Load data in Spark notebooks

Note:

You are not required to complete the processes, tasks, activities, or steps presented in this example. The various samples provided are for illustrative purposes only and it's likely that if you try this out you will encounter issues in your system.

There are several options available for ingesting data into a notebook. Currently, it is possible to load data from an Azure Storage Account, and an Azure Synapse Analytics dedicated SQL pool.

Some examples for reading data in a notebook are as follows:

- Read a CSV from Azure Data Lake Store Gen2 as an Apache Spark DataFrame
- Read a CSV from Azure Storage Account as an Apache Spark DataFrame
- Read data from the primary storage account

Example 1: Read a CSV file from an Azure Data Lake Store Gen2 store as an Apache Spark DataFrame.

The following code is used to read a CSV file from an Azure Data Lake Store Gen2 store as an Apache Spark DataFrame.

```
1
2
3
4
5
6
7
8
9
10
11
12
13
```

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
account_name = "Your account name"
container_name = "Your container name"
relative_path = "Your path"
adls_path = 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name, account_na
me, relative_path)

spark.conf.set("fs.azure.account.auth.type.%s.dfs.core.windows.net" %account_name
, "SharedKey")
```

```
spark.conf.set("fs.azure.account.key.%s.dfs.core.windows.net" %account_name
,"Your ADLS Gen2 Primary Key")

df1 = spark.read.option('header', 'true') \
                .option('delimiter', ',') \
                .csv(adls_path + '/Testfile.csv')
```

There are parameter name values that you need to replace in the above code to ensure that it works, including:

- **account_name:** Replace "Your account name" with the storage account name you wish to use
- **container_name:** Replace "Your container name" with the storage container you wish to use
- **relative_path:** Replace "Your path" with the relative path of where the file is stored
- **adls_path:** The adls_path is defined by passing through the above parameters.

Example 2: Read a CSV file from Azure Storage Account as a Spark DataFrame.

The following code is used to read a CSV file from Azure Storage Account as an Apache Spark DataFrame.

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *

blob_account_name = "Your blob account name"
blob_container_name = "Your blob container name"
blob_relative_path = "Your blob relative path"
blob_sas_token = "Your blob sas token"
```

```
wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blo
b_account_name, blob_relative_path)
spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name,
blob_account_name), blob_sas_token)

df = spark.read.option("header", "true") \
            .option("delimiter","|") \
            .schema(schema) \
            .csv(wasbs_path)
```

There are parameter name values that you need to replace in the above code to ensure that it works, including:

- **blob_account_name:** Replace "Your blob account name" with the name of your blob account.
- **blob_container_name:** Replace "Your blob container" with the name of the blob container the file is in.
- **blob_relative_path:** Replace "Your blob relative path" with the name of the relative path pointing to the csv you want to read.
- **blob_sas_token:** Replace "Your blob sas token" with the blob SAS key.

Example 3: Read data from the primary storage account

The third possibility is to read data from the primary storage account using the Data tab in the Azure Synapse Studio environment.

Right-click on a file and select **New notebook** to view a new notebook with the data generated.