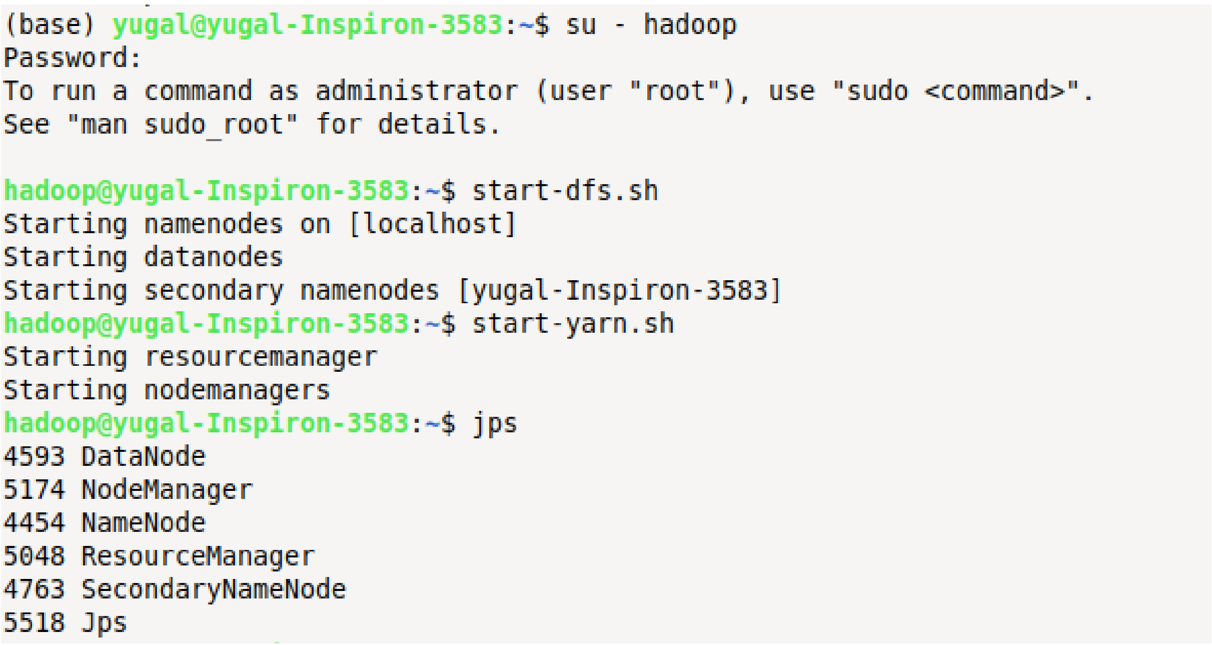
Assignment 3: Implement MapReduce programs for Data Processing.

Objective: Apply distributed computing concepts by implementing a MapReduce program in Hadoop to process data efficiently across distributed nodes.

3.1 Step 1: Start Hadoop Services

```
su - hadoop
start-dfs.sh
start-yarn.sh
jps
```

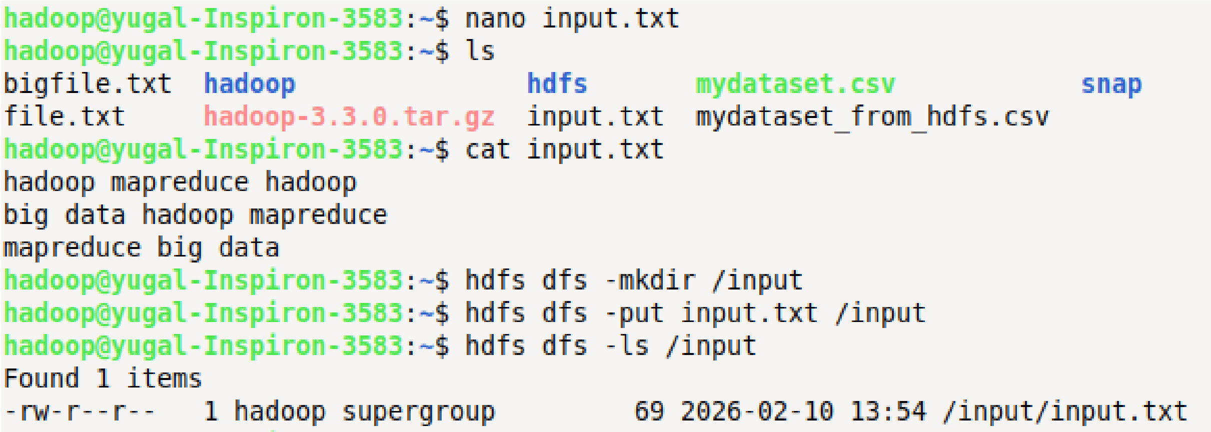


```
(base) yugal@yugal-Inspiron-3583:~$ su - hadoop
Password:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

hadoop@yugal-Inspiron-3583:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [yugal-Inspiron-3583]
hadoop@yugal-Inspiron-3583:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@yugal-Inspiron-3583:~$ jps
4593 DataNode
5174 NodeManager
4454 NameNode
5048 ResourceManager
4763 SecondaryNameNode
5518 Jps
```

