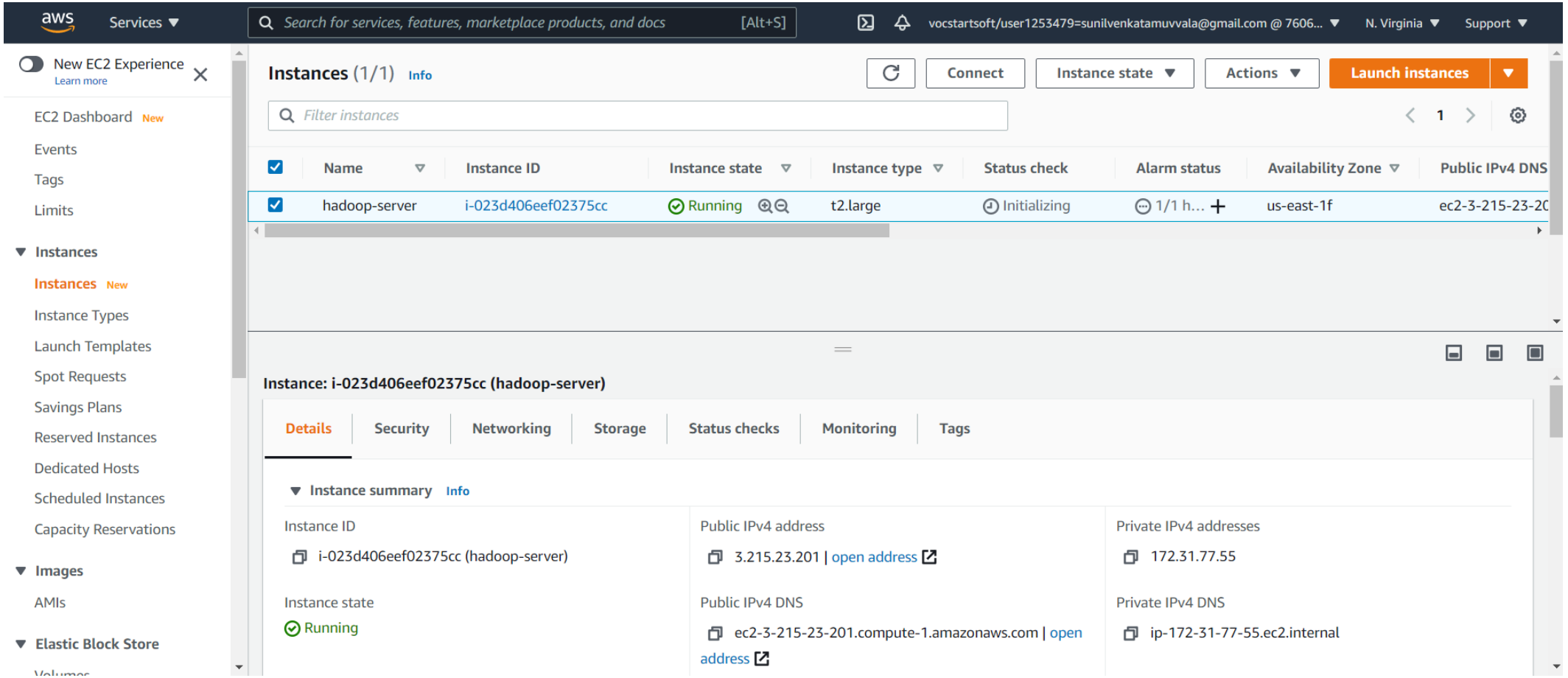
Hadoop and SPARK Installation on aws

Create aws account and launch an ubuntu 18.04 instance with 30 GB SSD, 8 gb RAM and 2 vcpus.
Create a key pair RSA for e.g. Hadoop.pem and download it.



