Practical 2

Installing and setting environment variables for working with apache Hadoop in linux

Theory: Apache Hadoop is an open-source framework designed for distributed storage and processing of large datasets using a cluster of computers. It is particularly suited for handling unstructured and semi-structured data. Here's an overview:

Core Components of Hadoop

1. **Hadoop Distributed File System (HDFS)**
   - **Purpose**: Distributed storage system that stores data across multiple nodes in a cluster.
   - **Features**:
     - Fault tolerance through data replication.
     - Handles large files by breaking them into smaller blocks (typically 128 MB or 256 MB).
     - Write-once, read-many model.
2. **MapReduce**
   - **Purpose**: Distributed data processing framework.
   - **Features**:
     - Breaks tasks into small chunks, processes them in parallel, and aggregates results.
     - Ideal for batch processing of large datasets.
3. **Yet Another Resource Negotiator (YARN)**
   - **Purpose**: Resource management layer that allocates system resources to various applications running on a cluster.
   - **Features**:
     - Supports multiple processing engines beyond MapReduce (e.g., Apache Spark, Flink).
4. **Hadoop Common**
   - Provides shared utilities, libraries, and services required by other Hadoop modules.

Steps:

A)Installing Virtual Box

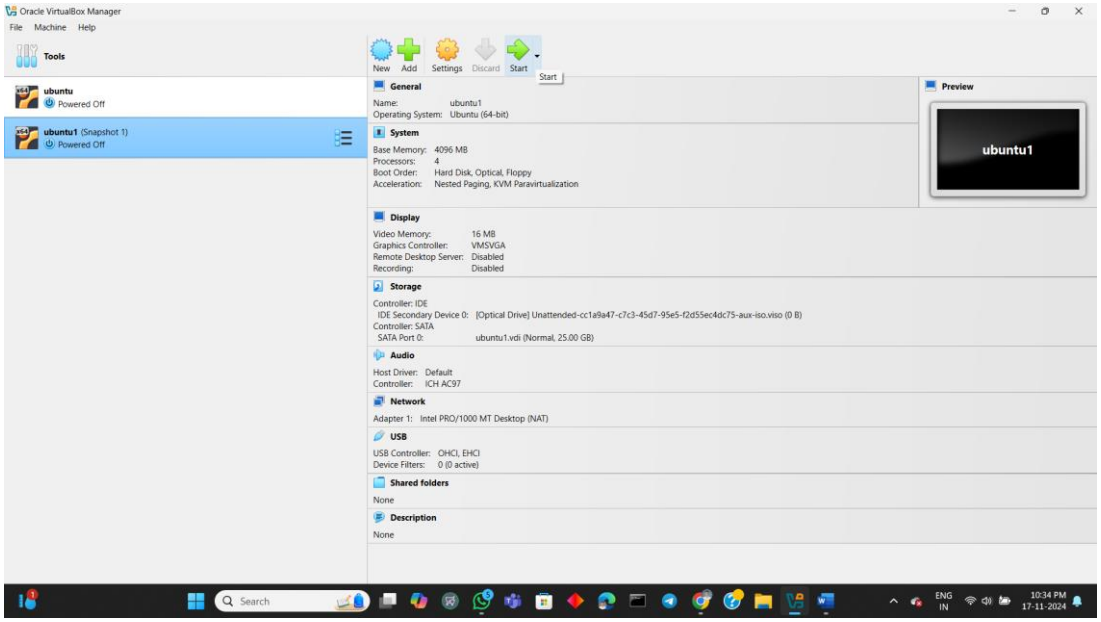
Link:https://download.virtualbox.org/virtualbox/7.1.4/VirtualBox-7.1.4-165100-Win.exe

B) Installing and Integrating Ubuntu in Virtual box

Link: https://ubuntu.com/download/desktop/thank-you?version=24.04.1&architecture=amd64&lts=true

