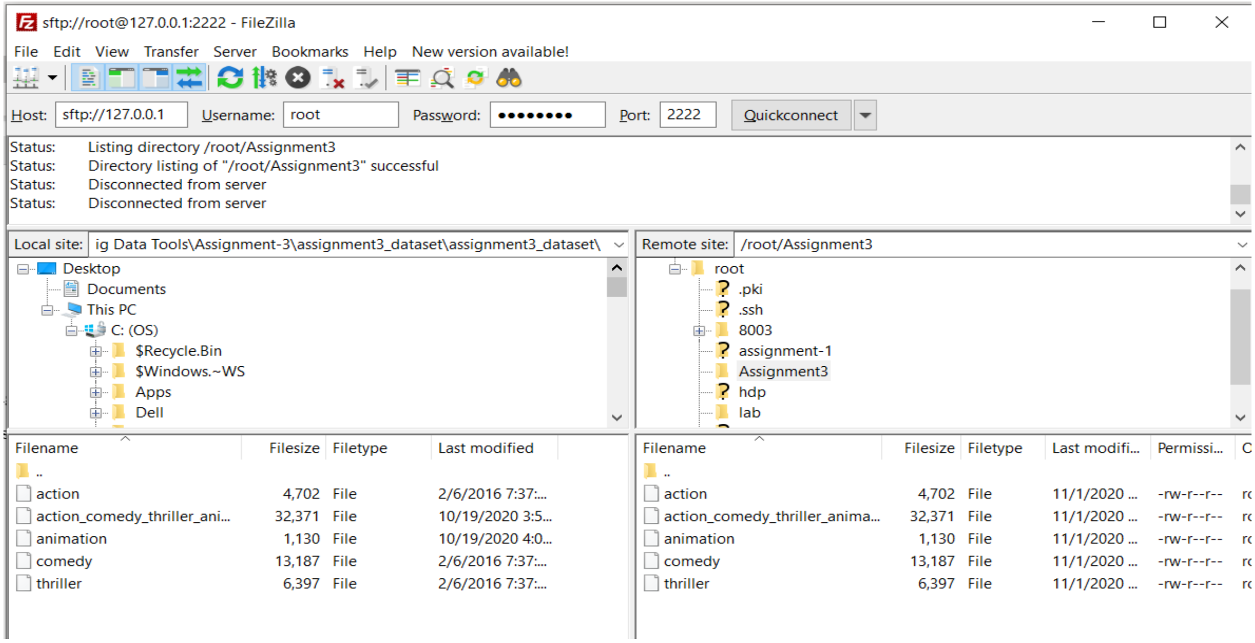
Dataset loading:

A new directory called Assignment3 is created in Linux under ‘/root/’

Provided datasets are transferred to the location ‘/root/Assignment3’ using FileZilla.



Host: sftp://127.0.0.1 Username: root Password: •••••••• Port: 2222

```
Status:     Listing directory /root/Assignment3
Status:     Directory listing of "/root/Assignment3" successful
Status:     Disconnected from server
Status:     Disconnected from server
```

Local site: ig Data Tools\Assignment-3\assignment3_dataset\assignment3_dataset\

Remote site: /root/Assignment3

Filename	Filesize	Filetype	Last modified
..			
action	4,702	File	2/6/2016 7:37:...
action_comedy_thriller_ani...	32,371	File	10/19/2020 3:5...
animation	1,130	File	10/19/2020 4:0...
comedy	13,187	File	2/6/2016 7:37:...
thriller	6,397	File	2/6/2016 7:37:...

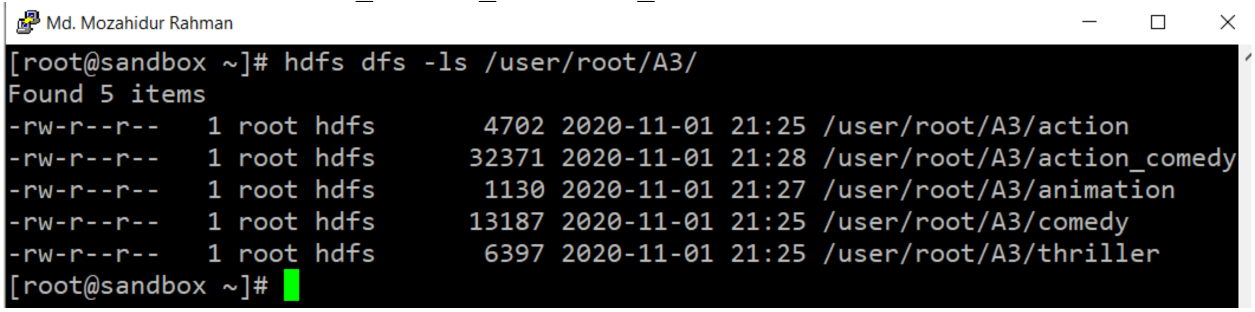
Filename	Filesize	Filetype	Last modifi...	Permissi...	C
..					
action	4,702	File	11/1/2020 ...	-rw-r--r--	rc
action_comedy_thriller_anima...	32,371	File	11/1/2020 ...	-rw-r--r--	rc
animation	1,130	File	11/1/2020 ...	-rw-r--r--	rc
comedy	13,187	File	11/1/2020 ...	-rw-r--r--	rc
thriller	6,397	File	11/1/2020 ...	-rw-r--r--	rc

A new directory ‘A3’ is created in Hadoop under ‘/user/root/’

```
[root@sandbox ~]# hdfs dfs -mkdir /user/root/A3
```

All data files are transferred to the A3 folder using the following command.

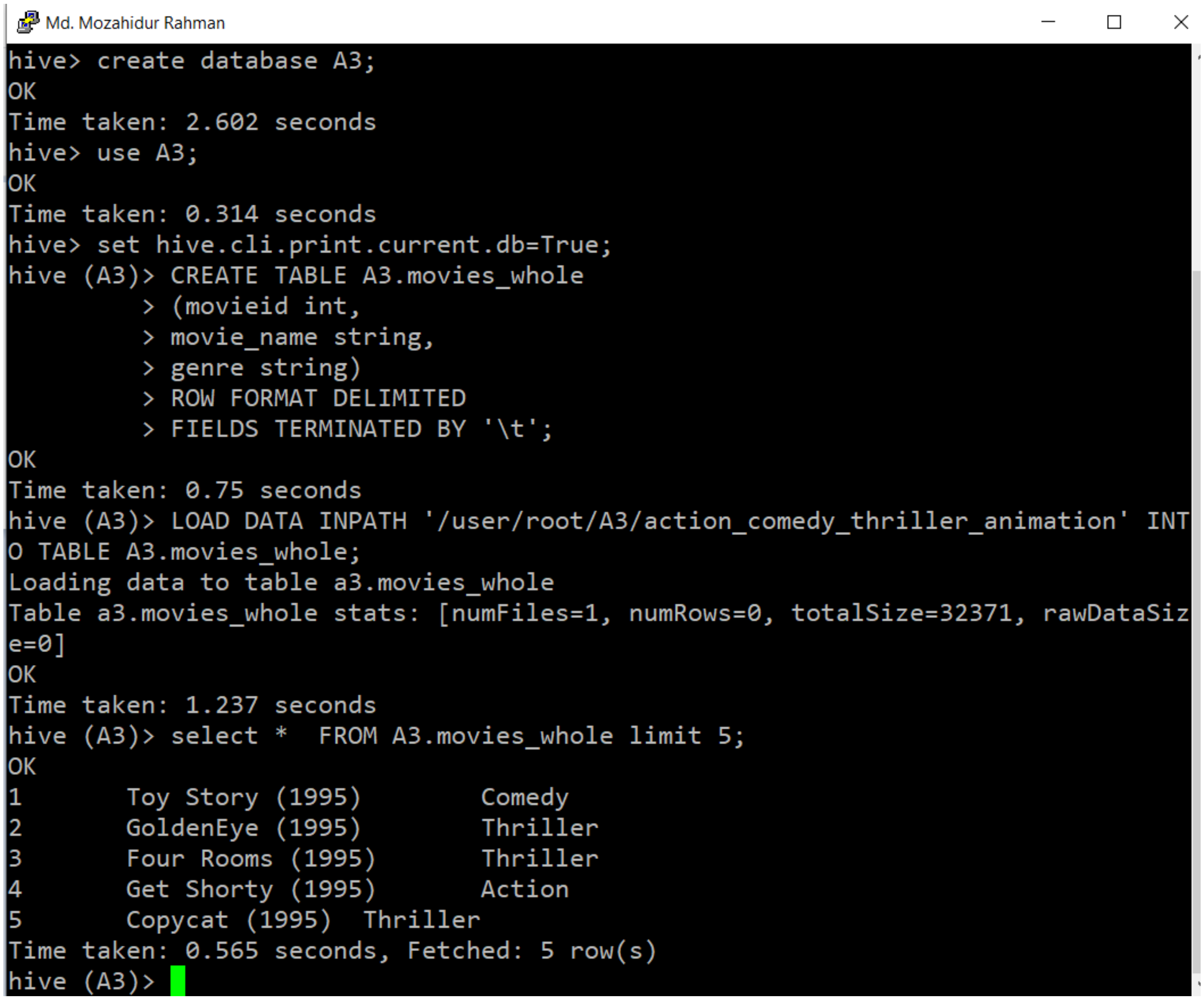
```
hdfs dfs -put /root/Assignment3/action /user/root/A3/action
hdfs dfs -put /root/Assignment3/comedy /user/root/A3/comedy
hdfs dfs -put /root/Assignment3/thriller /user/root/A3/thriller
hdfs dfs -put /root/Assignment3/animation /user/root/A3/animation
hdfs dfs -put /root/Assignment3/action_comedy_thriller_animation
/user/root/A3/action_comedy_thriller_animation
```



```
[root@sandbox ~]# hdfs dfs -ls /user/root/A3/
Found 5 items
-rw-r--r--   1 root hdfs       4702 2020-11-01 21:25 /user/root/A3/action
-rw-r--r--   1 root hdfs      32371 2020-11-01 21:28 /user/root/A3/action_comedy
-rw-r--r--   1 root hdfs       1130 2020-11-01 21:27 /user/root/A3/animation
-rw-r--r--   1 root hdfs      13187 2020-11-01 21:25 /user/root/A3/comedy
-rw-r--r--   1 root hdfs       6397 2020-11-01 21:25 /user/root/A3/thriller
[root@sandbox ~]#
```

Store complete information of all movies into a hive table.

```
create database A3;
use A3;
set hive.cli.print.current.db=True;

CREATE TABLE A3.movies_whole
(movieid int,
movie_name string,
genre string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';

LOAD DATA INPATH
'/user/root/A3/action_comedy_thriller_animation' INTO TABLE
A3.movies_whole;
select *  FROM A3.movies_whole limit 5;
```

