Project

Guidelines and deadline

You are asked to form groups of 2/3 students max. One student per group has to communicate by email (to dario.colazzo@polytechnique.edu) the group composition, by February 12, 2021

Each group has to chose one of the following problems (see next slides), design algorithms for the solution, perform experimental analys as indicated and write a report including:

1. a description of the adopted solution 4 points
2. designed algorithms plus related global comments/description 4 points; comments to main fragments of code 3 points
3. experimental analysis, concerning in particular scalability 3 points
4. comments about the experimental analysis outlining weak and strong points of the algorithms. 3 points
5. an appendix including all the code the code. 2 points

A pdf version of the report has to be sent via email (dario.colazzo@polytechnique.edu) before March 30, 2021.

A pdf version of a pre-report has to be sent via email (dario.colazzo@polytechnique.edu) before March 10, 2021

New!

Finding connected components in graph

- The algorithm is descibed in this paper
  - https://www.cse.unr.edu/~hkardes/pdfs/ccf.pdf
- The work to consists of understanding the MapReduce algorithm, and coding it into Spark by using both RDD and DataFrames
- Both Python and Scala implementations must be provided
- Experimental analysis comparing the RDD and DataFrame versions has to be conducted on graphs of increasing size
- For small graphs use Databricks, for bigger ones (<20GB) use the AWS cluster
- Pre-report: presentation of the problem, solution and at least one RDD implementation.

Data preparation for US flight delay prediction

- Consider the following paper
  - https://www.dropbox.com/s/4rqnjueuqi5e0uo/TIST-Flight-Delay-final.pdf
- Half of the paper is dedicated to data preparation by preprocesing and opportunely *joining* complex datasets about flights and weather conditions
- The project consists in porting in Spark (DataFrame) the data preparation process.
  - Scala or Python
  - RDD and DataFrames for the data wrangling part (the optional ML part only in DataFrames)
- The report should detail each step, comment encountered difficulties and how these have been overcome.
- Optional: the group can then opt for performing in SparkMLinb the prediction analysis by using decision trees.
- Suggestion: a group of three students would be preferable, dataprep may be time consuming.
- Why not: once data is prepared use Spark ML for prediction, by using ML techniques adopted in the paper, or other ones you deem more efficient.
- Pre-report: presentation of the problem, solution and at least one RDD implementation.

Fast Matrix Factorization for Online Recommendation with Implicit Feedback*

- Consider this paper
  - https://www.dropbox.com/s/1nqw7zvmo91gq3a/IMPORTANT-70805024%20copy.pdf
- It descibes/presents an efficient matrix factorization ALS algorithm
- The paper shows that the algorithm can be easily parallelized, but no MR implementation is provided
- The work consists of finding a Spark version of the algorithm (both RDD and DataFrames) and perform experimental analysis in one of the dataset used by the paper.
- Pre-report: presentation of the problem, adopted solution and at least one RDD implementation.

Project on Kmeans

- Make experimental analysis of the basic Python version, eventually by lowering the number of iterations
- Find, describe, and implement otpimizations
- Use bigger input data instead of Iris, or bigger Iris instances, and perform experiments
- Switch to Scala
    - RDD
    - Dataframes
    - Datasets
    - Pick a sufficiently large input and make experiments to compare Scala implementations
- Write a technical report, around 20 pages max, with main points about implementations and experiments.
- **Pre-report: presentation of the problem, solution and at least one RDD implementation.**

You can mount your own project

- Find an interesting problem admitting MapReduce implementation in Spark
- Write a one-page proposal and submit it to me by February 12 (by email).
- I will let you know wether the proposal is good and eventually how to adapt it.
- Both RDD and DataFrame implementation.
- Pre-report: presentation of the problem, solution and at least one RDD implementation.