

Name: Krushnakumar Patle

Email: [krishnapatle128@gmail.com](mailto: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

The screenshot shows a Databricks Notebook interface. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, Features, Models, and Services. The main area displays two commands. Command 1 sets up a Spark session with Delta Lake extensions and creates a catalog. Command 2 generates a range of data and writes it to a Delta table.

```
Cmd 1
1 import pyspark
2 from delta import *
3
4 builder = pyspark.sql.SparkSession.builder.appName("MyApp") \
5     .config("spark.sql.extensions", "io.delta.sql.DeltaSparkSessionExtension") \
6     .config("spark.sql.catalog.spark_catalog", "org.apache.spark.sql.delta.catalog.DeltaCatalog")
7
8 spark = configure_spark_with_delta_pip(builder).getOrCreate()

Command took 2.38 seconds -- by azuser1072_mml.local@ih1t1.omnicrossoft.com at 2/14/2024, 3:37:14 PM on azuser1072_mml.local's Cluster

Cmd 2
1 data = spark.range(0, 5)
2 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@ih1t1.omnicrossoft.com at 2/14/2024, 3:41:53 PM on azuser1072_mml.local's Cluster

Cmd 3
1 df = spark.read.format("delta").load("/tmp/delta-table")
2 df.show()

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

The screenshot shows the continuation of the Databricks Notebook. Command 3 displays the contents of the Delta table, showing a single row with id=3. Command 4 generates a larger range of data (0 to 10) and writes it to the same Delta table, overwriting the previous data. Command 5 imports the Delta Lake tables and functions modules.

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

Command took 3.25 seconds -- by azuser1072_mml.local@ih1t1.omnicrossoft.com at 2/14/2024, 3:42:49 PM on azuser1072_mml.local's Cluster

Cmd 4
1 data = spark.range(5, 10)
2 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@ih1t1.omnicrossoft.com at 2/14/2024, 3:43:08 PM on azuser1072_mml.local's Cluster

Cmd 5
1 from delta.tables import *
2 from pyspark.sql.functions import *

Command took 0.12 seconds -- by azuser1072_mml.local@ih1t1.omnicrossoft.com at 2/14/2024, 3:44:28 PM on azuser1072_mml.local's Cluster

Cmd 6
```

## Day 18 Assessment

The screenshot shows a Databricks workspace interface. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, Features, Models, and Recent. The main area displays a notebook titled "Untitled Notebook 2024-02-14 15:36:38" in Python. The notebook content includes a command to create a Delta table and a subsequent update operation. The output shows a list of Spark jobs.

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

Command took 0.20 seconds -- by azuser1072\_mml.local@ih1t1.onmicrosoft.com at 2/14/2024, 3:44:39 PM on azuser1072\_mml.local's Cluster

```
1 # Update every even value by adding 100 to it
2 deltaTable.update(
3     condition = expr("id % 2 == 0"),
4     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@ih1t1.onmicrosoft.com at 2/14/2024, 3:44:59 PM on azuser1072\_mml.local's Cluster

The screenshot shows a Databricks workspace interface. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Experiments, Features, Models, and Recent. The main area displays a notebook titled "Untitled Notebook 2024-02-14 15:36:38" in Python. The notebook content includes a command to delete even values, a command to create new data, and a command to merge the new data into the existing Delta table. The output shows a list of Spark jobs.

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

(11) Spark Jobs

Command took 4.18 seconds -- by azuser1072\_mml.local@ih1t1.onmicrosoft.com at 2/14/2024, 3:45:48 PM on azuser1072\_mml.local's Cluster

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

newData: pyspark.sql.dataframe.DataFrame

id: long

Command took 0.15 seconds -- by azuser1072\_mml.local@ih1t1.onmicrosoft.com at 2/14/2024, 3:47:12 PM on azuser1072\_mml.local's Cluster

```
1 deltaTable.alias("oldData") \
2     .merge(
3         newData.alias("newData"),
4         "oldData.id = newData.id" ) \
5     .whenMatchedUpdate(set = { "id": col("newData.id") }) \
6     .whenNotMatchedInsert(values = { "id": col("newData.id") }) \
7     .execute()
8
9 deltaTable.toDF().show()
```

(12) Spark Jobs

6

## Day 18 Assessment

The screenshot displays the Databricks web interface in a browser. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, and a search bar. The left sidebar contains a 'New' button and a 'Workspace' section with various navigation options like Recents, Catalog, Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, and SQL Warehouses. The main area shows an 'Untitled Notebook 2024-02-14 15:36:38' with a Python language selector. The notebook content consists of a list of numbers from 6 to 19, followed by 0 to 4, and a final line with four plus signs. Below the code, a command execution box shows the command took 6.37 seconds. The bottom of the interface includes a 'Run all' button, a user dropdown, and a 'Schedule' button.

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P hexa-deb-1072 azuser1072\_mml.local@ihti.onmicr...

New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments Features Models Session

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 azuser1072\_mml.local's... 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@ihti.onmicrsoft.com at 2/14/2024, 3:47:47 PM on azuser1072\_mml.local's Cluster

Cmd 11

Python Run Stop Close

1

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