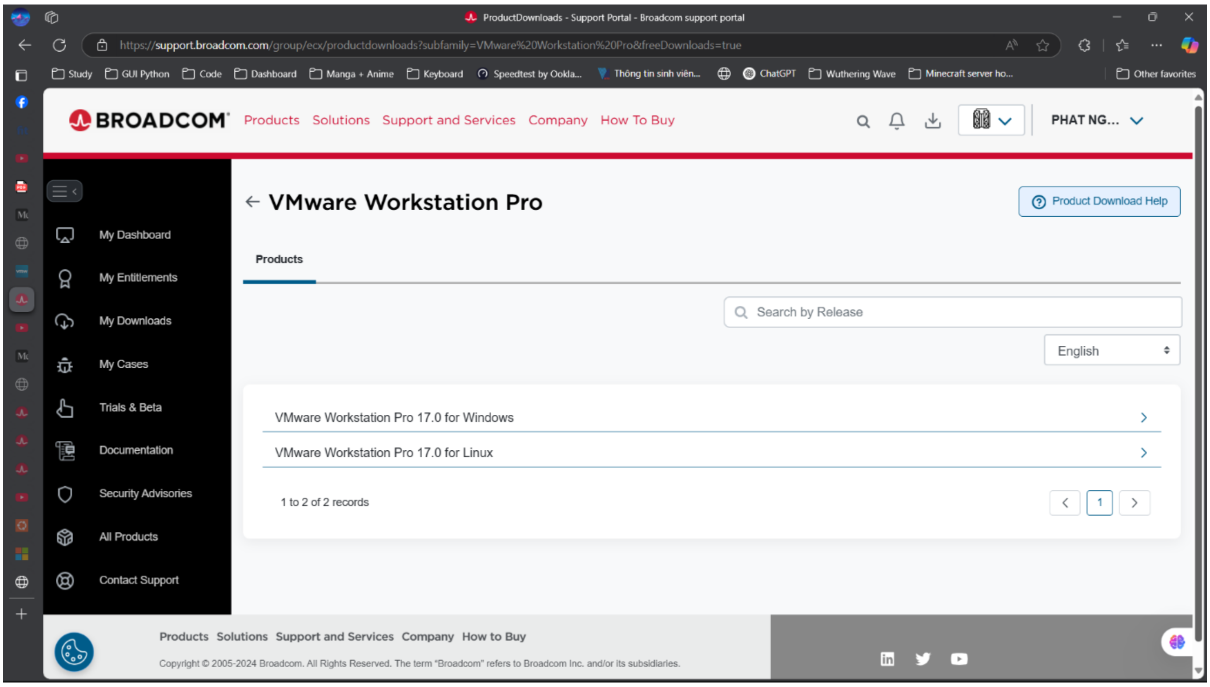
I. Setup virtual machine (VMWare Workstation Pro)

1. Install VMWare Workstation Pro and Ubuntu.

a. Preparation

Step 1: Download VMWare Workstation Pro.

- Go to Vmware Workstation Pro website to download the lastest version.



VMware Workstation Pro 17.0 for Windows

VMware Workstation Pro 17.0 for Linux

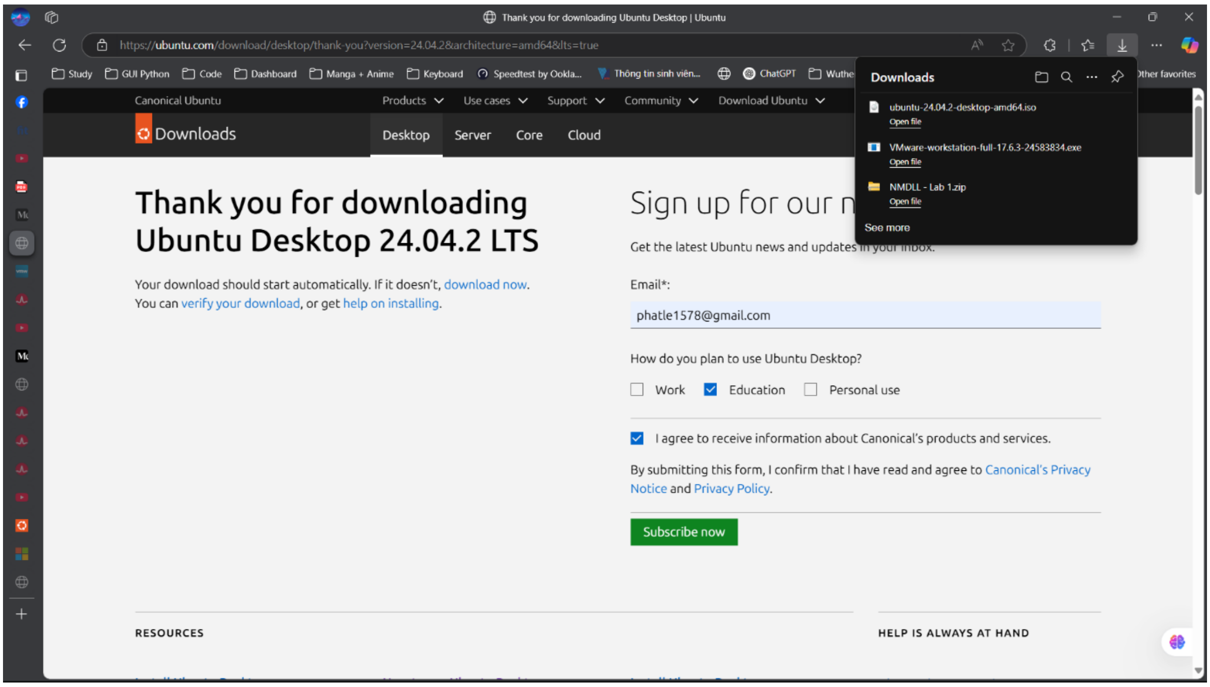
1 to 2 of 2 records

Copyright © 2005-2024 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

Figure 1: VMware Workstation Pro official website.

Step 2: Download Ubuntu.

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



Downloads

ubuntu-24.04.2-desktop-amd64.iso
Open file

VMware-workstation-full-17.6.3-24583834.exe
Open file

NMDLL - Lab 1.zip
Open file

See more

Thank you for downloading Ubuntu Desktop 24.04.2 LTS

Your download should start automatically. If it doesn't, download now.
You can verify your download, or get help on installing.

Sign up for our n

Get the latest Ubuntu news and updates in your inbox.

Email*:
phatle1578@gmail.com

How do you plan to use Ubuntu Desktop?

□ Work ☑ Education □ Personal use

☑ I agree to receive information about Canonical's products and services.

By submitting this form, I confirm that I have read and agree to Canonical's Privacy Notice and Privacy Policy.

Figure 2: Ubuntu official download website.

b. Setup the VMware Workstation Pro

Step 1: Run VMware Workstation Pro.

Choose "*Create a New Virtual Mahchine"* option on the main menu.

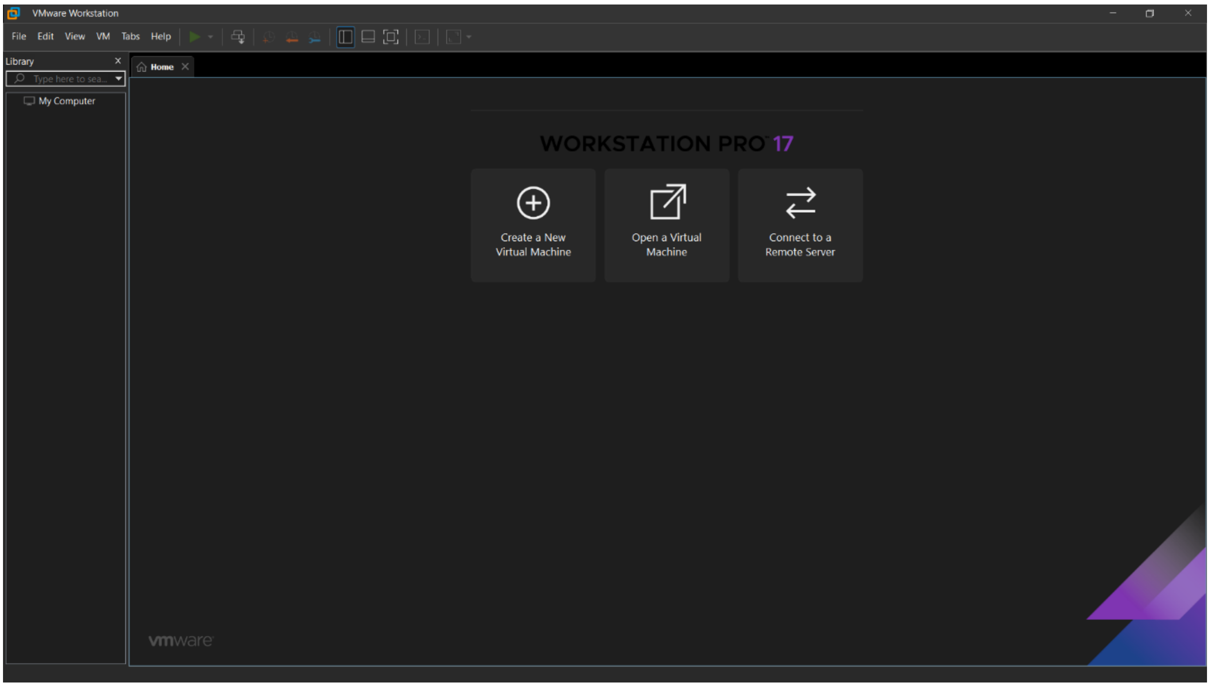
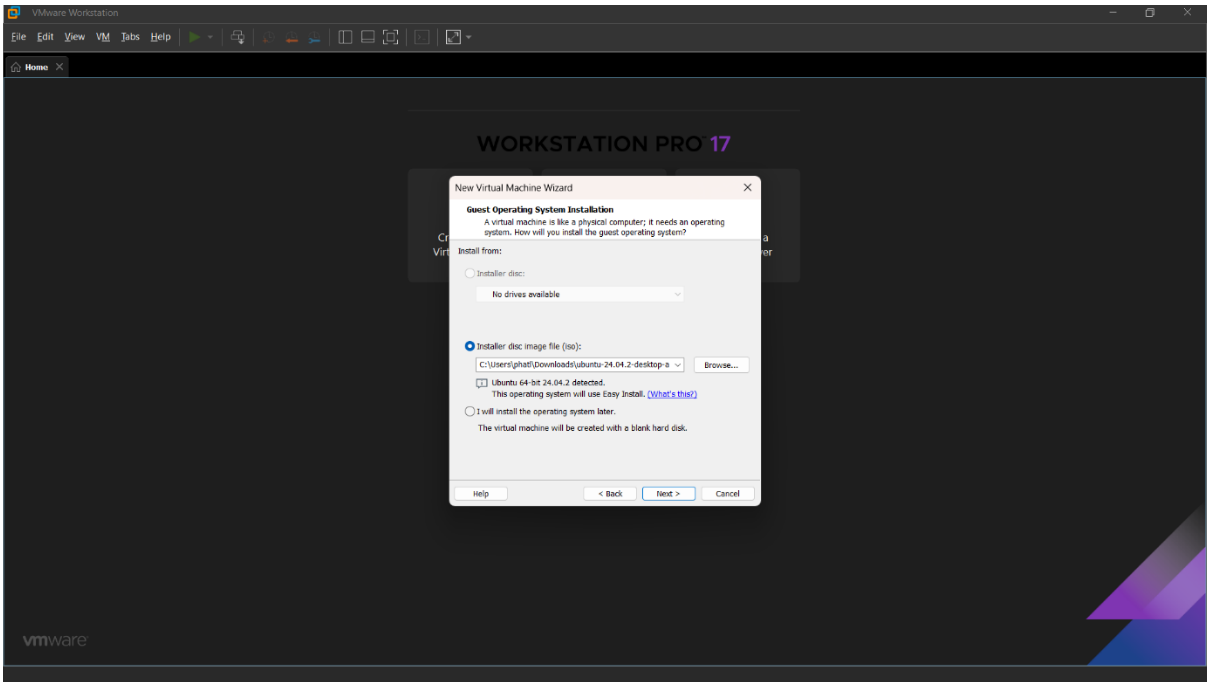


Figure 3: VMware Workstation Pro main menu.

