Data Warehousing

CS-452

Spring 2023

Assignment # 3

Due Date: July 5th, 2023

Question 1:

Data warehouse for your University consists of the five dimensions' student, course, semester, administration staff and your instructor, and two measures count and average grade of all students. At the lowest conceptual level, the average grade measure stores the actual course grade of the student. At higher conceptual levels, average grade stores the average grade for the given combination.

a. Draw a snowflake schema diagram for the data warehouse of your institute.

b. 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 *University* student.

c. 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)?

Question 2:

Suppose a company wants to design a data warehouse to facilitate the analysis of moving vehicles in an online analytical processing manner. The company registers huge amounts of auto movement data in the format of (*Auto ID, location, speed, time*). Each *Auto ID* represents a vehicle associated with information (e.g., *vehicle category, driver category*), and each location may be associated with a street in a city. Assume that a street map is available for the city.

a. Design such a data warehouse to facilitate effective online analytical processing in multidimensional space.

b. The movement data may contain noise. Discuss how you would develop a method to automatically discover data records that were likely erroneously registered in the data repository.

c. The movement data may be sparse. Discuss how you would develop a method that constructs a reliable data warehouse despite the sparsity of data.

d. If you want to drive from A to B starting at a particular time, discuss how a system may use the data in this warehouse to work out a fast route.

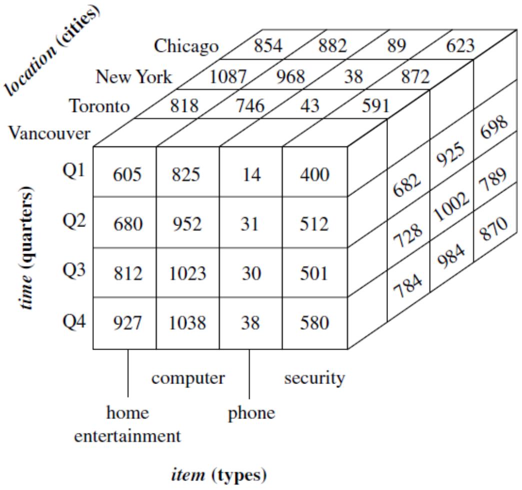
Question 3:

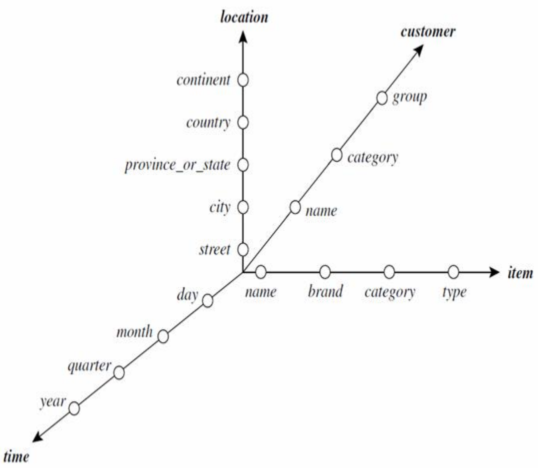
Radio-frequency identification is commonly used to trace object movement and perform inventory control. An RFID reader can successfully read an RFID tag from a limited distance at any scheduled time. Suppose a company wants to design a data warehouse to facilitate the analysis of objects with RFID tags in an online analytical processing manner. The company registers huge amounts of RFID data in the format of *(RFID, at location, time*), and also has some information about the objects carrying the RFID tag, for example, *(RFID, product name, product category, producer, date produced, price*).

a. Design a data warehouse to facilitate effective registration and online analytical processing of such data.
b. The RFID data may contain lots of redundant information. Discuss a method that maximally reduces redundancy during data registration in the RFID data warehouse.
c. The RFID data may contain lots of noise such as missing registration and misread IDs. Discuss a method that effectively cleans up the noisy data in the RFID data warehouse.
d. You may want to perform online analytical processing to determine how many TV sets were shipped from the LA seaport to BestBuy in Champaign, IL, by *month*, *brand*, and *price range*. Outline how this could be done efficiently if you were to store such RFID data in the warehouse.
e. If a customer returns a jug of milk and complains that is has spoiled before its expiration date, discuss how you can investigate such a case in the warehouse to find out what the problem is, either in shipping or in storage.

