Q0 [10 marks]: Secure Virtual Machines Setup on the Cloud

In this task, you are required to set up virtual machines (VM) on a cloud-computing platform. You are free to use any cloud platform, but Google Cloud is recommended. References [1] and [2] provide the tutorial for Google Cloud and Amazon AWS, respectively.

The default network settings in each cloud platform are *insecure*. **Your VM can be hacked by external users, resulting in resource overuse which may charge your credit card a big bill of up to $5000 USD.** To protect your VMs from being hacked and avoid economic losses, you should set up secure network configurations for all your VMs.

In this part, you need to set up a white list for your VMs: *only users in the CUHK network can access your VMs via SSH.* Traffic outside CUHK should be blocked. You can connect to CUHK VPN to ensure you are in the CUHK network. Reference [3] provides the CUHK VPN setup information from ITSC.

a. [**10 marks**] Secure Virtual Machine Setup
   Reference [4] and [5] are the user guides for the network security configuration of AWS and Google Cloud respectively. You can go through the document with respect to the cloud platform you use. Then follow the listed steps to configure your VM's network:
   i. find/ create the security group/ firewall of your VM;
   ii. remove all rules of inbound/ ingress and outbound/ egress;
   iii. add a new rule to the inbound/ ingress, with the SSH port(s) of VMs (default: 22) and source '137.189.0.0/16' specified.

Q1 [90 marks + 20 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 the VMs you instantiated in Q0.

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 [6]. Then modify the configuration file in each node to set up a Hadoop cluster with multiple nodes. References [7] 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. Set up a single-node Hadoop cluster (**Hadoop version: 2.9.x**, all available versions can be found in [13]). Print the page of http://localhost:50070 (or http://<VM ip>:50070 in the browser of your local machine) to verify that your installation is successful.

ii. After installing a single-node hadoop cluster, you need to run the Terasort example[8] on it. You need to record all your key steps, including your commands and output. The 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
```

Notes: In case you will need to monitor the Hadoop service via Hadoop NameNode WebUI (http://<VM ip>:50070) on your local browser, based on steps in Q0, you may further allow traffic from CUHK network to access port 50070 of VMs.

b. [**40 marks**] Multi-node Hadoop Cluster Setup

After the setup of a single-node Hadoop cluster in each VM, you can modify the configuration files in each node to setup 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 datasets 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.

Notes: You may need to add each VM to the white list of your security group/ firewall, and further permit traffic towards more ports needed by Hadoop/YARN services (reference [14] [15]).

c. [**30 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[9] via Hadoop streaming.

i. Reference [10] introduces the method to run a Python wordcount script via Hadoop streaming. You can also download the script from the reference [11].

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 20 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 [9].
   i. In order to compile the Java MapReduce program, you may need to use "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 [12] 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 [11].
   ii. Run the Java wordcount program and compare the running time with part c.

IMPORTANT NOTES:

1. Since AWS will not provide free credits anymore, we recommend you to use Google Cloud (which offers a 90-day, $300 free trial) for this homework. For those who still prefer using AWS, please submit your applications to apply for a dedicated AWS account in this form: https://forms.gle/pPSbCUWmgpAwJSEe8 TAs will review your application and approve the request if it is reasonable. The assigned AWS account will be strictly monitored and only provided limited services.
2. If you use Putty for SSH client, please download 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.
3. Launching instances with Ubuntu 18.04 LTS is recommended. Hadoop version 2.9.x is recommended. Older versions of Hadoop may have vulnerabilities that can be exploited by hackers to launch DoS attacks.
4. (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.
5. (Google) For the VM, you are recommended to use the n2-standard-2 instance type with 100GB hard disk, which consists of 2 CPU cores and 8GB RAM.
6. When following the given references, you may need to modify the commands according to your own environment, e.g., file location, etc.
7. 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.
8. Keep an image for your Hadoop cluster. You would need to use the Hadoop cluster again for subsequent homework assignments.

Submission Requirements:

1. Include all the key steps, source codes of your programs, 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. CUHK VPN user guide: https://www.itsc.cuhk.edu.hk/all-it/wifi-and-network/cuhk-vpn/
4. User guide of Amazon EC2 security groups for Linux instances: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-security-groups.html
5. User guide of Google Cloud firewall rules: https://cloud.google.com/vpc/docs/firewalls
6. Single-Node Hadoop setup: https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/SingleCluster.html
7. Multi-node Hadoop cluster setup: https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/ClusterSetup.html
8. Terasort example: https://hadoop.apache.org/docs/r2.9.2/api/org/apache/hadoop/examples/terasort/package-summary.html
9. Shakespeare dataset https://mobitec.ie.cuhk.edu.hk/iems5730Spring2023/static_files/assignments/shakespeare.zip
10. Writing a Hadoop MapReduce program in python http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
11. MapReduce wordcount program https://www.dropbox.com/s/kdhlzkcajq1g5h1/MapReduce_WordCount.zip?dl=0
12. Compile and run Java MapReduce program https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
13. The archive of different versions of Hadoop https://archive.apache.org/dist/hadoop/core/
14. HDFS Service Ports https://docs.cloudera.com/HDPDocuments/HDP2/HDP-2.6.5/bk_reference/content/hdfs-ports.html
15. YARN service ports https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.0/administration/content/yarn-ports.html