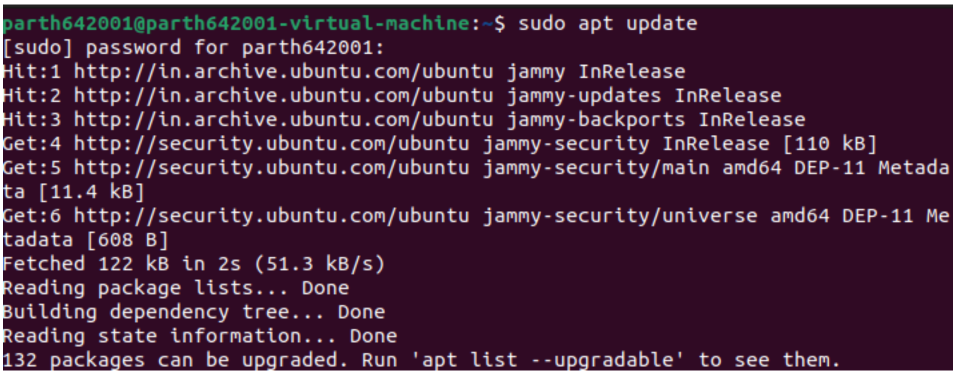
PRACTICAL-1

AIM:

To install Hadoop framework, configure it and setup a single node cluster. Use web based tools to monitor your Hadoop setup

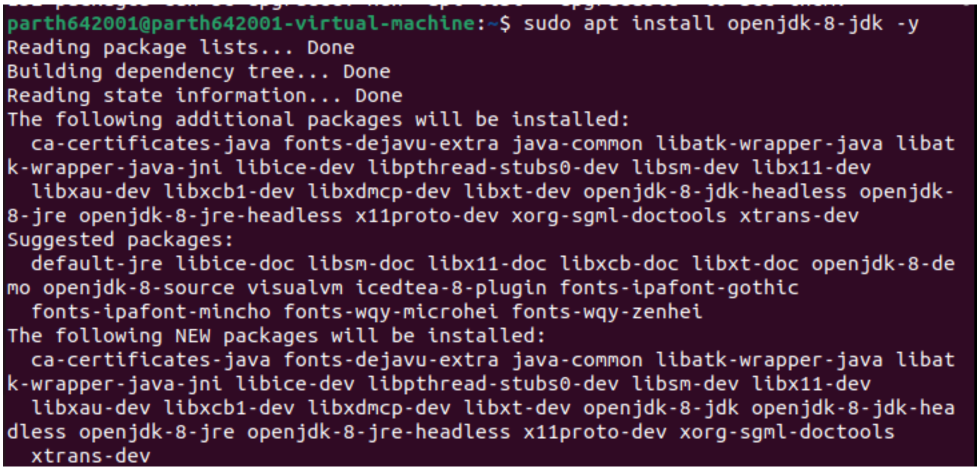
IMPLEMENTATION:

The Hadoop framework is written in Java, and its services require a compatible Java Runtime Environment (JRE) and Java Development Kit (JDK).