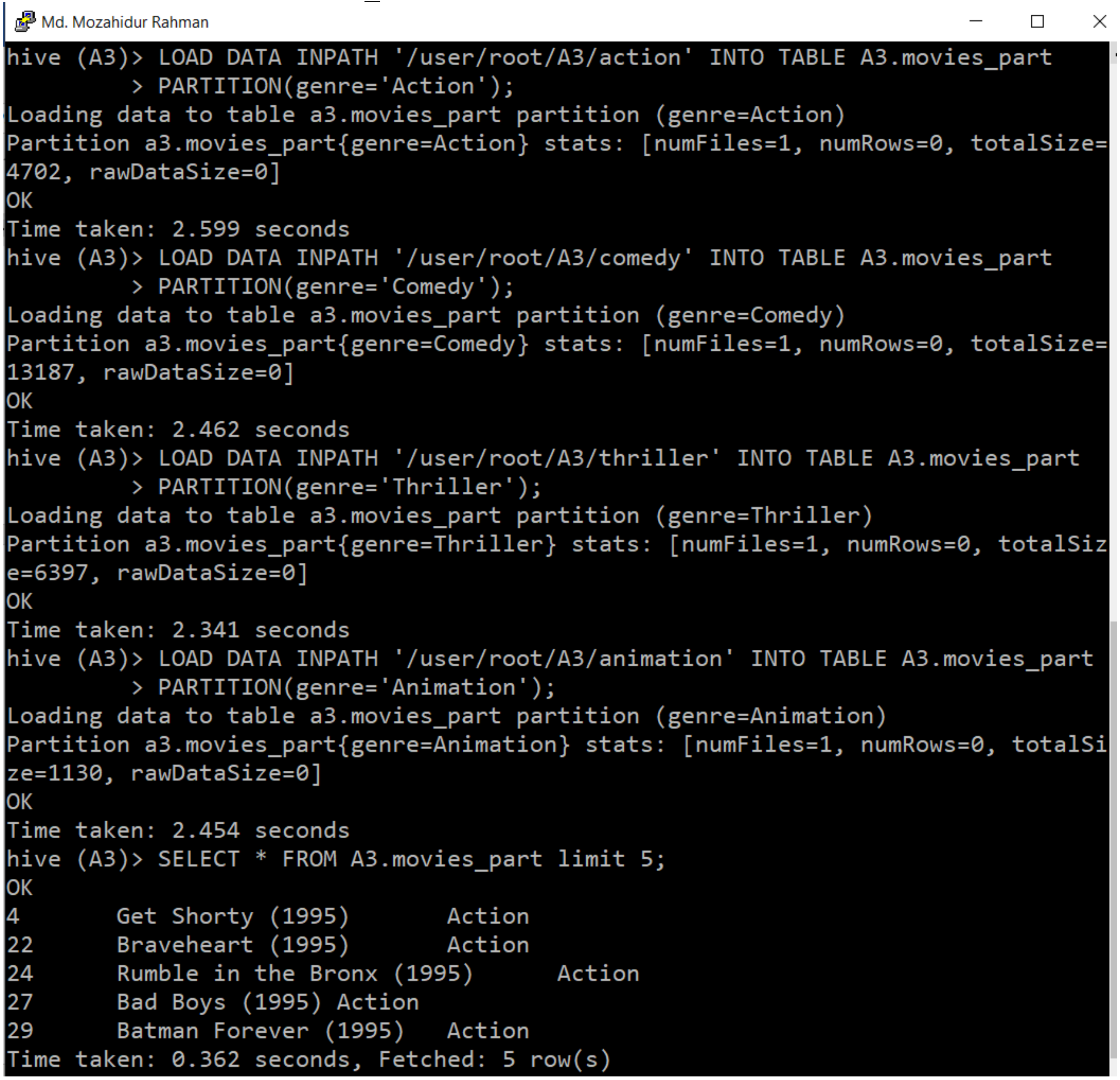


Md. Mozahidur Rahman

```
hive> create database A3;
OK
Time taken: 2.602 seconds
hive> use A3;
OK
Time taken: 0.314 seconds
hive> set hive.cli.print.current.db=True;
hive (A3)> CREATE TABLE A3.movies_whole
         > (movieid int,
         > movie_name string,
         > genre string)
         > ROW FORMAT DELIMITED
         > FIELDS TERMINATED BY '\t';
OK
Time taken: 0.75 seconds
hive (A3)> LOAD DATA INPATH '/user/root/A3/action_comedy_thriller_animation' INT
O TABLE A3.movies_whole;
Loading data to table a3.movies_whole
Table a3.movies_whole stats: [numFiles=1, numRows=0, totalSize=32371, rawDataSiz
e=0]
OK
Time taken: 1.237 seconds
hive (A3)> select *  FROM A3.movies_whole limit 5;
OK
1       Toy Story (1995)        Comedy
2       GoldenEye (1995)        Thriller
3       Four Rooms (1995)       Thriller
4       Get Shorty (1995)       Action
5       Copycat (1995)  Thriller
Time taken: 0.565 seconds, Fetched: 5 row(s)
hive (A3)>
```

Store data into a hive table that is partitioned on genre.

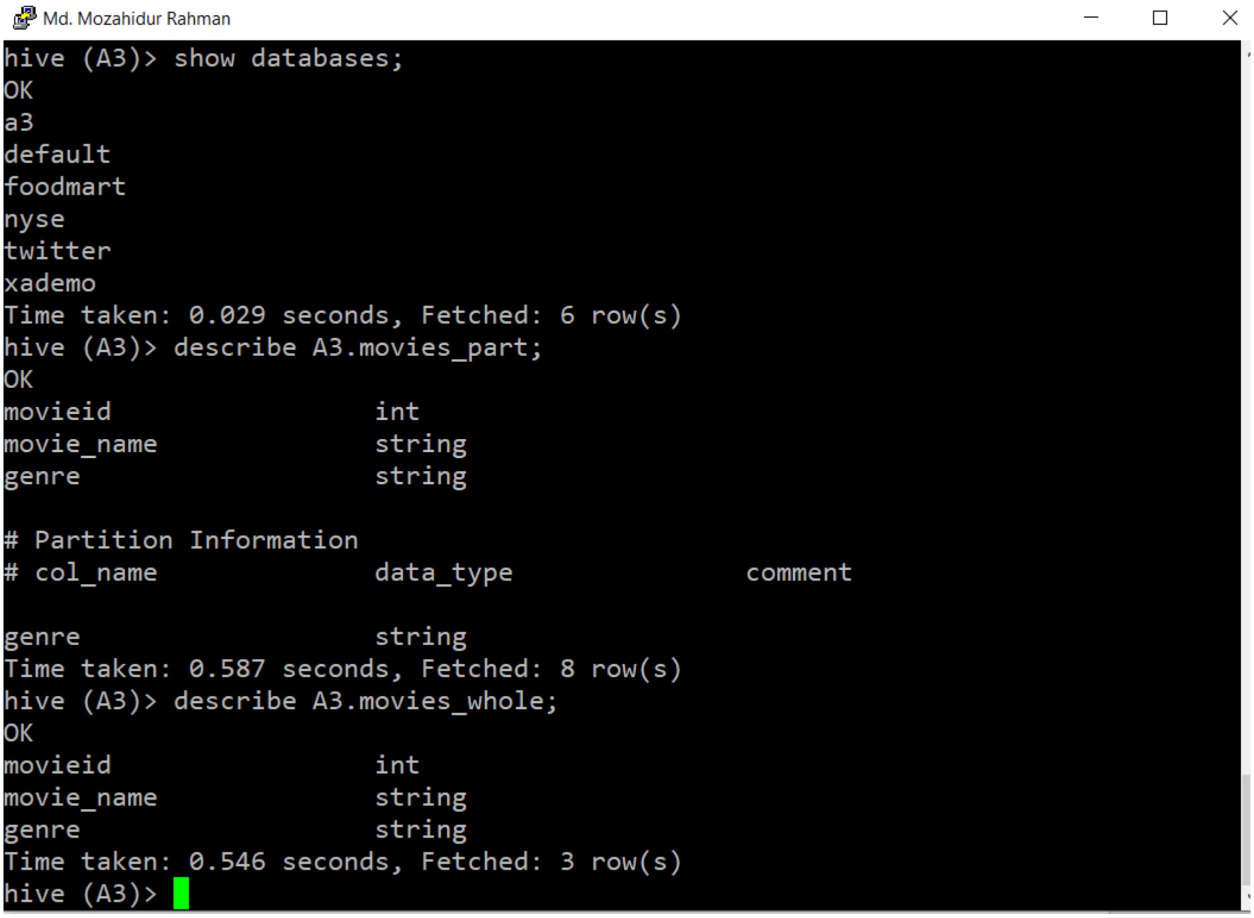
```
CREATE TABLE A3.movies_part (movieid int, movie_name string)
PARTITIONED BY (genre string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
LOAD DATA INPATH '/user/root/A3/action' INTO TABLE A3.movies_part
PARTITION(genre='Action');
LOAD DATA INPATH '/user/root/A3/comedy' INTO TABLE A3.movies_part
PARTITION(genre='Comedy');
LOAD DATA INPATH '/user/root/A3/thriller' INTO TABLE A3.movies_part
PARTITION(genre='Thriller');
LOAD DATA INPATH '/user/root/A3/animation' INTO TABLE A3.movies_part
PARTITION(genre='Animation');
SELECT * FROM A3.movies_part limit 5;
```



```
hive (A3)> LOAD DATA INPATH '/user/root/A3/action' INTO TABLE A3.movies_part
         > PARTITION(genre='Action');
Loading data to table a3.movies_part partition (genre=Action)
Partition a3.movies_part{genre=Action} stats: [numFiles=1, numRows=0, totalSize=
4702, rawDataSize=0]
OK
Time taken: 2.599 seconds
hive (A3)> LOAD DATA INPATH '/user/root/A3/comedy' INTO TABLE A3.movies_part
         > PARTITION(genre='Comedy');
Loading data to table a3.movies_part partition (genre=Comedy)
Partition a3.movies_part{genre=Comedy} stats: [numFiles=1, numRows=0, totalSize=
13187, rawDataSize=0]
OK
Time taken: 2.462 seconds
hive (A3)> LOAD DATA INPATH '/user/root/A3/thriller' INTO TABLE A3.movies_part
         > PARTITION(genre='Thriller');
Loading data to table a3.movies_part partition (genre=Thriller)
Partition a3.movies_part{genre=Thriller} stats: [numFiles=1, numRows=0, totalSiz
e=6397, rawDataSize=0]
OK
Time taken: 2.341 seconds
hive (A3)> LOAD DATA INPATH '/user/root/A3/animation' INTO TABLE A3.movies_part
         > PARTITION(genre='Animation');
Loading data to table a3.movies_part partition (genre=Animation)
Partition a3.movies_part{genre=Animation} stats: [numFiles=1, numRows=0, totalSi
ze=1130, rawDataSize=0]
OK
Time taken: 2.454 seconds
hive (A3)> SELECT * FROM A3.movies_part limit 5;
OK
4       Get Shorty (1995)       Action
22      Braveheart (1995)       Action
24      Rumble in the Bronx (1995)      Action
27      Bad Boys (1995) Action
29      Batman Forever (1995)   Action
Time taken: 0.362 seconds, Fetched: 5 row(s)
```