Figure 31: Starting Hadoop services and verifying with jps

3.2 Step 2: Create Input Directory in HDFS

```
nano input.txt
ls
cat input.txt
hdfs dfs -mkdir /input
hdfs dfs -put input.txt /input
hdfs dfs -ls /input
```



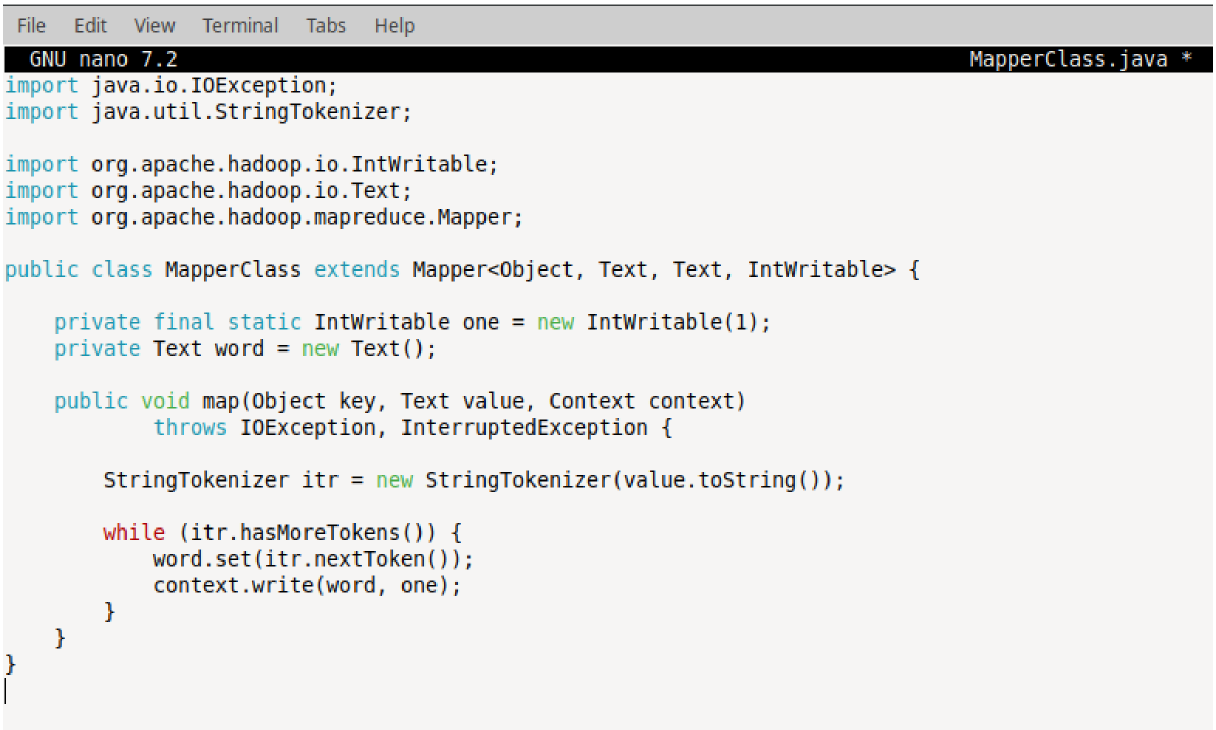
```
hadoop@yugal-Inspiron-3583:~$ nano input.txt
hadoop@yugal-Inspiron-3583:~$ ls
bigfile.txt  hadoop               hdfs       mydataset.csv            snap
file.txt     hadoop-3.3.0.tar.gz  input.txt  mydataset_from_hdfs.csv
hadoop@yugal-Inspiron-3583:~$ cat input.txt
hadoop mapreduce hadoop
big data hadoop mapreduce
mapreduce big data
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -mkdir /input
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -put input.txt /input
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /input
Found 1 items
-rw-r--r--   1 hadoop supergroup         69 2026-02-10 13:54 /input/input.txt
```

