What is Data warehouse

DB :- All companies maintains databases to store data

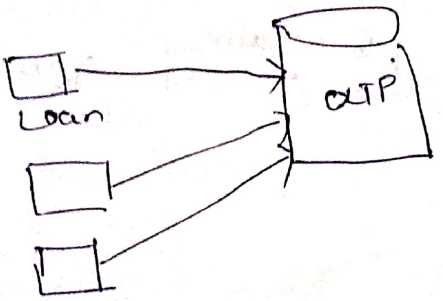
Banks, Online, Store, Organizations

=> They are storing day-to-day transaction.

Ex in bank depositing, many transhering

=> This data base is known Online transactional Processing (OLTP) because they are storing day-to-day transactions.

=> The source for OLTP data are the applications.



Applications are generating data and data are stored into OLTP DB.

=> Now days every organization maintains two databases OLTP and OLAP.

=> Online Analytical Processing (OLAP) is also known as Data warehous (DW) OR DSS (Decision support support).

=> OLAP is used for analysis of data. Because we need historical data. That historical data are placed in OLAP.

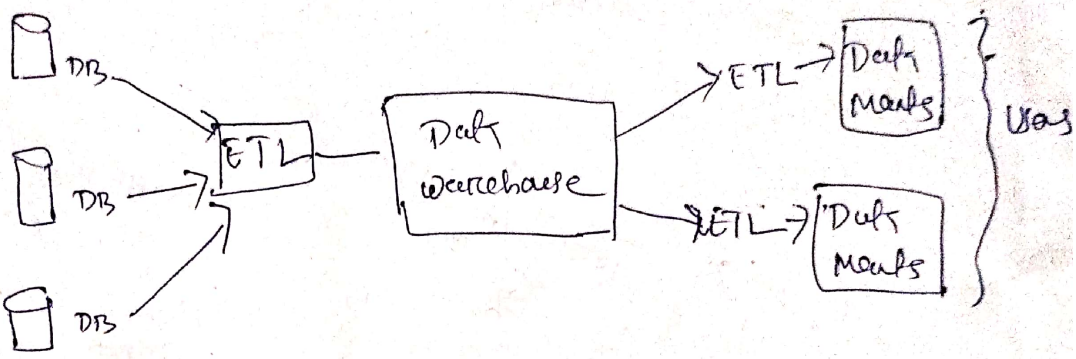
Ex If ten years data are avicable then we can find who is the best players

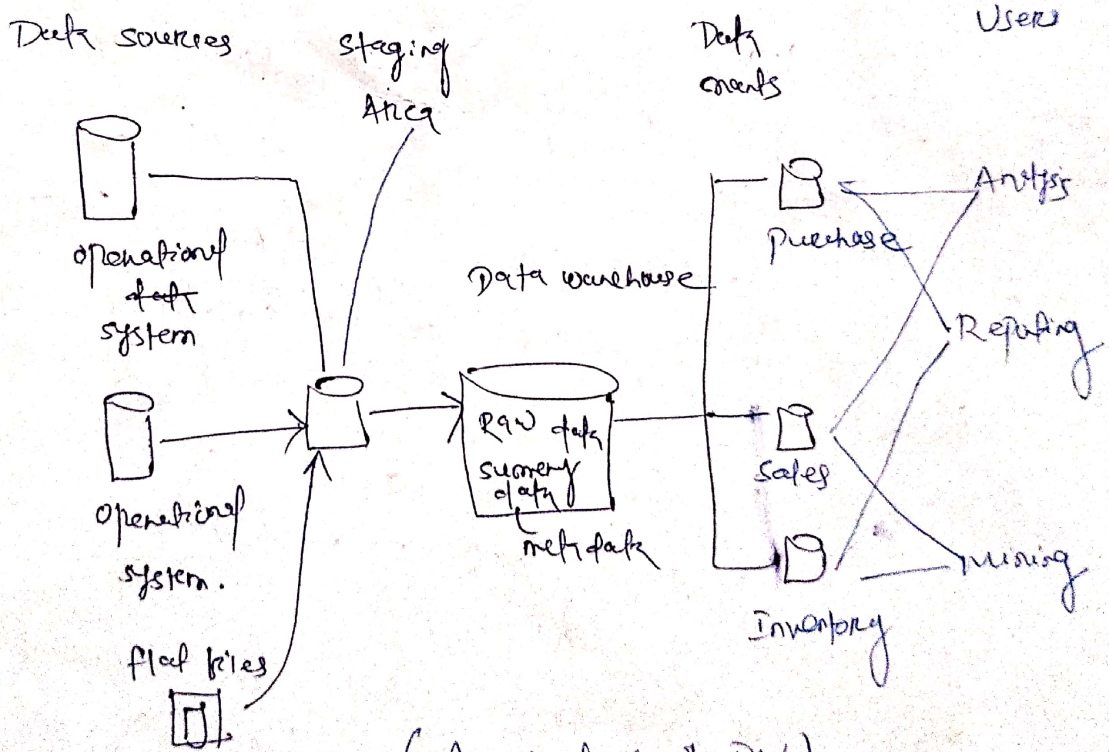
=> Source of OLAP are OLTP.

→ The process of moving the data from OLTP to OLAP is done by ETL (Extraction, transformation, Loading) process.

→ Data warehouse is the process of collecting and managing data from various sources to provide meaningful business insights.

→ DW is the central repositories of integrated from one or more sources.





(Architecture of DW)

Staging Area → intermediate Area. used for data processing during ETL process.

Data marts → is a single form of a DW that focus on a single subject. Ex sales. These are built and controlled by a single department with in an organization.

ETL Process

→ Extraction:- Data from various various source systems are extracted which can be in different formats.

→ Like Relational DB, NO-SQL, XML and flat files.

→ collected and stored into staging area.

→ first collect data from different sources. then store into staging (intermediate) area.

→ not directly to data warehouse because data may be corrupted.

→ transformation:- In this phase set of functions or rules are applied on the extracted data.

→ that function convert all data into a single standard format.

→ sub processes: filtering, joining, splitting, sorting

→ Filtering:- loading some of attributes into the data warehouse.

→ cleaning:- filling the NULL values with some default values, delete incorrect records.

→ splitting:- Spliting a single attribute into multiple attributes.

