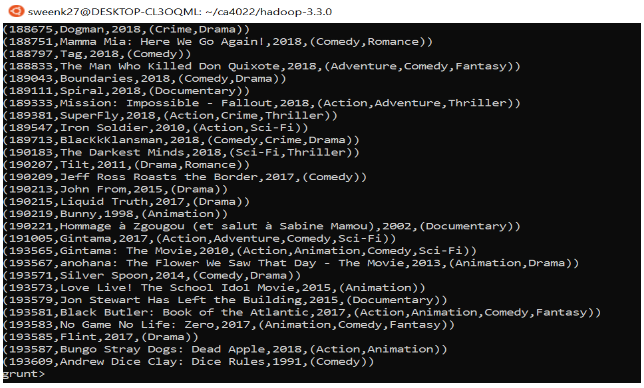
CA4022 Assignment 1 – Kian Sweeney

- ***GitHub link: https://github.com/kiansweeney11/ca4022-assignment1***

Data Cleaning

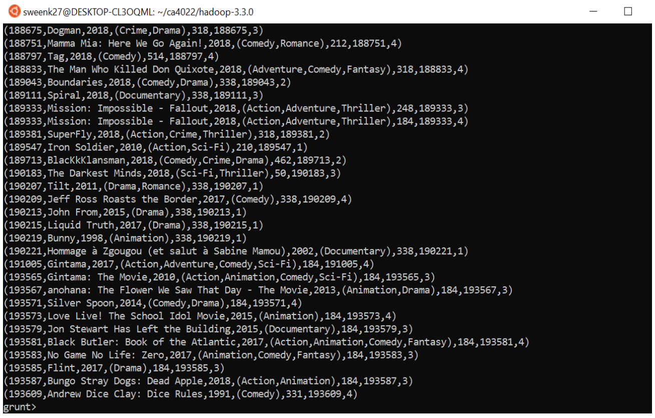
We use the index notation for each heading here ($, with the first index being 0 and so forth) to pick our title heading. We split it up accordingly. Every year is denoted as " (2017)" in the title so we subtract the number of characters present here (7) from the length of the entire tuple. This will give us our title. Then we use regex_extract to find a sequence of digits after the title. This will only target the string after the title of the movie due to our previous regex_extract on the title tuple. Lastly, we split our genres tuple on "|" to split up the genres.

Our next task is to look at merging the ratings file with our cleaned movies. Personally, I didn't see the need to utilise the timestamp heading in this file so I dropped it. This was due to the fact nothing was asked about timestamps in our questions.



```
(188675,Dogman,2018,(Crime,Drama))
(188751,Mamma Mia: Here We Go Again!,2018,(Comedy,Romance))
(188797,Tag,2018,(Comedy))
(188833,The Man Who Killed Don Quixote,2018,(Adventure,Comedy,Fantasy))
(189043,Boundaries,2018,(Comedy,Drama))
(189111,Spiral,2018,(Documentary))
(189333,Mission: Impossible - Fallout,2018,(Action,Adventure,Thriller))
(189381,SuperFly,2018,(Action,Crime,Thriller))
(189547,Iron Soldier,2010,(Action,Sci-Fi))
(189713,BlacKkKlansman,2018,(Comedy,Crime,Drama))
(190183,The Darkest Minds,2018,(Sci-Fi,Thriller))
(190207,Tilt,2011,(Drama,Romance))
(190209,Jeff Ross Roasts the Border,2017,(Comedy))
(190213,John From,2015,(Drama))
(190215,Liquid Truth,2017,(Drama))
(190219,Bunny,1998,(Animation))
(190221,Hommage à Zgougou (et salut à Sabine Mamou),2002,(Documentary))
(191005,Gintama,2017,(Action,Adventure,Comedy,Sci-Fi))
(193565,Gintama: The Movie,2010,(Action,Animation,Comedy,Sci-Fi))
(193567,anohana: The Flower We Saw That Day - The Movie,2013,(Animation,Drama))
(193571,Silver Spoon,2014,(Comedy,Drama))
(193573,Love Live! The School Idol Movie,2015,(Animation))
(193579,Jon Stewart Has Left the Building,2015,(Documentary))
(193581,Black Butler: Book of the Atlantic,2017,(Action,Animation,Comedy,Fantasy))
(193583,No Game No Life: Zero,2017,(Animation,Comedy,Fantasy))
(193585,Flint,2017,(Drama))
(193587,Bungo Stray Dogs: Dead Apple,2018,(Action,Animation))
(193609,Andrew Dice Clay: Dice Rules,1991,(Comedy))
grunt>
```

