Installation of Single Node Hadoop Machine in Ubuntu VM

Install OpenJDK on Ubuntu

To update your system before installing any other installations.

```
sudo apt update
```

Use the below following command to install openjdk8

```
sudo apt install openjdk-8-jdk -y
```

```
hduser@bigdata-VirtualBox:~$ java -version;
openjdk version "1.8.0_275"
OpenJDK Runtime Environment (build 1.8.0_275-8u275-b01-0ubuntu1~18.04-b01)
OpenJDK 64-Bit Server VM (build 25.275-b01, mixed mode)
hduser@bigdata-VirtualBox:~$
```

Set up a root user for hadoop environment:

Install open SSH on ubuntu using below command

```
sudo apt install openssh-server openssh-client -y
```

```
See http://www.oracle.com/technetwork/java/javase/documentation/index.html for more details.
bigdata@bigdata-VirtualBox:~$ sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:7.6p1-4ubuntu0.3).
openssh-client set to manually installed.
The following additional packages will be installed:
  ncurses-term openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere rssh ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
0 upgraded, 4 newly installed, 0 to remove and 415 not upgraded.
Need to get 637 kB of archives.
After this operation, 5,316 kB of additional disk space will be used.
Get:1 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ncurses-term all 6.1-1ubuntu1.18.04 [248 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-sftp-server amd64 1:7.6p1-4ubuntu0.3 [45.6 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-server amd64 1:7.6p1-4ubuntu0.3 [333 kB]
Get:4 http://in.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ssh-import-id all 5.7-0ubuntu1.1 [10.9 kB]
Fetched 637 kB in 1s (901 kB/s)
```

Create hadoop user

To create user “add user ” command to create a new hadoop user:

```
Sudo adduser hduser
```

```
bigdata@bigdata-VirtualBox:~$ sudo adduser hduser
Adding user `hduser' ...
Adding new group `hduser' (1001) ...
Adding new user `hduser' (1001) with group `hduser' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
```

The username, in this example, is hduser. You are free to use any username and password you see fit. Switch to the newly created user and enter the corresponding password:

```
su - hduser
```

To enable the passwordless SSH for the hduser(hadoop) user :
To generate ssh keypair and define the location to store

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
```

The system proceeds to generate and save the SSH key pair.

```
hduser@bigdata-VirtualBox:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:kDrFTIgtBIJhuLwVGtAeSWgRgbVBVyxIzPBapFOOY3w hduser@bigdata-VirtualBox
The key's randomart image is:
+---[RSA 2048]----+
|@^&=.+o          |
|OXX=++..         |
|B=BE..*          |
|oBo. o .         |
|. o o   S        |
| .   .           |
|                 |
|                 |
|                 |
+----[SHA256]-----+
```

Use the cat command to store the public key as authorized_keys in the ssh directory:

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

Set the permissions for your user with the chmod command:

```
chmod 0600 ~/.ssh/authorized_keys
```

the hduser user to SSH to localhost:

```
ssh localhost
```

Installation of the Hadoop

Use the below command to install the apache hadoop with wget command:

```
wget
https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
```

```
hduser@bigdata-VirtualBox:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
--2020-12-06 22:33:37--  https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 359196911 (343M) [application/x-gzip]
Saving to: 'hadoop-3.2.1.tar.gz'

hadoop-3.2.1.tar.gz 100%[===================>] 342.56M  4.18MB/s    in 97s

2020-12-06 22:35:14 (3.53 MB/s) - 'hadoop-3.2.1.tar.gz' saved [359196911/359196911]
```

After download is complete , extract the hadoop files to install

```
tar xzf hadoop-3.2.1.tar.gz
```

```
hduser@bigdata-VirtualBox:~$ tar xzf hadoop-3.2.1.tar.gz
```

To configure hadoop on single node :

Configure the hadoop environment variables in bashrc
Edit the .bashrc shell configuration in editor using below command:

```
sudo nano .bashrc
```

Update the hadoop environment variable by adding below content

