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 Eddie Christopher Fox III) Date: January 23, 2022

Name Eddie Christopher Fox III SID 1155160788

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.

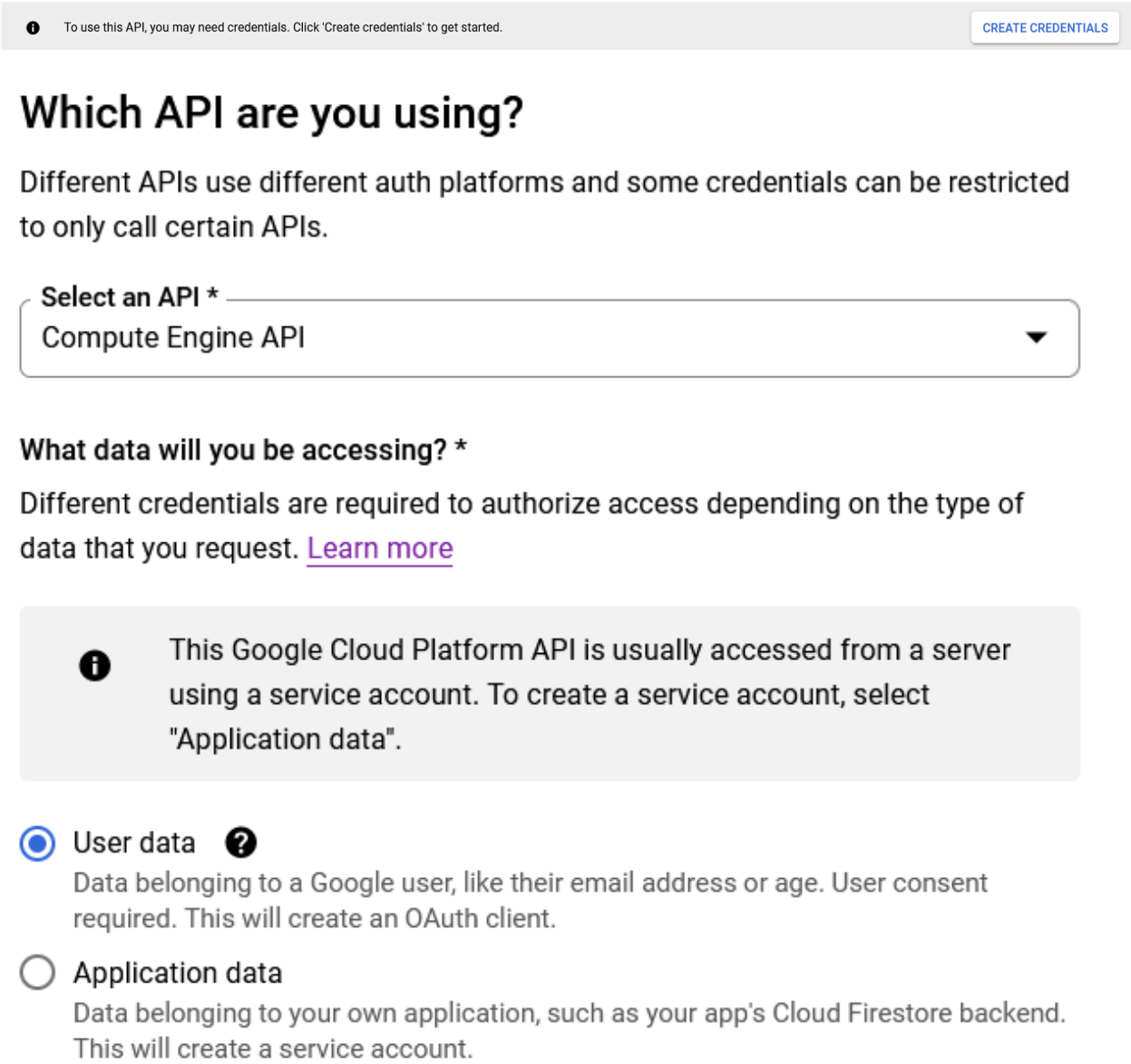
IEMS 5730: Big Data Systems and Information Processing: Spring 2022, Assignment #0

Part 1: Setting up Google Cloud and Ubuntu VM

First, I register for Google Cloud as suggested by homework 0 PDF. I did create an AWS educate account only to confirm they are no longer offering the 100 USD credit.

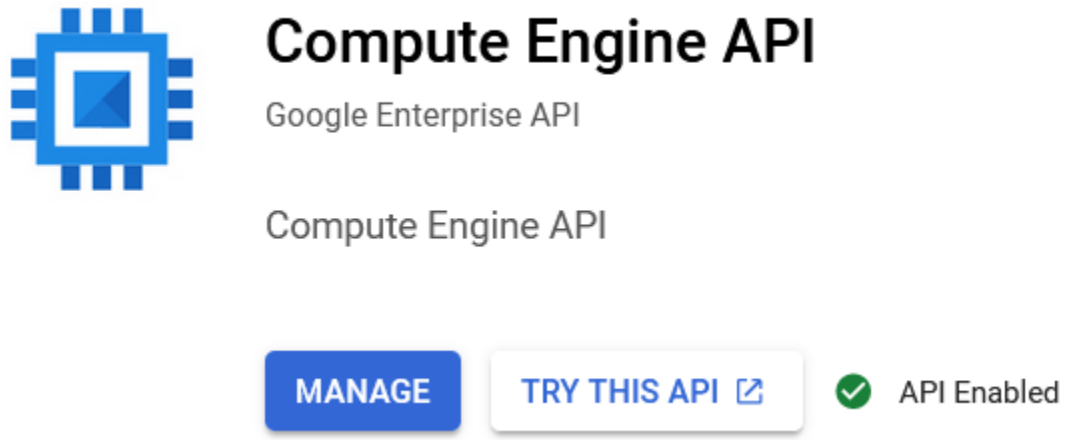
I use the free trial to get 2339 hkd in credit for 91 days.

I use the link provided in the homework 0 PDF for Google Cloud [1] to create the Linux VM. Before I begin, I need to enable the Compute Engine API, and I am prompted to create credentials to do so.



I choose user data. I continue through the registration process, filling in placeholder application names and using my Email address where one is requested. I do not add any scopes.

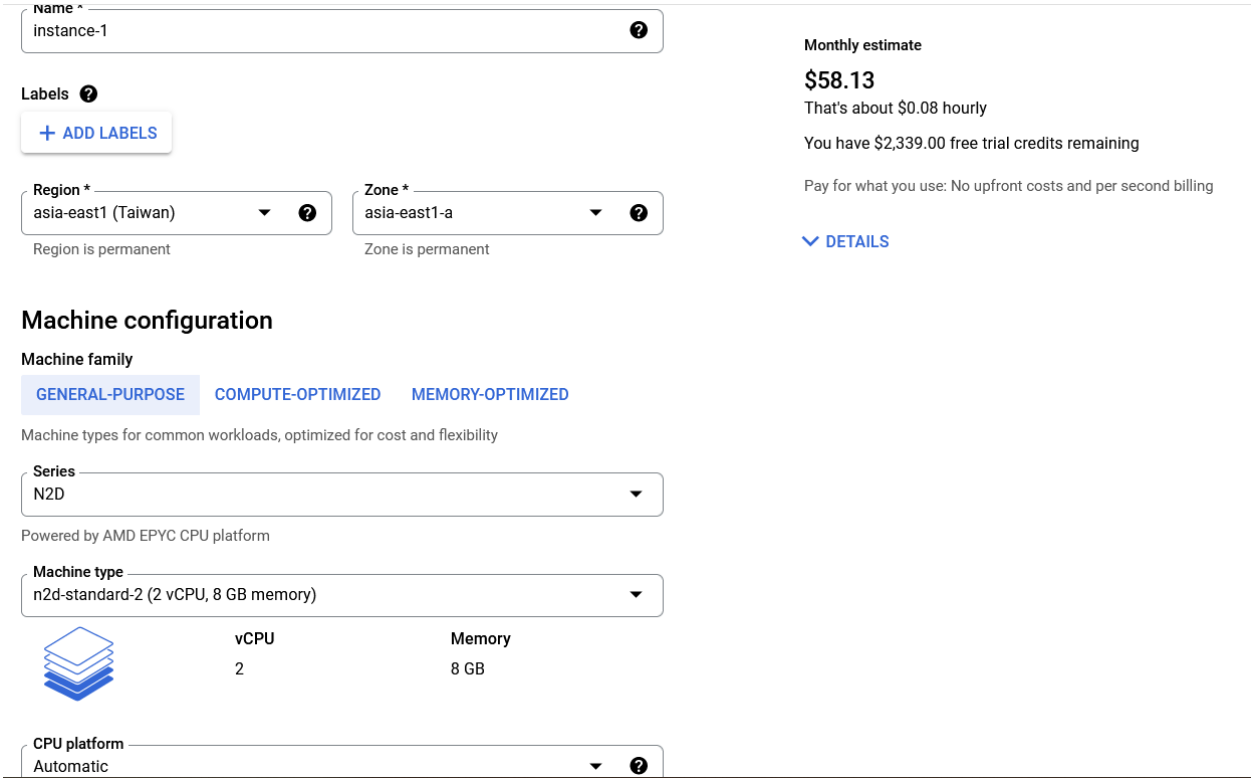
In step 4 I am prompted to create an OAuth Client ID so I choose Desktop app type application, and use a placeholder client name. I finally download my credential in JSON format, as provided by Google.

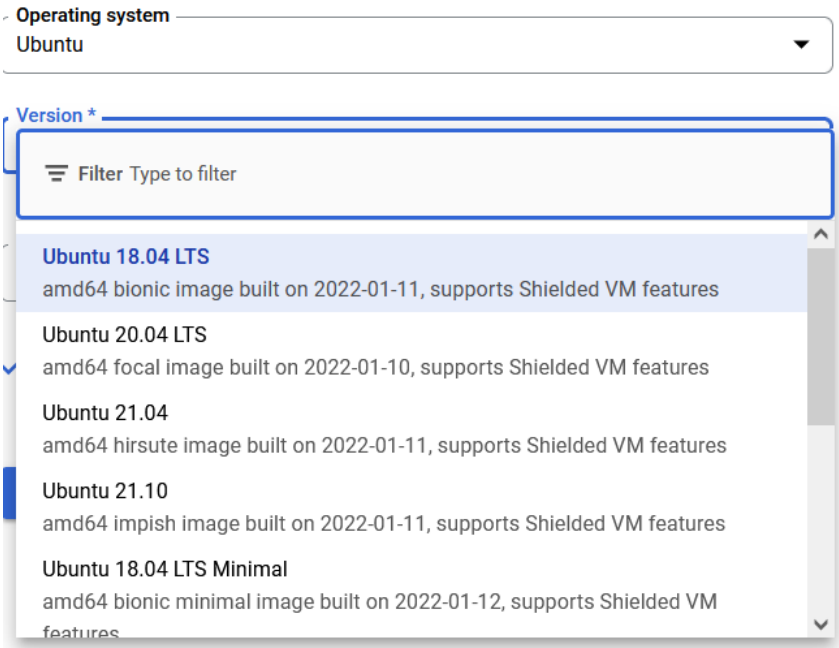


I confirm the API is enabled afterwards, then proceed with creating a Linux VM. I go to "Create an instance" and fill out the information as suggested in [2], the IERG 4300 fall tutorial 1, detailed VM creation PDF. I use asia-east1, Taiwan instead of asia-east2, Hong Kong, to match the PDF [2].

We are recommended to use N1 type with 2 cores and 7.5 GB memory in Homework 0 pdf, but it seems there is now a N2D type, second generation N type that uses AMD EPYC CPU. I choose the n2d-standard-2 type that has 2 cores and 8 GB memory to closely match. According to [3], n2d has superior efficiency, price-performance, and raw performance compared to N1, and is only 2hkd more per month.

Continued on next page.





I attempt to choose Ubuntu 16.04 LTS as suggested in Assignment 0 PDF, but this is the oldest version allowed, so I choose it. I choose balanced persistent disk, the default, and 100 GB size as recommended by assignment 0.

I enable display device (screen capturing and recording tools allowed), and http and https traffic as recommended in [2].

I choose the compute engine default service account under Identity and API access, and allow default access for access scopes. Unlike [2], I didn't see a menu to input public key, so I disregard this at the time.

I follow [4] to create a public key and [5] to provide my public key to Google and connect to my VM through Putty on my Windows 10 laptop.

EddieCFox@instance-1: ~

```
Using username "EddieCFox".
Authenticating with public key "EddieCFox"
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-1060-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Jan 19 03:58:47 UTC 2022

  System load:  0.0               Processes:           108
  Usage of /:   1.7% of 96.75GB   Users logged in:     0
  Memory usage: 2%                IP address for ens4: 10.140.0.2
  Swap usage:   0%

0 updates can be applied immediately.



The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

EddieCFox@instance-1:~$
```

I am now following the Fall 2020 Tutorial 1 [6] to configure things appropriately. There is no need to install SSH server as it is already on the latest version.

```
EddieCFox@instance-1:~$ sudo apt-get install openssh-server
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-server is already the newest version (1:7.6p1-4ubuntu0.5).
The following package was automatically installed and is no longer required:
  libnuma1
Use 'sudo apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
EddieCFox@instance-1:~$ sudo addgroup hadoop
Adding group `hadoop' (GID 1003) ...
Done.
```

```
EddieCFox@instance-1:~$ sudo usermod -a -G hadoop EddieCFox
EddieCFox@instance-1:~$ groups
EddieCFox adm dialout cdrom floppy audio dip video plugdev lxd netdev ubuntu google-sudoers
EddieCFox@instance-1:~$ groups EddieCFox
EddieCFox : EddieCFox adm dialout cdrom floppy audio dip video plugdev lxd netdev ubuntu goog
le-sudoers hadoop
```

I add myself to the Hadoop group.

Part 2: Single-node Hadoop setup (standalone)

Using [7], I see that I need to install Java onto the linux VM before I can run Hadoop. [8] notes that for version 2.9.2 (the latest 2.9.X) both Java 7 and Java 8 are supported, but versions 3+ will support Java 8 only. The documentation provided in assignment 0, [9], uses version 2.9.2, and [7] uses Java 7, so I will do the same.

```
Fetched 24.2 MB in 4s (6133 kB/s)
Reading package lists... Done
EddieCFox@instance-1:~$ sudo apt-get install openjdk-7-jre
Reading package lists... Done
Building dependency tree
Reading state information... Done
Package openjdk-7-jre is not available, but is referred to by another package.
This may mean that the package is missing, has been obsoleted, or
is only available from another source

E: Package 'openjdk-7-jre' has no installation candidate
EddieCFox@instance-1:~$ ^C
EddieCFox@instance-1:~$ java -version

Command 'java' not found, but can be installed with:

apt install default-jre
apt install openjdk-11-jre-headless
apt install openjdk-8-jre-headless

Ask your administrator to install one of them.

EddieCFox@instance-1:~$ sudo apt-get install openjdk-7-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
Package openjdk-7-jdk is not available, but is referred to by another package.
This may mean that the package is missing, has been obsoleted, or
is only available from another source

E: Package 'openjdk-7-jdk' has no installation candidate
```

After running sudo apt-get update, I try the other two commands, but this does not work. Probably because Java 7 is considered obsolete and no longer offered by this.

I replace java 7 and java 8 for the openjdk jre and openjdk jdk commands and they both work after some processing.

```
EddieCFox@instance-1:~$ java -version
openjdk version "1.8.0_312"
OpenJDK Runtime Environment (build 1.8.0_312-8u312-b07-0ubuntu1~18.04-b07)
OpenJDK 64-Bit Server VM (build 25.312-b07, mixed mode)
EddieCFox@instance-1:~$
```

I confirm Java 8 is installed using java -version, whereas before this command didn't work. According to [8], Hadoop 2.9.X still supports Java 8, so this should be fine.

```
# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
-- INSERT --
```

I run vim ~/.bashrc as specified in [7], but change the 7 to an 8 to reflect the installation of java 8. I then save and exit VIM using escape and :wq and run source ~/.bashrc [7]. I confirm that java -version still works appropriately.

Openssh-server [7] is already installed. Rsync is already installed, but ssh is either not installed or needed to be updated [9].

I proceed with [7] as [9] already assumes Hadoop is downloaded and installed.

```
EddieCFox@instance-1:~$ sudo useradd -m hadoop -s /bin/bash
useradd: group hadoop exists - if you want to add this user to that group, use -g.
EddieCFox@instance-1:~$ sudo passwd hadoop
passwd: user 'hadoop' does not exist
EddieCFox@instance-1:~$ sudo adduser hadoop sudo
adduser: The user `hadoop' does not exist.
```

I try the commands in [7], but it fails as I already created a Hadoop group in tutorial 1, it seems.

From [10], I gather to add -g to specify adding into an already exiting group.

```
EddieCFox@instance-1:~$ sudo useradd -m hadoop -s /bin/bash -g hadoop
EddieCFox@instance-1:~$ sudo passwd hadoop
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
```

Per [11], I run "sudo apt-get install members" to install a package that will list the members of a group, then "members Hadoop" to confirm that the Hadoop user is a member of the Hadoop group.

```
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
EddieCFox@instance-1:~$ members hadoop
hduser hadoop EddieCFox
```

```
sudo apt-get install openssh-server
cd ~/.ssh/
ssh-keygen -t rsa
cat ./id_rsa.pub >> ./authorized_keys
```

I run the above code from [7] to set up another SSH key.

```
Last login: Wed Jan 19 05:46:44 2022 from 218.102.217.204
EddieCFox@instance-1:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:0Eold/qpvvKn05tqN/qbRaZWQ8eyvYDjkDzK/KlmN8Q.
Are you sure you want to continue connecting (yes/no)? y
Please type 'yes' or 'no': yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-1060-gcp x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Jan 19 06:44:04 UTC 2022

  System load:  0.0               Processes:           110
  Usage of /:   2.3% of 96.75GB   Users logged in:     1
  Memory usage: 4%                IP address for ens4: 10.140.0.2
  Swap usage:   0%


9 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

New release '20.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Wed Jan 19 06:43:00 2022 from 218.102.217.204
EddieCFox@instance-1:~$
```

I run "ssh localhost" in [7] to confirm I can ssh in.

I download Hadoop 2.9.2 from [12] per Assignment 0 PDF, adapting the "wget http://ftp.cuhk.edu.hk/pub/packages/apache.org/hadoop/common/hadoop-2.7.1/hadoop-

2.7.1.tar.gz" command from [7], using the https://archive.apache.org/dist/hadoop/core/hadoop-2.9.2/hadoop-2.9.2.tar.gz address from the directory in [12] for 2.9.2.