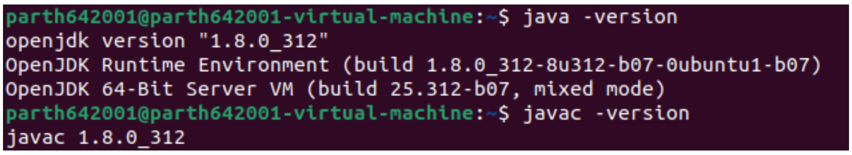
```
parth642001@parth642001-virtual-machine:~$ sudo apt update
[sudo] password for parth642001:
Hit:1 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://in.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:5 http://security.ubuntu.com/ubuntu jammy-security/main amd64 DEP-11 Metada
ta [11.4 kB]
Get:6 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 DEP-11 Me
tadata [608 B]
Fetched 122 kB in 2s (51.3 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
132 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

Apache Hadoop 3.x fully supports Java 8. The OpenJDK 8 package in Ubuntu contains both the runtime environment and development kit.



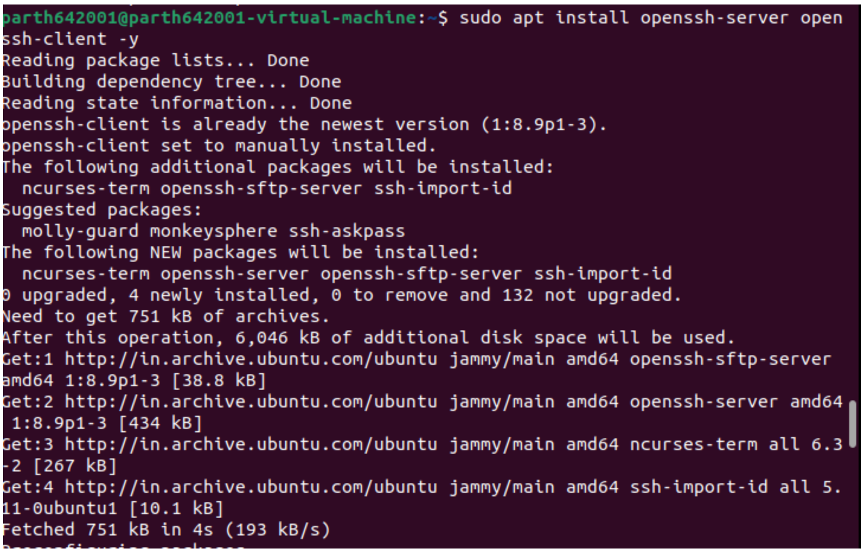
```
parth642001@parth642001-virtual-machine:~$ sudo apt install openjdk-8-jdk -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libat
k-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-de
mo openjdk-8-source visualvm icedtea-8-plugin fonts-ipafont-gothic
  fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libat
k-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-hea
dless openjdk-8-jre openjdk-8-jre-headless x11proto-dev xorg-sgml-doctools
  xtrans-dev
```

The OpenJDK or Oracle Java version can affect how elements of a Hadoop ecosystem interact. Hence, we need to be specific.



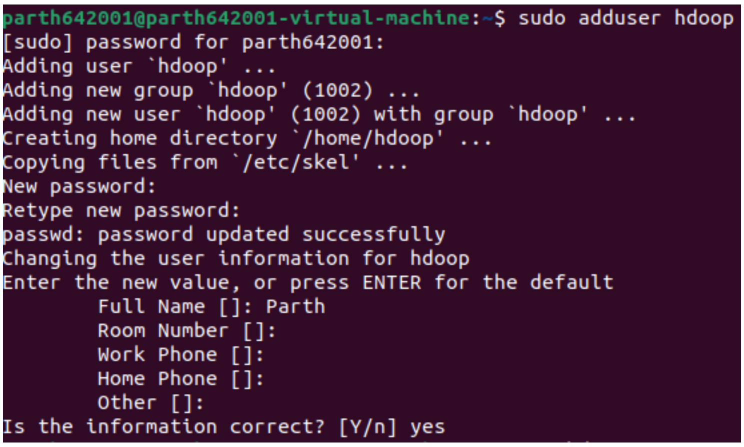
```
parth642001@parth642001-virtual-machine:~$ java -version
openjdk version "1.8.0_312"
OpenJDK Runtime Environment (build 1.8.0_312-8u312-b07-0ubuntu1-b07)
OpenJDK 64-Bit Server VM (build 25.312-b07, mixed mode)
parth642001@parth642001-virtual-machine:~$ javac -version
javac 1.8.0_312
```

Install the OpenSSH server and client.



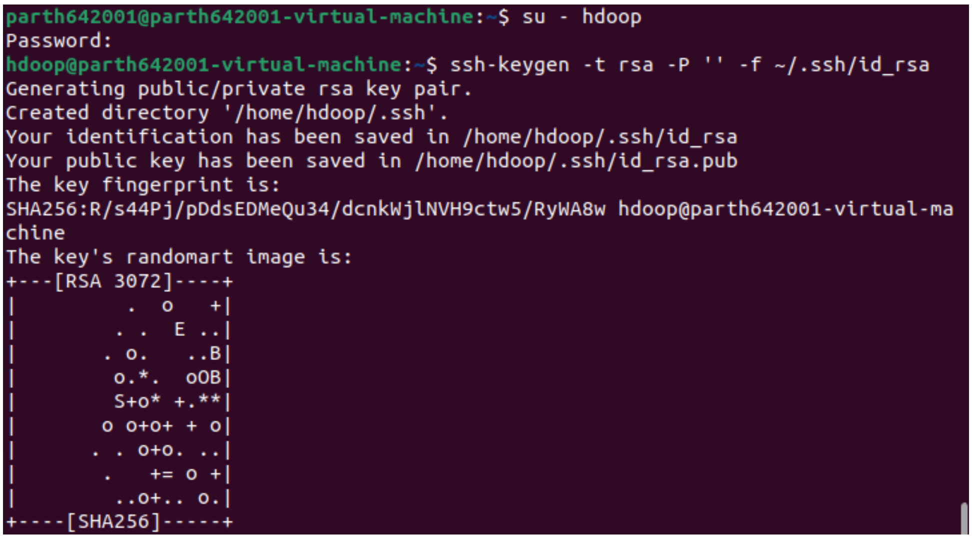
```
parth642001@parth642001-virtual-machine:~$ sudo apt install openssh-server open
ssh-client -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-client is already the newest version (1:8.9p1-3).
openssh-client set to manually installed.
The following additional packages will be installed:
  ncurses-term openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere 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 132 not upgraded.
Need to get 751 kB of archives.
After this operation, 6,046 kB of additional disk space will be used.
Get:1 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 openssh-sftp-server
amd64 1:8.9p1-3 [38.8 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 openssh-server amd64
 1:8.9p1-3 [434 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 ncurses-term all 6.3
-2 [267 kB]
Get:4 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 ssh-import-id all 5.
11-0ubuntu1 [10.1 kB]
Fetched 751 kB in 4s (193 kB/s)
```

Utilize the adduser command to create a new Hadoop user. The username, in this example, is **hdoop**. You are free the use any username and password you see fit. Switch to the newly created user and enter the corresponding password.



