1. Getting the Environment Ready doing R Analytics on Hadoop

In the section, students shall be introduced to the basic platforms and setup which should be installed to be able to do R analytics using Big Data. The three main setups include:

1. The Virtual Machine (VM) (VMWare Workstation player)
2. Hadoop Environment (Hortonworks)
3. R/RStudio on Hadoop

2. Installing the Virtual Machine

VMWare Workstation player (non-commercial) edition has been used. Follow the following steps to download and install the software.

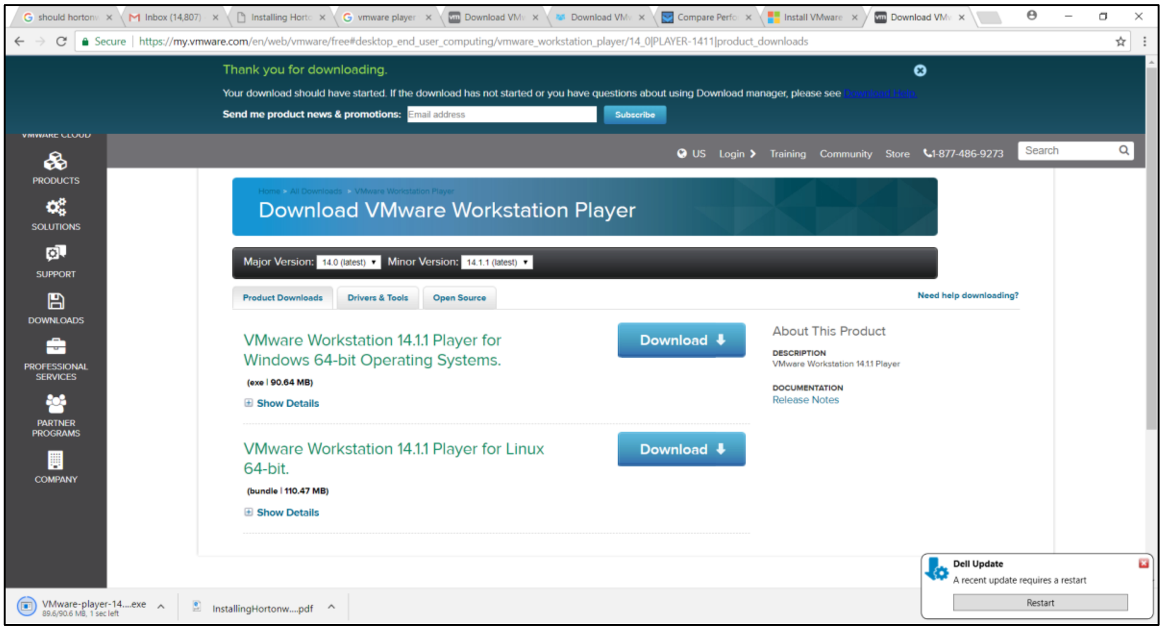
Note it!

- *There might be slight variations when you download and install depending on the version that you are using and the OS that you have.*

Steps

1. Download VMWare Workstation player by searching on the web.

VMware Player on Windows



https://my.vmware.com/en/web/vmware/free#desktop_end_user_computing/vmware_workstation_player/14_0|PLAYER-1411|product_downloads

Thank you for downloading.

Your download should have started. If the download has not started or you have questions about using Download manager, please see Download Help.

Send me product news & promotions: Email address

Download VMware Workstation Player

Major Version: 14.0 (latest) Minor Version: 14.1.1 (latest)

VMware Workstation 14.1.1 Player for Windows 64-bit Operating Systems.
(exe | 90.64 MB)

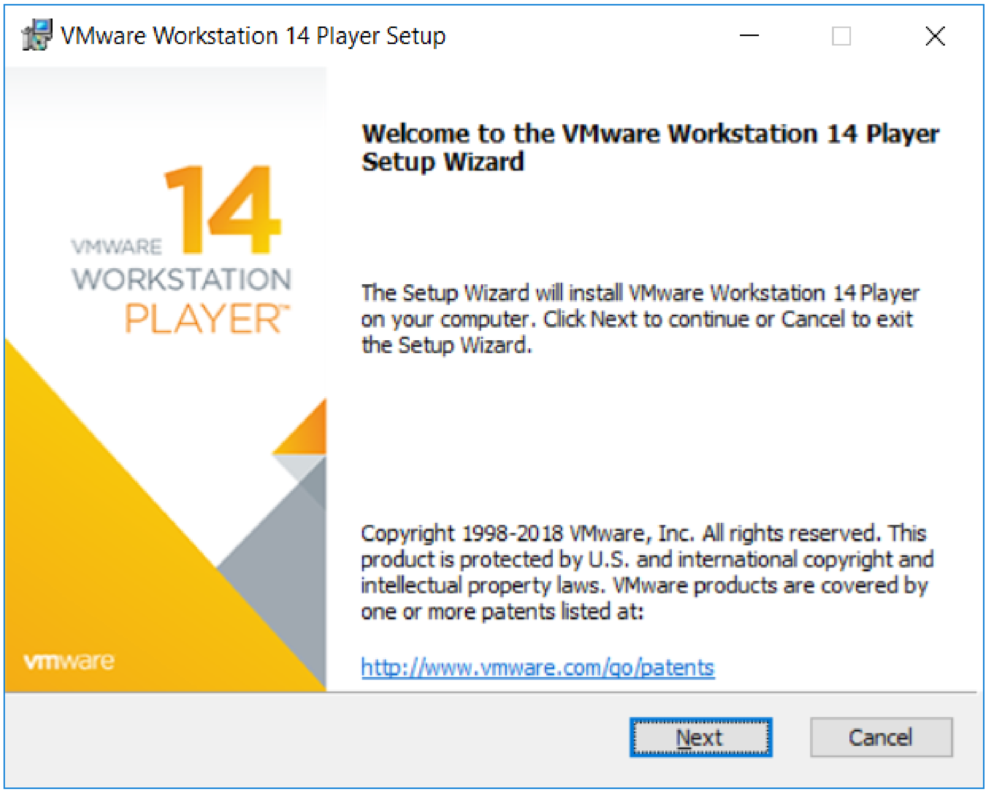
VMware Workstation 14.1.1 Player for Linux 64-bit.
(bundle | 110.47 MB)

About This Product
DESCRIPTION
VMware Workstation 14.1.1 Player
DOCUMENTATION
Release Notes

Dell Update
A recent update requires a restart

VMware-player-14....exe 89.6/90.6 MB, 1 sec left
InstallingHortonw....pdf

2. Install once download is complete. Click on Next and Accept the recommended settings.



VMware Workstation 14 Player Setup

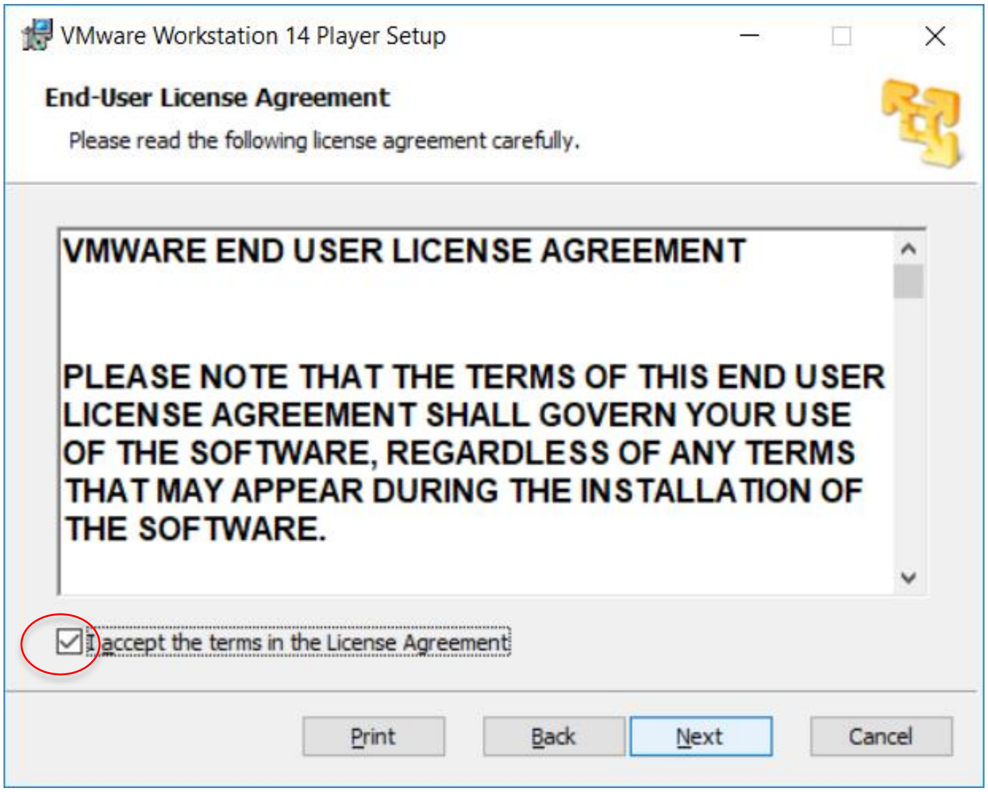
Welcome to the VMware Workstation 14 Player Setup Wizard

