PROGRAMMING ASSIGNMENT - 2

Spring 2023

DUE DATE: 05/14/2023 11:59 PM

GOAL: The goal of this programming assignment is to process large data sets using MapReduce. The problem is almost the same as the programming assignment-1, i.e., determining the most frequent words except you use MapReduce.

THE PROBLEM STATEMENT:

1. Determine 100 most frequent/repeated words in the given dataset using MapReduce.
2. Determine 100 most frequent/repeated words in the given dataset considering only the words having more than 6 characters using MapReduce.

Input:

1. Use Previously used datasets
2. For testing you could use the smaller datasets(50MB or 300MB dataset)
3. **But finally the algorithm should be implemented on the largest dataset (i.e.16GB dataset).**

Note-1: Please note that you need to skip the “stop” words such as ‘the’ – please see the list of stop words here: **https://gist.github.com/sebleier/554280**

Note-2: **Make sure the text in the input datasets is converted to lowercase. (If the word is in upper case convert it into lower case)**

Improve the performance by tuning the configurations of the job. Please refer to the links provided below for understanding job configuration.

Using MapReduce: You can write code in any language supported by MapReduce such as Java, Python, or Scala.

DELIVERABLES: Source code including a readme file to run your code, Output screenshots and also a report (PDF format) should be included in a SINGLE tar/zip file uploaded on Camino. Naming convention for the zip/tar file is "LastName_FirstName.zip/tar"

Source Code: (10 points):

Java (or any language) files for MapReduce.

Please don't upload JAR files as it won't be considered for evaluation.

Note: Code should be documented and commented on. Good coding practices will get some extra points.

PDF Report (5 points):

Code Design & Tuning

- Name the file properly (use your names)
- Execution time with the default configuration. **You can also provide other metrics** -- totally depends on your exploration and creativity. Explain the meaning and significance of all the metrics that you list in the report.
    - Configuration Tuning: List out all the properties (with values) that were modified to achieve performance improvement.
- Data Analysis and Presentation (5 points)
    - You can use a wide variety of ways to present data. e.g. you can plot graphs of speed up vs configuration. **Proper reasoning is required** for all observations listed in your pdf.

Note:

1. One submission per team
2. Reference links MapReduce. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Job_Configuration

Typical criteria for grading:

1. How well the basic implementation is analyzed using different experiments and presenting different insights?
2. How well the data is presented and overall how well the analysis is conveyed?
3. What are different optimization techniques applied & Analize how well the impact of these techniques are.