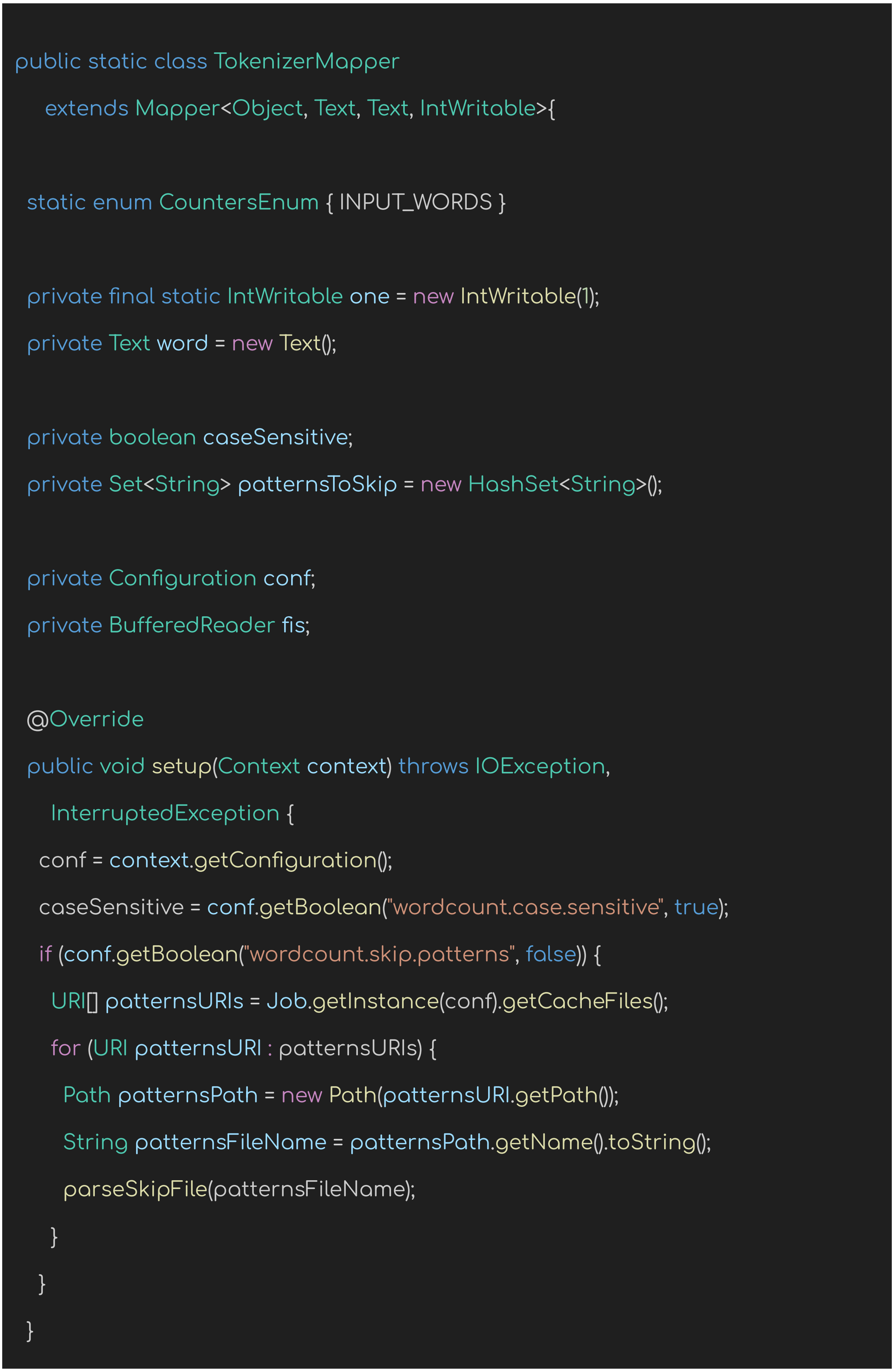
PRACTICAL 4

BDA 20BCE057

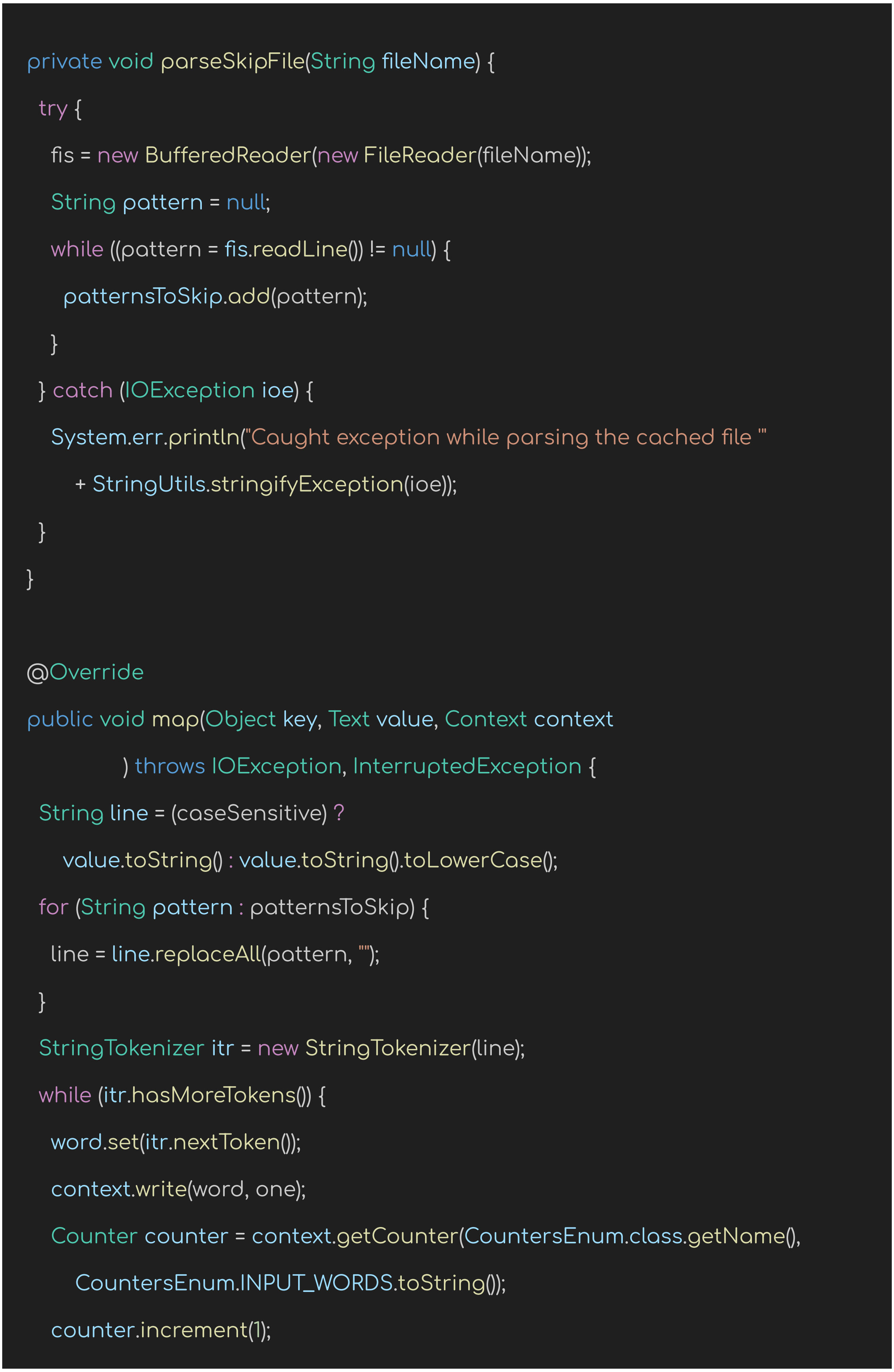
Aim : Mapreduce programme for wordcount

Code:

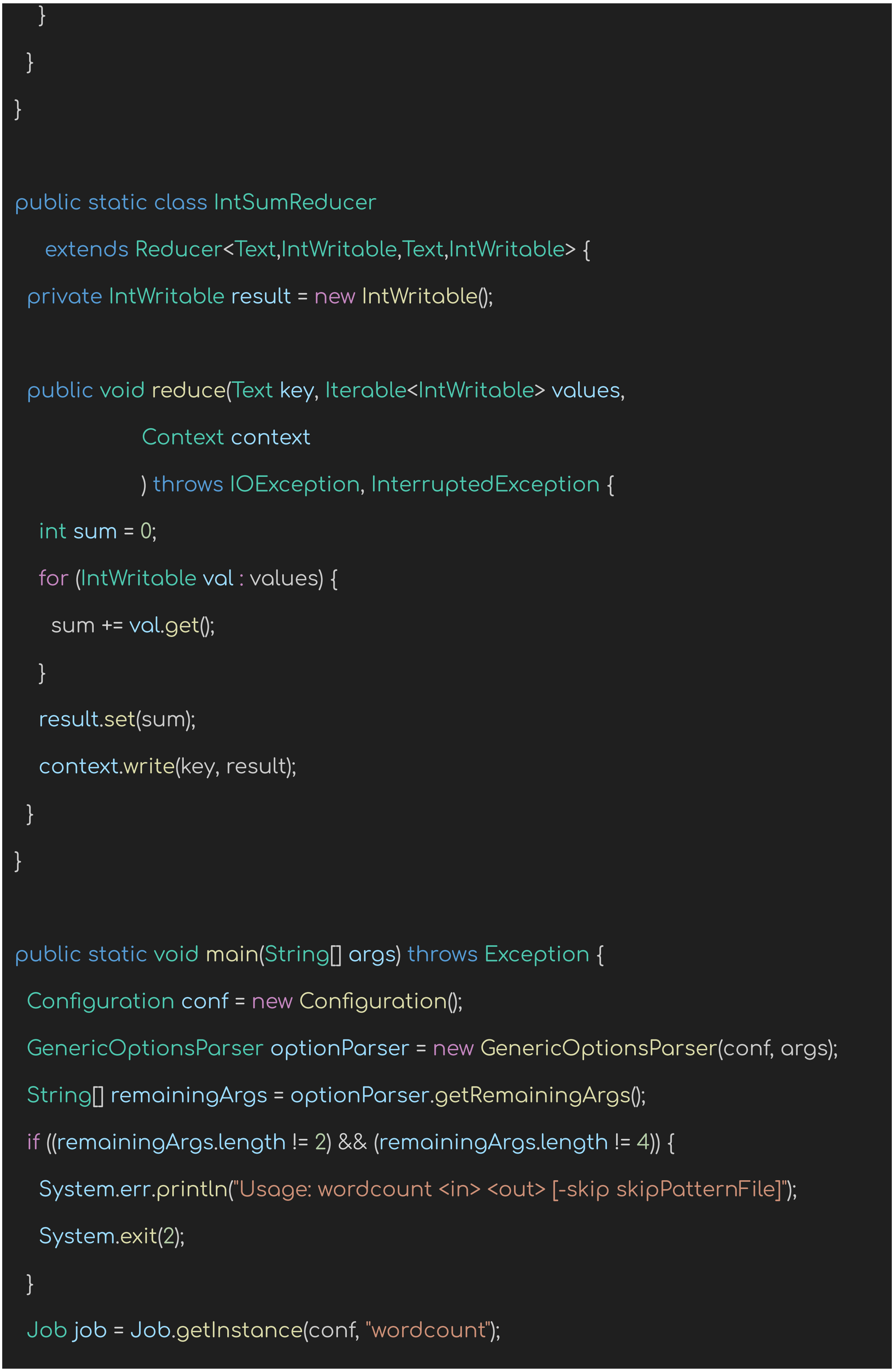
```
import java.io.BufferedReader;
import java.io.FileReader;
import java.io.IOException;
import java.net.URI;
import java.util.ArrayList;
import java.util.HashSet;
import java.util.List;
import java.util.Set;
import java.util.StringTokenizer;

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;
import org.apache.hadoop.mapreduce.Counter;
import org.apache.hadoop.util.GenericOptionsParser;
import org.apache.hadoop.util.StringUtils;

public class WordCount2 {
```



```
public static class TokenizerMapper
     extends Mapper<Object, Text, Text, IntWritable>{

  static enum CountersEnum { INPUT_WORDS }

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  private boolean caseSensitive;
  private Set<String> patternsToSkip = new HashSet<String>();

  private Configuration conf;
  private BufferedReader fis;

  @Override
  public void setup(Context context) throws IOException,
      InterruptedException {
    conf = context.getConfiguration();
