Data Mining & Data Warehousing

Introduction:

Companies are desperate for strategic decision to counter fiercer competition, extends market, share & improve profitability. Information needed for strategic decision making is not readily available in spite of tons of data accumulated by enterprises. We need different types of decision support system to provide strategic information. The types of information needed for strategic decision making is different from that available from operational system. We need a new type of system environmental for the purpose of providing strategic information for analysis, discerning trends & monitoring performance.

What is an operational system & informational system?

e.g.:- a railway reservation system generated vast amount of data which each day on train bookings. It is probably used for audit purposes. But it can be effective used for strategic management.

Operational system are called as online transaction system(OLAP).

Informational system are called as online analytical processing system(OLAP).

The new system environment happens to be the new paradigm of data warehousing. This new environment is kept separate from the system environment supporting day to day operation.

What is data warehouse?

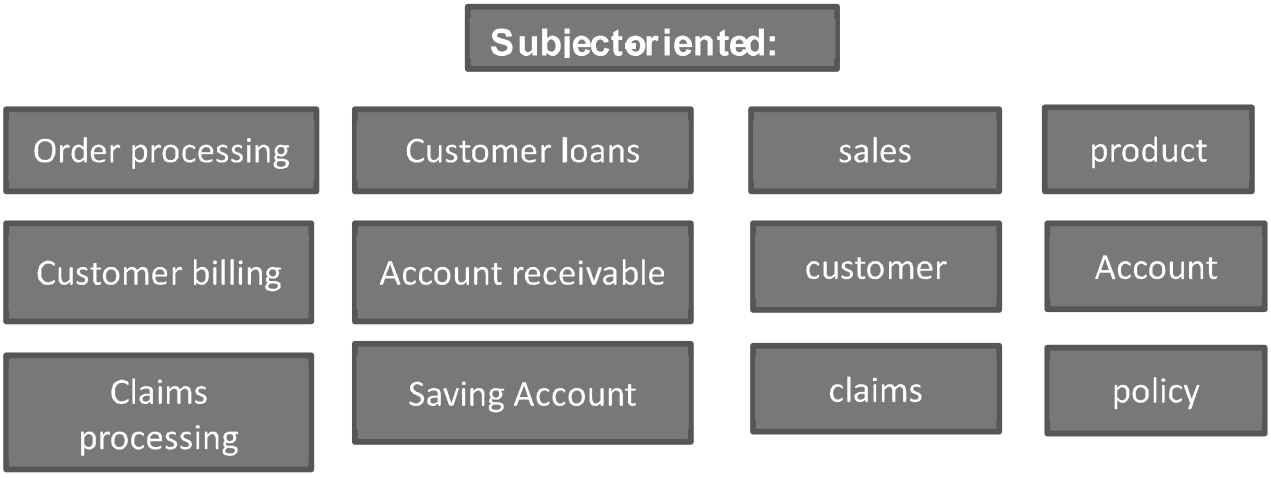
Data warehouse provides architecture & tools for business executives to systematically organize, understand, & use their data to make strategic decisions. Data warehousing is latest marketing weapon a way to retain customers by learning more about their need.

Definition:

A data warehouse is a subject oriented, integrated, time-variant & non-volatile collection of data in support of management's decision making process.

Above keyword distinguish data warehouse from other data repository system, such as relational database systems transaction processing system & file System.

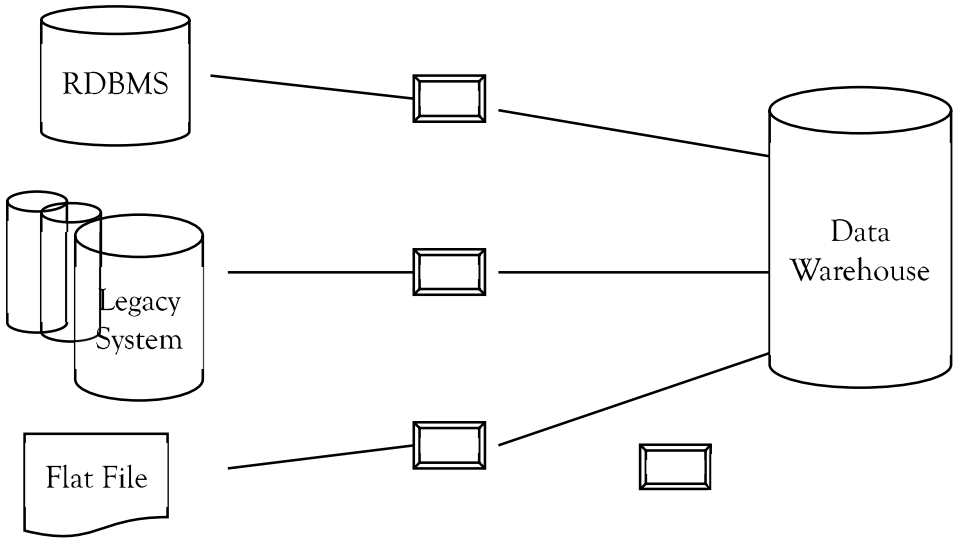
Subject Oriented



Data warehouse is organized around major subjects, such as customer, supplier, product & sales. Data warehouse typically provides a simple & concise view around particular subject issues by excluding data that are not useful in the decision support process.
i.e. data warehouse data is stored by subject, not by application.