The Setup Wizard will install VMware Workstation 14 Player on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Copyright 1998-2018 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at:

http://www.vmware.com/go/patents

3. Accept the License agreement by clicking on the option.



VMware Workstation 14 Player Setup

End-User License Agreement

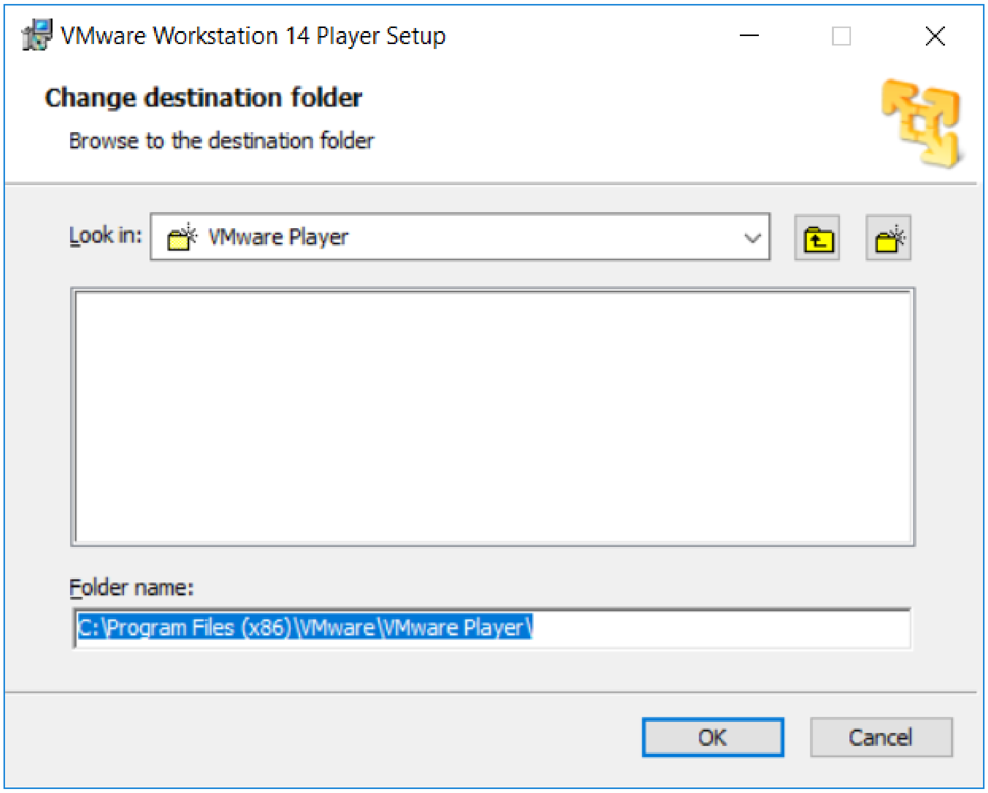
Please read the following license agreement carefully.

VMWARE END USER LICENSE AGREEMENT

PLEASE NOTE THAT THE TERMS OF THIS END USER LICENSE AGREEMENT SHALL GOVERN YOUR USE OF THE SOFTWARE, REGARDLESS OF ANY TERMS THAT MAY APPEAR DURING THE INSTALLATION OF THE SOFTWARE.

☑ I accept the terms in the License Agreement

4. Select folder to install and Click on “Ok”



VMware Workstation 14 Player Setup

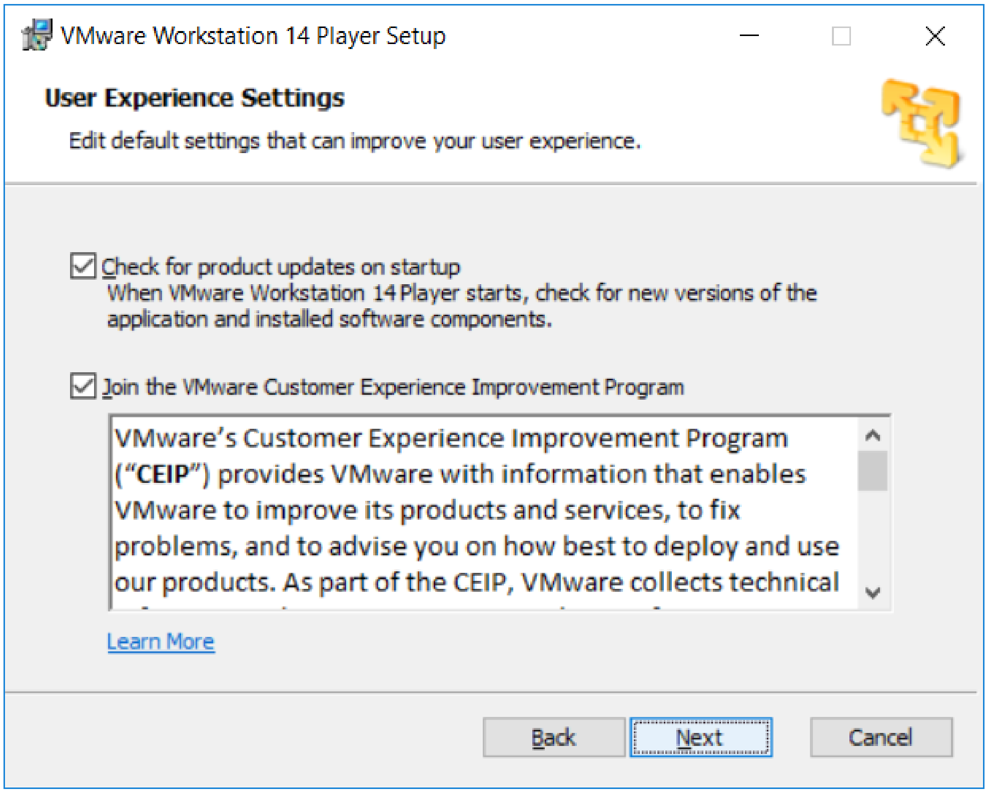
Change destination folder

Browse to the destination folder

Look in: VMware Player

Folder name: C:\Program Files (x86)\VMware\VMware Player\

5. Continue to follow the instructions.



VMware Workstation 14 Player Setup

User Experience Settings

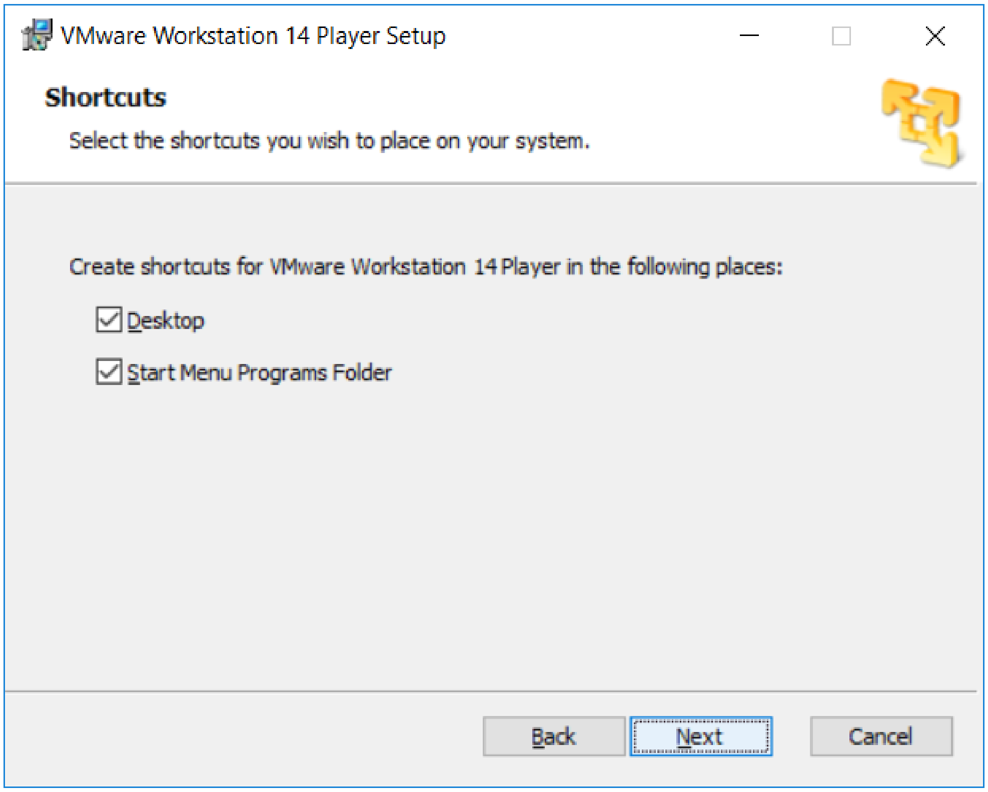
Edit default settings that can improve your user experience.

☑ Check for product updates on startup
When VMware Workstation 14 Player starts, check for new versions of the application and installed software components.

