2CS702 - Big Data Analytics

Practical 5

Aim: **Apply MapReduce algorithms to find phrase frequency from a given dataset. Prepare a report to guide design of mapper and reducer**

Author: Darshil Maru 20BCE514

Guide: Dr. Purnima Gandhi

WCDriver.java

```
import java.io.IOException;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class WCDriver extends Configured implements Tool {
    public int run(String args[]) throws IOException {
        if (args.length < 2) {
            System.out.println("Please give valid inputs");
            return -1;
        }
        JobConf conf = new JobConf(WCDriver.class);
        FileInputFormat.setInputPaths(conf, new Path(args[0]));
        FileOutputFormat.setOutputPath(conf, new Path(args[1]));
        conf.setMapperClass(WCMapper.class);
        conf.setReducerClass(WCReducer.class);
        conf.setMapOutputKeyClass(Text.class);
        conf.setMapOutputValueClass(IntWritable.class);
        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);
        JobClient.runJob(conf);
        return 0;
    }

    // Main Method public static void main(String args[]) throws Exception
    {
        int exitCode = ToolRunner.run(new WCDriver(), args);
        System.out.println(exitCode);
    }
}
```

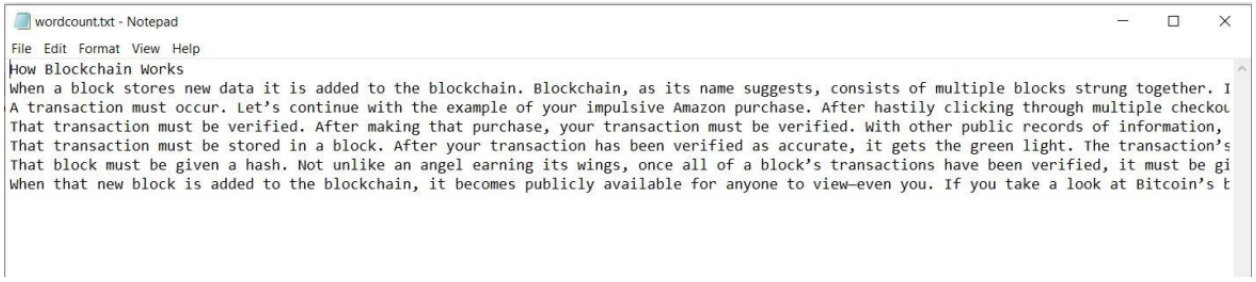
WCMapper.java

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;

public class WCMapper extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable> {
    // Map function
    public void map(LongWritable key, Text value, OutputCollector<Text,
IntWritable> output, Reporter rep)
            throws IOException {
        String line = value.toString();
        // Splitting the line on spaces for (String word : line.split("
"))
        {
            if (word.length() > 0) {
                output.collect(new Text(word), new IntWritable(1));
            }
        }
    }
}
```

WCReducer.java

```
import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class WCReducer extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable> {
    // Reduce function public void reduce(Text key, Iterator<IntWritable>
value,
OutputCollector (<Text, IntWritable> output, Reporter rep) throws
IOException
{
int count = 0;
// Counting the frequency of each words while (value.hasNext())
{
IntWritable i = value.next();
count += i.get();
}
output.collect(key, new IntWritable(count));
}
}
```

WordCount.txt



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

```
E:\Desktop>hadoop fs -put wordcount.txt wordcountinput.txt
2020-09-21 11:38:09,564 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

Hadoop jar WordCount.jar WCDriver wordcountinput.txt wordcountoutput

```
E:\Desktop>hadoop fs -put wordcount.txt wordcountinput.txt
2020-09-21 11:38:09,564 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

E:\Desktop>hadoop jar WordCount.jar WCDriver wordcountinput.txt wordcountoutput
2020-09-21 11:39:45,498 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-09-21 11:39:45,696 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-09-21 11:39:47,003 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this
2020-09-21 11:39:47,401 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/HETSHAH/.staging/job_1600666255309_0003
2020-09-21 11:39:47,609 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-09-21 11:39:47,842 INFO mapred.FileInputFormat: Total input files to process : 1
2020-09-21 11:39:48,021 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-09-21 11:39:48,205 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-09-21 11:39:48,254 INFO mapreduce.JobSubmitter: number of splits:2
2020-09-21 11:39:48,498 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-09-21 11:39:48,548 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1600666255309_0003
2020-09-21 11:39:48,549 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-09-21 11:39:48,757 INFO conf.Configuration: resource-types.xml not found
2020-09-21 11:39:48,757 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-09-21 11:39:48,830 INFO impl.YarnClientImpl: Submitted application application_1600666255309_0003
2020-09-21 11:39:48,874 INFO mapreduce.Job: The url to track the job: http://LAPTOP-0KQ9PBN6:8088/proxy/application_1600666255309_0003/
2020-09-21 11:39:48,877 INFO mapreduce.Job: Running job: job_1600666255309_0003
2020-09-21 11:40:08,200 INFO mapreduce.Job: Job job_1600666255309_0003 running in uber mode : false
2020-09-21 11:40:08,201 INFO mapreduce.Job:  map 0% reduce 0%
2020-09-21 11:40:16,376 INFO mapreduce.Job:  map 100% reduce 0%
2020-09-21 11:40:24,489 INFO mapreduce.Job:  map 100% reduce 100%
2020-09-21 11:40:34,629 INFO mapreduce.Job: Job job_1600666255309_0003 completed successfully
2020-09-21 11:40:34,719 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=4578
                FILE: Number of bytes written=689450
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3692
                HDFS: Number of bytes written=1842
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
```

Hadoop fs -cat wordcountoutput/part-00000

```
E:\Desktop>hadoop fs -cat wordcountoutput/part-00000
2020-09-21 11:40:59,990 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
(More   1
(ΓÇ£HeightΓÇ¥), 1
(ΓÇ£Relayed     1
(ΓÇ£TimeΓÇ¥),   1
A       1
After   3
Amazon  2
Amazon, 1
AmazonΓÇÖs      1
As      1
BitcoinΓÇÖs     1
Blockchain      1
Blockchain,     1
ByΓÇ¥)  1
Commission,     1
Exchange        1
How     1
If      1
In      1
LetΓÇÖs 1
Not     1
Once    1
Securities      1
That    4
The     2
There,  1
When    3
Wikipedia,      1
With    2
Works   1
a       13
about   1
above,  1
access  1
```

```
your    9

E:\Desktop>hadoop fs -get wordcountoutput/part-00000 wordcountoutput
2020-09-21 11:42:18,748 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

E:\Desktop>
```

wordcountoutput.txt - Notepad

File Edit Format View Help

```
how     1
however,        2
hundreds,       1
identifying     1
impulsive       1
in      7
including       1
information     2
information,    1
is      4
is,     1
it      5
it.     1
its     2
job     1
join    1
judgment        1
left    1
library,        1
light.  1
like    2
likely  1
local   1
look    1
make    2
making  1
many    1
most    1
multiple        2
must    7
name    1
network 2
new     3
occur.  1
of      11
on      1
once    1
or      2
order   1
other   2
others  1
packaged        1
```