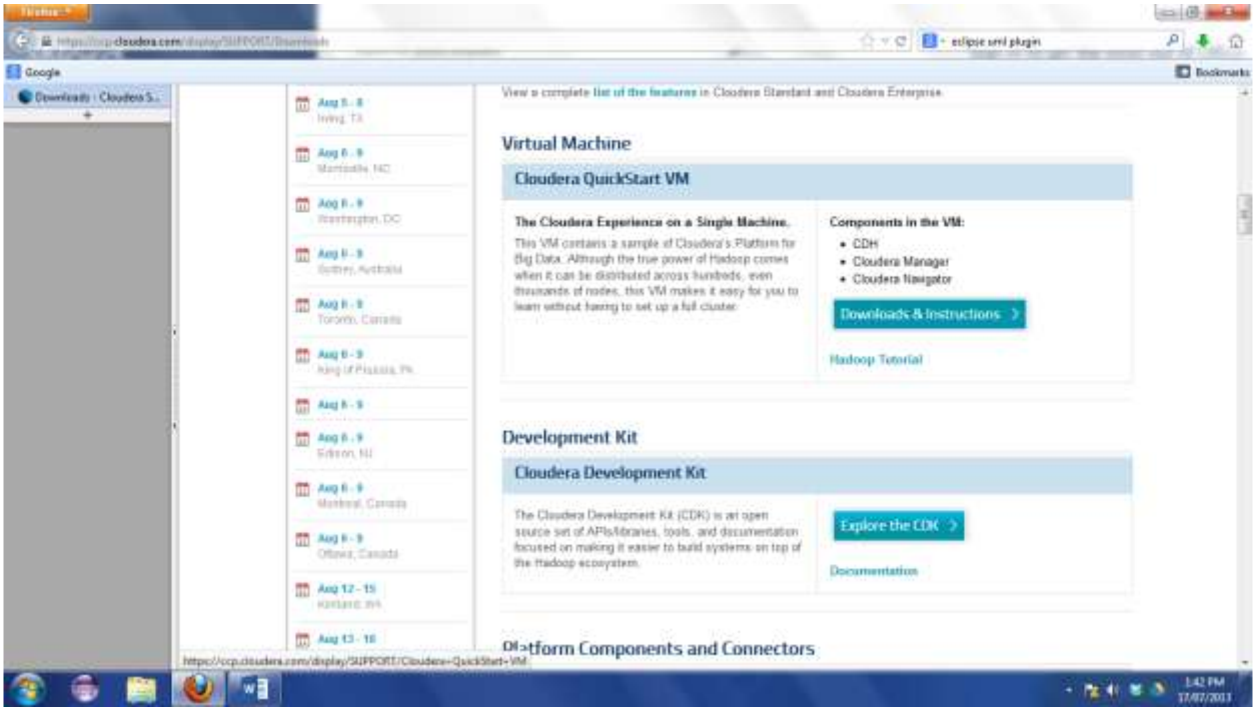
Cloudera Hadoop Installation and Configuration

1. Go to Cloudera Quickstart VM to download a pre-setup CDH virtual machine.



View a complete list of the features in Cloudera Standard and Cloudera Enterprise.

Virtual Machine

Cloudera QuickStart VM

The Cloudera Experience on a Single Machine.
This VM contains a sample of Cloudera's Platform for Big Data. Although the true power of Hadoop comes when it can be distributed across hundreds, even thousands of nodes, this VM makes it easy for you to learn without having to set up a full cluster.

Components in the VM:

- CDH
- Cloudera Manager
- Cloudera Navigator

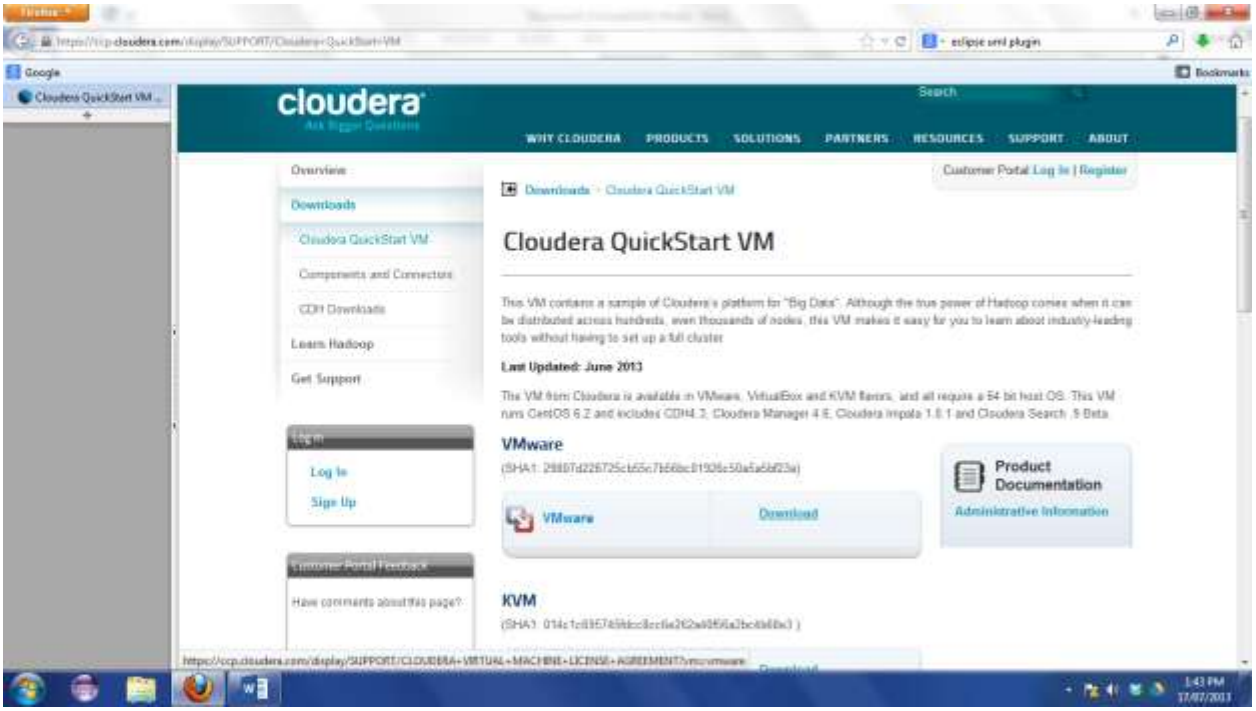
Development Kit

Cloudera Development Kit

The Cloudera Development Kit (CDK) is an open source set of APIs/libraries, tools, and documentation focused on making it easier to build systems on top of the Hadoop ecosystem.

Platform Components and Connectors

2. Select a VM you wish to download. For purpose of this assignment, I have used VMware Player.



Cloudera QuickStart VM

This VM contains a sample of Cloudera's platform for "Big Data". Although the true power of Hadoop comes when it can be distributed across hundreds, even thousands of nodes, this VM makes it easy for you to learn about industry-leading tools without having to set up a full cluster.

Last Updated: June 2013

The VM from Cloudera is available in VMware, VirtualBox and KVM flavors, and all require a 64 bit host OS. This VM runs CentOS 6.2 and includes CDH4.3, Cloudera Manager 4.6, Cloudera Impala 1.0.1 and Cloudera Search .9 Beta.

VMware

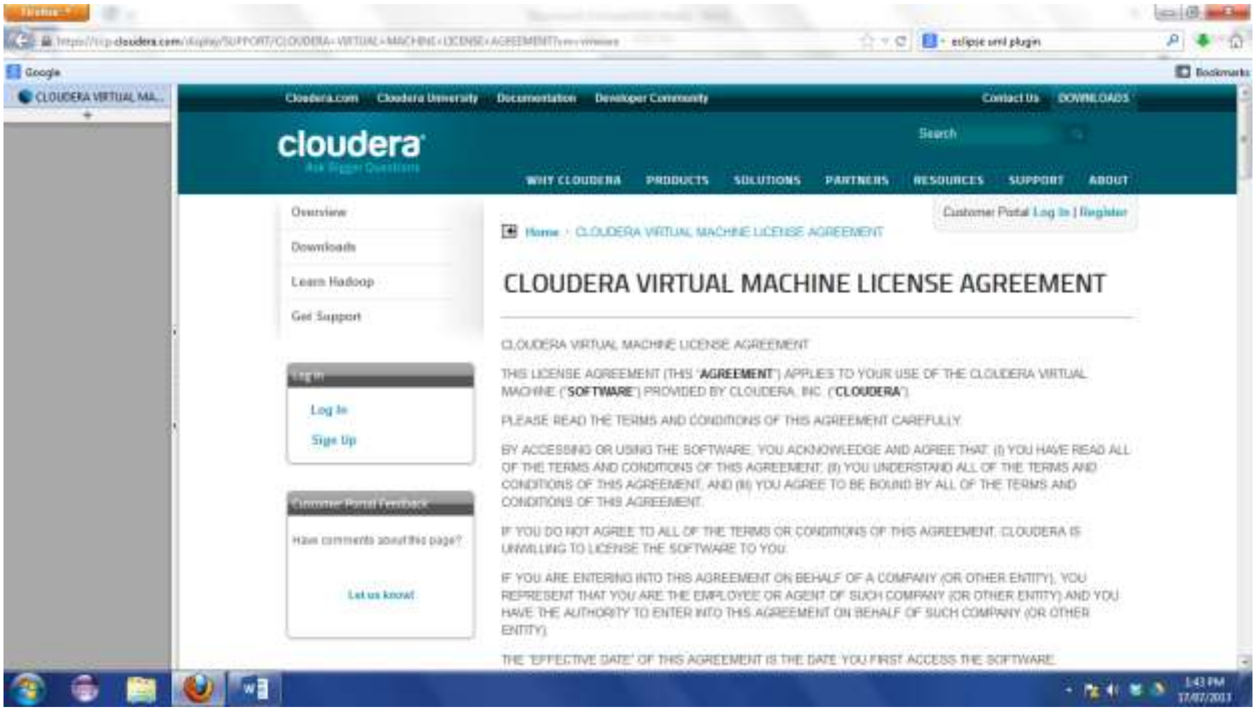
(SHA1: 28807d226725cb55c7b56bc01926c50a5a5bf23e)

KVM

(SHA1: 014c1c8357456dcc8ccfa262a4856a2bc4b68e3)

Have comments about this page?

3. To get download file, accept the agreement.



CLOUDERA VIRTUAL MACHINE LICENSE AGREEMENT

CLOUDERA VIRTUAL MACHINE LICENSE AGREEMENT

THIS LICENSE AGREEMENT (THIS "**AGREEMENT**") APPLIES TO YOUR USE OF THE CLOUDERA VIRTUAL MACHINE ("**SOFTWARE**") PROVIDED BY CLOUDERA, INC. ("**CLOUDERA**").

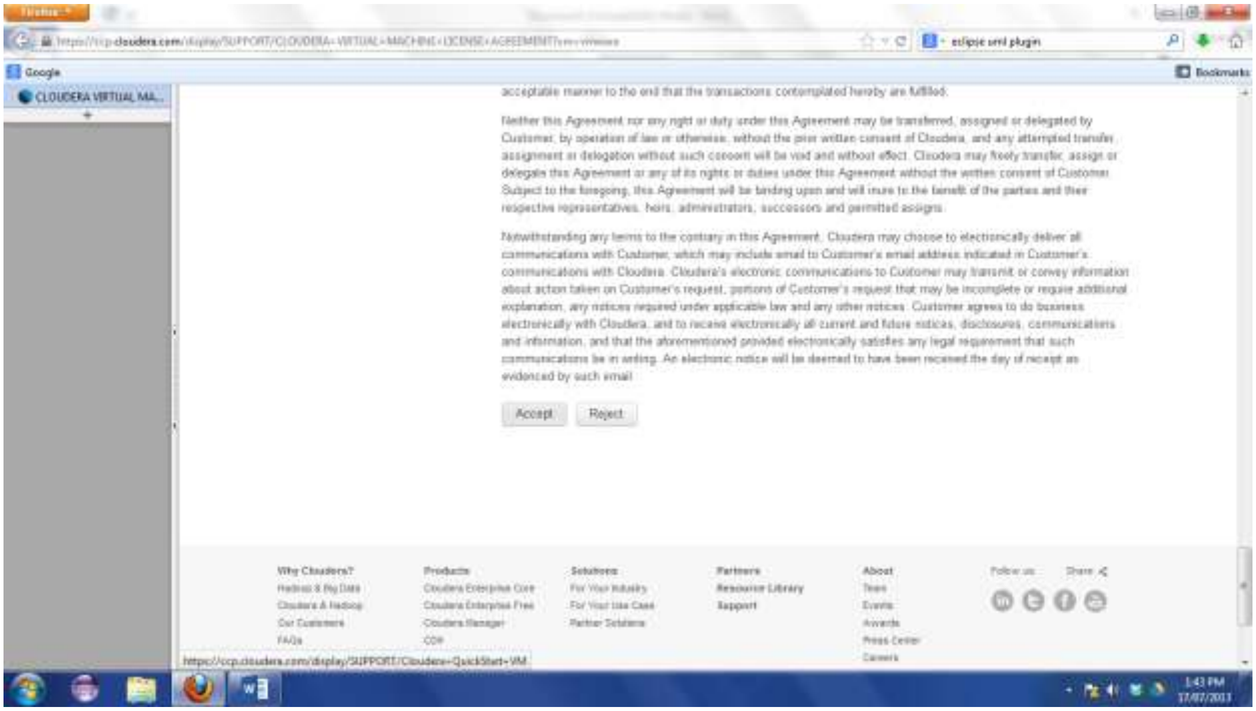
PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY.

BY ACCESSING OR USING THE SOFTWARE, YOU ACKNOWLEDGE AND AGREE THAT: (I) YOU HAVE READ ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT, (II) YOU UNDERSTAND ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT, AND (III) YOU AGREE TO BE BOUND BY ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT.

IF YOU DO NOT AGREE TO ALL OF THE TERMS OR CONDITIONS OF THIS AGREEMENT, CLOUDERA IS UNWILLING TO LICENSE THE SOFTWARE TO YOU.

IF YOU ARE ENTERING INTO THIS AGREEMENT ON BEHALF OF A COMPANY (OR OTHER ENTITY), YOU REPRESENT THAT YOU ARE THE EMPLOYEE OR AGENT OF SUCH COMPANY (OR OTHER ENTITY) AND YOU HAVE THE AUTHORITY TO ENTER INTO THIS AGREEMENT ON BEHALF OF SUCH COMPANY (OR OTHER ENTITY).

THE "EFFECTIVE DATE" OF THIS AGREEMENT IS THE DATE YOU FIRST ACCESS THE SOFTWARE.

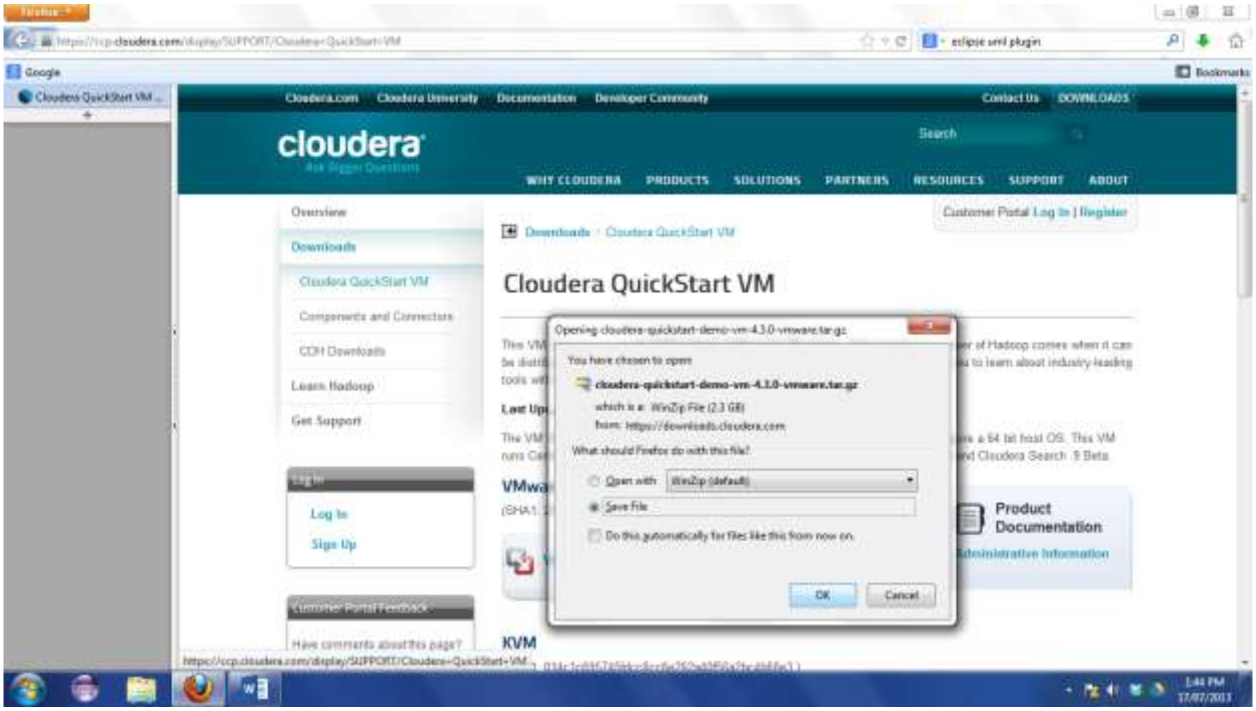


acceptable manner to the end that the transactions contemplated hereby are fulfilled.