```
#Hadoop Related Options
export HADOOP_HOME=/home/hduser/hadoop-3.2.1
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

After adding the variables, save and exit the .bashrc file.
Using below command:

```
source ~/.bashrc
```

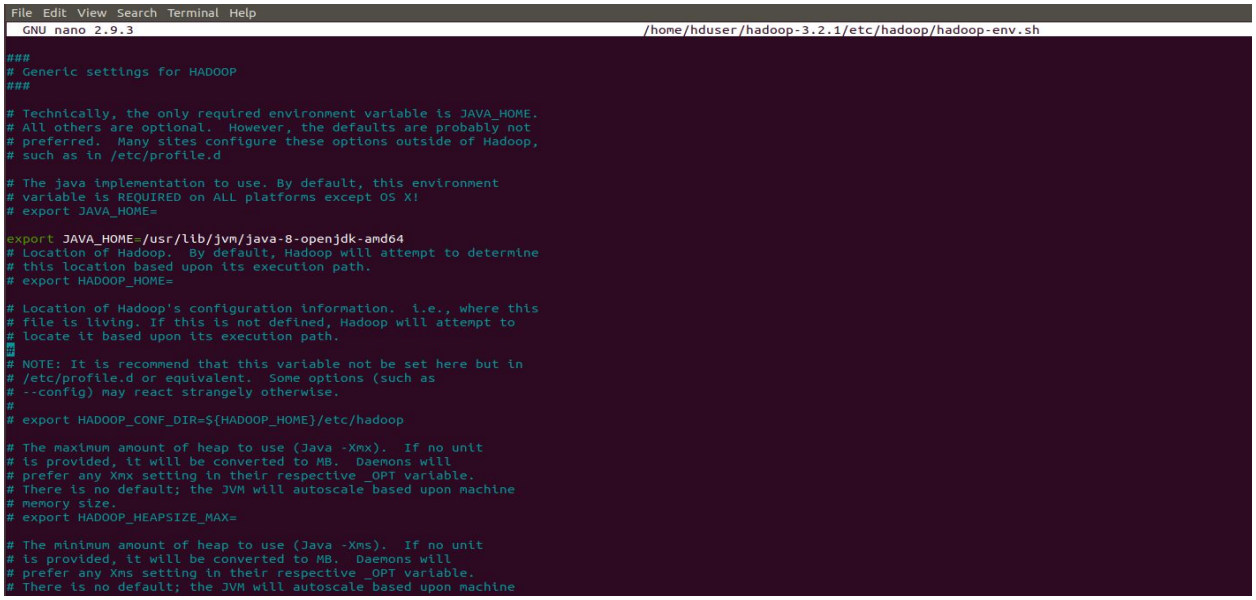
To edit hadoop-env.sh file:

The hadoop-env.sh file serves as a master file to configure YARN, HDFS, MapReduce and Hadoop-related project settings.

```
sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

Then update the java home by adding follow line

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

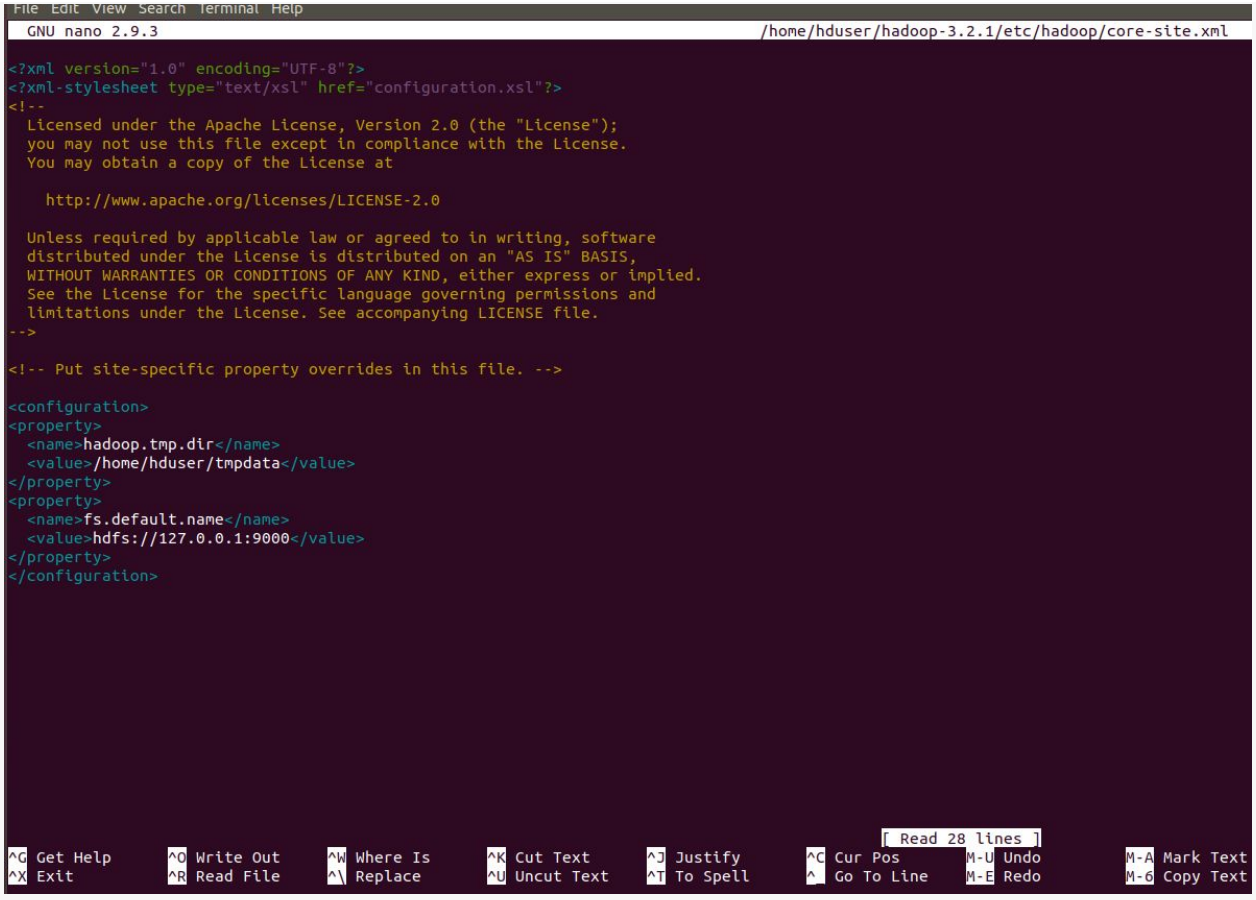


Edit core-site.xml File

The core-site.xml file defines HDFS and Hadoop core properties
Open the core-site.xml file and edit and add the properties

```
sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hduser/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```



