Institutionen för datavetenskap
Linköpings universitet

EXAM
TDDD74 Databases for Bioinformatics

August 20, 2016, 8.00-12.00

Help

No help such as dictionary, calculator, notes, books, etc. is allowed.

Grades

You can get max 32 points. To pass the exam, grade 3, you need 7.5 and 8.5 points in the practical and theoretical parts of the exam, respectively. For grade 4 and 5, you need 22 and 29 points, respectively.

Questions

Jose M. Peña will be available by phone.

Instructions

You can answer in Swedish or English. Write clearly. 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 EER diagram (5 p):

We want to create a database to store information about the allergies a group of people suffer. Specifically, we want to store the allergies each person has. For food allergies, we want to store the ingredient that causes the allergy as well as the products that contain such an ingredient, so that the person is aware of them when shopping. We also want to store the family relations that may exist between these people: Who is married to whom, and who is parent to whom.

Draw an EER diagram for the description above. Feel free to add the attributes that you consider necessary. Clearly write down your choices and assumptions in case you find that something in the information above is not clear.

Question 2. MySQL (1 + 2 + 2 = 5 p):

Consider the Jonson Brothers' relational schema used in the labs. The following relations should suffice to answer the queries below. However, you are free to use any other relation in the relational model used in the labs.

Relation: jbemployee
An employee is identified by an id and described by name, salary, birthyear and startyear. The id of the manager of each employee is also supplied. A null value means that the employee has no manager.

Relation: jbitem
An item is identified by an id and described by its name, the department where it is sold, its price, the quantity on hand (qoh) and the identifier of the supplier that supplied it.

Relation: jbsupplier
A supplier (of items and parts) is identified by its id and described by its name and the city in which it is located.

Relation:jbparts
A part, used internally by the store, not sold to customers, is identified by its id and described by its name, color, weight, and the quantity on hand (qoh).

Produce the MySQL code to answer the following queries:

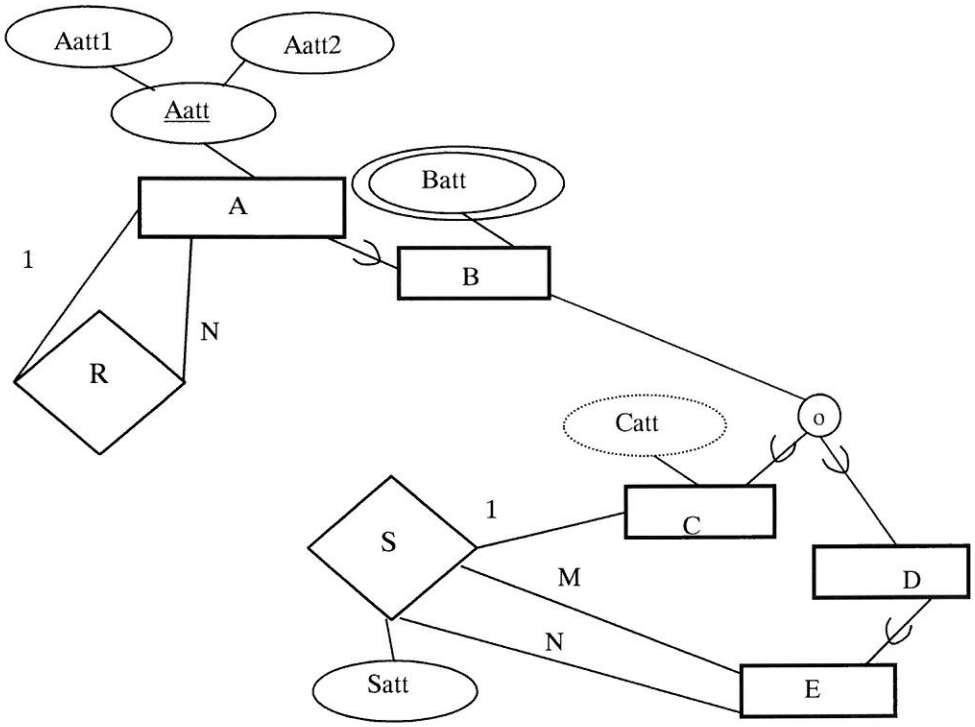
1. What was the age of each employee when they started working (startyear)?

2. Which items (note items, not parts) have been delivered by a supplier called Fisher-Price? Formulate this query using a subquery in the where-clause.

3. What is the name and color of the parts that are heavier than a card reader? Formulate this query without using a subquery in the where-clause.

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

Translate the EER diagram below into a relational schema. Use the algorithm you have seen in the course.



Theoretical part (17 points)

Question 4. Normalization (3 p):

Describe second, third and Boyce-Codd normal forms. Use just one sentence per normal form. Do not give examples.

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

We have a file with 1000000 records. Each record is 10 bytes long. The records have two key attributes X and Y. The file is ordered on X. The database uses a block size of B=1000 bytes and unspanning allocation. Each index record is 2 bytes long.

1. Calculate the average (or the maximum, if you prefer) number of block access needed to find a record with a given value for X when using (a) the primary access method and (b) a multi-level index (with as many levels as required).
2. Calculate the average (or the maximum, if you prefer) number of block access needed to find a record with a given value for Y when using (a) the primary access method and (b) a multi-level index (with as many levels as required).
3. Is there any maintenance costs associated with an index ? If so, can you name them ?

Question 6. Transactions and concurrency control (2 + 1 + 1 = 4 p):

1. Describe the two-phase locking protocol. Do not give examples but describe the protocol in general terms.

2. What is the purpose of the two-phase locking protocol ?

3. Complete the following sentences:

    a. Two schedules are conflict equivalent if ...

    b. A schedule is serializable if ...

Question 7. Database recovery (3 p):

Describe the three recovery methods you have seen in the course. Do not give examples but describe the methods in general terms.

Question 8. Information retrieval (2 p):

A model for information retrieval is described using 4 components.
D: how are documents represented?
Q: which queries can be asked and how are they represented?
F: how to connect document representations and query representations to answer the queries?
R: can answers be ranked, and if so, how?

Describe *and* give examples for D, Q, F and R for the boolean model.
For F make sure to show and exemplify each step in the calculation.