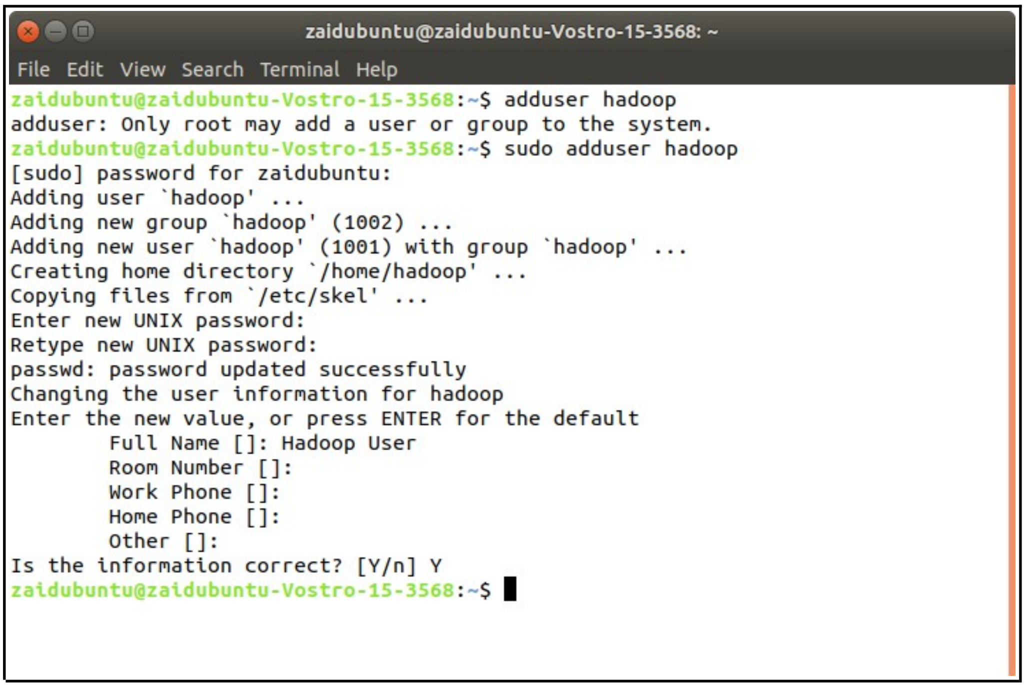
Practical No. 2

AIM: Install Hadoop, Hadoop Word Count and Mutligram.

Category	Requirements, Conventions or Software Version Used
System	Ubuntu 18.04
Software	Hadoop 2.8.5, Oracle JDK 1.8
Other	Privileged access to your Linux system as root or via the `sudo` command.
Conventions	**#** - requires given linux commands to be executed with root privileges either directly as a root user or by use of `sudo` command **$** - requires given linux commands to be executed as a regular non-privileged user

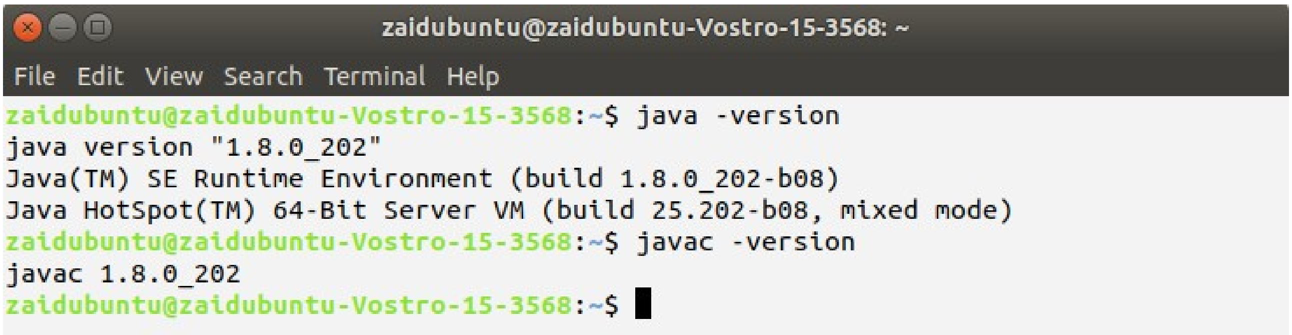


zaidubuntu@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ adduser hadoop
adduser: Only root may add a user or group to the system.
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ sudo adduser hadoop
[sudo] password for zaidubuntu:
Adding user `hadoop' ...
Adding new group `hadoop' (1002) ...
Adding new user `hadoop' (1001) with group `hadoop' ...
Creating home directory `/home/hadoop' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
        Full Name []: Hadoop User
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] Y
zaidubuntu@zaidubuntu-Vostro-15-3568:~$
```

Install and configure the Oracle JDK

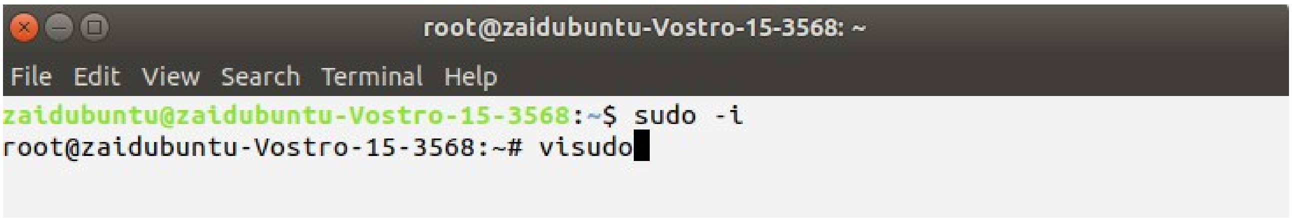
Make sure the installed Java is that of Oracle’s and not OpenJDK’s.



zaidubuntu@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ javac -version
javac 1.8.0_202
zaidubuntu@zaidubuntu-Vostro-15-3568:~$
```

Switch to root user:

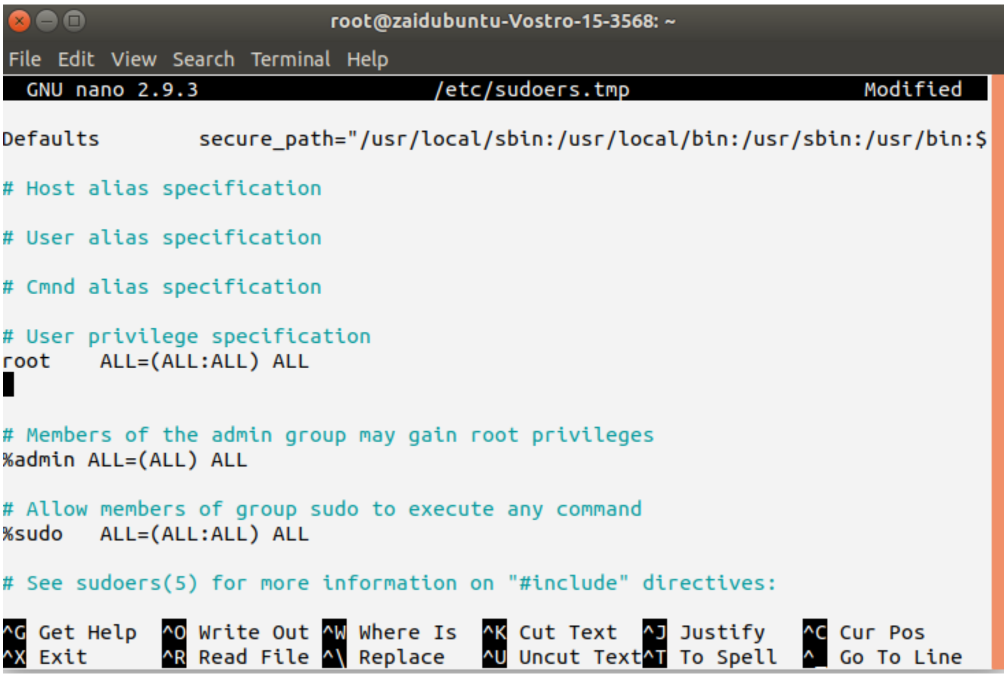


root@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ sudo -i
root@zaidubuntu-Vostro-15-3568:~# visudo
```

And then enter ‘**visudo**’:

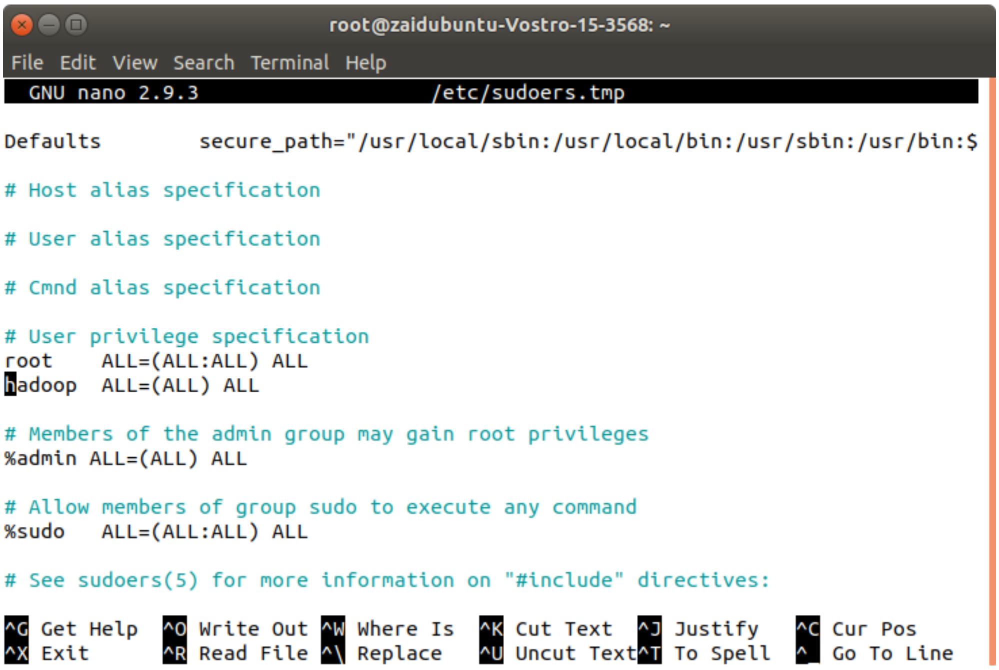
BEFORE:



root@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
  GNU nano 2.9.3                 /etc/sudoers.tmp                      Modified

Defaults        secure_path="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:$

# Host alias specification

# User alias specification

# Cmnd alias specification

# User privilege specification
root    ALL=(ALL:ALL) ALL


# Members of the admin group may gain root privileges
%admin ALL=(ALL) ALL

# Allow members of group sudo to execute any command
%sudo   ALL=(ALL:ALL) ALL

# See sudoers(5) for more information on "#include" directives:

^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos
^X Exit       ^R Read File  ^\ Replace    ^U Uncut Text ^T To Spell   ^_ Go To Line
```

AFTER ADDING:hadoop ALL=(ALL) ALL


root@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

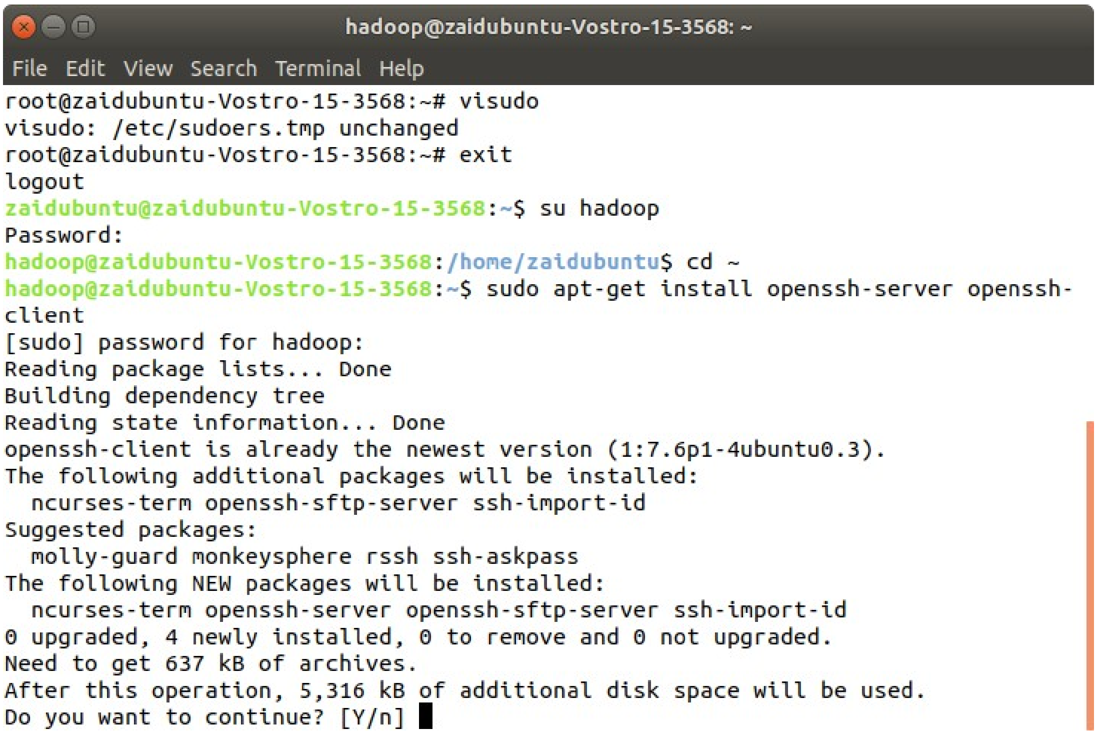
```
  GNU nano 2.9.3                    /etc/sudoers.tmp

Defaults        secure_path="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:$

# Host alias specification

# User alias specification

# Cmnd alias specification

# User privilege specification
root    ALL=(ALL:ALL) ALL
hadoop  ALL=(ALL) ALL

# Members of the admin group may gain root privileges
%admin ALL=(ALL) ALL

# Allow members of group sudo to execute any command
%sudo   ALL=(ALL:ALL) ALL

# See sudoers(5) for more information on "#include" directives:

^G Get Help  ^O Write Out ^W Where Is  ^K Cut Text  ^J Justify   ^C Cur Pos
^X Exit      ^R Read File ^\ Replace   ^U Uncut Text^T To Spell  ^_ Go To Line
```

Save by pressing CTRL + X and then Save Yes:

Switch to hadoop user and enter:

This part is required for Accessing and Running MapReduce programs remotely using SSH like in the Lab.

