Teacher's Sign / Remarks

e No.

UNIT-I

Data warehouse

Data:- Data is information that has been translated into a form that is efficient for movement or processing.

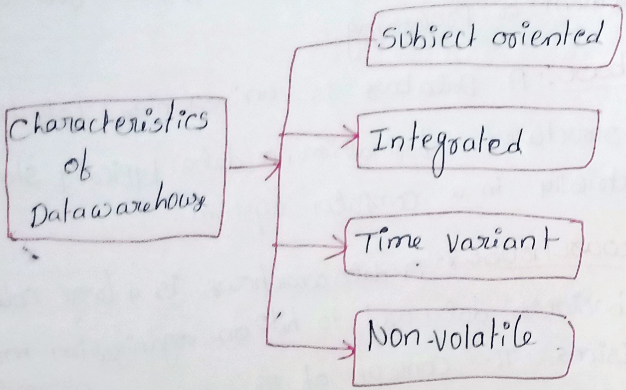
Database:- A Database is an organized collection of structured information, or data typically stored electrically in a computer system.

Data warehouse:- A data warehouse is a large collection of business data used to help an organization make decisions. The concept of the data warehouse has existed since the 1980's, when it was developed to help transition data from merely powering operations to fueling decision support systems that reveal business intelligence. The large amount of data in data warehouses comes from different places such as internal applications such as marketing, sales & finance, customer facing apps and external partner systems among others.

Data ware house is a subject oriented, integrated, non-volatile and time varient collection of data in support of management's decision.

Characteristics or Functions of Datawarehouse:-

Datawarehouse can be controlled when the user has a shared way of explaining the trends that are introduced as specific subject.



1. **Subject-oriented:-**

A datawarehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations. It can be achived on specific theme. That means the datawarehousing process is proposed to handle with a specific theme which is more defined.

These themes can be sales, distributions, marketing etc.

A datawarehouse never put emphasis only current operations it focus on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

2. **Integrated:-** It is somewhere same as subject orientation which is made in a reliable format integration means bounding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various datawarehouse in shared and generally granted manner. Integration of datawarehouse handles various subject related warehouse.

3. **Time-variant:-** In this data is maintained via different intervals of time such as weekly, monthly or annually etc. It bounds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for datawarehouse is wide ranged than that of operational systems.

The data resided in datawarehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time variance is that once data is stored in the datawarehouse then it cannot be modified, alter or updated.

4. Non-volatile:-

As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the quality of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis with in the technologies of warehouse. In this data is read-only and refered at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. Two types of data operations done in the data warehouse are:

→ Data Loading
→ Data Access.

Functions of Datawarehouse:-

It works as a collection of data and here is organized by various communities that endures the features to recover the data functions. It has stocked facts about the tables which has high transaction levels which are observed so as to define the datawarehousing techniques and major functions which are involved in this are

1. Data Consolidation
2. Data Cleaning
3. Data Integration.

Differences between operational Database systems and Datawarehouses:-

The operational Database is the source of information for the Data warehouse. It includes detailed information used to run the day to day operations of the business. The data frequently changes as updates are made and reflect the current value of the last transactions. Operational Database management systems also called as OLTP (online Transaction processing Database

are used to manage dynamic data in real time. Data warehouse systems serve users or knowledge workers in the purpose of data analysis and decision making such systems can organize and present information in specific formats to accommodate the diverse needs of various users. These systems are called as online Analytical Processing systems (OLAP).

Operational Database	Data warehouse
1. Operational systems are designed to support high-volume transaction processing	1. Data warehousing systems are typically designed to support high-volume analytical processing (OLAP)
2. Operational systems are usually concerned with current data.	2. Data warehousing systems are usualy concerned with historical Data.
3 Data with in operational systems are mainly update regularly according to need.	3. Non volatile, new data may be added regularly. once Added rarely changed.
4. It is designed for real time business dealings and processes.	4. It is designed for analysis of business measures by subject area, categories, and attributes.
5. It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	5. It is optimized for extent loads and high complex, unpredictable queries that access many rows per table.
6. It supports Thousends of concurrent clients	6. It supports a few concurrent clients relative to OLTP.
7. Operational systems are widely process-oriented	7. Data warehousing systems are widely subject oriented
8. Data in	8. Data out.
9. Less number of Data accessed	9. Large number of data accessed.
10. Relational Databases are created for online transactional processing (OLTP)	10. Data warehouse designed for on-line Analytical Processing (OLAP)

Differences Between OLTP & OLAP :-

OLTP :- OLTP system handle with operational data. operational data are those data contained in the operation of a particular system. Example. ATM transaction and Bank transactions etc.

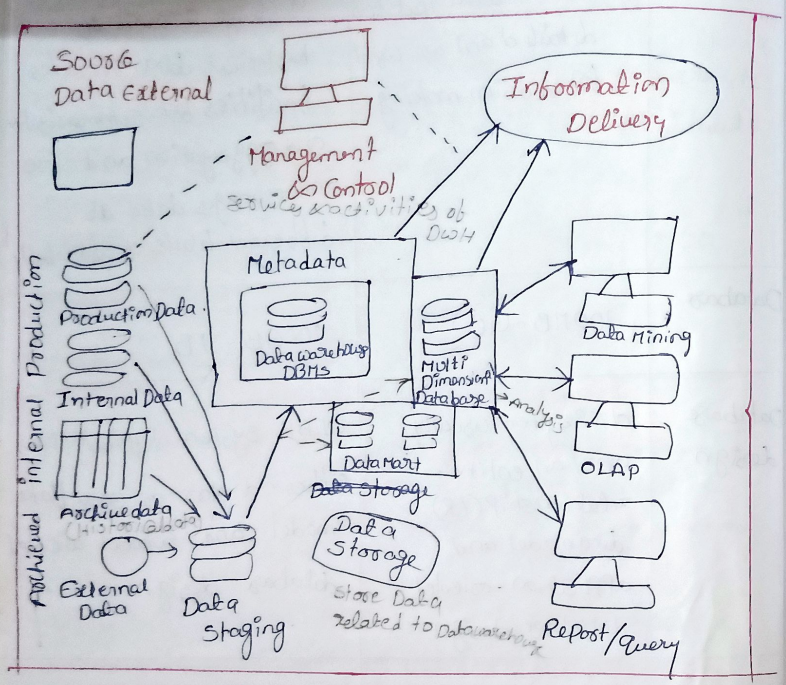
OLAP :- OLAP handle with historical Data or Archival data. Historical data are those data that are achieved over a long period. For example if we collect the last 10 years information about Flight reservation. The data can give us much meaningful data such as it trends in the reservation.

