Program 3 – WordCount

submitted by :
Shubham Garg
1NT19IS151
6th sem A3 batch

Code used in eclipse:

```
package three.two.one;

import java.io.IOException;
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.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;

public class WordCount
{
	public static class Reduce 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())
```

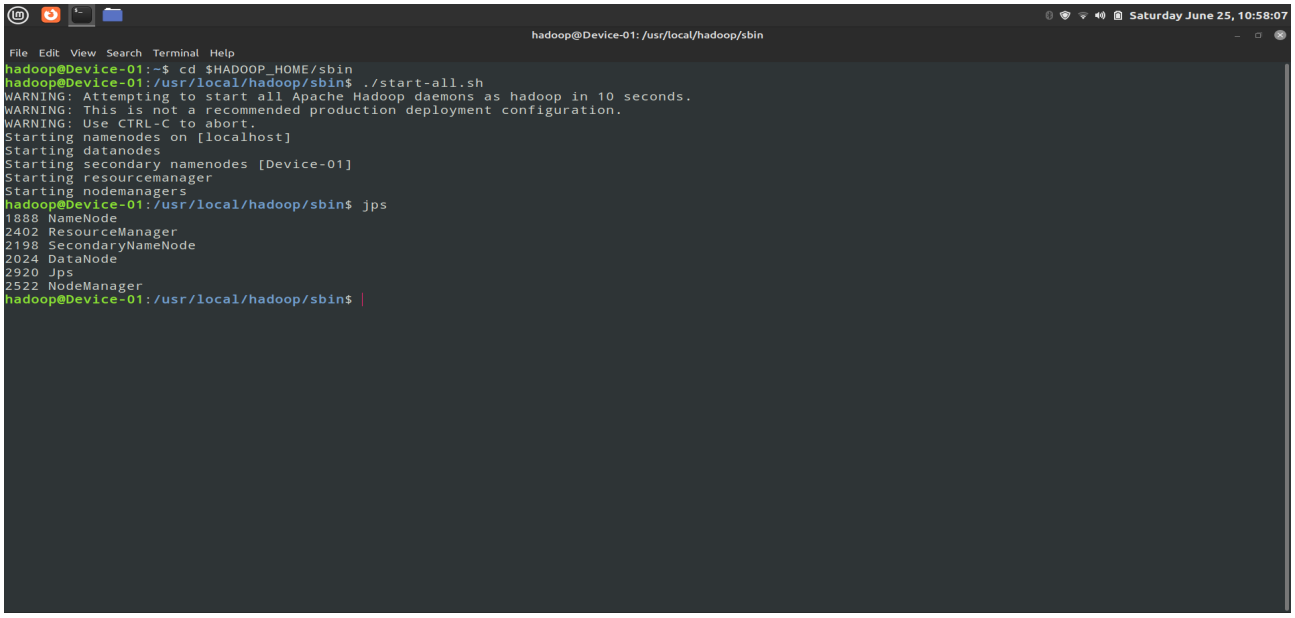
```
                        {
                                sum += values.next().get();
                        }
                        output.collect(key, new IntWritable(sum));
                }
        }



        public static class Map extends MapReduceBase implements Mapper <
LongWritable , Text , Text , IntWritable >
        {
                private final static IntWritable one = new IntWritable(1);
                private Text word = new Text();

                public void map(LongWritable key,Text
value,OutputCollector<Text,IntWritable>output,Reporter reporter) throws IOException
                {
                        String line = value.toString();
                        StringTokenizer tokenizer = new StringTokenizer(line);
                        while(tokenizer.hasMoreTokens())
                        {
                                word.set(tokenizer.nextToken());
                                output.collect(word,one);
                        }

                }
        }







        public static void main(String[] args) throws Exception
        {
                JobConf conf = new JobConf(WordCount.class);
                conf.setJobName("WordCount");

                conf.setOutputKeyClass(Text.class);
                conf.setOutputValueClass(IntWritable.class);

                conf.setMapperClass(Map.class);

                conf.setCombinerClass(Reduce.class);
                conf.setReducerClass(Reduce.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);
        }

}
```

This was exported as a .jar file and saved locally.

