Big Data Analytics

**Lab Practical and date** – Practical 5, Monday 14th September 2020

**Name and Roll Number**- Het Shah, 17BIT103

**Practical Objective**- Apply MapReduce algorithms to perform analytics on single node cluster:

a) Analyse phrase frequency from given dataset

**Steps Involved-**

We installed designed MapReduce Algorithms to perform the analytic of finding the wordcount in a input file consisting of a corpus of words and the output is a txt file showing the frequency of each words

**Background**

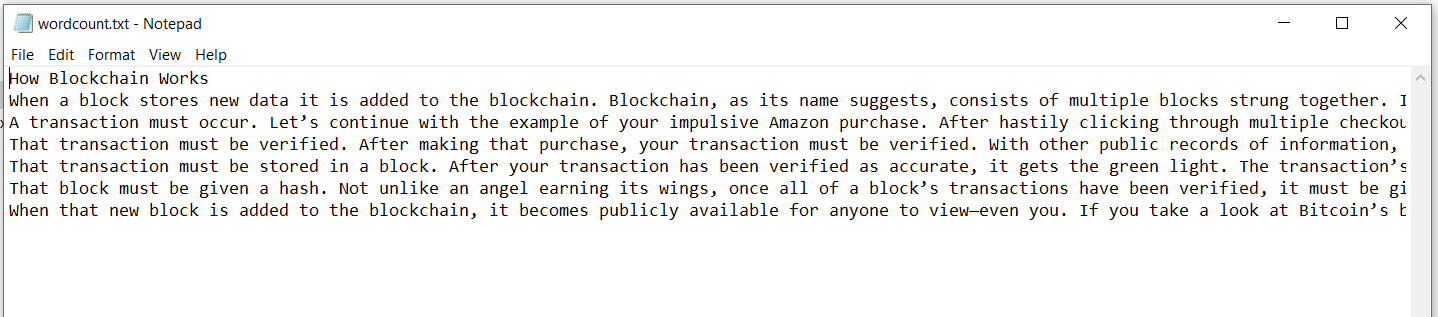
**Hadoop**

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model

**MapReduce**

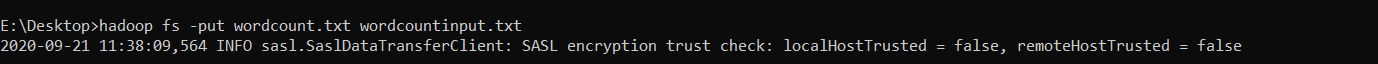
MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs filtering and sorting, and a reduce method, which performs a summary operation.

**Input File**

****

Input file consisted of words many of which were repeated and the file was copied to the HDFS system, by using the command