Neither this Agreement nor any right or duty under this Agreement may be transferred, assigned or delegated by Customer, by operation of law or otherwise, without the prior written consent of Cloudera, and any attempted transfer, assignment or delegation without such consent will be void and without effect. Cloudera may freely transfer, assign or delegate this Agreement or any of its rights or duties under this Agreement without the written consent of Customer. Subject to the foregoing, this Agreement will be binding upon and will inure to the benefit of the parties and their respective representatives, heirs, administrators, successors and permitted assigns.

Notwithstanding any terms to the contrary in this Agreement, Cloudera may choose to electronically deliver all communications with Customer, which may include email to Customer's email address indicated in Customer's communications with Cloudera. Cloudera's electronic communications to Customer may transmit or convey information about action taken on Customer's request, portions of Customer's request that may be incomplete or require additional explanation, any notices required under applicable law and any other notices. Customer agrees to do business electronically with Cloudera, and to receive electronically all current and future notices, disclosures, communications and information, and that the aforementioned provided electronically satisfies any legal requirement that such communications be in writing. An electronic notice will be deemed to have been received the day of receipt as evidenced by such email.

4. Save the downloadable file.



Cloudera QuickStart VM

Opening cloudera-quickstart-demo-vm-4.3.0-vmware.tar.gz

You have chosen to open

cloudera-quickstart-demo-vm-4.3.0-vmware.tar.gz

which is a: WinZip File (2.3 GB)

from: https://downloads.cloudera.com

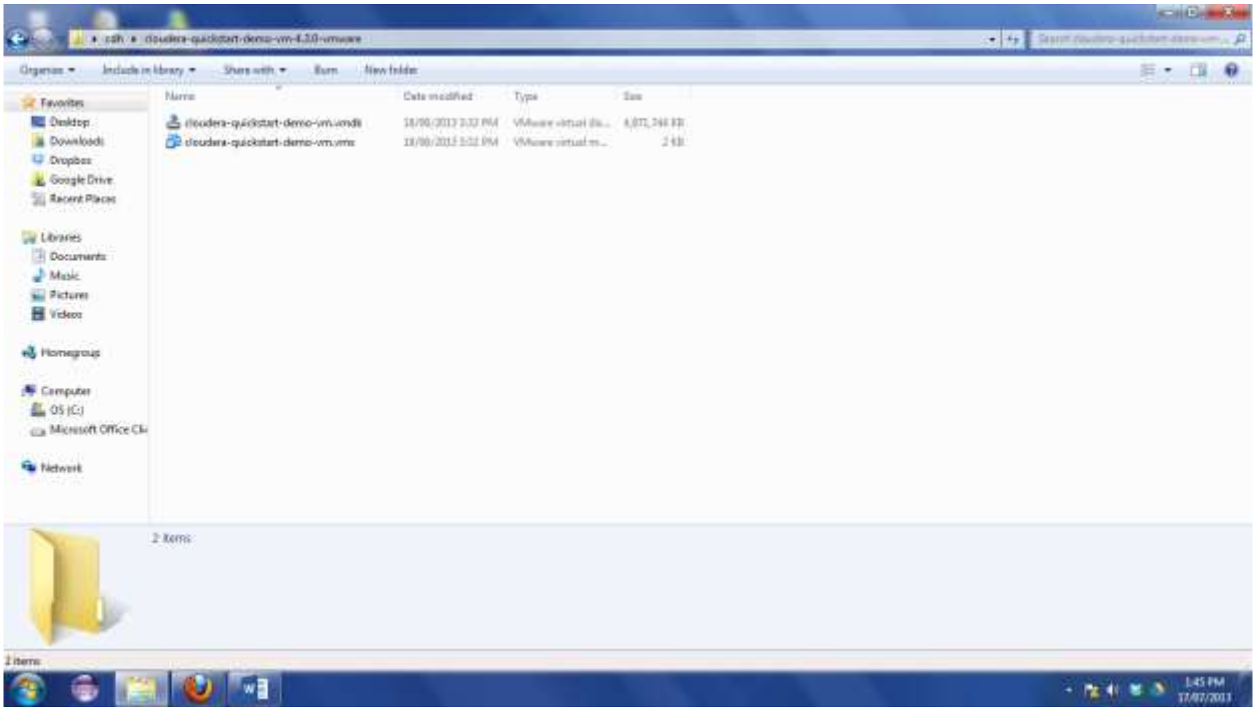
What should Firefox do with this file?

○ Open with WinZip (default)

● Save File

□ Do this automatically for files like this from now on.

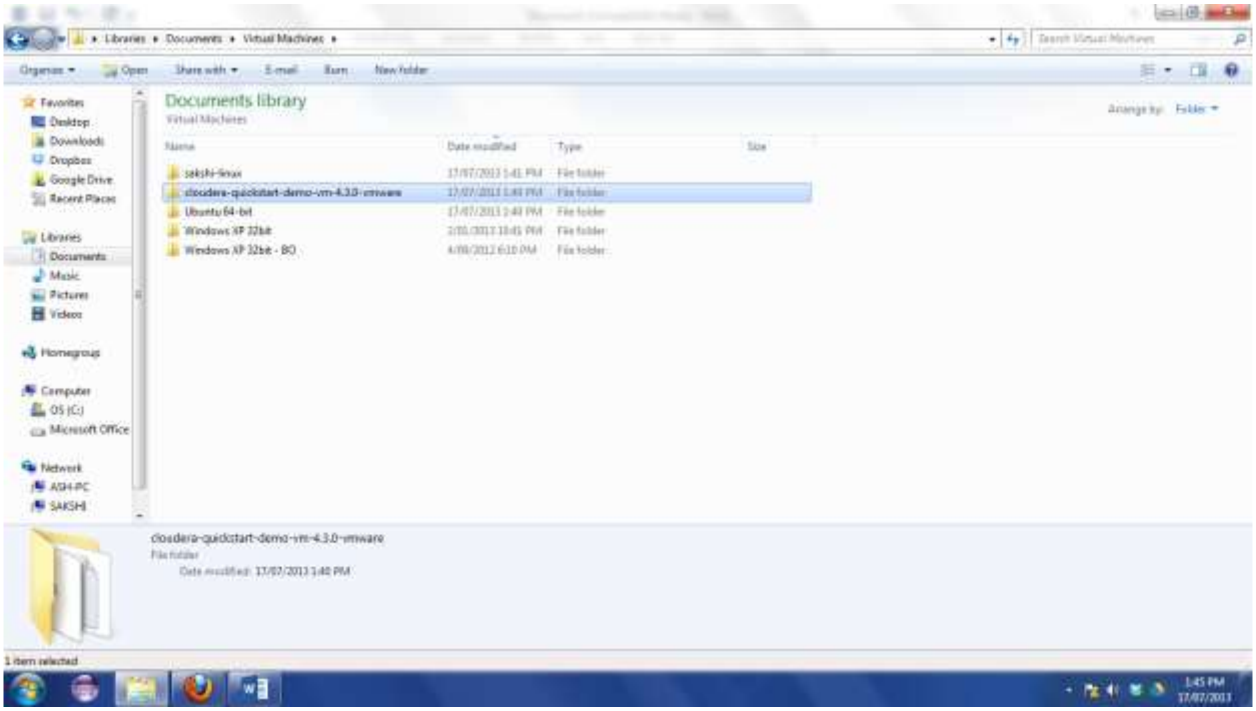
5. Unzip the downloaded file. You will get 2 files - .vmx (Virtual machine) and .vmdk



Name	Date modified	Type	Size
cloudera-quickstart-demo-vm.vmdk	18/06/2013 3:32 PM	VMware virtual di...	4,071,744 KB
cloudera-quickstart-demo-vm.vmx	18/06/2013 3:32 PM	VMware virtual m...	2 KB

2 items

6. Save this folder in the directory where VMPlayer stores all virtual machine files.

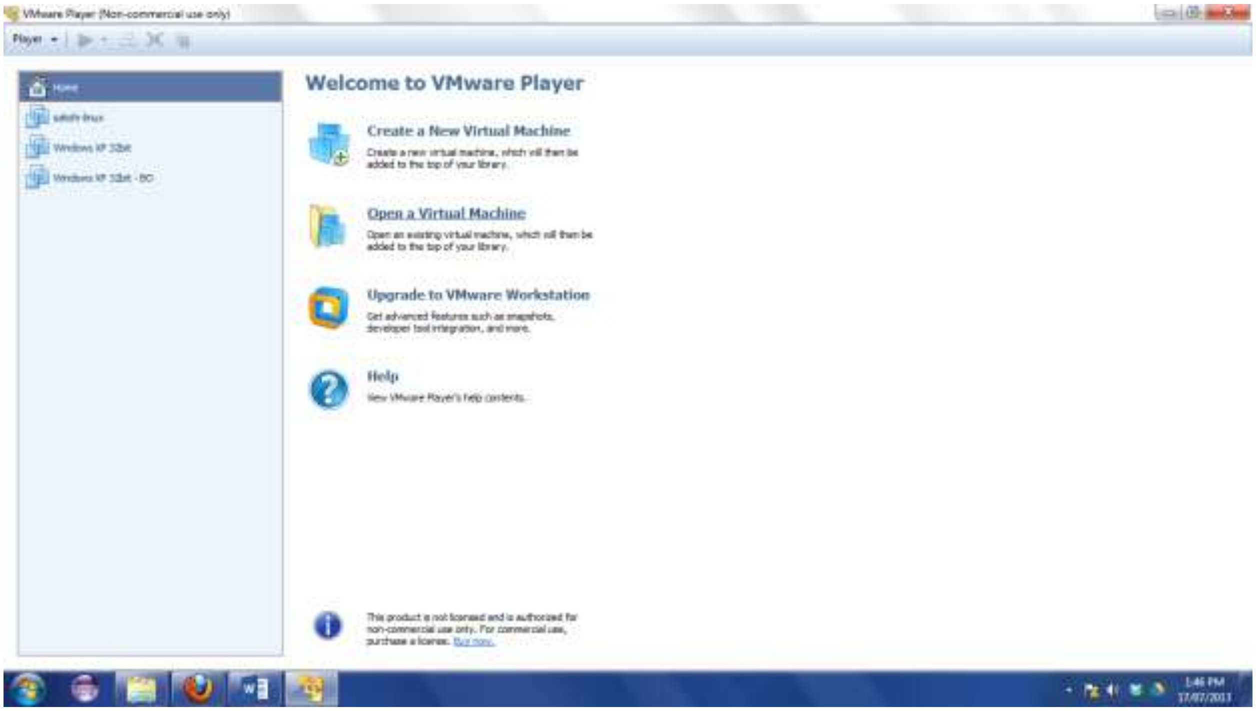


Name	Date modified	Type	Size
sakshi-linux	17/07/2013 1:41 PM	File folder	
cloudera-quickstart-demo-vm-4.3.0-vmware	17/07/2013 1:40 PM	File folder	
Ubuntu 64-bit	17/07/2013 1:40 PM	File folder	
Windows XP 32bit	1/01/2013 10:41 PM	File folder	
Windows XP 32bit - BO	4/08/2012 6:10 PM	File folder	

cloudera-quickstart-demo-vm-4.3.0-vmware
File folder
Date modified: 17/07/2013 1:40 PM

1 item selected

7. Go to VMWare player and click on Open a Virtual Machine.



Welcome to VMware Player

Create a New Virtual Machine
Create a new virtual machine, which will then be added to the top of your library.

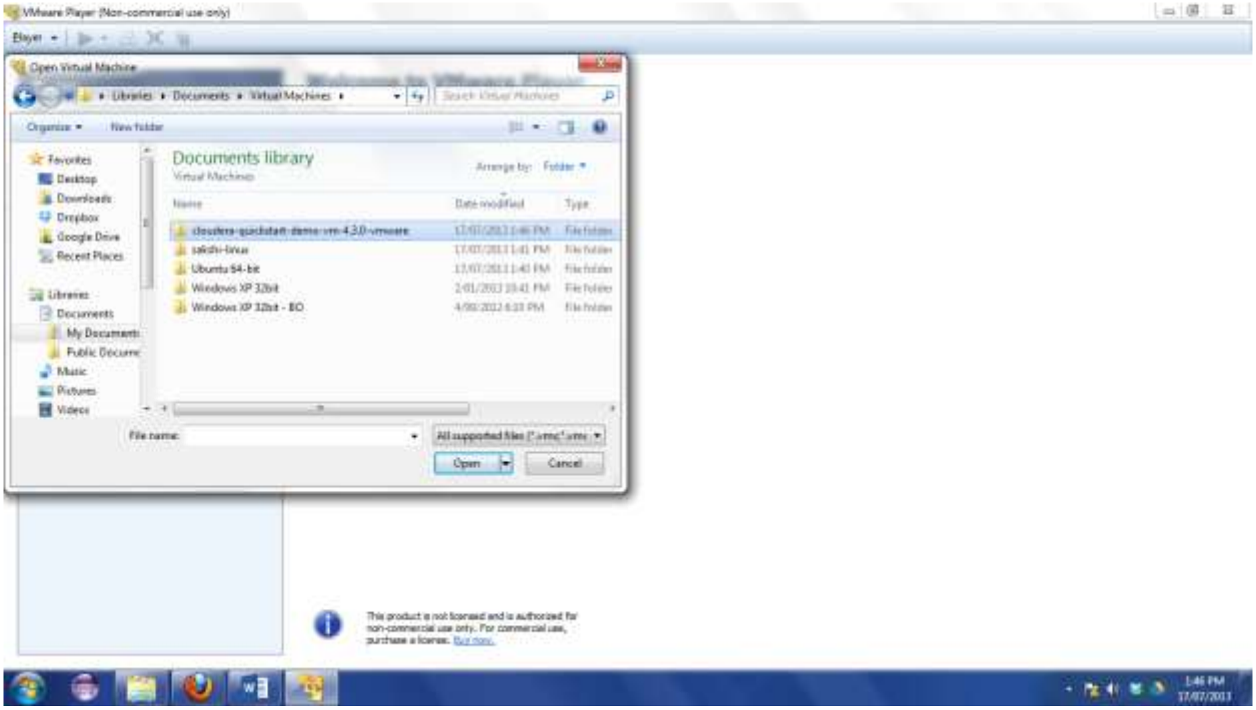
Open a Virtual Machine
Open an existing virtual machine, which will then be added to the top of your library.

Upgrade to VMware Workstation
Get advanced features such as snapshots, developer tool integration, and more.

Help
View VMware Player's help contents.

This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.

