**Practical 2**

**Aim: WordCount Application using Java and MapReduce**

**What is MapReduce ?**

A MapReduce is a data processing tool which is used to process the data parallely in a

distributed form. It was developed in 2004, on the basis of paper titled as "MapReduce:

Simplified Data Processing on Large Clusters," published by Google.

The MapReduce is a paradigm which has two phases, the mapper phase, and the reducer phase.

In the Mapper, the input is given in the form of a key-value pair. The output of the Mapper is fed

to the reducer as input. The reducer runs only after the Mapper is over. The reducer too takes

input in key-value format, and the output of reducer is the final output.

**Steps in Map Reduce**

• The map takes data in the form of pairs and returns a list of <key, value> pairs. The keys

will not be unique in this case.

• Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This

sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a

list of values associated with this unique key <key, list(values)>.

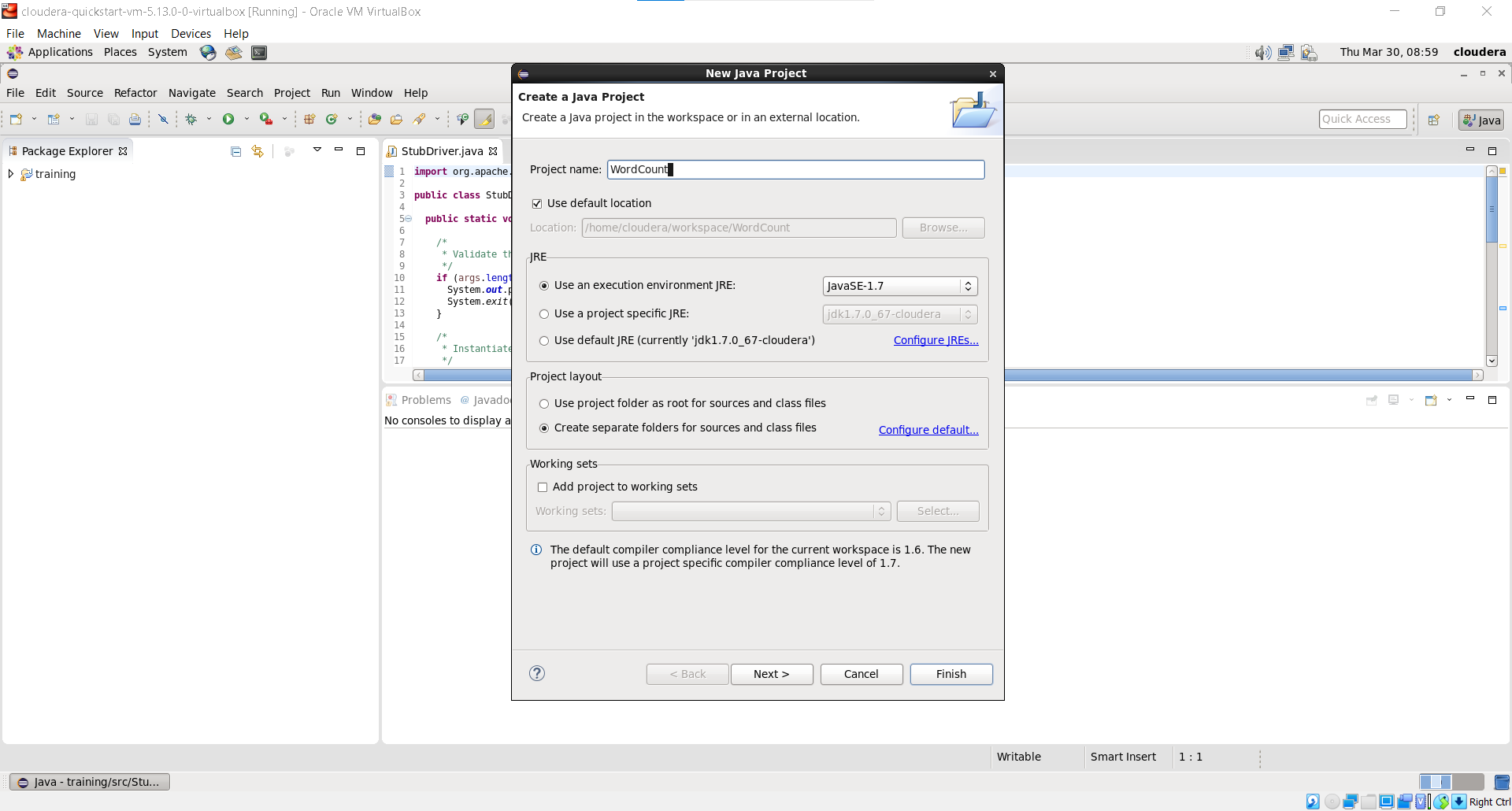
• An output of sort and shuffle sent to the reducer phase. The reducer performs a defined

function on a list of values for unique keys, and Final output <key, value> will be

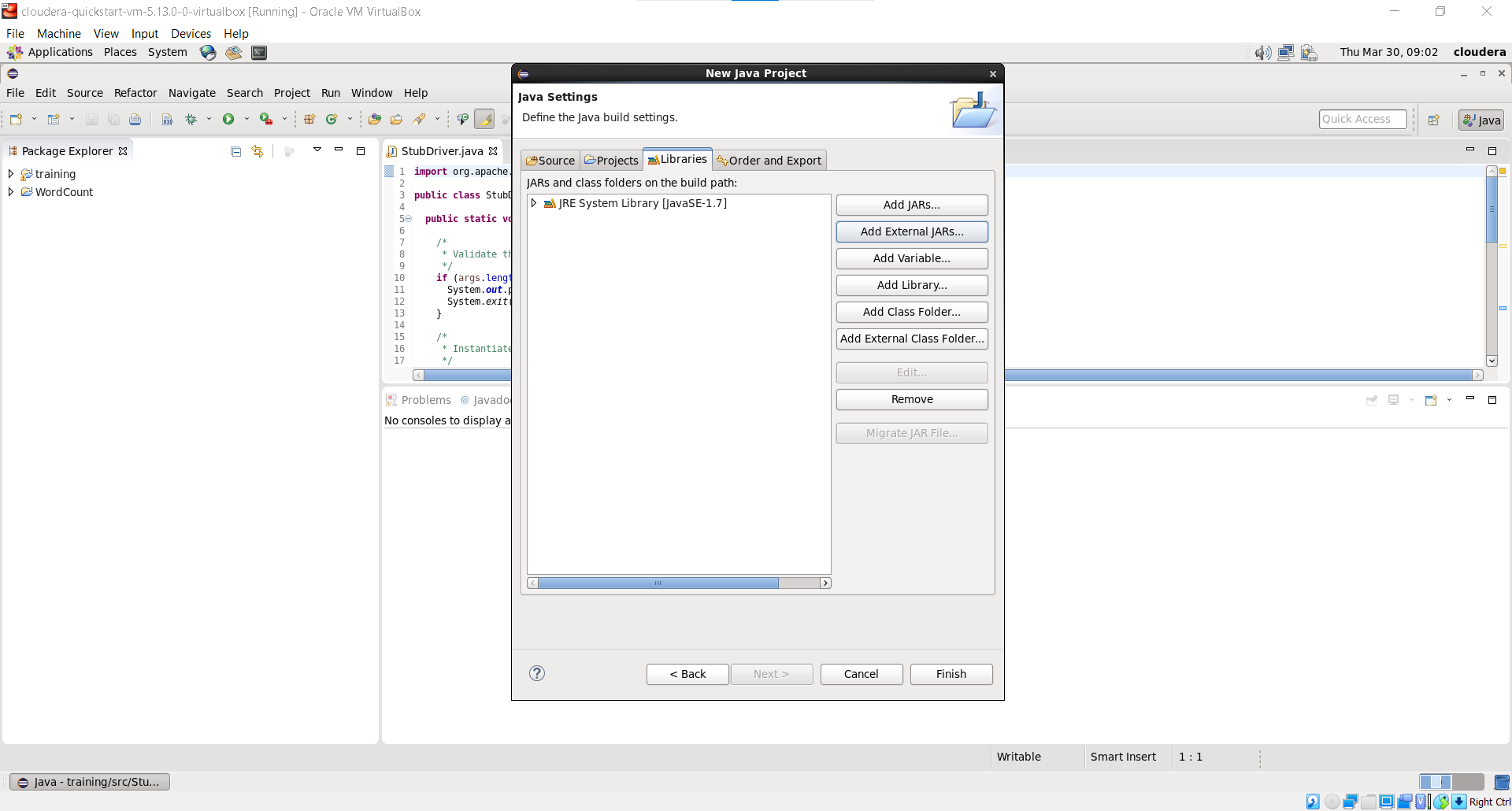
stored/displayed.