Question 4:

Draw the final resultant data cube and explain the roll-up OLAP operation on "time" dimension using following data cube.





Question 5:

A data warehouse of a telephone provider consists of five dimensions: caller customer, callee customer, time, call type, and call program and three measures: number of calls, duration, and amount. Define the OLAP operations to be performed in order to answer the following queries. Propose the dimension hierarchies when needed.

a) Total amount collected by each call program in 2012.
b) Total duration of calls made by customers from Brussels in 2012.
c) Total number of weekend calls made by customers from Brussels to customers in Antwerp in 2012.
d) Total duration of international calls started by customers in Belgium in 2012.
e) Total amount collected from customers in Brussels who are enrolled in the corporate program in 2012.

Question 6:

a. Divide your Data Warehouse into four stages and put the following tasks in their respective stage(s).

i. Sales
ii. ETL Process
iii. ODS
iv. Staging Area
v. External Data
vi. Data Marts
vii. Database
viii. ERP
ix. Text Data
x. Data Mining
xi. Marketing

b. By using above question answer, explain in your words with logic, why you assigned a task in that particular stage of the data warehouse.

Question 7:

A data warehouse of a train company contains information about train segments. It consists of six dimensions, namely, departure station, arrival station, trip, train, arrival time, and departure time, and three measures, namely, number of passengers, duration, and number of kilometers. Define the OLAP operations to be performed in order to answer the following queries. Propose the dimension hierarchies when needed.

a. Total number of kilometers made by Alstom trains during 2012 departing from French or Belgian stations.
b. Total duration of international trips during 2012, that is, trips departing from a station located in a country and arriving at a station located in another country.
c. Total number of trips that departed from or arrived at Paris during July 2012.
d. Average duration of train segments in Belgium in 2012.
e. For each trip, average number of passengers per segment, which means take all the segments of each trip and average the number of passengers.

Question 8:

Consider the data warehouse of a university that contains information about teaching and research activities. On the one hand, the information about teaching activities is related to dimensions' department, professor, course, and time, the latter at a granularity of academic semester. Measures for teaching activities are number of hours and number of credits. On the other hand, the information about research activities is related to dimensions' professor, funding agency, project, and time, the latter twice for the start date and the end date, both at a granularity of day. In this case, professors are related to the department to which they are affiliated. Measures for research activities are the number of person months and amount. Define the OLAP operations to be performed in order to answer the following queries. For this, propose the necessary dimension hierarchies.

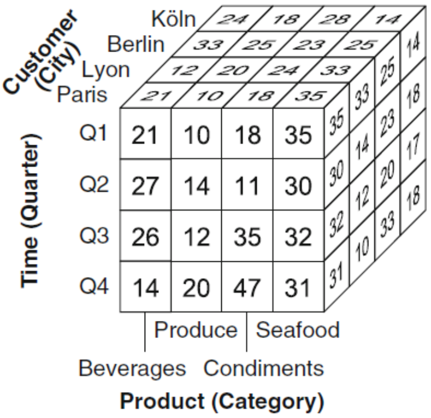
a. By department the total number of teaching hours during the academic year 2012–2013.

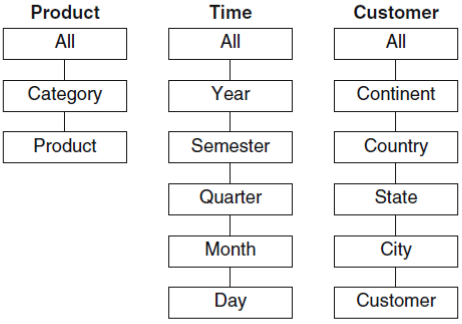
b. By department the total amount of research projects during the calendar year 2012.
c. By department the total number of professors involved in research projects during the calendar year 2012.
d. By professor the total number of courses delivered during the academic year 2012–2013.
e. By department and funding agency the total number of projects started in 2012.

Question 9:

Perform the following OLAP operations on the following data cube and show the final result (data cube). Hint: you can use dummy data if the cell in the resultant data cube is empty.

i. Drill-down on 'Time' dimension.
ii. Slice on Customer = 'Paris'.
iii. Dice on Customer = 'Paris' and 'Lyon' and Time = 'Q1' and 'Q2'.
iv. Dice on Customer = 'Paris' and Time = 'Q1'.
v. Dice on Customer = 'Paris' and Time = 'Q1' and 'Q2' and Product = 'Seafood' and 'Condiments'.





