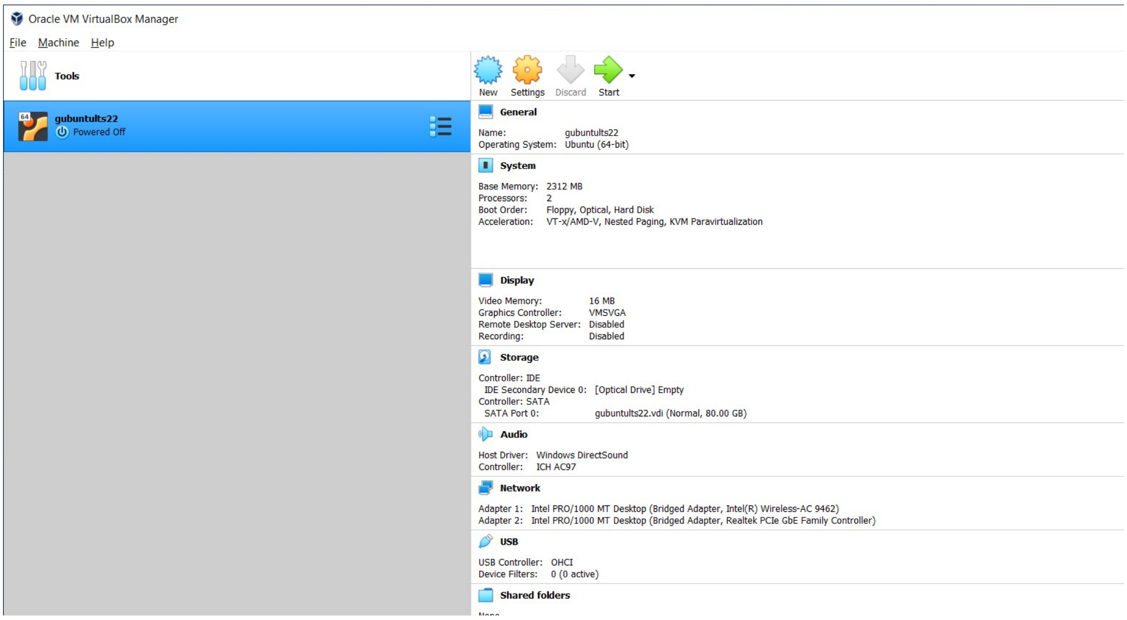
Week 4: Homework 1: MapReduce Program + Full Inverted Index ==> Extra

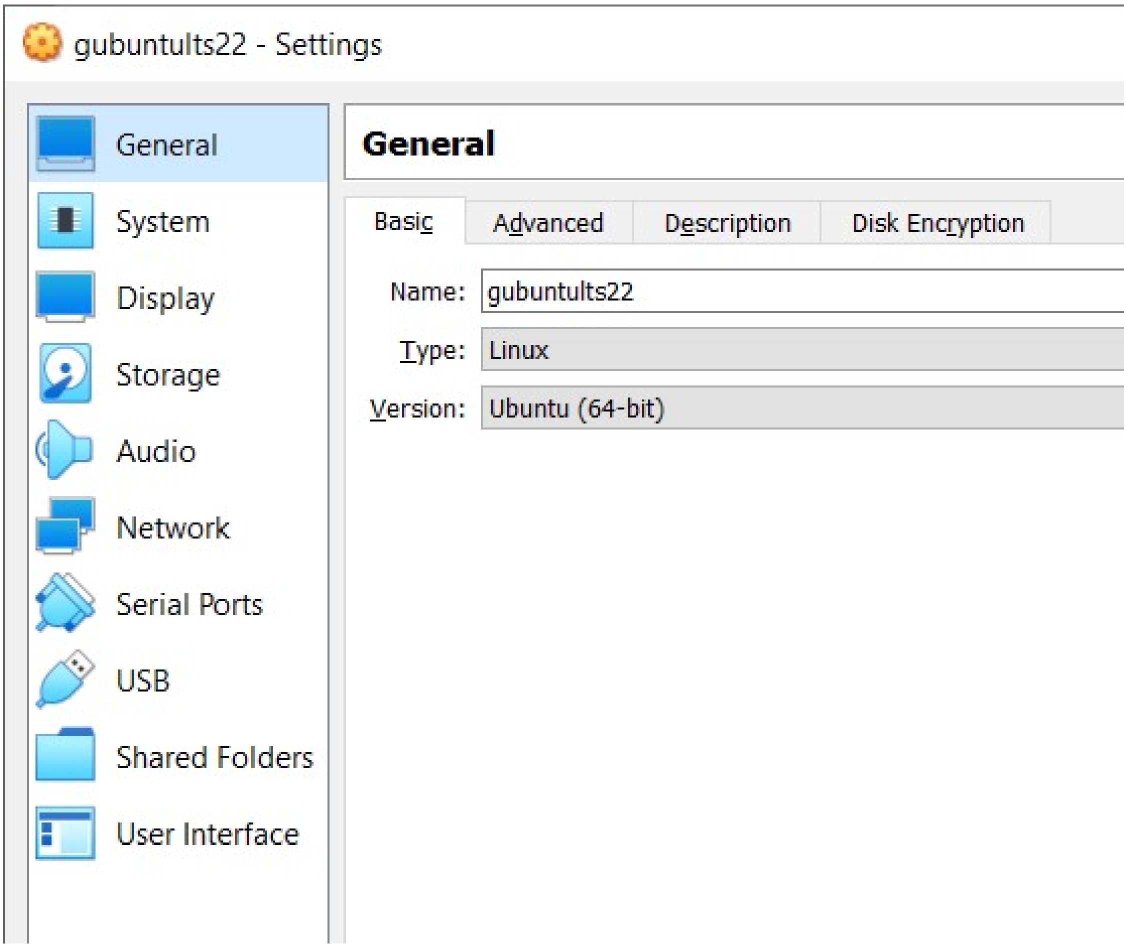
1) Installation of Oracle Virtual Machine
2) Installation Java
3) Installation of Hadoop
4) Configuration of Hadoop
5) Running Inverted index program

MapReduce on Ubuntu on Oracle Virtual Machine local:

Installation Of Oracle Vitual machine on local host



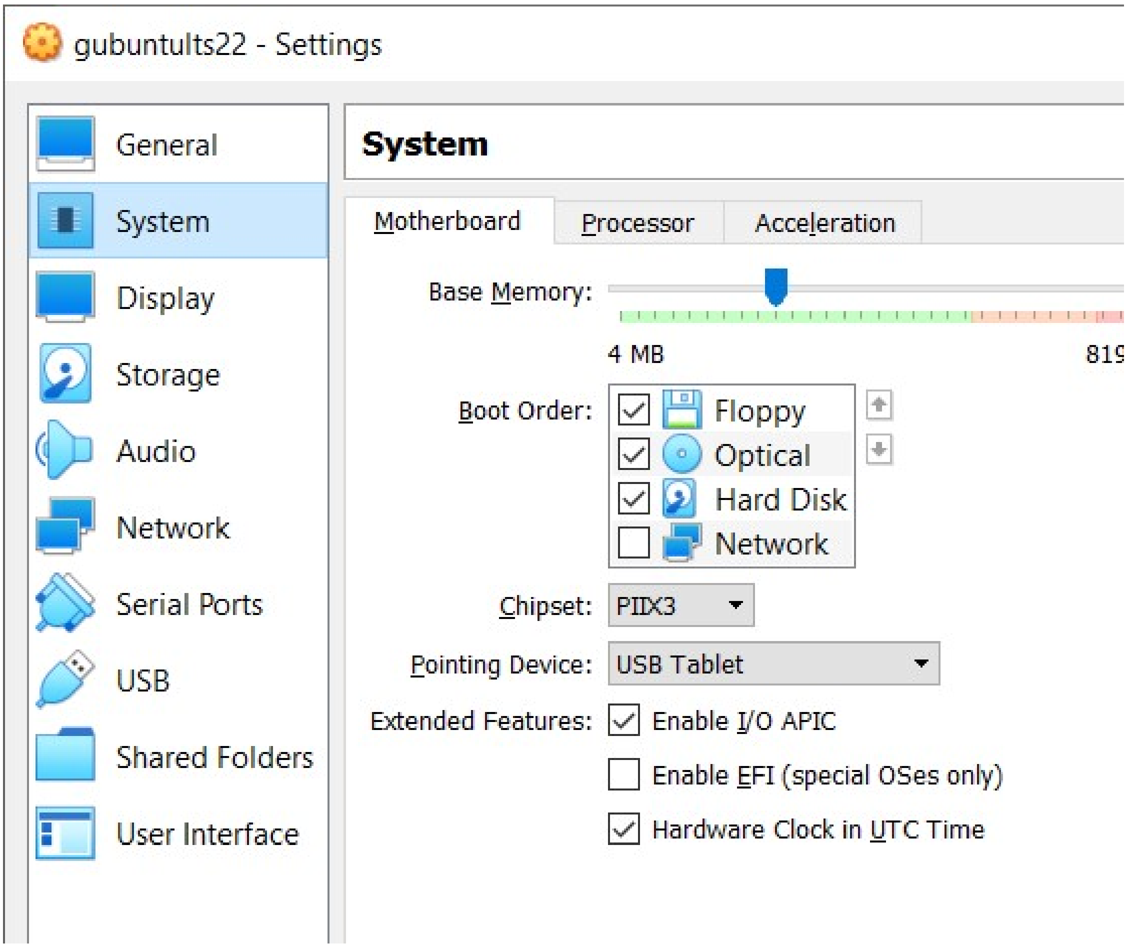
```
General
Name:               gubuntults22
Operating System:   Ubuntu (64-bit)

System
Base Memory:   2312 MB
Processors:    2
Boot Order:    Floppy, Optical, Hard Disk
Acceleration:  VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display
Video Memory:            16 MB
Graphics Controller:     VMSVGA
Remote Desktop Server:   Disabled
Recording:               Disabled

Storage
Controller: IDE
  IDE Secondary Device 0:   [Optical Drive] Empty
Controller: SATA
  SATA Port 0:              gubuntults22.vdi (Normal, 80.00 GB)

Audio
Host Driver:   Windows DirectSound
Controller:    ICH AC97

Network
Adapter 1:   Intel PRO/1000 MT Desktop (Bridged Adapter, Intel(R) Wireless-AC 9462)
Adapter 2:   Intel PRO/1000 MT Desktop (Bridged Adapter, Realtek PCIe GbE Family Controller)

USB
USB Controller:   OHCI
Device Filters:   0 (0 active)

Shared folders
```



