```
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);
```

```java
      context.write(key, result);
    }
  }

  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(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);
  }
}
```

HADOOP COMMANDS

# Hadoop Word Count Assignment

Write a code in JAVA for a simple Word Count application that counts the number of occurrences of each word in a given input set using the Hadoop Map-Reduce framework on local-standalone set-up.

## Steps

1. ssh into localhost

   ```bash
   ssh localhost
   ```

   ```bash
   Welcome to Ubuntu 18.04 LTS (bison-elk-cougar-mlk X54) (GNU/Linux
   ```

```
5.4.0-125-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

279 updates can be applied immediately.
236 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

Last login: Thu Apr 25 11:20:38 2024 from 127.0.0.1
```

2. Start hadoop services

```bash
start-all.sh
```

```bash
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 2795.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 2951.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [pict-OptiPlex-5070]
pict-OptiPlex-5070: secondarynamenode is running as process 3184.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 6467.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
```

```
localhost: nodemanager is running as process 6646.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
```

3. Check the status of the services

    ```bash
    jps
    ```

    ```bash
    3184 SecondaryNameNode
    6467 ResourceManager
    6646 NodeManager
    2951 DataNode
    9145 Jps
    2795 NameNode
    ```

4. Crrate a directory in HDFS

    ```bash
    hadoop dfs -mkdir /user/<roll no.>
    ```

    ```bash

    WARNING: Use of this script to execute dfs is deprecated.
    WARNING: Attempting to execute replacement "hdfs dfs" instead.
    ```

5. Export Hadoop classpath and echo it.

    ```bash
    export HADOOP_CLASSPATH=$(hadoop classpath)
    echo $HADOOP_CLASSPATH
    ```

    ```bash

    /home/hadoop/hadoop-3.3.5/etc/hadoop:/home/hadoop/hadoop-3.3.5/share/hadoop/common/lib/*:/home/hadoop/hadoop-3.3.5/share/hadoop/common/*:/home/hadoop/hadoop-3.3.5/share/hadoop/hdfs:/home/hadoop/
    ```

```
hadoop-3.3.5/share/hadoop/hdfs/lib/*:/home/hadoop/hadoop-3.3.5/share/hadoop/hdfs/*:/home/hadoop/hadoop-3.3.5/share/hadoop/mapreduce/*:/home/hadoop/hadoop-3.3.5/share/hadoop/yarn:/home/hadoop/hadoop-3.3.5/share/hadoop/yarn/lib/*:/home/hadoop/hadoop-3.3.5/share/hadoop/yarn/*
```

6. Create an input directory and put the `input.txt` file in it.

```bash
hadoop dfs -mkdir /user/<roll no.>/input
hadoop dfs -put input.txt /user/<roll no.>/input
```

```bash
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.
```

7. Compile the java code

```bash
# Directory of the compiled files
javac -classpath ${HADOOP_CLASSPATH} -d "</home/hadoop/<roll no.>/tut>" '/home/hadoop/<roll no.>/WordCount.java'
```

8. Create a jar file

```bash
# cd into the folder in which java files are present, then run
jar -cvf stutorial.jar -C tut/ .
```

```bash
added manifest
adding: WordCount$IntSumReducer.class(in = 1755) (out= 749)(deflated 57%)
adding: WordCount$TokenizerMapper.class(in = 1752) (out= 764)(deflated 56%)
adding: WordCount.class(in = 1511) (out= 825)(deflated 45%)
```

9. Run the jar file

```bash
hadoop jar '/home/hadoop/<roll no.>/stutorial.jar' WordCount /user/<roll no.>/input /user/<roll no.>/output
```

```bash
2024-04-25 11:26:06,585 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:8032
2024-04-25 11:26:06,840 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-25 11:26:06,922 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1714024261854_0001
2024-04-25 11:26:07,138 INFO input.FileInputFormat: Total input files to process : 1
2024-04-25 11:26:07,381 INFO mapreduce.JobSubmitter: number of splits:1
2024-04-25 11:26:07,538 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1714024261854_0001
2024-04-25 11:26:07,539 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-25 11:26:07,647 INFO conf.Configuration: resource-types.xml not found
2024-04-25 11:26:07,648 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-25 11:26:07,818 INFO impl.YarnClientImpl: Submitted application application_1714024261854_0001
2024-04-25 11:26:07,866 INFO mapreduce.Job: The url to track the job: http://pict-OptiPlex-5070:8088/proxy/application_1714024261854_0001/
2024-04-25 11:26:07,866 INFO mapreduce.Job: Running job: job_1714024261854_0001
2024-04-25 11:26:13,924 INFO mapreduce.Job: Job job_1714024261854_0001 running in uber mode : false
2024-04-25 11:26:13,926 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-25 11:26:17,987 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-25 11:26:22,012 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-25 11:26:23,036 INFO mapreduce.Job: Job job_1714024261854_0001 completed successfully
2024-04-25 11:26:23,090 INFO mapreduce.Job: Counters: 54
```

```
File System Counters
    FILE: Number of bytes read=29
    FILE: Number of bytes written=551415
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=139
    HDFS: Number of bytes written=15
    HDFS: Number of read operations=8
    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=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=2101
    Total time spent by all reduces in occupied slots (ms)=1554
    Total time spent by all map tasks (ms)=2101
    Total time spent by all reduce tasks (ms)=1554
    Total vcore-milliseconds taken by all map tasks=2101
    Total vcore-milliseconds taken by all reduce tasks=1554
    Total megabyte-milliseconds taken by all map tasks=2151424
    Total megabyte-milliseconds taken by all reduce tasks=1591296
Map-Reduce Framework
    Map input records=1
    Map output records=5
    Map output bytes=47
    Map output materialized bytes=29
    Input split bytes=112
    Combine input records=5
    Combine output records=2
    Reduce input groups=2
    Reduce shuffle bytes=29
    Reduce input records=2
    Reduce output records=2
    Spilled Records=4
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=30
    CPU time spent (ms)=870
    Physical memory (bytes) snapshot=496246784
```

```
      Virtual memory (bytes) snapshot=5576851456
      Total committed heap usage (bytes)=392167424
      Peak Map Physical memory (bytes)=287973376
      Peak Map Virtual memory (bytes)=2783084544
      Peak Reduce Physical memory (bytes)=208273408
      Peak Reduce Virtual memory (bytes)=2793766912
   Shuffle Errors
      BAD_ID=0
      CONNECTION=0
      IO_ERROR=0
      WRONG_LENGTH=0
      WRONG_MAP=0
      WRONG_REDUCE=0
   File Input Format Counters
      Bytes Read=27
   File Output Format Counters
      Bytes Written=15
```

10. Check the output

   ```bash
   hadoop dfs -cat /user/<roll no.>/output/*
   ```

   ```bash
   WARNING: Use of this script to execute dfs is deprecated.
   WARNING: Attempting to execute replacement "hdfs dfs" instead.

   hello   2
   pict    3
   ```