Cloudera Distributed Hadoop (CDH) Installation and Configuration on Virtual Box

By

Kavya Mugadur

W1014808

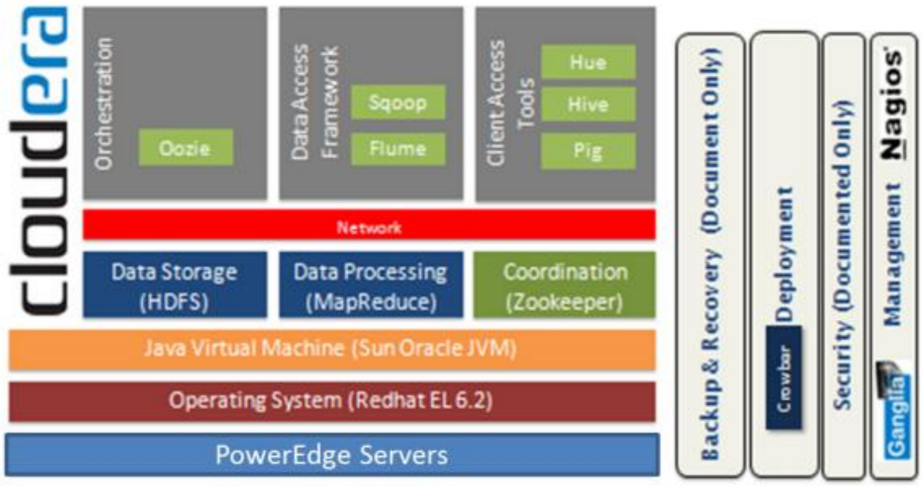
Table of contents

1.What is CDH ?

CDH (Cloudera Distribution Hadoop) is open-source Apache Hadoop distribution provided by Cloudera Inc which is a Palo Alto-based American enterprise software company.

CDH (Cloudera's Distribution Including Apache Hadoop) is the most complete, tested, and widely deployed distribution of Apache Hadoop. CDH is 100% open source and is the only Hadoop solution to offer batch processing, interactive SQL and interactive search as well as enterprise-grade continuous availability. More enterprises have downloaded CDH than all other distributions combined.

CLOUDERA TAXONOMY



The PowerEdge servers, the operating system, and the Java Virtual Machine make up the foundation on which the Hadoop software stack runs. The dark blue layer, depicting the core Hadoop components, comprises two frameworks:

• The Data Storage Framework is the file system that Hadoop uses to store data on the cluster nodes. Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file system.

• The Data Processing Framework (MapReduce) is a massively-parallel compute framework inspired by Google's MapReduce papers.

The next layer of the stack is the network layer. This is a dedicated cluster network, implemented from a blueprint using tested and qualified components. This implementation provides predictable high performance without interference from other applications.

The next three frameworks—the Orchestration, the Data Access Framework, and the Client Access Tools—are utilities that are part of the Hadoop ecosystem and provided by the CDH distribution.

2. Hadoop Basics

The Hadoop platform was designed to solve problems where you have a big data. It’s for situations where you want to run analytics that are deep and computationally extensive, like clustering and targeting. The majority of this data will be “unstructured” – complex data poorly suited to management by structured storage systems like relational database.

Unstructured data comes from many sources and takes many forms web logs, text files, sensor readings, user-generated content like product reviews or text messages, audio, video and still imagery and more

Dealing with big data requires two things:

- Inexpensive, reliable storage; and
- New tools for analyzing unstructured and structured data.

Apache Hadoop is a powerful open source software platform that addresses both of these problems. Hadoop is an Apache Software Foundation project. **Cloudera** offers commercial support and services to Hadoop users.

2.1 Reliable Storage: HDFS

Hadoop includes a fault-tolerant storage system called the **Hadoop Distributed File System, or HDFS**. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data.

Hadoop creates clusters of machines and coordinates work among them. Clusters can be builtwith inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster.

HDFS manages storage on the cluster by breaking incoming files into pieces, called “blocks,” and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers

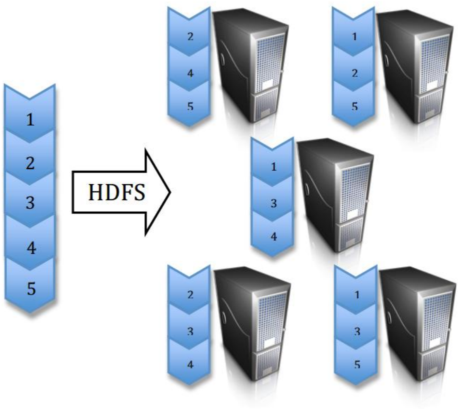


Figure 1: HDFS distributes file blocks among servers

2.2 Hadoop for Big Data Analysis

Many popular tools for enterprise data management -relational database systems, for example – are designed to make simple queries run quickly. They use techniques like indexing to examine just a small portion of all the available data in order to answer a question.

Hadoop is a different sort of tool. Hadoop is aimed at problems that require examination of all the available data. For example, text analysis and image processing generally require that every single record be read, and often interpreted in the context of similar records. Hadoop uses a technique called **MapReduce** to carry out this exhaustive analysis quickly.

In the previous section, we saw that HDFS distributes blocks from a single file among a large number of servers for reliability. Hadoop takes advantage of this data distribution by pushing the work involved in an analysis out to many different servers. Each of the servers runs the analysis on its own block from the file. Results are collated and digested into a single result after each piece has been analyzed.

Running the analysis on the nodes that actually store the data delivers much much better performance than reading data over the network from a single centralized server. Hadoop monitors jobs during execution, and will restart work lost due to node failure if necessary. In fact, if a particular node is running very slowly, Hadoop will restart its work on another server with a copy of the data.

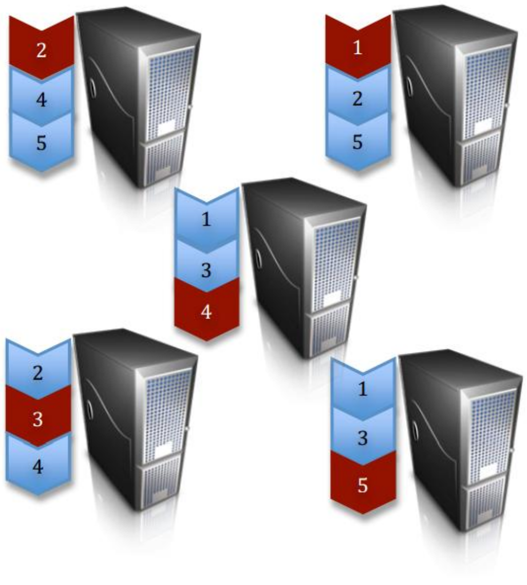


Figure 2: Hadoop pushes work out to the data

Hadoop's MapReduce and HDFS use simple, robust techniques on inexpensive computer systems to delivevery high data availability and to analyze enormous amounts of information quickly. Hadoop offers enterprises a powerful new tool for managing big data.

3.Ways to Install CDH4

You can install CDH4 in any of the following ways:

a. Installing using Cloudera quickstart vm.
b. Automated method using Cloudera Manager. Cloudera Manager Free Edition automates the installation and configuration of CDH4 on an entire cluster if you have root or password-less sudo SSH access to your cluster's machines.
c. Manual methods described below:
    - Download and install the CDH4 "1-click Install" package
    - Add the CDH4 repository
    - Build your own CDH4 repository
    - Install from a CDH4 tarball

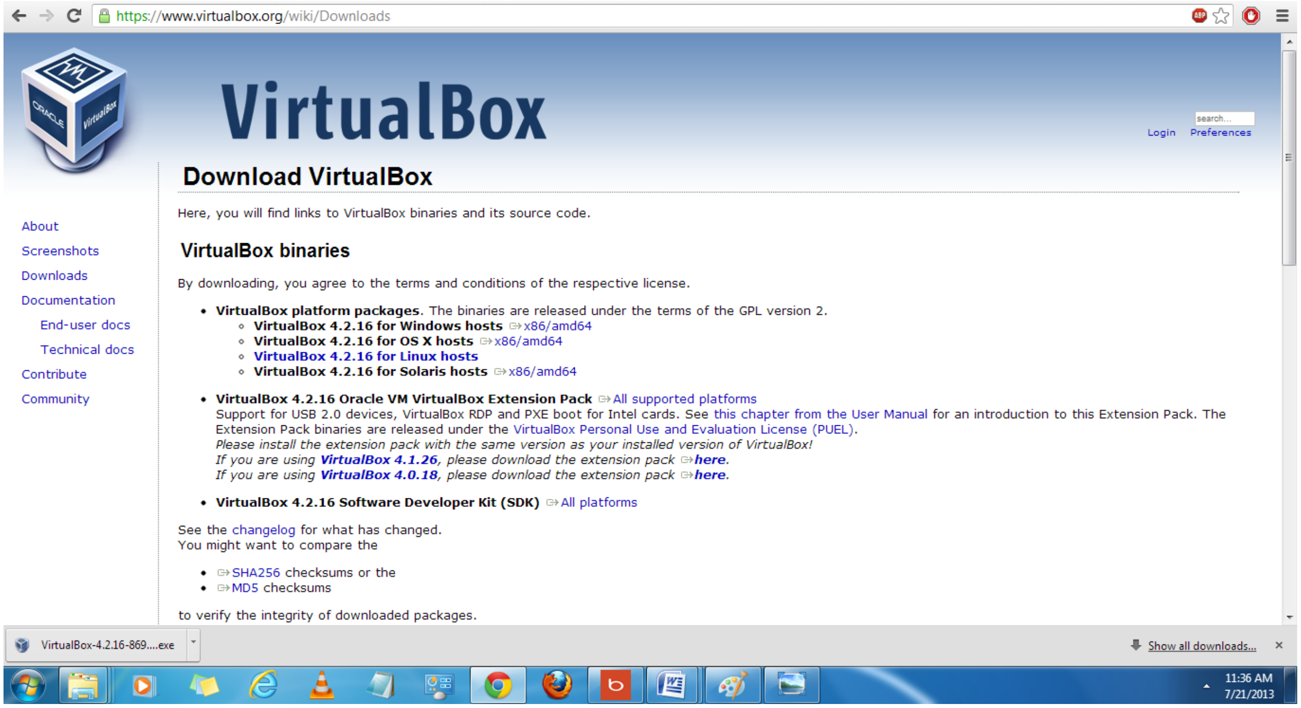
In this document we will explain the installation of CDH using first method.

4. Installation and Configuration of CDH on Virtual machine using Cloudera quickstart vm

Cloudera quickstart VM contains a sample of Cloudera's platform for "Big Data". 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

In this document we have installed CDH on VirtualBox. Below are the steps to install CDH using Cloudera quickstart vm on Virtual box

Step 1: Download the virtual box- executable file from https://www.virtualbox.org/wiki/Downloads
Download VirtualBox 4.2.16 for Windows hosts



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.

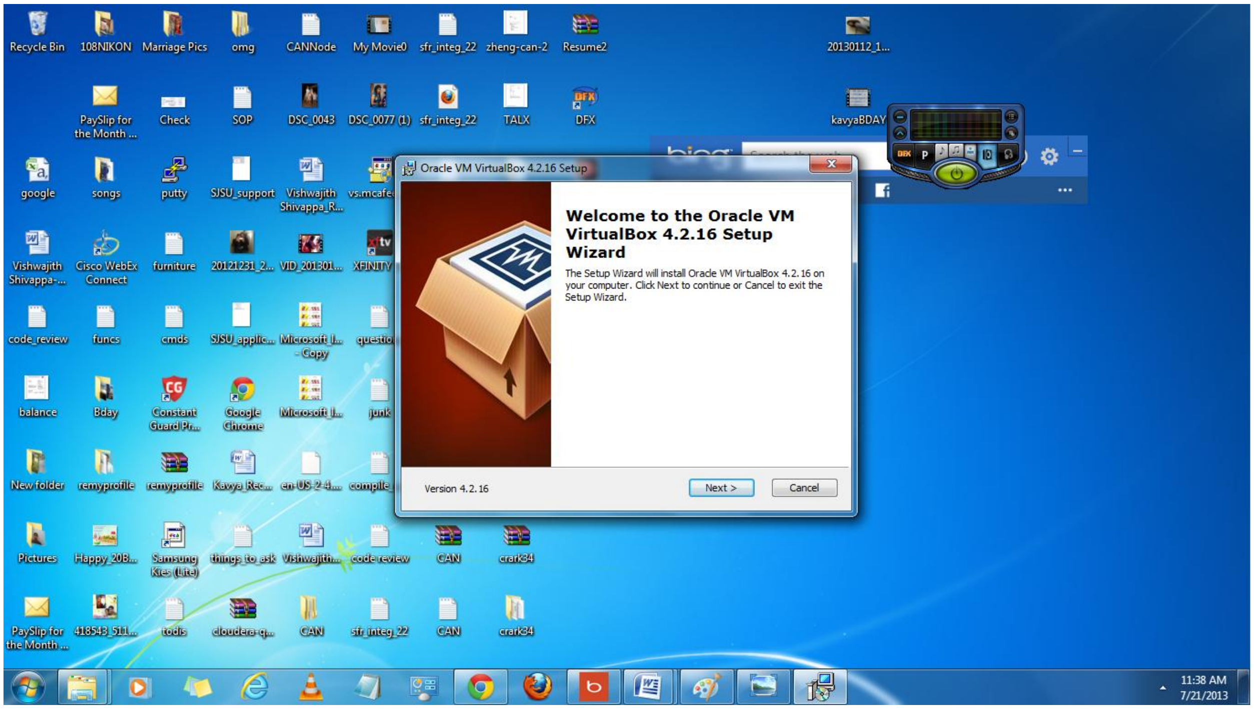
- **VirtualBox platform packages**. The binaries are released under the terms of the GPL version 2.
  - **VirtualBox 4.2.16 for Windows hosts** ⇒x86/amd64
  - **VirtualBox 4.2.16 for OS X hosts** ⇒x86/amd64
  - **VirtualBox 4.2.16 for Linux hosts**
  - **VirtualBox 4.2.16 for Solaris hosts** ⇒x86/amd64
- **VirtualBox 4.2.16 Oracle VM VirtualBox Extension Pack** ⇒All supported platforms
  Support for USB 2.0 devices, VirtualBox RDP and PXE boot for Intel cards. See this chapter from the User Manual for an introduction to this Extension Pack. The Extension Pack binaries are released under the VirtualBox Personal Use and Evaluation License (PUEL).
  Please install the extension pack with the same version as your installed version of VirtualBox!
  *If you are using **VirtualBox 4.1.26**, please download the extension pack ⇒**here**.*
  *If you are using **VirtualBox 4.0.18**, please download the extension pack ⇒**here**.*
- **VirtualBox 4.2.16 Software Developer Kit (SDK)** ⇒All platforms

See the changelog for what has changed.
You might want to compare the

- ⇒SHA256 checksums or the
- ⇒MD5 checksums

to verify the integrity of downloaded packages.

Step 2: Install VirtualBox by double clicking on the downloaded file.

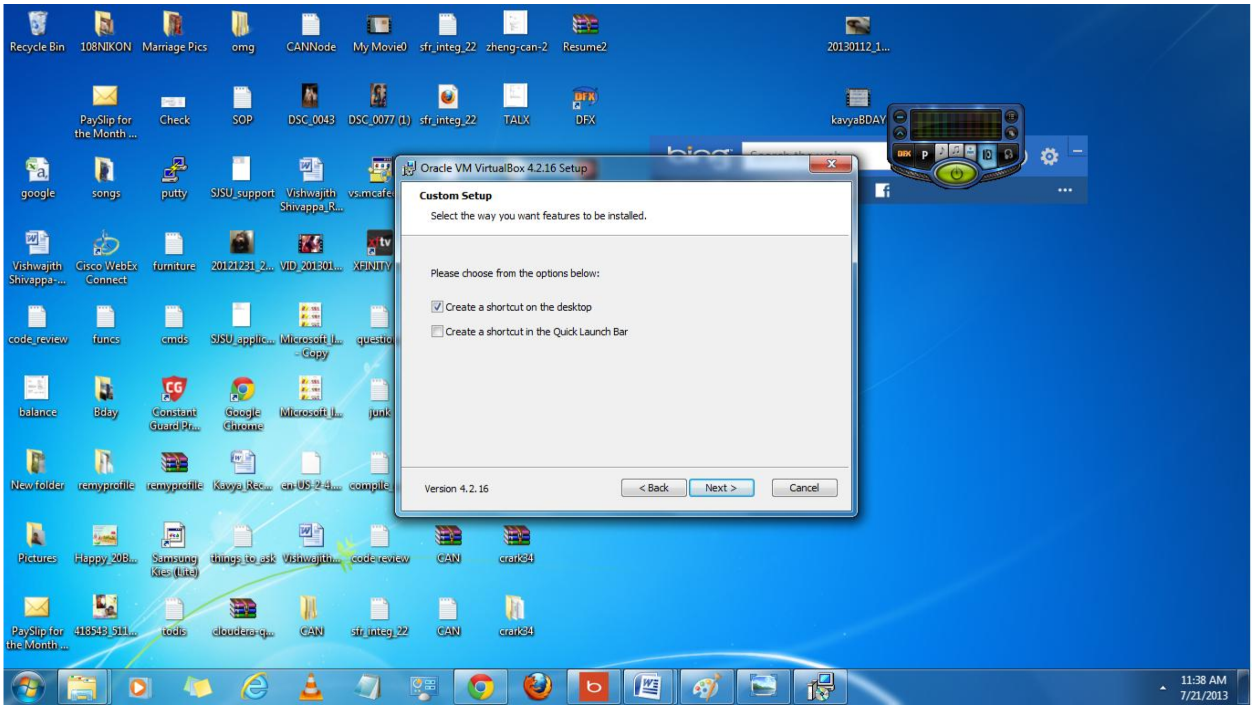


Oracle VM VirtualBox 4.2.16 Setup

Welcome to the Oracle VM VirtualBox 4.2.16 Setup Wizard

The Setup Wizard will install Oracle VM VirtualBox 4.2.16 on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Version 4.2.16



Oracle VM VirtualBox 4.2.16 Setup

Custom Setup

Select the way you want features to be installed.

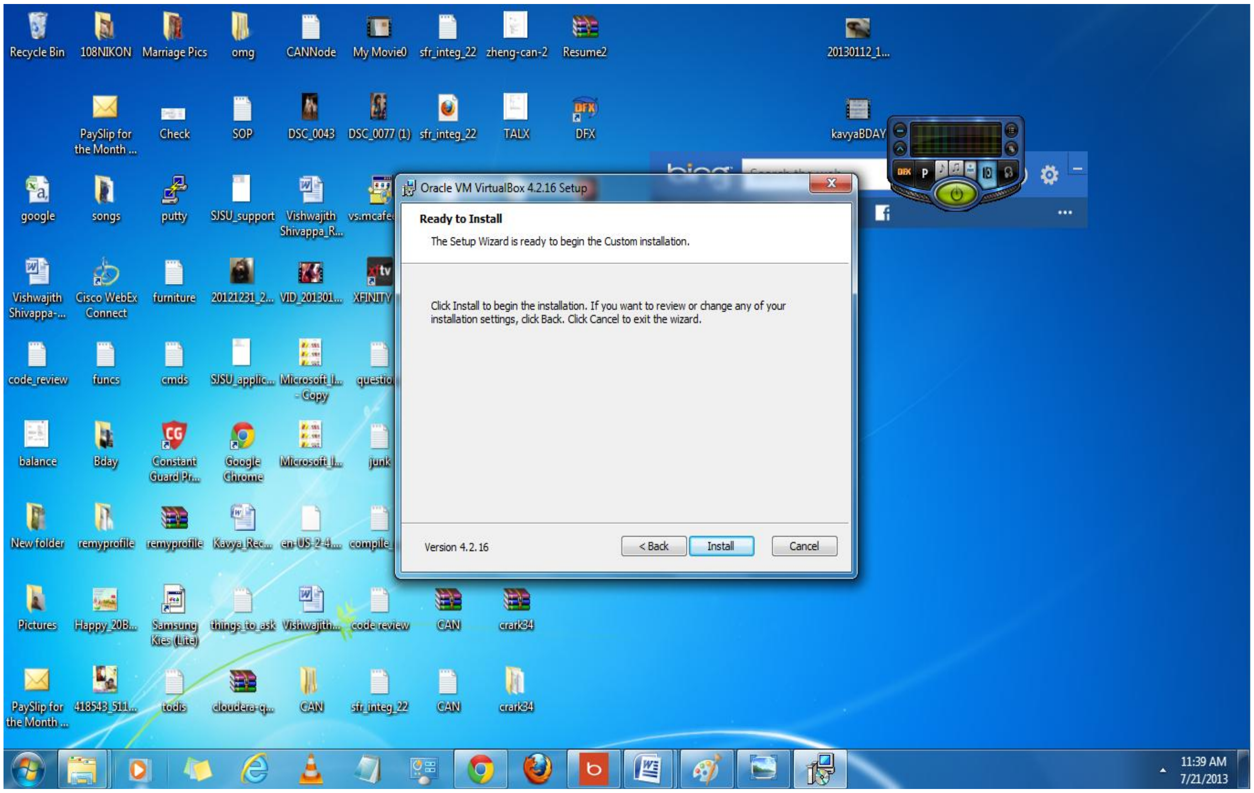
Please choose from the options below:

☑ Create a shortcut on the desktop

☐ Create a shortcut in the Quick Launch Bar

Version 4.2.16

Click install to install VirtualBox with default settings. Installation is shown as below:



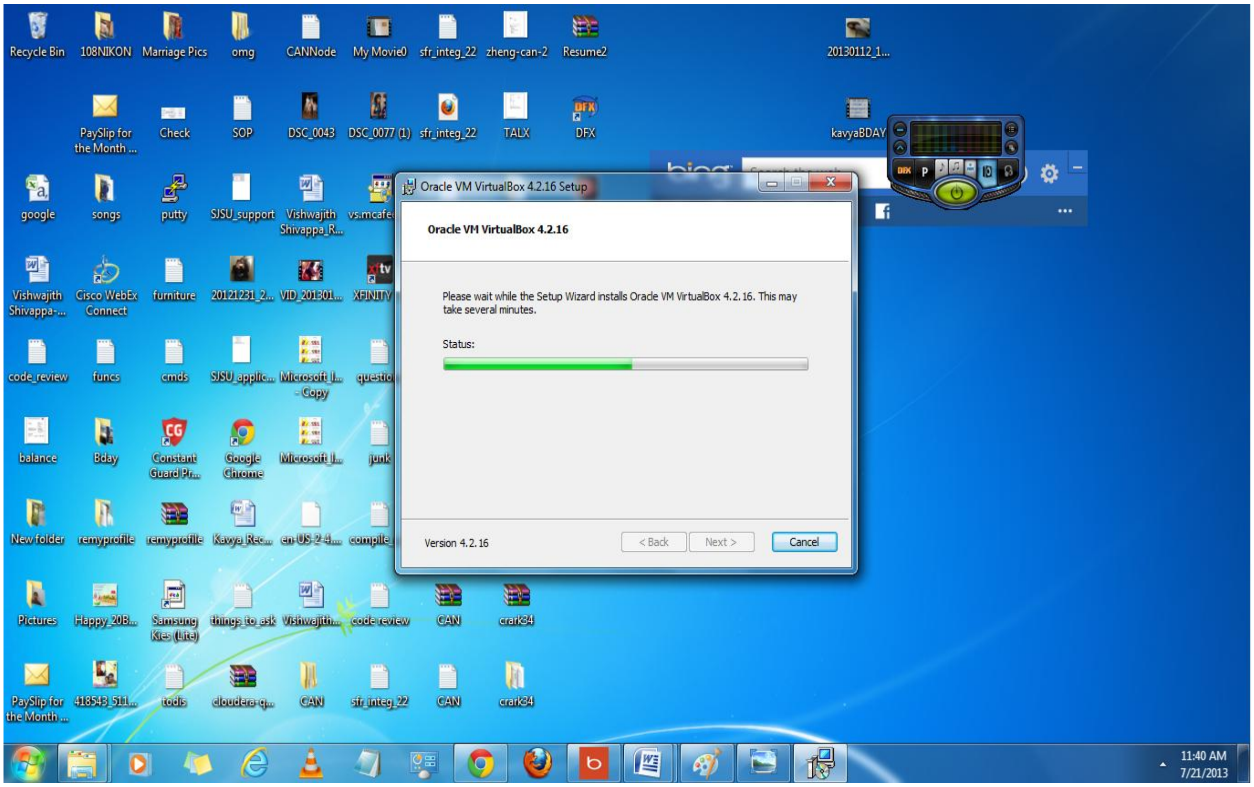
Oracle VM VirtualBox 4.2.16 Setup

Ready to Install

The Setup Wizard is ready to begin the Custom installation.

Click Install to begin the installation. If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.

Version 4.2.16



Oracle VM VirtualBox 4.2.16 Setup

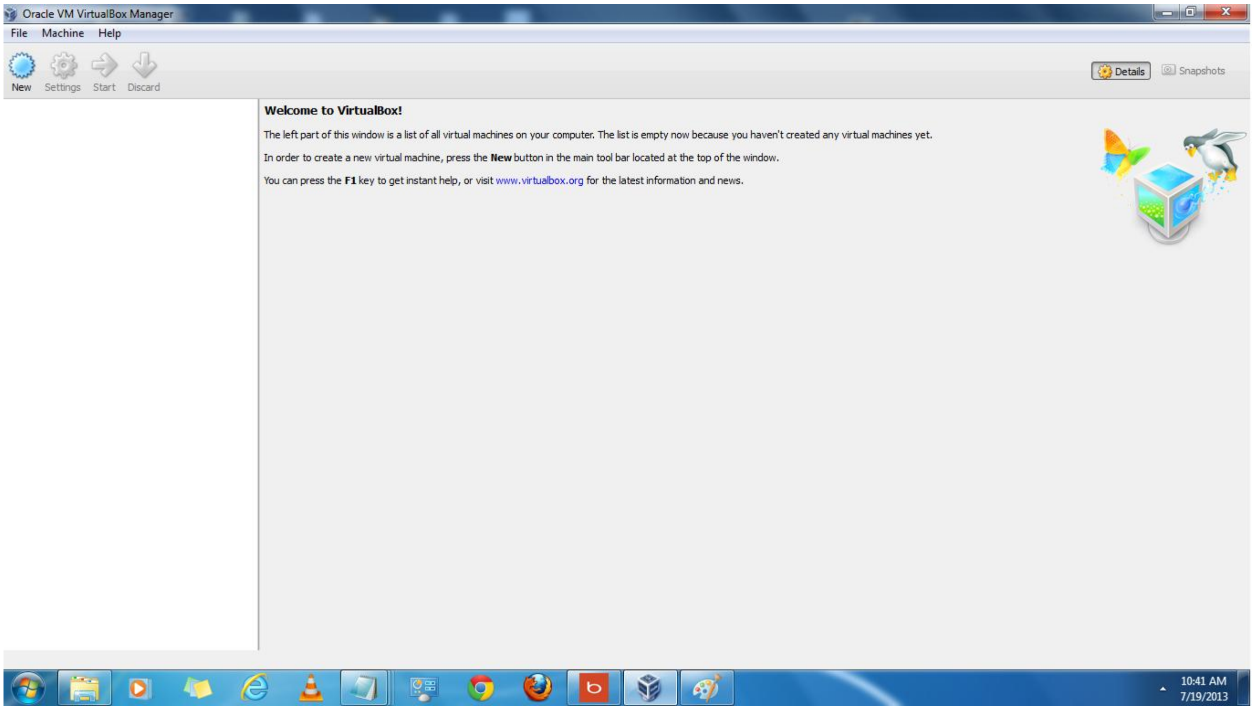
Oracle VM VirtualBox 4.2.16

Please wait while the Setup Wizard installs Oracle VM VirtualBox 4.2.16. This may take several minutes.

Status:

Version 4.2.16

If the installation is successful you will see a Virtual Manager window to manage VMs.



Welcome to VirtualBox!

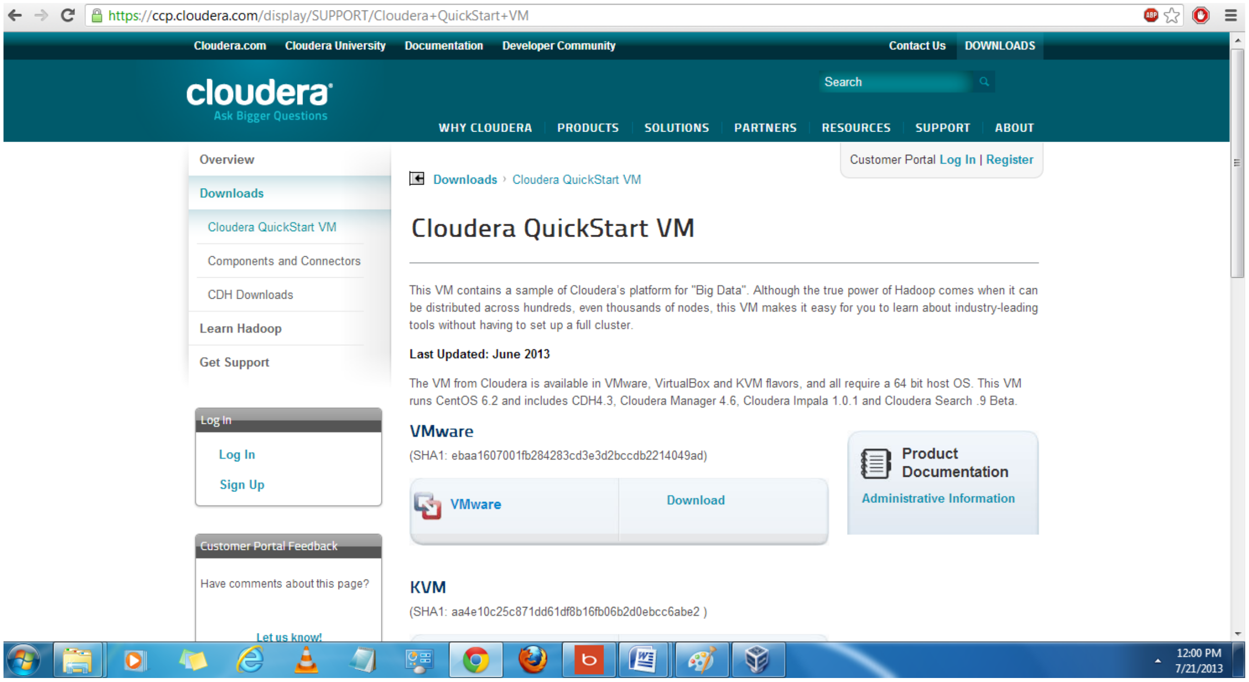
The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.

