What is Data Warehouse?

Data Warehousing is a technique that is mainly used to collect and manage data from various sources to give the business a meaningful business insight. A data warehouse is specifically designed to support management decisions.

In simple terms, a data warehouse defines a database that is maintained independently from an organization’s operational databases. Data warehouse systems enable the integration of multiple application systems. They provide data processing by offering a solid platform of consolidated, historical information for analysis.

Data warehouses generalize and centralize data in multidimensional space. The construction of data warehouses contains data cleaning, data integration, and data transformation and can be looked at as an important preprocessing step for data mining.

It provides online analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining. There are several data mining functions, including association, classification, prediction, and clustering can be integrated with OLAP operations to build up interactive mining of knowledge at various levels of abstraction.

There are three main types of Data Warehouses which are as follows −

Enterprise Data Warehouse (EDW) − Enterprise Data Warehouse is a centralized warehouse. It is used for organizing and representing the data. With the help of EDW, the user can classify data based on the subject.

Operational Data Store − In Operational Data Store, the Data warehouse is refreshed in real-time. Thus, it is more generally used for routine activities including storing records, etc.

Data Mart − A data mart can be defined as a subset of the data warehouse. It is designed for sales, finance, and so on.

What are the elements of a data warehouse system?

There are various elements of a data warehouse system which are as follows −

Source System − An operational system of data whose service it is to capture the transactions of the business. A source system is known as a “legacy system” in a mainframe environment.

The features of the source system are uptime and availability. Queries opposite to source systems are definite, "account-based" queries that are elements of the normal transaction flow and firmly restricted in their demands on the legacy system.

Data Staging Area − A storage area and group of processes that simple, transform, combine, de-duplicate, household, archive, and produce source records for use in the data warehouse.

The data staging area is dominated by the smooth activities of sorting and sequential processing and the data staging area does not need to be based on relational technology. After it can check the data for conformance with all the one-to-one and many-to-one business rules it has defined, it can be pointless to take the last phase of building a completely blown entity-relation-based physical database design.

Business Process − A coherent group of business activities that create sense to the business users of our data warehouses. A business process is generally a group of activities such as "order processing" or "user pipeline management," but business processes can overlap, and absolutely the definition of a single business process will develop over time.

Reporting − The data in the data warehouse should apply to the organization's staff if the data warehouse is to be beneficial. There is a large number of software applications that implement this function, or reporting can be custom-developed.

There are various reporting tools are as follows −

Business intelligence tools − These are software applications that analyze the process of development and management of business documents based on data warehouse data.

Executive information systems (known more widely as Dashboard (business) − These are software applications that can show complex business metrics and data graphically to enable rapid understanding.

Data Mining − Data mining tools are software that allows users to implement detailed numerical and statistical computations on detailed data warehouse data to identify trends, identify a pattern and interpret data.

Metadata − Metadata is data about the data which is required by the users. It can be used not only to inform operators and users of the data warehouse about its condition and the data held within the data warehouse but as an integration of incoming data and a tool to upgrade and clarify the fundamental data warehouse model.

Operations − A data warehouse operation includes the processes of loading, manipulating, and extracting data from the data warehouse. Operations also protect user administration, security, capacity management, and associated services.

What are the components of a data warehouse?

The major components of a data warehouse are as follows −

Data Sources − Data sources define an electronic repository of records that includes data of interest for administration use or analytics. The mainframe of databases (e.g. IBM DB2, ISAM, Adabas, Teradata, etc.), client-server databases (e.g. Teradata, IBM DB2, Oracle database, Informix, Microsoft SQL Server, etc.), PC databases (e.g. Microsoft Access, Alpha Five), spreadsheets (e.g. Microsoft Excel) and any other electronic storage of data.

Data Warehouse − The data warehouse is normally a relational database. It should be organized to hold data in a structure that best supports not only query and documenting but also advanced analysis techniques, such as data mining.

Reporting − The data in the data warehouse must be available to the organisations staff if the data warehouse is to be useful. There is a huge number of software applications that execute this function, or reporting can be custom-developed. Reporting tools includes are as follows:

- **Business intelligence tools** − These are software applications that clarify the process of development and production of business documents based on data warehouse information.
- **Executive information systems (known more widely as Dashboard (business)** − These are software applications that are used to display complex business metrics and information graphically to allow rapid understanding.
- **Data Mining** − Data mining tools are software that enables users to implement detailed numerical and statistical calculations on detailed data warehouse data to detect trends, identify design and analyze data.

Metadata − Metadata is data about the data which is needed by the users. It is used not only to instruct operators and users of the data warehouse about its status and the data held inside the data warehouse but also as a means of integration of incoming information and a tool to upgrade and refine the basic data warehouse model.

Operations − A data warehouse operation is consists of the processes of loading, manipulating, and extracting information from the data warehouse. Operations also cover user management, security, capacity management, and related functions.

Optional Components − There are the following components exist in some data warehouses which are as follows −

- **Dependent Data Marts** − A dependent data mart is a physical database (either on the same hardware as the data warehouse or on a separate hardware platform) that receives all its information from the data warehouse.

- **Logical Data Marts** − A logical data mart is a filtered view of the main data warehouse but does not physically exist as an independent data copy.
- **Operational Data Store** − An ODS is an integrated database of operational data. Its sources contain legacy systems, and it includes current or near-term information.

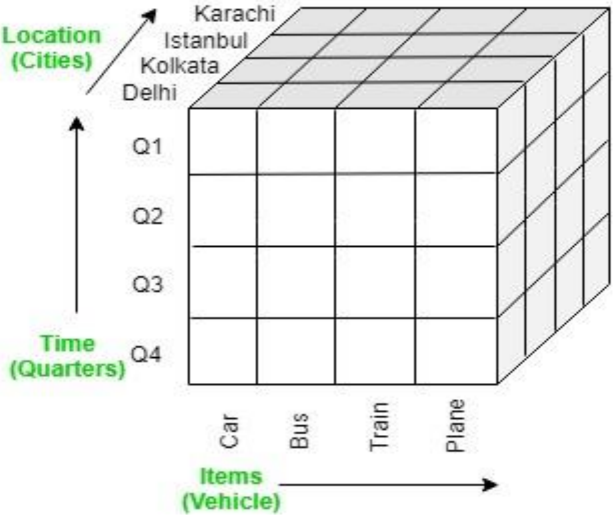
Characteristics of Data Warehouse

There are various characteristics of a data warehouse which are as follows −

- **Subject-oriented** − A data warehouse targets the modeling and analysis of information for decision-makers. Thus, data warehouses generally provide a simple and concise view of specific subject issues by excluding information that is not beneficial in the decision support process.
- **Integrated** − As the data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records, the data cleaning and data integration techniques need to be used to provide consistency in naming conventions, encoding mechanisms, attribute measures, etc.
- **Time-variant** − Data is saved to provide data from a historical perspective (e.g., the past 5-10 years). Each key mechanism in the data warehouse includes, either implicitly or explicitly, an element of time.
- **Non-volatile** − A data warehouse is always a physically independent store of data transformed from the software data found in the operational environment. Because of this separation, a data warehouse does not need transaction processing, recovery, and concurrency control structure. It usually requires only two operations in data accessing − initial loading of data and access of data.

OLAP

OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.



OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:

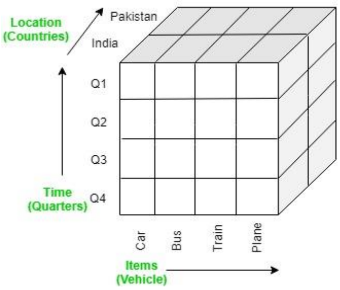
- Moving down in the concept hierarchy
- Adding a new dimension

In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:

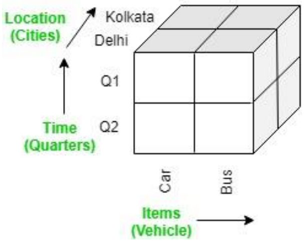
- Climbing up in the concept hierarchy
- Reducing the dimensions

In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).

