Assignment 5.1

Dataset is sample data of songs heard by users on an online streaming platform. The

Description of data set attached in musicdata.txt is as follows: -

1st Column – User ID

2nd Column – TrackId

3rd Column – Songs share status (1 for shared, 0 for not)

4th Column – Listening platform (0 for Radio, 1 for Web)

5th Column – Song listening status (0 for skipped, 1 for fully heard)

Problem Statement

Write Map Reduce program for following tasks.

Task 1

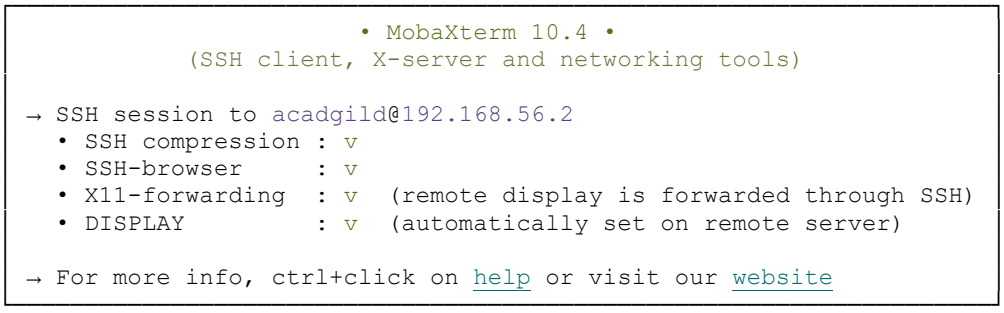
Find the number of unique listeners in the data set.

Task 2

What are the number of times a song was heard fully

Task 3

What are the number of times a song was shared




```
[acadgild.mmisra ~]$

#Copy the musicdata.txt file to HDFS under /files
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop -fs copyFromLocal musicdata.txt /files
Error: No command named `-fs' was found. Perhaps you meant `hadoop fs'
```

```
[acadgild.mmisra ~]$ hadoop fs -copyFromLocal musicdata.txt /files
18/07/05 14:57:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
```

#See if it is copied

```
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop fs -ls /files
18/07/05 14:58:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 acadgild supergroup  327155712 2018-06-24 10:51 /files/file327.txt
-rw-r--r--   1 acadgild supergroup        252 2018-07-05 14:57 /files/musicdata.txt
-rw-r--r--   1 acadgild supergroup        735 2018-07-04 13:21 /files/television.txt
-rw-r--r--   1 acadgild supergroup        336 2018-06-24 10:51 /files/test.txt
```

```
[acadgild.mmisra ~]$
```

#Dump the content of the file
#As per assignment
#1st Column – User ID
#2nd Column – TrackId
#3rd Column – Songs share status (1 for shared, 0 for not)
#4th Column – Listening platform (0 for Radio, 1 for Web)
#5th Column – Song listening status (0 for skipped, 1 for fully heard)

```
[acadgild.mmisra ~]$ hadoop fs -cat /files/musicdata.txt
18/07/05 15:09:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
111115|222|0|1|0
111113|225|1|0|0
111117|223|0|1|1
111115|225|1|0|0
222980|229|0|1|0
278338|908|1|0|1
```

```
122902|300|1|0|0
222980|222|1|1|1
333838|115|0|0|0
933833|112|1|0|0
333838|225|0|1|1
733838|223|1|0|1
225336|223|1|1|0
633533|225|0|1|1
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$

[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar SongExample.jar
Song database Example
Valid options are:
SongExample task1 <input path> <output path>
SongExample task2 <input path> <output path>
SongExample task3 <input path> <output path>
```

