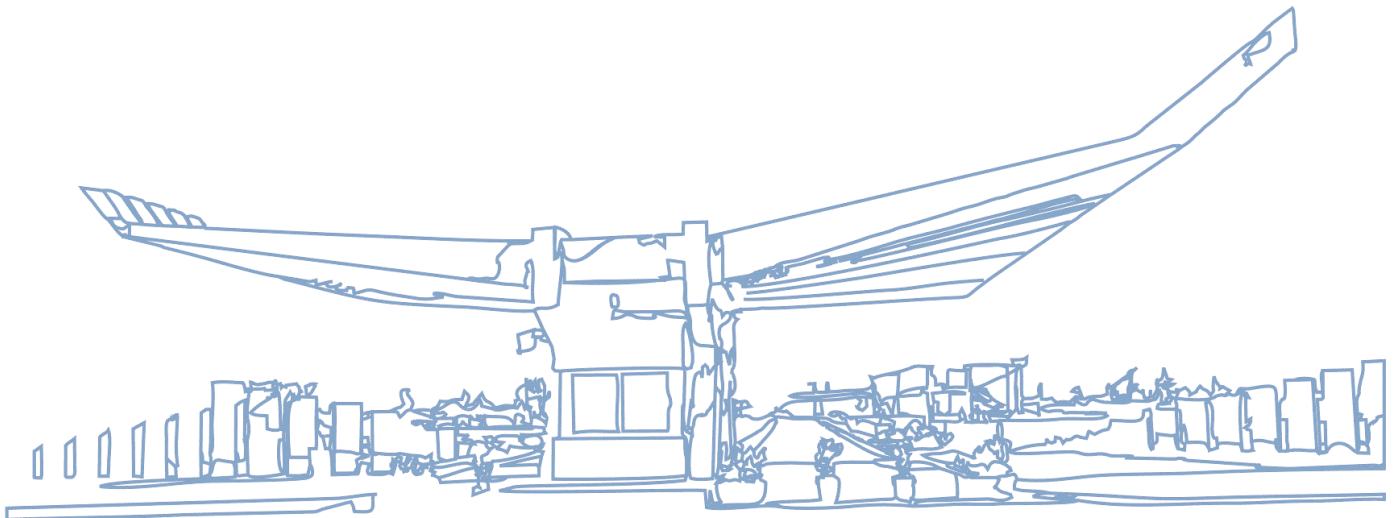


CEN 571 – Data Mining

Data Analysis with MapReduce and (Pig or Hive)

Project Documentation



PREPARED:
Baftjar TABAKU

31.05.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

1. Dataset

The dataset was taken by Kaggle.com



IMDb movies extensive dataset
81k+ movies and 175k+ cast members scraped from IMDb

Stefano Leone • updated 6 months ago (Version 1)

Sun Nov 24 2019 21:59:17 GMT+0100 (Central European Standard Time)

Data Tasks Kernels (6) Discussion Activity Metadata Download (164 MB) New Notebook

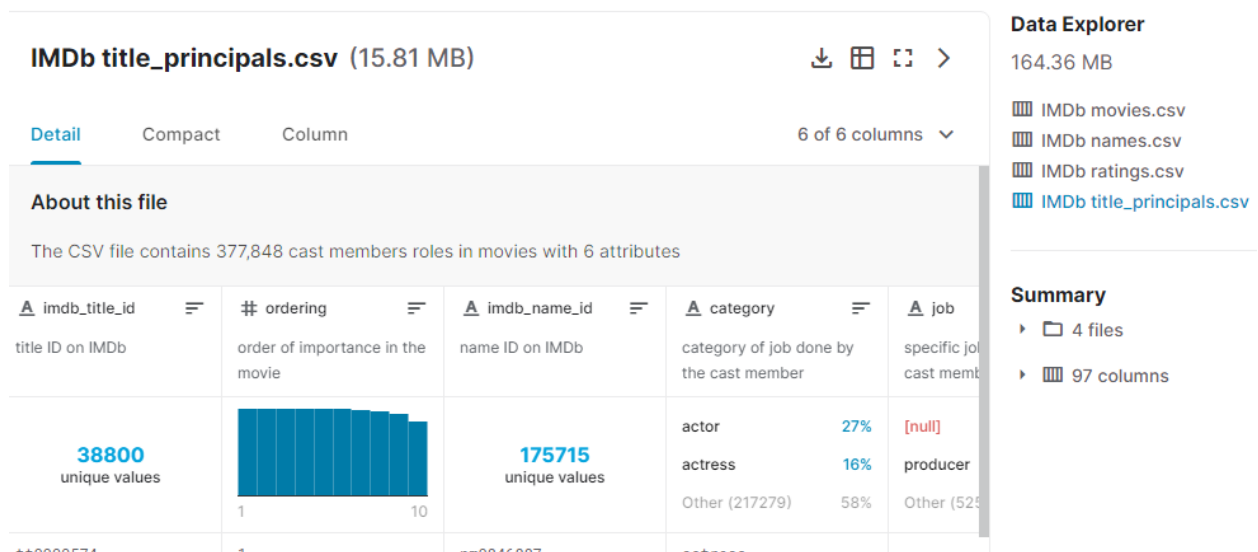
Usability 10.0 **License** CC0: Public Domain **Tags** computing, arts and entertainment, internet, film, popular culture

Description

Context

IMDb is the most popular movie website and it combines movie plot description. Metasore ratinas. critic and user ratinas and reviews. release dates.

With a size of 168 MB, the latest one, of 6 months, composed of 4 tables, by Stefano Leone. All in CSV format as shown below.




IMDb title_principals.csv (15.81 MB)

Detail Compact Column 6 of 6 columns

About this file

The CSV file contains 377,848 cast members roles in movies with 6 attributes

imdb_title_id	ordering	imdb_name_id	category	job
title ID on IMDb	order of importance in the movie	name ID on IMDb	category of job done by the cast member	specific job of cast member
38800 unique values		175715 unique values	actor 27% actress 16% Other (217279) 58%	[null] producer Other (525)

Data Explorer
164.36 MB

- IMDb movies.csv
- IMDb names.csv
- IMDb ratings.csv
- IMDb title_principals.csv

Summary

- 4 files
- 97 columns

2. Cleaning the unwanted data, the data selection according to the projects goals.

Some redundant features will be removed, and data will be processed using Map Reduce, with corresponding code and jar files.

Removing some columns was used the Microsoft Excel, where the data was displayed better and modified.

Ex: According to my goals, I don't need a movie Description, cast list, reviews numbers from users and anything to do with the price, or the user's that rate professions, spouses number and so on.

We also delete the cast data from dataset.

