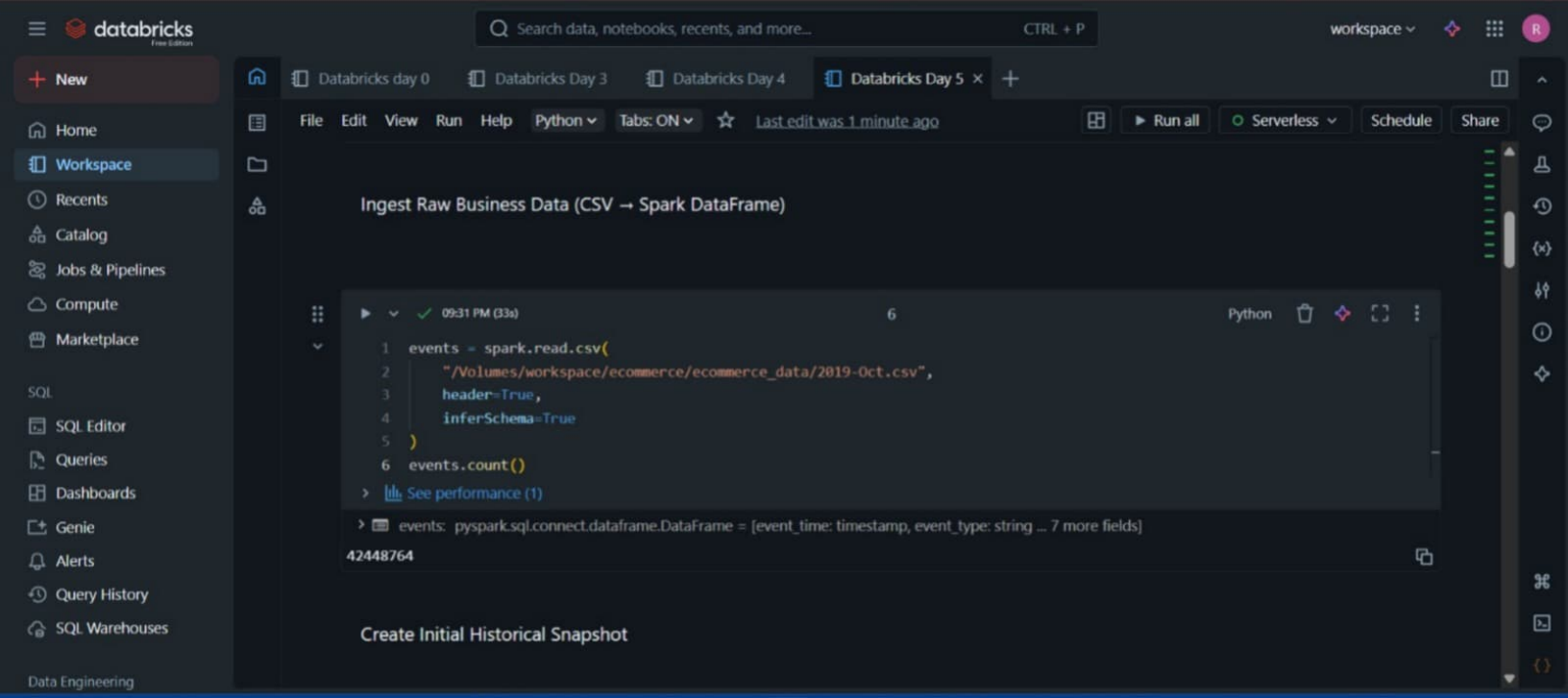
Challenge:

1. Implement incremental MERGE
2. Query historical versions
3. Optimize tables
4. Clean old files

#DatabricksWithIDC


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Databricks day 0
Databricks Day 3
Databricks Day 4
Databricks Day 5
File Edit View Run Help Python Tabs: ON Last edit was now
Run all Serverless Schedule Share
PHASE 2: DATA ENGINEERING
Day 5 : Delta Lake Advanced
Initialize Spark Session
09:30 PM (<1s)
4
1 spark
<pyspark.sql.connect.session.SparkSession at 0xff1d74c87680>
Ingest Raw Business Data (CSV → Spark DataFrame)


