**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. SELECT count(\*), gender FROM users GROUP BY gender;  
      273 F | 670 M
2. Give the number of men and women per occupation.
   1. SELECT count(\*), occupation, gender FROM users GROUP BY occupation, gender ORDER BY occupation;
3. Give the name of the movie with the highest ratings given by male students. Same question for female students.
   1. select t.gender, t.avgrating, i.title from

(SELECT u.gender as gender, AVG(d.rating) as avgrating, i.id as movieid, rank() over (order by avg(d.rating) desc) as r

FROM users u, useritem i, userdata d

WHERE u.userid = d.userid

AND d.itemid = i.id

GROUP BY u.gender, i.id) as t,

useritem as i

where t.movieid = i.id

and t.r = 1;

* 1. Multiple movies per gender (27 results total)

1. (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?

AVG is to me the best options since it will take every opinion equally into account. Eventhough in the last case you see one 5.0 is unfairly higher then 5 + 4 + 5 + 5…

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?

IMDB uses something called the weighted average, which means they weigh more votes of the same value higher then other values with less votes.

*Activity 6 (optional)*

Implement the IMDB formula in HiveQL.

SELECT d1.itemid, SUM(d1.rating \* d1.ratingcnt) / SUM(d1.ratingcnt) AS wrating

FROM (

SELECT d2.itemid as itemid, d2.rating AS rating, COUNT(d2.userid) AS ratingcnt

FROM userdata d2

GROUP BY d2.itemid, d2.rating

) d1

GROUP BY d1.itemid

ORDER BY wrating DESC

(Not gender specific or connected to other tables for ease, could be done using subqueries or even better (cached) views)