Programming Assignment 7
A Simple Map-Reduce Program using Hive

Due on Friday April 24 before midnight

Description

The purpose of this project is to develop a simple program using Apache Hive.

This project must be done individually. No copying is permitted. **Note: We will use a system for detecting software plagiarism, called 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.

Platform

As in the previous projects, you will develop your program on SDSC Comet.

Setting up your Project

Login into Comet and download and untar project7:

```
wget http://lambda.uta.edu/cse6331/project7.tgz
tar xfz project7.tgz
chmod -R g-wrx,o-wrx project7
```

You need to create an empty metastore database first (this must be done only once):

```
cd
rm -rf metastore_db  warehouse
export HIVE_HOME=/oasis/projects/nsf/uot143/fegaras/apache-hive-2.1.0-bin
export JAVA_HOME=/lib/jvm/java
export HADOOP_CONF_DIR=$HOME/cometcluster
module load hadoop/2.6.0
$HIVE_HOME/bin/schematool -dbType derby -initSchema
```

Go to project7/example and look at the join.hql example. You can run it in local mode using:

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

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 Comet.

To install Hive and the project:

```
cd
wget https://downloads.apache.org/hive/stable-2/apache-hive-2.3.7-bin.tar.gz
tar xfz apache-hive-2.3.7-bin.tar.gz
wget http://lambda.uta.edu/cse6331/project7.tgz
tar xfz project7.tgz
```

Every time you login, you need to execute these:

```
export HIVE_HOME=$HOME/apache-hive-2.3.7-bin
export HADOOP_HOME=$HOME/hadoop-2.6.5
export PATH=$HIVE_HOME/bin:$PATH
export HIVE_OPTS="--hiveconf mapreduce.framework.name=local --hiveconf fs.default.name=file://$HOME --hiveconf hive.metastore.warehouse.dir=file://
```

You also need to create an empty metastore database first (this must be done only once):

```
cd
rm -rf metastore_db  warehouse
schematool -dbType derby -initSchema
```

Then, to evaluate Hive commands interactively, do:

```
hive
```

Go to project7/example and look at the join.hql example. You can run it in local mode (after you setup your PATH) using:

```
hive -f join.hql
```

Project Description

You are asked to implement Project 1 (histogram of pixels) using Apache Pig. An empty `histogram.hql` is provided as well as a script to run this code on Comet. Your hive script should read an input pixel file as in Project 1. The data files are exactly the same as in Project 1. Note: you can access the input pixel file in Hive (which are passed as a parameter) as '${hiveconf:P}'.

To run it in local mode over the small pixel file do:

```
sbatch histogram.local.run
```

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

```
sbatch histogram.distr.run
```

This will work on `pixels-large.txt`.

Documentation

You can learn more about Hive at:

- Hive: Getting Started
- Hive Tutorial

What to Submit

You need to submit the following files only:

```
project7/histogram.hql
project7/histogram.local.out
project7/histogram.distr.out
```

Submit Programming Assignment #7:

Select a file: Choose File no file selected

Send File

Last modified: 15/04/2020 by Leonidas Fegaras