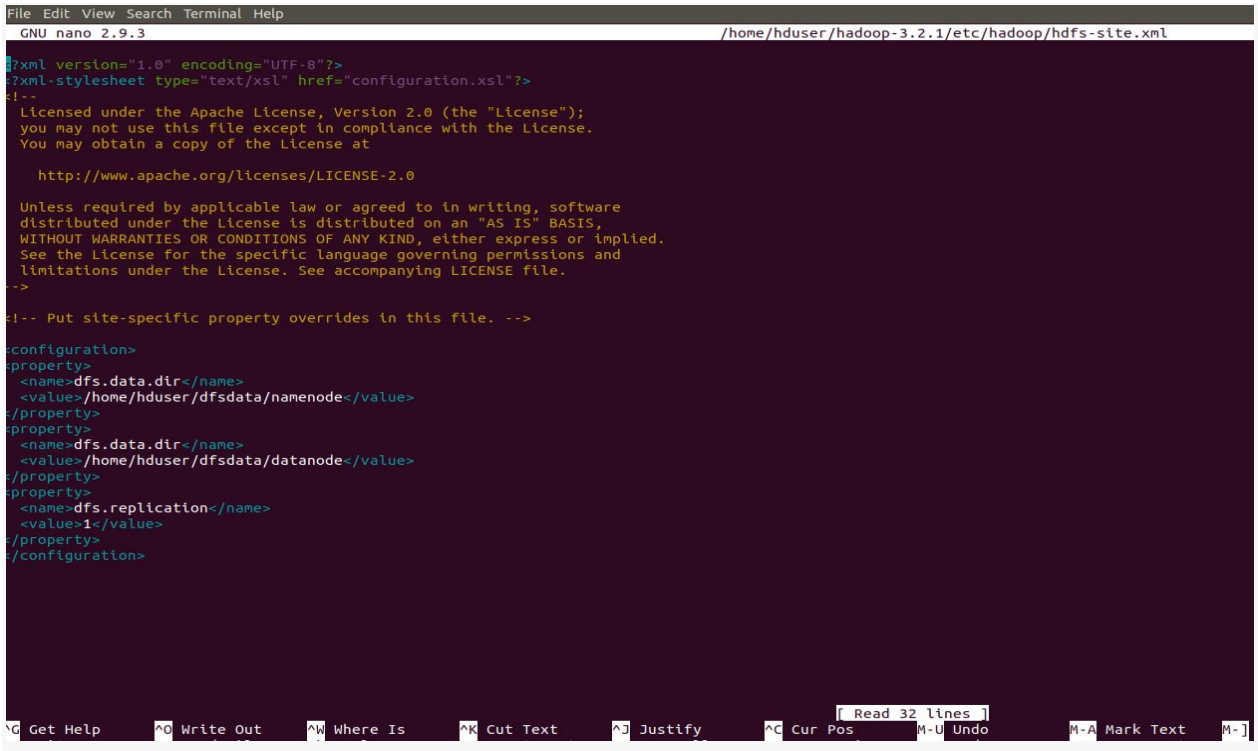
Edit hdfs-site.xml File :

The properties in the hdfs-site.xml file govern the location for storing node metadata, fsimage file, and edit log file. Configure the file by defining the NameNode and DataNode storage directorie
Use the following command to open the hdfs-site.xml file for editing:

```
sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

```
<configuration>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hduser/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hduser/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
```

```
  <value>1</value>
</property>
</configuration>
```



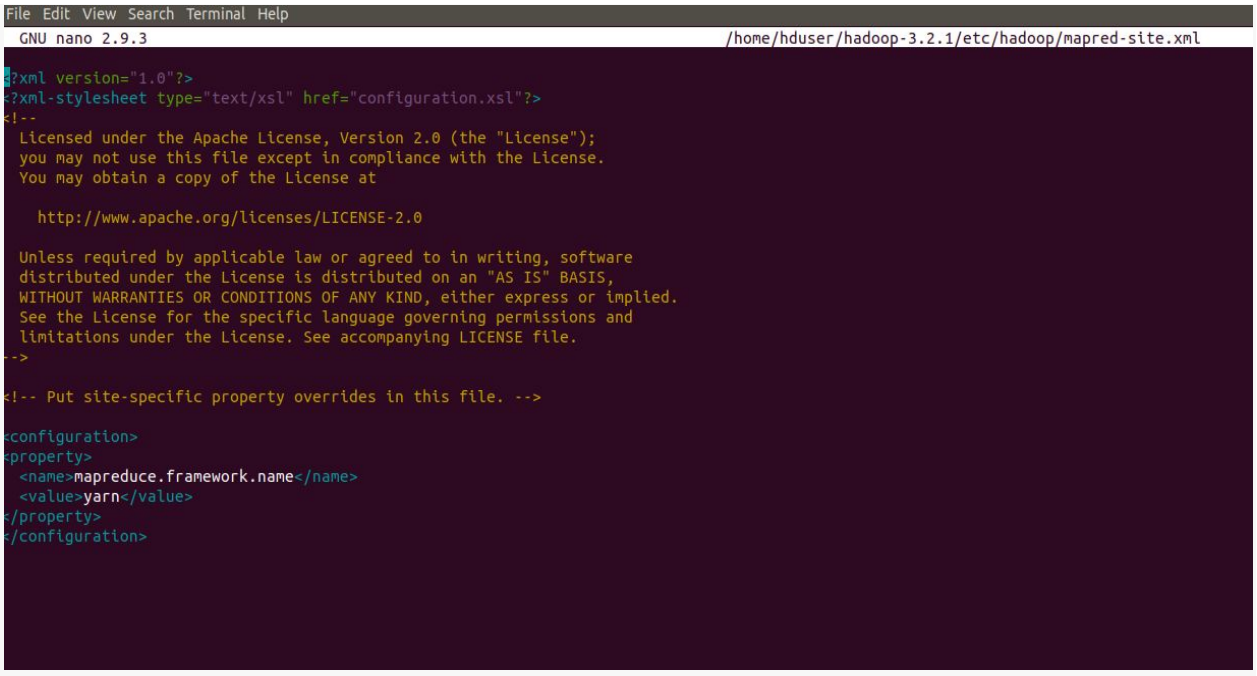
Edit mapred-site.xml File

Use the following command to access the *mapred-site.xml* file and define MapReduce values:

```
sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

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

```
</configuration>
```



Edit yarn-site.xml File

The yarn-site.xml file is used to define settings relevant to YARN. It contains configurations for the Node Manager, Resource Manager, Containers, and Application Master. Open the yarn-site.xml file in a text editor:

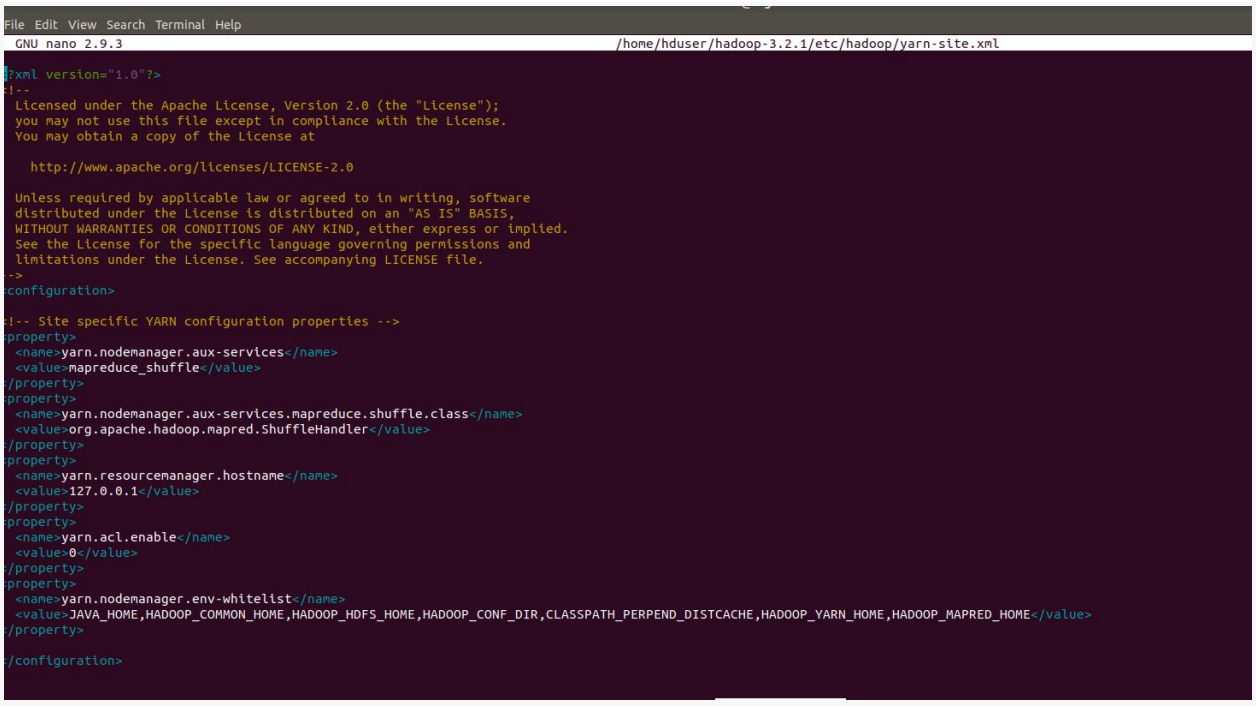
```
sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
```

