Hadoop Installation Guide and HDFS Hands On - H1

Every step is to be executed on the home directory. Use `cd` to move to home directory.

The commands in the guide use `USER` as the notation for your username. If you have executed A0 correctly, then this should be your SRN in lowercase. This is important since the auto-evaluation depends on it. Verify your username by running `whoami` on the terminal.

Change any `/home/USER/` to `/home/<your SRN>/`

This manual includes steps that you will be doing in the classroom. It assumes that you have completed steps 1 and 3 from your home which was emailed earlier. If you have not completed these steps, then click here to do so.

Execute the following commands to move to the home directory and updating the package list and the system. This guide assumes that you are working with Ubuntu or a Debian based distribution.

```
cd
sudo apt update -y
sudo apt upgrade -y
```

Step 1 - Setup passwordless SSH for Hadoop

We install the following packages to allow us to setup an ssh server on the system as well as a client to remote into it with the following commands.

```
sudo apt install openssh-server openssh-client -y
```

Enable passwordless SSH

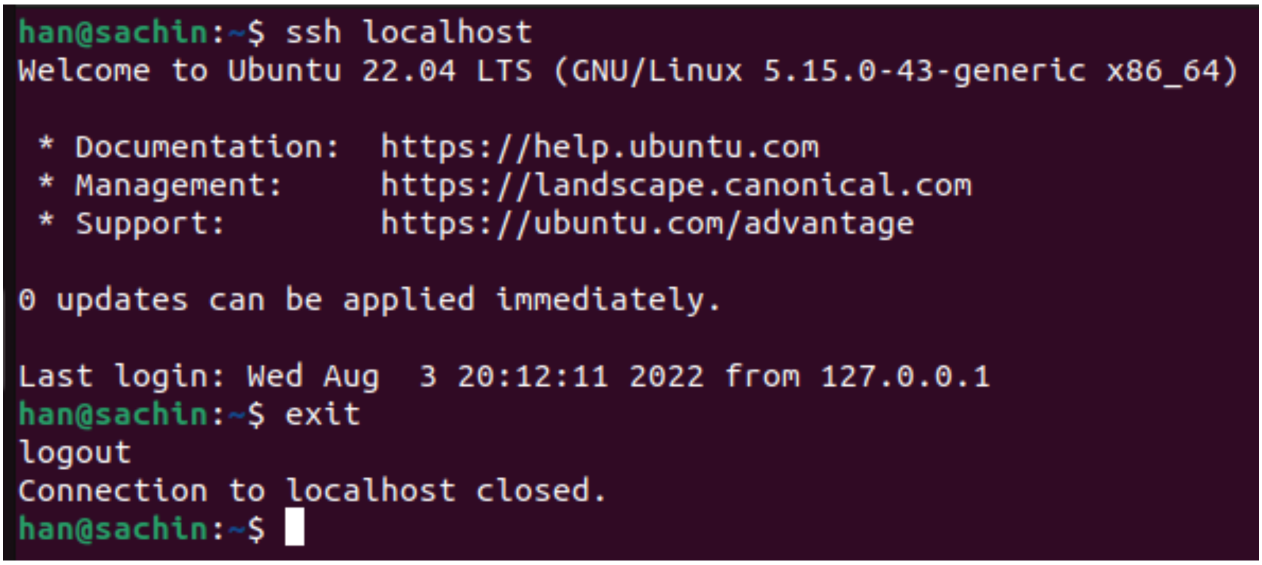
Generate an SSH key pair and define the location is is to be stored in id_rsa. Then use the cat command to store the public key as authorized_keys in the ssh directory. Follow these exact commands with change in permissions.

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```

Verify passwordless SSH is setup and working with

```
ssh localhost
```

If the above command does not ask you for a password, you have successfully setup passwordless SSH. Take a screenshot of the terminal and name is 1a.png.



```
han@sachin:~$ ssh localhost
Welcome to Ubuntu 22.04 LTS (GNU/Linux 5.15.0-43-generic x86_64)

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

0 updates can be applied immediately.

Last login: Wed Aug  3 20:12:11 2022 from 127.0.0.1
han@sachin:~$ exit
logout
Connection to localhost closed.
han@sachin:~$ 
```

Type `exit` or press `Ctrl+d` to quit the SSH session.

Step 2 - Single Node Deployment

The current setup is called pseudo-distributed mode, allows each Hadoop daemon to run as a single Java process. A Hadoop environment is configured by editing the following list of configuration files:

- .bashrc
- hadoop-env.sh
- core-site.xml
- hdfs-site.xml
- mapred-site-xml
- yarn-site.xml

Before editing the above mentioned files, we need to make a few directories for our namenodes and datanodes along with the required permissions.

```
cd
mkdir dfsdata
mkdir tmpdata
mkdir dfsdata/datanode
mkdir dfsdata/namenode
```

Change permissions for the directories using the following commands. Remember to replace `USER` with your username.

```
sudo chown -R USER:USER /home/USER/dfsdata/
sudo chown -R USER:USER /home/USER/dfsdata/datanode/
sudo chown -R USER:USER /home/USER/dfsdata/namenode/
```