Now on Hadoop:

1. Staring/Checking status of daemons:



2. The WordCount program takes a text file as an input from and input directory.
   Creating an input directory and within that creating a text file and appending text to it.

```
hadoop@Device-01: /usr/local/hadoop/sbin
File  Edit  View  Search  Terminal  Help
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -mkdir -p /InputDir1
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2022-06-25 11:09 /InputDir1
drwx------   - hadoop supergroup          0 2022-06-20 15:08 /tmp
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -touchz /InputDir1/Test.txt
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -appendToFile - /InputDir1/Test.txt
hii this a boring text written to be counted in wordcount program
hii i missed a 'is' in above statement, if you noticed.
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -cat /InputDir1/Test.txt
hii this a boring text written to be counted in wordcount program
hii i missed a 'is' in above statement, if you noticed.
hadoop@Device-01:/usr/local/hadoop/sbin$
```

3. Now we run the jar file in hadoop specifying the input dir. and output dir.

```
hadoop@Device-01: /usr/local/hadoop/sbin
File  Edit  View  Search  Terminal  Help
hadoop@Device-01:/usr/local/hadoop/sbin$ hadoop jar /home/files/eclipse/Eclipse_Workspace/WordCount.jar /InputDir1 /OutputDir1
2022-06-25 11:16:30,980 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-06-25 11:16:31,109 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2022-06-25 11:16:31,255 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
cute your application with ToolRunner to remedy this.
2022-06-25 11:16:31,271 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_165613
4868633_0001
2022-06-25 11:16:31,426 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-25 11:16:31,461 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-25 11:16:31,553 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1656134868633_0001
2022-06-25 11:16:31,555 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-25 11:16:31,681 INFO conf.Configuration: resource-types.xml not found
2022-06-25 11:16:31,681 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-25 11:16:32,039 INFO impl.YarnClientImpl: Submitted application application_1656134868633_0001
2022-06-25 11:16:32,076 INFO mapreduce.Job: The url to track the job: http://Device-01:8088/proxy/application_1656134868633_0001/
2022-06-25 11:16:32,078 INFO mapreduce.Job: Running job: job_1656134868633_0001
2022-06-25 11:16:37,209 INFO mapreduce.Job: Job job_1656134868633_0001 running in uber mode : false
2022-06-25 11:16:37,211 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-25 11:16:41,302 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-25 11:16:48,420 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-25 11:16:48,432 INFO mapreduce.Job: Job job_1656134868633_0001 completed successfully
2022-06-25 11:16:48,505 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=266
                FILE: Number of bytes written=665115
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=367
                HDFS: Number of bytes written=153
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=4174
                Total time spent by all reduces in occupied slots (ms)=3810
                Total time spent by all map tasks (ms)=4174
                Total time spent by all reduce tasks (ms)=3810
```

4. The output will be stored in Output Directory’s file.

Saturday June 25, 11:19:26

hadoop@Device-01: /usr/local/hadoop/sbin

File Edit View Search Terminal Help

```
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=183
        File Output Format Counters
                Bytes Written=153
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2022-06-25 11:09 /InputDir1
drwxr-xr-x   - hadoop supergroup          0 2022-06-25 11:16 /OutputDir1
drwx------   - hadoop supergroup          0 2022-06-20 15:08 /tmp
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -ls /OutputDir1
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2022-06-25 11:16 /OutputDir1/_SUCCESS
-rw-r--r--   1 hadoop supergroup        153 2022-06-25 11:16 /OutputDir1/part-00000
hadoop@Device-01:/usr/local/hadoop/sbin$ hdfs dfs -cat /OutputDir1/part*
'is'    1
a       2
above   1
be      1
boring  1
counted 1
hii     2
i       1
if      1
in      2
missed  1
noticed.        1
program 1
statement,      1
text    1
this    1
to      1
wordcount       1
written 1
you     1
hadoop@Device-01:/usr/local/hadoop/sbin$
```