

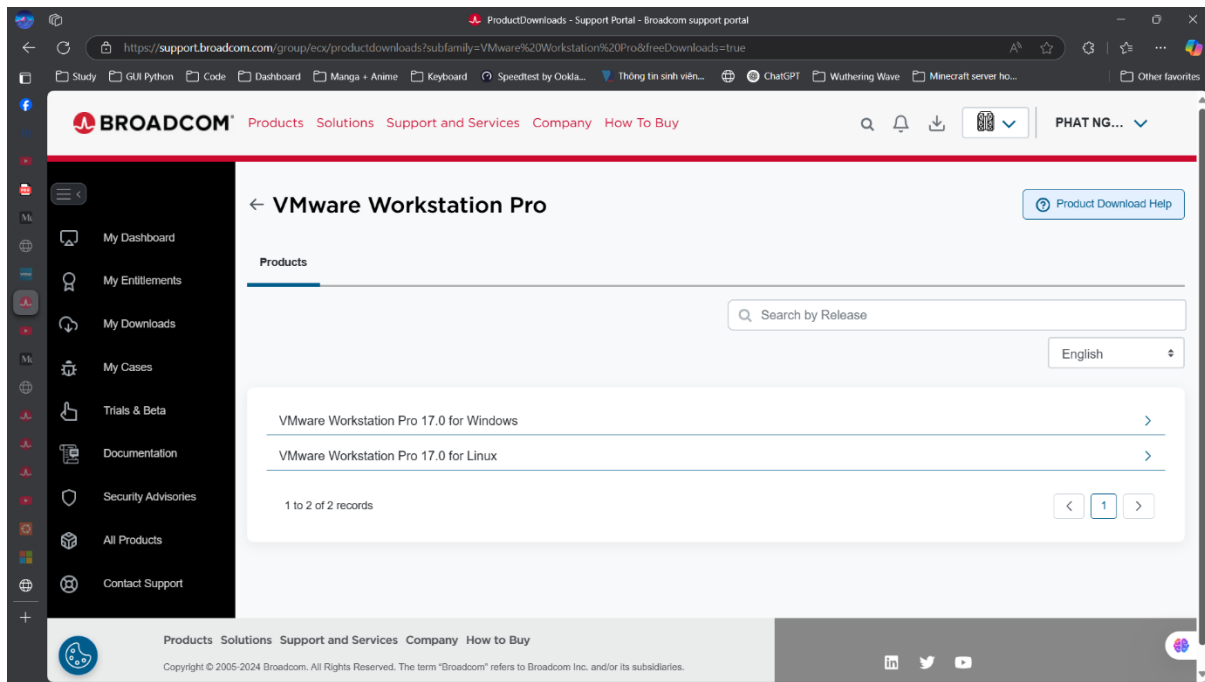
# I. Setup virtual machine (VMWare Workstation Pro)

## 1. Install VMWare Workstation Pro and Ubuntu.

### a. Preparation

*Step 1: Download VMWare Workstation Pro.*

- Go to VMware Workstation Pro website to download the latest version.



*Figure 1: VMware Workstation Pro official website.*

## Step 2: Download Ubuntu.

- Go to Ubuntu official website to download the suitable (x64) OS

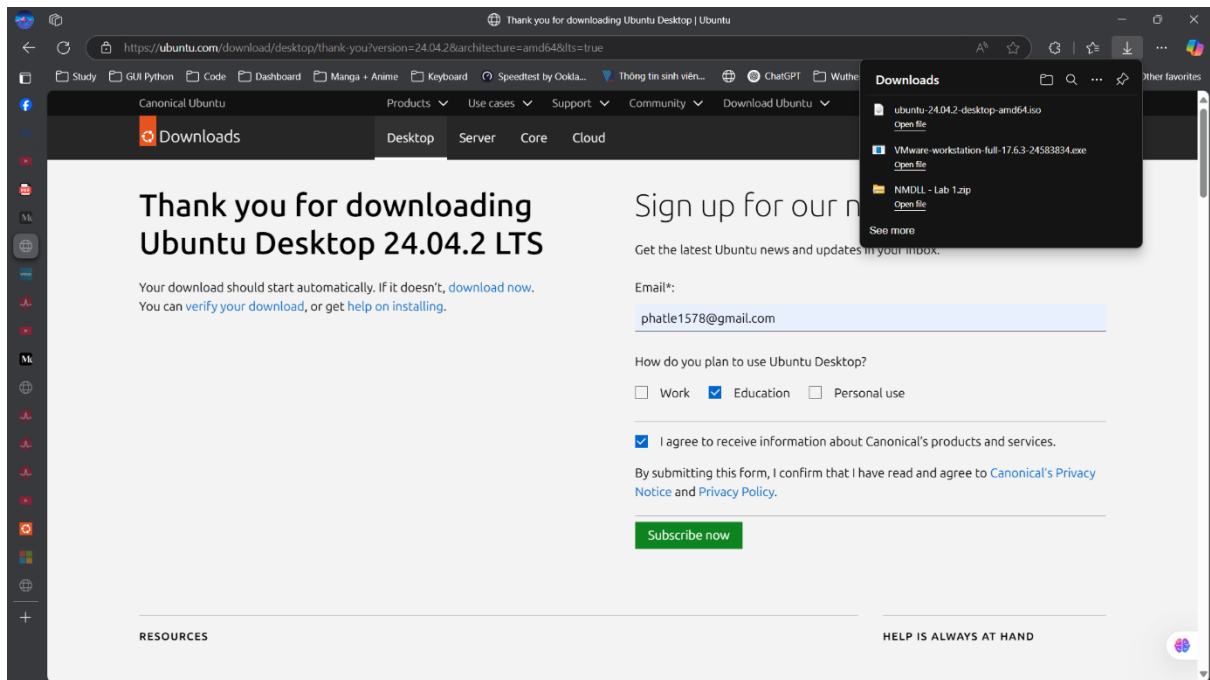


Figure 2: Ubuntu official download website.

## b. Setup the VMware Workstation Pro

### Step 1: Run VMware Workstation Pro.

Choose “Create a New Virtual Machine” option on the main menu.

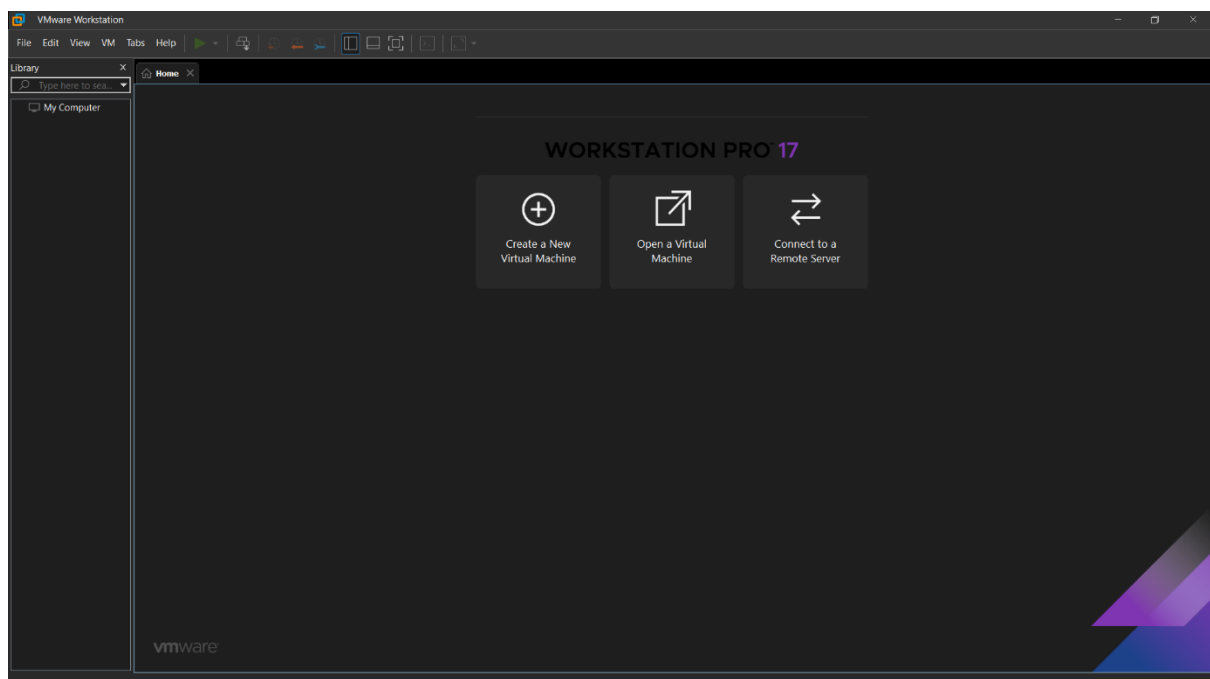


Figure 3: VMware Workstation Pro main menu.

