Case Study 1

Problem Statement:

a. **What are the movie titles that the user has rated?**
b. **How many times a movie has been rated by the user?**
c. **In question 2 above, what is the average rating given for a movie?**

Codes

Movie Mapper code

```
import java.io.IOException;
import java.io.IOException;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class CaseStudyIUseCasesMoviesMapper extends
                        Mapper<LongWritable, Text, Text, Text> {


    public void map(LongWritable key, Text value, Context context)
                        throws IOException, InterruptedException {

            try {
        if (key.get() == 0 && value.toString().contains("movieId")){
            return;
        } else {
            String record = value.toString();
                    String[] parts = record.split(",");
                    context.write(new Text(parts[0]), new Text("movies\t" + parts[1]));
        }
      } catch (Exception e) {
        e.printStackTrace();
      }
    }
}
```

Ratings Mapper Code

```
    import java.io.IOException;

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

    public class CaseStudyIUseCasesRatingsMapper extends
                            Mapper<LongWritable, Text, Text, Text> {

        public void map(LongWritable key, Text value, Context context)
                                throws IOException, InterruptedException {

                            try {
                   if (key.get() == 0 && value.toString().contains("userId")){
                     return;
                   } else {
                                String record = value.toString();
                                String[] parts = record.split(",");
                                context.write(new Text(parts[1]), new Text("ratings\t" +
    parts[2]));
                   }
                } catch (Exception e) {
                   e.printStackTrace();
                }

        }
    }
```

Reducer code

```
import java.io.IOException;

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


public class CaseStudyIUseCasesReducer extends
                            Reducer<Text, Text, Text, Text> {

                public void reduce(Text key, Iterable<Text> values, Context context)
                                throws IOException, InterruptedException {
                            String titles = "";
                            double total = 0.0;
                            int count = 0;
                            System.out.println("Text Key    =>"+key.toString());
                            for (Text t : values) {
```

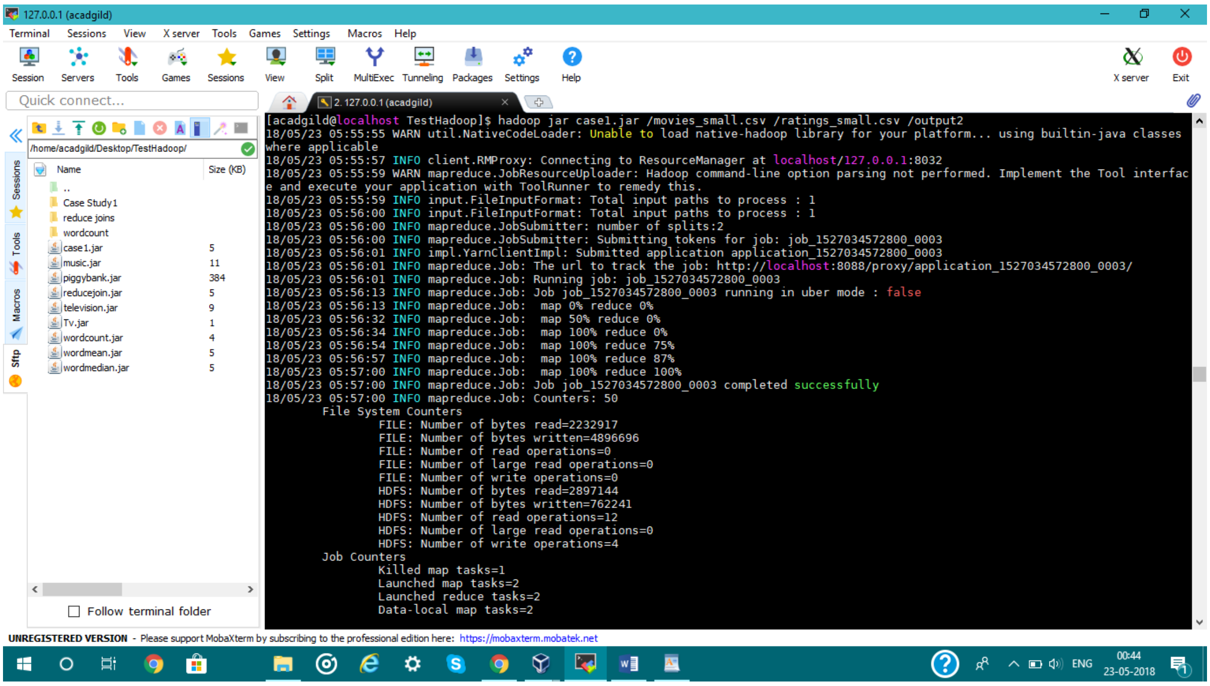
```
                    String parts[] = t.toString().split("\t");
                    System.out.println("Text values =>"+t.toString());
                    if (parts[0].equals("ratings")) {
                        count++;
                        String rating = parts[1].trim();
                        System.out.println("Rating is =>"+rating);
                        total += Double.parseDouble(rating);
                    } else if (parts[0].equals("movies")) {
                        titles = parts[1];
                    }
                }

                double average = total / count;            //for calculating average
                String str = String.format("Number of times rated = %d and average
rated time = %f", count, average);
             //String str = String.format("%d", count);

                context.write(new Text(titles), new Text(str));
            }
        }
```

