Vaibhav Kumar **Roll No: 19**

Assignment 1 Big Data

Date :08-08-2022

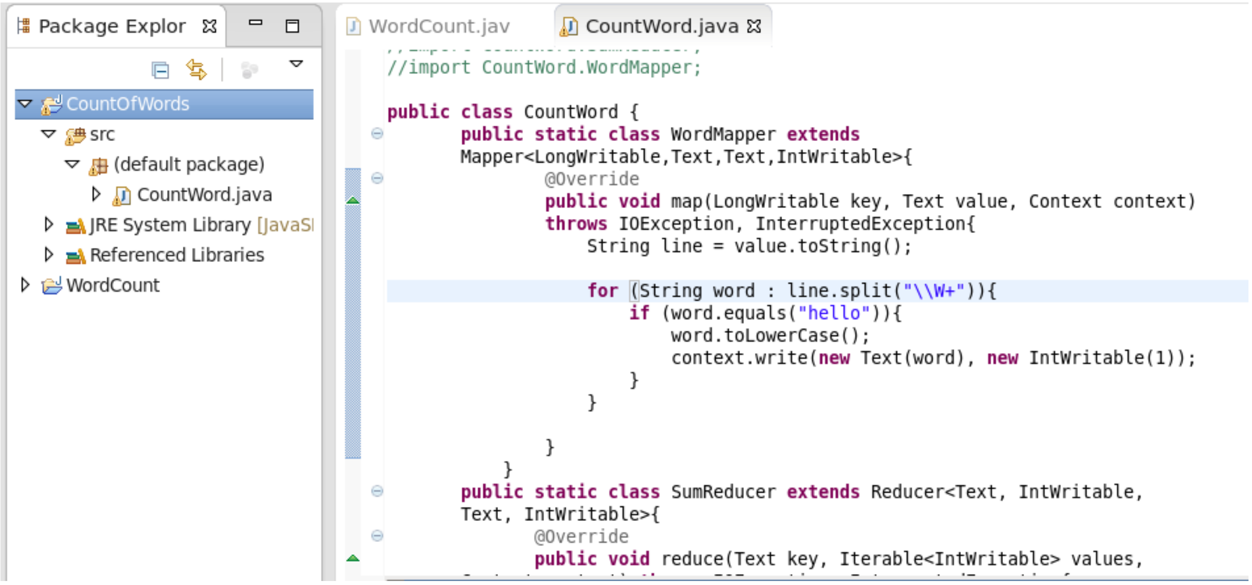
Problem Statement:

Write a program to find occurrence of any specific word

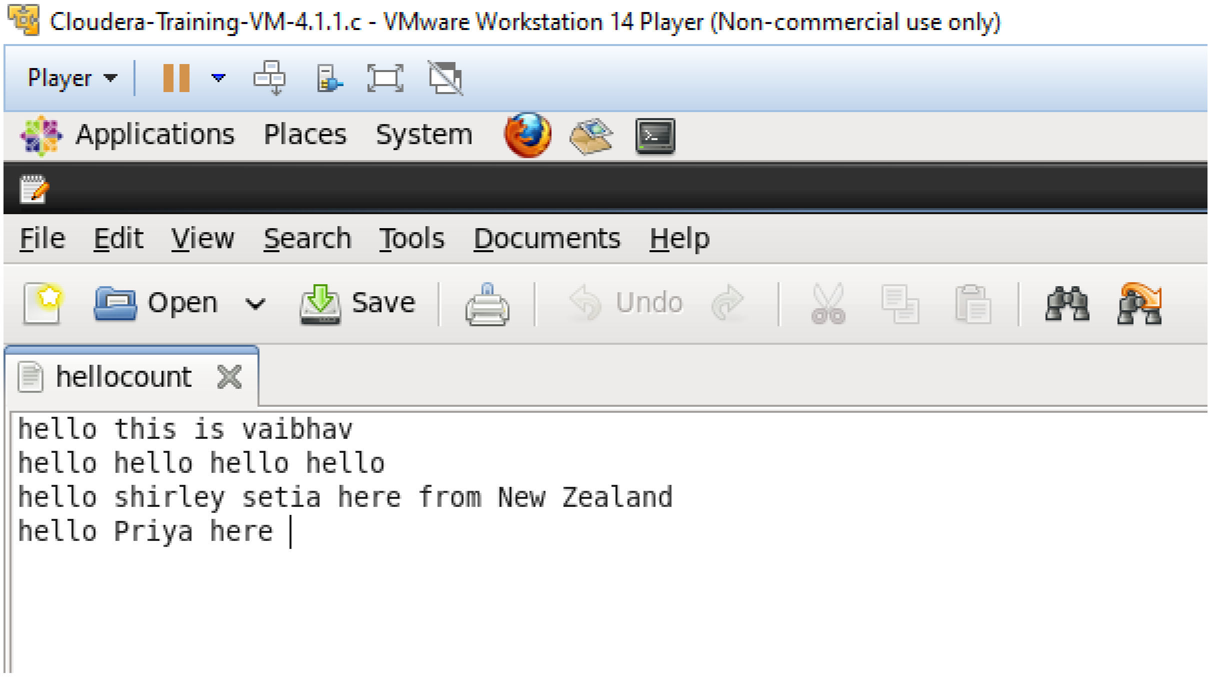
This is the File created by me

```
import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;//type casted to Int
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper.Context;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Mapper;//mapper class
import org.apache.hadoop.mapreduce.Reducer;

//import CountWord.SumReducer;
//import CountWord.WordMapper;

public class CountWord {
        public static class WordMapper extends
        Mapper<LongWritable,Text,Text,IntWritable>{
            @Override
            public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException{
              String line = value.toString();

              for (String word : line.split("\\W+")){
                if (word.equals("hello")){
                    word.toLowerCase();
                  context.write(new Text(word), new IntWritable(1));
                }
              }

            }
          }
        public static class SumReducer extends Reducer<Text, IntWritable,
        Text, IntWritable>{
```