→ Sorting :- sorting type on the basis of some attribute (generally key attribute)

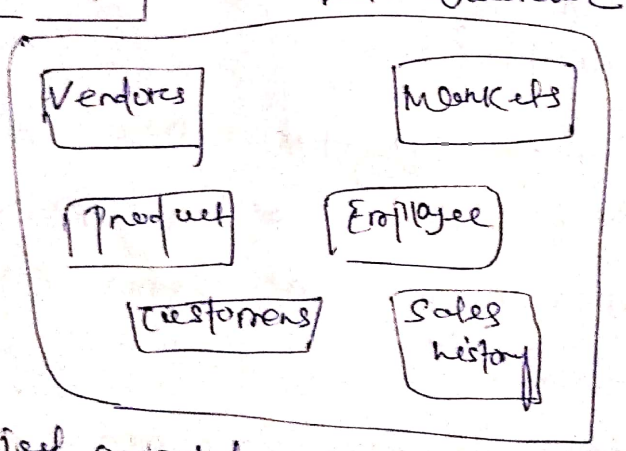
→ Loading :- the transformed data is finally stored / loaded into data warehouse.

→ ETL tools examples :- sybase, oracle warehouse builder / mark logic.

Characteristics of DW

- Subject-oriented
- Integrated
- Time-variant
- non-volatile

Subject-oriented



→ DW is subject oriented because it provides information in subject wise.

→ Means data are stored in Data warehouse in subject wise. so that data can analyzed easily.

Integrated

→ Data are collected from different sources and the combined and placed in DW. (Operational DB)

Time variant

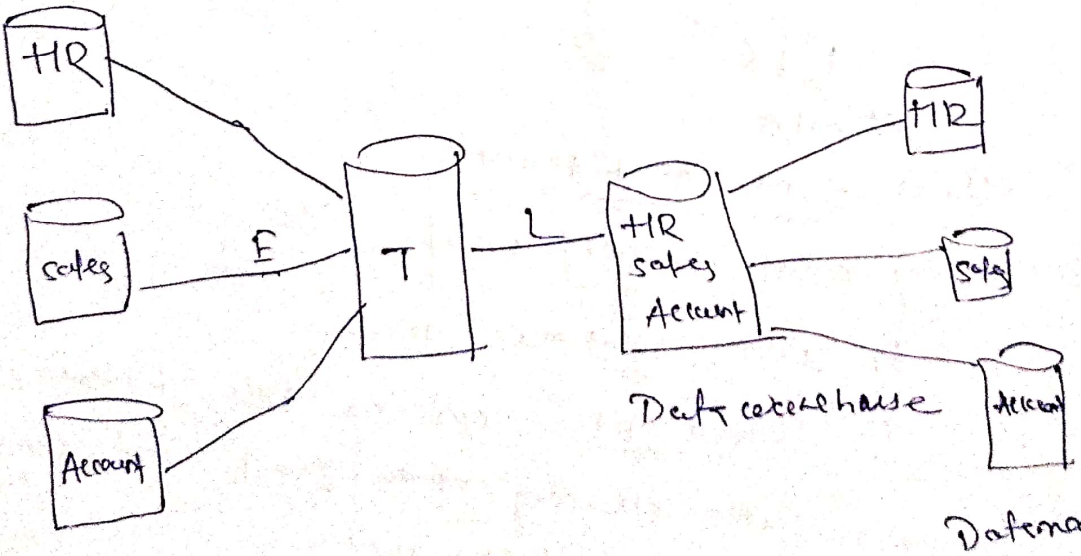
→ Historical data is kept in DW.
→ for analyzing data DW must contain historical data.
ex one can retrive data from 3 months / 6 months / 12 months older data from a DW.

Non volatile

→ once data is in data warehouse it will not change.
→ so historical data in a DW should never be altered.
→ The new data can be added into the DW.
→ Non-volatile means we can't update the historical or any data which is present in DW.

Data marts

→ subset of Data warehouse is known as data marts.



→ It focus on single subject.

→ It is subject-oriented and it is designed to meet the needs of a specific group of users.

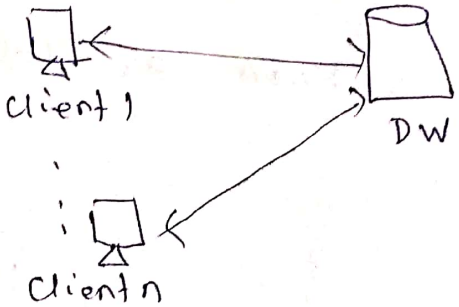
→ Data marts are fast and easy to use, as they make use of small amount of data.

Components of Data warehouse

DW Architecture

1) Single-tier Architecture
   - → objective to minimize the amount of data
   - → goal is to remove redundancy.
   - → this architecture is not used in practice.

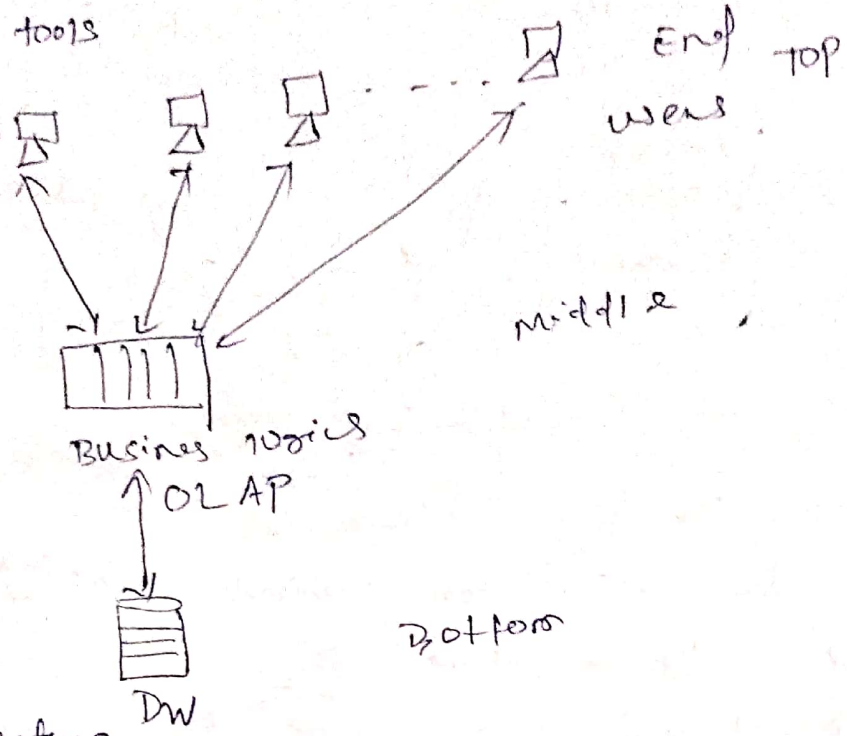
2) two-tier Architecture (client/server)
   - → it separates physically available resources and data warehouse.
   - → it can't support more number of users.



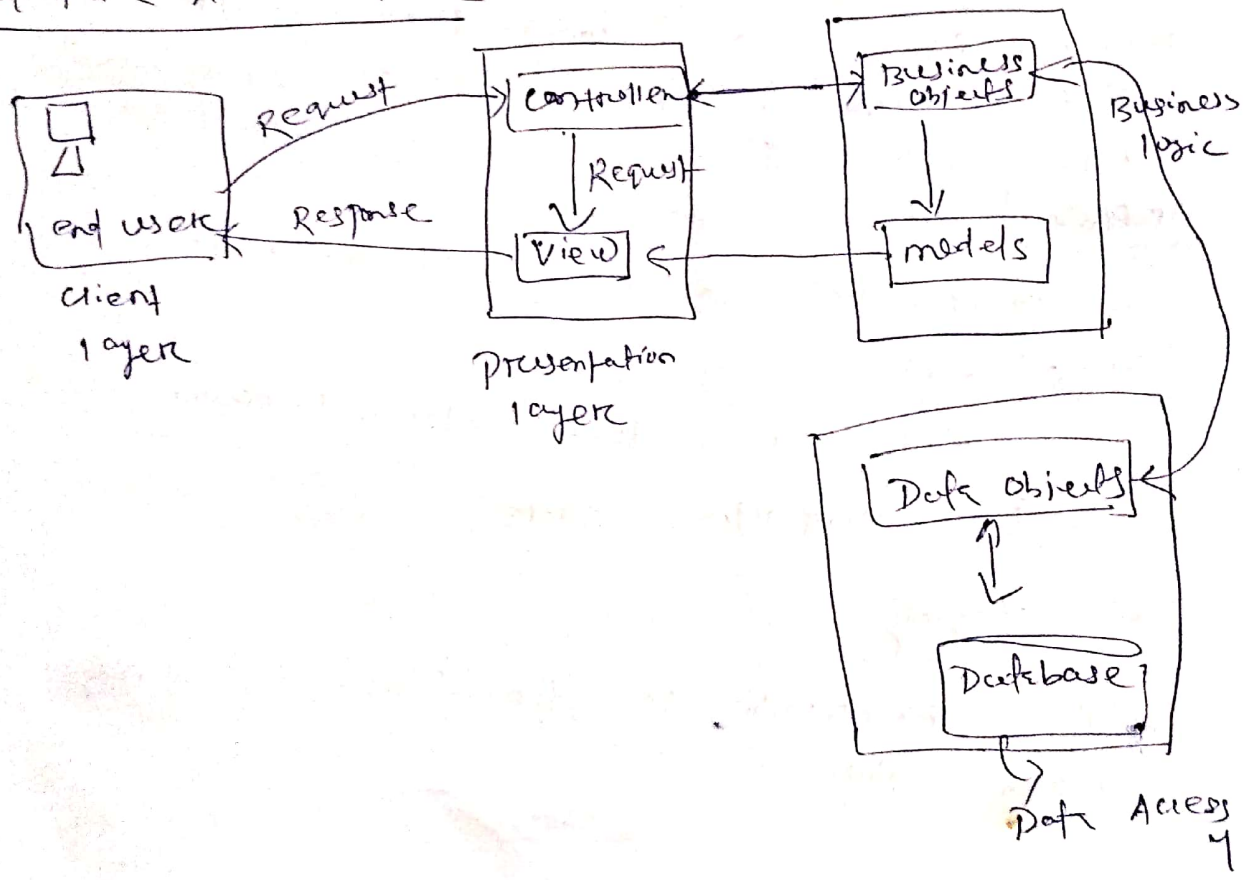
3) three-tier Architecture
   - → most widely used.
   - → contains three tiers.
   - → Bottom tier contains DBs (RDBMS). After cleaning, transformed and loaded into bottom tier.

⇒ Middle tier contains OLAP server. It gives an abstract view of database.

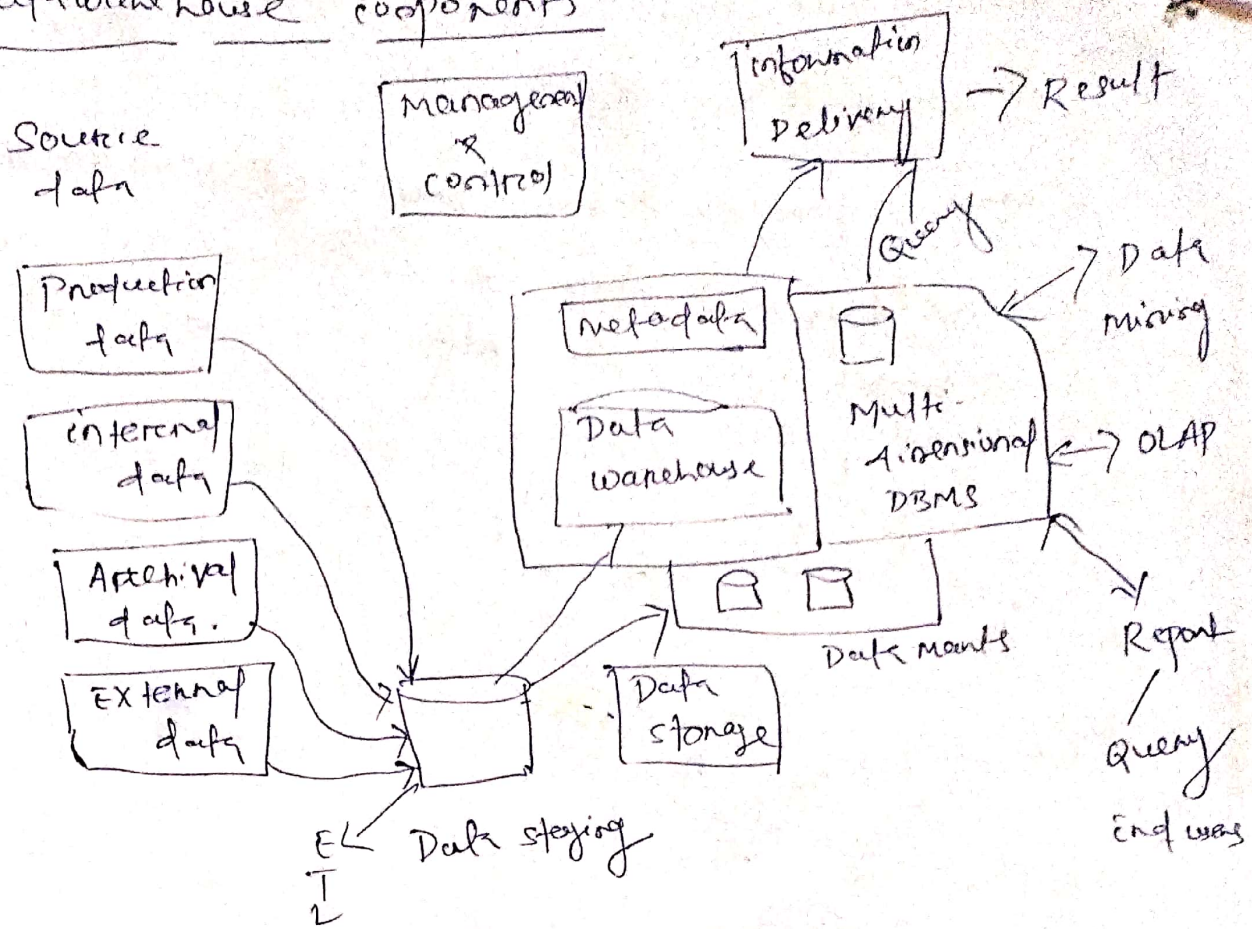
⇒ Top layer/tier is a front-end client layer. Here tools and APIs are available that connect and get data from data warehouse.
Ex: query tools