Editing and Setting up the ~/.bashrc config file

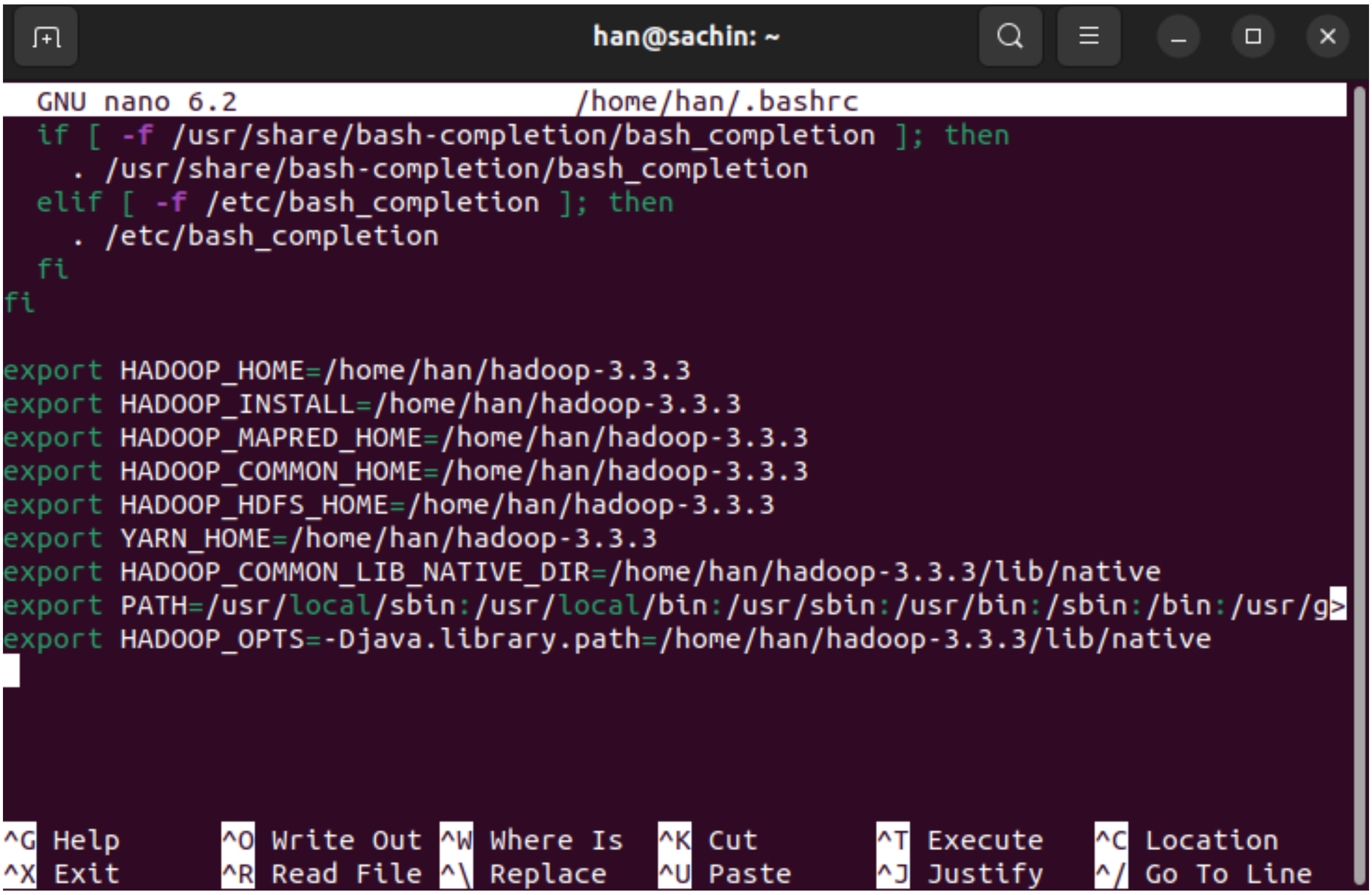
Open .bashrc with any text editor of your choice. This guide recommends using `nano`.

```
sudo nano ~/.bashrc
```

Scroll to the bottom of the file. Copy and paste the below mentioned statements to the end of the file.

```
#Hadoop Path Configs
export HADOOP_HOME=/home/USER/hadoop-3.3.3
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="Djava.library.path=$HADOOP_HOME/lib/native"
```

Take a screenshot at this stage and name it 2a.png. Make sure the filename on the top with the path is visible in the screenshot.



```
  GNU nano 6.2                    /home/han/.bashrc
  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 HADOOP_HOME=/home/han/hadoop-3.3.3
export HADOOP_INSTALL=/home/han/hadoop-3.3.3
export HADOOP_MAPRED_HOME=/home/han/hadoop-3.3.3
export HADOOP_COMMON_HOME=/home/han/hadoop-3.3.3
export HADOOP_HDFS_HOME=/home/han/hadoop-3.3.3
export YARN_HOME=/home/han/hadoop-3.3.3
export HADOOP_COMMON_LIB_NATIVE_DIR=/home/han/hadoop-3.3.3/lib/native
export PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/g>
export HADOOP_OPTS=-Djava.library.path=/home/han/hadoop-3.3.3/lib/native
```

Press `Ctrl+o` to save and `Ctrl+x` to exit nano . Apply changes to `bash` with the following command.

```
source ~/.bashrc
```

You can verify if the changes have been made by using the `echo` command and checking if the corresponding path gets printed in the terminal.

```
echo $HADOOP_HOME
echo $PATH
```

