Here is the Scala code that you will need to run to load the 2019 NYCTaxi data into a data frame.

import org.apache.spark.sql.functions.\_

val Data = "/databricks-datasets/nyctaxi/tripdata/yellow/yellow\_tripdata\_2019-\*"

val SchemaDF = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("/databricks-datasets/nyctaxi/tripdata/yellow/yellow\_tripdata\_2019-02.csv.gz")

val df = spark.read.format("csv").option("header", "true").schema(SchemaDF.schema).load(Data)

val nyctaxiDF = df

.withColumn("Year", year(col("tpep\_pickup\_datetime")))

.withColumn("Year\_Month", date\_format(col("tpep\_pickup\_datetime"),"yyyyMM"))

.withColumn("Year\_Month\_Day", date\_format(col("tpep\_pickup\_datetime"),"yyyyMMdd"))

Here is the code that you will need to run to persist your NYCTaxi data in delta format to your ADLS gen2 account.

val Factnyctaxi = nyctaxiDF.write

.format("delta")

.mode("overwrite")

.partitionBy(("Year"), ("Year\_Month"), ("Year\_Month\_Day"))

.save("/mnt/raw/delta/Factnyctaxi")

Here is the SQL code that you will need to run to create a delta Spark SQL table.

%sql

CREATE TABLE Factnyctaxi

USING DELTA

LOCATION '/mnt/raw/delta/Factnyctaxi'

Here is the code that you will need to run to infer the file’s schema and load the file to a data frame.

val dimedateDF = spark.read.format("csv")

.option("header", "true")

.option("inferSchema", "true")

.load("/mnt/raw/dimdates.csv")

Once the file is loaded to a data frame, run the following code to persist the data as Delta format to your ADLS gen2 account.

val DimDate = dimedateDF.write

.format("delta")

.mode("overwrite")

.save("/mnt/raw/delta/DimDate")

Also, execute the following SQL code to create a delta table called DimDate which will be joined to your Factnyctaxi table on a specified partition key.

%sql

CREATE TABLE DimDate

USING DELTA

LOCATION '/mnt/raw/delta/DimDate'

Here is the SQL query that you will need to run to generate the results shown in the figure above.

%sql

SELECT \* FROM Factnyctaxi F

INNER JOIN DimDate D

ON F.Year\_Month\_Day = D.DateNum

Here is the SQL query that you will need to run to generate the results shown in the figure above.

%sql

SELECT \* FROM Factnyctaxi F

INNER JOIN DimDate D

ON F.Year\_Month\_Day = D.DateNum

WHERE D.Calendar\_Quarter = 'Qtr 1' AND D.DayName = 'Friday'