CDH installation & Application Test Report

He Shouchun (SCUID: 00001008350, Email: she@scu.edu)

Read Cloudera installation guide: http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM4Ent/latest/Cloudera-Manager-Installation-Guide/cmig_intro_to_cm_install.html

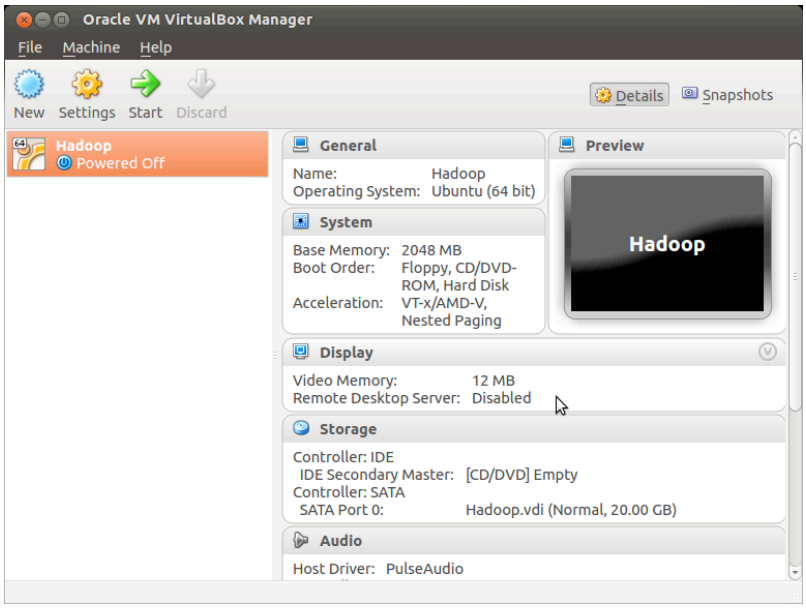
Chapter 1. Prepare the virtual machine

1.1 Download virtual machine software

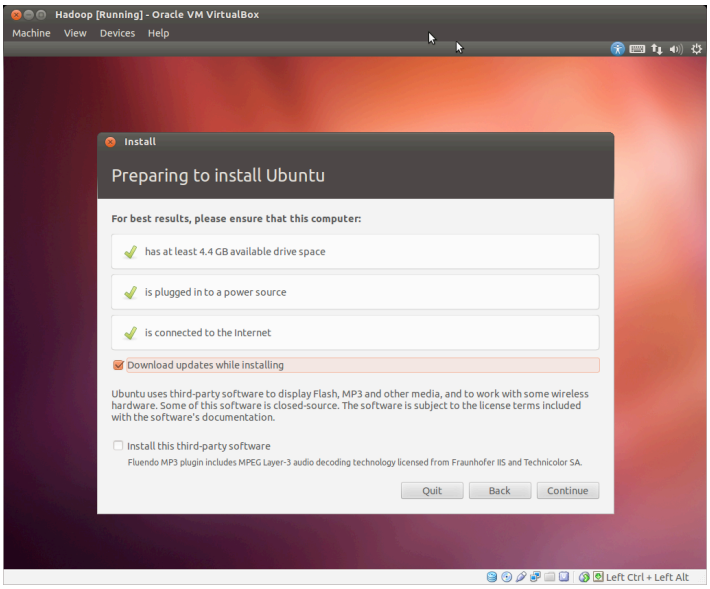
Download VirtualBox from oracle website: http://www.virtualbox.org/, and install it on host machine. Then create a new virtual machine:

1.2 Plan the guest operation system

Selected the Ubuntu 12.04 (64bit). According to the Cloudera installation guide, plan 20GB virtual HD space and 2GB RAM for the guest OS. Just follow the Linux installation steps to install it in the VM.



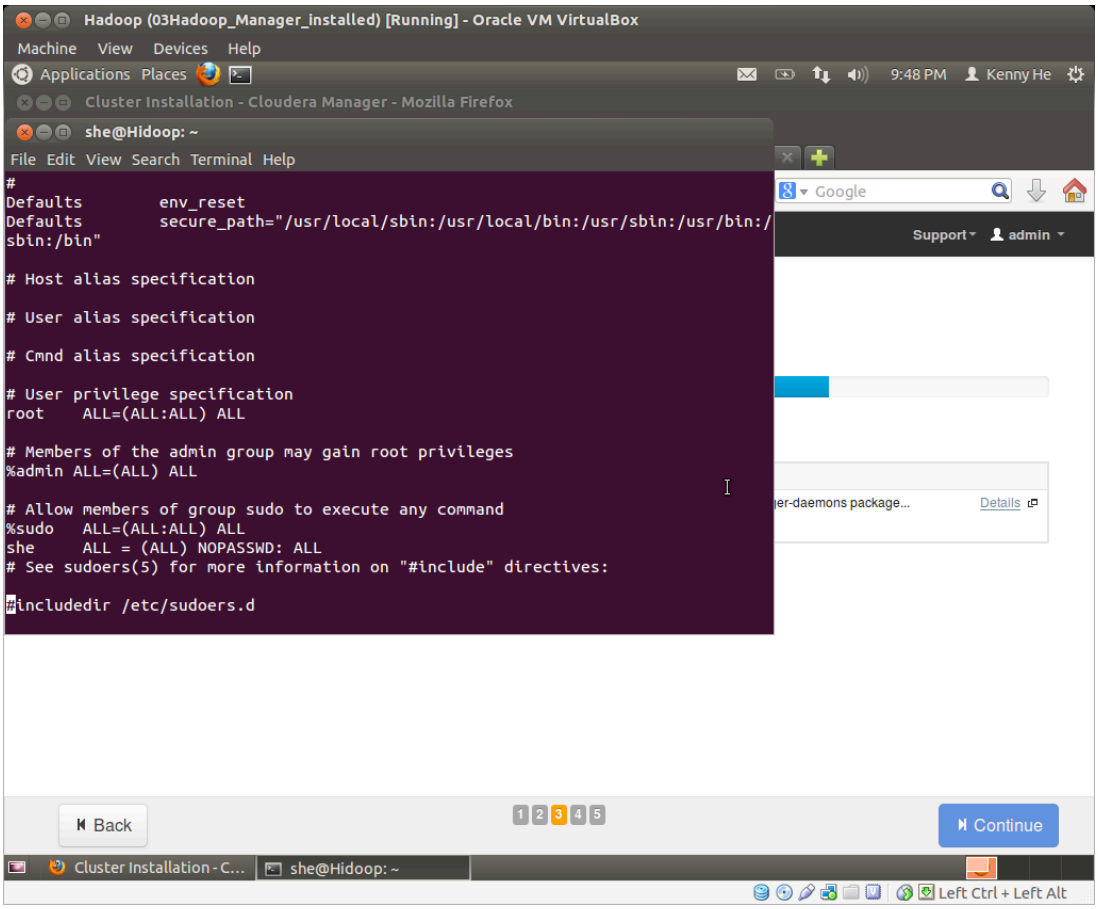
Ubuntu 12.04 installation:



1.3 Post-install configurations

Change the root password: sudo -i, input password and then switch to root mode, and passwd root to change the password of root.

Create a new account “she”, and make it a password-less sudo user by edit /etc/sudoers:



In the installation of CDH the system needs a account which has sudo privileges without input a password. So this step is very important.

Run software update manager to install all the patches: the English languages , the security patches, bug fixes, etc. After finished, save a snapshot called “Ubuntu-Install”.

1.4 Network setting

The default virtual network adapter for VirtualBox guest OS is NAT, which allows the guest OS users to access outside network but it does not allow the host OS users and other outside users to access the guest OS. To enable the host OS users to use/test the service on the guest OS, we need to add another network adapter whose type is “Bridged Adapter” and make this virtual network adapter on guest OS to be bridged to the Ethernet adapter “eth0”. However there is a limitation that only when the network cable is connected can both the guest virtual network adapter and the host Ethernet adapter work well. If the cable is pulled out, both adapters will be “ifdown” and can not connect to each other.

Setup static IP address for the Ethernet adapters of both the host and the guest OS:

```
Host: 192.168.1.100/24
Guest: 192.168.1.101/24
```

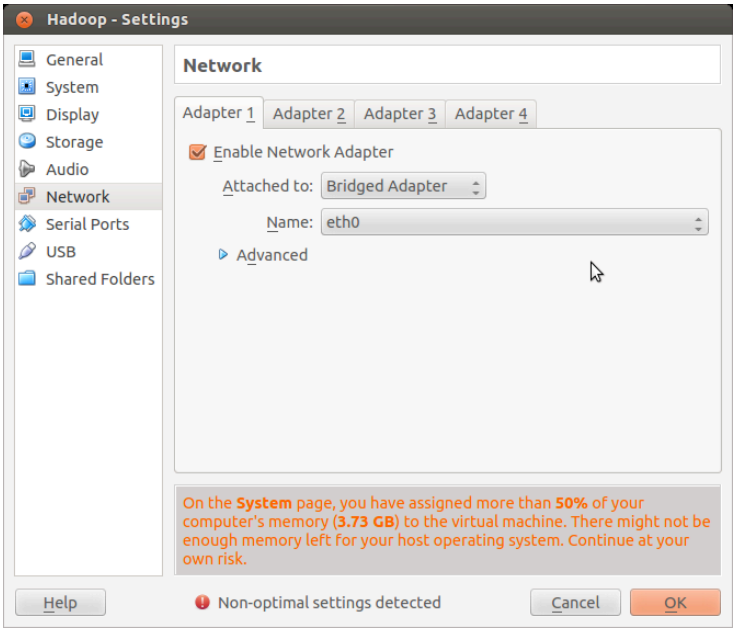
And must set a host name for the guest OS. Or we will meet problems during the installation.

I also tried to bridge the guest virtual network adapter to the wireless adapter. The VirtualBox official website shows this solution can work but on my laptop it does not work at all (the possible reason is that my host OS is Ubuntu and there are some limitations). I know it works on VMWare but I prefer using free VirtualBox VM.

