C0576 Advanced Databases Course Work 3: Pig

Due in 12noon Friday 23rd December 2018

Background Material

The tables of the **uscensus1990** database are available as a set of TSV files, accessed from CSG Linux Lab machines in the directory[1] /vol/automed/data/uscensus1990. A frament of the data is illustrated below.

county

state_code	fips_code	name	type?	population	housing_units	land_area	water_area	latitude	longitude
36	103	Suffolk County	county	1321864	481317	2360093	3786993	40.90536	-72.679044
6	75	San Francisco County	county	723959	328471	120955	479639	37.793250	-122.554783
55	29	Door County	county	25690	18037	1250317	4887822	45.020683	-87.009973
55	59	Kenosha County	county	128181	51262	706605	1247185	42.582298	-87.805528
55	61	Kewaunee County	county	18878	7544	887475	1921659	44.589317	-87.440146
55	71	Manitowoc County	county	80421	31843	1532148	2336924	44.145467	-87.553328
34	9	Cape May County	county	95089	85537	661007	945551	39.077466	-74.858609
55	79	Milwaukee County	county	959275	390715	625649	2455935	42.975611	-87.671417
55	89	Ozaukee County	county	72831	26482	600784	2290434	43.249500	-87.501558
55	101	Racine County	county	175034	66945	862811	1188420	42.784761	-87.755094

mcd

state code	fips code	fips subdivision code	name	type?	population	housing units	land area	water area	latitude	longitude
1	3	91053	Fairhope	division	16331	7361	172078	194445	30.466407	-87.913337
72	117	26846	Ensenada	barrio	763	410	2881	10213	18.332828	-67.284330
1	3	91152	Foley	division	20687	17587	453800	674407	30.292000	-87.763677
72	97	1820	Algarrobos	barrio	5074	1649	4301	31018	18.209253	-67.194724
1	97	90216	Bayou La Batre	division	9705	4580	216863	671076	30.301019	-88.192562
72	3	70921	Ri]o Grande	barrio	864	292	2596	9242	18.395570	-67.235495
10	1	92220	Milford North	division	6758	2938	168299	215641	38.998743	-75.333951
66	10	7250	Agat	district	4960	1300	27192	48149	13.356057	144.633899
5	23	93750	Valley	township	749	660	18250	29743	35.496497	-92.105746

place

state_code	county_code	name	type?	population	housing_units	land_area	water_area	latitude	longitude
2	2025	Amchitka	CDP	25	0	299980	417405	51.567103	178.877380
60	60100	Olosega	village	201	47	2969	42842	-14.201212	-169.599688
2	4210	Atka	city	73	26	23772	70025	52.242218	-174.205154
53	56304	Priest Point	CDP	703	313	2470	7336	48.036906	-122.249727
2	13860	Chiniak	CDP	69	36	103269	192748	57.631863	-152.182537
72	65589	Puerto Real	comunidad	3429	1206	1116	1666	18.072680	-67.191123
2	82750	Wainwright	city	492	160	10557	30331	70.599953	-160.071563
48	72989	Tiki Island	village	537	441	1679	1814	29.298700	-94.914177
2	86490	Yakutat	city	534	189	7572	12124	59.557526	-139.762121

state

code	abbr	name
1	AL	ALABAMA
2	AK	ALASKA
4	AZ	ARIZONA
5	AR	ARKANSAS
6	CA	CALIFORNIA
8	CO	COLORADO
9	CT	CONNECTICUT
10	DE	DELAWARE
11	DC	DISTRICT OF COLUMBIA
12	FL	FLORIDA

zip

state_code	zip_code	zip_name	longitude	latitude	population	allocation_factor
1	35004	ACMAR	-86.51557	33.584132	6055	0.001499
1	35005	ADAMSVILLE	-86.959727	33.588437	10616	0.002627
1	35006	ADGER	-87.167455	33.434277	3205	0.000793
1	35007	KEYSTONE	-86.812861	33.236868	14218	0.003519
1	35010	NEW SITE	-85.951086	32.941445	19942	0.004935
1	35014	ALPINE	-86.208934	33.331165	3062	0.000758
1	35016	ARAB	-86.489638	34.328339	13650	0.003378
1	35019	BAILEYTON	-86.621299	34.268298	1781	0.000441
1	35020	BESSEMER	-86.947547	33.409002	40549	0.010035
1	35023	HUEYTOWN	-86.999607	33.414625	39677	0.00982

The state table contains all states and some territories of the USA, which for the purpose of this exercise will be all referred to as states. Each state is divided into counties or adminstratively equivalent units, which are stored in the county table. The type column of county identifies the type of adminstrative unit. Counties are further divivided into **minor civil divisions** (**mcd**) or adminsitratively equivalent ares held in the mcd table, and again each is associated with the type of unit held.

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

[1] Since the /vol/automed is automounted, you must use cd to change to the directory, rather than use a filer to navigate down to the directory.

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 lab machines is either 0.12.0 or 0.15.0, but for the exercises we are doing there are no practical differences between the versions. 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 /vol/automed/data/uscensus1990/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