Management of Big Data and Tools – DS8003 – Fall 2016

Assignment 2

NAJLIS, BERNARDO - Student Number #500744793

Dataset:

1. We are going to use the Dataset : **movielens_less.zip** (The entire dataset is available here Dataset and Scripts -> Movielens & IMDB-> movielens.zip. I selected a 2 files for this assignment)
2. Unzip the movielens_less.zip.
3. The README file in the zip file will give you information of the project that collected the data
4. The zip contains files: **u.data** and **u.user**
5. **u.data** -- The dataset has 100000 ratings by 943 users on 1682 movies. The file has **4 tab** ("\t") separated columns. The first column is the **user id,** the second column is the **movie id,** the third column is the **rating,** and the fourth column is a **timestamp**.
6. **u.user** - Demographic information about the users; this has **5** p**ipe "|"** separated columns. the first column is the **user id,** the second column is the **age,** the third column is the **gender (**Male denoted by 'M' and Female denoted by 'F'**),** fourth column is the **occupation, and the fifth column is the zip code**. The user ids are the ones used in the u.data data set.
7. Copy **u.data and u.user** to the virtual machine (**Filezilla**)
8. Copy **u.data and u.user** from virtual machine into HDFS (**hadoop fs -put**)
9. Create one table for u.data and one table for u.user in Hive

Submission:

1. Submit both the hive commands and the results (copy it into a file and submit)

Example: if you submit a file called assignment1.txt. It should contain the following information for every question.

Question XX: (XX is the question number)

HiveQL: Select count (*) from u.data;

Result: 100000

2. Submit using Assessment -> Assignments -> Assignment 2 - Hive

Assignment (Total of 10)

1. Find the user id who has rated the most number of movies (3)
2. Find average rating received by movie with id 178. (3)
3. The users belonging to which 3 occupations provided the most number of ratings (2)
4. How many unique male users provided at least one rating of 5 (2)

Resolution

Hive script for data import

```
CREATE DATABASE movielens;
USE movielens;

CREATE TABLE movielens.users (
      userid INT,
      age INT,
      gender STRING,
      occupation STRING,
      zipcode STRING
) ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|';

LOAD DATA INPATH '/user/root/assignment2/u.user'
OVERWRITE INTO TABLE movielens.users;

SELECT * FROM movielens.users LIMIT 10;

CREATE TABLE movielens.data (
      userid INT,
      movieid INT,
      rating INT,
      ts TIMESTAMP
) ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';

LOAD DATA INPATH '/user/root/assignment2/u.data'
OVERWRITE INTO TABLE movielens.data;

SELECT * FROM movielens.data LIMIT 10;
```

1. Find the user id who has rated the most number of movies (3)

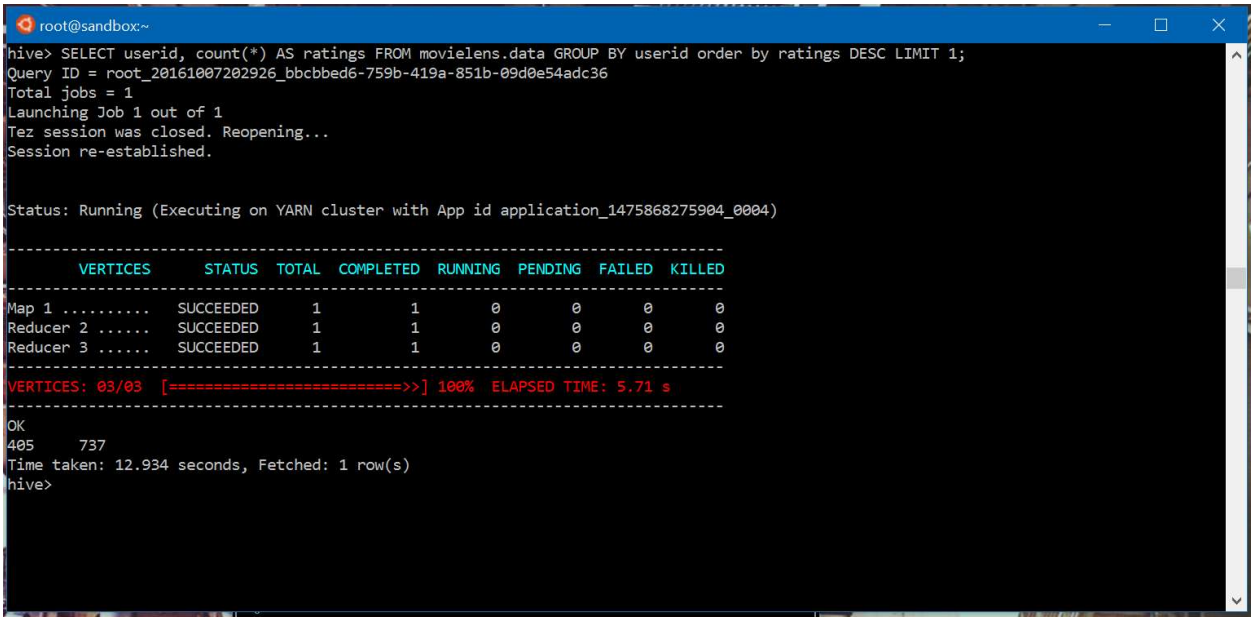
QUERY

```
SELECT userid, count(*) AS ratings FROM movielens.data GROUP BY userid
order by ratings DESC LIMIT 1;
```

