

## Guideline to Using Hadoop on CSE Department's Hadoop Cluster

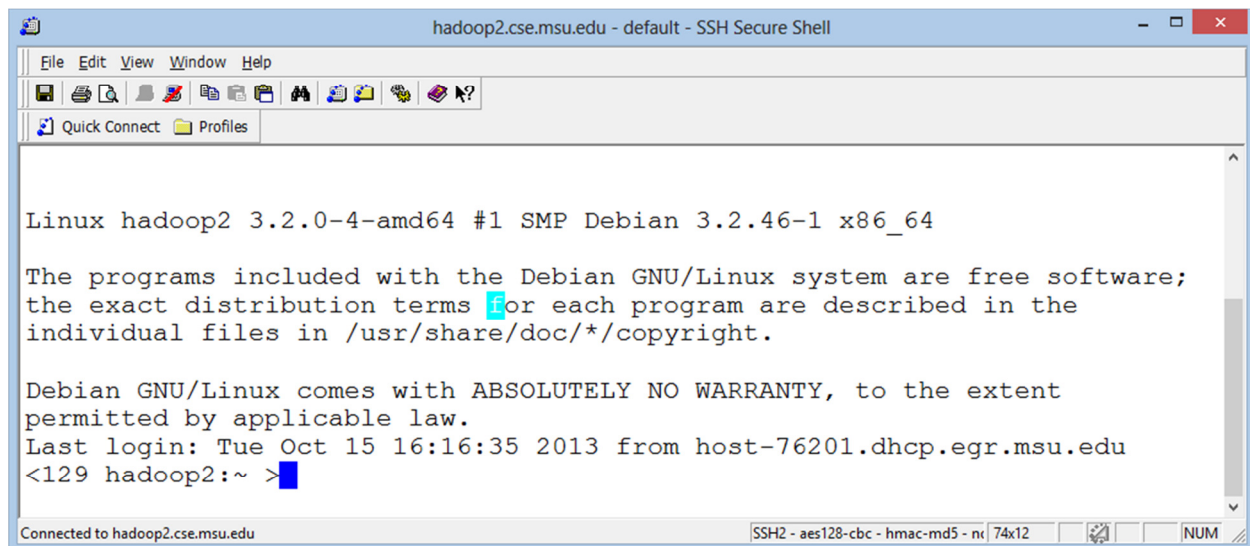
This documentation provides a step-by-step guideline on how to use the Hadoop cluster in the CSE department. This document is divided into 2 parts:

PART 1: Getting Started

PART 2: Running a Hadoop Program

### PART 1: GETTING STARTED

**STEP 1:** Open an SSH connection to `hadoop1.cse.msu.edu` or `hadoop2.cse.msu.edu`. You should be able to login using your CSE account. The disk drives are NSF-mounted, so you're basically logged in to your CSE home directory.



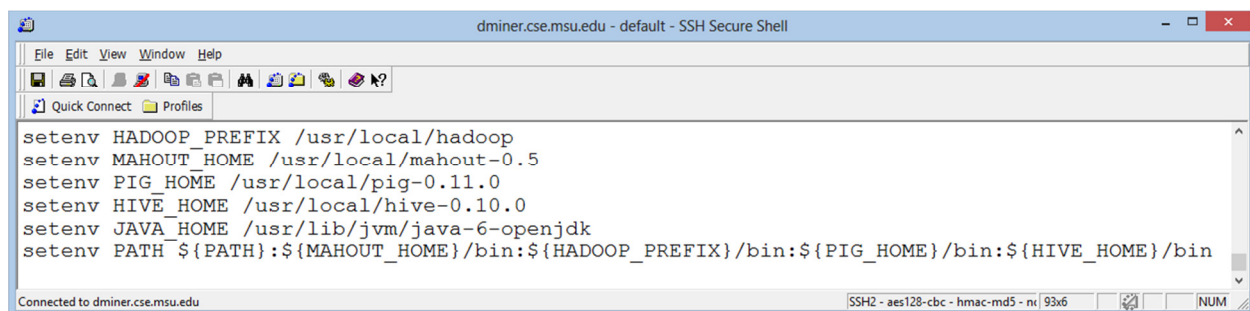
```
hadoop2.cse.msu.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

Linux hadoop2 3.2.0-4-amd64 #1 SMP Debian 3.2.46-1 x86_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.
Last login: Tue Oct 15 16:16:35 2013 from host-76201.dhcp.egr.msu.edu
<129 hadoop2:~ >
```