Integrated

It is usually constructed by integrating multiple heterogeneous sources, such as relational database, flat files & on-line transaction records.



Time variant

Data are stored to provide information from historical perspective. Every key structure in data warehouse contains either implicitly or explicitly, an element of time. Time variant nature of the data in a data warehouse :-

-allows for analysis of the past.
-relates information to the present.
-enables forecasts for the future.

Non volatile

It is always a physically separate store of data transformed from the application data found in the operational environment. Data warehousing doesn't require transaction processing, recovery and control mechanisms. It actually requires only two operations in data accessing: initial loading of data and access of data.

The construction of data warehouse requires data cleaning data integration, & data consolidation the utilization of a data warehouse often necessitates a collection of decision support technologies, this allows "knowledge workers" to use the warehouse to quickly & conveniently obtain overview of the data, & to make sound decision based on information in the warehouse.

Many organization use this information to support business decision making activities including:

1. Increasing customer focus.
2. analyzing operations & looking for sources of profit.
3. Managing customer relationship making environmental corrections & managing the cost of corporate assets.

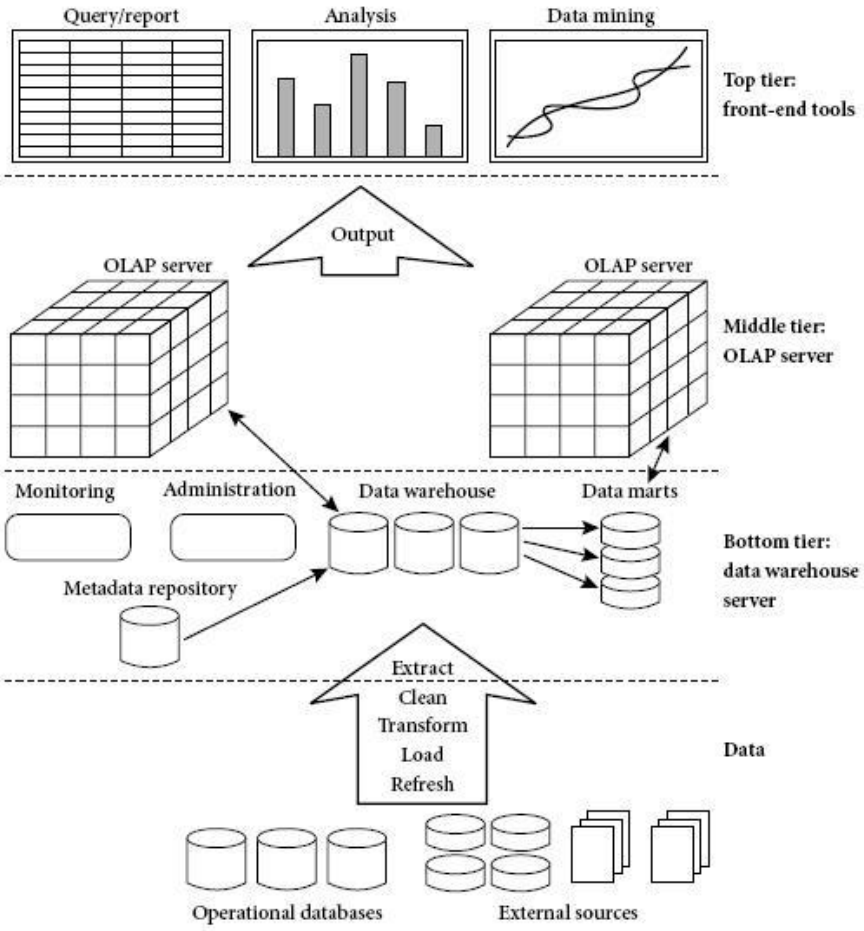
It is also useful from the point of view of heterogeneous database integration.

Data Warehouse architecture

1.The Bottom tier is a warehouse database server which is almost always a relationship database system. Back-end tools & utilities are used to feed data into the bottom tier from operational database or other external sources. These tools & utilities perform data extraction, cleaning & transformation as well as load & refresh functions to up to date the data warehouse. The tier also contains a metadata depository which stores information about the data warehouse & its contents.

2.Middle tier is an OLAP server that is typically implemented using either (a) a relational OLAP (ROLAP) model Or (b) a multidimensional OLAP (MOLAP) model.

3.The top tier front-end client layer, which contains query & reporting tools, analysis tools, & data mining tools.



3.12 A three-tier data warehousing architecture.

Data warehouse Model

from architecture point of view, there are three data warehouse models. The enterprise warehouse, the data mart, & virtual warehouse.

