Introduction to Database Design / Data Management MSc and BSc Exams

Björn Thór Jónsson

December 18, 2019

Instructions

You have 4 hours to answer 6 problems described in the following. There are 7 problems in the exam, but problem 2 is only for BSc students and problem 3 is only for MSc students. The exam consists of 10 numbered pages.

Instructions for SQL Queries in Question 1

Queries must work for any database instance and should be avoid system-specific language features. Queries should not return anything except the answer; a query that returns more information will not receive full points, even if the answer is part of the returned result. A sequence of several queries that answer the question will not receive full points, but subqueries and views can be used. Queries should be as simple as possible; queries that are unnecessarily complex may not get full marks, despite returning the correct answer. If you are unable to complete the query you can still submit your attempt, along with a brief description of the problem with the query, and it may be given partial points.

Database Description for Questions 1–3

In this exam you will work with a fictional database of countries, cities and languages. To start working with the database, run the commands in `idb-december-2019-DB.sql` found in LearnIT using the PostgreSQL DBMS on your laptop. It is recommended to use `psql` for this purpose.

The database contains a variety of information on countries in the following schema:

```
continents(Continent)
countries(Code, Name, Region, ..., Population, ...)
countries_continents(CountryCode, Continent, Percentage)

cities(ID, Name, CountryCode, District, Population)
empires(CountryCode, Empire)
countries_languages(CountryCode, Language, IsOfficial, Percentage)
```

Most attributes are self-explanatory. Primary and foreign keys are correctly defined, but you must study the DDL commands to understand the details of these. Some additional notes are in order:

- Some countries are present on more than one continent, and therefore have two entries in `countries_continents`; the `Percentage` attribute refers to the percentage of the population that lives on that continent.
- The table `empires` lists the constituent countries of some (fictional) empires. Countries that are not present in this table are not considered part of any empire.
- The `Percentage` data for languages in `countries_languages` also refers to the percentage of the population that speaks the language. The data is not complete, as the sum of percentages for countries is not 100.0 in all cases; this may be due to rounding errors or due to missing data.
- The data has various other errors, partly by design and partly because it is based on a publicly available dataset that has some errors in it.
- In `cities`, the `District` attribute refers to the region of the country where the city is located (rather than a district of the city).

1 SQL (40 points)

Answer each of the following questions using a single SQL query on the examination database:

(a) The empire 'Great Britain' consists of 4 countries. How many countries does the empire 'Danish Empire' consist of?

(b) In the database there are 3,264 cities which have a population that is less than 1% of the population of their country. How many cities in the database have a population that is more than 50% of the population of their country?

(c) There are 4 countries that are present on more than one continent. How many of these countries are partially in Europe?

(d) There are two countries for which the percentages of languages spoken add up to more than 100%. For how many countries in the `countries` table do the percentages add up to less than 100%?

(e) In the countries of North America that have more than 1 million inhabitants, there are a total of 164,688,674 people that speak Spanish, according to the statistics in the database. What is the corresponding number for South America?

(f) In France, the largest city is Paris with 2,125,246 inhabitants, while the smallest city is Montreuil with 90,674; the size ratio between the two is about 23.4. What is the ID of the city that has the highest size ratio relative to the smallest city from the same country?

Note: This query returns an ID of a city, not a count.

(g) According to the database, one language is spoken in all the countries of the 'Danish Empire'. How many languages are spoken in all the countries of 'Benelux'?

Note: This is a *division* query; points will only be awarded if division is attempted.

(h) Let us define the 'urban population' of a country as the population that lives in one of the country's cities, according to the database. For example, the urban population of the Netherlands is 5,180,049. Write a query to find the code of the country with more than 1 million inhabitants that has the highest *ratio* of urban population.

Note: The return value of this query is the three character country `Code` for a single country, not a number.

Enter each query, along with its numerical answer, in LearnIT. Queries must adhere to the detailed guidelines given on Page 1.

2 (BSc ONLY) SQL programming (5 points)

Consider the SQL trigger code in Figure 1.

```
-- Trigger function
CREATE FUNCTION CheckContinents() RETURNS TRIGGER
AS $$ BEGIN
  IF (2 < (SELECT COUNT(*)     -- Check 1: Number of continents
           FROM countries_continents
           WHERE CountryCode = NEW.CountryCode)) THEN
    RAISE EXCEPTION 'Too many continents'
    USING ERRCODE = '45000';
  END IF;
  IF (100 < (SELECT SUM(Percentage) -- Check 2: Percentage
             FROM countries_continents
             WHERE CountryCode = NEW.CountryCode)) THEN
    RAISE EXCEPTION 'Too large percentage for all occurrences'
    USING ERRCODE = '45000';
  END IF;
  RETURN NEW;
END; $$ LANGUAGE plpgsql;

-- Trigger code
CREATE TRIGGER CheckContinents
AFTER INSERT ON countries_continents
FOR EACH ROW EXECUTE PROCEDURE CheckContinents();

-- Test the trigger
INSERT INTO countries_continents VALUES ('ISL', 'Asia', 60);
INSERT INTO countries_continents VALUES ('ISL', 'Europe', 60);
INSERT INTO countries_continents VALUES ('ISL', 'Africa', 60);
```

Figure 1: Insertion trigger `CheckContinents` for the `countries_continents` relation.

a) Select the true statements:

(a) Check 1 cannot be replaced by a CHECK constraint on the `countries_continents` relation.

