```
Connected, host fingerprint: ssh-rsa 0 D9:81:D0:73:11:88:D7:96:63:01:70:C6:AA:AE
:71:43:06:6E:97:1A:0C:97:C1:CC:37:54:BF:92:CC:5D:4D:A4
HADOOP_CLASSPATH=${hadoop classpath}
Linux hadoop-cluster-1-m 4.9.0-8-amd64 #1 SMP Debian 4.9.144-3.1 (2019-02-19) x8
export HADOOP_CLASSPATH=$(hadoop classpath)
6_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
daveankin@hadoop-cluster-1-m:~$ git clone https://github.com/CS1699-2019-mba-hw4.git
Cloning into 'CS1699-2019-mba-hw4'...
remote: Not Found
fatal: repository 'https://github.com/CS1699-2019-mba-hw4.git/' not found
daveankin@hadoop-cluster-1-m:~$ git clone https://github.com/alexanderankin/CS1699-2019-mba-hw4.git
Cloning into 'CS1699-2019-mba-hw4'...
remote: Enumerating objects: 21, done.
remote: Counting objects: 100% (21/21), done.
remote: Compressing objects: 100% (18/18), done.
remote: Total 21 (delta 4), reused 20 (delta 3), pack-reused 0
Unpacking objects: 100% (21/21), done.
daveankin@hadoop-cluster-1-m:~$ cd CS1699-2019-mba-hw4/
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ l
-bash: l: command not found
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ls
acommons-cli-1.2.jar  compile.sh  gen_input.py  H4.java  input.sh  input.txt  mba_local.py  printlogs.sh  run.sh
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./compile.sh H4
javac -cp $HADOOP_CLASSPATH:acommons-cli-1.2.jar WordCount2.java -d WordCount2
javac: directory not found: H4
Usage: javac <options> <source files>
use -help for a list of possible options
jar cvf H4.jar -C H4/ .
H4/. : no such file or directory
added manifest
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ hadoop classpath
/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//*:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr
/lib/hadoop-hdfs/.//*:/usr/lib/hadoop-yarn/lib/*:/usr/lib/hadoop-yarn/.//*:/usr/lib/hadoop-mapreduce/lib/*:/usr/lib
/hadoop-mapreduce/.//*
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cd ..
daveankin@hadoop-cluster-1-m:~$ vim .bashrc
daveankin@hadoop-cluster-1-m:~$ vim .bashrc
daveankin@hadoop-cluster-1-m:~$ . $_
daveankin@hadoop-cluster-1-m:~$ cd CS1699-2019-mba-hw4/
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cd ..
```

```
daveankin@hadoop-cluster-1-m:~$ vim .bashrc
daveankin@hadoop-cluster-1-m:~$ vim .bashrc
daveankin@hadoop-cluster-1-m:~$ . $_
daveankin@hadoop-cluster-1-m:~$ cd CS1699-2019-mba-hw4/
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cd ..
daveankin@hadoop-cluster-1-m:~$ echo $JAVA_HOME
/usr/lib/jvm/java-8-openjdk-amd64
daveankin@hadoop-cluster-1-m:~$ cd
CS1699-2019-mba-hw4/ .ssh/
daveankin@hadoop-cluster-1-m:~$ cd CS1699-2019-mba-hw4/
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./compile.sh H4.ja
H4.jar   H4.java
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./compile.sh H4.java
javac -cp $HADOOP_CLASSPATH:acommons-cli-1.2.jar WordCount2.java -d WordCount2
javac: file not found: H4.java.java
Usage: javac <options> <source files>
use -help for a list of possible options
jar cvf H4.java.jar -C H4.java/ .
H4.java/. : no such file or directory
added manifest
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./compile.sh H4
javac -cp $HADOOP_CLASSPATH:acommons-cli-1.2.jar WordCount2.java -d WordCount2
javac: directory not found: H4
Usage: javac <options> <source files>
use -help for a list of possible options
jar cvf H4.jar -C H4/ .