⇒ 4-tier Architecture



Data warehouse components



Production Data

data comes from various operational data comes from the enterprise.

Internal Data

→ Organization, client keeps private spreadsheets, reports, customer profile. These are the internal data.

Archival Data

It contains old and historical data.

External data

→ data comes from external source.

EX competitors data

Data staging

→ Apply Extraction, transformation and Loading (ETL) Process.

→ After ETL process store the data into some intermediate storage known as data staging.

Metadata

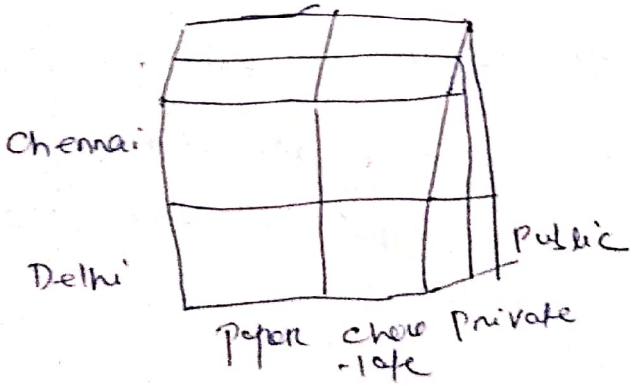
→ It is a data dictionary or the data catalog.
→ It keeps informations like indexes, record and addresses and internal details.

Data Marts

→ Data marts are sub parts of the data warehouse.
→ It is designed for a particular line of business such as sales, marketing.
→ Data mart is an access layer which is used to get data out to the users.
→ Access data from data marts take less time and takes less cost to build because it is coming from data warehouse.

Multidimensional Database

→ It is used mostly for OLAP (online Analytical Processing)
→ It allow users to quickly get answers to their requests. by generating and analysing the data.



Information delivery

→ It is used to enable the process of subscribing for data warehousing.

→ It will distributes warehouse stored data and other information to other data warehouse and end user's product such as spreadsheets and other database.

Management & control

→ It will coordinate the service and functions with in the DW.

→ The control component controls the data transformation and data transform into DW.

→ Manages the data delivery to clients.

→ Security management

DW Access tools

→ Different tools are available of end use site by which they can perform different operations.

→ Query & Reporting:- generate business reports for analysis.

→ Application development:- Application also developed.

→ Data mining:- used to find patterns and correlations in large amounts of data.

→ OLAP tools:- helps to build multi-dimensional dw Analyze the enterprise data from multiple perspective

Metadata in Data warehouse

-) Data dictionary contains metadata.

-) Metadata means data about data.

-) It contains
- information about the logical Data structure
- information about files and addresses
- information about indexes.

Ex: student's all information are stored in one place. Means called the previous records.

-) Metadata component serves as a dictionary of the contents which are present in data warehouse.

Types of Metadata

-) operational Metadata
-) Extraction and transformation metadata
-) End user metadata.

operational Metadata

-) Data comes from several operational systems of enterprise.

-) These source contain different data structures.

-) The elements selected for data warehouse from different sources have different fields and different length, different data types.

-) we can splitting the records, combining the records from different sources files and deal with multiple coding schemes and field lengths.

→ When the data is delivered to the end users that data must be in original form which exists in the source data sets.

Extraction & transformation metadata

→ It contains

1) the extraction of data from the source system
   - the extraction frequencies
   - extraction methods
   - Business rules for the data extraction.

2) Also contains information about all data transformation that take place at the data staging area.

End-user metadata

→ It enables the end-users to find data from the data warehouse.

→ It helps end users to understand the various types of information resources are available in data warehouse.

→ These resources can take many forms like data elements, queries, reports and published documents.

Data cube / OLAP cube

→ When data is grouped or combined in multidimensional matrices then is called as data cubes.

→ Data cube is a structure that enable OLAP operation to achive the multidimensional functionality.

→ It is the extension of 2-D.

→ whenever there are lots of complex data to be aggregated and there is a need to abstract the relevent/important data.

→ Data cube is created from a subset of attributes in the database.

→ specific attributes are selected to be measured attribute

Ex: xyz create a sales data warehouse to keep Records of sales for the dimensions time, item, branch & location.

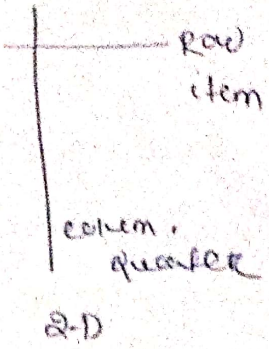
Each dimension may have table known as dimension/ table. Ex:- dimension table for item contains item-name brand and type.

