Exam Datawarehousing INFOH419 January 2015

Lecturer: Toon Calders

Student name: .....

The exam is **open book**, so all books and notes can be used. "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 maximal time to complete this exam is 3h and will be strictly observed. Plan your exam accordingly.

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

Success!

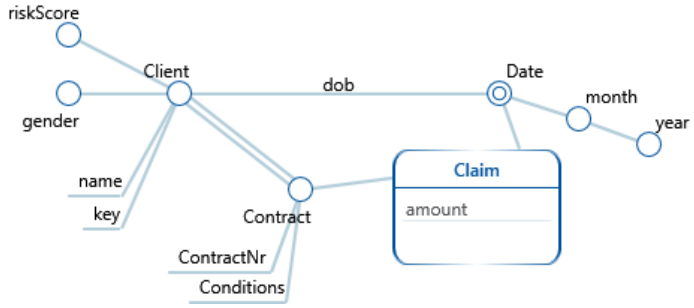
Question	Score	Max
1		6
2		5
3		5
4		3
5		1
Total		20

1. (6p) Carefully read the following data description and create a **dimensional fact model** that expresses as faithfully as possible this description:

   A university wants to analyze the results of its students and the quality of its courses over time. For this purpose they want to create a data warehouse that will store the results of all students. A student is registered for a program which consists of several courses taught by one or more lecturers. Every course has a course code, a name, the number of lecturing hours and the number of hours for exercises and lab sessions. A lecturer belongs to a department and a department to a faculty. A student can take one or more exams for a course. In case of multiple exam trials, all results should be stored. For a student, his or her name, student identifier, gender, and date of birth are stored. Furthermore, every course offering is evaluated (anonymously) by the students. This evaluation results in three scores: one for course delivery, one for course content and one for overall appreciation. Based on the data in the data warehouse it should be possible to evaluate the exam results by student, by program, by academic year, by course, by department, etc. Furthermore courses will be evaluated on the basis of their evaluations by year, by lecturer, by program, etc.

Name:

2. (5p) The following fragment of a fact model describes insurance claims. Every claim is associated to a contract and a contract involves one or more clients. For every client his or her name, gender, date of birth and a risk score that is regularly updated (for instance: risk decreases after one year without claim and increases every time the client is involved in a claim) are recorded.



It is the insurance company's policy to never change any property of a contract. In case a change is needed, the contract is terminated and a new one is signed. Consider the following (possibly incorrect) implementations of this DFM into the relational model. CID and ConID represent surrogate keys that were added during data warehouse design. Client_OLTP_key and ContractNr represent the primary keys of respectively clients and contracts in the source database system. Start and end are attributes that indicate the valid time of a tuple.

Solution 1:

- dimClient(CID, Client_OLTP_key, name, gender, dob, risk_score, start, end)
- dimContract(ConID, ContractNr, conditions, start, end)
- bridgeContractClient(CID,ConID, start, end)
- factClaim(DateID,ConID,amount)

Solution 2:

- dimClient(CID, Client_OLTP_key, name, gender, dob, RID, start, end)
- dimContract(ConID, ContractNr, conditions)
- dimRisk(RID, risk_score)
- bridgeContractClient(CID, ConID)
- factClaim(DateID, ConID, amount)

Solution 3:

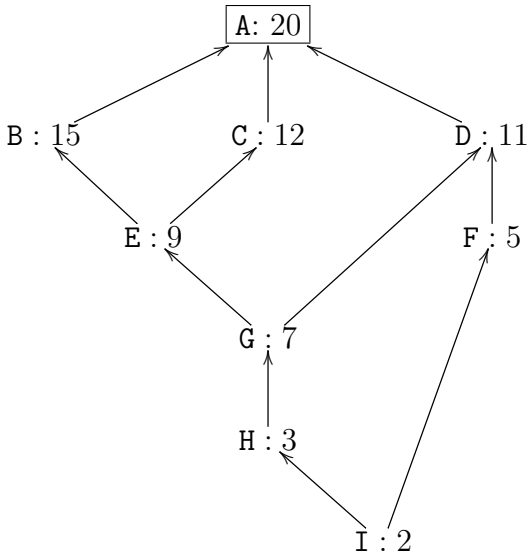
- dimClient(CID, Client_OLTP_key, name, gender, dob, start, end)
- dimContract(ContractNr, Client_Group_ID, conditions)
- dimRisk(RID, risk_score)
- ClientGroup(Client_Group_ID, CID)
- factClaim(DateID, ConID, RID,amount)

(a) Discuss the advantages and disadvantages of the three solutions, and point out the errors (if any).

(b) Select one of the solutions as the preferred solution, or propose an alternative solution yourself if none of the solutions above is satisfactory.

