C312 Advanced Databases Course Work 3: Pig

Due in 12noon Thursday 1st December 2016

Background Material

The tables of the **uscensus1990** database (as used in Course Work 2) are available as a set of TSV files, accessed from CSG Linux Lab machines in the directory /vol/automed/data/uscensus1990. The exercise below requires that Pig scripts be written that are named after the question number, and generate the result in a directory also named after the question number.

Running an Example Script

For example, suppose there was a question 0 which asks

> *Produce a CSV file with the scheme (state_name,county_name,density), that contains the names of states with their corresponding counties and population density of counties, for all counties with a density of at least 1.0; and the file should include one entry for each state without any such counties, with the county name and density left empty. The results should be sorted by the state name and county name.*

Then a suitable script to store in file q0.pig would be:

```
-- Load the state, county, mcd, place and zip
RUN /vol/automed/data/uscensus1990/load_tables.pig

-- Information about county population density
county_density =
    FOREACH county
    GENERATE state_code,
             name AS county_name,
             ROUND(10.0*population/land_area)/10.0 AS density;

-- Counties with high population density
high_density_county =
    FILTER county_density
    BY density>1.0;

-- Find the counties of all states, but include states without any counties
state_and_county =
    JOIN state BY code LEFT OUTER,
         high_density_county BY state_code;

-- Project just those columns necessary for the query
state_and_county_projected =
    FOREACH state_and_county
    GENERATE name AS state_name,
             county_name,
             density;

-- Sort by state name
state_and_county_ordered =
    ORDER state_and_county_projected
    BY state_name,
       county_name;
```

```
STORE state_and_county_ordered INTO 'q0' USING PigStorage(',');
```

To run this file, copy it into your working directory by typing on the command line:

```
cp /vol/automed/data/uscensus1990/q0.pig .
```

Then run the q0.pig script using the command:

```
pig -x local q0.pig
```

After building the pipeline for the job, and running the pipeline, you should have a q0 directory created, containing a file part-r-0000 with the result of the STORE command in the script.
Note that if you wish to rerun the script, since the STORE command expects the directory is saves into not to exist, you must first delete the q0 directory using the command:

```
rm -Rf q0
```

Further information about Pig

The official manuals for Pig are found at `pig.apache.org`, the version of Pig installed in the labs is 0.9.2 and hence you should look the documentation for that release. Versions of Pig are available to install on your own linux computers from `www.cloudera.com`. There is information on various development environment options at cloudera.
For those users that like using emacs as their text editor, there is a syntax highlighting option for emacs, which you can enable on a CSG linux machine by adding the following two lines to your .emacs configuration file in your home directory:

```
(setq load-path (nconc '("/vol/automed/emacs") load-path))
(require 'pig-mode)
```

Debugging Pig Scripts

To debug a Pig script, it is normally better to use Pig in interactive mode. If you enter the command:

```
pig -x local
```

then you should have the interactive prompt returned:

```
grunt>
```

You may then cut and paste sections of the script you are editing in a text editor into the command line to determine if they parse correctly, and then use the DUMP command to see the intermediate results.
For example, you enter the command at the grunt prompt:

```
RUN load_tables.pig
```

you can view the contents of the state.tsv file using the command:

```
DUMP state;
```

If you then cut and paste from q0.pig the definitions of the county_density and high_density_county aliases, you may then type at the grunt prompt:

```
DUMP high_density_county;
```

to view what are the contents of the alias, and then enter subsequent aliases, and view the contents of those aliases with other DUMP commands.

Submission

To gain full marks, answers to the following questions should make full use of Pig commands to write compact and efficient scripts, and be laid out such that structure of the scripts is clear. The queries must also run correctly on the Pig installation provided by CSG Linux lab machines, and be submitted electronically by the coursework deadline to CATE four Pig scripts q1.pig, q2.pig, q3.pig, and q4.pig providing the answer to corresponding question below. Each script should output its answer to a corresponding directory named q1, q2, q3, and q4.

Questions

1. Write a Pig script that writes a CSV file with the scheme (state_name) containing all those state names in state for which there are no corresponding records in county. The result must be ordered by state name.

   marks 20

2. Write a Pig script which writes a CSV file with the scheme (state_name,population,land_area, where state_name is a name of a state, population is the total population of all county records for the state, and land_area is the total land area of the county records for the state. You should not include states with no county records. The result must be ordered by state name.

   marks 20

3. Write a Pig script that writes a CSV file with the scheme (state_name,no_city,no_town,no_village) where state_name is the name of a state, and no_town is the number of places with type equal to 'town', no_city is the number of places with type equal to 'city', and no_village is the number of places with type equal to 'village', in each state. The result must be ordered by state name. You should not include states where there are no place records.

   marks 30

4. Write a Pig script that writes a CSV file with the scheme (state_name,city,population) containing the state name and corresponding names and population of cities in place, returning only the five largest cities in each state. The result must be ordered by state name, with cities in each state listed in declining order of population.

   marks 30