8. Select the newly copied virtual machine.



VMware Player (Non-commercial use only)

Open Virtual Machine

Libraries ▸ Documents ▸ Virtual Machines ▸

Documents library
Virtual Machines

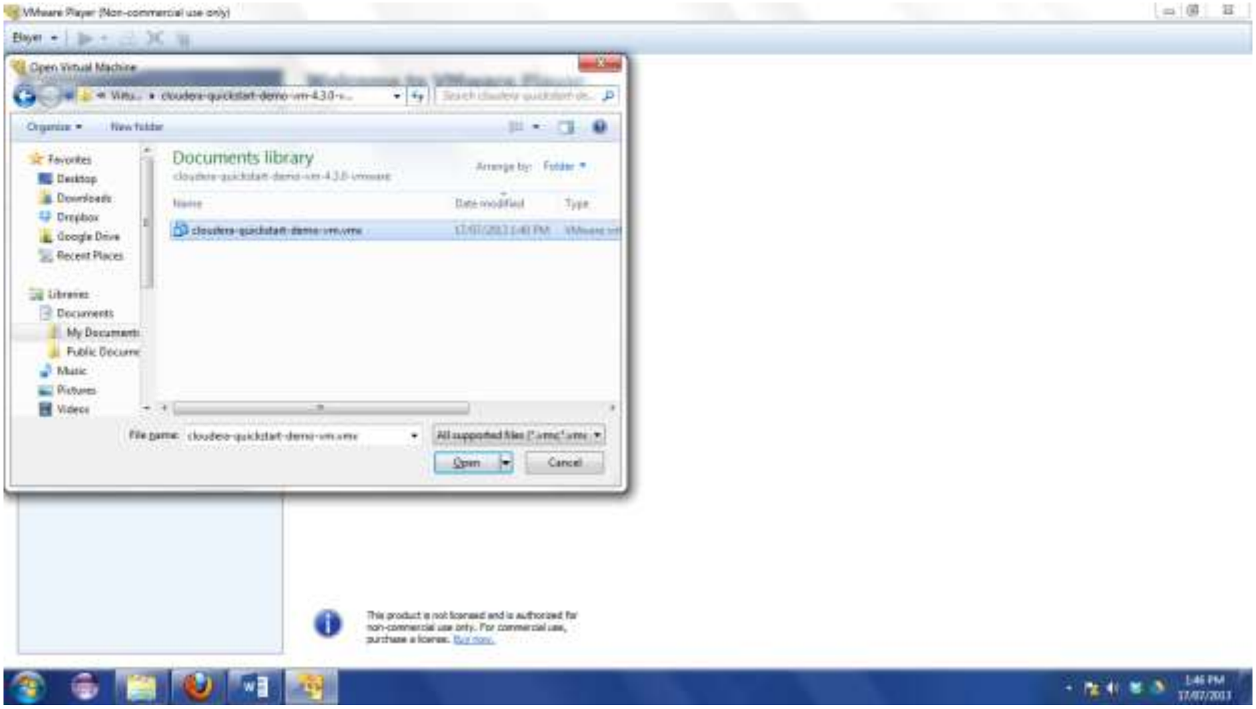
Name	Date modified	Type
cloudera-quickstart-demo-vm-4.3.0-vmware	17/07/2013 1:46 PM	File folder
sakshi-linux	17/07/2013 1:41 PM	File folder
Ubuntu 64-bit	17/07/2013 1:40 PM	File folder
Windows XP 32bit	2/01/2013 10:41 PM	File folder
Windows XP 32bit - BO	4/09/2012 6:33 PM	File folder

File name:

All supported files (*.vmc;*.vmx

This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.

9. Select .vmx file which is the virtual machine file.



VMware Player (Non-commercial use only)

Open Virtual Machine

Virtu... ▸ cloudera-quickstart-demo-vm-4.3.0-v...

Documents library
cloudera-quickstart-demo-vm-4.3.0-vmware

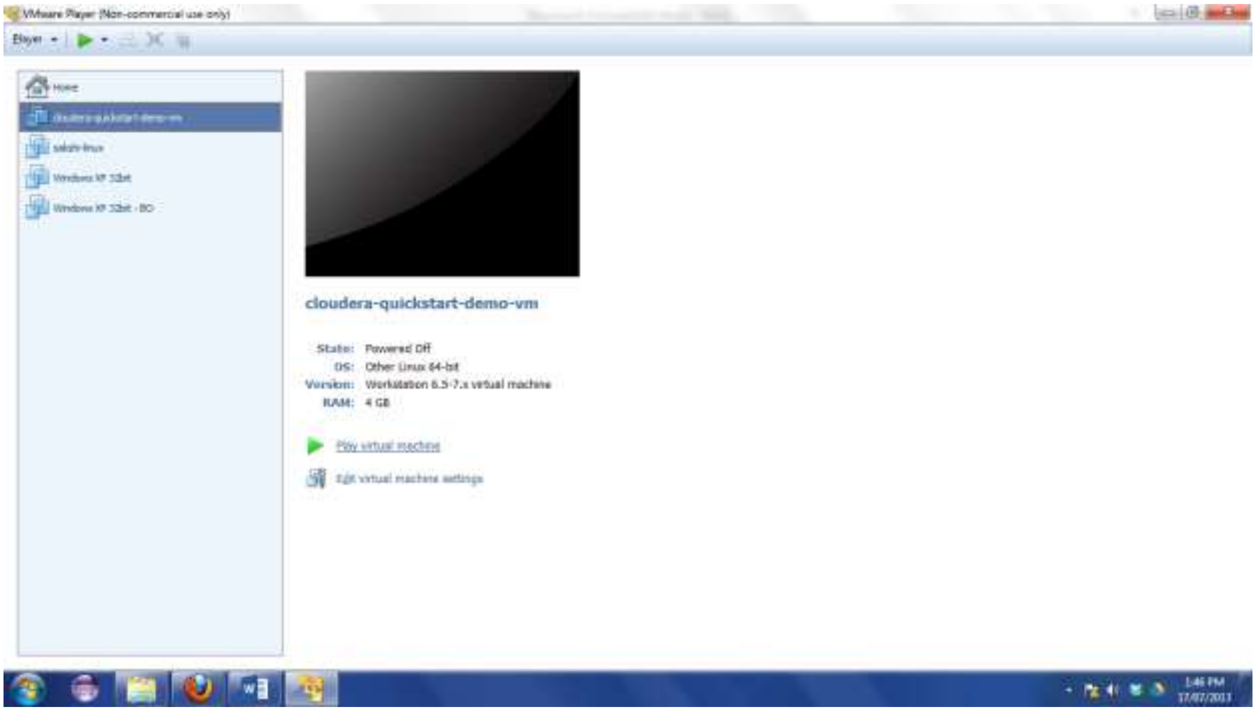
Name	Date modified	Type
cloudera-quickstart-demo-vm.vmx	17/07/2013 1:40 PM	VMware vir

File name: cloudera-quickstart-demo-vm.vmx

All supported files (*.vmc;*.vmx

This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.

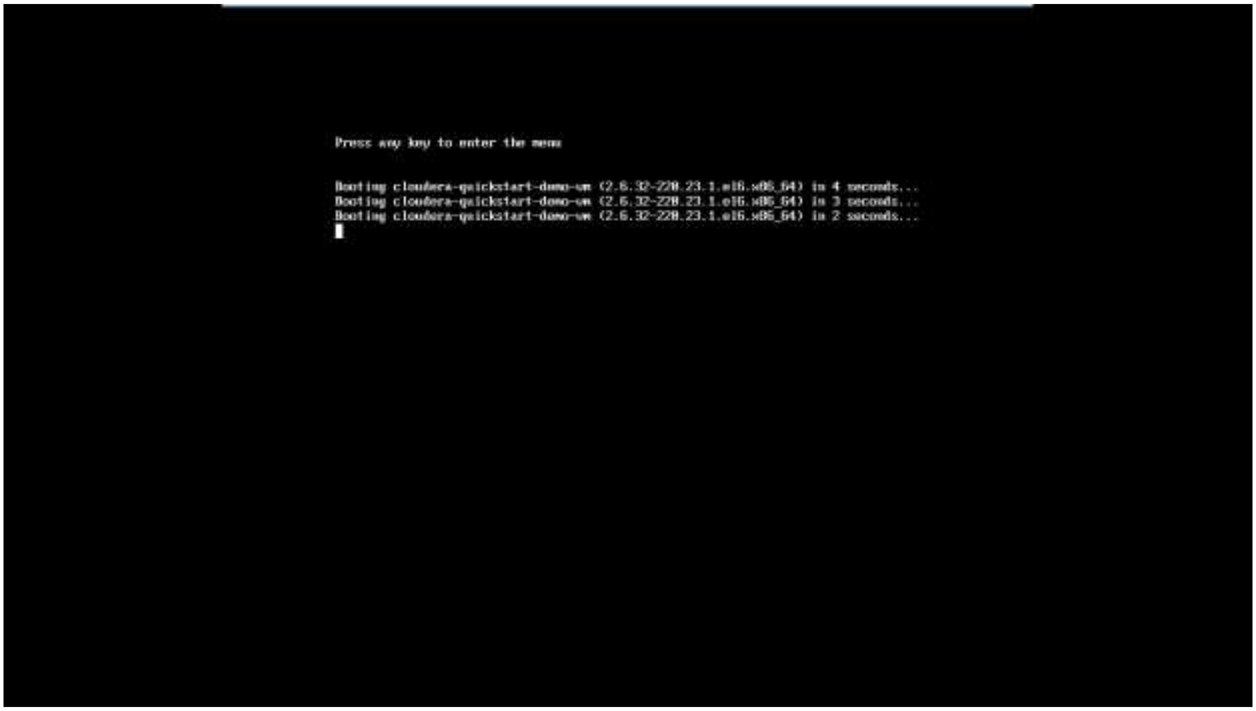
10. Once the VM is available, click on Play virtual machine.



cloudera-quickstart-demo-vm

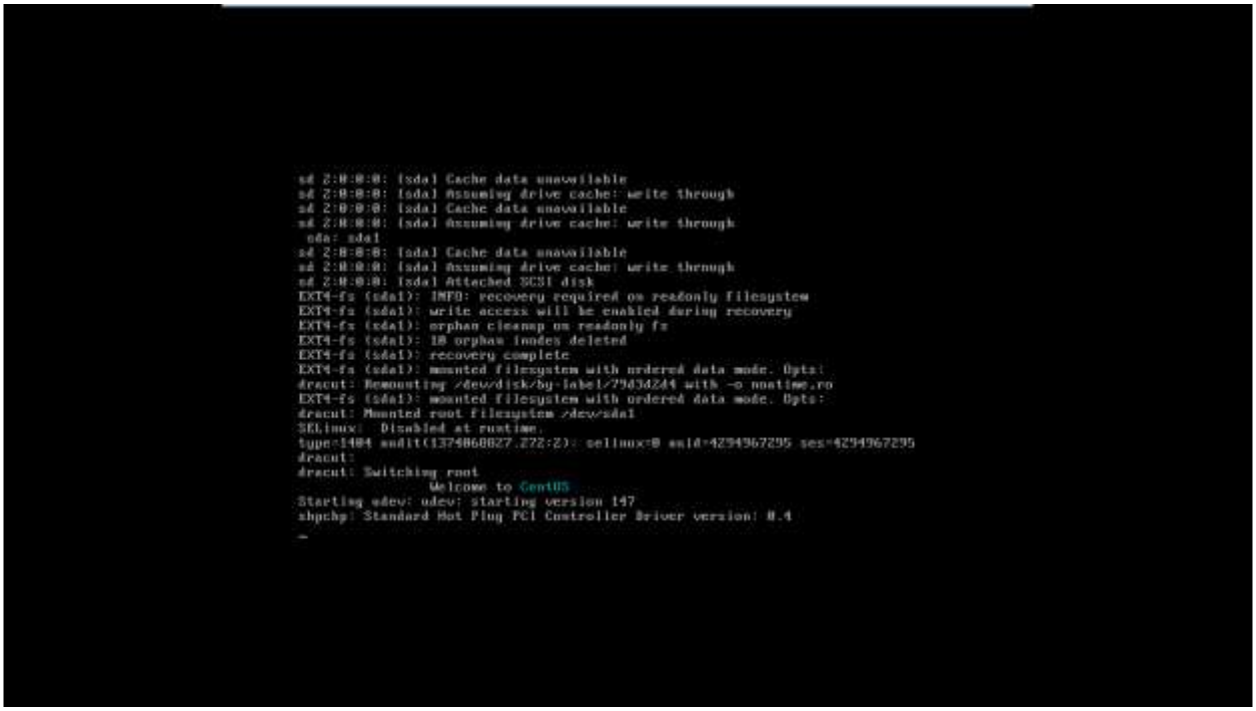
State: Powered Off
OS: Other Linux 64-bit
Version: Workstation 6.5-7.x virtual machine
RAM: 4 GB

11. As this is a new virtual machine, It will start deploying.



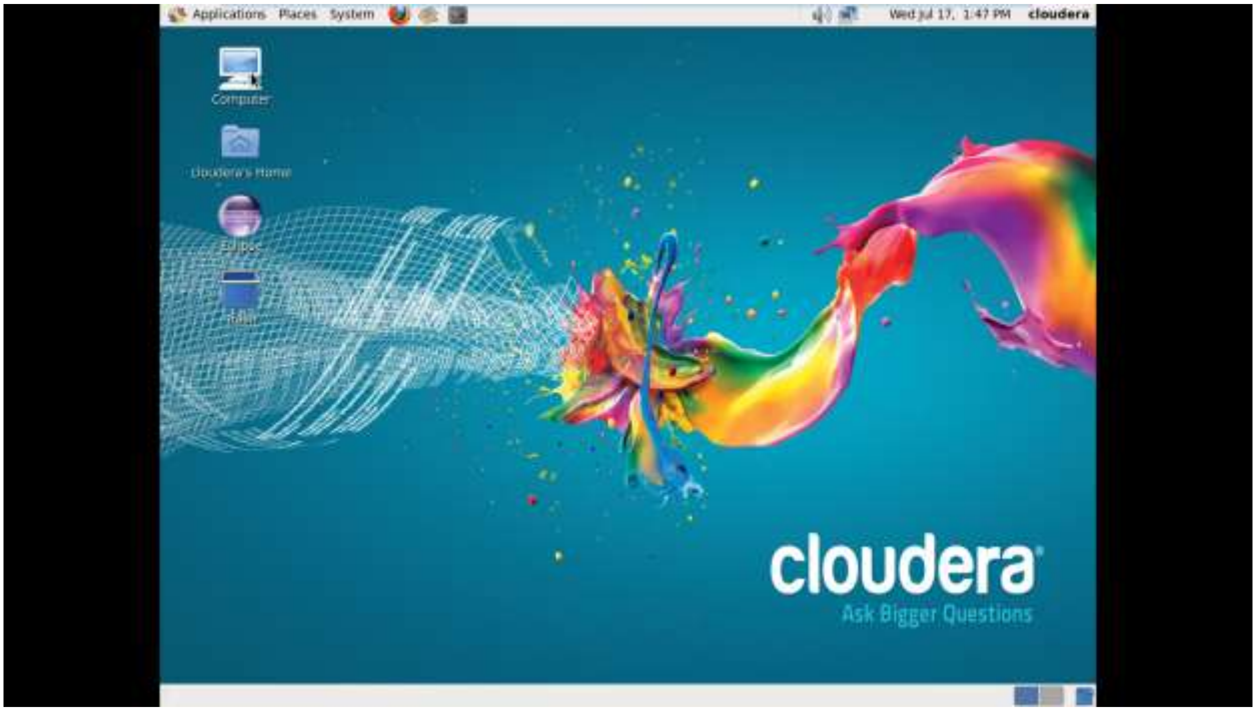
```
Press any key to enter the menu


Booting cloudera-quickstart-demo-vm (2.6.32-220.23.1.el6.x86_64) in 4 seconds...
Booting cloudera-quickstart-demo-vm (2.6.32-220.23.1.el6.x86_64) in 3 seconds...
Booting cloudera-quickstart-demo-vm (2.6.32-220.23.1.el6.x86_64) in 2 seconds...
█
```



