HW 3: Hadoop using Docker

HW 3B:Wordcount using Hadoop inside Docker container

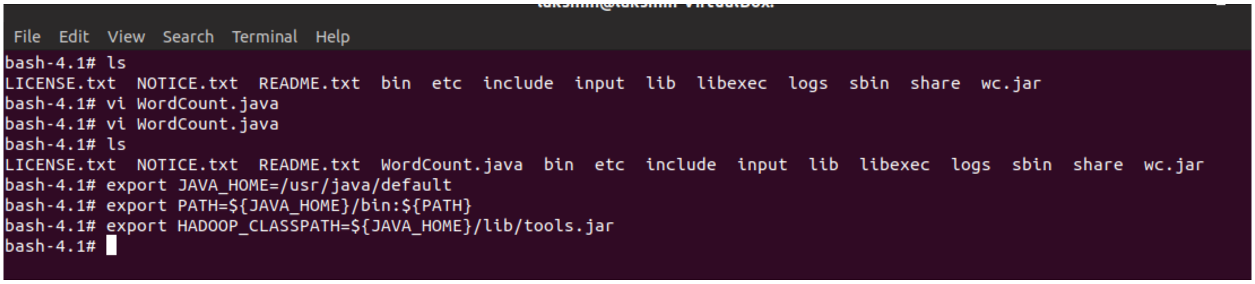
Steps:

1. Go to the Hadoop local directory and check list of files. Create java file for WordCount program by copying thee program from the link provided in the question and execute the following steps.

   bash-4.1# export JAVA_HOME=/usr/java/default

   bash-4.1#export PATH=${JAVA_HOME}/bin:${PATH}

   bash-4.1# export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar



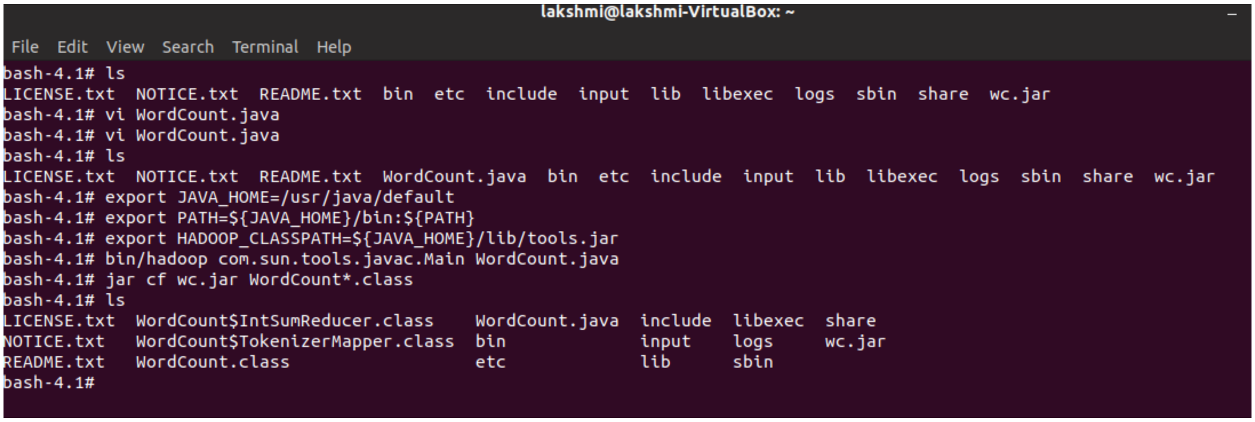
```
bash-4.1# ls
LICENSE.txt  NOTICE.txt  README.txt  bin  etc  include  input  lib  libexec  logs  sbin  share  wc.jar
bash-4.1# vi WordCount.java
bash-4.1# vi WordCount.java
bash-4.1# ls
LICENSE.txt  NOTICE.txt  README.txt  WordCount.java  bin  etc  include  input  lib  libexec  logs  sbin  share  wc.jar
bash-4.1# export JAVA_HOME=/usr/java/default
bash-4.1# export PATH=${JAVA_HOME}/bin:${PATH}
bash-4.1# export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
bash-4.1# 
```