Append the following configuration to the file:

```
<configuration>
<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>
<property>
```

```
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSP
ATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```



Format HDFS NameNode :

It is important to format the NameNode before starting Hadoop services for the first time:

```
hdfs namenode -format
```

```
hduser@bigdata-VirtualBox:~$ hdfs namenode -format
WARNING: /home/hduser/hadoop-3.2.1/logs does not exist. Creating.
2020-12-06 23:04:09,046 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = bigdata-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /home/hduser/hadoop-3.2.1/etc/hadoop:/home/hduser/hadoop-3.2.
/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/hduser/ha
x2-api-3.1.4.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/audience-annotations-
/hadoop/common/lib/avro-1.7.7.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/zooke
share/hadoop/common/lib/commons-net-3.6.jar:/home/hduser/hadoop-3.2.1/share/hadoop/commo
ser/hadoop-3.2.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/hduser/hado
3.24.v20180605.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/paranamer-2.3.jar:/
dnsjava-2.1.7.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/jackson-jaxrs-1.9.13
ommon/lib/hadoop-auth-3.2.1.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/jackso
e/hadoop/common/lib/gson-2.2.4.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib/ker
/share/hadoop/common/lib/jetty-webapp-9.3.24.v20180605.jar:/home/hduser/hadoop-3.2.1/sha
empty-to-avoid-conflict-with-guava.jar:/home/hduser/hadoop-3.2.1/share/hadoop/common/lib
```

Start Hadoop Cluster

Goto the hadoop install directory and execute the below command to start the namenode and datanode

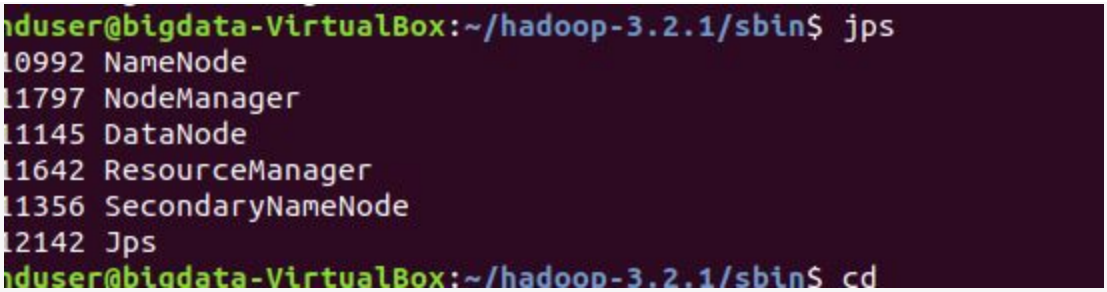
```
./start-dfs.sh
```

```
hduser@bigdata-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bigdata-VirtualBox]
bigdata-VirtualBox: Warning: Permanently added 'bigdata-virtualbox' (ECDSA) to the list of known hosts.
```

To start the YARN resource and nodemanagers

```
./start-yarn.sh
```

```
bigdata-VirtualBox: Warning: Permanently added 'bigdata-virtualbox' (ECDSA) to t
hduser@bigdata-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```



Access Hadoop UI from Browser

The default port number 9870 gives you access to the Hadoop NameNode UI:

```
http://localhost:9870
```