Setup hadoop-env.sh

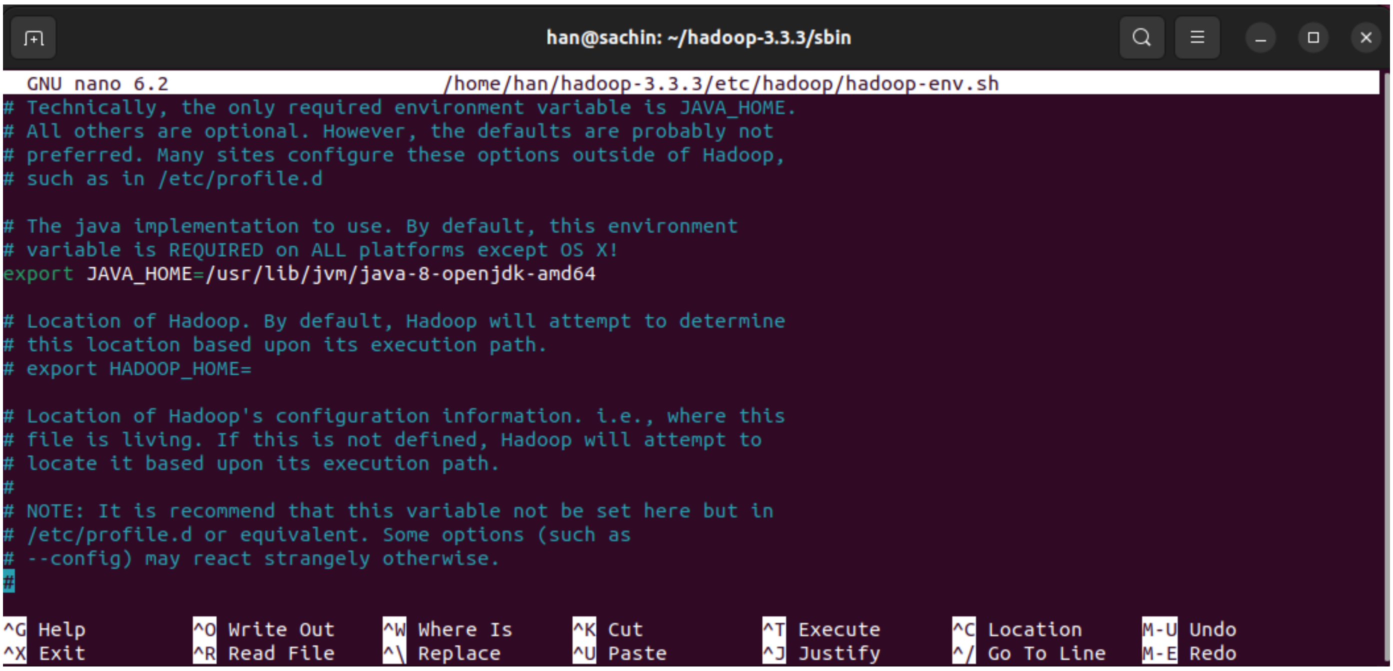
Open the file with

```
sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

Scroll down until you find the commented line `# export JAVA_HOME=`. Uncomment the line and replace the path with your Java path. The final line should look like this

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

Take a screenshot at this stage and name it 2b.png. Make sure the filename on the top with the path is visible in the screenshot.



```
han@sachin: ~/hadoop-3.3.3/sbin
  GNU nano 6.2                     /home/han/hadoop-3.3.3/etc/hadoop/hadoop-env.sh
# Technically, the only required environment variable is JAVA_HOME.
# All others are optional. However, the defaults are probably not
# preferred. Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Location of Hadoop. By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=

# Location of Hadoop's configuration information. i.e., where this
# file is living. If this is not defined, Hadoop will attempt to
# locate it based upon its execution path.
#
# NOTE: It is recommend that this variable not be set here but in
# /etc/profile.d or equivalent. Some options (such as
# --config) may react strangely otherwise.
#