Step 2: Setup Virtual Machine.

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



New Virtual Machine Wizard

Guest Operating System Installation

A virtual machine is like a physical computer; it needs an operating system. How will you install the guest operating system?

Install from:

○ Installer disc:

No drives available

● Installer disc image file (iso):

C:\Users\phatl\Downloads\ubuntu-24.04.2-desktop-a

Ubuntu 64-bit 24.04.2 detected.
This operating system will use Easy Install. (What's this?)

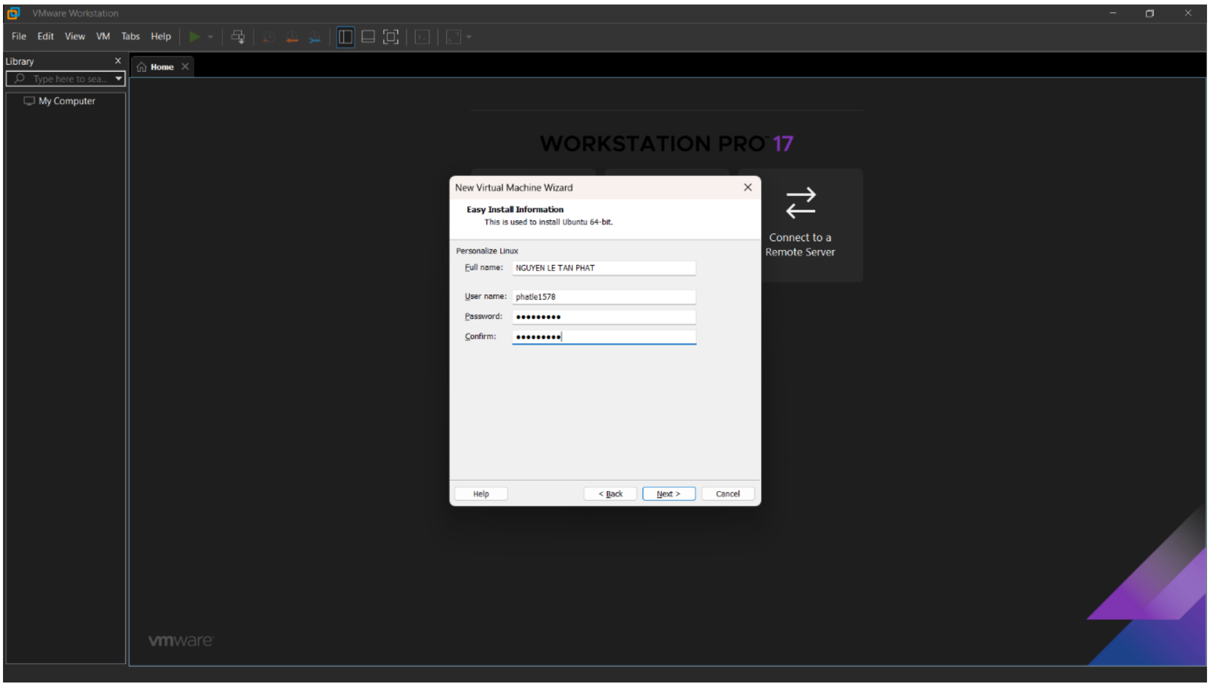
○ I will install the operating system later.

The virtual machine will be created with a blank hard disk.

Figure 4: VMware Workstation Pro virtual machine installer.

Step 3: Setup Linux profile.

- Fill in some information.



New Virtual Machine Wizard

Easy Install Information

This is used to install Ubuntu 64-bit.

Personalize Linux

Full name: NGUYEN LE TAN PHAT

User name: phatle1578

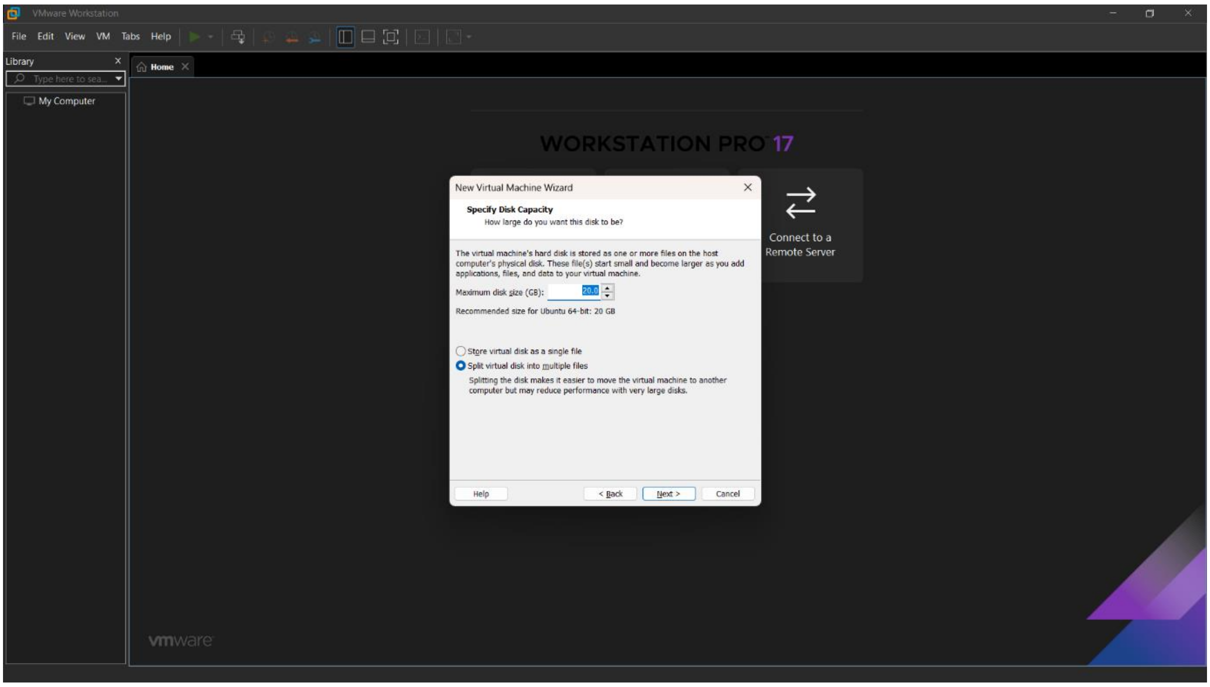
Password: •••••••••

Confirm: •••••••••

Figure 5: Personalize Linux.

Step 4: Setup virtual machine specs.

- Keep default and select "*next*"



New Virtual Machine Wizard

Specify Disk Capacity
How large do you want this disk to be?

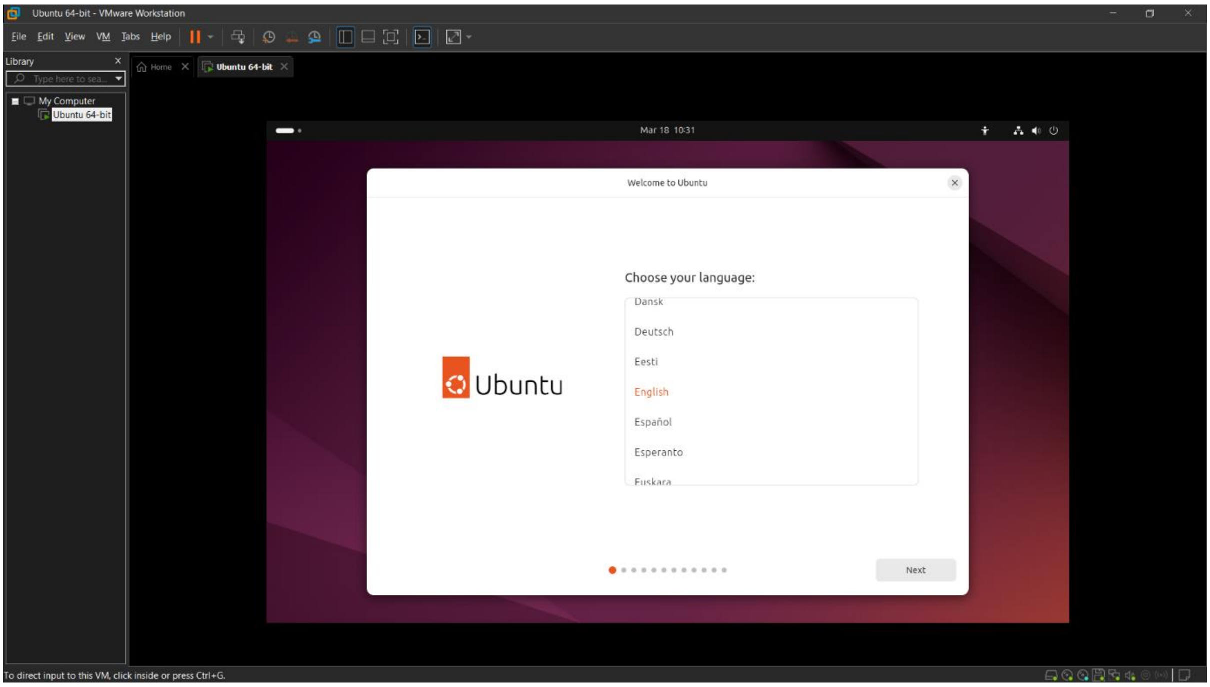
The virtual machine's hard disk is stored as one or more files on the host computer's physical disk. These file(s) start small and become larger as you add applications, files, and data to your virtual machine.

Maximum disk size (GB): 20.0

Recommended size for Ubuntu 64-bit: 20 GB

○ Store virtual disk as a single file
● Split virtual disk into multiple files
Splitting the disk makes it easier to move the virtual machine to another computer but may reduce performance with very large disks.

Figure 6: Virtual machine spec.