In order to create a new virtual machine, press the **New** button in the main tool bar located at the top of the window.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for the latest information and news.

Step 3: Download the Cloudera quickstart vm for VirtualBox

Go to the link - https://ccp.cloudera.com/display/SUPPORT/Cloudera+QuickStart+VM



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.

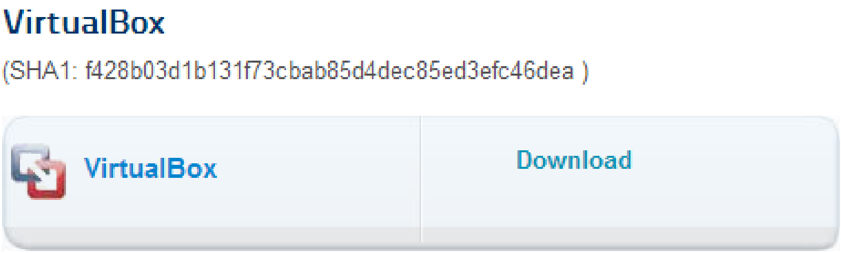
VMware

(SHA1: ebaa1607001fb284283cd3e3d2bccdb2214049ad)

KVM

(SHA1: aa4e10c25c871dd61df8b16fb06b2d0ebcc6abe2)

Select quickVM for VirtualBox and click on download

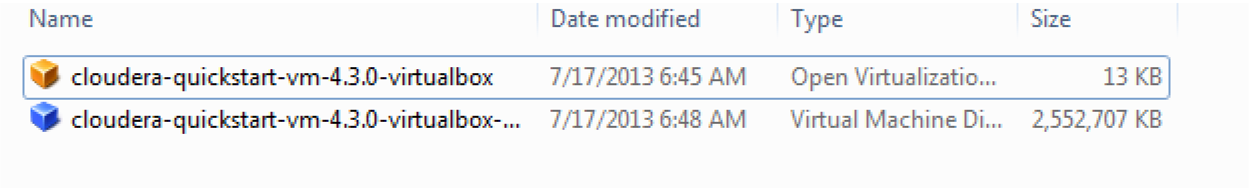


VirtualBox

(SHA1: f428b03d1b131f73cbab85d4dec85ed3efc46dea)

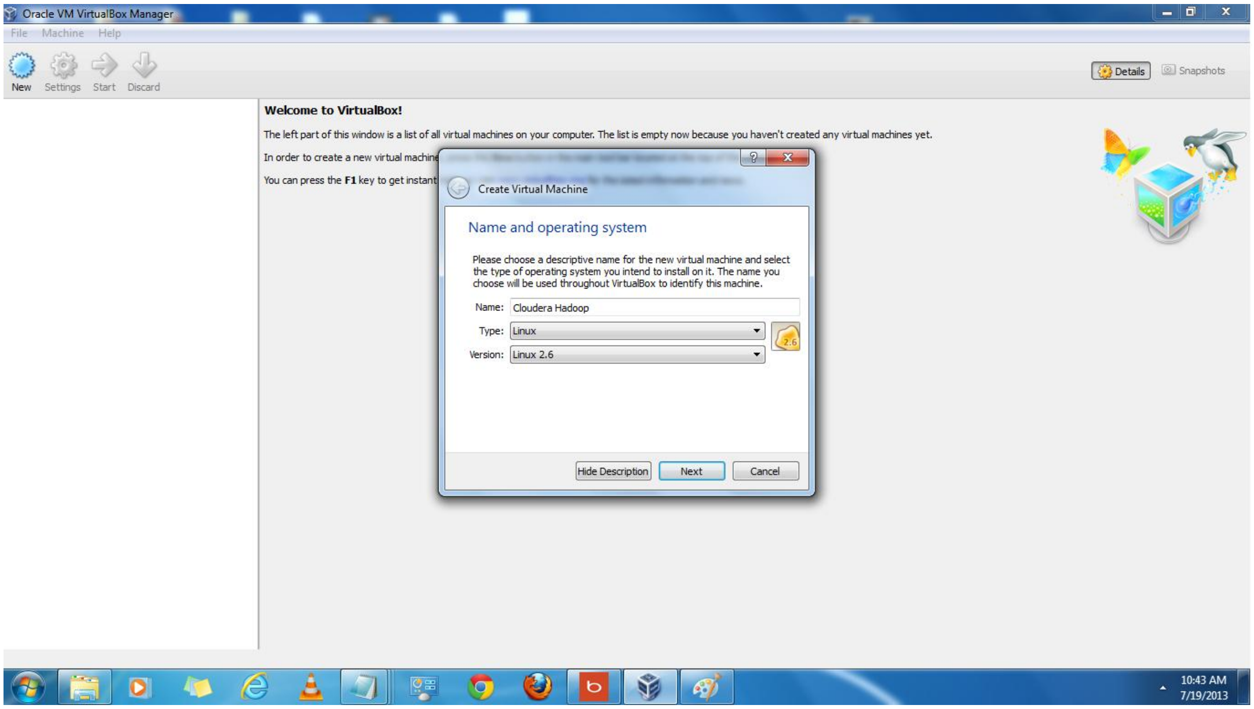
Step 4: Unzip the downloaded file.

When you unzip the file cloudera-quickstart-vm-4.3.0-virtualbox.tar you will find these two files in the directory.



Name	Date modified	Type	Size
cloudera-quickstart-vm-4.3.0-virtualbox	7/17/2013 6:45 AM	Open Virtualizatio...	13 KB
cloudera-quickstart-vm-4.3.0-virtualbox-...	7/17/2013 6:48 AM	Virtual Machine Di...	2,552,707 KB

Step 5: Open VirualBox and click on “New” to create new virtual box



Welcome to VirtualBox!

The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.

In order to create a new virtual machine

You can press the **F1** key to get instant

Create Virtual Machine

Name and operating system

Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

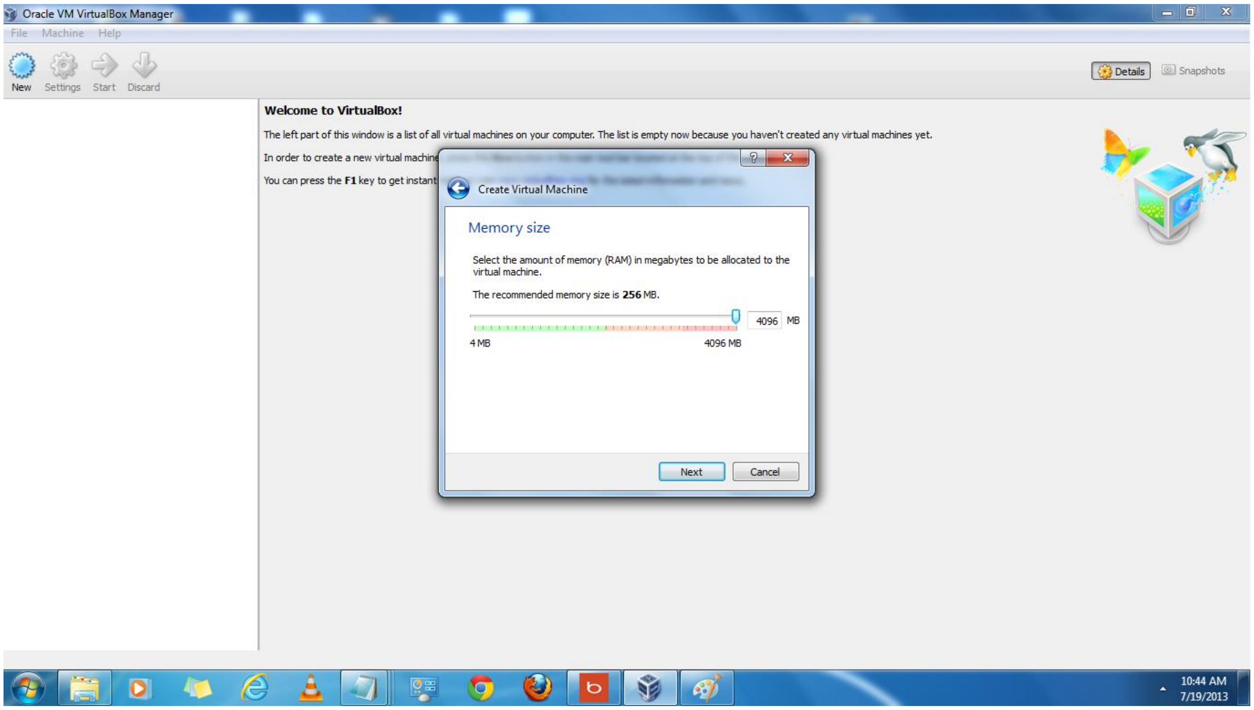
Name: Cloudera Hadoop

Type: Linux

Version: Linux 2.6

Give name for new virtual machine and select type as Linux and versions as Linux 2.6

Step 6 : Select Memory Size as 4GB and click Next.



Welcome to VirtualBox!

The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.

In order to create a new virtual machine

You can press the **F1** key to get instant

Memory size

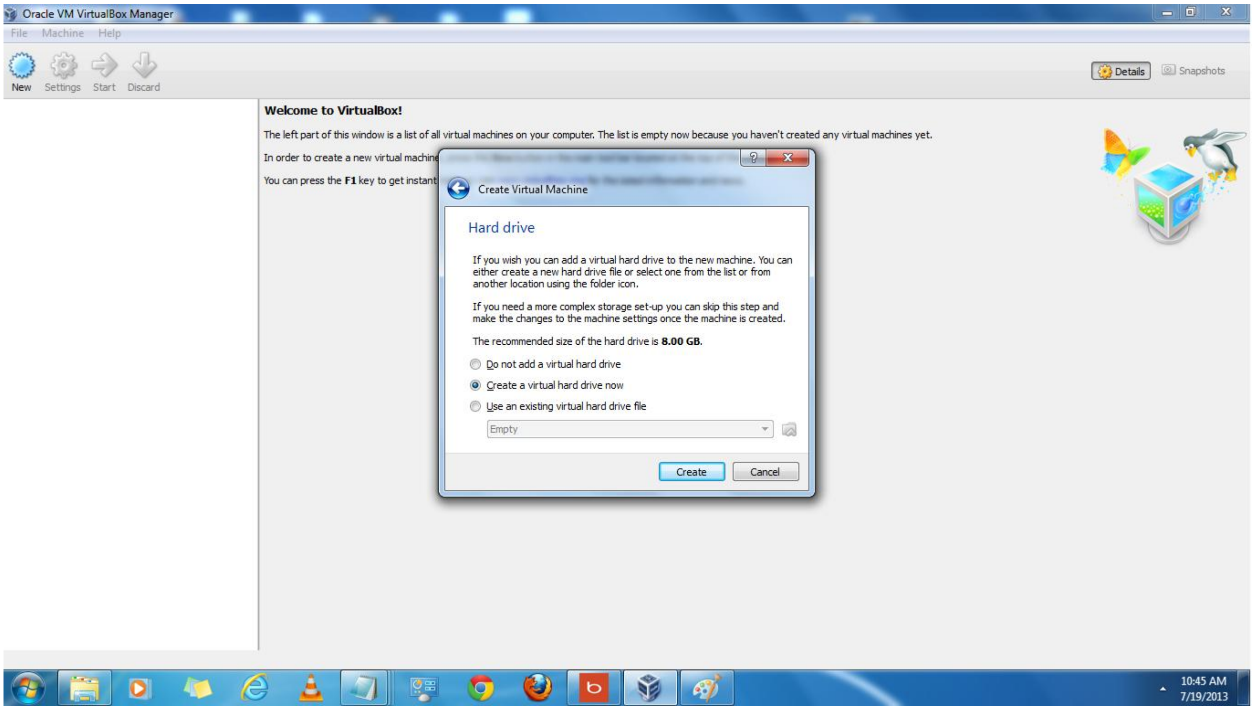
Select the amount of memory (RAM) in megabytes to be allocated to the virtual machine.

The recommended memory size is **256** MB.

4096 MB

4 MB 4096 MB

Step 7: In the next page, VirtualBox asks to select Hard Drive for new VirualBox as shown in the screenshot. Create a virtual hard drive now is selected by default. But you have to select "Use an existing virtual hard drive file" option.



Welcome to VirtualBox!

The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.

In order to create a new virtual machine

You can press the **F1** key to get instant

Hard drive

If you wish you can add a virtual hard drive to the new machine. You can either create a new hard drive file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard drive is **8.00 GB.**

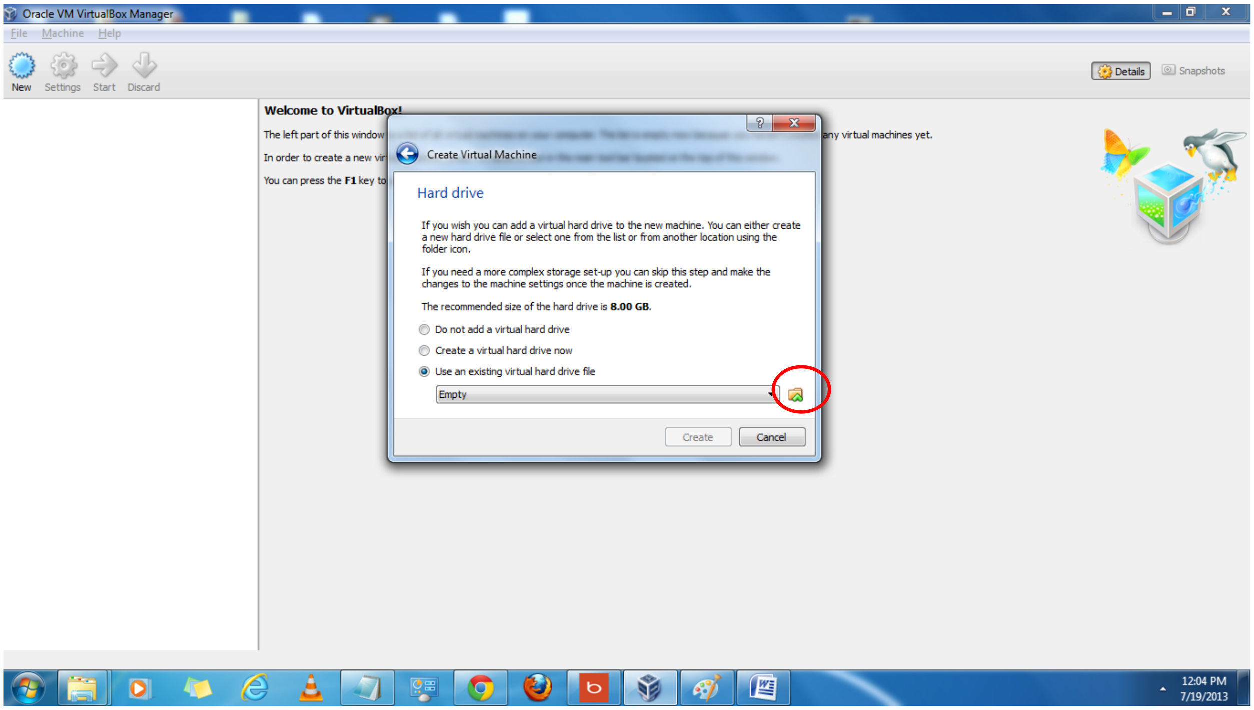
○ Do not add a virtual hard drive

● Create a virtual hard drive now

○ Use an existing virtual hard drive file

Empty

Select “Use an existing virtual hard drive file”.


Oracle VM VirtualBox Manager
File Machine Help
New Settings Start Discard
Details Snapshots
Welcome to VirtualBox!
The left part of this window
any virtual machines yet.
In order to create a new vir
You can press the F1 key to
12:04 PM
7/19/2013

Create Virtual Machine

Hard drive

If you wish you can add a virtual hard drive to the new machine. You can either create a new hard drive file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard drive is **8.00 GB**.

○ Do not add a virtual hard drive

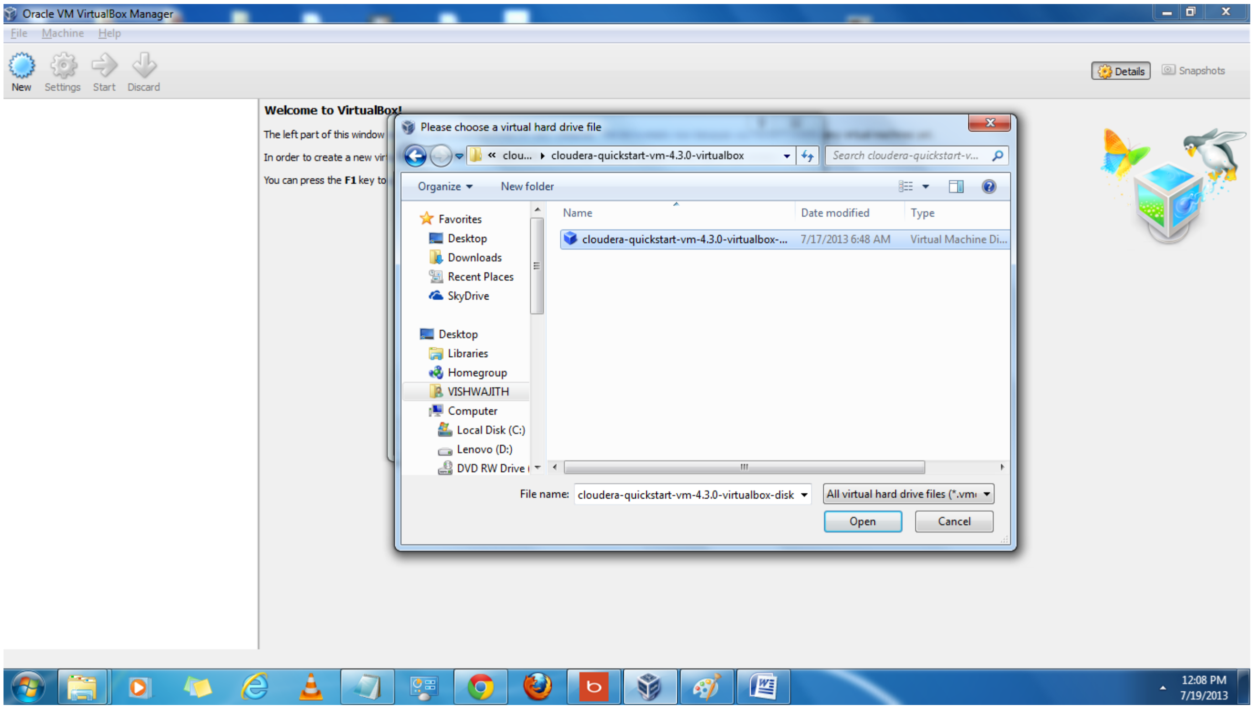
○ Create a virtual hard drive now

● Use an existing virtual hard drive file

Empty

Create Cancel

Step 8: Click on the small yellow icon beside the dropdown to browse and select the cloudera-quickstart-vm-4.3.0-virtualbox-disk1.vm file (which is download in step 4).


Oracle VM VirtualBox Manager
File Machine Help
New Settings Start Discard
Details Snapshots
Welcome to VirtualBox!
The left part of this window
In order to create a new vir
You can press the F1 key to
Please choose a virtual hard drive file
« clou... ▸ cloudera-quickstart-vm-4.3.0-virtualbox
Search cloudera-quickstart-v...
Organize ▾
New folder
Favorites
Desktop
Downloads
Recent Places
SkyDrive
Desktop
Libraries
Homegroup
VISHWAJITH
Computer
Local Disk (C:)
Lenovo (D:)
DVD RW Drive
12:08 PM
7/19/2013

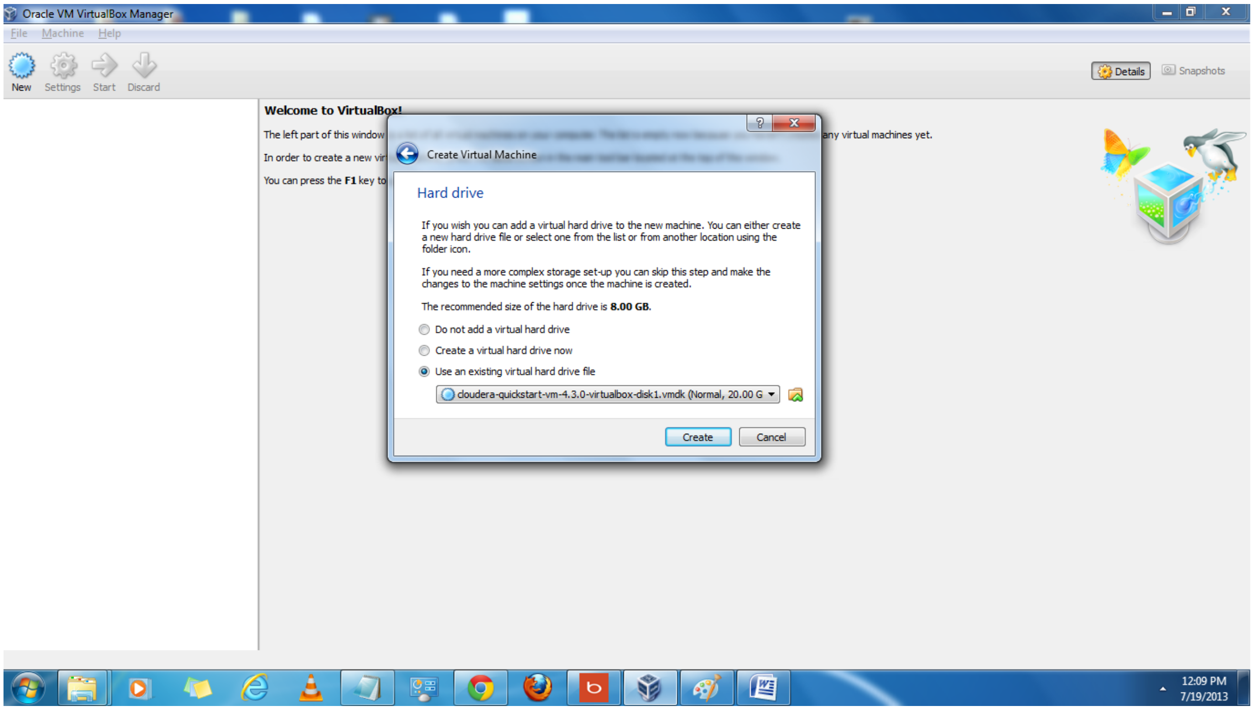
Name	Date modified	Type
cloudera-quickstart-vm-4.3.0-virtualbox-...	7/17/2013 6:48 AM	Virtual Machine Di...

File name: cloudera-quickstart-vm-4.3.0-virtualbox-disk

All virtual hard drive files (*.vmı

Open Cancel

Click on create to create Cloudera quickstart vm.



Welcome to VirtualBox!

The left part of this window … any virtual machines yet.

In order to create a new vir…

You can press the **F1** key to…

Create Virtual Machine

Hard drive

If you wish you can add a virtual hard drive to the new machine. You can either create a new hard drive file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard drive is **8.00 GB**.

○ Do not add a virtual hard drive

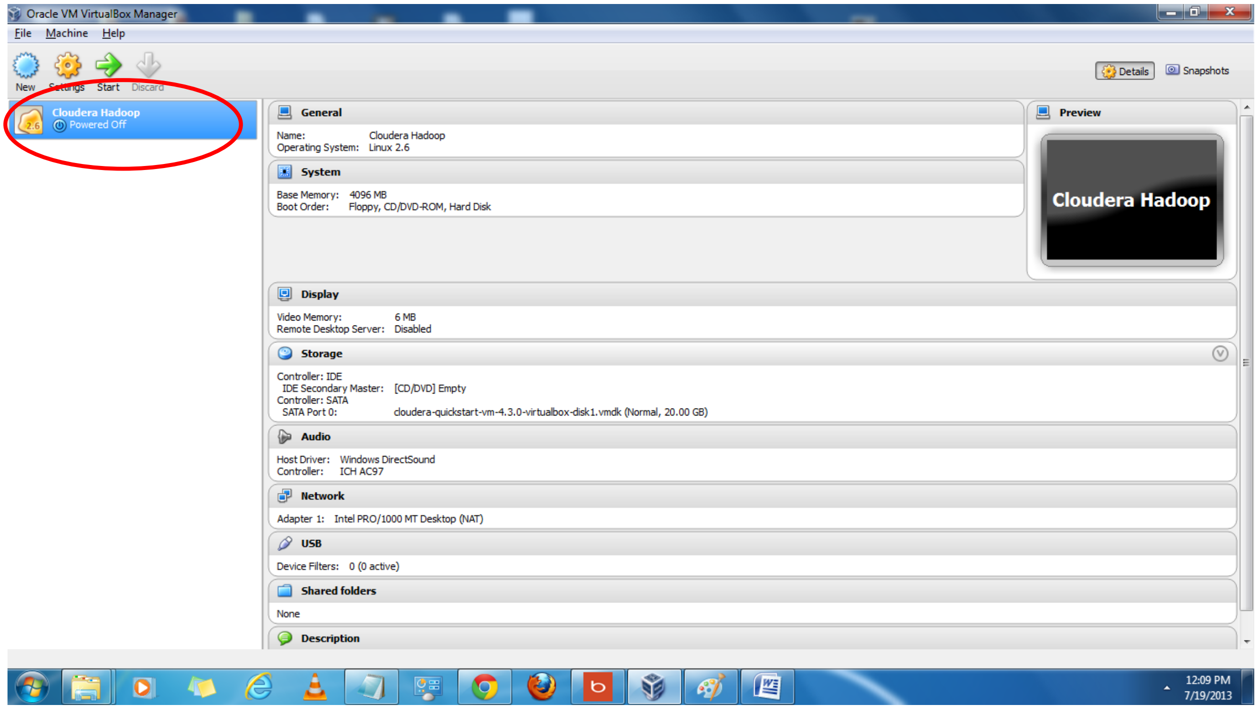
○ Create a virtual hard drive now

● Use an existing virtual hard drive file

cloudera-quickstart-vm-4.3.0-virtualbox-disk1.vmdk (Normal, 20.00 G ▾

Create Cancel

Step 9 : Your virtual box should look like following screen shots. We can see the new virtual machine named Cloudera Hadoop on the left side.