#task1 is to find number of Find the number of unique listeners in the data set.

```
[acadgild.mmisra ~]$ hadoop jar SongExample.jar task1 /files/musicdata.txt /out1
Song database Example
18/07/05 15:04:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 15:04:33 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 15:04:34 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
18/07/05 15:04:34 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 15:04:35 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 15:04:35 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0006
18/07/05 15:04:35 INFO impl.YarnClientImpl: Submitted application
application_1530772907907_0006
18/07/05 15:04:35 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0006/
18/07/05 15:04:35 INFO mapreduce.Job: Running job: job_1530772907907_0006
18/07/05 15:04:41 INFO mapreduce.Job: Job job_1530772907907_0006 running in uber mode
: false
18/07/05 15:04:41 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 15:04:46 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 15:04:52 INFO mapreduce.Job:  map 100% reduce 100%
18/07/05 15:04:53 INFO mapreduce.Job: Job job_1530772907907_0006 completed
successfully
18/07/05 15:04:53 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=188
                FILE: Number of bytes written=215635
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=358
                HDFS: Number of bytes written=99
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2842
                Total time spent by all reduces in occupied slots (ms)=3077
                Total time spent by all map tasks (ms)=2842
                Total time spent by all reduce tasks (ms)=3077
                Total vcore-milliseconds taken by all map tasks=2842
                Total vcore-milliseconds taken by all reduce tasks=3077
                Total megabyte-milliseconds taken by all map tasks=2910208
                Total megabyte-milliseconds taken by all reduce tasks=3150848
        Map-Reduce Framework
                Map input records=14
                Map output records=14
                Map output bytes=154
                Map output materialized bytes=188
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=11
                Reduce shuffle bytes=188
                Reduce input records=14
                Reduce output records=11
                Spilled Records=28
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=115
                CPU time spent (ms)=1460
                Physical memory (bytes) snapshot=417832960
                Virtual memory (bytes) snapshot=4167868416
                Total committed heap usage (bytes)=326107136
        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=252
        File Output Format Counters
                Bytes Written=99
Job success
Song Example Success
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$ hadoop fs -ls /out1
18/07/05 15:05:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-05 15:04 /out1/_SUCCESS
-rw-r--r--   1 acadgild supergroup         99 2018-07-05 15:04 /out1/part-r-00000
```

#output – List of unique Listener IDs

```
[acadgild.mmisra ~]$ hadoop fs -cat /out1/part-r-00000
18/07/05 15:05:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
111113  1
111115  2
111117  1
122902  1
222980  2
225336  1
278338  1
```

```
333838  2
633533  1
733838  1
933833  1
```

#task2 is to find number of of times a song was heard fully

```
[acadgild.mmisra ~]$ hadoop jar SongExample.jar task2 /files/musicdata.txt /out2
Song database Example
18/07/05 15:07:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 15:07:08 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 15:07:09 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
18/07/05 15:07:10 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 15:07:10 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 15:07:10 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0007
18/07/05 15:07:10 INFO impl.YarnClientImpl: Submitted application
application_1530772907907_0007
18/07/05 15:07:10 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0007/
18/07/05 15:07:10 INFO mapreduce.Job: Running job: job_1530772907907_0007
18/07/05 15:07:17 INFO mapreduce.Job: Job job_1530772907907_0007 running in uber mode
: false
18/07/05 15:07:17 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 15:07:22 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 15:07:29 INFO mapreduce.Job:  map 100% reduce 100%
18/07/05 15:07:29 INFO mapreduce.Job: Job job_1530772907907_0007 completed
successfully
18/07/05 15:07:29 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=146
                FILE: Number of bytes written=215551
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=358
                HDFS: Number of bytes written=48
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2893
                Total time spent by all reduces in occupied slots (ms)=3164
                Total time spent by all map tasks (ms)=2893
                Total time spent by all reduce tasks (ms)=3164
                Total vcore-milliseconds taken by all map tasks=2893
                Total vcore-milliseconds taken by all reduce tasks=3164
                Total megabyte-milliseconds taken by all map tasks=2962432
                Total megabyte-milliseconds taken by all reduce tasks=3239936
        Map-Reduce Framework
                Map input records=14
                Map output records=14
                Map output bytes=112
                Map output materialized bytes=146
                Input split bytes=106
```

```
                Combine input records=0
                Combine output records=0
                Reduce input groups=8
                Reduce shuffle bytes=146
                Reduce input records=14
                Reduce output records=8
                Spilled Records=28
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=119
                CPU time spent (ms)=1460
                Physical memory (bytes) snapshot=418664448
                Virtual memory (bytes) snapshot=4164575232
                Total committed heap usage (bytes)=324534272
        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=252
        File Output Format Counters
                Bytes Written=48
Job success
Song Example Success
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar SongExfs -ls /out2
18/07/05 15:07:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-05 15:07 /out2/_SUCCESS
-rw-r--r--   1 acadgild supergroup         48 2018-07-05 15:07 /out2/part-r-00000
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop fs -ls /out2
18/07/05 15:07:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-05 15:07 /out2/_SUCCESS
-rw-r--r--   1 acadgild supergroup         48 2018-07-05 15:07 /out2/part-r-00000
```