So there are two adapters for the guest OS:

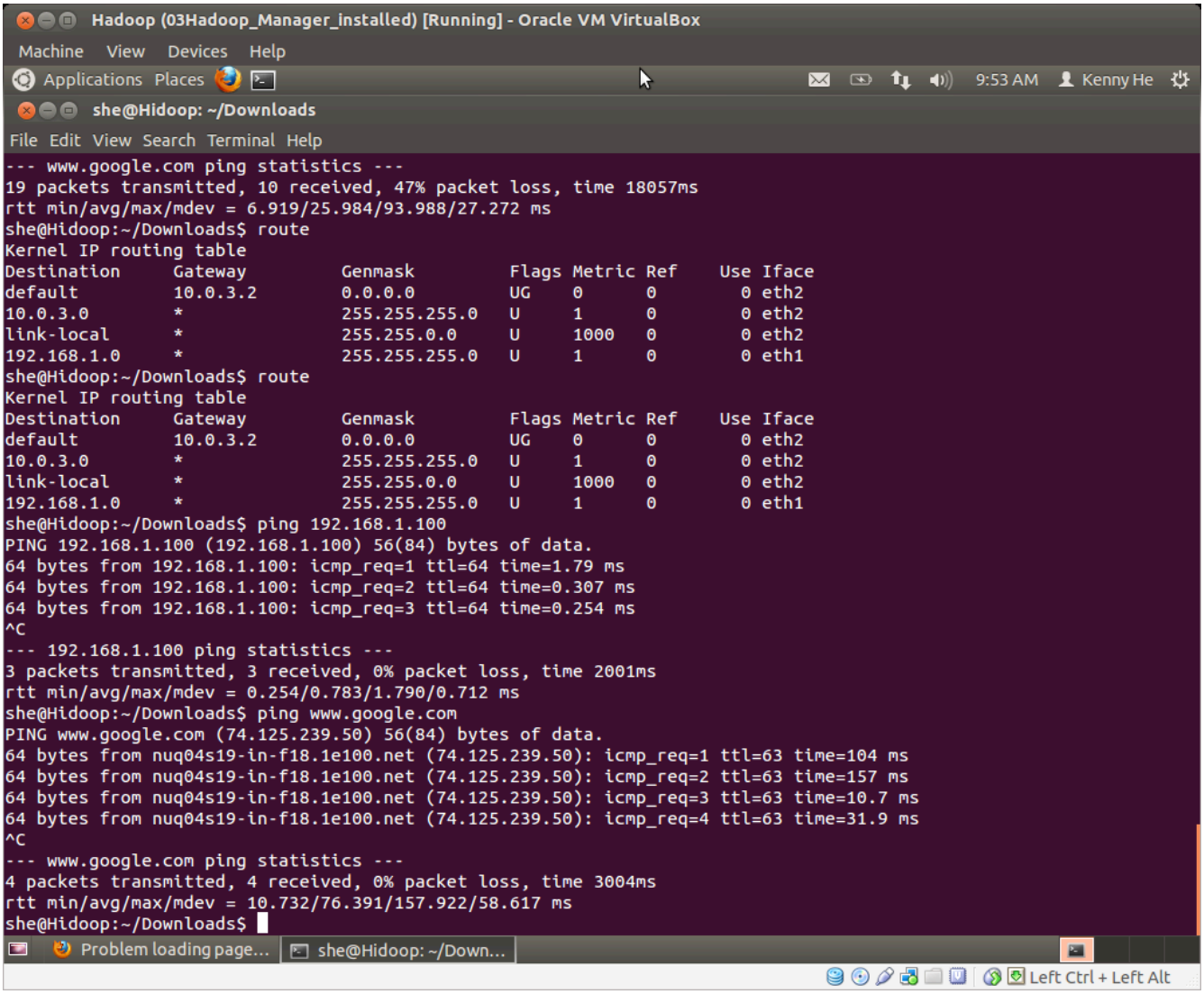
Adapter 1: “Bridged Adapter”, bridge to host Ethernet adapter eth0 for connecting to the host when Ethernet cable is connected;

Adapter 2: NAT, for Internet access via any Internet connection (e.g., through wlan0) on the host.



And then configure the guest OS route table to make sure that the route to Internet will go through Adapter 2 only and the Adapter 1 interface is for routing to local network 192.168.1.0/24 only.

If the configuration is correct, we can ping between the guest OS and host OS. And we can also open an Internet website (e.g. www.google.com, www.cloudera.com) from the Firefox web browser on the guest OS.



1.5 Install the sshd service

To maintain the server running in the guest, it is necessary to install SSH service:

sudo apt-get install openssh-server

Then change the firewall setting on the guest OS. To make life easier, I simply disabled the firewall service: sudo ufw disable. The firewall setting may cause not only the SSH but many other services are not accessible from the host OS and the outside.

After the installation finished, try to connect to the guest OS with putty by the IP address of the bridged virtual Ethernet adapter.

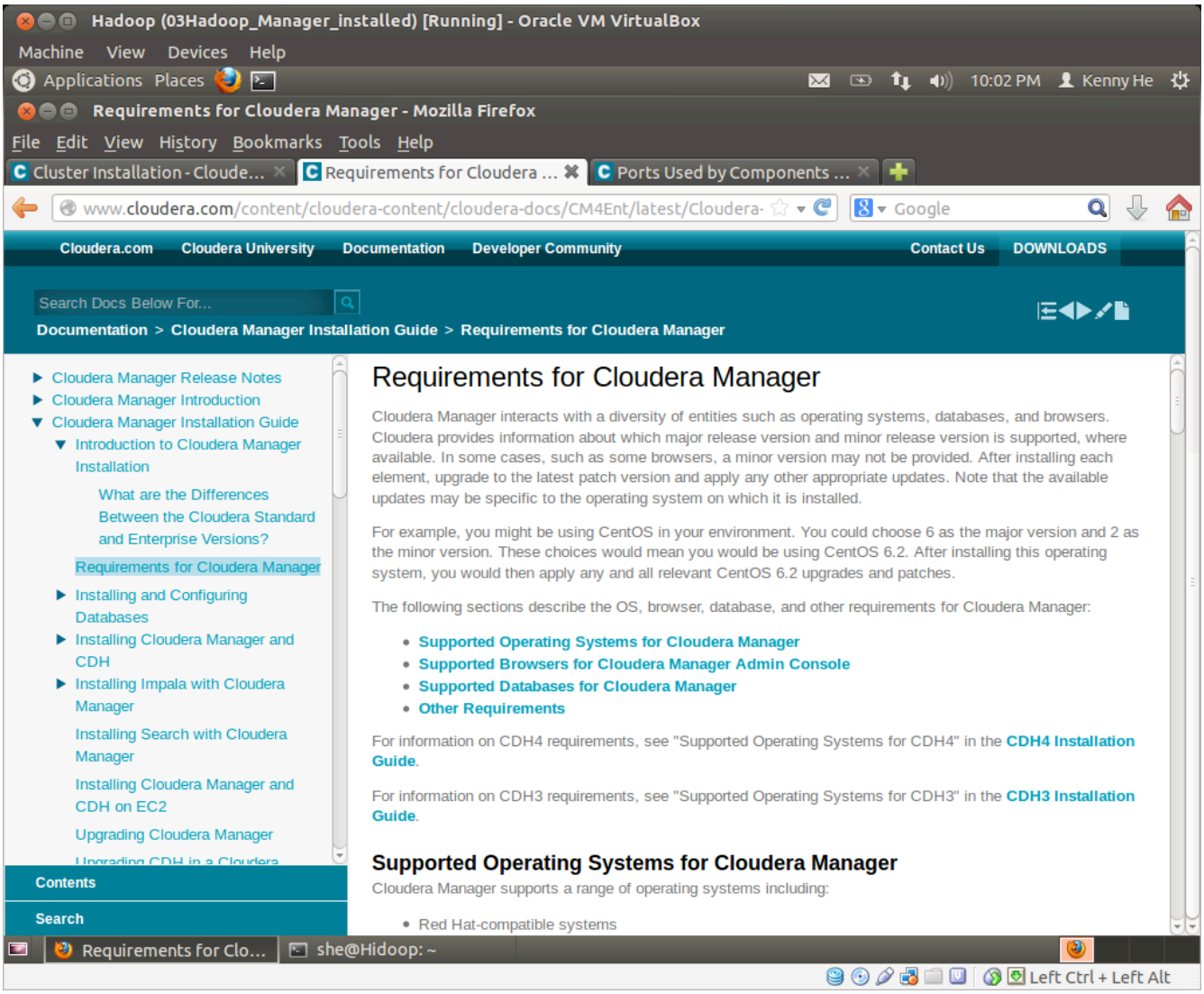
1.6 Database installation choice

Following the instructions on http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM4Ent/latest/Cloudera-Manager-Installation-Guide/cmig_installing_configuring_dbs.html. To make life easier, and get an opportunity to try and learn PostgreSQL DB (I am quite familiar with ORACLE, Sybase and MS SQL Server and MySQL), I prefer "Path A" automatically installs embedded PostgreSQL database.

Chapter 2. Install Cloudera Manager and CDH

2.1 Preparation

Before start, I read the requirements of the installation carefully. All the details need to be read thoroughly and strictly followed. Or, the installation may fail.



To make life easier, I chose Path A in the installation guide at the first installation. Follow the steps in this page:

http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM4Ent/latest/Cloudera-Manager-Installation-Guide/cmig_install_path_A.html?scroll=cmig_topic_6_5

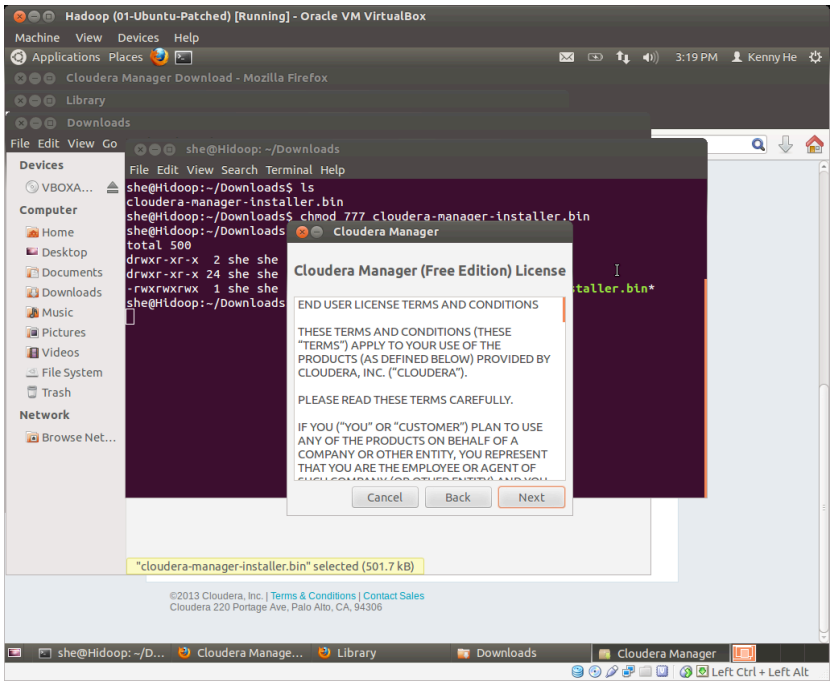
2.2 Download and install the Cloudera Manager

Download the installer “cloudera-manager-installer.bin”, change its attributes to “500”, and then run it

with root previledge:

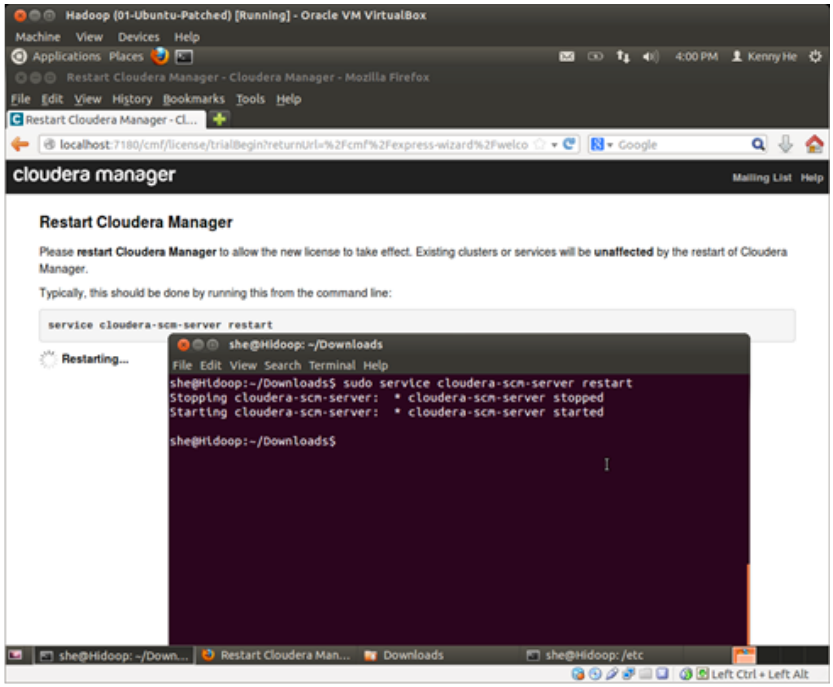
```
sudo  cloudera-manager-installer.bin
```

The installation application will download the application from http://www.cloudera.com/ and install. Depends on the network speed, the installation may take very long time. After successfully install it, create a VM snapshot.

