Exam Datawarehousing INFOH419 July 2014

Lecturer: Toon Calders

Student name: .....

The exam is **open book**, so all books and notes can be used. You are also allowed to use your laptop computer for consulting the slides, but you are not allowed to use any network or bluetooth connection. Any attempt to communicate to each other during the exam, electronically or in any other way, is strictly not allowed. Being online with your device will be considered as sufficient proof of fraud.

"Open book" implies that you can refer to a specific slide, book page, or exercise. Verbatim copying lecturing material will not be rewarded.

Use the empty spaces directly following the questions to write down your answers. These spaces should in principle 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. Stay focused on the question and avoid excessively long answers; succinct, to the point answers will be rewarded.

The total time foreseen for this exam is 2h30 and will be strictly observed. Plan your exam accordingly and avoid excessively long responses!

Please, do not forget to complete your name on every page.

Success!

Question	Score	Max
1		7
2		6
3		5
4		2
Total		20

1. (7p) Read the following description stating the requirements for a data warehouse for football statistics: *In order to professionalize their scouting activities, a football team decides to keep detailed statistics of all football players of all major European and Southern American leagues. For each player it is recorded for which team he plays and for every match, how many goals he scored, number of faults committed, position he played, number of passes completed, number of assists (last pass leading to a goal), yellow and red cards received during the match, number of minutes played, etc. If a player is transferred from one team to another, the transfer amount and the duration of the new contract are stored as well. Even if a player does not play a single match for a team, his membership in that team needs to be registered. For the matches it needs to be stored what was the home and what the away team, and in what competition the match was played (national competition, UEFA league, Champions league, confederations cup, ...).*

   Based on the data stored in the data warehouse, it should be possible to answer, for instance, the following questions:

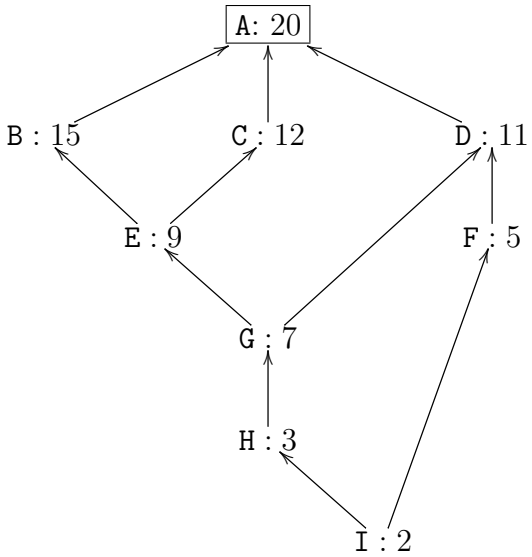
   - *Give the total minutes played for all players of Bayern Munich in national competition matches during the season 2013-2014.*
   - *Give the total number of goals scored in January 2013 by French teams.*
   - *Give per team the sum of all transfer prices received and paid.*
   - *Give per match of Barcelona the number of goals scored by Neymar or Messi.*

   (a) Draw a dimensional fact model (DFM) for the data warehouse. Notice that there is not a single correct solution and that multiple facts may be needed. For every fact in your model explain in one sentence what it represents (for instance: "Fact (c, p, s, d, u) in FactSales = customer c bought product p in store s on date d for unit price u"). Not all information in the description is necessarily relevant for the dimensional fact model. Your model should represent the description as faithfully as possible.

   (b) Translate the DFM you constructed to an appropriate relational model. Clearly indicate primary and foreign keys in your tables. *In order to gain full points for (b) the model given in your answer to (a) must be reasonable. For instance: not answering (a) and giving an empty relational schema as a (correct) answer for (b) will obviously not gain you any points.*

Name:

2. (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 size of 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. There is space available to additionally materialize views for a total size of 32. Select additional views from the views `B`, `C`, `D`, `E`, `F`, `G`, `H`, and `I` to materialize. Apply the modification for a space upper bound of 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 views under the cost model introduced by these authors?

Name:

3. (5p) Consider the following dimension table. For reference also part of the fact table FactSales is given:

DimProduct
operationalProductKey
productName
productDescription
productCategory
brand
brandManagerName

FactSales
operationalProductKey
CustomerID
...
price
discount

(a) Assume that productName, productCategory, productDescription and brandManagerName may change during the lifetime of the data warehouse. How can you adapt this dimensional table such that such changes can be dealt with in the data warehouse? (Show only the one way to adapt the schema that is the best choice according to your opinion.)

(b) Show the content of the dimension table after first product (0001,"brazooka","official ball of the World Cup in Brazil","football","Adidas",EMP00178) is added, then the product name is updated to "brazuca" and later on the category is updated to "soccer."

Name:

4. (2p) Deduplication is an important task in data cleaning. To do deduplication it is useful to have a measure of distance between string values such as names, addresses, phone numbers, etc. The edit distance is one such distance measure. Compute the edit distance between “Brussel” and “Bruges”.

Name:

Extra page

Extra page