Welcome to Ubuntu

Choose your language:

Dansk
Deutsch
Eesti
English
Español
Esperanto
Euskara

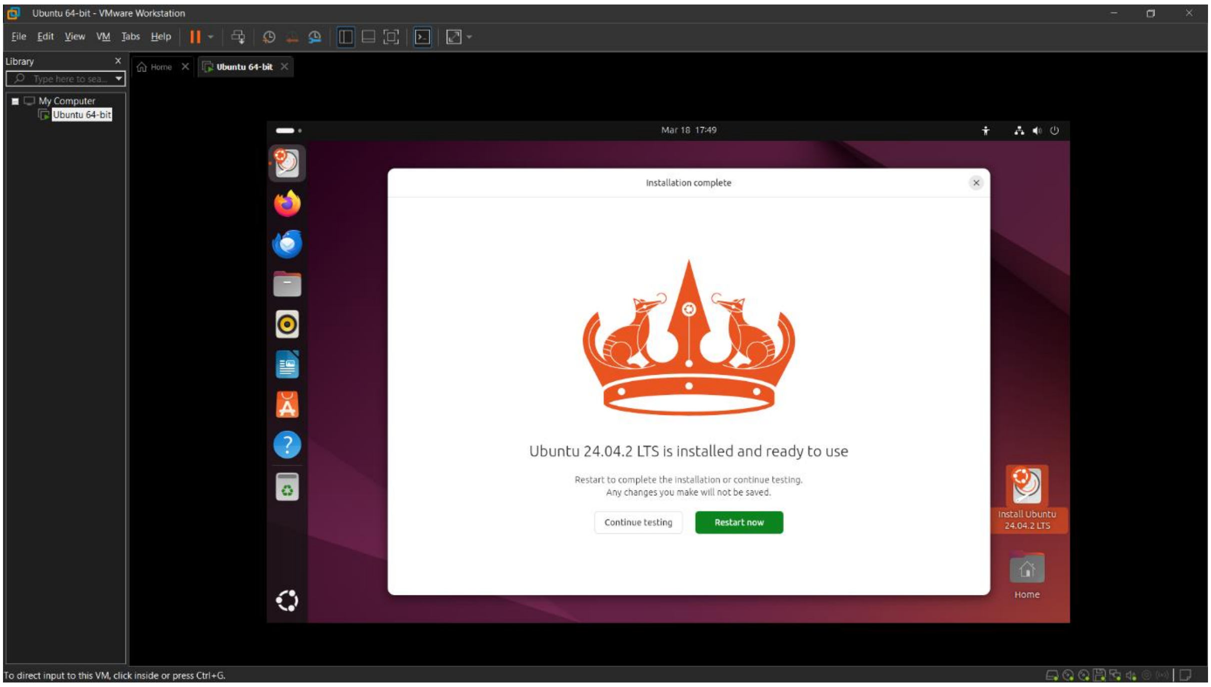
To direct input to this VM, click inside or press Ctrl+G.

Figure 7: Ubuntu setup (begin)

c. Setup Linux system in the virtual machine.

Step 1: Setup Ubuntu.

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



Installation complete

Ubuntu 24.04.2 LTS is installed and ready to use

Restart to complete the installation or continue testing.
Any changes you make will not be saved.

To direct input to this VM, click inside or press Ctrl+G.

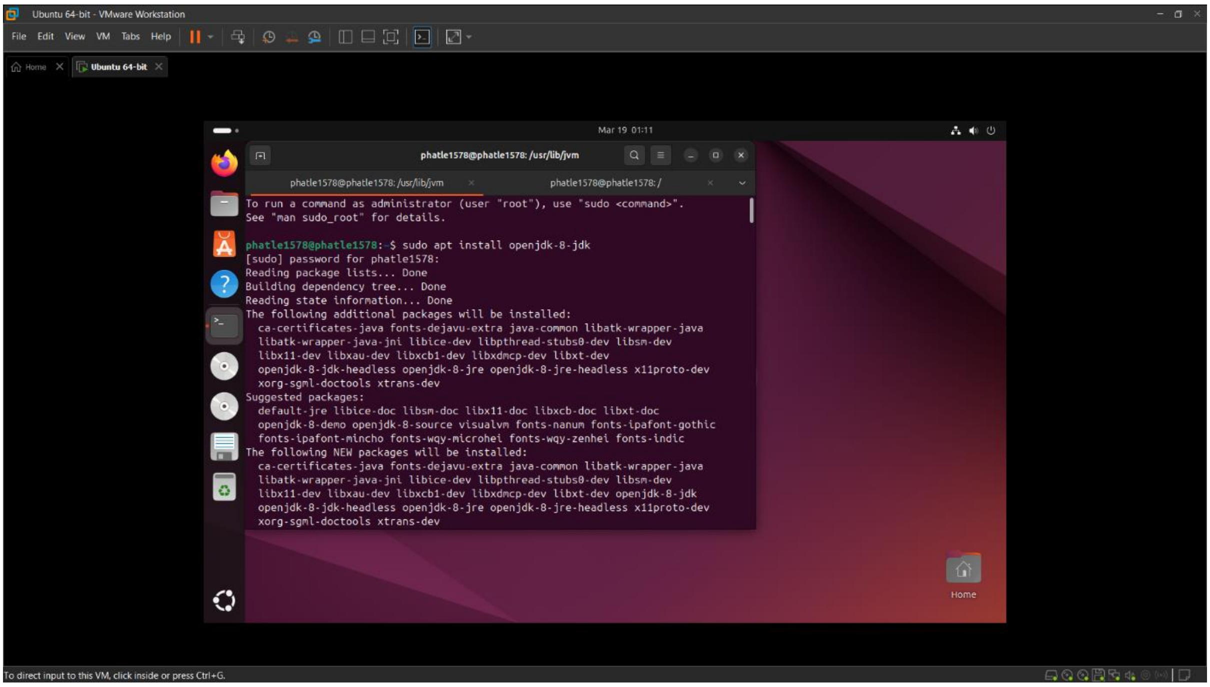
Figure 8: Ubuntu setup (finish)

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

a. Install Hadoop (VMware Workstation Pro).

Step 1: System preparation.

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

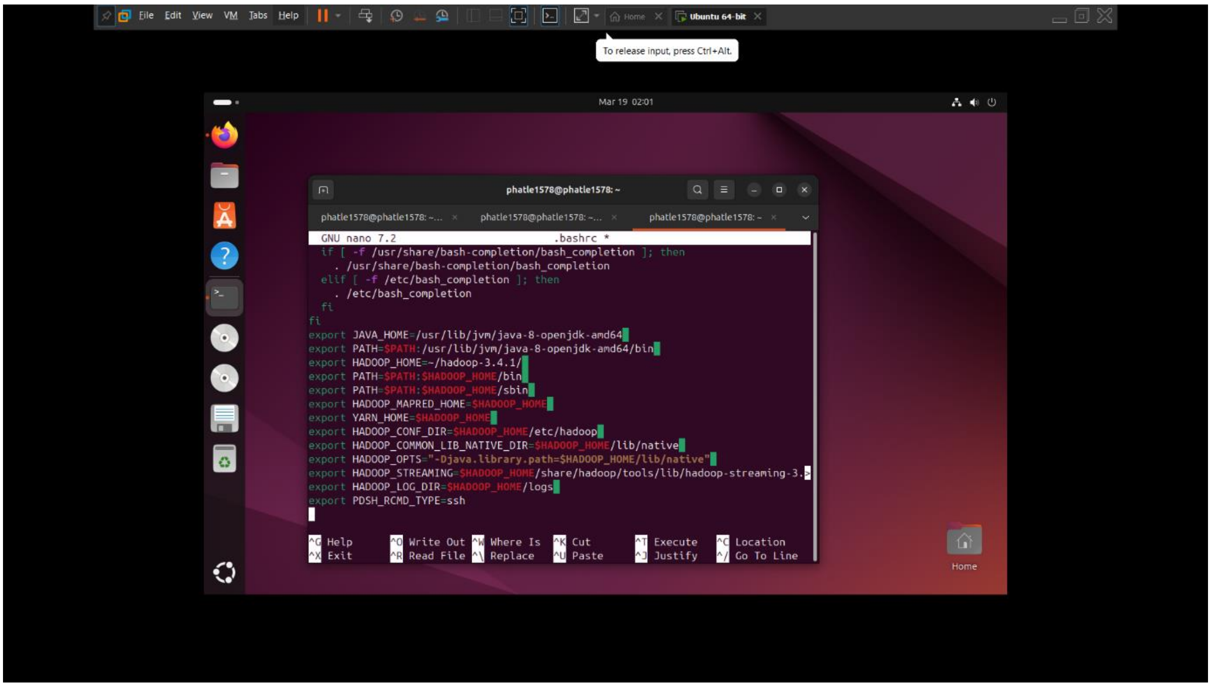

Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:11
phatle1578@phatle1578: /usr/lib/jvm
phatle1578@phatle1578: /usr/lib/jvm
phatle1578@phatle1578: /
Home
To direct input to this VM, click inside or press Ctrl+G.

```
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

phatle1578@phatle1578:~$ sudo apt install openjdk-8-jdk
[sudo] password for phatle1578:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev
  openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-dev
  xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc
  openjdk-8-demo openjdk-8-source visualvm fonts-nanum fonts-ipafont-gothic
  fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk
  openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-dev
  xorg-sgml-doctools xtrans-dev
```

Figure 9: Install OpenJDK 8

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

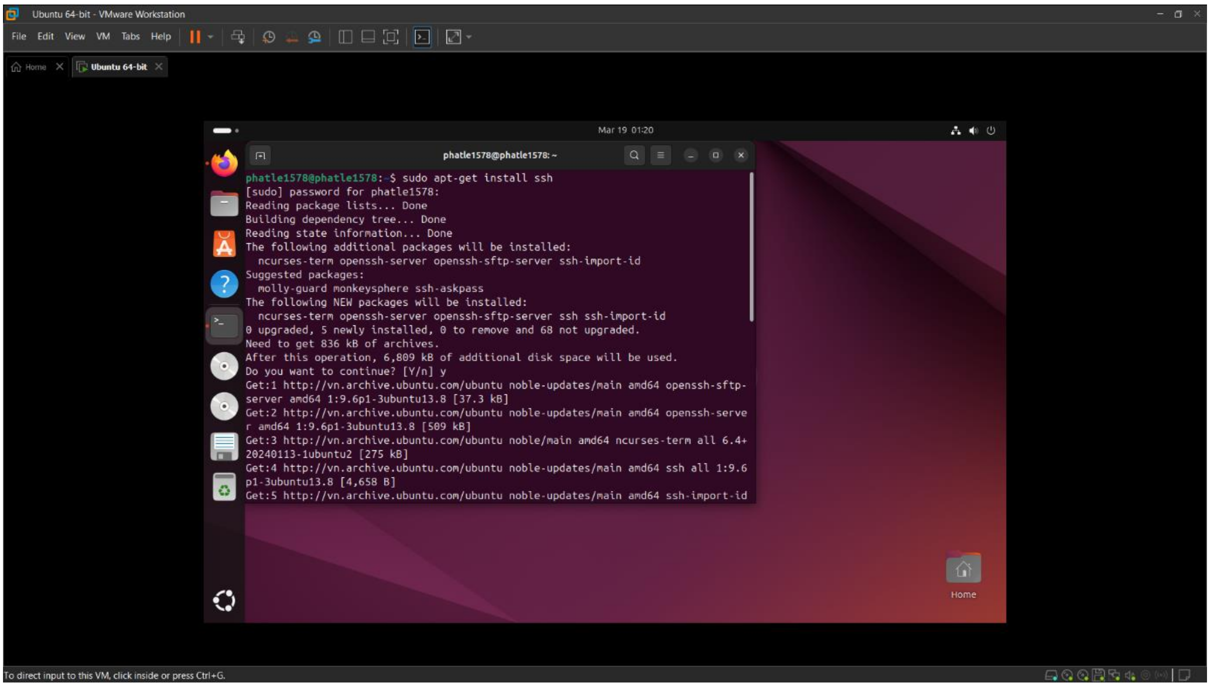
```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/hadoop-3.4.1 /
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
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
```


File Edit View VM Tabs Help
Home
Ubuntu 64-bit
To release input, press Ctrl+Alt.
Mar 19 02:01
phatle1578@phatle1578: ~
phatle1578@phatle1578: ~...
phatle1578@phatle1578: ~...
phatle1578@phatle1578: ~
Home

```
  GNU nano 7.2                          .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 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=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.>
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh


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

Figure 10: Configure bashrc file

- sudo apt-get install ssh: Install ssh.

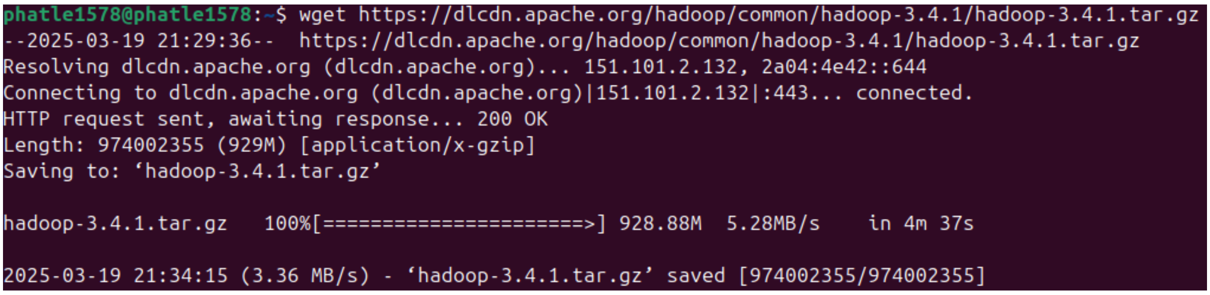

Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:20
phatle1578@phatle1578: ~
Home
To direct input to this VM, click inside or press Ctrl+G.