Name: gubuntults22

Type: Linux

Version: Ubuntu (64-bit)



gubuntults22 - Settings

System

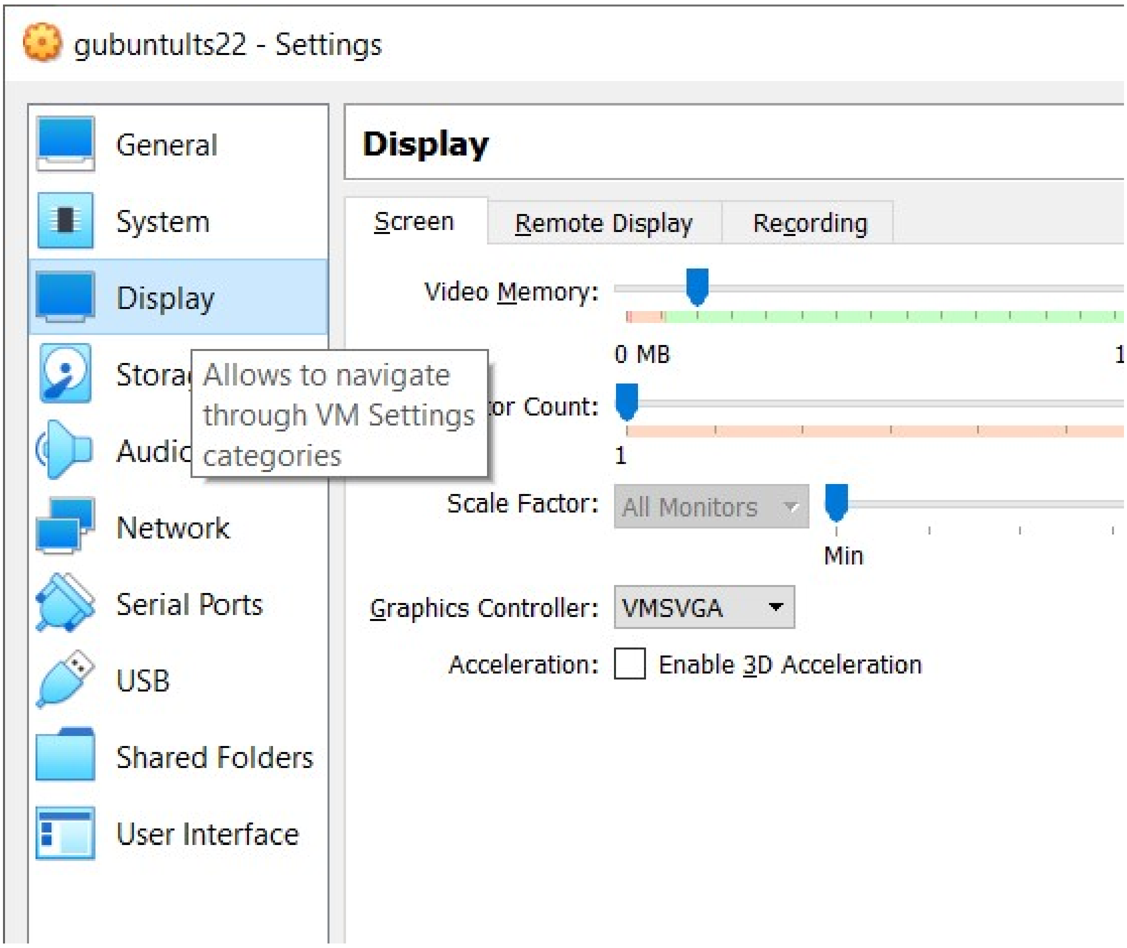
Base Memory: 4 MB 819

Boot Order:
☑ Floppy
☑ Optical
☑ Hard Disk
□ Network

Chipset: PIIX3

Pointing Device: USB Tablet

Extended Features:
☑ Enable I/O APIC
□ Enable EFI (special OSes only)
☑ Hardware Clock in UTC Time



gubuntults22 - Settings

Display

Video Memory: 0 MB

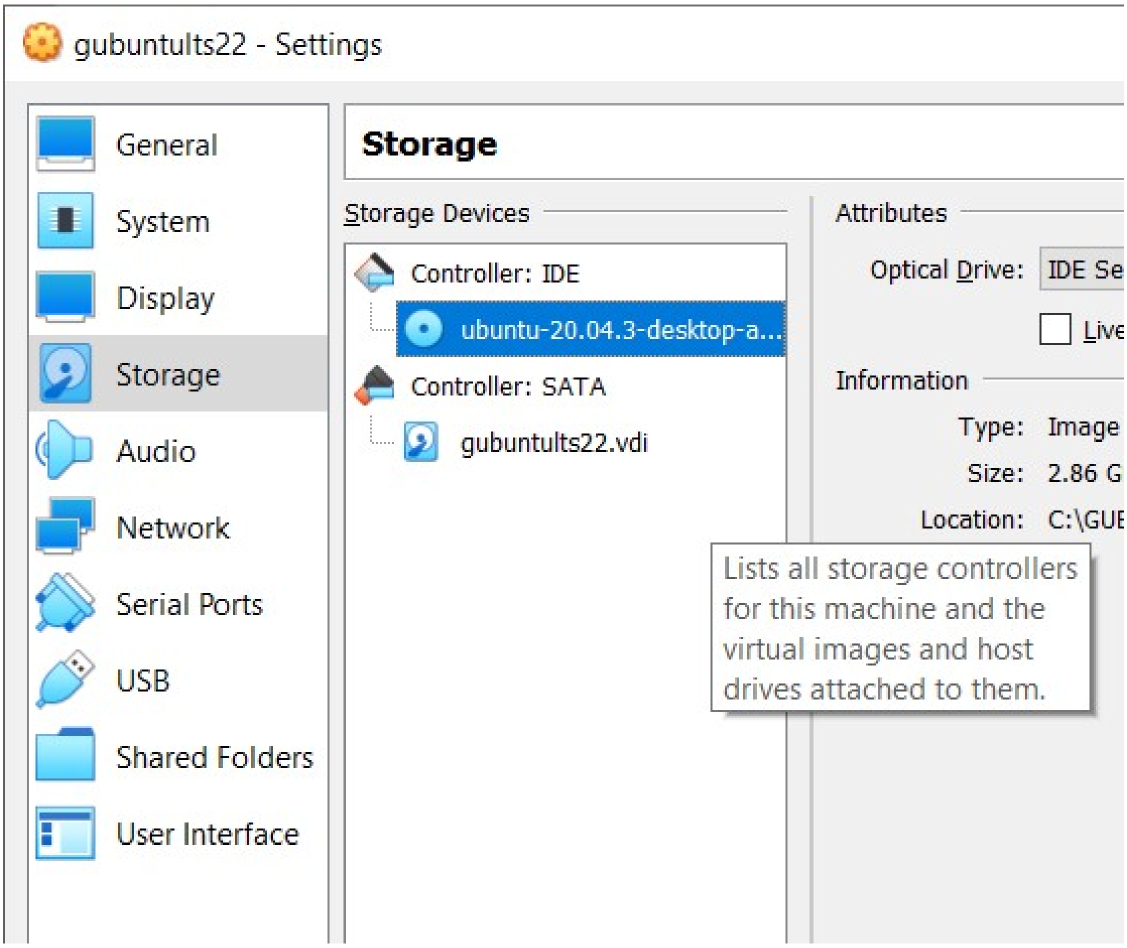
or Count: 1

Scale Factor: All Monitors, Min

Graphics Controller: VMSVGA

Acceleration: □ Enable 3D Acceleration

Allows to navigate through VM Settings categories



gubuntults22 - Settings

Storage

Storage Devices

- Controller: IDE
  - ubuntu-20.04.3-desktop-a...
- Controller: SATA
  - gubuntults22.vdi

