Homework 6

Due: Friday April 1 at 7pm

For each exercise, submit the necessary script files, and the results generated. **Script files should be the actual script files**, not code pasted into some other document. To show results, you can write the necessary relation to a file from pig. **All files should be submitted in one zip file.**

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 variations of the weather data we have worked with in class. The data files are available in the course space. The `19011910.txt` file is a tab-delimited file containing usaf identifier, wban identifier, and temperature for observations from 1901 to 1910. The `stationlistshort.txt` file contains the usaf identifier, wban identifier, and location name for the weather stations with recordings from 1901 and 1910. `UNKNOWN*` has been substituted for locations with names missing in the original data set (see `input/ncdc/metadata/stations-fixed-width.txt` in the Hadoop Book source files).

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 10 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 and maximum temperatures by station for each station from in the data.

Exercise 3:

For the station with the highest maximum temperature, obtain the minimum and maximum temperatures for each year from 1901 to 1910.

Additional Exercise for Graduate Students

Exercise 4:

Obtain the temperature range (max temperature – min temperature) for each recorded station location for the period from 1901 to 1910. Programmatically find the station name and temperature range for the station with the smallest temperature range for the time period (`ORDER` and `LIMIT` should be useful for getting this information from the range data), and then obtain that station's temperature ranges by year for each year from 1901 to 1910.