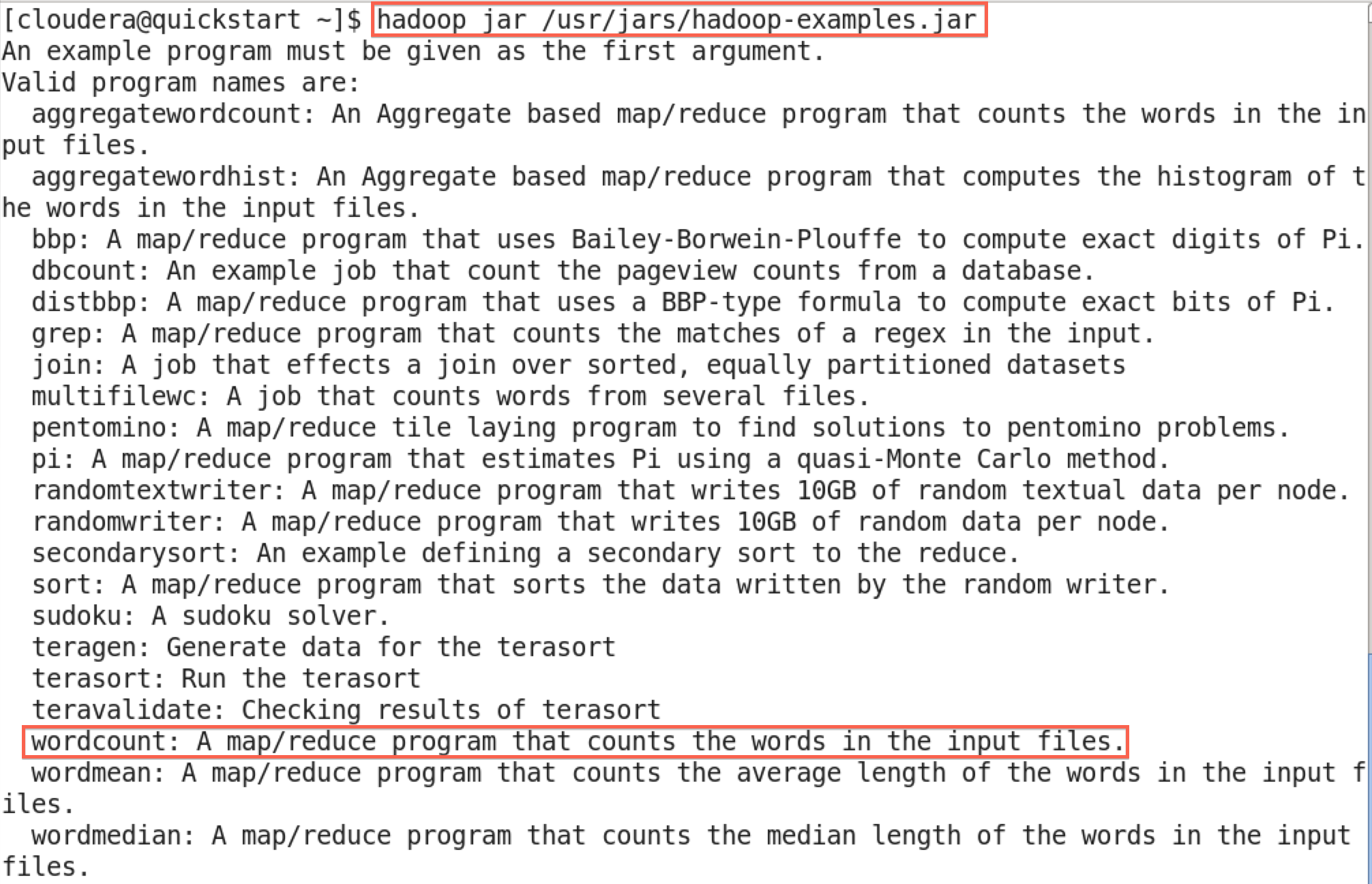
Learning Goals

By the end of this activity, you will be able to:

* Execute the WordCount application.
* Copy the results from WordCount out of HDFS.

1. Open a terminal shell. Start the Cloudera VM in VirtualBox, if not already running, and open a terminal shell. Detailed instructions for these steps can be found in the previous Readings.

2.**See example MapReduce programs.** Hadoop comes with several example MapReduce applications. You can see a list of them by running *hadoop jar /usr/jars/hadoop-examples.jar.*We are interested in running WordCount.

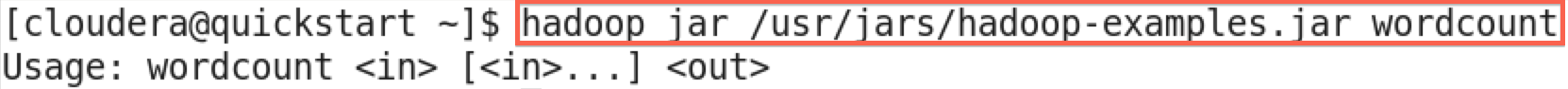


The output says that WordCount takes the name of one or more input files and the name of the output directory. Note that these files are in HDFS, not the local file system.

