Data Management for Big Data

September 23, 2021

2020–2021, 2nd fall session
Teachers: Dario Della Monica and Andrea Brunello

Surname and name: ____

Student ID (matricola): __________ email: ___________________________

The exam is divided in 3 parts. Each part of the exam will contribute equally (one third) to determine the final score. Thus, total score of each part might be normalized so that you get at most 10 points from each part. Teachers can assign extra bonus points at their own discretion.

Be careful: use a neat handwriting. It is particularly important in this emergency situation, since the test will be scanned and emailed to the teachers.

Part I: Fundamentals of database systems

Exercise 1. Design an ER conceptual schema for the management of cinematographic data. The database has to deal with movies, actors, and directors. Each movie is characterized by its title, release date, nation (or nations if more than one) in which it has been filmed, and the set of actors that act in it, with their roles (assume that each actor may play just a single role in a movie). There cannot be two movies released in the same year with the same title. Nevertheless, movies released in different years may have the same title. For each actor, we would like to store its name and surname (that together univocally identify her/him), the birth date, the gender, and the nationality. For each director, we are interested in its name and surname (that together univocally identify her/him), the birth date, the gender, and the nationality. In addition, we would like to keep track of the movies directed by her/him. Finally, assume that a same person can be both an actor and a director.

Build an ER schema that describes the above mentioned requirements, clearly explaining any assumptions you make. In particular, for each entity, identify its attributes, candidate keys, and carefully specify the constraints associated with each relation. Also, make sure to correctly specify generalization relationships, if any.

Exercise 2. Let us consider the following relational schema about writers writing books:

Writer(Name, Surname, Birth_date, Nationality)
Book(Title, ISBN, Genre)
Wrote(Name, Surname, ISBN)

Let us assume each writer to be univocally identified by its name and surname, and characterized by a birth date and a nationality. Each book has a title and a genre, and is univocally identified by its ISBN. A book may be written by one or more writers, and a writer may write more than one book.

Define preliminary primary keys, other candidate keys (if any), and foreign keys (if any). Then, formulate an SQL query to compute the following data (exploiting aggregate functions only if they are strictly necessary):

> *The writers of German nationality that have participated in the writing of only science fiction books (and of at least two of them).*

Part II: Advanced database models, languages, and systems

Instructions for multiple-choice questions.

- A right answer is worth +1.
- A wrong answer is worth -0.33.
- It is possible to give a short explanations for multiple-choice questions. It should be **very** short (no more than 2 lines) and should be written in a neat handwriting. They are optional and used **at teacher's discretion**, mainly to increase the score of a wrong answer; seldom, to lower the score of a right one.

Instructions for open questions.

- The score assigned to an open question can range from 0 to 3 points.
- No negative score is assigned to open questions.
- Be careful: use a neat handwriting.
- Try to be succinct also for open questions.

1. Let t_S = "time for one seek", t_T = "time for one-block transfer", and h be the height of a secondary index (a tree) over attribute A of relation R. Which is the estimated cost for accessing all tuples where $A > X$? (Assume that A is a key.) Briefly argument your answer.

2. Select the correct statement.

□ The Catalog stores results of queries that are executed frequently

□ The Catalog stores indices to be used to optimize query executions

□ The Catalog stores statistical information useful for cost estimations

Short explanation (optional):

3. Briefly describe the cost model we adopted in the context of classic (centralized) DBMS and the one for distributed DBMS, emphasizing the differences between them?

4. Consider the 2 transactions T_1 (over operations $R_1(y), W_1(y), R_1(x), W_1(x)$) and T_2 (over operations $R_2(x), W_2(y), W_2(x)$) formalized through the 2 following partial orders, respectively:

   $T_1 = \{W_1(x) \prec R_1(x), R_1(x) \prec R_1(y), W_1(x) \prec W_1(y), W_1(y) \prec R_1(y)\}$
   $T_2 = \{W_2(y) \prec R_2(x), R_2(x) \prec W_2(x)\}$.

   Mark the right statement among the following ones:

   □ There are exactly 2 serial histories over $\{T_1, T_2\}$

   □ Every history over $\{T_1, T_2\}$ that is serializable is also serial

   □ There is a history over $\{T_1, T_2\}$ that is serializable but not serial

   Short explanation (optional):

   Hint: recall that serial histories are always concurrent transaction executions, that is, they are formalized through linear orders.

Part III: Data analysis and big data

1. With respect to the Data Warehouse context, briefly describe the ETL process phases.

2. Describe the operations *slice* and *dice* for OLAP cube, also emphasising how they relate each other.

3. Briefly describe the text indexing and its three main steps.

4. Briefly characterize the most suitable domains for the usage of graph databases and mention a couple of such domains.

5. Briefly characterize *stationary* and *non-stationary* time series, emphasizing their differences (also use examples if it helps you).