**PySpark Assignments**

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

**Spark Core – RDD**

1. From the given **Baby\_Names.csv** dataset, get the top 10 most frequently occurred **First Name** by count along with average age from year 2000 to 2010 (both years inclusive). Arrange the data in the descending order of count (i.e. frequency of occurrence). The output should have the following fields: LastName, Count, and AverageAge.   
     
   Sample Output:   
   LOGAN, 307, 25.6

GAVIN, 210, 32.1

* Dataset: **Baby\_Names.csv**

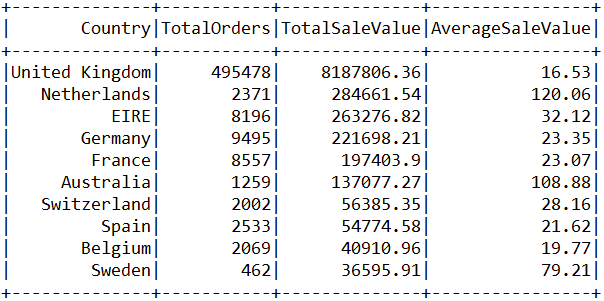
**Spark SQL – DataFrames**

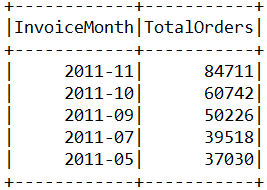
Solve the following two assignments (2 and 3) using the dataset:   
**online-retail-dataset.csv**

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. Show the **top 10 countries with highest TotalSaleValue** along with other details. Here, TotalSaleValue is the sum of all SaleValue in each country. Use DataFrame API methods only (do not use SQL queries).
   1. Data required: Country, TotalOrders, TotalSaleValue and AverageSaleValue.
   2. Filter all the customers with NULL value
   3. SaleValue is derived as UnitPrice \* Quantity
   4. Round the values to 2 decimal places only.
   5. Sample Output:



1. Find out the **top five months with highest number of orders** by count in the year 2011.
   1. Data required: InvoiceMonth, TotalOrders
   2. Arrange the data in the descending order of TotalOrders.
   3. Sample Output:  
      

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

1. Create a real time data pipeline using File streams to convert JSON files into CSV files in real time. The CSV files should have only those rows that have age > 21.  
   Streaming Source: **JSON** (File Source), Sink: **CSV** (File Sink)
   1. Create a directory called “source\_json\_files” in your home path.
   2. Create a directory called “json\_files” in your home path
   3. Create a directory called “csv\_files” in your home path
   4. Create 5 sample JSON files with the following columns: id, name, age (age should be an INT) with each file containing 4 or 5 JSON objects in “json\_files” directory.
   5. As you copy the JSON files from “source\_json\_files” to “json\_files”, your application should listen to these files in real time, get only those rows with age > 21 and write them as CSV files in “csv\_files” directory.
2. Create a streaming ‘word count’ program to create real time **CSV files** reading the data (of lines of text) from a **socket input stream**.  
   Streaming Source: **Socket** (localhost:9999), Sink: **CSV** (File Sink)
   1. Ingest data into socket using the netcat linux utility (nc –lk 9999)
   2. As you write lines of text to the socket, the CSV files have to be generated in a designated directory (directory name: csv\_output) with the word-counts of the data read from the socket.
   3. Use “Update” output mode.
   4. Use a trigger interval of 5 seconds.

**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>