```
parth642001@parth642001-virtual-machine:~$ sudo adduser hdoop
[sudo] password for parth642001:
Adding user `hdoop' ...
Adding new group `hdoop' (1002) ...
Adding new user `hdoop' (1002) with group `hdoop' ...
Creating home directory `/home/hdoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hdoop
Enter the new value, or press ENTER for the default
        Full Name []: Parth
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] yes
```

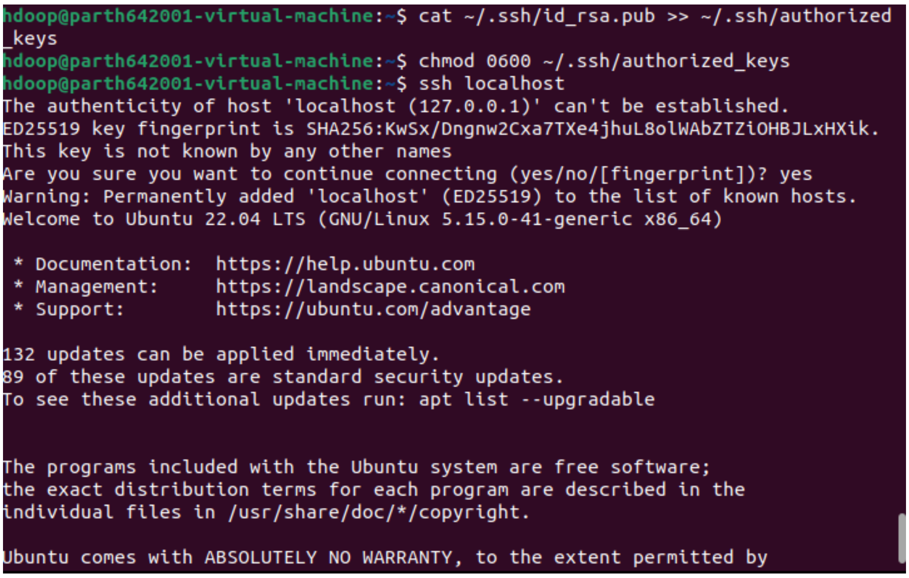
The user now needs to be able to SSH to the localhost without being prompted for a password.



