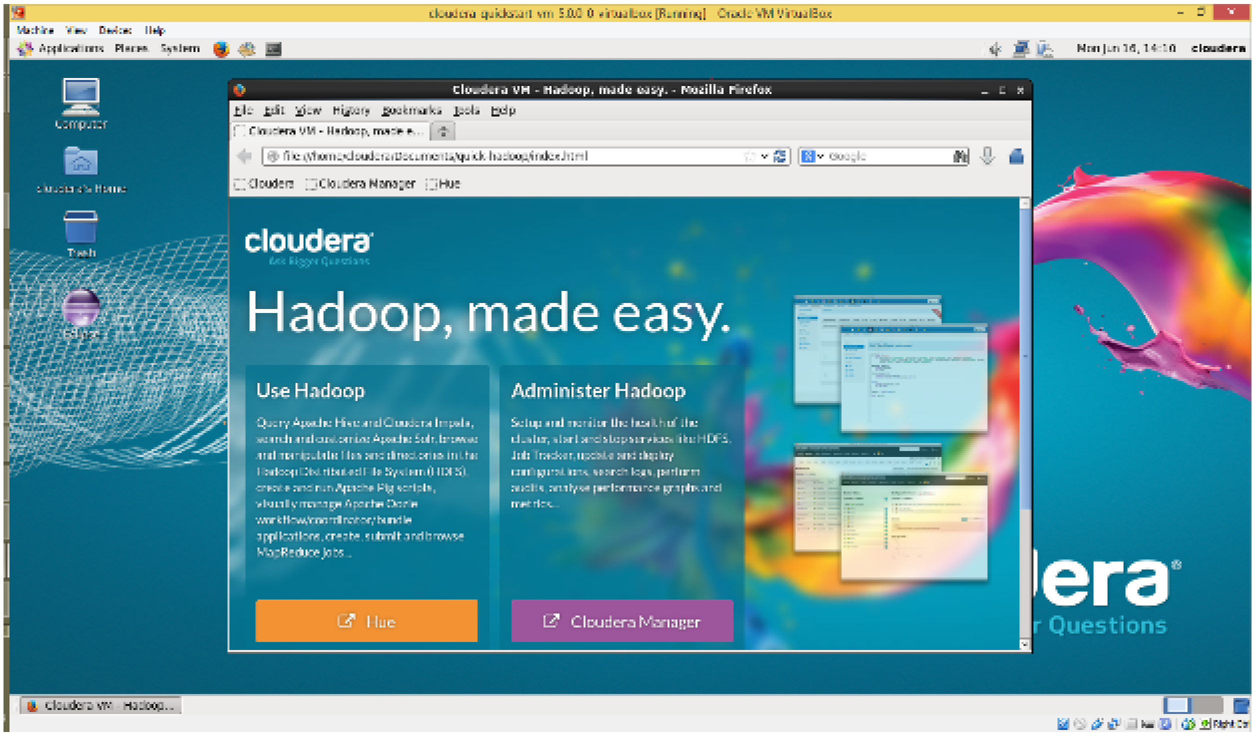
Lab 1 Part 2 Cloudera

Installing Cloudera on LocalMachine

A Cloudera 5.0 VM image is download from
http://www.cloudera.com/content/support/en/downloads/quickstart_vms/cdh-5-0-x.html

Now the VM is booted up from **VirtualBox** and is ready to perform Hadoop jobs.