2. Compile java file and create a jar file.

   bash-4.1# bin/hadoop com.sun.tools.javac.Main WordCount.java

   bash-4.1# jar cf wc.jar WordCount*.class

   bash-4.1# ls


lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
bash-4.1# ls
LICENSE.txt  NOTICE.txt  README.txt  bin  etc  include  input  lib  libexec  logs  sbin  share  wc.jar
bash-4.1# vi WordCount.java
bash-4.1# vi WordCount.java
bash-4.1# ls
LICENSE.txt  NOTICE.txt  README.txt  WordCount.java  bin  etc  include  input  lib  libexec  logs  sbin  share  wc.jar
bash-4.1# export JAVA_HOME=/usr/java/default
bash-4.1# export PATH=${JAVA_HOME}/bin:${PATH}
bash-4.1# export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
bash-4.1# bin/hadoop com.sun.tools.javac.Main WordCount.java
bash-4.1# jar cf wc.jar WordCount*.class
bash-4.1# ls
LICENSE.txt  WordCount$IntSumReducer.class    WordCount.java  include  libexec  share
NOTICE.txt   WordCount$TokenizerMapper.class  bin             input    logs     wc.jar
README.txt   WordCount.class                  etc             lib      sbin
bash-4.1#
```

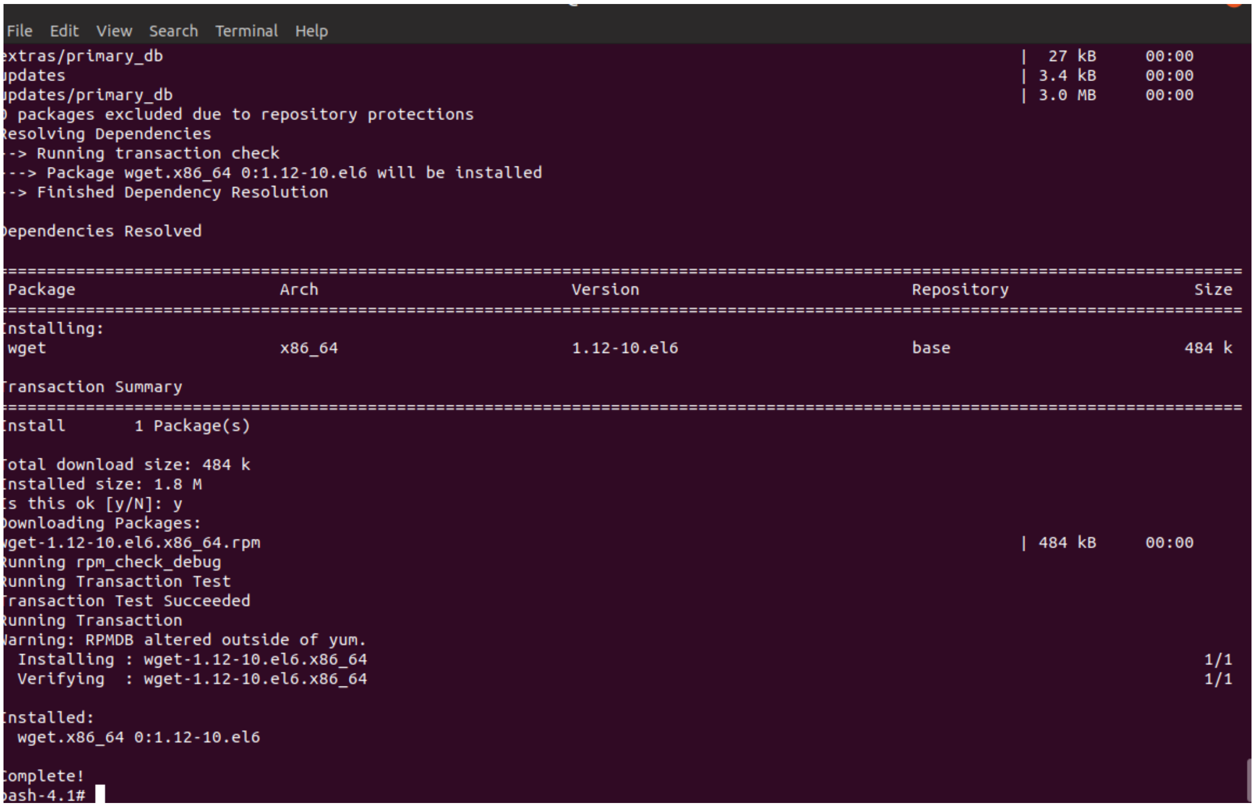
3. Take the input files from web and move them to input files

bash-4.1# sudo yum install wget

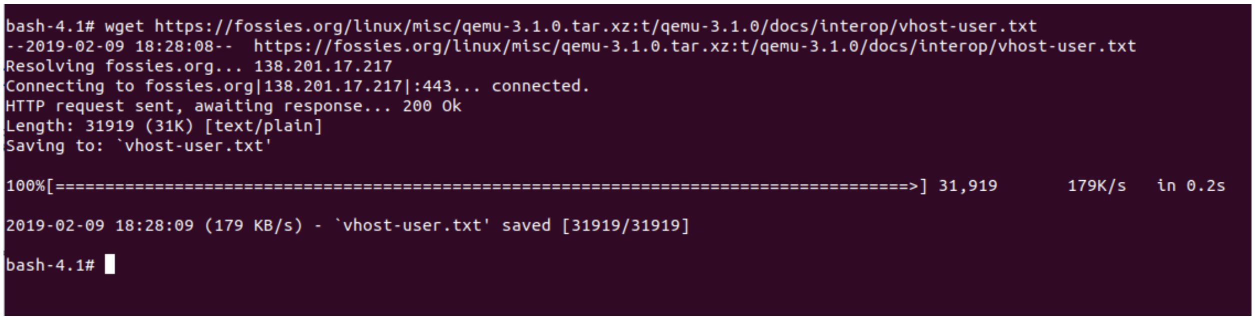
bash-4.1# wget https://fossies.org/linux/misc/qemu-3.1.0.tar.xz:t/qemu-3.1.0/docs/interop/vhost-user.txt

bash-4.1# ls

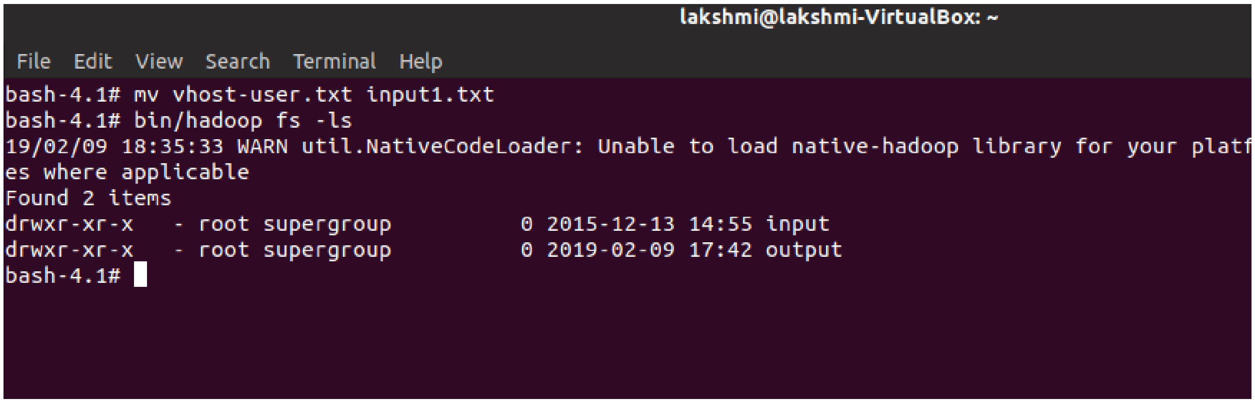
bash-4.1# mv vhost-user.txt input1.txt