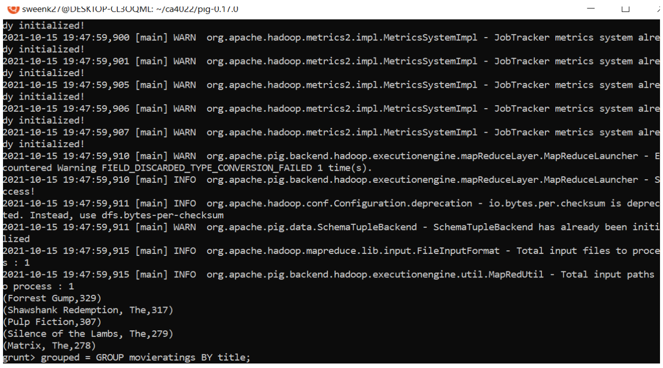
One limitation here was the duplicate movieId columns. I did try to remove this but I appeared to lose some output when I removed the column. As a result, I left this column in the output and denoted it as mov2, ignoring for it my Pig / Hive analysis.



```
(188675,Dogman,2018,(Crime,Drama),318,188675,3)
(188751,Mamma Mia: Here We Go Again!,2018,(Comedy,Romance),212,188751,4)
(188797,Tag,2018,(Comedy),514,188797,4)
(188833,The Man Who Killed Don Quixote,2018,(Adventure,Comedy,Fantasy),318,188833,4)
(189043,Boundaries,2018,(Comedy,Drama),338,189043,2)
(189111,Spiral,2018,(Documentary),338,189111,3)
(189333,Mission: Impossible - Fallout,2018,(Action,Adventure,Thriller),248,189333,3)
(189333,Mission: Impossible - Fallout,2018,(Action,Adventure,Thriller),184,189333,4)
(189381,SuperFly,2018,(Action,Crime,Thriller),318,189381,2)
(189547,Iron Soldier,2010,(Action,Sci-Fi),210,189547,1)
(189713,BlacKkKlansman,2018,(Comedy,Crime,Drama),462,189713,2)
(190183,The Darkest Minds,2018,(Sci-Fi,Thriller),50,190183,3)
(190207,Tilt,2011,(Drama,Romance),338,190207,1)
(190209,Jeff Ross Roasts the Border,2017,(Comedy),338,190209,4)
(190213,John From,2015,(Drama),338,190213,1)
(190215,Liquid Truth,2017,(Drama),338,190215,1)
(190219,Bunny,1998,(Animation),338,190219,1)
(190221,Hommage à Zgougou (et salut à Sabine Mamou),2002,(Documentary),338,190221,1)
(191005,Gintama,2017,(Action,Adventure,Comedy,Sci-Fi),184,191005,4)
(193565,Gintama: The Movie,2010,(Action,Animation,Comedy,Sci-Fi),184,193565,3)
(193567,anohana: The Flower We Saw That Day - The Movie,2013,(Animation,Drama),184,193567,3)
(193571,Silver Spoon,2014,(Comedy,Drama),184,193571,4)
(193573,Love Live! The School Idol Movie,2015,(Animation),184,193573,4)
(193579,Jon Stewart Has Left the Building,2015,(Documentary),184,193579,3)
(193581,Black Butler: Book of the Atlantic,2017,(Action,Animation,Comedy,Fantasy),184,193581,4)
(193583,No Game No Life: Zero,2017,(Animation,Comedy,Fantasy),184,193583,3)
(193585,Flint,2017,(Drama),184,193585,3)
(193587,Bungo Stray Dogs: Dead Apple,2018,(Action,Animation),184,193587,3)
(193609,Andrew Dice Clay: Dice Rules,1991,(Comedy),331,193609,4)
grunt>
```

