Name: Krushnakumar Patle

Email: krishnapatle128@gmail.com

Batch: Data Engineering Batch-1

Delta Lake:

Delta Lake is the optimized storage layer that provides the foundation for storing data and tables in the Databricks lakehouse. Delta Lake is open source software that extends Parquet data files with a file-based transaction log for ACID transactions and scalable metadata handling. Delta Lake is fully compatible with Apache Spark APIs, and was developed for tight integration with Structured Streaming, allowing you to easily use a single copy of data for both batch and streaming operations and providing incremental processing at scale.

To set up a Python project (for example, for unit testing), you can install Delta Lake using pip install delta-spark==3.1.0 and then configure the SparkSession with the configure_spark_with_delta_pip() utility function in Delta Lake.

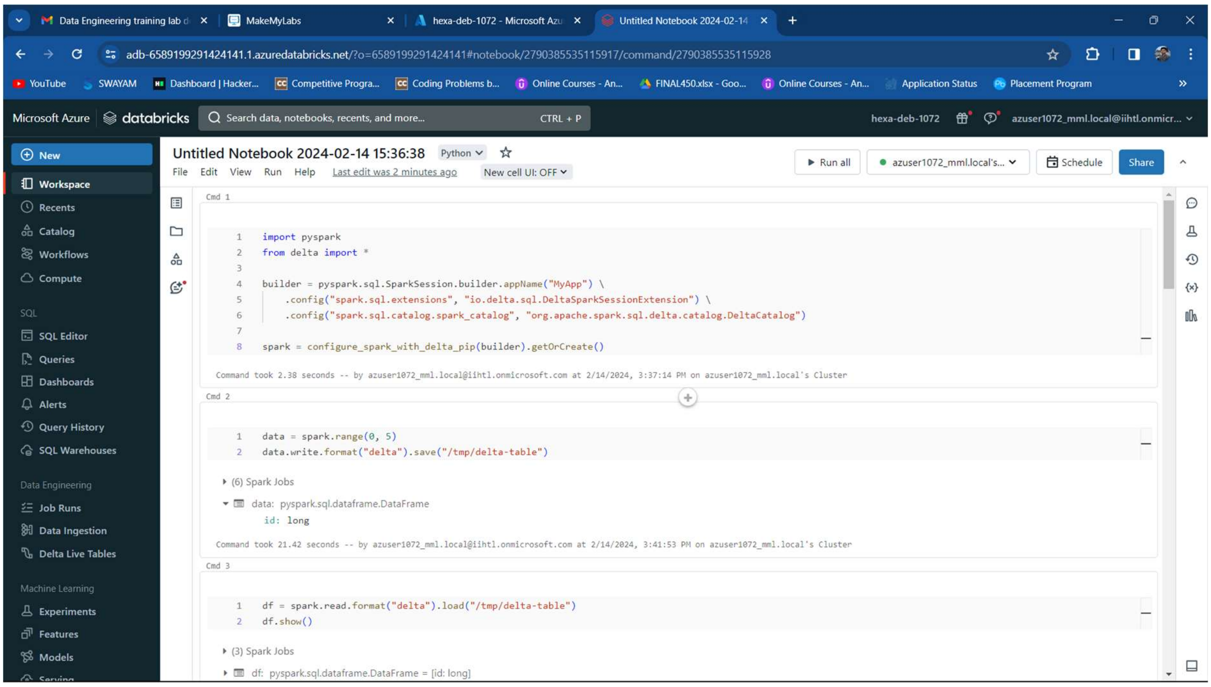
Create a table:

To create a Delta table, write a DataFrame out in the delta format. You can use existing Spark SQL code and change the format from parquet, csv, json, and so on, to delta.

