Exp – 4

Name: Namansh Singh Maurya
Roll: 22MIA1034

Aim: Word Count example on Hadoop using MapReduce

Algorithm/Procedure: following are the commands used to run the map reduce part and some images are also added.

ponny@ubuntu:~$ start-all.sh
ponny@ubuntu:~$ jps
ponny@ubuntu:~$ hadoop version
ponny@ubuntu:~$ javac -version
**ponny@ubuntu:~$** export HADOOP_CLASSPATH=$(hadoop classpath)
ponny@ubuntu:~$ echo $HADOOP_CLASSPATH
ponny@ubuntu:~$ hadoop fs -mkdir /WordCountTutorial
ponny@ubuntu:~$ hadoop fs -mkdir /WordCountTutorial/Input
ponny@ubuntu:~$ hadoop fs -put'/home/ponny/Desktop/WordCountTutorial/input_data/ex.txt' /WordCountTutorial/Input
ponny@ubuntu:~$ hadoop fs -put '/home/ponny/Desktop/WordCountTutorial/input_data/ex.txt' /WordCountTutorial/Input
ponny@ubuntu:~$ cd /home/ponny/Desktop/WordCountTutorial
ponny@ubuntu:~/Desktop/WordCountTutorial$ javac -classpath ${HADOOP_CLASSPATH} -d '/home/ponny/Desktop/WordCountTutorial/tutorial_classes' '/home/ponny/Desktop/WordCountTutorial/WordCount.java'
ponny@ubuntu:~/Desktop/WordCountTutorial$ jar -cvf firsttutorial.jar -C tutorial_classes/ .
ponny@ubuntu:~/Desktop/WordCountTutorial$ hadoop jar '/home/ponny/Desktop/WordCountTutorial/firsttutorial.jar' WordCount /WordCountTutorial/Input /WordCountTutorial/Output
ponny@ubuntu:~/Desktop/WordCountTutorial$ hadoop dfs -cat /WordCountTutorial/Output/*

```
ponny@ubuntu:~$ start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-namenode-ubuntu.out
localhost: starting datanode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-jobtracker-ubuntu.out
localhost: starting tasktracker, logging to /home/ponny/hadoop/libexec/../logs/hadoop-ponny-tasktracker-ubuntu.out
ponny@ubuntu:~$ jps
6919 DataNode
7331 TaskTracker
6754 NameNode
7086 SecondaryNameNode
7169 JobTracker
7370 Jps
ponny@ubuntu:~$ hadoop version
Warning: $HADOOP_HOME is deprecated.

Hadoop 1.0.4
Subversion https://svn.apache.org/repos/asf/hadoop/common/branches/branch-1.0 -r 1393290
Compiled by hortonfo on Wed Oct  3 05:13:58 UTC 2012
From source with checksum fe2baea87c4c81a2c505767f3f9b71f4
ponny@ubuntu:~$ javac -version
javac 1.7.0_201
ponny@ubuntu:~$
```

```
ponny@ubuntu:~$ export HADOOP_CLASSPATH=$(hadoop classpath)
Warning: $HADOOP_HOME is deprecated.

ponny@ubuntu:~$ echo $HADOOP_CLASSPATH
/home/ponny/hadoop/libexec/../conf:/usr/lib/jvm/java-7-openjdk-i386/lib/tools.jar:/home/ponny/hadoop/libexec/..:/home/ponny/hadoop/libexec/../hadoop-core-1.0.4.jar:/home/ponny/had
oop/libexec/../lib/asm-3.2.jar:/home/ponny/hadoop/libexec/../lib/aspectjrt-1.6.5.jar:/home/ponny/hadoop/libexec/../lib/aspectjtools-1.6.5.jar:/home/ponny/hadoop/libexec/../lib/com
mons-beanutils-1.7.0.jar:/home/ponny/hadoop/libexec/../lib/commons-beanutils-core-1.8.0.jar:/home/ponny/hadoop/libexec/../lib/commons-cli-1.2.jar:/home/ponny/hadoop/libexec/../lib
/commons-codec-1.4.jar:/home/ponny/hadoop/libexec/../lib/commons-collections-3.2.1.jar:/home/ponny/hadoop/libexec/../lib/commons-configuration-1.6.jar:/home/ponny/hadoop/libexec/.
