Introduction to Database Systems
BSc and MSc Trial Exam
Spring 2023

Jorge Quiané

May 8, 2023

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. Unless instructed otherwise, your answers must be provided in the LearnIT quiz *Trial Exam*.

Database Description for Questions 1–3

In this exam you will work with the database `cargo`. To start working with the database, import/run `idb-fall-2022-trial-DB.sql` found in LearnIT using the PostgreSQL DBMS on your laptop. The database contains information on flights, aircrafts and airports. Note that you may already have a database called `cargo` on your server, but the exam version is different, so you must import/run the script.

```
acgroup(ag, agfullname)
aircraft(actype, actypefullname, capacity, ag)
airport(airport, country)
country(country, region)
flights(id, al, dep, arr, actype, start_op, end_op, ...)
```

Most attributes are self-explanatory, and primary and foreign keys are generally defined as expected. Countries are represented by a two-letter country code; you will need to use the codes for Denmark (DK), Germany (DE) and Netherlands (NL). Regions are essentially continents, also represented by two-letter codes; you will need to use the codes for Europe (EU) and Asia (AS).

The `flights` relation is the most complicated one and warrants further explanation:

- There are more attributes in the relation, but you will only need the attributes given above in the exam.
- The `al` attribute is a two-letter code for the airline in question. The actual names of the airlines are not relevant.
- The `dep` and `arr` attributes are the departure and arrival airports, respectively. Both attributes have foreign key constraints to the airport relation.
- The `actype` is the type of the aircraft, with a foreign key constraint to the aircraft relation.
- The `start_op` and `end_op` attributes indicate how long the flight has been (or was) running. In these queries we make no distinction between flight routes that are currently running and flight routes that have stopped.
- You can compute the total running time of a flight route in days using (end_op - start_op).

1 SQL (40 points)

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

(a) In Denmark, there are 2 registered airports. How many airports are registered in Germany?

(b) In Asia, there are 57 airports that have both departing and arriving flights. How many airports in Europe have both departing and arriving flights?

(c) The average number of days that a flight route has been running is 42.77. For how many days has the longest running flight route been running?

(d) There are 6126 flights that a) depart from an airport within Europe and b) have an aircraft capacity of more than 300 passengers. How many flights with more capacity than 300 passengers depart from an airport within Asia?

(e) Each aircraft has a registered aircraft group (aircraft.ag). The smallest such aircraft group has 2 members. How many members does the largest group have?

Hint: Using a view can simplify the query significantly. If you do, include the view creation statement in your answer.

(f) According to the flights relation, there are 124 airports with more departing flights than arriving flights. How many airports have more arriving flights than departing flights?

(g) Only 1 airline has flights departing from every registered airport in Denmark. How many airlines have flights departing from every registered airport in the Netherlands?

Instructions for SQL Queries in Question 1

Enter each query, as well as the numerical answer to each question in LearnIT. Queries must return correct results for any database instance. They should avoid system-specific features, including the LIMIT keyword. 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, and it may be given partial points.

2 (BSc ONLY) SQL Programming (5 points)

Consider the SQL trigger code in Figure 1.

```
DROP TRIGGER IF EXISTS CheckDate ON flights;
DROP FUNCTION IF EXISTS CheckDate();

-- Trigger function
CREATE FUNCTION CheckDate() RETURNS TRIGGER
AS $$ BEGIN
  -- Check: Is the operating time range OK?
  IF (NEW.END_OP < NEW.START_OP) THEN
    RAISE EXCEPTION 'Cannot operate for a negative duration'
    USING ERRCODE = '45000';
  END IF;
  RETURN NEW;
END; $$ LANGUAGE plpgsql;

-- Trigger code
CREATE TRIGGER CheckDate
BEFORE INSERT ON flights
FOR EACH ROW EXECUTE PROCEDURE CheckDate();

-- Test the trigger
INSERT INTO flights VALUES (500000,'LH','LH001','HAM','FRA',610,720,'321','2010-10-16','2010-10-05','MO,TH,SA,SU');
INSERT INTO flights VALUES (500001,'LH','LH001','HAM','FRA',610,720,'321','2010-10-16','2011-10-05','MO,TH,SA,SU');
INSERT INTO flights VALUES (500002,'LH','LH001','HAM','FRA',610,720,'321','2010-10-16',NULL,'MO,TH,SA,SU');
```

Figure 1: Trigger `CheckDate` for the `flights` relation.

Select the true statements:

(a) The check can be replaced by a CHECK constraint on the `flights` relation.

(b) The first INSERT statement will give an error.

(c) The second INSERT statement will give an error.

(d) The third INSERT statement will give an error.

