Week 3

Felipe Rodriguez

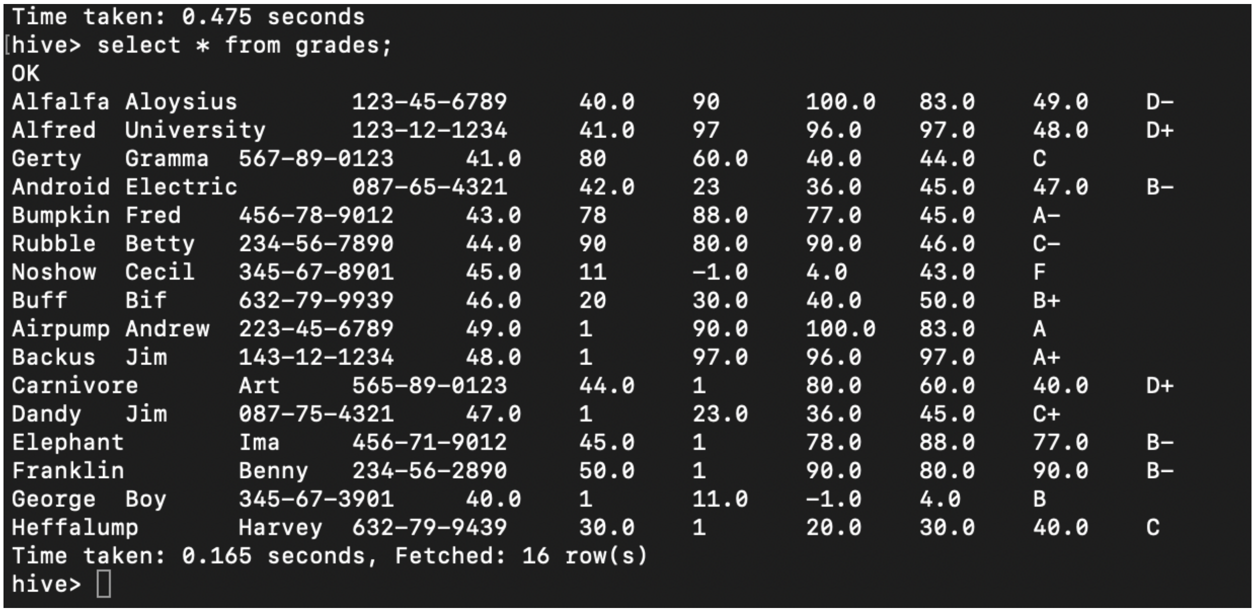
Bellevue University

DSC 650 Big Data

Professor Nasheb Ismaily

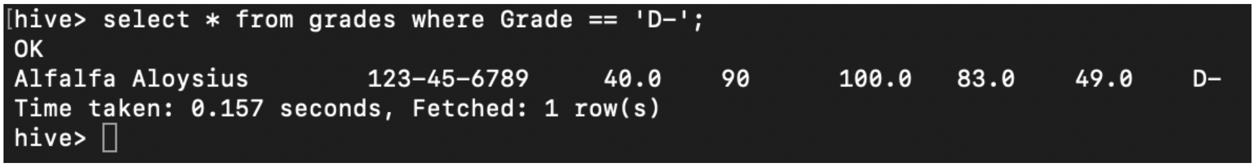
December 17, 2023

Query Results



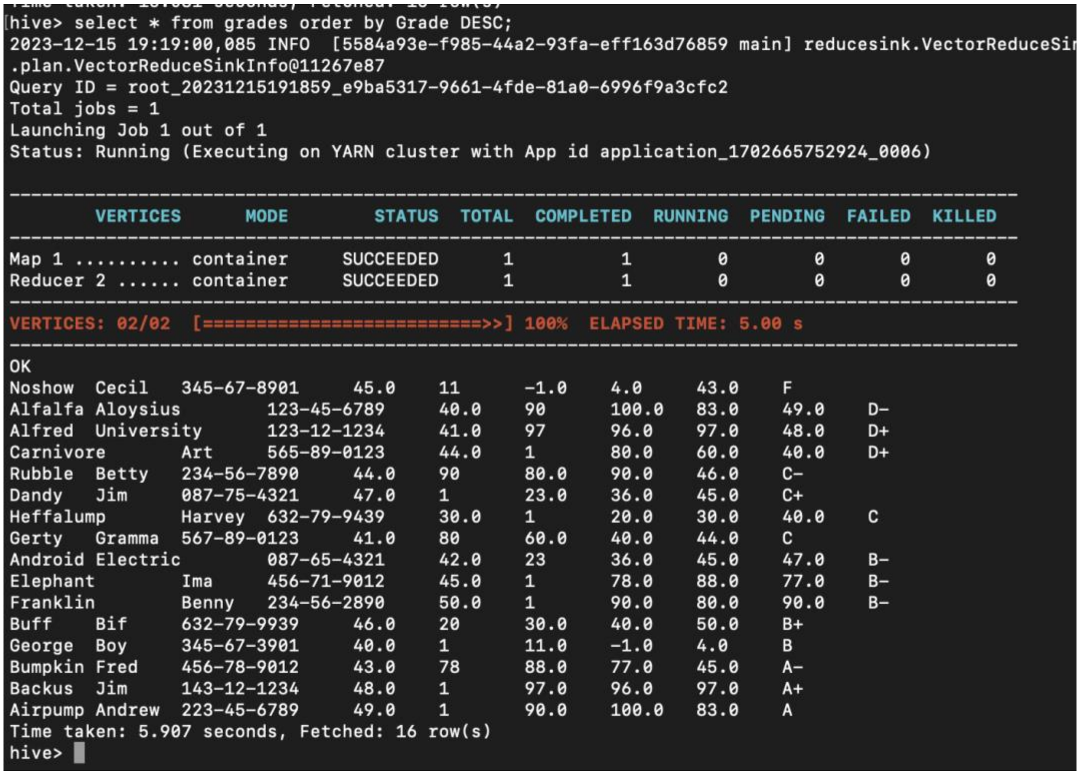
```
Time taken: 0.475 seconds
[hive> select * from grades;
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      123-12-1234     41.0    97      96.0    97.0    48.0    D+
Gerty   Gramma  567-89-0123     41.0    80      60.0    40.0    44.0    C
Android Electric        087-65-4321     42.0    23      36.0    45.0    47.0    B-
Bumpkin Fred    456-78-9012     43.0    78      88.0    77.0    45.0    A-
Rubble  Betty   234-56-7890     44.0    90      80.0    90.0    46.0    C-
Noshow  Cecil   345-67-8901     45.0    11      -1.0    4.0     43.0    F
Buff    Bif     632-79-9939     46.0    20      30.0    40.0    50.0    B+
Airpump Andrew  223-45-6789     49.0    1       90.0    100.0   83.0    A
Backus  Jim     143-12-1234     48.0    1       97.0    96.0    97.0    A+
Carnivore       Art     565-89-0123     44.0    1       80.0    60.0    40.0    D+
Dandy   Jim     087-75-4321     47.0    1       23.0    36.0    45.0    C+
Elephant        Ima     456-71-9012     45.0    1       78.0    88.0    77.0    B-
Franklin        Benny   234-56-2890     50.0    1       90.0    80.0    90.0    B-
George  Boy     345-67-3901     40.0    1       11.0    -1.0    4.0     B
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Time taken: 0.165 seconds, Fetched: 16 row(s)
hive> 
```

Query 1



```
[hive> select * from grades where Grade == 'D-';
OK
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Time taken: 0.157 seconds, Fetched: 1 row(s)
hive> 
```

