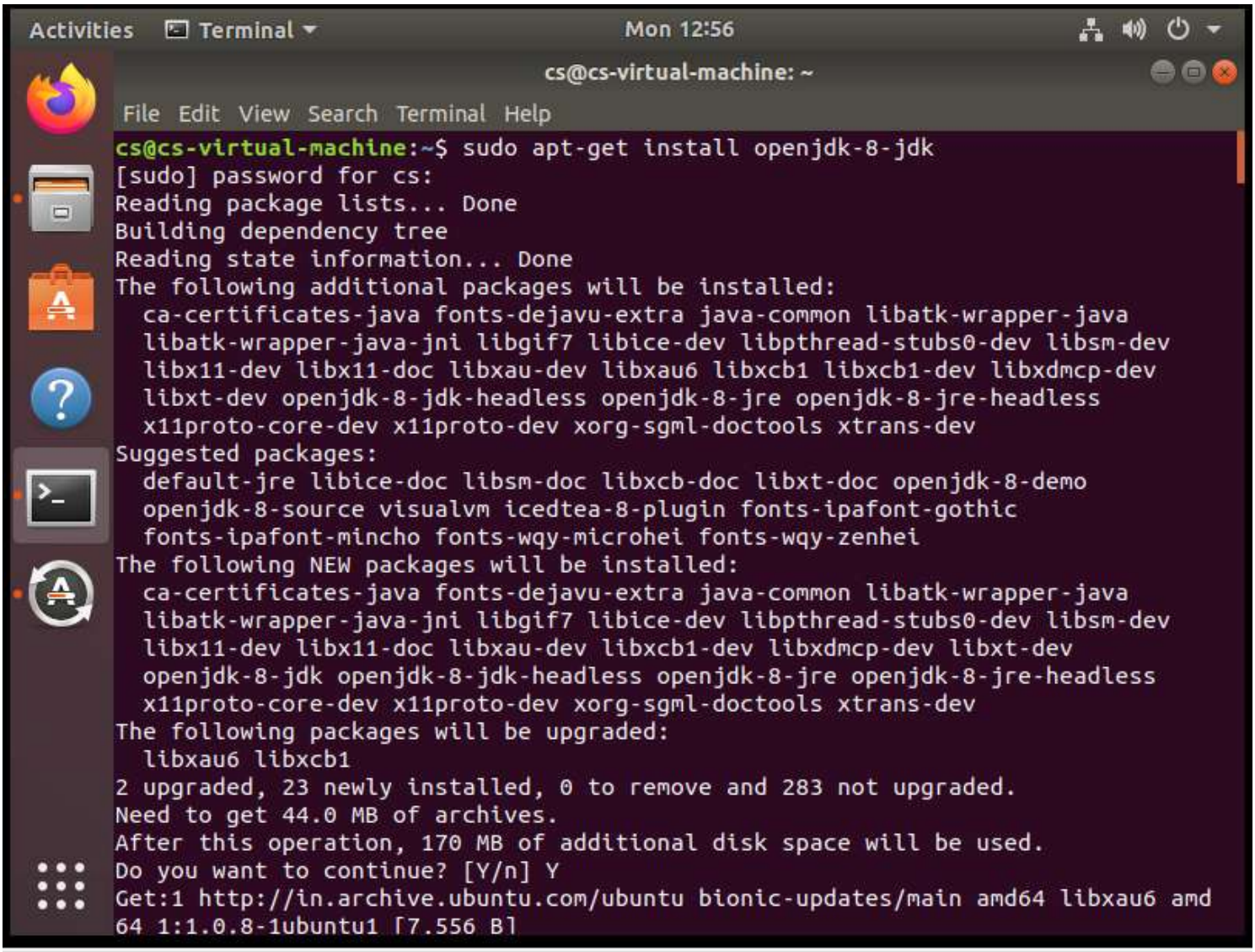
Practical No.1.

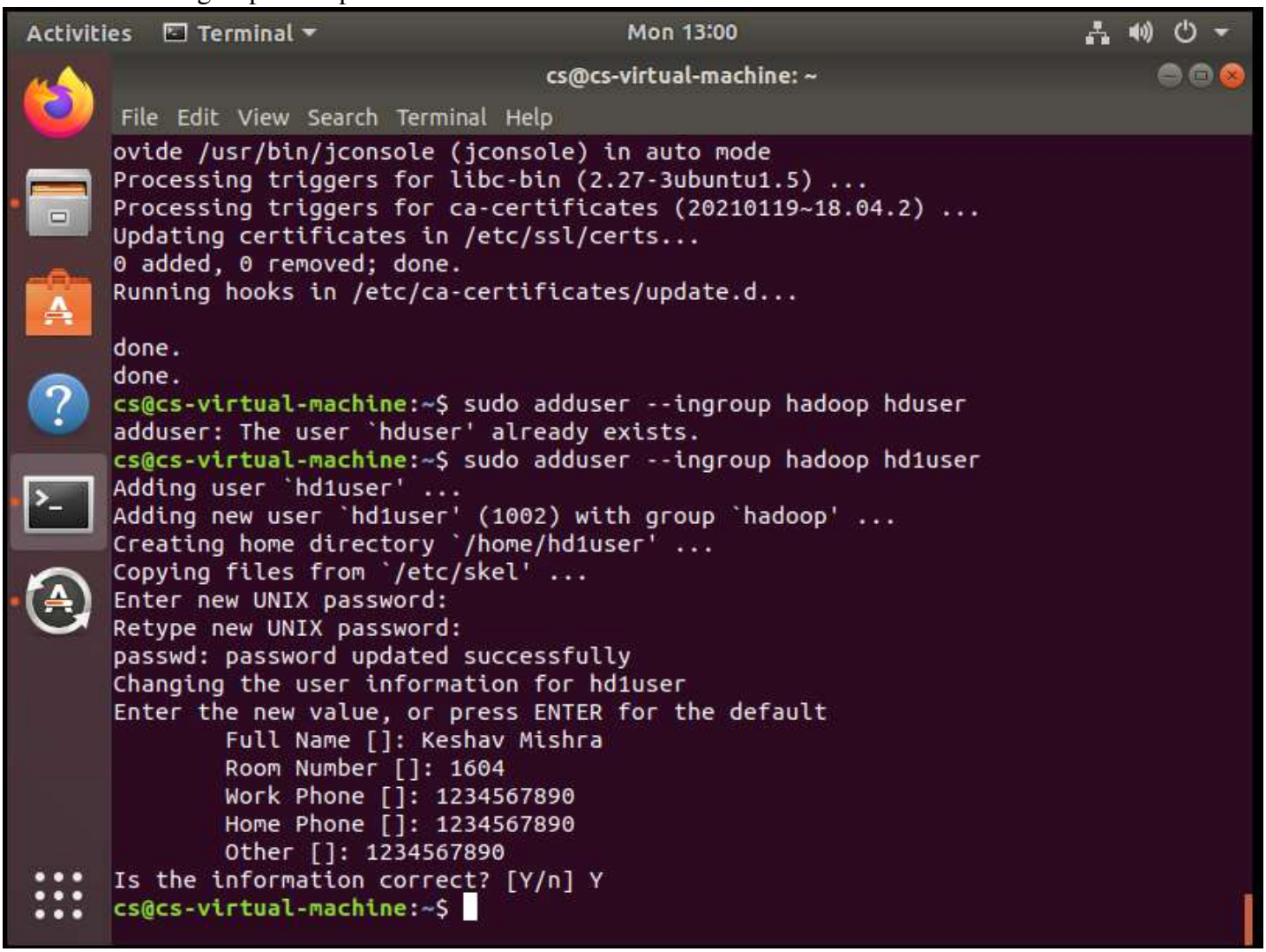
Aim: Install, configure and run Hadoop and HDFS ad explore HDFS
>Install Virtual Box in your system
https://download.virtualbox.org/virtualbox/6.1.22/VirtualBox-6.1.22-144080-Win.exe

>Create a ubuntu VM inside Virtaul box using the below ISO
https://ubuntu.com/download/desktop/thank-you?version=20.04.2.0&architecture=amd64

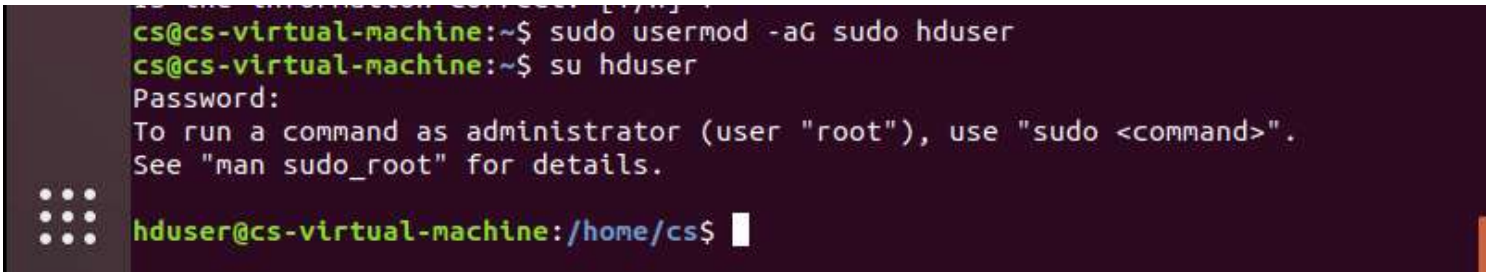
>use the following link to check ubuntu installation process
https://youtu.be/x5MhydijWmc
1. First you need to perform some prerequisites
sudo apt-get install openjdk-8-jdk



