Försättsblad till skriftlig tentamen vid Linköpings universitet

Datum för tentamen	2019-10-29
Sal (1)	G37(16)
Tid	8-12
Utb. kod	TDDD74
Modul	TEN1
Utb. kodnamn/benämning **Modulnamn/benämning**	Databaser för bioinformatik Skriftlig tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	8
Jour/Kursansvarig Ange vem som besöker salen	Patrick Lambrix
Telefon under skrivtiden	013-28 26 05
Besöker salen ca klockan	9.00 and 10.30
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Annelie Almquist, 013-28 29 34 annelie.almquist@liu.se
Tillåtna hjälpmedel	Dictionary
Övrigt	
Antal exemplar i påsen	

Institutionen för datavetenskap
Linköpings universitet

RETAKE EXAM
Databases for Bioinformatics
TDDD74

October 29, 2019
8.00 – 12.00

Grades

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

Questions

Patrick Lambrix 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!

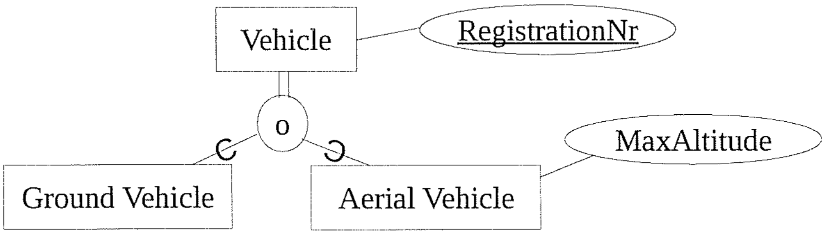
Practical part (15 points)

Question 1. Data modeling with an EER diagram (1 + 4 = 5 p):

(a) Consider the following EER diagram. Answer the following two questions with respect to the constraints captured by the diagram. Write only "yes" or "no" for each of the two questions.

i) There can be vehicles that are both a ground vehicle and an aerial vehicle. yes/no?

ii) There can be vehicles that are neither a ground vehicle nor an aerial vehicle. yes/no?



(b) We want to create a database with the following information about music concerts.

- Every concert has a unique number, a date, a start time, and an end time.
- Every concert takes place in a location, where each location has a name and an address; the address is composed of street name, number, postal code, and city. While different locations may have the same name, the address of each location is unique.
- Concert goers attend concerts (at least one!). A concert goer is identified by a social insurance number (SIN) and has a first name and a last name. For every concert that a concert goer attends we want to record the ticket number of the concert goer's ticket.
- There may be concerts that nobody attends. Of course, most concerts are attended by multiple concert goers.
- Concert goers may have one or more favorite locations, but they don't have to. As with concerts, some locations may not be favored by any concert goer.

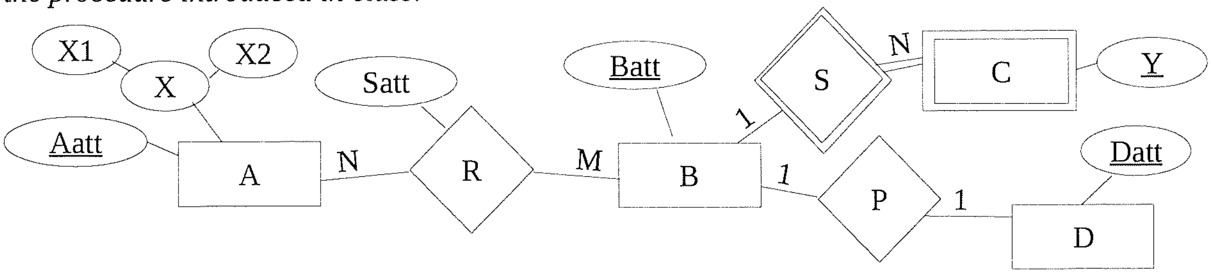
Draw an EER diagram that captures the aforementioned information (including cardinality constraints and participation constraints). Use the *notation as introduced in class*. Clearly write down your choices and assumptions in case you find that something in the information above is not clear.

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

For both of the following questions, your answers should be given in the form of diagrams that show the relation schemas, including primary keys and foreign keys.

(a) Recall that there exist different approaches to translate specializations of entity types (i.e., superclasses with their sub-classes). Apply *two* possible approaches (from the approaches discussed in class) to the EER diagram in Question 1.a) given above. 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 + 1 + 1 + 2 = 5 p):

Consider the following database schema

Country(*Name*, Code, *Capital*, *Area*, *Population*)
Organization(*Name*, *Abbreviation*, *Established*)
IsMember(*Organization*, *Country*, *YearJoined*)

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)

(a) Describe what exactly the following SQL statement aims to achieve.

```
SELECT Name, Country
FROM Organization LEFT OUTER JOIN IsMember ON Abbreviation=Organization
WHERE YearJoined > 1990;
```

(b) Describe what exactly the following SQL statement aims to achieve.

```
UPDATE Country
SET Population = Population + ( SELECT Population FROM Country
                                WHERE Code="GDR" )
WHERE Code = "FRG";
```

(c) Provide an SQL query to list the names of all countries that have joined some organization in 1990 (it does not matter which organization it is).

(d) Provide an SQL query to list the country names of all the European Union members that joined the European Union in 1990 or later. Hint: Such countries are members of the EU *and* have not been an EU member before 1990.

Theoretical part (15 points)

Question 4. Normalization (1 + 1 + 2 = 4 p):

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

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

(a) What is attribute closure X^+ of the set $X = \{B\}$ with respect to these three FDs? Provide only the answer to the question; that is, write only the resulting set X^+ without any explanation.

(b) Show that R is not in Boyce-Codd normal form (BCNF).

(c) Normalize R to BCNF. Explain your solution step by step. Bear in mind that a relation may have several candidate keys.

Question 5. Data structures (1 + 1 = 2 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 400 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 records **i)** with a given value for X, and **ii)** with a given value for Y (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: $r_2(X)$, $w_2(X)$

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

pair 3: $r_3(Z)$, $w_2(Z)$

pair 4: $w_3(Z)$, $w_2(X)$

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

Step	T1	T2	T3
1	read(x)		
2	x=x+1		
3	write(x)		
4		read(x)	
5		x=x+1	
6		write(x)	
7		read(z)	
8		z=z+1	
9		write(z)	
10			read(z)
11			z=z+1
12			write(z)
13			read(y)
14			y=y+1
15			write(y)
16	read(y)		
17	y=y+1		
18	write(y)		

(c) Specify the two-phase locking (2PL) protocol; what does a transaction have to do to follow the protocol? (This is a general question; i.e., it is independent of the aforementioned schedule.)

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

(a) List the steps that a DBMS performs to create a checkpoint of a database.

(b) Assume a DBMS applies the deferred update strategy *without* checkpointing. Then, **i)** when does the DBMS write out to disk the data blocks that a transaction has updated, and **ii)** when does the DBMS write out the log buffers that contain log records about the transaction?

(c) 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.

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. Information Retrieval (2 p):

Assume that we use the vector model for information retrieval. Assume that we are only interested in the words “dog”, “mouse”, “pig”, and “rooster”. Assume that we have two documents in our document base.

- Document 1 contains “dog” 5 times, “mouse” 10 times, “pig” 0 times, and “rooster” 8 times.
- Document 2 contains “dog” 0 times, “mouse” 0 times, “pig” 7 times, and “rooster” 1 time.

Give the document representations for Document 1 and Document 2 according to the vector model *and* show/explain how you computed them.