Project 6

100 Possible Points

| 11/15/2022

Attempt 1

11/15/2022
Next Up: Review Feedback

Attempt 1 Score:
N/A

Add Comment

Unlimited Attempts Allowed

Details

Netflix Data Analysis using Pig

Description

The purpose of this project is to develop a simple analysis program for Netflix Data using Apache Pig.

This project must be done individually. No copying is permitted. **Note: We will use a system for detecting software plagiarism, called Moss (http://theory.stanford.edu/~aiken/moss/) , which is an automatic system for determining the similarity of programs.** That is, your program will be compared with the programs of the other students in class as well as with the programs submitted in previous years. This program will find similarities even if you rename variables, move code, change code structure, etc.

Note that, if you use a Search Engine to find similar programs on the web, we will find these programs too. So don't do it because you will get caught and you will get an F in the course (this is cheating). Don't look for code to use for your project on the web or from other students (current or past). Just do your project alone using the help given in this project description and from your instructor and GTA only.

Setting up your Project

As in the previous projects, you will develop your program on SDSC Expanse. Login into Expanse and download and untar project6:

```
wget http
tar xfz pr
```

Try Again

```
chmod -R g-wrx,o-wrx project6
```

Go to project6/examples and look at the join.pig example. You can run it in standalone mode on Expanse using:

```
sbatch join.local.run
```

The results will be in the directory output.

Optional: Use your laptop to develop your project

If you'd prefer, you may use your laptop to develop your program and then test it and run it on Expanse.

To install Pig and the project:

```
cd
wget https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
tar xfz pig-0.17.0.tar.gz
wget http://lambda.uta.edu/cse6332/project6.tgz
tar xfz project6.tgz
```

You may use Pig on your laptop in local mode interactively:

```
~/pig-0.17.0/bin/pig -x local
```

Go to project6/examples and look at the join.pig example. You can run it in local mode using:

```
rm -rf output
~/pig-0.17.0/bin/pig -x local join.pig
```

The results will be in the directory output. To run project6 in local mode:

```
cd ~/project6
rm -rf output
~/pig-0.17.0/bin/pig -x local -param G=small-netflix.csv -param O=output netflix.pig
```

Project

Try Again

You are asked to re-implement Project #1 (Netflix data analysis) using Apache Pig. In your Pig script, you can access the path of the input dataset as '$G' and the output path as '$O'. That is, you can use LOAD '$G' USING ..., to load the dataset and STORE X INTO '$O' ..., to write the relation X to the output directory. Also, if there is a read error in the input, the LOAD operation will set the result values to NULL, which can be tested using the condition: IS NOT NULL.

To run your program in local mode over the small dataset use:

```
sbatch netflix.local.run
```

These results should be similar to the results in the file `small-solution.txt`.

After you make sure that your program runs correctly in local mode, you run it in distributed mode using:

```
sbatch netflix.distr.run
```

This will process the netflix on the large Netflix dataset and will write the result in the directory `output-distr`. These results should be similar to the results in the file `large-solution.txt`.

Documentation

You can learn more about Pig at:

- [Pig: Getting Started](http://pig.apache.org/docs/r0.17.0/start.html) (http://pig.apache.org/docs/r0.17.0/start.html)
- [Pig Latin Basics](http://pig.apache.org/docs/r0.17.0/basic.html) (http://pig.apache.org/docs/r0.17.0/basic.html)

What to Submit

Submit the zipped project6 directory, which must contain the files:

```
project6/netflix.pig
project6/netflix.local.out
project6/output/part-r-00000
project6/netflix.distr.out
project6/
```

Try Again

Preview Unavailable

project6.tgz

Download
(https://uta.instructure.com/files/23280518/download?download_frd=1&verifier=TKSFkXpDIa9PGeb6pIM0X8v8UG4crJnU5T2FOZ8z)

Try Again