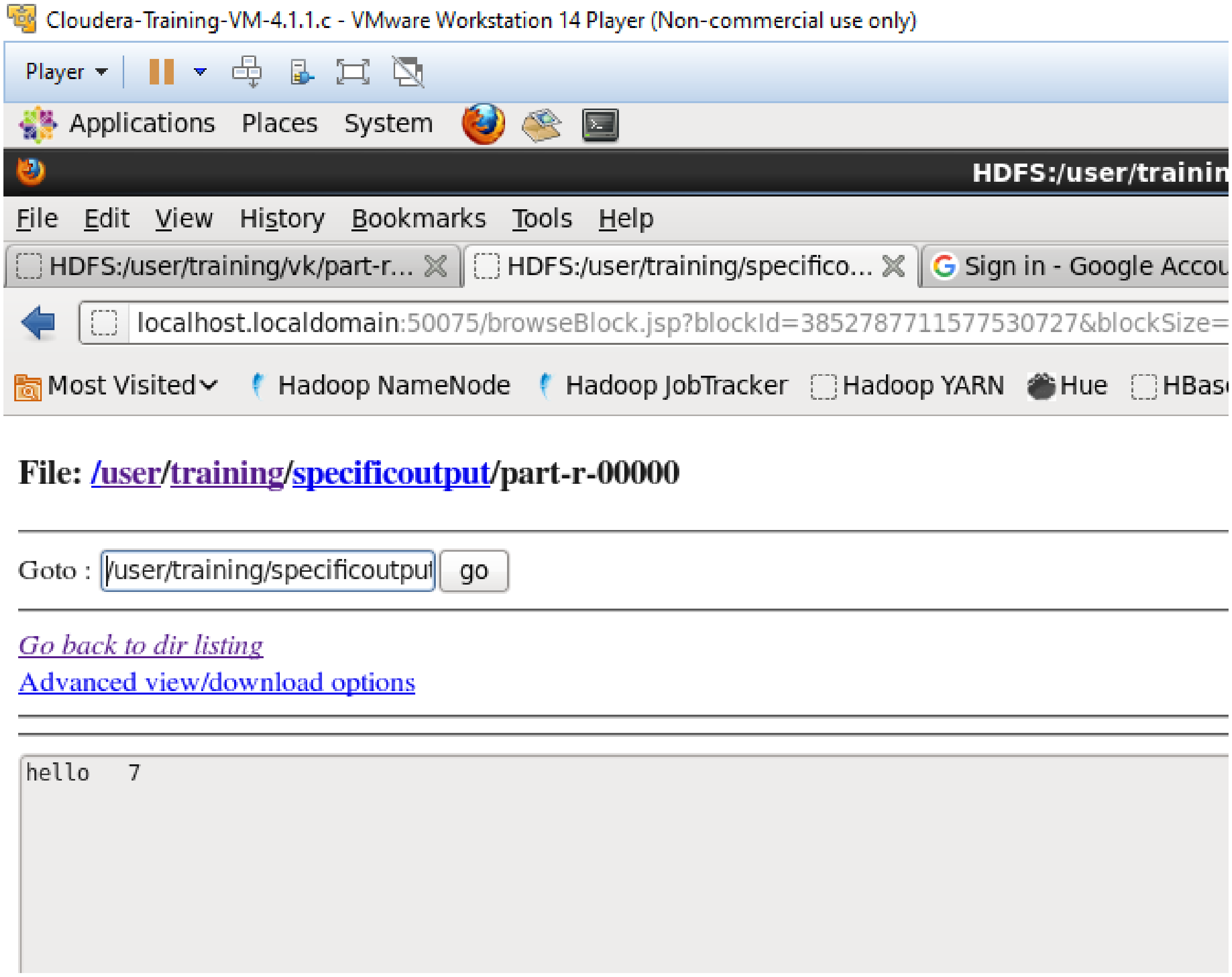
```
        @Override
        public void reduce(Text key, Iterable<IntWritable> values,
  Context context) throws IOException, InterruptedException{
           int wordCount = 0;

           for (IntWritable value: values){
              wordCount += value.get();

              }
           context.write(key, new IntWritable(wordCount));
           }

        }
  public static void main(String[] args) throws Exception{
      if(args.length != 2){
         System.out.printf("Usage: CountWord <input dir> <output dir>\n");
         System.exit(-1);

      }
      Job job = new Job();
      job.setJarByClass(CountWord.class);
      job.setJobName("CountWord");

      FileInputFormat.setInputPaths(job, new Path(args[0]));
      FileOutputFormat.setOutputPath(job, new Path(args[1]));

      job.setMapperClass(WordMapper.class);
      job.setReducerClass(SumReducer.class);

      job.setOutputKeyClass(Text.class);
      job.setOutputValueClass(IntWritable.class);

      boolean success = job.waitForCompletion(true);
      System.exit(success ? 0 : 1);

   }
}
```



```java
//import CountWord.WordMapper;

public class CountWord {
        public static class WordMapper extends
        Mapper<LongWritable,Text,Text,IntWritable>{
                @Override
                public void map(LongWritable key, Text value, Context context)
                throws IOException, InterruptedException{
                    String line = value.toString();

                    for (String word : line.split("\\W+")){
                        if (word.equals("hello")){
                            word.toLowerCase();
                            context.write(new Text(word), new IntWritable(1));
                        }
                    }

                }
            }
        public static class SumReducer extends Reducer<Text, IntWritable,
        Text, IntWritable>{
               @Override
               public void reduce(Text key, Iterable<IntWritable> values,
```



```
hello this is vaibhav
hello hello hello hello
hello shirley setia here from New Zealand
hello Priya here
```



