
Hadoop Installation

The University of Texas at Dallas

Big Data Course CS6350

Professor: Dr. Latifur Khan

Author:

Ayoade Gbadebo (gga110020@utdallas.edu)

Vishal Karande (vmk130030@utdallas.edu)

Revision: 1

Release Date: 22-Jan-2015(Spring 2015)

Introduction

This document will guide students to install hadoop using Hortonworks sandbox. If you have hadoop already on your system you may skip this exercise.

Process

We will be using hortonworks sandbox for our hadoop installation. Hortonworks sandbox contains bigdata applications which include hadoop mapreduce, hive and pig.

We will be making use of the raw hadoop component.

Step 1: Download and install virtualbox.

Please use this url to download and install virtualbox.

For **windows users** : <http://download.virtualbox.org/virtualbox/4.3.20/VirtualBox-4.3.20-96997-Win.exe>

For **ubuntu users**: copy and paste this command to your terminal.

```
sudo sh -c "echo 'deb
http://download.virtualbox.org/virtualbox/debian '$
(lsb_release -cs)' contrib non-free' >
/etc/apt/sources.list.d/virtualbox.list" && wget -q
http://download.virtualbox.org/virtualbox/debian/oracle_vbo
x.asc -O- | sudo apt-key add - && sudo apt-get update &&
sudo apt-get install virtualbox-4.3 dkms
```

Step 2: Download Hortonworks sandbox image.

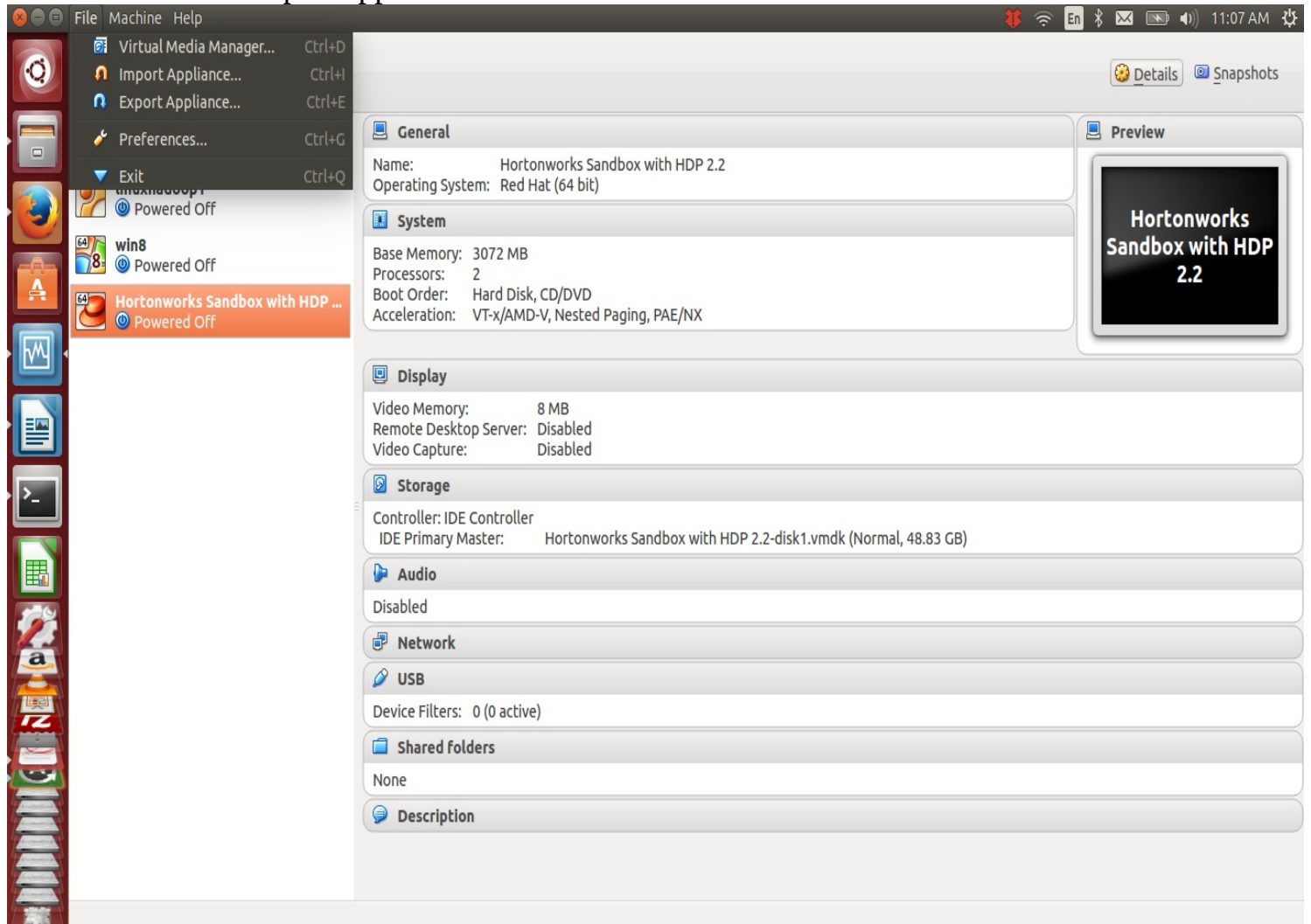
Use the url below to download Hortonworks sandbox for **virtualbox**.

http://hortonassets.s3.amazonaws.com/2.2/Sandbox_HDP_2.2_VirtualBox.ova

Step 3: Install Hortonworks sandbox virtual appliance.

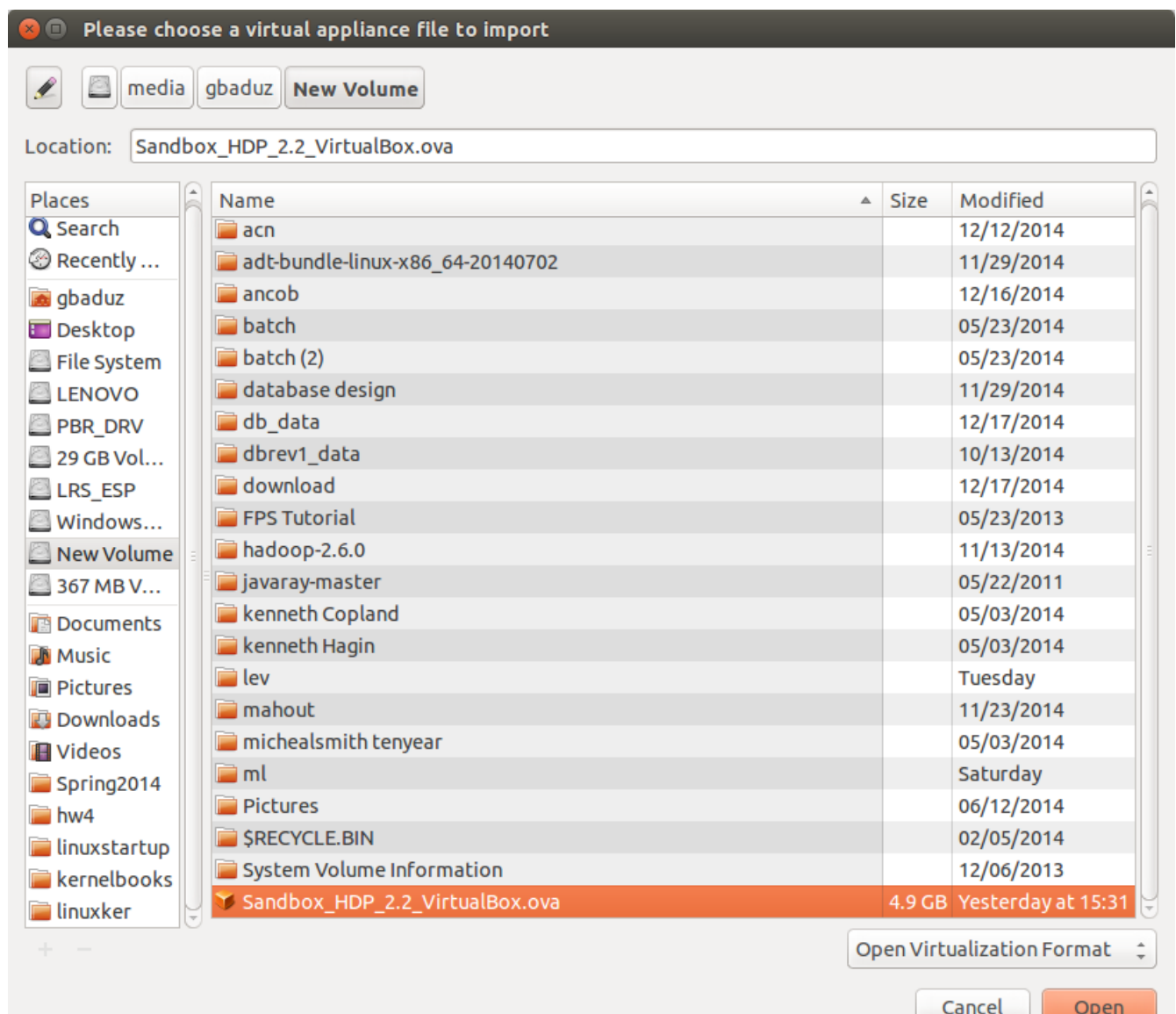
Start **virtualbox** application.

