CLOUD COMPUTING

School of Computer Science
University College Dublin

Practical 5: Implementation of Map-Reduce Programming Model

To be Graded: YES

MapReduce model is a framework for designing solution in terms of parallel tasks, which are then combined to give the final desired output result. In this practical, we will implement MapReduce solutions for a classical and very popular application, i.e., matrix-matrix multiplication.

Important Links

Setup Cloudera HDP Sandbox

- You may want to read the following tutorial. Getting started with HDP sandbox https://www.cloudera.com/tutorials/getting-started-with-hdp-sandbox.html
- Download HDP sandbox (https://www.cloudera.com/downloads/hortonworks-sandbox.html). Choose the virtual machine image that fits your operating system. You may need to install a virtual machine player (VMWare or VirtualBox).
- Connect to the virtual machine using browser or secure shell (SSH). The following tutorial will help you to learn how to connect with the HDP sandbox https://www.cloudera.com/tutorials/learning-the-ropes-of-the-hdp-sandbox.html

You may also choose to Install standalone Hadoop on Ubuntu

How to Install Hadoop On Ubuntu 18.04 Or 20.04[1]

Exercise 1

Implement Map and Reduce functions of the matrix-matrix multiplication in Python.

Q1: Describe the setup architecture of your exercise.
Q2: Describe Mapper and Reducer function in Python.

Exercise 2

Execute and test the implemented matrix multiplication application on Hadoop using MapReduce. You may choose to install Cloudera HDP Sandbox with a pre-installed Hadoop setup or install a standalone version of Hadoop on your local machine.

Q3: Describe your experience step-by-step in your own words and provide screenshots of executed MapReduce programs.

[1] https://phoenixnap.com/kb/install-hadoop-ubuntu

NOTE: ***How to Run the Map-Reduce Job on Hadoop***

Explore Hadoop Streaming[2] to execute Python MapReduce Jobs on Hadoop. Before you run the actual MapReduce job, you must first copy the files from our local file system to Hadoop's HDFS.

Example:

```
$ chmod +x <path>/Mapper.py
$ chmod +x <path>/Reducer.py

$ $HADOOP_HOME/bin/hadoop jar <path>/<streaming>.jar \
-input input_dirs \
-output output_dir \
-mapper <path/mapper.py> \
-reducer <path/reducer.py>
Where "\" is used for line continuation for clear readability
```

Submission Instructions

All submissions must be done via **Brightspace** with deadline: **1 November 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****05*.pdf

Example: *(**COMP**41110_Smith_John_12345 _**Practical**05.pdf)*

[2] https://hadoop.apache.org/docs/current/hadoop-streaming/HadoopStreaming.html