Uppsala University
Department of Information Technology

2014-20-21
Database Design I (1DL300)

Instructions:

Read through the complete exam and note any unclear directives before you start solving the questions.

Make sure you observe the following guidelines:

- The exam is divided into 3 parts:
  - PART I: COMMON QUESTIONS
  - PART II: TECHNO-TRACK (for non-X programme: IT, DV, etc.)
  - PART III: BIO-TRACK (for X programme).
- PART I must be answered by everyone. Then, you should choose to solve either PART II or PART III, depending on your study programme.
- Even if you feel like doing it, we recommend NOT TO solve both PART II and PART III.
- Some of the tasks are annotated with the symbol Ⓟ.
  - To achieve a grade of (3), you must answer correctly all the Ⓟ tasks.

    These cover the required learning outcomes of the course.
  - To achieve a grade of (4), you must answer correctly all the Ⓟ tasks and at least 40% of the other tasks.
  - To achieve a grade of (5), you must answer correctly all the Ⓟ tasks and at least 80% of the other tasks.
  - The teacher reserves the right to lower these thresholds.
- You are allowed to use dictionaries to and from English and a calculator, but **no other material**.
- Assumptions outside of what is stated in the question must be explained. Any assumptions made should not alter the given question.
- Write legibly and clearly. Solutions failing to follow this guideline will be marked as wrong. Well, should I even mention this?...

PART I: Common tasks

1 - ER modeling

The owners of a real estate agency want to digitalize their paper archive, to reduce their environmental impact. To allow the efficient extraction of information about the properties managed by the agency, they have decided to build a relational database and asked you to design it, using the EER model and according to the following requirements:

a) Ⓟ For each property, we need to know the extension of its land (expressed in square meters), a unique property identifier, and some text with a general description.

b) Ⓟ Each property belongs to one or more provinces. For each province we want to store its name, population, and a general description.

c) Ⓟ Each property contains zero, one or more buildings.

d) Ⓟ Each building has a unique identifier and can be of different types: a "villa" (in which case we want to store its name, e.g., "Villa Magnana"), an "apartment block" (in which case we want to store the number of apartments) or "other" (in which case we want to have a textual description of the building). No other types of building are possible.

(For the following tasks, please draw a separate EER diagram)

e) Modify the previous solution so that the identifier of the buildings is no longer globally unique, but is only unique with respect to the property to which it belongs.

f) In case a property belongs to more than one region, we want to store the percentage of the property present in each of them (the sum of these percentages for each property must be 100).

g) A province can be administratively dependent on another.

h) No province can have more than three other provinces depending on it.

HINT: clearly indicate all requirements that cannot be represented in the diagram or that you prefer not to represent in the diagram, to keep it simple, if any. Please, specify the correct cardinalities on the relationships (if you do not want to disappoint the owners of the agency).

2 - Translation to the Relational Model

Consider the part of the diagram corresponding to Exercise 1, task (d).

a) Ⓟ Translate it to a relational model schema (i.e., indicate the relations and their attributes).

b) Indicate all the primary keys.

c) Indicate all the foreign keys, if any.

d) Indicate all the attributes that allow NULL values, if any.

e) Indicate all the UNIQUE (sets of) attributes that are not part of a primary key, if any.

3 - Data Quality and Normalization

It turns out that the owners of the agency did not trust your ability to design a good database, so they also asked a professional information technologist called ”El Gargantua”. El Gargantua has deployed for them a relational database containing, among others, this relation:

PropertyID	BuildingID	BType	Province	Population	LandSize
PR001	B001	Villa	Uppland	743 000	200
PR001	B002	Apart	Uppland	743 000	200
PR001	B003	Apart	Nedland	253 000	200
PR002	B001	Villa	Östland	564 000	132

Now, you have to convince the owners of the agency that they have made a mistake hiring El Gargantua! (if you can, and if the relation has problems…)

a) Ⓟ Present an example of a problem that can occur after updating a value in the relation. (if you think that this relation cannot suffer from update/deletion problems, just say it)
b) Ⓟ Enumerate all the candidate keys.
c) Ⓟ Enumerate all the functional dependencies that are necessary to perform the normalization of this table (or to show that it does not need to be normalized).
d) Ⓟ What Normal Form is this table in (1NF, 2NF, 3NF, BCNF)? Why? If some FFDs violate the criteria for a certain normal form, identify them.
e) If the relation is not in BCNF, present an example of a problem that can occur after removing one or more rows from the table.
f) If the relation is not in BCNF, normalize it such that it fulfills the requirements of BCNF, list the resulting relations and mark primary keys and foreign keys. No extra attributes should be added.
g) Have all the original functional dependencies been preserved after normalization?
h) If the relation is not in BCNF, populate the relations obtained after normalization with the data in the original relation. You do not need to use SQL: just draw the tables and their content.

4 - Database concepts

a) Ⓟ Explain the concept of view (MAX one sentence).
b) Provide a motivation to use a view instead of a stored relation (MAX one sentence).
c) Provide a motivation to use a stored relation instead of a view (MAX one sentence).