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location    M-U Undo
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line  M-E Redo
```

Save and exit the file as shown previously.

Setup core-site.xml

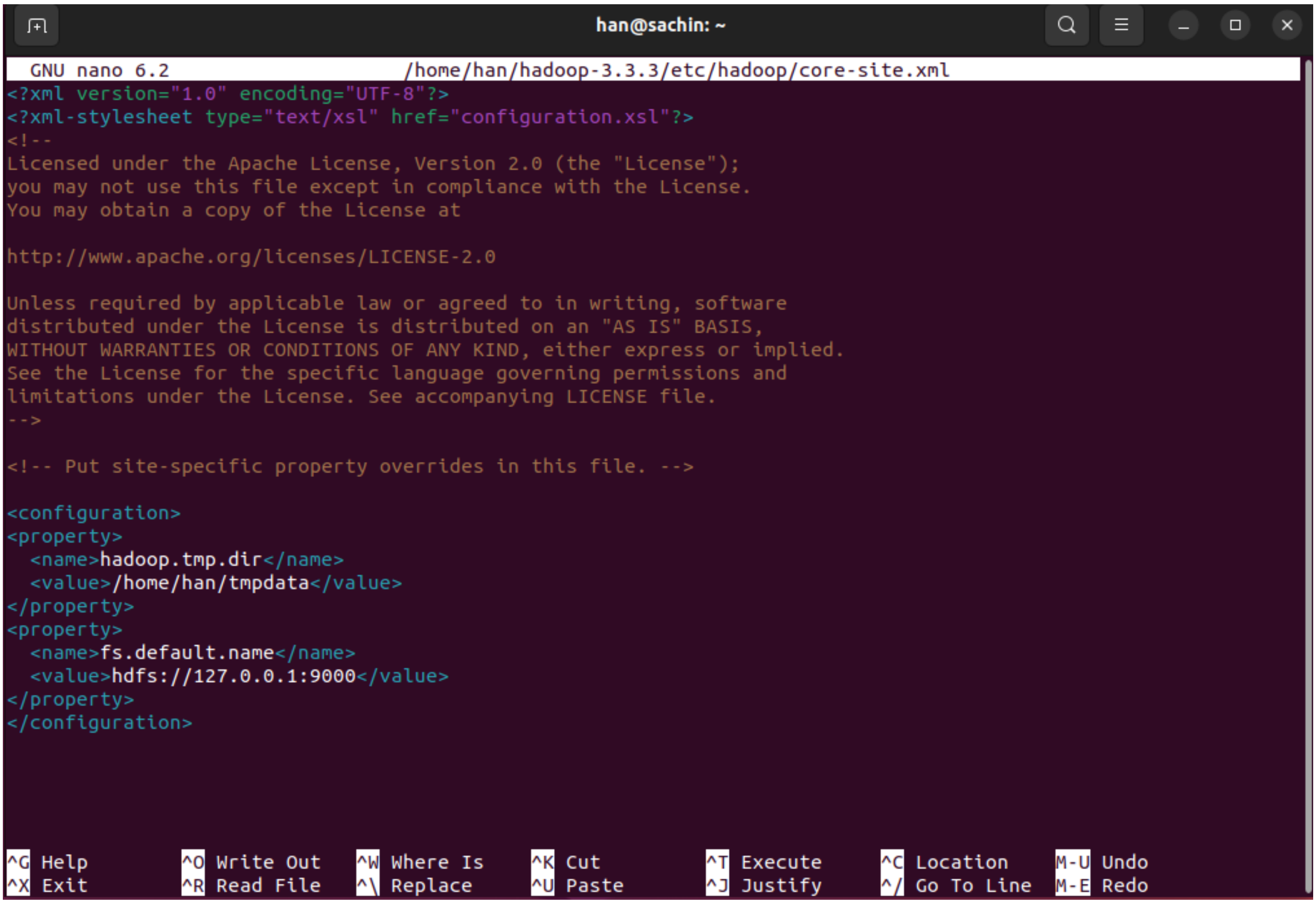
Open the file with

```
sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

Replace the existing configuration tags with the following

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/USER/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

Take a screenshot of the file and name it 2c.png. Make sure the filename on the top with the path is visible in the screenshot.



```
  GNU nano 6.2                  /home/han/hadoop-3.3.3/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

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

<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/han/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

Save and exit the file.

Setup hdfs-site.xml

Open the file using

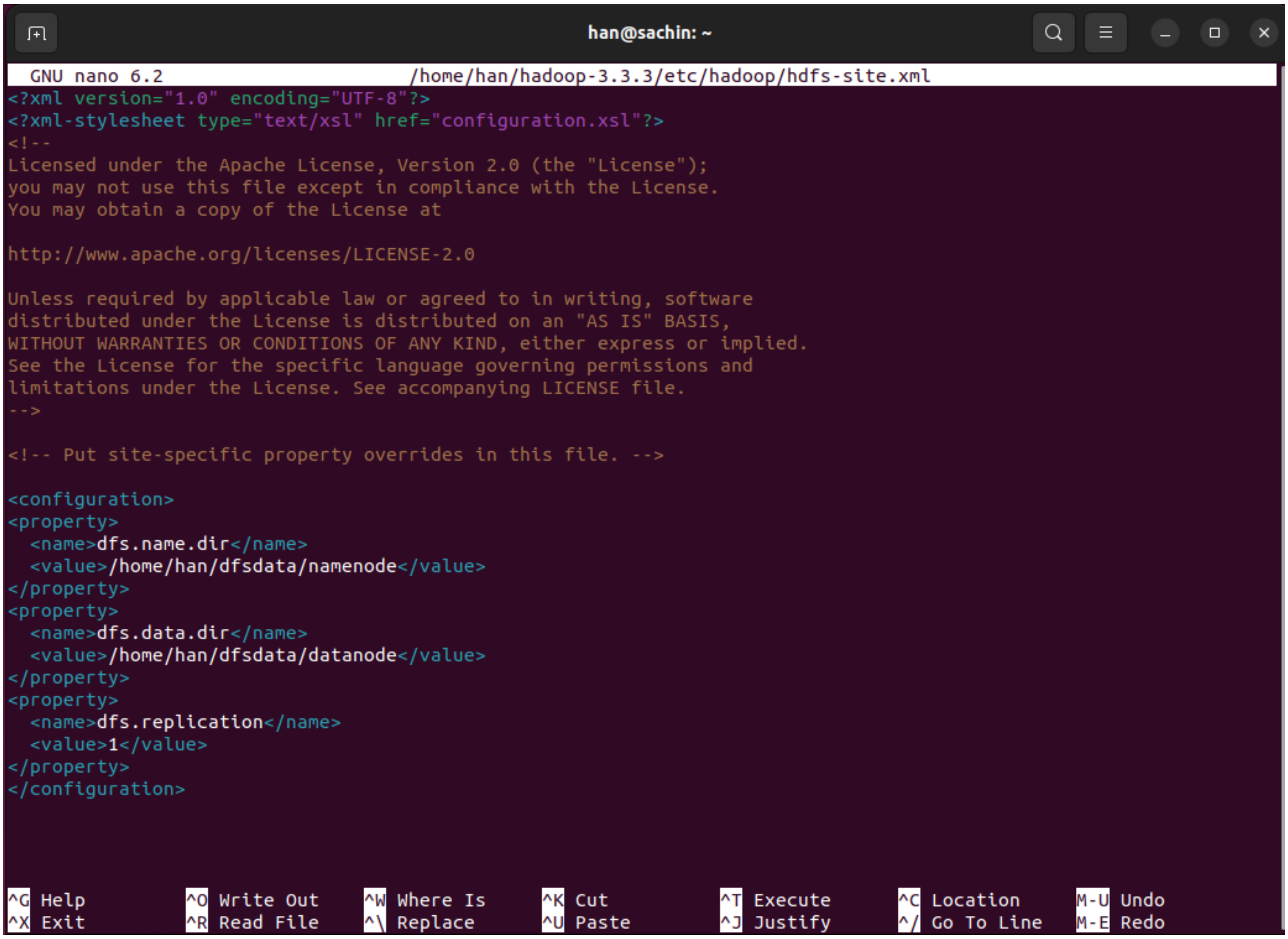
```
sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

