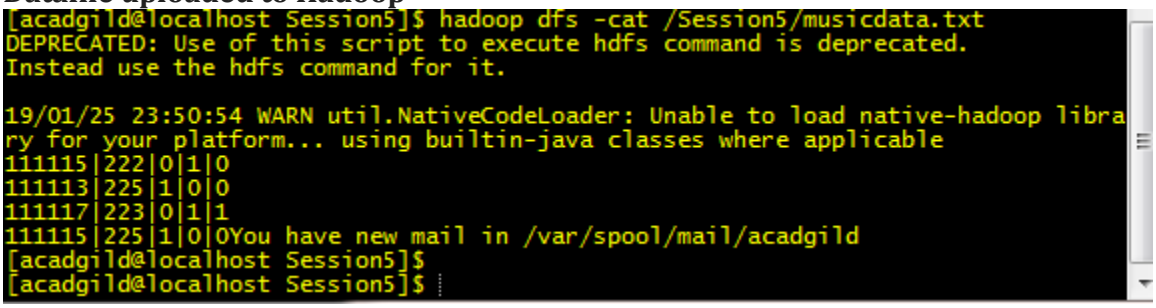
Session 5: Advance Map Reduce

5. Problem Statement

Write Map Reduce program for following tasks.

Datafile uploaded to Hadoop



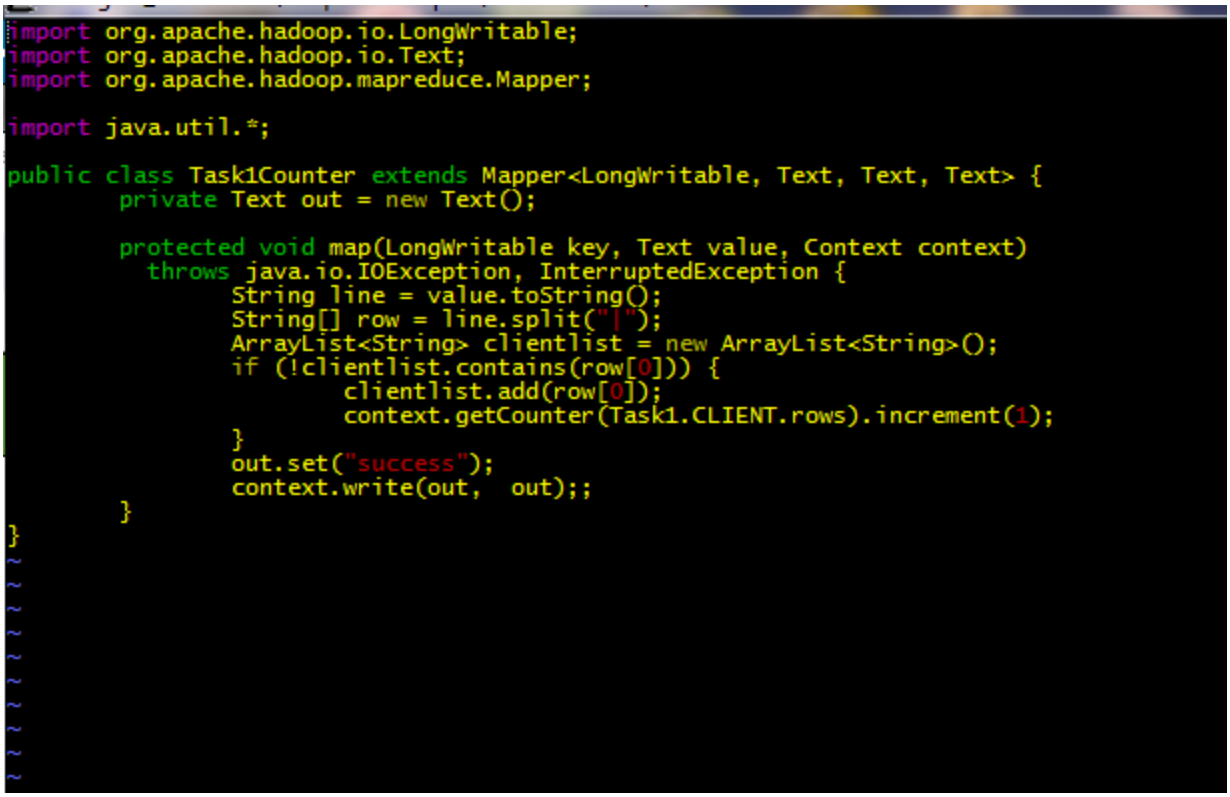
Task 1

Find the number of unique listeners in the data set.