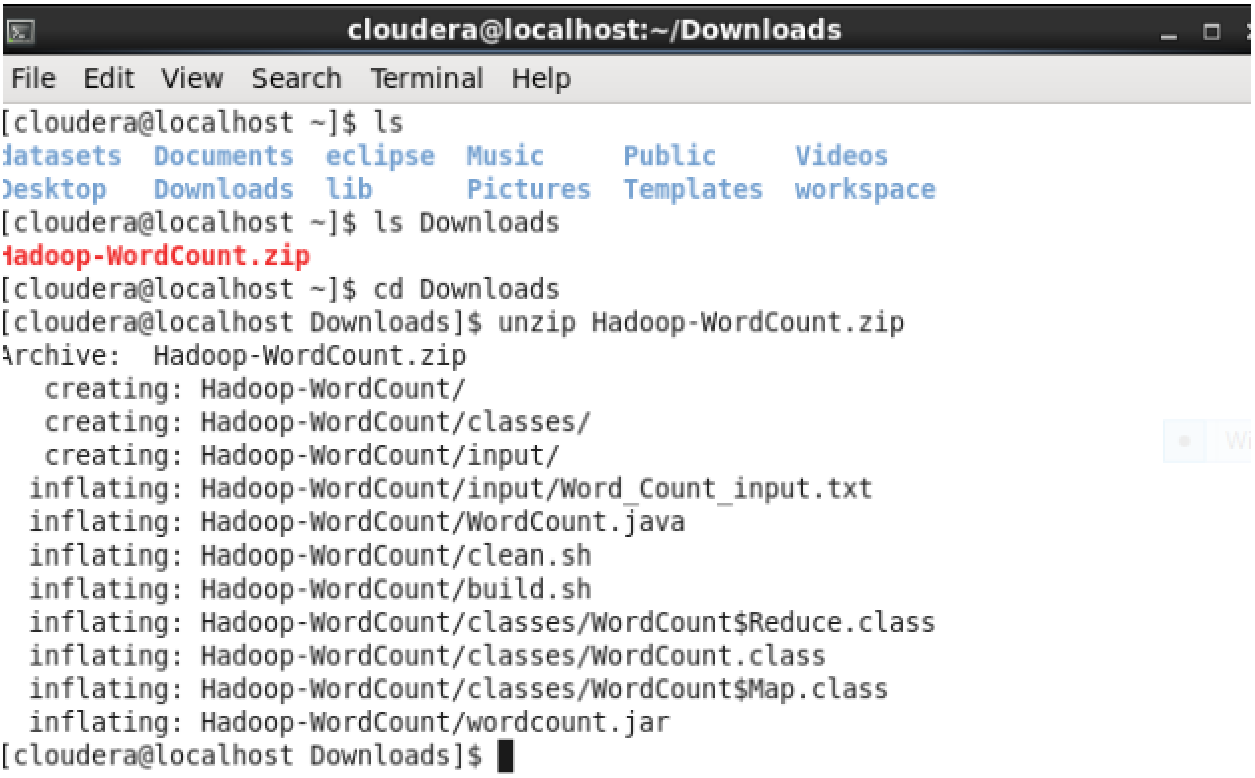

cloudera quickstart vm 5.0.0 0 virtualbox [Running] - Oracle VM VirtualBox
Machine View Devices Help
Applications Places System
Mon Jun 16, 14:10 cloudera
Computer
cloudera's Home
Trash
Eclipse
Cloudera VM - Hadoop, made easy. - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Cloudera VM - Hadoop, made e...
file:///home/cloudera/Documents/quick-hadoop/index.html
Google
Cloudera Cloudera Manager Hue
cloudera
Ask Bigger Questions
Hadoop, made easy.
Use Hadoop
Administer Hadoop
Hue
Cloudera Manager
Cloudera VM - 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...

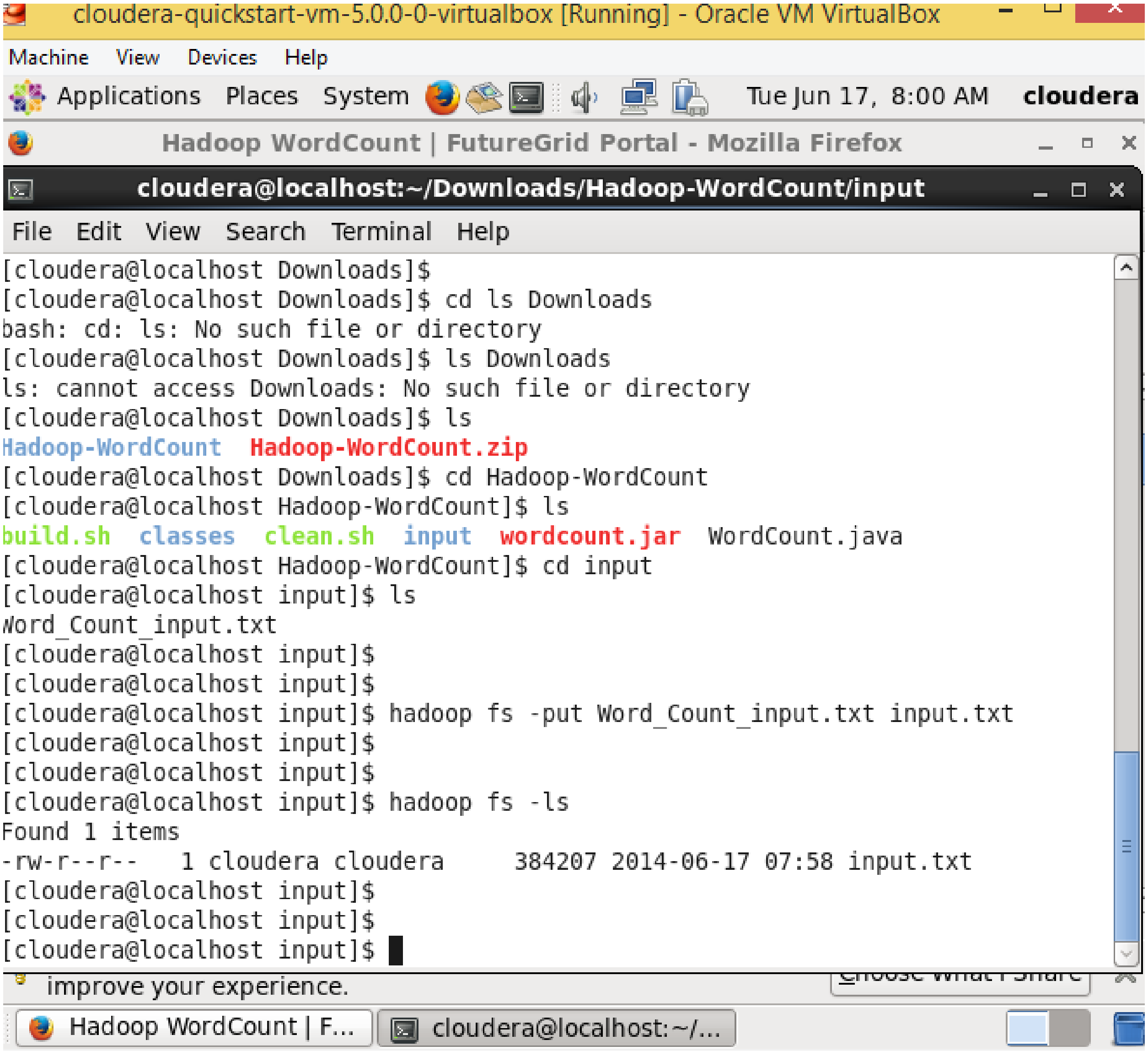
Setup and monitor the health of the cluster, start and stop services like HDFS, Job Tracker, update and deploy configurations, search logs, perform audits, analyse performance graphs and metrics...

