Institutionen för datavetenskap
Linköpings universitet

EXAM
Databases for Bioinformatics
TDDD74

May 27, 2019
14.00 – 18.00

Grades

You can get max 30 points. To pass the exam, grade 3, you need 7.5 points in both the practical part (questions 1–3) and the theoretical part (questions 4–8) of the exam. For grade 4 and 5, you need 21 and 27 points, respectively.

Questions

Olaf Hartig will visit the room at 15.00 and 16.30.

Instructions

- Write clearly.
- Use a separate page for every question.
- Answer in English.
- Give relevant and motivated answers only to the questions asked.
- State the assumptions you make besides those in the questions. None of these additional assumptions should change the spirit of the exercises.

Good luck!

Practical part (15 points)

Question 1. Data modeling with an EER diagram (5 p):

We want to create a database with information about figure skating events, skaters, and spectators.

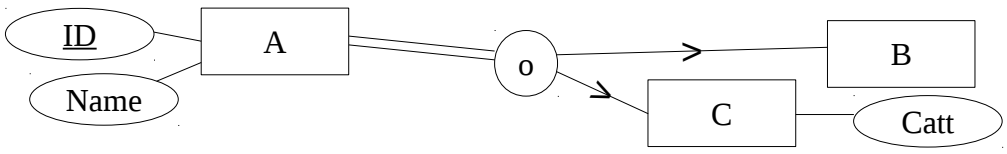
- For the purpose of our database, each figure skating event has a unique event number. Furthermore, such events have a date, a start time, and an end time.
- Skaters and spectators are persons. Every person is identified by a social insurance number (SIN). Moreover, every person has a name and a birth date; the birth date is composed of a year, a month, and a day.
- Some persons are skaters who may perform in figure skating events.
- While not every skater performs in figure skating events, those who do, may perform in more than one of these events. On the other hand, every figure skating event must have one or more skaters performing in it.
- Every person (including skaters) may attend figure skating events as a spectator. However, not every person has to do so, and there may be events without any spectator. Of course, most events are attended by multiple spectators.
- For every figure skating event that a spectator attends, we want to record a ticket number of the ticket that the spectator used for entering the event.

Please draw an EER diagram that captures the aforementioned information (including cardinality constraints and participation constraints for participation of entities in relationships, as well as totalness constraints and disjointness constraints for specializations). Use the *notation as introduced in class*. Clearly write down your choices and assumptions in case you find that something in the information above is not clear.

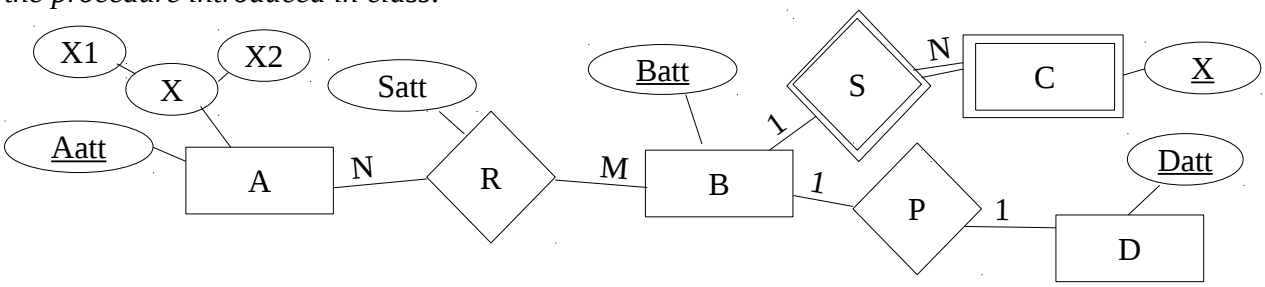
Question 2. EER diagram and relational schema (2 + 3 = 5 p):

For both of the following questions, your answer should be given in the form of a diagram that shows the relation schemas, including primary keys and foreign keys.

(a) Recall that there exist different approaches to translate specializations of entity types (i.e., super-classes with their sub-classes). Apply ***two*** possible approaches (from the approaches discussed in class) to the following example of such a specialization. That is, create two separate relational database schemas such that each of them illustrates the application of one of the approaches.



(b) Translate the following EER diagram into an equivalent relational database schema, *by using the procedure introduced in class*.



Question 3. SQL (1 + 2 + 2 = 5 p):

Consider the following database schema

Country(*Name*, *Code*, *Capital*, *Area*, *Population*)

Organization(*Name*, *Abbreviation*, *Established*)

IsMember(*Organization*, *Country*, *Joined*)

The attribute *Organization* in the table *IsMember* is a foreign key reference to *Abbreviation* in the table *Organization*. The attribute *Country* in table *IsMember* is a foreign key reference to *Code* in the table *Country*. Examples of the tuples for the above relational schema are as follows:

Country(Sweden, SWE, Stockholm, 449964, 9514000)

Organization(European Union, EU, 1952)

IsMember(EU, SWE, 1995-01-01)

Provide SQL queries to answer the following questions.

(a) List the country names of all the European Union (Abbreviation: 'EU') members.