```
phatle1578@phatle1578:~$ sudo apt-get install ssh
[sudo] password for phatle1578:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 upgraded, 5 newly installed, 0 to remove and 68 not upgraded.
Need to get 836 kB of archives.
After this operation, 6,809 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://vn.archive.ubuntu.com/ubuntu noble-updates/main amd64 openssh-sftp-
server amd64 1:9.6p1-3ubuntu13.8 [37.3 kB]
Get:2 http://vn.archive.ubuntu.com/ubuntu noble-updates/main amd64 openssh-serve
r amd64 1:9.6p1-3ubuntu13.8 [509 kB]
Get:3 http://vn.archive.ubuntu.com/ubuntu noble/main amd64 ncurses-term all 6.4+
20240113-1ubuntu2 [275 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu noble-updates/main amd64 ssh all 1:9.6
p1-3ubuntu13.8 [4,658 B]
Get:5 http://vn.archive.ubuntu.com/ubuntu noble-updates/main amd64 ssh-import-id
```

Figure 11: Install ssh

Step 2: Download Apache Hadoop

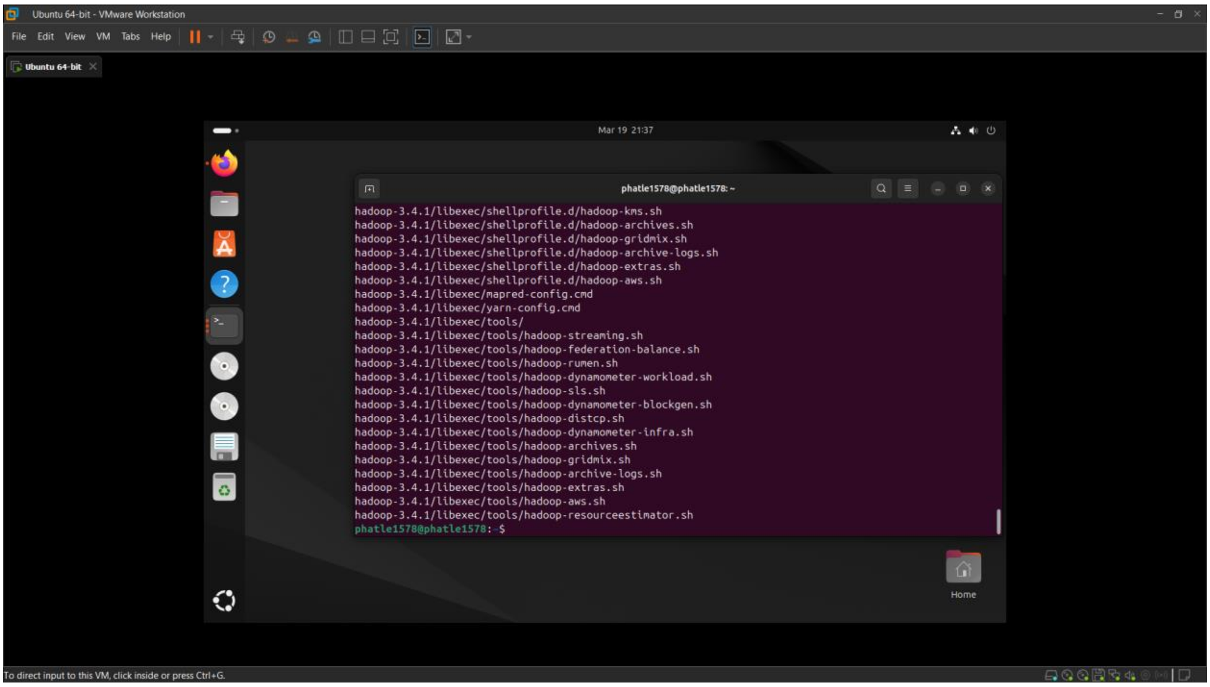
- wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz: download Apache Hadoop lastest version.



```
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 lastest version

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



Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Ubuntu 64-bit
Mar 19 21:37
phatle1578@phatle1578: ~
Home
To direct input to this VM, click inside or press Ctrl+G.

```
hadoop-3.4.1/libexec/shellprofile.d/hadoop-kms.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-archives.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-gridmix.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-archive-logs.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-extras.sh
hadoop-3.4.1/libexec/shellprofile.d/hadoop-aws.sh
hadoop-3.4.1/libexec/mapred-config.cmd
hadoop-3.4.1/libexec/yarn-config.cmd
hadoop-3.4.1/libexec/tools/
hadoop-3.4.1/libexec/tools/hadoop-streaming.sh
hadoop-3.4.1/libexec/tools/hadoop-federation-balance.sh
hadoop-3.4.1/libexec/tools/hadoop-rumen.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-workload.sh
hadoop-3.4.1/libexec/tools/hadoop-sls.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-blockgen.sh
hadoop-3.4.1/libexec/tools/hadoop-distcp.sh
hadoop-3.4.1/libexec/tools/hadoop-dynamometer-infra.sh
hadoop-3.4.1/libexec/tools/hadoop-archives.sh
hadoop-3.4.1/libexec/tools/hadoop-gridmix.sh
hadoop-3.4.1/libexec/tools/hadoop-archive-logs.sh
hadoop-3.4.1/libexec/tools/hadoop-extras.sh
hadoop-3.4.1/libexec/tools/hadoop-aws.sh
hadoop-3.4.1/libexec/tools/hadoop-resourceestimator.sh
phatle1578@phatle1578:~$
```

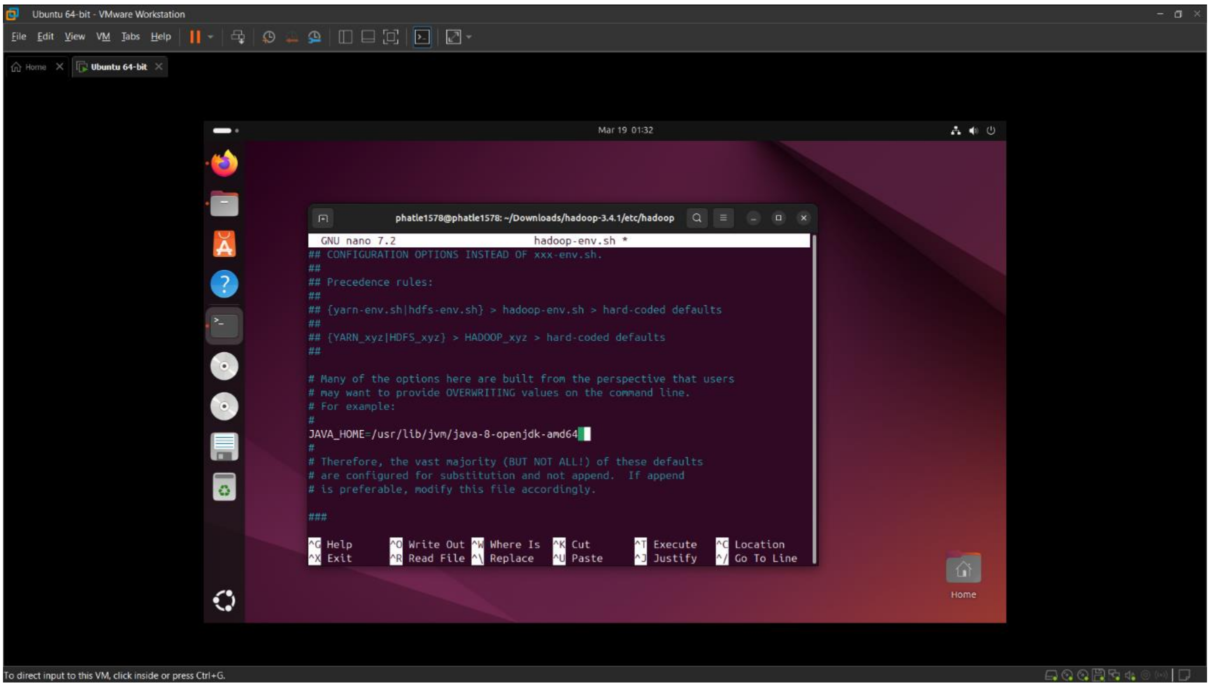
Figure 13: Extract the binary Hadoop file.

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
```


Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:32
Home
To direct input to this VM, click inside or press Ctrl+G.

phatle1578@phatle1578: ~/Downloads/hadoop-3.4.1/etc/hadoop

```
  GNU nano 7.2                          hadoop-env.sh *
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules:
##
## {yarn-env.sh|hdfs-env.sh} > hadoop-env.sh > hard-coded defaults
##
## {YARN_xyz|HDFS_xyz} > HADOOP_xyz > hard-coded defaults
##

# Many of the options here are built from the perspective that users
# may want to provide OVERWRITING values on the command line.
# For example:
#
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.