My final command is "wget https://archive.apache.org/dist/hadoop/core/hadoop-2.9.2/hadoop-2.9.2.tar.gz"

```
EddieCFox@instance-1:~$ wget https://archive.apache.org/dist/hadoop/core/hadoop-2.9.2/hadoop-2.9.2.tar.gz
--2022-01-19 06:50:30--  https://archive.apache.org/dist/hadoop/core/hadoop-2.9.2/hadoop-2.9.2.tar.gz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 366447449 (349M) [application/x-gzip]
Saving to: 'hadoop-2.9.2.tar.gz'

hadoop-2.9.2.tar.gz                 100%[===================================================================================>] 349.47M  9.04MB/s    in 36s

2022-01-19 06:51:07 (9.75 MB/s) - 'hadoop-2.9.2.tar.gz' saved [366447449/366447449]

EddieCFox@instance-1:~$ ls -a
.  ..  .bash_history  .bash_logout  .bashrc  .cache  .gnupg  .profile  .ssh  .viminfo  hadoop-2.9.2.tar.gz
EddieCFox@instance-1:~$
```

I then adapt sudo tar -zxf hadoop-2.7.1.tar.gz -C /usr/local from [7], changing the version number for a final command for "sudo tar -zxf hadoop-2.9.2.tar.gz -C /usr/local"

```
EddieCFox@instance-1:~$ ls -a
.  ..  .bash_history  .bash_logout  .bashrc  .cache  .gnupg  .profile  .ssh  .viminfo  hadoop-2.9.2.tar.gz
EddieCFox@instance-1:~$ sudo tar -zxf hadoop-2.9.2.tar.gz -C /usr/local
EddieCFox@instance-1:~$ cd /usr/local
EddieCFox@instance-1:/usr/local$ ls -a
.  ..  bin  etc  games  hadoop-2.9.2  include  lib  man  sbin  share  src
EddieCFox@instance-1:/usr/local$ sudo mv ./hadoop-2.9.2 ./hadoop
EddieCFox@instance-1:/usr/local$ ls -a
.  ..  bin  etc  games  hadoop  include  lib  man  sbin  share  src
EddieCFox@instance-1:/usr/local$ sudo chown -R hadoop ./hadoop
EddieCFox@instance-1:/usr/local$
```

I finish running the other commands in [7], changing the version number where relevant.

```
hadoop@instance-1:/usr/local/hadoop$ ./bin/hadoop version
Hadoop 2.9.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled by ajisaka on 2018-11-13T12:42Z
Compiled with protoc 2.5.0
From source with checksum 3a9939967262218aa556c684d107985
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.9.2.jar
hadoop@instance-1:/usr/local/hadoop$
```

Version confirmed by running "./bin/hadoop version" per [7] after running su Hadoop to change to the Hadoop user and changing to the usr/local/Hadoop directory.

```
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=123
        File Output Format Counters
                Bytes Written=23
hadoop@instance-1:/usr/local/hadoop$ ls
LICENSE.txt  NOTICE.txt  README.txt  bin  etc  include  input  lib  libexec  output  sbin  share
hadoop@instance-1:/usr/local/hadoop$ cat ./output/*
1       dfsadmin
hadoop@instance-1:/usr/local/hadoop$ rm -r ./output
hadoop@instance-1:/usr/local/hadoop$
```

I run the first map reduce job referenced in [7] and the Standalone operation section of [9] to confirm that standalone Hadoop cluster operation is functional. For the assignment, I now need to setup pseudo-distributed Hadoop cluster.

Assignment #0, Part A, Section I: Setting up Hadoop with Pseudo-distributed Single Node

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://localhost:9000</value>
        </property>
</configuration>
~
```

First I use vim to modify the etc/Hadoop/core-site.xml as directed in [9]. I initially get an error that it is read only, but realize this is because I am currently logged in as myself instead of the Hadoop user I gave ownership of Hadoop to. After using su Hadoop I am able to modify the file appropriately.

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
</configuration>
```

I use vim to modify etc/Hadoop/hdfs-site.xml as suggested in [9].

Since I believe the multi node Hadoop cluster setup involves YARN, I opt to do the Yarn pseudo distributed setup for single node Hadoop cluster. As a result, I also need to make configuration changes in etc/hadoop/mapred-site.xml and etc/hadoop/yarn-site.xml , using the configurations in [9] under “YARN on a Single Node”.

In Hadoop 2.9.2, there is not a mapred-site.xml in this folder, but there is a mapred-site.xml.template. Per [13], I can rename the template file to mapred-site.xml to proceed.

Continued on next page.

```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
</configuration>
~
~
```

```
-->
<configuration>

<!-- Site specific YARN configuration properties -->

        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>
~
```

Yarn-site.xml configuration above. I left the comment that was already included.

I now proceed through the first 4 steps in [9] for psudeodistributed setup under “Execution”, as the steps are the same as in the YARN case.

Continued on next page.

```
hadoop@instance-1:/usr/local/hadoop$ bin/hdfs namenode -format
22/01/20 03:54:21 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = instance-1.asia-east1-a.c.planar-ray-338702.internal/10.140.0.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.9.2
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common
/lib/commons-logging-1.1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/htrace-core4-4.1.0-i
ncubating.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-6.1.26.jar:/usr/local/hadoop/sh
are/hadoop/common/lib/commons-digester-1.8.jar:/usr/local/hadoop/share/hadoop/common/lib/cura
tor-client-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/us
r/local/hadoop/share/hadoop/common/lib/hamcrest-core-1.3.jar:/usr/local/hadoop/share/hadoop/c
ommon/lib/log4j-1.2.17.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-core-1.9.jar:/usr
/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.9.2.jar:/usr/local/hadoop/share/hadoop/co
mmon/lib/curator-framework-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/xmlenc-0.52.ja
r:/usr/local/hadoop/share/hadoop/common/lib/jackson-jaxrs-1.9.13.jar:/usr/local/hadoop/share/
hadoop/common/lib/slf4j-log4j12-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-
configuration-1.6.jar:/usr/local/hadoop/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar:/u
sr/local/hadoop/share/hadoop/common/lib/curator-recipes-2.7.1.jar:/usr/local/hadoop/share/had
oop/common/lib/apacheds-i18n-2.0.0-M15.jar:/usr/local/hadoop/share/hadoop/common/lib/stax2-ap
i-3.1.4.jar:/usr/local/hadoop/share/hadoop/common/lib/httpcore-4.4.4.jar:/usr/local/hadoop/sh
are/hadoop/common/lib/jets3t-0.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jsch-0.1.54.
jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop/share/had
oop/common/lib/commons-net-3.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-codec-1.
4.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/usr/local/hadoop/share
/hadoop/common/lib/paranamer-2.3.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-server-
1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-util-6.1.26.jar:/usr/local/hadoop/sha
re/hadoop/common/lib/jetty-sslengine-6.1.26.jar:/usr/local/hadoop/share/hadoop/common/lib/api
-util-1.0.0-M20.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-annotations-2.9.2.jar:/u
sr/local/hadoop/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/local/hadoop/share/hadoop/com
mon/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jettison-1
.1.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/shar
e/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop/share/hadoop/common/lib/jsp-api-2.
1.jar:/usr/local/hadoop/share/hadoop/common/lib/zookeeper-3.4.6.jar:/usr/local/hadoop/share/h
adoop/common/lib/commons-compress-1.4.1.jar:/usr/local/hadoop/share/hadoop/common/lib/json-sm
art-1.3.1.jar:/usr/local/hadoop/share/hadoop/common/lib/asm-3.2.jar:/usr/local/hadoop/share/h
adoop/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/usr/local/hadoop/share/hadoop/common/
lib/junit-4.11.jar:/usr/local/hadoop/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/loc
al/hadoop/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop/share/hadoop/
```

Continued on next page.

```
22/01/20 03:54:21 INFO util.GSet: VM type       = 64-bit
22/01/20 03:54:21 INFO util.GSet: 1.0% max memory 889 MB = 8.9 MB
22/01/20 03:54:21 INFO util.GSet: capacity      = 2^20 = 1048576 entries
22/01/20 03:54:21 INFO namenode.FSDirectory: ACLs enabled? false
22/01/20 03:54:21 INFO namenode.FSDirectory: XAttrs enabled? true
22/01/20 03:54:21 INFO namenode.NameNode: Caching file names occurring more than 10 times
22/01/20 03:54:21 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: falseskipCap
tureAccessTimeOnlyChange: false
22/01/20 03:54:21 INFO util.GSet: Computing capacity for map cachedBlocks
22/01/20 03:54:21 INFO util.GSet: VM type       = 64-bit
22/01/20 03:54:21 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
22/01/20 03:54:21 INFO util.GSet: capacity      = 2^18 = 262144 entries
22/01/20 03:54:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets =
10
22/01/20 03:54:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
22/01/20 03:54:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5
,25
22/01/20 03:54:21 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
22/01/20 03:54:21 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and ret
ry cache entry expiry time is 600000 millis
22/01/20 03:54:21 INFO util.GSet: Computing capacity for map NameNodeRetryCache
22/01/20 03:54:21 INFO util.GSet: VM type       = 64-bit
22/01/20 03:54:21 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
22/01/20 03:54:21 INFO util.GSet: capacity      = 2^15 = 32768 entries
22/01/20 03:54:21 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1001498076-10.140.0.2-
1642650861802
22/01/20 03:54:21 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/name has been
 successfully formatted.
22/01/20 03:54:21 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-hadoop/d
fs/name/current/fsimage.ckpt_0000000000000000000 using no compression
22/01/20 03:54:21 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-hadoop/dfs/name
/current/fsimage.ckpt_0000000000000000000 of size 325 bytes saved in 0 seconds .
22/01/20 03:54:21 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid
 >= 0
22/01/20 03:54:21 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at instance-1.asia-east1-a.c.planar-ray-338702.internal/
10.140.0.2
************************************************************/
hadoop@instance-1:/usr/local/hadoop$
```

Lots of text after running “bin/hdfs namenode -format”. Not all text included.

```
hadoop@instance-1:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [localhost]
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:0Eold/qpvvKn05tqN/qbRaZWQ8eyvYDjkDzK/KlmN8Q.
Are you sure you want to continue connecting (yes/no)? y
Please type 'yes' or 'no': yes
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: hadoop@localhost: Permission denied (publickey).
localhost: hadoop@localhost: Permission denied (publickey).
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
```

I receive an error related to publickey when running “sbin-start.dfs.sh”. I realize I ran SSH commands to give a key for the username when I log in, but not for Hadoop, so I run the commands suggested in [9] if you can’t ssh into localhost.

```
hadoop@instance-1:/usr/local/hadoop$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:ClQpG3/wXanQhooM3CRo/a1H8807I1G6MePQNDPJCEw hadoop@instance-1
The key's randomart image is:
+---[RSA 2048]----+
| ooo+E.. o   .   |
|..oo++o o o o    |
|.  oo*o+o+.o     |
|   .+.o=oBo.     |
|    . o.S O      |
|     o + O o     |
|      o o * .    |
|         + +     |
|          . o    |
+----[SHA256]-----+
hadoop@instance-1:/usr/local/hadoop$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@instance-1:/usr/local/hadoop$ chmod 0600 ~/.ssh/authorized_keys
hadoop@instance-1:/usr/local/hadoop$
```

```
hadoop@instance-1:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: Error: JAVA_HOME is not set and could not be found.
```

When running “sbin/start-dfs.sh”, the next error I get is that Java_Home is not set and could not be found. Again, this is because I did this process on the EddieCFox username and not Hadoop, so I do it again according to [7], modifying .bashrc to include “export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64”, then running “source ~/.bashrc” to reload bash.

I still received the error afterwards. Per [14], this can be resolved by modifying /etc/hadoop/hadoop-env.sh from “export JAVA_HOME=${JAVA_HOME}” to “export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64”

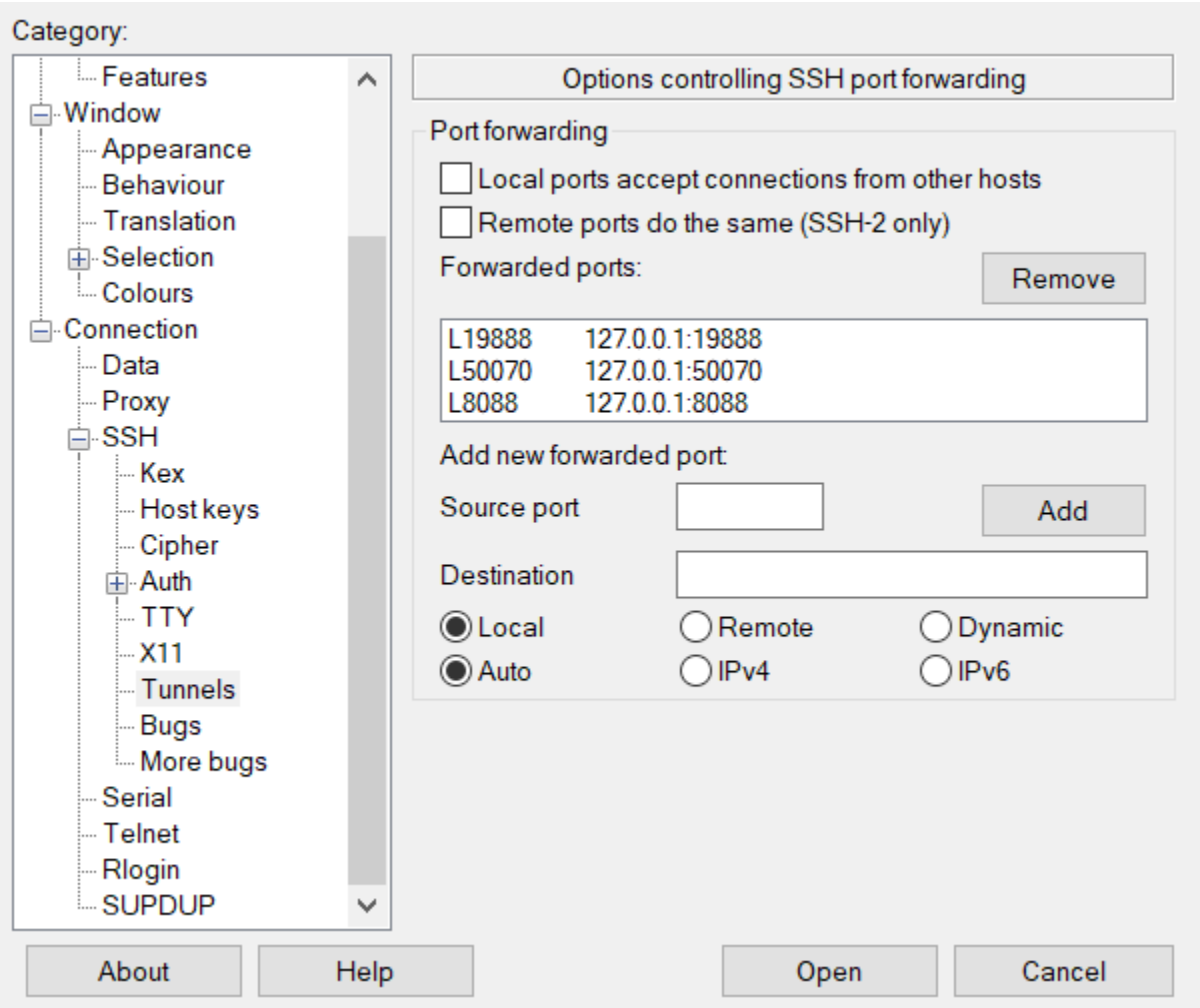
```
hadoop@instance-1:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-namenode-instan
ce-1.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-instan
ce-1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-secondar
ynamenode-instance-1.out
hadoop@instance-1:/usr/local/hadoop$
```

Finally, “sbin/start-dfs.sh” works.

```
hadoop@instance-1:/usr/local/hadoop$ jps
7575 SecondaryNameNode
7352 DataNode
7165 NameNode
7742 Jps
```

Jps command confirms the namenode and datanodes are working appropriately, however I need to view localhost:50070 for the assignment, which initially proves difficult because my virtual machine has no GPU, and I am only connected to it through SSH.

Using the methodology in [15], I stop the namenode and datanode [9], create SSH tunnels to localhost in Putty for ports 50070 (namenode) 8088 (yarn resource manager), and 19888 (MapReduce JobHistory Server) [16], then log in again and restart the name node.



This works successfully.

Below is the printout from http://localhost:50070 as requested by assignment 0.

Assignment #0 Part A, Section I: localhost:50070 printout

localhost:50070/dfshealth.html#tab-overview

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Overview 'localhost:9000' (active)

Started:	Thu Jan 20 13:01:12 +0800 2022
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-95604e28-4743-431d-8269-d6f87f68bf4c
Block Pool ID:	BP-731876950-10.140.0.2-1642651882866

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 79.87 MB of 186 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 39.3 MB of 40.31 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	96.75 GB
DFS Used:	28 KB (0%)

DFS Used:	28 KB (0%)

Non DFS Used:	4.09 GB
DFS Remaining:	92.64 GB (95.76%)
Block Pool Used:	28 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Thu Jan 20 13:01:12 +0800 2022
Last Checkpoint Time	Thu Jan 20 13:02:23 +0800 2022

NameNode Journal Status

Current transaction ID: 6

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-hadoop/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-hadoop/dfs/name/current/edits_inprogress_0000000000000000006)

NameNode Storage

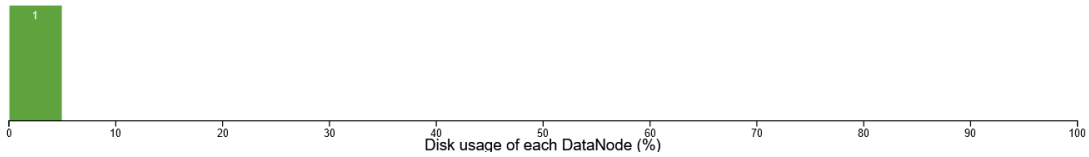
Storage Directory	Type	State
/tmp/hadoop-hadoop/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	96.75 GB	28 KB (0%)	92.64 GB (95.76%)	28 KB	1

Hadoop, 2018.

Datanode usage histogram



In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✔instance-1.asia-east1-a.c.planar-ray-338702.internal:50010 (127.0.0.1:50010)	http://instance-1.asia-east1-a.c.planar-ray-338702.internal:50075	2s	2m	96.75 GB	0	28 KB (0%)	2.9.2

Showing 1 to 1 of 1 entries

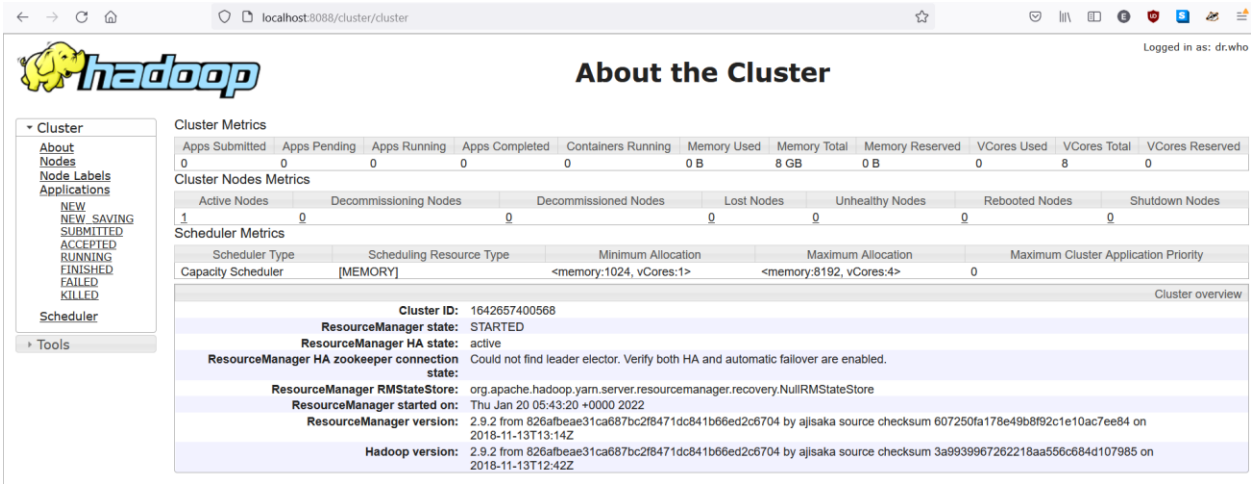
Previous 1 Next

Assignment #0 Part A, Section II: Teragen, terasort, teravalidate

```
hadoop@instance-1:/usr/local/hadoop$ bin/hdfs dfs -mkdir /user
hadoop@instance-1:/usr/local/hadoop$ bin/hdfs dfs -mkdir /user/hadoop
```

I perform Step 4 of [9] to make directories in the HDFS system necessary for mapreduce jobs. I then follow the steps for setting up a YARN node, besides the configuration files which I already edited.