```
xtras/primary_db                                                                                          |  27 kB     00:00
pdates                                                                                                    | 3.4 kB     00:00
pdates/primary_db                                                                                         | 3.0 MB     00:00
 packages excluded due to repository protections
esolving Dependencies
-> Running transaction check
--> Package wget.x86_64 0:1.12-10.el6 will be installed
-> Finished Dependency Resolution

ependencies Resolved

================================================================================================================================
 Package                     Arch                          Version                            Repository                   Size
================================================================================================================================
nstalling:
 wget                        x86_64                        1.12-10.el6                        base                        484 k

ransaction Summary
================================================================================================================================
nstall       1 Package(s)

otal download size: 484 k
nstalled size: 1.8 M
s this ok [y/N]: y
ownloading Packages:
get-1.12-10.el6.x86_64.rpm                                                                                | 484 kB     00:00
unning rpm_check_debug
unning Transaction Test
ransaction Test Succeeded
unning Transaction
arning: RPMDB altered outside of yum.
  Installing : wget-1.12-10.el6.x86_64                                                                                      1/1
  Verifying  : wget-1.12-10.el6.x86_64                                                                                      1/1

nstalled:
  wget.x86_64 0:1.12-10.el6

omplete!
ash-4.1#
```



```
bash-4.1# wget https://fossies.org/linux/misc/qemu-3.1.0.tar.xz:t/qemu-3.1.0/docs/interop/vhost-user.txt
--2019-02-09 18:28:08--  https://fossies.org/linux/misc/qemu-3.1.0.tar.xz:t/qemu-3.1.0/docs/interop/vhost-user.txt
Resolving fossies.org... 138.201.17.217
Connecting to fossies.org|138.201.17.217|:443... connected.
HTTP request sent, awaiting response... 200 Ok
Length: 31919 (31K) [text/plain]
Saving to: `vhost-user.txt'

100%[=====================================================================================>] 31,919       179K/s   in 0.2s

2019-02-09 18:28:09 (179 KB/s) - `vhost-user.txt' saved [31919/31919]

bash-4.1#
```

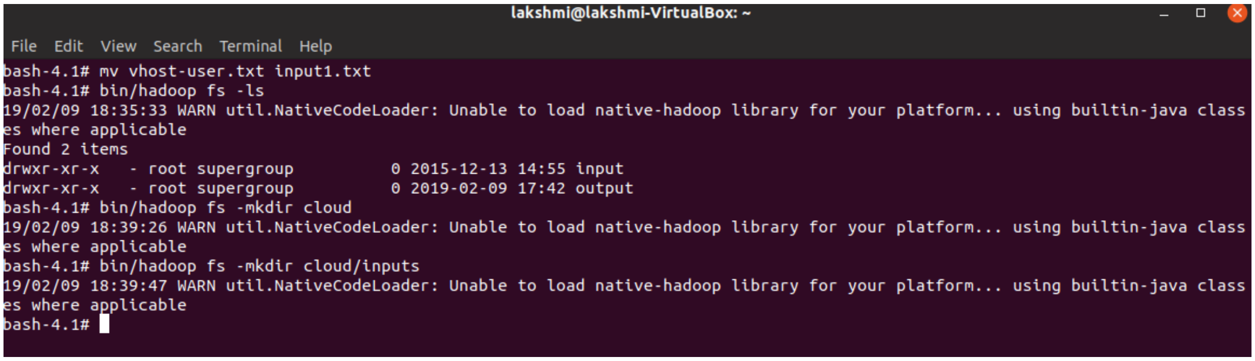

lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
bash-4.1# mv vhost-user.txt input1.txt
bash-4.1# bin/hadoop fs -ls
19/02/09 18:35:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platf
es where applicable
Found 2 items
drwxr-xr-x   - root supergroup          0 2015-12-13 14:55 input
drwxr-xr-x   - root supergroup          0 2019-02-09 17:42 output
bash-4.1# 
```

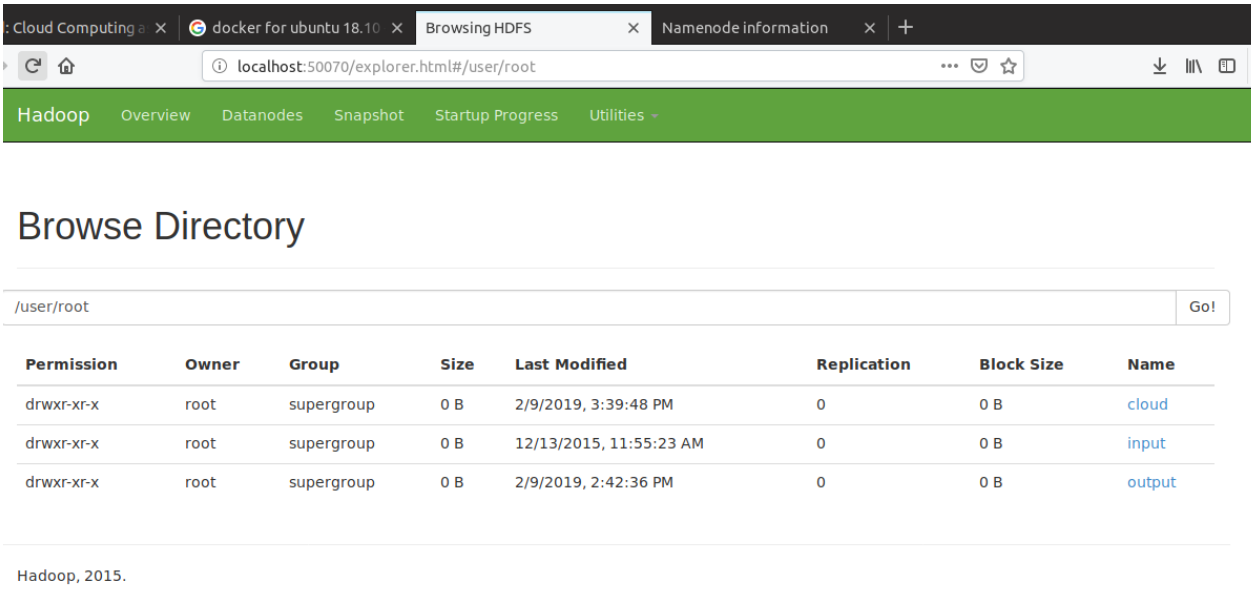
4. Go to Hadoop file system, create our own directory and folders for input files

