Research School of Computer Science, Australian National University

COMP3420 Advanced Database and Data Mining

Tutorial 4

Note: Questions with an asterisk (*) are roughly more challenging than 85% of the course material, i.e., you are especially encouraged to master them if you'd like to achieve more than 85/100 in this course.

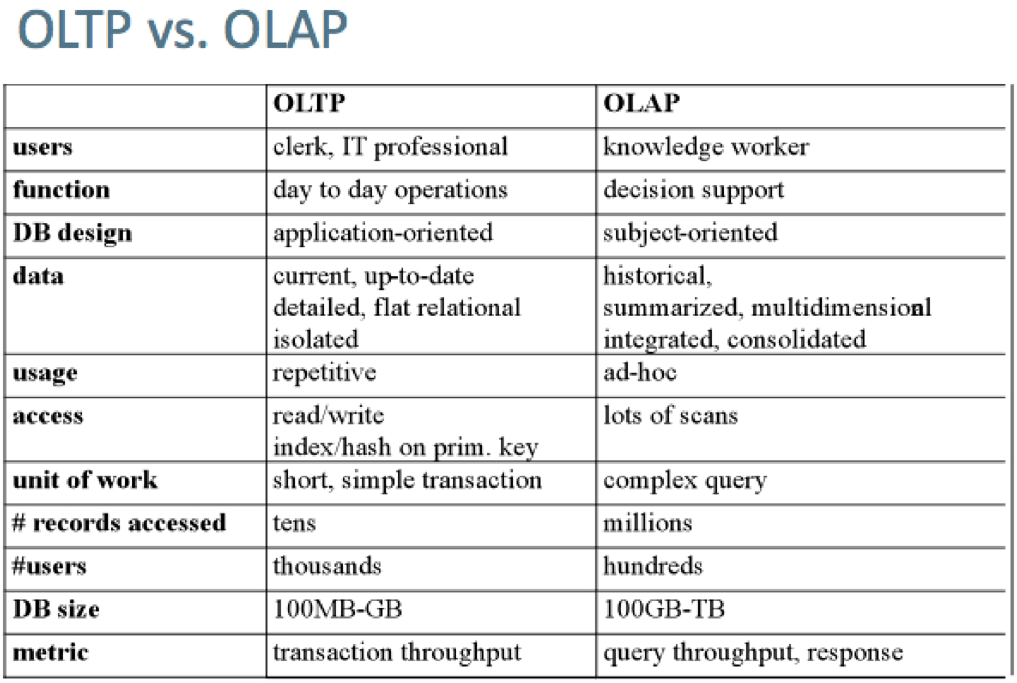
Question 1 What are OLAP and OLTP and what distinguishes them?

Answer:

OLTP (on-line transaction processing) is one of the major operations in traditional relational DBMS, it can be used to deal with day-to-day operations, purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

OLAP (on-line analytical processing) is the major analysis engine in data warehousing. It deals with multidimensional data analysis and provides the support to enterprise strategic decision making.

The comparison between OLAP and OLTP is shown in the Figure below.



OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Question 2 Can you name two examples of OLAP processing in the real-world?

Answer: For example Google search is an example of OLTP where people search and interact with the system. Google trends on the other hand is an OLAP where the aggregated data is presented and the historical aspect of the database is exploited.

Another example is ATM machines where customers perform transactions. The system that shows the user transactions and aggregated balance for all their accounts would be an OLAP.

Question 3 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.

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 4 Data mining and data warehousing concepts

Suppose you are the data analyst of the road transportation authority (RTA) in a medium-sized city. RTA wants to optimize its physical asset maintenance process. This process is responsible for ensuring the normal operation of all transportation assets, such as roadwork, signs, equipment and facilities. Your mission is to help on this goal by examining the following three databases.

(i) Physical assets. These include transportation assets such as traffic lights, streetlights, road signs, traffic cameras, road/sidewalk/bikepath segments, roundabouts; operational assets such as trucks, as well as repairing tools.

(ii) Crew. These include information about each worker: name, title, address, contact number, responsibility, skill, hourly pay rate, reporting structure, shift and availability.

(iii) Maintenance. These include work orders being executed every day: the type of order (e.g. repair, re-paint street lines, install street signs, pave new roads), its priority (critical/urgent/high/medium/low), location, day/time of different status updates (reported/dispatched/completed), which crew performed each order, which asset(s) was used and/or worked on, ...