1.Enterprise data warehouse:

It collects all of the information about subject spanning the entire organization. It provides co-operate wide data integration, usually from one or more operational system or external information provides & is cross-functional in scope. It may implemented on traditional mainframes, computer super servers or parallel architecture platforms.

2.Data mart:

It contains a subject of corporate wide data that is of value to a specific group of users. Data mart are usually implemented on low cost departmental servers. Data mart can be categorized as independent or dependent.

3.Virtual warehouse:

It is set of views over operational databases. It is easy to build but requires excess capacity on operational servers.

Difference between OLTP & OLAP:-

OLTP	OLAP
1. It uses operational processing.	1. It uses informational processing.
2. Transaction oriented i.e. customer oriented.	2. Analysis oriented that is market oriented.
3. Used by clerk, DBA & Database professional.	3. Used by knowledge workers (manager, executive, analyst).
4. It is used for day to day operations.	4. It is used for long term informational requirements, decision support.
5. It uses E-R based data model & an application oriented database design.	5. It uses store /snowflake, model & subject oriented database design.
6. It focuses on the current data & guaranteed up to data.	6. It focuses on historical data in which accuracy is maintained over time.
7. Data is primitive & highly detailed.	7. Data in summarized & consolidated.
8. Access patterns of this system consist mainly of short, atomic transactions & also read/write	8. Accesses patterns of OLAP are read only of operations, although many could be complex queries.
9. It requires index or hash operation on primary key	9. It requires lots of scan operations.
10.Number of records accessed in tens	10.Number of records accessed in millions.
11.DB size is 100Mb to Gb	11.DB size is 100 Gb to Tb.

Components of Data Warehouse :-

Figure shows the basic components of a typical warehouse.

You see the ***Source Data*** component shown on the left.

The ***Data Staging*** component serves as the next building block.

In the middle, you see the ***Data Storage*** component that manages the data warehouse data. This component not only stores and manages the data, it also keeps track of the data by means of the metadata repository.

The ***Information Delivery*** component shown on the right consists of all the different ways of making the information from the data warehouse available to the users.

Source Data Component

Source data coming into the data warehouse may be grouped into four broad categories.

Production Data. This category of data comes from the various operational systems of the enterprise. Based on the information requirements in the data warehouse, you choose segments of data from the different operational systems. While dealing with this data, you come across many variations in the data formats. You also notice that the data resides on different hardware platforms. Further, the data is supported by different database systems and operating systems. This is data from many vertical applications.

Internal Data. In every organization, users keep their "private" spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse.

Internal data adds additional complexity to the process of transforming and integrating the data before it can be stored in the data warehouse. You have to determine strategies for collecting data from spreadsheets, find ways of taking data from textual documents, and tie into departmental databases to gather pertinent data from those sources. Again, you may want to schedule the acquisition of internal data.