Show database and table structures

```
show databases;
describe A3.movies_part;
describe A3.movies_whole;
```



```
hive (A3)> show databases;
OK
a3
default
foodmart
nyse
twitter
xademo
Time taken: 0.029 seconds, Fetched: 6 row(s)
hive (A3)> describe A3.movies_part;
OK
movieid                 int
movie_name              string
genre                   string

# Partition Information
# col_name              data_type               comment

genre                   string
Time taken: 0.587 seconds, Fetched: 8 row(s)
hive (A3)> describe A3.movies_whole;
OK
movieid                 int
movie_name              string
genre                   string
Time taken: 0.546 seconds, Fetched: 3 row(s)
hive (A3)>
```

Question 1: Write the following queries, report results and execution time on both partitioned and complete data:

a) Write a hive query to select last 50 distinct records from *table* after sorting it by movie_name in descending order

Query: (for the whole data)

```
SELECT distinct * FROM A3.movies_whole ORDER BY movie_name
desc LIMIT 50;
```

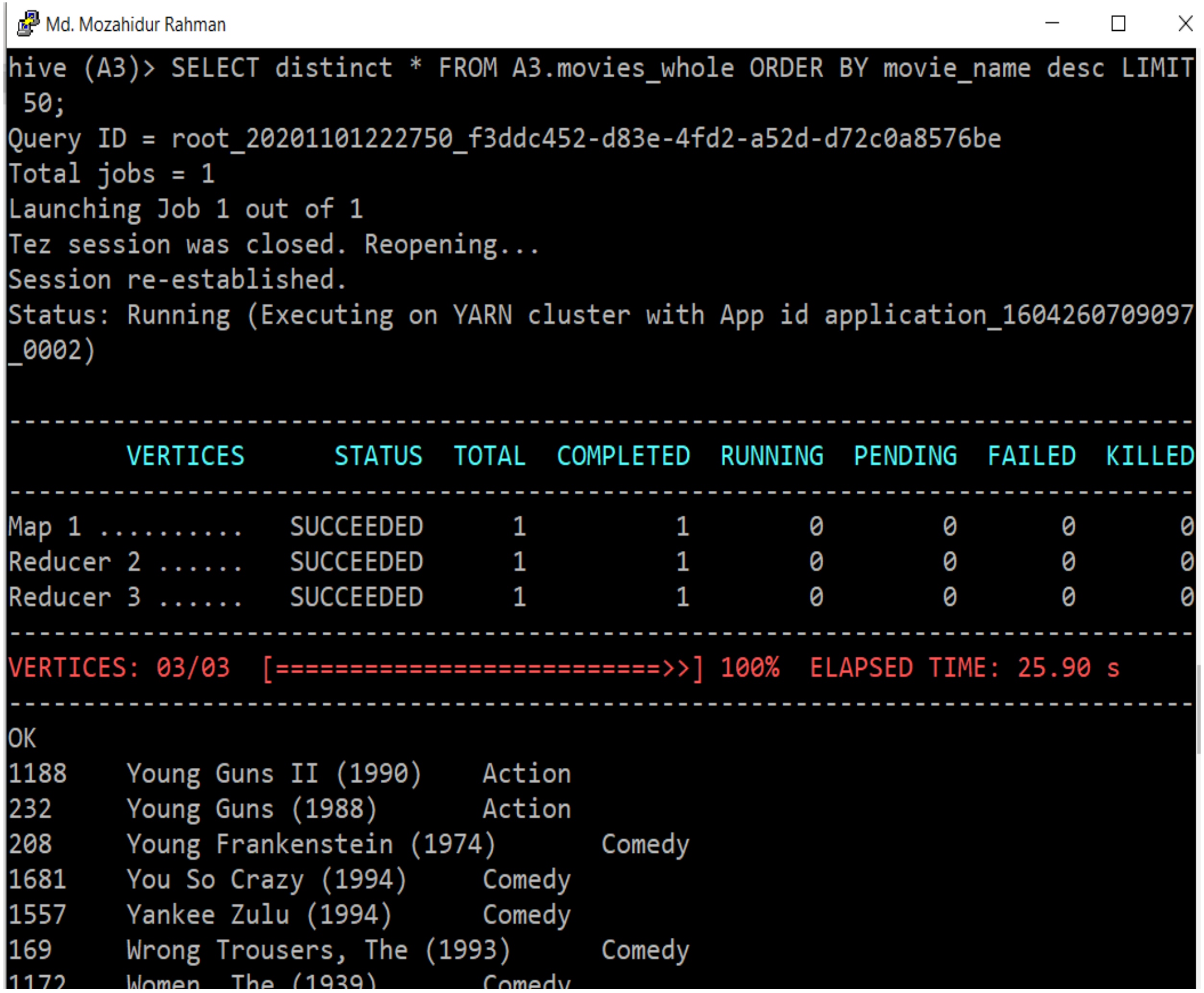
(Execution time = 5.883 seconds)

Result:

```
1188    Young Guns II (1990)    Action
232     Young Guns (1988)       Action
208     Young Frankenstein (1974)       Comedy
1681    You So Crazy (1994)     Comedy
1557    Yankee Zulu (1994)      Comedy
169     Wrong Trousers, The (1993)      Comedy
1172    Women, The (1939)       Comedy
1707    Wolf and Calf (1984)    Animation
1526    Witness (1985)  Thriller
1404    Withnail and I (1987)   Comedy
941     With Honors (1994)      Comedy
1704    Winter in Prostokvashino (1984) Animation
1699    Winnie the Pooh and the Day of Concern (1972)   Animation
1698    Winnie the Pooh Goes Visiting (1971)    Animation
311     Wings of the Dove, The (1997)   Thriller
512     Wings of Desire (1987)  Comedy
1492    Window to Paris (1994)  Comedy
151     Willy Wonka and the Chocolate Factory (1971)    Comedy
914     Wild Things (1998)      Thriller
1627    Wife, The (1995)        Comedy
66      While You Were Sleeping (1995)  Comedy
256     When the Cats Away (Chacun cherche son chat) (1996)     Comedy
216     When Harry Met Sally... (1989)  Comedy
65      What's Eating Gilbert Grape (1993)      Comedy
1100    What Happened Was... (1994)     Comedy
93      Welcome to the Dollhouse (1995) Comedy
158     Weekend at Bernie's (1989)      Comedy
354     Wedding Singer, The (1998)      Comedy
1668    Wedding Bell Blues (1996)       Comedy
554     Waterworld (1995)       Action
1615    Warriors of Virtue (1997)       Action
1311    Waiting to Exhale (1995)        Comedy
1007    Waiting for Guffman (1996)      Comedy
347     Wag the Dog (1997)      Comedy
```

```
678     Volcano (1997)  Thriller
1472    Visitors, The (Visiteurs, Les) (1993)   Comedy
1210    Virtuosity (1995)       Thriller
565     Village of the Damned (1995)    Thriller
1569    Vie est belle, La (Life is Rosey) (1987)        Comedy
629     Victor/Victoria (1982)  Comedy
412     Very Brady Sequel, A (1996)     Comedy
479     Vertigo (1958)  Thriller
1568    Vermont Is For Lovers (1992)    Comedy
907     Vermin (1998)   Comedy
871     Vegas Vacation (1997)   Comedy
1241    Van, The (1996) Comedy
545     Vampire in Brooklyn (1995)      Comedy
1703    Vacations in Prostokvashino (1980)      Animation
12      Usual Suspects, The (1995)      Thriller
1118    Up in Smoke (1978)      Comedy
```

Screenshot: (Partial as the output is long)



```
hive (A3)> SELECT distinct * FROM A3.movies_whole ORDER BY movie_name desc LIMIT
 50;
Query ID = root_20201101222750_f3ddc452-d83e-4fd2-a52d-d72c0a8576be
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_1604260709097
_0002)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.90 s
----------------------------------------------------------------------------------
OK
1188    Young Guns II (1990)    Action
232     Young Guns (1988)       Action
208     Young Frankenstein (1974)       Comedy
1681    You So Crazy (1994)     Comedy
1557    Yankee Zulu (1994)      Comedy
169     Wrong Trousers, The (1993)      Comedy
1172    Women, The (1939)       Comedy
```

Query for the partitioned data:

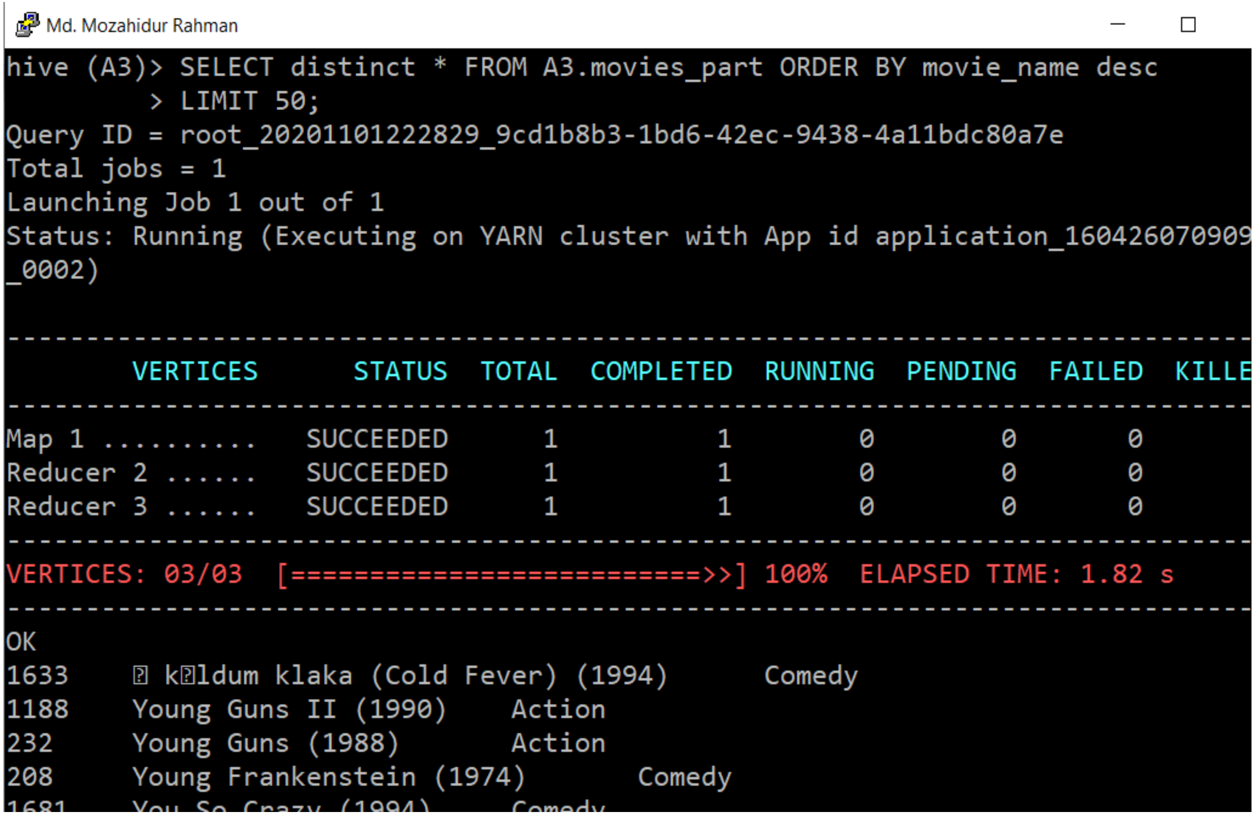
```
SELECT distinct * FROM A3.movies_part ORDER BY movie_name desc
LIMIT 50;
```