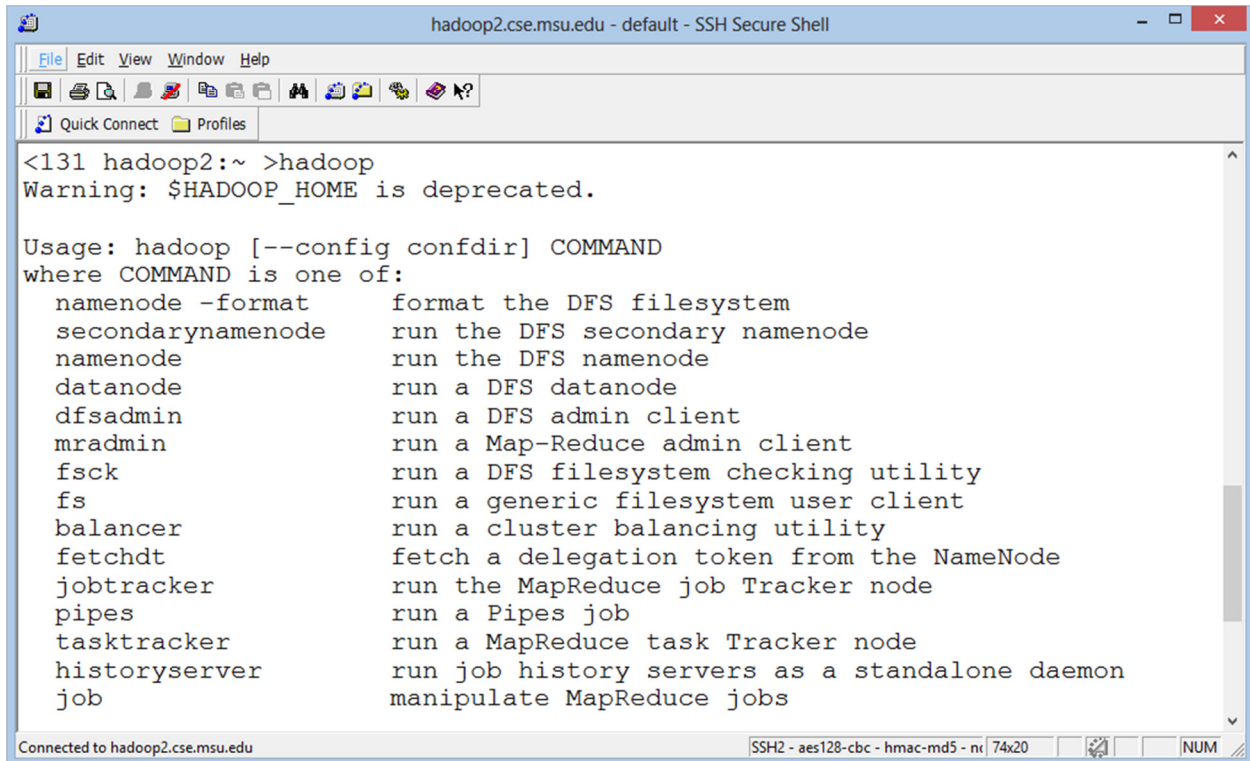
**STEP 2:** Try typing `hadoop` on the Linux command line. If it doesn't work, you'll need to setup the environment variables for `PATH`, `HADOOP_PREFIX`, `MAHOUT_HOME`, `PIG_HOME`, `JAVA_HOME`, and `HIVE_HOME`. You can set it up in your profile to make sure these variables are set each time you logged in to the cluster. For `tcsh` shell, you can modify the variables by adding the following lines to your `.personal` file in your home directory:



```
dminer.cse.msu.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

setenv HADOOP_PREFIX /usr/local/hadoop
setenv MAHOUT_HOME /usr/local/mahout-0.5
setenv PIG_HOME /usr/local/pig-0.11.0
setenv HIVE_HOME /usr/local/hive-0.10.0
setenv JAVA_HOME /usr/lib/jvm/java-6-openjdk
setenv PATH ${PATH}:${MAHOUT_HOME}/bin:${HADOOP_PREFIX}/bin:${PIG_HOME}/bin
```

**STEP 3:** Now you're ready to run hadoop. Type hadoop on command line and it will display a list of commands you can use.