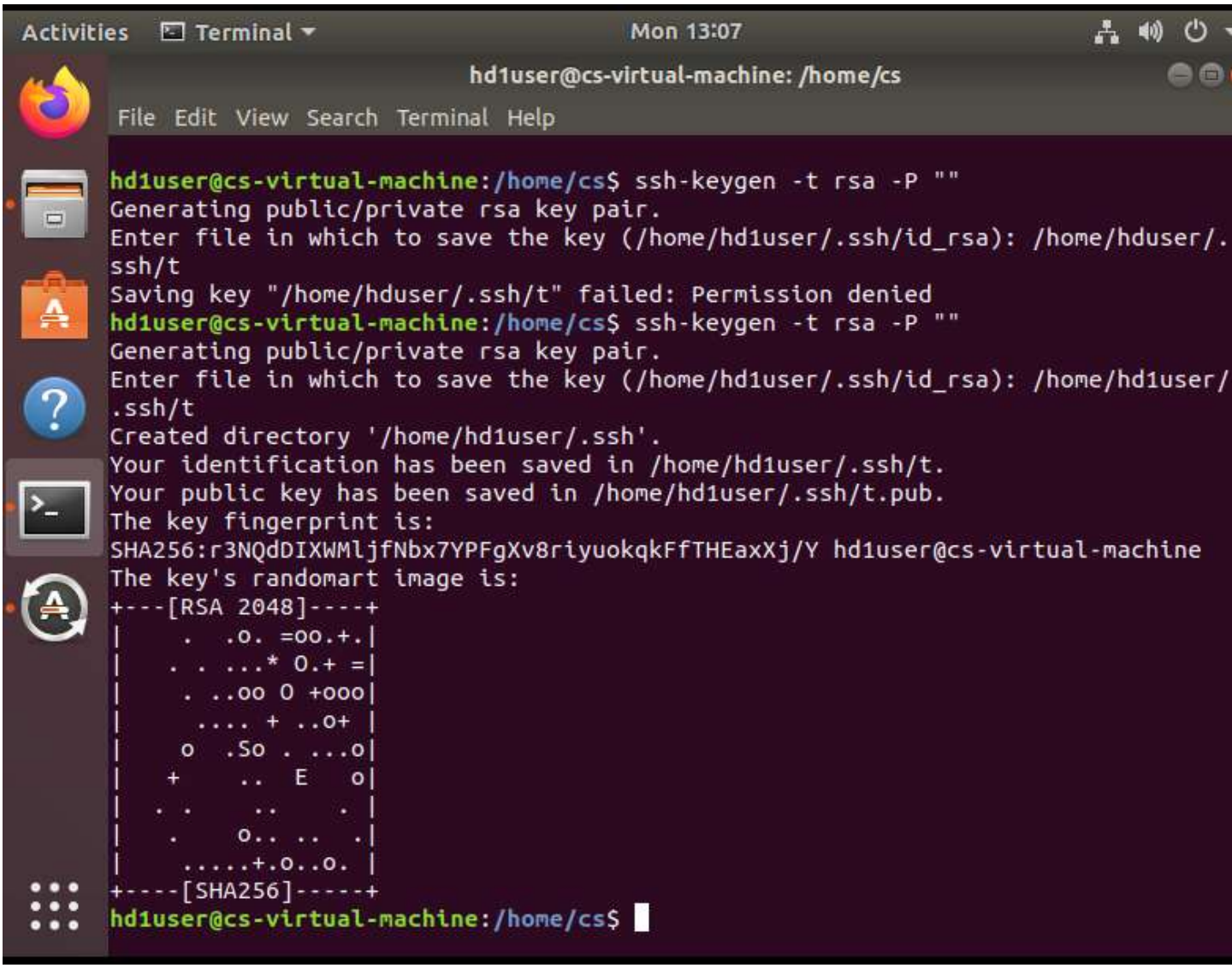
sudo addgroup hadoop
sudo adduser --ingroup hadoop hduser



usermod -aG sudo hduser
su hduser

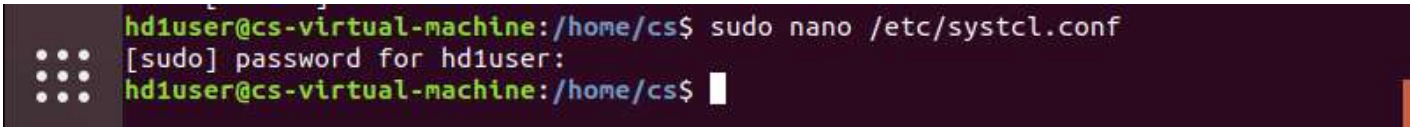


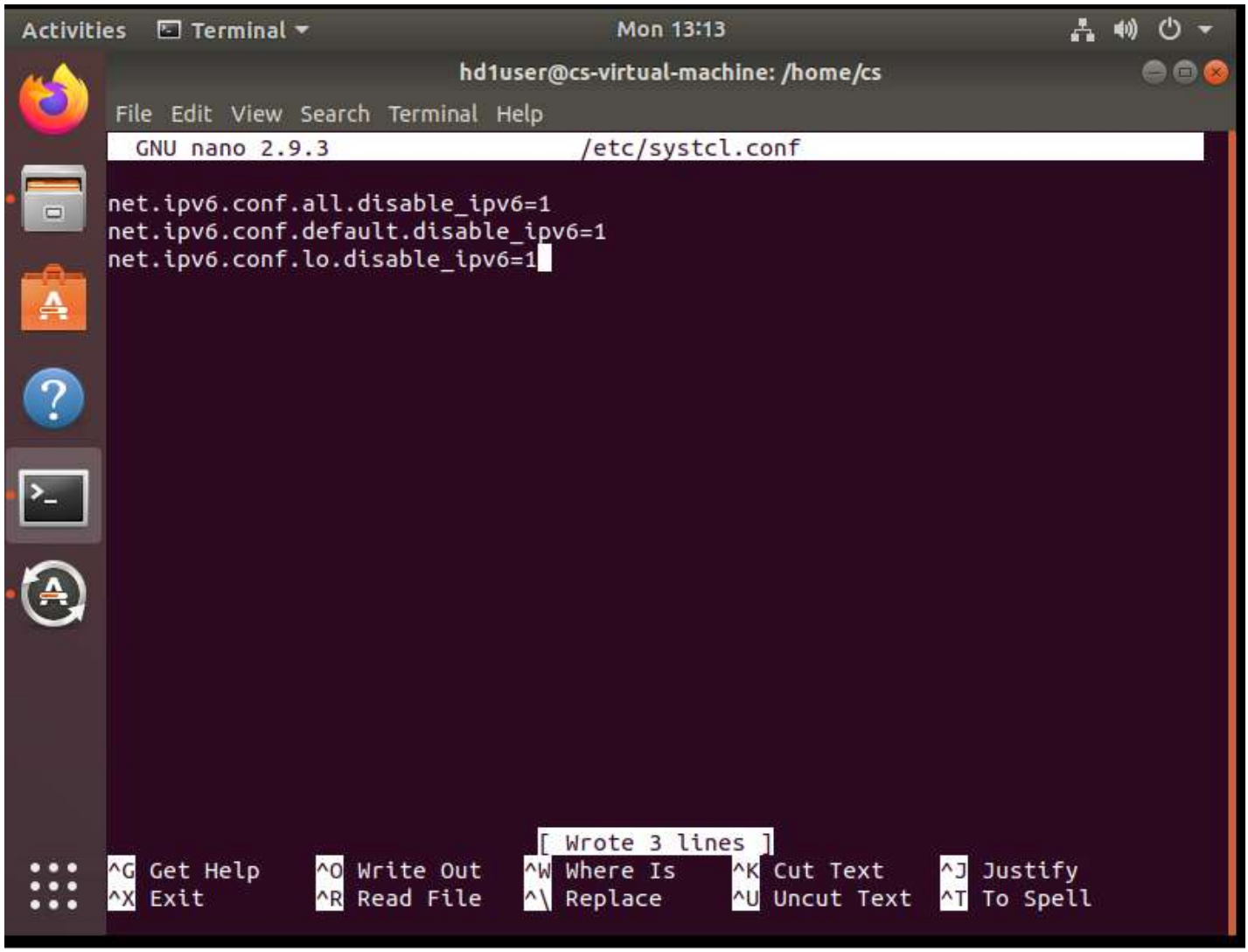
ssh-keygen -t rsa -P ""
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys



2. Now you need to disable IPv6. Open the /etc/sysctl.conf file and add the following lines to the end of the file and save it. (One way of opening the file is sudo nano /etc/sysctl.conf, after you add the lines you need to press Ctrl+X, Shift Y and Enter)