C) Installing JDK & Hadoop in Ubuntu

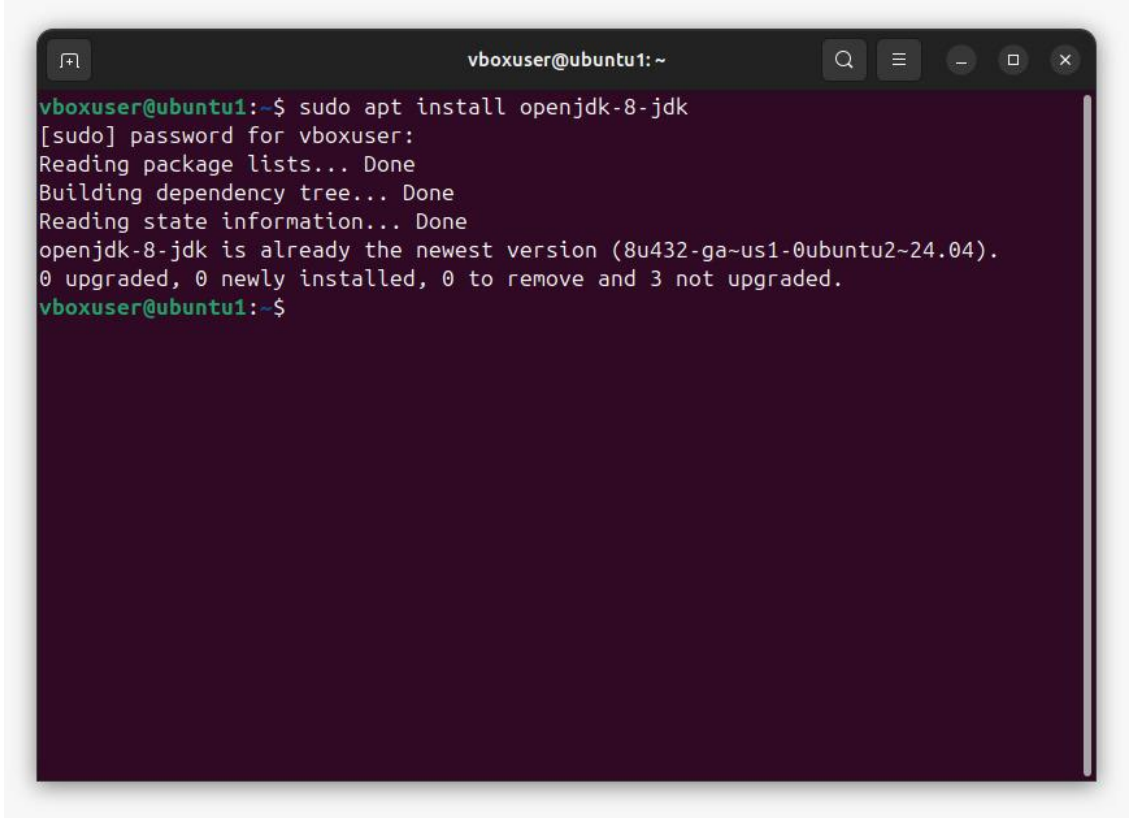
Open virtual box and start ubuntu machine



Step 1 : Install java jdk 8

First of all you must install Java JDK 8 on your system. You can just type this command to install java jdk on your system.

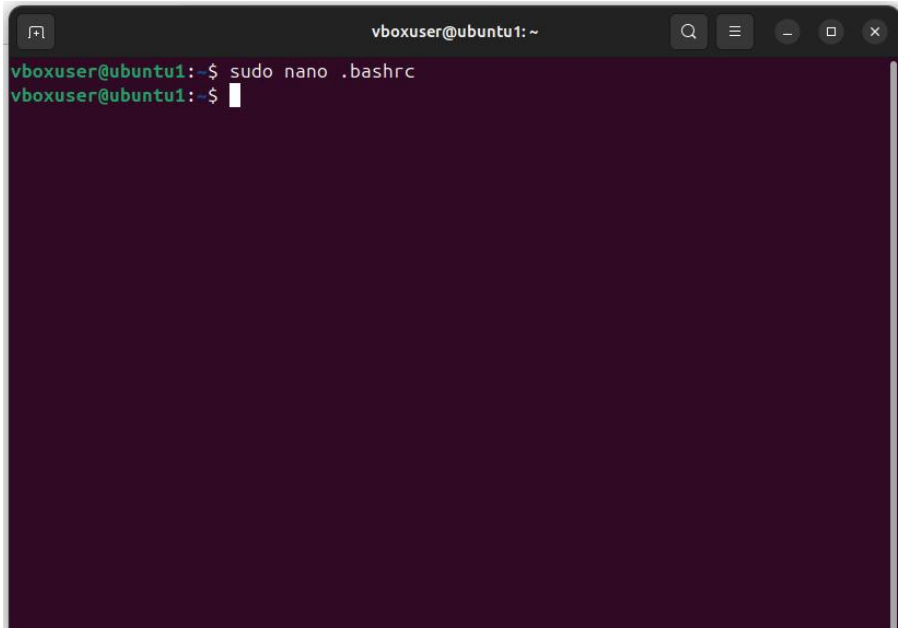
Cmd: **sudo apt install openjdk-8-jdk**



Step 2 : Add this configuration on you bash file

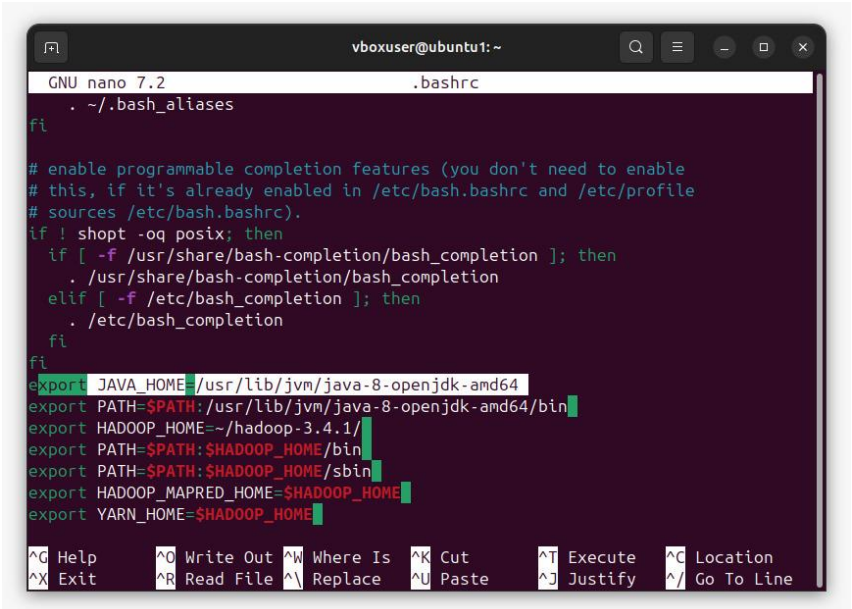
Now just open .bashrc file and paste these commands.

