Setting up Hadoop made Easy

Word of Motivation

Hadoop installation is the most complex step when you start out to learn Hadoop, especially when you are new to Linux as well. At some point of time it may test you, please be patient and follow the steps below. Many have installed it following the same steps as below.

Although I have tried to cover installation which should be applicable to all scenarios, but some strange situation specific error can spring up at your end. When Hadoop tests you with a challenge, please try to resolve it through internet.

Just in case, if you fail to get the right advice on internet and are stuck for long (2 days or more), please contact me. I would help you out.

Basic Idea in a Nutshell

Following are the steps that would be taken in a nutshell:

1. Install virtual machine on windows or OS.
2. Install Ubuntu on the virtual machine.
3. Download and untar Hadoop package on Ubuntu.
4. Download and install Java on Ubuntu. (Hadoop is written completely in Java).
5. Tell Ubuntu where the Java installation has been done.
6. Tell Hadoop where Java installation has been done. At this point Standalone is done.
7. For pseudo-distribution mode, change the configuration files to configure:
    a. Core-site.xml -> to set default Schema and authority.
    b. Hdfs-site.xml -> to set def.replication to 1 rather than the default three, otherwise all the blocks would always be alarmed with under replication.
    c. Mapred-site.xml -> To let know of host and port pair where the Jobtrackers runs at.
8. Format the name node and you are ready.

Version details

Following are the details of components used, all license free:

1. Hadoop 1.2.1
2. Ubuntu LTS 12.04 (running on virtual Machine) 64 Bit
3. Windows 8. (The same thing can be done on mac, i.e., install a virtual machine on mac and follow the below procedure). Any windows machine would do well.

Step 1. Installing Virtual Machine

Step 1.1 Download

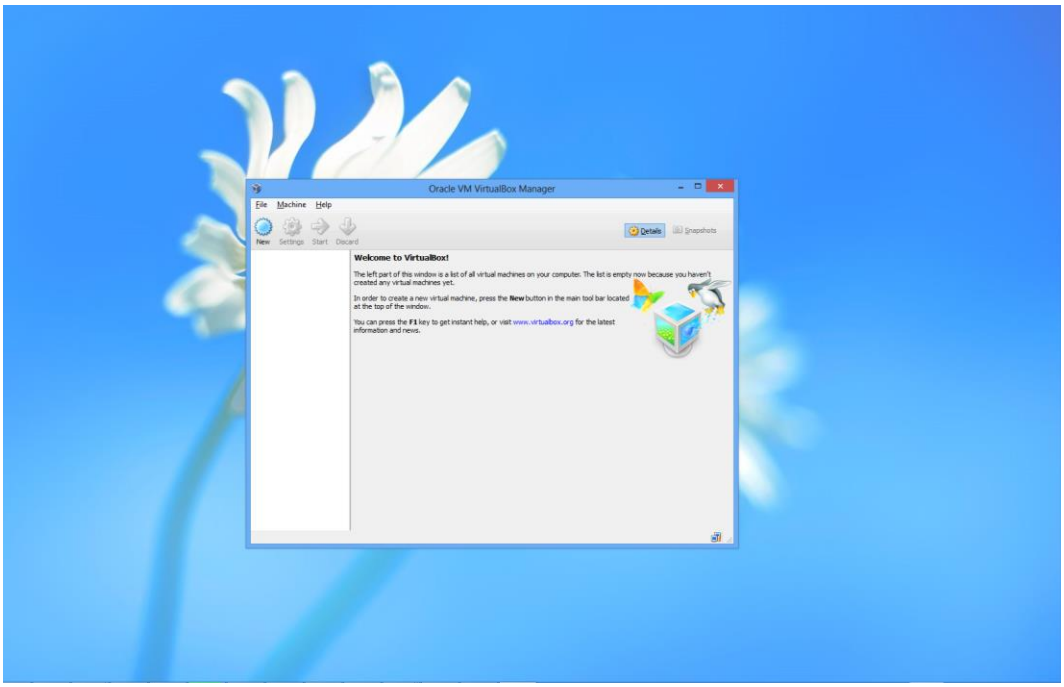
Free version of Oracle VirtualBox can be downloaded from:

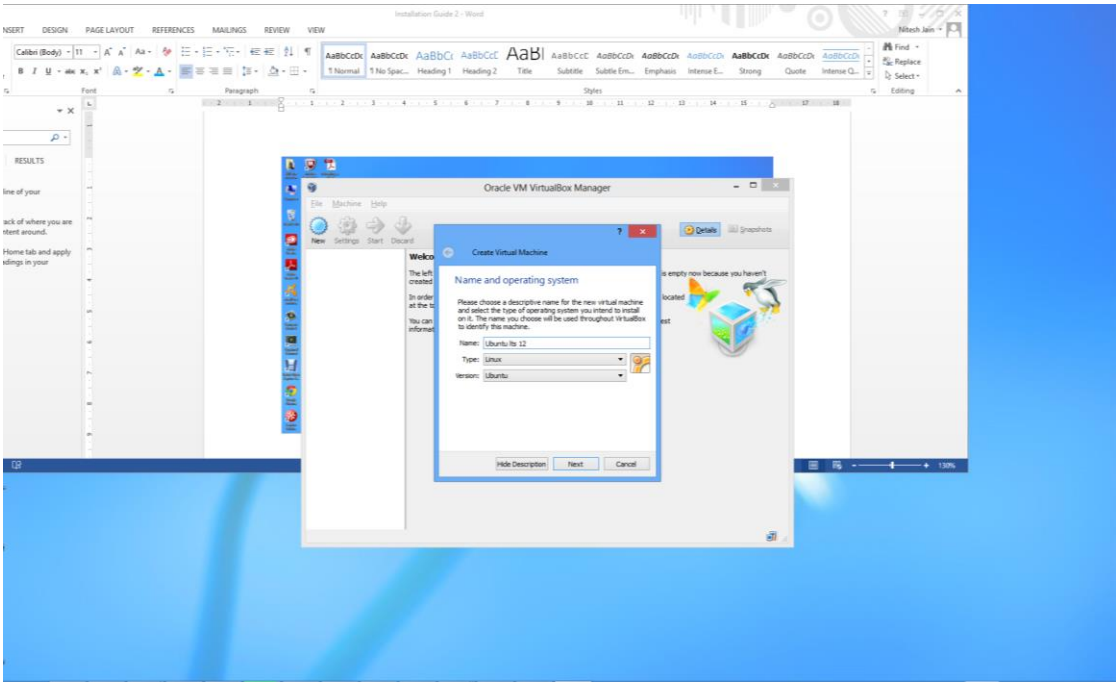
https://www.virtualbox.org/wiki/Downloads