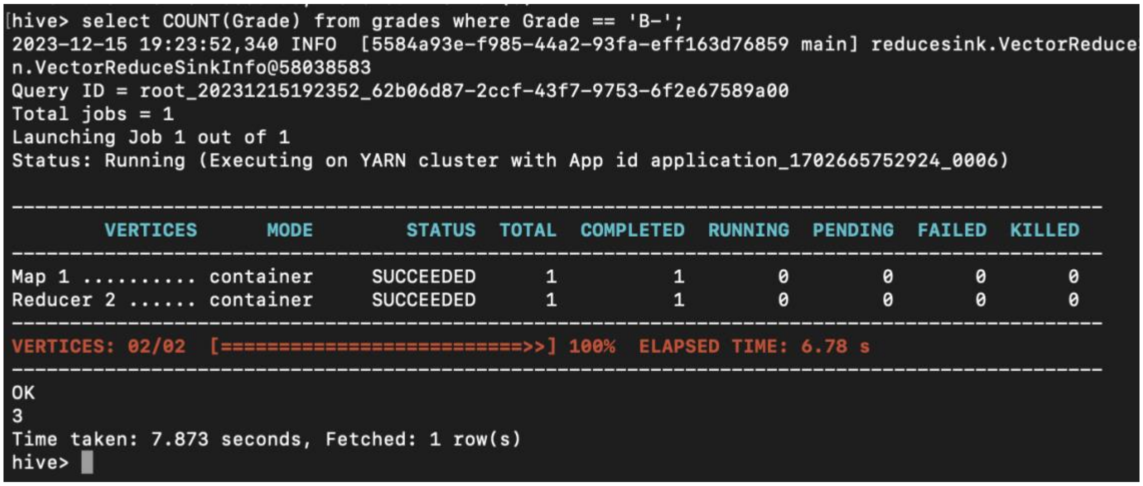
Query 2



```
[hive> select * from grades order by Grade DESC;
2023-12-15 19:19:00,085 INFO  [5584a93e-f985-44a2-93fa-eff163d76859 main] reducesink.VectorReduceSi
.plan.VectorReduceSinkInfo@11267e87
Query ID = root_20231215191859_e9ba5317-9661-4fde-81a0-6996f9a3cfc2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702665752924_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.00 s
----------------------------------------------------------------------------------------------
OK
Noshow  Cecil   345-67-8901     45.0    11      -1.0    4.0     43.0    F
Alfalfa Aloysius        123-45-6789     40.0    90      100.0   83.0    49.0    D-
Alfred  University      123-12-1234     41.0    97      96.0    97.0    48.0    D+
Carnivore       Art     565-89-0123     44.0    1       80.0    60.0    40.0    D+
Rubble  Betty   234-56-7890     44.0    90      80.0    90.0    46.0    C-
Dandy   Jim     087-75-4321     47.0    1       23.0    36.0    45.0    C+
Heffalump       Harvey  632-79-9439     30.0    1       20.0    30.0    40.0    C
Gerty   Gramma  567-89-0123     41.0    80      60.0    40.0    44.0    C
Android Electric        087-65-4321     42.0    23      36.0    45.0    47.0    B-
Elephant        Ima     456-71-9012     45.0    1       78.0    88.0    77.0    B-
Franklin        Benny   234-56-2890     50.0    1       90.0    80.0    90.0    B-
Buff    Bif     632-79-9939     46.0    20      30.0    40.0    50.0    B+
George  Boy     345-67-3901     40.0    1       11.0    -1.0    4.0     B
Bumpkin Fred    456-78-9012     43.0    78      88.0    77.0    45.0    A-
Backus  Jim     143-12-1234     48.0    1       97.0    96.0    97.0    A+
Airpump Andrew  223-45-6789     49.0    1       90.0    100.0   83.0    A
Time taken: 5.907 seconds, Fetched: 16 row(s)
hive> 
```

Query 3



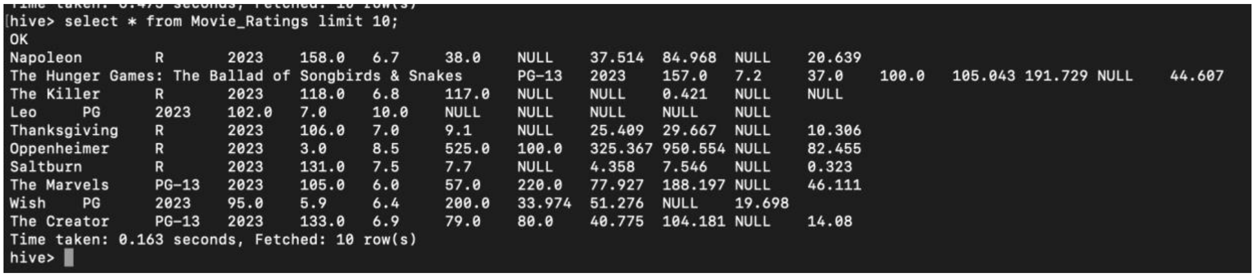
```
[hive> select COUNT(Grade) from grades where Grade == 'B-';
2023-12-15 19:23:52,340 INFO  [5584a93e-f985-44a2-93fa-eff163d76859 main] reducesink.VectorReduce
n.VectorReduceSinkInfo@58038583
Query ID = root_20231215192352_62b06d87-2ccf-43f7-9753-6f2e67589a00
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702665752924_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.78 s
----------------------------------------------------------------------------------------------
OK
3
Time taken: 7.873 seconds, Fetched: 1 row(s)
hive> 
```

Summary

The data set selected contains IMDb Movie rating. This data set was chosen to analyze Ratings and the Gross generated by the movies. The first query confirms if the data has been loaded correctly prior to beginning analysis. There is a limit set to 10 so that there is not an excess of data being omitted.