The screenshot shows an SSH Secure Shell window titled "hadoop2.cse.msu.edu - default - SSH Secure Shell". The window has a menu bar (File, Edit, View, Window, Help) and a toolbar. Below the toolbar is a "Quick Connect" button and a "Profiles" button. The main text area displays the output of the command `<131 hadoop2:~ >hadoop`. The output is as follows:

```
<131 hadoop2:~ >hadoop
Warning: $HADOOP_HOME is deprecated.

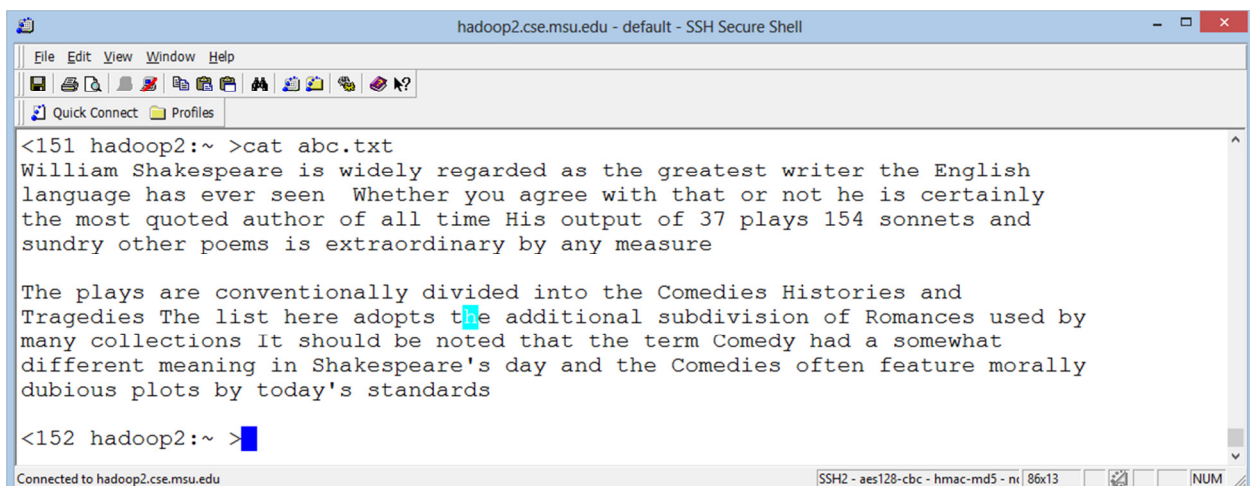
Usage: hadoop [--config confdir] COMMAND
where COMMAND is one of:
    namenode -format          format the DFS filesystem
    secondarynamenode        run the DFS secondary namenode
    namenode                  run the DFS namenode
    datanode                  run a DFS datanode
    dfsadmin                  run a DFS admin client
    mradmin                   run a Map-Reduce admin client
    fsck                      run a DFS filesystem checking utility
    fs                        run a generic filesystem user client
    balancer                  run a cluster balancing utility
    fetchdt                   fetch a delegation token from the NameNode
    jobtracker                run the MapReduce job Tracker node
    pipes                     run a Pipes job
    tasktracker               run a MapReduce task Tracker node
    historyserver             run job history servers as a standalone daemon
    job                       manipulate MapReduce jobs
```

The status bar at the bottom of the window shows "Connected to hadoop2.cse.msu.edu", "SSH2 - aes128-cbc - hmac-md5 - nr", "74x20", and a "NUM" button.

## PART 2: RUNNING A HADOOP PROGRAM

The hadoop installation comes with example programs and datasets. In this example, we will illustrate how to run the Hadoop program for counting words in a collection of documents.

