**HW 6: Using MapReduce for word count in JAVA**

**Following the example in the class, implement a MapReduce to do word count in Java code. You have to use a source data file different as the one used in class. Show the detailed process in a Word file.**

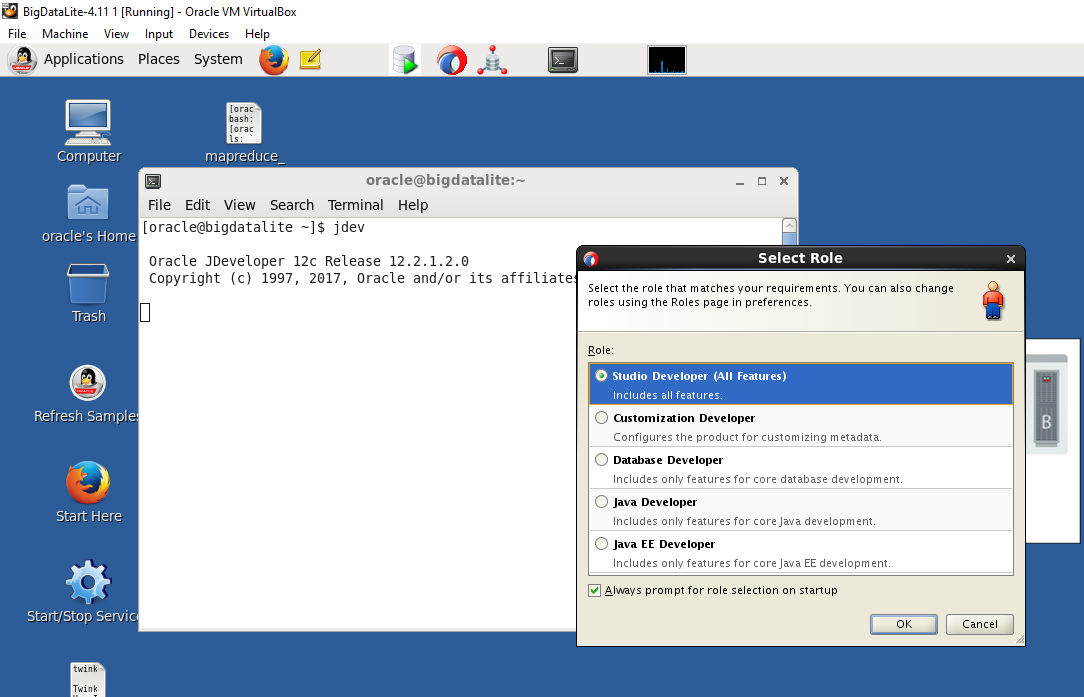
**Answer:**

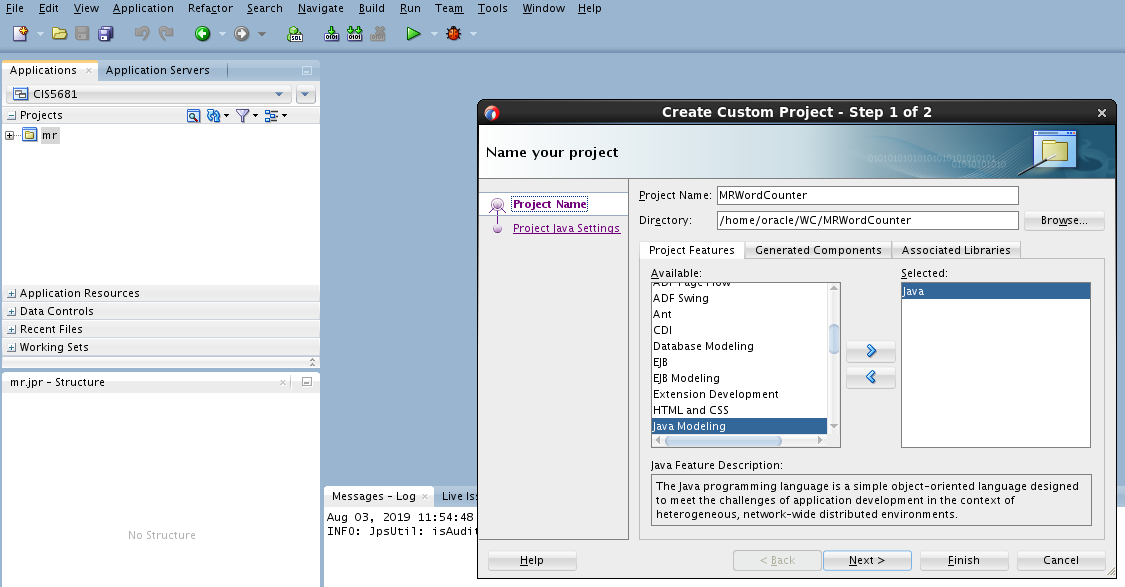
Following steps are followed while creating MapReduce interface in Java:

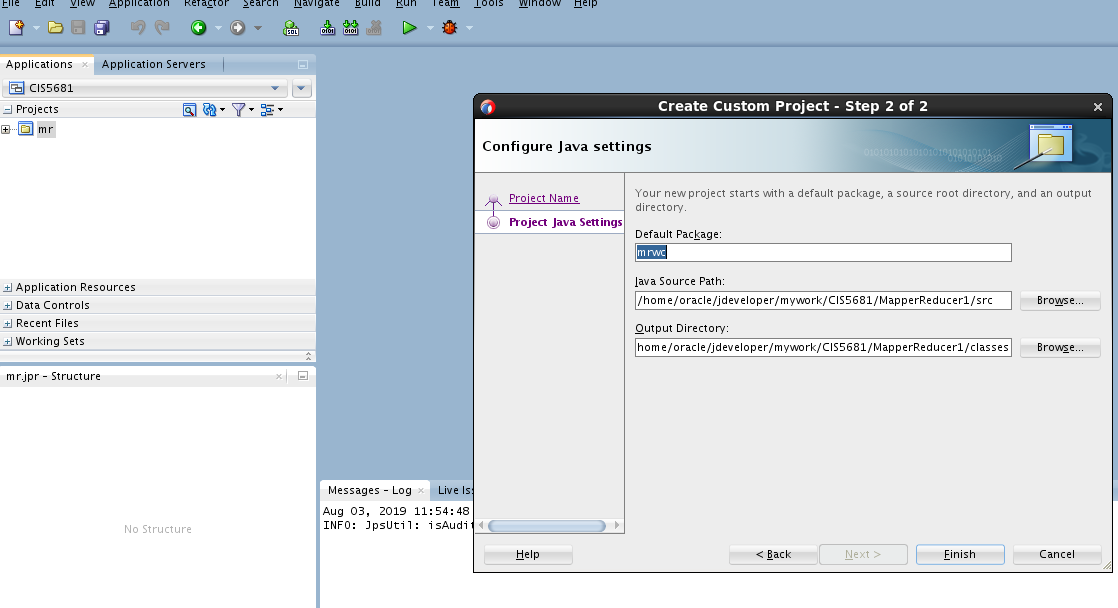
1. Create Java project “MRWordCounter” using Jdev
2. Hadoop Mapreduce Word counter has 3 phases, ie Mapper, Shuffler and Reducer
3. Create 3 java class files ie. WordCounterMapper.java, WordCounterReducer.java and WordCounterjob.java
4. In WordCounterMapper.java file , the input .txt file is tokenized into words to form a key value pair with all the words present in the input text file. The key is the word from the input file and value is ‘1’.
5. After execution of Mapper, shuffler is executed automatically. It sorts the key value pairs generated in mapper phase automatically.
6. In WordCounterReducer.java file, all the key value pairs are aggregated. In this phase, output of shuffler is input and all the keys are grouped together and values of similar keys are added up to find the number of occurrences of words