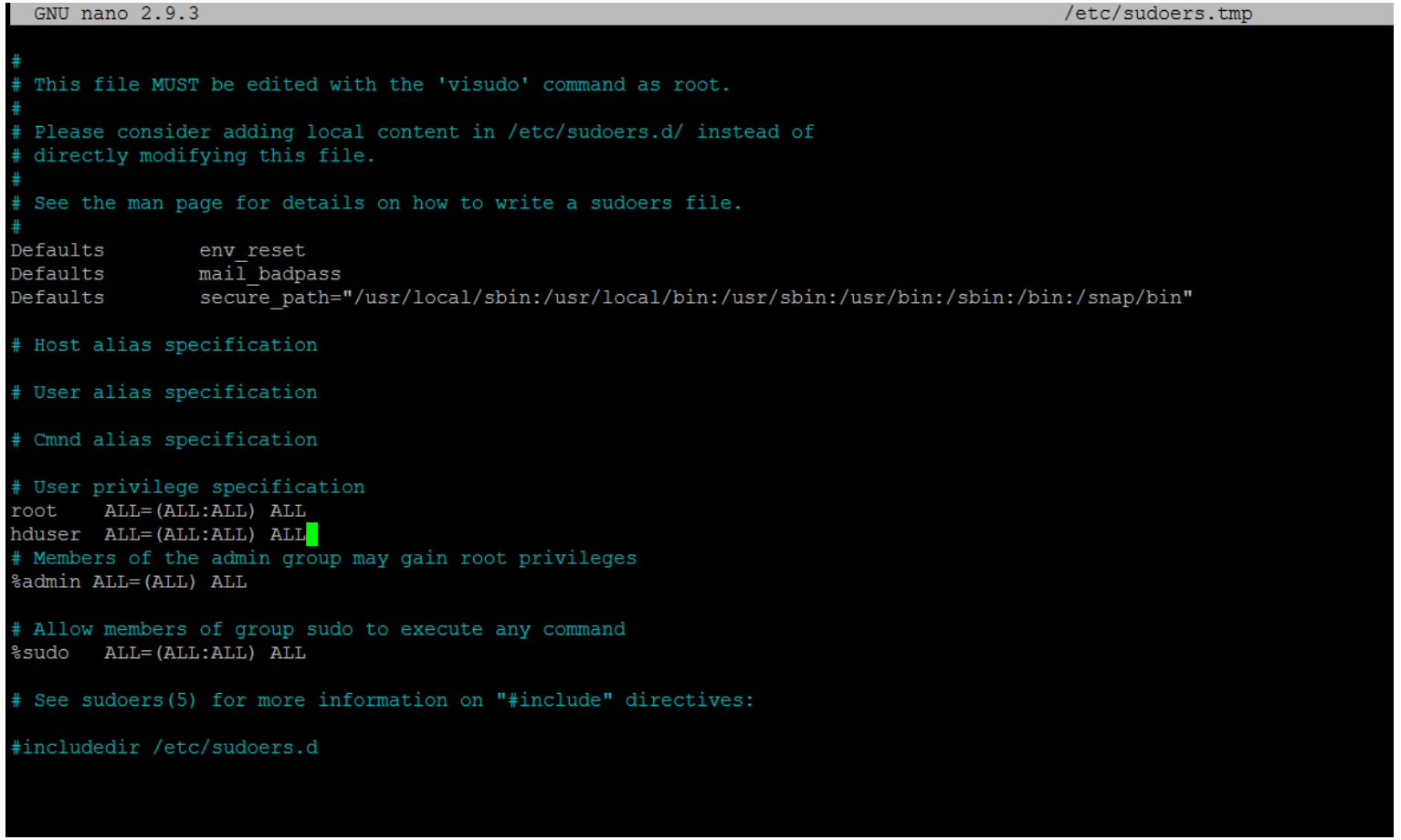
Login using putty to ubuntu instance on cloud with Hadoop.ppk key.
Create a group Hadoop and add user hduser to group

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

  System information as of Mon Feb 15 06:41:30 UTC 2021

  System load:  0.0               Processes:           100
  Usage of /:   3.9% of 29.02GB   Users logged in:     0
  Memory usage: 2%                IP address for eth0: 172.31.77.55
  Swap usage:   0%

 * 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 of these updates are security updates.