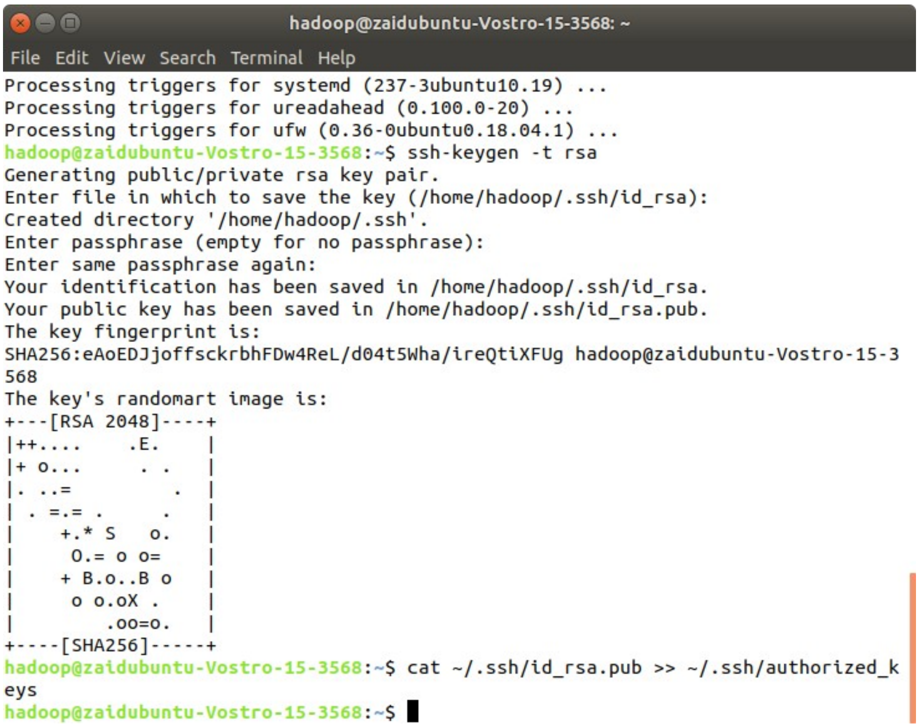
```
sudo apt-get install openssh-server openssh-client
```


hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
root@zaidubuntu-Vostro-15-3568:~# visudo
visudo: /etc/sudoers.tmp unchanged
root@zaidubuntu-Vostro-15-3568:~# exit
logout
zaidubuntu@zaidubuntu-Vostro-15-3568:~$ su hadoop
Password:
hadoop@zaidubuntu-Vostro-15-3568:/home/zaidubuntu$ cd ~
hadoop@zaidubuntu-Vostro-15-3568:~$ sudo apt-get install openssh-server openssh-
client
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:7.6p1-4ubuntu0.3).
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 0 not upgraded.
Need to get 637 kB of archives.
After this operation, 5,316 kB of additional disk space will be used.
Do you want to continue? [Y/n]
```

Generate Public and Private Key Pairs with the following command. The terminal will prompt for entering the file name. Press `ENTER` and proceed. After that copy the public keys form `id_rsa.pub` to `authorized_keys`.

```
$ ssh-keygen -t rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

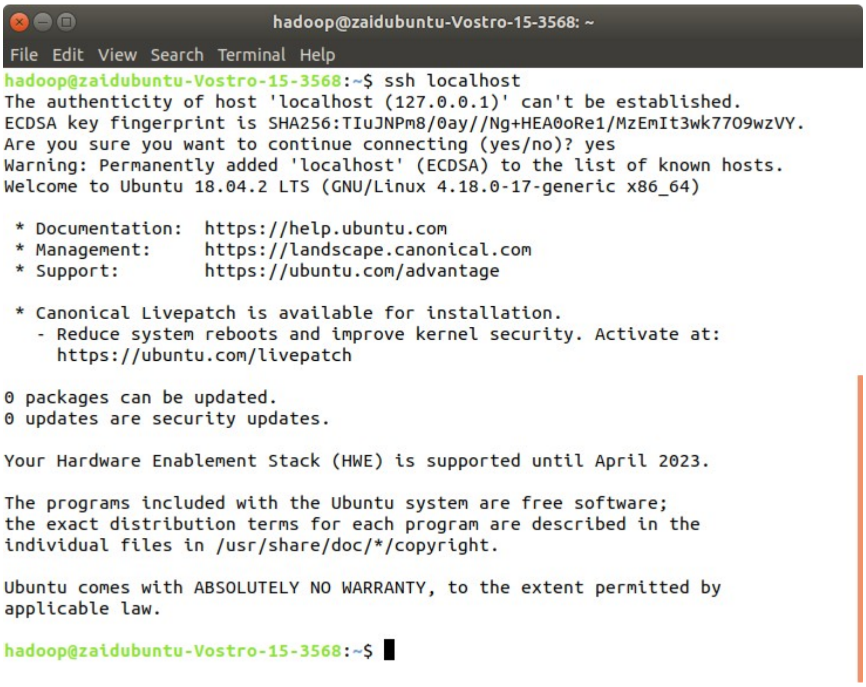


hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
Processing triggers for systemd (237-3ubuntu10.19) ...
Processing triggers for ureadahead (0.100.0-20) ...
Processing triggers for ufw (0.36-0ubuntu0.18.04.1) ...
hadoop@zaidubuntu-Vostro-15-3568:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
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:eAoEDJjoffsckrbhFDw4ReL/d04t5Wha/ireQtiXFUg hadoop@zaidubuntu-Vostro-15-3
568
The key's randomart image is:
+---[RSA 2048]----+
|++....    .E.    |
|+ o...      . .  |
|. ..=          . |
| . =.= .      .  |
|     +.* S   o.  |
|      O.= o o=   |
|     + B.o..B o  |
|      o o.oX .   |
|         .oo=o.  |
+----[SHA256]-----+
hadoop@zaidubuntu-Vostro-15-3568:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_k
eys
hadoop@zaidubuntu-Vostro-15-3568:~$ 
```

Verify the password-less ssh configuration with the command :

```
$ ssh localhost
```



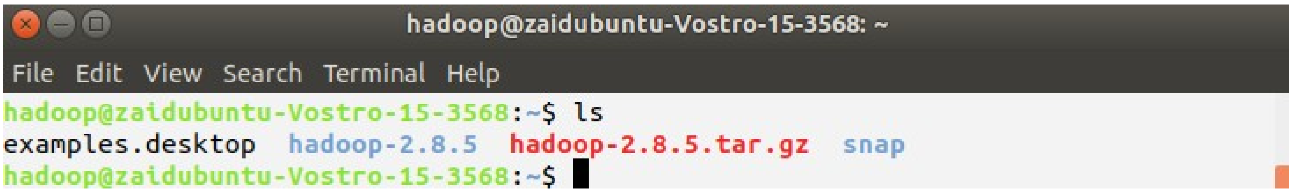
hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
hadoop@zaidubuntu-Vostro-15-3568:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:TIuJNPm8/0ay//Ng+HEA0oRe1/MzEmIt3wk77O9wzVY.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.18.0-17-generic x86_64)

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

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

0 packages can be updated.
0 updates are security updates.

Your Hardware Enablement Stack (HWE) is supported until April 2023.

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
applicable law.

hadoop@zaidubuntu-Vostro-15-3568:~$ 
```

Install Hadoop and configure related xml files

Download and extract Hadoop 2.8.5 from Apache official website.