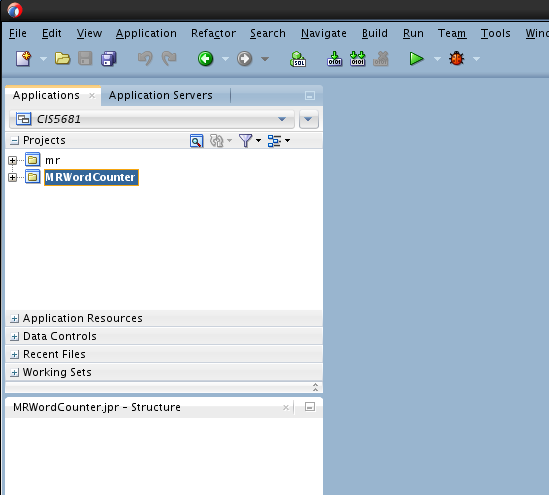
**Word Count example in Hadoop using Java API:**

1. Create the java project MRWordCounter

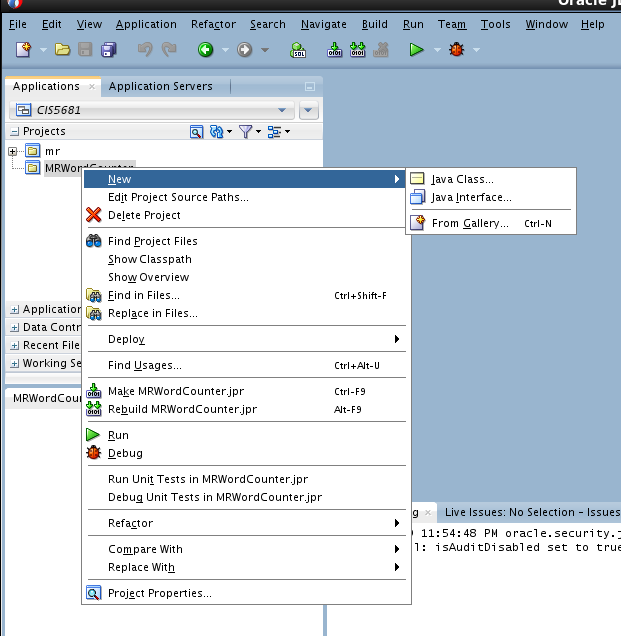


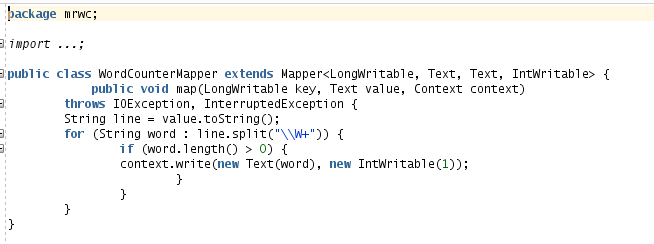




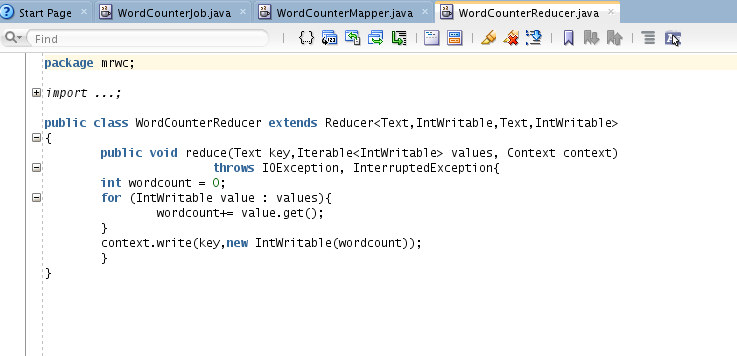


1. Create a new java Class WordCounterMapper.java. This is mapper class

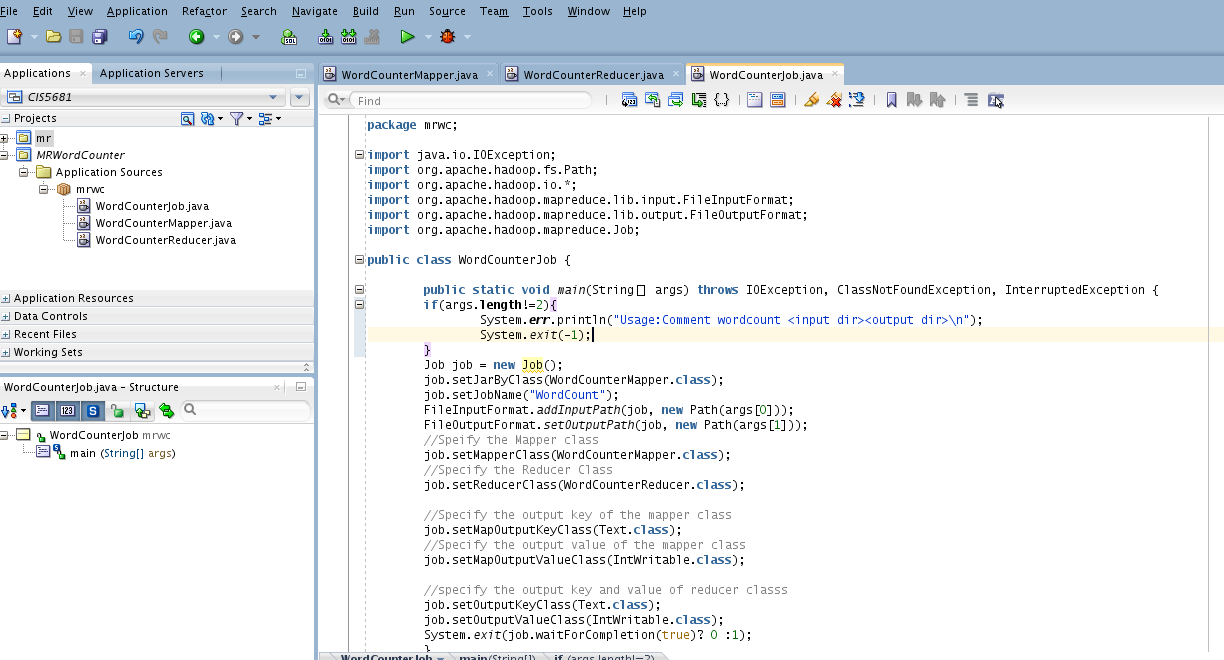




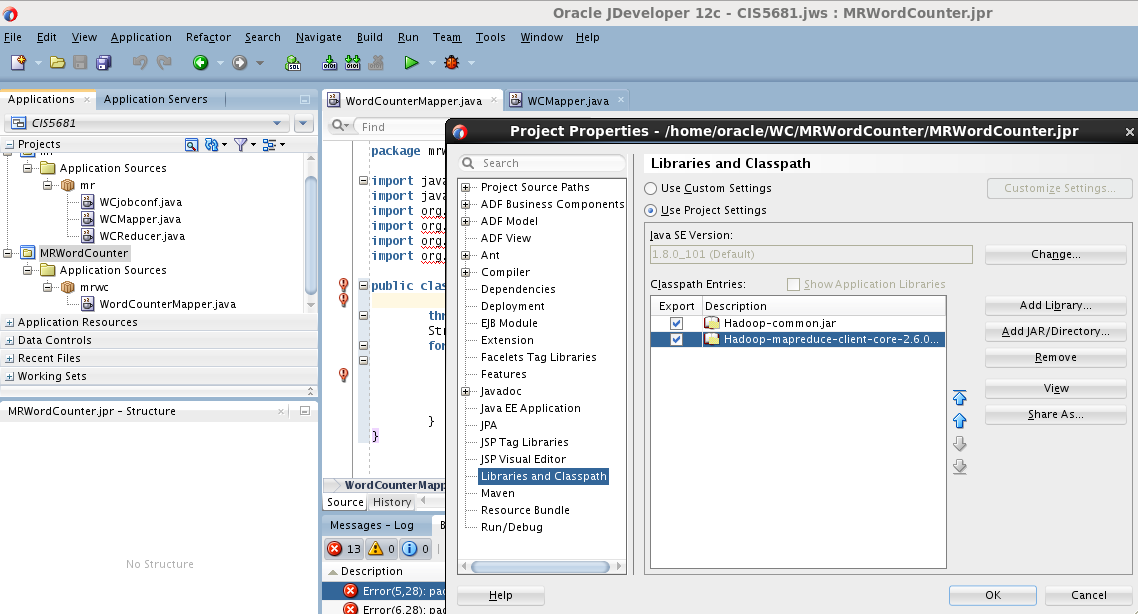
1. Create a new java file WordCounterReducer.java. This is the reducer class.



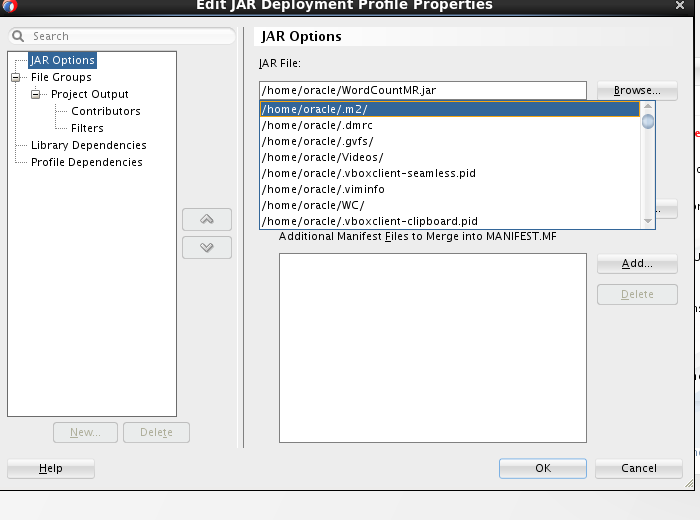
