

Assignment 1

I. Installation

- Cloudera provides Apache Hadoop Ecosystem as QuickStarts, Cloudera Manager and Cloudera Director.
- Before selecting any of the packages, we need to make sure we have VMware or Virtual Box installed.
- We selected QuickStarts as our download package. Using QuickStarts, we can start using Cloudera's VM or Docker image in a sandbox environment on our local machine.
- After downloading the QuickStarts installation package, we just need to run it in order to launch the VM instance.
- We can configure the VM as per our requirement; we can increase the memory to be allotted, the network adapter to use etc., in the VMware settings.
- Due to low disk space and just 8 GB of RAM, my VM was slow to boot.
- I increased the memory allotted to the VM from 4 GB to 6 GB and it started working smoothly.

Exercise 1: Hadoop Pi

The first test with Hadoop will be to run an existing Hadoop program, to make sure you can launch the program, monitor progress, and get/put files on the HDFS. The simplest program computes pi in parallel on 5 nodes with 5 samples:

The command used for this was:

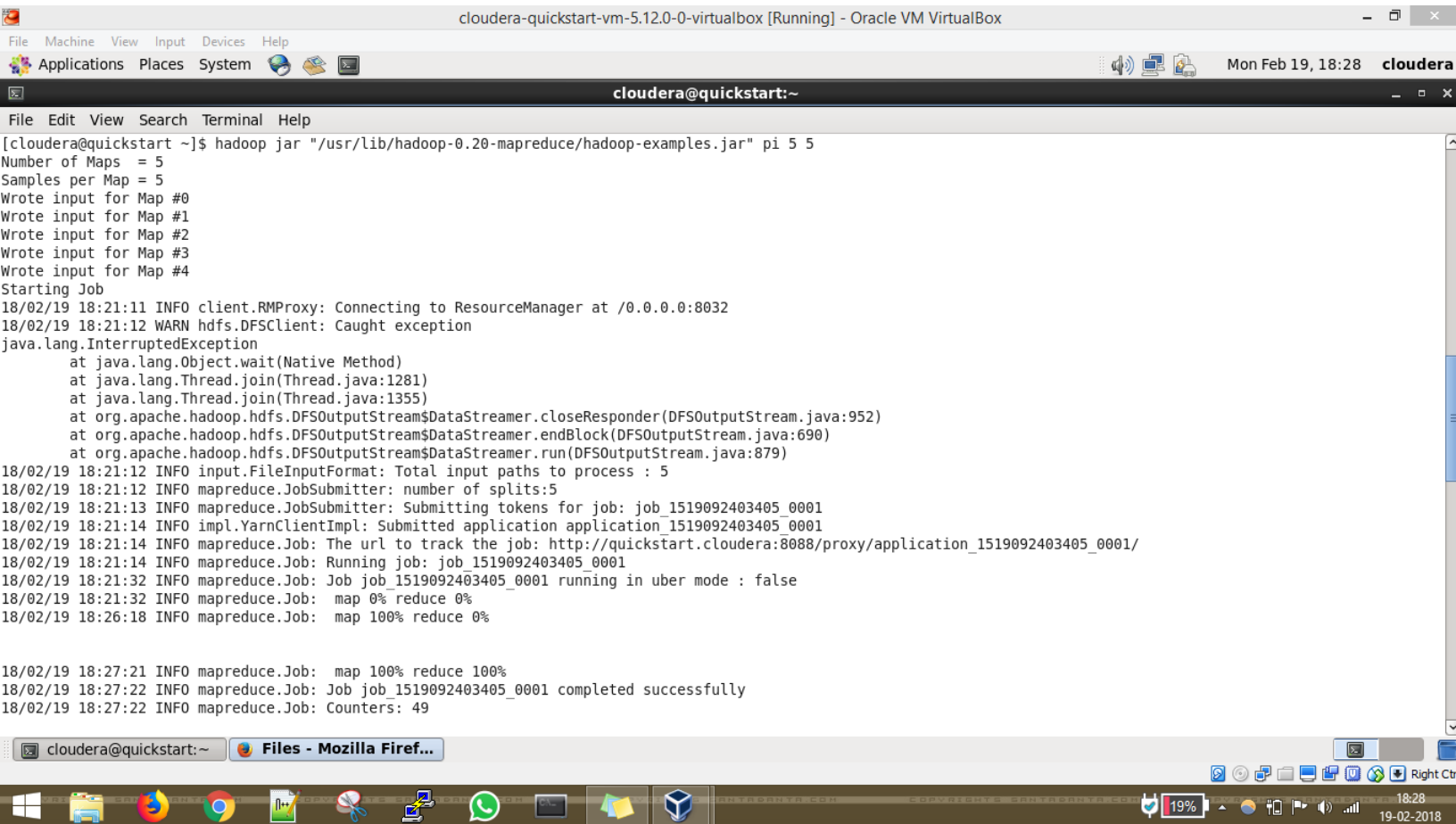
```
$ Hadoop jar "/usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar" pi 5 5
```

Answer 1:

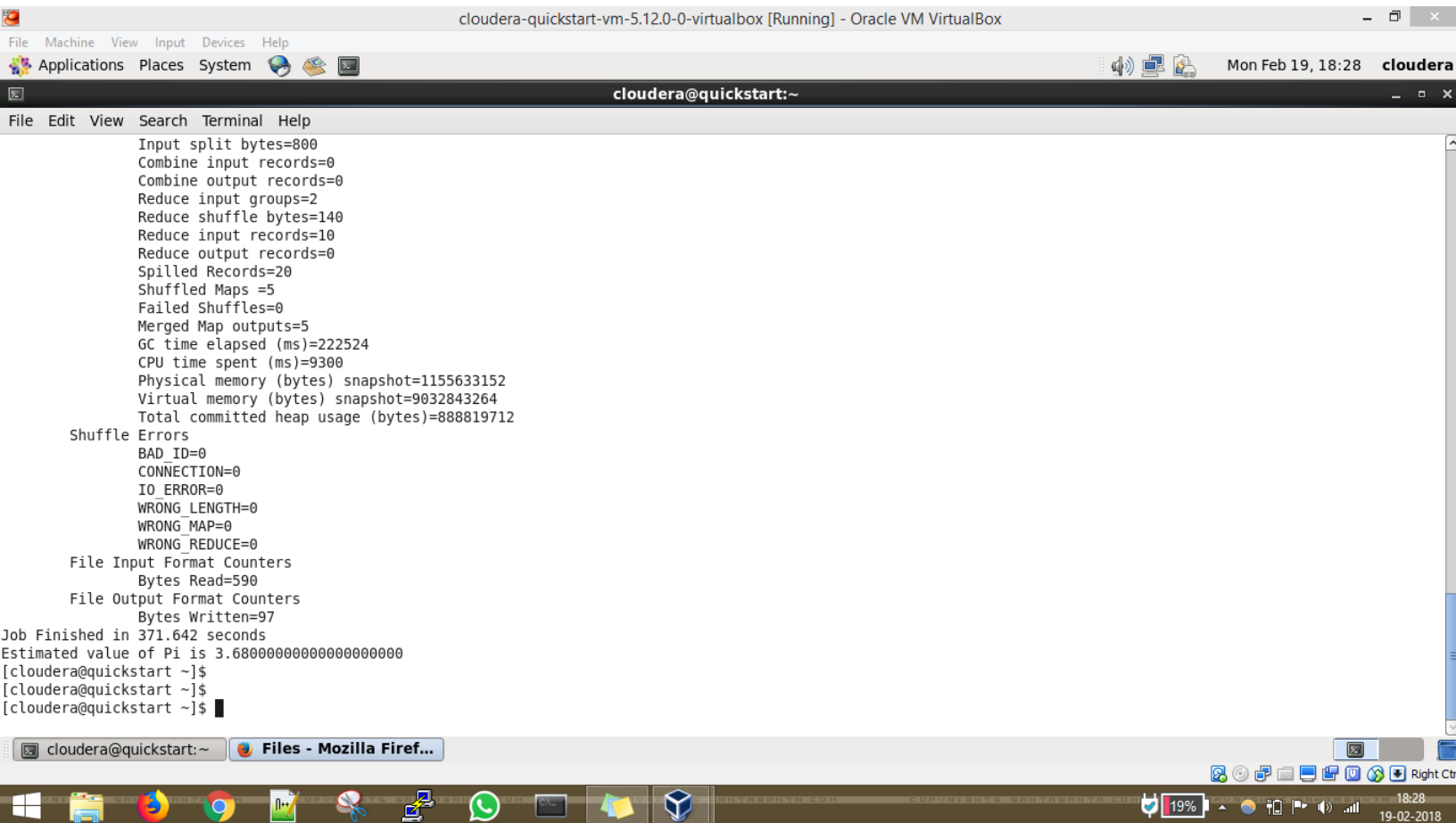
The Output value received was:

3.680000000000000000000000000000

Screenshots:



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop jar "/usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar" pi 5 5
Number of Maps = 5
Samples per Map = 5
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Starting Job
18/02/19 18:21:11 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/19 18:21:12 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
18/02/19 18:21:12 INFO input.FileInputFormat: Total input paths to process : 5
18/02/19 18:21:12 INFO mapreduce.JobSubmitter: number of splits:5
18/02/19 18:21:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519092403405_0001
18/02/19 18:21:14 INFO impl.YarnClientImpl: Submitted application application_1519092403405_0001
18/02/19 18:21:14 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519092403405_0001/
18/02/19 18:21:14 INFO mapreduce.Job: Running job: job_1519092403405_0001
18/02/19 18:21:32 INFO mapreduce.Job: Job job_1519092403405_0001 running in uber mode : false
18/02/19 18:21:32 INFO mapreduce.Job: map 0% reduce 0%
18/02/19 18:26:18 INFO mapreduce.Job: map 100% reduce 0%
18/02/19 18:27:21 INFO mapreduce.Job: map 100% reduce 100%
18/02/19 18:27:22 INFO mapreduce.Job: Job job_1519092403405_0001 completed successfully
18/02/19 18:27:22 INFO mapreduce.Job: Counters: 49
cloudera@quickstart:~
Files - Mozilla Firef...
```

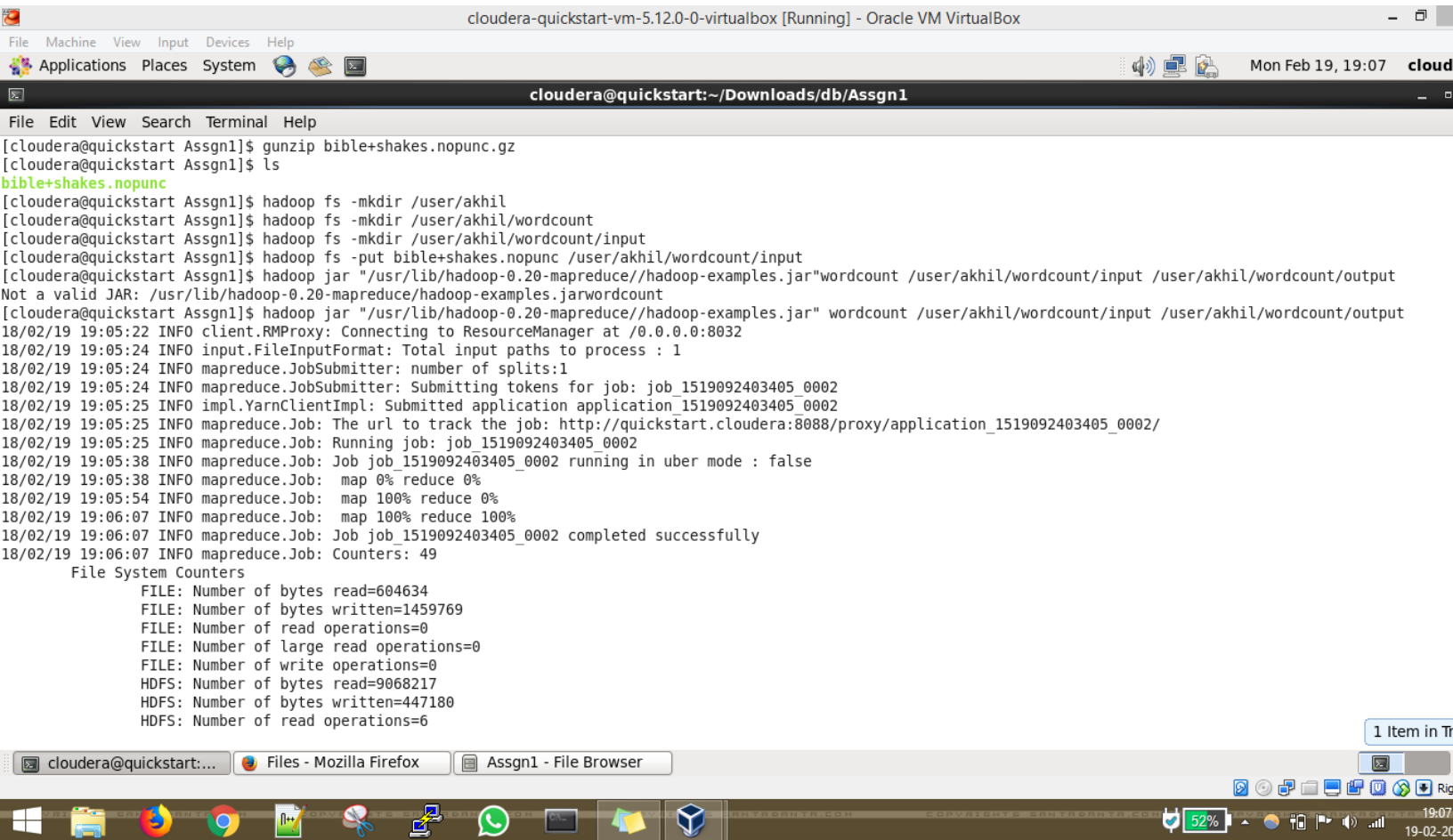