###

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

Figure 14: Configure hadoop-env.sh

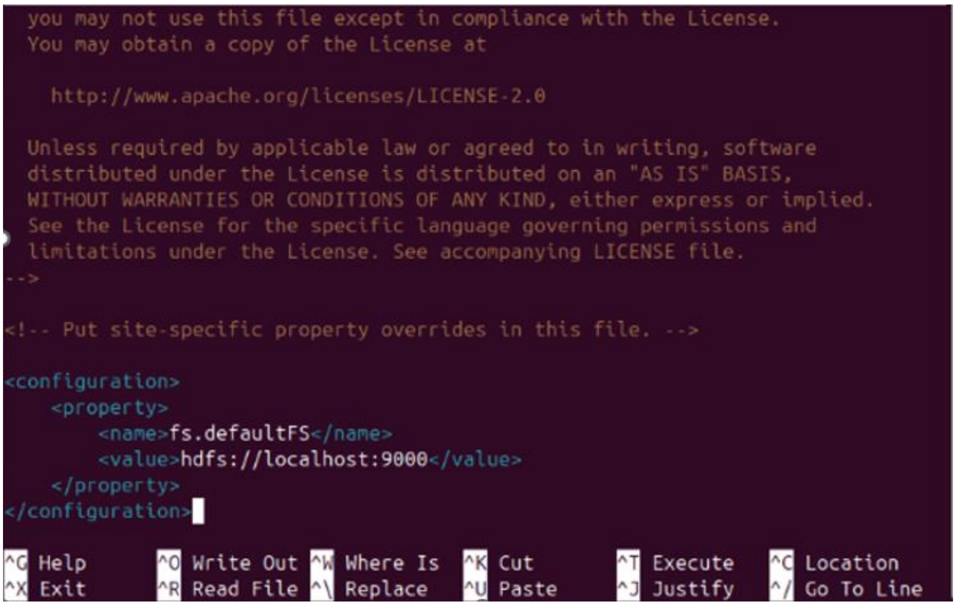
- Ctrl + O: Save
- Ctrl + X: Exit

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>

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

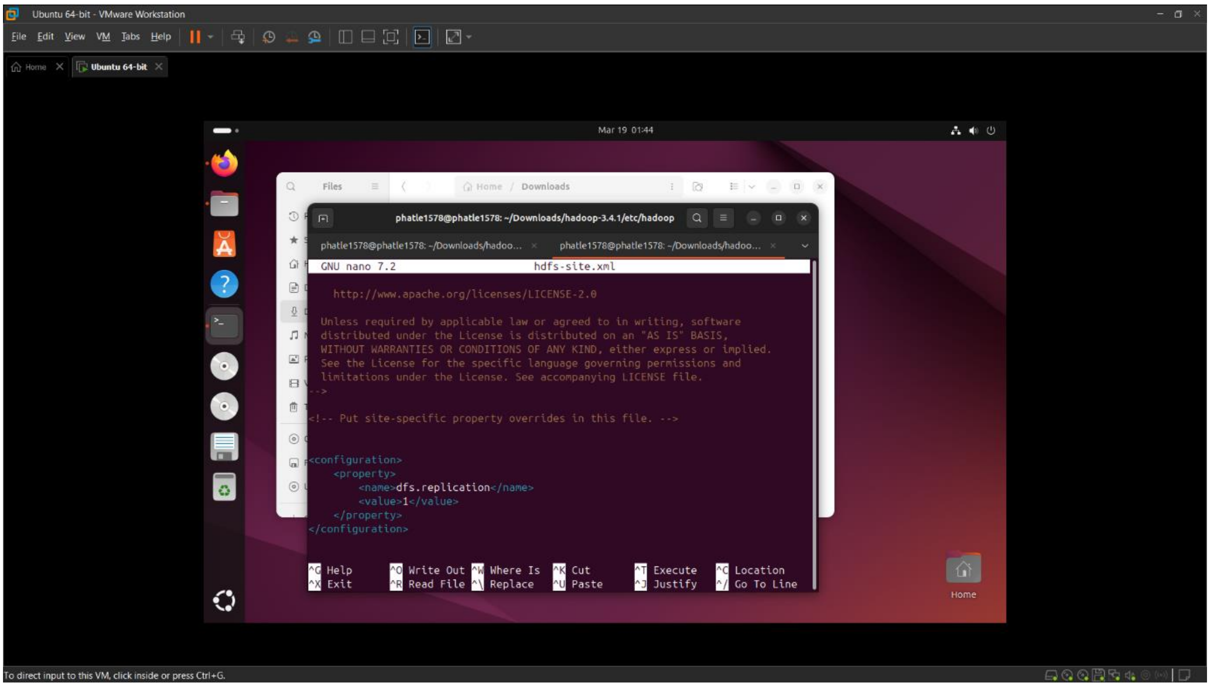
Figure 15: Configure core-site.xml

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.



Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:44
Files
Home / Downloads
Home
To direct input to this VM, click inside or press Ctrl+G.

phatle1578@phatle1578: ~/Downloads/hadoop-3.4.1/etc/hadoop

```
  GNU nano 7.2                    hdfs-site.xml

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


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

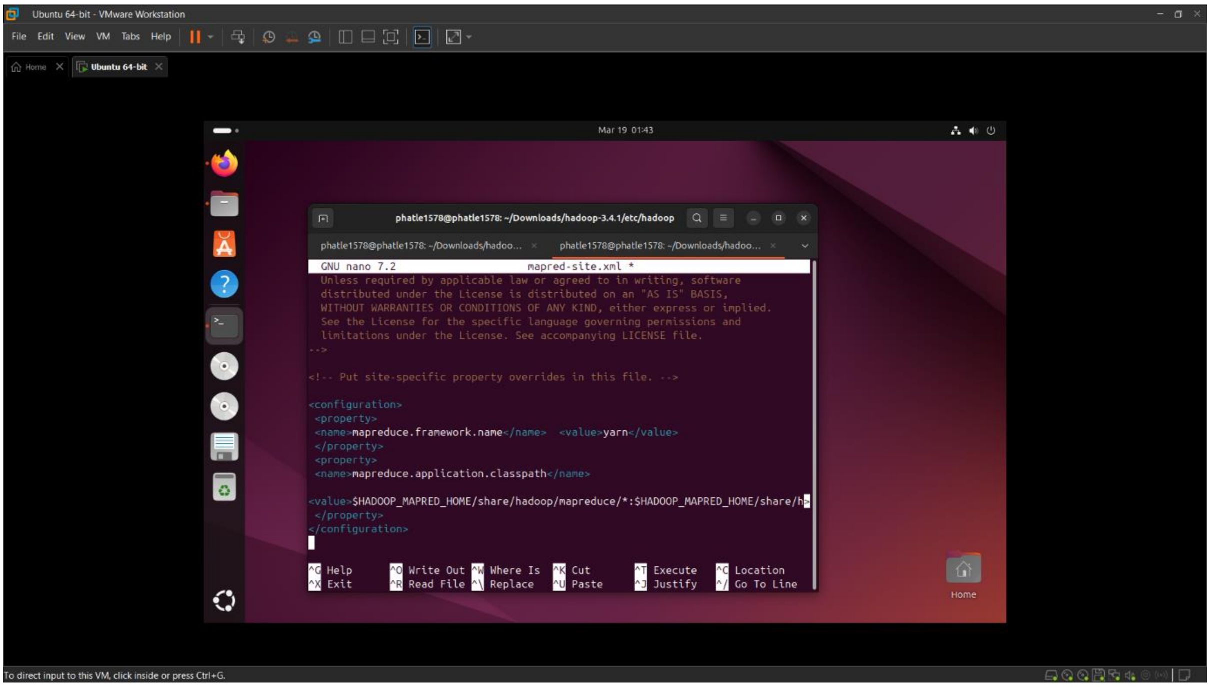
Figurte 16: Configure hdfs-site.xml

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.


Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:43
phatle1578@phatle1578: ~/Downloads/hadoop-3.4.1/etc/hadoop
phatle1578@phatle1578: ~/Downloads/hadoo...
phatle1578@phatle1578: ~/Downloads/hadoo...
Home
To direct input to this VM, click inside or press Ctrl+G.

```
  GNU nano 7.2                  mapred-site.xml *
  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>
 <property>
 <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/h>
 </property>
</configuration>


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

Figure 17: Configure mapred-stie.xml

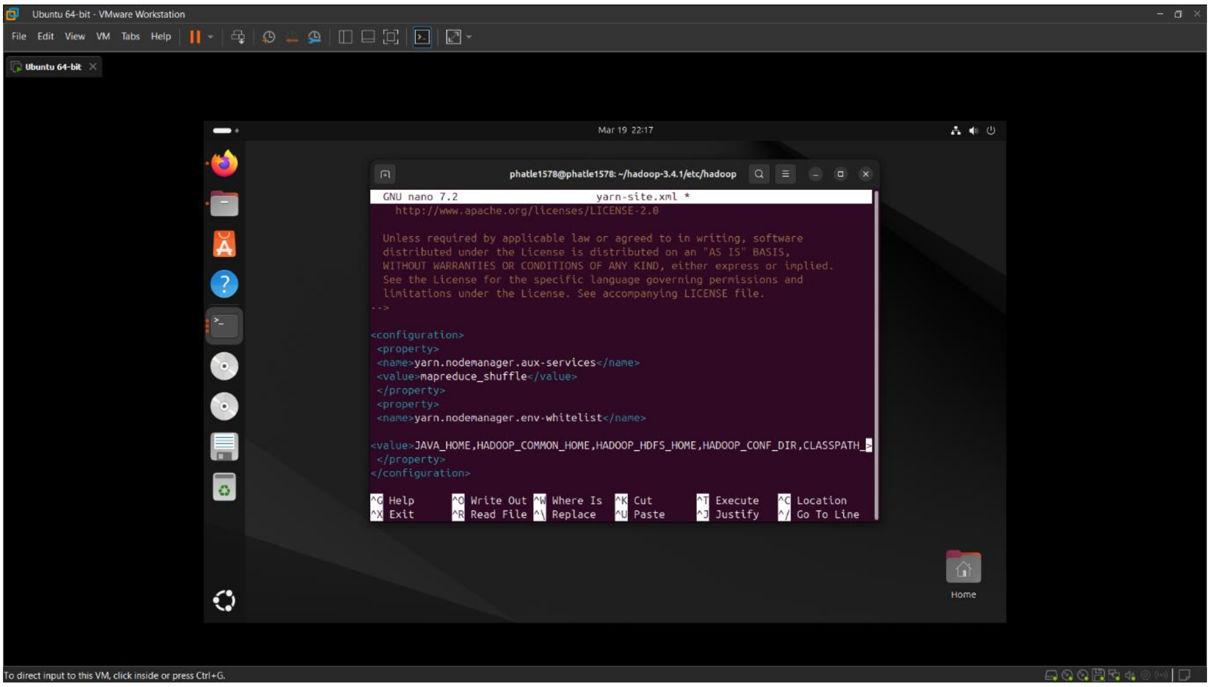
v. yarn-site.xml

- 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.

