Exam Datawarehousing INFOH419 July 2013

Lecturer: Toon Calders

Student name: .....

The exam is **open book**, so all books and notes can be used. The use of a basic calculator is allowed. *The use of a laptop computer, tablet, cell phone or any other electronic device that can be used for communication is strictly not allowed.*

Please use the empty space on the forms to formulate your answers to the questions. These spaces should be sufficient for answering the questions. If you need more space, use the extra empty pages at the end and **clearly** indicate where your answer can be found. Please, do not forget to complete your name on every page.

Success!

Question	Score	Max
1		6
2		4
3		6
4		4
Total		20

1. (6p) In order to analyze the delays of their trains, a railway company decides to create a data warehouse in which they store all information relevant to the train delays. For every trip of a train that took place, the database should contain:

   - The departure and destination station;
   - The date of the trip;
   - The *planned* departure and arrival times;
   - The delay in minutes at arrival and at departure;
   - The locomotive with which the trip was executed. Every locomotive has a unique number, a type, engine type (diesel or electricity), and total horsepower. There can be different locomotives of the same type. The type determines the engine type and the total horsepower.
   - The conductor. For the conductor, his or her name, birth date, place of living, salary, and the types of trains he or she is allowed to conduct are stored as well.

   Based on this data, the railway management would like to analyze, on a regular basis, the delays of the trains. In such analysis the train delays will typically be aggregated by time of the day, day of the week, by departure or destination station, or line (source-destination pair), and when systematic problems are detected on one or more lines, even an overview of the delays per conductor on these line may be requested.

   Question: Propose a dimensional model for the data warehouse at the conceptual level. Indicate which cube(s) are needed, what are the dimensions, measures, hierarchies, etc. Show the tables for storing the data cube in a relational database; that is, in a ROLAP solution; make sure that your tables can accommodate the following changes: a conductor may learn how to drive a type of train he or she could not drive before, or the conductor's salary and/or place of living may change.

Name:

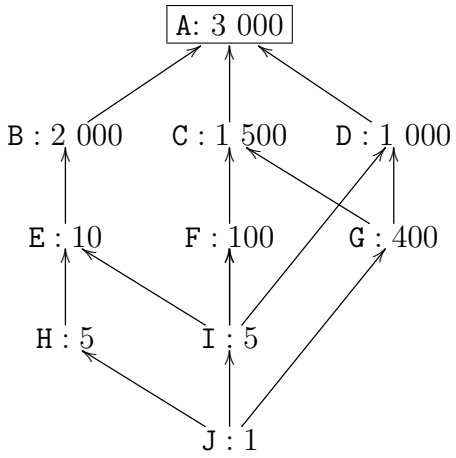
2. (4p) *Jensen et al.* make in their book *Multidimensional Databases and Data Warehousing* the following statement regarding the use of a surrogate key as the primary key in a dimension table of a star-schema:

> [A surrogate key as key column in a dimensional table] has several advantages over the option of information-bearing keys from the source systems, including better storage use, prevention of problems associated with key re-use, better support for dimension updates, and more efficient query processing.

Explain, **in your own words, and with an example of your own** the second and the third benefit; i.e., (a) why a surrogate key prevents problems associated with key re-use, and (b) why it better supports dimension updates.

Name:

3. (6p) An important technique to speed up analytical queries is by pre-computing and materializing aggregations. Consider the following lattice of views that can be requested by the user, along with the number of rows in each view.



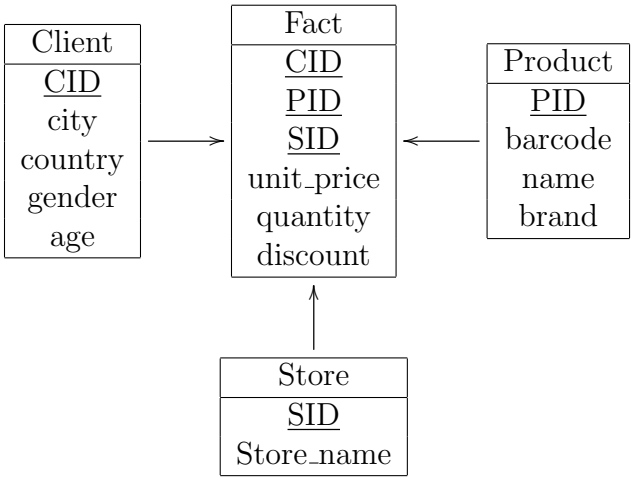
A is the view representing the base relation. The edges indicate the relation "can be computed from."

(a) Suppose that only the top-view `A` has been materialized. Select three additional views from the views `B`, `C`, `D`, `E`, `F`, `G`, `H`, `I`, and `J` to materialize. Apply the greedy method described by *Harinarayan, Rajaraman, and Ullman* in their seminal paper "Implementing Data Cubes Efficiently" (SIGMOD 1996)

(b) What benefit gives the additional materialization of these three views under the cost model introduced by these authors?

Name:

4. (4pt) Consider the following Star schema for a ROLAP database. A fact (c, p, s, up, q, d) represents that customer c bought q units of product p in store s at unit price up and received a discount d.



(a) Discuss the advantages and disadvantages of the following indices; for instance, describe under what circumstances and for which types of queries are these indices interesting.

i. a bitmap join index for the attribute country from the table Client into the fact table (mapping countries to tuples in the fact table that are about this country);

ii. a bitmap index on table Client for the attribute gender; and,

iii. a bitmap index for the attribute unit_price in table Fact.

(b) Consider the following query:

```
SELECT Client.Gender, SUM(Fact.quantity)
FROM Fact, Client
WHERE Fact.CID=Client.CID AND
      (Client.country = "The Netherlands"
       OR Client.country = "Belgium")
group by Client.Gender
```

Select one of the indices from (a) and explain how this index may help to answer this query efficiently.

Name:

Extra page