SMBUD Project - 2023/2024

Choose a DB Technology of interest among Neo4j, MongoDB, and Elasticsearch, and a Dataset (**different** from those presented during the course), including at least 20’000 data points. Import the dataset in the chosen DBMS and perform various queries of different complexities with a good complexity variety.

N.B. Queries should represent standard and interesting analyses that one may want to perform on the datasets (e.g., considering the AirBnB dataset discussed during the lectures, one may ask themselves, “Which are the best AirBnBs with a set of desired amenities?”, etc.). Creation, update, or delete queries **do not count**!

Based on the number of group members, the requirements are as follows.

- **Individual Projects (1 Student)** must provide 10 queries.
- **Groups of 2 Students** must provide 20 queries.
- **Groups of 3 Students** must work on **two** different database technologies instead of one, finally providing 10 queries each. Two different datasets can be used for the two technologies.

I’d like to remind you that you can only work in groups of up to 3 people or individually.
No exceptions are allowed.

The delivery should include the following.

- A **.pdf** document describing the work done (more details later on), named “SMBUD Project - Name1 Surname1, Name2 Surname2, etc.”
- A folder with **high-resolution** pictures of all the pictures included in the delivery document (for readability purposes), named “Pictures”.
    - **N.B.** This doesn’t mean that the pictures in the deliverables can be unreadable!
- A **dump** of the database, or a **dump.txt** file including a download link in case the dataset is huge, within a folder named “dump”.
- A **readme.txt** file if any further specification of the delivery folder structure is made.

The delivery document is to be organised as follows

- Frontpage
    - Title, name and codice_persona of all the group members
- Introduction
    - Clearly describe the problem you’d like to face using the chosen dataset and technology and why you picked a specific DB technology based on the dataset features.
- (Optional) Data Wrangling/Data Generation
    - If any data wrangling operation is performed on the dataset, clearly describe the process.
- Dataset
    - Describe the dataset you chose, its non-relational DB implementation (showing a suited **non-relational** data schema), and its structure (for each

entity/relationship, provide its attributes, their types, a description for each attribute, etc.)

    - You cannot generate your own datasets. The dataset must be an open dataset found online and publicly accessible. You must provide the official link to the original dataset. Examples of places where to find open datasets: https://dataverse.harvard.edu/, https://www.kaggle.com/, https://data.gov/, https://data.europa.eu/en, …
    - Based on your choices or limitations, you may import a subset of the original set (still including at least 20’000 data points).
- Queries
    - For each query, provide a title, a description of what the query is supposed to do, the (text of the) query, and a screenshot/table representing the (possibly partial) outcome.
- Extra
    - You can perform some extra work (e.g., providing a Kibana dashboard representing some of your queries, some analyses in Python by connecting to the DB with their representation, an app, etc.) to get an extra 0.5 points on top of the 2 points you may get from the project. Providing extra queries does not count as “extra”. A CERTAIN LEVEL OF COMPLEXITY AND COMPLETENESS IS EXPECTED for the extra work (i.e., providing a single query representation is insufficient).

To write the final report, you can use any text tool (e.g., Word, Overleaf, etc.). I recommend using Latex through Overleaf, and downloading the template at the following link. The document **structure, understandability,** and **appearance** will also affect the evaluation. The final document is to be delivered in **.pdf** format.

Deadline: 8th January 2024 at 00:01:00

- A zip file named “SMBUD Project - Name1 Surname1, Name2 Surname2, etc.” (e.g., “SMBUD Project - Andrea Tocchetti, Stefano Brusadelli, Filippo Rezzonico”) is to be uploaded in the dedicated section on WeBeeP.
- Deliveries via e-mail (or other channels) and late deliveries will not be evaluated and will be automatically assigned a score of 0. **No exceptions** are allowed.
- If your name is not on the list of people who registered for the project, you are not allowed to get a mark, regardless of the quality and completeness of the work you delivered.
- People who register for the project and won’t deliver the required documents will be automatically assigned a score of 0.
- Only a single member of the group is supposed to deliver the project.
- Do not deliver your project at the very last moment! That “extra” minute is not there to deliver the project. Be sure to deliver the project at least one hour before the deadline. As just mentioned, I will not allow any delivery through other channels, and the system is very strict with the deadlines!