Homework - Big Data Computing With Apache Spark

In this assignment, you'll find a dataset, upload it to your S3 bucket, and use Spark to analyze it. Of course, you are free to bring your own dataset. If you can't find a dataset you like, you can grab one we've collected for you.

You all must work on different datasets. Regardless if you bring your own dataset or the one we provided, indicate it in this spreadsheet to make sure no one will use the same dataset. Dataset selection works on a first-come, first-served basis.

The task

Flawless delivery of the requirements in the next section will take you to a grade of B+, A-, or A, respectively.

For a B+ grade, analyze the dataset selected and make sure that you use these Spark/Databricks features:

- For every command you execute, add a markdown %md cell above the command explaining what the command does. You can keep the explanations short.
- Use dbutils (or %fs) ls to explore the dataset on S3
- Load the dataset. If your dataset comes as CSV or JSON, load it **using both** automatic schema inference and also manually specifying the schema. If your dataset is stored in Parquet or Delta, just load it in a way you'd like to.
- Create a table (or view or tempView) from the dataset and execute at least two SQL queries.
- Add at least two different visualizations (different meaning that you can't have two bar charts; you need at least two different kinds of charts)
- Apply the filter, select, limit, and orderBy (or sort) transformations on the dataset (all of these, in a meaningful way).
- Use grouping where you use at least two different aggregation functions (like avg and sum)
- Use at least two pyspark.sql.functions
- You must have at least 15 command cells in your analytics (meaning total cell count minus markdown cell count)

A-:

- Create a Widget
- Use the *groupBy.“agg”* function
- Use at least one of the DataframeNAFunctions
- Write one of your Dataframes in delta format to a folder under /tmp and use the “ls” dbutils function to list the folder containing the parquet files.
- Use the *withColumn* function
- Use the *df.createOrReplaceTempView* and the *spark.table* functions

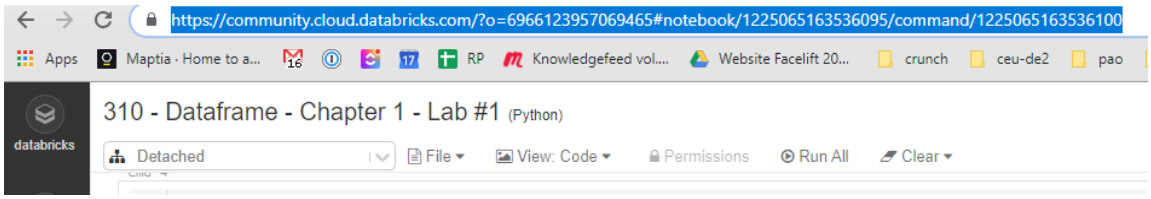
A:

- Display the number of partitions in the input Dataframe, execute a “.count()” on your dataframe, and try to match it to the Spark UI’s Dataframe/SQL graph. (remember to disable the adaptive query executor as we did in class for easier matching of actions and Spark jobs)
- Make a screenshot of that part of the graph that shows the reading of the input data, upload it to S3, make it public, and link it from a Databricks cell (or embed it as an inline image if you’d like to). Next, explain (that is, write down in markdown in the notebook) what you saw and whether the partitions of the Dataframe and metrics in the graph match up (for smaller datasets, the partitions reported in the notebook and the partitions on the UI might differ as Spark optimizes the execution!). Finally, explain your understanding in the markdown cell.

Delivery

You need to go through all the steps below for us to be able to accept your delivery:

1. Ensure that you invited tothz@ceu.edu to your Databricks workspace (Click your email in the top right corner -> Admin Settings -> Users/Manage)
2. Ensure you show your results in the Databricks Notebooks (don’t delete the resulting tables/charts/ …).
3. Copy the url of your notebook (highlighted in blue below) and send it to ceu-data@googlegroups.com



Make the subject of the email “CEU Spark Homework 1 - YOUR_STUDENT_ID”

Getting help

@ mention Zoltan Slack before 3 pm Mon-Thu. Expect up to a day to receive a response.

Deadline: Sunday, 2 Jun 2024, 23:59

Late submissions

1% of “homework completeness” points are deducted every hour between the deadline and the submission date. You can deliver any time before 8 Jun for a minimum passing score.