Figure 32: Creating input directory and uploading dataset

3.3 Step 3: Write MapReduce Program

Mapper Class: Processes input data and generates intermediate key-value pairs.

```
nano Mapper.java
```

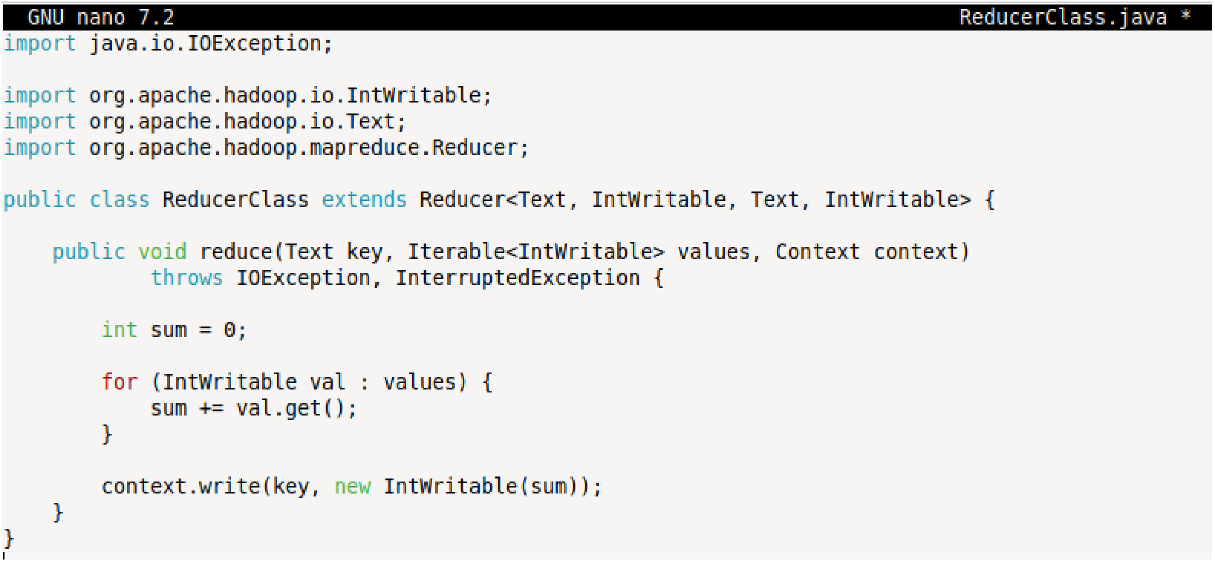


```
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class MapperClass 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);
        }
    }
}
```

Figure 33: Mapper class implementation

Reducer Class: Aggregates values for each key produced by the mapper.

```
nano Reducer.java
```