I run “sbin/start-yarn.sh” and can view it on http://localhost:8088/ [9]




```
hadoop@instance-1:/usr/local/hadoop$ jps
9762 SecondaryNameNode
10774 NodeManager
10523 ResourceManager
11035 Jps
9532 DataNode
9340 NameNode
```

Running jps properly shows all processes running, so I can proceed with the Teragen job.

I run the following command, as shown in Assignment 0 PDF: "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 100000 terasort/input"

```
hadoop@instance-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 100000 terasort/input
22/01/20 05:54:37 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/20 05:54:38 INFO terasort.TeraGen: Generating 100000 using 2
22/01/20 05:54:38 INFO mapreduce.JobSubmitter: number of splits:2
22/01/20 05:54:38 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/20 05:54:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642657400568_0001
22/01/20 05:54:39 INFO impl.YarnClientImpl: Submitted application application_1642657400568_0001
22/01/20 05:54:39 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1642657400568_0001/
22/01/20 05:54:39 INFO mapreduce.Job: Running job: job_1642657400568_0001
22/01/20 05:54:46 INFO mapreduce.Job: Job job_1642657400568_0001 running in uber mode : false
22/01/20 05:54:46 INFO mapreduce.Job:  map 0% reduce 0%
22/01/20 05:54:53 INFO mapreduce.Job:  map 50% reduce 0%
22/01/20 05:54:54 INFO mapreduce.Job:  map 100% reduce 0%
22/01/20 05:54:54 INFO mapreduce.Job: Job job_1642657400568_0001 completed successfully
22/01/20 05:54:54 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396188
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=164
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=8876
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=8876
                Total vcore-milliseconds taken by all map tasks=8876
                Total megabyte-milliseconds taken by all map tasks=9089024
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Input split bytes=164
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=218
                CPU time spent (ms)=1720
                Physical memory (bytes) snapshot=380260352
                Virtual memory (bytes) snapshot=3866890240
                Total committed heap usage (bytes)=207618048
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=214574985129000
        File Input Format Counters
```

```
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=10000000
hadoop@instance-1:/usr/local/hadoop$
```

I then run the Terasort command, also in the Assignment 0 PDF:

"./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/input terasort/output"

```
hadoop@instance-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/input terasort/output
22/01/20 06:01:28 INFO terasort.TeraSort: starting
22/01/20 06:01:29 INFO input.FileInputFormat: Total input files to process : 2
Spent 109ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
Computing input splits took 112ms
Sampling 2 splits of 2
Making 1 from 100000 sampled records
Computing paritition took 501ms
Spent 616ms computing partitions.
22/01/20 06:01:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/20 06:01:30 INFO mapreduce.JobSubmitter: number of splits:2
22/01/20 06:01:30 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/20 06:01:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642657400568_0002
22/01/20 06:01:31 INFO impl.YarnClientImpl: Submitted application application_1642657400568_0002
22/01/20 06:01:31 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1642657400568_0002/
22/01/20 06:01:31 INFO mapreduce.Job: Running job: job_1642657400568_0002
22/01/20 06:01:37 INFO mapreduce.Job: Job job_1642657400568_0002 running in uber mode : false
22/01/20 06:01:37 INFO mapreduce.Job:  map 0% reduce 0%
22/01/20 06:01:45 INFO mapreduce.Job:  map 100% reduce 0%
22/01/20 06:01:50 INFO mapreduce.Job:  map 100% reduce 100%
22/01/20 06:01:51 INFO mapreduce.Job: Job job_1642657400568_0002 completed successfully
22/01/20 06:01:51 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=10400006
                FILE: Number of bytes written=21398487
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000252
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=12004
                Total time spent by all reduces in occupied slots (ms)=3143
                Total time spent by all map tasks (ms)=12004
                Total time spent by all reduce tasks (ms)=3143
                Total vcore-milliseconds taken by all map tasks=12004
                Total vcore-milliseconds taken by all reduce tasks=3143
                Total megabyte-milliseconds taken by all map tasks=12292096
                Total megabyte-milliseconds taken by all reduce tasks=3218432
        Map-Reduce Framework
                Map input records=100000
```

```
                Total megabyte-milliseconds taken by all reduce tasks=3218432
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Map output bytes=10200000
                Map output materialized bytes=10400012
                Input split bytes=252
                Combine input records=0
                Combine output records=0
                Reduce input groups=100000
                Reduce shuffle bytes=10400012
                Reduce input records=100000
                Reduce output records=100000
                Spilled Records=200000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=379
                CPU time spent (ms)=4350
                Physical memory (bytes) snapshot=779145216
                Virtual memory (bytes) snapshot=5781962752
                Total committed heap usage (bytes)=507510784
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=10000000
        File Output Format Counters
                Bytes Written=10000000
22/01/20 06:01:51 INFO terasort.TeraSort: done
hadoop@instance-1:/usr/local/hadoop$
```

Finally, I run the teravalidate command, from Assignment 0 PDF. "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/output terasort/check"

Continued on next page.

```
hadoop@instance-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/output terasort/check
22/01/20 06:13:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/20 06:13:59 INFO input.FileInputFormat: Total input files to process : 1
Spent 21ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
22/01/20 06:13:59 INFO mapreduce.JobSubmitter: number of splits:1
22/01/20 06:13:59 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/20 06:13:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642657400568_0003
22/01/20 06:13:59 INFO impl.YarnClientImpl: Submitted application application_1642657400568_0003
22/01/20 06:13:59 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1642657400568_0003/
22/01/20 06:13:59 INFO mapreduce.Job: Running job: job_1642657400568_0003
22/01/20 06:14:05 INFO mapreduce.Job: Job job_1642657400568_0003 running in uber mode : false
22/01/20 06:14:05 INFO mapreduce.Job:  map 0% reduce 0%
22/01/20 06:14:11 INFO mapreduce.Job:  map 100% reduce 0%
22/01/20 06:14:15 INFO mapreduce.Job:  map 100% reduce 100%
22/01/20 06:14:16 INFO mapreduce.Job: Job job_1642657400568_0003 completed successfully
22/01/20 06:14:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=397195
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000127
                HDFS: Number of bytes written=22
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2806
                Total time spent by all reduces in occupied slots (ms)=2559
                Total time spent by all map tasks (ms)=2806
                Total time spent by all reduce tasks (ms)=2559
                Total vcore-milliseconds taken by all map tasks=2806
                Total vcore-milliseconds taken by all reduce tasks=2559
                Total megabyte-milliseconds taken by all map tasks=2873344
                Total megabyte-milliseconds taken by all reduce tasks=2620416
        Map-Reduce Framework
                Map input records=100000
                Map output records=3
                Map output bytes=80
                Map output materialized bytes=92
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
```

```
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Map output bytes=10200000
                Map output materialized bytes=10400012
                Input split bytes=252
                Combine input records=0
                Combine output records=0
                Reduce input groups=100000
                Reduce shuffle bytes=10400012
                Reduce input records=100000
                Reduce output records=100000
                Spilled Records=200000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=379
                CPU time spent (ms)=4350
                Physical memory (bytes) snapshot=779145216
                Virtual memory (bytes) snapshot=5781962752
                Total committed heap usage (bytes)=507510784
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=10000000
        File Output Format Counters
                Bytes Written=10000000
22/01/20 06:01:51 INFO terasort.TeraSort: done
hadoop@instance-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/output terasort/check
22/01/20 06:13:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/20 06:13:59 INFO input.FileInputFormat: Total input files to process : 1
Spent 21ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
22/01/20 06:13:59 INFO mapreduce.JobSubmitter: number of splits:1
22/01/20 06:13:59 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/20 06:13:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642657400568_0003
22/01/20 06:13:59 INFO impl.YarnClientImpl: Submitted application application_1642657400568_0003
22/01/20 06:13:59 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1642657400568_0003/
22/01/20 06:13:59 INFO mapreduce.Job: Running job: job_1642657400568_0003
22/01/20 06:14:05 INFO mapreduce.Job: Job job_1642657400568_0003 running in uber mode : false
22/01/20 06:14:05 INFO mapreduce.Job:  map 0% reduce 0%
22/01/20 06:14:11 INFO mapreduce.Job:  map 100% reduce 0%
22/01/20 06:14:15 INFO mapreduce.Job:  map 100% reduce 100%
22/01/20 06:14:16 INFO mapreduce.Job: Job job_1642657400568_0003 completed successfully
22/01/20 06:14:16 INFO mapreduce.Job: Counters: 49
```

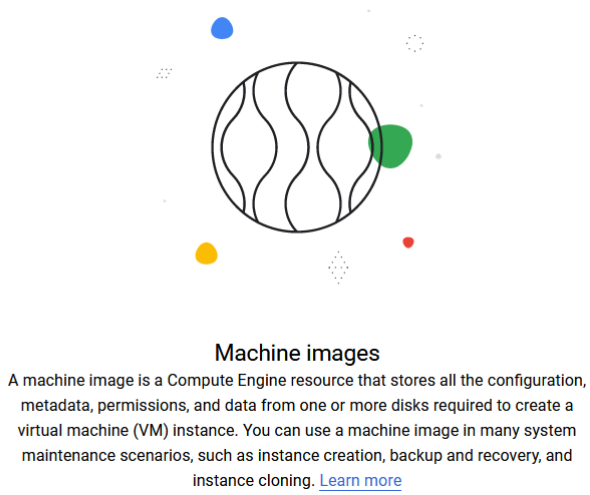
```
22/01/20 06:14:16 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=397195
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000127
                HDFS: Number of bytes written=22
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2806
                Total time spent by all reduces in occupied slots (ms)=2559
                Total time spent by all map tasks (ms)=2806
                Total time spent by all reduce tasks (ms)=2559
                Total vcore-milliseconds taken by all map tasks=2806
                Total vcore-milliseconds taken by all reduce tasks=2559
                Total megabyte-milliseconds taken by all map tasks=2873344
                Total megabyte-milliseconds taken by all reduce tasks=2620416
        Map-Reduce Framework
                Map input records=100000
                Map output records=3
                Map output bytes=80
                Map output materialized bytes=92
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=92
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=182
                CPU time spent (ms)=1570
                Physical memory (bytes) snapshot=480858112
                Virtual memory (bytes) snapshot=3869270016
                Total committed heap usage (bytes)=301465600
        Shuffle Errors
```

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=10000000
        File Output Format Counters
                Bytes Written=22
hadoop@instance-1:/usr/local/hadoop$
```

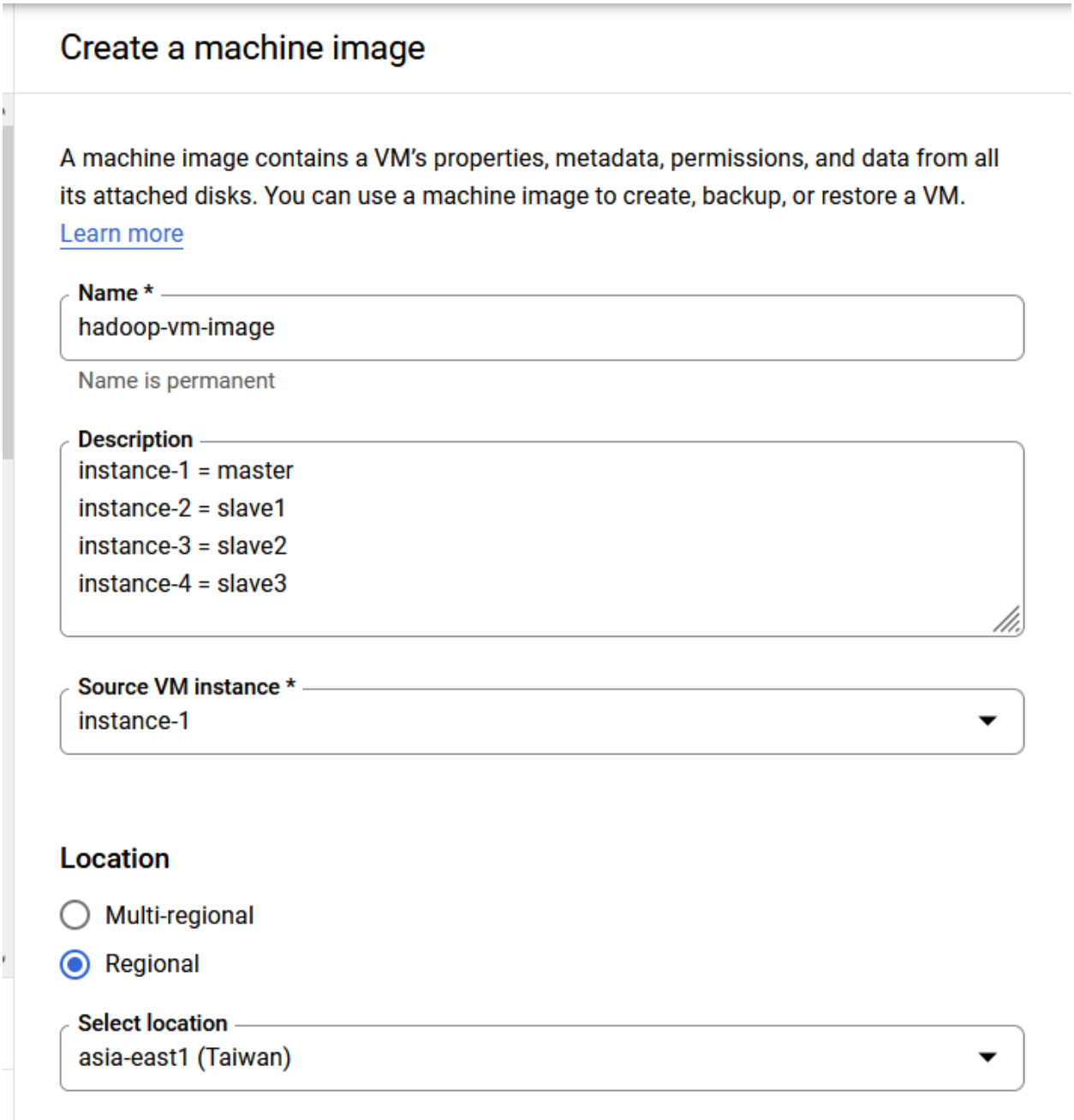
Everything is functioning correctly with Terasort on pseudo-distributed single node Hadoop.

Assignment #0, Part B, Section I Multi-node Hadoop Cluster Setup (4 VMs)

As the single node is working correctly, I need to create an image of my instance-1 and create 3 more VMs from it.



I go to “Create machine image” from Virtual machines under Compute Engine



You can also see my planned naming scheme in the description field of the screenshot above. I wanted to rename instance-1, but it seems you need to create a new instance using the original instance as a source, and I don't want to do that in case it introduces bugs, so I will just name them instance-1, instance-2, instance-3, and instance-4, with the other 3 being the slaves.

	Status	Name	Source instance	Machine type	Storage location	Creation time
☐	✔	hadoop-vm-image	instance-1	n2d-standard-2	asia-east1	Jan 21, 2022, 10:27:53 AM UTC+08:00

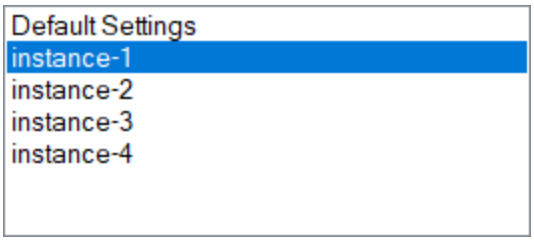
I then go to actions on the right of creation time and click "Create instance", filling out the details 3 times for 3 slave VMs. I do not much customization besides changing the region to asia-

east1 and the zone to 1a to be consistent with the master VM. This should also minimize the loss of performance from communicating across the world. I only realize on instance-4 that I didn't enable display device on instances 2 and 3, but it should be fine as I am only capturing screenshots from the master instance.

	Status	Name ↑	Zone
☐	✓	instance-1	asia-east1-a
☐	✓	instance-2	asia-east1-a
☐	✓	instance-3	asia-east1-a
☐	✓	instance-4	asia-east1-a

Internal IP	External IP
10.140.0.2 (nic0)	35.229.202.64
10.140.0.3 (nic0)	35.236.184.18
10.140.0.4 (nic0)	35.194.186.251
10.140.0.5 (nic0)	34.81.23.107

While I use the external IP address for connecting with Putty, I believe that I use the internal IP addresses when modifying the Hadoop configurations because they are being treated as on the same network internally, which also allows them to connect with each other.



I create 3 more sessions in Putty for the other 3 VMs, ensuring that all the details are the same except the IP address I am connecting to.

```
EddieCFox@instance-2:~$ cd /
EddieCFox@instance-2:/$ cd usr/local
EddieCFox@instance-2:/usr/local$ ls
bin  etc  games  hadoop  include  lib  man  sbin  share  src
EddieCFox@instance-2:/usr/local$
```

I ensure that I can log into each VM with Putty and that Hadoop is installed.

After doing this, I realize I should modify my putty configuration so that only instance-1 (the master) has SSH tunneling for the web interfaces of name node and resource manager. It could lead to potential bugs if there is tunnels created by every instance for the same ports, even when they have nothing to contribute.

As suggested in [17] I modify the /etc/hostname file in each VM to have the name that I later intend to have for them in the hosts files, i.e. renaming instance-1 to master, instance-2 to slave1, etc. The hostname file is read only, so I need to use sudo to edit it.

```
EddieCFox@instance-1:/etc$ sudo vim hostname
EddieCFox@instance-1:/etc$ cat hostname
master
EddieCFox@instance-1:/etc$ hostname
instance-1
```

Even after doing so, the hostname doesn't change when I use the hostname command. I assume it may only take effect after restarting, so before I do that, I also modify the hosts file.

I modify the hosts file for each VM according to [18] so that each VM is aware of the other VMs in the Hadoop cluster by IP address. I need to later designate them specifically as slaves and masters in other configuration files.

EddieCFox@instance-1:/etc

```
127.0.0.1 localhost
10.140.0.2 master
10.140.0.3 slave1
10.140.0.4 slave2
10.140.0.5 slave3

# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
169.254.169.254 metadata.google.internal metadata
```

After editing the hosts and hostname file on instance-1, I consult [19] from Google to determine I should restart it using the stop and start operations on the Google Compute Engine Console.

```
EddieCFox@master:~$ hostname
master
```

After restarting the VM, the hostname changes appropriately. I then proceed to make the same changes in the hostname and hosts file of instances 2-4, naming them slave1-3, and restart the VMs to confirm the hostname changes. After restarting the VMs, the external IP addresses for all of them changes except for master, so I needed to change the session configuration in Putty. While doing this, I also took the opportunity to change the name from instance1-4 to master and slave1-3.

[16] notes that ssh trusts are required to run the start and stop shell scripts across all the machines, and that by ssh trusts, they mean the ability to do passwordless ssh between the servers. This is what changing the hostnames, adding every VM into the hosts files, and taking care of the SSH keys on the master VM before cloning it was meant to achieve.

As suggested in [18], I made sure that the VMs could communicate with each other by going into each VM and attempting to SSH into all of the other ones. For example, from master I would type “ssh slave 1” “ssh slave2” “ssh slave3”, from slave1 I would type “ssh master” “ssh slave2” “ssh slave3”. In each attempt, I needed to type yes to accept the key, but afterwards, everything was accepted and I could smoothly SSH from any VM to all other VMs.

Based on [16], I knew I would need to modify the fs.defaultFS property in core-site.xml to point to master instead of local, since I changed all the hostnames / host files. [18] suggested to add a temporary directory, so I did that as well. I am making these changes on master VM, and then I will copy it to the slaves through the network.

Core-site.xml changes below.

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://master:9000</value>
        </property>

        <property>
                <name>hadoop.tmp.dir</name>
                <value>file:/usr/local/hadoop/tmp</value>
        </property>
</configuration>
```

```
<configuration>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>

        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/usr/local/hadoop/hadoop_data/hdfs/namenode</value>
        </property>

        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/usr/local/hadoop/hadoop_data/hdfs/datanode</value>
        </property>
</configuration>
```

Above is my changes to hdfs-site.xml. I already had dfs.replication at 1 because I don't need any redundancy and want to reduce the runtime instead of storing copies in 3 machines. [18] suggested adding specifiers for the namenode and datanode directories, which makes sense because there needs to be coordination in the HDFS across multiple machines instead of everything being self contained in one node in pseudo distributed mode.

```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>

        <property>
                <name>mapreduce.jobtracker.address</name>
                <value>master:5431</value>
        </property>
</configuration>
```

In mapred-site.xml above, I already had the mapreduce framework line since I opted to run the single node in YARN mode, but I am now adding mapreduce.jobtracker.address as master:5431 per [18]. This parameter was also mentioned in [16].

```
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>

        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>master</value>
```

In yarn-site.xml above, I already had the aux-services line, but as [16] mentions, it is a good idea to specify the hostname for the resourcemanager since this a multi-node setup.

