Programming Assignment 6
A Simple Map-Reduce Program using Pig

Due on Thursday April 16 before midnight

Description

The purpose of this project is to develop a simple program 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, 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 Comet. Login into Comet and download and untar project6:

```
wget http://lambda.uta.edu/cse6331/project6.tgz
tar xfz project6.tgz
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 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 Comet.

To install Pig and the project:

```
cd
wget http://mirrors.gigenet.com/apache/pig/pig-0.16.0/pig-0.16.0.tar.gz
tar xfz pig-0.16.0.tar.gz
```

```
wget http://lambda.uta.edu/cse6331/project6.tgz
tar xfz project6.tgz
```

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

```
~/pig-0.16.0/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-0.16.0/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-0.16.0/bin/pig -x local -param P=pixels-small.txt -param O=output histogram.pig
```

Project Description

You are asked to implement Project 1 (histogram of pixels) using Apache Pig. Your Pig script should read an input pixel file as in Project 1. The data files are exactly the same as in Project 1.

In your Pig script, you can access the path of the input pixels file as '$P' and the output path as '$O'. That is, you can use LOAD '$P' USING ..., to load the pixels and STORE X INTO '$O' ..., to write the relation X to the output directory.

To run it in local mode over the small pixels file use:

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

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

Documentation

You can learn more about Pig at:

- Pig: Getting Started
- Pig Latin Basics

What to Submit