Transferring files to Hadoop

Every Command on Hadoop can be performed by
`hadoop -fs`
command, and is same as Unix (Linux) Command


cloudera@localhost:~/Downloads
File Edit View Search Terminal Help

```
[cloudera@localhost ~]$ ls
datasets  Documents  eclipse  Music     Public     Videos
Desktop   Downloads  lib      Pictures  Templates  workspace
[cloudera@localhost ~]$ ls Downloads
Hadoop-WordCount.zip
[cloudera@localhost ~]$ cd Downloads
[cloudera@localhost Downloads]$ unzip Hadoop-WordCount.zip
Archive:  Hadoop-WordCount.zip
   creating: Hadoop-WordCount/
   creating: Hadoop-WordCount/classes/
   creating: Hadoop-WordCount/input/
  inflating: Hadoop-WordCount/input/Word_Count_input.txt
  inflating: Hadoop-WordCount/WordCount.java
  inflating: Hadoop-WordCount/clean.sh
  inflating: Hadoop-WordCount/build.sh
  inflating: Hadoop-WordCount/classes/WordCount$Reduce.class
  inflating: Hadoop-WordCount/classes/WordCount.class
  inflating: Hadoop-WordCount/classes/WordCount$Map.class
  inflating: Hadoop-WordCount/wordcount.jar
[cloudera@localhost Downloads]$ 
```


cloudera-quickstart-vm-5.0.0-0-virtualbox [Running] - Oracle VM VirtualBox
Machine View Devices Help
Applications Places System
Tue Jun 17, 8:00 AM cloudera
Hadoop WordCount | FutureGrid Portal - Mozilla Firefox
cloudera@localhost:~/Downloads/Hadoop-WordCount/input
File Edit View Search Terminal Help
improve your experience.
Hadoop WordCount | F...
cloudera@localhost:~/...

```
[cloudera@localhost Downloads]$
[cloudera@localhost Downloads]$ cd ls Downloads
bash: cd: ls: No such file or directory
[cloudera@localhost Downloads]$ ls Downloads
ls: cannot access Downloads: No such file or directory
[cloudera@localhost Downloads]$ ls
Hadoop-WordCount  Hadoop-WordCount.zip
[cloudera@localhost Downloads]$ cd Hadoop-WordCount
[cloudera@localhost Hadoop-WordCount]$ ls
build.sh  classes  clean.sh  input  wordcount.jar  WordCount.java
[cloudera@localhost Hadoop-WordCount]$ cd input
[cloudera@localhost input]$ ls
Word_Count_input.txt
[cloudera@localhost input]$
[cloudera@localhost input]$
[cloudera@localhost input]$ hadoop fs -put Word_Count_input.txt input.txt
[cloudera@localhost input]$
[cloudera@localhost input]$
[cloudera@localhost input]$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera     384207 2014-06-17 07:58 input.txt
[cloudera@localhost input]$
[cloudera@localhost input]$
[cloudera@localhost input]$
```