Your first task is to build a data warehouse from these three databases to analyze maintenance needs and cost.

(a) Which one of the above is the transaction database?

Answer: A transactional database is where a database transaction might consist of one or more data-manipulation statements and queries, each reading and/or writing information in the database. ”Maintenance” is transaction DB.

(b) In one sentence give one example scenario where entries are **added** to this database daily? Give another example scenario where entries are **updated** in this database daily?

Answer: Add entry: call center operator adds workorders on as phone reports arrive.
Update entry: each workorder gets updated when it's assigned, dispatched, and completed.

(c) Your data warehouse design starts from the three-tier data warehouse diagram below.

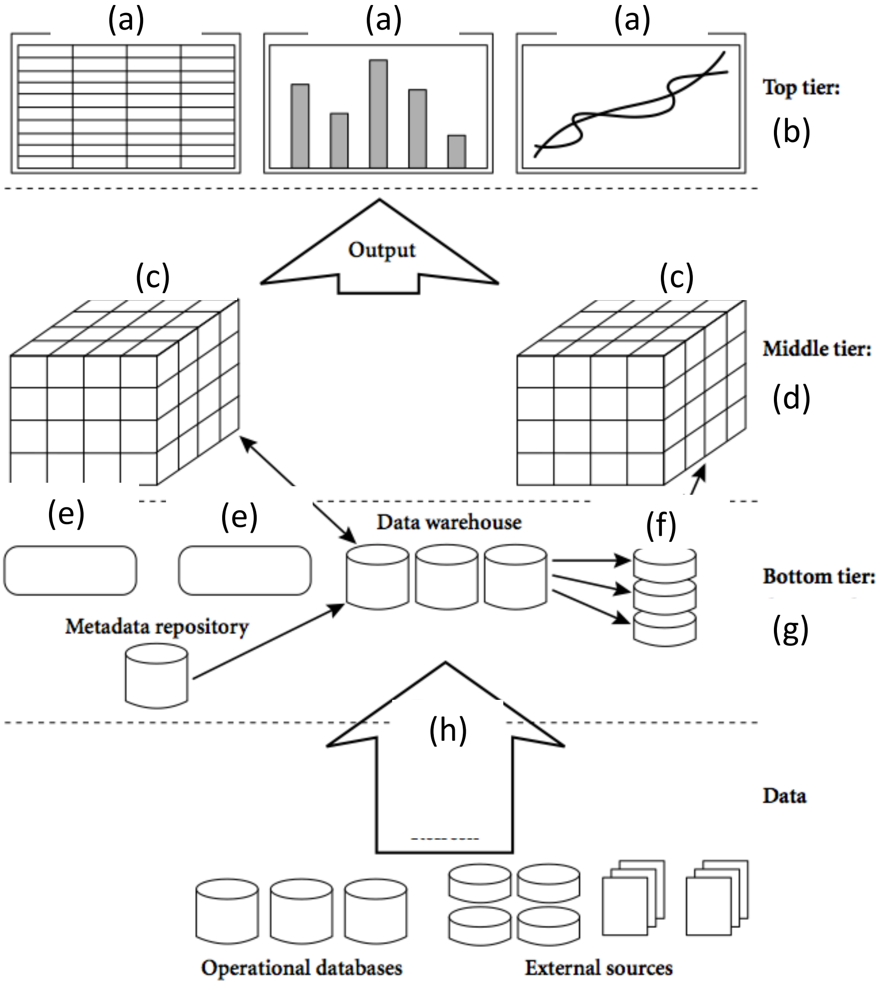


Figure 1: A three-tier data warehouse architecture.

Name the three different layers in Figure 1:

Top tier (b):
Middle tier (d):
Bottom tier (g):

Answer: top: front-end tools; mid: OLAP server; bottom: data warehouse server.

(d) Then label the different components in Figure 1. You can choose from the following components: Administration, Monitoring, OLAP Server, Querying, Reporting, Data

Marts, Visualization, Data Mining.
Note: there may be more than one labels that is appropriate for each letter.

(a) in the top tier;

(c) in the middle tier;

(e) in the bottom tier;

(f) in the bottom tier.

Answer: (a) in the top tier: Querying, Reporting, Data Mining.

(c) in the middle tier: OLAP Server.