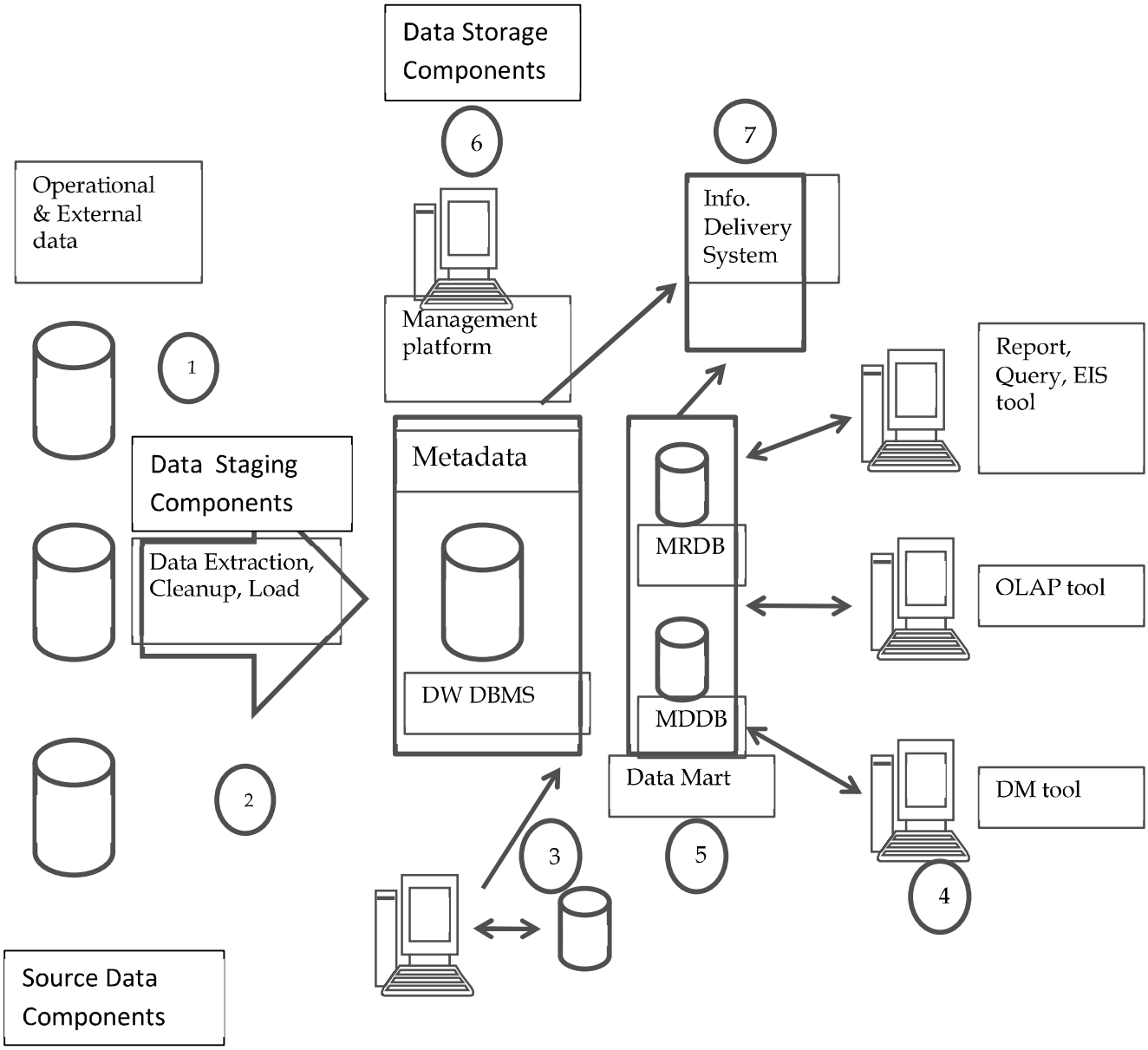


Fig : Components of Data Ware House

Archived Data. Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. The circumstances in your organization dictate how often and which portions of the operational databases are archived for storage. Some data is archived after a year. Sometimes data is left in the operational system databases for as long as five years.

For getting historical information, you look into your archived data sets. Depending on your data warehouse requirements, you have to include sufficient historical data. This type of data is useful for discerning patterns and analysing trends.

External Data. Most executives depend on data from external sources for a high percentage of the information they use. They use statistics relating to their industry produced by external agencies. They use market share data of competitors. They use standard values of financial indicators for their business to check on their performance.

Data Staging Component

After you have extracted data from various operational systems and from external sources, you have to prepare the data for storing in the data warehouse. The extracted data coming from several different sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.

Three major functions need to be performed for getting the data ready. You have to extract the data, transform the data, and then load the data into the data warehouse storage. These three major functions take place in a staging area. The data staging component consists of a workbench for these functions. Data staging provides a place and an area with a set of functions to clean, change, combine, convert, deduplicate, and prepare source data for storage and use in the data warehouse.

Data Extraction. This function has to deal with numerous data sources. You have to employ the appropriate technique for each data source. Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Some data may be on other legacy network and hierarchical data models. Many data sources may still be in flat files. You may want to include data from spreadsheets and local departmental data sets. Data extraction may become quite complex.

More frequently, data warehouse implementation teams extract the source into a separate physical environment from which moving the data into the data warehouse would be easier. In the separate environment, you may extract the source data into a group of flat files, or a data-staging relational database, or a combination of both.

Data Transformation. In every system implementation, data conversion is an important function. You perform a number of individual tasks as part of data transformation. First, you clean the data extracted from each source. Cleaning may just be correction of misspellings, or may include resolution of conflicts between state codes and zip codes in the source data, or may deal with providing default values for missing data elements, or elimination of duplicates when you bring in the same data from multiple source systems.

Standardization of data elements forms a large part of data transformation. You standardize the data types and field lengths for same data elements retrieved from the various sources. Data transformation involves many forms of combining pieces of data from the different sources.

When the data transformation function ends, you have a collection of integrated data that is cleaned, standardized, and summarized. You now have data ready to load into each data set in your data warehouse.

Data Loading. Two distinct groups of tasks form the data loading function. When you complete the design and construction of the data warehouse and go live for the first time, you do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time. As the data warehouse starts functioning, you continue to extract the changes to the source data, transform the data revisions, and feed the incremental data revisions on an on going basis.

Data Storage Component

The data storage for the data warehouse is a separate repository. In the data repository for a data warehouse, you need to keep large volumes of historical data for analysis. Further, you have to keep the data in the data warehouse in structures suitable for analysis, and not for quick retrieval of individual pieces of information. Therefore, the data storage for the data warehouse is kept separate from the data storage for operational systems.

When your analysts use the data in the data warehouse for analysis, they need to know that the data is stable and that it represents snapshots at specified periods. As they are working with the data, the data storage must not be in a state of continual updating. For this reason, the data warehouses are "read-only" data repositories.

