Lab 3: MapReduce Programming

# Objectives

* Understand the MapReduce concept.
* Get familiar with the Hadoop framework.
* Experience working with small Hadoop cluster using VMs.

# Equipment Needs

* Computers
* Internet

# Experiments

## Basics

1. Go through the Apache Hadoop introduction to get the general idea about Hadoop:

<http://hadoop.apache.org/>

1. Go through the Apache Hadoop release notes to understand the evolution of Hadoop:

<http://hadoop.apache.org/releases.html>

## Hadoop Single Node Mode.

1. Follow the instructions on   
   <http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>  
   to set up Hadoop environment on your own Linux machine.
2. Follow the instructions/tutorials from the above link to run simple practice with single node mode.

## Hadoop “cluster”

Create two VMs (one master and one slave) on your own computer (ex: Virtualbox) and construct a small hadoop cluster for running the word count program. You can also use Docker containers to perform this assignment.

1. Create two VMs.
2. Configure the VM network so that the VMs can ping and communicate with each other.
3. Download and install hadoop and all required tools.
4. Configure hadoop configure files for your two-VM cluster
5. Run the WordCount Hadoop job on the file (wordCountText.txt) provided by the TA.

# Reports

1. What are the differences between Hadoop 0.X, 1.X, 2.X?
2. What is YARN? Why do we need YARN?
3. What is Hadoop streaming?
4. Screenshots of the practice of single node mode Hadoop on your own computer. The screenshot should show the output and result of Hadoop execution as well as the files in HDFS.   
   (ref: http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html)
5. Run WordCount on the Hadoop cluster with 2 VMs, what are the top 5 most frequent word in the provided txt file?
6. Use jps commands on both VMs to show running Hadoop daemons and provide and screenshots.
7. Screenshots of configuration files and IP addresses for Master node and Slave node of your small cluster as well as the MapReduce execution result. For each configuration file, please also briefly explain what it does.
8. What are the differences between Hadoop master and slave nodes? Also name what functionalities are performed on each node.
9. Write a pseudo code to multiply large matrices using Hadoop. Also explain the function of your Mappers and Reducers.
10. What is a combiner? Add a combiner to the last question and explain its function.

**We have zero tolerance to forged or fabricated data!!** A single piece of forged/fabricated data would bring the total score down to zero.