

3804ICT/3031ICT/7031ICT - Data Mining

**Workshop / Lab Week 3**

***By Dr Can Wang***

Q1. Briefly compare the following concepts. You may use an example to explain your point(s). (15 mins)

1. Snowflake schema, fact constellation (4 mins)

* Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
* Fact constellation: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

1. Data cleaning, data transformation, refresh. (4 mins)

* Data cleaning: detect errors in the data and rectify them when possible
* Data transformation: convert data from legacy or host format to warehouse format
* Refresh: propagate the updates from the data sources to the warehouse

1. Enterprise warehouse, data mart, virtual warehouse. (7 mins)

Q2. 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. (20 mins)

1. Enumerate three classes of schemas that are popularly used for modeling data warehouses. (2 mins)
2. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a). (10 mins)
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 2017? (8 mins)

Q3. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When 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. (25 mins)

* Draw a snowflake schema diagram for the data warehouse. (12 mins)
* Starting with the base cuboid [student; course; semester; instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student. (10 mins)
* 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)? (3 mins)

Q4. Answering the following questions (20 mins):

* Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube. (2 mins)
* What are the main types of data warehouse usage? And what are the differences between them? (5 mins)
* Given the data in the following table, please draw a Bitmap Index table for column “Course”. (8 mins)

|  |  |  |
| --- | --- | --- |
| StudentID | Course | Grade |
| S1 | 2802ICT | C |
| S2 | 3802ICT | D |
| S3 | 3804ICT | D |
| S4 | 3803ICT | E |
| S5 | 4030ICT | H |

* What is data generalization? Why we need data generalization? (5 mins)

Discussion on results (20 - 30 mins).