Assignment

Choose any datasets of your choice in the public folder of the external lab to demonstrate the following

1. Analyze the datasets chosen and come up with an example use-case.

2. Develop a Pyspark code to meet the use-case requirement.

3. Execute the Pyspark code using the spark-submit utility

4. Now, try to vary the number of executors, executor-memory, executor-cores and present your inference with relevant explanation and screenshots of the results. (Perform this by disabling the dynamic-memory allocation feature of pyspark)

5. Provide a detailed explanation with diagrams for executor memory distribution in the above example use-case considered.

6. Create the following scenarios and explain the spark execution behaviour
    a. Storage Memory is full and it is extended to the execution memory that is free.
    b. Then a need arises for more execution memory due to more jobs lined up for execution. What happens to the storage that has extended into the execution memory space.

7. Consider an example use-case on the dataset chosen previously to demonstrate when the spark engine chooses to use Hash / Sort Aggregation.

8. Demonstrate the following optimization in Spark's logical and physical execution plan
    a. Predicate Pushdown.
    b. Merging of multiple projections into one.
    c. Merging of multiple filters into one.

9. Demonstrate Schema Evolution on the dataset considered by
    a. Adding a new column
    b. Dropping a column
    c. Changing the datatype

10. Research and explore the different file formats furthermore. Depict your inferences with relevant diagrams and explanations. (Parquet, ORC, AVRO)

11. Apply the different generalised compression techniques explained in the course on the example datasets and illustrate the differences noticed.

Process to Submit the Assignment -

You need to create a Google Document consisting of answers to all the above questions. Name the Google Document as **yourname_week11_assignment**
Please upload your solution by filling the following form -
https://forms.gle/8tPzTXwvMGxBKWjm9

Top 5 answers will be selected and they will be compiled into a solution document and added to the Learning portal.