(Execution time = 5.836 seconds)

```
Result:
1633    � k�ldum klaka (Cold Fever) (1994)      Comedy
1188    Young Guns II (1990)    Action
232     Young Guns (1988)       Action
208     Young Frankenstein (1974)       Comedy
1681    You So Crazy (1994)     Comedy
1557    Yankee Zulu (1994)      Comedy
169     Wrong Trousers, The (1993)      Comedy
1172    Women, The (1939)       Comedy
1707    Wolf and Calf (1984)    Animation
1526    Witness (1985)  Thriller
1404    Withnail and I (1987)   Comedy
941     With Honors (1994)      Comedy
1704    Winter in Prostokvashino (1984) Animation
1699    Winnie the Pooh and the Day of Concern (1972)
Animation
1698    Winnie the Pooh Goes Visiting (1971)    Animation
311     Wings of the Dove, The (1997)   Thriller
512     Wings of Desire (1987)  Comedy
1492    Window to Paris (1994)  Comedy
151     Willy Wonka and the Chocolate Factory (1971)    Comedy
914     Wild Things (1998)      Thriller
1627    Wife, The (1995)        Comedy
66      While You Were Sleeping (1995)  Comedy
256     When the Cats Away (Chacun cherche son chat) (1996)
Comedy
216     When Harry Met Sally... (1989)  Comedy
65      What's Eating Gilbert Grape (1993)      Comedy
1100    What Happened Was... (1994)     Comedy
93      Welcome to the Dollhouse (1995) Comedy
158     Weekend at Bernie's (1989)      Comedy
354     Wedding Singer, The (1998)      Comedy
1668    Wedding Bell Blues (1996)       Comedy
554     Waterworld (1995)       Action
1615    Warriors of Virtue (1997)       Action
1311    Waiting to Exhale (1995)        Comedy
1007    Waiting for Guffman (1996)      Comedy
347     Wag the Dog (1997)      Comedy
678     Volcano (1997)  Thriller
1472    Visitors, The (Visiteurs, Les) (1993)   Comedy
1210    Virtuosity (1995)       Thriller
```

```
565     Village of the Damned (1995)    Thriller
1569    Vie est belle, La (Life is Rosey) (1987)        Comedy
629     Victor/Victoria (1982)  Comedy
412     Very Brady Sequel, A (1996)     Comedy
479     Vertigo (1958)  Thriller
1568    Vermont Is For Lovers (1992)    Comedy
907     Vermin (1998)   Comedy
871     Vegas Vacation (1997)   Comedy
1241    Van, The (1996) Comedy
545     Vampire in Brooklyn (1995)      Comedy
1703    Vacations in Prostokvashino (1980)      Animation
12      Usual Suspects, The (1995)      Thriller
```

Screenshot: (partial)



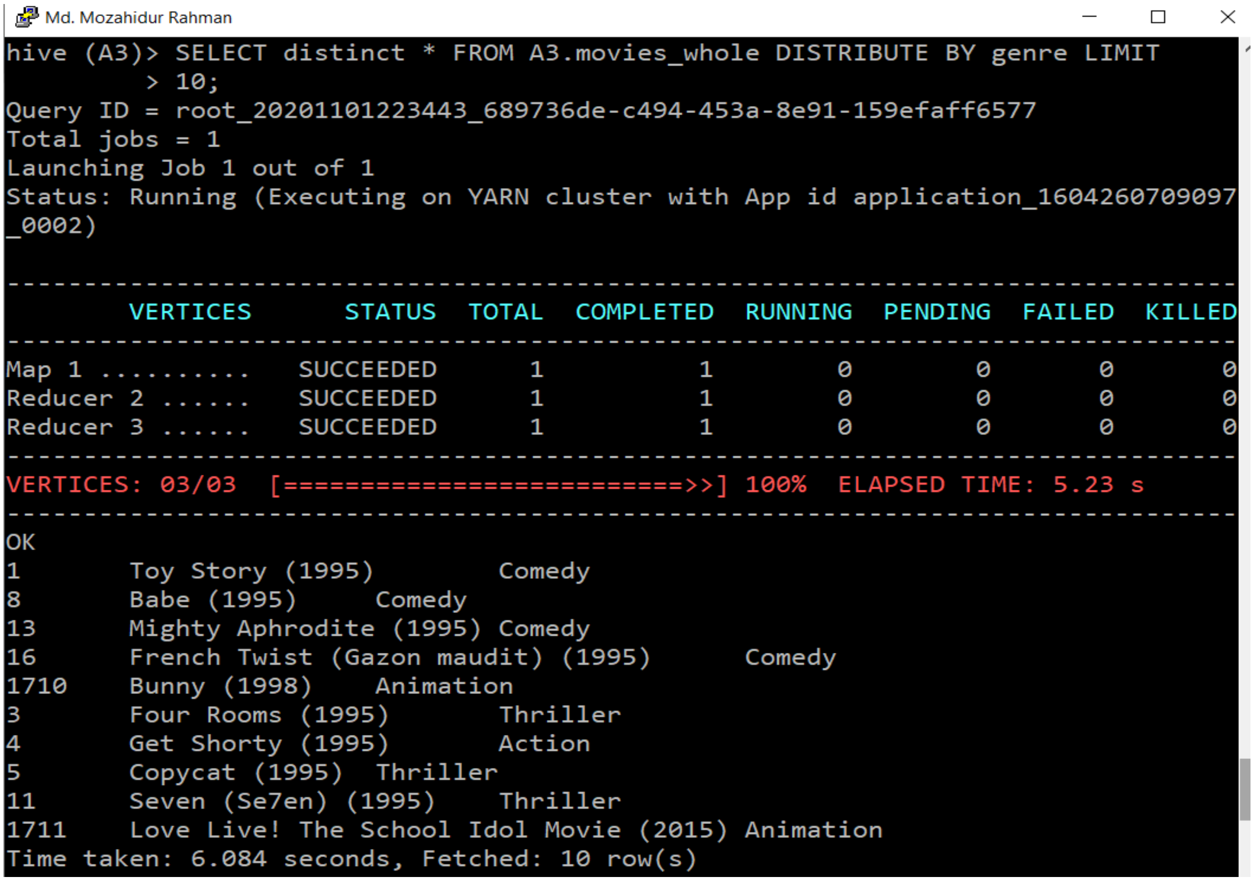
Md. Mozahidur Rahman

```
hive (A3)> SELECT distinct * FROM A3.movies_part ORDER BY movie_name desc
         > LIMIT 50;
Query ID = root_20201101222829_9cd1b8b3-1bd6-42ec-9438-4a11bdc80a7e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_160426070909
_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLE
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 1.82 s
--------------------------------------------------------------------------------
OK
1633    ⍰ k⍰ldum klaka (Cold Fever) (1994)      Comedy
1188    Young Guns II (1990)    Action
232     Young Guns (1988)       Action
208     Young Frankenstein (1974)       Comedy
1681    You So Crazy (1994)     Comedy
```

b) Write a hive query to select to 10 distinct records from *table* after distributing it by genre

Query for the whole data: (Execution time = 6.084 seconds)

```
SELECT distinct * FROM A3.movies_whole DISTRIBUTE BY genre LIMIT
10;

Result:

1       Toy Story (1995)        Comedy
8       Babe (1995)     Comedy
13      Mighty Aphrodite (1995) Comedy
16      French Twist (Gazon maudit) (1995)      Comedy
1710    Bunny (1998)    Animation
3       Four Rooms (1995)       Thriller
4       Get Shorty (1995)       Action
5       Copycat (1995)  Thriller
11      Seven (Se7en) (1995)    Thriller
1711    Love Live! The School Idol Movie (2015) Animation


Screenshots:
```



```
hive (A3)> SELECT distinct * FROM A3.movies_whole DISTRIBUTE BY genre LIMIT
         > 10;
Query ID = root_20201101223443_689736de-c494-453a-8e91-159efaff6577
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.23 s
--------------------------------------------------------------------------------
OK
1       Toy Story (1995)        Comedy
8       Babe (1995)     Comedy
13      Mighty Aphrodite (1995) Comedy
16      French Twist (Gazon maudit) (1995)      Comedy
1710    Bunny (1998)    Animation
3       Four Rooms (1995)       Thriller
4       Get Shorty (1995)       Action
5       Copycat (1995)  Thriller
11      Seven (Se7en) (1995)    Thriller
1711    Love Live! The School Idol Movie (2015) Animation
Time taken: 6.084 seconds, Fetched: 10 row(s)
```

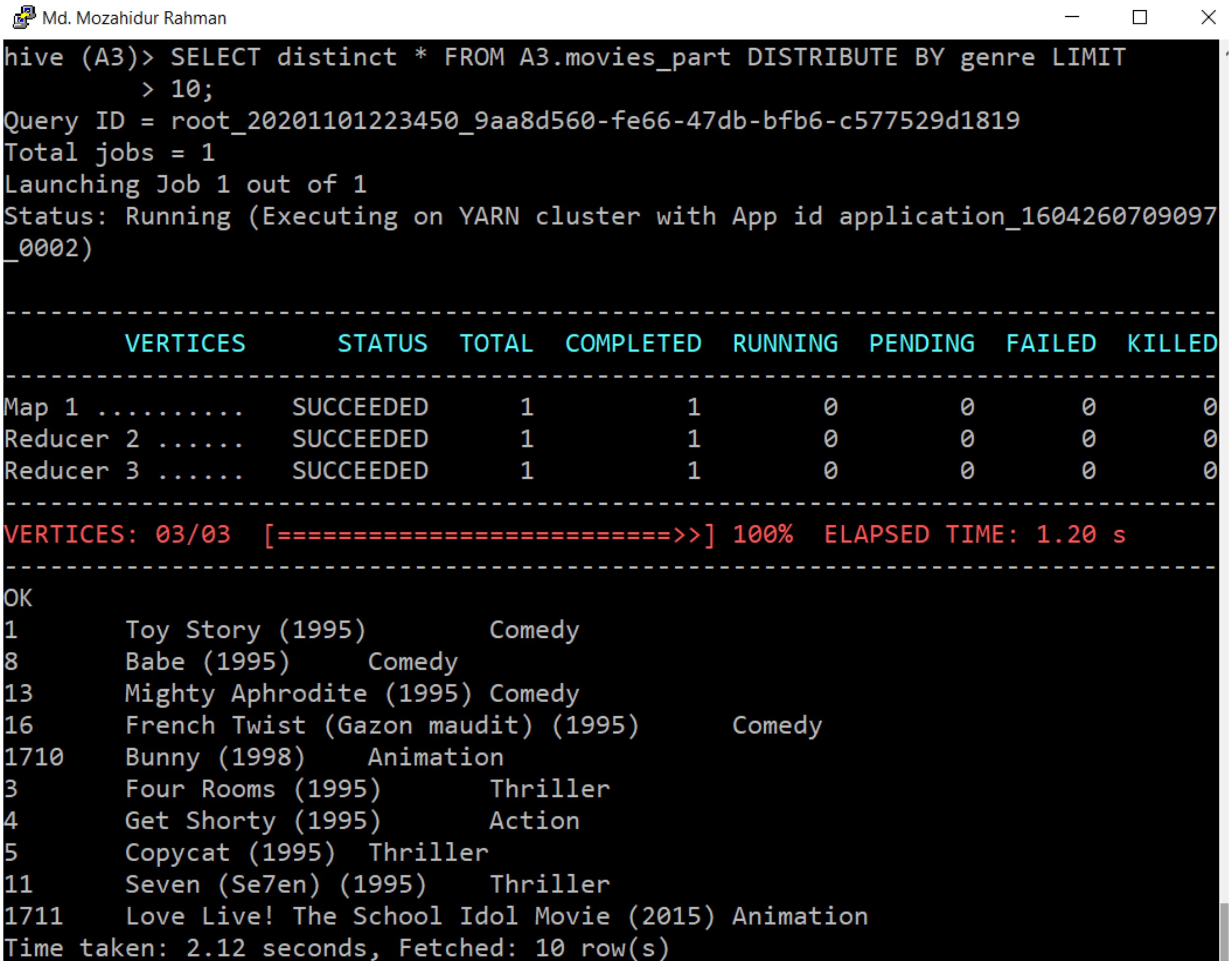
Query on the partitioned data: (Execution time = 2.12 seconds)

```
SELECT distinct * FROM A3.movies_part DISTRIBUTE BY genre LIMIT
10;
```