For running WordCount Example, Two files from **Local FileSystem** to **HDFS** are transferred

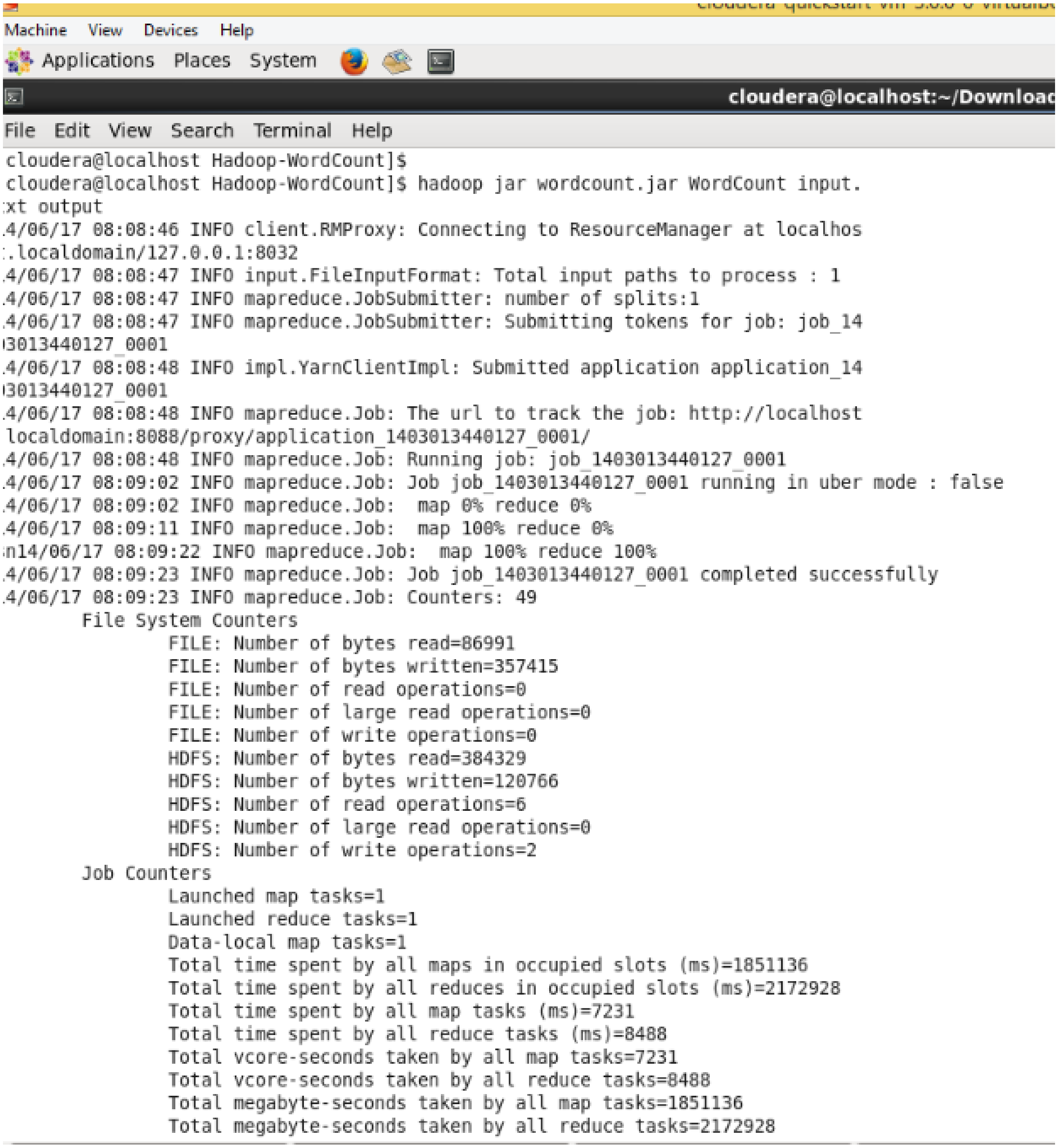
```
hadoop fs -put Word_Count_input.txt input.txt
```

```
hadoop fs -put wordcount.jar worcount.jar
```

