Session 5 - ADVANCE MAPREDUCE AND INTRODUCTION TO UNIX CONCEPTS

Assignment 1

There are 3 jar files for each of the task. Each task has a separate Package and a separate driver program.

The first task is a mapreduce job, while the 2nd and 3rd tasks are map only jobs.

Syntax for the hadoop command used :

Hadoop jar <location of jar in local FS> <HDFS location of input text file> <HDFS location of the output to be created.>

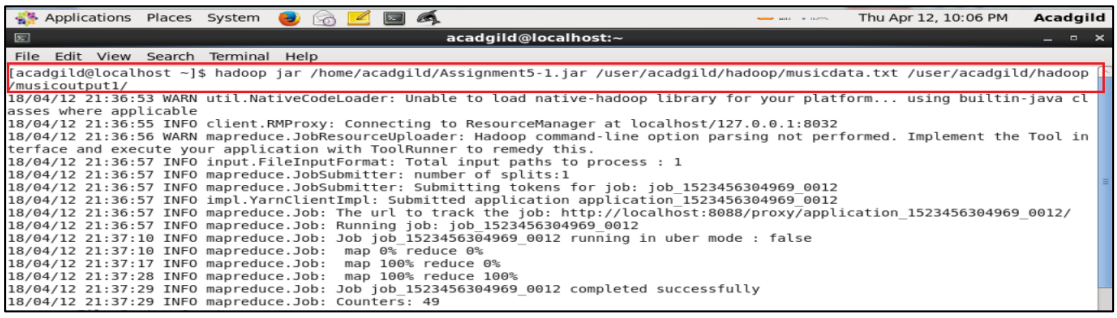
Main class for each jar file is set while exporting the jar in Eclipse.

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

Jar name : Assignment5-1.jar

Hadoop command :

hadoop jar /home/acadgild/Assignment5-1.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop/musicoutput1



```
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Assignment5-1.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop
/musicoutput1/
18/04/12 21:36:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/04/12 21:36:55 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/04/12 21:36:56 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
18/04/12 21:36:57 INFO input.FileInputFormat: Total input paths to process : 1
18/04/12 21:36:57 INFO mapreduce.JobSubmitter: number of splits:1
18/04/12 21:36:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1523456304969_0012
18/04/12 21:36:57 INFO impl.YarnClientImpl: Submitted application application_1523456304969_0012
18/04/12 21:36:57 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1523456304969_0012/
18/04/12 21:36:57 INFO mapreduce.Job: Running job: job_1523456304969_0012
18/04/12 21:37:10 INFO mapreduce.Job: Job job_1523456304969_0012 running in uber mode : false
18/04/12 21:37:10 INFO mapreduce.Job:  map 0% reduce 0%
18/04/12 21:37:17 INFO mapreduce.Job:  map 100% reduce 0%
18/04/12 21:37:28 INFO mapreduce.Job:  map 100% reduce 100%
18/04/12 21:37:29 INFO mapreduce.Job: Job job_1523456304969_0012 completed successfully
18/04/12 21:37:29 INFO mapreduce.Job: Counters: 49
```

Output folder : musicoutput1

A '_SUCCESS' file is created indicating the successful execution of the job.

The output of the job id stored in the part file.

```
                CPU time spent (ms)=1810
                Physical memory (bytes) snapshot=306098176
                Virtual memory (bytes) snapshot=4118196224
                Total committed heap usage (bytes)=202379264
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        com.acadgild.Assgn5Task1.Assgn5Task1$COUNTERS1
                RECORD_COUNTER=2
        File Input Format Counters
                Bytes Read=68
        File Output Format Counters
                Bytes Written=18
number of unique listeners  : 2
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/hadoop/musicoutput1/
18/04/13 05:36:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-04-13 05:36 /user/acadgild/hadoop/musicoutput1/_SUCCESS
-rw-r--r--   1 acadgild supergroup         18 2018-04-13 05:36 /user/acadgild/hadoop/musicoutput1/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/hadoop/musicoutput1/part-r-00000
18/04/13 05:36:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
111113  1
111117  1
[acadgild@localhost ~]$
```

The output shows that the number of unique listeners is 2. (which is equal to the value of counter).

The part file comtains the UserIds that occurred once in the file, and hence are unique listeners.

Task2 : What are the number of times a song was heard fully.