	K	L	M	N	O	P	Q
	writer	production_company	avg_vote	votes	budget	usa_gross_income	worldwide_gross_income
1	Charles Tait	J. and N. Tait	6.1	537	\$2,250		
2	Urban Gad, Gebhard Schützler-Perasini	Fotorama	5.9	171			
3	Victorien Sardou	Helen Gardner Picture	5.2	420	\$45,000		
4	Dante Alighieri	Milano Film	7	2019			
5	Gene Gauntier	Kalem Company	5.7	438			
6	Norbert Falk, Hanns Kräpely	Projektions-AG Union	6.8	709			
7	Henryk Sienkiewicz, Enrico Guazzoni	Societ� Italiana Cines	6.2	241	ITL 45000		
8	Aristide Demetriade, Petre Liciu	Societatea Filmului de	6.7	187	ROL 400000		
9	James Keane, William Shakespeare	Le Film d'Art	5.5	211	\$30,000		
10	Axel Garde, Gerhart Hauptmann	Nordisk Film	6.7	310			
11	Marcel Allain, Louis Feuillade	Soci�t� des Etabliss	7	1853			

From all data of 4 tables, I reduced it to 2 and removed the unnecessary features for all of them.

Note: I could have done them in MapReduce too by just ignoring their position, but data was very large and hard to manage according to my goals.

3. Writing the map reduce code

By analyzing the data, for each movie, it was on my interest goal to see the difference of rating between males and females, I modified a map reduce code according to my dataset and what I want. A jar file was also exported for later use.

```
import java.io.IOException;
import java.util.*;
```

```

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class TotalOpinionDifference {
    public static class MyMapper extends Mapper<LongWritable, Text, Text,
DoubleWritable> {

        private Text imdb_title_id = new Text();
        private DoubleWritable opinion_difference = new DoubleWritable(1);

        public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {
            String line = value.toString();
            String[] tokens = line.split(",");
            {

                try {
                    float males_allages_avg_vote =
Float.parseFloat(tokens[5]);
                    float females_allages_avg_vote =
Float.parseFloat(tokens[6]);

                    float avg_difference = Math.abs(males_allages_avg_vote-
females_allages_avg_vote); //difference rating

                    imdb_title_id.set(tokens[0]);
                    opinion_difference.set(avg_difference);

                    context.write(imdb_title_id, opinion_difference);

                } catch (Exception ex) {
                    System.out.println(imdb_title_id);
                    System.out.println(ex.toString());
                }

            }
        }

        public static class MyReducer extends Reducer<Text, DoubleWritable, Text,
DoubleWritable> {
            public void reduce(Text key, Iterable<DoubleWritable> values, Context
context)
                throws IOException, InterruptedException {

                double sum = 0;
                int count = 0;

                for (DoubleWritable val : values) {
                    sum = sum + val.get();
                    count++;
                }
                double rate_dif_avg = 0;
                rate_dif_avg = sum / count;
                context.write(key, new DoubleWritable(rate_dif_avg));
            }
        }
    }
}

```

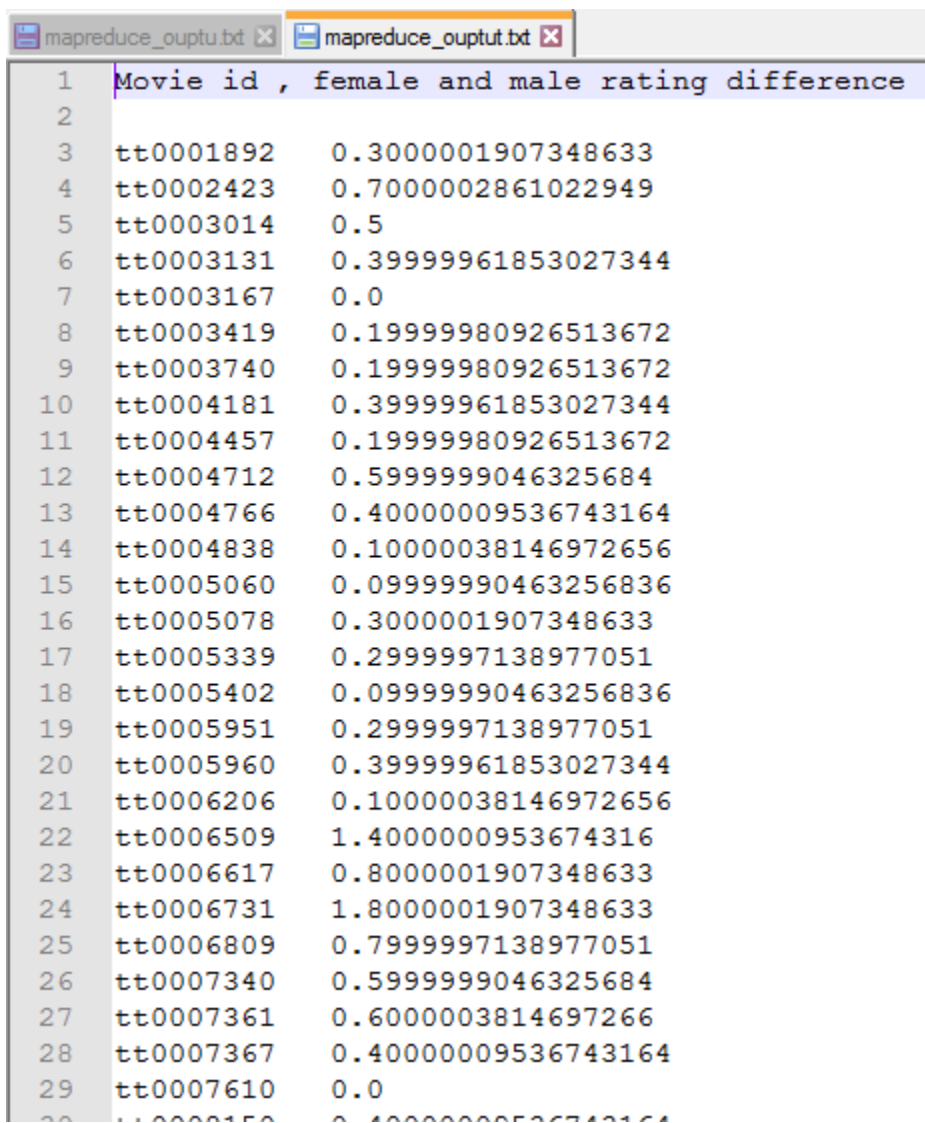
```

    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "TotalOpinionDifference");
        job.setJarByClass(TotalOpinionDifference.class);
        job.setMapperClass(MyMapper.class);
        job.setReducerClass(MyReducer.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(DoubleWritable.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
    }
}