1)Cmd: sudo nano .bashrc

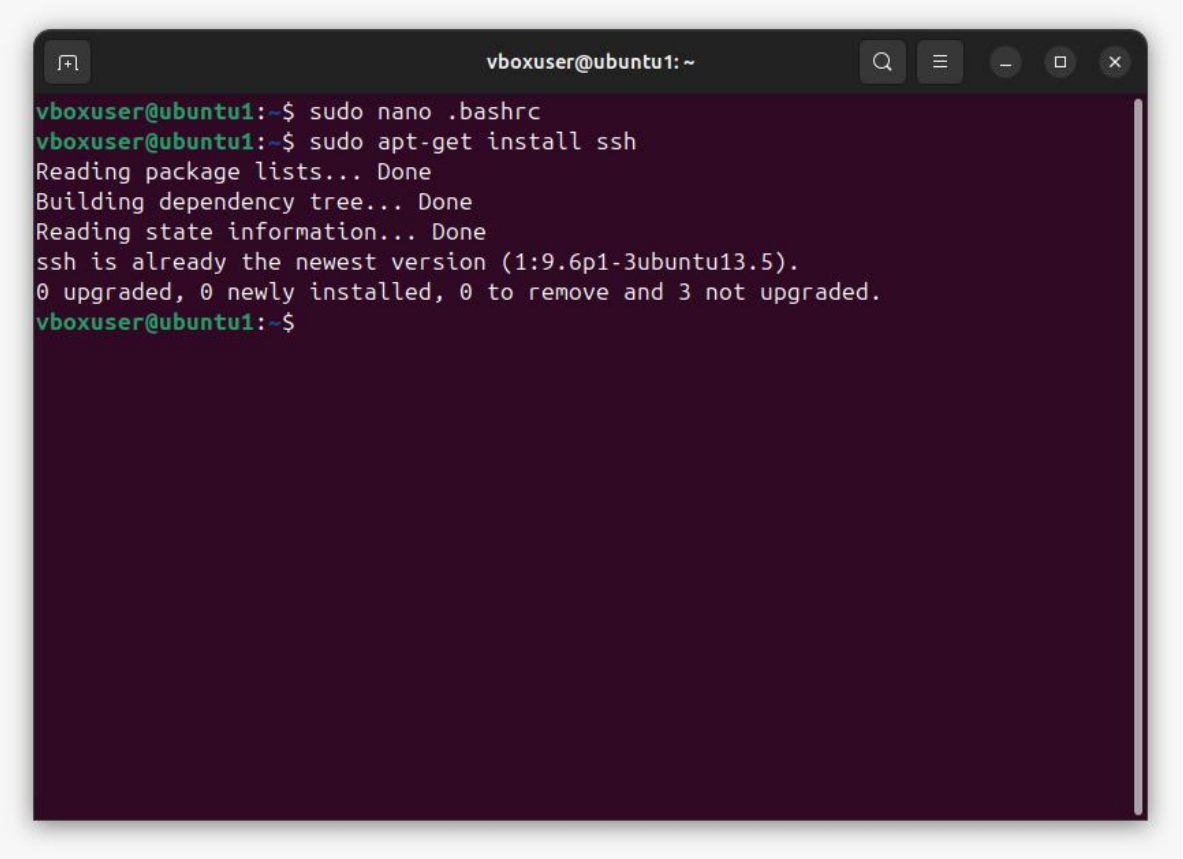


Copy paste:

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

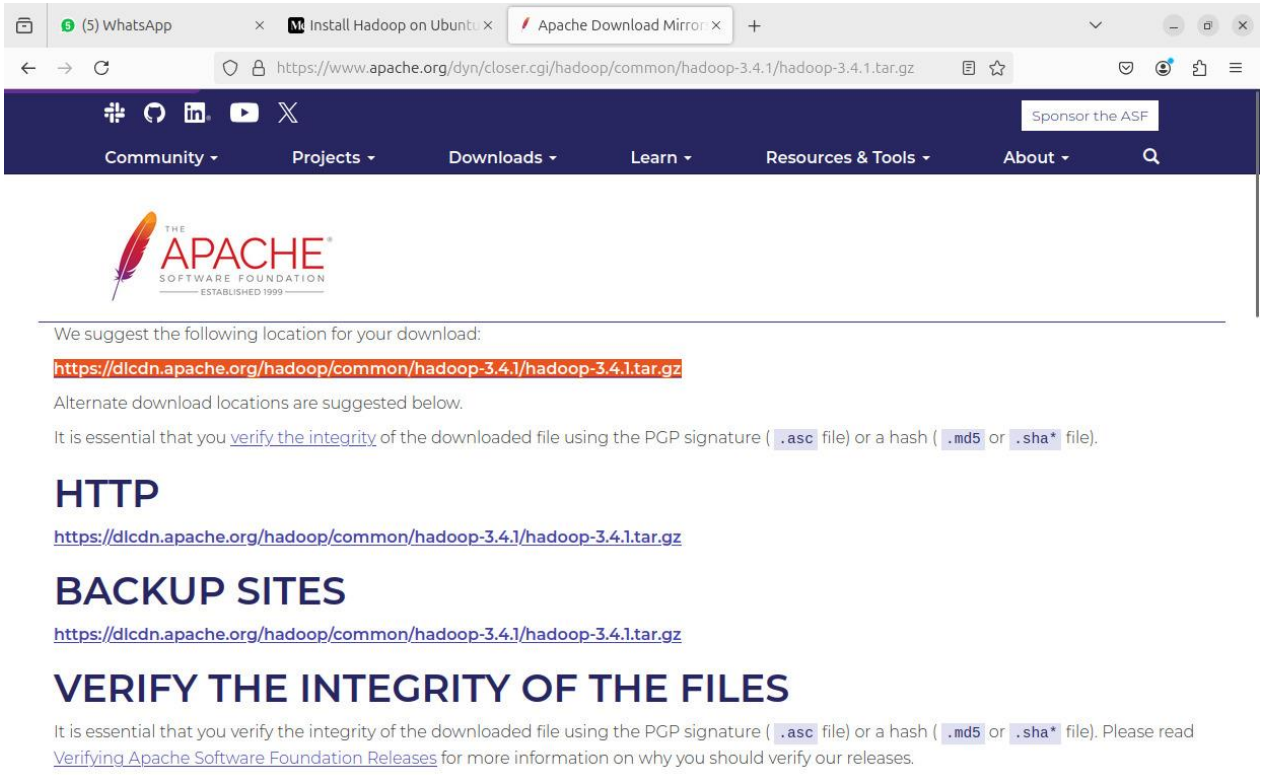


2)Cmd: sudo apt-get install ssh



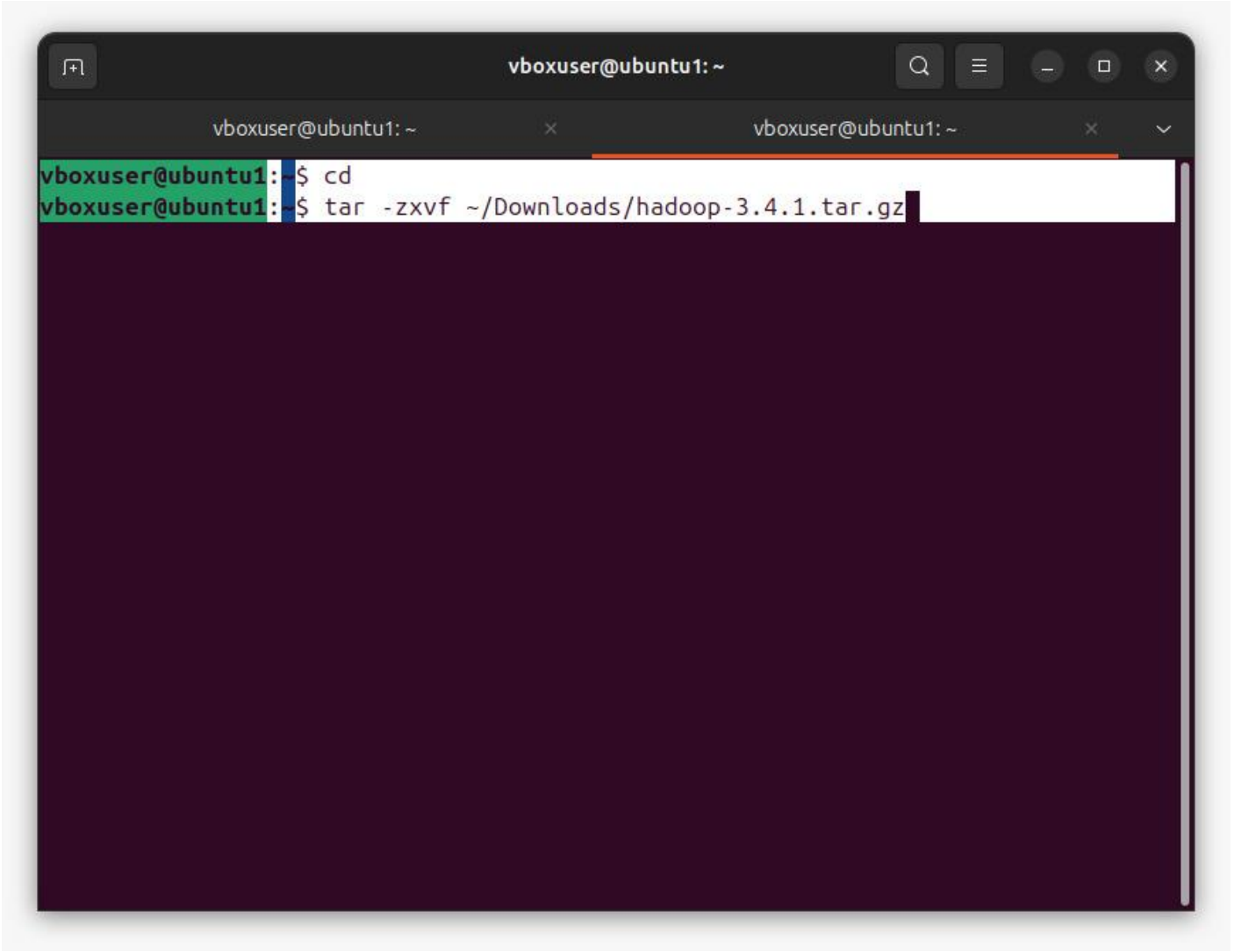
3) now go to hadoop.apache.org website download the tar file (hadoop.apache.org — download tar file of hadoop.)