```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
Input split bytes=800
Combine input records=0
Combine output records=0
Reduce input groups=2
Reduce shuffle bytes=140
Reduce input records=10
Reduce output records=0
Spilled Records=20
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=222524
CPU time spent (ms)=9300
Physical memory (bytes) snapshot=1155633152
Virtual memory (bytes) snapshot=9032843264
Total committed heap usage (bytes)=888819712
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=590
File Output Format Counters
Bytes Written=97
Job Finished in 371.642 seconds
Estimated value of Pi is 3.68000000000000000000
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
cloudera@quickstart:~
Files - Mozilla Firef...
```

Exercise 2: Hadoop Word Count:

The next program to test is the Hadoop word count program. This example reads text files and counts how often words occur. The input is text files and the output is text files, each line of which contains a word and the count of how often it occurred, separated by a tab.

Before we can run the example, we'll have to copy some data into the distributed file system (HDFS). Here we will create an input directory, and copy in the complete works of Shakespeare and the bible (a standard large corpus for text mining).



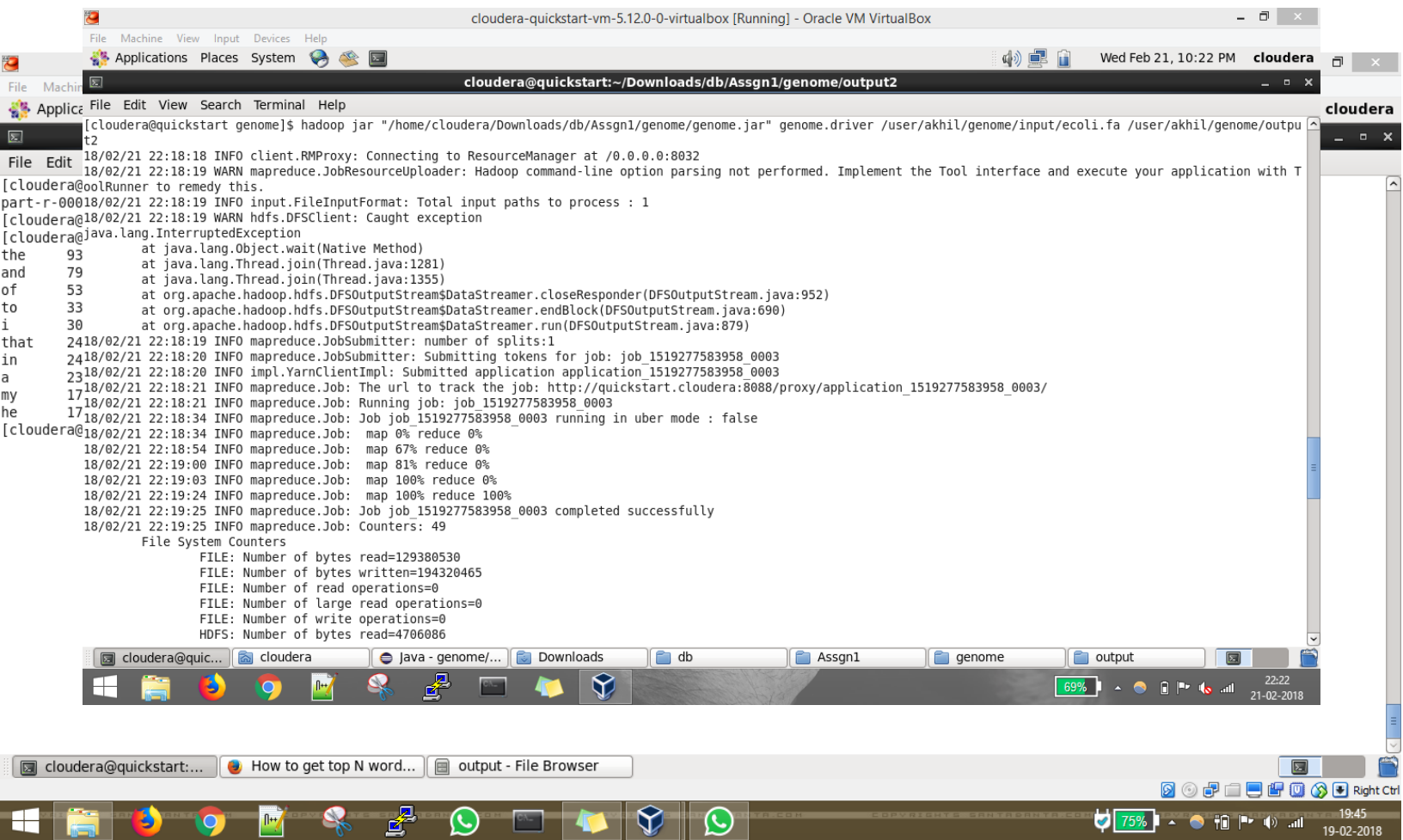
The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal is running commands to set up the Hadoop word count environment. It starts with "gunzip bible+shakes.nopunc.gz" and "ls" to verify the files. Then, it creates directories "user/akhil" and "user/akhil/wordcount", and "user/akhil/wordcount/input". The main command is "hadoop jar /usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount /user/akhil/wordcount/input /user/akhil/wordcount/output". The output shows the job running successfully, with counters indicating 49 files and 604634 bytes read. The bottom of the screen shows the Windows taskbar with various icons and the system clock at 19:07 on 19-02-20.

