CELEBRATING 25 YEARS
CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A+ GRADE BY NAAC

Scheme of Evaluation
Internal Assessment Test 1 – April 2023

Sub:	DATA MINING AND DATA WAREHOUSING							**Code:**	18CS641
Date**:**	25/4/2023	Duration:	90mins	Max Marks:	50	**Sem:**	VI	**Branch:**	ISE

Note: Answer Any five full questions.

Question #		Description	Marks Distribution		Max Marks
1	a)	Define i) Dimension ii) Measures iii) Fact Tables iv) Data Mining	1M*4	4M	10M
1	b)	Compare OLTP and OLAP systems. Any 6 differences	1M*6	6M	
2	a)	With suitable example, explain Star Schema, Snow Flakes Schema, Fact Constellation Schema for multidimensional database. Star Schema Snow Flakes Schema Fact Constellation	4M 3M 3M	10M	10M
3	a)	Explain OLAP operations in multidimensional data model with suitable example and diagram Diagram Roll Up Drill Down Slice/Dice Pivote	2M 2M 2M 2M 2M	10M	10M
4	a)	Differentiate ROLAP, MOLAP and HOLAP servers ROLAP MOLAP HOLAP	4M 3M 3M	10M	10M

5	a)	Explain the concept of Materialization for the selected computation of cuboids materialization No materialization Full materialization Partial materialization	1M 1M 1M 1M	4M	10M
5	b)	Write a short note on compute cube operator and curse of dimensionality Diagram Explanation + query	2M 4M	6M	
6	a)	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. a. Enumerate three classes of schemas that are popularly used for modelling data warehouses using Star Schema. b. Draw star and snowflake schema diagram for the above data warehouse. c. 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 2010. d. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge). a. Three classes b. Star and snowflakes schema c. Roll up operation d. query	2M 4M 2M 2M	10M	10M

Scheme Of Evaluation
Internal Assessment Test 1 – Oct 2022

Sub:	DATA MINING AND DATA WAREHOUSING							**Code:**	18CS641
Date**:**	25/4/2023	Duration:	90mins	Max Marks:	50	**Sem:**	VI	**Branch:**	ISE

Note: Answer Any full five questions

Q. 1 a) Define i) Dimension ii) Measures iii) Fact Tables iv) Data Mining

Dimension

A Data Dimension is a set of data attributes pertaining to something of interest to a business. Examples of dimensions are things like "customers", "products", "stores" and "time".

For users of Data Warehouses, data dimensions are entry points to numeric facts (e.g. sale, profit, revenue) that a business wishes to monitor.

Measures

A measure might be qualitative, such as a Product ID, or quantitative, such as a product's price. Following the example above, the act of selling a hat online entails several factors, including:

- It was sold at 1 p.m. the day before yesterday.
- It was purchased for $35.
- SFAF3423 is the product ID.

A measure is a name given to each of these facts about reality. Applying computations or aggregations to quantitative measures as needed for your data model is a part of data modeling. The qualitative measures can then be linked to the measure's specific properties, which are referred to as dimensions.

Fact Tables

A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information.

Data Mining

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Q 1b) Compare OLTP and OLAP systems