Go to file → import appliance





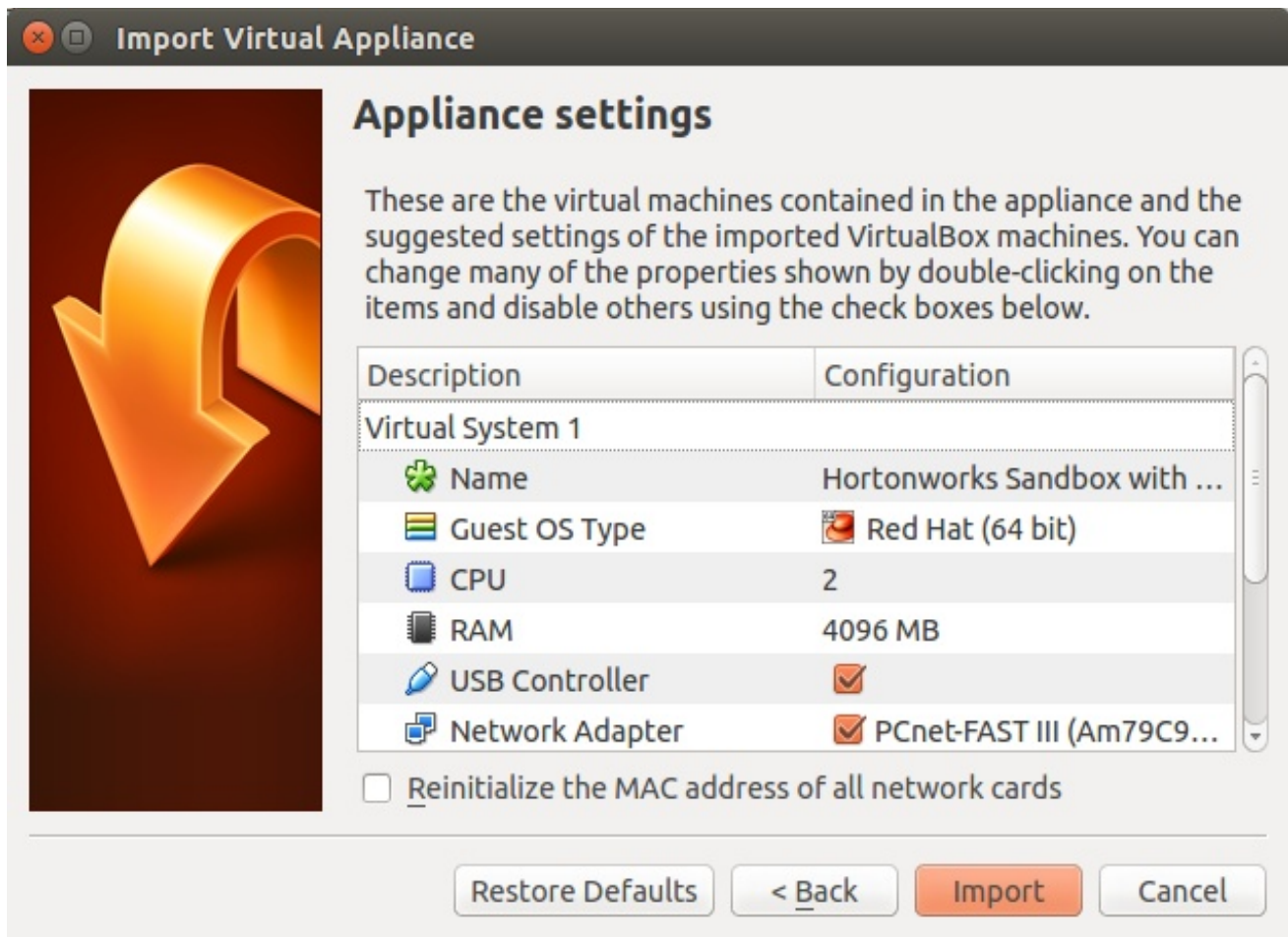
Click on the folder icon at the right to navigate to where you downloaded the hortonworks sandbox appliance file.



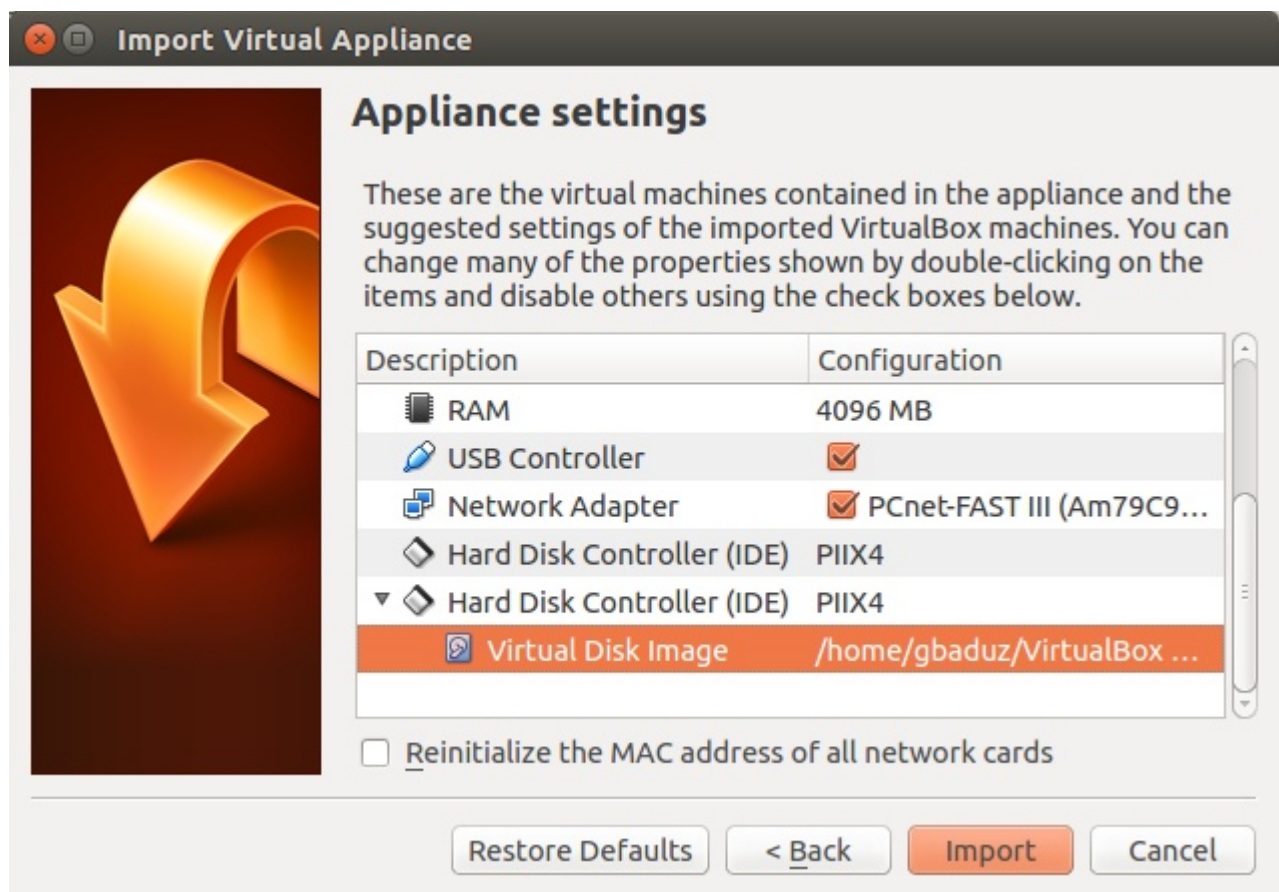
Select the Sandbox file and Click open.