```

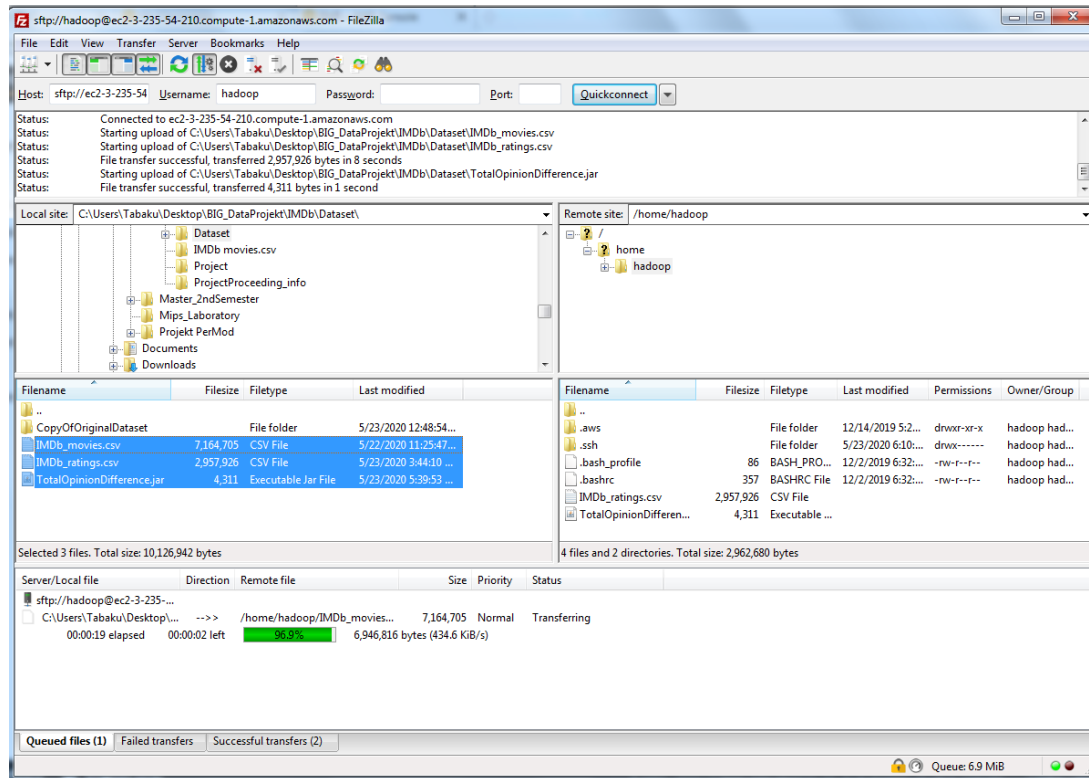
Output part,



Movie id	female and male rating difference
tt0001892	0.3000001907348633
tt0002423	0.7000002861022949
tt0003014	0.5
tt0003131	0.39999961853027344
tt0003167	0.0
tt0003419	0.19999980926513672
tt0003740	0.19999980926513672
tt0004181	0.39999961853027344
tt0004457	0.19999980926513672
tt0004712	0.5999999046325684
tt0004766	0.40000009536743164
tt0004838	0.10000038146972656
tt0005060	0.09999990463256836
tt0005078	0.3000001907348633
tt0005339	0.2999997138977051
tt0005402	0.09999990463256836
tt0005951	0.2999997138977051
tt0005960	0.39999961853027344
tt0006206	0.10000038146972656
tt0006509	1.4000000953674316
tt0006617	0.8000001907348633
tt0006731	1.8000001907348633
tt0006809	0.7999997138977051
tt0007340	0.5999999046325684
tt0007361	0.6000003814697266
tt0007367	0.40000009536743164
tt0007610	0.0

4. Uploading the data into the cluster and processing calculations.

The data tables.csv were successfully added to the AWS cluster for further analyzing, using MapReduce and Hive.



```

Bytes Written=2147355
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hadoop hadoop          0 2020-05-23 16:17 /exam
drwxrwxrwt - hdfs hadoop            0 2020-05-23 16:11 /tmp
drwxr-xr-x - hdfs hadoop            0 2020-05-23 16:11 /user
drwxr-xr-x - hdfs hadoop            0 2020-05-23 16:11 /var
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/
Found 2 items
drwxr-xr-x - hadoop hadoop          0 2020-05-23 16:18 /exam/input
drwxr-xr-x - hadoop hadoop          0 2020-05-23 16:21 /exam/output
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output
ls: '/exam/output': No such file or directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/outputtu
ls: '/exam/outputtu': No such file or directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/outputtut
ls: '/exam/outputtut': No such file or directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output
Found 1 items
drwxr-xr-x - hadoop hadoop          0 2020-05-23 16:21 /exam/output/results
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -cat /exam/output/results
cat: '/exam/output/results': Is a directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output/results
Found 4 items
-rw-r--r-- 1 hadoop hadoop          0 2020-05-23 16:21 /exam/output/results/_SUCCESS
-rw-r--r-- 1 hadoop hadoop    715174 2020-05-23 16:21 /exam/output/results/part-r-00000
-rw-r--r-- 1 hadoop hadoop    716987 2020-05-23 16:21 /exam/output/results/part-r-00001
-rw-r--r-- 1 hadoop hadoop    715194 2020-05-23 16:21 /exam/output/results/part-r-00002
[hadoop@ip-172-31-2-188 ~]$

```

```
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output
ls: '/exam/output': No such file or directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output
Found 1 items
drwxr-xr-x - hadoop hadoop 0 2020-05-23 16:21 /exam/output/results
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -cat /exam/output/results
cat: '/exam/output/results': Is a directory
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -ls /exam/output/results
Found 4 items
-rw-r--r-- 1 hadoop hadoop 0 2020-05-23 16:21 /exam/output/results/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 715174 2020-05-23 16:21 /exam/output/results/part-r-00000
-rw-r--r-- 1 hadoop hadoop 716987 2020-05-23 16:21 /exam/output/results/part-r-00001
-rw-r--r-- 1 hadoop hadoop 715194 2020-05-23 16:21 /exam/output/results/part-r-00002
[hadoop@ip-172-31-2-188 ~]$ hadoop fs -copyToLocal /exam/output/results
[hadoop@ip-172-31-2-188 ~]$ ls -a
. . .aws .bash_profile .bashrc IMDb_movies.csv IMDb_ratings.csv results .ssh TotalOpinionDifference.jar
[hadoop@ip-172-31-2-188 ~]$
```

```
hadoop@ip-172-31-2-188:~$
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
[hadoop@ip-172-31-2-188 ~]$ hadoop jar TotalOpinionDifference.jar /exam/input/IMDb_ratings.csv /exam/output/results
20/05/23 16:21:18 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-2-188.ec2.internal:172.31.2.188:8032
20/05/23 16:21:18 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
20/05/23 16:21:19 INFO input.FileInputFormat: Total input files to process : 1
20/05/23 16:21:19 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
20/05/23 16:21:19 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 5f788d5e8f90539ee331702c753fa250727128f4]
20/05/23 16:21:19 INFO mapreduce.JobSubmitter: number of splits=1
20/05/23 16:21:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1590250282427_0001
20/05/23 16:21:19 INFO impl.YarnClientImpl: Submitted application application_1590250282427_0001
20/05/23 16:21:19 INFO mapreduce.Job: The url to track the job: http://ip-172-31-2-188.ec2.internal:20888/proxy/application_1590250282427_0001/
20/05/23 16:21:19 INFO mapreduce.Job: Running job: job_1590250282427_0001
20/05/23 16:21:26 INFO mapreduce.Job: Job job_1590250282427_0001 running in uber mode : false
20/05/23 16:21:26 INFO mapreduce.Job: map 0% reduce 0%
20/05/23 16:21:32 INFO mapreduce.Job: map 100% reduce 0%
20/05/23 16:21:36 INFO mapreduce.Job: map 100% reduce 33%
20/05/23 16:21:38 INFO mapreduce.Job: map 100% reduce 100%
20/05/23 16:21:38 INFO mapreduce.Job: Job job_1590250282427_0001 completed successfully
20/05/23 16:21:38 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=654409
FILE: Number of bytes written=1986528
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2958059
HDFS: Number of bytes written=2149355
HDFS: Number of read operations=12
HDFS: Number of large read operations=0
HDFS: Number of write operations=6
Job Counters
Launched map tasks=1
Launched reduce tasks=3
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=310560
Total time spent by all reduces in occupied slots (ms)=1795200
Total time spent by all map tasks (ms)=3235
Total time spent by all reduce tasks (ms)=9350
```