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.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-77-55:~$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.
ubuntu@ip-172-31-77-55:~$ sudo adduser --ingroup hadoop hduser
Adding user `hduser' ...
Adding new user `hduser' (1001) with group `hadoop' ...
Creating home directory `/home/hduser' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] Y
ubuntu@ip-172-31-77-55:~$
```

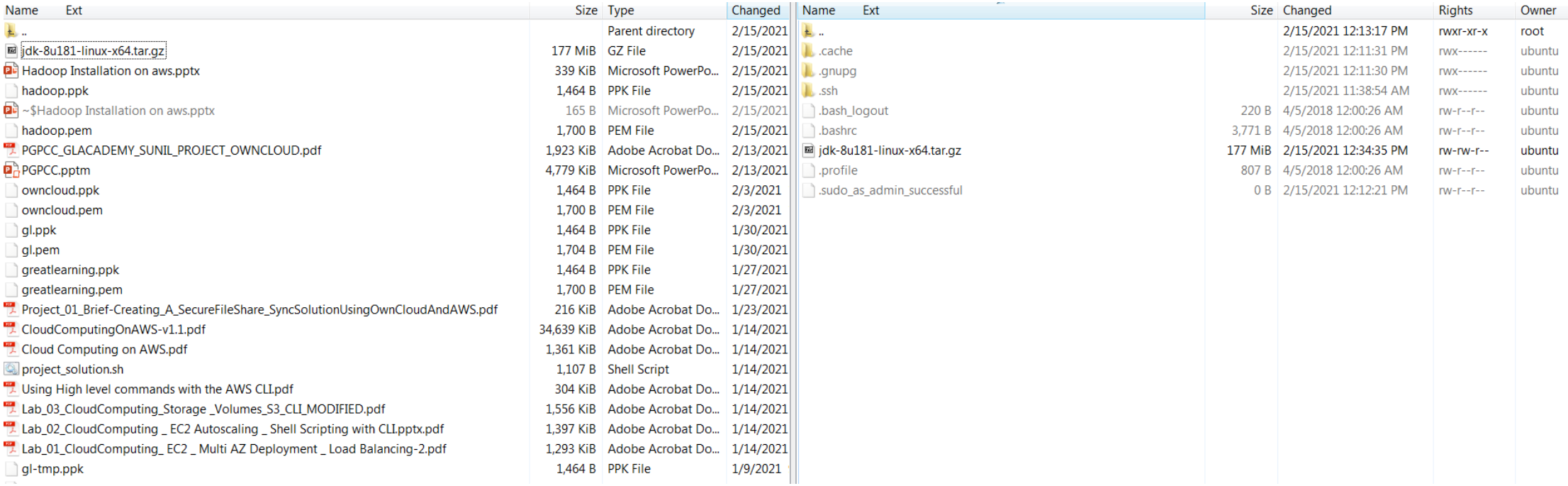
Add hduser to sudoers list to have root priveleges

#sudo visudo

add hduser as shown below.



Download and copy the jdk 8.1 tar file to home/ubuntu location. Minimum pre requisite for Hadoop 3 is jdk 8.1



Name	Ext	Size	Type	Changed
..			Parent directory	2/15/2021
jdk-8u181-linux-x64.tar.gz		177 MiB	GZ File	2/15/2021
Hadoop Installation on aws.pptx		339 KiB	Microsoft PowerPo...	2/15/2021
hadoop.ppk		1,464 B	PPK File	2/15/2021
~$Hadoop Installation on aws.pptx		165 B	Microsoft PowerPo...	2/15/2021
hadoop.pem		1,700 B	PEM File	2/15/2021
PGPCC_GLACADEMY_SUNIL_PROJECT_OWNCLOUD.pdf		1,923 KiB	Adobe Acrobat Do...	2/13/2021
PGPCC.pptm		4,779 KiB	Microsoft PowerPo...	2/13/2021
owncloud.ppk		1,464 B	PPK File	2/3/2021
owncloud.pem		1,700 B	PEM File	2/3/2021
gl.ppk		1,464 B	PPK File	1/30/2021
gl.pem		1,704 B	PEM File	1/30/2021
greatlearning.ppk		1,464 B	PPK File	1/27/2021
greatlearning.pem		1,700 B	PEM File	1/27/2021
Project_01_Brief-Creating_A_SecureFileShare_SyncSolutionUsingOwnCloudAndAWS.pdf		216 KiB	Adobe Acrobat Do...	1/23/2021
CloudComputingOnAWS-v1.1.pdf		34,639 KiB	Adobe Acrobat Do...	1/14/2021
Cloud Computing on AWS.pdf		1,361 KiB	Adobe Acrobat Do...	1/14/2021
project_solution.sh		1,107 B	Shell Script	1/14/2021
Using High level commands with the AWS CLI.pdf		304 KiB	Adobe Acrobat Do...	1/14/2021
Lab_03_CloudComputing_Storage _Volumes_S3_CLI_MODIFIED.pdf		1,556 KiB	Adobe Acrobat Do...	1/14/2021
Lab_02_CloudComputing _ EC2 Autoscaling _ Shell Scripting with CLI.pptx.pdf		1,397 KiB	Adobe Acrobat Do...	1/14/2021
Lab_01_CloudComputing_ EC2 _ Multi AZ Deployment _ Load Balancing-2.pdf		1,293 KiB	Adobe Acrobat Do...	1/14/2021
gl-tmp.ppk		1,464 B	PPK File	1/9/2021

Name	Ext	Size	Changed	Rights	Owner
..			2/15/2021 12:13:17 PM	rwxr-xr-x	root
.cache			2/15/2021 12:11:31 PM	rwx------	ubuntu
.gnupg			2/15/2021 12:11:30 PM	rwx------	ubuntu
.ssh			2/15/2021 11:38:54 AM	rwx------	ubuntu
.bash_logout		220 B	4/5/2018 12:00:26 AM	rw-r--r--	ubuntu
.bashrc		3,771 B	4/5/2018 12:00:26 AM	rw-r--r--	ubuntu
jdk-8u181-linux-x64.tar.gz		177 MiB	2/15/2021 12:34:35 PM	rw-rw-r--	ubuntu
.profile		807 B	4/5/2018 12:00:26 AM	rw-r--r--	ubuntu
.sudo_as_admin_successful		0 B	2/15/2021 12:12:21 PM	rw-r--r--	ubuntu

copy the jdk tar file to usr/local where we can keep all installations. Here sudo command helps to copy to usr/local folder with root priveleges.