1. Create a new java file WordCountJob.java file. This file will include the main() function.

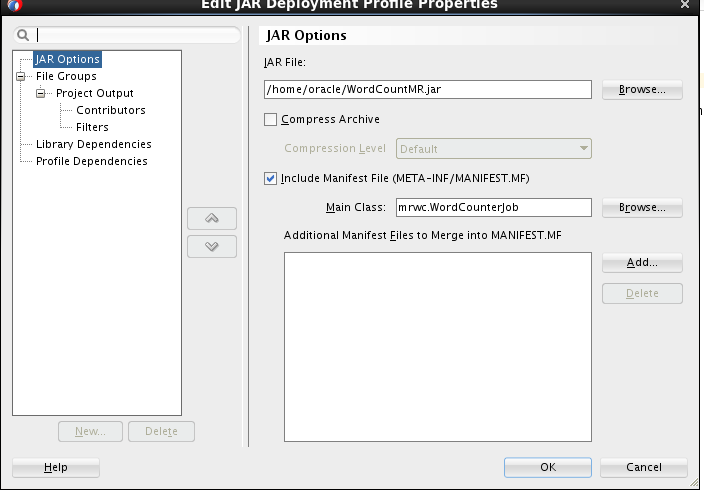


1. Select the necessary Hadoop libraries. Right click on MRWordCounter and select Project Proerties -> Add jar/ direcoties. Select required jar files.