    caseSensitive = conf.getBoolean("wordcount.case.sensitive", true);
    if (conf.getBoolean("wordcount.skip.patterns", false)) {
      URI[] patternsURIs = Job.getInstance(conf).getCacheFiles();
      for (URI patternsURI : patternsURIs) {
        Path patternsPath = new Path(patternsURI.getPath());
        String patternsFileName = patternsPath.getName().toString();
        parseSkipFile(patternsFileName);
      }
    }
  }
```

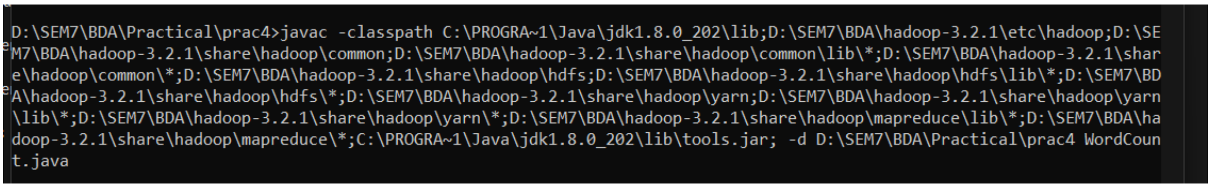


```
private void parseSkipFile(String fileName) {
  try {
    fis = new BufferedReader(new FileReader(fileName));
    String pattern = null;
    while ((pattern = fis.readLine()) != null) {
      patternsToSkip.add(pattern);
    }
  } catch (IOException ioe) {
    System.err.println("Caught exception while parsing the cached file '"
        + StringUtils.stringifyException(ioe));
  }
}

@Override
public void map(Object key, Text value, Context context
                ) throws IOException, InterruptedException {
  String line = (caseSensitive) ?
      value.toString() : value.toString().toLowerCase();
  for (String pattern : patternsToSkip) {
    line = line.replaceAll(pattern, "");