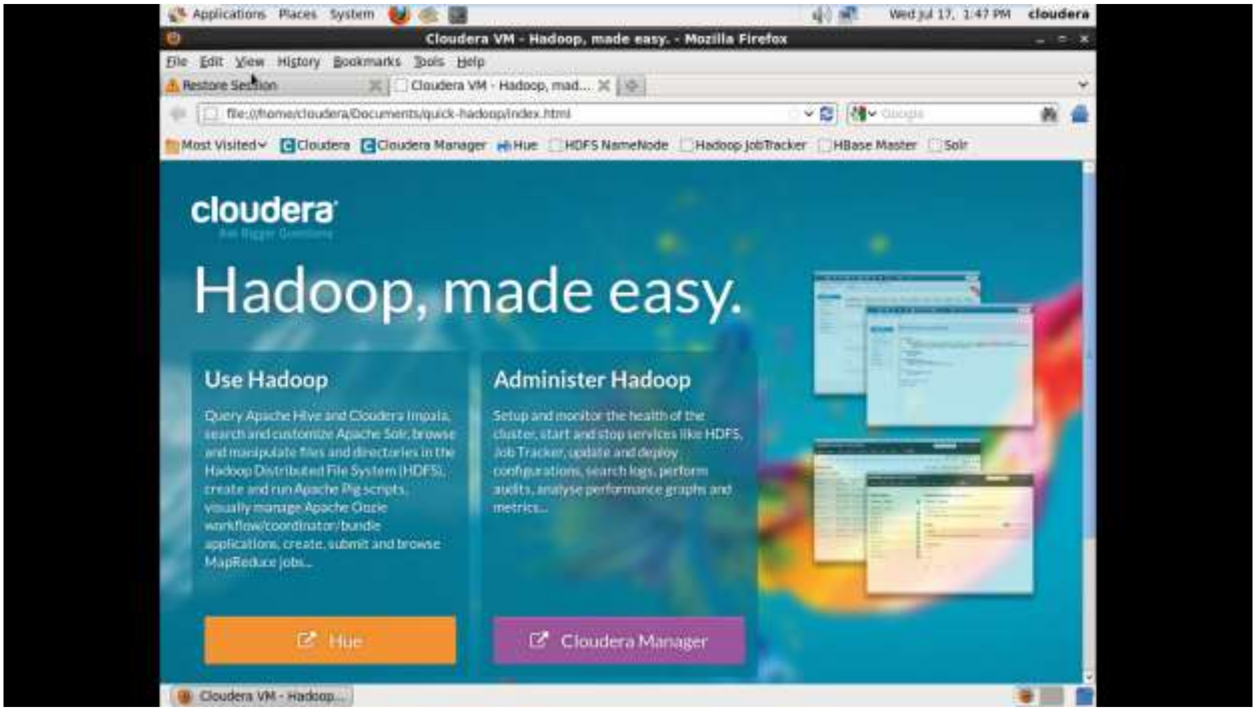
```
sd 2:0:0:0: [sda] Cache data unavailable
sd 2:0:0:0: [sda] Assuming drive cache: write through
sd 2:0:0:0: [sda] Cache data unavailable
sd 2:0:0:0: [sda] Assuming drive cache: write through
 sda: sda1
sd 2:0:0:0: [sda] Cache data unavailable
sd 2:0:0:0: [sda] Assuming drive cache: write through
sd 2:0:0:0: [sda] Attached SCSI disk
EXT4-fs (sda1): INFO: recovery required on readonly filesystem
EXT4-fs (sda1): write access will be enabled during recovery
EXT4-fs (sda1): orphan cleanup on readonly fs
EXT4-fs (sda1): 10 orphan inodes deleted
EXT4-fs (sda1): recovery complete
EXT4-fs (sda1): mounted filesystem with ordered data mode. Opts:
dracut: Remounting /dev/disk/by-label/79d3d2d4 with -o noatime,ro
EXT4-fs (sda1): mounted filesystem with ordered data mode. Opts:
dracut: Mounted root filesystem /dev/sda1
SELinux:  Disabled at runtime.
type=1404 audit(1374060027.272:2): selinux=0 auid=4294967295 ses=4294967295
dracut:
dracut: Switching root
                Welcome to CentOS
Starting udev: udev: starting version 147
shpchp: Standard Hot Plug PCI Controller Driver version: 0.4
_
```

12. After the installation/ deployment is complete, you get the virtual machine desktop.



Hadoop Configuration

1. Once the installation is complete, you will get below welcome page.



Applications Places System
Wed Jul 17, 1:47 PM cloudera
Cloudera VM - Hadoop, made easy. - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Restore Session
Cloudera VM - Hadoop, mad...
file:///home/cloudera/Documents/quick-hadoop/index.html
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
cloudera
Hue
Cloudera Manager
Cloudera VM - Hadoop...

Hadoop, made easy.

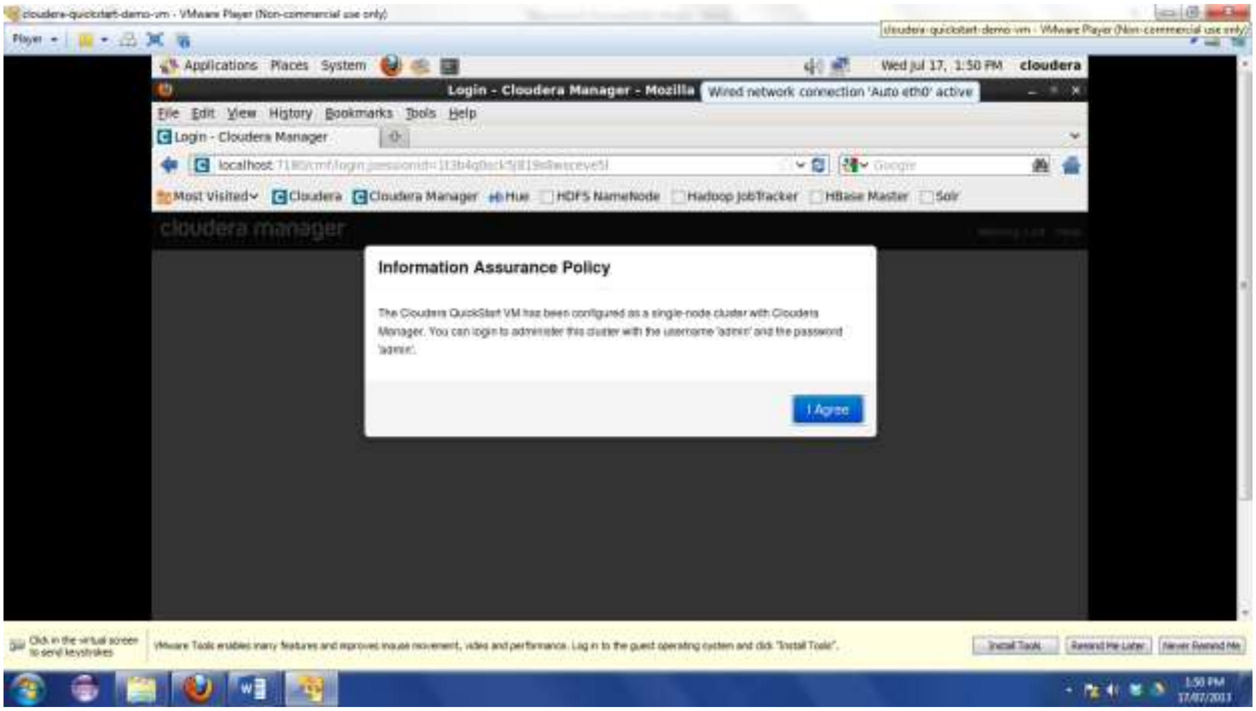
Use Hadoop

Query Apache Hive and Cloudera Impala, search and customize Apache Solr, browse and manipulate files and directories in the Hadoop Distributed File System (HDFS), create and run Apache Pig scripts, visually manage Apache Oozie workflow/coordinator/bundle applications, create, submit and browse MapReduce jobs...

Administer Hadoop

Setup and monitor the health of the cluster, start and stop services like HDFS, Job Tracker, update and deploy configurations, search logs, perform audits, analyse performance graphs and metrics...

2. Click on Cloudera Manager, as this is your first login, agree to Information Assurance policy.



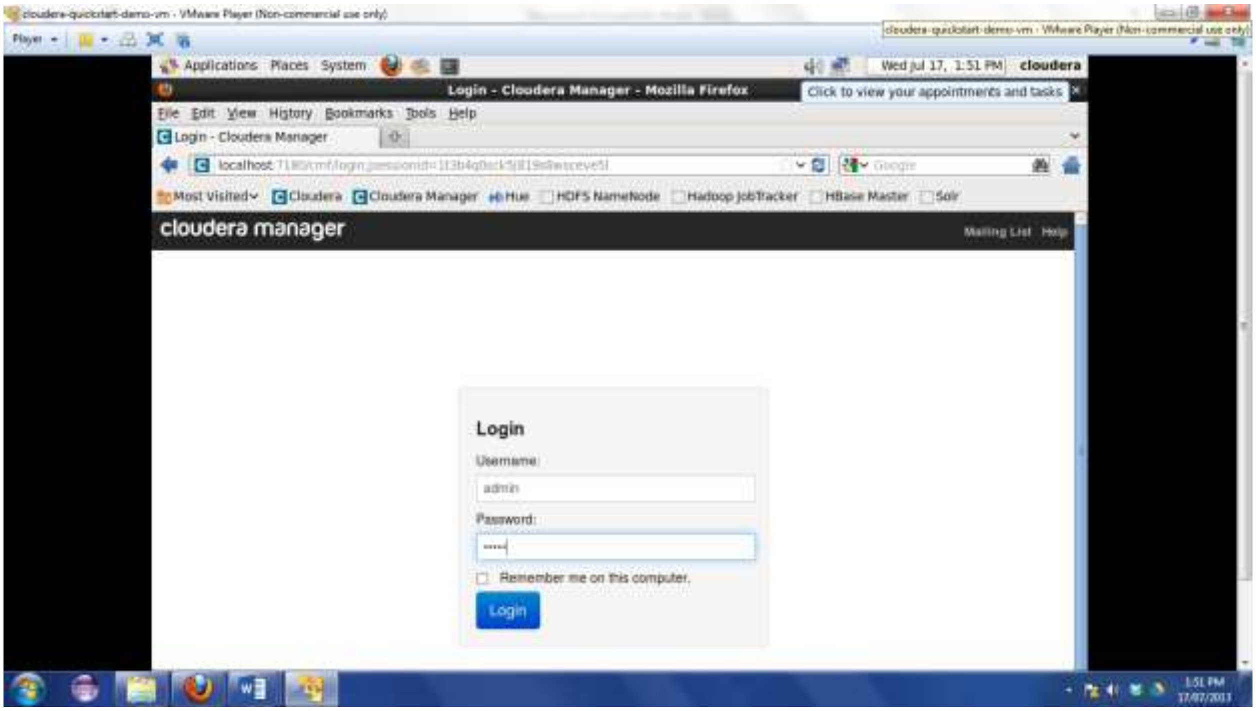
cloudera-quickstart-demo-vm - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 17, 1:50 PM cloudera
Login - Cloudera Manager - Mozilla
Wired network connection 'Auto eth0' active
File Edit View History Bookmarks Tools Help
Login - Cloudera Manager
localhost:7180/cmf/login;jsessionid=1t3b4q0eck5j819s8wsceve5l
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
cloudera manager
I Agree
Click in the virtual screen to send keystrokes
Install Tools
Remind Me Later
Never Remind Me
1:50 PM
17/07/2013

Information Assurance Policy

The Cloudera QuickStart VM has been configured as a single-node cluster with Cloudera Manager. You can login to administer this cluster with the username 'admin' and the password 'admin'.

VMware Tools enables many features and improves mouse movement, video and performance. Log in to the guest operating system and click "Install Tools".

3. Login to CDH using username: admin, password: admin.



cloudera manager

Login

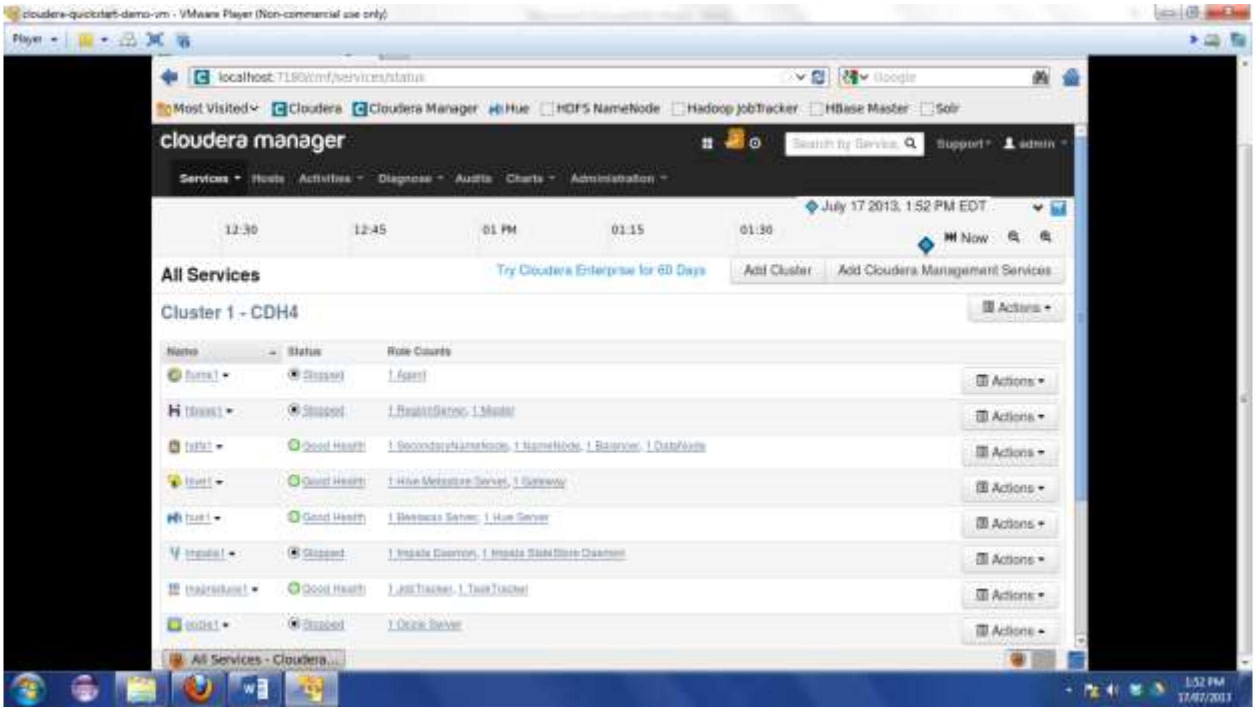
Username: admin

Password: •••••

□ Remember me on this computer.

Login

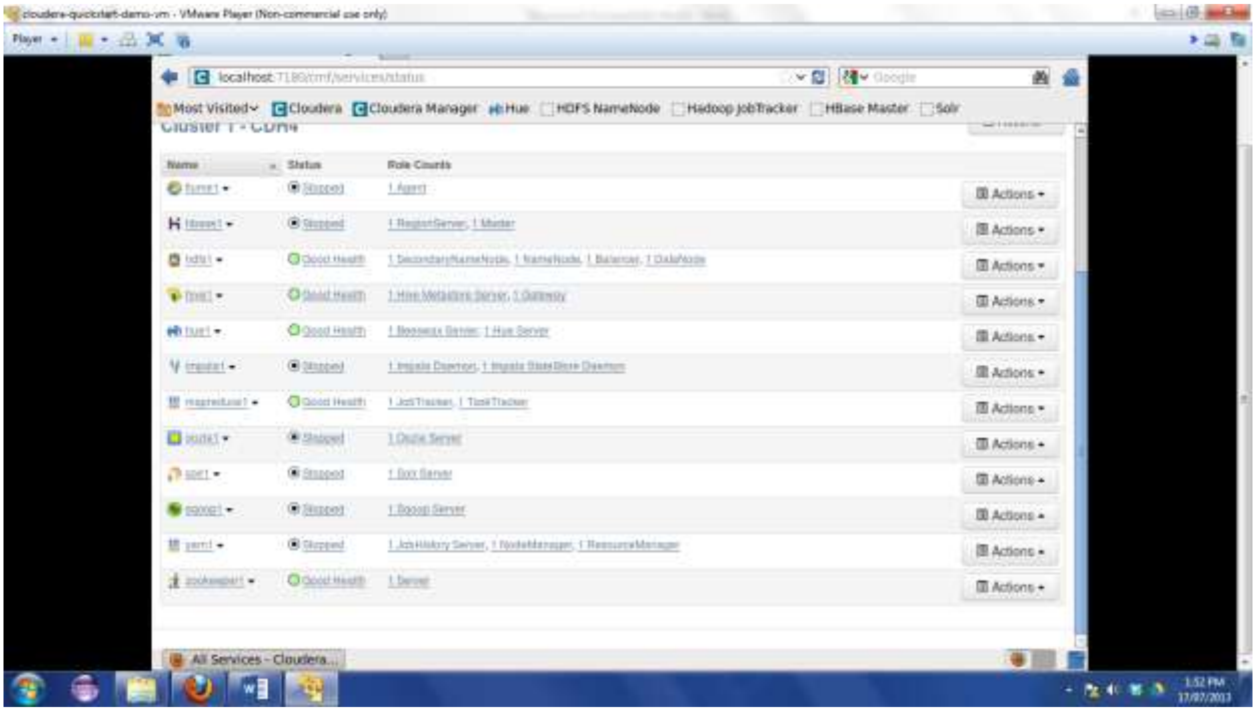
4. Once you login, you will get list of all services available. Check and confirm there is no unhealthy service that needs to be fixed.



