Distributed grep[1]

Summarization pattern (Challenge 3_a)

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

Grep is a popular text filtering utility that scans through a file line-by-line and only outputs lines that match a specific pattern. We want to parallelize the regular expression search across a larger body of text. In this example, we'll show how to apply a regular expression to every line in MapReduce.

1.2 Implementation

Look at page 47 of the book "Map Reduce design patterns" and see the proposed `Map` and `Reduce` codes. Prepare a data collection of your choice and implement the solution.

- Explain the use of simple random sampling. Why does the solution proposes to grab a subset of a larger dataset?
- 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.
- Explain and discuss why is it not possible (as stated in the book) to have a combiner. Support your arguments with examples.

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 47.