5 - Theory

Indicate the correct answer. Every correct answer counts as "minus 1" – that is, the number of tasks used to get a grade of (4) or (5) you have solved correctly will be decremented. If you do not answer you do not gain anything and you do not lose anything.

a) Consider the following sequence of GRANTs (we indicate the user executing it at the bend of the statement, and Odin is the owner of the relation Valhalla):

- GRANT select(region) ON Valhalla TO Loki WITH GRANT OPTION; **-- *Odin***
- GRANT select(region) ON Valhalla TO Thor; **-- *Odin***
- REVOKE select(region) ON Valhalla FROM Thor; **-- *Loki***

What happens if these statements are executed in a relational database system?

1) Only Odin keeps the select privilege.
2) Odin and Loki are the only ones keeping the select privilege.
3) Odin, Loki and Thor all keep the select privilege.
4) None of the previous answers.

a) If A is an attribute of table T, which of the following SQL queries cannot return a value which is higher than the value returned by any of the two other queries?

1) SELECT COUNT(*) FROM T.
2) SELECT COUNT(distinct A) FROM T.
3) SELECT COUNT(A) from T.
4) None: they all return the same value.

b) In the relational model, if a set of attributes K is a candidate key of a relation r and X is an attribute of r not in K, then:

1) K $\cup$ {X} is also a candidate key.
2) K / {X} is also a candidate key.
3) K is also a primary key of the entity corresponding to r.
4) None of the previous answers.

(with "/" we notate the *set difference* operator)

c) In the context of normalization theory, assume that:

- $R(A_1, ..., A_n)$, $X \subseteq \{A_1, ..., A_n\}$, $Y \subseteq \{A_1, ..., A_n\}$, $Z \subseteq \{A_1, ..., A_n\}$, $W \subseteq \{A_1, ..., A_n\}$
- $X \rightarrow Y$
- $WY \rightarrow Z$

Then:

1) $X \rightarrow WZ$
2) $WX \rightarrow Z$
3) $X \rightarrow WY$
4) None of the previous answers

PART II: Technological track

6 - SQL queries

Consider the following database schema (where Registration indicates that a student attends a course):

Student(SID, Name, Surname, Age)

Registration(StudentID, CourseID)

Course(CID, Name, Cost)

a) Ⓟ Write an SQL query to extract the IDs of the courses attended by Carol Linneus.
b) Ⓟ Write an SQL query to extract the average cost of the courses attended by Carol Linneus.
c) Write a view corresponding to the query in task (a). (use this view to solve the next queries)
d) Write an SQL query to extract the most expensive among the courses attended by Carol Linneus.
e) Write an SQL query to extract the most expensive among the courses attended by Carol Linneus, not considering courses costing more than 135 kr.
f) Write an SQL query to extract the most expensive among the courses attended by Carol Linneus, not considering courses with less than 30 registered students.

7 - Query execution and optimization

Consider again the following database schema:

Student(SID, Name, Surname, Age)

Registration(StudentID, CourseID)

Course(CID, Name, Cost)

And the following SQL query:

```
CREATE VIEW NAMES AS
SELECT Name FROM Student UNION ALL SELECT Name FROM Course;
SELECT * FROM NAMES ORDER BY Name;
```

a) Write a relational algebra expression to extract all the names present in the database, as in the SQL view definition (you can use the RA operator $\cup_{ALL}$ to write the expression).
b) Describe a possible execution plan for the second SQL query, and estimate how long it would take to execute the query according to this plan, with the following parameters:
   - Tuples in Students: 30 000. Tuples in Registration: 100 000. Tuples in Course: 200.
   - Bytes to store the Name attributes: 100.
   - Data stored on a SSD with average data transfer rate: 600 MB/s.
   - Available main memory: 10 MB.

8 - Theory

Indicate the correct answer. Every correct answer counts as "minus 1" – that is, the number of tasks used to get a grade of (4) or (5) you have solved correctly will be decremented. If you do not answer you do not gain anything and you do not lose anything.

a) The cardinality of a left outer join between two relations r1 and r2 is:
   1) Always in [min(|r1|, |r2|), |r1| x |r2|].
   2) Always in [0, min(|r1|, |r2|)].
   3) Always in [0, max(|r1|, |r2|)].
   4) None of the previous answers.