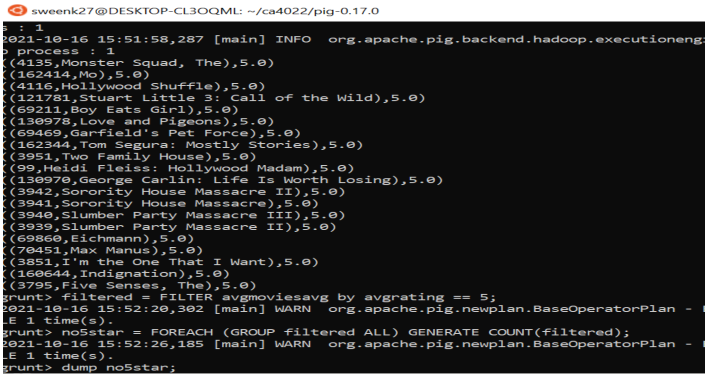
Pig Analysis

Q3Part1 - We see the movie with the greatest number of ratings here is Forrest Gump with 329 reviews.

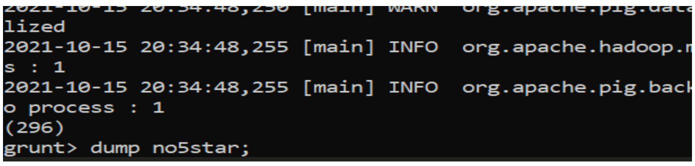


```
dy initialized!
2021-10-15 19:47:59,900 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
dy initialized!
2021-10-15 19:47:59,901 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
dy initialized!
2021-10-15 19:47:59,905 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
dy initialized!
2021-10-15 19:47:59,906 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
dy initialized!
2021-10-15 19:47:59,907 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system alre
dy initialized!
2021-10-15 19:47:59,910 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - E
countered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2021-10-15 19:47:59,910 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - S
ccess!
2021-10-15 19:47:59,911 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprec
ted. Instead, use dfs.bytes-per-checksum
2021-10-15 19:47:59,911 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initi
lized
2021-10-15 19:47:59,915 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to proce
s : 1
2021-10-15 19:47:59,915 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths 
o process : 1
(Forrest Gump,329)
(Shawshank Redemption, The,317)
(Pulp Fiction,307)
(Silence of the Lambs, The,279)
(Matrix, The,278)
grunt> grouped = GROUP movieratings BY title;
```

Q3Part2 - Check how many movies had a 5-star average, 296 movies had an average 5-star rating! I analysed this further using Hive later on looking at >= 4.0 average rating and how many ratings these 5.0 averages actually had.