(e) in the bottom tier: Monitoring, Administration.

(f) in the bottom tier: Data Marts.

(e) There are several operations being performed at the arrow labeled as (h) in Figure 1. Name at least two of them.

Answer: Extract; Clean; Transform; Load; Refresh.

(f) What are two examples of **external** data sources you may use for the RTA data warehouse in Figure 1. Given that RTA staff often perform work outdoors in a complex urban environment, that other city agencies (such as water, electric and telecom suppliers) may also need to work on road segments.

Answer: weather data, water pipe repair schedule, street wiring schedule for electric and telcom, public event schedule, etc

Question 5 Review of data pre-processing

Your next task is to pre-process the RTA data in Question 1 to construct the data warehouse.

(a) There are a number of different forms of data pre-processing that you can perform, including data cleaning and data transformation. Can you name two more forms?

Answer: Data integration: Data may be kept in several sources that should be integrated and unified.

Data reduction: Can be used to obtain a smaller representation of the dataset that is easier to use in data mining algorithms.

(b) In the RTA scenario, work-orders are either called in by citizens or recorded by RTA crew. The location of the work-orders can be in the form of an address, an intersection, or a (latitude, longitude) tuple from GPS devices carried by the crew. Give one example in which this process will generate noise in the Maintenance database, and write down your proposed method to clean such noise

Answer: duplicate removal, clustering, etc.

(c) You visualize the Maintenance database using a series of plotting tools. Name two of such plots. Specify the input of each plot, and its purpose. Your answers can be from both within the textbook and from real-world practices.

Answer: box plot, scatter plot on a map, histogram, time series, etc.

(d) You examine a sample of the Maintenance database. It has a column recording the number of crew members needed for each job (`num_crew`), that takes integer values such as 1, 2, ... Over 1000 different work-orders, there were:
200 work-orders with `num_crew`=1;
300 work-orders with `num_crew`=2;
220 work-orders with `num_crew`=3;
200 work-orders with `num_crew`=5;
70 work-orders with `num_crew`=10;
8 work-orders with `num_crew`=15;
2 work-orders with `num_crew`=25;
If you were asked to convert numeric values of `num_crew` to four different types: *single-crew*, *small-team*, *large-team* and *multiple-teams*.
Write down your mapping from each of the observed `num_crew` value to one of the four target types. Briefly state why.

Answer: *single-crew*:1, *small-team*: 2 and 3, *large-team*: 5 and *multiple-teams*: the rest

(e) Compute the **mean** and **median** of `num_crew` among the 1000 work-orders in the previous part.

Answer: median: 2.5;
mean: (200+300*2+220*3+200*5+700+120+50)/1000=3.33

Question 6 Data cubes and OLAP operations

You selected four dimensions from the RTA databases in Question 4 to build a basecube. There dimensions are: `date, work-order type, priority, location`. There are three measures over these dimensions: `count, time_taken`, and `cost`. Here `time_taken` is computed as the number of hours elapsed between job dispatch time and job completion time; and `cost` is computed as the sum of salary rates for all crew members involved, plus truck milage and consumables (e.g. petrol).

(a) Draw a star schema for this data warehouse, with the three measures and four dimensions specified above.

Answer: The star schema for this database is shown in Figure 3.

(b) Specify how you can use all or a subset of these OLAP operations to get the number of road-repair work-orders that were urgent or critical during the first quarter of 2011.

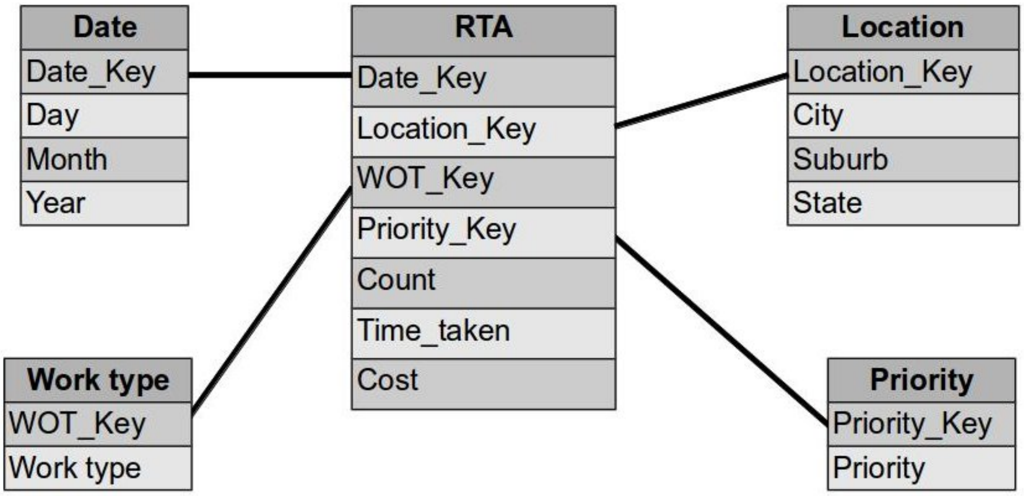
Answer: Roll-up on `location`, and then slice/dice the data on `date`, `work-order type`, `priority`.

Question 7 Data cube computation

Use the dataset in Question 6.

(a) In this base cuboid, `date` has three levels: day, month and quarter; `work-order type` has two levels; `priority` and `location` each has one level. How many cuboids are there in the full data cube?

Figure 2: Star Schema for RTA database



Answer: #cuboids $= (3 + 1) \times (2 + 1) \times (1 + 1) \times (1 + 1) = 48$.

(b) Assume that you are testing this datacube with a small data sample, containing all `dates` (365) in 2011, 20 different `work-order type`, 5 different `priority` values and the entire city partitioned into 10 distinct `location` zones based on postcode. There are two work-orders in this data sample:

`work-order-id`	`date`	`work-order type`	`priority`	`location`
00123	2011-06-09	traffic_light_repair	urgent	4217
00234	2011-06-09	traffic_sign_install	medium	4222

What is the fraction of empty (zero) cells in the cuboid with dimensions `work-order type, priority, location` ?

Answer: There are totally $20 \times 5 \times 10 = 1000$ cells in this cuboid, while there are only 2 records, so the fraction of empty cells is $1 - \frac{2}{1000} = 99.8\%$.

(c) (*) Take the data cube in part (b) of this question. How many non-zero cells will you need to compute for the entire data cube? This includes the base cubiod, the apex cuboid, and everything in between.

Answer: There are totally 5 levels in the lattice. The number of non-empty cells in each level is: 0-D cuboid: 1

1-D cuboid: 1 (date) + 2 + 2 + 2 = 7

2-D cuboid: 2 + 2 + 2 + 2 + 2 + 2 = 12

3-D cuboid: 2 + 2 + 2 + 2 = 8

4-D cuboid: 2.

So the total number of non-zeros cells = 1 + 7 + 12 + 8 + 2 = 30.

(d) (*) Give two different examples about how you could use the available data in the RTA databases to help improve transportation service quality or save operating cost.

For each example specify the following: which database(s) to take data from, what are the dimensions and measures of the data cube, what kind of mining or analysis you will do on the data cube, and what is the possible action(s) that can be taken from the analysis results.
Note: you can (but do not have to) use the data cube given in this problem in one of your examples.

Answer: Example 1: because different work-order types require different materials, suppose we want to know which area has the most often traffic light repairs in the last 1 year. To do this, we need the Maintenance database, the dimension of this data cube would be 3, *e.g.*, `date, work-order type, location`. The measure would be *counts*, and we need to do the counting for each area and each work-order type. Based on this information, we then can dispatch different materials to different area based on the historical statistics.

Example 2: suppose in one day the road maintenance is only allowed to be on for several hours due to some important affair to be held in a specific area. As a result, we have to only perform the most urgent maintenance activities, *e.g.*, install street signs; and at the same time inform the worker of the other maintenances to lay off. In this case, we need the 'Maintenance' database and 'Crew' database, the dimension of 'Maintenance' is 4, *e.g.*, `date, work-order type, priority, location`, and we only need 2 dimension of the 'Crew' database, *e.g.*, `name, contact number`. We can take the `count` as the measure (counts of activities that are urgent). The task is to retrieve the tasks that are urgent and also find the corresponding workers on 'Crew' for the un-urgent tasks and get them informed.

Question 8 A "Flat" *Data Cube* about Cars.

Download a local copy of the "Car Evaluation" dataset http://cecs.anu.edu.au/~xlx/comp3420/cars.csv or http://wattlecourses.anu.edu.au/mod/resource/view.php?id=184679. Information about the data can be found at the UCI repository http://archive.ics.uci.edu/ml/datasets/Car+Evaluation. Load it into Excel or open office.

