**Analyze with Apache Spark**

**Create a serverless Apache Spark pool**

1. In Synapse Studio, on the left-side pane, select **Manage** > **Apache Spark pools**.
2. Select **New**
3. For **Apache Spark pool name** enter **Spark1**.
4. For **Node size** enter **Small**.
5. For **Number of nodes** Set the minimum to 3 and the maximum to 3
6. Select **Review + create** > **Create**. Your Apache Spark pool will be ready in a few seconds.

**Understanding serverless Apache Spark pools**

A serverless Spark pool is a way of indicating how a user wants to work with Spark. When you start using a pool, a Spark session is created if needed. The pool controls how many Spark resources will be used by that session and how long the session will last before it automatically pauses. You pay for spark resources used during that session not for the pool itself. In this way a Spark pool lets you work with Spark, without having to worry managing clusters. This is similar to how a serverless SQL pool works.

**Analyze NYC Taxi data with a Spark pool**

**Note**

Make sure you have **placed the sample data in the primary storage account**.

1. In Synapse Studio, go to the **Develop** hub.
2. Create a new notebook.
3. Create a new code cell and paste the following code in that cell:

pyCopy

%%pyspark

df = spark.read.load('abfss://demo@ckctestdbmount1.dfs.core.windows.net/demo/NYCTripSmall.parquet', format='parquet')

display(df.limit(10))

1. Modify the load URI, so it references the sample file in your storage account according to the abfss URI scheme.
2. In the notebook, in the **Attach to** menu, choose the **Spark1** serverless Spark pool that we created earlier.
3. Select **Run** on the cell. Synapse will start a new Spark session to run this cell if needed. If a new Spark session is needed, initially it will take about two seconds to be created.
4. If you just want to see the schema of the dataframe run a cell with the following code:

%%pyspark

df.printSchema()

**Load the NYC Taxi data into the Spark nyctaxi database**

Data is available via the dataframe named **df**. Load it into a Spark database named **nyctaxi**.

1. Add a new code cell to the notebook, and then enter the following code:

%%pyspark

spark.sql("CREATE DATABASE IF NOT EXISTS nyctaxi")

df.write.mode("overwrite").saveAsTable("nyctaxi.trip")

**Analyze the NYC Taxi data using Spark and notebooks**

1. Create a new code cell and enter the following code.

%%pyspark

df = spark.sql("SELECT \* FROM nyctaxi.trip")

display(df)

1. Run the cell to show the NYC Taxi data we loaded into the **nyctaxi** Spark database.
2. Create a new code cell and enter the following code. We will analyze this data and save the results into a table called **nyctaxi.passengercountstats**.

%%pyspark

df = spark.sql("""

SELECT PassengerCount,

SUM(TripDistanceMiles) as SumTripDistance,

AVG(TripDistanceMiles) as AvgTripDistance

FROM nyctaxi.trip

WHERE TripDistanceMiles > 0 AND PassengerCount > 0

GROUP BY PassengerCount

ORDER BY PassengerCount

""")

display(df)

df.write.saveAsTable("nyctaxi.passengercountstats")