The major difference between an OLTP and OLAP system is the amount of data analyzed in a single transaction. an OLTP manage many concurrent customers and queries touching only an individual record or limited group of files at a time.

An OLAP system must have the capability to operate on million of files to answer a single query.

Feature	OLTP	OLAP
Characteristic	It is a system which is used to manage operational Data	It is a system which is used to manage informational Data.
Users	Clerks, clients, and information technology Professionals	Knowledge workers including managers, executives and analysts
system orientation	OLTP system is a customer oriented, transaction an query processing are done by Clerks, clients and information technology professionals.	OLAP system is market oriented, knowledge workers including managers, do data analysts executive and analysts.
Data Contents	OLTP system manages current data that too detailed and are used for decision making.	OLAP system manages a large amount of historical data. Provides facilities for summarization and aggregation and stores and manages data at different levels of granularity
Database	100MB - GB	100GB - TB.
Database design	OLTP system usually uses an entity - relationship (ER) data model and application - oriented database design.	OLAP system typically uses either a star or snowflake model and subject oriented database design.
Volume of Data	Not very large	Because of their large volume, OLAP data are stored on multiple storage media
Access Mode	Read/write	Mostly write.
Number of records accessed	Tens	Millions.

Data warehouse Components & Architecture :-



Components or building blocks of Data warehouse & Architecture.

We build a data warehouse with software & hardware components. to suit the requirements of our organizations, we arrange these building we may want to boost up another part with extra tools and services.

Source Data Component :-

Source data comming into data warehouses may be grouped into four broad Categories.

Productiondata :- This type of data comes from the different operating systems of the enterpoise. Based on the data requirments in the data warehouse, we Choose segments of the data from the various operational modes.

Internal Data :- In each organization the client keeps their "Poivate" spread sheets, reports, customer Poofiles, and sometimes even department databases This is the Internal data, Part of which could be useful in a data warehouse.

Archieved Data :- Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in achieved files.

External Data :-

Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department

Data Stagging Component :-

Extracted Data from various operational systems and external sources, we have to prepare the files for storing in the datawarehouse. The extracted data comming from several different sources need to be changed, converted, and made ready in a format that is relevant to be saved for querying and analysis

Three primary functions that takes place in the stagging area

1) Extraction
2) Transformation
3) Loading.

1) Data Extraction :-

This method has to deal with numerous data sources. we have to employ the appropriate techniques for each data source.

2) Data transformation :

Data for the data warehouse comes from many different sources. If data extraction for a data warehouse posture big challenges, data transformation present even significant challenges. we perform several individual tasks as part of data transformation.

First we clean the data extracted from each source. Cleaning may be the correction of misspellings or may deal with providing default values for missing data elements, or elimination of duplicates when we bring in the same data from various source systems. Standardization of Data components from a large part of data transformation. Data transformation contains many forms of combining pieces of data from different sources. we combine data from single source record or related data parts from many source records.

3) Data Loading :-

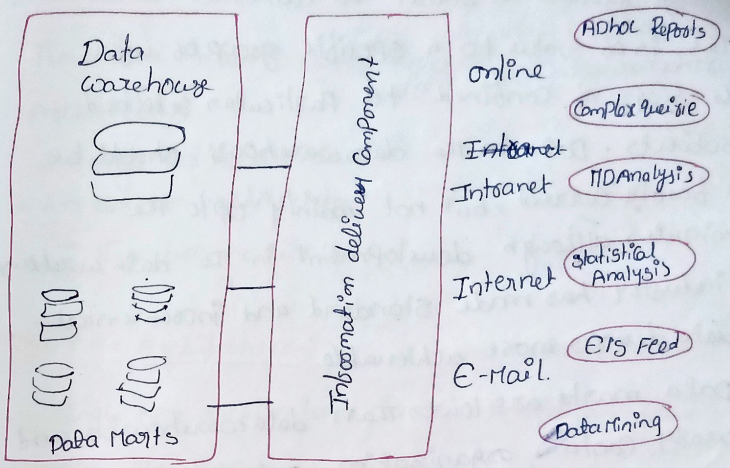
Two distinct categories of tasks form data loading functions. when we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the information into the data warehouse storage. The initial Load moves high volumes of data using up a substantial amount of time.

Data storage components :-

Data storage for the data warehousing is a split repository. The data repositories for the operational systems generally include only the current data. Also, these data repositories include the data structured in highly normalized for fast and efficient processing.

Information Delivery :-

The information delivery element is used to enable the process of subscribing for data warehouse files and having it transferred to one or more destinations according to some customer specific scheduling algorithm.



Information delivery Component.

Metadata Components :

Metadata is a data warehouse is equal to the data dictionary or the data catalog in a database management system. In the data dictionary, we keep the data about the logical data structures, the data about the records and addresses, the information about the indexes.

Metadata is nothing but Data about Data.

Data Marts:-

It includes a subset of corporate-wide data that is of value to a specific group of users. The scope of confined to particular selected subjects. Data in the data warehouse should be a fairly current, but not mainly upto the minute, although development in the data warehouse industry has made standard and incremental data dumps more achievable.

Data marts are lower than data warehouses and usually contain organization. Data warehousing developed a data warehouse with several smaller related data marts for particular kinds of queries and reports.

Management & Control:-

The management & control elements coordinate the services and functions within the data warehouse. These components control the data transformation and the data transfer into the data warehouse storage. It moderates the data delivery to the clients. It's work with the DBMS and authorized data to be correctly saved in to the repositories.

