Databases and Big Data 2023-2024: Project

Professors: Blerina Sinaimeri, Valerio Rughetti and Marco Esposito
(50% of the grade + bonus point 10% of the grade)

October 4, 2023

1 Key Dates

Handed out: October 4, 2023.
Submission of ER Diagram (Deliverable 1): **October 27, 2023** (before 11:59 PM).
Submission of SQL and Python Application (Deliverable 2): **December 1, 2023** (before 11:59 PM)
Submission of CRUD on MongoDB and Python Application (Deliverable 2): **December 1, 2023** (before 11:59 PM)

2 Objective of the project

The goal of this project is to provide students with hands-on experience in designing and implementing relational databases and NoSQL databases, as well as developing a Python application to query and analyze data from real-world datasets. Through this project, students will learn to model data using Entity-Relationship (ER) diagrams, use SQL to create relational databases based on their designs, use CRUD calls to create a MongoDB database and build user-friendly Python applications for data retrieval and analysis.

3 Requirements

3.1 Deliverable 1 - ER Diagram

3.1.1 Select a Dataset

Each group must select a dataset for their project. This dataset can either be chosen from a the list of datasets below or independently sourced. In the case of an independent dataset selection, you **must** obtain approval from the instructor or teaching assistants to ensure its suitability.

1. Crimes in Boston
   https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston
2. Airlines Delay
   https://www.kaggle.com/datasets/giovamata/airlinedelaycauses
3. Formula 1 World Championships
   https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020
4. Billionaires Statistics
   https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset
5. Jobs
   https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset
6. Employees
   https://www.kaggle.com/datasets/ravindrasinghrana/employeedataset
7. IMDb
   https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset

8. Wine Reviews
   https://www.kaggle.com/datasets/zynicide/wine-reviews

9. Netflix Movies and TV Shows
   https://www.kaggle.com/datasets/shivamb/netflix-shows

10. FIFA World Cup
    https://www.kaggle.com/datasets/abecklas/fifa-world-cup

11. SpaceX Missions
    https://www.kaggle.com/datasets/spacex/spacex-missions

12. Pokémon
    https://www.kaggle.com/datasets/rounakbanik/pokemon

13. NYC Restaurant Inspections
    https://www.kaggle.com/datasets/new-york-city/nyc-inspections

14. California University Courses History
    https://www.kaggle.com/datasets/sujaykapadnis/california-university-history

15. Anime Recommendations
    https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database?select=rating.csv

16. UFC Fights
    https://www.kaggle.com/datasets/rajeevw/ufcdata

17. Homicide Reports
    https://www.kaggle.com/datasets/murderaccountability/homicide-reports

18. Popular Video Games
    https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023/

19. COVID-19 Clinical Trials
    https://www.kaggle.com/datasets/parulpandey/covid19-clinical-trials-dataset

3.1.2 Design an ER Diagram

Using the selected dataset, the group will design an ER diagram that accurately represents the dataset's structure. The ER diagram should include entities, attributes, relationships, and cardinalities.

3.1.3 Submit ER Diagram

Deliverable 1 consists of submitting the ER diagram as a digital document. Ensure that it is clear and well-labeled. Do not forget to add all the cardinality constraints. The file should be submitted by one of the members of your group on luiss.learn. Then, each group will have a short meeting (max 15 mins) with the instructor to discuss the deliverable. The meeting can be online or in person. All the members of the group should participate. The dates for these meetings will be decided when the deadline approaches.

3.2 Deliverable 2 - SQL and Python Applications

3.2.1 SQL Code

- Generate Database: Write SQL code that creates a database schema based on the ER diagram designed in Deliverable 1. The code should effectively define tables, relationships, and constraints.

- Query Code: Develop SQL queries (at least 4 for groups of ≥ 3 students, and at least 3 for groups of < 3 students) that provide **insightful** analysis of the dataset. Queries should extract interesting insight from the data. At least 2 queries should make use of aggregation. Feel free to use even more complex constructs that were not covered in the course. The complexity of the queries will also be considered in the evaluation.

3.2.2 NoSQL aggregations and grouping

- Generate Database: Write CRUD code that creates a database schema.
- Query Code: Develop some NoSQL queries (at least 4 for groups of ≥ 3 students, and at least 2 for groups of < 3 students) that provide **insightful** analysis of the dataset. Queries should extract interesting insight from the data. At least 2 queries should make use of aggregations and grouping (1 for groups of < 3 students). Feel free to use even more complex constructs that were not covered in the course. The complexity of the queries will also be considered in the evaluation.

3.2.3 Python Applications

- Load Data: Develop a Python application that loads the raw dataset into the SQL database, implementing the relational schema defined in the ER diagram.
- Load Data: Develop a Python application that loads the raw dataset into the NoSQL database.
- Query Interface: Create a simple, user-friendly interface within the Python application that allows users to execute the SQL queries selected during the analysis. While a fancy interface is not necessary, the application should be intuitive and functional and we should be able to use the application without contacting you.
- Query Interface: Create a simple, user-friendly interface within the Python application that allows users to execute the NoSQL queries selected during the analysis. While a fancy interface is not necessary, the application should be intuitive and functional and we should be able to use the application without contacting you.
- Instructions: Include a small text file with clear instructions on how to run the Python application and import the dataset into the database.

It is important that all work is completed and uploaded before the deadline. Before the deadline, the group should submit on luiss.learn a single zip folder containing all the code of the implementation as well as the presentation pdf.

3.3 Prepare for in-class presentation

The group should prepare a presentation to show the projects to their classmates and instructor. The presentation should include an overview of the project and its goals, a description of the dataset and task analysis, a demonstration of the visualization, and a discussion of the main findings. The group should also be prepared to answer questions and discuss their design and implementation decisions.

- The presentation should be approximately 8 minutes in duration (cannot exceed 15 minutes including Q&A).
- Each member of the group must speak for approximately equal portions of the presentation.
- The presentation must include a live-demo of the interactive visualizations.

4 Submission and Evaluation:

The evaluation of your project will take into account various aspects, including:

- **Correctness of ER diagram :** Your ER diagram should accurately reflect the structure of the dataset and the relationships between entities.
- **Effectiveness of SQL Queries:** The SQL queries you write should be effective in extracting valuable insights from the data. The complexity and creativity of the queries will also be considered.
- **Effectiveness of NoSQL Queries:** The NoSQL queries you write should be effective in extracting valuable insights from the data. The complexity and creativity of the queries will also be considered.

- **Functionality and Usability of Python Applications:** The Python applications should function smoothly and allow users to execute the SQL and NoSQL queries with ease. While a sophisticated interface is not required, it should be user-friendly.
- **Handling of Presentation and Questions:** Your group's presentation and the way you handle questions during the presentation will be assessed. Clear communication of your project's goals, dataset, analysis, and findings is essential for a successful presentation.

Good luck!