3 (MSc ONLY) Database programming (5 points)

Consider the Java code in Figure 2.

```
public static void insertAcgroup(
    Connection conn,
    String AG,
    String agFullname) throws SQLException
{
    PreparedStatement st = conn.prepareStatement(
      "INSERT INTO acgroup (ag, agfullname) VALUES (?,?)");
    st.setString(1, AG);
    st.setString(2, agFullname);
    st.executeQuery();
    st.close();
    conn.close();
}
```

Figure 2: Code for inserting into the `acgroup` 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.

4 ER Diagrams and Normalization (25 points)

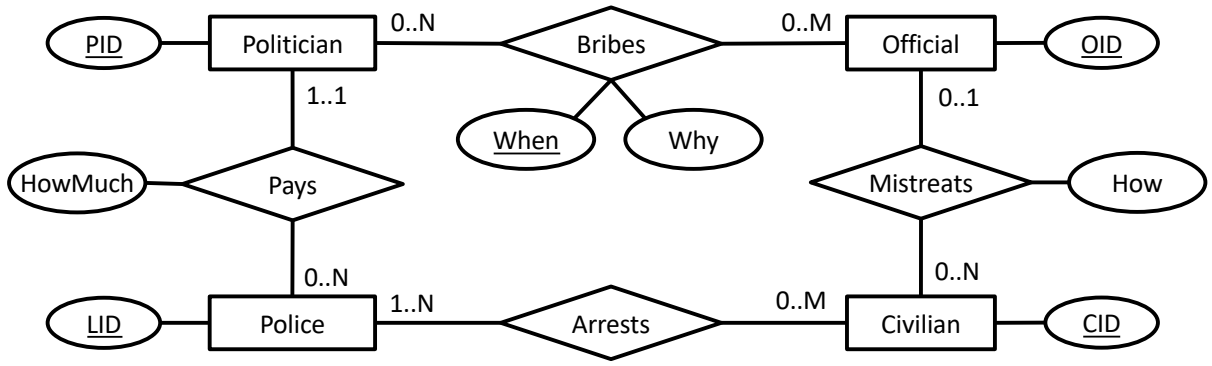

Figure 3: ER Diagram for the corruption database.

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 policemen have made an arrest.

(b) All politicians have taken a bribe.

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

(d) Civilians can be mistreated multiple times.

(e) When converted to SQL DDL according to the methodology presented in the class, the resulting database should have exactly 7 tables.

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

(g) All civilians have been arrested.

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

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

Select the true statements:

(a) A is the only key of R.

(b) $BCD \rightarrow D$ is a trivial functional dependency.

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

(d) The relation $Z(A, B, C, D)$ is in BCNF.

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

$$\begin{aligned} AC &\rightarrow B \\ DE &\rightarrow ABC \\ DE &\rightarrow D \\ A &\rightarrow C \end{aligned}$$

Normalize R to the highest possible normal form (3NF or BCNF), based on functional dependencies, while allowing all functional dependencies (excluding trivial, unavoidable, and redundant dependencies) to be checked within a single relation. For each resulting relation, write its columns and clearly indicate whether it is in BCNF.

5 Index Selection (10 points)

Consider the following large relation with information on employees:

Emp (id, name, age, sal, <many long attributes>)

Assume that the attribute sal follows a normal distribution with a maximum value of 50, and the attribute age follows a uniform distribution between 25 and 75. Additionally, consider the following three SQL queries:

Query 1

```
select id, name
from Emp;
```

Query 2

```
select id
from Emp
where age > (select max(sal) from Emp);
```

Query 3

```
select age
from Emp
where sal = (select avg(sal) from Emp);
```

Answer each of the following questions:

(i) Select the correct statements in the following:

- (a) Query 1 can only benefit from a covering index. Otherwise, a full table scan is preferable.
- (b) Query 2 will benefit from an index on sal.
- (c) For Query 2, an unclustered index on age will perform better than a clustered index on age.
- (d) Query 3 will benefit from an index on age.
- (e) For Query 3, a clustered index on sal will perform the same as an unclustered index on sal.

(ii) Indicate for each query whether a covering index would be preferable to a clustered index. Explain your answer and define the indexes you consider.

(iii) Considering all three queries, which single clustered index would you define on the relation? Explain your answer.

6 Hardware and DBMS Design (10 points)

(i) Select the correct statements below:

(a) It is easier to achieve serializability when transactions take a short time to be executed.

(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 notion of consistency in the CAP theorem is the same as the notion of consistency in ACID.

(ii) 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 replace RAM with this new storage medium?

7 Data Systems for Analytics (10 points)

(i) Select the correct statements below:

(a) Sequential disk writes 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) MapReduce is a parallel programming model.

(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.

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