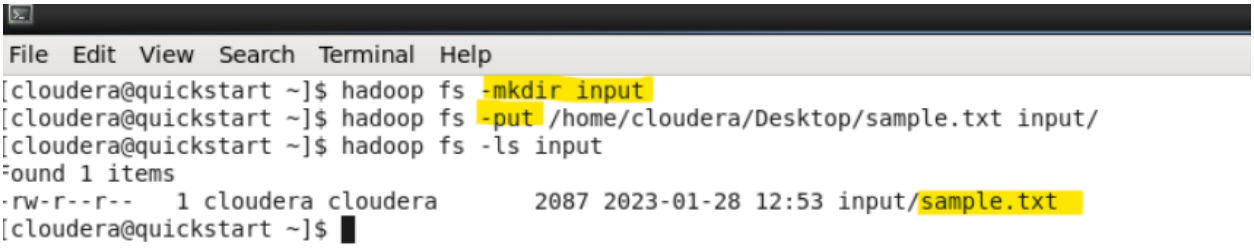
CSCE 5300 Section 006 - Introduction to Big Data and Data Science

Assignment 2

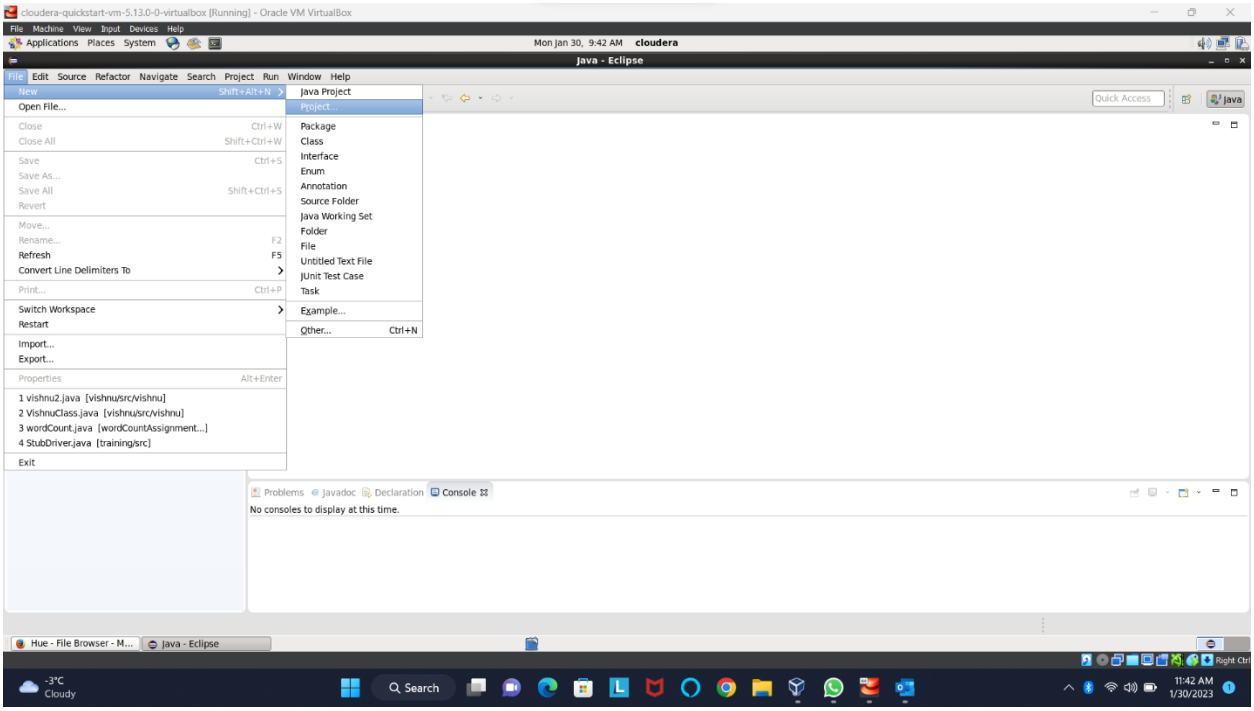
Downloaded the file from canvas and moved the file to hadoop by creating a folder named input using hadoop fs -mkdir

