CS6350
Big data Management Analytics and Management
Spring 2022
Homework 2
Submission Deadline: March 7th, 2022

Q1. What is the difference between Hadoop and traditional database management systems?

Q2. Shuffling and sorting phase requires two types of sorting. Why and in which case are those sorting used?

Q3. During fault tolerance, in-progress jobs in the reducer side need to be rerun whereas completed ones do not. But in the mapper side, regardless of whether the jobs are in-progress or completed they need to be rerun. Why?

Q4. Suppose you have a file that stores all sales related information of a store. It contains the columns product name, price, payment mode, city, and country of client. The goal is to find out number of products sold in each country. Write a map/reduce pseudocode that computes the goal.

Q5. Write the pseudocode for a MapReduce program to perform matrix multiplication. You use the following link for reference:

http://www.mathcs.emory.edu/~cheung/Courses/554/Syllabus/9-parallel/matrix-mult.html