Big Data Analysis

Practical 5: MapReduce

Objective :

Apply MapReduce algorithms to find phrase frequency from given dataset.

- Prepare a report to guide design of mapper and reducer.

Roll No. & Name : 18bce183 & Prince Prajapati

Submitted to: Prof. Jaiprakash Verma

What is MapReduce?

MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop. The term "MapReduce" refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

Entire Class (with Mapper ans Reducer Method):

```
import java.io.IOException;
import java.io.InvalidObjectException;
import java.util.Iterator;
import java.util.StringTokenizer;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class MapReducer {

	public static class MyMap extends MapReduceBase implements Mapper<LongWritable,
Text,Text,IntWritable>
	{
		private Text mykey = new Text();
		public void map (LongWritable key, Text value, OutputCollector<Text,
IntWritable> output, Reporter reporter) throws IOException
		{
			String line = value.toString();
			String phrase[]=line.split(" ");
			int i=1;String sts="";
			while (i<phrase.length+1)
			{
				sts+=" "+phrase[i-1];
				if(i%3==0) {
					mykey.set(sts);
					output.collect (mykey, new IntWritable(1));
					sts="";
				}
				else if(i==phrase.length) {
					mykey.set(sts);
					output.collect (mykey, new IntWritable(1));
					sts="";
				}
				i++;
			}

		}
	}
```

```
	public static class MyReduce extends MapReduceBase implements Reducer<Text,
IntWritable, Text,IntWritable>
	{
		public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text ,
IntWritable> output, Reporter reporter) throws IOException
		{
			int sum=0;
			while (values.hasNext ())
			{
				sum +=values.next().get();

			}
			output.collect(key, new IntWritable(sum));
		}

	}
	public static void main(String[] args) throws Exception {
		JobConf conf = new JobConf (MapReducer.class);
		conf.setJobName( "MyFirstProgram");
		conf.setMapperClass (MyMap.class);
		conf.setReducerClass (MyReduce.class);
		conf.setOutputKeyClass (Text.class);
		conf.setOutputValueClass(IntWritable.class);
		conf.setInputFormat (TextInputFormat.class);
		conf.setOutputFormat (TextOutputFormat.class);
		FileInputFormat.setInputPaths (conf, new Path(args[0]));
		FileOutputFormat.setOutputPath(conf, new Path(args[1]));
		JobClient.runJob(conf);
	}


}
```

OUTPUT:

Command : hadoop jar C:\big-data\mapReduce.jar MapReducer /practical5/pp.txt /output10

```
C:\big-data\hadoop-3.3.0\sbin>hadoop jar C:\big-data\mapReduce.jar MapReducer /practical5/pp.txt /output10
2021-10-14 09:54:24,905 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-14 09:54:25,271 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-10-14 09:54:26,110 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inte
2021-10-14 09:54:26,234 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/91926/.stagin
2021-10-14 09:54:27,772 INFO mapred.FileInputFormat: Total input files to process : 1
2021-10-14 09:54:28,271 INFO mapreduce.JobSubmitter: number of splits:2
2021-10-14 09:54:29,106 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634185324567_0001
2021-10-14 09:54:29,106 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 09:54:29,391 INFO conf.Configuration: resource-types.xml not found
2021-10-14 09:54:29,392 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 09:54:29,931 INFO impl.YarnClientImpl: Submitted application application_1634185324567_0001
2021-10-14 09:54:30,073 INFO mapreduce.Job: The url to track the job: http://LAPTOP-0MDL8L9R:8088/proxy/application_1634185324567_000
2021-10-14 09:54:30,076 INFO mapreduce.Job: Running job: job_1634185324567_0001
2021-10-14 09:54:43,334 INFO mapreduce.Job: Job job_1634185324567_0001 running in uber mode : false
2021-10-14 09:54:43,338 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-14 09:54:52,763 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-14 09:55:01,912 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-14 09:55:02,926 INFO mapreduce.Job: Job job_1634185324567_0001 completed successfully
2021-10-14 09:55:03,258 INFO mapreduce.Job: Counters: 54
```

Txt file:

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.
Big data analytics is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.
Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

OUTPUT of MapReducer:

```
C:\big-data\hadoop-3.3.0\sbin>hadoop fs -ls /output10
Found 2 items
-rw-r--r--   1 91926 supergroup          0 2021-10-14 09:55 /output10/_SUCCESS
-rw-r--r--   1 91926 supergroup       1503 2021-10-14 09:54 /output10/part-00000

C:\big-data\hadoop-3.3.0\sbin>hadoop fs -cat /output10/part-00000
 (AI), mobile, social   1
 Analysis of big        1
 Big data analytics     2
 Businesses can use     1
 a very large   1
 advanced analytics techniques  1
 analytics, machine learning,   1
 and the Internet       1
 applications, web, and 1
 applied to data        1
 are driving data       1
 better and faster      1
 beyond the ability     1
 comes from sensors,    1
 complexity through new 1
 data allows analysts,  1
 data has one   1
 data.  1
 databases to capture,  1
 decisions using data   1
 devices, video/audio, networks,        1
 different sources, and 1
 example, big data      1
 forms and sources      1
```

CONCLUSION:

After implementing this practical now I have complete knowledge of how MapReducer works and how to code it.