./lib/commons-daemon-1.0.1.jar:/home/ponny/hadoop/libexec/../lib/commons-digester-1.8.jar:/home/ponny/hadoop/libexec/../lib/commons-el-1.0.jar:/home/ponny/hadoop/libexec/../lib/co
mmons-httpclient-3.0.1.jar:/home/ponny/hadoop/libexec/../lib/commons-io-2.1.jar:/home/ponny/hadoop/libexec/../lib/commons-lang-2.4.jar:/home/ponny/hadoop/libexec/../lib/commons-lo
gging-1.1.1.jar:/home/ponny/hadoop/libexec/../lib/commons-logging-api-1.0.4.jar:/home/ponny/hadoop/libexec/../lib/commons-math-2.1.jar:/home/ponny/hadoop/libexec/../lib/commons-ne
t-1.4.1.jar:/home/ponny/hadoop/libexec/../lib/core-3.1.1.jar:/home/ponny/hadoop/libexec/../lib/hadoop-capacity-scheduler-1.0.4.jar:/home/ponny/hadoop/libexec/../lib/hadoop-fairsch
eduler-1.0.4.jar:/home/ponny/hadoop/libexec/../lib/hadoop-thriftfs-1.0.4.jar:/home/ponny/hadoop/libexec/../lib/hsqldb-1.8.0.10.jar:/home/ponny/hadoop/libexec/../lib/jackson-core-a
sl-1.8.8.jar:/home/ponny/hadoop/libexec/../lib/jackson-mapper-asl-1.8.8.jar:/home/ponny/hadoop/libexec/../lib/jasper-compiler-5.5.12.jar:/home/ponny/hadoop/libexec/../lib/jasper-r
untime-5.5.12.jar:/home/ponny/hadoop/libexec/../lib/jdeb-0.8.jar:/home/ponny/hadoop/libexec/../lib/jersey-core-1.8.jar:/home/ponny/hadoop/libexec/../lib/jersey-json-1.8.jar:/home/
ponny/hadoop/libexec/../lib/jersey-server-1.8.jar:/home/ponny/hadoop/libexec/../lib/jets3t-0.6.1.jar:/home/ponny/hadoop/libexec/../lib/jetty-6.1.26.jar:/home/ponny/hadoop/libexec/
../lib/jetty-util-6.1.26.jar:/home/ponny/hadoop/libexec/../lib/jsch-0.1.42.jar:/home/ponny/hadoop/libexec/../lib/junit-4.5.jar:/home/ponny/hadoop/libexec/../lib/kfs-0.2.2.jar:/hom
e/ponny/hadoop/libexec/../lib/log4j-1.2.15.jar:/home/ponny/hadoop/libexec/../lib/mockito-all-1.8.5.jar:/home/ponny/hadoop/libexec/../lib/oro-2.0.8.jar:/home/ponny/hadoop/libexec/.
./lib/servlet-api-2.5-20081211.jar:/home/ponny/hadoop/libexec/../lib/slf4j-api-1.4.3.jar:/home/ponny/hadoop/libexec/../lib/slf4j-log4j12-1.4.3.jar:/home/ponny/hadoop/libexec/../li
b/xmlenc-0.52.jar:/home/ponny/hadoop/libexec/../lib/jsp-2.1/jsp-2.1.jar:/home/ponny/hadoop/libexec/../lib/jsp-2.1/jsp-api-2.1.jar
ponny@ubuntu:~$
```

```
ponny@ubuntu:~/Desktop/WordCountTutorial$ jar -cvf firsttutorial.jar -C t
utorial_classes/ .
added manifest
adding: WordCount$IntSumReducer.class(in = 1739) (out= 740)(deflated 57%)
adding: WordCount$TokenizerMapper.class(in = 1736) (out= 754)(deflated 56
%)
adding: WordCount.class(in = 1459) (out= 798)(deflated 45%)
```

```
ponny@ubuntu:~/Desktop/WordCountTutorial$ hadoop jar  '/home/ponny/Desktop/WordC
ountTutorial/firsttutorial.jar' WordCount /WordCountTutorial/Input /WordCountTut
orial/Output
Warning: $HADOOP_HOME is deprecated.

25/02/03 11:02:13 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
25/02/03 11:02:13 INFO input.FileInputFormat: Total input paths to process : 1
25/02/03 11:02:13 INFO util.NativeCodeLoader: Loaded the native-hadoop library
25/02/03 11:02:13 WARN snappy.LoadSnappy: Snappy native library not loaded
25/02/03 11:02:13 INFO mapred.JobClient: Running job: job_202502030957_0001