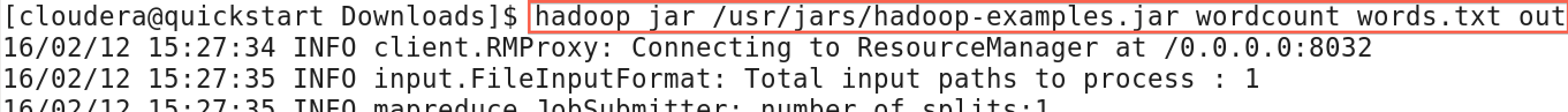
3.**Verify input file exists.**In the previous Reading, we downloaded the complete works of Shakespeare and copied them into HDFS. Let's make sure this file is still in HDFS so we can run WordCount on it. Run *hadoop fs -ls*



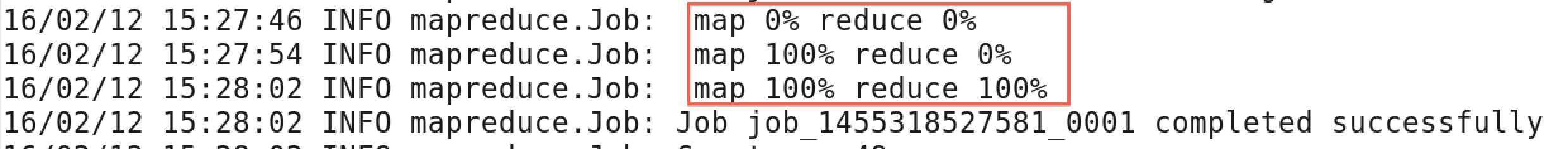
4.**See WordCount command line arguments.** We can learn how to run WordCount by examining its command-line arguments. Run *hadoop jar /usr/jars/hadoop-examples.jar wordcount*.



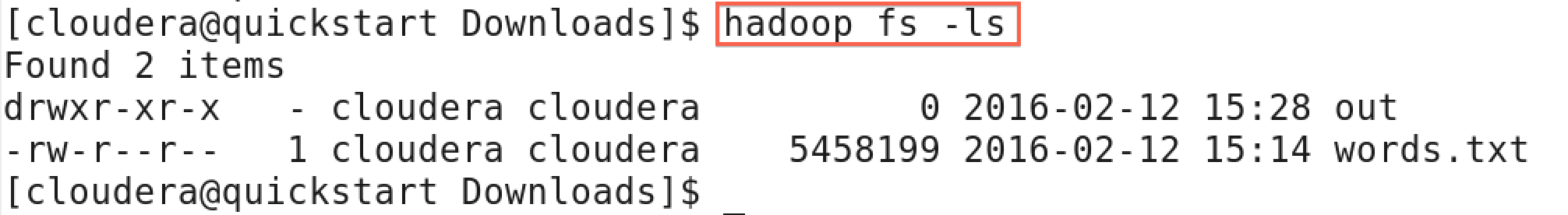
5.**Run WordCount.** Run WordCount for hadoop.txt: *hadoop jar /usr/jars/hadoop-examples.jar wordcount words.txt out*



As WordCount executes, the Hadoop prints the progress in terms of Map and Reduce. When the WordCount is complete, both will say *100%*.

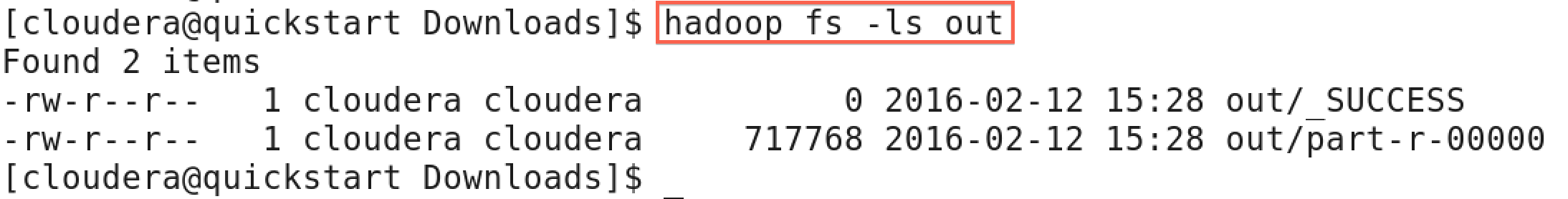


6.**See WordCount output directory.** Once WordCount is finished, let's verify the output was created. First, let's see that the output directory, *out,*was created in HDFS by running *hadoop fs –ls*



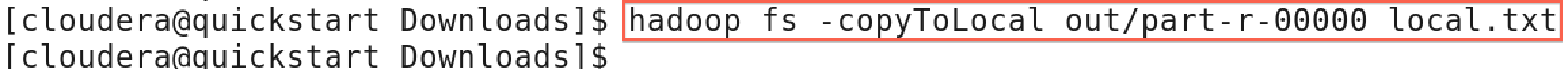
We can see there are now two items in HDFS: *words.txt*is the text file that we previously created, and *out*is the directory created by WordCount.

7.**Look inside output directory.** The directory created by WordCount contains several files. Look inside the directory by running *hadoop –fs ls out*



The file *part-r-00000* contains the results from WordCount. The file \_SUCCESS means WordCount executed successfully.

8.**Copy WordCount results to local file system.** Copy *part-r-00000* to the local file system by running *hadoop fs –copyToLocal out/part-r-00000 local.txt*



9.**View the WordCount results.** View the contents of the results: *more local.txt*



Each line of the results file shows the number of occurrences for a word in the input file. For example, *Accuse*appears four times in the input, but *Accusing* appears only once.