Read data:

You read data in your Delta table by specifying the path to the files: "/tmp/delta-table":

Day 18 Assessment



Untitled Notebook 2024-02-14 15:36:38

Cmd 1

```python
import pyspark
from delta import *

builder = pyspark.sql.SparkSession.builder.appName("MyApp") \
    .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
    .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog")

spark = configure_spark_with_delta_pip(builder).getOrCreate()
```

```
Command took 2.38 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:37:14 PM on azuser1072_mml.local's Cluster
```

Cmd 2

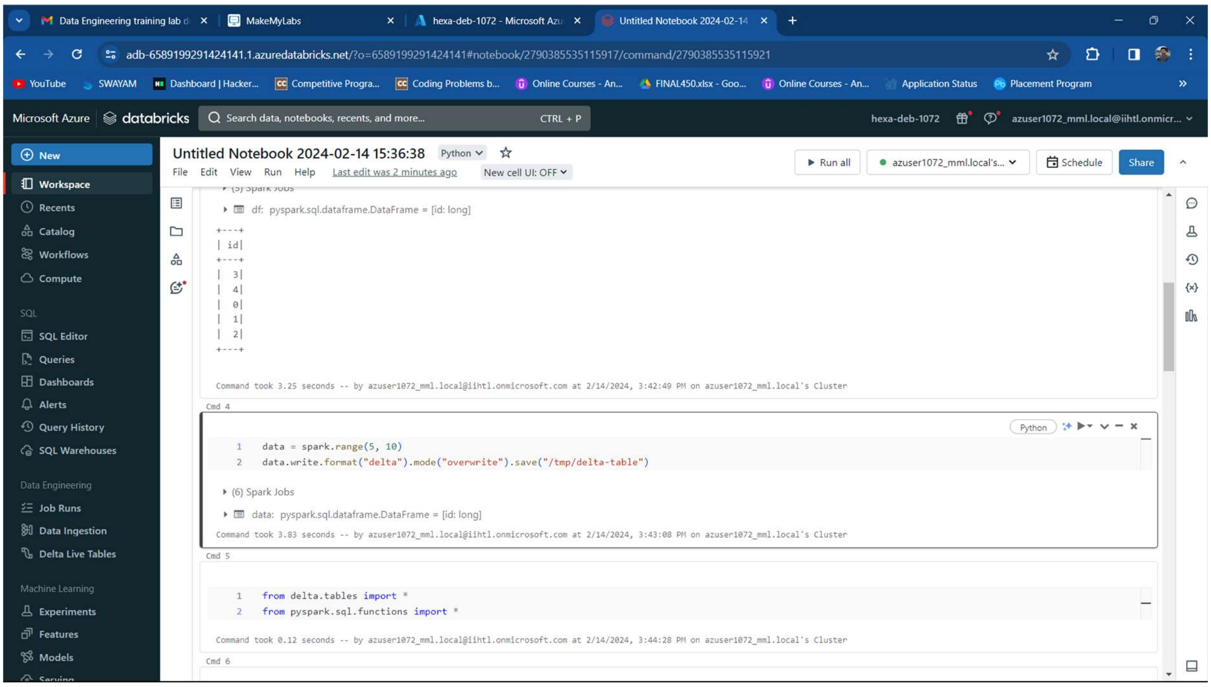
```python
data = spark.range(0, 5)
data.write.format("delta").save("/tmp/delta-table")
```

```
(6) Spark Jobs
data:  pyspark.sql.dataframe.DataFrame
    id:  long
Command took 21.42 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:41:53 PM on azuser1072_mml.local's Cluster
```

Cmd 3

```python
df = spark.read.format("delta").load("/tmp/delta-table")
df.show()
```

```
(3) Spark Jobs
df:  pyspark.sql.dataframe.DataFrame = [id: long]
```



Untitled Notebook 2024-02-14
adb-6589199291424141.1.azuredatabricks.net/?o=6589199291424141#notebook/2790385535115917/command/2790385535115921
Microsoft Azure
databricks
Run all
Schedule
Share

Untitled Notebook 2024-02-14 15:36:38

```
df:  pyspark.sql.dataframe.DataFrame = [id: long]
+---+
| id|
+---+
|  3|
|  4|
|  0|
|  1|
|  2|
+---+

Command took 3.25 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:42:49 PM on azuser1072_mml.local's Cluster
```