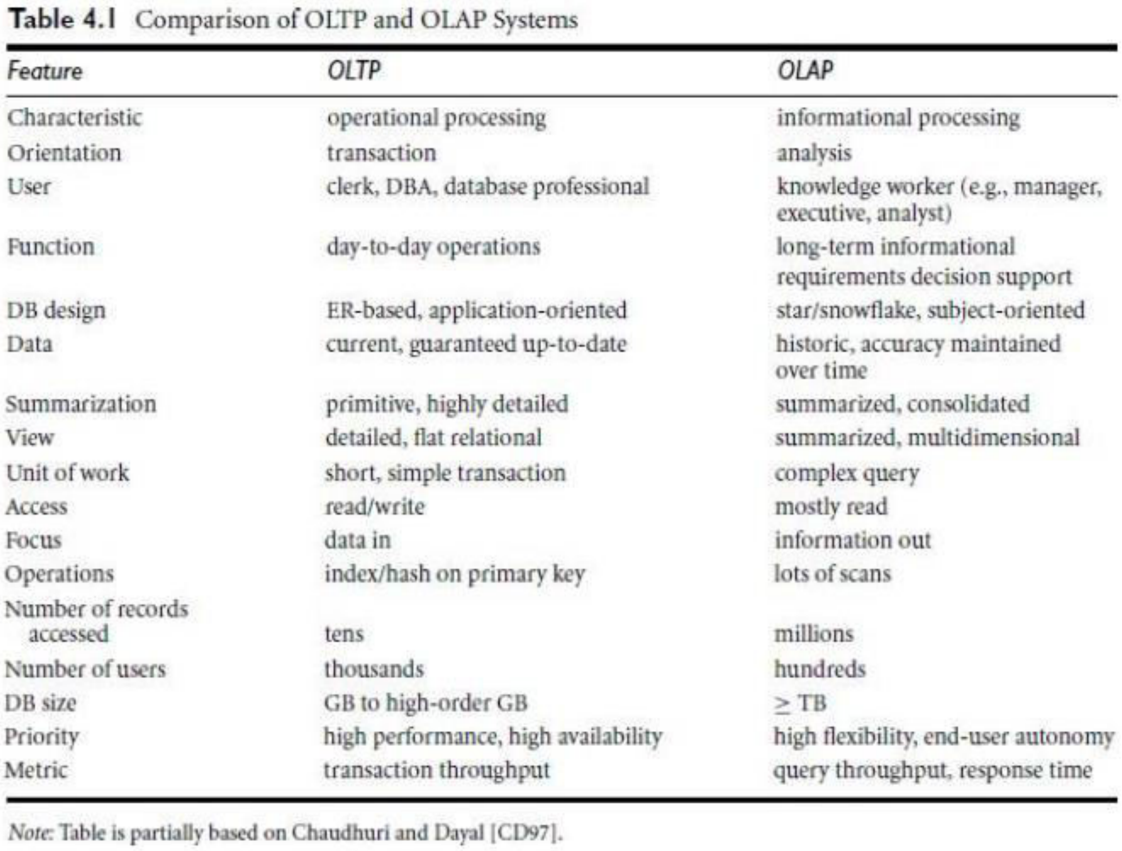


Table 4.1 Comparison of OLTP and OLAP Systems

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Note: Table is partially based on Chaudhuri and Dayal [CD97].

Differences between Operational Database Systems and Data Warehouses

- The major task of online operational database systems is to perform online transaction and query processing. These systems are called **online transaction processing (OLTP)** systems. They cover most of the day-to-day operations of an organization such as purchasing, inventory,manufacturing, banking, payroll, registration, and accounting.

- Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users. These systems are known as **online analytical processing(OLAP**) systems. The major distinguishing features of OLTP and OLAP are summarized as follows:

- **Users and system orientation:** An OLTP system is **customer-oriented** and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is **market-oriented** and is used for data analysis by knowledge workers, including managers, executives, and analysts.

- **Data contents:** An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier to use for informed decision making.

- **Database design**: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design.

- **View:** An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historic data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

- **Access patterns:** The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information), although many could be complex queries.

Q. 2 a) With suitable example, explain Star Schema, Snow Flakes Schema, Fact Constellation Schema for multidimensional database.

Star schema:

- A star schema for AllElectronics sales is shown in Figure 4.6. Sales are considered along four dimensions: time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (e.g., time key and item key) are system-generated identifiers
- The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

