Yara Weening - 642287

Hadoop  23-6-2022

Assignment 1

**GitHub url with my source code:** <https://github.com/yaraweening/Hadoop>

For Hadoop an assignment is made to use HDFS for the data of a movie dataset. This dataset is also used in combination with Python, MrJob and Nano.

To complete the assignment the following steps took place:

**Step 1:** First a Virtual Machine (VM) has to be downloaded. Therefore the Oracle VM is used. Downloaded from: <https://www.virtualbox.org>.

**Step 2:** Then the Hortonworks Data Platform (HDP) on Hortonworks Sandbox can be installed. Downloaded from: <https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html>.

**Step 3:** Now that the VM and Hortonworks are installed. The Hortonworks can be imported in the VM to start it after it is imported.

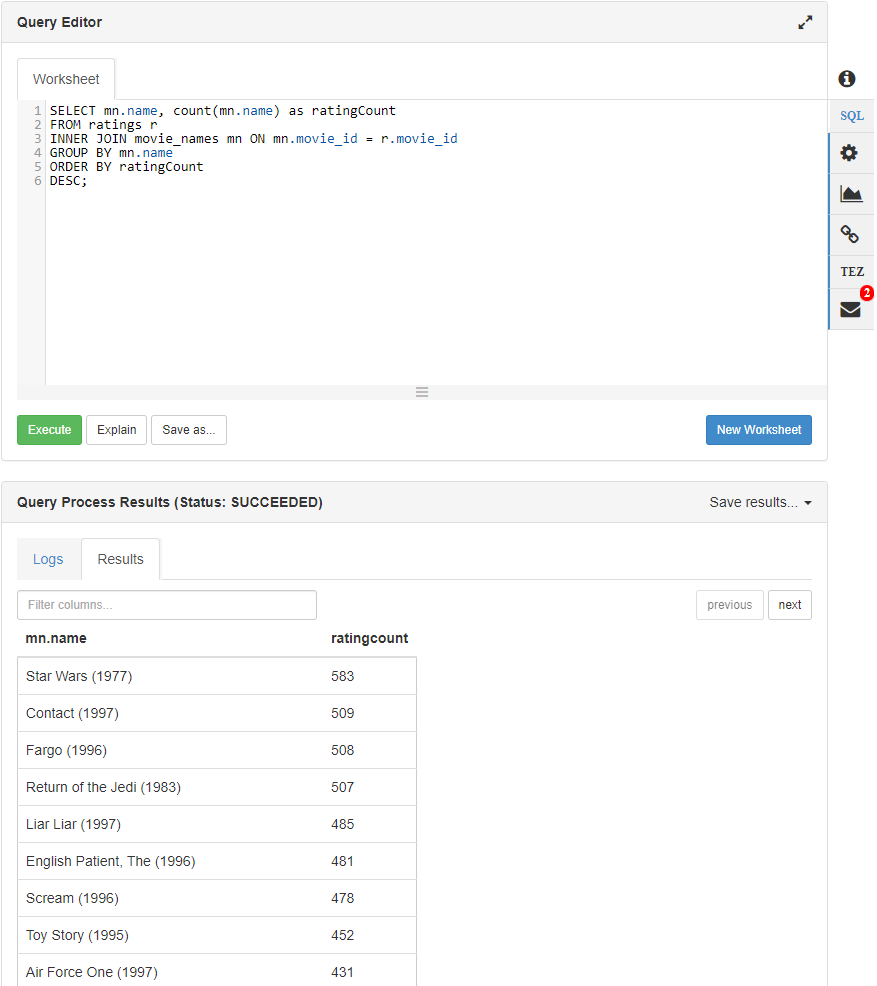
**Step 4:** From <https://grouplens.org> a Movie Lens dataset is downloaded and unpacked locally on my machine to use it for the assignment.

**Step 5:** When the Hortonworks is started up an url is shown. At a new internet tab the following link can be filled in to launch the Ambari Dashboard: <http://127.0.0.1:8888>.

**Step 6:** To sign in the maria\_dev credentials are used. After signing in the Dashboard is shown.

**Step 7:** The downloaded Movie Lens dataset is imported in the Dashboard with the use of some extra needed settings for showing the data. The table has to get a name and also the colums has to get some proper column names.

**Step 8:** In the Hive View a query is made to retrieve the movie name with their rating count from heighest to lowest. To get the name of the movies an INNER JOIN is used, based on the movie\_id.



**Step 9:** PuTTY is installed to do the assignment also with Python, MrJob and Nano. As host 127.0.0.1:2222 is used with maria\_dev is to login.

**Step 10:** With only PuTTY and the currently running Hortonworks it does not work. A new Hortonworks needs to be installed provided by the teacher. This already includes Python, MrJob and Nano can be used to edit the MrJob code.

**Step 11:** The new Hortonbox is imported to the VM and can be started on the same url as the previous Hortonbox.

**Step 12:** Because this is one is new the dataset has to be inported again.

**Step 13:** Then PuTTY can be started to connect to the host.

**Step 14:** Now python HD\_rating.py u.data is typed in to check the output. The data still needs to be counted by the ratings given for each movie and sorted by their number of ratings.

**Step 15:** To do so su root is typed to have the rights to modify the code with Nano.

**Step 16:** The following code is added to the MrJob code that was already there and saved as a new file HD\_rating\_sorted.py:

Old:

Afbeelding met tekst

Automatisch gegenereerde beschrijving

New:

Afbeelding met tekst

Automatisch gegenereerde beschrijving

So the the new code now counts the ratings and returns the movies wich have the heighest count of ratings first.

**Step 17:** Last step is to check if the code works. Herefore the following commands are used.



Old:

Afbeelding met tekst

Automatisch gegenereerde beschrijving

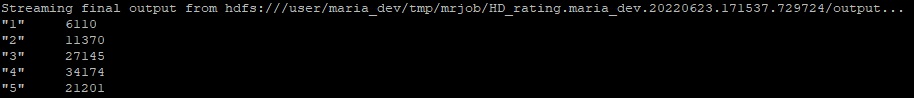
New:

Afbeelding met tekst

Automatisch gegenereerde beschrijving



Old:



New:

