Advanced Databases/Databases Technologies
2024/2025 Project

This project aims to compare a relational database and a NoSQL database regarding data modeling, querying, and optimizations.

Infrastructure

Relational

MySQL: It is recommended to use a local installation of the MySQL.The simplest and recommended method is to download MySQL Installer (for Windows) and let it install and configure a specific version of MySQL Server. Other DBMS can be used but they will have no support. You can use MySQL Workbench as a frontend for MySQL.

NoSQL

Document database: MongoDB (use a local installation or a free cloud installation with Atlas - up to 500Mb)

Data

Go to Kaggle (https://www.kaggle.com/datasets)

1. Select a dataset from a field of your choice
2. The dataset must be in CSV, and it must have a minimum of 3 CSVs
3. Each CSV must have at least one column in common
4. Examples:
    a. https://www.kaggle.com/datasets/thedevastator/udemy-courses-revenue-generation-and-course-anal?select=3.1-data-sheet-udemy-courses-busines s-courses.csv
    b. https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017

Deliverables

Submit a single file in zip format containing all material produced from each phase at Moodle. The zip file should contain the code files used and a PDF with the report (maximum 6 pages, not including references)

ATTENTION: the filename has to be BDA2425_<Group>_<Phase>.zip, example **BDA2425_G001_P1.zip**

Phase 1 delivery date: **1 Dez (23h59)**

Phase 2 delivery date: **15 Dez (23h59)**

Evaluations

At each checkpoint, groups receive a formative evaluation in the form of feedback on the aspects they should improve before the delivery

Until the final evaluation, groups can always improve on the previous phase of the project even after the respective delivery. This means a group can improve on Phase 1 during Phase 2. These improvements should be detailed in the report, but very summarized during the final discussion since these should focus on the main topic

Phase 1: Data Modelling and Querying

1. Select the datasets you will be working on
2. Write the specifications for two simple data operations for relational and NoSQL databases
    a. Example: Select the customers that are VIP and have a discount higher than 10%
    b. This is a simple operation because it include one query, involving only one table/collection
3. Write the specifications for two fairly complex data operations that are able to showcase the differences between relational and NoSQL databases

    a. Example: Insert a new product called 'Potato' for a customer who made their first purchase in the 90s and spent the most in the 2000s
    b. This is a complex operation because it includes multiple queries, includes write and read operations, and includes heavy queries (sort by, group by, range queries)
4. Define the relational schema :
    a. You can draw an Entity-Relationship model
    b. You can draw a Relational Diagram
    c. You MUST write the CREATE TABLES statements
5. Build a relational database in MySQL to store your data and implement the operations designed in 2 and 3.
6. Build a NoSQL database in MongoDB to store your data and implement the operations designed in 2 and 3.

Notes

You should use python and the libraries studied in the classes to create the databases (mysql.connector/sqlalchemy, pandas, pymongo)

Your report for this phase should include:

1. Group number, student name, and student number
2. For each student, include a detailed description of their specific contribution to the project and provide the percentage of contribution by each student
3. A description of the dataset
4. Relational Schema and MongoDB collections structure

5. Discussion of additional features done in the project
6. Discussion of points not done in the project
7. Discussion of known errors

Phase 2: Indexing and Optimization

1. Rewrite the queries developed in Phase 1 in case they can be optimized
2. Apply indexes to both your databases (relational and NoSQL) to improve the performance of your complex operations implemented in Phase 1
3. Introduce changes to the relational schema to improve the performance
4. Consider alterations to the data model in NoSQL to improve the performance
5. Test the performance of your queries in your databases with prior optimization and after optimization
6. Demonstrate the impact of the options 1-4 in each query performance
7. Discuss the trade-offs (if any) between each design choice for each query

Your report for this phase should include:

1. Group number, student name, and student number
2. For each student, include a detailed description of their specific contribution to the project and provide the percentage of contribution by each student

3. Detailed descriptions of changes made regarding the goals for Phase 1
4. Provide a comparative analysis of your operations in both databases with prior optimization and after optimization
5. Discussion about the optimization and indexes used
6. Discussion of additional features done in the project
7. Discussion of points not done in the project
8. Discussion of known errors