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: Wednesday 5 June 2013
Duration of paper: 10.00-12.00

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. "It is well-known that data warehouses are focused on decision support rather than on transaction support, and that they are prevalently characterized by an OLAP workload."

   (a) Describe the functional and performance requirements of an OLAP workload and how these differ from those of the workload of a traditional OLTP database system.

   (8 marks)

   (b) 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)

   (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.

   (8 marks)

   [Total: 25 marks]

2. In a relational data warehouse, the basic unit of access at the logical level is that of a table consisting of rows. However, at the physical level, additional or alternative low-level storage structures may be supported, for example *partitioned tables* or *column stores.*

   (a) Explain how a partitioned table differs from a non-partitioned table, and the different bases on which partitioning can be made.

   (7 marks)

   (b) Explain in what circumstances a partitioned table may give improved performance or other advantages over a non-partioned table.

   (7 marks)

   (c) Explain how column stores differ from traditional row-oriented architectures and why they may offer advantages for data warehousing applications.

   (8 marks)

   (d) Briefly note any circumstances in which partitioned tables or column stores may give rise to reduced performance of OLTP applications.

   (3 marks)

   [Total: 25 marks]

3. "The construction of a data warehouse is primarily a process of *information integration.*"

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

   (5 marks)

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

   (16 marks)

   (c) Conventional data warehouse construction and maintenance techniques may not have the flexibiliy and performance to satisfy the needs of applications which rely on the daily loading of very large amounts of data. Some recent warehouse architectures are based on the *MapReduce* programming model. Briefly explain what is meant by MapReduce and why architectures based on MapReduce may have advantages compared to a conventional warehouse architecture.

   (4 marks)

[Total: 25 marks]

4. A college department uses a relational database to record the exam results of its students on taught modules. The tables include:

   STUDENT_MODULE (S_ID, S_NAME, M_ID, M_NAME)
   MODULE_QUESTION_MARK (S_ID, M_ID, M_QNO, M_QMARK)
   MODULE_TOTAL_MARK (S_ID, M_ID, M_TMARK)

   Students are uniqely identified by a student id S_ID value while modules are uniquely identified by a module id M_ID value.

   The STUDENT_MODULE table has a row recording the modules enrolled for by each student. As well as the identifying S_ID and M_ID values, a row records the name of the student S_NAME, and the name of the module M_NAME, for example (10293, 'Adam Ant', 'CS102', 'Java Programming')

   The MODULE_QUESTION_MARK table has a row recording the mark obtained by a student on each question of the module exam. No row is stored if a student does not attempt a particular question. For example, (10293, 'CS102', 3, 18) records that student 10293 got 18 marks for question 3 of the exam for the module with id CS102.

   The MODULE_TOTAL_MARK table has a row recording the total mark obtained by a student on a module exam. For example, (10293, 'CS102', 74) records that student 10293 got 74 marks in total on the exam for the module with id CS102.

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

(i) For each question on the Java Programming module exam, find the average mark obtained by students on that question. In each case give the question number and the average mark obtained.

(ii) Find each module which had less than 10 students sitting the module exam. In each case, give the module id, module name and the number of students sitting that exam.

(iii) Find each module with an average module total mark obtained by students which is more than 10 higher than the average module total mark obtained by students on modules overall. In each case, give the module id, module name, and average total mark for that module.

(15 marks)

(b) SQL:1999 introduced the *ROLLUP* extension to GROUP BY. Explain what extra capabilities the *ROLLUP* extension supports. Illustrate your answer with a query providing a more detailed analysis of the module exam results. Briefly explain how the *CUBE* extension to GROUP BY differs from *ROLLUP*.

(10 marks)

[Total: 25 marks]

5. A database recording rainfall readings each day at various sites in geographical regions over a period of time includes the following table.

RAINFALL_READINGS (SITE_ID, SITE_REGION, YEAR, MONTH, DAY, RAINFALL)

For example, the row (734, 'East Anglia', 2013, 1, 3, 2) records that at site 734 in the East Anglia region on 3 January 2013 rainfall of 2cm was recorded.

(a) SQL/OLAP introduced new features to support more powerful capabilities for analysing data such as in RAINFALL_READINGS. Explain what extra capabilities are supported by SQL/OLAP *RANK* and *WINDOW* features.

(5 marks)

(b) Write down SQL queries using these SQL/OLAP features to answer the following queries.

(i) Find the 5 days in January 2013 with the highest rainfall readings at site 734 - the same rainfall reading may appear more than once. In each case give the day and rainfall recorded. Explain what your query would produce in the case that there are more than 5 days with the same heavy rainfall.

(ii) Find the 3 months in 2012 with the highest average daily rainfall recorded at sites in East Anglia during that month. In each case give the month and average daily rainfall recorded that month at sites in East Anglia.

(iii) For each month of 2012, find the total rainfall from the start of the year to the end of that month at site 734.

(iv) Find the total rainfall recorded at each site in East Anglia in the 3 months June to August 2012 as well as the 3 months July to September 2012. In each case give the site id, final month of the 3 month period (i.e. 8 or 9) and the total rainfall recorded in that period at that site.

(20 marks)

[Total: 25 marks]

6. Data mining is normally only one step in a wider *knowledge discovery* process.

(a) Briefly explain what steps compose the knowledge discovery process.

(4 marks)

(b) *Classification* is a common data mining task. Explain what the purpose of this task is, and how *decision trees* are constructed and used as a classification method.

(7 marks)

(c) *Clustering* is another common data mining task. Explain what the purpose of clustering is, and how a clustering algorithm which is hierarchical differs from one which is partitional. Illustrate your answer by writing down the steps of the *k-means* clustering algorithm and briefly explain whether it is an example of a hierarchical or partitional algorithm. (7 marks)

(d) Explain why traditional data mining algorithms are likely to be inefficient for analysing warehouse data, and describe approaches which have been proposed for increasing the efficiency of such algorithms. Your answer should describe the overall aims of such approaches and give examples of their application to the k-means algorithm, but detailed algorithms are not required.

(7 marks)

[Total: 25 marks]