Result:

```
1       Toy Story (1995)        Comedy
8       Babe (1995)     Comedy
13      Mighty Aphrodite (1995) Comedy
16      French Twist (Gazon maudit) (1995)      Comedy
1710    Bunny (1998)    Animation
3       Four Rooms (1995)       Thriller
4       Get Shorty (1995)       Action
5       Copycat (1995)  Thriller
11      Seven (Se7en) (1995)    Thriller
1711    Love Live! The School Idol Movie (2015) Animation
```

Scrennshot:



```
hive (A3)> SELECT distinct * FROM A3.movies_part DISTRIBUTE BY genre LIMIT
         > 10;
Query ID = root_20201101223450_9aa8d560-fe66-47db-bfb6-c577529d1819
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 1.20 s
--------------------------------------------------------------------------------
OK
1       Toy Story (1995)        Comedy
8       Babe (1995)     Comedy
13      Mighty Aphrodite (1995) Comedy
16      French Twist (Gazon maudit) (1995)      Comedy
1710    Bunny (1998)    Animation
3       Four Rooms (1995)       Thriller
4       Get Shorty (1995)       Action
5       Copycat (1995)  Thriller
11      Seven (Se7en) (1995)    Thriller
1711    Love Live! The School Idol Movie (2015) Animation
Time taken: 2.12 seconds, Fetched: 10 row(s)
```

c) Write a hive query to count movies released by years from *table*

Query: (Whole data)

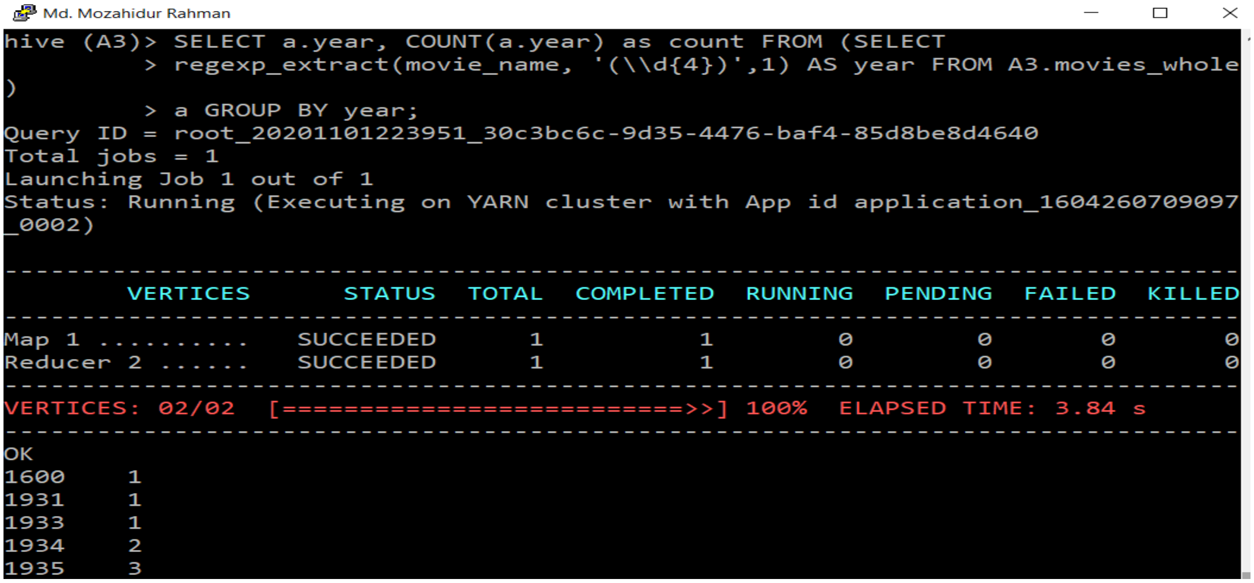
```
SELECT a.year, COUNT(a.year) as count FROM (SELECT
regexp_extract(movie_name, '(\\d{4})',1) AS year FROM A3.movies_whole)
a GROUP BY year;
```

Result:

```
 1600    1    1952    1        1972    3          1991    11
 1931    1    1953    2        1973    2          1992    23
 1933    1    1954    5        1974    6          1993    69
 1934    2    1955    1        1975    3          1994    133
 1935    3    1956    1        1976    1          1995    145
 1936    1    1957    2        1977    2          1996    163
 1937    2    1958    3        1978    4          1997    133
 1938    2    1959    4        1979    5          1998    32
 1939    2    1960    4        1980    5          2000    1
 1940    5    1961    1        1981    10         2001    1
 1942    1    1962    2        1982    4          2003    1
 1943    1    1963    1        1983    4          2007    2
 1944    3    1964    2        1984    6          2009    5
 1945    1    1965    3        1985    4          2010    1
 1946    1    1966    1        1986    9          2014    1
 1948    1    1967    1        1987    11         2015    2
 1949    1    1969    2        1988    6          2016    3
 1950    1    1970    1        1989    9          2017    2
 1951    1    1971    5        1990    13         3000    1
```

Time taken: 4.848 seconds, Fetched: 76 row(s)

Screenshots: (partial)



```
hive (A3)> SELECT a.year, COUNT(a.year) as count FROM (SELECT
         > regexp_extract(movie_name, '(\\d{4})',1) AS year FROM A3.movies_whole
)
         > a GROUP BY year;
Query ID = root_20201101223951_30c3bc6c-9d35-4476-baf4-85d8be8d4640
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0002)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.84 s
----------------------------------------------------------------------------------
OK
1600    1
1931    1
1933    1
1934    2
1935    3
```

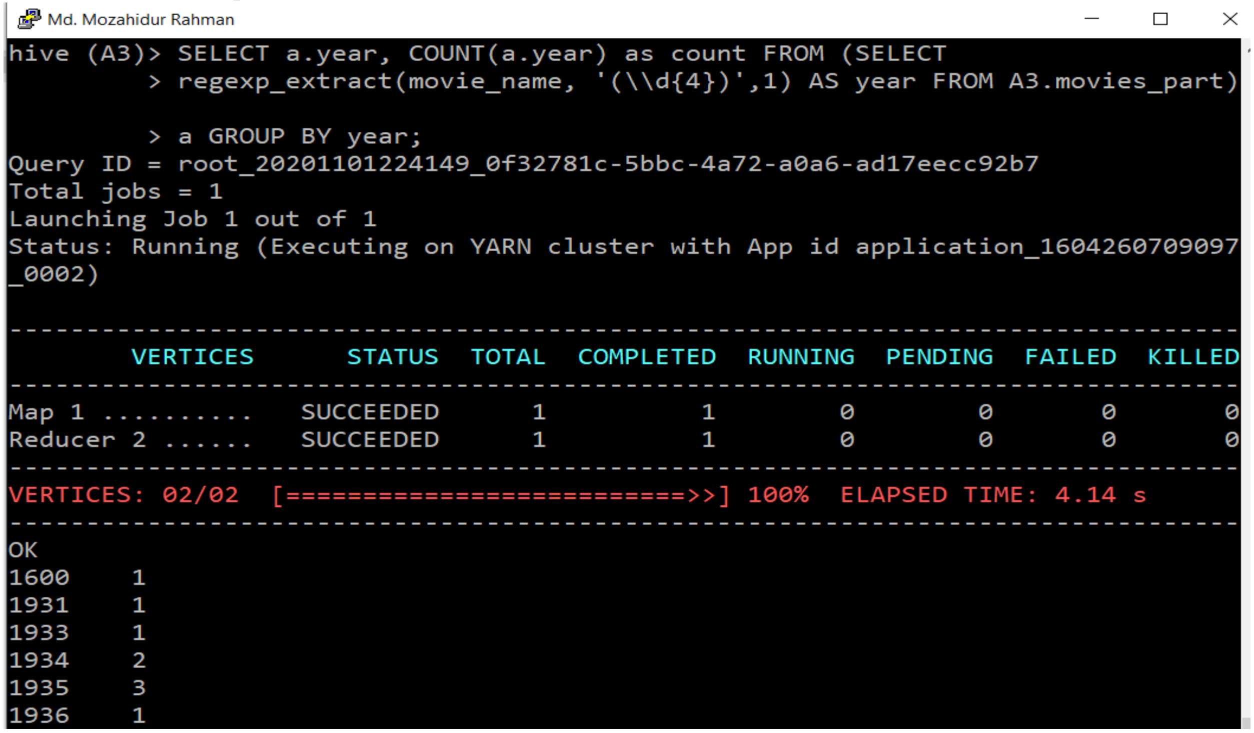
Query (Part data)

```
SELECT a.year, COUNT(a.year) as count FROM (SELECT
regexp_extract(movie_name, '(\\d{4})',1) AS year FROM
A3.movies_part)a GROUP BY year;
```

```
Result:
 1600    1       1952    1        1972    3          1991    11
 1931    1       1953    2        1973    2          1992    23
 1933    1       1954    5        1974    6          1993    69
 1934    2       1955    1        1975    3          1994    133
 1935    3       1956    1        1976    1          1995    145
 1936    1       1957    2        1977    2          1996    163
 1937    2       1958    3        1978    4          1997    133
 1938    2       1959    4        1979    5          1998    32
 1939    2       1960    4        1980    5          2000    1
 1940    5       1961    1        1981    10         2001    1
 1942    1       1962    2        1982    4          2003    1
 1943    1       1963    1        1983    4          2007    2
 1944    3       1964    2        1984    6          2009    5
 1945    1       1965    3        1985    4          2010    1
 1946    1       1966    1        1986    9          2014    1
 1948    1       1967    1        1987    11         2015    2
 1949    1       1969    2        1988    6          2016    3
 1950    1       1970    1        1989    9          2017    2
 1951    1       1971    5        1990    13         3000    1
```