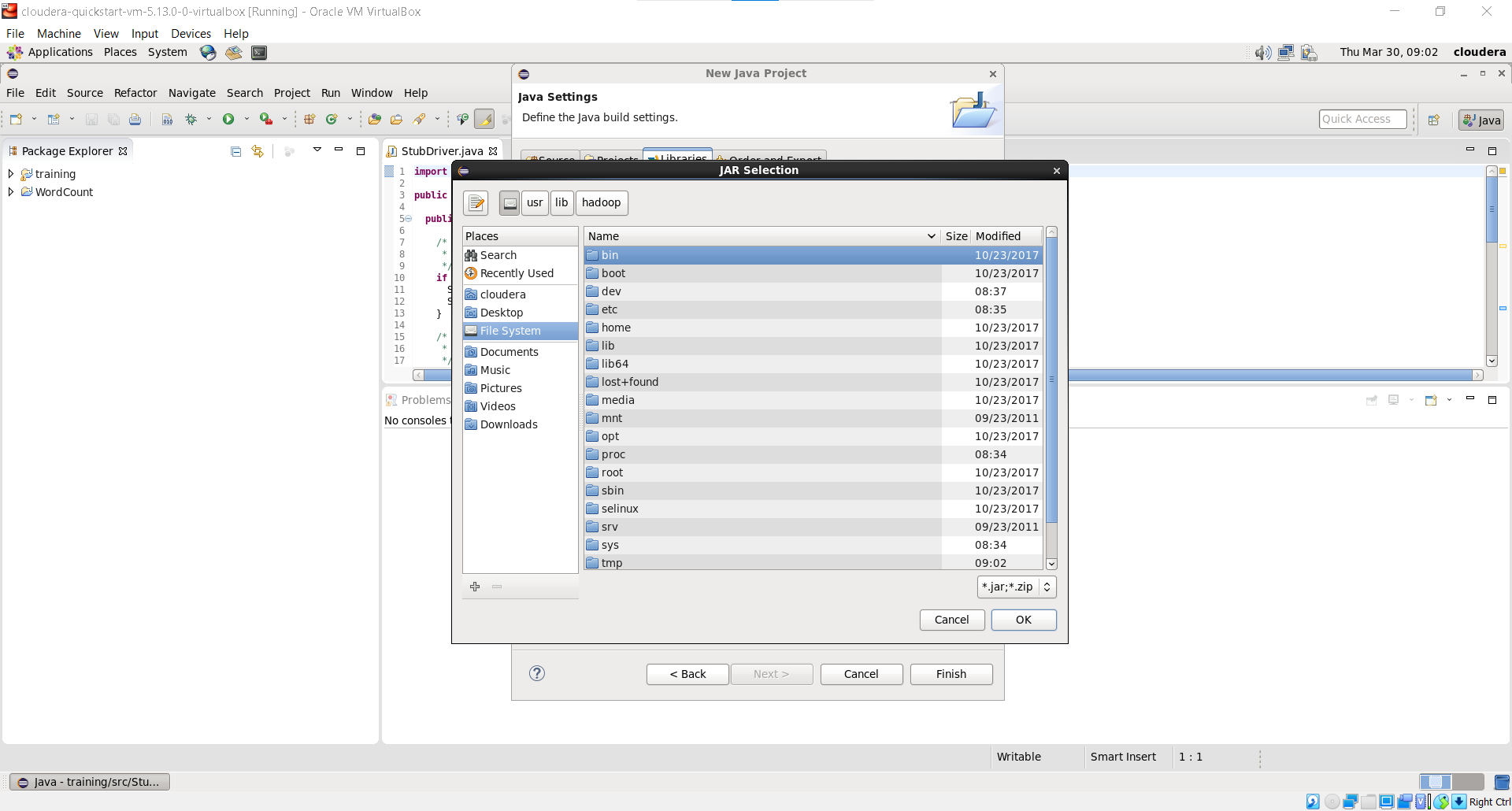
**Steps:**

1. File -> New -> Java Project -> Name: WordCount

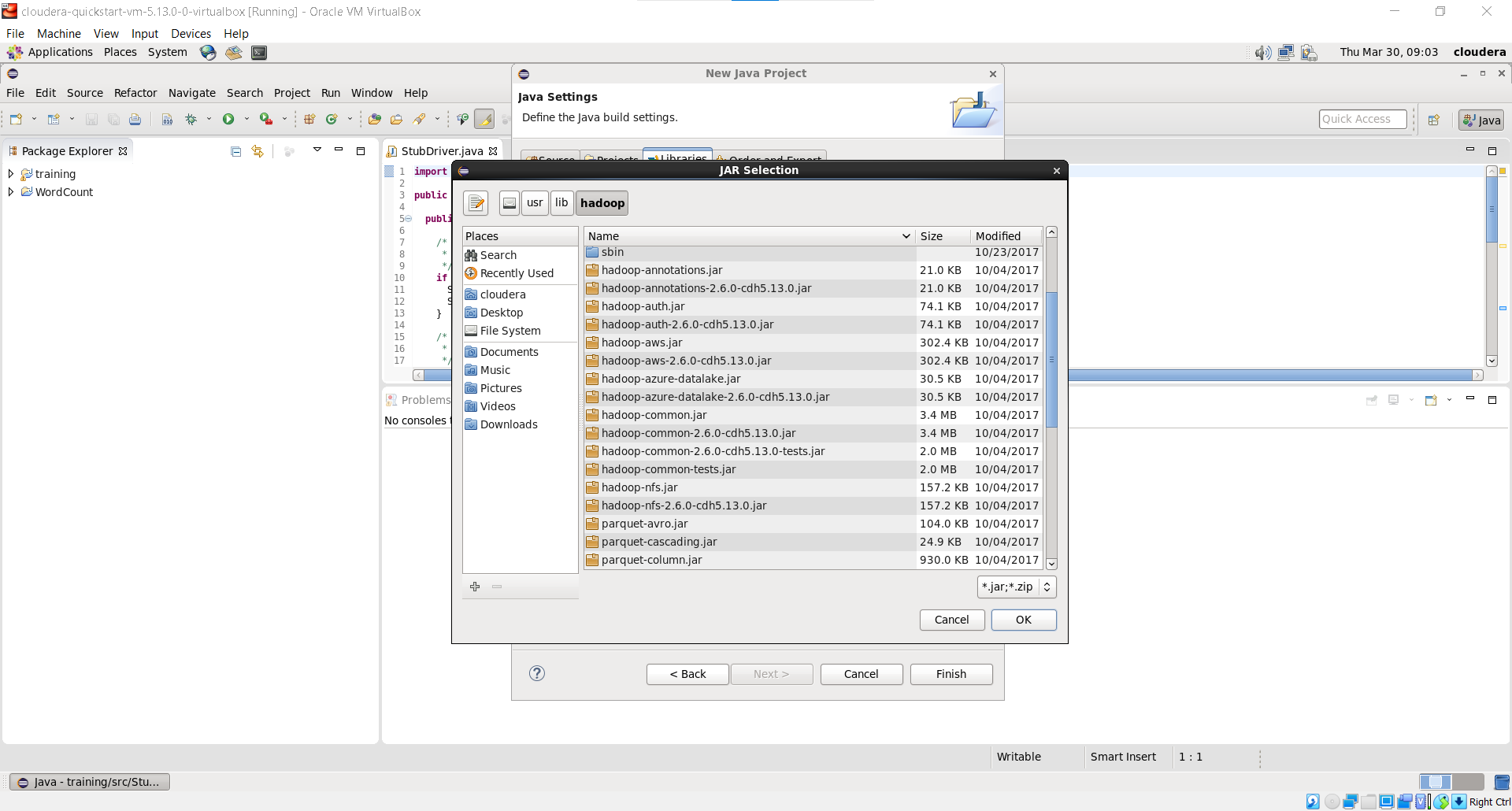


1. Add Hadoop Libraries to the Project

