COEN 242 Big Data

Project 1

Movie Reviews in MapReduce and Hive

Immanuel Amirtharaj
Jackson Beadle

May 21, 2018

Methodology

Hive

Importing Tables

One consideration we had to make when importing tables was making sure that the movie ratings were not integers. Therefore we imported the ratings as floats along with all other fields being integers. In addition, some of the movie titles contained commas causing Hive to incorrectly parse some lines in the CSV file. Our tables were defined with Hive SerDe properties to ignore these commas and to remove any double-quotes that padded the beginning or end of each data field.

Part 1 Query

```sql
SELECT COUNT(*) C, title
FROM imdb_bigdata17.movies_large M, imdb_bigdata17.reviews_large R
WHERE M.movieId = R.movieId
GROUP BY M.title
ORDER BY C ASC, title ASC;
```

Part 2 Query

```sql
SELECT title, AVG(rating) A, COUNT(*) reviews
FROM imdb_bigdata17.movies_large M, imdb_bigdata17.reviews_large R
WHERE M.movieid = R.movieid
GROUP BY M.title
HAVING A > 4 AND reviews > 10
ORDER BY A ASC, title ASC;
```

MapReduce

Query 1

This part consisted of two Hadoop jobs - Popularity Calculator and Popularity Sorter. The first job had two mappers and one reducer. The second job had one mapper and one reducer.

Popularity Calculator Job

The first two mappers **`MapForPopularity()`** and **`MapForTitle()`** were used to read the reviews.csv and movies.csv file respectively. To make sure the reviews and movie title were grouped together in the intermediate values, we set our output key for both as the **`movieid`**. The output value for **`MapForPopularity()`** was the string *"pop 1"* to denote that it was a review and the output value for **`MapForTitle()`** was the string *"title <movie_name>"* which we would later on parse out in the reduce stage.

In the reduce stage **`ReduceForPopularity()`** we receive in a key (**`movieid`**) and an iterable of values in a format similar to this.

```
["title "War of the Worlds"", "pop 1", "pop 1", "pop 1"];
```

We iterate through this array and split the string by the first tab. If the first element in the resulting array is "title", we set our output value to be the second element. Otherwise we increment our running count by 1. The output key is then set to the running count.

Our output will now be a text file in the format **`[count, title]`** where the output is not ordered.

```
10    War of the Worlds
30    Terminator 2
5     Citizen Kane
```

Popularity Sorter Job

The second mapper **`DumbMapper()`** reads from the output file of the previous Hadoop job. Each line comes in the format of **`[count, title]`** as shown:

```
10    War of the Worlds
```

We then split the string to extract the movie title. We set the output value to the movie title and the output key to the original line in order to ensure that all generated keys are unique.

```
("10  War of the Worlds", "War of the Worlds");
```

To sort the intermediate results, we then overload the **`Writable Comparable`** class to sort the keys by both review count and movie title. This ensures two movies with the same review count, such as all the movies with only 1 review, do not get grouped together. If two keys have the same review count, the keys are sorted by the title in ascending order.

In the reducer stage we split the key to extract the review count and extract the title from the iterable. We then just write both of those values to file.

```
 5     Citizen Kane
10    War of the Worlds
30    Terminator 2
```

Query 2

For this part, we took a similar approach as Part 1, where we had two Hadoop jobs, Review Calculator and Review Sorter.

Review Calculator Job

The first two mappers **`MapForReview()`** and **`MapForTitle()`** were used to read the reviews.csv and movies.csv file respectively. They are written to the reducer the same way as Part 1, except the value for **`MapForReview()`** is the rating.

In the reducer stage, we count up all the reviews similarly to Part 1. We receive a key (**`movieid`**) and an iterable of values like this format.

```
["title "War of the Worlds"", "pop 3.0", "pop 3.5", "pop 4.0"];
```

We extract the review and add that to our running total score and increment our review count. We then check if the number of reviews is greater than 10 and if the average rating is greater than 4.0. If so, we write the result to the file. The output value is in the format "<movie_title> <average_rating> <review_count>".

```
War of the Worlds     4.125      11
```