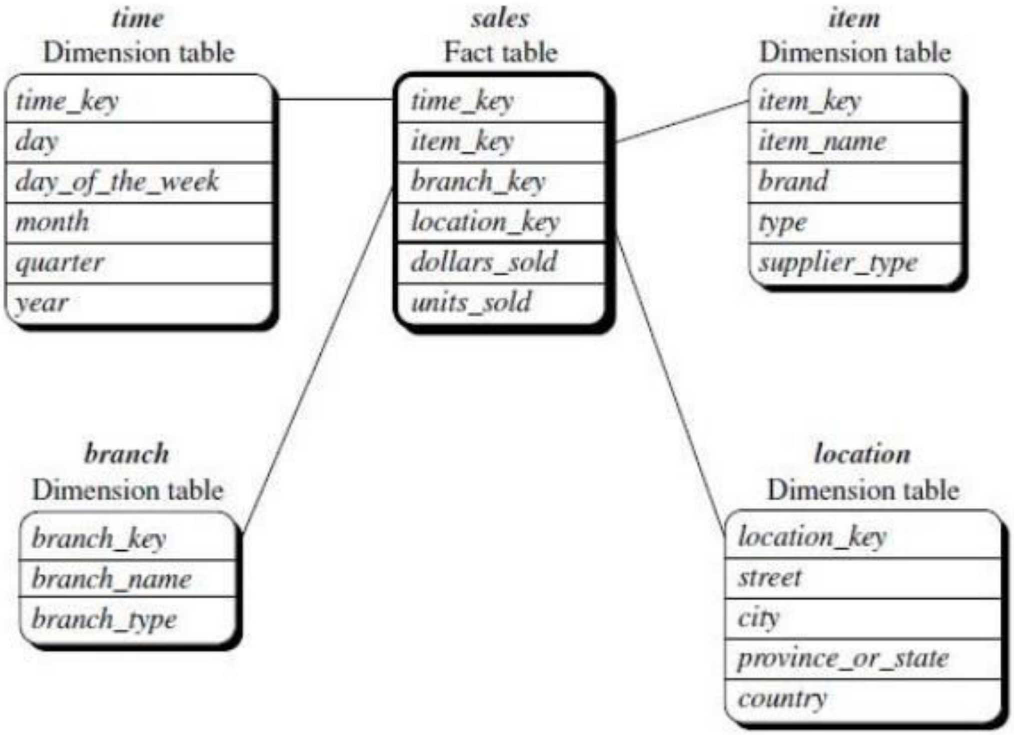
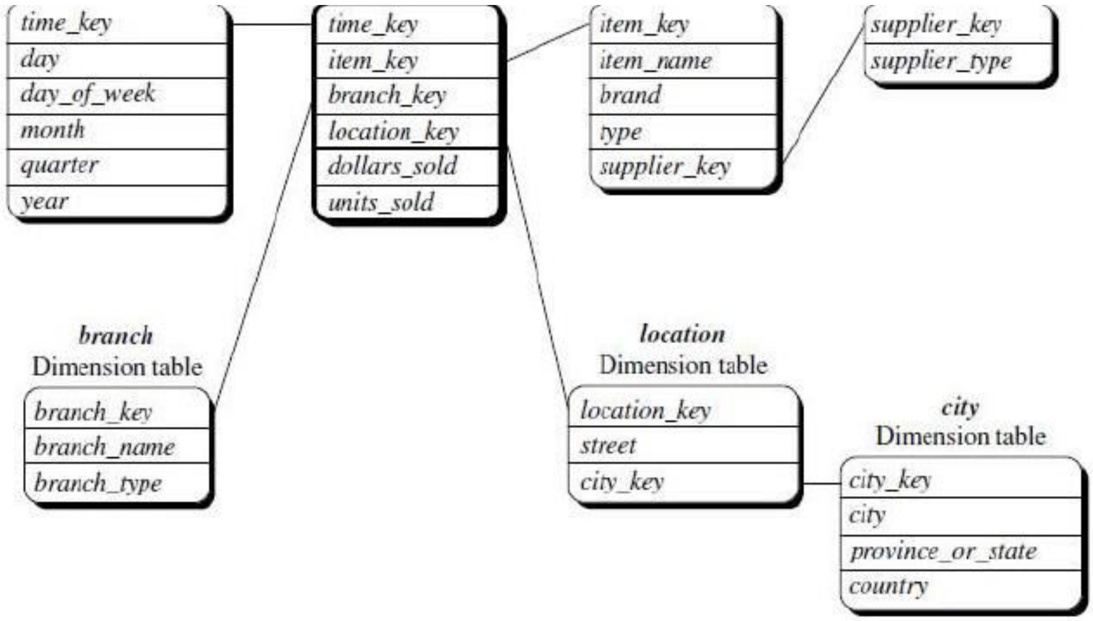


Figure 4.6 Star schema of *sales* data warehouse.

Snowflake schema:

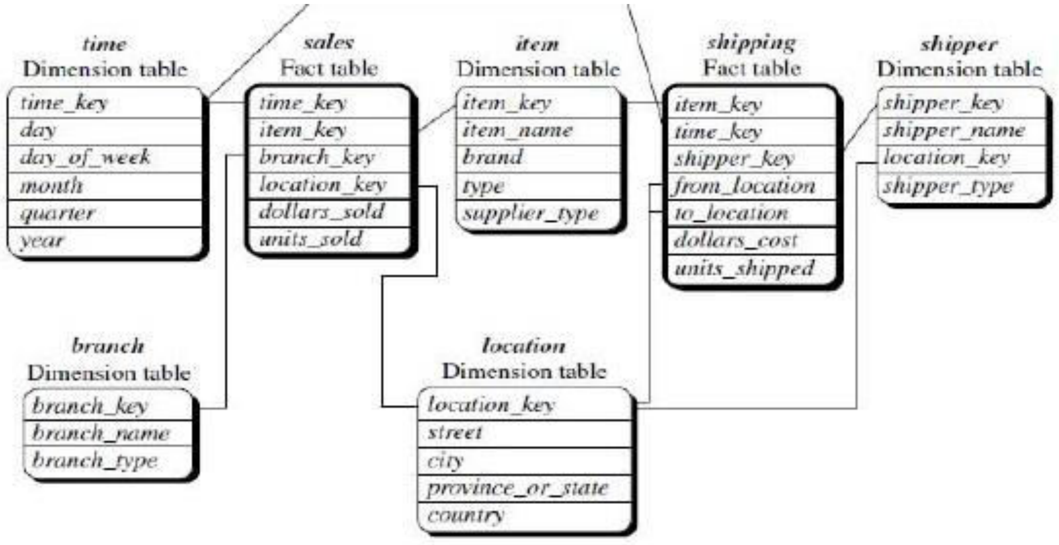
- The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.
- The sales fact table is identical to that of the star schema. The main difference between the two schemas is in the definition of dimension tables. The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables.
- For example, the item dimension table now contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information.
- Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension. Notice that, when desirable, further normalizationcan be performed on province or state and country in the snowflake schema shown in Figure 4.7.



4.7 Snowflake schema of a *sales* data warehouse.

Fact constellation:

- Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.
- Fact constellation schema specifies two fact tables, sales and shipping. The sales table definition is identical to that of the star schema (Figure 4.6).
- The shipping table has five dimensions, or keys—item key, time key, shipper key, from location, and to location—and two measures—dollars cost and units shipped.
- A fact constellation schema allows dimension tables to be shared between fact tables. For example, the dimensions tables for time, item, and location are shared between the sales and shipping fact tables.
- In data warehousing, there is a distinction between a data warehouse and a data mart. A data warehouse collects information about subjects that span the entire organization, such as customers, items, sales, assets, and personnel, and thus its scope is enterprise-wide.
- For data warehouses, the fact constellation schema is commonly used, since it can model multiple, interrelated subjects. A data mart, on the other hand, is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is department wide.
- For data marts, the star or snowflake schema is commonly used, since both are geared toward modelling single subjects, although the star schema is more popular and efficient.



ure 4.8 Fact constellation schema of a sales and shipping data warehouse.