Click next.



You can double click on the RAM size to increase or reduce as the case may be. You may have to reduce if you do not have a lot of RAM. Note (Virtualbox mostly does not use more than 1/2 of your physical ram if you have 6GB, reduce the RAM entry to 3000MB.) Use as much RAM your system can allow.



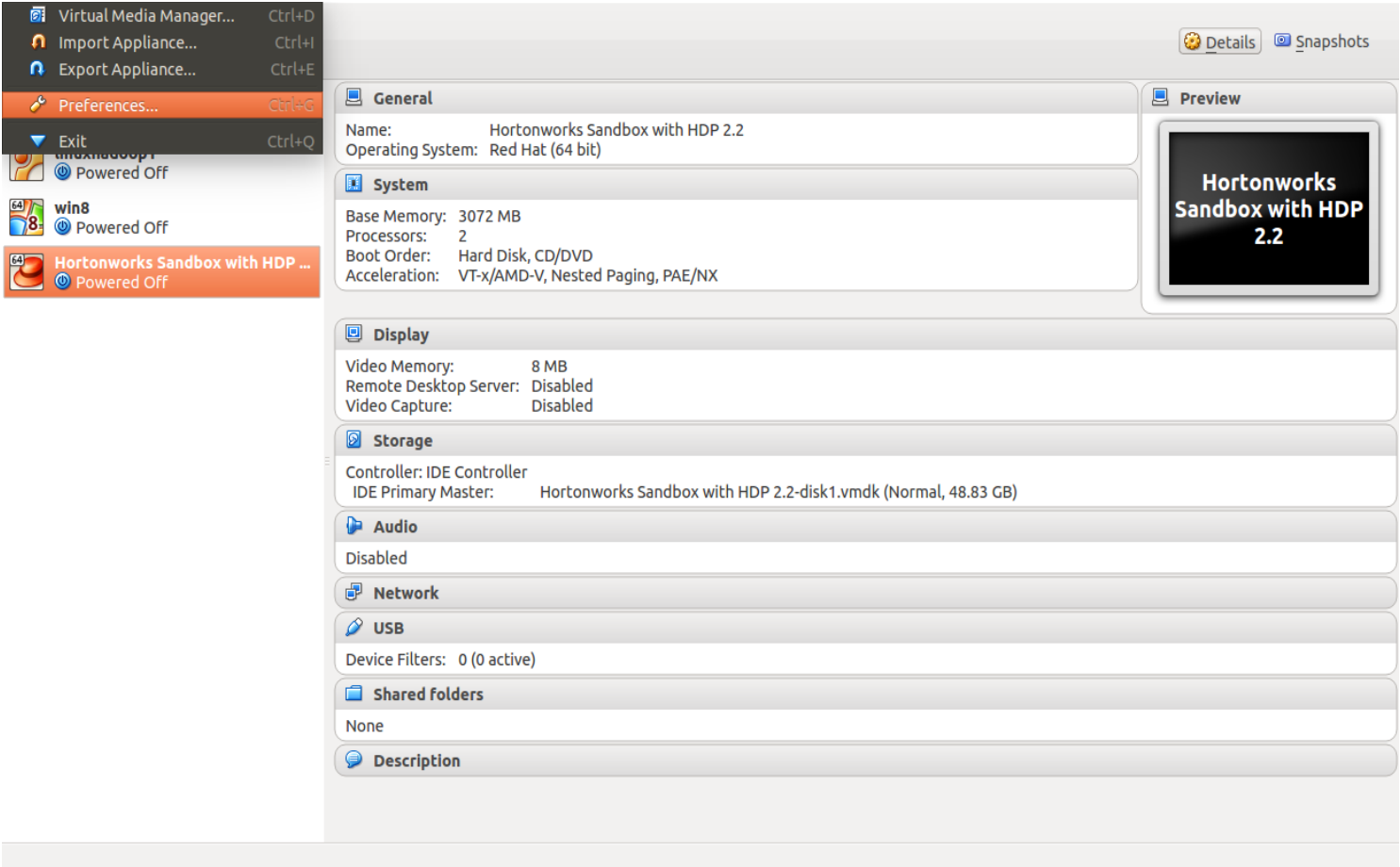
You can also scroll down to set the location of where you want the disk image installed.

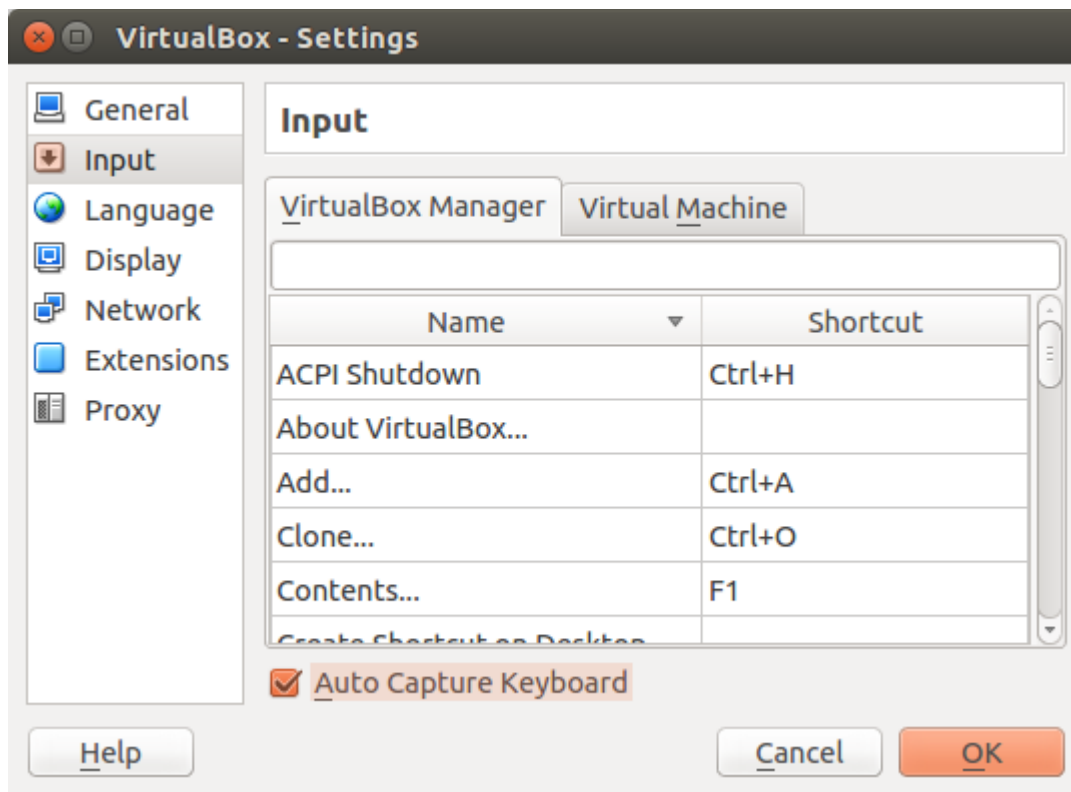
Double click on the virtualdisk image value and change to your preferred location.

Click on import.

