Challenge HADAS-01

Look for the highest temperature of the year[1]

Summarization design pattern

1.1 Problem statement

Problem: Given a list of temperature readings done during two years, look for the highest temperature measured during those years.

Computing the maximum of a given field is an excellent application of the numerical summarization pattern. After a grouping operation, the reducer simply iterates through all the values associated with the group and finds the max (or min), as well as counts the number of members in the key grouping. Due to the associative and commutative properties, a combiner can be used to vastly cut down on the number of intermediate key/value pairs that need to be shuffled to the reducers. If implemented correctly, the code used for your reducer can be identical to that of a combiner.

For this exercise:

- Use the previously collected data 1901, 1902, sample.txt and sample2.txt from http://vargas-solar.com/bigdata-fest/challenges/mr-patterns-on-an-elephant/
- Use the given code files in the same URL to complete them with your map & reduce functions and test them on the hortonworks hadoop environment.
- Results: present the top temperature for each input.

1.2 Implementation

At the end you should be able to:

- Explain the principles and utility of the maximum (summarization) pattern
- For more data download a dump from NCDC - National Climatic Data Center is available at http://nomads.ncdc.noaa.gov/data.php?name=access (50MB at least).

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 63.