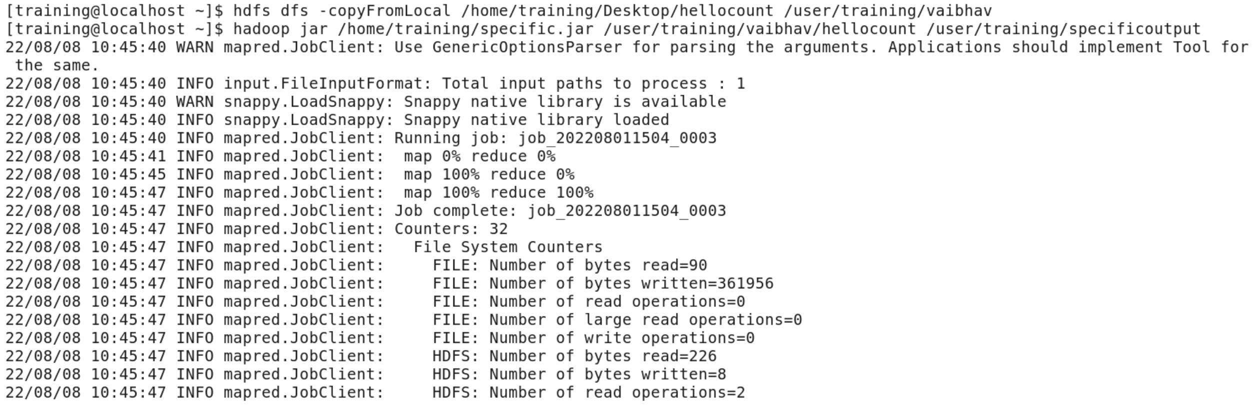
Cloudera-Training-VM-4.1.1.c - VMware Workstation 14 Player (Non-commercial use only)
Player
Applications Places System
HDFS:/user/trainin
File Edit View History Bookmarks Tools Help
HDFS:/user/training/vk/part-r...
HDFS:/user/training/specifico...
Sign in - Google Accou
localhost.localdomain:50075/browseBlock.jsp?blockId=3852787711577530727&blockSize=
Most Visited
Hadoop NameNode
Hadoop JobTracker
Hadoop YARN
Hue
HBas

