**HW#2: Big Data w/ Hadoop**

**Introduction:**

1. Weather data from 2010 to 2019 is available at /data/weather/<year>
2. Each year data is in the respective directory (i*.e., 2010 directory have data for year 2010*)
3. Additional data information can be found in the files:
   1. ish-history.txt
   2. readme.txt
   3. country-list.txt
4. Use spark to analyze the above-mentioned data.
5. Example commands used in the class is also attached: Temp\_Test.html
6. Ignore the missing values for the dataset (readme.txt files attached).
7. Drop columns which we dropped in the class (refer Temp\_Test.html file).
8. If you are unable to load all the data at once, load data for one year at a time.

**Tasks:**

1. Find the hottest and coldest day along the station code and date for each year.
2. Find the hottest and coldest day across all years (2010 - 2019) along with station code and date.
3. Maximum and minimum precipitation with station code and date for year 2015.
4. Count percentage missing values for mean station pressure (STP) for year 2019 and stations.
5. Station code with maximum wind gust and date for year 2019

Submission:

1. Your result.txt file with all the above information
2. Your spark code/command

----------------------------------------------------------------------------------------------------------------

**Hints (to minimize confusion):**

1. Find the hottest and coldest day along with the station code and date for each year.

* *Use column “MAX" and “MIN” to find the Hottest and Coldest temperature.  This is clearly stated in README.TXT.*
* *There can be a situation where there is more than one date with hottest or coldest temperature. Do LIMIT of 1 in your query and share that value, along with the corresponding Station code.  Each student could have different Station code in your answers, but MAX or MIN temperature would be the same.*

1. Find the hottest and coldest day across all years (2010 - 2019) along with station code and date.

* *Similar to the hints in Question 1, however, the year might also differ, also with the Station code.  However, the MAX or MIN temperature would be the same.*

1. Maximum and minimum precipitation with the station code and date for the year 2015.

* *Column PRCP represents precipitation.  Similar to the hint in Question1, the Station code could be different for each student if there is more than one MAX or MIN precipitation for the year.*

1. Count percentage missing values for mean station pressure (STP) for the year 2019 and stations.

* *Look at the README.TXT.  The missing value is given some version of 9999.9’s.  Do not worry about the station code.*

1. Station code with maximum wind gust and date for the year 2019

* *Use the "Maximum wind gust" column and the corresponding Station code.  As before Station code could differ for each student depending on the code is written.*