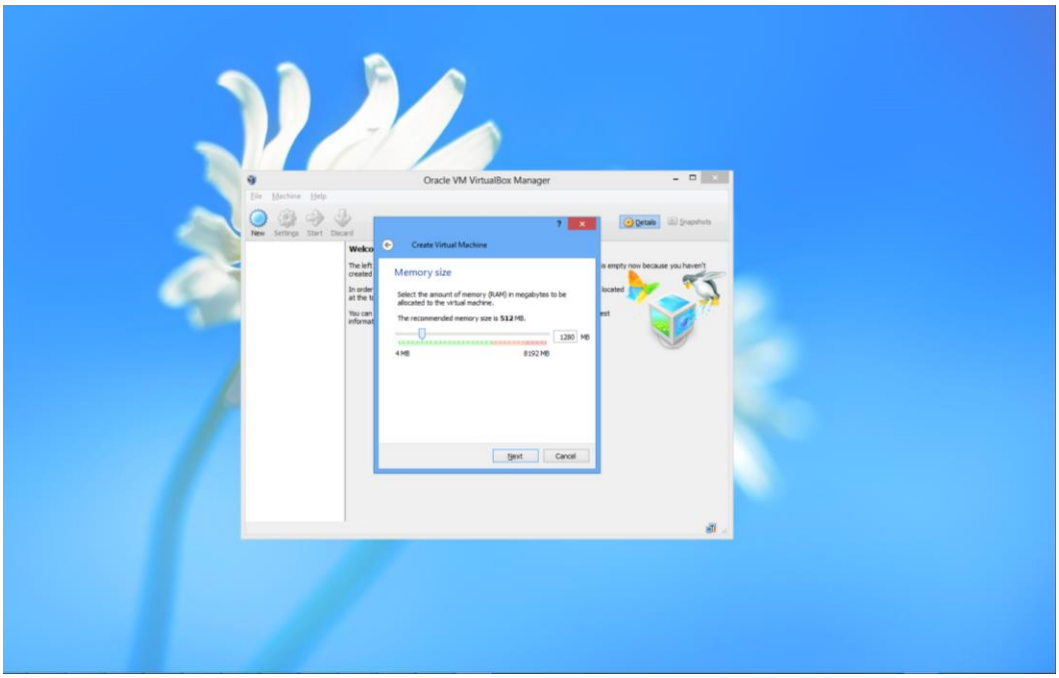
Download UBUNTU LTS 64 bit from the following link (Make sure its ISO format and for 64 bit):