File: /user/training/specificoutput/part-r-00000

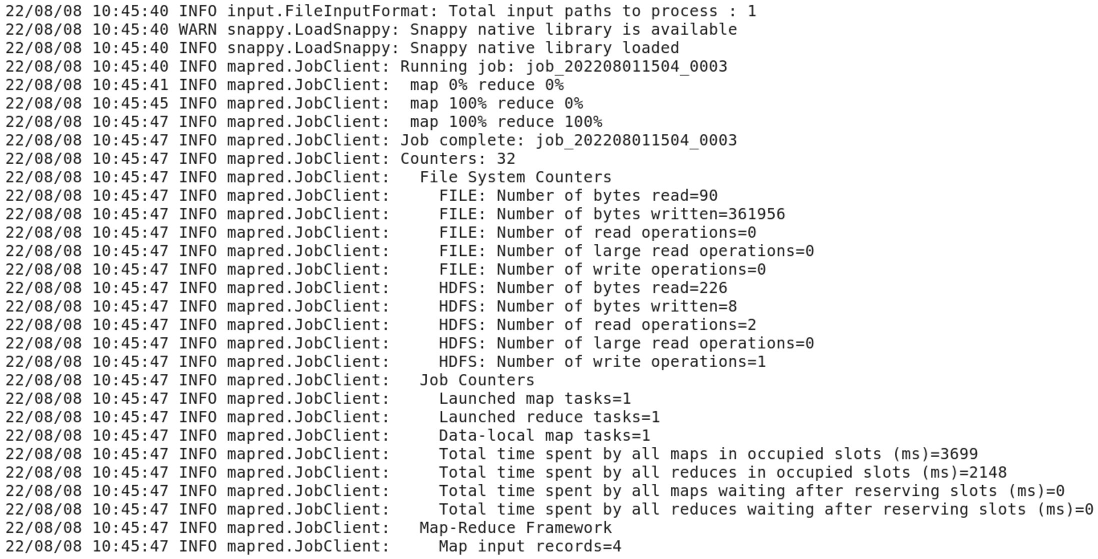
Goto : /user/training/specificoutput go

Go back to dir listing
Advanced view/download options

```
hello	7
```



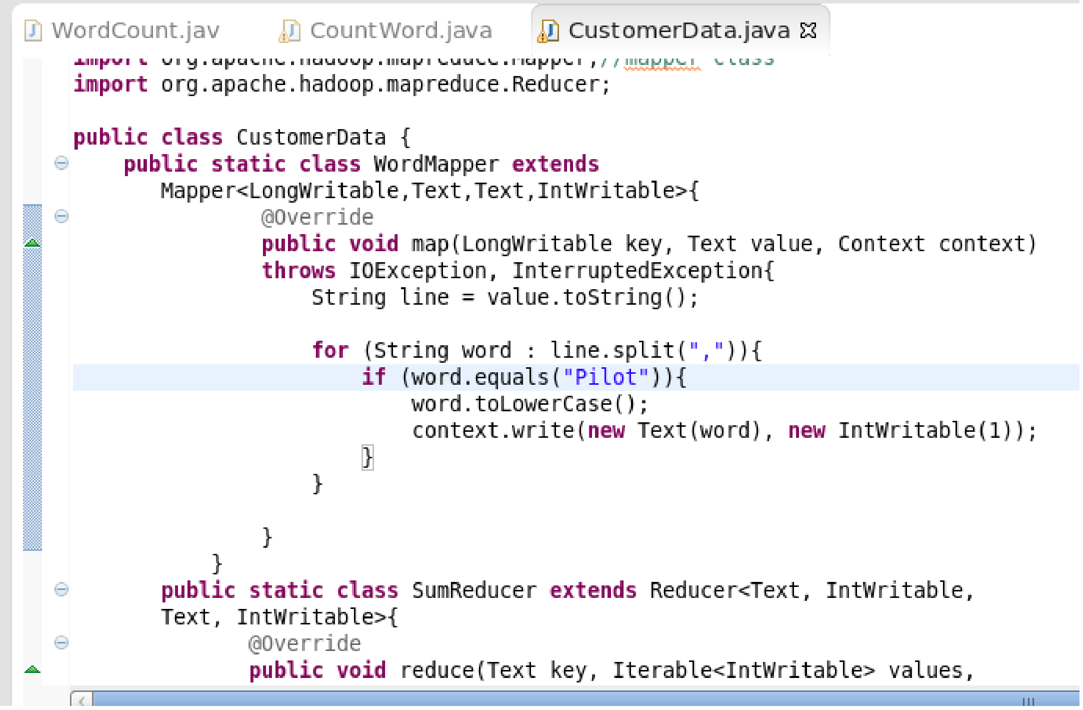
```
[training@localhost ~]$ hdfs dfs -copyFromLocal /home/training/Desktop/hellocount /user/training/vaibhav
[training@localhost ~]$ hadoop jar /home/training/specific.jar /user/training/vaibhav/hellocount /user/training/specificoutput
22/08/08 10:45:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for
 the same.
22/08/08 10:45:40 INFO input.FileInputFormat: Total input paths to process : 1
22/08/08 10:45:40 WARN snappy.LoadSnappy: Snappy native library is available
22/08/08 10:45:40 INFO snappy.LoadSnappy: Snappy native library loaded
22/08/08 10:45:40 INFO mapred.JobClient: Running job: job_202208011504_0003
22/08/08 10:45:41 INFO mapred.JobClient:  map 0% reduce 0%
22/08/08 10:45:45 INFO mapred.JobClient:  map 100% reduce 0%
22/08/08 10:45:47 INFO mapred.JobClient:  map 100% reduce 100%
22/08/08 10:45:47 INFO mapred.JobClient: Job complete: job_202208011504_0003
22/08/08 10:45:47 INFO mapred.JobClient: Counters: 32
22/08/08 10:45:47 INFO mapred.JobClient:   File System Counters
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of bytes read=90
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of bytes written=361956
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of read operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of large read operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of write operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of bytes read=226
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of bytes written=8
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of read operations=2
```



