CMI/Jan-Apr 2024/Assignment 4 Mar 2024

Chennai Mathematical Institute

Distributed Computing and Big Data Deadline: Apr 10, 2024 10:00 PM. Max Marks: 10.

Instructions:

(1) Submit your assignment solution as a single pdf file on moodle. Clearly mention your roll number and name in the solution pdf.
(2) You may write and scan your work or use tools like Word or Overleaf.

---

(1) Use the dataset in https://www.kaggle.com/datasets/datasnaek/chess. Include inline comments to explain the pig script.
   (a) Write a pig script to save the data rows where winner is white (see winner column) into a separate file.
   (b) Write a pig script to print the average rating of the winner when the winner is white.
   (c) Write a pig script to print the count of games that took more than 100 turns.
(2) You need to count the frequency of length of words in a given text file using map reduce paradigm. For example, if the input file has "hello world", the output should be a single line:

$$5, 2$$

meaning that there were two words of length five each. For another input, say, "I love India", the output should carry three lines:

$$1, 1$$

$$4, 1$$

$$5, 1$$

The input will be a large text document. You do not need to write the map reduce code. Describe the map reduce pattern that fits this work.