Tutorial II: Data & Data Warehousing

Prof. M-Tahar Kechadi

School of Computer Science
University College Dublin.

Question 1

Consider a dataset D that consists of customers purchasing tour packages to various places at different prices. We can perform different kinds of data operations on the dataset D. Try to categorise the following operations into a) simple query of data retrieval, b) Online Analytical Processing, or c) Data Mining.

1. Find names of customers who have purchased tours that cost less than €500.
2. List the names of the customers, the number of tour packages that the customers have purchased, and the total cost for the tours.
3. Calculate the difference in quarterly sales of tours between this year and the previous two years.
4. Find a rule such as "**If** customers purchase a tour package to France, **then** it is 80% likely that the same customers also purchase a tour package to Spain.
5. From the customer purchase history, build a model for predicting the kinds of customer who are likely to purchase tours to a certain country.

Question 2 To be Submitted

Consider the dataset given in Table 1, representing the employee data of a company.

1. If the attribute "*Salary*" needs to be discretised into three pay bands, suggest a simple yet sensible solution for the discretisation backed with a valid argument.
2. Miss Davis's salary is unknown and the unknown value needs to be imputed, what is a sensible replacement value and why?
3. Among the employee records, which record can be considered as an outlier? What harm can an outlier cause to the understanding of the dataset?

Online Analytical Processing (OLAP) perceives a dataset in a multi-dimensional space. For the dataset given in Table 1, perform the following tasks/operations:

4. Draw a diagram of a 3D view using the following attributes: `Year of Birth`, `Status`, and `Salary`.

Table 1: Data Set

Emp ID	Name	Year of Birth	Gender	Status	Salary
100	Smith	1964	M	Director	€10000
125	Jones	1977	F	Technician	€1800
167	Davis	1985	F	Senior Technician	
200	O'Bien	1997	M	Technician	€1600
205	Edward	1995	M	Technician	€1700
216	Evans	1995	F	Technician	€1700
220	Moore	1996	F	Senior Technician	€3300
301	Rogers	1965	M	Deputy Director	€8000

5. What do the data points inside the cube represent?
6. Use the cube as an example to discuss the meaning of OLAP operations such as pivoting, slicing and dicing, rolling up and drilling down.

Question 3

Suppose that a data warehouse consists of the three dimensions *time, doctor*,and *patient*, and the two measures *count()* and *charge()*, where charge is the fee that a doctor charges a patient for a visit.

1. Enumerate three classes of schemas that are popularly used for modeling data warehouses.
2. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (1).
3. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010?
4. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (*day, month, year, doctor, hospital, patient, count, charge*).

Question 4 To be submitted

Suppose that a data warehouse for Big University consists of the four dimensions *student, course, semester*, and *instructor*, and two measures *count* and *avg_grade* . At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg_grade()* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

1. Draw a snowflake schema diagram for the data warehouse.
2. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student.
3. If each dimension has five levels (including all), such as student $<$ major $<$ status $<$ university $<$ all, how many cuboids will this cube contain (including the base and apex cuboids)?