Project: IMDb > MR

Status: File transfer successful, transferred 115,114 bytes in 5 seconds

Status: Starting download of /home/hadoop/results/part-r-00001

Status: File transfer successful, transferred 715,194 bytes in 3 seconds

Status: Starting download of /home/hadoop/results/_SUCCESS

Local site: C:\Users\Tabaku\Desktop\BIG_DataProject\IMDb\Dataset\

Remote site: /home/hadoop

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Group
..					
CopyOfOriginalDataset		File folder	5/23/2020 12:48:54...		
IMDb_movies.csv	7,164,705	CSV File	5/22/2020 11:25:47...		
IMDb_ratings.csv	2,957,926	CSV File	5/23/2020 3:44:10...		
TotalOpinionDifference.jar	4,311	Executable Jar File	5/23/2020 5:39:53...		

Selected 3 files. Total size: 10,126,942 bytes

Selected 1 directory.

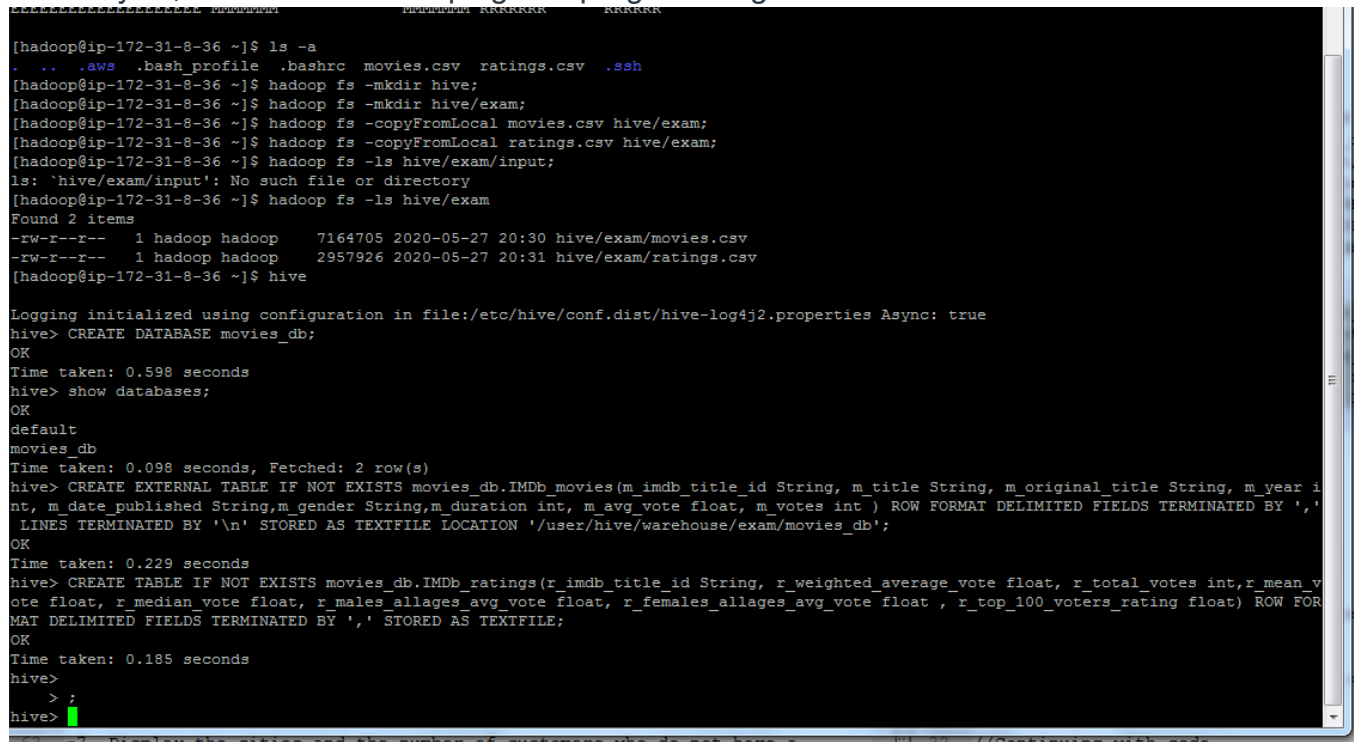
Server/Local file	Direction	Remote file	Size	Priority	Status
sftp://hadoop@ec2-3-235-...	<--	/home/hadoop/results/part-r-...	716,987	Normal	Transferring
C:\Users\Tabaku\Desktop\...	<--	/home/hadoop/results/_SUC...	0	Normal	Transferring

6:26 PM
5/23/2020

The MapReduce output will be stored in the output folder, according to the ordering of the submission, below are the corresponding output screenshots.

5. Hive used commands and their output

I choose HIVE because the characteristic of my project goals are the reports and data analysis, rather than developing and programming.



```

[had00p@ip-172-31-8-36 ~]$ ls -a
. . . .aws .bash_profile .bashrc movies.csv ratings.csv .ssh
[had00p@ip-172-31-8-36 ~]$ hadoop fs -mkdir hive;
[had00p@ip-172-31-8-36 ~]$ hadoop fs -mkdir hive/exam;
[had00p@ip-172-31-8-36 ~]$ hadoop fs -copyFromLocal movies.csv hive/exam;
[had00p@ip-172-31-8-36 ~]$ hadoop fs -copyFromLocal ratings.csv hive/exam;
[had00p@ip-172-31-8-36 ~]$ hadoop fs -ls hive/exam/input;
ls: 'hive/exam/input': No such file or directory
[had00p@ip-172-31-8-36 ~]$ hadoop fs -ls hive/exam
Found 2 items
-rw-r--r-- 1 hadoop hadoop 7164705 2020-05-27 20:30 hive/exam/movies.csv
-rw-r--r-- 1 hadoop hadoop 2957926 2020-05-27 20:31 hive/exam/ratings.csv
[had00p@ip-172-31-8-36 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> CREATE DATABASE movies_db;
OK
Time taken: 0.598 seconds
hive> show databases;
OK
default
movies_db
Time taken: 0.098 seconds, Fetched: 2 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS movies_db.IMDb_movies(m_imdb_title_id String, m_title String, m_original_title String, m_year int, m_date_published String, m_gender String, m_duration int, m_avg_vote float, m_votes int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '/user/hive/warehouse/exam/movies_db';
OK
Time taken: 0.229 seconds
hive> CREATE TABLE IF NOT EXISTS movies_db.IMDb_ratings(r_imdb_title_id String, r_weighted_average_vote float, r_total_votes int, r_mean_vote float, r_median_vote float, r_males_allages_avg_vote float, r_females_allages_avg_vote float, r_top_100_voters_rating float) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.185 seconds
hive>
> ;
hive>

```

The data successfully moved into the Hadoop and the hive databases and tables are successfully created.

Describing the data structure of the table

```
imdb_movies
imdb_ratings
Time taken: 0.031 seconds, Fetched: 2 row(s)
hive> describe IMDb_movies;
OK
m_imdb_title_id      string
m_title              string
m_original_title      string
m_year               int
m_date_published      string
m_gender              string
m_duration            int
m_avg_vote            float
m_votes              int
Time taken: 0.081 seconds, Fetched: 9 row(s)
hive> describe IMDb_ratings;
OK
r_imdb_title_id      string
r_weighted_average_vote float
r_total_votes         int
r_mean_vote           float
r_median_vote         float
r_males_allages_avg_vote float
r_females_allages_avg_vote float
r_top_100_voters_rating float
Time taken: 0.032 seconds, Fetched: 8 row(s)
hive>
```

Populating the data into the respective tables

```
Time taken: 0.032 seconds, Fetched: 8 row(s)
hive> LOAD DATA LOCAL INPATH './movies.csv' OVERWRITE INTO TABLE IMDb_movies;
Loading data to table movies_db.imdb_movies
OK
Time taken: 0.972 seconds
hive> LOAD DATA LOCAL INPATH './ratings.csv' OVERWRITE INTO TABLE IMDb_ratings;
Loading data to table movies_db.imdb_ratings
OK
Time taken: 0.75 seconds
hive>
```

Selecting all the movie data.

```
hadoop@ip-172-31-8-36:~$
tt9789686      Un rubio      Un rubio      2019      2/26/2019      "Drama NULL 108.0 7
tt9793334      Athiran Athiran 2019      4/26/2019      "Mystery NULL 135.0 6
tt9795368      Tobol Tobol 2019      2/21/2019      "Action NULL 108
tt9799984      Falaknuma Das Falaknuma Das 2019      5/31/2019      "Action NULL 145.0 6
tt9799992      Krasue: Inhuman Kiss Krasue: Inhuman Kiss 2019      3/14/2019      "Drama NULL NULL 122
tt9801736      "Hvitur hvitur dagur" NULL hvitur dagur" 2019      NULL NULL 109
tt9805504      Otryv Otryv 2019      2/14/2019      Thriller 85 4.1 134
tt9806192      J'ai perdu mon corps J'ai perdu mon corps 2019      11/6/2019      "Animation NULL 81.0 7
tt9811374      Inséparables Inséparables 2019      9/4/2019      Comedy 94 6.0 101
tt9815714      The Hard Way The Hard Way 2019      3/5/2019      Action 92 4.8 1734
tt9816970      Majarar Nimrooz: Radde Khoon Majarar Nimrooz: Radde Khoon 2019      9/25/2019      "Crime NULL NULL 125
tt9817018      Shabi Ke Mah Kamel Shod Shabi Ke Mah Kamel Shod 2019      6/5/2019      "Crime NULL NULL 137
tt9817070      Metri Shesh Va Nim Metri Shesh Va Nim 2019      3/17/2019      "Crime NULL NULL 131
tt9817300      15-Aug 15-Aug 2019      3/29/2019      Drama 124 5.9 128
tt9818102      Yowis Ben 2 Yowis Ben 2 2019      3/14/2019      "Comedy NULL 109.0 7
tt9820594      Misteri Dilaila Misteri Dilaila 2019      2/21/2019      "Horror NULL 82.0 5
tt9831136      The Banana Splits Movie The Banana Splits Movie 2019      8/13/2019      "Comedy NULL NULL 89
tt9838372      Fei fen shu nü Fei fen shu nü 2019      4/4/2019      Drama 108 4.9 167
tt9839040      Murphy's Law: Ghanooone Morfi Murphy's Law: Ghanooone Morfi 2019      1/2/2019      "Action NULL NULL 98
tt9840958      Love Struck Sick Love Struck Sick 2019      8/8/2019      Romance 92 7.0 284
tt9845398      Fin de siglo Fin de siglo 2019      8/16/2019      Drama 84 7.2 242
tt9850064      Kaijû no kodomo Kaijû no kodomo 2019      6/7/2019      "Animation NULL NULL 110
tt9855990      Nightmare Tenant Nightmare Tenant 2018      12/5/2018      Thriller 90 5.5 129
tt9860728      Falling Inn Love Falling Inn Love 2019      8/29/2019      "Comedy NULL 98.0 5
tt9860860      Abduction 101 Abduction 101 2019      1/1/2019      Horror 77 2.4 146
tt9861522      Ali Ali 2019      3/22/2019      Drama 110 5.0 119
tt9866208      Beyond the Line Beyond the Line 2019      6/4/2019      War 78 3.6 150
tt9866700      Paranormal Investigation Paranormal Investigation 2018      12/1/2018      "Horror NULL 92.0 3
tt9870726      Gholamreza Takhti Gholamreza Takhti 2019      3/16/2019      "Biography NULL NULL 113
tt9872556      Momenti di trascurabile felicità Momenti di trascurabile felicità 2019      3/14/2019      Comedy 93
.3 413
tt9875852      Demovoy Demovoy 2019      4/11/2019      "Comedy NULL NULL 90
tt9878242      Subharathri Subharathri 2019      7/6/2019      "Drama NULL 130.0 6
tt9880982      Dulce Familia Dulce Familia 2019      5/10/2019      Comedy 101 4.6 171
tt9894098      Sathru Sathru 2019      3/7/2019      Thriller 129 6.1 163
tt9894394      Upin & Ipin: Keris Siamang Tunggal Upin & Ipin: Keris Siamang Tunggal 2019      3/21/2019      Animation
00 8.0 367
tt9896916      The Pilgrim's Progress The Pilgrim's Progress 2019      4/18/2019      "Animation NULL NULL 108
tt9899880      Columbus Columbus 2018      12/5/2018      "Comedy NULL 82.0 4
tt9900782      Kaithi Kaithi 2019      10/25/2019      "Action NULL 145.0 8
tt9903716      Jessie Jessie 2019      3/15/2019      "Horror NULL 106.0 7
tt9905412      Ottam Ottam 2019      3/8/2019      Drama 120 7.8 510
tt9905462      Pengalila Pengalila 2019      3/8/2019      Drama 111 8.4 604
tt9911774      Padmavyuhathile Abhimanyu Padmavyuhathile Abhimanyu 2019      3/8/2019      Drama 130 8.4 369
tt9914286      Sokagin Çocuklari Sokagin Çocuklari 2019      3/15/2019      "Drama NULL 98.0 7
Time taken: 1.201 seconds, Fetched: 81274 row(s)
hive>
```

