Assignment 1 – mrJob script

Tim Gras

630259

URL => <https://github.com/dantim1997/PDP-Assignment1-630259>

# Setup

to run the code, install a VM with Hadoop.

Make sure to have MRJob, Pip and the u.data file installed.

# execute

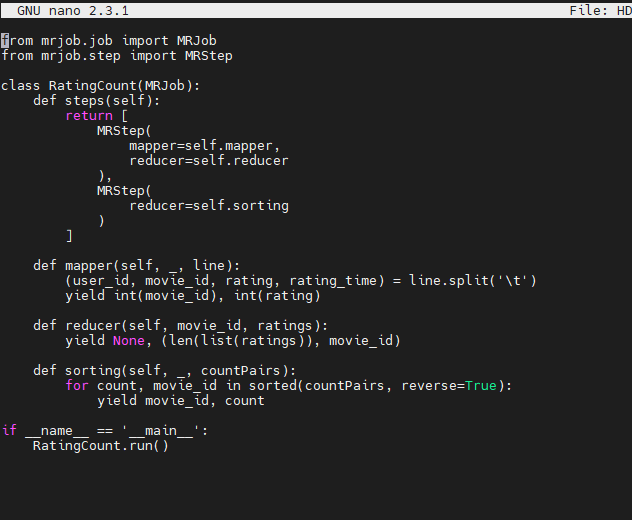
Log in as su root

Run the python file with the following command:

python HD\_rating.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar u.data

or can be run locally with:

python HD\_rating.py u.data



# Information

For this code to get every rating by the movies id. I made 2 steps, one to give every rating by the movie\_id and the other one to order them from high to low.

So first it will come to the main when running the python application it will go to the function steps. There I declared 2 steps the one to get the data and one to order the data.

So first the mapper function: in this function it will split every line based on the tab split. This will create a table of the u.data and giving the movie\_id and the rating back.

Then it will go the the reducer. This will count the amount of rating based on the key that is the movie\_id. Now I have a list of movie\_ids with the amount of rating each movie has.

For the second stap the sorter function will order the movies based on the amount of ratings each movie has. And I did this descending to get the most first.

# Github

HD\_rating.py This is the code to run the file

Assignment 1.docx This is the documentation

Result.jpg This is the result image of the code

# Result:

