Data Management for Big Data

February 15, 2021

2019–2020, 2nd winter session
Teachers: Angelo Montanari, Dario Della Monica, Andrea Brunello

Surname and name: _____

Student ID (matricola): _____ email: _____

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:

Let us consider the following relational schema about films and film festivals:
$FILM(CodF, Title, FilmMaker, ProductionYear)$;
$FESTIVAL(FestivalName, Place, Year)$;
$PARTICIPATION(Film, Festival, Year, Position)$.

Let us assume each film to be univocally identified by a code, and characterized by a title, a film maker (for the sake of simplicity, let us assume that each film has one film maker only), and a production year. We assume that there may exist distinct films with the same title (with the same film maker or with a different one), but they cannot be produced the same year.

Let us assume each film festival to be characterized by a name, that distinguishes it (for instance, Silent Film Festival), and by a place where it takes place. Moreover, let us assume that each film festival is organized once per year (film festival edition), always in the same place, and that different film festivals can be organized in the same place.

Let us assume each edition of a film festival to produce a ranking of presented films (ex aequo are excluded). We assume also that each film maker may present more than one film at the same edition of a festival. Finally, let us assume that

the same film may be presented at more than one festival, but not at different editions of the same festival.

Define preliminarily 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 film maker (the film makers if more than one) that presented the maximum number of films at the same edition of some festival.

Exercise 2:

Sinthesize the ER conceptual schema of a database for the management of a book shop.

- The book shop has a catalogue of books. For each book, the following pieces of bibliographical information are recorded: the ISBN code, that univocally identifies the book, the title, the authors, the publisher, and the year of publication.
- Not all the books in the catalogue are available in the book shop. Of the available ones, we record the number of copies and the selling price. Moreover, we keep track of the location of each available copy: closet, shelf, and position in the shelf (from left to right).
- Some of the missing books have been ordered. Each order may include various books, it is univocally identified by a code, and it is characterized by the date of issue, the supplier to which the order has been sent, and the number of ordered copies for each book included in it.
- Each supplier is univocally identified by his/her fiscal code, and it is characterized by a name, a postal address, an email, and a telephone number.

Build an ER schema that describes the above requirements, clearly explaining any assumption you made. In particular, for each entity, identify its possible keys, and carefully specify the constraints associated with each relation.

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; make suitable assumptions for possibly missing pieces of information that you need for your answer.) Briefly argument your answer.

2. Which of the following statements applies to the context of Distributed DB Systems?

   ☐ Information is distributed among the nodes and data replication is allowed for efficiency purposes

   ☐ Some of the nodes are devoted to information storage and other ones to query management

   ☐ Information is distributed among the nodes and no data replications is allowed to avoid data inconsistency issues

Short explanation (optional): _____

3. Let R be a relation whose primary key is the attribute *key*, M be a partition of attributes of R, and S be a relation such that there is a link L with *owner(L)* $= R$ and *member(L)* $= S$. In other word, one of the attributes of S is a foreign key referring to attribute *key* of relation R.

   Which is the vertical fragmentation over relation R induced by M?

   ☐ $\{R_i \mid R_i = \sigma_{m_i}(R), m_i \in M\}$

   ☐ $\{R_i \mid R_i = \Pi_{m_i \cup \{key\}}(R), m_i \in M\}$

   ☐ $\{R_i \mid R_i = R \ltimes S_{m_i}, m_i \in M\}$

   Short explanation (optional): _____

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

   Is there a history over $\{T_1, T_2\}$ that is serializable but not serial? If yes, write down one such history. If not, write down a history that is both serializable and serial.

   Is there a history over $\{T_1, T_2\}$ that is serial but not serializable? If yes, write down one such history. If not, write down a history that is neither serializable nor serial.

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

- With respect to the Data Warehouse context, briefly describe the ETL process phases.
- Name the *3 Vs of Big Data*.
- Briefly describe what is a *graph database*, highlighting the difference between a *native* and a *non native* solution. Then, name a possible context of usage.
- Briefly describe *text indexing* and name its main phases.
- What are the advantages in using Hadoop, with respect to solutions based on HPC and grid computing?