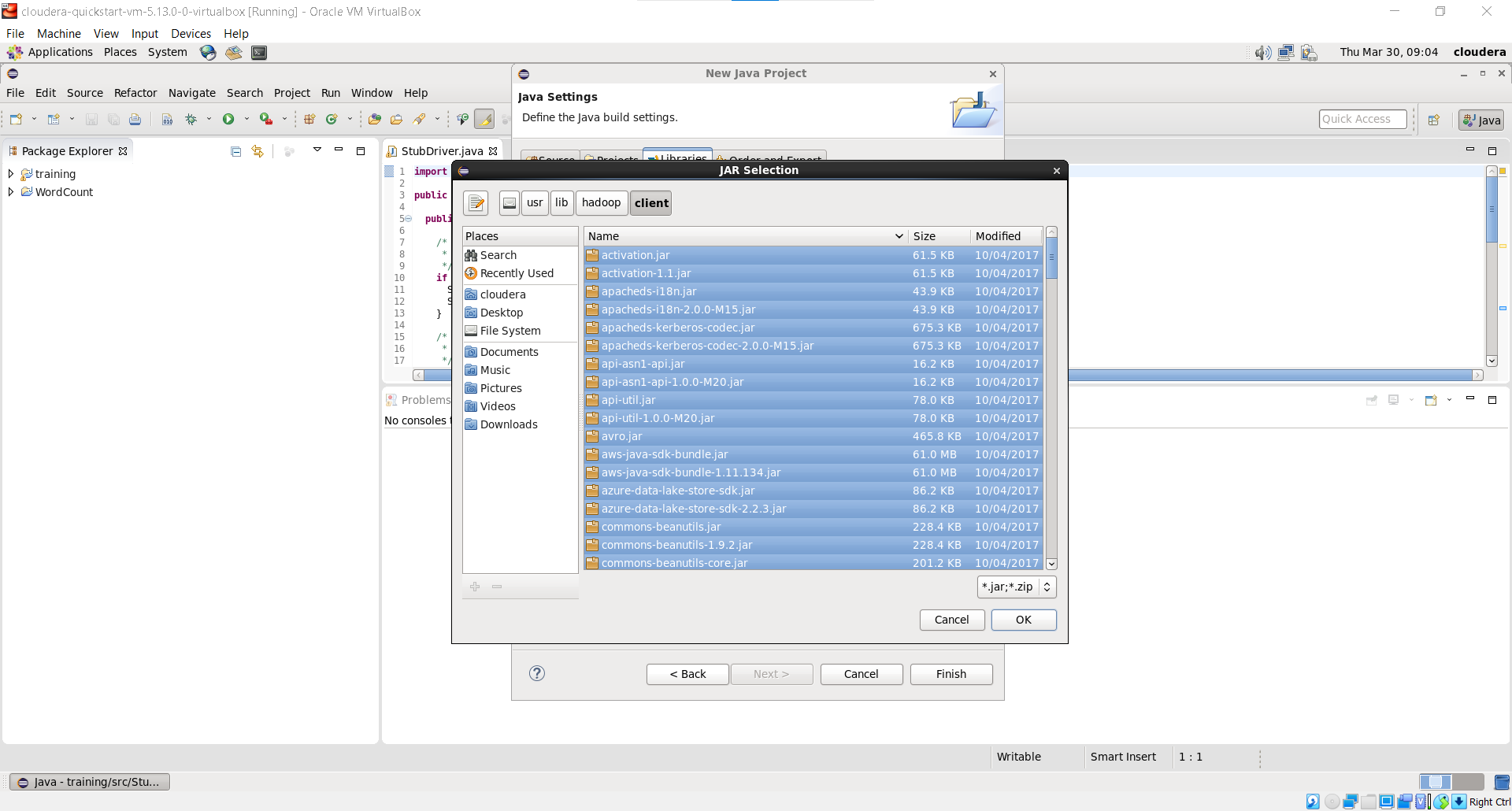




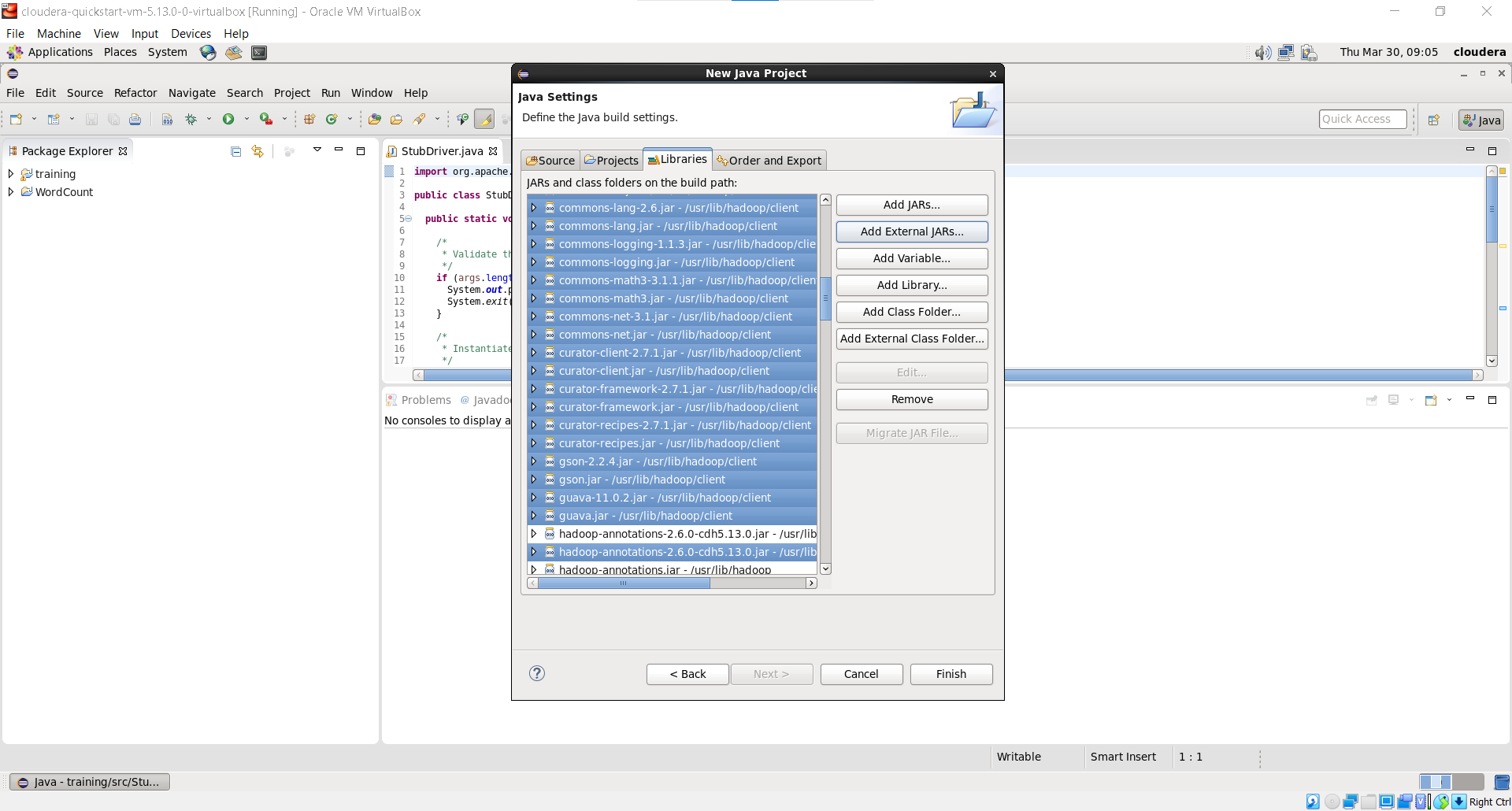
1. JAR Files -> usr/lib/hadoop/



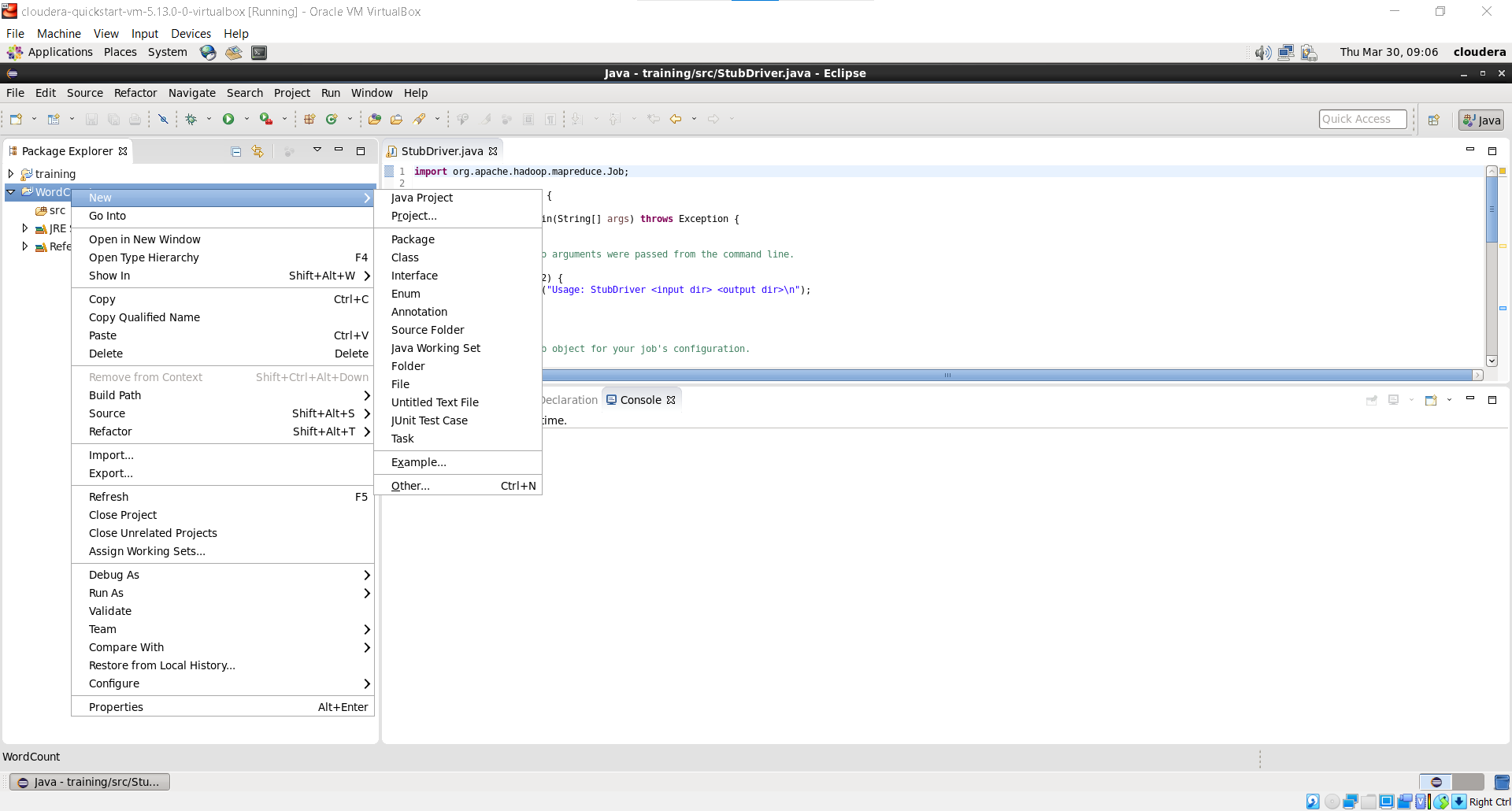
1. JAR File -> usr/lib/hadoop/client