Most of the data warehouses employ relational database management systems. Many of the data warehouses also employ multidimensional database management systems. Data extracted from the data warehouse storage is aggregated in many ways and the summary data is kept in the multidimensional databases (MDDBs). Such multidimensional database systems are usually proprietary products.

Information Delivery Component

In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery. The novice user comes to the data warehouse with no training and, therefore, needs prefabricated reports and preset queries. The casual user needs information once in a while, not regularly. This type of user also needs prepackaged information. The business analyst looks for ability to do complex analysis using the information in the data warehouse. The power user wants to be able to navigate throughout the data warehouse, pick up interesting data, format his or her own queries, drill through the data layers, and create custom reports and ad hoc queries.

Ad hoc reports are predefined reports primarily meant for novice and casual users. Provision for complex queries, multidimensional (MD) analysis, and statistical analysis cater to the needs of the business analysts and power users. Information fed into Executive Information Systems (EIS) is meant for senior executives and high-level managers. Some data warehouses also provide data to data-mining applications. Data-mining applications are knowledge discovery systems where the mining algorithms help you discover trends and patterns from the usage of your data.

Metadata Component

The metadata component is the data about the data in the data warehouse. It is useful to define data warehouse objects. It is similar to the data dictionary or the data catalogue in a database management system. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for timestamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

Types of Metadata

Metadata in a data warehouse fall into three major categories:

_ Operational Metadata

_ Extraction and Transformation Metadata

_ End-User Metadata

Operational Metadata. Data for the data warehouse comes from several operational systems of the enterprise. These source systems contain different data structures. The data elements selected for the data warehouse have various field lengths and data types. In selecting data from the source systems for the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths. When you deliver information to the end-users, you must be able to tie that back to the original source data sets. Operational metadata contain all of this information about the operational data sources.

Extraction and Transformation Metadata. Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction. Also, this category of metadata contains information about all the data transformations that take place in the data staging area.

End-User Metadata. The end-user metadata is the navigational map of the data warehouse. It enables the end-users to find information from the data warehouse. The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

Management and Control Component

This component of the data warehouse architecture sits on top of all the other components. The management and control component coordinates the services and activities within the data warehouse. This component controls the data transformation and the data transfer into the data warehouse storage. On the other hand, it moderates the information delivery to the users. It works with the database management systems and enables data to be properly stored in the repositories. It monitors the movement of data into the staging area and from there into the data warehouse storage itself.

The management and control component interacts with the metadata component to perform the management and control functions. As the metadata component contains information about the data warehouse itself, the metadata is the source of information for the management module.

Multidimensional Data Model:-

Date warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of data cube. Data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

Dimensions are the entities with respect to which an organization wants to keep records.

E.g. Sales data warehouse is created to keep records of stores sales with respect to the dimensions time, item, branch and location.

Each dimension may have table associated with it, called a **dimensional table**. E.g. dimension table for item may contain the attributes item_name, brand and type.

Facts are numerical measures, which are quantities by which we want to analyze relations hips between dimensions.

e.g. facts for a sales amount ,rupees, number of units sold amount_budgeted.

The **fact table** contains the name of facts, or measures, as well as keys to each of the related dimension tables.

In data warehousing the data cube is ***n-dimensional***. The data cube is metaphor for multidimensional data storage. The actual physical storage of such data may differ from its logical representation.

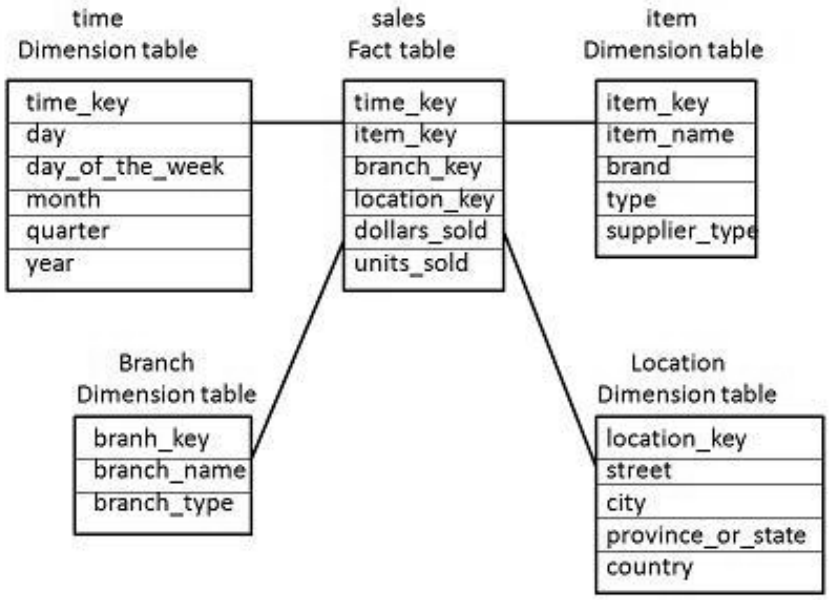
Schemas for multidimensional databases:

Multidimensional model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

1.Star schema : -

In star schema, the data warehouse contains (1) a large central table (Fact table) containing the bulk of the data, with or redundancy, and (2) a set of smaller attendant tables, one for each dimension. Each tuple in the fact table consists of a key pointing to each of the dimension tables that provide its multidimensional coordinates. It also stores numerical values for those attributes. The advantage of a star schema is that it is easy to understand, easy to define hierarchies, reduces the number of physical joins, requires low maintenance and very simple metadata.

