Data Management for Big Data

September 16, 2019

2018–2019, 2nd fall session
Teachers: Angelo Montanari, Dario Della Monica, and Paolo Gallo

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.

Part I: Fundamentals of database systems

Exercise 1:

Let us consider the following relational schema about actors and films:
$FILM(FilmCode, Title, Filmmaker, Year)$;
$ACTOR(ActorCode, Surname, Name, Sex, BirthDate, Nationality)$;
$INTERPREATATION(Film, Actor, Role)$
Let every film be univocally identified by a code and characterized by a title, a filmmaker, and the year when it has been released. For the sake of simplicity, let as assume each film to be directed by a single filmmaker and each filmmaker to be univocally identified by his/her surname. Let us consider the possibility that distinct films have the same title (this is the case, for instance, with remakes), but exclude the possibility for two films with the same title to be released the same year. Let every actor be univocally identified by a code and characterized by a name, a surname, a sex, a birth date, and a nationality. Let us assume that more than one actor may act in a film and that the same actor may act in more than one film. For the sake of simplicity, let us assume that, in a given film, each actor may play one role only.
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 actors that acted in exactly two films (not necessarily the same) directed by the filmmaker Ken Loach.

Exercise 2:

Let us sinthesize the ER conceptual schema of a database to be used by a municipality to manage the artistic events it organizes on the basis of the following set of requirements.

- Each artistic event is univocally identified by a code, and it is characterized by a name and a date. It consists of one or more shows.
- Each show is identified by a number, which univocally identifies it within the artistic event it belongs to (the first show of event E27, the second show of event E27, and so on), and it is characterized by the hour at which it starts and the duration.
- A show features one or more artists (an artist can make at most one exibition within a given show an he/she receives a payment for that). The same artists can participate in more than one show of a given artistic event. Every artist is univocally identified by his/her SIAE code, and he/she is characterized by a stage name. For each artist, there exists another one who can replace him/her in case of unavailability. An artist can be indicated as a possible replacement for more than one other artist.
- Shows are hosted in suitable locations. Each day, a location can host at most 3 different shows, that may belong to same artistic event or to different ones.

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.
- Short explanations 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.
- You can use additional sheets 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.)

   ☐ $h * (t_T + t_S) + t_S$

   ☐ $h * (t_T + t_S) + t_T$

   ☐ $h * (t_T + t_S) + t_S + t_T$

   Short explanation (optional): _____

   _____

   Hint: recall that
   - *a primary index is defined over the attribute(s) used to physically order the file in the filesystem;*
   - *a secondary index is defined over any (subset) of the other attributes.*

2. If a relation R is fragmented, according to primary horizontal fragmentation, into $\{R_1, \ldots, R_n\}$, then we have:

   ☐ $R = R_1 \bigcup R_2 \bigcup \ldots \bigcup R_n$

   ☐ $R = R_1 \bowtie R_2 \bowtie \ldots \bowtie R_n$

   ☐ $R = R_1 \ltimes R_2 \ltimes \ldots \ltimes R_n$

   Short explanation (optional): _____

   _____

3. Select the statement that better describes what a Query Execution Plan (QEP) is.

   ☐ A QEP establishes how different nodes of a distributed DBMS must cooperate to store data efficiently

   ☐ A QEP describes the order of execution of a set of queries to minimize the global execution time

   ☐ A QEP is a decorated tree-like structure obtained from a relational algebra expression

Short explanation (optional): _____

4. Which sentence matches better the notion of allocation in the context of distributed DB systems?

   ☐ Allocation is mainly about how far to locate different nodes of a distributed DB system

   ☐ Allocation is mainly about whether or not to replicate relations and fragments

   ☐ Allocation is mainly about the quantity of memory to allocate for each unit of data in each node of a distributed DB system

   Short explanation (optional): _____

5. Explain what the ACID properties are. Be concise but clear and exhaustive.

6. Consider the 2 transactions T_1 (over operations $R_1(y), R_1(x), W_1(x)$) and T_2 (over operations $W_2(x), W_2(y)$) formalized through the 2 following partial orders, respectively:
   $$T_1 = \{R_1(x) \prec W_1(x), R_1(y) \prec W_1(x)\}$$
   $$T_2 = \{W_2(x) \prec W_2(y)\}.$$
   Is there a history over $\{T_1, T_2\}$ that is neither serializable nor 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 both serializable and serial.

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

7. Explain the difference between checking that an XML document is well-formed and validating it.

Hint: which is the difference between an XML parser and an XML validator?

Part III: Data analysis and big data

1. Explain why the analyst job in the data warehouse environment is easier than using OLTP solutions.

2. Briefly introduce HADOOP.

3. List the four main families of NoSQL databases.

4. Briefly introduce the decomposition of time series using the Additive Model.

5. Consider a time series from a single sensor which is transmitting a pressure value every minute. Using a document database you can store a single JSON document for each reading:

```
{
  pressID: "I8008",
  pressure: 1024,
  ts: ISODate("2018-11-10T22:56:00.00-0500")
 }
```

alternatively, you can store a JSON document containing a nested object ready to store one hour of observations and update it as soon as a new measure comes (every minute):

```
  {
    pressID: "I8008",
    pressure: {0:1003, 1:1023, 2:1012, ..., 59:1002},
    ts: ISODate("2018-11-10T22:00:00.00-0500")
 }
```

For each of the two models above write down:

- How many document writes will typically occur in one hour?
- How many documents updates will typically occur in one hour?
- Suppose that each document will include a 100 bytes long field for indexing purposes, what will be the space used only by those fields for storing an entire day of readings?

6. Briefly describe main differences between text mining and retrieval.

Hint: Retrieval looks for an exact ...

7. Explain why dbRef(s) (also named references) across documents in MongoDB differ from foreign keys used by relational databases.

Hint: Please keep in mind that some properties are guaranteed by default by relational databases.