#Output – listing Track IDs and how many times they were heard fully

```
[acadgild.mmisra ~]$ hadoop fs -cat  /out2/part-r-00000
18/07/05 15:07:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
112     0
115     0
222     1
223     2
225     2
229     0
300     0
908     1
```

#task3 is to find number of times a song was shared

```
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop jar SongExample.jar task3 /files/musicdata.txt /out3
Song database Example
18/07/05 15:08:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
18/07/05 15:08:18 INFO client.RMProxy: Connecting to ResourceManager at
localhost/127.0.0.1:8032
18/07/05 15:08:18 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
18/07/05 15:08:19 INFO input.FileInputFormat: Total input paths to process : 1
18/07/05 15:08:19 INFO mapreduce.JobSubmitter: number of splits:1
18/07/05 15:08:19 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1530772907907_0008
18/07/05 15:08:19 INFO impl.YarnClientImpl: Submitted application
application_1530772907907_0008
18/07/05 15:08:19 INFO mapreduce.Job: The url to track the job:
http://localhost:8088/proxy/application_1530772907907_0008/
18/07/05 15:08:19 INFO mapreduce.Job: Running job: job_1530772907907_0008
18/07/05 15:08:27 INFO mapreduce.Job: Job job_1530772907907_0008 running in uber mode
: false
18/07/05 15:08:27 INFO mapreduce.Job:  map 0% reduce 0%
18/07/05 15:08:32 INFO mapreduce.Job:  map 100% reduce 0%
18/07/05 15:08:37 INFO mapreduce.Job:  map 100% reduce 100%
18/07/05 15:08:37 INFO mapreduce.Job: Job job_1530772907907_0008 completed
successfully
18/07/05 15:08:37 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=146
                FILE: Number of bytes written=215551
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=358
                HDFS: Number of bytes written=48
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2813
                Total time spent by all reduces in occupied slots (ms)=3164
                Total time spent by all map tasks (ms)=2813
                Total time spent by all reduce tasks (ms)=3164
                Total vcore-milliseconds taken by all map tasks=2813
                Total vcore-milliseconds taken by all reduce tasks=3164
                Total megabyte-milliseconds taken by all map tasks=2880512
                Total megabyte-milliseconds taken by all reduce tasks=3239936
        Map-Reduce Framework
                Map input records=14
                Map output records=14
                Map output bytes=112
                Map output materialized bytes=146
                Input split bytes=106
                Combine input records=0
```

```
                Combine output records=0
                Reduce input groups=8
                Reduce shuffle bytes=146
                Reduce input records=14
                Reduce output records=8
                Spilled Records=28
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=134
                CPU time spent (ms)=1450
                Physical memory (bytes) snapshot=419172352
                Virtual memory (bytes) snapshot=4164554752
                Total committed heap usage (bytes)=310378496
        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=252
        File Output Format Counters
                Bytes Written=48
Job success
Song Example Success
You have new mail in /var/spool/mail/acadgild
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$
[acadgild.mmisra ~]$ hadoop fs -ls /out3
18/07/05 15:08:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-07-05 15:08 /out3/_SUCCESS
-rw-r--r--   1 acadgild supergroup         48 2018-07-05 15:08 /out3/part-r-00000
```

#Output – listing Track IDs and how many times it was shared

```
[acadgild.mmisra ~]$ hadoop fs -cat /out3/part-r-00000
18/07/05 15:08:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for
your platform... using builtin-java classes where applicable
112     1
115     0
222     1
223     2
225     2
229     0
300     1
908     1
[acadgild.mmisra ~]$
```