Project description. The project cosists of 3 parts.

(\*) The first part consists of extending and consolidating the k-means implementation in Spark Scala, and more specifically

-

-

-

-

-

producing a baseline version with no optimization

Identifying weaks points concerning performance, by using your datasets and first experimental results

Implementing optimization, providing justifications, and experiments showing the imrovement in terms of execution time

Implementing k-means ++ and comparing performances with the previous implementation Providing an implementation using DataFrames (you can use RDD to load the data and for some initial step, zipWithINdex for instance)

Comparing performances between the iprimizied RDD version and the DataFrame version, by using you data stes

All experiments should be well commented (without providing very long comments anyway)

IMPORTANT: groups consists of two students at most - the same data set cannot be used in more than one group

You can use DataBricks.

(\*\*)The second part is still about K-means, and this time you are asked to implement the algorithm by using DataFrames or DataSets.

Perform the same experiments done with the RDD-based implementation and then compare results about time execution. Discuss differences in terms of performance between the RDD and DataFrame/DataSet implementation.

(\*\*\*)The third part is about Gradient descendant and in this part you are asked to provide RDD based immplementaiton of the Momentum and Nesterow variants of SGD (see slides).

Deadline September 1st 2024. Please send an HTML file including everything: text descriptions, as well as code and comments, introduction, final discussion, conclusion.