```
cloudera@quickstart:~/Downloads/db/Assgn1
File Edit View Search Terminal Help
[cloudera@quickstart Assgn1]$ gunzip bible+shakes.nopunc.gz
[cloudera@quickstart Assgn1]$ ls
bible+shakes.nopunc
[cloudera@quickstart Assgn1]$ hadoop fs -mkdir /user/akhil
[cloudera@quickstart Assgn1]$ hadoop fs -mkdir /user/akhil/wordcount
[cloudera@quickstart Assgn1]$ hadoop fs -mkdir /user/akhil/wordcount/input
[cloudera@quickstart Assgn1]$ hadoop fs -put bible+shakes.nopunc /user/akhil/wordcount/input
[cloudera@quickstart Assgn1]$ hadoop jar "/usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar" wordcount /user/akhil/wordcount/input /user/akhil/wordcount/output
Not a valid JAR: /usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jarwordcount
[cloudera@quickstart Assgn1]$ hadoop jar "/usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar" wordcount /user/akhil/wordcount/input /user/akhil/wordcount/output
18/02/19 19:05:22 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/19 19:05:24 INFO input.FileInputFormat: Total input paths to process : 1
18/02/19 19:05:24 INFO mapreduce.JobSubmitter: number of splits:1
18/02/19 19:05:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519092403405_0002
18/02/19 19:05:25 INFO impl.YarnClientImpl: Submitted application application_1519092403405_0002
18/02/19 19:05:25 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519092403405_0002/
18/02/19 19:05:25 INFO mapreduce.Job: Running job: job_1519092403405_0002
18/02/19 19:05:38 INFO mapreduce.Job: Job job_1519092403405_0002 running in uber mode : false
18/02/19 19:05:38 INFO mapreduce.Job: map 0% reduce 0%
18/02/19 19:05:54 INFO mapreduce.Job: map 100% reduce 0%
18/02/19 19:06:07 INFO mapreduce.Job: map 100% reduce 100%
18/02/19 19:06:07 INFO mapreduce.Job: Job job_1519092403405_0002 completed successfully
18/02/19 19:06:07 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=604634
  FILE: Number of bytes written=1459769
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=9068217
  HDFS: Number of bytes written=447180
  HDFS: Number of read operations=6
```

Answer 2:

The top 10 most frequently used words were:

the	93739
and	79182
of	53121
to	33929
i	30240
that	24407
in	24350
a	23504
my	17312
he	17887



Exercise 3: Hadoop KMER counting:

The next exercise was to implement a KMER counter using Hadoop. Conceptually this is very similar to the word count program, but since there are no spaces in the human genome, we counted the overlapping KMERS instead of discrete words.

Output:

Top 10 most frequently occurring 9-mers in E coli:

CCAGCGCCA	252
CAGCGCCAG	247
GCGCTGGCG	234
CGCCAGCAG	220
CCGTAGCGG	219
CGCTGGACC	212
CGCCAGGCC	211
GCGGTCGCA	207
TCCAGCGCG	200
CAGGTCGGC	199

```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Downloads/db/Assgn1/genome/twoBit
File Edit View Search Terminal Help

[cloudera@quickstart output2]$ cd ..
[cloudera@quickstart genome]$ ls
ecoli.fa ecoli.fa~ genome.jar output output2 twoBit
[cloudera@quickstart genome]$ cd twoBit/
[cloudera@quickstart twoBit]$ ls
hg19.2bit hg19.fa twoBitToFa
[cloudera@quickstart twoBit]$ hadoop fs -put hg19.fa /user/akhil/genome/input
[cloudera@quickstart twoBit]$ hadoop jar "/home/cloudera/Downloads/db/Assgn1/genome/genome.jar" genome.driver /user/akhil/genome/input/hg19.fa /user/akhil/genome/output
3
18/02/21 23:03:58 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/21 23:03:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/02/21 23:03:59 INFO input.FileInputFormat: Total input paths to process : 1
18/02/21 23:04:00 INFO mapreduce.JobSubmitter: number of splits:24
18/02/21 23:04:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519277583958_0004
18/02/21 23:04:01 INFO impl.YarnClientImpl: Submitted application application_1519277583958_0004
18/02/21 23:04:01 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519277583958_0004/
18/02/21 23:04:01 INFO mapreduce.Job: Running job: job_1519277583958_0004
18/02/21 23:04:15 INFO mapreduce.Job: Job job_1519277583958_0004 running in uber mode : false
18/02/21 23:04:15 INFO mapreduce.Job: map 0% reduce 0%
18/02/21 23:05:31 INFO mapreduce.Job: map 1% reduce 0%
18/02/21 23:08:02 INFO mapreduce.Job: map 2% reduce 0%
18/02/21 23:11:31 INFO mapreduce.Job: map 3% reduce 0%

cloudera@quickst... cloudera java - genome/src... Downloads db Assgn1 genome
```

```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Downloads/db/Assgn1/genome/output2
File Edit View Search Terminal Help

Spilled Records=12129423
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=389
CPU time spent (ms)=26940
Physical memory (bytes) snapshot=464912384
Virtual memory (bytes) snapshot=3019251712
Total committed heap usage (bytes)=355282944

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=4705957
File Output Format Counters
Bytes Written=3220369

[cloudera@quickstart genome]$ hadoop fs -get /user/akhil/genome/output2 ~/Downloads/db/Assgn1/genome/output2
[cloudera@quickstart genome]$ cd output2/
[cloudera@quickstart output2]$ cat part-r-00000 | sort -k2 -n -r | head -n10
CCAGCGCCA 252
CAGCGCCAG 247
GCGCTGGCG 234
CGCCAGCAG 220
CGCTGGCGG 219
CTGGCGCTG 212
CGCCAGCGC 211
GCCAGCGCC 207
TGGCGCTGG 200
CCGCCAGCA 199
[cloudera@quickstart output2]$

Assgn1
cloudera@quic... cloudera java - genome/... Downloads db Assgn1 genome output
```


Exercise 4: Hadoop Playing Cards Counting

Use the same Hadoop/Mapreduce framework to write your own mapper and reducer codes in Java to count the numeric number for each suit of playing cards. (as the demo in the class)

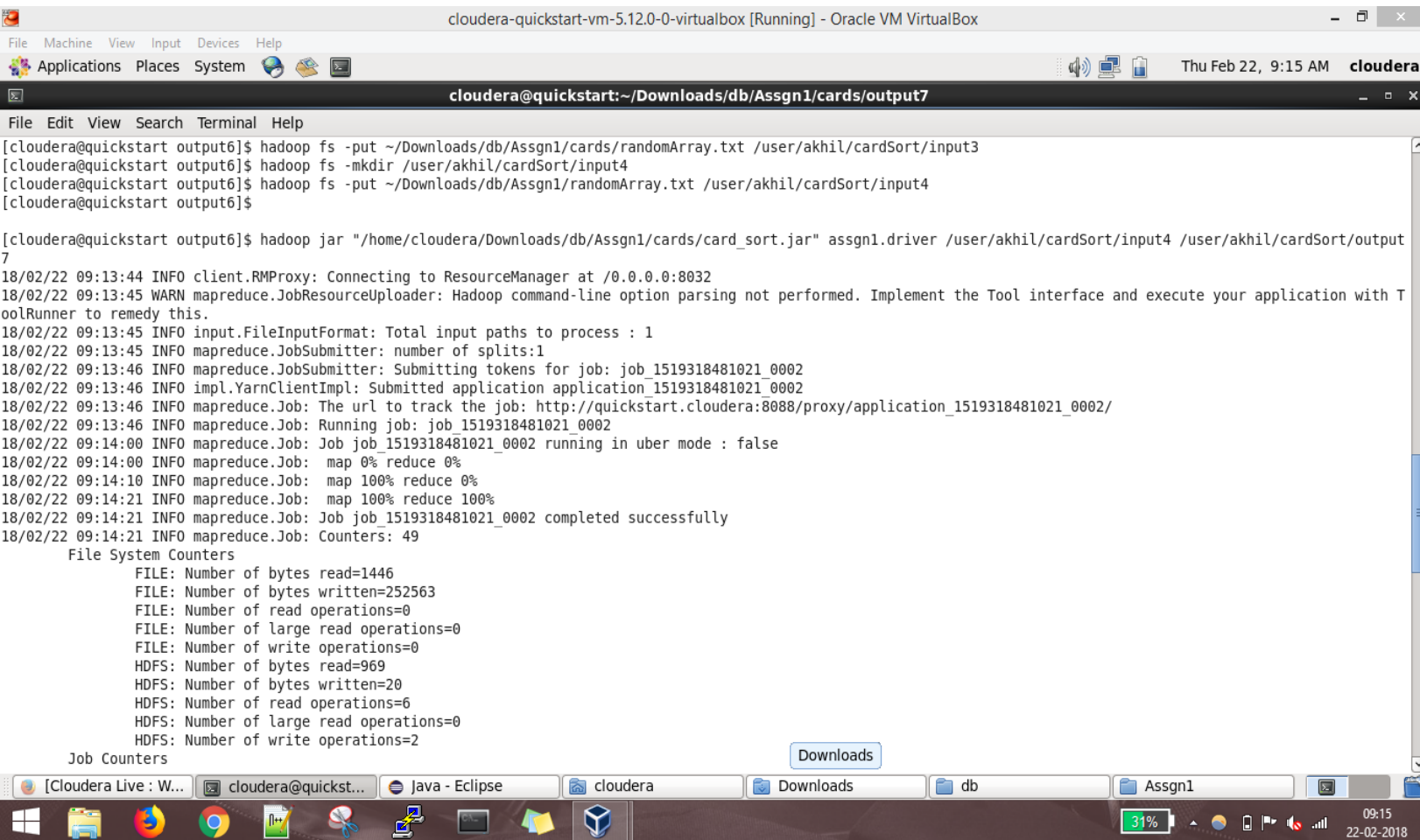
For data file, we needed to write java codes to generate an input file containing shuffled 100 decks of 54 cards.

NOTE: *I was confused about whether we have to count the total cards in a suit or we have to sum it, so I wrote program for both, addition and counting. I've attached screenshots for both.*

For a deck of 5:

First we generated an input file containing 5 shuffled decks of cards.

We will be ignoring all the face cards and the joker cards.



The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal displays the following commands and output:

```
[cloudera@quickstart output6]$ hadoop fs -put ~/Downloads/db/Assgn1/cards/randomArray.txt /user/akhil/cardSort/input3
[cloudera@quickstart output6]$ hadoop fs -mkdir /user/akhil/cardSort/input4
[cloudera@quickstart output6]$ hadoop fs -put ~/Downloads/db/Assgn1/randomArray.txt /user/akhil/cardSort/input4
[cloudera@quickstart output6]$

[cloudera@quickstart output6]$ hadoop jar "/home/cloudera/Downloads/db/Assgn1/cards/card_sort.jar" assgn1.driver /user/akhil/cardSort/input4 /user/akhil/cardSort/output7

18/02/22 09:13:44 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 09:13:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/02/22 09:13:45 INFO input.FileInputFormat: Total input paths to process : 1
18/02/22 09:13:45 INFO mapreduce.JobSubmitter: number of splits:1
18/02/22 09:13:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519318481021_0002
18/02/22 09:13:46 INFO impl.YarnClientImpl: Submitted application application_1519318481021_0002
18/02/22 09:13:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519318481021_0002/
18/02/22 09:13:46 INFO mapreduce.Job: Running job: job_1519318481021_0002
18/02/22 09:14:00 INFO mapreduce.Job: Job job_1519318481021_0002 running in uber mode : false
18/02/22 09:14:00 INFO mapreduce.Job: map 0% reduce 0%
18/02/22 09:14:10 INFO mapreduce.Job: map 100% reduce 0%
18/02/22 09:14:21 INFO mapreduce.Job: map 100% reduce 100%
18/02/22 09:14:21 INFO mapreduce.Job: Job job_1519318481021_0002 completed successfully
18/02/22 09:14:21 INFO mapreduce.Job: Counters: 49