databricks
Search data, notebooks, recents, and more...
CTRL + P
workspace
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Databricks Day 3
Databricks Day 4
Databricks Day 5
File Edit View Run Help Python Tabs: ON
Last edit was 3 minutes ago
Interrupt
Serverless
Schedule
Share
+ Code
+ Text
Assistant
DAY 5 – Delta Lake Advanced Delta Lake Incremental Engineering
Interrupt 03:17
2: MERGE for incremental updates (fixed column)
Python
from delta.tables import DeltaTable
# MERGE for incremental updates
deltaTable = DeltaTable.forPath(spark, "/Volumes/workspace/ecommerce/delta/events")
updates = spark.read.csv("/Volumes/workspace/ecommerce/ecommerce_data/2019-Oct.csv", header=True, inferSchema=True)
deltaTable.alias("target").merge(
updates.alias("source"),
"target.product_id = source.product_id"
).whenMatchedUpdateAll() \
.whenNotMatchedInsertAll() \
.execute()
See performance
Statement 0/1 1
Tasks
172/438 (168 running)


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Databricks day 0
Databricks Day 3
Databricks Day 4
Databricks Day 5
File Edit View Run Help Python Tabs: ON Last edit was 1 minute ago
Run all Serverless Schedule Share
Ingest Raw Business Data (CSV → Spark DataFrame)
09:31 PM (33s)
6
Python
events = spark.read.csv(
"/Volumes/workspace/ecommerce/ecommerce_data/2019-Oct.csv",
header=True,
inferSchema=True
)
events.count()
See performance (1)
events: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]
42448764
Create Initial Historical Snapshot

databricks
Free Edition

Search data, notebooks, recents, and more... CTRL + P

workspace

New

- Home
- Workspace
- Recents
- Catalog
- Jobs & Pipelines
- Compute
- Marketplace

SQL

- SQL Editor
- Queries
- Dashboards
- Genie
- Alerts
- Query History
- SQL Warehouses

Data Engineering

Databricks day 0 | Databricks Day 3 | Databricks Day 4 | Databricks Day 5 ×

File Edit View Run Help Python Tabs: ON Last edit was 2 minutes ago

Run all | Serverless | Schedule | Share

Create Initial Historical Snapshot

09:32 PM (2s) 8

```
1 batch_1 = events.limit(100000)
2 batch_1.count()
```

See performance (1)