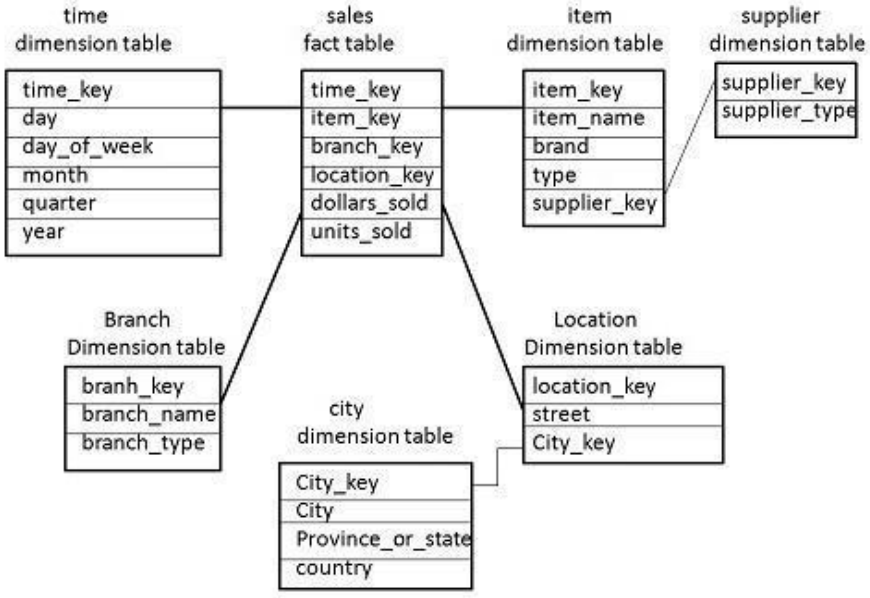
E.g. star schema of data warehouse for sales.



2. Snowflake schema: -

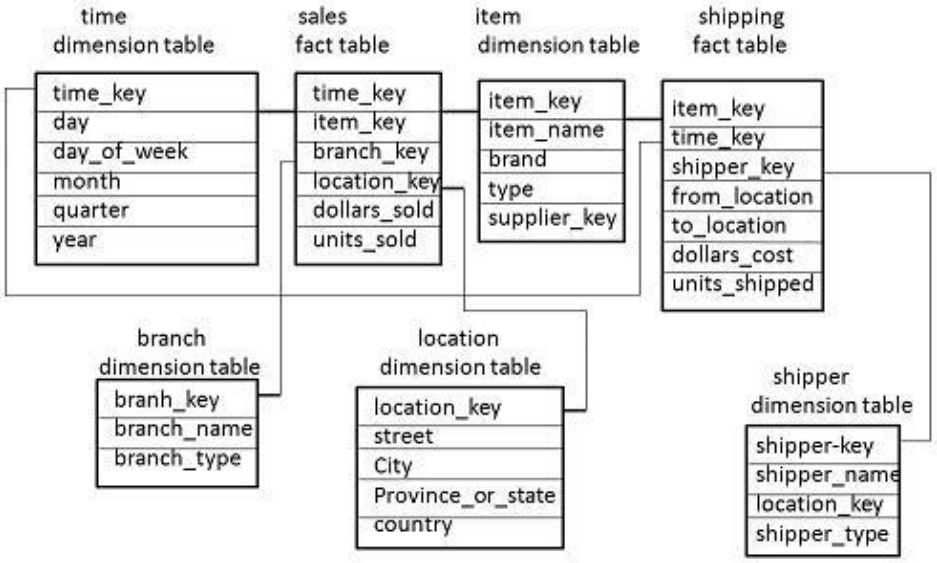
It is a variant of the star schema model, where some dimension tables are normalized, there by further splitting the data into additional tables.

Major difference between snowflake & star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. System performance may be adversely impacted. Due to this it is not popular as the star schema.



3. 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, & hence is called a galaxy schema or fact constellation.



Measures:-

Multidimensional point in the data cube space can be defined by asset of dimension value pairs. A data cube measures is a numerical function that can be evaluated at each point in the data cube space.

A measures value is computed for a given point by aggregating the data corresponding to the respective dimension value pairs defining the given point.

Measures can be organized into three categories:-

Based on the kind of aggregate functions.

1) **Distributive:**- It can be computed in distributed manner.