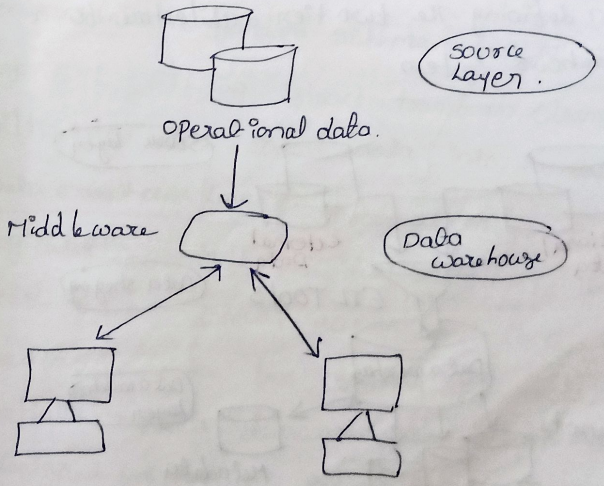
Types of Datawarehouse Architecture:-

There are mainly three types of Datawarehouse Architectures

- → Single tier Architecture
- → Two tier architecture
- → Three tier Architecture.

Single tier Architecture:-

Single tier architecture is not periodically used in practice. Its purpose is to minimize the amount of data stored to reach this goal. It removes data redundancies.

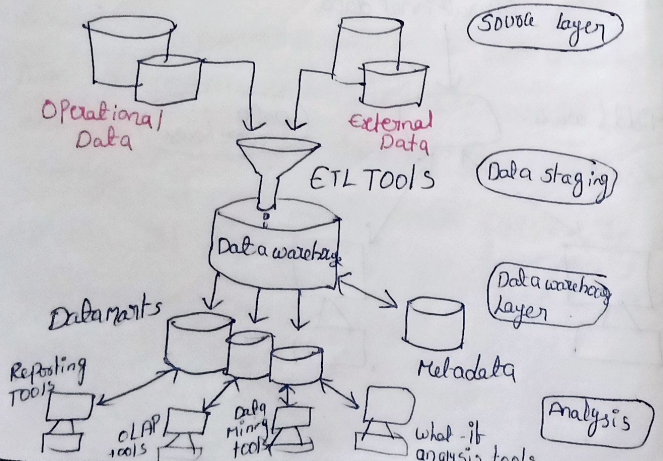


The only Layer physically available is the source Layer. In this method data warehouse are virtual. This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware or an intermediate Processing Layer.

The vulnerability of this architecture lies in its failure to meet the requirment for seperation between analytical and transactional Processing.

Two-tier Architecture :-

The requirment for seperation plays an essential role in defining the two tier architecture for a data warehouse system



it is typically called two-layer architecture to highlight a seperation between physically available sources & data warehouses.

It consists of four subsequent data flow stages.

1. Source Layer :- A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases or it may come from an information system outside the corporate walls.

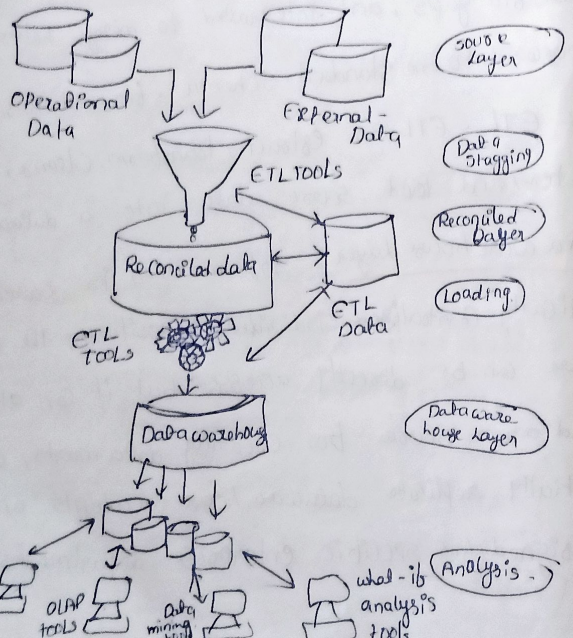
2. Data Stagging :- The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. For all this we use ETL. ETL can extract, transform, cleanse, validate filter, and load source data into a data warehouse.

3. Data ware house Layer :- Information is saved to one logically centralized individual repository. The data ware house can be directly accessed but it can also be used as a source for creating data marts, which partially replicate data ware house contents and are designed for specific enterprise departments.

4. Analysis :- On this layer integrated data is efficiently and flexible accessed to issue reports, dynamically analyze information and simulate hypothetical business scenarios. It should feature aggregate information navigators, complex query optimizers, and customer friendly GUI's.

Three-tier Architecture :

The three tier architecture consists of the source layer, the reconciled layer and the data warehouse layer. The reconciled layer sits between the source data & data warehouse.

