Screenshots of Pig Cleaning

```
grunt> REGISTER '/home/niamh/hadoop/pig-0.17.0/contrib/piggybank/java/piggybank.jar';
2021-10-23 19:58:00,142 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
grunt> DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;
grunt>
grunt> -- load movies.csv
grunt> movies_orig = LOAD './ml-latest-small/movies.csv' using CSVExcelStorage() AS (movieId:int, title:chararray
, genres:chararray);
2021-10-23 19:58:00,689 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
grunt>
grunt> -- current state: (movieId, title (year), list of genres seperated by '|')
grunt> -- e.g. (1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy)
grunt>
grunt> -- Split genres on '|', and store in new table with only one genre per row (from: movieId, genre)
grunt> mId_genres = FOREACH movies_orig GENERATE movieId,FLATTEN(STRSPLIT(genres,'\\|'));
grunt> genres = FOREACH mId_genres GENERATE $0 AS movieId,FLATTEN(TOBAG($1..));
grunt>
grunt> -- seperate title and year, and remove genres
grunt> movies_clean = FOREACH movies_orig GENERATE
grunt>     movieId,
>>     REGEX_EXTRACT(title,'(.+?(?= \\([0-9][0-9][0-9][0-9]\\)))|(.+)',0) as title,
>>     REGEX_EXTRACT(title,'(?<=\\()[1-2][0-9][0-9][0-9](?=\\))',0) as year;
grunt>
```

Movies Cleaned:

```
(1,Toy Story,1995)
(2,Jumanji,1995)
(3,Grumpier Old Men,1995)
(4,Waiting to Exhale,1995)
(5,Father of the Bride Part II,1995)
(6,Heat,1995)
(7,Sabrina,1995)
(8,Tom and Huck,1995)
(9,Sudden Death,1995)
grunt>
```

Genres:

```
(1,Comedy)
(1,Fantasy)
(1,Children)
(1,Adventure)
(1,Animation)
(2,Fantasy)
(2,Children)
(2,Adventure)
(3,Comedy)
```

Screenshots of Pig Queries

Query 1:

```
l input paths to process : 1
(356,Forrest Gump,1994,356,329)
(318,Shawshank Redemption, The,1994,318,317)
(296,Pulp Fiction,1994,296,307)
(593,Silence of the Lambs, The,1991,593,279)
(2571,Matrix, The,1999,2571,278)
(260,Star Wars: Episode IV - A New Hope,1977,260,251)
(480,Jurassic Park,1993,480,238)
(110,Braveheart,1995,110,237)
(589,Terminator 2: Judgment Day,1991,589,224)
(527,Schindler's List,1993,527,220)
grunt>
```

Query 2:

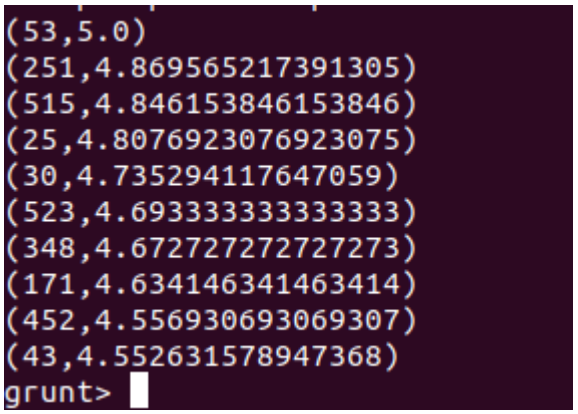
(movie with most 5 star ratings)

```
(318,Shawshank Redemption, The,1994,153,274)
(296,Pulp Fiction,1994,123,244)
(356,Forrest Gump,1994,116,249)
(2571,Matrix, The,1999,109,222)
(260,Star Wars: Episode IV - A New Hope,1977,104,201)
(527,Schindler's List,1993,92,175)
(593,Silence of the Lambs, The,1991,92,225)
(858,Godfather, The,1972,88,158)
(2959,Fight Club,1999,81,179)
(1196,Star Wars: Episode V - The Empire Strikes Back,1980,80,168)
grunt>
```

(movie with most ratings of 4 or more stars)

```
(318,Shawshank Redemption, The,1994,153,274)
(356,Forrest Gump,1994,116,249)
(296,Pulp Fiction,1994,123,244)
(593,Silence of the Lambs, The,1991,92,225)
(2571,Matrix, The,1999,109,222)
(260,Star Wars: Episode IV - A New Hope,1977,104,201)
(2959,Fight Club,1999,81,179)
(527,Schindler's List,1993,92,175)
(1196,Star Wars: Episode V - The Empire Strikes Back,1980,80,168)
(110,Braveheart,1995,80,166)
grunt>
```

Query 3:


(53,5.0)
(251,4.869565217391305)
(515,4.846153846153846)
(25,4.8076923076923075)
(30,4.735294117647059)
(523,4.693333333333333)
(348,4.672727272727273)
(171,4.634146341463414)
(452,4.556930693069307)
(43,4.552631578947368)
grunt>