CS6350

Big data Management Analytics and Management

Spring 2018

Homework 1

Submission Deadline: 20th February, 2018

In this homework, you will use hadoop/mapreduce to solve the following problems.

Q1

Write a MapReduce program in Hadoop that implements a simple "Mutual/Common friend list of two friends". The key idea is that if two people are friend then they have a lot of mutual/common friends. This question will give any two Users as input, output the list of the user id of their mutual friends.

For example,

Alice's friends are Bob, Sam, Sara, Nancy

Bob's friends are Alice, Sam, Clara, Nancy

Sara's friends are Alice, Sam, Clara, Nancy

As Alice and Bob are friend and so, their mutual friend list is [Sam, Nancy]

As Sara and Bob are not friend and so, their mutual friend list is empty. (In this case you may exclude them from your output).

Input:

Input files

1. **soc-LiveJournal1Adj.txt**

The input contains the adjacency list and has multiple lines in the following format:

<User><TAB><Friends>

Here, <User> is a unique integer ID corresponding to a unique user and <Friends> is a comma-separated list of unique IDs (<User> ID) corresponding to the friends of the user. Note that the friendships are mutual (i.e., edges are undirected): if A is friend with B then B is also friend with A. The data provided is consistent with that rule as there is an explicit entry for each side of each edge. So when you make the pair, always consider (A, B) or (B, A) for user A and B but not both.

Output:

The output should contain one line per user in the following format:

<User_A><TAB><User_B><TAB><Mutual/Common Friend List>

where <User_A> & <User_B> are unique IDs corresponding to a user A and B (A and B are friend). < Mutual/Common Friend List > is a comma-separated list of unique IDs corresponding to mutual friend list of User A and B. Please find the above output for the following pairs.

(0,4), (20, 22939), (1, 29826), (6222, 19272), (28041, 28056)

Q2.

Please answer this question by using dataset from Q1.

Now Let's consider all the users (i.e. every line) in **soc-LiveJournal1Adj.txt.** Find friend pairs whose common friend number are within the top-10 in all the pairs, i.e. Sort friend pairs in descending order based on the number of common friends they have and output the top-10 pairs.

Output Format:

<User_A><TAB><User_B><TAB><Mutual/Common Friend Number>

Requirement: Please solve this question using **JOB CHAINING** technique.

Other solutions are not accepted!

Q3.

Numeric Operations.

Given a list of numbers, please find the mean value μ, variance σ

Output Format:

<Mean Value><TAB><Variance Value>

Input:

numbers.txt

Requirement: This program should be a one-pass efficient map-reduce program. Hint: you can use combiner to improve the efficiency.

Q4

Use the above dataset, implement an efficient map-reduce program to find the minimum value, maximum value and median value of these numbers.

Requirement: To answer all these 3 problems, a single program should be written and it should be a one-pass efficient map-reduce program. Hint: you can use either customized compareTo method or customized Partitioner.

Q5

Given two sparse matrix A and B ("p5-input.txt") whose format in the file is given as follows:

"A", i,j,val

"B", i,j,val

For example, A with two rows [0,1] and [2,0] is given as

"A", 0,1,1

“A”, 1,0,2

write a map-reduce program to implement the matrix operation A*B (matrix multiplication).

Submission Instructions: You have to upload your submission via e-learning before due date.

Please upload the following to eLearning:

1. The jar files, one for each problem.

2. Java files which have the source code.

3. An output of your program

4. ***A Readme text file about how to run your jar file. Give the command to run

your jar file.