```
parth642001@parth642001-virtual-machine:~$ su - hdoop
Password:
hdoop@parth642001-virtual-machine:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hdoop/.ssh'.
Your identification has been saved in /home/hdoop/.ssh/id_rsa
Your public key has been saved in /home/hdoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:R/s44Pj/pDdsEDMeQu34/dcnkWjlNVH9ctw5/RyWA8w hdoop@parth642001-virtual-ma
chine
The key's randomart image is:
+---[RSA 3072]----+
|         .  o   +|
|        . .  E ..|
|       . o.   ..B|
|        o.*.  oOB|
|        S+o* +.**|
|       o o+o+ + o|
|      . . o+o. ..|
|       .   += o +|
|        ..o+.. o.|
+----[SHA256]-----+
```

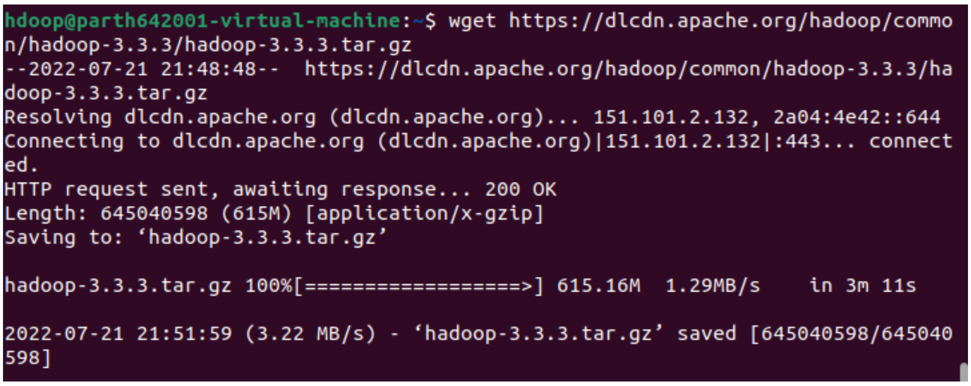
Use the cat command to store the public key as **authorized_keys** in the *ssh* directory

Set the permissions for your user with the chmod command

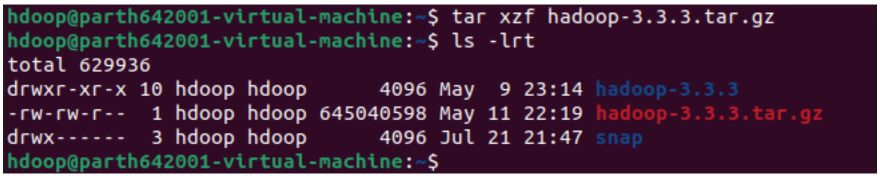


```
hdoop@parth642001-virtual-machine:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized
_keys
hdoop@parth642001-virtual-machine:~$ chmod 0600 ~/.ssh/authorized_keys
hdoop@parth642001-virtual-machine:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:KwSx/Dngnw2Cxa7TXe4jhuL8olWAbZTZiOHBJLxHXik.
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.
Welcome to Ubuntu 22.04 LTS (GNU/Linux 5.15.0-41-generic x86_64)

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

132 updates can be applied immediately.
89 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable


The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
```

Download and extract the Hadoop setup.



```
hdoop@parth642001-virtual-machine:~$ wget https://dlcdn.apache.org/hadoop/commo
n/hadoop-3.3.3/hadoop-3.3.3.tar.gz
--2022-07-21 21:48:48--  https://dlcdn.apache.org/hadoop/common/hadoop-3.3.3/ha
doop-3.3.3.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... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 645040598 (615M) [application/x-gzip]
Saving to: ‘hadoop-3.3.3.tar.gz’