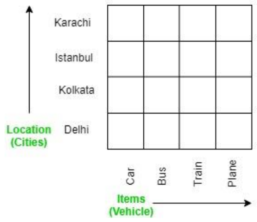


3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:

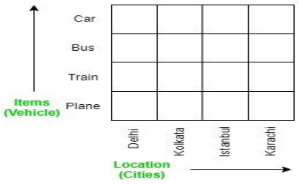
- Location = “Delhi” or “Kolkata”
- Time = “Q1” or “Q2”
- Item = “Car” or “Bus”



4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = “Q1”.

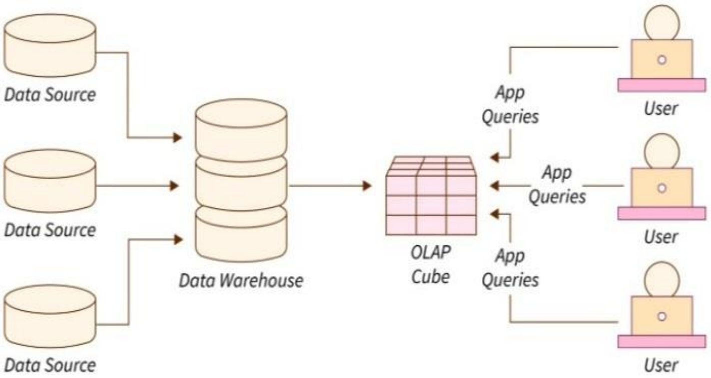


5. Pivot: It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



Online Analytical Processing Server (OLAP)

Online Analytical Processing Server (OLAP) is a software. Users can analyze information from many different databases all at once. It uses a multidimensional data model where users can ask questions based on multiple dimensions at the same time. OLAP databases are split up into cubes, which are also called hyper-cubes.



OLAP Servers

Online Analytical Processing(OLAP) refers to a set of software tools used for data analysis in order to make business decisions. OLAP provides a platform for gaining insights from databases retrieved from multiple database systems at the same time. It is based on a multidimensional data model, which enables users to extract and view data from various perspectives. A multidimensional database is used to store OLAP data. Many Business Intelligence (BI) applications rely on OLAP technology.

Type of OLAP servers:

The three major types of OLAP servers are as follows:

- **ROLAP**
- **MOLAP**
- **HOLAP**

Relational OLAP (ROLAP):

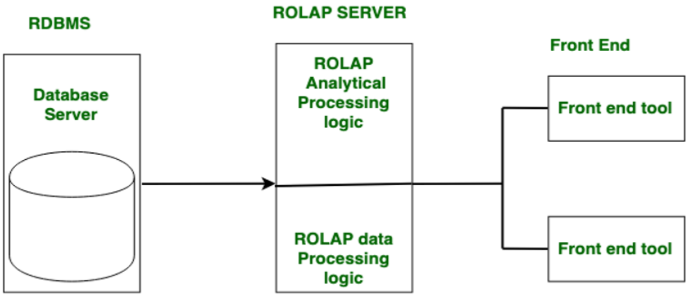
Relational On-Line Analytical Processing (ROLAP) is primarily used for data stored in a relational database, where both the base data and dimension tables are stored as relational tables. ROLAP servers are used to bridge the gap between the relational back-end server and the client's front-end tools. ROLAP servers store and manage warehouse data using RDBMS, and OLAP middleware fills in the gaps.

Benefits:

- It is compatible with data warehouses and OLTP systems.
- The data size limitation of ROLAP technology is determined by the underlying RDBMS. As a result, ROLAP does not limit the amount of data that can be stored.

Limitations:

- SQL functionality is constrained.
- It's difficult to keep aggregate tables up to date.



Multidimensional OLAP (MOLAP):

Through array-based multidimensional storage engines, Multidimensional On-Line Analytical Processing (MOLAP) supports multidimensional views of data. Storage utilization in multidimensional data stores may be low if the data set is sparse.