Attributes

Optical Drive: IDE Se

□ Live

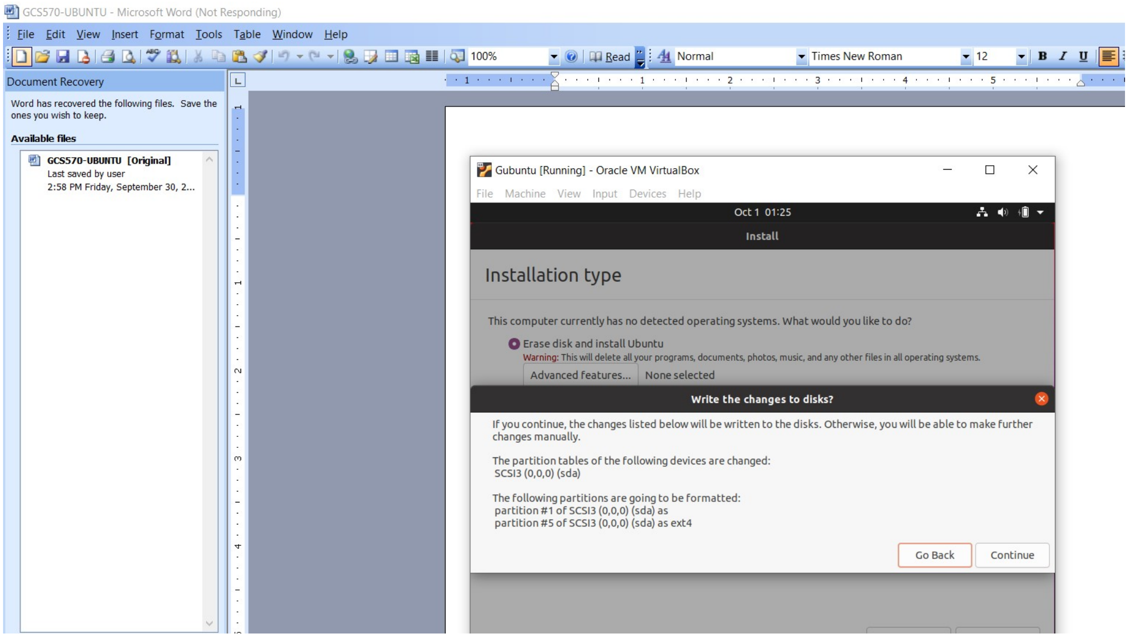
Information

Type: Image

Size: 2.86 G

Location: C:\GUE

Lists all storage controllers for this machine and the virtual images and host drives attached to them.


GCS570-UBUNTU - Microsoft Word (Not Responding)
File Edit View Insert Format Tools Table Window Help
100%
Read
Normal
Times New Roman
12
Document Recovery
Available files
GCS570-UBUNTU [Original]
Last saved by user
2:58 PM Friday, September 30, 2...
Gubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Oct 1 01:25
Install
Go Back
Continue

Word has recovered the following files. Save the ones you wish to keep.

Installation type

This computer currently has no detected operating systems. What would you like to do?

● Erase disk and install Ubuntu

Warning: This will delete all your programs, documents, photos, music, and any other files in all operating systems.

Advanced features... None selected

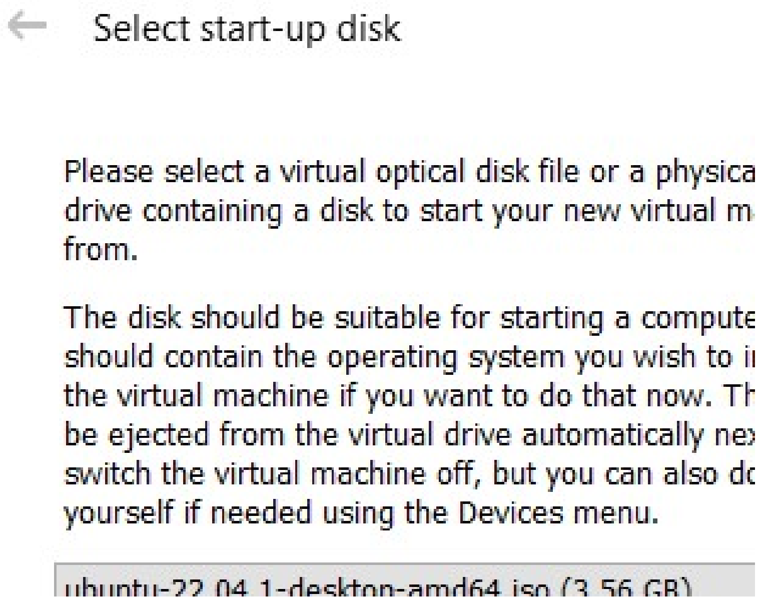
Write the changes to disks?

If you continue, the changes listed below will be written to the disks. Otherwise, you will be able to make further changes manually.

The partition tables of the following devices are changed:
SCSI3 (0,0,0) (sda)

The following partitions are going to be formatted:
partition #1 of SCSI3 (0,0,0) (sda) as
partition #5 of SCSI3 (0,0,0) (sda) as ext4

?


Select start-up disk

Please select a virtual optical disk file or a physica
drive containing a disk to start your new virtual m
from.

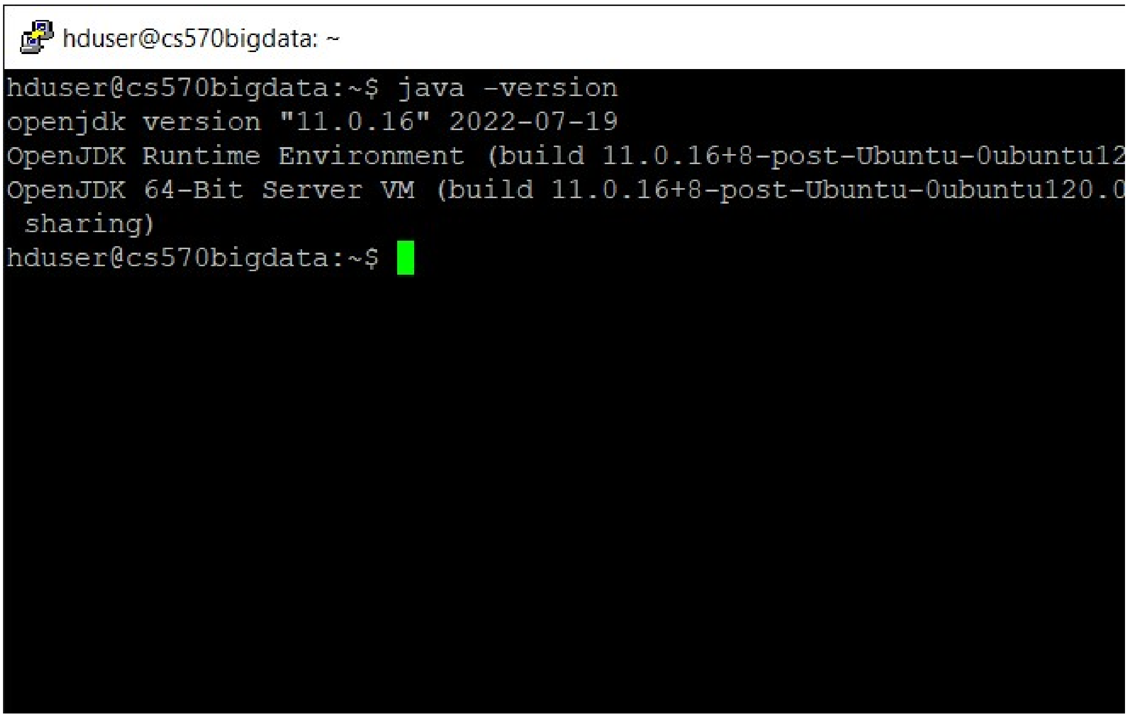
The disk should be suitable for starting a compute
should contain the operating system you wish to i
the virtual machine if you want to do that now. Th
be ejected from the virtual drive automatically ne
switch the virtual machine off, but you can also dc
yourself if needed using the Devices menu.

ubuntu-22.04.1-desktop-amd64.iso (3.56 GB)

After completion of installation Oracle VM , install Java , Hadoop and the script files
Download hadoop
Download Java
Extract hadoop

Java installation steps

sudo apt-get install openjdk-11-jre
sudo apt-get install openjdk-11-jdk
java -version



hduser@cs570bigdata: ~

```
hduser@cs570bigdata:~$ java -version
openjdk version "11.0.16" 2022-07-19
OpenJDK Runtime Environment (build 11.0.16+8-post-Ubuntu-0ubuntu12
OpenJDK 64-Bit Server VM (build 11.0.16+8-post-Ubuntu-0ubuntu120.0
 sharing)
hduser@cs570bigdata:~$
```

