

Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 14-02-24

DAY 18 – Azure Databricks-Delta tables

Creating a session and importing a table to delta table

```
1 from pyspark.sql import SparkSession
2
3 spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
4 df = spark.read.option("inferSchema", "true").option("header", "true").csv("/FileStore/tables/emp.txt")
5
6
7 df.write.format("delta").mode("overwrite").save("/FileStore/tables/delta_train/")
8
9
10
```

▶ (8) Spark Jobs

▶  df: pyspark.sql.dataframe.DataFrame = [123234877: integer, Michael: string ... 2 more fields]

Display delta table

```
1 display(dbutils.fs.ls("/FileStore/tables/delta_train/"))
```

▶ (3) Spark Jobs

Table ▾ +

	path	name
1	dbfs:/FileStore/tables/delta_train/_delta_log/	_delta_log/
2	dbfs:/FileStore/tables/delta_train/part-00000-25119bca-3aa9-4418-b90e-53b8f5491705-c000.snappy.parquet	part-00000-25119bca-3aa9-4418-

⬇ 2 rows | 4.51 seconds runtime

Command took 4.51 seconds -- by kakilesh123@gmail.com at 2/14/2024, 12:02:34 PM on My Cluster

Create the database

```
1 spark.sql("CREATE DATABASE IF NOT EXISTS delta_training")
2
```

Out[6]: DataFrame[]

Command took 1.40 seconds -- by kakilesh123@gmail.com at 2/14/2024, 12:04:36 PM on My Cluster

Execute a spark SQL

```
1
2  ddl_query = ""
3  CREATE TABLE IF NOT EXISTS delta_training.emp_file
4  USING DELTA
5  LOCATION '/FileStore/tables/delta_train/'
6  ""
7
8  spark.sql(ddl_query)
```

Out[8]: DataFrame[]

Command took 1.87 seconds -- by kakilesh123@gmail.com at 2/14/2024, 12:05:50 PM on My Cluster

View the table

```
1  spark.catalog.setCurrentDatabase("delta_training")
2  |
3  tables = spark.catalog.listTables()
4  for table in tables:
5  |     print(table)
```

► (2) Spark Jobs

Table(name='emp_file', catalog='spark_catalog', namespace=['delta_training'], description=None, tableType='EXTERNAL', isTemporary=False)

Command took 0.46 seconds -- by kakilesh123@gmail.com at 2/14/2024, 12:19:31 PM on My Cluster

Viewing the table

```
1 spark.sql("select * from delta_training.emp_file").show()
```

► (2) Spark Jobs

```
+-----+-----+-----+----+
|123234877| Michael| Rogers| 14|
+-----+-----+-----+----+
|152934485| Anand|Manikutty| 14|
|222364883| Carol| Smith| 37|
|326587417| Joe| Stevens| 37|
|332154719| Mary-Anne| Foster| 14|
|332569843| George| ODonnell| 77|
|546523478| John| Doe| 59|
|631231482| David| Smith| 77|
|654873219| Zacary| Efron| 59|
|745685214| Eric| Goldsmith| 59|
|845657245| Elizabeth| Doe| 14|
|845657246| Kumar| Swamy| 14|
+-----+-----+-----+----+
```

Command took 6.77 seconds -- by kakilesh123@gmail.com at 2/14/2024, 12:07:10 PM on My Cluster

Building a session

```
1 import pyspark
2 from delta import *
3
4 builder = pyspark.sql.Session.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 2.22 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:18:11 PM on hexaCluster

Creating a delta table

```
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 36.90 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:18:23 PM on hexaCluster

Load the value and displaying it

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

▶ (3) Spark Jobs

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


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

Command took 9.53 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:19:27 PM on hexaCluster

Overwriting on the table

```
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 7.27 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:26:32 PM on hexaCluster

Display the delta table

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

▶ (3) Spark Jobs

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

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

Command took 2.73 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:26:57 PM on hexaCluster

Reverting the table by using versionAsOf

```
1 df = spark.read.format("delta").option("versionAsOf", 0).load("/tmp/delta-table")
2 df.show()
```

▶ (3) Spark Jobs

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

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

Command took 2.84 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:27:18 PM on hexaCluster

Streaming the data



Update the delta table using a condition



Viewing the table

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

▶ (5) Spark Jobs

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

```
| 73 |
| 213 |
| 243 |
| 35 |
| 217 |
| 255 |
| 157 |
| 15 |
| 199 |
| 135 |
| 11 |
| 281 |
| 149 |
| 249 |
```

Deleting the rows with certain condition

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

▶ (7) Spark Jobs

Command took 16.07 seconds -- by kakilesh123@gmail.com at 2/14/2024, 3:35:53 PM on hexaCluster

Viewing the table

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

► (5) Spark Jobs


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

```
| 73|
|213|
|243|
| 35|
|217|
|255|
|157|
| 15|
|199|
|135|
| 11|
|281|
.
```

Merging the delta table

```
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()
```

► (26) Spark Jobs

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

```
+---+
| id|
+---+
| 21|
| 23|
| 25|
| 27|
| 29|
| 31|
| 33|
| 35|
| 37|
| 39|
| 41|
```