(b) Triggers in RDBMSs are only useful for checking complex constraints.

b) If the `countries_continents` relation is empty when the three INSERT statements are issued, while the relevant country and continents exist, which INSERT statement will be the first to give an error:

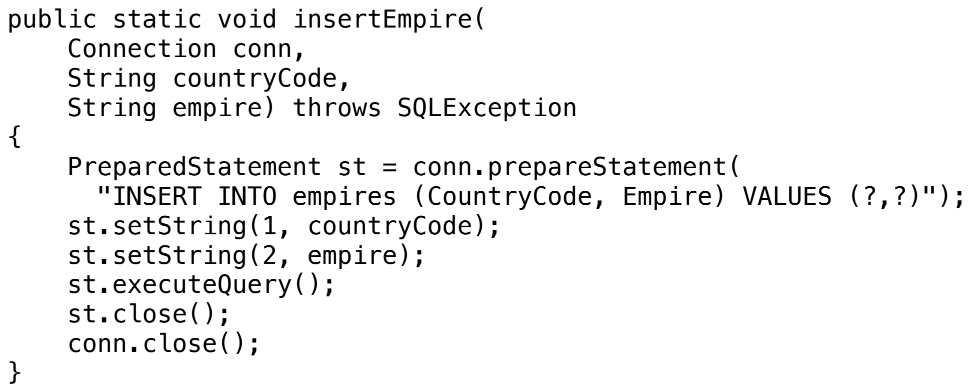
(a) The first INSERT statement will be the first to give an error.

(b) The second INSERT statement will be the first to give an error.

(c) The third INSERT statement will be the first to give an error.

3 (MSc ONLY) Database programming (5 points)

To answer this question, you will need to study the `empires` relation of the exam database. Consider the Java code in Figure 2 (it is an image, so the code cannot be copied from the PDF).



```
public static void insertEmpire(
    Connection conn,
    String countryCode,
    String empire) throws SQLException
{
    PreparedStatement st = conn.prepareStatement(
      "INSERT INTO empires (CountryCode, Empire) VALUES (?,?)");
    st.setString(1, countryCode);
    st.setString(2, empire);
    st.executeQuery();
    st.close();
    conn.close();
}
```

Figure 2: Code for inserting a country into the `empires` relation.

Select the true statements:

(a) Closing the database connection inside the function is important, because it frees up resources.

(b) An advantage of prepared statements is that they reduce the likelihood of introducing security vulnerabilities.

(c) Using executeQuery will throw an exception for an INSERT statement.

(d) The code is using Object Relational Mapping.

4 ER Diagrams and Normalization (25 points)

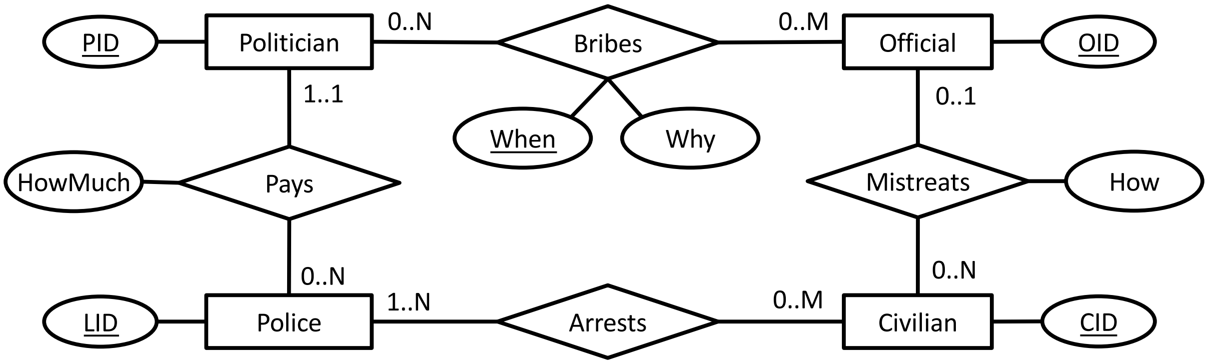


Figure 3: ER Diagram for the corruption database.

a) The ER diagram in Figure 3 shows a database for a very corrupt society. Select the true statements. You should base your answers **only** on the ER diagram:

(a) All civilians have been arrested.