**STEP 1:** In this example, you will count the number of times each word appears in a given input file named `abc.txt`. You can replace this with any files that you want.



The screenshot shows an SSH Secure Shell window titled "hadoop2.cse.msu.edu - default - SSH Secure Shell". The window has a menu bar (File, Edit, View, Window, Help) and a toolbar. Below the toolbar is a "Quick Connect" button and a "Profiles" button. The main text area displays the output of the command `<151 hadoop2:~ >cat abc.txt`. The output is as follows:

```
<151 hadoop2:~ >cat abc.txt
William Shakespeare is widely regarded as the greatest writer the English
language has ever seen Whether you agree with that or not he is certainly
the most quoted author of all time His output of 37 plays 154 sonnets and
sundry other poems is extraordinary by any measure

The plays are conventionally divided into the Comedies Histories and
Tragedies The list here adopts the additional subdivision of Romances used by
many collections It should be noted that the term Comedy had a somewhat
different meaning in Shakespeare's day and the Comedies often feature morally
dubious plots by today's standards

<152 hadoop2:~ >
```

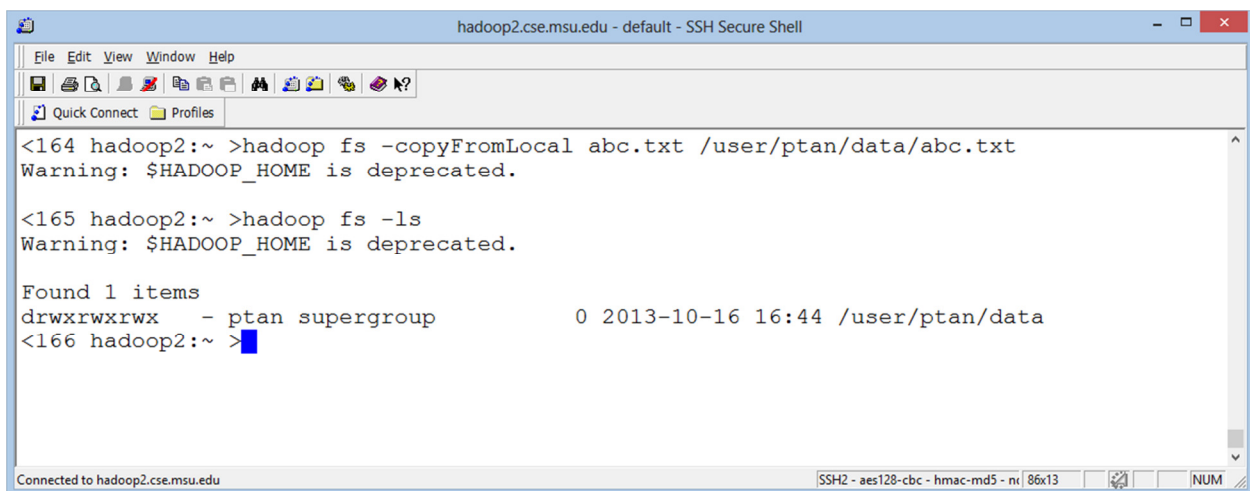
The status bar at the bottom of the window shows "Connected to hadoop2.cse.msu.edu", "SSH2 - aes128-cbc - hmac-md5 - nr", "86x13", and a "NUM" button.

**STEP 2:** Upload the data from your local directory to Hadoop Distributed File System (HDFS). If you're using HDFS for the first time, make sure you type the full path of the destination directory. The command for copying the file is:

```
> hadoop fs -copyFromLocal <source> <destination>
```

In the example below, the file `abc.txt` will be uploaded to HDFS and stored in the directory named `/user/ptan/data`. To make sure the file is copied correctly, you can use the `ls` command to list the content of your Hadoop working directory

```
> hadoop fs -ls
```

A screenshot of an SSH Secure Shell window titled "hadoop2.cse.msu.edu - default - SSH Secure Shell". The window has a menu bar (File, Edit, View, Window, Help) and a toolbar. Below the toolbar is a "Quick Connect" section with a "Profiles" button. The main area is a terminal window showing the following commands and output:

```
<164 hadoop2:~ >hadoop fs -copyFromLocal abc.txt /user/ptan/data/abc.txt
Warning: $HADOOP_HOME is deprecated.

<165 hadoop2:~ >hadoop fs -ls
Warning: $HADOOP_HOME is deprecated.

Found 1 items
drwxrwxrwx  - ptan supergroup          0 2013-10-16 16:44 /user/ptan/data
<166 hadoop2:~ >
```

The status bar at the bottom indicates "Connected to hadoop2.cse.msu.edu" and "SSH2 - aes128-cbc - hmac-md5 - nr 86x13".

