Big Data 2019 17CS313

Assignment 1
IPL Analysis with MapReduce

1. Assignment Objectives and Outcomes

a. The objective of this assignment is for the students to install Hadoop in pseudo distributed mode and to become familiar with the Map Reduce programming environment and the HDFS file system.

   For this assignment, students will load data that is provided to them into HDFS and then write Map-Reduce code to perform a sequence of simple analyses on the IPL cricket dataset. The objective of the analysis is to find the most impactful players in IPL.

b. At the end of this assignment, the student will be able to write and debug Map-Reduce code.

2. Ethical practices

Please submit original code only. You can discuss your approach with your friends but you must write original code. All solutions must be submitted through the portal. We will perform a plagiarism check on the code.

3. IPL Dataset:

a. The dataset is obtained from https://cricsheet.org/

b. Use the given dataset.

   i. The CSV contains information about each IPL match played till 2018

      1. Info - 19 lines

      2. Ball by Ball data

   ii. All the matches are appended into one CSV file.

   iii. Use the provided dataset **ONLY.**

c. Use this dataset for the tasks given below.

4. Software/Languages to be used:

a. Python (Version **3.6**) (Use Hadoop Streaming) or Java (Version **1.8** Only)

b. Please make sure to use Hadoop Version **3.2.0** only.

5. Marks:

a. 20 (Scaled down to 5)

b. Each Task is for 5 marks.

6. Submission Date:

a. September 6 (Friday), 2019

7. Tasks:

a. These four tasks are to be computed, each of which carries one mark.
   i. Change the replication factor.
   ii. Batsman vulnerability to a bowler - # times that a batsman got out to a bowler.
   iii. Bowler vulnerability to a batsman - # runs given by a bowler to a batsman.
   iv. Most prolific batsman at each venue.

b. Load the data into HDFS.
c. Write the output of the mapper and reducer for each question in the report.
d. Submit one page report based on template and answer the questions on the report.

8. Submission Link:

Will be shared with you on the portal at a later date.

9. Task Specifications:

a. Task 0:

i. Install Hadoop in Pseudo-Distributed mode.

ii. You can refer the following few links (but not limited to) to help you get started - (Make sure that you install the correct versions of any software - as mentioned above)

iii. https://medium.com/@nidhinmahesh/getting-started-hadoop-mapreduce-hdfs-and-yarn-configuration-and-sample-program-febb1415f945

iv. https://netjs.blogspot.com/2018/02/installing-hadoop-single-node-pseudo-distributed-mode.html

v. https://medium.com/@jayden.chua/installing-hadoop-on-macos-a334ab45bb3 (This link should help MacOS users)

vi. You might need to use StackOverflow extensively to complete this task.

b. Task 1:

i. Problem Statement:

Change the replication factor.

ii. Description:

Change the replication factor of your Hadoop cluster to 2 and record which file(s) need(s) to be edited and what parameter(s) need(s) to be changed.

iii. Comments:

1. You need to submit any files that were edited to achieve this task.

c. Task 2:

i. Problem Statement:

Batsman Vulnerability to a Bowler.

ii. Description:

For every batsman-bowler pair to have more than 5 deliveries, in total, against each other, find the # of times a batsman got out to that bowler **using MapReduce**.

iii. Comments:

1. Sort the output in **descending** order of # of wickets.
2. In the case of 2 batsman bowler pairs having the same # of wicket occurrences, the pair who have played lesser number of balls should come first in the list. In case of a tie even after the previous condition, print the list in alphabetical order (based on Batsman name).
3. Wickets **do not include run-outs** as run-outs are not credited to the bowler
4. While counting number of deliveries, take into account any extras bowled by the bowler.
5. Output any numbers as integers only

iv. Output Format:

1. Comma separated values in the form of -
   Batsman, Bowler, # of wickets, # of deliveries
2. Each tuple/record must be in a new line with there should be **no spaces** in between the values.
   Eg. MS Dhoni,RD Chahar,7,24
   SR Watson,SN Thakur,6,33
   ...

d. Task 3:

i. Problem Statement:

Bowler Vulnerability to a Batsman.

ii. Description:

For every batsman-bowler pair to have more than 5 deliveries, in total, against each other, find the # of runs conceded by a bowler against a particular batsman **using MapReduce**.

iii. Comments:

1. Sort the output in **descending** order of # of runs.
2. In the case of 2 batsman bowler pairs having the same # of runs conceded, the pair who have played lesser number of balls should come first in the list. In case of a tie even after the previous condition, print the list in alphabetical order (based on Bowler name).
3. Runs conceded by a bowler **include** any extras conceded (Wides, No Balls, etc.)
4. While counting number of deliveries, take into account any extras bowled by the bowler.
5. Output any numbers as integers only.

iv. Output Format:

1. Comma separated values in the form of -
   Bowler, Batsman, # of runs conceded, # of deliveries
2. Each pair must be in a new line with there should be **no spaces** in between the values. (The values below are just for representational purposes)
   Eg. RD Chahar,MS Dhoni,43,24
   SN Thakur,SR Watson,41,33
   ...

e. Task 4:

i. Problem Statement:

Most prolific batsman at each venue. (Use the "Venue" column is present in the dataset).

ii. Description:

For each venue at which an IPL match has been played, find the batsman who has been the most prolific there **using MapReduce**. Use the strike rate of the batsman at that venue as a performance measure.
[Strike Rate = (Total No. of runs scored * 100)/ Total No. Of deliveries faced]
(The batsman should have faced a **minimum of 10 deliveries** at the venue)
(The batsman can have played in the same venue in different matches, consider the **overall** runs and deliveries)

iii. Comments:

1. Sort the output **alphabetically** based on the venue. In the case of 2 batsmen having the same strike rate, choose the batsman who has scored more number of runs.
2. **Ignore** all extras in the dataset, **do not consider** the runs scored or balls faced during all types of extras.
3. If the venue name contains commas or quotes, we want you to consider the entire name.
4. Consider the venue names **as is.**
   For example
   A. "M. Chinnaswamy Stadium" and "M Chinnaswamy Stadium" should be considered as two different venues (There's a difference of "dot") .
   B. "Punjab Cricket Association Stadium" and "Punjab Cricket Association Stadium IS Bindra Stadium" should be considered as two different venues.

iv. Output Format:

1. Comma separated values in the form of -
   Venue,Batsman
2. Each pair must be in a new line with **no spaces** between the values.
   Eg. Eden Gardens,SC Hussey
   Wankhede Stadium,MS Dhoni
   ...