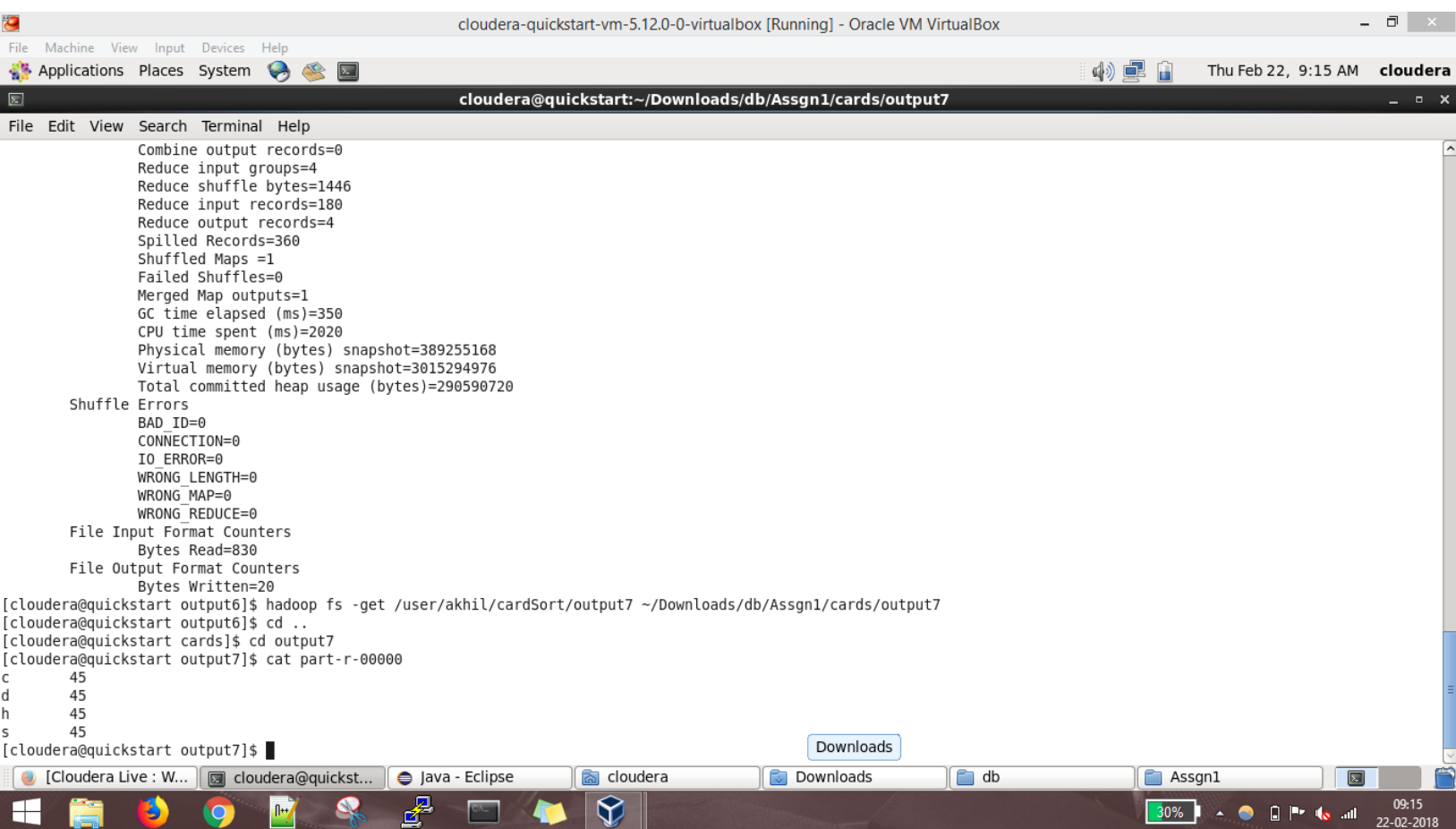
File System Counters
  FILE: Number of bytes read=1446
  FILE: Number of bytes written=252563
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=969
  HDFS: Number of bytes written=20
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2

