FACULTY OF ENGINEERING

Word Count application

Distributed Systems

Presented by

1- **Seif Eldeen Ehab Mostafa 32**
2- **Islam Yousry Abdelwahid 14**
3- **Andrew Adel Sanad 17**

6/3/2022

Table of Contents

1 Problem Definition

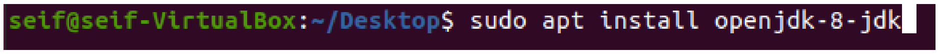
After installing Hadoop as a one node cluster in a pseudo distributed mode, it is required to download the Plain Text UTF-8 format data set then create a java word count application that runs as a Hadoop job on the downloaded data in order to find the count of each word found in the documents.

2 Algorithms

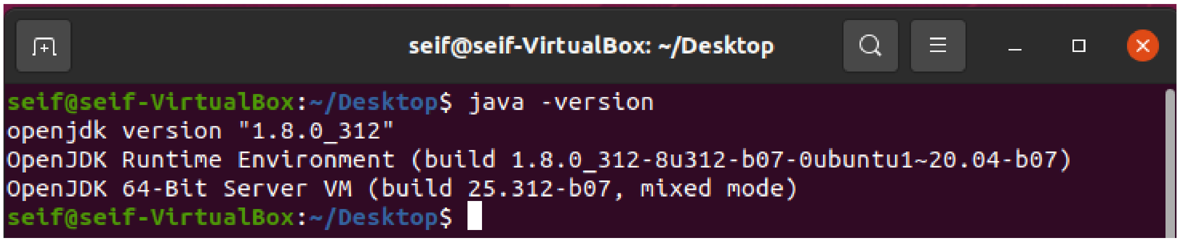
- Hadoop installation is done on Ubuntu 20.04 VM with the appropriate commands.
- The WordCount java application is implemented as described for the Hadoop job.
- After creating the appropriate directories on both Hadoop and locally we compile the java application and run the jar program on the text data we have.
- Output is saved and compared to the desired output.

3 Implementation

- Hadoop installation:
    - Download and extract the downloaded Hadoop file
    - Installing java

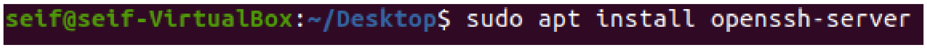


```
seif@seif-VirtualBox:~/Desktop$ sudo apt install openjdk-8-jdk
```

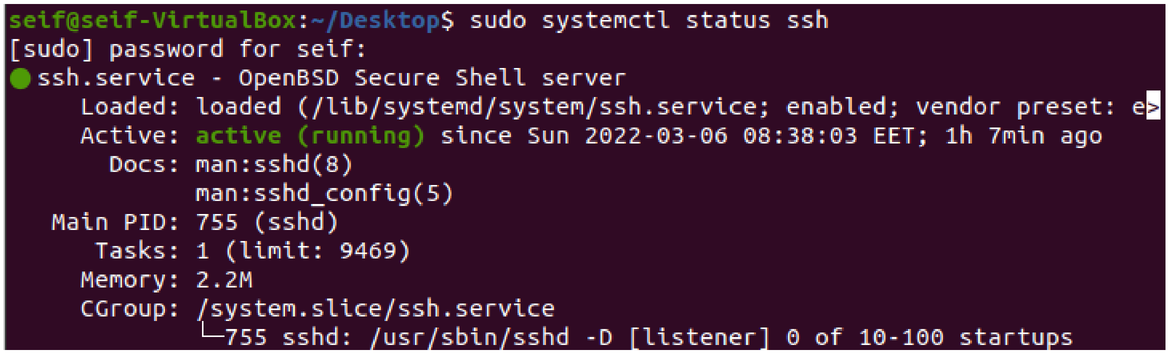


```
seif@seif-VirtualBox:~/Desktop$ java -version
openjdk version "1.8.0_312"
OpenJDK Runtime Environment (build 1.8.0_312-8u312-b07-0ubuntu1~20.04-b07)
OpenJDK 64-Bit Server VM (build 25.312-b07, mixed mode)
seif@seif-VirtualBox:~/Desktop$
```

- Installing ssh

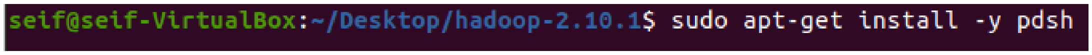


```
seif@seif-VirtualBox:~/Desktop$ sudo apt install openssh-server
```



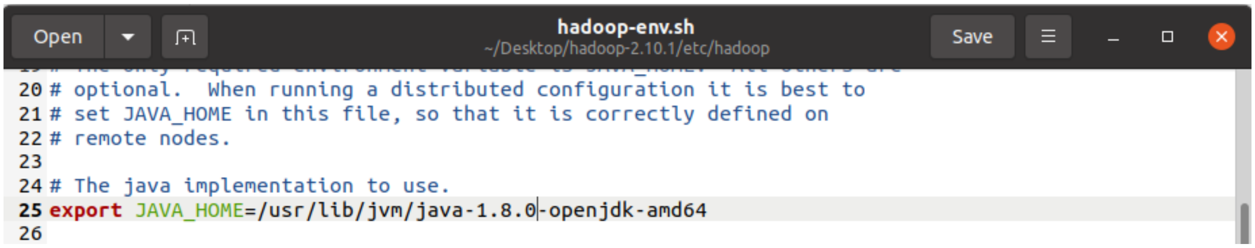
```
seif@seif-VirtualBox:~/Desktop$ sudo systemctl status ssh
[sudo] password for seif:
● ssh.service - OpenBSD Secure Shell server
     Loaded: loaded (/lib/systemd/system/ssh.service; enabled; vendor preset: e>
     Active: active (running) since Sun 2022-03-06 08:38:03 EET; 1h 7min ago
       Docs: man:sshd(8)
             man:sshd_config(5)
   Main PID: 755 (sshd)
      Tasks: 1 (limit: 9469)
     Memory: 2.2M
     CGroup: /system.slice/ssh.service
             └─755 sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups
```

