Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Integration and Analysis

Date: Monday 2\textsuperscript{nd} June 2008
Time: 14:00 – 16:00

Please answer any THREE Questions from the FIVE questions provided

This is a CLOSED book examination

The use of electronic calculators is permitted provided they
are not programmable and do not store text.
1. The following two databases were independently developed. They both record information from the personnel departments of universities that are to merge. In all parts of the question, state any assumptions you make.

**UMIST:**

Employee(EmpId, DoB, EmpName, ContractStartDate, ContractEndDate, JobTitle, Department)
Grade(EmpId, GradeId)
Salary(GradeId, Salary, IncrementDate)

**MANU:**

ContractStaff(EmpId, Name, ContractStartDate, ContractEndDate, JobTitle)
PermanentStaff(EmpId, Name, Dob, JobTitle)
WorksFor(EmpId, Dept)
Salary(EmpId, Salary)

a) Describe 4 examples of schema conflicts between the above schemas. Each of the conflicts should be of a different kind (e.g. you should not describe more than one Table-Attribute conflict) and in each case you should state the kind of conflict involved. (8 marks)

b) Views are often used to bridge across independently developed schemas. They protect users from the schematic differences that may arise as a result of independent development. You are using Oracle MANAGEMENTDB to provide a database that brings together existing UMIST and MANU databases.

i) Give the Oracle distributed database commands that would be required to link the MANAGEMENTDB with the UMIST and MANU databases, and explain why you have chosen public or private, and connected user or fixed user links. (6 marks)

ii) Produce an SQL view in MANAGEMENTDB that derives a table salary(EmpName, Salary) that provides information on how much each employee is paid, and which contains data derived from both the UMIST and MANU databases. (6 marks)
2. You have been asked to design a data warehouse for an international bookstore, which sells books in several countries across the world. Data from individual shops need to be integrated before being stored in the data warehouse to support aggregation of sales data (both as the number of books and sale amount) by regions, countries and genres, in a given fiscal quarter. A book is assigned a maximum of one genre.

   a) Design an appropriate data warehouse schema using the multidimensional model, explaining the fact and dimension tables that you would need to support the above aggregations. Explain and discuss any assumptions that you have made. (9 marks)

   b) In the context of data warehouses, explain briefly the five main tasks of the load manager. (5 marks)

   c) Give two examples of the main challenges of the load manager that you would need for this task and discuss how these could be addressed. (6 marks)

3. Consider the following table (car-sales) containing data on sale amounts for a car dealer:

<table>
<thead>
<tr>
<th>shop-id</th>
<th>make</th>
<th>colour</th>
<th>amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>BMW</td>
<td>blue</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>VW</td>
<td>red</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>VW</td>
<td>white</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>Nissan</td>
<td>white</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>Ford</td>
<td>blue</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>Ford</td>
<td>red</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>BMW</td>
<td>red</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>VW</td>
<td>white</td>
<td>5</td>
</tr>
</tbody>
</table>

   a) In the context of online analytical processing (OLAP), explain roll-up and drill-down operations. (4 marks)

   b) Write an SQL statement that will retrieve the total amount of cars sold, rolled-up along the colour dimension. Calculate the result of the query using the above table. (7 marks)

   c) Calculate the result of the following query using the above table:

   ```sql
   SELECT shop-id, make, count(*), sum(amount)
   FROM car-sales
   GROUP BY CUBE(shop-id, make)
   ```

   (9 marks)
4. a) Association rules describe relationships between two itemsets in the form $A \rightarrow B$. Explain the two measures that are used in the generation of association rules and how they are calculated. Given an example of market based data below, calculate the respective values of the measures for the rule $\{\text{pretzels}\} \rightarrow \{\text{beer}\}$.

<table>
<thead>
<tr>
<th>transaction-id</th>
<th>items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>beer, diapers</td>
</tr>
<tr>
<td>2</td>
<td>pretzels, diapers, beer</td>
</tr>
<tr>
<td>3</td>
<td>soda, pretzels</td>
</tr>
</tbody>
</table>

(6 marks)

b) Some attributes (e.g. numeric values such as salary or price) in a dataset can have a large set of values. Explain why this might be a problem for association rule mining. What are the approaches that can be used to address it? Justify your answer by examples.

(8 marks)

c) Decision trees are used as a classification technique in data mining. Explain the process of representing resulting decision trees as IF-THEN classification rules. Explain the main differences between association rules and classification rules.

(6 marks)

5. a) What is clustering? Describe the two classes of clustering algorithms (with respect to the approach to building clusters).

(5 marks)

b) Briefly explain the main steps in the k-means clustering algorithm. Why can the algorithm terminate in a local optimum?

(7 marks)

c) An example dataset consists of six points whose values represent the total amount sold in six different stores on a given date: A(1), B(5), C(6), D(2), E(3) and F(7). The distance between these points is defined using the Manhattan distance, i.e. the absolute value of the difference between the respective sale amounts (e.g. the distance between points A and B is 4, while the distance between C and F is 1). Cluster these points into two groups using the k-means algorithm, the Manhattan distance, and points A and F as initial clusters.

(8 marks)

END OF EXAMINATION