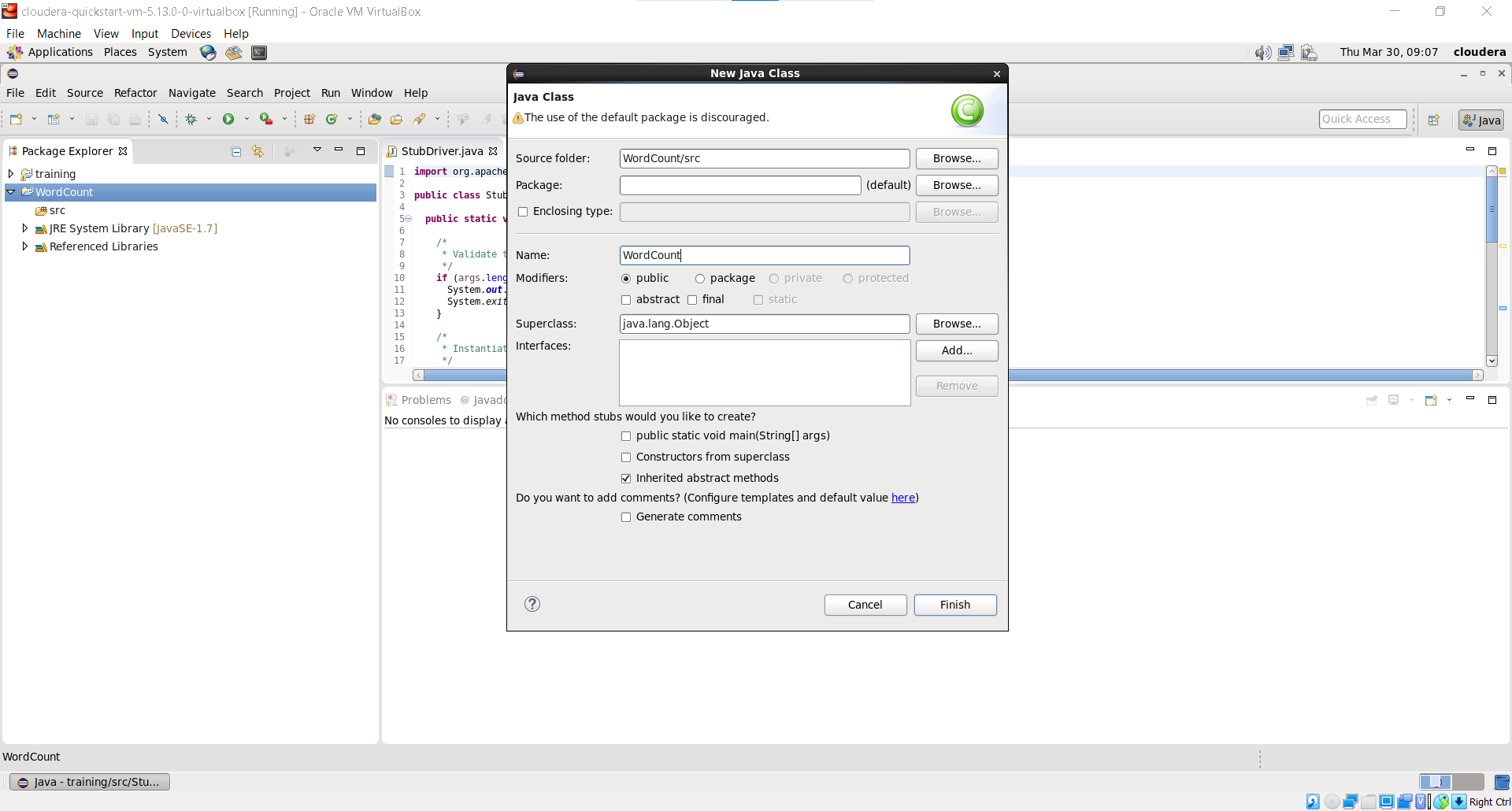
1. Added All Jar Files



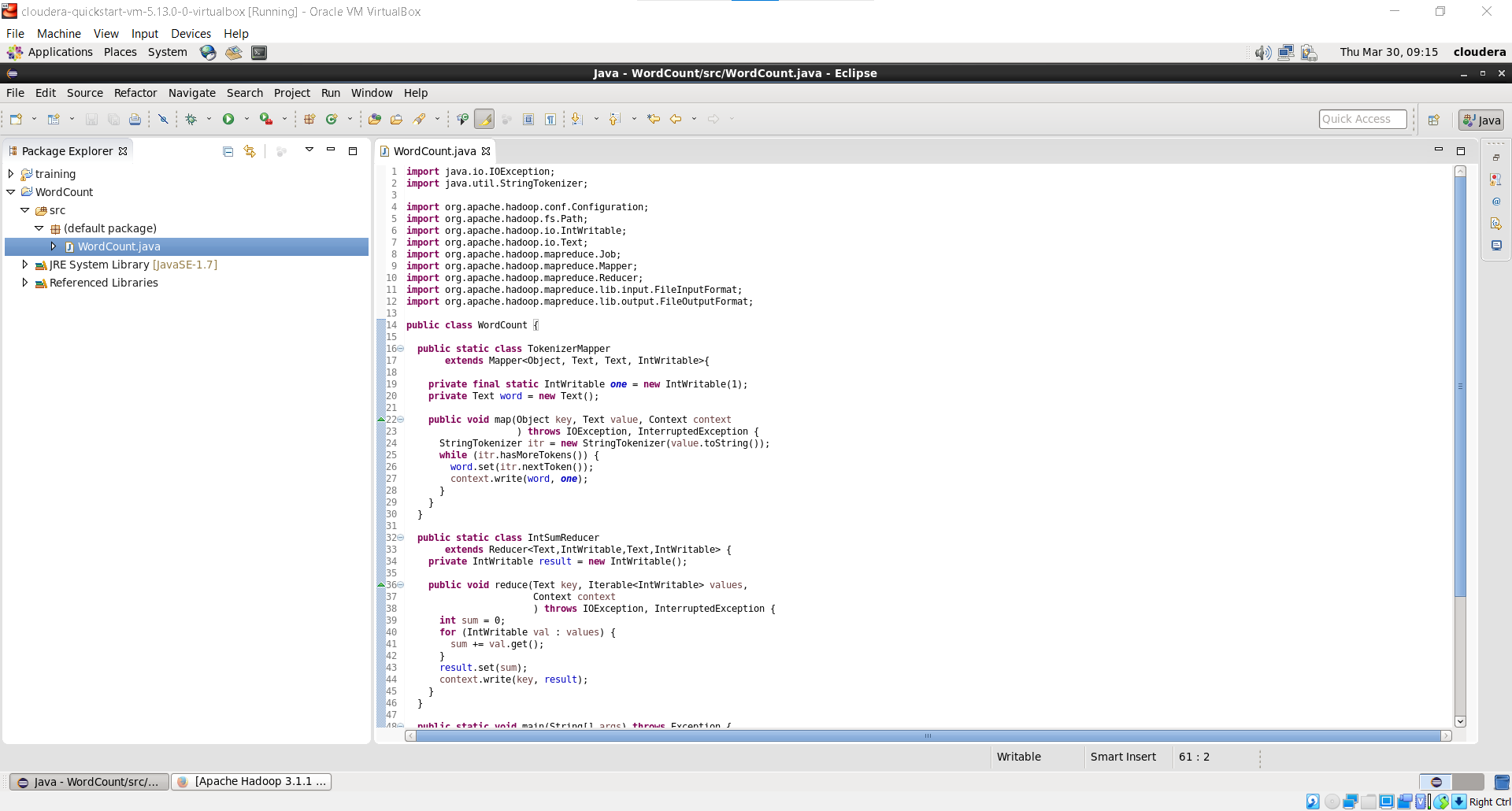
1. Create Class: Right Click on Project -> New -> Class



