Homework 6

Due: Friday March 15 at 11:59pm in compass2g.illinois.edu

For this assignment, submit one `.pig` script file containing all the necessary Pig code to generate your results and one text-based report file (Word doc, pdf, or .txt file). The **Pig script file should be an executable `.pig` file** (If using the Hue editor or run your code in steps at command line, you may copy the code and paste it into a Notepad++ file with extension `.pig`.).

To show results, write the necessary relation to a file from Pig using `STORE` within your code. This will make it easier to show results, and would also allow you to do exercises in different scripts if desired (by loading the stored results from a previous exercise). You can then run the necessary `cat` commands at command line to show the results, copy the command line commands and results into a text-based file and add comments to explain results.

Any code based on code from elsewhere (e.g. code provided with the text) must reference in comments the source of the original code.

Use Pig for all exercises. All exercises are based on processed variations of the weather data we have worked with. The data files are available in the course space.

Be sure to state units for results. (You do not need to convert from tenths of degrees Celsius, but you do need to comment on the units when presenting your results.)

The `Data1920s.txt` file is a tab-delimited file containing usaf identifier, wban identifier, and temperature for observations from 1920 to 1929.

The `StationCodes.txt` file contains the usaf identifier, wban identifier, and location name for the weather stations in the metadata from the text set (see `input/ncdc/metadata/stations-fixed-width.txt` in the Hadoop Book source files). `UNKNOWN*` names have been substituted for locations with names missing in the original data and locations with numbers for names in the original data have been renamed `CODEDLOCATION*` where `*` is the number in the original data.

To get the data into the virtual machine, you can upload through the File Browser in the Hue web interface and move the files as needed.

Exercises for All Students

Exercise 1:

Create a relation that joins the observed temperature data with the station name data, so the location name will be included within each observation in the relation.

Rather than show the entire relation, use the `LIMIT` keyword to show 5 entries from the relation (see the end of the Sorting Data section on page 408 of the text to see how to use `LIMIT`).

Exercise 2:

Obtain the number of trusted temperature observations and the minimum, average, and maximum temperatures by station for each station from in the data. Show the first 10 results.

Note: The data has already been filtered for missing temperatures and bad quality codes, so you do not need to do filtering to get trusted observations.

Exercise 3:

For the station with the lowest minimum temperature, obtain the name, the minimum, average, and maximum temperatures for each year from 1920 to 1929, and show your results.

Note: Not all stations have temperature observations for every year, thus the station with largest temperature might not have results for every year.

Additional Exercise for Graduate Students

Exercise 4:

Obtain the maximum temperature deviation above the average (max temperature – average temperature) and the maximum temperature deviation below the average (average temperature – min temperature) for each recorded station location for the period from 1920 to 1929 (you do not need to show this result).

Programmatically find the station name and maximum temperature deviation above the average for the station with the largest maximum deviation above the average for the time period (`ORDER` and `LIMIT` should be useful for getting this information from the range data). Then obtain that station's maximum temperature deviation below the average by year for each year from 1920 to 1929.

Note: Again, not all stations have temperature observations for every year, thus the station with largest temperature range might not have results for every year.