Hadoop fs -mkdir <folder name> is used to create a new directory in hadoop

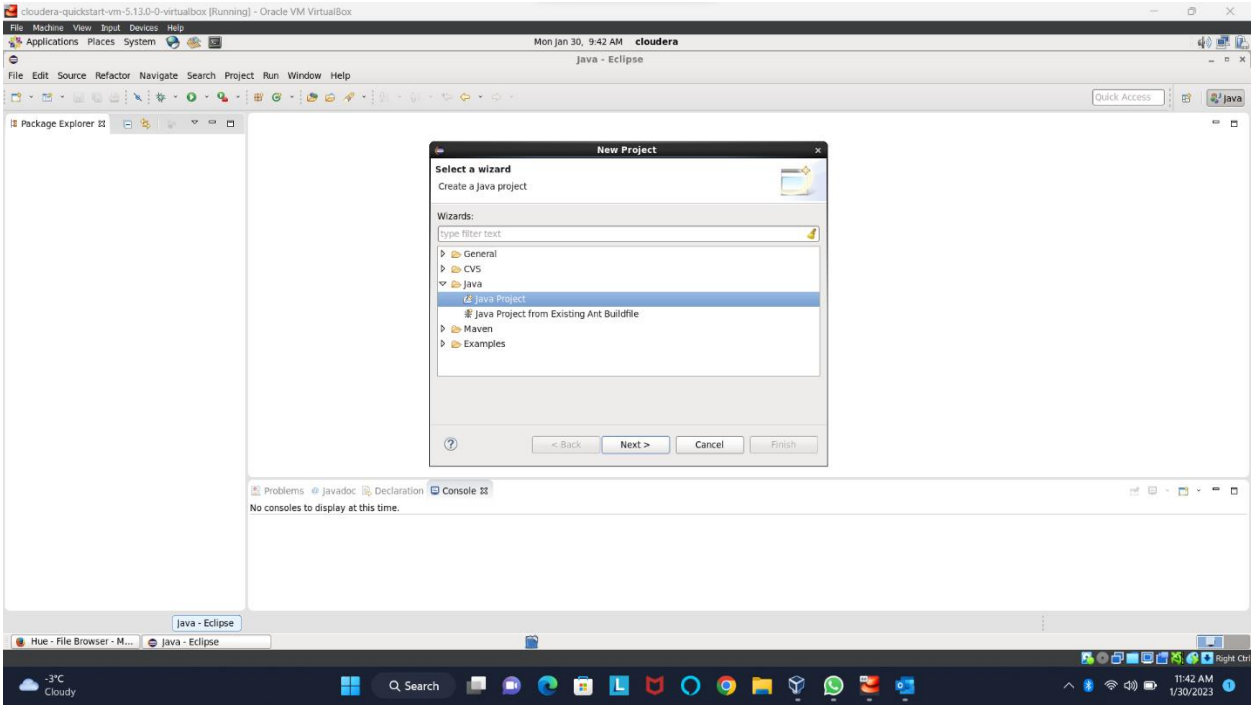


Creating a new project in Eclipse:

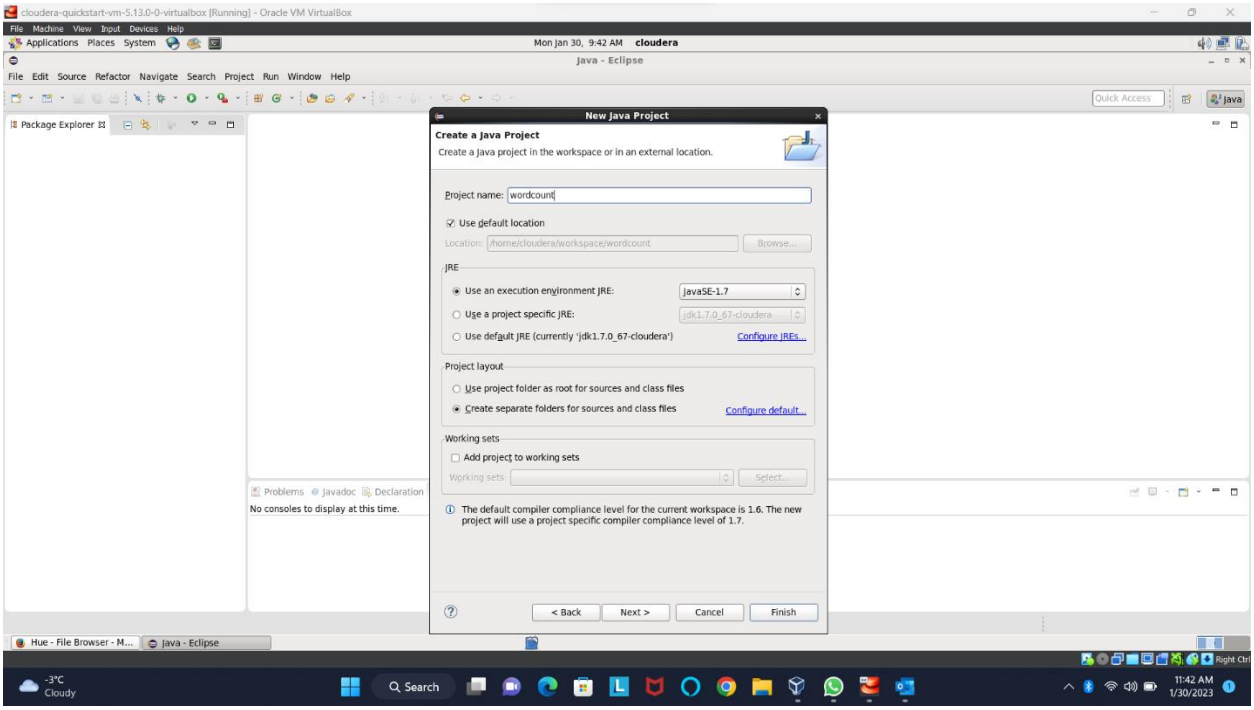
Created a new java project by clicking on file > new > Project



