Assignment #3

All the assignments must be submitted through Blackboard, in a single pdf format. Rename your file as "CPTS415_"+your last name+"_i.pdf", for your i-th homework.

1. **[Parallel Data Models]** (30)
   a. What is speedup and scaleup? Give three reasons why we cannot do better than linear speedup.
   b. Describe and compare the pros and cons of the three architecture for parallel systems.

2. **[MapReduce]** (40) This set of questions test the understanding and application of MapReduce framework.
   a. (20) 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 User i and User 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 1 main function, and 1 Map() and 1 Reduce() function. Specify the key/value pair and their semantics (what are they referring to?).

   b. (20) Most frequent keyword. Search engine companies like Google maintains hot webpages in a set R for keyword search. Each record $r \in R$ is an article, stored as a sequence of keywords. Write a MapReduce program to report the most frequent keyword appeared in the webpages in R. Give the pseudo-code of your MR program.
   Hit: You may need two rounds of MR processes for (b)

3. **[Apache Spark]** (30) This set of questions relate to Apache Spark

   a. Explain the definition of RDD and how the lineage retrieval works

   b. List the reasons why Spark can be faster than MapReduce.

   c. Explain the definitions of narrow dependencies and wide dependencies. In addition, explain how Spark determines the boundary of each stage in a DAG and why put operators into stages will improve the performance.