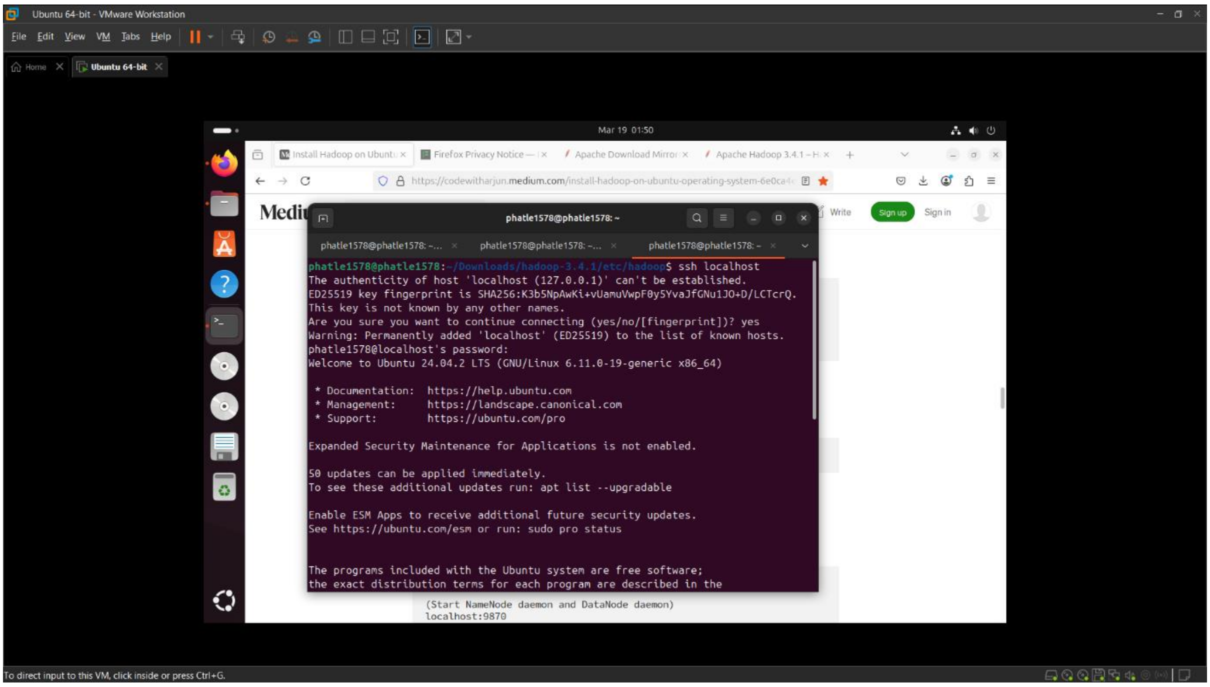

Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Ubuntu 64-bit
Mar 19 22:17
phatle1578@phatle1578: ~/hadoop-3.4.1/etc/hadoop
Home
To direct input to this VM, click inside or press Ctrl+G.

```
  GNU nano 7.2                          yarn-site.xml *
    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.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_>
 </property>
</configuration>

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

Figure 18: Configure yarn-site.xml

Step 4: Setup ssh

- ssh localhost: Start the ssh localhost.

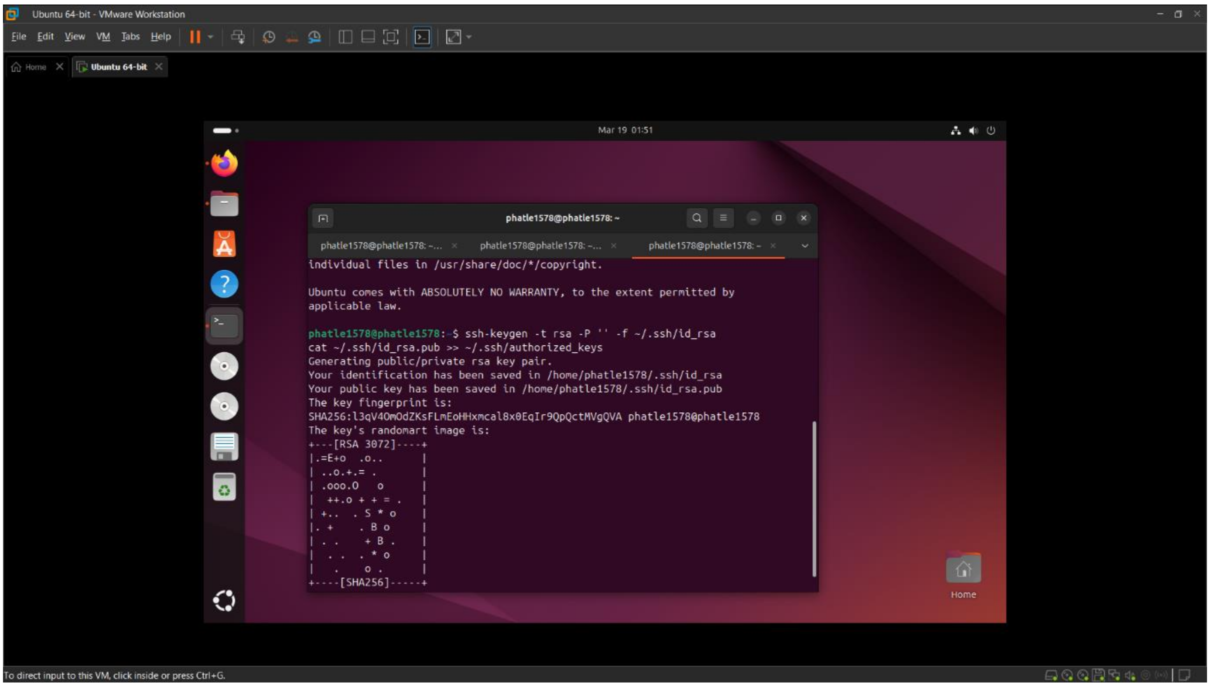

Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:50
Install Hadoop on Ubuntu
Firefox Privacy Notice —
Apache Download Mirror
Apache Hadoop 3.4.1 – H
https://codewitharjun.medium.com/install-hadoop-on-ubuntu-operating-system-6e0ca4
Medi
Write
Sign up
Sign in
phatle1578@phatle1578: ~
phatle1578@phatle1578: ~...
phatle1578@phatle1578: ~...
phatle1578@phatle1578: ~
To direct input to this VM, click inside or press Ctrl+G.

```
phatle1578@phatle1578:~/Downloads/hadoop-3.4.1/etc/hadoop$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:K3b5NpAwKi+vUamuVwpF0y5YvaJfGNu1JO+D/LCTcrQ.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
phatle1578@localhost's password:
Welcome to Ubuntu 24.04.2 LTS (GNU/Linux 6.11.0-19-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

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

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
```

```
(Start NameNode daemon and DataNode daemon)
localhost:9870
```

Figure 19: Start ssh localhost

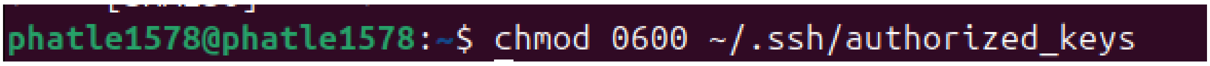
- ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa: Generate private key at ~/.ssh/id_rsa and public key at ~/.ssh/id_rsa.pub
- cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys: add public key we just create previously to the authorized _keys list.


Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Home
Ubuntu 64-bit
Mar 19 01:51
phatle1578@phatle1578: ~
Home
To direct input to this VM, click inside or press Ctrl+G.

```
individual files in /usr/share/doc/*/copyright.

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

phatle1578@phatle1578:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
Generating public/private rsa key pair.
Your identification has been saved in /home/phatle1578/.ssh/id_rsa
Your public key has been saved in /home/phatle1578/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:l3qV4OmOdZKsFLmEoHHxmcal8x0EqIr9QpQctMVgQVA phatle1578@phatle1578
The key's randomart image is:
+---[RSA 3072]----+
|.=E+o  .o..      |
| ..o.+.= .       |
| .ooo.O   o      |
|  ++.o + + = .   |
| +..  . S * o    |
|. +    . B o     |
| . .    + B .    |
|  . .  . * o     |
|   .    o .      |
+----[SHA256]-----+
```

Figure 20: Generate ssh key.

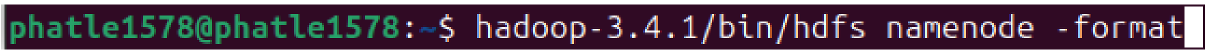
- chmod 0600 ~/.ssh/authorized_keys: Set the access permission to the key. Only the owner can read and write.



```
phatle1578@phatle1578:~$ chmod 0600 ~/.ssh/authorized_keys
```

Figure 21: Set key permission.

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

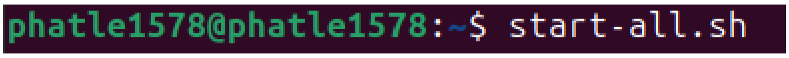


```
phatle1578@phatle1578:~$ hadoop-3.4.1/bin/hdfs namenode -format
```

Figure 22: Restart the distributed system.

b. Start Hadoop Cluster.

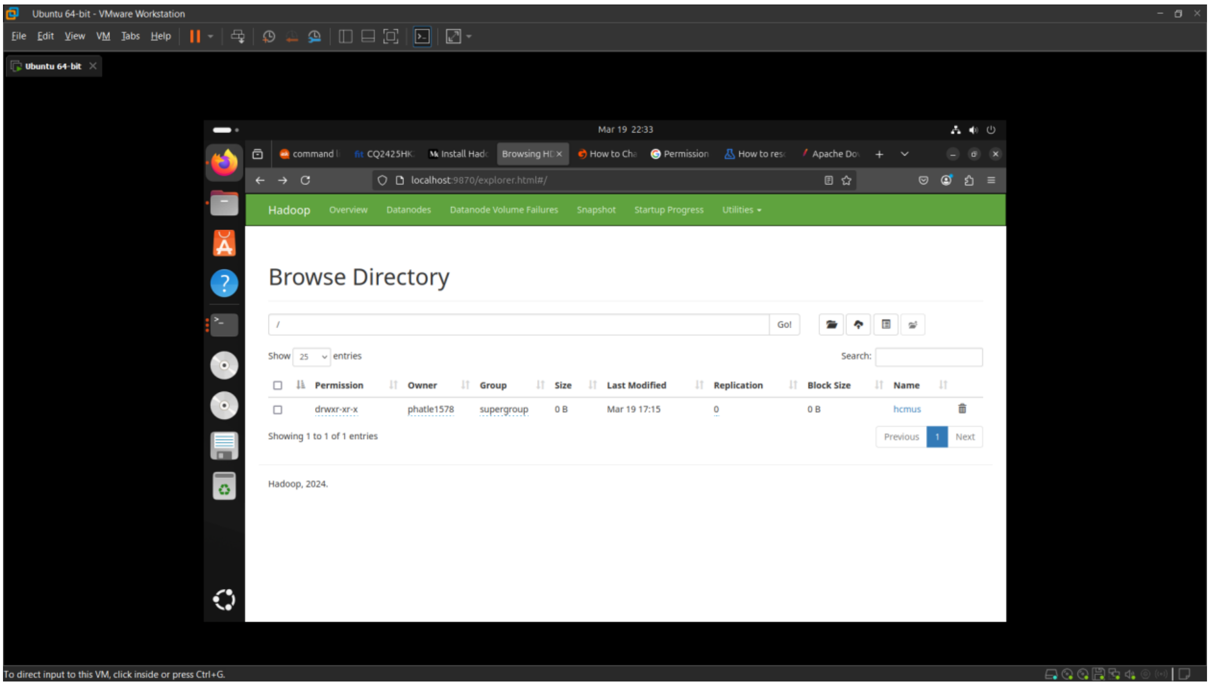
- start-all.sh: Start all hadoop services.