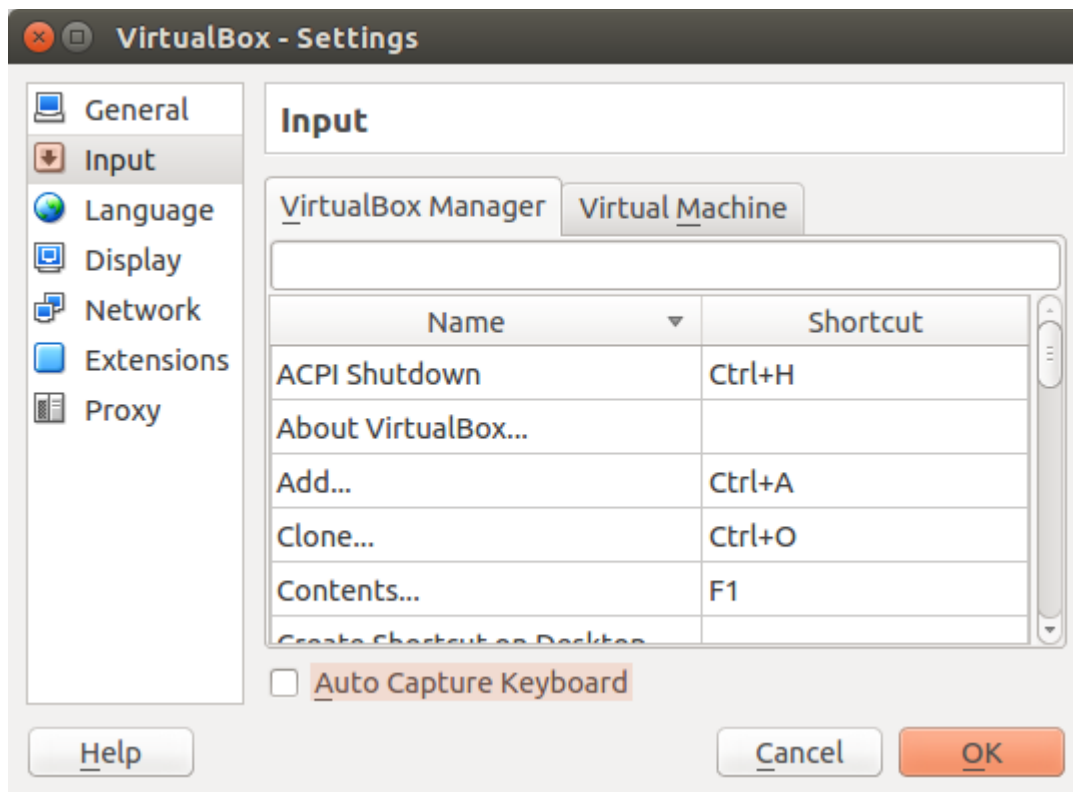
Wait for the import to finish.

Go to file → preferences





Click Input → Uncheck Auto capture keyboard and click OK.



Uncheck Auto capture keyboard and click OK.

Step 4: Configure the Hortonworks Sandbox with HDP 2.2 VM

Click on the **Hortonworks Sandbox with HDP 2.2 VM** as shown below on the left panel.

Click on the setting button at the left tab.

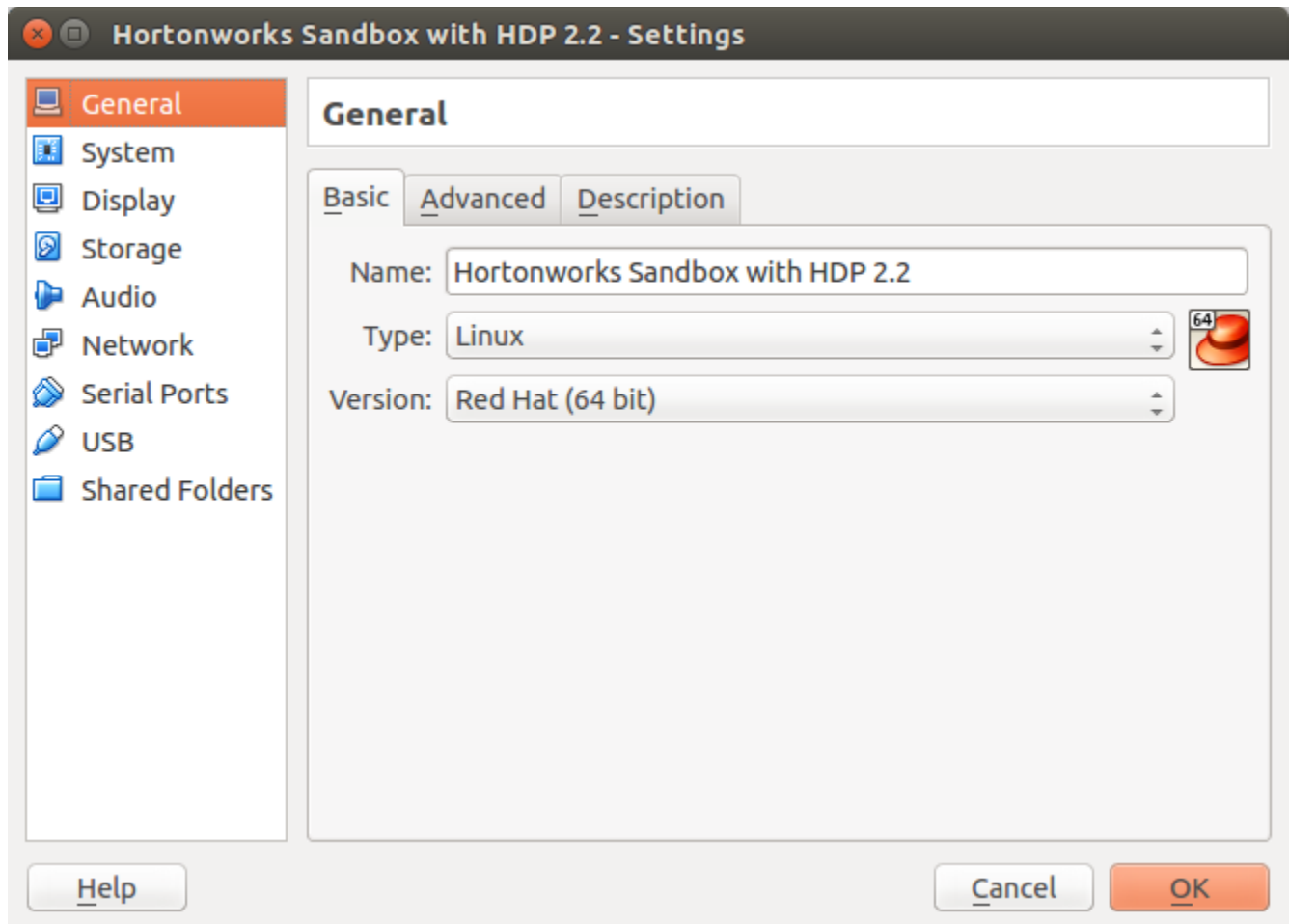
The screenshot displays the VMware Workstation interface. On the left, a list of virtual machines is shown: 'ubuntu' (Powered Off), 'linuxhadoop1' (Powered Off), 'win8' (Powered Off), and 'Hortonworks Sandbox with HDP ...' (Powered Off). The 'Hortonworks Sandbox with HDP ...' VM is selected and highlighted in orange. Above the list are buttons for 'New', 'Settings', 'Start', and 'Discard'. To the right of the VM list, there are buttons for 'Details' and 'Snapshots'. The main panel shows the configuration for the selected VM, organized into tabs: 'General', 'System', 'Display', 'Storage', 'Audio', 'Network', 'USB', 'Shared folders', and 'Description'. The 'General' tab is active, showing the VM's name, operating system, and system specifications. A 'Preview' window on the right shows a thumbnail of the VM's display.

VM List:

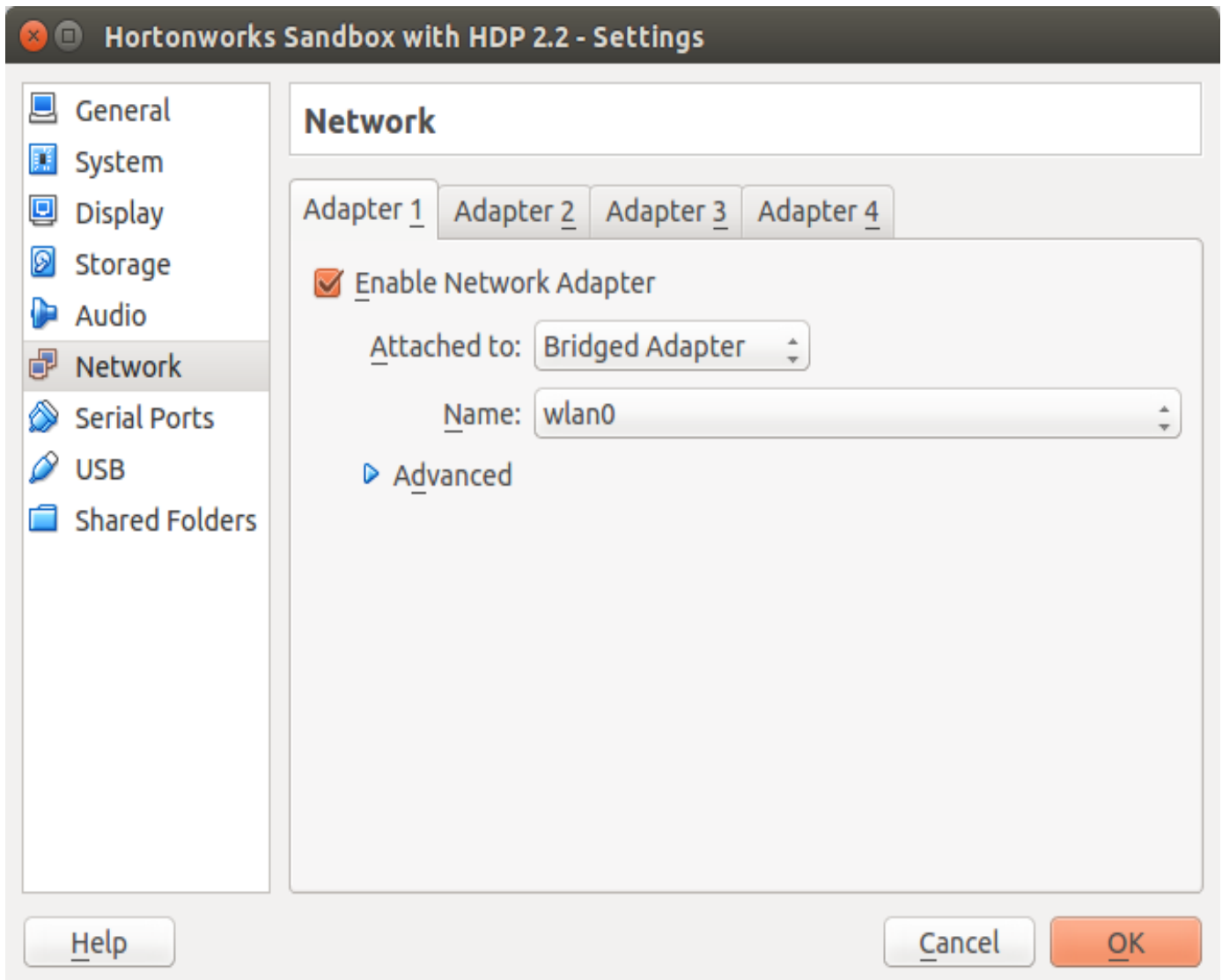
- ubuntu (Powered Off)
- linuxhadoop1 (Powered Off)
- win8 (Powered Off)
- Hortonworks Sandbox with HDP ... (Powered Off)**

Configuration Details for Hortonworks Sandbox with HDP 2.2:

- General:** Name: Hortonworks Sandbox with HDP 2.2, Operating System: Red Hat (64 bit)
- System:** Base Memory: 3072 MB, Processors: 2, Boot Order: Hard Disk, CD/DVD, Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX
- Display:** Video Memory: 8 MB, Remote Desktop Server: Disabled, Video Capture: Disabled
- Storage:** Controller: IDE Controller, IDE Primary Master: Hortonworks Sandbox with HDP 2.2-disk1.vmdk (Normal, 48.83 GB)
- Audio:** Disabled
- Network:**
- USB:** Device Filters: 0 (0 active)
- Shared folders:** None
- Description:**



Click on the network



Click on Adapter 1 tab.

Attached to: Change to bridged adapter.

Name: use the name of the network device on your machine. In my case **wlan0**.

Click OK

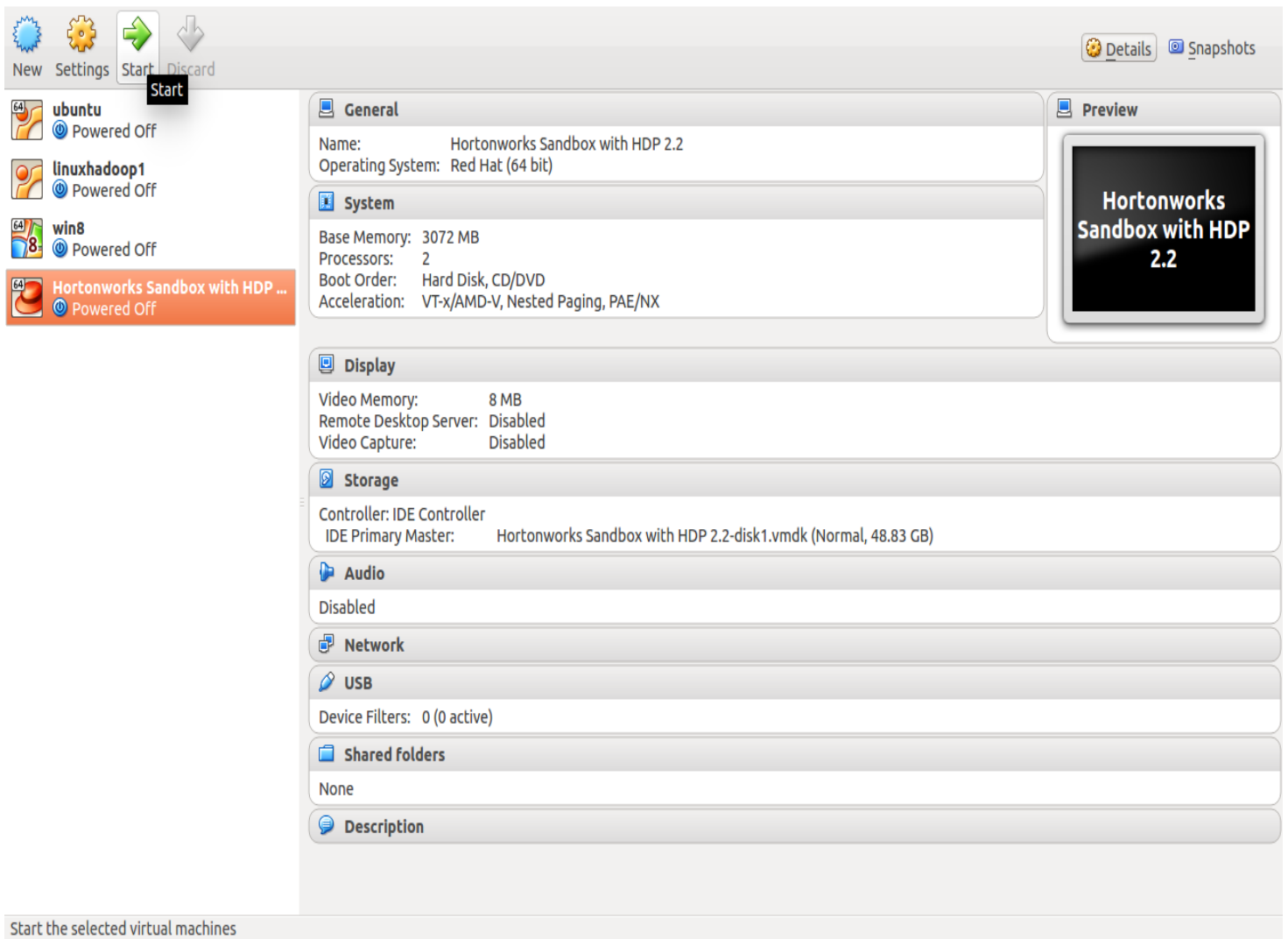
To get the name of the network device on your machine

→ go to terminal and type 'ifconfig'

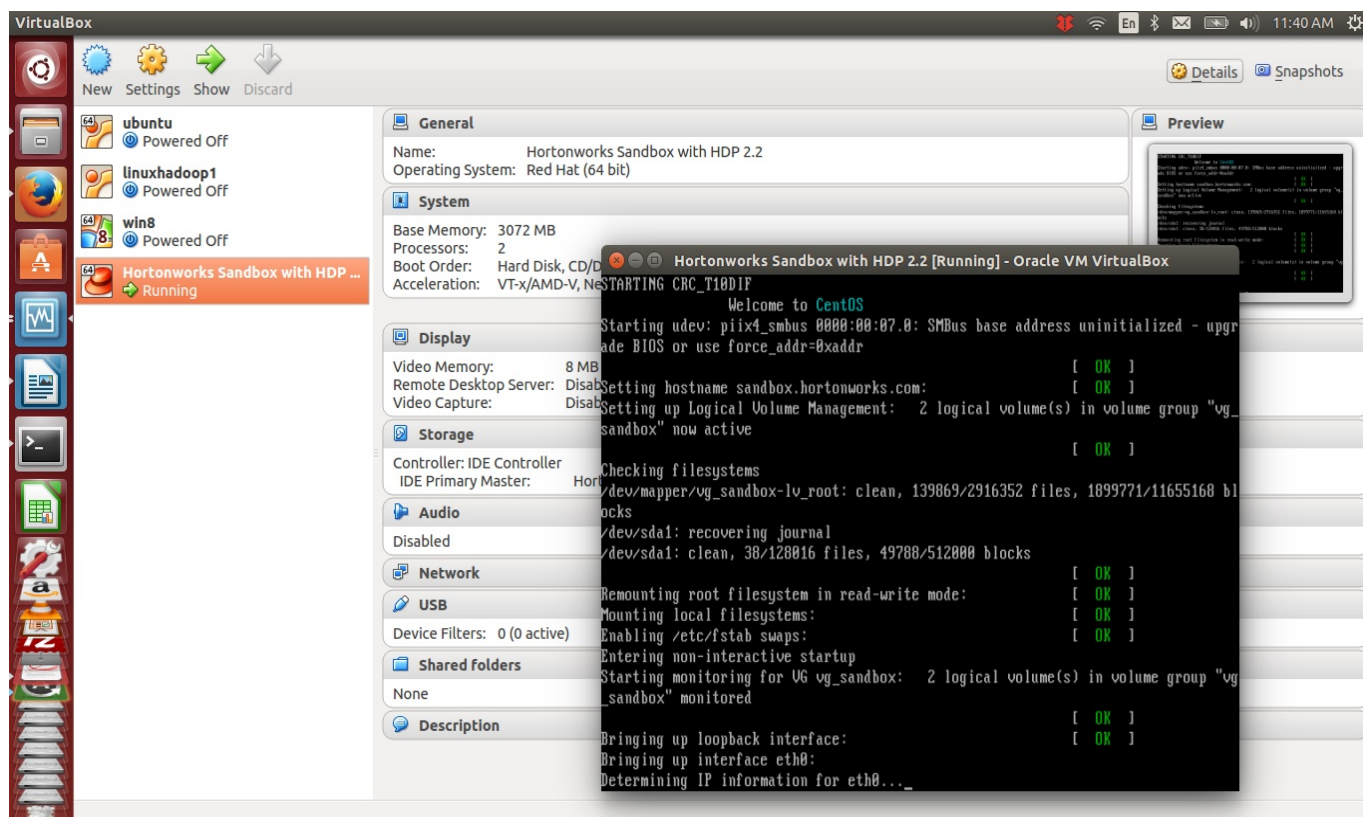
→ Identify your connection.

```
gbaduz@gbaduz-Lenovo-IdeaPad-P500: ~  
RX packets:0 errors:0 dropped:0 overruns:0 frame:0  
TX packets:0 errors:0 dropped:0 overruns:0 carrier:0  
collisions:0 txqueuelen:1000  
RX bytes:0 (0.0 B) TX bytes:0 (0.0 B)  
  
lo      Link encap:Local Loopback  
        inet addr:127.0.0.1  Mask:255.0.0.0  
        inet6 addr: ::1/128 Scope:Host  
        UP LOOPBACK RUNNING  MTU:65536  Metric:1  
        RX packets:63771 errors:0 dropped:0 overruns:0 frame:0  
        TX packets:63771 errors:0 dropped:0 overruns:0 carrier:0  
        collisions:0 txqueuelen:0  
        RX bytes:20336384 (20.3 MB)  TX bytes:20336384 (20.3 MB)  
  
wlan0   Link encap:Ethernet  HWaddr 84:a6:c8:7e:09:82  
        inet addr:192.168.1.102  Bcast:192.168.1.255  Mask:255.255.255.0  
        inet6 addr: fe80::86a6:c8ff:fe7e:982/64 Scope:Link  
        UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1  
        RX packets:1009942 errors:0 dropped:0 overruns:0 frame:0  
        TX packets:596431 errors:0 dropped:0 overruns:0 carrier:0  
        collisions:0 txqueuelen:1000  
        RX bytes:1299817508 (1.2 GB)  TX bytes:67463478 (67.4 MB)  
  
gbaduz@gbaduz-Lenovo-IdeaPad-P500:~$
```

See connection to the internet is with **wlan0**, in this example.

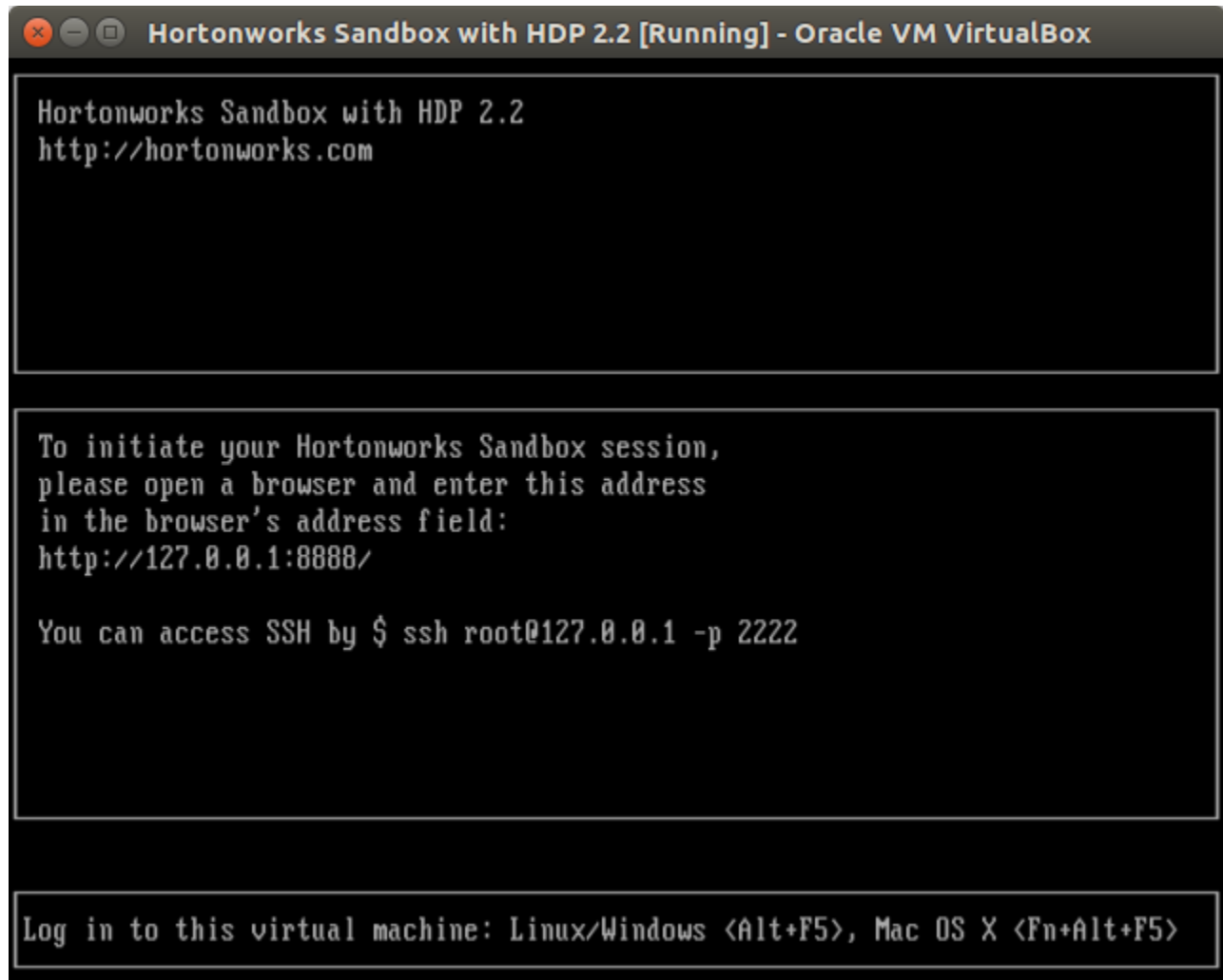


Click on **start** to launch the hortonworks sandbox vm as shown above.



This page shows the vm start up.

Press Alt + F5 to login to the VM. Please disregard the IP address shown below,
Next step shows how to get the real IP address of this VM.



Username : root
password : hadoop

You can also login as hue.

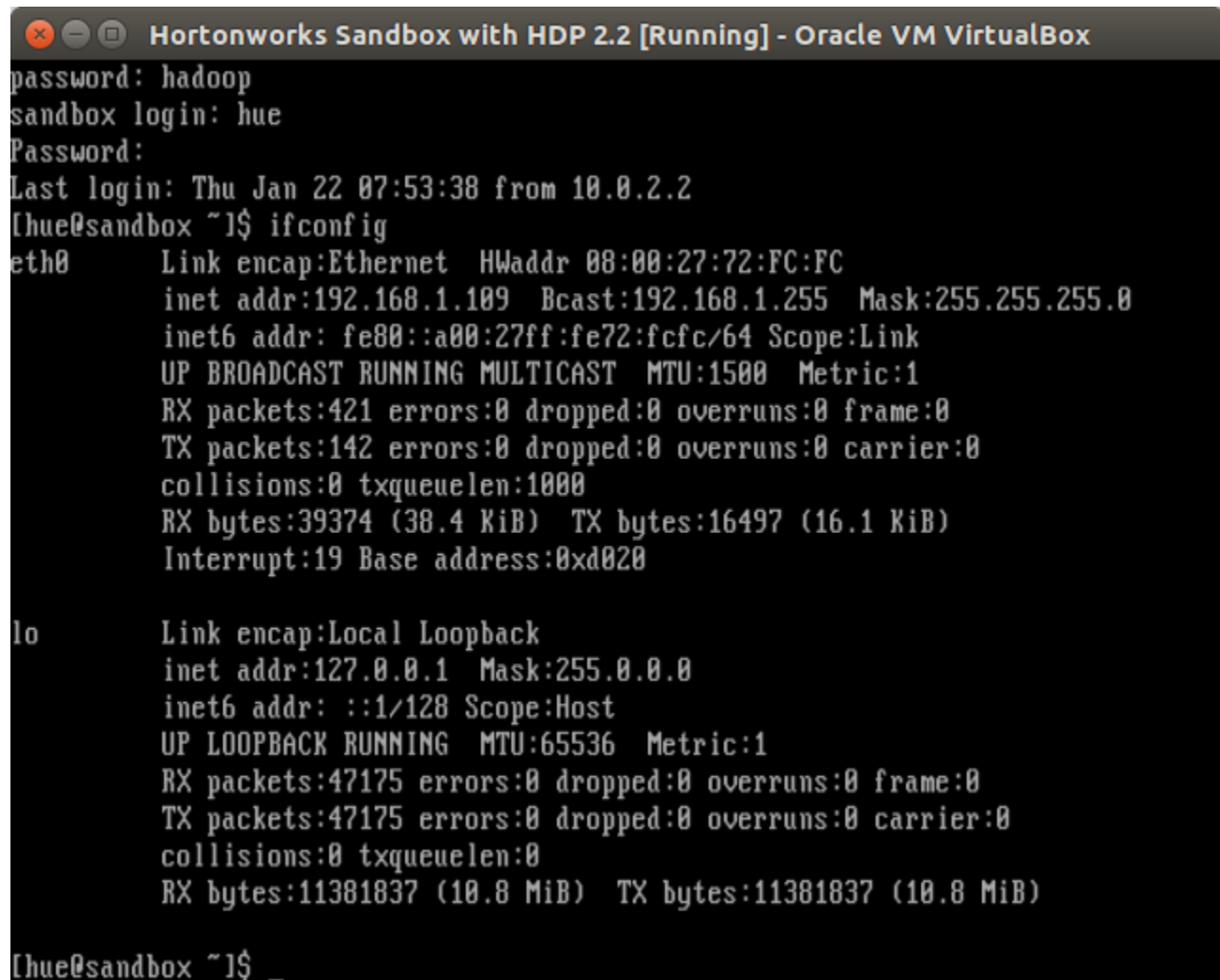
Username : hue
password : hadoop

For this tutorial use the User as hue

Login as **hue**, password is **hadoop**.

Determine the ip address of the vm

run the command `ifconfig` in the vm terminal prompt as shown below



```
Hortonworks Sandbox with HDP 2.2 [Running] - Oracle VM VirtualBox
password: hadoop
sandbox login: hue
Password:
Last login: Thu Jan 22 07:53:38 from 10.0.2.2
[hue@sandbox ~]$ ifconfig
eth0      Link encap:Ethernet  HWaddr 08:00:27:72:FC:FC
          inet addr:192.168.1.109  Bcast:192.168.1.255  Mask:255.255.255.0
          inet6 addr: fe80::a00:27ff:fe72:fcfc/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:421 errors:0 dropped:0 overruns:0 frame:0
          TX packets:142 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:39374 (38.4 KiB)  TX bytes:16497 (16.1 KiB)
          Interrupt:19 Base address:0xd020

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:47175 errors:0 dropped:0 overruns:0 frame:0
          TX packets:47175 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:11381837 (10.8 MiB)  TX bytes:11381837 (10.8 MiB)

[hue@sandbox ~]$ _
```

Note the inet address of the `eth0` interface which is 192.168.1.109 in the above snapshot.

This is the ip address of your vm. You can use `ssh` to access the VM from your host machine and run `hadoop` commands.

E.g

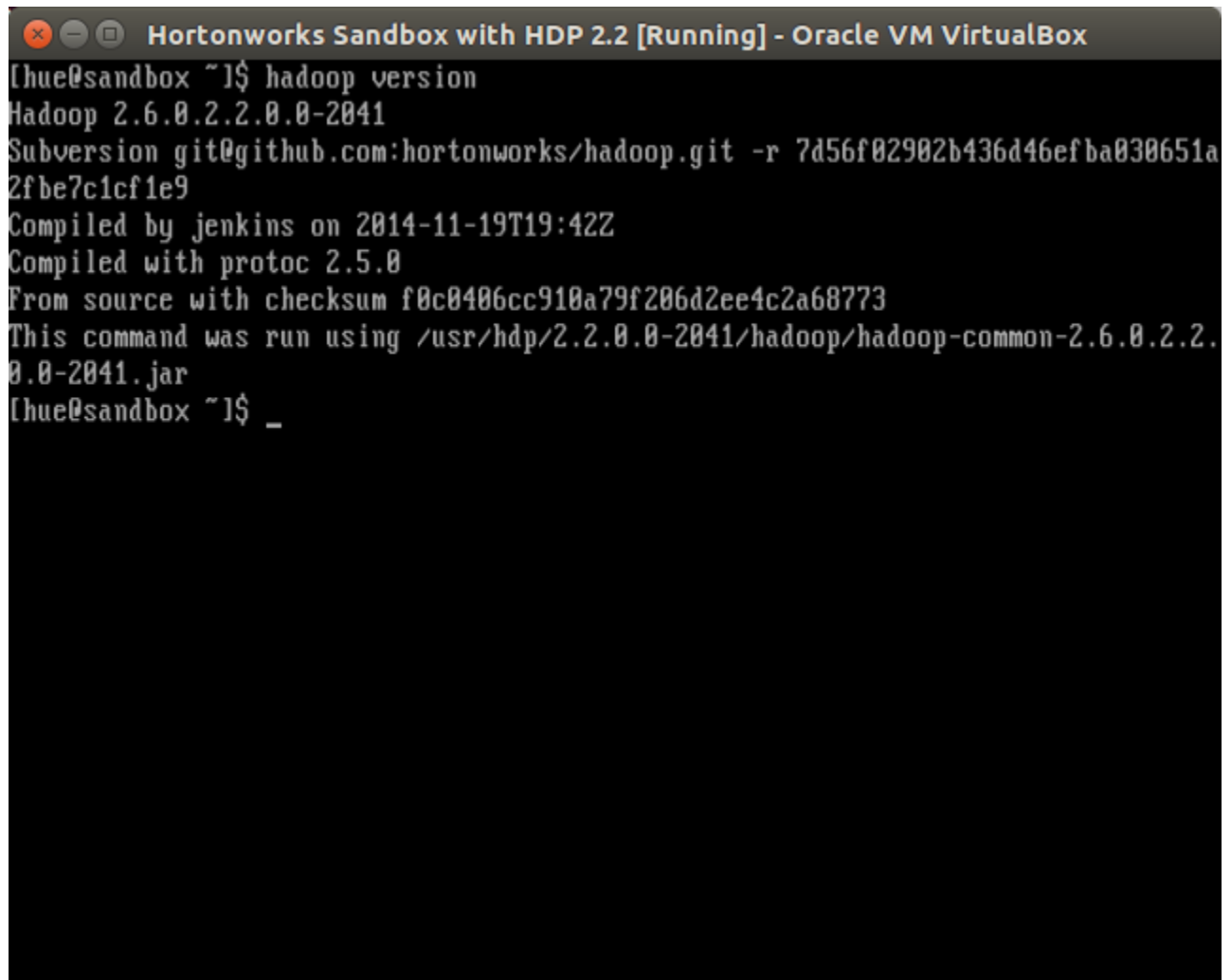
you can login using ssh or putty

ssh hue@192.168.1.109 -p 2222

Step 4:

Run hadoop commands

To test hadoop, run the following command to check hadoop version as shown below.

A screenshot of a terminal window titled "Hortonworks Sandbox with HDP 2.2 [Running] - Oracle VM VirtualBox". The terminal shows the command "hadoop version" being executed. The output displays the Hadoop version as 2.6.0.2.2.0.0-2041, the subversion as git@github.com:hortonworks/hadoop.git -r 7d56f02902b436d46efba030651a2fbe7c1cf1e9, and the compilation details: compiled by jenkins on 2014-11-19T19:42Z, compiled with protoc 2.5.0, and from source with checksum f0c0406cc910a79f206d2ee4c2a68773. It also states the command was run using /usr/hdp/2.2.0.0-2041/hadoop/hadoop-common-2.6.0.2.2.0.0-2041.jar. The prompt returns to the user's shell.

```
[hue@sandbox ~]$ hadoop version
Hadoop 2.6.0.2.2.0.0-2041
Subversion git@github.com:hortonworks/hadoop.git -r 7d56f02902b436d46efba030651a2fbe7c1cf1e9
Compiled by jenkins on 2014-11-19T19:42Z
Compiled with protoc 2.5.0
From source with checksum f0c0406cc910a79f206d2ee4c2a68773
This command was run using /usr/hdp/2.2.0.0-2041/hadoop/hadoop-common-2.6.0.2.2.0.0-2041.jar
[hue@sandbox ~]$ _
```

Step 5 Run hadoop wordcount program.

An example jar is in the hadoop distribution from apache but not in the hortonworks vm. We will need to download hadoop distribution from apache

At the VM terminal

→ **type**

wget <http://mirror.tcpdiag.net/apache/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz>

→ **untar the compressed file with**

tar xvzf hadoop-2.6.0.tar.gz.

(Note: we only need the example jar in this download.)

→ **Use the following command to create a input folder in hdfs.**

hdfs dfs -mkdir input

→ **Copy any txt file to the input folder, in this case 'Makefile' is the file used in this example**

hdfs dfs -put Makefile input

→ **Run the hadoop example jar**

hadoop jar hadoop-2.6.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.0.jar wordcount input output2

(note your output2 folder must not exist).

```
Hortonworks Sandbox with HDP 2.2 [Running] - Oracle VM VirtualBox
[hue@sandbox ~]$ hdfs dfs -rm -r -f output2
15/01/22 18:07:20 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 360 minutes, Empty interval = 0 minutes.
Moved: 'hdfs://sandbox.hortonworks.com:8020/user/hue/output2' to trash at: hdfs:
//sandbox.hortonworks.com:8020/user/hue/.Trash/Current
[hue@sandbox ~]$ hadoop jar hadoop-2.6.0/share/hadoop/mapreduce/hadoop-mapreduce
-examples-2.6.0.jar wordcount input output2
15/01/22 18:07:35 INFO impl.TimelineClientImpl: Timeline service address: http://
/sandbox.hortonworks.com:8188/ws/v1/timeline/
15/01/22 18:07:36 INFO client.RMProxy: Connecting to ResourceManager at sandbox.
hortonworks.com/192.168.1.109:8050
15/01/22 18:07:39 INFO input.FileInputFormat: Total input paths to process : 1
15/01/22 18:07:39 INFO mapreduce.JobSubmitter: number of splits:1
15/01/22 18:07:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
21948515836_0001
15/01/22 18:07:42 INFO impl.YarnClientImpl: Submitted application application_14
21948515836_0001
15/01/22 18:07:42 INFO mapreduce.Job: The url to track the job: http://sandbox.h
ortonworks.com:8088/proxy/application_1421948515836_0001/
15/01/22 18:07:42 INFO mapreduce.Job: Running job: job_1421948515836_0001
15/01/22 18:09:14 INFO mapreduce.Job: Job job_1421948515836_0001 running in uber
mode : false
15/01/22 18:09:14 INFO mapreduce.Job: map 0% reduce 0%
15/01/22 18:09:38 INFO mapreduce.Job: map 100% reduce 0%
```

See Mapreduce in action.

→ Show the output of the hadoop job use

*hdfs dfs -cat output2/**

```
Hortonworks Sandbox with HDP 2.2 [Running] - Oracle VM VirtualBox
types 1
typically 1
under 4
up 3
update 1
use 3
used 1
using 1
variables 2
various 1
virtual 3
virtual-bootstrap.py 1
virtual-env 6
virtual-env: 1
virtualenv 1
we 2
where 1
which 6
with 2
work 1
writing, 1
xml 1
you 2
i 12
[hue@sandbox ~]$ _
```

→ To rerun the hadoop job please remove 'output2' directory from hdfs using the command below

```
hdfs dfs -rm -r -f output2
```

→ To list the files in hdfs type

```
hdfs dfs -ls /user/hue
```

You can browse hortonworks web UI using


<http://192.168.1.109:8000> in your host browser.

Browser tabs: Zimbra: Co... Hortonwor... Linux_Down... VirtualBox/... VMware/W... About Hort... HDP 2.2 Jordi Burgo...


Address bar: 192.168.1.109:8000/about/ Search

Navigation: Configuration Check for misconfiguration Server Logs

Hortonworks Sandbox with HDP 2.2


[Leave Feedback](#)

Component	Version
Hue	2.6.1-2041
HDP	2.2.0
Hadoop	2.6.0
Pig	0.14.0
Hive-Hcatalog	0.14.0
Oozie	4.1.0
Ambari	1.7-169 <input type="button" value="Enable"/>
HBase	0.98.4
Knox	0.5.0
Storm	0.9.3

 Copyright © 2013 The Apache Software Foundation.
Apache Hadoop, Hadoop, HDFS, HBase, Hive, Mahout, Pig, Zookeeper are trademarks of the Apache Software Foundation.
Hue and the Hue logo are trademarks of Cloudera, Inc. and licensed under the Apache 2 license. For more information: gethue.com