H4/. : no such file or directory
added manifest
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ mkdir H4
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./compile.sh H4
javac -cp $HADOOP_CLASSPATH:acommons-cli-1.2.jar WordCount2.java -d WordCount2
jar cvf H4.jar -C H4/ .
added manifest
adding: H4$IntSumReducer.class(in = 1718) (out= 735)(deflated 57%)
adding: H4$TokenizerMapper$CountersEnum.class(in = 958) (out= 500)(deflated 47%)
adding: H4$TokenizerMapper.class(in = 4710) (out= 2061)(deflated 56%)
adding: H4.class(in = 2401) (out= 1317)(deflated 45%)
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cat input.sh
#!/usr/bin/env bash


source_dir=inputData
rm -rf $soure_dir
mkdir -p $source_dir
input_files='shakespeare.tar.gz'

for file in $input_files; do
```

```
for file in $input_files; do
  echo "cp $file $source_dir"
  cp $file $source_dir
  echo "(cd $source_dir; tar xvzf $file)"
  (cd $source_dir; tar xvzf $file; find . -mindepth 2 -type f -exec mv -f '{}' . ';'; find . -type d -delete)
  echo "rm $source_dir/$file"
  rm $source_dir/$file
done

daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ mkdir inputData
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cp input.txt inputData/
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ll
-bash: ll: command not found
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ls -la
total 112
drwxr-xr-x 5 daveankin daveankin  4096 Nov 20 20:03 .
drwxr-xr-x 4 daveankin daveankin  4096 Nov 20 20:02 ..
-rw-r--r-- 1 daveankin daveankin 41123 Nov 20 20:01 acommons-cli-1.2.jar
-rwxr-xr-x 1 daveankin daveankin   226 Nov 20 20:01 compile.sh
-rwxr-xr-x 1 daveankin daveankin  1344 Nov 20 20:01 gen_input.py
drwxr-xr-x 8 daveankin daveankin  4096 Nov 20 20:01 .git
drwxr-xr-x 2 daveankin daveankin  4096 Nov 20 20:03 H4
-rw-r--r-- 1 daveankin daveankin  5505 Nov 20 20:03 H4.jar
-rw-r--r-- 1 daveankin daveankin  5048 Nov 20 20:01 H4.java
-rw-r--r-- 1 daveankin daveankin   342 Nov 20 20:03 H4.java.jar
drwxr-xr-x 2 daveankin daveankin  4096 Nov 20 20:04 inputData
-rwxr-xr-x 1 daveankin daveankin   409 Nov 20 20:01 input.sh
-rw-r--r-- 1 daveankin daveankin   991 Nov 20 20:01 input.txt
-rwxr-xr-x 1 daveankin daveankin  1360 Nov 20 20:01 mba_local.py
-rwxr-xr-x 1 daveankin daveankin    84 Nov 20 20:01 printlogs.sh
-rwxr-xr-x 1 daveankin daveankin   316 Nov 20 20:01 run.sh
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cat input.sh
#!/usr/bin/env bash


source_dir=inputData
rm -rf $soure_dir
mkdir -p $source_dir
input_files='shakespeare.tar.gz'

for file in $input_files; do
  echo "cp $file $source_dir"
  cp $file $source_dir
  echo "(cd $source_dir; tar xvzf $file)"
  (cd $source_dir; tar xvzf $file; find . -mindepth 2 -type f -exec mv -f '{}' . ';'; find . -type d -delete)
  echo "rm $source_dir/$file"
```

```
  (cd $source_dir; tar xvzf $file; find . -mindepth 2 -type f -exec mv -f '{}' . ';'; find . -type d -delete)
  echo "rm $source_dir/$file"
  rm $source_dir/$file
done

daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ cat run.sh
#!/usr/bin/env bash

source_dir="inputData"

hadoop fs -rm -r -f -skipTrash $source_dir output
hadoop fs -put $source_dir/ .
hadoop jar $1.jar $1 $source_dir output

hadoop fs -getmerge -nl output collectedResults
#You can add -nl to enable adding newline char after the end of each file
echo cat collectedResults

daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ^C
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ head -n 3 run.sh | tail -n 1 > s
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ source s
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ echo $source_dir
inputData
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ head -n 6 run.sh | tail -n 1 > s
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ source s
19/11/20 20:06:47 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
put: `.': No such file or directory
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ hadoop fs -ls
19/11/20 20:07:29 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
ls: `.': No such file or directory
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ hadoop fs -mkdir -p .