```
phatle1578@phatle1578:~$ start-all.sh
```

Figure 23: Start all hadoop services

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


Ubuntu 64-bit - VMware Workstation
File Edit View VM Tabs Help
Ubuntu 64-bit
Mar 19 22:33
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
To direct input to this VM, click inside or press Ctrl+G.

Browse Directory

/ Go!

Show 25 entries Search:

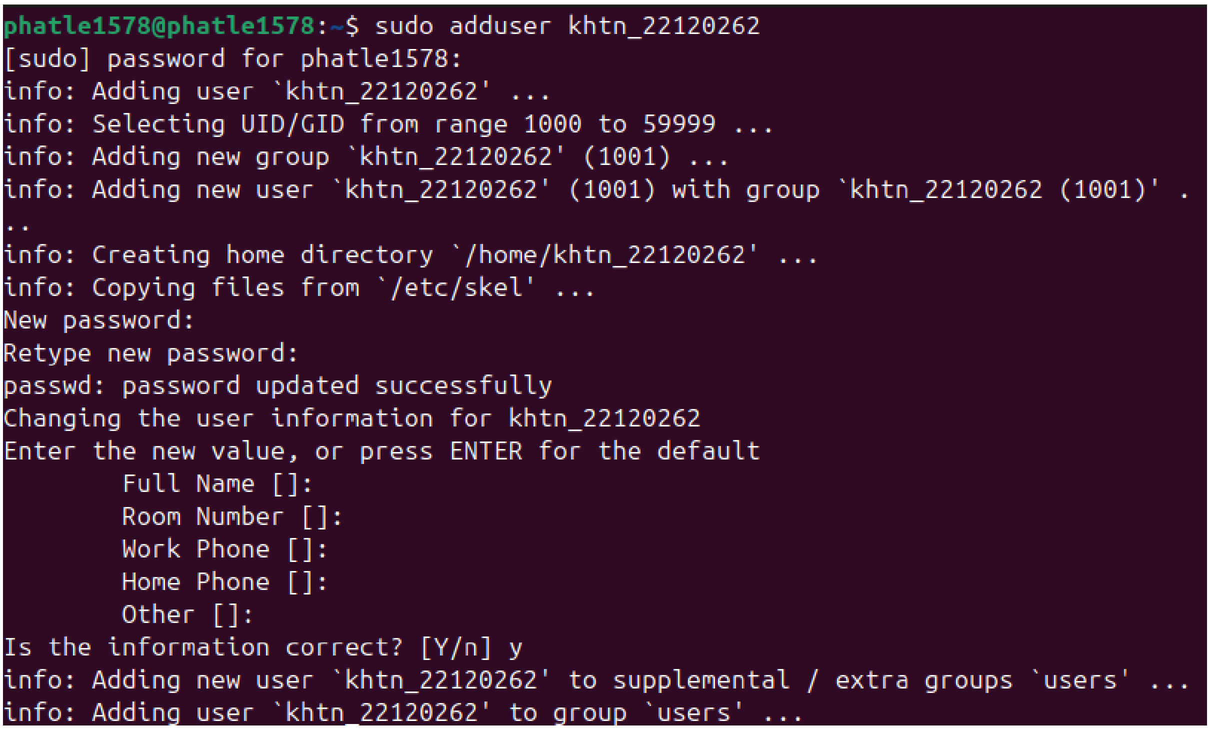
□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	phatle1578	supergroup	0 B	Mar 19 17:15	0	0 B	hcmus

Showing 1 to 1 of 1 entries Previous 1 Next

Hadoop, 2024.

Figure 24: Create hcmus folder in hadoop.

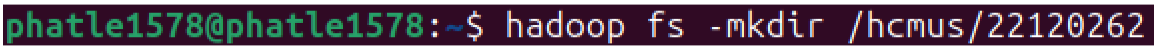
- sudo adduser khtn_22120262: Create new user with name khtn_22120262



```
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 hcmus folder.

- Download the require file from courses.fit.hcmus.edu.vn
- Right click in the folder Downloads to open in terminal

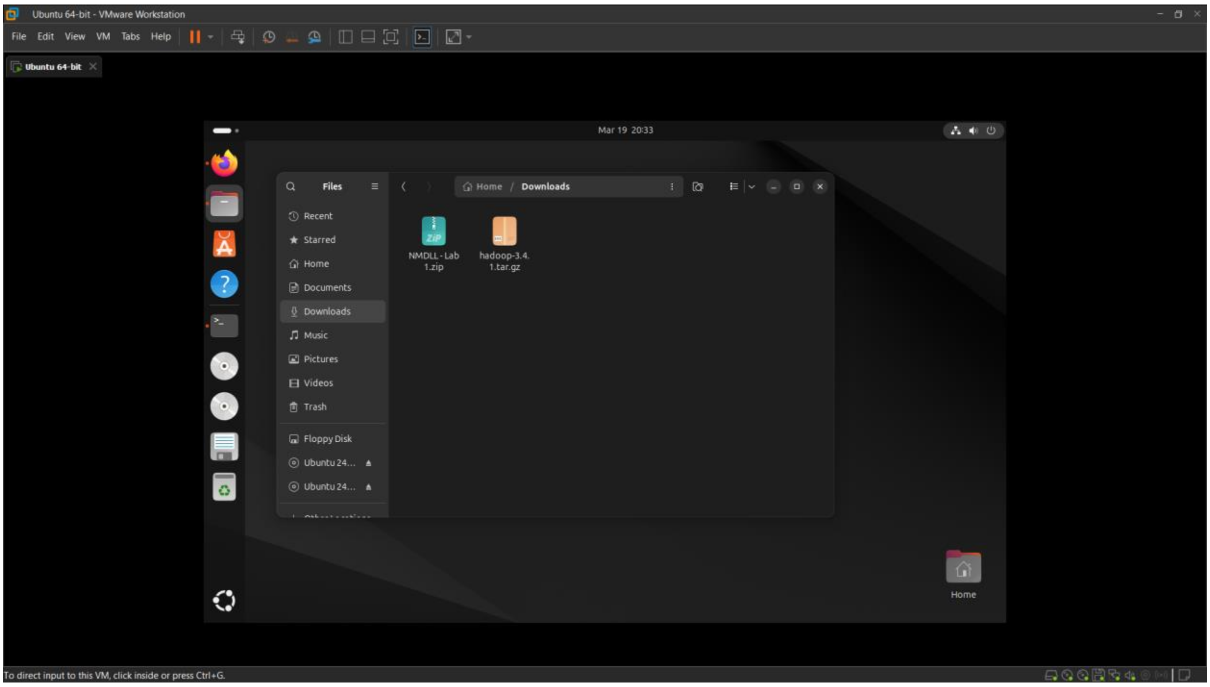
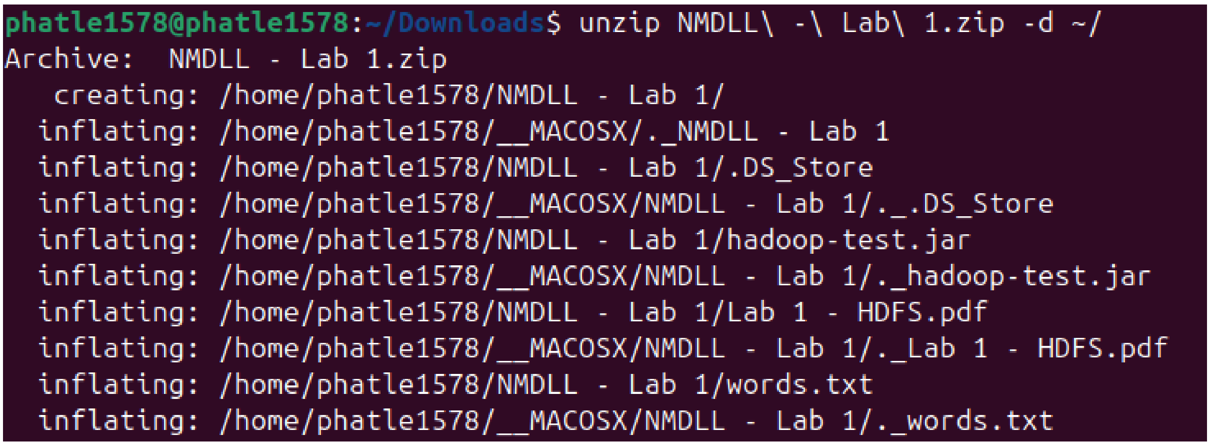


Figure 26: Downloads require file.

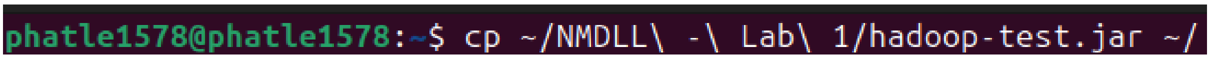
- unzip NMDLL\ -\ Lab\ 1.zip - ~\: Unzip the file we just download into the /home/phatle1578 directory..



```
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.

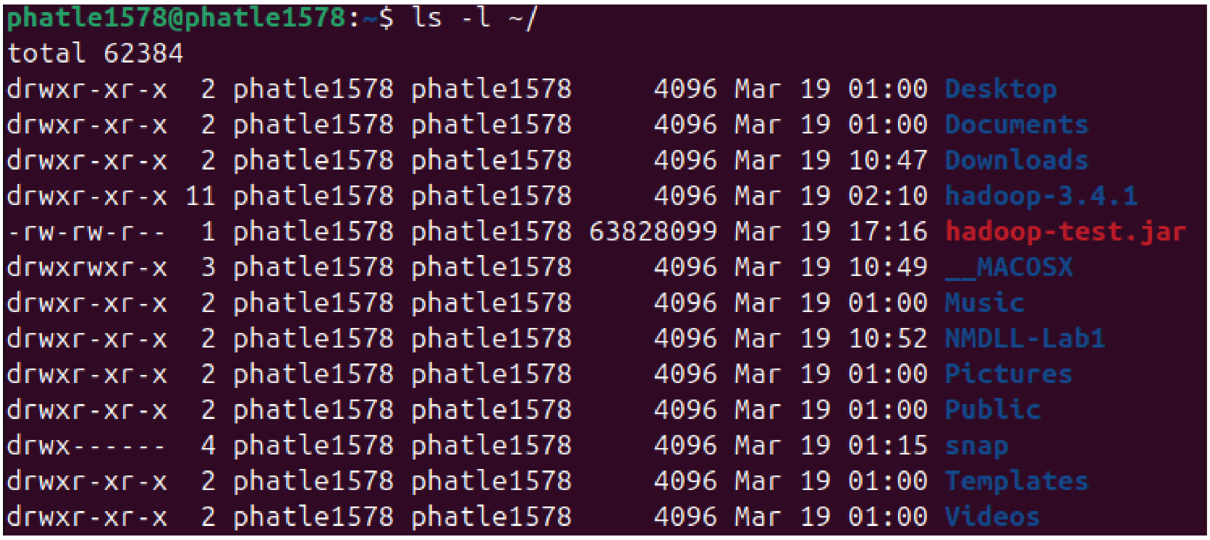
- cp ~/NMDLL\ -\ Lab\ 1/hadoop-test.jar ~/: Move the hadoop-test.jar file out to /home/phatle1578 directory for more convenient.



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

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

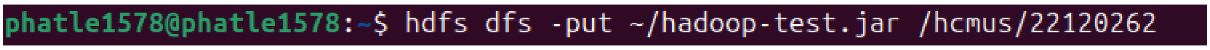
- ls -l ~/: Verify it has been copy to the home directory



```
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.

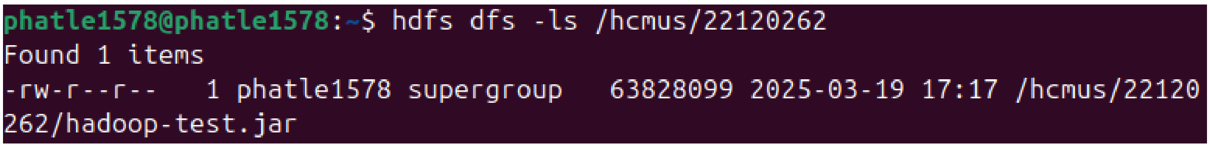
- hdfs dfs -put ~/hadoop-test.jar /hcmus/22120262: Move the hadoop-test.jar into the path /hcmus/22120262 in hdfs.