Replace the existing configuration tags with the following

```
<configuration>
<property>
  <name>dfs.name.dir</name>
  <value>/home/USER/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/USER/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
```

```
</property>
</configuration>
```

Take a screenshot of the file and name it 2d.png. Make sure the filename on the top with the path is visible in the screenshot.



```
  GNU nano 6.2                    /home/han/hadoop-3.3.3/etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

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

<configuration>
<property>
  <name>dfs.name.dir</name>
  <value>/home/han/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/han/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

Save and exit the file after making all the changes.

Setup mapred-site.xml

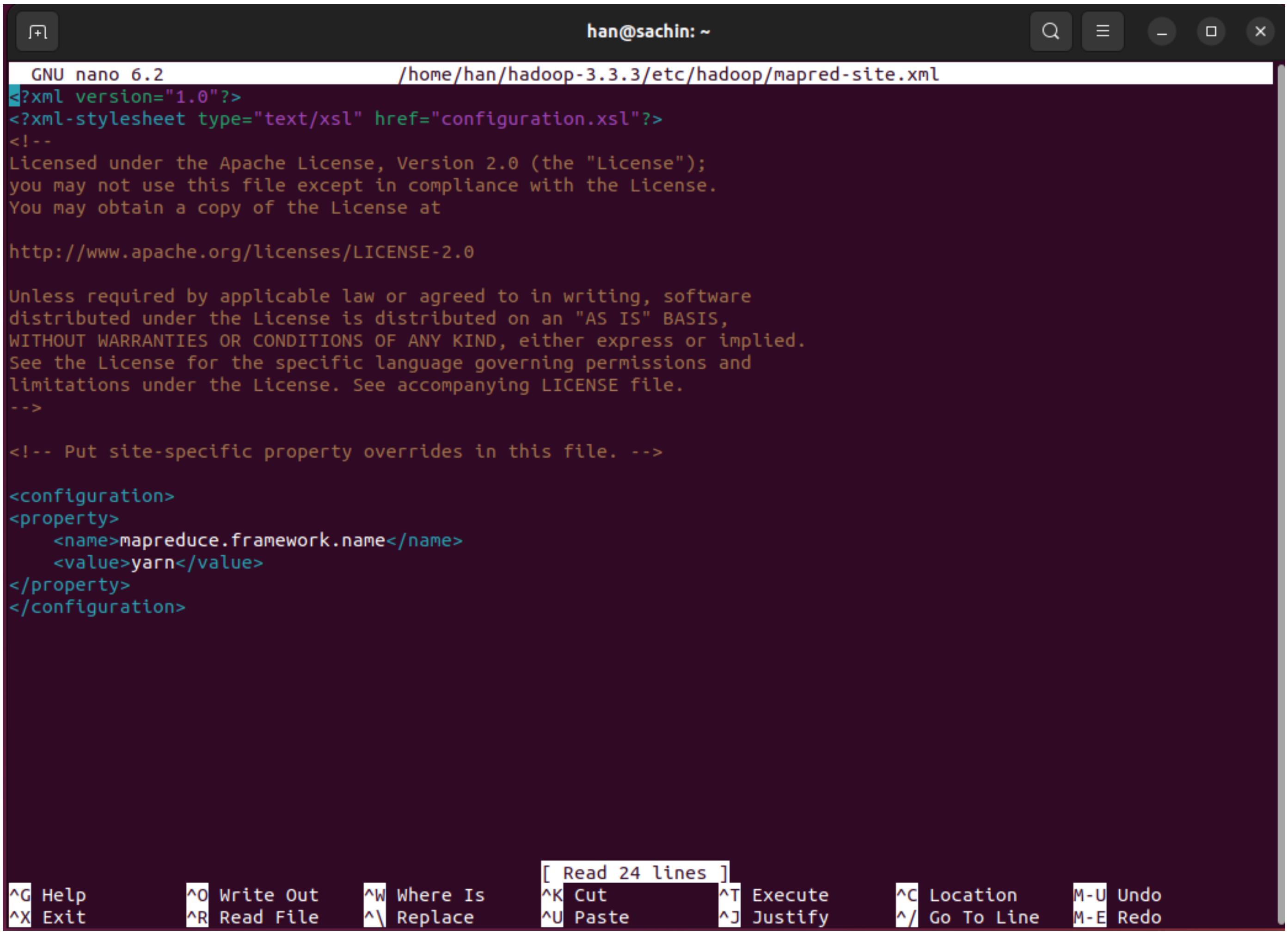
Open the file with