Cloudera Hadoop
Powered Off

General

Name: Cloudera Hadoop
Operating System: Linux 2.6

System

Base Memory: 4096 MB
Boot Order: Floppy, CD/DVD-ROM, Hard Disk

Display

Video Memory: 6 MB
Remote Desktop Server: Disabled

Storage

Controller: IDE
IDE Secondary Master: [CD/DVD] Empty
Controller: SATA
SATA Port 0: cloudera-quickstart-vm-4.3.0-virtualbox-disk1.vmdk (Normal, 20.00 GB)

Audio

Host Driver: Windows DirectSound
Controller: ICH AC97

Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

USB

Device Filters: 0 (0 active)

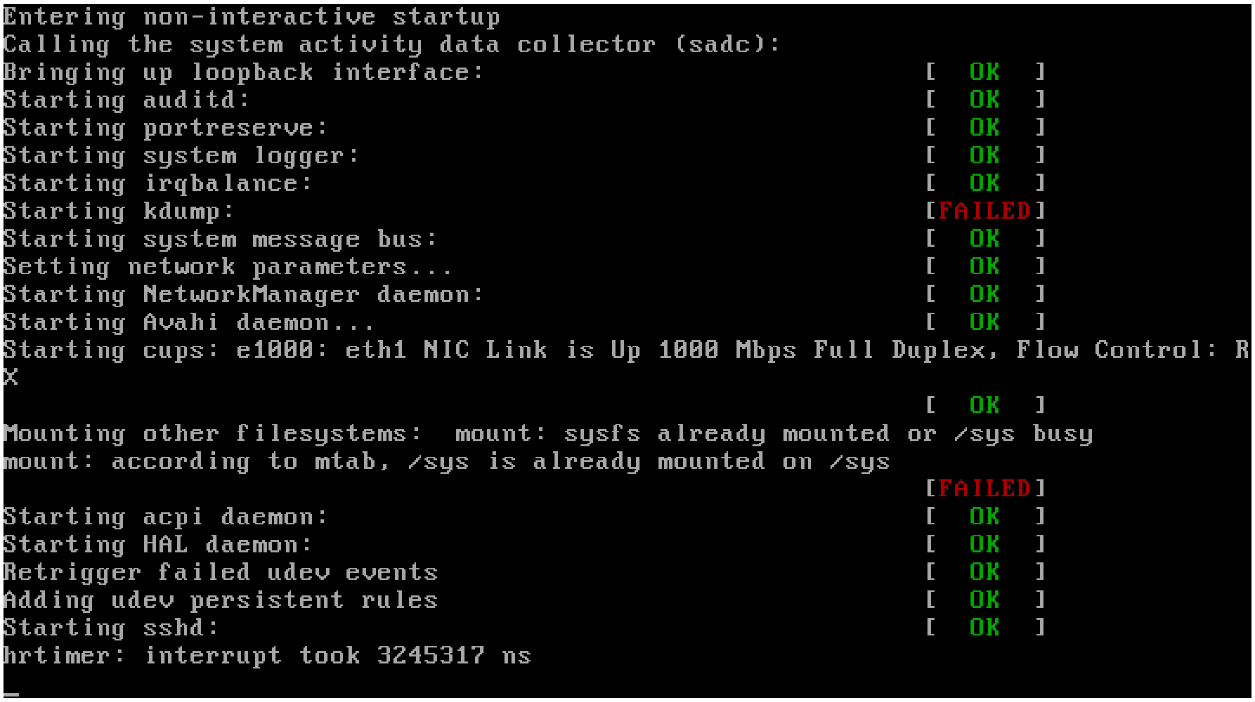
Shared folders

None

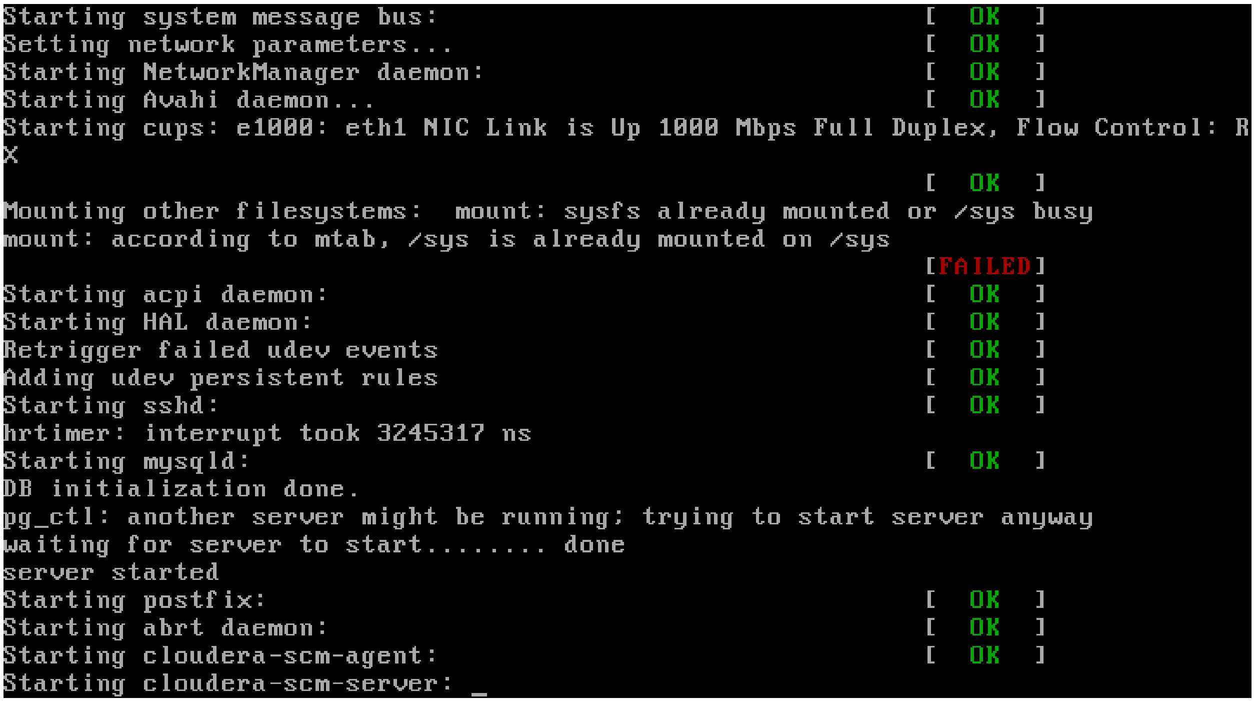
Description

Step 10: Select Cloudera vm and click on “Start”

Virtual Machine starts to boot

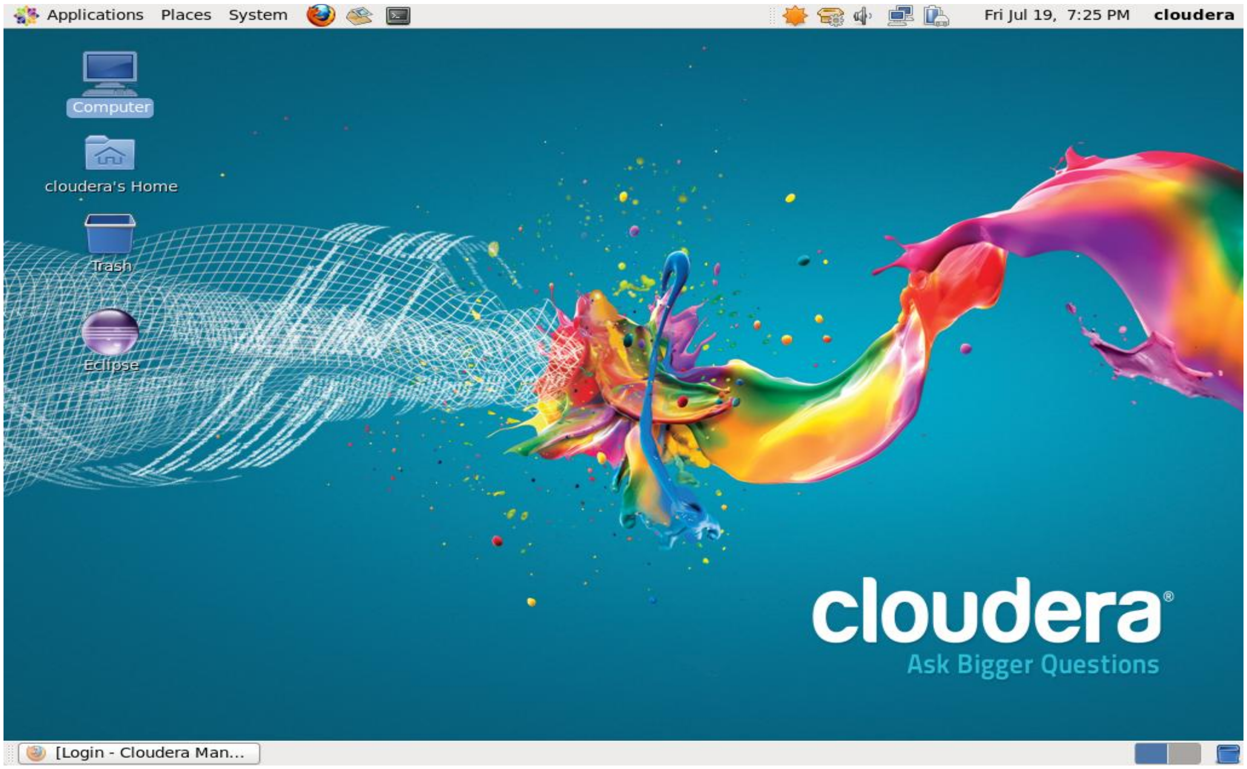


```
Entering non-interactive startup
Calling the system activity data collector (sadc):
Bringing up loopback interface:                           [  OK  ]
Starting auditd:                                          [  OK  ]
Starting portreserve:                                     [  OK  ]
Starting system logger:                                   [  OK  ]
Starting irqbalance:                                      [  OK  ]
Starting kdump:                                           [FAILED]
Starting system message bus:                              [  OK  ]
Setting network parameters...                             [  OK  ]
Starting NetworkManager daemon:                           [  OK  ]
Starting Avahi daemon...                                  [  OK  ]
Starting cups: e1000: eth1 NIC Link is Up 1000 Mbps Full Duplex, Flow Control: R
X
                                                          [  OK  ]
Mounting other filesystems:  mount: sysfs already mounted or /sys busy
mount: according to mtab, /sys is already mounted on /sys
                                                          [FAILED]
Starting acpi daemon:                                     [  OK  ]
Starting HAL daemon:                                      [  OK  ]
Retrigger failed udev events                              [  OK  ]
Adding udev persistent rules                              [  OK  ]
Starting sshd:                                            [  OK  ]
hrtimer: interrupt took 3245317 ns
_
```

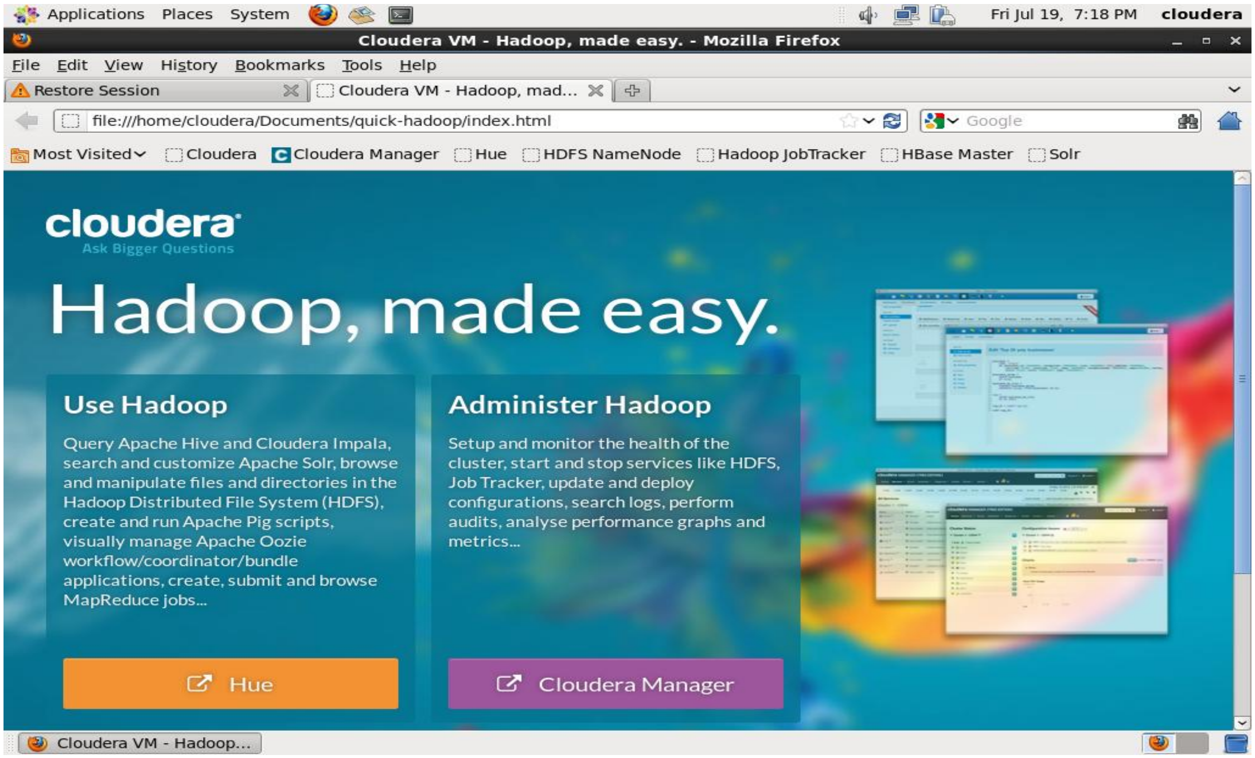


```
Starting system message bus:                              [  OK  ]
Setting network parameters...                             [  OK  ]
Starting NetworkManager daemon:                           [  OK  ]
Starting Avahi daemon...                                  [  OK  ]
Starting cups: e1000: eth1 NIC Link is Up 1000 Mbps Full Duplex, Flow Control: R
X
                                                          [  OK  ]
Mounting other filesystems:  mount: sysfs already mounted or /sys busy
mount: according to mtab, /sys is already mounted on /sys
                                                          [FAILED]
Starting acpi daemon:                                     [  OK  ]
Starting HAL daemon:                                      [  OK  ]
Retrigger failed udev events                              [  OK  ]
Adding udev persistent rules                              [  OK  ]
Starting sshd:                                            [  OK  ]
hrtimer: interrupt took 3245317 ns
Starting mysqld:                                          [  OK  ]
DB initialization done.
pg_ctl: another server might be running; trying to start server anyway
waiting for server to start........ done
server started
Starting postfix:                                         [  OK  ]
Starting abrt daemon:                                     [  OK  ]
Starting cloudera-scm-agent:                              [  OK  ]
Starting cloudera-scm-server: _
```

Step 11: System is loaded and CDH is installed on virtual machine.



Step 12: System redirects you to the index page of Cloudera.


Applications Places System
Fri Jul 19, 7:18 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
Google
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
cloudera
Ask Bigger Questions
Cloudera VM - Hadoop...

Hadoop, made easy.

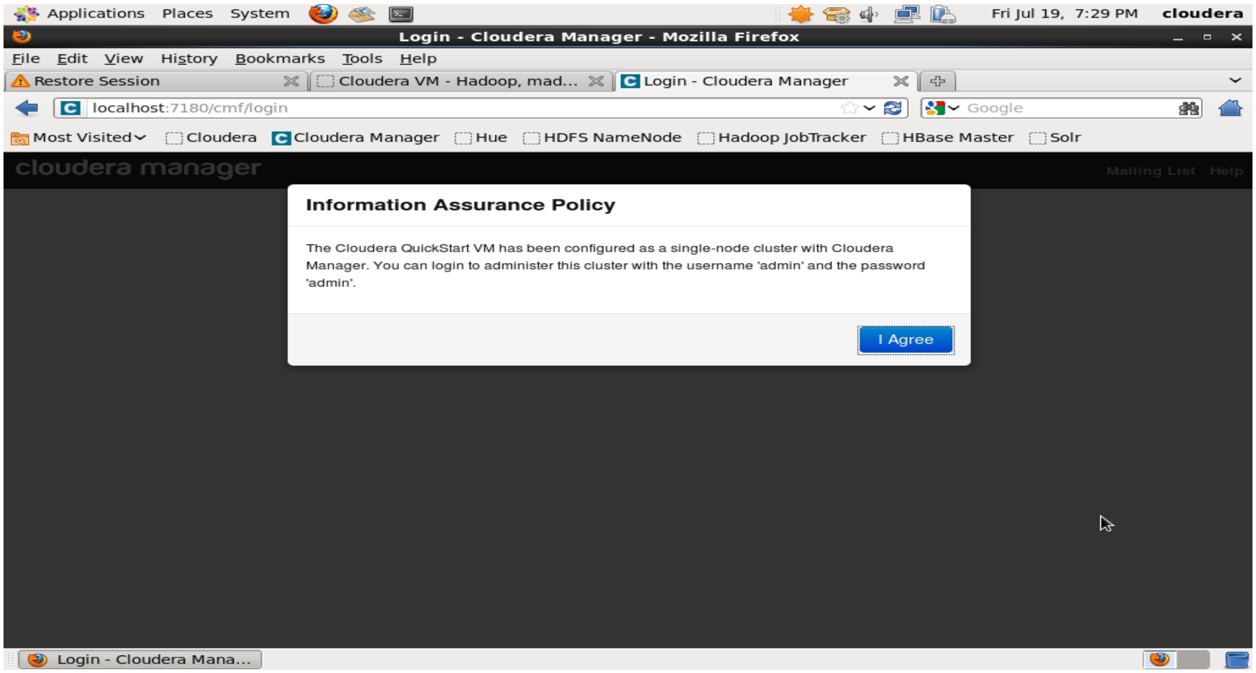
Use Hadoop

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...

Step 13: Select Cloudera Manager and Agree to the information assurance policy.

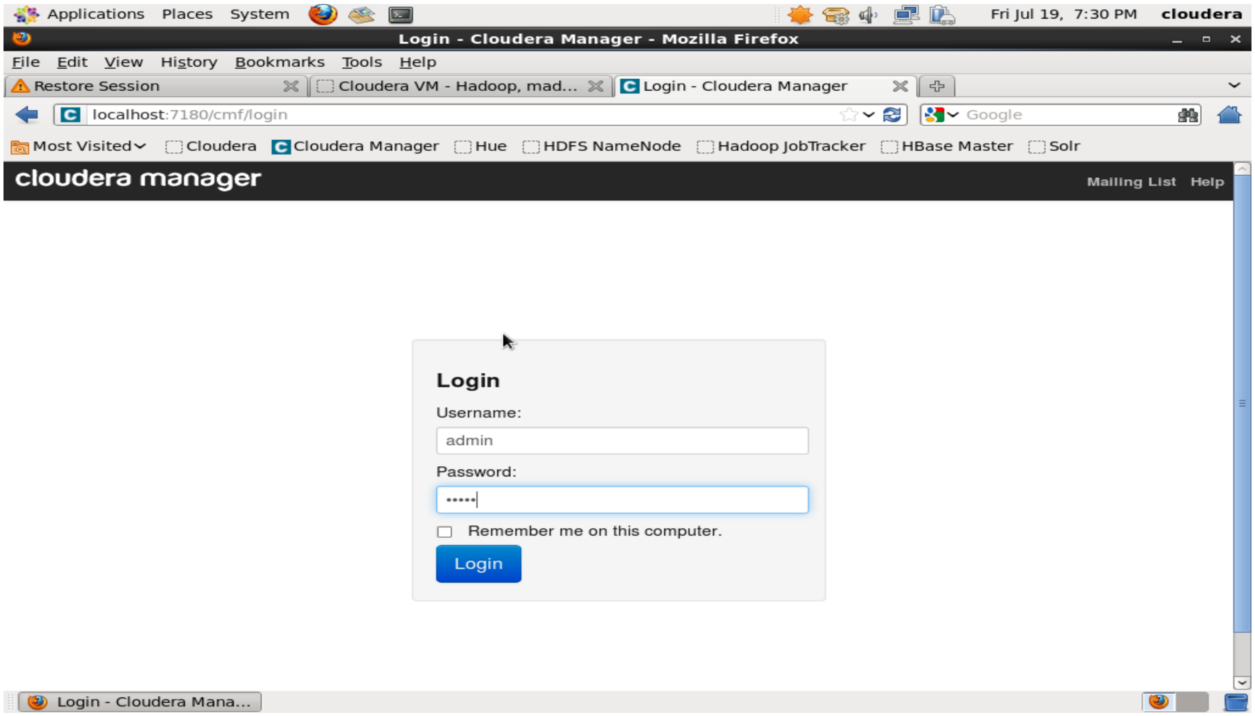


Information Assurance Policy

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'.

I Agree

Step 14: Login to Cloudera Manager as admin. Password is admin.



Login

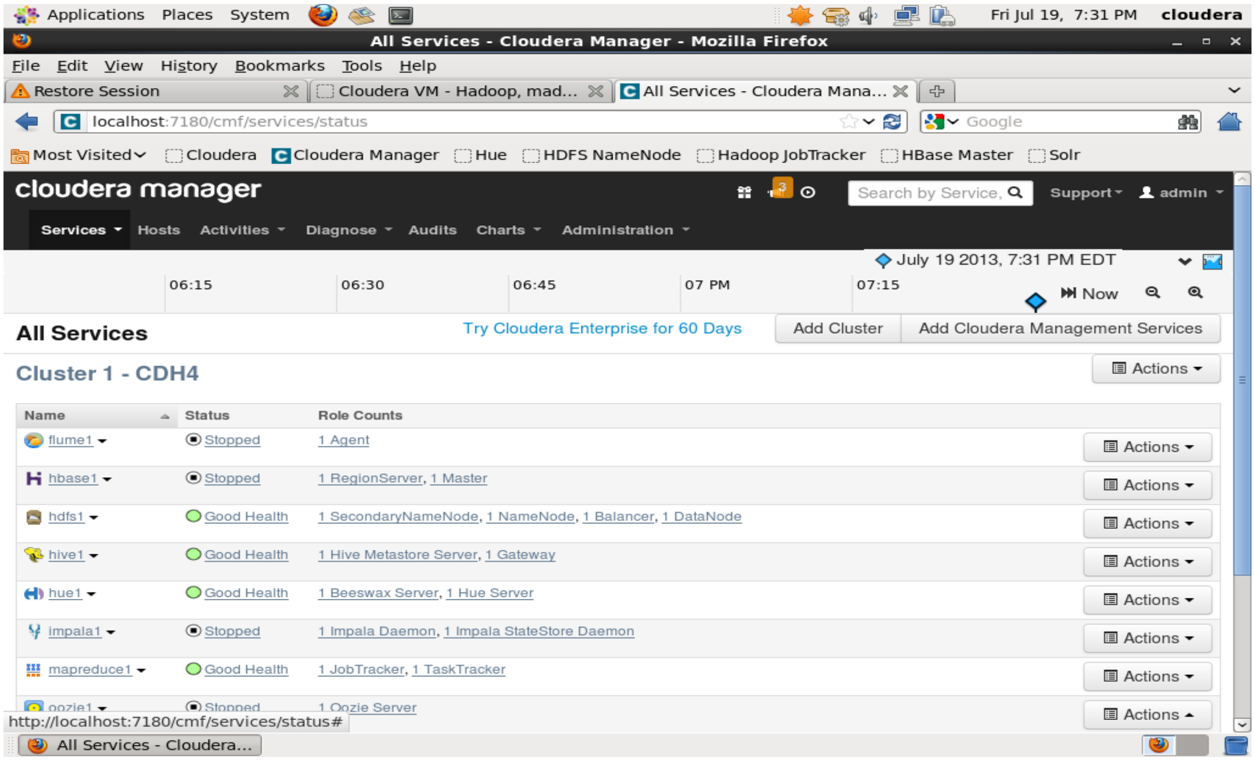
Username: admin

Password: •••••

□ Remember me on this computer.

Login

Step 15: We can see all the services running on our single node cluster.

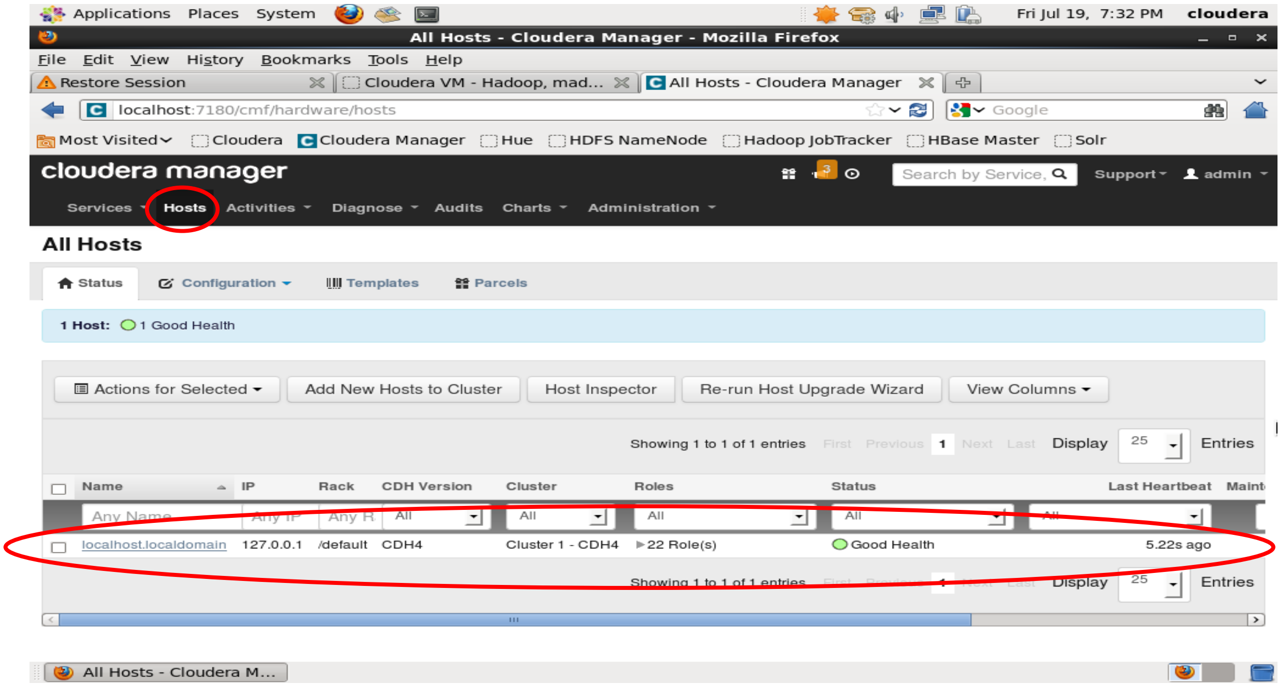

Applications Places System
Fri Jul 19, 7:31 PM cloudera
All Services - Cloudera Manager - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Restore Session
Cloudera VM - Hadoop, mad...
All Services - Cloudera Mana...
localhost:7180/cmf/services/status
Google
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
cloudera manager
Search by Service,
Support admin
Services Hosts Activities Diagnose Audits Charts Administration
July 19 2013, 7:31 PM EDT
06:15
06:30
06:45
07 PM
07:15
Now
Try Cloudera Enterprise for 60 Days
Add Cluster
Add Cloudera Management Services
Actions
http://localhost:7180/cmf/services/status#
All Services - Cloudera...

All Services

Cluster 1 - CDH4

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

Step 16: Click on the Hosts tab and we can see that one host is running , version of CDH installed on it is 4 , health of the host is good and last heart beat was listened 5.22s ago.


Applications Places System
Fri Jul 19, 7:32 PM cloudera
All Hosts - Cloudera Manager - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Restore Session
Cloudera VM - Hadoop, mad...
All Hosts - Cloudera Manager
localhost:7180/cmf/hardware/hosts
Google
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
cloudera manager
Search by Service,
Support admin
Services Hosts Activities Diagnose Audits Charts Administration
Status Configuration Templates Parcels
Actions for Selected
Add New Hosts to Cluster
Host Inspector
Re-run Host Upgrade Wizard
View Columns
First Previous 1 Next Last Display 25 Entries
All Hosts - Cloudera M...

All Hosts

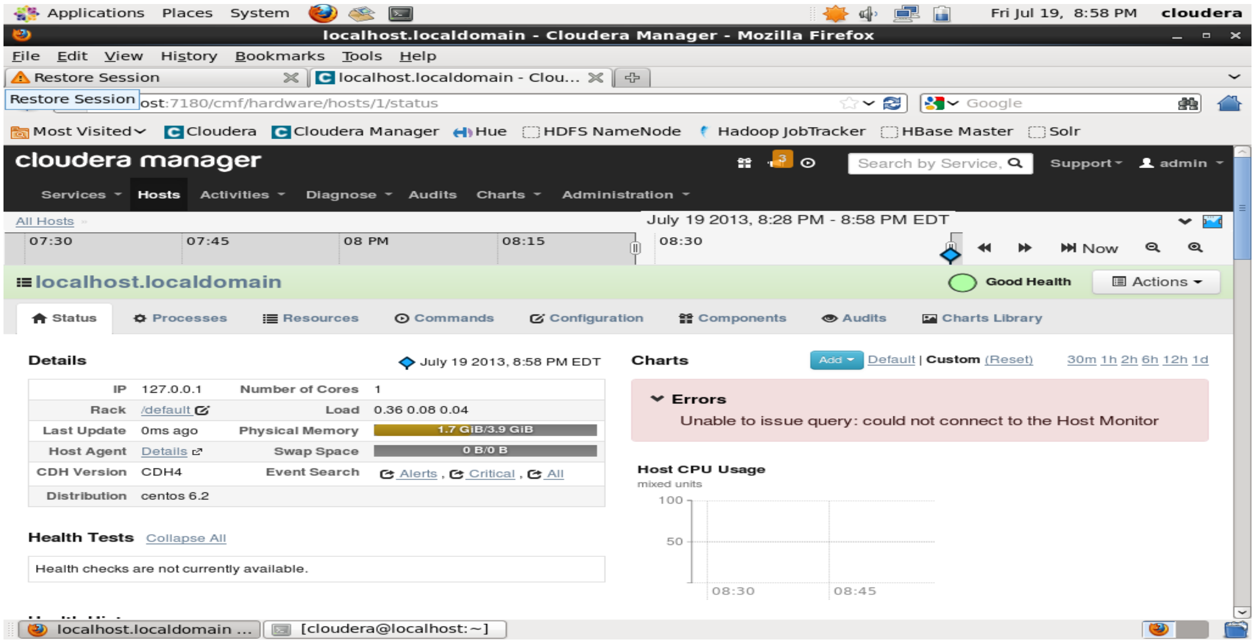
1 Host: 1 Good Health

Showing 1 to 1 of 1 entries

□	Name	IP	Rack	CDH Version	Cluster	Roles	Status	Last Heartbeat	Maint
	Any Name	Any IP	Any R	All	All	All	All	All	
□	localhost.localdomain	127.0.0.1	/default	CDH4	Cluster 1 - CDH4	▶22 Role(s)	Good Health	5.22s ago	

Showing 1 to 1 of 1 entries

Step 17 : Click on the localhost.localdomain to se the detail information about the host



Details July 19 2013, 8:58 PM EDT

IP	127.0.0.1	Number of Cores	1
Rack	/default	Load	0.36 0.08 0.04
Last Update	0ms ago	Physical Memory	1.7 GiB/3.9 GiB
Host Agent	Details	Swap Space	0 B/0 B
CDH Version	CDH4	Event Search	Alerts , Critical , All
Distribution	centos 6.2		

Health Tests Collapse All

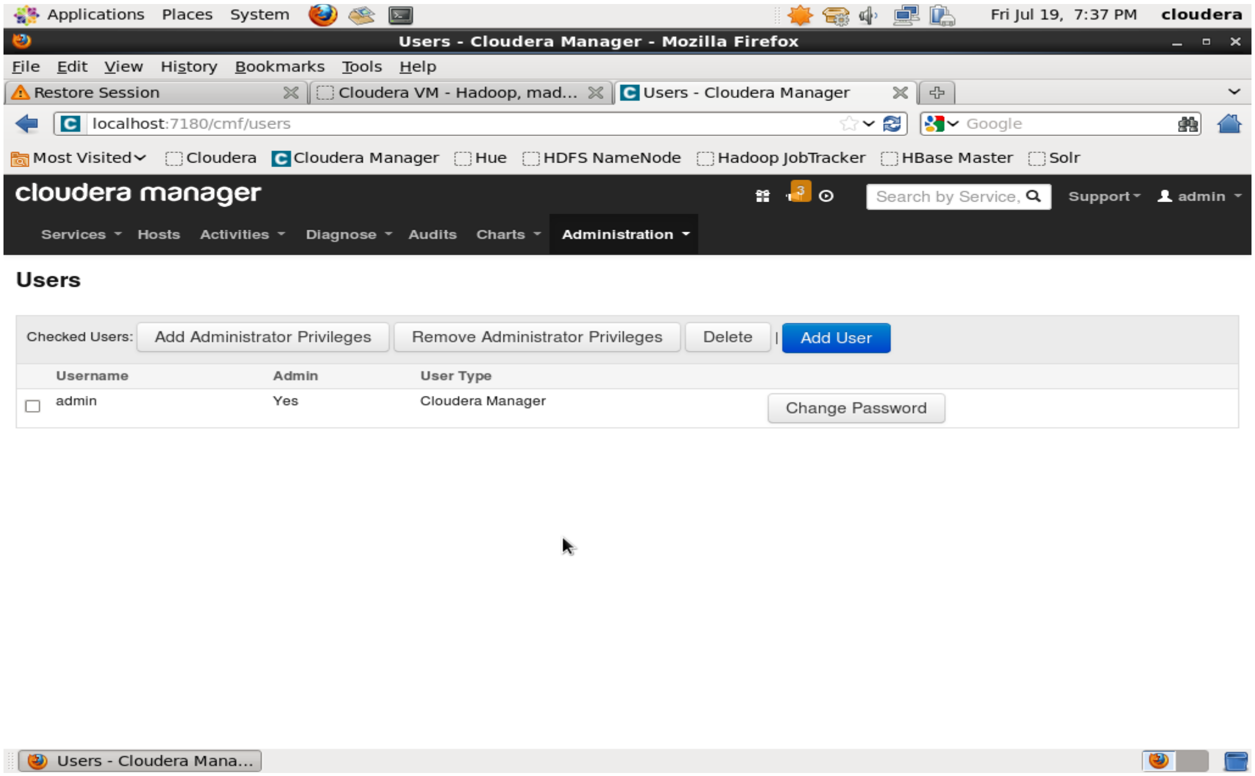
Health checks are not currently available.