25/02/03 11:02:14 INFO mapred.JobClient:  map 0% reduce 0%
25/02/03 11:02:28 INFO mapred.JobClient:  map 100% reduce 0%
25/02/03 11:02:43 INFO mapred.JobClient:  map 100% reduce 100%
25/02/03 11:02:48 INFO mapred.JobClient: Job complete: job_202502030957_0001
25/02/03 11:02:48 INFO mapred.JobClient: Counters: 29
25/02/03 11:02:48 INFO mapred.JobClient:   Job Counters
25/02/03 11:02:48 INFO mapred.JobClient:     Launched reduce tasks=1
25/02/03 11:02:48 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=10329
25/02/03 11:02:48 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
25/02/03 11:02:48 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
25/02/03 11:02:48 INFO mapred.JobClient:     Launched map tasks=1
25/02/03 11:02:48 INFO mapred.JobClient:     Data-local map tasks=1
25/02/03 11:02:48 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=12523
25/02/03 11:02:48 INFO mapred.JobClient:   File Output Format Counters
25/02/03 11:02:48 INFO mapred.JobClient:     Bytes Written=54
25/02/03 11:02:48 INFO mapred.JobClient:   FileSystemCounters
25/02/03 11:02:48 INFO mapred.JobClient:     FILE_BYTES_READ=92
25/02/03 11:02:48 INFO mapred.JobClient:     HDFS_BYTES_READ=161
25/02/03 11:02:48 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=42989
25/02/03 11:02:48 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=54
25/02/03 11:02:48 INFO mapred.JobClient:   File Input Format Counters
25/02/03 11:02:48 INFO mapred.JobClient:     Bytes Read=43
25/02/03 11:02:48 INFO mapred.JobClient:   Map-Reduce Framework
25/02/03 11:02:48 INFO mapred.JobClient:     Map output materialized bytes=92
25/02/03 11:02:48 INFO mapred.JobClient:     Map input records=4
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce shuffle bytes=92
```

```
25/02/03 11:02:48 INFO mapred.JobClient:   Map-Reduce Framework
25/02/03 11:02:48 INFO mapred.JobClient:     Map output materialized bytes=92
25/02/03 11:02:48 INFO mapred.JobClient:     Map input records=4
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce shuffle bytes=92
25/02/03 11:02:48 INFO mapred.JobClient:     Spilled Records=16
25/02/03 11:02:48 INFO mapred.JobClient:     Map output bytes=78
25/02/03 11:02:48 INFO mapred.JobClient:     Total committed heap usage (bytes)=
177016832
25/02/03 11:02:48 INFO mapred.JobClient:     CPU time spent (ms)=480
25/02/03 11:02:48 INFO mapred.JobClient:     Combine input records=9
25/02/03 11:02:48 INFO mapred.JobClient:     SPLIT_RAW_BYTES=118
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce input records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce input groups=8
25/02/03 11:02:48 INFO mapred.JobClient:     Combine output records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Physical memory (bytes) snapshot=19
5129344
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce output records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=770
924544
25/02/03 11:02:48 INFO mapred.JobClient:     Map output records=9
```

