Data Management for Big Data

July 13, 2021

2020–2021, 2nd summer 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 a hospital. The database has to keep track of information regarding people, which can be patients or doctors. Each person is identified by its VAT code, and is characterized by a name, a surname, a gender, and a birth date. For simplicity, let us assume that a doctor might also be a patient himself.

Patients receives diagnoses (at most one per day). A diagnosis is made by a doctor and is characterized by a description of the symptoms exhibited by the patient. For each diagnosis, a treatment may be prescribed to the patient, consisting of one or more drugs with the corresponding quantities. A drug is univocally identified by its code, and is characterized by a name and one or more side-effects.

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 singers participating to musicals:

Singer(Art_name, Name, Surname, Nationality)
Musical(Name, Year)
Participates_to(Singer, Musical, Year, Role)

Let us assume assume each singer to be univocally identified by its art name, and characterized by a name, a surname, and a nationality. Each musical has a name and is first released in a given year. Note that different musicals may have the same name (this is the case, for instance, of remakes); nevertheless, we assume that, given a year, no homonymies may happen. We further assume that a singer may sing in more than one musical (but can just play a single role per musical), and that a musical has, in general, more than one singer.

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 singers that participated in at least two musicals released in 2020, but not in any musical released in 2019.

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 R_1 and R_2 be relations and τ be a predicate (e.g., "$age = 20$") over attributes of R_1. Consider the following equivalent relational algebra expressions:

   (i) $\sigma_\tau(R_1) \bowtie R_2$,

   (ii) $\sigma_\tau(R_1 \bowtie R_2)$.

   Which is the correct statement among the following:

   ☐ "pushing" the selection inside a join is always more efficient, i.e., (i) is more efficient than (ii)

   ☐ "pushing" the selection inside a join is always less efficient, i.e., (ii) is more efficient than (i)

   ☐ there is no general rule, i.e., the relative efficiency of (i) and (ii) depends on concrete instances of τ, R_1, and R_2

   Short explanation (optional): _____

   _____

2. What is a cost model and why is it necessary?

   _____

   _____

3. Why is semijoin important, especially in the context of distributed DB systems?

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 is a history over $\{T_1, T_2\}$ that is neither serializable nor serial

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

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

   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. Name and describe the *3 Vs of Big Data*.
3. Briefly describe *text indexing* and name its main phases.
4. What are the advantages in using Hadoop, with respect to solutions based on HPC and grid computing?
5. Briefly describe the concept of *stationary time series*, and provide an example of such a time series.