Driver Code

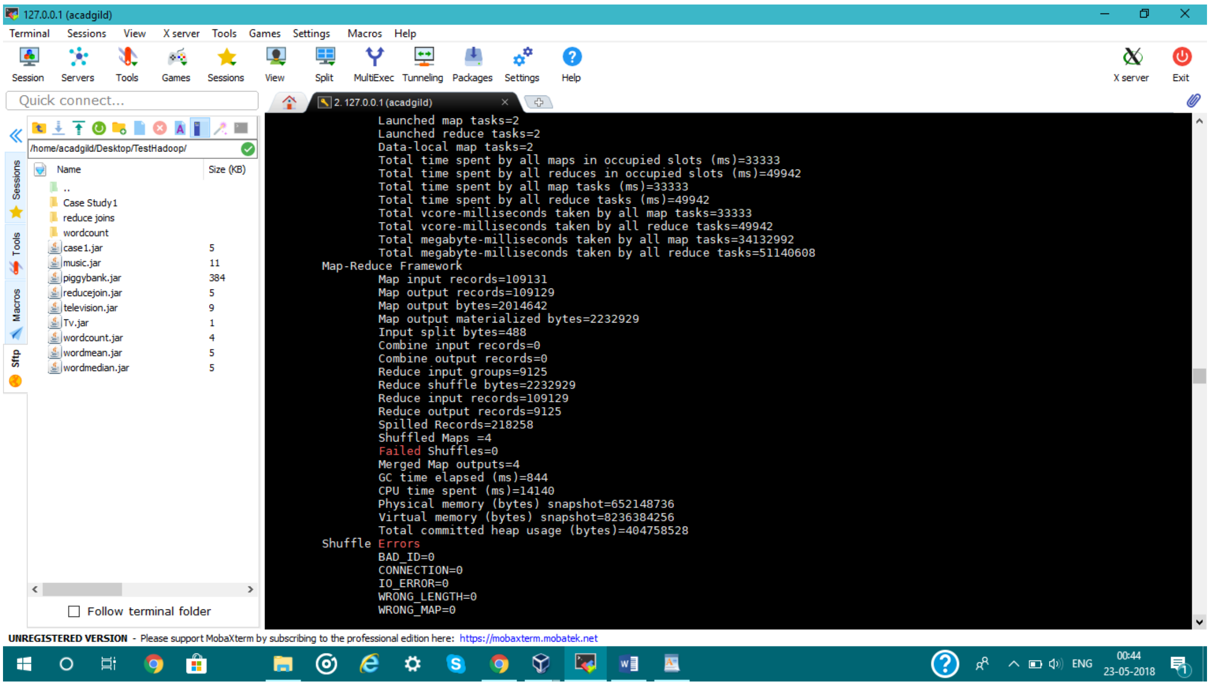
```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class CaseStudyIUseCasesDriver {

    @SuppressWarnings("deprecation")
    public static void main(String[] args) throws Exception {
  if (args.length != 3) {
    System.err.println("Usage: CaseStudyIUseCase2Driver <input path1> <input path2>
<output path>");
    System.exit(-1);
  }

    //Job Related Configurations
    Configuration conf = new Configuration();
```

```
        Job job = new Job(conf, "CaseStudyIUseCase2Driver");
        job.setJarByClass(CaseStudyIUseCasesDriver.class);

        job.setNumReduceTasks(2);

        //Since there are multiple input, there is a slightly different way of specifying input
path, input format and mapper
        MultipleInputs.addInputPath(job, new Path(args[0]),TextInputFormat.class,
CaseStudyIUseCasesMoviesMapper.class);
        MultipleInputs.addInputPath(job, new Path(args[1]),TextInputFormat.class,
CaseStudyIUseCasesRatingsMapper.class);

        //Set the reducer
        job.setReducerClass(CaseStudyIUseCasesReducer.class);

   //set the out path
        Path outputPath = new Path(args[2]);
        FileOutputFormat.setOutputPath(job, outputPath);
        outputPath.getFileSystem(conf).delete(outputPath, true);

   //set up the output key and value classes
   job.setOutputKeyClass(Text.class);
   job.setOutputValueClass(Text.class);

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

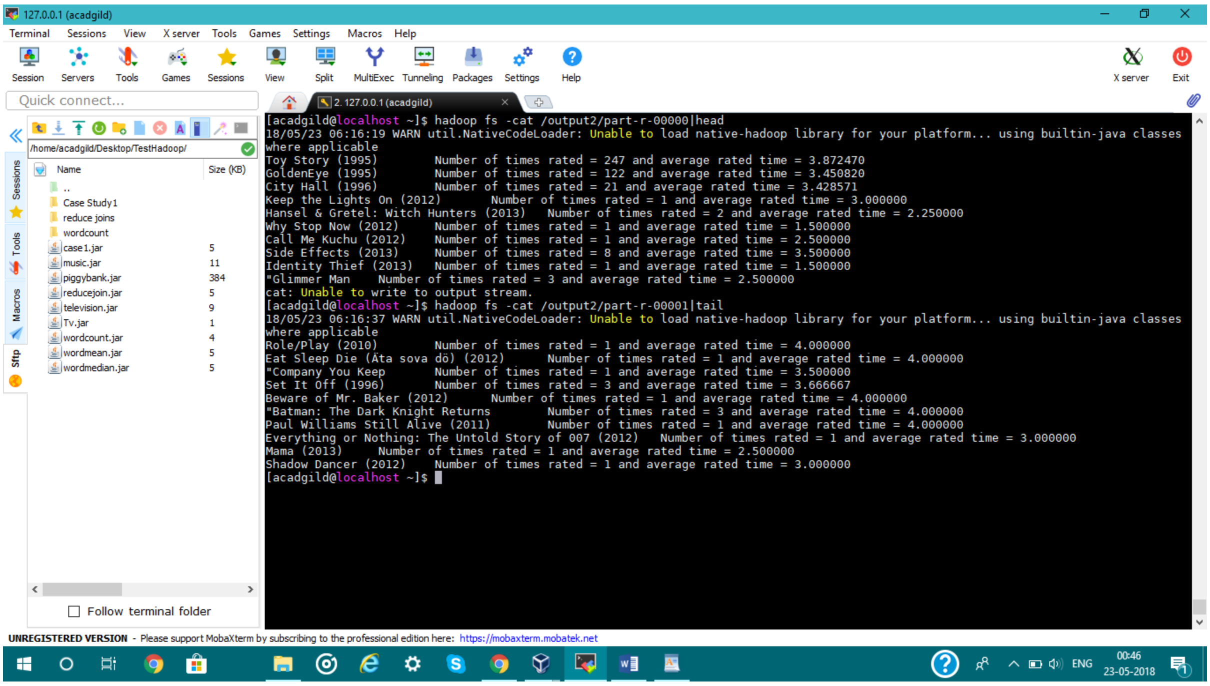
Screenshots


127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. 127.0.0.1 (acadgild)
/home/acadgild/Desktop/TestHadoop/
Name Size (KB)
..
Case Study1
reduce joins
wordcount
case1.jar 5
music.jar 11
piggybank.jar 384
reducejoin.jar 5
television.jar 9
Tv.jar 1
wordcount.jar 4
wordmean.jar 5
wordmedian.jar 5
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
ENG
00:44
23-05-2018