☑ Join the VMware Customer Experience Improvement Program

VMware's Customer Experience Improvement Program ("**CEIP**") provides VMware with information that enables VMware to improve its products and services, to fix problems, and to advise you on how best to deploy and use our products. As part of the CEIP, VMware collects technical

5. Select "Next".



VMware Workstation 14 Player Setup

Shortcuts

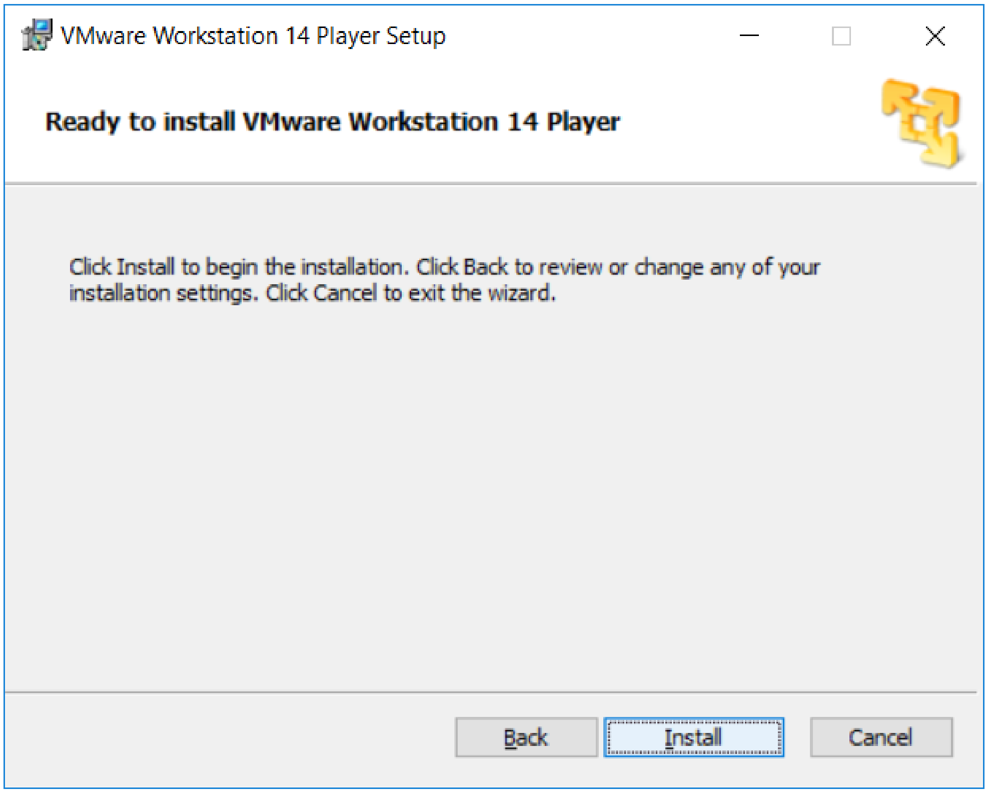
Select the shortcuts you wish to place on your system.

Create shortcuts for VMware Workstation 14 Player in the following places:

☑ Desktop

☑ Start Menu Programs Folder

6. Click “Install”.

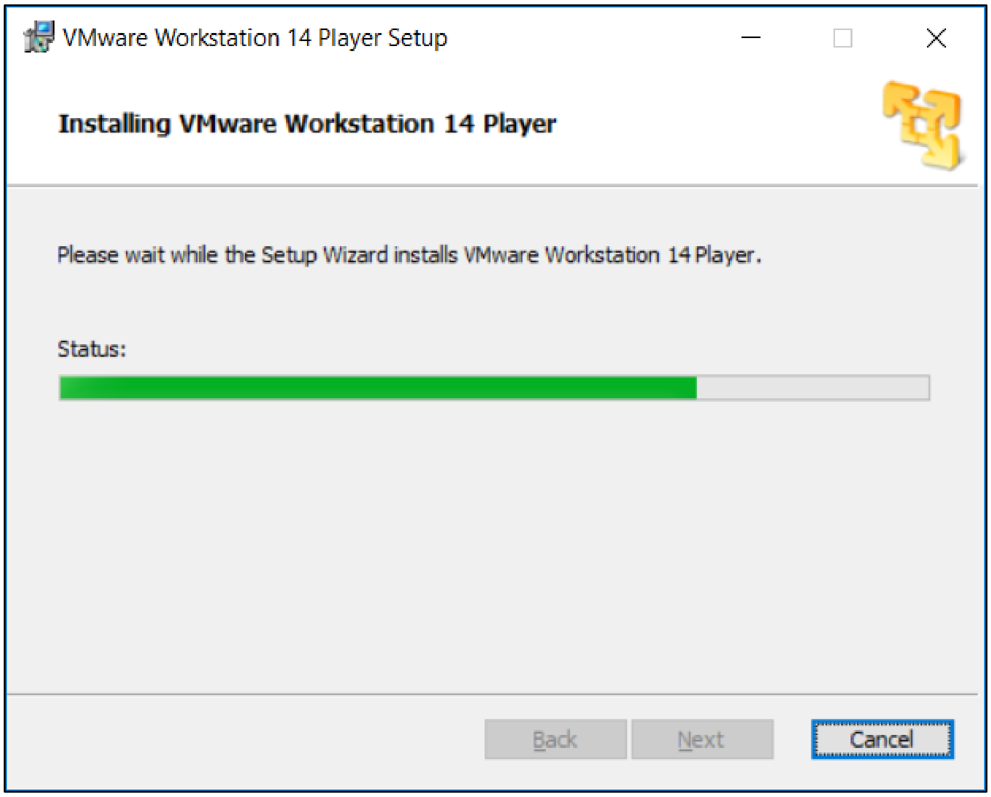


VMware Workstation 14 Player Setup

Ready to install VMware Workstation 14 Player

Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.

7. Wait for the progress window.



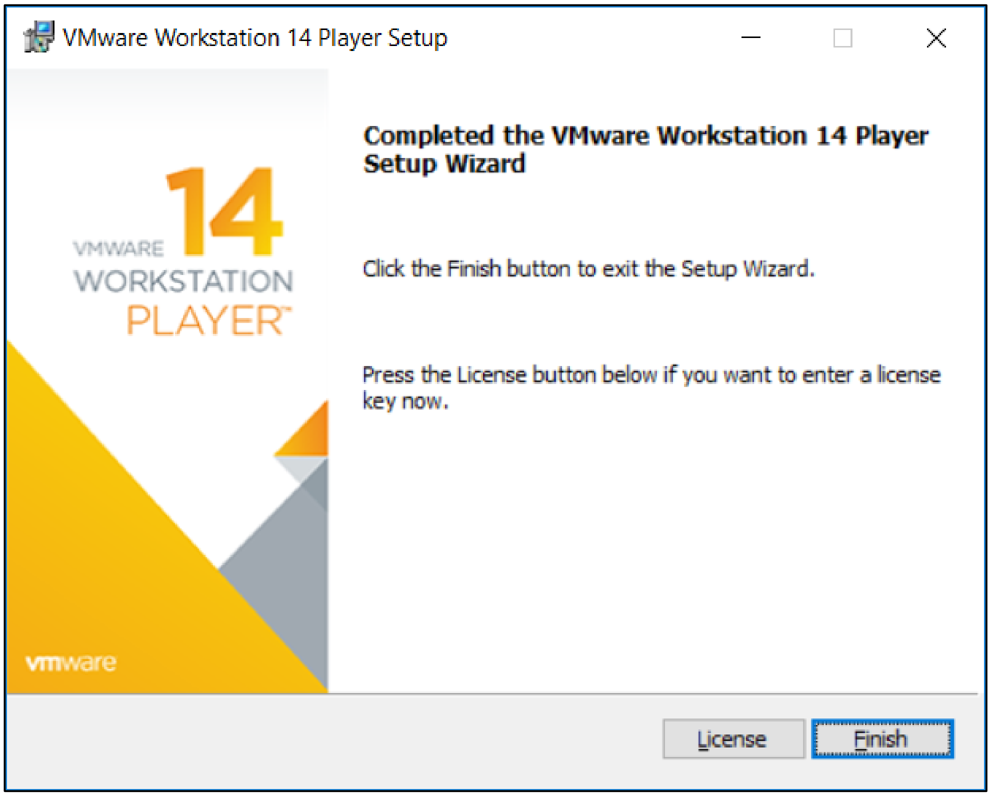
VMware Workstation 14 Player Setup

Installing VMware Workstation 14 Player

Please wait while the Setup Wizard installs VMware Workstation 14 Player.

Status:

8. Click “Finish”.



VMware Workstation 14 Player Setup

Completed the VMware Workstation 14 Player Setup Wizard

Click the Finish button to exit the Setup Wizard.

Press the License button below if you want to enter a license key now.

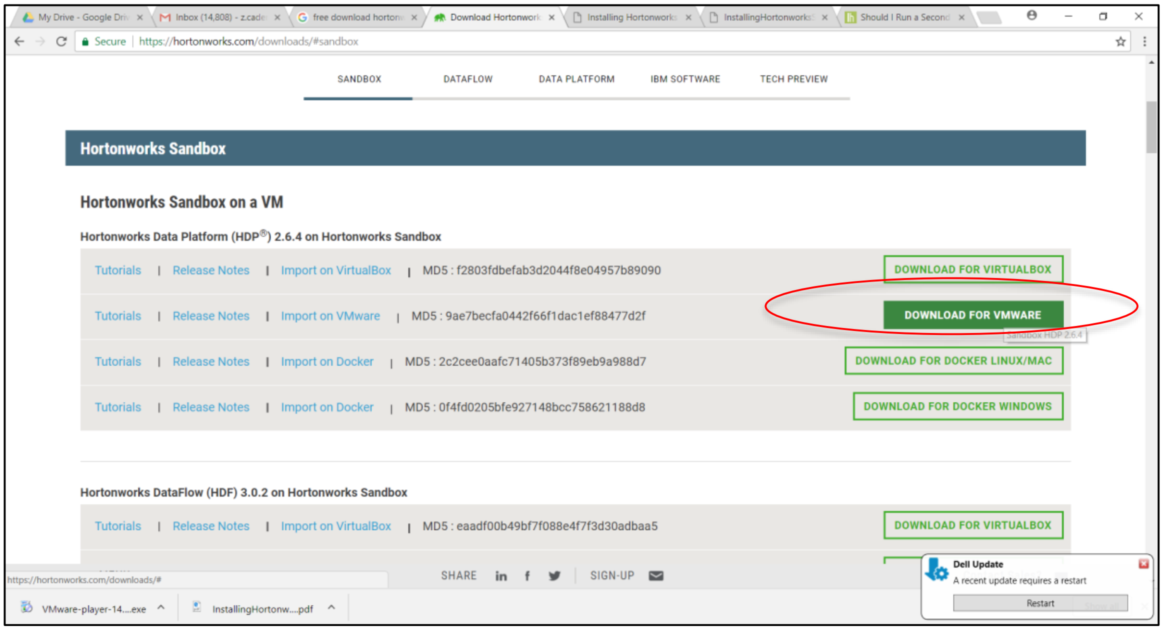
Your VM should now be ready and you can proceed with the next setup to install your Hadoop environment.