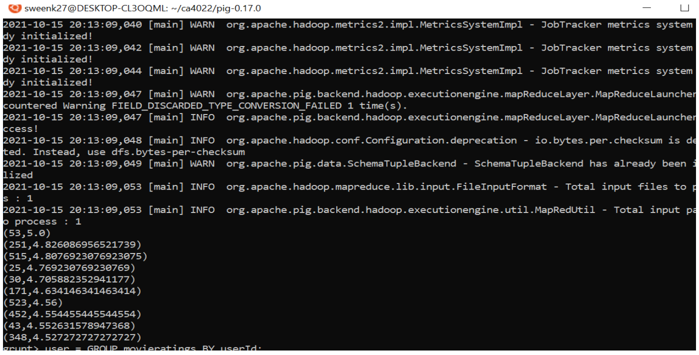

sweenk27@DESKTOP-CL3OQML: ~/ca4022/pig-0.17.0

```
s : 1
2021-10-16 15:51:58,287 [main] INFO  org.apache.pig.backend.hadoop.executioneng
o process : 1
((4135,Monster Squad, The),5.0)
((162414,Mo),5.0)
((4116,Hollywood Shuffle),5.0)
((121781,Stuart Little 3: Call of the Wild),5.0)
((69211,Boy Eats Girl),5.0)
((130978,Love and Pigeons),5.0)
((69469,Garfield's Pet Force),5.0)
((162344,Tom Segura: Mostly Stories),5.0)
((3951,Two Family House),5.0)
((99,Heidi Fleiss: Hollywood Madam),5.0)
((130970,George Carlin: Life Is Worth Losing),5.0)
((3942,Sorority House Massacre II),5.0)
((3941,Sorority House Massacre),5.0)
((3940,Slumber Party Massacre III),5.0)
((3939,Slumber Party Massacre II),5.0)
((69860,Eichmann),5.0)
((70451,Max Manus),5.0)
((3851,I'm the One That I Want),5.0)
((160644,Indignation),5.0)
((3795,Five Senses, The),5.0)
grunt> filtered = FILTER avgmoviesavg by avgrating == 5;
2021-10-16 15:52:20,302 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan -
LE 1 time(s).
grunt> no5star = FOREACH (GROUP filtered ALL) GENERATE COUNT(filtered);
2021-10-16 15:52:26,185 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan -
LE 1 time(s).
grunt> dump no5star;
```



```
2021-10-15 20:34:48,250 [main] WARN  org.apache.pig.data
lized
2021-10-15 20:34:48,255 [main] INFO  org.apache.hadoop.m
s : 1
2021-10-15 20:34:48,255 [main] INFO  org.apache.pig.back
o process : 1
(296)
grunt> dump no5star;
```

Q3Part3 - UserID 53 is the only user with a 5-star average rating here. We will expand on this later in Hive and confirm that this user did indeed have the highest average rating. It was interesting to note that a lot of these reviewers had very high average ratings (>= 4.5).

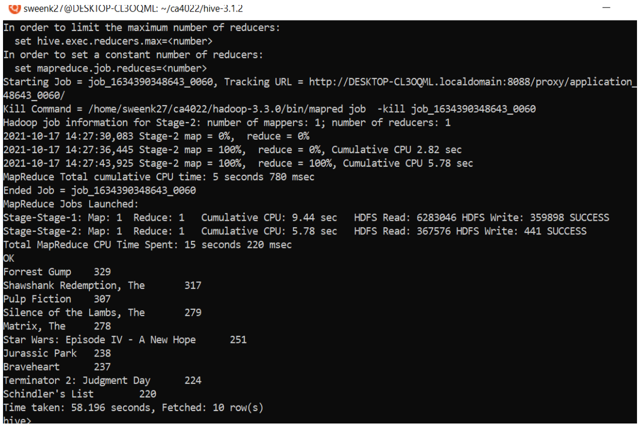

sweenk27@DESKTOP-CL3OQML: ~/ca4022/pig-0.17.0

```
2021-10-15 20:13:09,040 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
dy initialized!
2021-10-15 20:13:09,042 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
dy initialized!
2021-10-15 20:13:09,044 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system
dy initialized!
2021-10-15 20:13:09,047 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
countered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2021-10-15 20:13:09,047 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher
ccess!
2021-10-15 20:13:09,048 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de
ted. Instead, use dfs.bytes-per-checksum
2021-10-15 20:13:09,049 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i
lized
2021-10-15 20:13:09,053 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to p
s : 1
2021-10-15 20:13:09,053 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa
o process : 1
(53,5.0)
(251,4.826086956521739)
(515,4.8076923076923075)
(25,4.769230769230769)
(30,4.705882352941177)
(171,4.634146341463414)
(523,4.56)
(452,4.554455445544554)
(43,4.552631578947368)
(348,4.527272727272727)
grunt> user = GROUP movieratings BY userId;
```