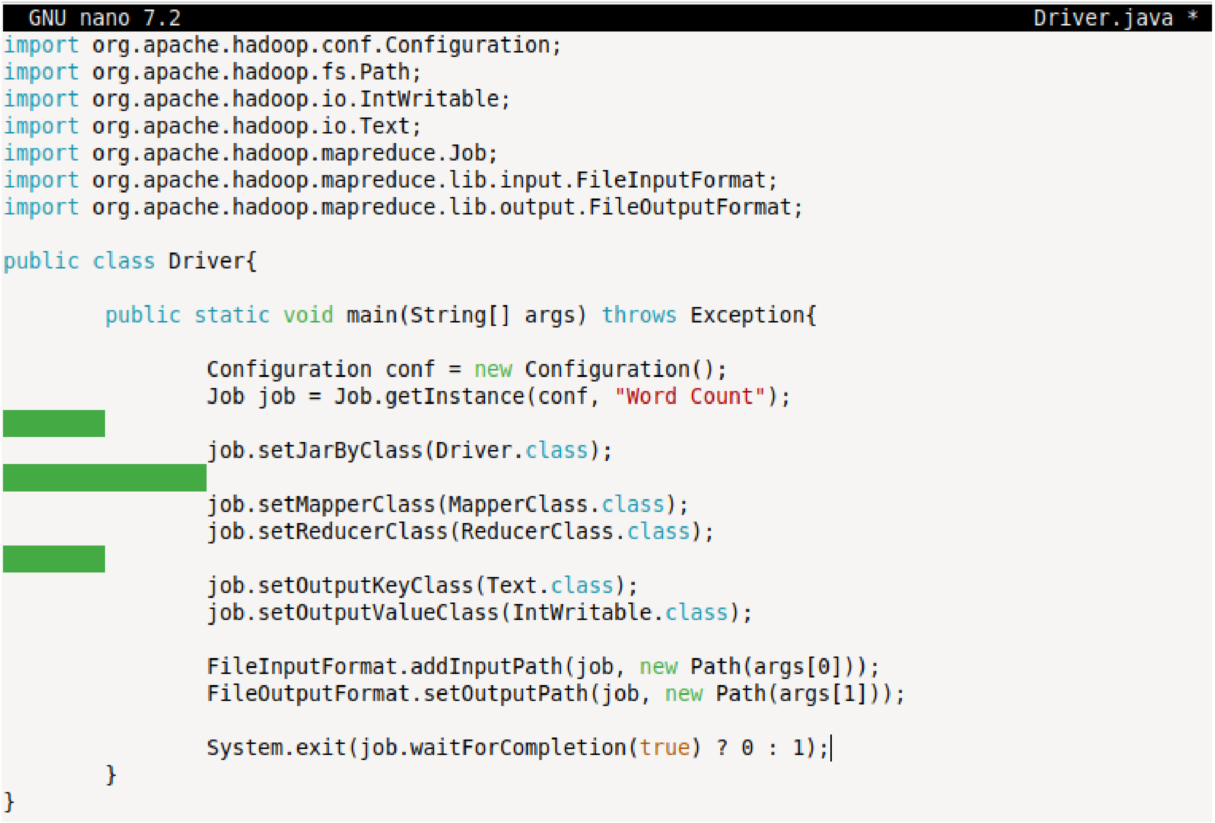


```
  GNU nano 7.2                                                         ReducerClass.java *
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class ReducerClass extends Reducer<Text, IntWritable, Text, IntWritable> {

    public void reduce(Text key, Iterable<IntWritable> values, Context context)
            throws IOException, InterruptedException {

        int sum = 0;

        for (IntWritable val : values) {
            sum += val.get();
        }

        context.write(key, new IntWritable(sum));
    }
}
```

Figure 34: Reducer class implementation

Driver Class: Configures job parameters such as input/output paths, mapper, reducer, and execution settings.

```
nano Driver.java
```

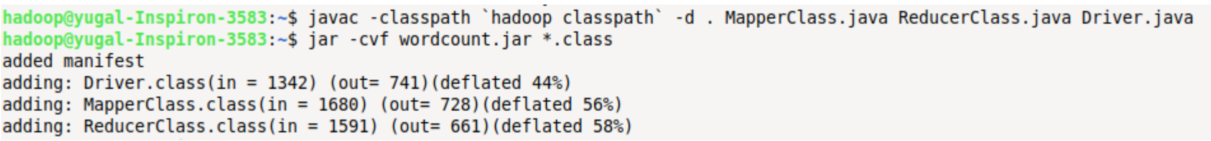


```
  GNU nano 7.2                                                              Driver.java *
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class Driver{

        public static void main(String[] args) throws Exception{

                Configuration conf = new Configuration();
                Job job = Job.getInstance(conf, "Word Count");

                job.setJarByClass(Driver.class);

                job.setMapperClass(MapperClass.class);
                job.setReducerClass(ReducerClass.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);
        }
}
```

