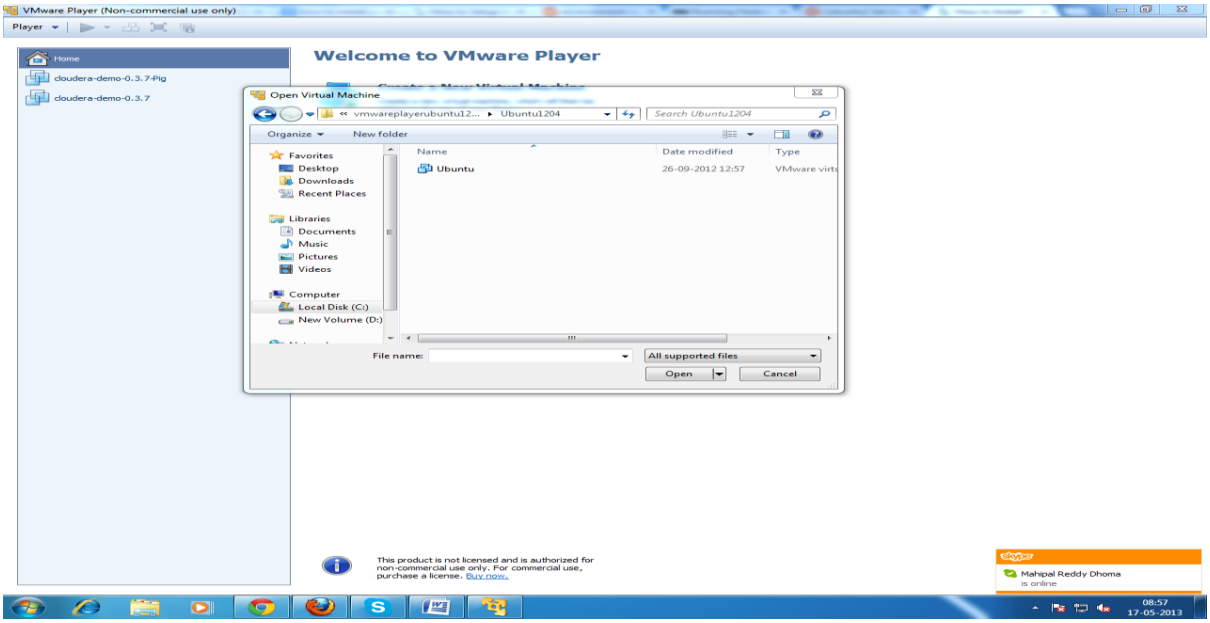
edureka!

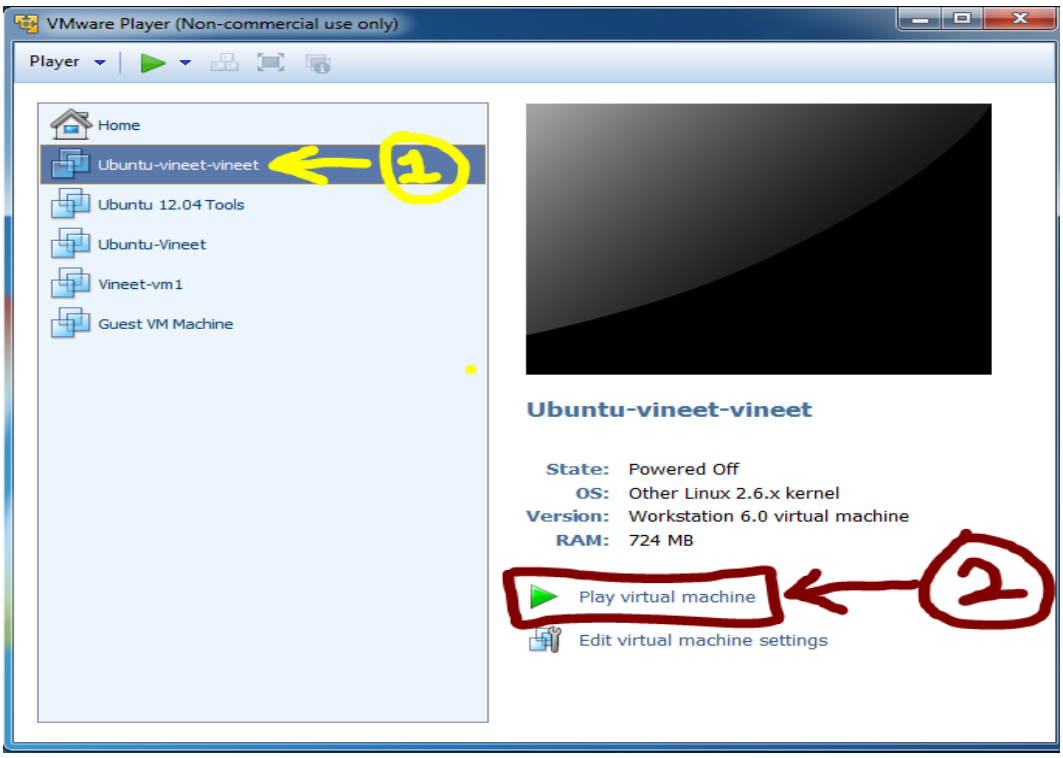
Hadoop Installation with Multiple DataNode

- **Go to the below link and download the image of ubuntu 12.04**
- http://www.traffictool.net/vmware/ubuntu1204t.html

- **Open VMware Player and click open virtual machine and select path where you have extracted image of Ubuntu. After that select the .vmx file and click ok.**

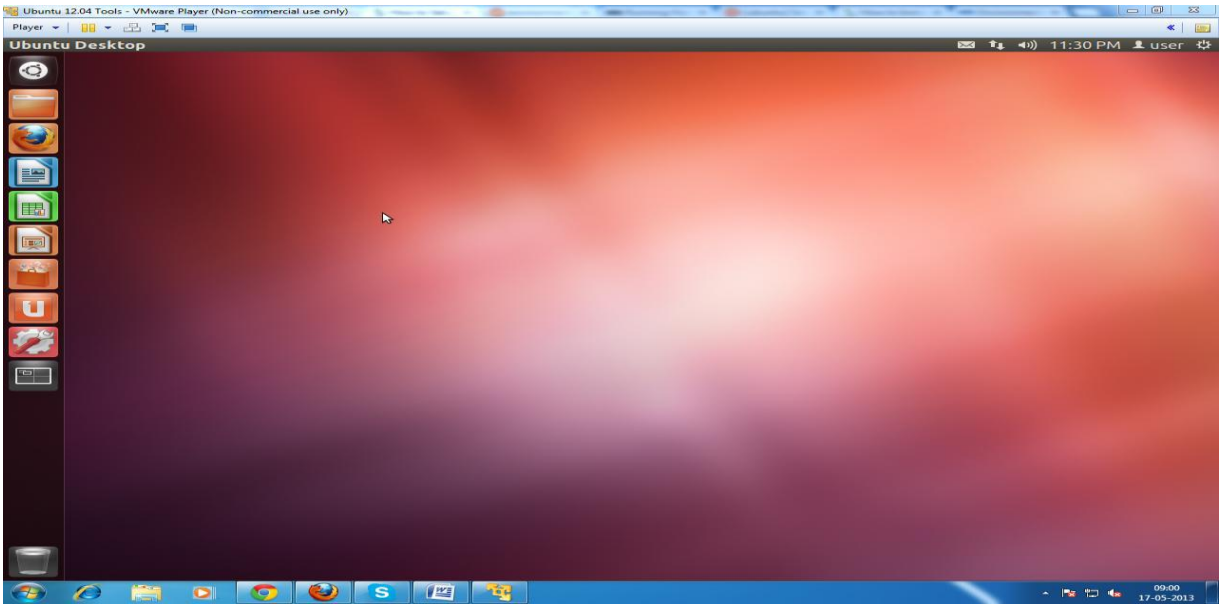


- **Now you can see the below screen in VMware Player.**

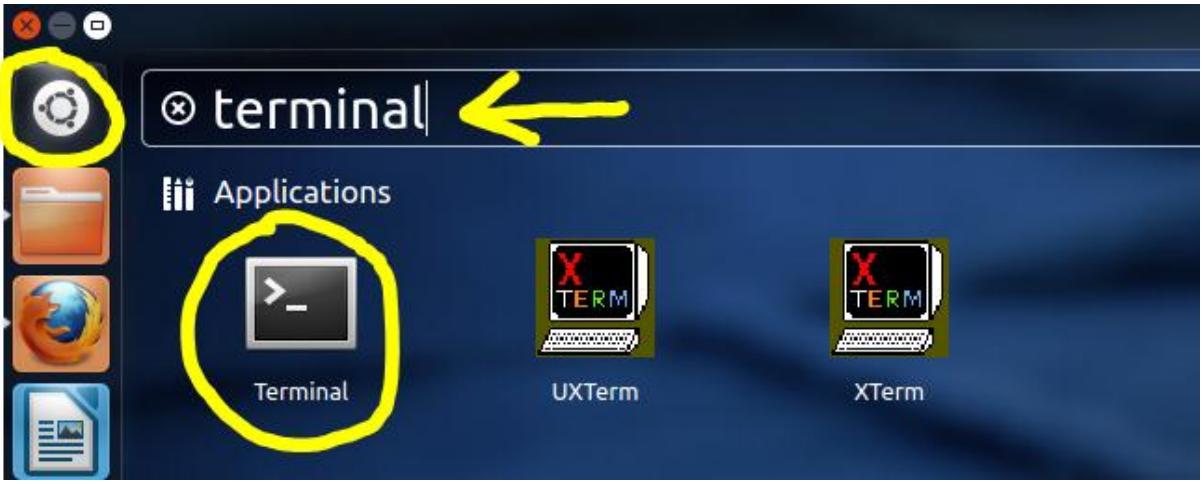


- **Double click on ubuntu present in VMware Player. You will get a screen of the below image.**

Username : **user**
Password : **password**



- **Open a Terminal**



- **Update the repository:**

- **Command:** sudo apt-get update

```
user@ubuntu:~$ sudo apt-get update
```

- **Once the Update is complete :**
- **Command:** sudo apt-get install openjdk-6-jdk

```
user@ubuntu:~$ sudo apt-get install openjdk-6-jdk
```

- **After Java has been Installed, To check whether Java is installed on your system or not give the below command :**
- **Command:** java -version

```
user@ubuntu:~$ java -version
java version "1.6.0_27"
OpenJDK Runtime Environment (IcedTea6 1.12.5) (6b27-1.12.5-0ubuntu0.12.04.1)
OpenJDK Client VM (build 20.0-b12, mixed mode, sharing)
user@ubuntu:~$
```

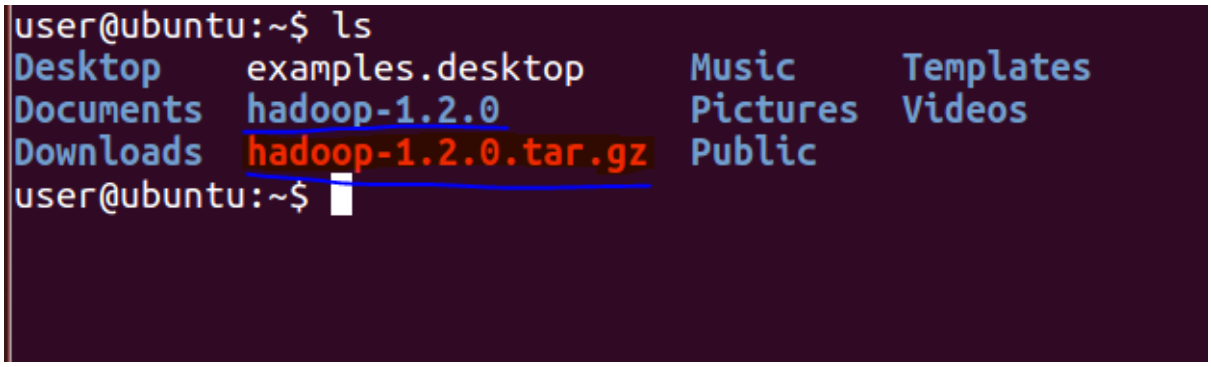
- **Install openssh-server:**
- **Command:** sudo apt-get install openssh-server

```
user@ubuntu: ~
user@ubuntu:~$ sudo apt-get install openssh-server
[sudo] password for user:
```

```
user@ubuntu: ~
user@ubuntu:~$ sudo apt-get install openssh-server
[sudo] password for user:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following extra packages will be installed:
  openssh-client ssh-import-id
Suggested packages:
  libpam-ssh keychain monkeysphere openssh-blacklist openssh-blacklist-extra
  rssh molly-guard
The following NEW packages will be installed:
  openssh-server ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 2 newly installed, 0 to remove and 562 not upgraded.
Need to get 1,309 kB of archives.
After this operation, 891 kB of additional disk space will be used.
Do you want to continue [Y/n]? Y
```

- **Download and extract Hadoop:**
- **Command:** wget http://archive.apache.org/dist/hadoop/core/hadoop-1.2.0/hadoop-1.2.0.tar.gz

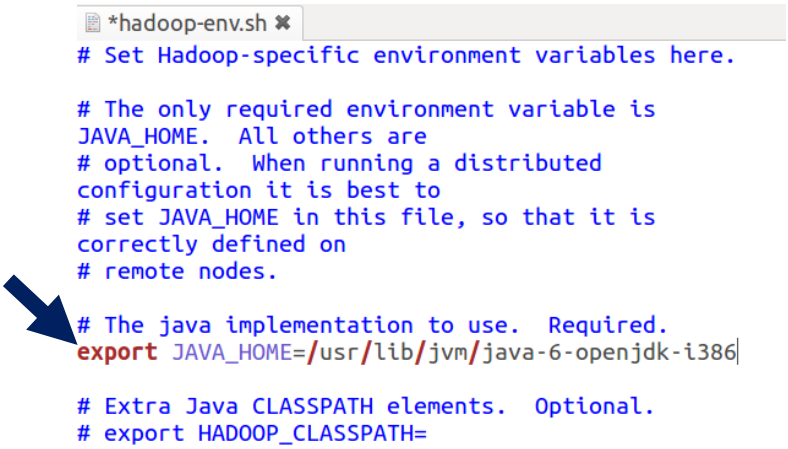
- **Command: tar -xvf hadoop-1.2.0.tar.gz**



Configuration

- **Add JAVA_HOME in hadoop-env.sh file:**
- **Command:** sudo gedit hadoop-1.2.0/conf/hadoop-env.sh
- **Type :** export JAVA_HOME=/usr/lib/jvm/java-6-openjdk-i386

Uncomment the below shown export and add the below the path to your JAVA_HOME:

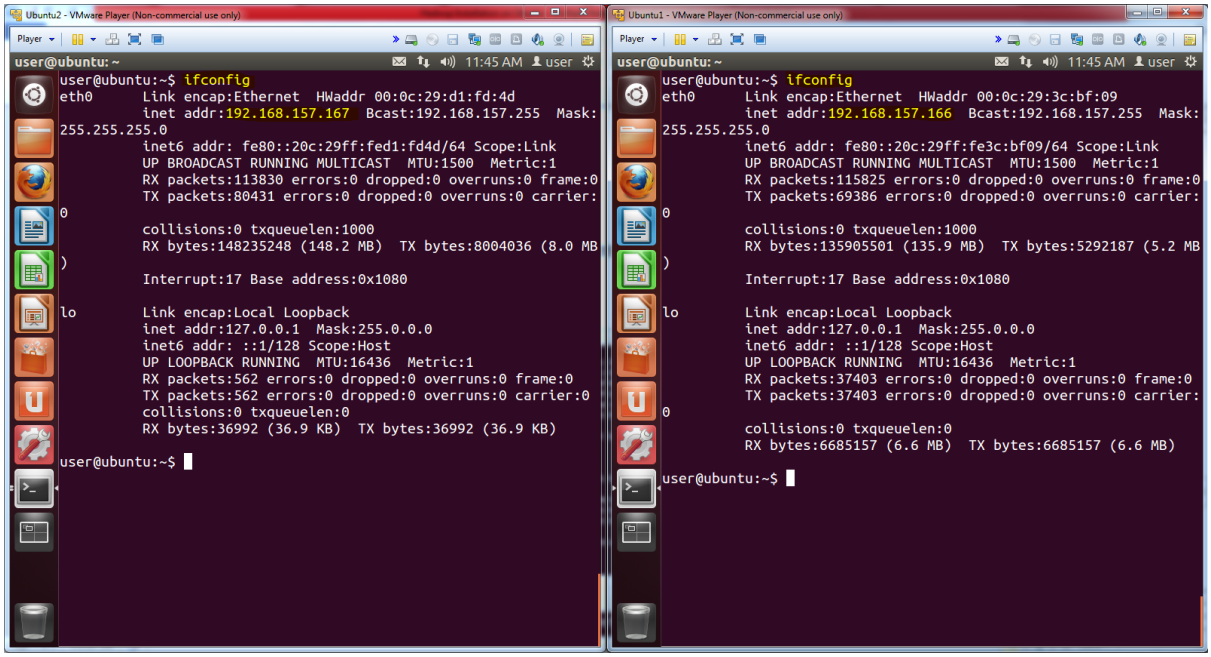


Now create another instance of Ubuntu VM and start it on VMPlayer.

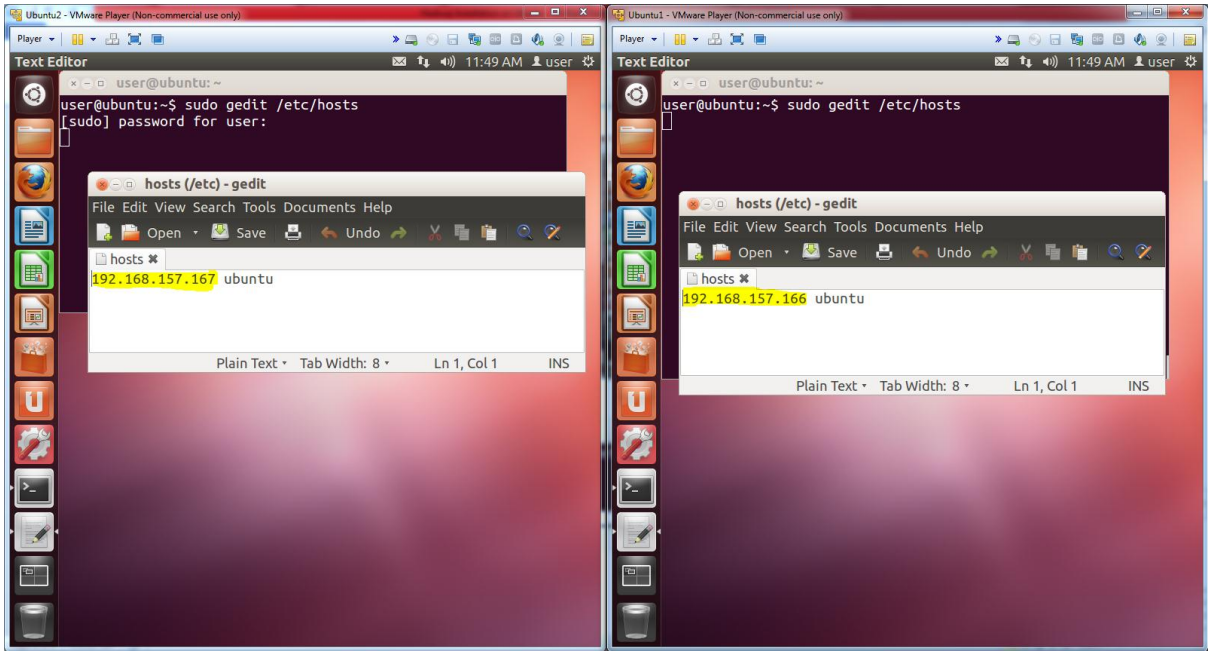
[Important – Do all the above steps for second Image of Ubuntu]

edureka!

- **Getting ip addresses for both VMs:**
- **Command:** ifconfig
- **Command:** sudo gedit /etc/hosts



Getting ip addresses for both VMs:



Select a VM to assign as master and note its ip address.

I have chosen VM 1 whose ip address is 192.168.157.166

[Note: core-site.xml, mapred-site.xml and hdfs-site.xml are same for both VMs, you need to set above 3 files on both VM]

- **Edit core-site.xml:**
- **Command:** sudo gedit Hadoop-1.2.0/conf/core-site.xml

<property>

<name>fs.default.name</name>

<value>hdfs:// 192.168.157.166:8020</value>

</property>

```
user@ubuntu:~$ sudo gedit hadoop-1.2.0/conf/core-site.xml
[sudo] password for user:
```

core-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl"
href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file.
-->

<configuration>
<property>
    <name>fs.default.name</name>
    <value>hdfs://192.168.157.166:8020</value>
  </property>
</configuration>
```

- **Edit hdfs-site.xml:**
- **Command:** sudo gedit Hadoop-1.2.0/conf/hdfs-site.xml

```
<property>
  <name>dfs.replication</name>
  <value>2</value>
 </property>
 <property>
  <name>dfs.permissions</name>
  <value>false</value>
 </property>
```

```
user@ubuntu:~$ sudo gedit hadoop-1.2.0/conf/hdfs-site.xml
user@ubuntu:~$
```

hdfs-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
     <name>dfs.permissions</name>
     <value>false</value>
  </property>
</configuration>
```

- **Edit mapred-site.xml:**
- **Command: sudo gedit Hadoop-1.2.0/conf/mapred -site.xml**

```
<property>
  <name>mapred.job.tracker</name>
  <value>192.168.157.166:8021</value>
 </property>
```

```
user@ubuntu:~$ sudo gedit hadoop-1.2.0/conf/mapred-site.xml
user@ubuntu:~$
```

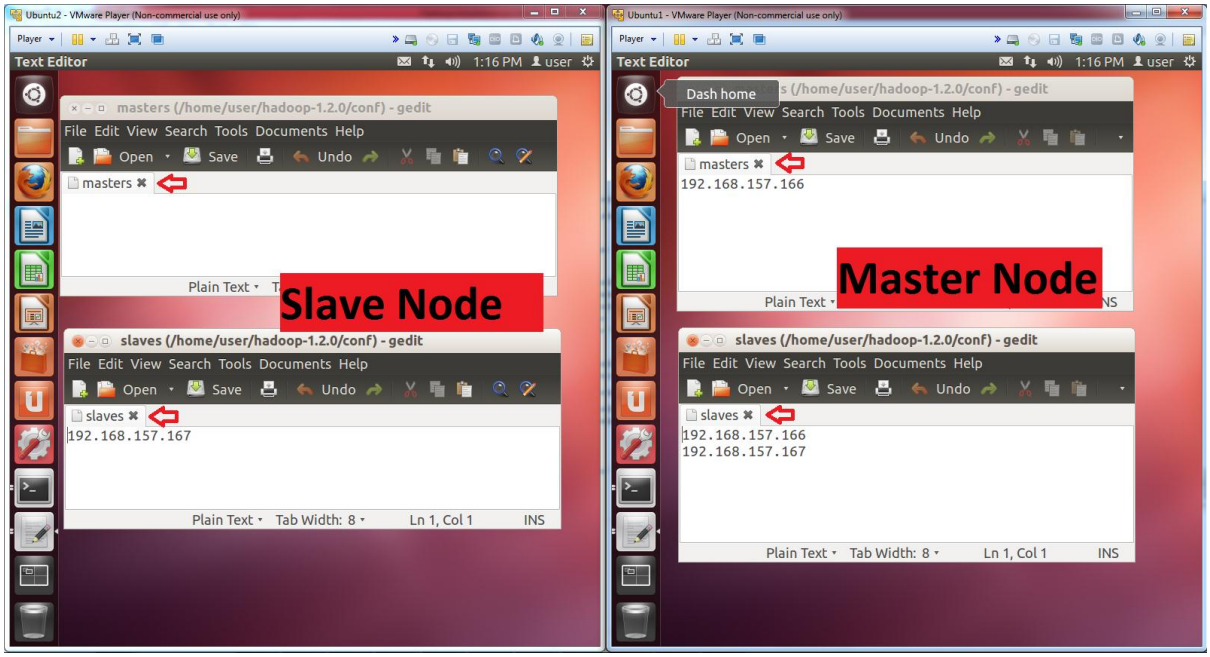
mapred-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl"
href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file.
-->

<configuration>
<property>
    <name>mapred.job.tracker</name>
    <value>192.168.157.166:8021</value>
  </property>
</configuration>
```

Now there is a slight difference in slaves and masters file for both VM.

On master node, masters file contains master node's ip address only and slaves file contains ip addresses of both vms.

On slave node, master file is blank and slaves file contains slave VM's ip address. See the image below.



- **Create a ssh key:**
- **Command:** ssh-keygen -t rsa -P ""

```
user@ubuntu:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/user/.ssh/id_rsa):
Your identification has been saved in /home/user/.ssh/id_rsa.
Your public key has been saved in /home/user/.ssh/id_rsa.pub.
The key fingerprint is:
77:2e:8e:c8:21:69:1e:04:b4:14:b5:12:6a:1a:9a:5f user@ubuntu
The key's randomart image is:
+--[ RSA 2048]----+
|   =o.           |
|  + o .          |
|o.+ .            |
|+o o             |
|+   E   S . .    |
| . o .   . o     |
|  . = .   . .    |
|   o + o o .     |
|    . o . .      |
+-----------------+
user@ubuntu:~$
```

- **Creating a password-less ssh login:**
- **Command:** ssh-copy-id -i $HOME/.ssh/id_rsa.pub user@192.168.157.166
- **Command:** ssh-copy-id -i $HOME/.ssh/id_rsa.pub user@192.168.157.167

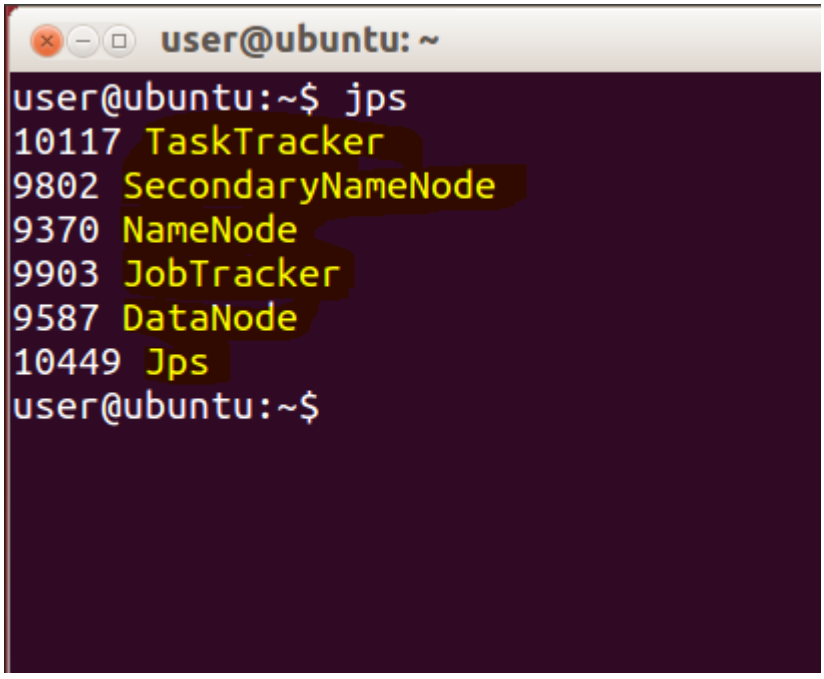
Run the below commands on Master node.

- **Format the name node**
- **Command:** bin/hadoop namenode -format

- **Start the namenode, datanode and job tracker**
- **Command:** bin/start-dfs.sh

- **Start the task tracker**
- **Command:** bin/start-mapred.sh

- **To check if Hadoop started correctly**
- **Command: jps**



Open browser and type http://localhost:50070/dfshealth.jsp to see the current live node.

edureka!

NameNode 'ubuntu:8020'

Started:	Mon Jun 03 10:50:11 EDT 2013
Version:	1.2.0, r1479473
Compiled:	Mon May 6 06:59:37 UTC 2013 by hortonfo
Upgrades:	There are no upgrades in progress.

Browse the filesystem
Namenode Logs

Cluster Summary

6 files and directories, 1 blocks = 7 total. Heap Size is 31.57 MB / 966.69 MB (3%)

Configured Capacity	:	78.56 GB
DFS Used	:	68 KB
Non DFS Used	:	9.9 GB
DFS Remaining	:	68.66 GB
DFS Used%	:	0 %
DFS Remaining%	:	87.39 %
Live Nodes	:	2
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	0

That's it….You did it. You can see both nodes running.