bash-4.1# bin/hadoop fs -mkdir cloud

bash-4.1# bin/hadoop fs -mkdir cloud/inputs


lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
bash-4.1# mv vhost-user.txt input1.txt
bash-4.1# bin/hadoop fs -ls
19/02/09 18:35:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 2 items
drwxr-xr-x   - root supergroup          0 2015-12-13 14:55 input
drwxr-xr-x   - root supergroup          0 2019-02-09 17:42 output
bash-4.1# bin/hadoop fs -mkdir cloud
19/02/09 18:39:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1# bin/hadoop fs -mkdir cloud/inputs
19/02/09 18:39:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1# 
```



Browse Directory

/user/root Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	2/9/2019, 3:39:48 PM	0	0 B	cloud
drwxr-xr-x	root	supergroup	0 B	12/13/2015, 11:55:23 AM	0	0 B	input
drwxr-xr-x	root	supergroup	0 B	2/9/2019, 2:42:36 PM	0	0 B	output

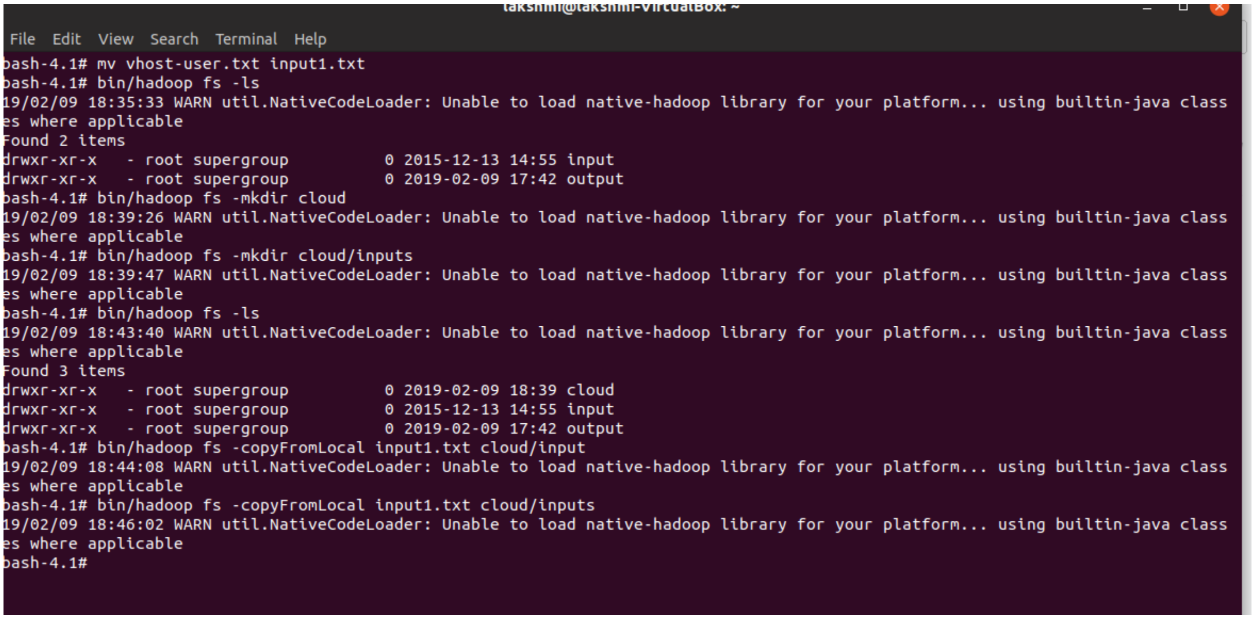
Hadoop, 2015.

5. Copy the input files to newly created directory and folder from local Hadoop directory

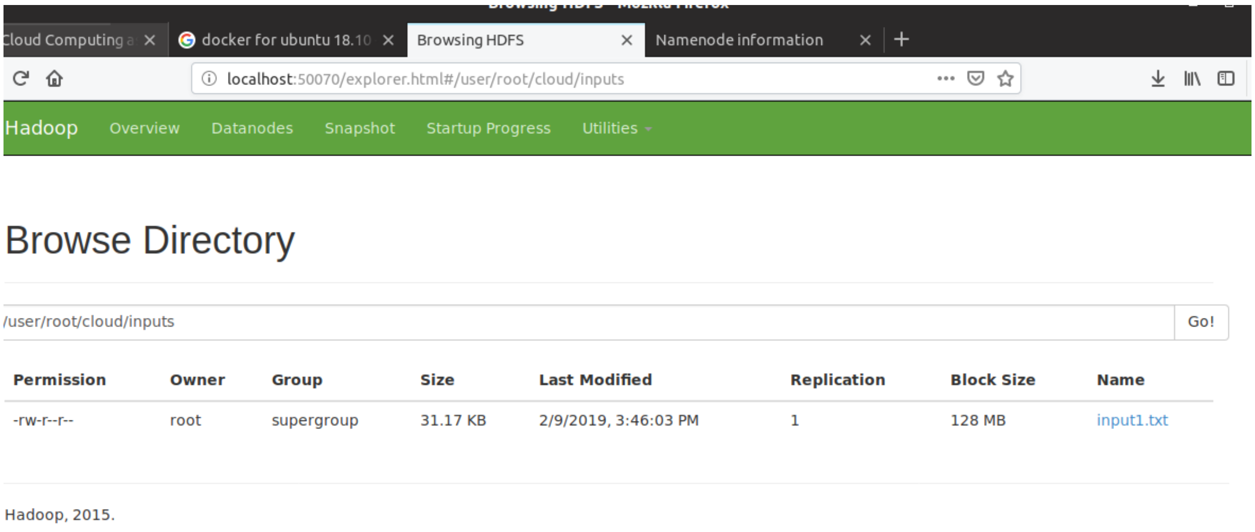
bash-4.1# bin/hadoop fs -ls

bash-4.1# bin/hadoop fs -copyFromLocal input1.txt cloud/input

bash-4.1# bin/hadoop fs -copyFromLocal input1.txt cloud/inputs


lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
bash-4.1# mv vhost-user.txt input1.txt
bash-4.1# bin/hadoop fs -ls
19/02/09 18:35:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 2 items
drwxr-xr-x   - root supergroup          0 2015-12-13 14:55 input
drwxr-xr-x   - root supergroup          0 2019-02-09 17:42 output
bash-4.1# bin/hadoop fs -mkdir cloud
19/02/09 18:39:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1# bin/hadoop fs -mkdir cloud/inputs
19/02/09 18:39:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1# bin/hadoop fs -ls
19/02/09 18:43:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 3 items
drwxr-xr-x   - root supergroup          0 2019-02-09 18:39 cloud
drwxr-xr-x   - root supergroup          0 2015-12-13 14:55 input
drwxr-xr-x   - root supergroup          0 2019-02-09 17:42 output
bash-4.1# bin/hadoop fs -copyFromLocal input1.txt cloud/input
19/02/09 18:44:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1# bin/hadoop fs -copyFromLocal input1.txt cloud/inputs
19/02/09 18:46:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
bash-4.1#
```