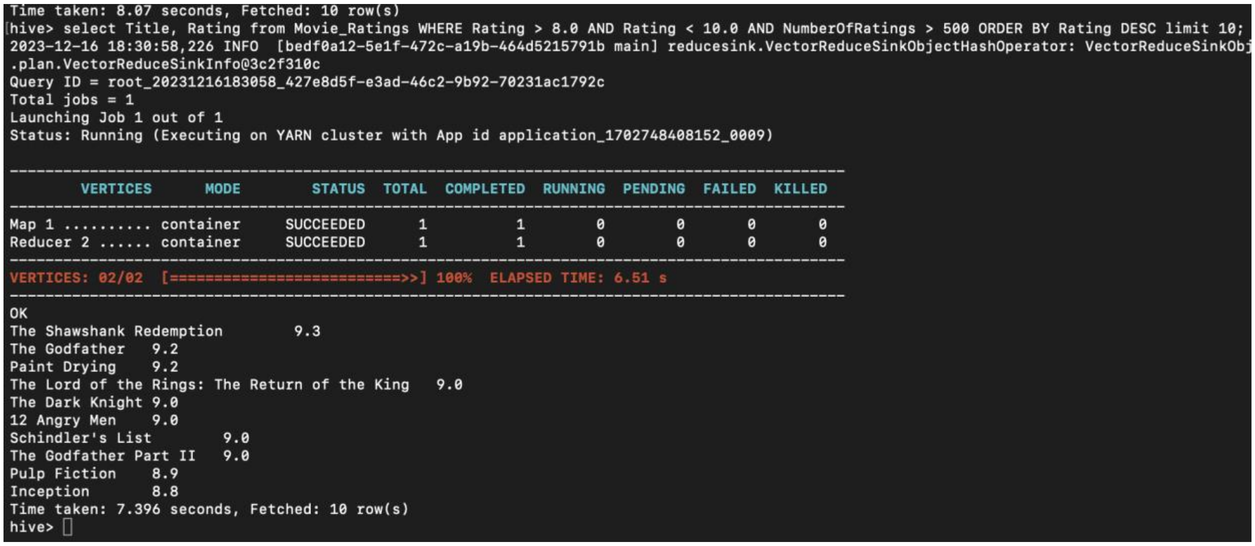
Query 1



```
Time taken: 0.473 seconds, Fetched: 10 row(s)
hive> select * from Movie_Ratings limit 10;
OK
Napoleon        R       2023    158.0   6.7     38.0    NULL    37.514  84.968  NULL    20.639
The Hunger Games: The Ballad of Songbirds & Snakes      PG-13   2023    157.0   7.2     37.0    100.0   105.043 191.729 NULL    44.607
The Killer      R       2023    118.0   6.8     117.0   NULL    NULL    0.421   NULL    NULL
Leo     PG      2023    102.0   7.0     10.0    NULL    NULL    NULL    NULL    NULL
Thanksgiving    R       2023    106.0   7.0     9.1     NULL    25.409  29.667  NULL    10.306
Oppenheimer     R       2023    3.0     8.5     525.0   100.0   325.367 950.554 NULL    82.455
Saltburn        R       2023    131.0   7.5     7.7     NULL    4.358   7.546   NULL    0.323
The Marvels     PG-13   2023    105.0   6.0     57.0    220.0   77.927  188.197 NULL    46.111
Wish    PG      2023    95.0    5.9     6.4     200.0   33.974  51.276  NULL    19.698
The Creator     PG-13   2023    133.0   6.9     79.0    80.0    40.775  104.181 NULL    14.08
Time taken: 0.163 seconds, Fetched: 10 row(s)
hive> 
```

The next query involves obtaining the movies with the highest ratings. Because the data is not yet cleansed, the filters are set to movies that are within 8 and 10 rating. This is done to ensure there are no movie with a rating above 10. Also, it is inclusive of movies with over 500 ratings. Again, it is limited to only the top 10 movies.