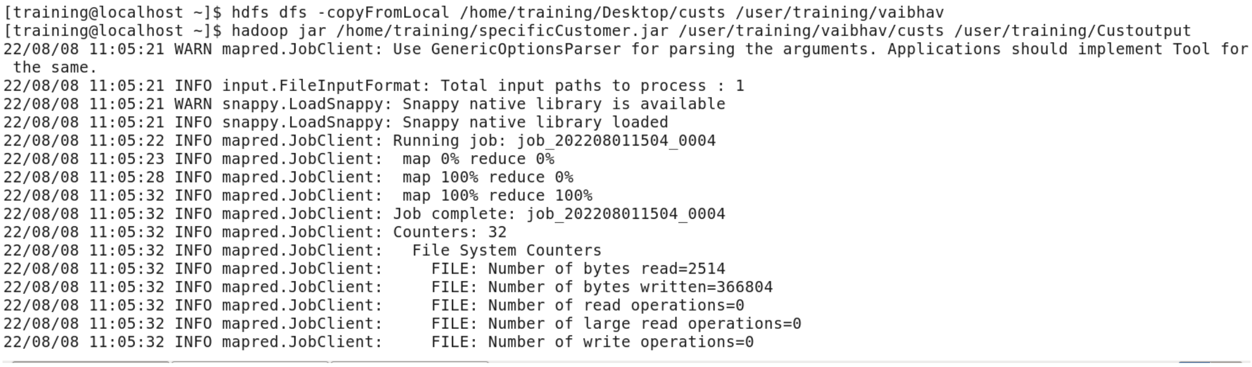
```
22/08/08 10:45:40 INFO input.FileInputFormat: Total input paths to process : 1
22/08/08 10:45:40 WARN snappy.LoadSnappy: Snappy native library is available
22/08/08 10:45:40 INFO snappy.LoadSnappy: Snappy native library loaded
22/08/08 10:45:40 INFO mapred.JobClient: Running job: job_202208011504_0003
22/08/08 10:45:41 INFO mapred.JobClient:  map 0% reduce 0%
22/08/08 10:45:45 INFO mapred.JobClient:  map 100% reduce 0%
22/08/08 10:45:47 INFO mapred.JobClient:  map 100% reduce 100%
22/08/08 10:45:47 INFO mapred.JobClient: Job complete: job_202208011504_0003
22/08/08 10:45:47 INFO mapred.JobClient: Counters: 32
22/08/08 10:45:47 INFO mapred.JobClient:   File System Counters
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of bytes read=90
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of bytes written=361956
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of read operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of large read operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     FILE: Number of write operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of bytes read=226
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of bytes written=8
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of read operations=2
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of large read operations=0
22/08/08 10:45:47 INFO mapred.JobClient:     HDFS: Number of write operations=1
22/08/08 10:45:47 INFO mapred.JobClient:   Job Counters
22/08/08 10:45:47 INFO mapred.JobClient:     Launched map tasks=1
22/08/08 10:45:47 INFO mapred.JobClient:     Launched reduce tasks=1
22/08/08 10:45:47 INFO mapred.JobClient:     Data-local map tasks=1
22/08/08 10:45:47 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=3699
22/08/08 10:45:47 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=2148
22/08/08 10:45:47 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
22/08/08 10:45:47 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
22/08/08 10:45:47 INFO mapred.JobClient:   Map-Reduce Framework
22/08/08 10:45:47 INFO mapred.JobClient:     Map input records=4
```