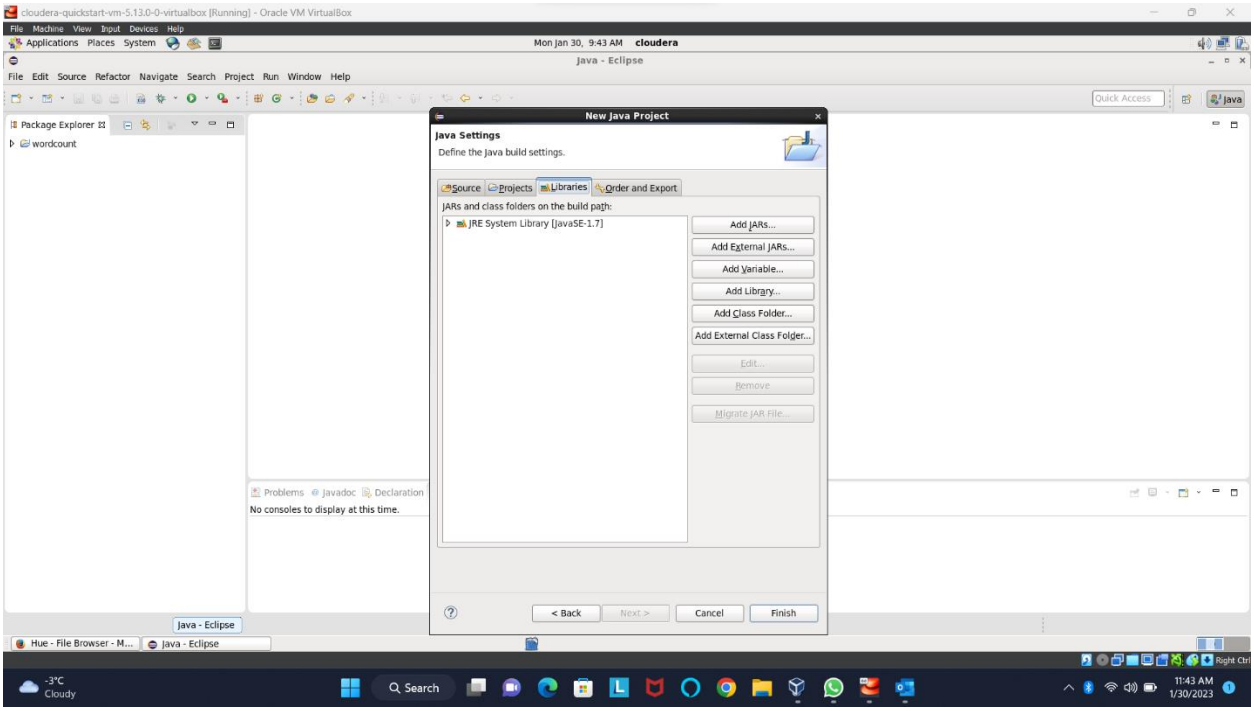
Select java project and click on next to the java project setup



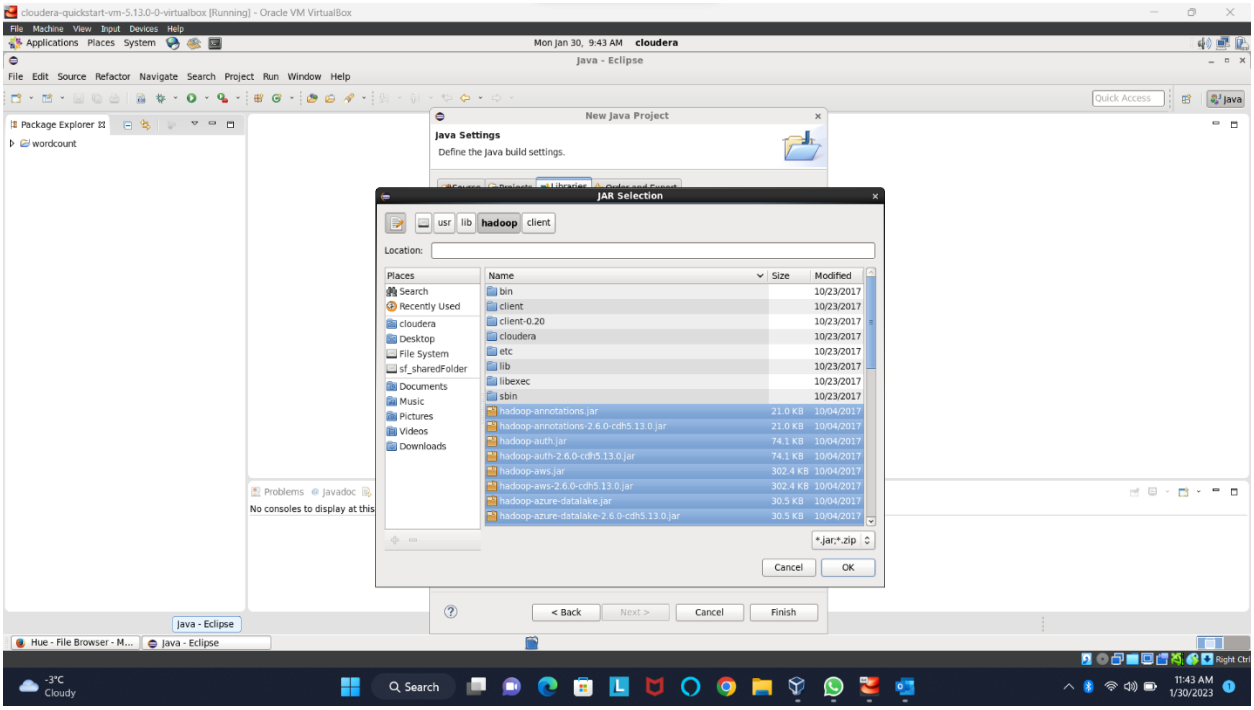
Give name of the project and clicked on next.



