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: Thursday 09 June 2016
Duration of paper: 14.30-16.30

INSTRUCTIONS

There are six questions on this paper.

Candidates should attempt any FOUR questions.

If you answer more than four questions, only the best four answers will count.

All questions carry 25 marks.

Candidates must NOT bring in any supplementary material into the examination.

Use of electronic calculators is not permitted.

This 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) A multidimensional model is often adopted as the conceptual model for a data warehouse. Describe the features of the multidimensional model and explain how the model helps support requirements of OLAP applications. (9 marks)

   (c) Explain how a multidimensional model is typically represented in the logical model of a relational database implementation of a data warehouse. (8 marks)

   [Total: 25 marks]

2. (a) 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)

   (b) Storage structures have also been developed to scale to very large amounts of data distributed on commodity hardware such as *HDFS (Hadoop Distributed File System)*. Describe the architecture of HDFS and how reads and writes to files are supported. (8 marks)

   (c) One technique used to allow more efficient access in warehouse architectures built on distributed file systems is use of a *Bloom Filter*. Explain the features of a Bloom Filter and how it supports efficient warehouse access. (9 marks)

   [Total: 25 marks]

3. (a) Briefly describe the three tasks - commonly referred to as ETL - that are involved in the construction of a data warehouse. (6 marks)

(b) For the two tasks represented by E and L in ETL, describe the difficulties in supporting these tasks that must be overcome in large data warehouse systems and the techniques used to achieve this.

(14 marks)

(c) A more scalable solution to the ETL tasks required to support data being streamed from operational data sources is to model ETL as a workflow of processes. Briefly explain how this provides a basis to support more efficient ETL tasks.

(5 marks)

[Total: 25 marks]

4. (a) One of the main architectures for OLAP systems is ROLAP. Briefly explain the characteristics of ROLAP noting any particular strengths or weaknesses.

(4 marks)

(b) Consider a ROLAP implementation which includes tables

PRODUCT (PRODUCT_ID, PRODUCT_NAME, PRODUCT_TYPE)
PROMOTION (PROMOTION_ID, PROMOTION_NAME,
PROMOTION_TYPE)
CUSTOMER (CUSTOMER_ID, CUSTOMER_NAME, CUSTOMER_CITY)
STORE_SALES (TRANSACTION_ID, PRODUCT_ID, PROMOTION_ID,
CUSTOMER_ID, QUANTITY, UNIT_PRICE)

recording sales transactions of products bought by customers when promotions are offered. For example, (15621, 9382, 144, 7734, 3, 4.95) in table STORE_SALES records that in transaction 15621, product 9382 on sale with promotion 144 was bought by customer 7734 who bought 3 at 4.95 each.

Common operations in OLAP systems include *slice*, *dice* and *roll-up*. Explain what the effect of each of these operations is, giving an example SQL query in each case for the sales data described above.

(9 marks)

(c) SQL:1999 introduced the *ROLLUP*, *CUBE* and *GROUPING SETS* 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 described above.

(12 marks)

[Total: 25 marks]

5. (a) SQL/OLAP introduced new functions useful in analytical queries including *RANK*, *DENSE_RANK* and *NTILE* functions. Explain what each function produces and how their results differ.

(5 marks)

(b) A relational database holds ratings made by reviewers eating meals at restaurants. The database includes a table:

RESTAURANT_REVIEWS (ID, YEAR, MONTH, DAY, REVIEWER, RESTAURANT, POSTCODE, TYPE, SCORE)

For example, the row (1921, 2016, 5, 12, 814, 'Millepini', 'W1', 'Italian', 9) records that a review with ID 1921 was undertaken by a reviewer with id 814 on 12 May 2016. The meal was at the Millepini restaurant with postcode W1, which serves Italian Food. The reviewer gave the meal a score of 9. You may assume that there is only one restaurant with a particular name in a postcode area.

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

(i) Find the 5 highest scoring meals eaten in Italian restaurants in 2015. In each case give the restaurant name, postcode, the day and month of the meal and the score obtained.

(ii) Find the 5 French restaurants with the highest average score for meals eaten during 2015. In each case give the restaurant name, postcode, and the average score obtained.

(iii) Find the average score obtained by the Golden Prawn restaurant in WC1 for the meal reviewed there on 21 April 2016 and the preceding nine dates on which a meal there was reviewed.

(iv) Find the average score obtained by each Italian restaurant in W1 reviewed in each of the three month periods in 2015 August-October, September-November and October-December. In each case give the restaurant name, postcode, final month number (e.g. 10 for October) and average score in the three month period ending that month.

(20 marks)

[Total: 25 marks]

6. (a) Choosing the *data mining task* is one step in the knowledge discovery process. Explain the purpose of the four different data mining tasks *prediction*, *classification*, *clustering*, and *link analysis (associations)*.

(6 marks)

(b) The *A-Priori* algorithm is one method used in association rule mining. Write down the steps of the A-Priori algorithm and describe how it is used to identify

association rules. Explain how *confidence*, *support* and *lift* measures are used to identify interesting rules in particular.

(10 marks)

(c) The principles underpinning the A-Priori algorithm can also be used to optimize a class of OLAP queries known as *Iceberg* queries. Explain what an Iceberg query is and how the principles underpinning the A-Priori algorithm may enable such queries to be efficiently executed.

(4 marks)

(d) Explain why traditional implementations of the A-Priori algorithm are likely to be inefficient for very large data warehouses. Briefly describe approaches which have been suggested to overcome such inefficiencies.

(5 marks)