Q. 3 Explain OLAP operations in multidimensional data model with suitable example and diagram.

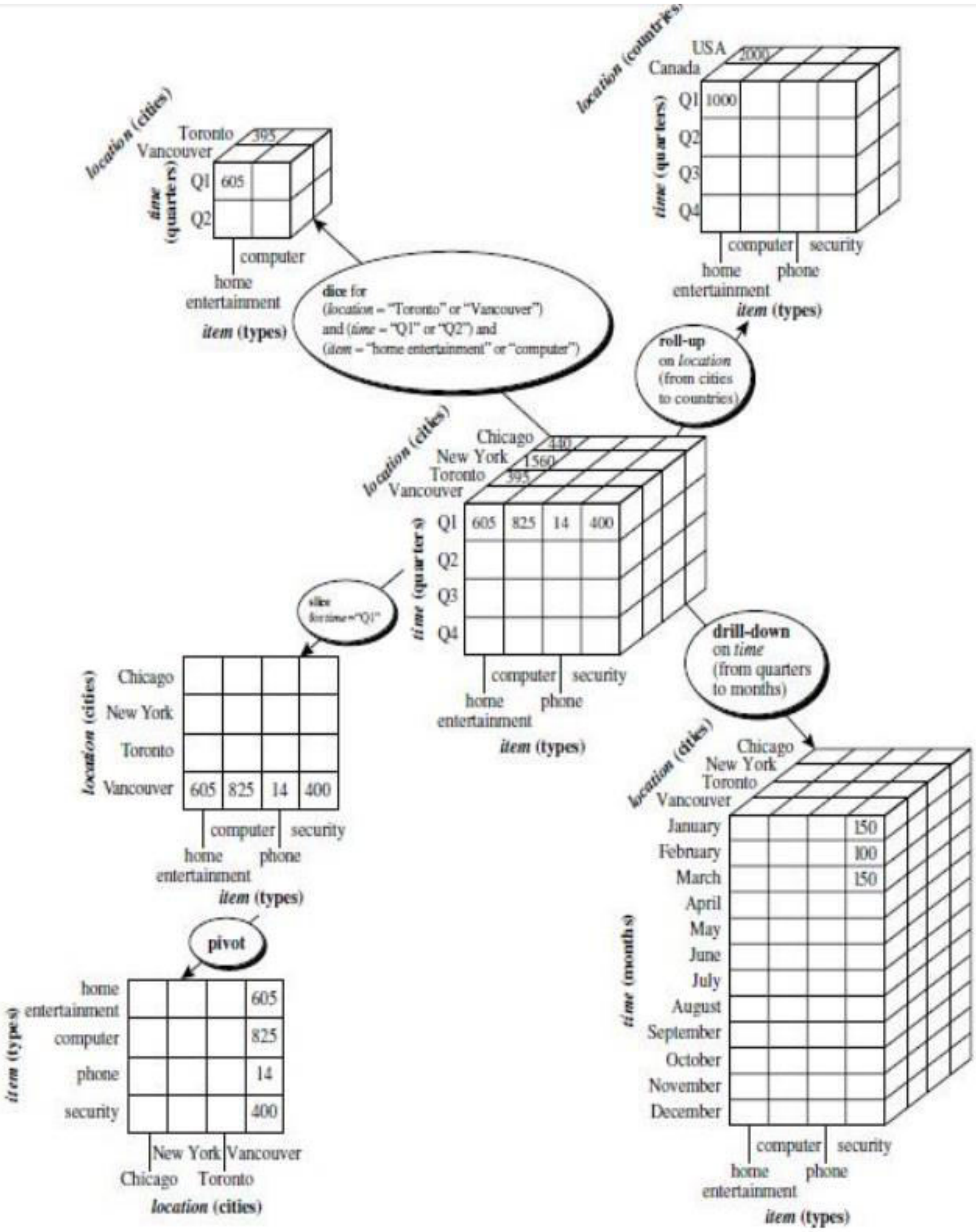
Typical OLAP Operations.

- Let's look at some typical OLAP operations for multidimensional data. Each of the following operations described is illustrated in Figure 4.12. At the center of the figure is a data cube for AllElectronics sales.
- The cube contains the dimensions location, time, and item, where location is aggregated with respect to city values, time is aggregated with respect to quarters, and item is aggregated with respect to item types.
- To aid in our explanation, we refer to this cube as the central cube. The measure displayed is dollars sold (in thousands). The data examined are for the cities Chicago, New York, Toronto, and Vancouver.

Roll up:

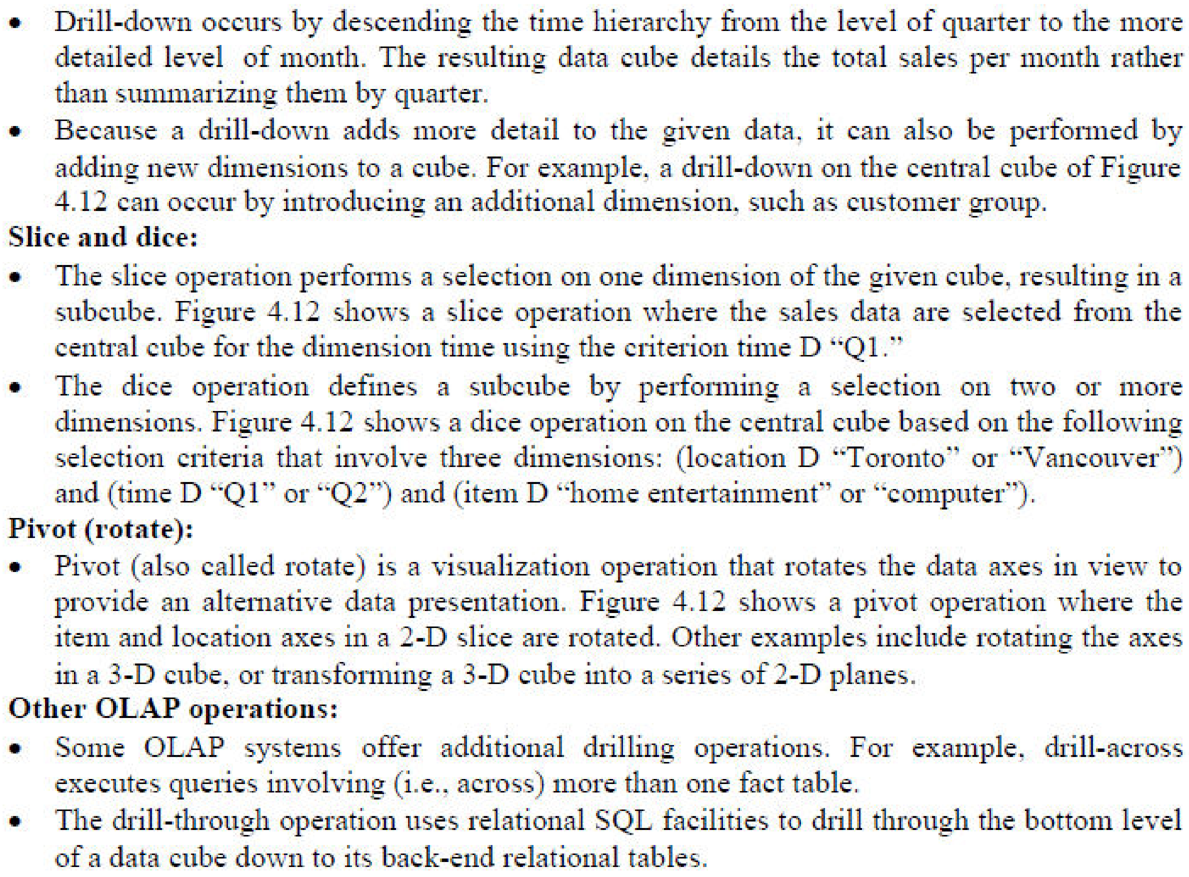
- The roll-up operation also called as the drill-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Figure 4.12 shows the result of a roll-up operation performed on the central cube by