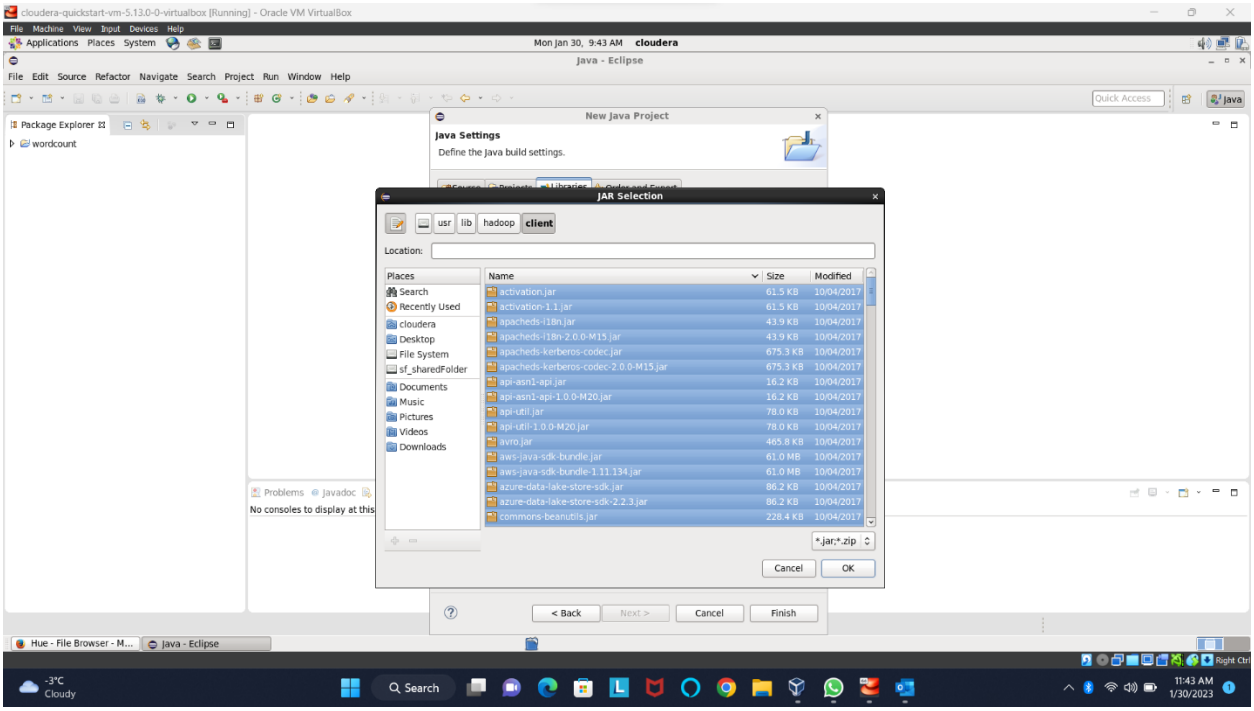
Select Libraries in the next pop up menu, New java Project Settings, and click on Add External Jars



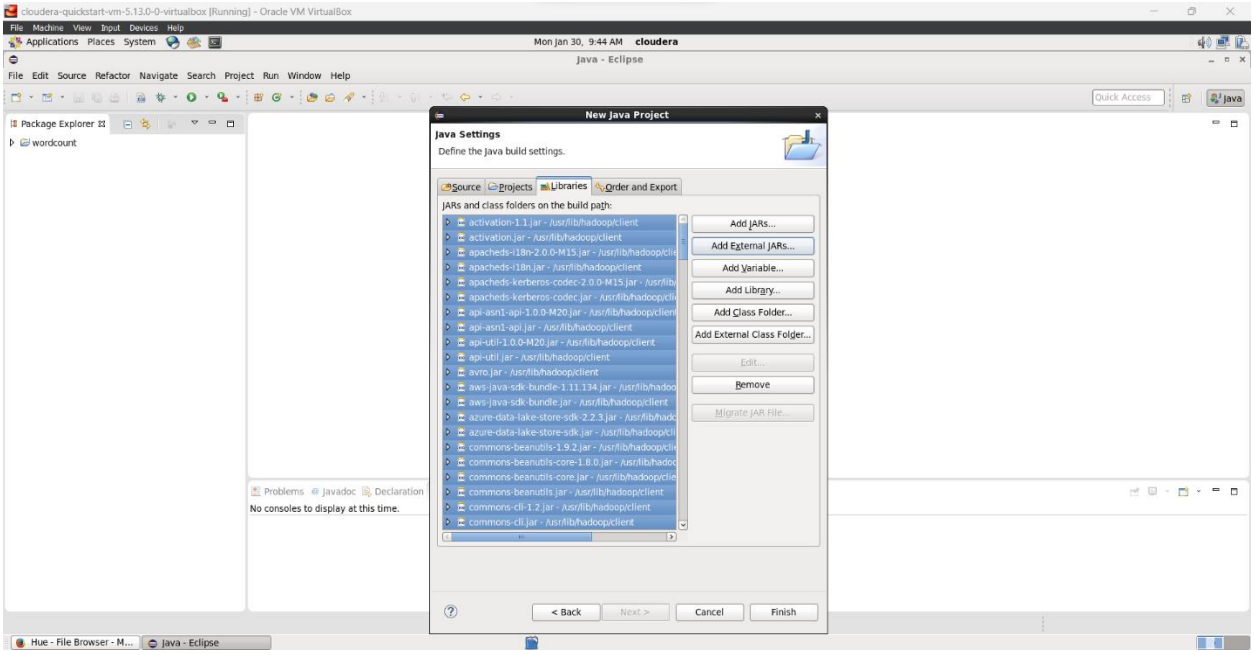
Select all the jars under hadoop folder and add them to jars by clicking on ok button.



Also select all the jars from the bin folder in hadoop and add them to the java jars by clicking on OK button.