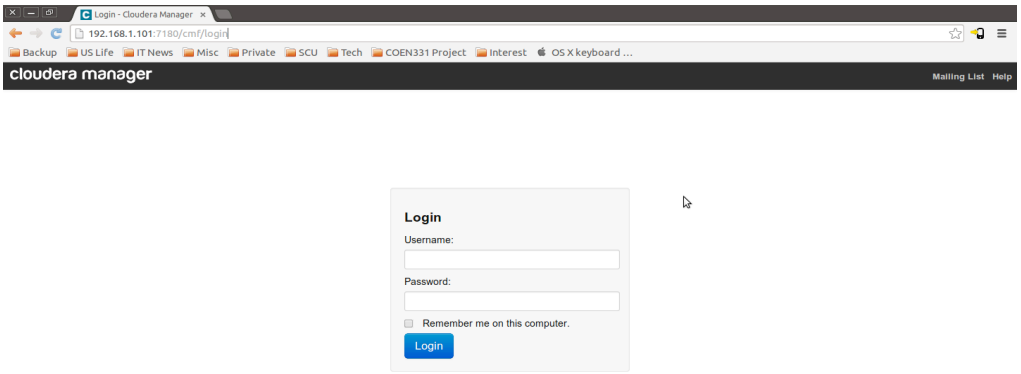


2.3 Start the Cloudera Manager Admin Console

Run the admin console by opening the URL http://localhost:7180/ in the web browser of the guest OS.

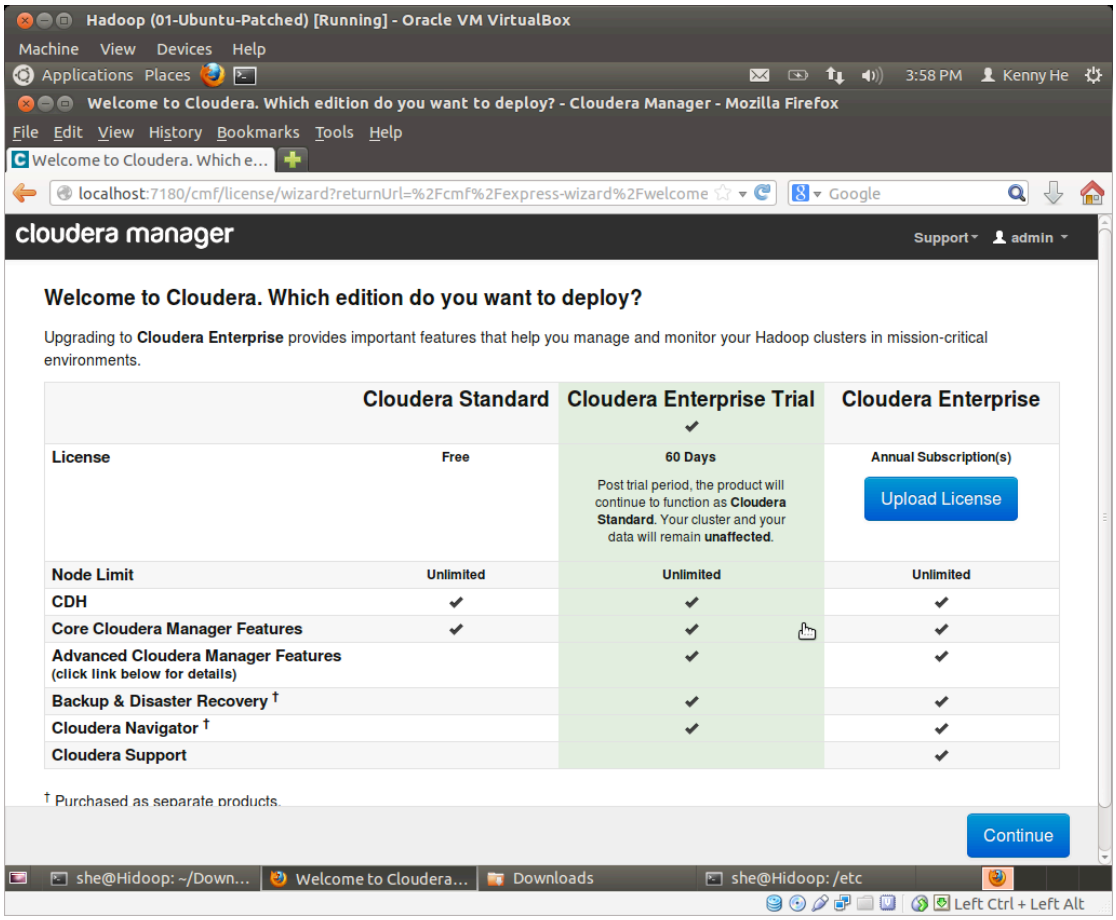


If the network connection between the host and guest OS has been setup correctly, the admin console on the host OS by opening the URL http://192.168.1.101:7180/ (the following screen shot is got from my host OS):



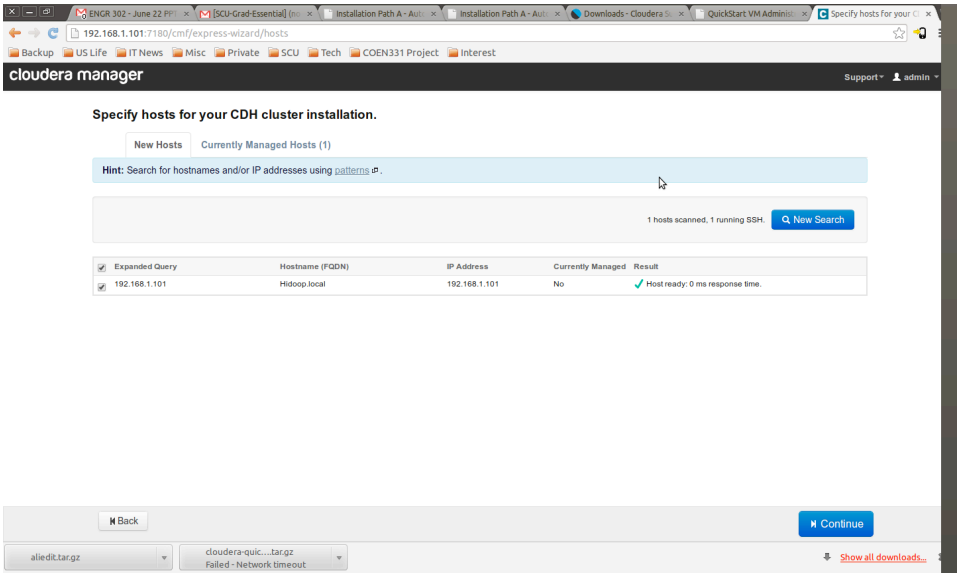
Then we can continue and restart the cloudera-scm-server service according to the instruction on the screen.

In the next page, the installation program will ask which edition to choose. Of course we can only choose the "Cloudera Standard" edition;



2.4 Install cluster nodes

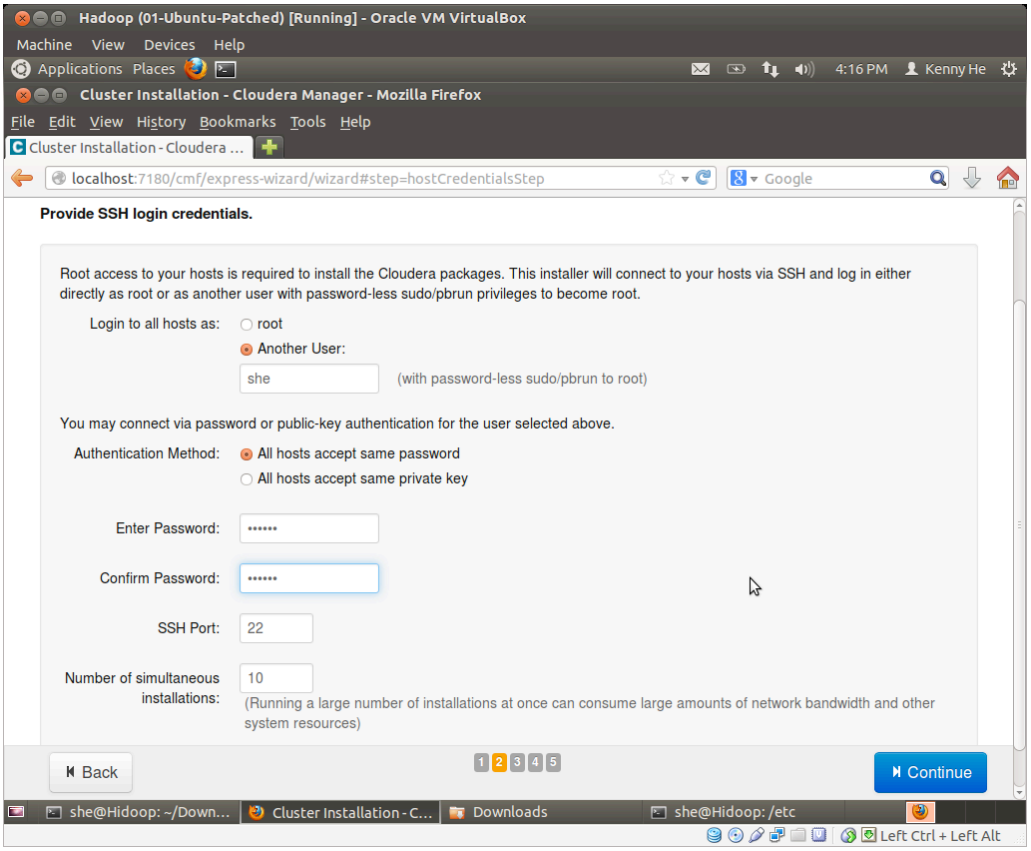
Click “Continue” and we can see a screen to ask for searching the nodes which will be installed as clusters. Input the address pattern 192.168.1.[100-101] (Refer to the IP address I set in chapter 1 when I was preparing the virtual machine network setup). And then go forward.