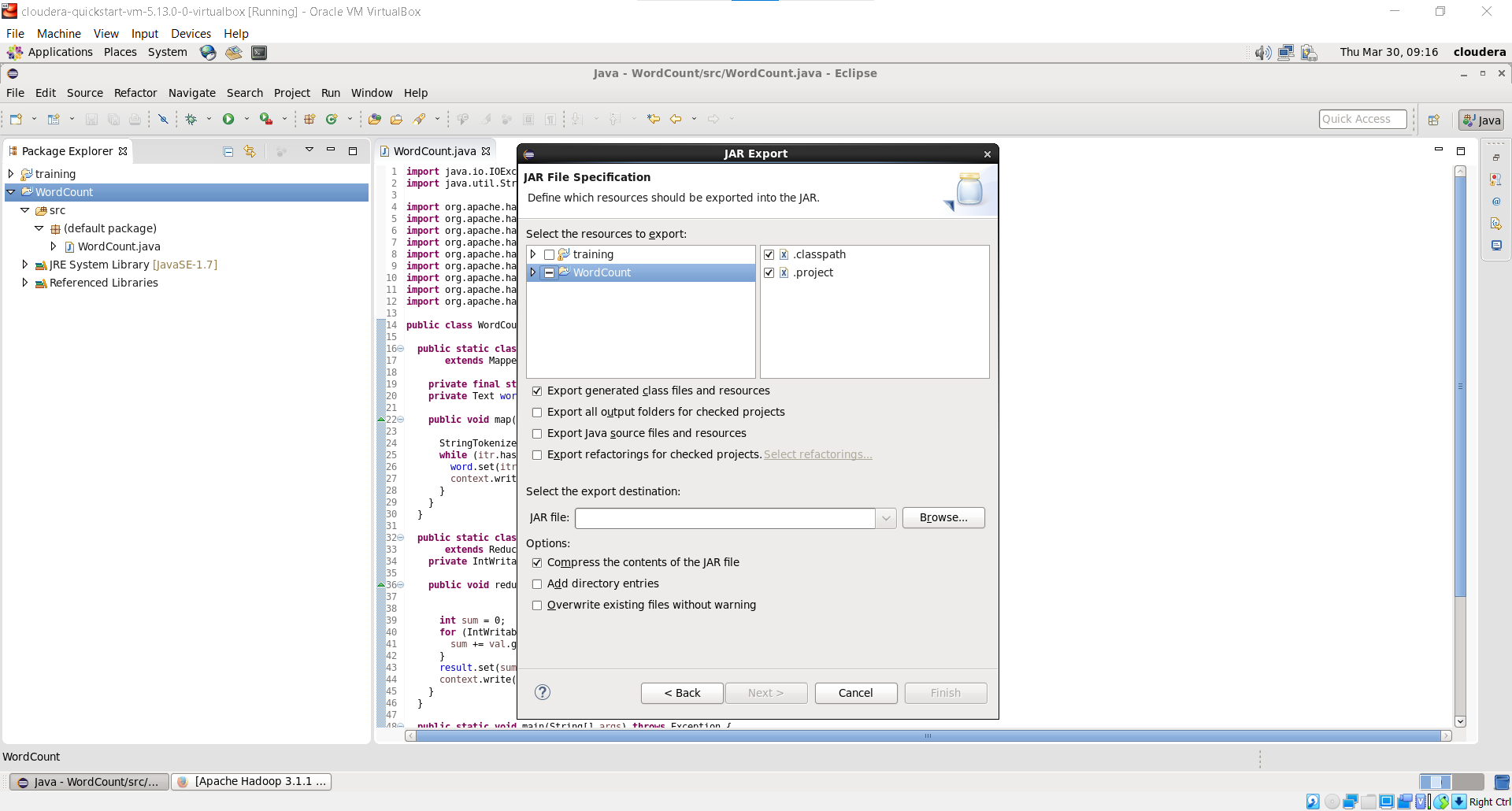
1. Give name to WordCount & Finish



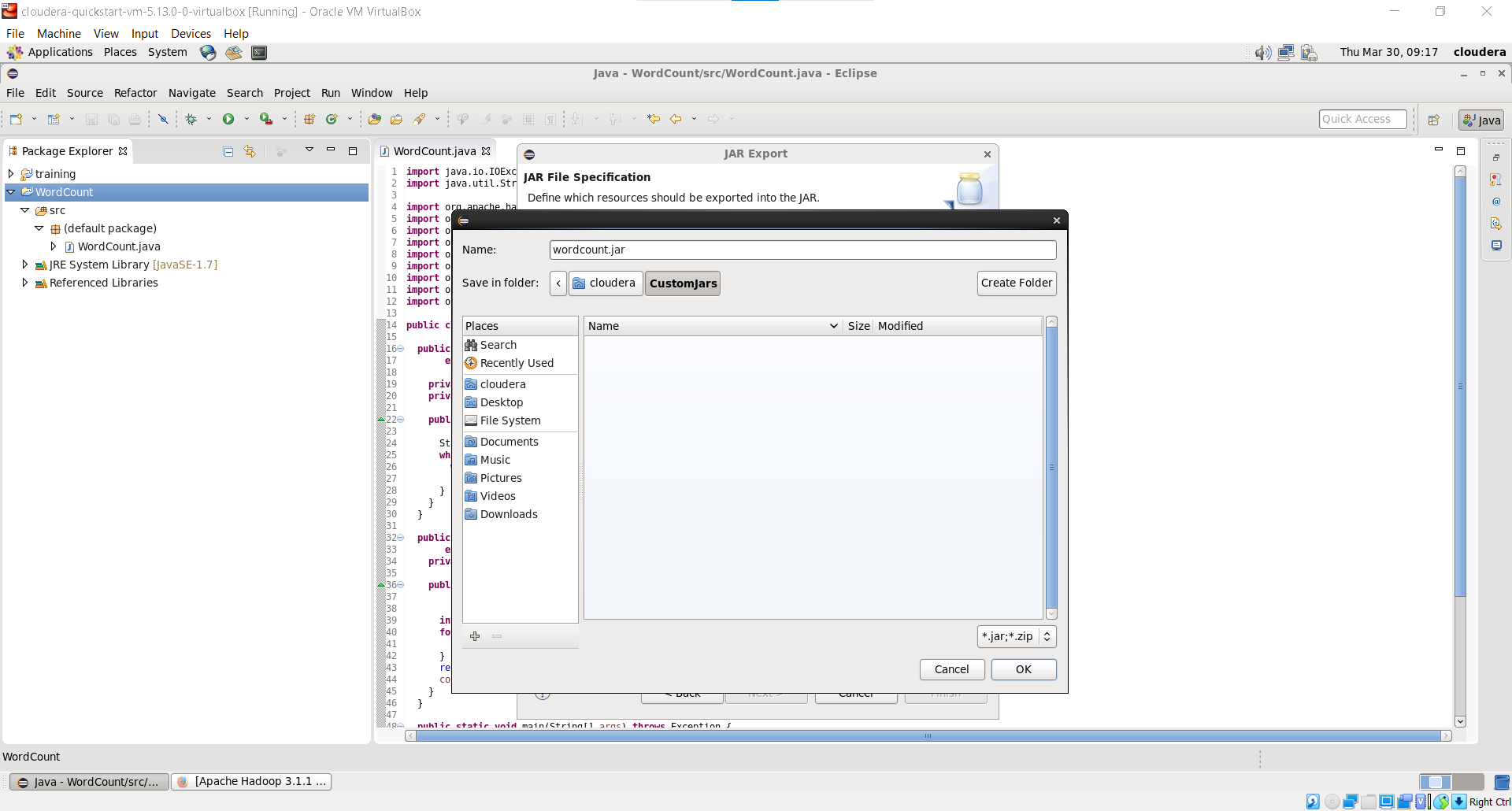