hadoop-3.3.3.tar.gz 100%[=================>] 615.16M  1.29MB/s    in 3m 11s

2022-07-21 21:51:59 (3.22 MB/s) - ‘hadoop-3.3.3.tar.gz’ saved [645040598/645040
598]
```



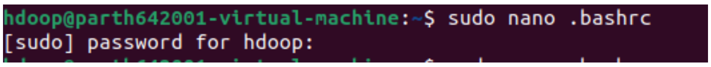
```
hdoop@parth642001-virtual-machine:~$ tar xzf hadoop-3.3.3.tar.gz
hdoop@parth642001-virtual-machine:~$ ls -lrt
total 629936
drwxr-xr-x 10 hdoop hdoop      4096 May  9 23:14 hadoop-3.3.3
-rw-rw-r--  1 hdoop hdoop 645040598 May 11 22:19 hadoop-3.3.3.tar.gz
drwx------  3 hdoop hdoop      4096 Jul 21 21:47 snap
hdoop@parth642001-virtual-machine:~$
```

Hadoop excels when deployed in a fully distributed mode on a large cluster of networked servers. However, if you are new to Hadoop and want to explore basic commands or test applications, you can configure Hadoop on a single node.

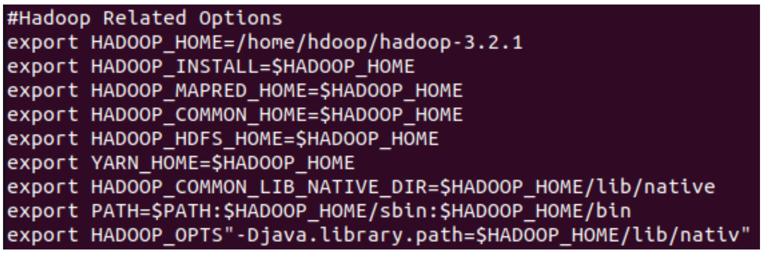
This setup, also called pseudo-distributed mode, allows each Hadoop daemon to run as a single Java process. A Hadoop environment is configured by editing a set of configuration files:

- bashrc
- hadoop-env.sh
- core-site.xml
- hdfs-site.xml
- mapred-site-xml

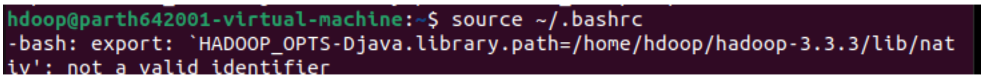
- yarn-site.xml



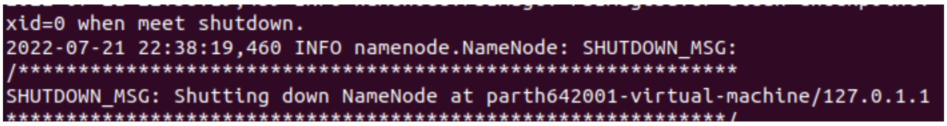
```
hdoop@parth642001-virtual-machine:~$ sudo nano .bashrc
[sudo] password for hdoop:
```



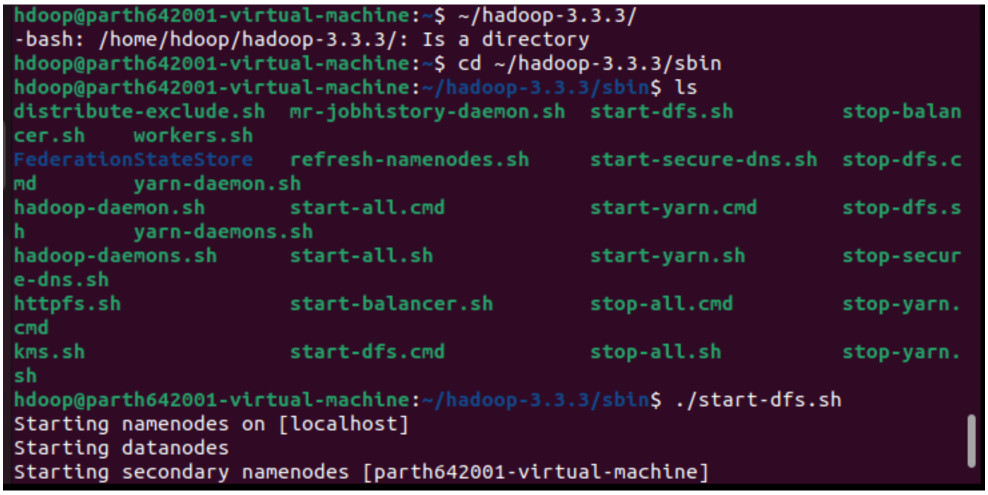
```
#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-3.2.1
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native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



```
hdoop@parth642001-virtual-machine:~$ source ~/.bashrc
-bash: export: `HADOOP_OPTS-Djava.library.path=/home/hdoop/hadoop-3.3.3/lib/nat
iv': not a valid identifier
```