  }
  StringTokenizer itr = new StringTokenizer(line);
  while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
    context.write(word, one);
    Counter counter = context.getCounter(CountersEnum.class.getName(),
        CountersEnum.INPUT_WORDS.toString());
    counter.increment(1);
```

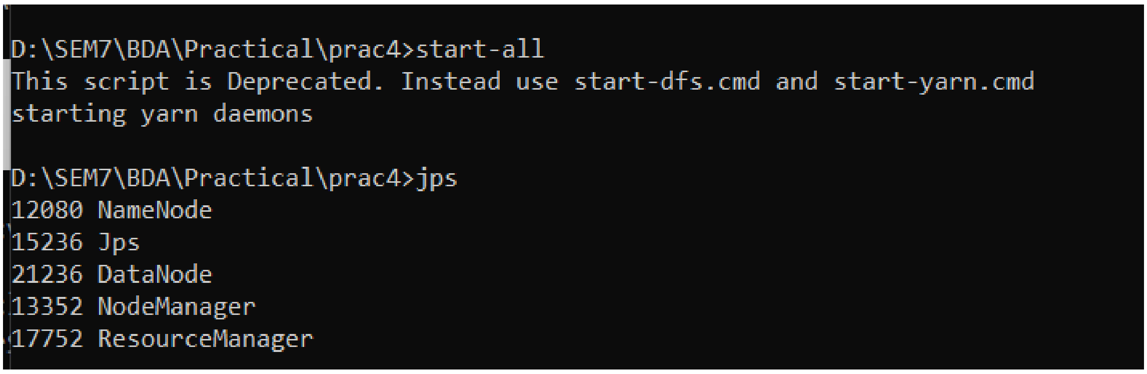


```
    }
  }
}

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();
  GenericOptionsParser optionParser = new GenericOptionsParser(conf, args);
  String[] remainingArgs = optionParser.getRemainingArgs();
  if ((remainingArgs.length != 2) && (remainingArgs.length != 4)) {
    System.err.println("Usage: wordcount <in> <out> [-skip skipPatternFile]");
    System.exit(2);
  }
  Job job = Job.getInstance(conf, "wordcount");
  job.setJarByClass(WordCount2.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);

  List<String> otherArgs = new ArrayList<String>();
  for (int i=0; i < remainingArgs.length; ++i) {
   if ("-skip".equals(remainingArgs[i])) {
    job.addCacheFile(new Path(remainingArgs[++i]).toUri());
    job.getConfiguration().setBoolean("wordcount.skip.patterns", true);
   } else {
    otherArgs.add(remainingArgs[i]);
