CPTS 415 Big Data

Assignment 3

Instructor: Yinghui Wu

---

1. [Parallel Data models] [30]

    a. What's speed-up and scale-up? Give three reasons why we cannot do better than linear speedup.

    b. Assume a program P running on a single-processor system takes time T to complete. Assume 40% of P can only be executed sequentially on a single processor, and the rest is "embarrassingly parallel" in that it can be easily divided to smaller tasks executing concurrently across multiple processors. What are the best time costs to execute P using 2,4 and 8 machines (expressed by T), and what are the speed-up respectively?

    c. Describe and compare the pros and cons of the three architectures for parallel systems.

2. [MapReduce] [40] This sets of questions test the understanding and application of MapReduce framework.

    a. Facebook updates the "common friends" of you and response to hundreds of millions of requests every day. The friendship information is stored as a pair (Person, [List of friends]) for every user in the social network. Write a MapReduce program to return a dictionary of common friends of the form ((User i, User j), [List of Common friends of i and j]) for all pairs of i and j who are friends. The order of i and j you returned should be the same as the lexicographical order of their names. You need to give the pseudo-code of a main function, and both Map() and Reduce() function. Specify the key/value pair and their semantics (what are they referring to?)

    b. Top-10 keywords. Search engine companies like Google maintains hot webpages in a set R for keyword search. Each record r from set R is an article, stored as a sequence of keywords. Write a MapReduce program to report the top 10 most frequent keywords appeared in the webpages in R. (hint: you may need two rounds of MR process). Give the pseudo-code of your MR program.

3. [Graph parallel models] [30] This sets of questions relates to MR for graph processing.

    a. Consider the common friends problem in problem 2.a. We study a "2-hop common contact problem", where a list should be returned for any pair of friends i and j, such that the list contains all the users that can reach both i and j within 2 hops. Write a MR algorithm to solve the problem and give the pseudo code.

    b. We described how to compute distances with MapReduce. Consider a class of d- bounded reachability queries as follows. Given a graph G, two nodes u and v and an integer d, it returns a Boolean answer YES, if the two nodes in G can be connected by a path of length no greater than d. Otherwise, it returns NO. Write a MR algorithm to compute the query G(u,v,d) and give the pseudo code. Provide necessary correctness and complexity analysis.