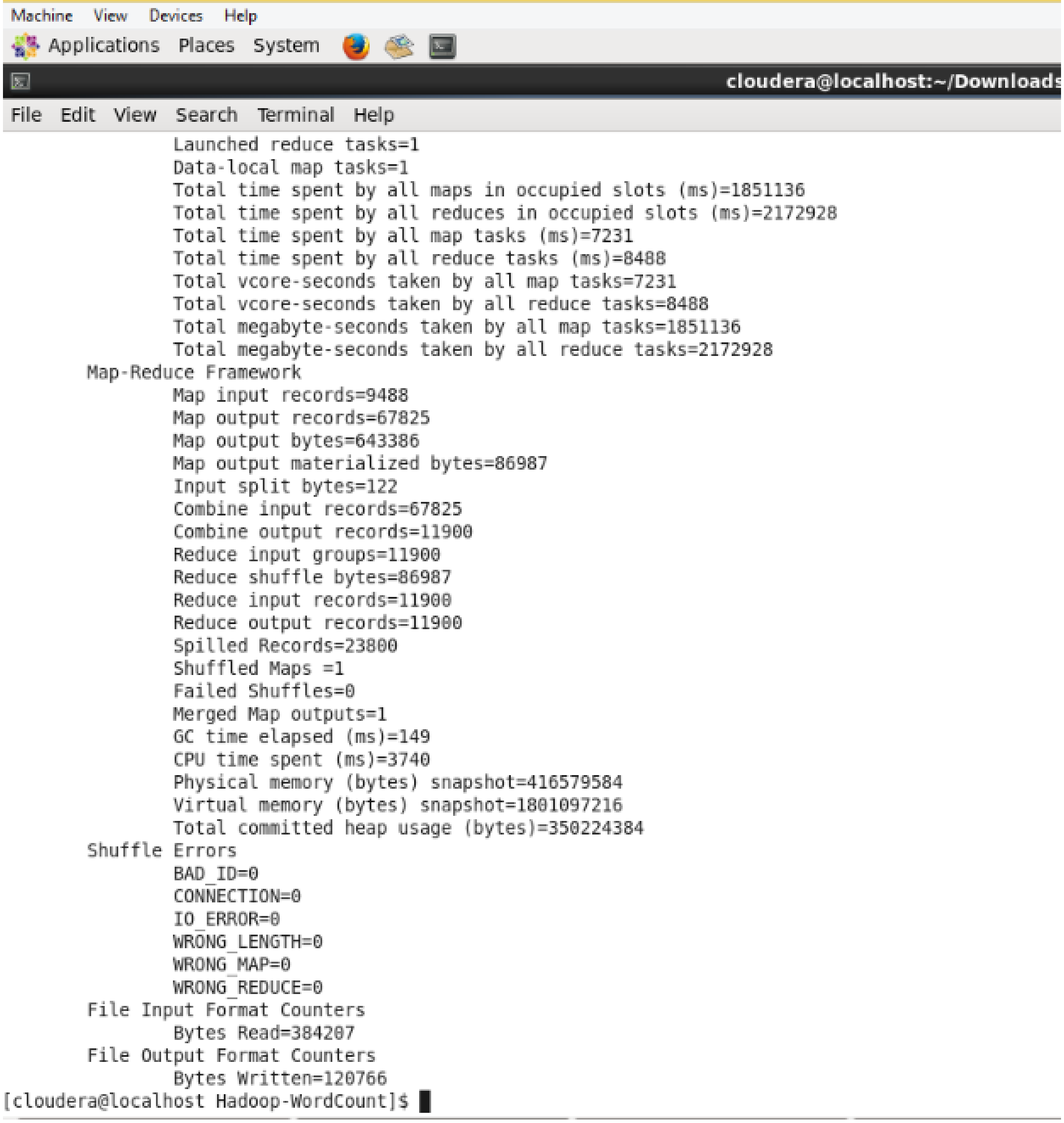
Running WordCount Program on Cloudera

Now as both the files required are moved to HDFS, we're ready to run the program

```
hadoop jar worcount.jar WordCount input.txt output
```


cloudera quickstart vm 5.0.0 0 virtualbo
Machine View Devices Help
Applications Places System
cloudera@localhost:~/Downloac
File Edit View Search Terminal Help

