Setting Up Hadoop on Ubuntu: A Step-by-Step Guide

Introduction:

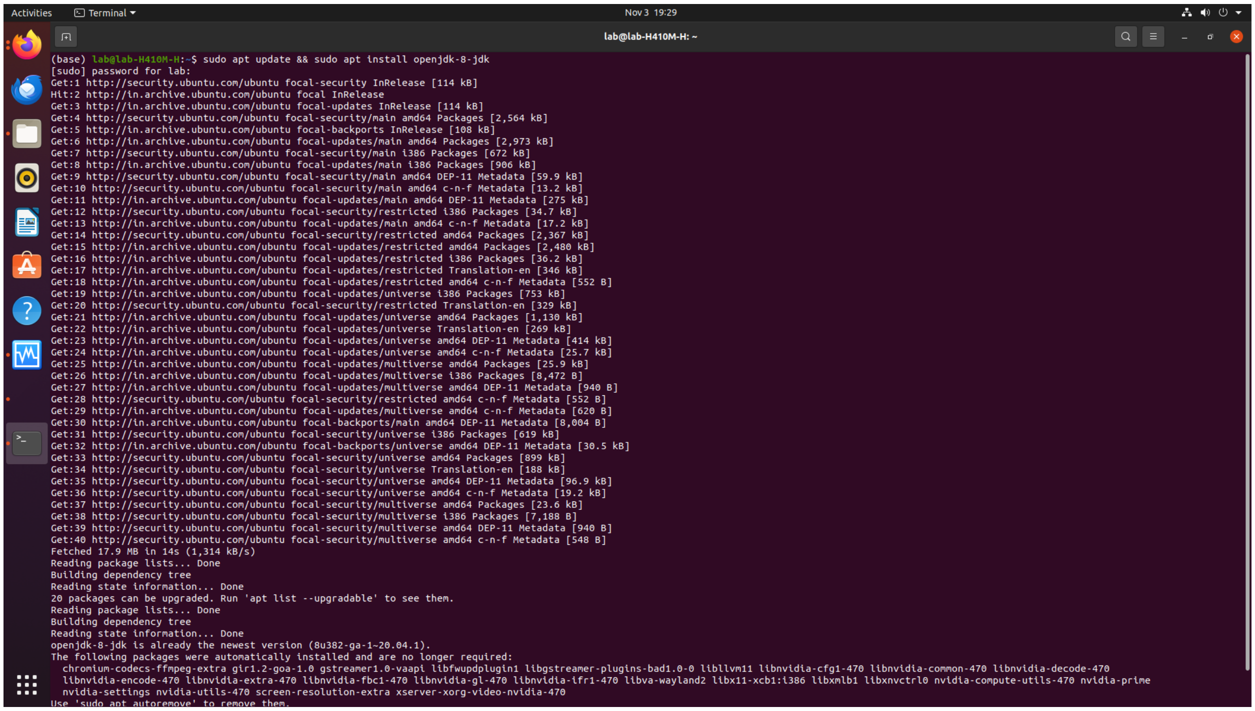
Hadoop is a powerful framework for distributed data storage and processing, and setting it up on your own system can be a valuable skill. In this guide, we'll walk you through the process of installing and configuring Hadoop on an Ubuntu machine, step by step.

Prerequisites:

Before we dive into the installation process, make sure you have the following prerequisites in place:
An Ubuntu system
Basic familiarity with the Linux command line

Step 1: Installing Java Development Kit

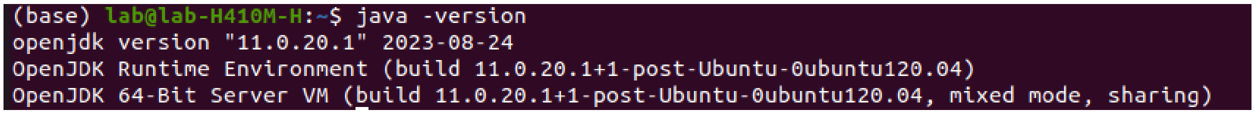
The first step in setting up Hadoop is to ensure you have the right Java version installed. Hive, a key component of the Hadoop ecosystem, requires Java 8. Here's how you can install it:
Command:- sudo apt update && sudo apt install openjdk-8-jdk



```
(base) lab@lab-H410M-H:~$ sudo apt update && sudo apt install openjdk-8-jdk
[sudo] password for lab:
Get:1 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Hit:2 http://in.archive.ubuntu.com/ubuntu focal InRelease
Get:3 http://in.archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:4 http://security.ubuntu.com/ubuntu focal-security/main amd64 Packages [2,564 kB]
Get:5 http://in.archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Get:6 http://in.archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [2,973 kB]
Get:7 http://security.ubuntu.com/ubuntu focal-security/main i386 Packages [672 kB]
Get:8 http://in.archive.ubuntu.com/ubuntu focal-updates/main i386 Packages [906 kB]
Get:9 http://security.ubuntu.com/ubuntu focal-security/main amd64 DEP-11 Metadata [59.9 kB]
Get:10 http://security.ubuntu.com/ubuntu focal-security/main amd64 c-n-f Metadata [13.2 kB]
Get:11 http://in.archive.ubuntu.com/ubuntu focal-updates/main amd64 DEP-11 Metadata [275 kB]
Get:12 http://security.ubuntu.com/ubuntu focal-security/restricted i386 Packages [34.7 kB]
Get:13 http://in.archive.ubuntu.com/ubuntu focal-updates/main amd64 c-n-f Metadata [17.2 kB]
Get:14 http://security.ubuntu.com/ubuntu focal-security/restricted amd64 Packages [2,367 kB]
Get:15 http://in.archive.ubuntu.com/ubuntu focal-updates/restricted amd64 Packages [2,480 kB]
Get:16 http://in.archive.ubuntu.com/ubuntu focal-updates/restricted i386 Packages [36.2 kB]
Get:17 http://in.archive.ubuntu.com/ubuntu focal-updates/restricted Translation-en [346 kB]
Get:18 http://in.archive.ubuntu.com/ubuntu focal-updates/restricted amd64 c-n-f Metadata [552 B]
Get:19 http://in.archive.ubuntu.com/ubuntu focal-updates/universe i386 Packages [753 kB]
Get:20 http://security.ubuntu.com/ubuntu focal-security/restricted Translation-en [329 kB]
Get:21 http://in.archive.ubuntu.com/ubuntu focal-updates/universe amd64 Packages [1,130 kB]
Get:22 http://in.archive.ubuntu.com/ubuntu focal-updates/universe Translation-en [269 kB]
Get:23 http://in.archive.ubuntu.com/ubuntu focal-updates/universe amd64 DEP-11 Metadata [414 kB]
Get:24 http://in.archive.ubuntu.com/ubuntu focal-updates/universe amd64 c-n-f Metadata [25.7 kB]
Get:25 http://in.archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 Packages [25.9 kB]
Get:26 http://in.archive.ubuntu.com/ubuntu focal-updates/multiverse i386 Packages [8,472 B]
Get:27 http://in.archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 DEP-11 Metadata [940 B]
Get:28 http://security.ubuntu.com/ubuntu focal-security/restricted amd64 c-n-f Metadata [552 B]
Get:29 http://in.archive.ubuntu.com/ubuntu focal-updates/multiverse amd64 c-n-f Metadata [620 B]
Get:30 http://in.archive.ubuntu.com/ubuntu focal-backports/main amd64 DEP-11 Metadata [8,004 B]
Get:31 http://security.ubuntu.com/ubuntu focal-security/universe i386 Packages [619 kB]
Get:32 http://in.archive.ubuntu.com/ubuntu focal-backports/universe amd64 DEP-11 Metadata [30.5 kB]
Get:33 http://security.ubuntu.com/ubuntu focal-security/universe amd64 Packages [899 kB]
Get:34 http://security.ubuntu.com/ubuntu focal-security/universe Translation-en [188 kB]
Get:35 http://security.ubuntu.com/ubuntu focal-security/universe amd64 DEP-11 Metadata [96.9 kB]
Get:36 http://security.ubuntu.com/ubuntu focal-security/universe amd64 c-n-f Metadata [19.2 kB]
Get:37 http://security.ubuntu.com/ubuntu focal-security/multiverse amd64 Packages [23.6 kB]
Get:38 http://security.ubuntu.com/ubuntu focal-security/multiverse i386 Packages [7,188 B]
Get:39 http://security.ubuntu.com/ubuntu focal-security/multiverse amd64 DEP-11 Metadata [940 B]
Get:40 http://security.ubuntu.com/ubuntu focal-security/multiverse amd64 c-n-f Metadata [548 B]
Fetched 17.9 MB in 14s (1,314 kB/s)
Reading package lists... Done
Building dependency tree
Reading state information... Done
20 packages can be upgraded. Run 'apt list --upgradable' to see them.
Reading package lists... Done
Building dependency tree
Reading state information... Done
openjdk-8-jdk is already the newest version (8u382-ga-1~20.04.1).
The following packages were automatically installed and are no longer required:
  chromium-codecs-ffmpeg-extra gir1.2-goa-1.0 gstreamer1.0-vaapi libfwupdplugin1 libgstreamer-plugins-bad1.0-0 libllvm11 libnvidia-cfg1-470 libnvidia-common-470 libnvidia-decode-470
  libnvidia-encode-470 libnvidia-extra-470 libnvidia-fbc1-470 libnvidia-gl-470 libnvidia-ifr1-470 libva-wayland2 libx11-xcb1:i386 libxmlb1 libxnvctrl0 nvidia-compute-utils-470 nvidia-prime
  nvidia-settings nvidia-utils-470 screen-resolution-extra xserver-xorg-video-nvidia-470
Use 'sudo apt autoremove' to remove them.
```