Hive Queries

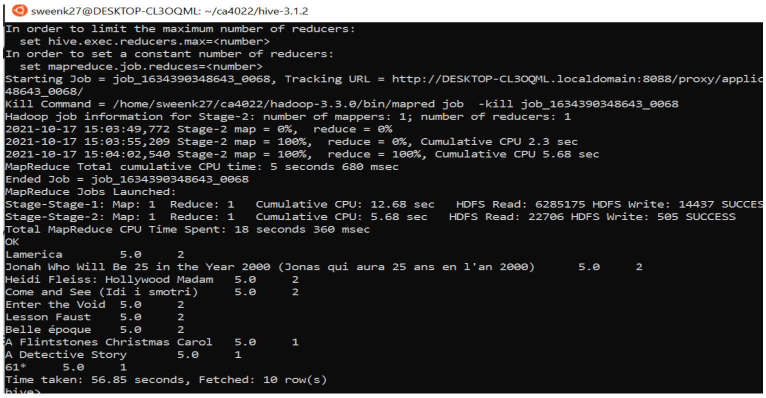
Originally, I just stored the file without delimiters. This meant when I tried to populate a table, I just got NULL values for each row. This is because the default separator in Hive is "^A" and because our merged file was not a CSV file, we could not delimit it using commas. This meant delimiting it using "|" instead. I had to go back to storing my merged file on Pig and specify delimiters and then this allowed me to successfully read in our data to a table in Hive. The table loaded in can be seen in the GitHub repository specified.

Q4Part1 – We can see our result here confirms our previous result in Pig. Forrest Gump does indeed have the highest number of ratings in 329. We use our GROUP BY function here which groups all the records in the result by our title column. I decide not to use movieID column in this as well as we only really need the title.


sweenk27@DESKTOP-CL3OQML: ~/ca4022/hive-3.1.2

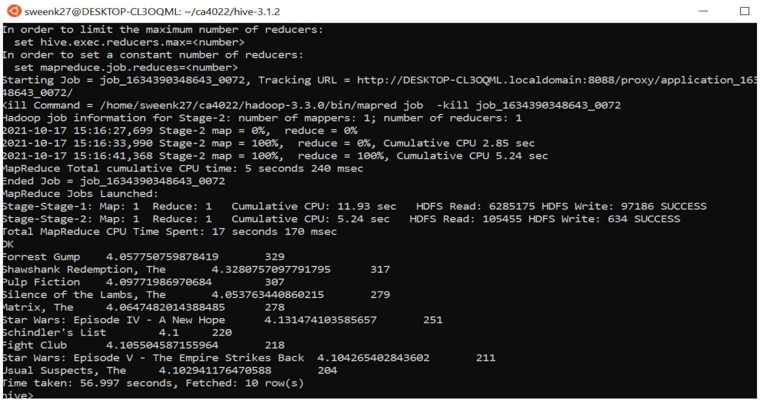
```
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634390348643_0060, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/application_
48643_0060/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634390348643_0060
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-17 14:27:30,083 Stage-2 map = 0%,  reduce = 0%
2021-10-17 14:27:36,445 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.82 sec
2021-10-17 14:27:43,925 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.78 sec
MapReduce Total cumulative CPU time: 5 seconds 780 msec
Ended Job = job_1634390348643_0060
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.44 sec   HDFS Read: 6283046 HDFS Write: 359898 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.78 sec   HDFS Read: 367576 HDFS Write: 441 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 220 msec
OK
Forrest Gump    329
Shawshank Redemption, The       317
Pulp Fiction    307
Silence of the Lambs, The       279
Matrix, The     278
Star Wars: Episode IV - A New Hope      251
Jurassic Park   238
Braveheart      237
Terminator 2: Judgment Day      224
Schindler's List        220
Time taken: 58.196 seconds, Fetched: 10 row(s)
hive>
```

