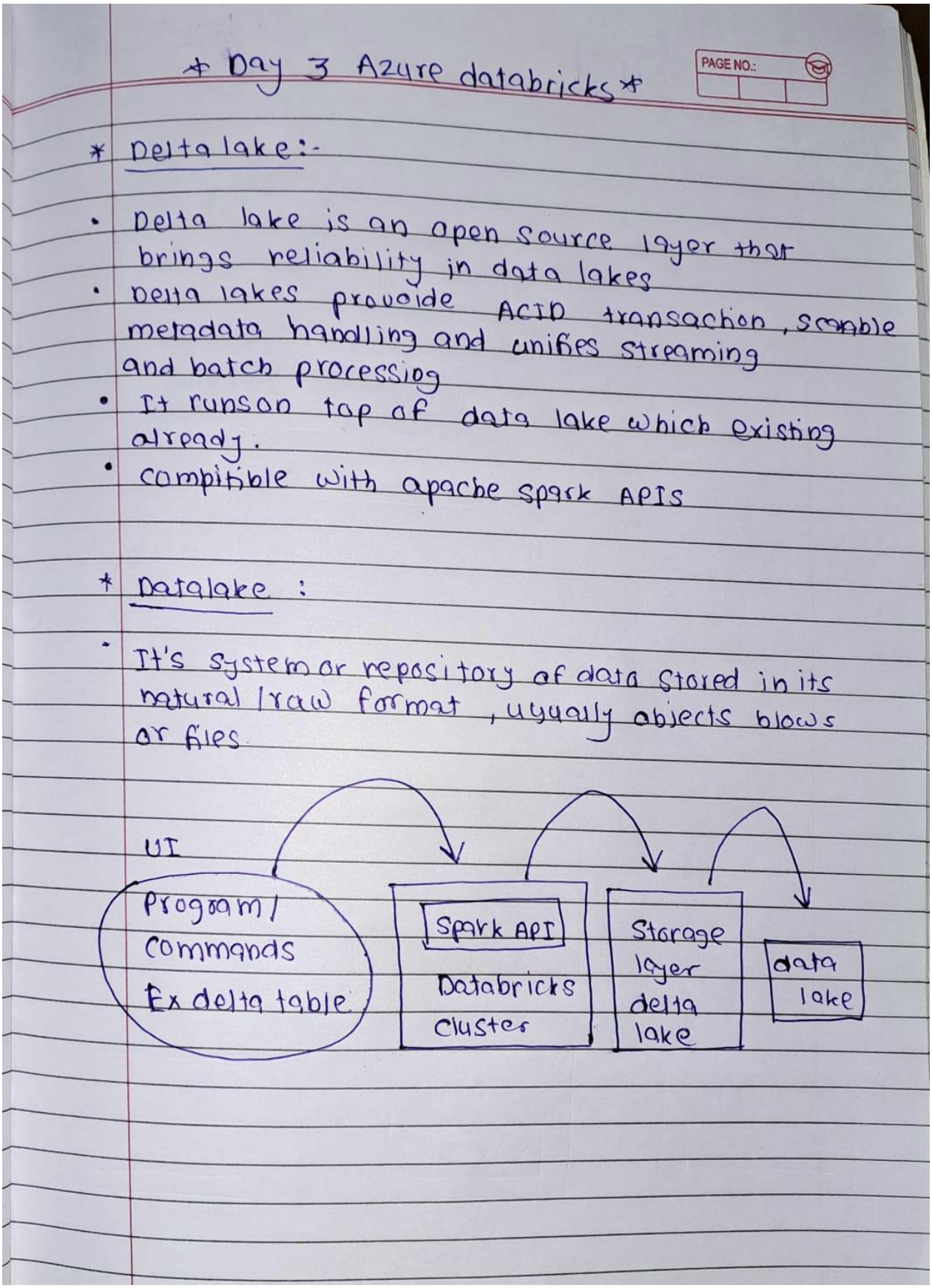
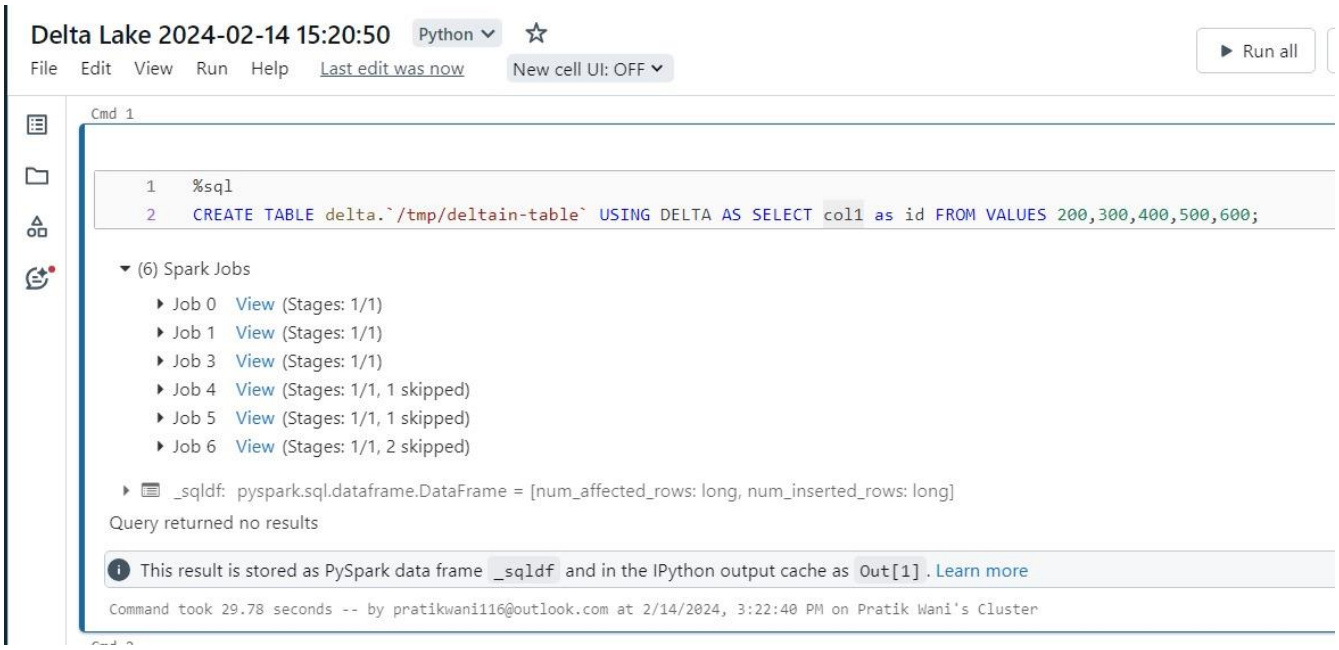
Azure Databricks Assignment 3

