Project 2 (FINAL) : BigData Analysis Project - Weather

Due: April 30 2018 @ 11:55pm

This project will have you perform Data Analysis and processing using MapReduce/ Apache Spark. The Project will use the weather dataset from https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/ .This project will use only 19 years of data (2000 - 2019) for all the stations starting with US and elements TMAX, TMIN. The dataset is available on the CEAS hadoop directory /user/tatavag/weather

If you prefer to generate the datasets on your own. Please use the following shell script

```
for i in `seq 2000 2019`
do
wget https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/${i}.csv.gz
gzip -cd ${i}.csv.gz  | grep -e TMIN -e TMAX | grep ^US > ${i}.csv
done
```

The following information serves as a definition of each field in one line of data covering one station-day. Each field described below is separated by a comma (,) and follows the order presented in this document.

ID = 11 character station identification code
YEAR/MONTH/DAY = 8 character date in YYYYMMDD format (e.g. 19860529 = May 29, 1986)
ELEMENT = 4 character indicator of element type
DATA VALUE = 5 character data value for ELEMENT
M-FLAG = 1 character Measurement Flag
Q-FLAG = 1 character Quality Flag
S-FLAG = 1 character Source Flag
OBS-TIME = 4-character time of observation in hour-minute format (i.e. 0700 =7:00 am)

See section III of the GHCN-Daily ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt file for an explanation of ELEMENT codes and their units as well as the M-FLAG, Q-FLAGS and S-FLAGS.

The OBS-TIME field is populated with the observation times contained in NOAA/NCDC's Multinetwork Metadata System (MMS).

In particular, it will have you build a hadoop map/reduce or Apache Spark that yields the following analysis.:

- Average TMIN, TMAX for each year excluding abnormalities or missing data
- Maximum TMAX, Minimum TMIN for each year excluding abnormalities or missing data
- 5 hottest , 5 coldest weather stations for each year excluding abnormalities or missing data
- Hottest and coldest day and corresponding weather stations in the entire dataset

BONUS QUESTION (10+15 POINTS)

- Median TMIN, TMAX for each year and corresponding weather stations
- Median TMIN, TMAX for the entire dataset

You may reuse any of the components that were used to build the homeworks, as well as the in-class examples, or you may build a new application from scratch - it is up to you. You can either use the CAES cluster or your own Hadoop VM. You can use one or many of the following techniques to solve the analysis problem

- Map/Reduce in Java
- Map/Reduce streaming (python or other languages)
- Apache spark - (RDD)
- Apache Spark (Spark Dataframes)
- Apache Spark (SparkSQL)
- Hive SQL (Not available on CAES cluster though)

The final deliverable would be a report (preferably PDF format) which answers the above analysis questions. You can either represent the results as Tabular or Plot or both. Please explain the method you have used to solve the analysis task and why you choose that method (eg. spark or map reduce etc). The appendix should include all the code used for analysis.

NOTES / HINTS:

http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

https://github.uc.edu/tatavag/CloudComputing2019/wiki/Spark-Examples

https://github.uc.edu/tatavag/CloudComputing2019/wiki/Hive-Examples