Using [20], I copy the configuration file to the slaves, just replacing everything in the /usr/local/hadoop/etc directory of the slaves with what is in the master.

I initially encounter errors with denied permission, and I realize this is probably because I am logged in with my EddieCFox account currently rather than Hadoop, which has ownership of the files. After using su to log in as Hadoop, the scp works appropriately.

Typical scp command: “scp -r etc/Hadoop hadoop@slaveX:/usr/local/hadoop/etc” from the /usr/local/hadoop directory.

```
hadoop@master:/usr/local/hadoop$ scp -r etc/hadoop hadoop@slave1:/usr/local/hadoop/etc
The authenticity of host 'slave1 (10.140.0.3)' can't be established.
ECDSA key fingerprint is SHA256:Zl8XD8Ws7peNyIWDh+oZfdEXtRb+iICb8tHGLEzZl14.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'slave1,10.140.0.3' (ECDSA) to the list of known hosts.
httpfs-log4j.properties                                   100% 1657   618.1KB/s   00:00
mapred-queues.xml.template                                100% 4113     2.0MB/s   00:00
hadoop-metrics.properties                                 100% 2490     1.3MB/s   00:00
hdfs-site.xml                                             100% 1107   943.8KB/s   00:00
kms-site.xml                                              100% 5939     3.6MB/s   00:00
kms-log4j.properties                                      100% 1788     1.0MB/s   00:00
core-site.xml                                             100%  963     1.0MB/s   00:00
kms-acls.xml                                              100% 3518     2.0MB/s   00:00
container-executor.cfg                                    100% 1211   869.2KB/s   00:00
hadoop-env.cmd                                            100% 4133     3.2MB/s   00:00
kms-env.sh                                                100% 3139     2.7MB/s   00:00
hadoop-metrics2.properties                                100% 2598     2.0MB/s   00:00
httpfs-site.xml                                           100%  620   503.6KB/s   00:00
mapred-site.xml                                           100%  943     1.1MB/s   00:00
yarn-env.cmd                                              100% 2250     1.6MB/s   00:00
ssl-server.xml.example                                    100% 2697     2.0MB/s   00:00
slaves                                                    100%   10     7.9KB/s   00:00
ssl-client.xml.example                                    100% 2316     2.1MB/s   00:00
httpfs-env.sh                                             100% 2230     1.7MB/s   00:00
configuration.xsl                                         100% 1335     1.2MB/s   00:00
capacity-scheduler.xml                                    100% 7861     4.4MB/s   00:00
hadoop-env.sh                                             100% 4990     4.8MB/s   00:00
log4j.properties                                          100%   14KB   8.2MB/s   00:00
yarn-site.xml                                             100%  890   967.4KB/s   00:00
mapred-env.sh                                             100% 1507     1.1MB/s   00:00
yarn-env.sh                                               100% 4876     4.0MB/s   00:00
hadoop-policy.xml                                         100%   10KB   8.7MB/s   00:00
mapred-env.cmd                                            100% 1076   877.8KB/s   00:00
httpfs-signature.secret                                   100%   21    19.0KB/s   00:00
hadoop@master:/usr/local/hadoop$
```

I use cat on each individual slave to confirm that the configuration files copied over appropriately.

Based on [16], [17], and [18], I now create masters and slaves files on the master node. This is located at /usr/local/hadoop/etc/Hadoop. There is already a slaves file but not master, so I create one. [18] says to use master while [17] uses masters, so I do both, create new files with vim with only the word master, referring to the hostname.

```
hadoop@master:/usr/local/hadoop/etc/hadoop$ cat master
master
hadoop@master:/usr/local/hadoop/etc/hadoop$ cat masters
master
```

When using vim to enter the slaves file, I see that it already is configured with the word "localhost". I take this to mean that by default, any Hadoop installation will be treated as a slave unless it is explicitly specified as a master.

I erase localhost and put slave1, 2, and 3 [18].

```
hadoop@master:/usr/local/hadoop/etc/hadoop$ cat slaves
slave1
slave2
slave3
```

After all this configuration, I can begin executing. Using [16], I run the namenode format command again after navigating back to the /usr/local/Hadoop directory.

"bin/hdfs namenode -format"

```
22/01/21 05:14:01 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid
 >= 0
22/01/21 05:14:01 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at master/10.140.0.2
************************************************************/
```

Most of messages omitted, just kept the end, which is the same as before, but now it explicitly specifies master as the hostname and the internal IP address matches up, so I proceed.

Per [16], because ssh trusted access is configured, I can now use start-dfs.sh and start-yarn.sh to start the appropriate processes on each machine.

"sbin/start-dfs.sh"

```
hadoop@master:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [master]
The authenticity of host 'master (10.140.0.2)' can't be established.
ECDSA key fingerprint is SHA256:0Eold/qpvvKn05tqN/qbRaZWQ8eyvYDjkDzK/KlmN8Q.
Are you sure you want to continue connecting (yes/no)? yes
master: Warning: Permanently added 'master,10.140.0.2' (ECDSA) to the list of known hosts.
master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-namenode-master.ou
t
slave3: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-slave3.ou
t
slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-slave1.ou
t
slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-slave2.ou
t
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-secondar
ynamenode-master.out
hadoop@master:/usr/local/hadoop$
```

After a brief yes command to add master to the list of known hosts, start-dfs.sh is able to start the namenode and secondary namenode on master, and datanodes on slaves1-3.

Finally, I start the yarn processes using start-yarn.sh [16]. “sbin/start-yarn.sh”

```
hadoop@master:/usr/local/hadoop$ sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-resourcemanager-master.out
slave3: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-slave3.out
slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-slave1.out
slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodemanager-slave2.out
hadoop@master:/usr/local/hadoop$
```

```
hadoop@master:/usr/local/hadoop$ jps
4256 SecondaryNameNode
4691 Jps
4003 NameNode
4421 ResourceManager
hadoop@master:/usr/local/hadoop$
```

Using JPS, I can confirm namenode, secondary namenode, and resourcemanager are running on the master machine.

```
EddieCFox@slave1:~$ jps
4425 Jps
EddieCFox@slave1:~$ su hadoop
Password:
hadoop@slave1:/home/EddieCFox$ jps
4448 Jps
4258 NodeManager
4042 DataNode
hadoop@slave1:/home/EddieCFox$
```

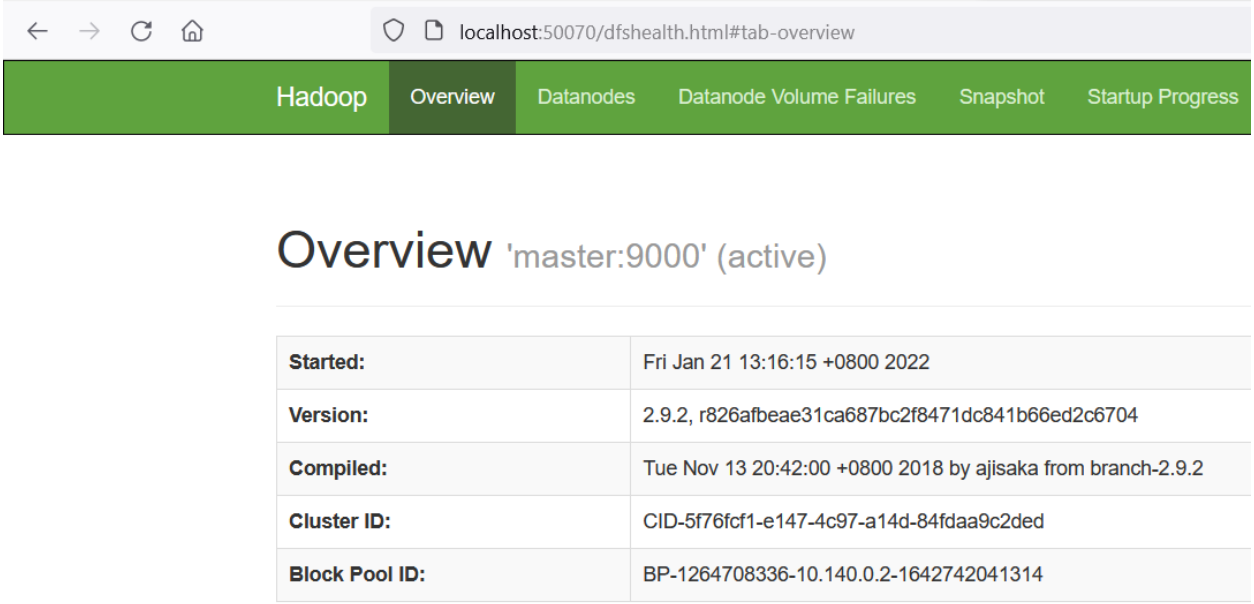
I initially despair with slave1 because jps only yielded jps, but that was because I was not logged in with Hadoop on slave1. After using su to switch to Hadoop, jps properly shows nodemanager and datanode.

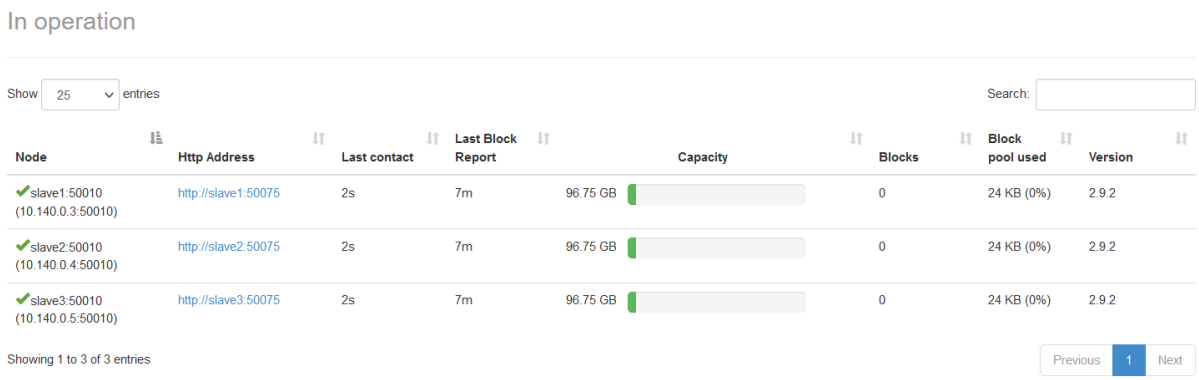
```
hadoop@slave2:/home/EddieCFox$ jps
3968 DataNode
4296 Jps
4142 NodeManager
hadoop@slave2:/home/EddieCFox$
```

```
hadoop@slave3:/home/EddieCFox$ jps
4275 Jps
4106 NodeManager
3932 DataNode
hadoop@slave3:/home/EddieCFox$
```

Slave2 and Slave3 are similarly successful.

I can also confirm that the localhost:50070 website is working and showing the datanodes appropriately.

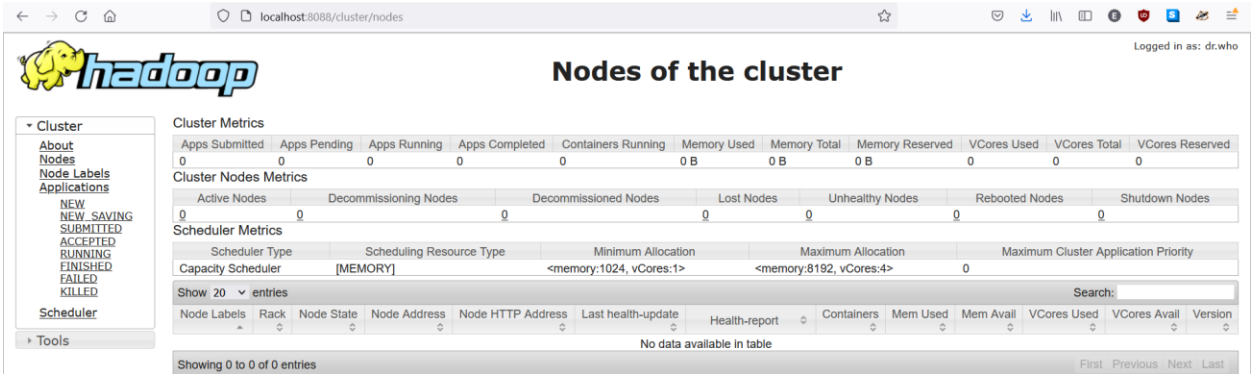




Although localhost:50070 was working, localhost:8088 for resource manager was not. I did research, and [21] says jobtracker no longer exists in YARN, so I went back to the configuration to delete jobtracker address from mapred-site.xml. Using [16] and [22], I looked at the default hostname for resource manager, and it was 0.0.0.0, so I changed the hostname in yarn-site.xml from master to 0.0.0.0. Afterwards, I copied the changes over to slaves 1-3 using scp.

Finally, Resource manager was working successfully.

Note: I realized later on, that even though the resource manager page was showing up in my browser, it was not configured correctly. I address that later.



Assignment #0, Part B, Section II: Multi-node Hadoop Cluster Terasortting (2GB and 20GB datasets)

For this part, we run the same command as we did in part 1, but change the parameter. Terasort uses 100 bytes per row. We can tell this because when we first ran it, we specified 100,000 as the parameter, and the output said 10,000,000 bytes were written in teragen.

2 GIB is $2^{30} * 2$ = 2,147,483,648, we then divide by 100 to get 21,474,837 (rounding up to be safe) lines that Terasort needs to do.

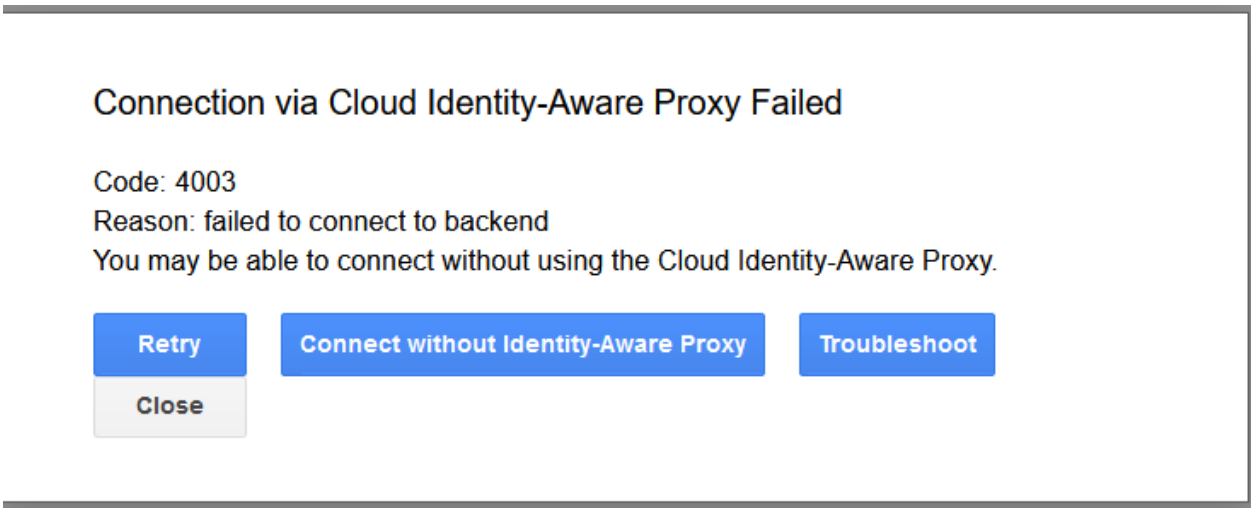
Besides input, I will also change the name of the output of teragen to 2GB-input.

The final teragen command is "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 21474837 terasort/2GB-input"

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-e
xamples-2.9.2.jar teragen 21474837 terasort/2GB-input
22/01/21 06:24:53 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/01/21 06:24:54 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /tmp/hadoop-yarn/staging/had
oop/.staging/job_1642745469275_0001/job.jar could only be replicated to 0 nodes instead of mi
nReplication (=1).  There are 0 datanode(s) running and no node(s) are excluded in this opera
tion.
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(B
lockManager.java:1814)
        at org.apache.hadoop.hdfs.server.namenode.FSDirWriteFileOp.chooseTargetForNewBlock(FS
DirWriteFileOp.java:265)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesyste
m.java:2569)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServe
r.java:846)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.add
Block(ClientNamenodeProtocolServerSideTranslatorPB.java:510)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeP
rotocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpc
Engine.java:503)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:989)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:871)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:817)
        at java.security.AccessController.doPrivileged(Native Method)
```

When I tried running the command, I got the following error. I also noticed that resource manager listed 1 active node for the single node setup but lists 0 datanodes now.

I tried logging into my master node but was unable to after it timed out, even after using SSH through Google Cloud, it would fail to connect. Following [23], I was able to make a startup script and pass it to the VM through Google Console to allow port 22. I probably disabled this port when enabling ufw on accident.



After getting back into the master node, there still was no datanodes showing up on the hdfs website or JPS on slave nodes. The node managers showed up in jps, but the resource manager website still showed 0 active nodes.

```
<property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
</property>

<property>
        <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

<property>
        <name>yarn.resourcemanager.resource-tracker.address</name>
        <value>master:8025</value>
</property>

<property>
        <name>yarn.resourcemanager.scheduler.address</name>
        <value>master:8030</value>
</property>

<property>
        <name>yarn.resourcemanager.address</name>
        <value>master:8040</value>
</property>
```

Yarn-site.xml above.

