Binning by Hadoop-related tags[1]

Binning pattern (Challenge 5_c)

For this exercise you have to prepare a report including:

- Your data collections comment describing what they represent.
- Your source codes: a file precising in a comment the compiling and execution commands.
- Propose a test battery and report execution measures.
- Compress everything and sent a .zip or a .tar to genoveva.vargas@gmail.com

1.1 Problem statement

Problem: Given a set of StackOverflow posts, bin the posts into four bins based on the tags hadoop, pig, hive, and hbase. Also, create a separate bin for posts mentioning hadoop in the text or title.

1.2 Implementation

Look at page 90 of the book "Map Reduce design patterns" and see the proposed `Map`, `Driver`, and `Reduce` codes. Prepare a data collection of your choice and implement the solution. As in previous challenges, you can use StackOverflow or any other source for generating your collection. Program the first part of the example.

- Explain the use of `MuttipleOutputs` and its role in the phase `map`.
- Explain the "absence" of a `reducer`.
- Prepare collections of different sizes to run your tests trying to get to the limits of your solution.
- Make comparisons. Do not hesitate to prepare graphics.

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 90.