   }
  }
  FileInputFormat.addInputPath(job, new Path(otherArgs.get(0)));
  FileOutputFormat.setOutputPath(job, new Path(otherArgs.get(1)));

  System.exit(job.waitForCompletion(true) ? 0 : 1);
 }
}
```

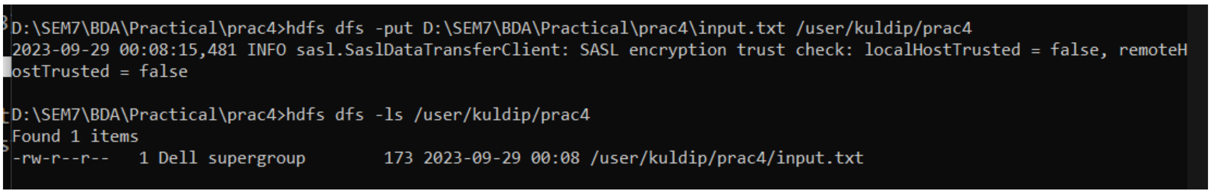
OUTPUT:



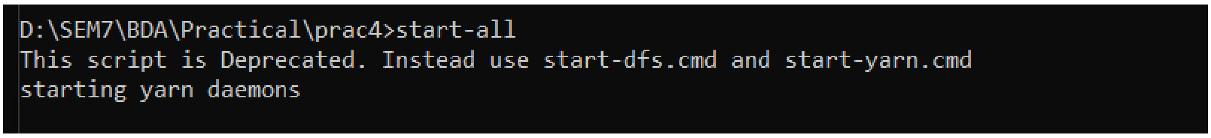
```
D:\SEM7\BDA\Practical\prac4>javac -classpath C:\PROGRA~1\Java\jdk1.8.0_202\lib;D:\SEM7\BDA\hadoop-3.2.1\etc\hadoop;D:\SE
M7\BDA\hadoop-3.2.1\share\hadoop\common;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\common\lib\*;D:\SEM7\BDA\hadoop-3.2.1\shar
e\hadoop\common\*;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\hdfs;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\hdfs\lib\*;D:\SEM7\BD
A\hadoop-3.2.1\share\hadoop\hdfs\*;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\yarn;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\yarn
\lib\*;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\yarn\*;D:\SEM7\BDA\hadoop-3.2.1\share\hadoop\mapreduce\lib\*;D:\SEM7\BDA\ha
doop-3.2.1\share\hadoop\mapreduce\*;C:\PROGRA~1\Java\jdk1.8.0_202\lib\tools.jar; -d D:\SEM7\BDA\Practical\prac4 WordCoun
t.java
```