Job Counters
```

Output for 5 decks of cards:

For counting total number of numeric cards in a suit.

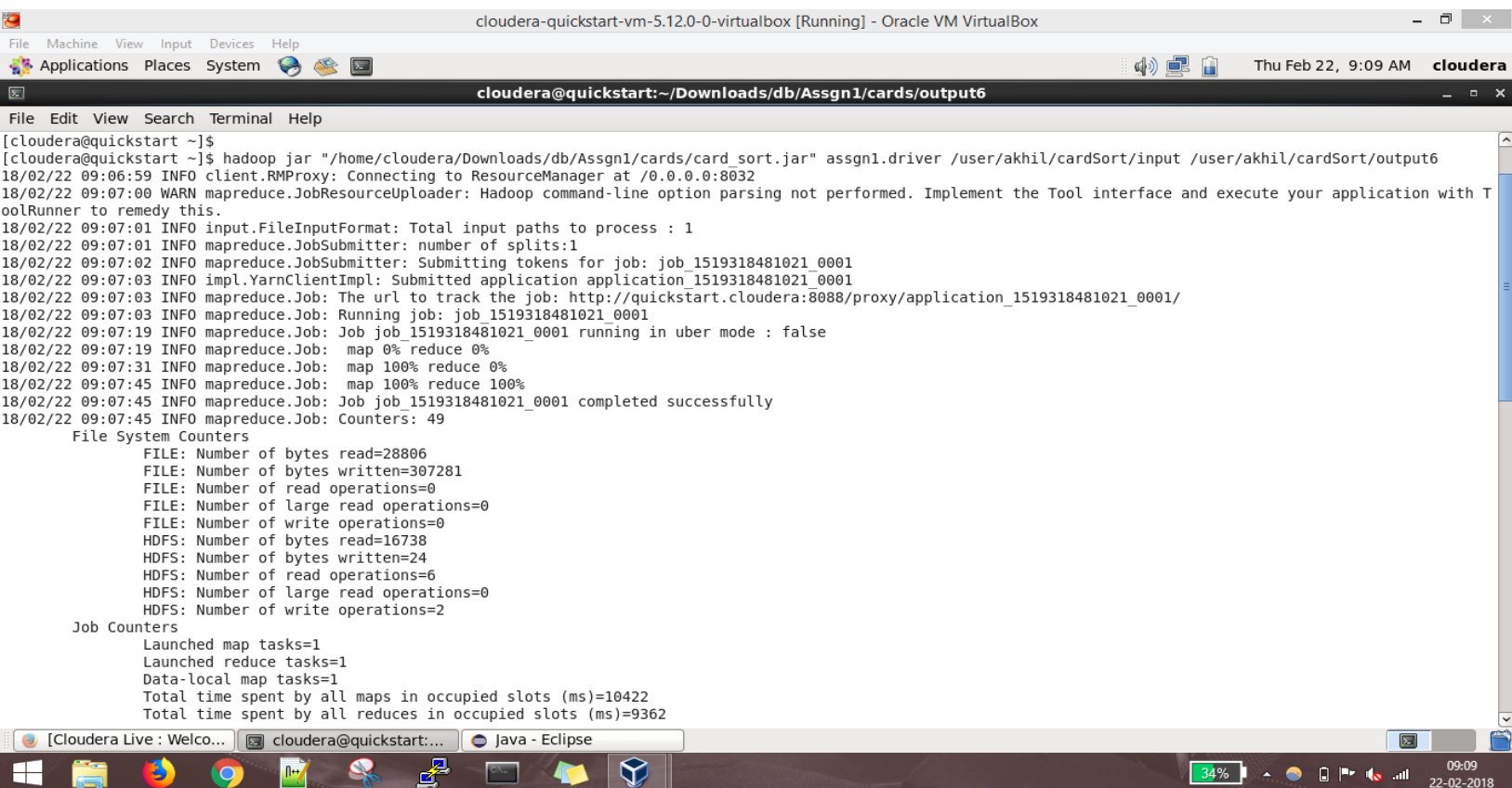
c	45	(clubs)
d	45	(diamonds)
h	45	(hearts)
s	45	(spades)



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart: ~/Downloads/db/Assgn1/cards/output7
File Edit View Search Terminal Help
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=1446
Reduce input records=180
Reduce output records=4
Spilled Records=360
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=350
CPU time spent (ms)=2020
Physical memory (bytes) snapshot=389255168
Virtual memory (bytes) snapshot=3015294976
Total committed heap usage (bytes)=290590720
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=830
File Output Format Counters
Bytes Written=20
[cloudera@quickstart output6]$ hadoop fs -get /user/akhil/cardSort/output7 ~/Downloads/db/Assgn1/cards/output7
[cloudera@quickstart output6]$ cd ..
[cloudera@quickstart cards]$ cd output7
[cloudera@quickstart output7]$ cat part-r-00000
c
45
d
45
h
45
s
45
[cloudera@quickstart output7]$
```

For Deck of 100 cards:

For counting total number of numeric cards in a suit.

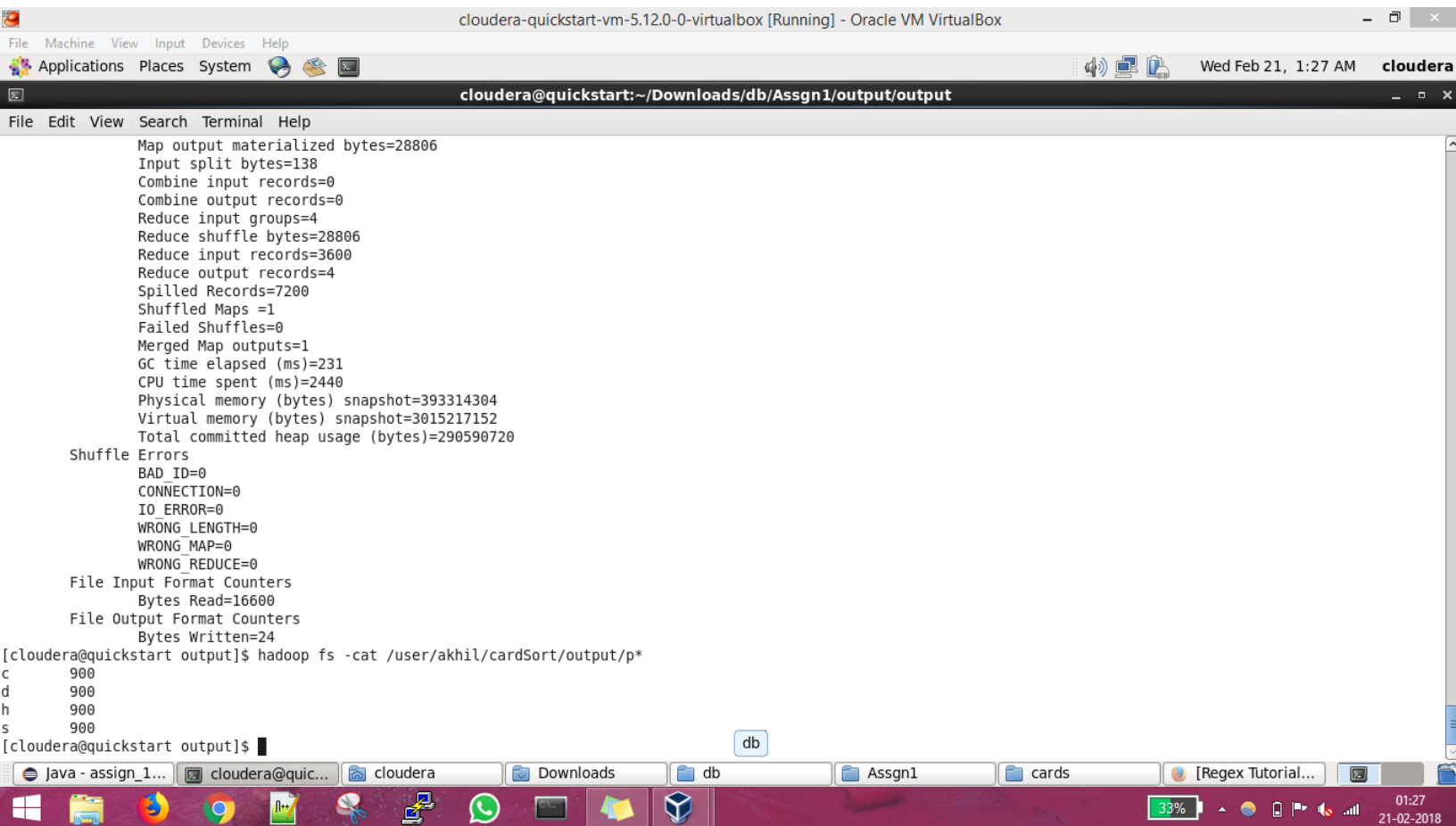


```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart: ~/Downloads/db/Assgn1/cards/output6
File Edit View Search Terminal Help
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop jar "/home/cloudera/Downloads/db/Assgn1/cards/card_sort.jar" assgn1.driver /user/akhil/cardSort/input /user/akhil/cardSort/output6
18/02/22 09:06:59 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 09:07:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/02/22 09:07:01 INFO input.FileInputFormat: Total input paths to process : 1
18/02/22 09:07:01 INFO mapreduce.JobSubmitter: number of splits:1
18/02/22 09:07:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519318481021_0001
18/02/22 09:07:03 INFO impl.YarnClientImpl: Submitted application application_1519318481021_0001
18/02/22 09:07:03 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519318481021_0001/
18/02/22 09:07:03 INFO mapreduce.Job: Running job: job_1519318481021_0001
18/02/22 09:07:19 INFO mapreduce.Job: Job job_1519318481021_0001 running in uber mode : false
18/02/22 09:07:19 INFO mapreduce.Job: map 0% reduce 0%
18/02/22 09:07:31 INFO mapreduce.Job: map 100% reduce 0%
18/02/22 09:07:45 INFO mapreduce.Job: map 100% reduce 100%
18/02/22 09:07:45 INFO mapreduce.Job: Job job_1519318481021_0001 completed successfully
18/02/22 09:07:45 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=28806
FILE: Number of bytes written=307281
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=16738
HDFS: Number of bytes written=24
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)=10422
Total time spent by all reduces in occupied slots (ms)=9362
```

