



By:

Rashmi Krishna

BIGDATA HADOOP & SPARK TRAINING

Assignment 5: Advance MapReduce Programs

Input data for all the tasks:

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

Description of data set is as follows: -

User ID	Track ID	Song Share Status	Listening Platform	Song Listening Status
111115	222	0	1	0
111113	225	1	0	0
111117	223	0	1	1
111115	225	1	0	0

Task1:Find the number of unique listeners in the data set:

- Created a jar file in eclipse with 1 mapper,1reducer and 1driver.
- In Mapper: Divided the input Strings by “|” and bifurcated User ID & Track ID from the input string
- In Reducer: Filtered out unique users by using **HashSet** function for the respective Track ID.
- Executed the jar file musicdata.jar to get the number of unique listeners in the data set.

```
[acadgild@localhost musicassgnt]$ hadoop jar musicdata.jar /hadoopdata/music/musicdata.txt musicout1
18/03/19 11:03:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/03/19 11:03:09 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/19 11:03:10 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/03/19 11:03:11 INFO input.FileInputFormat: Total input paths to process : 1
18/03/19 11:03:11 INFO mapreduce.JobSubmitter: number of splits:1
18/03/19 11:03:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521435307904_0005
18/03/19 11:03:12 INFO impl.YarnClientImpl: Submitted application application_1521435307904_0005
18/03/19 11:03:12 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1521435307904_0005/
18/03/19 11:03:12 INFO mapreduce.Job: Running job: job_1521435307904_0005
18/03/19 11:03:25 INFO mapreduce.Job: Job job_1521435307904_0005 running in uber mode : false
18/03/19 11:03:25 INFO mapreduce.Job:  map 0% reduce 0%
18/03/19 11:03:33 INFO mapreduce.Job:  map 100% reduce 0%
18/03/19 11:03:45 INFO mapreduce.Job:  map 100% reduce 100%
18/03/19 11:03:46 INFO mapreduce.Job: Job job_1521435307904_0005 completed successfully
18/03/19 11:03:46 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=46
      FILE: Number of bytes written=216141
      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
      HDFS: Number of bytes written=18
      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)=6295
      Total time spent by all reduces in occupied slots (ms)=8671
```

```



Total vcore-milliseconds taken by all map tasks=6295
Total vcore-milliseconds taken by all reduce tasks=8671
Total megabyte-milliseconds taken by all map tasks=6446080
Total megabyte-milliseconds taken by all reduce tasks=8879104
Map-Reduce Framework
  Map input records=5
  Map output records=4
  Map output bytes=32
  Map output materialized bytes=46
  Input split bytes=117
  Combine input records=0
  Combine output records=0
  Reduce input groups=3
  Reduce shuffle bytes=46
  Reduce input records=4
  Reduce output records=3
  Spilled Records=0
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=260
  CPU time spent (ms)=2010
  Physical memory (bytes) snapshot=299339776
  Virtual memory (bytes) snapshot=4118206320
  Total committed heap usage (bytes)=170004480
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=74
File Output Format Counters
  Bytes Written=18
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost musicassgnt]$ clear

```

```

[acadgild@localhost musicassgnt]$ hadoop fs -cat musicout1/part-r-00000
18/03/19 11:05:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
222 1
223 1
225 2
[acadgild@localhost musicassgnt]$

```


 List of Uniques listeners of respective Song Tracks
 
 2 Listeners indicate, there are 2 unique listeners of song track 225

222	1
223	1
225	2

Task 2: What are the number of times a song was heard fully:

- Created a jar file in eclipse with 1 mapper, 1 reducer and 1 driver.
- In Mapper: Divided the input Strings by “|” and bifurcated User ID, Track ID & Song Listening Status from the input string.
 - Filtered the song listening status with value “1” by using **equalsIgnoreCase** function.
 - Using counter counted the number of times the song was heard fully by a User.
- In Reducer: passed the data to the Driver to print the outputs.
- Executed the jar file musicdata1.jar to get the number times a song was heard fully.

```

[acadgild@localhost musicassgnt]$ hadoop jar musicdata.jar /hadoopdata/music/musicdata.txt music
18/03/20 11:34:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/03/20 11:34:07 INFO client.RMPProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/20 11:34:09 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/03/20 11:34:10 INFO input.FileInputFormat: Total input paths to process : 1
18/03/20 11:34:11 INFO mapreduce.JobSubmitter: number of splits:1
18/03/20 11:34:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521525676195_0001
18/03/20 11:34:13 INFO impl.YarnClientImpl: Submitted application application_1521525676195_0001
18/03/20 11:34:13 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1521525676195_0001/
18/03/20 11:34:13 INFO mapreduce.Job: Running job: job_1521525676195_0001
18/03/20 11:34:38 INFO mapreduce.Job: Job job_1521525676195_0001 running in uber mode : false
18/03/20 11:34:38 INFO mapreduce.Job: map 0% reduce 0%
18/03/20 11:34:47 INFO mapreduce.Job: map 100% reduce 0%
18/03/20 11:34:57 INFO mapreduce.Job: map 100% reduce 100%
18/03/20 11:34:58 INFO mapreduce.Job: Job job_1521525676195_0001 completed successfully
18/03/20 11:34:58 INFO mapreduce.Job: Counters: 49

File System Counters
  FILE: Number of bytes read=19
  FILE: Number of bytes written=216337
  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
  HDFS: Number of bytes written=11
  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)=7081
  Total time spent by all reduces in occupied slots (ms)=8039
  Total time spent by all map tasks (ms)=7081
  Total time spent by all reduce tasks (ms)=8039