- Installing pdsh



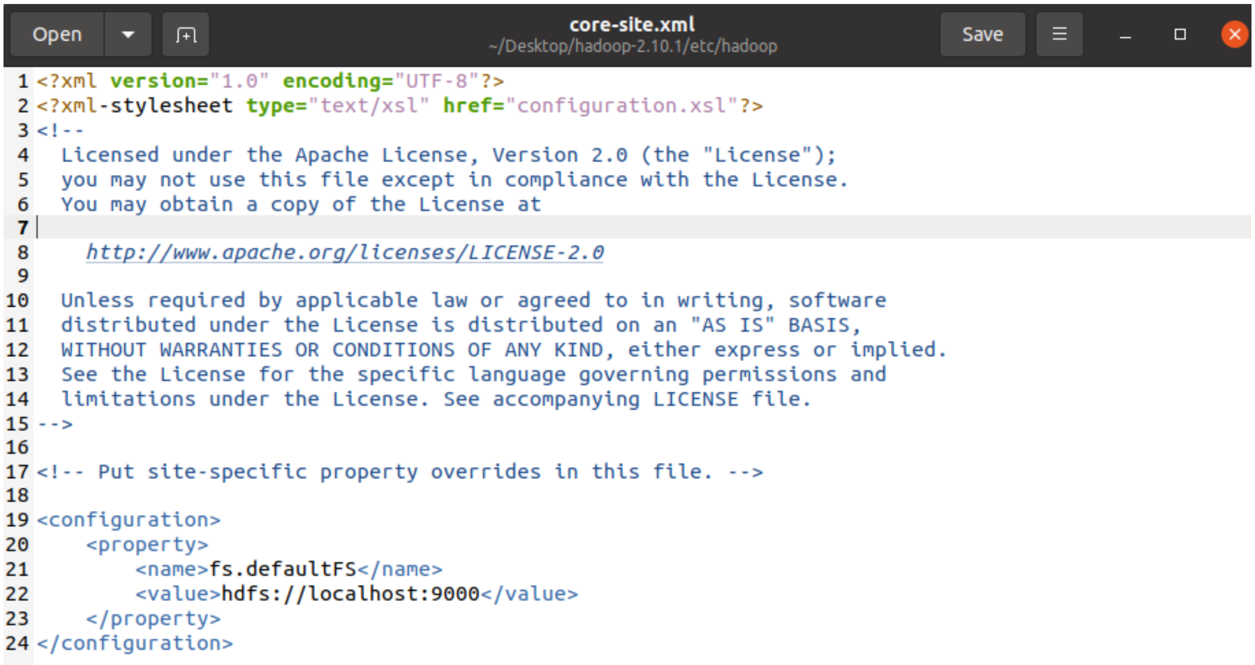
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ sudo apt-get install -y pdsh
```

- Assigning JAVA_HOME

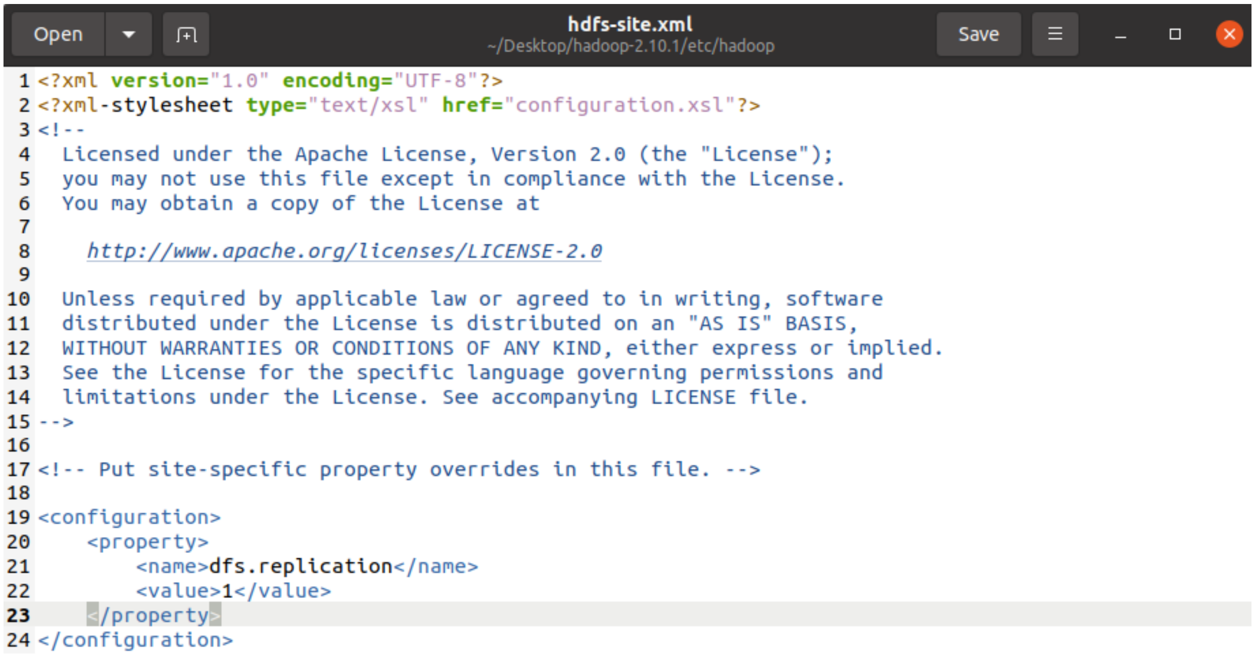