RESULT

```
405   737
```

JOB RESULTS



2. Find average rating received by movie with id 178. (3)

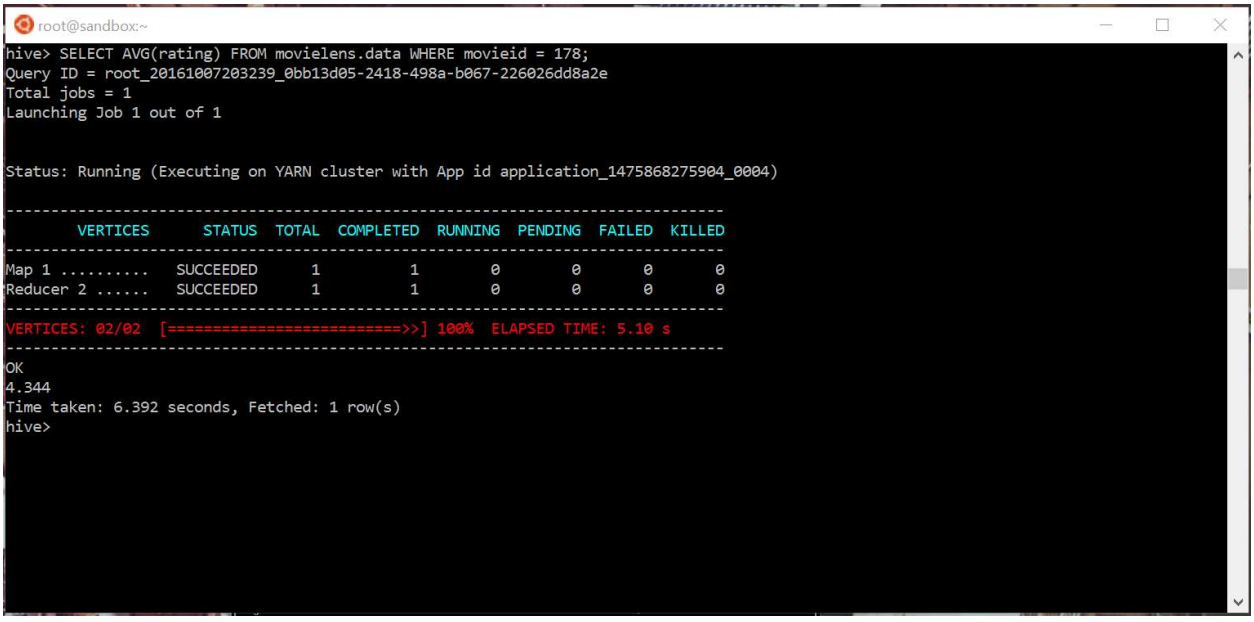
QUERY

```
SELECT AVG(rating) FROM movielens.data WHERE movieid = 178;
```