  climbing up the concept hierarchy for location given in Figure 4.9. This hierarchy was defined as the total order "street < city < province or state < country."
- The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country. In other words, rather than grouping the data by city, the resulting cube groups the data by country.
- When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube. For example, consider a sales data cube containing only the location and time dimensions. Roll-up may be performed by removing, say, the time dimension,



Drill-down:

- 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.
- Figure 4.12 shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as "day < month < quarter < year."



- Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month. The resulting data cube details the total sales per month rather than summarizing them by quarter.
- Because a drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube. For example, a drill-down on the central cube of Figure 4.12 can occur by introducing an additional dimension, such as customer group.

Slice and dice:

- The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Figure 4.12 shows a slice operation where the sales data are selected from the central cube for the dimension time using the criterion time D "Q1."
- The dice operation defines a subcube by performing a selection on two or more dimensions. Figure 4.12 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions: (location D "Toronto" or "Vancouver") and (time D "Q1" or "Q2") and (item D "home entertainment" or "computer").

Pivot (rotate):

- Pivot (also called rotate) is a visualization operation that rotates the data axes in view to provide an alternative data presentation. Figure 4.12 shows a pivot operation where the item and location axes in a 2-D slice are rotated. Other examples include rotating the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes.

Other OLAP operations:

- Some OLAP systems offer additional drilling operations. For example, drill-across executes queries involving (i.e., across) more than one fact table.
- The drill-through operation uses relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.

Q. 4 Differentiate ROLAP, MOLAP and HOLAP servers

OLAP Server Architectures: ROLAP versus MOLAP versus HOLAP

- Logically, OLAP servers present business users with multidimensional data from data warehouses or data marts, without concerns regarding how or where the data are stored.
- However, the physical architecture and implementation of OLAP servers must consider data storage issues. Implementations of a warehouse server for OLAP processing include the following:

1. Relational OLAP (ROLAP) Servers:

- These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
- ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.
- ROLAP technology **tends to have greater scalability than MOLAP technology**. The DSS server of Micro strategy, for example, adopts the ROLAP approach.
- ROLAP works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information.
- It depends on a specialized schema design. This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality.
- In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.
- ROLAP **tools do not use pre-calculated data cubes but instead pose the query to the standard relational database and its tables in order to** bring back the data required to answer the question.
- ROLAP tools feature the ability to ask any question because the methodology does not limit to the contents of a cube. ROLAP also has the ability to drill down to the lowest level of detail in the database.

2. Multidimensional OLAP (MOLAP) servers:

- These servers support multidimensional data views through **array-based multidimensional storage engines.**
- They map multidimensional views directly to data cube array structures. The advantage of using a data cube is that it allows **fast indexing to pre-computed summarized data.**
- Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse.