<u>Jar name :</u> Assignment5-2.jar

<u>Hadoop command :</u>

hadoop jar /home/acadgild/Assignment5-2.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop/musicoutput2

```
Applications  Places  System                                   Thu Apr 12, 10:09 PM   Acadgild
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Assignment5-2.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop
/musicoutput2/
18/04/12 22:07:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/04/12 22:07:39 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/04/12 22:07:41 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
18/04/12 22:07:41 INFO input.FileInputFormat: Total input paths to process : 1
18/04/12 22:07:41 INFO mapreduce.JobSubmitter: number of splits:1
18/04/12 22:07:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1523456304969_0013
18/04/12 22:07:42 INFO impl.YarnClientImpl: Submitted application application_1523456304969_0013
18/04/12 22:07:42 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1523456304969_0013/
18/04/12 22:07:42 INFO mapreduce.Job: Running job: job_1523456304969_0013
18/04/12 22:07:54 INFO mapreduce.Job: Job job_1523456304969_0013 running in uber mode : false
18/04/12 22:07:54 INFO mapreduce.Job:  map 0% reduce 0%
18/04/12 22:08:02 INFO mapreduce.Job:  map 100% reduce 0%
18/04/12 22:08:11 INFO mapreduce.Job:  map 100% reduce 100%
18/04/12 22:08:11 INFO mapreduce.Job: Job job_1523456304969_0013 completed successfully
18/04/12 22:08:11 INFO mapreduce.Job: Counters: 49
```

<u>Output folder :</u> musicoutput2

A '_SUCCESS' file is created indicating the successful execution of the job.

The output of the job id stored in the part file.

```
                Map input records=4
                Map output records=1
                Input split bytes=121
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=67
                CPU time spent (ms)=800
                Physical memory (bytes) snapshot=101892096
                Virtual memory (bytes) snapshot=2056761344
                Total committed heap usage (bytes)=48758784
        com.acadgild.Assgn5Task2.Assgn5Task2$COUNTERS
                RECORD_COUNTER=1
        File Input Format Counters
                Bytes Read=68
        File Output Format Counters
                Bytes Written=6
number of times a song was heard fully :1
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/hadoop/musicoutput2/
18/04/13 05:27:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-04-13 05:26 /user/acadgild/hadoop/musicoutput2/_SUCCESS
-rw-r--r--   1 acadgild supergroup          6 2018-04-13 05:26 /user/acadgild/hadoop/musicoutput2/part-m-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/hadoop/musicoutput2/part-m-00000
18/04/13 05:28:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
223     1
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

The output shows that the number of times a song was heard fully without skipping is 1. (which is equal to the value of counter).

The part file contains the TrackIds that had the value in the 5th column as 1 indicating that the song was fully heard and not skipped.

Task3 : What are the number of times a song was shared.

<u>Jar name :</u> Assignment5-3.jar

<u>Hadoop command :</u>

hadoop jar /home/acadgild/Assignment5-3.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop/musicoutput3

```
Applications  Places  System                                   Thu Apr 12, 10:11 PM   Acadgild
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ hadoop jar /home/acadgild/Assignment5-3.jar /user/acadgild/hadoop/musicdata.txt /user/acadgild/hadoop
/musicoutput3/
18/04/12 22:10:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/04/12 22:10:27 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/04/12 22:10:28 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool in
terface and execute your application with ToolRunner to remedy this.
18/04/12 22:10:28 INFO input.FileInputFormat: Total input paths to process : 1
18/04/12 22:10:29 INFO mapreduce.JobSubmitter: number of splits:1
18/04/12 22:10:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1523456304969_0014
18/04/12 22:10:29 INFO impl.YarnClientImpl: Submitted application application_1523456304969_0014
18/04/12 22:10:29 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1523456304969_0014/
18/04/12 22:10:29 INFO mapreduce.Job: Running job: job_1523456304969_0014
18/04/12 22:10:40 INFO mapreduce.Job: Job job_1523456304969_0014 running in uber mode : false
18/04/12 22:10:40 INFO mapreduce.Job:  map 0% reduce 0%
18/04/12 22:10:48 INFO mapreduce.Job:  map 100% reduce 0%
18/04/12 22:10:58 INFO mapreduce.Job:  map 100% reduce 100%
18/04/12 22:10:58 INFO mapreduce.Job: Job job_1523456304969_0014 completed successfully
18/04/12 22:10:59 INFO mapreduce.Job: Counters: 49
```

<u>Output folder :</u> musicoutput3

A '_SUCCESS' file is created indicating the successful execution of the job.

The output of the job id stored in the part file.

```
        Map-Reduce Framework
                Map input records=4
                Map output records=2
                Input split bytes=121
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=60
                CPU time spent (ms)=690
                Physical memory (bytes) snapshot=101982208
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=48758784
        com.acadgild.Assign5Task3.Assgn5Task3$COUNTERS3
                RECORD_COUNTER=2
        File Input Format Counters
                Bytes Read=68
        File Output Format Counters
                Bytes Written=12
Number of times a song was shared : 2
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/hadoop/musicoutput3/
18/04/13 05:29:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-04-13 05:29 /user/acadgild/hadoop/musicoutput3/_SUCCESS
-rw-r--r--   1 acadgild supergroup         12 2018-04-13 05:29 /user/acadgild/hadoop/musicoutput3/part-m-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/hadoop/musicoutput3/part-m-00000
18/04/13 05:30:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
225     1
225     1
```

The output shows that the number of times a song was shared is 2. (which is equal to the value of counter).

The part file contains the TrackIds that had the value in the 3rd column as 1 indicating that the song was shared.