```
<configuration>

        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>

        <property>
                <name>mapreduce.jobhistory.address</name>
                <value>master:10020</value>
        </property>
</configuration>
```

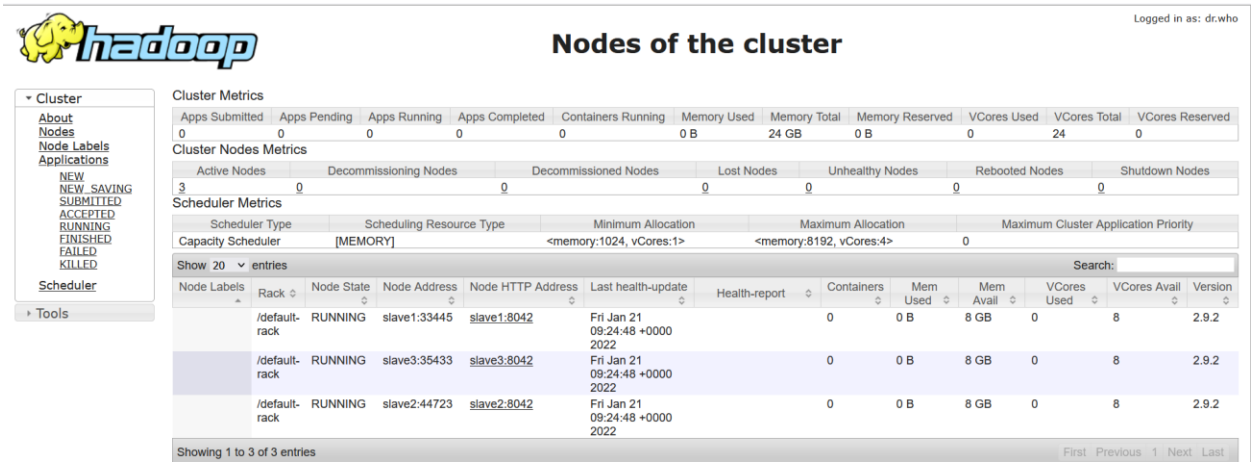
Mapred-site.xml above.

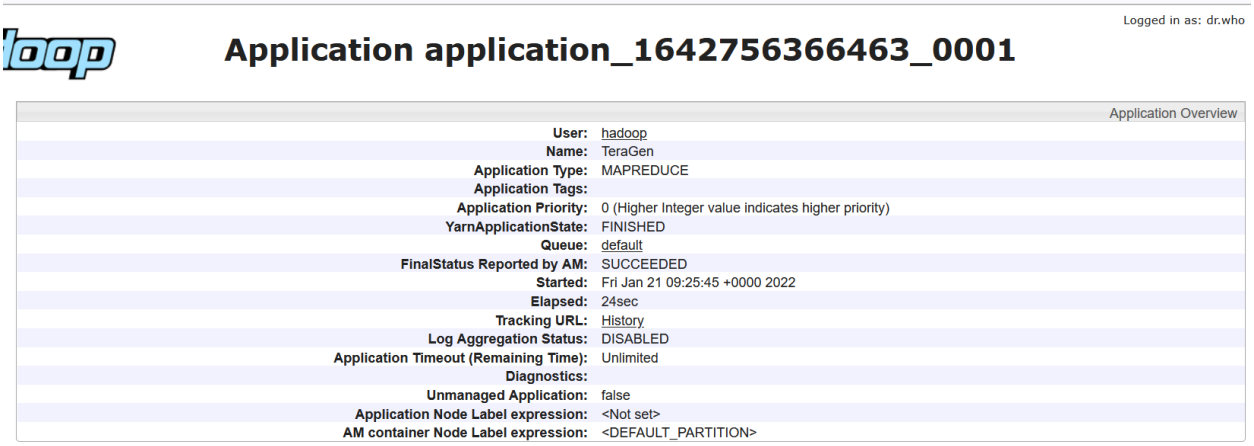
I modified these configurations to be the same as [24]. Although the resource manager website was finally showing 3 active nodes for the first time ever, HDFS website and jps still showed no datanodes.

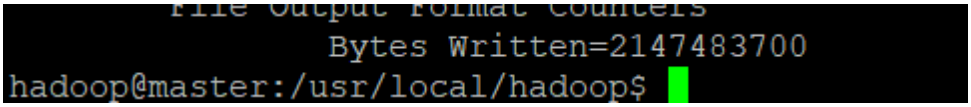
I managed to fix this by deleting both the temporary files and all files under Hadoop_data on every node, both masters and slaves, before reformatting the distributed file system. I also realized that when copying the entire /etc/Hadoop directory from master to the slave nodes, I

inadvertently copied master's master and slave files, which may have messed up the behavior. After deleting the master files and changing the contents of the slave file on the slave nodes back to just "localhost", in combination with the above, everything worked fine.

Here is a screenshot of the resource manager website finally showing 3 active nodes.







Running Teragen using the previously mentioned command, the job successfully completed in 24 seconds with the 2GB dataset.

Running the following command for Terasort:

"./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/2GB-input terasort/2GB-output"

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/2GB-input terasort/2GB-output
22/01/21 09:32:29 INFO terasort.TeraSort: starting
22/01/21 09:32:30 INFO input.FileInputFormat: Total input files to process : 2
Spent 98ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
Computing input splits took 102ms
Sampling 10 splits of 16
Making 1 from 100000 sampled records
Computing paritition took 553ms
Spent 657ms computing partitions.
22/01/21 09:32:30 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/21 09:32:31 INFO mapreduce.JobSubmitter: number of splits:16
22/01/21 09:32:31 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/21 09:32:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642756366463_0002
22/01/21 09:32:31 INFO impl.YarnClientImpl: Submitted application application_1642756366463_0002
22/01/21 09:32:31 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1642756366463_0002/
22/01/21 09:32:31 INFO mapreduce.Job: Running job: job_1642756366463_0002
22/01/21 09:32:37 INFO mapreduce.Job: Job job_1642756366463_0002 running in uber mode : false
22/01/21 09:32:37 INFO mapreduce.Job:  map 0% reduce 0%
22/01/21 09:32:53 INFO mapreduce.Job:  map 6% reduce 0%
22/01/21 09:32:54 INFO mapreduce.Job:  map 13% reduce 0%
22/01/21 09:33:06 INFO mapreduce.Job:  map 21% reduce 0%
22/01/21 09:33:07 INFO mapreduce.Job:  map 29% reduce 0%
22/01/21 09:33:09 INFO mapreduce.Job:  map 34% reduce 0%
22/01/21 09:33:10 INFO mapreduce.Job:  map 51% reduce 4%
22/01/21 09:33:12 INFO mapreduce.Job:  map 52% reduce 4%
22/01/21 09:33:13 INFO mapreduce.Job:  map 55% reduce 4%
22/01/21 09:33:17 INFO mapreduce.Job:  map 59% reduce 4%
22/01/21 09:33:18 INFO mapreduce.Job:  map 60% reduce 4%
22/01/21 09:33:19 INFO mapreduce.Job:  map 63% reduce 4%
22/01/21 09:33:21 INFO mapreduce.Job:  map 64% reduce 4%
22/01/21 09:33:22 INFO mapreduce.Job:  map 71% reduce 6%
22/01/21 09:33:24 INFO mapreduce.Job:  map 74% reduce 6%
22/01/21 09:33:25 INFO mapreduce.Job:  map 80% reduce 6%
22/01/21 09:33:26 INFO mapreduce.Job:  map 82% reduce 6%
22/01/21 09:33:27 INFO mapreduce.Job:  map 84% reduce 6%
22/01/21 09:33:28 INFO mapreduce.Job:  map 85% reduce 10%
22/01/21 09:33:34 INFO mapreduce.Job:  map 88% reduce 15%
22/01/21 09:33:39 INFO mapreduce.Job:  map 94% reduce 15%
22/01/21 09:33:40 INFO mapreduce.Job:  map 99% reduce 17%
22/01/21 09:33:41 INFO mapreduce.Job:  map 100% reduce 17%
22/01/21 09:33:46 INFO mapreduce.Job:  map 100% reduce 25%
22/01/21 09:33:58 INFO mapreduce.Job:  map 100% reduce 31%
22/01/21 09:34:04 INFO mapreduce.Job:  map 100% reduce 47%
22/01/21 09:34:10 INFO mapreduce.Job:  map 100% reduce 67%
22/01/21 09:34:16 INFO mapreduce.Job:  map 100% reduce 74%
22/01/21 09:34:22 INFO mapreduce.Job:  map 100% reduce 79%
22/01/21 09:34:28 INFO mapreduce.Job:  map 100% reduce 81%
```

Continued on next page for screenshot sizing reasons.

```
22/01/21 09:34:28 INFO mapreduce.Job:  map 100% reduce 81%
22/01/21 09:34:34 INFO mapreduce.Job:  map 100% reduce 84%
22/01/21 09:34:40 INFO mapreduce.Job:  map 100% reduce 86%
22/01/21 09:34:46 INFO mapreduce.Job:  map 100% reduce 88%
22/01/21 09:34:52 INFO mapreduce.Job:  map 100% reduce 92%
22/01/21 09:34:58 INFO mapreduce.Job:  map 100% reduce 95%
22/01/21 09:35:04 INFO mapreduce.Job:  map 100% reduce 97%
22/01/21 09:35:10 INFO mapreduce.Job:  map 100% reduce 100%
22/01/21 09:35:12 INFO mapreduce.Job: Job job_1642756366463_0002 completed successfully
22/01/21 09:35:12 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=5443871246
                FILE: Number of bytes written=7680647881
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2147485732
                HDFS: Number of bytes written=2147483700
                HDFS: Number of read operations=51
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=3
                Launched map tasks=18
                Launched reduce tasks=1
                Data-local map tasks=15
                Rack-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=781725
                Total time spent by all reduces in occupied slots (ms)=135971
                Total time spent by all map tasks (ms)=781725
                Total time spent by all reduce tasks (ms)=135971
                Total vcore-milliseconds taken by all map tasks=781725
                Total vcore-milliseconds taken by all reduce tasks=135971
                Total megabyte-milliseconds taken by all map tasks=800486400
                Total megabyte-milliseconds taken by all reduce tasks=139234304
        Map-Reduce Framework
                Map input records=21474837
                Map output records=21474837
                Map output bytes=2190433374
                Map output materialized bytes=2233383144
                Input split bytes=2032
                Combine input records=0
                Combine output records=0
                Reduce input groups=21474837
```

```
                Total megabyte-milliseconds taken by all reduce tasks=139234304
        Map-Reduce Framework
                Map input records=21474837
                Map output records=21474837
                Map output bytes=2190433374
                Map output materialized bytes=2233383144
                Input split bytes=2032
                Combine input records=0
                Combine output records=0
                Reduce input groups=21474837
                Reduce shuffle bytes=2233383144
                Reduce input records=21474837
                Reduce output records=21474837
                Spilled Records=73819750
                Shuffled Maps =16
                Failed Shuffles=0
                Merged Map outputs=16
                GC time elapsed (ms)=6328
                CPU time spent (ms)=192270
                Physical memory (bytes) snapshot=4694466560
                Virtual memory (bytes) snapshot=32718409728
                Total committed heap usage (bytes)=3272605696
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2147483700
        File Output Format Counters
                Bytes Written=2147483700
22/01/21 09:35:12 INFO terasort.TeraSort: done
hadoop@master:/usr/local/hadoop$
```

Application application_1642756366463_0002

User:	hadoop
Name:	TeraSort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Fri Jan 21 09:32:31 +0000 2022
Elapsed:	2mins, 39sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Terasort finishes successfully. According to resource manager, it took 2 minutes and 39 seconds to run.

Using the following command for Teravalidate: "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/2GB-output terasort/check"

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/2GB-output terasort/check
22/01/21 09:40:17 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/21 09:40:18 INFO input.FileInputFormat: Total input files to process : 1
Spent 21ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
22/01/21 09:40:18 INFO mapreduce.JobSubmitter: number of splits:1
22/01/21 09:40:18 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/21 09:40:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642756366463_0003
22/01/21 09:40:18 INFO impl.YarnClientImpl: Submitted application application_1642756366463_0003
22/01/21 09:40:18 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1642756366463_0003/
22/01/21 09:40:18 INFO mapreduce.Job: Running job: job_1642756366463_0003
22/01/21 09:40:25 INFO mapreduce.Job: Job job_1642756366463_0003 running in uber mode : false
22/01/21 09:40:25 INFO mapreduce.Job:  map 0% reduce 0%
22/01/21 09:40:41 INFO mapreduce.Job:  map 100% reduce 0%
22/01/21 09:40:46 INFO mapreduce.Job:  map 100% reduce 100%
22/01/21 09:40:46 INFO mapreduce.Job: Job job_1642756366463_0003 completed successfully
22/01/21 09:40:46 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=94
                FILE: Number of bytes written=397399
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2147483828
                HDFS: Number of bytes written=24
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=13555
                Total time spent by all reduces in occupied slots (ms)=2577
                Total time spent by all map tasks (ms)=13555
                Total time spent by all reduce tasks (ms)=2577
                Total vcore-milliseconds taken by all map tasks=13555
                Total vcore-milliseconds taken by all reduce tasks=2577
                Total megabyte-milliseconds taken by all map tasks=13880320
                Total megabyte-milliseconds taken by all reduce tasks=2638848
        Map-Reduce Framework
                Map input records=21474837
                Map output records=3
                Map output bytes=82
                Map output materialized bytes=94
                Input split bytes=128
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
```

```
        Map-Reduce Framework
                Map input records=21474837
                Map output records=3
                Map output bytes=82
                Map output materialized bytes=94
                Input split bytes=128
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=94
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=310
                CPU time spent (ms)=12750
                Physical memory (bytes) snapshot=466481152
                Virtual memory (bytes) snapshot=3845881856
                Total committed heap usage (bytes)=329252864
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2147483700
        File Output Format Counters
                Bytes Written=24
hadoop@master:/usr/local/hadoop$
```

Application application_1642756366463_0003

User:	hadoop
Name:	TeraValidate
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Fri Jan 21 09:40:18 +0000 2022
Elapsed:	26sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Teravalidate finishes completely for the 2GB dataset in 26 seconds.

20 GB dataset = $2^{30} * 20$ = 21,474,836,480 bytes divided by 100 bytes per terasort row = 214,748,365 terasort rows.

20 GB Teragen command: "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 214748365 terasort/20GB-input"

Image on next page.

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-e
xamples-2.9.2.jar teragen 214748365 terasort/20GB-input
22/01/22 01:36:21 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:804
0
22/01/22 01:36:22 INFO terasort.TeraGen: Generating 214748365 using 2
22/01/22 01:36:22 INFO mapreduce.JobSubmitter: number of splits:2
22/01/22 01:36:22 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publish
er.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/22 01:36:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642815277676_0
001
22/01/22 01:36:23 INFO impl.YarnClientImpl: Submitted application application_1642815277676_0
001
22/01/22 01:36:23 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/appl
ication_1642815277676_0001/
22/01/22 01:36:23 INFO mapreduce.Job: Running job: job_1642815277676_0001
22/01/22 01:36:29 INFO mapreduce.Job: Job job_1642815277676_0001 running in uber mode : false
22/01/22 01:36:29 INFO mapreduce.Job:  map 0% reduce 0%
22/01/22 01:36:45 INFO mapreduce.Job:  map 7% reduce 0%
22/01/22 01:36:47 INFO mapreduce.Job:  map 13% reduce 0%
22/01/22 01:36:51 INFO mapreduce.Job:  map 17% reduce 0%
22/01/22 01:36:53 INFO mapreduce.Job:  map 20% reduce 0%
22/01/22 01:36:57 INFO mapreduce.Job:  map 23% reduce 0%
22/01/22 01:36:59 INFO mapreduce.Job:  map 26% reduce 0%
22/01/22 01:37:03 INFO mapreduce.Job:  map 27% reduce 0%
22/01/22 01:37:05 INFO mapreduce.Job:  map 28% reduce 0%
22/01/22 01:37:09 INFO mapreduce.Job:  map 29% reduce 0%
22/01/22 01:37:15 INFO mapreduce.Job:  map 30% reduce 0%
22/01/22 01:37:17 INFO mapreduce.Job:  map 31% reduce 0%
22/01/22 01:37:23 INFO mapreduce.Job:  map 32% reduce 0%
22/01/22 01:37:28 INFO mapreduce.Job:  map 33% reduce 0%
22/01/22 01:37:29 INFO mapreduce.Job:  map 34% reduce 0%
22/01/22 01:37:34 INFO mapreduce.Job:  map 35% reduce 0%
22/01/22 01:37:35 INFO mapreduce.Job:  map 36% reduce 0%
22/01/22 01:37:40 INFO mapreduce.Job:  map 37% reduce 0%
22/01/22 01:37:46 INFO mapreduce.Job:  map 38% reduce 0%
22/01/22 01:37:48 INFO mapreduce.Job:  map 39% reduce 0%
22/01/22 01:37:52 INFO mapreduce.Job:  map 40% reduce 0%
22/01/22 01:37:54 INFO mapreduce.Job:  map 41% reduce 0%
22/01/22 01:37:58 INFO mapreduce.Job:  map 42% reduce 0%
22/01/22 01:38:00 INFO mapreduce.Job:  map 43% reduce 0%
22/01/22 01:38:04 INFO mapreduce.Job:  map 44% reduce 0%
22/01/22 01:38:06 INFO mapreduce.Job:  map 45% reduce 0%
22/01/22 01:38:14 INFO mapreduce.Job:  map 46% reduce 0%
22/01/22 01:38:17 INFO mapreduce.Job:  map 47% reduce 0%
22/01/22 01:38:20 INFO mapreduce.Job:  map 48% reduce 0%
22/01/22 01:38:24 INFO mapreduce.Job:  map 49% reduce 0%
22/01/22 01:38:26 INFO mapreduce.Job:  map 50% reduce 0%
22/01/22 01:38:32 INFO mapreduce.Job:  map 51% reduce 0%
22/01/22 01:38:36 INFO mapreduce.Job:  map 52% reduce 0%
```

```
22/01/22 01:38:26 INFO mapreduce.Job:  map 50% reduce 0%
22/01/22 01:38:32 INFO mapreduce.Job:  map 51% reduce 0%
22/01/22 01:38:36 INFO mapreduce.Job:  map 52% reduce 0%
22/01/22 01:38:39 INFO mapreduce.Job:  map 53% reduce 0%
22/01/22 01:38:42 INFO mapreduce.Job:  map 54% reduce 0%
22/01/22 01:38:45 INFO mapreduce.Job:  map 55% reduce 0%
22/01/22 01:38:51 INFO mapreduce.Job:  map 56% reduce 0%
22/01/22 01:38:55 INFO mapreduce.Job:  map 57% reduce 0%
22/01/22 01:38:57 INFO mapreduce.Job:  map 58% reduce 0%
22/01/22 01:39:01 INFO mapreduce.Job:  map 59% reduce 0%
22/01/22 01:39:04 INFO mapreduce.Job:  map 60% reduce 0%
22/01/22 01:39:10 INFO mapreduce.Job:  map 61% reduce 0%
22/01/22 01:39:13 INFO mapreduce.Job:  map 62% reduce 0%
22/01/22 01:39:16 INFO mapreduce.Job:  map 63% reduce 0%
22/01/22 01:39:19 INFO mapreduce.Job:  map 64% reduce 0%
22/01/22 01:39:22 INFO mapreduce.Job:  map 65% reduce 0%
22/01/22 01:39:28 INFO mapreduce.Job:  map 66% reduce 0%
22/01/22 01:39:31 INFO mapreduce.Job:  map 67% reduce 0%
22/01/22 01:39:34 INFO mapreduce.Job:  map 68% reduce 0%
22/01/22 01:39:37 INFO mapreduce.Job:  map 69% reduce 0%
22/01/22 01:39:40 INFO mapreduce.Job:  map 70% reduce 0%
22/01/22 01:39:46 INFO mapreduce.Job:  map 71% reduce 0%
22/01/22 01:39:49 INFO mapreduce.Job:  map 72% reduce 0%
22/01/22 01:39:52 INFO mapreduce.Job:  map 73% reduce 0%
22/01/22 01:39:56 INFO mapreduce.Job:  map 74% reduce 0%
22/01/22 01:40:02 INFO mapreduce.Job:  map 75% reduce 0%
22/01/22 01:40:04 INFO mapreduce.Job:  map 76% reduce 0%
22/01/22 01:40:08 INFO mapreduce.Job:  map 77% reduce 0%
22/01/22 01:40:11 INFO mapreduce.Job:  map 78% reduce 0%
22/01/22 01:40:14 INFO mapreduce.Job:  map 79% reduce 0%
22/01/22 01:40:17 INFO mapreduce.Job:  map 80% reduce 0%
22/01/22 01:40:23 INFO mapreduce.Job:  map 81% reduce 0%
22/01/22 01:40:27 INFO mapreduce.Job:  map 82% reduce 0%
22/01/22 01:40:29 INFO mapreduce.Job:  map 83% reduce 0%
22/01/22 01:40:33 INFO mapreduce.Job:  map 84% reduce 0%
22/01/22 01:40:35 INFO mapreduce.Job:  map 85% reduce 0%
22/01/22 01:40:41 INFO mapreduce.Job:  map 86% reduce 0%
22/01/22 01:40:45 INFO mapreduce.Job:  map 87% reduce 0%
22/01/22 01:40:47 INFO mapreduce.Job:  map 88% reduce 0%
22/01/22 01:40:52 INFO mapreduce.Job:  map 89% reduce 0%
22/01/22 01:40:58 INFO mapreduce.Job:  map 90% reduce 0%
22/01/22 01:40:59 INFO mapreduce.Job:  map 91% reduce 0%
22/01/22 01:41:04 INFO mapreduce.Job:  map 92% reduce 0%
22/01/22 01:41:05 INFO mapreduce.Job:  map 93% reduce 0%
22/01/22 01:41:12 INFO mapreduce.Job:  map 94% reduce 0%
22/01/22 01:41:16 INFO mapreduce.Job:  map 95% reduce 0%
22/01/22 01:41:18 INFO mapreduce.Job:  map 96% reduce 0%
22/01/22 01:41:22 INFO mapreduce.Job:  map 97% reduce 0%
22/01/22 01:41:24 INFO mapreduce.Job:  map 98% reduce 0%
22/01/22 01:41:28 INFO mapreduce.Job:  map 99% reduce 0%
```

```
22/01/22 01:41:18 INFO mapreduce.Job:  map 96% reduce 0%
22/01/22 01:41:22 INFO mapreduce.Job:  map 97% reduce 0%
22/01/22 01:41:24 INFO mapreduce.Job:  map 98% reduce 0%
22/01/22 01:41:28 INFO mapreduce.Job:  map 99% reduce 0%
22/01/22 01:41:32 INFO mapreduce.Job:  map 100% reduce 0%
22/01/22 01:41:34 INFO mapreduce.Job: Job job_1642815277676_0001 completed successfully
22/01/22 01:41:34 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396410
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=170
                HDFS: Number of bytes written=21474836500
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=601851
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=601851
                Total vcore-milliseconds taken by all map tasks=601851
                Total megabyte-milliseconds taken by all map tasks=616295424
        Map-Reduce Framework
                Map input records=214748365
                Map output records=214748365
                Input split bytes=170
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2745
                CPU time spent (ms)=179640
                Physical memory (bytes) snapshot=477057024
                Virtual memory (bytes) snapshot=3860226048
                Total committed heap usage (bytes)=288358400
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=461200258163748239
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=21474836500
hadoop@master:/usr/local/hadoop$
```

Application application_1642815277676_0001

User:	hadoop
Name:	TeraGen
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Jan 22 01:36:23 +0000 2022
Elapsed:	5mins, 9sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

The 20 GB Teragen command executed successfully in 5 minutes and 9 seconds.