Review Reducer Job

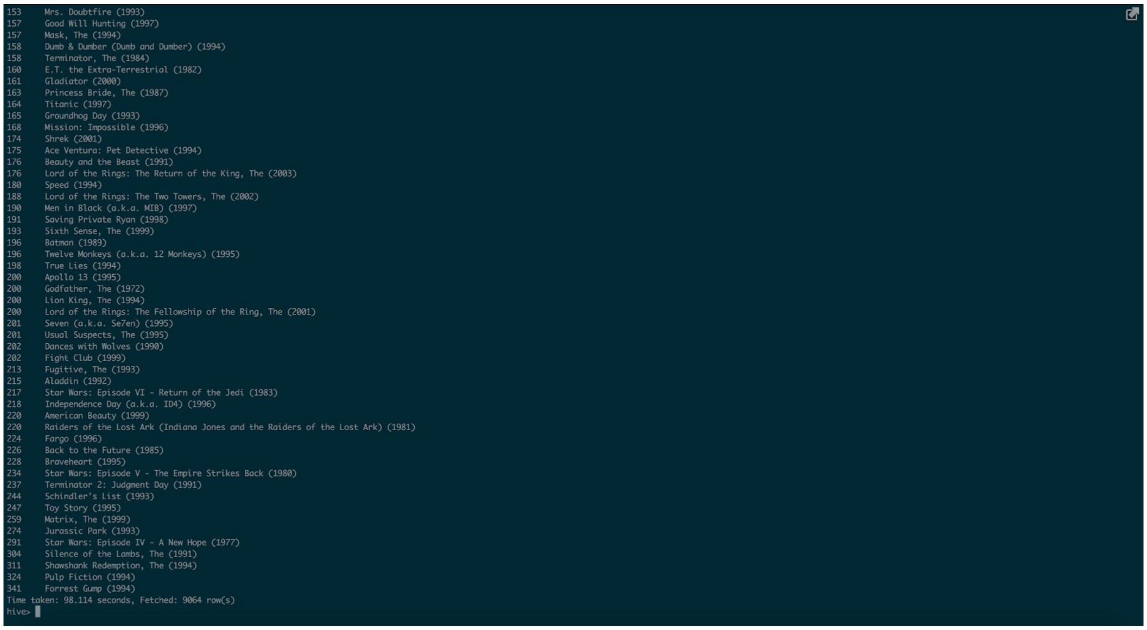
This job is the exact same implementation as the **Popularity Sorter Job** in Part 1, except the Comparator function sorts the intermediate results by average rating then movie title in ascending order. The final output is a list of values, sorted by average review. .

```
War of the Worlds     4.125      11
   Citizen Kane     4.2593     30
   Terminator 2     4.6667     15
```

Results

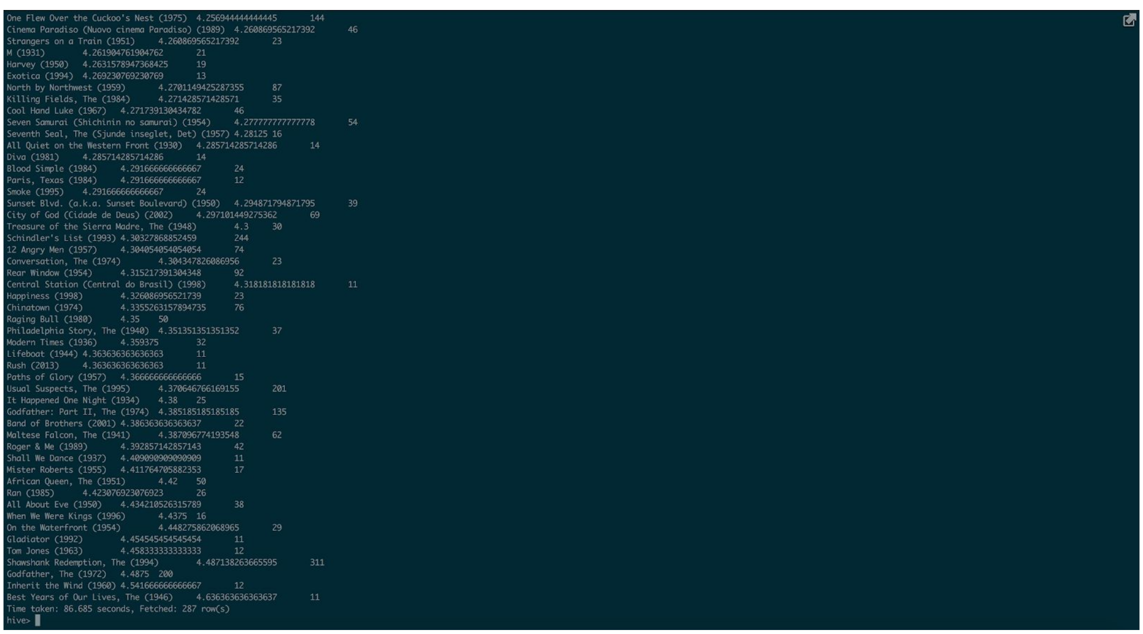
Hive

Small Dataset Part 1



```
153     Mrs. Doubtfire (1993)
157     Good Will Hunting (1997)
157     Mask, The (1994)
158     Dumb & Dumber (Dumb and Dumber) (1994)
158     Terminator, The (1984)
160     E.T. the Extra-Terrestrial (1982)
161     Gladiator (2000)
163     Princess Bride, The (1987)
164     Titanic (1997)
165     Groundhog Day (1993)
168     Mission: Impossible (1996)
174     Shrek (2001)
175     Ace Ventura: Pet Detective (1994)
176     Beauty and the Beast (1991)
176     Lord of the Rings: The Return of the King, The (2003)
180     Speed (1994)
188     Lord of the Rings: The Two Towers, The (2002)
190     Men in Black (a.k.a. MIB) (1997)
191     Saving Private Ryan (1998)
193     Sixth Sense, The (1999)
196     Batman (1989)
196     Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
198     True Lies (1994)
200     Apollo 13 (1995)
200     Godfather, The (1972)
200     Lion King, The (1994)
200     Lord of the Rings: The Fellowship of the Ring, The (2001)
201     Seven (a.k.a. Se7en) (1995)
201     Usual Suspects, The (1995)
202     Dances with Wolves (1990)
202     Fight Club (1999)
213     Fugitive, The (1993)
215     Aladdin (1992)
217     Star Wars: Episode VI - Return of the Jedi (1983)
218     Independence Day (a.k.a. ID4) (1996)
220     American Beauty (1999)
220     Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
224     Fargo (1996)
226     Back to the Future (1985)
228     Braveheart (1995)
234     Star Wars: Episode V - The Empire Strikes Back (1980)
237     Terminator 2: Judgment Day (1991)
244     Schindler's List (1993)
247     Toy Story (1995)
259     Matrix, The (1999)
274     Jurassic Park (1993)
291     Star Wars: Episode IV - A New Hope (1977)
304     Silence of the Lambs, The (1991)
311     Shawshank Redemption, The (1994)
324     Pulp Fiction (1994)
341     Forrest Gump (1994)
Time taken: 98.114 seconds, Fetched: 9064 row(s)
hive>
```