Name:

3. (5p) 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 size of each view.



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

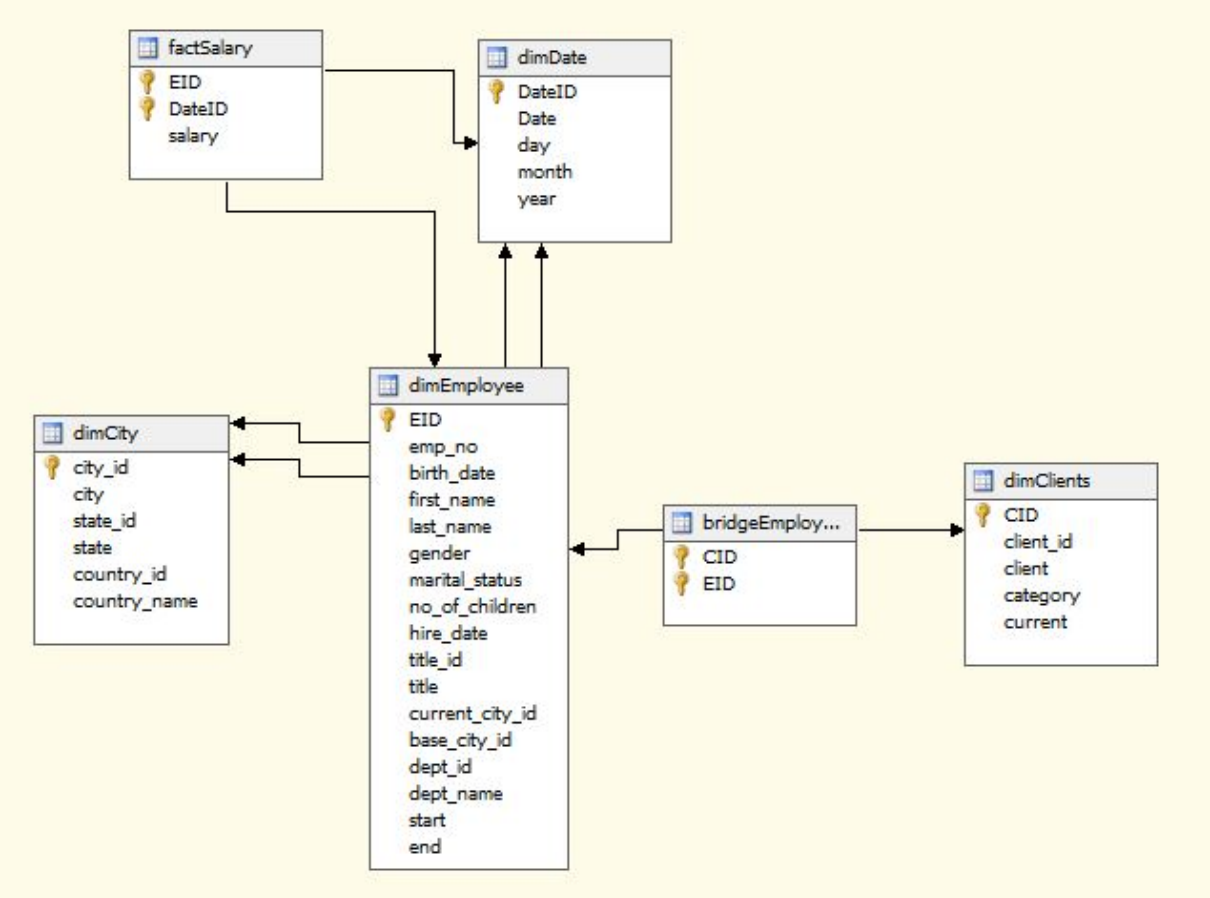
Furthermore, we have the following usage statistics of the views. The views are requested with the following frequencies: A:0%, B:10%, C:10%, D:20%, E:10%, F:20%, G:10%, H:10%, I:10%.

(a) Suppose that only the top-view A has been materialized. Select 3 additional views from the views B, C, D, E, F, G, H, and I to materialize. Apply the greedy method described by *Harinarayan, Rajaraman, and Ullman* in their seminal paper "Implementing Data Cubes Efficiently" (SIGMOD 1996) in order to optimize the *expected* query time.

(b) What is the expected relative gain in speed under the cost model of *Harinarayan et al.*?

Name:

4. (3p) Consider the database scheme of part II of the course assignment:



(a) Give an example of one or more bitmap and/or bitmap join indices for this database and a query that would substantially benefit from these indices.

(b) Explain how the indices you propose should be exploited during the execution of this query.

Name:

5. (1p) 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 “Brussels” and “Blois”.

Name:

Extra page

Extra page