



HADOOP INSTALLATION GUIDE-step by step

Installation Instructions

This document will help to install Ubuntu and hadoop 2.x

Sangam Biradar

smbiradar14@gmail.com

sangambiradar@codexplus.in

codexplus.in



Ubuntu is a Linux operating system that has a reputation for being basic and easy to use. Ubuntu was also the world's most popular Linux operating system for the year 2012. Ubuntu is based on Debian Linux operating system and has many derivatives that come from it. Examples include: Kubuntu, Edubuntu, Xubuntu, and many others.

This installation guide shows you how to install Ubuntu 12.04 /13.10/14 along side Windows 7. This means that when you boot your computer you will be choose between booting into Windows 7 or Ubuntu version 12.04/13/14

STEP 1 :

The first step you need to do to install Ubuntu Linux is to head on over to the Ubuntu website and download the ISO image for the installation. The link to the download page for the Ubuntu Desktop edition is: <http://www.ubuntu.com/download/desktop>

Download Link for 12.04 : <http://releases.ubuntu.com/precise/> and choose Ubuntu-12.04.2 64bit

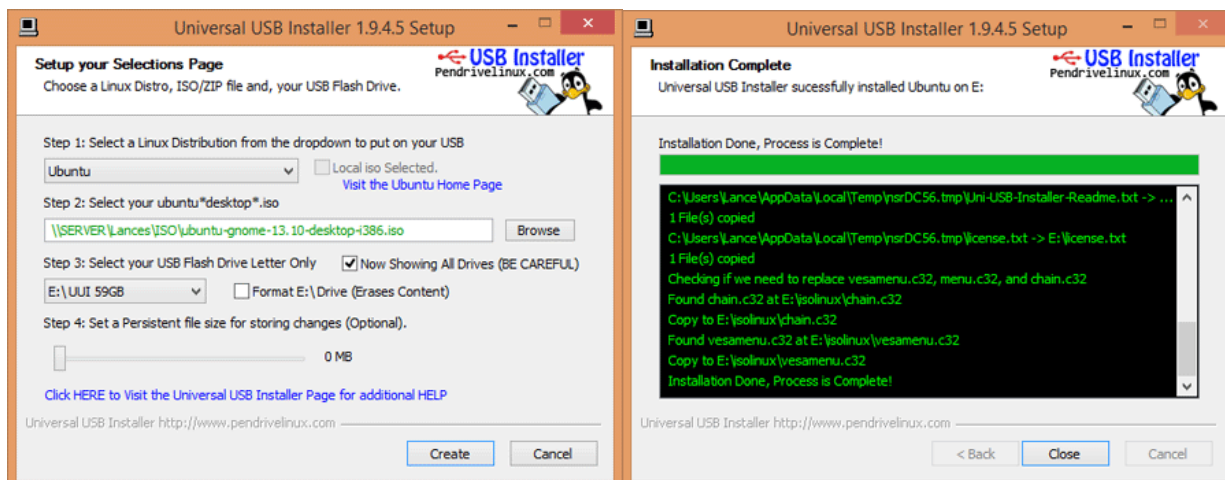
Direct Download Link: <http://releases.ubuntu.com/precise/ubuntu-12.04.4-alternate-amd64.iso>

STEP 2:

URL : <http://www.pendrivelinux.com/universal-usb-installer-easy-as-1-2-3/>

Download: <http://www.pendrivelinux.com/downloads/Universal-USB-Installer/Universal-USB-Installer-1.9.5.5.exe>

Universal USB Installer (UII) Screenshots

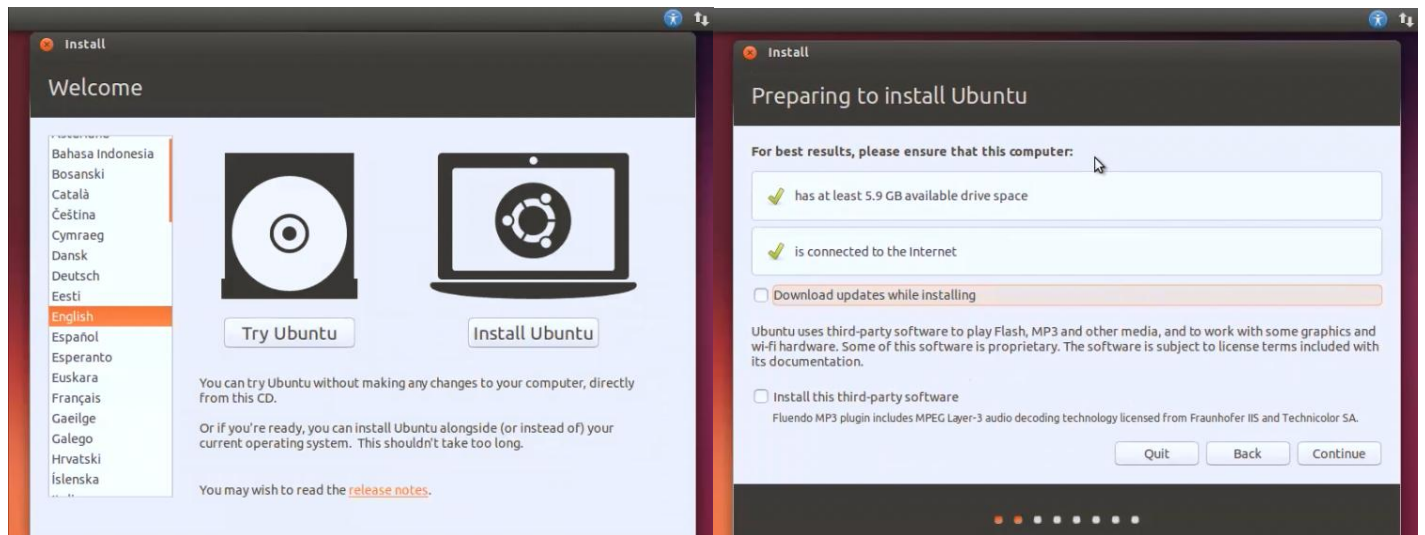


Once you have downloaded the ISO image for the installation then you need to burn the ISO image to a blank DVD/USB. If you do not know how to do this then do a Google search for “burning ISO image” using your operating system.

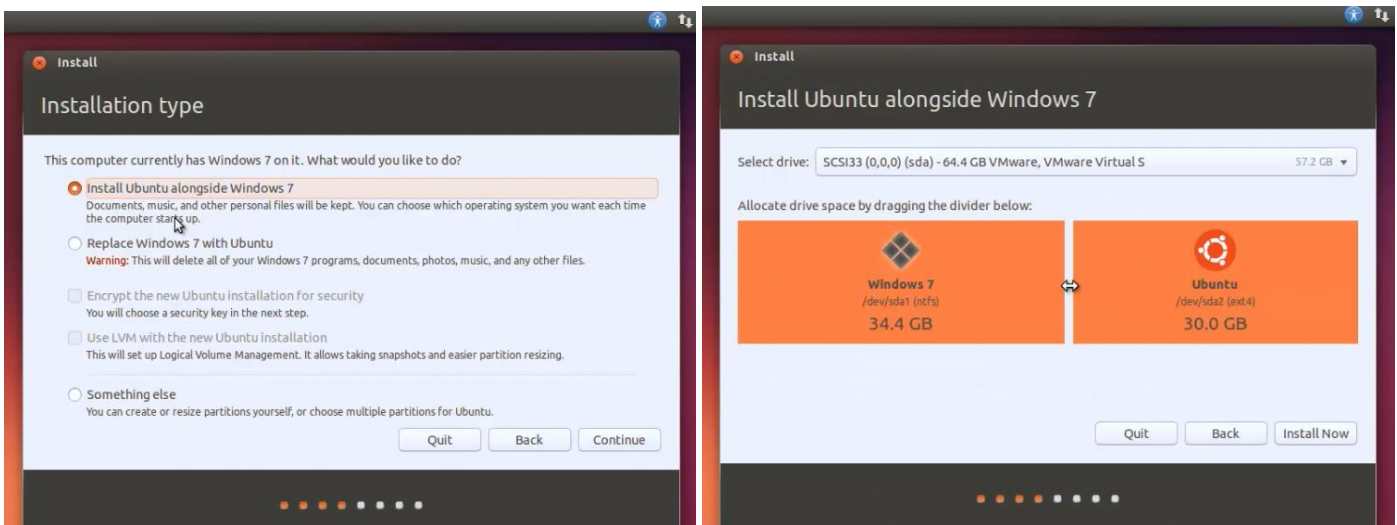
Once you have burned the ISO image to a blank DVD then it is time to start the installation. Put the disk into the disk drive if it isn't already and reboot your computer. When your computer reboots be sure to boot off of the Disk and not

back into your current operating system. If your computer does not boot off of the disk then you may have to change your bios settings to boot off of a CD/DVD like the one below.

Once you boot off of the Ubuntu Disk then your screen should look like the image below.

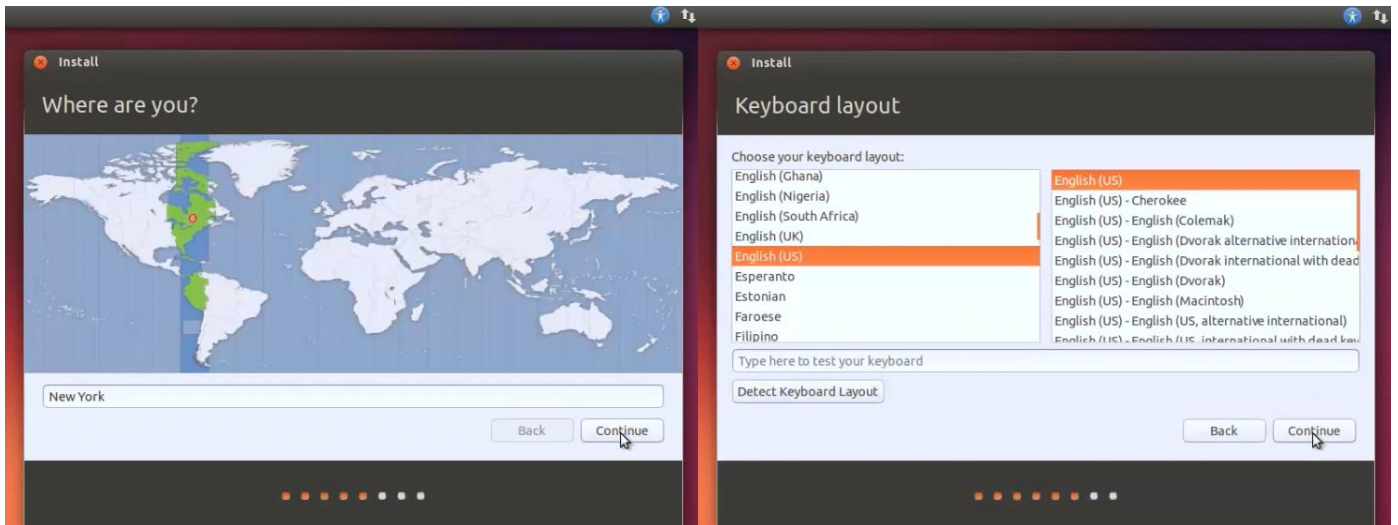


On the next screen you will be selecting the “Install Ubuntu alongside Windows 7” option. As noted above this will make your computer a dual-boot meaning you will have the option to choose between Windows 7 or Ubuntu 12.04 when you boot. Click Continue.



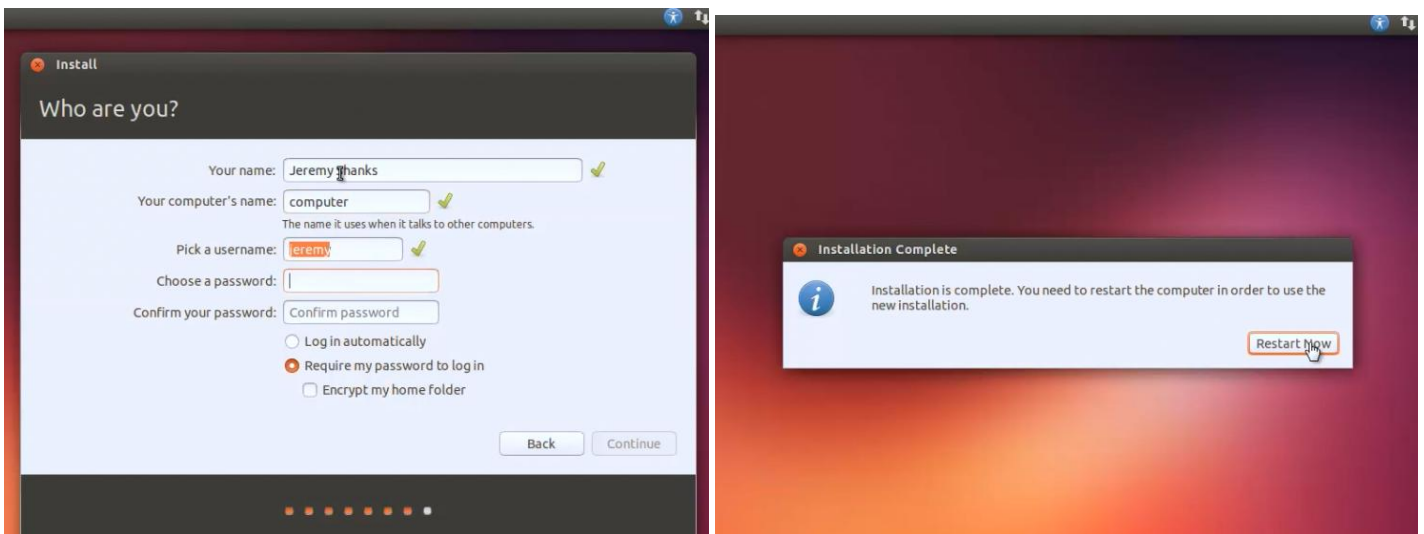
You will then be given the option to choose how much disk space you want each operating system to have. This option is completely up to you. You can drag the bar both ways to change the disk space for each. In this tutorial I will be allowing Ubuntu to have 30 GB of hard disk space.

On the next screen you will be selecting your time zone. New York happens to be in my time zone and therefore I selected New York.

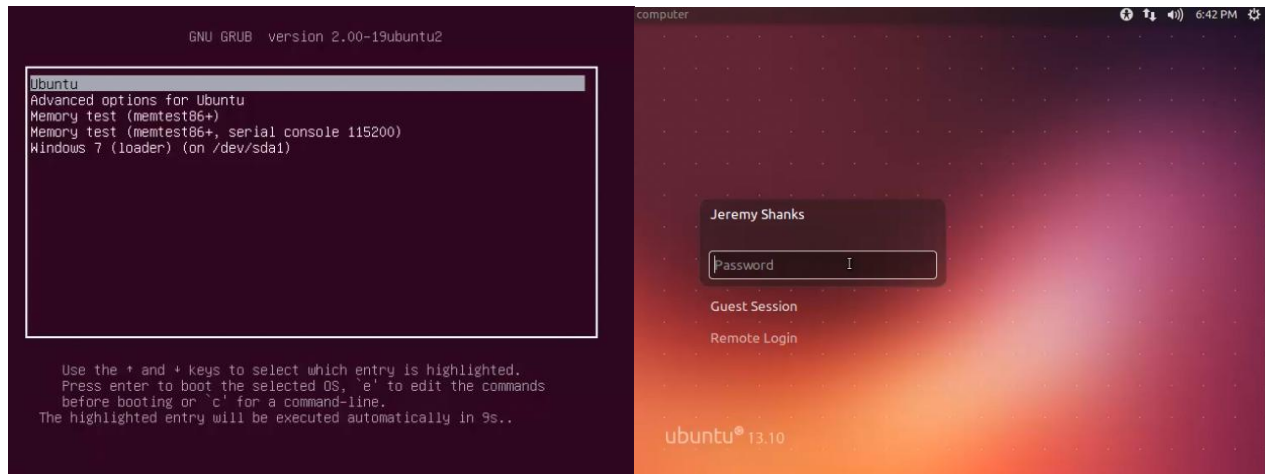


Now you will choose your keyboard layout. If you know what it is already then you may select it from the list. If not, then you may click “Detect Keyboard Layout” and follow the steps on the screen.

Now it is time to Type your Name or user name as Bigdata, Computer name, username, and password. You can choose to have your computer log in automatically to the username you supplied or to require you to log in every time you turn your computer on.

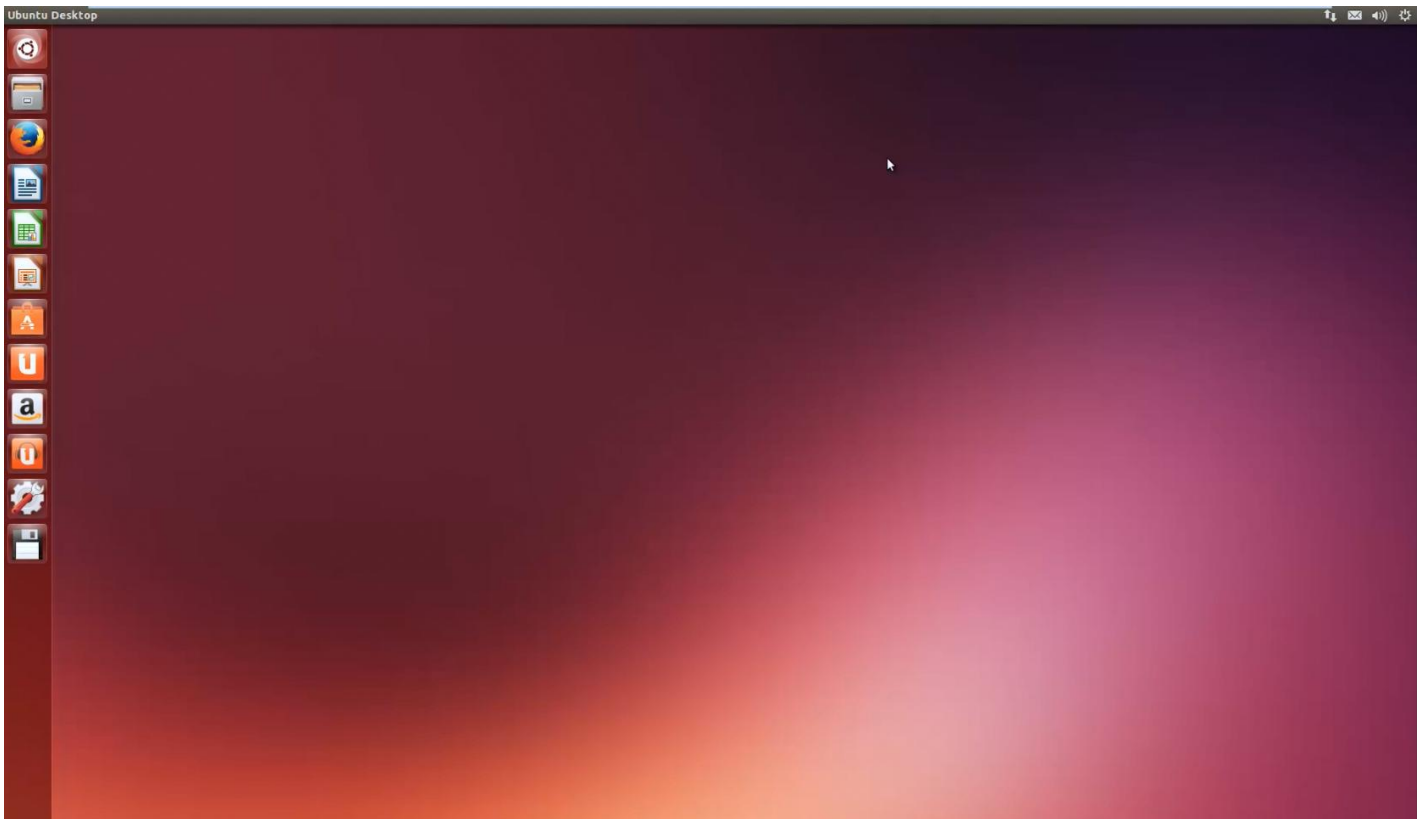


When your computer reboots you can choose between Ubuntu 13.10 (Top) or Windows 7 (Bottom). Choose Ubuntu.



Your computer will now boot into Ubuntu 13.10 and prompt you to login with your username and password provided earlier if you chose to require password at every boot. If you choose to automatically login then you will not see the login screen below but rather boot strait to the desktop.

After logging in you should see the desktop shown below. Welcome to Ubuntu 13.10.



Congratulations you have successfully installed Ubuntu 12.04 as a Dual-boot with Windows 7 onto your computer. Be sure to check out my site and find other helpful articles to get you started with Ubuntu 12.04 or 13/14 based on your choice.

Method 2 : Alternative method to install hadoop on windows using virtual box



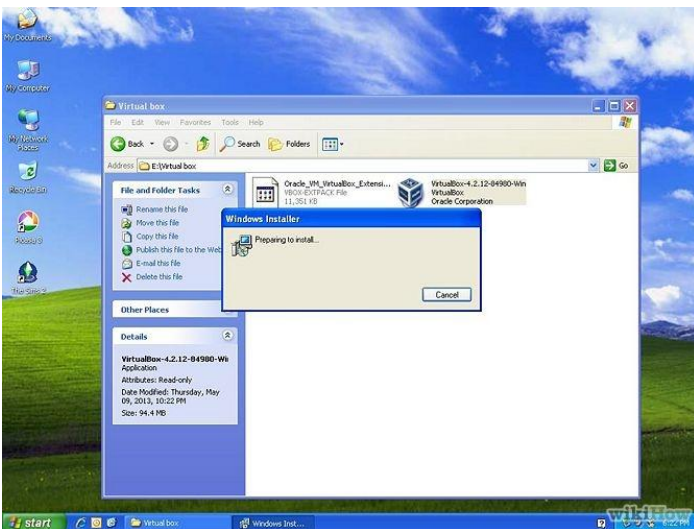
We have used following tools to setup our single node cluster

- Virtual Box 4.2.6 – 64 bit
- Ubuntu 12.04 – 64 bit Host & Guest OS
- Oracle Java 1.7 – 64 bit
- Hadoop 1.0.4

Installing VirtualBox

URL : <http://www.oracle.com/technetwork/server-storage/virtualbox/downloads/index.html>

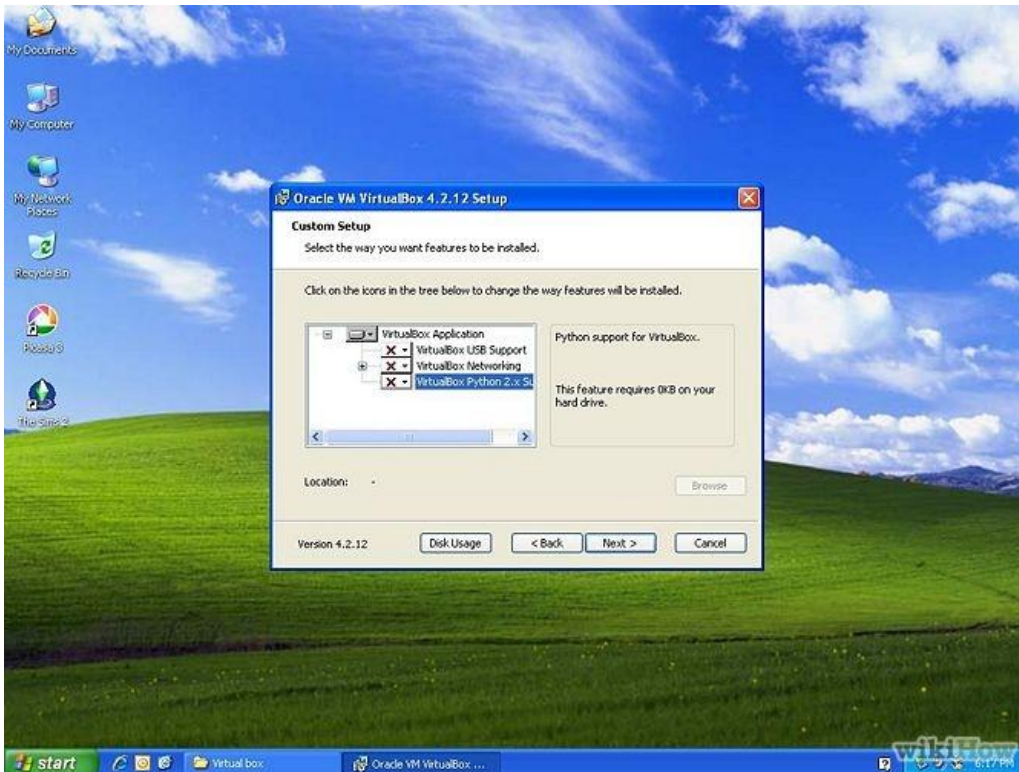
1 Download the latest version of VirtualBox. Go the VirtualBox website and go to the download section. There are versions available for Windows, Mac, and Linux. If you're not sure what operating system you're running, download the Windows version at the top by clicking "x86/amd64".



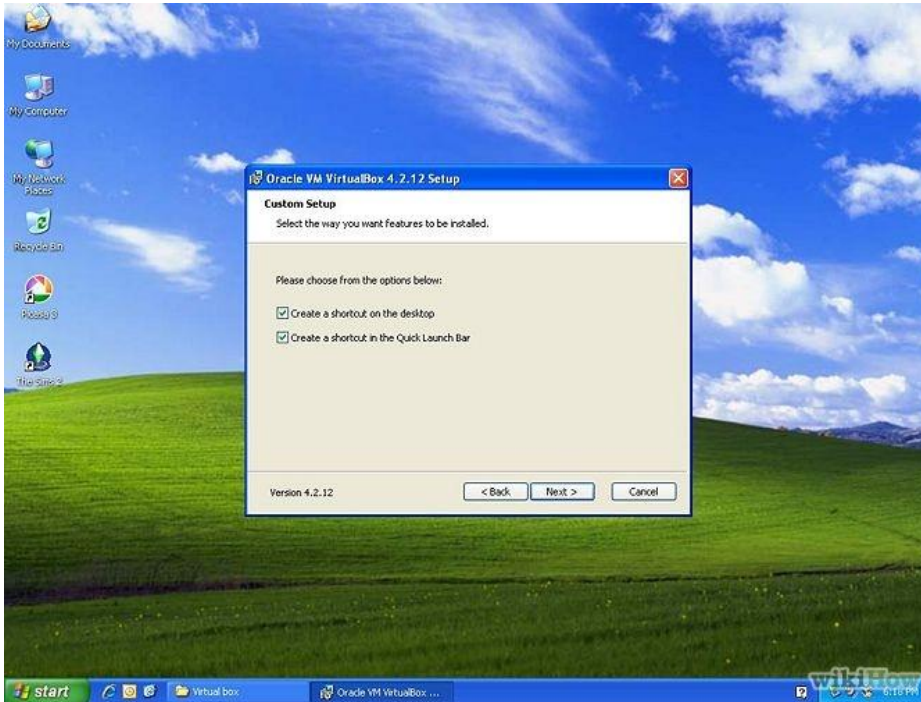
2 Start the installation and click “Next” to reach the license agreement.



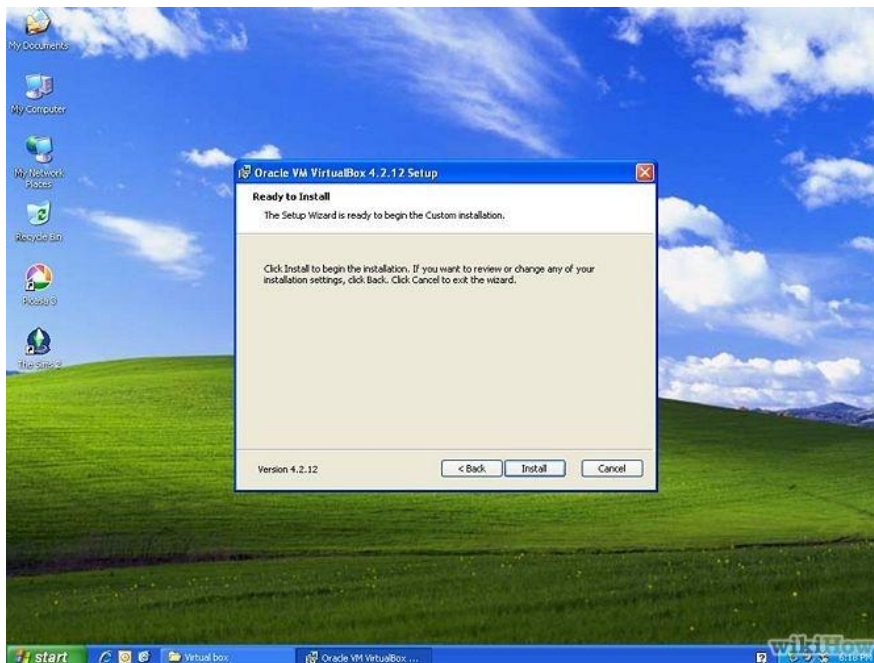
3 Choose the “I accept” option and click “Next” to continue.



4 Choose not to install USB support, networking, or Python support. Do this by clicking the grey icon near each option and selecting the red X or “Entire feature will be unavailable”. Then click “Next” to continue. If this is your first time dealing with virtual machines, this will eliminate the need to install custom drivers, which makes installing and uninstalling VirtualBox easier. If you’ve worked with virtual machines before, you can choose to keep these options selected.

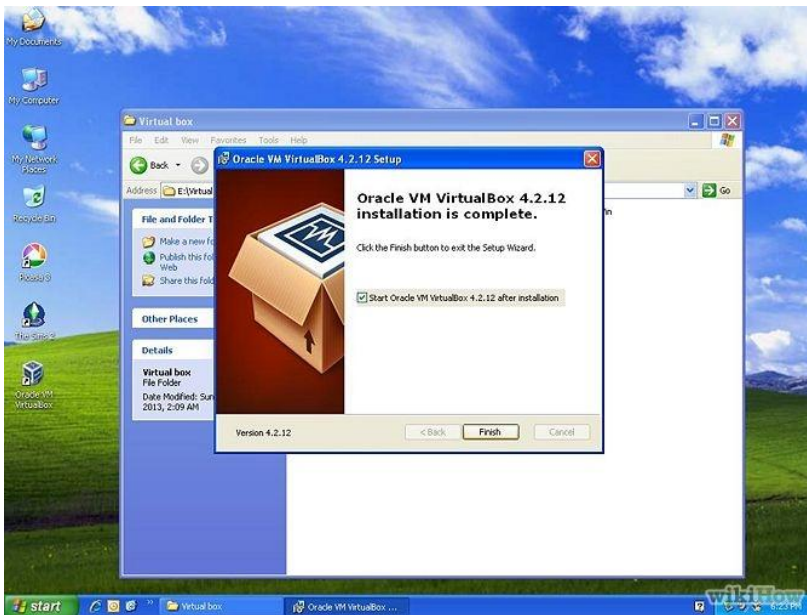


5 Deselect the Quick Launch Bar checkbox and click “Next” to



continue.

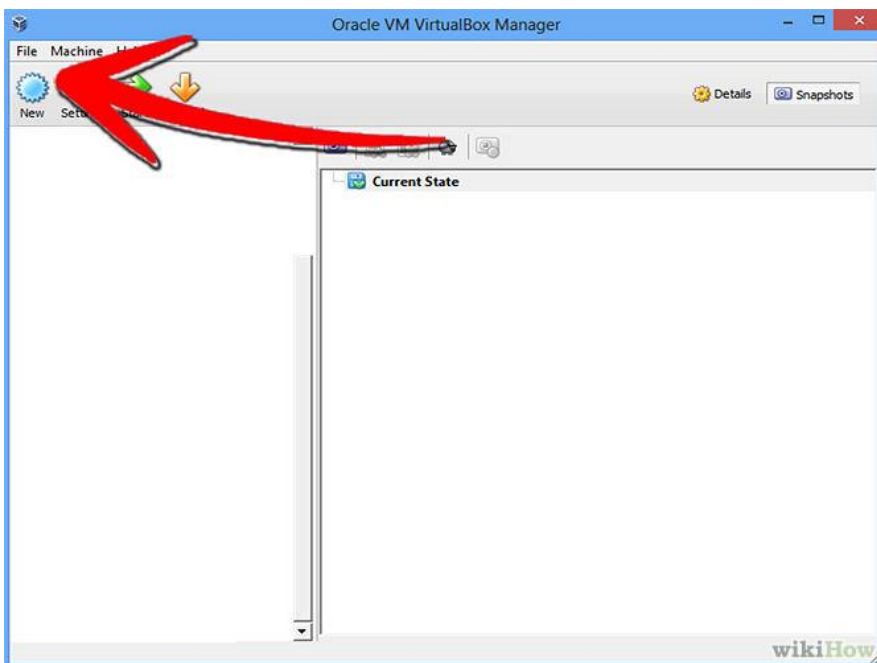
6 Click “Install” to install VirtualBox.



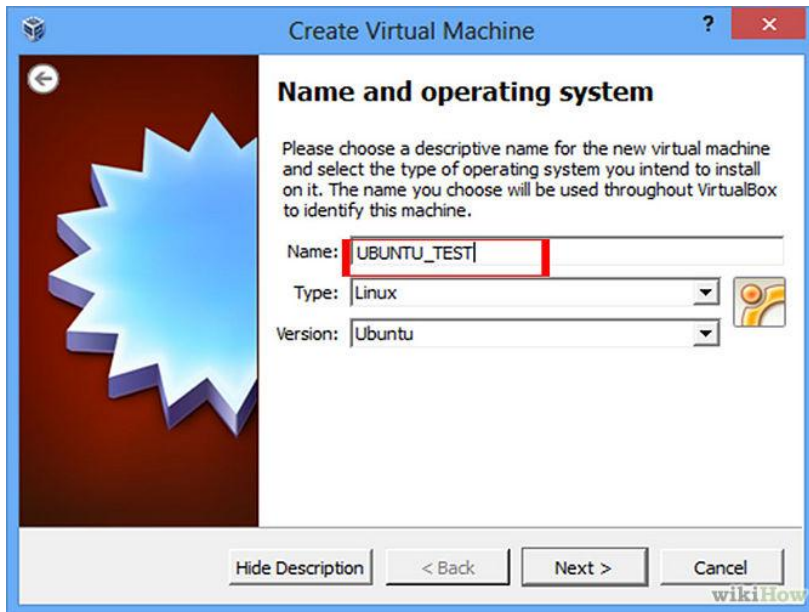
7 Click “Finish” to open VirtualBox.

Creating A New Virtual Machine

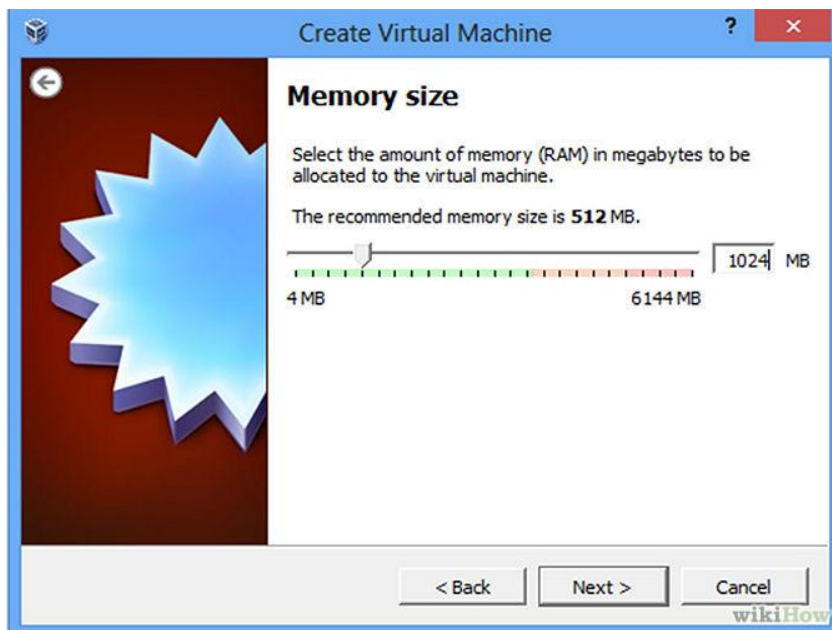
1 In VirtualBox, click the “New” button to start the virtual machine wizard.

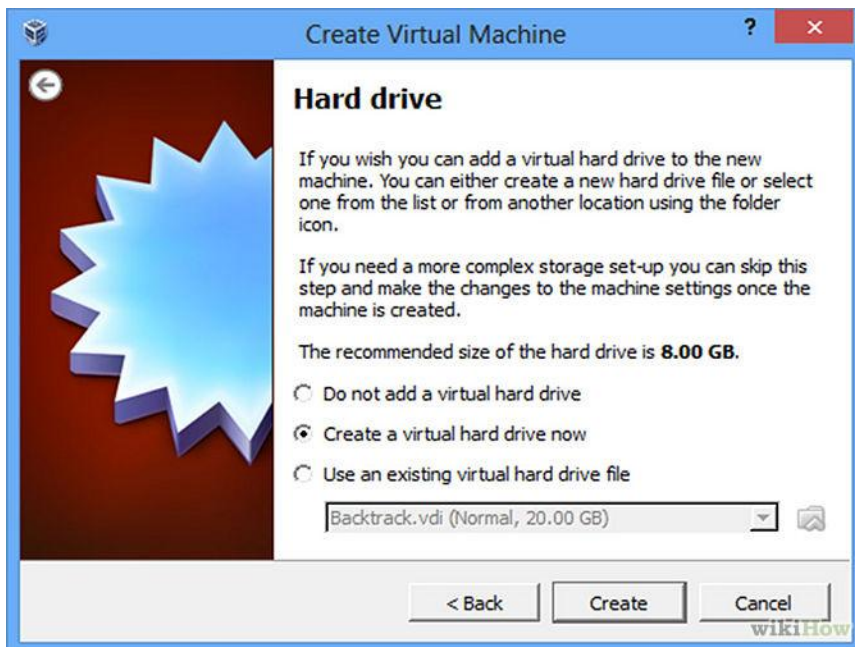


2 Give your virtual machine a name and select the operating system you'll be running. Click "Next". For this example, you'll be installing Ubuntu. Type any name in the Name field (such as Ubuntu or Linux). For "Operating System," choose "Linux." The version will automatically default to "Ubuntu." Click "Next" when you're done.



3 Select the amount of memory your VM will use and click "Next." When you chose your operating system in the previous step, VirtualBox automatically recommends the proper amount of memory to use. If you feel this amount isn't correct, you can move the slider or type a new amount in the box. Click "Next" when you're done.

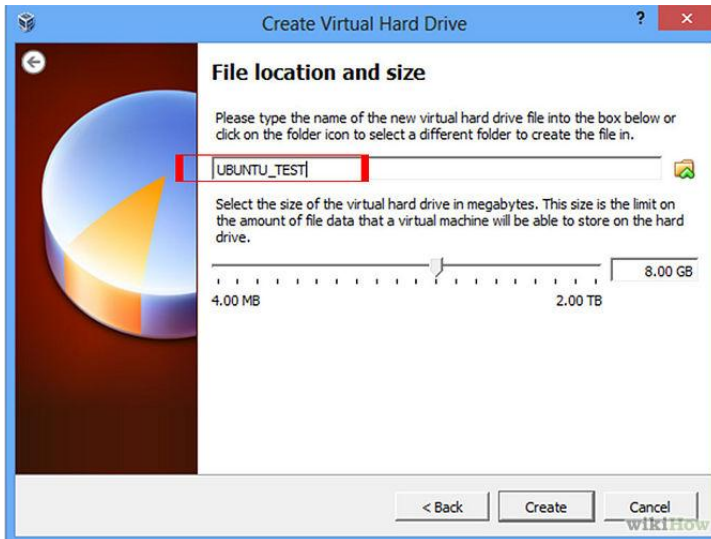




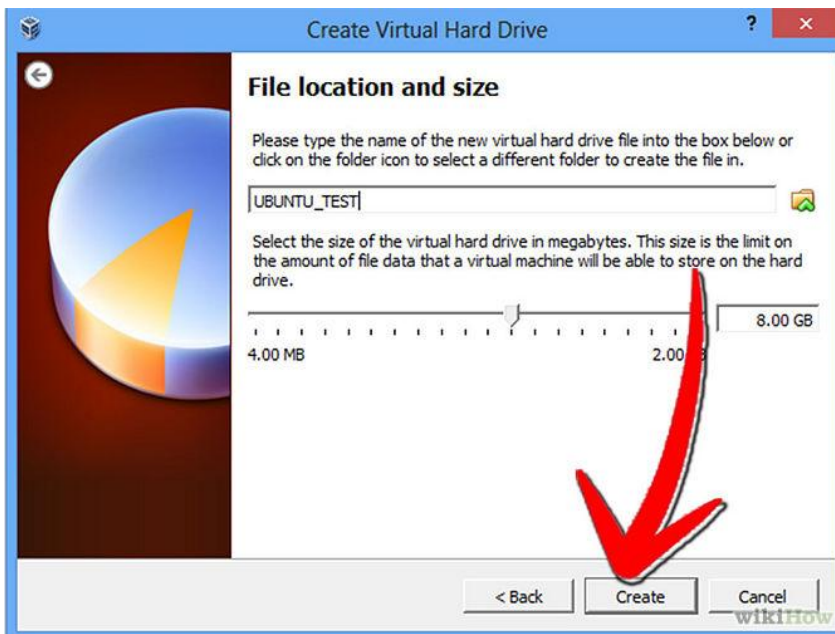
4 Click “Next” to create a new virtual hard disk, then click “Next” again. This opens a second wizard to create a new virtual hard disk.



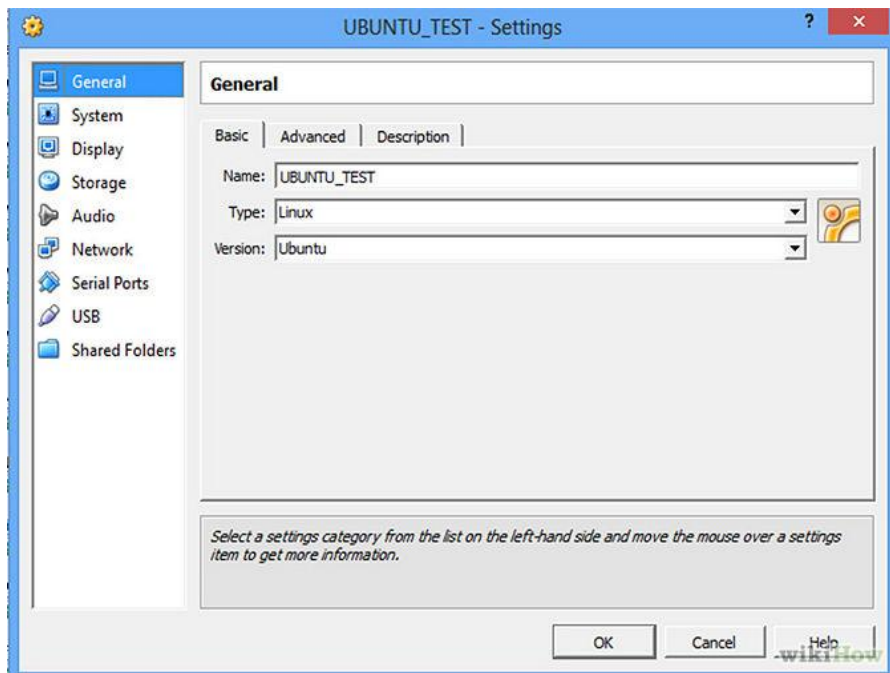
5 Select either **"Fixed-Size Storage"** or **"Dynamically Expanding Storage"** depending upon your needs. A fixed size storage is going to be the size of the virtual hard disk on the host OS (e.g.: a virtual disk 8 GB will be 8 GB on the host OS's hard disk). A dynamically expanding storage will be only the size of Ubuntu on your hard disk, but will grow in size as files are added to it until it reaches its limit (e.g.: virtual disk is created and has a 1 megabyte file on it. The size of the virtual disk is 1 megabyte. Then, another 1 megabyte file is added. The size of the virtual disk is 2 megabytes. This goes on until it reaches the specified size of the disk).



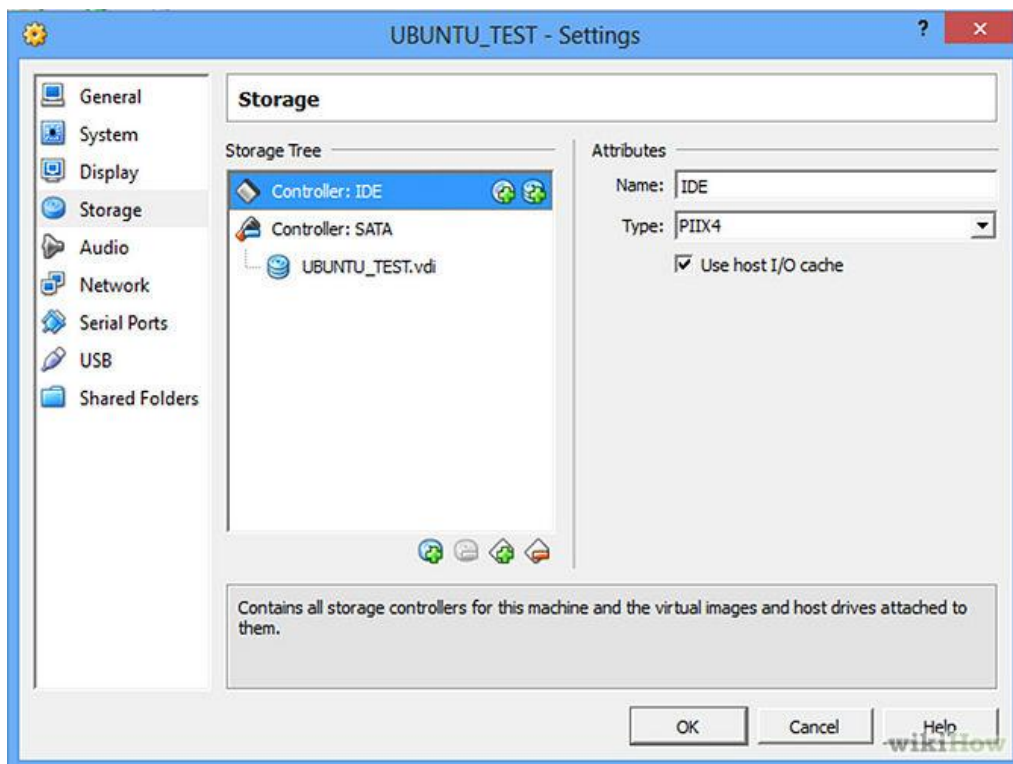
6 Click **"Next"** to accept the default name and size of the virtual hard disk. Again, VirtualBox recommends the proper size of your virtual hard disk. If you feel this amount isn't correct, you can move the slider or type a new amount in the box. Click **"Next"** when you're done.



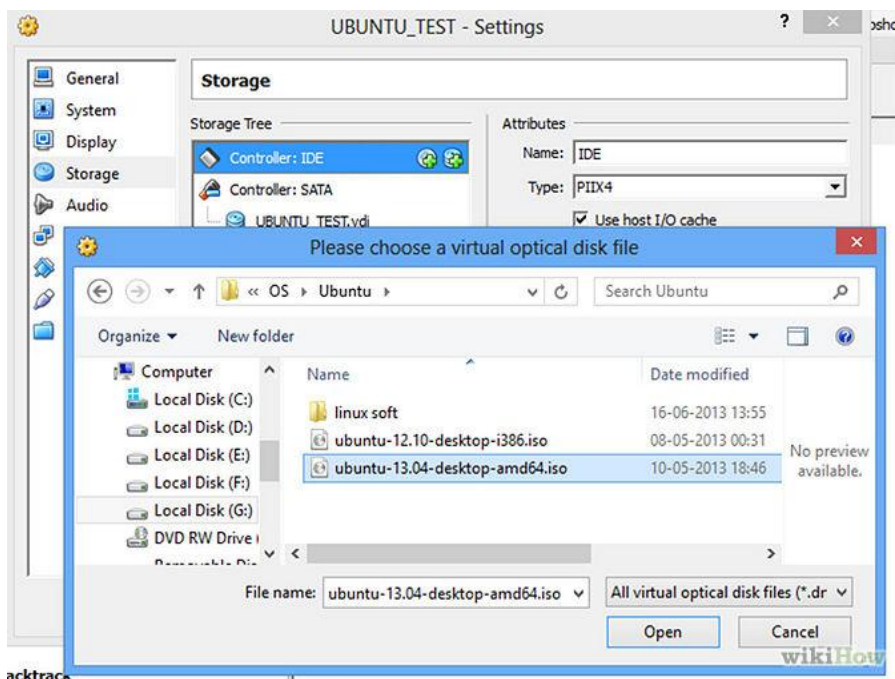
- 7** Click “Finish” and wait while VirtualBox creates the new virtual hard disk. You will see your new virtual machine in list.
- ## Setting The CD To Start



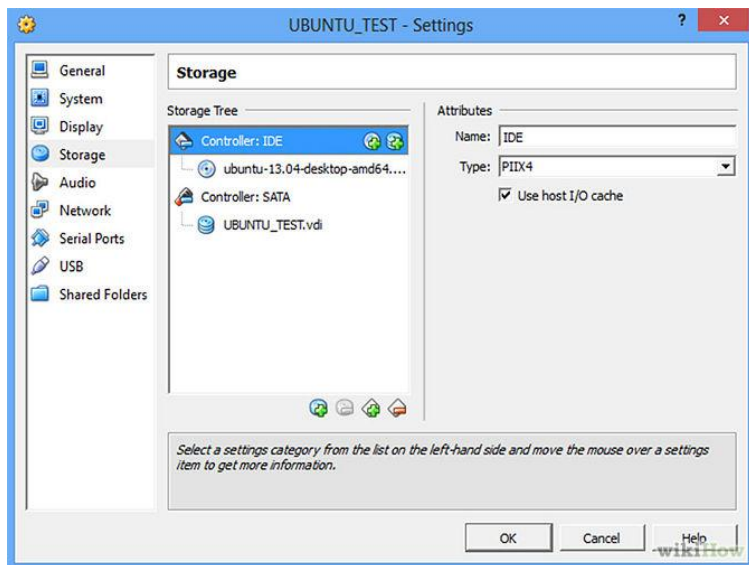
- 1** Select your new virtual machine. Once you've done this, click the “Settings” button.



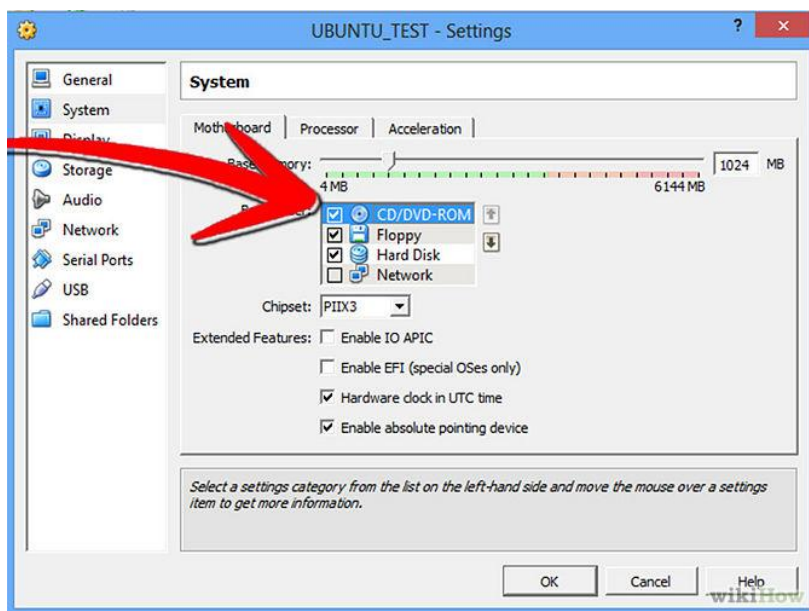
2 Click "Storage" tab.



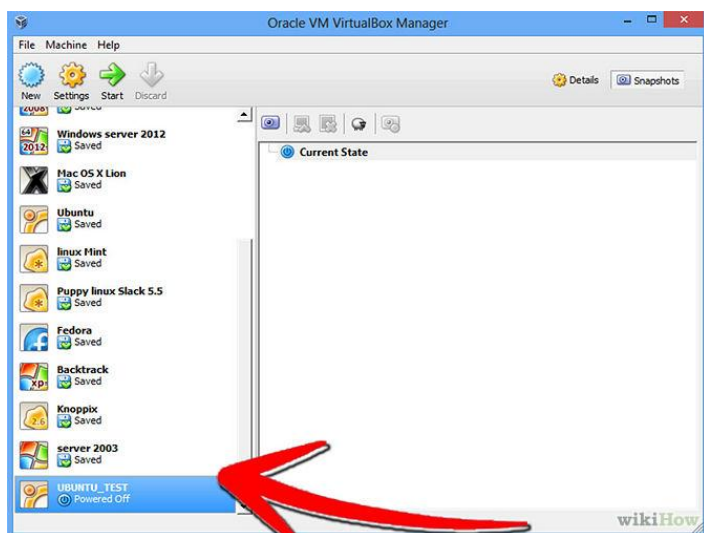
3 Click the "CD/DVD icon" having "+" on it and select ISO to mount



4 Ubuntu ISO will be mounted under controller device.

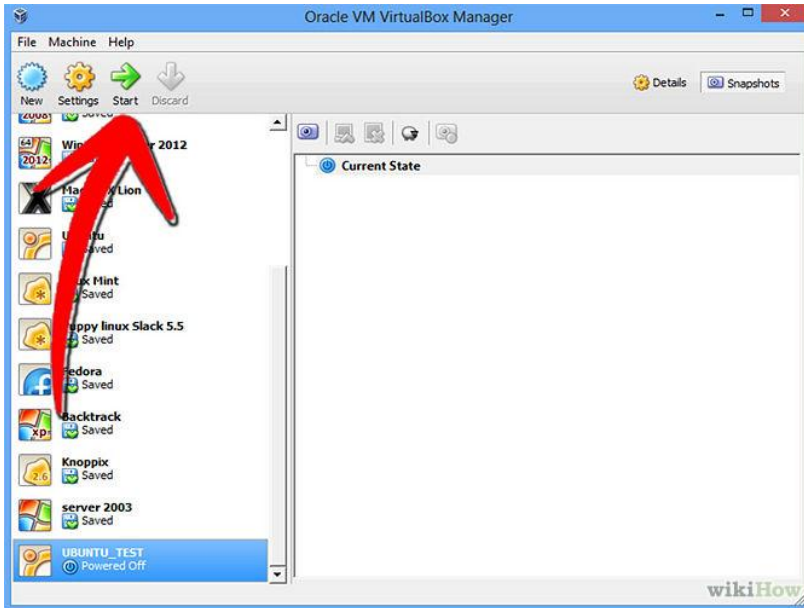


5 Click on the system tab on the left. Choose boot order and keep CD/DVD on the top as first priority.



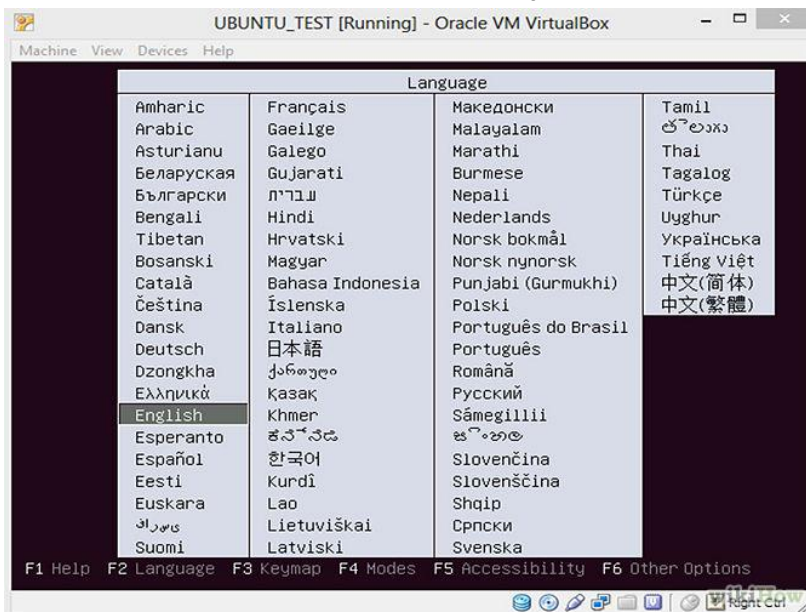
6 You may now close the settings window and return to the main window. Your Ubuntu machine is ready to boot now.

Installing Ubuntu

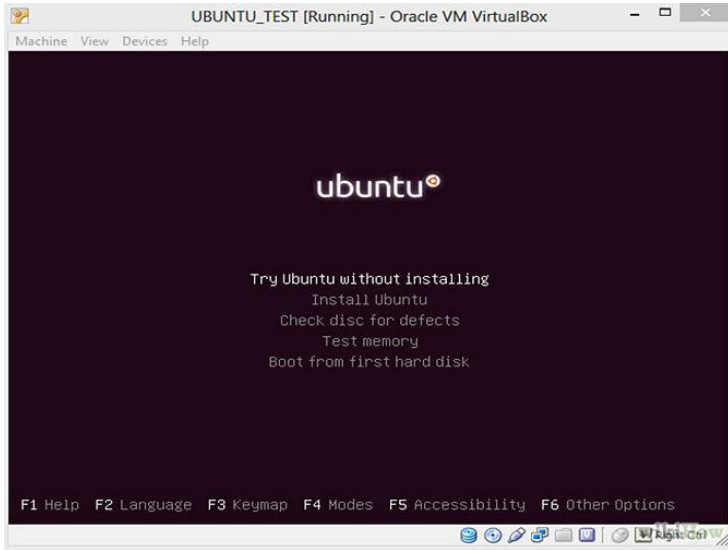


1 Select your virtual machine. Then click the "Start" button

Ubuntu Virtual machine will start in a separate window.



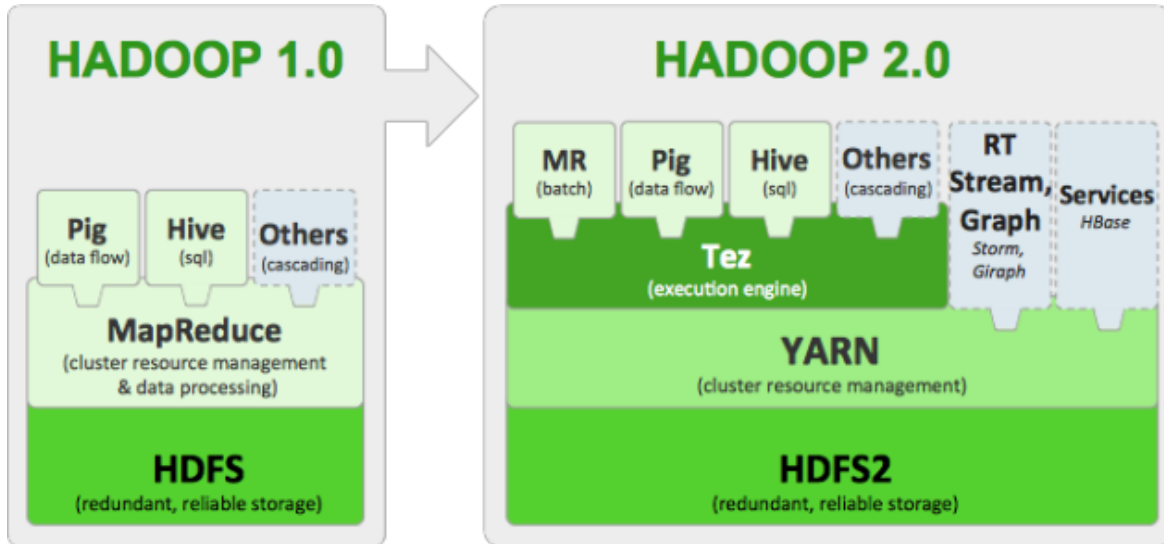
Machine will boot from selected ISO and you will see language option. Choose your preferred language and press Enter.



Follow Page Number 3 On Ubuntu installation instruction



Installing single node Hadoop 2.X on Ubuntu



Apache Hadoop 2.x release has significant changes compared to its previous stable release, which is Apache Hadoop 1.2.1

In short , this release has a number of changes compared to its earlier version 1.2.1:

YARN – A general purpose resource management system for Hadoop to allow MapReduce and other data processing frameworks like Hive, Pig and Services

High Availability for HDFS

HDFS Federation, Snapshots

NFSv3 access to data in HDFS

Introduced Application Manager to manage the application life cycle

Support for running Hadoop on Microsoft Windows

HDFS – Symlinks feature is disabled & will be taken out in future versions

Jobtracker has been replaced with Resource Manager and Node Manager

In this tutorial you will know step by step process for setting up a Hadoop Single Node cluster, so that you can play around with the framework and learn more about it.

In This tutorial we are using following Software versions, you can download same by clicking the hyperlinks:

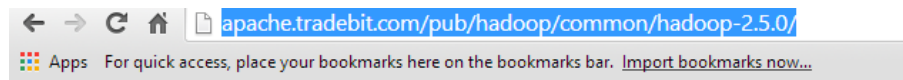


- [Ubuntu Linux](#) 12.04.3 LTS
- [Hadoop 2.X](#) (any version 2.0, 2.4, 2.5)






HADOOP DOWNLOAD URL : <http://hadoop.apache.org/releases.html#Download>

Direct Download URL : <http://apache.tradefbit.com/pub/hadoop/common/>

Hadoop 2.5 Download : <http://apache.tradefbit.com/pub/hadoop/common/hadoop-2.5.0/>



Index of /pub/hadoop/common/hadoop-2.5.0

Name	Last modified	Size	Description
 Parent Directory		-	
 hadoop-2.5.0-src.tar.gz	06-Aug-2014 12:51	15M	
 hadoop-2.5.0-src.tar.gz.mds	06-Aug-2014 15:54	1.2K	
 hadoop-2.5.0.tar.gz	12-Aug-2014 07:00	297M	
 hadoop-2.5.0.tar.gz.mds	12-Aug-2014 07:04	1.1K	

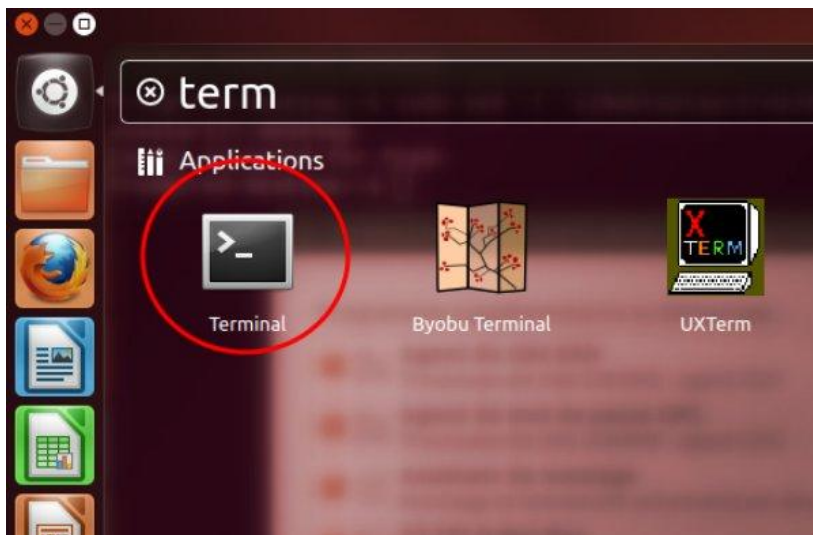
Apache/2.2.15 (CentOS) Server at apache.tradefbit.com Port 80

Step 1 : Download the hadoop-2.5.0.tar.gz file

Size : 297 MB



Step 2 : Click Ubuntu Terminal Windows



```
$ sudo addgroup hadoop
$ sudo adduser --ingroup hadoop hduser
$ su hduser
$ sudo tar vxzf hadoop-2.2.0.tar.gz -C /usr/local
$ cd /usr/local
$ sudo mv hadoop-2.2.0 hadoop
$ sudo chown -R hduser:hadoop hadoop
```

Activate SSH without password

Hadoop master node remotely control its sub-nodes using SSH. In single node cluster, master and sub-nodes run on the same machine but Hadoop is not aware of that. It will still use the exact same way to communicate between them using SSH.

Install SSH server:

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

Create a ssh-key without password:

```
$ cd ~  
$ ssh-keygen -t rsa -P ""
```

Set the key as trusted key for remote login:

```
$ cat .ssh/id_rsa.pub >> .ssh/authorized_keys
```

Try to connect on localhost and accept the connection (mandatory)

```
$ ssh localhost
```

```
doduck@doduck:~$ ssh-keygen -t rsa -P ""  
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/doduck/.ssh/id_rsa):  
Created directory '/home/doduck/.ssh'.  
Your identification has been saved in /home/doduck/.ssh/id_rsa.  
Your public key has been saved in /home/doduck/.ssh/id_rsa.pub.  
The key fingerprint is:  
e3:87:9c:89:7b:ce:db:6b:89:39:0e:f1:ae:38:84:35 doduck@doduck  
The key's randomart image is:  
+--[ RSA 2048 ]-----+  
|  
|   E  
| o .. S  
| . . * =  
| . o Bo..  
| .. ==oo  
| ..o=*+o.  
+-----+  
doduck@doduck:~$ cat .ssh/id_rsa.pub >> .ssh/authorized_keys  
doduck@doduck:~$ ssh localhost  
The authenticity of host 'localhost (127.0.0.1)' can't be established.  
ECDSA key fingerprint is 17:11:4f:9b:1b:6c:5f:4a:70:02:12:68:82:22:7f:4d.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.  
■
```

open `/etc/sysctl.conf` in the editor of your choice and add the following lines to the end of the file
`/etc/sysctl.conf` add the following lines and save.

STEP 7: Disabling IPv6

```
# disable ipv6  
net.ipv6.conf.all.disable_ipv6 = 1  
net.ipv6.conf.default.disable_ipv6 = 1  
net.ipv6.conf.lo.disable_ipv6 = 1
```


You have to reboot your machine in order to make the changes take effect.

You can check whether IPv6 is enabled on your machine with the following command:

```
1$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

A return value of 0 means IPv6 is enabled, a value of 1 means disabled (that's what we want).

Given the fact that Apache Hadoop is not currently supported on IPv6 networks (see [Hadoop and IPv6](#)) we will disable IPv6 in Java by editing **hadoop-env.sh** again.

```
1 $ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

Add the following line at the bottom of the file

```
1 HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
```

STEP 8: Configuring HDFS

The Hadoop Distributed File System (HDFS) is a reliable distributed file system designed to run on ordinary hardware and to store very large amounts of data (terabytes or even petabytes). HDFS is highly fault-tolerant because from a practical standpoint it was built upon the premise that **hardware failure is the norm rather than the exception** (see [HDFS Architecture Guide](#)). Thus, failure detection, distributed replication and quick recovery are in its core architecture.

The main configurations are stored in the 3 files below: **these files are located on etc/hadoop folder**

core-site.xml – contains default values for core Hadoop properties.

mapred-site.xml – contains configuration information for MapReduce properties.

hdfs-site.xml – contains server side configuration of your distributed file system.

core-site.xml

In between `<configuration> ... </configuration>` put the below code:

```
<property>
  <name>hadoop.tmp.dir</name>
  <value>/tmp/hadoop_data</value>
  <description>directory for hadoop data</description>
</property>

<property>
  <name>fs.default.name</name>
```

```
<value>hdfs://localhost:54310</value>
<description> data to be put on this URI</description>
</property>
```

mapred-site.xml

```
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
  <description>...
</description>
</property>
```

hdfs-site.xml

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

Setting Bashrc.sh Open bashrc from home and add the following lines

```
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
```

Formatting and Starting the Single Node Cluster.

So if you are done till now successfully, you are done with the installation part. Now we just have to format the namenode and start the cluster.

```
hduser@ubuntu:~$ /usr/local/hadoop/bin/hadoop namenode -format
```

the output will be something like:

Starting the single node cluster:

```
hduser@ubuntu:~$ /usr/local/hadoop/bin/start-all.sh
```

After the start-up you will get an output like:

```
hduser@com:/$ /usr/local/hadoop/bin/start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduser-
namenode-com.out
localhost: Warning: $HADOOP_HOME is deprecated.
localhost:
localhost: starting datanode, logging to /usr/local/hadoop/libexec/../logs/had
oop-hduser-datanode-com.out
localhost: Warning: $HADOOP_HOME is deprecated.
localhost:
localhost: starting secondarynamenode, logging to /usr/local/hadoop/libexec/..
/logs/hadoop-hduser-secondarynamenode-com.out
starting jobtracker, logging to /usr/local/hadoop/libexec/../logs/hadoop-hduse
r-jobtracker-com.out
localhost: Warning: $HADOOP_HOME is deprecated.
localhost:
localhost: starting tasktracker, logging to /usr/local/hadoop/libexec/../logs/
hadoop-hduser-tasktracker-com.out
hduser@com:/$
```

The above command starts the Namenode, Datanode, Secondary Namenode, Job Tracker and Task Tracker on your local machine.

you can try using the **JPS** command to see if these services are running or not.

```
hduser@ubuntu:/usr/local/hadoop$ jps
```

```
2246 TaskTracker
1927 JobTracker
1944 DataNode
2091 SecondaryNameNode
2311 Jps
1993 NameNode
```

So here you are done with the Single node installation of hadoop on your local machine.



Step 1 » Download the latest eclipse package from
here <http://www.eclipse.org/downloads/?osType=linux>

Step 2 » move the package to the /opt directory

```
$:~$ sudo mv eclipse-SDK-4.2.2-linux-gtk.tar.gz /opt/
```

Step 3 » Unzip the package by typing the below command

```
$:~$ cd /opt
```

```
$ /opt:~$ sudo tar -xvf eclipse-SDK-4.2.2-linux-gtk.tar.gz
```

Hadoop comes with several web interfaces which are by default (see conf/hadoop-default.xml) available at these locations:

- <http://localhost:50070/> – web UI of the NameNode daemon
- <http://localhost:50030/> – web UI of the JobTracker daemon
- <http://localhost:50060/> – web UI of the TaskTracker daemon
- <http://localhost:50090/> – web UI of Yarn