## Step 2: Setup Virtual Machine.

- Choose “*Installer disc image file (iso)*” option and select the Ubuntu file we download previously.

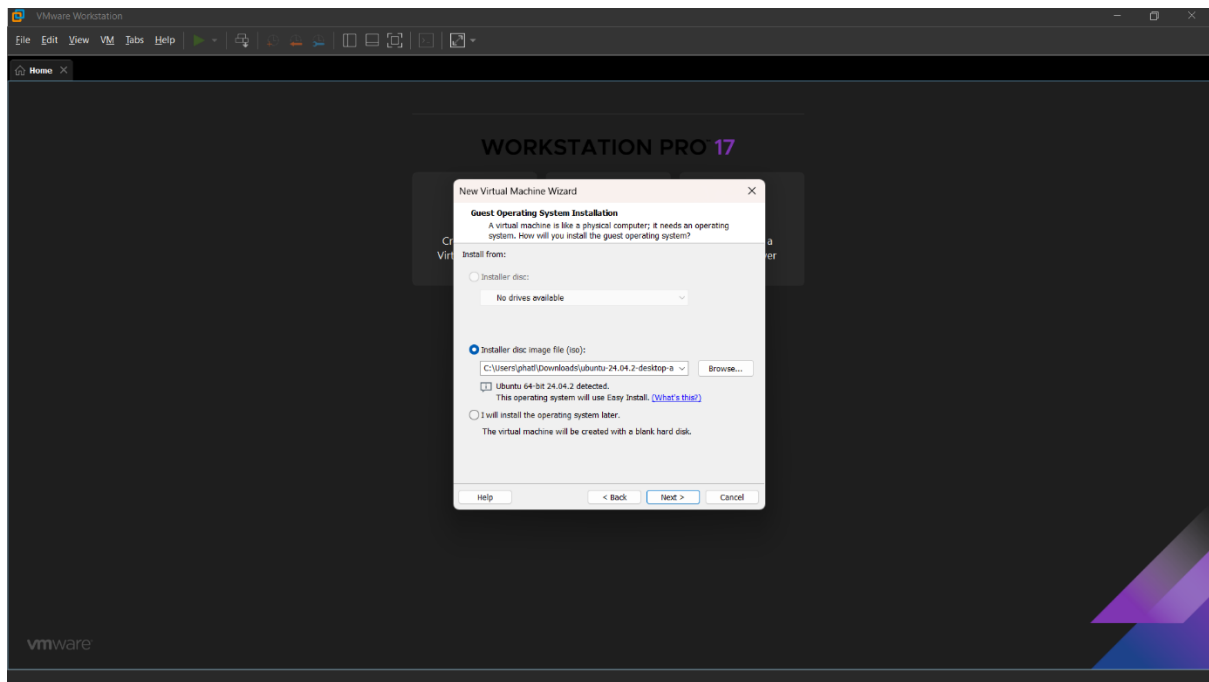


Figure 4: VMware Workstation Pro virtual machine installer.

## Step 3: Setup Linux profile.

- Fill in some information.

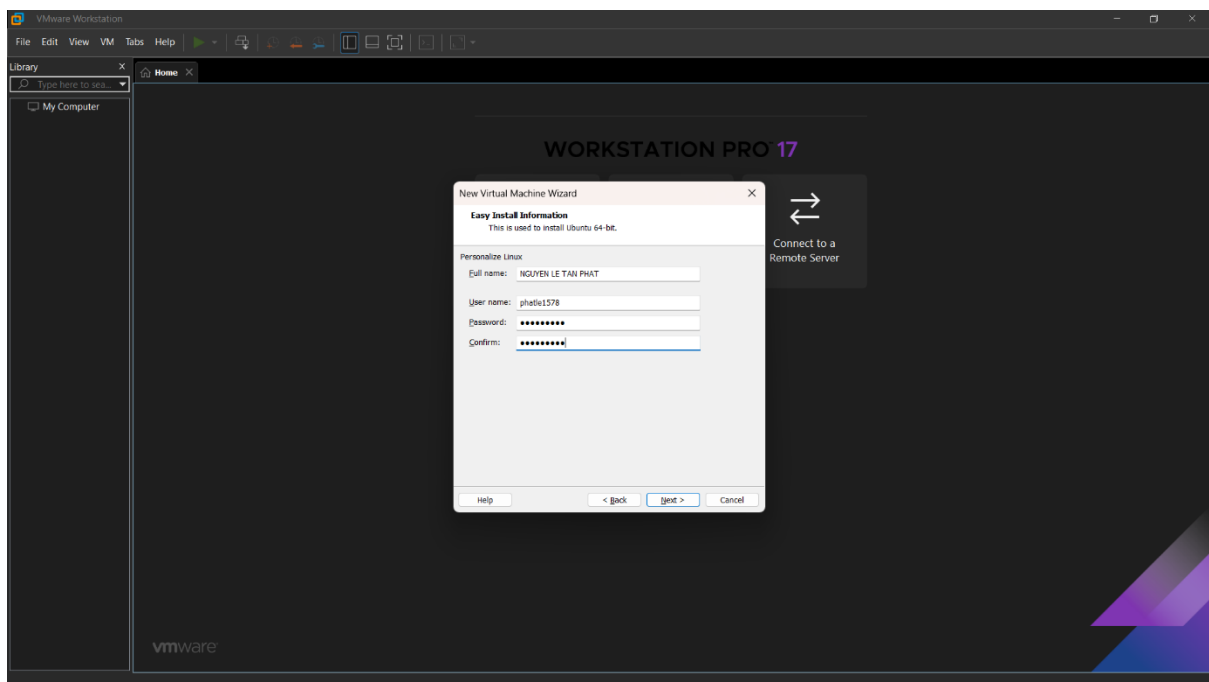


Figure 5: Personalize Linux.

#### Step 4: Setup virtual machine specs.

- Keep default and select “next”

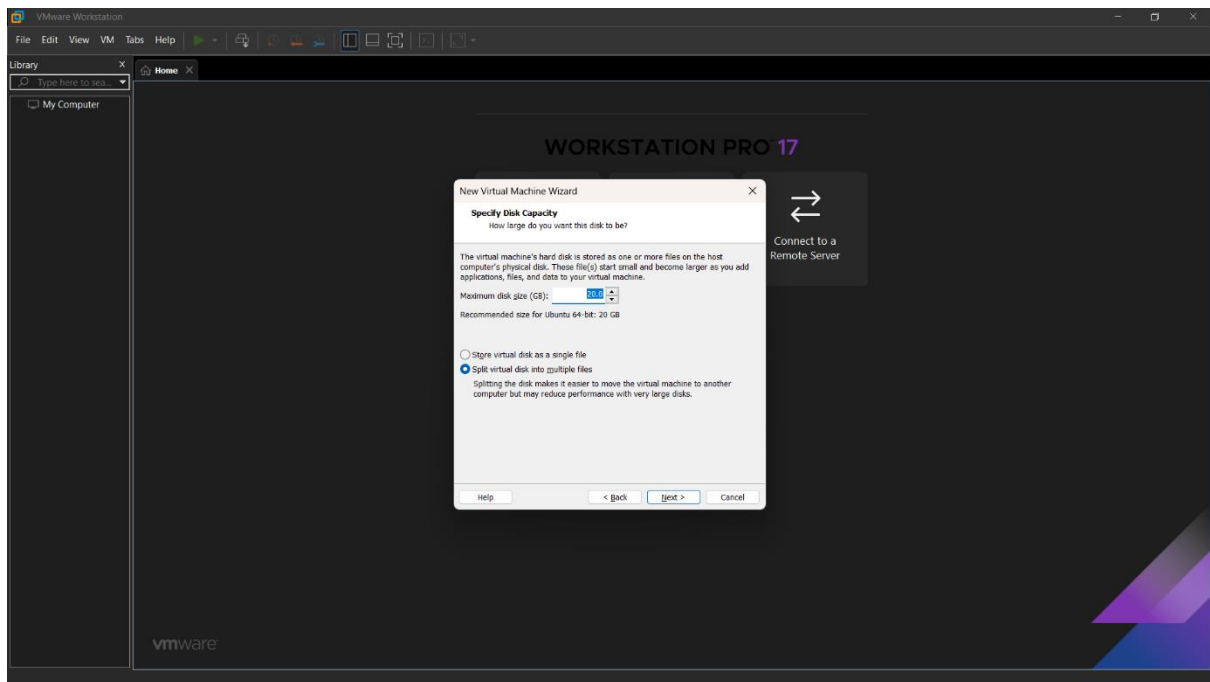


Figure 6: Virtual machine spec.