Charts

Errors

Unable to issue query: could not connect to the Host Monitor

Step 18 : We can also chang the password for admin by selecting the adminstration tab and clicking on "Change Password" button.



Users

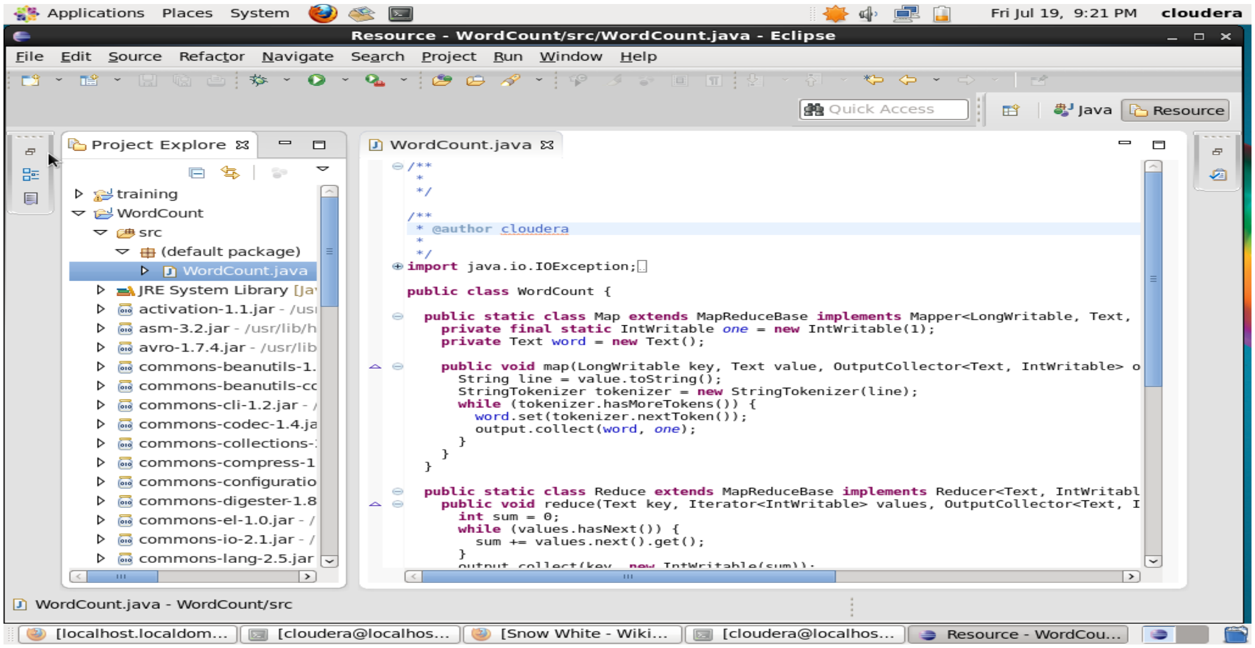
Checked Users: Add Administrator Privileges | Remove Administrator Privileges | Delete | Add User

	Username	Admin	User Type	
□	admin	Yes	Cloudera Manager	Change Password

5. Running MapReduce Program

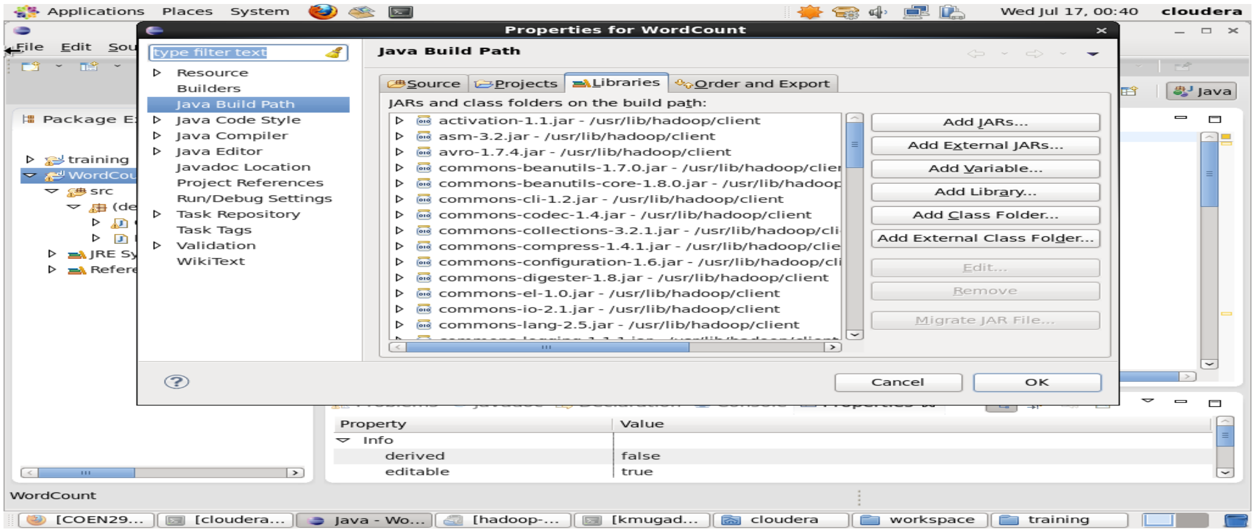
Step 1: Write a MapReduce program. We have used the word count program for testing the CDH installation.

Write a java program using eclipse.



```
/**
 *
 */

/**
 * @author cloudera
 *
 */
import java.io.IOException;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text,
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> o
      String line = value.toString();
      StringTokenizer tokenizer = new StringTokenizer(line);
      while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        output.collect(word, one);
      }
    }
  }

  public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritabl
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, I
      int sum = 0;
      while (values.hasNext()) {
        sum += values.next().get();
      }
      output.collect(key, new IntWritable(sum));
```

Step 2: Add external jars to compile the code. Right click on the project and select properties. Select the Java build path and then click on "Add External Jars". Select all the jars present in folder /usr/bin/hadoop/client

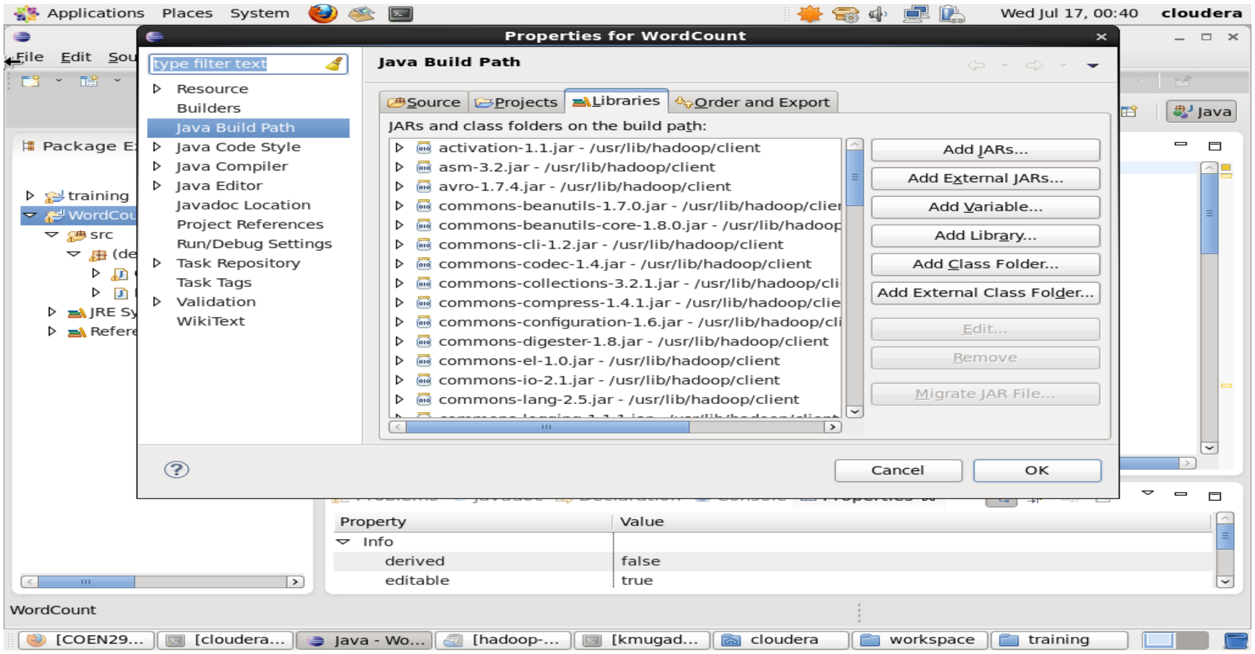


JARs and class folders on the build path:

```
activation-1.1.jar - /usr/lib/hadoop/client
asm-3.2.jar - /usr/lib/hadoop/client
avro-1.7.4.jar - /usr/lib/hadoop/client
commons-beanutils-1.7.0.jar - /usr/lib/hadoop/clier
commons-beanutils-core-1.8.0.jar - /usr/lib/hadoop
commons-cli-1.2.jar - /usr/lib/hadoop/client
commons-codec-1.4.jar - /usr/lib/hadoop/client
commons-collections-3.2.1.jar - /usr/lib/hadoop/cli
commons-compress-1.4.1.jar - /usr/lib/hadoop/clie
commons-configuration-1.6.jar - /usr/lib/hadoop/cli
commons-digester-1.8.jar - /usr/lib/hadoop/client
commons-el-1.0.jar - /usr/lib/hadoop/client
commons-io-2.1.jar - /usr/lib/hadoop/client
commons-lang-2.5.jar - /usr/lib/hadoop/client
```

Property	Value
Info	
derived	false
editable	true

Step 3: Click “OK” o add those jars to your program.



Java Build Path

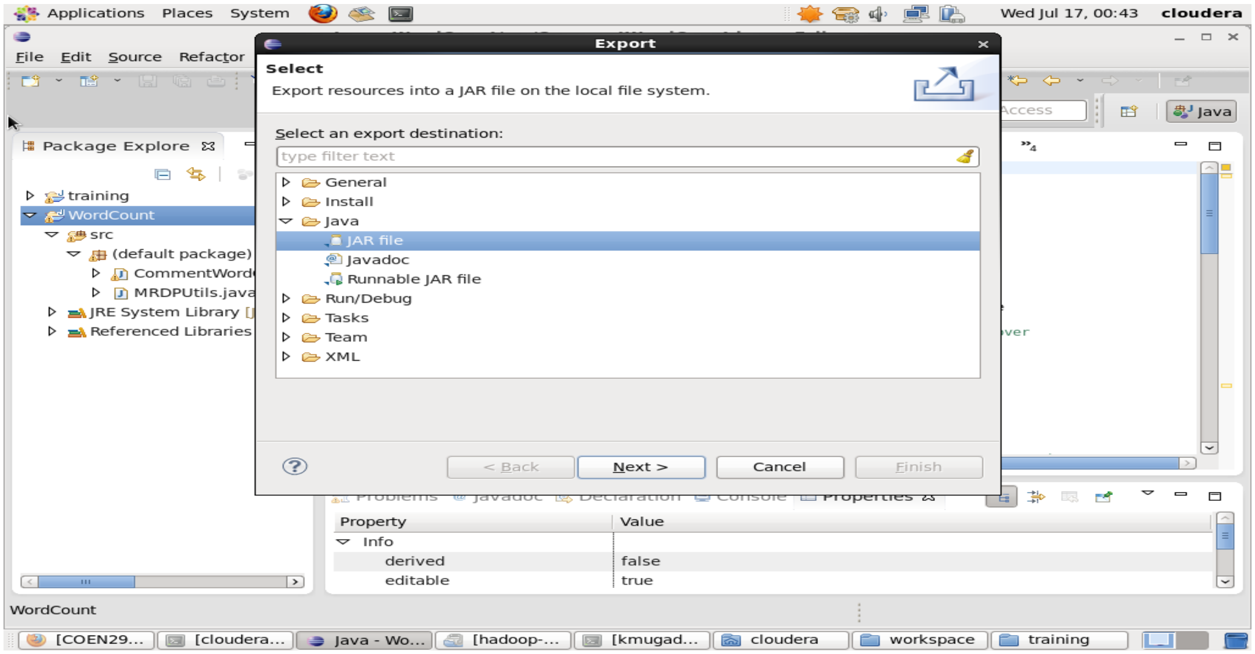
JARs and class folders on the build path:

```
activation-1.1.jar - /usr/lib/hadoop/client
asm-3.2.jar - /usr/lib/hadoop/client
avro-1.7.4.jar - /usr/lib/hadoop/client
commons-beanutils-1.7.0.jar - /usr/lib/hadoop/clier
commons-beanutils-core-1.8.0.jar - /usr/lib/hadoop
commons-cli-1.2.jar - /usr/lib/hadoop/client
commons-codec-1.4.jar - /usr/lib/hadoop/client
commons-collections-3.2.1.jar - /usr/lib/hadoop/cli
commons-compress-1.4.1.jar - /usr/lib/hadoop/clie
commons-configuration-1.6.jar - /usr/lib/hadoop/cli
commons-digester-1.8.jar - /usr/lib/hadoop/client
commons-el-1.0.jar - /usr/lib/hadoop/client
commons-io-2.1.jar - /usr/lib/hadoop/client
commons-lang-2.5.jar - /usr/lib/hadoop/client
```

Property	Value
Info	
derived	false
editable	true

Step 4 : Create a jar file of your program.

Right click on project select “Export” and then click on “Jar File” under Java folder.



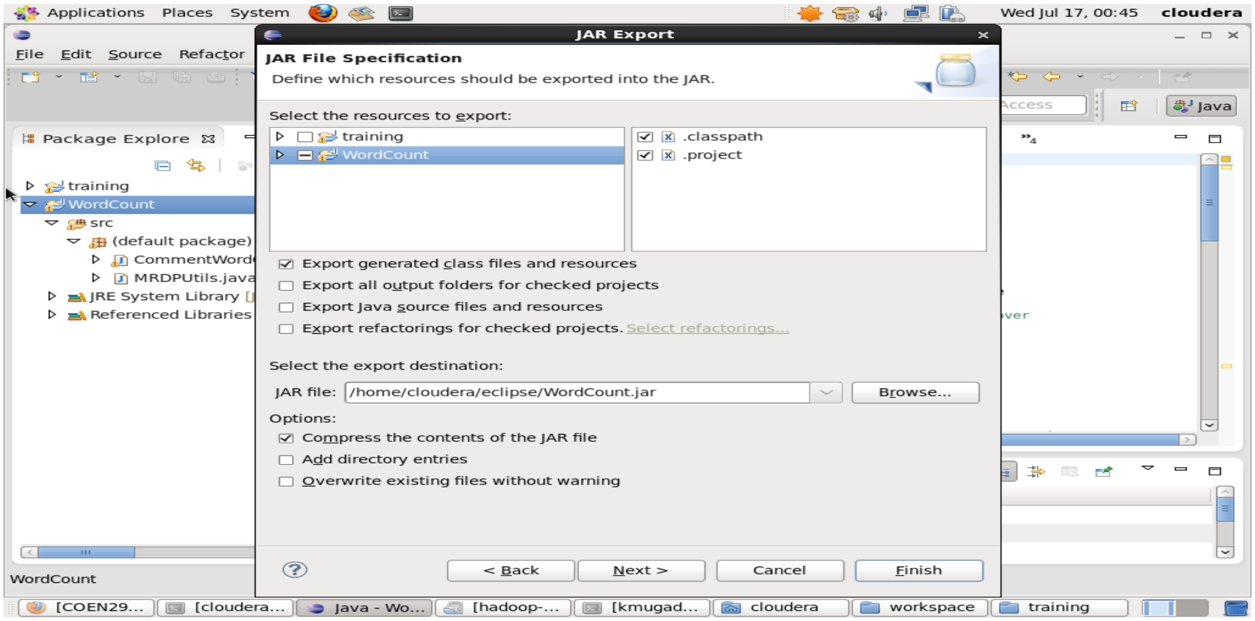
Select

Export resources into a JAR file on the local file system.

Select an export destination:

Property	Value
Info	
derived	false
editable	true

Give the location path where you want to store your .jar file.



JAR File Specification

Define which resources should be exported into the JAR.

Select the resources to export:

- ☐ training
- ☐ WordCount
- ☑ .classpath
- ☑ .project

☑ Export generated class files and resources
☐ Export all output folders for checked projects
☐ Export Java source files and resources
☐ Export refactorings for checked projects. Select refactorings...

Select the export destination:

JAR file: /home/cloudera/eclipse/WordCount.jar

Options:

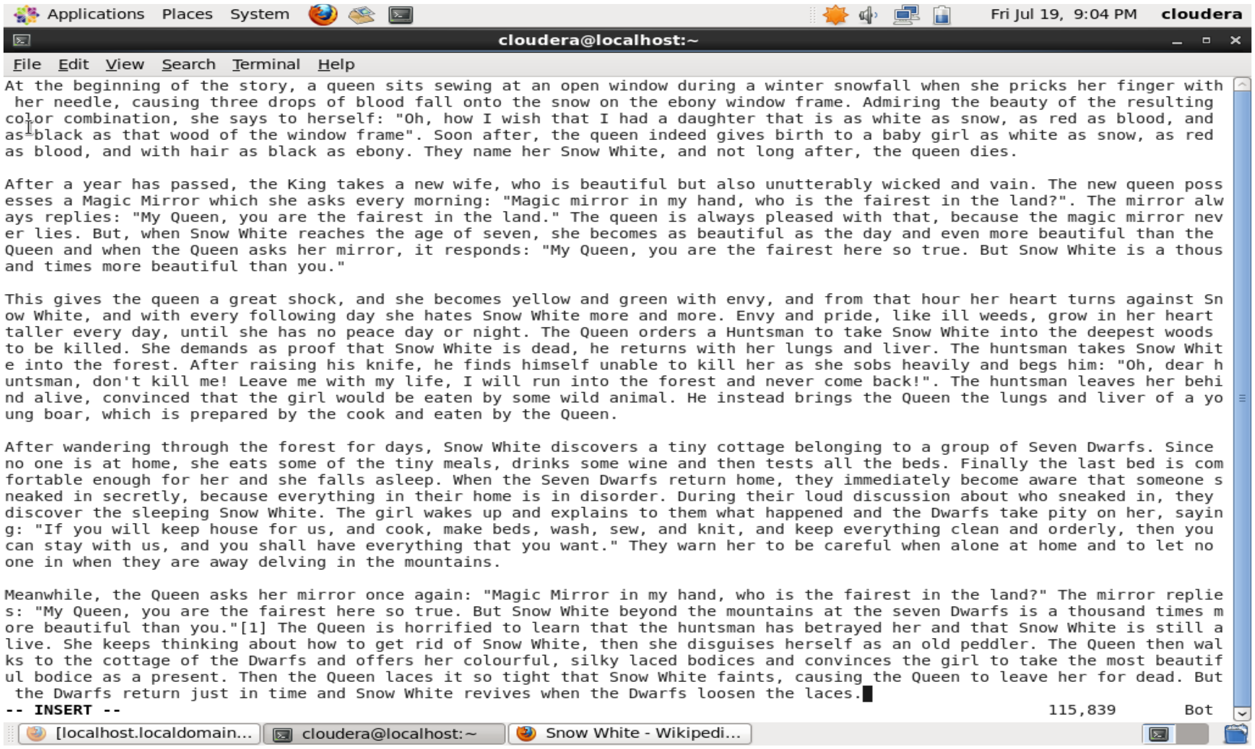
☑ Compress the contents of the JAR file
☐ Add directory entries
☐ Overwrite existing files without warning

Now we can see the jar file in the given location.

Step 5: Input file

Open the terminal and create a input file which is a huge text file.

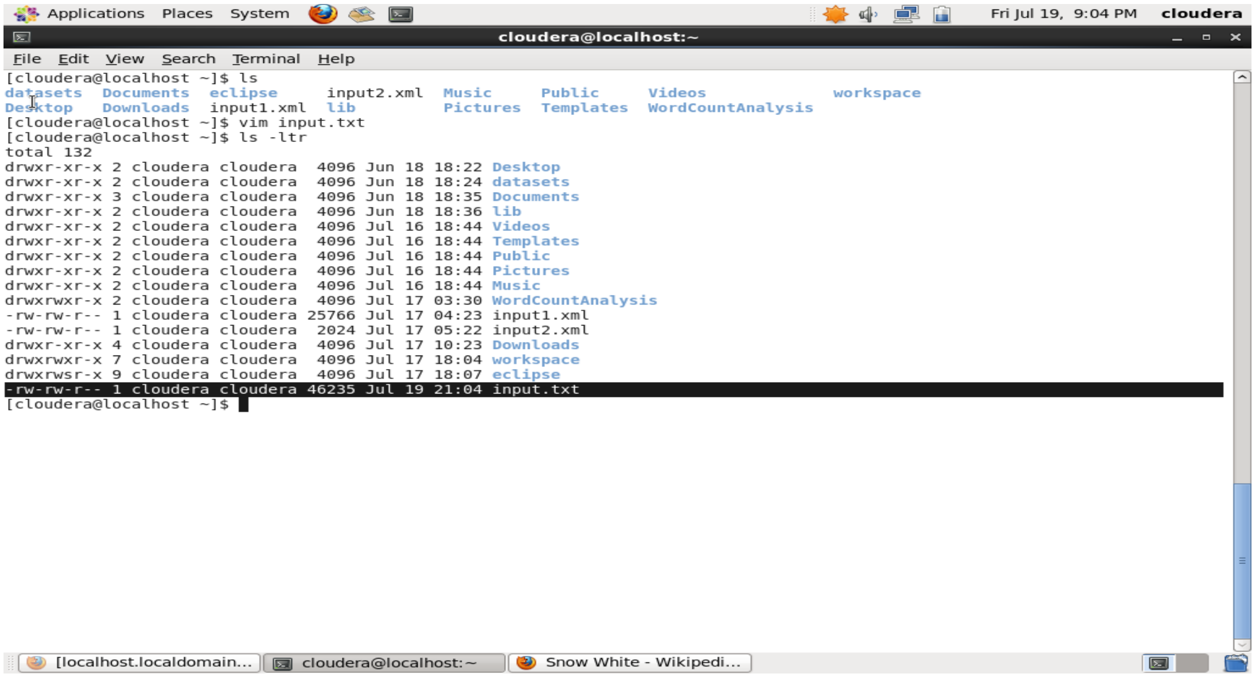
$vim input.txt



```
At the beginning of the story, a queen sits sewing at an open window during a winter snowfall when she pricks her finger with
 her needle, causing three drops of blood fall onto the snow on the ebony window frame. Admiring the beauty of the resulting
color combination, she says to herself: "Oh, how I wish that I had a daughter that is as white as snow, as red as blood, and
as black as that wood of the window frame". Soon after, the queen indeed gives birth to a baby girl as white as snow, as red
as blood, and with hair as black as ebony. They name her Snow White, and not long after, the queen dies.

After a year has passed, the King takes a new wife, who is beautiful but also unutterably wicked and vain. The new queen poss
esses a Magic Mirror which she asks every morning: "Magic mirror in my hand, who is the fairest in the land?". The mirror alw
ays replies: "My Queen, you are the fairest in the land." The queen is always pleased with that, because the magic mirror nev
er lies. But, when Snow White reaches the age of seven, she becomes as beautiful as the day and even more beautiful than the
Queen and when the Queen asks her mirror, it responds: "My Queen, you are the fairest here so true. But Snow White is a thous
and times more beautiful than you."

This gives the queen a great shock, and she becomes yellow and green with envy, and from that hour her heart turns against Sn
ow White, and with every following day she hates Snow White more and more. Envy and pride, like ill weeds, grow in her heart
taller every day, until she has no peace day or night. The Queen orders a Huntsman to take Snow White into the deepest woods
to be killed. She demands as proof that Snow White is dead, he returns with her lungs and liver. The huntsman takes Snow Whit
e into the forest. After raising his knife, he finds himself unable to kill her as she sobs heavily and begs him: "Oh, dear h
untsman, don't kill me! Leave me with my life, I will run into the forest and never come back!". The huntsman leaves her behi
nd alive, convinced that the girl would be eaten by some wild animal. He instead brings the Queen the lungs and liver of a yo
ung boar, which is prepared by the cook and eaten by the Queen.

After wandering through the forest for days, Snow White discovers a tiny cottage belonging to a group of Seven Dwarfs. Since
no one is at home, she eats some of the tiny meals, drinks some wine and then tests all the beds. Finally the last bed is com
fortable enough for her and she falls asleep. When the Seven Dwarfs return home, they immediately become aware that someone s
neaked in secretly, because everything in their home is in disorder. During their loud discussion about who sneaked in, they
discover the sleeping Snow White. The girl wakes up and explains to them what happened and the Dwarfs take pity on her, sayin
g: "If you will keep house for us, and cook, make beds, wash, sew, and knit, and keep everything clean and orderly, then you
can stay with us, and you shall have everything that you want." They warn her to be careful when alone at home and to let no
one in when they are away delving in the mountains.

Meanwhile, the Queen asks her mirror once again: "Magic Mirror in my hand, who is the fairest in the land?" The mirror replie
s: "My Queen, you are the fairest here so true. But Snow White beyond the mountains at the seven Dwarfs is a thousand times m
ore beautiful than you."[1] The Queen is horrified to learn that the huntsman has betrayed her and that Snow White is still a
live. She keeps thinking about how to get rid of Snow White, then she disguises herself as an old peddler. The Queen then wal
ks to the cottage of the Dwarfs and offers her colourful, silky laced bodices and convinces the girl to take the most beautif
ul bodice as a present. Then the Queen laces it so tight that Snow White faints, causing the Queen to leave her for dead. But
 the Dwarfs return just in time and Snow White revives when the Dwarfs loosen the laces.
-- INSERT --                                                                                          115,839       Bot
```

Applications Places System
Fri Jul 19, 9:04 PM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
[localhost.localdomain...
cloudera@localhost:~
Snow White - Wikipedi...

Step 6 : Check the input.txt. Execute command $ ls –ltr. The highlighted file is our input file.



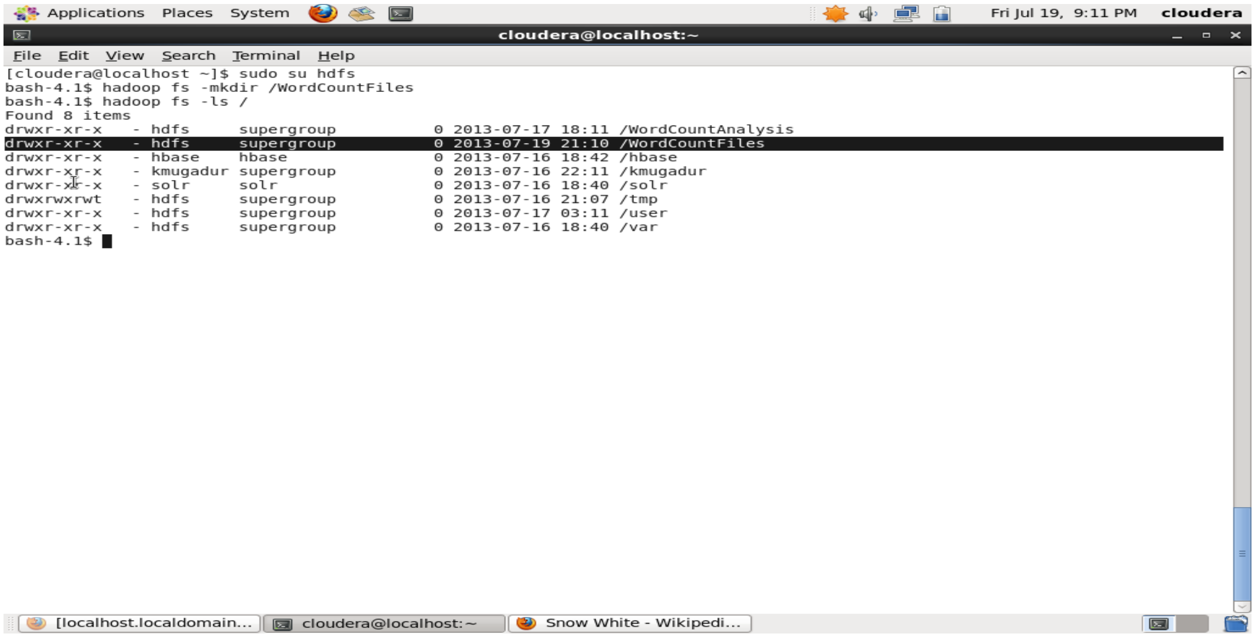
```
[cloudera@localhost ~]$ ls
datasets  Documents  eclipse     input2.xml  Music     Public     Videos              workspace
Desktop   Downloads  input1.xml  lib         Pictures  Templates  WordCountAnalysis
[cloudera@localhost ~]$ vim input.txt
[cloudera@localhost ~]$ ls -ltr
total 132
drwxr-xr-x 2 cloudera cloudera  4096 Jun 18 18:22 Desktop
drwxr-xr-x 2 cloudera cloudera  4096 Jun 18 18:24 datasets
drwxr-xr-x 3 cloudera cloudera  4096 Jun 18 18:35 Documents
drwxr-xr-x 2 cloudera cloudera  4096 Jun 18 18:36 lib
drwxr-xr-x 2 cloudera cloudera  4096 Jul 16 18:44 Videos
drwxr-xr-x 2 cloudera cloudera  4096 Jul 16 18:44 Templates
drwxr-xr-x 2 cloudera cloudera  4096 Jul 16 18:44 Public
drwxr-xr-x 2 cloudera cloudera  4096 Jul 16 18:44 Pictures
drwxr-xr-x 2 cloudera cloudera  4096 Jul 16 18:44 Music
drwxrwxr-x 2 cloudera cloudera  4096 Jul 17 03:30 WordCountAnalysis
-rw-rw-r-- 1 cloudera cloudera 25766 Jul 17 04:23 input1.xml
-rw-rw-r-- 1 cloudera cloudera  2024 Jul 17 05:22 input2.xml
drwxr-xr-x 4 cloudera cloudera  4096 Jul 17 10:23 Downloads
drwxrwxr-x 7 cloudera cloudera  4096 Jul 17 18:04 workspace
drwxrwsr-x 9 cloudera cloudera  4096 Jul 17 18:07 eclipse
-rw-rw-r-- 1 cloudera cloudera 46235 Jul 19 21:04 input.txt
[cloudera@localhost ~]$ 
```

Step 7 : Make a new file directory on HDFS (Hadoop Distributed File System)

$ sudo su hdfs

hadoop fs –mkdir /WordCountFiles

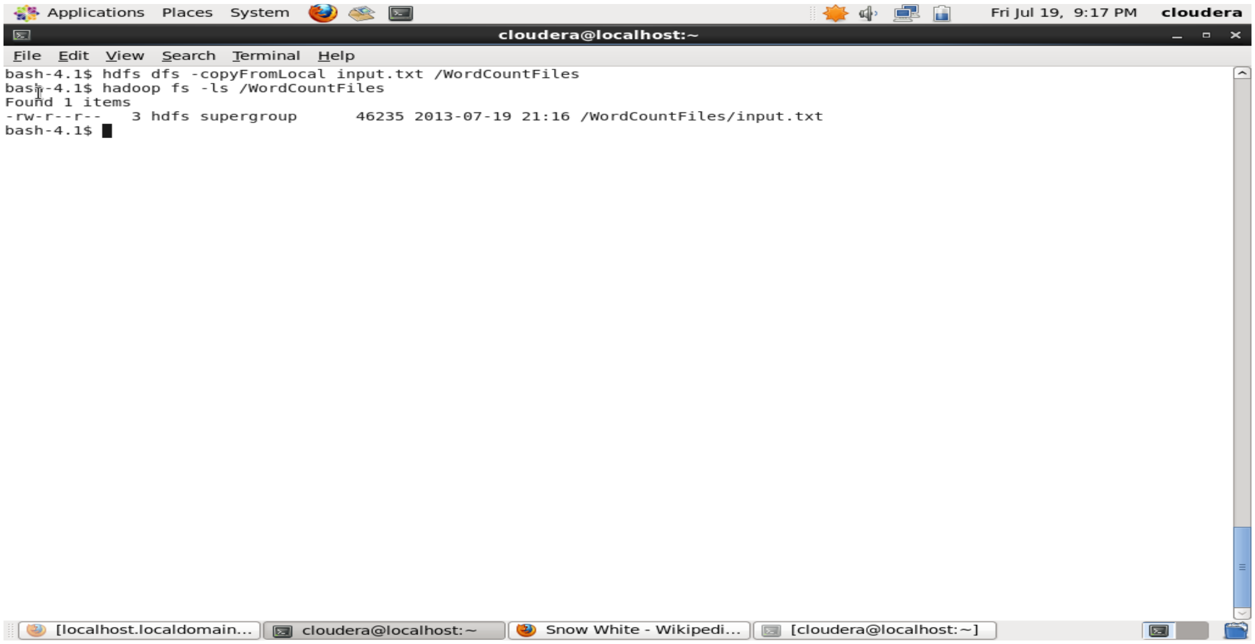
hadoop fs –ls /



```
[cloudera@localhost ~]$ sudo su hdfs
bash-4.1$ hadoop fs -mkdir /WordCountFiles
bash-4.1$ hadoop fs -ls /
Found 8 items
drwxr-xr-x   - hdfs     supergroup          0 2013-07-17 18:11 /WordCountAnalysis
drwxr-xr-x   - hdfs     supergroup          0 2013-07-19 21:10 /WordCountFiles
drwxr-xr-x   - hbase    hbase               0 2013-07-16 18:42 /hbase
drwxr-xr-x   - kmugadur supergroup          0 2013-07-16 22:11 /kmugadur
drwxr-xr-x   - solr     solr                0 2013-07-16 18:40 /solr
drwxrwxrwt   - hdfs     supergroup          0 2013-07-16 21:07 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2013-07-17 03:11 /user
drwxr-xr-x   - hdfs     supergroup          0 2013-07-16 18:40 /var
bash-4.1$ 
```

Step 8 :Copy this file on the NameNode i.e., on HDFS

$ hdfs dfs –copyFromLocal input.txt /WordCountFiles



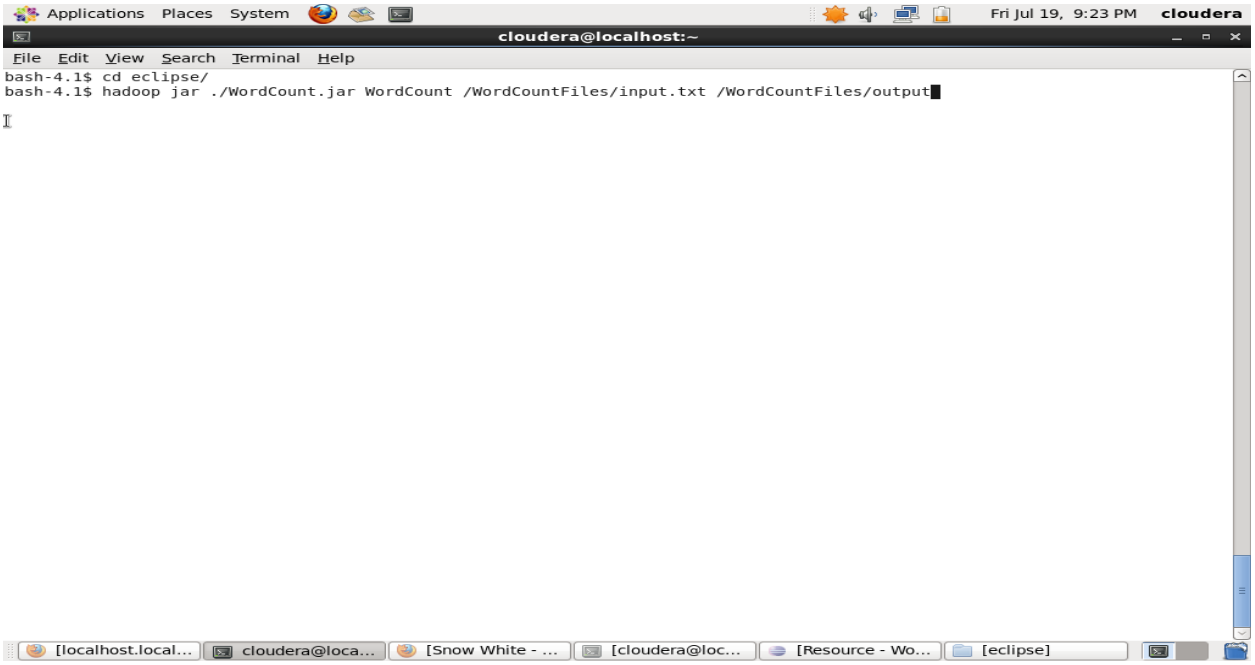
Applications Places System
Fri Jul 19, 9:17 PM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
[localhost.localdomain...
cloudera@localhost:~
Snow White - Wikipedi...
[cloudera@localhost:~]

```
bash-4.1$ hdfs dfs -copyFromLocal input.txt /WordCountFiles
bash-4.1$ hadoop fs -ls /WordCountFiles
Found 1 items
-rw-r--r--   3 hdfs supergroup      46235 2013-07-19 21:16 /WordCountFiles/input.txt
bash-4.1$ 
```

Step 9: Run the program using the hadoop command

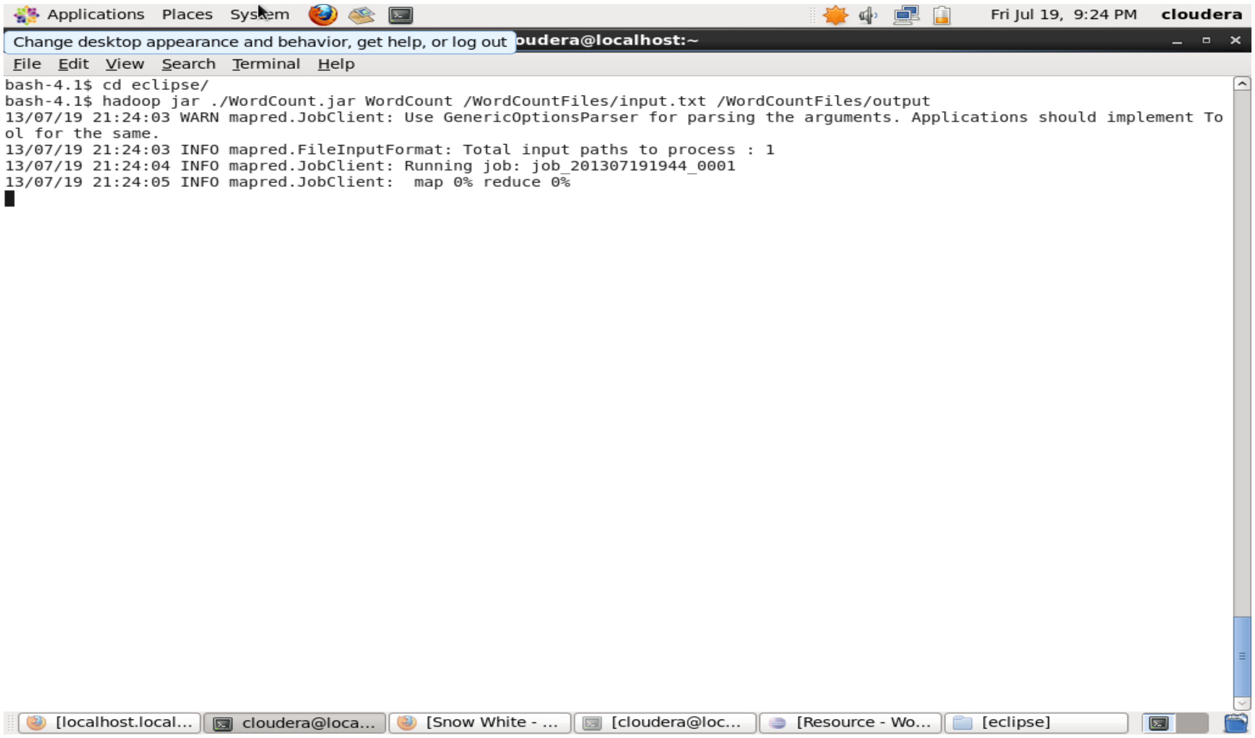
$ hadoop jar ./WordCount.jar WordCount /WordCountFiles/input.txt /WordCountFiles/output

Where ./WordCount.jar is the path and name of the jar file we created in Step 4 and WordCount is the name of the program.


Applications Places System
Fri Jul 19, 9:23 PM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
[localhost.local... cloudera@loca... [Snow White - ... [cloudera@loc... [Resource - Wo... [eclipse]

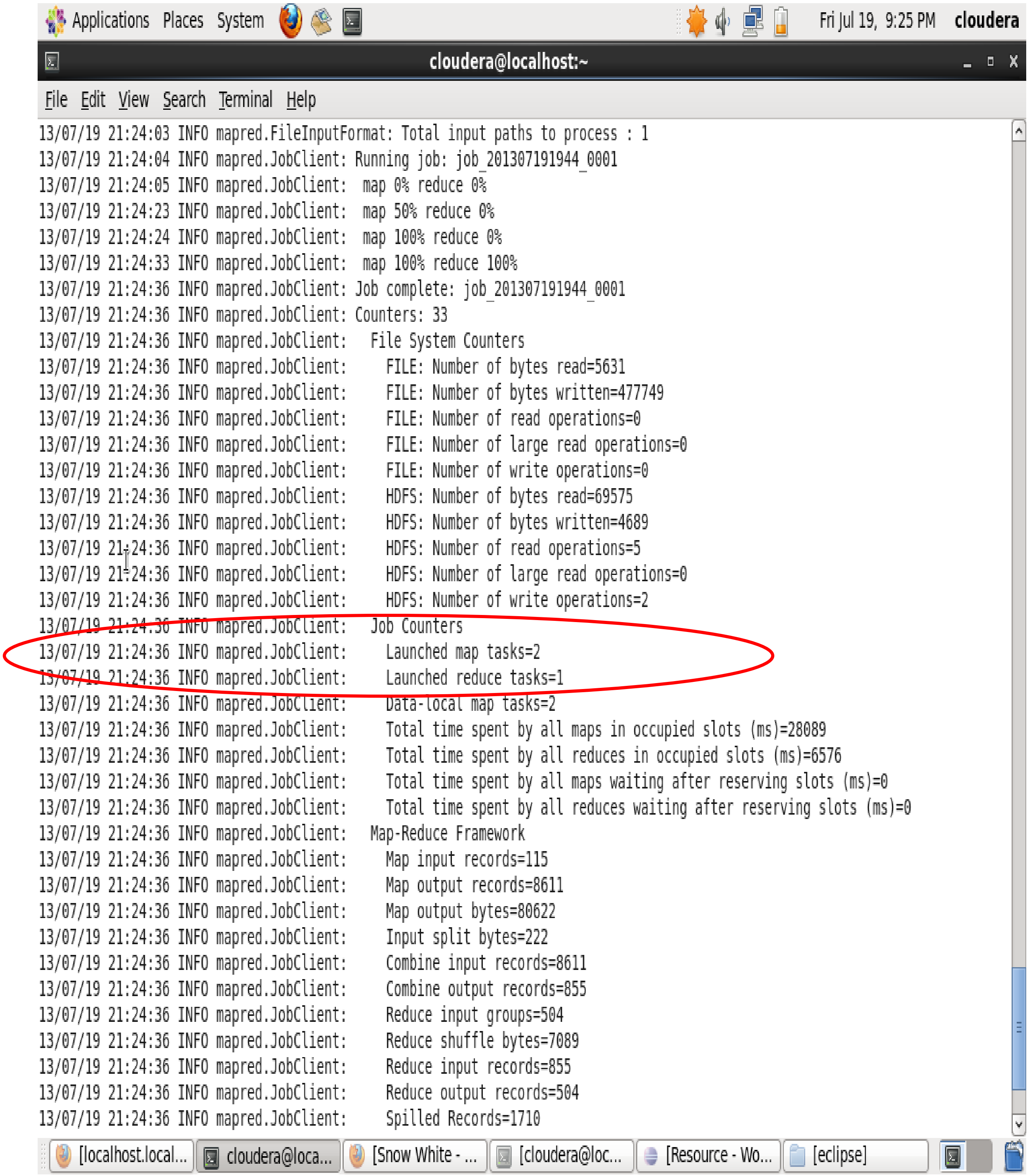
```
bash-4.1$ cd eclipse/
bash-4.1$ hadoop jar ./WordCount.jar WordCount /WordCountFiles/input.txt /WordCountFiles/output
```

MapReduce program starts to run.We can see the percentage of mapping and reducing the program is doing on the command line.


Applications Places System
Fri Jul 19, 9:24 PM cloudera
Change desktop appearance and behavior, get help, or log out
cloudera@localhost:~
File Edit View Search Terminal Help
[localhost.local... cloudera@loca... [Snow White - ... [cloudera@loc... [Resource - Wo... [eclipse]

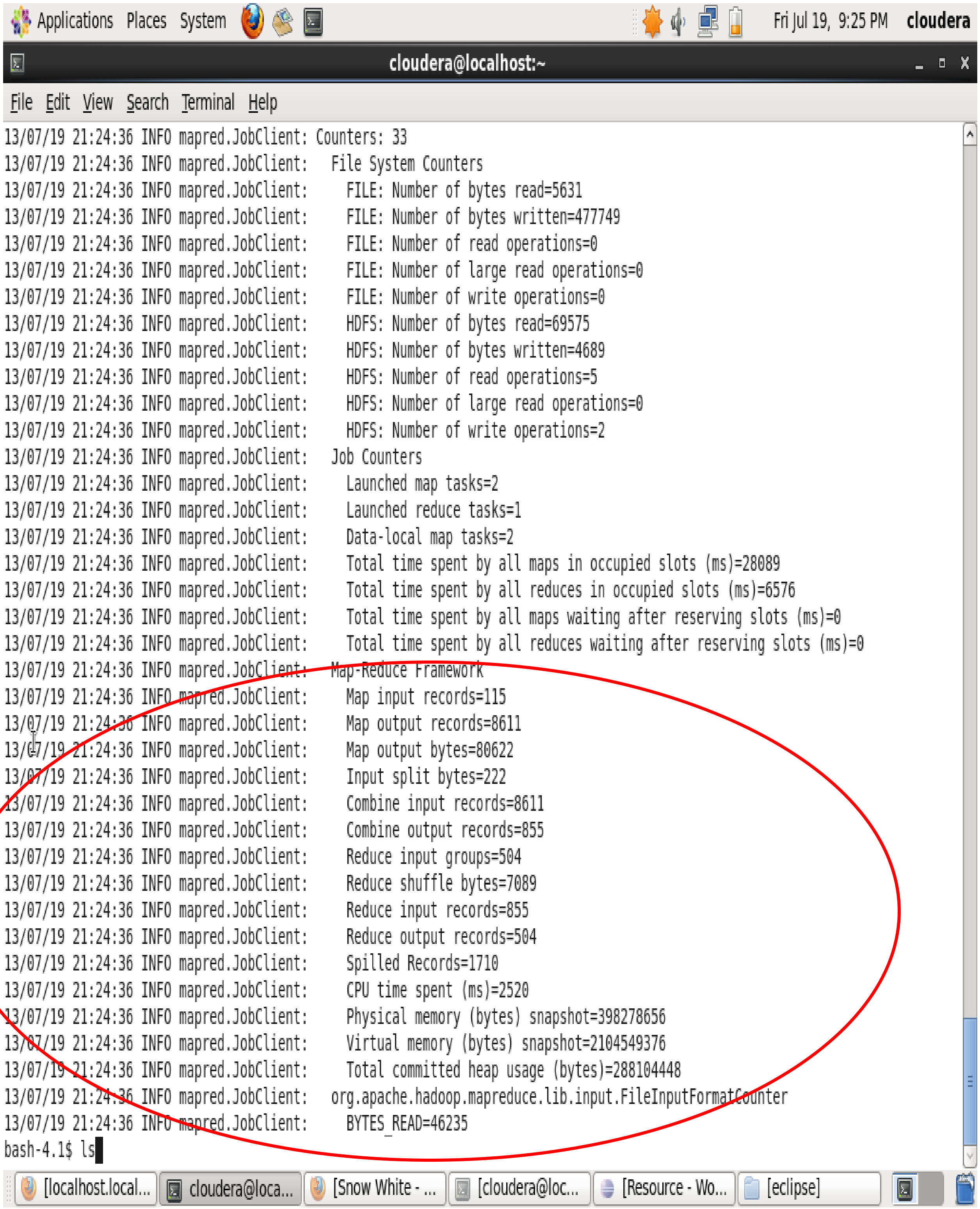
```
bash-4.1$ cd eclipse/
bash-4.1$ hadoop jar ./WordCount.jar WordCount /WordCountFiles/input.txt /WordCountFiles/output
13/07/19 21:24:03 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement To
ol for the same.
13/07/19 21:24:03 INFO mapred.FileInputFormat: Total input paths to process : 1
13/07/19 21:24:04 INFO mapred.JobClient: Running job: job_201307191944_0001
13/07/19 21:24:05 INFO mapred.JobClient:  map 0% reduce 0%
```

We can see that the map and reduce percentage gradually increasing, this shows that the program is successfully running on CDH and using its MapReduce technique to count the frequency of each words.



```
13/07/19 21:24:03 INFO mapred.FileInputFormat: Total input paths to process : 1
13/07/19 21:24:04 INFO mapred.JobClient: Running job: job_201307191944_0001
13/07/19 21:24:05 INFO mapred.JobClient:  map 0% reduce 0%
13/07/19 21:24:23 INFO mapred.JobClient:  map 50% reduce 0%
13/07/19 21:24:24 INFO mapred.JobClient:  map 100% reduce 0%
13/07/19 21:24:33 INFO mapred.JobClient:  map 100% reduce 100%
13/07/19 21:24:36 INFO mapred.JobClient: Job complete: job_201307191944_0001
13/07/19 21:24:36 INFO mapred.JobClient: Counters: 33
13/07/19 21:24:36 INFO mapred.JobClient:   File System Counters
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of bytes read=5631
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of bytes written=477749
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of large read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of write operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of bytes read=69575
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of bytes written=4689
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of read operations=5
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of large read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of write operations=2
13/07/19 21:24:36 INFO mapred.JobClient:   Job Counters
13/07/19 21:24:36 INFO mapred.JobClient:     Launched map tasks=2
13/07/19 21:24:36 INFO mapred.JobClient:     Launched reduce tasks=1
13/07/19 21:24:36 INFO mapred.JobClient:     Data-local map tasks=2
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=28089
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=6576
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/07/19 21:24:36 INFO mapred.JobClient:   Map-Reduce Framework
13/07/19 21:24:36 INFO mapred.JobClient:     Map input records=115
13/07/19 21:24:36 INFO mapred.JobClient:     Map output records=8611
13/07/19 21:24:36 INFO mapred.JobClient:     Map output bytes=80622
13/07/19 21:24:36 INFO mapred.JobClient:     Input split bytes=222
13/07/19 21:24:36 INFO mapred.JobClient:     Combine input records=8611
13/07/19 21:24:36 INFO mapred.JobClient:     Combine output records=855
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce input groups=504
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce shuffle bytes=7089
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce input records=855
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce output records=504
13/07/19 21:24:36 INFO mapred.JobClient:     Spilled Records=1710
```

When the program runs we can see on the command line the number of input bytes the program has read and number of tasks launched and other useful information on the command line.



```
13/07/19 21:24:36 INFO mapred.JobClient: Counters: 33
13/07/19 21:24:36 INFO mapred.JobClient:   File System Counters
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of bytes read=5631
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of bytes written=477749
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of large read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     FILE: Number of write operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of bytes read=69575
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of bytes written=4689
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of read operations=5
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of large read operations=0
13/07/19 21:24:36 INFO mapred.JobClient:     HDFS: Number of write operations=2
13/07/19 21:24:36 INFO mapred.JobClient:   Job Counters
13/07/19 21:24:36 INFO mapred.JobClient:     Launched map tasks=2
13/07/19 21:24:36 INFO mapred.JobClient:     Launched reduce tasks=1
13/07/19 21:24:36 INFO mapred.JobClient:     Data-local map tasks=2
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=28089
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=6576
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/07/19 21:24:36 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/07/19 21:24:36 INFO mapred.JobClient:   Map-Reduce Framework
13/07/19 21:24:36 INFO mapred.JobClient:     Map input records=115
13/07/19 21:24:36 INFO mapred.JobClient:     Map output records=8611
13/07/19 21:24:36 INFO mapred.JobClient:     Map output bytes=80622
13/07/19 21:24:36 INFO mapred.JobClient:     Input split bytes=222
13/07/19 21:24:36 INFO mapred.JobClient:     Combine input records=8611
13/07/19 21:24:36 INFO mapred.JobClient:     Combine output records=855
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce input groups=504
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce shuffle bytes=7089
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce input records=855
13/07/19 21:24:36 INFO mapred.JobClient:     Reduce output records=504
13/07/19 21:24:36 INFO mapred.JobClient:     Spilled Records=1710
13/07/19 21:24:36 INFO mapred.JobClient:     CPU time spent (ms)=2520
13/07/19 21:24:36 INFO mapred.JobClient:     Physical memory (bytes) snapshot=398278656
13/07/19 21:24:36 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=2104549376
13/07/19 21:24:36 INFO mapred.JobClient:     Total committed heap usage (bytes)=288104448
13/07/19 21:24:36 INFO mapred.JobClient:   org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter
13/07/19 21:24:36 INFO mapred.JobClient:     BYTES_READ=46235
bash-4.1$ ls
```

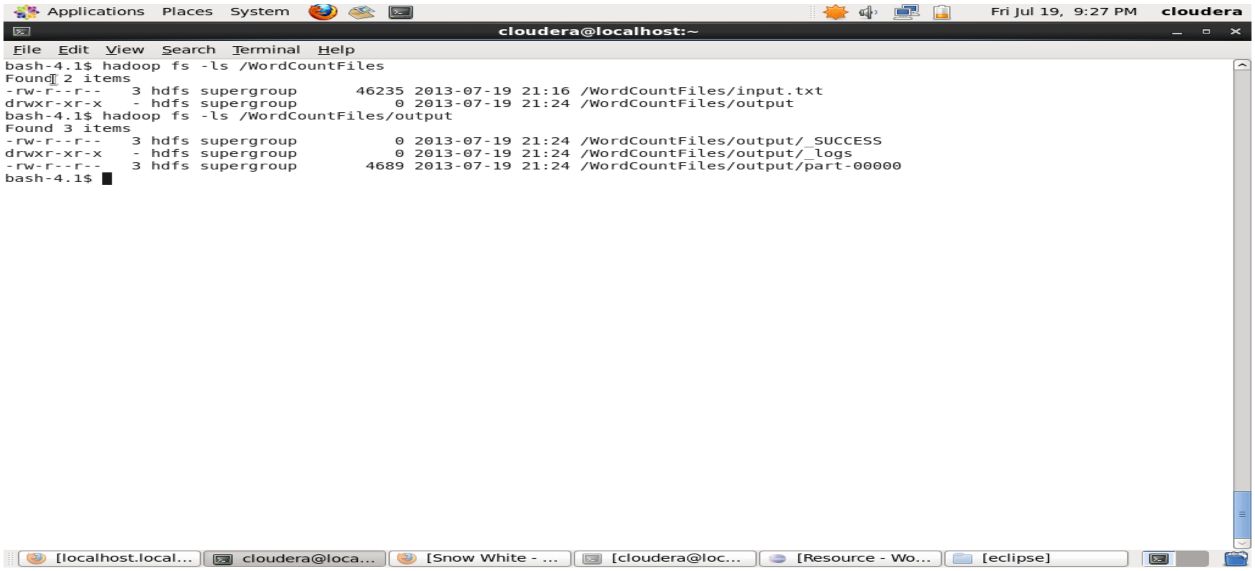
Step 10 : Check the output

When the program runs successfully output directory is created. In our case the output directory name is "output" (as mentioned in the command)

$ hadoop fs –ls /WordCountFiles

$ hadoop fs –ls /WordCountFiles/ouput

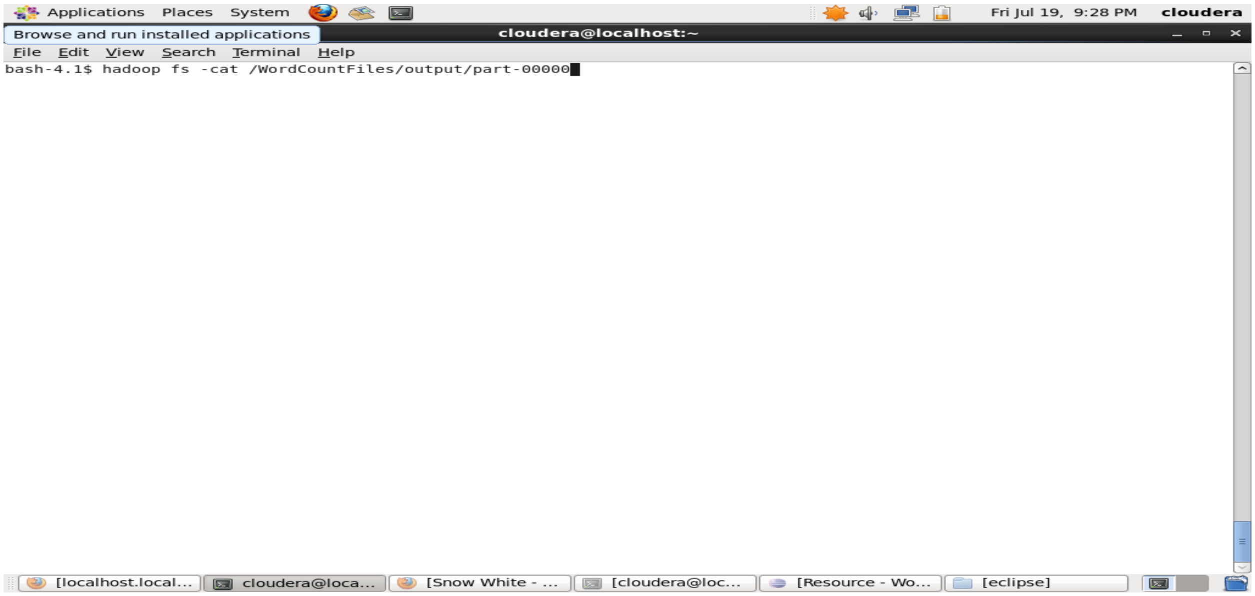
We can see that there are three files/directory in the output directory. Our output is present in part-0000 file



```
bash-4.1$ hadoop fs -ls /WordCountFiles
Found 2 items
-rw-r--r--   3 hdfs supergroup      46235 2013-07-19 21:16 /WordCountFiles/input.txt
drwxr-xr-x   - hdfs supergroup          0 2013-07-19 21:24 /WordCountFiles/output
bash-4.1$ hadoop fs -ls /WordCountFiles/output
Found 3 items
-rw-r--r--   3 hdfs supergroup          0 2013-07-19 21:24 /WordCountFiles/output/_SUCCESS
drwxr-xr-x   - hdfs supergroup          0 2013-07-19 21:24 /WordCountFiles/output/_logs
-rw-r--r--   3 hdfs supergroup       4689 2013-07-19 21:24 /WordCountFiles/output/part-00000
bash-4.1$ 
```

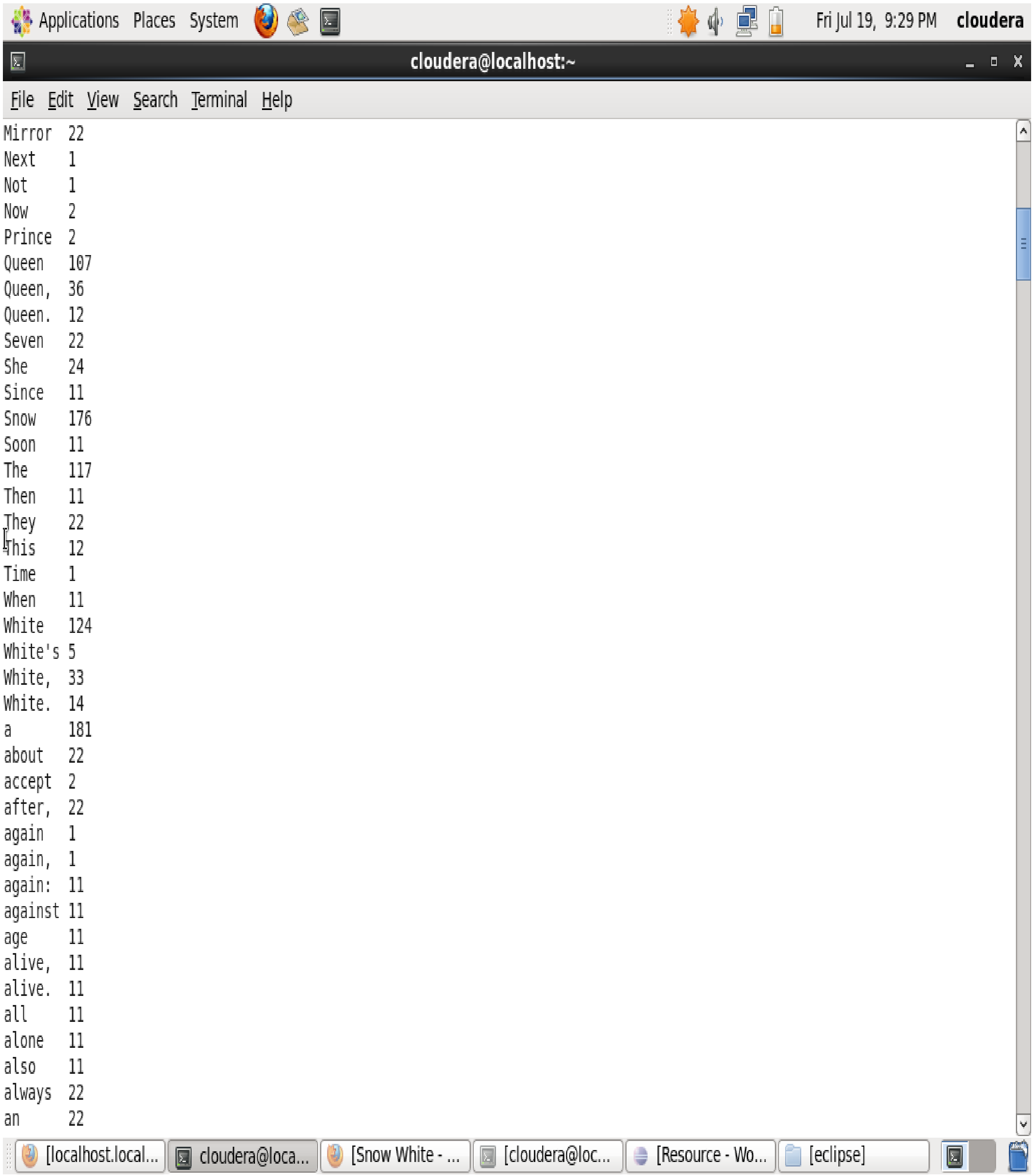
Step 11 : Display the output

$hadoop fs –cat /WordCountFiles/ouput/part-0000



```
bash-4.1$ hadoop fs -cat /WordCountFiles/output/part-00000
```

Output file shows the word and the number of times it has occurred in the file. For example , word "Mirror" has occurred 22 times in the given input file.