Figure 35: Driver class implementation

3.4 Step 4: Compile Program and Create JAR

```
javac -classpath ‘hadoop classpath‘ -d . Mapper.java Reducer.java Driver.java
jar -cvf wordcount.jar *.class
```

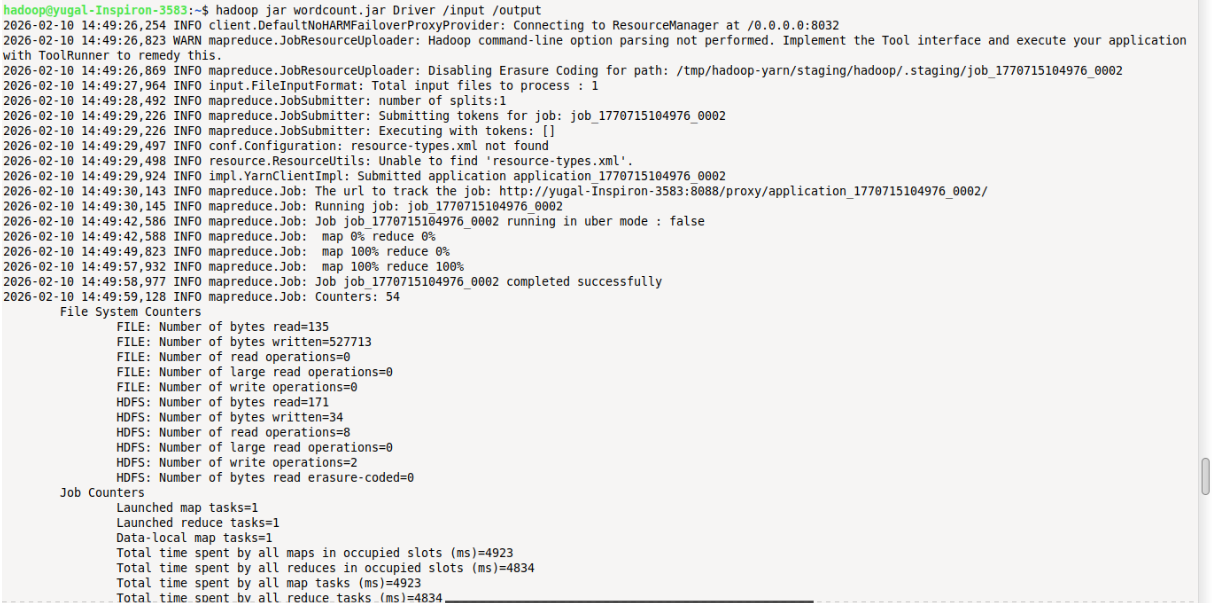


```
hadoop@yugal-Inspiron-3583:~$ javac -classpath `hadoop classpath` -d . MapperClass.java ReducerClass.java Driver.java
hadoop@yugal-Inspiron-3583:~$ jar -cvf wordcount.jar *.class
added manifest
adding: Driver.class(in = 1342) (out= 741)(deflated 44%)
adding: MapperClass.class(in = 1680) (out= 728)(deflated 56%)
adding: ReducerClass.class(in = 1591) (out= 661)(deflated 58%)
```