Code:
Task1.java

```java
//Write Map Reduce program:
// - Find the number of unique listeners in the data set

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
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.mapreduce.Counters;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;


public class Task1 {

        public static enum CLIENT {
                rows
        };

        public static void main(String[] args)
                throws IOException, ClassNotFoundException, InterruptedException{
                //Job Related Configuration
                Configuration conf = new Configuration();
                Job job = new Job(conf, "Task 1 Client Counter");
                job.setJarByClass(Task1.class);

                //set reduce to zero to perform no reducer task
                job.setNumReduceTasks(0);

                //set the mapper class
                job.setMapperClass(Task1Counter.class);

                //Set the output key class and values
                job.setMapOutputKeyClass(Text.class);
                job.setMapOutputValueClass(Text.class);

                FileInputFormat.addInputPath(job,  new Path(args[0]));

                Path outputPath = new Path(args[1]);
                FileOutputFormat.setOutputPath(job,  outputPath);
                outputPath.getFileSystem(conf).delete(outputPath, true);

                //execute
                job.waitForCompletion(true);
                Counters counters = job.getCounters();
                Counter c1 = counters.findCounter(CLIENT.rows);
                System.out.println(c1.getDisplayName() + " : " + c1.getValue());


        }

}
~
```

Task1Counter.java // Mapper



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

import java.util.*;

public class Task1Counter extends Mapper<LongWritable, Text, Text, Text> {
        private Text out = new Text();

        protected void map(LongWritable key, Text value, Context context)
          throws java.io.IOException, InterruptedException {
                String line = value.toString();
                String[] row = line.split("|");
                ArrayList<String> clientlist = new ArrayList<String>();
                if (!clientlist.contains(row[0])) {
                        clientlist.add(row[0]);
                        context.getCounter(Task1.CLIENT.rows).increment(1);
                }
                out.set("success");
                context.write(out,  out);;
        }
}
~
~
~
~
~
~
~
~
~
~
```

Result:

```
[acadgild@localhost Session5]$ hadoop jar Session5Task1.jar /Session5/musicdata.
txt /Session5/task1out1.txt
19/01/25 23:46:35 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
19/01/25 23:46:38 INFO client.RMProxy: Connecting to ResourceManager at localhos
t/127.0.0.1:8032
19/01/25 23:46:41 WARN mapreduce.JobResourceUploader: Hadoop command-line option
 parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
19/01/25 23:46:42 INFO input.FileInputFormat: Total input paths to process : 1
19/01/25 23:46:42 INFO mapreduce.JobSubmitter: number of splits:1
19/01/25 23:46:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
48424476353_0001
19/01/25 23:46:44 INFO impl.YarnClientImpl: Submitted application application_15
48424476353_0001
19/01/25 23:46:44 INFO mapreduce.Job: The url to track the job: http://localhost
:8088/proxy/application_1548424476353_0001/
19/01/25 23:46:44 INFO mapreduce.Job: Running job: job_1548424476353_0001
19/01/25 23:47:09 INFO mapreduce.Job: Job job_1548424476353_0001 running in uber
 mode : false
19/01/25 23:47:09 INFO mapreduce.Job:  map 0% reduce 0%
19/01/25 23:47:19 INFO mapreduce.Job:  map 100% reduce 0%
19/01/25 23:47:20 INFO mapreduce.Job: Job job_1548424476353_0001 completed succe
ssfully
19/01/25 23:47:20 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=106984
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=182
                HDFS: Number of bytes written=64
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7231
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7231
                Total vcore-milliseconds taken by all map tasks=7231
                Total megabyte-milliseconds taken by all map tasks=7404544
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Input split bytes=109
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=96
                CPU time spent (ms)=640
                Physical memory (bytes) snapshot=85438464
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=32571392
        Task1$CLIENT
                rows=4
        File Input Format Counters
                Bytes Read=73
        File Output Format Counters
                Bytes Written=64
rows : 4
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Session5]$
```

Task 2

What are the number of times a song was heard fully.

***Please note that the program name is the same as task 1 because I decided to expand the original program.**

Task1.java

```java
//Write Map Reduce program:
// - Find the number of unique listeners in the data set