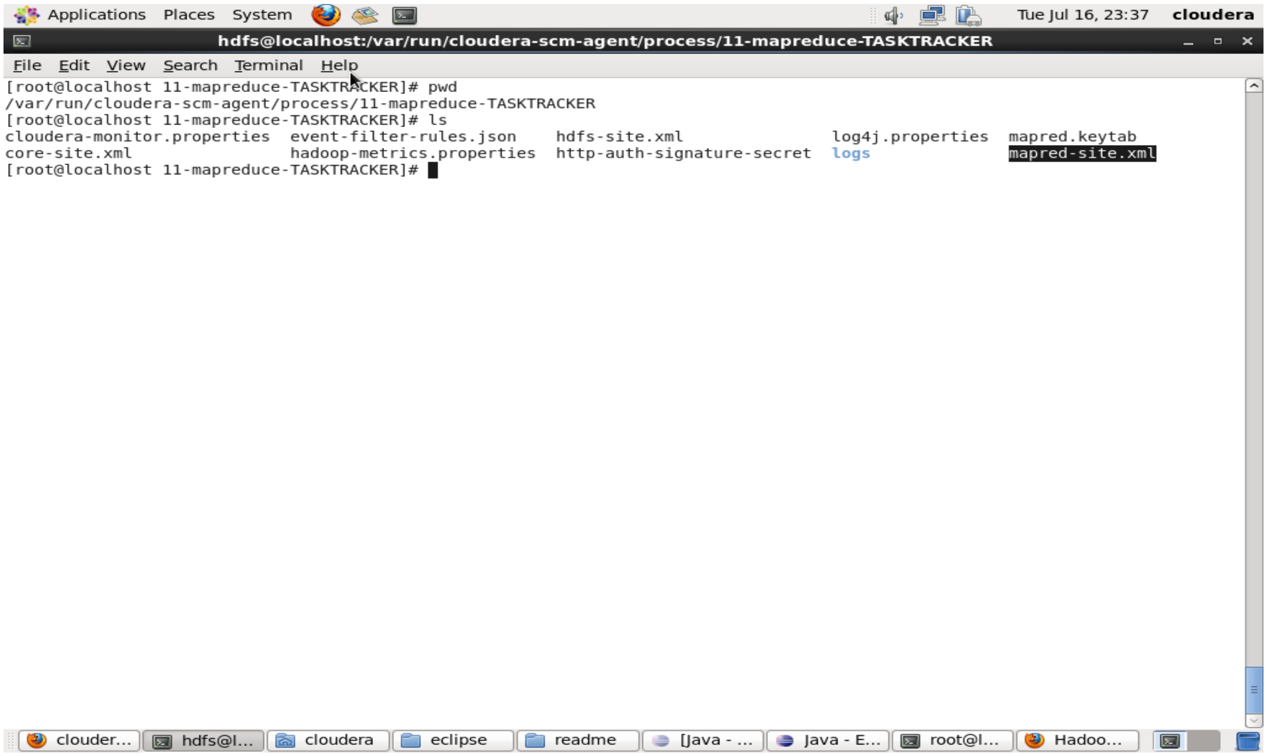


```
Mirror  22
Next    1
Not     1
Now     2
Prince  2
Queen   107
Queen,  36
Queen.  12
Seven   22
She     24
Since   11
Snow    176
Soon    11
The     117
Then    11
They    22
This    12
Time    1
When    11
White   124
White's 5
White,  33
White.  14
a       181
about   22
accept  2
after,  22
again   1
again,  1
again:  11
against 11
age     11
alive,  11
alive.  11
all     11
alone   11
also    11
always  22
an      22
```

6. Configuring hadoop in multi-tasking mode (Multi -Thread)

Step1 : Go to directory where task tracker’s map reduce configuration file is found.

$ cd /var/run/Cloudera-scm-agent/process/11-mapreduce-TASKTRACKER

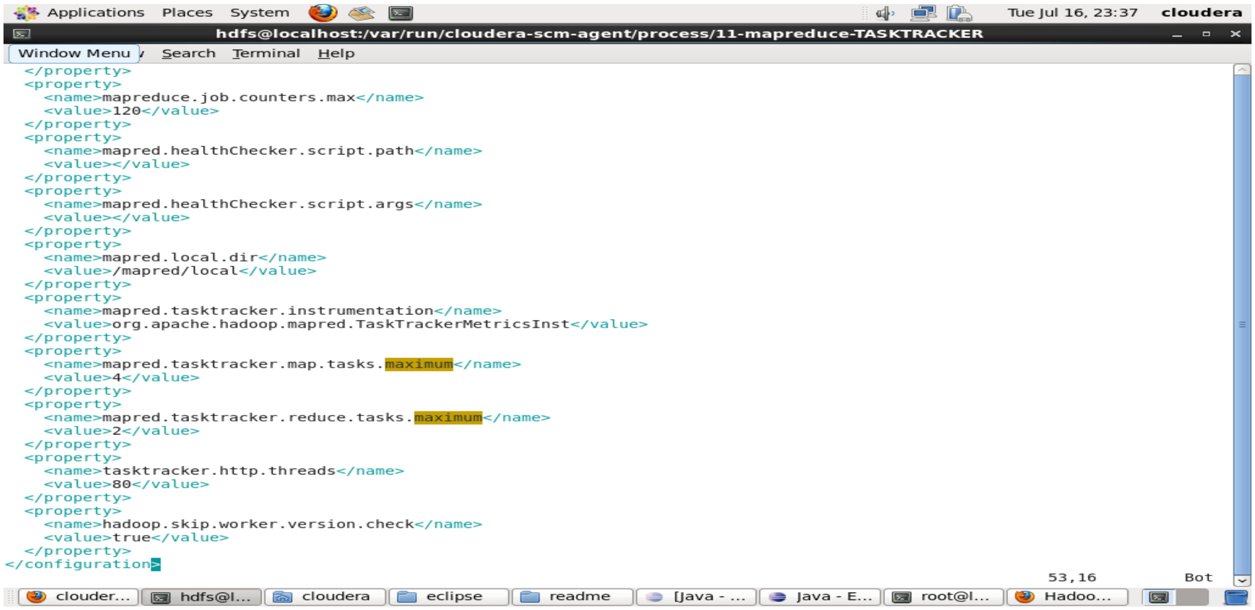


Applications Places System
Tue Jul 16, 23:37 cloudera
hdfs@localhost:/var/run/cloudera-scm-agent/process/11-mapreduce-TASKTRACKER
File Edit View Search Terminal Help
clouder… hdfs@l… cloudera eclipse readme [Java - … Java - E… root@l… Hadoo…

```
[root@localhost 11-mapreduce-TASKTRACKER]# pwd
/var/run/cloudera-scm-agent/process/11-mapreduce-TASKTRACKER
[root@localhost 11-mapreduce-TASKTRACKER]# ls
cloudera-monitor.properties  event-filter-rules.json    hdfs-site.xml              log4j.properties  mapred.keytab
core-site.xml                hadoop-metrics.properties  http-auth-signature-secret logs              mapred-site.xml
[root@localhost 11-mapreduce-TASKTRACKER]# 
```

Step 2: Open the file mapred-site.xml for editing.

Default file looks as follows:

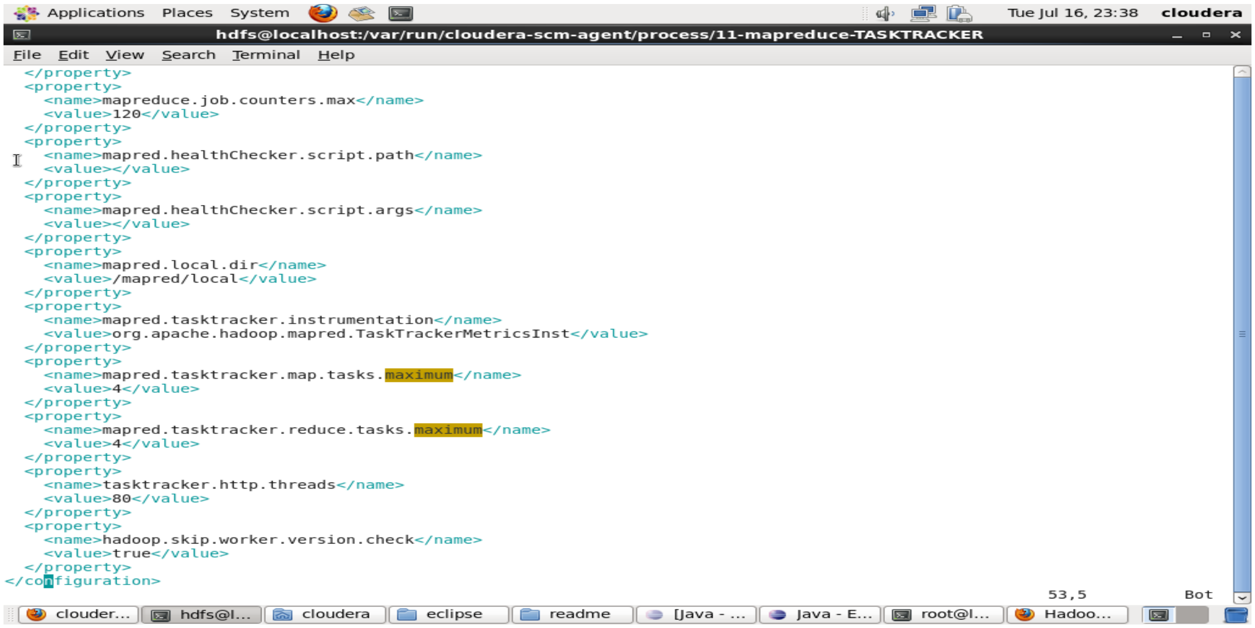


Applications Places System
Tue Jul 16, 23:37 cloudera
hdfs@localhost:/var/run/cloudera-scm-agent/process/11-mapreduce-TASKTRACKER
Window Menu Search Terminal Help
clouder… hdfs@l… cloudera eclipse readme [Java - … Java - E… root@l… Hadoo…

```
  </property>
  <property>
    <name>mapreduce.job.counters.max</name>
    <value>120</value>
  </property>
  <property>
    <name>mapred.healthChecker.script.path</name>
    <value></value>
  </property>
  <property>
    <name>mapred.healthChecker.script.args</name>
    <value></value>
  </property>
  <property>
    <name>mapred.local.dir</name>
    <value>/mapred/local</value>
  </property>
  <property>
    <name>mapred.tasktracker.instrumentation</name>
    <value>org.apache.hadoop.mapred.TaskTrackerMetricsInst</value>
  </property>
  <property>
    <name>mapred.tasktracker.map.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>mapred.tasktracker.reduce.tasks.maximum</name>
    <value>2</value>
  </property>
  <property>
    <name>tasktracker.http.threads</name>
    <value>80</value>
  </property>
  <property>
    <name>hadoop.skip.worker.version.check</name>
    <value>true</value>
  </property>
</configuration>
                                                                  53,16         Bot
```

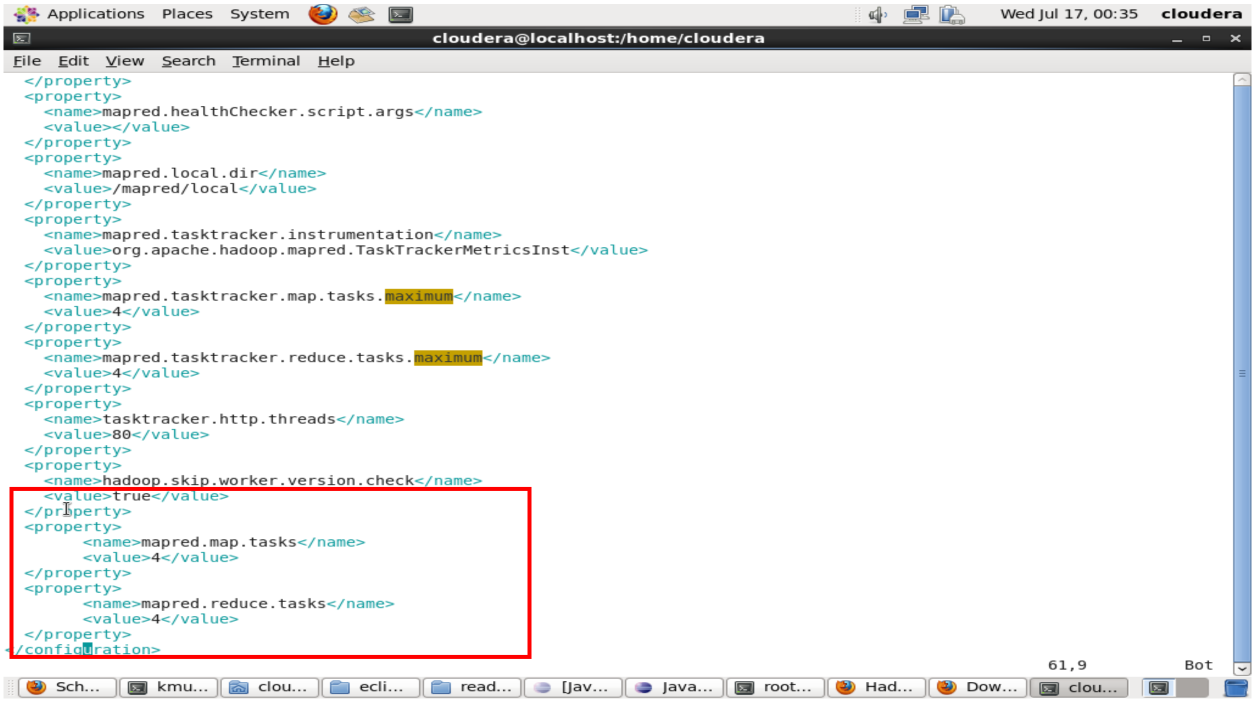
Step 3 : Change the Configuration.

Change mapreduce.tasktracker.map.tasks.maximum and mapreduce.tasktracker.reduce.tasks.maximum to 4. These parameters determine the maximum number of map/reduce tasks that will be run by task tracker in pseudo distributed mode.


Applications Places System
Tue Jul 16, 23:38 cloudera
hdfs@localhost:/var/run/cloudera-scm-agent/process/11-mapreduce-TASKTRACKER
File Edit View Search Terminal Help
53,5
Bot
clouder... hdfs@l... cloudera eclipse readme [Java - ... Java - E... root@l... Hadoo...

```
  </property>
  <property>
    <name>mapreduce.job.counters.max</name>
    <value>120</value>
  </property>
  <property>
    <name>mapred.healthChecker.script.path</name>
    <value></value>
  </property>
  <property>
    <name>mapred.healthChecker.script.args</name>
    <value></value>
  </property>
  <property>
    <name>mapred.local.dir</name>
    <value>/mapred/local</value>
  </property>
  <property>
    <name>mapred.tasktracker.instrumentation</name>
    <value>org.apache.hadoop.mapred.TaskTrackerMetricsInst</value>
  </property>
  <property>
    <name>mapred.tasktracker.map.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>mapred.tasktracker.reduce.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>tasktracker.http.threads</name>
    <value>80</value>
  </property>
  <property>
    <name>hadoop.skip.worker.version.check</name>
    <value>true</value>
  </property>
</configuration>
```

Step 4 : The above params are just a hint for the tracker. If you want to enforce the number then add new properties mapred.map.tasks and mapred.reduce.tasks and set them to the desired value.


Applications Places System
Wed Jul 17, 00:35 cloudera
cloudera@localhost:/home/cloudera
File Edit View Search Terminal Help
61,9
Bot
Sch... kmu... clou... ecli... read... [Jav... Java... root... Had... Dow... clou...

```
  </property>
  <property>
    <name>mapred.healthChecker.script.args</name>
    <value></value>
  </property>
  <property>
    <name>mapred.local.dir</name>
    <value>/mapred/local</value>
  </property>
  <property>
    <name>mapred.tasktracker.instrumentation</name>
    <value>org.apache.hadoop.mapred.TaskTrackerMetricsInst</value>
  </property>
  <property>
    <name>mapred.tasktracker.map.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>mapred.tasktracker.reduce.tasks.maximum</name>
    <value>4</value>
  </property>
  <property>
    <name>tasktracker.http.threads</name>
    <value>80</value>
  </property>
  <property>
    <name>hadoop.skip.worker.version.check</name>
    <value>true</value>
  </property>
  <property>
        <name>mapred.map.tasks</name>
        <value>4</value>
  </property>
  <property>
        <name>mapred.reduce.tasks</name>
        <value>4</value>
  </property>
</configuration>
```

Step 5 : Restart the hadoop daemons

```
$ bin/hadoop-daemon.sh stop cloudera-scm-server

$ bin/hadoop-daemon.sh start cloudera-scm-server

$ bin/hadoop-daemon.sh stop cloudera-scm-agent

$ bin/hadoop-daemon.sh start cloudera-scm-agent
```

Now the new configurations will load and when the huge file data is given, the number of task to map and reduce will be increased.

7. Configuring Flume (Monitoring Configuration)

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

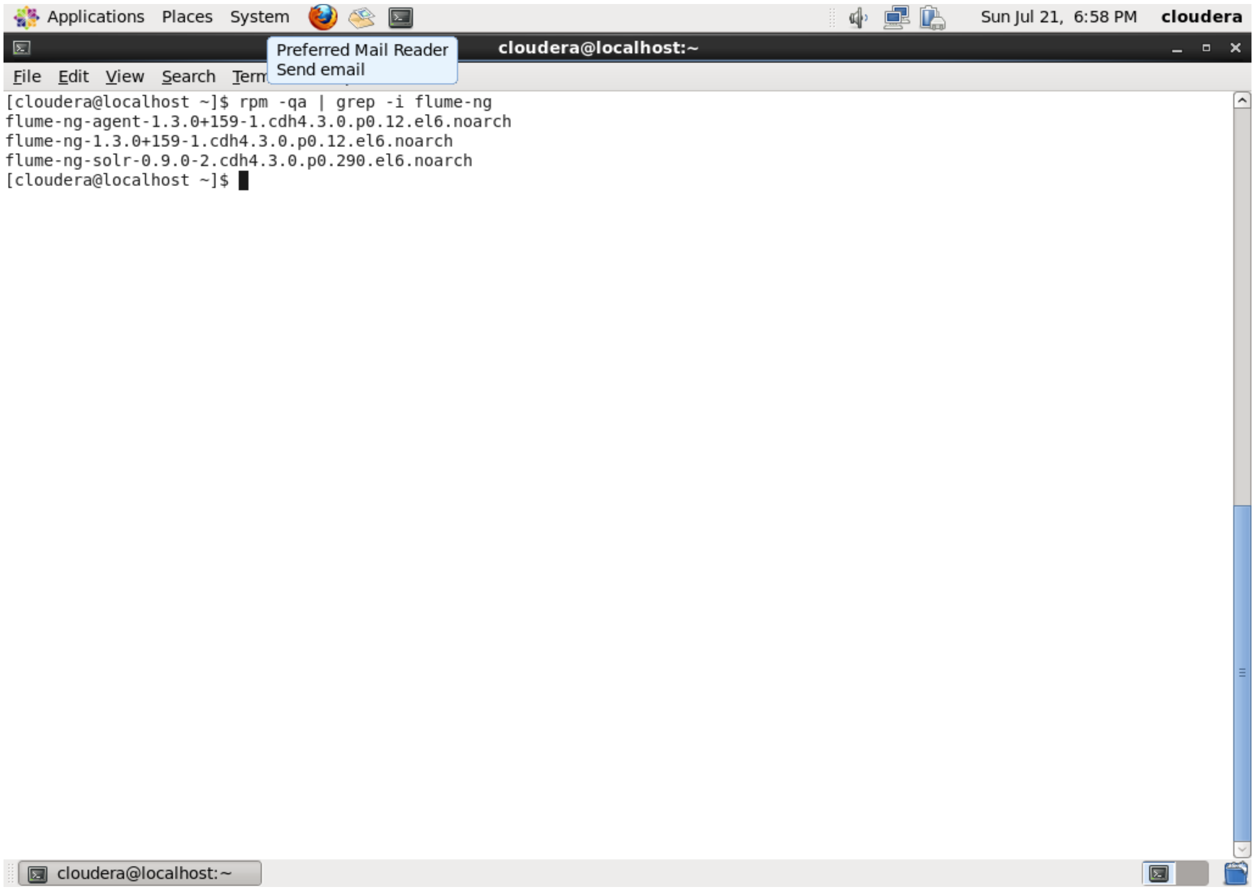
The use of Apache Flume is not only restricted to log data aggregation. Since data sources are customizable, Flume can be used to transport massive quantities of event data including but not limited to network traffic data, social-media-generated data, email messages and pretty much any data source possible.

Apache Flume is a top level project at the Apache Software Foundation.

Step 1 : Check Flume installation. Flume is installed as a part of quickstart VM.

$ rpm –qa | grep –i flume-ng

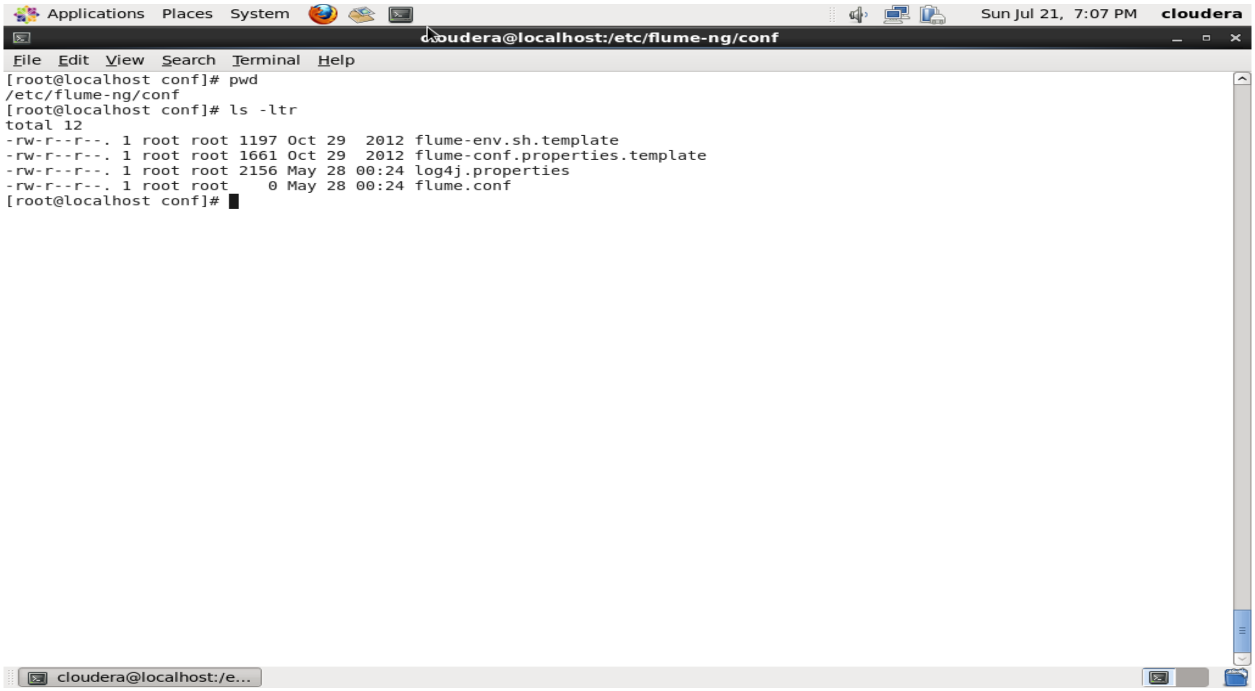
This displays the flume installation files



Applications Places System
Sun Jul 21, 6:58 PM cloudera
cloudera@localhost:~
Preferred Mail Reader
Send email
File Edit View Search Term
cloudera@localhost:~

```
[cloudera@localhost ~]$ rpm -qa | grep -i flume-ng
flume-ng-agent-1.3.0+159-1.cdh4.3.0.p0.12.el6.noarch
flume-ng-1.3.0+159-1.cdh4.3.0.p0.12.el6.noarch
flume-ng-solr-0.9.0-2.cdh4.3.0.p0.290.el6.noarch
[cloudera@localhost ~]$
```

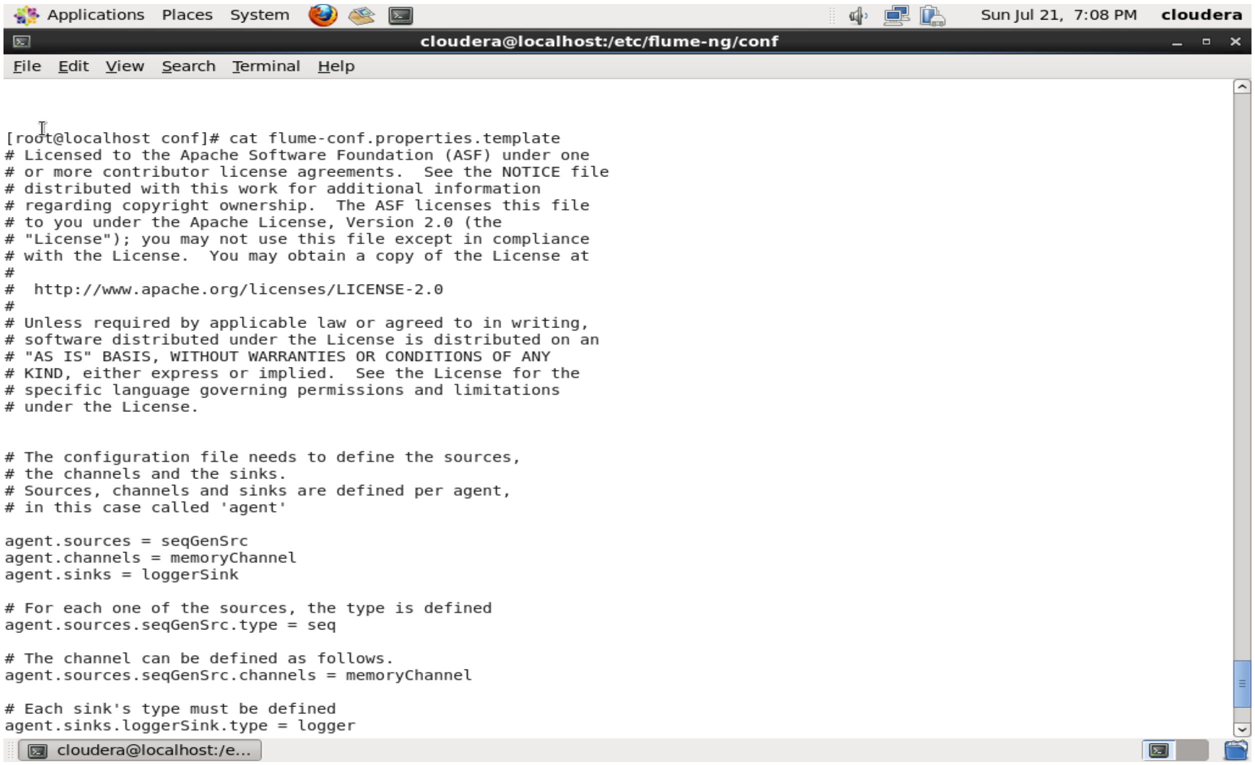
Step 3 :Check the flume template files under /etc/flume-ng/conf


Applications Places System
Sun Jul 21, 7:07 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help
cloudera@localhost:/e...

```
[root@localhost conf]# pwd
/etc/flume-ng/conf
[root@localhost conf]# ls -ltr
total 12
-rw-r--r--. 1 root root 1197 Oct 29  2012 flume-env.sh.template
-rw-r--r--. 1 root root 1661 Oct 29  2012 flume-conf.properties.template
-rw-r--r--. 1 root root 2156 May 28 00:24 log4j.properties
-rw-r--r--. 1 root root    0 May 28 00:24 flume.conf
[root@localhost conf]#
```

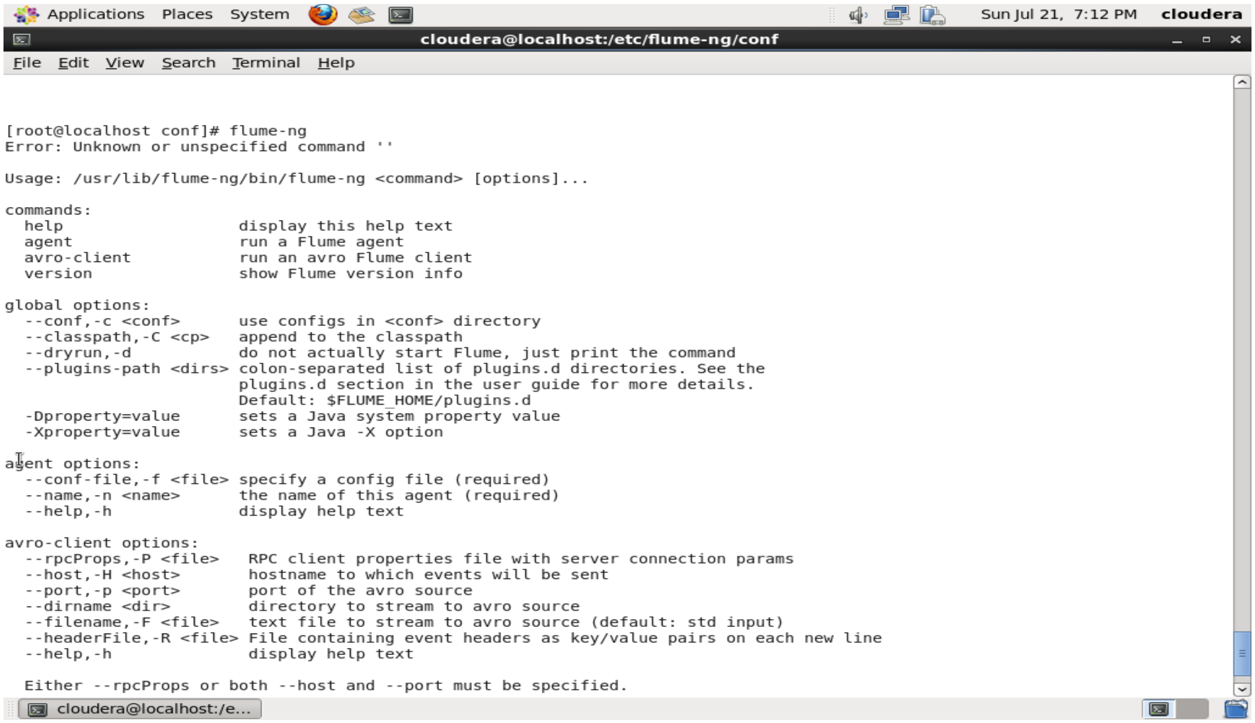
Step 4: Flume default configuration file.