20 GB Terasort command: "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/20GB-input terasort/20GB-output"

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/20GB-input terasort/20GB-output
22/01/22 01:45:47 INFO terasort.TeraSort: starting
22/01/22 01:45:48 INFO input.FileInputFormat: Total input files to process : 2
Spent 140ms computing base-splits.
Spent 4ms computing TeraScheduler splits.
Computing input splits took 145ms
Sampling 10 splits of 160
Making 1 from 100000 sampled records
Computing paritition took 520ms
Spent 667ms computing partitions.
22/01/22 01:45:48 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/22 01:45:49 INFO mapreduce.JobSubmitter: number of splits:160
22/01/22 01:45:49 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/22 01:45:49 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642815277676_0002
22/01/22 01:45:49 INFO impl.YarnClientImpl: Submitted application application_1642815277676_0002
22/01/22 01:45:49 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1642815277676_0002/
22/01/22 01:45:49 INFO mapreduce.Job: Running job: job_1642815277676_0002
22/01/22 01:45:54 INFO mapreduce.Job: Job job_1642815277676_0002 running in uber mode : false
22/01/22 01:45:54 INFO mapreduce.Job:  map 0% reduce 0%
22/01/22 01:46:20 INFO mapreduce.Job:  map 2% reduce 0%
22/01/22 01:46:23 INFO mapreduce.Job:  map 3% reduce 0%
22/01/22 01:46:25 INFO mapreduce.Job:  map 4% reduce 0%
22/01/22 01:46:26 INFO mapreduce.Job:  map 5% reduce 0%
22/01/22 01:46:27 INFO mapreduce.Job:  map 6% reduce 0%
22/01/22 01:46:32 INFO mapreduce.Job:  map 7% reduce 0%
22/01/22 01:46:34 INFO mapreduce.Job:  map 8% reduce 0%
22/01/22 01:46:38 INFO mapreduce.Job:  map 9% reduce 0%
22/01/22 01:46:40 INFO mapreduce.Job:  map 10% reduce 0%
22/01/22 01:46:47 INFO mapreduce.Job:  map 11% reduce 0%
22/01/22 01:46:49 INFO mapreduce.Job:  map 12% reduce 0%
22/01/22 01:46:52 INFO mapreduce.Job:  map 13% reduce 0%
22/01/22 01:46:57 INFO mapreduce.Job:  map 14% reduce 0%
22/01/22 01:46:59 INFO mapreduce.Job:  map 15% reduce 0%
22/01/22 01:47:02 INFO mapreduce.Job:  map 15% reduce 2%
22/01/22 01:47:03 INFO mapreduce.Job:  map 16% reduce 2%
22/01/22 01:47:09 INFO mapreduce.Job:  map 17% reduce 2%
22/01/22 01:47:12 INFO mapreduce.Job:  map 18% reduce 2%
22/01/22 01:47:17 INFO mapreduce.Job:  map 19% reduce 2%
22/01/22 01:47:18 INFO mapreduce.Job:  map 20% reduce 2%
22/01/22 01:47:23 INFO mapreduce.Job:  map 21% reduce 2%
22/01/22 01:47:30 INFO mapreduce.Job:  map 22% reduce 2%
22/01/22 01:47:33 INFO mapreduce.Job:  map 23% reduce 2%
22/01/22 01:47:39 INFO mapreduce.Job:  map 24% reduce 2%
22/01/22 01:47:42 INFO mapreduce.Job:  map 25% reduce 2%
22/01/22 01:47:45 INFO mapreduce.Job:  map 26% reduce 2%
22/01/22 01:47:51 INFO mapreduce.Job:  map 27% reduce 2%
22/01/22 01:48:02 INFO mapreduce.Job:  map 27% reduce 3%
22/01/22 01:48:10 INFO mapreduce.Job:  map 28% reduce 3%
22/01/22 01:48:14 INFO mapreduce.Job:  map 29% reduce 3%
```

```
22/01/22 01:47:51 INFO mapreduce.Job:  map 27% reduce 2%
22/01/22 01:48:02 INFO mapreduce.Job:  map 27% reduce 3%
22/01/22 01:48:10 INFO mapreduce.Job:  map 28% reduce 3%
22/01/22 01:48:14 INFO mapreduce.Job:  map 29% reduce 3%
22/01/22 01:48:19 INFO mapreduce.Job:  map 30% reduce 3%
22/01/22 01:48:23 INFO mapreduce.Job:  map 31% reduce 3%
22/01/22 01:48:28 INFO mapreduce.Job:  map 32% reduce 3%
22/01/22 01:48:33 INFO mapreduce.Job:  map 33% reduce 3%
22/01/22 01:48:49 INFO mapreduce.Job:  map 34% reduce 3%
22/01/22 01:48:56 INFO mapreduce.Job:  map 35% reduce 3%
22/01/22 01:49:02 INFO mapreduce.Job:  map 36% reduce 3%
22/01/22 01:49:06 INFO mapreduce.Job:  map 37% reduce 3%
22/01/22 01:49:09 INFO mapreduce.Job:  map 37% reduce 4%
22/01/22 01:49:12 INFO mapreduce.Job:  map 38% reduce 4%
22/01/22 01:49:19 INFO mapreduce.Job:  map 39% reduce 4%
22/01/22 01:49:26 INFO mapreduce.Job:  map 40% reduce 4%
22/01/22 01:49:47 INFO mapreduce.Job:  map 41% reduce 4%
22/01/22 01:49:56 INFO mapreduce.Job:  map 42% reduce 4%
22/01/22 01:49:57 INFO mapreduce.Job:  map 43% reduce 4%
22/01/22 01:50:02 INFO mapreduce.Job:  map 44% reduce 4%
22/01/22 01:50:04 INFO mapreduce.Job:  map 44% reduce 5%
22/01/22 01:50:08 INFO mapreduce.Job:  map 45% reduce 5%
22/01/22 01:50:20 INFO mapreduce.Job:  map 46% reduce 5%
22/01/22 01:50:26 INFO mapreduce.Job:  map 47% reduce 5%
22/01/22 01:50:32 INFO mapreduce.Job:  map 48% reduce 5%
22/01/22 01:50:40 INFO mapreduce.Job:  map 49% reduce 5%
22/01/22 01:50:51 INFO mapreduce.Job:  map 50% reduce 5%
22/01/22 01:50:56 INFO mapreduce.Job:  map 51% reduce 5%
22/01/22 01:51:02 INFO mapreduce.Job:  map 52% reduce 5%
22/01/22 01:51:11 INFO mapreduce.Job:  map 53% reduce 5%
22/01/22 01:51:23 INFO mapreduce.Job:  map 54% reduce 5%
22/01/22 01:51:28 INFO mapreduce.Job:  map 54% reduce 6%
22/01/22 01:51:33 INFO mapreduce.Job:  map 55% reduce 6%
22/01/22 01:51:42 INFO mapreduce.Job:  map 56% reduce 6%
22/01/22 01:51:45 INFO mapreduce.Job:  map 57% reduce 6%
22/01/22 01:51:53 INFO mapreduce.Job:  map 58% reduce 6%
22/01/22 01:51:56 INFO mapreduce.Job:  map 59% reduce 6%
22/01/22 01:52:12 INFO mapreduce.Job:  map 60% reduce 6%
22/01/22 01:52:19 INFO mapreduce.Job:  map 61% reduce 6%
22/01/22 01:52:31 INFO mapreduce.Job:  map 62% reduce 6%
22/01/22 01:52:39 INFO mapreduce.Job:  map 63% reduce 6%
22/01/22 01:52:49 INFO mapreduce.Job:  map 64% reduce 6%
22/01/22 01:52:53 INFO mapreduce.Job:  map 64% reduce 7%
22/01/22 01:52:57 INFO mapreduce.Job:  map 65% reduce 7%
22/01/22 01:53:04 INFO mapreduce.Job:  map 66% reduce 7%
22/01/22 01:53:09 INFO mapreduce.Job:  map 67% reduce 7%
22/01/22 01:53:16 INFO mapreduce.Job:  map 68% reduce 7%
22/01/22 01:53:24 INFO mapreduce.Job:  map 69% reduce 7%
22/01/22 01:53:32 INFO mapreduce.Job:  map 70% reduce 7%
22/01/22 01:53:38 INFO mapreduce.Job:  map 71% reduce 7%
```

```
22/01/22 01:53:32 INFO mapreduce.Job:  map 70% reduce 7%
22/01/22 01:53:38 INFO mapreduce.Job:  map 71% reduce 7%
22/01/22 01:53:46 INFO mapreduce.Job:  map 72% reduce 7%
22/01/22 01:53:48 INFO mapreduce.Job:  map 72% reduce 8%
22/01/22 01:53:54 INFO mapreduce.Job:  map 73% reduce 8%
22/01/22 01:54:02 INFO mapreduce.Job:  map 74% reduce 8%
22/01/22 01:54:16 INFO mapreduce.Job:  map 75% reduce 8%
22/01/22 01:54:24 INFO mapreduce.Job:  map 76% reduce 8%
22/01/22 01:54:36 INFO mapreduce.Job:  map 77% reduce 8%
22/01/22 01:54:42 INFO mapreduce.Job:  map 78% reduce 8%
22/01/22 01:54:48 INFO mapreduce.Job:  map 79% reduce 8%
22/01/22 01:54:54 INFO mapreduce.Job:  map 80% reduce 8%
22/01/22 01:55:02 INFO mapreduce.Job:  map 81% reduce 8%
22/01/22 01:55:06 INFO mapreduce.Job:  map 81% reduce 9%
22/01/22 01:55:11 INFO mapreduce.Job:  map 82% reduce 9%
22/01/22 01:55:22 INFO mapreduce.Job:  map 83% reduce 9%
22/01/22 01:55:30 INFO mapreduce.Job:  map 84% reduce 9%
22/01/22 01:55:37 INFO mapreduce.Job:  map 85% reduce 9%
22/01/22 01:55:47 INFO mapreduce.Job:  map 86% reduce 9%
22/01/22 01:55:55 INFO mapreduce.Job:  map 87% reduce 9%
22/01/22 01:56:06 INFO mapreduce.Job:  map 88% reduce 9%
22/01/22 01:56:11 INFO mapreduce.Job:  map 89% reduce 9%
22/01/22 01:56:19 INFO mapreduce.Job:  map 90% reduce 9%
22/01/22 01:56:27 INFO mapreduce.Job:  map 91% reduce 9%
22/01/22 01:56:36 INFO mapreduce.Job:  map 91% reduce 10%
22/01/22 01:56:37 INFO mapreduce.Job:  map 92% reduce 10%
22/01/22 01:56:45 INFO mapreduce.Job:  map 93% reduce 10%
22/01/22 01:56:52 INFO mapreduce.Job:  map 94% reduce 10%
22/01/22 01:56:58 INFO mapreduce.Job:  map 95% reduce 10%
22/01/22 01:57:05 INFO mapreduce.Job:  map 96% reduce 10%
22/01/22 01:57:13 INFO mapreduce.Job:  map 97% reduce 10%
22/01/22 01:57:21 INFO mapreduce.Job:  map 98% reduce 10%
22/01/22 01:57:40 INFO mapreduce.Job:  map 99% reduce 10%
22/01/22 01:57:48 INFO mapreduce.Job:  map 99% reduce 11%
22/01/22 01:58:04 INFO mapreduce.Job:  map 100% reduce 11%
22/01/22 01:58:36 INFO mapreduce.Job:  map 100% reduce 12%
22/01/22 01:59:12 INFO mapreduce.Job:  map 100% reduce 13%
22/01/22 01:59:54 INFO mapreduce.Job:  map 100% reduce 14%
22/01/22 02:00:43 INFO mapreduce.Job:  map 100% reduce 15%
22/01/22 02:01:25 INFO mapreduce.Job:  map 100% reduce 16%
22/01/22 02:02:07 INFO mapreduce.Job:  map 100% reduce 17%
22/01/22 02:02:55 INFO mapreduce.Job:  map 100% reduce 18%
22/01/22 02:03:31 INFO mapreduce.Job:  map 100% reduce 19%
22/01/22 02:04:19 INFO mapreduce.Job:  map 100% reduce 20%
22/01/22 02:05:01 INFO mapreduce.Job:  map 100% reduce 21%
22/01/22 02:05:43 INFO mapreduce.Job:  map 100% reduce 22%
22/01/22 02:06:19 INFO mapreduce.Job:  map 100% reduce 23%
22/01/22 02:07:07 INFO mapreduce.Job:  map 100% reduce 24%
22/01/22 02:07:54 INFO mapreduce.Job:  map 100% reduce 25%
22/01/22 02:08:36 INFO mapreduce.Job:  map 100% reduce 26%
```

```
22/01/22 02:07:54 INFO mapreduce.Job:  map 100% reduce 25%
22/01/22 02:08:36 INFO mapreduce.Job:  map 100% reduce 26%
22/01/22 02:09:18 INFO mapreduce.Job:  map 100% reduce 27%
22/01/22 02:09:48 INFO mapreduce.Job:  map 100% reduce 28%
22/01/22 02:10:36 INFO mapreduce.Job:  map 100% reduce 29%
22/01/22 02:11:19 INFO mapreduce.Job:  map 100% reduce 30%
22/01/22 02:12:07 INFO mapreduce.Job:  map 100% reduce 31%
22/01/22 02:12:49 INFO mapreduce.Job:  map 100% reduce 32%
22/01/22 02:13:25 INFO mapreduce.Job:  map 100% reduce 33%

22/01/22 02:28:50 INFO mapreduce.Job:  map 100% reduce 34%
22/01/22 02:29:14 INFO mapreduce.Job:  map 100% reduce 35%
22/01/22 02:29:32 INFO mapreduce.Job:  map 100% reduce 36%
22/01/22 02:29:56 INFO mapreduce.Job:  map 100% reduce 37%
```

When the map was 100% done but the reduce was at 33%, it paused for a long time, so I was worried. According to [25], 0-33% of reduce is the shuffle, 33-67% is sort, and 67-100% is the actual reduce code. I knew that the reduce code itself should be fine because the reduce code was in the Hadoop map examples jar included with Hadoop, and terasort reduction should be well vetted by the community / whoever made it, but I was worried that I might have made a configuration error with the multi node cluster setup that prevented efficient scaling when the dataset was as large as 20 GB, although the reduction of Terasort worked fine for a 2GB dataset. The black line in the middle is me pressing the enter key. After waiting 15 minutes, it moved on from 33 to 34%, then proceeded through the rest of reduction with no issues.

Continued on next page.