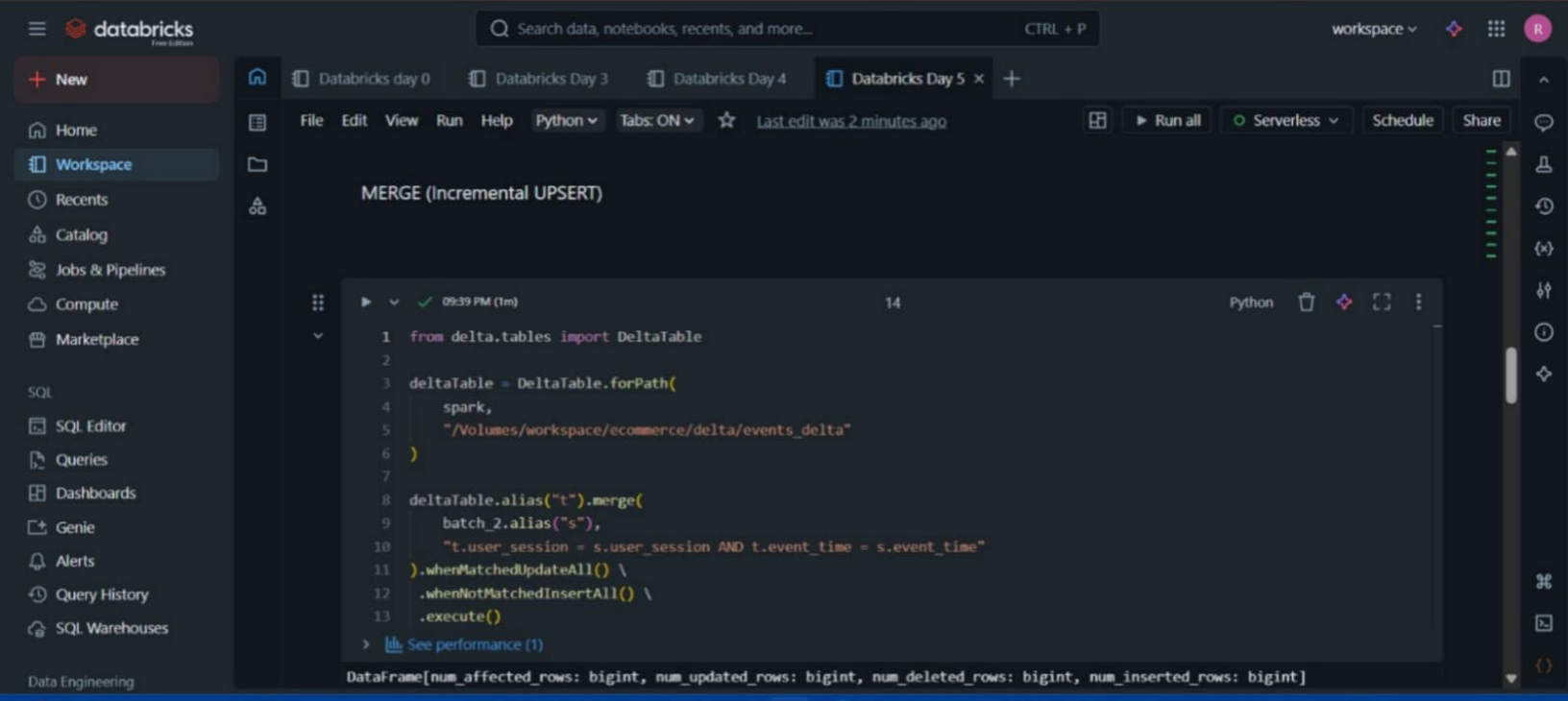
batch_1: pyspark.sql.connect.dataframe.DataFrame = [event_time: timestamp, event_type: string ... 7 more fields]