Output for 100 decks of cards:

For counting total number of numeric cards in a suit.

c 900 (clubs)
d 900 (diamonds)
h 900 (hearts)
s 900 (spades)



The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal displays the output of a Hadoop MapReduce job. The output includes various metrics such as "Map output materialized bytes=28806", "Input split bytes=138", "Combine input records=0", "Combine output records=0", "Reduce input groups=4", "Reduce shuffle bytes=28806", "Reduce input records=3600", "Reduce output records=4", "Spilled Records=7200", "Shuffled Maps =1", "Failed Shuffles=0", "Merged Map outputs=1", "GC time elapsed (ms)=231", "CPU time spent (ms)=2440", "Physical memory (bytes) snapshot=393314304", "Virtual memory (bytes) snapshot=3015217152", and "Total committed heap usage (bytes)=290590720". Below these metrics, the "Shuffle Errors" section shows "BAD_ID=0", "CONNECTION=0", "IO_ERROR=0", "WRONG_LENGTH=0", "WRONG_MAP=0", and "WRONG_REDUCE=0". The "File Input Format Counters" section shows "Bytes Read=16600", and the "File Output Format Counters" section shows "Bytes Written=24". The terminal then shows the command "[cloudera@quickstart output]\$ hadoop fs -cat /user/akhil/cardSort/output/p*" and its output, which lists the counts for each suit: "c 900", "d 900", "h 900", and "s 900". The terminal prompt is "[cloudera@quickstart output]\$". The taskbar at the bottom shows various applications, including Java, cloudera, Downloads, db, Assgn1, cards, and [Regex Tutorial...]. The system tray shows the date and time as "01:27 21-02-2018".

```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Downloads/db/Assgn1/output/output
File Edit View Search Terminal Help
Map output materialized bytes=28806
Input split bytes=138
Combine input records=0
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=28806
Reduce input records=3600
Reduce output records=4
Spilled Records=7200
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=231
CPU time spent (ms)=2440
Physical memory (bytes) snapshot=393314304
Virtual memory (bytes) snapshot=3015217152
Total committed heap usage (bytes)=290590720
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=16600
File Output Format Counters
Bytes Written=24
[cloudera@quickstart output]$ hadoop fs -cat /user/akhil/cardSort/output/p*
c 900
d 900
h 900
s 900
[cloudera@quickstart output]$
```


For addition of total values of numeric cards in a suit:

c 5400 (clubs)
d 5400 (diamonds)
h 5400 (hearts)
s 5400 (spades)

The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal displays the output of a Hadoop MapReduce job, including statistics like "Map output materialized bytes=28806", "Input split bytes=138", and "Total committed heap usage (bytes)=290590720". It also shows "Shuffle Errors" and "File Input Format Counters". Below the job output, the user runs the command "hadoop fs -cat /user/akhil/cardSort/output11/p*", which lists the values "c 5400", "d 5400", "h 5400", and "s 5400". The terminal window is part of a desktop environment with a taskbar at the bottom showing various applications and system icons.

```
cloudera@quickstart:~/Downloads/db/Assign1/cards/output7
Map output materialized bytes=28806
Input split bytes=138
Combine input records=0
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=28806
Reduce input records=3600
Reduce output records=4
Spilled Records=7200
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=248
CPU time spent (ms)=2740
Physical memory (bytes) snapshot=391278592
Virtual memory (bytes) snapshot=3015290880
Total committed heap usage (bytes)=290590720
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=16600
File Output Format Counters
Bytes Written=28
[cloudera@quickstart output7]$ hadoop fs -cat /user/akhil/cardSort/output11/p*
c 5400
d 5400
h 5400
s 5400
[cloudera@quickstart output7]$
```

Lessons learnt:

- How to use Hadoop file system.
- The purpose of MapReduce program.
- How MapReduce eases mining of massive data sets.
- To write programs or applications in eclipse for using Hadoop MapReduce and how to modify the program as per our requirement.
- Learnt some advance Linux commands to use with HDFS.
- Executing java applications and providing input to get the expected results.