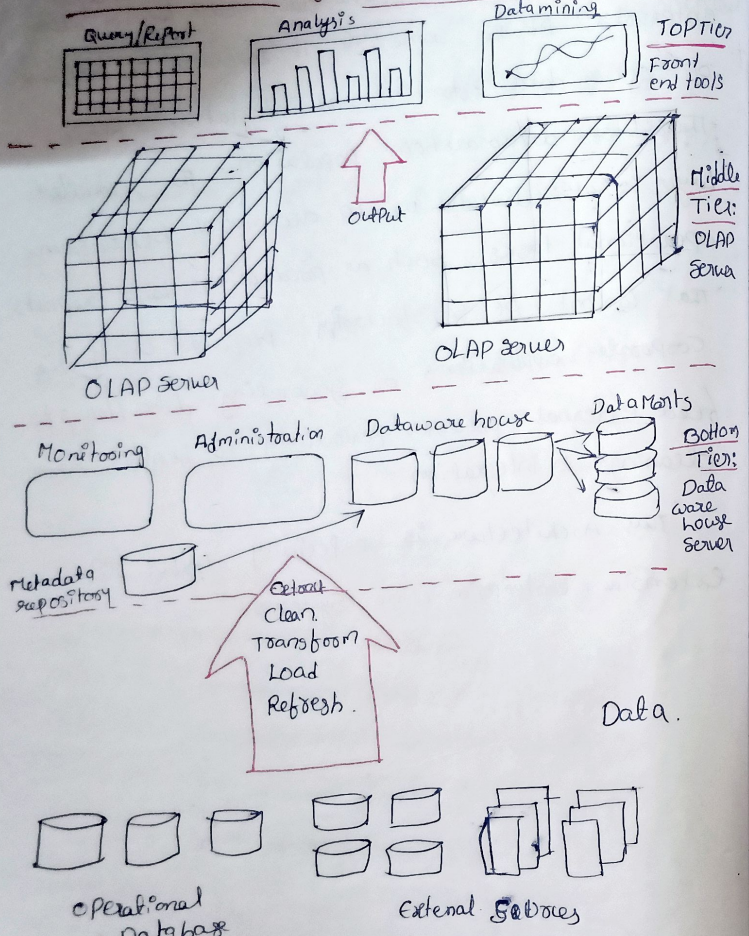


The main advantage of the reconciled layer is that it creates a standard reference data model for a whole enterprise. At the same time, it separates the problems of source data extraction and integration from those of data warehouse population. The reconciled layer is also directly used to accomplish better some operational tasks. such as producing daily reports that cannot be satisfactorily prepared using the corporate applications or generating data flows to feed external processes periodically to benefit from cleaning & integration.

This Architecture is especially useful for extensive, enterprise-wide systems.

TYPES

Data warehousing : A Multi tiered Architecture



A Three tier Datawarehouse Architecture

1. The Bottom tier is warehouse database server that is almost always a RDBMS (Relational DB) Back end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (e.g. Customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning and transformation as well as Load and refresh the functions to update data warehouse.

The data are extracted using application Program interfaces known as gateways. A gateway is provided by the underlying DBMS and allows customer programs to generate SQL code to be executed at a server. Examples of gateways contain ODBC (Open Database Connection) and OLE-DB (Open-Linking and Embedding for Databases) by Microsoft, and JDBC (Java Database Connection).

2. The OLAP server is implemented using either

① A Relational OLAP (ROLAP) model i.e an extended relational DBMS that maps functions on multidimensional data to standard relational operations.

(2) A multidimensional OLAP (MOLAP) model - i.e a particular purpose server that directly implements multidimensional information and operations.

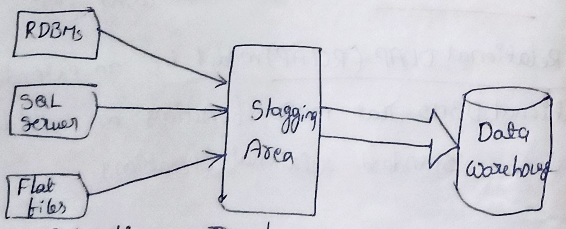
(3) A top-tier that contains front-end tools for displaying result provided by OLAP, as well as additional tools for data mining of the OLAP generated data.

The metadata repository stores information that defines DW objects. Meta data repository located in Bottom layer.

Extraction, Transformation and Loading :-

Data warehouse systems use back-end tools and utilities to populate and refresh their data.

ETL is a process in Data warehousing and it stands for Extract, transform and Load. ETL tool extracts the data from various data source systems, transforms it in the stagging area and then load it into the data warehouse systems.



→ Data Extraction :- which typically gathers data from multiple, heterogeneous and external sources.

→ Transformation :- Transformation having set of rules or functions are applied on extracted data to convert it into a single standard format

- → Filtering – loading only certain attributes into the data warehouse
- → Cleaning – Filling up the NULL values with some default values, mapping U.S.A, United States and America into USA etc
- → Joining - joining multiple attributes into one.
- → splitting - splitting a single attribute into multiple attribute
- → Sorting - sorting tuples on the basis of some attributes

→ Loading :- In the Loading, the transformed data is finally loaded into the data warehouse. Some times the data is updated by loading into data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading depends on the requirments and varies from system to system.

Metadata Repository:-

Metadata is data about data. Meta data is the data that define datawarehouse objects. The Metadata Repository with in bottom tier of the data warehouse Architecture Metadata created for the data names in definitions of the given warehouse. Additional metadata are created and captured for timestamping any extracted data.

A meta data repository should contain.

→ The datawarehouse structure, which includes the warehouse schema view, dimensions, hierarchies, and derived data definitions as well as data mart locations in contents.

→ Operational metadata which include data lineage (history of migrated data and the sequence of transactions applied for it) currency of data (active, archived, or purged) and monitoring information (warehouse usage statistics, error reports, and audit trails)

→ The algorithm used for summarization which include measure and dimension definition algorithms, data in granularity, partitions, subject areas, aggregation. summarization and predefined queries and reports

→ Mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security.

→ Data related to system performance which includ indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

→ Business metadata which include business terms and definitions, data ownership information and charging policies.

Logical (multi-dimensional) Data Model:-

The multidimensional data model is a method which is used for ordering data in the database along with good arrangement and assembling of the content in the database.

The multidimensional Data model allows customers to interrogate analytical questions associated with market or business trends, unlike relational databases which allow customers to access data in the form

They allow users to rapidly receive answers to the requests which they made by creating and examining the data comparatively fast.

OLAP (online analytical processing) and data warehousing uses multidimensional databases. It is used to show multidimensions of the data to users.

It represents data in the form of data cubes. Data cubes allow to model & view the data from many dimensions & perspectives. It is defined by dimensions & facts and is represented by a fact table. Facts are numerical measures and fact tables contain measures of the related dimensional tables or names of the facts.

A Data cube allows data to be modeled & viewed in multiple dimensions. It is defined by dimensions & facts.

Working on a Multi-dimensional Data Model:-

The following stages should be followed by every project for building a multi dimensional Data model.

Stage 1: Assembling data from the Client:-

A multidimensional data model collects correct Data from the client. Mostly, software professionals provide simplicity to the client about the range of data which can be gained with the selected technology & collect the complete data in detail.

Stage 2: Grouping different segments of the system:-

The multidimensional Data model recognizes and classifies all the data to the respective section they belong to and also build it problem-free to apply step by step.

Stage 3:- Noticing the different proportions:-

it is the basis on which the design of the system is based. In this stage the main factors are recognized according to the user's point of view. These factors are also known as "Dimensions".

stage 4 :- Preparing the actual-time factors and their respective qualities :-

The factors which are recognized in the previous step are used further for identifying the related qualities. These qualities also known as "attributes" in the database.