Setup hadoop user for Hadoop Installation

sudo addgroup hadoop

sudo adduser --ingroup hadoop hduser

sudo su hduser

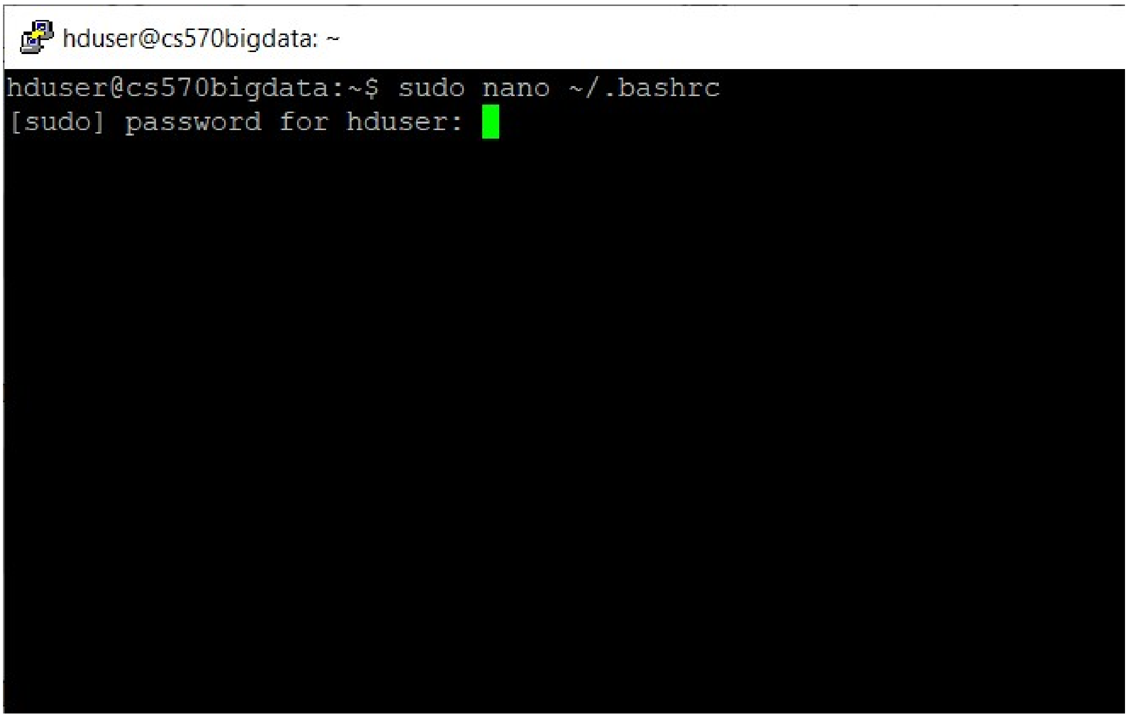
For this I used hadoop version hadoop-2.10.2.tar.gz

Sudo tar vxzf hadoop-2.10.2.tar.gz –C /usr/local

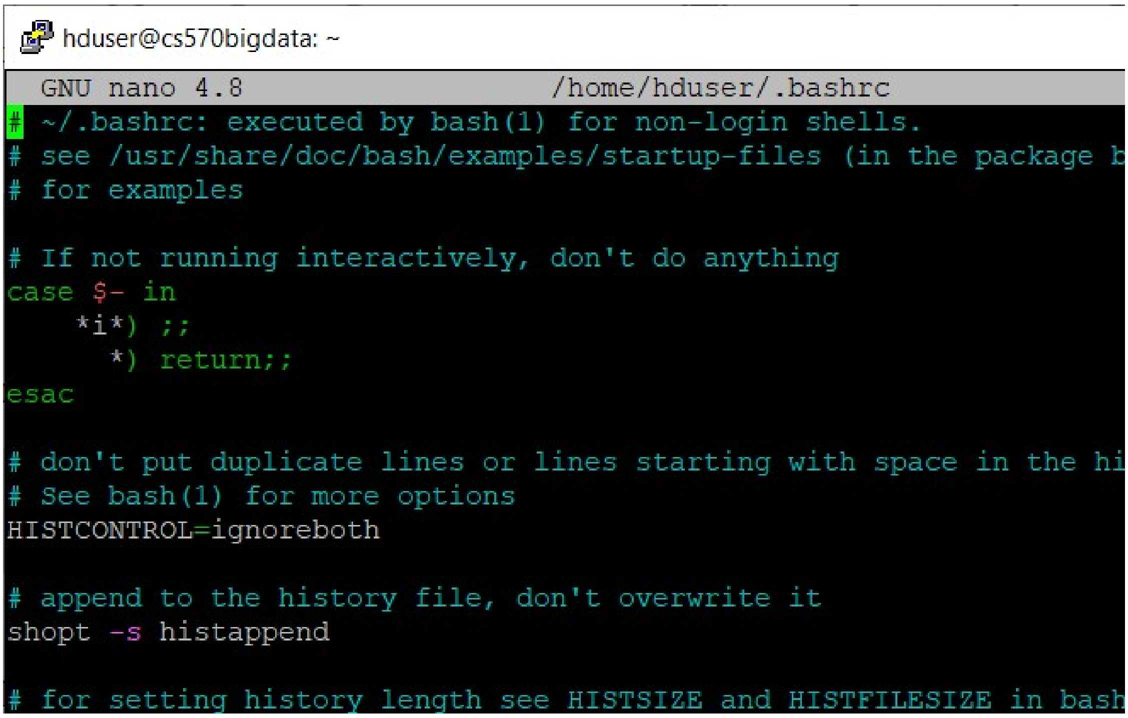
Cd / usr/local

Sudo mv hadoop-2.10.2 hadoop

Sudo chown –R hduser:hadoop hadoop



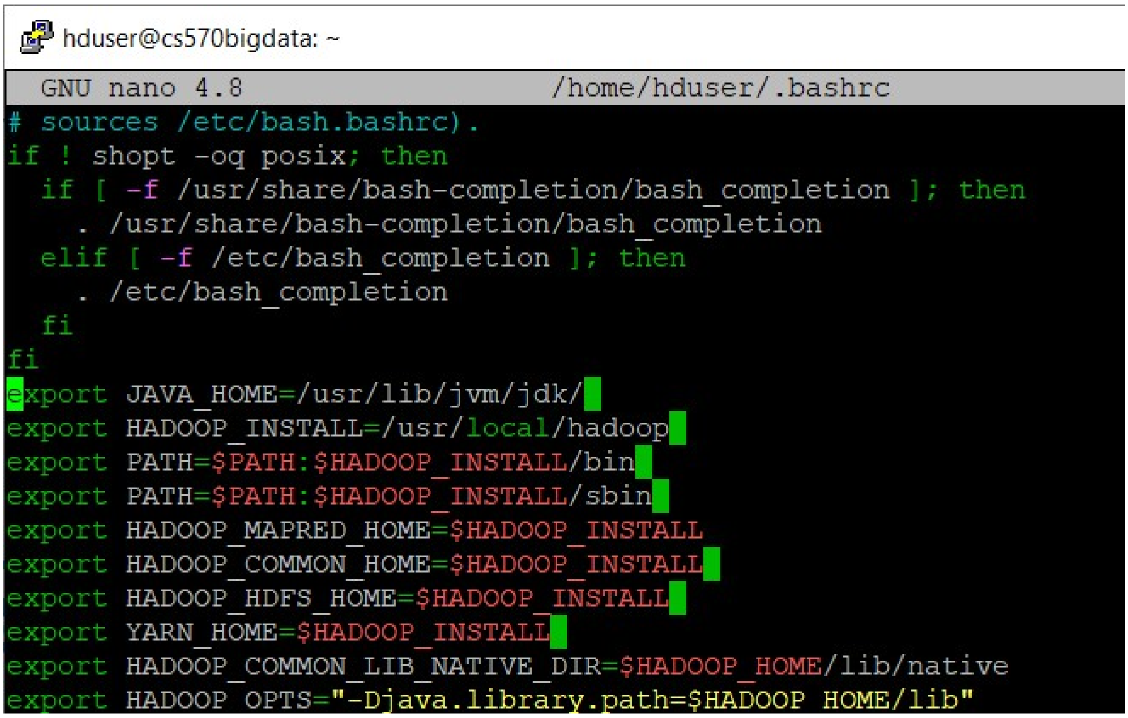
```
hduser@cs570bigdata:~$ sudo nano ~/.bashrc
[sudo] password for hduser:
```



hduser@cs570bigdata: ~

```
  GNU nano 4.8                     /home/hduser/.bashrc
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package b
# for examples

# If not running interactively, don't do anything
case $- in
    *i*) ;;
      *) return;;
esac

# don't put duplicate lines or lines starting with space in the hi
# See bash(1) for more options
HISTCONTROL=ignoreboth

# append to the history file, don't overwrite it
shopt -s histappend

# for setting history length see HISTSIZE and HISTFILESIZE in bash
```

