**Metropolitan State University**

**ICS 490-01: Big Data Storage**

**Running a map-reduce job in Cloudera’s Virtual Machine**

This tutorial includes a step-by-step procedure to run the Word Count Map Reduce job on Hadoop.

**Step 1: Creating and compiling a MapReduce job in Java:**

1. A zip file, called WordCountMRJob, is posted on D2L that includes the following:
   1. Three .java files that includes the MapReduce implementation of the Word Count example.
   2. A directory called shakespeare that includes the input text files to be processed by the Word Count MapReduce Job.
2. Copy the zip file and paste it on the desktop of your Cloudera VM.
3. Open Eclipse from inside the virtual machine and create a new project called WordCountProject
4. Import the following three java files into the ‘src’ directory of Eclipse’s ‘WordCountProject’ project:
   1. WordCount.java
   2. WordCountMapper.java
   3. WordCountReducer.java
5. The file will initially have compilation errors because of missing libraries. To resove this issue, add all the Hadoop libraries to your project build path. This can be done as follows:
   1. Right clock on WordCountProject project
   2. Click on ‘Properties’
   3. From the left menu, choode ‘Java Build Path’
   4. Click on ‘Add External Jars’
   5. Browse to:
      1. File System/usr/lib/Hadoop
         1. choose hadoop-common.jar
      2. File System/usr/lib/Hadoop-0.20-mapreduce
         1. choose hadoop-core-mr1.jar

**Now all compilation erros should be resoved.**

1. Eclipse automatically compiles your files and generates .class files in the following location:

desktop/cloudera’shome/workspace/WordCountProject

1. Open a command window and browse to the directory using the following:

$cd workspace/WordCountProject

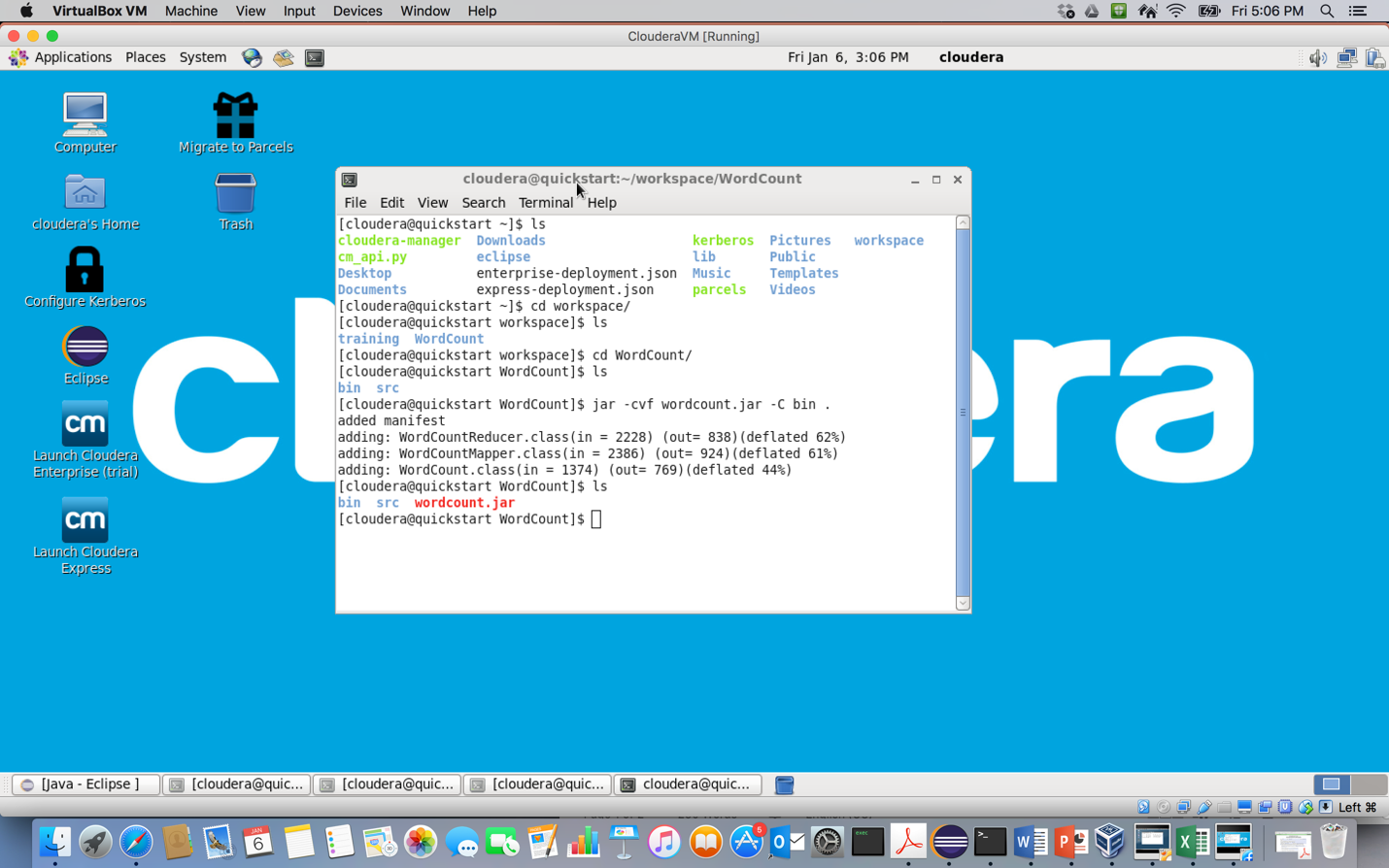
you can use ls to list the contents of your current folder.

$ls

1. From the command window, create a jar file using the following command (note that the dot at the end of the command is part of the command)

**$jar –cvf wordcount.jar –C bin .**

1. Once you do that, now you have wordcountjob.jar created and this is the file that is used by Hadoop to run the WordCount MapReduce job.



**Step 2: Moving the input data files to HDFS:**

1. The input data files is given to you in a directory called shakepeare in the attached zip file. The word count job is supposed to count how many times each word appears in shakepeare data set. Copy the shakepeare directory and paste under the cloudera directory on your desktop. Note that the location of the directory is important as you are going to use this location in the next step.
2. MapReduce jobs can only read from and write to HDFS. So in order to run the word count job, you have to copy shakepeare directory to HDFS. You can do that by following these steps:

Open a new command window.

$hadoop fs –mkdir /user/<your-name>

$hadoop fs –mkdir /user/<your-name>/wordcount

$hadoop fs –mkdir /user/<your-name>/wordcount/input

$hadoop fs –copyFromLocal shakespeare

/user/<your-name>/wordcount/input

**Step 3: Running the MapReduce job and checking the output:**

1. By now you created the word count jar wordcount.jar and copied the input data directory shkaespeare to DHFS. The next step is to run the job.
2. You can run the map reduce job using the following command. Make sure to run this command from inside the directory at which the jar file is located. You can do that as follows:

Open a new command window.

$cd workspace/WordCountProject

$hadoop jar wordcount.jar WordCount

/user/<your-name>/wordcount/input/shakespeare

/user/<your-name>/wordcount/output

1. The output of the job is saved to the HDFS file /user/<your-name>/wordcount/output
2. You can check the output file by using the following command:

$hadoop fs –cat /user/<your-name>/wordcount/output/part-r-00000

Or you can copy the file from HDFS to your VM and open the file using any file browser.