DC- DSE Assignment 2024/25

Each group (1-3 students) has to submit a short document (max 5 pages, notebook or PDF) divided into 3 parts:

P1: A presentation of a data processing problem to be solved using PySpark. The presentation must describe the dataset(s) and at least 5 queries/tasks of interest for the considered problem. The problem must be complementary to the labs presented in the course.

P2: A presentation of the PySpark solutions to the tasks described in P1.
The source code must be copied to a folder, named DC24, in your **user_dc_XY** home.

P3: A summary of the execution statistics for the PySpark programs in P2, e.g. using local[K] for K>=1, local[*], and on the cluster mode (if applicable).

Submission and Deadlines

The presentation (P1+P2+P3) must be uploaded to Aulaweb via a subat least 3 days before the date of the written exam in the session in which you plan to take the exam.