Cmd 4

```python
data = spark.range(5, 10)
data.write.format("delta").mode("overwrite").save("/tmp/delta-table")
```

```
(6) Spark Jobs
data:  pyspark.sql.dataframe.DataFrame = [id: long]
Command took 3.83 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:43:08 PM on azuser1072_mml.local's Cluster
```

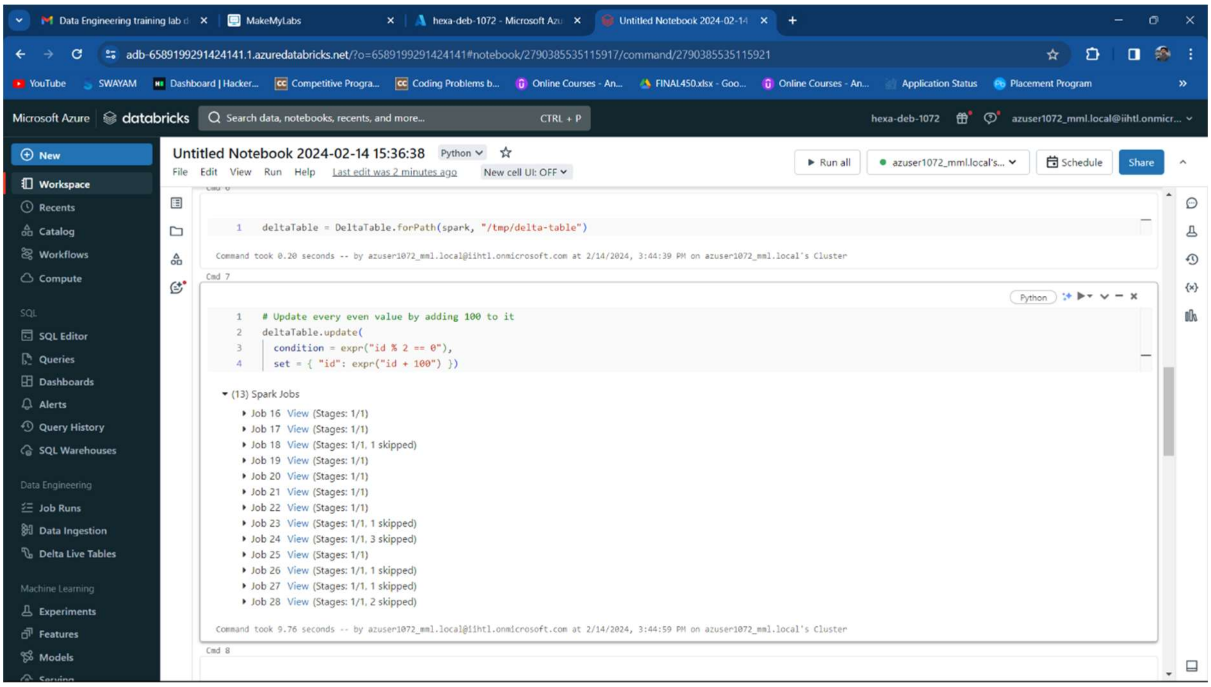
Cmd 5

```python
from delta.tables import *
from pyspark.sql.functions import *
```

```
Command took 0.12 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:44:28 PM on azuser1072_mml.local's Cluster
```

Cmd 6

Day 18 Assessment



```
deltaTable = DeltaTable.forPath(spark, "/tmp/delta-table")
```

```
Command took 0.20 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:44:39 PM on azuser1072_mml.local's Cluster
```