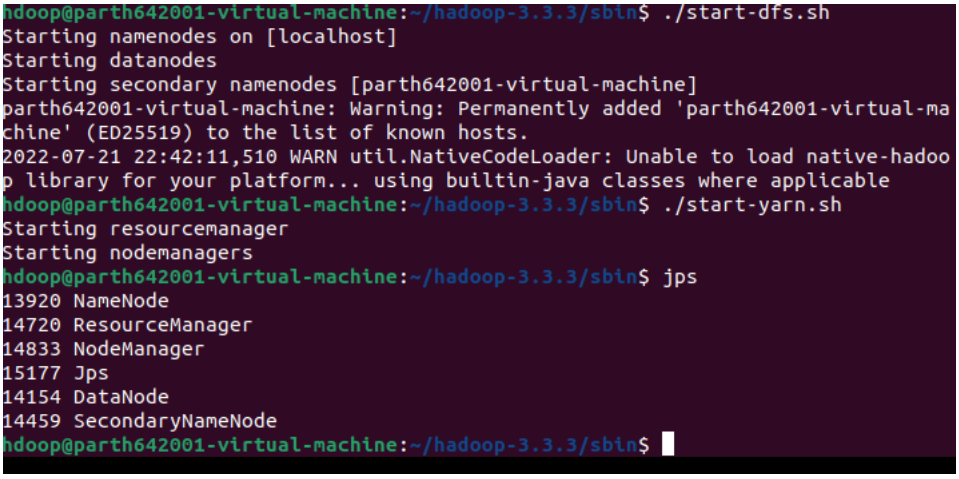
```
xid=0 when meet shutdown.
2022-07-21 22:38:19,460 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at parth642001-virtual-machine/127.0.1.1
************************************************************/
```



```
hdoop@parth642001-virtual-machine:~$ ~/hadoop-3.3.3/
-bash: /home/hdoop/hadoop-3.3.3/: Is a directory
hdoop@parth642001-virtual-machine:~$ cd ~/hadoop-3.3.3/sbin
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$ ls
distribute-exclude.sh  mr-jobhistory-daemon.sh  start-dfs.sh         stop-balan
cer.sh     workers.sh
FederationStateStore   refresh-namenodes.sh     start-secure-dns.sh  stop-dfs.c
md         yarn-daemon.sh
hadoop-daemon.sh       start-all.cmd            start-yarn.cmd       stop-dfs.s
h          yarn-daemons.sh
hadoop-daemons.sh      start-all.sh             start-yarn.sh        stop-secur
e-dns.sh
httpfs.sh              start-balancer.sh        stop-all.cmd         stop-yarn.
cmd
kms.sh                 start-dfs.cmd            stop-all.sh          stop-yarn.
sh
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [parth642001-virtual-machine]
```

Navigate to the *hadoop-3.2.1/sbin* directory and execute the following commands to start the NameNode and DataNode.

