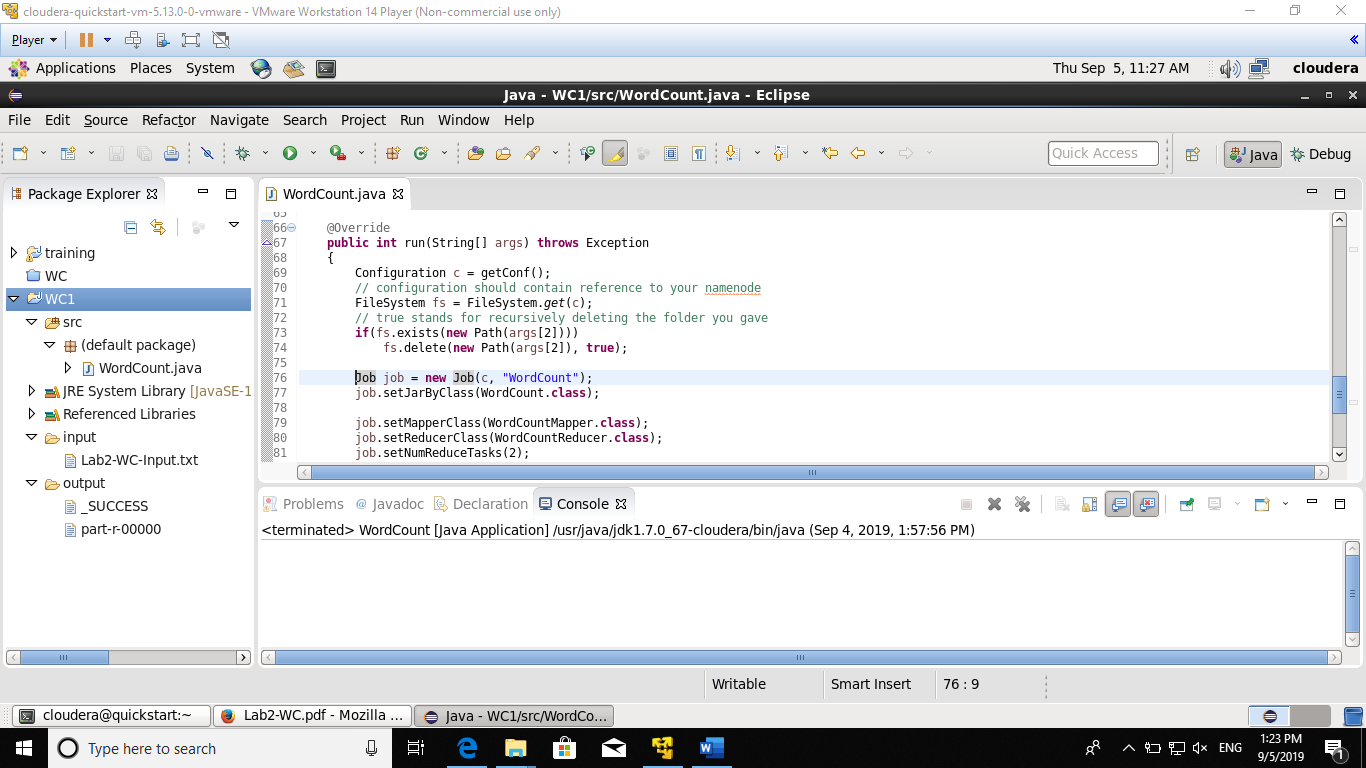
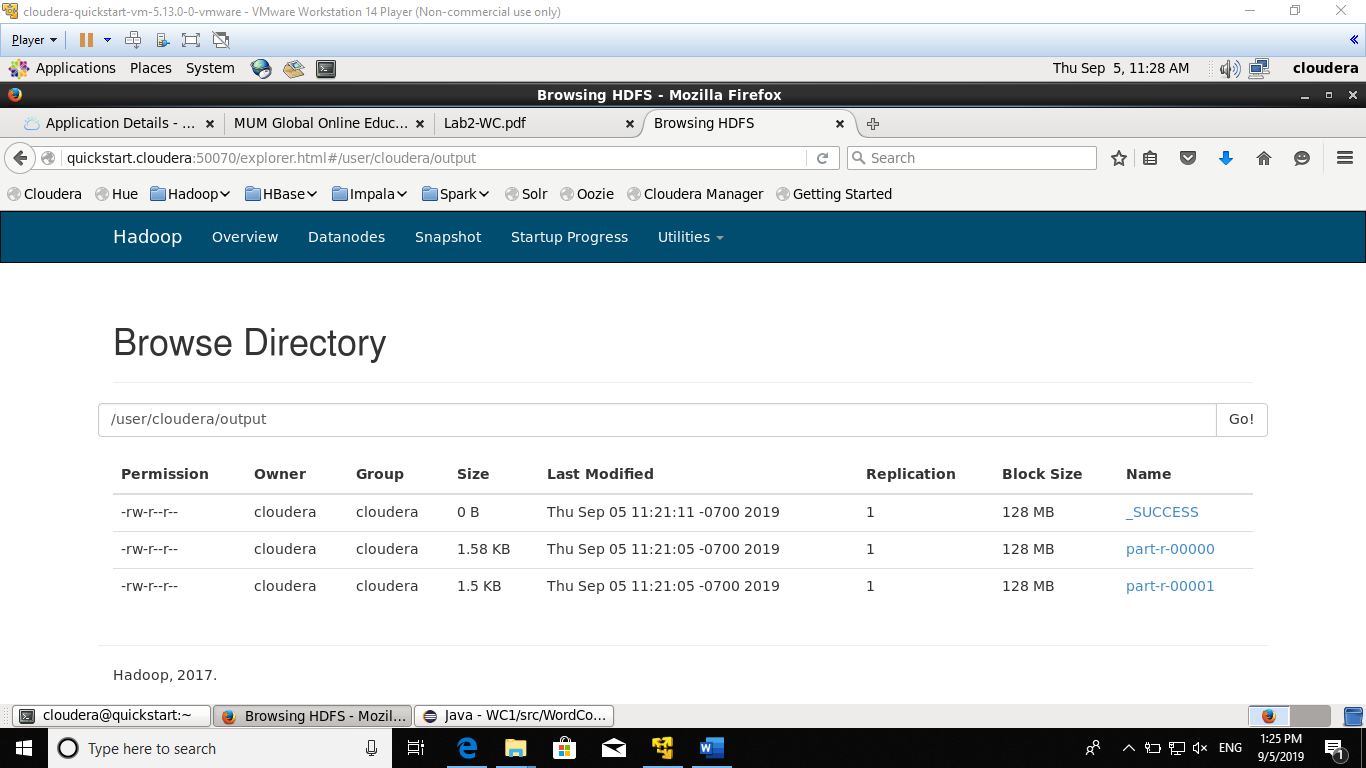
**Homework**

a. Do some research and find out what code needs to be added to the Word Count program for automatic removal of “output” directory before job execution.



b. Run the above basic Word Count program in pseudo distributed mode with 2 reducers. Use setNumReduceTask method of Job object. Paste the screenshot of the two partr-\* files created in HDFS. (Note: multiple reducers only work in pseudo-distributed mode and not local mode)



c. Modify the WordCount program to output the counts of only the words “Hadoop” and “Java”. (Count Hadoop and hadoop as same word!) Submit your output file.

File submitted under folder ProblemC

d. Modify the WordCount program to output the counts of only those words which appear in the document at least 25 times. Submit your output file.

File submitted under folder ProblemD

e. Write a program to find out how many distinct words are there in the input file. Submit your output file.

File submitted under folder ProblemE