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
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.mapreduce.Counters;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;


public class Task1 {

	public static enum CLIENT {
		rows,
		listenfully
	};

	public static void main(String[] args)
		throws IOException, ClassNotFoundException, InterruptedException{
		//Job Related Configuration
		Configuration conf = new Configuration();
		Job job = new Job(conf, "Task 1 Client Counter");
		job.setJarByClass(Task1.class);

		//set reduce to zero to perform no reducer task
		job.setNumReduceTasks(0);

		//set the mapper class
		job.setMapperClass(Task1Counter.class);

		//Set the output key class and values
		job.setMapOutputKeyClass(Text.class);
		job.setMapOutputValueClass(Text.class);

		FileInputFormat.addInputPath(job,  new Path(args[0]));

		Path outputPath = new Path(args[1]);
		FileOutputFormat.setOutputPath(job,  outputPath);
		outputPath.getFileSystem(conf).delete(outputPath, true);

		//execute
		job.waitForCompletion(true);
		Counters counters = job.getCounters();
		Counter c1 = counters.findCounter(CLIENT.rows);
		Counter c2 = counters.findCounter(CLIENT.listenfully);
		System.out.println(c1.getDisplayName() + " : " + c1.getValue());
		System.out.println(c2.getDisplayName() + " : " + c2.getValue());

	}

}
~
~
```

Task1Counter.java

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

import java.util.*;

public class Task1Counter extends Mapper<LongWritable, Text, Text, Text> {
        private Text out1 = new Text();
        private Text out2 = new Text();

        protected void map(LongWritable key, Text value, Context context)
          throws java.io.IOException, InterruptedException {
                String line = value.toString();
                String[] row = line.split("\\|");
                ArrayList<String> clientlist = new ArrayList<String>();
                if (!clientlist.contains(row[0])) {
                        clientlist.add(row[0]);
                        context.getCounter(Task1.CLIENT.rows).increment(1);
                }

                if (row[4].equals("1")) { //Field 5 - Song listening status (0 for skipped, 1 for fully heard)
                        context.getCounter(Task1.CLIENT.listenfully).increment(1);
                }
                //out.set("success");
                out1.set(row[0]);
                out2.set(row[4]);

                context.write(out1,  out2);
        }
}
~
~
~
~
~
~
```

Results

```
[acadgild@localhost Session5]$ hadoop jar Session5Task1.jar /Session5/musicdata.txt /Session5/task2out6
19/01/26 00:24:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/26 00:24:44 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/26 00:24:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/01/26 00:24:46 INFO input.FileInputFormat: Total input paths to process : 1
19/01/26 00:24:46 INFO mapreduce.JobSubmitter: number of splits:1
19/01/26 00:24:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1548424476353_0008
19/01/26 00:24:46 INFO impl.YarnClientImpl: Submitted application application_1548424476353_0008
19/01/26 00:24:47 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1548424476353_0008/
19/01/26 00:24:47 INFO mapreduce.Job: Running job: job_1548424476353_0008
19/01/26 00:24:57 INFO mapreduce.Job: Job job_1548424476353_0008 running in uber mode : false
19/01/26 00:24:57 INFO mapreduce.Job:  map 0% reduce 0%
19/01/26 00:25:05 INFO mapreduce.Job:  map 100% reduce 0%
19/01/26 00:25:06 INFO mapreduce.Job: Job job_1548424476353_0008 completed successfully
19/01/26 00:25:06 INFO mapreduce.Job: Counters: 32
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=106980
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=182
                HDFS: Number of bytes written=36
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6178
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6178
                Total vcore-milliseconds taken by all map tasks=6178
                Total megabyte-milliseconds taken by all map tasks=6326272
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Input split bytes=109
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=71
                CPU time spent (ms)=660
                Physical memory (bytes) snapshot=86921216
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=32571392
        Task1$CLIENT
                listenfully=1
                rows=4
        File Input Format Counters
                Bytes Read=73
        File Output Format Counters
                Bytes Written=36
rows : 4
listenfully : 1
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Session5]$ hadoop dfs -cat /Session5/task2out6/part-m-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/26 00:25:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
111115  0
111113  0
111117  1
111115  0
[acadgild@localhost Session5]$
```

Task 3

What are the number of times a song was shared.