3. Installing the Hadoop Environment

Hortonworks Sandbox has been chosen as the Big Data Platform. Follow the following steps to get Hortonworks ready for VMWare.

Steps

1. Download Hortonworks Sandbox for VMWare.



Hortonworks Sandbox

Hortonworks Sandbox on a VM

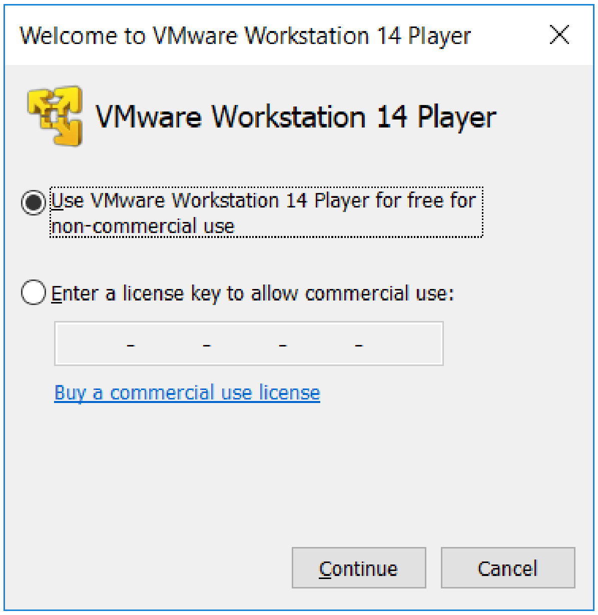
Hortonworks Data Platform (HDP®) 2.6.4 on Hortonworks Sandbox

Tutorials	Release Notes	Import on VirtualBox	MD5 : f2803fdbefab3d2044f8e04957b89090	DOWNLOAD FOR VIRTUALBOX
Tutorials	Release Notes	Import on VMware	MD5 : 9ae7becfa0442f66f1dac1ef88477d2f	DOWNLOAD FOR VMWARE
Tutorials	Release Notes	Import on Docker	MD5 : 2c2cee0aafc71405b373f89eb9a988d7	DOWNLOAD FOR DOCKER LINUX/MAC
Tutorials	Release Notes	Import on Docker	MD5 : 0f4fd0205bfe927148bcc758621188d8	DOWNLOAD FOR DOCKER WINDOWS

Hortonworks DataFlow (HDF) 3.0.2 on Hortonworks Sandbox

Tutorials	Release Notes	Import on VirtualBox	MD5 : eaadf00b49bf7f088e4f7f3d30adbaa5	DOWNLOAD FOR VIRTUALBOX

2. Choose for free and click on “Continue”.

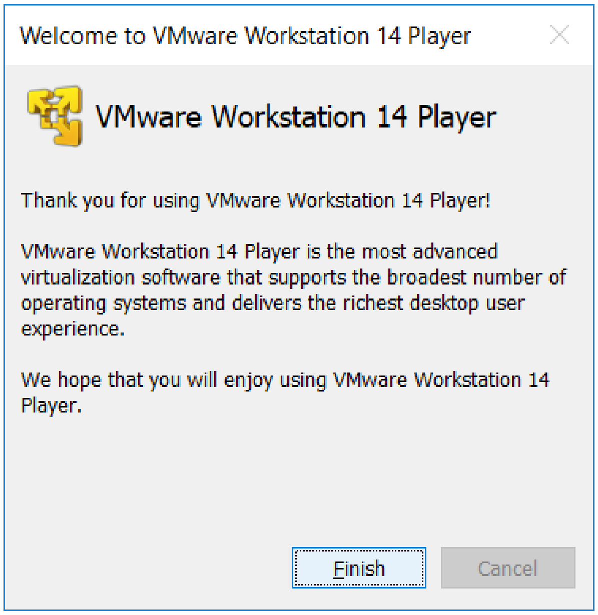

Welcome to VMware Workstation 14 Player

VMware Workstation 14 Player

● Use VMware Workstation 14 Player for free for non-commercial use

○ Enter a license key to allow commercial use:

3. Click “Finish”.



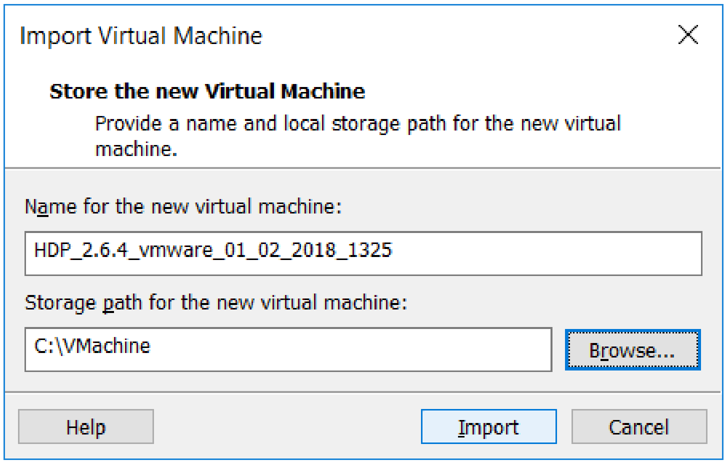
VMware Workstation 14 Player

Thank you for using VMware Workstation 14 Player!

VMware Workstation 14 Player is the most advanced virtualization software that supports the broadest number of operating systems and delivers the richest desktop user experience.

We hope that you will enjoy using VMware Workstation 14 Player.

4. Choose the storage location for the VM and leave the defaults settings (unless you have specific needs). Click on "Import".



Store the new Virtual Machine

Provide a name and local storage path for the new virtual machine.

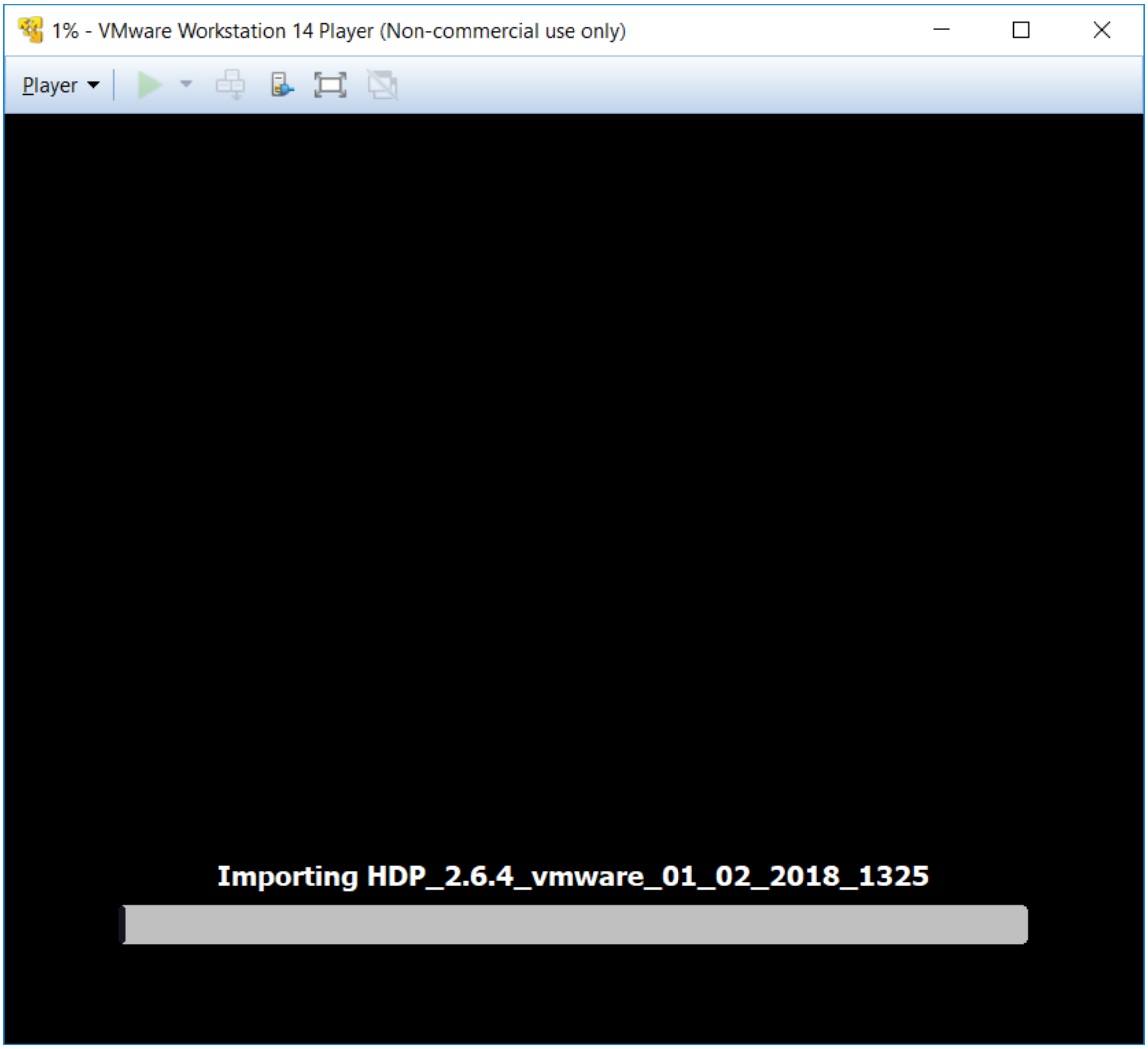
Name for the new virtual machine: HDP_2.6.4_vmware_01_02_2018_1325

Storage path for the new virtual machine: C:\VMachine

5. You will see that Hortonworks being imported automatically in the virtual machine.

Note it!