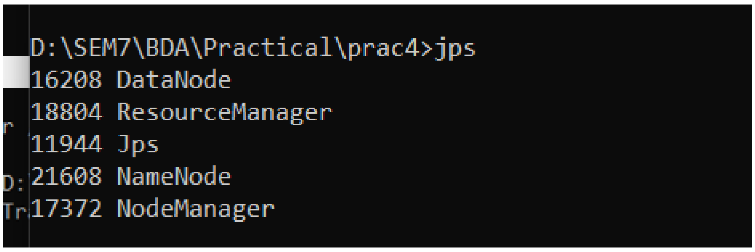
```
D:\SEM7\BDA\Practical\prac4>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

D:\SEM7\BDA\Practical\prac4>jps
12080 NameNode
15236 Jps
21236 DataNode
13352 NodeManager
17752 ResourceManager
```



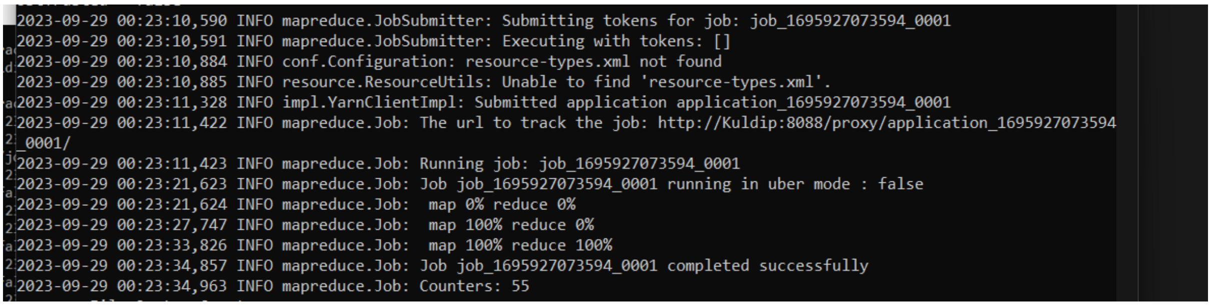
```
D:\SEM7\BDA\Practical\prac4>hdfs dfs -put D:\SEM7\BDA\Practical\prac4\input.txt /user/kuldip/prac4
2023-09-29 00:08:15,481 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false

D:\SEM7\BDA\Practical\prac4>hdfs dfs -ls /user/kuldip/prac4
Found 1 items
-rw-r--r--   1 Dell supergroup        173 2023-09-29 00:08 /user/kuldip/prac4/input.txt
```



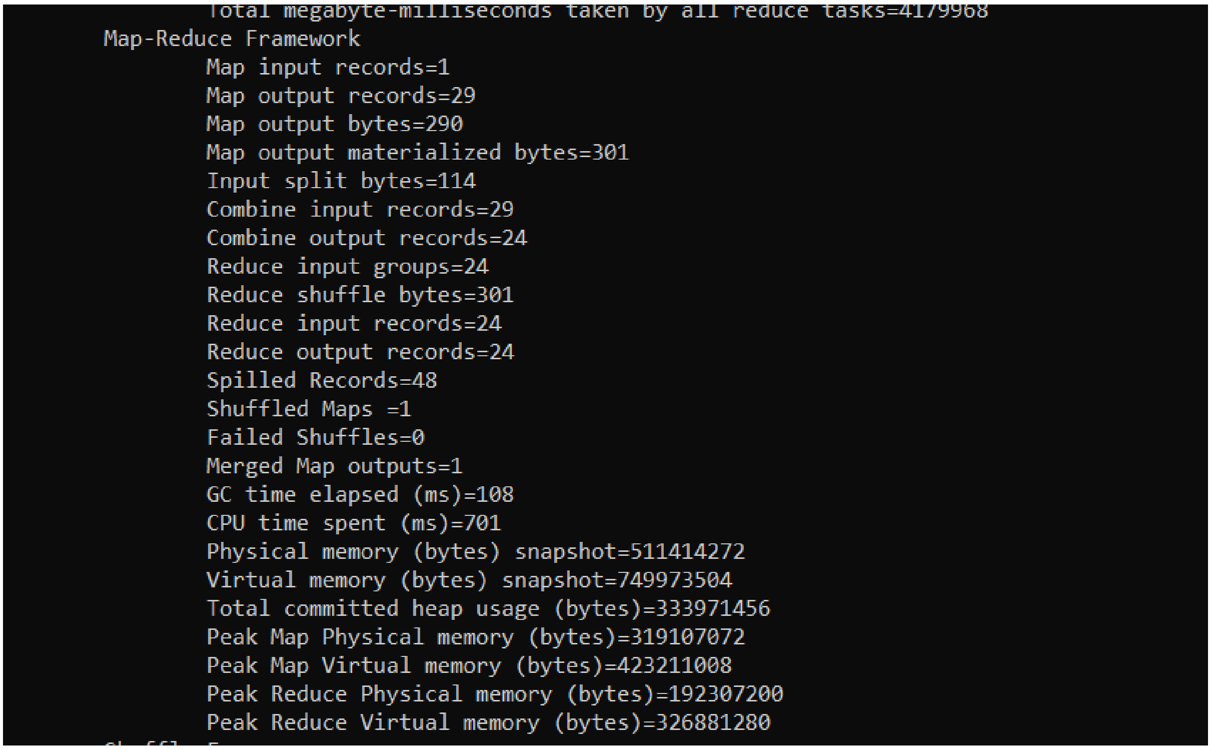
```
D:\SEM7\BDA\Practical\prac4>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
```



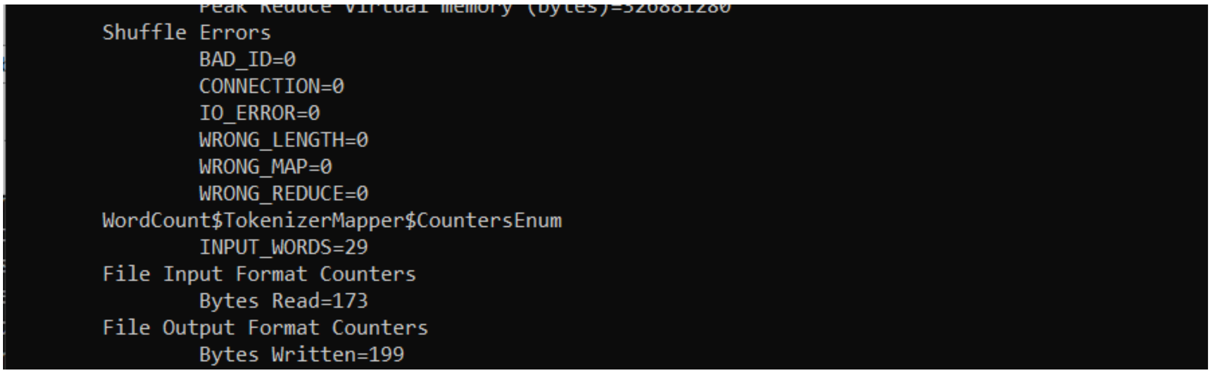
```
D:\SEM7\BDA\Practical\prac4>jps
16208 DataNode
18804 ResourceManager
11944 Jps
21608 NameNode
17372 NodeManager
```



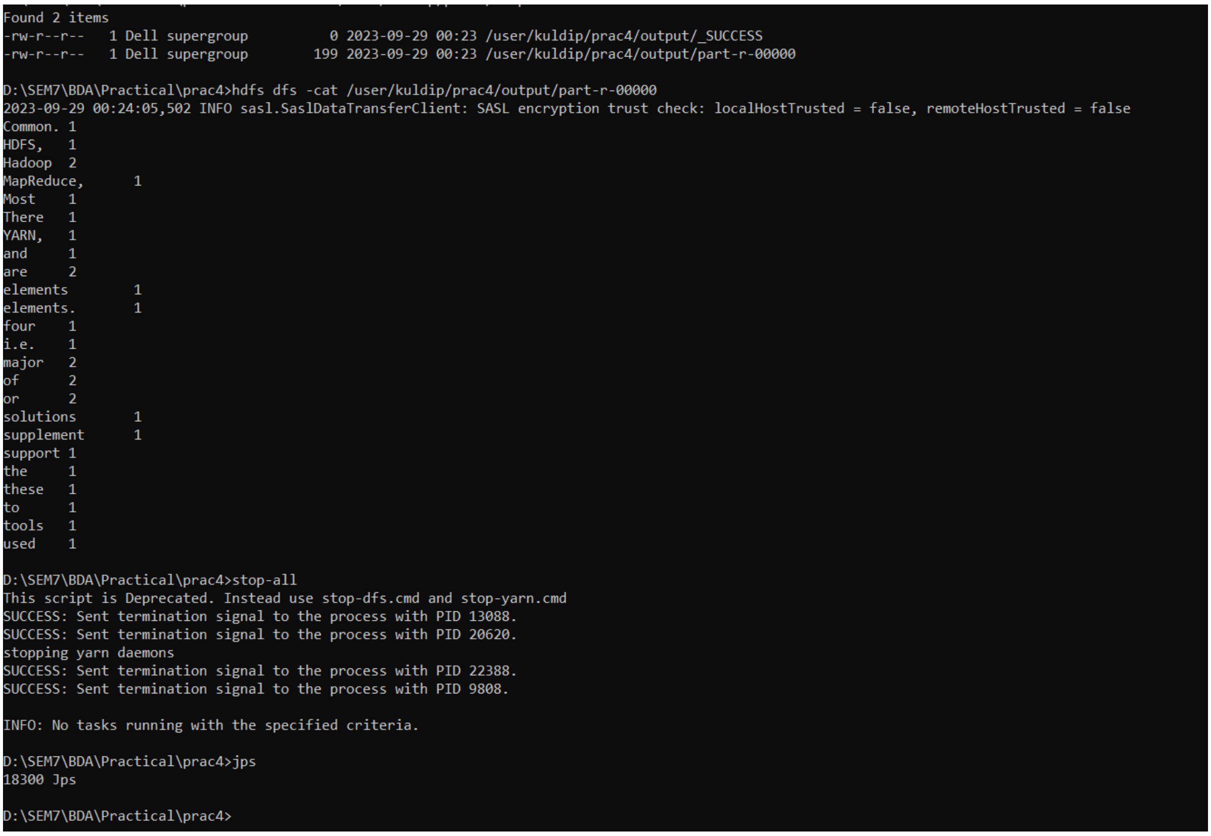
```
2023-09-29 00:23:10,590 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1695927073594_0001
2023-09-29 00:23:10,591 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-29 00:23:10,884 INFO conf.Configuration: resource-types.xml not found
