Corso di laurea magistrale in Data Science
Facoltà di Ingegneria dell'Informazione, Informatica e Statistica, Sapienza Università di Roma

Data Management for Data Science - Homework Assignments

2023/2024

Prof. Domenico Lembo
Prof. Riccardo Rosati

Students can present homeworks during the lectures. There will be three homework assignments, which will be announced on this web page.

Assignment 1 - SQL

Choose an application domain and, using a relational DBMS, build a database. This can be done in two ways:

- (recommended) use an interesting existing dataset, i.e.:
  1. get interesting data from the Web or other sources (e.g., use the Web to look for a whole database, or data that can be easily imported into a relational DBMS) and build a relational database using such data
  2. formulate a set of SQL queries (about 10) over the relational schema
  3. execute such queries over the database and analyze the results
  4. NOTICE: all datasets are potentially fine EXCEPT MOVIE DATASETS (too many projects used movie DBs in the previous years). If, however, there will be overlapping projects (that is, projects using the same dataset) we will contact the interested groups
- create the schema and the dataset from scratch, i.e.:
  1. define the relational schema (i.e., write SQL statements to create tables defining attributes, domains, and possibly integrity constaints);
  2. insert tuples into tables (through SQL statements)
  3. formulate a set of SQL queries (about 10) over the relational schema
  4. execute such queries over the database and analyze the results

Students can use publicly available DBMSs like MySQL or PostgreSQL (see below), or other, commercial DBMSs.

The queries defined by the students should comprise all the aspects of SQL queries analyzed during the course lectures and exercises (joins, aggregations, nested queries, queries with negated subqueries). The complexity of the queries produced should be at least comparable to the specification appearing in this exercise on SQL.

Assignment 2 - SQL evaluation and optimization

Starting from the database developed in the first homework, every group has to identify at least 4 SQL queries that pose performance problems to the DBMS. The students have to show both the "slow" and the "fast" execution of the queries, where the fast version is obtained by:

- adding integrity constraints to one or more tables
- rewriting the SQL query (without changing its meaning)
- adding indices to one or more tables
- modifying the schema of the database
- adding views or new (materialized) tables derived from the existing database tables

Ideally, these queries should be picked from the queries created for the first homework; however, new queries can be considered if none of the previous queries poses performance problems to the DBMS.

Rules

GROUPS: **The homework must be done by groups of two students.**

GROUP REGISTRATION: Every group must send an email to both prof. Lembo and prof. Rosati no later than **March 27, 2024** (strict deadline), with subject: "DMDS homework group" and containing:

- last name, first name and matricola of every group member
- the DBMS chosen
- a brief description of the domain of the dataset
- the link to such a dataset

The teachers will create and maintain a list of the projects officially registered, and such a list will be accessible on Classroom (so the students can check on that list if their registration has been successful).

PRESENTATION OF 1ST AND 2ND HOMEWORK: The presentation of the work done will consist of a short (15 minutes) session in which the students will show the work done by directly interacting with the relational DBMS on their own laptop.

PRESENTATION DATES: The first homework, together with the second one, will be presented **online** during the week of **April 15-19, 2024** (the exact schedule will be published a few days before April 15).

EVALUATION: For every homework, every student will get a score ranging from -4 to +1. The final exam score will be computed as follows:

final_score = hw_1 + hw_2 + hw_3 + 30

where hw_n is the score of homework n (if final_score > 30, then the final score is 30 cum laude).

USEFUL LINKS:

- Link to download the MySQL Server and Workbench
- There are many tutorials on how to install MySQL Server and Workbench, see e.g. this one.
- Link to download the PostgreSQL DBMS and pgAdmin
- There are many tutorials on how to install PostgreSQL and pgAdmin, see e.g. this one.

Assignment 3 - NoSQL

Use a NoSQL tool (property-graph database, RDF triple store, document-based database, key-value database, column-family database) to manage and query a dataset. Ideally, the groups should use the same dataset used in the first and second homework. Examples of such systems include (but are not limited to) Neo4J, MongoDB, Redis, Cassandra (see the course material on aggregate databases and graph databases for more details).

The work must be done by the same student groups who presented the first and second homework assignments (i.e., current project groups cannot be modified). The presentation of the work done will consist of a short session (15 minutes at most) in which the students will show the work done by directly interacting with the NoSQL system on their own laptop, highlighting the differences with respect to a standard (SQL) relational database system.

Additional notes:

(1) If you have difficulties in importing the entire original database (e.g., because of limited computing power of your PC), you can use only a (significative) portion of it (in this case, you should try to import some data from all the original tables, but you can avoid to transform all the tuples of the original tables).

(2) Try to reformulate each query used in homework 1 over the new database (if you have done more than 10 queries in hw1, you can select just 10 among them); there could be queries for which the reformulation is overly complicated (and it may happen that the NoSQL system is not able to process it), or it is even not possible (depending on the way in which you design the NoSQL DB, in particular if you are working on document or key-value databases); in this case, substitute the original query with a new one, possibly with similar aims of the original query, and explain which were the difficulties in the reformulation the original query;

(3) It is not necessary to optimize the queries, that is, we are not asking for a comparison between two different versions of a query (non-optimized vs. optimized), as done for hw2.

(4) If you think that the database you have used in the first Homework is not suited for transformation into a NoSQL database, you may ask (via email) to change the database. Your request must be accompanied with valid motivations. You will then receive an email saying whether your request has been accepted or not.

The presentations of the third homework will be held during the lectures of **May 27 and May 28, 2024.** The order of presentations will be the same as the one for homeworks 1 and 2: so, the groups who presented HW 1 and 2 on April 15 will present HW 3 on May 27, and the groups who presented HW 1 and 2 on April 16 will present HW 3 on May 28 (a list with a more detailed presentation schedule will be published on Google Classroom).
