Week 4

Kevin Jetten

# Activity 1

Because I extracted the files from the zip in the documents folder of the system I used the following command to make the m visible to the HDFS:

hadoop fs -put /home/cloudera/Documents/ml-data/ /ml-data

# Acitivity 2

Separators used for each file:

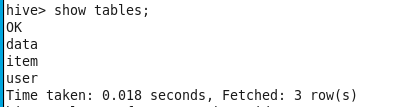
u.user: |

u.item: |

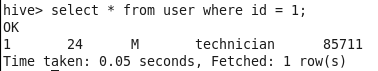
u.data: tab-sign

While importing the u.data I had to use a tab sign as previously mentioned. After doing some research online I found that the default separator used by hive is “ASCII \001”. The ASCII octal for a horizontal tab is 011. Thus making the create statement the following :

FIELD TERMINATED BY ‘\011’



*Image 1: show tables*

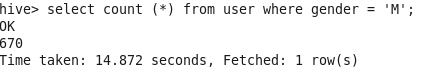


*Image 2: get user with id 1*

I’ve added some screenshots from really simple queries to prove that I have imported all the data.

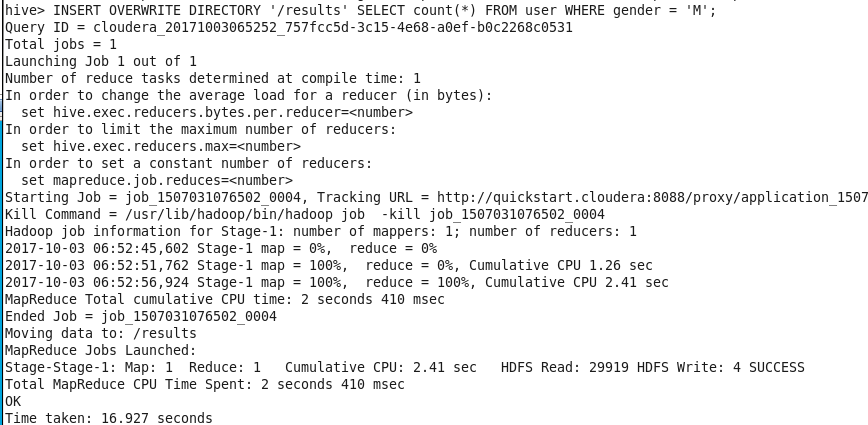
# Activity 3

**Give the number of male and female users.**



*Image 3: get male users*

As you can see in the image above there are 670 male users, if you replace the ‘M’ with ‘F’ you will see there are 273 women.



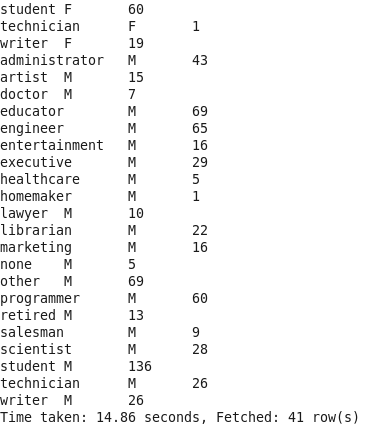
*Image 4: Storing the results on HDFS*

Image 4 shows how I store the results on the HDFS. Because this is a step that has to be done for every question and it has no variation I won’t make a screenshot of it for every question.

**Give the number of men and women per occupation.**

****

*Image 5: Query*



*Image 6: Results (cropped for formatting)*

**Give the name of the movie with the highest ratings given by male students. Same question for female students.**

****



*Image 7: Query*

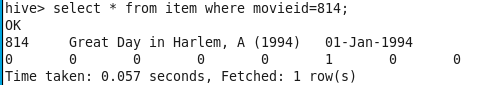
To avoid making an even bigger query by making this into a subquery to get the movie name I just did an extra query getting the name by its id.



*Image 8: Result for males*



*Image 9: Result for females*



*Image 10: Movie for males*



*Images 11: Movie for females*

# Activity 4

I calculated the highest rating by finding the max on the average grouped by all movies. This seems to me like the fairest way to judge a movie.

# Activity 5

They use a “weighted average” which they explain very clearly.

“A weighted average means that some votes have more weight than others in our calculations. If you want more information, see this [detailed explanation](http://www.imdb.com/help/show_leaf?votes).”