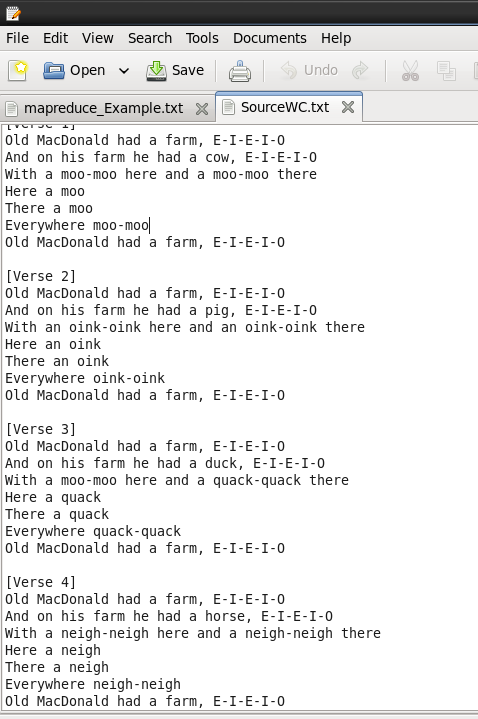
1. Create JAR file of the above project.



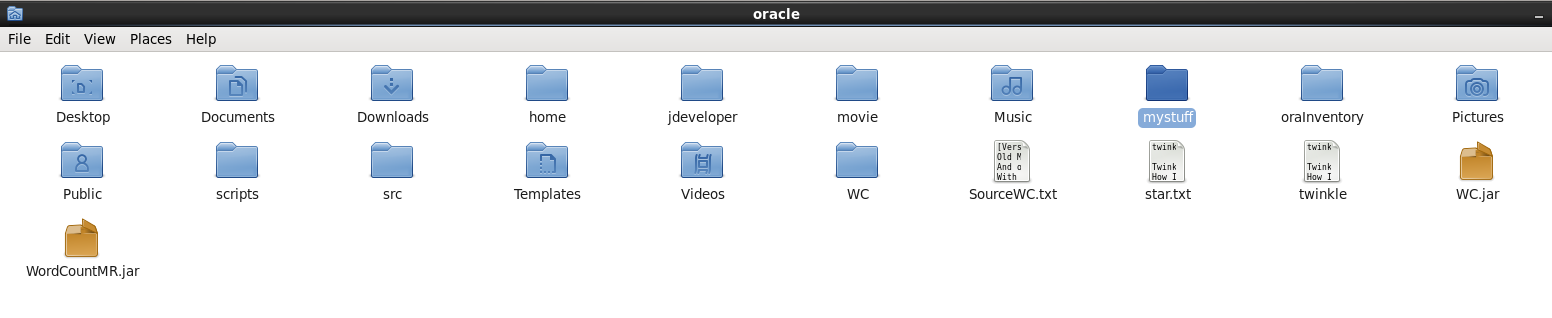


\

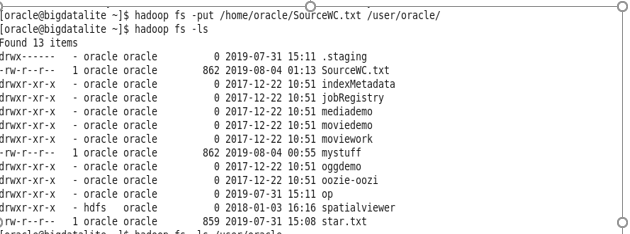
1. The SourceWC.txt file is the input file and contains required text to be processed.