RESULT

```
4.344
```

JOB RESULTS



3. The users belonging to which 3 occupations provided the most number of ratings (2)

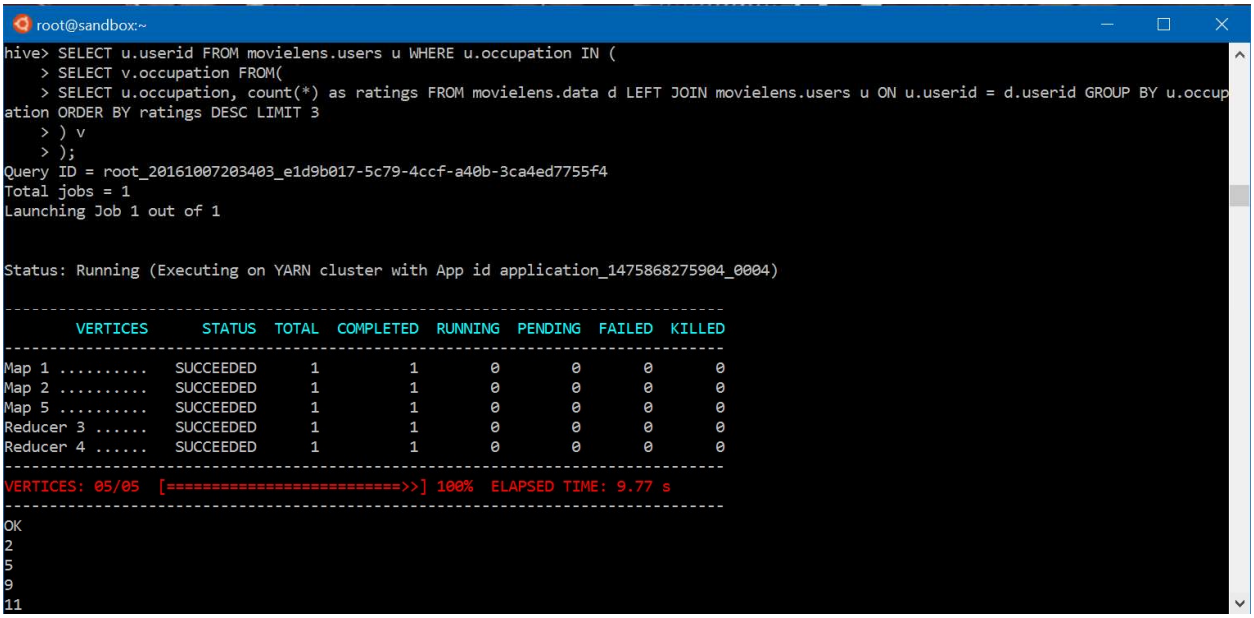
QUERY

```
SELECT u.userid FROM movielens.users u WHERE u.occupation IN (
SELECT v.occupation FROM(
SELECT u.occupation, count(*) as ratings FROM movielens.data d LEFT
JOIN movielens.users u ON u.userid = d.userid GROUP BY u.occupation
ORDER BY ratings DESC LIMIT 3
) v
);
```

RESULT

```
2, 5, 9, 11, 12, 13, 15, 18, 30, 32, 33, 36, 37, 38, 49, 51, 52, 59,
64, 65, 66, 67, 68, 73, 76, 81, 83, 85, 90, 94, 99, 101, 103, 104,
108, 109, 110, 117, 120, 124, 132, 135, 136, 137, 139, 140, 142, 152,
153, 154, 155, 156, 158, 159, 165, 166, 167, 168, 169, 171, 173, 178,
187, 188, 192, 193, 198, 202, 203, 206, 209, 212, 217, 221, 223, 224,
226, 228, 230, 235, 240, 241, 242, 243, 245, 246, 248, 249, 254, 257,
258, 259, 262, 270, 273, 274, 276, 281, 286, 291, 295, 297, 301, 302,
303, 304, 306, 307, 310, 312, 314, 315, 316, 320, 321, 322, 323, 324,
327, 329, 330, 332, 333, 341, 342, 346, 347, 348, 350, 351, 355, 358,
359, 360, 361, 363, 366, 367, 368, 369, 372, 373, 376, 377, 378, 388,
391, 393, 395, 397, 398, 399, 403, 406, 408, 411, 412, 413, 415, 416,
417, 420, 423, 425, 426, 428, 429, 434, 437, 440, 442, 446, 450, 451,
453, 454, 459, 460, 461, 462, 465, 466, 469, 471, 472, 473, 476, 477,
478, 479, 482, 484, 485, 486, 489, 492, 496, 497, 501, 502, 505, 510,
511, 512, 517, 519, 521, 524, 528, 532, 534, 535, 541, 542, 544, 547,
550, 552, 553, 555, 556, 560, 565, 566, 568, 569, 570, 572, 574, 577,
579, 580, 581, 582, 584, 586, 587, 588, 590, 592, 593, 594, 597, 599,
602, 604, 608, 609, 610, 612, 614, 615, 618, 619, 621, 623, 624, 629,
631, 632, 635, 636, 637, 640, 641, 642, 646, 647, 649, 652, 654, 656,
659, 660, 663, 669, 673, 674, 675, 677, 678, 679, 684, 686, 689, 691,
696, 697, 699, 700, 702, 703, 705, 706, 709, 710, 711, 712, 713, 719,
725, 727, 729, 731, 732, 733, 734, 740, 742, 747, 749, 751, 755, 757,
758, 759, 760, 761, 764, 765, 766, 770, 771, 773, 774, 778, 779, 781,
787, 789, 791, 793, 794, 797, 804, 805, 810, 811, 813, 814, 815, 816,
817, 820, 824, 831, 834, 838, 847, 849, 851, 854, 858, 859, 861, 863,
866, 869, 870, 872, 875, 876, 877, 878, 880, 885, 886, 887, 890, 892,
893, 894, 897, 899, 903, 904, 905, 907, 909, 912, 913, 914, 917, 919,
921, 923, 924, 928, 931, 932, 933, 936, 937, 939, 941, 943
```