All Services

Cluster 1 - CDH4

Name	Status	Role Counts
flume1	Stopped	1 Agent
hbase1	Stopped	1 RegionServer, 1 Master
hdfs1	Good Health	1 SecondaryNameNode, 1 NameNode, 1 Balancer, 1 DataNode
hive1	Good Health	1 Hive Metastore Server, 1 Gateway
hue1	Good Health	1 Beeswax Server, 1 Hue Server
impala1	Stopped	1 Impala Daemon, 1 Impala StateStore Daemon
mapreduce1	Good Health	1 JobTracker, 1 TaskTracker
oozie1	Stopped	1 Oozie Server

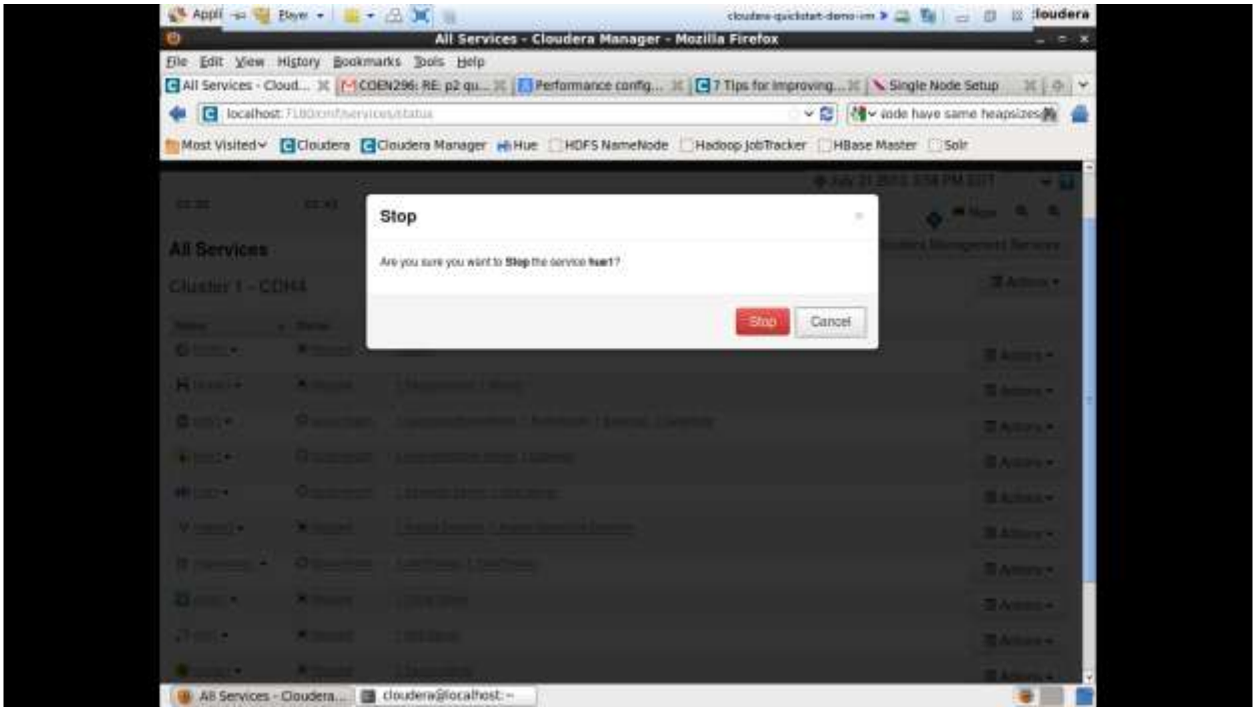


Cluster 1 - CDH4

Name	Status	Role Counts
flume1	Stopped	1 Agent
hbase1	Stopped	1 RegionServer, 1 Master
hdfs1	Good Health	1 SecondaryNameNode, 1 NameNode, 1 Balancer, 1 DataNode
hive1	Good Health	1 Hive Metastore Server, 1 Gateway
hue1	Good Health	1 Beeswax Server, 1 Hue Server
impala1	Stopped	1 Impala Daemon, 1 Impala StateStore Daemon
mapreduce1	Good Health	1 JobTracker, 1 TaskTracker
oozie1	Stopped	1 Oozie Server
solr1	Stopped	1 Solr Server
sqoop1	Stopped	1 Sqoop Server
yarn1	Stopped	1 JobHistory Server, 1 NodeManager, 1 ResourceManager
zookeeper1	Good Health	1 Server

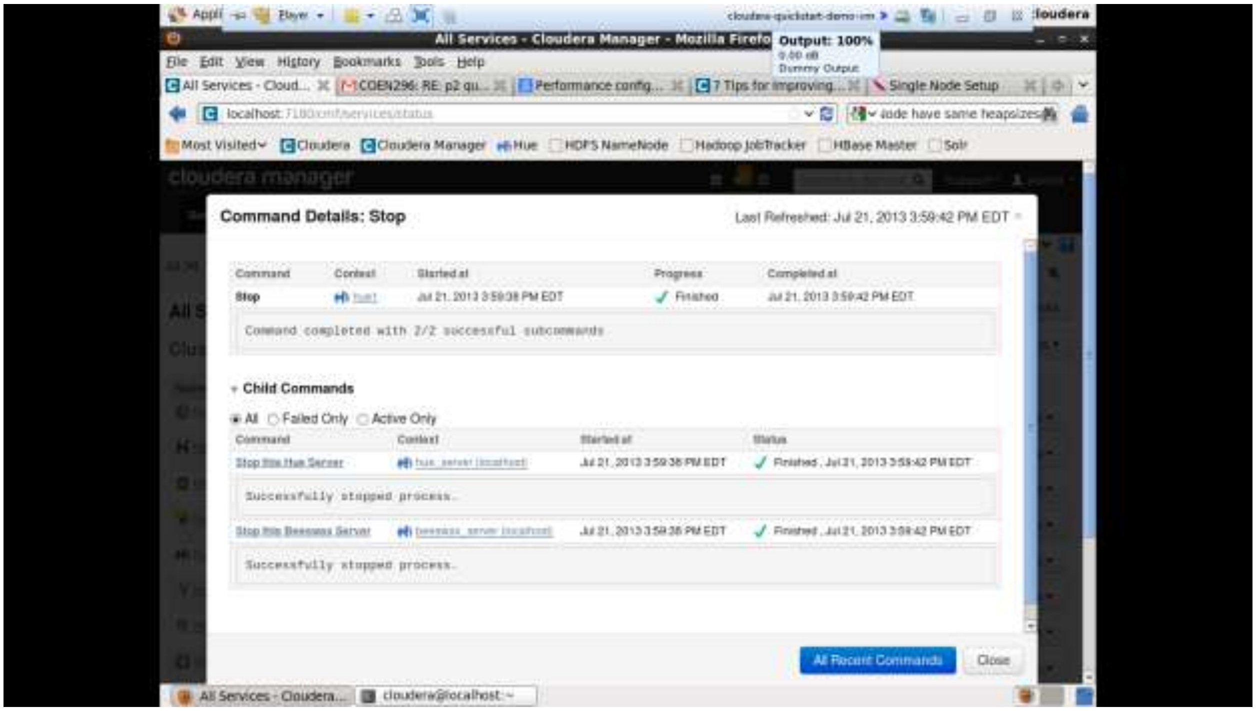
There are some services that are not required to be running to run the wordcount program. These services can be stopped for better performance.

5. Stop Hue as it is not required to run word count program.



Stop

Are you sure you want to Stop the service hue1?


All Services - Cloudera Manager - Mozilla Firefox
Output: 100%
File Edit View History Bookmarks Tools Help
localhost:7180/cmf/services/status
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
cloudera manager
All Recent Commands
Close

Command Details: Stop

Last Refreshed: Jul 21, 2013 3:59:42 PM EDT

Command	Context	Started at	Progress	Completed at
Stop	hue1	Jul 21, 2013 3:59:38 PM EDT	✓ Finished	Jul 21, 2013 3:59:42 PM EDT

```
Command completed with 2/2 successful subcommands
```

Child Commands

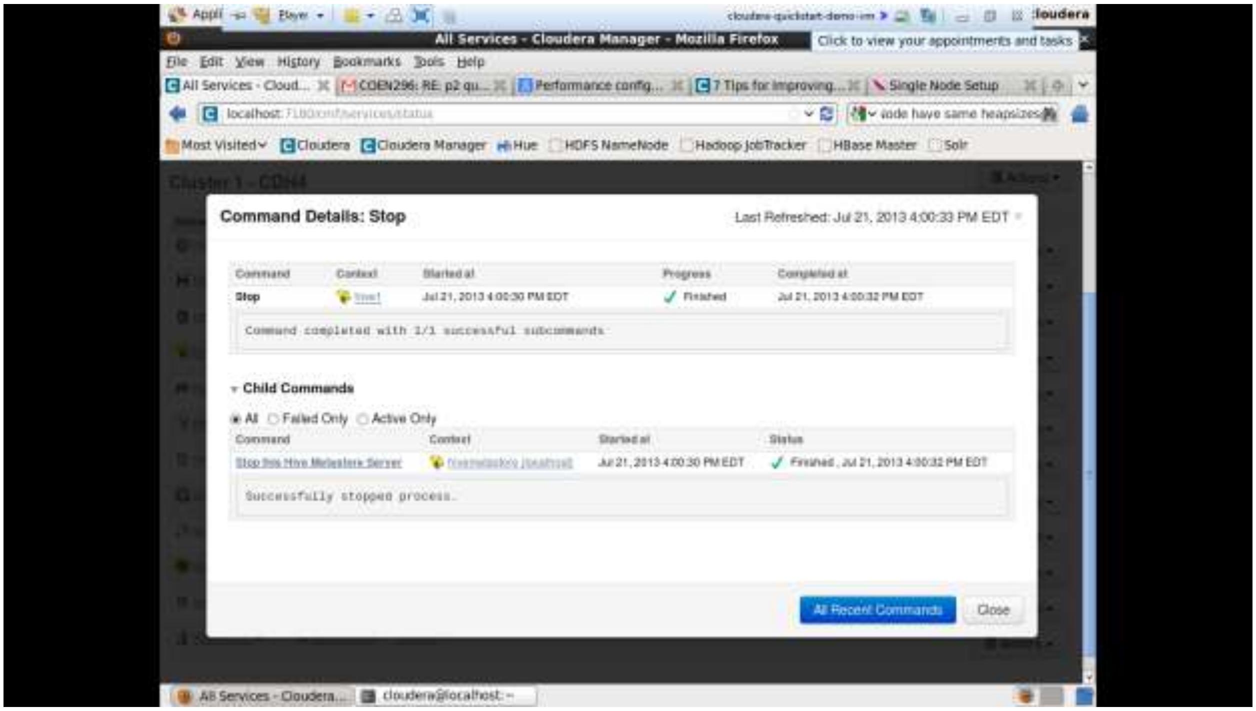
● All ○ Failed Only ○ Active Only

Command	Context	Started at	Status
Stop this Hue Server	hue_server (localhost)	Jul 21, 2013 3:59:38 PM EDT	✓ Finished , Jul 21, 2013 3:59:42 PM EDT
Stop this Beeswax Server	beeswax_server (localhost)	Jul 21, 2013 3:59:38 PM EDT	✓ Finished , Jul 21, 2013 3:59:42 PM EDT

```
Successfully stopped process.
```

```
Successfully stopped process.
```

6. Stop Hive


All Services - Cloudera Manager - Mozilla Firefox
Click to view your appointments and tasks
File Edit View History Bookmarks Tools Help
localhost:7180/cmf/services/status
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
Cluster 1 - CDH4
All Recent Commands
Close

Command Details: Stop

Last Refreshed: Jul 21, 2013 4:00:33 PM EDT

Command	Context	Started at	Progress	Completed at
Stop	hive1	Jul 21, 2013 4:00:30 PM EDT	✓ Finished	Jul 21, 2013 4:00:32 PM EDT

```
Command completed with 1/1 successful subcommands
```

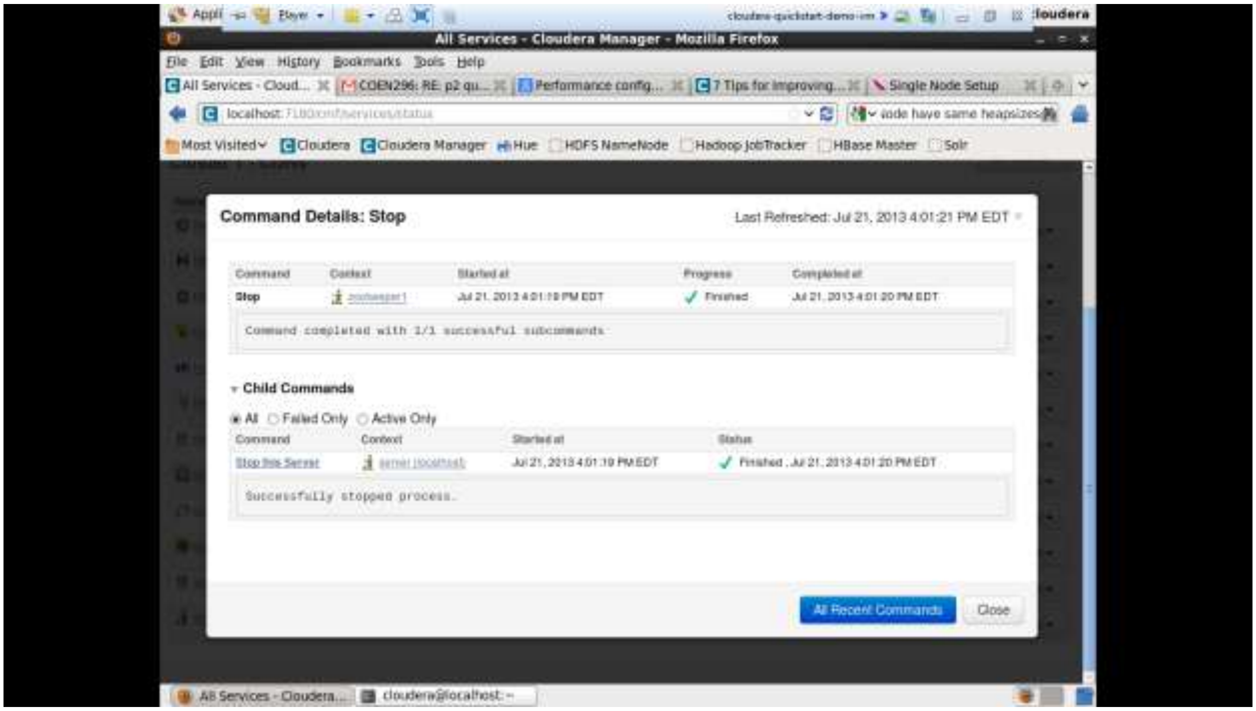
Child Commands

● All ○ Failed Only ○ Active Only

Command	Context	Started at	Status
Stop this Hive Metastore Server	hivemetastore (localhost)	Jul 21, 2013 4:00:30 PM EDT	✓ Finished , Jul 21, 2013 4:00:32 PM EDT

```
Successfully stopped process.
```

7. Stop service Zookeeper



Command Details: Stop

Last Refreshed: Jul 21, 2013 4:01:21 PM EDT

Command	Context	Started at	Progress	Completed at
Stop	zookeeper1	Jul 21, 2013 4:01:19 PM EDT	✓ Finished	Jul 21, 2013 4:01:20 PM EDT

```
Command completed with 1/1 successful subcommands
```

Child Commands

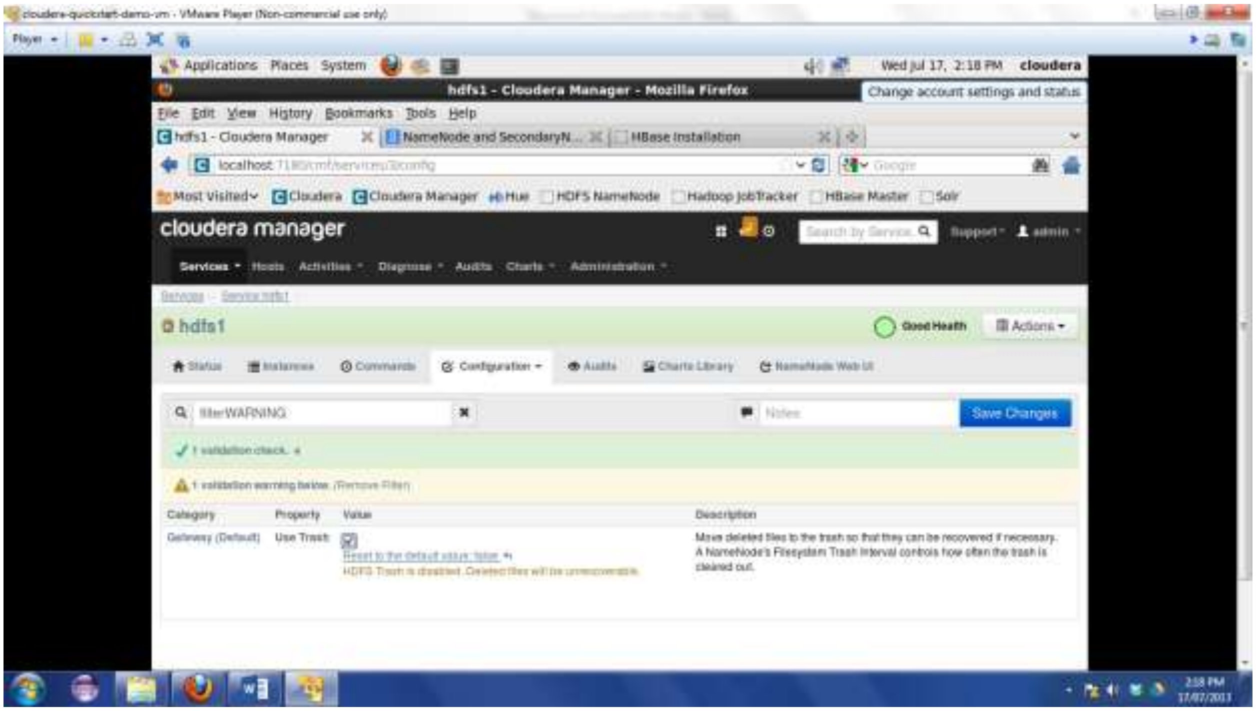
● All ○ Failed Only ○ Active Only

Command	Context	Started at	Status
Stop this Server	server (localhost)	Jul 21, 2013 4:01:19 PM EDT	✓ Finished , Jul 21, 2013 4:01:20 PM EDT

```
Successfully stopped process.
```