Step 2: Verifying the Java Version

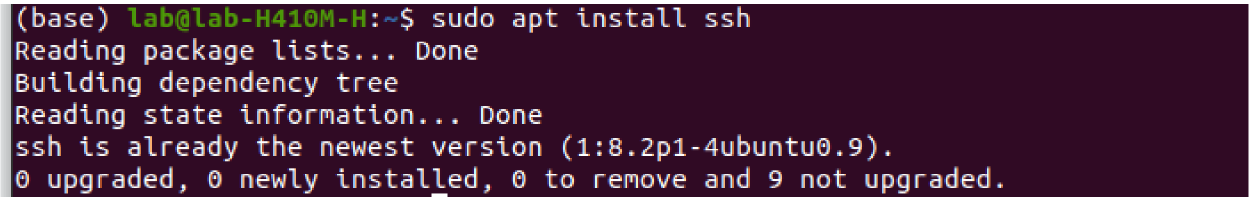
After installing Java, it's a good practice to check the installed version using:
Command:- java -version



```
(base) lab@lab-H410M-H:~$ java -version
openjdk version "11.0.20.1" 2023-08-24
OpenJDK Runtime Environment (build 11.0.20.1+1-post-Ubuntu-0ubuntu120.04)
OpenJDK 64-Bit Server VM (build 11.0.20.1+1-post-Ubuntu-0ubuntu120.04, mixed mode, sharing)
```

Step 3: Installing SSH

SSH is essential for secure communication in a Hadoop cluster. Install it with:
Command:- sudo apt install ssh



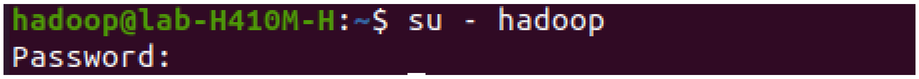
```
(base) lab@lab-H410M-H:~$ sudo apt install ssh
Reading package lists... Done
Building dependency tree
Reading state information... Done
ssh is already the newest version (1:8.2p1-4ubuntu0.9).
0 upgraded, 0 newly installed, 0 to remove and 9 not upgraded.
```

Step 4: Creating the Hadoop User

All Hadoop components run under a dedicated user. Create the 'hadoop' user:
Command:- sudo adduser hadoop

Step 5: Switching to the Hadoop User

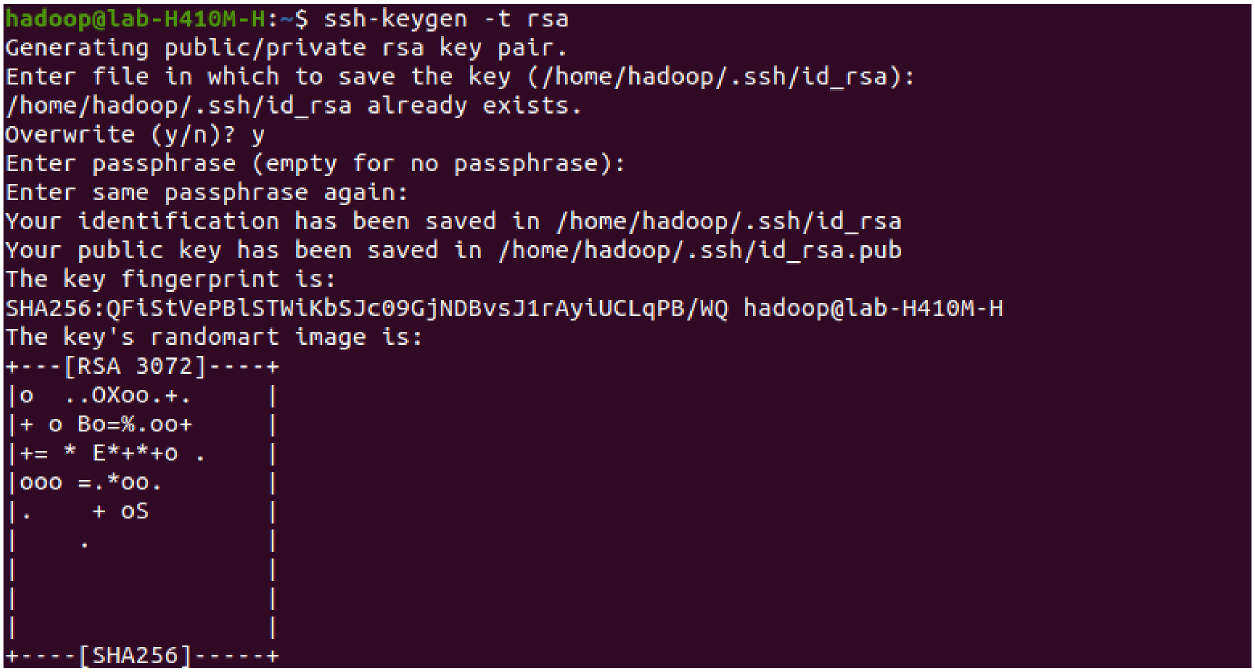
Switch to the newly created 'hadoop' user:
Command:- su - hadoop



```
hadoop@lab-H410M-H:~$ su - hadoop
Password:
```

Step 6: Configuring SSH for Passwordless Access

To enable passwordless SSH access for 'hadoop', generate an SSH key pair:
Command:- ssh-keygen -t rsa



```
hadoop@lab-H410M-H:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
/home/hadoop/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:QFiStVePBlSTWiKbSJc09GjNDBvsJ1rAyiUCLqPB/WQ hadoop@lab-H410M-H
The key's randomart image is:
+---[RSA 3072]----+
|o  ..OXoo.+.     |
|+ o Bo=%.oo+     |
|+= * E*+*+o .    |
|ooo =.*oo.       |
|.    + oS        |
|    .            |
|                 |
|                 |
|                 |
+----[SHA256]-----+
```

Step 7: Setting Permissions for SSH

Set permissions for the generated public keys:

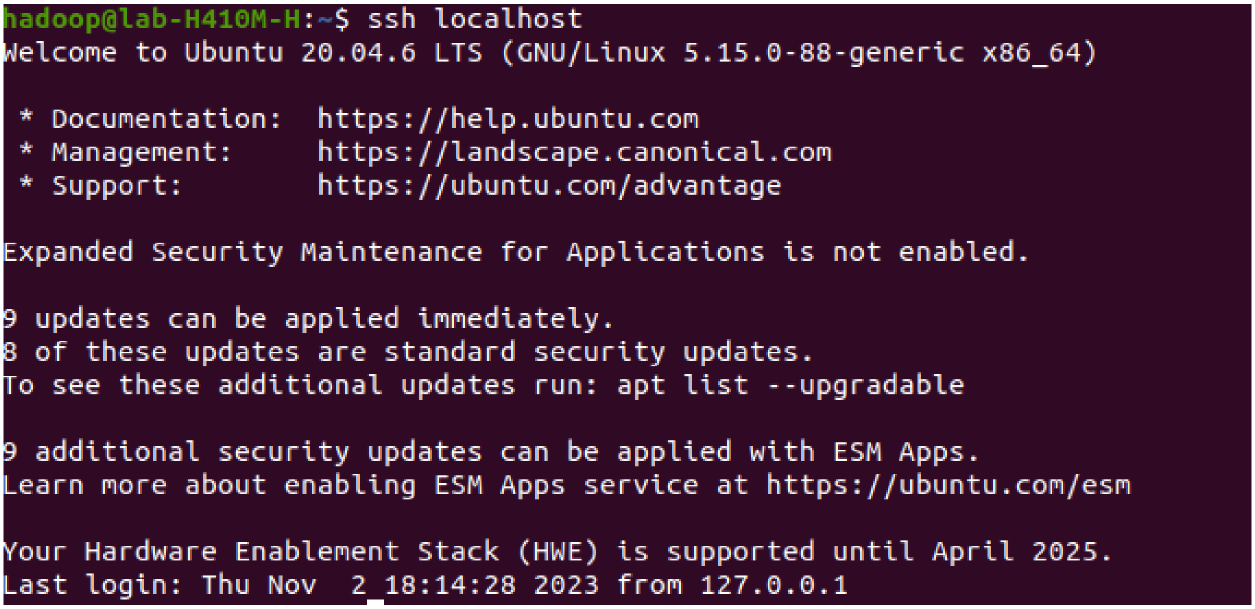
Command:- cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

chmod 640 ~/.ssh/authorized_keys

Step 8: Testing SSH to localhost

Test the SSH setup by connecting to localhost:

Command:- ssh localhost



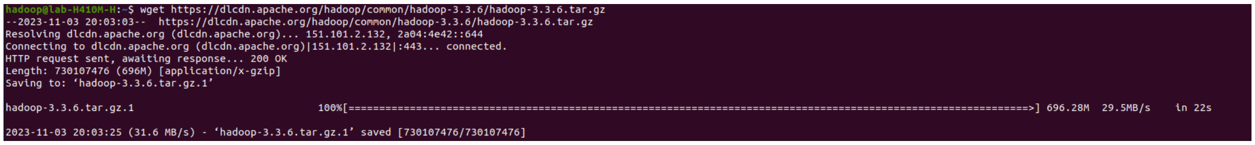
```
hadoop@lab-H410M-H:~$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-88-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

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

9 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Thu Nov  2 18:14:28 2023 from 127.0.0.1
```

Step 9: Switch Back to the Hadoop User

After successful SSH testing, switch back to the 'hadoop' user:
Command:- su - hadoop