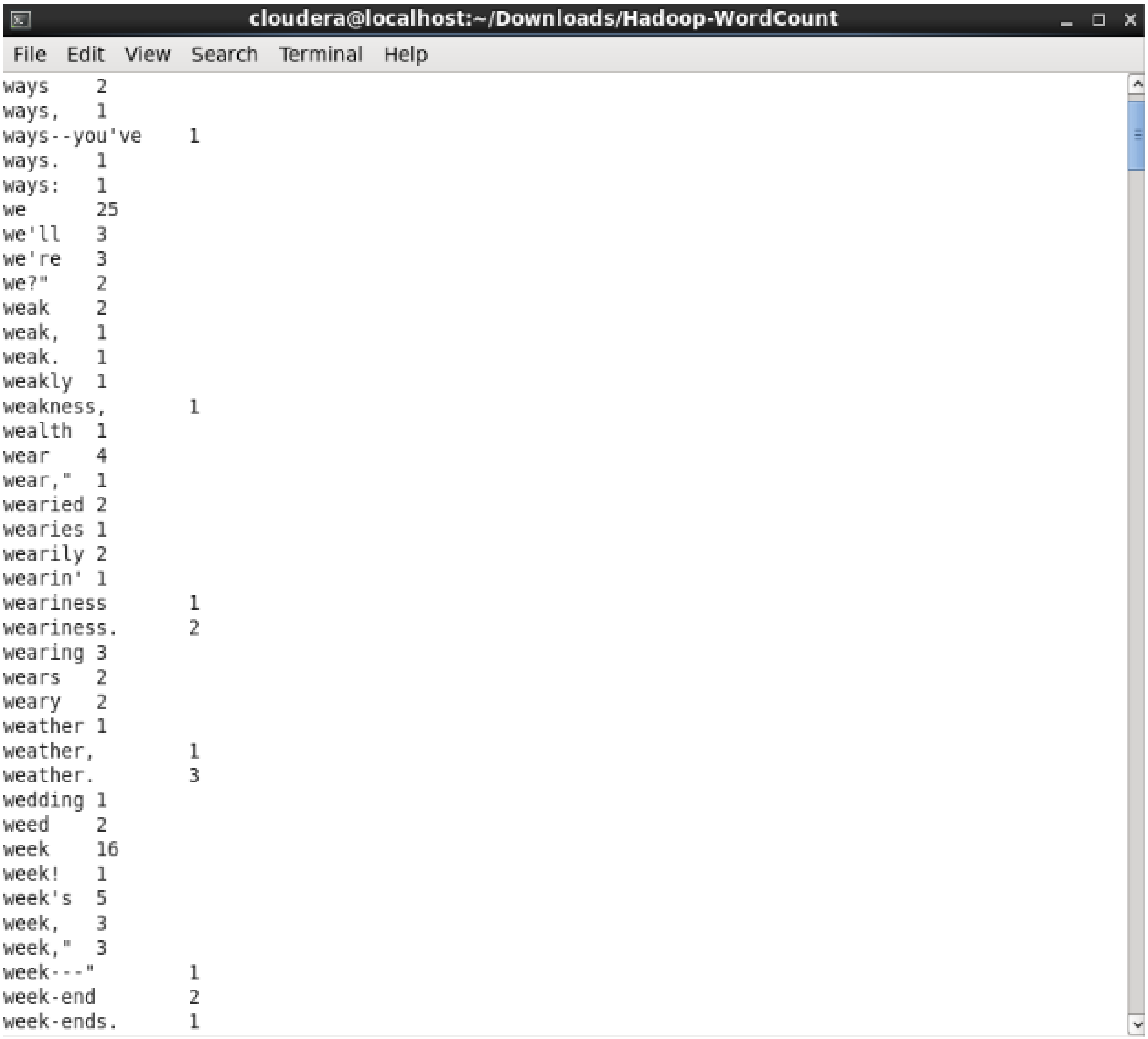
```
cloudera@localhost Hadoop-WordCount]$
cloudera@localhost Hadoop-WordCount]$ hadoop jar wordcount.jar WordCount input.
:xt output
.4/06/17 08:08:46 INFO client.RMProxy: Connecting to ResourceManager at localhos
:.localdomain/127.0.0.1:8032
.4/06/17 08:08:47 INFO input.FileInputFormat: Total input paths to process : 1
.4/06/17 08:08:47 INFO mapreduce.JobSubmitter: number of splits:1
.4/06/17 08:08:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
3013440127_0001
.4/06/17 08:08:48 INFO impl.YarnClientImpl: Submitted application application_14
3013440127_0001
.4/06/17 08:08:48 INFO mapreduce.Job: The url to track the job: http://localhost
localdomain:8088/proxy/application_1403013440127_0001/
.4/06/17 08:08:48 INFO mapreduce.Job: Running job: job_1403013440127_0001
.4/06/17 08:09:02 INFO mapreduce.Job: Job job_1403013440127_0001 running in uber mode : false
.4/06/17 08:09:02 INFO mapreduce.Job:  map 0% reduce 0%
.4/06/17 08:09:11 INFO mapreduce.Job:  map 100% reduce 0%
n14/06/17 08:09:22 INFO mapreduce.Job:  map 100% reduce 100%
.4/06/17 08:09:23 INFO mapreduce.Job: Job job_1403013440127_0001 completed successfully
.4/06/17 08:09:23 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=86991
                FILE: Number of bytes written=357415
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=384329
                HDFS: Number of bytes written=120766
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1851136
                Total time spent by all reduces in occupied slots (ms)=2172928
                Total time spent by all map tasks (ms)=7231
                Total time spent by all reduce tasks (ms)=8488
                Total vcore-seconds taken by all map tasks=7231
                Total vcore-seconds taken by all reduce tasks=8488
                Total megabyte-seconds taken by all map tasks=1851136
                Total megabyte-seconds taken by all reduce tasks=2172928
```