Stage 5 :- Finding the actuality of factors which are listed previously and their qualities :-

A multidimensional Data model seperates and differentiates the actuality from the factors which are collected by it. These actually play a significant role in the arrangement of a Multidimensional data model.

stage 6 :- Building the schema to place the data with respect to the information collected from the steps above :-

On the Basis of the data which was collected previously a schema is built.

Let us take the example of the data of a factory which sells products per quarter in Banglore. The Data is represented in the table

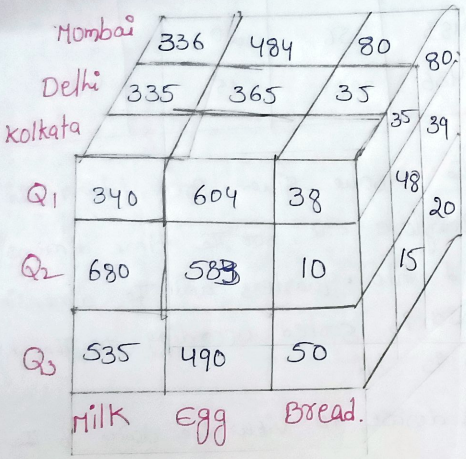
Location = "Banglore"				
	Type of item			
Time (quarter)	Jam	Bread	Sugar	Milk
Q1	350	389	35	50
Q2	260	528	50	90
Q3	483	256	20	60
Q4	436	396	15	40

In the above given presentation. The factory's sales for Banglore are, for the time dimension which is organized into quarters and the dimension of items, which is sorted according to the kind of item which is sold.

If we desire to view the data of the sales in a three dimensional table, then it is represented in a diagram. Here the data of the sales is represented as a two dimensional table Let us consider the data according to item, time and location

	Location = "Kolkata"			Location = "Delhi"			Location = "Mumbai"		
	Item			Item			Item		
Time	Milk	Egg	Bread	Milk	Egg	Bread	Milk	Egg	Bread
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	594	39
Q3	535	490	50	389	385	15	366	385	20

This Data Reprasented in the form of 3-dimensions conceptually.



Advantages of Multidimensional Data Model:-

→ A multidimensional data model is easy to handle.

→ It is easy to maintain.

→ Its Performance is better than that of normal data base.

→ The representation of data is better than traditional databases. That is because the multi-dimensional databases are multi-viewed and carry different types of factors.

→ It is workable on complex systems and applications, contrary to the simple one-dimensional data base systems.

→ The compatibility in this type of database is an upliftment for projects having lower bandwidth for maintenance staff.

Disadvantages of Multi Dimensional Data model:-

→ The multidimensional Data model is slightly complicated in nature and it requires professional to recognize and examine the data in the database.

→ It is complicated in nature due to which the databases are generally dynamic in design.

→ The path to achieving the end product is complicated most of the time

→ As the multidimensional data model has complicated systems, databases have a large number of databases due to which the system is very insecure when there is a security break.

Data warehouse Modeling:-

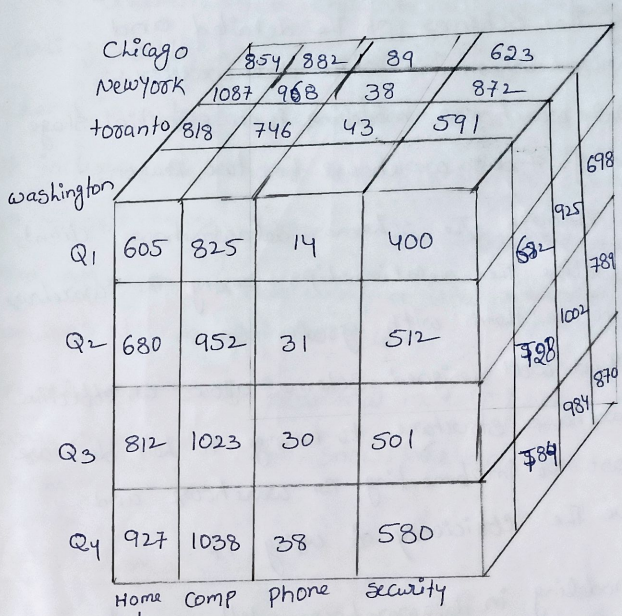
Example:-

2D view of sales Data for All Electronics According to time & item.

	location = "washington"			
	item			
Time	Home Entertainment	Computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1032	38	580

3D view of sales data for All Electronics According Time, item, Location.

	location = "Chicago"				location = "newyork"				location = "toronto"				location = "washington"			
	item				item				item				item			
Time	home ent	comp	phone	sec	home	comp	phone	sec	home	comp	phone	sec	C	P	S	
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580



Data Modeling :-

Data warehouse modeling is the process of designing the schemas of the detailed and summarized information of the data warehouse

Data warehouse modeling is an essential stage of building a data warehouse for two Reasons.

First, Through the schema, data warehouse clients can visualize the relationships among the warehouse data, to use them with greater ease.

Secondly, a well designed schema allows an effective data warehouse structure to emerge, to help decrease the cost of implementing the warehouse and improve the efficiency of using it.

Data modeling in data warehouses different from data modeling in operational Database.

Data modeling is a process for defining and ordering data for use and analysis by certain business processes. The goal of data modeling is to produce high quality, consistent, structured, data for running business applications and achieving consistent result.

Schema design :-

A schema is a collection of database objects including tables, views, indexes, and synonyms

schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. A database, a data warehouse also requires to maintain a schema.

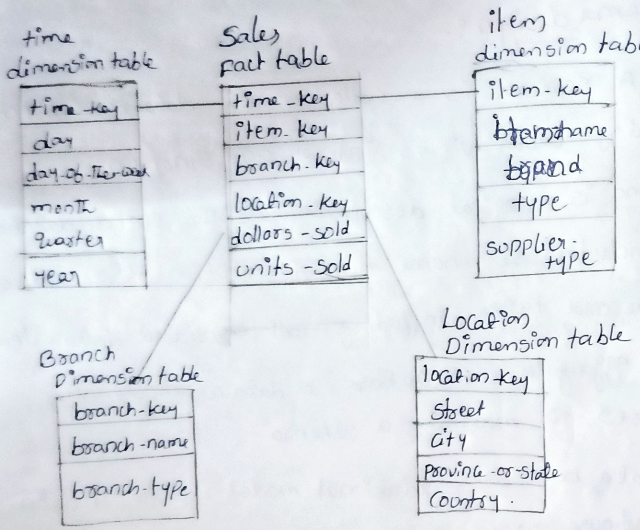
A data base uses relational model while A data warehouse uses star, snowflake and fact constellation schema.