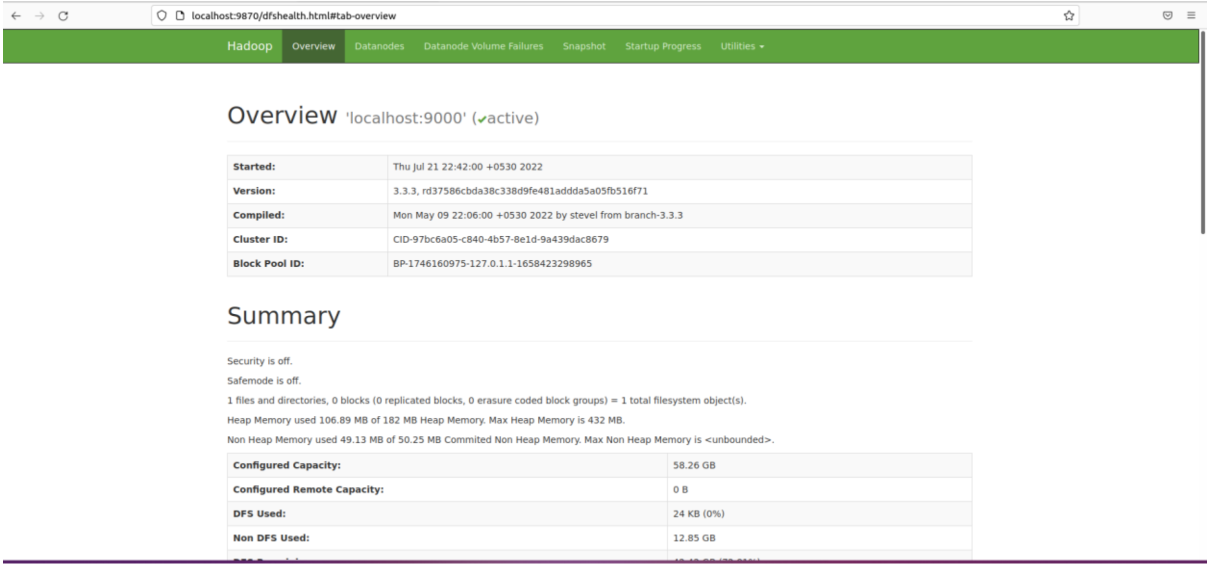
Once the namenode, datanodes, and secondary namenode are up and running, start the YARN resource and nodemanagers



```
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [parth642001-virtual-machine]
parth642001-virtual-machine: Warning: Permanently added 'parth642001-virtual-ma
chine' (ED25519) to the list of known hosts.
2022-07-21 22:42:11,510 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$ jps
13920 NameNode
14720 ResourceManager
14833 NodeManager
15177 Jps
14154 DataNode
14459 SecondaryNameNode
hdoop@parth642001-virtual-machine:~/hadoop-3.3.3/sbin$
```

One can also use localhost to access the Hadoop overview.

http://localhost:9870



localhost:9870/dfshealth.html#tab-overview

Overview 'localhost:9000' (✓active)

Started:	Thu Jul 21 22:42:00 +0530 2022
Version:	3.3.3, rd37586cbda38c338d9fe481addda5a05fb516f71
Compiled:	Mon May 09 22:06:00 +0530 2022 by stevel from branch-3.3.3
Cluster ID:	CID-97bc6a05-c840-4b57-8e1d-9a439dac8679
Block Pool ID:	BP-1746160975-127.0.1.1-1658423298965

Summary

Security is off.

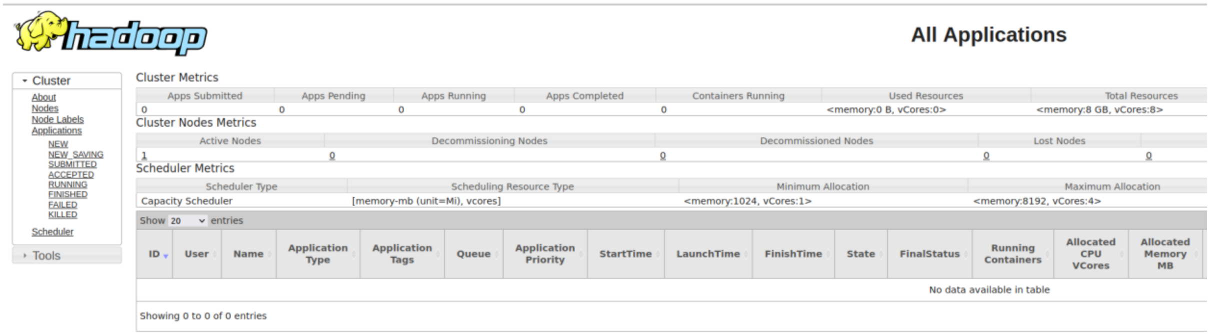
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 106.89 MB of 182 MB Heap Memory. Max Heap Memory is 432 MB.

Non Heap Memory used 49.13 MB of 50.25 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	58.26 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	12.85 GB



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:8>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	
1	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB
No data available in table														

Showing 0 to 0 of 0 entries

CONCLUSION:

By performing this practical, I learnt how to install and configure Hadoop.