**STEP 3:** We will use a sample Hadoop program called wordcount to do this. The Java program is archived in a file called `hadoop-examples-1.1.1.jar` located in `/usr/local/hadoop` directory. This program expects two input arguments: (1) source directory (in HDFS) that contains the input data, and (2) destination directory (in HDFS) that contains the results. The typical syntax for executing a Hadoop java program (stored in a jar file) is as follows:

```
> hadoop jar <name-of-jar-file> <name-of-java-program> <input arguments>
```

In the example shown below, the wordcount program will read input documents stored in `/user/ptan/data` directory and write the output to `/user/ptan/output` directory.

```
hadoop2.cse.msu.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

<167 hadoop2:~ >hadoop jar ${HADOOP_HOME}/hadoop-examples-1.1.1.jar wordcount /user/ptan/data /user/ptan/output
Warning: $HADOOP_HOME is deprecated.

13/10/16 16:46:29 INFO input.FileInputFormat: Total input paths to process : 1
13/10/16 16:46:29 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/10/16 16:46:29 WARN snappy.LoadSnappy: Snappy native library not loaded
13/10/16 16:46:30 INFO mapred.JobClient: Running job: job_201308211055_0058
13/10/16 16:46:31 INFO mapred.JobClient: map 0% reduce 0%
13/10/16 16:46:36 INFO mapred.JobClient: map 100% reduce 0%
13/10/16 16:46:44 INFO mapred.JobClient: map 100% reduce 33%
13/10/16 16:46:45 INFO mapred.JobClient: map 100% reduce 100%
13/10/16 16:46:46 INFO mapred.JobClient: Job complete: job_201308211055_0058
13/10/16 16:46:46 INFO mapred.JobClient: Counters: 29
13/10/16 16:46:46 INFO mapred.JobClient: Job Counters
13/10/16 16:46:46 INFO mapred.JobClient: Launched reduce tasks=12
13/10/16 16:46:46 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=5380
13/10/16 16:46:46 INFO mapred.JobClient: Total time spent by all reduces waiting a
fter reserving slots (ms)=0
13/10/16 16:46:46 INFO mapred.JobClient: Total time spent by all maps waiting afte
r reserving slots (ms)=0

Connected to hadoop2.cse.msu.edu
SSH2 - aes128-cbc - hmac-md5 - nr 86x21 NUM
```