1. How many dimensions does the base cuboid has? How many cells are in the base cuboid? How many different 3-dimensional cuboids can this data have? How many cuboids are in the full lattice?

2. (*) How many data objects (i.e. cars) are in this dataset? At least how many cells in the base cuboids are empty?

3. Build a 3-dimensional data cube/cuboid using any three dimensions you choose. The tool is called "Pivot Table" in Excel or "Data Pilots" in OpenOffice, look at instructions in the following few links if you need.

   - Google Spreadsheet https://support.google.com/drive/bin/answer.py?hl=en&answer=1272898&topic=1258755&rd=1
   - OpenOffice http://openoffice.blogs.com/openoffice/2006/11/data_pilots_in_.html
   - Microsoft Excel http://office.microsoft.com/en-us/excel-help/pivottable-reports-101-HA00103463 aspx
   - Wikipedia narrative about pivot tables http://en.wikipedia.org/wiki/Pivot_table

4. How many cars have the highest class value – "vgood" while at the same time has "low" buying cost, "low" maintenance cost, and "high" safety?

5. How do you do slicing, dicing, drilling down and rolling up on this data cube?

6. How else will you use this dataset to support car-purchasing decisions and find a "dream car"?

7. Compared to Pivot Tables, what additional things can a *real* data cube in a data base system do? When shall we choose to use such *spreadsheet cubes* versus *database cubes*?

8. Can you derive/guess something about how "ClassValue" gets assigned based on the other six attributes?

9. Take another UCI dataset (http://archive.ics.uci.edu/ml/index.html, candidates can be: Wine, Pittsburgh Bridges, Adult, ...), or one from your own sources. Build data cubes as above, and try to imagine a few scenarios that it could be useful for knowledge discovery or decision-making.

Answer:

1. The base cuboid has 7 dimensions (from the UCI page, 6 attributes plus "class value"). There are 4 dimensions with 4 distinct values, and 3 dimensions with 3 distinct values. The total number of cells is therefore $4^4 \times 3^3 = 6912$ in the base cuboid; There are $C_7^3 = 35$ ways of picking any 3 dimensions from the total 7. There are $2^7 = 128$ cuboids in the lattice.

2. There are 1728 instances in the set which was supposed to be 6912, thus, $6912 - 1728 = 5184$ cells are empty.

3. An example of pivot table is shown in Figure 4.

4. 13.

5. Using the pivot table. Slicing and dicing: select subsets of values in any dimension or a number of dimensions. Drilling down: add dimension(s) to the column or row. Rolling up: deleting dimension(s) from the column/row.

