READ ME

1. Jar file name is “saavntrendproject.jar”. This was used for execution of Map Reduce program.
2. “MROutput.zip” contains output of Map Reduce program. It has seven files namely part-r-*. One for each day.
3. Folder “finaloutput” contains final output for this project, 7 files, each containing top 100 trending songs (songIds) for each day. Trending songs for each day are present in the file for that day. Trending songs for 25-12-2017 are present in 25.txt and so on.
4. Use notepad++ or other suitable editor to view the contents of these files.
5. Zip file “saavntrendproject.zip” contains project exported from eclipse. It has all the folder structure along with *.java files used for creating saavntrendproject.jar file.
6. Algorithm used for Map Reduce program is provided in file “Algorithm used for MapReduce Program.docx”
7. I have also provided a shell script “matchsaavntrend.sh” to do all the work needed. Using this shell script, I have automated whole process.
8. Copy jar file saavntrendproject.jar and shell script matchsaavntrend.sh in a particular directory on ec2 instance and ensure shell script has executable permission. Otherwise run chmod 755 matchsaavntrend.sh on unix prompt.
9. Since you might have copied file from windows to ec2 instance, run this command to remove special characters from shell script:

   perl -pi -e ‘s/^M$//’ matchsaavntrend.sh

   Note: You will have to type this command as its is ctrl+v+ctrl+m character and not just ^M so please take care and ensure that it is removed from the file as this character appears when script is copied from windows to linux. Alternatively, you can create new file on ec2 instance and copy in it, contents of file on windows and save it and provide executable permission.
10. Execute shell script from inside the directory in which both jar file and shell script were copied, using below command:
    ./matchsaavntrend.sh
11. And then just wait for it get over.

Steps followed in shell script:

(a) It first cleans up output directory (if exist already) on local file system as well as on s3.
(b) It then executes map reduce program using hadoop jar command.
(c) After successful completion of map reduce program, it copies files from s3 to local file system.
(d) It also copies trending list provided by Saavn from s3 to local file system. Trending list provided by Saavn is a “,” separated list of values having data for all days in December 2017.
(e) We are interested in data only from 25th Dec to 31st Dec 2017 so script extracts data from trending list provided by Saavn only for these days using grep command and then cut on “,” and takes out only 1st field which is the songId and then sort the data and store in separate files for each day like data for 25th Dec will be present in saavn25.txt and so on.

(f) MapReduce output files are sorted on 2nd key (which is the count for each key) in reverse numeric order so key having maximum count will appear on top and output is stored in separate files namely new25.txt and so on. Unix sort command is used for the same.
(g) After sorting, first 100 records are taken from these files, only songIds are extracted using cut command and then sorted and stored in separate files for each day like 25.txt and so on.
(h) After all this, match is calculated between trending list provided by Saavn and files produced by my program using comm command and number of songs matching between both lists are then shown as output.
(i) Finally, MapReduce output files are removed from s3.
(j) Script generates appropriate messages at regular times to show the progress.

Below is the output shown when I executed matchsaavntrend.sh on my ec2 instance:

[ec2-user@ip-172-31-91-95 scriptsaavn]$ ls -lrt

total 16

-rw-rw-r-- 1 ec2-user ec2-user 7127 Aug 3 16:47 saavntrendproject.jar

-rwxrwxr-x 1 ec2-user ec2-user 4449 Aug 12 03:52 matchsaavntrend.sh

[ec2-user@ip-172-31-91-95 scriptsaavn]$ ./matchsaavntrend.sh

Cleaning up

Clean up completed

Executing MapReduce

18/08/12 03:52:43 WARN impl.MetricsConfig: Cannot locate configuration: tried hadoop-metrics2-s3a-file-system.properties,hadoop-metrics2.properties

18/08/12 03:52:43 INFO impl.MetricsSystemImpl: Scheduled snapshot period at 10 second(s).

18/08/12 03:52:43 INFO impl.MetricsSystemImpl: s3a-file-system metrics system started

18/08/12 03:52:45 INFO Configuration.deprecation: fs.s3a.server-side-encryption-key is deprecated. Instead, use fs.s3a.server-side-encryption.key

18/08/12 03:52:45 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-91-95.ec2.internal/172.31.91.95:8032

18/08/12 03:52:46 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

18/08/12 03:52:46 INFO input.FileInputFormat: Total input paths to process : 1

18/08/12 03:52:46 INFO mapreduce.JobSubmitter: number of splits:176

18/08/12 03:52:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1534045789091_0001

18/08/12 03:52:47 INFO impl.YarnClientImpl: Submitted application application_1534045789091_0001