All Services - Cloudera Manager - Mozilla Firefox
localhost:7180/cmf/services/status
All Recent Commands
Close

8. Check Use trash property. This keeps a copy of every file that is deleted in _trash folder. Setting this property helps in better monitoring.



hdfs1

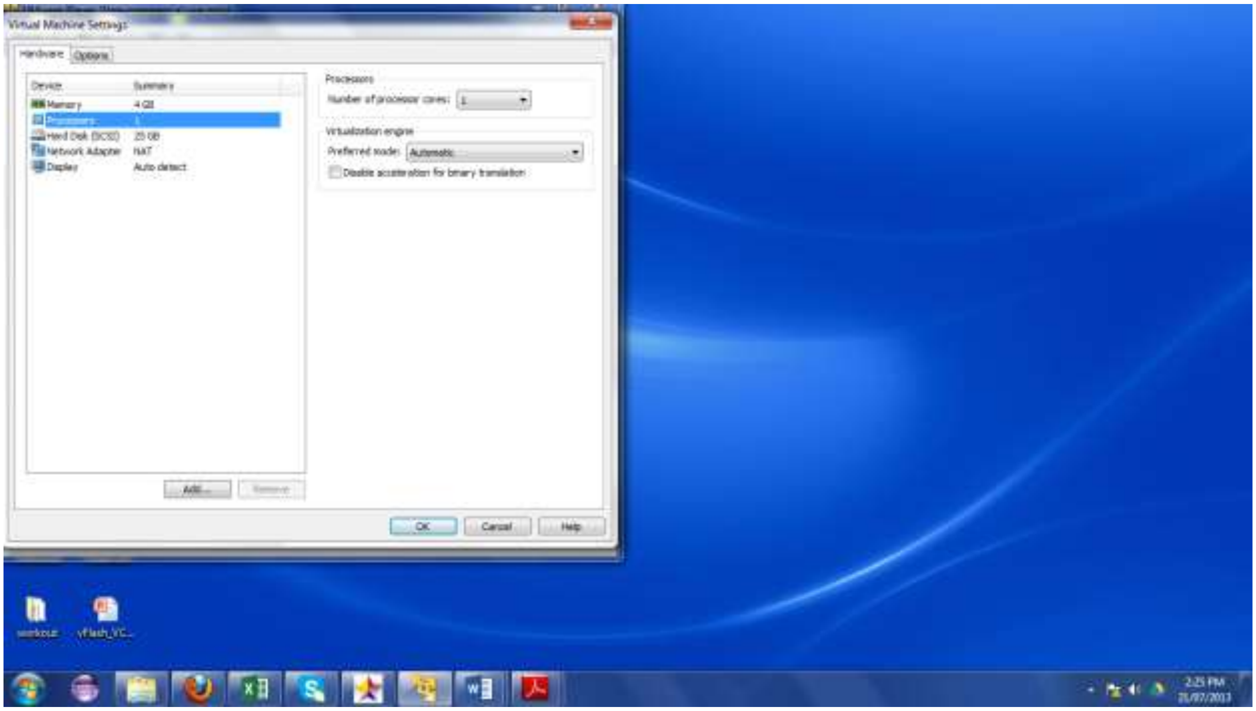
filterWARNING

✓ 1 validation check.

⚠ 1 validation warning below. (Remove Filter)

Category	Property	Value	Description
Gateway (Default)	Use Trash	☑ Reset to the default value: false HDFS Trash is disabled. Deleted files will be unrecoverable.	Move deleted files to the trash so that they can be recovered if necessary. A NameNode's Filesystem Trash Interval controls how often the trash is cleared out.

9. Also, By default each VM has a single processor, as shown in the below screenshot.



Virtual Machine Settings

Device	Summary
Memory	4 GB
Processors	1
Hard Disk (SCSI)	25 GB
Network Adapter	NAT
Display	Auto detect

Processors

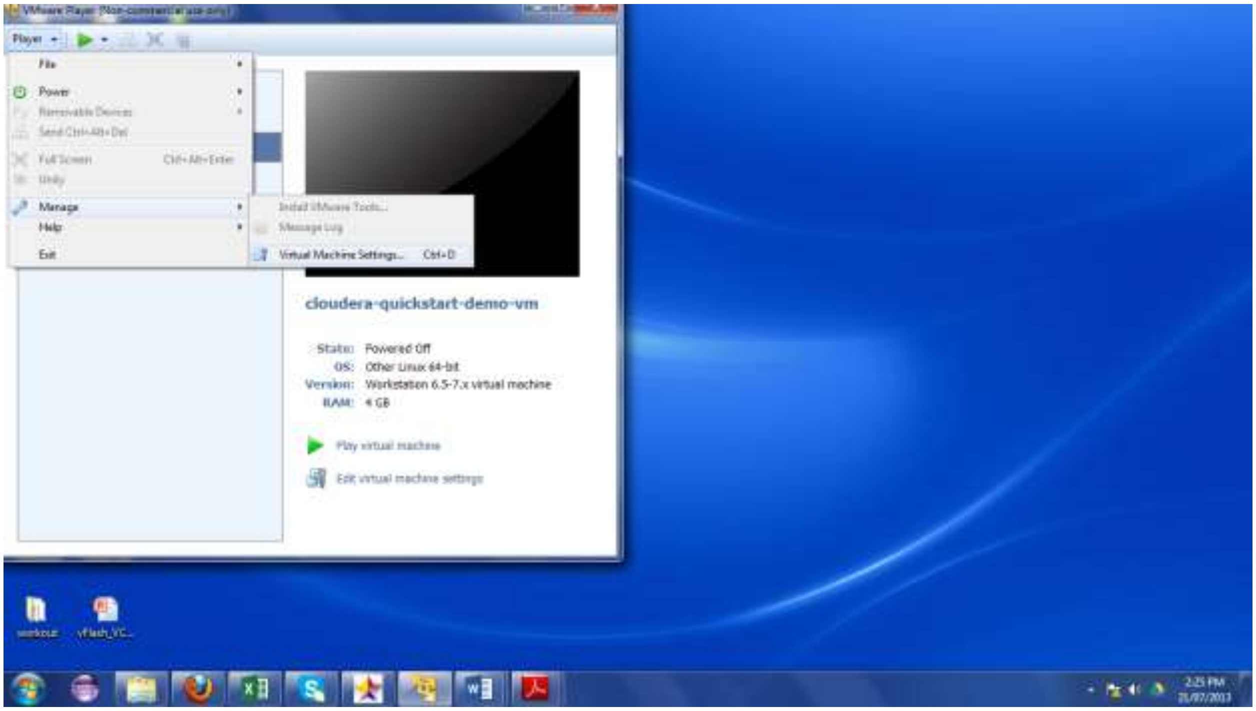
Number of processor cores: 1

Virtualization engine

Preferred mode: Automatic

□ Disable acceleration for binary translation

10. To enhance performance VM can be configured to be multi CPU. TO change the setting, Go to Player → Manage → Virtual Machine Settings…

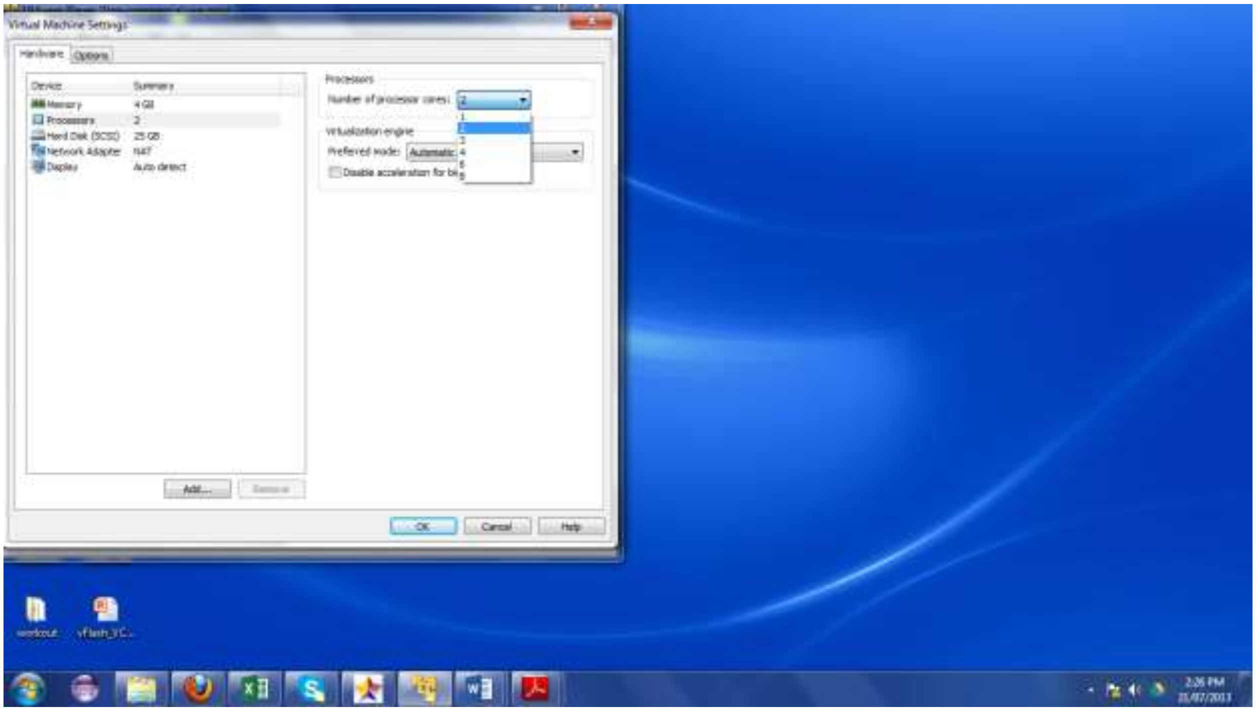


VMware Player (Non-commercial use only)

cloudera-quickstart-demo-vm

State: Powered Off
OS: Other Linux 64-bit
Version: Workstation 6.5-7.x virtual machine
RAM: 4 GB

11. Click on Processors and change the number to desired number. For our purposes, I have changed it to 2.



Virtual Machine Settings

Device	Summary
Memory	4 GB
Processors	2
Hard Disk (SCSI)	25 GB
Network Adapter	NAT
Display	Auto detect

Processors

Number of processor cores: 2

Virtualization engine

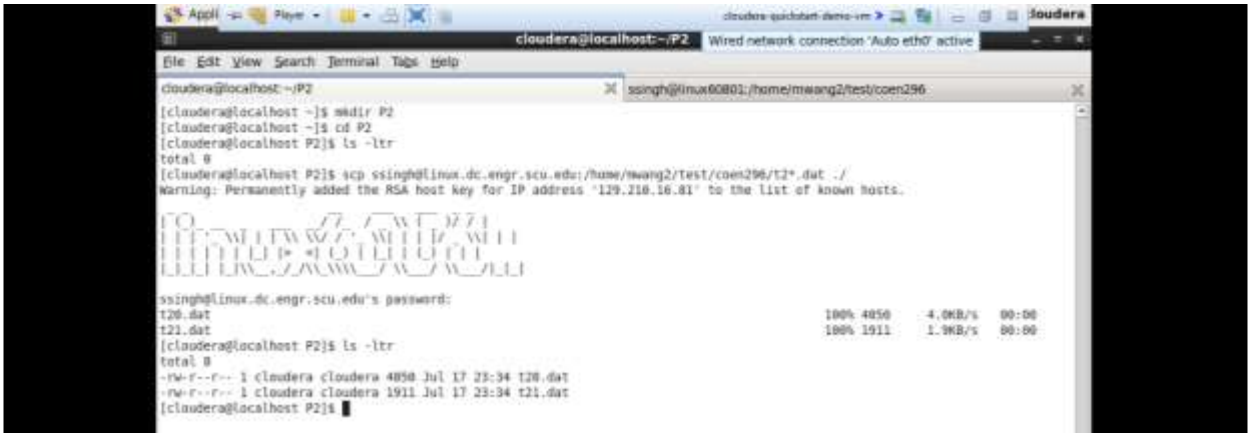
Preferred mode: Automatic

□ Disable acceleration for bi

Running a prgoram on Hadoop

1. Create a directory P2 and copy the sample program files to current working directory.

Command: mkdir P2
cd P2
ls –ltr
scp ssingh@linux.dc.engr/scu/edu:/home/mwang2/test/coen296/t2*.dat ./
ls -ltr

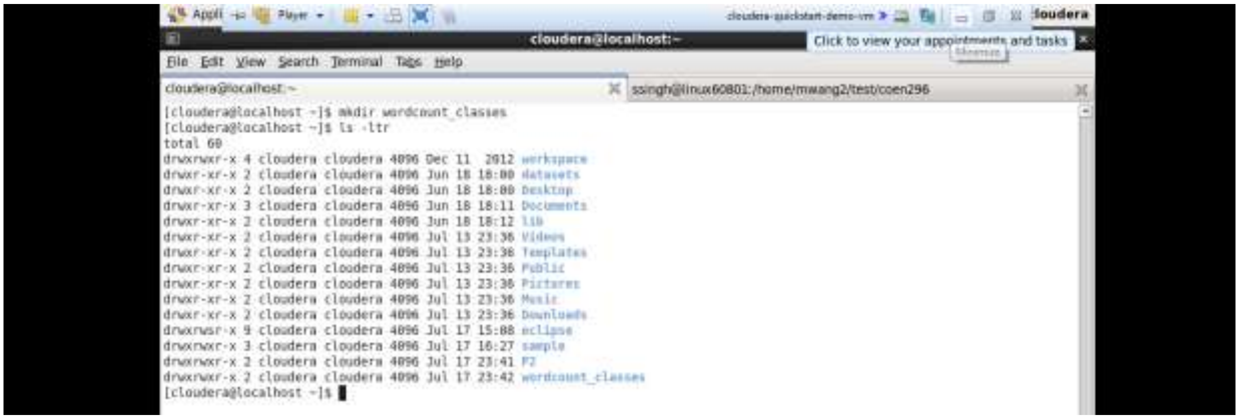

Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~/P2
Wired network connection 'Auto eth0' active
File Edit View Search Terminal Tabs Help
cloudera@localhost:~/P2
ssingh@linux60801:/home/mwang2/test/coen296

```
[cloudera@localhost ~]$ mkdir P2
[cloudera@localhost ~]$ cd P2
[cloudera@localhost P2]$ ls -ltr
total 0
[cloudera@localhost P2]$ scp ssingh@linux.dc.engr.scu.edu:/home/mwang2/test/coen296/t2*.dat ./
Warning: Permanently added the RSA host key for IP address '129.210.16.81' to the list of known hosts.

ssingh@linux.dc.engr.scu.edu's password:
t20.dat                                                                   100% 4050     4.0KB/s   00:00
t21.dat                                                                   100% 1911     1.9KB/s   00:00
[cloudera@localhost P2]$ ls -ltr
total 8
-rw-r--r-- 1 cloudera cloudera 4050 Jul 17 23:34 t20.dat
-rw-r--r-- 1 cloudera cloudera 1911 Jul 17 23:34 t21.dat
[cloudera@localhost P2]$
```

