Databases and Data Mining 2016
Final Exam

LIACS Room 407/409
Friday December 23rd 2016
10.00 – 13.00

- State your name and student number and affiliation (e.g. LU (Leiden University), TUD (TU Delft), etc.) on every page of your answers.
- Every assignment has the same weight. There are 10 assignments.
- **Always fully motivate and explain your answers!** (Answering a question with only 'yes' or 'no' will never be counted as a correct answer.) Use examples and/or sketches whenever / wherever you consider them useful to explain your answers.
- Please note that you have a total of 3 hours to answer the questions.
- It is an open book exam: you are allowed to use your book and course notes (slides).
- **All electronic equipment should be off the table and switched off.**

1. Consider a base cuboid with five dimensions *A, B, C, D, E,* with the following numbers of distinct values per dimension:
   A: 10,000
   B: 100
   C: 10
   D: 1,000
   E: 100,000

   Each cell contains three 4-byte integer measures. Suppose the 0-d cuboid ("apex"/"all") needs to be computed with only the base cuboid precomputed and materialized.
   a) How often do you need to traverse the base cuboid to compute the aggregates for all three measures?
   b) In which order would you traverse the base cuboid to require the least amount of memory for computing the 0-d cuboid?
   c) What is the minimum amount of memory required to compute the 0-d cuboid directly from the base cuboid?

2. For modern data analytics workloads (OLAP, BI, Data Mining), columnar data storage show considerable performance advantages over row-wise data storage.
   a) Explain the principle differences between columnar and row-wise data storage.
   b) Explain why and how columnar data storage provides performance advantages over row-wise data storage for analytical workloads.
   c) Name other workload types that favor row-wise storage and explain why.

3. Data in data warehouses is commonly modeled using a so-called "Star schema", or one of its variants, i.e., "Snowflake schema" or "Fact Constellation".
   a) Name and explain the different roles / types of tables and their attributes in these schemas.
   b) While different, "Snowflake schema" and "Fact Constellation" share a common intension how to improve over plain "Star schema", though in different scenarios. What is this common intension and how do both variants realize it differently?
   c) Consider a company's data warehouse that consists of the five dimensions *product, product_material, time, supplier, location,* and the three measures *total_costs, average_costs, count*. Draw a "Star schema" or a "Snowflake schema" for this data warehouse (you will have to invent your own attributes). Explain your choice.

4. In data cubes, we distinguish three categories of measurement aggregation. Name and describe all three categories, and give at least one example for each category.

5. Name and explain four typical OLAP operations. Where possible, relate them to an equivalent (combination of) relational algebra operations.

6. Describe how a succinct constraint can be exploited in the FP-Growth Algorithm. Give two examples of a succinct constraint.

7. Assume a database *DB* containing population survey data is given with items/events {A, B, C, …}. (Note: *A* can be something like *"has a pet", B = "walks more than 5 km a day", C = "has a company", etc.*) After data mining the database the pattern *A* => *B* is found.

   a) Give two examples of null-invariant interestingness measures for the correlation/dependence of the two items/events *A* and *B*.
   b) What values of your measures would you expect, if the items/events *A* and *B* are negatively correlated/dependent in *DB*?
   c) What values of your measures would you expect, if the item sets in *DB* are imbalanced and positive with respect to items/events A and B?
   d) Is it important to use null-invariant interestingness measures for mined rules for this particular kind of database?

8. A database has five transactions. Let min_sup = 50%.

TID	*Items_bought*
T100	{A,D,E}
T200	{A,C,D,E,F}
T300	{B,C,F,G}
T400	{B,C,F,G,H}
T500	{A,D,E,F,I}

   a. Find all frequent item-sets using the Apriori-Algorithm (give the intermediate results for the different steps of the algorithm).
   b. Assume you will need to find the frequent item-sets of 10^6 transactions with a given minimum support, each transaction containing the items bought from a set of 10^3 items. Would the Apriori-Algorithm be useful for calculating these frequent item sets? Explain your answer.

9. An optimization in frequent item set mining is mining closed patterns.

   a) Describe why mining closed patterns can be done more efficiently than mining frequent item sets.
   b) Would the FP-Growth method be useful for mining closed patterns in a 10000-dimensional data set of 10^7 elements?
   c) Which method would be suitable to mine very long frequent patterns in a 10000-dimensional data set of 10^7 elements?

10. Let a graph *G = (V, E)* be defined by *V* the set of vertices being equal to the set of Facebook members, and *E* the set of edges between vertices of *V*, where an edge *(v, w)* is in *E*, if and only if the Facebook member represented by vertex *v* is Facebook-friend with the Facebook member represented by vertex *w*.
   a) Give 3 important graph theoretic characteristics of graph *G*.
   b) Give 2 other examples of natural occurring networks that also have these characteristics.
   c) Describe an algorithm that produces a synthetic graph *H* with at least 2 similar graph theoretic characteristics as *G*.