Figure 36: Compilation and JAR creation

3.5 Step 5: Execute MapReduce Job

```
hadoop jar wordcount.jar Driver /input /output
```

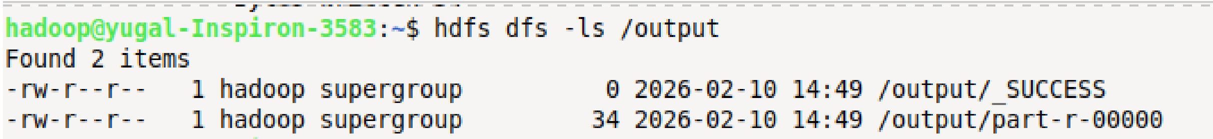


```
hadoop@yugal-Inspiron-3583:~$ hadoop jar wordcount.jar Driver /input /output
2026-02-10 14:49:26,254 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2026-02-10 14:49:26,823 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
with ToolRunner to remedy this.
2026-02-10 14:49:26,869 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1770715104976_0002
2026-02-10 14:49:27,964 INFO input.FileInputFormat: Total input files to process : 1
2026-02-10 14:49:28,492 INFO mapreduce.JobSubmitter: number of splits:1
2026-02-10 14:49:29,226 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1770715104976_0002
2026-02-10 14:49:29,226 INFO mapreduce.JobSubmitter: Executing with tokens: []
2026-02-10 14:49:29,497 INFO conf.Configuration: resource-types.xml not found
2026-02-10 14:49:29,498 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2026-02-10 14:49:29,924 INFO impl.YarnClientImpl: Submitted application application_1770715104976_0002
2026-02-10 14:49:30,143 INFO mapreduce.Job: The url to track the job: http://yugal-Inspiron-3583:8088/proxy/application_1770715104976_0002/
2026-02-10 14:49:30,145 INFO mapreduce.Job: Running job: job_1770715104976_0002
2026-02-10 14:49:42,586 INFO mapreduce.Job: Job job_1770715104976_0002 running in uber mode : false
2026-02-10 14:49:42,588 INFO mapreduce.Job:  map 0% reduce 0%
2026-02-10 14:49:49,823 INFO mapreduce.Job:  map 100% reduce 0%
2026-02-10 14:49:57,932 INFO mapreduce.Job:  map 100% reduce 100%
2026-02-10 14:49:58,977 INFO mapreduce.Job: Job job_1770715104976_0002 completed successfully
2026-02-10 14:49:59,128 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=135
                FILE: Number of bytes written=527713
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=171
                HDFS: Number of bytes written=34
                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)=4923
                Total time spent by all reduces in occupied slots (ms)=4834
                Total time spent by all map tasks (ms)=4923
                Total time spent by all reduce tasks (ms)=4834
```

Figure 37: Executing MapReduce job

3.6 Step 6: Check Output Directory in HDFS

```
hdfs dfs -ls /output
```

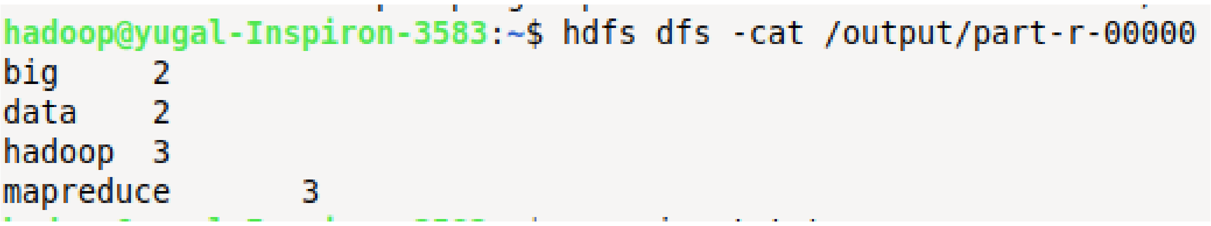


```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -ls /output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2026-02-10 14:49 /output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         34 2026-02-10 14:49 /output/part-r-00000
```

Figure 38: Output directory created in HDFS

3.7 Step 7: Display Result File

```
hdfs dfs -cat /mapreduce_output/part-r-00000
```

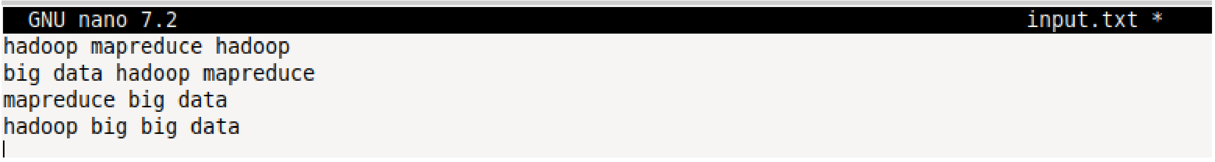


```
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -cat /output/part-r-00000
big     2
data    2
hadoop  3
mapreduce       3
```