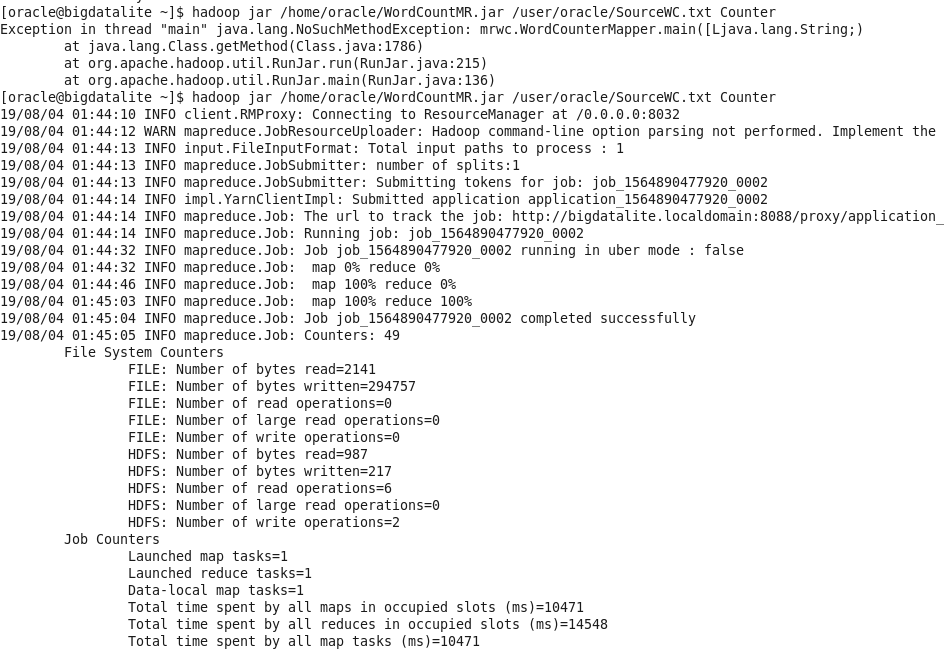
1. The WordCountMR.jar and SourceWC.txt files are in /home/oracle/ folder

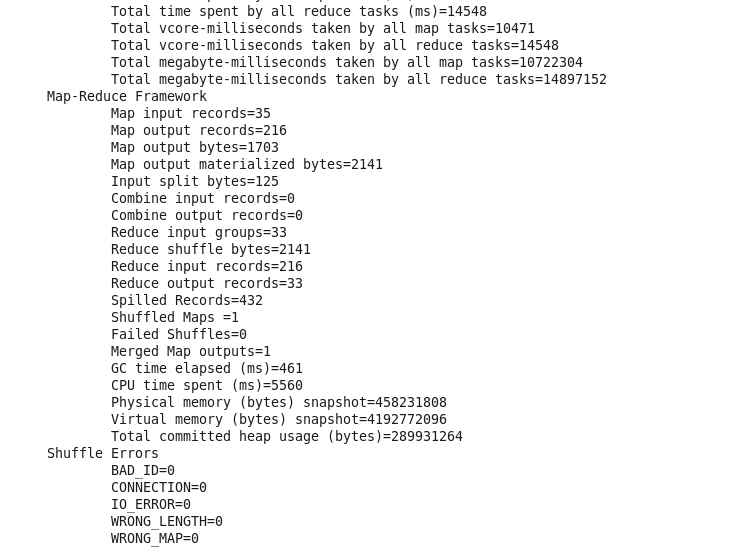


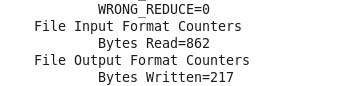
1. Upload SourceWC.txt on HDFS using the command



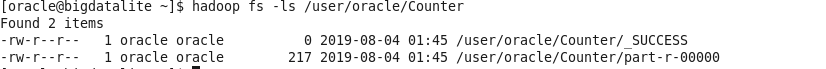
1. Execute the jar file and the output is redirected to “Counter” folder







1. The counter folder has two files. \_SUCCESS records the status and part-r-00000 contains the aggregated key value pairs (words) from the source file.



1. The part-r-00000 file contains all the words and respective counters of each word in SourceWC.txt file