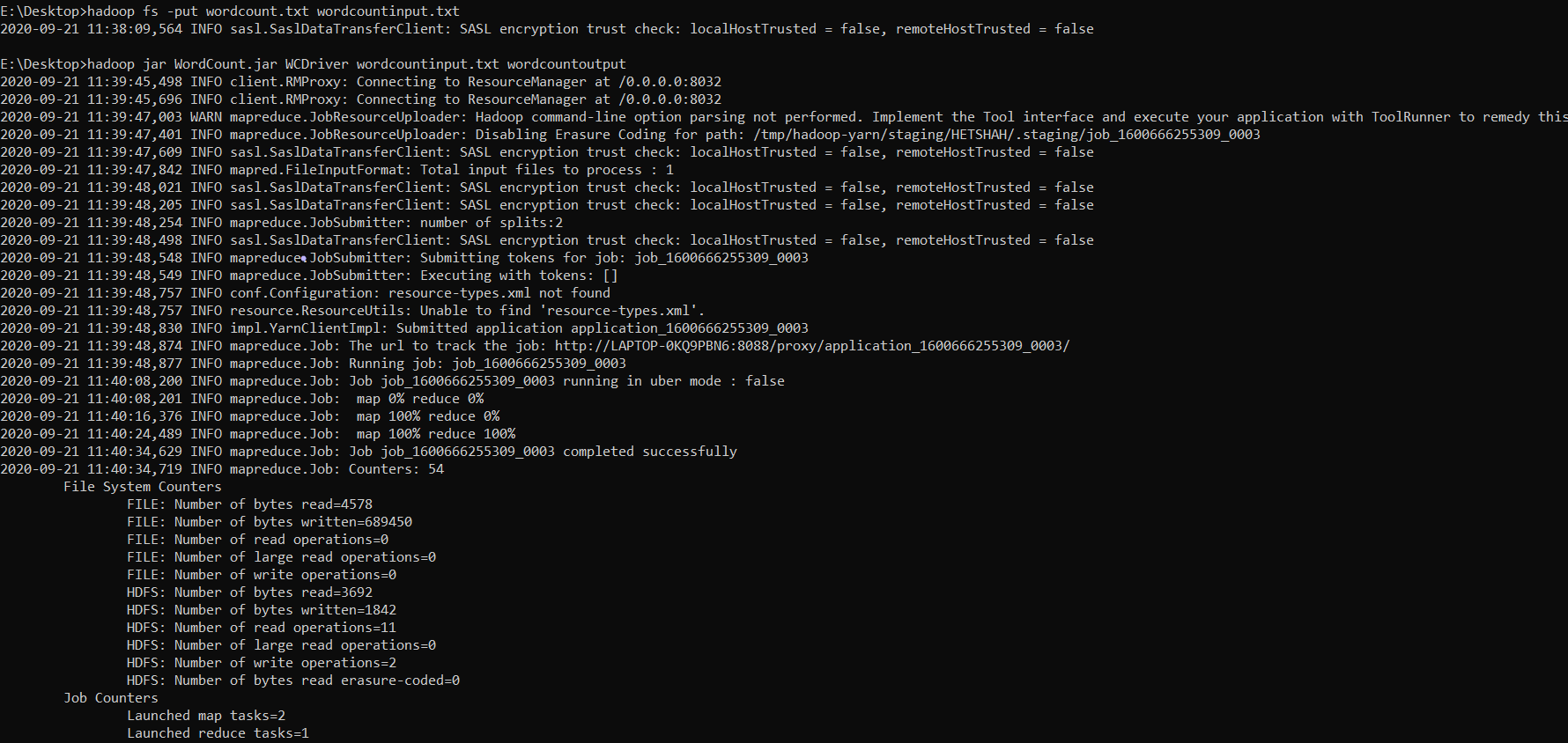
**Hadoop fs -put wordcount.txt wordcountinput.txt**