$ vim flume.conf


Applications Places System
Sun Jul 21, 7:08 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help
cloudera@localhost:/e...

```
[root@localhost conf]# cat flume-conf.properties.template
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License.  You may obtain a copy of the License at
#
#  http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing,
# software distributed under the License is distributed on an
# "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY
# KIND, either express or implied.  See the License for the
# specific language governing permissions and limitations
# under the License.


# The configuration file needs to define the sources,
# the channels and the sinks.
# Sources, channels and sinks are defined per agent,
# in this case called 'agent'

agent.sources = seqGenSrc
agent.channels = memoryChannel
agent.sinks = loggerSink

# For each one of the sources, the type is defined
agent.sources.seqGenSrc.type = seq

# The channel can be defined as follows.
agent.sources.seqGenSrc.channels = memoryChannel

# Each sink's type must be defined
agent.sinks.loggerSink.type = logger
```

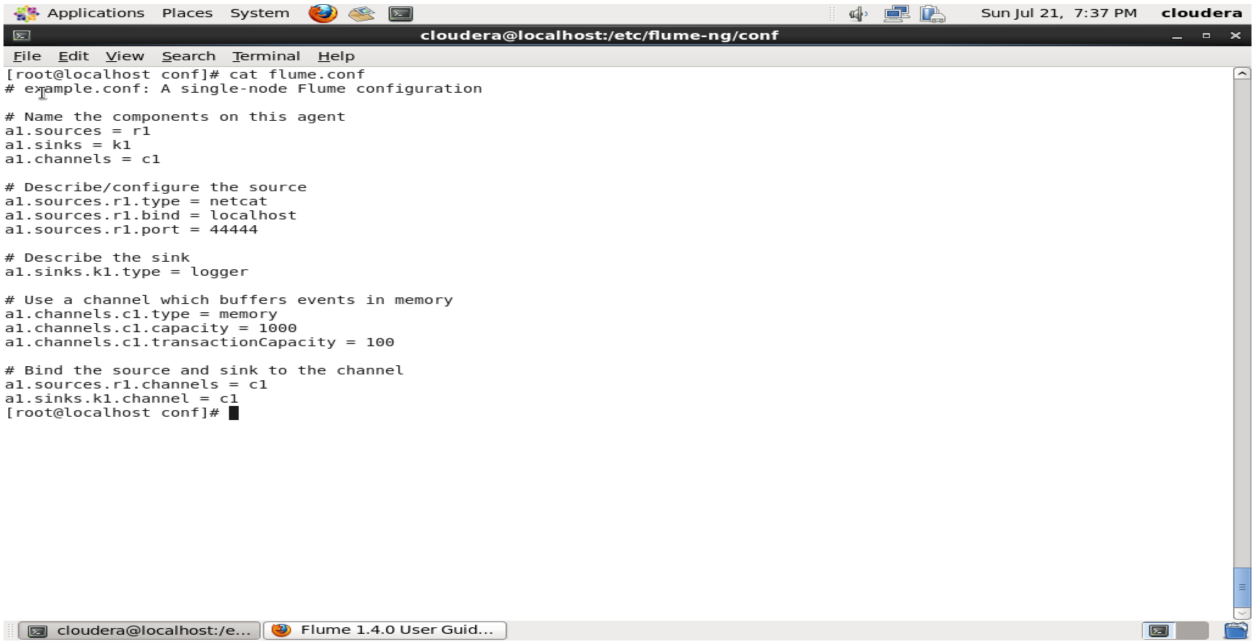
Step 5 :Flume command and its options. We will be using the agent command.


Applications Places System
Sun Jul 21, 7:12 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help
cloudera@localhost:/e...

```
[root@localhost conf]# flume-ng
Error: Unknown or unspecified command ''

Usage: /usr/lib/flume-ng/bin/flume-ng <command> [options]...

commands:
  help                  display this help text
  agent                 run a Flume agent
  avro-client           run an avro Flume client
  version               show Flume version info

global options:
  --conf,-c <conf>      use configs in <conf> directory
  --classpath,-C <cp>   append to the classpath
  --dryrun,-d           do not actually start Flume, just print the command
  --plugins-path <dirs> colon-separated list of plugins.d directories. See the
                        plugins.d section in the user guide for more details.
                        Default: $FLUME_HOME/plugins.d
  -Dproperty=value      sets a Java system property value
  -Xproperty=value      sets a Java -X option

agent options:
  --conf-file,-f <file> specify a config file (required)
  --name,-n <name>      the name of this agent (required)
  --help,-h             display help text

avro-client options:
  --rpcProps,-P <file>   RPC client properties file with server connection params
  --host,-H <host>       hostname to which events will be sent
  --port,-p <port>       port of the avro source
  --dirname <dir>        directory to stream to avro source
  --filename,-F <file>   text file to stream to avro source (default: std input)
  --headerFile,-R <file> File containing event headers as key/value pairs on each new line
  --help,-h              display help text

  Either --rpcProps or both --host and --port must be specified.
```

Step 6:Changing flume.conf file

This configuration defines a single agent named a1. a1 has a source that listens for data on port 44444, a channel that buffers event data in memory, and a sink that logs event data to the console. The configuration file names the various components, then describes their types and configuration parameters. A given configuration file might define several named agents; when a given Flume process is launched a flag is passed telling it which named agent to manifest.

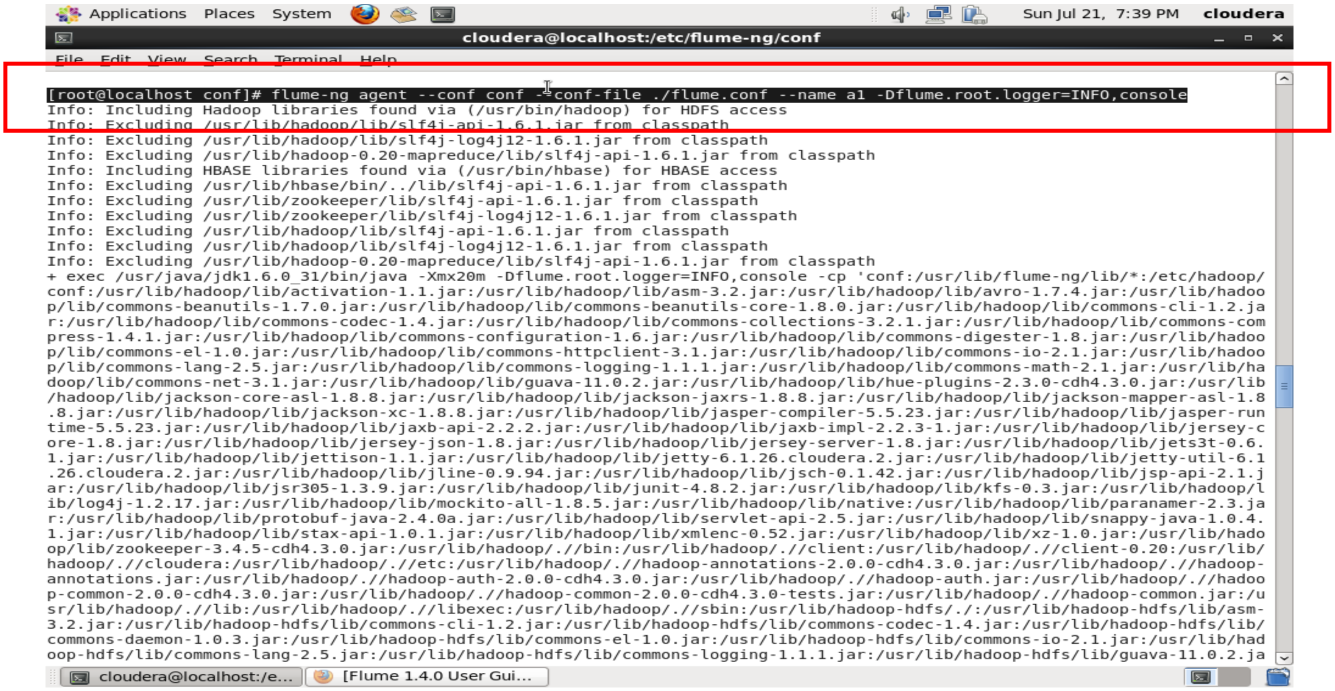

Applications Places System
Sun Jul 21, 7:37 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help
cloudera@localhost:/e...
Flume 1.4.0 User Guid...

```
[root@localhost conf]# cat flume.conf
# example.conf: A single-node Flume configuration

# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# Describe/configure the source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# Describe the sink
a1.sinks.k1.type = logger

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
[root@localhost conf]# 
```

Step 7:Starting of a flume agent

An agent is started using a shell script called flume-ng which is located in the bin directory of the Flume distribution.

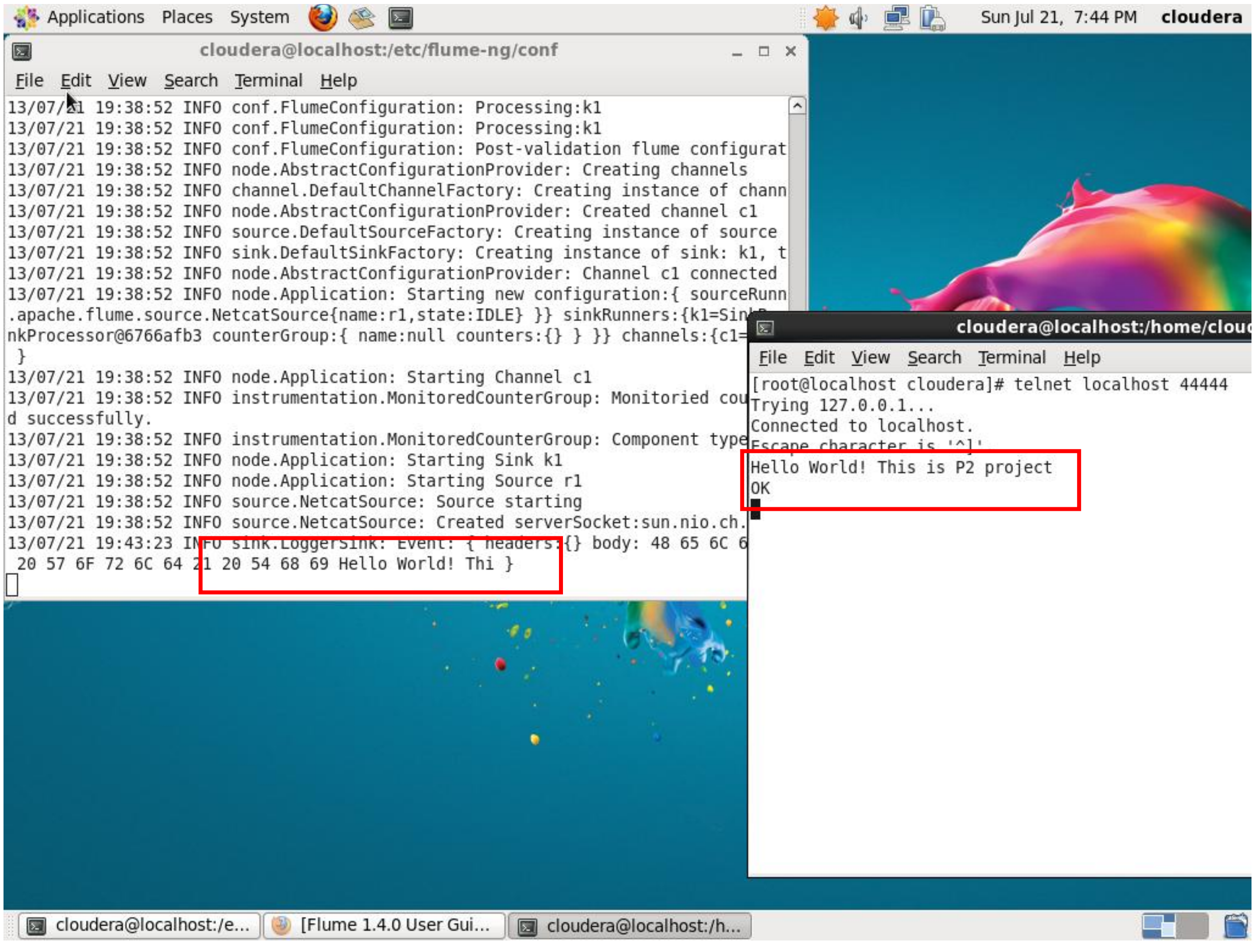
$ bin/flume-ng agent –conf conf –conf-file ./flume.conf –name a1 –Dflume.root.logger=INFO,console


Applications Places System
Sun Jul 21, 7:39 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help
cloudera@localhost:/e...
[Flume 1.4.0 User Gui...

```
[root@localhost conf]# flume-ng agent --conf conf --conf-file ./flume.conf --name a1 -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/usr/bin/hadoop) for HDFS access
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.6.1.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12-1.6.1.jar from classpath
Info: Excluding /usr/lib/hadoop-0.20-mapreduce/lib/slf4j-api-1.6.1.jar from classpath
Info: Including HBASE libraries found via (/usr/bin/hbase) for HBASE access
Info: Excluding /usr/lib/hbase/bin/../lib/slf4j-api-1.6.1.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-api-1.6.1.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-log4j12-1.6.1.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.6.1.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12-1.6.1.jar from classpath
Info: Excluding /usr/lib/hadoop-0.20-mapreduce/lib/slf4j-api-1.6.1.jar from classpath
+ exec /usr/java/jdk1.6.0_31/bin/java -Xmx20m -Dflume.root.logger=INFO,console -cp 'conf:/usr/lib/flume-ng/lib/*:/etc/hadoop/
conf:/usr/lib/hadoop/lib/activation-1.1.jar:/usr/lib/hadoop/lib/asm-3.2.jar:/usr/lib/hadoop/lib/avro-1.7.4.jar:/usr/lib/hadoo
p/lib/commons-beanutils-1.7.0.jar:/usr/lib/hadoop/lib/commons-beanutils-core-1.8.0.jar:/usr/lib/hadoop/lib/commons-cli-1.2.ja
r:/usr/lib/hadoop/lib/commons-codec-1.4.jar:/usr/lib/hadoop/lib/commons-collections-3.2.1.jar:/usr/lib/hadoop/lib/commons-com
press-1.4.1.jar:/usr/lib/hadoop/lib/commons-configuration-1.6.jar:/usr/lib/hadoop/lib/commons-digester-1.8.jar:/usr/lib/hadoo
p/lib/commons-el-1.0.jar:/usr/lib/hadoop/lib/commons-httpclient-3.1.jar:/usr/lib/hadoop/lib/commons-io-2.1.jar:/usr/lib/hadoo
p/lib/commons-lang-2.5.jar:/usr/lib/hadoop/lib/commons-logging-1.1.1.jar:/usr/lib/hadoop/lib/commons-math-2.1.jar:/usr/lib/ha
doop/lib/commons-net-3.1.jar:/usr/lib/hadoop/lib/guava-11.0.2.jar:/usr/lib/hadoop/lib/hue-plugins-2.3.0-cdh4.3.0.jar:/usr/lib
/hadoop/lib/jackson-core-asl-1.8.8.jar:/usr/lib/hadoop/lib/jackson-jaxrs-1.8.8.jar:/usr/lib/hadoop/lib/jackson-mapper-asl-1.8
.8.jar:/usr/lib/hadoop/lib/jackson-xc-1.8.8.jar:/usr/lib/hadoop/lib/jasper-compiler-5.5.23.jar:/usr/lib/hadoop/lib/jasper-run
time-5.5.23.jar:/usr/lib/hadoop/lib/jaxb-api-2.2.2.jar:/usr/lib/hadoop/lib/jaxb-impl-2.2.3-1.jar:/usr/lib/hadoop/lib/jersey-c
ore-1.8.jar:/usr/lib/hadoop/lib/jersey-json-1.8.jar:/usr/lib/hadoop/lib/jersey-server-1.8.jar:/usr/lib/hadoop/lib/jets3t-0.6.
1.jar:/usr/lib/hadoop/lib/jettison-1.1.jar:/usr/lib/hadoop/lib/jetty-6.1.26.cloudera.2.jar:/usr/lib/hadoop/lib/jetty-util-6.1
.26.cloudera.2.jar:/usr/lib/hadoop/lib/jline-0.9.94.jar:/usr/lib/hadoop/lib/jsch-0.1.42.jar:/usr/lib/hadoop/lib/jsp-api-2.1.j
ar:/usr/lib/hadoop/lib/jsr305-1.3.9.jar:/usr/lib/hadoop/lib/junit-4.8.2.jar:/usr/lib/hadoop/lib/kfs-0.3.jar:/usr/lib/hadoop/l
ib/log4j-1.2.17.jar:/usr/lib/hadoop/lib/mockito-all-1.8.5.jar:/usr/lib/hadoop/lib/native:/usr/lib/hadoop/lib/paranamer-2.3.ja
r:/usr/lib/hadoop/lib/protobuf-java-2.4.0a.jar:/usr/lib/hadoop/lib/servlet-api-2.5.jar:/usr/lib/hadoop/lib/snappy-java-1.0.4.
1.jar:/usr/lib/hadoop/lib/stax-api-1.0.1.jar:/usr/lib/hadoop/lib/xmlenc-0.52.jar:/usr/lib/hadoop/lib/xz-1.0.jar:/usr/lib/hado
op/lib/zookeeper-3.4.5-cdh4.3.0.jar:/usr/lib/hadoop/.//bin:/usr/lib/hadoop/.//client:/usr/lib/hadoop/.//client-0.20:/usr/lib/
hadoop/.//cloudera:/usr/lib/hadoop/.//etc:/usr/lib/hadoop/.//hadoop-annotations-2.0.0-cdh4.3.0.jar:/usr/lib/hadoop/.//hadoop-
annotations.jar:/usr/lib/hadoop/.//hadoop-auth-2.0.0-cdh4.3.0.jar:/usr/lib/hadoop/.//hadoop-auth.jar:/usr/lib/hadoop/.//hadoo
p-common-2.0.0-cdh4.3.0.jar:/usr/lib/hadoop/.//hadoop-common-2.0.0-cdh4.3.0-tests.jar:/usr/lib/hadoop/.//hadoop-common.jar:/u
sr/lib/hadoop/.//lib:/usr/lib/hadoop/.//libexec:/usr/lib/hadoop/.//sbin:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/asm-
3.2.jar:/usr/lib/hadoop-hdfs/lib/commons-cli-1.2.jar:/usr/lib/hadoop-hdfs/lib/commons-codec-1.4.jar:/usr/lib/hadoop-hdfs/lib/
commons-daemon-1.0.3.jar:/usr/lib/hadoop-hdfs/lib/commons-el-1.0.jar:/usr/lib/hadoop-hdfs/lib/commons-io-2.1.jar:/usr/lib/had
oop-hdfs/lib/commons-lang-2.5.jar:/usr/lib/hadoop-hdfs/lib/commons-logging-1.1.1.jar:/usr/lib/hadoop-hdfs/lib/guava-11.0.2.ja
```

Step 8 : Testing the new Configuration

Open the new terminal and then we can then telnet port 44444 and send Flume an event. The original Flume terminal will output the event in a log message.



Applications Places System
Sun Jul 21, 7:44 PM cloudera
cloudera@localhost:/etc/flume-ng/conf
File Edit View Search Terminal Help

```
13/07/21 19:38:52 INFO conf.FlumeConfiguration: Processing:k1
13/07/21 19:38:52 INFO conf.FlumeConfiguration: Processing:k1
13/07/21 19:38:52 INFO conf.FlumeConfiguration: Post-validation flume configurat
13/07/21 19:38:52 INFO node.AbstractConfigurationProvider: Creating channels
13/07/21 19:38:52 INFO channel.DefaultChannelFactory: Creating instance of chann
13/07/21 19:38:52 INFO node.AbstractConfigurationProvider: Created channel c1
13/07/21 19:38:52 INFO source.DefaultSourceFactory: Creating instance of source
13/07/21 19:38:52 INFO sink.DefaultSinkFactory: Creating instance of sink: k1, t
13/07/21 19:38:52 INFO node.AbstractConfigurationProvider: Channel c1 connected
13/07/21 19:38:52 INFO node.Application: Starting new configuration:{ sourceRunn
.apache.flume.source.NetcatSource{name:r1,state:IDLE} }} sinkRunners:{k1=Sin
nkProcessor@6766afb3 counterGroup:{ name:null counters:{} } }} channels:{c1=
 }
13/07/21 19:38:52 INFO node.Application: Starting Channel c1
13/07/21 19:38:52 INFO instrumentation.MonitoredCounterGroup: Monitoried cou
d successfully.
13/07/21 19:38:52 INFO instrumentation.MonitoredCounterGroup: Component type
13/07/21 19:38:52 INFO node.Application: Starting Sink k1
13/07/21 19:38:52 INFO node.Application: Starting Source r1
13/07/21 19:38:52 INFO source.NetcatSource: Source starting
13/07/21 19:38:52 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.
13/07/21 19:43:23 INFO sink.LoggerSink: Event: { headers:{} body: 48 65 6C 6
 20 57 6F 72 6C 64 21 20 54 68 69 Hello World! Thi }
```

cloudera@localhost:/home/clou
File Edit View Search Terminal Help

```
[root@localhost cloudera]# telnet localhost 44444
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
Hello World! This is P2 project
OK
```

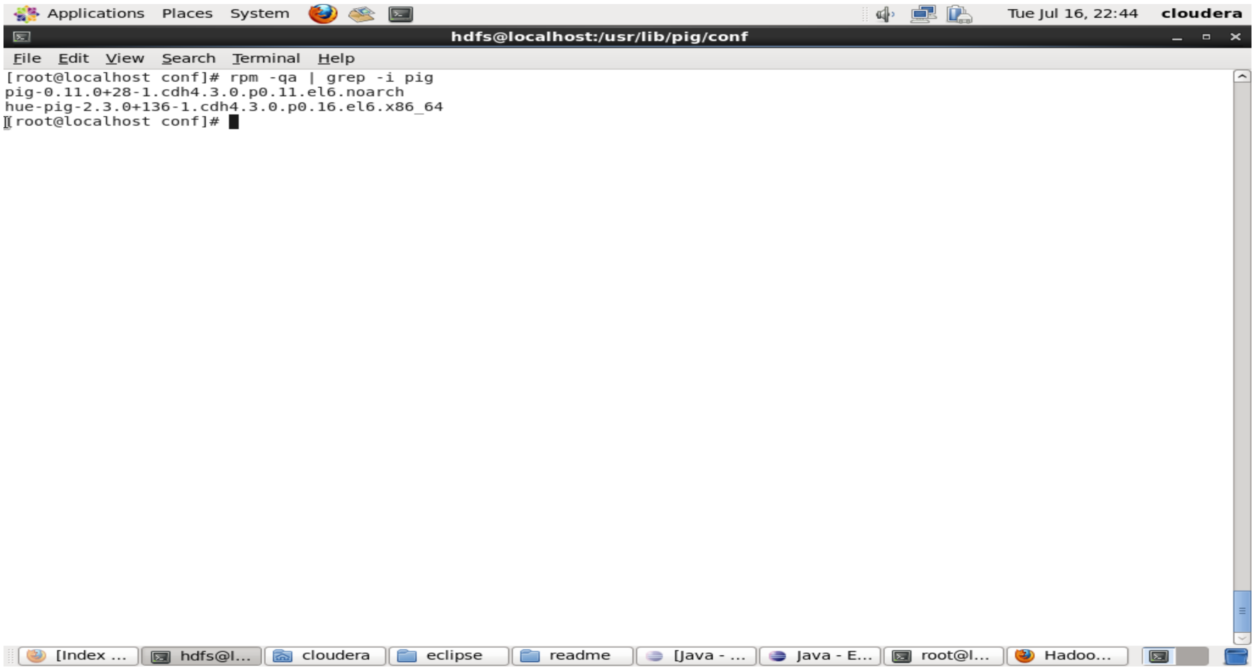
8. Running Pig project

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist . Pig's language layer currently consists of a textual language called Pig Latin which can be used to write queries.