***Please note that the program name is the same as task 1 because I decided to expand the original program.**

Programs:

Task1.java

```java
//Write Map Reduce program:
// - Find the number of unique listeners in the data set.
// - What are the number of times a song was fully heard?
// - What are the number of times a song was shared?

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
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.mapreduce.Counters;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;


public class Task1 {

        public static enum CLIENT {
                rows,
                listenfully,
                songshared
        };

        public static void main(String[] args)
                throws IOException, ClassNotFoundException, InterruptedException{
                //Job Related Configuration
                Configuration conf = new Configuration();
                Job job = new Job(conf, "Task 1 Client Counter");
                job.setJarByClass(Task1.class);

                //set reduce to zero to perform no reducer task
                job.setNumReduceTasks(0);

                //set the mapper class
                job.setMapperClass(Task1Counter.class);

                //Set the output key class and values
                job.setMapOutputKeyClass(Text.class);
                job.setMapOutputValueClass(Text.class);

                FileInputFormat.addInputPath(job,  new Path(args[0]));

                Path outputPath = new Path(args[1]);
                FileOutputFormat.setOutputPath(job,  outputPath);
                outputPath.getFileSystem(conf).delete(outputPath, true);

                //execute
                job.waitForCompletion(true);
                Counters counters = job.getCounters();
                Counter c1 = counters.findCounter(CLIENT.rows);
                Counter c2 = counters.findCounter(CLIENT.listenfully);
                Counter c3 = counters.findCounter(CLIENT.songshared);
                System.out.println(c1.getDisplayName() + " : " + c1.getValue());
                System.out.println(c2.getDisplayName() + " : " + c2.getValue());
                System.out.println(c3.getDisplayName() + " : " + c3.getValue());
                }
}
~
~
~
~
~
```

Task1Count.java

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

import java.util.*;

public class Task1Counter extends Mapper<LongWritable, Text, Text, Text> {
        private Text out1 = new Text();
        private Text out2 = new Text();

        protected void map(LongWritable key, Text value, Context context)
          throws java.io.IOException, InterruptedException {
                String line = value.toString();
                String[] row = line.split("\\|");
                ArrayList<String> clientlist = new ArrayList<String>();
                if (!clientlist.contains(row[0])) {
                        clientlist.add(row[0]);
                        context.getCounter(Task1.CLIENT.rows).increment(1);
                }

                if (row[4].equals("1")) { //Field 5 - Song listening status (0 for skipped, 1 for fully heard)
                        context.getCounter(Task1.CLIENT.listenfully).increment(1);
                }
                if (row[2].equals("1")) {
                        context.getCounter(Task1.CLIENT.songshared).increment(1);
                }
                //out.set("success");
                out1.set(row[0]);
                out2.set(row[2]+" "+row[4]);

                context.write(out1,  out2);
        }
}
~
~
~
~
~
~
```

Results

```
[acadgild@localhost Session5]$ hadoop jar Session5Task1.jar /Session5/musicdata.txt /Session5/task3out1
19/01/26 00:38:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/26 00:38:09 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/26 00:38:10 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
19/01/26 00:38:11 INFO input.FileInputFormat: Total input paths to process : 1
19/01/26 00:38:11 INFO mapreduce.JobSubmitter: number of splits:1
19/01/26 00:38:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1548424476353_0009
19/01/26 00:38:12 INFO impl.YarnClientImpl: Submitted application application_1548424476353_0009
19/01/26 00:38:12 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1548424476353_0009/
19/01/26 00:38:12 INFO mapreduce.Job: Running job: job_1548424476353_0009
19/01/26 00:38:22 INFO mapreduce.Job: Job job_1548424476353_0009 running in uber mode : false
19/01/26 00:38:22 INFO mapreduce.Job:  map 0% reduce 0%
19/01/26 00:38:31 INFO mapreduce.Job:  map 100% reduce 0%
19/01/26 00:38:32 INFO mapreduce.Job: Job job_1548424476353_0009 completed successfully
19/01/26 00:38:32 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=106980
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=182
                HDFS: Number of bytes written=44
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6617
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6617
                Total vcore-milliseconds taken by all map tasks=6617
                Total megabyte-milliseconds taken by all map tasks=6775808
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Input split bytes=109
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=75
                CPU time spent (ms)=690
                Physical memory (bytes) snapshot=86892544
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=32571392
        Task1$CLIENT
                listenfully=1
                rows=4
                songshared=2
        File Input Format Counters
                Bytes Read=73
        File Output Format Counters
                Bytes Written=44
rows : 4
listenfully : 1
songshared : 2
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Session5]$ hadoop dfs -cat /Session5/task3out1/part-m-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

19/01/26 00:38:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
111115  0 0
111113  1 0
111117  0 1
111115  1 0
[acadgild@localhost Session5]$
```