Small Dataset Part 1



```
One Flew Over the Cuckoo's Nest (1975)  4.256944444444445       144
Cinema Paradiso (Nuovo cinema Paradiso) (1989)  4.260869565217392       46
Strangers on a Train (1951)     4.260869565217392       23
M (1931)        4.261904761904762       21
Harvey (1950)   4.2631578947368425      19
Exotica (1994)  4.269230769230769       13
North by Northwest (1959)       4.2701149425287355      87
Killing Fields, The (1984)      4.271428571428571       35
Cool Hand Luke (1967)   4.271739130434782       46
Seven Samurai (Shichinin no samurai) (1954)     4.277777777777778       54
Seventh Seal, The (Sjunde inseglet, Det) (1957) 4.28125 16
All Quiet on the Western Front (1930)   4.285714285714286       14
Diva (1981)     4.285714285714286       14
Blood Simple (1984)     4.291666666666667       24
Paris, Texas (1984)     4.291666666666667       12
Smoke (1995)    4.291666666666667       24
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)   4.294871794871795       39
City of God (Cidade de Deus) (2002)     4.297101449275362       69
Treasure of the Sierra Madre, The (1948)        4.3     30
Schindler's List (1993) 4.30327868852459        244
12 Angry Men (1957)     4.304054054054054       74
Conversation, The (1974)        4.304347826086956       23
Rear Window (1954)      4.315217391304348       92
Central Station (Central do Brasil) (1998)      4.318181818181818       11
Happiness (1998)        4.326086956521739       23
Chinatown (1974)        4.3355263157894735      76
Raging Bull (1980)      4.35    50
Philadelphia Story, The (1940)  4.351351351351352       37
Modern Times (1936)     4.359375        32
Lifeboat (1944) 4.363636363636363       11
Rush (2013)     4.363636363636363       11
Paths of Glory (1957)   4.366666666666666       15
Usual Suspects, The (1995)      4.370646766169155       201
It Happened One Night (1934)    4.38    25
Godfather: Part II, The (1974)  4.385185185185185       135
Band of Brothers (2001) 4.386363636363637       22
Maltese Falcon, The (1941)      4.387096774193548       62
Roger & Me (1989)       4.392857142857143       42
Shall We Dance (1937)   4.409090909090909       11
Mister Roberts (1955)   4.411764705882353       17
African Queen, The (1951)       4.42    50
Ran (1985)      4.423076923076923       26
All About Eve (1950)    4.434210526315789       38
When We Were Kings (1996)       4.4375  16
On the Waterfront (1954)        4.448275862068965       29
Gladiator (1992)        4.454545454545454       11
Tom Jones (1963)        4.458333333333333       12
Shawshank Redemption, The (1994)        4.487138263665595       311
Godfather, The (1972)   4.4875  200
Inherit the Wind (1960) 4.541666666666667       12
Best Years of Our Lives, The (1946)     4.636363636363637       11
Time taken: 86.685 seconds, Fetched: 287 row(s)
hive>
```

For the small dataset, we got an expected 9064 rows for the first query. For the second query, we got a total of 287 rows. We also verified that the information was in ascending order.

Large Dataset Part 1

```
39227   Good Will Hunting (1997)
39600   Dark Knight, The (2008)
39725   Indiana Jones and the Last Crusade (1989)
40103   One Flew Over the Cuckoo's Nest (1975)
40250   Terminator, The (1984)
40436   Die Hard: With a Vengeance (1995)
40706   Memento (2000)
40931   Princess Bride, The (1987)
41283   Beauty and the Beast (1991)
42193   Men in Black (a.k.a. MIB) (1997)
42558   Titanic (1997)
43660   Shrek (2001)
44121   Mission: Impossible (1996)
44742   Ace Ventura: Pet Detective (1994)
45303   Lion King, The (1994)
45413   Gladiator (2000)
45544   Speed (1994)
49643   Sixth Sense, The (1999)
50168   True Lies (1994)
50387   Aladdin (1992)
50809   Saving Private Ryan (1998)
51338   Dances with Wolves (1990)
51837   Lord of the Rings: The Return of the King, The (2003)
51882   Lord of the Rings: The Two Towers, The (2002)
52474   Fargo (1996)
52658   Seven (a.k.a. Se7en) (1995)
53398   Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
53717   Batman (1989)
54783   Back to the Future (1985)
56820   Fugitive, The (1993)
56827   Lord of the Rings: The Fellowship of the Ring, The (2001)
57070   Godfather, The (1972)
57232   Independence Day (a.k.a. ID4) (1996)
57416   Apollo 13 (1995)
57879   American Beauty (1999)
59271   Usual Suspects, The (1995)
59693   Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
60024   Fight Club (1999)
61672   Star Wars: Episode V - The Empire Strikes Back (1980)
61836   Terminator 2: Judgment Day (1991)
62714   Star Wars: Episode VI - Return of the Jedi (1983)
66008   Toy Story (1995)
66512   Braveheart (1995)
67662   Schindler's List (1993)
74355   Jurassic Park (1993)
77045   Star Wars: Episode IV - A New Hope (1977)
77960   Matrix, The (1999)
84078   Silence of the Lambs, The (1991)
87901   Pulp Fiction (1994)
91082   Shawshank Redemption, The (1994)
91921   Forrest Gump (1994)
Time taken: 98.047 seconds, Fetched: 45069 row(s)
hive>
```

