Partitioning users by access date[1]

Partitioning pattern (Challenge 5_b)

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

Given a set of user information, partition the records based on the year of last access date, one partition per year.

In the StackOverflow data set, users are stored in the order in which they registered. Instead, we want to organize the data into partitions based on the year of the last access date.

1.2 Implementation

Look at page 86 of the book "Map Reduce design patterns" and see the proposed `Map`, `Partitioner`, 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.

- Why does the `partitioner` must be configured? And what about the number of `reducers`? Explain using examples.
- Explain the role of the `Partioner`.
- 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. 86.