Move to the end and add following lines for hadoop



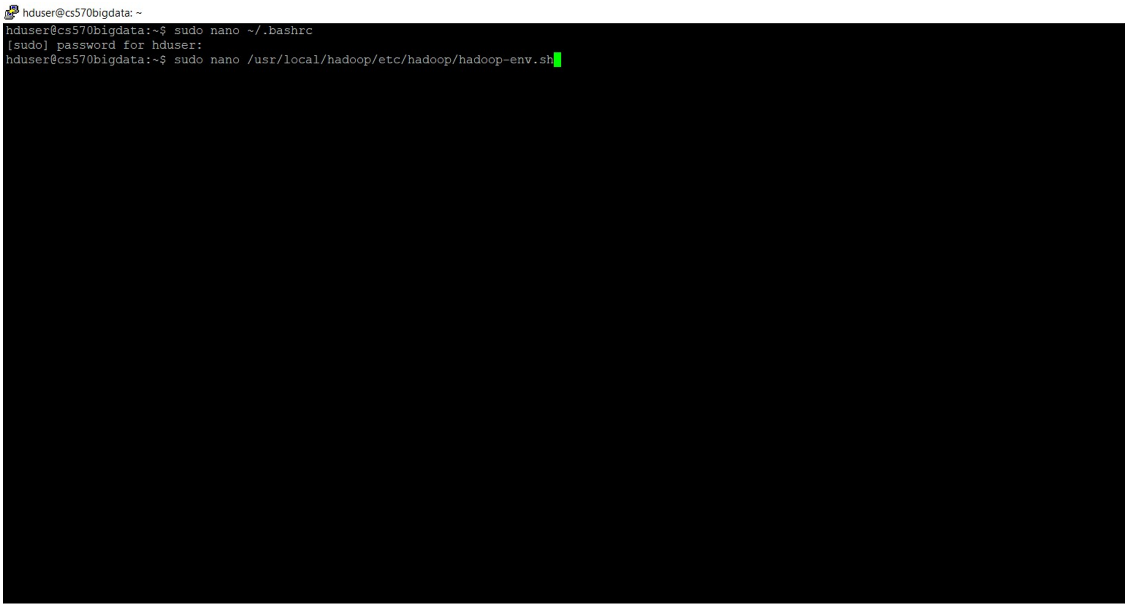
hduser@cs570bigdata: ~

```
  GNU nano 4.8                     /home/hduser/.bashrc
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export JAVA_HOME=/usr/lib/jvm/jdk/
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
```

To save use source ~/.bashrc

systemctl reboot –i

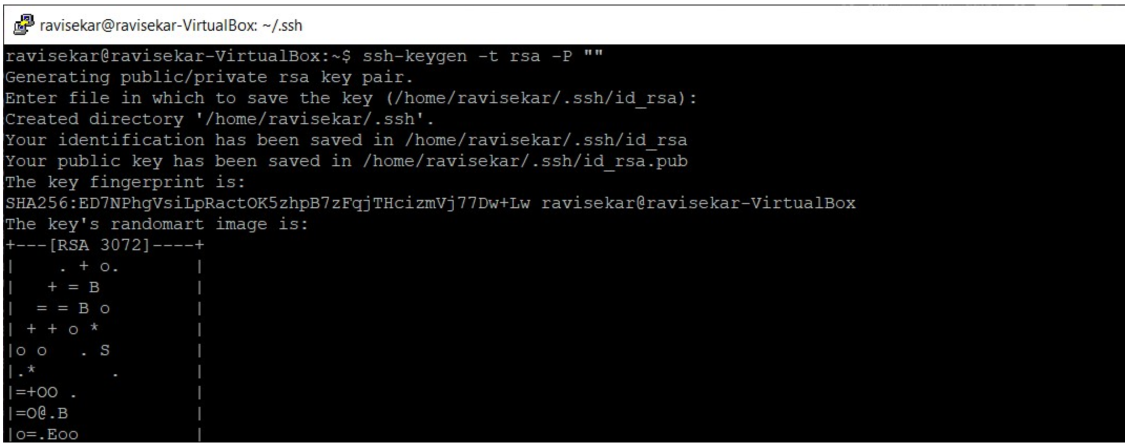
Now do the following setting for hduser



```
hduser@cs570bigdata:~$ sudo nano ~/.bashrc
[sudo] password for hduser:
hduser@cs570bigdata:~$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

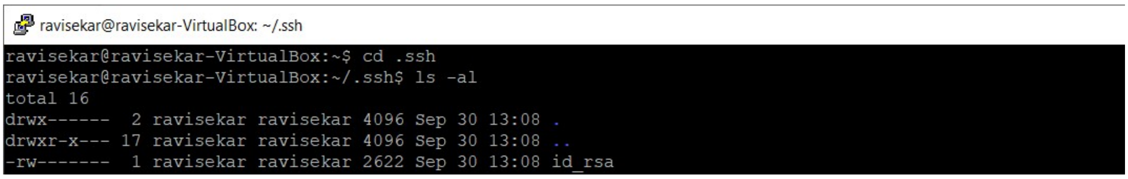
export JAVA_HOME=/usr/lib/jvm/jdk

Ssh generation and creation of authorized keys from public ssh keys



ravisekar@ravisekar-VirtualBox: ~/.ssh

```
ravisekar@ravisekar-VirtualBox:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ravisekar/.ssh/id_rsa):
Created directory '/home/ravisekar/.ssh'.
Your identification has been saved in /home/ravisekar/.ssh/id_rsa
Your public key has been saved in /home/ravisekar/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:ED7NPhgVsiLpRactOK5zhpB7zFqjTHcizmVj77Dw+Lw ravisekar@ravisekar-VirtualBox
The key's randomart image is:
+---[RSA 3072]----+
|    . + o.       |
|   + = B         |
|  = = B o        |
| + + o *         |
|o o   . S        |
|.*        .      |
|=+OO .           |
|=O@.B            |
|o=.Eoo           |
```

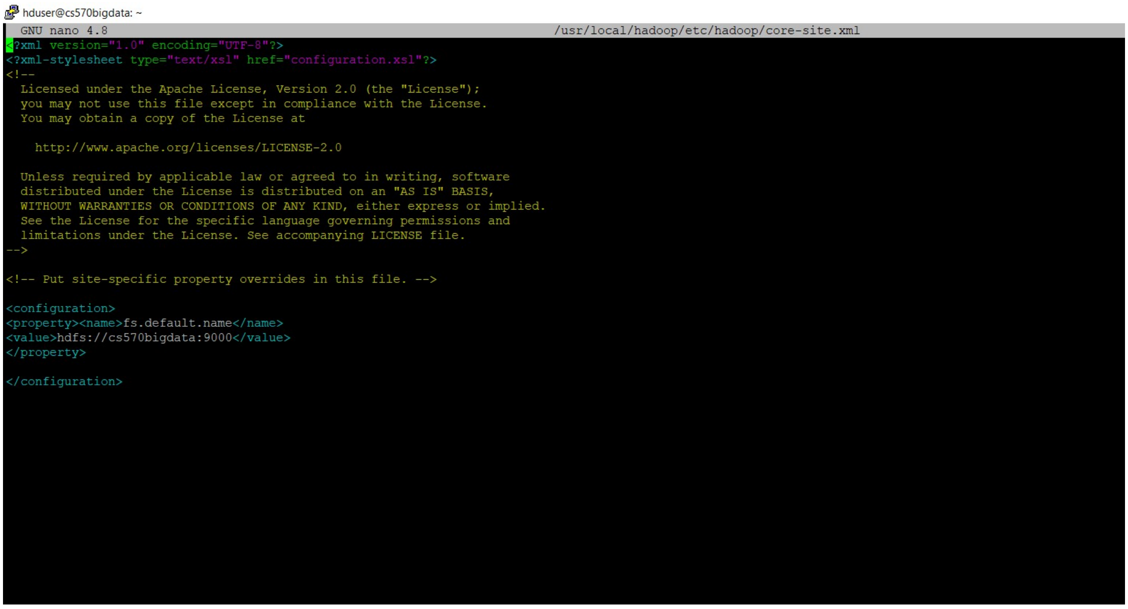


ravisekar@ravisekar-VirtualBox: ~/.ssh

```
ravisekar@ravisekar-VirtualBox:~$ cd .ssh
ravisekar@ravisekar-VirtualBox:~/.ssh$ ls -al
total 16
drwx------  2 ravisekar ravisekar 4096 Sep 30 13:08 .
drwxr-x--- 17 ravisekar ravisekar 4096 Sep 30 13:08 ..
-rw-------  1 ravisekar ravisekar 2622 Sep 30 13:08 id_rsa
```

Now we need following files to be set for hadoop :

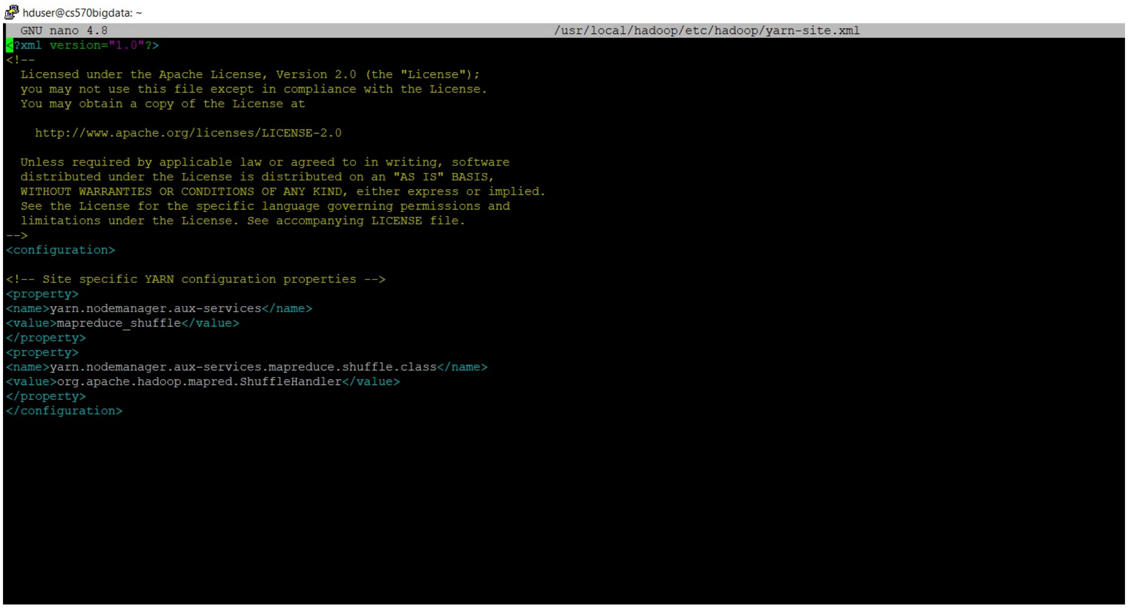
1 . sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml



hduser@cs570bigdata: ~