Machine View Devices Help
Applications Places System
cloudera@localhost:~/Downloads
File Edit View Search Terminal Help

```
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1851136
                Total time spent by all reduces in occupied slots (ms)=2172928
                Total time spent by all map tasks (ms)=7231
                Total time spent by all reduce tasks (ms)=8488
                Total vcore-seconds taken by all map tasks=7231
                Total vcore-seconds taken by all reduce tasks=8488
                Total megabyte-seconds taken by all map tasks=1851136
                Total megabyte-seconds taken by all reduce tasks=2172928
        Map-Reduce Framework
                Map input records=9488
                Map output records=67825
                Map output bytes=643386
                Map output materialized bytes=86987
                Input split bytes=122
                Combine input records=67825
                Combine output records=11900
                Reduce input groups=11900
                Reduce shuffle bytes=86987
                Reduce input records=11900
                Reduce output records=11900
                Spilled Records=23800
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=149
                CPU time spent (ms)=3740
                Physical memory (bytes) snapshot=416579584
                Virtual memory (bytes) snapshot=1801097216
                Total committed heap usage (bytes)=350224384
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=384207
        File Output Format Counters
                Bytes Written=120766
[cloudera@localhost Hadoop-WordCount]$ █
```

The word count is run on the input.txt and the Output is generated on output folder