2023-09-29 00:23:10,885 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-09-29 00:23:11,328 INFO impl.YarnClientImpl: Submitted application application_1695927073594_0001
2023-09-29 00:23:11,422 INFO mapreduce.Job: The url to track the job: http://Kuldip:8088/proxy/application_1695927073594
_0001/
2023-09-29 00:23:11,423 INFO mapreduce.Job: Running job: job_1695927073594_0001
2023-09-29 00:23:21,623 INFO mapreduce.Job: Job job_1695927073594_0001 running in uber mode : false
2023-09-29 00:23:21,624 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-29 00:23:27,747 INFO mapreduce.Job:  map 100% reduce 0%
2023-09-29 00:23:33,826 INFO mapreduce.Job:  map 100% reduce 100%
2023-09-29 00:23:34,857 INFO mapreduce.Job: Job job_1695927073594_0001 completed successfully
2023-09-29 00:23:34,963 INFO mapreduce.Job: Counters: 55
```



```
                Total megabyte-milliseconds taken by all reduce tasks=4179968
        Map-Reduce Framework
                Map input records=1
                Map output records=29
                Map output bytes=290
                Map output materialized bytes=301
                Input split bytes=114
                Combine input records=29
                Combine output records=24
                Reduce input groups=24
                Reduce shuffle bytes=301
                Reduce input records=24
                Reduce output records=24
                Spilled Records=48
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=108
                CPU time spent (ms)=701
                Physical memory (bytes) snapshot=511414272
                Virtual memory (bytes) snapshot=749973504
                Total committed heap usage (bytes)=333971456
                Peak Map Physical memory (bytes)=319107072
                Peak Map Virtual memory (bytes)=423211008
                Peak Reduce Physical memory (bytes)=192307200
                Peak Reduce Virtual memory (bytes)=326881280
```



```
                Peak Reduce Virtual memory (bytes)=326881280
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        WordCount$TokenizerMapper$CountersEnum
                INPUT_WORDS=29
        File Input Format Counters
                Bytes Read=173
        File Output Format Counters
