Challenge HADAS-02

Top ten users by reputation[1]

Filtering design pattern

1.1 Problem statement

Problem: *Given a list of user information, output the information of the top ten users based on reputation*.

Determining the top ten records of a data set is an interesting use of MapReduce. Each mapper determines the top ten records of its input split and outputs them to the reduce phase. The mappers are essentially filtering their input split to the top ten records, and the reducer is responsible for the final ten.

For this exercise:

- Use the previously collected data Users.xml 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: list the top Users.

1.2 Implementation

At the end you should be able to:

- Explain the principles and utility of the top-k (filtering) patterns.
- Test your implementation with different data collection sizes and compare the results.
- For more data download a dump from https://archive.org/details/stackexchange (50MB at least).

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