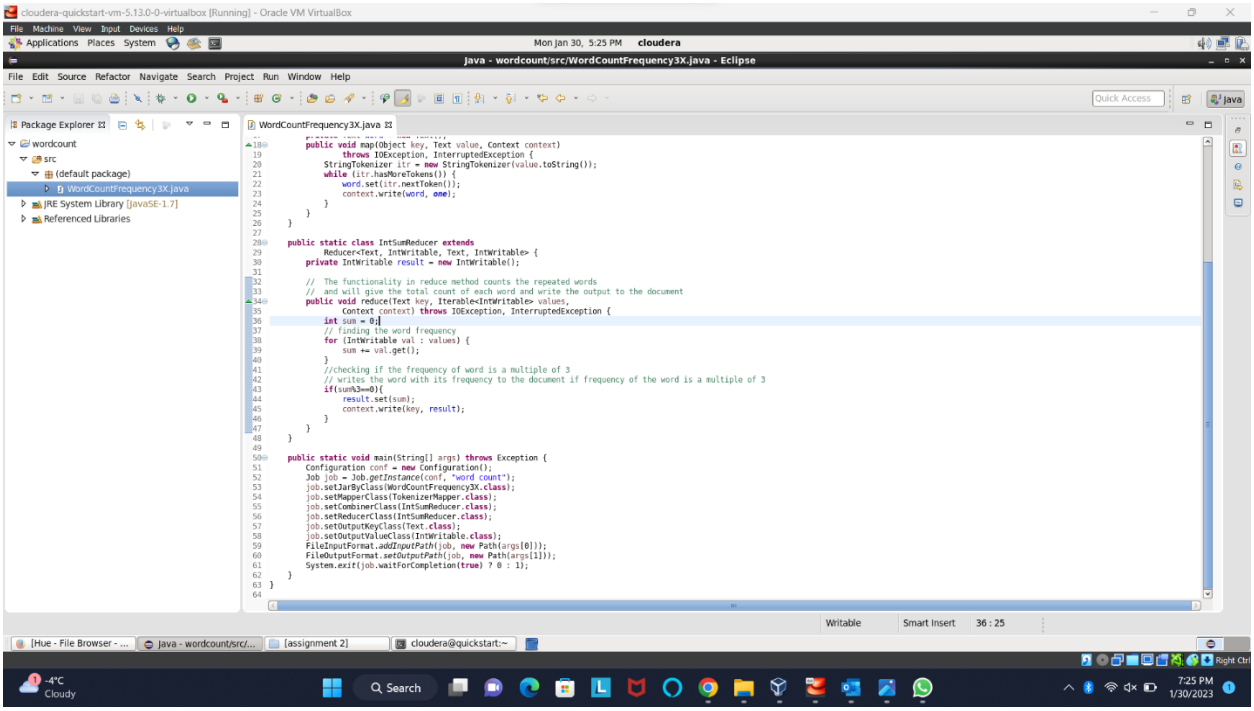
After selecting and adding all the necessary jar files to the java project, finish the project setup by clicking on finish button in the popup menu



Task 1:

Finding all the words that has the frequency of multiples of 3 in the given sample.txt file.

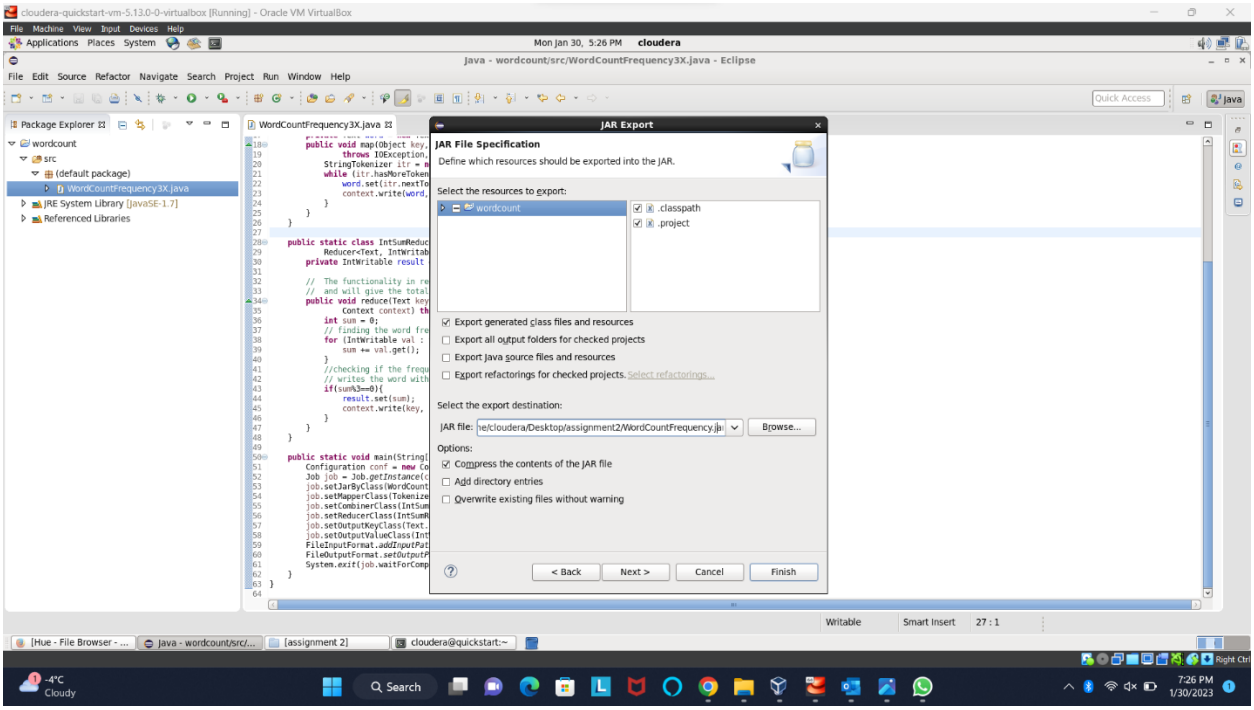
Got the sample code from canvas and add necessary logic to generate the file that has all the words that are repeated in multiples of 3s.



