**EXPERIMENT 3**

**AIM:** Write a map reduce program to count words in a text file.

**THEORY:**

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

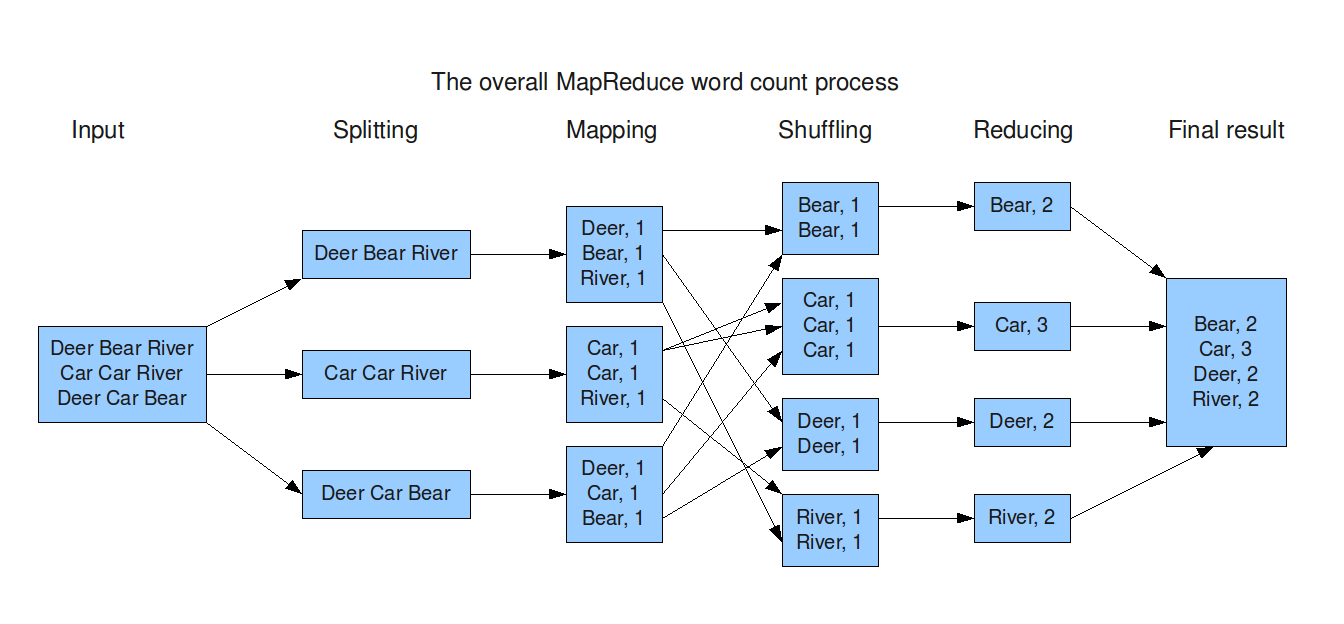
The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

**Usage:**

* It can be used in various applications like document clustering, distributed sorting, and web link-graph reversal.
* It can be used for distributed pattern-based searching.
* We can also use MapReduce in machine learning.
* It was used by Google to regenerate Google's index of the World Wide Web.
* It can be used in multiple computing environments such as multi-cluster, multi-core, and mobile environments.

**Stages:** MapReduce works in 2 stages:

1. Map stage − The map or mapper’s job is to process the input data. Generally, the input data is in the form of a file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
2. Reduce stage − This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.



**CODE:**

**mapper.java**

**Graphical user interface, text, application, email

Description automatically generated**

**reducer.java**

**Graphical user interface, text, application, email

Description automatically generated**

**driver code**

**Graphical user interface, text, application

Description automatically generated**

**Text file**

**Graphical user interface, text, application

Description automatically generated**

**Execute mapReduce**

**Graphical user interface, text

Description automatically generated**

**Store output**

**Text

Description automatically generated with medium confidence**

**OUTPUT:**

Graphical user interface, text, application, Word

Description automatically generated

**CONCLUSION:**

In this experiment, I have implemented MapReduce program in Hadoop for getting the word count of the data in the text file. I first coded the mapper and the reducer stages and then handled both the functions through the driver function. The mapper converts the given data into key-value pair and then the reducer shuffles and re-arranges the data from the mapper and reduces it according to the frequency of the words. Finally, we get the word count of the given data which is dumped into the output folder.