

Computer Science COMP3420 (2010) – Tutorial Two

Question 1. A data warehouse can be modeled by either a *star schema* or a *snowflake schema*. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answers.

Answer: The most common modeling paradigm is the *star schema*, in which data warehouse contains (1) a large central table (fact table) containing the bulk of the data without redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension.

The *snowflake schema* is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The major difference between them is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space. However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be carried out to execute a query. Consequently, the system performance may be adversely impacted. Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

Question 2 Suppose a group of 12 *sales price* records has been sorted as follows.

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods.

- (a) equal-frequency (equi-depth) partitioning
- (b) equal-width partitioning.

Can you use the histogram method to allocate the data in the set to 3 equal intervals?

Answer: We allocate the data into 3 bins as follows.

- (a) Equal-frequency partitioning
 - Bin 1: 5, 10, 11, 13
 - Bin 2: 15, 35, 50, 55
 - Bin 3: 72, 92, 204, 215
- (b) Equal-width partitioning: $R_{min} = 5$, $R_{max} = 215$, $N = 3$, and $W = \frac{R_{max} - R_{min}}{N} = 70$. Therefore, the three intervals are $[5, 75]$, $[76, 146]$, $[147, 217]$.
 - Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72
 - Bin 2: 92
 - Bin 3: 204, 215.

The Histogram method is presented as follows. $min = 5$ and $max = 215$, then the width of each bucket is $w = \frac{215-5}{3} \approx 70$, thus, the three intervals are $A_1 = [5, 75]$, $A_2 = [76, 146]$, and $A_3 = [147, 215]$, the corresponding buckets B_1 , B_2 and B_3 contain the elements $\{5, 10, 11, 13, 15, 35, 72\}$, $\{92\}$, and $\{201, 215\}$ respectively.

Question 3. What is the multidimensional data model? What are the OLAP operations?

Answer: The multidimensional data model is a model used to express the data from different combination of dimensions, in which the data are usually stored in datacubes. In multidimensional model, data are organized into multiple dimensions and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides the users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. OLAP provides a user friendly environment for interactive data analysis. The OLAP operations are as follows.

- **Roll-up** performs aggregation on a data cube, either claiming up a concept hierarchy for a dimension or by dimension reduction. e.g `street < city < state < country`.
- **Drill-down** is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. e.g time concept `day < month < quarter < year`. Because a drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube.
- **Slide and dice** performs a selection on one dimension of the given cube, resulting in a subcube.
- **Pivot (rotate)** is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.
- **Other operations:** Drill-across executes queries involving more than one fact table. The drill-through operation used relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables. Top- k listing operation is to list the top k items, etc.

Question 4. What is the three-tier data warehouse architecture? Can you describe the basic components of the architecture and the connection of these components?

Answer: Refer to Fig. 3.12 on page 131 in the textbook (or Lecture 2 slide 27). The three-tier data warehouse architecture consists of the bottom tier (data warehouse server), the middle tier (OLAP server), and the top-tier (front-end tools).

- The bottom tier is almost a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources. These tools and utilities perform data extraction, cleaning, and transformation, as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows clients program to generate SQL code to be executed at a server. This tier also contains a metadata repository, which stores information about the

data warehouse and its contents (source, staging, data integration [ETL/ELT], metadata repository).

- The middle tier is typically implemented using either a relational OLAP model (ROLAP) or a multidimensional OLAP model (MOLAP), that is, a special purpose server directly implements multidimensional data and operations (presentation area).
- The top tier contains query and reporting tools, analysis tools, and/or data mining tools (access).

Question 5. What is the meta-data? Can you list three common types of meta-data?

Answer: Meta-data often is referred to as being the data about data, which defines all aspects of the data contained in a data warehouse including where it originally comes from, its type, what transformations it has been subjected to, where it has been used and what it means from a business perspective.

There are three types of common meta-data:

- Technical meta-data
- Operational meta-data
- Business meta-data.

Question 6. Why most data warehouses stored lots materialized views, what is the benefit by doing so? What are the possible challenges to maintain these materialized views ?

Answer: There are three choices for data cube materialization given a base cuboid.

- No materialization. This leads to computing expensive multidimensional aggregates on the fly, which can be extremely slow.
- Full materialization: Precompute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice requires amounts of memory space in order to store all of the precomputed cuboids.
- Partial materialization: Selectively compute a proper subset of the whole set of possible cuboids. The partial materialization of cuboid or subcubes should consider three factors: (a) identify the subset of cuboids or subcubes to materialize; (b) exploit the materialized cuboids or subcubes during the query processing; (c) efficiently update the materialized cuboids or subcubes during load and refresh.

The proposed method of materialized cuboids is to speed up query processing in data cubes. Given materialized views, query processing should proceed as follows.

1. Determine which operations should be performed on the available cuboids. This involves transforming any selection, projection, roll-up and drill-down operations in the query into corresponding SQL and/or OLAP operations.
2. Determine to which materialized cuboids the relevant operations should be applied. This involves identifying all of the materialized cuboids that may potentially be used to answer the query, pruning the above set using knowledge of “ dominance” relationships among the cuboids, estimating the costs of using the remaining materialized cuboids, and selecting the cuboid with the least cost.

The difficulties of the materialized view maintenance are (1) the data consistency in the materialized view with the remote sources; (2) how long the view maintenance requires; (3) which materialized views should be replaced when a new view is required to be materialized, which optimization metric or metrics will be used (maintenance time, space requirement, and the frequency of such type of query in the future, etc).

Question 7. Robust data loading poses a challenge in database systems because the input data are often dirty. In many cases, an input record may have several missing values and some records could be contaminated. Work out an automated data cleaning and loading algorithm so that the erroneous data will be marked and contaminated data will not be mistakenly inserted into the database during data loading.

Answer: do it by yourself!

Question 8. Use the three methods below to normalize the following group of data: 200, 300, 400, 600, 1000

- min-max normalization by setting $min_{new} = 0$ and $max_{new} = 1$
- z-score normalization
- normalization by decimal scaling.

Answer: (a) min-max normalization, following the formula $v' = \frac{v-min}{max-min} \cdot (max_{new} - min_{new}) + min_{new}$, we have

$$0, 0.125, 0.25, 0.5, 1$$

(b) z-score normalization $v' = \frac{v-\bar{A}}{\sigma_A}$, we have $\bar{A} = 25,00/5 = 500$. $\sigma_A = \frac{1}{N} \sum_{i=1}^N (v_i - \bar{A})^2 = \frac{90,000+40,000+10,000+10,000+250,000}{5} = 80,000$, we then have

$$-\frac{3}{800}, \quad -\frac{1}{400}, \quad -\frac{1}{800}, \quad \frac{1}{800}, \quad \frac{1}{160}$$

(c) normalization by decimal scaling: $v'_i = \frac{v_i}{10^j} \leq 1$, the resulting sequence is

$$0.2(j=3), 0.3(j=3), 0.4(j=3), 0.6(j=3), 1(j=3)$$

Question 9. Both Pearson's product moment coefficient and χ^2 methods are used to determine whether two groups of data are independent, or positively or negatively correlated. Can you point out where they will be used in data preprocessing?

Answer: Data reduction by reducing the number of features (attributes) or data discretization (applied to the data in the same column (a numeric attribute) by reducing the detailed data values into several intervals and using the interval label to represent the values in an interval.