Customer Data

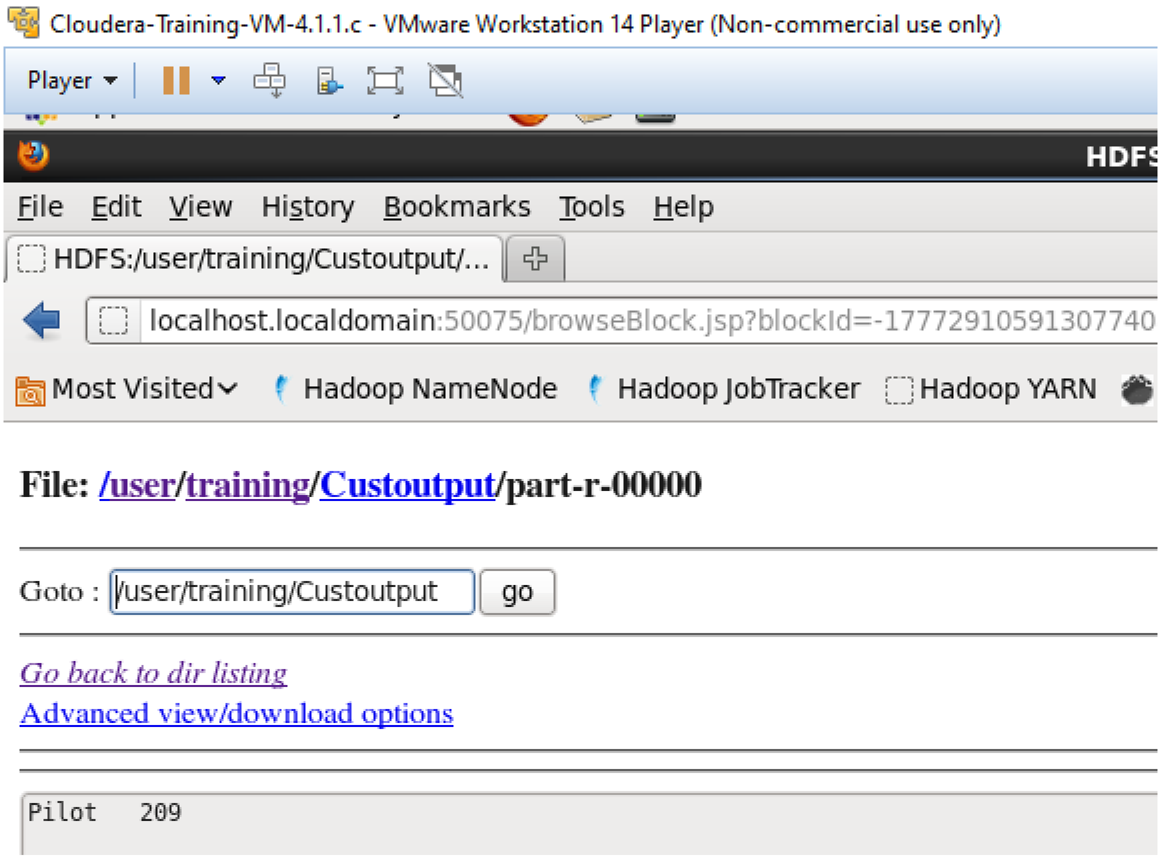
```
public class CustomerData {
        public static class WordMapper extends
          Mapper<LongWritable,Text,Text,IntWritable>{
                @Override
                public void map(LongWritable key, Text value, Context context)
                throws IOException, InterruptedException{
                   String line = value.toString();

                   for (String word : line.split(",")){
                      if (word.equals("Pilot")){
                           word.toLowerCase();
                        context.write(new Text(word), new IntWritable(1));
                      }
                   }

                }
             }
          public static class SumReducer extends Reducer<Text, IntWritable,
          Text, IntWritable>{
               @Override
               public void reduce(Text key, Iterable<IntWritable> values,
          Context context) throws IOException, InterruptedException{
                  int wordCount = 0;

                  for (IntWritable value: values){
                     wordCount += value.get();

                     }
                  context.write(key, new IntWritable(wordCount));
                  }

               }
          public static void main(String[] args) throws Exception{
              if(args.length != 2){
                 System.out.printf("Usage: CountWord <input dir> <output dir>\n");
                 System.exit(-1);

              }
              Job job = new Job();
              job.setJarByClass(CustomerData.class);
              job.setJobName("CustomerData");
              FileInputFormat.setInputPaths(job, new Path(args[0]));
              FileOutputFormat.setOutputPath(job, new Path(args[1]));

              job.setMapperClass(WordMapper.class);
              job.setReducerClass(SumReducer.class);

              job.setOutputKeyClass(Text.class);
              job.setOutputValueClass(IntWritable.class);

              boolean success = job.waitForCompletion(true);
              System.exit(success ? 0 : 1);

          }
        }
```



```
import org.apache.hadoop.mapreduce.Mapper;//mapper class
import org.apache.hadoop.mapreduce.Reducer;

public class CustomerData {
    public static class WordMapper extends
       Mapper<LongWritable,Text,Text,IntWritable>{
               @Override
               public void map(LongWritable key, Text value, Context context)