1. Write the WordCount Code



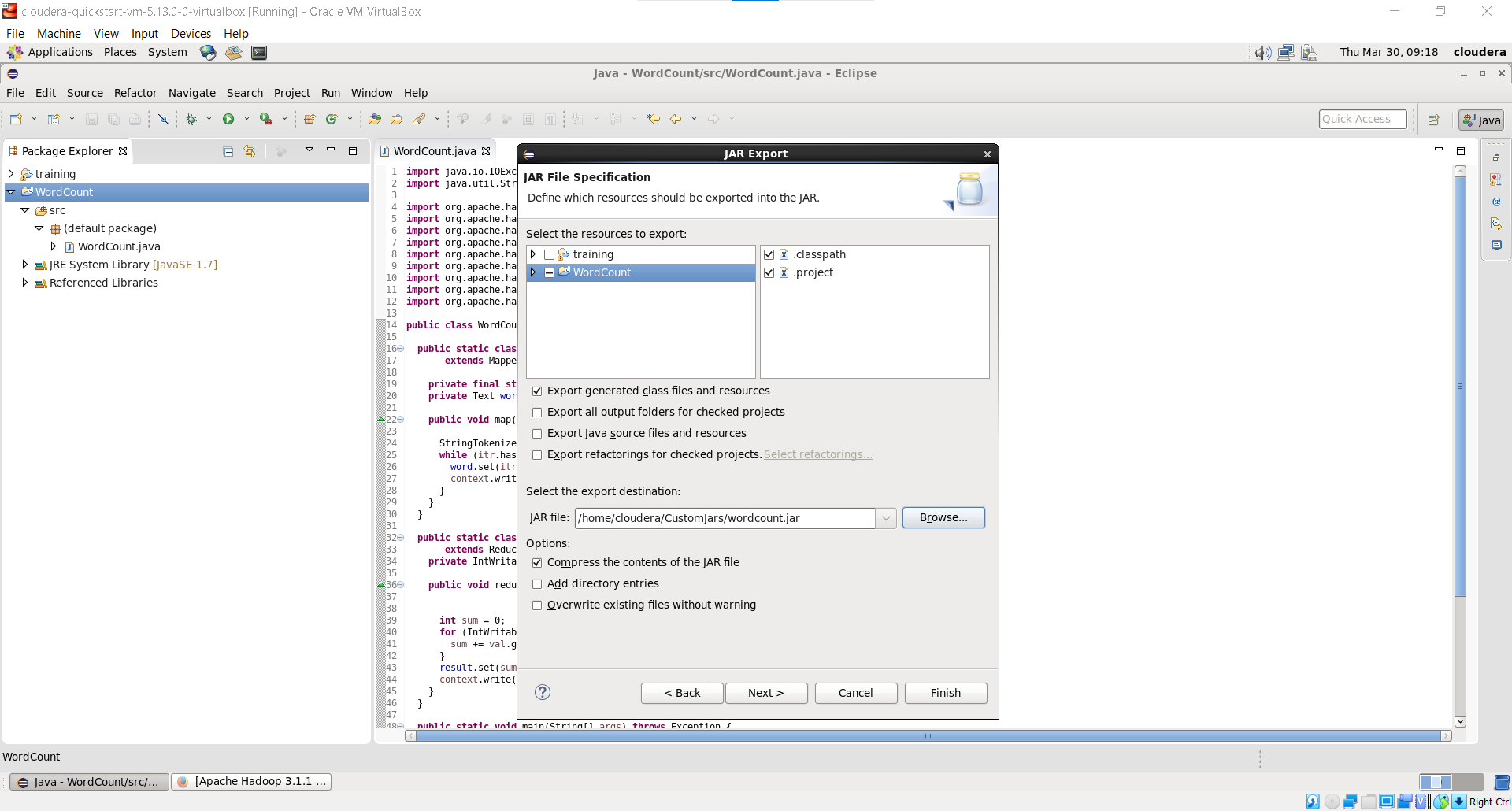
1. Right Click on Project -> Export -> JAR File