Link: https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz



5)now type

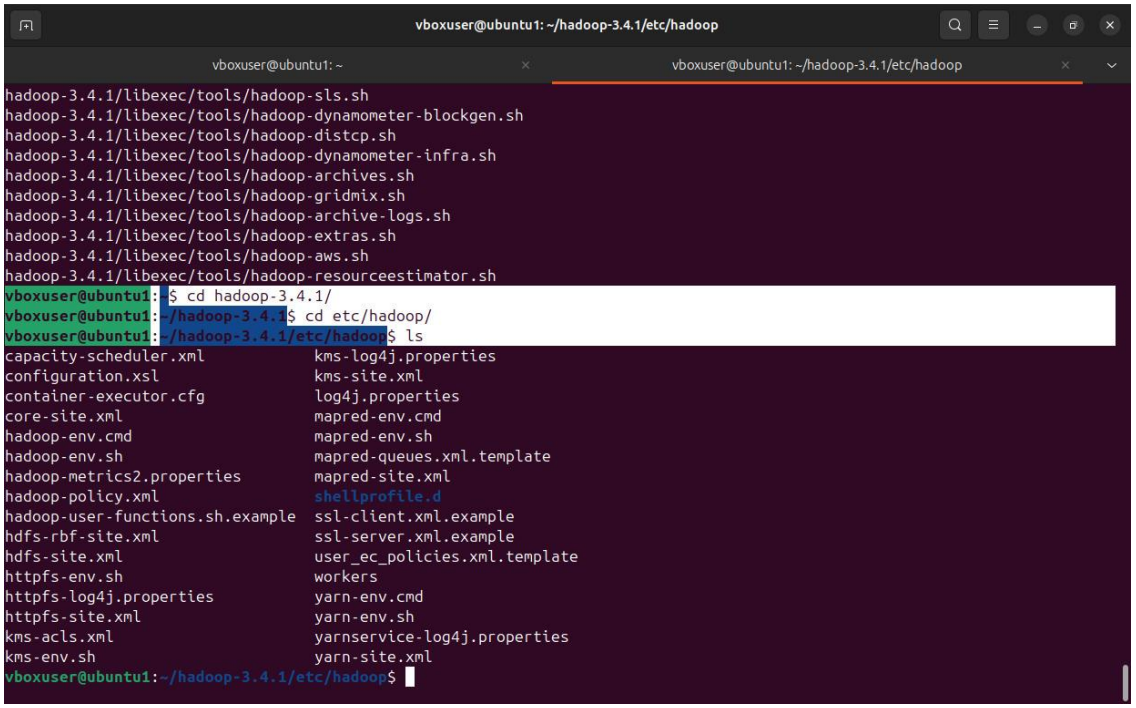
Cmd: tar -zxvf ~/Downloads/hadoop-3.4.1.tar.gz



6)Now change the directory to edit the files

(Extract the tar file)