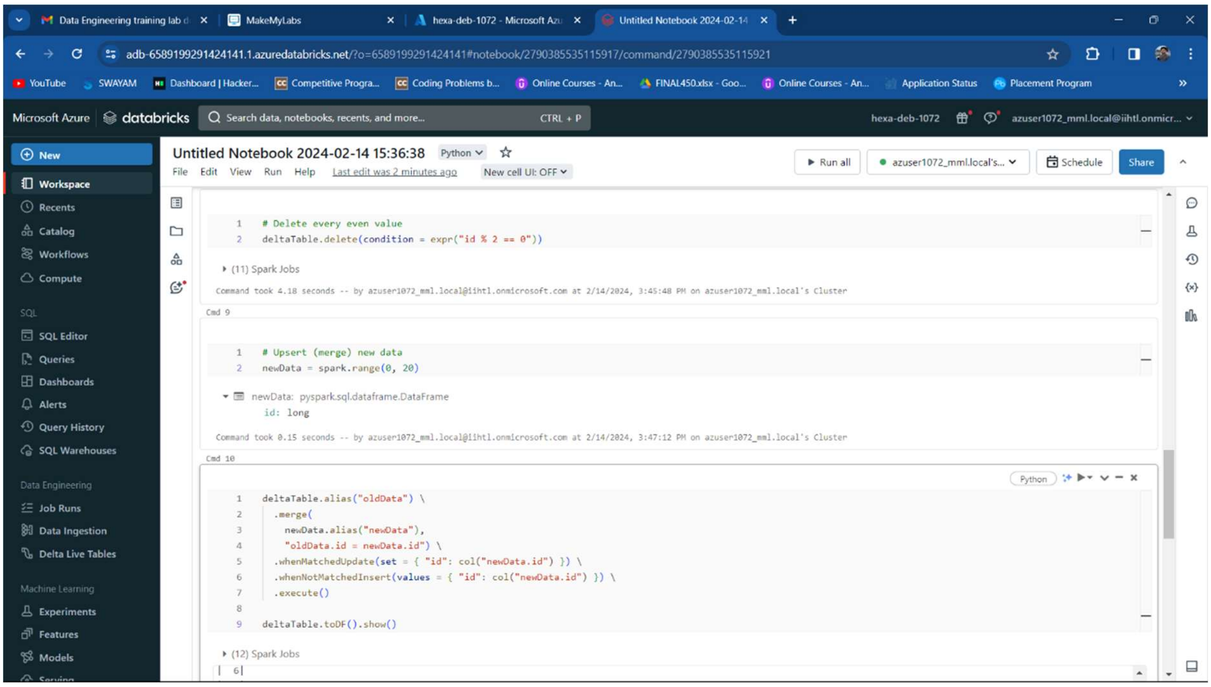
Cmd 7

```
# Update every even value by adding 100 to it
deltaTable.update(
  condition = expr("id % 2 == 0"),
  set = { "id": expr("id + 100") })
```

```
(13) Spark Jobs
  Job 16  View (Stages: 1/1)
  Job 17  View (Stages: 1/1)
  Job 18  View (Stages: 1/1, 1 skipped)
  Job 19  View (Stages: 1/1)
  Job 20  View (Stages: 1/1)
  Job 21  View (Stages: 1/1)
  Job 22  View (Stages: 1/1)
  Job 23  View (Stages: 1/1, 1 skipped)
  Job 24  View (Stages: 1/1, 3 skipped)
  Job 25  View (Stages: 1/1)
  Job 26  View (Stages: 1/1, 1 skipped)
  Job 27  View (Stages: 1/1, 1 skipped)
  Job 28  View (Stages: 1/1, 2 skipped)

Command took 9.76 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:44:59 PM on azuser1072_mml.local's Cluster
```

Cmd 8


Data Engineering training lab
MakeMyLabs
hexa-deb-1072 - Microsoft Azure
Untitled Notebook 2024-02-14
adb-6589199291424141.1.azuredatabricks.net/?o=6589199291424141#notebook/2790385535115917/command/2790385535115921
Microsoft Azure
databricks
Untitled Notebook 2024-02-14 15:36:38
Run all
Schedule
Share
Python

```
# Delete every even value
deltaTable.delete(condition = expr("id % 2 == 0"))
```

```
(11) Spark Jobs
Command took 4.18 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:45:48 PM on azuser1072_mml.local's Cluster
```

Cmd 9

```
# Upsert (merge) new data
newData = spark.range(0, 20)
```

```
newData:  pyspark.sql.dataframe.DataFrame
    id:  long
Command took 0.15 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:47:12 PM on azuser1072_mml.local's Cluster
```

