“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()
```