Sorting the movies according to the duration,

```
hive> SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration FROM IMDB_movies SORT BY m_duration ASC;
Query ID = hadoop_20200527205930_56f1a947-3edd-4149-bd27-3928189efb60
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0002)

-----
VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FAILED      KILLED
-----
Map 1 ..... container      SUCCEEDED      1            1            0            0            0            0
Reducer 2     container      RUNNING        1            0            1            0            0            0
-----
/VERTICES: 01/02 [=====]>>-----] 50% ELAPSED TIME: 4.66 s
-----
```

hadoop@ip-172-31-8-36:~

tt0385878	Monica la mitraille	Drama	2004	125		
tt0385842	El Lobo	Thriller	2004	125		
tt0054271	Scent of Mystery	Mystery	1960	125		
tt0384537	Silent Hill	Horror	2006	125		
tt0303858	Le collectionneur	Thriller		2002	125	
tt0101362	Al-aragoz	Drama	1989	125		
tt0177022	Mazhki vremena	Drama	1977	125		
tt8718580	Eghantham	Drama	2018	125		
tt0177016	Mulher Objeto	Drama	1981	125		
tt0382197	Nienasycenie	Drama	2003	125		
tt0100581	La setta	Horror	1991	125		
tt1830645	Igillena Maluwo	Drama	2011	125		
tt1579943	Danger	Thriller	2005	125		
tt5104080	Surga Yang Tak Dirindukan	Drama		2015	125	
tt2912144	Dast-neveshtehaa nemisoosand	Drama		2013	125	
tt1345728	Nankyoku ryôrinin	Comedy	2009	125		
tt8590992	Paper Boy	Romance	2018	125		
tt0189796	Necista krv	Drama	1996	125		
tt4811564	Las furias	Drama	2016	125		
tt6263618	In den Gängen	Drama	2018	125		
tt0311072	En la ciudad sin límites	Drama		2002	125	
tt9007142	The Dark Side of Life: Mumbai City	Drama		2018	125	
tt4691804	Mohican kokyô ni kaeru	Comedy	2016	125		
tt0356019	Seraa fil Nil	Drama	1959	125		
tt5125576	Rajwade and Sons	Drama	2015	125		
tt0089057	Le due vite di Mattia Pascal	Comedy	1985	125		
tt0040288	Le diable boiteux	History	1948	125		
tt4623812	Luo man di ke xiao wang shi	Drama	2016	125		
tt8581230	B.A. Pass 2	Drama	2017	125		
tt0431814	Hoteru binasu	Drama	2004	125		
tt1662634	Bentein men misr	Drama	2010	125		
tt0118687	Le bassin de J.W.	Comedy	1997	125		
tt1417032	Amarufi: Megami no hôshû	Drama		2009	125	
tt7201744	Philophobia	Drama	2019	125		
tt1844016	Rathinirvedam	Drama	2011	125		
tt1860318	Tantei wa bar ni iru	Drama	2011	125		
tt2389344	The Mistress	Drama	2012	125		
tt6289452	Kötü Çocuk	Romance	2017	125		
tt0211126	Zakhm	Drama	1998	125		
tt0077326	The Children of Sanchez	Drama	1978	126		
tt0119608	Mandragora	Drama	1997	126		
tt0395972	North Country	Drama	2005	126		
tt0049746	Shabab emraa	Drama	1956	126		
tt4964580	Mae Bia	Drama	2015	126		
tt0081632	Tian yun shan chuan qi	Drama	1980	126		
tt7104984	La sombra de la ley	Thriller		2018	126	
tt0089275	Himatsuri	Drama	1985	126		

Selecting all the movies from year 2016, and analyzing according to their duration.

```
hive> SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration FROM IMDB_movies where m_year>=2016 SORT BY m_duration ASC;
Query ID = hadoop_20200527211126_10b9c986-bd33-4f8c-8708-2602e27360bf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0003)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1         container   INITED    1       0           0        1       0       0
Reducer 2     container   INITED    1       0           0        1       0       0
-----
VERTICES: 00/02  [>>-----] 0%  ELAPSED TIME: 2.84 s
-----
```

```
hadoop@ip-172-31-8-36:~
tt5934788      The Ouija Possession      Horror  2016    95
tt4508542      Nacida para ganar         Comedy  2016    95
tt0451201      Sludge Horror             2017    95
tt5599702      Kizkaçıran                Comedy  2016    95
tt4523910      Chicuarotes               Drama   2019    95
tt5338174      Les naufragés             Comedy  2016    95
tt4529162      Even Lovers Get the Blues Drama    2016    95
tt4901356      Forever Young              Comedy  2016    95
tt5485466      Kharms Biography          2017    95
tt6255230      Clown Motel: Spirits Arise Horror    2019    95
tt5189440      Tour de France            Drama   2016    95
tt9185066      Human Capital              Drama   2019    95
tt8506874      Komola Rocket             Drama   2018    95
tt2769828      Peelers Horror            2016    95
tt4562260      O Amor é Lindo ... Porque Sim! Comedy  2016    95
tt8898544      Khalaweas                 Family  2018    95
tt8898648      Yabani Asly               Family  2017    95
tt5176486      Grandma's House           Drama   2016    95
tt6143422      Kadvi Hawa                Drama   2017    95
tt3984432      Ghost Source Zero         Sci-Fi  2017    95
tt6259946      2 Cool 2 Be 4gotten       Drama   2016    95
tt8065772      Benyamin Biang Kerok      Comedy  2018    95
tt9144672      Kafalar Karisik           Comedy  2018    95
tt7106968      Ride                      Drama   2018    95
tt9378320      Watchman                  Thriller 2019    95
tt5517708      Lola Pater                Drama   2017    95
tt6803924      L'intrusa                 Drama   2017    95
tt6680312      Who We Are Now            Drama   2017    95
tt8696440      Sibel                    Drama   2018    95
tt6809094      Eaten by Lions            Comedy  2018    95
tt4894198      Death Pool                Thriller 2017    95
tt5793184      The Spy and the Poet      Drama   2016    95
tt6081632      Marie-Francine            Comedy  2017    95
tt6515420      La deuxième étoile        Comedy  2017    95
tt6736198      Solteras                  Comedy  2019    95
tt8305806      The Wretched              Horror  2019    95
tt5029602      Brice 3 Comedy            2016    95
tt7535756      Bölük                    Drama   2017    95
tt8495600      Borç                      Drama   2018    95
tt8098548      Larguées                  Comedy  2018    95
tt4788744      Unhinged                  Thriller 2018    95
tt8308496      Córka trenera            Drama   2018    95
tt8795582      Posledice                 Drama   2018    95
tt5033018      In the Radiant City       Drama   2016    95
tt6446550      The Parting Glass         Drama   2018    95
tt1949605      Who Gets the Dog?         Comedy  2016    95
tt5725284      Patriotai                 Comedy  2016    95
```

Counting the movies from 2016 until today,

```
hive> SELECT m_imdb_title_id, m_title, m_gender, m_year, m_duration, COUNT(m_imdb_title_id) as total FROM IMdb_movies where m_year>=2016 GROUP BY
m_imdb_title_id, m_title, m_gender, m_year, m_duration ;
Query ID= hadoop_20200527211646_34574e0d-907f-4831-80d8-1089e44f578b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0003)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1		container	INITED	1	0	0	1	0	0
Reducer 2		container	INITED	1	0	0	1	0	0

VERTICES: 00/02 [>>-----] 0% ELAPSED TIME: 1.41 s