19/11/20 20:07:40 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ hadoop fs -ls
19/11/20 20:07:44 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./run.sh
19/11/20 20:07:52 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
19/11/20 20:07:54 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
Not a valid JAR: /home/daveankin/CS1699-2019-mba-hw4/.jar
19/11/20 20:07:56 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
getmerge: `output': No such file or directory
cat collectedResults
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ ./run.sh H4
19/11/20 20:08:01 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
Deleted inputData
19/11/20 20:08:04 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
19/11/20 20:08:06 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
19/11/20 20:08:06 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-1-m/10.138.0.2:8032
```

```
19/11/20 20:08:06 INFO client.RMProxy: Connecting to ResourceManager at hadoop-cluster-1-m/10.138.0.2:8032
19/11/20 20:08:07 INFO input.FileInputFormat: Total input paths to process : 1
19/11/20 20:08:07 INFO mapreduce.JobSubmitter: number of splits:1
19/11/20 20:08:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1574278458300_0001
19/11/20 20:08:08 INFO impl.YarnClientImpl: Submitted application application_1574278458300_0001
19/11/20 20:08:08 INFO mapreduce.Job: The url to track the job: http://hadoop-cluster-1-m:8088/proxy/application_15
74278458300_0001/
19/11/20 20:08:08 INFO mapreduce.Job: Running job: job_1574278458300_0001
19/11/20 20:08:18 INFO mapreduce.Job: Job job_1574278458300_0001 running in uber mode : false
19/11/20 20:08:18 INFO mapreduce.Job:  map 0% reduce 0%
19/11/20 20:08:27 INFO mapreduce.Job:  map 100% reduce 0%
19/11/20 20:08:34 INFO mapreduce.Job:  map 100% reduce 8%
19/11/20 20:08:37 INFO mapreduce.Job:  map 100% reduce 17%
19/11/20 20:08:39 INFO mapreduce.Job:  map 100% reduce 25%
19/11/20 20:08:43 INFO mapreduce.Job:  map 100% reduce 33%
19/11/20 20:08:44 INFO mapreduce.Job:  map 100% reduce 42%
19/11/20 20:08:49 INFO mapreduce.Job:  map 100% reduce 58%
19/11/20 20:08:54 INFO mapreduce.Job:  map 100% reduce 67%
19/11/20 20:08:56 INFO mapreduce.Job:  map 100% reduce 75%
19/11/20 20:08:59 INFO mapreduce.Job:  map 100% reduce 83%
19/11/20 20:09:00 INFO mapreduce.Job:  map 100% reduce 92%
19/11/20 20:09:04 INFO mapreduce.Job:  map 100% reduce 100%
19/11/20 20:09:05 INFO mapreduce.Job: Job job_1574278458300_0001 completed successfully
19/11/20 20:09:06 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=3072
                FILE: Number of bytes written=1647585
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1116
                HDFS: Number of bytes written=70
                HDFS: Number of read operations=63
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=36
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=12
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=75024
                Total time spent by all reduces in occupied slots (ms)=699900
                Total time spent by all map tasks (ms)=6252
                Total time spent by all reduce tasks (ms)=58325
                Total vcore-milliseconds taken by all map tasks=6252
                Total vcore-milliseconds taken by all reduce tasks=58325
                Total megabyte-milliseconds taken by all map tasks=19206144
                Total megabyte-milliseconds taken by all reduce tasks=179174400
```

```
                Total megabyte-milliseconds taken by all reduce tasks=179174400
        Map-Reduce Framework
                Map input records=100
                Map output records=300
                Map output bytes=2400
                Map output materialized bytes=3072
                Input split bytes=125
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=3072
                Reduce input records=300
                Reduce output records=10
                Spilled Records=600
                Shuffled Maps =12
                Failed Shuffles=0
                Merged Map outputs=12
                GC time elapsed (ms)=901
                CPU time spent (ms)=8060
                Physical memory (bytes) snapshot=1714704384
                Virtual memory (bytes) snapshot=55502950400
                Total committed heap usage (bytes)=869928960
        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=991
        File Output Format Counters
                Bytes Written=70
19/11/20 20:09:07 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.10-hadoop2
cat collectedResults
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$ awk 'NF' collectedResults
A B     31
A C     35
A D     25
B C     34
A E     25
B D     31
B E     34
C D     34
C E     27
D E     24
daveankin@hadoop-cluster-1-m:~/CS1699-2019-mba-hw4$
```