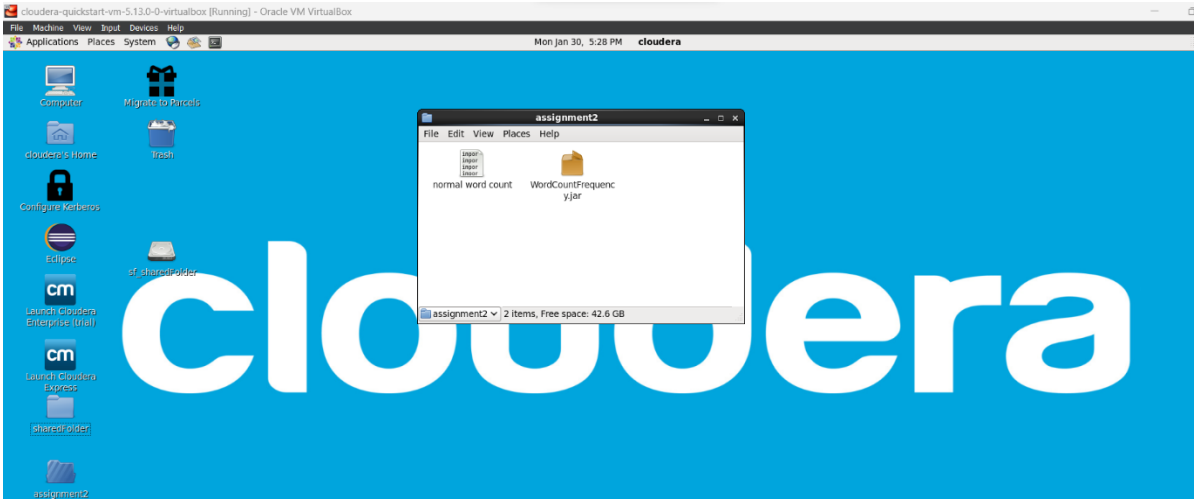
Added the login in IntSumReducer class. The purpose of this class is to count the number of occurrences of words in the file. Here an if condition is being added after 42 to check the count of word. If the count of the word is a multiple of 3, then the if condition in line 43 gets succeeded and the lines inside the if condition will gets executed. These lines will appended the word and its frequency to the the text file that will be generated at the end.

Exporting the project and creating a jar file

By right clicking on the project name on the right navigation and selecting properties in the menu, a new pop up for jar export will appear on the screen. Select java jar file export and then provide the location for with name of the jar file and click on finish.



A new jar file at the prescribed location is generated as shown in the below figure.

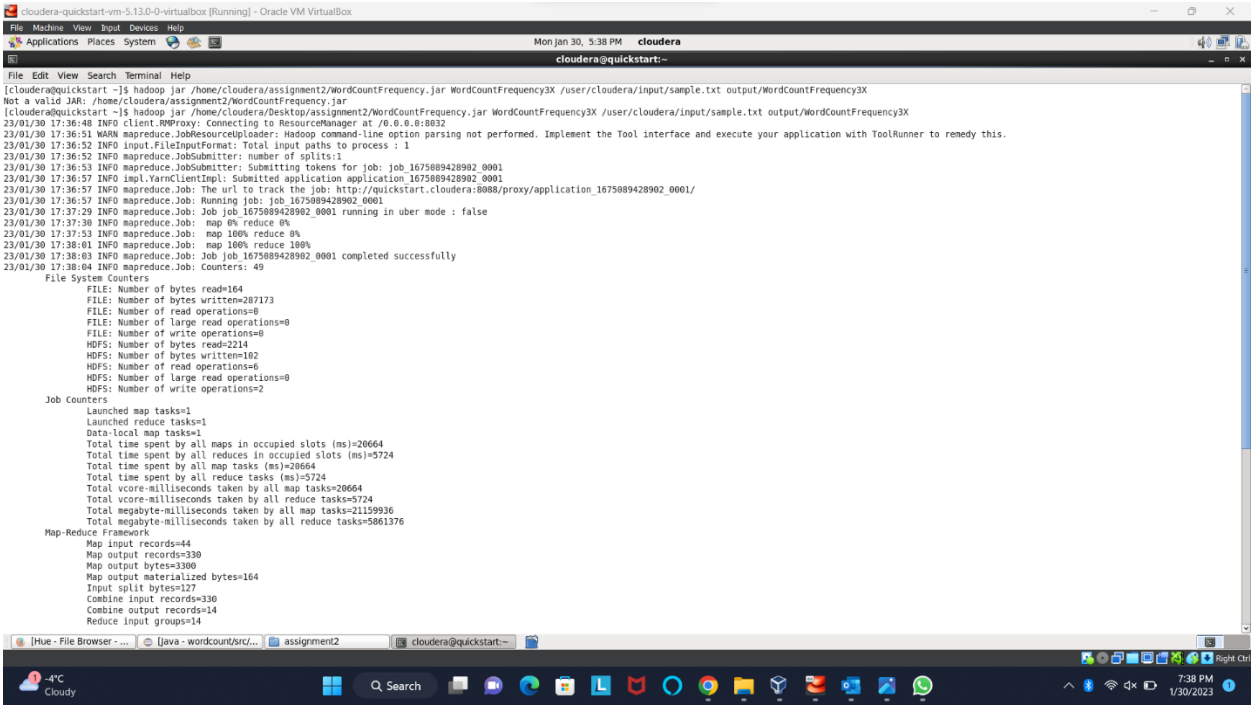


Running the Map Reduce job in the terminal.