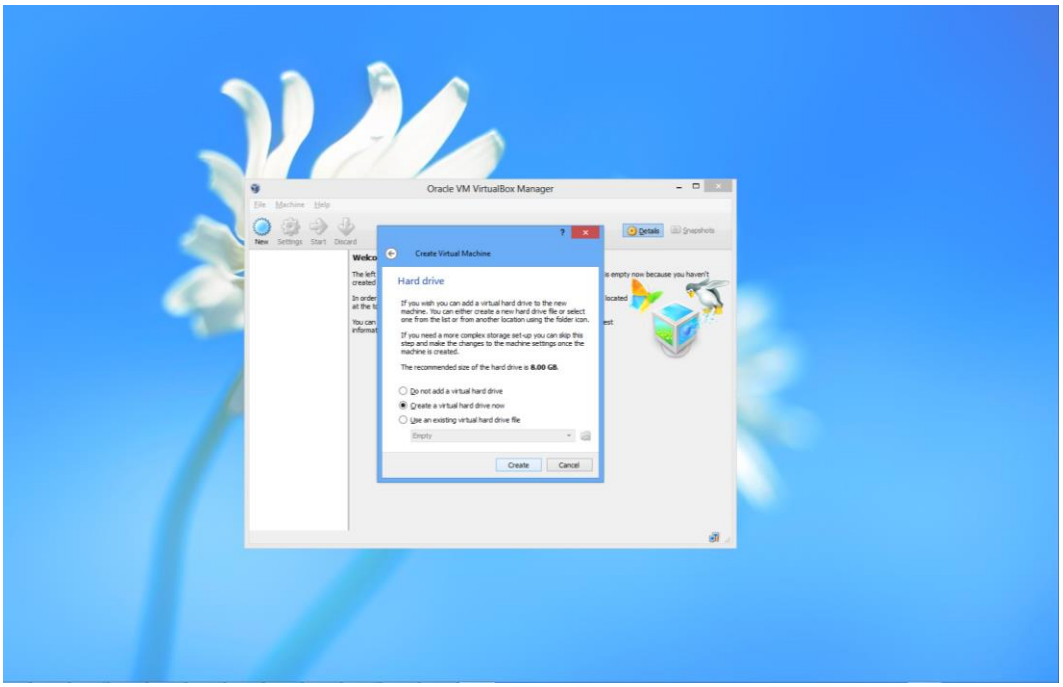
http://www.ubuntu.com/download/desktop

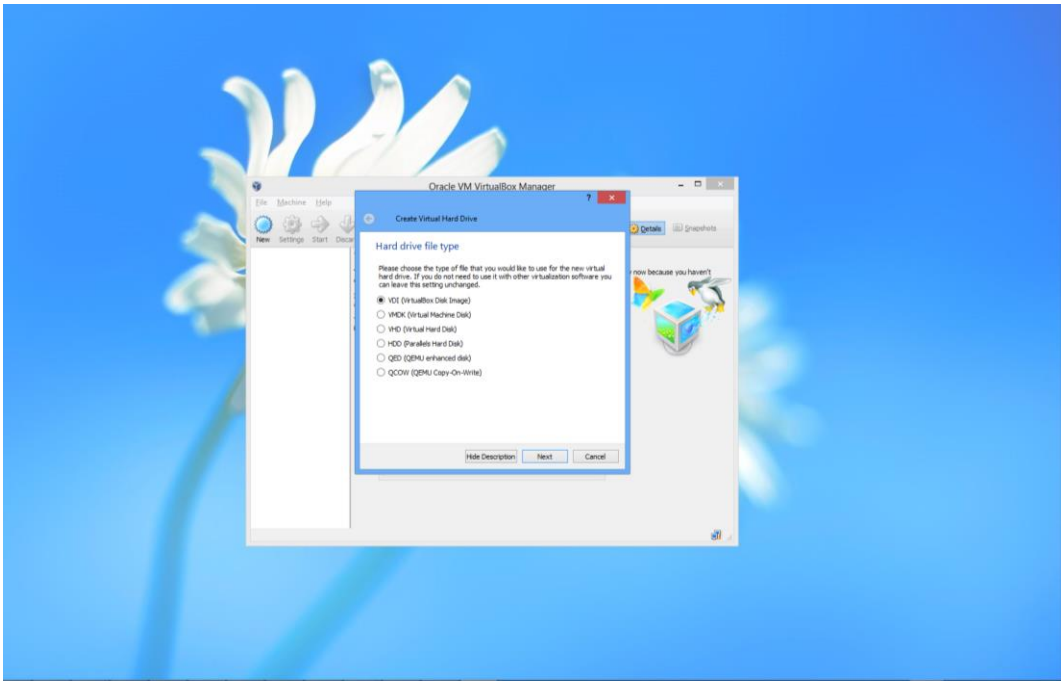
Step 1.2 Installation

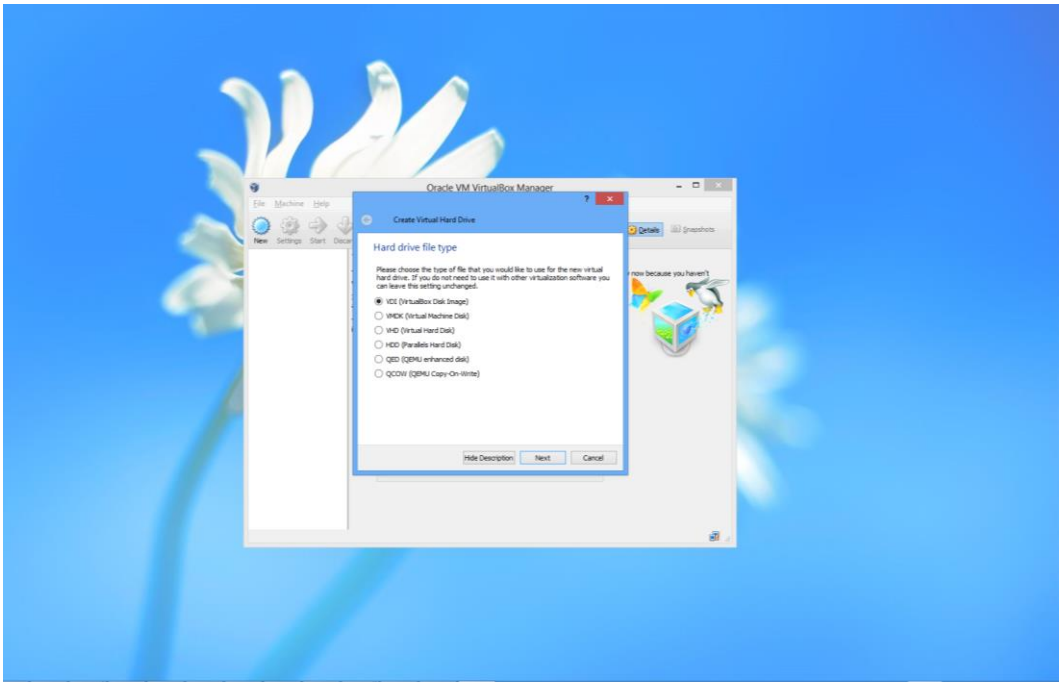


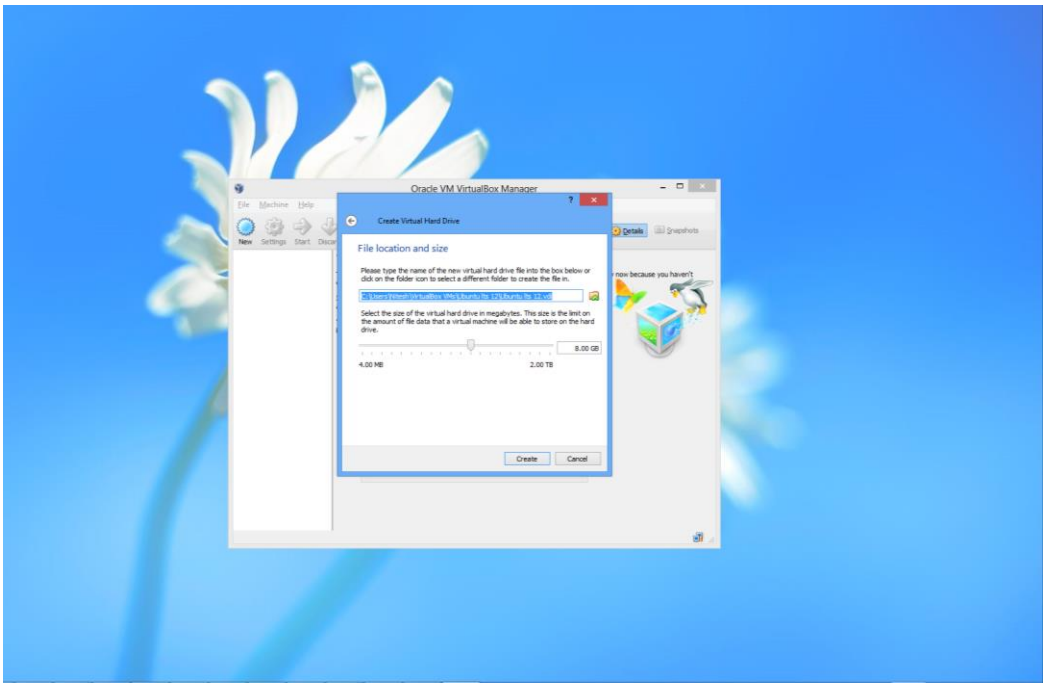

Oracle VM VirtualBox Manager
Create Virtual Machine
Name and operating system
Name: Ubuntu lts 12
Type: Linux
Version: Ubuntu
Hide Description
Next
Cancel

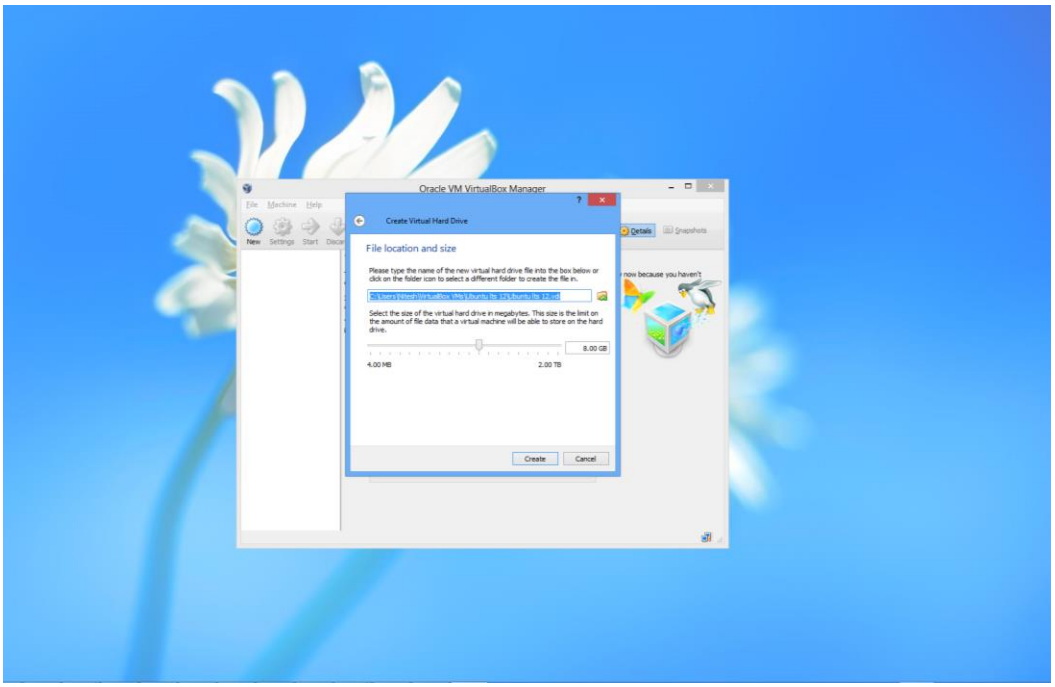

Oracle VM VirtualBox Manager
Create Virtual Machine
Memory size
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.
The recommended memory size is 512 MB.
1280 MB
4 MB
8192 MB
Next
Cancel