Ex 2-D view of data

Electronic item sales data for the quarters in the city delhi.

location = "Delhi"			
	item (type)		
time (Quarter)	home entertainment	computer	phone
Q1	605	825	14
Q2	680	952	31
Q3	812	1023	50
Q4	715	250	38

in dollars



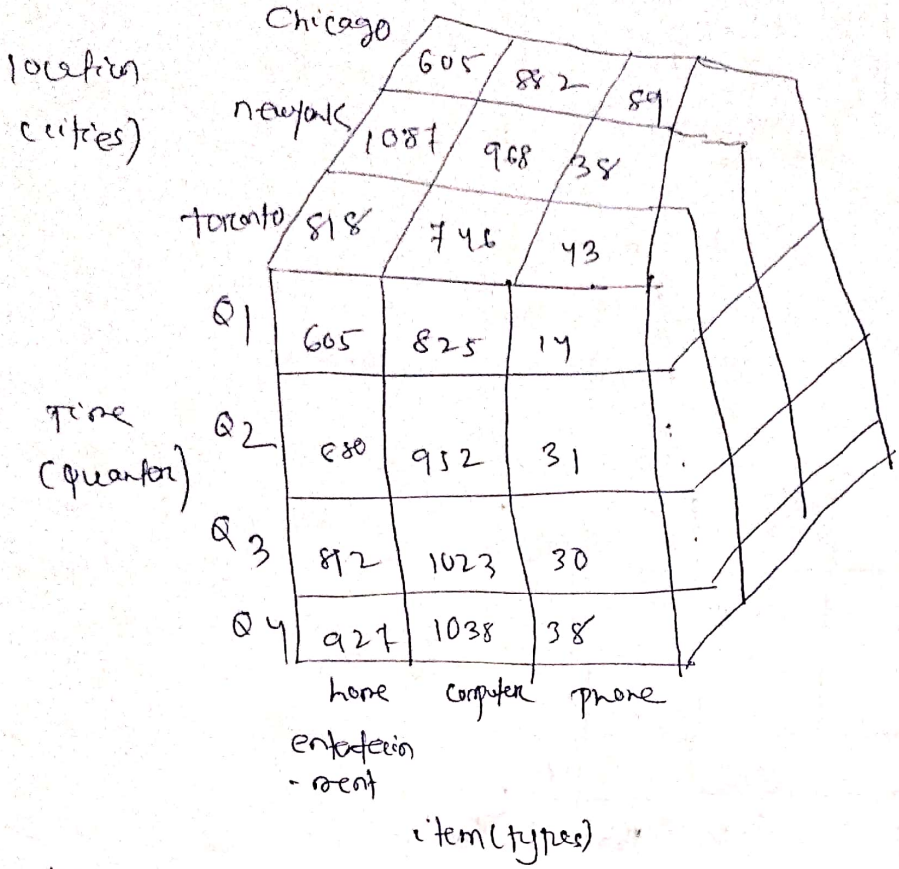
2-D

3-D cubes

→ To view data according to time, item, location for cities chicago, NEWYORK, TORONTO, ...

location = "chicago"				location = "new york"				location = "toronto"			
item				item				item			
time	home ent.	comp.	phone	time	home ent.	comp.	phone	time	home ent.	comp.	phone
Q1	605	812	89	Q1	1087	968	38	Q1	818	746	43
Q2	680	890	64	Q2	1130	1024	41	Q2	894	769	52
Q3	812	924	59	Q3	1034	1048	45	Q3	940	795	58
Q4	715	992	43	Q4	1142	1091	54	Q4	978	864	59

Here we are viewing data in three dimensions like time, items and locations.



OLAP operations

4 types of operations are there

- Roll-up
- Drill-down
- slice and dice
- Pivot (Rotate)

Pivot

→ Here, rotate the data axes to provide a substitution of data.

→ Here, rotation is done.

Ex

Location (cities)	PC	Book	Shoe	Clothes
New jersey				
Los Angeles				
Perth				
Sydney	605	825	14	400

(Item types)

↓

Pivot

↓

Location (cities)	New Jersey	Los Angeles	Perth	Sydney
PC				605
BOOK				825
Shoe				14
Clothes				400

(Location (cities))

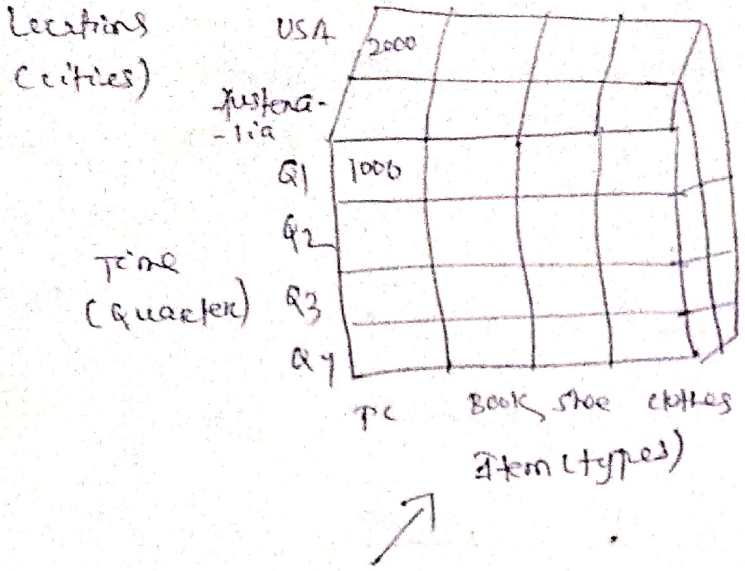
Roll up

→ It is also known as aggregation.

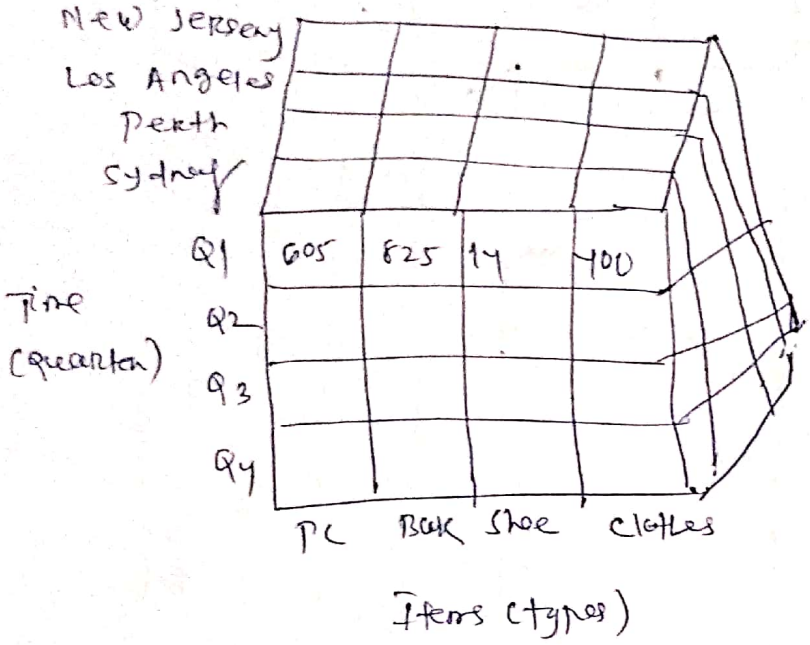
→ Roll up operation can be performed in 2 ways.