Cmd: cd hadoop-3.4.1/etc/Hadoop

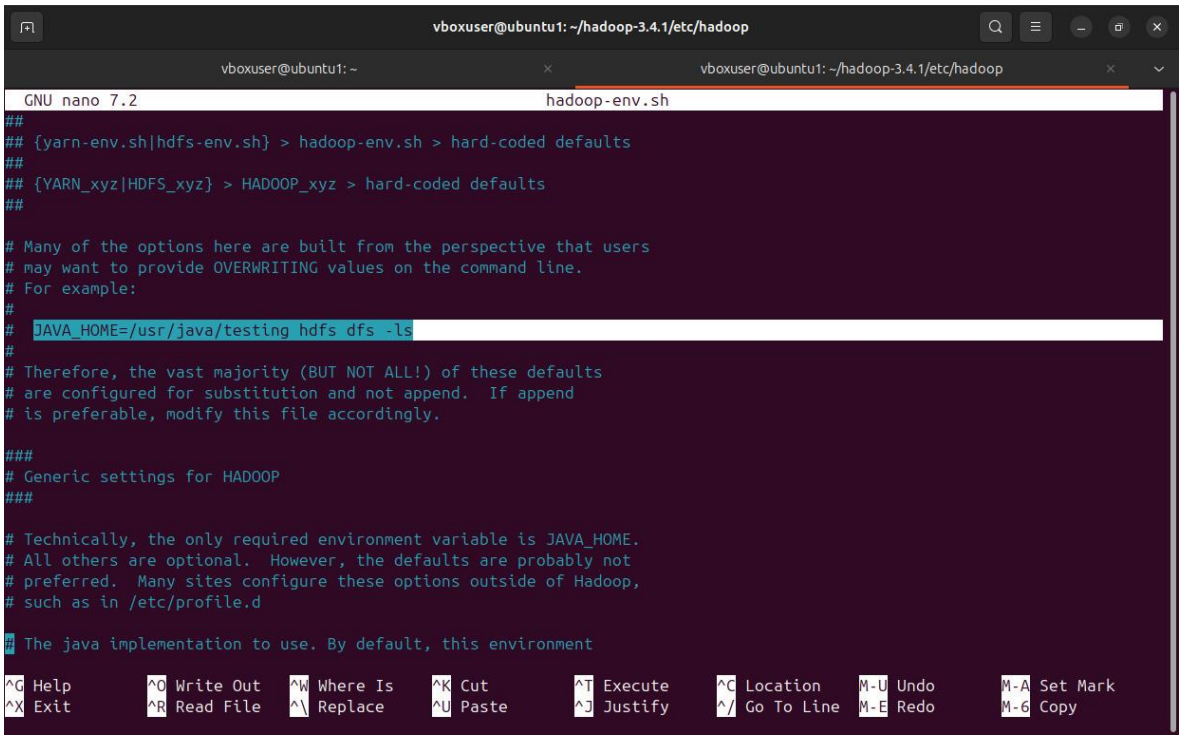


7) now open hadoop-env.h

Cmd: sudo nano hadoop-env.h

In that change the java home path i.e

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



Step 3 : Add this the following code in the respective xml Files

Cmd: sudo nano core-site.xml

```
 GNU nano 6.2                                                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.defaultFS</name>
 <value>hdfs://localhost:9000</value>  </property>
 <property>
<name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.groups</name> <value>*</value>
 </property>
</configuration>
```

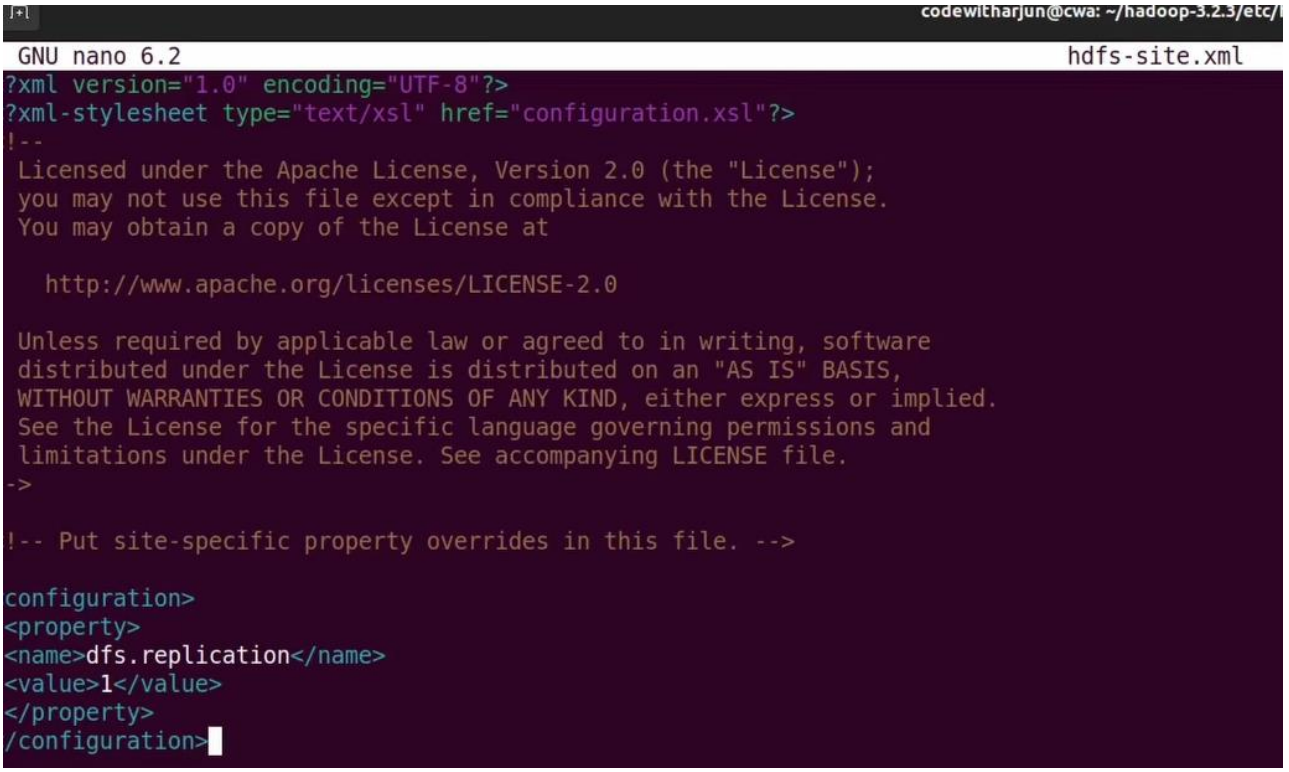
Now add this configuration in core-site.xml file.

