Programming Assignment I

CS 5433: Big Data Management

MapReduce Jobs

Part 1: Using flume collect twitter data based on any two keywords. In other words collect tweets that contain keywords 'a' or 'b'. Store the data in HDFS in the department's Hadoop cluster. Take screen shots of

i. files and directories where the tweet data is stored in HDFS
ii. contents of a file in HDFS that stores the tweets. Do not display all the contents. A snapshot of one file will suffice.

Before you start collecting twitter data, post on the discussion board in canvas the keywords you will be using. Do not use both keywords posted by another student, but you may use one of them. For example if 'Good and 'morning' have been posted on the discussion board, you may use one of them, but not both. [10 marks]

Part 2: Count the number of rows using MapReduce.
Take a screen shot that shows the number of rows in the dataset collected.
[10 marks]

Part 3: Partition the tweets based on the hashtag. Partition on a maximum of 10 hashtages. Your code should therefore include Map code, Partitioner code and Reducer code. Count the number of rows in each partiton based on the keywords 'a' and 'b'. For example,
#x, a, 10, #y, a, 20, …, #x, b, 5, #y, b, 17, ..
#x is a hashtag, **a** is a keyword and **10** is the count. Results of the MapReduce job may be output in any format.
Note: some of the tweets may contain both keywords 'a' and 'b', whereas others may contain only one keyword.
[20 marks]

Collaboration Policy:

You should complete this programming assignment individually. Any doubts/clarification about the questions should be directed to either the instructor or the TA. Make sure you acknowledge web & other resources that you have used in your work.

Note:

2. All Computer Science students should write the source code in Java. Non Computer Science students may use Java or Python.
3. All the source code you submit should be well commented [Penalty for not commenting adequately 25%]

4. Your source code should run on the Hadoop cluster in the department. Instructions to log in and collect twitter data using Flume are outlined in the document named "Using Hadoop and Flume.pdf".
5. Submissions
    a. README File for each part [FirstName_LastName_README_x]. 'x' is the part number. The readme file will give instructions to run your code and list the relevant files.
    b. Commented source code for each part [FirstName_LastName_Program_x]. 'x' is the part number.
    c. Report
        i. A maximum of one page per part that describes your approach (3 pages in total)
        ii. Up to two pages showing for each question screenshots of results. Include the part number and the figure number as well as a caption that will explain what a figure refers to. For example: 'Figure 2,3' refers to figure 3 in part 2.
    d. All the source files zipped as a single zip file [FirstName_LastName_Code.zip]. Each file should be named Code_x_i where' x' is the part number and 'i' is the i^{th} file for part x.

Deadline: Monday March 7th, 2022

Submit all your deliverables on Canvas