Cloud Computing
COMP30520/COMP41110/COMP41610
Prof. M-Tahar Kechadi

Practical 3: MapReduce Programming Model

Due: 27 March 2015 Time 23:45 (Dublin Time)

Deliverable: ***ClassNo_Surname_FirstName_StudentNo_ Pracitcal3.pdf***

(or .doc, Ex: COMP41110_Smith_John_12345 _Practical3.pdf)

Important Notice:

- The programming problems I and II are compulsory for all students.
- The programming problems III and IV are only for students who could not submit Practical 2.

A. Description

Let's consider the following programming problems:

I. Matrix multiplication

Let A be a n×m matrix and B be a m×n matrix. Write a program that allows users to enter A and B and then computes the product of A and B.

II. Dissimilarity Matrix

Suppose a datasets A contains n objects (x_1, x_2, …, x_n) of p dimensions (e.g. A = {x_1(5, 3), x_2(2,6), x_3(4,1)}, A is a dataset of 3 objects (x_1, x_2 and x_3) of 2 dimensions), a dissimilarity matrix of A is a collection of distances for all pairs of n objects. It is often represented by an n × n table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

where d(i, j) is the distance between objects i and j. The distance (Euclidean distance) between objects i and j is defined as:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (1)$$

For example, with A = {x_1(5, 3), x_2(2,6), x_3(4,1)}:

$$d(x_2, x_1) = \sqrt{(2-5)^2 + (6-3)^2} = 4.24$$

$$Dis(A) = \begin{bmatrix} 0 & & \\ 4.24 & 0 & \\ 2.24 & 5.39 & 0 \end{bmatrix}$$

Write a program that allows user to enter a dataset A of n objects of p dimensions and then computes the dissimilarity matrix of A.

III. Log File Analysis

A common task in digital forensics is the analysis of log files. Each line (entry) of a web server's log file normally contains important information such as: IP address, data and time of request, request line, HTTP status, etc. Write a program that takes a log file as an input and then extracts the following information from the log files: the total number of connections to the server (i.e. total numbers of entries), the number of distinct IPs and the number of entries for each IP.

IV. Distance of two clusters

There are many methods used to measure the distance between two clusters P_1 and P_2. One of the widely used measures is mean distance that is defined as:

$$d_{mean}(P_1,P_2) = d(c_1, c_2)$$

where c_1 and c_2 are centroids of clusters P_1 and P_2 respectively, $d(c_1, c_2)$ is the Euclidean distance of two objects c_1 and c_2 (cf. formula (1)). For example, $d_{mean}(P_1,P_2)$ of two cluster P_1 and P_2 in the example of Lecture 9 (Part 3, slide 9) is:

$$d(c''_1, c''_2) = \sqrt{(22.67 - 12)^2 + (20.67 - 12.67)^2} = 13.34$$

Write a program that takes two clusters of objects as an input, computes the mean distance of these two clusters.

B. Question:

Suppose that we use MapReduce programming model to code the programming problems above, answer the following questions:

1. Define the input and the output of the Map function(s) for each programming problem. Justify your answer.
2. Define the input and the output of the Reduce function(s) for each programming problem. Justify your answer.

Note that for each programming problem, you can have one or many Map /Reduce functions.

C. Submission:

Submission should take place via moodle on or before the deadline. Submission should consist of one file (MS Word or PDFs) which contains the answer of all questions