               throws IOException, InterruptedException{
                   String line = value.toString();

                   for (String word : line.split(",")){
                       if (word.equals("Pilot")){
                           word.toLowerCase();
                           context.write(new Text(word), new IntWritable(1));
                       }
                   }

               }
           }
       public static class SumReducer extends Reducer<Text, IntWritable,
       Text, IntWritable>{
              @Override
              public void reduce(Text key, Iterable<IntWritable> values,
```



```
[training@localhost ~]$ hdfs dfs -copyFromLocal /home/training/Desktop/custs /user/training/vaibhav
[training@localhost ~]$ hadoop jar /home/training/specificCustomer.jar /user/training/vaibhav/custs /user/training/Custoutput
22/08/08 11:05:21 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for
 the same.
22/08/08 11:05:21 INFO input.FileInputFormat: Total input paths to process : 1
22/08/08 11:05:21 WARN snappy.LoadSnappy: Snappy native library is available
22/08/08 11:05:21 INFO snappy.LoadSnappy: Snappy native library loaded
22/08/08 11:05:22 INFO mapred.JobClient: Running job: job_202208011504_0004
22/08/08 11:05:23 INFO mapred.JobClient:  map 0% reduce 0%
22/08/08 11:05:28 INFO mapred.JobClient:  map 100% reduce 0%
22/08/08 11:05:32 INFO mapred.JobClient:  map 100% reduce 100%
22/08/08 11:05:32 INFO mapred.JobClient: Job complete: job_202208011504_0004
22/08/08 11:05:32 INFO mapred.JobClient: Counters: 32
22/08/08 11:05:32 INFO mapred.JobClient:   File System Counters
22/08/08 11:05:32 INFO mapred.JobClient:     FILE: Number of bytes read=2514
22/08/08 11:05:32 INFO mapred.JobClient:     FILE: Number of bytes written=366804
22/08/08 11:05:32 INFO mapred.JobClient:     FILE: Number of read operations=0
22/08/08 11:05:32 INFO mapred.JobClient:     FILE: Number of large read operations=0
22/08/08 11:05:32 INFO mapred.JobClient:     FILE: Number of write operations=0
```



Cloudera-Training-VM-4.1.1.c - VMware Workstation 14 Player (Non-commercial use only)
Player
HDFS
File
Edit
View
History
Bookmarks
Tools
Help
HDFS:/user/training/Custoutput/...
localhost.localdomain:50075/browseBlock.jsp?blockId=-177729105913077740
Most Visited
Hadoop NameNode
Hadoop JobTracker
Hadoop YARN

File: /user/training/Custoutput/part-r-00000

Goto : /user/training/Custoutput go

Go back to dir listing
Advanced view/download options

```
Pilot	209
```