Data Engineering Batch 1   **Day 18:14/02/2024**

Name: Pradip Bochare

🪟 SQL Delta

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

Delta Python ⌄ ☆
File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ⌄   ▶ Run all   ■ Terminated ⌄   📅 Schedule   Share

5 rows | 3.85 seconds runtime   Refreshed 7 hours ago

ⓘ This result is stored as PySpark data frame `_sqldf` and in the IPython output cache as `Out[3]` . Learn more

Command took 3.85 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:00:27 on azuser1077_mml.local's Cluster

Cmd 3

```sql
1   %sql
2   INSERT OVERWRITE delta.`/tmp/deltain-table` SELECT col1 as id FROM VALUES 5,6,7,8,9;
```

▸ (6) Spark Jobs

▸ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [num_affected_rows: long, num_inserted_rows: long]

Table ⌄ +   New result table: OFF ⌄

	num_affected_rows ▲	num_inserted_rows ▲	
1	5	5	

1 row | 4.35 seconds runtime   Refreshed 7 hours ago

ⓘ This result is stored as PySpark data frame `_sqldf` and in the IPython output cache as `Out[4]` . Learn more

Command took 4.35 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:05:31 on azuser1077_mml.local's Cluster

Cmd 4

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

Delta Python ⌄ ☆
File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ⌄   ▶ Run all   ■ Terminated ⌄   📅 Schedule   Share

Cmd 4

SQL

```sql
1   %sql
2   SELECT * FROM delta.`/tmp/deltain-table`;
```

▸ (2) Spark Jobs

▸ ▦ _sqldf: pyspark.sql.dataframe.DataFrame = [id: integer]

Table ⌄ +   New result table: OFF ⌄

	id ▲
1	5
2	6
3	7
4	8
5	9

5 rows | 1.36 seconds runtime   Refreshed 7 hours ago

ⓘ This result is stored as PySpark data frame `_sqldf` and in the IPython output cache as `Out[5]` . Learn more

Command took 1.36 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:06:41 on azuser1077_mml.local's Cluster

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

🪟 Python Delta

python-delta Python ⌄ ☆
File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ⌄   ▶ Run all   ■ Terminated ⌄   📅 Schedule   Share ⌃

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

▸ (4) Spark Jobs
▸ 🔲 df: pyspark.sql.dataframe.DataFrame = [id: long]

```
+---+
| id|
+---+
|  3|
|  4|
|  0|
|  1|
|  2|
+---+
```

Table ⌄ +   New result table: OFF ⌄

	id ▲
1	3
2	4
3	0
4	1
5	2

---

python-delta Python ⌄ ☆
File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ⌄   ▶ Run all   ■ Terminated ⌄   📅 Schedule   Share ⌃

Cmd 6

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

▸ (10) Spark Jobs
▸ 🔲 data: pyspark.sql.dataframe.DataFrame = [id: long]

```
+---+
| id|
+---+
|  8|
|  9|
|  5|
|  6|
|  7|
+---+
```

Command took 4.15 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:15:16 on azuser1077_mml.local's Cluster

Cmd 7

```python
1  streamingDf = spark.readStream.format("rate").load()
2  stream = streamingDf.selectExpr("value as id").writeStream.format("delta").option("checkpointLocation", "/tmp/checkpoint").
   start("/tmp/delta-table")
```

python-delta   Python ∨   ☆
File   Edit   View   Run   Help   Last edit was 6 hours ago   New cell UI: OFF ∨   ▶ Run all   ■ Terminated ∨   🗓 Schedule   Share

Cmd 7

```
1   streamingDf = spark.readStream.format("rate").load()
2   stream = streamingDf.selectExpr("value as id").writeStream.format("delta").option("checkpointLocation", "/tmp/checkpoint").
        start("/tmp/delta-table")
```

▶ (1) Spark Jobs

▶ ⊘ cd184367-5b66-4a7d-8007-6cd6fec28ca1   Last updated: 6 hours ago

▶ ▦ streamingDf: pyspark.sql.dataframe.DataFrame = [timestamp: timestamp, value: long]

Command complete

Cmd 8

```
1   from delta.tables import *
2   from pyspark.sql.functions import *
```

Command took 0.04 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:22:48 on azuser1077_mml.local's Cluster

Cmd 9

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

Command took 0.24 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:22:51 on azuser1077_mml.local's Cluster