```
# optional.  When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64

```

- Pseudo Distribution configuration

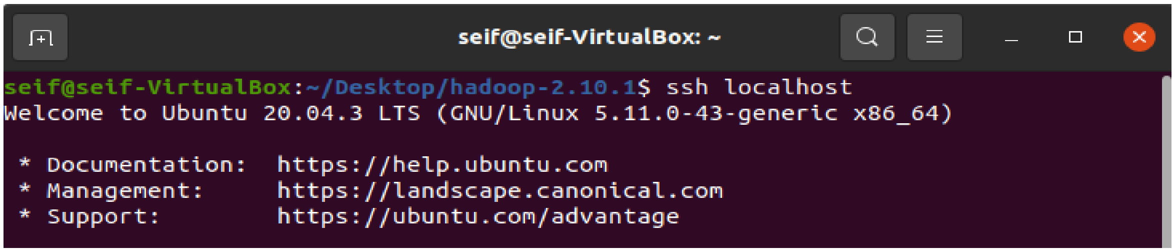


```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```



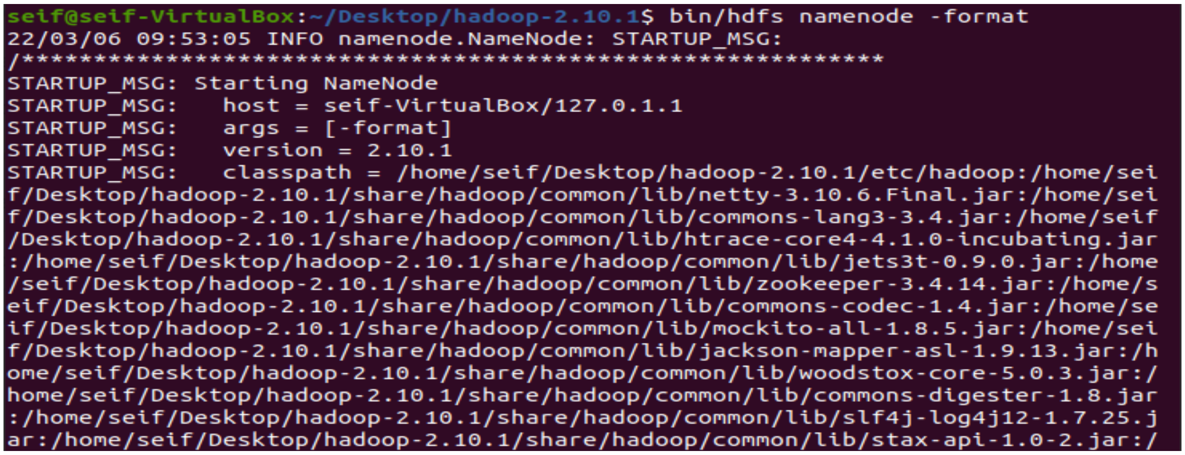
```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

- Passphraseless ssh



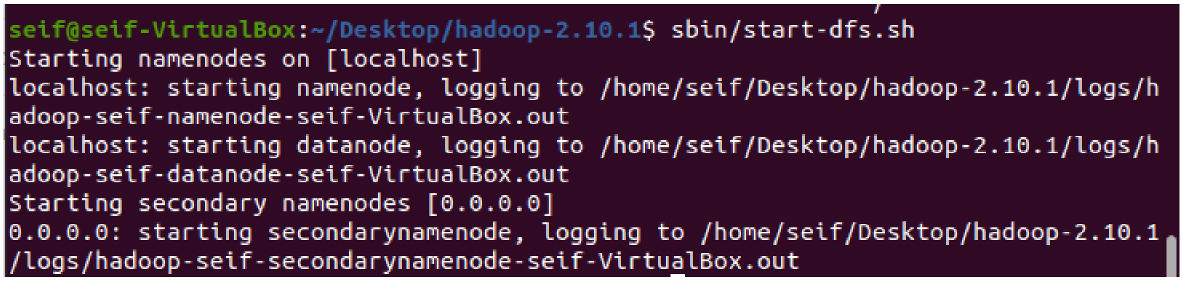
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ ssh localhost
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.11.0-43-generic x86_64)

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

- HDFS setup:
  - Format the filesystem



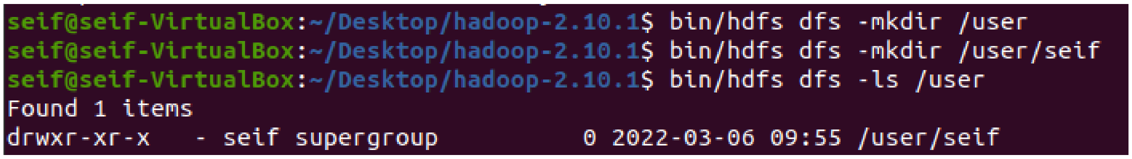
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs namenode -format
22/03/06 09:53:05 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = seif-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.10.1
STARTUP_MSG:   classpath = /home/seif/Desktop/hadoop-2.10.1/etc/hadoop:/home/sei
f/Desktop/hadoop-2.10.1/share/hadoop/common/lib/netty-3.10.6.Final.jar:/home/sei
f/Desktop/hadoop-2.10.1/share/hadoop/common/lib/commons-lang3-3.4.jar:/home/seif
/Desktop/hadoop-2.10.1/share/hadoop/common/lib/htrace-core4-4.1.0-incubating.jar
:/home/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/jets3t-0.9.0.jar:/home
/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/zookeeper-3.4.14.jar:/home/s
eif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/commons-codec-1.4.jar:/home/se
if/Desktop/hadoop-2.10.1/share/hadoop/common/lib/mockito-all-1.8.5.jar:/home/sei
f/Desktop/hadoop-2.10.1/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/h
ome/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/woodstox-core-5.0.3.jar:/
home/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/commons-digester-1.8.jar
:/home/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.j
ar:/home/seif/Desktop/hadoop-2.10.1/share/hadoop/common/lib/stax-api-1.0-2.jar:/
```