- Many MOLAP servers adopt a **two-level storage representation to handle dense and sparse data sets**: Denser subcubes are identified and stored as array structures, whereas sparse subcubes employ compression technology for efficient storage utilization.
- MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database. **Therefore it requires the pre-computation and storage of information in the cube - the operation known as processing.**
- MOLAP tools generally utilize a pre-calculated data set referred to as a data cube. The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a **very fast response time and the ability to quickly write back data into the data set.**

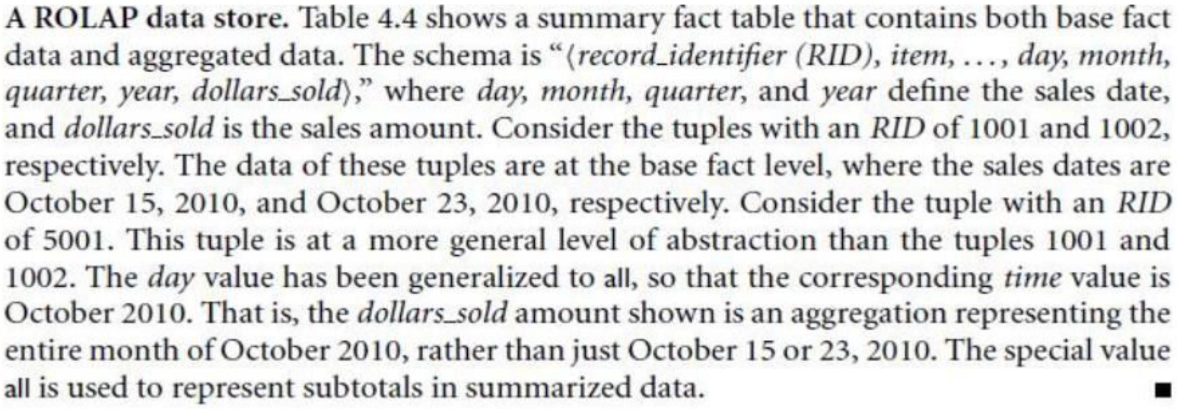
3.Hybrid OLAP (HOLAP) Servers:

- The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP. For example, a **HOLAP server may allow large volumes of detailed data to be stored in a relational database, while aggregations are kept in a separate MOLAP store**.
- The Microsoft SQL Server 2000 supports a hybrid OLAP server.
- There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage.

4. Specialized SQL servers:

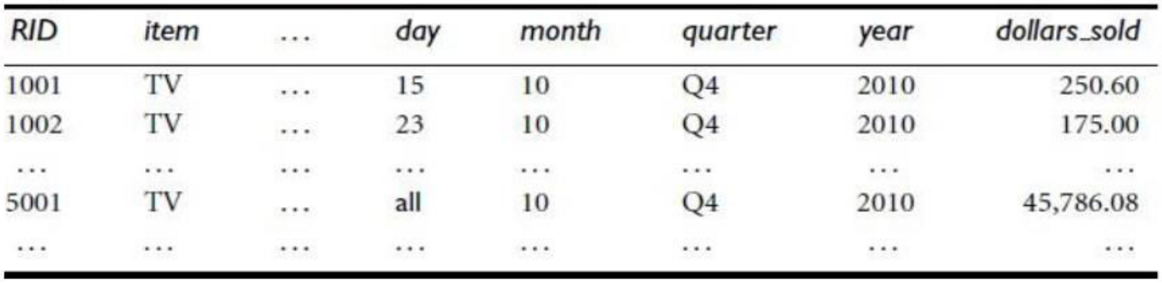
- To meet the growing demand of OLAP processing in relational databases, some database system vendors implement specialized SQL servers that provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

  Example:



A ROLAP data store. Table 4.4 shows a summary fact table that contains both base fact data and aggregated data. The schema is "⟨*record_identifier (RID), item, ..., day, month, quarter, year, dollars_sold*⟩," where *day, month, quarter,* and *year* define the sales date, and *dollars_sold* is the sales amount. Consider the tuples with an *RID* of 1001 and 1002, respectively. The data of these tuples are at the base fact level, where the sales dates are October 15, 2010, and October 23, 2010, respectively. Consider the tuple with an *RID* of 5001. This tuple is at a more general level of abstraction than the tuples 1001 and 1002. The *day* value has been generalized to all, so that the corresponding *time* value is October 2010. That is, the *dollars_sold* amount shown is an aggregation representing the entire month of October 2010, rather than just October 15 or 23, 2010. The special value all is used to represent subtotals in summarized data. ■

Single Table for Base and Summary Facts



RID	item	...	day	month	quarter	year	dollars_sold
1001	TV	...	15	10	Q4	2010	250.60
1002	TV	...	23	10	Q4	2010	175.00
...	...	...	...	...	...	...	...
5001	TV	...	all	10	Q4	2010	45,786.08
...	...	...	...	...	...	...	...