Time taken: 4.831 seconds, Fetched: 76 row(s)

Screenshot: (partial)

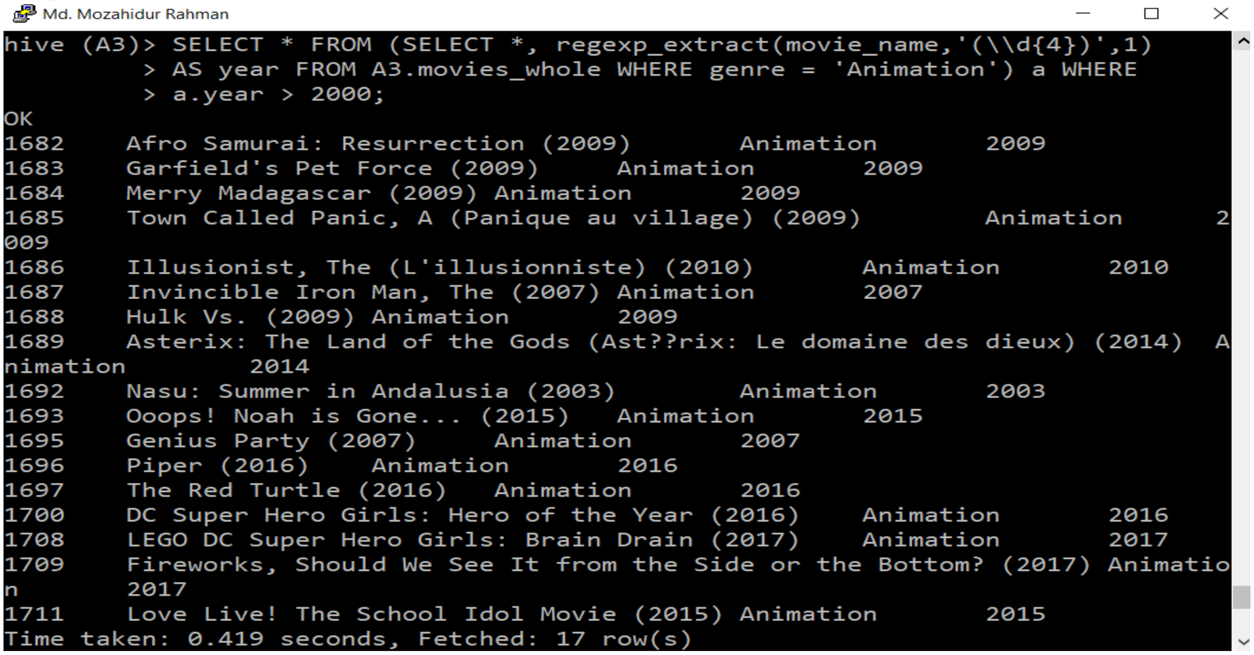


```
hive (A3)> SELECT a.year, COUNT(a.year) as count FROM (SELECT
         > regexp_extract(movie_name, '(\\d{4})',1) AS year FROM A3.movies_part)

         > a GROUP BY year;
Query ID = root_20201101224149_0f32781c-5bbc-4a72-a0a6-ad17eecc92b7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0002)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.14 s
----------------------------------------------------------------------------------
OK
1600    1
1931    1
1933    1
1934    2
1935    3
1936    1
```

d) Write a hive query to find all animation movies released after year 2000 from *table*

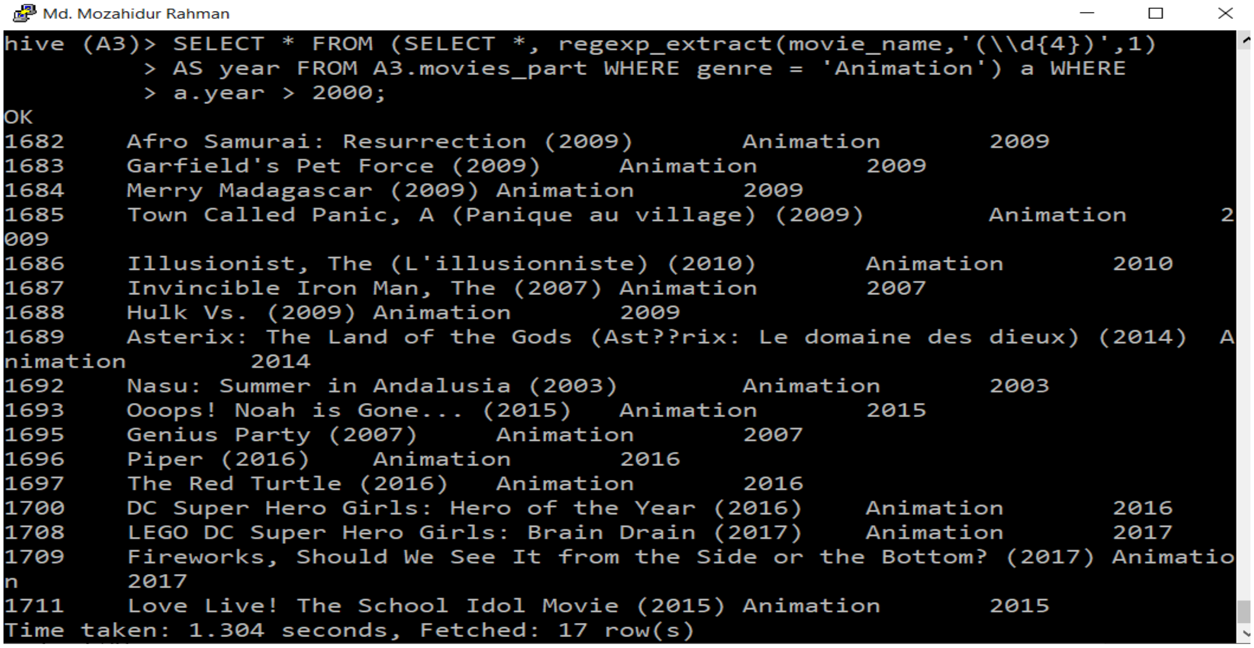
Query and screenshots: (Execution times are shown in the screenshots)

```
SELECT * FROM (SELECT *, regexp_extract(movie_name,'(\\d{4})',1)
AS year FROM A3.movies_whole WHERE genre = 'Animation') a WHERE
a.year > 2000;
```



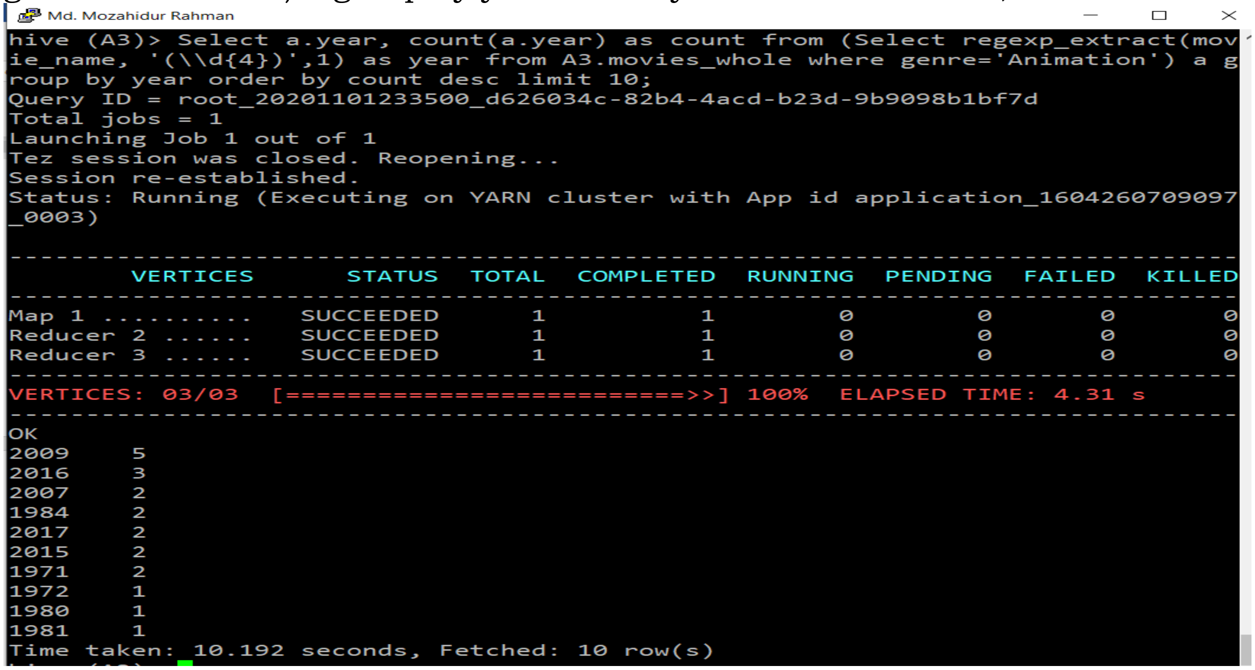
```
hive (A3)> SELECT * FROM (SELECT *, regexp_extract(movie_name,'(\\d{4})',1)
         > AS year FROM A3.movies_whole WHERE genre = 'Animation') a WHERE
         > a.year > 2000;
OK
1682    Afro Samurai: Resurrection (2009)       Animation       2009
1683    Garfield's Pet Force (2009)     Animation       2009
1684    Merry Madagascar (2009) Animation       2009
1685    Town Called Panic, A (Panique au village) (2009)        Animation       2
009
1686    Illusionist, The (L'illusionniste) (2010)       Animation       2010
1687    Invincible Iron Man, The (2007) Animation       2007
1688    Hulk Vs. (2009) Animation       2009
1689    Asterix: The Land of the Gods (Ast??rix: Le domaine des dieux) (2014)   A
nimation        2014
1692    Nasu: Summer in Andalusia (2003)        Animation       2003
1693    Ooops! Noah is Gone... (2015)   Animation       2015
1695    Genius Party (2007)     Animation       2007
1696    Piper (2016)    Animation       2016
1697    The Red Turtle (2016)   Animation       2016
1700    DC Super Hero Girls: Hero of the Year (2016)    Animation       2016
1708    LEGO DC Super Hero Girls: Brain Drain (2017)    Animation       2017
1709    Fireworks, Should We See It from the Side or the Bottom? (2017) Animatio
n       2017
1711    Love Live! The School Idol Movie (2015) Animation       2015
Time taken: 0.419 seconds, Fetched: 17 row(s)
```

```
SELECT * FROM (SELECT *, regexp_extract(movie_name,'(\\d{4})',1)
AS year FROM A3.movies_part WHERE genre = 'Animation') a WHERE
a.year > 2000;
```



