TRENDYTECH

UPLIFT YOUR CAREER!

Assignment

Week13: Apache Spark - Structured API Part-1

IMPORTANT

Self-assessment enables students to develop:

1. A sense of responsibility for their own learning and the ability & desire to continue learning,
2. Self-knowledge & capacity to assess their own performance critically & accurately, and
3. An understanding of how to apply their knowledge and abilities in different contexts.

All assignments are for self assessment. Solutions will be released on every subsequent week.Once the solution is out, evaluate yourself.

No discussions/queries allowed on assignment questions in slack channel.

***Note**: You can raise your doubts in the subsequent week once the solution is released*

Problem:

Download the *bigLogNew.txt* file provided already as part of course material. Size of the file is approx. 1.4 GB. Transfer the file to the edge node of the cluster. Open multiple instances of this complete file using cat command and then append the contents of these instances to a new file named *bigLogLatest.txt*, so that the size of the new file genertaed becomes 10 gb approx.

We need this big file of 10 GB to proceed further.

Copy this 10gb file to hdfs (refer video session 5 & 6). Process this file in spark and count occurrences of each logging level. Try using **groupByKey** transformation. Dynamic allocation can be used.

Refer Spark UI to see how many tasks are launched at each stage.

See what is the level of parallelism achieved.

See the out of memory issues that occur and key-skew problem because of which job might fail.

(Refer video session 6 & 7 for the related explanation.)

Try to handle this problem using process of salting and see the impact (refer video session 9).

Note: *No solution will be released for this assignment as it is already explained in the referred videos.*

Google REVIEWS

5 Star Google Rated Big Data Course

LEARN FROM THE EXPERT

9108179578

Call for more details