- Reducing dimensions
- climbing up concept hierarchy

→ concept hierarchy is a system of grouping things based on their order or level.



Roll-up on location (from cities to countries)



→ Here, cities Jersey, Los Angeles are aggregated into one country USA.

→ Perth, sydney into Australia.

→ Data is location hierarchy moves up from city to country.

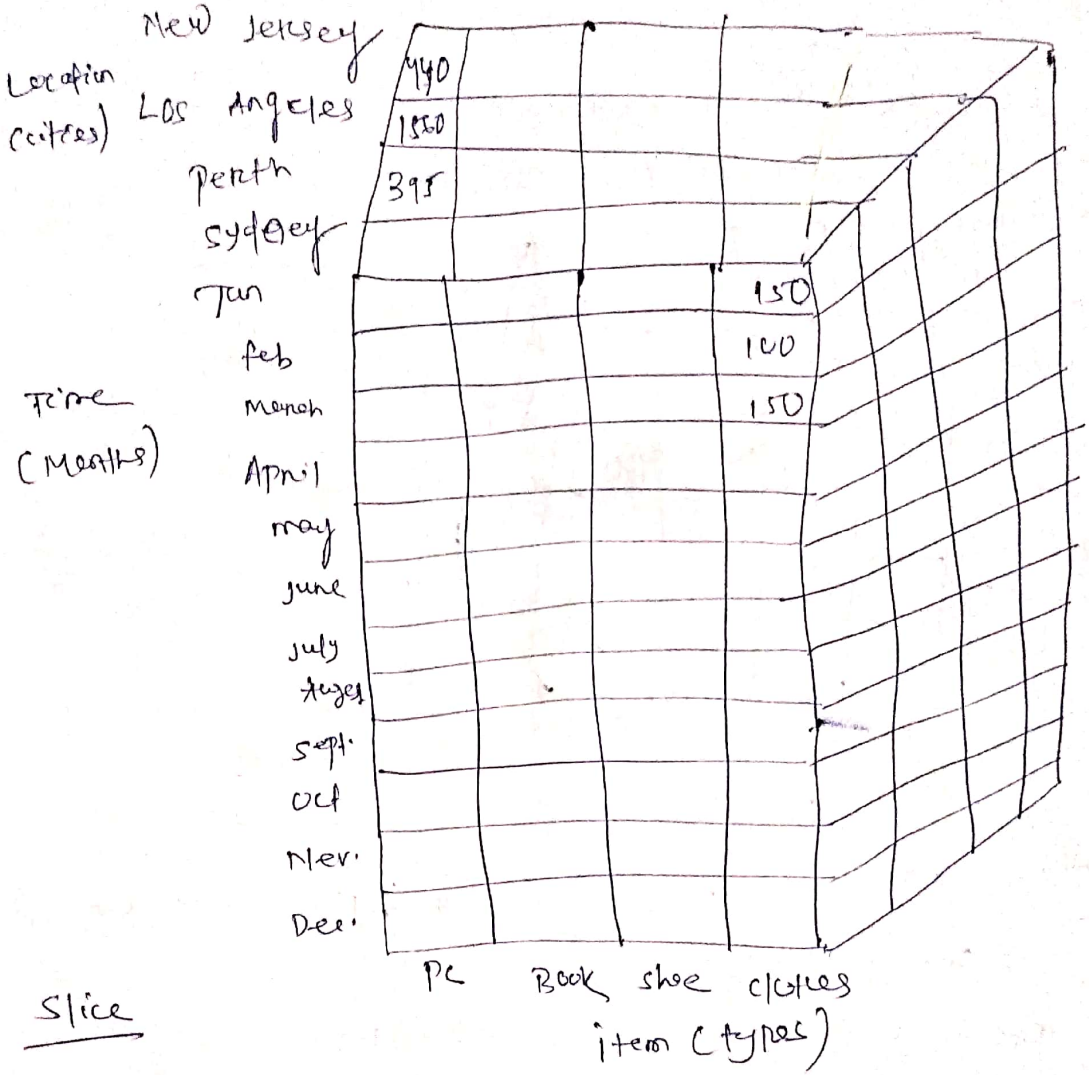
Roll - down

-) Here, data is fragmented into smaller parts.

-) It is the reverse of the roll up process.

-) It can be done by

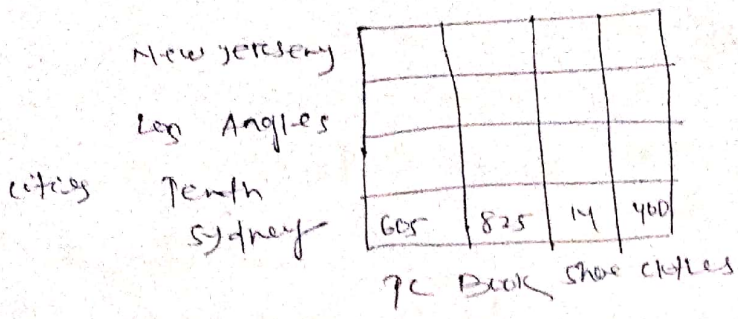
- moving down the hierachy
- increasing a dimension.



Slice

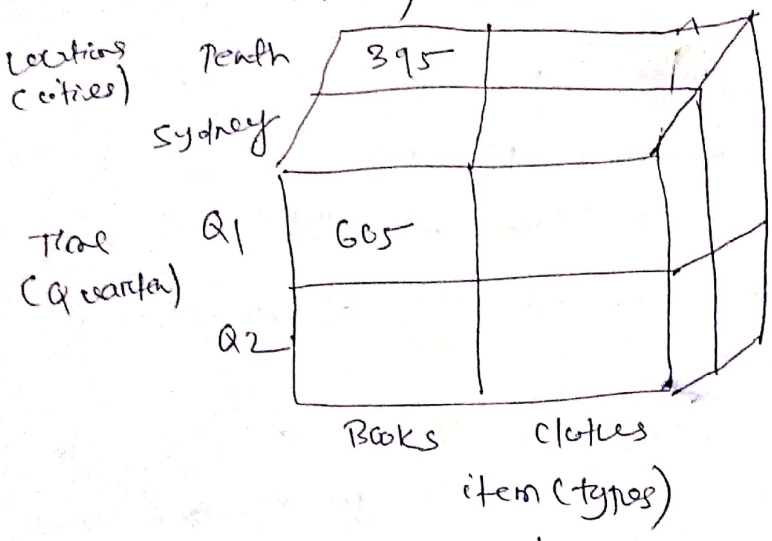
-) Here, one dimension is selected and a new sub-cube is created.

