Name of Faculty: Prof. Puneet Nema

Designation: Assistant Professor

Department: CSE

Subject: Data mining

Unit: II

Topic: OLAP, Characteristics of OLAP System, Motivation for using OLAP, Multidimensional View and Data Cube, Data Cube Implementations, Data Cube Operations, Guidelines for OLAP Implementation, Difference between OLAP & OLTP, OLAP Servers:- ROLAP, MOLAP, HOLAP Queries.

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula
Computer Science and
Engineering,VIII-Semester

CS-8003 Data Mining

UNIT-II

Topic Covered:

OLAP, Characteristics of OLAP System, Motivation for using OLAP, Multidimensional View and Data Cube, Data Cube Implementations, Data Cube Operations, Guidelines for OLAP Implementation, Difference between OLAP & OLTP, OLAP Servers:- ROLAP, MOLAP, HOLAP Queries.

OLAP:

OLAP (Online Analytical Processing) is the technology support the multidimensional view of data for many Business Intelligence (BI) applications. OLAP provides fast, steady and proficient access, powerful technology for data discovery, including capabilities to handle complex queries, analytical calculations, and predictive "what if" scenario planning.

OLAP is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making.

Characteristics of OLAP System :

The need for more intensive decision support prompted the introduction of a new generation of tools. Generally used to analyze the information where huge amount of historical data is stored. Those new tools, called online analytical processing (OLAP), create an advanced data analysis environment that supports decision making, business modeling, and operations research.

Its four main characteristics are:

1. Multidimensional Data Analysis Techniques:

Multidimensional analysis are inherently representative of an actual business model. The most distinctive characteristic of modern OLAP tools is their capacity for multidimensional analysis (for example actual vs budget). In multidimensional analysis, data are processed and viewed as part of a multidimensional structure. This type of data analysis is particularly attractive to business decision makers because they tend to view business data as data that are related to other business data.

2. Advanced Database Support:

For efficient decision support, OLAP tools must have advanced data access features. Access to many different kinds of DBMSs, flat files, and internal and external data sources.

- Access to aggregated data warehouse data as well as to the detail data found in operational databases.
- Advanced data navigation features such as drill-down and roll-up.
- Rapid and consistent query response times.
- The ability to map end-user requests, expressed in either business or model terms, to the appropriate data source and then to the proper data access language (usually SQL).
- Support for very large databases. As already explained the data warehouse can easily and quickly grow to multiple gigabytes and even terabytes.

3. Easy-to-Use End-User Interface:

Advanced OLAP features become more useful when access to them is kept simple. OLAP tools have equipped their sophisticated data extraction and analysis tools with easy-to-use graphical interfaces. Many of the interface features are "borrowed" from previous generations of data analysis tools that are already familiar to end users. This familiarity makes OLAP easily accepted and readily used.

4. Client/Server Architecture:

Conform the system to the principals of Client/server architecture to provide a framework within which new systems can be designed, developed, and implemented. The client/server environment enables an OLAP system to be divided into several components that define its architecture. Those components can then be placed on the same computer, or they can be distributed among several computers. Thus, OLAP is designed to meet ease-of-use requirements while keeping the system flexible.

Motivation for using OLAP:

I). Understanding and improving sales: For an enterprise that has many products and uses a number of channels for selling the products, OLAP can assist in finding the most popular products and the most popular channels. In some cases it may be possible to find the most profitable customers.

II). Understanding and reducing costs of doing business: Improving sales is one aspect of improving a business, the other aspect is to analyze costs and to control them as much as possible without affecting sales. OLAP can assist in analyzing the costs associated with sales.

Multidimensional View and Data Cube

Multidimensional Views:

The ability to quickly switch between one slice of data and another allows users to analyze their information in small palatable chunks instead of a giant report that is confusing.
Looking at data in several dimensions; for example, sales by region, sales by sales rep, sales by product category, sales by month, etc. Such capability is provided in numerous decision support applications under various function names. Multidimensional approach that time is an important dimension, and that time can have many different attributes. For example, in a spreadsheet or database, a pivot table provides these views and enables quick switching between them.

