CE/CZ 4123 Big Data Management Tutorial 7
Distributed Systems and MapReduce

Question 1

Amazon wants to estimate the Top-*K* best sold products from *S* purchase records of *L* products in the form of a list of (User id, Product id) pairs. Assume that *L* is a multiple of *M*. Suppose there is a distributed system with 1 master machine and *M* slave machines. Design a distributed computation procedure to finish the task. Please describe

(1) how the data is distributed, computed and aggregated?

(2) how much data is sent across machines?

Question 2:

Consider the MapReduce paradigm and answer the following questions.

(1) In a MapReduce job, the output of Map phase is a list of key-value pairs: (A, 1) (C, 2), (A, 5), (C, 6), (B, 3), (E, 3), (C, 8). Please list the possible input to the Reduce function.

(2) Based on the answer to Q3(a), write a Reduce function (in pseudocode) so that the MapReduce output is: (2, A), (3, C), (6, A), (7, C), (4, B), (4, E), (9, C).

(3) Consider an employee table containing three columns (EmployeeID, age, monthly-salary) where age and monthly-salary are integers. Use MapReduce to collect the number of employees falling into each of the following two categories:

- Category 1: The age of the employee is between 30 and 40 (including 30 and 40). His/her monthly salary is at most 7000.

- Category 2: The age of the employee is between 40 and 50 (including 40 and 50). His/her monthly salary is more than 7000.

Please use only one MapReduce Job to achieve this task and write down the pseudocode of the Map function and Reduce function. The input key and value for Map function are an employee's age and monthly-salary respectively.

(Example: if there are 100 employees in Category 1 and 50 employees in Category 2, then the MapReduce output will contain two key-value pairs:
(1, 100), (2, 50))

Question 3:

Design MapReduce algorithms for the multiplication of two matrices A, B of n by n. Elements in the matrices are integers. The input key for *Map* function is *MatrixName*; the input value for *Map* function is in the form of *i;j;v*, indicating that the value of the *i*-th row and *j*-th column is *v*.

(Matrix Multiplication: Given matrix A[nxn] and matrix B[nxn], compute matrix C such that $C[i][z] = \sum_{j=0}^{n-1} A[i][j]xB[j][z]$, for i,z in [0, n-1]).

(1) Use at most two MapReduce jobs to finish the computation.
(2) Furthermore, can the multiplication be finished using one MapReduce job?