```
hadoop@ip-172-31-8-36:~
tt9783778      Adventures of Aladdin      "Adventure      2019      NULL      1
tt9789440      Ninu Veedani Needanu Nene      Comedy      2019      123      1
tt9789670      Porinju Mariam Jose      "Action      2019      NULL      1
tt9789686      Un rubio      "Drama      2019      NULL      1
tt9793334      Athiran "Mystery      2019      NULL      1
tt9795368      Tobol "Action      2019      NULL      1
tt9799984      Falaknuma Das      "Action      2019      NULL      1
tt9799992      Krasue: Inhuman Kiss      "Drama      2019      NULL      1
tt9805504      Otryv Thriller      2019      85      1
tt9806192      J'ai perdu mon corps      "Animation      2019      NULL      1
tt9811374      Inséparables      Comedy      2019      94      1
tt9815714      The Hard Way      Action      2019      92      1
tt9816970      Majarary Nimrooz: Radde Khoon      "Crime      2019      NULL      1
tt9817018      Shabi Ke Mah Kamel Shod      "Crime      2019      NULL      1
tt9817070      Metri Shesh Va Nim      "Crime      2019      NULL      1
tt9817300      15-Aug Drama      2019      124      1
tt9818102      Yowis Ben 2      "Comedy      2019      NULL      1
tt9820594      Misteri Dilaila "Horror      2019      NULL      1
tt9831136      The Banana Splits Movie      "Comedy      2019      NULL      1
tt9838372      Fei fen shu nü Drama      2019      108      1
tt9839040      Murphy's Law: Ghanoone Morfi      "Action      2019      NULL      1
tt9840958      Love Struck Sick      Romance      2019      92      1
tt9845398      Fin de siglo Drama      2019      84      1
tt9850064      Kaijû no kodomo "Animation      2019      NULL      1
tt9855990      Nightmare Tenant      Thriller      2018      90      1
tt9860728      Falling Inn Love      "Comedy      2019      NULL      1
tt9860860      Abduction 101 Horror      2019      77      1
tt9861522      Ali Drama      2019      110      1
tt9866208      Beyond the Line War      2019      78      1
tt9866700      Paranormal Investigation      "Horror      2018      NULL      1
tt9870726      Gholamreza Takhti      "Biography      2019      NULL      1
tt9872556      Momenti di trascurabile felicità      Comedy      2019      93      1
tt9875852      Domovoy "Comedy      2019      NULL      1
tt9878242      Subharathri      "Drama      2019      NULL      1
tt9880982      Dulce Familia Comedy      2019      101      1
tt9894098      Sathru Thriller      2019      129      1
tt9894394      Upin & Ipin: Keris Siamang Tunggal      Animation      2019      100      1
tt9896916      The Pilgrim's Progress      "Animation      2019      NULL      1
tt9899880      Columbus      "Comedy      2018      NULL      1
tt9900782      Kaithi "Action      2019      NULL      1
tt9903716      Jessie "Horror      2019      NULL      1
tt9905412      Ottam Drama      2019      120      1
tt9905462      Pengalila Drama      2019      111      1
tt9911774      Padmavyuhathile Abhimanyu      Drama      2019      130      1
tt9914286      Sokagin Çocuklari      "Drama      2019      NULL      1
Time taken: 5.24 seconds, Fetched: 10438 row(s)
hive>
```

Counting the movies for year 2015, 2016, 2017, and 2018.

2015 movies counted,

```

tt6055162      Oglan Evi: 1      Comedy  2015    105    1
tt6108002      Tabula Rosa        "Drama  2015    NULL    1
tt6159026      Gatao      Action  2015    105    1
tt6217766      Marge Mahi        Drama    2015    101    1
tt6239540      Apocalypse Will Not Happen      Drama    2015    79    1
tt7323332      Risunâ      Drama    2015    111    1
tt8216024      Nun-gil "Drama  2015    NULL    1
Time taken: 5.405 seconds, Fetched: 2863 row(s)
hive>

```

2016 movies counted,

```

hive> SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration , COUNT(m_imdb_title_id) as total FROM IMDB_movies where m_year==2016 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
Query ID = hadoop_20200527211945_a7447aa8-57d2-4a48-b3d9-11b165667c20
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0003)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1          container  INITED      1         0         0         1         0         0
Reducer 2      container  INITED      1         0         0         1         0         0
-----
VERTICES: 00/02  [>>-----] 0%  ELAPSED TIME: 1.01 s
-----

tt6857268      El paseo 4      Comedy  2016    95    1
tt6892250      Absurd Accident "Comedy  2016    NULL    1
tt7032958      Mashina lyubvi  "Drama  2016    NULL    1
tt7318290      Xiaozhen's Story      Musical  2016    112    1
tt7600396      Gospel Movie: Who Is He That Has Returned      Drama    2016    164    1
tt8297300      Fractured      "Mystery  2016    NULL    1
tt8434682      Nasil Yani      Comedy  2016    96    1
tt9011132      Diamond Sword  "Action  2016    NULL    1
Time taken: 5.254 seconds, Fetched: 2997 row(s)
hive>

```

2017 movies counted,

```

> SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration , COUNT(m_imdb_title_id) as total FROM IMDB_movies where m_year=2017 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
Query ID = hadoop_20200527212359_624a3651-3851-488f-9633-f65317ccf944
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0003)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1          container  INITED      1         0         0         1         0         0
Reducer 2      container  INITED      1         0         0         1         0         0
-----
VERTICES: 00/02  [>>-----] 0%  ELAPSED TIME: 1.01 s
-----

tt8335776      SI made tette demokristiani      Comedy  2017    89    1
tt8387464      House Shark      "Action  2017    NULL    1
tt8581230      B.A. Pass 2      Drama    2017    125    1
tt8898648      Yabani Asly      Family   2017    95    1
tt8967900      Shomareh 17 Soheila      "Drama  2017    NULL    1
tt9555284      Bekar Bekir      Comedy  2017    94    1
Time taken: 4.905 seconds, Fetched: 3056 row(s)
hive>

```


2018 movies counted,

```

tt9647330      G Saat "Animation      2018      NULL      1
tt9652322      Chief Daddy      Comedy  2018      99        1
tt9742422      Xue bao "Action  2018      NULL      1
tt9855990      Nightmare Tenant  Thriller 2018      90        1
tt9866700      Paranormal Investigation "Horror 2018      NULL      1
tt9899880      Columbus "Comedy  2018      NULL      1
Time taken: 5.97 seconds, Fetched: 2839 row(s)
hive>

```

The movies total average rating, according to the years (2016, 2017, and 2018).

The year 2016.

```

hive> SELECT AVG (r_weighted_average_vote) as total FROM IMDb_movies JOIN IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
Query ID = hadoop_20200527223454_68d44e2-dee5-49a0-879a-e823d9d880bc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0007)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 5.43 s
-----
OK
5.694461128733179
Time taken: 5.923 seconds, Fetched: 1 row(s)
hive>

```

The year 2017.

```

hive> SELECT AVG (r_weighted_average_vote) as total FROM IMDb_movies JOIN IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
Query ID = hadoop_20200527223549_2ba18dca-5c9a-4cb1-a909-c3c22b135fa2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0007)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 6.05 s
-----
OK
5.753632198102499
Time taken: 6.472 seconds, Fetched: 1 row(s)
hive>

```

The year 2018.