```
[acadgild@localhost TestHadoop]$ hadoop jar case1.jar /movies_small.csv /ratings_small.csv /output2
18/05/23 05:55:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
18/05/23 05:55:57 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/05/23 05:55:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
18/05/23 05:55:59 INFO input.FileInputFormat: Total input paths to process : 1
18/05/23 05:56:00 INFO input.FileInputFormat: Total input paths to process : 1
18/05/23 05:56:00 INFO mapreduce.JobSubmitter: number of splits:2
18/05/23 05:56:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1527034572800_0003
18/05/23 05:56:01 INFO impl.YarnClientImpl: Submitted application application_1527034572800_0003
18/05/23 05:56:01 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1527034572800_0003/
18/05/23 05:56:01 INFO mapreduce.Job: Running job: job_1527034572800_0003
18/05/23 05:56:13 INFO mapreduce.Job: Job job_1527034572800_0003 running in uber mode : false
18/05/23 05:56:13 INFO mapreduce.Job:  map 0% reduce 0%
18/05/23 05:56:32 INFO mapreduce.Job:  map 50% reduce 0%
18/05/23 05:56:34 INFO mapreduce.Job:  map 100% reduce 0%
18/05/23 05:56:54 INFO mapreduce.Job:  map 100% reduce 75%
18/05/23 05:56:57 INFO mapreduce.Job:  map 100% reduce 87%
18/05/23 05:57:00 INFO mapreduce.Job:  map 100% reduce 100%
18/05/23 05:57:00 INFO mapreduce.Job: Job job_1527034572800_0003 completed successfully
18/05/23 05:57:00 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=2232917
                FILE: Number of bytes written=4896696
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2897144
                HDFS: Number of bytes written=762241
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=2
                Data-local map tasks=2
```


127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. 127.0.0.1 (acadgild)
/home/acadgild/Desktop/TestHadoop/
Name Size (KB)
..
Case Study1
reduce joins
wordcount
case1.jar 5
music.jar 11
piggybank.jar 384
reducejoin.jar 5
television.jar 9
Tv.jar 1
wordcount.jar 4
wordmean.jar 5
wordmedian.jar 5
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
ENG
00:44
23-05-2018

