Data Management for Big Data

June 28, 2021

2020–2021, 1st 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 brewery.

The brewery produces several beers, each identified by its name. Also, for each beer, we would like to keep track of its kind (e.g., IPA, Stout, Lager, ...), alcohol content, and price-per-liter.

A beer needs several ingredients (each identified by a code and described by a name) for its production, that are stored in an inventory. For each ingredient it is necessary to keep track of its currently stored quantity.

The brewery has a set of customers, which may be private citizens or companies. Each customer is identified by its VAT number, and described by a name, an address, one or more phone numbers, and possibly an email address. For companies, the database has also to keep track of their business area.

A customer may place multiple orders to the brewery, and an order may be composed of several beers. For each order, univocally identified by a number, it is necessary to keep track of the date of receipt, the shipping date, the quantity (measured in liters) of each ordered beer, and possible discount rate.

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

Exercise 2 Let us consider the following relational schema about teacher and courses:

> *TEACHER(VAT, Name, Surname, Salary)*
> *COURSE(Code, Name, Description, Hours, Prerequisite)*
> *TEACHES(Teacher, Course, Year)*

Let us assume assume each teacher to be univocally identified by its VAT, and characterized by a name, surname, and salary. Each course is univocally identified by its code, and is characterized by its name, description, and duration (in hours). A course may have at most one prerequisite. We assume that a teacher can teach in more than one course, and that a course may be taught by more than one teacher. Teacher assignations may be different over the years.

Define the 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 teacher(s) having the highest salary among those that have taught in course* MAT01.

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. Why is it useful to keep in the catalog information about the number of different values occurring in a column A of a relation R?

2. Distributed DB Systems are a necessity for several reasons. They come with benefits and drawbacks. Name three such benefits.

3. Which sentence match better the notion of reliability in the context of Distributed DB Systems?

   □ Reliability is the capability of being tolerant to node failures and concurrent accesses

   □ Reliability concerns the absence of bugs in the software that controls the DBMS

   □ Reliability means that communications over the network will never fail

   Short explanation (optional):

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

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

   Mark the right statement among the following ones:

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

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

   □ There is a history over $\{T_1, T_2\}$ that is neither serializable nor 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. Explain the main differences between OLTP and OLAP

2. Describe the differences between *descriptive*, *predictive* and *prescriptive analytics*, providing also some examples.

3. In the context of text indexing in text analytics, briefly describe *stopword removal* and explain why it is needed.

4. Name a benefit of adopting a column based DB solution.

5. Which of the time series in Figure 1 would you decompose using an *additive* model? And which using a *multiplicative* model? Why?

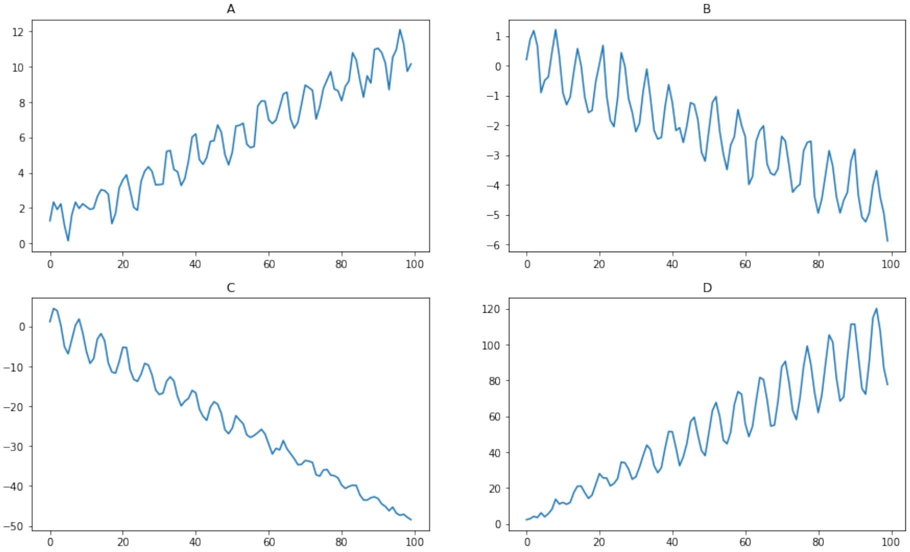

Figure 1: Time series