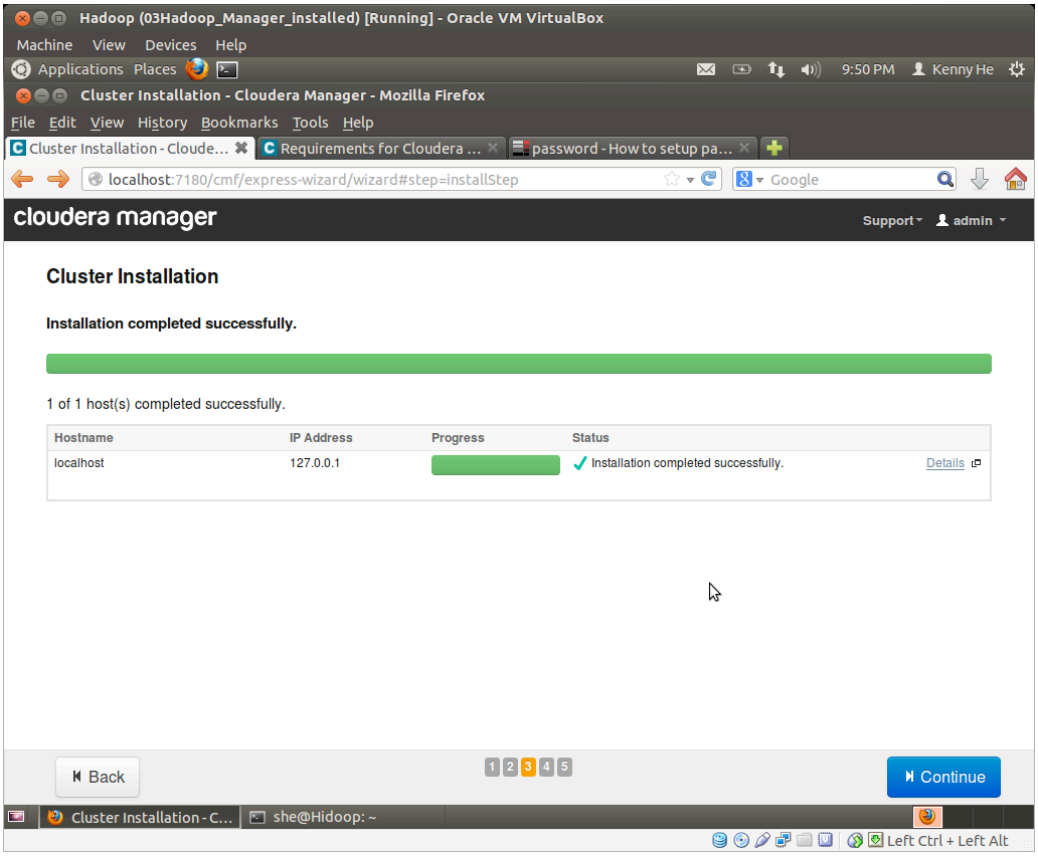
A valid host name and IP address should be provided in this step. We can also use “localhost” or “127.0.0.1”, but that is not recommended since the loop back IP address can be used for testing only and should never use them in real business network.

Fill in the installation options in the next 5 steps.

If the system cannot continue due to “failed to sudo”, that means you have not created an account which has password-less sudo privilege. Please go back to see how to create such account in “3. Post-install configurations” of chapter 1.

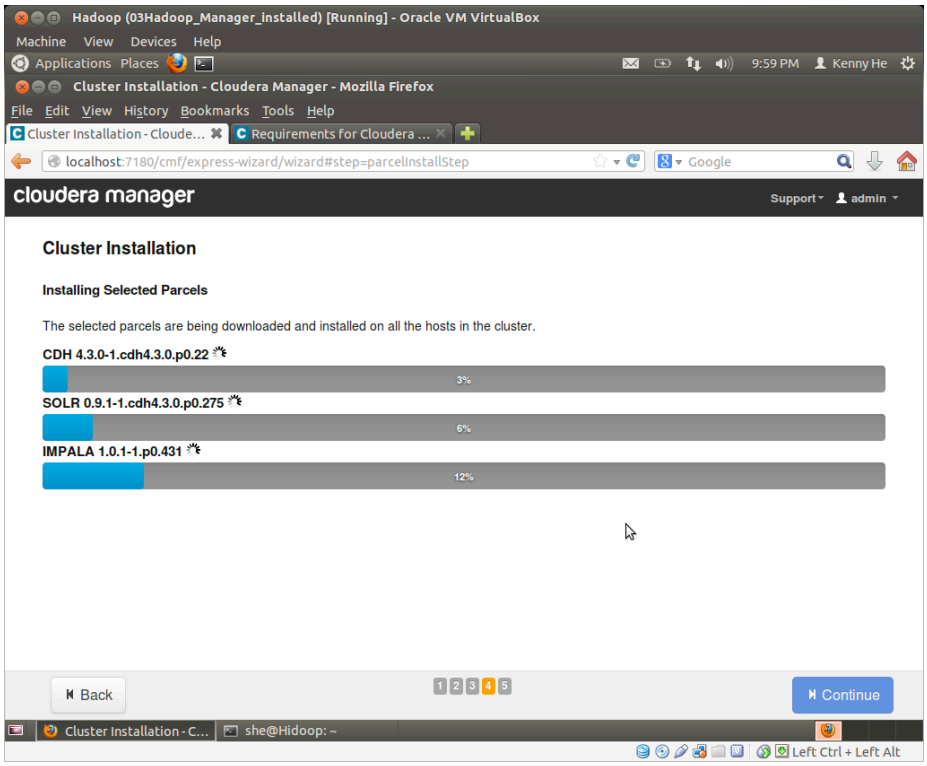


Then you can see the “Installation complete successfully” screen as below. Create a VM snapshot.



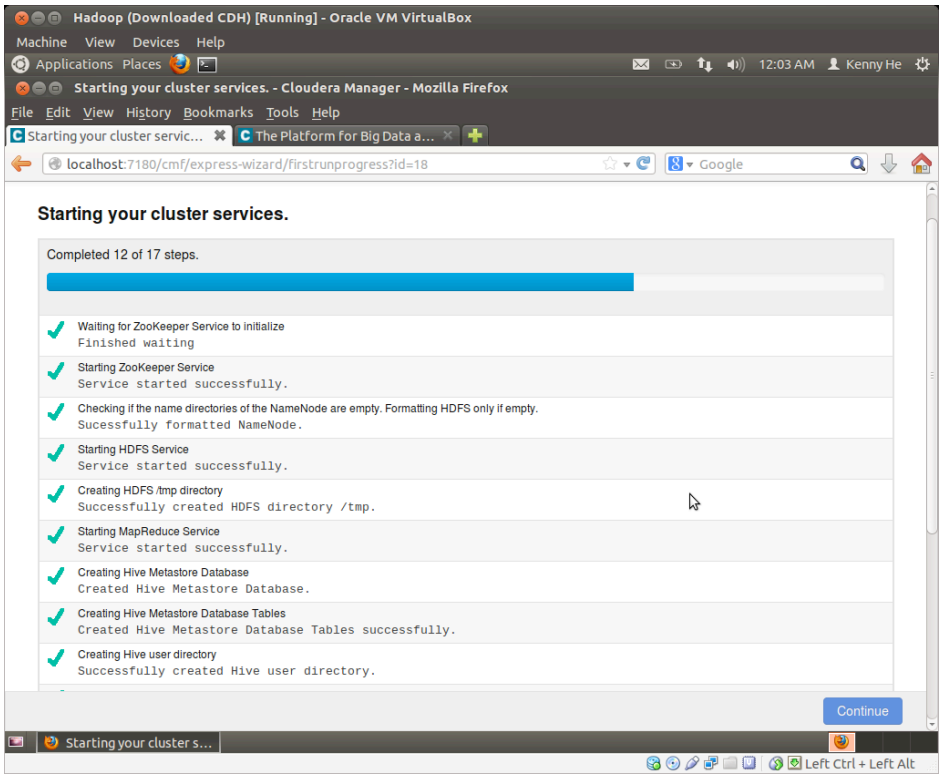
2.5 Install the parcels

Move forward by click the “Continue” button and install the parcels. The installation of Parcels may take a long time (about one hour on my computer). Create another snapshot of the VM once parcels install finished.

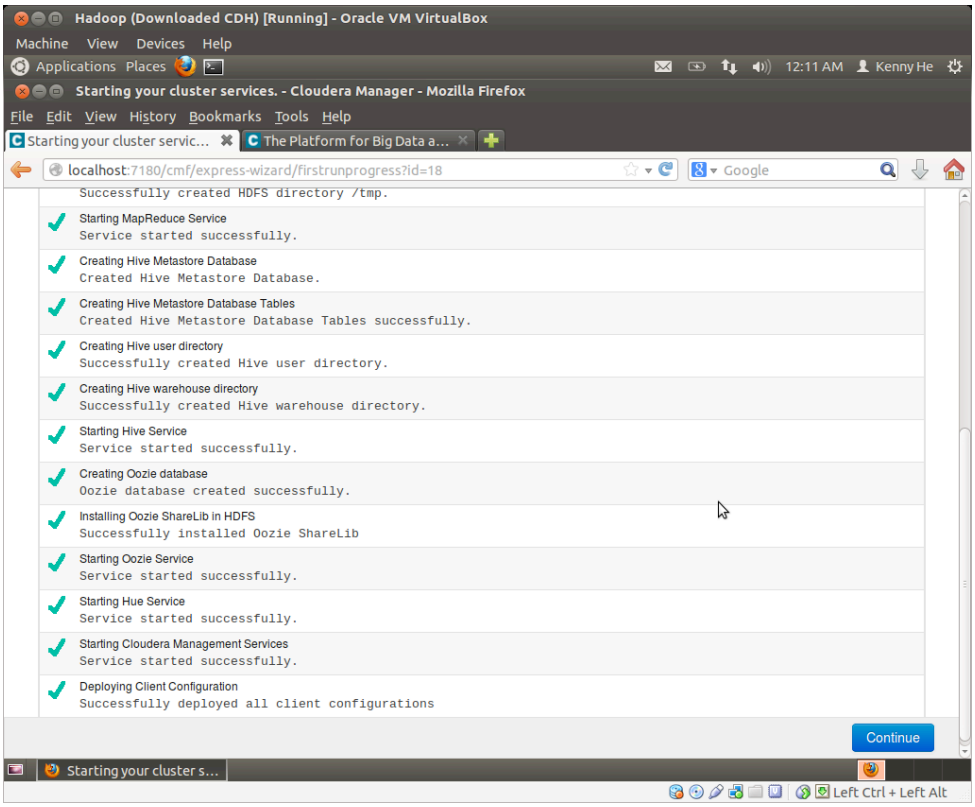


2.6 Install Hadoop services

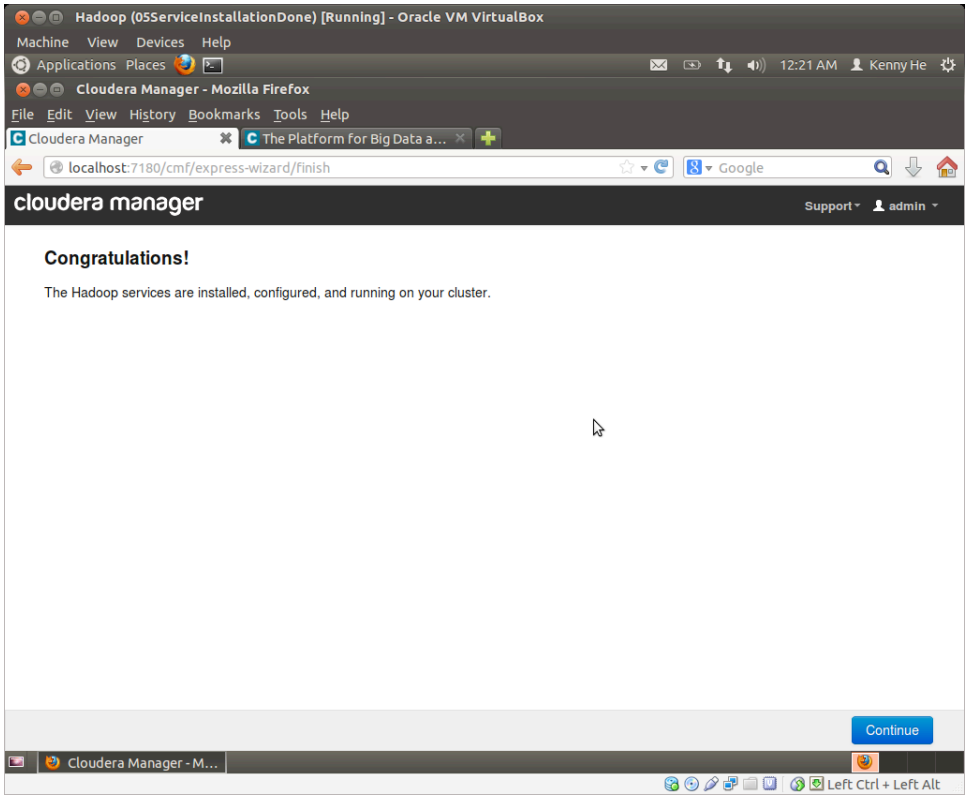
After installation finished, select “Core Hadoop services”, test the database connection, and then continue. The system will start the Hadoop cluster services:



Once all the services have been successfully started, that means the system installation finished. Click the “Continue” button. Create a snapshot of the virtual machine.

