INFS3200 Advanced Database Systems

Tutorial 5: DW Implementation

Semester 1, 2021

Question 1: Consider a data warehouse with *d* dimensions. The fact table *T* contains |*T*| records, and each dimension A_i contains $|A_i|$ distinct values.

(a) Assume that we construct a bitmap index for each dimension. What is the total size (i.e., number of bits) of the bitmap indices?

(b) Below is a sample fact table of the *AllElectronics* data warehouse with two dimensions *item* and *location*, and one measure *sales*. Suppose the dimension *item* at the top level has three values (representing item types): "computer", "phone", and "security", and the dimension *location* has four values (representing cities): "Chicago", "New York", "Toronto", and "Vancouver". Please create bitmap indices for both dimensions.

(c) How can we use the bitmap indices to answer the following queries?

- Find the total sales of each item type.
- Find the total sales of "computer" and "phone" in "New York".

RID	*item*	*location*	*sales*
R1	computer	Chicago	882
R2	computer	New York	968
R3	computer	Toronto	746
R4	computer	Vancouver	825
R5	phone	Chicago	89
R6	phone	New York	38
R7	phone	Toronto	43
R8	phone	Vancouver	14
R9	security	Chicago	623
R10	security	New York	872
R11	security	Toronto	591
R12	security	Vancouver	400

Question 2: Consider a data warehouse with *d* dimensions, and a data cube constructed on all these dimensions $\{A_1, \ldots, A_d\}$.

(a) How many cuboids will be created if the dimensions have no hierarchies, and why?

(b) Suppose that each dimension A_i contains L_i levels in its hierarchy. How many cuboids will be created, and why?

(c) Consider the *AllElectronics* data warehouse which consists of three dimensions *time*, *item*, and *location*, and one measure *sales*. The dimension hierarchies used are [day < month < quarter < year] for *time*; [item name < brand < type] for *item*; and [street < city < state < country] for *location*. Given a group-by query on {brand, state}, can we use each of the following cuboids to answer the query, and why?

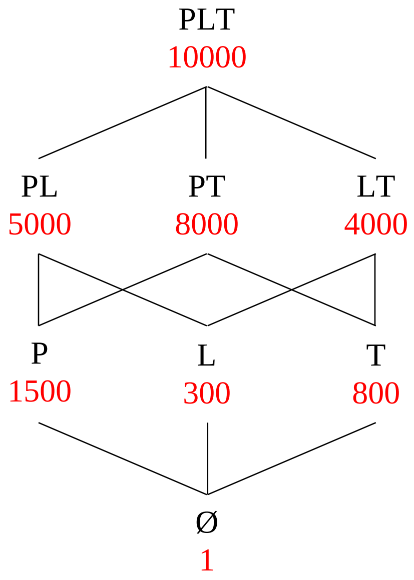
- Cuboid1: {year, item, city}
- Cuboid2: {year, brand, country}
- Cuboid3: {year, brand, state}
- Cuboid4: {item name, state}

(d) Which of the above cuboids is the best, in terms of query efficiency, to answer the group-by query on {brand, state}, and why?

Question 3: Consider a data warehouse which contains three dimensions *product*, *location*, and *time* with no hierarchies. Below is a lattice of all possible cuboids created on the data warehouse, where P, L, and T represent *product*, *location*, and *time*, respectively. Each of the red numbers shows the cost of using the corresponding cuboid, if materialized, to answer a group-by query. Suppose that the frequency distribution of group-by queries is as follows:

{PTL (0.05), PL (0.25), PT (0.15), LT (0.1), P (0.2), L (0.1), T (0.1), Ø (0.05)}.

What are the first two cuboids that should be materialized in order to minimize total query cost, and why?



---ooo000O000ooo---