Programming Assignment 1

CS 5433: Big Data Management

MapReduce Jobs

Part 1: using flume collect 2 sets of twitter data for 5 days (120 hours). The two sets will be collected separately and use different keywords, but should be related. For example, the word "Election" for the first set and "2020" for the second set. Store the data in HDFS in CSH [10 marks]

Part 2: count the number of rows in both datasets. Use MapReduce [10 marks]

Part 3: join the 2 twitter data sets based on common user information (such as author, ID or something else). Use MapReduce [20 marks]

Part 4: count the number of rows in the joined dataset. Use MapReduce [10 marks]

Part 5: Combine parts 2,3 and 4 into one MapReduce job [20 marks]

Collaboration Policy:

You should complete this programming assignment individually. Any doubts/clarification about the questions should be directed to either the instructor/TA. Using code from web & other resources partially/completely is prohibited and will be considered as plagiarism.

Note:

1. Source code should be written in Java
2. All the source code you submit should be well commented [Penalty for not commenting adequately 25%]
3. Your source code should run on CSH
4. Submissions
    a. README File for each question [FirstName_LastName_README_x]. The readme file will (a) describe your approach and (b) give instructions to run your code
    b. Commented source code for each question [FirstName_LastName_Program_x]
    c. All the source files zipped as a single zip file [FirstName_LastName_PA1.zip].

Deadline: Monday February 24th, 2020
Submit all your deliverables on Canvas