```
  GNU nano 4.8                                                      /usr/local/hadoop/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>fs.default.name</name>
<value>hdfs://cs570bigdata:9000</value>
</property>

</configuration>
```

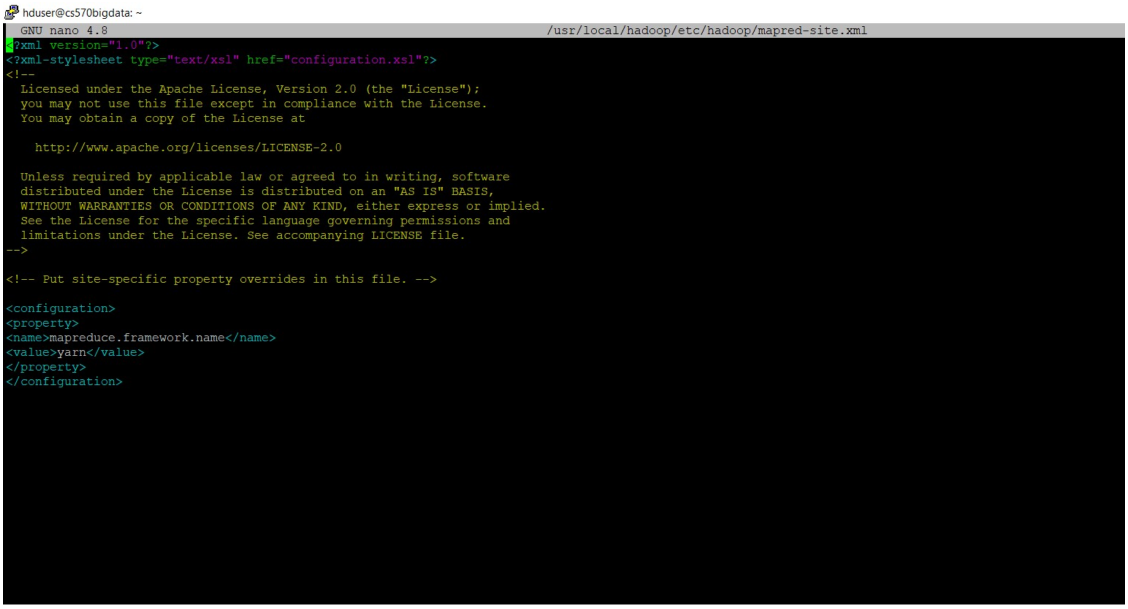
2. sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml



hduser@cs570bigdata: ~

```
  GNU nano 4.8                                                      /usr/local/hadoop/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>

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

3. sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml



```
hduser@cs570bigdata: ~
  GNU nano 4.8                                                                         /usr/local/hadoop/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>
```

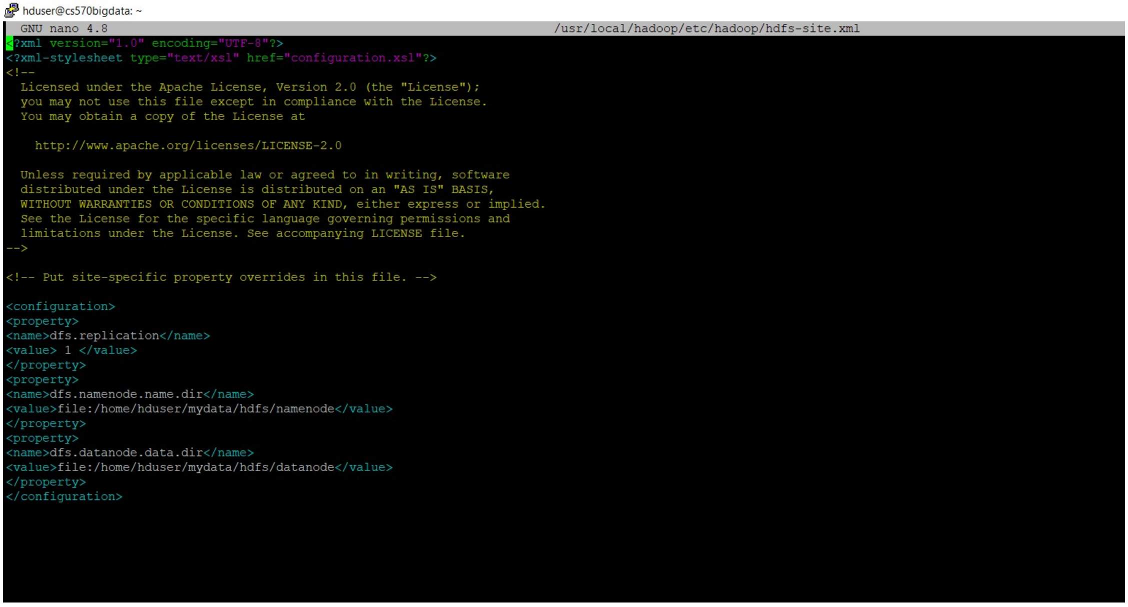
4 For doing hdfs.xml

First complete this task's

mkdir –p mydata/hdfs/namenode

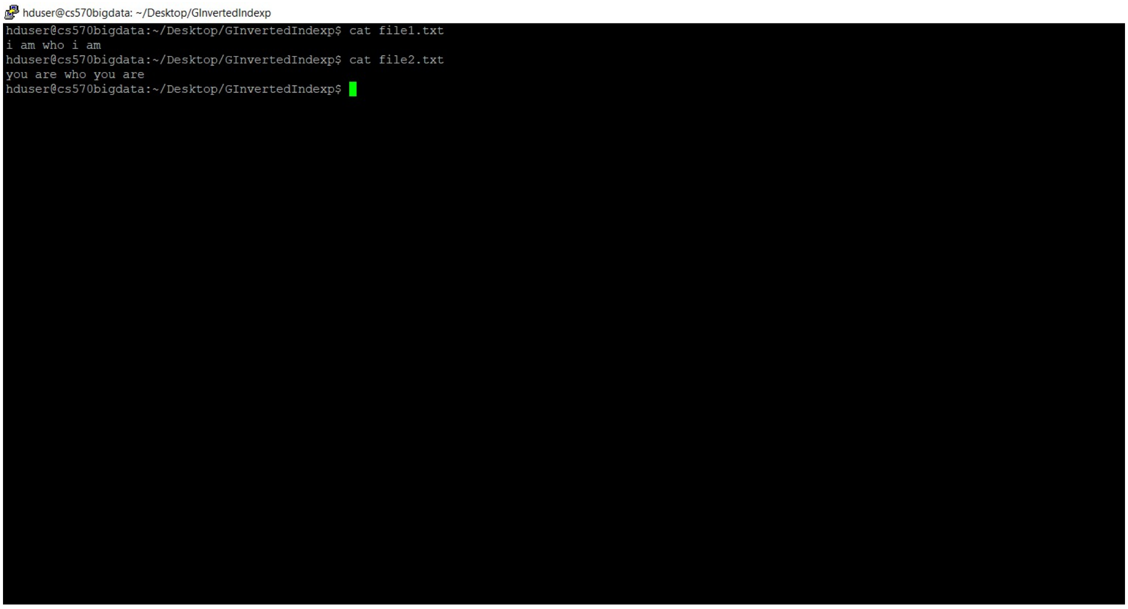
mkdir –p mydata/hdfs/datanode

sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml



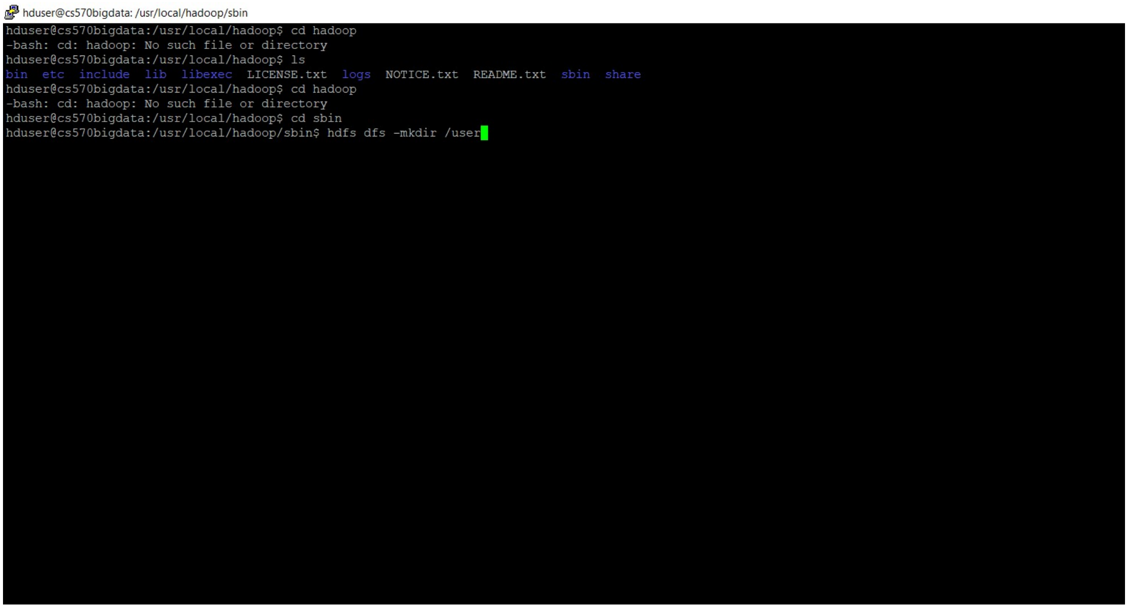
```
hduser@cs570bigdata: ~
  GNU nano 4.8                                                                          /usr/local/hadoop/etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 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.replication</name>
<value> 1 </value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/home/hduser/mydata/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/home/hduser/mydata/hdfs/datanode</value>
</property>
</configuration>
```