```
sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

Replace the existing configuration tags with the following

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

Take a screenshot of the file and name it 2e.png. Make sure the filename on the top with the path is visible in the screenshot.



```
  GNU nano 6.2                    /home/han/hadoop-3.3.3/etc/hadoop/mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

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

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

                                                 [ Read 24 lines ]
^G Help         ^O Write Out    ^W Where Is     ^K Cut          ^T Execute      ^C Location     M-U Undo
^X Exit         ^R Read File    ^\ Replace      ^U Paste        ^J Justify      ^/ Go To Line   M-E Redo
```

Save and exit the file.

Setup yarn-site.xml

Open the file with

```
sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
```

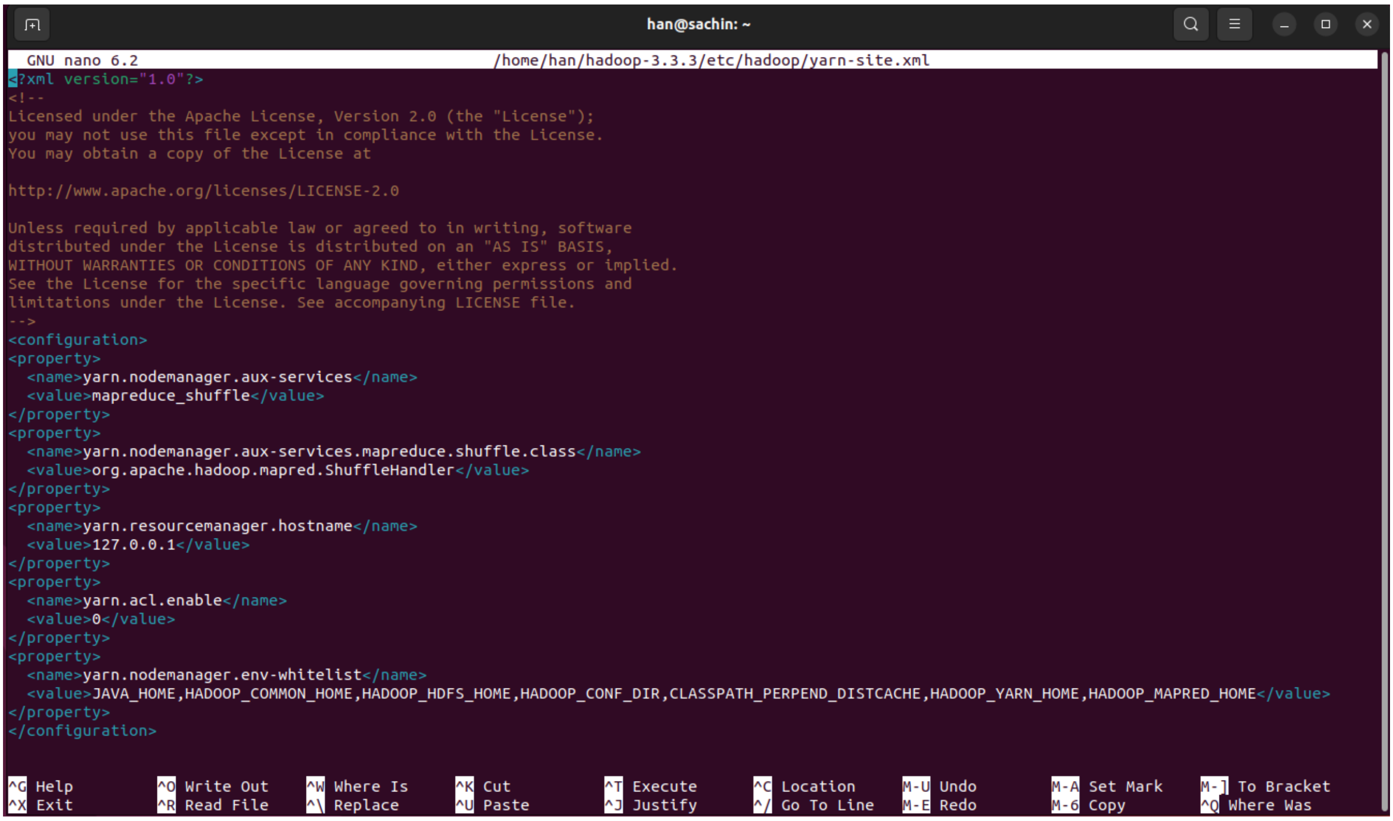
Replace the existing configuration tags with the following

```
<configuration>
```

```
  <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.hostname</name>
    <value>127.0.0.1</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>

  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLA
  SSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
  </configuration>
```

Take a screenshot of the file and name it 2f.png. Make sure the filename on the top with the path is visible in the screenshot.



```
  GNU nano 6.2                                     /home/han/hadoop-3.3.3/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<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.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>
  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>


^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute     ^C Location    M-U Undo       M-A Set Mark   M-] To Bracket
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify     ^/ Go To Line  M-E Redo       M-6 Copy       ^Q Where Was
```

Save and exit the file.