MOLAP stores data on discs in the form of a specialized multidimensional array structure. It is used for OLAP, which is based on the arrays' random access capability. Dimension instances determine array elements, and the data or measured value associated with each cell is typically stored in the corresponding array element. The multidimensional array is typically stored in MOLAP in a linear allocation based on nested traversal of the axes in some predetermined order.

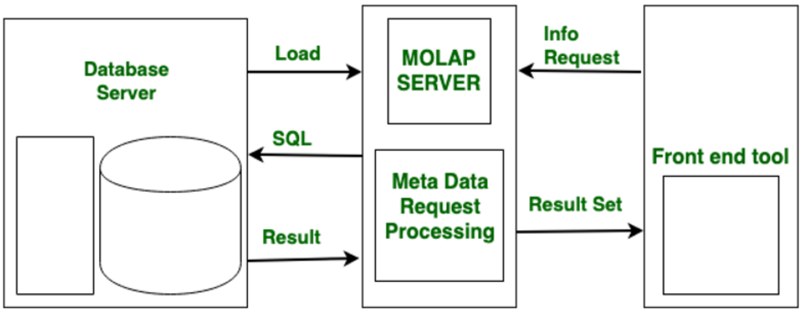
However, unlike ROLAP, which stores only records with non-zero facts, all array elements are defined in MOLAP, and as a result, the arrays tend to be sparse, with empty elements occupying a larger portion of them. MOLAP systems typically include provisions such as advanced indexing and hashing to locate data while performing queries for handling sparse arrays, because both storage and retrieval costs are important when evaluating online performance. MOLAP cubes are ideal for slicing and dicing data and can perform complex calculations. When the cube is created, all calculations are pre-generated.

Benefits:

- Suitable for slicing and dicing operations.
- Outperforms ROLAP when data is dense.
- Capable of performing complex calculations.

Limitations:

- It is difficult to change the dimensions without re-aggregating.
- Since all calculations are performed when the cube is built, a large amount of data cannot be stored in the cube itself.



Hybrid OLAP (HOLAP):

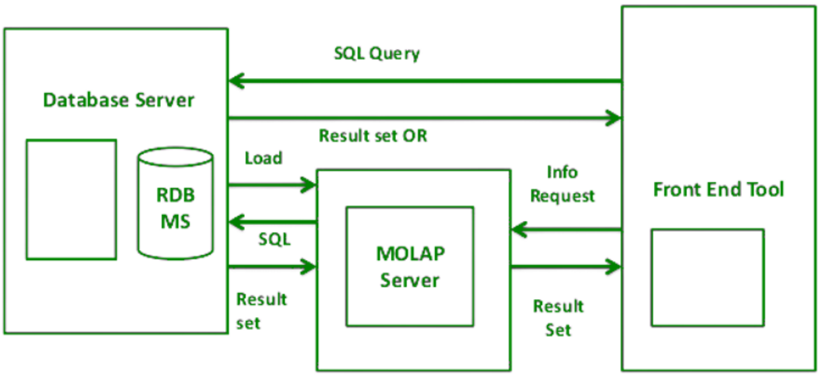
ROLAP and MOLAP are combined in Hybrid On-Line Analytical Processing (HOLAP). HOLAP offers greater scalability than ROLAP and faster computation than MOLAP.HOLAP is a hybrid of ROLAP and MOLAP. HOLAP servers are capable of storing large amounts of detailed data. On the one hand, HOLAP benefits from ROLAP's greater scalability. HOLAP, on the other hand, makes use of cube technology for faster performance and summary-type information. Because detailed data is stored in a relational database, cubes are smaller than MOLAP.

Benefits:

- HOLAP combines the benefits of MOLAP and ROLAP.
- Provide quick access at all aggregation levels.

Limitations

- Because it supports both MOLAP and ROLAP servers, HOLAP architecture is extremely complex.
- There is a greater likelihood of overlap, particularly in their functionalities.


SQL Query
Database Server
Result set OR
Load
RDB MS
Info Request
Front End Tool
SQL
MOLAP Server
Result set
Result Set