net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1

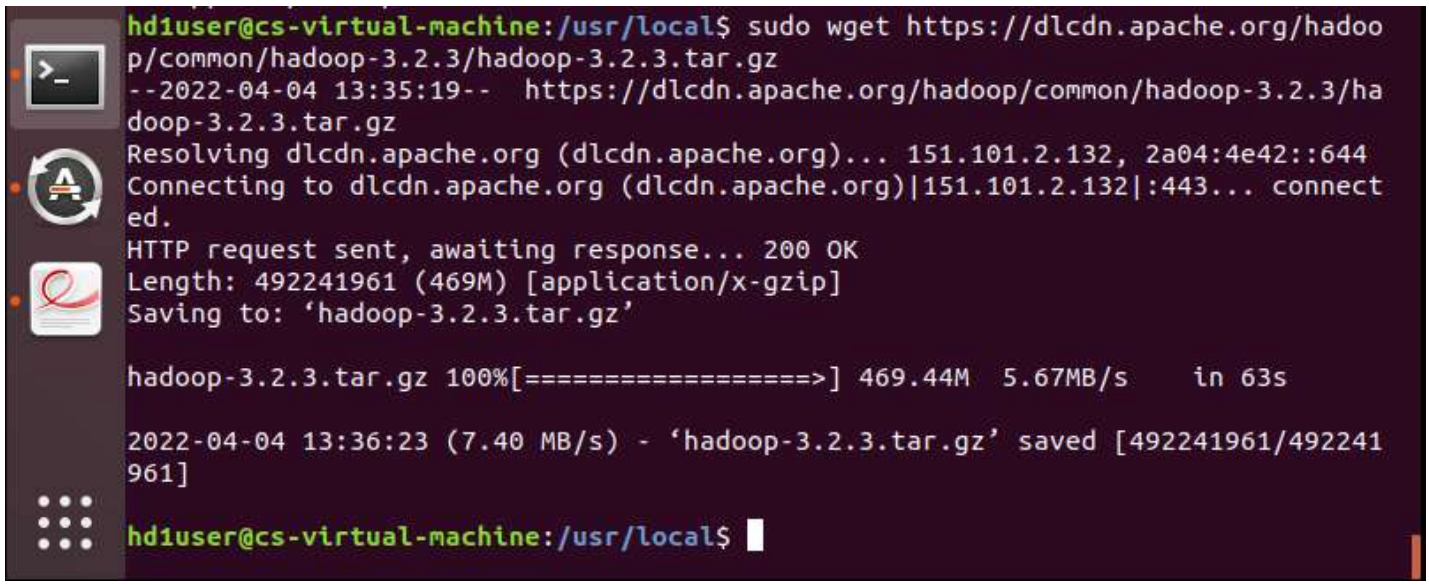




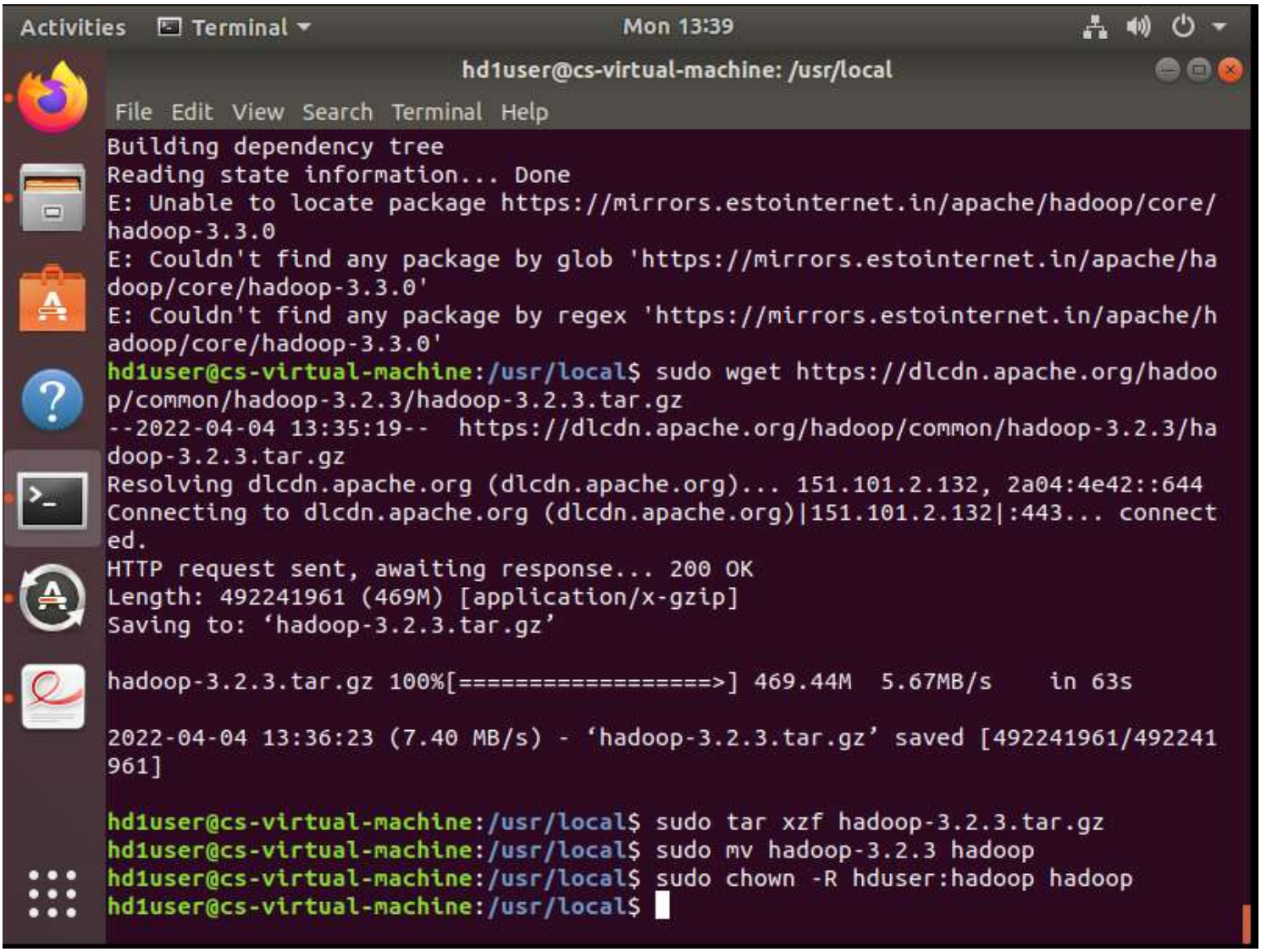
3. Now we download hadoop

cd /usr/local

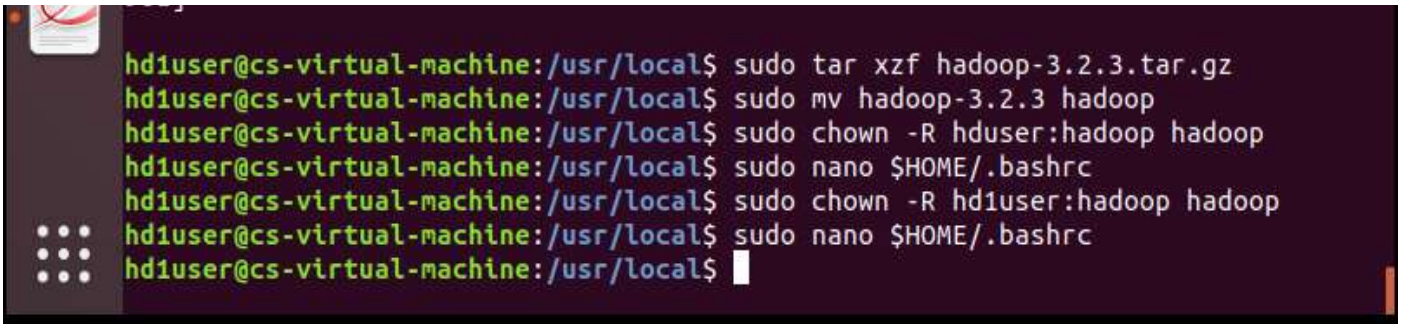
Wget https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.gz

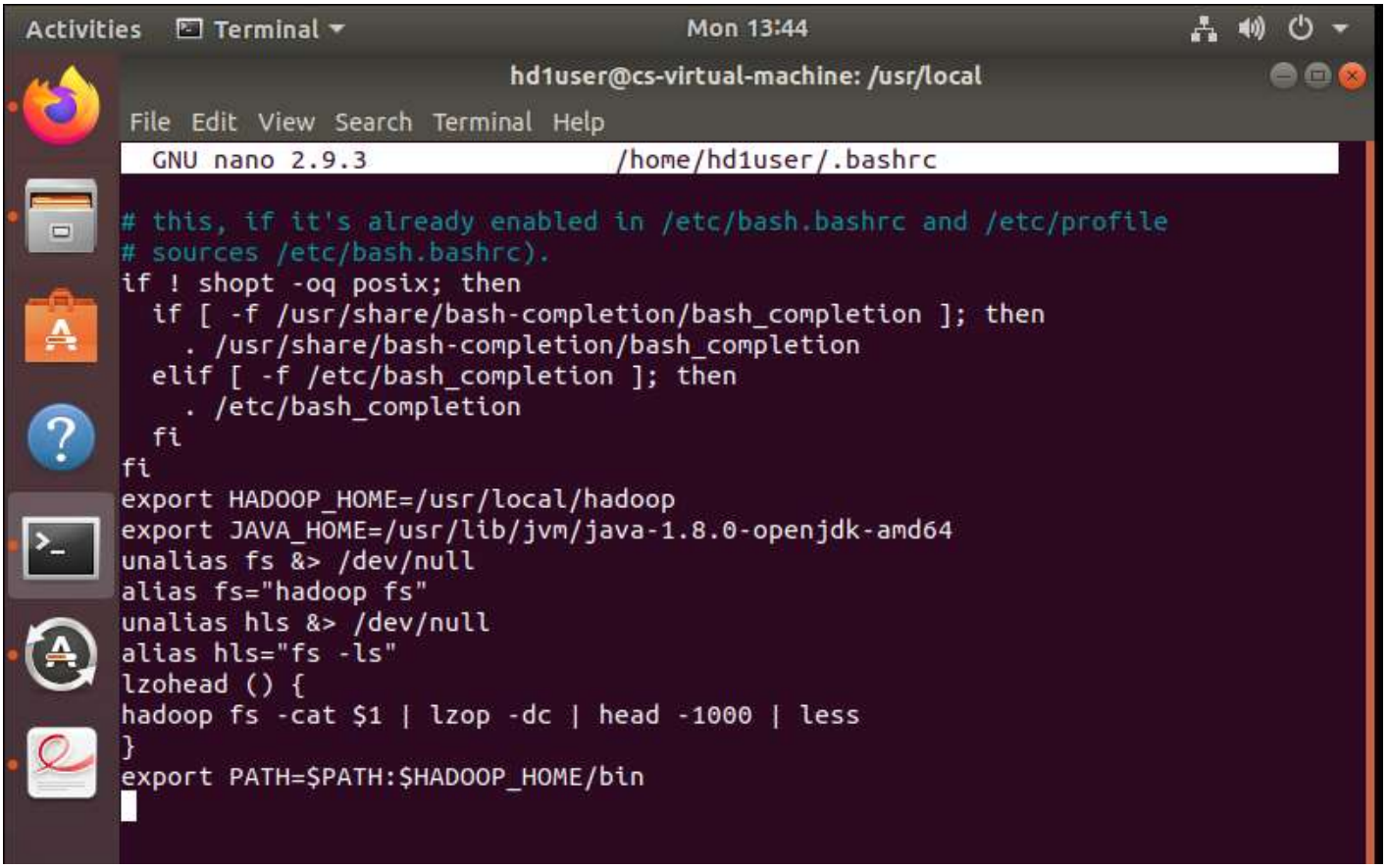


sudo tar xzf hadoop-3.3.0.tar.gz
sudo mv hadoop-3.3.0 hadoop
sudo chown -R hduser:hadoop hadoop

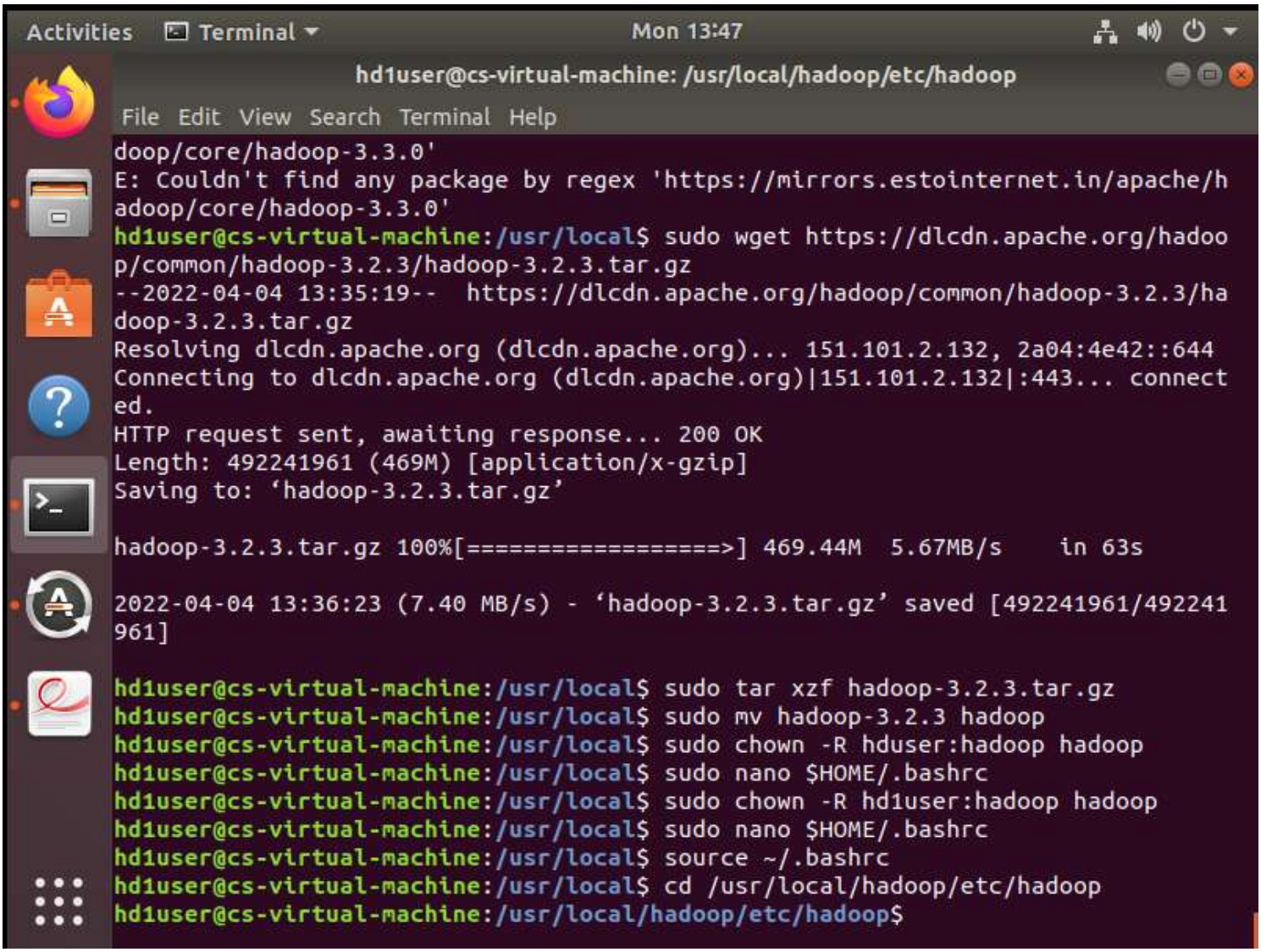


4. Now open $HOME/.bashrc and add the following lines:
export HADOOP_HOME=/usr/local/hadoop
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
unalias fs &> /dev/null
alias fs="hadoop fs"
unalias hls &> /dev/null
alias hls="fs -ls"
lzohead () {
 hadoop fs -cat $1 | lzop -dc | head -1000 | less
 }
export PATH=$PATH:$HADOOP_HOME/bin