Create following test data file for hdfs



```
hduser@cs570bigdata:~/Desktop/GInvertedIndexp$ cat file1.txt
i am who i am
hduser@cs570bigdata:~/Desktop/GInvertedIndexp$ cat file2.txt
you are who you are
hduser@cs570bigdata:~/Desktop/GInvertedIndexp$
```

Create directory as mentioned



```
hduser@cs570bigdata:/usr/local/hadoop$ cd hadoop
-bash: cd: hadoop: No such file or directory
hduser@cs570bigdata:/usr/local/hadoop$ ls
bin  etc  include  lib  libexec  LICENSE.txt  logs  NOTICE.txt  README.txt  sbin  share
hduser@cs570bigdata:/usr/local/hadoop$ cd hadoop
-bash: cd: hadoop: No such file or directory
hduser@cs570bigdata:/usr/local/hadoop$ cd sbin
hduser@cs570bigdata:/usr/local/hadoop/sbin$ hdfs dfs -mkdir /user
```

Now copy to file created for input

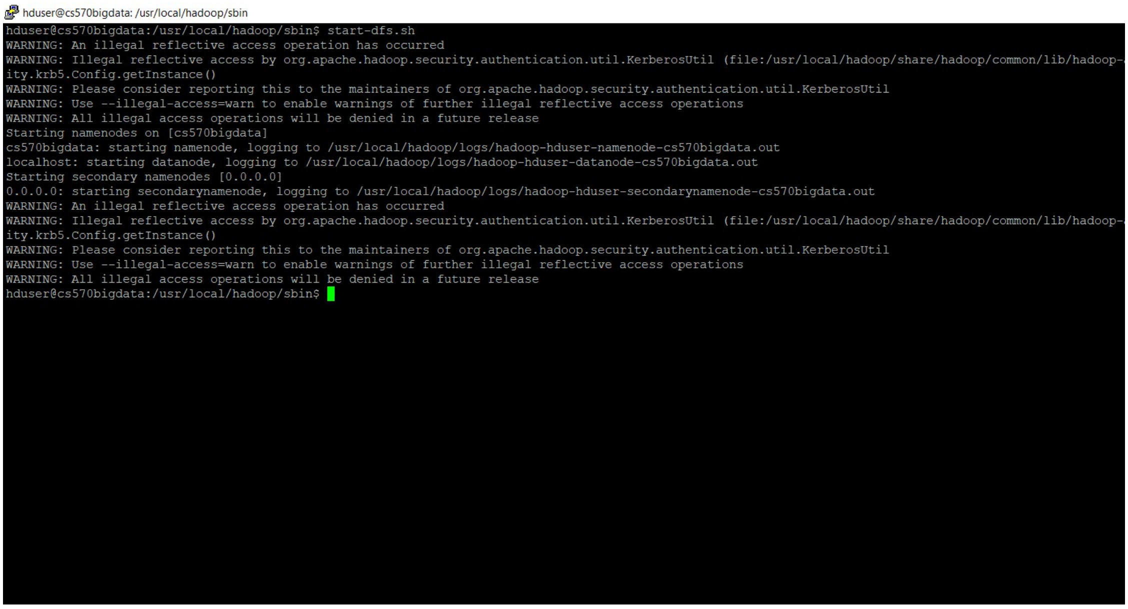
```
/usr/local/hadoop/bin/hdfs dfs -put '/home/hduser/Desktop/inputdata' /user
```

Next final step for hadoop

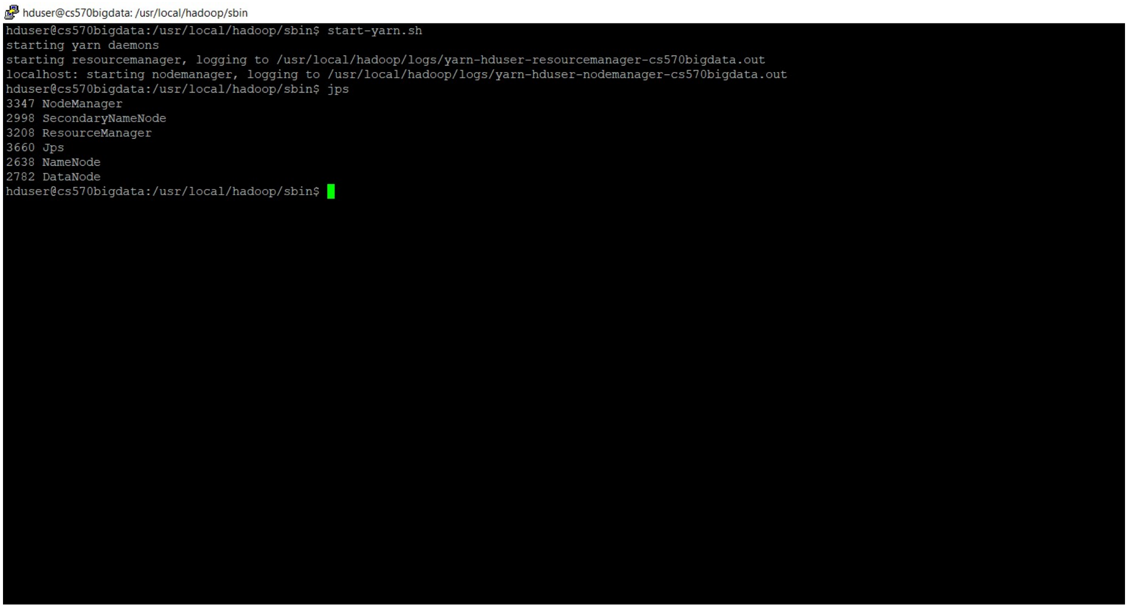
```
hdfs namenode –format
```

now everything is set we can start hadoop and run the jar for the word count and get the output as shown in below screen

now start hadoop as shown below by running
start-dfs.sh
start-yarn.sh and jps to check the status as shown in below screens

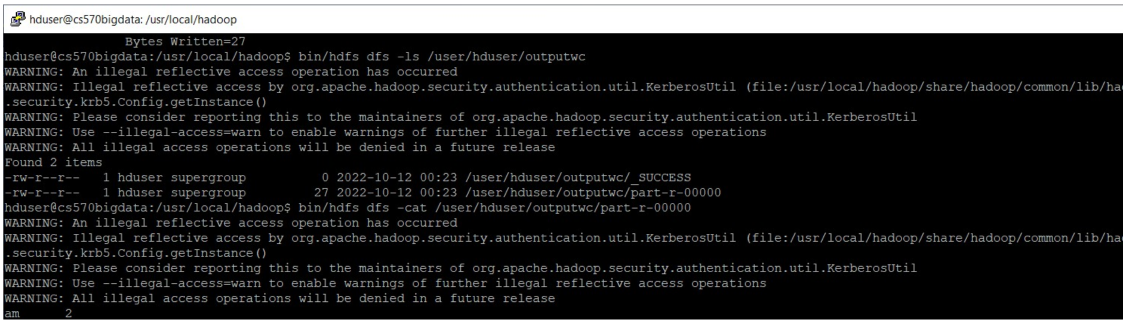


```
hduser@cs570bigdata:/usr/local/hadoop/sbin$ start-dfs.sh
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-
ity.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Starting namenodes on [cs570bigdata]
cs570bigdata: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-cs570bigdata.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-cs570bigdata.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-cs570bigdata.out
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-
ity.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
hduser@cs570bigdata:/usr/local/hadoop/sbin$
```



