Data Management

2023/2024

Projects

Instructions and Guidelines

SAPIENZA
UNIVERSITÀ DI ROMA

Tutor
Roberto Maria Delfino

Topics

- General instructions
- Project approval and presentation modalities
- How to write a project proposal
- How to ask for project presentation
- Evaluation
- Possible ideas for projects
- Useful resources
- General advices

Options to pass the exam

Two modalities exist to pass the Data Management exam: written test only (Option 1) or project + written test (Option 2).

Differences between the two options:

1. Option 1: full standard written exam (passing score $x \in [18,30]$)
2. Option 2: shortened written exam (passing score $x \in [15,24]$) plus a practical project (passing score $y \in [4,8]$). Final grade $z = x + y$ ($z \in [19,32]$)

NB: projects can be carried out either **individually** or in **groups of two students**

More details can be found in the course website

Projects approval and presentation

In case you choose Option 2 (written test + project), please do the following:

1. Send me your project proposal* (delfino@diag.uniroma1.it) and wait for it to be approved (do not start working before receiving my feedback)

2. After completing the project, prepare a 10 to 15 mins slideshow presentation (stick to the time!)

3. Send me an e-mail containing the material* of your project specifying if you prefer to have your presentation in person or remotely

* *details in the next slides*

Project proposal

Write a single-page pdf document including:

- Names and student IDs (matricola) of the group members
- Chosen type of project (DW, DBMS comparison, use of a NoSQL tool, etc.)
- Description of the dataset which you intend to use (include a reference link)
- Brief description (5-10 lines) of the work you intend to do

Send me an e-mail having:

- **[DM] Project Proposal** as the subject
- The pdf document as described before as an attachment
- All group members participating in the conversation (either as the sender or in cc)

Project presentation

Once you are ready to have your project discussion, send me an e-mail having:

- **[DM] Project Discussion** as the subject
- Your slides as an attachment
- Any additional material you think might be useful (link to repository, data, etc.)

I will reply by proposing you an appointment to have your project discussion.

Projects evaluation

Projects will be evaluated with a score ranging from **0** to **8** points. Points will be added to the grade obtained in the **written test** (**0** to **24**).
In order to have your exam registered you have to pass both parts (obtain at least 4 points for the project and at least 15 for the written test).

If project requirements are fulfilled you will have your project marked as passed and all group members will receive an e-mail assessing the achievement and the obtained grade.

If your project does not meet all the requirements or lacks in some aspects you will be asked to address them and come back for another appointment later on.

There is no expiration date for the grades you get for the project or for the written exam. You can do them whenever you like, in the order you prefer.

Ideas for possible projects

[DW] Select a set of data sources, integrate them by means of ETL operations (you can either use an integration tool or write simple scripts to solve common ETL problems), define the DFM model for your data, define and populate the corresponding star schema or snowflake schema, by using a relational DB (e.g., Postgres)
Focus: ability to combine data from different datasets in order to produce a relevant unified source of analysis

[DW] Identify a problem concerning data analysis and try to compare different approaches to address the proposed problem (e.g., carrying on the analysis via a relational DBMS and through a DW tool)
Focus: ability to compare different technologies

[DW] Select a dataset concerning a domain of your interest, carry out some relevant analyses and produce some report to show the results
Focus: ability to single out an interesting phenomenon (not explicitly evident from the data) and show how technology can help in communicating the findings of the analysis

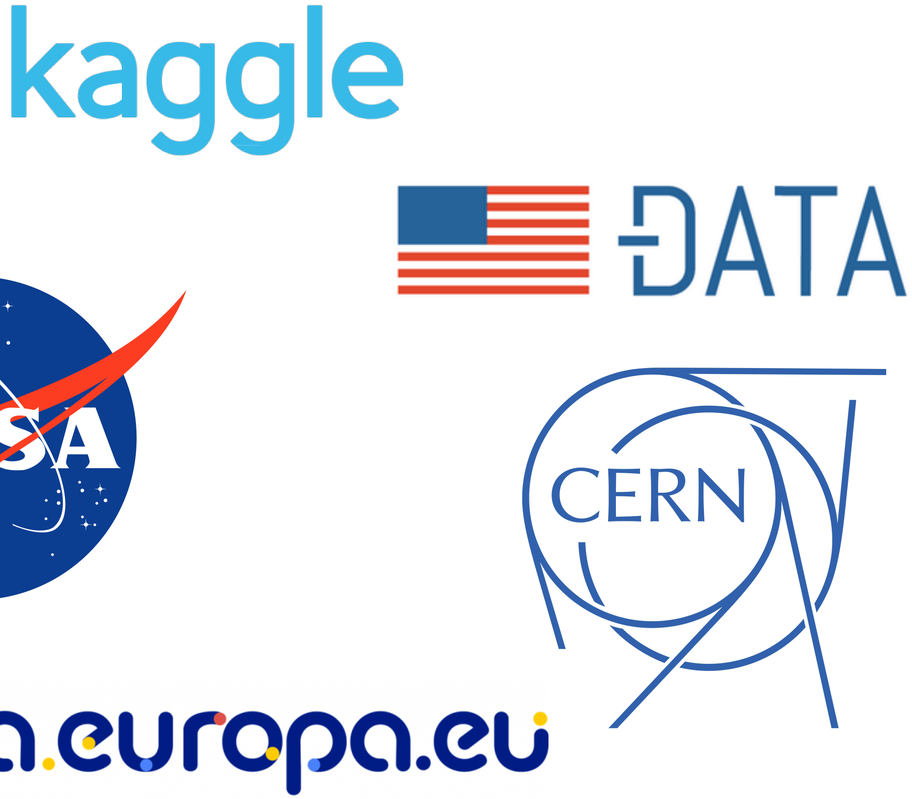
[NoSQL] Select a dataset, a NoSQL tool (graph database, document-based, etc.) and a relational DBMS, and compare the results of analyses on different systems, highlighting advantages and disadvantages of each approach in terms of efficiency
Focus: ability to compare different technologies

[NoSQL] Select a NoSQL Data Management tool (e.g., MongoDB), acquire a reasonable knowledge on it and develop a simple DB project using such a tool. If you select Neo4J (shown in class), the project should exhibit a reasonable level of complexity
Focus: ability to acquire new knowledge about a DM tool and use it in a “real case” scenario)

Where to find datasets?

This is a non-exhaustive list of sources where you can find interesting datasets:

- kaggle.com/datasets
- data.europa.eu
- opendata.cern.ch
- earthdata.nasa.gov
- dati.gov.it (Italian only)
- data.gov
- datasetsearch.research.google.com



Some advices

- Try to challenge yourselves, but do not underestimate the workload
- Don’t let your partner do all the work and don’t do all the work yourself: cooperate!
- Put your focus on the content, not on the tool (I won’t buy anything)
- Take it as an opportunity to improve your writing and your communication skills