(b) List the names of all organizations that do *not* have any member.

(c) For every organization, return the organization's name and the sum of the population size of all its member countries. (Do not assume that organization names are unique.)

Theoretical part (15 points)

Question 4. Normalization (1 + 1 + 1 = 3 p):

Consider a relation schema *R*(*A*, *B*, *C*, *D*) for which the following functional dependencies exist:

FD1: {*A*,*C*} → {B} FD2: {B} → {*D*} FD3: {B} → {*A*}

(a) Assume a relation state of *R* that contains the tuples (a1,b2,c6,d1) and (a2,b2,c6,d4). Is such a state a valid state of *R* (taking into account the FDs)? Explain your answer briefly (just writing "yes" or "no" without any further explanation does not earn you any points).

(b) What is attribute closure X^+ of the set $X = \{B\}$ w.r.t. the aforementioned three FDs? Provide only the answer to the question; that is, write only the resulting set X^+ without any explanation.

(c) Show that *R* is not in Boyce-Codd normal form (BCNF). There is **no** need to normalize *R*!

Question 5. Data structures (1 + 2 = 3 p):

(a) Recall the notion of *un*spanned allocation of records to file blocks. What problem does it cause?

(b) Assume we have a *heap file* with 1,000,000 records, a block size of 40,000 bytes, and unspanned allocation. Each record has a size of 400 bytes. The records contain only one field, X, which is a key field. For each of the following *four* points, provide only the numbers that are asked for; that is, *do **not** write any explanation/justification!*

Calculate: **i)** the blocking factor of the file and **ii)** the overall number of blocks that the file has.

Moreover, assume we want to find a record with a given value for X. How many block accesses are needed **iii)** in the best case and **iv)** in the worst case? (do not assume the existence of any index)

Recall that $\log_2(2^x) = x$. That is, $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(4) = 2$, $\log_2(8) = 3$, $\log_2(16) = 4$, $\log_2(32) = 5$, $\log_2(64) = 6$, $\log_2(128) = 7$, $\log_2(256) = 8$, $\log_2(512) = 9$, $\log_2(1024) = 10$, $\log_2(2048) = 11$, $\log_2(4096) = 12$, $\log_2(8192) = 13$, $\log_2(16384) = 14$, etc.

Question 6. Transactions and concurrency control (1 + 1 + 1 = 3 p):

(a) For each of the following four pairs of operations, indicate whether the pair conflicts. (Points will be deduced for wrong answers!)

pair 1: $r_7(A)$, $w_7(A)$

pair 2: $w_9(B)$, $w_5(B)$

pair 3: $w_4(E)$, $w_8(D)$

pair 4: $r_6(G)$, $w_{10}(G)$

(b) Consider the following schedule *S*. Is it *serializable*? Justify your claim.

S : $b_1, r_1(X), b_2, r_2(Y), w_1(X), b_3, w_2(Y), e_2, r_1(Y), r_3(X), e_3, w_1(Y), e_1$

(Notice that this task is independent from the pairs of operations in the previous task)

(c) Consider again the same schedule *S*. Is this schedule *serial*? Justify your claim.

Question 7. Database recovery (1 + 2 + 1 = 4 p):

(a) In the case of the deferred update strategy, one of the following is true: there is either no need to undo changes of non-committed transactions or no need to redo changes of committed transactions. *Which* one is it (no need to undo or no need to redo), and *why*?

(b) Given the following log, apply each of the two recovery algorithms for the two immediate update strategies described in the course. In each of the two cases, list the operations that are performed during recovery in the order in which they are performed. For each operation in these two lists, indicate explicitly which value is written by the operation; you can do this by specifying the (new) log record resulting from the operation.

Start-transaction T2
Write-item T2, B, 3, 4
Start-transaction T3
Write-item T3, A, 7, 8
Checkpoint
Write-item T3, A, 8, 1
Commit T2
Checkpoint
Write-item T3, A, 1, 5
Start-transaction T4
Write-item T4, B, 4, 5
Write-item T4, B, 5, 10
Commit T3
Checkpoint
Start-transaction T1
Write-item T1, C, 8, 9
Commit T4
* system crash *

(c) Something is wrong with the following log (on the next page). What is it?
(Note that this question and log is separate from the previous one)

Start-transaction T1
Write-item T1, A, 4, 62
Start-transaction T2
Write-item T2, B, 8, 91
Checkpoint
Write-item T1, C, 91, 1
Commit T1
Checkpoint
Write-item T2, A, 1, 5
Commit T2

Question 8. Information Retrieval (1 + 1 = 2 p):

Assume that we use the vector model for information retrieval.

(a) Explain tf and idf in the vector model.

(b) Suppose that we are only interested in the words ‘gene’, ‘enzyme’, ‘protein’ and ‘signal’, and that we have two documents in our document base such that:

- Document 1 contains 'enzyme' 8 times, 'gene' 5 times, 'protein' 0 times and 'signal' 10 times.
- Document 2 contains 'enzyme' 1 times, 'gene' 0 times, 'protein' 6 times and 'signal' 0 time.

Give the document representations for Document 1 and Document 2 according to the tf-idf model.