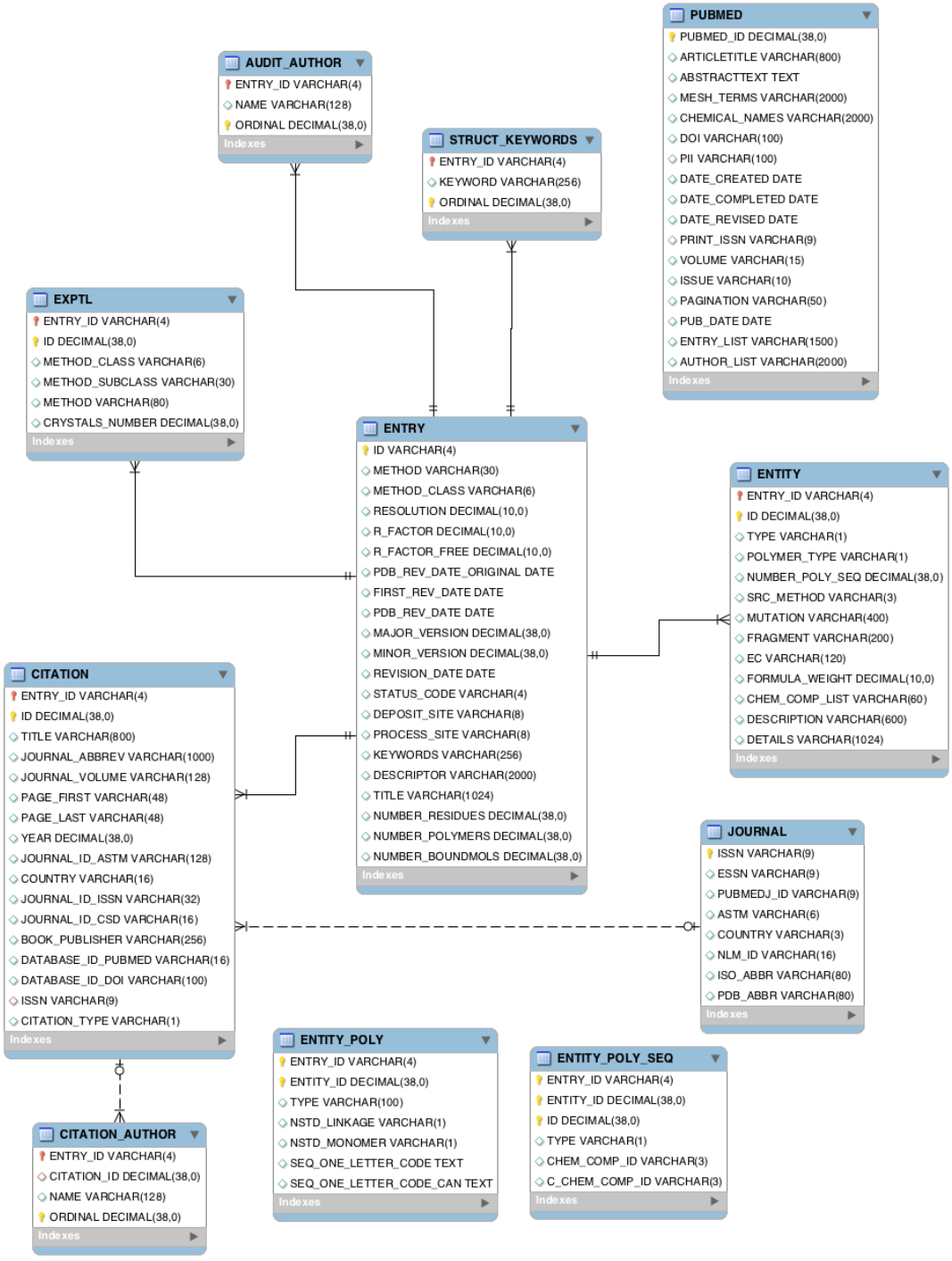
   (with |r| we indicate the cardinality or r, i.e., the number of tuples in r)

b) In relational algebra, assuming that E1 and E2 are union-compatible and contain the attribute X:
   1) $\boldsymbol{\sigma}_{X=35}(E1 \cup E2) = \boldsymbol{\sigma}_{X=35}(E1) \cup \boldsymbol{\sigma}_{X=35}(E2)$
   2) $\boldsymbol{\sigma}_{X=35}(E1 \cup E2)$ can be a proper subset of $\boldsymbol{\sigma}_{X=35}(E1) \cup \boldsymbol{\sigma}_{X=35}(E2)$
   3) $\boldsymbol{\sigma}_{X=35}(E1) \cup \boldsymbol{\sigma}_{X=35}(E2)$ can be a proper subset of $\boldsymbol{\sigma}_{X=35}(E1 \cup E2)$
   4) None of the previous answers

PART III: Bio-track

9 - PDB data extraction

Consider the following portion of the PDB database schema. In the following SQL queries we will assume that only amino acids are present in the stored sequences.



a) Ⓟ For each protein in the database return method, resolution and the title of each citation associated to it.

b) Ⓟ For protein 4r3g, count how many amino acids are present in each of its entities.

c) For each protein in the database, count how many amino acids are present in each of its entities.

d) Select the citation titles that are not present in PubMed.

e) Select the citation titles that are also present in PubMed.

f) Select all the proteins where NUMBER_RESIDUES does not correspond to the total number of amino acids in their sequences.

g) Create a view containing all protein ids of proteins having at least two associated entities.