**GCD Assignments week 4 – Hadoop Hive**

Do the following activities and put the statements and answers in your portfolio.

The overall process for these exercises is as follows.

ml-data

Output

Hive

download

copy

import

SQL

Cut -d

*Activity 1*

Copy ml-data.tar.gz from the intranet (at Assignments) to your virtual machine, unzip the files and store them in **HDFS**.

The file README contains a description of the files.

You will use the following files:

* u.data 100000 movie ratings
* u.user information about users that posted ratings
* u.item information about the movies

*Activity 2*

Import the HDFS files about the users (u.user), movies (u.item) and ratings (u.data) in Hive. Create tables first.

Tip 1: use something like

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘#'

STORED AS TEXTFILE

TIP 2: In the exercises below you won’t need all columns of these files. To save work you could remove unnecessary columns. To do this use the Unix cut command.

Example:

**cut -d'|' –f3,4,17,18,19 u.item > movies**

This copies only columns 3, 4, etc to the file ‘movies’

*Activity 3*

Answer the following questions using Hive queries. Give the queries (keep them simple, the answer counts, not the query). **Store the results on HDFS.**

1. Give the number of male and female users.
   1. **273 females and 670 males**
2. Give the number of men and women per occupation.
   1. **F administrator 36**
   2. **F artist 13**
   3. **F educator 26**
   4. **F engineer 2**
   5. **F entertainment 2**
   6. **F executive 3**
   7. **F healthcare 11**
   8. **F homemaker 6**
   9. **F lawyer 2**
   10. **F librarian 29**
   11. **F marketing 10**
   12. **F none 4**
   13. **F other 36**
   14. **F programmer 6**
   15. **F retired 1**
   16. **F salesman 3**
   17. **F scientist 3**
   18. **F student 60**
   19. **F technician 1**
   20. **F writer 19**
   21. **M administrator 43**
   22. **M artist 15**
   23. **M doctor 7**
   24. **M educator 69**
   25. **M engineer 65**
   26. **M entertainment 16**
   27. **M executive 29**
   28. **M healthcare 5**
   29. **M homemaker 1**
   30. **M lawyer 10**
   31. **M librarian 22**
   32. **M marketing 16**
   33. **M none 5**
   34. **M other 69**
   35. **M programmer 60**
   36. **M retired 13**
   37. **M salesman 9**
   38. **M scientist 28**
   39. **M student 136**
   40. **M technician 26**
   41. **M writer 26**
3. Give the name of the movie with the highest ratings given by male students. Same question for female students.
   1. **F** **– Santa with Muscles 5.0**
   2. **M – Love Serenade 5.0**
4. (Optional) Give the names of the movies in each of the genres ‘Action’, ‘Romance’ and ‘Horror’ with the highest ratings given by male students. Same question for female students.

*Activity 4*

In question 3: how do you calculate “highest rating”? Do you use MAX, SUM, AVG? Or an other function?

**I used the AVG function. However, the function is not suitable when only one person rated it with 5 stars, then it will come out on top. The sum function is not suited when a lot of people voted. The movies who have less votes will be out-voted then. The max function is not suited, because it will misplace movies when they had just one rating.**

E.g.

MAX not suitable (wrong answer) if:

Movie A scores 5,1,1,1,1,1,1,..

Movie B scores 5,5,4,4,5,5,4,4,5,5,4,4,...

SUM not suitable (wrong answer) if:

Movie A scores 500x 1.0

Movie B scores 10x 5.0

AVG not suitable (wrong answer) if:

Movie A scores 5.0 (only one value)

Movie B scores 5,5,4,4,5,5,4,4,5,5,4,4,...

*Activity 5*

How does IMDB rate its movies?

**It is a star-based system where the user rates a movie between one- and five stars.**

*Activity 6 (optional)*

Implement the IMDB formula in HiveQL.