```
                Launched map tasks=2
                Launched reduce tasks=2
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=33333
                Total time spent by all reduces in occupied slots (ms)=49942
                Total time spent by all map tasks (ms)=33333
                Total time spent by all reduce tasks (ms)=49942
                Total vcore-milliseconds taken by all map tasks=33333
                Total vcore-milliseconds taken by all reduce tasks=49942
                Total megabyte-milliseconds taken by all map tasks=34132992
                Total megabyte-milliseconds taken by all reduce tasks=51140608
        Map-Reduce Framework
                Map input records=109131
                Map output records=109129
                Map output bytes=2014642
                Map output materialized bytes=2232929
                Input split bytes=488
                Combine input records=0
                Combine output records=0
                Reduce input groups=9125
                Reduce shuffle bytes=2232929
                Reduce input records=109129
                Reduce output records=9125
                Spilled Records=218258
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=844
                CPU time spent (ms)=14140
                Physical memory (bytes) snapshot=652148736
                Virtual memory (bytes) snapshot=8236384256
                Total committed heap usage (bytes)=404758528
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```


127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. 127.0.0.1 (acadgild)
Sessions Tools Macros Sftp
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
ENG
00:46
23-05-2018

/home/acadgild/Desktop/TestHadoop/

Name	Size (KB)
..	
Case Study1	
reduce joins	
wordcount	
case1.jar	5
music.jar	11
piggybank.jar	384
reducejoin.jar	5
television.jar	9
Tv.jar	1
wordcount.jar	4
wordmean.jar	5
wordmedian.jar	5

```
[acadgild@localhost ~]$ hadoop fs -cat /output2/part-r-00000|head
18/05/23 06:16:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Toy Story (1995)        Number of times rated = 247 and average rated time = 3.872470
GoldenEye (1995)        Number of times rated = 122 and average rated time = 3.450820
City Hall (1996)        Number of times rated = 21 and average rated time = 3.428571
Keep the Lights On (2012)       Number of times rated = 1 and average rated time = 3.000000
Hansel & Gretel: Witch Hunters (2013)   Number of times rated = 2 and average rated time = 2.250000
Why Stop Now (2012)     Number of times rated = 1 and average rated time = 1.500000
Call Me Kuchu (2012)    Number of times rated = 1 and average rated time = 2.500000
Side Effects (2013)     Number of times rated = 8 and average rated time = 3.500000
Identity Thief (2013)   Number of times rated = 1 and average rated time = 1.500000
"Glimmer Man    Number of times rated = 3 and average rated time = 2.500000
cat: Unable to write to output stream.
[acadgild@localhost ~]$ hadoop fs -cat /output2/part-r-00001|tail
18/05/23 06:16:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
where applicable
Role/Play (2010)        Number of times rated = 1 and average rated time = 4.000000
Eat Sleep Die (Äta sova dö) (2012)      Number of times rated = 1 and average rated time = 4.000000
"Company You Keep       Number of times rated = 1 and average rated time = 3.500000
Set It Off (1996)       Number of times rated = 3 and average rated time = 3.666667
Beware of Mr. Baker (2012)      Number of times rated = 1 and average rated time = 4.000000
"Batman: The Dark Knight Returns        Number of times rated = 3 and average rated time = 4.000000
Paul Williams Still Alive (2011)        Number of times rated = 1 and average rated time = 4.000000
Everything or Nothing: The Untold Story of 007 (2012)   Number of times rated = 1 and average rated time = 3.000000
Mama (2013)     Number of times rated = 1 and average rated time = 2.500000
Shadow Dancer (2012)    Number of times rated = 1 and average rated time = 3.000000
[acadgild@localhost ~]$
```