  - Start NameNode daemon and DataNode daemon



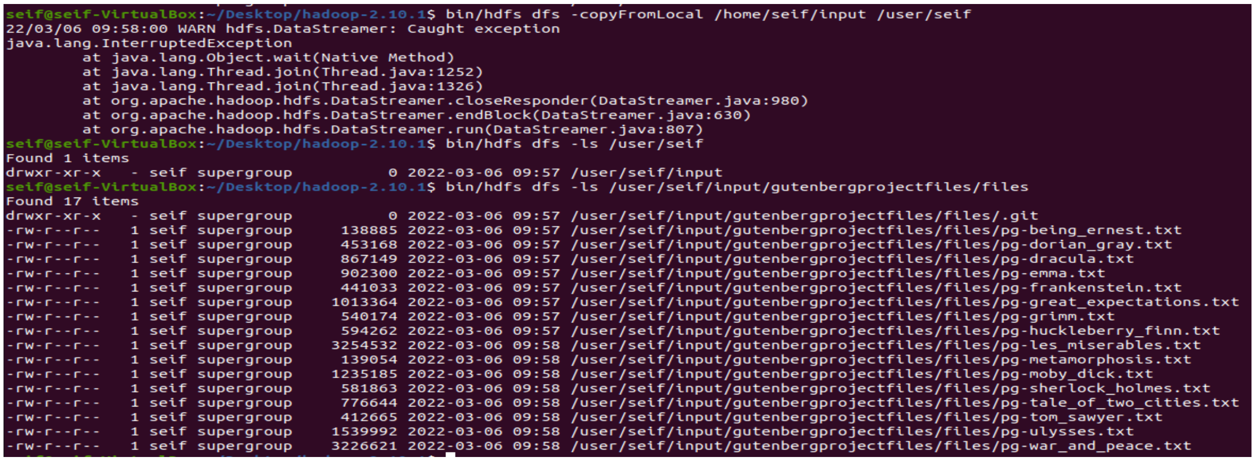
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/seif/Desktop/hadoop-2.10.1/logs/h
adoop-seif-namenode-seif-VirtualBox.out
localhost: starting datanode, logging to /home/seif/Desktop/hadoop-2.10.1/logs/h
adoop-seif-datanode-seif-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/seif/Desktop/hadoop-2.10.1
/logs/hadoop-seif-secondarynamenode-seif-VirtualBox.out
```

  - User HDFS directories



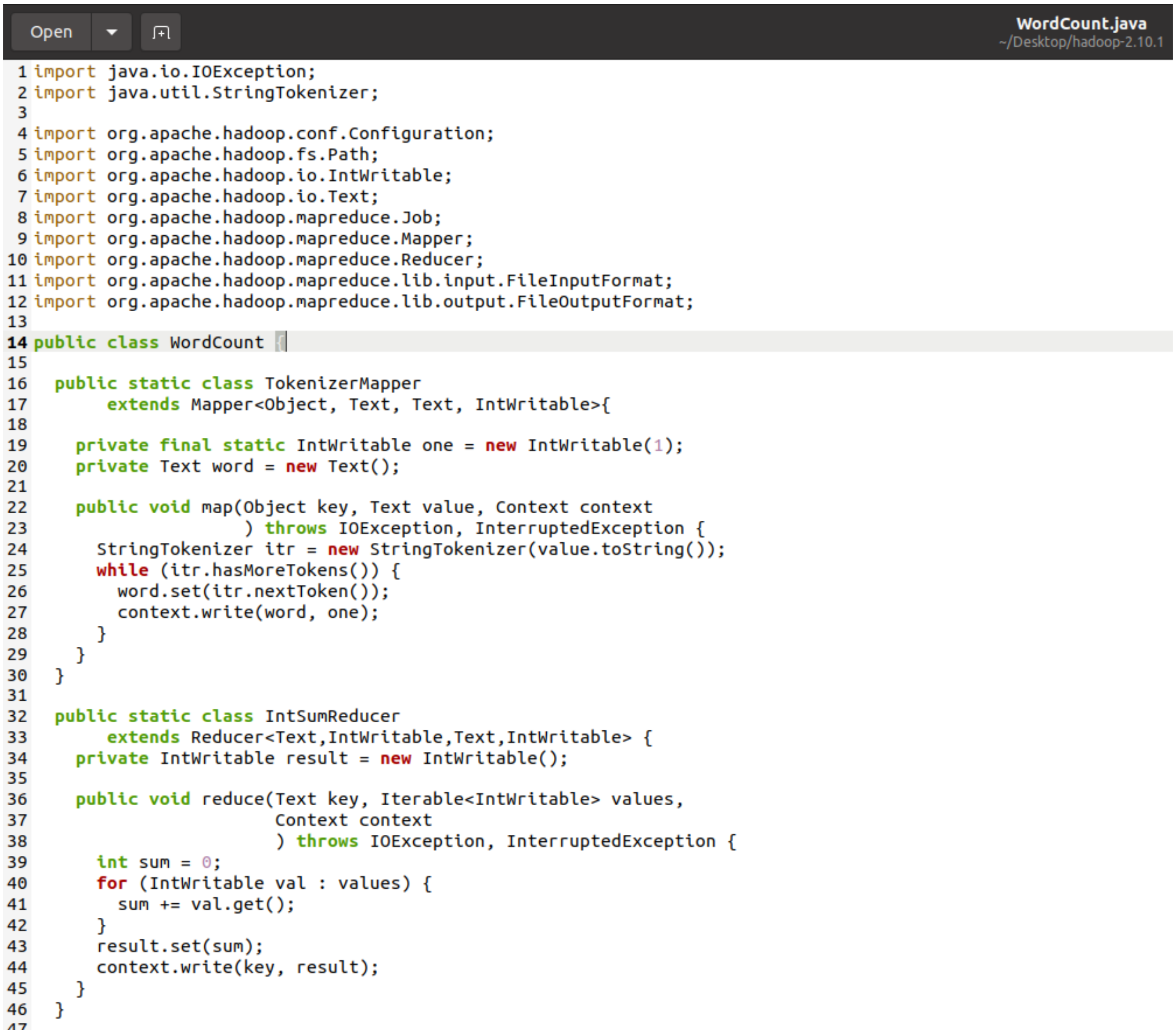
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -mkdir /user
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -mkdir /user/seif
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - seif supergroup          0 2022-03-06 09:55 /user/seif
```