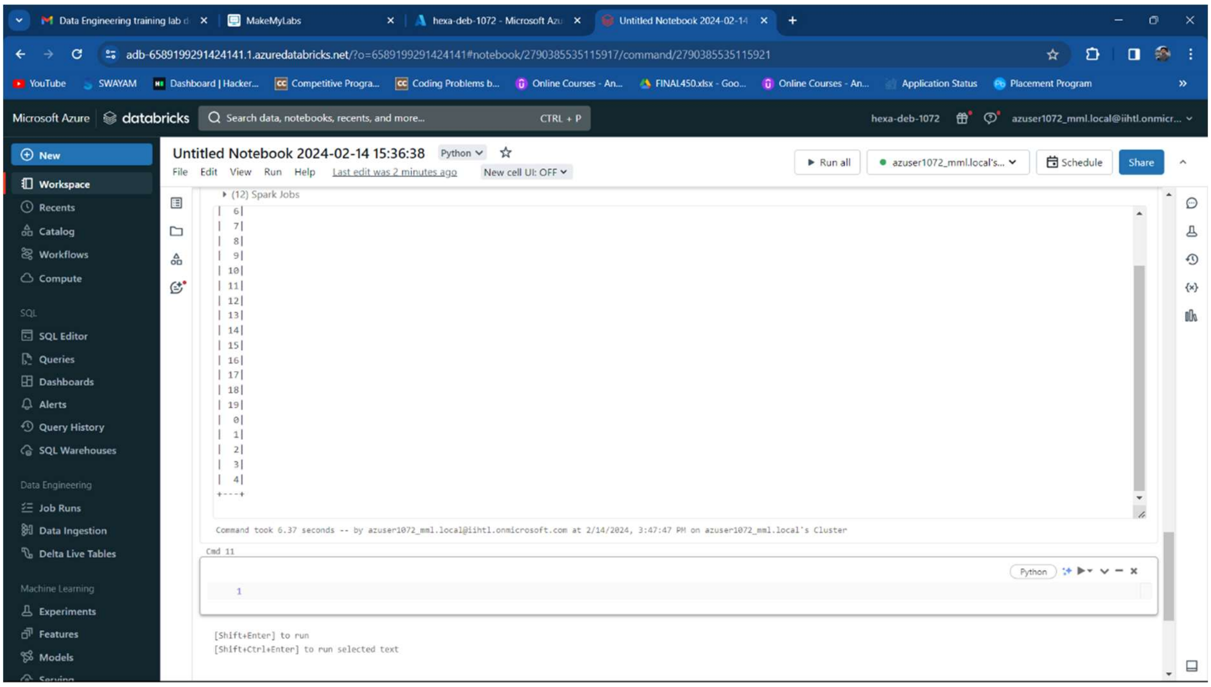
Cmd 10

```
deltaTable.alias("oldData") \
  .merge(
    newData.alias("newData"),
    "oldData.id = newData.id") \
  .whenMatchedUpdate(set = { "id": col("newData.id") }) \
  .whenNotMatchedInsert(values = { "id": col("newData.id") }) \
  .execute()

deltaTable.toDF().show()
```

```
(12) Spark Jobs
|  6|
```

Day 18 Assessment



Untitled Notebook 2024-02-14 15:36:38
Python
File
Edit
View
Run
Help
Last edit was 2 minutes ago
New cell UI: OFF
Run all
Schedule
Share

(12) Spark Jobs

```
|  6|
|  7|
|  8|
|  9|
| 10|
| 11|
| 12|
| 13|
| 14|
| 15|
| 16|
| 17|
| 18|
| 19|
|  0|
|  1|
|  2|
|  3|
|  4|
+---+
```

Command took 6.37 seconds -- by azuser1072_mml.local@iihtl.onmicrosoft.com at 2/14/2024, 3:47:47 PM on azuser1072_mml.local's Cluster

Cmd 11

```
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
```