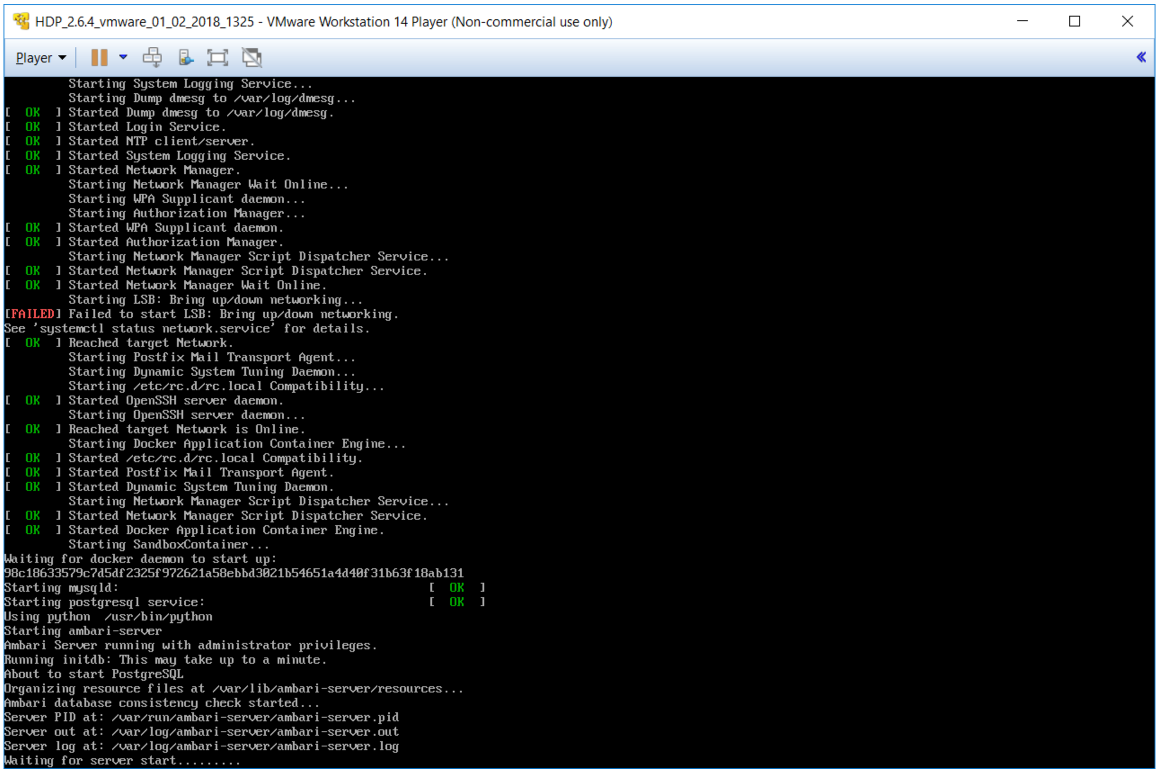
- *In this lab, HDP2..6.4 for VMWare has been used.*



1% - VMware Workstation 14 Player (Non-commercial use only)

Importing HDP_2.6.4_vmware_01_02_2018_1325

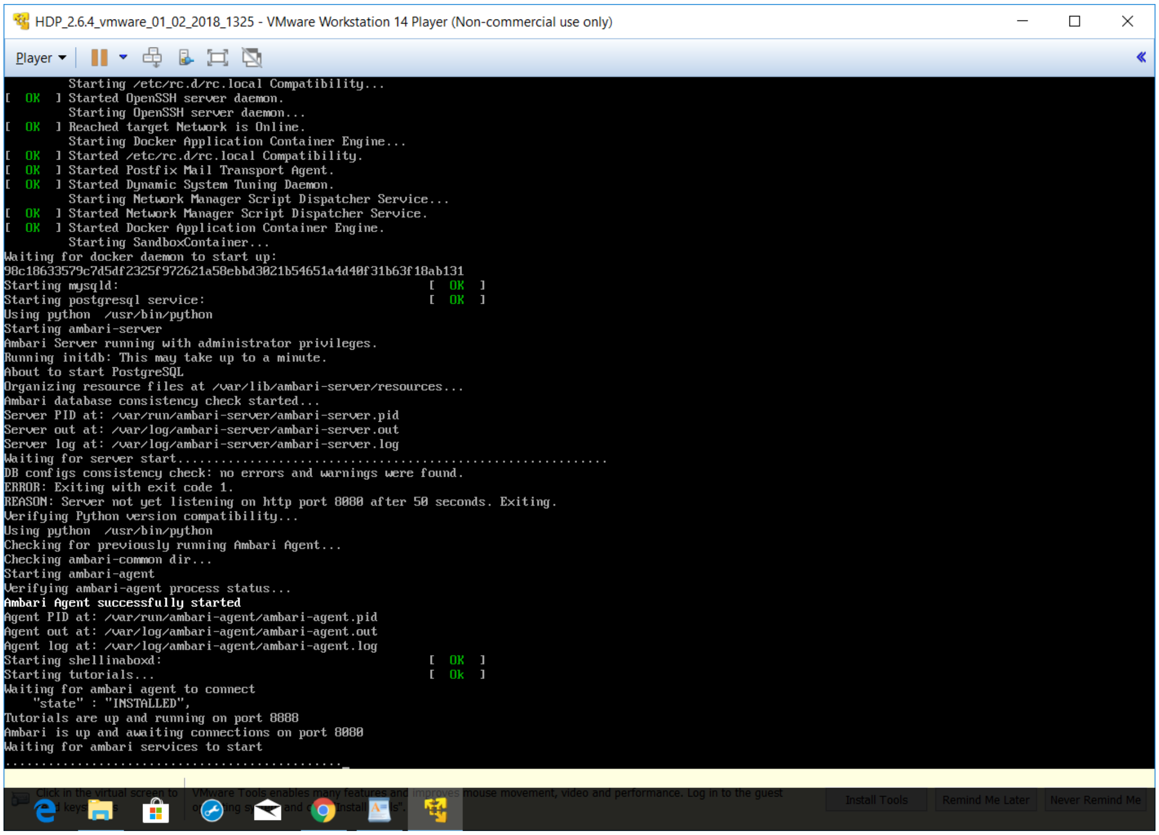
6. You will see the progress being shown and the packages being installed



HDP_2.6.4_vmware_01_02_2018_1325 - VMware Workstation 14 Player (Non-commercial use only)
Player

```
         Starting System Logging Service...
         Starting Dump dmesg to /var/log/dmesg...
[  OK  ] Started Dump dmesg to /var/log/dmesg.
[  OK  ] Started Login Service.
[  OK  ] Started NTP client/server.
[  OK  ] Started System Logging Service.
[  OK  ] Started Network Manager.
         Starting Network Manager Wait Online...
         Starting WPA Supplicant daemon...
         Starting Authorization Manager...
[  OK  ] Started WPA Supplicant daemon.
[  OK  ] Started Authorization Manager.
         Starting Network Manager Script Dispatcher Service...
[  OK  ] Started Network Manager Script Dispatcher Service.
[  OK  ] Started Network Manager Wait Online.
         Starting LSB: Bring up/down networking...
[FAILED] Failed to start LSB: Bring up/down networking.
See 'systemctl status network.service' for details.
[  OK  ] Reached target Network.
         Starting Postfix Mail Transport Agent...
         Starting Dynamic System Tuning Daemon...
         Starting /etc/rc.d/rc.local Compatibility...
[  OK  ] Started OpenSSH server daemon.
         Starting OpenSSH server daemon...
[  OK  ] Reached target Network is Online.
         Starting Docker Application Container Engine...
[  OK  ] Started /etc/rc.d/rc.local Compatibility.
[  OK  ] Started Postfix Mail Transport Agent.
[  OK  ] Started Dynamic System Tuning Daemon.
         Starting Network Manager Script Dispatcher Service...
[  OK  ] Started Network Manager Script Dispatcher Service.
[  OK  ] Started Docker Application Container Engine.
         Starting SandboxContainer...
Waiting for docker daemon to start up:
98c18633579c7d5df2325f972621a58ebbd3021b54651a4d40f31b63f18ab131
Starting mysqld:                                           [  OK  ]
Starting postgresql service:                               [  OK  ]
Using python  /usr/bin/python
Starting ambari-server
Ambari Server running with administrator privileges.
Running initdb: This may take up to a minute.
About to start PostgreSQL
Organizing resource files at /var/lib/ambari-server/resources...
Ambari database consistency check started...
Server PID at: /var/run/ambari-server/ambari-server.pid
Server out at: /var/log/ambari-server/ambari-server.out
Server log at: /var/log/ambari-server/ambari-server.log
Waiting for server start.........
```

6. This might take some time. Wait for the progress


HDP_2.6.4_vmware_01_02_2018_1325 - VMware Workstation 14 Player (Non-commercial use only)
Player
Install Tools
Remind Me Later
Never Remind Me