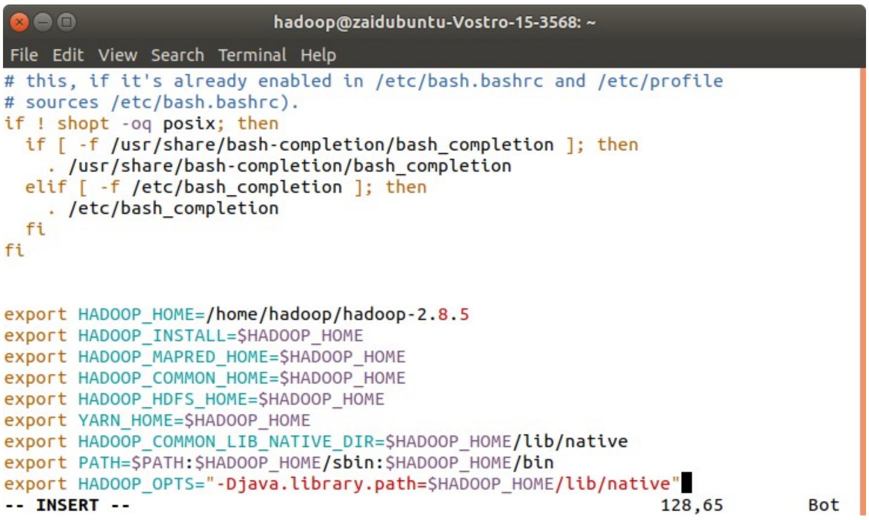
```
# tar -xzvf hadoop-2.8.5.tar.gz
```


hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

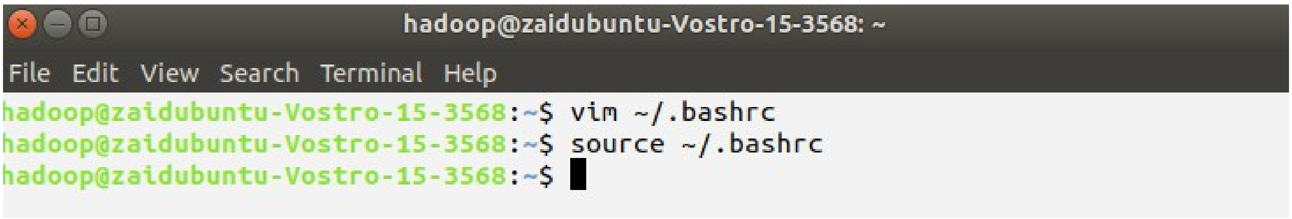
```
hadoop@zaidubuntu-Vostro-15-3568:~$ ls
examples.desktop  hadoop-2.8.5  hadoop-2.8.5.tar.gz  snap
hadoop@zaidubuntu-Vostro-15-3568:~$
```

Edit the `bashrc` for the Hadoop user via setting up the following Hadoop environment variables :

```
export HADOOP_HOME=/home/hadoop/hadoop-2.8.5
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"
```


hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# 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 HADOOP_HOME=/home/hadoop/hadoop-2.8.5
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"
-- INSERT --                                            128,65        Bot
```

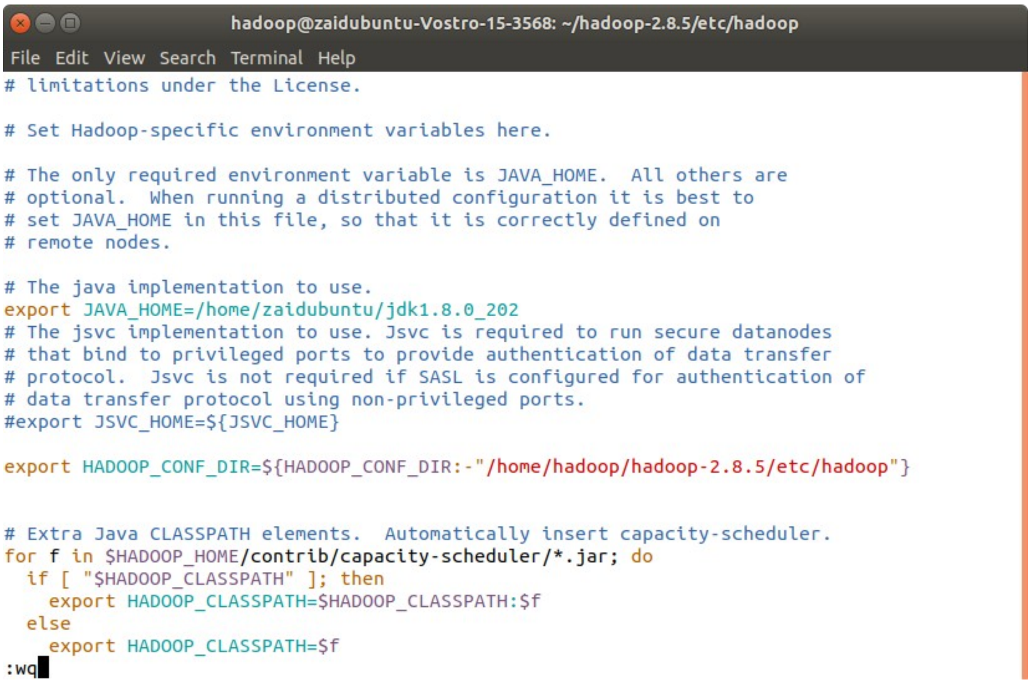

hadoop@zaidubuntu-Vostro-15-3568: ~
File Edit View Search Terminal Help

```
hadoop@zaidubuntu-Vostro-15-3568:~$ vim ~/.bashrc
hadoop@zaidubuntu-Vostro-15-3568:~$ source ~/.bashrc
hadoop@zaidubuntu-Vostro-15-3568:~$
```

Edit the `hadoop-env.sh` file which is in `/etc/hadoop` inside the Hadoop installation directory and make the following changes and check if you want to change any other configurations.