  - Copy the text data file from local system to HDFS input directory

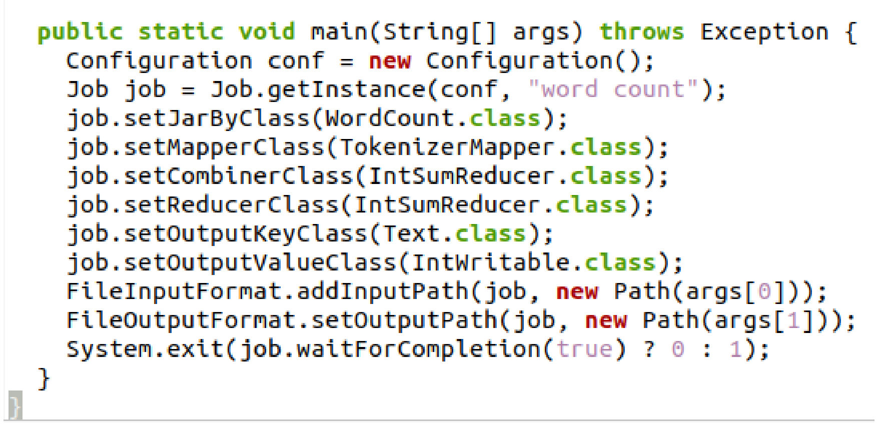


```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -copyFromLocal /home/seif/input /user/seif
22/03/06 09:58:00 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -ls /user/seif
Found 1 items
drwxr-xr-x   - seif supergroup          0 2022-03-06 09:57 /user/seif/input
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hdfs dfs -ls /user/seif/input/gutenbergprojectfiles/files
Found 17 items
drwxr-xr-x   - seif supergroup          0 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/.git
-rw-r--r--   1 seif supergroup     138885 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-being_ernest.txt
-rw-r--r--   1 seif supergroup     453168 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-dorian_gray.txt
-rw-r--r--   1 seif supergroup     867149 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-dracula.txt
-rw-r--r--   1 seif supergroup     902300 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-emma.txt
-rw-r--r--   1 seif supergroup     441033 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-frankenstein.txt
-rw-r--r--   1 seif supergroup    1013364 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-great_expectations.txt
-rw-r--r--   1 seif supergroup     540174 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-grimm.txt
-rw-r--r--   1 seif supergroup     594262 2022-03-06 09:57 /user/seif/input/gutenbergprojectfiles/files/pg-huckleberry_finn.txt
-rw-r--r--   1 seif supergroup    3254532 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-les_miserables.txt
-rw-r--r--   1 seif supergroup     139054 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-metamorphosis.txt
-rw-r--r--   1 seif supergroup    1235185 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-moby_dick.txt
-rw-r--r--   1 seif supergroup     581863 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-sherlock_holmes.txt
-rw-r--r--   1 seif supergroup     776644 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-tale_of_two_cities.txt
-rw-r--r--   1 seif supergroup     412665 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-tom_sawyer.txt
-rw-r--r--   1 seif supergroup    1539992 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-ulysses.txt
-rw-r--r--   1 seif supergroup    3226621 2022-03-06 09:58 /user/seif/input/gutenbergprojectfiles/files/pg-war_and_peace.txt
```

- WordCount application:
    - The code

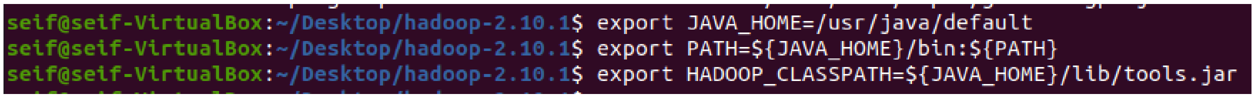


```java
import java.io.IOException;
import java.util.StringTokenizer;

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);
    }
  }
```



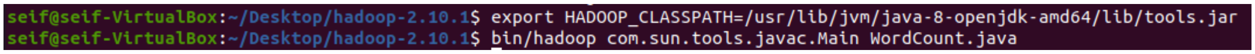
```java
  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = Job.getInstance(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    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);
  }
}
```