1. **core-site.xml**

```
<configuration>
 <property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>  </property>
 <property>
<name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.groups</name> <value>*</value>
 </property>
</configuration>
```

2. hdfs-site.xml:

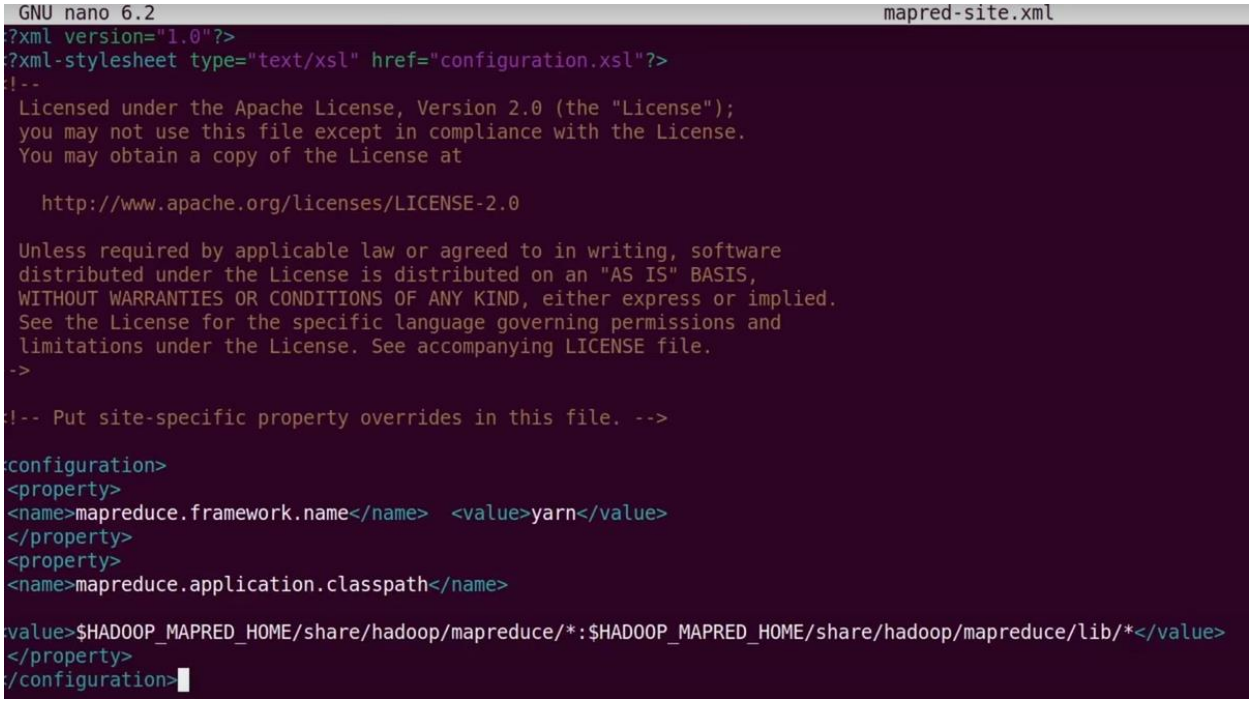
Cmd: sudo nano hdfs-site.xml




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

3. mapred-site.xml

Cmd: sudo nano mapred-site.xml

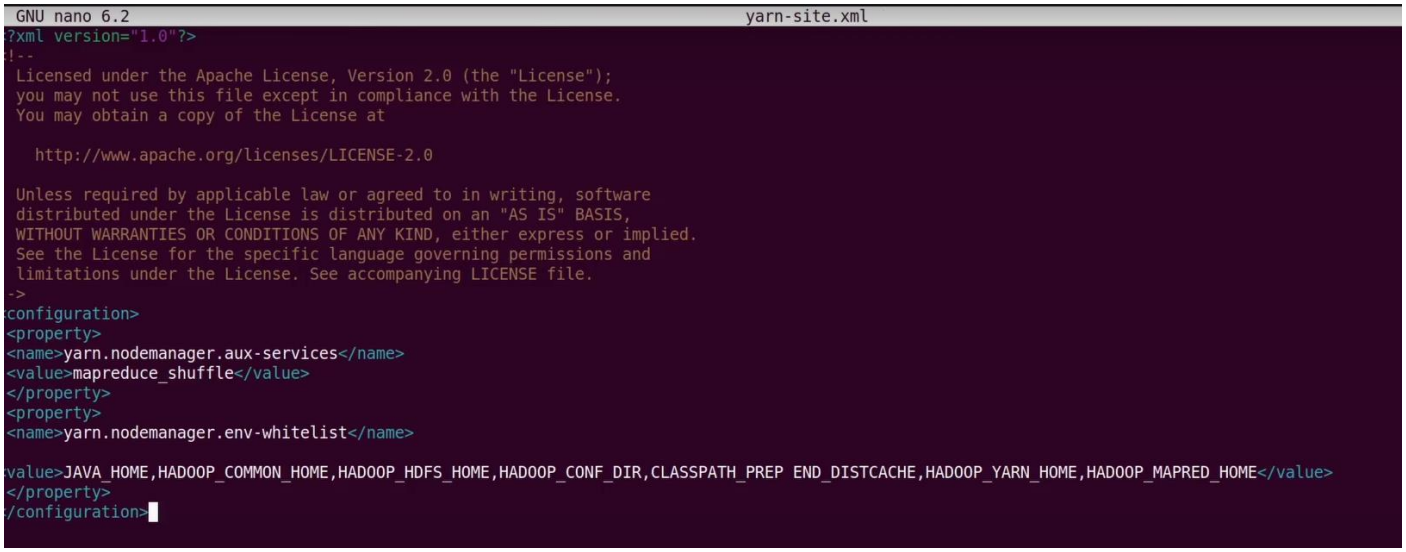


<configuration>
<property>
<name>mapreduce.framework.name</name> <value>yarn</value>
</property>
<property>
<name>mapreduce.application.classpath</name>