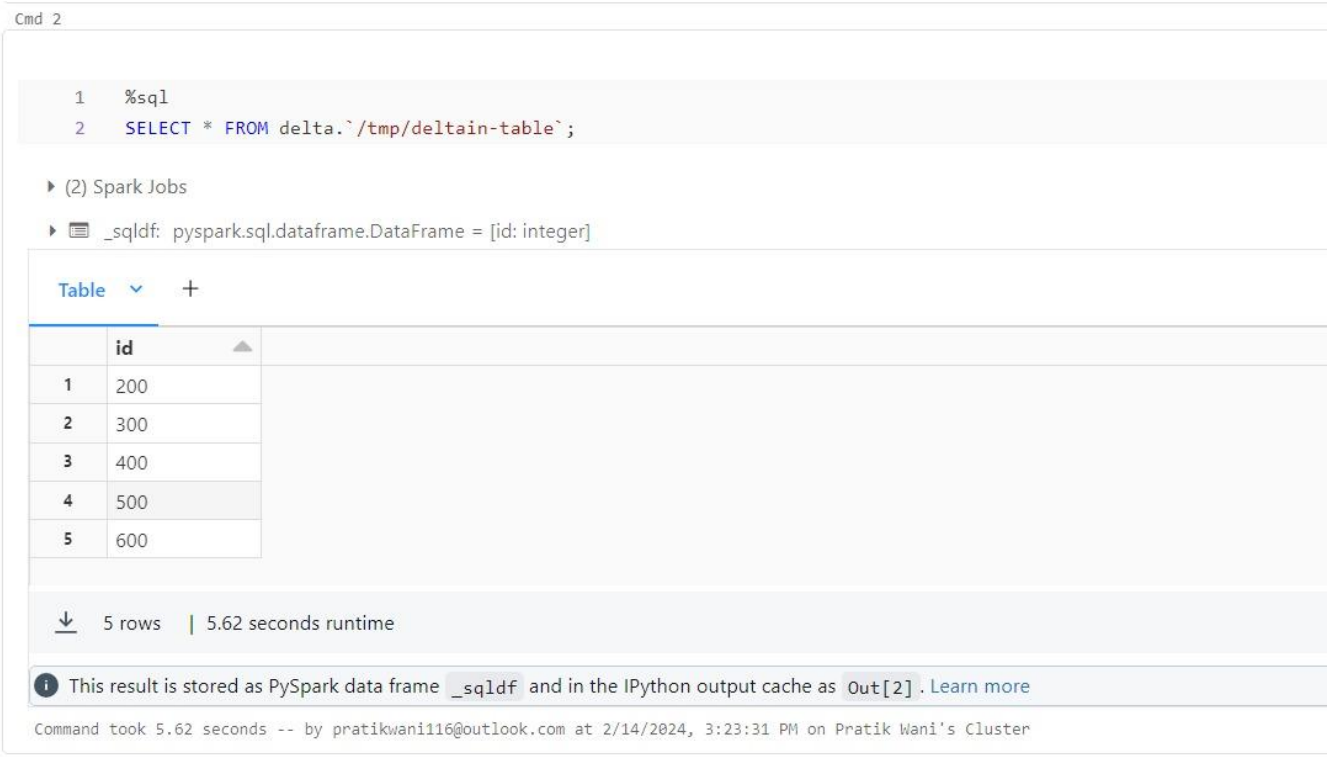
Pratik Wani

- Notes



- Delta Lake:


Delta Lake 2024-02-14 15:20:50
Python
File Edit View Run Help Last edit was now
New cell UI: OFF
Run all
Cmd 1
1 %sql
2 CREATE TABLE delta.`/tmp/deltain-table` USING DELTA AS SELECT col1 as id FROM VALUES 200,300,400,500,600;
(6) Spark Jobs
Job 0 View (Stages: 1/1)
Job 1 View (Stages: 1/1)
Job 3 View (Stages: 1/1)
Job 4 View (Stages: 1/1, 1 skipped)
Job 5 View (Stages: 1/1, 1 skipped)
Job 6 View (Stages: 1/1, 2 skipped)
_sqldf: pyspark.sql.dataframe.DataFrame = [num_affected_rows: long, num_inserted_rows: long]
Query returned no results
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[1]. Learn more
Command took 29.78 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:22:40 PM on Pratik Wani's Cluster


Cmd 2
1 %sql
2 SELECT * FROM delta.`/tmp/deltain-table`;
(2) Spark Jobs
_sqldf: pyspark.sql.dataframe.DataFrame = [id: integer]
Table
id
1 200
2 300
3 400
4 500
5 600
5 rows | 5.62 seconds runtime
This result is stored as PySpark data frame _sqldf and in the IPython output cache as Out[2]. Learn more
Command took 5.62 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:23:31 PM on Pratik Wani's Cluster

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

▸ (1) Spark Jobs

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

```
+---+
| id|
+---+
|200|
|300|
|400|
|500|
|600|
+---+
```

Command took 0.40 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:35:47 PM on Pratik Wani's Cluster

Cmd 4

```
# 2nd example

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 3.71 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:32:11 PM on Pratik Wani's Cluster

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

▸ (3) Spark Jobs

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

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

Command took 0.92 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:32:25 PM on Pratik Wani's Cluster

Cmd 6

```
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 2.59 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:34:08 PM on Pratik Wani's Cluster

Cmd 7

```
1  data.show()
2
```

▸ (1) Spark Jobs

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

Command took 0.35 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:34:19 PM on Pratik Wani's Cluster

