◄ Copy of the small test about Spark Dataframes (not graded)   Report 3 (about TD 6 - Top K in Spark) ►

✔ Terminé

(For each of the following questions, you will need to run experiments. You may choose the data to use (among the ones available at LSD or you may create/obtain others). However, it is important to notice the results may depend on the data size.)

We have studied and discussed the best algorithm for finding the top K in MapReduce.

1. In Spark, we may use the `.top(K)` action to obtain directly the top K from an RDD. Is it the most efficient way? To test it, you are asked to implement top K "by hand" by using `mapPartitions` to keep the top K of each partition and then calling `top` on the resulting RDD, and compare that to simply using `top(K)`. As a baseline, also compare it to doing `sortByKey` followed by `take`.
2. There is no `top` action for dataframes, meaning the only way of finding the top K is to sort the data using `orderBy` and then obtaining the first K. However, a quick Google search seems to indicate that if we do `orderBy` followed by `limit(K)`, Spark will be smart enough to only find the top K and not sort the whole data set. Is it true?To test it, compare it to `orderBy` followed by `show(K)` and to converting it to an RDD and then doing `top`.

Report

For the third report, which will make for 30% of your grade for this course, you are asked to write about this activity. You are asked to:

- answer the questions above and explain your study in details: what experiments you designed to answer your questions (including code and what data set you used) and why, your experimental methodology, the obtained results, your conclusions from them.
- Respect a page count limit of 3 and submit your work in .pdf format.
- Work individually and submit your own report (i.e. one report per person).
- As long as not excessive, figures and code snippets may also be added to an appendix.
- Not copy anything from the internet or from colleagues, to not ask for solutions in online forums, and to cite all external sources you use.
- Write in French or English (or Portuguese :).

Grades will be based on:

- quality of the report: readability, clarity, organization, etc;
- scientific rigor applied to the study, including but not limited to experimental methodology;
- correctness of the provided arguments and explanations.

Modifié le: Friday 2 December 2022, 11:32