Figure 39: Displaying MapReduce output

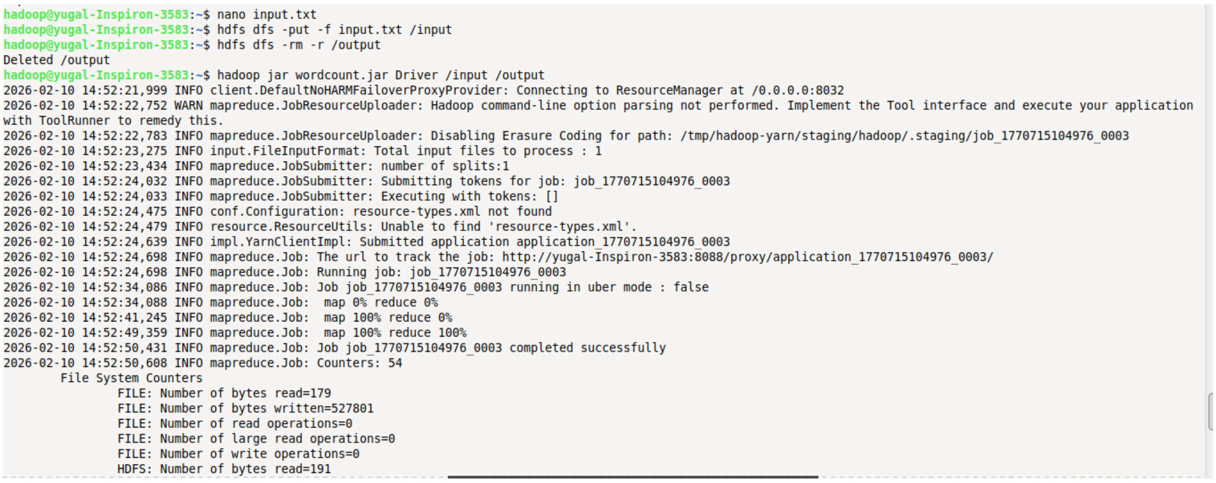
3.8 Step 8: Modify Input and Re-run Job

```
nano input.text
hdfs dfs -put input.txt /input
hdfs dfs -rm -r /output
hadoop jar wordcount.jar Driver /input /output
hdfs dfs -cat /output/part-r-00000
```



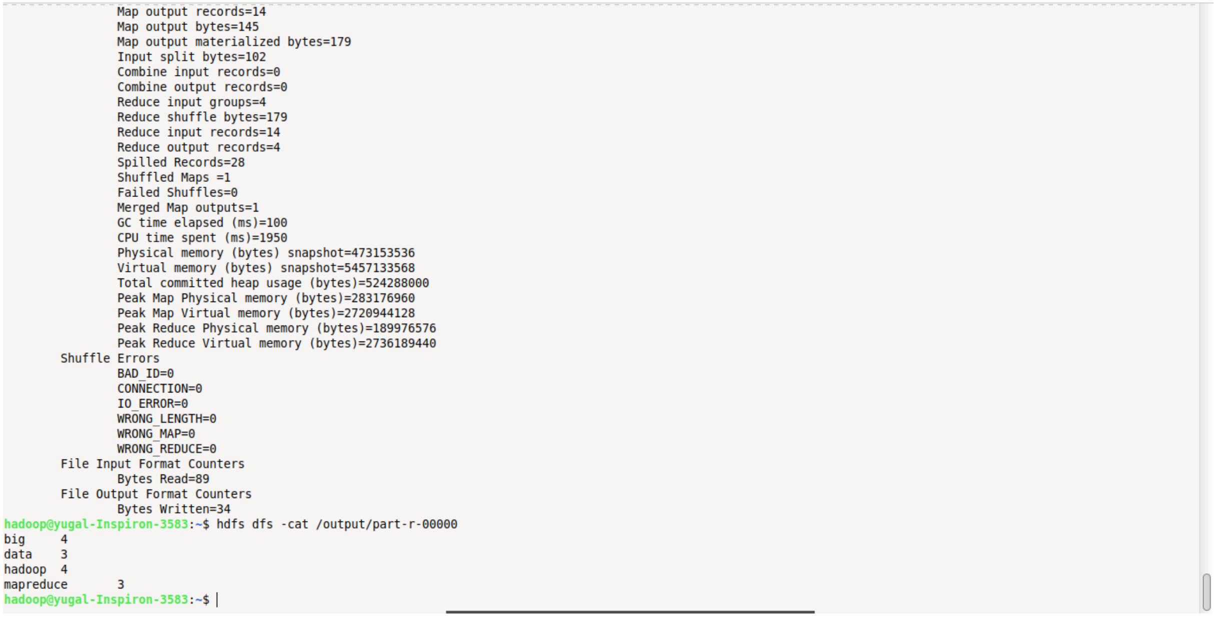
```
  GNU nano 7.2                                                       input.txt *
hadoop mapreduce hadoop
big data hadoop mapreduce
mapreduce big data
hadoop big big data
```

