“Azure - Databricks - Cheat Sheet”

databricks

12 3 2022

Reading Data

```
fileName = "dbfs:/mnt/training/wikipedia/clickstream/2015_02_clickstream.tsv"
csvSchema = StructType([
  StructField("prev_id", IntegerType(), False),
  StructField("curr_id", IntegerType(), False),
  StructField("n", IntegerType(), False),
  StructField("prev_title", StringType(), False),
  StructField("curr_title", StringType(), False),
  StructField("type", StringType(), False)
])

testDF = (spark.read          #The DataFrameReader
  .option('header', 'true')   #Ignore line #1 - it's a header
  .option('sep', "\t")        #Use tab delimiter (default is comma-separator)
  .schema(csvSchema)          #Use the specified schema
  .csv(fileName)              #Creates a DataFrame from CSV after reading in the file
)

#Display data
testDF.printSchema()
```

DataFrames vs SQL & Temporary Views

```
dataDF.createOrReplaceTempView("name_of_db")

#Use simple sql

%sql
SELECT * FROM pagecounts
```

Convert from SQL back to a `DataFrame`

```
tableDF = spark.sql("SELECT DISTINCT project FROM pagecounts ORDER BY project")
display(tableDF)
```

Exercise

```
(source, sasEntity, sasToken) = getAzureDataSource()
spark.conf.set(sasEntity, sasToken)

path = source + "/wikipedia/pagecounts/staging_parquet_en_only_clean/"

# 1. Define data frame
#
df = (spark                   # Our SparkSession & Entry Point
  .read                       # Our DataFrameReader
  .parquet(path)              # Read in the parquet files
  .select("article")          # Reduce the columns to just the one
```

```
  .distinct()                    # Produce a unique set of values
)
totalArticles = df.count() # Identify the total number of records remaining.

print("Distinct Articles: {0:,}".format(totalArticles))
```