Step 3 - Format HDFS NameNode

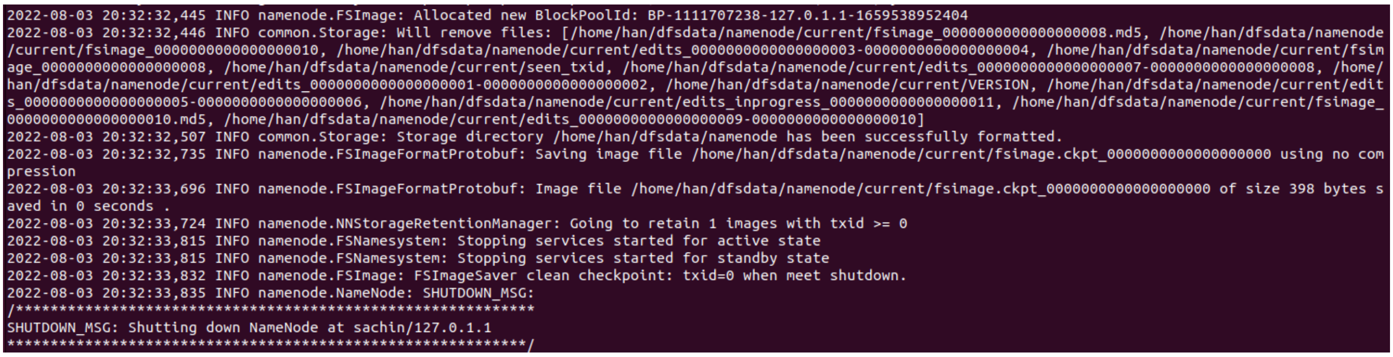
Before starting Hadoop for the first time, the namenode must be formatted. Use the following command.

```
hdfs namenode -format
```

A SHUTDOWN message will signify the end of the formatting process.

If you have reached this stage, it signifies that you have successfully installed hadoop.

Take a screenshot of the terminal output indicating the shutdown message and name it 3a.png.



```
2022-08-03 20:32:32,445 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1111707238-127.0.1.1-1659538952404
2022-08-03 20:32:32,446 INFO common.Storage: Will remove files: [/home/han/dfsdata/namenode/current/fsimage_0000000000000000008.md5, /home/han/dfsdata/namenode
/current/fsimage_0000000000000000010, /home/han/dfsdata/namenode/current/edits_0000000000000000003-0000000000000000004, /home/han/dfsdata/namenode/current/fsim
age_0000000000000000008, /home/han/dfsdata/namenode/current/seen_txid, /home/han/dfsdata/namenode/current/edits_0000000000000000007-0000000000000000008, /home/
han/dfsdata/namenode/current/edits_0000000000000000001-0000000000000000002, /home/han/dfsdata/namenode/current/VERSION, /home/han/dfsdata/namenode/current/edit
s_0000000000000000005-0000000000000000006, /home/han/dfsdata/namenode/current/edits_inprogress_0000000000000000011, /home/han/dfsdata/namenode/current/fsimage_
0000000000000000010.md5, /home/han/dfsdata/namenode/current/edits_0000000000000000009-0000000000000000010]
2022-08-03 20:32:32,507 INFO common.Storage: Storage directory /home/han/dfsdata/namenode has been successfully formatted.
2022-08-03 20:32:32,735 INFO namenode.FSImageFormatProtobuf: Saving image file /home/han/dfsdata/namenode/current/fsimage.ckpt_0000000000000000000 using no com
pression
2022-08-03 20:32:33,696 INFO namenode.FSImageFormatProtobuf: Image file /home/han/dfsdata/namenode/current/fsimage.ckpt_0000000000000000000 of size 398 bytes s
aved in 0 seconds .
2022-08-03 20:32:33,724 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2022-08-03 20:32:33,815 INFO namenode.FSNamesystem: Stopping services started for active state
2022-08-03 20:32:33,815 INFO namenode.FSNamesystem: Stopping services started for standby state
2022-08-03 20:32:33,832 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2022-08-03 20:32:33,835 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at sachin/127.0.1.1
************************************************************/
```

Step 4 - Starting Hadoop

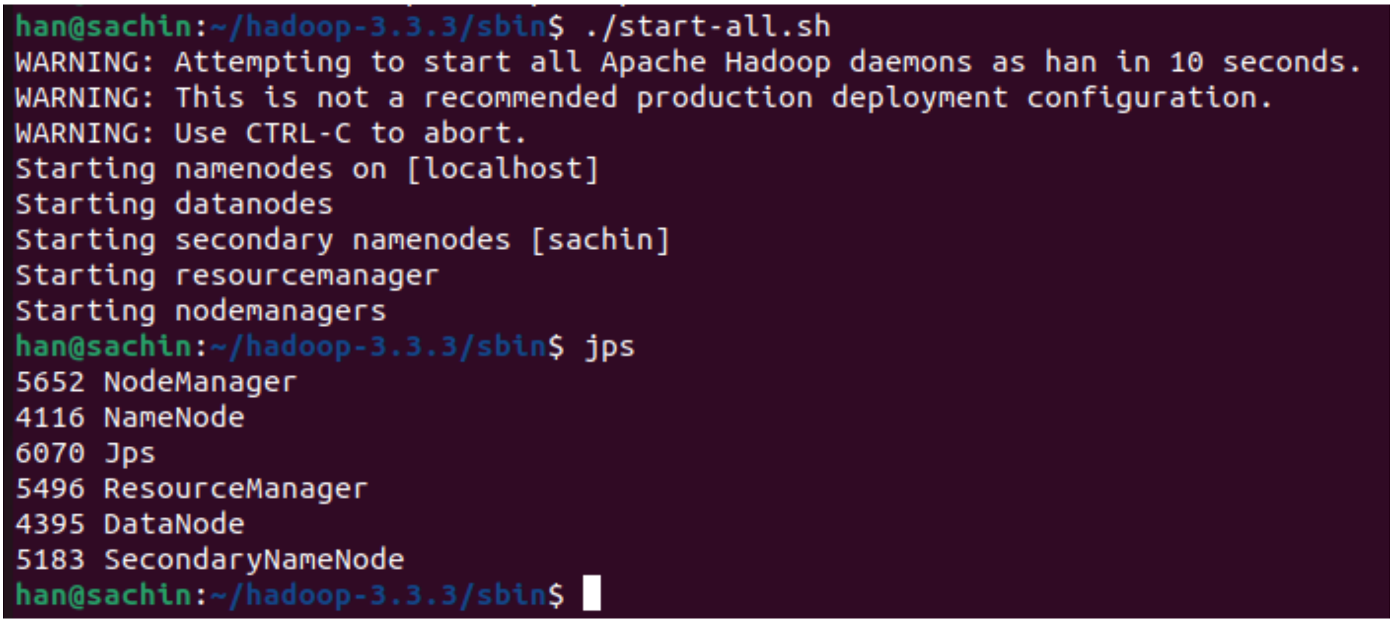
Navigate to the `hadoop` folder and execute the following commands. `start-all.sh` is a shell script that is used to start all the processes that hadoop requires.

```
cd
cd hadoop-3.3.3/sbin/
./start-all.sh
```

Type `jps` to find all the Java Processes started by the shell script. You should see a total of 6 processes, including the `jps` process. Note that the order of the items and the process IDs will be different.

```
2994 DataNode
3219 SecondaryNameNode
3927 Jps
3431 ResourceManager
2856 NameNode
3566 NodeManager
```

Take a screenshot of the terminal output and name it 4a.png.



```
han@sachin:~/hadoop-3.3.3/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as han in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [sachin]
Starting resourcemanager
Starting nodemanagers
han@sachin:~/hadoop-3.3.3/sbin$ jps
5652 NodeManager
4116 NameNode
6070 Jps
5496 ResourceManager
4395 DataNode
5183 SecondaryNameNode
han@sachin:~/hadoop-3.3.3/sbin$
```

Step 5 - Accessing Hadoop from the Browser

You can access Hadoop on `localhost` on the following ports

- NameNode - http://localhost:9870
- DataNode - http://localhost:9864

- YARN Manager - http://localhost:8088

Step 6 - Running WordCount

Now, we will run a sample HDFS commmand to calculate the frequency of a particular word in a text file.
Firstly, clone the GitHub repository.

```
git clone https://github.com/Cloud-Computing-Big-Data/UE20CS322-H1.git
```

This repo contains a sample mapper, reducer and a dataset file named tech.txt. Run the following commands to setup HDFS directories and copy the dataset file to the HDFS.

```
cd UE20CS322-H1/
hdfs dfs -mkdir /handson
hdfs dfs -put tech.txt /handson
chmod +x *.py
```

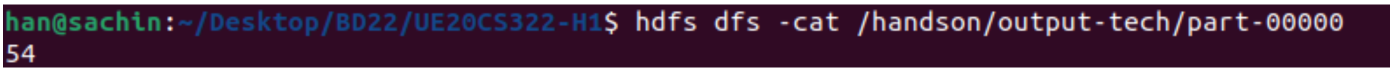
Next, run the following command to run the wordcount program.

```
hadoop jar /home/USER/hadoop-3.3.3/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar \
    -mapper "$PWD/mapper.py" \
    -reducer "$PWD/reducer.py 'perseus'" \
    -input /handson/tech.txt \
    -output /handson/output-tech
```

To check the output, execute the following command.

```
hdfs dfs -cat /handson/output-tech/part-00000
```

Take a screenshot of the terminal output and name it 6a.png.



```
han@sachin:~/Desktop/BD22/UE20CS322-H1$ hdfs dfs -cat /handson/output-tech/part-00000
54
```

Step 7 - Auto-evaluation

Auto-evaluation is allowed only once. So make sure you have the following checklist ticked before proceeding.

- ○ JPS has 6 processes(including the JPS process)
- ○ Step 6 - Running WordCount - Output is correct

For RR campus students, run the following command

```
python3 eval-rr.pyc
```

For EC campus students, run the following command

```
python3 eval-ec.pyc
```

You can see your score in the terminal output after the program finishes.

To stop all processes when you are done with your work, execute the following command.

```
cd
cd hadoop-3.3.3/sbin/
./stop-all.sh
```

Step 8 - Final Assessment

Make a word document with all the screenshots and upload the PDF to this link. Your file should be named with the format `PES1UG20CS999.pdf` with your SRN.