(b) All policemen have made an arrest.

(c) Every civilian is linked to at least one politician via the relationships.

(d) Civilians can be mistreated multiple times.

(e) When converted to SQL DDL according to the methodology presented in class, the resulting database should have exactly 8 tables.

(f) When converted to SQL DDL, the table for the Bribes relationship will have a primary key with two attributes.

b) Write SQL DDL commands to create the research database based on the ER diagram in Figure 3. The DDL script must run in PostgreSQL. The relations must include all primary key, canditate key, foreign key and NOT NULL constraints. Constraints that cannot be enforced with standard primary key and foreign key constraints can be omitted. Make reasonable assumptions on the attribute types.

c) Write an ER diagram for a medical database. The diagram should clearly show the entities, relationships and participation constrains described below. Use the notation presented in the textbook and lectures. Attributes are not important. If you need to make additional assumptions put them in the box below.

- Products may be crafted or grown, some may be both.
- Each product is verified by one vendor.
- Products are sold by vendors to customers, and the date, quantity and price are stored.
- The same product can be sold by the same vendor to the same customer, but only once per each day.
- Occasionally, an auditor checks the validity of a sale.

d) Consider a table $R(A, B, C, D, E)$ with the following dependencies:

$$\begin{aligned} AB &\rightarrow CDE \\ A &\rightarrow C \\ D &\rightarrow E \\ A &\rightarrow A \end{aligned}$$

Select the true statements:

(a) AB is the only (candidate) key of R.

(b) $A \rightarrow C$ is an unavoidable functional dependency.

(c) Normalizing to 3NF or BCNF results in exactly two relations.

(d) The relation can be normalized to BCNF without losing functional dependencies (excluding trivial, unavoidable, and derivable dependencies) .

e) Consider a table $R(A, B, C, D, E)$ with the following dependencies:

$$\begin{aligned} A &\rightarrow BCD \\ E &\rightarrow A \\ C &\rightarrow D \\ CB &\rightarrow E \end{aligned}$$

Normalize R to the highest possible normal form based on functional dependencies (3NF or BCNF), while allowing all functional dependencies (excluding trivial, unavoidable, and derivable dependencies) to be checked within a single relation, and write down the resulting relations.

5 Index Selection (10 points)

Consider the following large relations with information on clients:

Clients(ID, name, birthday, salary, <many long attributes>)
Represents(clientID, agentID, role)

For each of the queries below, select the index that a good query optimiser is most likely to use for the Clients table to process the query. Assume that all indexes are unclustered B+-trees. Also, assume that attribute values are non-nullable and correspond to reality, and that the query optimizer has basic (approximate) statistics, such as smallest and largest value of each attribute. Each each case, select the best index, or select "no index" if a full table scan would yield better performance than any index.

(a) Clients(ID)

(b) Clients(birthday)

(c) Clients(salary)

(d) Clients(salary, name)

(e) Clients(birthday, salary, name)

(f) No index

The queries are:

Query 1

```
select C.ID, C.salary
from C.Clients
where C.birthday = 12-12-1989;
```

Query 2

```
select C.Name
from Clients C join Represents R on C.ID = R.clientID
where R.agentID = 99856 and R.role = 'manager';
```

Query 3

```
select avg(C.salary)
from Clients C
where C.salary > 0;
```

Query 4

```
select C.name
from Clients
where C.salary > 1234;
```

6 Hardware and DBMS Design (10 points)

a) Select the correct statements below:

(a) Transaction isolation is easier to manage with very short transactions than with very long transactions.

(b) Data replication in a distributed system eliminates the risk of losing data.

(c) Compared to older persistent storage technology, solid state disks (SSDs) are particularly effective for small random reads.

(d) The CAP theorem applies to normal operation of large-scale distributed systems.

b) Imagine that 10 years from now a new type of persistent storage emerges that is a) as fast as regular memory, and b) similarly priced, making it feasible to replace main memory with this new storage medium. Compared to traditional relational management systems, how could the implementation of ACID transaction processing be simplified for servers that using this new storage medium as RAM replacement.

7 Data Systems for Analytics (10 points)

a) Select the correct statements below:

(a) Sequential disk reads are the most important disk access pattern in big data analytics.

(b) In Big Data applications, “velocity” has two potential meanings: a) that data is added very rapidly, and b) that one must react rapidly to the added data in many cases.

(c) The novelty of Hadoop MapReduce was primarily the invention of the Map and Reduce operations.

(d) In Big Data applications, it is important to verify that the data is clean and applicable to the analysis that is to be undertaken.

b) Discuss the pros and cons of using Spark to implement interactive big data applications.