**How to run**

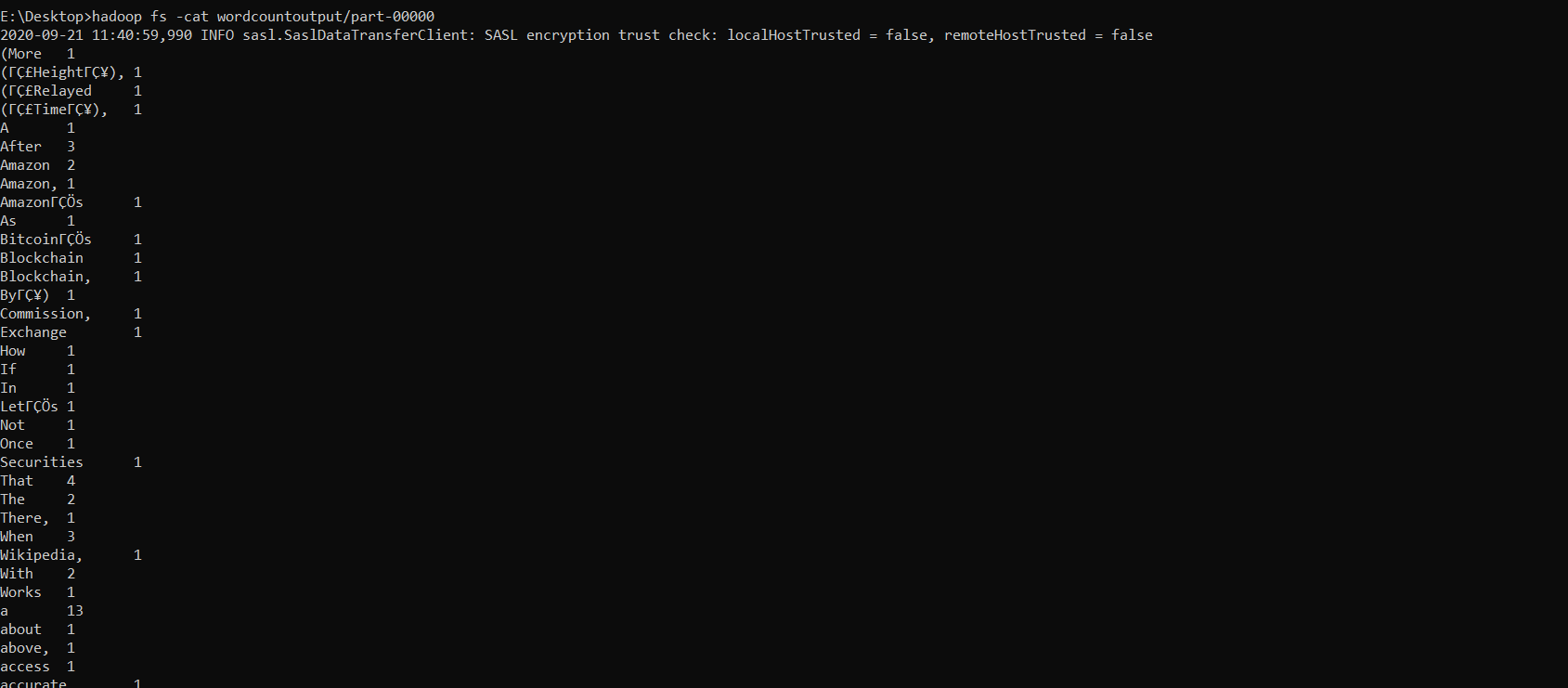
Each MapReduce task had 3 class files associated with it WCDriver, WCMapper and WCReducer. WCDriver was the main file and it would call the mapper and reducer

All the files were kept in the same package and after importing all the Hadoop extensions, the files were exported as a jar file which was then used to run the mapreduce program

**To run the program we write**

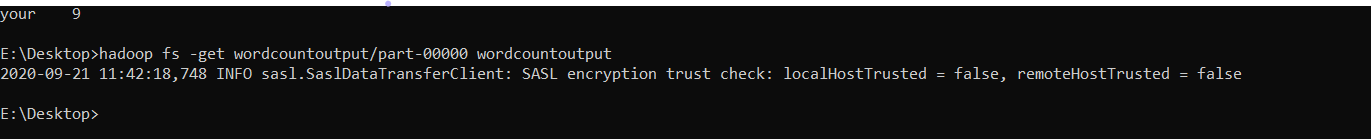
Hadoop jar WordCount.jar WCDriver wordcountinput.txt wordcountoutput 