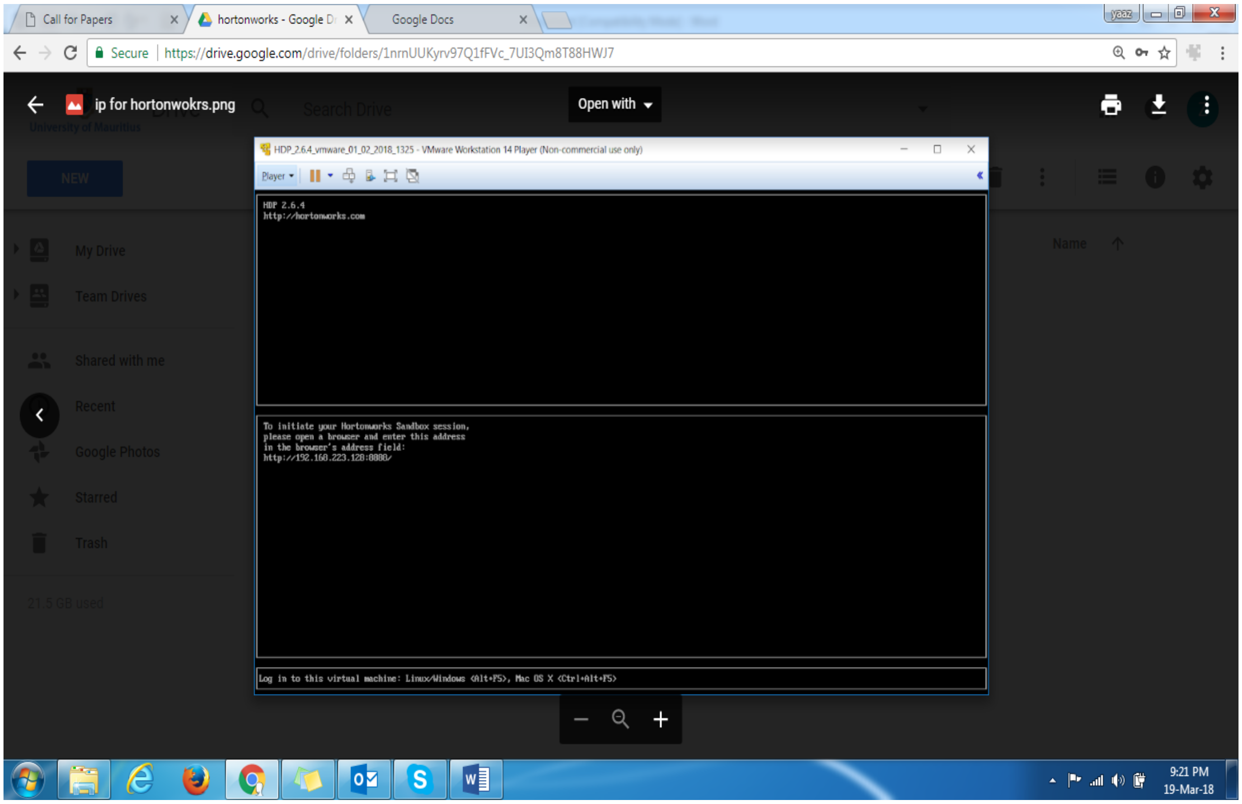
```
         Starting /etc/rc.d/rc.local Compatibility...
[  OK  ] Started OpenSSH server daemon.
         Starting OpenSSH server daemon...
[  OK  ] Reached target Network is Online.
         Starting Docker Application Container Engine...
[  OK  ] Started /etc/rc.d/rc.local Compatibility.
[  OK  ] Started Postfix Mail Transport Agent.
[  OK  ] Started Dynamic System Tuning Daemon.
         Starting Network Manager Script Dispatcher Service...
[  OK  ] Started Network Manager Script Dispatcher Service.
[  OK  ] Started Docker Application Container Engine.
         Starting SandboxContainer...
Waiting for docker daemon to start up:
98c18633579c7d5df2325f972621a58ebbd3021b54651a4d40f31b63f18ab131
Starting mysqld:                                           [  OK  ]
Starting postgresql service:                               [  OK  ]
Using python  /usr/bin/python
Starting ambari-server
Ambari Server running with administrator privileges.
Running initdb: This may take up to a minute.
About to start PostgreSQL
Organizing resource files at /var/lib/ambari-server/resources...
Ambari database consistency check started...
Server PID at: /var/run/ambari-server/ambari-server.pid
Server out at: /var/log/ambari-server/ambari-server.out
Server log at: /var/log/ambari-server/ambari-server.log
Waiting for server start............................................................
DB configs consistency check: no errors and warnings were found.
ERROR: Exiting with exit code 1.
REASON: Server not yet listening on http port 8080 after 50 seconds. Exiting.
Verifying Python version compatibility...
Using python  /usr/bin/python
Checking for previously running Ambari Agent...
Checking ambari-common dir...
Starting ambari-agent
Verifying ambari-agent process status...
Ambari Agent successfully started
Agent PID at: /var/run/ambari-agent/ambari-agent.pid
Agent out at: /var/log/ambari-agent/ambari-agent.out
Agent log at: /var/log/ambari-agent/ambari-agent.log
Starting shellinaboxd:                                     [  OK  ]
Starting tutorials...                                      [  Ok  ]
Waiting for ambari agent to connect
    "state" : "INSTALLED",
Tutorials are up and running on port 8888
Ambari is up and awaiting connections on port 8080
Waiting for ambari services to start
...............................................
```

7. The following screen will appear after installation. Note the IP which is displayed which will have to be opened from your host computer.

http://192.168.223.108:8888

Note it!

- *You shall be accessing the VM from your host computer by using this IP.*



```
HDP 2.6.4
http://hortonworks.com
```

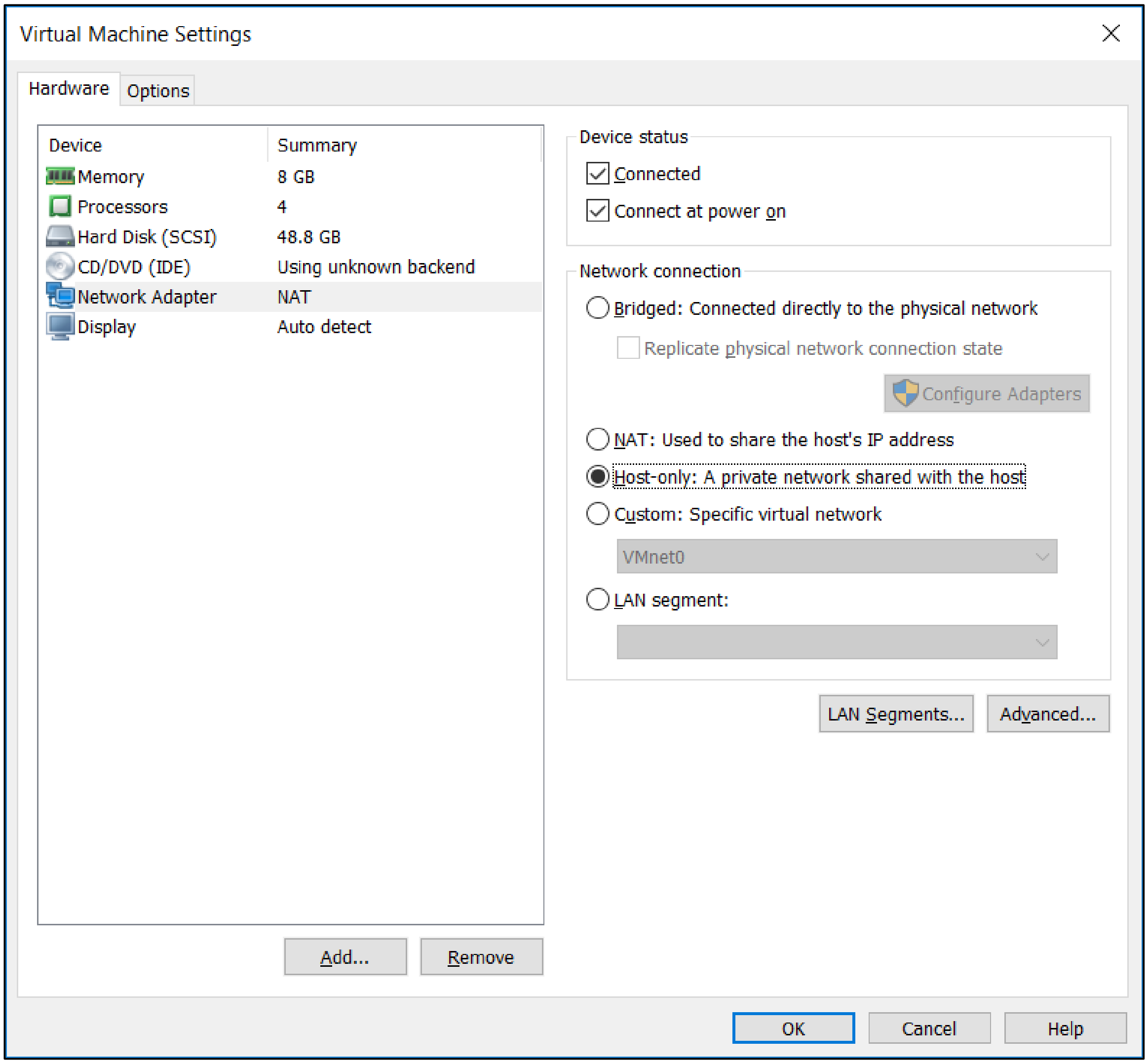
```
To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://192.168.223.128:8888/
```

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

8. From the VM, press Alt+F5 to log into your virtual machine. The user name and password is root and hadoop respectively

9. From the VM, go to player > manage > virtual machine setting and press the tab “Hardware”.

10. The following screen will appear. Click on Network adapter. Check the Network connection option. You can keep to NAT or host only depending on your requirements. For this setup, NAT was selected.