18/08/12 03:52:47 INFO mapreduce.Job: The url to track the job: http://ip-172-31-91-95.ec2.internal:8088/proxy/application_1534045789091_0001/

18/08/12 03:52:47 INFO mapreduce.Job: Running job: job_1534045789091_0001

18/08/12 03:52:58 INFO mapreduce.Job: Job job_1534045789091_0001 running in uber mode : false

18/08/12 03:52:58 INFO mapreduce.Job: map 0% reduce 0%

18/08/12 03:53:19 INFO mapreduce.Job: map 1% reduce 0%

18/08/12 03:53:21 INFO mapreduce.Job: map 2% reduce 0%

18/08/12 03:53:39 INFO mapreduce.Job: map 3% reduce 0%

18/08/12 03:53:56 INFO mapreduce.Job: map 4% reduce 0%

18/08/12 03:54:04 INFO mapreduce.Job: map 5% reduce 0%

18/08/12 03:54:19 INFO mapreduce.Job: map 6% reduce 0%

18/08/12 03:54:25 INFO mapreduce.Job: map 7% reduce 0%

18/08/12 03:54:43 INFO mapreduce.Job: map 8% reduce 0%

18/08/12 03:54:53 INFO mapreduce.Job: map 9% reduce 0%

18/08/12 03:55:04 INFO mapreduce.Job: map 10% reduce 0%

18/08/12 03:55:16 INFO mapreduce.Job: map 11% reduce 0%

18/08/12 03:55:29 INFO mapreduce.Job: map 12% reduce 0%

18/08/12 03:55:38 INFO mapreduce.Job: map 13% reduce 0%

18/08/12 03:55:52 INFO mapreduce.Job: map 14% reduce 0%

18/08/12 03:56:05 INFO mapreduce.Job: map 15% reduce 0%

18/08/12 03:56:15 INFO mapreduce.Job: map 16% reduce 0%

18/08/12 03:56:29 INFO mapreduce.Job: map 17% reduce 0%

18/08/12 03:56:38 INFO mapreduce.Job: map 18% reduce 0%

18/08/12 03:56:50 INFO mapreduce.Job: map 19% reduce 0%

18/08/12 03:57:04 INFO mapreduce.Job: map 20% reduce 0%

18/08/12 03:57:16 INFO mapreduce.Job: map 21% reduce 0%

18/08/12 03:57:28 INFO mapreduce.Job: map 22% reduce 0%

18/08/12 03:57:36 INFO mapreduce.Job: map 23% reduce 0%

18/08/12 03:57:52 INFO mapreduce.Job: map 24% reduce 0%

18/08/12 03:58:09 INFO mapreduce.Job: map 25% reduce 0%

18/08/12 03:58:13 INFO mapreduce.Job: map 26% reduce 0%

18/08/12 03:58:30 INFO mapreduce.Job: map 27% reduce 0%

18/08/12 03:58:35 INFO mapreduce.Job: map 28% reduce 0%

18/08/12 03:58:51 INFO mapreduce.Job: map 29% reduce 0%

18/08/12 03:58:58 INFO mapreduce.Job: map 30% reduce 0%

18/08/12 03:59:13 INFO mapreduce.Job: map 31% reduce 0%

18/08/12 03:59:26 INFO mapreduce.Job: map 32% reduce 0%

18/08/12 03:59:36 INFO mapreduce.Job: map 33% reduce 0%

18/08/12 03:59:50 INFO mapreduce.Job: map 34% reduce 0%

18/08/12 03:59:56 INFO mapreduce.Job: map 35% reduce 0%

18/08/12 04:00:12 INFO mapreduce.Job: map 36% reduce 0%

18/08/12 04:00:27 INFO mapreduce.Job: map 37% reduce 0%

18/08/12 04:00:31 INFO mapreduce.Job: map 38% reduce 0%

18/08/12 04:00:51 INFO mapreduce.Job: map 39% reduce 0%

18/08/12 04:01:00 INFO mapreduce.Job: map 40% reduce 0%

18/08/12 04:01:13 INFO mapreduce.Job: map 41% reduce 0%

18/08/12 04:01:29 INFO mapreduce.Job: map 42% reduce 0%

18/08/12 04:01:38 INFO mapreduce.Job: map 43% reduce 0%

18/08/12 04:01:49 INFO mapreduce.Job: map 44% reduce 0%

18/08/12 04:01:59 INFO mapreduce.Job: map 45% reduce 0%

18/08/12 04:02:16 INFO mapreduce.Job: map 46% reduce 0%

18/08/12 04:02:22 INFO mapreduce.Job: map 47% reduce 0%

18/08/12 04:02:36 INFO mapreduce.Job: map 48% reduce 0%

18/08/12 04:02:57 INFO mapreduce.Job: map 49% reduce 0%

18/08/12 04:03:06 INFO mapreduce.Job: map 50% reduce 0%

18/08/12 04:03:15 INFO mapreduce.Job: map 51% reduce 0%

18/08/12 04:03:27 INFO mapreduce.Job: map 52% reduce 0%

18/08/12 04:03:44 INFO mapreduce.Job: map 53% reduce 0%

18/08/12 04:03:55 INFO mapreduce.Job: map 54% reduce 0%

18/08/12 04:04:05 INFO mapreduce.Job: map 55% reduce 0%

18/08/12 04:04:24 INFO mapreduce.Job: map 56% reduce 0%

18/08/12 04:04:29 INFO mapreduce.Job: map 57% reduce 0%

18/08/12 04:04:44 INFO mapreduce.Job: map 58% reduce 0%