localhost:50070/explorer.html#/user/root/cloud/inputs

Browse Directory

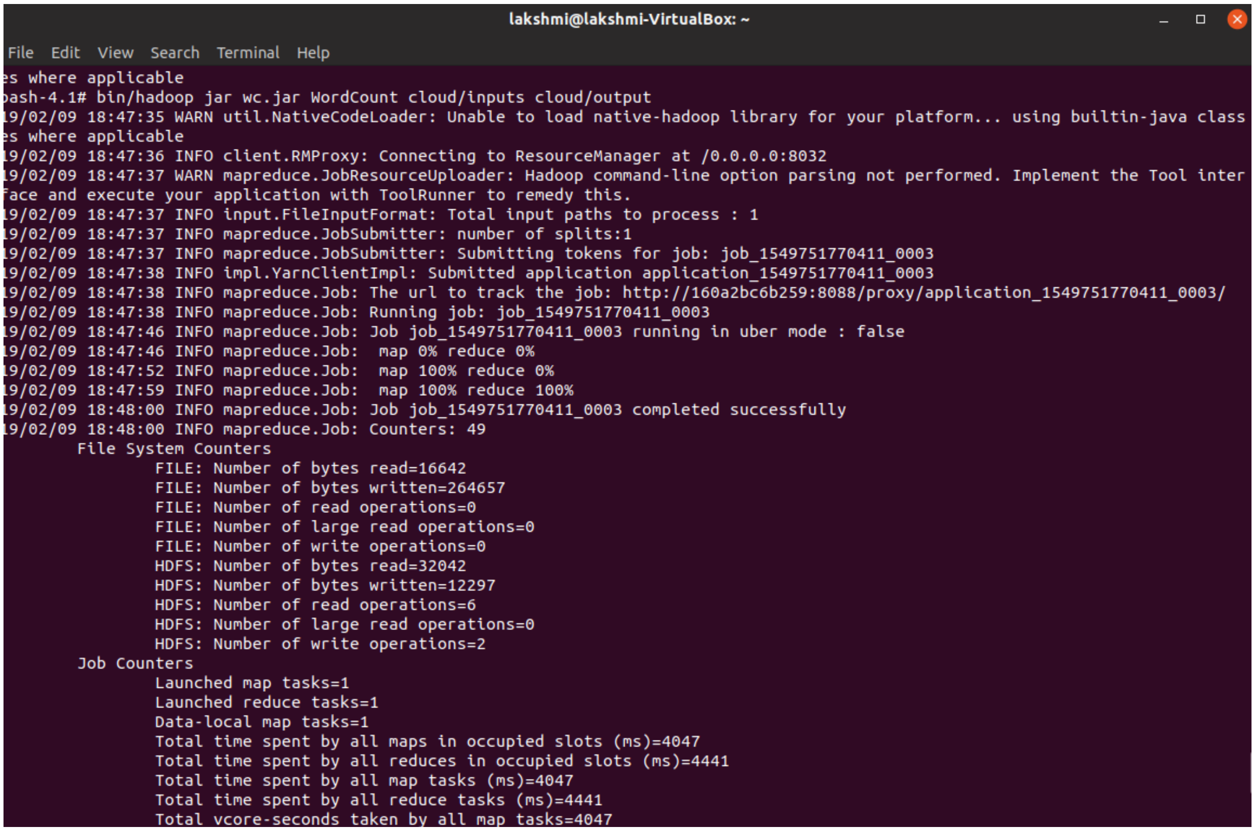
/user/root/cloud/inputs

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	31.17 KB	2/9/2019, 3:46:03 PM	1	128 MB	input1.txt