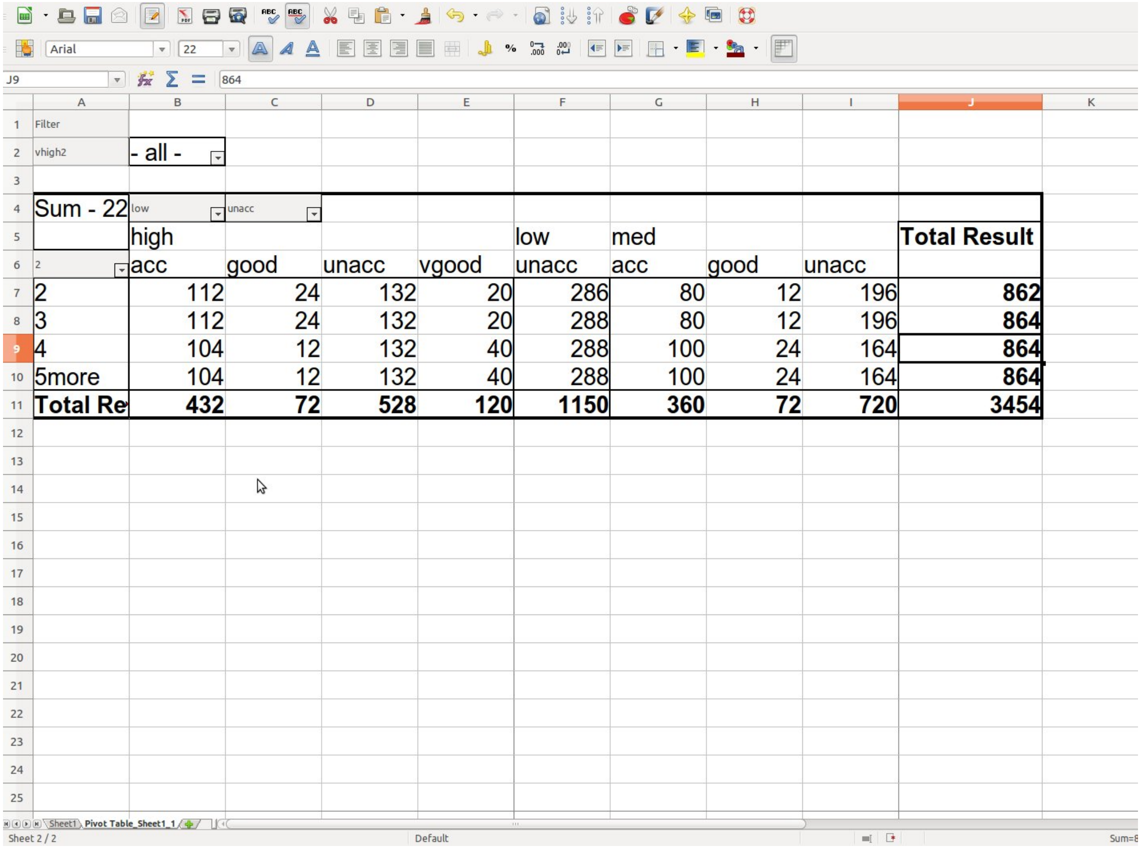
6. You may want to find a car which has low cost, low maintainance cost, size (capacity for people and luggages), safety, luxury and comfort (not covered in this data) etc. This is a open question.

7. Pivot table does not support OLAP ops such as roll-up, drill down, we'd have to re-do the pivot table. Pivot table supports slicing because it aggregate all the data.

8. (use your observations, also see classification, especially decision trees in the 2nd half of this course) If safety equals low, then it gets unacc in ClassValue; buying, maintain are low,safety is high, lug-boot is not small, it gets vgood; buying, maintain are low,safety is high, lug-boot is small, it gets goodm, etc.

9. left as self-paced exploration.

Figure 3: Pivot table example



	A	B	C	D	E	F	G	H	I	J
1	Filter									
2	vhigh2	- all -								
3										
4	Sum - 22	low	unacc							
5		high				low	med			Total Result
6	2	acc	good	unacc	vgood	unacc	acc	good	unacc	
7	2	112	24	132	20	286	80	12	196	**862**
8	3	112	24	132	20	288	80	12	196	**864**
9	4	104	12	132	40	288	100	24	164	**864**
10	5more	104	12	132	40	288	100	24	164	**864**
11	**Total Re**	**432**	**72**	**528**	**120**	**1150**	**360**	**72**	**720**	**3454**

Question 9 Given a data cube consisting four dimensions D_1,D_2, D_3 and D_4, assume that the numbers of levels associated these dimensions are 10, 4, 3, and 15. How many cuboids does the data cube contain?

Answer: The number of cuboids contained in the data cube is $\prod_{i=1}^{4}(L_i + 1) = 11 * 5 * 4 * 16 = 3,520$.

How is the answer for this question different from that of Question 8.1?

Answer: In Question 1 every dimension is assumed to have only $L = 1$ level. The total number of cuboids are $\prod_{i}^{n}(L + 1)\|_{L=1} = 2^n$. See book page 139.

Question 10 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 11 What are examples of the three types of metadata above, for the RTA database in Question 4?

Answer:

- Technical meta-data: Data types, size of data types, Machine on which it will be stored, IP of the machine.
- Operational meta-data: Last update date/time, Last user updated data, warehouse usage, error report, etc.
- System performance metadata: retrieval performance, rules for timing and scheduling of refresh, etc.
- Business meta-data: Data ownership policies, Charging policies.

Question 12 What is the lattice of data cube? what is the cuboid? what is the cuboid cells? and what is the base cell or aggregate cell?

Answer: Each table derived from a different dimensional combination of the base table is a cuboid in the data cube. All cuboids in a data cube form a lattice of data cube. Each cell in the base table is a base cell. Otherwise, the cell in a non-base table (an aggregated cuboid) is the aggregate cell.

Question 13 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 be 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.
- **Slice 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.