Device	Summary
Memory	8 GB
Processors	4
Hard Disk (SCSI)	48.8 GB
CD/DVD (IDE)	Using unknown backend
Network Adapter	NAT
Display	Auto detect

Device status

☑ Connected

☑ Connect at power on

Network connection

○ Bridged: Connected directly to the physical network

□ Replicate physical network connection state

○ NAT: Used to share the host's IP address

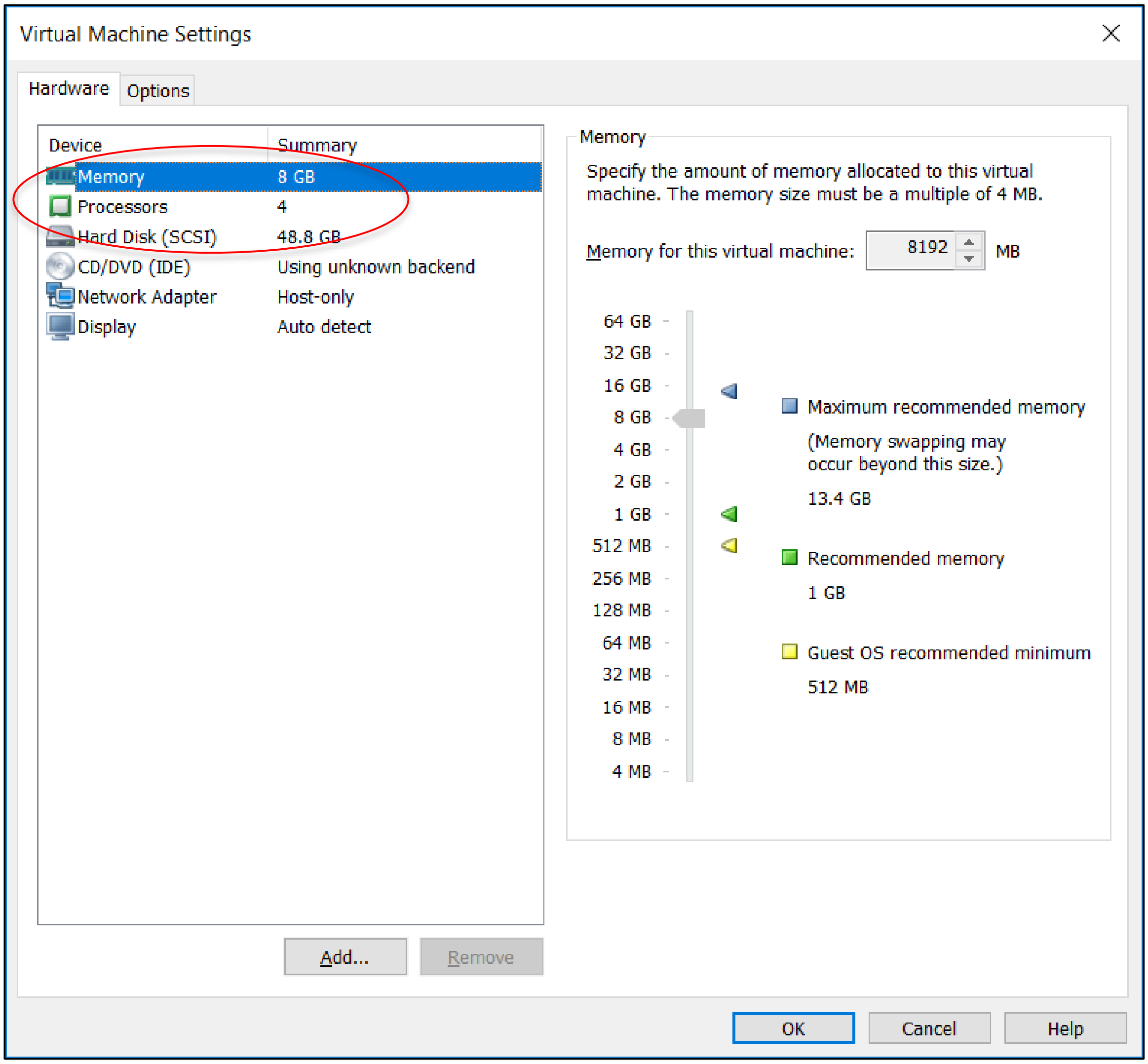
● Host-only: A private network shared with the host

○ Custom: Specific virtual network

VMnet0

○ LAN segment:

11. You can check the other settings. Select Memory as shown below. You will see that 8 GB memory has automatically been allocated for this VM as well as a portion of the hard disk.



Device	Summary
Memory	8 GB
Processors	4
Hard Disk (SCSI)	48.8 GB
CD/DVD (IDE)	Using unknown backend
Network Adapter	Host-only
Display	Auto detect

Memory

Specify the amount of memory allocated to this virtual machine. The memory size must be a multiple of 4 MB.

Memory for this virtual machine: 8192 MB

🟦 Maximum recommended memory

(Memory swapping may occur beyond this size.)

13.4 GB

🟩 Recommended memory

1 GB

🟨 Guest OS recommended minimum

512 MB

12. Go to your host computer and type the IP saved from Step 7 above.

http://192.168.223.108:8888

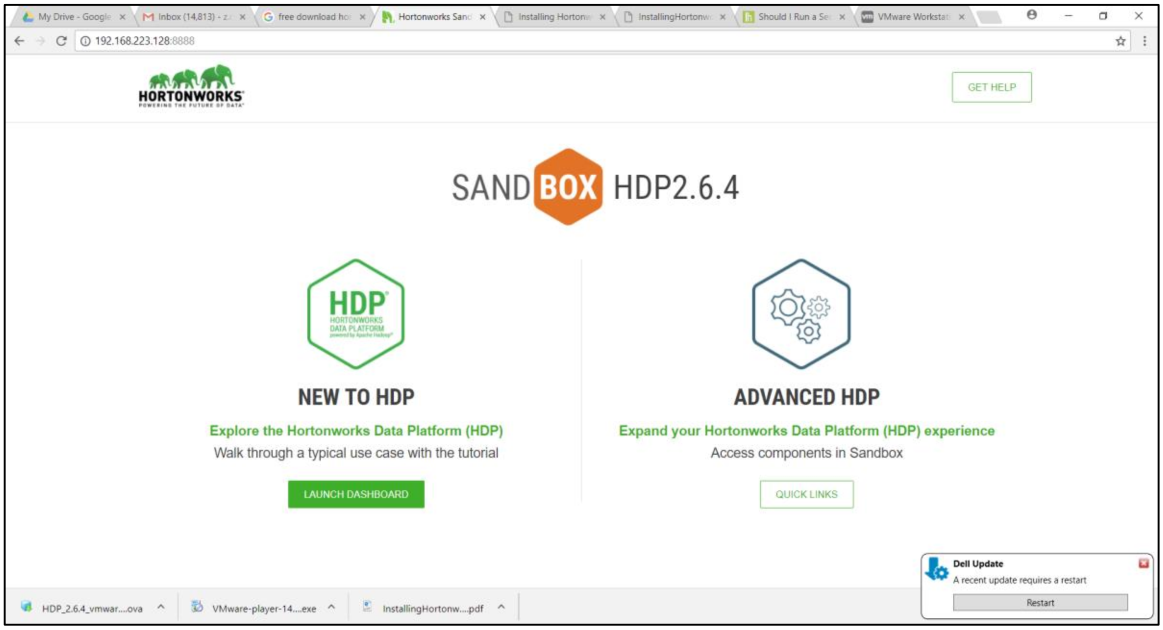
13. The following screen should appear and your Big Data platform should be ready to use.

You might not have the same IP and port number. In case you are not able

Note it!

to open your Hortonworks check the following:

1. *You VM should be in status open and not shut down*
2. *You might have to use other port numbers. Please search for the options*



NEW TO HDP

Explore the Hortonworks Data Platform (HDP)

