CLOUD COMPUTING

Practical 04: Map-Reduce Programming Model

MapReduce model is a framework for designing solution in terms of parallel tasks, which are them combined to give the final desired output result. In this practical, we will build MapReduce solutions for some classical and very popular applications such as matrix-matrix multiplication, k-means algorithm, etc.

Exercise 1

Suppose that we have a large web corpus. Its metadata file has lines of the form (URL, size, date, ...). For each host, find the total number of bytes, i.e. the sum of the page sizes for all URLs from that host.

1. Define the input and the output of the Map and Reduce functions.
2. Write Map and Reduce functions. (pseudo-code)

Assume that we have generated two matrices from the previous file; A (n x n) and B (n x n). One wants to calculate another matrix M (n x n), which the product of the first two.

3. Define the input and the output of the Map and Reduce functions.
4. Write Map and Reduce functions of the matrix-matrix multiplication.
5. What if the matrices A and B do not fit in the mappers' memory? (How would you implement your Map-Reduce program?)

Exercise 2

K-Means is a simple clustering algorithm that aims to partition a set of objects into k clusters in which each object belongs to the cluster with the nearest mean. K-Means algorithm has 4 steps:

- **Step 1:** Given k, partition objects into k nonempty subsets
- **Step 2:** Compute seed points as the centroids of the clusters of the current partition. The centroid is the centre (mean point) of the cluster
- **Step 3:** Assign each object to the cluster with the nearest seed point
- **Step 4:** Go back to Step 2 until no more new assignment Define the input and the output of the Map and Reduce functions of KH Mean algorithm

Submission Instructions

All submissions must be done via Moodle. The submission deadline is Friday October 9th at 17:00. Your submission should consist of one file (MS Word or PDFs), which contains the answers to the above questions. The submitted file should be named following the format:

COMP*xxxxx**_**Surname_FirstName_StudentNo**_ **Pracitcal**04*.{doc, pdf}

Example: *(* ***COMP****41110_Smith_John_12345* ***_Practical****04.pdf)*