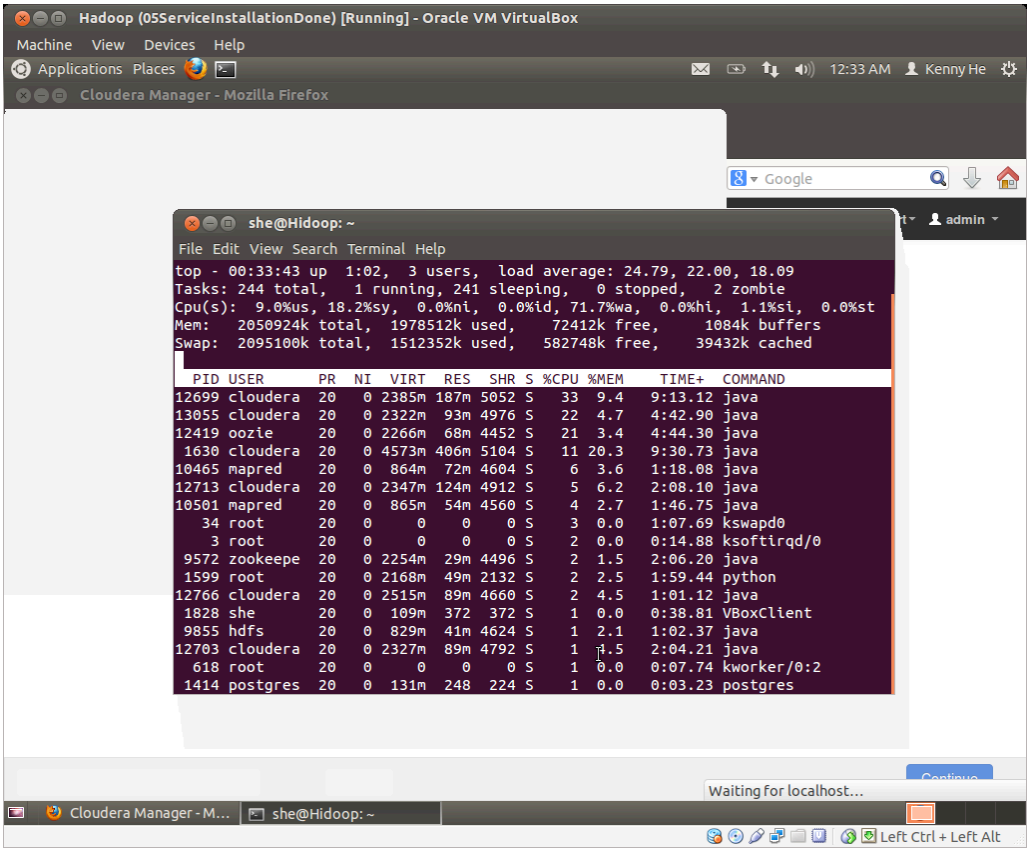


Click the “Continue”, wait for a few minutes, the page will be redirected to the “Congratulations”:

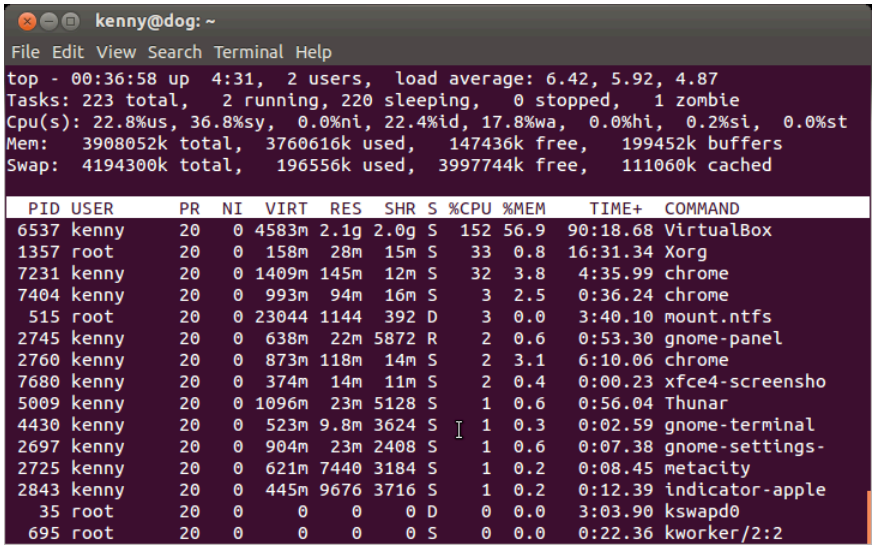


2.7 Try to tune the performance of the server

The virtual machine is running Hadoop now and it becomes very busy and almost freezes. Open a terminal and run “top” command and see that how busy the cluster runs:



And I ran “top” command on my host OS, the VirtualBox consumes almost 150%-300% CPU time (my CPU has 4 cores):



The virtual machine is almost freezes so I can only SSH from my host machine to it and do some tuning.

First, change the init mode so the VM starts only in character UI and disable the GUI.

Secondly, disable all the services.

Thirdly, shut down the VM, disable the USB, sound card, and other useless HW from the VM, and give it more resources such as CPU cores and RAM.

Finally, copy the virtual machine to a faster computer.

Chapter 3. Run a testing program on the MapReduce

3.1 Preparation

Just follow the guide of "Hadoop Tutorial": http://www.cloudera.com/content/cloudera-content/cloudera-docs/HadoopTutorial/CDH4/Hadoop-Tutorial.html

Since I cannot get enough hardware resource to run a cluster with three or more virtual machines with Hadoop installed, I can only try to run the test program on a single node cluster. Then I need to make some configurations to the Hadoop virtual machine I installed to single node mode.

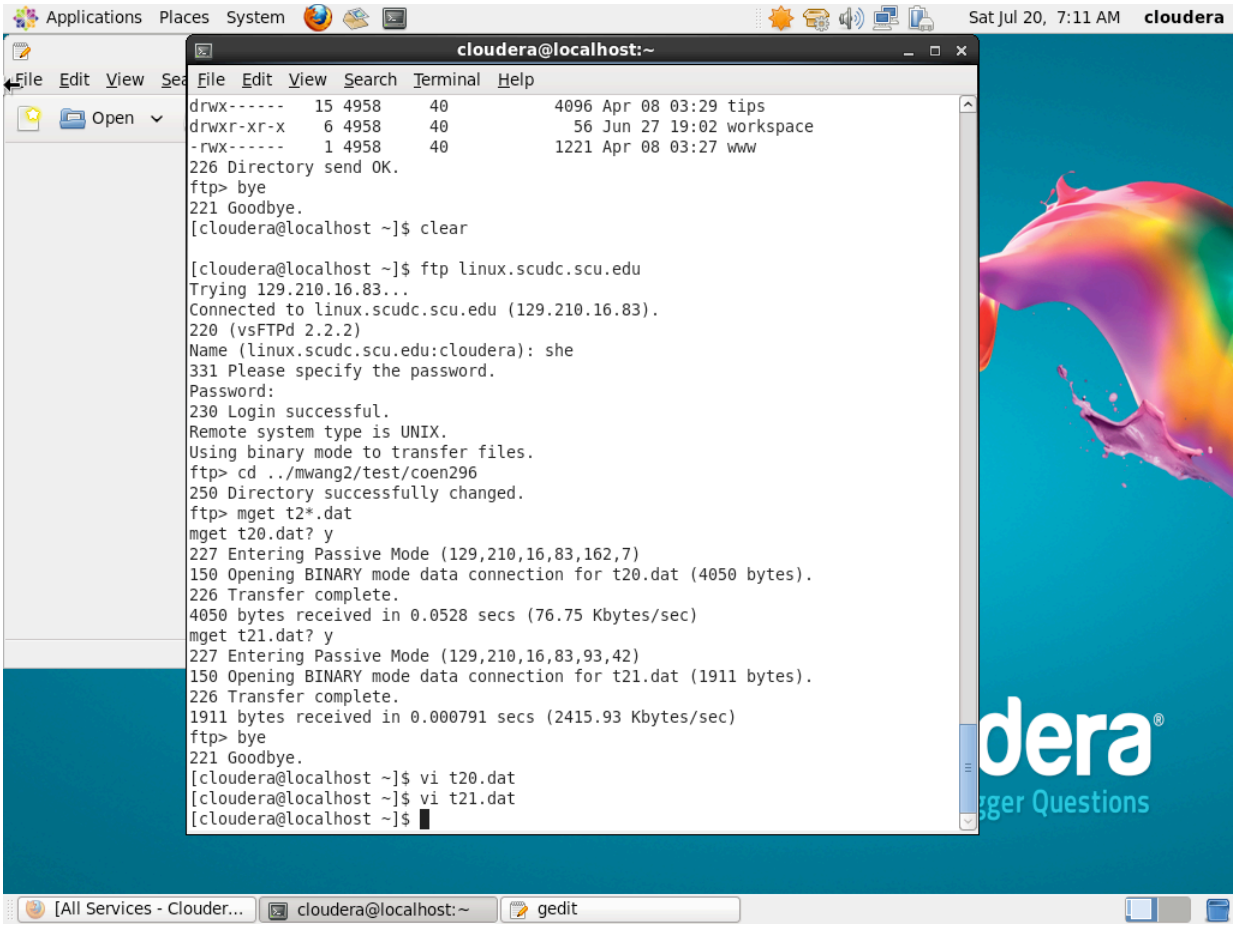
To make the life easier, I simply use the preinstalled quick VM provided by cloudera: https://ccp.cloudera.com/display/SUPPORT/Cloudera+QuickStart+VM

Download the pre-build VM for VirtualBox and import.