2. Create a directory wordcount_classes that will contain .jar file of the program.

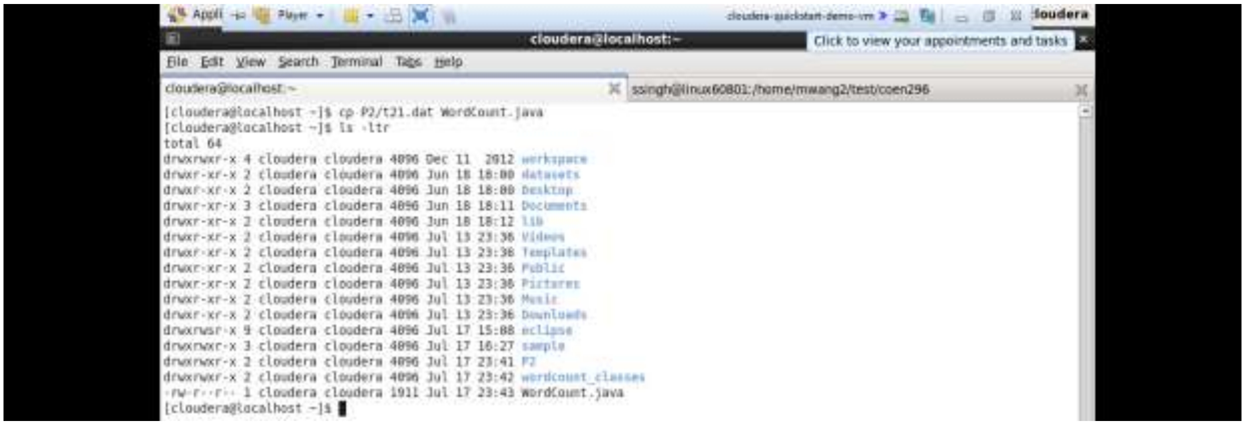
Command: cd ..
mkdir wordcount_classes
ls -ltr


Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296

```
[cloudera@localhost ~]$ mkdir wordcount_classes
[cloudera@localhost ~]$ ls -ltr
total 60
drwxrwxr-x 4 cloudera cloudera 4096 Dec 11  2012 workspace
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:00 datasets
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:00 Desktop
drwxr-xr-x 3 cloudera cloudera 4096 Jun 18 18:11 Documents
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:12 lib
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Videos
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Templates
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Public
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Pictures
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Music
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Downloads
drwxrwsr-x 9 cloudera cloudera 4096 Jul 17 15:08 eclipse
drwxrwxr-x 3 cloudera cloudera 4096 Jul 17 16:27 sample
drwxrwxr-x 2 cloudera cloudera 4096 Jul 17 23:41 P2
drwxrwxr-x 2 cloudera cloudera 4096 Jul 17 23:42 wordcount_classes
[cloudera@localhost ~]$
```

3. Copy one of the program files to WordCount.java file

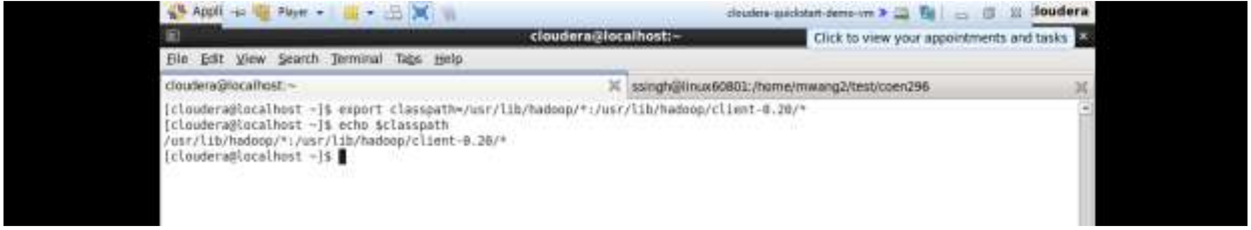
Command: cp P2/t21.dat WordCount.java
ls -ltr


Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296

```
[cloudera@localhost ~]$ cp P2/t21.dat WordCount.java
[cloudera@localhost ~]$ ls -ltr
total 64
drwxrwxr-x 4 cloudera cloudera 4096 Dec 11  2012 workspace
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:00 datasets
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:00 Desktop
drwxr-xr-x 3 cloudera cloudera 4096 Jun 18 18:11 Documents
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:12 lib
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Videos
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Templates
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Public
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Pictures
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Music
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Downloads
drwxrwsr-x 9 cloudera cloudera 4096 Jul 17 15:08 eclipse
drwxrwxr-x 3 cloudera cloudera 4096 Jul 17 16:27 sample
drwxrwxr-x 2 cloudera cloudera 4096 Jul 17 23:41 P2
drwxrwxr-x 2 cloudera cloudera 4096 Jul 17 23:42 wordcount_classes
-rw-r--r-- 1 cloudera cloudera 1911 Jul 17 23:43 WordCount.java
[cloudera@localhost ~]$
```

4. Set classpath variable to access Hadoop libraries.

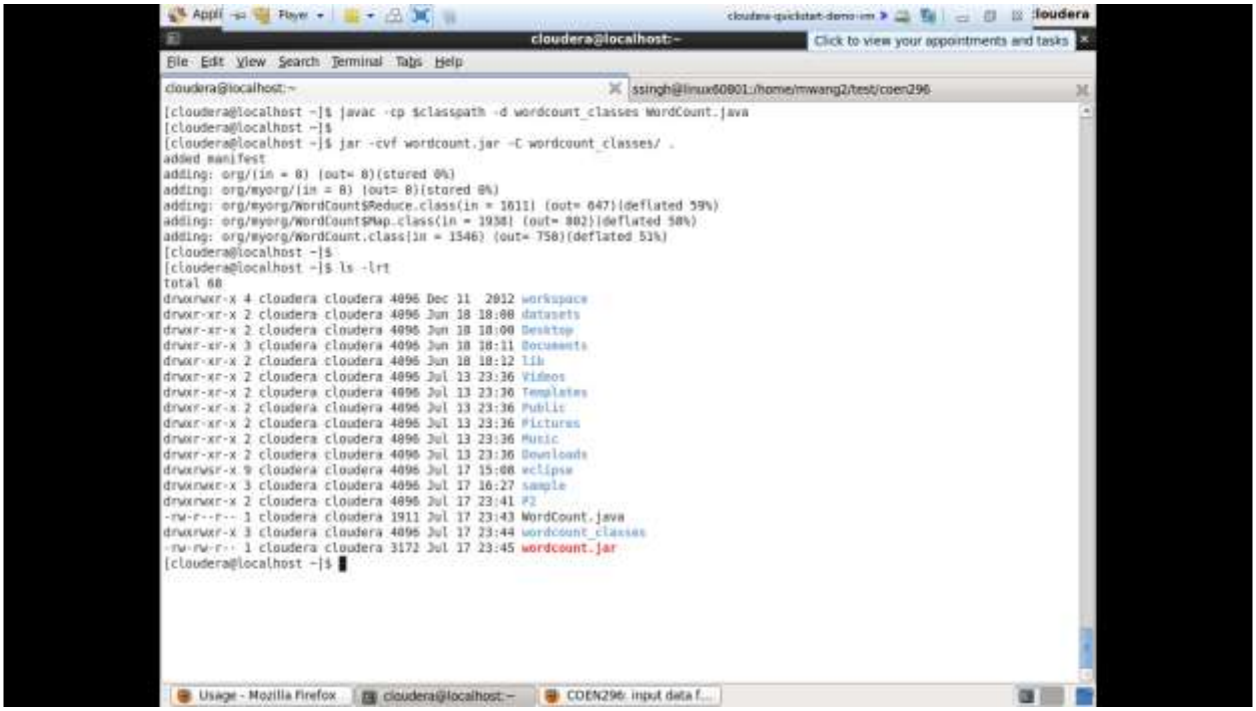
Command:
Export classpath= “/usr/lib/hadoop/*:/usr/lib/hadoop/client-0.20/*”
Echo $classpath


Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296

```
[cloudera@localhost ~]$ export classpath=/usr/lib/hadoop/*:/usr/lib/hadoop/client-0.20/*
[cloudera@localhost ~]$ echo $classpath
/usr/lib/hadoop/*:/usr/lib/hadoop/client-0.20/*
[cloudera@localhost ~]$
```

5. Compile WordCount.java and save class files in wordcount_classes directory. Create a JAR file for all classes.

Command:
javac -cp $*classpath* -d wordcount_classes WordCount.java
jar -cvf wordcount.jar -C wordcount_classes/ .
ls -lrt

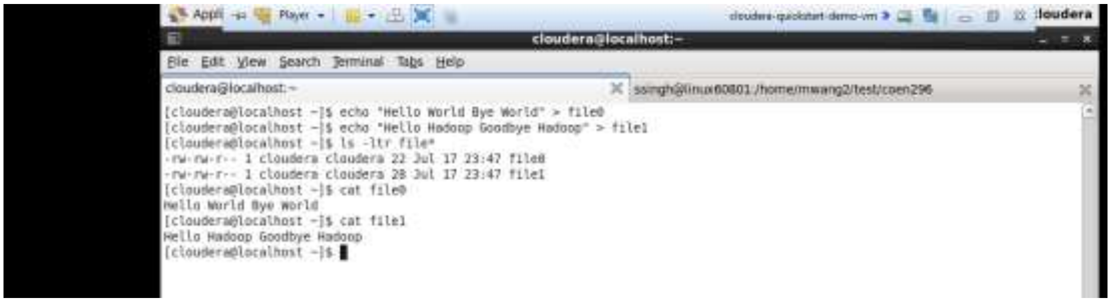


```
[cloudera@localhost ~]$ javac -cp $classpath -d wordcount_classes WordCount.java
[cloudera@localhost ~]$
[cloudera@localhost ~]$ jar -cvf wordcount.jar -C wordcount_classes/ .
added manifest
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/(in = 0) (out= 0)(stored 0%)
adding: org/myorg/WordCount$Reduce.class(in = 1611) (out= 647)(deflated 59%)
adding: org/myorg/WordCount$Map.class(in = 1938) (out= 802)(deflated 58%)
adding: org/myorg/WordCount.class(in = 1546) (out= 758)(deflated 51%)
[cloudera@localhost ~]$
[cloudera@localhost ~]$ ls -lrt
total 68
drwxrwxr-x 4 cloudera cloudera 4096 Dec 11  2012 workspace
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:08 datasets
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:08 Desktop
drwxr-xr-x 3 cloudera cloudera 4096 Jun 18 18:11 Documents
drwxr-xr-x 2 cloudera cloudera 4096 Jun 18 18:12 lib
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Videos
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Templates
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Public
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Pictures
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Music
drwxr-xr-x 2 cloudera cloudera 4096 Jul 13 23:36 Downloads
drwxrwsr-x 9 cloudera cloudera 4096 Jul 17 15:08 eclipse
drwxrwxr-x 3 cloudera cloudera 4096 Jul 17 16:27 sample
drwxrwxr-x 2 cloudera cloudera 4096 Jul 17 23:41 P2
-rw-r--r-- 1 cloudera cloudera 1911 Jul 17 23:43 WordCount.java
drwxrwxr-x 3 cloudera cloudera 4096 Jul 17 23:44 wordcount_classes
-rw-rw-r-- 1 cloudera cloudera 3172 Jul 17 23:45 wordcount.jar
[cloudera@localhost ~]$
```

Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296
Usage - Mozilla Firefox
cloudera@localhost:~
COEN296: input data f...

6. Create 2 sample files – file0 and file1.

Command:
echo "Hello World Bye World" > file0
echo "Hello Hadoop Goodbye Hadoop" > file1
ls –ltr file*
cat file0
cat file1

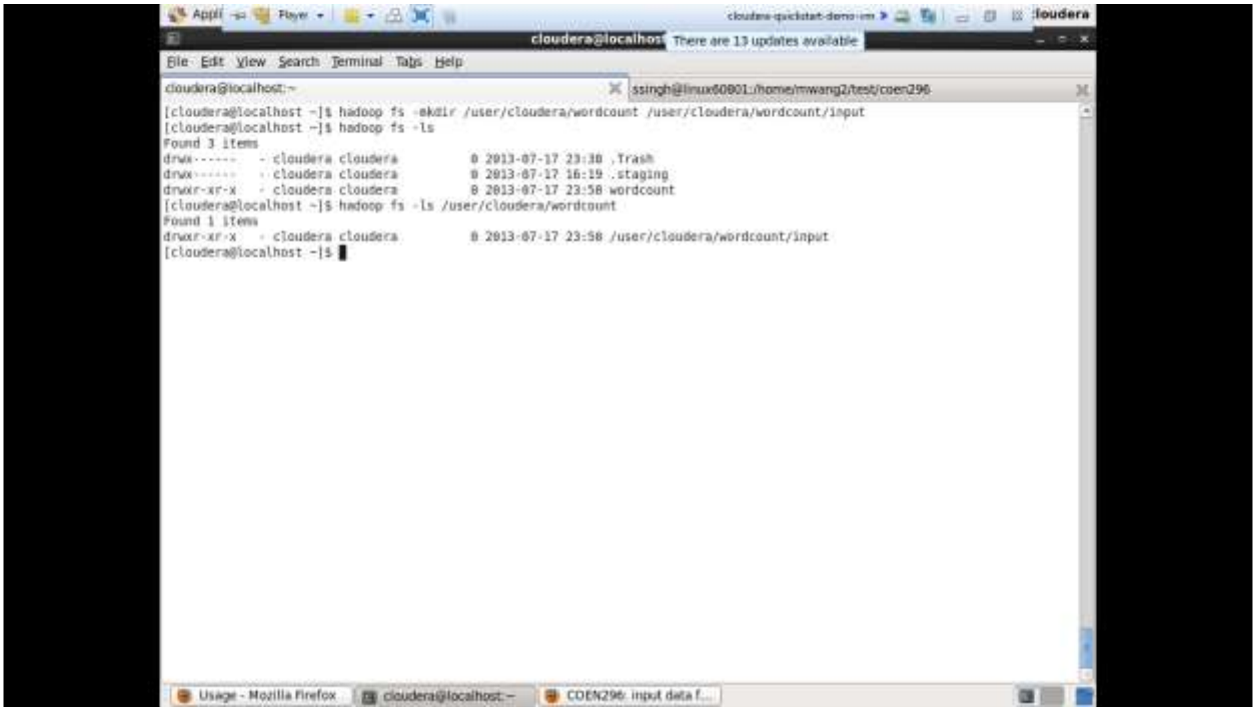


```
[cloudera@localhost ~]$ echo "Hello World Bye World" > file0
[cloudera@localhost ~]$ echo "Hello Hadoop Goodbye Hadoop" > file1
[cloudera@localhost ~]$ ls -ltr file*
-rw-rw-r-- 1 cloudera cloudera 22 Jul 17 23:47 file0
-rw-rw-r-- 1 cloudera cloudera 28 Jul 17 23:47 file1
[cloudera@localhost ~]$ cat file0
Hello World Bye World
[cloudera@localhost ~]$ cat file1
Hello Hadoop Goodbye Hadoop
[cloudera@localhost ~]$
```

Appli
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296

7. Create a wordcount directory and wordcount/input directory at Hadoop server.

Command:
hadoop fs -mkdir /user/cloudera/wordcount /user/cloudera/wordcount/input
hadoop fs -ls /user/cloudera/wordcount
hadoop fs -ls /user/cloudera/wordcount/input