Star schema :-

→ Each dimension in a star schema is represented with only one dimension table.

→ This dimension table contains the set of attributes

→ The diagram shows the sales data of a company with respect to the four dimensions, namely. time, item, branch, and location.



→ There is a fact table at the center, it contains the keys to each of four dimensions

→ The fact table also contains the attributes, namely dollors sold and units sold.

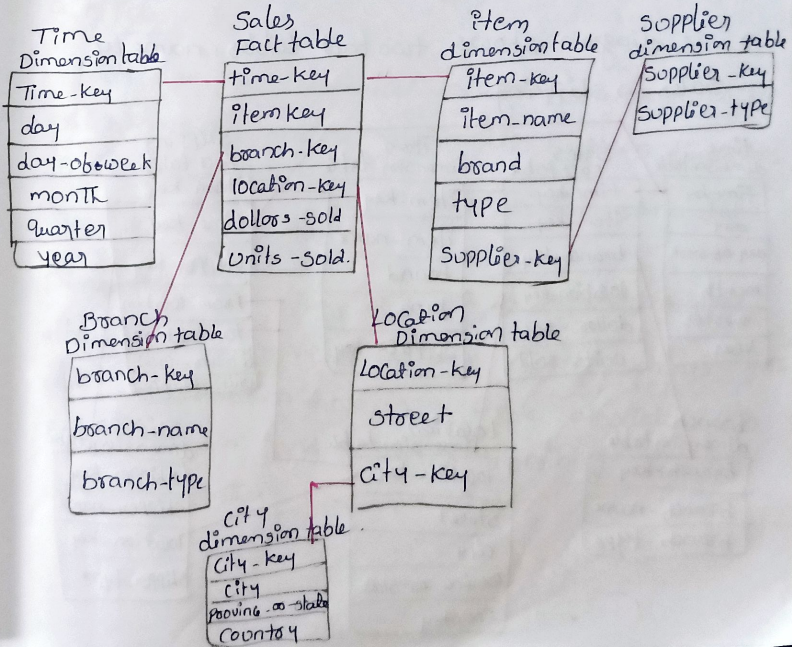
Each dimension has only one dimension table and each table holds a set of attributes.

Snow Flake Schema :-

→ Some dimension tables in the snowflake schema are normalized.

→ The normalization splits up the data into additional tables.

→ Unlike star schema, the dimensions table in a snowflake schema are normalized. For example the item dimension table in star schema is normalized and split into two dimension tables, namely item & supplier table.



→ Now the item dimension table contains the attributes item-key, item-name, type, brand and supplier-key.

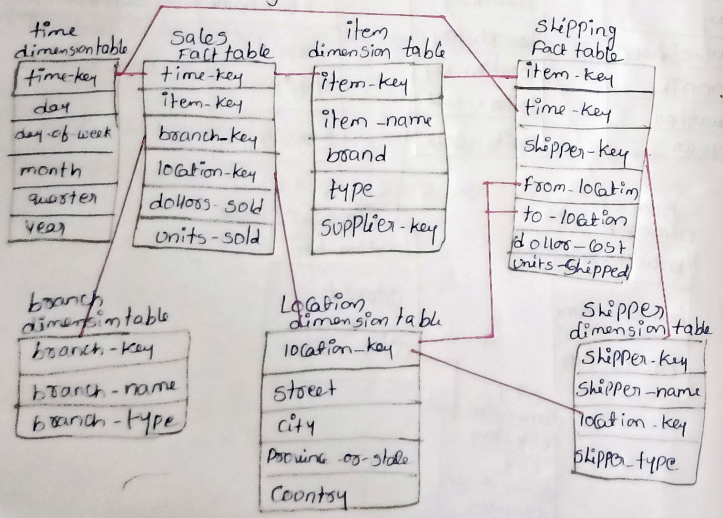
→ The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier-key & supplier-type.

Due to normalization in the snowflake schema the redundancy is reduced and therefore it becomes easy to maintain and the save storage space.

Fact Constellation schema:-

→ A Fact constellation has multiple fact tables. It is also known as galaxy schema.

→ The diagram shows two fact tables, namely sales & shipping.



→ The sales fact table is same as in the star schema

→ The shipping fact table has the five dimensions, namely item-key, time-key, shipper-key, from location, to-location.

→ The shipping fact table also contains two measures namely dollars sold and units sold.

→ It is also possible to share dimension tables between fact tables. For example time, item and location dimension tables are shared between the sales & shipping fact table.

Fact Table:-

Facts are the numerical measures (or) quantities by which one can analyze relationships between dimensions The relation containing such multi dimensional data are called Fact-tables.

Dimensions are the collection of logically related attributes and is views as an axis for modelling the data. A Dimension table is a table associated with each dimension and helps is discribing the dimension further.

Example Sale of Books

BookID	transactionID	number
B1	1	25
B2	2	26

Fact table.

BookID	author	Price
B1	XYZ	40
B2	ABC	70

Dimension table.

Fact Measures :-

Fact table holds the measures data for measuring the performance of your business.
Your business might be sales, purchasing, inventory, logistics, banking, telephony data and many more normally one fact table represents a line of core business and it usually takes more fact tables to more complex business aspects (purchasing, selling, sales etc)

Based on line of business measures ~~tables~~ types in fact table can be :

→ Fully-additive
→ Semi-additive
→ non-additive.

Fully-Additive :-

Fact table holds measures that are grouped summed through all dimensions.

Semi-Additive :-

Fact table holds measures that can be grouped summed or aggregated through some dimensions but not all the dimensions.