JOB RESULTS



4. How many unique male users provided at least one rating of 5 (2)

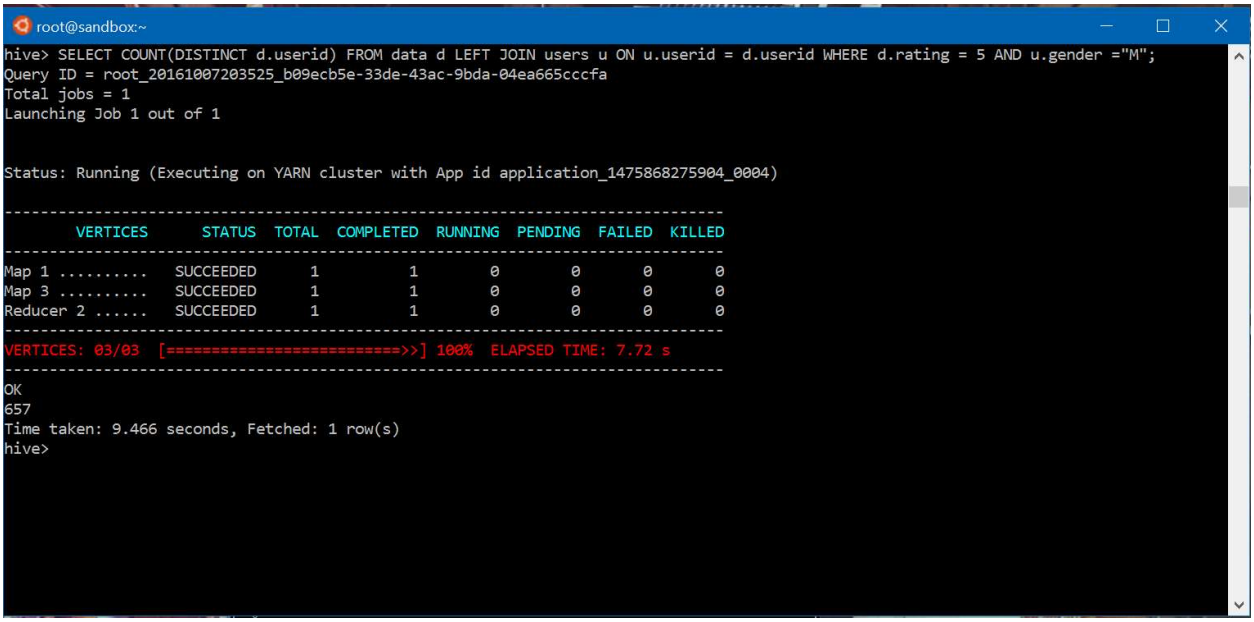
QUERY

```
SELECT COUNT(DISTINCT d.userid) FROM data d LEFT JOIN users u ON
u.userid = d.userid WHERE d.rating = 5 AND u.gender ="M";
```

RESULT

```
657
```

JOB RESULTS


root@sandbox:~
hive> SELECT COUNT(DISTINCT d.userid) FROM data d LEFT JOIN users u ON u.userid = d.userid WHERE d.rating = 5 AND u.gender ="M";
Query ID = root_20161007203525_b09ecb5e-33de-43ac-9bda-04ea665cccfa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1475868275904_0004)
VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... SUCCEEDED 1 1 0 0 0 0
Map 3 ..... SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==========>>] 100% ELAPSED TIME: 7.72 s
OK
657
Time taken: 9.466 seconds, Fetched: 1 row(s)
hive>