- Execution:
    - Set environment variables



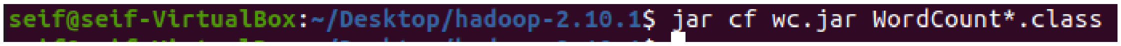
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ export JAVA_HOME=/usr/java/default
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ export PATH=${JAVA_HOME}/bin:${PATH}
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
```

    - Compiling the java application



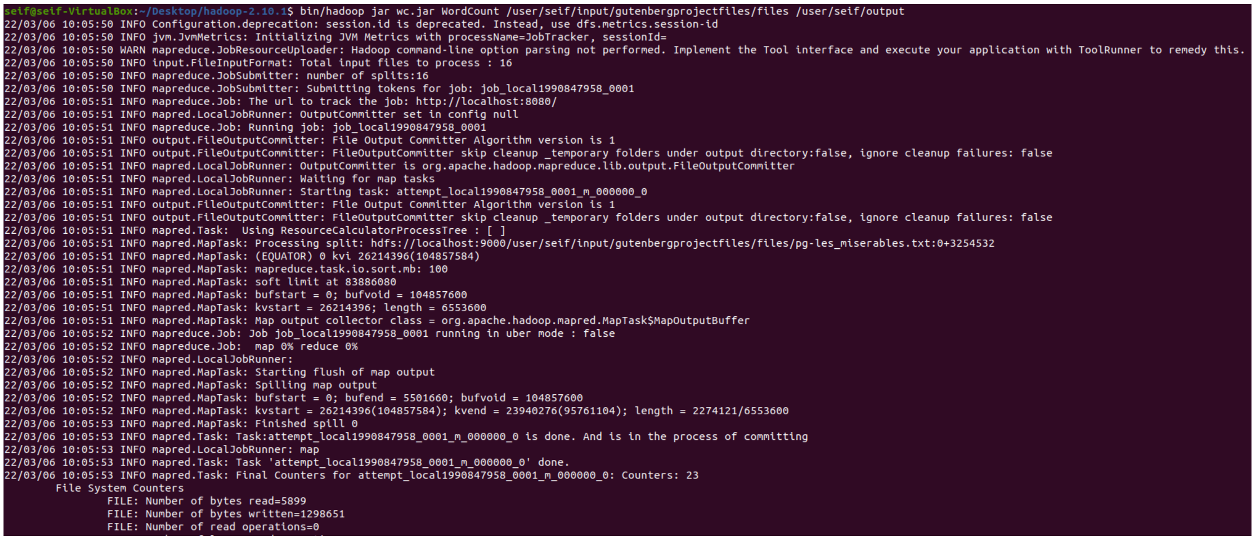
```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ export HADOOP_CLASSPATH=/usr/lib/jvm/java-8-openjdk-amd64/lib/tools.jar
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hadoop com.sun.tools.javac.Main WordCount.java
```

    - Create JAR

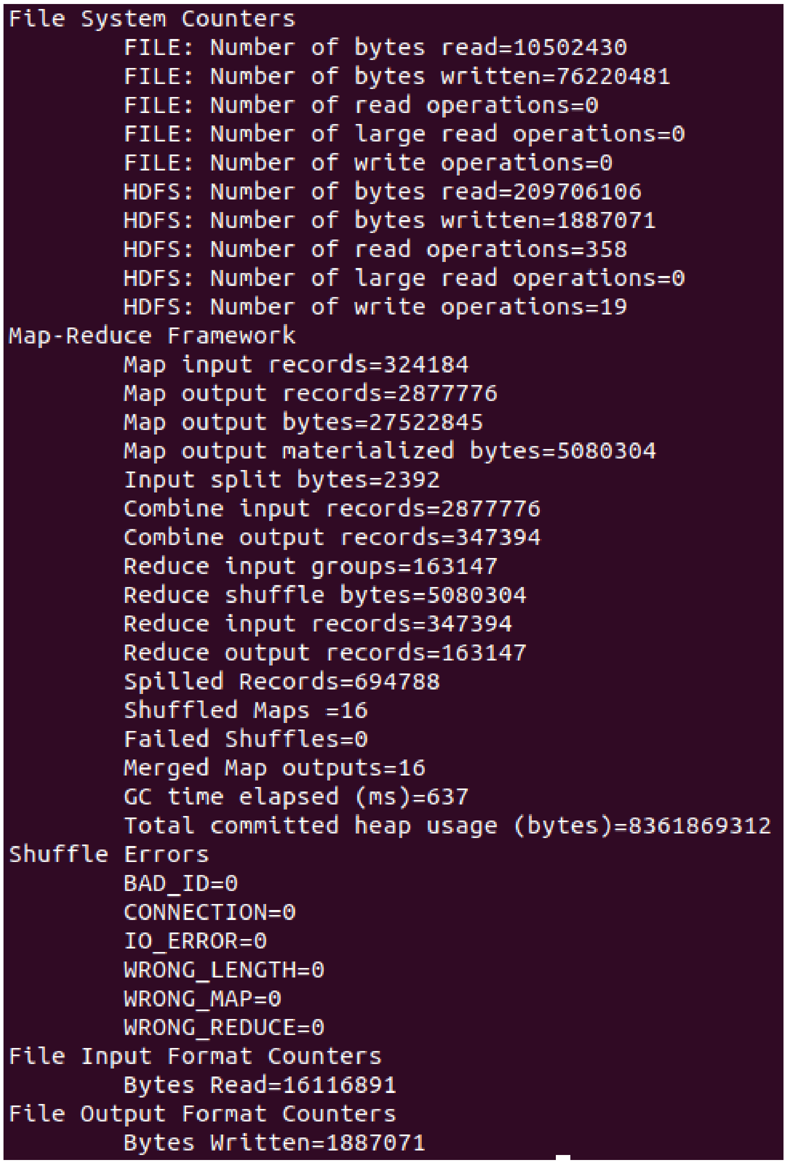


```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ jar cf wc.jar WordCount*.class
```

    - Run the WordCount application

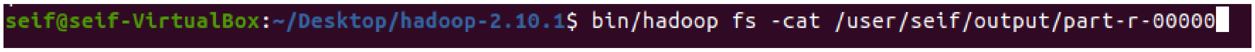