```
22/01/22 02:29:32 INFO mapreduce.Job:  map 100% reduce 36%
22/01/22 02:29:56 INFO mapreduce.Job:  map 100% reduce 37%
22/01/22 02:30:14 INFO mapreduce.Job:  map 100% reduce 38%
22/01/22 02:30:32 INFO mapreduce.Job:  map 100% reduce 39%
22/01/22 02:30:50 INFO mapreduce.Job:  map 100% reduce 40%
22/01/22 02:31:08 INFO mapreduce.Job:  map 100% reduce 41%
22/01/22 02:31:26 INFO mapreduce.Job:  map 100% reduce 42%
22/01/22 02:31:44 INFO mapreduce.Job:  map 100% reduce 43%
22/01/22 02:32:08 INFO mapreduce.Job:  map 100% reduce 44%
22/01/22 02:32:26 INFO mapreduce.Job:  map 100% reduce 45%
22/01/22 02:32:44 INFO mapreduce.Job:  map 100% reduce 46%
22/01/22 02:33:08 INFO mapreduce.Job:  map 100% reduce 47%
22/01/22 02:33:26 INFO mapreduce.Job:  map 100% reduce 48%
22/01/22 02:33:50 INFO mapreduce.Job:  map 100% reduce 49%
22/01/22 02:34:08 INFO mapreduce.Job:  map 100% reduce 50%
22/01/22 02:34:33 INFO mapreduce.Job:  map 100% reduce 51%
22/01/22 02:34:51 INFO mapreduce.Job:  map 100% reduce 52%
22/01/22 02:35:15 INFO mapreduce.Job:  map 100% reduce 53%
22/01/22 02:35:33 INFO mapreduce.Job:  map 100% reduce 54%
22/01/22 02:35:57 INFO mapreduce.Job:  map 100% reduce 55%
22/01/22 02:36:15 INFO mapreduce.Job:  map 100% reduce 56%
22/01/22 02:36:39 INFO mapreduce.Job:  map 100% reduce 57%
22/01/22 02:36:57 INFO mapreduce.Job:  map 100% reduce 58%
22/01/22 02:37:21 INFO mapreduce.Job:  map 100% reduce 59%
22/01/22 02:37:39 INFO mapreduce.Job:  map 100% reduce 60%
22/01/22 02:38:03 INFO mapreduce.Job:  map 100% reduce 61%
22/01/22 02:38:21 INFO mapreduce.Job:  map 100% reduce 62%
22/01/22 02:38:45 INFO mapreduce.Job:  map 100% reduce 63%
22/01/22 02:39:09 INFO mapreduce.Job:  map 100% reduce 64%
22/01/22 02:39:27 INFO mapreduce.Job:  map 100% reduce 65%
22/01/22 02:39:51 INFO mapreduce.Job:  map 100% reduce 66%
22/01/22 02:40:09 INFO mapreduce.Job:  map 100% reduce 67%
22/01/22 02:40:51 INFO mapreduce.Job:  map 100% reduce 68%
22/01/22 02:41:33 INFO mapreduce.Job:  map 100% reduce 69%
22/01/22 02:42:15 INFO mapreduce.Job:  map 100% reduce 70%
22/01/22 02:43:03 INFO mapreduce.Job:  map 100% reduce 71%
22/01/22 02:43:45 INFO mapreduce.Job:  map 100% reduce 72%
22/01/22 02:44:27 INFO mapreduce.Job:  map 100% reduce 73%
22/01/22 02:45:09 INFO mapreduce.Job:  map 100% reduce 74%
22/01/22 02:45:51 INFO mapreduce.Job:  map 100% reduce 75%
22/01/22 02:46:33 INFO mapreduce.Job:  map 100% reduce 76%
22/01/22 02:47:21 INFO mapreduce.Job:  map 100% reduce 77%
22/01/22 02:48:04 INFO mapreduce.Job:  map 100% reduce 78%
22/01/22 02:48:46 INFO mapreduce.Job:  map 100% reduce 79%
22/01/22 02:49:28 INFO mapreduce.Job:  map 100% reduce 80%
22/01/22 02:50:10 INFO mapreduce.Job:  map 100% reduce 81%
22/01/22 02:50:52 INFO mapreduce.Job:  map 100% reduce 82%
22/01/22 02:51:34 INFO mapreduce.Job:  map 100% reduce 83%
22/01/22 02:52:16 INFO mapreduce.Job:  map 100% reduce 84%
22/01/22 02:52:58 INFO mapreduce.Job:  map 100% reduce 85%
```

```
22/01/22 02:51:34 INFO mapreduce.Job:  map 100% reduce 83%
22/01/22 02:52:16 INFO mapreduce.Job:  map 100% reduce 84%
22/01/22 02:52:58 INFO mapreduce.Job:  map 100% reduce 85%
22/01/22 02:53:46 INFO mapreduce.Job:  map 100% reduce 86%
22/01/22 02:54:28 INFO mapreduce.Job:  map 100% reduce 87%
22/01/22 02:55:10 INFO mapreduce.Job:  map 100% reduce 88%
22/01/22 02:55:52 INFO mapreduce.Job:  map 100% reduce 89%
22/01/22 02:56:34 INFO mapreduce.Job:  map 100% reduce 90%
22/01/22 02:57:22 INFO mapreduce.Job:  map 100% reduce 91%
22/01/22 02:58:04 INFO mapreduce.Job:  map 100% reduce 92%
22/01/22 02:58:46 INFO mapreduce.Job:  map 100% reduce 93%
22/01/22 02:59:28 INFO mapreduce.Job:  map 100% reduce 94%
22/01/22 03:00:10 INFO mapreduce.Job:  map 100% reduce 95%
22/01/22 03:00:53 INFO mapreduce.Job:  map 100% reduce 96%
22/01/22 03:01:35 INFO mapreduce.Job:  map 100% reduce 97%
22/01/22 03:02:17 INFO mapreduce.Job:  map 100% reduce 98%
22/01/22 03:02:59 INFO mapreduce.Job:  map 100% reduce 99%
22/01/22 03:03:47 INFO mapreduce.Job:  map 100% reduce 100%
22/01/22 03:04:07 INFO mapreduce.Job: Job job_1642815277676_0002 completed successfully
22/01/22 03:04:07 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=76353783446
                FILE: Number of bytes written=98719753985
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=21474856980
                HDFS: Number of bytes written=21474836500
                HDFS: Number of read operations=483
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
                Launched map tasks=162
                Launched reduce tasks=1
                Data-local map tasks=101
                Rack-local map tasks=61
                Total time spent by all maps in occupied slots (ms)=14267044
                Total time spent by all reduces in occupied slots (ms)=4647692
                Total time spent by all map tasks (ms)=14267044
                Total time spent by all reduce tasks (ms)=4647692
                Total vcore-milliseconds taken by all map tasks=14267044
                Total vcore-milliseconds taken by all reduce tasks=4647692
                Total megabyte-milliseconds taken by all map tasks=14609453056
                Total megabyte-milliseconds taken by all reduce tasks=4759236608
        Map-Reduce Framework
                Map input records=214748365
                Map output records=214748365
                Map output bytes=21904333230
                Map output materialized bytes=22333830920
```

```
        Job Counters
                Killed map tasks=2
                Launched map tasks=162
                Launched reduce tasks=1
                Data-local map tasks=101
                Rack-local map tasks=61
                Total time spent by all maps in occupied slots (ms)=14267044
                Total time spent by all reduces in occupied slots (ms)=4647692
                Total time spent by all map tasks (ms)=14267044
                Total time spent by all reduce tasks (ms)=4647692
                Total vcore-milliseconds taken by all map tasks=14267044
                Total vcore-milliseconds taken by all reduce tasks=4647692
                Total megabyte-milliseconds taken by all map tasks=14609453056
                Total megabyte-milliseconds taken by all reduce tasks=4759236608
        Map-Reduce Framework
                Map input records=214748365
                Map output records=214748365
                Map output bytes=21904333230
                Map output materialized bytes=22333830920
                Input split bytes=20480
                Combine input records=0
                Combine output records=0
                Reduce input groups=214748365
                Reduce shuffle bytes=22333830920
                Reduce input records=214748365
                Reduce output records=214748365
                Spilled Records=948919331
                Shuffled Maps =160
                Failed Shuffles=0
                Merged Map outputs=160
                GC time elapsed (ms)=47517
                CPU time spent (ms)=1708820
                Physical memory (bytes) snapshot=45331124224
                Virtual memory (bytes) snapshot=309435449344
                Total committed heap usage (bytes)=31685345280
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=21474836500
        File Output Format Counters
                Bytes Written=21474836500
22/01/22 03:04:07 INFO terasort.TeraSort: done
hadoop@master:/usr/local/hadoop$
hadoop@master:/usr/local/hadoop$
```

Application application_1642815277676_000

User:	hadoop
Name:	TeraSort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Jan 22 01:45:49 +0000 2022
Elapsed:	1hrs, 18mins, 16sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

The 20 GB Terasort command executed successfully in 1 hour, 18 minutes, and 16 seconds.

20 GB TeraValidate command: "./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/20GB-output terasort/20GBcheck"

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/20GB-output terasort/check
22/01/22 04:19:32 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://master:9000/user/hadoop/terasort/check already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:279)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:145)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1570)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1567)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1893)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
        at org.apache.hadoop.examples.terasort.TeraValidate.run(TeraValidate.java:178)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
        at org.apache.hadoop.examples.terasort.TeraValidate.main(TeraValidate.java:185)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:71)
        at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
        at org.apache.hadoop.examples.ExampleDriver.main(ExampleDriver.java:74)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:244)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
hadoop@master:/usr/local/hadoop$
```

I initially encountered errors because instead of "terasort/20GBcheck", I had "terasort/check", thinking I was invoking a command within a suite provided by terasort rather than specifying some filename for the potential output of Teravalidate. By reading the error log which said the terasort/check file already existed (due to the 2GB dataset), I modified my Terasort command to the one above where I put "terasort/20GBcheck" and it executed with no issues.

```
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/20GB-output terasort/20GBcheck
22/01/22 04:21:53 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/22 04:21:54 INFO input.FileInputFormat: Total input files to process : 1
Spent 38ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
22/01/22 04:21:54 INFO mapreduce.JobSubmitter: number of splits:1
22/01/22 04:21:54 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/22 04:21:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642815277676_0003
22/01/22 04:21:54 INFO impl.YarnClientImpl: Submitted application application_1642815277676_0003
22/01/22 04:21:54 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1642815277676_0003/
22/01/22 04:21:54 INFO mapreduce.Job: Running job: job_1642815277676_0003
22/01/22 04:22:00 INFO mapreduce.Job: Job job_1642815277676_0003 running in uber mode : false
22/01/22 04:22:00 INFO mapreduce.Job:  map 0% reduce 0%
22/01/22 04:22:16 INFO mapreduce.Job:  map 4% reduce 0%
22/01/22 04:22:22 INFO mapreduce.Job:  map 5% reduce 0%
22/01/22 04:22:34 INFO mapreduce.Job:  map 6% reduce 0%
22/01/22 04:22:41 INFO mapreduce.Job:  map 7% reduce 0%
22/01/22 04:22:53 INFO mapreduce.Job:  map 8% reduce 0%
22/01/22 04:23:05 INFO mapreduce.Job:  map 9% reduce 0%
22/01/22 04:23:17 INFO mapreduce.Job:  map 10% reduce 0%
22/01/22 04:23:29 INFO mapreduce.Job:  map 11% reduce 0%
22/01/22 04:23:41 INFO mapreduce.Job:  map 12% reduce 0%
22/01/22 04:23:47 INFO mapreduce.Job:  map 13% reduce 0%
22/01/22 04:23:59 INFO mapreduce.Job:  map 14% reduce 0%
22/01/22 04:24:11 INFO mapreduce.Job:  map 15% reduce 0%
22/01/22 04:24:23 INFO mapreduce.Job:  map 16% reduce 0%
22/01/22 04:24:35 INFO mapreduce.Job:  map 17% reduce 0%
22/01/22 04:24:47 INFO mapreduce.Job:  map 18% reduce 0%
22/01/22 04:24:53 INFO mapreduce.Job:  map 19% reduce 0%
22/01/22 04:25:05 INFO mapreduce.Job:  map 20% reduce 0%
22/01/22 04:25:17 INFO mapreduce.Job:  map 21% reduce 0%
22/01/22 04:25:29 INFO mapreduce.Job:  map 22% reduce 0%
22/01/22 04:25:41 INFO mapreduce.Job:  map 23% reduce 0%
22/01/22 04:25:53 INFO mapreduce.Job:  map 24% reduce 0%
22/01/22 04:25:59 INFO mapreduce.Job:  map 25% reduce 0%
22/01/22 04:26:11 INFO mapreduce.Job:  map 26% reduce 0%
22/01/22 04:26:23 INFO mapreduce.Job:  map 27% reduce 0%
22/01/22 04:26:35 INFO mapreduce.Job:  map 28% reduce 0%
22/01/22 04:26:47 INFO mapreduce.Job:  map 29% reduce 0%
22/01/22 04:26:59 INFO mapreduce.Job:  map 30% reduce 0%
22/01/22 04:27:11 INFO mapreduce.Job:  map 31% reduce 0%
22/01/22 04:27:17 INFO mapreduce.Job:  map 32% reduce 0%
22/01/22 04:27:29 INFO mapreduce.Job:  map 33% reduce 0%
22/01/22 04:27:41 INFO mapreduce.Job:  map 34% reduce 0%
22/01/22 04:27:53 INFO mapreduce.Job:  map 35% reduce 0%
22/01/22 04:28:05 INFO mapreduce.Job:  map 36% reduce 0%
22/01/22 04:28:17 INFO mapreduce.Job:  map 37% reduce 0%
22/01/22 04:28:23 INFO mapreduce.Job:  map 38% reduce 0%
```

Continued on next page for screenshot sizing reasons.

```
22/01/22 04:28:17 INFO mapreduce.Job:  map 37% reduce 0%
22/01/22 04:28:23 INFO mapreduce.Job:  map 38% reduce 0%
22/01/22 04:28:36 INFO mapreduce.Job:  map 39% reduce 0%
22/01/22 04:28:48 INFO mapreduce.Job:  map 40% reduce 0%
22/01/22 04:29:00 INFO mapreduce.Job:  map 41% reduce 0%
22/01/22 04:29:12 INFO mapreduce.Job:  map 42% reduce 0%
22/01/22 04:29:24 INFO mapreduce.Job:  map 43% reduce 0%
22/01/22 04:29:30 INFO mapreduce.Job:  map 44% reduce 0%
22/01/22 04:29:42 INFO mapreduce.Job:  map 45% reduce 0%
22/01/22 04:29:54 INFO mapreduce.Job:  map 46% reduce 0%
22/01/22 04:30:06 INFO mapreduce.Job:  map 47% reduce 0%
22/01/22 04:30:18 INFO mapreduce.Job:  map 48% reduce 0%
22/01/22 04:30:30 INFO mapreduce.Job:  map 49% reduce 0%
22/01/22 04:30:42 INFO mapreduce.Job:  map 50% reduce 0%
22/01/22 04:30:47 INFO mapreduce.Job:  map 51% reduce 0%
22/01/22 04:30:59 INFO mapreduce.Job:  map 52% reduce 0%
22/01/22 04:31:11 INFO mapreduce.Job:  map 53% reduce 0%
22/01/22 04:31:23 INFO mapreduce.Job:  map 54% reduce 0%
22/01/22 04:31:35 INFO mapreduce.Job:  map 55% reduce 0%
22/01/22 04:31:47 INFO mapreduce.Job:  map 56% reduce 0%
22/01/22 04:31:53 INFO mapreduce.Job:  map 57% reduce 0%
22/01/22 04:32:05 INFO mapreduce.Job:  map 58% reduce 0%
22/01/22 04:32:17 INFO mapreduce.Job:  map 59% reduce 0%
22/01/22 04:32:29 INFO mapreduce.Job:  map 60% reduce 0%
22/01/22 04:32:41 INFO mapreduce.Job:  map 61% reduce 0%
22/01/22 04:32:53 INFO mapreduce.Job:  map 62% reduce 0%
22/01/22 04:32:59 INFO mapreduce.Job:  map 63% reduce 0%
22/01/22 04:33:11 INFO mapreduce.Job:  map 64% reduce 0%
22/01/22 04:33:23 INFO mapreduce.Job:  map 65% reduce 0%
22/01/22 04:33:35 INFO mapreduce.Job:  map 66% reduce 0%
22/01/22 04:33:46 INFO mapreduce.Job:  map 100% reduce 0%
22/01/22 04:33:51 INFO mapreduce.Job:  map 100% reduce 100%
22/01/22 04:33:52 INFO mapreduce.Job: Job job_1642815277676_0003 completed successfully
22/01/22 04:33:52 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=95
                FILE: Number of bytes written=397411
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=21474836629
                HDFS: Number of bytes written=25
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=703938
```

```
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=703938
                Total time spent by all reduces in occupied slots (ms)=2411
                Total time spent by all map tasks (ms)=703938
                Total time spent by all reduce tasks (ms)=2411
                Total vcore-milliseconds taken by all map tasks=703938
                Total vcore-milliseconds taken by all reduce tasks=2411
                Total megabyte-milliseconds taken by all map tasks=720832512
                Total megabyte-milliseconds taken by all reduce tasks=2468864
        Map-Reduce Framework
                Map input records=214748365
                Map output records=3
                Map output bytes=83
                Map output materialized bytes=95
                Input split bytes=129
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=95
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=3448
                CPU time spent (ms)=103730
                Physical memory (bytes) snapshot=453599232
                Virtual memory (bytes) snapshot=3835408384
                Total committed heap usage (bytes)=270008320
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=21474836500
        File Output Format Counters
                Bytes Written=25
hadoop@master:/usr/local/hadoop$
```

Application application_1642815277676_0003

User:	hadoop
Name:	TeraValidate
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sat Jan 22 04:21:54 +0000 2022
Elapsed:	11mins, 56sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

The 20 GB TeraValidate operation completes successfully in 11 minutes and 56 seconds.

Assignment #0, Part B, Section II: Runtime Comparison

Although I noted the runtime of each process above, along with screenshots for evidence, I decided to make a comparison table as well.

Process	Teragen	Terasort	Teravalidate	Total
2 GB Dataset	24 Seconds	159 Seconds	26 Seconds	209 Seconds
20 GB Dataset	309 Seconds	4,696 Seconds	716 Seconds	5,721 Seconds
20 GB runtime as percentage of 2GB runtime (1000% expected if linear scaling)	1,287.5% (12.875x length)	2,953.46% (29.534x length)	2,753.84% (27.538x length)	2,737.32% (27.373x length)

Interestingly, Teragen seems to be close to linear in scaling, with not much more than 10x the runtime for a dataset that is 10x as long. Terasort and Teravalidate, and consequently the total runtime of all 3 commands, appear to be closer to O(n log n) in complexity (base 2), which would have a scaling of 33.219x longer runtime for 10x the dataset.

Assignment #0, Part C Section I and II: Running Python Code on Hadoop

I refer to [26], provided by Assignment 0 PDF for this section of the assignment. Shakespeare dataset derived from [27] from Assignment 0 PDF.

```
#!/usr/bin/env pytho
"""mapper.py"""

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print '%s\t%s' % (word, 1)
```

```
hadoop@master:~$ pwd
/home/hadoop
hadoop@master:~$ chmod +x /home/hadoop/mapper.py
hadoop@master:~$ ls -al
total 60
drwxr-xr-x 5 hadoop hadoop  4096 Jan 23 00:45 .
drwxr-xr-x 7 root   root    4096 Jan 20 04:45 ..
-rw------- 1 hadoop hadoop  7997 Jan 22 04:36 .bash_history
-rw-r--r-- 1 hadoop hadoop   220 Apr  4  2018 .bash_logout
-rw-r--r-- 1 hadoop hadoop  3822 Jan 20 04:05 .bashrc
drwx------ 2 hadoop hadoop  4096 Jan 20 04:02 .cache
drwx------ 3 hadoop hadoop  4096 Jan 20 04:02 .gnupg
-rw-r--r-- 1 hadoop hadoop   807 Apr  4  2018 .profile
drwx------ 2 hadoop hadoop  4096 Jan 20 04:01 .ssh
-rw------- 1 hadoop hadoop 15637 Jan 23 00:45 .viminfo
-rwxrwxr-x 1 hadoop hadoop   553 Jan 23 00:44 mapper.py
hadoop@master:~$
```

I saved the mapper.py file from [26] in home/Hadoop and used “chmod +x /home/hadoop/mapper.py” to give the script executable permissions.

```python
#!/usr/bin/env python
"""reducer.py"""

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

```
hadoop@master:~$ chmod +x /home/hadoop/reducer.py
hadoop@master:~$ ls -l
total 8
-rwxrwxr-x 1 hadoop hadoop  553 Jan 23 00:44 mapper.py
-rwxrwxr-x 1 hadoop hadoop 1044 Jan 23 01:17 reducer.py
hadoop@master:~$
```

I use the reducer.py code from [26] and save it in /home/Hadoop, using “chmod +x /home/hadoop/reducer.py” to give it execution permissions to prevent issues later.

```
hadoop@master:/usr/bin$ echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py
/usr/bin/env: ‘python’: No such file or directory
```

I try to run the suggested local test of mapper.py in [26], but I get the above error.

```
EddieCFox@master:~$ whereis python3
python3: /usr/bin/python3.6m /usr/bin/python3 /usr/bin/python3.6 /usr/lib/python3.7 /usr/lib/python3 /usr/lib/python3.8 /usr/lib/python3.6 /etc/python3 /etc/python3.6 /usr/local/lib/python3
.6 /usr/share/python3 /usr/share/man/man1/python3.1.gz
EddieCFox@master:~$ sudo ln -s /usr/bin/python3 /usr/bin/python
```

[28] suggested I find python3 using whereis python3, and then create a symbolic link between python3 and the /usr/bin/python. This was successful.

```
hadoop@master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py
  File "/home/hadoop/mapper.py", line 19
    print '%s\t%s' % (word, 1)
                 ^
SyntaxError: invalid syntax
```

This was the next error I encountered. I did some searching and discovered in [29] that the code written in [26] may have done so on Python 2, while I linked the /usr/bin to Python 3, which uses print as a function rather than a statement. I redid my code for both mapper.py and reducer.py to include parentheses around the single print statement in mapper.py and the two print statements in reducer.py.

Screenshots starting next page due to sizing.

