Birkbeck
(University of London)

MSc and MRes Examination for Internal Students

Department of Computer Science and Information Systems

Data Warehousing and Data Mining (COIY026H7)

Credit Value: 15 credits

Date of examination: Tuesday 03 June 2014
Duration of paper: 14.30-16.30

There are six questions on this paper. Candidates should attempt any FOUR of them. Use of electronic calculators is not permitted.

The paper is not prior-disclosed.

1. "Typically, a data warehouse is maintained separately from an organization's operational databases. There are many reasons for doing this."

   (a) Explain what the requirements are for a data warehouse and why these do typically lead to a data warehouse being maintained separately from operational databases.

   (9 marks)

   (b) In maintaining separation of operational data, some warehouse architectures use a materialized approach whereas others use a virtual approach. Explain what is meant by these terms and the advantages and disadvantages of each approach.

   (8 marks)

   (c) While relational database management system implementations have traditionally been row-oriented, there has been increasing interest in column-oriented architectures often referred to as *column stores*. Explain how column stores differ from traditional row-oriented architectures and why they may offer advantages for data warehousing applications.

   (8 marks)

   [Total: 25 marks]

2. Many data warehouse applications load, analyse and finally remove data from a warehouse on the basis of time periods for which the data is of interest.

   What issues would you consider, and what data warehouse capabilities and techniques would you exploit in the design of a data warehouse system in order to support loading, analysis and removal of time-dependent data as efficiently as possible?

   Your answer should cover both the

   (a) logical design and (15 marks)

   (b) physical design (10 marks)

   of a warehouse.

   [Total: 25 marks]

3. You are an administrator for an Oracle data warehouse which is implemented through conventional relational tables. Queries against the warehouse are running too slowly and it has been suggested that *materialized views* and *dimensions* may help performance.

(a) Briefly explain what relational DBMS materialized views and dimensions are. Give an example of a materialized view and a dimension which illustrates the features which may be defined with each. (Your examples should make clear what may be defined with materialized views and dimensions, but precise SQL syntax is not required.)

(10 marks)

(b) Explain in what ways query performance may be improved through the use of materialized views and dimensions. Illustrate your answer with the types of SQL queries which would benefit from materialized views and dimensions.

(8 marks)

(c) A particular issue when refreshing warehouse data is the basis on which materialized views are refreshed. Explain the different ways a materialized view may be refreshed in a system such as Oracle and the performance implications of each.

(7 marks)

[Total: 25 marks]

4. A company sells products in its chain of stores. Information recorded about products includes an identifying product id, and the product's name and class. Information recorded about suppliers of products includes an identifying supplier id, and the supplier's name, status and location (street address, postcode, city and country). Information recorded about stores includes an identifying store id, and the store type and location (street address, postcode, city and country). Information related to the sale of products is also recorded, including the date of the sale, the product, supplier and store involved, as well as information about the sale such as the number of products sold (unit_sales) and the amount paid by the customer (store_sales).

   A multidimensional model is often used as a conceptual model in which such data can be visualized as being stored in an n-dimensional OLAP cube.

   (a) Common operations on an OLAP cube are *slice*, *dice*, *roll-up*, *drill-down* and *pivot*. Explain what the effect of each of these operations is, giving an example in each case for the company sales data described above.

   (10 marks)

   (b) Data such as unit_sales and store_sales are examples of *measures* which may be classified as *additive*, *semi-additive* or *non-additive*. Briefly explain the difference between these three types of measure.

   (6 marks)

   (c) SQL:1999 introduced the *ROLLUP* and *CUBE* extensions to GROUP BY. Explain what extra capabilities each extension supports, illustrating your answer

with an example query in each case for a suitable relational table representation of the company sales data.

(9 marks)

[Total: 25 marks]

5. A TV broadcaster uses a relational database to record the numbers of people watching its programmes.

The database includes the following tables.

PROG_DETAILS (P_NO, P_NAME, P_CATEGORY)
AUDIENCE_LOG (L_NO, P_NO, YEAR, MONTH, DAY, AUDIENCE)

The PROG_DETAILS table has a row for each programme broadcast with an identifying number, programme name and category, for example (854, 'Casualty', 'Drama'). You may assume that all programmes are in long-running series and each episode of a programme within a series has a separate identifying number but the same name.

A row in the AUDIENCE_LOG table records the number of viewers (in millions) watching the broadcast of a programme. For example row (18496, 854, 2014, 2, 22, 6.1) is a log entry with identifying number 18496 recording that the episode of Casualty broadcast on 22 February 2014 had 6.1 million viewers.

(a) SQL:1999 introduced the *WITH* clause which can be useful in OLAP queries. Explain what the advantages are of using this clause, and give an example SQL query illustrating its use.

(5 marks)

(b) Write down SQL queries using SQL/OLAP features to answer the following queries.

(i) Find the 10 programmes broadcast on 10 April 2014 with the highest number of viewers. In each case give the programme name, category, and number of viewers.

(ii) Considering the episode of Casualty broadcast on 1 March 2014 and the previous 7 episodes, find the total number of viewers of those 8 episodes.

(iii) Find the average number of viewers of episodes of Casualty broadcast in each of the 2 month periods January/February and February/March 2014. In each case give the last month of the 2 month period and the average number of viewers of episodes of Casualty broadcast in that period.

(iv) For each programme category, find the average number of viewers of programmes in that category in the 3 months ending October, November and December 2013. In each case, give the programme category, the last month

of the 3 month period and the average number of viewers of programmes of that category in that period.

(20 marks)

[Total: 25 marks]

6. The *A-Priori* algorithm is one method used in *association rule mining*.

   (a) Explain what association rule mining is, and how *confidence* and *support* measures are used to identify interesting rules.

   (5 marks)

   (b) Write down the steps of the A-Priori algorithm and explain how it is used in association rule mining. Illustrate your answer by showing what is produced by each step of the algorithm with the following set of transactions, assuming that frequent itemsets with minimum support of 25% are required.

   $T1$ $\{I2, I4\}$
   $T2$ $\{I1, I4, I5\}$
   $T3$ $\{I1, I2, I3, I4, I5\}$
   $T4$ $\{I3, I4, I5\}$
   $T5$ $\{I1, I3\}$
   $T6$ $\{I3, I5\}$
   $T7$ $\{I1, I3, I4, I5\}$
   $T8$ $\{I1, I5\}$

   (10 marks)

   (c) Briefly explain how the basic approach of finding frequent itemsets can be extended to identify rules which might otherwise not be identified as interesting.

   (5 marks)

   (d) The use of confidence and support measures may lead to rules of limited use being identified. Explain in what circumstances this can happen and how the *lift* measure improves on confidence and support in such circumstances.

   (5 marks)

[Total: 25 marks]