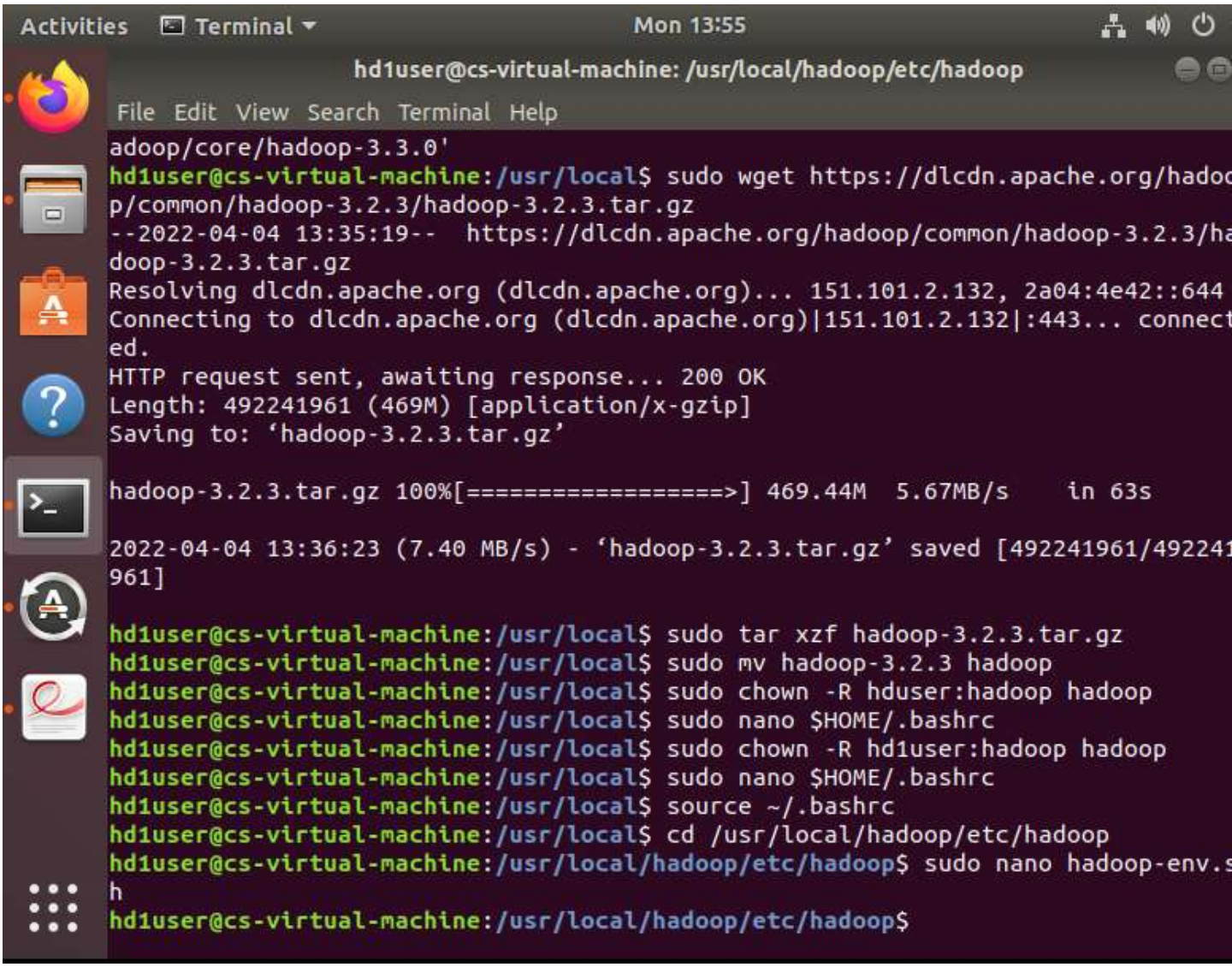


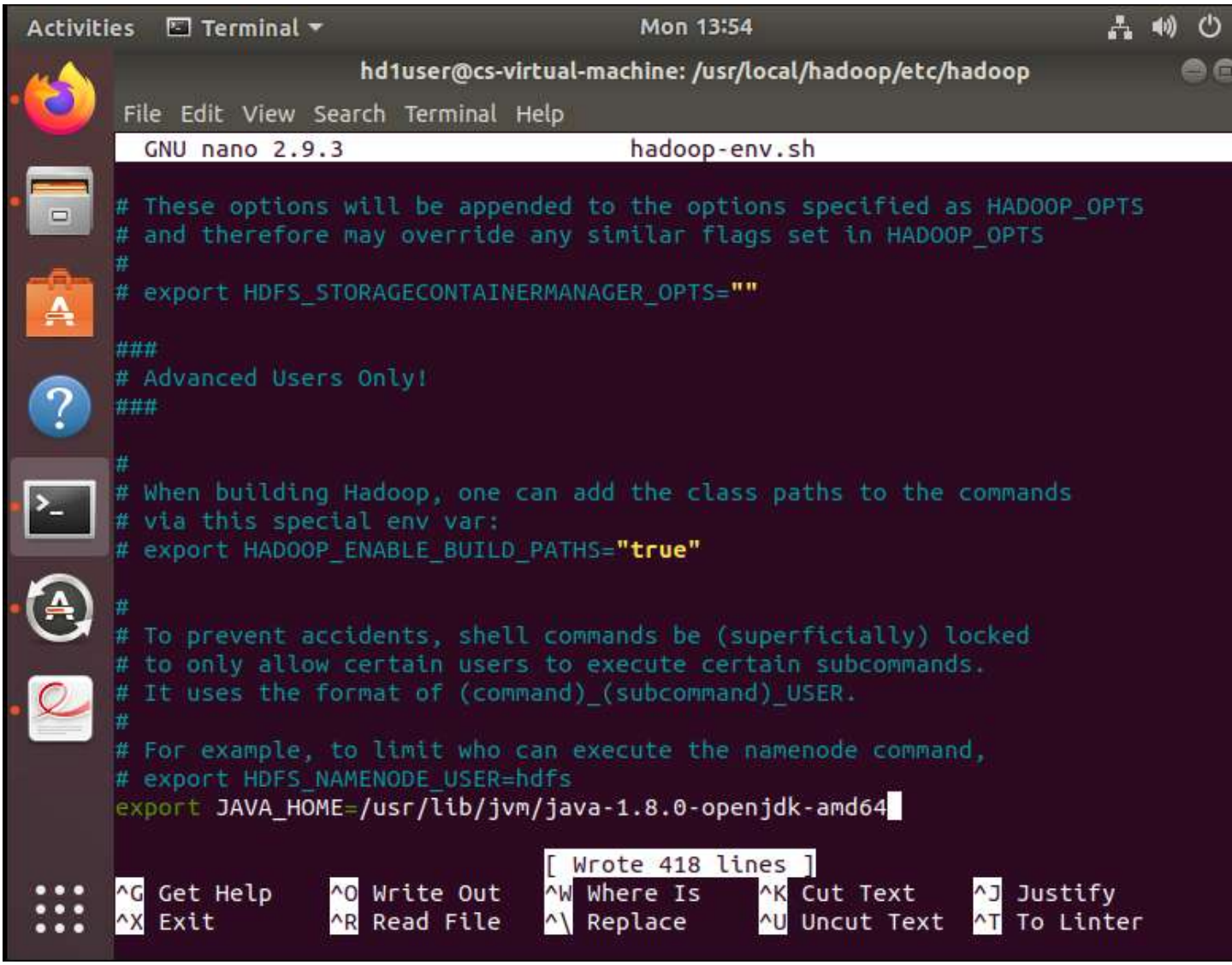


5. Enter following commands:
source ~/.bashrc
cd /usr/local/hadoop/etc/hadoop

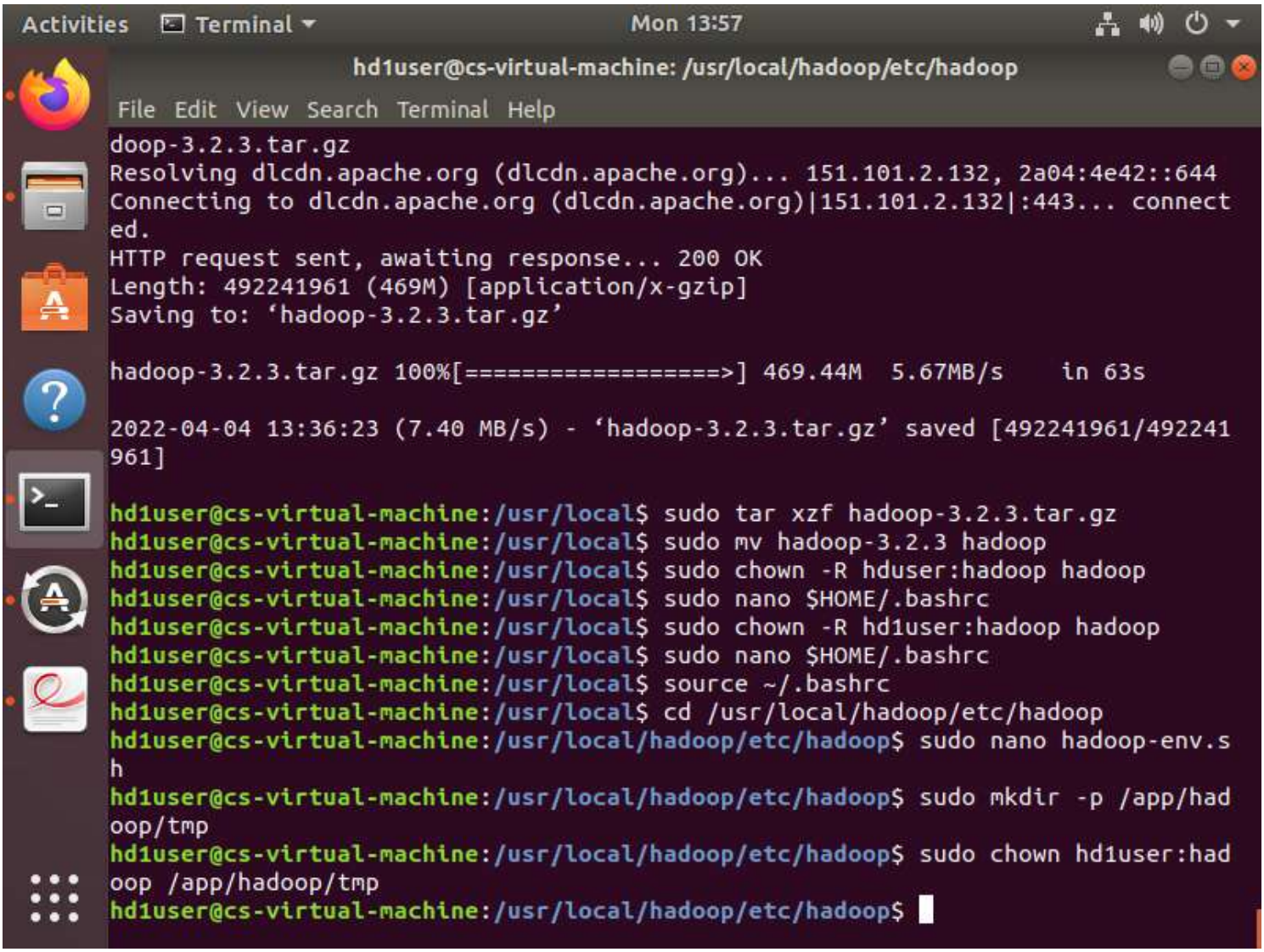


6. Add the following line to hadoop-env.sh
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64


Activities
Terminal
Mon 13:55
hd1user@cs-virtual-machine: /usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
adoop/core/hadoop-3.3.0'
hd1user@cs-virtual-machine:/usr/local$ sudo wget https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.gz
--2022-04-04 13:35:19-- https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.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: 492241961 (469M) [application/x-gzip]
Saving to: 'hadoop-3.2.3.tar.gz'
hadoop-3.2.3.tar.gz 100%[==========>] 469.44M 5.67MB/s in 63s
2022-04-04 13:36:23 (7.40 MB/s) - 'hadoop-3.2.3.tar.gz' saved [492241961/492241961]
hd1user@cs-virtual-machine:/usr/local$ sudo tar xzf hadoop-3.2.3.tar.gz
hd1user@cs-virtual-machine:/usr/local$ sudo mv hadoop-3.2.3 hadoop
hd1user@cs-virtual-machine:/usr/local$ sudo chown -R hduser:hadoop hadoop
hd1user@cs-virtual-machine:/usr/local$ sudo nano $HOME/.bashrc