Cmd 10

---

python-delta   Python ∨   ☆
File   Edit   View   Run   Help   Last edit was 6 hours ago   New cell UI: OFF ∨   ▶ Run all   ■ Terminated ∨   🗓 Schedule   Share

Cmd 10

```
1   deltaTable.update(
2     condition = expr("id % 2 == 0"),
3     set = { "id": expr("id + 100") })
4
```

▶ (12) Spark Jobs

Command took 14.07 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:22:53 on azuser1077_mml.local's Cluster

Cmd 11

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

▶ (6) Spark Jobs

▶ ▦ df: pyspark.sql.dataframe.DataFrame = [id: long]

|259|
|247|
|303|
|267|
|279|
|305|
|271|
|293|

python-delta   Python ∨ ☆

File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ∨   ▶ Run all   ■ Terminated ∨   ▤ Schedule   Share

Cmd 11

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

▶ (6) Spark Jobs

▶ ▣ df: pyspark.sql.dataframe.DataFrame = [id: long]

```
|259|
|247|
|303|
|267|
|279|
|305|
|271|
|293|
|295|
|283|
|307|
|281|
|275|
|285|
|287|
|273|
|253|
|309|
+---+
only showing top 20 rows
```

python-delta   Python ∨ ☆

File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ∨   ▶ Run all   ■ Terminated ∨   ▤ Schedule   Share

	id
1	5
2	106
3	108
4	100
5	102
6	104
7	106

⤓ 376 rows | 2.16 seconds runtime   Refreshed 6 hours ago

Command took 2.16 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:23:51 on azuser1077_mml.local's Cluster

Cmd 12

```
1  deltaTable.delete(condition = expr("id % 2 == 0"))
```

▶ (9) Spark Jobs

Command took 4.75 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:25:32 on azuser1077_mml.local's Cluster

Cmd 13

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

python-delta Python ∨ ☆
File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ∨   ▶ Run all   ■ Terminated ∨   Schedule   Share

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

▶ (4) Spark Jobs
▶ ▦ df: pyspark.sql.dataframe.DataFrame = [id: long]

```
|  9|
|  1|
|  3|
|  5|
|  7|
|  9|
| 11|
| 13|
| 15|
| 17|
| 19|
| 21|
| 23|
| 25|
| 27|
| 29|
| 31|
| 33|
+---+
only showing top 20 rows
```

25°C Mostly cloudy — Search — ENG IN — 9:43 PM 2/14/2024

---

	id
1	5
2	7
3	9
4	1
5	3
6	5
7	7

188 rows | 1.11 seconds runtime   Refreshed 6 hours ago

Command took 1.11 seconds -- by azuser1077_mml.local@iihtl.onmicrosoft.com at 14/02/2024, 15:25:52 on azuser1077_mml.local's Cluster

Cmd 14

```
1   newData = spark.range(0, 20)
2
3   deltaTable.alias("oldData") \
4     .merge(
5       newData.alias("newData"),
6       "oldData.id = newData.id") \
7     .whenMatchedUpdate(set = { "id": col("newData.id") }) \
8     .whenNotMatchedInsert(values = { "id": col("newData.id") }) \
9     .execute()
10
11  deltaTable.toDF().show()
```

25°C Mostly cloudy — Search — ENG IN — 9:43 PM 2/14/2024

python-delta Python ⌄ ☆

File Edit View Run Help   Last edit was 6 hours ago   New cell UI: OFF ⌄

▶ Run all   ■ Terminated ⌄   📅 Schedule   Share

```python
1    newData = spark.range(0, 20)
2
3    deltaTable.alias("oldData") \
4      .merge(
5        newData.alias("newData"),
6        "oldData.id = newData.id") \
7      .whenMatchedUpdate(set = { "id": col("newData.id") }) \
8      .whenNotMatchedInsert(values = { "id": col("newData.id") }) \
9      .execute()
10
11   deltaTable.toDF().show()
```

▶ (18) Spark Jobs

▶ ▤ newData: pyspark.sql.dataframe.DataFrame = [id: long]

```
+---+
| id|
+---+
|  5|
|  5|
|  6|
|  7|
|  7|
|  8|
|  9|
|  9|
| 10|
| 11|
| 12|
| 13|
| 14|
| 15|
| 16|
| 17|
| 18|
| 19|
```