Q. 5 a) Explain the concept of Materialization for the selected computation of cuboids

1.No materialization: Do not pre-compute any of the "non base" cuboids. This leads to computing expensive multidimensional aggregates on-the-fly, which can be extremely slow.
2.Full materialization: Pre-compute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice typically requires huge amounts of memory space in order to store all of the pre-computed cuboids.
3.Partial materialization: Selectively compute a proper subset of the whole set of possible cuboids. Alternatively, we may compute a subset of the cube, which contains only those cells that satisfy some user-specified criterion, such as where the tuple count of each cell is above some threshold
The partial materialization of cuboids or subcubes should consider three factors:
(1) identify the subset of cuboids or subcubes to materialize;
(2) exploit the materialized cuboids or subcubes during query processing; and

(3) efficiently update the materialized cuboids or subcubes during load and refresh

Q. 5 b) Write a short note on compute cube operator and curse of dimensionality

A **cube operator** on n dimensions is equivalent to a collection of group-by statements, one for each subset of the n dimensions
Ex:- define cube sales_cube [city, item, year]: sum(sales in dollars)
Curse Of Dimensionality:-
How many cuboids in an n-dimensional cube with L levels?
• If there were no hierarchies associated with each dimension, then the total number of cuboids for an n-dimensional data cube, as we have seen, is 2^n . However, in practice, many dimensions do have hierarchies.
For example, time is usually explored not at only one conceptual level (e.g., year), but rather at multiple conceptual levels such as in the hierarchy "day < month < quarter < year."

$$Total\ number\ of\ cuboids = \prod_{i=1}^{n} (L_i + 1),$$

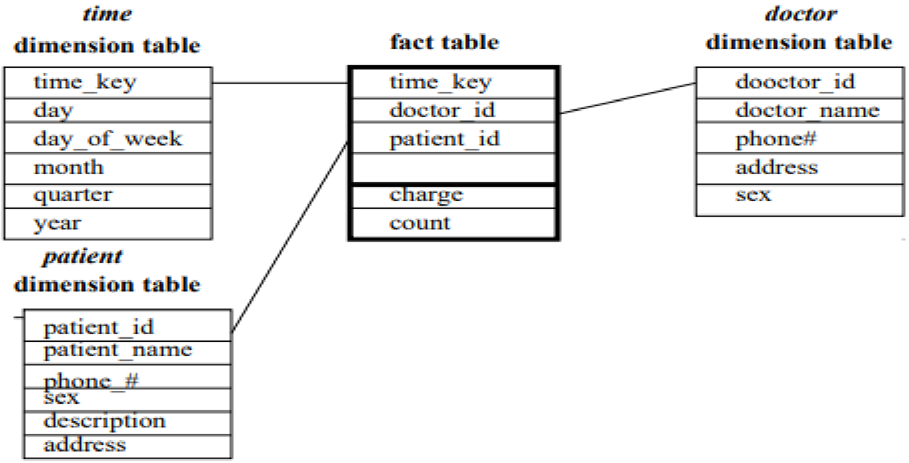
Q. 6 a) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

a. Enumerate three classes of schemas that are popularly used for modelling data warehouses using Star Schema.
b. Draw star and snowflake schema diagram for the above data warehouse.
c. 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 2010.

To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

a. Enumerate three classes of schemas that are popularly used for modelling data warehouses using Star Schema.
   Scheme:- Star Schema Definition – 1 Mark
   Solution:-A fact table in the middle connected to a set of dimension tables.
b. Draw star and snowflake schema diagram for the above data warehouse.

Scheme:- Star and Snowflake Schema for Doctor Warehouse- 4Marks



a. 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 2010.
   Scheme:-Defining OLAP Operations for cuboid – 3 Marks
   Solution:-The operations to be performed are:
   • Roll-up on time from day to year.
   • Slice for time = 2010.
   • Roll-up on patient from individual patient to all.
b. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).
   Scheme:-Writing SQL Query – 2 Marks
   Solution:-

select doctor, SUM(charge) from fee where year = 2010 group by doctor