Here is the code that has been executed in the notebook illustration shown in Figure 22-7.

# Read NYC yellow cab data from Azure Open Datasets

from azureml.opendatasets import NycTlcYellow

from datetime import datetime

from dateutil import parser

end\_date = parser.parse('2018-05-08 00:00:00')

start\_date = parser.parse('2018-05-01 00:00:00')

nyc\_tlc = NycTlcYellow(start\_date=start\_date, end\_date=end\_date)

nyc\_tlc\_df = nyc\_tlc.to\_spark\_dataframe()

Here is the code that has been executed in the notebook illustration shown in Figure 22-8.

nyc\_tlc\_df.createOrReplaceTempView('nyc\_tlc\_df')

nyc\_tlc\_df.write.csv('nyc\_tlc\_df\_csv', mode='overwrite')

nyc\_tlc\_df.write.parquet('nyc\_tlc\_df\_parquet', mode='overwrite')

Here is the code that has been executed in the notebook illustration shown in Figure 22-11.

%%sql

SELECT \* FROM myc\_tlc\_df

Here is the code that has been executed in the notebook illustration shown in Figure 22-12.

from pyspark.sql import functions as F

df\_nyc = nyc\_tlc\_df.groupBy("passengerCount").agg(F.avg('tripDistance').alias('AvgTripDistance'), F.sum('tripDistance').alias('SumTripDistance'))

display(df\_nyc)

Here is the code that has been executed in the notebook illustration shown in Figure 22-13.

import matplotlib.pyplot

import seaborn

seaborn.set(style = "whitegrid")

pdf\_nyc = df\_nyc.toPandas()

seaborn.lineplot(x="passengerCount", y="SumTripDistance" , data = pdf\_nyc)

seaborn.lineplot(x="passengerCount", y="AvgTripDistance" , data = pdf\_nyc)

matplotlib.pyplot.show()

Here is the code that has been executed in the notebook illustration shown in Figure 22-15.

SELECT TOP 100 \* FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=2019/puMonth=\*/\*.parquet',

        FORMAT='PARQUET'

    ) AS [nyc];

Here is the code that has been executed in the notebook illustration shown in Figure 22-16.

SELECT

    YEAR(tpepPickupDateTime) AS current\_year,

    COUNT(\*) AS rides\_per\_year

FROM

    OPENROWSET(

        BULK 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/puYear=\*/puMonth=\*/\*.parquet',

        FORMAT='PARQUET'

    ) AS [nyc]

WHERE nyc.filepath(1) >= '2014' AND nyc.filepath(1) <= '2019'

GROUP BY YEAR(tpepPickupDateTime)

ORDER BY 1 ASC;

Here is the code that has been executed in the notebook illustration shown in Figure 22-18.

/\* Note: this script is filtered on a specific month. You can modify the location to read the entire dataset. \*/

IF NOT EXISTS (SELECT \* FROM sys.external\_file\_formats WHERE name = 'SynapseParquetFormat')

    CREATE EXTERNAL FILE FORMAT [SynapseParquetFormat]

    WITH ( FORMAT\_TYPE = PARQUET)

GO

IF NOT EXISTS (SELECT \* FROM sys.external\_data\_sources WHERE name = 'nyctlc\_azureopendatastorage\_blob\_core\_windows\_net')

    CREATE EXTERNAL DATA SOURCE [nyctlc\_azureopendatastorage\_blob\_core\_windows\_net]

    WITH (

        LOCATION = 'wasbs://nyctlc@azureopendatastorage.blob.core.windows.net',

        TYPE     = HADOOP

    )

GO

CREATE EXTERNAL TABLE nyc\_tlc\_yellow\_trip\_ext (

    [vendorID] varchar(8000),

    [tpepPickupDateTime] datetime2(7),

    [tpepDropoffDateTime] datetime2(7),

    [passengerCount] int,

    [tripDistance] float,

    [puLocationId] varchar(8000),

    [doLocationId] varchar(8000),

    [startLon] float,

    [startLat] float,

    [endLon] float,

    [endLat] float,

    [rateCodeId] int,

    [storeAndFwdFlag] varchar(8000),

    [paymentType] varchar(8000),

    [fareAmount] float,

    [extra] float,

    [mtaTax] float,

    [improvementSurcharge] varchar(8000),

    [tipAmount] float,

    [tollsAmount] float,

    [totalAmount] float

    )

    WITH (

    LOCATION = 'yellow/puYear=2014/puMonth=3/',

    -- LOCATION = 'yellow'

    DATA\_SOURCE = [nyctlc\_azureopendatastorage\_blob\_core\_windows\_net],

    FILE\_FORMAT = [SynapseParquetFormat],

    REJECT\_TYPE = VALUE,

    REJECT\_VALUE = 0

    )

GO

SELECT TOP 100 \* FROM nyc\_tlc\_yellow\_trip\_ext

GO

Here is the code that has been executed in the notebook illustration shown in Figure 22-19.

SELECT TOP 100 \* FROM nyc\_tlc\_yellow\_trip\_ext

GO