Birkbeck
(University of London)

MSc and MRes Examination for Internal Students
MSc in Advanced Information Systems
MSc in Intelligent Technologies
MSc in Information and Web Technologies
MRes in Computer Science

Department of Computer Science and Information Systems

Data Warehousing and Data Mining (COIY026H7)

Credit Value: 15 credits

Date of examination: Wednesday 8 June 2011
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. A conventional database system typically supports on-line transaction processing (OLTP) whereas a data warehouse supports on-line analytical processing (OLAP).

   (a) Explain in what ways the functional and performance requirements of OLTP applications differ from those of OLAP applications, and why a typical OLTP database system may be inappropriate to support the needs of OLAP applications.

   (8 marks)

   (b) Data in a data warehouse often reflects a multidimensional model supporting *aggregations* of *measures* depending on *dimensions* described by *attributes*. Explain what is meant by these terms, and the ways in which they may be represented if a relational model is used to implement a warehouse schema.

   (9 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 companies maintain data warehouses with a relational DBMS such as Oracle which they wish to use to analyse data such as sales receipts by different criteria. For example, they may wish to analyse by reference to individual sales outlets or the cities, regions or countries in which those outlets are situated. Equally, they may wish to analyse such sales data by reference to time slots which may vary from individual hours through to years or longer. Materialized views and dimensions may be exploited by a query optimizer to give improved query performance accessing such data warehouses.

   (a) Briefly explain what relational DBMS *materialized views* and *dimensions* are.

   (4 marks)

   (b) Explain how you would determine which materialized views and dimensions would be beneficial for the sort of data warehouse described above. Give an example of a materialized view and a dimension which illustrates the features which may be defined with each and explain what those features are. (Your examples should make clear what may be defined with materialized views and dimensions, but precise SQL syntax is not required.)

   (12 marks)

   (c) Explain in what ways a query optimizer may exploit materialized views and dimensions to give improved query performance, illustrating your answer with the types of SQL queries which would benefit from the example materialized view and dimension you have given in part (b).

   (9 marks)

   [Total: 25 marks]

3. "*ETL* process operations are the most complex and technically challenging among all the data warehouse process phases".

   (a) Briefly describe which phases of data warehousing ETL refers to.

   (5 marks)

(b) For each of these phases, describe the difficulties that must be overcome and the techniques used in large data warehouse systems to achieve this.

(16 marks)

(c) Briefly explain why keeping track of metadata is fundamental to supporting ETL phases and give examples of metadata which could be used in supporting the techniques you have described in (b).

(4 marks)

[Total: 25 marks]

4. A company runs a telephone helpline which customers of its products can ring for help with use of the company's products. A relational database is used to log information about calls received. The tables include:

PRODUCT (PROD_ID, PROD_NAME, PROD_CLASS)
PROBLEM (PRBLM_ID, PRBLM_DESC, PRBLM_CLASS)
SOLUTION (SOL_ID, SOL_DESC, SOL_CLASS)
TIME (TIME_ID, YEAR, WEEK, DAY)
CALL_LOG (TEL_NO, TIME_ID, PROD_ID, PRBLM_ID, SOL_ID, CALL_DURATION)

The PRODUCT table records a product id for each product together with the name and class of the product.

Each problem which can occur and each solution which may be suggested is identified by a unique PRBLM_ID and SOL_ID respectively. The PROBLEM and SOLUTION tables record these together with a description and class for each possible problem and solution.

The TIME table associates an identifier with each day of each week of each year, for example (873, 2011, 4, 3) associates an identifier 873 with day 3 of week 4 of 2011.

Rows in the CALL_LOG table record for each call received, the telephone number of the caller, as well as identifiers for the time of the call, the product involved, the problem with that product, the solution given, as well as the duration in minutes of the telephone call.

(a) Write down SQL statements to answer the following queries.

(i) For each identified problem, find the total number of calls relating to that problem in week 9 of 2011. In each case give the problem id, problem description, and total number of calls about that problem.

(ii) Find the products in respect of which there were calls to the helpline totalling more than 300 minutes during at least one week in 2010. For each such product and week give the product id, product name, week number, and the total duration of calls relating to that product in that week.

(iii) Find any class of problem in respect of which more than 5% of the total number of calls were received in 2010. In each case give the problem class, total number of calls received in 2010 and average duration of each call.

(15 marks)

(b) Data such as that held in respect of helpline calls can be considered as an OLAP cube with *slice*, *dice*, *roll-up* and *drill-down* operations being possible on the cube. Explain what the effect of each of these operations is, giving an example in each case for the helpline data. Which of these operations did SQL:1999 introduce direct support for? Briefly explain how that support is provided in SQL:1999.

(10 marks)

[Total: 25 marks]

5. (a) There are three main architectures for OLAP systems depending on the underlying database technnology: ROLAP, MOLAP and HOLAP. Briefly explain the characteristics of each of these noting any particular strengths or weaknesses.

(5 marks)

(b) A market research company interviews members of the public on a regular basis regarding their level of support for different policy areas of political parties. The results are held in a summarised form in a relational database which includes the following table.

SURVEY_RESULTS (YEAR, MONTH, DAY, POLICY_AREA, PARTY, SUPPORT)

The table records the interview date year, month and day, the policy area (e.g crime), the party (e.g. lab) and the level of support which is a number between 0 (low) and 100 (high) reflecting the overall level of support for the party's policy amongst all people interviewed on that day. For example, the row (2011, 4, 28, 'crime', 'lab', 78) records a level of support of 78 for the Labour party's policy on crime amongst people interviewed on 28 April 2011.

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

(i) Find the 5 policy areas of the Labour party which received the highest level of support amongst people inverviewed on 28 April 2011. In each case give the policy area and level of support. Explain what your query would produce in a situation where more than one policy receives the same high level of support.

(ii) Find the 3 parties which got the highest average level of support for their policies on crime amongst people interviewed in April 2011. In each case give the party, and average level of support for their crime policy in April 2011.

(iii) Find the average level support for the Labour party policy on crime in interviews carried out in the 3 month period January to March 2011, and also the period February to April 2011.

(iv) Find the average level of support for each party's policy area in the 3 month period January to March 2011, and also the period February to April 2011.

(20 marks)

[Total: 25 marks]

6. "*Data mining* is the application of specific algorithms for extracting patterns from data."

(a) OLAP methods provide tools for extracting data from warehouses. Briefly explain why OLAP methods are inadequate for more generally extracting patterns from data as envisaged by the quote.

(2 marks)

(b) Data mining is normally only one step in a wider *knowledge discovery* process. Briefly explain what steps are undertaken in this wider process.

(4 marks)

(c) The *A-Priori* algorithm is one method used in *association rule mining*. Explain what association rule mining is, and how *confidence*, *support* and *lift* measures are used to identify interesting rules.

(7 marks)

(d) 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.

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

(12 marks)

[Total: 25 marks]