CS 4740 Programming Assignment 5

**Hadoop YARN Configuration on EC2**

1. Use four t2.micro instances to configure a 4-node Hadoop cluster on EC2 (*do not use Amazon EMR directly, we need you to configure the Hadoop cluster by yourselves*). You may refer to any online tutorials. Please use Hadoop version 2.7.X.

Tutorial link: <https://www.youtube.com/watch?v=cr5RmnyWbYw>

Please refer to file *commands.txt* for the explanation of each command of the steps in the above video.

2. Test the configure Hadoop YARN cluster with Teragen and Terasort benchmark using 1GB data.

3. You may need 3-4 hours or longer time to finish PA5.

Notes:

1. Going through videos part1 to part8 will give you a better understanding of what Hadoop is and what HDFS is (please listen to explanation, it is helpful).

2. The videos are based on Windows OS. But the only difference for Macbook is ignoring part 2 video and instead searching how to connect to EC2 instance for Mac online.

3. Java version: if you are using Ubuntu 14.04, following the tutorial is fine. But if you are using ubuntu 16.04, you should use "sudo apt-get install openjdk-8-jdk-headless". Pay attention to the ubuntu version you request for EC2 instance.

4. In video part3, you use command ssh-keygen -f ~/.ssh/sshkey\_rsa -t rsa -P "" to generate a key file called sshkey\_rsa. In this step, replace the name sshkey\_rsa with id\_rsa and also replace the name for the following steps that use sshkey\_rsa.

5. Don't skip any of the steps in the tutorial.

6. How to run the teragen and terasort

hadoop jar $HADOOP\_HOME/share/hadoop/mapreduce/hadoop-\*examples\*.jar teragen 10000000 /terasort-input

hadoop jar $HADOOP\_HOME/share/hadoop/mapreduce/hadoop-\*examples\*.jar terasort /terasort-input /terasort-output

**What to submit:** You need to submit screenshots for some configuration files, screenshots for testing some simple Hadoop command, and screenshot showing Teragen, Terasort is running to show that you successfully configure the cluster and the benchmark runs successfully.

You **MUST** submit one screenshot of your cluster health by going to ec2-{instance-address}.amazonaws.com:50070 when your cluster is running. See the example below.