hd1user@cs-virtual-machine:/usr/local$ sudo chown -R hd1user:hadoop hadoop
hd1user@cs-virtual-machine:/usr/local$ sudo nano $HOME/.bashrc
hd1user@cs-virtual-machine:/usr/local$ source ~/.bashrc
hd1user@cs-virtual-machine:/usr/local$ cd /usr/local/hadoop/etc/hadoop
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.sh
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$



7. Run the following commands:
sudo mkdir -p /app/hadoop/tmp
sudo chown hduser:hadoop /app/hadoop/tmp

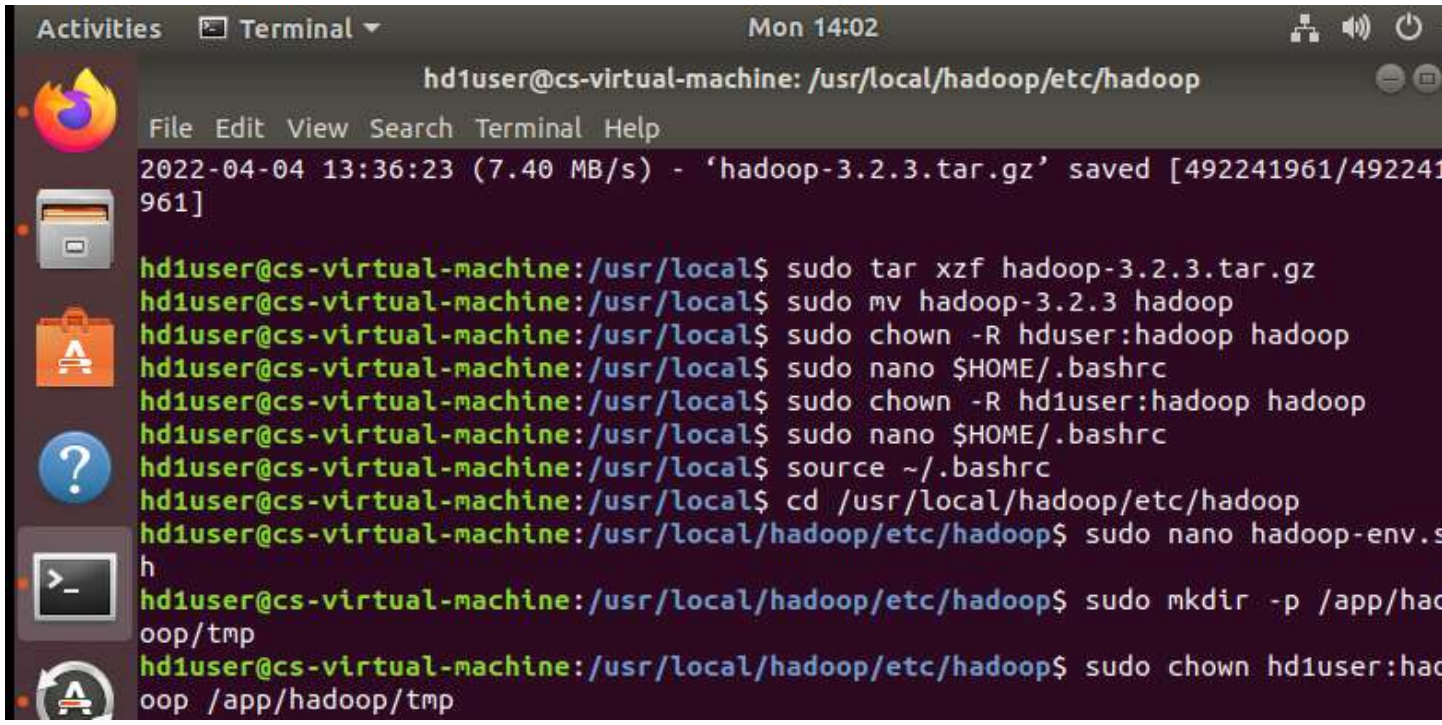


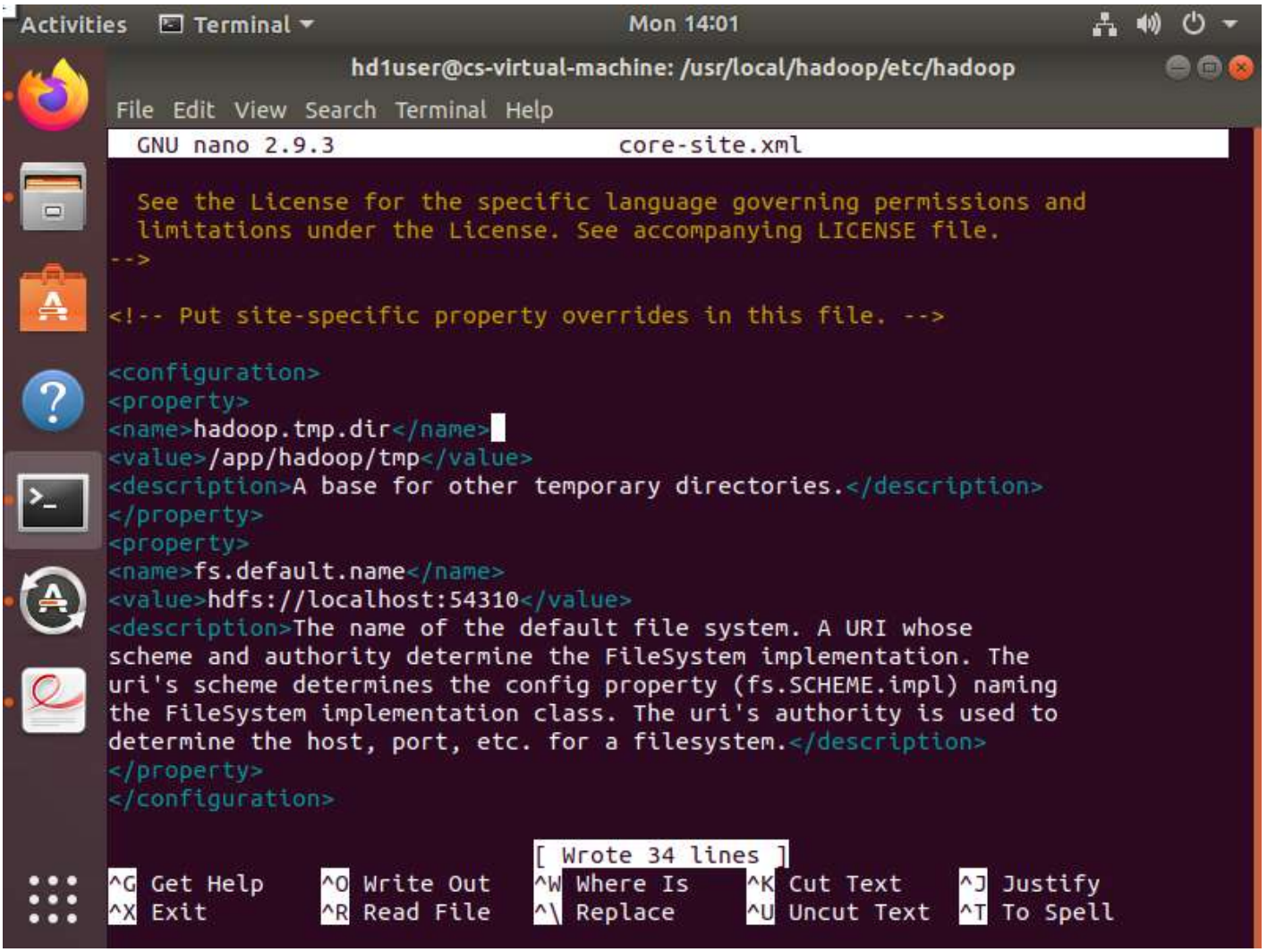
8. Make the following changes in core-site.xml (this file is present in /usr/local/hadoop/etc/hadoop)
Add the following between <configuration> and </configuration>

```
<property>
 <name>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
 <description>A base for other temporary directories.</description>
</property>

<property>
 <name>fs.default.name</name>
 <value>hdfs://localhost:54310</value>
 <description>The name of the default file system. A URI whose
 scheme and authority determine the FileSystem implementation. The
 uri's scheme determines the config property (fs.SCHEME.impl) naming
 the FileSystem implementation class. The uri's authority is used to
```