   E.g. Count () can be computed for a data cube by first partitioning cube into a set of sub cubes, computing count () for each sub cube, & Then summing up the counts obtained for each sub cube. Therefore count() is a distributive aggregate function.

   A measures is distributive is obtained by applying a distributive aggregate function.

2) **Algebraic:**-It can be computed by an algebraic function with M arguments, each of which is obtained by applying a distributive aggregate function.

   E.g. avg () can be computed by sum ()/count (),

   Where both sum () & count () are distributive aggregate functions.

   A measures is algebraic if it is obtained by applying an algebraic aggregate function.

3) Holistic:-

An aggregate function is holistic if there is no constant bound on the storage size needed to describe a sub aggregate.

E.g. Median () , Mode() , Rank() .

A measure is holistic if it is obtained by applying a holistic aggregate function.

Concept Hierarchies:-

A concept hierarchies defines a sequence of mappings from a set of low-level concepts to higher-level.

Eg. Concept hierarchy for dimension location.

Many concept hierarchies are implicit within the database schema.

A schema hierarchy that is total or practical order among attributes in a database schema is called a schema hierarchy. Concept hierarchies that are common to many applications may be predefined in the data mining system.

Eg. For time.

IT is also defined by descripting or grouping values for a given dimension or attribute , resulting in a set-grouping hierarchy.

Concept hierarchies may be provided manually by system users , domain experts , or knowledge engineers.

OLAP Operations in Multidimensional Data Model:-

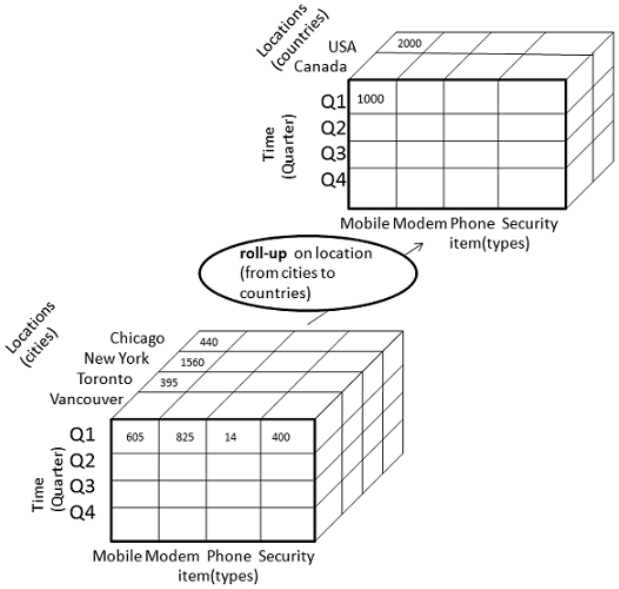
Data are organized into multiple dimensions , & each dimension contains multiple levels of abstraction defined by concept hierarchies.

OLAP provides a user friendly environment for interactive data analysis by using numbers of OLAP data cube operations.

1.**Roll-up(Drill-up):-**

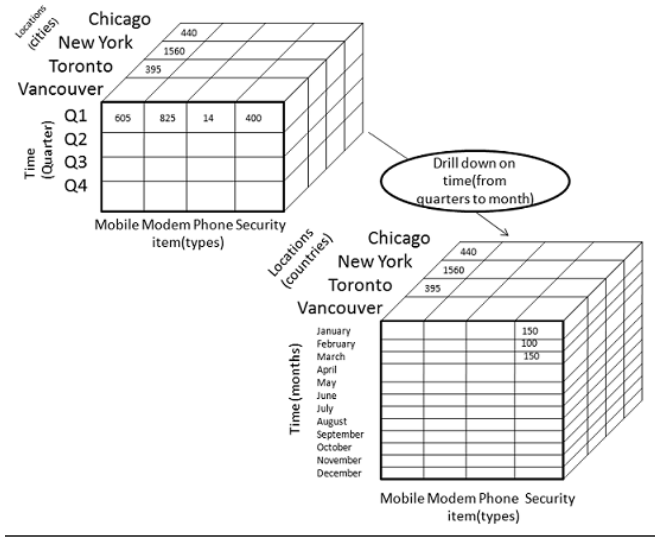
This operation performance aggregation an a data cube , either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Eg. Roll-up operation performed on the central cube by climbing up the concept hierarchy for location. The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country.