18/08/12 04:04:51 INFO mapreduce.Job: map 59% reduce 0%

18/08/12 04:05:09 INFO mapreduce.Job: map 60% reduce 0%

18/08/12 04:05:21 INFO mapreduce.Job: map 61% reduce 0%

18/08/12 04:05:32 INFO mapreduce.Job: map 62% reduce 0%

18/08/12 04:05:44 INFO mapreduce.Job: map 63% reduce 0%

18/08/12 04:05:55 INFO mapreduce.Job: map 64% reduce 0%

18/08/12 04:06:13 INFO mapreduce.Job: map 65% reduce 0%

18/08/12 04:06:22 INFO mapreduce.Job: map 66% reduce 0%

18/08/12 04:06:34 INFO mapreduce.Job: map 67% reduce 0%

18/08/12 04:06:45 INFO mapreduce.Job: map 68% reduce 0%

18/08/12 04:06:56 INFO mapreduce.Job: map 69% reduce 0%

18/08/12 04:07:14 INFO mapreduce.Job: map 70% reduce 0%

18/08/12 04:07:21 INFO mapreduce.Job: map 71% reduce 0%

18/08/12 04:07:34 INFO mapreduce.Job: map 72% reduce 0%

18/08/12 04:07:40 INFO mapreduce.Job: map 73% reduce 0%

18/08/12 04:07:55 INFO mapreduce.Job: map 74% reduce 0%

18/08/12 04:08:11 INFO mapreduce.Job: map 75% reduce 0%

18/08/12 04:08:23 INFO mapreduce.Job: map 76% reduce 0%

18/08/12 04:08:36 INFO mapreduce.Job: map 77% reduce 0%

18/08/12 04:08:44 INFO mapreduce.Job: map 78% reduce 0%

18/08/12 04:09:02 INFO mapreduce.Job: map 79% reduce 0%

18/08/12 04:09:04 INFO mapreduce.Job: map 80% reduce 0%

18/08/12 04:09:22 INFO mapreduce.Job: map 81% reduce 0%

18/08/12 04:09:33 INFO mapreduce.Job: map 82% reduce 0%

18/08/12 04:09:41 INFO mapreduce.Job: map 82% reduce 4%

18/08/12 04:09:53 INFO mapreduce.Job: map 83% reduce 4%

18/08/12 04:09:58 INFO mapreduce.Job: map 84% reduce 4%

18/08/12 04:10:14 INFO mapreduce.Job: map 85% reduce 4%

18/08/12 04:10:30 INFO mapreduce.Job: map 86% reduce 4%

18/08/12 04:10:45 INFO mapreduce.Job: map 87% reduce 4%

18/08/12 04:11:01 INFO mapreduce.Job: map 88% reduce 4%

18/08/12 04:11:16 INFO mapreduce.Job: map 89% reduce 4%

18/08/12 04:11:32 INFO mapreduce.Job: map 90% reduce 4%

18/08/12 04:11:48 INFO mapreduce.Job: map 91% reduce 4%

18/08/12 04:12:04 INFO mapreduce.Job: map 92% reduce 4%

18/08/12 04:12:09 INFO mapreduce.Job: map 93% reduce 4%

18/08/12 04:12:25 INFO mapreduce.Job: map 94% reduce 4%

18/08/12 04:12:41 INFO mapreduce.Job: map 95% reduce 4%

18/08/12 04:12:43 INFO mapreduce.Job: map 95% reduce 5%

18/08/12 04:12:59 INFO mapreduce.Job: map 96% reduce 5%

18/08/12 04:13:10 INFO mapreduce.Job: map 97% reduce 5%

18/08/12 04:13:25 INFO mapreduce.Job: map 98% reduce 5%

18/08/12 04:13:41 INFO mapreduce.Job: map 99% reduce 5%

18/08/12 04:13:53 INFO mapreduce.Job: map 100% reduce 5%

18/08/12 04:13:56 INFO mapreduce.Job: map 100% reduce 14%

18/08/12 04:14:05 INFO mapreduce.Job: map 100% reduce 29%

18/08/12 04:14:06 INFO mapreduce.Job: map 100% reduce 43%

18/08/12 04:14:10 INFO mapreduce.Job: map 100% reduce 57%

18/08/12 04:14:18 INFO mapreduce.Job: map 100% reduce 71%

18/08/12 04:14:20 INFO mapreduce.Job: map 100% reduce 100%

18/08/12 04:14:25 INFO mapreduce.Job: Job job_1534045789091_0001 completed successfully

18/08/12 04:14:25 INFO mapreduce.Job: Counters: 54

    File System Counters

        FILE: Number of bytes read=27768102

FILE: Number of bytes written=116133357

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=18128

HDFS: Number of bytes written=0

HDFS: Number of read operations=176

HDFS: Number of large read operations=0

HDFS: Number of write operations=0

S3A: Number of bytes read=47087654647

S3A: Number of bytes written=32526830

S3A: Number of read operations=429

S3A: Number of large read operations=0

S3A: Number of write operations=4099

Job Counters

Launched map tasks=176

Launched reduce tasks=7

Rack-local map tasks=176