Data Cube:

A data cube is generally used to easily interpret data. It is especially useful when representing data together with dimensions as certain measures of business requirements. A cube's every dimension represents certain characteristic of the database, for example, daily, monthly or yearly sales. The data included inside a data cube makes it possible analyze almost all the figures for virtually any or all customers, sales agents, products, and much more. Thus, a data cube can help to establish trends and analyze performance.

Data cubes are mainly categorized into two categories:

- **Multidimensional Data Cube:** Most OLAP products are developed based on a structure where the cube is patterned as a multidimensional array. These multidimensional OLAP (MOLAP) products usually offers improved performance when compared to other approaches mainly because they can be indexed directly into the structure of the data cube to gather subsets of data. When the number of dimensions is greater, the cube becomes sparser. That means that several cells that represent particular attribute combinations will not contain any aggregated data. This in turn boosts the storage requirements, which may reach undesirable levels at times, making the MOLAP solution untenable for huge data sets with many dimensions. Compression techniques might help; however, their use can damage the natural indexing of MOLAP.

- **Relational OLAP:** Relational OLAP make use of the relational database model. The ROLAP data cube is employed as a bunch of relational tables (approximately twice as many as the quantity of dimensions) compared to a multidimensional array. Each one of these tables, known as a cuboid, signifies a specific view.

Data Cube Implementations :

Cube implementation involves the procedures of computation, storage, and manipulation of a data cube, which is a disk structure that stores the results of the aggregate queries that group the tuples of a fact table on all possible combinations of its dimension attributes. Let us take example assuming that R is a fact table that consists of three dimensions (A, B, C) and one measure M (see definitional entry for *Measure*), the corresponding cube of R. Each cube node (i.e., view that belongs to the data cube) stores the results of a particular aggregate query. Clearly, if D denotes the number of dimensions of a fact table, the number of all possible aggregate queries is 2^D; hence, in the worst case, the size of the data cube is exponentially larger with respect to D than the size of the original fact table. In typical applications, this may be in the order of gigabytes or even more.

Data Cube Operations :

The most popular end user operations on dimensional data are:

Roll up:

The roll-up operation (also called drill-up or aggregation operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by climbing down a concept hierarchy, i.e. dimension reduction. Let me explain roll up with an example:
Consider the following cube illustrating temperature of certain days recorded weekly:

temperature	64	65	68	69	70	71	72	75	80	81	83	85
week 1	1	0	1	0	1	0	0	0	0	0	1	0
week 2	0	0	0	1	0	0	1	2	0	1	0	0

Figure 2.1: Example data for Roll-up

Assume we want to set up levels (hot(80-85), mild(70-75), cold(64-69)) in temperature from the above cube. To do this we have to group columns and add up the values according to the concept hierarchy. This operation is called roll-up. By doing this we obtain the following cube.

temperature	cool	mild	hot
week 1	2	1	1
week 2	1	3	1

Figure 2.2: Rollup.

The concept hierarchy can be defined as hot-->day-->week. The roll-up operation groups the data by levels of temperature.

Roll Down:

The roll down operation (also called drill down) is the reverse of roll up. It navigates from less detailed data to more detailed data. It can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Drill down adds more detail to the given data, it can also be performed by adding new dimensions to a cube. Performing roll down operation on the same cube mentioned above:

The result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for temperature can be defined as day<--week<--cool. Drill-down occurs by descending the time hierarchy from the level of week to the more detailed level of day. Also new dimensions can be added to the cube, because drill-down adds more detail to the given data.

	cool	mild	hot
day 1	0	0	0
day 2	0	0	0
day 3	0	0	1
day 4	0	1	0
day 5	1	0	0
day 6	0	0	0
day 7	1	0	0
day 8	0	0	0
day 9	1	0	0
day 10	0	1	0
day 11	0	1	0
day 12	0	1	0
day 13	0	0	1
day 14	0	0	0

Figure 2.3: Roll down.