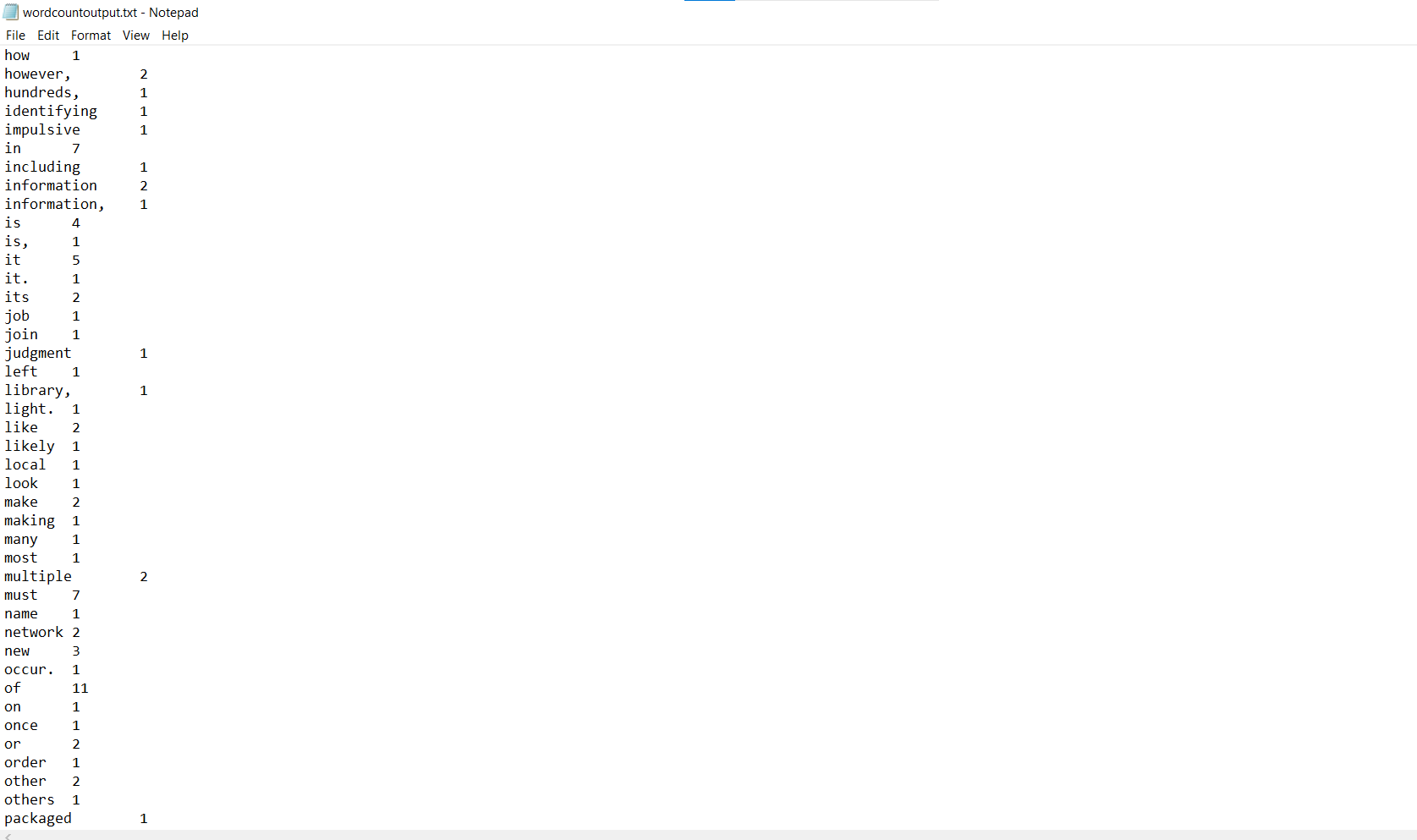
**To print the output we write**

Hadoop fs -cat wordcountoutput/part-00000

**To transfer the file from Hdfs to the LocalFile system we write the following command**

Hadoop fs -get wordcountoutput/part-00000 wordcountoutput



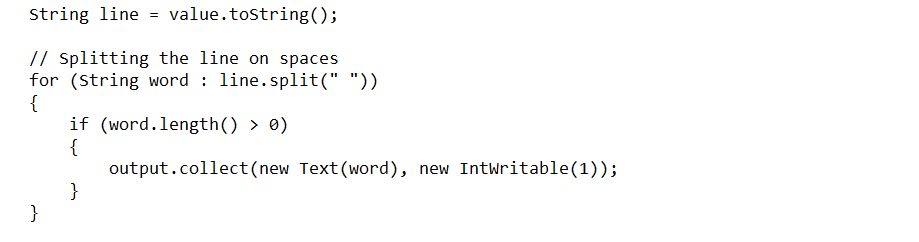


Output File

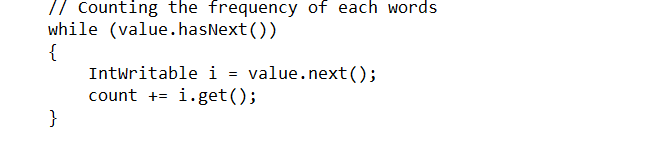
**Logic of Mapper and Reducer**

The text file is read and the split into the words using the split

The mapper emits the (word,1)



The reducer collects the word and then counts the number of instances that word has occurs and then emits (word,count)



**Conclusion**

In this practical we learned how to program using the MapReduce programming paradigm and used it to perform analytics on a set of word corpus