**PySpark Assignments**

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

**Spark Core – RDD**

1. From the given dataset of **server\_log.tsv**, find out how many times each type of error occurred. Consider only “error” messages and group the errors by the type. (Ex: how many php errors, how many mysql errors etc.)

Dataset: **server\_log.tsv**

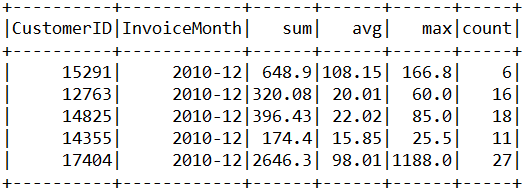
**Spark SQL – DataFrames**

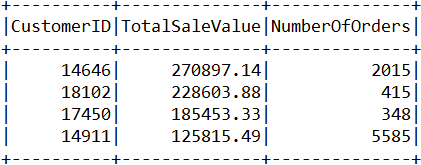
Solve the following two assignments (2 and 3) using the dataset:   
**online-retail-dataset.csv**

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.

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



1. Find out the top 10 customers with highest SaleValue in the year 2011. Use DataFrame transformation methods only (do not use SQL).
   1. Data required: CustomerID, TotalSaleValue, NumberOfOrders
   2. Create InvoiceYear as a derived column from InvoiceDate
   3. Filter all the customers with NULL value
   4. SaleValue is derived as UnitPrice \* Quantity
   5. Arrange the data in the descending order of SaleValue
   6. Shown below is a sample of the output:  
      

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

1. 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)
   1. Create a directory called “source\_csv\_files” in your home path.
   2. Create a directory called “csv\_files” in your home path
   3. Create a directory called “parquet\_files” in your home path
   4. 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.
   5. 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.
2. 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**
   1. Create an input stream from a rate source to create a stream of 5 records per seconds
   2. Rename the columns of the input stream as ‘ts’ and ‘message’
   3. 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-Nov-2022 to 10-Nov-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>