Question 10:

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

i. Draw a snowflake schema diagram for the data warehouse.
ii. 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 University student.

Question 11:

List and explain the properties of aggregation operators.

Question 12:

Design a data warehouse for a regional weather bureau. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for more than 10 years. Your design should facilitate efficient querying and online analytical processing, and derive general weather patterns in multidimensional space.

Question 13:

Regarding the *computation of measures* in a data cube:

a. Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube.

b. For a data cube with the three dimensions' *time, location*, and *item*, which category does the function *variance* belong to? Describe how to compute it if the cube is partitioned into many chunks.

   Hint: The formula for computing *variance* is $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_i)^2$, where $\bar{x}_i$ is the average if $x_i s$.

c. Suppose the function is "*top 10 sales*." Discuss how to efficiently compute this measure in a data cube.

Question 14:

Suppose that a data warehouse contains 20 dimensions, each with about five levels of granularity.

a. Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to support this preference efficiently?

b. At times, a user may want to *drill through* the cube to the raw data for one or two particular dimensions. How would you support this feature?

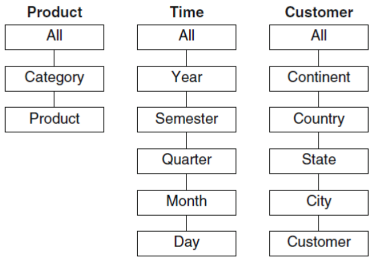
Question 15:

If the cube has 12 dimensions and each dimension has total five levels, what is the total number of cuboids that can be generated? Make sure to show all steps when solving.

Question 16:

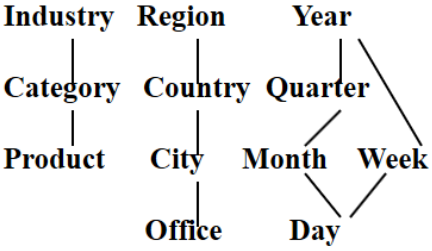
If the cube has 3 dimensions and each dimension has different levels (as shown in figure), what is the total number of cuboids that can be generated? Make sure to show all steps when solving. Hint: Formula,

$$T = \prod_{i=1}^{n}(L_i + 1)$$



Question 17:

If the cube has 3 dimensions and each dimension has different levels (as shown in figure), what is the total number of cuboids that can be generated? Make sure to show all steps when solving.



Question 18:

Explain the following efficient data accessing methods

a. Index Structures.
b. Materialized Views (using cuboids).

Question 19:

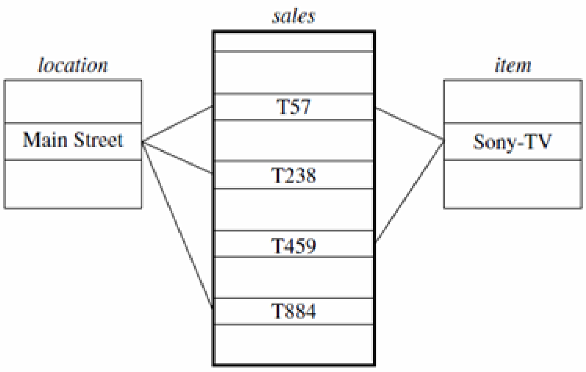
a. Explain bitmap indexing using suitable example.
b. Apply bitmap indexing method and draw the *item bitmap index table.*
c. Apply bitmap indexing method and draw the *city bitmap index table.*

R_ID	***Item***	***City***
R1	Home Entertainment	Vancouver
R2	Computer	Vancouver
R3	Phone	Vancouver
R4	Security	Vancouver
R5	Home Entertainment	Toronto
R6	Computer	Toronto
R7	Phone	Toronto
R8	Security	Toronto

Question 20:

a. Explain join indexing using suitable example.
b. Using following data shown in figure,

   i. Apply join indexing method and draw the *join index table for location/sales.*
   ii. Apply join indexing method and draw the *join index table for item/sales*
   iii. Apply join indexing and draw *join index table linking location and item to sales.*


sales
location
item
Main Street
Sony-TV
T57
T238
T459
T884