CS236 Class Project (Due December 8, 2015 before midnight.)
Fall 2015
TA: Joobin Gharibshah (jghar002@ucr.edu)

Project Summary:

For this project, we will have two datasets. The first provides station information for weather stations across the world. The second provides individual recordings for the stations over a 4-year period. The goal of the project is to find out which states in the US have the most stable temperature (i.e. their hottest month and coldest month have the least difference).

Formal Problem:

For stations within the United States, group the stations by state. For each state with readings, find the average temperature recorded for each month (ignoring year)
Find the months with the highest and lowest averages for that state. Order the states by the difference between the highest and lowest month average, ascending.
For each state, return:
The state abbreviation, e.g. "CA"
The average temperature and name of the highest month, e.g. "90, July"
The average temperature and name of the lowest month, e.g. "50, January"
The difference between the two, e.g. "40"

Dataset Information:

The locations dataset is a single .csv file, containing the metadata for every station across the world (We only care about the stations with "ST" information). Keep in mind that the first row of this file is Header. Here are the fields for this dataset:
USAF = Air Force station ID. May contain a letter in the first position.
WBAN = NCDC WBAN number
CTRY = FIPS country ID
ST = State for US stations
LAT = Latitude in thousandths of decimal degrees
LON = Longitude in thousandths of decimal degrees
ELEV = Elevation in meters
BEGIN = Beginning Period Of Record (YYYYMMDD).
END = Ending Period Of Record (YYYYMMDD).
Sample Row:
"724920","23237","STOCKTON METROPOLITAN AIRPORT","US","CA","+37.889","-121.226","+0007.9","20050101","20140403"

The recordings dataset is contained in four files, one for each year. The "STN---" value will match with the "USAF" field in the locations dataset. These files are concatenated from many small files, so keep in mind that there will be Header lines through the files. Here are the fields for this dataset:
STN--- = The station ID (USAF)
WBAN = NCDC WBAN number

YEARMODA = The datestamp
TEMP = The average temperature for the day, followed by the number of recordings
DEWP = Ignore for this project
SLP = Ignore for this project
STP = Ignore for this project
VISIB = Ignore for this project (Visibility)
WDSP = Ignore for this project
MXSPD = Ignore for this project
GUST = Ignore for this project
MAX = Ignore for this project (Max Temperature for the day)
MIN = Ignore for this project (Min Temperature for the day)
PRCP = Ignore for this project (Precipitation)
NDP = Ignore for this project
FRSHTT = Ignore for this project
Sample Row:
997781 99999 20061121 42.4 13 9999.9 0 9999.9 0 9999.9 0 999.9 0 17.5 13 22.0 999.9 46.2* 39.0* 0.00I 999.9 000000

Due Date:

This project will be due Tuesday, December 8, before Midnight. Email your submissions to the TA, one email per group.

Deliverables:

A zipped file containing:
The script to run your job
A README describing your project, including:

0. Your usernames and node number.
1. An overall description of how you chose to separate the problem into different mapreduce jobs, with reasoning.
2. A description of each mapreduce job including:
    - What the job does
    - An estimate of runtime for the pass
3. A description of how you chose to do the join(s)
4. A description of anything you did extra for the project, such as adding a combiner. If there is anything that you feel deserves extra credit, put it here.

All files needed for the script to run (Don't include the two original datasets).

The script should take as input three things:

- A folder containing the Locations file
- A folder containing the Recordings files
- An output folder

The TA will provide these when running your script. Please change the hdfs interactions to be based on /user/jgahr002 before you submit (it should be your own folder when you run it).

Think About:

1.There are different ways to do joins in mapreduce. What can you use from the datasets to inform your decision (e.g. size)? Please Google joins in mapreduce for more information

2. Make sure your parse correctly. One file is a csv, the other is not. One file has a single row of Headers, the other has many.

3. How many passes should you do? How mush work should each pass do?

4. Make sure to start early. The cluster will get slower as more people use it. A single pass over the data might take about 2 minutes, even while the cluster is not loaded.

Potential For Extra Credit:

Please feel free to try to beef up your project for extra credit. There are many ways that you can do this. Here are a few examples:

A good use of combiners

A clever way to achieve faster execution time

Enriching the data, e.g. including the average precipitation for the two months

Bigger Bonus:

Include the stations with “CTRY” as “US” that don’t have a state tag, finding a way to estimate the state using a spatial distance with the known stations. There are some stations that are Ocean Buoys so you may want to have a maximum distance to be required in order to be included in a state, or you could create a separate “state” representing the “pacific” and “atlantic” ocean (checked by using coordinates). There is a lot of potential work here so the extra credit could be large.

Whatever you try to do let the TA know in your README. If you aren’t sure whether your idea is worth extra credit or not, just email the TA.