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

4. yarn-site xml

Cmd: **sudo nano yarn-site.xml**



<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>

Step 4: Now ssh part

```
codewitharjun@cwa:~/hadoop-3.2.3/etc/hadoop$ sudo nano core-site.xml
codewitharjun@cwa:~/hadoop-3.2.3/etc/hadoop$ sudo nano hdfs-site.xml
codewitharjun@cwa:~/hadoop-3.2.3/etc/hadoop$ sudo nano mapred-site.xml
codewitharjun@cwa:~/hadoop-3.2.3/etc/hadoop$ sudo nano yarn-site.xml
codewitharjun@cwa:~/hadoop-3.2.3/etc/hadoop$ ssh localhost
Welcome to Ubuntu 22.04 LTS (GNU/Linux 5.15.0-27-generic x86_64)

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

0 updates can be applied immediately.
```

CMD:ssh localhost

CMD: ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

CMD: chmod 0600 ~/.ssh/authorized_keys

CMD: hadoop-3.4.1/bin/hdfs namenode -format

STEP:5 FORMAT THE FILE SYSTEM

CMD: export PDSH_RCMD_TYPE=ssh

```
codewitharjun@cwa: ~
2022-05-10 20:45:03,790 INFO namenode.FSDirectory: XAttrs enabled? true
2022-05-10 20:45:03,790 INFO namenode.NameNode: Caching file names occurring more than 10 times
2022-05-10 20:45:03,799 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescend
ant: true, maxSnapshotLimit: 65536
2022-05-10 20:45:03,801 INFO snapshot.SnapshotManager: SkipList is disabled
2022-05-10 20:45:03,807 INFO util.GSet: Computing capacity for map cachedBlocks
2022-05-10 20:45:03,807 INFO util.GSet: VM type       = 64-bit
2022-05-10 20:45:03,807 INFO util.GSet: 0.25% max memory 3.4 GB = 8.8 MB
2022-05-10 20:45:03,807 INFO util.GSet: capacity      = 2^20 = 1048576 entries
2022-05-10 20:45:03,821 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2022-05-10 20:45:03,821 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2022-05-10 20:45:03,821 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2022-05-10 20:45:03,826 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2022-05-10 20:45:03,826 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2022-05-10 20:45:03,828 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2022-05-10 20:45:03,828 INFO util.GSet: VM type       = 64-bit
2022-05-10 20:45:03,828 INFO util.GSet: 0.029999999329447746% max memory 3.4 GB = 1.1 MB
2022-05-10 20:45:03,828 INFO util.GSet: capacity      = 2^17 = 131072 entries
Re-format filesystem in Storage Directory root= /tmp/hadoop-codewitharjun/dfs/name; location= null ? (Y or N) Y
2022-05-10 20:45:06,845 INFO namenode.FSImage: Allocated new BlockPoolId: BP-228066331-127.0.1.1-1652194806827
2022-05-10 20:45:06,845 INFO common.Storage: Will remove files: [/tmp/hadoop-codewitharjun/dfs/name/current/edits_inprogress_0000000000000000004, /tmp/hadoop-codewith
arjun/dfs/name/current/VERSION, /tmp/hadoop-codewitharjun/dfs/name/current/seen_txid, /tmp/hadoop-codewitharjun/dfs/name/current/edits_0000000000000000001-000000000000
00000001, /tmp/hadoop-codewitharjun/dfs/name/current/fsimage_0000000000000000000.md5, /tmp/hadoop-codewitharjun/dfs/name/current/edits_0000000000000000002-000000000000
00000003, /tmp/hadoop-codewitharjun/dfs/name/current/fsimage_0000000000000000003]
2022-05-10 20:45:06,863 INFO common.Storage: Storage directory /tmp/hadoop-codewitharjun/dfs/name has been successfully formatted.
2022-05-10 20:45:06,898 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-codewitharjun/dfs/name/current/fsimage.ckpt_0000000000000000000 using no co
mpression
2022-05-10 20:45:07,005 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-codewitharjun/dfs/name/current/fsimage.ckpt_0000000000000000000 of size 408 bytes
saved in 0 seconds .
2022-05-10 20:45:07,013 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2022-05-10 20:45:07,045 INFO namenode.FSNamesystem: Stopping services started for active state
2022-05-10 20:45:07,045 INFO namenode.FSNamesystem: Stopping services started for standby state
2022-05-10 20:45:07,050 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2022-05-10 20:45:07,050 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at cwa/127.0.1.1
************************************************************/
codewitharjun@cwa: $ export PDSH_RCMD_TYPE=ssh
codewitharjun@cwa: $ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as codewitharjun in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
```

STEP6: START HADOOP

CMD: start-all.sh