Lab 6 – IMDB's Most Frequent Co-stars (with Pig)

CC5212-1

April 22, 2015

Today we do the same task as we did with Java, but this time using Pig. The goal is the same as in the previous lab: `http://aidanhogan.com/teaching/cc5212-1/doc/lab5.pdf`; we again want to count the number of times each pair of co-stars have appeared in a theatrical movie together. The input data are also the same.

- The instructions for logging into the server, for accessing and uploading data to HDFS, for building your code and running it, etc. are available from Lab 4: `http://aidanhogan.com/teaching/cc5212-1/doc/lab4.pdf`.
- The data you need are on HDFS in the `/uhadoop/shared/imdb/` folder. The `imdb-stars.tsv` file (as before) is about 1GB and contains 13 million roles. There is also a smaller file, `imdb-stars-100k.tsv`, which contains 100,000 roles; we will use this smaller file for testing first.
- Download the code project from `http://aidanhogan.com/teaching/cc5212-1/code/mdp-lab6.zip` and open it in Eclipse. There's a single file in there. Open it in the text editor of your choice.
- The exercise for today is to code an Apache Pig script to run the same job as last week. To help in this, you should check out the user guide at `http://pig.apache.org/docs/r0.14.0/basic.html`. There is in total nine lines of code ... we will go through it line by line in the lab.
- Replace "`ahogan`" on the last line with your username!!
- Once your script is done, try testing it! Upload it to the local directory on `cluster.dcc.uchile.cl`. Call `pig actor_count.pig`. The first output line should be `28 Akpinar, Metin##Alasya, Zeki`.
- If your script did not work, try debugging it by putting a `STORE` command after each line in turn and checking the output.
- If your script is working, try testing it over the big file. Change the input file as indicated in the comments at the top of the Pig script to `hdfs://cm:9000/uhadoop/shared/imdb/imdb-stars.tsv` and likewise change the output to `/uhadoop/[username]/imdb-costars/`. Since it will take a while to run, you should run it on a `screen`, which is a virtual terminal session that will remain active on the server even when you log out:
  - Type `screen`. A blank terminal starts. Run your Pig script as normal in that terminal.
  - If you want to go back to your main terminal, press Ctrl + A, then D (hold Ctrl while pressing A, release, then press D). Your virtual screen is still running in the background (even if you log out, it will remain). Make sure to take careful note of the number `NUM` in the message here: `[detached from NUM.pts-3.cluster-01]`. Write it down! This is your screen. There are many like it, but this one is yours. You'll need the number to find the right screen later.
  - If you want to go back to your virtual screen from the main terminal, type `screen -d -r NUM`.
  - If you're done with a virtual screen, while in that screen, press Ctrl + D to kill it (and whatever it was running). Needless to say, don't kill other people's screens; it's very impolite.
- It's unlikely that you will get a result over the full data before the end of the lab but you can check up on it later. Which pairs of stars appeared together the most times? In how many movies did Al Pacino and Robert De Niro co-star? Who has co-starred the most times with Uma Thurman?
- Once the script works for the smaller file, feel free to submit.