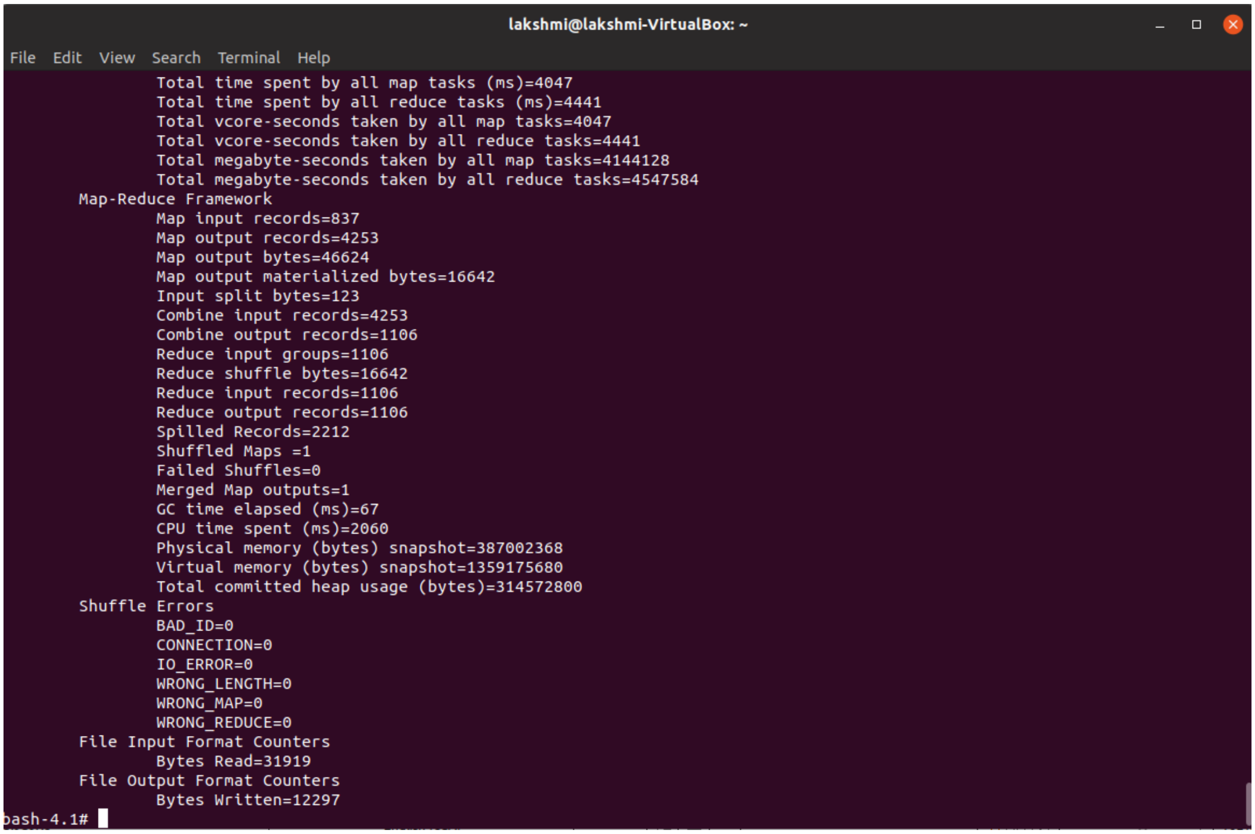
Hadoop, 2015.

6. Run the WordCount.java program taking input from input files. Also, save the output files to output folder

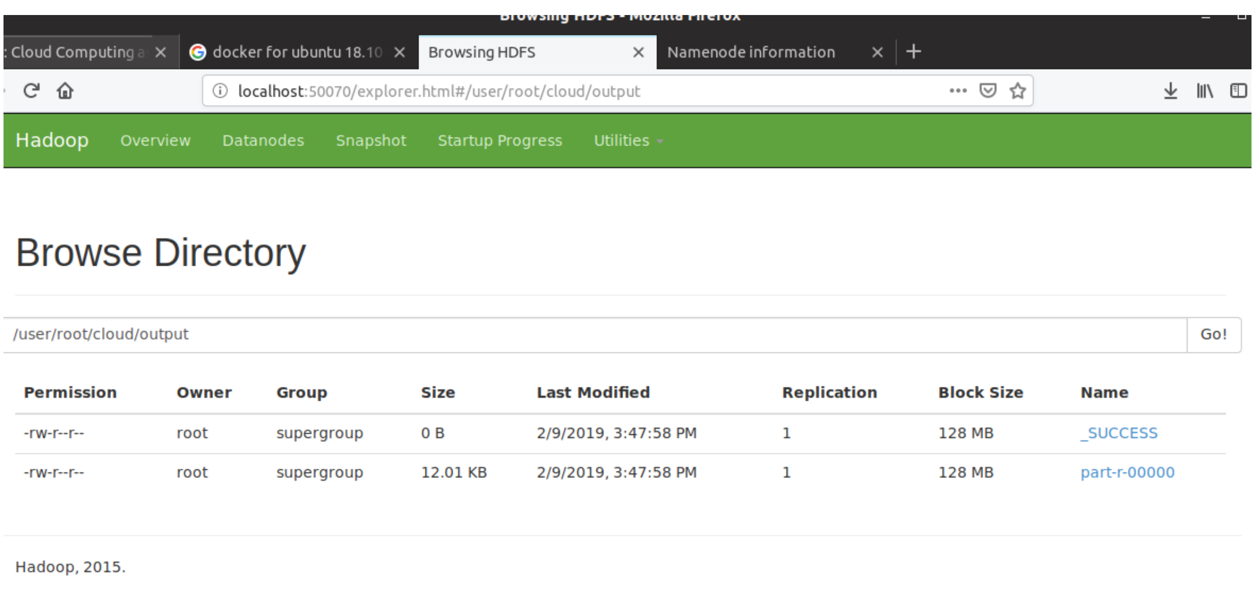
bash-4.1# bin/hadoop jar wc.jar WordCount cloud/inputs cloud/output


lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
es where applicable
bash-4.1# bin/hadoop jar wc.jar WordCount cloud/inputs cloud/output
19/02/09 18:47:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
19/02/09 18:47:36 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/02/09 18:47:37 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inter
face and execute your application with ToolRunner to remedy this.
19/02/09 18:47:37 INFO input.FileInputFormat: Total input paths to process : 1
19/02/09 18:47:37 INFO mapreduce.JobSubmitter: number of splits:1
19/02/09 18:47:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1549751770411_0003
19/02/09 18:47:38 INFO impl.YarnClientImpl: Submitted application application_1549751770411_0003
19/02/09 18:47:38 INFO mapreduce.Job: The url to track the job: http://160a2bc6b259:8088/proxy/application_1549751770411_0003/
19/02/09 18:47:38 INFO mapreduce.Job: Running job: job_1549751770411_0003
19/02/09 18:47:46 INFO mapreduce.Job: Job job_1549751770411_0003 running in uber mode : false
19/02/09 18:47:46 INFO mapreduce.Job:  map 0% reduce 0%
19/02/09 18:47:52 INFO mapreduce.Job:  map 100% reduce 0%
19/02/09 18:47:59 INFO mapreduce.Job:  map 100% reduce 100%
19/02/09 18:48:00 INFO mapreduce.Job: Job job_1549751770411_0003 completed successfully
19/02/09 18:48:00 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=16642
                FILE: Number of bytes written=264657
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=32042
                HDFS: Number of bytes written=12297
                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)=4047
                Total time spent by all reduces in occupied slots (ms)=4441
                Total time spent by all map tasks (ms)=4047
                Total time spent by all reduce tasks (ms)=4441
                Total vcore-seconds taken by all map tasks=4047
```


lakshmi@lakshmi-VirtualBox: ~
File Edit View Search Terminal Help

```
		Total time spent by all map tasks (ms)=4047
		Total time spent by all reduce tasks (ms)=4441
		Total vcore-seconds taken by all map tasks=4047
		Total vcore-seconds taken by all reduce tasks=4441
		Total megabyte-seconds taken by all map tasks=4144128
		Total megabyte-seconds taken by all reduce tasks=4547584
	Map-Reduce Framework
		Map input records=837
		Map output records=4253
		Map output bytes=46624
		Map output materialized bytes=16642
		Input split bytes=123
		Combine input records=4253
		Combine output records=1106
		Reduce input groups=1106
		Reduce shuffle bytes=16642
		Reduce input records=1106
		Reduce output records=1106
		Spilled Records=2212
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=67
		CPU time spent (ms)=2060
		Physical memory (bytes) snapshot=387002368
		Virtual memory (bytes) snapshot=1359175680
		Total committed heap usage (bytes)=314572800
	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=31919
	File Output Format Counters
		Bytes Written=12297
bash-4.1#
```



Browse Directory

/user/root/cloud/output Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	2/9/2019, 3:47:58 PM	1	128 MB	_SUCCESS
-rw-r--r--	root	supergroup	12.01 KB	2/9/2019, 3:47:58 PM	1	128 MB	part-r-00000

Hadoop, 2015.

7. Copy output files to the local directory

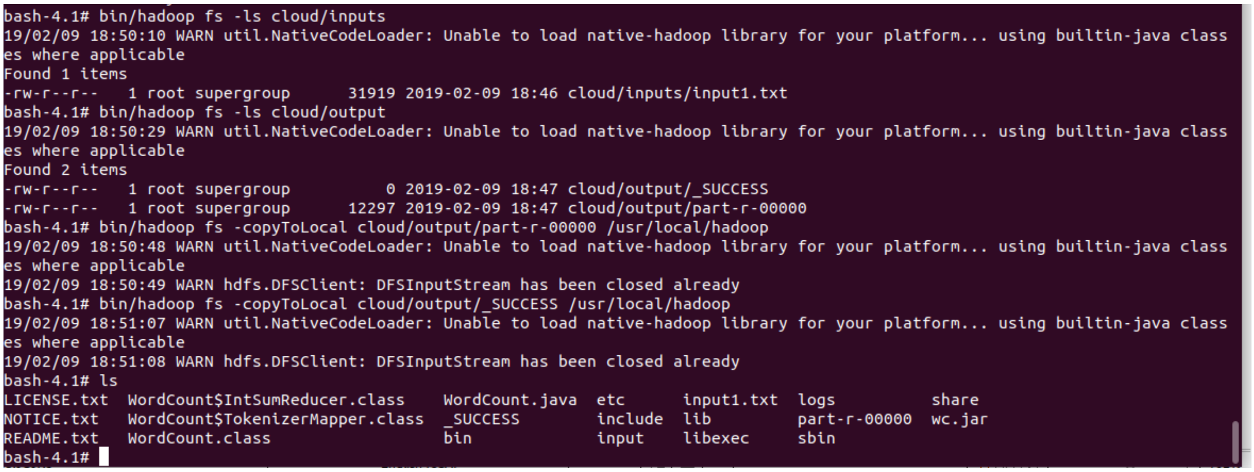
bash-4.1# bin/hadoop fs -ls cloud/inputs

bash-4.1# bin/hadoop fs -ls cloud/output

bash-4.1# bin/hadoop fs -copyToLocal cloud/output/part-r-00000 /usr/local/hadoop

bash-4.1# bin/hadoop fs -copyToLocal cloud/output/_SUCCESS /usr/local/hadoop

bash-4.1# ls



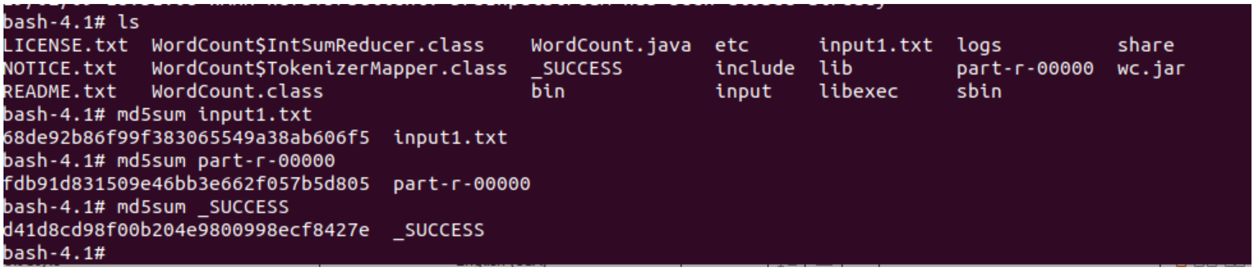
```
bash-4.1# bin/hadoop fs -ls cloud/inputs
19/02/09 18:50:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 1 items
-rw-r--r--   1 root supergroup      31919 2019-02-09 18:46 cloud/inputs/input1.txt
bash-4.1# bin/hadoop fs -ls cloud/output
19/02/09 18:50:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Found 2 items
-rw-r--r--   1 root supergroup          0 2019-02-09 18:47 cloud/output/_SUCCESS
-rw-r--r--   1 root supergroup      12297 2019-02-09 18:47 cloud/output/part-r-00000
bash-4.1# bin/hadoop fs -copyToLocal cloud/output/part-r-00000 /usr/local/hadoop
19/02/09 18:50:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
19/02/09 18:50:49 WARN hdfs.DFSClient: DFSInputStream has been closed already
bash-4.1# bin/hadoop fs -copyToLocal cloud/output/_SUCCESS /usr/local/hadoop
19/02/09 18:51:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
19/02/09 18:51:08 WARN hdfs.DFSClient: DFSInputStream has been closed already
bash-4.1# ls
LICENSE.txt  WordCount$IntSumReducer.class    WordCount.java  etc      input1.txt  logs          share
NOTICE.txt   WordCount$TokenizerMapper.class  _SUCCESS        include  lib         part-r-00000  wc.jar
README.txt   WordCount.class                  bin             input    libexec     sbin
bash-4.1# 
```

8. Run md5sum command on both input and output files

bash-4.1# md5sum input1.txt

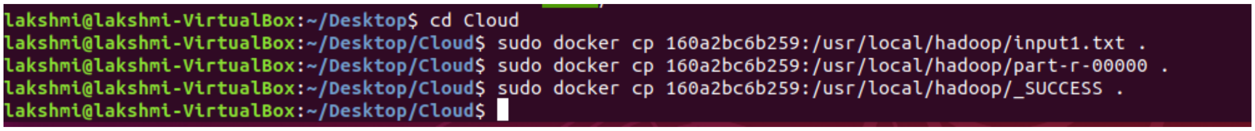
bash-4.1# md5sum part-r-00000

bash-4.1# md5sum _SUCCESS

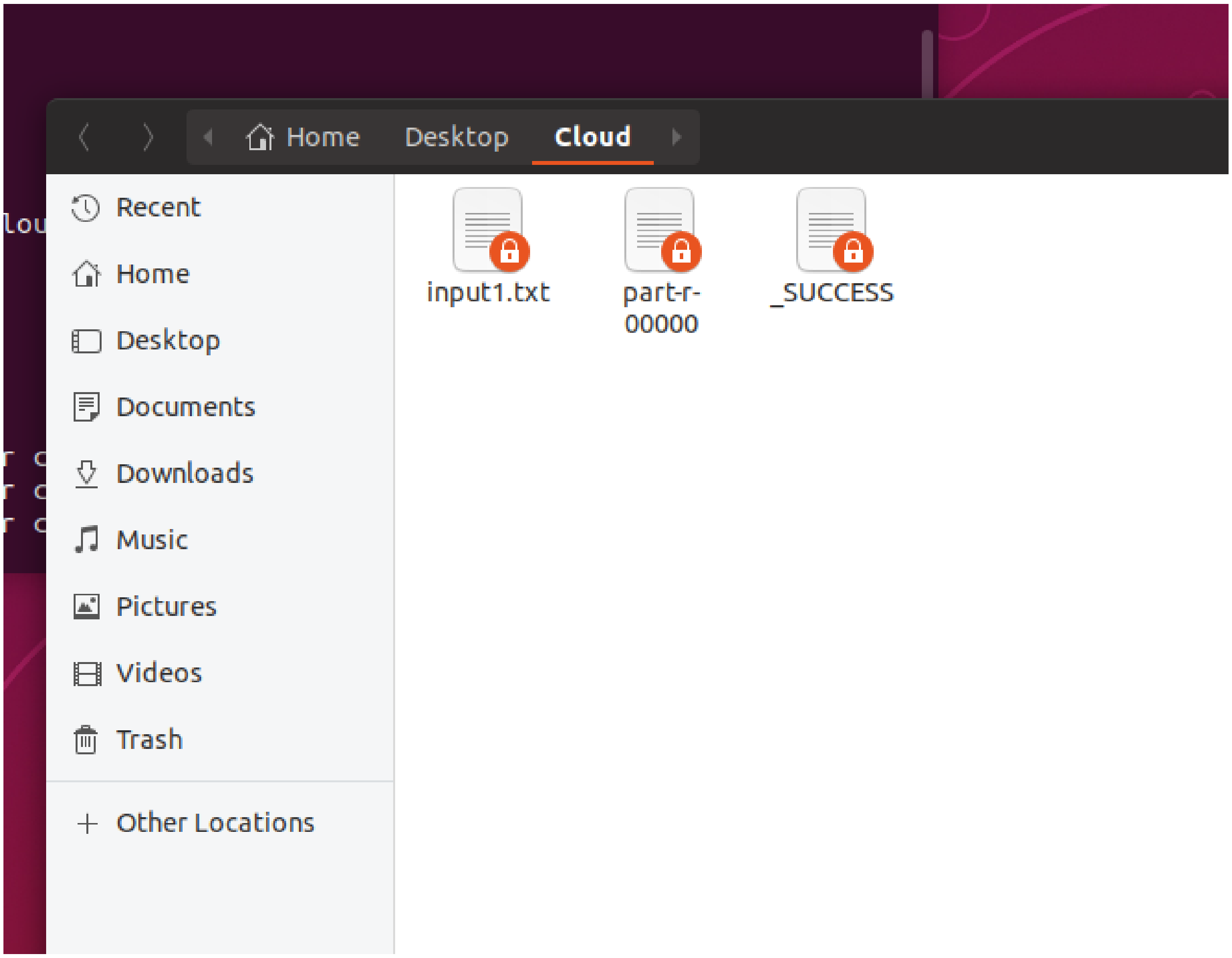