Oracle VM VirtualBox Manager
Create Virtual Machine
Hard drive
If you wish you can add a virtual hard drive to the new machine. You can either create a new hard drive file or select one from the list or from another location using the folder icon.
If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.
The recommended size of the hard drive is 8.00 GB.
Do not add a virtual hard drive
Create a virtual hard drive now
Use an existing virtual hard drive file
Empty
Create
Cancel

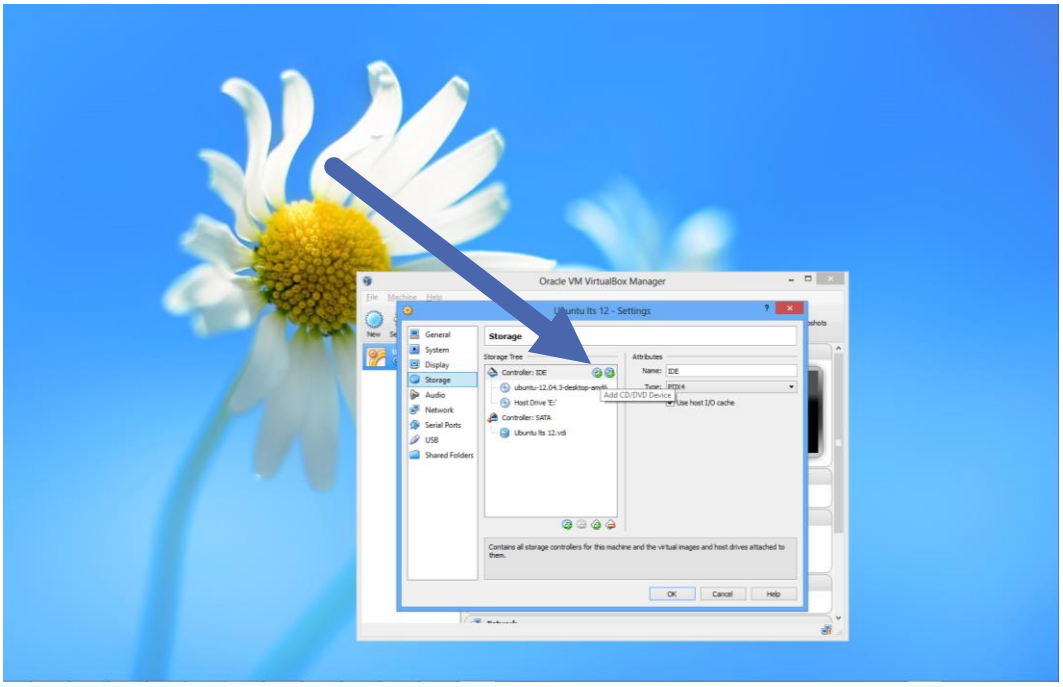

Oracle VM VirtualBox Manager
Create Virtual Hard Drive
Hard drive file type
Please choose the type of file that you would like to use for the new virtual hard drive. If you do not need to use it with other virtualization software you can leave this setting unchanged.
VDI (VirtualBox Disk Image)
VMDK (Virtual Machine Disk)
VHD (Virtual Hard Disk)
HDD (Parallels Hard Disk)
QED (QEMU enhanced disk)
QCOW (QEMU Copy-On-Write)
Hide Description
Next
Cancel


Oracle VM VirtualBox Manager
Create Virtual Hard Drive
Hard drive file type
Please choose the type of file that you would like to use for the new virtual hard drive. If you do not need to use it with other virtualization software you can leave this setting unchanged.
VDI (VirtualBox Disk Image)
VMDK (Virtual Machine Disk)
VHD (Virtual Hard Disk)
HDD (Parallels Hard Disk)
QED (QEMU enhanced disk)
QCOW (QEMU Copy-On-Write)
Hide Description
Next
Cancel

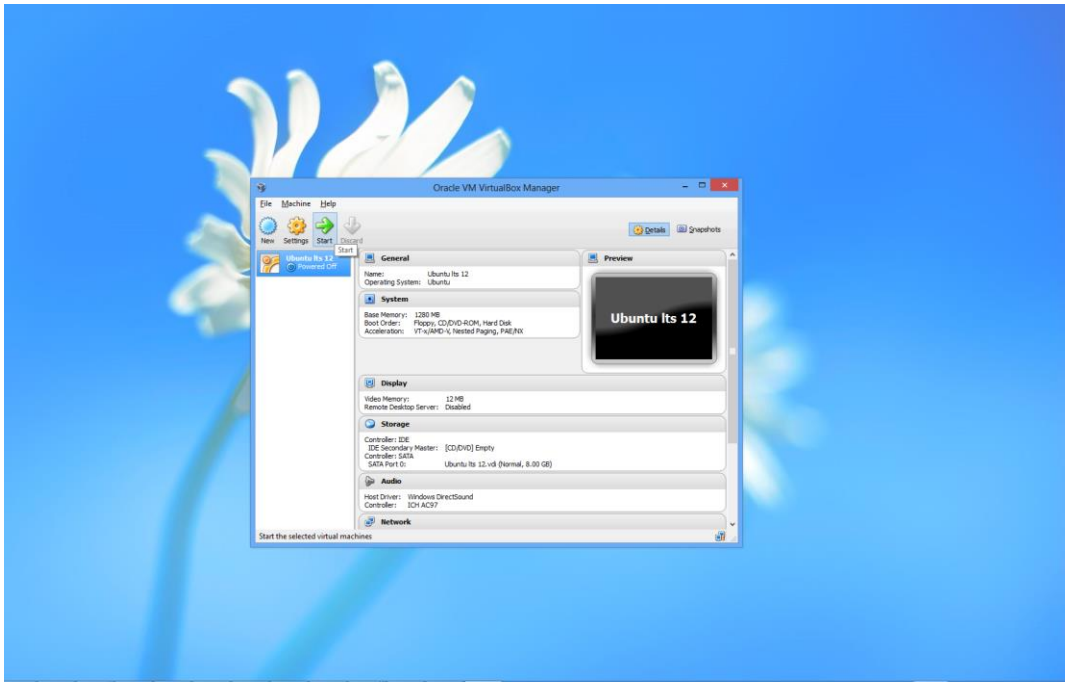

Oracle VM VirtualBox Manager
Create Virtual Hard Drive
File location and size
Please type the name of the new virtual hard drive file into the box below or click on the folder icon to select a different folder to create the file in.
Select the size of the virtual hard drive in megabytes. This size is the limit on the amount of file data that a virtual machine will be able to store on the hard drive.
8.00 GB
4.00 MB
2.00 TB
Create
Cancel



In the below screen shot click on the '+' sign to add ISO which you have already downloaded to be loaded as CD drive.