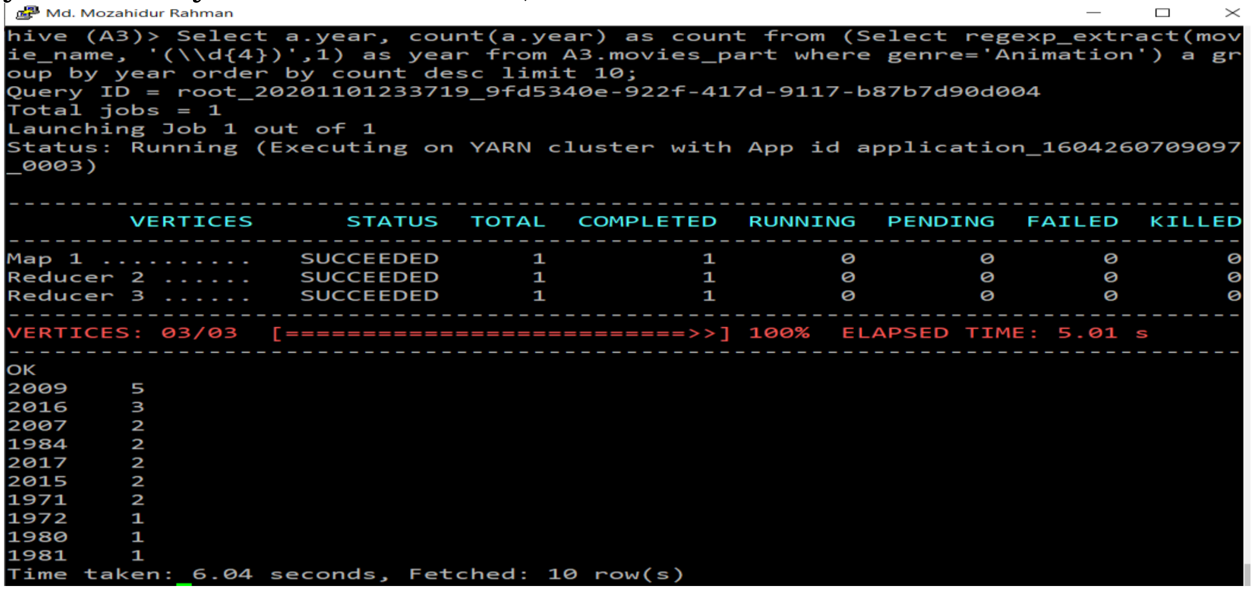
```
hive (A3)> SELECT * FROM (SELECT *, regexp_extract(movie_name,'(\\d{4})',1)
         > AS year FROM A3.movies_part WHERE genre = 'Animation') a WHERE
         > a.year > 2000;
OK
1682    Afro Samurai: Resurrection (2009)       Animation       2009
1683    Garfield's Pet Force (2009)     Animation       2009
1684    Merry Madagascar (2009) Animation       2009
1685    Town Called Panic, A (Panique au village) (2009)        Animation       2
009
1686    Illusionist, The (L'illusionniste) (2010)       Animation       2010
1687    Invincible Iron Man, The (2007) Animation       2007
1688    Hulk Vs. (2009) Animation       2009
1689    Asterix: The Land of the Gods (Ast??rix: Le domaine des dieux) (2014)   A
nimation        2014
1692    Nasu: Summer in Andalusia (2003)        Animation       2003
1693    Ooops! Noah is Gone... (2015)   Animation       2015
1695    Genius Party (2007)     Animation       2007
1696    Piper (2016)    Animation       2016
1697    The Red Turtle (2016)   Animation       2016
1700    DC Super Hero Girls: Hero of the Year (2016)    Animation       2016
1708    LEGO DC Super Hero Girls: Brain Drain (2017)    Animation       2017
1709    Fireworks, Should We See It from the Side or the Bottom? (2017) Animatio
n       2017
1711    Love Live! The School Idol Movie (2015) Animation       2015
Time taken: 1.304 seconds, Fetched: 17 row(s)
```

e) Select a.year, count(a.year) as count from (Select regexp_extract(movie_name, '(\\d{4})',1) as year from A3.movies_whole where genre='Animation') a group by year order by count desc limit 10;



```
hive (A3)> Select a.year, count(a.year) as count from (Select regexp_extract(mov
ie_name, '(\\d{4})',1) as year from A3.movies_whole where genre='Animation') a g
roup by year order by count desc limit 10;
Query ID = root_20201101233500_d626034c-82b4-4acd-b23d-9b9098b1bf7d
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_1604260709097
_0003)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.31 s
----------------------------------------------------------------------------------
OK
2009    5
2016    3
2007    2
1984    2
2017    2
2015    2
1971    2
1972    1
1980    1
1981    1
Time taken: 10.192 seconds, Fetched: 10 row(s)
```

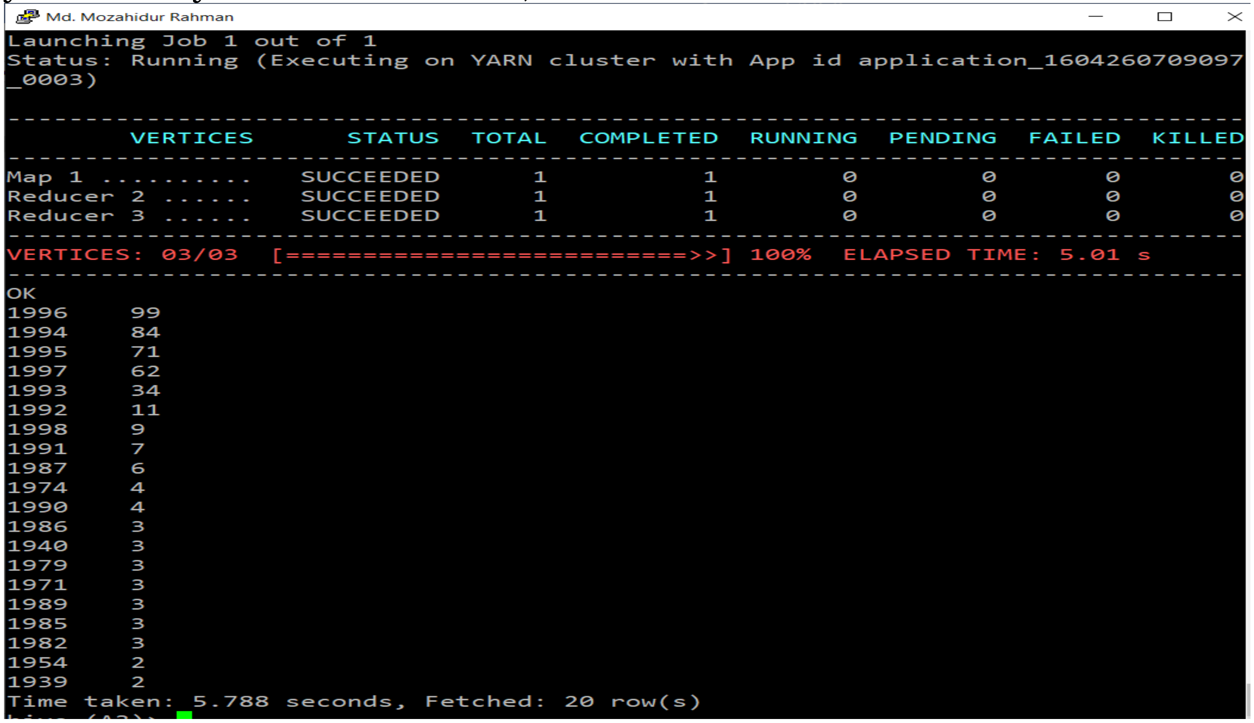
Select a.year, count(a.year) as count from (Select regexp_extract(movie_name, '(\\d{4})',1) as year from A3.movies_part where genre='Animation') a group by year order by count desc limit 10;



```
hive (A3)> Select a.year, count(a.year) as count from (Select regexp_extract(mov
ie_name, '(\\d{4})',1) as year from A3.movies_part where genre='Animation') a gr
oup by year order by count desc limit 10;
Query ID = root_20201101233719_9fd5340e-922f-417d-9117-b87b7d90d004
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0003)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.01 s
----------------------------------------------------------------------------------
OK
2009    5
2016    3
2007    2
1984    2
2017    2
2015    2
1971    2
1972    1
1980    1
1981    1
Time taken: 6.04 seconds, Fetched: 10 row(s)
```

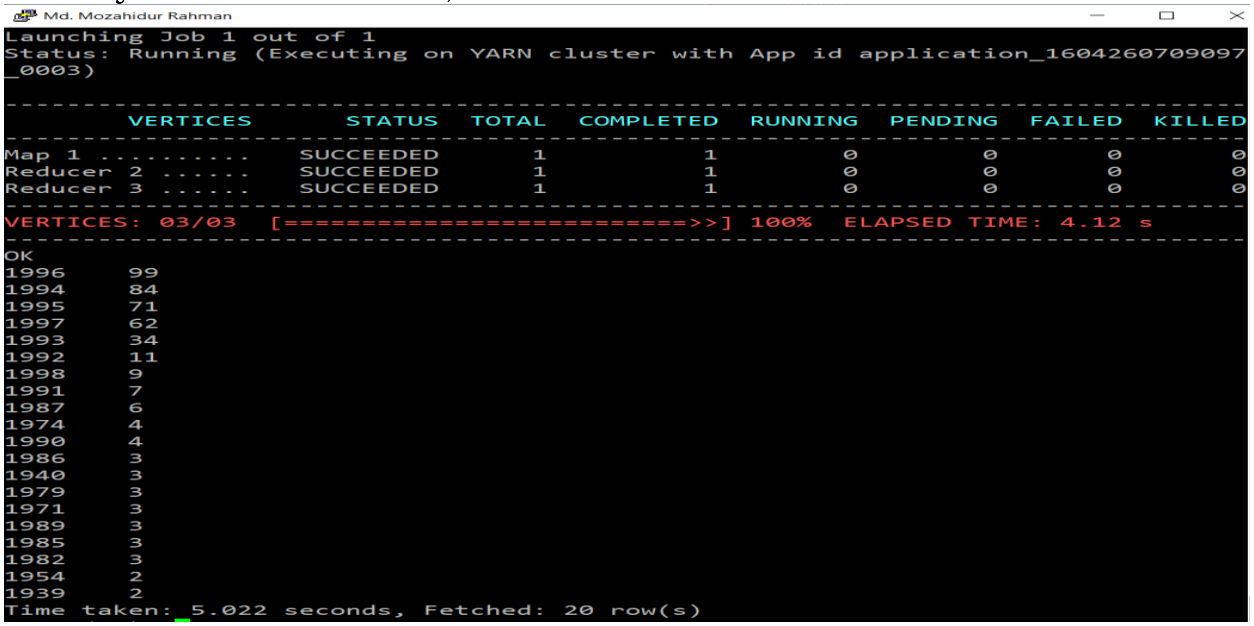
Findings: In my machine, movies_part (6.04 seconds) table is executed faster than movies_whole (10.192 seconds) table, however, it may vary based on the machine configurations.

f) Select a.year, count(a.year) as count from (Select regexp_extract(movie_name, '(\\d{4})',1) as year from A3.movies_whole where genre='Comedy') a group by year order by count desc limit 20;



```
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0003)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.01 s
----------------------------------------------------------------------------------
OK
1996    99
1994    84
1995    71
1997    62
1993    34
1992    11
1998    9
1991    7
1987    6
1974    4
1990    4
1986    3
1940    3
1979    3
1971    3
1989    3
1985    3
1982    3
1954    2
1939    2
Time taken: 5.788 seconds, Fetched: 20 row(s)
```