Figure 40: Updating input text file



```
hadoop@yugal-Inspiron-3583:~$ nano input.txt
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -put -f input.txt /input
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -rm -r /output
Deleted /output
hadoop@yugal-Inspiron-3583:~$ hadoop jar wordcount.jar Driver /input /output
2026-02-10 14:52:21,999 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2026-02-10 14:52:22,752 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
with ToolRunner to remedy this.
2026-02-10 14:52:22,783 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1770715104976_0003
2026-02-10 14:52:23,275 INFO input.FileInputFormat: Total input files to process : 1
2026-02-10 14:52:23,434 INFO mapreduce.JobSubmitter: number of splits:1
2026-02-10 14:52:24,032 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1770715104976_0003
2026-02-10 14:52:24,033 INFO mapreduce.JobSubmitter: Executing with tokens: []
2026-02-10 14:52:24,475 INFO conf.Configuration: resource-types.xml not found
2026-02-10 14:52:24,479 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2026-02-10 14:52:24,639 INFO impl.YarnClientImpl: Submitted application application_1770715104976_0003
2026-02-10 14:52:24,698 INFO mapreduce.Job: The url to track the job: http://yugal-Inspiron-3583:8088/proxy/application_1770715104976_0003/
2026-02-10 14:52:24,698 INFO mapreduce.Job: Running job: job_1770715104976_0003
2026-02-10 14:52:34,086 INFO mapreduce.Job: Job job_1770715104976_0003 running in uber mode : false
2026-02-10 14:52:34,088 INFO mapreduce.Job:  map 0% reduce 0%
2026-02-10 14:52:41,245 INFO mapreduce.Job:  map 100% reduce 0%
2026-02-10 14:52:49,359 INFO mapreduce.Job:  map 100% reduce 100%
2026-02-10 14:52:50,431 INFO mapreduce.Job: Job job_1770715104976_0003 completed successfully
2026-02-10 14:52:50,608 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=179
                FILE: Number of bytes written=527801
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=191
```

Figure 41: Re-running MapReduce with modified dataset



```
                Map output records=14
                Map output bytes=145
                Map output materialized bytes=179
                Input split bytes=102
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=179
                Reduce input records=14
                Reduce output records=4
                Spilled Records=28
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=100
                CPU time spent (ms)=1950
                Physical memory (bytes) snapshot=473153536
                Virtual memory (bytes) snapshot=5457133568
                Total committed heap usage (bytes)=524288000
                Peak Map Physical memory (bytes)=283176960
                Peak Map Virtual memory (bytes)=2720944128
                Peak Reduce Physical memory (bytes)=189976576
                Peak Reduce Virtual memory (bytes)=2736189440
        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=89
        File Output Format Counters
                Bytes Written=34
hadoop@yugal-Inspiron-3583:~$ hdfs dfs -cat /output/part-r-00000
big     4
data    3
hadoop  4
mapreduce       3
hadoop@yugal-Inspiron-3583:~$ 
```

Figure 42: New output after re-running MapReduce job

3.9 Result

The MapReduce job was successfully executed on Hadoop. Input data stored in HDFS was processed using Mapper and Reducer classes, and the output was generated in the HDFS output directory. Re-running the job with modified input demonstrated changes in output based on data size and content.

3.10 Conclusion

This assignment demonstrated the practical implementation of the MapReduce programming model in Hadoop. It showed how distributed computing enables scalable data processing by dividing tasks into mapping and reducing phases, improving efficiency for large datasets.