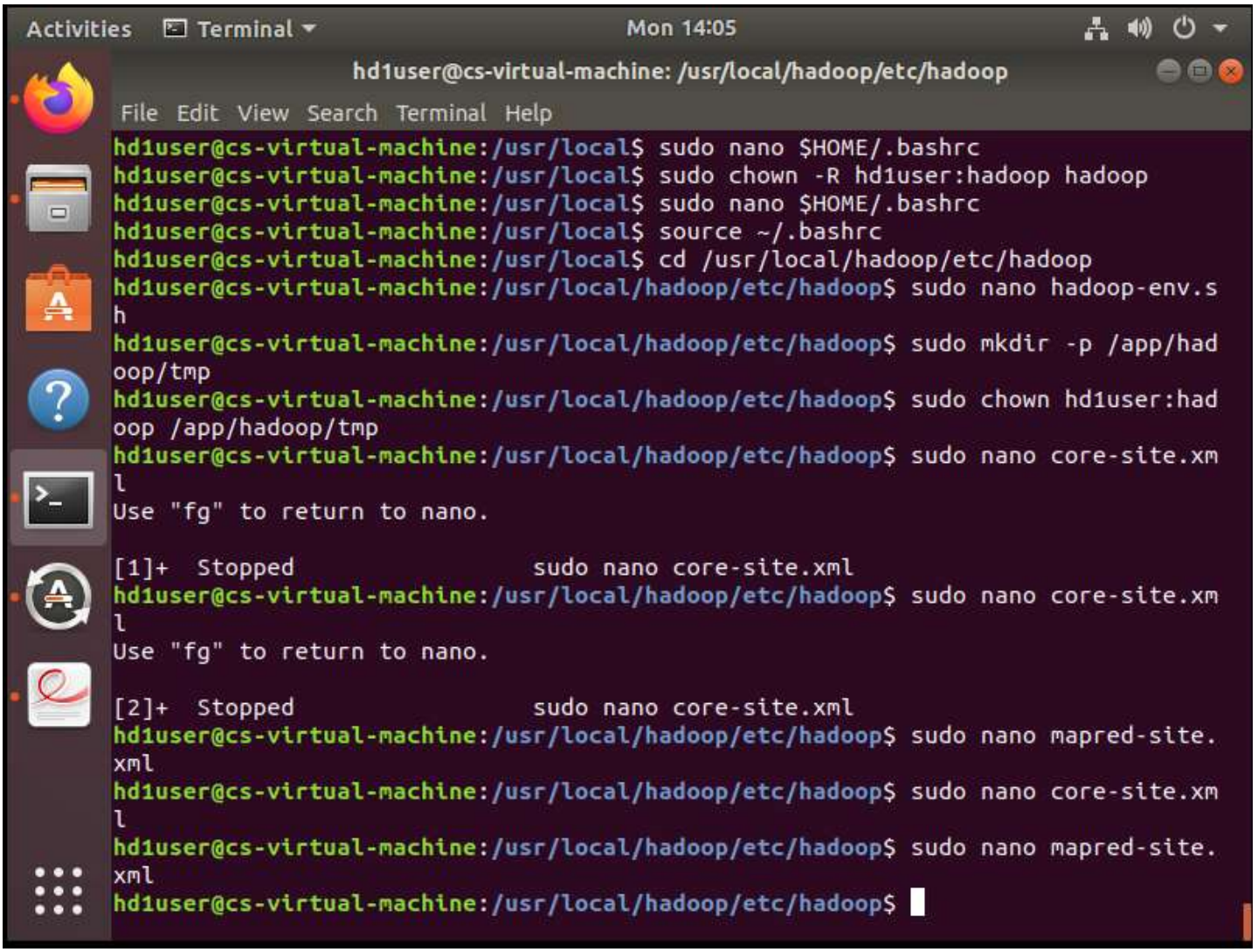
determine the host, port, etc. for a filesystem.</description>
</property>

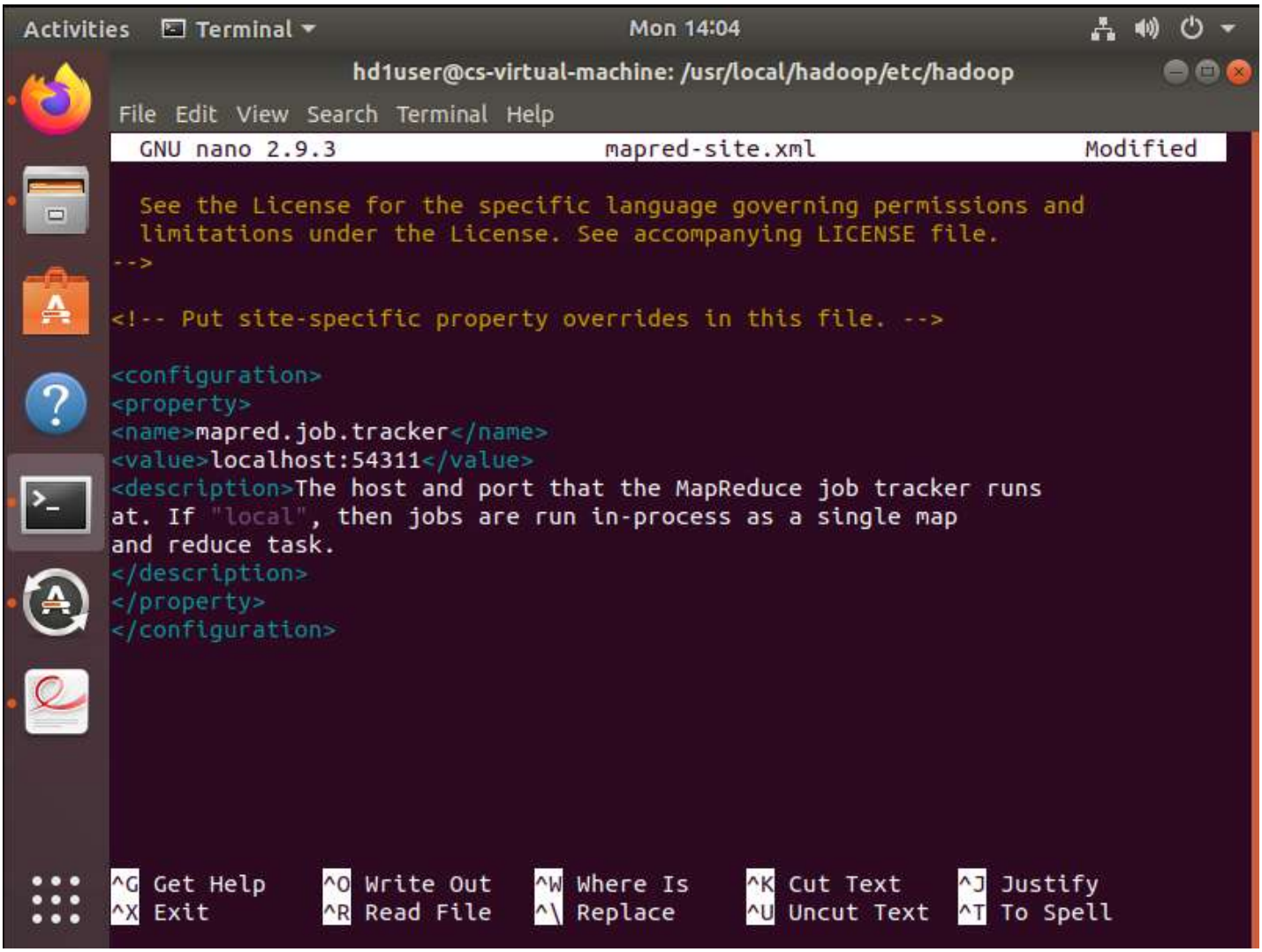




9. In the file mapred-site.xml Add the following between <configuration> and </configuration>

```
<property>
 <name>mapred.job.tracker</name>
 <value>localhost:54311</value>
 <description>The host and port that the MapReduce job tracker runs
 at. If "local", then jobs are run in-process as a single map
 and reduce task.
 </description>
</property>
```

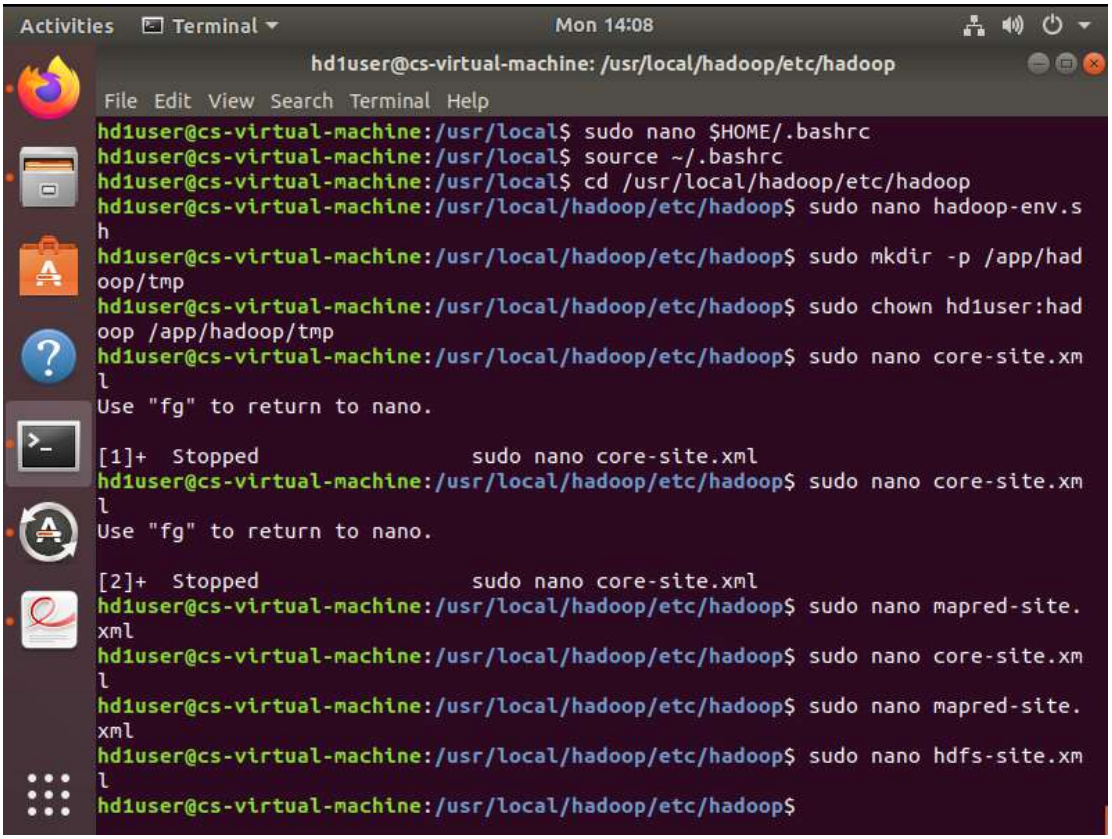

Activities
Terminal
Mon 14:05
hd1user@cs-virtual-machine: /usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
hd1user@cs-virtual-machine:/usr/local$ sudo nano $HOME/.bashrc
hd1user@cs-virtual-machine:/usr/local$ sudo chown -R hd1user:hadoop hadoop
hd1user@cs-virtual-machine:/usr/local$ sudo nano $HOME/.bashrc
hd1user@cs-virtual-machine:/usr/local$ source ~/.bashrc
hd1user@cs-virtual-machine:/usr/local$ cd /usr/local/hadoop/etc/hadoop
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.s
h
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo mkdir -p /app/had
oop/tmp
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo chown hd1user:had
oop /app/hadoop/tmp
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
Use "fg" to return to nano.
[1]+ Stopped sudo nano core-site.xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
Use "fg" to return to nano.
[2]+ Stopped sudo nano core-site.xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.
xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.
xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$

