Information and Guidelines for the Final Project
Distributed Processing Exam (MAPD-B)

Latest revision: **15-May-2025**

General Information

- For the Data Processing final project, students must work in groups of 3 or 4 to complete a distributed processing task.
- Once a group is formed and ready to start working on the project, the students must fill out the spreadsheet available on the Moodle page with their member names, project title, and computing resources used.
- Upon completion, **all project materials** (code, Jupyter notebooks, plots) **must be submitted at least 3 days before the exam date**.
- The oral exam includes a 20–25 minute group presentation followed by individual questions. Each student must clearly describe their contributions and answer questions on both the project and course topics.

Project Guidelines

1. Projects will be evaluated based on complexity, methodology, completeness, and ingenuity.
2. Projects must **focus on using distributed frameworks** covered in the course (e.g., PySpark, Dask, MapReduce, Kafka) and involve setting up and managing a small computing cluster.
3. Groups may either use physical machines (e.g., a cluster of laptops, or Raspberry Pi/Jetson Nano boards) or CloudVeneto Virtual Machines. **Each group is responsible for configuring and maintaining its own cluster**.
4. A list of predefined (”boilerplate”) projects is available on Moodle, labeled by difficulty. Students may choose among these, or propose custom projects (see point 6).
5. **Performance benchmarking is mandatory**. Groups must study and analyze how the performances (at least the execution time) is affected by the cluster’s parameters (at least the number of dataset partitions and the number of processing units). **Projects submitted without benchmarks will be considered incomplete**.
6. Groups may propose custom, non-default projects. Non-default projects may also stem from topics related to other courses (e.g. Laboratory of Computational Physics), but **require prior approval** by the Professor to ensure they meet distributed processing objectives.