**GCD Assigments week 4 – Hadoop Hive**

*Activity 1*

Unzip ml-data.tar.gz (see [link](https://portal.fhict.nl/erp-bi/Lesmateriaal/ml-data.tgz)) 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: You probably 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.
2. Give the number of men and women per occupation.
3. Give the name of the movie with the highest ratings given by male students. Same question for female students.
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?

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?