-) Here, time dimension is sliced with Q1 as the filter.

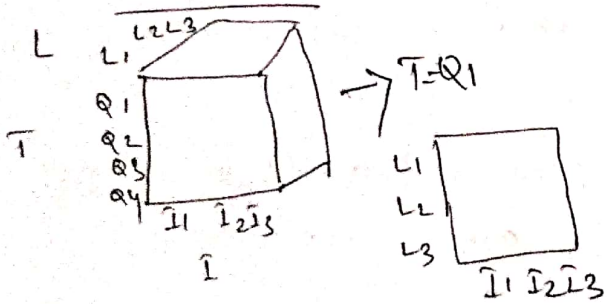


Dice :- this operation is similar to a slice. only difference is dice selected 2 or more dimensions that gives a new sub-cube.

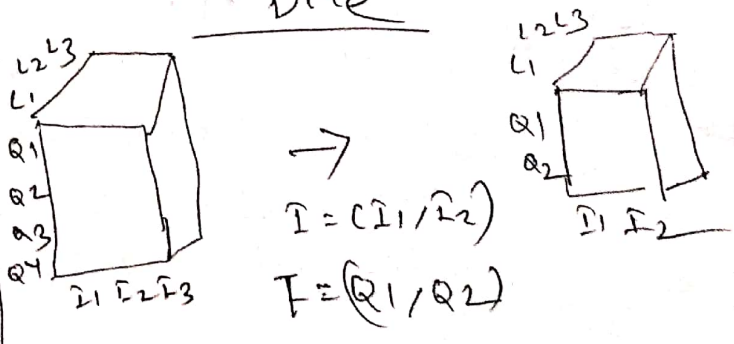
Ex Dice for (location = 'perth' or 'sydney') and (time = Q1 or Q2) and (item = 'books' or 'clothes')



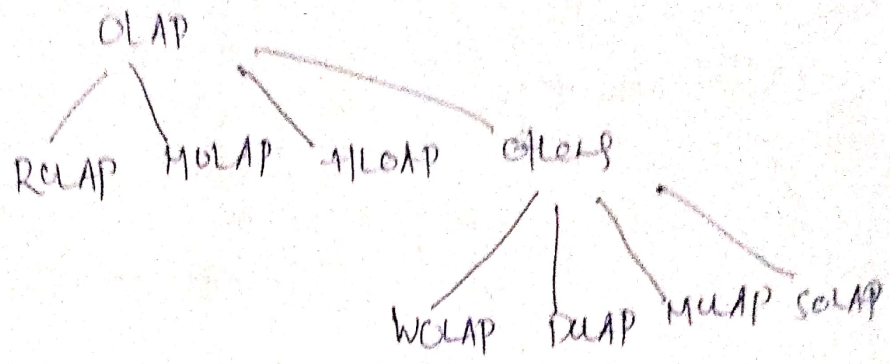
Slice



Dice



OLAP Types



→ ROLAP:-

* extension of RDBMS. * Highly scalable.
* Multi dimensional data mapping to perform the standard relational operations.

→ MOLAP (Multidimensional OLAP):

* It implements operations on multidimensional data. * information retrival is fast. * doesn't contain detailed description.

→ Hybrid OLAP:-

* Aggregate totals are stored in a multidimensional DB.
* details are stored in RDBMS
* Offers data efficiency of ROLAP and MOLAP.

→ Desktop OLAP:- User downloads a part of the data from DB locally and analyze it.

→ Web OLAP:- The OLAP which is accessible via web browser.

→ Mobile OLAP:- User will access and analyze OLAP data using mobile devices.

→ Spatial OLAP:- It is created to facilitate management of both spatial and non-spatial data in geographic information system.

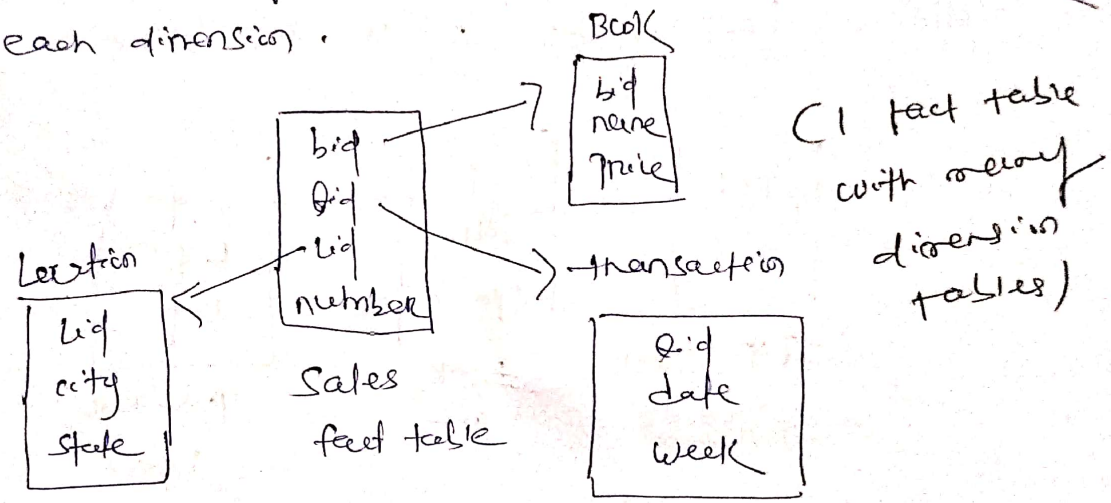
Spatial data:- large amount of historical data

Data warehouse schemas

→ schema is a collection of database objects including tables, views, indexes etc.

1) Star schema

It consists of FAct table with a single table for each dimension.

