Assignment

Week12: Apache Spark - Structured API Part-2

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 once the solution is released*

Problem 1:

Given 2 Datasets employee.json and dept.json

We need to calculate the count of employees against each department. Use Structured API's.

Sample output

depName,deptid,empcount
IT,11,1
HR,21,1
Marketing,31,1
Fin,41,2
Admin,51,0

Problem 2:

Find the top movies as shown in spark practical 18 using broadcast join. Use Dataframes or Datasets to solve it this time.

Problem 3:

File A is a text file of size 1.2 GB in HDFS at location /loc/x. It contains match by match statistics of runs scored by all the batsman in the history of cricket.

File B is a text file of size 1.2 MB present in local dir /loc/y. It contains list of batsman playing in cricket world cup 2019.

File A:

MatchNumber Batsman Team RunsScored StrikeRate

1 Rohit Sharma India 200 100.2

1 Virat Kohli India 100 98.02

1 Steven Smith Aus 77 79.23

35 Clive Lloyd WI 29 37.00

243 Rohit Sharma India 23 150.00

243 Faf du Plesis SA 17 35.06

File B:

Batsman Team

Rohit_Sharma India

Steven_Smith Aus

Virat_Kohli India

Question: Find the batsman participating in 2019 who has the best average of scoring runs in his career. Solve using Dataframes or Datasets.

Google REVIEWS

5 Star Google Rated
Big Data Course

LEARN FROM THE EXPERT

9108179578

Call for more details