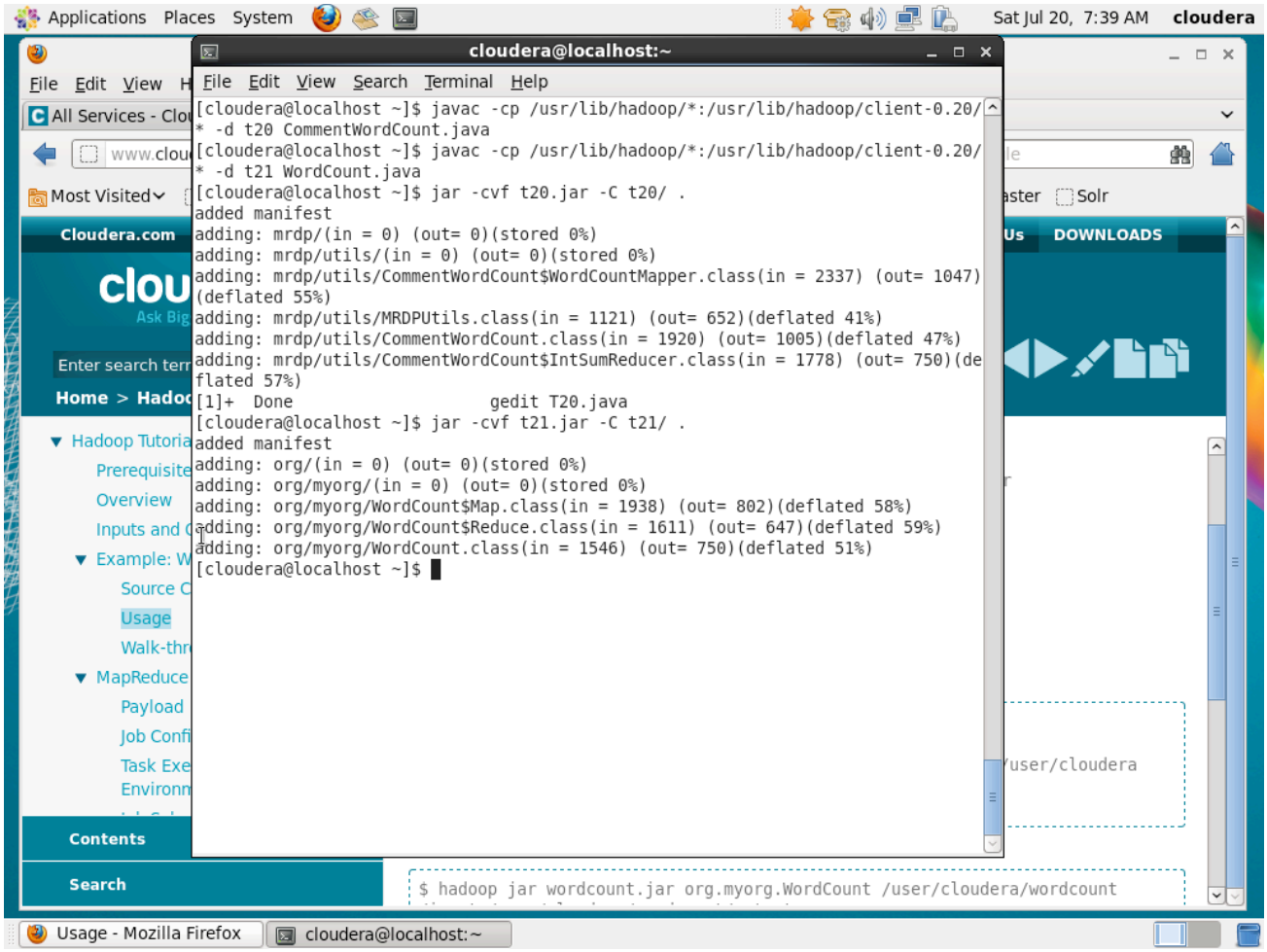
Download the latest stable (version 1.1.2) Hadoop MapReduce library (for testing Java application) from Apache:

3.2 Download testing source code

Ftp to linux.scudc.scu.edu, and download source code from /users/mwang2/test/coen296/:

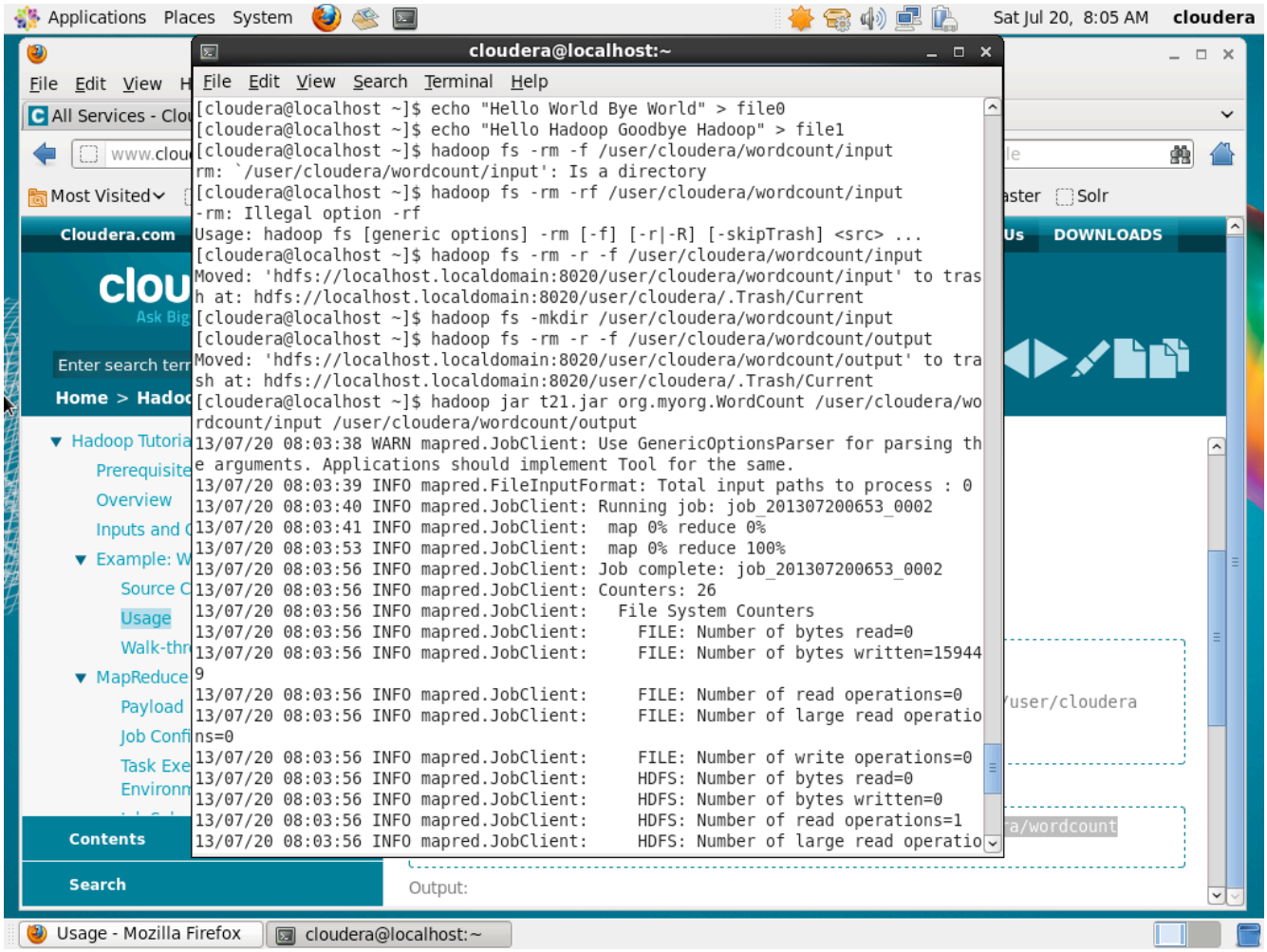


Read the source code, and correct the errors in the source code. And rename the files to major class names. Then compile the test java applications and create the .jar packages:



3.3 Prepare the test data and input/output folder in HDFS

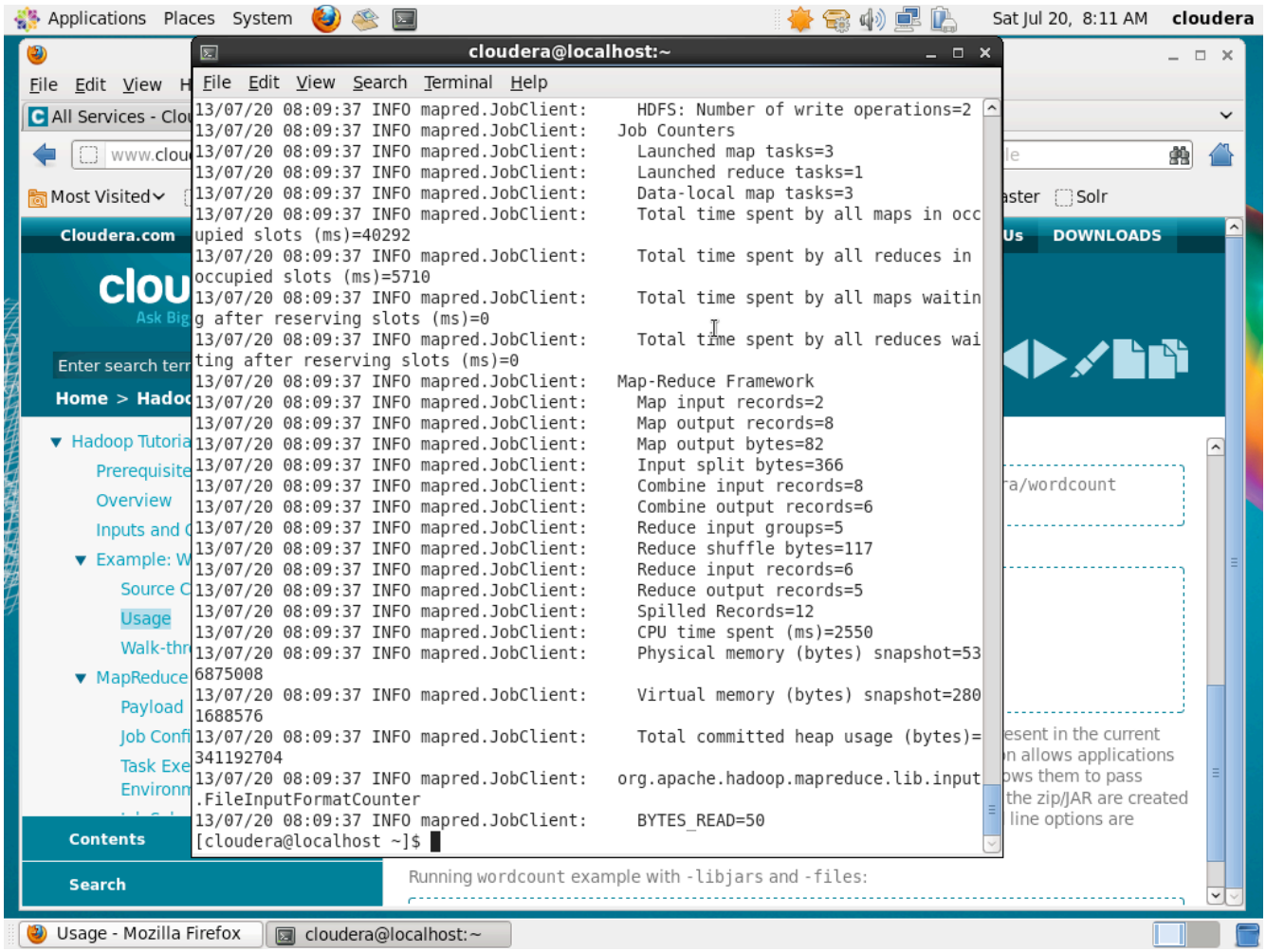
Follow the tutorial, create two input text files, file0 and file1, and upload them into the Hadoop HDFS file system. Since I had run the test before, I need to remove the existing input and output folders first.