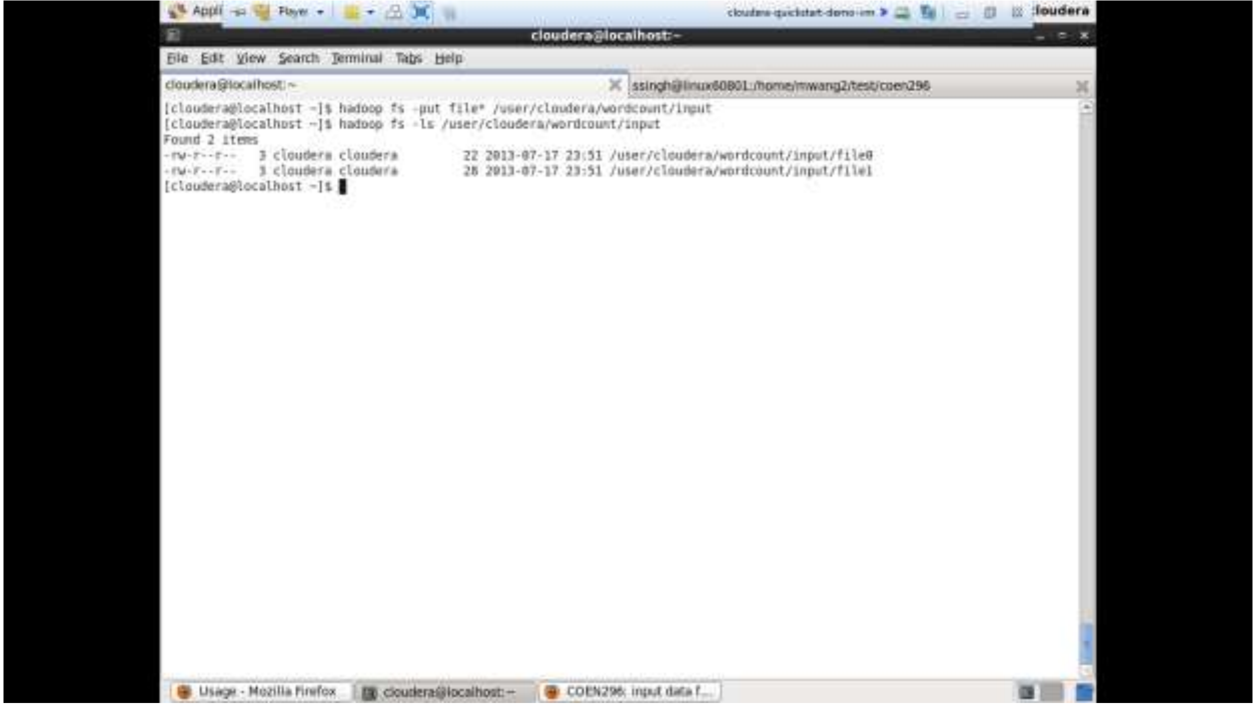


Appl
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost
There are 13 updates available
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296
Usage - Mozilla Firefox
cloudera@localhost:~
COEN296: input data f...

```
[cloudera@localhost ~]$ hadoop fs -mkdir /user/cloudera/wordcount /user/cloudera/wordcount/input
[cloudera@localhost ~]$ hadoop fs -ls
Found 3 items
drwx------   - cloudera cloudera          0 2013-07-17 23:30 .Trash
drwx------   - cloudera cloudera          0 2013-07-17 16:19 .staging
drwxr-xr-x   - cloudera cloudera          0 2013-07-17 23:50 wordcount
[cloudera@localhost ~]$ hadoop fs -ls /user/cloudera/wordcount
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2013-07-17 23:50 /user/cloudera/wordcount/input
[cloudera@localhost ~]$ 
```

8. Copy the files file0 and file1 to the Hadoop input files directory /user/cloudera/wordcount/input.

Command:
hadoop fs -put file* /user/cloudera/wordcount/input
hadoop fs -ls /user/cloudera/wordcount/input

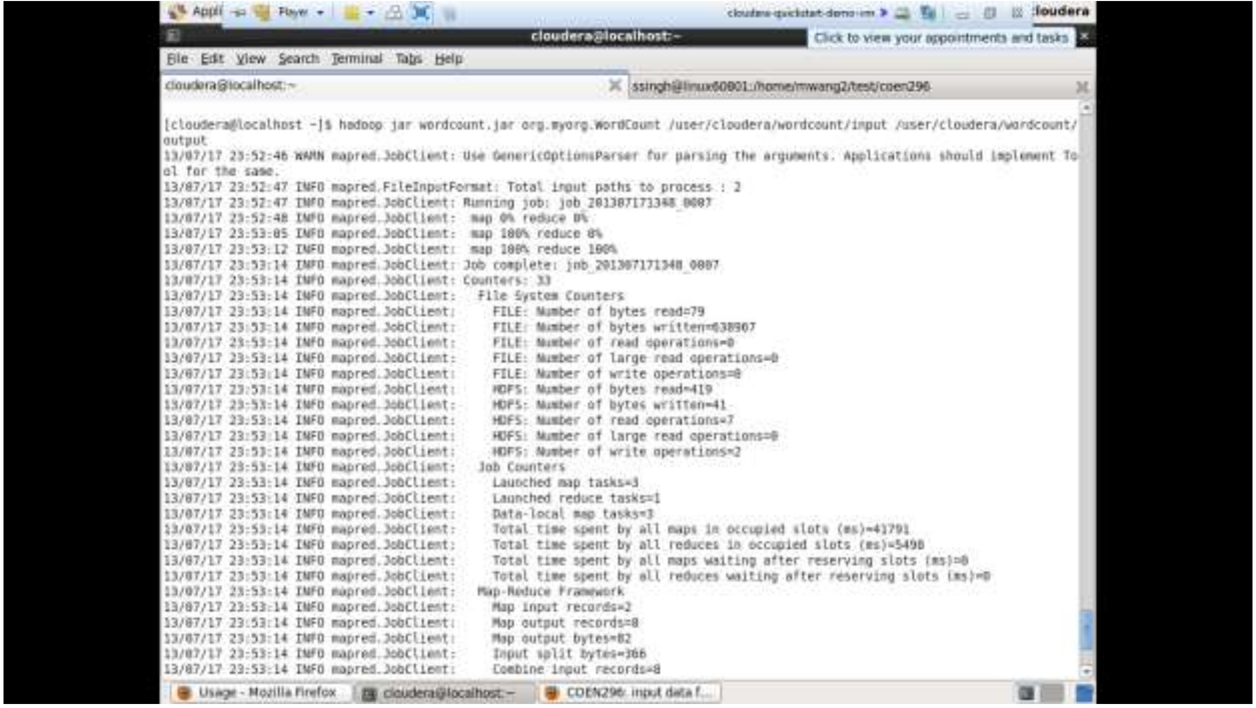


```
[cloudera@localhost ~]$ hadoop fs -put file* /user/cloudera/wordcount/input
[cloudera@localhost ~]$ hadoop fs -ls /user/cloudera/wordcount/input
Found 2 items
-rw-r--r--   3 cloudera cloudera         22 2013-07-17 23:51 /user/cloudera/wordcount/input/file0
-rw-r--r--   3 cloudera cloudera         28 2013-07-17 23:51 /user/cloudera/wordcount/input/file1
[cloudera@localhost ~]$ 
```

9. Run the jar file wordcount.jar, reading input files from /user/cloudera/wordcount/input and providing output at /user/cloudera/wordcount/output.

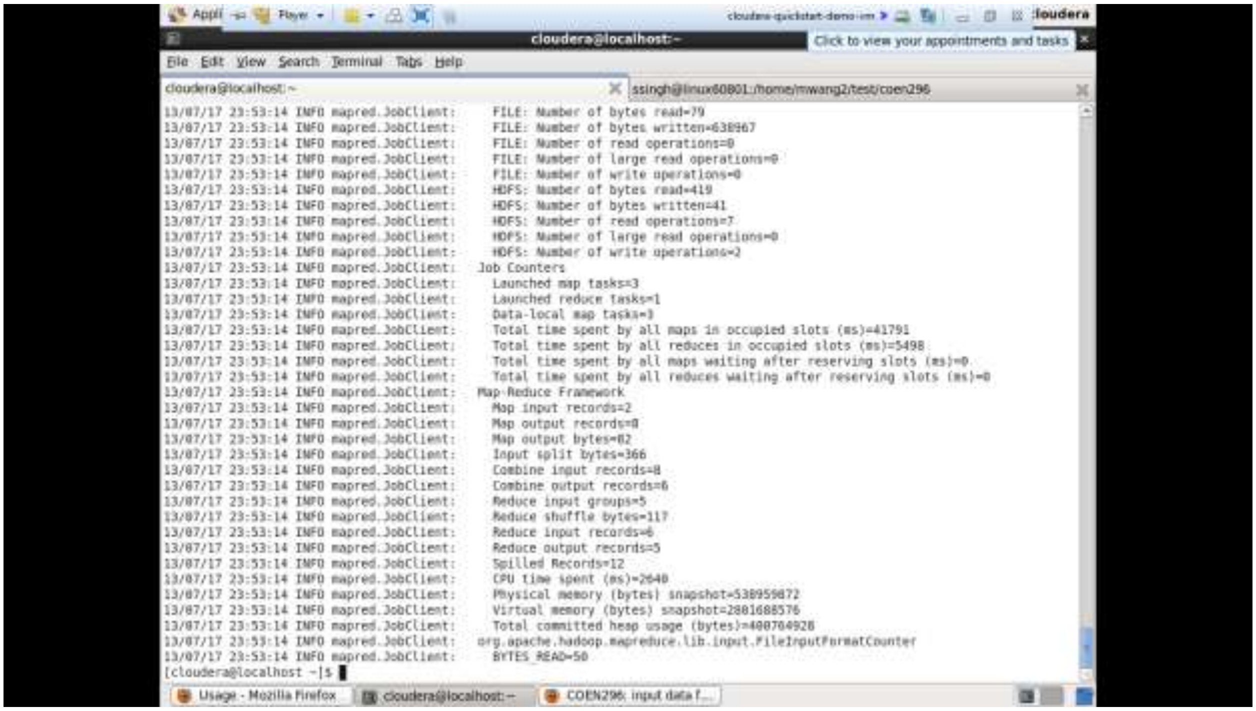
Command:

```
hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input
/user/cloudera/wordcount/output
```



Appl
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296
Usage - Mozilla Firefox
cloudera@localhost:~
COEN296: input data f...

```
[cloudera@localhost ~]$ hadoop jar wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input /user/cloudera/wordcount/
output
13/07/17 23:52:46 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement To
ol for the same.
13/07/17 23:52:47 INFO mapred.FileInputFormat: Total input paths to process : 2
13/07/17 23:52:47 INFO mapred.JobClient: Running job: job_201307171348_0007
13/07/17 23:52:48 INFO mapred.JobClient:  map 0% reduce 0%
13/07/17 23:53:05 INFO mapred.JobClient:  map 100% reduce 0%
13/07/17 23:53:12 INFO mapred.JobClient:  map 100% reduce 100%
13/07/17 23:53:14 INFO mapred.JobClient: Job complete: job_201307171348_0007
13/07/17 23:53:14 INFO mapred.JobClient: Counters: 33
13/07/17 23:53:14 INFO mapred.JobClient:   File System Counters
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of bytes read=79
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of bytes written=638907
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of large read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of write operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of bytes read=419
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of bytes written=41
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of read operations=7
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of large read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of write operations=2
13/07/17 23:53:14 INFO mapred.JobClient:   Job Counters
13/07/17 23:53:14 INFO mapred.JobClient:     Launched map tasks=3
13/07/17 23:53:14 INFO mapred.JobClient:     Launched reduce tasks=1
13/07/17 23:53:14 INFO mapred.JobClient:     Data-local map tasks=3
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=41791
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=5498
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/07/17 23:53:14 INFO mapred.JobClient:   Map-Reduce Framework
13/07/17 23:53:14 INFO mapred.JobClient:     Map input records=2
13/07/17 23:53:14 INFO mapred.JobClient:     Map output records=8
13/07/17 23:53:14 INFO mapred.JobClient:     Map output bytes=82
13/07/17 23:53:14 INFO mapred.JobClient:     Input split bytes=366
13/07/17 23:53:14 INFO mapred.JobClient:     Combine input records=8
```



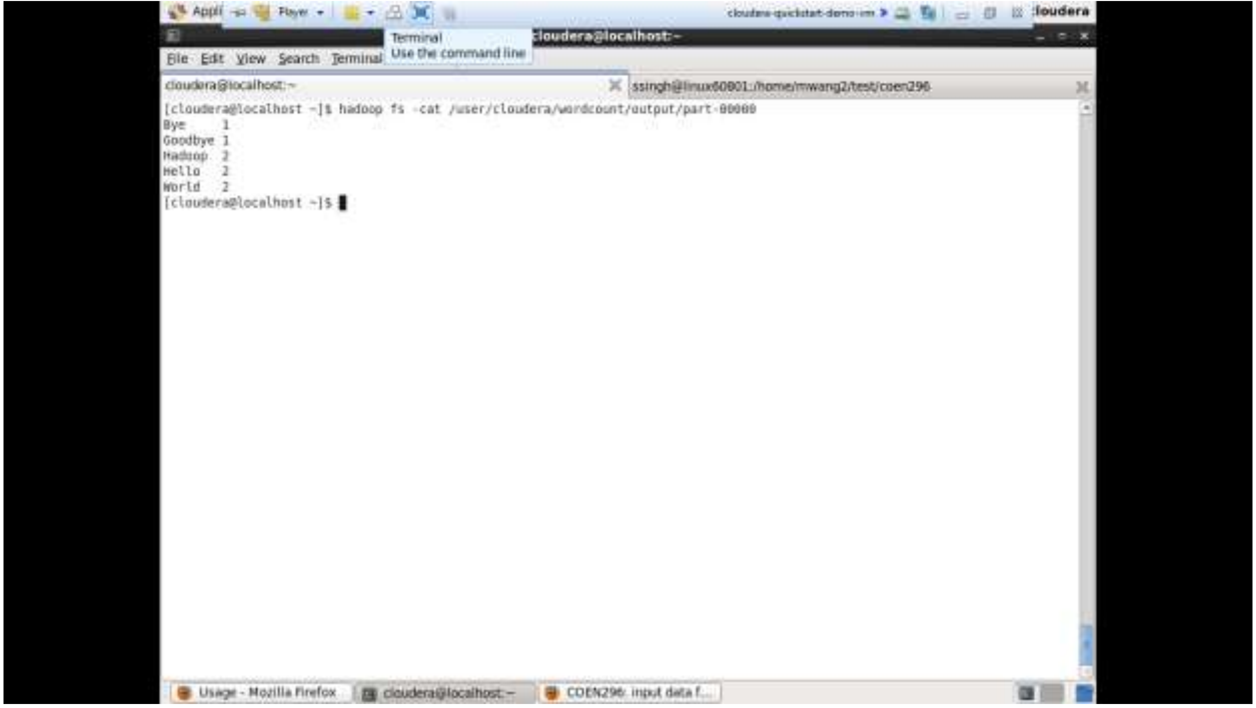
Appl
Player
cloudera-quickstart-demo-vm
cloudera
cloudera@localhost:~
Click to view your appointments and tasks
File Edit View Search Terminal Tabs Help
cloudera@localhost:~
ssingh@linux60801:/home/mwang2/test/coen296
Usage - Mozilla Firefox
cloudera@localhost:~
COEN296: input data f...

```
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of bytes read=79
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of bytes written=638967
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of large read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     FILE: Number of write operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of bytes read=419
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of bytes written=41
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of read operations=7
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of large read operations=0
13/07/17 23:53:14 INFO mapred.JobClient:     HDFS: Number of write operations=2
13/07/17 23:53:14 INFO mapred.JobClient:   Job Counters
13/07/17 23:53:14 INFO mapred.JobClient:     Launched map tasks=3
13/07/17 23:53:14 INFO mapred.JobClient:     Launched reduce tasks=1
13/07/17 23:53:14 INFO mapred.JobClient:     Data-local map tasks=3
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=41791
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=5498
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/07/17 23:53:14 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/07/17 23:53:14 INFO mapred.JobClient:   Map-Reduce Framework
13/07/17 23:53:14 INFO mapred.JobClient:     Map input records=2
13/07/17 23:53:14 INFO mapred.JobClient:     Map output records=8
13/07/17 23:53:14 INFO mapred.JobClient:     Map output bytes=82
13/07/17 23:53:14 INFO mapred.JobClient:     Input split bytes=366
13/07/17 23:53:14 INFO mapred.JobClient:     Combine input records=8
13/07/17 23:53:14 INFO mapred.JobClient:     Combine output records=6
13/07/17 23:53:14 INFO mapred.JobClient:     Reduce input groups=5
13/07/17 23:53:14 INFO mapred.JobClient:     Reduce shuffle bytes=117
13/07/17 23:53:14 INFO mapred.JobClient:     Reduce input records=6
13/07/17 23:53:14 INFO mapred.JobClient:     Reduce output records=5
13/07/17 23:53:14 INFO mapred.JobClient:     Spilled Records=12
13/07/17 23:53:14 INFO mapred.JobClient:     CPU time spent (ms)=2640
13/07/17 23:53:14 INFO mapred.JobClient:     Physical memory (bytes) snapshot=538959872
13/07/17 23:53:14 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=2801688576
13/07/17 23:53:14 INFO mapred.JobClient:     Total committed heap usage (bytes)=400764928
13/07/17 23:53:14 INFO mapred.JobClient:   org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter
13/07/17 23:53:14 INFO mapred.JobClient:     BYTES_READ=50
[cloudera@localhost ~]$ 
```

10. See the output byh displaying contents of file part-00000 in output directory /user/cloudera/wordcount/output

Command:

hadoop fs -cat /user/cloudera/wordcount/output/part-00000



```
[cloudera@localhost ~]$ hadoop fs -cat /user/cloudera/wordcount/output/part-00000
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
[cloudera@localhost ~]$ 
```