```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hadoop jar wc.jar WordCount /user/seif/input/gutenbergprojectfiles/files /user/seif/output
22/03/06 10:05:50 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
22/03/06 10:05:50 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
22/03/06 10:05:50 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/03/06 10:05:50 INFO input.FileInputFormat: Total input files to process : 16
22/03/06 10:05:50 INFO mapreduce.JobSubmitter: number of splits:16
22/03/06 10:05:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1990847958_0001
22/03/06 10:05:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
22/03/06 10:05:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
22/03/06 10:05:51 INFO mapreduce.Job: Running job: job_local1990847958_0001
22/03/06 10:05:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
22/03/06 10:05:51 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
22/03/06 10:05:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
22/03/06 10:05:51 INFO mapred.LocalJobRunner: Waiting for map tasks
22/03/06 10:05:51 INFO mapred.LocalJobRunner: Starting task: attempt_local1990847958_0001_m_000000_0
22/03/06 10:05:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
22/03/06 10:05:51 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
22/03/06 10:05:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/03/06 10:05:51 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/seif/input/gutenbergprojectfiles/files/pg-les_miserables.txt:0+3254532
22/03/06 10:05:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/03/06 10:05:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
22/03/06 10:05:51 INFO mapred.MapTask: soft limit at 83886080
22/03/06 10:05:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/03/06 10:05:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
22/03/06 10:05:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/03/06 10:05:52 INFO mapreduce.Job: Job job_local1990847958_0001 running in uber mode : false
22/03/06 10:05:52 INFO mapreduce.Job:  map 0% reduce 0%
22/03/06 10:05:52 INFO mapred.LocalJobRunner: 
22/03/06 10:05:52 INFO mapred.MapTask: Starting flush of map output
22/03/06 10:05:52 INFO mapred.MapTask: Spilling map output
22/03/06 10:05:52 INFO mapred.MapTask: bufstart = 0; bufend = 5501660; bufvoid = 104857600
22/03/06 10:05:52 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 23940276(95761104); length = 2274121/6553600
22/03/06 10:05:53 INFO mapred.MapTask: Finished spill 0
22/03/06 10:05:53 INFO mapred.Task: Task:attempt_local1990847958_0001_m_000000_0 is done. And is in the process of committing
22/03/06 10:05:53 INFO mapred.LocalJobRunner: map
22/03/06 10:05:53 INFO mapred.Task: Task 'attempt_local1990847958_0001_m_000000_0' done.
22/03/06 10:05:53 INFO mapred.Task: Final Counters for attempt_local1990847958_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=5899
                FILE: Number of bytes written=1298651
                FILE: Number of read operations=0
```

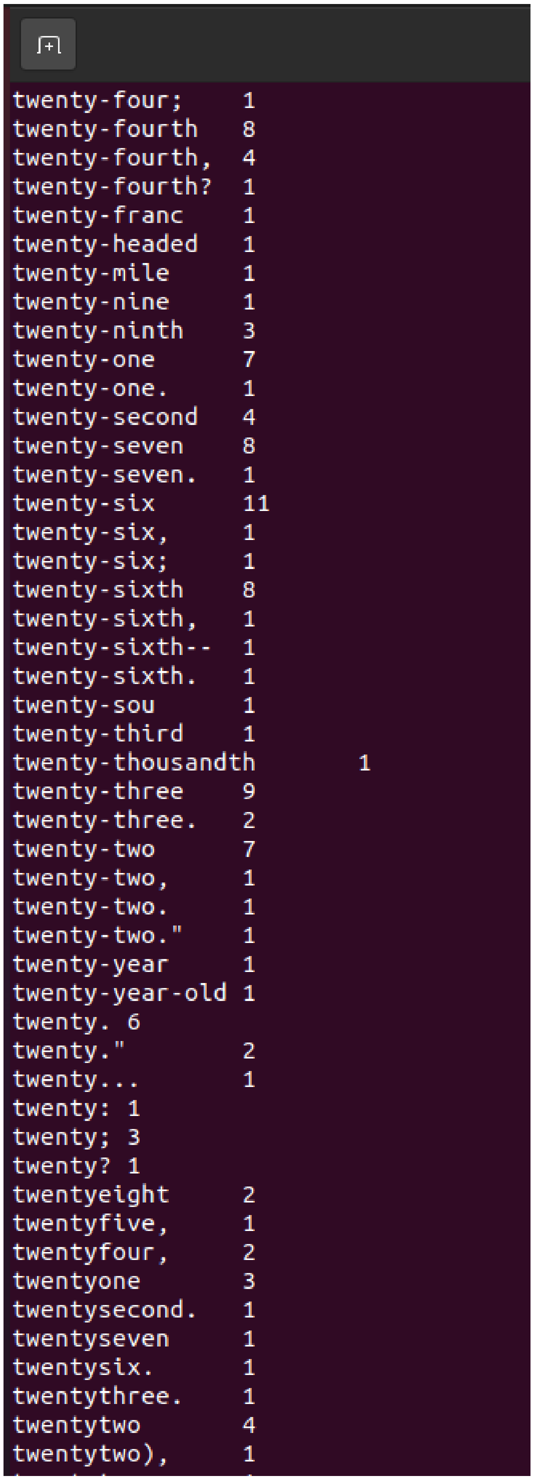