Total time spent by all maps in occupied slots (ms)=6581868

Total time spent by all reduces in occupied slots (ms)=685228

Total time spent by all map tasks (ms)=3290934

Total time spent by all reduce tasks (ms)=342614

Total vcore-milliseconds taken by all map tasks=3290934

Total vcore-milliseconds taken by all reduce tasks=342614

Total megabyte-milliseconds taken by all map tasks=6739832832

Total megabyte-milliseconds taken by all reduce tasks=701673472

Map-Reduce Framework

Map input records=702782657

Map output records=48826811

Map output bytes=1171819221

Map output materialized bytes=60843389

Input split bytes=18128

Combine input records=48826811

Combine output records=5349334

Reduce input groups=1464694

Reduce shuffle bytes=60843389

Reduce input records=5349334

Reduce output records=1464694

Spilled Records=10698668

Shuffled Maps =1232

Failed Shuffles=0

Merged Map outputs=1232

GC time elapsed (ms)=62322

CPU time spent (ms)=2242400

Physical memory (bytes) snapshot=130424139776

Virtual memory (bytes) snapshot=331297660928

Total committed heap usage (bytes)=140611420160

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=47087654647

File Output Format Counters

Bytes Written=32526830

Total Input Record : 702782657

Total Output Record : 1464694

MapReduce Job is Success

18/08/12 04:14:25 INFO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system...

18/08/12 04:14:25 INFO impl.MetricsSystemImpl: s3a-file-system metrics system stopped.

18/08/12 04:14:25 INFO impl.MetricsSystemImpl: s3a-file-system metrics system shutdown complete.

Create directory (/home/ec2-user/scriptsaavn/amitgoeloutput) for further processing

Copy MapReduce output files from s3 to local directory /home/ec2-user/scriptsaavn/amitgoeloutput

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00000 to amitgoeloutput/part-r-00000

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00006 to amitgoeloutput/part-r-00006

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00003 to amitgoeloutput/part-r-00003

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00002 to amitgoeloutput/part-r-00002

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00001 to amitgoeloutput/part-r-00001

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00004 to amitgoeloutput/part-r-00004

download: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00005 to amitgoeloutput/part-r-00005

Copy Saavn Treding list from s3 to local directory /home/ec2-user/scriptsaavn/amitgoeloutput

download: s3://mapreduce-project-bde/trending_data_daily.csv to amitgoeloutput/trending_data_daily.csv

Start further processing

Processing over

Lets try to find matching songs

Number of songs matching for 25-DEC-2017 : 68

Number of songs matching for 26-DEC-2017 : 69

Number of songs matching for 27-DEC-2017 : 70

Number of songs matching for 28-DEC-2017 : 72

Number of songs matching for 29-DEC-2017 : 72

Number of songs matching for 30-DEC-2017 : 73

Number of songs matching for 31-DEC-2017 : 73

Removing MapReduce output from s3 as not needed there anymore

delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00002
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00005
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00001
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00003
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00006
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00004
delete: s3://amitgoelc2bdesaavnproject/mroutput/part-r-00000
delete: s3://amitgoelc2bdesaavnproject/mroutput/_SUCCESS

Total time taken approximately (in minutes): 21

Thank you!!!

[ec2-user@ip-172-31-91-95 scriptsaavn]$