1. Created wordcount.jar inside CustomJars

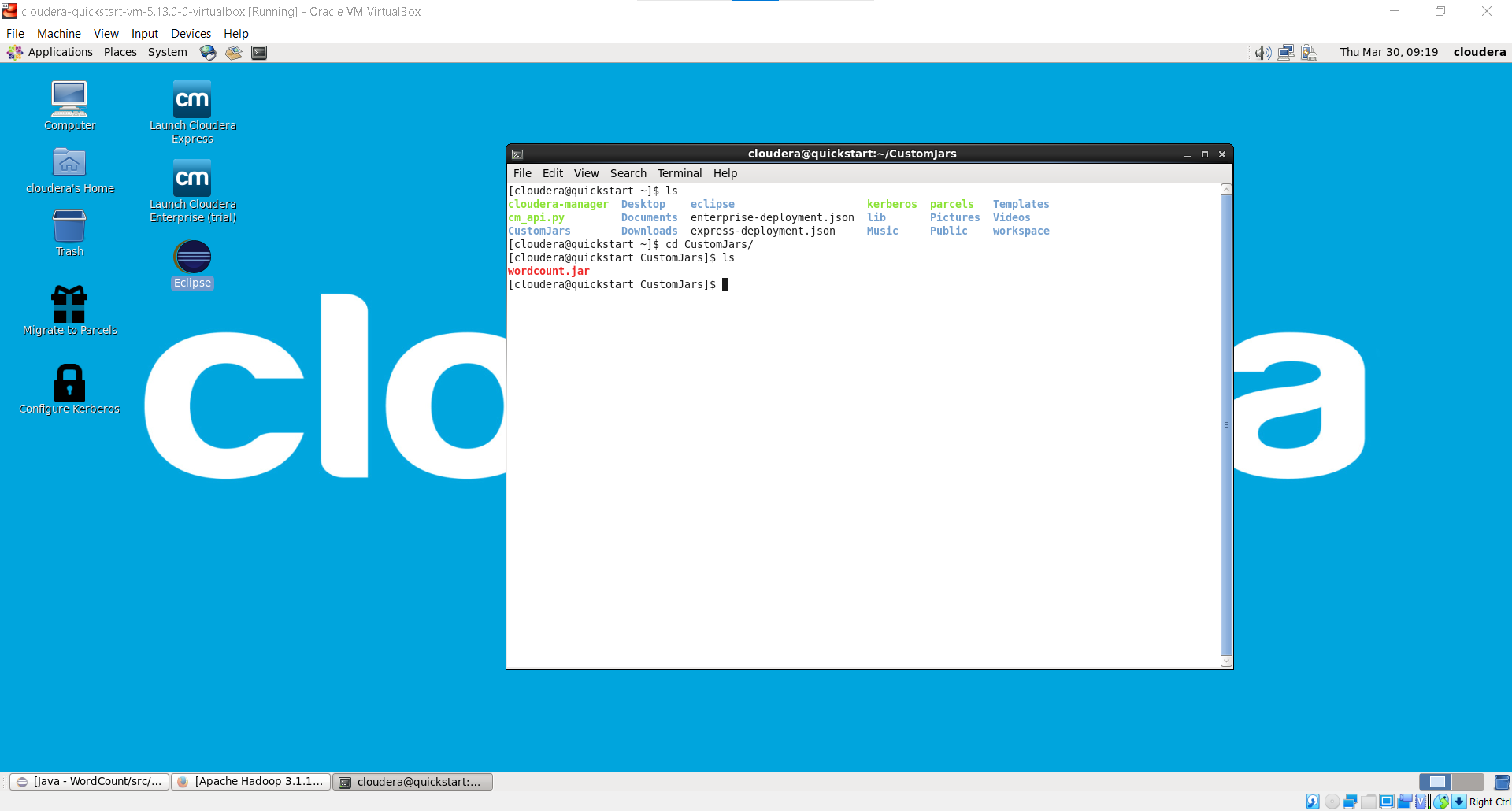


1. Click on Finish

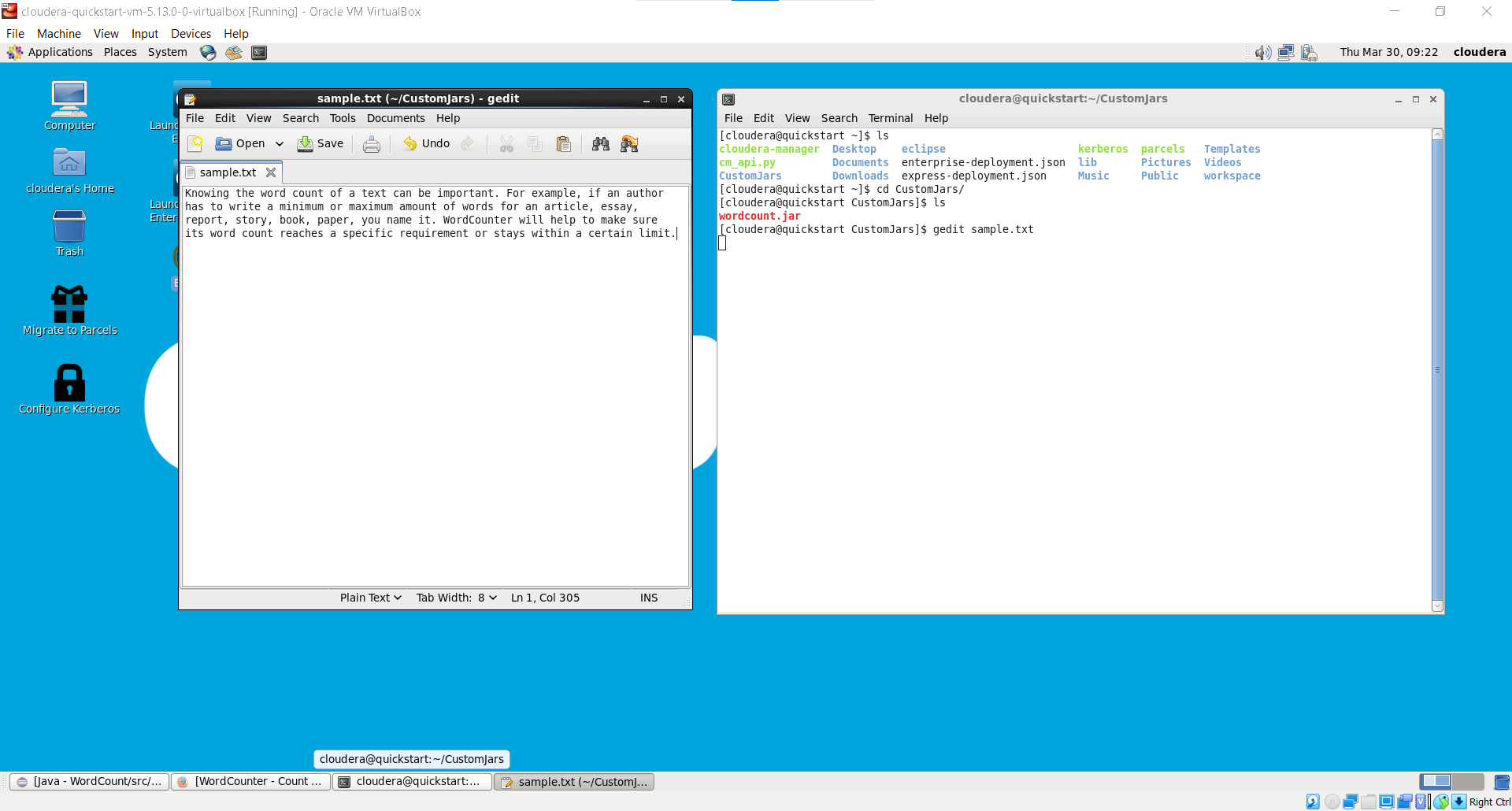


1. Basic Commands

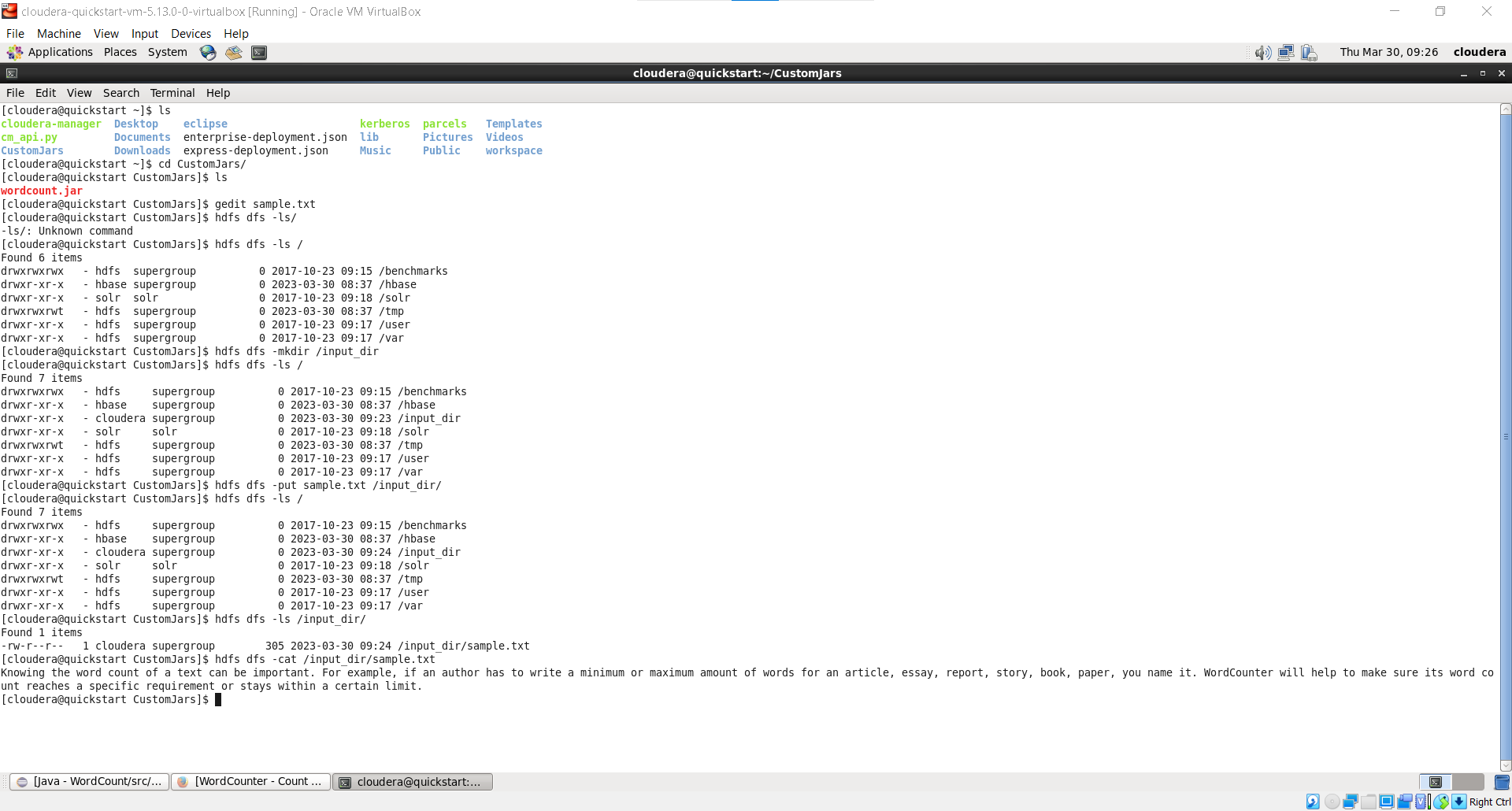
* listing all the directory present



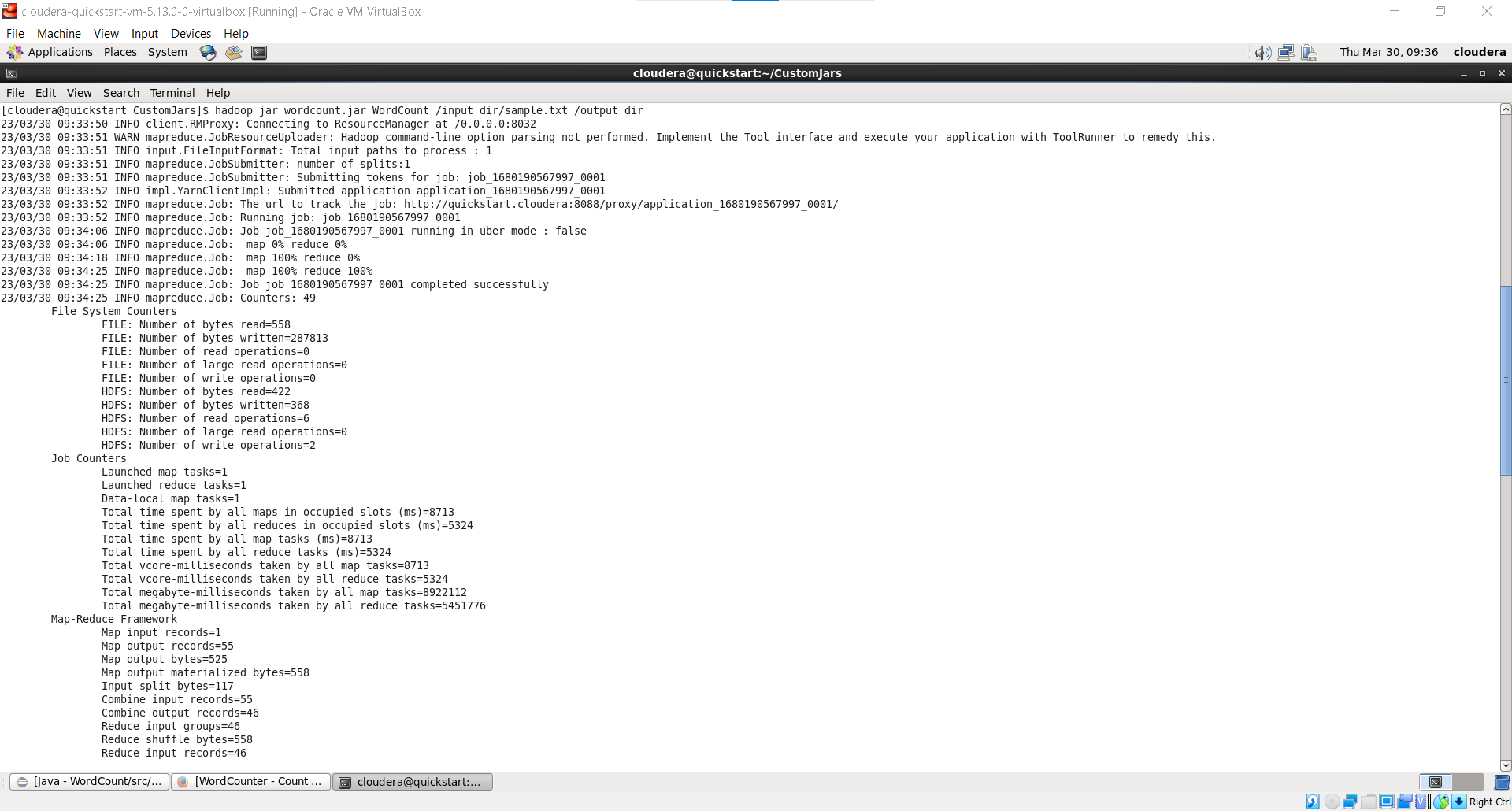
* Created File : sample.txt



* listing all the directory present in hdfs using ***ls*** command
* create directory in hdfs using ***mkdir*** command
* move sample.txt to input\_dir using ***put*** command
* ***cat*** command to print text present inside sample.txt



* run wordcount.jar on hadoop



* print the output