Large Dataset Part 2

```
Monty Python and the Holy Grail (1975)  4.155153873726253       39058
Memento (2000)  4.1570775807006335      40706
Touch of Evil (1958)    4.157241777264859       5199
To Kill a Mockingbird (1962)    4.157649361114308       17374
Chinatown (1974)        4.157715931945426       18397
Inception (2010)        4.161755956596878       35297
M (1931)        4.163554278678432       4873
Big Sleep, The (1946)   4.163652229097255       6303
A Song of Lisbon (1933) 4.166666666666667       15
Life Is Beautiful (La Vita è bella) (1997)      4.167062784709843       25245
Notorious (1946)        4.167152752460809       5486
Pulp Fiction (1994)     4.169975313136369       87901
All About Eve (1950)    4.174582798459563       5453
Black Mirror: White Christmas (2014)    4.174603174603175       63
Goodfellas (1990)       4.17828875746609        33987
Yojimbo (1961)  4.180264741275572       4155
Guten Tag, Ramón (2013) 4.181818181818182       11
Dark Knight, The (2008) 4.182070707070707       39600
Still Bill (2009)       4.1875  16
City of God (Cidade de Deus) (2002)     4.187872863087181       19947
O Auto da Compadecida (Dog's Will, A) (2000)    4.191558441558442       154
Lives of Others, The (Das leben der Anderen) (2006)     4.199038891372374       8948
A Silent Voice (2016)   4.2     20
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)   4.200819672131147       7930
Spirited Away (Sen to Chihiro no kamikakushi) (2001)    4.202589307120594       20855
Double Indemnity (1944) 4.202603887997147       5607
Paths of Glory (1957)   4.20264575040974        4271
North by Northwest (1959)       4.205228001893441       19013
Third Man, The (1949)   4.209418968212611       7676
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)     4.213030410183875       28280
Casablanca (1942)       4.2143927037912325      30043
The Blue Planet (2001)  4.217948717948718       273
Over the Garden Wall (2013)     4.219745222929936       157
Whiplash (2013) 4.226775956284153       183
One Flew Over the Cuckoo's Nest (1975)  4.22913497743311        40103
Fight Club (1999)       4.2307160469145675      60024
12 Angry Men (1957)     4.231208570075758       16896
Rear Window (1954)      4.232552144363722       21335
Seven Samurai (Shichinin no samurai) (1954)     4.255073602972702       13994
Godfather: Part II, The (1974)  4.263475012950189       36679
Human (2015)    4.264705882352941       34
Schindler's List (1993) 4.266530696698294       67662
Human Planet (2011)     4.271573604060913       197
Usual Suspects, The (1995)      4.300188962561792       59271
O Pátio das Cantigas (1942)     4.3076923076923075      13
Welfare (1975)  4.318181818181818       11
Godfather, The (1972)   4.339810758717364       57070
Planet Earth II (2016)  4.352631578947369       95
Band of Brothers (2001) 4.394366197183099       284
Shawshank Redemption, The (1994)        4.429014514393623       91082
Planet Earth (2006)     4.478779840848806       754
Time taken: 156.067 seconds, Fetched: 381 row(s)
hive>
```

For the small dataset, we got an expected 45069 rows for the first query. For the second query, we got a total of 381 rows. We also verified that the information was in ascending order.

MapReduce

Small Dataset Part 1

```
151 Die Hard (1988)
153 Mrs. Doubtfire (1993)
157 Good Will Hunting (1997)
157 Mask, The (1994)
158 Dumb & Dumber (Dumb and Dumber) (1994)
158 Terminator, The (1984)
160 E.T. the Extra-Terrestrial (1982)
161 Gladiator (2000)
163 Princess Bride, The (1987)
164 Titanic (1997)
165 Groundhog Day (1993)
168 Mission: Impossible (1996)
174 Shrek (2001)
175 Ace Ventura: Pet Detective (1994)
176 Beauty and the Beast (1991)
176 Lord of the Rings: The Return of the King, The (2003)
180 Speed (1994)
188 Lord of the Rings: The Two Towers, The (2002)
190 Men in Black (a.k.a. MIB) (1997)
191 Saving Private Ryan (1998)
193 Sixth Sense, The (1999)
196 Batman (1989)
196 Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
198 True Lies (1994)
200 Apollo 13 (1995)
200 Godfather, The (1972)
200 Lion King, The (1994)
200 Lord of the Rings: The Fellowship of the Ring, The (2001)
201 Seven (a.k.a. Se7en) (1995)
201 Usual Suspects, The (1995)
202 Dances with Wolves (1990)
202 Fight Club (1999)
213 Fugitive, The (1993)
215 Aladdin (1992)
217 Star Wars: Episode VI - Return of the Jedi (1983)
218 Independence Day (a.k.a. ID4) (1996)
220 American Beauty (1999)
220 Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
224 Fargo (1996)
226 Back to the Future (1985)
228 Braveheart (1995)
234 Star Wars: Episode V - The Empire Strikes Back (1980)
237 Terminator 2: Judgment Day (1991)
244 Schindler's List (1993)
247 Toy Story (1995)
259 Matrix, The (1999)
274 Jurassic Park (1993)
291 Star Wars: Episode IV - A New Hope (1977)
304 Silence of the Lambs, The (1991)
311 Shawshank Redemption, The (1994)
324 Pulp Fiction (1994)
341 Forrest Gump (1994)
```

