CSEE5590-0001/490-0003: Big Data Programming

Lesson Plan # 10

https://forms.gle/xMAmr3zATrtMG5cX7

Lesson Title: *Data Frame and SQL*

Lesson Description: *Distributed Collection of Data*

Lesson Overview:

- Data frames
- Construction of Data Frames
- SparkSQL
- Transformation
- Laziness
- Actions
- Basic Commands on Data frames
- Basic commands of SQL on Data frames

In Class Exercise

Dataset: https://umkc.box.com/s/tg08jqi9circsjpycwxoujnux85kbxw6

DataFrames & SQL in Scala/Pyspark
Consider the dataset attached above:

References:

https://stackoverflow.com/questions/51689460/select-specific-columns-from-spark-dataframe

https://jaceklaskowski.gitbooks.io/mastering-spark-sql/spark-sql-aggregate-functions.html

How to read CSV into dataframe:

https://stackoverflow.com/questions/29704333/spark-load-csv-file-as-dataframe

Part – 1

1. Import the dataset and create data frames directly on import.
2. Save data to file.
3. Check for Duplicate records in the dataset.
4. Apply Union operation on the dataset and order the output by Country Name alphabetically.
5. Use Groupby Query based on treatment.

Part – 2

1. Apply the basic queries related to Joins and aggregate functions (at least 2)
2. Write a query to fetch 13th Row in the dataset.

Part –3:(bonus)

1. Write a parse Line method to split the comma-delimited row and create a Data frame.
2. Apply Covariance on the data frame and explain the understanding.
3. Apply Correlation on the data frame and explain the understanding.

ICP Submission Guidelines:

1. ICP Submission is individual however, it can be completed as a Team during session.
2. If completed, should be presented to TA or Instructor before the completion of the class
3. Submission after the deadline is considered as late submission. (Check the late submission policy in the syllabus)

4. ICP Code with brief explanation should be pushed to GitHub.
5. Submit your screenshots as well to GitHub and documentation. The screenshot should have both the code and the output.
6. Submit a demo video 2-3 min showing your assignment with a voice over explaining your work if you are unable to complete ICP within the deadline due to genuine reason.
7. Provide the video submission link through the GitHub and submission form https://forms.gle/xMAmr3zATrtMG5cX7

Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: *https://catalog.umkc.edu/special-notices/academic-honesty/*