Cmd 8

```
1  # Update Without overwrite
2
3  from delta.tables import *
4  from pyspark.sql.functions import *
```

Command took 0.11 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:37:02 PM on Pratik Wani's Cluster

Cmd 9

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

Command took 0.18 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:41:17 PM on Pratik Wani's Cluster

Cmd 10

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

▸ (3) Spark Jobs

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

Command took 0.83 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:42:02 PM on Pratik Wani's Cluster

Cmd 11

Cmd 11

```
# update values

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

```
1
```

▸ (2) Spark Jobs

```
+---+
| id|
+---+
|  5|
|106|
|  7|
|108|
|  9|
+---+
```

Command took 0.72 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:42:28 PM on Pratik Wani's Cluster

Cmd 13

```
# delete values

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

▸ (10) Spark Jobs

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

Command took 4.00 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:45:32 PM on Pratik Wani's Cluster

Cmd 14

```
# merge
newData = spark.range(0, 20)
newData.show()
```

▸ (1) Spark Jobs

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

```
| 0|
| 1|
| 2|
| 3|
| 4|
| 5|
| 6|
| 7|
| 8|
| 9|
| 10|
| 11|
| 12|
| 13|
| 14|
| 15|
| 16|
| 17|
| 18|
| 19|
+---+
```

Command took 0.46 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:46:09 PM on Pratik Wani's Cluster

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

▸ (9) Spark Jobs

Command took 5.08 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:48:49 PM on Pratik Wani's Cluster

Cmd 16

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

▸ (3) Spark Jobs

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

Cmd 17

```
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

▸ 3dfdeb69-8c56-47ea-bcad-27a224785b8c *Last updated: 7 minutes ago*

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

Command complete

Cmd 18

```
1  stream.stop()
```

Command took 0.11 seconds -- by pratikwani116@outlook.com at 2/14/2024, 3:53:50 PM on Pratik Wani's Cluster

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