Slicing:

A Slice is a subset of multidimensional array corresponding to a single value for one or more members of the dimensions. Slice performs a selection on one dimension of the given cube, thus resulting in a subcube. For example, in the cube example above, if we make the selection, temperature=cool we will obtain the following cube:

	cool
day 1	0
day 2	0
day 3	0
day 4	0
day 5	1
day 6	0
day 7	1
day 8	0
day 9	1
day 10	0
day 11	0
day 12	0
day 13	0
day 14	0

Figure 2.4: Slicing.

Dicing:

A related operation to slicing is dicing. The dice operation defines a subcube by performing a selection on two or more dimensions. For example, applying the selection (time = day 3 OR time = day 4) AND (temperature = cool OR temperature = hot) to the original cube we get the following subcube (still two-dimensional): Dicing provides you the smallest available slice.

	cool	hot
day 3	0	1
day 4	0	0

Figure 2.5: Dicing

Pivot/Rotate:

Pivot or rotate is a visualization operation that rotates the data axes in view in order to provide an alternate presentation of the data. Rotating changes the dimensional orientation of the cube, i.e. rotates the data axes to view the data from different perspectives. Pivot groups data with different dimensions. The below cubes shows 2D represntation of Pivot.
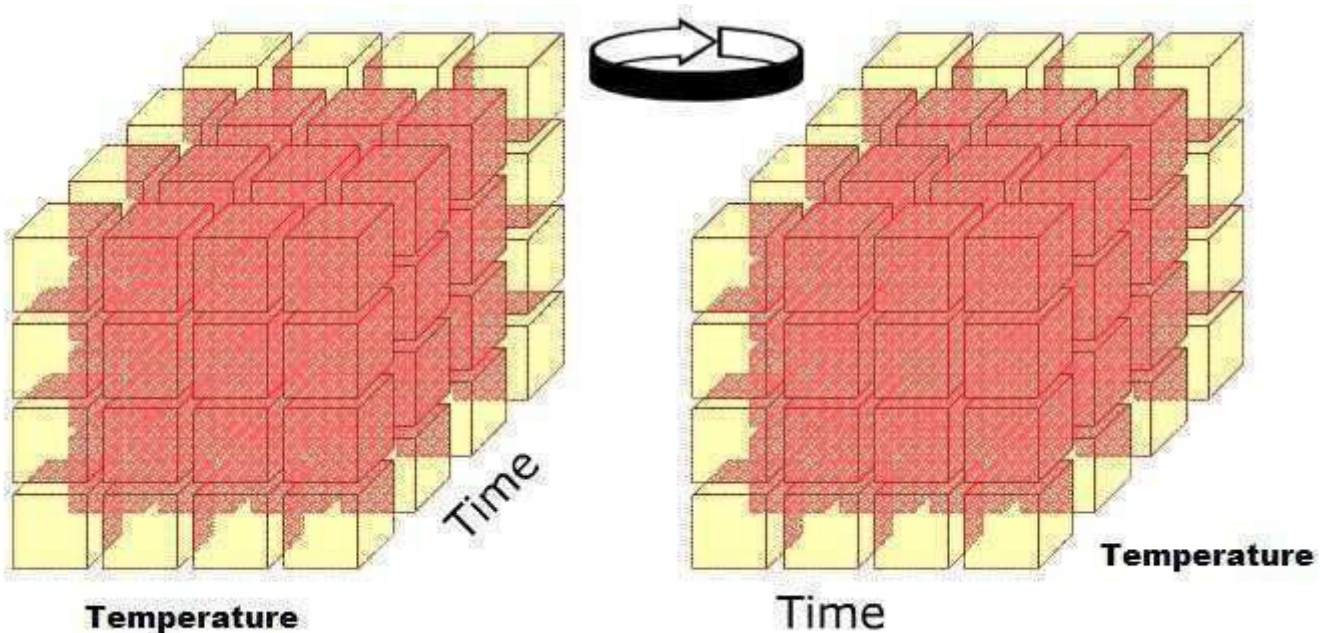
Figure 2.6: Pivot

Some more OLAP operations include:

SCOPING: Restricting the view of database objects to a specified subset is called scoping. Scoping will allow users to receive and update some data values they wish to receive and update.

