LAB 1 INSTRUCTIONS

DS8003 – MGT OF BIG DATA AND TOOLS

Ryerson University

Instructor: Kanchana Padmanabhan

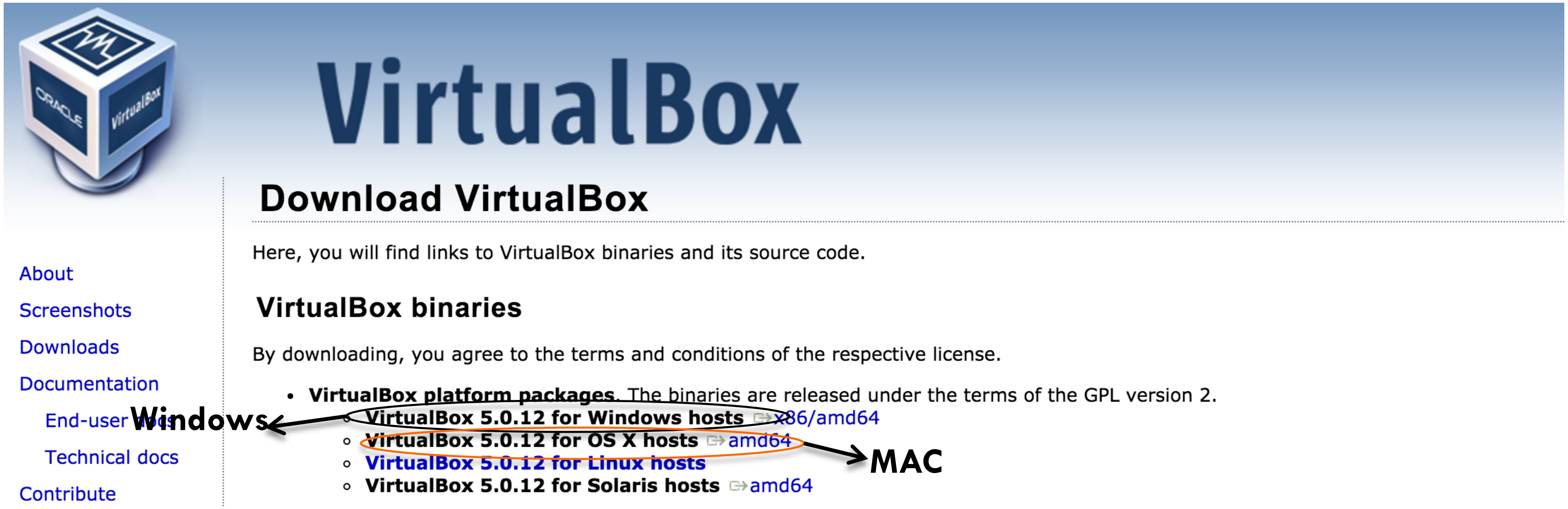
Lab & Assignments

- Lab Computer
  - Username: same as ryerson
  - Password: same as ryerson
- Lab 1
  - Virtualbox
  - Hadoop setup
  - SSH login

Lab 1 – Environment Setup

Install VirtualBox

- Download VirtualBox for your respective Operating System
  - https://www.virtualbox.org/wiki/Downloads



Download VirtualBox

Here, you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

- **VirtualBox platform packages**. The binaries are released under the terms of the GPL version 2.
  - **VirtualBox 5.0.12 for Windows hosts** x86/amd64
  - **VirtualBox 5.0.12 for OS X hosts** amd64
  - **VirtualBox 5.0.12 for Linux hosts**
  - **VirtualBox 5.0.12 for Solaris hosts** amd64

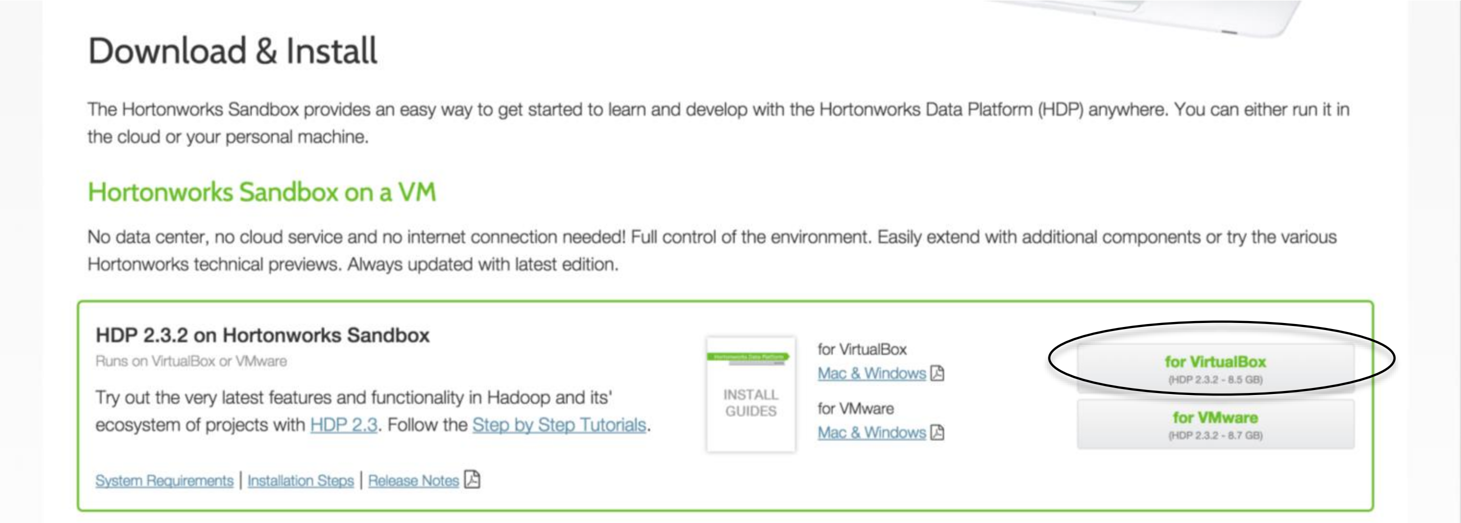
Windows

MAC

- For MAC – file has extension “.dmg”
- For Windows – file has extensiion “.exe”
- Double click on the downloaded file and follow instructions

Download Hortonworks HDP Sandbox

- □ Download HDP hadoop vm image: http://hortonworks.com/products/hortonworks-sandbox/
- □ The file will be called “HDP_2.3.2_virtualbox.ova” (or download the version available)



Download & Install

The Hortonworks Sandbox provides an easy way to get started to learn and develop with the Hortonworks Data Platform (HDP) anywhere. You can either run it in the cloud or your personal machine.

Hortonworks Sandbox on a VM

No data center, no cloud service and no internet connection needed! Full control of the environment. Easily extend with additional components or try the various Hortonworks technical previews. Always updated with latest edition.

HDP 2.3.2 on Hortonworks Sandbox

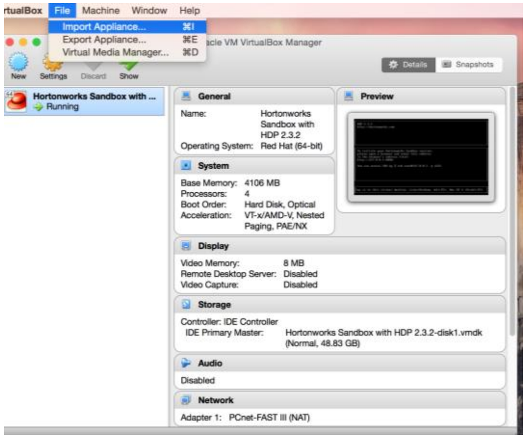
Runs on VirtualBox or VMware

Try out the very latest features and functionality in Hadoop and its' ecosystem of projects with HDP 2.3. Follow the Step by Step Tutorials.

System Requirements | Installation Steps | Release Notes

Install Hortonworks HDP Sandbox

- Sandbox installation tutorial for Windows: http://hortonworks.com/wp-content/uploads/unversioned/pdfs/InstallingHortonworksSandbox2onWindowsusingVB.pdf
- Double-click on the virtual box
- Import the file "HDP_2.3.2_virtualbox.ova" (or latest version your downloaded) into the Virtual box (Follow pictures below)



General
Name: Hortonworks Sandbox with HDP 2.3.2
Operating System: Red Hat (64-bit)

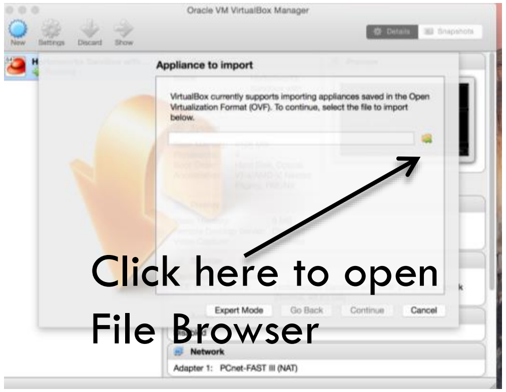
System
Base Memory: 4106 MB
Processors: 4
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Display
Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage
Controller: IDE Controller
IDE Primary Master: Hortonworks Sandbox with HDP 2.3.2-disk1.vmdk (Normal, 48.83 GB)

Audio
Disabled

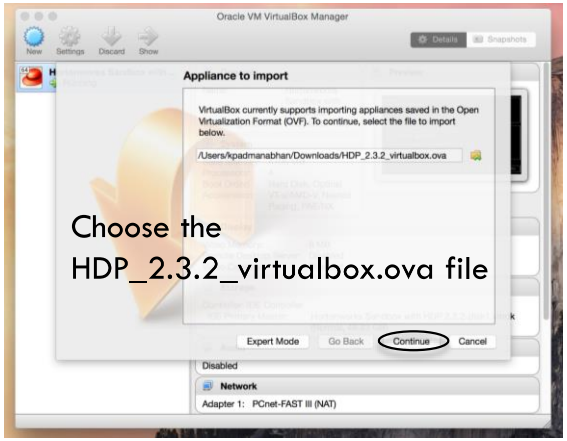
Network
Adapter 1: PCnet-FAST III (NAT)



Appliance to import

VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

Click here to open File Browser

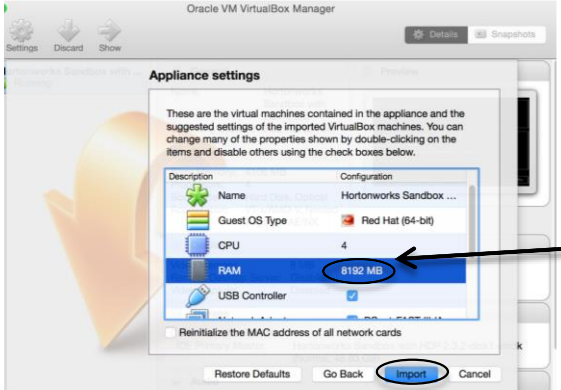


Appliance to import

VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

/Users/kpadmanabhan/Downloads/HDP_2.3.2_virtualbox.ova

Choose the HDP_2.3.2_virtualbox.ova file

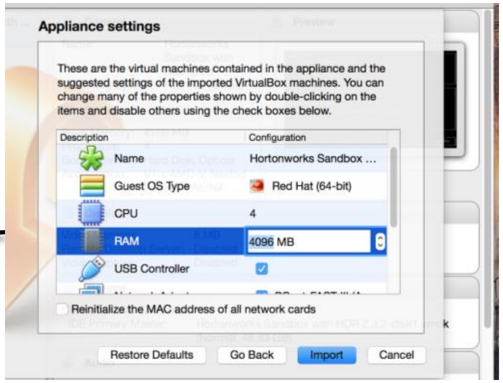


Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Description	Configuration
Name	Hortonworks Sandbox ...
Guest OS Type	Red Hat (64-bit)
CPU	4
RAM	8192 MB
USB Controller	☑

□ Reinitialize the MAC address of all network cards



Appliance settings

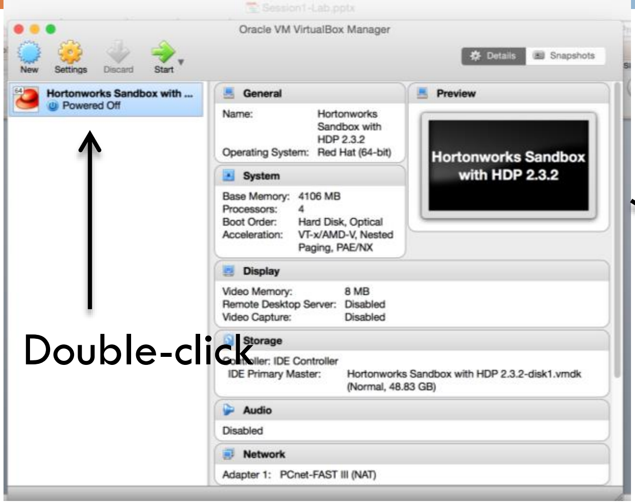
These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Description	Configuration
Name	Hortonworks Sandbox ...
Guest OS Type	Red Hat (64-bit)
CPU	4
RAM	4096 MB
USB Controller	☑

□ Reinitialize the MAC address of all network cards

You could change the RAM value to 4096 MB (4GB instead of 8 GB)

After Installation
Hortonworks HDP Sandbox



Hortonworks Sandbox with ...
Powered Off

Name: Hortonworks Sandbox with HDP 2.3.2
Operating System: Red Hat (64-bit)

Base Memory: 4106 MB
Processors: 4
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Hortonworks Sandbox with HDP 2.3.2

Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

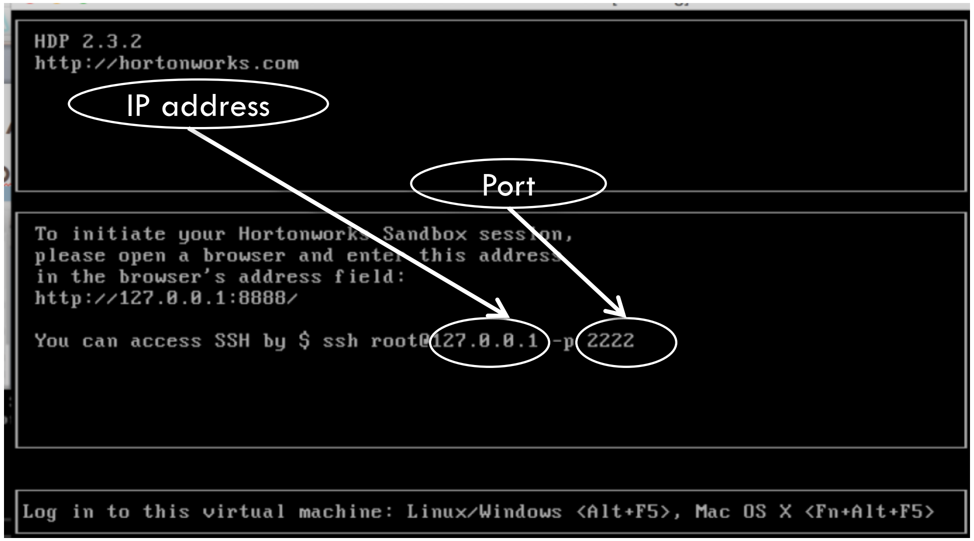
Controller: IDE Controller
IDE Primary Master: Hortonworks Sandbox with HDP 2.3.2-disk1.vmdk (Normal, 48.83 GB)

Disabled

Adapter 1: PCnet-FAST III (NAT)

Double-click

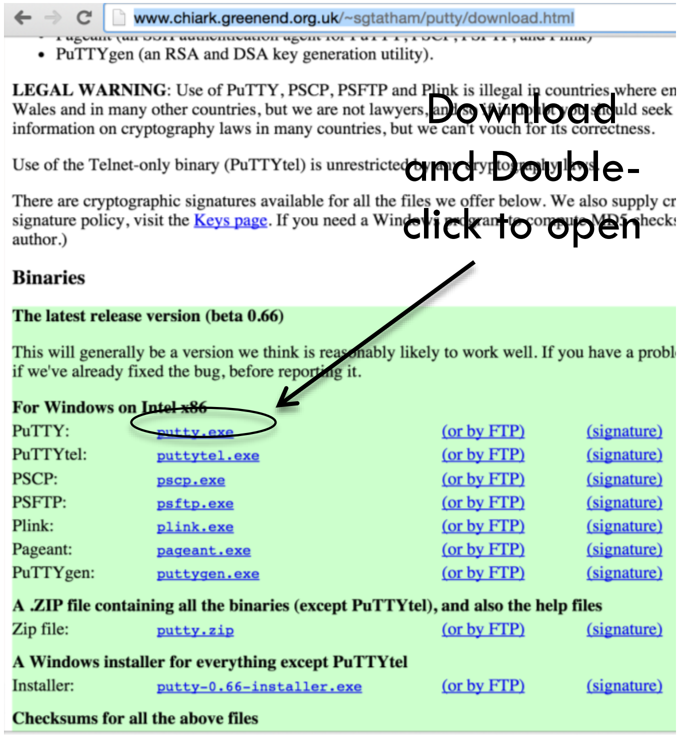
It will lead to the following screen



```
HDP 2.3.2
http://hortonworks.com

To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://127.0.0.1:8888/

You can access SSH by $ ssh root@127.0.0.1 -p 2222

Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Fn+Alt+F5>
```

SSH into the loaded Virtual Machine WINDOWS

- □ http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html



www.chiark.greenend.org.uk/~sgtatham/putty/download.html

- Pageant (an SSH authentication agent for PuTTY, PSCP, PSFTP, and Plink)
- PuTTYgen (an RSA and DSA key generation utility).

LEGAL WARNING: Use of PuTTY, PSCP, PSFTP and Plink is illegal in countries where en
Wales and in many other countries, but we are not lawyers, and so if in doubt you should seek
information on cryptography laws in many countries, but we can't vouch for its correctness.

Use of the Telnet-only binary (PuTTYtel) is unrestricted by any cryptography laws.

There are cryptographic signatures available for all the files we offer below. We also supply cr
signature policy, visit the Keys page. If you need a Windows program to compute MD5 checks
author.)

Binaries

The latest release version (beta 0.66)

This will generally be a version we think is reasonably likely to work well. If you have a probl
if we've already fixed the bug, before reporting it.

For Windows on Intel x86

PuTTY:	putty.exe	(or by FTP)	(signature)
PuTTYtel:	puttytel.exe	(or by FTP)	(signature)
PSCP:	pscp.exe	(or by FTP)	(signature)
PSFTP:	psftp.exe	(or by FTP)	(signature)
Plink:	plink.exe	(or by FTP)	(signature)
Pageant:	pageant.exe	(or by FTP)	(signature)
PuTTYgen:	puttygen.exe	(or by FTP)	(signature)

A .ZIP file containing all the binaries (except PuTTYtel), and also the help files

Zip file:	putty.zip	(or by FTP)	(signature)

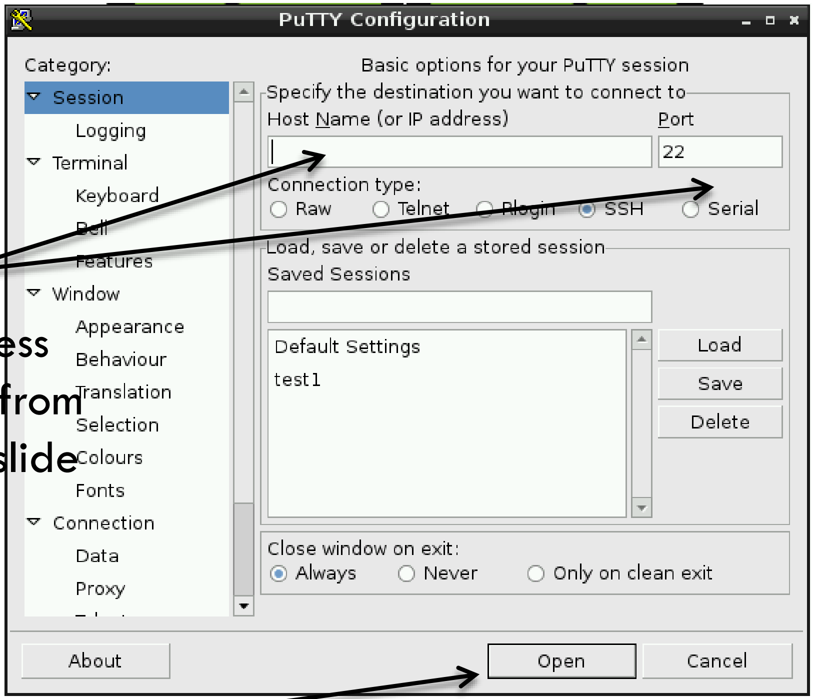
A Windows installer for everything except PuTTYtel

Installer:	putty-0.66-installer.exe	(or by FTP)	(signature)

Checksums for all the above files

Download and Double-click to open

IP address and port from previous slide



PuTTY Configuration

Category:

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address):

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions

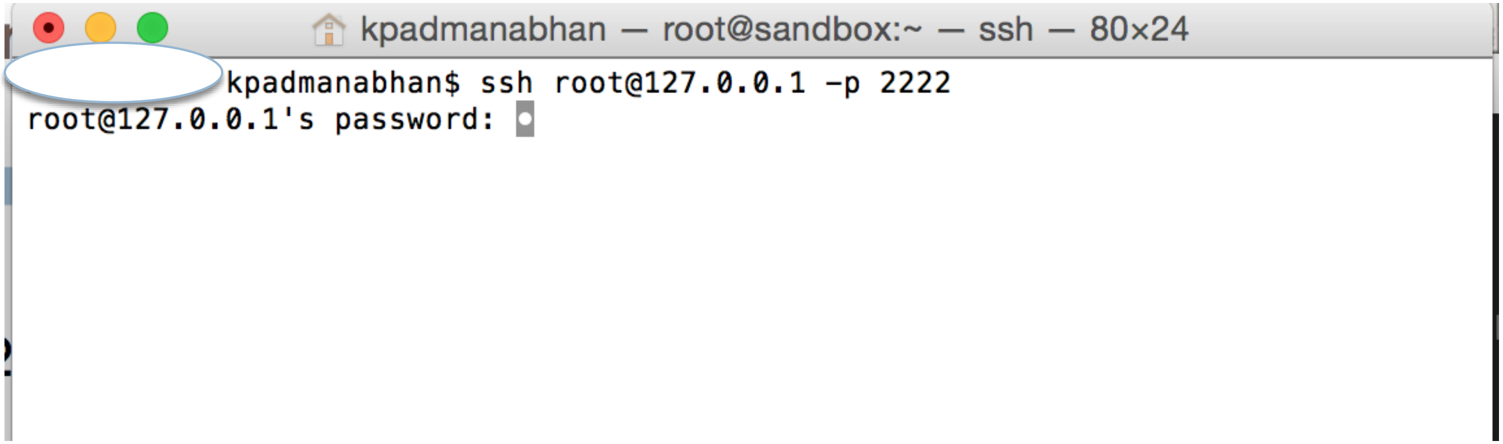
Default Settings
test1

Close window on exit: ● Always ○ Never ○ Only on clean exit

Click

SSH into the loaded Virtual Machine MAC

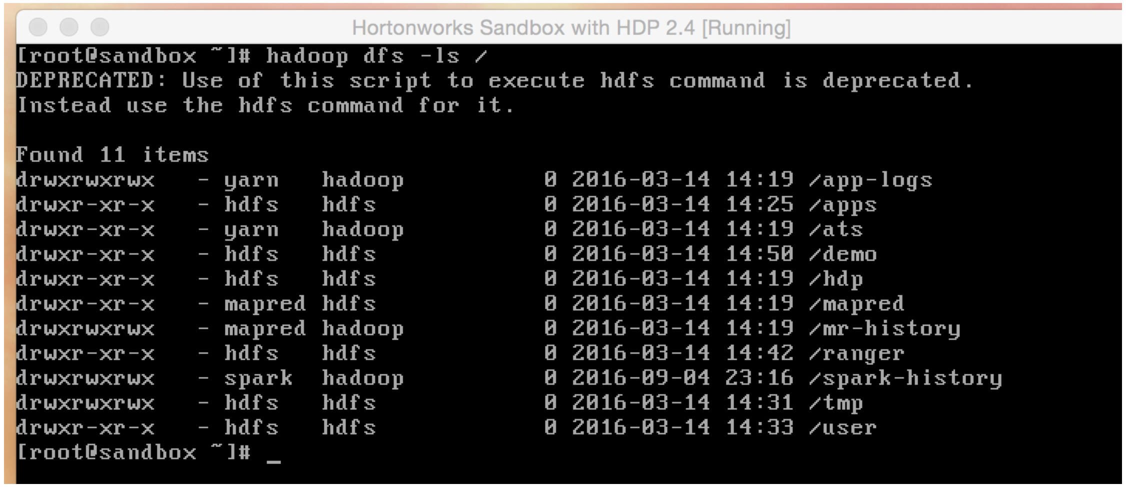
- Open Terminal
- Type “ssh root@127.0.0.1 –p 2222



```
kpadmanabhan$ ssh root@127.0.0.1 -p 2222
root@127.0.0.1's password:
```

Log On and Test HDFS

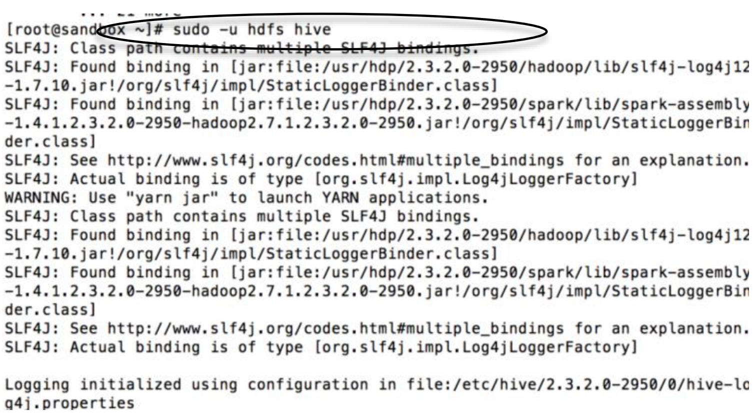
1. Login Info
   - Username: root
   - Password: hadoop
2. You will be asked to change your password after you login
3. Test HDFS



```
[root@sandbox ~]# hadoop dfs -ls /
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 11 items
drwxrwxrwx   - yarn    hadoop          0 2016-03-14 14:19 /app-logs
drwxr-xr-x   - hdfs    hdfs            0 2016-03-14 14:25 /apps
drwxr-xr-x   - yarn    hadoop          0 2016-03-14 14:19 /ats
drwxr-xr-x   - hdfs    hdfs            0 2016-03-14 14:50 /demo
drwxr-xr-x   - hdfs    hdfs            0 2016-03-14 14:19 /hdp
drwxr-xr-x   - mapred  hdfs            0 2016-03-14 14:19 /mapred
drwxrwxrwx   - mapred  hadoop          0 2016-03-14 14:19 /mr-history
drwxr-xr-x   - hdfs    hdfs            0 2016-03-14 14:42 /ranger
drwxrwxrwx   - spark   hadoop          0 2016-09-04 23:16 /spark-history
drwxrwxrwx   - hdfs    hdfs            0 2016-03-14 14:31 /tmp
drwxr-xr-x   - hdfs    hdfs            0 2016-03-14 14:33 /user
[root@sandbox ~]# _
```

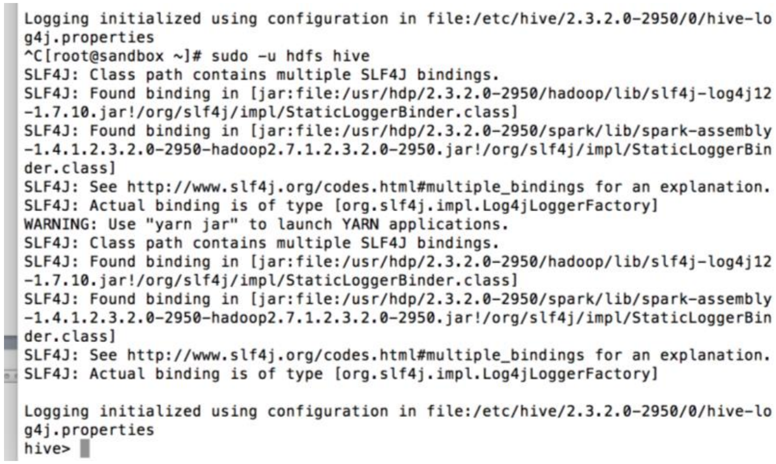
Test Hive

1. Test Hive



```
    ... 21 more
[root@sandbox ~]# sudo -u hdfs hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/hadoop/lib/slf4j-log4j12
-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/spark/lib/spark-assembly
-1.4.1.2.3.2.0-2950-hadoop2.7.1.2.3.2.0-2950.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
WARNING: Use "yarn jar" to launch YARN applications.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/hadoop/lib/slf4j-log4j12
-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/spark/lib/spark-assembly
-1.4.1.2.3.2.0-2950-hadoop2.7.1.2.3.2.0-2950.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

Logging initialized using configuration in file:/etc/hive/2.3.2.0-2950/0/hive-lo
g4j.properties
```

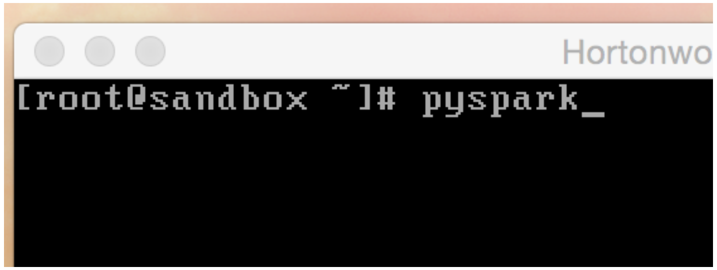
2. You will end up with a screen below



```
Logging initialized using configuration in file:/etc/hive/2.3.2.0-2950/0/hive-lo
g4j.properties
^C[root@sandbox ~]# sudo -u hdfs hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/hadoop/lib/slf4j-log4j12
-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/spark/lib/spark-assembly
-1.4.1.2.3.2.0-2950-hadoop2.7.1.2.3.2.0-2950.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
WARNING: Use "yarn jar" to launch YARN applications.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/hadoop/lib/slf4j-log4j12
-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.3.2.0-2950/spark/lib/spark-assembly
-1.4.1.2.3.2.0-2950-hadoop2.7.1.2.3.2.0-2950.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

Logging initialized using configuration in file:/etc/hive/2.3.2.0-2950/0/hive-lo
g4j.properties
hive> 
```

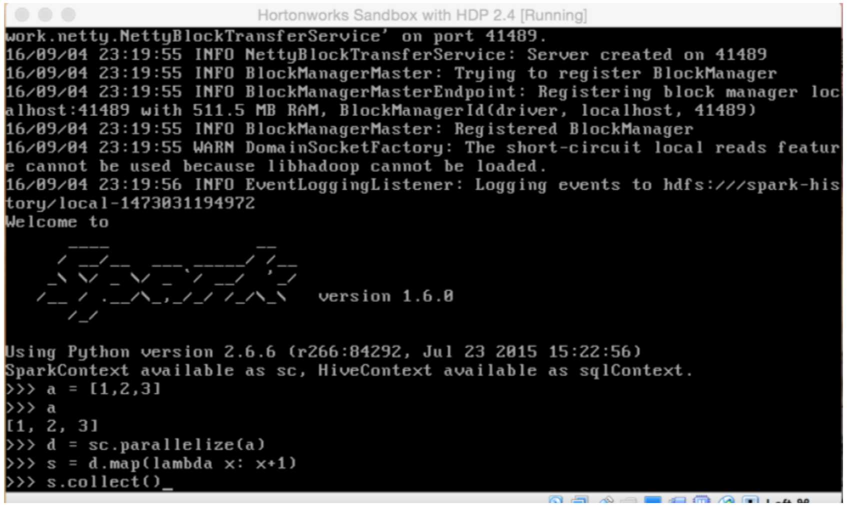
Test Spark

1. Test Spark



```
[root@sandbox ~]# pyspark_
```

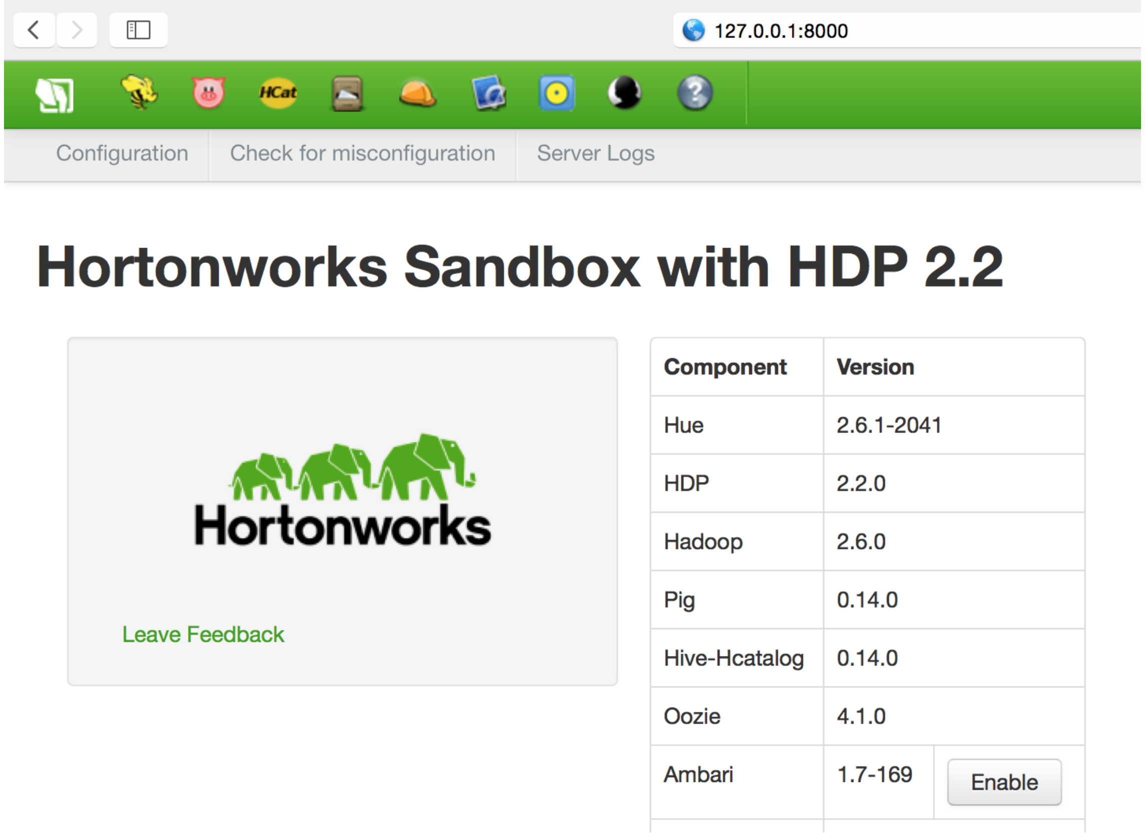
2. You will end up with a screen below



```
work.netty.NettyBlockTransferService' on port 41489.
16/09/04 23:19:55 INFO NettyBlockTransferService: Server created on 41489
16/09/04 23:19:55 INFO BlockManagerMaster: Trying to register BlockManager
16/09/04 23:19:55 INFO BlockManagerMasterEndpoint: Registering block manager loc
alhost:41489 with 511.5 MB RAM, BlockManagerId(driver, localhost, 41489)
16/09/04 23:19:55 INFO BlockManagerMaster: Registered BlockManager
16/09/04 23:19:55 WARN DomainSocketFactory: The short-circuit local reads featur
e cannot be used because libhadoop cannot be loaded.
16/09/04 23:19:56 INFO EventLoggingListener: Logging events to hdfs:///spark-his
tory/local-1473031194972
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkContext available as sc, HiveContext available as sqlContext.
>>> a = [1,2,3]
>>> a
[1, 2, 3]
>>> d = sc.parallelize(a)
>>> s = d.map(lambda x: x+1)
>>> s.collect()_
```

Access Hadoop via Browser Using Hue

- Access Hue via Browser
  - http://127.0.0.1:8000

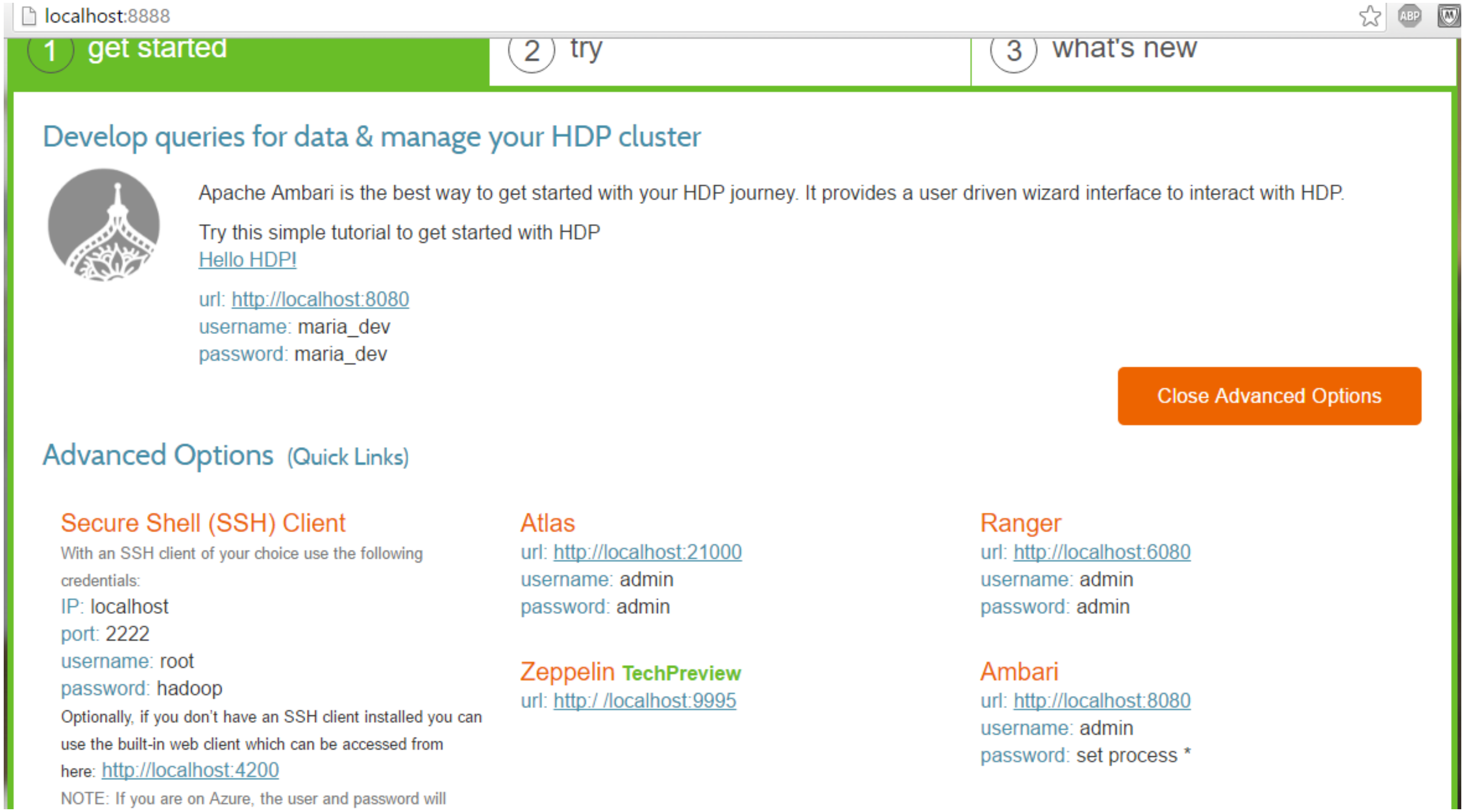


Hortonworks Sandbox with HDP 2.2

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

Welcome Message and Username Information

Go to http://localhost:8888 for welcome screen and Ambari username and password
Click on “Advanced Options” to get SSH and other information



Develop queries for data & manage your HDP cluster

Apache Ambari is the best way to get started with your HDP journey. It provides a user driven wizard interface to interact with HDP.

Try this simple tutorial to get started with HDP
Hello HDP!

url: http://localhost:8080
username: maria_dev
password: maria_dev

Advanced Options (Quick Links)

Secure Shell (SSH) Client
With an SSH client of your choice use the following credentials:
IP: localhost
port: 2222
username: root
password: hadoop
Optionally, if you don't have an SSH client installed you can use the built-in web client which can be accessed from here: http://localhost:4200
NOTE: If you are on Azure, the user and password will

Atlas
url: http://localhost:21000
username: admin
password: admin

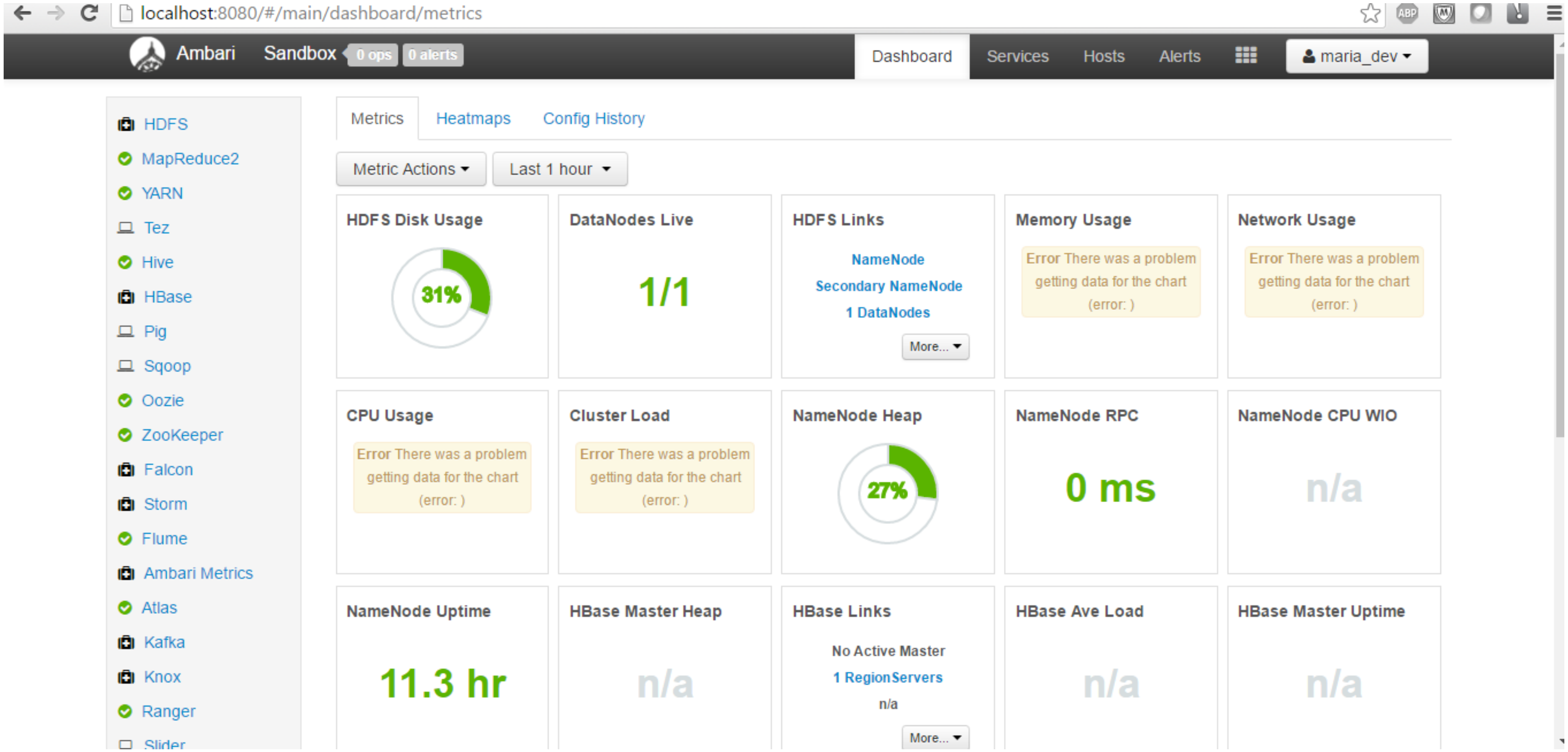
Zeppelin TechPreview
url: http:/ /localhost:9995

Ranger
url: http://localhost:6080
username: admin
password: admin

Ambari
url: http://localhost:8080
username: admin
password: set process *

Browse Through System Setup using Ambari

Go to http://localhost:8080 for Ambari



localhost:8080/#/main/dashboard/metrics

HDFS Disk Usage: 31%

DataNodes Live: 1/1

HDFS Links: NameNode, Secondary NameNode, 1 DataNodes

Memory Usage: Error There was a problem getting data for the chart (error:)

Network Usage: Error There was a problem getting data for the chart (error:)

CPU Usage: Error There was a problem getting data for the chart (error:)

Cluster Load: Error There was a problem getting data for the chart (error:)

NameNode Heap: 27%

NameNode RPC: 0 ms

NameNode CPU WIO: n/a

NameNode Uptime: 11.3 hr

HBase Master Heap: n/a

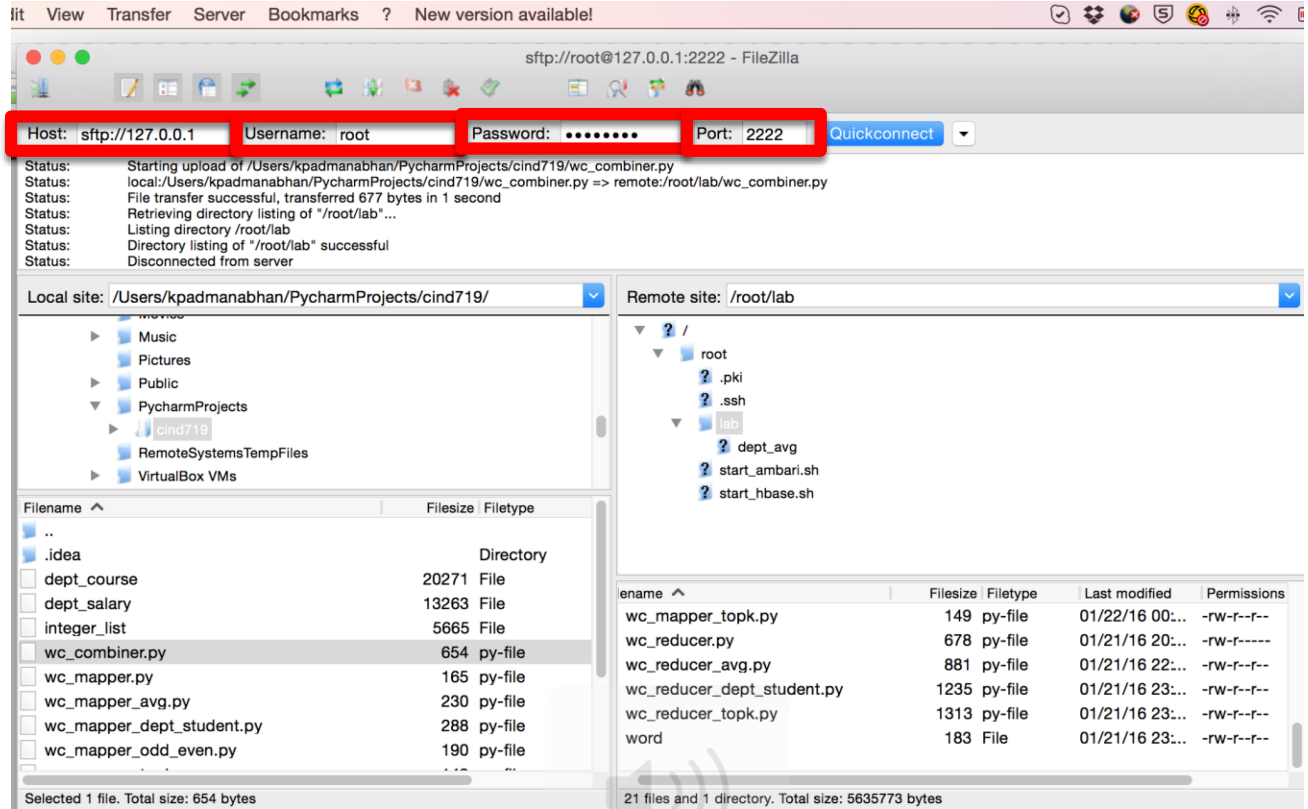
HBase Links: No Active Master, 1 RegionServers, n/a

HBase Ave Load: n/a

HBase Master Uptime: n/a

Connect to Virtual box using Filezilla

- □ Open Filezilla
- □ Enter Host: sftp://127.0.0.1 or sftp://localhost
- □ Port: 2222
- □ Username & Password – same as your virtual box
- □ Click QuickConnect



Host: sftp://127.0.0.1 Username: root Password: •••••••• Port: 2222

```
Status:     Starting upload of /Users/kpadmanabhan/PycharmProjects/cind719/wc_combiner.py
Status:     local:/Users/kpadmanabhan/PycharmProjects/cind719/wc_combiner.py => remote:/root/lab/wc_combiner.py
Status:     File transfer successful, transferred 677 bytes in 1 second
Status:     Retrieving directory listing of "/root/lab"...
Status:     Listing directory /root/lab
Status:     Directory listing of "/root/lab" successful
Status:     Disconnected from server
```

Local site: /Users/kpadmanabhan/PycharmProjects/cind719/

Remote site: /root/lab

Filename	Filesize	Filetype
..		
.idea		Directory
dept_course	20271	File
dept_salary	13263	File
integer_list	5665	File
wc_combiner.py	654	py-file
wc_mapper.py	165	py-file
wc_mapper_avg.py	230	py-file
wc_mapper_dept_student.py	288	py-file
wc_mapper_odd_even.py	190	py-file

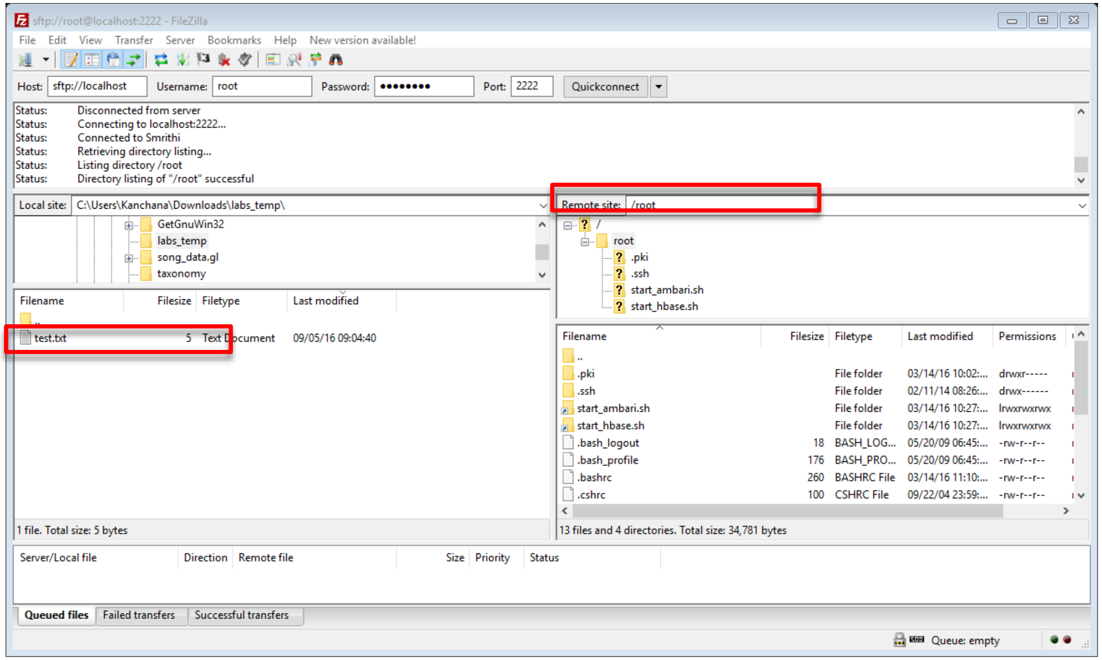
Selected 1 file. Total size: 654 bytes

ename	Filesize	Filetype	Last modified	Permissions
wc_mapper_topk.py	149	py-file	01/22/16 00:...	-rw-r--r--
wc_reducer.py	678	py-file	01/21/16 20:...	-rw-r-----
wc_reducer_avg.py	881	py-file	01/21/16 22:...	-rw-r--r--
wc_reducer_dept_student.py	1235	py-file	01/21/16 23:...	-rw-r--r--
wc_reducer_topk.py	1313	py-file	01/21/16 23:...	-rw-r--r--
word	183	File	01/21/16 23:...	-rw-r--r--

21 files and 1 directory. Total size: 5635773 bytes

Upload Files to the VirtualBox

- After connecting to the Sandbox access node in Filezilla…
  - The left side shows directories of your local computer (WINDOWS COMPUTER IF USING LAB MACHINE)
  - The right side box shows directories of your remote machine on Linux
    - In this case the HDP Sandbox (Virtual Box)
- Create a text file called test.txt on your machine with the numbers 1,2,3 written inside
- Upload test.txt to Sandbox
  - On the right-side box, navigate to /root
  - On the left-side box, navigate and find the test.txt file you downloaded
  - Drag and drop "test.txt" into /root/ on the right to the Sandbox



sftp://root@localhost:2222 - FileZilla
File Edit View Transfer Server Bookmarks Help New version available!
Quickconnect

Host: sftp://localhost Username: root Password: •••••••• Port: 2222

```
Status:	Disconnected from server
Status:	Connecting to localhost:2222...
Status:	Connected to Smrithi
Status:	Retrieving directory listing...
Status:	Listing directory /root
Status:	Directory listing of "/root" successful
```

Local site: C:\Users\Kanchana\Downloads\labs_temp\

GetGnuWin32
labs_temp
song_data.gl
taxonomy

Filename	Filesize	Filetype	Last modified
..			
test.txt	5	Text Document	09/05/16 09:04:40

1 file. Total size: 5 bytes

Remote site: /root

/
root
.pki
.ssh
start_ambari.sh
start_hbase.sh

Filename	Filesize	Filetype	Last modified	Permissions
..				
.pki		File folder	03/14/16 10:02:...	drwxr-----
.ssh		File folder	02/11/14 08:26:...	drwx------
start_ambari.sh		File folder	03/14/16 10:27:...	lrwxrwxrwx
start_hbase.sh		File folder	03/14/16 10:27:...	lrwxrwxrwx
.bash_logout	18	BASH_LOG...	05/20/09 06:45:...	-rw-r--r--
.bash_profile	176	BASH_PRO...	05/20/09 06:45:...	-rw-r--r--
.bashrc	260	BASHRC File	03/14/16 11:10:...	-rw-r--r--
.cshrc	100	CSHRC File	09/22/04 23:59:...	-rw-r--r--

13 files and 4 directories. Total size: 34,781 bytes

Linux Command Line

Try out the following tutorial

- Playing around with big data tools becomes easier with linux command line
- From http://www.ee.surrey.ac.uk/Teaching/Unix/
- Try out Tutorial Sections 1, 2, 3, 4, 5 (Section 5.5), and 6

Python

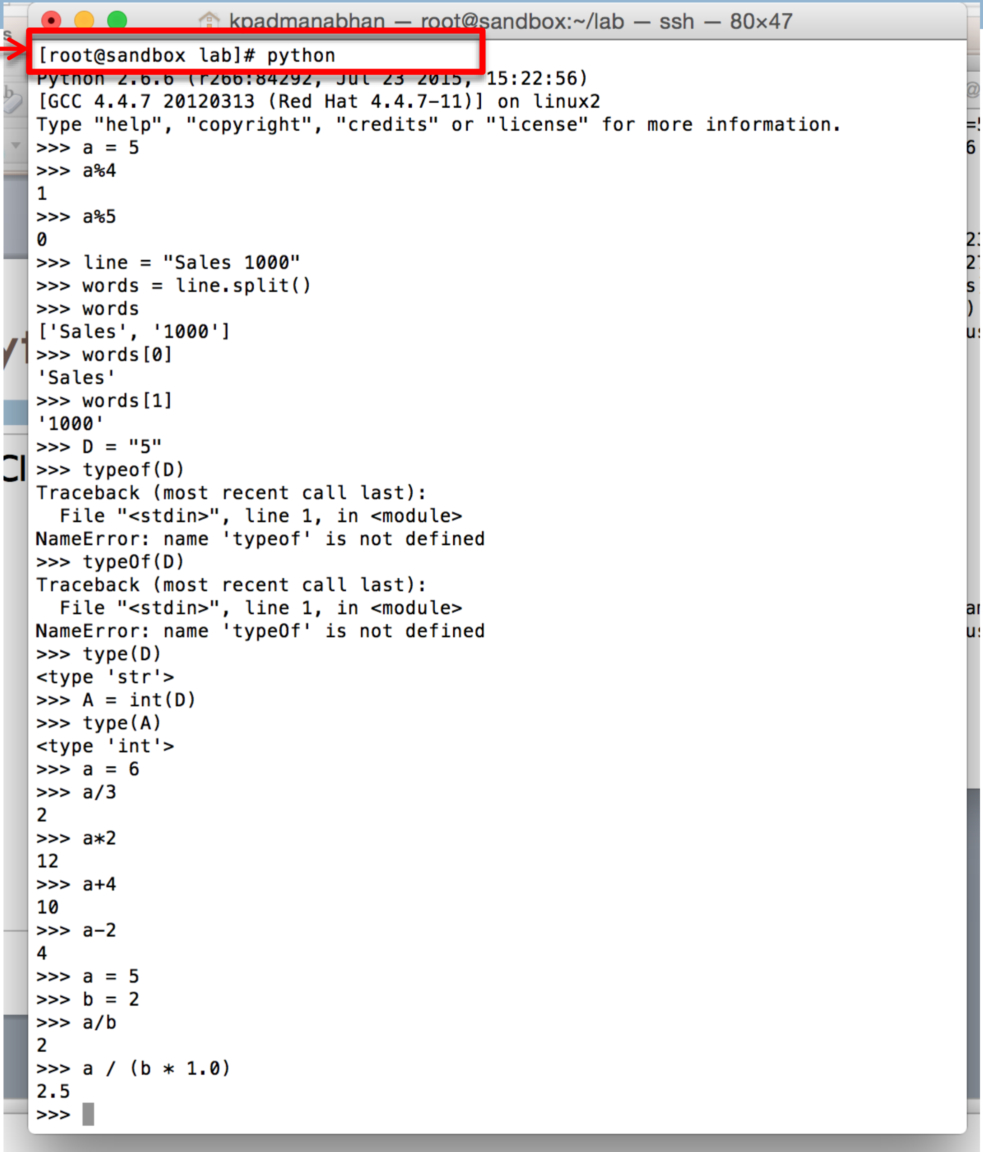
Practice Python Commands

"python" command opens python shell where we can try out some commands. Similar to the "R" shell

1. a = 5; assigning the number 5 to variable a
2. a%5; is math modulo operator; it will give the value of the reminder when a is divided by 5
3. line = "Sales 1000"; assigns the string to variable
4. line.split(); Splits the string into multiple strings; Uses "space" to decide where to split
5. words = line.split(); splits the string and assigns to words
6. A = int(D); Convert string "5" to number 5
7. type(D); outputs type of D; "string" or "int" (integer)
8. +, -, *, / - same mathematical operators
9. Notice difference between a/b and a/ (b * 1.0)

http://thepythonguru.com/getting-started-with-python/

http://www.afterhoursprogramming.com/tutorial/Python/Introduction/



```
[root@sandbox lab]# python
Python 2.6.6 (r266:84292, Jul 23 2015, 15:22:56)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-11)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> a = 5
>>> a%4
1
>>> a%5
0
>>> line = "Sales 1000"
>>> words = line.split()
>>> words
['Sales', '1000']
>>> words[0]
'Sales'
>>> words[1]
'1000'
>>> D = "5"
>>> typeof(D)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'typeof' is not defined
>>> typeOf(D)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'typeOf' is not defined
>>> type(D)
<type 'str'>
>>> A = int(D)
>>> type(A)
<type 'int'>
>>> a = 6
>>> a/3
2
>>> a*2
12
>>> a+4
10
>>> a-2
4
>>> a = 5
>>> b = 2
>>> a/b
2
>>> a / (b * 1.0)
2.5
>>> 
```

Writing a python script

- Copy the following piece of code into ***test.py (Keep track of indentation)***

```
import sys
import math
def returnSquare (x):
    return x**2

def main(a):
    print returnSquare(int(a))

if __name__ == "__main__":
    if len(sys.argv) >= 2:
        try:
            main(sys.argv[1])
        except:
            print "Not an Integer"
```

- ***Execute: python test.py 2***
- ***Output: 4***

Try yourself

- Write a python script that will read file shakespere_100.txt and print first 100 lines
- Write a python script that can take a as input from command line and print “even” if even and “odd” if odd
- Write a python script that can take two numbers a & b as input, if a is less than b then calculate a divided by b and if a > b calculate a * b , and print the result rounded to 2 digits