```
bash-4.1# ls
LICENSE.txt  WordCount$IntSumReducer.class    WordCount.java  etc      input1.txt  logs          share
NOTICE.txt   WordCount$TokenizerMapper.class  _SUCCESS        include  lib         part-r-00000  wc.jar
README.txt   WordCount.class                  bin             input    libexec     sbin
bash-4.1# md5sum input1.txt
68de92b86f99f383065549a38ab606f5  input1.txt
bash-4.1# md5sum part-r-00000
fdb91d831509e46bb3e662f057b5d805  part-r-00000
bash-4.1# md5sum _SUCCESS
d41d8cd98f00b204e9800998ecf8427e  _SUCCESS
bash-4.1#
```

9. Copy the output files from Hadoop file system to local system and output files.



```
lakshmi@lakshmi-VirtualBox:~/Desktop$ cd Cloud
lakshmi@lakshmi-VirtualBox:~/Desktop/Cloud$ sudo docker cp 160a2bc6b259:/usr/local/hadoop/input1.txt .
lakshmi@lakshmi-VirtualBox:~/Desktop/Cloud$ sudo docker cp 160a2bc6b259:/usr/local/hadoop/part-r-00000 .
lakshmi@lakshmi-VirtualBox:~/Desktop/Cloud$ sudo docker cp 160a2bc6b259:/usr/local/hadoop/_SUCCESS .
lakshmi@lakshmi-VirtualBox:~/Desktop/Cloud$ 
```


Home
Desktop
Cloud
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Other Locations
input1.txt
part-r-00000
_SUCCESS