Select a.year, count(a.year) as count from (Select regexp_extract(movie_name, '(\\d{4})',1) as year from A3.movies_part where genre='Comedy') a group by year order by count desc limit 20;



```
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0003)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.12 s
----------------------------------------------------------------------------------
OK
1996    99
1994    84
1995    71
1997    62
1993    34
1992    11
1998    9
1991    7
1987    6
1974    4
1990    4
1986    3
1940    3
1979    3
1971    3
1989    3
1985    3
1982    3
1954    2
1939    2
Time taken: 5.022 seconds, Fetched: 20 row(s)
```

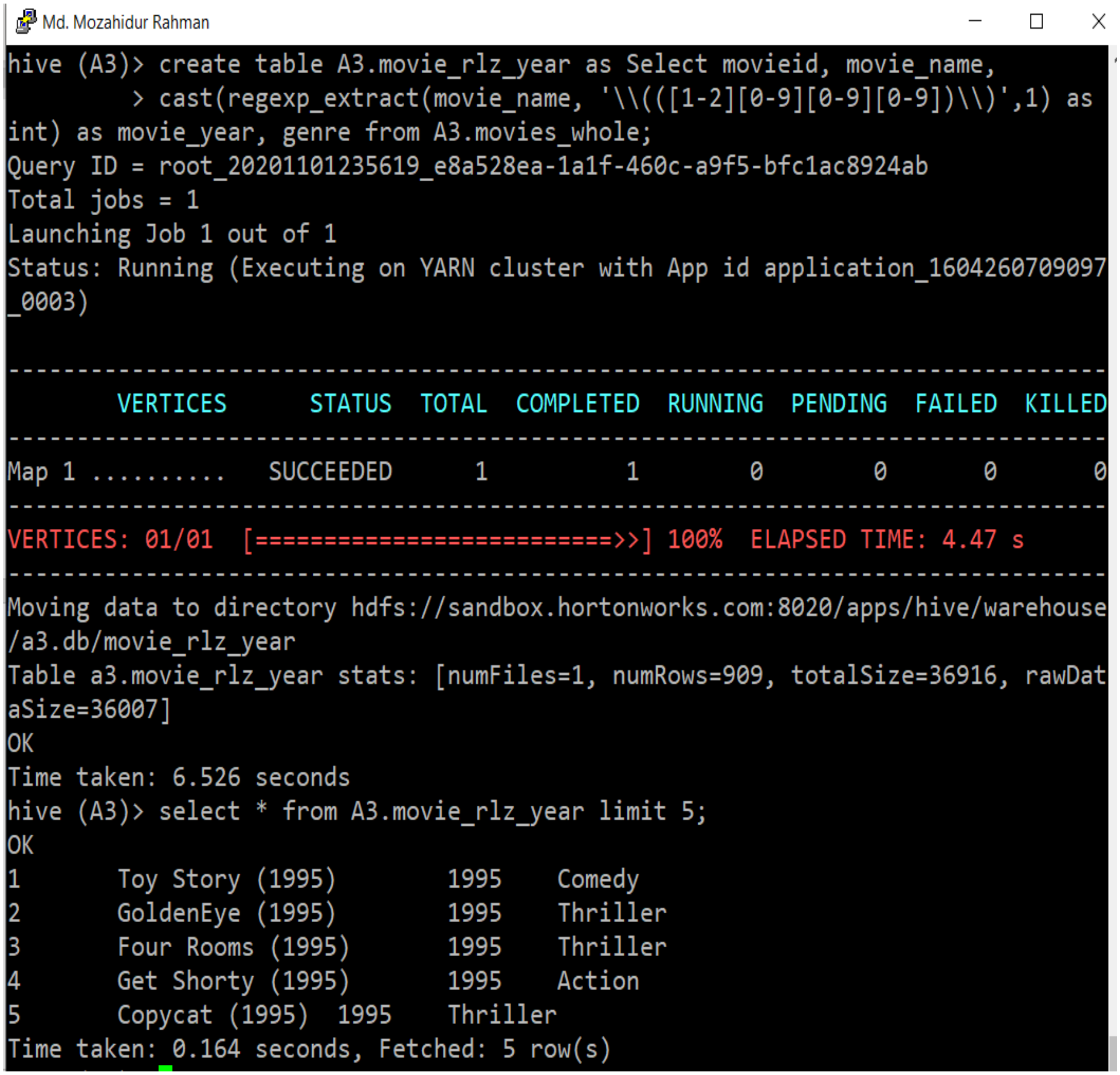
Findings: In my machine, movies_part (5.022 seconds) table is executed slightly faster than movies_whole (5.788 seconds) table, however, it may vary based on the machine configurations.

g) Extract movie released year from the movie title and store it by creating an additional attribute in a new table (Hint: Use regular expression and table with all information)

Query:

```
create table A3.movie_rlz_year as Select movieid, movie_name,
cast(regexp_extract(movie_name, '\\(([1-2][0-9][0-9][0-
9])\\)',1) as int) as movie_year, genre from A3.movies_whole;
select * from movie_rlz_year limit 5;
```

Screenshot:

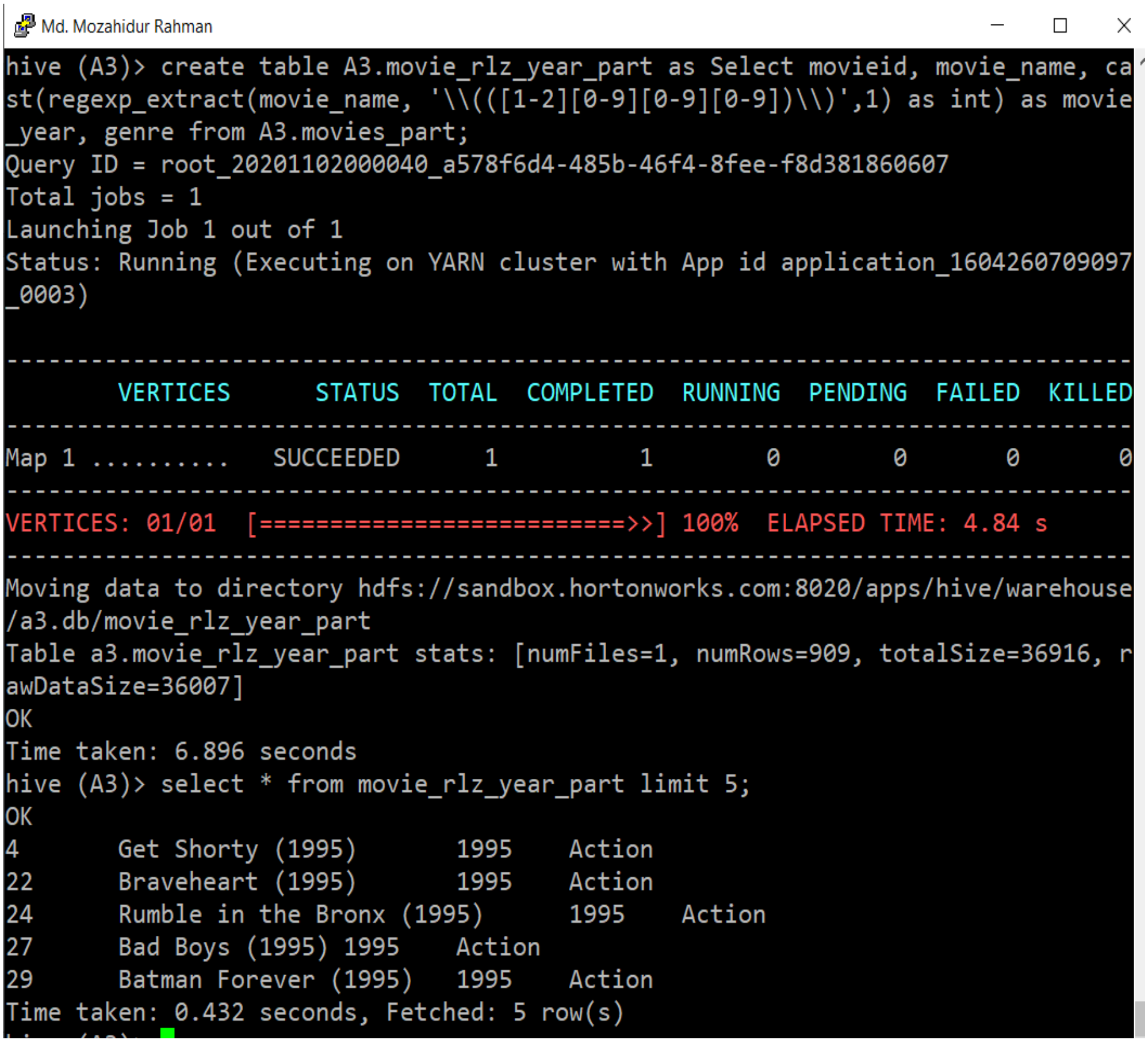


```
hive (A3)> create table A3.movie_rlz_year as Select movieid, movie_name,
         > cast(regexp_extract(movie_name, '\\(([1-2][0-9][0-9][0-9])\\)',1) as
int) as movie_year, genre from A3.movies_whole;
Query ID = root_20201101235619_e8a528ea-1a1f-460c-a9f5-bfc1ac8924ab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0003)


--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.47 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse
/a3.db/movie_rlz_year
Table a3.movie_rlz_year stats: [numFiles=1, numRows=909, totalSize=36916, rawDat
aSize=36007]
OK
Time taken: 6.526 seconds
hive (A3)> select * from A3.movie_rlz_year limit 5;
OK
1       Toy Story (1995)        1995    Comedy
2       GoldenEye (1995)        1995    Thriller
3       Four Rooms (1995)       1995    Thriller
4       Get Shorty (1995)       1995    Action
5       Copycat (1995)  1995    Thriller
Time taken: 0.164 seconds, Fetched: 5 row(s)
```

For the table movies_part:

```
create table A3.movie_rlz_year_part as Select movieid,
movie_name, cast(regexp_extract(movie_name, '\\(([1-2][0-9][0-
9][0-9])\\)',1) as int) as movie_year, genre from
A3.movies_part;
select * from movie_rlz_year_part limit 5;
```

Screenshot



```
hive (A3)> create table A3.movie_rlz_year_part as Select movieid, movie_name, ca
st(regexp_extract(movie_name, '\\(([1-2][0-9][0-9][0-9])\\)',1) as int) as movie
_year, genre from A3.movies_part;
Query ID = root_20201102000040_a578f6d4-485b-46f4-8fee-f8d381860607
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1604260709097
_0003)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.84 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse
/a3.db/movie_rlz_year_part
Table a3.movie_rlz_year_part stats: [numFiles=1, numRows=909, totalSize=36916, r
awDataSize=36007]
OK
Time taken: 6.896 seconds
hive (A3)> select * from movie_rlz_year_part limit 5;
OK
4       Get Shorty (1995)       1995    Action
22      Braveheart (1995)       1995    Action
24      Rumble in the Bronx (1995)      1995    Action
27      Bad Boys (1995) 1995    Action
29      Batman Forever (1995)   1995    Action
Time taken: 0.432 seconds, Fetched: 5 row(s)
```