Program:

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {
  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }
  public static class IntSumReducer
       extends Reducer<Text,IntWritable,Text,IntWritable> {
    private IntWritable result = new IntWritable();
    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context
                       ) throws IOException, InterruptedException {
      int sum = 0;
      for (IntWritable val : values) {
        sum += val.get();
      }
      result.set(sum);
      context.write(key, result);
    }
  }
  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
```

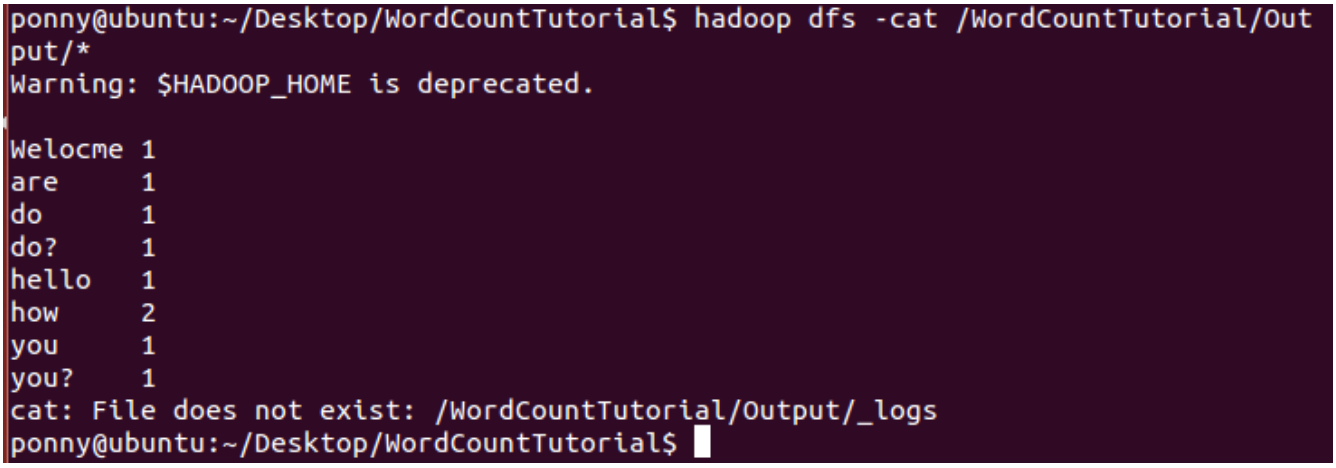
```
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

Output:

```
25/02/03 11:02:48 INFO mapred.JobClient:   File Output Format Counters
25/02/03 11:02:48 INFO mapred.JobClient:     Bytes Written=54
25/02/03 11:02:48 INFO mapred.JobClient:   FileSystemCounters
25/02/03 11:02:48 INFO mapred.JobClient:     FILE_BYTES_READ=92
25/02/03 11:02:48 INFO mapred.JobClient:     HDFS_BYTES_READ=161
25/02/03 11:02:48 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=42989
25/02/03 11:02:48 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=54
25/02/03 11:02:48 INFO mapred.JobClient:   File Input Format Counters
25/02/03 11:02:48 INFO mapred.JobClient:     Bytes Read=43
25/02/03 11:02:48 INFO mapred.JobClient:   Map-Reduce Framework
25/02/03 11:02:48 INFO mapred.JobClient:     Map output materialized bytes=92
25/02/03 11:02:48 INFO mapred.JobClient:     Map input records=4
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce shuffle bytes=92
25/02/03 11:02:48 INFO mapred.JobClient:     Spilled Records=16
25/02/03 11:02:48 INFO mapred.JobClient:     Map output bytes=78
25/02/03 11:02:48 INFO mapred.JobClient:     Total committed heap usage (bytes)=
177016832
25/02/03 11:02:48 INFO mapred.JobClient:     CPU time spent (ms)=480
25/02/03 11:02:48 INFO mapred.JobClient:     Combine input records=9
25/02/03 11:02:48 INFO mapred.JobClient:     SPLIT_RAW_BYTES=118
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce input records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce input groups=8
25/02/03 11:02:48 INFO mapred.JobClient:     Combine output records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Physical memory (bytes) snapshot=19
5129344
25/02/03 11:02:48 INFO mapred.JobClient:     Reduce output records=8
25/02/03 11:02:48 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=770
924544
25/02/03 11:02:48 INFO mapred.JobClient:     Map output records=9
ponny@ubuntu:~/Desktop/WordCountTutorial$ hadoop dfs -cat /WordCountTutorial/Out
put
Warning: $HADOOP_HOME is deprecated.

cat: File does not exist: /WordCountTutorial/Output
ponny@ubuntu:~/Desktop/WordCountTutorial$ hadoop dfs -cat /WordCountTutorial/Out
put/*
Warning: $HADOOP_HOME is deprecated.

Welocme 1
are     1
do      1
do?     1
hello   1
how     2
you     1
you?    1
cat: File does not exist: /WordCountTutorial/Output/_logs
ponny@ubuntu:~/Desktop/WordCountTutorial$
```



Result: Word Count was successfully calculated and printed on the screen.