Walk through a typical use case with the tutorial

ADVANCED HDP

Expand your Hortonworks Data Platform (HDP) experience

Access components in Sandbox

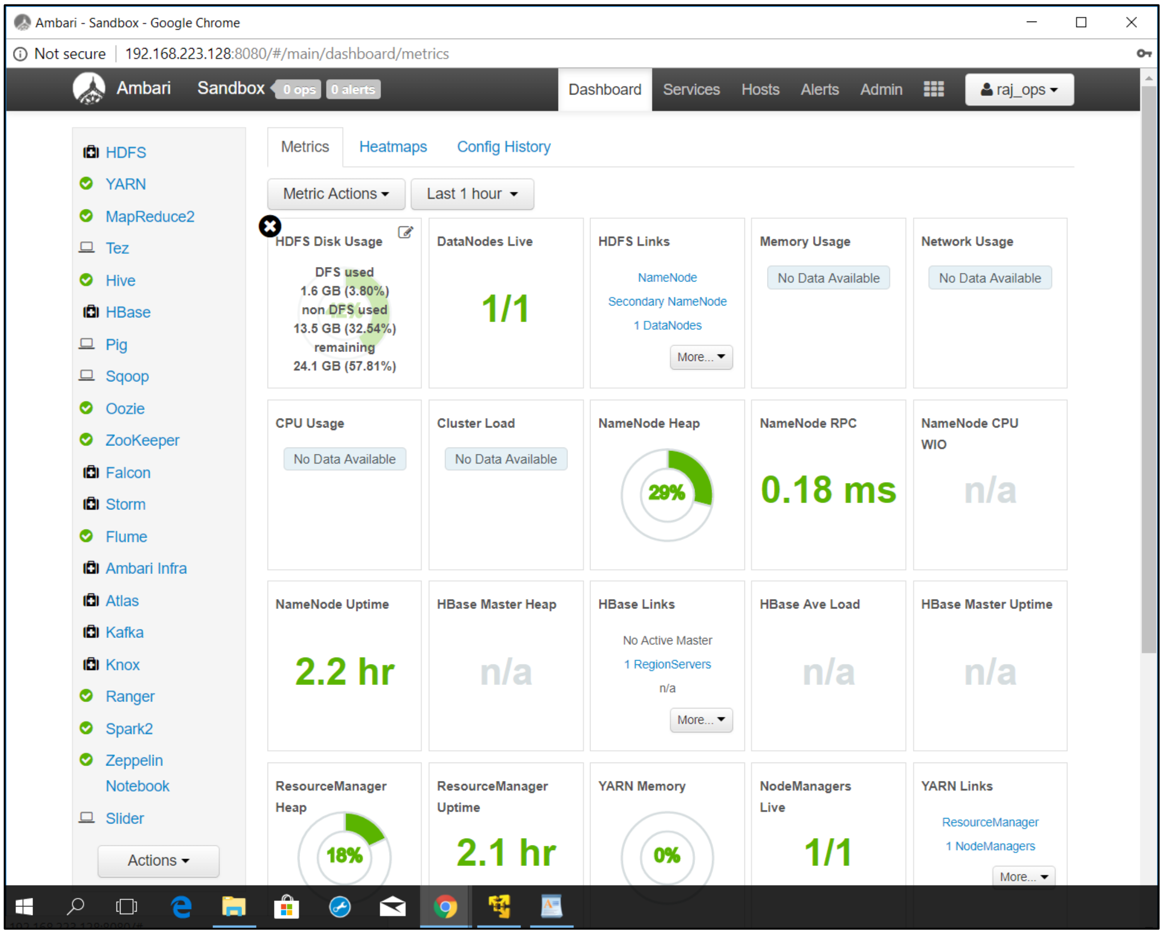
Dell Update

A recent update requires a restart

11. Click on launch dashboard and select “Ambari”. The following user name and password have to be used.
    raj_ops (same username and password)

Note it!

- *You might have a different username and password depending on the version used.*



4. Installing R/RStudio on Hadoop

To be able to run R on the Big Data platform, the related R packages and IDE have to be installed. The following should be setup in your Big Data platform:

1. R
2. R studio

4.1 Installing R and RStudio on the VM

- *For each installation/download, you should check whether the*

Note it!

installation or download is successful. Otherwise, you have to look for other versions depending on the OS/version/virtual machine that you are using.

- *For this VM, linux centos version 5 has been used.*
- *When installing select "y" whenever prompted to do so to continue with the installation*

Steps

1. Start by installing R. Login to your virtual machine (if not already login). We should first enable the epel repository using the following command. Type the following command and press enter.

```
yum install epel-release -y
```

2. Install R by typing the following and press enter. R version 3.4.4 has been used in this setup.

```
yum install R -y
```

3. Install RServer IDE for RStudio. For this, you should first download RStudio and then install.

4. Use command wget to download RStudio.

Note it!

- *If you use wget and you get the error message "command not found", you should first install wget using yum install wget*

```
yum install wget
```

5. Download RStudio using

```
wget https://download2.rstudio.org/rstudio-server-rhel-1.0.44-x86_64.rpm
```

6. Install RStudio after successful download using

```
yum install --nogpgcheck rstudio-server-rhel-1.0.44-x86_64.rpm
```

7. Once the package has been get and installed, it will be automatically starting the service. To check this use

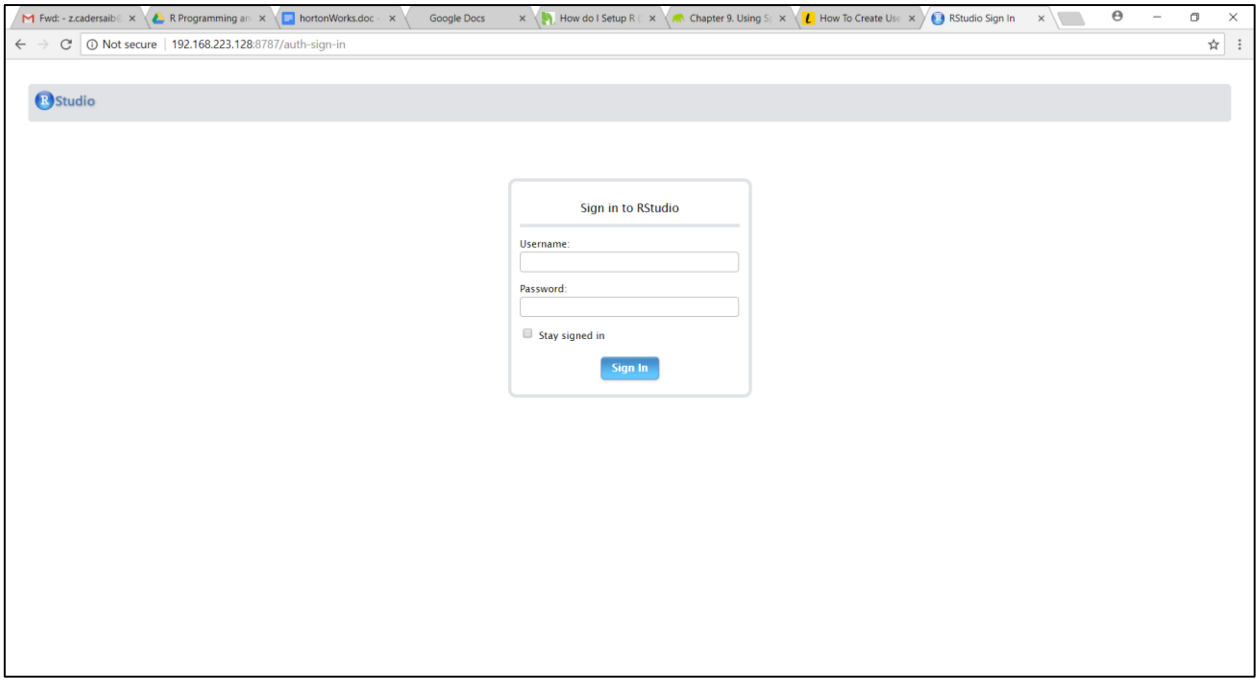
```
systemctl status rstudio-server.service
```

Note it!	- *You should have the message Active (running).*

8. Now you can open Rstudio from the browser of your host computer using port 8787 by typing the following IP

   http://192.168.223.128:8787

9. The following window should be opened prompting you to enter the user name and password.



192.168.223.128:8787/auth-sign-in

Sign in to RStudio

Username:

Password:

□ Stay signed in

Sign In

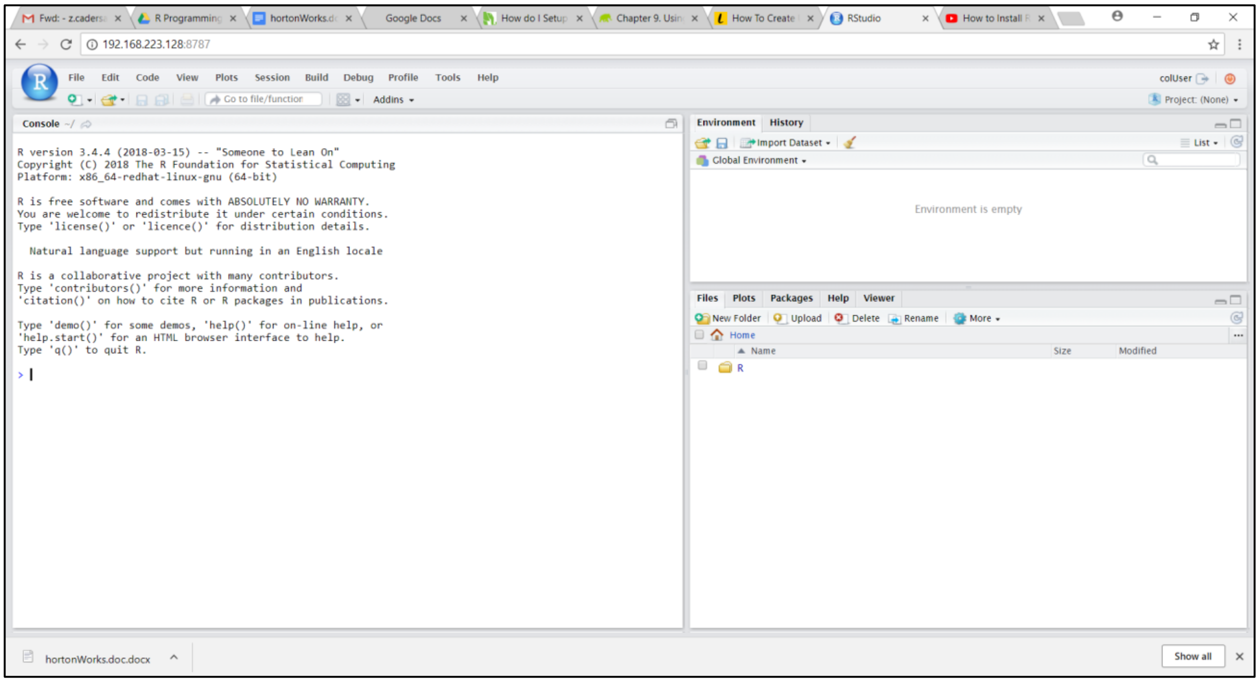
10. You should create **another user than root**. See instructions below for creating a new user from the VM command line. Go back to the VM to create a new user. Creating a new user, **colUser (or any other user)** using the command

```
useradd -m colUser
```

11. Set the password of the new user created (in this case colUser) using

```
passwd colUser
```

12. Back in the browser of your host computer, use the new user (colUser) created to login and RStudio should be opened successfully as below



```
R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

13. Go back to the terminal and install the developers' tool using

```
yum groupinstall "Development Tools"
```

14. After successful installation of the development packages, install packages in RStudio by going back to browser where RStudio is opened and install the following packages.

```
>install.packages('quantmod')

>install.packages('txtplot')
```

Note it!

- *Refer to chapter 1 to know how to install packages in RStudio.*