# Homework 5

**Due: Friday March 8 at 11:59pm in compass2g.illinois.edu**

For this assignment, submit one zip file containing all the necessary script files (map and reduce scripts) and one report file outside of the zip file. **Script files in your zip file should be the actual script files**, not code pasted into another document.

Your report should be a Word, pdf or plain text file (e.g. `.txt`; Compass will not allow you to upload `.sh` files). For your report, include the command line code needed to run the job (including any command line commands to download data files, copy data to HDFS, and show results obtained on HDFS), and the results generated. Results must be clearly labeled (e.g. year in the results, and column labels in the results or explanation of the columns in a report file so it is clear to the reader what the quantities are). There should be short descriptive text for describing the results. Be sure to state units.

Use python for the map and reduce scripts. You may need to make your script files executable by `all` using **`chmod a+x`** ***`filename`*** so they can be run by Hadoop. Any code based on code from elsewhere (e.g. code provided with the text) must reference in comments the source of the original code.

All exercises are based on NCDC weather data like the data we have worked with in class. You will need to download the files for the specified year ranges below.

## Exercises for All Students

Note: In python, you will also want to use `float` for non-integer arithmetic, rather than `int` when dividing numbers. Using `int` will result in integer arithmetic, so remainders will be dropped instead of resulting in decimals so there would be truncation errors with `int`.

### Exercise 1:

Using Hadoop and MapReduce, find the minimum monthly recorded air temperature from 1915 to 1924 and return those minimum values in degrees Celsius. (You should have 12 values total, one for each month).

### Exercise 2:

Using Hadoop and MapReduce, obtain the number of trusted temperature observations and the minimum and maximum monthly temperatures in degrees Fahrenheit over the period of 1915 to 1924. Make sure your code only goes through the data once to get these results (to do this you will need to update the minimum, maximum, and count at the same step in the code).

### Exercise 3:

Using Hadoop and MapReduce, obtain the total number of air temperature observations that are not missing for each month during the period from 1915 to 1924 and the total number of observations with acceptable quality codes for each month during that period. Make sure your code only goes through the data once to get these results (to do this, you could have the mapper return (month, tempcount, validqcount) for each observation, and have the reducer aggregate).

## Additional Exercise for Graduate Students

### Exercise 4:

Using Hadoop and MapReduce, obtain the monthly mean air temperature in degrees Celsius for the period from 1915 to 1924. If you use a combiner, make sure your code will work when data needs to be recombined from samples of different sizes.