Q4Part2 – Following on from our pig analysis we look at movies with high average ratings on different intervals. Here we focus on average rating >= 4.0 and 4.9. We can see any movie with an average rating of 5 has a very low number of reviews. This points towards these users having a keen interest in this film and the films being targeted to a specific audience (which these users appear to be part of). However, when we lower this value to average of 4 and order by the number of ratings, we can get a much clearer picture of the most liked films. The two moves with the greatest number of reviews have very high averages of >= 4.05. By ordering on number of reviews we and combining with our average rating we get a very complete view of the most liked movies in this dataset.



sweenk27@DESKTOP-CL3OQML: ~/ca4022/hive-3.1.2

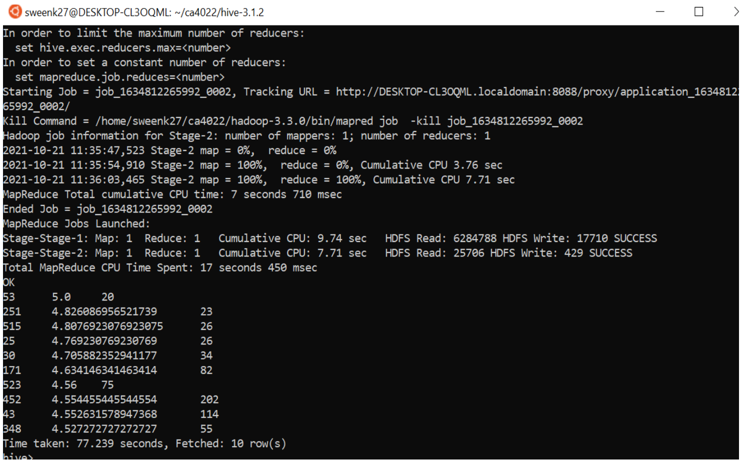
```
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634390348643_0068, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/applic
48643_0068/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634390348643_0068
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-17 15:03:49,772 Stage-2 map = 0%,  reduce = 0%
2021-10-17 15:03:55,209 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.3 sec
2021-10-17 15:04:02,540 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.68 sec
MapReduce Total cumulative CPU time: 5 seconds 680 msec
Ended Job = job_1634390348643_0068
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.68 sec   HDFS Read: 6285175 HDFS Write: 14437 SUCCES
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.68 sec   HDFS Read: 22706 HDFS Write: 505 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 360 msec
OK
Lamerica        5.0     2
Jonah Who Will Be 25 in the Year 2000 (Jonas qui aura 25 ans en l'an 2000)      5.0     2
Heidi Fleiss: Hollywood Madam   5.0     2
Come and See (Idi i smotri)     5.0     2
Enter the Void  5.0     2
Lesson Faust    5.0     2
Belle époque    5.0     2
A Flintstones Christmas Carol   5.0     1
A Detective Story       5.0     1
61*     5.0     1
Time taken: 56.85 seconds, Fetched: 10 row(s)
hive>
```



sweenk27@DESKTOP-CL3OQML: ~/ca4022/hive-3.1.2

```
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634390348643_0072, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/application_163
48643_0072/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634390348643_0072
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-17 15:16:27,699 Stage-2 map = 0%,  reduce = 0%
2021-10-17 15:16:33,990 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.85 sec
2021-10-17 15:16:41,368 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.24 sec
MapReduce Total cumulative CPU time: 5 seconds 240 msec
Ended Job = job_1634390348643_0072
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.93 sec   HDFS Read: 6285175 HDFS Write: 97186 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.24 sec   HDFS Read: 105455 HDFS Write: 634 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 170 msec
OK
Forrest Gump    4.057750759878419       329
Shawshank Redemption, The       4.3280757097791795      317
Pulp Fiction    4.09771986970684        307
Silence of the Lambs, The       4.053763440860215       279
Matrix, The     4.0647482014388485      278
Star Wars: Episode IV - A New Hope      4.131474103585657       251
Schindler's List        4.1     220
Fight Club      4.105504587155964       218
Star Wars: Episode V - The Empire Strikes Back  4.104265402843602       211
Usual Suspects, The     4.102941176470588       204
Time taken: 56.997 seconds, Fetched: 10 row(s)
hive>
```

Q4Part3 – We see this also confirms our earlier result in that userID 53 has the highest average rating. Here I added in the extra context of how many reviews these users gave. While none of them gave an insignificant number of reviews it was interesting to see users 452 and 43 maintain such a high average rating with so many reviews. These could easily be regarded as the most "generous" reviewers due to the number of reviews they gave and the high averages they had.

