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 12 June 2012
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) Describe the architecture of a typical data warehouse system, explaining the function of each component of the architecture.

(9 marks)

(b) 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) Some recent warehouse architectures are based on the *MapReduce* programming model. Explain what is meant by MapReduce and how it has been developed for analysing large data sets and 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.

(8 marks)

(b) Modeling of dimensions in a relational database 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) Some relational database management software such as Oracle supports an SQL *CREATE DIMENSION* statement. Explain what properties of a dimension are defined with this statement, and how creating a dimension in this way may give performance advantages.

(8 marks)

[Total: 25 marks]

3. Large data warehouses often have index structures to support efficient access to data.

(a) Explain the features of a bitmap index and how such indexes may be used to support both single table queries and multiple table join queries efficiently.

(8 marks)

(b) Explain why bitmap indexes may be used rather than B-tree indexes to support data warehouse applications.

(4 marks)

(c) *Encoded bitmap indexes* may have advantages over simple bitmap indexes. Describe two forms of encoded bitmap indexes, and explain the advantages of each.

(7 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 chain of supermarkets uses a relational database to record sales information in respect of 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,
QUANTITY, UNIT_PRICE)

Each store within the chain is identified by a unique STORE_ID. Each product sold by the chain is identified by a unique PRODUCT_ID.

The STORE table has a row for each of the company's stores recording the postcode, city and region the store is in, for example (82, 'SE26 4DE', 'London', 'South_East'). The PRODUCT table records the id, name and type of each product, for example (14523, 'Sardines in Tomato Sauce 125g', 'Grocery').

The TIME table associates an identifier with each day of each week of each year, for example (999, 2012, 6, 1) associates an identifier 999 with day 1 of week 6 of 2012. Rows in the STORE_PRODUCT_SALES table record on a daily basis how many of each product is sold in each store at what unit price. For example, (999, 82, 14523, 46, 0.85) records that on the day with identifier 999, store 82 sold 46 of product 14523 at 0.85 each.

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

(i) For each product, find the total number sold in the stores of each region in week 5 of 2012. In each case give the product id, region and total number of that product sold in that region.

(ii) Find the products of type Dairy which had total sales receipts in excess of 40000.00 in at least one week in 2011. For each such product and week give the product id, product name, the week, and the total sales receipts for that product in that week.

(iii) For each product type with more than 20 products, find any product generating more than 20% of that product type's total sales receipts in 2011. For each such product give the product id, product name and total sales receipts for that product in 2011.

(15 marks)

(b) 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 the store sales database.

(10 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) Data relating to the punctuality of train operating company rail services on different routes is held in a relational database which includes the following table.

PUNCTUALITY_RESULTS (YEAR, MONTH, DAY, OPERATOR,
ROUTE, PERCENT_ON_TIME)

The table records for each day the percentage of trains arriving on time for each train operator's route. For example, the row (2012, 3, 28, 'Sussex Connect', 'London:Hastings', 92.5) records that on 28 March 2012, 92.5% of operator Sussex Connect's rail services from London to Hastings arrived on time.

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

(i) Find the 3 routes of operator Sussex Connect which had the lowest percentage of trains on time on 3 May 2012. In each case give the route and percentage of Sussex Connect's trains on time that day. Explain what your query would produce in a situation where more than one route had the same low level of punctuality.

(ii) Find the 5 train operators which had the highest average percentage of trains on time across all their routes in February 2012. In each case give the train operator, and average percentage of trains on time in February 2012.

(iii) Find the average percentage of operator Hampshire Trains' services between London and Portsmouth which were on time in the 3 month period September to November 2011, and also the period October to December 2011. In each case give the final month of the period and the average percentage of trains on time in that period.

(iv) Find the average percentage of each train operator's trains which were on time in the 3 month periods August to October 2011, September to November 2011, and October to December 2011. In each case give the operator, final month of the period, and the average percentage of trains of that operator on time in that period.

(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)

[Total: 25 marks]