```
hduser@ip-172-31-77-55:/home/ubuntu$ sudo su
[sudo] password for hduser:
root@ip-172-31-77-55:/home/ubuntu# cp /home/ubuntu/jdk-8u181-linux-x64.tar.gz /usr/local/
root@ip-172-31-77-55:/home/ubuntu# ls /usr/local/
bin  etc  games  include  jdk-8u181-linux-x64.tar.gz  lib  man  sbin  share  src
root@ip-172-31-77-55:/home/ubuntu# exit
exit
```

Extract tar file with sudo priveleges.

```
hduser@ip-172-31-77-55:/home/ubuntu$ cd /usr/local
hduser@ip-172-31-77-55:/usr/local$ sudo tar jdk-8u181-linux-x64.tar.gz
tar: Old option 'g' requires an argument.
Try 'tar --help' or 'tar --usage' for more information.
hduser@ip-172-31-77-55:/usr/local$ sudo tar xvzf jdk-8u181-linux-x64.tar.gz
jdk1.8.0_181/
jdk1.8.0_181/javafx-src.zip
jdk1.8.0_181/bin/
jdk1.8.0_181/bin/jmc
jdk1.8.0_181/bin/serialver
jdk1.8.0_181/bin/jmc.ini
jdk1.8.0_181/bin/jstack
jdk1.8.0_181/bin/rmiregistry
jdk1.8.0_181/bin/unpack200
jdk1.8.0_181/bin/jar
jdk1.8.0_181/bin/jps
jdk1.8.0_181/bin/wsimport
jdk1.8.0_181/bin/rmic
jdk1.8.0_181/bin/jdeps
jdk1.8.0_181/bin/jcontrol
jdk1.8.0_181/bin/javafxpackager
jdk1.8.0_181/bin/schemagen
jdk1.8.0_181/bin/jcmd
jdk1.8.0_181/bin/servertool
jdk1.8.0_181/bin/xjc
jdk1.8.0_181/bin/jmap
jdk1.8.0_181/bin/jvisualvm
jdk1.8.0_181/bin/policytool
jdk1.8.0_181/bin/jstat
jdk1.8.0_181/bin/jconsole
jdk1.8.0_181/bin/jdb
```

Rename java.1.8.0_141 to java for ease.
Add environment variables and export the variables in the .bashrc file using source command for the logged in session.

```
hduser@ip-172-31-77-55:/usr/local$ sudo mv jdk1.8.0_181 java
hduser@ip-172-31-77-55:/usr/local$ ls
bin  etc  games  include  java  jdk-8u181-linux-x64.tar.gz  lib  man  sbin  share  src
hduser@ip-172-31-77-55:/usr/local$ cd ~
hduser@ip-172-31-77-55:~$ sudo nano ~/bashrc
hduser@ip-172-31-77-55:~$ sudo nano ~/.bashrc
hduser@ip-172-31-77-55:~$ source ~/.bashrc
hduser@ip-172-31-77-55:~$
```

Update alternatives so that the current version of the java is used if they are multiple versions are used.

```
hduser@ip-172-31-77-55:~$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/local/java/bin/java" 1
update-alternatives: using /usr/local/java/bin/java to provide /usr/bin/java (java) in auto mode
hduser@ip-172-31-77-55:~$ sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/local/java/bin/javac" 1
update-alternatives: using /usr/local/java/bin/javac to provide /usr/bin/javac (javac) in auto mode
hduser@ip-172-31-77-55:~$ sudo update-alternatives --install "usr/bin/javaws" "javaws" "/usr/local/java/bin/javaws" 1
update-alternatives: error: alternative link is not absolute as it should be: usr/bin/javaws
hduser@ip-172-31-77-55:~$ sudo update-alternatives --install "/usr/bin/javaws" "javaws" "/usr/local/java/bin/javaws" 1
update-alternatives: using /usr/local/java/bin/javaws to provide /usr/bin/javaws (javaws) in auto mode
hduser@ip-172-31-77-55:~$ #verify java installation
hduser@ip-172-31-77-55:~$ sudo update-alternatives --set java /usr/local/java/bin/java
hduser@ip-172-31-77-55:~$ sudo update-alternatives --set javac /usr/local/java/bin/javac
hduser@ip-172-31-77-55:~$ sudo update-alternatives --set javaws /usr/local/java/bin/javaws
hduser@ip-172-31-77-55:~$ java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)
```

Password less ssh is required when Hadoop is installed because node manager, resource manager will be running in different JVMs will get on to local host where password less SSH is required.
Generate a public/private key and store the public key on to authorized keys and change the permissions.

```
ubuntu@ip-172-31-77-55:~$ su hduser
Password:
hduser@ip-172-31-77-55:/home/ubuntu$ cd /home/hduser/
hduser@ip-172-31-77-55:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:nms9jc2BvhJHoZrMs2ujhobHpV9lJn5dmN0Eee+1g4k.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
hduser@localhost: Permission denied (publickey).
hduser@ip-172-31-77-55:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:2WxNU3YIabvFKg7oSJqHTgJyap/X/qSCePxSI6lNbf8 hduser@ip-172-31-77-55
The key's randomart image is:
+---[RSA 2048]----+
|             .oo..|
|            oo.. |
|           .oo   |
|         + o..o  |
|o . o  .S + .+   |
|oo +.=. ... o    |
|o.B==o+  + .     |
|.==*+o.oo .      |
| .o++.oooE       |
+----[SHA256]-----+
hduser@ip-172-31-77-55:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hduser@ip-172-31-77-55:~$ chmod 0600 ~/.ssh/authorized_keys
hduser@ip-172-31-77-55:~$ ssh localhost
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 5.4.0-1037-aws x86_64)

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

  System information as of Mon Feb 15 10:15:13 UTC 2021

  System load:  0.0               Processes:           118
  Usage of /:   6.4% of 29.02GB   Users logged in:     1
  Memory usage: 3%                IP address for eth0: 172.31.77.55
  Swap usage:   0%
```

Download Hadoop-3.0.2 on to usr/local directory

```
hduser@ip-172-31-77-55:/usr/local$ sudo wget https://www-eu.apache.org/dist/haddop/common/hadoop-3.0.2/hadoop-3.0.2.tar.gz
--2021-02-15 10:27:35--  https://www-eu.apache.org/dist/haddop/common/hadoop-3.0.2/hadoop-3.0.2.tar.gz
Resolving www-eu.apache.org (www-eu.apache.org)... 95.216.26.30, 2a01:4f9:2a:1a61::2
Connecting to www-eu.apache.org (www-eu.apache.org)|95.216.26.30|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://downloads.apache.org/haddop/common/hadoop-3.0.2/hadoop-3.0.2.tar.gz [following]
--2021-02-15 10:27:36--  https://downloads.apache.org/haddop/common/hadoop-3.0.2/hadoop-3.0.2.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2021-02-15 10:27:36 ERROR 404: Not Found.

hduser@ip-172-31-77-55:/usr/local$ sudo wget https://archive.apache.org/dist/hadoop/core/hadoop-3.0.2/hadoop-3.0.2.tar.gz
--2021-02-15 10:31:22--  https://archive.apache.org/dist/hadoop/core/hadoop-3.0.2/hadoop-3.0.2.tar.gz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 307618649 (293M) [application/x-gzip]
Saving to: 'hadoop-3.0.2.tar.gz'

hadoop-3.0.2.tar.gz                         100%[=============================================================================================>] 293.37M  15.8MB/s    in 20s

2021-02-15 10:31:43 (14.3 MB/s) - 'hadoop-3.0.2.tar.gz' saved [307618649/307618649]
```

Extract the tar file and rename the folder to Hadoop
give complete permissions and ownership to hduser on
/usr/local/Hadoop folder

```
hadoop-3.0.2/bin/mapred.cmd
hadoop-3.0.2/bin/hdfs
hduser@ip-172-31-77-55:/usr/local$ sudo mv hadoop-3.0.2 hadoop
hduser@ip-172-31-77-55:/usr/local$ sudo chown -R hduser:hadoop /usr/local/hadoop
hduser@ip-172-31-77-55:/usr/local$ sudo chmod -R 777 /usr/local/hadoop
hduser@ip-172-31-77-55:/usr/local$
```

Open /etc/sysctl.conf file and add the below lines and save to disable ipv6 as Hadoop works with ipv4

```
  GNU nano 2.9.3

# Accept ICMP redirects only for gateways listed in our default
# gateway list (enabled by default)
# net.ipv4.conf.all.secure_redirects = 1
#
# Do not send ICMP redirects (we are not a router)
#net.ipv4.conf.all.send_redirects = 0
#
# Do not accept IP source route packets (we are not a router)
#net.ipv4.conf.all.accept_source_route = 0
#net.ipv6.conf.all.accept_source_route = 0
#
# Log Martian Packets
#net.ipv4.conf.all.log_martians = 1
#

###################################################################
# Magic system request Key
# 0=disable, 1=enable all
# Debian kernels have this set to 0 (disable the key)
# See https://www.kernel.org/doc/Documentation/sysrq.txt
# for what other values do
#kernel.sysrq=1

###################################################################
# Protected links
#
# Protects against creating or following links under certain conditions
# Debian kernels have both set to 1 (restricted)
# See https://www.kernel.org/doc/Documentation/sysctl/fs.txt
#fs.protected_hardlinks=0
#fs.protected_symlinks=0
net.ipv6.conf.all.disable_ipv6=1
net.ipv6.conf.default.disable_ipv6=1
net.ipv6.conf.lo.disable_ipv6=1
```

Run the command to see that ipv6 is disabled.
Open the /.bashrc and save the environment variables for Hadoop.
Source the environment variables.

```
hduser@ip-172-31-77-55:/usr/local$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
0
hduser@ip-172-31-77-55:/usr/local$
```

Add environment variables to ~/.bashrc file and source the file.

```
 GNU nano 2.9.3                                                    /home/hduser/.bashrc

alias l='ls -CF'

# Add an "alert" alias for long running commands.  Use like so:
#   sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || echo error)" "$(history|tail -n1|sed -e '\''s/^\s*[0-9]\+\s*//;s/[;&|]\s*alert$//'\'')"'

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# 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 JAVA_HOME=/usr/local/java
export PATH=$PATH:/usr/local/java/bin
#HADOOP ENVIRONMENT
export HADOOP_PREFIX=/usr/local/hadoop
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
export HADOOP_MAPRED_HOME=/usr/local/hadoop
export HADOOP_COMMON_HOME=/usr/local/hadoop
export HADOOP_HDFS_HOME=/usr/local/hadoop
export YARN_HOME=/usr/local/hadoop
export PATH=$PATH:/usr/local/hadoop/bin
export PATH=$PATH:/usr/local/hadoop/sbin
#HADOOP NATIVE PATH:
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_PREFIX/lib"
```

Set the hdaoop-env.sh,yran-site,hdfs-site,core-site,mapred-site files

```
[sudo] password for hduser:
hduser@ip-172-31-77-55:/usr/local$ cd /usr/local/hadoop/etc/hadoop/
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.sh
[sudo] password for hduser:
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano yarn-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ source ~/.bashrc
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano hdfs-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$
```

Hadoop-env.
Set java_home and ipv4 enabled,suppress warnings

```
  GNU nano 2.9.3                                                    hadoop-env.sh

# Specify the JVM options to be used when starting the HDFS Mover.
# These options will be appended to the options specified as HADOOP_OPTS
# and therefore may override any similar flags set in HADOOP_OPTS
#
# export HDFS_MOVER_OPTS=""

###
# Router-based HDFS Federation specific parameters
# Specify the JVM options to be used when starting the RBF Routers.
# These options will be appended to the options specified as HADOOP_OPTS
# and therefore may override any similar flags set in HADOOP_OPTS
#
# export HDFS_DFSROUTER_OPTS=""
###

###
# Advanced Users Only!
###

#
# When building Hadoop, one can add the class paths to the commands
# via this special env var:
# export HADOOP_ENABLE_BUILD_PATHS="true"

#
# To prevent accidents, shell commands be (superficially) locked
# to only allow certain users to execute certain subcommands.
# It uses the format of (command)_(subcommand)_USER.
#
# For example, to limit who can execute the namenode command,
# export HDFS_NAMENODE_USER=hdfs
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
export JAVA_HOME=/usr/local/java
export HADOOP_HOME_WARN_SUPPRESS="TRUE"
export HADOOP_ROOT_LOGGER="WARN,DRFA"
```

Set map_reduce in Hadoop 3 which is taken care by yarn

```
 GNU nano 2.9.3                                                    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>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

</configuration>
```

Set replication factor,name node and data node directories. Here it is on single machine which is pseudo distributed mode i.e. on single machines the processes run on multiple JVMs

GNU nano 2.9.3 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:/usr/local/hadoop/yarn_data/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop/yarn_data/hdfs/datanode</value>
</property>
</configuration>
```

Core-site.xml
set Hadoop temp directory and name node port to default 9000

```
 GNU nano 2.9.3                                                  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>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

Mapred-site.xml
set map-red framework is taken care by yarn and map-red job history port is 10020.

```
 GNU nano 2.9.3                                                  mapred-site.xml

<?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>mapred.framework.name</name>
<value>yarn</value>
</property>
<property>
<name>mapreduce.jobhistory.address</name>
<value>localhost:10020</value>
</property>
</configuration>
```

Create tmp,datanode and namenode directories with appropriate permissions. Also format namenode.

```
hduser@ip-172-31-77-55:/usr/local$ cd /usr/local/hadoop/etc/hadoop/
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.sh
[sudo] password for hduser:
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano yarn-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ source ~/.bashrc
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano hdfs-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano core-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo mkdir -p /app/hadoop/tmp
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chown -R hduser:hadoop /app/hadoop/tmp
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chmod -R 777 /app/hadoop/tmp
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo mkdir -p /usr/local/hadoop/yarn_data/hdfs/namenode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo mkdir -p /usr/local/hadoop/yarn_data/hdfs/datanode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chmod -R 777 /usr/local/hadoop/yarn_data/hdfs/namenode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chmod -R 700 /usr/local/hadoop/yarn_data/hdfs/datanode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chown -R hduser:hadoop /usr/local/hadoop/yarn_data/hdfs/namenode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ sudo chown -R hduser:hadoop /usr/local/hadoop/yarn_data/hdfs/datanode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ hdfs namenode -format
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
WARNING: /usr/local/hadoop/logs does not exist. Creating.
Formatting using clusterid: CID-fd625b6c-fb50-4a6e-8f3f-776e05ec5984
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$
```

check Hadoop version
Hadoop version
Start-dfs.sh and check the java processes running using jps command. It will show
DataNode
Jps
NameNode
SecondaryNameNode

```
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ start-dfs.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting namenodes on [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting datanodes
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting secondary namenodes [ip-172-31-77-55]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ip-172-31-77-55: Warning: Permanently added 'ip-172-31-77-55,172.31.77.55' (ECDSA) to the list of known hosts.
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ jps
17156 DataNode
17541 Jps
16939 NameNode
17420 SecondaryNameNode
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$
```

start-yarn.sh and verify the processes.
Also verify the name node using public ip of ubuntu instance.
For e.g. in browser paste. In Hadoop 2 port is 5870 and here it is 9870
http://3.215.23.201:9870/
resource manager UI
http://3.215.23.201:8088/

```
nduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ start-yarn.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting resourcemanager
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting nodemanagers
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
nduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ jps
18065 NodeManager
17156 DataNode
17700 ResourceManager
16939 NameNode
17420 SecondaryNameNode
18222 Jps
nduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$
```

Spark installation: download spark-2.3.3 and extract spark to /home/hduser directory

```
hduser@ip-172-31-77-55:/usr/local/hadoop/etc/hadoop$ cd ~
hduser@ip-172-31-77-55:~$ pwd
/home/hduser
hduser@ip-172-31-77-55:~$ wget https://archive.apache.org/dist/spark/spark-2.3.3/spark-2.3.3-bin-hadoop2.7.tgz
--2021-02-15 12:59:55--  https://archive.apache.org/dist/spark/spark-2.3.3/spark-2.3.3-bin-hadoop2.7.tgz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 226027370 (216M) [application/x-gzip]
Saving to: 'spark-2.3.3-bin-hadoop2.7.tgz'

spark-2.3.3-bin-hadoop2.7.tgz                 100%[=============================================================================================>] 215.56M  15.6MB/s    in 14s

2021-02-15 13:00:10 (15.1 MB/s) - 'spark-2.3.3-bin-hadoop2.7.tgz' saved [226027370/226027370]

hduser@ip-172-31-77-55:~$ tar -zxvf spark-2.3.3-bin-hadoop2.7.tgz
spark-2.3.3-bin-hadoop2.7/
spark-2.3.3-bin-hadoop2.7/bin/
spark-2.3.3-bin-hadoop2.7/bin/beeline
spark-2.3.3-bin-hadoop2.7/bin/beeline.cmd
spark-2.3.3-bin-hadoop2.7/bin/docker-image-tool.sh
spark-2.3.3-bin-hadoop2.7/bin/find-spark-home
spark-2.3.3-bin-hadoop2.7/bin/find-spark-home.cmd
spark-2.3.3-bin-hadoop2.7/bin/load-spark-env.cmd
spark-2.3.3-bin-hadoop2.7/bin/load-spark-env.sh
spark-2.3.3-bin-hadoop2.7/bin/pyspark
```

Export HADOOP_CONF_DIR and YARN_CONF_DIR in spark bin folder and start the spark shell.

```
hduser@ip-172-31-77-55:~$ cd spark-2.3.3-bin-hadoop2.7
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7$ ls
LICENSE  NOTICE  R  README.md  RELEASE  bin  conf  data  examples  jars  kubernetes  licenses  python  sbin  yarn
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7$ cd bin
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop/
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ export YARN_CONF_DIR=/usr/local/hadoop/etc/hadoop/
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ ls
beeline              find-spark-home      load-spark-env.sh  pyspark2.cmd     spark-class       spark-shell       spark-sql       spark-submit       sparkR
beeline.cmd          find-spark-home.cmd  pyspark            run-example      spark-class.cmd   spark-shell.cmd   spark-sql.cmd   spark-submit.cmd   sparkR.cmd
docker-image-tool.sh load-spark-env.cmd   pyspark.cmd        run-example.cmd  spark-class2.cmd  spark-shell2.cmd  spark-sql2.cmd  spark-submit2.cmd  sparkR2.cmd
```

```
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ ./spark-shell
2021-02-15 13:10:01 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://ip-172-31-77-55.ec2.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1613394608438).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.3
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Install pyspark and run pyspark. Python version check.
python --version

```
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ python

Command 'python' not found, but can be installed with:

apt install python3
apt install python
apt install python-minimal

Ask your administrator to install one of them.

You also have python3 installed, you can run 'python3' instead.

hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ sudo apt install python
[sudo] password for hduser:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  libpython-stdlib libpython2.7-minimal libpython2.7-stdlib python-minimal python2.7 python2.7-minimal
Suggested packages:
  python-doc python-tk python2.7-doc binutils binfmt-support
The following NEW packages will be installed:
  libpython-stdlib libpython2.7-minimal libpython2.7-stdlib python python-minimal python2.7 python2.7-minimal
0 upgraded, 7 newly installed, 0 to remove and 0 not upgraded.
Need to get 3965 kB of archives.
After this operation, 16.8 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 libpython2.7-minimal amd64 2.7.17-1~18.04ubuntu1.2 [335 kB]
Get:2 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 python2.7-minimal amd64 2.7.17-1~18.04ubuntu1.2 [1290 kB]
Get:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic/main amd64 python-minimal amd64 2.7.15~rc1-1 [28.1 kB]
Get:4 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 libpython2.7-stdlib amd64 2.7.17-1~18.04ubuntu1.2 [1916 kB]
Get:5 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 python2.7 amd64 2.7.17-1~18.04ubuntu1.2 [248 kB]
Get:6 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic/main amd64 libpython-stdlib amd64 2.7.15~rc1-1 [7620 B]
Get:7 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic/main amd64 python amd64 2.7.15~rc1-1 [140 kB]
```

```
hduser@ip-172-31-77-55:~/spark-2.3.3-bin-hadoop2.7/bin$ ./pyspark
Python 2.7.17 (default, Sep 30 2020, 13:38:04)
[GCC 7.5.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
2021-02-15 13:13:10 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.3
      /_/

Using Python version 2.7.17 (default, Sep 30 2020 13:38:04)
SparkSession available as 'spark'.
```