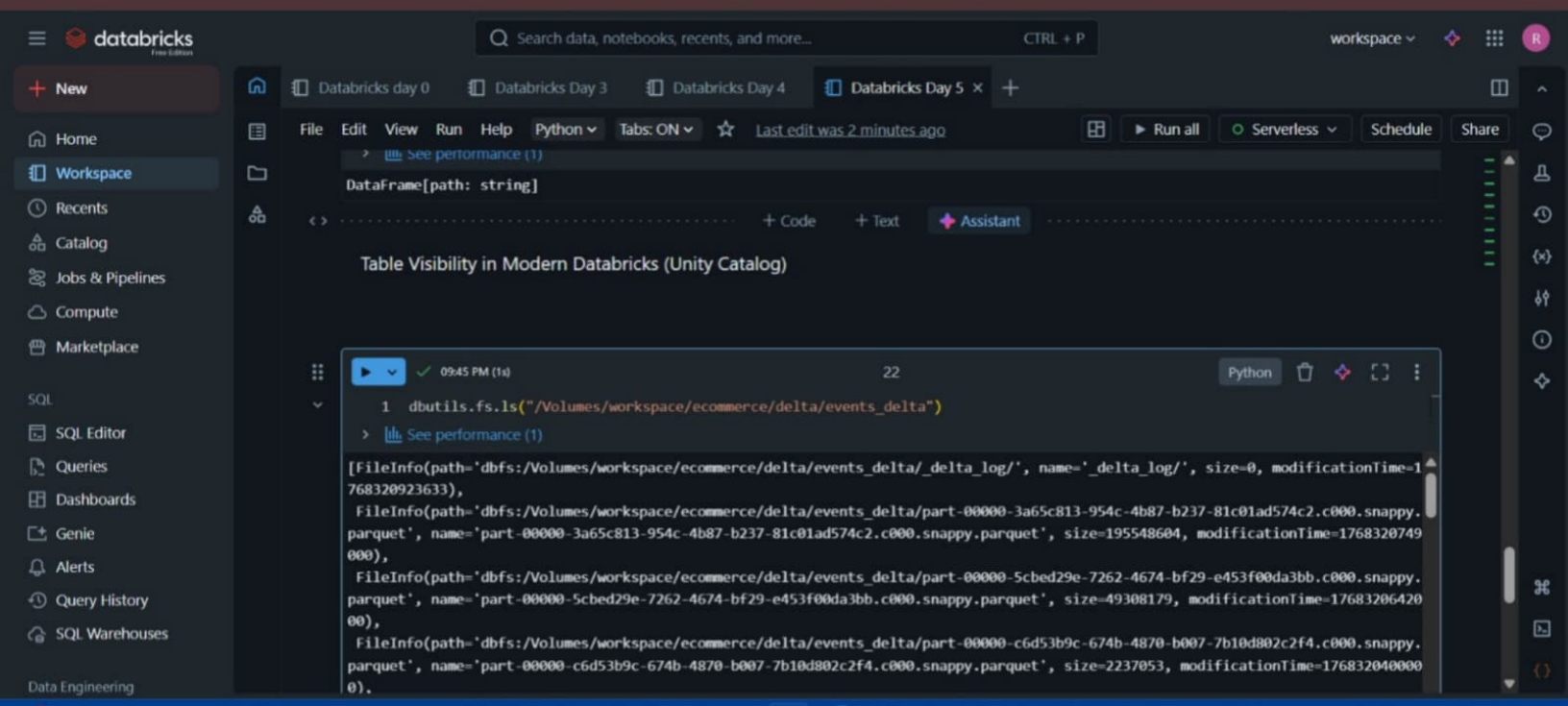
```
100000
```

Create Delta Table

09:36 PM (6s) 10: Create Delta Table (fixed path)

```
1 batch_1.write.format("delta").mode("overwrite") \
```


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Databricks day 0
Databricks Day 3
Databricks Day 4
Databricks Day 5
File Edit View Run Help Python Tabs: ON Last edit was 2 minutes ago
Run all Serverless Schedule Share
MERGE (Incremental UPSERT)
09:39 PM (1m)
14
Python
from delta.tables import DeltaTable
deltaTable = DeltaTable.forPath(
spark,
"/Volumes/workspace/ecommerce/delta/events_delta"
)
deltaTable.alias("t").merge(
batch_2.alias("s"),
"t.user_session = s.user_session AND t.event_time = s.event_time"
).whenMatchedUpdateAll() \
.whenNotMatchedInsertAll() \
.execute()
See performance (1)
DataFrame[num_affected_rows: bigint, num_updated_rows: bigint, num_deleted_rows: bigint, num_inserted_rows: bigint]


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
New
Home
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Databricks day 0
Databricks Day 3
Databricks Day 4
Databricks Day 5
File Edit View Run Help Python Tabs: ON Last edit was 2 minutes ago
Run all
Serverless
Schedule
Share
See performance (1)
DataFrame[path: string]
+ Code
+ Text
Assistant
Table Visibility in Modern Databricks (Unity Catalog)
09:45 PM (1s)
22
Python
1 dbutils.fs.ls("/Volumes/workspace/ecommerce/delta/events_delta")
See performance (1)
[FileInfo(path='dbfs:/Volumes/workspace/ecommerce/delta/events_delta/_delta_log/', name='_delta_log/', size=0, modificationTime=1768320923633),
FileInfo(path='dbfs:/Volumes/workspace/ecommerce/delta/events_delta/part-00000-3a65c813-954c-4b87-b237-81c01ad574c2.c000.snappy.parquet', name='part-00000-3a65c813-954c-4b87-b237-81c01ad574c2.c000.snappy.parquet', size=195548604, modificationTime=1768320749000),
FileInfo(path='dbfs:/Volumes/workspace/ecommerce/delta/events_delta/part-00000-5cbed29e-7262-4674-bf29-e453f00da3bb.c000.snappy.parquet', name='part-00000-5cbed29e-7262-4674-bf29-e453f00da3bb.c000.snappy.parquet', size=49308179, modificationTime=1768320642000),
FileInfo(path='dbfs:/Volumes/workspace/ecommerce/delta/events_delta/part-00000-c6d53b9c-674b-4870-b007-7b10d802c2f4.c000.snappy.parquet', name='part-00000-c6d53b9c-674b-4870-b007-7b10d802c2f4.c000.snappy.parquet', size=2237053, modificationTime=1768320400000),