Small Dataset Part 2

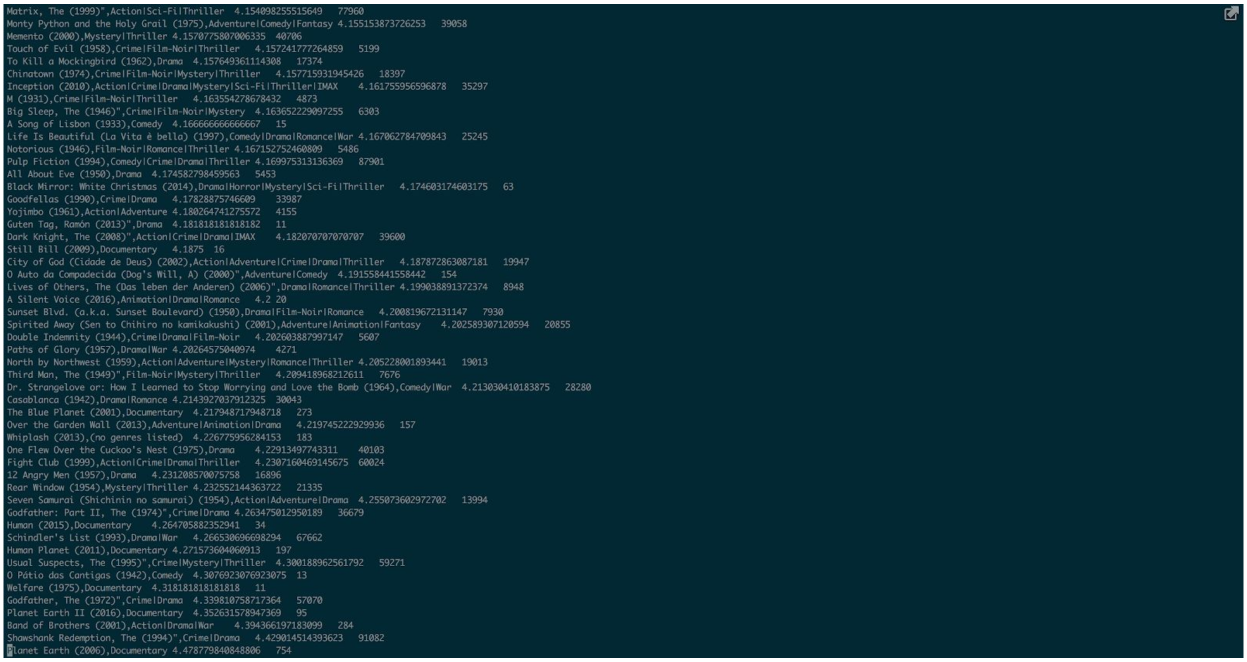
```
Fargo (1996),Comedy|Crime|Drama|Thriller    4.256696428571429   224
One Flew Over the Cuckoo's Nest (1975),Drama    4.256944444444445   144
Cinema Paradiso (Nuovo cinema Paradiso) (1989),Drama    4.260869565217392   46
Strangers on a Train (1951),Crime|Drama|Film-Noir|Thriller  4.260869565217392   23
M (1931),Crime|Film-Noir|Thriller   4.261904761904762   21
Harvey (1950),Comedy|Fantasy    4.2631578947368425  19
Exotica (1994),Drama    4.269230769230769   13
North by Northwest (1959),Action|Adventure|Mystery|Romance|Thriller 4.2701149425287355  87
Killing Fields, The (1984)",Drama|War   4.271428571428571   35
Cool Hand Luke (1967),Drama 4.271739130434782   46
Seven Samurai (Shichinin no samurai) (1954),Action|Adventure|Drama  4.277777777777778   54
Seventh Seal, The (Sjunde inseglet, Det) (1957)",Drama  4.28125 16
All Quiet on the Western Front (1930),Action|Drama|War  4.285714285714286   14
Diva (1981),Action|Drama|Mystery|Romance|Thriller   4.285714285714286   14
Blood Simple (1984),Crime|Drama|Film-Noir   4.291666666666667   24
Paris, Texas (1984)",Drama|Romance  4.291666666666667   12
Smoke (1995),Comedy|Drama   4.291666666666667   24
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950),Drama|Film-Noir|Romance   4.294871794871795   39
City of God (Cidade de Deus) (2002),Action|Adventure|Crime|Drama|Thriller   4.297101449275362   69
Treasure of the Sierra Madre, The (1948)",Action|Adventure|Drama|Western    4.3 30
Schindler's List (1993),Drama|War   4.30327868852459    244
12 Angry Men (1957),Drama   4.304054054054054   74
Conversation, The (1974)",Drama|Mystery 4.304347826086956   23
Rear Window (1954),Mystery|Thriller 4.315217391304348   92
Central Station (Central do Brasil) (1998),Drama    4.318181818181818   11
Happiness (1998),Comedy|Drama   4.326086956521739   23
Chinatown (1974),Crime|Film-Noir|Mystery|Thriller   4.3355263157894735  76
Raging Bull (1980),Drama    4.35    50
Philadelphia Story, The (1940)",Comedy|Drama|Romance    4.351351351351352   37
Modern Times (1936),Comedy|Drama|Romance    4.359375    32
Lifeboat (1944),Drama|War   4.363636363636363   11
Rush (2013),Action|Drama    4.363636363636363   11
Paths of Glory (1957),Drama|War 4.366666666666666   15
Usual Suspects, The (1995)",Crime|Mystery|Thriller  4.370646766169155   201
It Happened One Night (1934),Comedy|Romance 4.38    25
Godfather: Part II, The (1974)",Crime|Drama 4.385185185185185   135
Band of Brothers (2001),Action|Drama|War    4.386363636363637   22
Maltese Falcon, The (1941)",Film-Noir|Mystery   4.387096774193548   62
Roger & Me (1989),Documentary   4.392857142857143   42
Shall We Dance (1937),Comedy|Musical|Romance    4.409090909090909   11
Mister Roberts (1955),Comedy|Drama|War  4.411764705882353   17
African Queen, The (1951)",Adventure|Comedy|Romance|War 4.42    50
Ran (1985),Drama|War    4.423076923076923   26
All About Eve (1950),Drama  4.434210526315789   38
When We Were Kings (1996),Documentary   4.4375  16
On the Waterfront (1954),Crime|Drama    4.448275862068965   29
Gladiator (1992),Action|Drama   4.454545454545454   11
Tom Jones (1963),Adventure|Comedy|Romance   4.458333333333333   12
Shawshank Redemption, The (1994)",Crime|Drama   4.487138263665595   311
Godfather, The (1972)",Crime|Drama  4.4875  200
Inherit the Wind (1960),Drama   4.541666666666667   12
Best Years of Our Lives, The (1946)",Drama|War  4.636363636363637   11
```