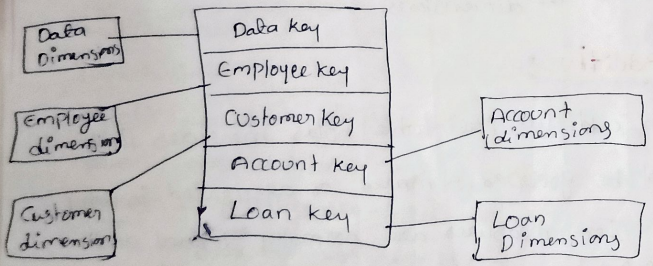
Non-Additive :-

Non-additive fact table holds measures that can not be grouped, summed or aggregated in any aspect. These measures are normally derived and calculated measures such as Percentages, ratios, running sums or any similar measures

Fact-Less-Facts :-

Fact-less-fact tables simply mean the key available in the fact that no remedies are available. Fact less fact tables are only used to establish relationships between elements of different dimensions. And are also useful for describing events and coverage, meaning tables contain information that nothing has happend. It often represents many-to-many relationships.

The only thing they have is an abbreviated key. They still represent a focal phenomenon that is identified by the combination referenced in the dimension table.



A Fact less Fact table

Dimension table characteristics :-

a dimension table characteristics is:

Dimension table key :-

Primary key of the dimension table uniquely identifies each row in the table.

Table is wide :-

a dimension table has many columns or attributes. It is not uncommon for some dimension tables to have more than fifty attributes. Therefore we say that the dimension table is wide. If you lay it out as a table with columns and rows, the table is spread out horizontally.

→ Dimension tabb contain textual information that represents the attributes of the business.

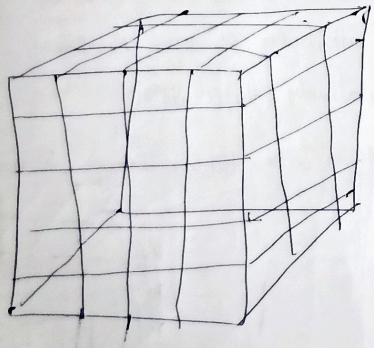
→ It contain relatively static data

→ Are joined to a fact table through a foreign key reference.

OLAP Cube :-

It is also called as Hyper Cube. The OLAP cube is a data structure optimized for very quick data analysis. The OLAP cube consists of numeric facts called measures which are categorized by dimensions.

Data operations and analysis are performed using single spreadsheet, where data values are arranged in a row and column format. This is ideal for two dimensional data. OLAP contains multidimensional data, with data usually obtain from a different and unrelated source. Using a spread sheet is not an optimal option. The cube can store and analyze multidimensional data in a logical and orderly manner.



OLAP Cube.

OLAP Operations :-

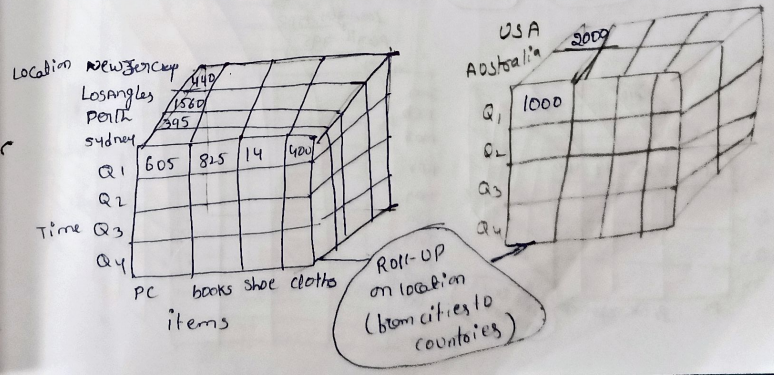
OLAP operations for multidimensional data is four types

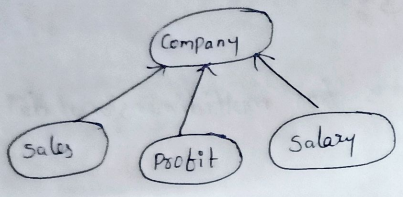
1. Roll-up
2. Drill-down
3. Slice-and-Dice
4. Pivot (rotate).

→ Roll-up :-

Roll-up is also known as "consolidation" or "aggregation". The Roll-up operation can be performed in two ways.

1. Reducing dimensions
2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their ordered level.





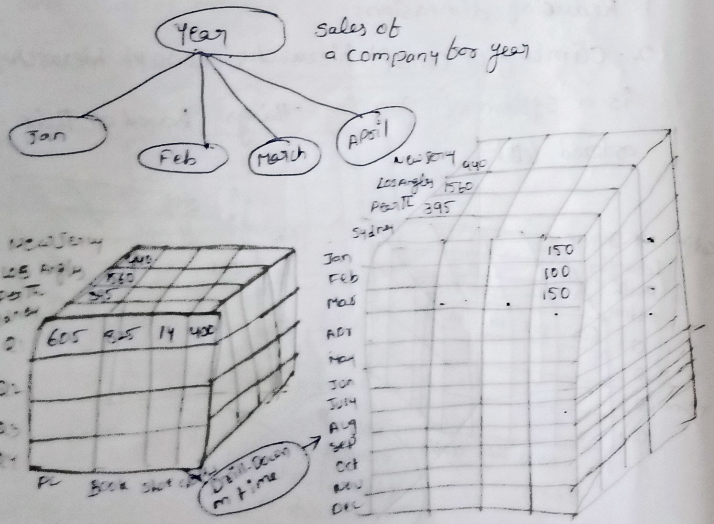
2) Drill-Down :-

In drill-down data is fragmented into smaller parts. It is the opposite of the roll up process.

It can be done via

→ Moving down the concept hierarchy

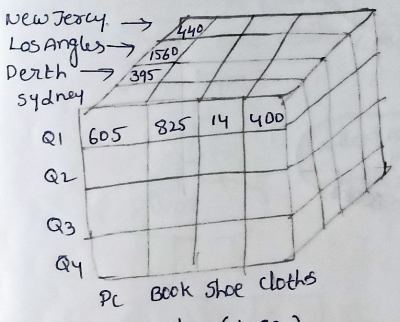
→ Increasing a dimension.