**STEP 4:** To verify that the job has been completed, list the content of /user/ptan/output directory and check whether there is a file named \_SUCCESS. If it is successful, the results will be stored in the files called part-r-XXXXX (where X is a digit from 0 to 9). The number of output files generated depends on the number of reducers used to run the Hadoop job.

```
hadoop2.cse.msu.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

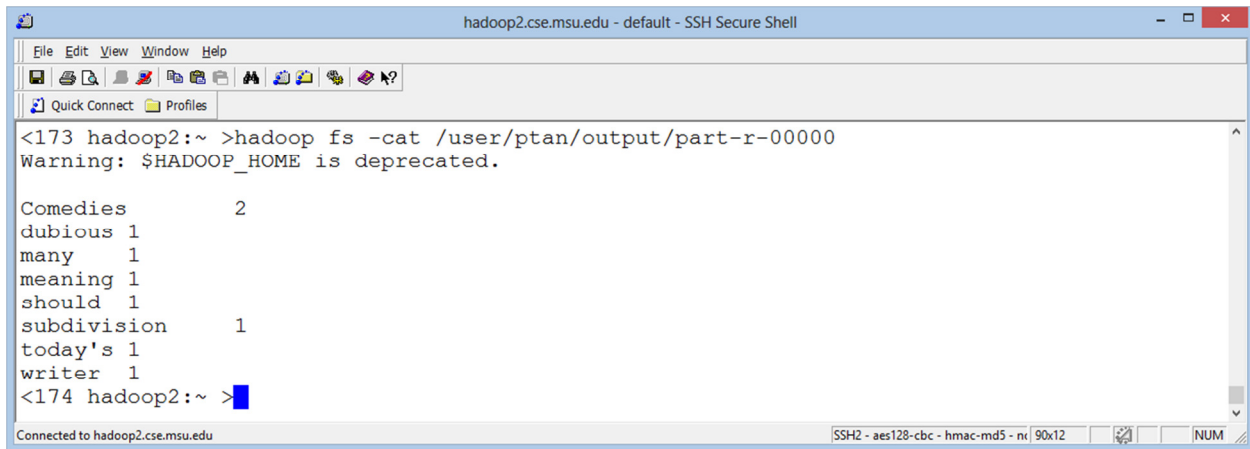
<171 hadoop2:~ >hadoop fs -ls /user/ptan/output
Warning: $HADOOP_HOME is deprecated.

Found 14 items
-rw-r--r-- 2 ptan supergroup 0 2013-10-16 16:46 /user/ptan/output/_SUCCESS
drwxrwxrwx - ptan supergroup 0 2013-10-16 16:46 /user/ptan/output/_logs
-rw-r--r-- 2 ptan supergroup 80 2013-10-16 16:46 /user/ptan/output/part-r-00000
-rw-r--r-- 2 ptan supergroup 87 2013-10-16 16:46 /user/ptan/output/part-r-00001
-rw-r--r-- 2 ptan supergroup 85 2013-10-16 16:46 /user/ptan/output/part-r-00002
-rw-r--r-- 2 ptan supergroup 52 2013-10-16 16:46 /user/ptan/output/part-r-00003
-rw-r--r-- 2 ptan supergroup 68 2013-10-16 16:46 /user/ptan/output/part-r-00004
-rw-r--r-- 2 ptan supergroup 52 2013-10-16 16:46 /user/ptan/output/part-r-00005
-rw-r--r-- 2 ptan supergroup 29 2013-10-16 16:46 /user/ptan/output/part-r-00006
-rw-r--r-- 2 ptan supergroup 63 2013-10-16 16:46 /user/ptan/output/part-r-00007
-rw-r--r-- 2 ptan supergroup 76 2013-10-16 16:46 /user/ptan/output/part-r-00008
-rw-r--r-- 2 ptan supergroup 64 2013-10-16 16:46 /user/ptan/output/part-r-00009
-rw-r--r-- 2 ptan supergroup 20 2013-10-16 16:46 /user/ptan/output/part-r-00010
-rw-r--r-- 2 ptan supergroup 17 2013-10-16 16:46 /user/ptan/output/part-r-00011

<172 hadoop2:~ >
```

**STEP 5:** You can view the content of a file by using the following command

> `hadoop fs -cat <filename>`



The image shows a terminal window titled "hadoop2.cse.msu.edu - default - SSH Secure Shell". The terminal displays the output of the command `hadoop fs -cat /user/ptan/output/part-r-00000`. The output is a list of words and their counts: "Comedies 2", "dubious 1", "many 1", "meaning 1", "should 1", "subdivision 1", "today's 1", and "writer 1". A warning message at the top states: "Warning: \$HADOOP\_HOME is deprecated." The terminal also shows the prompt `<173 hadoop2:~ >` and the status bar at the bottom indicates "Connected to hadoop2.cse.msu.edu" and "SSH2 - aes128-cbc - hmac-md5 - nx | 90x12".

