Information and guidelines concerning the final project for the MAPD-B distributed processing exam

Latest revision: 24-05-2022

General information

- As data-processing final project, students are required to work in small groups (either 3 or 4 people) and work on a distributed processing task.
- All students must fill out the spreadsheet available on the Moodle to specify their group's composition and the title of the project.
- The project code and/or other material (e.g. Jupyter-notebooks) will have to be submitted at least 3 days before the exam date.
- The data-processing oral discussion will consist of a short presentation (20-25 mins) about the group's project, followed by individual questions on the topics covered in the Data Processing part of the course.
- During the exam each student in the group should highlight his/her specific contribution to the project; questions will be asked individually to each student to assess the understanding of the topics discussed during the lectures.

Projects' guidelines

1. All projects will be evaluated for their complexity, methodology, ingenuity, and completeness.
2. A small set of "default/boilerplate" projects are provided on the Moodle webpage of the course. Students are free to choose them as their final project, although students are also allowed (and invited) to propose different projects for evaluation.
3. The list of "boilerplate" projects offers a degree of difficulty levels, clearly mentioned in the project description, which will be taken into account when assessing the final grade.
4. All projects must revolve around the application of the tools and techniques related to the distributed nature of the frameworks discussed throughout the course (pySpark / Dask / MapReduce / Kafka / ...)
5. All projects will have to be carried out by setting up and managing small clusters of VirtualMachines, using CloudVeneto computing resources.
6. Once a group is ready to start working on their project, they have to contact the Professor AND (not OR) the relevant Teaching Assistants, to be assigned a few (3) dedicated VMs for the project.
7. As all projects are centered on the topic of distributed processing, all groups will have to benchmark and discuss the performances of their computing projects. This includes the change of the cluster parameters such as the number of workers, the number of partitions, etc.

8. It is possible to use as a starting point for the development of a new project the topics assigned for other courses (e.g. Laboratory of Computational Physics - Module B exam projects).

9. Typical problems might include the use of techniques such as clustering, classification, regressions, etc, and/or using large datasets. Data from all fields of application are suitable (physics / astro / industrial / ...).

10. It is however strictly required to contact the Professor and Teaching Assistant to illustrate any ”non-default” project *before* starting working on it. The Professor and TA will decide whether the proposed project is going to be considered suitable for the exam or not.