Step 10: Installing Hadoop

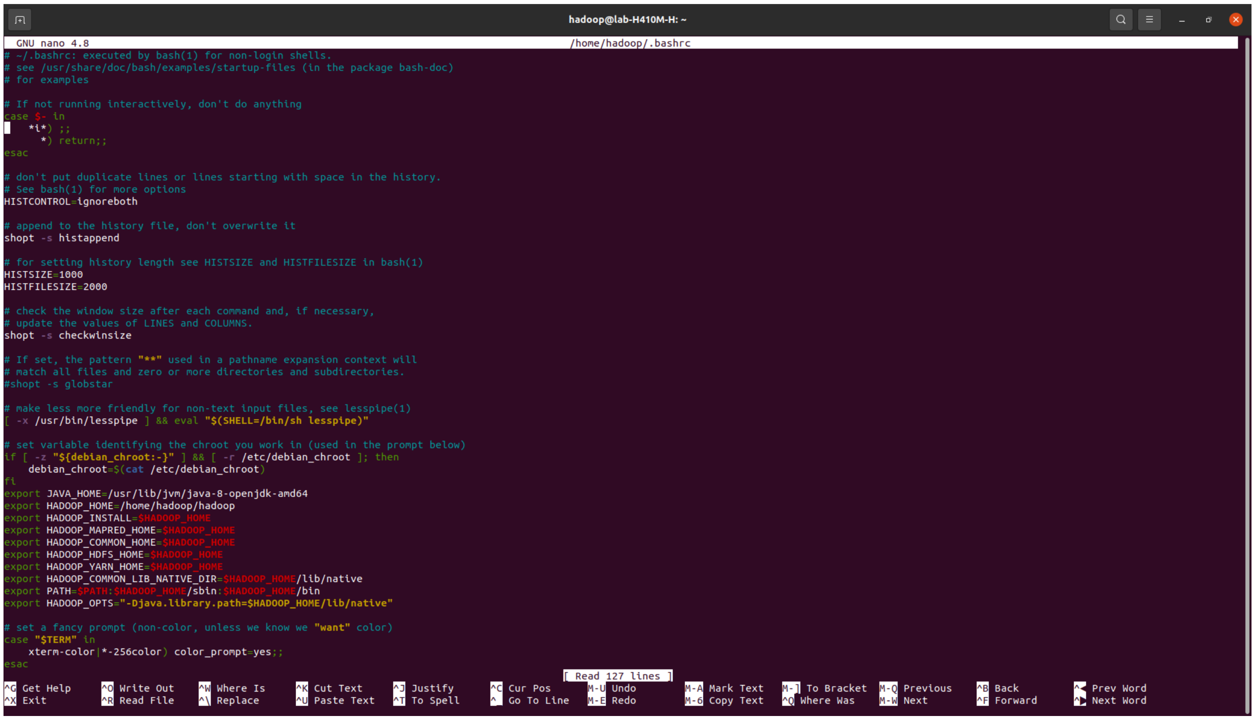
Download and install Hadoop 3.3.6
Command:- wget
https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xvzf hadoop-3.3.6.tar.gz
mv hadoop-3.3.6 hadoop



```
hadoop@lab-H410M-H:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
--2023-11-03 20:03:03--  https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.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: 730107476 (696M) [application/x-gzip]
Saving to: ‘hadoop-3.3.6.tar.gz.1’

hadoop-3.3.6.tar.gz.1                     100%[==============================================================================================================>] 696.28M  29.5MB/s    in 22s

2023-11-03 20:03:25 (31.6 MB/s) - ‘hadoop-3.3.6.tar.gz.1’ saved [730107476/730107476]
```

- In next step we need to configure Hadoop and Java Environment Variables on our system. Open the ~/.bashrc file in your favorite text editor.Here we are nano editor , to pasting the code we use ctrl+shift+v for saving the file ctrl+x and ctrl+y ,then hit enter:
  nano ~/.bashrc

- We need to append below lines to the file.
  export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
  export HADOOP_HOME=/home/hadoop/hadoop
  export HADOOP_INSTALL=$HADOOP_HOME
  export HADOOP_MAPRED_HOME=$HADOOP_HOME
  export HADOOP_COMMON_HOME=$HADOOP_HOME
  export HADOOP_HDFS_HOME=$HADOOP_HOME
  export HADOOP_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"



```
  GNU nano 4.8                                  /home/hadoop/.bashrc
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# 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 history.
# 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(1)
HISTSIZE=1000
HISTFILESIZE=2000

# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize

# If set, the pattern "**" used in a pathname expansion context will
# match all files and zero or more directories and subdirectories.
#shopt -s globstar

# make less more friendly for non-text input files, see lesspipe(1)
[ -x /usr/bin/lesspipe ] && eval "$(SHELL=/bin/sh lesspipe)"

# set variable identifying the chroot you work in (used in the prompt below)
if [ -z "${debian_chroot:-}" ] && [ -r /etc/debian_chroot ]; then
    debian_chroot=$(cat /etc/debian_chroot)
fi
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_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"

# set a fancy prompt (non-color, unless we know we "want" color)
case "$TERM" in
    xterm-color|*-256color) color_prompt=yes;;
esac
                                          [ Read 127 lines ]
^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text  M-] To Bracket M-Q Previous   ^B Back        ^◀ Prev Word
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text  ^Q Where Was   M-W Next       ^F Forward     ^▶ Next Word
```

Step 11: Configuring Environment Variables

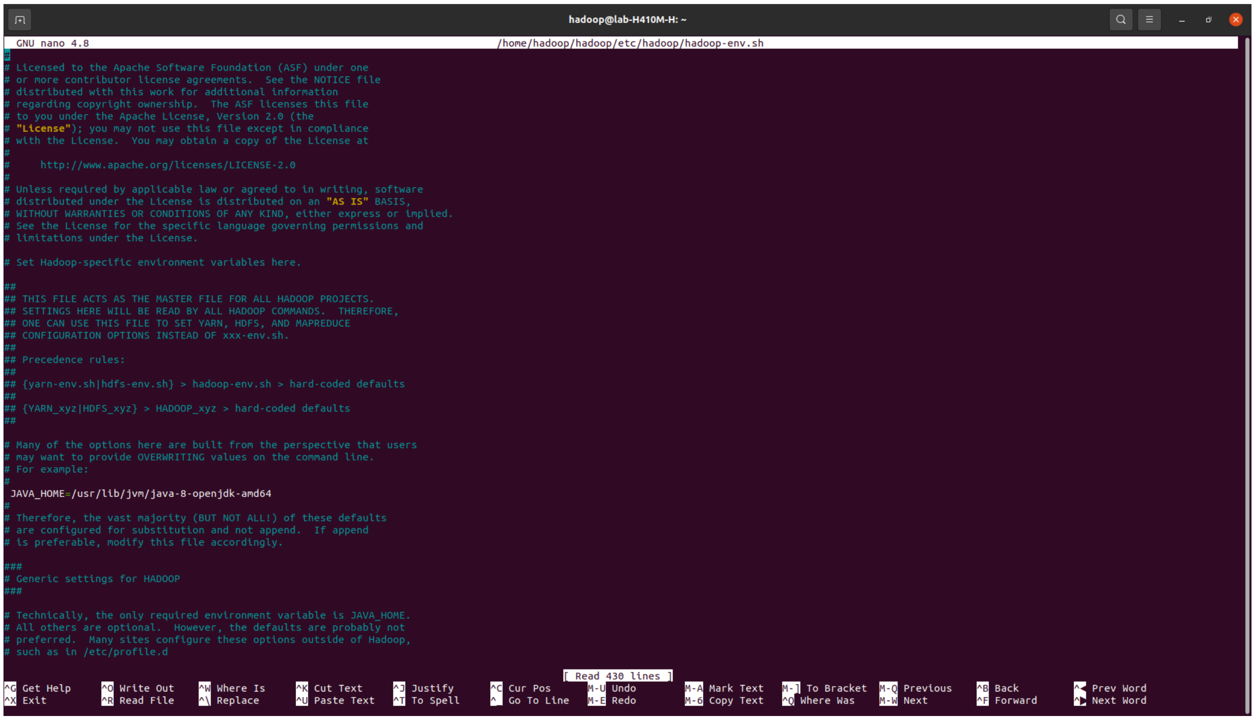
Edit the ~/.bashrc file to configure Java and Hadoop environment variables:

Command:- nano ~/.bashrc

Add the following lines and then run:

Command:- source ~/.bashrc

- We also need to configure JAVA_HOME in hadoop-env.sh file. Edit the Hadoop environment variable file in the text editor:
  nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
- In the above Configuration file we need to edit the following command :
  JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

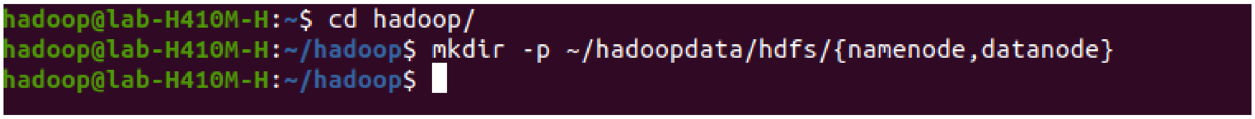


```
#
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "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.

# Set Hadoop-specific environment variables here.

##
## THIS FILE ACTS AS THE MASTER FILE FOR ALL HADOOP PROJECTS.
## SETTINGS HERE WILL BE READ BY ALL HADOOP COMMANDS.  THEREFORE,
## ONE CAN USE THIS FILE TO SET YARN, HDFS, AND MAPREDUCE
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules:
##
## {yarn-env.sh|hdfs-env.sh} > hadoop-env.sh > hard-coded defaults
##
## {YARN_xyz|HDFS_xyz} > HADOOP_xyz > hard-coded defaults
##

# Many of the options here are built from the perspective that users
# may want to provide OVERWRITING values on the command line.
# For example:
#
 JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.

###
# Generic settings for HADOOP
###

# Technically, the only required environment variable is JAVA_HOME.
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d
```

Step 12: Configuring Hadoop

Create Namenode and Datanode directories and edit Hadoop configuration files.
Ensure you set the correct values according to your system.

Firstly we need to configure Hadoop and also we need to create Namenode and Datanode directories inside the user directory.
cd hadoop/
mkdir -p ~/hadoopdata/hdfs/{namenode,datanode}



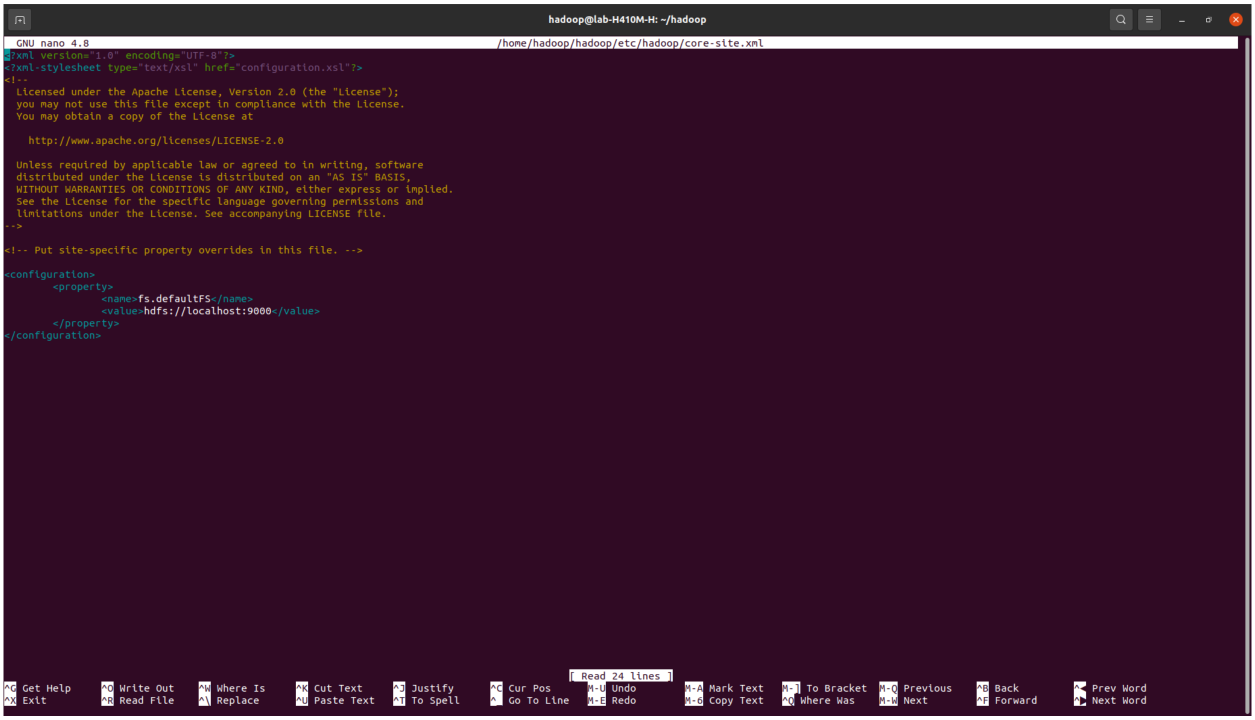
```
hadoop@lab-H410M-H:~$ cd hadoop/
hadoop@lab-H410M-H:~/hadoop$ mkdir -p ~/hadoopdata/hdfs/{namenode,datanode}
hadoop@lab-H410M-H:~/hadoop$ 
```

- Now, next edit core-site.xml file and update with your system hostname:
  nano $HADOOP_HOME/etc/hadoop/core-site.xml

  Also, we change the following names as per our system hostname:

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://localhost:9000</value>
        </property>
</configuration>
```

Later, on save (Ctrl+S) and close(Ctrl+X) the file.

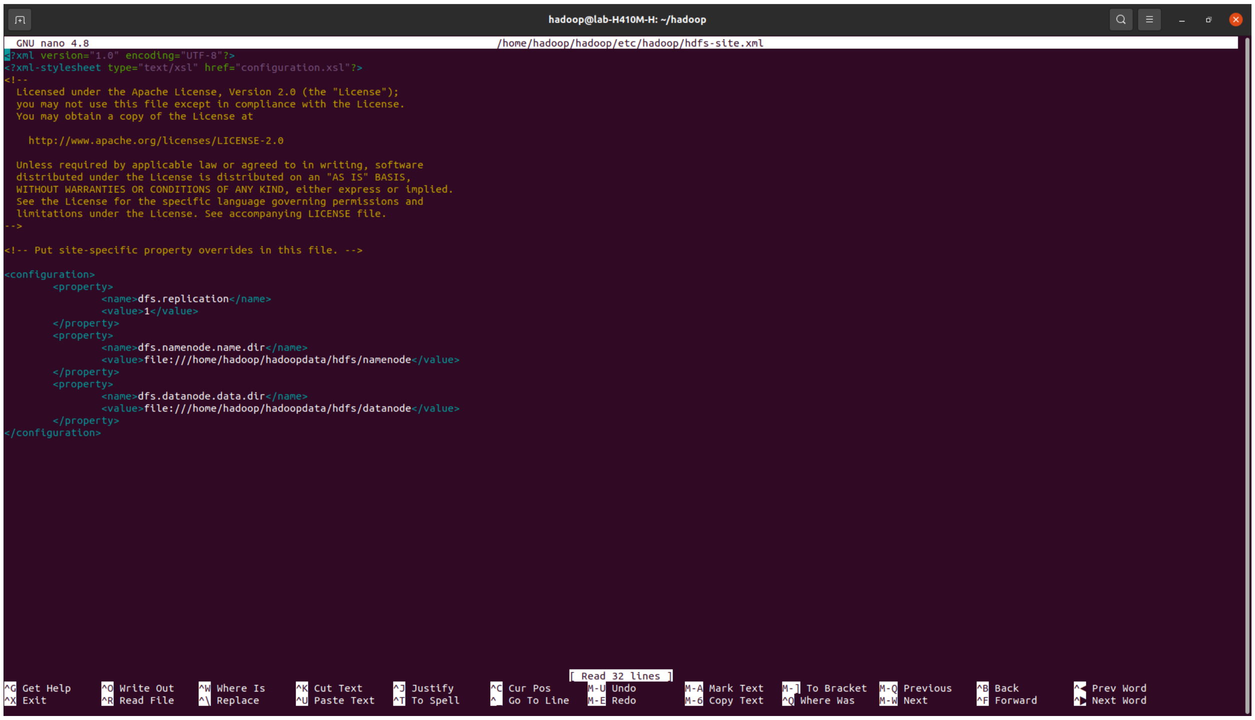


```
hadoop@lab-H410M-H: ~/hadoop
  GNU nano 4.8                                /home/hadoop/hadoop/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may 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.defaultFS</name>
                <value>hdfs://localhost:9000</value>
        </property>
</configuration>

                                                    [ Read 24 lines ]
^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo      M-A Mark Text  M-] To Bracket  M-Q Previous  ^B Back       ^◀ Prev Word
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text ^T To Spell   ^_ Go To Line M-E Redo      M-6 Copy Text  ^Q Where Was    M-W Next      ^F Forward    ^▶ Next Word
```

- Then, edit the hdfs-site.xml file:
  nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

  Change the NameNode and DataNode directory paths as shown below:

  ```
  <configuration>
          <property>
                  <name>dfs.replication</name>
                  <value>1</value>
          </property>
          <property>
                  <name>dfs.namenode.name.dir</name>
                  <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
          </property>
          <property>
                  <name>dfs.datanode.data.dir</name>
                  <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
          </property>
  </configuration>
  ```

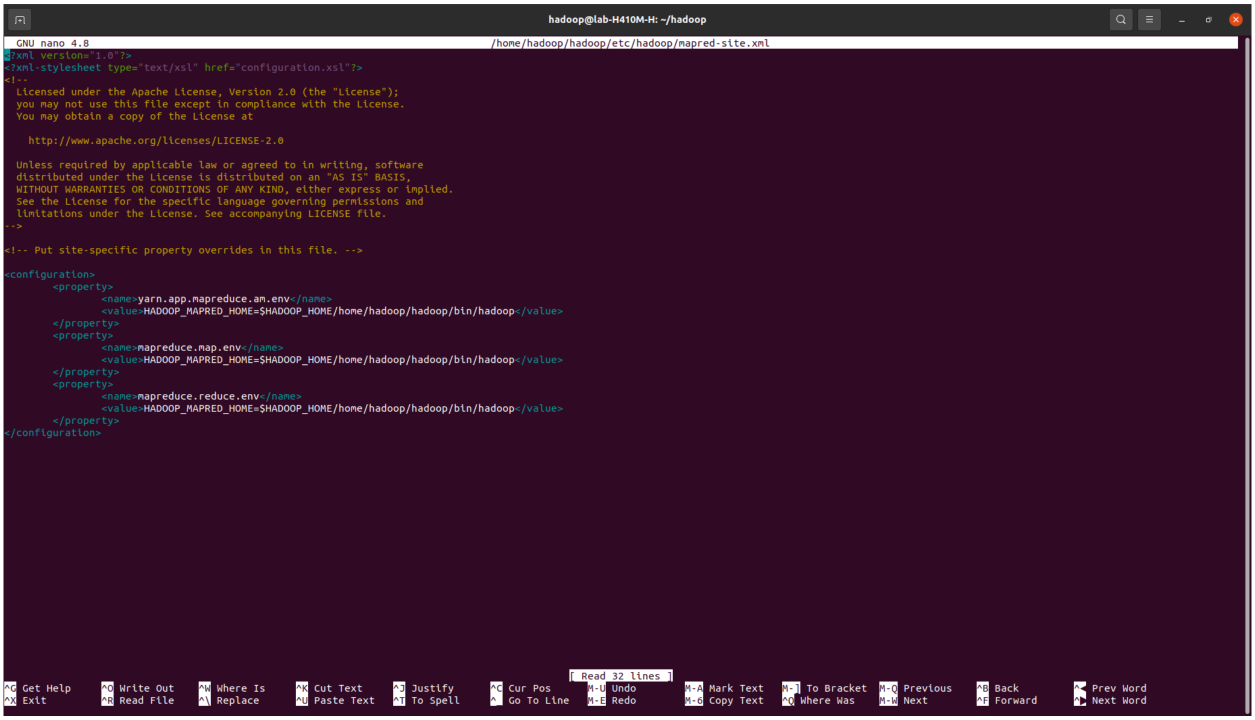


```
  GNU nano 4.8                                        /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file 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/hadoop/hadoopdata/hdfs/namenode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
        </property>
</configuration>

                                        [ Read 32 lines ]
```

- Then, edit the mapred-site.xml file:
  nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

  Make the following changes:

```
<configuration>
        <property>
                <name>yarn.app.mapreduce.am.env</name>

<value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
        <property>
                <name>mapreduce.map.env</name>

<value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
        <property>
                <name>mapreduce.reduce.env</name>

<value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
</configuration>
```



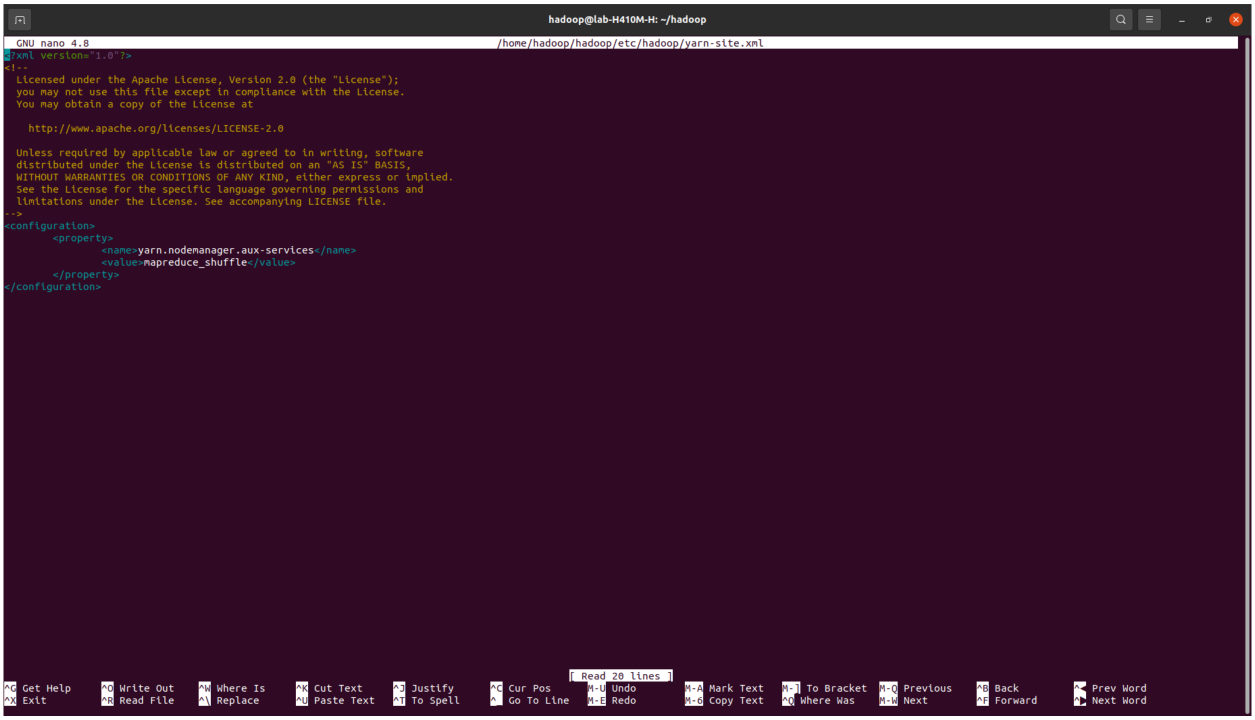
```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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>yarn.app.mapreduce.am.env</name>
                <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
        <property>
                <name>mapreduce.map.env</name>
                <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
        <property>
                <name>mapreduce.reduce.env</name>
                <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
        </property>
</configuration>

[ Read 32 lines ]
```

- Then, edit the yarn-site.xml file:
  nano $HADOOP_HOME/etc/hadoop/yarn-site.xml

  Make, the following changes:

```
<configuration>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>
```

  Now, save the file and close it



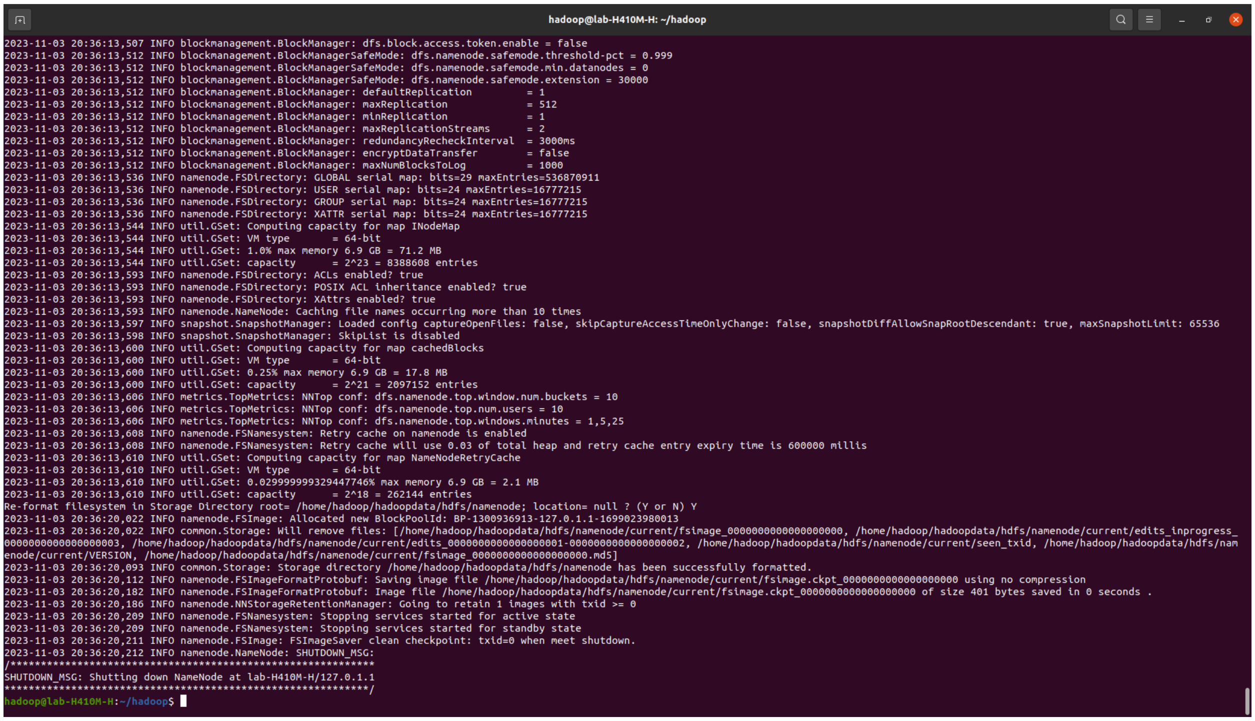
```
  GNU nano 4.8                                    /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>

                                                   [ Read 20 lines ]
^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos     M-U Undo       M-A Mark Text  M-] To Bracket M-Q Previous   ^B Back        ^◀ Prev Word
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line  M-E Redo       M-6 Copy Text  ^Q Where Was   M-W Next       ^F Forward     ^▶ Next Word
```

Step 13: Starting the Hadoop Cluster

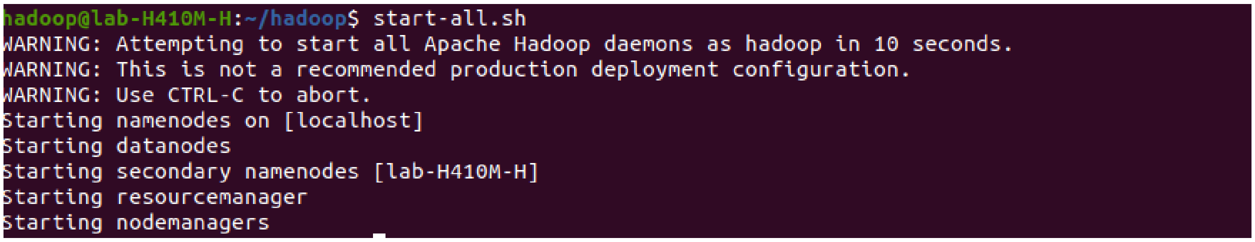
Format the Hadoop Namenode and start the Hadoop cluster:

Command:- hdfs namenode -format

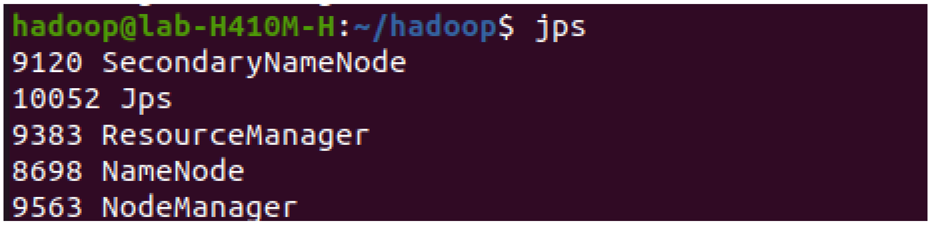
Start-all.sh



```
2023-11-03 20:36:13,507 INFO blockmanagement.BlockManager: dfs.block.access.token.enable = false
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.threshold-pct = 0.999
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.min.datanodes = 0
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.extension = 30000
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: defaultReplication         = 1
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: maxReplication             = 512
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: minReplication             = 1
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: maxReplicationStreams      = 2
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: redundancyRecheckInterval  = 3000ms
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: encryptDataTransfer        = false
2023-11-03 20:36:13,512 INFO blockmanagement.BlockManager: maxNumBlocksToLog          = 1000
2023-11-03 20:36:13,536 INFO namenode.FSDirectory: GLOBAL serial map: bits=29 maxEntries=536870911
2023-11-03 20:36:13,536 INFO namenode.FSDirectory: USER serial map: bits=24 maxEntries=16777215
2023-11-03 20:36:13,536 INFO namenode.FSDirectory: GROUP serial map: bits=24 maxEntries=16777215
2023-11-03 20:36:13,536 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2023-11-03 20:36:13,544 INFO util.GSet: Computing capacity for map INodeMap
2023-11-03 20:36:13,544 INFO util.GSet: VM type       = 64-bit
2023-11-03 20:36:13,544 INFO util.GSet: 1.0% max memory 6.9 GB = 71.2 MB
2023-11-03 20:36:13,544 INFO util.GSet: capacity      = 2^23 = 8388608 entries
2023-11-03 20:36:13,593 INFO namenode.FSDirectory: ACLs enabled? true
2023-11-03 20:36:13,593 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2023-11-03 20:36:13,593 INFO namenode.FSDirectory: XAttrs enabled? true
2023-11-03 20:36:13,593 INFO namenode.NameNode: Caching file names occurring more than 10 times
2023-11-03 20:36:13,597 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2023-11-03 20:36:13,598 INFO snapshot.SnapshotManager: SkipList is disabled
2023-11-03 20:36:13,600 INFO util.GSet: Computing capacity for map cachedBlocks
2023-11-03 20:36:13,600 INFO util.GSet: VM type       = 64-bit
2023-11-03 20:36:13,600 INFO util.GSet: 0.25% max memory 6.9 GB = 17.8 MB
2023-11-03 20:36:13,600 INFO util.GSet: capacity      = 2^21 = 2097152 entries
2023-11-03 20:36:13,606 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2023-11-03 20:36:13,606 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2023-11-03 20:36:13,606 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2023-11-03 20:36:13,608 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-11-03 20:36:13,608 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2023-11-03 20:36:13,610 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-11-03 20:36:13,610 INFO util.GSet: VM type       = 64-bit
2023-11-03 20:36:13,610 INFO util.GSet: 0.029999999329447746% max memory 6.9 GB = 2.1 MB
2023-11-03 20:36:13,610 INFO util.GSet: capacity      = 2^18 = 262144 entries
Re-format filesystem in Storage Directory root= /home/hadoop/hadoopdata/hdfs/namenode; location= null ? (Y or N) Y
2023-11-03 20:36:20,022 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1300936913-127.0.1.1-1699023980013
2023-11-03 20:36:20,022 INFO common.Storage: Will remove files: [/home/hadoop/hadoopdata/hdfs/namenode/current/fsimage_0000000000000000000, /home/hadoop/hadoopdata/hdfs/namenode/current/edits_inprogress_
0000000000000000003, /home/hadoop/hadoopdata/hdfs/namenode/current/edits_0000000000000000001-0000000000000000002, /home/hadoop/hadoopdata/hdfs/namenode/current/seen_txid, /home/hadoop/hadoopdata/hdfs/nam
enode/current/VERSION, /home/hadoop/hadoopdata/hdfs/namenode/current/fsimage_0000000000000000000.md5]
2023-11-03 20:36:20,093 INFO common.Storage: Storage directory /home/hadoop/hadoopdata/hdfs/namenode has been successfully formatted.
2023-11-03 20:36:20,112 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000 using no compression
2023-11-03 20:36:20,182 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/hadoopdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000 of size 401 bytes saved in 0 seconds .
2023-11-03 20:36:20,186 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-11-03 20:36:20,209 INFO namenode.FSNamesystem: Stopping services started for active state
2023-11-03 20:36:20,209 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-11-03 20:36:20,211 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-11-03 20:36:20,212 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at lab-H410M-H/127.0.1.1
************************************************************/
hadoop@lab-H410M-H:~/hadoop$ 
```



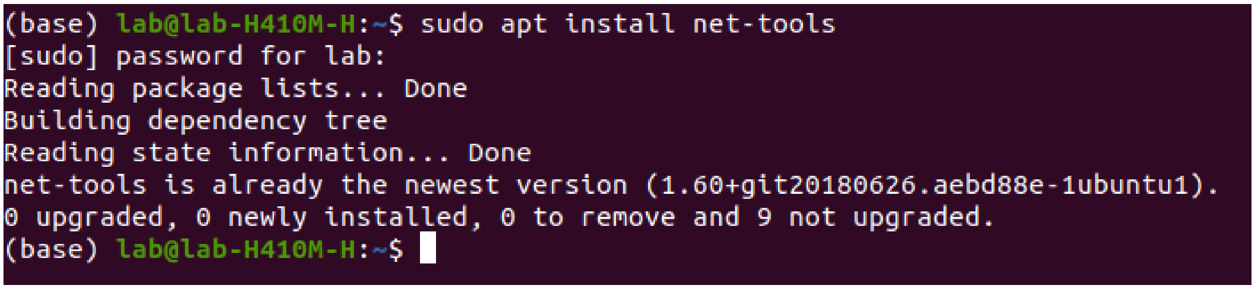
```
hadoop@lab-H410M-H:~/hadoop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [lab-H410M-H]
Starting resourcemanager
Starting nodemanagers
```



```
hadoop@lab-H410M-H:~/hadoop$ jps
9120 SecondaryNameNode
10052 Jps
9383 ResourceManager
8698 NameNode
9563 NodeManager
```

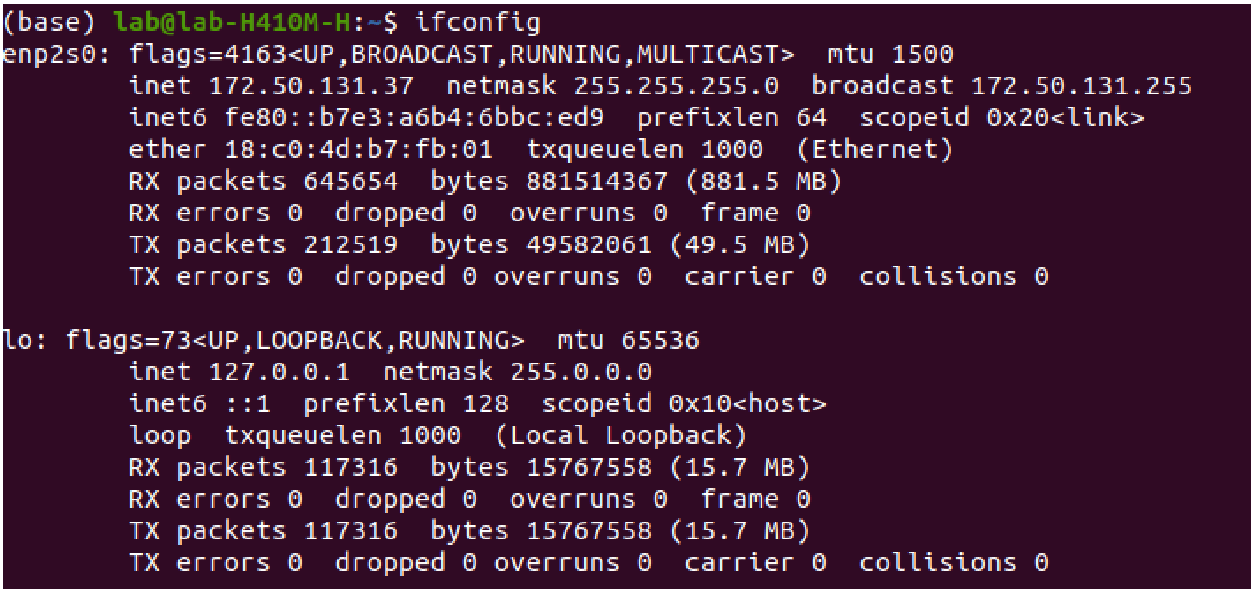
Step 14: Accessing Hadoop Services

Access the Hadoop Namenode and Resource Manager using your server's IP address.



```
(base) lab@lab-H410M-H:~$ sudo apt install net-tools
[sudo] password for lab:
Reading package lists... Done
Building dependency tree
Reading state information... Done
net-tools is already the newest version (1.60+git20180626.aebd88e-1ubuntu1).
0 upgraded, 0 newly installed, 0 to remove and 9 not upgraded.
(base) lab@lab-H410M-H:~$
```

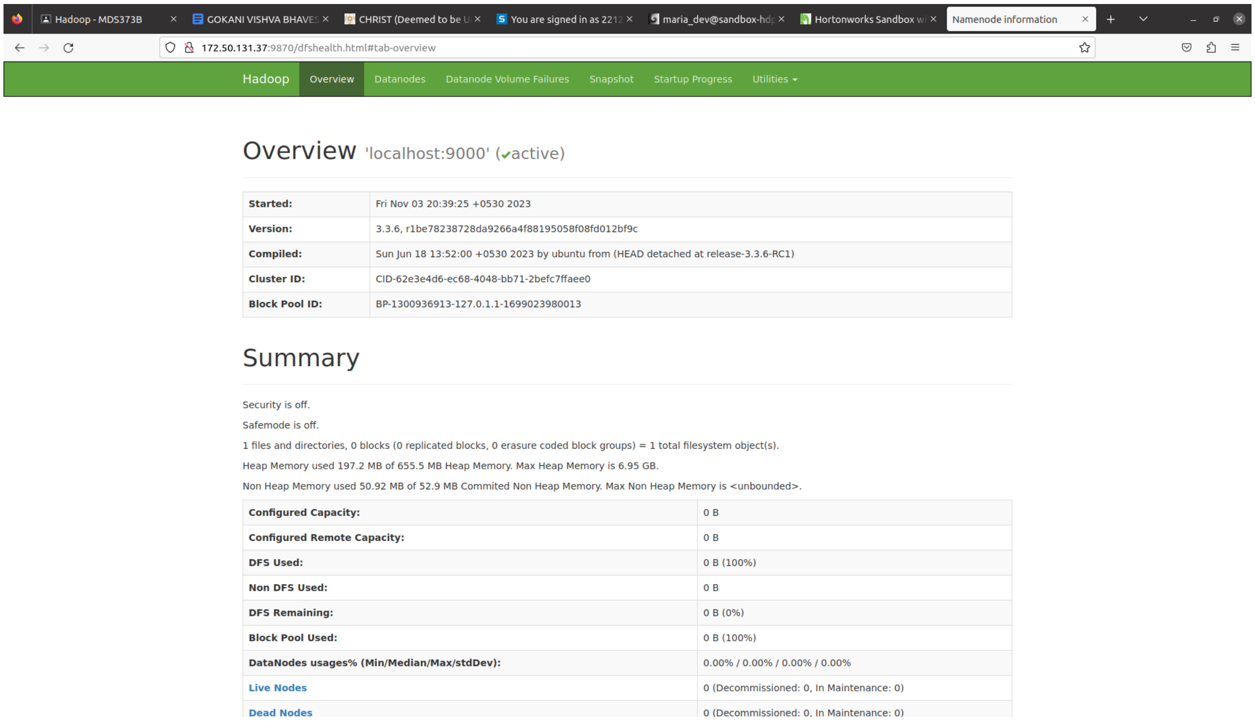
Then run ifconfig command to know our ip address:
Ifconfig



```
(base) lab@lab-H410M-H:~$ ifconfig
enp2s0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 172.50.131.37  netmask 255.255.255.0  broadcast 172.50.131.255
        inet6 fe80::b7e3:a6b4:6bbc:ed9  prefixlen 64  scopeid 0x20<link>
        ether 18:c0:4d:b7:fb:01  txqueuelen 1000  (Ethernet)
        RX packets 645654  bytes 881514367 (881.5 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 212519  bytes 49582061 (49.5 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 117316  bytes 15767558 (15.7 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 117316  bytes 15767558 (15.7 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

Here, my IP address is 172.50.131.37

So, in new tab enter 172.50.131.37:9870 we will get the following output:



Overview 'localhost:9000' (✓active)

Started:	Fri Nov 03 20:39:25 +0530 2023
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-62e3e4d6-ec68-4048-bb71-2befc7ffaee0
Block Pool ID:	BP-1300936913-127.0.1.1-1699023980013

Summary

Security is off.

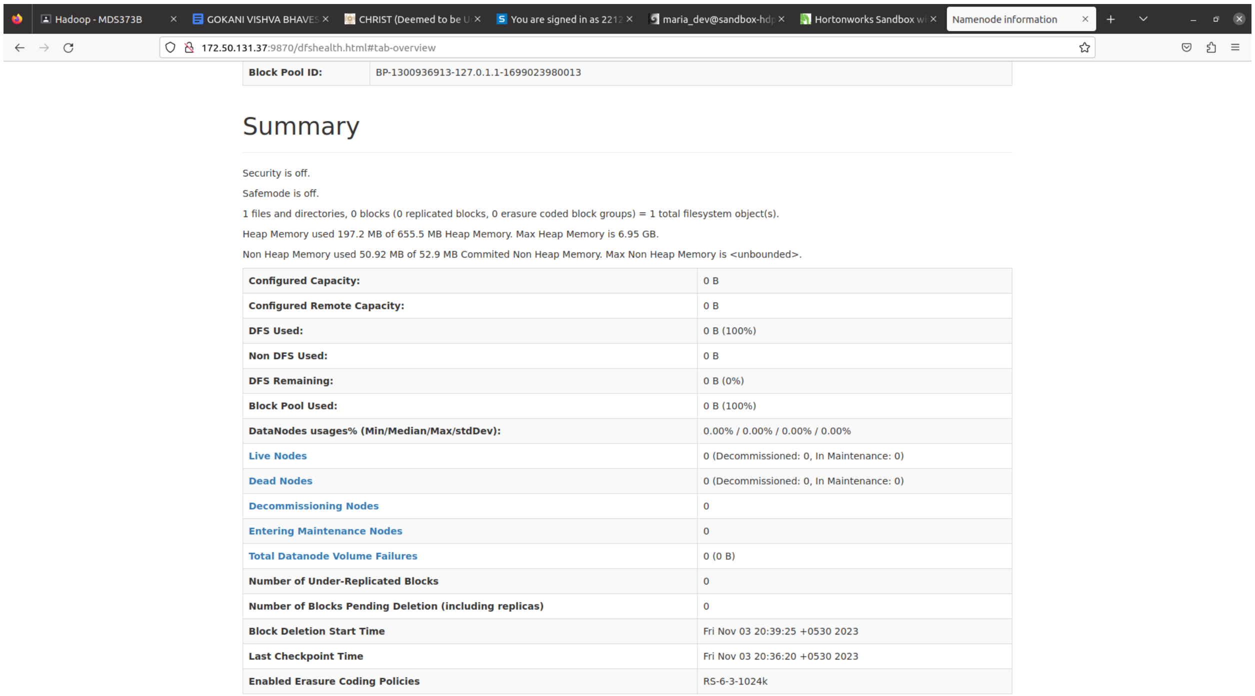
Safemode is off.

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

Heap Memory used 197.2 MB of 655.5 MB Heap Memory. Max Heap Memory is 6.95 GB.

Non Heap Memory used 50.92 MB of 52.9 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B
DFS Used:	0 B (100%)
Non DFS Used:	0 B
DFS Remaining:	0 B (0%)
Block Pool Used:	0 B (100%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)



Block Pool ID:	BP-1300936913-127.0.1.1-1699023980013

Summary

Security is off.

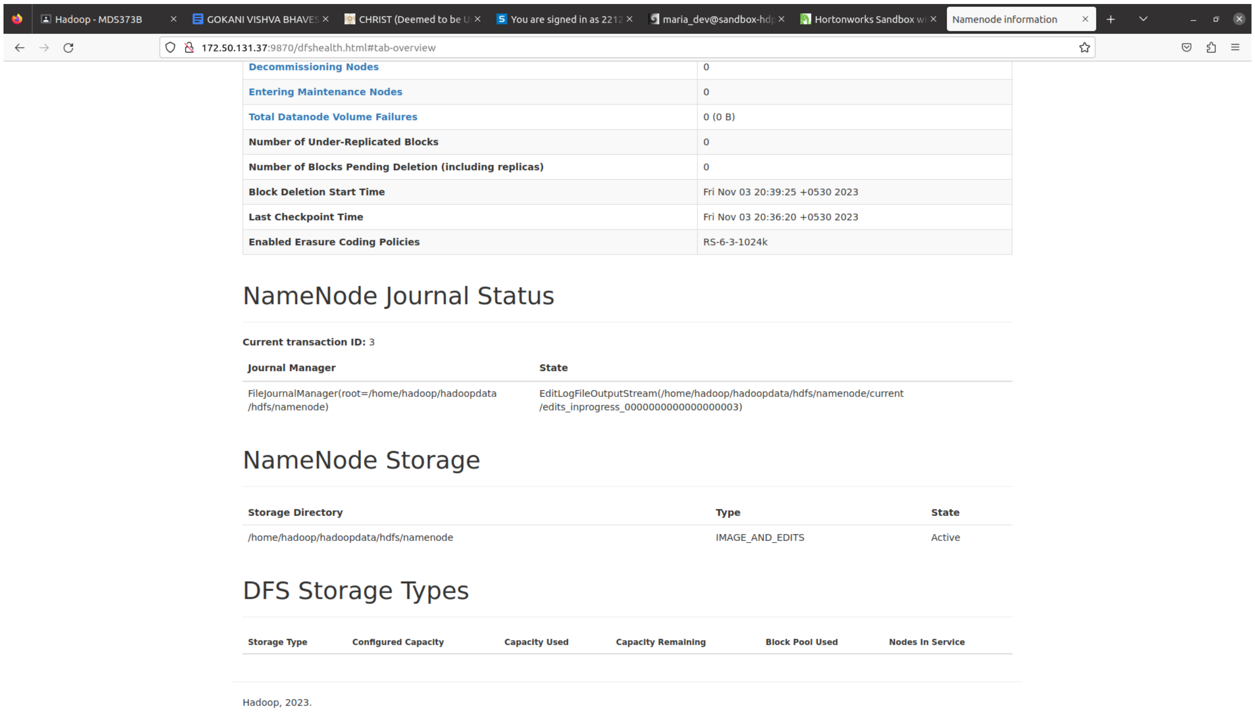
Safemode is off.

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

Heap Memory used 197.2 MB of 655.5 MB Heap Memory. Max Heap Memory is 6.95 GB.

Non Heap Memory used 50.92 MB of 52.9 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B
DFS Used:	0 B (100%)
Non DFS Used:	0 B
DFS Remaining:	0 B (0%)
Block Pool Used:	0 B (100%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Fri Nov 03 20:39:25 +0530 2023
Last Checkpoint Time	Fri Nov 03 20:36:20 +0530 2023
Enabled Erasure Coding Policies	RS-6-3-1024k



172.50.131.37:9870/dfshealth.html#tab-overview

Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Fri Nov 03 20:39:25 +0530 2023
Last Checkpoint Time	Fri Nov 03 20:36:20 +0530 2023
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 3

Journal Manager	State
FileJournalManager(root=/home/hadoop/hadoopdata/hdfs/namenode)	EditLogFileOutputStream(/home/hadoop/hadoopdata/hdfs/namenode/current/edits_inprogress_0000000000000000003)

NameNode Storage

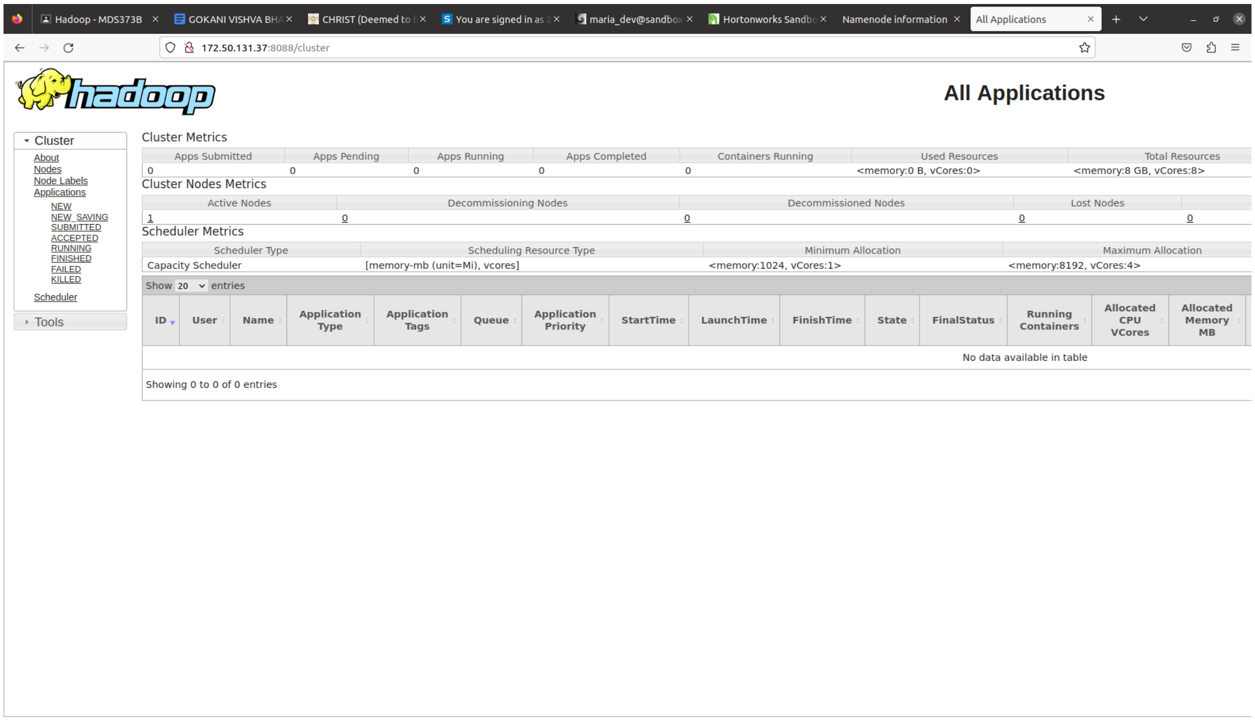
Storage Directory	Type	State
/home/hadoop/hadoopdata/hdfs/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service

Hadoop, 2023.

To access Resource Manage, open your web browser and visit the URL http://your-server-ip:8088. You should see the following screen: 172.50.131.37:8088



172.50.131.37:8088/cluster

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

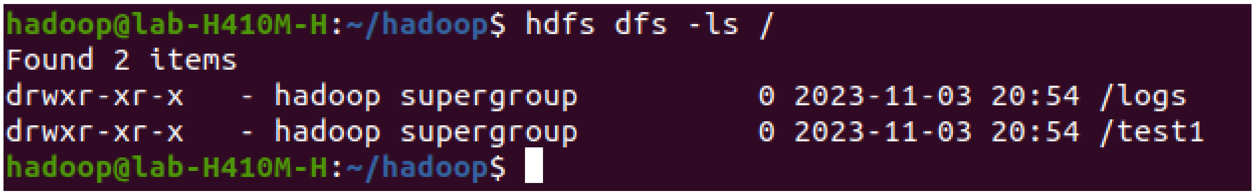
ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB
No data available in table														

Showing 0 to 0 of 0 entries

Step 15: Verifying the Hadoop Cluster

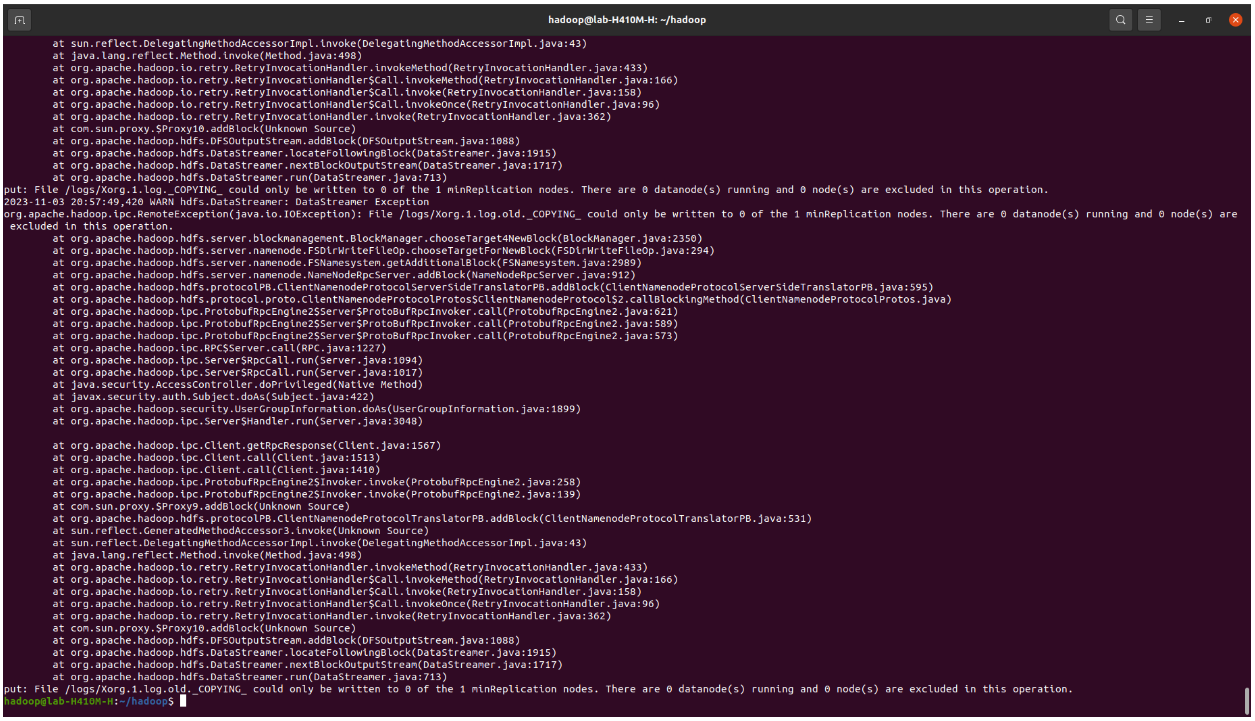
After setting up Hadoop, create directories and put some files in the HDFS filesystem to verify its functionality.
At this point, the Hadoop cluster is installed and configured. Next, we will create some directories in the HDFS filesystem to test the Hadoop.
Let’s create some directories in the HDFS filesystem using the following command:
hdfs dfs -mkdir /test1
hdfs dfs -mkdir /logs

Next, run the following command to list the above directory:
hdfs dfs -ls /



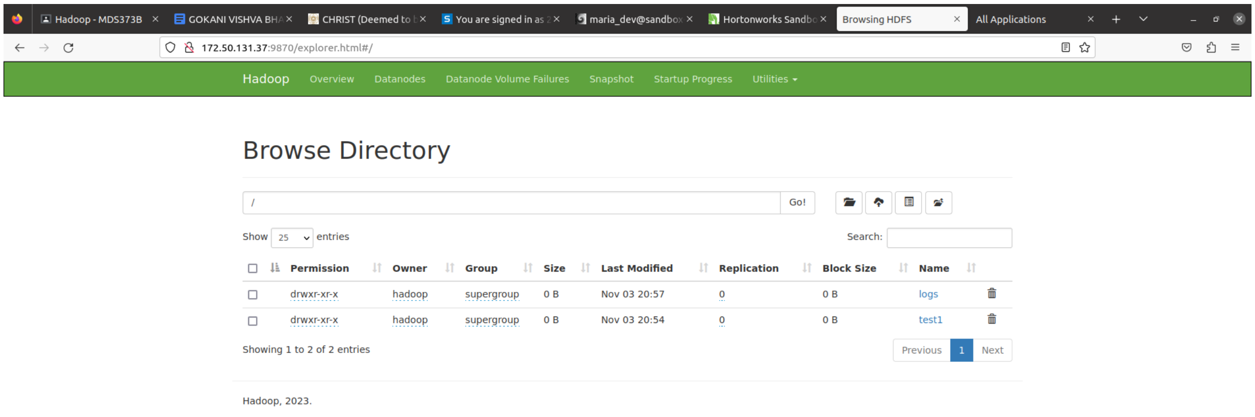
```
hadoop@lab-H410M-H:~/hadoop$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2023-11-03 20:54 /logs
drwxr-xr-x   - hadoop supergroup          0 2023-11-03 20:54 /test1
hadoop@lab-H410M-H:~/hadoop$ 
```

Also, put some files to hadoop file system. For the example, putting log files
from host machine to hadoop file system.
hdfs dfs -put /var/log/* /logs/



```
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:433)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeMethod(RetryInvocationHandler.java:166)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invoke(RetryInvocationHandler.java:158)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeOnce(RetryInvocationHandler.java:96)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:362)
        at com.sun.proxy.$Proxy10.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.DFSOutputStream.addBlock(DFSOutputStream.java:1088)
        at org.apache.hadoop.hdfs.DataStreamer.locateFollowingBlock(DataStreamer.java:1915)
        at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1717)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:713)
put: File /logs/Xorg.1.log._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 0 datanode(s) running and 0 node(s) are excluded in this operation.
2023-11-03 20:57:49,420 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /logs/Xorg.1.log.old._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 0 datanode(s) running and 0 node(s) are
 excluded in this operation.
        at org.apache.hadoop.hdfs.server.blockmanagement.BlockManager.chooseTarget4NewBlock(BlockManager.java:2350)
        at org.apache.hadoop.hdfs.server.namenode.FSDirWriteFileOp.chooseTargetForNewBlock(FSDirWriteFileOp.java:294)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.getAdditionalBlock(FSNamesystem.java:2989)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.addBlock(NameNodeRpcServer.java:912)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.addBlock(ClientNamenodeProtocolServerSideTranslatorPB.java:595)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:621)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:589)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine2.java:573)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:1227)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1094)
        at org.apache.hadoop.ipc.Server$RpcCall.run(Server.java:1017)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1899)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:3048)

        at org.apache.hadoop.ipc.Client.getRpcResponse(Client.java:1567)
        at org.apache.hadoop.ipc.Client.call(Client.java:1513)
        at org.apache.hadoop.ipc.Client.call(Client.java:1410)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:258)
        at org.apache.hadoop.ipc.ProtobufRpcEngine2$Invoker.invoke(ProtobufRpcEngine2.java:139)
        at com.sun.proxy.$Proxy9.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.addBlock(ClientNamenodeProtocolTranslatorPB.java:531)
        at sun.reflect.GeneratedMethodAccessor3.invoke(Unknown Source)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:433)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeMethod(RetryInvocationHandler.java:166)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invoke(RetryInvocationHandler.java:158)
        at org.apache.hadoop.io.retry.RetryInvocationHandler$Call.invokeOnce(RetryInvocationHandler.java:96)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:362)
        at com.sun.proxy.$Proxy10.addBlock(Unknown Source)
        at org.apache.hadoop.hdfs.DFSOutputStream.addBlock(DFSOutputStream.java:1088)
        at org.apache.hadoop.hdfs.DataStreamer.locateFollowingBlock(DataStreamer.java:1915)
        at org.apache.hadoop.hdfs.DataStreamer.nextBlockOutputStream(DataStreamer.java:1717)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:713)
put: File /logs/Xorg.1.log.old._COPYING_ could only be written to 0 of the 1 minReplication nodes. There are 0 datanode(s) running and 0 node(s) are excluded in this operation.
hadoop@lab-H410M-H:~/hadoop$ 
```

To verify the above files and directories. Go to the web interface, click on the Utilities => Browse the file system. We should see your directories which you have created earlier in the following screen:



Browse Directory

/

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwxr-xr-x	hadoop	supergroup	0 B	Nov 03 20:57	0	0 B	logs
☐	drwxr-xr-x	hadoop	supergroup	0 B	Nov 03 20:54	0	0 B	test1

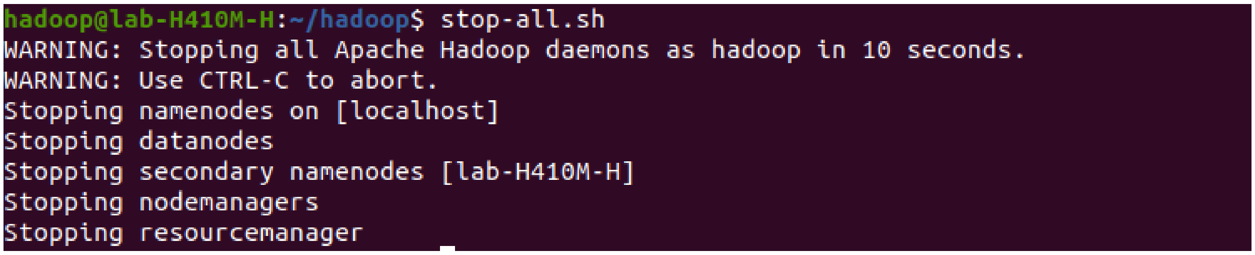
Showing 1 to 2 of 2 entries

Hadoop, 2023.

Step 16: Stopping Hadoop

To stop the Hadoop services, use the following command.

Command:- stop-all.sh



```
hadoop@lab-H410M-H:~/hadoop$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [lab-H410M-H]
Stopping nodemanagers
Stopping resourcemanager
```