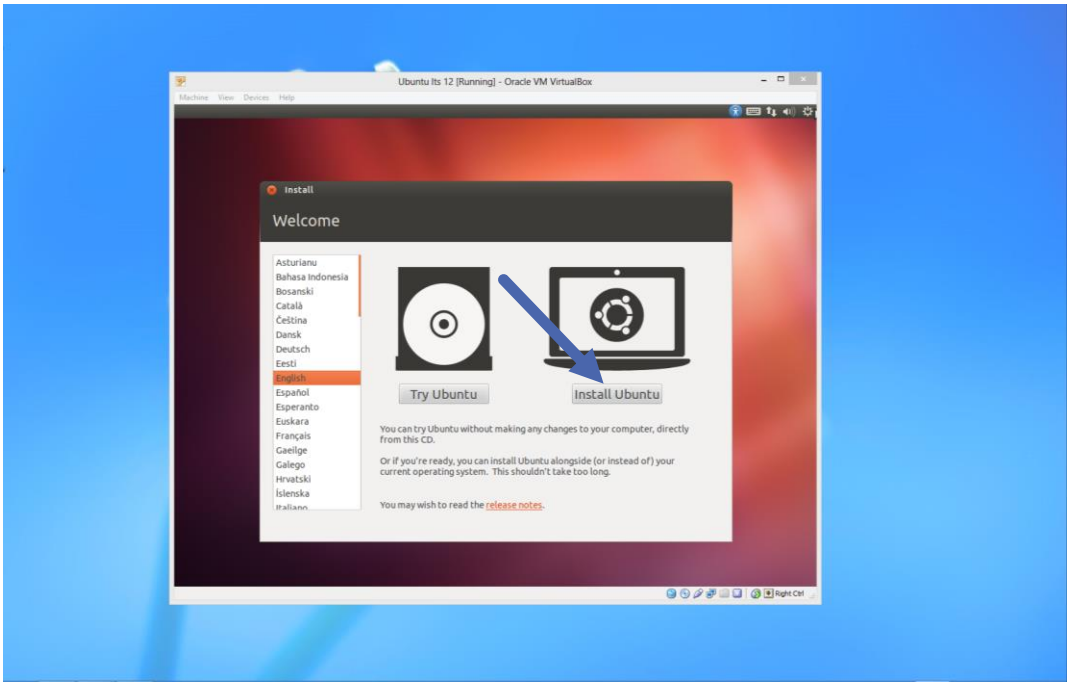


Press Start.

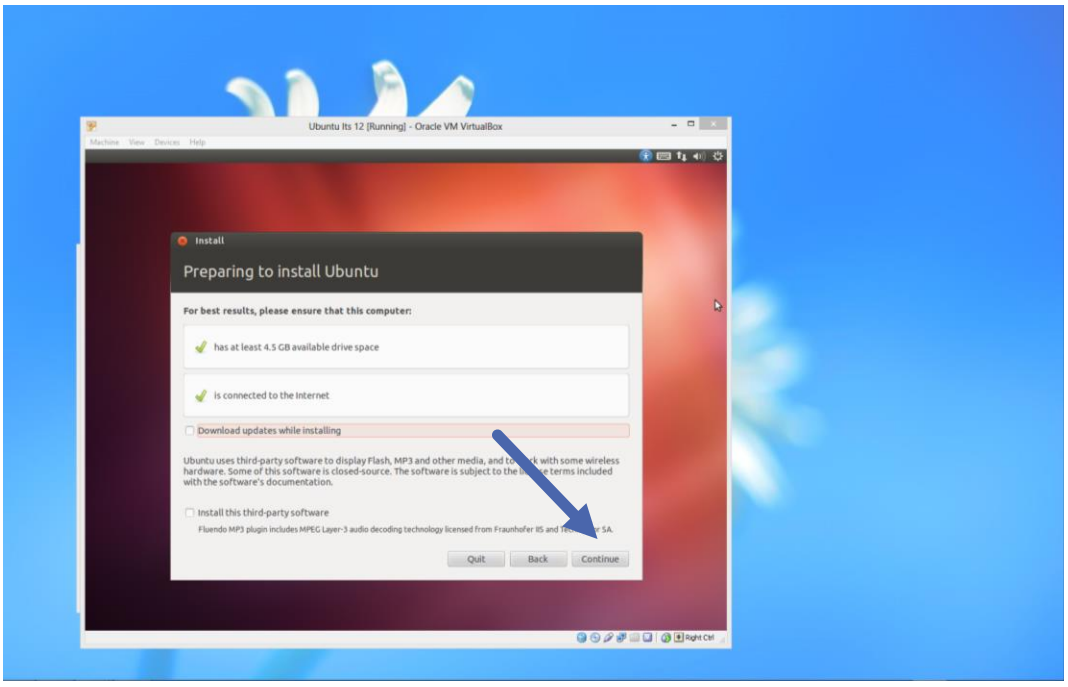


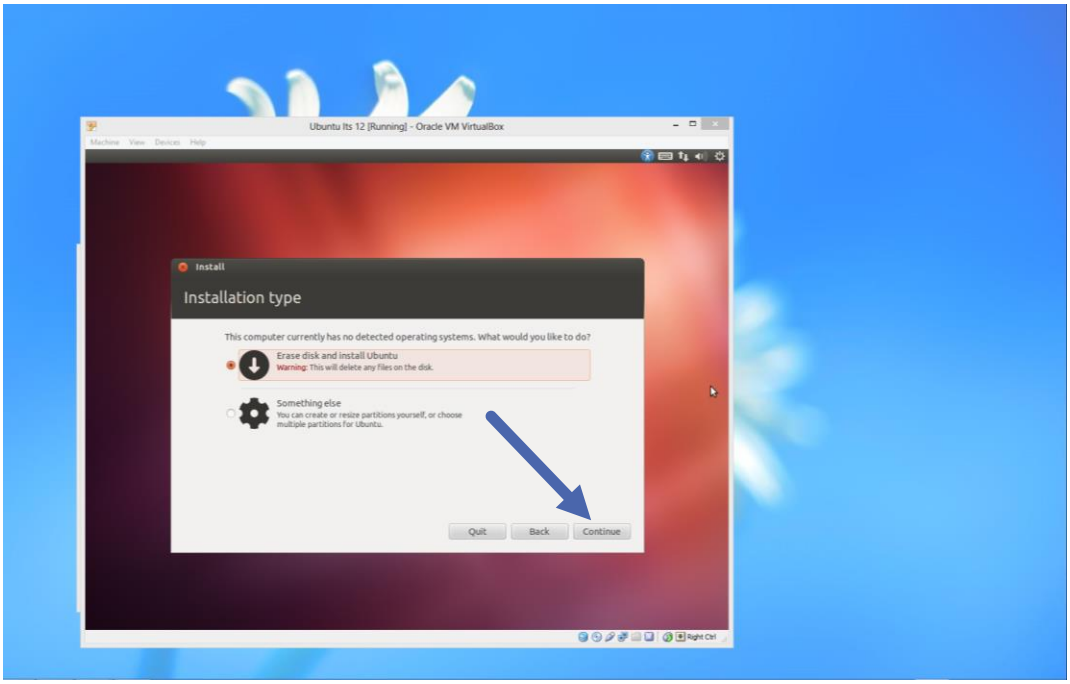
If throws an error, saying something about that 64 bit support and about VT-x/AMD-V,

- It means that your BIOS doesn't support visualization.
- Perform the following steps. This is for my configuration yours may be a little different:
  - Restart you computer and go to BIOS setup
  - Goto UEFI Firmware>>Advaced>>CPU Setup >> Intel ® Virtualization Techonlogy. Enable this.
  - Save and exit.
- Now try to start the Ubuntu boot with the ISO image and it should work.



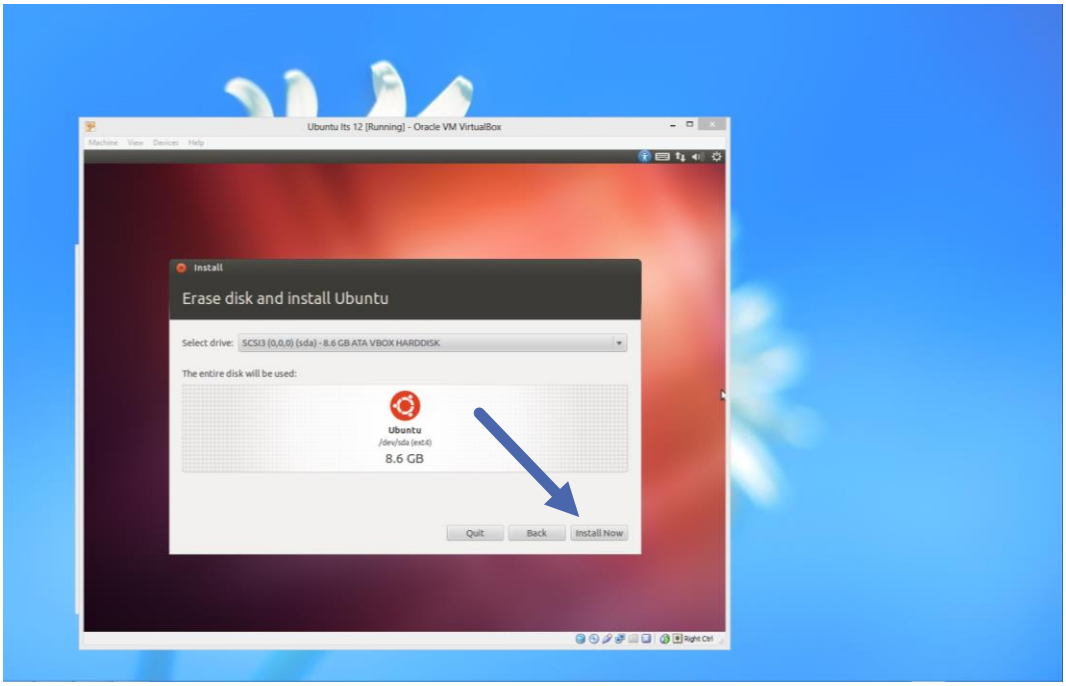
Click on install Ubuntu.

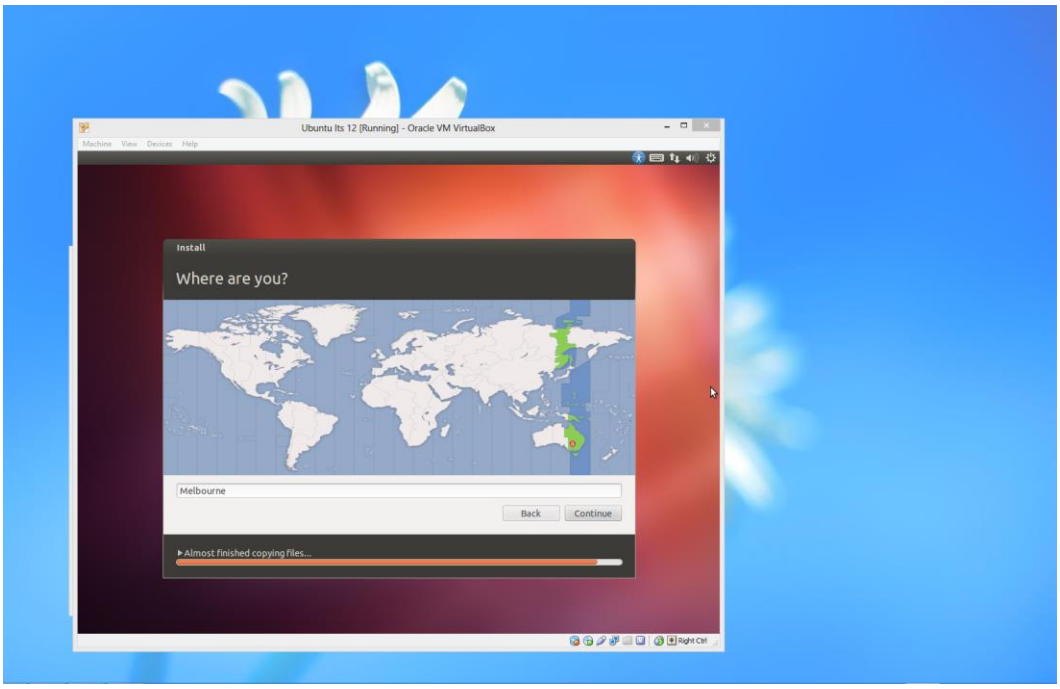




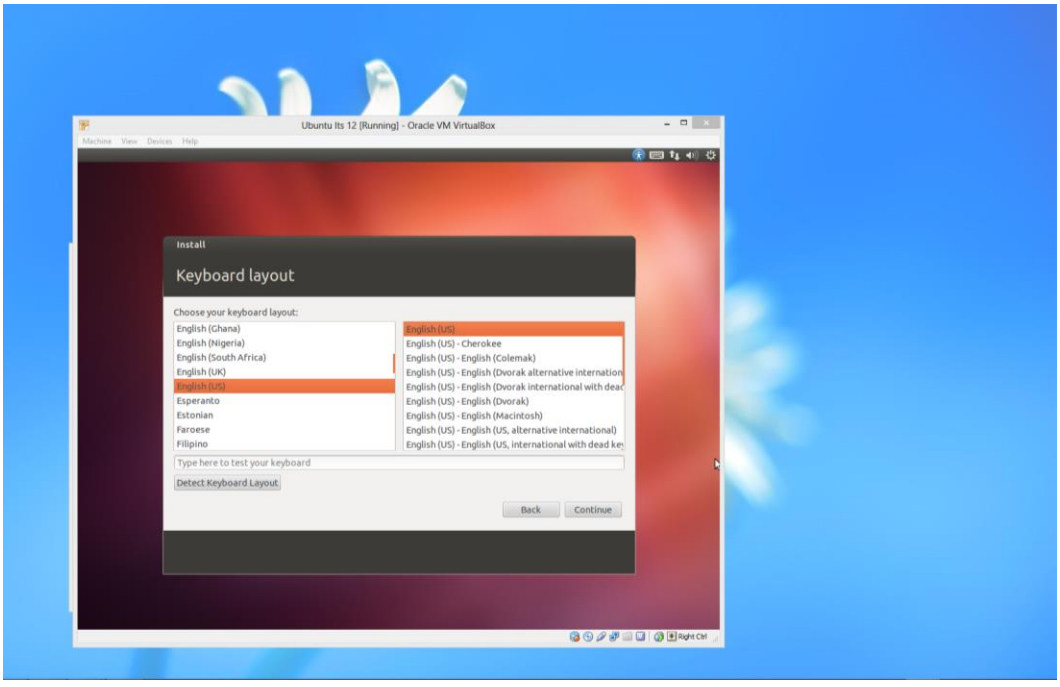
And after you have pressed continue the whole disk would be formatted!

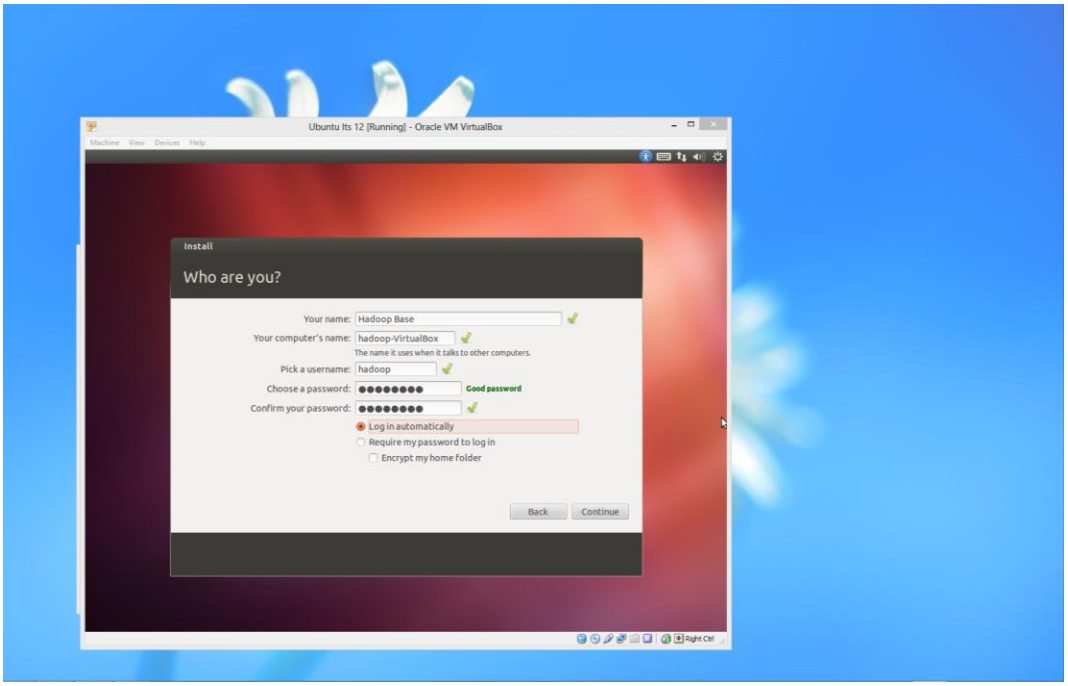
Nope just joking! (: Only the dynamic Disk allocated would be formatted.





(I live in Melbourne. One of the loveliest cities in the world.)





Step 2 Download Hadoop tar.gz

At this point you would like to reopen this document on Ubuntu. You can transfer it by internet.

(Most of the following steps are referred from apache docs:

http://hadoop.apache.org/docs/stable/single_node_setup.html

The link above might get obsolete with time. If so, please google search Apache hadoop installation to find apache installation guide)

The main idea behind the following steps is to create a folder for hadoop and untar (or unzip) the tar file that has been downloaded.

1. Downloading a stable release copy ending with tar.gz
2. Create a new folder /home/hadoop
3. Move the file hadoop.x.y.z.tar.gz to the folder /home/Hadoop
4. *Type/Copy/Paste:* cd /home/hadoop
5. *Type/Copy/Paste*: tar xzf hadoop*tar.gz

Step 3 Downloading and setting up Java

For more refer: http://www.wikihow.com/Install-Oracle-Java-on-Ubuntu-Linux

(The link above might get obsolete with time. If so, please google search wikihow install java Ubuntu Linux)

1. Check if Java is already present, by

   Type/Copy/Paste : java –version

2. If it is 1.7.* then you can setup the JAVA_HOME Variable according to where it is setup.
3. If you are confident to setup the JAVA_HOME variable please go ahead to step 9 (in this section itself). If not don't worry and follow the following steps:
4. First we will purge the Java installed.

   Type/Copy/Paste : sudo apt-get purge openjdk-*

5. Make the directory where java would installed, by:

   sudo mkdir -p /usr/local/java

6. Download Java JDK and JRE from the link, look for linux, 64 bit and tar.gz ending file:
   http://www.oracle.com/technetwork/java/javase/downloads/index.html

7. Goto downloads folder and then copy to the folder we created for java:
   Type/Copy/Paste: sudo cp -r jdk-*.tar.gz /usr/local/java
   Type/Copy/Paste: sudo cp -r jre-*.tar.gz /usr/local/java

8. Extract and install Java:
   Type/Copy/Paste: cd /usr/local/java
   Type/Copy/Paste: sudo tar xvzf jdk*.tar.gz
   Type/Copy/Paste: sudo tar xvzf jre*.tar.gz

9. Now put all the variables in the profile.

   Type/Copy/Paste: sudo gedit /etc/profile

   At the end copy paste the following.(Note: change the highlighted paths according to your installations. Version number would have changed from making this guide to your installation. So just make sure that the path you mention actually exists)
   [Tip: It is important that you specify complete and correct path while declaring HADOOP_INSTALL variable. To get the right value, navigate to the folder hadoop-1.2.1 and then type in command 'pwd' which would return the complete present working directory. Copy and paste that to avoid any typo errors.]

   ```
   JAVA_HOME=/usr/local/java/jdk1.7.0_40
   PATH=$PATH:$JAVA_HOME/bin
   JRE_HOME=/usr/local/java/jre1.7.0_40
   PATH=$PATH:$JRE_HOME/bin
   HADOOP_INSTALL=/home/{user_name}/hadoop/hadoop-1.2.1
   PATH=$PATH:$HADOOP_INSTALL/bin
   export JAVA_HOME
   export JRE_HOME
   export PATH
   ```

10. Do the following so that Linux knows where Java is, (Note that the highlighted following paths may be needed to changed in accordance to your installation):

    ```
    sudo update-alternatives --install "/usr/bin/java" "java" "/usr/local/java/jre1.7.0_40/bin/java" 1
    ```

```
sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/local/java/jdk1.7.0_40/bin/javac" 1
sudo update-alternatives --install "/usr/bin/javaws" "javaws" "/usr/local/java/jre1.7.0_40/bin/javaws" 1
sudo update-alternatives --set java /usr/local/java/jre1.7.0_40/bin/java
sudo update-alternatives --set javac /usr/local/java/jdk1.7.0_40/bin/javac
sudo update-alternatives --set javaws /usr/local/java/jre1.7.0_40/bin/javaws
```

11. Refresh the profile by:

    Type/Copy/Paste: `. /etc/profile`

12. Test by typing Java –version.

Step 4 Stand Alone mode installed! Congratulations!

At this point you should have had got to the point that you can run Hadoop in Stand Alone mode. You can practice almost anything for practicing developments in Map Reduce. Test if you are successful:

Type/Copy/Paste: cd /home/hadoop (going to the Hadoop directory)
Type/copy/Paste: mkdir input
Type/copy/Paste: bin/hadoop jar hadoop-examples-*.jar grep input output 'dfs[a-z.]+'
Or the above can be typed in without 'bin' as well.
Type/copy/Paste: hadoop jar hadoop-examples-*.jar grep input output 'dfs[a-z.]+'
Type/copy/Paste: ls output/*

Step 5 Pseudo Distribution Mode

1. Type/Copy/Paste: sudo apt-get install ssh (to install ssh)
2. Type/Copy/Paste: sudo apt-get install rsync
3. Change conf/core-site.xml to:

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

4. Change conf/hdfs-site.xml to:

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

5. Change conf/mapred-site.xml to:

```
<configuration>
    <property>
       <name>mapred.job.tracker</name>
       <value>localhost:9001</value>
    </property>
</configuration>
```

6. Edit conf/hadoop-env.sh look for JAVA_HOME and set it up
   export JAVA_HOME=/usr/local/java/jdk1.7.0_40

7. Setup passwordless ssh by:
   Type/copy/paste: ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
   Type/copy/paste: cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

8. To confirm that passwordless ssh has been setup type the following and you shouod not be prompted for a password.
   Type/copy/paste: ssh localhost

9. Format the name node:
   Type/copy/paste: bin/hadoop namenode –format

10. Start the all the demons:
    Type/copy/paste: bin/start-all.sh

11. On web browser navigate to http//localhost:50070/ and then to http://localhost:50030/

Make sure hadoop started properly.

http://localhost:50030/ should forward to http://localhost:50030/jobtracker.jsp localhost Hadoop Map/Reduce Administration page

http://localhost:50070/ should forward to http://localhost:50070/dfshealth.jsp NameNode 'localhost:9000' page

If any of url doesn't work than make sure that namenode and datanode started successfully by running the command 'jps' (show java processes) and the output should look like the following:

```
2310 SecondaryNameNode
1833 NameNode
2068 DataNode
2397 JobTracker
2635 TaskTracker
2723 Jps
```

If NameNode or DataNode is not listed than it might happen that the namenode's or datanode's root directory which is set by the property 'dfs.name.dir' is getting messed up. It by default points to the /tmp directory which operating system changes from time to time. Thus, HDFS when comes up after some changes by OS, gets confused and namenode doesn't start.

Solution:

a) Stop hadoop by running 'stop-all.sh'

   We need to explicitly set the 'dfs.name.dir' and 'dfs.data.dir'.

Perform the following steps and the issue should resolve (You can of course create any folders and give that path, but below I would be giving an example. You can create your own folder your way)

b) Goto hadoop folder and create a folder 'dfs'. So now the folder '/home/hadoop/dfs' would exist. The idea is to make two folders inside it which would be used for datanode demon and namenode demon.

Create 'data' and 'name' folders inside '/home/{user_name}/hadoop/dfs' folder)

c) Change the configuration file hdfs-site.xml to set properties 'dfs.name.dir' and 'dfs.data.dir' as follows. Two points to be noted. First, change the indentation. Second, change the username portion (/{user_name} in the case below, it should be your's) of path according to your system. Giving incomplete path is a common error:

(TIP: go to newly created dfs folder thorough command prompt and type in command 'pwd' to get exact path. Copy paste to avoid typos)

configuration file hdfs-site.xml should look like below:

```
<configuration>
        <property>
                <name>dfs.data.dir</name>
                <value>/home/{user_name}/hadoop/dfs/data/</value>
        </property>
        <property>
                <name>dfs.name.dir</name>
                <value>/home/{user_name}/hadoop/dfs/name/</value>
        </property>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
</configuration>
```

d) Run command

hadoop namenode -format

Look for the following output to confirm that the format has been successful. If you do not see the message, format command is having some problems.
(I am pasting the output of one of the course taker Vadim and so you see the username as Vadim here)

```
14/02/04 22:56:12 INFO namenode.NameNode: STARTUP_MSG:

/************************************************************

STARTUP_MSG: Starting NameNode

STARTUP_MSG: host = vadim-VirtualBox/127.0.1.1

STARTUP_MSG: args = [-format]
```

STARTUP_MSG: version = 1.2.1

STARTUP_MSG: build =https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.2 -r 1503152; compiled by 'mattf' on Mon Jul 22 15:23:09 PDT 2013

STARTUP_MSG: java = 1.7.0_51

**/

Re-format filesystem in /home/vadim/hadoop/dfs/name ? (Y or N) Y

14/02/04 22:56:17 INFO util.GSet: Computing capacity for map BlocksMap

14/02/04 22:56:17 INFO util.GSet: VM type = 64-bit

14/02/04 22:56:17 INFO util.GSet: 2.0% max memory = 1013645312

14/02/04 22:56:17 INFO util.GSet: capacity = 2^21 = 2097152 entries

14/02/04 22:56:17 INFO util.GSet: recommended=2097152, actual=2097152

14/02/04 22:56:18 INFO namenode.FSNamesystem: fsOwner=vadim

14/02/04 22:56:18 INFO namenode.FSNamesystem: supergroup=supergroup

14/02/04 22:56:18 INFO namenode.FSNamesystem: isPermissionEnabled=true

14/02/04 22:56:18 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100

14/02/04 22:56:18 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessKeyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)

14/02/04 22:56:18 INFO namenode.FSEditLog: dfs.namenode.edits.toleration.length = 0

14/02/04 22:56:18 INFO namenode.NameNode: Caching file names occuring more than 10 times

14/02/04 22:56:19 INFO common.Storage: Image file /home/vadim/hadoop/dfs/name/current/fsimage of size 111 bytes saved in 0 seconds.

14/02/04 22:56:19 INFO namenode.FSEditLog: closing edit log: position=4, editlog=/home/vadim/hadoop/dfs/name/current/edits

14/02/04 22:56:19 INFO namenode.FSEditLog: close success: truncate to 4, editlog=/home/vadim/hadoop/dfs/name/current/edits

14/02/04 22:56:19 INFO common.Storage: Storage directory /home/vadim/hadoop/dfs/name has been successfully formatted.

14/02/04 22:56:19 INFO namenode.NameNode: SHUTDOWN_MSG:

/**

SHUTDOWN_MSG: Shutting down NameNode at vadim-VirtualBox/127.0.1.1

**/

e) Run command

start-all.sh

f) Run command

jps

and this will now show all the demons running like the below:

2310 SecondaryNameNode

1833 NameNode

2068 DataNode

2397 JobTracker

2635 TaskTracker

2723 Jps

g) Run command
stop-all.sh

and you should see the output as:

```
topping jobtracker

localhost: stopping tasktracker

stopping namenode

localhost: stopping datanode

localhost: stopping secondarynamenode
```

*there should be no message as "no namenode to stop" or "no datanode to stop"

Once this is done, you have successfully installed Hadoop in pseudo-distribution mode and this is one of the most difficult things to do.

If you face any issues, please feel free to post the problem on the question forum so that everyone can look at it and help.

As well please answer to others problems so that you sharpen your knowledge and others get the needed help as well.

Best,
Nitesh