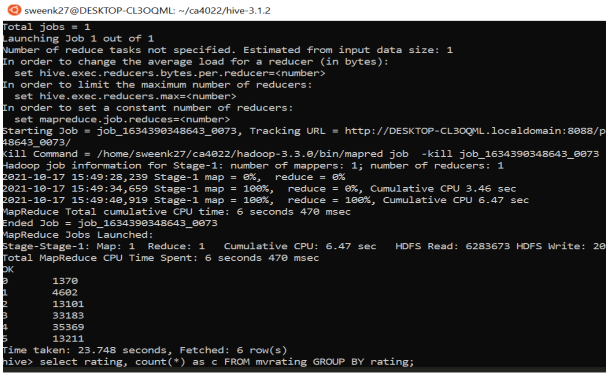


sweenk27@DESKTOP-CL3OQML: ~/ca4022/hive-3.1.2

```
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634812265992_0002, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/application_1634812
65992_0002/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634812265992_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-10-21 11:35:47,523 Stage-2 map = 0%,  reduce = 0%
2021-10-21 11:35:54,910 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.76 sec
2021-10-21 11:36:03,465 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.71 sec
MapReduce Total cumulative CPU time: 7 seconds 710 msec
Ended Job = job_1634812265992_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.74 sec   HDFS Read: 6284788 HDFS Write: 17710 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.71 sec   HDFS Read: 25706 HDFS Write: 429 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 450 msec
OK
53      5.0     20
251     4.826086956521739       23
515     4.8076923076923075      26
25      4.769230769230769       26
30      4.705882352941177       34
171     4.634146341463414       82
523     4.56    75
452     4.554455445544554       202
43      4.552631578947368       114
348     4.527272727272727       55
Time taken: 77.239 seconds, Fetched: 10 row(s)
hive>
```

Hive- Advanced Queries

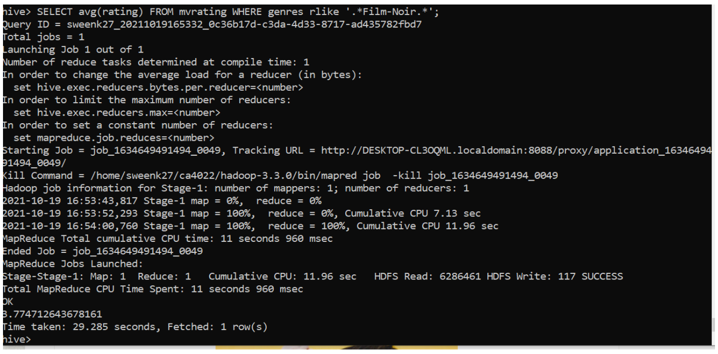
Q5Part1&Part2 – Again we use our GROUP by function to filter our results by rating. Count(*) refers to the count of all subjects related to the group by. We see a large majority of ratings are around the 3- and 4-star ratings. There is a similar level of 2- and 5-star ratings also with very few 0- and 1-star ratings handed. We then sort this on the number of ratings in descending order and see 4-star ratings are the most common. It will be interesting to see when we look at genres do the averages align with these ratings i.e., are the averages all within the 3-4 scale?



