Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

Data Warehousing and Data Mining (COIY026H7)

CREDIT VALUE: 15 credits

Date of examination: Wednesday 07 June 2017
Duration of paper: (10.00-12.00)

There are **six** questions on this paper.
Answer only **four** of the six questions.
Each question carries **25** marks in total.
If you answer more than four questions, only the best four answers will count.
Candidates must **NOT** bring in any supplementary material into the examination.
This paper is not prior-disclosed.
Use of electronic calculators is not permitted.

1. "The architectures of OLAP data warehouses differ significantly from those of database systems supporting OLTP"

   (a) Briefly describe the requirements for a data warehouse architecture.

   (**5** marks)

   (b) Many data warehouses follow a *two-layer architecture*. Explain what is separated in a two-level architecture and the advantages of this separation.

   (**7** marks)

   (c) A *three-layer architecture* for data warehouses is also commonly found. Explain how this differs from a two-level architecture.

   (**6** marks)

   (d) *Data marts* are often used in both two and three layer architectures. Explain what a data mart is, the two main approaches to building a data mart and the advantages of each.

   (**7** marks)

   [Total: **25** marks]

2. (a) In a relational database implementation of a data warehouse, a *star* schema might be used.

   Explain what a star schema is illustrating your answer with an example of a star schema appropriate for a data warehouse used by a College to store information relating to the examination entries and results of students on module exams over a number of years as outlined below.

   Information is stored about each student including their status e.g. part-time, and fee liability e.g. overseas. Each student is enrolled on a particular programme e.g. MSc Computer Science and additional information is stored about each programme such as the programme director.

   Information is stored about each module including the module leader and the module level e.g. postgraduate level 7. Module results are considered by a particular exam board, and additional information is stored about each exam board such as the chair and external examiners.

   For each student's module examination entry, the information stored includes the year of the exam, the exam entry date, and the mark out of 100 that the student got on that exam.

   (**8** marks)

   (b) Alternatives to a star schema include a *snowflake* schema or *fact constellation*. Explain how these differ from a star schema illustrating your answer with an example in each case of how they might be used for the examination system described in (a) above.

   (**6** marks)

(c) One decision to be made in a relational database implementation of a data warehouse is how to model *slowly changing dimensions*. Describe the possible solutions to this issue using a slowly changing dimension appropriate for your example star schema in answer to (a).

(**6** marks)

(d) *Slowly changing measures* may also need to be modeled in a star schema. Explain how a slowly changing measure may arise and the possible ways one can be modeled.

(**5** marks)

[Total: **25** marks]

3. A company uses a relational database implementation of a warehouse recording information about sales of products in its stores. The tables include:

STORE (STORE_ID, POSTCODE, CITY, REGION)
PRODUCT (PRODUCT_ID, PRODUCT_NAME, PRODUCT_TYPE)
TIME (TIME_ID, YEAR, WEEK, DAY)
STORE_PRODUCT_SALES (TIME_ID, STORE_ID, PRODUCT_ID,
STORE_SALES, STORE_COST, UNIT_SALES)

The STORE_PRODUCT_SALES table records each sale of a product in a store. STORE_SALES holds the store's sale receipt for the sale, STORE_COST holds the cost to the company, and UNIT_SALES holds the number of units sold.

A type of query which is run frequently analyses receipts and costs for sales of different products of particular types in stores in particular cities or regions. This type of query is running too slowly and it has been suggested that *materialized views* and *bitmap indexes* may help performance.

(a) Briefly explain what materialized views and bitmap indexes are.

(**6** marks)

(b) Explain how materialized views may be used to support efficiently queries of the type described in the question. Illustrate your answer with a materialized view that you would propose defining for the tables shown. (Your illustration should make clear what may be defined with a materialized view, but precise SQL syntax is not required.)

(**6** marks)

(c) Explain how bitmap indexes may be used to support efficiently queries of the type described in the question. State, with brief reasons, what bitmap indexes you would propose defining for the tables shown.

(**6** marks)

(d) Bitmap indexes may be used in a technique known as *star transformation* or *star join optimization*. Explain what bitmap indexes are needed and how they are used in this optimization technique. State, with brief reasons, which of the bitmap indexes if any in your answer to (c) could be used in a star transformation.

(**7** marks)

[Total: **25** marks]

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

(**7** marks)

(b) Consider the tables of the warehouse described in Question 3:

STORE (STORE_ID, POSTCODE, CITY, REGION)
PRODUCT (PRODUCT_ID, PRODUCT_NAME, PRODUCT_TYPE)
TIME (TIME_ID, YEAR, WEEK, DAY)
STORE_PRODUCT_SALES (TIME_ID, STORE_ID, PRODUCT_ID,
STORE_SALES, STORE_COST, UNIT_SALES)

Explain how the OLAP operations of *slice*, *dice* and *roll-up* are achieved using SQL in a ROLAP system. Explain what the effect of each of these operations is, giving an example SQL query in each case for the sales data above.

(**9** 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 SQL query in each case for the sales data above.

(**9** marks)

[Total: **25** marks]

5. (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) A database recording air pollution readings hourly each day at various sites in central London over a period of time includes the following table.

POLLUTION_READINGS (SITE_ID, SITE_AREA, SITE_LOCATION, YEAR,
MONTH, DAY, HOUR, POLLUTANT, READING)

For example, the row (141, 'Camden', 'Euston Road', 2017, 4, 27, 14, 'NO2', 170) records that at site 141 in Camden located at Euston Road, the reading on 27 April 2017 at 14.00 recorded the NO2 level as 170.

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

(i) Find the 5 highest NO2 readings recorded at the Camden Euston Road site in April 2017. In each case give the reading and the day and hour it was recorded. You may assume that there are 5 unique highest readings, but explain what your query would produce in a situation where the same high reading occurred at the site on more than one occasion in April 2017.

(**5** marks)

(ii) Find the 5 sites in Camden with the highest average NO2 reading in April 2017. In each case give the site id, site location, and the average NO2 reading recorded.

(**5** marks)

(iii) Find the average NO2 reading at the Camden Euston Road site starting with the reading at midday on 1 May 2017 and averaged over that and the next 100 readings at the site.

(**5** marks)

(iv) Find the average N02 reading at each site in Camden in each of the three month periods January-March, February-April and March-May in 2017. In each case give the site id, location, final month number (e.g. 3 for March) and average NO2 reading at the site in the three month period ending that month.

(**5** 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 defined and used to identify interesting rules. (**8** 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 explaining what is produced by each iteration of the steps of the algorithm with the following set of transactions, assuming that frequent itemsets with minimum support of 25% are required.

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

(**10** marks)

(c) One approach supporting a more efficient way of identifying frequent itemsets is to use a *FP-tree*. Explain why the FP-tree approach is more efficient than A-Priori, what a FP-tree is, and how it is built given a set of transactions. (You do not need to draw out the tree structure which results.) (**7** marks)

[Total: **25** marks]

