Institutionen för datavetenskap
Linköpings universitet

EXAM
PART 2
Database Technology
TDDD12 – TDDD81

May 29, 2020
16:05 – 18:00

Instructions

See: https://www.ida.liu.se/divisions/adit/dwis/dbcourse/distanceexam.shtml

Grades

You can get up to 15 points for this second part of the exam. Hence, together with the max 15 points for the first part, you thus may get an overall of up to 30 points.

To pass the exam (grade 3) you have to meet both of the following two conditions: First, you need to achieve at least 8 of the 15 points that can be achieved in this first part of the exam. Second, for both parts together, you need to achieve at least 15 of the 30 points that can be achieved overall. If you do not meet the first condition, your second part will *not* be considered for grading.

After fulfilling the aforementioned requirements to pass the exam, then for grade 4, you need at least 21 points (for both parts together), and for grade 5, you need at least 27 points.

Questions

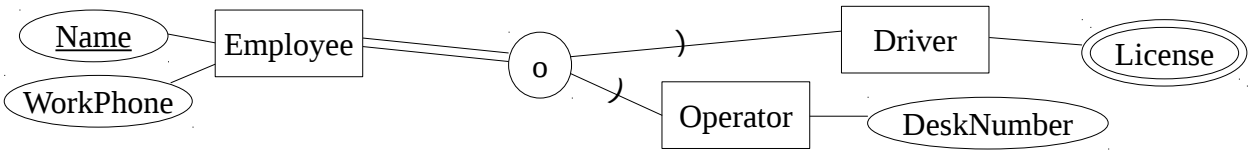
If you have clarification questions regarding some of the exercises in the exam, the examiner will be available in the following Zoom meeting room throughout the whole time of the exam.

https://liu-se.zoom.us/j/67490211293?pwd=YkNHTEpLMS9BNGJWOXdtOGh0K2JHdz09
Meeting ID: 67490211293
Password: 556432

Notice that this Zoom meeting room has been set up using the waiting room feature of Zoom. Hence, when you enter, you will be put into the waiting room and, from there, you will then be admitted to the meeting room to ask your question.

Question 9. Data modeling with an EER diagram (2 p):

Consider the following EER diagram. Reconstruct the written data requirements that may have led to the creation of this diagram. Hence, write a set of bullet points with these data requirements (similar to the bullet points in Question 1 of part 1 of the exam). Be as accurate as possible.



Question 10. EER diagram and relational schema (2 + 1 = 3 p):

(a) Translate the EER diagram from Question 9 into an equivalent relational database schema. Do not forget to include primary keys and foreign keys.

(b) There might be a constraint in the given EER diagram that is impossible to capture in your relational database schema. If so, specify which constraint that is. On the other hand, if all the constraints of the EER diagram are present in your relation database schema, then explicitly state here that this is the case.

Question 11. SQL (1 + 1 + 1 = 3 p):

Consider the same database as in Question 3 of part 1 of the exam. Recall that this database has been created by the following SQL statements.

```
CREATE TABLE Continent ( cid INTEGER  PRIMARY KEY,
                         name VARCHAR(30)  );
CREATE TABLE Country ( code INTEGER  PRIMARY KEY,
                       name VARCHAR(30),
                       continent INTEGER,
                       CONSTRAINT fk_cont FOREIGN KEY (continent)
                                          REFERENCES Continent(cid)  );
CREATE TABLE IsMember ( country INTEGER,
                        organization VARCHAR(30),
                        CONSTRAINT PRIMARY KEY (country, organization),
                        CONSTRAINT fk_ism FOREIGN KEY (country)
                                          REFERENCES Country(code)  );
```

(a) Remember the following faulty query from Question 3 where you were asked to identify a mistake that has been made in this query. In fact, this query contains two mistakes. What is the other one? To answer this question, please copy your earlier answer from Question 3(a) and then add one or two sentences that describe the second mistake. (If you did not answer Question 3(a), then you have another chance to describe one mistake here, but you will get only 1 point even if you find both mistakes now.)

```
SELECT continent, code, COUNT(*)
FROM Continent, Country
WHERE cid = continent AND name LIKE "A%"
GROUP BY continent;
```

(b) For the given database, write an SQL query that lists all countries (given by their name) together with the organizations in which they are members. Hence, the result of this query should be a two-column table with pairs of country names and organizations. Countries that are not a member of any organization must also be included in this list (with a NULL value for the organization).

(c) Write another query for the given database to list all organizations with more than five members.

Question 12. Functional Dependencies and Normalization (2 p):

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

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

First show that this schema is not in BCNF and, then, normalize it to BCNF. Explain your solution step by step.

Question 13. Data structures (2 p):

Assume we have a ***sorted data file*** with 1 000 blocks and 10 000 records. These records contain two fields, *Name* and *SSN*, where *SSN* is a key field and *Name* is not. The file is sorted on *Name*.

Suppose we have created a clustering index on the *Name* field where the index records have the same size as the data records in the aforementioned sorted data file, and the block size of the index file is also the same as the block size of the sorted data file. We may use this index to find records that have a specific value in their *Name* field (e.g., the value "Alice Smith"). Describe both the best case scenario and the worst case scenario for using this index and, for each of these two scenarios, compare the number of block reads needed in this scenario to the number of block reads needed for a binary search over the sorted data file instead.

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 14. Transactions and concurrency control (1 p):

Recall transaction T_1 from Question 6, but now with the necessary locking-related operations:

T_1 : lock write(X), $r_1(X)$, $w_1(X)$, unlock(X), lock write(Y), $r_1(Y)$, $w_1(Y)$, unlock(Y)

Explain in one or two sentences why this transaction does *not* follow the two-phase locking (2PL) protocol and modify the transaction so that it does follow the 2PL protocol.

Question 15. Database recovery (1 p):

In class you have learned that write-item log records are of the following general form:

[write_item, *T*, *X*, *old_value*, *new_value*]

Observe that the *old_value* may actually not be needed in some database systems. Specify for what systems this is the case and why.

Question 16. Query Processing (1 p):

Remember from the lecture that there are three desirable properties for a query optimizer:

1. its cost estimation is accurate,
2. it search space has low-cost QEPs, and
3. its enumeration algorithm is efficient.

Pick *one* of these properties, indicate which one you pick, and then explain in two to three sentences why this is a desirable property for query optimizers.