Query 2

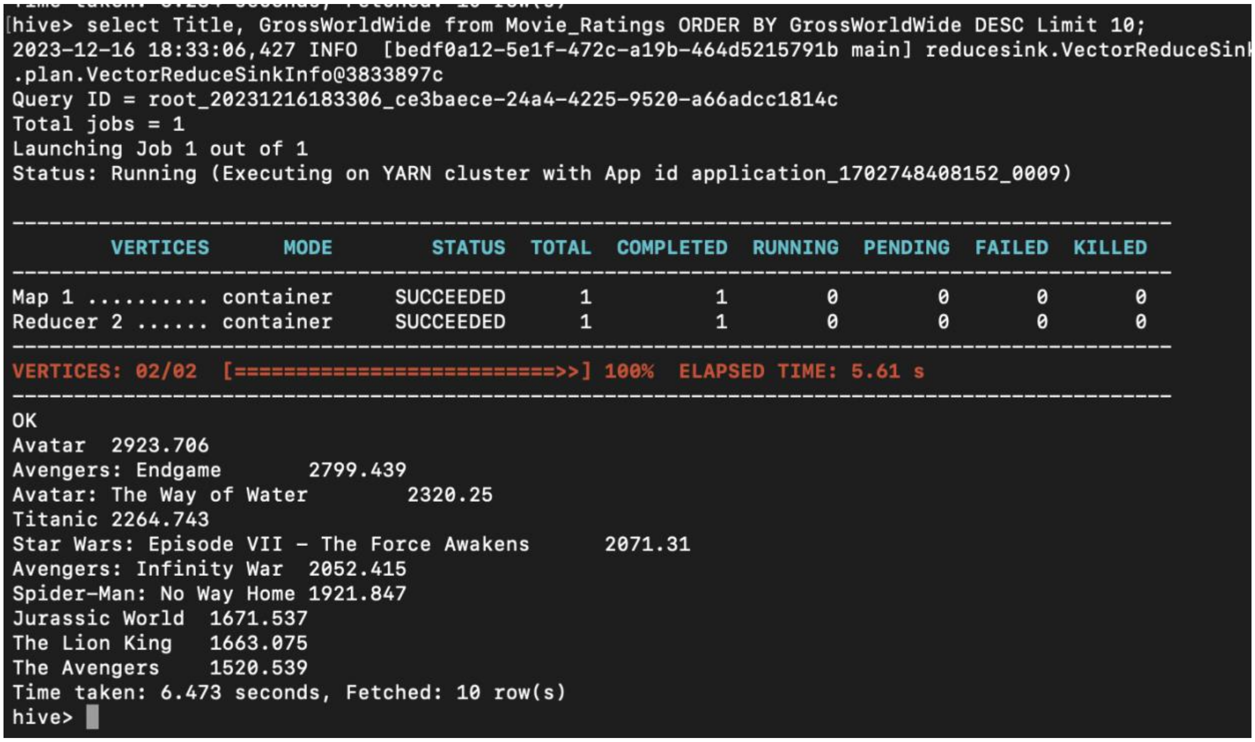


```
Time taken: 8.07 seconds, Fetched: 10 row(s)
hive> select Title, Rating from Movie_Ratings WHERE Rating > 8.0 AND Rating < 10.0 AND NumberOfRatings > 500 ORDER BY Rating DESC limit 10;
2023-12-16 18:30:58,226 INFO  [bedf0a12-5e1f-472c-a19b-464d5215791b main] reducesink.VectorReduceSinkObjectHashOperator: VectorReduceSinkObj
.plan.VectorReduceSinkInfo@3c2f310c
Query ID = root_20231216183058_427e8d5f-e3ad-46c2-9b92-70231ac1792c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702748408152_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.51 s
----------------------------------------------------------------------------------------------
OK
The Shawshank Redemption        9.3
The Godfather   9.2
Paint Drying    9.2
The Lord of the Rings: The Return of the King   9.0
The Dark Knight 9.0
12 Angry Men    9.0
Schindler's List        9.0
The Godfather Part II   9.0
Pulp Fiction    8.9
Inception       8.8
Time taken: 7.396 seconds, Fetched: 10 row(s)
hive> 
```

Lastly, the final query involves pulling the gross worldwide to see which movies have the highest gross. This quick analysis is meant to discover if the movies with the best rating have equally high world-wide gross. Based on the data and filters, there are not the same top 10 for ratings and revenue.

Query 3



```
[hive> select Title, GrossWorldWide from Movie_Ratings ORDER BY GrossWorldWide DESC Limit 10;
2023-12-16 18:33:06,427 INFO  [bedf0a12-5e1f-472c-a19b-464d5215791b main] reducesink.VectorReduceSink
.plan.VectorReduceSinkInfo@3833897c
Query ID = root_20231216183306_ce3baece-24a4-4225-9520-a66adcc1814c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1702748408152_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.61 s
----------------------------------------------------------------------------------------------
OK
Avatar  2923.706
Avengers: Endgame       2799.439
Avatar: The Way of Water        2320.25
Titanic 2264.743
Star Wars: Episode VII - The Force Awakens      2071.31
Avengers: Infinity War  2052.415
Spider-Man: No Way Home 1921.847
Jurassic World  1671.537
The Lion King   1663.075
The Avengers    1520.539
Time taken: 6.473 seconds, Fetched: 10 row(s)
hive> 
```