Electrical and Computer Engineering / Software Engineering

CPRE/SE 419 Software Tools for Large Scale Data Analytics

Spring 2022

Homework 1

Due: Wednesday, February 16, 11:59PM

Preamble:
The purpose of this homework is for you to get some elementary "quantitative" practice on the topics covered in class so far. You homework has two parts: (1) General questions, to be answered very briefly; (2) Actual problem (plus – since you are great younger colleagues, you get the bonus of Part III: a reading assignment (just reading, nothing to report in this homework[1]))

Part I:

1. (12 pts.) Would you say that a Namenode machine should be on the same machine as a DataNode in terms of hardware?

2. (12 pts) Would you say that a Secondary node is a substitute/back up node for the NameNode?

3. (12 pts.) Briefly describe at least one advantage of YARN over Hadoop.

Part II:

1. Consider the setting of a machine with a single disk, with the following properties:
    i. The throughput is 256MB/sec.
    ii. There is a file of size 500GB

(a) (8 pts.) How long does it take to read the entire file into the main memory?

(b) (10 pts.) Now, consider distributed setting, where the cluster has the following properties:
    i. Each node has the same throughput (256MB/sec.).
    ii. There is a file of size 500GB which is split into blocks, and each block is 128MB.
    iii. There are a total of 20000 nodes.

    What would be the best-case scenario of distributing the blocks among the nodes, for the purpose of optimizing the throughput (in terms of making sure that the entire file is read (i.e., each portion is read from disk -> main memory)? What is the speed-up with respect to Problem 1?

(c) (12 pts.) For the final variant, consider settings (almost) same as in problem (b) – the only difference being that now the cluster has 100 nodes. What is the time to read the file (best-case scenario in terms of blocks distribution)?

2. (34 pts.) Assume that you are given a large file named *sales_data.txt* in which the lines are of the format: *(store, item, total_sales)*.
   Write a MapReduce pseudo-code that will provide the average sales per store.

[1] But, then – who knows what the future holds… maybe a pop-quiz, together with Vivaldi or Haydn cello concertos?

Part III:
Reading assignment: https://www.usenix.org/system/files/conference/atc13/atc13-cidon.pdf

What to turnin: Typed solutions are strongly preferred. If, for whatever reason, you are prevented from using any editor, then we may accept hand-written solution – only if they are legible.

You can work in teams of two students for this assignment (of course, we will honor individual submissions). If you are working in a team – please make sure to put the names of the team members in the beginning of the document.