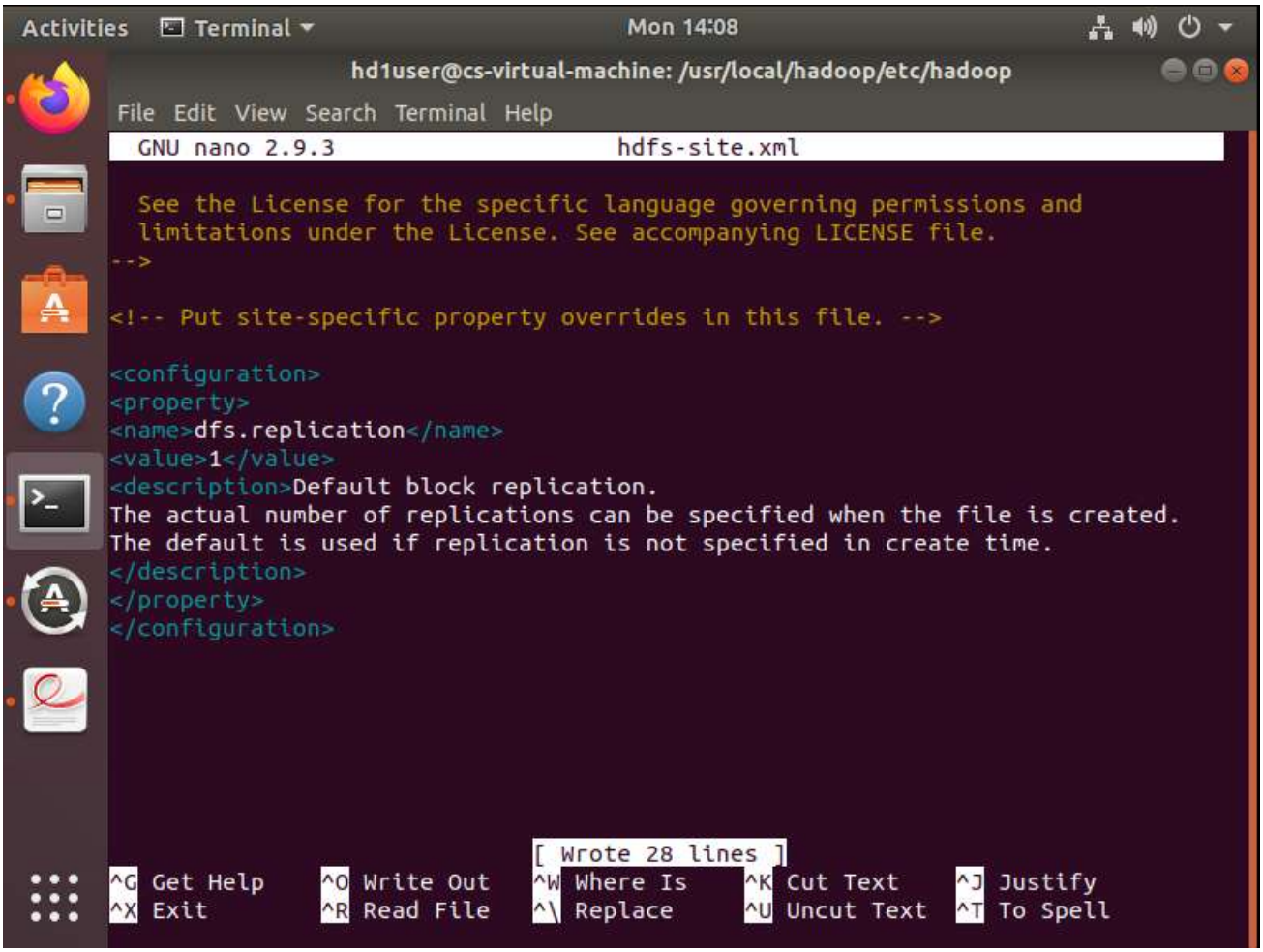


10. In the file hdfs-site.xml Add the following between <configuration> and </configuration>

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

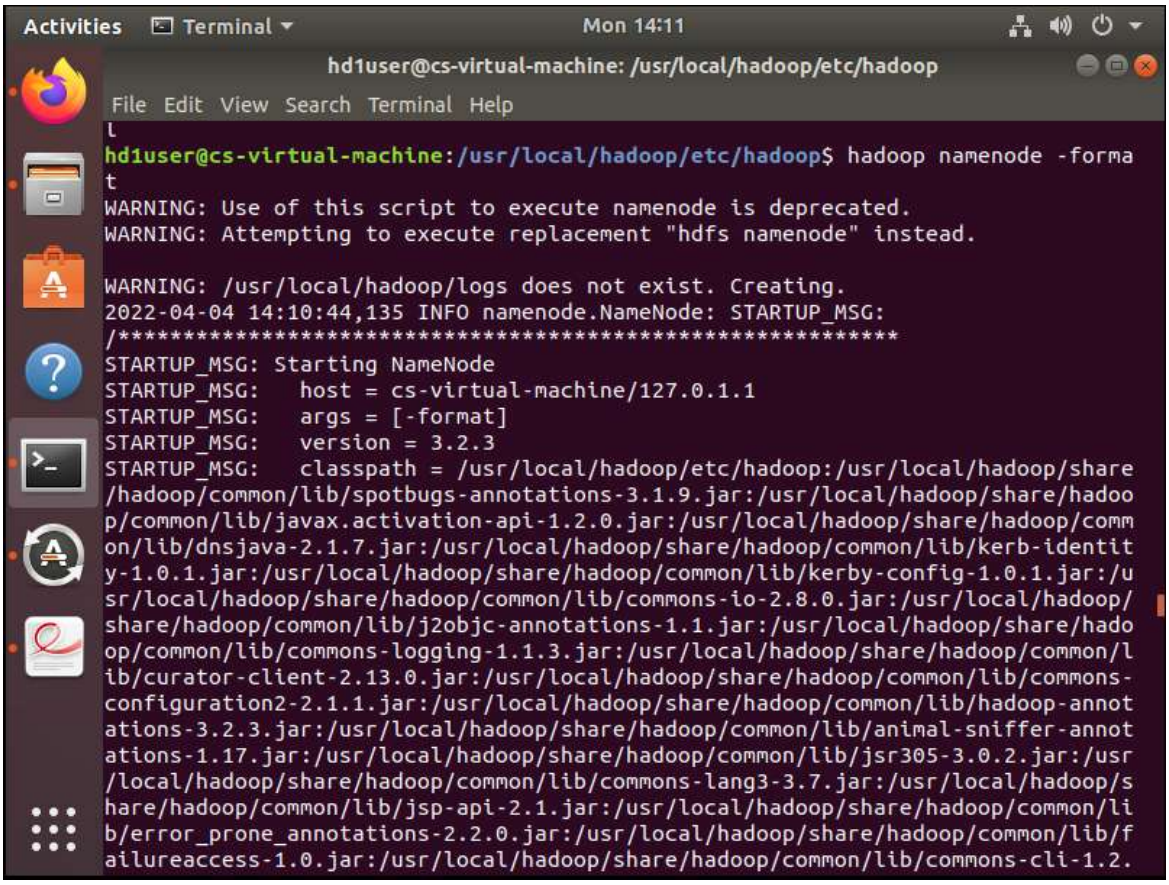
<description>Default block replication.
The actual number of replications can be specified when the file is created.
The default is used if replication is not specified in create time.
</description>
</property>