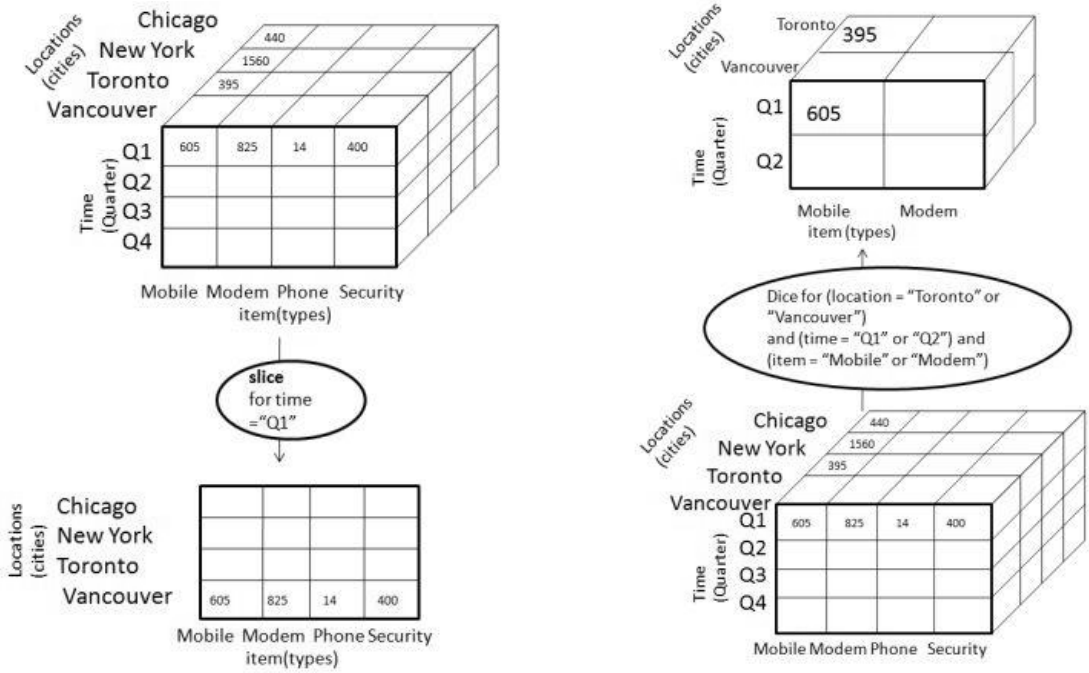
2.Drill-down:-

It 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.



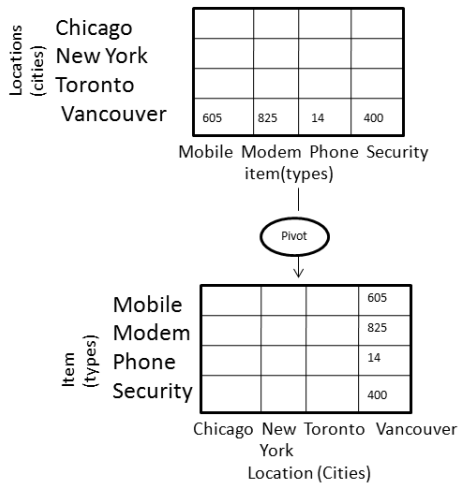
3.Slice & Dice :-

Slice operation performs a selection an one dimension of the given cube, resulting in a sub cube.The dice operation defines a sub cube by performing a selection of two or more dimensions.



4.Pivot :-

Pivot(Rotate) is a visualization operation that rotates the data axes in view in order to provide an alternatives presentation of data.



Types of OLAP servers:

Relational OLAP servers(ROLAP)

The Relational OLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data the Relational OLAP use relational or extended-relational DBMS. The ROLAP tools need to analyze large volume of data across multiple dimensions.The ROLAP tools need to store and analyze highly volatile and changeable data.

ROLAP includes the following.

implementation of aggregation navigation logic.
optimization for each DBMS back end.
additional tools and services.

Advantages

The ROLAP servers are highly scalable.
They can be easily used with the existing RDBMS.
Data Can be stored efficiently since no zero facts can be stored.
ROLAP tools do not use pre-calculated data cubes.
DSS server of microstrategy adopts the ROLAP approach.

Disadvantages

Poor query performance.
Some limitations of scalability depending on the technology architecture that is utilized.

Multidimensional OLAP (MOLAP)

Multidimensional OLAP (MOLAP) uses the array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore many MOLAP Server uses the two level of data storage representation to handle dense and sparse data sets. MOLAP tools need to process information with consistent response time regardless of level of summarizing or calculations selected. The MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis. The MOLAP tools need fastest possible performance.

Advantages

Here is the list of advantages of Multidimensional OLAP
MOLAP allows fastest indexing to the precomputed summarized data.
Helps the user who are connected to a network and need to analyze larger, less defined data.
Easier to use therefore MOLAP is best suitable for inexperienced user.

Disadvantages

MOLAP are not capable of containing detailed data.

The storage utilization may be low if the data set is sparse.

MOLAP vs ROLAP

SN	MOLAP	ROLAP
1	The information retrieval is fast.	Information retrieval is comparatively slow.
2	It uses the sparse array to store the data sets.	It uses relational table.
3	MOLAP is best suited for inexperienced users since it is very easy to use.	ROLAP is best suited for experienced users.
4	The separate database for data cube.	It may not require space other than available in Data warehouse.
5	DBMS facility is weak.	DBMS facility is strong.