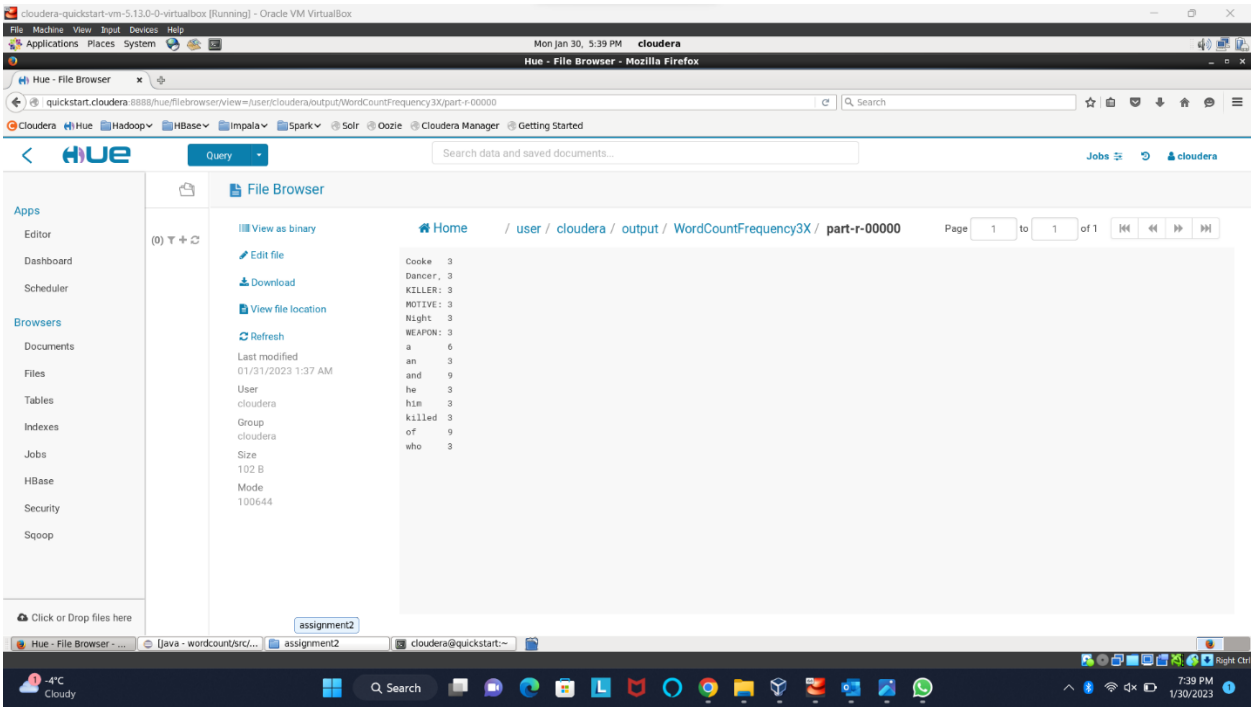
To run the jar file in the terminal the following command is being used.

hadoop jar <jar file location><space> <class name of the java file in jar><space><input file path><space><output file path>

On giving the command in the above format, the java program along with hadoop will pick the input sample.txt file from the given input file path and runs the map reduce job in that file and generates an out file at the given output location.



The below picture is the visualization of the newly generated file in hue.



Task 2

Finding all the words which ends with letter ‘S’ and their frequency.

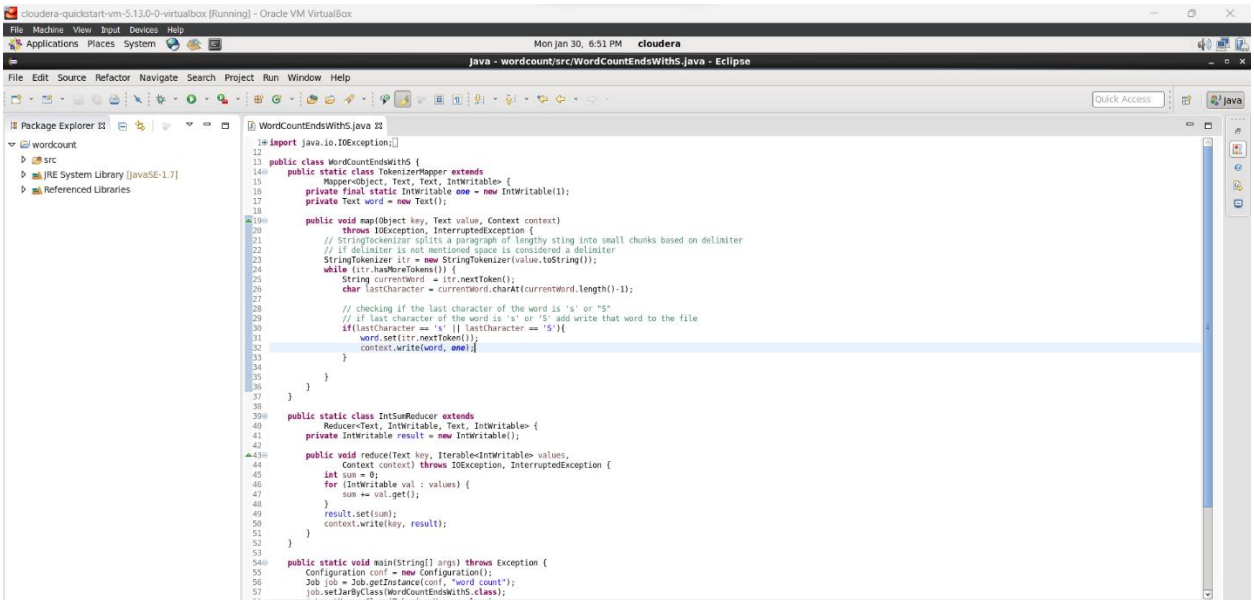
In the provided java code, all the words in sample.txt that are ending with letter words are found out and are added to a file and save in hadoop.

In the logic, StringTockenizer is used to break down the paragraphs in the file to words.

