BIRKBECK
(University of London)

MSc and MRes EXAMINATION

SCHOOL OF BUSINESS, ECONOMICS AND INFORMATICS

DATA WAREHOUSING AND DATA MINING

COIY026H7

15 credits

Friday 05 June 2015
14.30-16.30

INSTRUCTIONS

There are six questions on this paper.

Candidates should attempt any FOUR questions.

No credit will be given for additional questions answered.

All questions carry 25 marks.

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. (a) Describe the functional and performance requirements of a data warehouse and how these differ from those of the workload of a traditional operational database.

(9 marks)

(b) Describe the features of an architecture of a typical data warehouse system and how those features support the requirements you have identified in (a) above.

(8 marks)

(c) Some recent warehouse architectures are based on the *MapReduce* programming model. Briefly explain what is meant by MapReduce and how it has been developed in the Hive project for supporting data warehouse infrastructures. Note the advantages and disadvantages of architectures based on MapReduce compared to a conventional warehouse architecture.

(8 marks)

[Total: 25 marks]

2. (a) Warehouse data is often represented by a *multidimensional* model supporting *aggregations* of *measures* depending on *dimensions* described by *attributes*. Explain what is meant by these terms and how they relate to a *star* schema in a relational database.

(9 marks)

(b) Modeling of dimensions in a star schema involves a number of design decisions. Two of these are whether or not to use *surrogate keys* and how to model *slowly changing dimensions*. Explain the possible design approaches to each of these and the advantages and disadvantages of each approach you describe.

(9 marks)

(c) In practice, a relational data warehouse design is likely to be more complex than a star schema. Explain in what ways a *snowflake* schema and *fact constellation* differ from a star schema, and the reasons which might lead to their use.

(7 marks)

[Total: 25 marks]

3. (a) Explain the features of a bitmap index and why bitmap indexes may be used rather than B-tree indexes to support data warehouse applications.

(7 marks)

(b) Explain how bitmap indexes can be used to process both single and multiple table SQL queries efficiently. In each case illustrate your answer with an example of a bitmap index which might be appropriate for tables containing information about customers and sales in a data warehouse.

(7 marks)

(c) Briefly explain why *encoded bitmap indexes* may have advantages over simple bitmap indexes. Illustrate your answer by describing a form of encoded bitmap index which can efficiently support queries which search for rows with a column value less than a specified value.

(5 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.

(6 marks)

[Total: 25 marks]

4. (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.

(7 marks)

(b) Consider a ROLAP implementation which includes tables

STORE (STORE_ID, CITY, REGION)
PRODUCT (PRODUCT_ID, PRODUCT_NAME, PRODUCT_TYPE)
STORE_SALES (TRANSACTION_ID, CUSTOMER_ID, STORE_ID,
PRODUCT_ID, QUANTITY, UNIT_PRICE)

recording sales transactions of customers buying products in stores. For example, (10921, 23482, 523, 3294, 4, 1.25) in table STORE_SALES records that in transaction 10921, customer 23482 in store 523 bought 4 of product 3294 at 1.25 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* 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 described above.

(9 marks)

[Total: 25 marks]

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

(5 marks)

(b) A relational database holding election results over a series of elections includes the following table.

VOTING_RESULTS (YEAR, MONTH, DAY, ELECTION_TYPE,
CONSTITUENCY, PARTY, VOTES)

The table records the year, month and day of the election, the type (e.g general or by-election), the constituency name, party and number of votes that party's candidate got in that constituency at that election. For example, the row (2010, 5, 6, 'general', 'witney', 'con', 33973) records that in the 6 May 2010 general election, the Conservative party received 33973 votes in the Witney constituency.

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

(i) Find the 10 largest votes received by candidates at the 7 May 2015 election. In each case give the candidate's party, constituency and number of votes received. You may assume that there are 10 unique largest votes, but explain what your query would produce in a situation where more than one candidate receives a particular"large" vote.

(ii) Find the 5 parties which got the smallest total number of people voting for them at the 7 May 2015 election. In each case give the party and the total number of votes that party received.

(iii) Find the average vote received by the Conservative party in Witney averaged over the 7 May 2015 and preceding 3 general elections. You may assume that the Conservative party has candidates in Witney at every election.

(iv) Find the average vote received by each party in Witney averaged over the 7 May 2015 and the preceding 3 general elections. In each case give the party and the average number of votes that party received. Explain what your query produces if some parties had candidates in some but not all of those elections.

(20 marks)

[Total: 25 marks]

6. (a) Choosing the data mining task is one step in the knowledge discovery process. Briefly explain the purpose of the *association rule mining* task.

(2 marks)

(b) Before data can be mined, it will need to be transformed into a suitable representation for efficient processing. Briefly describe methods which might be used to transform data into a suitable representation for association rule mining.

(6 marks)

(c) 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.

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

(10 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.

(7 marks)

[Total: 25 marks]