Step 1: Checking the pig installation. Pig is installed as a part of quickstart vm

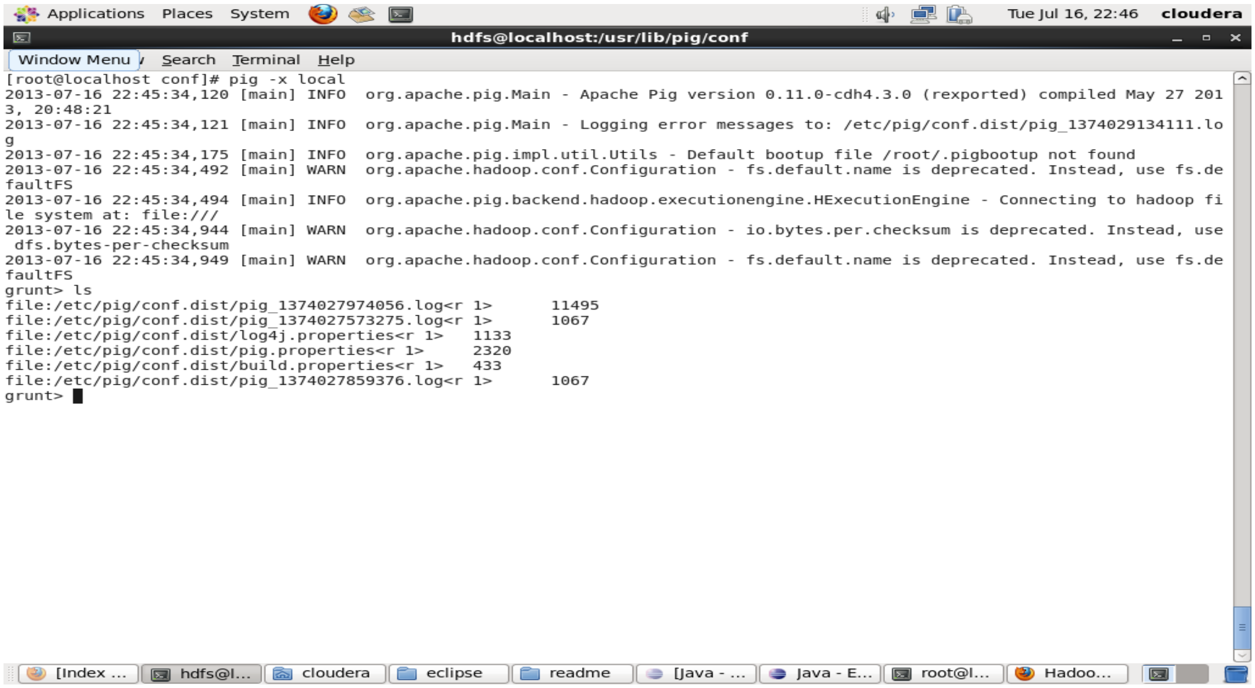
$rpm –qa |grep –i pig



```
[root@localhost conf]# rpm -qa | grep -i pig
pig-0.11.0+28-1.cdh4.3.0.p0.11.el6.noarch
hue-pig-2.3.0+136-1.cdh4.3.0.p0.16.el6.x86_64
[root@localhost conf]# 
```

Step 2: Open Pig Shell which display the grunt prompt. This is used to run user queries in high level query language called Pig Latin

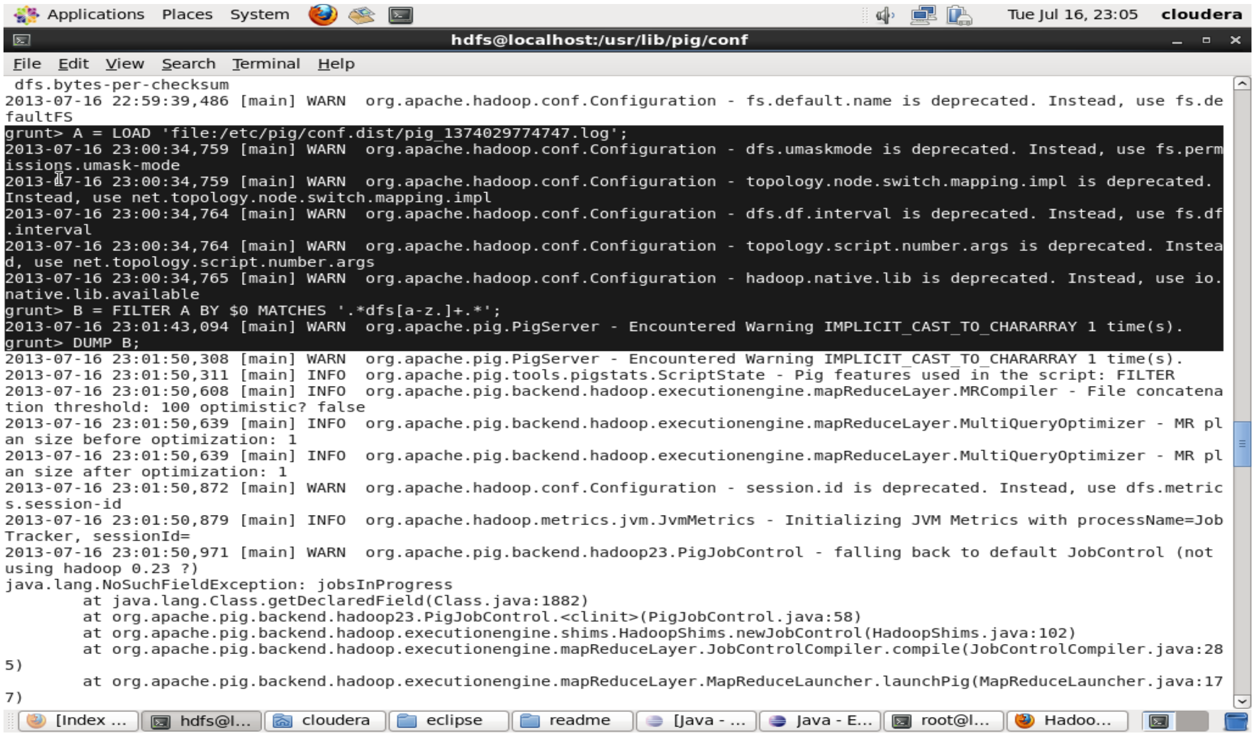
$ pig –x local



Applications Places System
Tue Jul 16, 22:46 cloudera
hdfs@localhost:/usr/lib/pig/conf
Window Menu Search Terminal Help
[Index ... hdfs@l... cloudera eclipse readme [Java - ... Java - E... root@l... Hadoo...

```
[root@localhost conf]# pig -x local
2013-07-16 22:45:34,120 [main] INFO  org.apache.pig.Main - Apache Pig version 0.11.0-cdh4.3.0 (rexported) compiled May 27 201
3, 20:48:21
2013-07-16 22:45:34,121 [main] INFO  org.apache.pig.Main - Logging error messages to: /etc/pig/conf.dist/pig_1374029134111.lo
g
2013-07-16 22:45:34,175 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2013-07-16 22:45:34,492 [main] WARN  org.apache.hadoop.conf.Configuration - fs.default.name is deprecated. Instead, use fs.de
faultFS
2013-07-16 22:45:34,494 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2013-07-16 22:45:34,944 [main] WARN  org.apache.hadoop.conf.Configuration - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2013-07-16 22:45:34,949 [main] WARN  org.apache.hadoop.conf.Configuration - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> ls
file:/etc/pig/conf.dist/pig_1374027974056.log<r 1>      11495
file:/etc/pig/conf.dist/pig_1374027573275.log<r 1>      1067
file:/etc/pig/conf.dist/log4j.properties<r 1>   1133
file:/etc/pig/conf.dist/pig.properties<r 1>     2320
file:/etc/pig/conf.dist/build.properties<r 1>   433
file:/etc/pig/conf.dist/pig_1374027859376.log<r 1>      1067
grunt> 
```

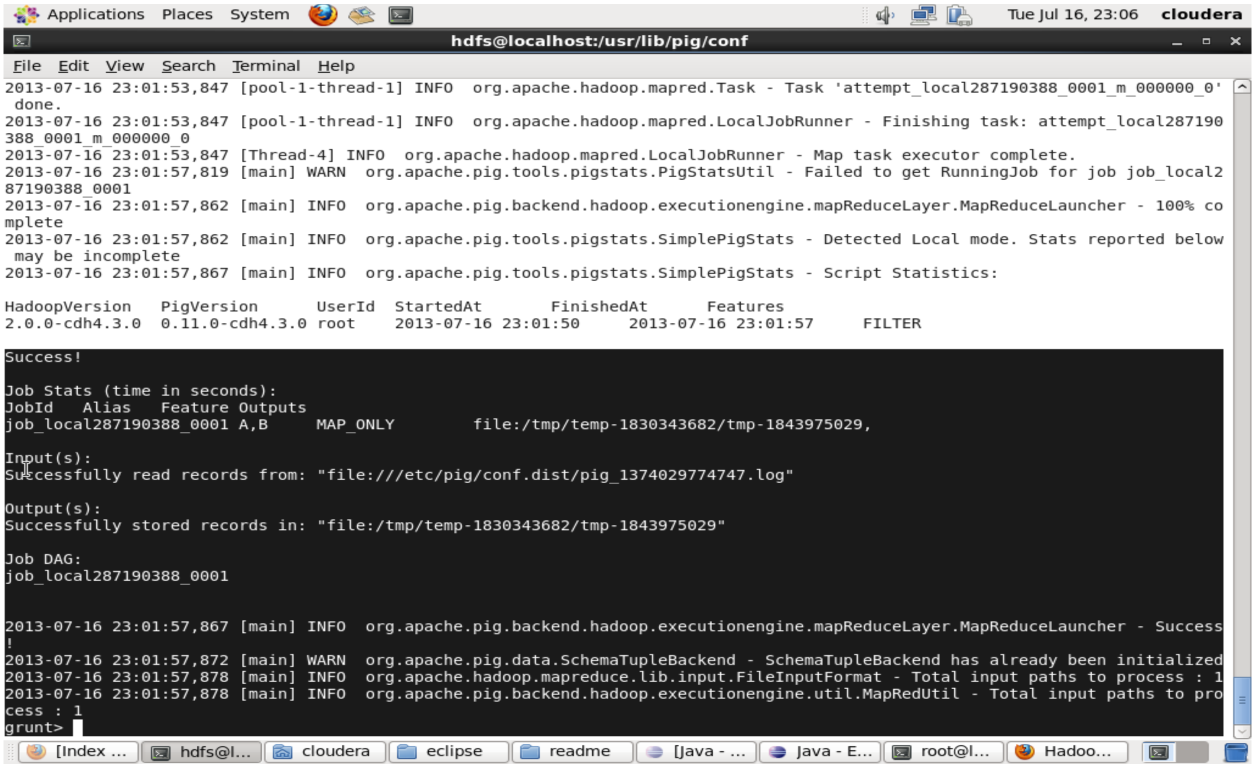
Step 3:Running the sample pig latin query: This greps for a particular pattern in a input file.



Applications Places System
Tue Jul 16, 23:05 cloudera
hdfs@localhost:/usr/lib/pig/conf
File Edit View Search Terminal Help
[Index ... hdfs@l... cloudera eclipse readme [Java - ... Java - E... root@l... Hadoo...

```
 dfs.bytes-per-checksum
2013-07-16 22:59:39,486 [main] WARN  org.apache.hadoop.conf.Configuration - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> A = LOAD 'file:/etc/pig/conf.dist/pig_1374029774747.log';
2013-07-16 23:00:34,759 [main] WARN  org.apache.hadoop.conf.Configuration - dfs.umaskmode is deprecated. Instead, use fs.perm
issions.umask-mode
2013-07-16 23:00:34,759 [main] WARN  org.apache.hadoop.conf.Configuration - topology.node.switch.mapping.impl is deprecated.
Instead, use net.topology.node.switch.mapping.impl
2013-07-16 23:00:34,764 [main] WARN  org.apache.hadoop.conf.Configuration - dfs.df.interval is deprecated. Instead, use fs.df
.interval
2013-07-16 23:00:34,764 [main] WARN  org.apache.hadoop.conf.Configuration - topology.script.number.args is deprecated. Instea
d, use net.topology.script.number.args
2013-07-16 23:00:34,765 [main] WARN  org.apache.hadoop.conf.Configuration - hadoop.native.lib is deprecated. Instead, use io.
native.lib.available
grunt> B = FILTER A BY $0 MATCHES '.*dfs[a-z.]+.*';
2013-07-16 23:01:43,094 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
grunt> DUMP B;
2013-07-16 23:01:50,308 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2013-07-16 23:01:50,311 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2013-07-16 23:01:50,608 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2013-07-16 23:01:50,639 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2013-07-16 23:01:50,639 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2013-07-16 23:01:50,872 [main] WARN  org.apache.hadoop.conf.Configuration - session.id is deprecated. Instead, use dfs.metric
s.session-id
2013-07-16 23:01:50,879 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Initializing JVM Metrics with processName=Job
Tracker, sessionId=
2013-07-16 23:01:50,971 [main] WARN  org.apache.pig.backend.hadoop23.PigJobControl - falling back to default JobControl (not
using hadoop 0.23 ?)
java.lang.NoSuchFieldException: jobsInProgress
        at java.lang.Class.getDeclaredField(Class.java:1882)
        at org.apache.pig.backend.hadoop23.PigJobControl.<clinit>(PigJobControl.java:58)
        at org.apache.pig.backend.hadoop.executionengine.shims.HadoopShims.newJobControl(HadoopShims.java:102)
        at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler.compile(JobControlCompiler.java:28
5)
        at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher.launchPig(MapReduceLauncher.java:17
7)
```

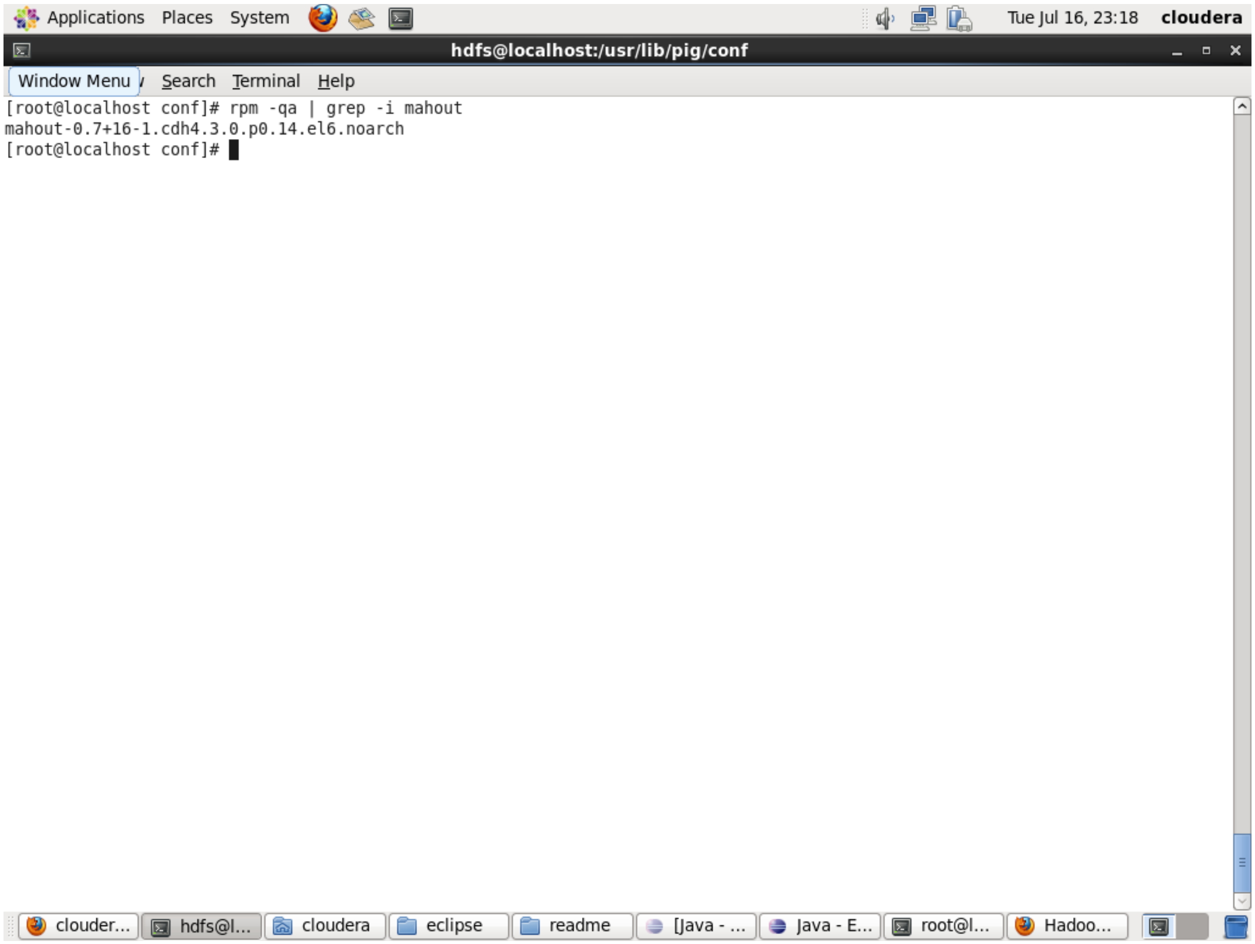
Step 4 : Displays the success message on successful completion of query.



```
2013-07-16 23:01:53,847 [pool-1-thread-1] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local287190388_0001_m_000000_0'
 done.
2013-07-16 23:01:53,847 [pool-1-thread-1] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local287190
388_0001_m_000000_0
2013-07-16 23:01:53,847 [Thread-4] INFO  org.apache.hadoop.mapred.LocalJobRunner - Map task executor complete.
2013-07-16 23:01:57,819 [main] WARN  org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local2
87190388_0001
2013-07-16 23:01:57,862 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% co
mplete
2013-07-16 23:01:57,862 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below
 may be incomplete
2013-07-16 23:01:57,867 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.0.0-cdh4.3.0  0.11.0-cdh4.3.0 root    2013-07-16 23:01:50     2013-07-16 23:01:57     FILTER

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local287190388_0001 A,B     MAP_ONLY        file:/tmp/temp-1830343682/tmp-1843975029,

Input(s):
Successfully read records from: "file:///etc/pig/conf.dist/pig_1374029774747.log"

Output(s):
Successfully stored records in: "file:/tmp/temp-1830343682/tmp-1843975029"

Job DAG:
job_local287190388_0001


2013-07-16 23:01:57,867 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2013-07-16 23:01:57,872 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2013-07-16 23:01:57,878 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2013-07-16 23:01:57,878 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
grunt> 
```

9. MAHOUT installation

The goal of Mahout is to build a vibrant, responsive, diverse community to facilitate discussions not only on the project itself but also on potential use cases. Mahout comes installed with quickstart vm. There is no additional configuration required.



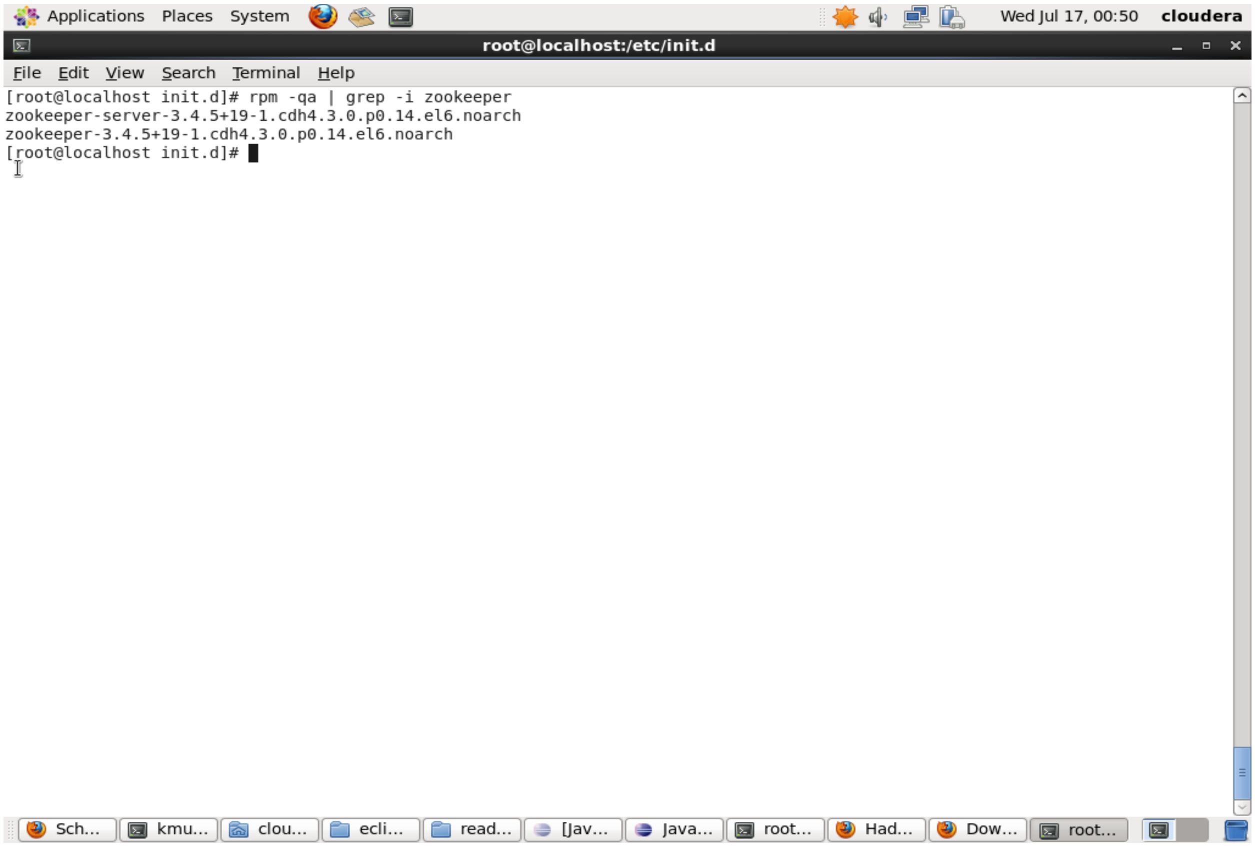
Applications Places System
Tue Jul 16, 23:18 cloudera
hdfs@localhost:/usr/lib/pig/conf
Window Menu Search Terminal Help
clouder... hdfs@l... cloudera eclipse readme [Java - ... Java - E... root@l... Hadoo...

```
[root@localhost conf]# rpm -qa | grep -i mahout
mahout-0.7+16-1.cdh4.3.0.p0.14.el6.noarch
[root@localhost conf]#
```

10. Zookeeper

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. ZooKeeper allows distributed processes to coordinate with each other through a shared hierarchal namespace which is organized similarly to a standard file system. All objects in this namespace are represented by znode which has data and attributes associated with it. We can perform all basic file operations like create, delete, modify etc on this node.

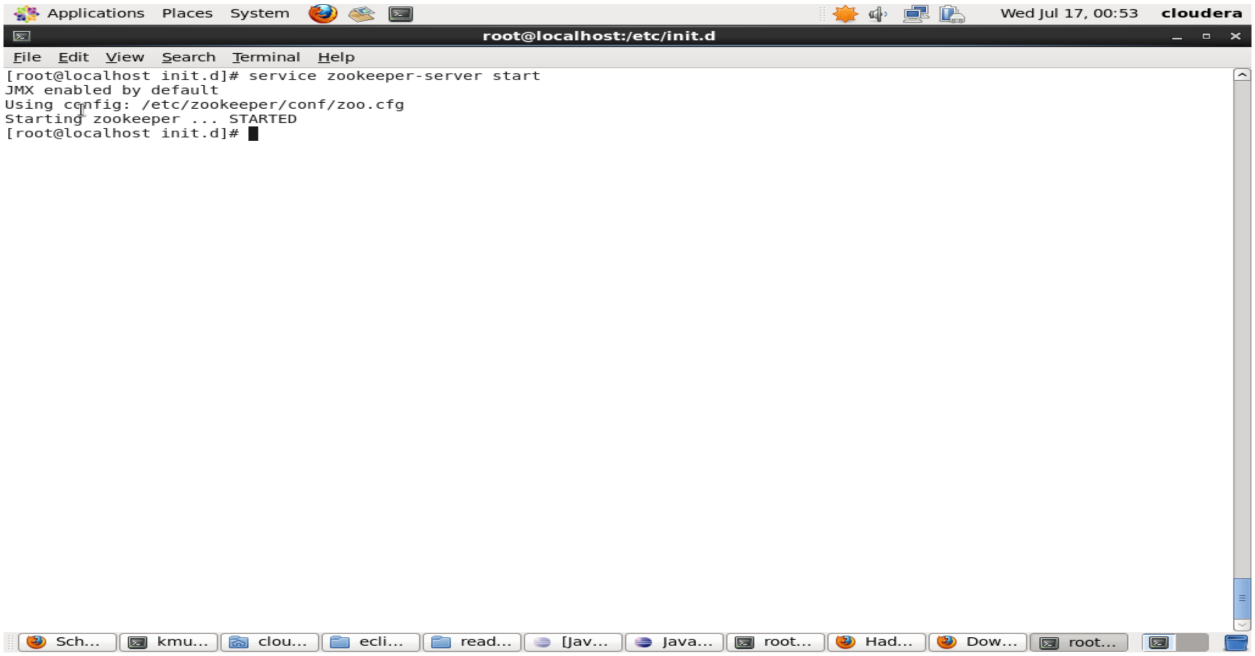
Step 1: Check for zookeeper installation. Zookeeper is installed as part of Cloudera quickstart vm



```
[root@localhost init.d]# rpm -qa | grep -i zookeeper
zookeeper-server-3.4.5+19-1.cdh4.3.0.p0.14.el6.noarch
zookeeper-3.4.5+19-1.cdh4.3.0.p0.14.el6.noarch
[root@localhost init.d]#
```

Step 2 : Start zookeeper

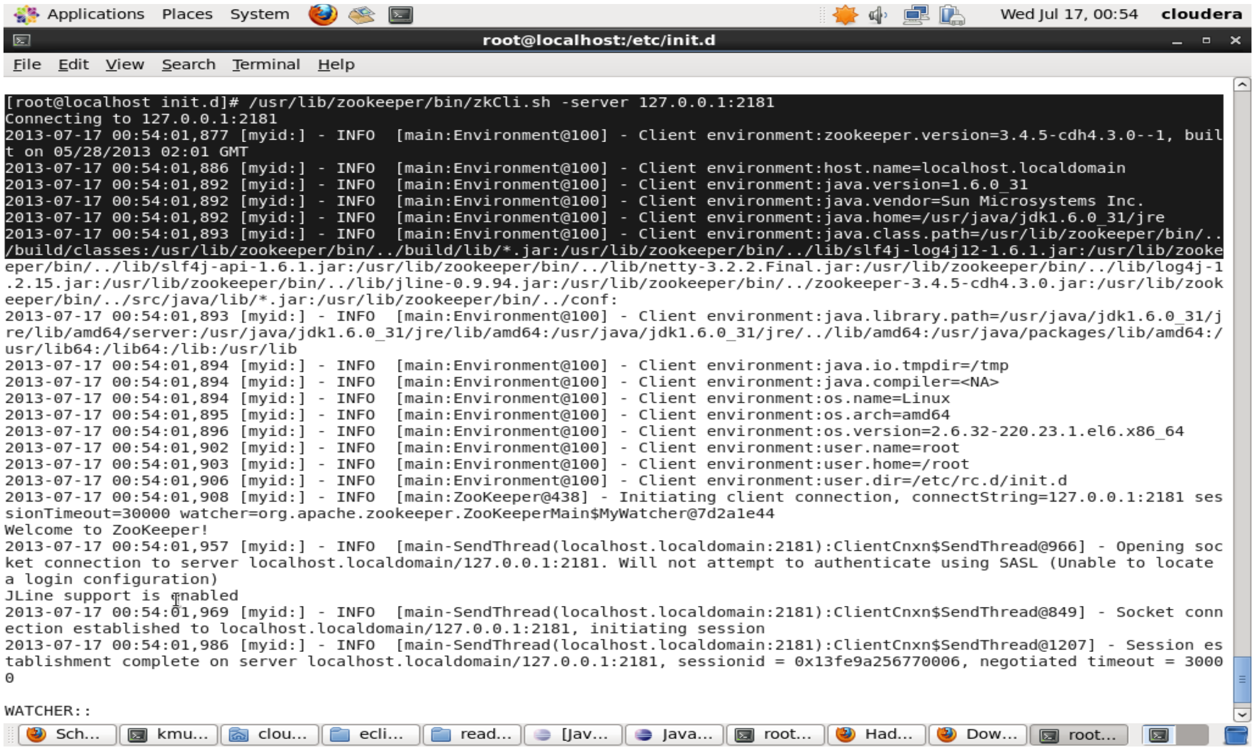
$service zookeeper.server start



Applications Places System
Wed Jul 17, 00:53 cloudera
root@localhost:/etc/init.d
File Edit View Search Terminal Help
Sch... kmu... clou... ecli... read... [Jav... Java... root... Had... Dow... root...

```
[root@localhost init.d]# service zookeeper-server start
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... STARTED
[root@localhost init.d]# 
```

Step 3: Creating a z-node and setting some data in that node



Applications Places System
Wed Jul 17, 00:54 cloudera
root@localhost:/etc/init.d
File Edit View Search Terminal Help
Sch... kmu... clou... ecli... read... [Jav... Java... root... Had... Dow... root...

```
[root@localhost init.d]# /usr/lib/zookeeper/bin/zkCli.sh -server 127.0.0.1:2181
Connecting to 127.0.0.1:2181
2013-07-17 00:54:01,877 [myid:] - INFO  [main:Environment@100] - Client environment:zookeeper.version=3.4.5-cdh4.3.0--1, buil
t on 05/28/2013 02:01 GMT
2013-07-17 00:54:01,886 [myid:] - INFO  [main:Environment@100] - Client environment:host.name=localhost.localdomain
2013-07-17 00:54:01,892 [myid:] - INFO  [main:Environment@100] - Client environment:java.version=1.6.0_31
2013-07-17 00:54:01,892 [myid:] - INFO  [main:Environment@100] - Client environment:java.vendor=Sun Microsystems Inc.
2013-07-17 00:54:01,892 [myid:] - INFO  [main:Environment@100] - Client environment:java.home=/usr/java/jdk1.6.0_31/jre
2013-07-17 00:54:01,893 [myid:] - INFO  [main:Environment@100] - Client environment:java.class.path=/usr/lib/zookeeper/bin/..
/build/classes:/usr/lib/zookeeper/bin/../build/lib/*.jar:/usr/lib/zookeeper/bin/../lib/slf4j-log4j12-1.6.1.jar:/usr/lib/zooke
eper/bin/../lib/slf4j-api-1.6.1.jar:/usr/lib/zookeeper/bin/../lib/netty-3.2.2.Final.jar:/usr/lib/zookeeper/bin/../lib/log4j-1
.2.15.jar:/usr/lib/zookeeper/bin/../lib/jline-0.9.94.jar:/usr/lib/zookeeper/bin/../zookeeper-3.4.5-cdh4.3.0.jar:/usr/lib/zook
eeper/bin/../src/java/lib/*.jar:/usr/lib/zookeeper/bin/../conf:
2013-07-17 00:54:01,893 [myid:] - INFO  [main:Environment@100] - Client environment:java.library.path=/usr/java/jdk1.6.0_31/j
re/lib/amd64/server:/usr/java/jdk1.6.0_31/jre/lib/amd64:/usr/java/jdk1.6.0_31/jre/../lib/amd64:/usr/java/packages/lib/amd64:/
usr/lib64:/lib64:/lib:/usr/lib
2013-07-17 00:54:01,894 [myid:] - INFO  [main:Environment@100] - Client environment:java.io.tmpdir=/tmp
2013-07-17 00:54:01,894 [myid:] - INFO  [main:Environment@100] - Client environment:java.compiler=<NA>
2013-07-17 00:54:01,894 [myid:] - INFO  [main:Environment@100] - Client environment:os.name=Linux
2013-07-17 00:54:01,895 [myid:] - INFO  [main:Environment@100] - Client environment:os.arch=amd64
2013-07-17 00:54:01,896 [myid:] - INFO  [main:Environment@100] - Client environment:os.version=2.6.32-220.23.1.el6.x86_64
2013-07-17 00:54:01,902 [myid:] - INFO  [main:Environment@100] - Client environment:user.name=root
2013-07-17 00:54:01,903 [myid:] - INFO  [main:Environment@100] - Client environment:user.home=/root
2013-07-17 00:54:01,906 [myid:] - INFO  [main:Environment@100] - Client environment:user.dir=/etc/rc.d/init.d
2013-07-17 00:54:01,908 [myid:] - INFO  [main:ZooKeeper@438] - Initiating client connection, connectString=127.0.0.1:2181 ses
sionTimeout=30000 watcher=org.apache.zookeeper.ZooKeeperMain$MyWatcher@7d2a1e44
Welcome to ZooKeeper!
2013-07-17 00:54:01,957 [myid:] - INFO  [main-SendThread(localhost.localdomain:2181):ClientCnxn$SendThread@966] - Opening soc
ket connection to server localhost.localdomain/127.0.0.1:2181. Will not attempt to authenticate using SASL (Unable to locate
a login configuration)
JLine support is enabled
2013-07-17 00:54:01,969 [myid:] - INFO  [main-SendThread(localhost.localdomain:2181):ClientCnxn$SendThread@849] - Socket conn
ection established to localhost.localdomain/127.0.0.1:2181, initiating session
2013-07-17 00:54:01,986 [myid:] - INFO  [main-SendThread(localhost.localdomain:2181):ClientCnxn$SendThread@1207] - Session es
tablishment complete on server localhost.localdomain/127.0.0.1:2181, sessionid = 0x13fe9a256770006, negotiated timeout = 3000
0

WATCHER::
```

References:

www.cloudera.com

http://hadoop.apache.org/

http://developer.yahoo.com/hadoop/tutorial/