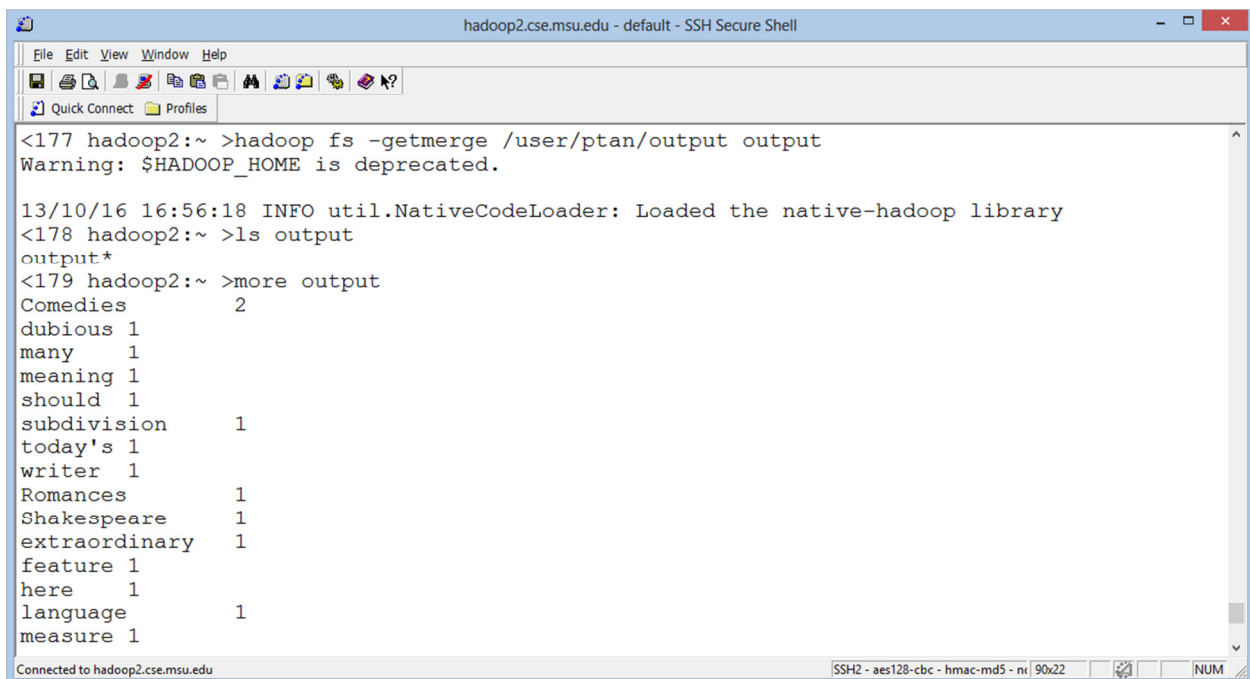
```
<173 hadoop2:~ >hadoop fs -cat /user/ptan/output/part-r-00000
Warning: $HADOOP_HOME is deprecated.

Comedies          2
dubious 1
many 1
meaning 1
should 1
subdivision      1
today's 1
writer 1
<174 hadoop2:~ >
```

**STEP 6:** You can merge the outputs of the program and write the output back to your local directory using the `getmerge` command.

`> hadoop fs -getmerge <source-directory> <destination-local-filename>`

For example, the following command will combine all the `part-r-XXXXXX` files in `/user/ptan/output` directory and store them into a single file called `output`.



The image shows a terminal window titled "hadoop2.cse.msu.edu - default - SSH Secure Shell". The terminal displays the output of the command `hadoop fs -getmerge /user/ptan/output output`. The output includes a warning message: "Warning: \$HADOOP\_HOME is deprecated." and a log message: "13/10/16 16:56:18 INFO util.NativeCodeLoader: Loaded the native-hadoop library". The terminal also shows the prompt `<177 hadoop2:~ >` and the status bar at the bottom indicates "Connected to hadoop2.cse.msu.edu" and "SSH2 - aes128-cbc - hmac-md5 - nx | 90x22".

```
<177 hadoop2:~ >hadoop fs -getmerge /user/ptan/output output
Warning: $HADOOP_HOME is deprecated.

13/10/16 16:56:18 INFO util.NativeCodeLoader: Loaded the native-hadoop library
<178 hadoop2:~ >ls output
output*
<179 hadoop2:~ >more output
Comedies          2
dubious 1
many 1
meaning 1
should 1
subdivision      1
today's 1
writer 1
Romances         1
Shakespeare      1
extraordinary    1
feature 1
here 1
language         1
measure 1
```