Big Data Analytics

**Lab Practical and date** – Practical 6, Monday 28th September 2020

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

**Practical Objective**-   
Implement PCY/Multi-Hash/SON algorithm for identification of frequent item set by handling larger datasets in main memory.

**Steps Involved-**

We installed designed MapReduce Algorithms to perform the analytic of implementing the apriori algorithm on the given data set to find the frequent itemset

**Background**

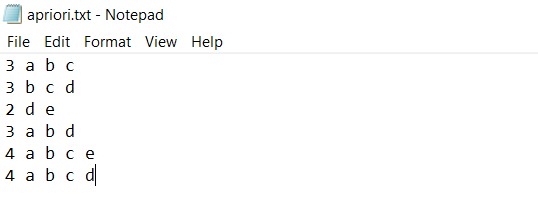
**Hadoop**

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**

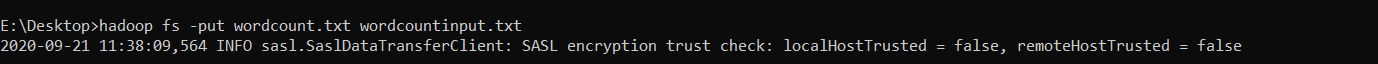
MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs filtering and sorting, and a reduce method, which performs a summary operation.

**Input File**

****

Input file consisted of multiple lines. Each line was a transaction where first word of each line was the number of items purchased. the file was copied to the HDFS system, by using the command

hadoop fs -put apriori.txt apriori

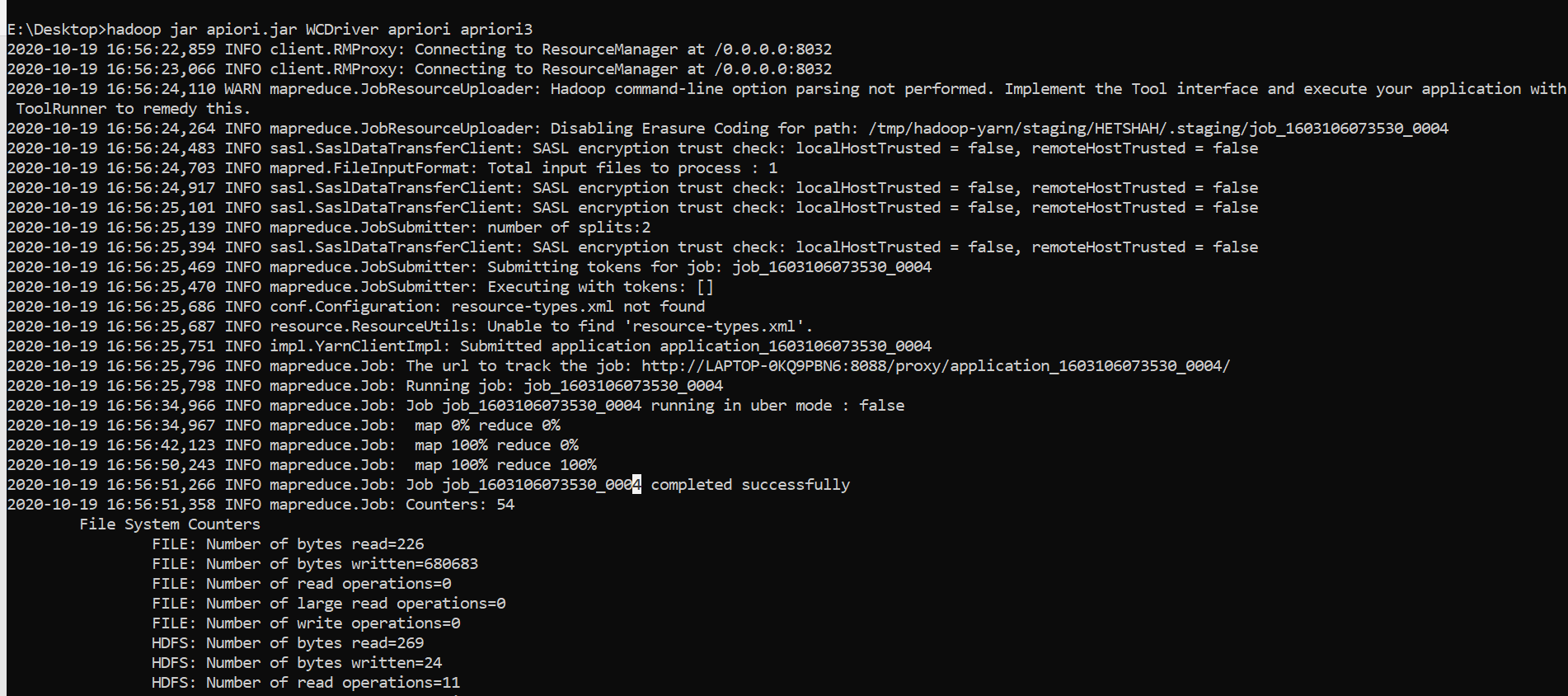
**How to run**

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

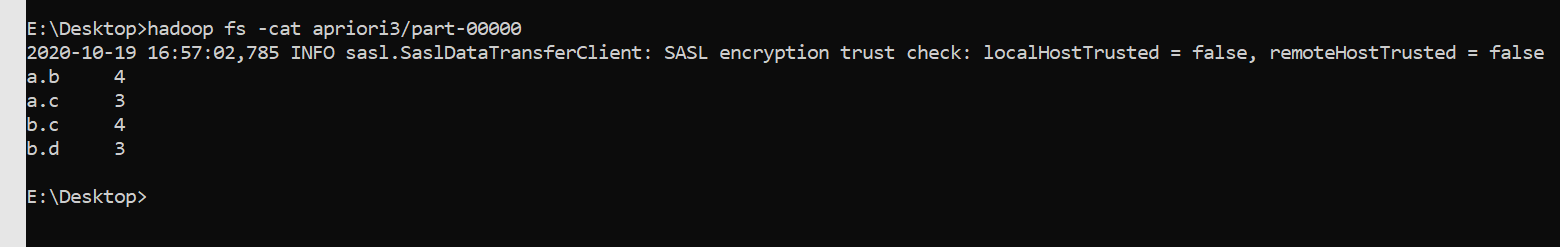
**To run the program we write**

hadoop jar apiori.jar WCDriver apriori apriori3

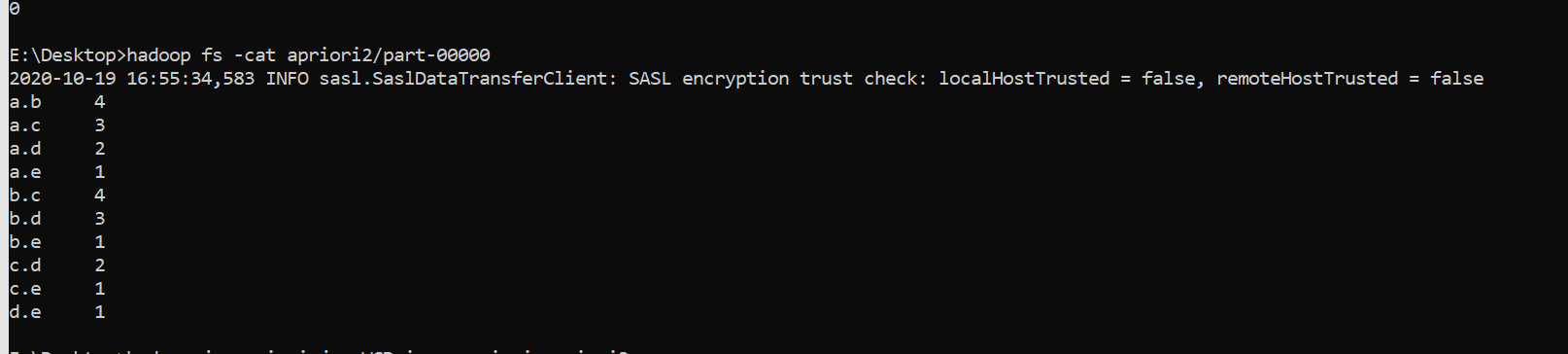


**To print the output we write**

hadoop fs -cat apriori3/part-00000

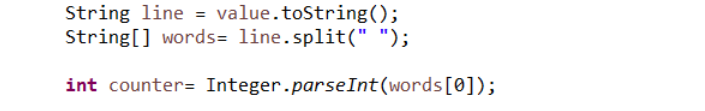


All the frequent data itemset without the support threshold

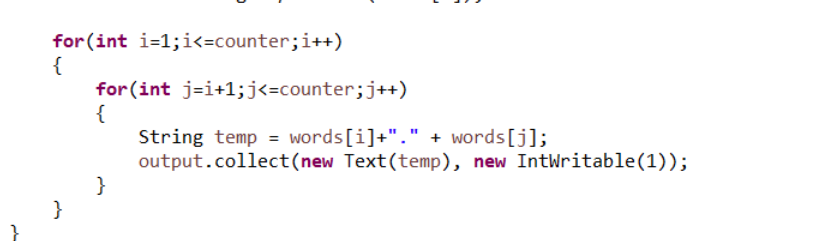


**Logic of Mapper and Reducer**

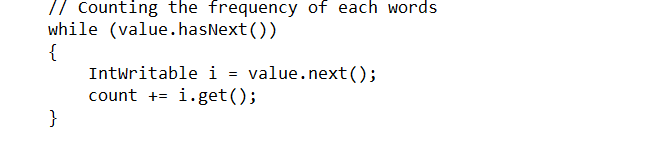
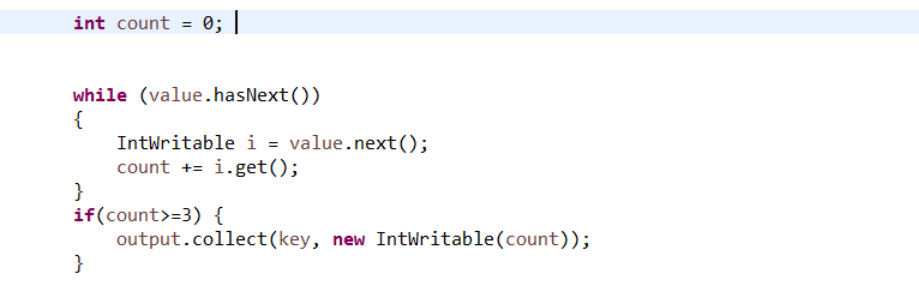
The text file is read and the split into the words using the split and the number of items brought are stored in the counter variable



Two for loops are used to find the association rules and the mapper emits the combination of the words and 1



The reducer collects the word and then counts the number of instances that word has occurs and then emits (combination ,count) if the count is more than support threshold



**Conclusion**

In this practical we learned how to program using the MapReduce programming paradigm and used it to perform the apriori algorithm on the given list of transactions.