CLOUD COMPUTING Practical

School of Computer Science
University College Dublin

Practical 4: Map-Reduce Programming Model

To be Graded: YES

MapReduce is a programming framework for designing simple data parallel solutions. The main idea is to divide the application at hand into a number of tasks, which can be executed on a parallel and/or distributed system. The MapReduce paradigm is simple, it consists of two main operations: Map and Reduce functions. The first function is to **map** the tasks onto processing nodes. The second operation is to combine (**reduce**) all the local results to obtain the final result.

In this practical, we will develop MapReduce solutions for some classical and very popular applications, such as matrix-matrix multiplication, k-means algorithm, etc.

Question 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. **This is an example of metadata file**:

```
...
https://www.ucd.ie/contact-details/contact-ucd/,         30KB,          21/02/2019
https://hadoop.apache.org/docs/r3.1.2/,            5.1KB,              04/05/2019
http://www.ucd.ie/events/,                    300KB,                   15/09/2019
...
```

1. Define the inputs and the outputs of the Map and Reduce functions.
2. Write Map and Reduce functions for this problem (pseudo-code).

Question 2

Assume that we have two matrices: 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.

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

Question 3

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. The 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

1. Define the input and the output of the Map and Reduce functions of K-Mean algorithm
2. Write Map and Reduce functions of the K-Means algorithm

________________________ o0o __________________________

Submission Instructions

All submissions must be done via **Brightspace.** The deadline is Monday **18th October 2021, 23:55**. Your submission should consist of one file (PDF), which contains the answers, to the above questions. The submitted file should be named following the format:

COMP_xxxxx_Surname_FirstName_StudentNo__ ***Pracitcal***_04_.pdf

Example: _(**COMP**47780_Smith_John_12345 _**Practical**04.pdf)_