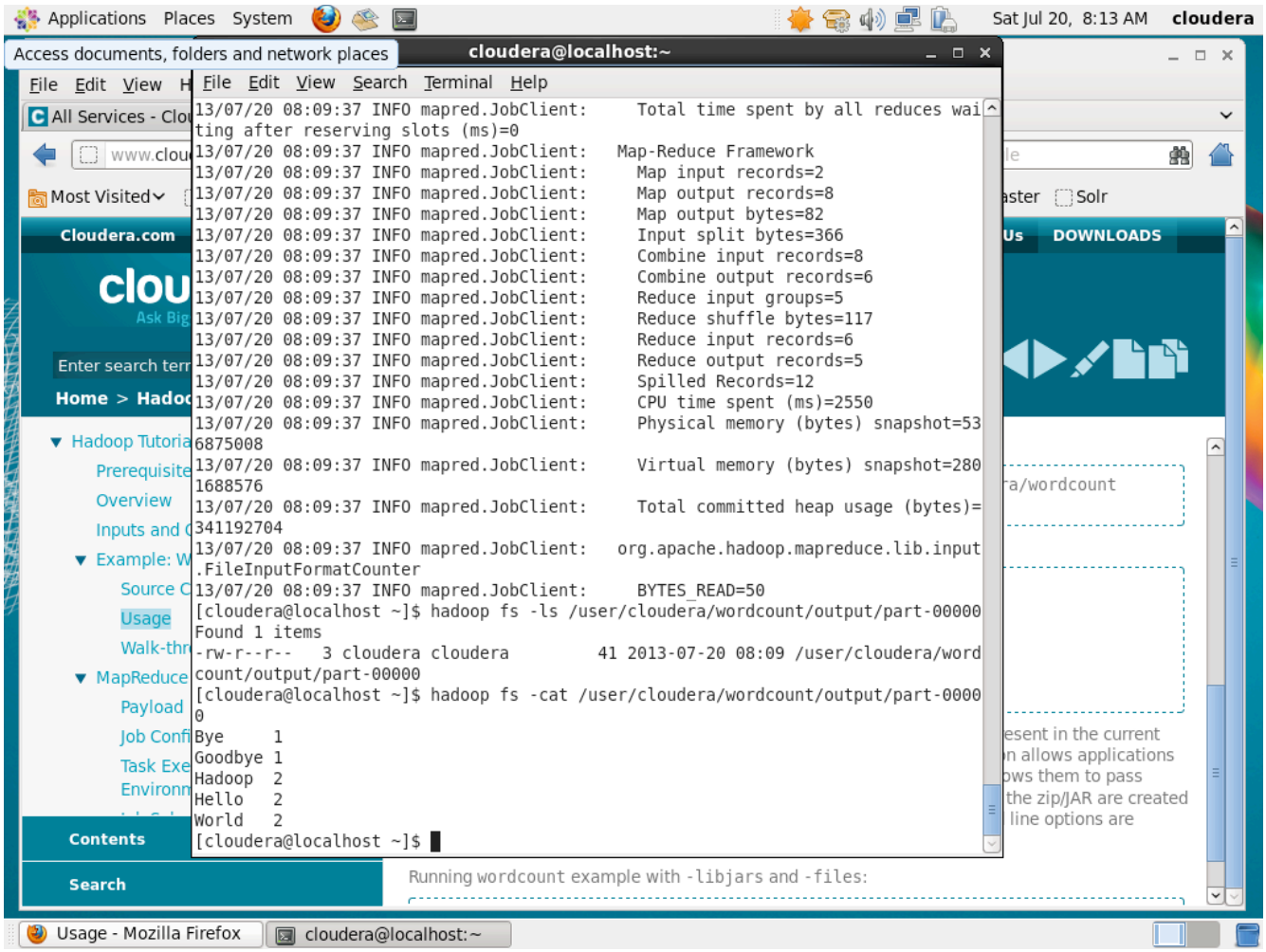
(In the picture above, I forgot upload the test files to HDFS before running the test. You can see the read=0 and definitely the output is nothing. I corrected the issue in the following operations and please refer to the following screen shots.)

3.4 Run test case t21.dat (successful)

Run the test application t21.dat (the sample Java test code from the Hadoop Tutorial, the famous WordCount.java) on Hadoop in the terminal. The Hadoop platform prints out the verbose information about how the job is divided and dispatched. Since I run Hadoop on a single node cluster, the system only emulate the task divide and dispatch and actually all tasks are run on the same node.

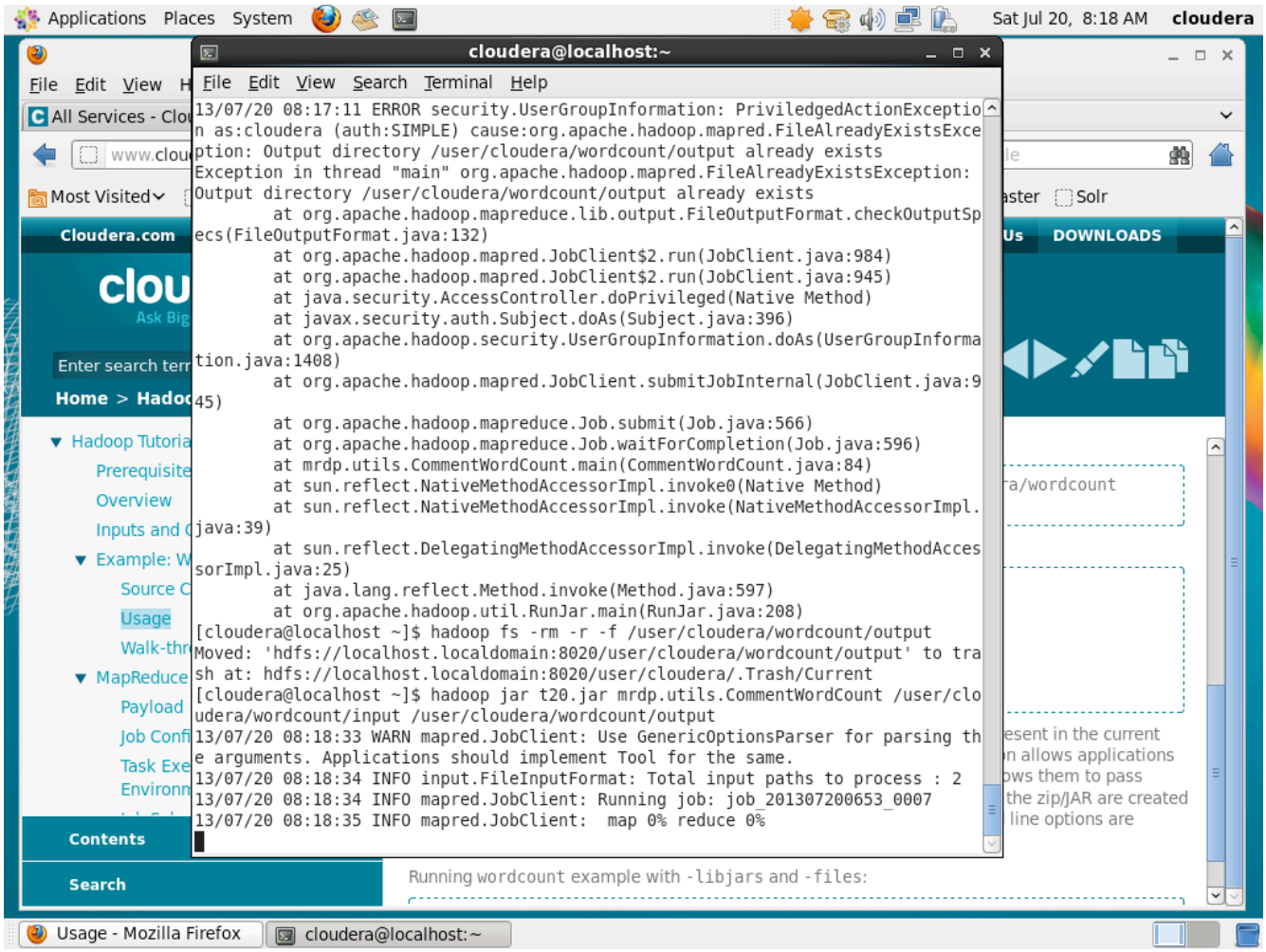


After finished, check the contents in the output files and found they are correct:

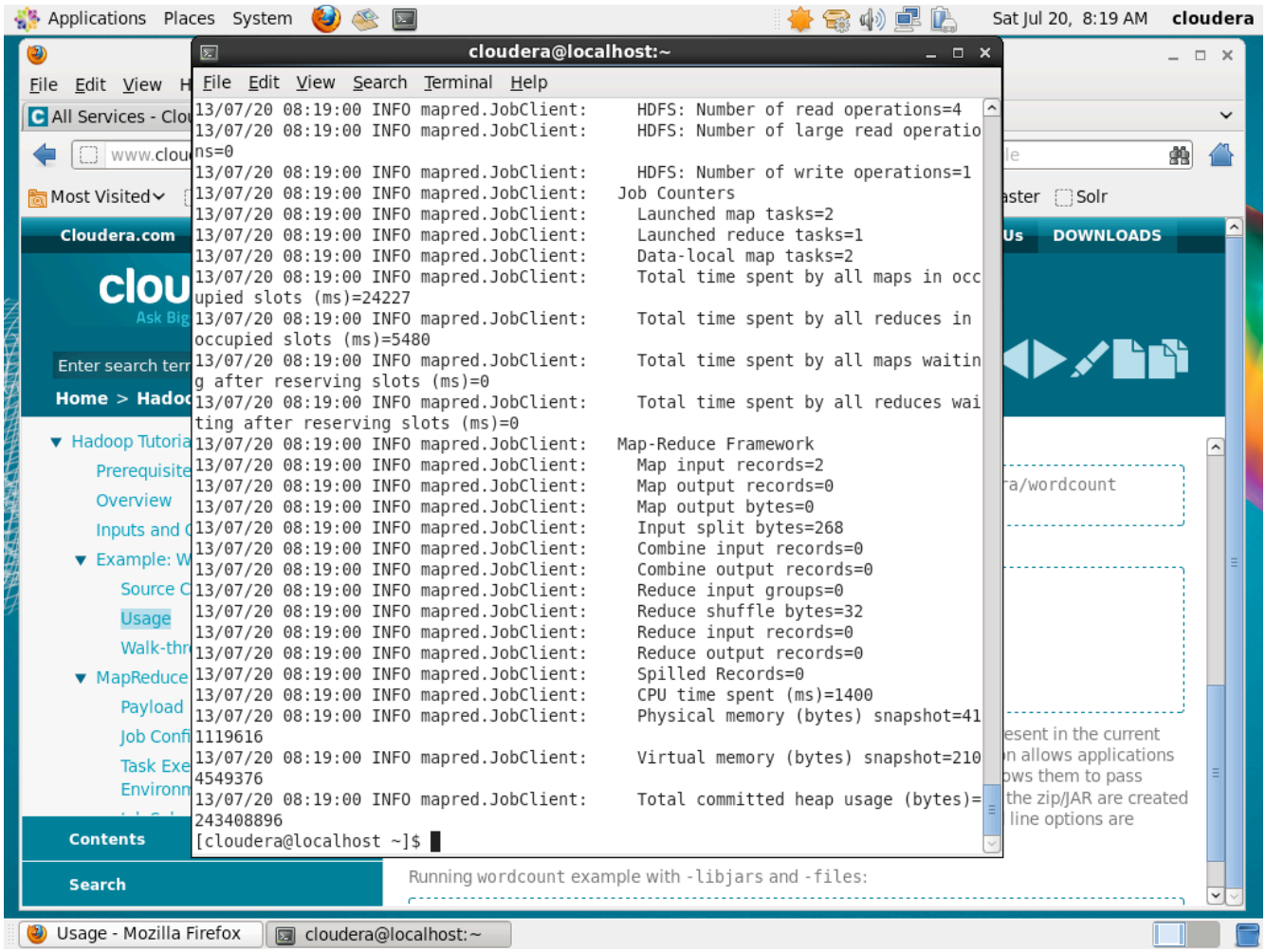


3.5 Run test case t20.dat (no output)

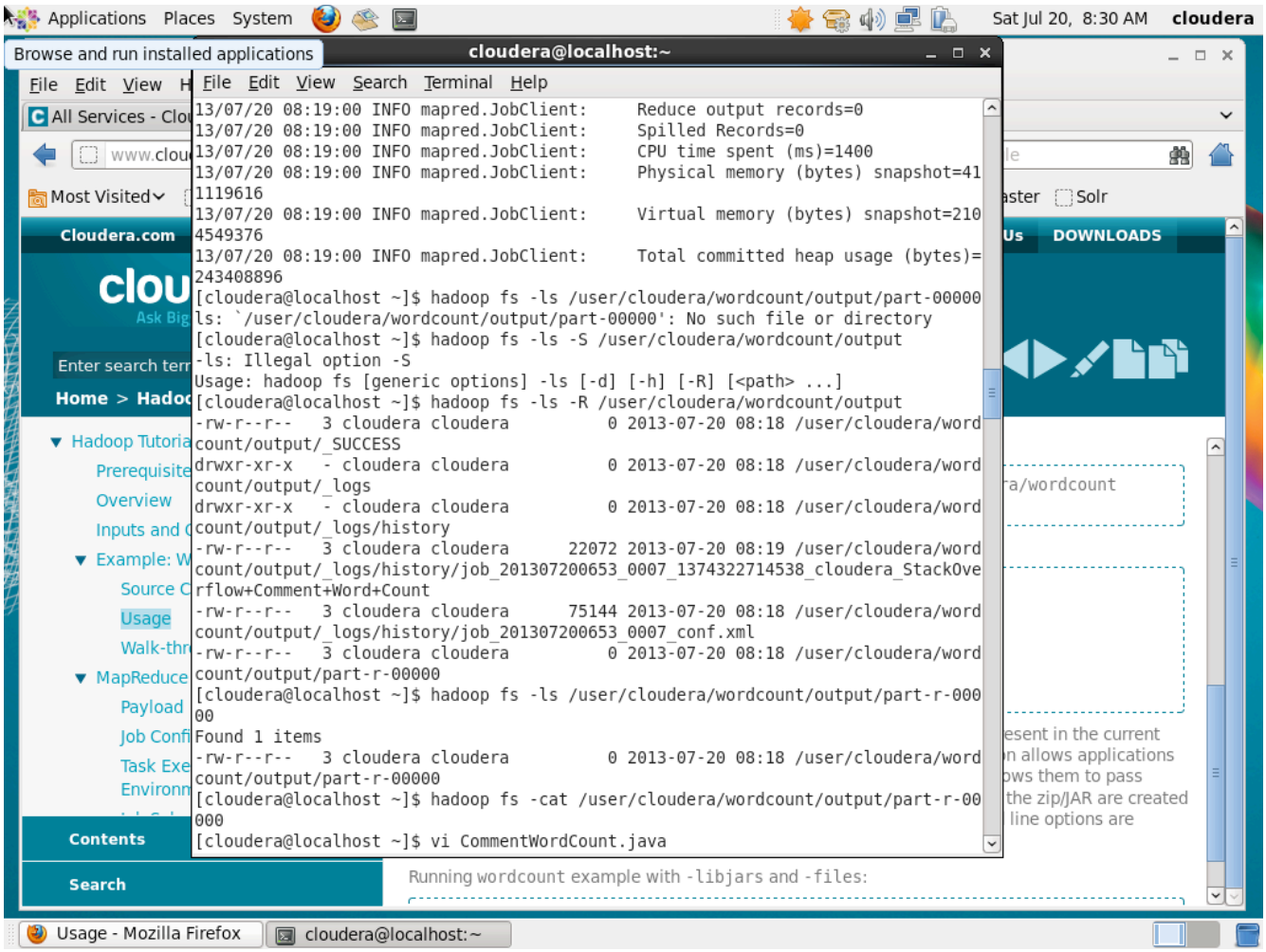
Run the test application t20.dat (The CommentWordCount.java provided by the professor) in the same way. Need to clear the output folder before running the application. Or, will see the “folder exist” exception.



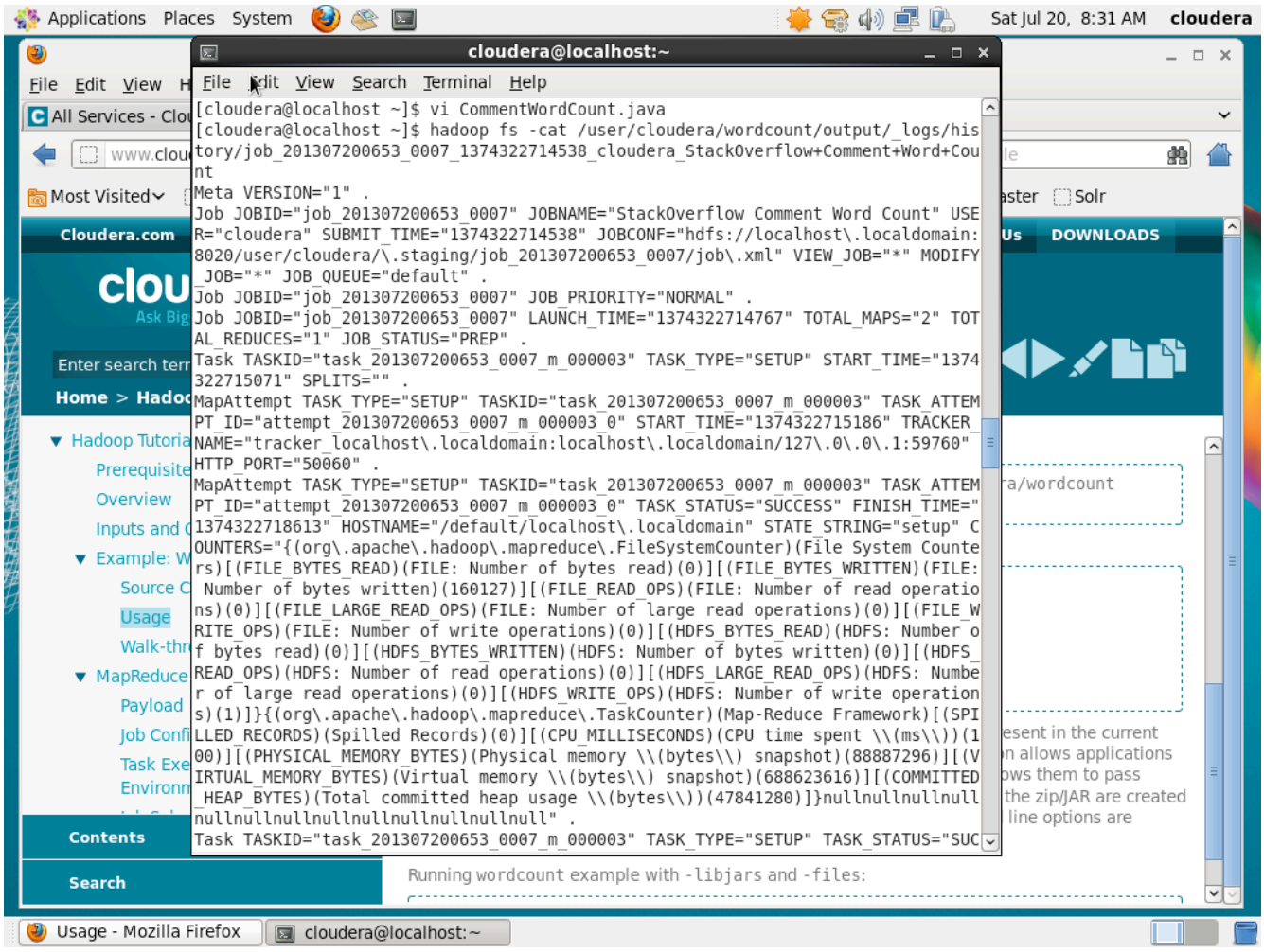
The application ran successfully. But there is no output. Am I wrong in any stage?



I tried to print the output file, but there is nothing inside. Am I wrong?



Checked the application source code and the Hadoop logs, and could not find out any clue about the issue.


Applications Places System
Sat Jul 20, 8:31 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ vi CommentWordCount.java
[cloudera@localhost ~]$ hadoop fs -cat /user/cloudera/wordcount/output/_logs/history/job_201307200653_0007_1374322714538_cloudera_StackOverflow+Comment+Word+Count
Meta VERSION="1" .
Job JOBID="job_201307200653_0007" JOBNAME="StackOverflow Comment Word Count" USER="cloudera" SUBMIT_TIME="1374322714538" JOBCONF="hdfs://localhost\.localdomain:8020/user/cloudera/\.staging/job_201307200653_0007/job\.xml" VIEW_JOB="*" MODIFY_JOB="*" JOB_QUEUE="default" .
Job JOBID="job_201307200653_0007" JOB_PRIORITY="NORMAL" .
Job JOBID="job_201307200653_0007" LAUNCH_TIME="1374322714767" TOTAL_MAPS="2" TOTAL_REDUCES="1" JOB_STATUS="PREP" .
Task TASKID="task_201307200653_0007_m_000003" TASK_TYPE="SETUP" START_TIME="1374322715071" SPLITS="" .
MapAttempt TASK_TYPE="SETUP" TASKID="task_201307200653_0007_m_000003" TASK_ATTEMPT_ID="attempt_201307200653_0007_m_000003_0" START_TIME="1374322715186" TRACKER_NAME="tracker_localhost\.localdomain:localhost\.localdomain/127\.0\.0\.1:59760" HTTP_PORT="50060" .
MapAttempt TASK_TYPE="SETUP" TASKID="task_201307200653_0007_m_000003" TASK_ATTEMPT_ID="attempt_201307200653_0007_m_000003_0" TASK_STATUS="SUCCESS" FINISH_TIME="1374322718613" HOSTNAME="/default/localhost\.localdomain" STATE_STRING="setup" COUNTERS="{(org\.apache\.hadoop\.mapreduce\.FileSystemCounter)(File System Counters)[(FILE_BYTES_READ)(FILE: Number of bytes read)(0)][(FILE_BYTES_WRITTEN)(FILE: Number of bytes written)(160127)][(FILE_READ_OPS)(FILE: Number of read operations)(0)][(FILE_LARGE_READ_OPS)(FILE: Number of large read operations)(0)][(FILE_WRITE_OPS)(FILE: Number of write operations)(0)][(HDFS_BYTES_READ)(HDFS: Number of bytes read)(0)][(HDFS_BYTES_WRITTEN)(HDFS: Number of bytes written)(0)][(HDFS_READ_OPS)(HDFS: Number of read operations)(0)][(HDFS_LARGE_READ_OPS)(HDFS: Number of large read operations)(0)][(HDFS_WRITE_OPS)(HDFS: Number of write operations)(1)]}{(org\.apache\.hadoop\.mapreduce\.TaskCounter)(Map-Reduce Framework)[(SPILLED_RECORDS)(Spilled Records)(0)][(CPU_MILLISECONDS)(CPU time spent \\(ms\\))(100)][(PHYSICAL_MEMORY_BYTES)(Physical memory \\(bytes\\) snapshot)(88887296)][(VIRTUAL_MEMORY_BYTES)(Virtual memory \\(bytes\\) snapshot)(688623616)][(COMMITTED_HEAP_BYTES)(Total committed heap usage \\(bytes\\))(47841280)]}nullnullnullnullnullnullnullnullnullnull" .
Task TASKID="task_201307200653_0007_m_000003" TASK_TYPE="SETUP" TASK_STATUS="SUC
All Services - Clou
Most Visited
Cloudera.com
DOWNLOADS
Home > Hadoo
Hadoop Tutoria
Prerequisite
Overview
Inputs and
Example: W
Source C
Usage
Walk-thr
MapReduce
Payload
Job Confi
Task Exe
Environn
Contents
Search
Running wordcount example with -libjars and -files:
Usage - Mozilla Firefox
cloudera@localhost:~