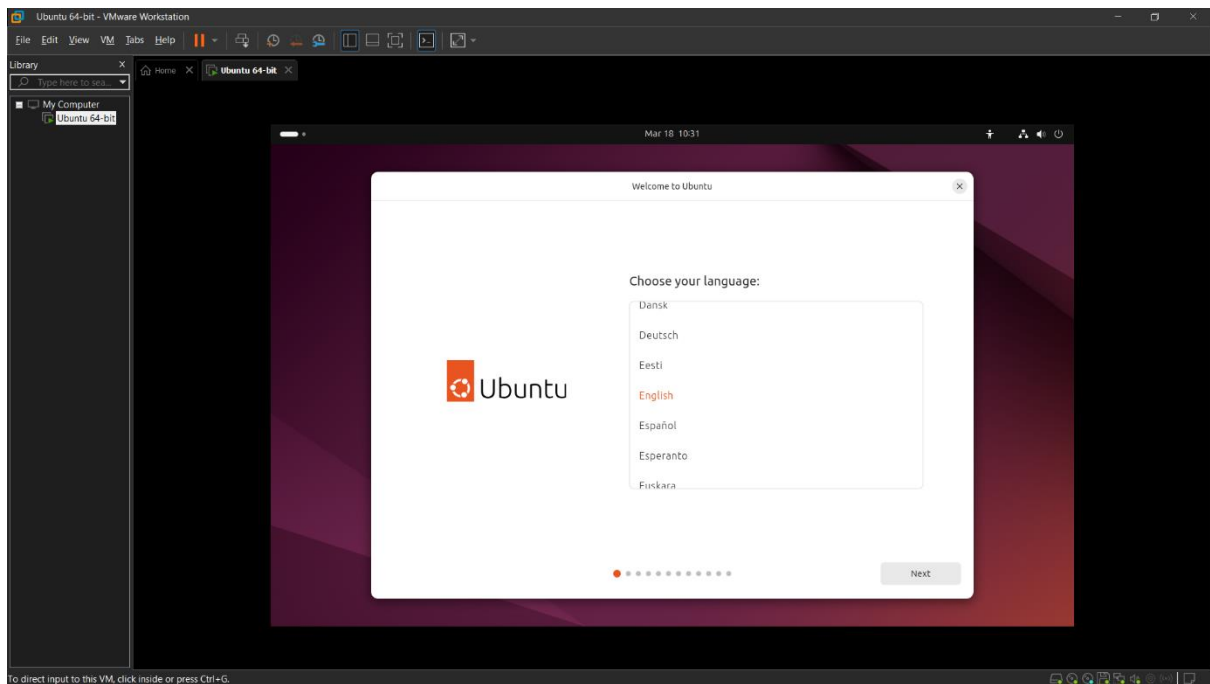
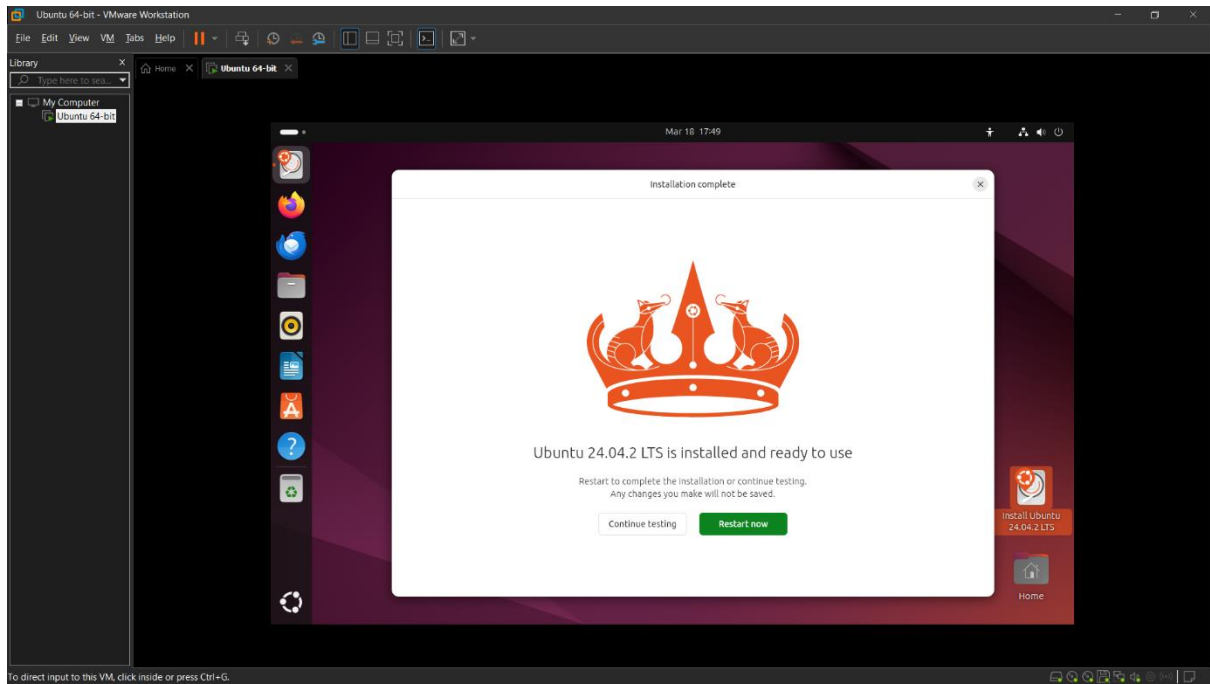


Figure 7: Ubuntu setup (begin)

c. Setup Linux system in the virtual machine.

*Step 1: Setup Ubuntu.*

- Choose default settings and just select “*next*”



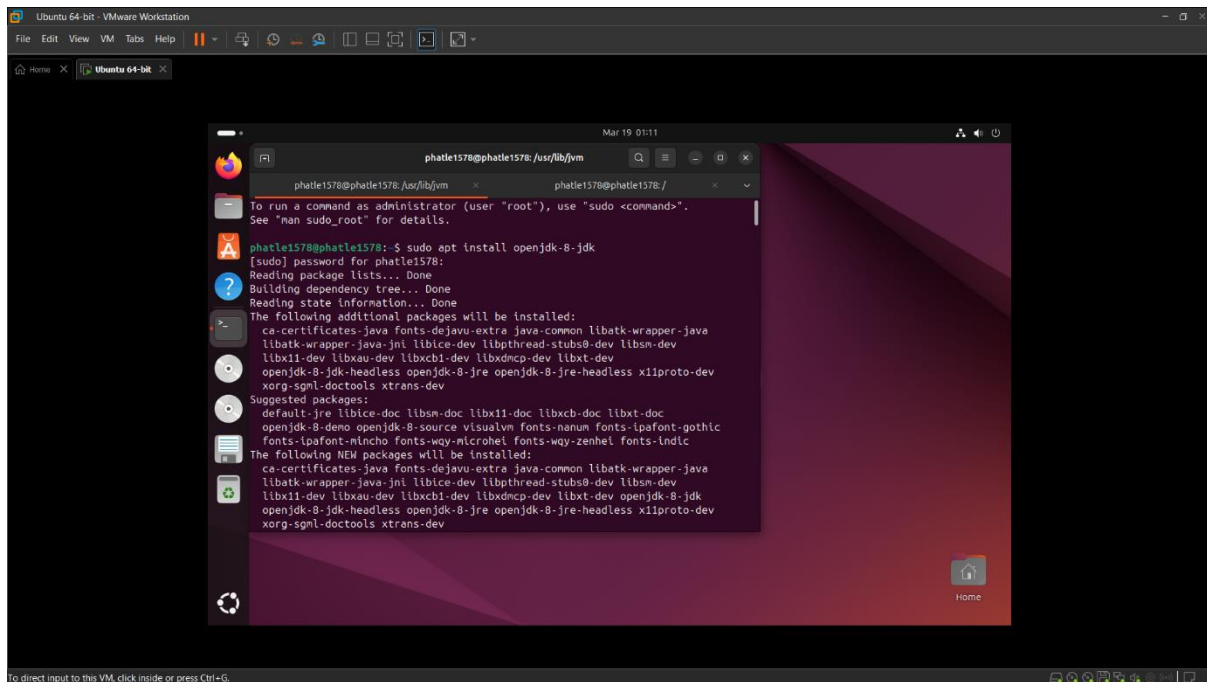
*Figure 8: Ubuntu setup (finish)*

## 2. Setup Hadoop Cluster (Pseudo – Distributed Mode).

### a. Install Hadoop (VMware Workstation Pro).

#### Step 1: System preparation.

- `sudo apt install openjdk-8-jdk`: Install OpenJDK 8, a requirement for Hadoop and the most stable version to run Hadoop



*Figure 9: Install OpenJDK 8*

- `sudo nano .bashrc`: Configure Hadoop environment variables.
- Go to the end of the file and add these line.

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/.hadoop-3.4.1 /
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh
```

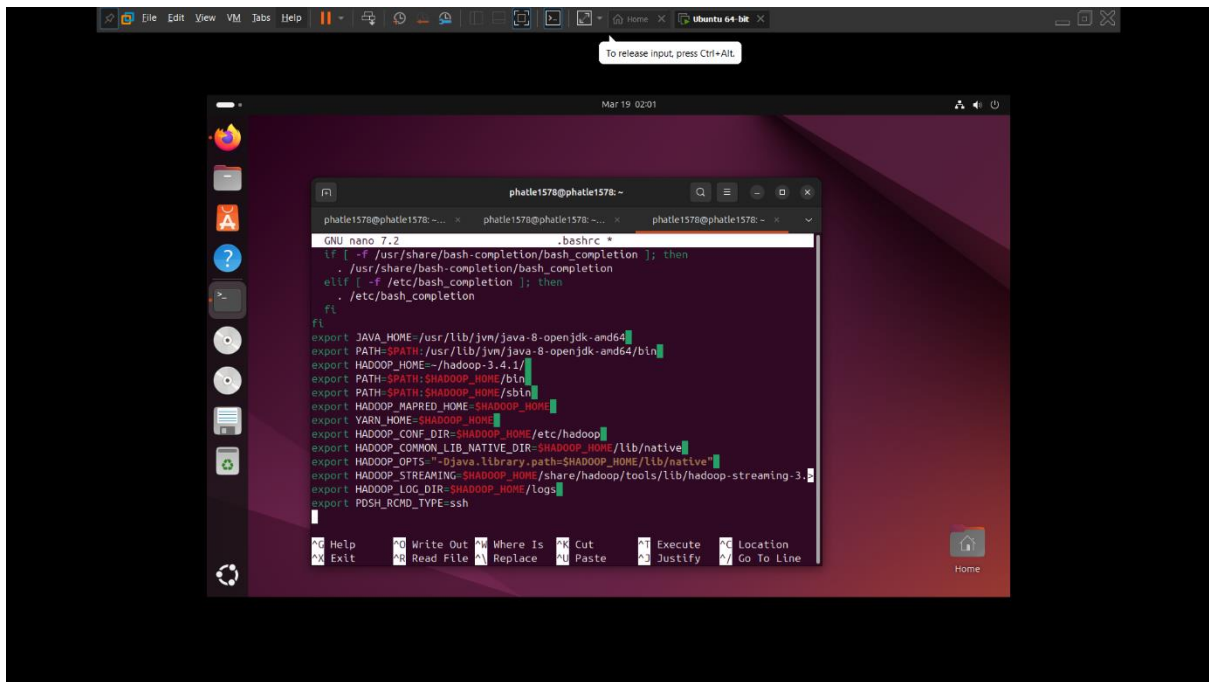


Figure 10: Configure bashrc file

- `sudo apt-get install ssh`: Install ssh.

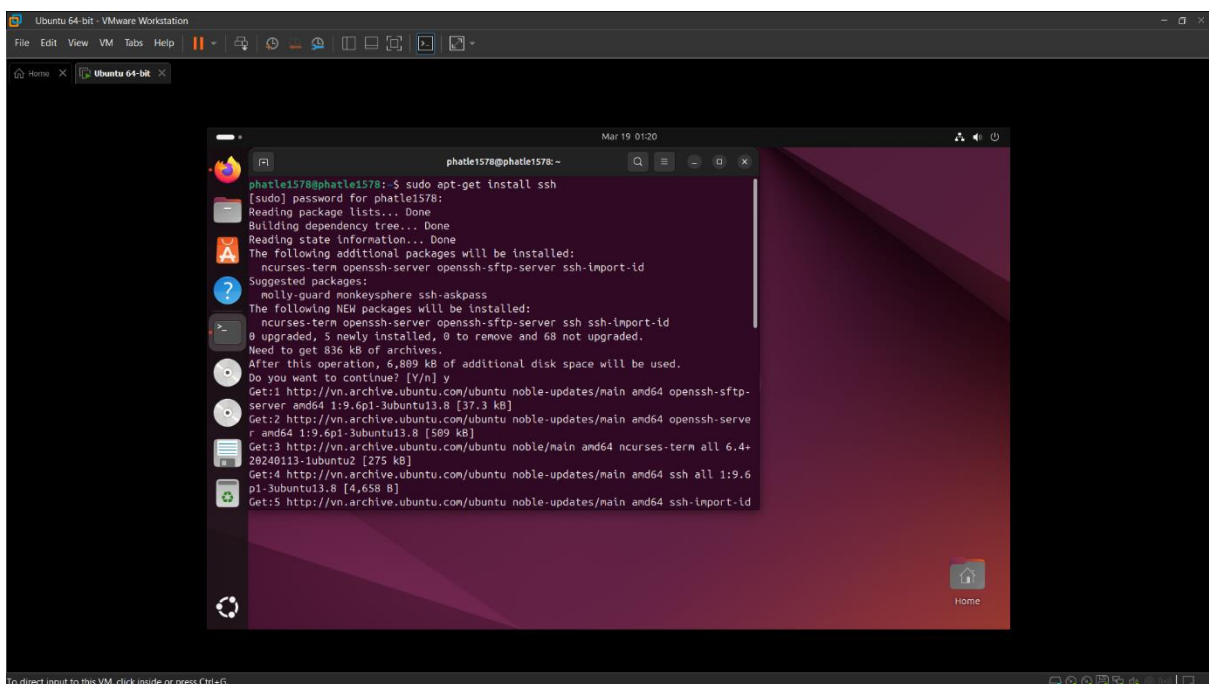


Figure 11: Install ssh

## Step 2: Download Apache Hadoop

- `wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz`: download Apache Hadoop latest version.

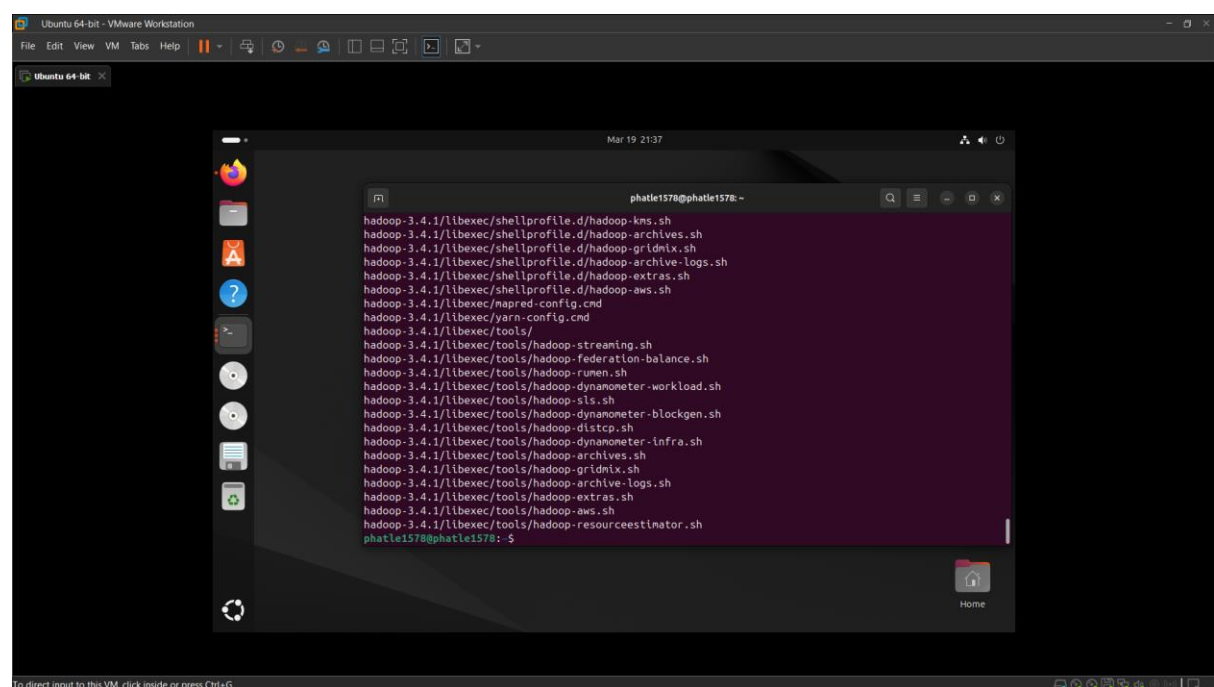
```
phatle1578@phatle1578:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
--2025-03-19 21:29:36-- https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 974002355 (929M) [application/x-gzip]
Saving to: 'hadoop-3.4.1.tar.gz'

hadoop-3.4.1.tar.gz  100%[=====] 928.88M  5.28MB/s   in 4m 37s

2025-03-19 21:34:15 (3.36 MB/s) - 'hadoop-3.4.1.tar.gz' saved [974002355/974002355]
```

*Figure 12: Download Apache Hadoop latest version*

- `tar -zxvf ~/Downloads/hadoop-3.4.1.tar.gz`: Extract the binary Hadoop file we just download.



*Figure 13: Extract the binary Hadoop file.*



### Step 3: Configure java environment variables and other xml files.

#### i. hadoop-env.sh

- `cd /etc/hadoop/`: go to /etc/hadoop directory
- `sudo nano hadoop-env.s`: open hadoop-env.sh to configure
- Search for the JAVA\_HOME and add this line to set the Java home path:

```
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd4
```

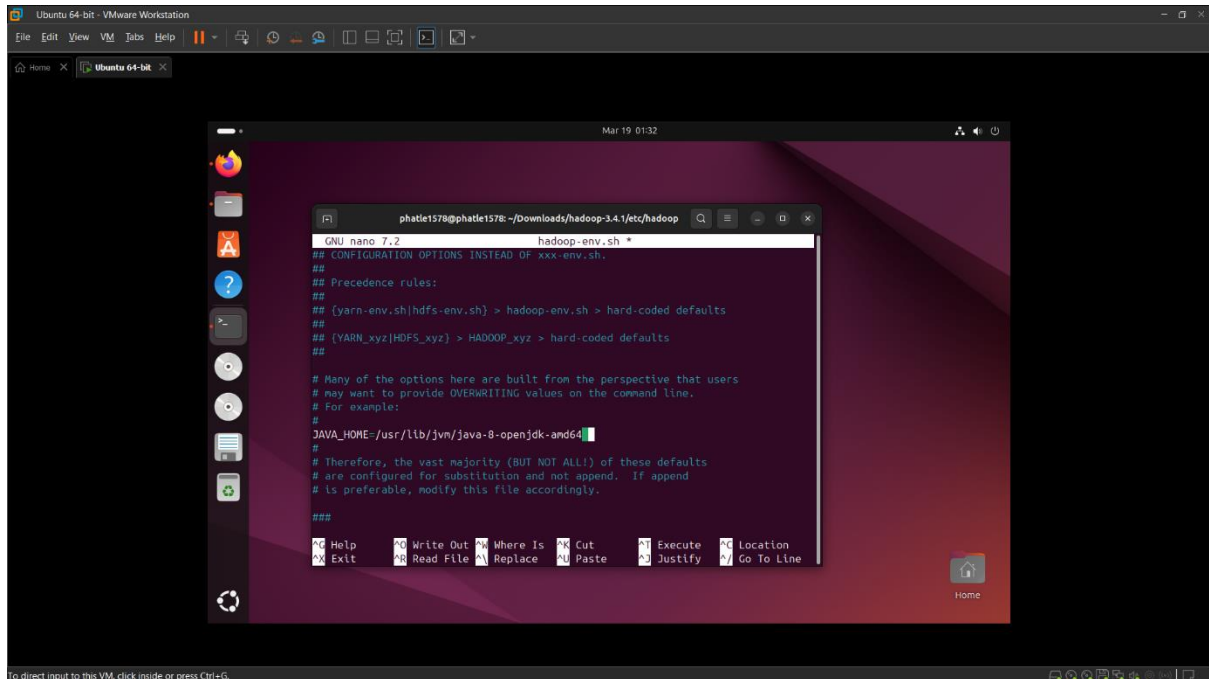


Figure 14: Configure hadoop-env.sh

- `Ctrl + O`: Save
- `Ctrl + X`: Exit

#### ii. core-site.xml

- `sudo nano core-site.xml`: Open and configure core-site.xml for pseudo-distributed mode
- Navigate to the end of the file and add these lines:

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

- `Ctrl + O`: Save.
- `Ctrl + X`: Exit

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

Help Write Out Where Is Cut Execute Location  
Exit Read File Replace Paste Justify Go To Line

Figure 15: Configure core-site.xml

### iii. hdfs-site.xml

- `sudo nano hdfs-site.xml`: Open and configure hdfs-site.xml.
- Navigate to the end of the file hdfs-site.xml and add these lines:

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

- `Ctrl + O`: Save.
- `Ctrl + X`: Exit.

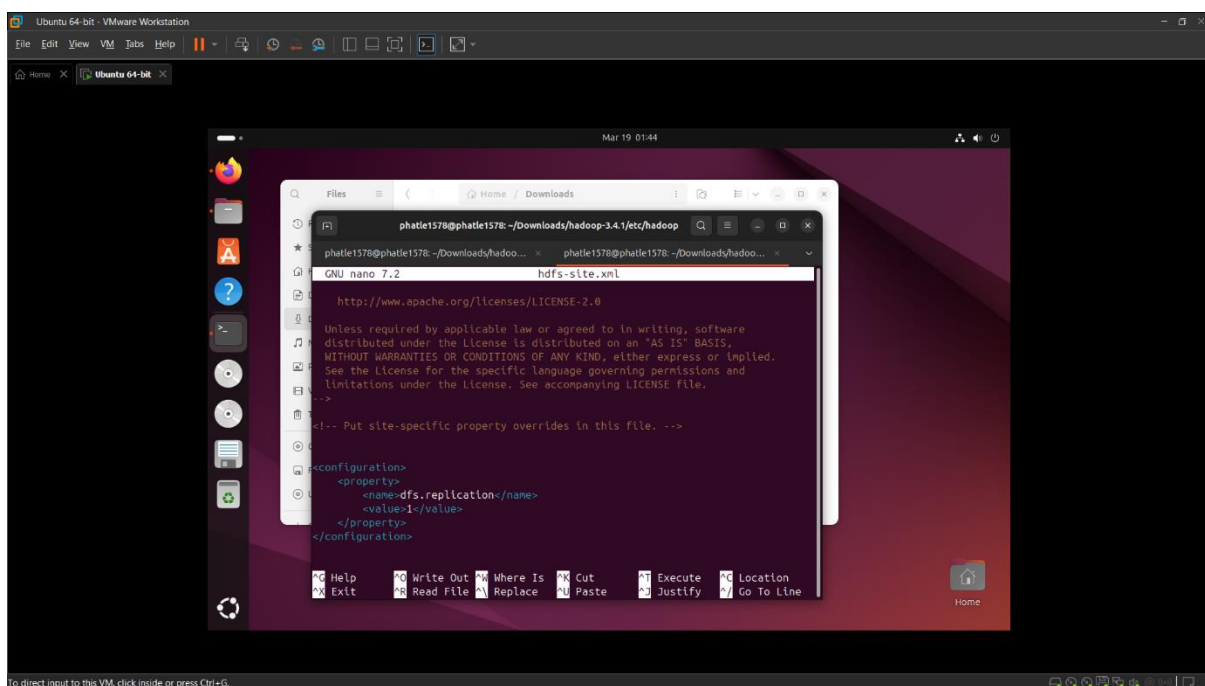


Figure 16: Configure hdfs-site.xml

#### iv. mapred-site.xml

- `sudo nano mapred-site.xml`: Open and configure mapred-site.xml.
- Navigate to the end of the file and add these lines:

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

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_H
OME/share/hadoop/mapreduce/lib/*</value>
</property>
</configuration>
```

- `Ctrl + O`: Save.
- `Ctrl + X`: Exit.

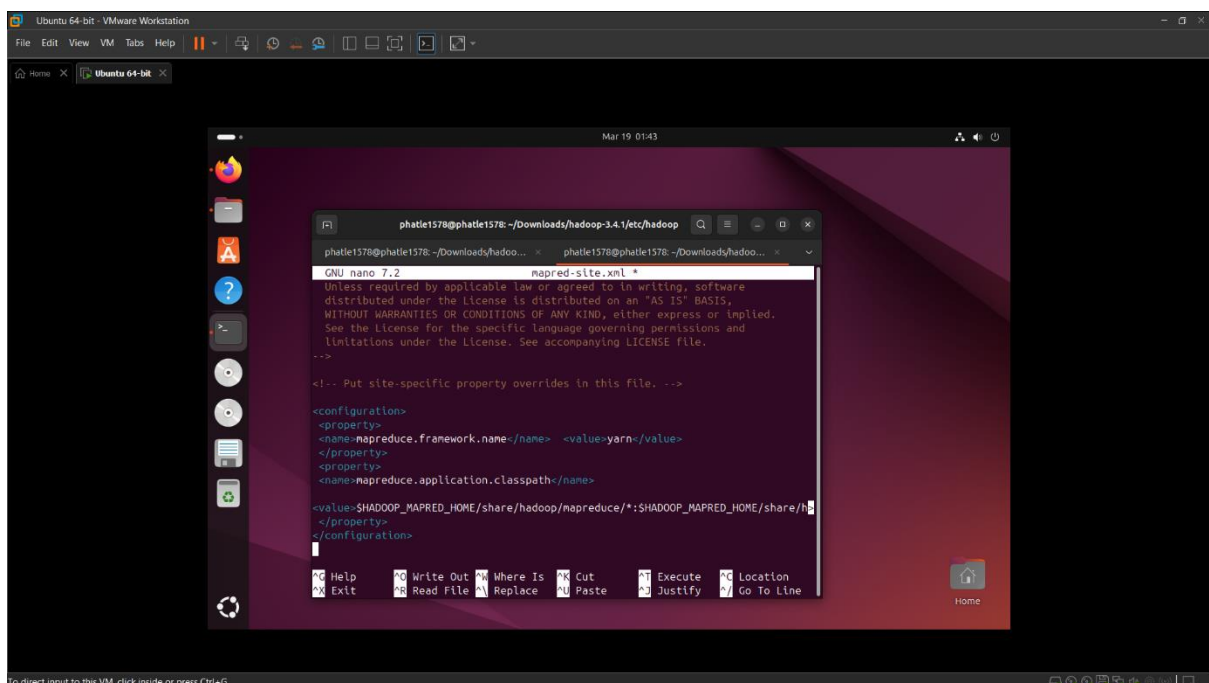


Figure 17: Configure mapred-site.xml

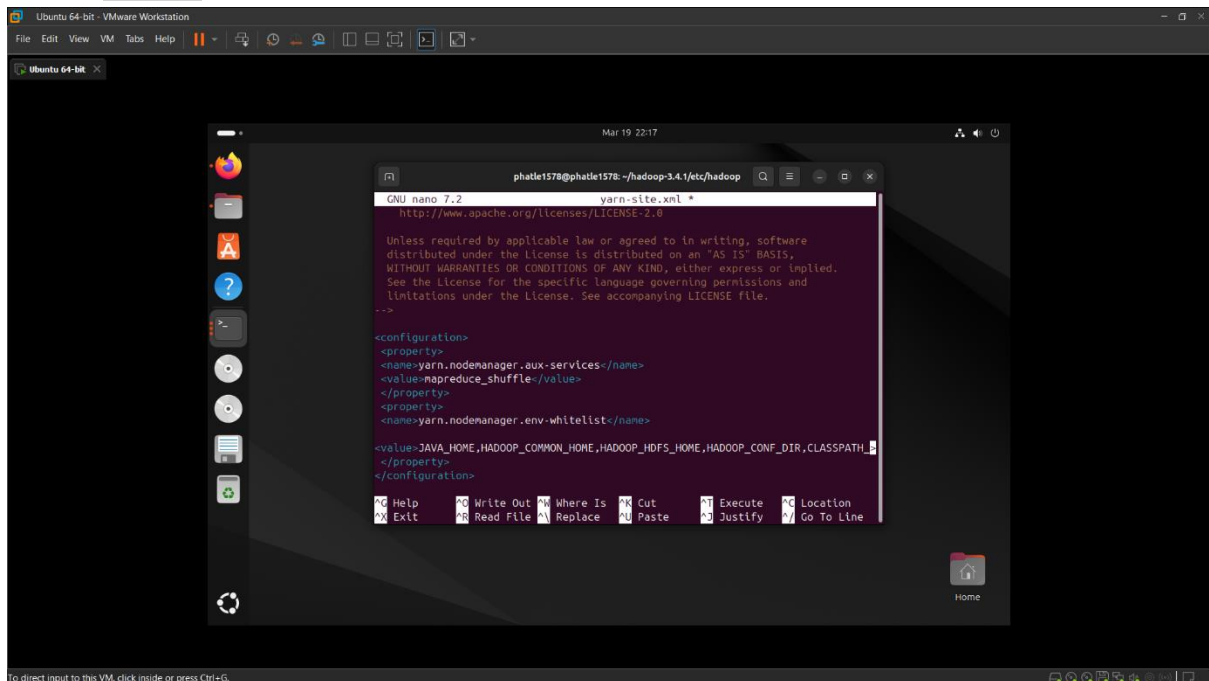
#### v. yarn-site.xml

- `sudo nano yarn-site.xml`: Open and configure yarn-site.xml.
- Navigate to the end of the file and add these lines:

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

```
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREP
END_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

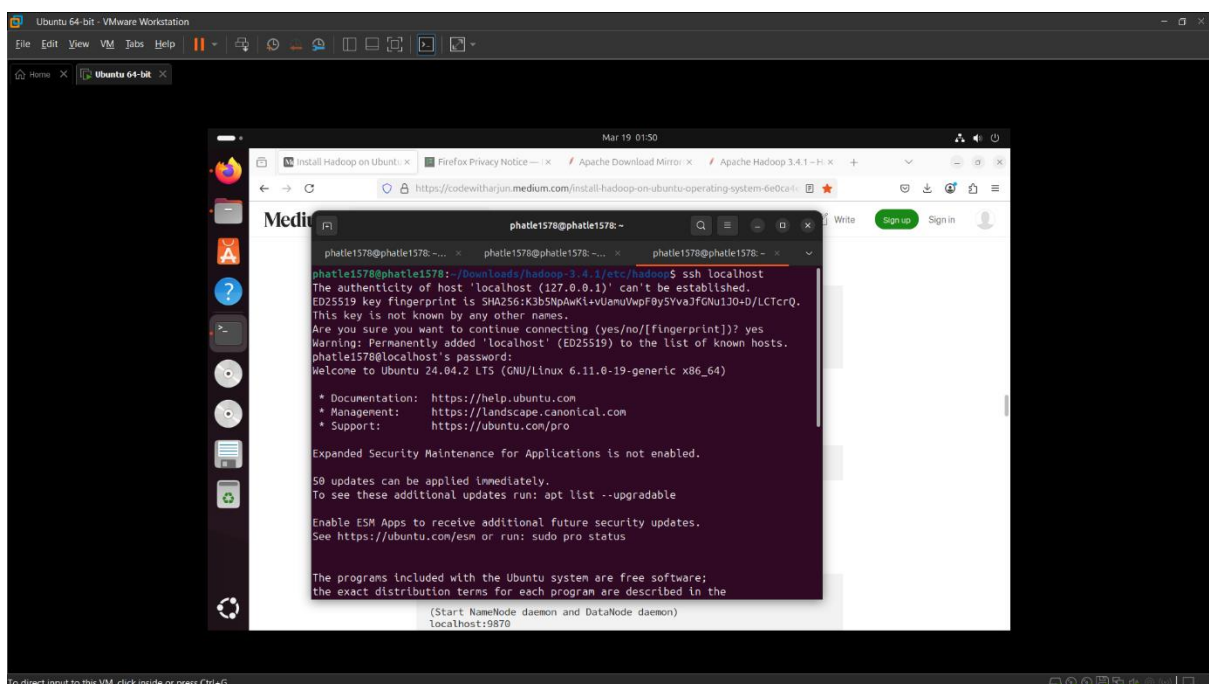
- **Ctrl + O**: Save.
- **Ctrl + X**: Exit.



*Figure 18: Configure yarn-site.xml*

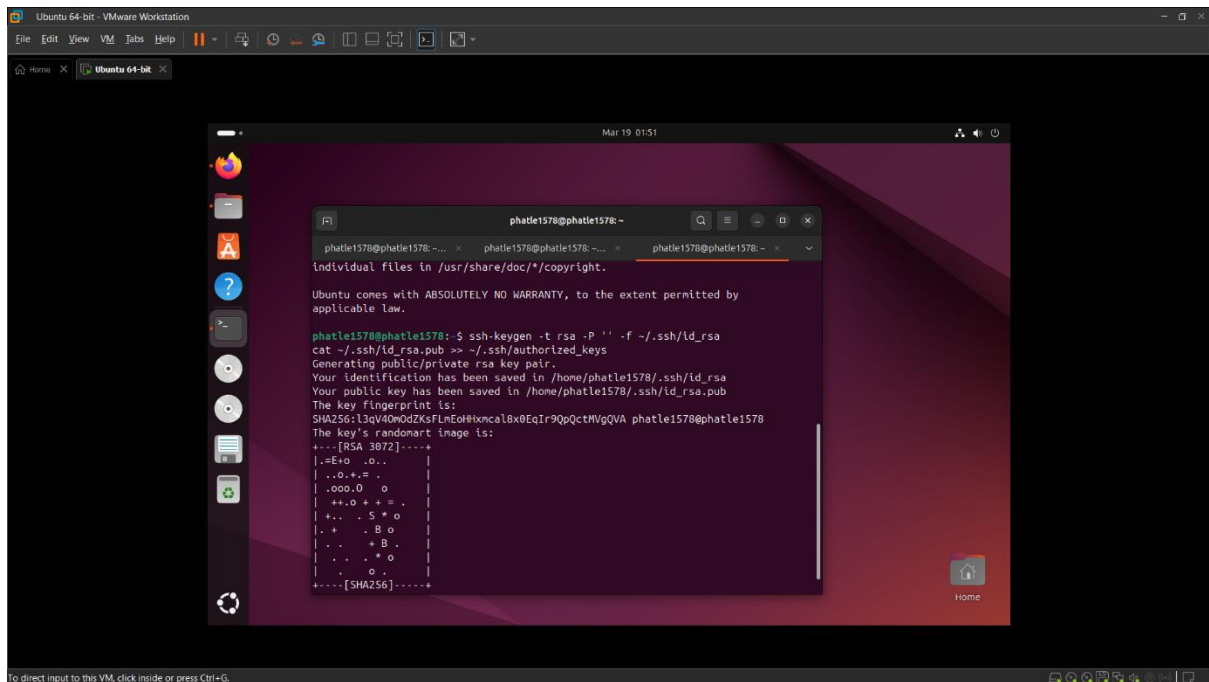
#### Step 4: Setup ssh

- **ssh localhost**: Start the ssh localhost.



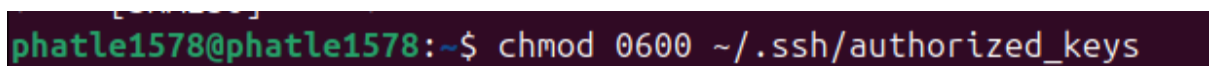
*Figure 19: Start ssh localhost*

- `ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa`: Generate private key at `~/.ssh/id_rsa` and public key at `~/.ssh/id_rsa.pub`
- `cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`: add public key we just create previously to the `authorized_keys` list.



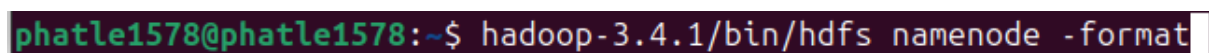
*Figure 20: Generate ssh key.*

- `chmod 0600 ~/.ssh/authorized_keys`: Set the access permission to the key. Only the owner can read and write.



*Figure 21: Set key permission.*

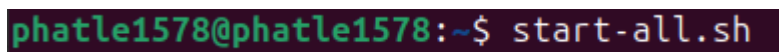
- `export PDSH_RCMD_TYPE=sshFormat`: delete all the current metadata of HDFS and restart the system.



*Figure 22: Restart the distributed system.*

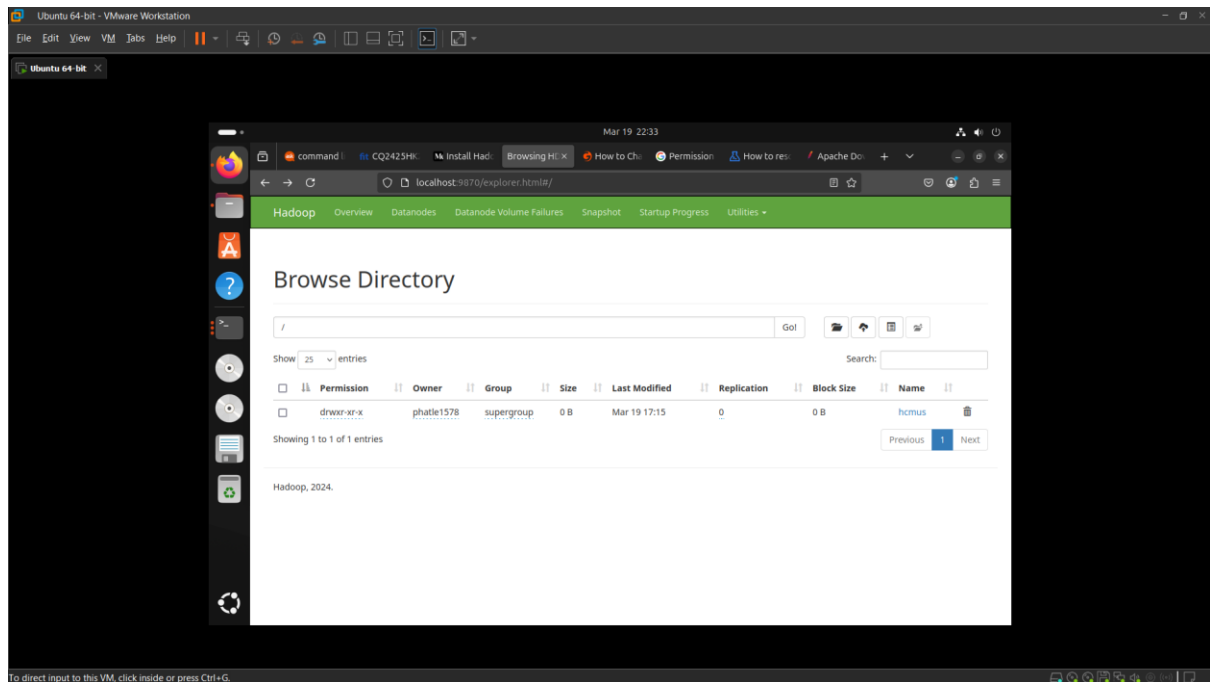
## b. Start Hadoop Cluster.

- `start-all.sh`: Start all hadoop services.



*Figure 23: Start all hadoop services*

- `hadoop fs -mkdir /hcmus`: Create hcmus folder in hadoop



*Figure 24: Create hcmus folder in hadoop.*

- `sudo adduser khtn_22120262`: Create new user with name khtn\_22120262

```
phatle1578@phatle1578:~$ sudo adduser khtn_22120262
[sudo] password for phatle1578:
info: Adding user `khtn_22120262' ...
info: Selecting UID/GID from range 1000 to 59999 ...
info: Adding new group `khtn_22120262' (1001) ...
info: Adding new user `khtn_22120262' (1001) with group `khtn_22120262 (1001)' .
..
info: Creating home directory `/home/khtn_22120262' ...
info: Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for khtn_22120262
Enter the new value, or press ENTER for the default
  Full Name []:
  Room Number []:
  Work Phone []:
  Home Phone []:
  Other []:
Is the information correct? [Y/n] y
info: Adding new user `khtn_22120262' to supplemental / extra groups `users' ...
info: Adding user `khtn_22120262' to group `users' ...
```

*Figure 24: Create new user with name khtn\_22120262*

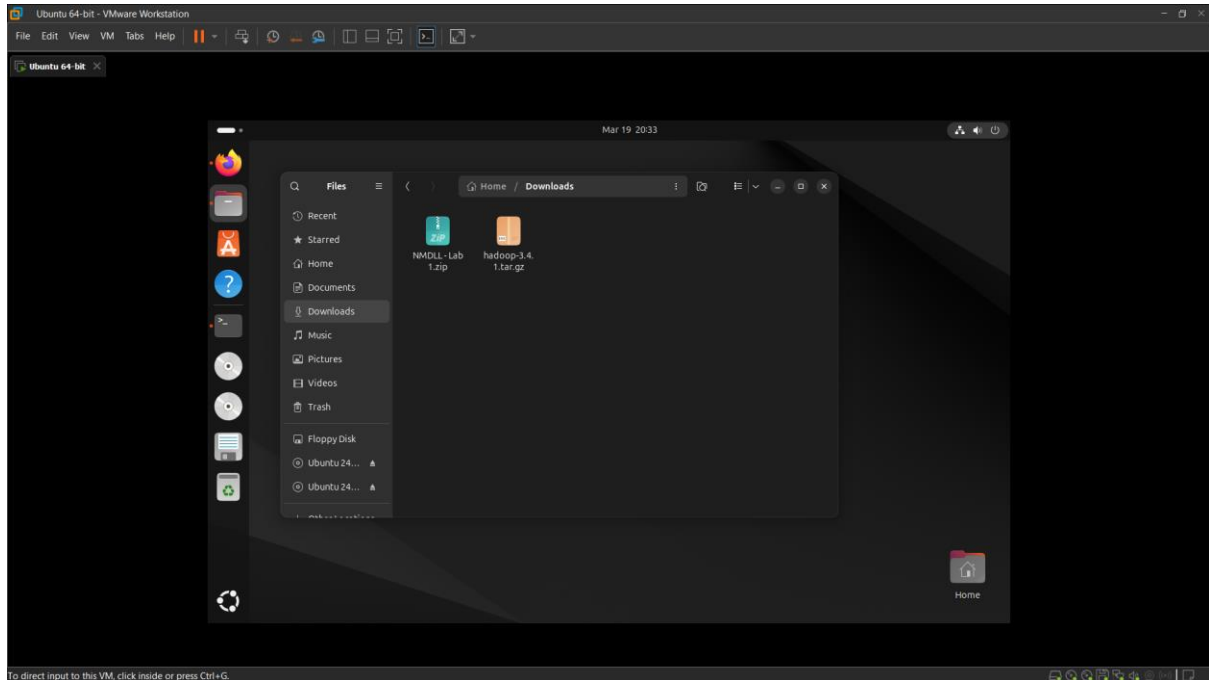
- `hadoop fs -mkdir /hcmus/22120262`: Create a subfolder 22120262 in hcmus folder.

```
phatle1578@phatle1578:~$ hadoop fs -mkdir /hcmus/22120262
```



*Figure 25: Create subfolder 22120262 in hemus folder.*

- Download the require file from [courses.fit.hcmus.edu.vn](http://courses.fit.hcmus.edu.vn)
- Right click in the folder Downloads to open in terminal



*Figure 26: Downloads require file.*

- `unzip NMDLL\ -\ Lab\ 1.zip - ~\`: Unzip the file we just download into the /home/phatle1578 directory..

```
phatle1578@phatle1578:~/Downloads$ unzip NMDLL\ -\ Lab\ 1.zip -d ~/
Archive:  NMDLL - Lab 1.zip
  creating: /home/phatle1578/NMDLL - Lab 1/
  inflating: /home/phatle1578/__MACOSX/._NMDLL - Lab 1
  inflating: /home/phatle1578/NMDLL - Lab 1/.DS_Store
  inflating: /home/phatle1578/__MACOSX/NMDLL - Lab 1/._.DS_Store
  inflating: /home/phatle1578/NMDLL - Lab 1/hadoop-test.jar
  inflating: /home/phatle1578/__MACOSX/NMDLL - Lab 1/._hadoop-test.jar
  inflating: /home/phatle1578/NMDLL - Lab 1/Lab 1 - HDFS.pdf
  inflating: /home/phatle1578/__MACOSX/NMDLL - Lab 1/._Lab 1 - HDFS.pdf
  inflating: /home/phatle1578/NMDLL - Lab 1/words.txt
  inflating: /home/phatle1578/__MACOSX/NMDLL - Lab 1/._words.txt
```

*Figure 27: Unzip the file.*

- `cp ~/NMDLL\ -\ Lab\ 1/hadoop-test.jar ~/`: Move the hadoop-test.jar file out to /home/phatle1578 directory for more convenient.

```
phatle1578@phatle1578:~$ cp ~/NMDLL\ -\ Lab\ 1/hadoop-test.jar ~/
```

*Figure 28: Move the hadoop-test.jar file to home directory.*

- `ls -l ~/`: Verify it has been copy to the home directory

```
phatle1578@phatle1578:~$ ls -l ~/
total 62384
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Desktop
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Documents
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 10:47 Downloads
drwxr-xr-x 11 phatle1578 phatle1578    4096 Mar 19 02:10 hadoop-3.4.1
-rw-rw-r--  1 phatle1578 phatle1578 63828099 Mar 19 17:16 hadoop-test.jar
drwxrwxr-x  3 phatle1578 phatle1578    4096 Mar 19 10:49 __MACOSX
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Music
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 10:52 NMDLL-Lab1
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Pictures
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Public
drwx----- 4 phatle1578 phatle1578    4096 Mar 19 01:15 snap
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Templates
drwxr-xr-x  2 phatle1578 phatle1578    4096 Mar 19 01:00 Videos
```

*Figure 29: Verify the file is exist in home directory.*

- `hdfs dfs -put ~/hadoop-test.jar /hcmus/22120262`: Move the hadoop-test.jar into the path /hcmus/22120262 in hdfs.

```
phatle1578@phatle1578:~$ hdfs dfs -put ~/hadoop-test.jar /hcmus/22120262
```

*Figure 30: Move the file hadoop-test.jar into /hcmus/22120262 in hdfs.*

- `hdfs dfs -ls /hcmus/22120262`: Verify it has been move to /hcmus/22120262 in hdfs

```
phatle1578@phatle1578:~$ hdfs dfs -ls /hcmus/22120262
Found 1 items
-rw-r--r--  1 phatle1578 supergroup    63828099 2025-03-19 17:17 /hcmus/22120262/hadoop-test.jar
```

*Figure 31: Verify the file is exist in /hcmus/22120262 directory.*

- `hdfs dfs -chown khtn_22120262 /hcmus/22120262`: Set ownership of /hcmus/22120262/ to khtn\_22120262 user.

```
phatle1578@phatle1578:~$ hdfs dfs -chown khtn_22120262 /hcmus/22120262
```

*Figure 32: Set ownership.*

- `hdfs dfs -chmod 744 /hcmus/22120262/hadoop-test.jar`: Set the file permissions to 744

```
phatle1578@phatle1578:~$ hdfs dfs -chmod 744 /hcmus/22120262/hadoop-test.jar
```

*Figure 33: Set the file permissions to 744.*



- `java -jar ~/hadoop-test.jar 9000 /hcmus/22120262`: Execute the hadoop-test.jar file

```
phatle1578@phatle1578:~$ java -jar ~/hadoop-test.jar 9000 /hcmus/22120262
Trying to read /hcmus/22120262
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Found hdfs://localhost:9000/hcmus/22120262/hadoop-test.jar
Your student ID: 22120262 (ensure it matches your student ID)
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
The first method to get MAC address is failed: Could not get network interface
Trying the alternative method
File written at /home/phatle1578/22120262_verification.txt
```

*Figure 35: Execute the file.*

```
MAC=00-0C-29-BB-66-1B
179f4cfb620bf7dcfa33f2f310c104889d3f04da846b5cfb97c98c5e95698100
```

*Figure 36: Verification file*

## II. Word Count

- Result.txt

a	30850
f	15870
j	4158
g	14466
h	18662
c	38934
m	25194
u	23790
s	50571

*Figure 37: Output*