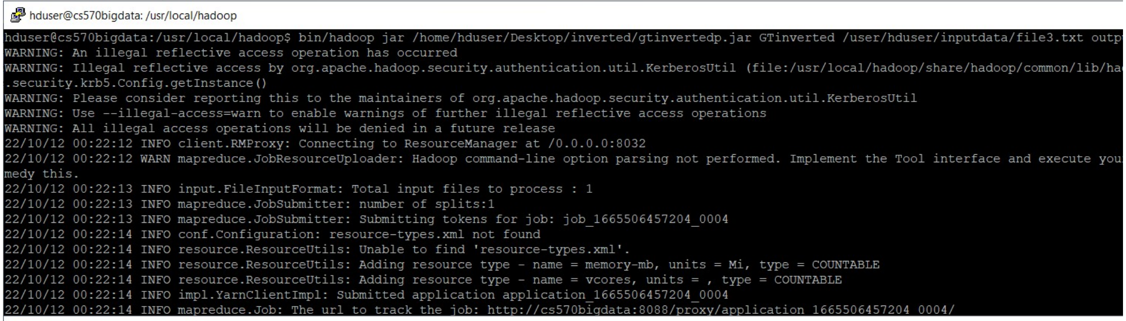
```
hduser@cs570bigdata:/usr/local/hadoop/sbin$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-cs570bigdata.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-cs570bigdata.out
hduser@cs570bigdata:/usr/local/hadoop/sbin$ jps
3347 NodeManager
2998 SecondaryNameNode
3208 ResourceManager
3660 Jps
2638 NameNode
2782 DataNode
hduser@cs570bigdata:/usr/local/hadoop/sbin$
```

Output generated by MapReduce Inverted Program

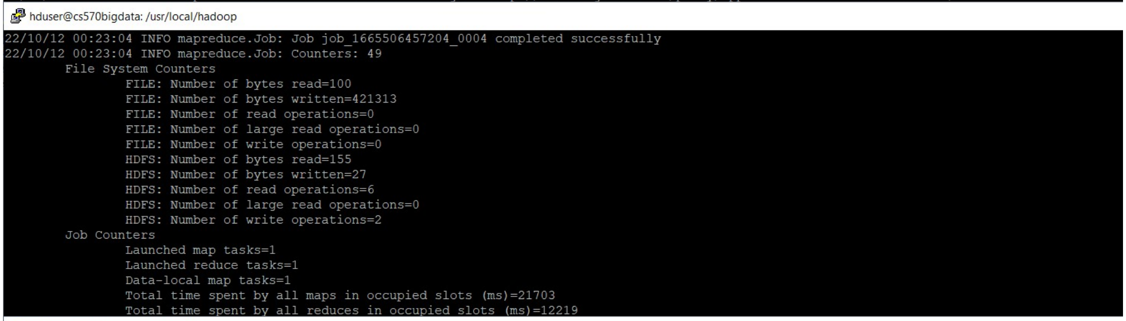


```
                Bytes Written=27
hduser@cs570bigdata:/usr/local/hadoop$ bin/hdfs dfs -ls /user/hduser/outputwc
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/ha
.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-10-12 00:23 /user/hduser/outputwc/_SUCCESS
-rw-r--r--   1 hduser supergroup         27 2022-10-12 00:23 /user/hduser/outputwc/part-r-00000
hduser@cs570bigdata:/usr/local/hadoop$ bin/hdfs dfs -cat /user/hduser/outputwc/part-r-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/ha
.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
am      2
```

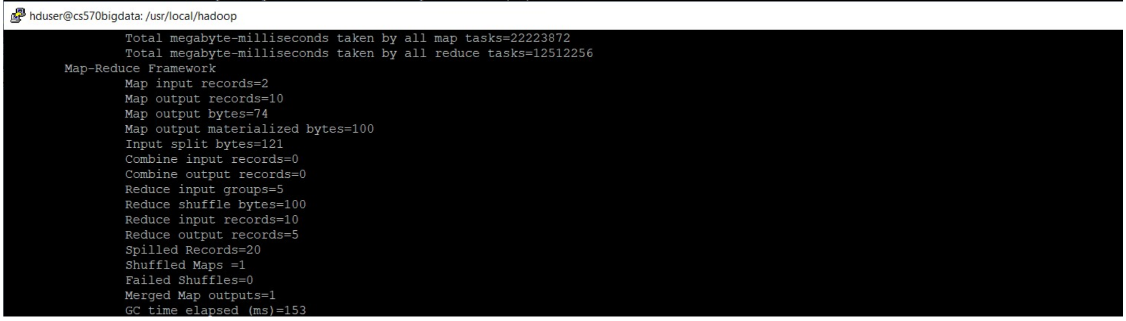
Run hadoop jar file as shown in the screen :



```
hduser@cs570bigdata:/usr/local/hadoop$ bin/hadoop jar /home/hduser/Desktop/inverted/gtinvertedp.jar GTinverted /user/hduser/inputdata/file3.txt outp
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/ha
.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/10/12 00:22:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/10/12 00:22:12 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute you
medy this.
22/10/12 00:22:13 INFO input.FileInputFormat: Total input files to process : 1
22/10/12 00:22:13 INFO mapreduce.JobSubmitter: number of splits:1
22/10/12 00:22:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1665506457204_0004
22/10/12 00:22:14 INFO conf.Configuration: resource-types.xml not found
22/10/12 00:22:14 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/10/12 00:22:14 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/10/12 00:22:14 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/10/12 00:22:14 INFO impl.YarnClientImpl: Submitted application application_1665506457204_0004
22/10/12 00:22:14 INFO mapreduce.Job: The url to track the job: http://cs570bigdata:8088/proxy/application_1665506457204_0004/
```

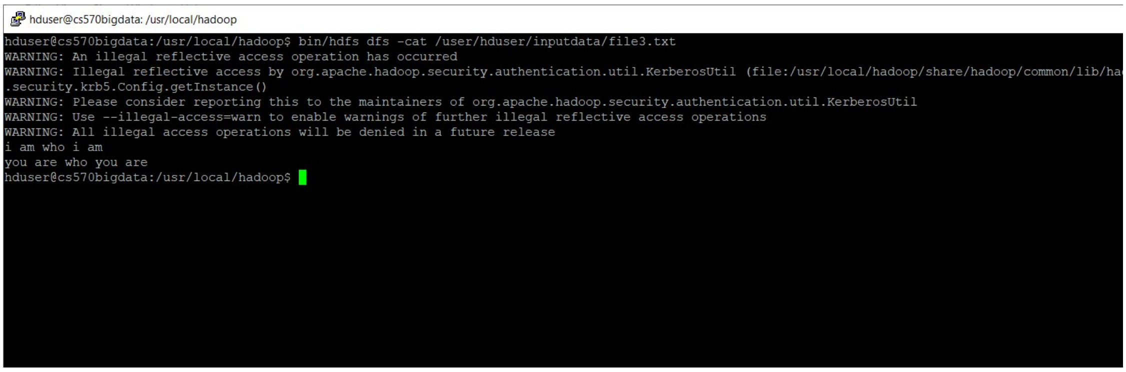


```
22/10/12 00:23:04 INFO mapreduce.Job: Job job_1665506457204_0004 completed successfully
22/10/12 00:23:04 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=100
                FILE: Number of bytes written=421313
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=155
                HDFS: Number of bytes written=27
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=21703
                Total time spent by all reduces in occupied slots (ms)=12219
```



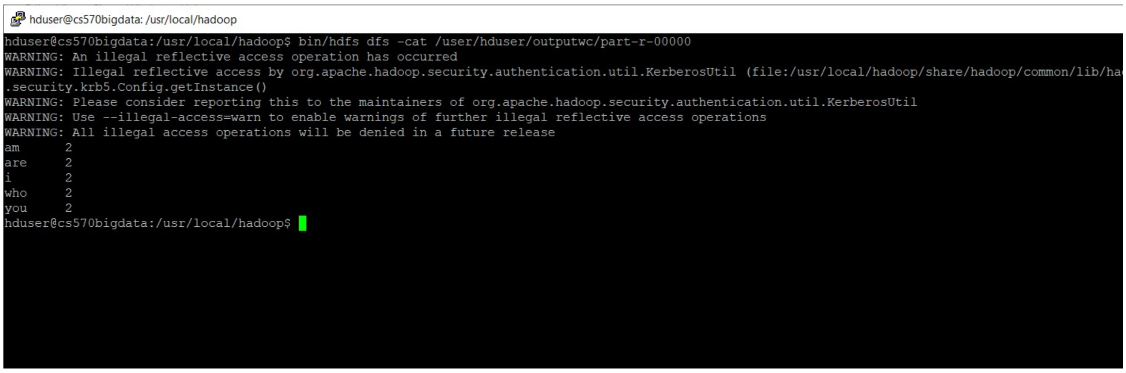
```
                Total megabyte-milliseconds taken by all map tasks=22223872
                Total megabyte-milliseconds taken by all reduce tasks=12512256
        Map-Reduce Framework
                Map input records=2
                Map output records=10
                Map output bytes=74
                Map output materialized bytes=100
                Input split bytes=121
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=100
                Reduce input records=10
                Reduce output records=5
                Spilled Records=20
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=153
```

Inverted index output for the inputfiles
File1
File2



```
hduser@cs570bigdata:/usr/local/hadoop$ bin/hdfs dfs -cat /user/hduser/inputdata/file3.txt
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/ha
.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
i am who i am
you are who you are
hduser@cs570bigdata:/usr/local/hadoop$
```

Output generated by MapReduce inverted program



```
hduser@cs570bigdata:/usr/local/hadoop$ bin/hdfs dfs -cat /user/hduser/outputwc/part-r-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/ha
.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
am      2
are     2
i       2
who     2
you     2
hduser@cs570bigdata:/usr/local/hadoop$
```