For the small dataset, we got 9066 rows for the first query and 287 for the second.

Large Dataset Part 1

```
39058   Monty Python and the Holy Grail (1975)
39227   Good Will Hunting (1997)
39600   Dark Knight, The (2008)
39725   Indiana Jones and the Last Crusade (1989)
40103   One Flew Over the Cuckoo's Nest (1975)
40250   Terminator, The (1984)
40436   Die Hard: With a Vengeance (1995)
40706   Memento (2000)
40931   Princess Bride, The (1987)
41283   Beauty and the Beast (1991)
42193   Men in Black (a.k.a. MIB) (1997)
42558   Titanic (1997)
43660   Shrek (2001)
44121   Mission: Impossible (1996)
44742   Ace Ventura: Pet Detective (1994)
45303   Lion King, The (1994)
45413   Gladiator (2000)
45544   Speed (1994)
49643   Sixth Sense, The (1999)
50168   True Lies (1994)
50375   Aladdin (1992)
50809   Saving Private Ryan (1998)
51338   Dances with Wolves (1990)
51837   Lord of the Rings: The Return of the King, The (2003)
51882   Lord of the Rings: The Two Towers, The (2002)
52474   Fargo (1996)
52658   Seven (a.k.a. Se7en) (1995)
53398   Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
53717   Batman (1989)
54783   Back to the Future (1985)
56820   Fugitive, The (1993)
56827   Lord of the Rings: The Fellowship of the Ring, The (2001)
57070   Godfather, The (1972)
57232   Independence Day (a.k.a. ID4) (1996)
57416   Apollo 13 (1995)
57879   American Beauty (1999)
59271   Usual Suspects, The (1995)
59693   Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
60024   Fight Club (1999)
61672   Star Wars: Episode V - The Empire Strikes Back (1980)
61836   Terminator 2: Judgment Day (1991)
62714   Star Wars: Episode VI - Return of the Jedi (1983)
66008   Toy Story (1995)
66512   Braveheart (1995)
67662   Schindler's List (1993)
74355   Jurassic Park (1993)
77045   Star Wars: Episode IV - A New Hope (1977)
77960   Matrix, The (1999)
84078   Silence of the Lambs, The (1991)
87901   Pulp Fiction (1994)
91082   Shawshank Redemption, The (1994)
91921   Forrest Gump (1994)
```

