PySpark Assignments

(By: Y. Kanakaraju, Created: May 2022, For: CTS Academy)

Spark Core – RDD

1. From the given Car details dataset, compute the 'Average Weight' of 'American Cars' for each 'Make'. Do not use 'groupBy' transformation The output should look like: (ford, 3540), (buick, 2800) etc.
    - Dataset: **cars.tsv**

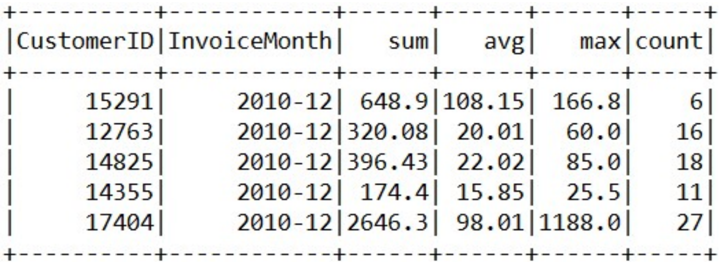
Spark SQL – DataFrames

Solve the following two assignments (2 and 3) using the dataset: **online-retail-dataset.csv**

The dataset may be downloaded from the following URL: https://archive.ics.uci.edu/ml/machine-learning-databases/00352/

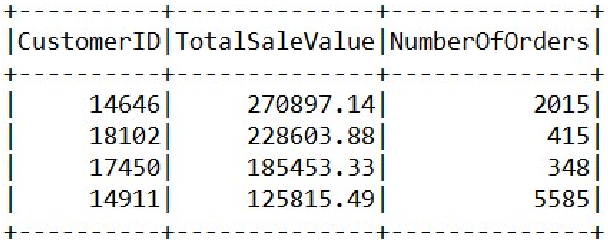
Save the excel file as CSV file.

2. Compute the total SUM, AVERAGE, MAX and COUNT of the **SaleValue** for **each customer** for **each month**.
    a. Filter all the customers with NULL value
    b. SaleValue is derived as UnitPrice * Quantity
    c. Create the month as an additional derived column from InvoiceDate in YYYY-MM format (ex: 2022-01)
    d. Shown below is a sample of the output:



```
+----------+------------+------+------+------+-----+
|CustomerID|InvoiceMonth|   sum|   avg|   max|count|
+----------+------------+------+------+------+-----+
|     15291|     2010-12| 648.9|108.15| 166.8|    6|
|     12763|     2010-12|320.08| 20.01|  60.0|   16|
|     14825|     2010-12|396.43| 22.02|  85.0|   18|
|     14355|     2010-12| 174.4| 15.85|  25.5|   11|
|     17404|     2010-12|2646.3| 98.01|1188.0|   27|
+----------+------------+------+------+------+-----+
```

3. Find out the top 10 customers with highest SaleValue in the year 2011. Use DataFrame transformation methods only (do not use SQL).
    a. Create InvoiceYear as a derived column from InvoiceDate
    b. Filter all the customers with NULL value
    c. SaleValue is derived as UnitPrice * Quantity
    d. Arrange the data in the DESC order of SaleValue
    e. Fetch the following data: CustomerID, TotalSaleValue, InvoiceYear, NumberOfOrders
    f. Shown below is a sample of the output:



```
+----------+--------------+--------------+
|CustomerID|TotalSaleValue|NumberOfOrders|
+----------+--------------+--------------+
|     14646|     270897.14|          2015|
|     18102|     228603.88|           415|
|     17450|     185453.33|           348|
|     14911|     125815.49|          5585|
+----------+--------------+--------------+
```

Structured Streaming

Solve the assignments 4 and 5 using Spark Structured Streaming API (do not use DStreams API).

4. Create a streaming file format conversion pipeline using File streams to convert CSV files into Parquet files in real time.
   Streaming Source: **CSV** (File Source), Sink: **Parquet** (File Sink)

    a. Create a directory called “source_csv_files” in your home path.
    b. Create a directory called “csv_files” in your home path
    c. Create a directory called “parquet_files” in your home path
    d. Create a 5 sample CSV files with the following columns: id, name, age (id INT, name STRING, age INT) with each file containing 4 or 5 rows in “source_csv_files” directory.
    e. As you copy the CSV files from “source_csv_files” to “csv_files”, your application should listen to these files in real time and write them as parquet files in “parquet_files” directory.

5. Create a simple data flow to ingest streaming data from a **Rate source** at a rate of 5 records per second into a MySQL table using **ForEachBatch sink**. Streaming Source: **Rate**, Sink: **ForEachBatch**
    a. Create an input stream from a rate source to create a stream of 5 records per seconds
    b. Rename the columns of the input stream as 'ts' and 'message'
    c. Write the stream into MySQL table with two columns – ts (varchar) and message (varchar) using forEachBatch sink.

Weightage:

- Assignment 1: 15% (RDD API)
- Assignment 2: 15% (Spark SQL)
- Assignment 3: 20% (Spark SQL)
- Assignment 4: 25% (Structured Streaming)
- Assignment 5: 25% (Structured Streaming)

Assignment Submission Guidelines

- Please submit all the solutions in **a single text file created using Notepad**.
- Clearly mention your Associate ID the dates of the training batch you attended towards the top of the submitted file.
- Mention the assignment number followed the by source-code. Simply put all your source-code in text format.
- Separate each assignment with a horizontal line.
- No need to show/print the output.
- Even if you practiced on Jupyter Notebook or Databricks, still submit the code in a notepad file only. Just copy and paste all the code in the text file.
- Do not submit notebook files (.pynb files), word documents and image files.

Sample submission format (for your understanding)

Associate ID: 123456
Dates: PySpark from 01-May-2022 to 10-May-2022

Assignment 1:

<Paste the source code here>

Assignment 2

<Paste the source code here>

Assignment 3

<Paste the source code here>

Assignment 4

<Paste the source code here>

Assignment 5

<Paste the source code here>