```
export JAVA_HOME=/home/zaidubuntu/jdk1.8.0_202

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/home/hadoop/hadoop-2.8.5/etc/
hadoop"}
```

NOTE: Delete the old `HADOOP_CONF_DIR and JAVA_HOME line`



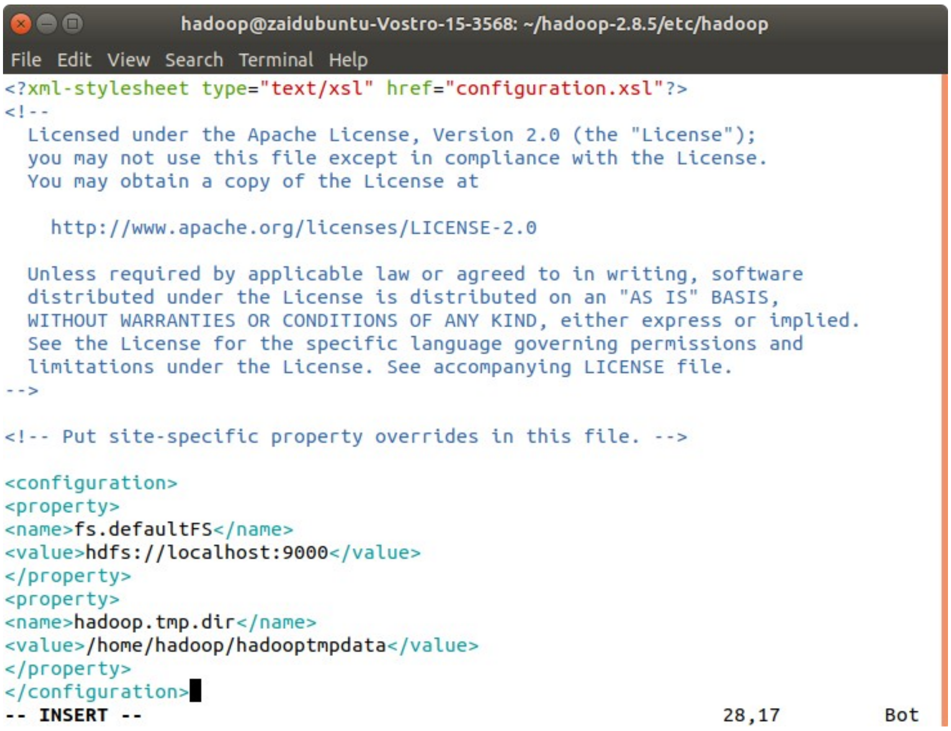
hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
# limitations under the License.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME.  All others are
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/home/zaidubuntu/jdk1.8.0_202
# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol.  Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/home/hadoop/hadoop-2.8.5/etc/hadoop"}


# Extra Java CLASSPATH elements.  Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
:wq
```

Configuration Changes in core-site.xml file

Edit the `core-site.xml` with vim or you can use any of the editors. The file is under `/etc/hadoop` inside `hadoop` home directory and add following entries.

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/hadoop/hadooptmpdata</value>
</property>
</configuration>
```


hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
<?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.tmp.dir</name>
<value>/home/hadoop/hadooptmpdata</value>
</property>
</configuration>
-- INSERT --                                          28,17         Bot
```

In addition, create the directory under hadoop home folder.

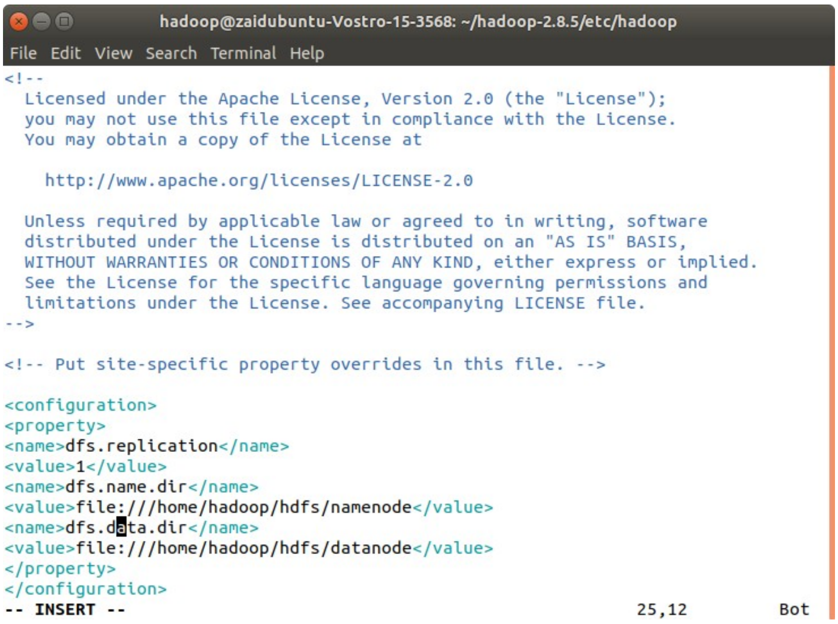
```
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ mkdir
/home/hadoop/hadooptmpdata
```

Configuration Changes in hdfs-site.xml file

Edit the `hdfs-site.xml` which is present under the same location i.e `/etc/hadoop` inside `hadoop` installation directory and create the `Namenode/Datanode` directories under `hadoop` user home directory.

```
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ mkdir -p
/home/hadoop/hdfs/namenode

hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ mkdir -p
/home/hadoop/hdfs/datanode
```

```
<configuration>

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

</configuration>
```


hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

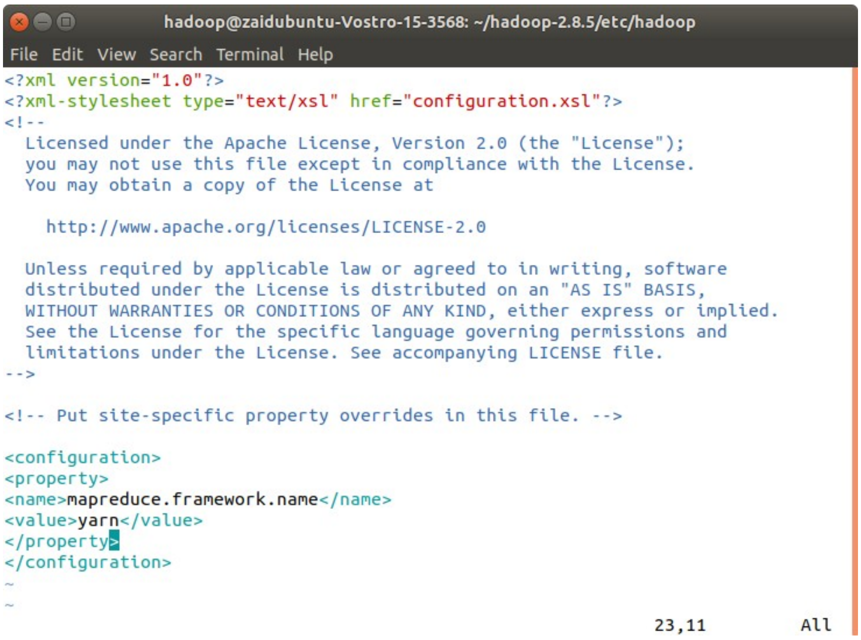
```
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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>
<name>dfs.name.dir</name>
<value>file:///home/hadoop/hdfs/namenode</value>
<name>dfs.data.dir</name>
<value>file:///home/hadoop/hdfs/datanode</value>
</property>
</configuration>
-- INSERT --                                              25,12         Bot
```

Configuration Changes in mapred-site.xml file

Copy the `mapred-site.xml` from `mapred-site.xml.template` using `cp` command and then edit the `mapred-site.xml` placed in `/etc/hadoop` under `hadoop` instillation directory with the following changes.

```
$ cp mapred-site.xml.template mapred-site.xml
```

```
<configuration>

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

</configuration>
```

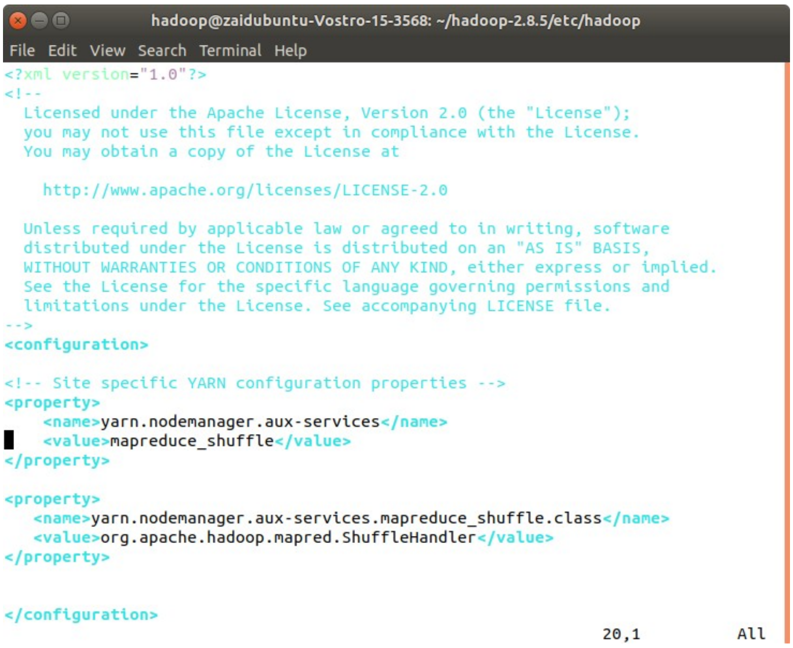

hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
<?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>
~
~
                                                          23,11         All
```

Configuration Changes in yarn-site.xml file

Edit `yarn-site.xml` with the following entries.

```
<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>
```



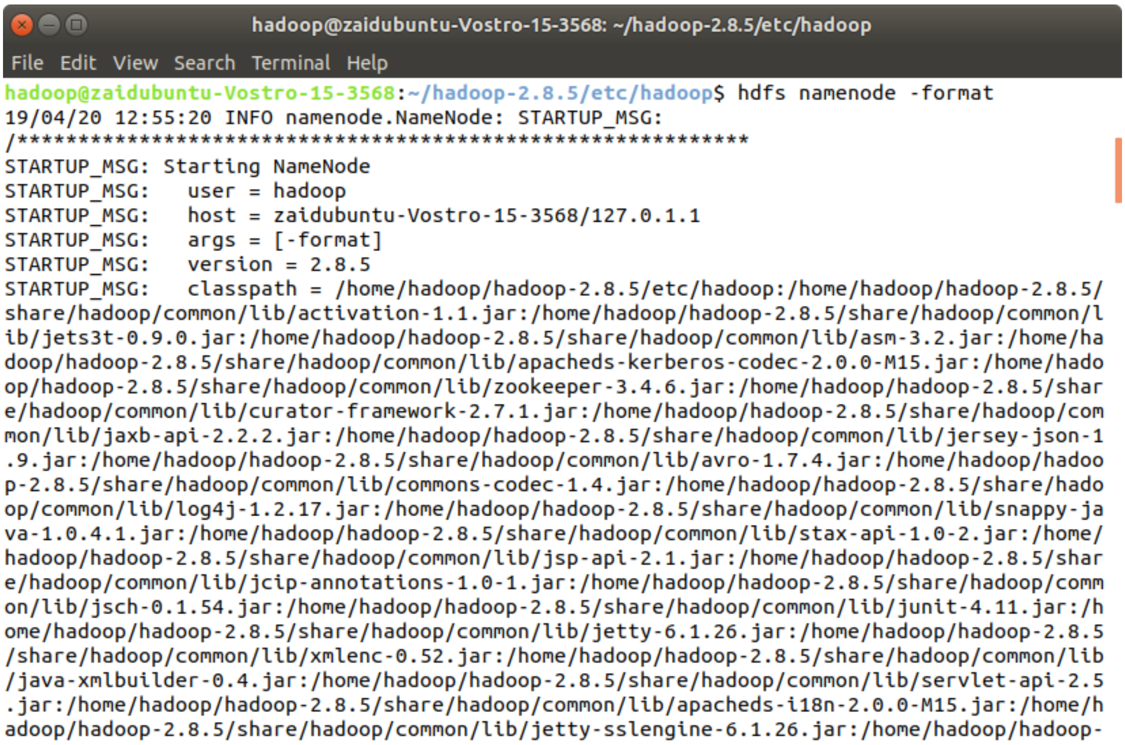
hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
<?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>
                                                              20,1          All
```

Starting the Hadoop Cluster

Format the namenode before using it for the first time. As HDFS user run the below command to format the Namenode.

```
$ hdfs namenode -format
```

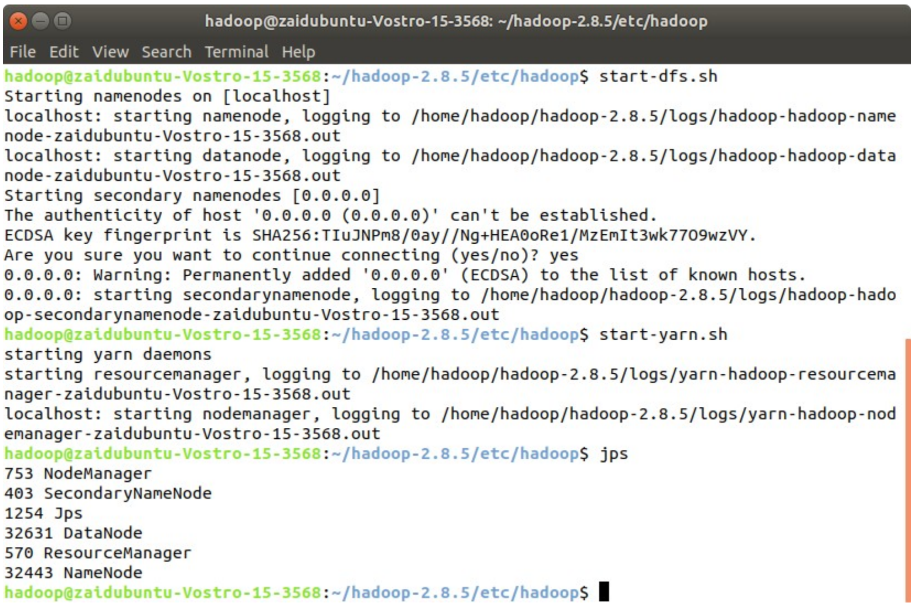

hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ hdfs namenode -format
19/04/20 12:55:20 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   user = hadoop
STARTUP_MSG:   host = zaidubuntu-Vostro-15-3568/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.8.5
STARTUP_MSG:   classpath = /home/hadoop/hadoop-2.8.5/etc/hadoop:/home/hadoop/hadoop-2.8.5/
share/hadoop/common/lib/activation-1.1.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/l
ib/jets3t-0.9.0.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/asm-3.2.jar:/home/ha
doop/hadoop-2.8.5/share/hadoop/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/hado
op/hadoop-2.8.5/share/hadoop/common/lib/zookeeper-3.4.6.jar:/home/hadoop/hadoop-2.8.5/shar
e/hadoop/common/lib/curator-framework-2.7.1.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/com
mon/lib/jaxb-api-2.2.2.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/jersey-json-1
.9.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/avro-1.7.4.jar:/home/hadoop/hadoo
p-2.8.5/share/hadoop/common/lib/commons-codec-1.4.jar:/home/hadoop/hadoop-2.8.5/share/hado
op/common/lib/log4j-1.2.17.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/snappy-ja
va-1.0.4.1.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/stax-api-1.0-2.jar:/home/
hadoop/hadoop-2.8.5/share/hadoop/common/lib/jsp-api-2.1.jar:/home/hadoop/hadoop-2.8.5/shar
e/hadoop/common/lib/jcip-annotations-1.0-1.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/comm
on/lib/jsch-0.1.54.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/junit-4.11.jar:/h
ome/hadoop/hadoop-2.8.5/share/hadoop/common/lib/jetty-6.1.26.jar:/home/hadoop/hadoop-2.8.5
/share/hadoop/common/lib/xmlenc-0.52.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib
/java-xmlbuilder-0.4.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/servlet-api-2.5
.jar:/home/hadoop/hadoop-2.8.5/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/home/h
adoop/hadoop-2.8.5/share/hadoop/common/lib/jetty-sslengine-6.1.26.jar:/home/hadoop/hadoop-
```

Once the Namenode has been formatted then start the HDFS using the start-dfs.sh script.

To start the YARN services you need to execute the yarn start script i.e. start-yarn.sh

To verify all the Hadoop services/daemons are started successfully you can use the jps command.


hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/hadoop/hadoop-2.8.5/logs/hadoop-hadoop-name
node-zaidubuntu-Vostro-15-3568.out
localhost: starting datanode, logging to /home/hadoop/hadoop-2.8.5/logs/hadoop-hadoop-data
node-zaidubuntu-Vostro-15-3568.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:TIuJNPm8/0ay//Ng+HEA0oRe1/MzEmIt3wk77O9wzVY.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/hadoop/hadoop-2.8.5/logs/hadoop-hado
op-secondarynamenode-zaidubuntu-Vostro-15-3568.out
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/hadoop/hadoop-2.8.5/logs/yarn-hadoop-resourcema
nager-zaidubuntu-Vostro-15-3568.out
localhost: starting nodemanager, logging to /home/hadoop/hadoop-2.8.5/logs/yarn-hadoop-nod
emanager-zaidubuntu-Vostro-15-3568.out
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ jps
753 NodeManager
403 SecondaryNameNode
1254 Jps
32631 DataNode
570 ResourceManager
32443 NameNode
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ █
```

Now we
can check the current Hadoop version you can use below command :

```
$ hadoop version
```

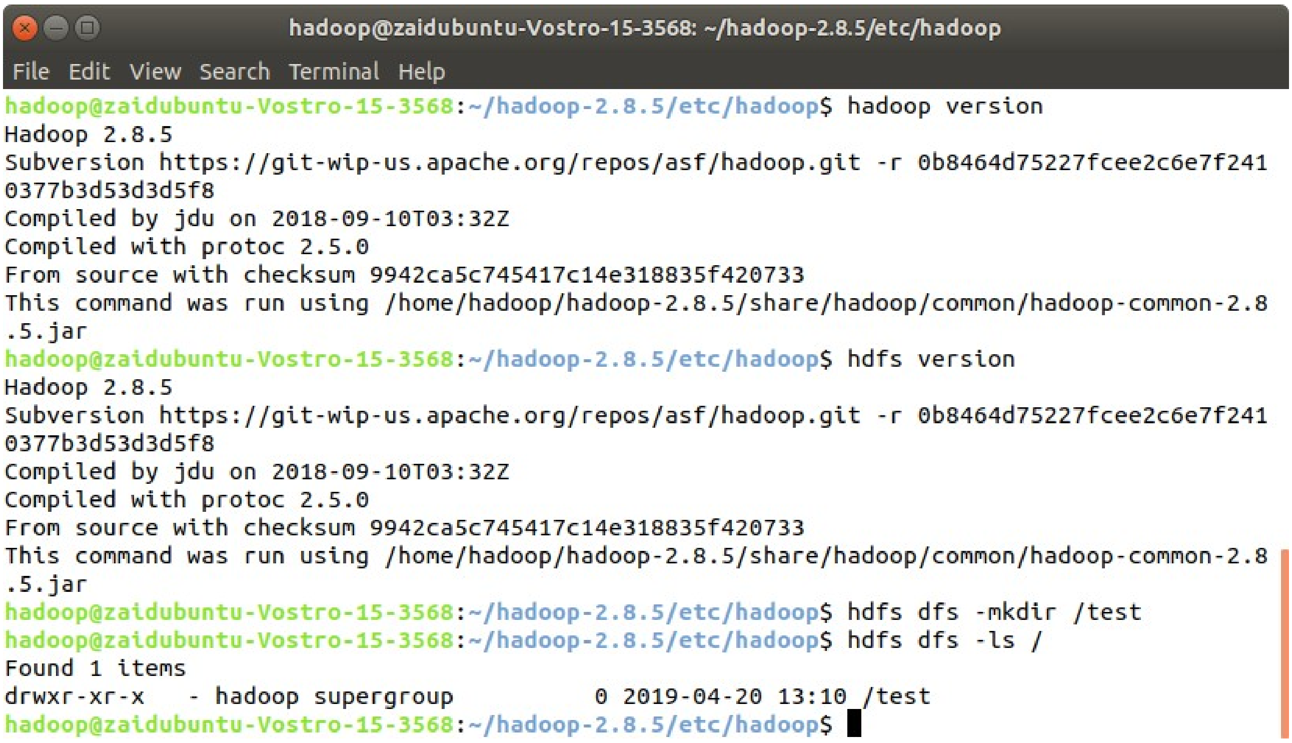
or

```
$ hdfs version
```

HDFS Command Line Interface

To access the HDFS and create some directories top of DFS you can use HDFS CLI.

```
$ hdfs dfs -mkdir /test


$ hdfs dfs -ls /
```


hadoop@zaidubuntu-Vostro-15-3568: ~/hadoop-2.8.5/etc/hadoop
File Edit View Search Terminal Help

```
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ hadoop version
Hadoop 2.8.5
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 0b8464d75227fcee2c6e7f241
0377b3d53d3d5f8
Compiled by jdu on 2018-09-10T03:32Z
Compiled with protoc 2.5.0
From source with checksum 9942ca5c745417c14e318835f420733
This command was run using /home/hadoop/hadoop-2.8.5/share/hadoop/common/hadoop-common-2.8
.5.jar
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ hdfs version
Hadoop 2.8.5
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 0b8464d75227fcee2c6e7f241
0377b3d53d3d5f8
Compiled by jdu on 2018-09-10T03:32Z
Compiled with protoc 2.5.0
From source with checksum 9942ca5c745417c14e318835f420733
This command was run using /home/hadoop/hadoop-2.8.5/share/hadoop/common/hadoop-common-2.8
.5.jar
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ hdfs dfs -mkdir /test
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - hadoop supergroup          0 2019-04-20 13:10 /test
hadoop@zaidubuntu-Vostro-15-3568:~/hadoop-2.8.5/etc/hadoop$
```

Overview of Hadoop Cluster :

INSTALLATION OF HADOOP ON MULTIPLE MACHINES

One (1) Name Node : 192.168.1.1 (hadoopmaster)

Three (3) Data Nodes : 192.168.1.2 (hadoopslave1), 192.168.1.3 (hadoopslave2) 192.168.4 (hadoop slave3)

After installation of SINGLE NODE HADOOP CLUSTER. You are going to CLONE that ubuntu image and named it hadoopmaster.

Open terminal and run "ifconfig" command to see the IPv4 Address.

If it is IPv6 then you have to disable the IPv6 address. Here is a link of tutorial of how to disable IPv6 address - http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/#disabling-ipv6.

Now change the host file -->> "$ sudo gedit /etc/hosts".

Add the following lines :

```
hadoopmaster  192.168.1.1

hadoopslave1   192.168.1.2

hadoopslave2   192.168.1.3

hadoopslave3   192.168.1.4
```

Change the hostname --->> "$ sudo gedit /etc/hostname'.

Add : hadoopmaster.

Go to hadoop director and do changes in its files : "$ cd /usr/local/hadoop/etc/hadoop".

Edit core site xml file ------>> "$ sudo gedit core-site.xml".

replace localhost as hadoopmaster.

Edit hdfs site xml file ------>> $ sudo gedit hdfs-site.xml.

replace value 1 as 3 (represents no of datanode).

Edit yarn site xml file ------>> $ sudo gedit yarn-site.xml.

Add these properties files inside Configuration tag :

```
<property>

 <name>yarn.resourcemanager.resource-tracker.address</name>

 <value>hadoopmaster:8025</value>
```

```
        <property>

        <property>

                <name>yarn.resourcemanager.scheduler.address</name>

                <value>hadoopmaster:8030</value>

        <property>

        <property>

                <name>yarn.resourcemanager.address</name>

                <value>hadoopmaster:8050</value>

         </property>
```

Edit couple of things in yarn site xml file ----------------->> $ sudo gedit yarn-site.xml.

replace mapreduce.framework.name as mapred.job.tracker

replace yarn as hadoopmaster:54311

SHUT DOWN the hadoopmaster Ubuntu Image.

----

master node setup complete

Clone hadoopmaster Node as hadoopslave1, hadoopslave2, hadoopslave3.

Change hadoop master host : $ sudo gedit /usr/local/hadoop/etc/hadoop/master.

replace localhost to hadoopmaster.

Change hadoop slaves : $ sudo gedit /usr/local/hadoop/etc/hadoop/slave.

replace localhost to hadoopslave1 \n hadoopslave2 \n hadoopslave3.

Change hdfs site xml file -------->> $ sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml.

remove dfs.datanode.data.dir property section.

Initial Network setup -

In a virtual machine IDE.

select hadoopmaster ubuntu image and go to its settings.

Go to network

choose attached to option as "internal network".

Give name : "hadoop multinode network".

Go to its advanced settings.

"Allow all" --- promiscuous mode.

Do this for all the slave machines as well.

Open all 3 slave nodes and run ------->> $ sudo gedit /etc/hostname.

replace hadoopmaster to hadoopslave1, hadoopslav2, hadoopslave3 respectively to all the three slave virtual machines.

Reboot all the slave nodes/machines.

On hadoopmaster node run below command to remove all the hadoop data:

"Remove hadoop data ------>> $ sudo rm -rf /usr/local/hadoop/hadoop_data."

On hadoopmaster node === >> $ sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/namenode

Run this command ------->> $ sudo chown -R username:username /usr/local/hadoop

On all hadoopslave nodes ===>> Run following commands ===>>

$ sudo rm -rf /usr/local/hadoop/hadoop_data

$ sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/datanode.

$ sudo chown -R username:username /usr/local/hadoop.

Change hdfs site xml file for all slave nodes ----->>$ sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml.

remove dfs.namenode.data.dir property section.

On hadoopmaster node,

Run the command ---->>$ sudo ssh-copy-id -i ~/.ssh/id_dsa.pub username@hadoopmaster.

If you get error then solution of your problem is :

OpenSSH is not installed. For installation : sudo apt-get install openssh-client.

OR you will get this error "permission denied for root@localhost for ssh connection" .

Solution of 2nd problem is : http://askubuntu.com/questions/497895/permission-denied-for-rootlocalhost-for-ssh-connection.

Next problem might be your Internal network is not setup.

-------------------------Internal Network Setup between all the 4 virtual machine---

Click on top right WIFI or Internet icon.

Go to edit connection.

click on add/edit connection

Give connection name : "master connection"

Go to IPv4 settings.

Change method from automatic to manual.

Enter IP Address like for master node : 192.168.1.1

Enter net mask address : 255.255.255.0

save it and do the above steps for all the nodes.

On hadoopmaster machine -

Run following commands -

```
$ sudo ssh-copy-id -i ~/.ssh/id_dsa.pub chaalpritam@hadoopmaster

$ sudo ssh-copy-id -i ~/.ssh/id_dsa.pub chaalpritam@hadoopslave1

$ sudo ssh-copy-id -i ~/.ssh/id_dsa.pub chaalpritam@hadoopslave2

$ sudo ssh-copy-id -i ~/.ssh/id_dsa.pub chaalpritam@hadoopslave3
```

Now we can access the machines using SSH -

```
$ sudo ssh hadoopmaster

$ exit

$ sudo ssh hadoopslave1
$ exit

$ sudo ssh hadoopslave2

$ exit

$ sudo ssh hadoopslave3

$ exit
```

If you able to do this you are accessing all the nodes using SSH.

Next, Format namenode and start hadoop -

```
$ hadoop namenode -format

$ start-all.sh

$ jps (check in all 3 datanodes)
```

http://hadoopmaster:8088/

http://hadoopmaster:50070/

http://hadoopmaster:50090/

http://hadoopmaster:50075/

Conclusion:

We started installing Hadoop on a pseudo node/single node on our laptops. Then we tried installing Hadoop on distributed machine(s). We learned that it is very important checking the versions of Java and that OpenJDK should not be used.

There are also the problem that Output directory should not exist before running and that even a program fails from executing it creates the output directory before terminating and throwing an exception.

To be on the safe side don’t use folders having spaces in them and don’t use names too similar to existing keywords and program names.

The Word count program is the classic Hello World program and does mostly executes without giving any errors.

For Multigrams, different strategies could be used. One being (Generalized) where we could use StringTokenizer to append string in runtime in an optimized manner.