Large Dataset Part 2



```
Matrix, The (1999)",Action|Sci-Fi|Thriller  4.154098255515649   77960
Monty Python and the Holy Grail (1975),Adventure|Comedy|Fantasy 4.155153873726253   39058
Memento (2000),Mystery|Thriller 4.1570775807006335  40706
Touch of Evil (1958),Crime|Film-Noir|Thriller   4.157241777264859   5199
To Kill a Mockingbird (1962),Drama  4.157649361114308   17374
Chinatown (1974),Crime|Film-Noir|Mystery|Thriller   4.157715931945426   18397
Inception (2010),Action|Crime|Drama|Mystery|Sci-Fi|Thriller|IMAX    4.161755956596878   35297
M (1931),Crime|Film-Noir|Thriller   4.163554278678432   4873
Big Sleep, The (1946)",Crime|Film-Noir|Mystery  4.163652229097255   6303
A Song of Lisbon (1933),Comedy  4.166666666666667   15
Life Is Beautiful (La Vita è bella) (1997),Comedy|Drama|Romance|War 4.167062784709843   25245
Notorious (1946),Film-Noir|Romance|Thriller 4.167152752460809   5486
Pulp Fiction (1994),Comedy|Crime|Drama|Thriller 4.169975313136369   87901
All About Eve (1950),Drama  4.174582798459563   5453
Black Mirror: White Christmas (2014),Drama|Horror|Mystery|Sci-Fi|Thriller   4.174603174603175   63
Goodfellas (1990),Crime|Drama   4.17828875746609    33987
Yojimbo (1961),Action|Adventure 4.180264741275572   4155
Guten Tag, Ramón (2013)",Drama  4.181818181818182   11
Dark Knight, The (2008)",Action|Crime|Drama|IMAX    4.182070707070707   39600
Still Bill (2009),Documentary   4.1875  16
City of God (Cidade de Deus) (2002),Action|Adventure|Crime|Drama|Thriller   4.187872863087181   19947
O Auto da Compadecida (Dog's Will, A) (2000)",Adventure|Comedy  4.191558441558442   154
Lives of Others, The (Das leben der Anderen) (2006)",Drama|Romance|Thriller 4.199038891372374   8948
A Silent Voice (2016),Animation|Drama|Romance   4.2 20
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950),Drama|Film-Noir|Romance   4.200819672131147   7930
Spirited Away (Sen to Chihiro no kamikakushi) (2001),Adventure|Animation|Fantasy    4.202589307120594   20855
Double Indemnity (1944),Crime|Drama|Film-Noir   4.202603887997147   5607
Paths of Glory (1957),Drama|War 4.20264575040974    4271
North by Northwest (1959),Action|Adventure|Mystery|Romance|Thriller 4.205228001893441   19013
Third Man, The (1949)",Film-Noir|Mystery|Thriller   4.209418968212611   7676
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964),Comedy|War  4.213030410183875   28280
Casablanca (1942),Drama|Romance 4.2143927037912325  30043
The Blue Planet (2001),Documentary  4.217948717948718   273
Over the Garden Wall (2013),Adventure|Animation|Drama   4.219745222929936   157
Whiplash (2013),(no genres listed)  4.226775956284153   183
One Flew Over the Cuckoo's Nest (1975),Drama    4.22913497743311    40103
Fight Club (1999),Action|Crime|Drama|Thriller   4.2307160469145675  60024
12 Angry Men (1957),Drama   4.231208570075758   16896
Rear Window (1954),Mystery|Thriller 4.232552144363722   21335
Seven Samurai (Shichinin no samurai) (1954),Action|Adventure|Drama  4.255073602972702   13994
Godfather: Part II, The (1974)",Crime|Drama 4.263475012950189   36679
Human (2015),Documentary    4.264705882352941   34
Schindler's List (1993),Drama|War   4.266530696698294   67662
Human Planet (2011),Documentary 4.271573604060913   197
Usual Suspects, The (1995)",Crime|Mystery|Thriller  4.300188962561792   59271
O Pátio das Cantigas (1942),Comedy  4.3076923076923075  13
Welfare (1975),Documentary  4.318181818181818   11
Godfather, The (1972)",Crime|Drama  4.339810758717364   57070
Planet Earth II (2016),Documentary  4.352631578947369   95
Band of Brothers (2001),Action|Drama|War    4.394366197183099   284
Shawshank Redemption, The (1994)",Crime|Drama   4.429014514393623   91082
Planet Earth (2006),Documentary 4.478779840848806   754
```

For the large dataset we got 45112 for the first query and 381 rows for the second.

Discrepancies in Row Counts and Explanation

	Dataset	# Rows Query 1	# Rows Query 2
Small Dataset	Hive	9064	287
	MapReduce	9066	287
Large Dataset	Hive	45069	381
	MapReduce	45112	381

As shown, the number of rows created matched for Query 2. However the number of rows for Query 1 differed between Hive and MapReduce. Upon further analysis we found out that that this is due to the differences in implementation for our Hadoop jobs and the Hive query.

Two movies, Hamlet (2000) and War of the Worlds (2005) in the small dataset that had two different movie ids. Out MapReduce job groups by movie id which creates two extra rows, while the Hive groups by title leading to all Hamlet and all War of the Worlds rows being correctly group. We can modify our MapReduce job to group by movie id to fix this.

Discussion (Part 3)

Results for the first Query (Default)

	Dataset	Jobs	Total mappers	Total reducers	Time Taken (s)
MapReduce	Small	2	2/96 => 98	96/1 => 97	147.550
	Large	2	7/96 => 103	96/1 => 97	343.130
Hive	Small	2	1/1 => 2	1/1 => 2	8.520
	Large	2	6/7 => 13	11/1 => 12	210.270

For time taken, we decided to use CPU time as our metric. We chose CPU time over other metrics because it has the best real-world application as it is experienced by the programmer. The Hive queries outperformed the MapReduce queries. We believe that this had to do with the difference in number of mappers and reducers allocated. For example, for the small dataset, the Hive query outperformed the MapReduce dataset by 139.03 seconds and executed 191 fewer tasks.

In the MapReduce code, we explicitly set the number of reduce tasks for the second job to 1, regardless of the dataset size. This guarantees the final output is always sorted. This behavior is also automatically set by Hive at compile time. There are a minimum of two mappers for the first job because we have two input paths. The rest is left to the JobSubmitter to decide. The number of reduces in the first job is equal to the number of mappers in the second job and by default is set to 96 for both the small and large dataset. The large dataset also used more mappers, as configured by the JobSubmitter.