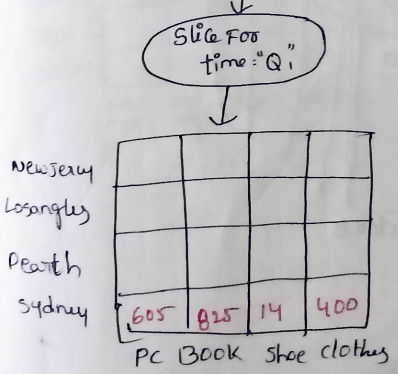


Quarter Q1 is drilled down to months January, February and march. Corresponding sales are also registers.

3) Slice :-

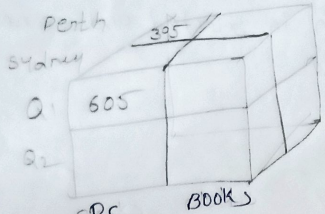
Here, one dimension is selected and a new sub-cube is created



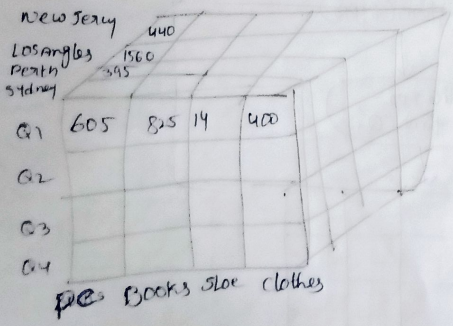


Dice :-

This operation similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube.

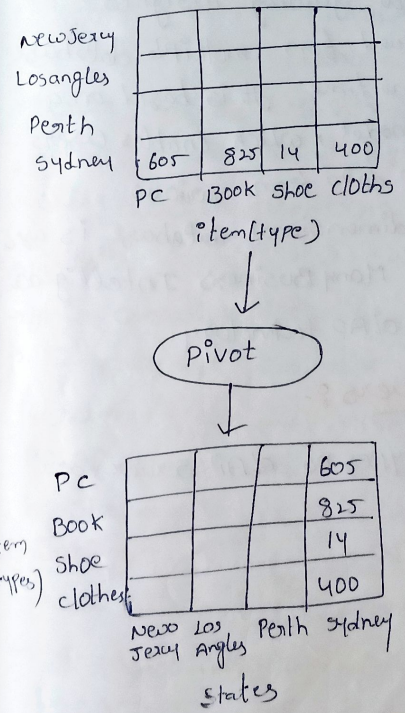


Dice for (location = "Perth" or "Sydney") and time = Q1 or Q2 and item = PC's or BOOKS



pivot :-

In Pivot you rotate the data axes to provide a substitute presentation of data.



OLAP Server Architecture :-

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 applications relay on OLAP technology.

Types of OLAP servers :-

The three major types of OLAP servers are

→ ROLAP

→ MOLAP

→ HOLAP.

Relational OLAP (ROLAP) :-

Relational OLAP (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

Benifits :-

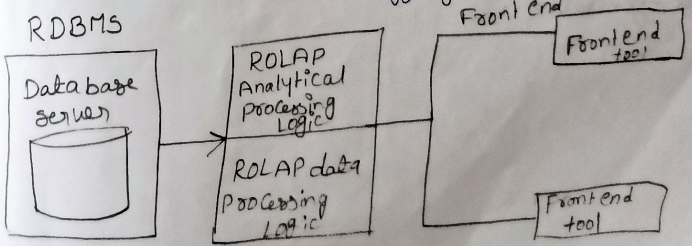
→ It is compatible with datawarehouses and OLTP systems

→ 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 is difficult to keep aggregate tables upto date.



Multidimensional OLAP :-(MOLAP)

MOLAP supports multidimensional views of data. storage utilization in multi dimensional 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 array's Random access capability. The multidimensional array is typically stored in MOLAP in a linear allocation based on nested traversal of the axes in some predetermined order.

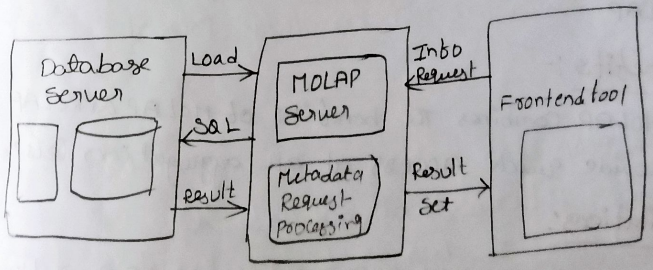
MOLAP systems typically include provisions such as advanced indexing and hashing to locate data while performing queries for handling sparse arrays because both storage and retrieval costs are important when evaluating online performance. MOLAP cubes are ideal for slicing and dicing data and can perform complex calculations. When the cube is created all calculations are pre-generated.

Benefits :-

→ suitable for slicing & 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) :-

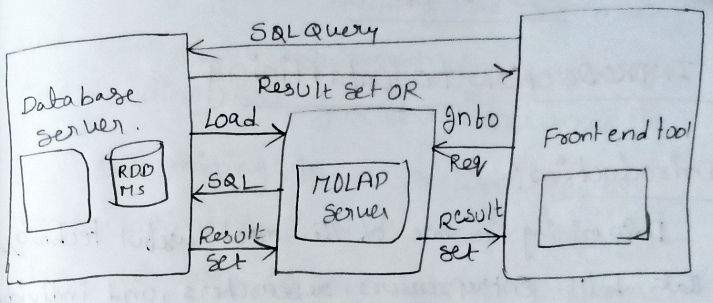
ROLAP & MOLAP are combined in Hybrid OLAP (HOLAP). HOLAP offers greter scalability than ROLAP and faster computation than MOLAP. HOLAP is a hybrid of ROLAP & MOLAP. HOLAP servers are capable of storing large amounts of detailed data. on the one hand HOLAP benefits from ROLAP's greater scalability. HOLAP,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

Benefits :-

→ HOLAP combines the benefits of MOLAP & ROLAP
→ provide quick access at all aggregation levels

Limitations :-

→ supports of MOLAP & ROLAP servers, HOLAP architecture is extremely complex
→ There is greater likelihood of overlap particularly in their functionality.



other types of OLAP are

→ web OLAP (WOLAP)
→ Desktop OLAP (DOLAP)
→ Mobile OLAP (MOLAP)
→ Spatial OLAP (SOLAP)