```python
#!/usr/bin/env python
"""mapper.py"""

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print('%s\t%s' % (word, 1))
```

```python
#!/usr/bin/env python
"""reducer.py"""

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print('%s\t%s' % (current_word, current_count))
```

```
hadoop@master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py
foo     1
foo     1
quux    1
labs    1
foo     1
bar     1
quux    1
hadoop@master:~$
```

```
hadoop@master:~$ echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py | sort -k1,1 | /home/hadoop/reducer.py
bar     1
foo     3
labs    1
quux    2
```

Afterwards, mapper.py and reducer.py achieved exactly the same results under test as under [26].

Mapper.py test command: “echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py”

Reducer.py test command: “echo "foo foo quux labs foo bar quux" | /home/hadoop/mapper.py | sort -k1,1 | /home/hadoop/reducer.py”

I use mkdir to create a directory under the path /tmp/Shakespeare to store the Shakespeare file. While in the directory, I run “wget https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/assignments/shakespeare.zip” to download the zip file.

```
hadoop@master:/tmp/Shakespeare$ wget https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/ass
ignments/shakespeare.zip
--2022-01-23 02:33:51--  https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/assignments/sha
kespeare.zip
Resolving mobitec.ie.cuhk.edu.hk (mobitec.ie.cuhk.edu.hk)... 137.189.99.12
Connecting to mobitec.ie.cuhk.edu.hk (mobitec.ie.cuhk.edu.hk)|137.189.99.12|:443... connected
.
HTTP request sent, awaiting response... 200 OK
Length: 1794068 (1.7M) [application/zip]
Saving to: ‘shakespeare.zip’

shakespeare.zip         100%[===========================>]   1.71M  --.-KB/s    in 0.1s

2022-01-23 02:33:51 (15.2 MB/s) - ‘shakespeare.zip’ saved [1794068/1794068]

hadoop@master:/tmp/Shakespeare$ ls -al
total 1764
drwxrwxr-x  2 hadoop hadoop    4096 Jan 23 02:33 .
drwxrwxrwt 16 root   root      4096 Jan 23 02:31 ..
-rw-rw-r--  1 hadoop hadoop 1794068 Jan 21 08:03 shakespeare.zip
```

```
EddieCFox@master:~$ sudo apt-get install unzip
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following package was automatically installed and is no longer required:
  libnuma1
Use 'sudo apt autoremove' to remove it.
Suggested packages:
  zip
The following NEW packages will be installed:
  unzip
0 upgraded, 1 newly installed, 0 to remove and 1 not upgraded.
Need to get 168 kB of archives.
After this operation, 567 kB of additional disk space will be used.
Get:1 http://asia-east1.gce.archive.ubuntu.com/ubuntu bionic-updates/main amd64 unzip amd64 6.0-21ubuntu1.1 [168 kB]
Fetched 168 kB in 0s (7330 kB/s)
Selecting previously unselected package unzip.
(Reading database ... 117201 files and directories currently installed.)
Preparing to unpack .../unzip_6.0-21ubuntu1.1_amd64.deb ...
Unpacking unzip (6.0-21ubuntu1.1) ...
Setting up unzip (6.0-21ubuntu1.1) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
```

```
hadoop@master:/tmp/Shakespeare$ unzip shakespeare.zip
Archive:  shakespeare.zip
  inflating: shakespeare
hadoop@master:/tmp/Shakespeare$ ls -al
total 6244
drwxrwxr-x  2 hadoop hadoop    4096 Jan 23 02:37 .
drwxrwxrwt 16 root   root      4096 Jan 23 02:37 ..
-rw-rw-r--  1 hadoop hadoop 4583798 Sep 20  2019 shakespeare
-rw-rw-r--  1 hadoop hadoop 1794068 Jan 21 08:03 shakespeare.zip
hadoop@master:/tmp/Shakespeare$
```

As suggested by [30], I install the unzip package, then run “unzip shakespeare.zip” to extract the Shakespeare.zip file.

After running “cat Shakespeare”, it scrolls through a lot of poetic English text, so presumably the file is in good order.

Screenshot on next page.

```
"She hangs about his neck:"
"If she pertain to life let her speak too."
"Ay, and make't manifest where she has lived,"
"Or how stolen from the dead."
"That she is living,"
"Were it but told you, should be hooted at"
"Like an old tale: but it appears she lives,"
"Though yet she speak not. Mark a little while."
"Please you to interpose, fair madam: kneel"
"And pray your mother's blessing. Turn, good lady,"
"Our Perdita is found."
"You gods, look down"
"And from your sacred vials pour your graces"
"Upon my daughter's head! Tell me, mine own."
"Where hast thou been preserved? where lived? how found"
"Thy father's court? for thou shalt hear that I,"
"Knowing by Paulina that the oracle"
"Gave hope thou wast in being, have preserved"
"Myself to see the issue."
"There's time enough for that,"
"Lest they desire upon this push to trouble"
"Your joys with like relation. Go together,"
"You precious winners all, your exultation"
"Partake to every one. I, an old turtle,"
"Will wing me to some wither'd bough and there"
"My mate, that's never to be found again,"
"Lament till I am lost."
"O, peace, Paulina!"
"Thou shouldst a husband take by my consent,"
"As I by thine a wife: this is a match,"
"And made between's by vows. Thou hast found mine,"
"But how, is to be question'd, for I saw her,"
"As I thought, dead, and have in vain said many"
"A prayer upon her grave. I'll not seek far--"
"For him, I partly know his mind--to find thee"
"An honourable husband. Come, Camillo,"
"And take her by the hand, whose worth and honesty"
"Is richly noted and here justified"
"By us, a pair of kings. Let's from this place."
"What! look upon my brother: both your pardons,"
"That e'er I put between your holy looks"
"My ill suspicion. This is your son-in-law,"
"And son unto the king, who, heavens directing,"
"Is troth-plight to your daughter. Good Paulina,"
"Lead us from hence, where we may leisurely"
"Each one demand an answer to his part"
"Perform'd in this wide gap of time since first"
"We were dissever'd: hastily lead away."
Exeunt
hadoop@master:/tmp/Shakespeare$
```

I navigate to the Hadoop directory again to execute commands.

“bin/hadoop dfs -copyFromLocal /tmp/Shakespeare /user/hadoop/Shakespeare”

```
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -copyFromLocal /tmp/Shakespeare /user/hadoop/Shakespeare
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

hadoop@master:/usr/local/hadoop$
hadoop@master:/usr/local/hadoop$
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2022-01-23 02:44 Shakespeare
drwxr-xr-x   - hadoop supergroup          0 2022-01-22 04:21 terasort
```

We use the commands in [26] to copy the contents of /tmp/Shakespeare to the hdfs. When you use the -ls commands, I see that there is both the Shakespeare and terasort items copied over. However, I realize that I also copied over the Shakespeare.zip, which could cause issues later.

Note: I realize it says that bin/hadoop dfs is deprecated but I couldn’t easily find the hdfs command to use instead, and it was still performing the intended function.

```
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -ls Shakespeare
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 hadoop supergroup    4583798 2022-01-23 02:44 Shakespeare/shakespeare
-rw-r--r--   1 hadoop supergroup    1794068 2022-01-23 02:44 Shakespeare/shakespeare.zip
hadoop@master:/usr/local/hadoop$
```

As I thought. Command derived through trial and error after the base -ls command.

```
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -rm Shakespeare/shakespeare.zip
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Deleted Shakespeare/shakespeare.zip
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -ls Shakespeare
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 1 items
-rw-r--r--   1 hadoop supergroup    4583798 2022-01-23 02:44 Shakespeare/shakespeare
```

I delete the Shakespeare.zip after using -rm instead of -ls and confirm that the Shakespeare direction now only has the raw Shakespeare file.

We are now ready to run the map reduce job. Command modified from [26] and Assignment 0 PDF command.

Command: “./bin/hadoop jar ./share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -file /home/hadoop/mapper.py -mapper /home/hadoop/mapper.py -file /home/hadoop/reducer.py -

reducer /home/hadoop/reducer.py -input /user/hadoop/Shakespeare/* -output /user/hadoop/Shakespeare-output”

```
22/01/23 03:07:27 INFO mapreduce.Job: Task Id : attempt_1642905725026_0001_m_000001_2, Status : FAILED
Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127
        at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)
        at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)
        at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)
        at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
        at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)
        at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:459)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:343)
        at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:177)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1893)
        at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:171)

22/01/23 03:07:31 INFO mapreduce.Job:  map 50% reduce 100%
22/01/23 03:07:32 INFO mapreduce.Job:  map 100% reduce 100%
22/01/23 03:07:32 INFO mapreduce.Job: Job job_1642905725026_0001 failed with state FAILED due to: Task failed task_1642905725026_0001_m_000001
Job failed as tasks failed. failedMaps:1 failedReduces:0

22/01/23 03:07:32 INFO mapreduce.Job: Counters: 14
        Job Counters
                Failed map tasks=7
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=8
                Other local map tasks=6
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=29959
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=29959
                Total vcore-milliseconds taken by all map tasks=29959
                Total megabyte-milliseconds taken by all map tasks=30678016
        Map-Reduce Framework
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
22/01/23 03:07:32 ERROR streaming.StreamJob: Job not successful!
Streaming Command Failed!
hadoop@master:/usr/local/hadoop$
```

The command fails with numerous errors that are not included here. The primary error is “Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127”.

I modify the -mapper and -reducer switches to be more like the Assignment 0 PDF instead of [26], then try the following command: “./bin/hadoop jar ./share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -file /home/hadoop/mapper.py -mapper ‘mapper.py’ -file /home/hadoop/reducer.py -reducer ‘reducer.py’ -input /user/hadoop/Shakespeare/* -output /user/hadoop/Shakespeare-output”

This command initially failed, saying Shakespeare-output already exists.

```
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -rm Shakespeare-output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

rm: `Shakespeare-output': Is a directory
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -rmr Shakespeare-output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

rmr: DEPRECATED: Please use '-rm -r' instead.
Deleted Shakespeare-output
```

I ran “bin/Hadoop dfs -rmr Shakespeare-output” to delete the directory from hdfs. I ended up needing a number of attempts to finally figure out the source of the error above, and every time I failed, I would get the exact same error code as previously mentioned, and also need to delete the Shakespeare-output directory from hdfs.

```
Deleted Shakespeare-output
hadoop@master:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -file /home/hadoop/mapper.py -mapper 'mapper.py' -file /home/hadoop/reducer.py -reducer
'reducer.py' -input /user/hadoop/Shakespeare/* -output /user/hadoop/Shakespeare-output
22/01/23 03:46:11 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/hadoop/mapper.py, /home/hadoop/reducer.py, /tmp/hadoop-unjar1230010688370356616/] [] /tmp/streamjob2083528573561914899.jar tmpDir=null
22/01/23 03:46:11 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/23 03:46:12 INFO client.RMProxy: Connecting to ResourceManager at master/10.140.0.2:8040
22/01/23 03:46:12 INFO mapred.FileInputFormat: Total input files to process : 1
22/01/23 03:46:12 INFO mapreduce.JobSubmitter: number of splits:2
22/01/23 03:46:13 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
22/01/23 03:46:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1642905725026_0005
22/01/23 03:46:13 INFO impl.YarnClientImpl: Submitted application application_1642905725026_0005
22/01/23 03:46:13 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1642905725026_0005/
22/01/23 03:46:13 INFO mapreduce.Job: Running job: job_1642905725026_0005
22/01/23 03:46:19 INFO mapreduce.Job: Job job_1642905725026_0005 running in uber mode : false
22/01/23 03:46:19 INFO mapreduce.Job:  map 0% reduce 0%
22/01/23 03:46:27 INFO mapreduce.Job:  map 100% reduce 0%
22/01/23 03:46:34 INFO mapreduce.Job:  map 100% reduce 100%
22/01/23 03:46:34 INFO mapreduce.Job: Job job_1642905725026_0005 completed successfully
22/01/23 03:46:34 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=7841073
                FILE: Number of bytes written=16288379
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4585871
                HDFS: Number of bytes written=878840
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=12852
                Total time spent by all reduces in occupied slots (ms)=4199
                Total time spent by all map tasks (ms)=12852
                Total time spent by all reduce tasks (ms)=4199
                Total vcore-milliseconds taken by all map tasks=12852
                Total vcore-milliseconds taken by all reduce tasks=4199
                Total megabyte-milliseconds taken by all map tasks=13160448
                Total megabyte-milliseconds taken by all reduce tasks=4299776
        Map-Reduce Framework
                Map input records=111396
                Map output records=814671
                Map output bytes=6211725
                Map output materialized bytes=7841079
                Input split bytes=212
```

Next screenshot on next page due to sizing.

```
        Map-Reduce Framework
                Map input records=111396
                Map output records=814671
                Map output bytes=6211725
                Map output materialized bytes=7841079
                Input split bytes=212
                Combine input records=0
                Combine output records=0
                Reduce input groups=80364
                Reduce shuffle bytes=7841079
                Reduce input records=814671
                Reduce output records=80364
                Spilled Records=1629342
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=358
                CPU time spent (ms)=5990
                Physical memory (bytes) snapshot=743759872
                Virtual memory (bytes) snapshot=5765464064
                Total committed heap usage (bytes)=505413632
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=4585659
        File Output Format Counters
                Bytes Written=878840
22/01/23 03:46:34 INFO streaming.StreamJob: Output directory: /user/hadoop/Shakespeare-output
hadoop@master:/usr/local/hadoop$
```

Application application_1642905725026_0005

User:	hadoop
Name:	streamjob2083528573561914899.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Jan 23 03:46:13 +0000 2022
Elapsed:	20sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

The Mapreduce command for the python script section completed successfully in 20 seconds.

The error was caused because slaves 1-3 did not have the symbolic link I created on the master node in [28] when testing the mapper.py and reducer.py scripts locally. I recalled the Big Data

lectures where the professor said that each node was running the same code, so the other nodes would have the same errors the master did before the symbolic link. In order to fix this, I ran "sudo ln -s /usr/bin/python3 /usr/bin/python" on all 3 slave nodes. Afterwards, the mapreduce job completed without any issues.

I had a hint this was the issue through [31], but I was initially worried that copy pasting the mapper and reducer code from [26] inserted invisible characters. Redoing the script manually did not help however, so I later thought of copying the solution from [28] to the other nodes by myself.

```
22/01/23 03:46:34 INFO streaming.StreamJob: Output directory: /user/hadoop/Shakespeare-output
hadoop@master:/usr/local/hadoop$ bin/hadoop dfs -ls /user/hadoop/Shakespeare-output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2022-01-23 03:46 /user/hadoop/Shakespeare-output/_SUCCESS
-rw-r--r--   1 hadoop supergroup     878840 2022-01-23 03:46 /user/hadoop/Shakespeare-output/part-00000
```

As suggested in [26], I use the dfs -ls command in the Shakespeare-output directory to figure out where I should direct the cat command. We can see that content was inserted into the Shakespeare-output directory successfully.

To verify the contents, I output it using (from [26]): "bin/hadoop dfs -cat /user/hadoop/Shakespeare-output/part-00000"

Screenshot on next page.

```
"'      1
"''Tis  2
"'--O   1
"'A     3
"'Above 1
"'Achilles      1
"'Ad    2
"'Affectionate  1
"'After 1
"'Against       1
"'Aio   1
"'Alack,        1
"'Alas, 1
"'Alla  1
"'An    2
"'And   6
"'Anon  1
"'Aroint        1
"'As    3
"'Ask   1
"'Ay,   1
"'B     1
"'Ban,  1
"'Be    2
"'Because       1
"'Besides,      1
"'Better        1
"'Boarding,'    1
"'Bove  1
"'Break 1
"'Bring 1
"'Brutus,       1
"'But   5
"'By    2
"'C     1
"'Caesar,       1
"'Call  1
"'Clarence      1
"'Come  1
"'Come, 1
"'Commend       1
"'Con   1
"'Confess'      1
"'Conscience,'  1
"'Convey,'      1
"'Couple        1
"'Cucullus      1
"'D     1
"'Dear  1
--More--
```

```
yourselves!"    1
yourselves"     8
yourselves,     8
yourselves,"    10
yourselves--"   1
yourselves.     1
yourselves."    9
yourselves:     3
yourselves:"    3
yourselves?     2
yourselves?"    2
youth   90
youth!  1
youth!" 5
youth"  27
youth's 4
youth,  36
youth," 32
youth.  4
youth." 12
youth.'"        1
youth:  8
youth:" 7
youth?" 5
youthful        24
youthful,       1
youths  5
zanies."        1
zany,"  1
zeal    16
zeal!   1
zeal"   3
zeal,   3
zeal,"  3
zeal."  2
zeal:"  2
zealous 5
zeals,  1
zed!    1
zenith  1
zephyrs 1
zir,    1
zir:    1
zo      1
zodiac  1
zodiacs 1
zone,"  1
zounds,"        1
zwaggered       1
hadoop@master:/usr/local/hadoop$
```

The result is a very long list (mostly omitted here) of every word and its respective result. The results look sensible at first glance. It splits words such as "water" 'water' and water differently, but this behavior is expected per [26].

After finishing the assignment, I took an image of each instance and powered them all off so I wouldn't be billed further (in my trial credits) until I need them for the next assignment. So far I have used 268 hkd of my 2399 hkd trial credit, leaving me with 2131 hkd and 87 days left.

References:

[1] https://cloud.google.com/compute/docs/quickstart-linux

[2] https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2020/tutorial/tutorial1/launch_VMs.pdf

[3] https://talentopia.global/web/google-cloud-compute-engine-how-to-choose-the-one-suits-your-specific-workloads

[4] https://cloud.google.com/compute/docs/connect/create-ssh-keys#windows

[5] https://cloud.google.com/compute/docs/instances/connecting-advanced#provide-key

[6] https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2020/tutorial/tutorial1/

[7] https://mobitec.ie.cuhk.edu.hk/ierg4300Fall2020/tutorial/tutorial2/

[8] https://cwiki.apache.org/confluence/display/HADOOP/Hadoop+Java+Versions

[9] https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/SingleCluster.html

[10] https://linux.die.net/man/8/useradd

[11] https://www.cyberciti.biz/faq/linux-list-all-members-of-a-group/

[12] https://archive.apache.org/dist/hadoop/core/hadoop-2.9.2/

[13] https://askubuntu.com/questions/636222/mapred-site-xml-file-not-present-in-hadoop-2-7-0

[14] https://stackoverflow.com/questions/14325594/working-with-hadoop-localhost-error-java-home-is-not-set

[15] https://www.godaddy.com/garage/how-to-set-up-an-ssh-tunnel-with-putty/

[16] https://hadoop.apache.org/docs/r2.9.2/hadoop-project-dist/hadoop-common/ClusterSetup.html

[17] https://stackoverflow.com/questions/22193051/how-to-configure-hosts-file-for-hadoop-ecosystem

[18] https://zhuanlan.zhihu.com/p/58968191

[19] https://cloud.google.com/compute/docs/instances/stop-start-instance

[20] https://haydenjames.io/linux-securely-copy-files-using-scp/

[21] https://stackoverflow.com/questions/29680155/why-there-is-a-mapreduce-jobtracker-address-configuration-on-yarn

[22] https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-common/yarn-default.xml

[23] https://stackoverflow.com/questions/58450539/error-4003-cant-ssh-login-into-the-instance-that-i-created-in-google-cloud-pla

[24] https://medium.com/@satyavinay456/setting-up-a-multi-node-cluster-in-hadoop-2-9-2-dba73c5465a9

[25] https://stackoverflow.com/questions/19592744/hadoop-reduce-stops-running

[26] https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

[27] https://mobitec.ie.cuhk.edu.hk/iems5730/static_files/assignments/shakespeare.zip

[28] https://stackoverflow.com/questions/3655306/ubuntu-usr-bin-env-python-no-such-file-or-directory

[29] https://stackoverflow.com/questions/937491/invalid-syntax-when-using-print

[30] https://askubuntu.com/questions/86849/how-to-unzip-a-zip-file-from-the-terminal

[31] https://stackoverflow.com/questions/4460522/hadoop-streaming-job-failed-error-in-python