IEMS 5730 Spring 2022 Homework #0

Release date: Jan 12, 2022 (Wednesday)
Due date: Jan 24, 2022 (Monday) 11:59:00am (i.e. noon-time)
No late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on the Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student_____) Date:_____

Name_____ SID_____

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [100 marks + 10 bonus marks]: Hadoop Cluster Setup

Hadoop is an open-source software framework used for distributed storage and processing. In this problem, you are required to set up a Hadoop cluster using Amazon EC2, or Google Compute Engine. References [1] and [2] provide the tutorial for each platform respectively.

In order to set up a Hadoop cluster with multiple virtual machines (VM), you can set up a single-node Hadoop cluster for each VM first [3]. Then modify the configuration file in each node to setup a Hadoop cluster with multiple nodes. References [4] provide the setup instruction for a Hadoop cluster.

a. [**20 marks**] Single-node Hadoop Setup
   In this part, you need to set up a single-node Hadoop cluster in a pseudo-distributed mode, and run the Terasort example on your Hadoop cluster.
   i. Setup a single-node Hadoop cluster (**Hadoop version: 2.9.x,** all available versions can be found in [10]). Print the page of http://localhost:50070 to verify that your installation is successful.
   ii. After installing a single-node hadoop cluster, you need to run the Terasort example[5] on it. You need to record all your key steps, including your commands and output. Following commands may be useful:

```
$ ./bin/hadoop jar  \
      ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar  \
      teragen 100000 terasort/input
                                  //generate the data for sorting
$ ./bin/hadoop jar  \
      ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar  \
      terasort terasort/input terasort/output
                                  //terasort the generated data
$ ./bin/hadoop jar  \
      ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar  \
      teravalidate terasort/output terasort/check
                                  //validate the output is sorted
```

b. [**60 marks**] Multi-node Hadoop Cluster Setup
   After you setup a single-node Hadoop cluster in each VM, you can modify the configuration files in each node to set up the multi-node Hadoop cluster.
   i. Install and set up a multi-node Hadoop cluster **with 4VMs (1 Master and 3 Slaves).** Use the 'jps' command to verify all the processes are running.
   ii. In this part, you need to use the 'teragen' command to generate 2 different data-sets of size 2GB and 20GB to serve as input for the Terasort program. Then, run the Terasort code again for these different datasets and compare their running time.

   Hints:

- Keep an image for your Hadoop cluster. You would need to use the Hadoop cluster again for subsequent homework assignments.
- *(Important !)* Keep in mind that any port on which you allow inbound traffic represents a potential security vulnerability. it is not recommended to set the Source IP as 0.0.0.0/0 (everywhere) as many online tutorials suggest. It is recommended that you only allow inbound traffic from your PC's IP address (or CUHK IP if you are using CUHK VPN)

c. [**20 marks**] Running the Python Code on Hadoop
Hadoop streaming is a utility that comes with the Hadoop distribution. This utility allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer. In this part, you need to run the Python wordcount script to handle the Shakespeare dataset[6] via Hadoop streaming.
   i. Reference [7] introduces the method to run a Python wordcount script via Hadoop streaming. You can also download the script from the reference[8].
   ii. Run the Python wordcount script and record the running time. Following command may be useful:

```
$ ./bin/hadoop jar  \
      ./share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar  \
      -file mapper.py -mapper mapper.py  \
      -file reducer.py -reducer reducer.py  \
      -input input/*  \
      -output output
                          //submit a Python program via Hadoop streaming
```

d. [**Bonus 10 marks**] Compiling the Java WordCount program for MapReduce
The Hadoop framework is written in Java. You can easily compile and submit a Java MapReduce job. In this part, you need to compile and run your own Java wordcount program to process the Shakespeare dataset [6].
   i. In order to compile the Java MapReduce program, you may need to use the "hadoop classpath" command to get all Hadoop jars. Or you can simply copy all dependency jars in a directory and use them for compilation. Reference [9] introduces the method to compile and run a Java wordcount program in the Hadoop cluster. You can also download the Java wordcount program from reference [8].
   ii. Run the Java wordcount program and compare the running time with the part c.

IMPORTANT NOTES:

1. Since AWS will not provide any free credits anymore, we recommend you to use Google Cloud for this homework.
2. Only if your AWS Educate account and Google Cloud account can not be approved by Amazon and Google, send an email to tds019@ie.cuhk.edu.hk to get a template AWS account for the Homework#0. There are 50 USD in the account, which is

enough for you to complete Homework#0. Since the school's temporary AWS accounts share the same EC2 resource, please don't modify others' instances (including the security group settings). Otherwise, we will remove your account without any further notice!!!

3. Please download Putty from the website https://www.putty.org/ and avoid using the default private key. Failure to do so will subject your AWS account/ Hadoop cluster to hijacking.
4. Launching instances with Ubuntu 16.04 LTS is recommended. Hadoop version 2.9.x is recommended. Older versions of Hadoop may have vulnerabilities which can be exploited by hackers to launch DoS attacks.
5. (AWS) For the VM, you are recommended to use the t2.large instance type with 100GB hard disk, which consists of 2 CPU cores and 8GB RAM.
6. (Google) For the VM, you are recommended to use the n1-standard-2 instance type with 100GB hard disk, which consists of 2 CPU cores and 7.5GB RAM.
7. When following the given references, you may need to modify the commands according to your own environment, e.g., file location, etc.
8. After installing a single-node Hadoop, you can save the system image and launch copies of the VM with that image. This can simplify your process of installing the single-node Hadoop cluster on each VM.
9. Keep an image for your Hadoop cluster. You would need to use the Hadoop cluster again for subsequent homework assignment.

Submission Requirements:

1. Include all the key steps, codes, together with screenshots, into a SINGLE PDF report.

References:

[1] Google Compute Engine Tutorial: https://cloud.google.com/compute/docs/quickstart
[2] AWS Tutorial: https://aws.amazon.com/getting-started
[3] Single-Node Hadoop setup:
https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/SingleCluster.html
[4] Multi-node Hadoop cluster setup:
https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/ClusterSetup.html
[5]Terasort example:
https://hadoop.apache.org/docs/r2.9.2/api/org/apache/hadoop/examples/terasort/package-summary.html
[6] Shakespeare dataset
https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/assignments/shakespeare.zip
[7] Writing an hadoop mapreduce program in python
http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
[8] MapReduce wordcount program

https://www.dropbox.com/s/kdhlzkcajq1g5h1/MapReduce_WordCount.zip?dl=0
[9] Compile and run Java MapReduce program
https://hadoop.apache.org/docs/r2.9.2/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
[10] The arxiv of different versions of Hadoop https://archive.apache.org/dist/hadoop/core/