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

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

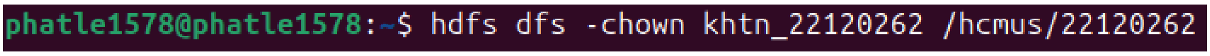
- hdfs dfs -ls /hcmus/22120262: Verify it has been move to /hcmus/22120262 in hdfs



```
phatle1578@phatle1578:~$ hdfs dfs -ls /hcmus/22120262
Found 1 items
-rw-r--r--   1 phatle1578 supergroup   63828099 2025-03-19 17:17 /hcmus/22120
262/hadoop-test.jar
```

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

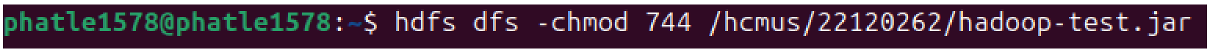
- hdfs dfs -chown khtn_22120262 /hcmus/22120262: Set ownership of /hcmus/22120262/ to khtn_22120262 user.



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

Figure 32: Set ownership.

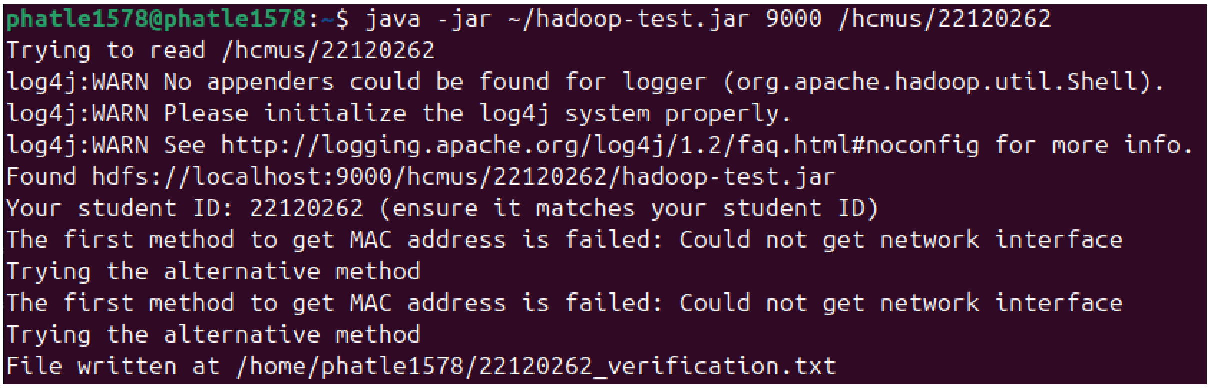
- hdfs dfs -chmod 744 /hcmus/22120262/hadoop-test.jar: Set the file permissions to 744



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

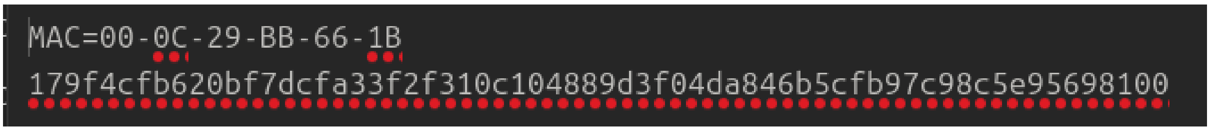
Figure 33: Set the file permissions to 744.

- 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.



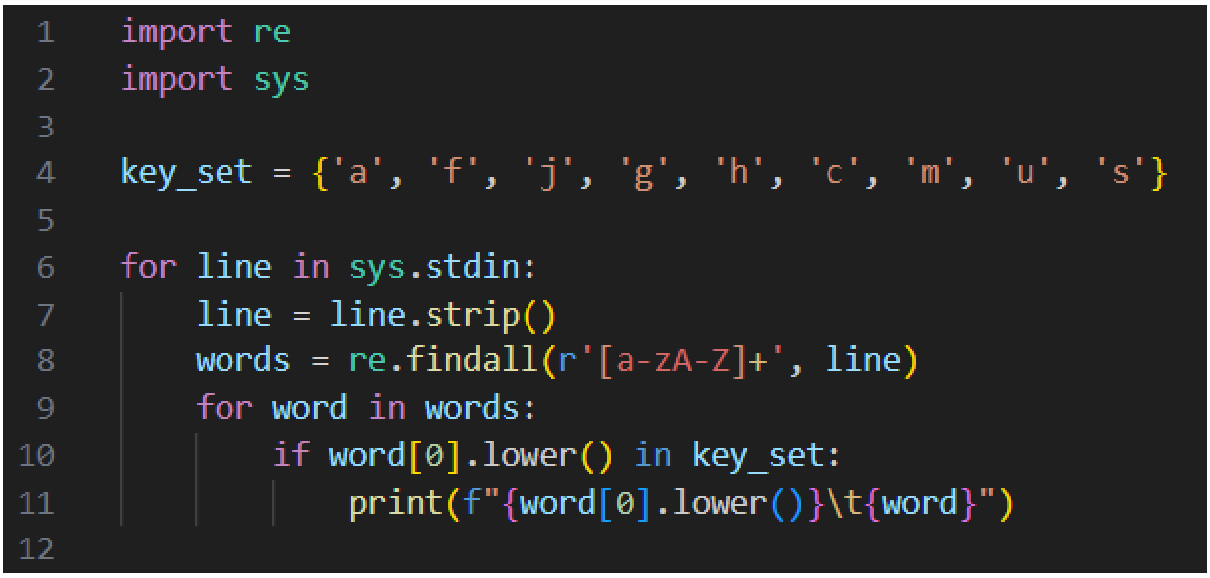
```
MAC=00-0C-29-BB-66-1B
179f4cfb620bf7dcfa33f2f310c104889d3f04da846b5cfb97c98c5e95698100
```

Figure 36: Verification file

II. Word Count

1. Mappper.

The script reads input from *sys.stdin*, extracts words that contain only character in the alphabet using *re.findall(r'[a-zA-Z]+', line)*, checks if the first letter (lowercased) is in *key_set*, and if so, prints the letter and the word separated by a tab (*\t*).

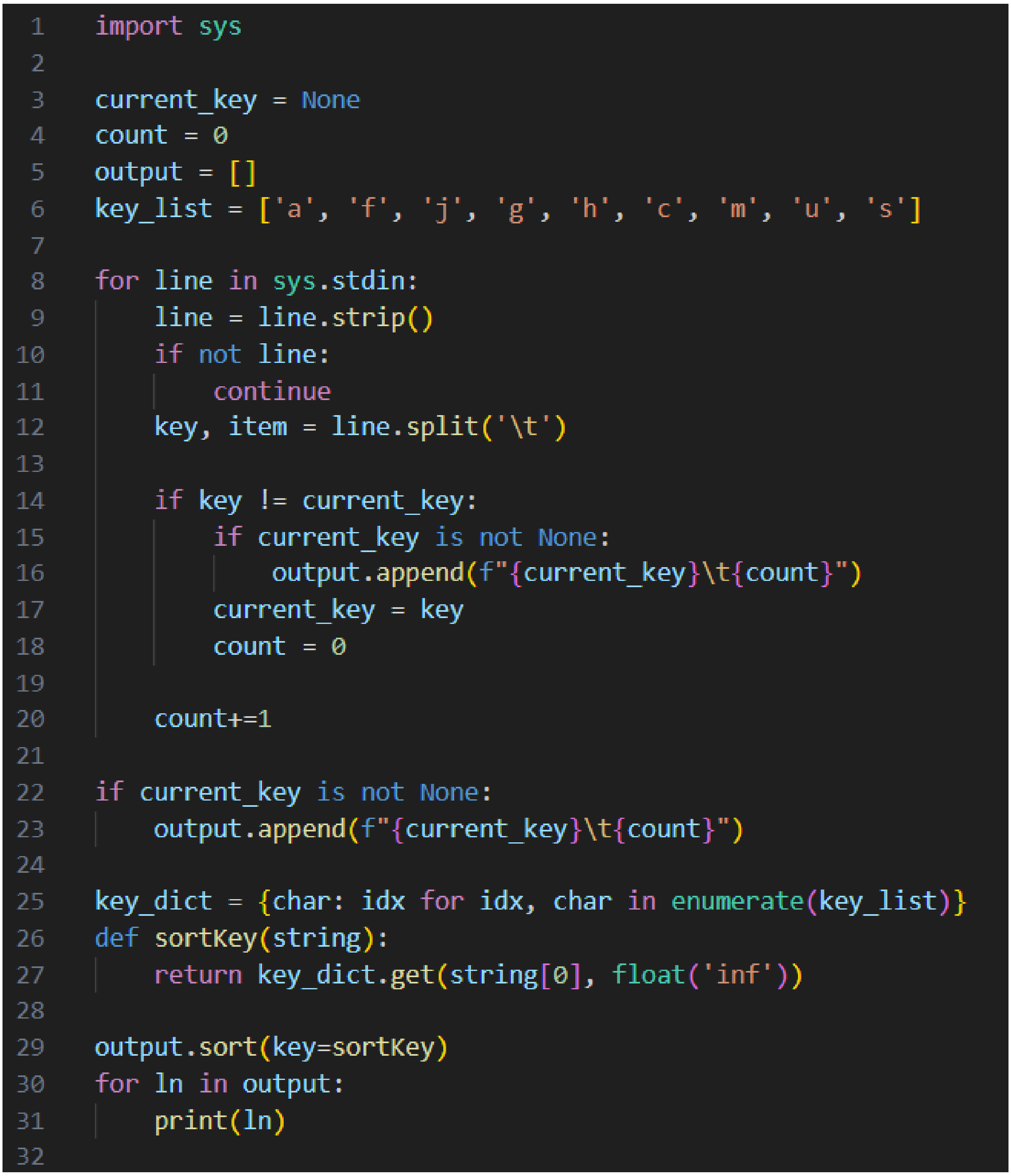


```python
import re
import sys

key_set = {'a', 'f', 'j', 'g', 'h', 'c', 'm', 'u', 's'}

for line in sys.stdin:
    line = line.strip()
    words = re.findall(r'[a-zA-Z]+', line)
    for word in words:
        if word[0].lower() in key_set:
            print(f"{word[0].lower()}\t{word}")

```

Figure 37: mapper.py

2. Reducer.

The script reads input from *sys.stdin*, processes key-value pairs (tab-separated), counts occurrences of each key, and stores results in output. It then sorts the output based on key_list order using a dictionary lookup (key_dict) and prints the sorted results.

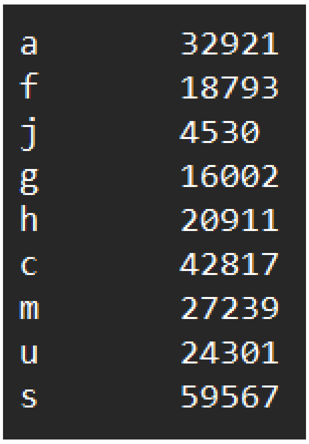


```python
import sys

current_key = None
count = 0
output = []
key_list = ['a', 'f', 'j', 'g', 'h', 'c', 'm', 'u', 's']

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue
    key, item = line.split('\t')

    if key != current_key:
        if current_key is not None:
            output.append(f"{current_key}\t{count}")
        current_key = key
        count = 0

    count+=1

if current_key is not None:
    output.append(f"{current_key}\t{count}")

key_dict = {char: idx for idx, char in enumerate(key_list)}
def sortKey(string):
    return key_dict.get(string[0], float('inf'))

output.sort(key=sortKey)
for ln in output:
    print(ln)

```

Figure 38: reducer.py

3. Result.

Result is the txt file with the TSV-formatted.



```
a	32921
f	18793
j	4530
g	16002
h	20911
c	42817
m	27239
u	24301
s	59567
```

Figure 38: result.txt