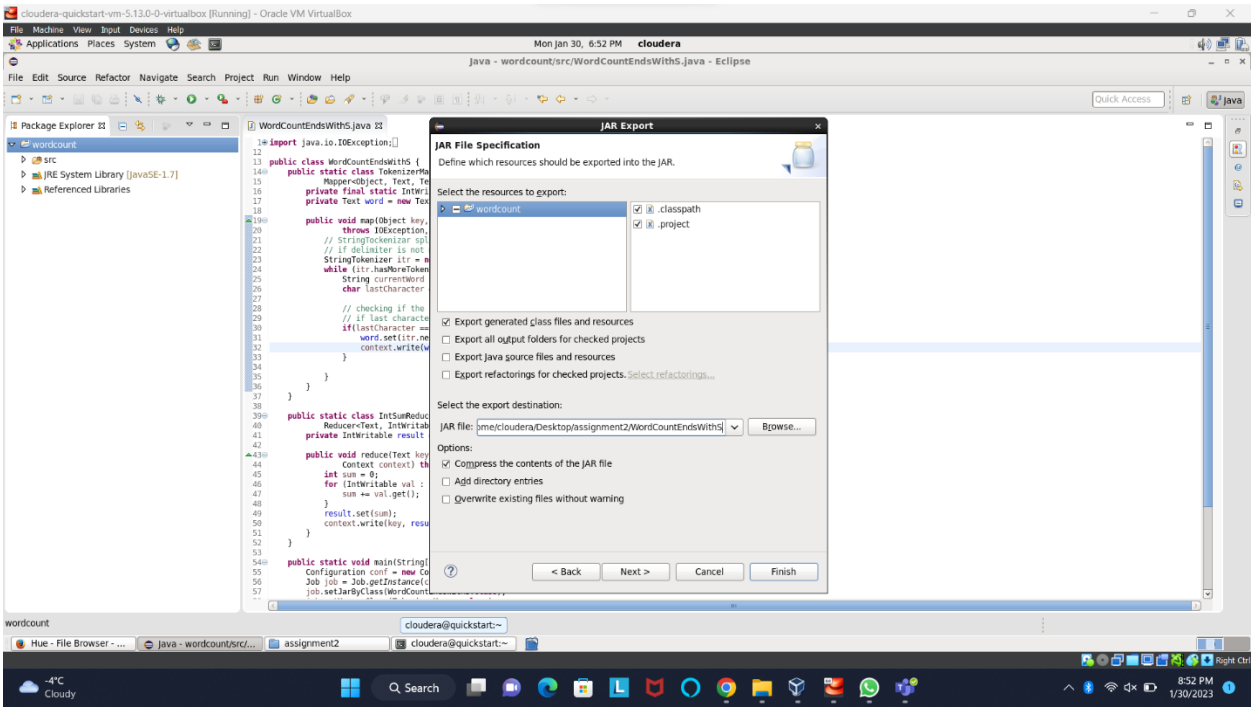
In line 25, checking availability of the word, if the word is available, the condition will turn true.

In line 26, getting the last letter in the word using charAt(lengthOfWord -1)

In line 30, checking if the letter is a ‘s’ or ‘S’. if they match, those words are sent to reduces and added to the file.



Exported the jar file and ran in terminal.



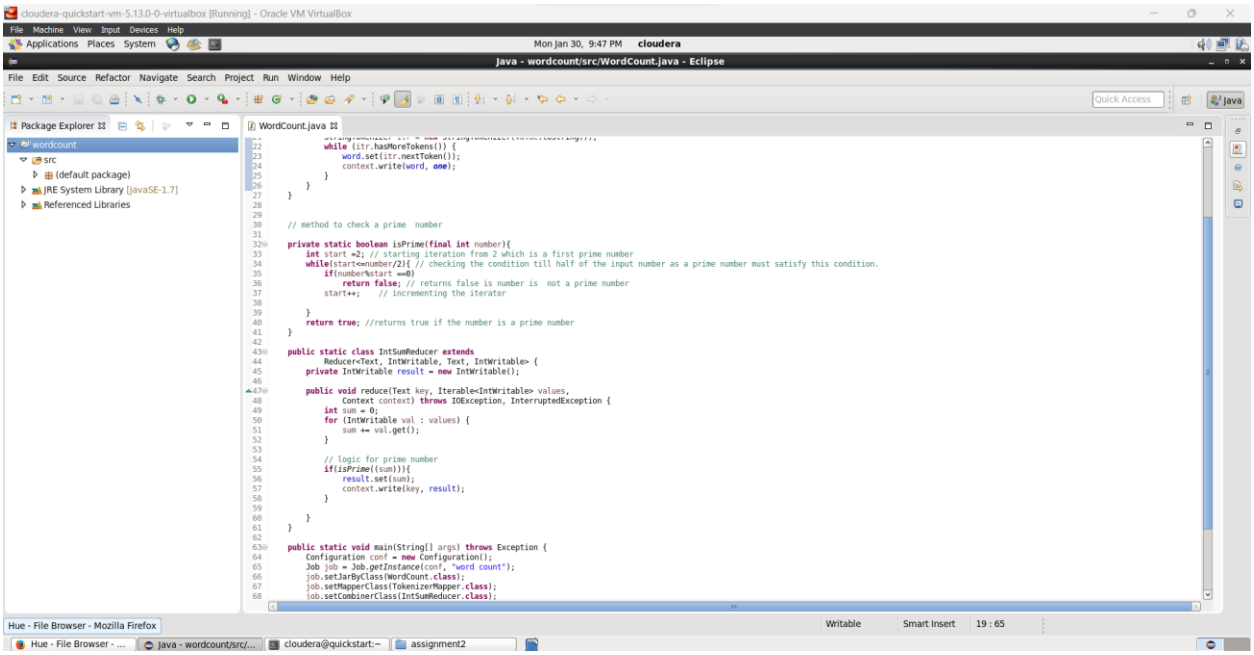
Task 3

Finding the words that are repeated prime numbers of times.

In the below program, logic is being added in the IntSumReducer class.

In line 55, checking if the sum is a prime number. If it is a prime number, the word and its frequency will be added to the output file.

This file is exported to a jar and ran in terminal to generate the output file in hadoop.



Generated output files for all the 3 tasks.