```

hive> SELECT AVG (r_weighted_average_vote) as total FROM IMDb_movies JOIN IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
Query ID = hadoop_20200527223649_c941da41-0291-4b38-b31b-b4dad39fec4a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0007)

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 6.20 s
-----
OK
5.768510034404283
Time taken: 6.651 seconds, Fetched: 1 row(s)
hive>

```

Males and females rating for the years 2015,2016,2017,2018, Year 2015, males and females rating average

```
hive> SELECT AVG(r_males_allages_avg_vote) as total FROM IMdb_movies JOIN IMdb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2015;
Query ID = hadoop_20200527224730_f515c401-428a-4b84-bd79-513871c11c09
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====]>>] 100% ELAPSED TIME: 6.27 s
OK
5.543870066039179
Time taken: 11.965 seconds, Fetched: 1 row(s)
hive> SELECT AVG(r_females_allages_avg_vote) as total FROM IMdb_movies JOIN IMdb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2015;
Query ID = hadoop_20200527224744_9e3df23a-7725-4a8d-8604-1317c29a76d8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====]>>] 100% ELAPSED TIME: 1.22 s
OK
5.80720028386421
Time taken: 1.666 seconds, Fetched: 1 row(s)
hive>
```

Year 2016, males and females,

```
hive> SELECT AVG(r_males_allages_avg_vote) as total FROM IMdb_movies JOIN IMdb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
Query ID = hadoop_20200527225005_0587c8c8-0daa-4c69-9206-e1ceccc6c52bd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====]>>] 100% ELAPSED TIME: 7.94 s
OK
5.586052720690712
Time taken: 8.405 seconds, Fetched: 1 row(s)
hive>
> SELECT AVG(r_females_allages_avg_vote) as total FROM IMdb_movies JOIN IMdb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
Query ID = hadoop_20200527225013_207195c9-a438-4a80-8364-c2d2d5c5a797
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 03/03 [=====]>>] 100% ELAPSED TIME: 1.81 s
OK
5.909729187602544
Time taken: 2.228 seconds, Fetched: 1 row(s)
hive>
```


Year 2017, males and females.

```
hive> SELECT AVG(r_males.allages_avg_vote) as total FROM IMDB_movies JOIN IMDB_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
Query ID = hadoop_20200527225210_835d36d1-9217-402e-94cb-a81327d4ce2d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 5.98 s
-----
OK
5.60844240704288
Time taken: 6.438 seconds, Fetched: 1 row(s)
hive>
> SELECT AVG(r_females.allages_avg_vote) as total FROM IMDB_movies JOIN IMDB_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
Query ID = hadoop_20200527225226_c2a4a549-e92d-4f85-b305-a9a08f3e645a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 5.50 s
-----
OK
5.983753688523623
Time taken: 5.932 seconds, Fetched: 1 row(s)
hive>
```

Year 2018, males and females votes average.

```
hive> SELECT AVG(r_males.allages_avg_vote) as total FROM IMDB_movies JOIN IMDB_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
Query ID = hadoop_20200527225349_1610ffb2-24b5-445e-b3b6-c6eefb02ab70
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 5.62 s
-----
OK
5.577069391495495
Time taken: 6.054 seconds, Fetched: 1 row(s)
hive> SELECT AVG(r_females.allages_avg_vote) as total FROM IMDB_movies JOIN IMDB_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
Query ID = hadoop_20200527225401_b2ab1545-9a92-4267-8499-362612827043
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590611198514_0008)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  1      1      0      0      0      0
Map 3 ..... container  SUCCEEDED  1      1      0      0      0      0
Reducer 2 ..... container  SUCCEEDED  1      1      0      0      0      0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 2.04 s
-----
OK
6.006104448856227
Time taken: 2.519 seconds, Fetched: 1 row(s)
hive>
```

Note: These queries do not show all kind of operations, ex. Count, and so on, the queries are performed according to the goals of the project only.

Queries

```

hadoop fs -mkdir hive;
hadoop fs -mkdir hive/exam;
hadoop fs -mkdir hive/exam/input;

hadoop fs -copyFromLocal movies.csv hive/exam/input;

hadoop fs -copyFromLocal ratings.csv hive/exam/input;

hadoop fs -ls hive/exam/input;

hive

CREATE DATABASE movies_db;

show databases;

CREATE EXTERNAL TABLE IF NOT EXISTS movies_db.IMDb_movies(m_imdb_title_id
String, m_title String, m_original_title String, m_year int, m_date_published
String,m_gender String,m_duration int, m_avg_vote float, m_votes int ) ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS
TEXTFILE LOCATION '/user/hive/warehouse/exam/movies_db';

CREATE TABLE IF NOT EXISTS movies_db.IMDb_ratings(r_imdb_title_id String,
r_weighted_average_vote float, r_total_votes int,r_mean_vote float,
r_median_vote float, r_males_allages_avg_vote float,
r_females_allages_avg_vote float , r_top_100_voters_rating float) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;

use movies_db;
show tables;

describe IMDb_movies;
describe IMDb_ratings;

//load data
LOAD DATA LOCAL INPATH './movies.csv' OVERWRITE INTO TABLE IMDb_movies;

LOAD DATA LOCAL INPATH './ratings.csv' OVERWRITE INTO TABLE IMDb_ratings;

//Continuing with code

-Sorting the movies according to lengths

SELECT * FROM IMDb_movies;
SELECT * FROM IMDb_ratings;

//Sorting all the movies according to the time
SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration FROM IMDb_movies
SORT BY m_duration ASC;

//selecting movies from year 2016 to current

SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration FROM IMDb_movies
where m_year>=2016 SORT BY m_duration ASC;

//counting movies from year 2016 and above

```

```
SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration ,
COUNT(m_imdb_title_id) as total FROM IMDb_movies where m_year>=2016 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
```

```
//counting movies of year 2016
```

```
SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration ,
COUNT(m_imdb_title_id) as total FROM IMDb_movies where m_year=2016 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
```

```
//counting movies of year 2017
```

```
SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration ,
COUNT(m_imdb_title_id) as total FROM IMDb_movies where m_year=2017 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
```

```
//counting movies of year 2018
```

```
SELECT m_imdb_title_id, m_title,m_gender, m_year ,m_duration ,
COUNT(m_imdb_title_id) as total FROM IMDb_movies where m_year=2018 GROUP BY
m_imdb_title_id, m_title,m_gender, m_year ,m_duration ;
```

```
//Calculating the rating average of the movies for the year 2016, 2017,2018
```

```
//year 2016 average rating
```

```
SELECT AVG (r_weighted_average_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
```

```
//year 2017 average rating
```

```
SELECT AVG (r_weighted_average_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
```

```
//year 2018 average rating
```

```
SELECT AVG(r_weighted_average_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
```

```
//Total Males and females rating average for years 2015,2016,2017,2018
```

```
//females
```

```
SELECT AVG(r_females_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2015;
```

```
//males
```

```
SELECT AVG(r_males_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2015;
```

```
//2016
```

```
//males
```

```
SELECT AVG(r_males_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
```

```
//females
```

```
SELECT AVG(r_females_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2016;
```

```
//2017
```

```
//males
```

```
SELECT AVG(r_males_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
```

```
//females
```

```
SELECT AVG(r_females_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2017;
```

```
//2018
//mles
SELECT AVG(r_males_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
//females
SELECT AVG(r_females_allages_avg_vote) as total FROM IMDb_movies JOIN
IMDb_ratings ON m_imdb_title_id=r_imdb_title_id where m_year=2018;
```