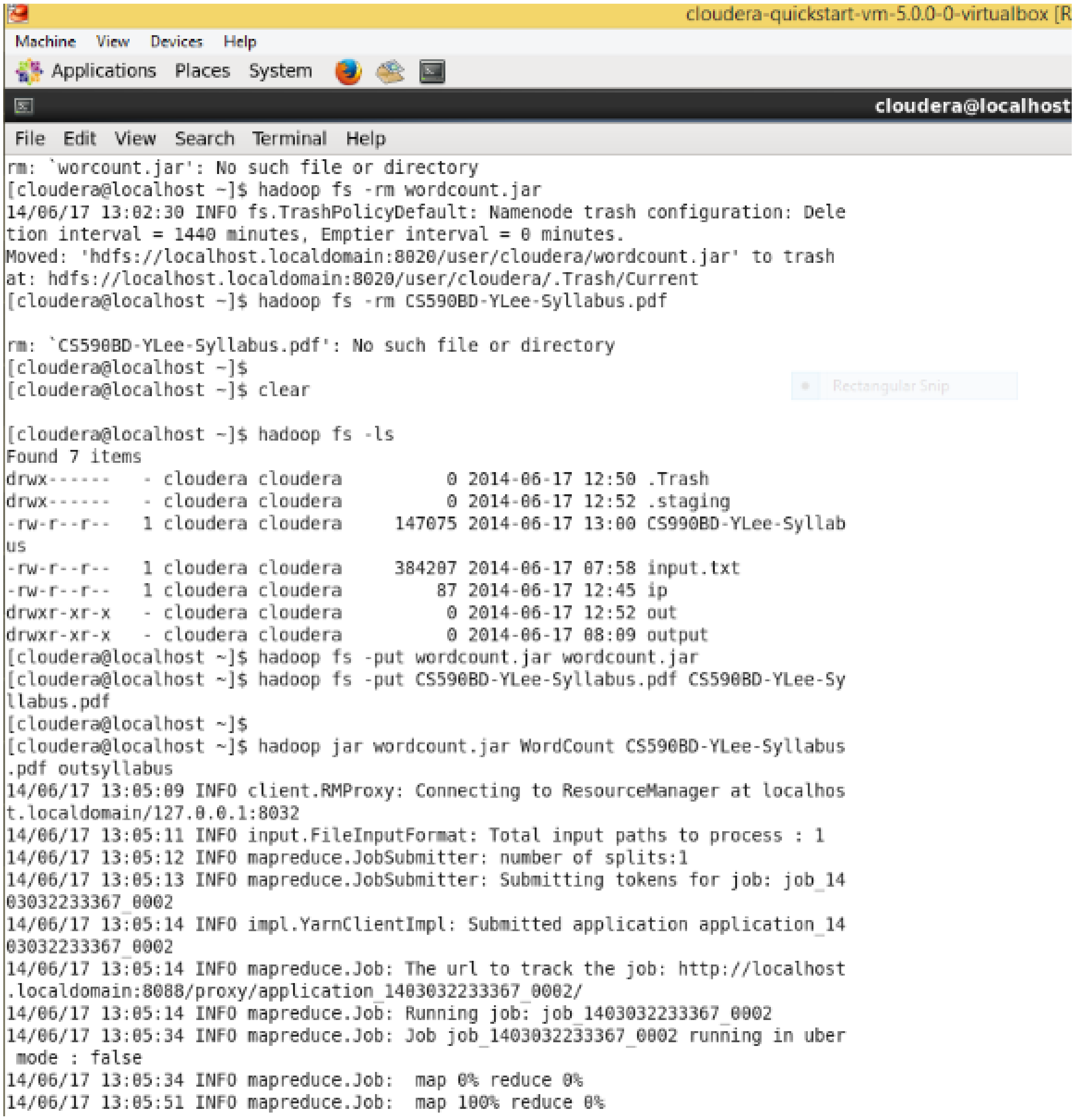
```
hadoop fs -cat output/*
```


cloudera@localhost:~/Downloads/Hadoop-WordCount
File Edit View Search Terminal Help

```
ways    2
ways,   1
ways--you've    1
ways.   1
ways:   1
we      25
we'll   3
we're   3
we?"    2
weak    2
weak,   1
weak.   1
weakly  1
weakness,       1
wealth  1
wear    4
wear,"  1
wearied 2
wearies 1
wearily 2
wearin' 1
weariness       1
weariness.      2
wearing 3
wears   2
weary   2
weather 1
weather,        1
weather.        3
wedding 1
weed    2
week    16
week!   1
week's  5
week,   3
week,"  3
week---"        1
week-end        2
week-ends.      1
```

The Terminal Shows the output, occurrence count of every word form the input.txt

WordCount Program is run on the **CS590BD Syllabus** file.


cloudera-quickstart-vm-5.0.0-0-virtualbox [R
Machine View Devices Help
Applications Places System
cloudera@localhost
File Edit View Search Terminal Help
Rectangular Snip

```
rm: `worcount.jar': No such file or directory
[cloudera@localhost ~]$ hadoop fs -rm wordcount.jar
14/06/17 13:02:30 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 1440 minutes, Emptier interval = 0 minutes.
Moved: 'hdfs://localhost.localdomain:8020/user/cloudera/wordcount.jar' to trash
at: hdfs://localhost.localdomain:8020/user/cloudera/.Trash/Current
[cloudera@localhost ~]$ hadoop fs -rm CS590BD-YLee-Syllabus.pdf

rm: `CS590BD-YLee-Syllabus.pdf': No such file or directory
[cloudera@localhost ~]$
[cloudera@localhost ~]$ clear

[cloudera@localhost ~]$ hadoop fs -ls
Found 7 items
drwx------   - cloudera cloudera          0 2014-06-17 12:50 .Trash
drwx------   - cloudera cloudera          0 2014-06-17 12:52 .staging
-rw-r--r--   1 cloudera cloudera     147075 2014-06-17 13:00 CS990BD-YLee-Syllab
us
-rw-r--r--   1 cloudera cloudera     384207 2014-06-17 07:58 input.txt
-rw-r--r--   1 cloudera cloudera         87 2014-06-17 12:45 ip
drwxr-xr-x   - cloudera cloudera          0 2014-06-17 12:52 out
drwxr-xr-x   - cloudera cloudera          0 2014-06-17 08:09 output
[cloudera@localhost ~]$ hadoop fs -put wordcount.jar wordcount.jar
[cloudera@localhost ~]$ hadoop fs -put CS590BD-YLee-Syllabus.pdf CS590BD-YLee-Sy
llabus.pdf
[cloudera@localhost ~]$
[cloudera@localhost ~]$ hadoop jar wordcount.jar WordCount CS590BD-YLee-Syllabus
.pdf outsyllabus
14/06/17 13:05:09 INFO client.RMProxy: Connecting to ResourceManager at localhos
t.localdomain/127.0.0.1:8032
14/06/17 13:05:11 INFO input.FileInputFormat: Total input paths to process : 1
14/06/17 13:05:12 INFO mapreduce.JobSubmitter: number of splits:1
14/06/17 13:05:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_14
03032233367_0002
14/06/17 13:05:14 INFO impl.YarnClientImpl: Submitted application application_14
03032233367_0002
14/06/17 13:05:14 INFO mapreduce.Job: The url to track the job: http://localhost
.localdomain:8088/proxy/application_1403032233367_0002/
14/06/17 13:05:14 INFO mapreduce.Job: Running job: job_1403032233367_0002
14/06/17 13:05:34 INFO mapreduce.Job: Job job_1403032233367_0002 running in uber
 mode : false
14/06/17 13:05:34 INFO mapreduce.Job:  map 0% reduce 0%
14/06/17 13:05:51 INFO mapreduce.Job:  map 100% reduce 0%
```

Terminal Log
https://github.com/pcn92-jedicode/CS5590BD/blob/master/Lab1/Terminal_Log.txt

Wordcount Output
https://github.com/pcn92-jedicode/CS5590BD/blob/master/Lab1/Syllabus_WordCount.txt