                Bytes Written=199
```



```
Found 2 items
-rw-r--r--   1 Dell supergroup          0 2023-09-29 00:23 /user/kuldip/prac4/output/_SUCCESS
-rw-r--r--   1 Dell supergroup        199 2023-09-29 00:23 /user/kuldip/prac4/output/part-r-00000

D:\SEM7\BDA\Practical\prac4>hdfs dfs -cat /user/kuldip/prac4/output/part-r-00000
2023-09-29 00:24:05,502 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Common. 1
HDFS,   1
Hadoop  2
MapReduce,      1
Most    1
There   1
YARN,   1
and     1
are     2
elements        1
elements.       1
four    1
i.e.    1
major   2
of      2
or      2
solutions       1
supplement      1
support 1
the     1
these   1
to      1
tools   1
used    1

D:\SEM7\BDA\Practical\prac4>stop-all
This script is Deprecated. Instead use stop-dfs.cmd and stop-yarn.cmd
SUCCESS: Sent termination signal to the process with PID 13088.
SUCCESS: Sent termination signal to the process with PID 20620.
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 22388.
SUCCESS: Sent termination signal to the process with PID 9808.

INFO: No tasks running with the specified criteria.

D:\SEM7\BDA\Practical\prac4>jps
18300 Jps

D:\SEM7\BDA\Practical\prac4>
```