Institutionen för datavetenskap
Linköpings universitet

# EXAM
# Database Technology
# TDDD37 – TDDD46

January 15, 2019
8.00 – 12.00

## 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 9.00 and at 10.30.

## 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 the following information about writers and readers of books.

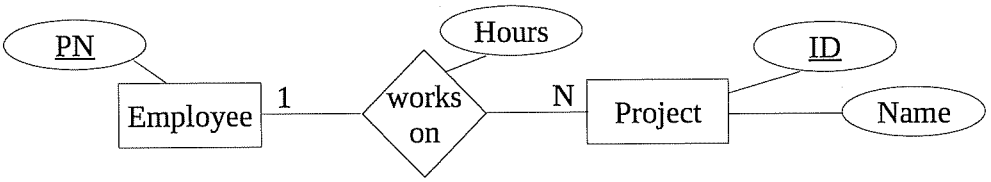
- Each book has an ISBN, which is unique. Furthermore, books have a title.
- A person is identified by a social insurance number (SIN), and has a name and a birth date; the birth date is composed of a year, a month, and a day.
- Some persons are writers, who may write books and may have multiple favorite topics, where different writers may have the same favorite topic(s). Just to avoid any confusion, this point about the topics has to do only with the authors, not with books and not with readers.
- While not every writer writes books, those who do, may write more than one book. On the other hand, every book must have one or more writers writing it.
- Some persons are readers; that is, they read books (at least one!). Writers may also be readers.
- While books typically have multiple readers, there may be books that nobody reads.

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.

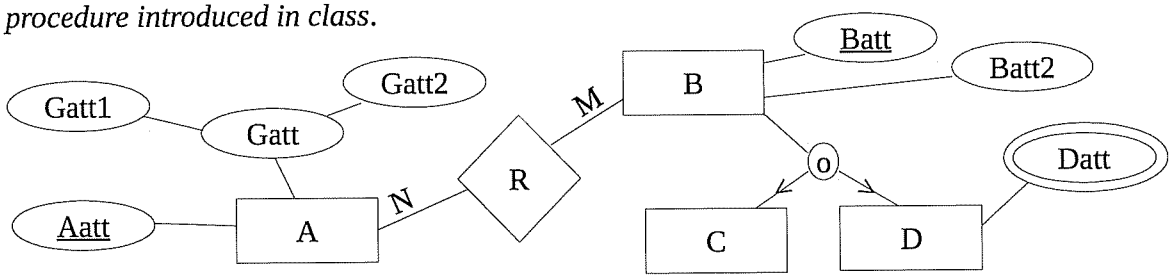
### Question 2. EER diagram and relational schema (2 + 3 = 5 p):

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 we have two different approaches to translate a 1:N relationship type during the translation of an ER diagram to a relational database schema. For the following example of such a 1:N relationship type, apply *both* approaches. That is, create two separate relational database schemas such that each of them illustrates the application of one of the two 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 + 1 + 1 = 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 statements to answer the following questions.

**(a)** List the names of all organizations that the country with code 'SWE' is a member of.

**(b)** 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.)

**(c)** Provide an SQL query whose result consist of *a single column* that contains the names of all organizations and of all countries.

**(d)** For the country with code 'SWE', increase the population size by 1000.

## Theoretical part (15 points)

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

Consider a relation schema $R(A, B, C, D)$ with the following four functional dependencies:

FD1: $\{A\} \rightarrow \{C\}$
FD2: $\{B\} \rightarrow \{D\}$
FD3: $\{C\} \rightarrow \{A\}$
FD4: $\{D\} \rightarrow \{B\}$

**(a)** Assume a relation state of $R$ that contains the two tuples (1,2,6,1) and (2,2,6,4). 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 $X = \{A\}$ w.r.t. the aforementioned four 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).

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

Assume we have a sorted file with 100,000 records, a block size of 40,000 bytes, and unspanned allocation. Each record has a size of 40 bytes. The records have two fields, *X* and *Y*, where *X* is a key field (and *Y* is not). The file is sorted on *X*. For each of the following points, provide only the numbers that are asked for; that is, *do not write any explanation/justification.*

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

**(b)** Calculate the average number of block accesses needed to find a record **i)** with a given value for *X*, and **ii)** with a given value for *Y* (do not assume the existence of any index).

**(c)** To speed up the retrieval we may use an index. Assume we want to speed up finding a record with a given value for *X*. Name **i)** the type of *single-level* index that we can use in this case and **ii)** the concrete number of index records that this index would have for our file.

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 + 1 = 4 p):**

**(a)** For each of the following four pairs of operations, indicate whether the pair conflicts. Hence, for each pair, say "yes" or "no". (Points will be deduced for wrong answers!)

pair 1: $r_2(X)$, $w_2(X)$

pair 2: $w_4(Y)$, $w_2(Y)$

pair 3: $w_3(Z)$, $w_2(X)$

pair 4: $r_3(Z)$, $w_2(Z)$

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

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

**(c)** Is this schedule *serial*? Justify your claim.

**(d)** Specify the two-phase locking (2PL) protocol; what does a transaction have to do to follow the protocol? (Note that this is a general question; it is independent of the aforementioned schedule.)

**Question 7. Database recovery (1 + 2 = 3 p):**

**(a)** Two things are guaranteed to have happened when a transaction reaches its *commit point* (independent of whether we use the deferred update strategy or any of the two immediate update strategies). What are these two things? (Note that this question is not about checkpoints.)

**(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 *

**Question 8. Query Processing (1 + 1 = 2 p):**

**(a)** Recall that, during query processing, queries are represented as logical plans with logical operators and, thereafter, as physical plans with physical operators. What is the major difference of physical operators in contrast to logical operators?

**(b)** Assume two relations, $R$ and $S$. Let $pr$ and $ps$ be the number of disk pages occupied by the file for relation $R$ and for $S$, respectively. What is the I/O cost (in terms of page reads) of using the nested loops join algorithm (NLJ) to join $R$ and $S$ if the outer loop of the NLJ iterates over $R$?