```

```

  Map input records=5
  Map output records=1
  Map output bytes=6
  Map output materialized bytes=14
  Input split bytes=117
  Combine input records=1
  Combine output records=1
  Reduce input groups=1
  Reduce shuffle bytes=14
  Reduce input records=1
  Reduce output records=1
  Spilled Records=2
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=272
  CPU time spent (ms)=1890
  Physical memory (bytes) snapshot=299044864
  Virtual memory (bytes) snapshot=4118192128
  Total committed heap usage (bytes)=170004480

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=74
File Output Format Counters
  Bytes Written=6
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost musicassgnt]$ hadoop fs -cat music/part-r-00000
18/03/20 11:55:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
223 1
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost musicassgnt]$

```

← This Output indicates that Track 223 was heard fully only once

Task 3: What are the number of times a song was shared.

- Created a jar file in eclipse with 1 mapper, 1 reducer and 1 driver.
- In Mapper: Divided the input Strings by “|” and bifurcated User ID, Track ID & Song Share Status from the input string.
 - Filtered the song listening status with value “1” by using **equalsIgnoreCase** function.
 - Filtered Track ID based on the above equalsIgnorecase function and passed it to the reducer.
- In Reducer: calculated the sum of values of Song Share Status to find the number of times a song was shared.
- Executed the jar file musicdata2.jar to get the number times a song was shared.

```
[acagild@localhost musicassgnt]$ hadoop jar musicdata2.jar /hadoopdata/music/musicdata.txt music out6
18/03/20 16:28:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/03/20 16:28:59 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/03/20 16:29:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/03/20 16:29:01 INFO input.FileInputFormat: Total input paths to process : 1
18/03/20 16:29:01 INFO mapreduce.JobSubmitter: number of splits:1
18/03/20 16:29:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1521541544818_0004
18/03/20 16:29:02 INFO impl.YarnClientImpl: Submitted application application_1521541544818_0004
18/03/20 16:29:02 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1521541544818_0004/
18/03/20 16:29:02 INFO mapreduce.Job: Running job: job_1521541544818_0004
18/03/20 16:29:16 INFO mapreduce.Job: Job job_1521541544818_0004 running in uber mode : false
18/03/20 16:29:16 INFO mapreduce.Job:  map 0% reduce 0%
18/03/20 16:29:27 INFO mapreduce.Job:  map 100% reduce 0%
18/03/20 16:29:39 INFO mapreduce.Job:  map 100% reduce 100%
18/03/20 16:29:40 INFO mapreduce.Job: Job job_1521541544818_0004 completed successfully
18/03/20 16:29:41 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=16
    FILE: Number of bytes written=216381
    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
    HDFS: Number of bytes written=6
    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)=8449
    Total time spent by all reduces in occupied slots (ms)=9193
    Total time spent by all map tasks (ms)=8449
    Total time spent by all reduce tasks (ms)=9193
```

```

Map-Reduce Framework
  Map input records=5
  Map output records=2
  Map output bytes=16
  Map output materialized bytes=16
  Input split bytes=117
  Combine input records=2
  Combine output records=1
  Reduce input groups=1
  Reduce shuffle bytes=16
  Reduce input records=1
  Reduce output records=1
  Spilled Records=2
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=350
  CPU time spent (ms)=2320
  Physical memory (bytes) snapshot=302219264
  Virtual memory (bytes) snapshot=4117913600
  Total committed heap usage (bytes)=170004480

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=74

File Output Format Counters
  Bytes Written=6

You have new mail in /var/spool/mail/acadgild
[acadgild@localhost musicassgnt]$ hadoop fs -cat music_out6/part-r-00000
18/03/20 16:30:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
225      2
[acadgild@localhost musicassgnt]$

```

This indicates that the Track 225 was shared twice in the given data set