CS236 Spring 2018
Project Description

Christina Pavlopoulou (cpavl001@ucr.edu)

PROJECT GOAL

You are given two types of datasets, one with information about weather stations across the world and one with weather readings for each station for a 4-year period. The final goal of the project is to find out the most stable, in terms of temperature, US states.

DETAILED DESCRIPTION

To complete this project you have to fill the following steps:

1. For stations within the United States:
    a. **(if there is STATE information)** Group the stations by state and calculate the average latitude and longitude for each one (State centroid).
    b. **(if there is not STATE information)** For the stations that belong to US but the State tag is empty find the suitable State by calculating which centroid is the closest (calculate the difference between the coordinates of each station with the unknown state and each centroid and find the minimum one)
2. For each state find the average temperature recorded for each month (ignore the year).
3. Find the months with the highest and lowest averages for each state.
4. Order the states by the difference, ascending.

Each row of your output should contain: The state abbreviation, the average temperature and name of the highest month, the average temperature and name of the lowest month and the difference between the two.

DATASET DESCRIPTION

The dataset with the station information is contained in a .txt file. Remember that you will only need the US stations. Following are the fields for this dataset:

`ID:` is the station identification code. Note that the first two characters denote the FIPS country code, the third character is a network code that identifies the station numbering system used, and the remaining eight characters contain the actual station ID.
LATITUDE: Ignore for the project
LONGITUDE: Ignore for the project
ELEVATION: Ignore for the project
STATE: is the U.S. postal code for the state (for U.S. stations only).
NAME: Ignore for the project
GSN FLAG: Ignore for the project
HCN/CRN FLAG: Ignore for the project

WMO ID: Ignore for the project
Note that the file will not have the above headers (only the raw data). Each field is separated by spaces but the number of spaces varies. So be careful while parsing the fields.

The dataset with the weather readings information is split into 4 .csv files (each represents 1-year period). Following are the fields for this dataset:
`ID:` is the station identification code.
YEARMODA: The datestamp
OBSERVATION TYPE: Type of the temperature
VALUE: Temperature
Ignore all the remaining fields of each file for this project. Note that for this project you will need only the temperatures that their type is TAVG (average temperature for the date).

THINK ABOUT

1. Take into account the size of the files when you do joins in mapreduce.
2. Pay attention that your files have two different formats, thus they are not parsed in the same way.
3. How many passed should you do? How much work for each pass?
4. Start early since the sever will get slower when more people use it.