Results for the second Query

	Dataset	Jobs	Total mappers	Total reducers	Time Taken (s)
MapReduce	Small	2	2/96 => 98	96/1 => 97	146.960
	Large	2	7/96 => 103	96/1 => 97	305.740
Hive	Small	2	1/1 => 2	1/1 => 2	7.470
	Large	2	6/4 => 10	11/1 => 12	196.060

Like the first query, the Hive queries significantly outperformed the MapReduce code. By default: the MapReduce code runs 96 reduce tasks in the first job and 96 map tasks in the second job.

Observations

With such a large difference in the number of tasks being run per job, compared to the number of tasks used in each Hive query, it is unsurprising the MapReduce jobs perform so much slower. Having so many tasks increases the memory and execution overhead. As will be seen and discussed in Part 4, specifying the number of reducers the MapReduce code uses significantly improves the performance of the programs.

Discussion (Part 4)

Since the Hive queries performed so much better, we decided to try optimizing the MapReduce jobs to minimize the performance gap. We decided to test using the Part 2 query.

We ran the Reviews program on all four datasets provided. The only setting we altered was the number of reduce tasks performed in the first job. The second job has an equal number of mappers since there is always one mapper for each reducer in the first job. The number of mappers for the first job is fixed, dictated by the JobSubmitter, and there is only ever 1 reducer in the second job to ensure sorted order of the final output.

For each iteration of the test, three time counters were recorded for each job: total CPU time spent for all map tasks; total CPU time spent for all reduce tasks; and the elapsed CPU time. The first time in each table field is for job 1, the second for job 2. We can use these two times to see how altering reducers affects job 1 and altering mappers affects job 2.

Small Dataset
There are two mappers used in the first job for each iteration of the test.

Reducers	Total Mapper Time (s)	Total Reducer Time (s)	CPU Time (s)
1	12.909+2.790=15.699	6.163+3.239=9.402	5.190+1.340=6.530
2	7.987+9.992=17.979	6.925+2.670=9.622	5.120+1.960=7.080
4	7.145+15.098=22.243	19.990+2.490=22.480	10.330+3.020=13.350
8	11.613+33.855=45.468	38.503+2.738=41.241	17.220+4.760=21.980
16	6.461+72.746=79.207	85.795+2.031=87.826	27.120+8.820=35.940

50MB Dataset
There are two mappers used in the first job for each iteration of the test.

Reducers	Total Mapper Time (s)	Total Reducer Time (s)	CPU Time (s)
1	13.446+2.772=16.218	4.562+2.951=7.513	9.710+1.350=11.060
2	8.520+8.960=17.480	8.671+7.841=16.512	10.400+1.870=12.270
4	18.872+16.895=35.767	29.705+7.815=37.520	13.280+3.200=16.480
8	9.149+34.897=44.046	55.868+2.750=58.618	20.410+5.120=25.530
16	8.725+75.147=83.872	92.315+2.416=94.731	36.010+8.740=44.750

300MB Dataset
There are four mappers used in the first job for each iteration of the test.

Reducers	Total Mapper Time (s)	Total Reducer Time (s)	CPU Time (s)
1	35.336+2.751=38.087	11.855+2.731=14.586	35.030+1.330=36.360
2	36.454+6.032=42.486	18.006+2.282=20.288	38.670+1.950=40.620
4	42.622+22.856=65.478	38.017+2.826=40.843	47.850+3.270=51.120
8	37.354+29.200=66.554	69.656+2.253=71.909	56.300+5.200=61.500
16	52.651+89.927=142.578	100.182+4.545=104.727	68.520+10.360=78.880

Large Dataset
There are seven mappers used in the first job for each iteration of the test.

Reducers	Total Mapper Time (s)	Total Reducer Time (s)	CPU Time (s)
1	75.386+4.201=79.587	23.939+3.252=27.191	66.450+1.280=67.730
2	101.909+6.369=108.278	26.582+2.003=28.585	88.570+1.920=90.490
4	66.441+16.013=82.454	57.006+2.781=59.787	75.200+2.990=78.190
8	89.167+36.379=125.546	64.210+2.928=67.138	102.520+5.240=107.760
16	69.888+72.649=142.537	93.113+2.941=96.054	104.280+8.950=113.230

Observations

With the exception of using 4 reducers on the large dataset, an increase in reducers causes n increase in execution time. In most iterations of the test, all three time counters significantly increased with the increase in tasks. We believe this to be due to a dramatic increase in overhead. With more reducers in the first job, more time is spent initiating reduce tasks and on I/O operations since more files are written. With more output files, the NameNode also has an increased workload. The same observations can be made about the increase in mappers in the second job. There is more time spent on I/O operations, network delays, and task initiation.

In all four datasets, the reducer time for the second job was always pretty low since there was only every one reducer. We believed it was not necessary to test with more reducers since having only one reducer was always the quickest configuration. With only one reducer, our MapReduce code begins to perform almost as well as our Hive queries.

Commands to Execute

Our datasets are stored in Hadoop in the folder structure suggested in the project description. Each MapReduce program takes four parameters: movies CSV file path; reviews CSV file path; output directory path for the intermediate results; and the output directory path for the final sorted results. Example commands are below.

```
> Hadoop jar Popularity.jar <movies.csv> <reviews.csv> <output_folder>
<output_sorted_folder>

> Hadoop jar Popularity.jar ./dataset/movies/movies.csv
./dataset/reviews/reviews.csv ./output1_small ./output1_small_sorted

> Hadoop jar Reviews.jar ./dataset/movies/movies.csv
./dataset/reviews/reviews.csv ./output1_small ./output1_small_sorted
```