SCREENING: Screening is performed against the data or members of a dimension in order to restrict the set of data retrieved.

DRILL ACROSS: Accesses more than one fact table that is linked by common dimensions. Combiens cubes that share one or more dimensions.

DRILL THROUGH: Drill down to the bottom level of a data cube down to its back end relational tables.

Guidelines for OLAP Implementation

Following are a number of guidelines for successful implementation of OLAP. The guidelines are, somewhat similar to those presented for data warehouse implementation.

1. Vision: The OLAP team must, in consultation with the users, develop a clear vision for the OLAP system. This vision including the business objectives should be clearly defined, understood, and shared by the stakeholders.

2. Senior management support: The OLAP project should be fully supported by the senior managers and multidimensional view of data. Since a data warehouse may have been developed already, this should not be difficult.

3. Selecting an OLAP tool: The OLAP team should familiarize themselves with the ROLAP and MOLAP tools available in the market. Since tools are quite different, careful planning may be required in selecting a tool that is appropriate for the enterprise. In some situations, a combination of ROLAP and MOLAP may be most effective.

4. Corporate strategy: The OLAP strategy should fit in with the enterprise strategy and business objectives. A good fit will result in the OLAP tools being used more widely.

5. Focus on the users: The OLAP project should be focused on the users. Users should, in consultation with the technical professional, decide what tasks will be done first and what will be done later. Attempts should be made to provide each user with a tool suitable for that person's skill level and information needs. A good GUI user interface should be provided to non-technical users. The project can only be successful with the full support of the users.

6. Joint management: The OLAP project must be managed by both the IT and business professionals. Many other people should be involved in supplying ideas. An appropriate committee structure may be necessary to channel these ideas.

7. Review and adapt: As noted in last chapter, organizations evolve and so must the OLAP systems. Regular reviews of the project may be required to ensure that the project is meeting the current needs of the enterprise.