```
File System Counters
        FILE: Number of bytes read=10502430
        FILE: Number of bytes written=76220481
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=209706106
        HDFS: Number of bytes written=1887071
        HDFS: Number of read operations=358
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=19
Map-Reduce Framework
        Map input records=324184
        Map output records=2877776
        Map output bytes=27522845
        Map output materialized bytes=5080304
        Input split bytes=2392
        Combine input records=2877776
        Combine output records=347394
        Reduce input groups=163147
        Reduce shuffle bytes=5080304
        Reduce input records=347394
        Reduce output records=163147
        Spilled Records=694788
        Shuffled Maps =16
        Failed Shuffles=0
        Merged Map outputs=16
        GC time elapsed (ms)=637
        Total committed heap usage (bytes)=8361869312
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=16116891
File Output Format Counters
        Bytes Written=1887071
```

○ Output

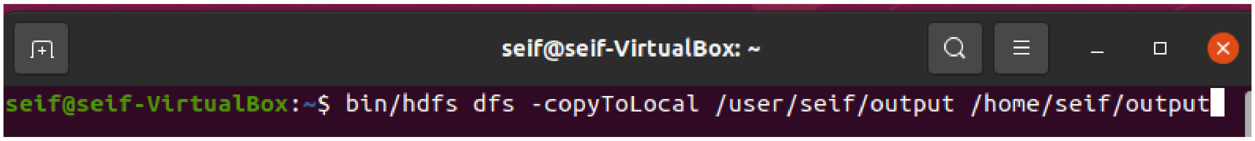


```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ bin/hadoop fs -cat /user/seif/output/part-r-00000
```



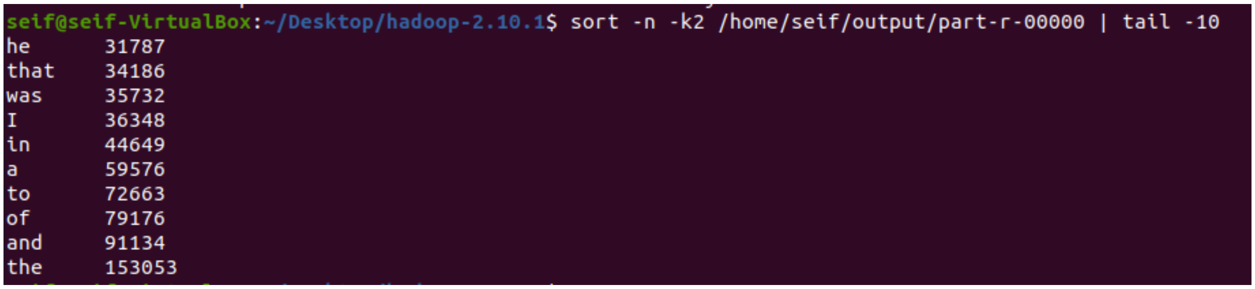
```
twenty-four;    1
twenty-fourth   8
twenty-fourth,  4
twenty-fourth?  1
twenty-franc    1
twenty-headed   1
twenty-mile     1
twenty-nine     1
twenty-ninth    3
twenty-one      7
twenty-one.     1
twenty-second   4
twenty-seven    8
twenty-seven.   1
twenty-six      11
twenty-six,     1
twenty-six;     1
twenty-sixth    8
twenty-sixth,   1
twenty-sixth--  1
twenty-sixth.   1
twenty-sou      1
twenty-third    1
twenty-thousandth       1
twenty-three    9
twenty-three.   2
twenty-two      7
twenty-two,     1
twenty-two.     1
twenty-two."    1
twenty-year     1
twenty-year-old 1
twenty. 6
twenty."        2
twenty...       1
twenty: 1
twenty; 3
twenty? 1
twentyeight     2
twentyfive,     1
twentyfour,     2
twentyone       3
twentysecond.   1
twentyseven     1
twentysix.      1
twentythree.    1
twentytwo       4
twentytwo),     1
```

- Copy output to local



```
seif@seif-VirtualBox:~$ bin/hdfs dfs -copyToLocal /user/seif/output /home/seif/output
```

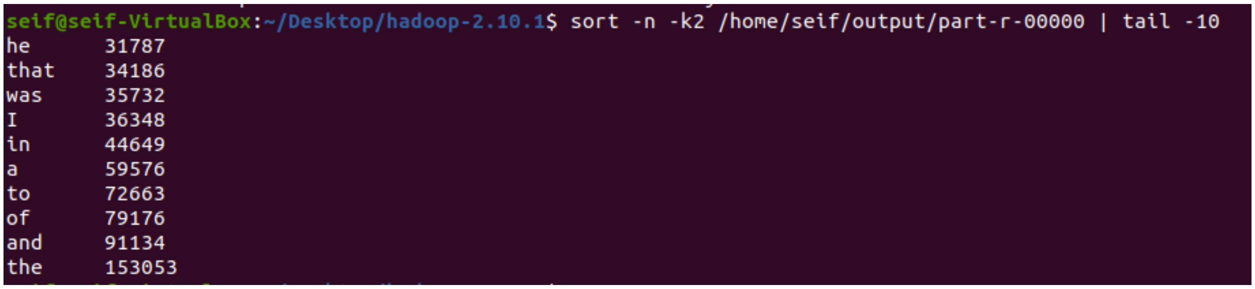
- Compare with desired output



```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ sort -n -k2 /home/seif/output/part-r-00000 | tail -10
he      31787
that    34186
was     35732
I       36348
in      44649
a       59576
to      72663
of      79176
and     91134
the     153053
```

4 Results

The sorted output matches the desired output.



```
seif@seif-VirtualBox:~/Desktop/hadoop-2.10.1$ sort -n -k2 /home/seif/output/part-r-00000 | tail -10
he      31787
that    34186
was     35732
I       36348
in      44649
a       59576
to      72663
of      79176
and     91134
the     153053
```

5 Conclusion

- We can now use a MapReduce programming model using Hadoop in pseudo distributed mode in creating many applications using Java programming language like the WordCount application and test it on as much data as we want.
- We would have changed the application to ignore some unimportant words and symbols and ignore words that are count less than a certain threshold.