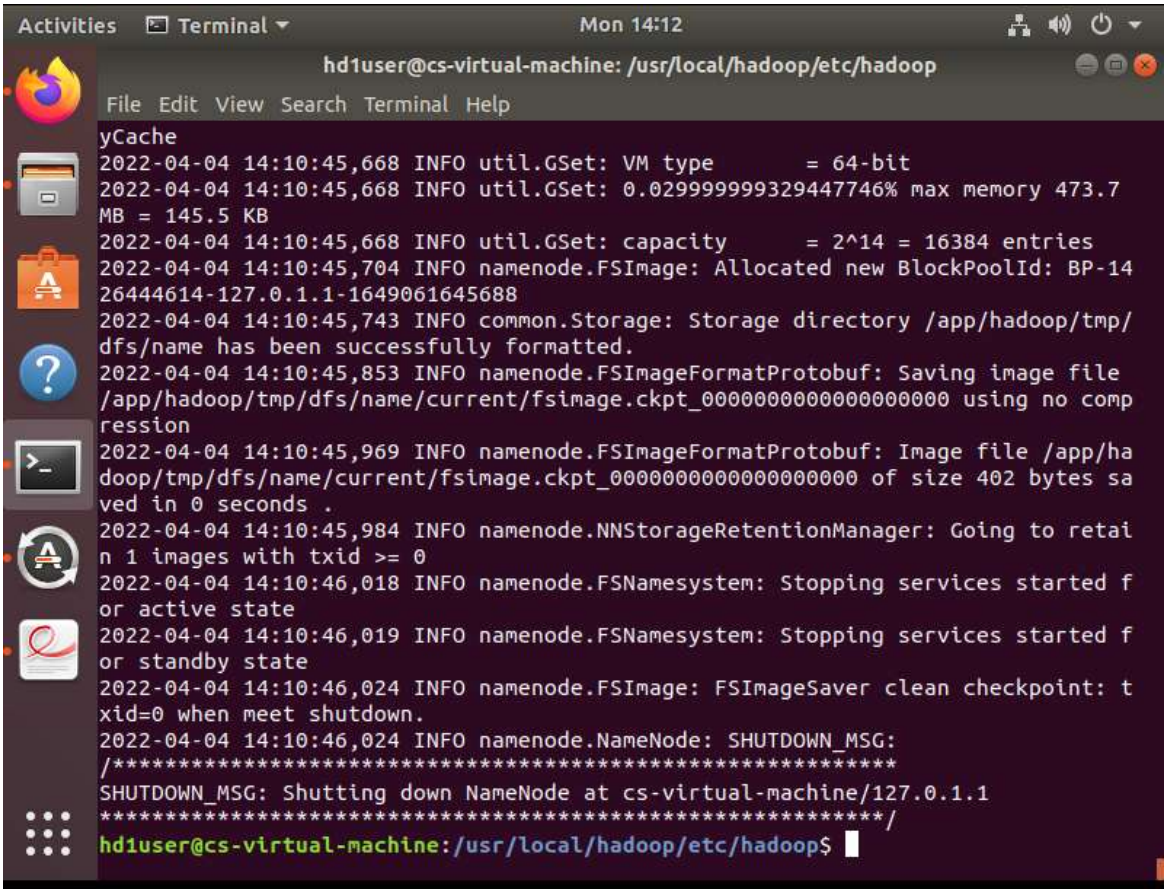

Activities
Terminal
Mon 14:08
hd1user@cs-virtual-machine: /usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
hd1user@cs-virtual-machine:/usr/local$ sudo nano $HOME/.bashrc
hd1user@cs-virtual-machine:/usr/local$ source ~/.bashrc
hd1user@cs-virtual-machine:/usr/local$ cd /usr/local/hadoop/etc/hadoop
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.s
h
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo mkdir -p /app/had
oop/tmp
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo chown hd1user:had
oop /app/hadoop/tmp
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
Use "fg" to return to nano.
[1]+ Stopped sudo nano core-site.xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
Use "fg" to return to nano.
[2]+ Stopped sudo nano core-site.xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.
xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xm
l
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.
xml
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$ sudo nano hdfs-site.xm
l
hd1user@cs-virtual-machine:/usr/local/hadoop/etc/hadoop$


Activities
Terminal
Mon 14:08
hd1user@cs-virtual-machine: /usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
GNU nano 2.9.3 hdfs-site.xml
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>
<description>Default block replication.
The actual number of replications can be specified when the file is created.
The default is used if replication is not specified in create time.
</description>
</property>
</configuration>
[Wrote 28 lines]
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Spell

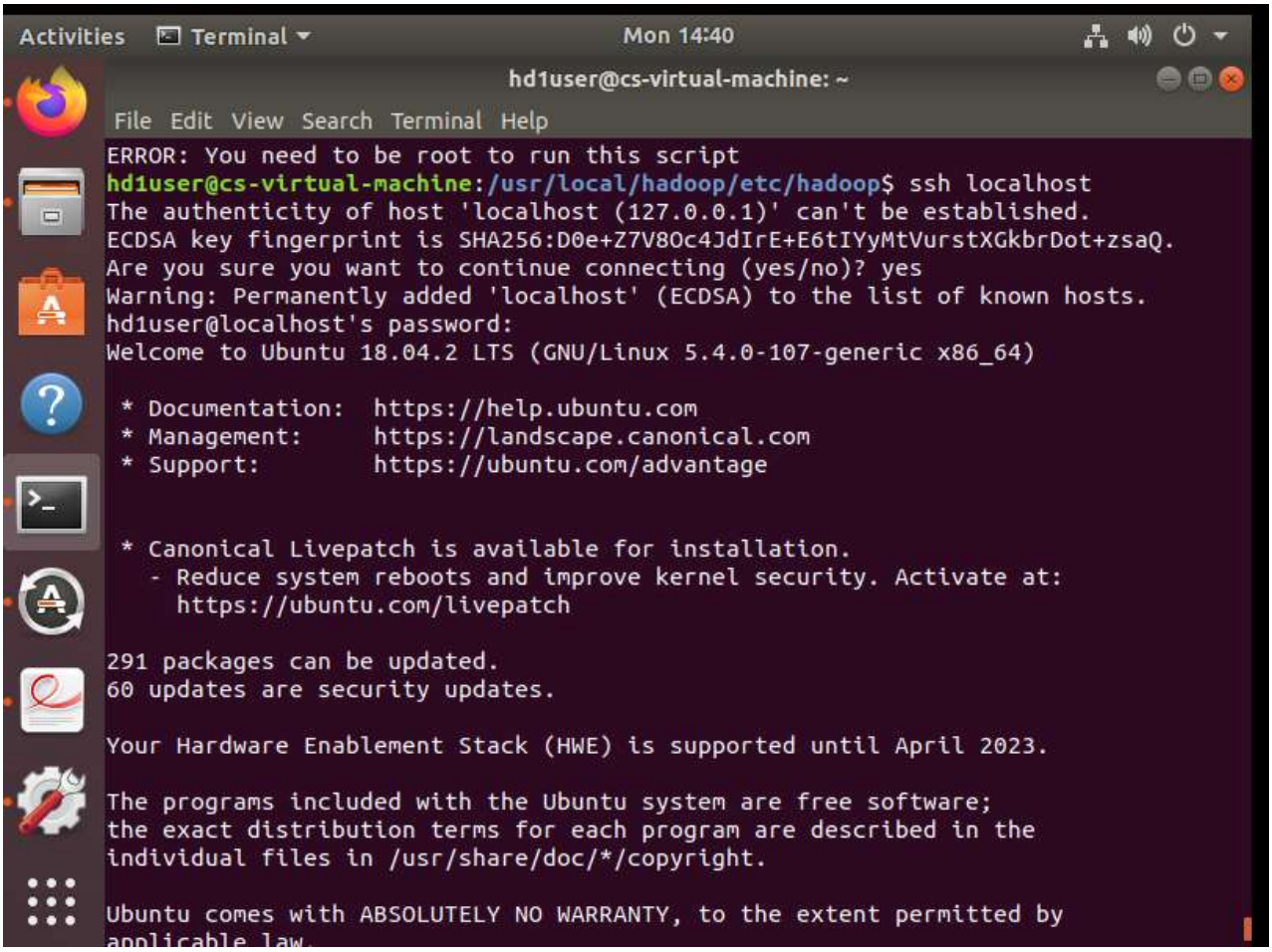
11. Finally we format namenode by the following commands:

hadoop namenode -format





Now to start hadoop, we need to run command: ssh localhost

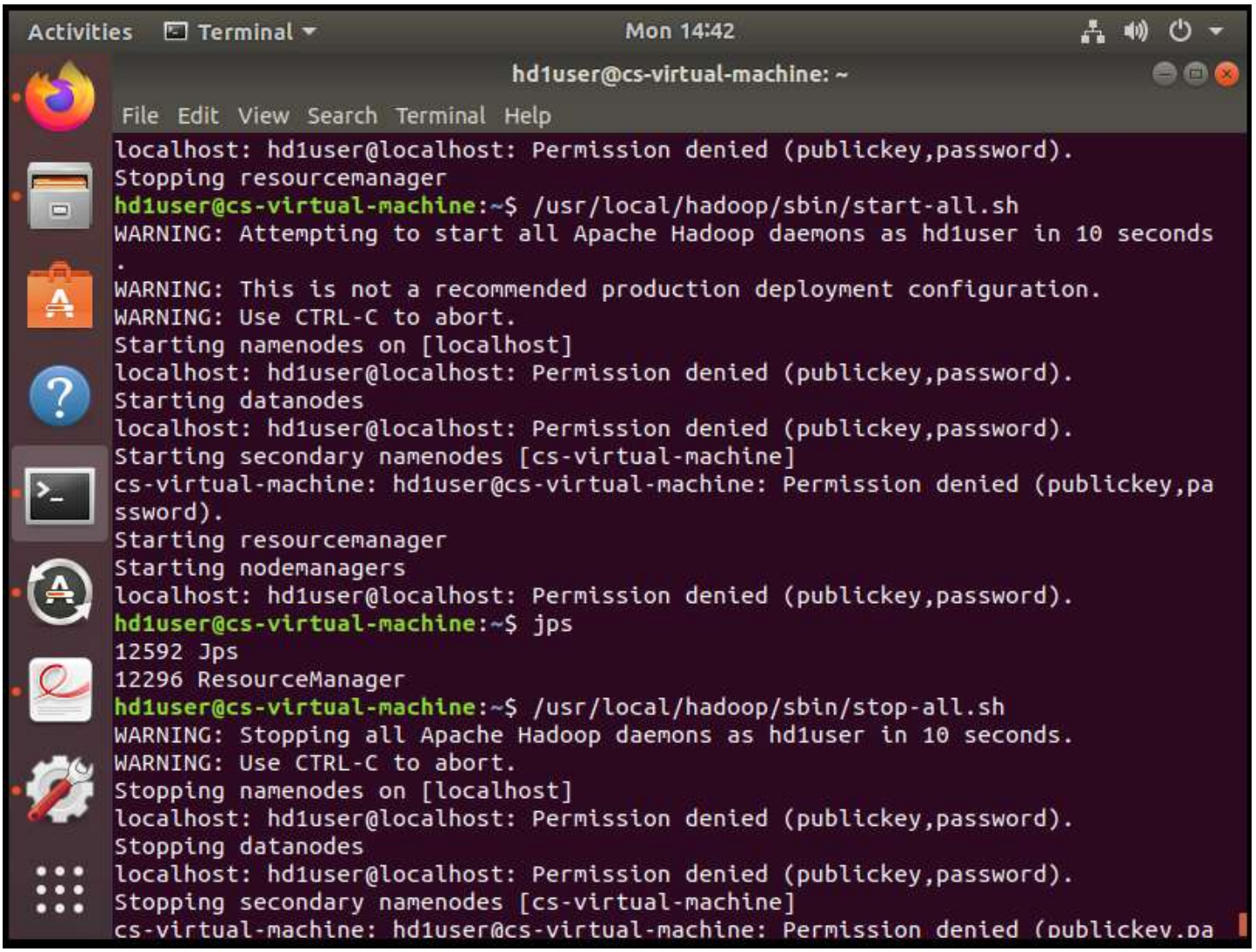


Now to start hadoop, we need to run command:
ssh localhost

This command is to start hadoop services:
/usr/local/hadoop/sbin/start-all.sh

This command is to check that all hadoop services are running (6 services should appear): jps

This command is to stop hadoop services:
/usr/local/hadoop/sbin/stop-all.sh



***Incase if ssh fails
https://linuxhint.com/fix_connection_refused_ubuntu/

$ sudo apt list --installed | grep openssh-server
$ sudo apt install openssh-server
$ sudo apt list --installed | grep openssh-server
$ sudo service ssh status
$ sudo service ssh start
$ sudo service ssh restart
**Incase if all services not starting for hadoop

https://gautambangalore.medium.com/resolve-permission-issue-among-datanodes-with-namenode-to-establish-secure-shell-ssh-without-a-1e76e2863111

$ ssh-keygen -t rsa -P „“ -f ~/.ssh/id_rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys

sudo service network-manager restart
sudo service ssh restart