Advanced Databases/Databases Technologies

2022/2023

Project description

This project part aims at comparing a relational database and a NoSQL database in terms of data modeling, querying, and optimizations.

Infrastructure:

- Relational:
    - SQLite:
- NoSQL:
    - Document database: MongoDB

Data:

- Go to kaggle (https://www.kaggle.com/datasets)
- Select a dataset from a field of your choice
- The dataset must be in CSV, and it must have a minimum of 3 CSVs
- Each CSV must have at least one column in common
- Example: https://www.kaggle.com/datasets/thedevastator/udemy-courses-revenue-generation-and-course-anal?select=3.1-data-sheet-udemy-courses-business-courses.csv

1. The first goal of the project is to select the dataset and the databases schemes
    a. Select the datasets
    b. Design the scheme of the databases:
        i. Several tables/collections

ii. Primary keys
iii. Secondary keys
iv. Relations between the tables
v. Data types

2. The second goal is to create the databases:
   a. Create the databases
   b. You should use python and the libraries studied in the classes (sqlite3, pandas, pymongo)

3. Create at least six queries for each database (relational and noSQL)
   a. Two simples queries, selecting data from one or two columns/fields
   b. Two complex queries, using joins and aggregates, involving at least 2 tables/collections of your database
   c. Two update/insert queries

4. Indexing and Optimization
   a. Implement optimizations and adequate indexing in your databases
   b. Test the performance of your queries in your databases with prior optimization vs after optimization
   c. Suggestions
      i. Rewrite the queries developed in part 1 and 2 in case they can be optimized.
      ii. Apply indexes to both your databases (relational and NoSQL) to improve the performance of your complex operations
      iii. Introduce changes to the relational schema to improve the performance
      iv. Consider alterations to the data model in NoSQL to improve the performance
      v. Demonstrate the impact of the options 1-4 in each query performance
      vi. Discuss the trade-offs (if any) between each design choice for each query.

Delivery:

Date: December 4th, 2022 (23:59)

Moodle

Zip file with code and report

Zip file name: BDA2223_GroupNumber.zip, example BDA2223_G01.zip

The report:

- Maximum of six pages
- Description of the dataset
- Scheme for both databases
- Discussion of point done/not done in the project
- Description of how to replicate the project: creation of the databases, and running the queries