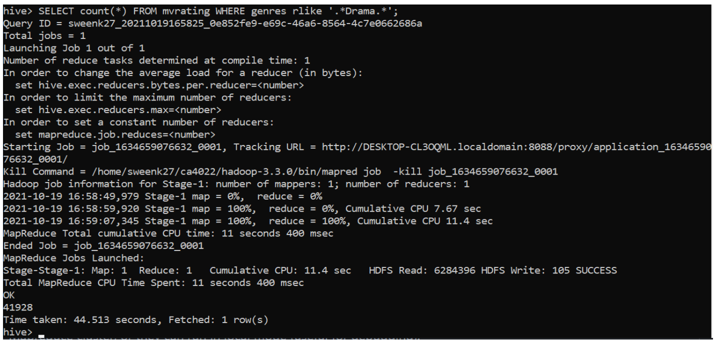
sweenk27@DESKTOP-CL3OQML: ~/ca4022/hive-3.1.2

```
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634390348643_0073, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/p
48643_0073/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634390348643_0073
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-17 15:49:28,239 Stage-1 map = 0%,  reduce = 0%
2021-10-17 15:49:34,659 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.46 sec
2021-10-17 15:49:40,919 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.47 sec
MapReduce Total cumulative CPU time: 6 seconds 470 msec
Ended Job = job_1634390348643_0073
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.47 sec   HDFS Read: 6283673 HDFS Write: 20
Total MapReduce CPU Time Spent: 6 seconds 470 msec
OK
0       1370
1       4602
2       13101
3       33183
4       35369
5       13211
Time taken: 23.748 seconds, Fetched: 6 row(s)
hive> select rating, count(*) as c FROM mvrating GROUP BY rating;
```

Q5Part3 – Our last part of the analysis looked at grouping ratings and number of reviews by genre. We see the genre with the most reviews is the “drama” genre. It has a comparatively mediocre average rating of ~3.5 in contrast to other genres but the sheer volume of ratings probably doesn’t help. This is in stark contrast to the ‘Film-Noir’ genre. It had the lowest number of ratings (870) but the largest average rating of ~3.77. This could potentially be attributed to the fact that with such a small reviewing audience these reviewers are more interested in this genre and thus more likely to give it a higher rating. And there is more than likely a much smaller collection of movies in this genre due to the smaller reviewing numbers hence adding to the theory that the reviewers are more interested in films of this genre and will give a higher rating as a result. We will look at this further in a Jupyter notebook.



```
hive> SELECT avg(rating) FROM mvrating WHERE genres rlike '.*Film-Noir.*';
Query ID = sweenk27_20211019165332_0c36b17d-c3da-4d33-8717-ad435782fbd7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634649491494_0049, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/application_16346494
91494_0049/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634649491494_0049
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-19 16:53:43,817 Stage-1 map = 0%,  reduce = 0%
2021-10-19 16:53:52,293 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.13 sec
2021-10-19 16:54:00,760 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.96 sec
MapReduce Total cumulative CPU time: 11 seconds 960 msec
Ended Job = job_1634649491494_0049
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.96 sec   HDFS Read: 6286461 HDFS Write: 117 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 960 msec
OK
3.774712643678161
Time taken: 29.285 seconds, Fetched: 1 row(s)
hive>
```



```
hive> SELECT count(*) FROM mvrating WHERE genres rlike '.*Drama.*';
Query ID = sweenk27_20211019165825_0e852fe9-e69c-46a6-8564-4c7e0662686a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634659076632_0001, Tracking URL = http://DESKTOP-CL3OQML.localdomain:8088/proxy/application_16346590
76632_0001/
Kill Command = /home/sweenk27/ca4022/hadoop-3.3.0/bin/mapred job  -kill job_1634659076632_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-19 16:58:49,979 Stage-1 map = 0%,  reduce = 0%
2021-10-19 16:58:59,920 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.67 sec
2021-10-19 16:59:07,345 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.4 sec
MapReduce Total cumulative CPU time: 11 seconds 400 msec
Ended Job = job_1634659076632_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.4 sec   HDFS Read: 6284396 HDFS Write: 105 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 400 msec
OK
41928
Time taken: 44.513 seconds, Fetched: 1 row(s)
hive>
```