OLTP vs. OLAP

1. Transaction oriented / Subject Oriented
2. High create, read, update delete activity / High Read activity
3. Many users / Few Users
4. Real time information / Historical Information
5. Operational Database / Information Database.
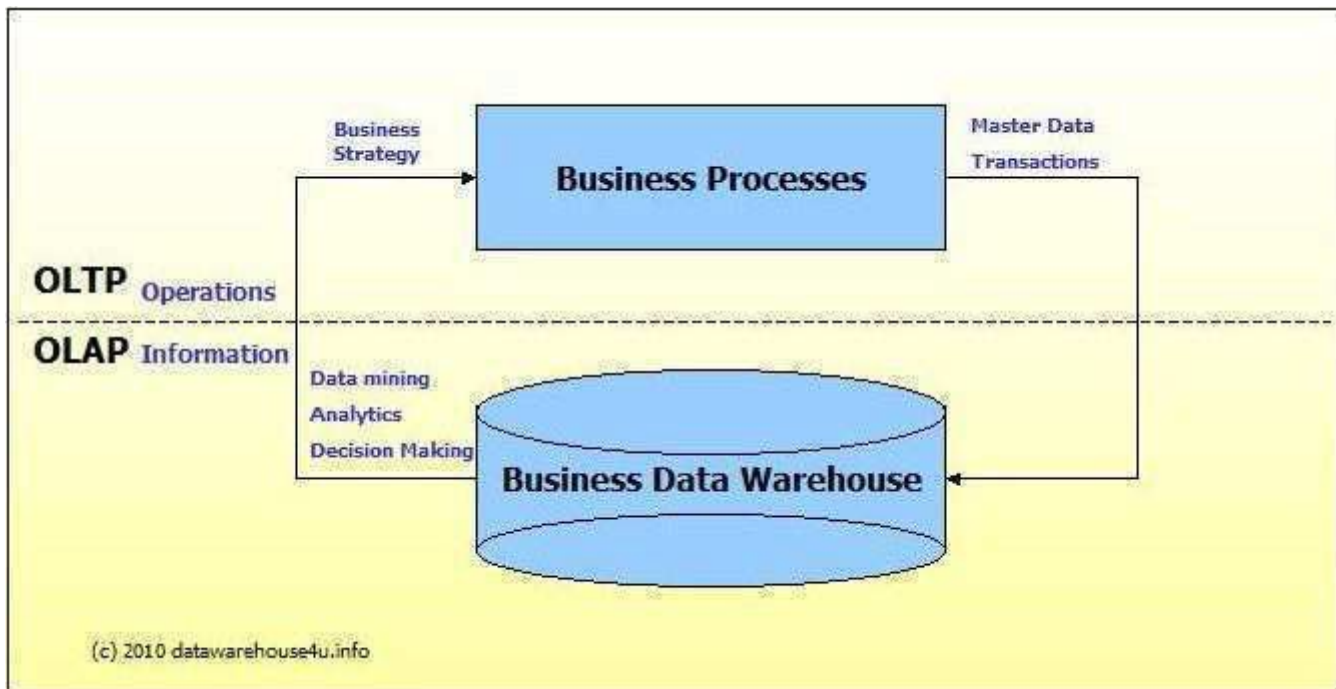
Figure 2.7: OLAP vs OLTP

OLTP (On-line Transaction Processing)

Using high transaction volumes at a time and high volatile data. Is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. In OLTP database there is detailed

and

current data, and schema used to store transactional databases is the entity model (usually 3NF). Uses
complex database designs used by IT panel.

- OLAP (On-line Analytical Processing)

Low transaction volumes using many records at a time. It is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response
time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star
schema).

The following table summarizes the major differences between OLTP and OLAP system design.

Topic	OLTP System - Online Transaction Processing (Operational System)	OLAP System - Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	OLAP: Periodic long-running batch jobs refresh the data

Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
DatabaseDesign	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

OLAP Servers

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

Types of OLAP Servers -
We have four types of OLAP servers −
1. Relational OLAP (ROLAP)
2. Multidimensional OLAP (MOLAP)
3. Hybrid OLAP (HOLAP)
4. Specialized SQL Servers

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.
ROLAP includes the following −

1. Implementation of aggregation navigation logic.
2. Optimization for each DBMS back end.
3. Additional tools and services.
4. Can handle large amounts of data
5. Performance can be slow

Since ROLAP uses a relational database, it requires more processing time and/or disk space to perform some of the tasks that multidimensional databases are designed for. However, ROLAP supports larger user groups and greater amounts of data and is often used when these capacities are crucial, such as in a large and complex department of an enterprise.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. Multidimensional data stores
The storage utilization may be low if the data set is sparse.
MOLAP server use two levels of data storage representation to handle dense and sparse data sets.Using a MOLAP, a user can use multidimensional view data with different facets. Multidimensional data analysis is also possible if a relational database is used. By that would require querying data from multiple tables. On the contrary, MOLAP has all possible combinations of data already stored in a multidimensional array. MOLAP can access this data directly. Hence, MOLAP is faster compared to Relational Online Analytical Processing (ROLAP).

Hybrid OLAP

Hybrid OLAP technologies attempt to combine the advantages of MOLAP and ROLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

HOLAP can use varying combinations of ROLAP and OLAP technology. It typically stores data in both a relational database and a multidimensional database, depending on the preferred type of processing. The databases are used to store data in the most functional way possible. For heavy data processing, the data is more efficiently stored in a relational database, whereas multidimensional bases are used for speculative processing.

Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment