CMI/Apr-Jun 2021/Assignment 3 Jun 2021

Chennai Mathematical Institute

Distributed Computing and Big Data Deadline: Jun 25, 2021 11:59 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.
(3) A group submission of up to three students is allowed. If you are working in a group, ensure that only one member of the group submits to moodle.

---

(1) Use the dataset in https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29.
   (a) Write a pig script to save the non-drawn data rows into a separate file. Include inline comments to explain the pig script.
   (b) Write a pig script to count the number of positions that lead to a win within five moves. Include inline comments to explain the pig script.
(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.