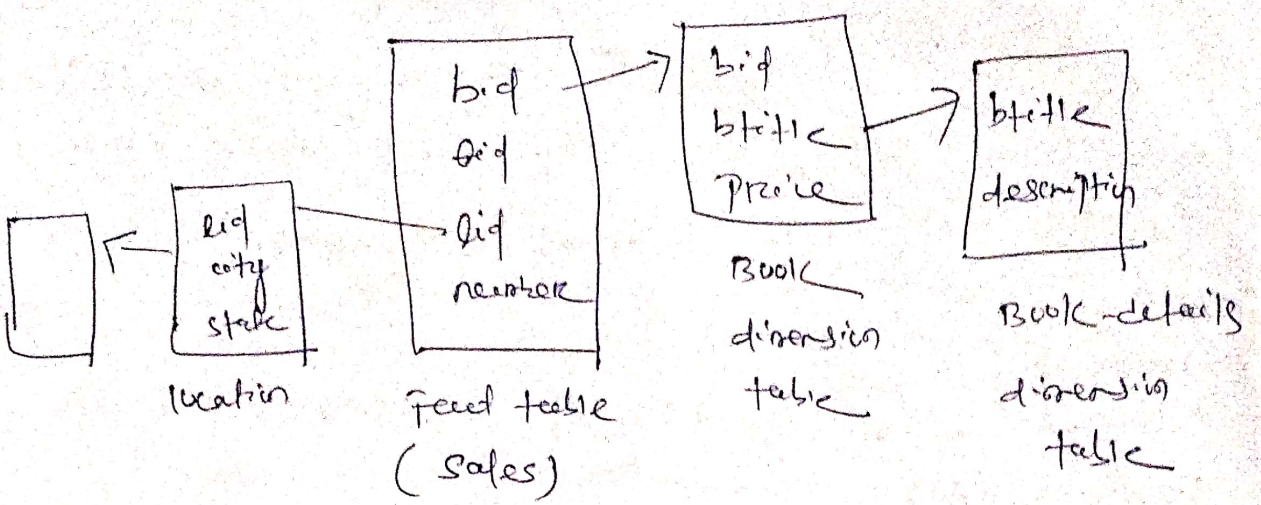


2) Snowflake schema

→ It is a variation of star schema which have multiple level of dimensions tables.

→ Dimension tables are normalized.

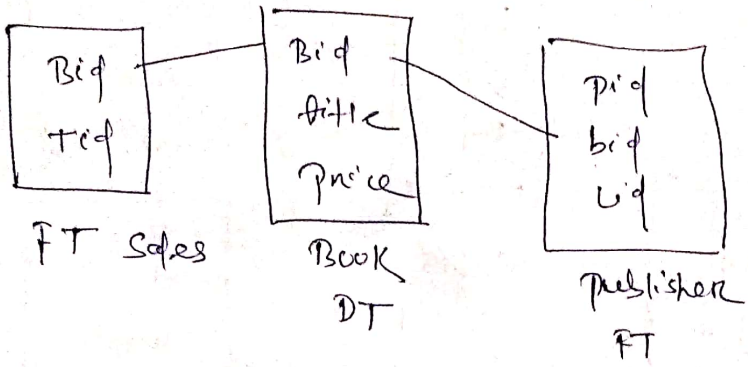
→ It is like ER-Diagram.

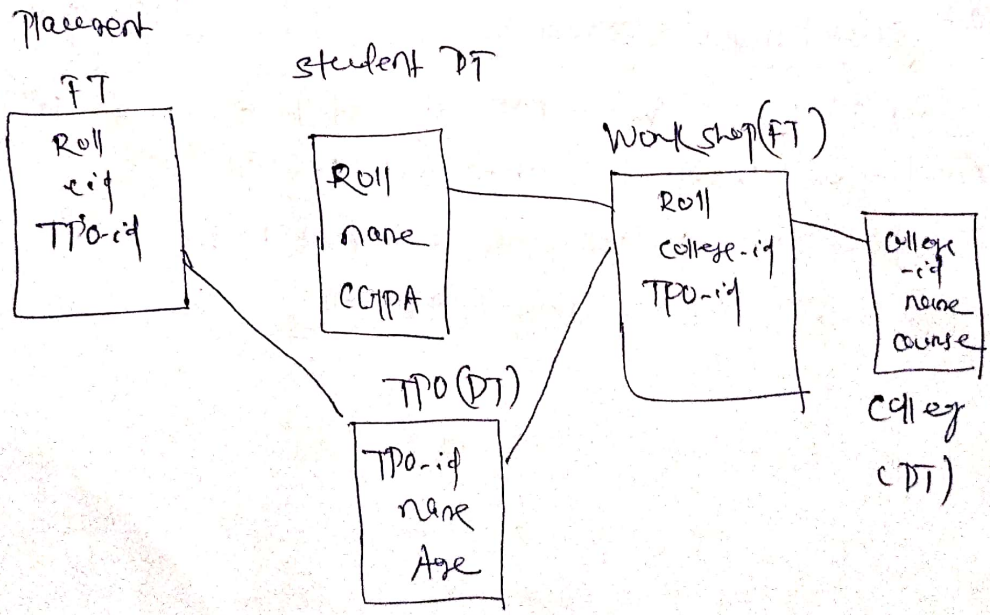


fact constellation schema

-) Also known as galaxy schema.

-) Multiple FACT tables shares the dimension tables.





OLAP vs OLTP

⇒ OLAP (online Analytical Processing) is a category of software tools that analyze data stored in a data warehouse.

Ex
A company might compare their mobile phone sales in september with sales in october.

⇒ OLTP (online transaction processing) supports transaction-oriented application. OLTP administers day-to-day transactions.

Ex ATM center
Withdraw of amount / deposit

⇒ It controls daily transactions of an organi-zation.

OLTP	OLAP
⇒ It manages DB modification.	⇒ OLAP is an analysis and data refining process.
⇒ It consists of large number of short online transactions.	⇒ It is characterized by large volume of data.
⇒ It is an online DB modifying system.	⇒ It is an online query management system.
⇒ It uses DBMS.	⇒ It uses Data warehouse.
⇒ Databases are normal-ized.	⇒ Tables in OLAP DB are not normalized.

OLTP

→ Different transactions are the sources of data.

→ It helps to control and run fundamental business task.

→ Simple queries are used.

→ Processing time is less.

→ OLTP must maintain data integrity constraint.

→ It focus on insert, delete, insert information from the database.

OLAP

→ Different OLTP database-ses are the sources of data for OLAP.

→ It helps with Planning problem-solving and decision support.

→ Complex queries, involving aggregations.

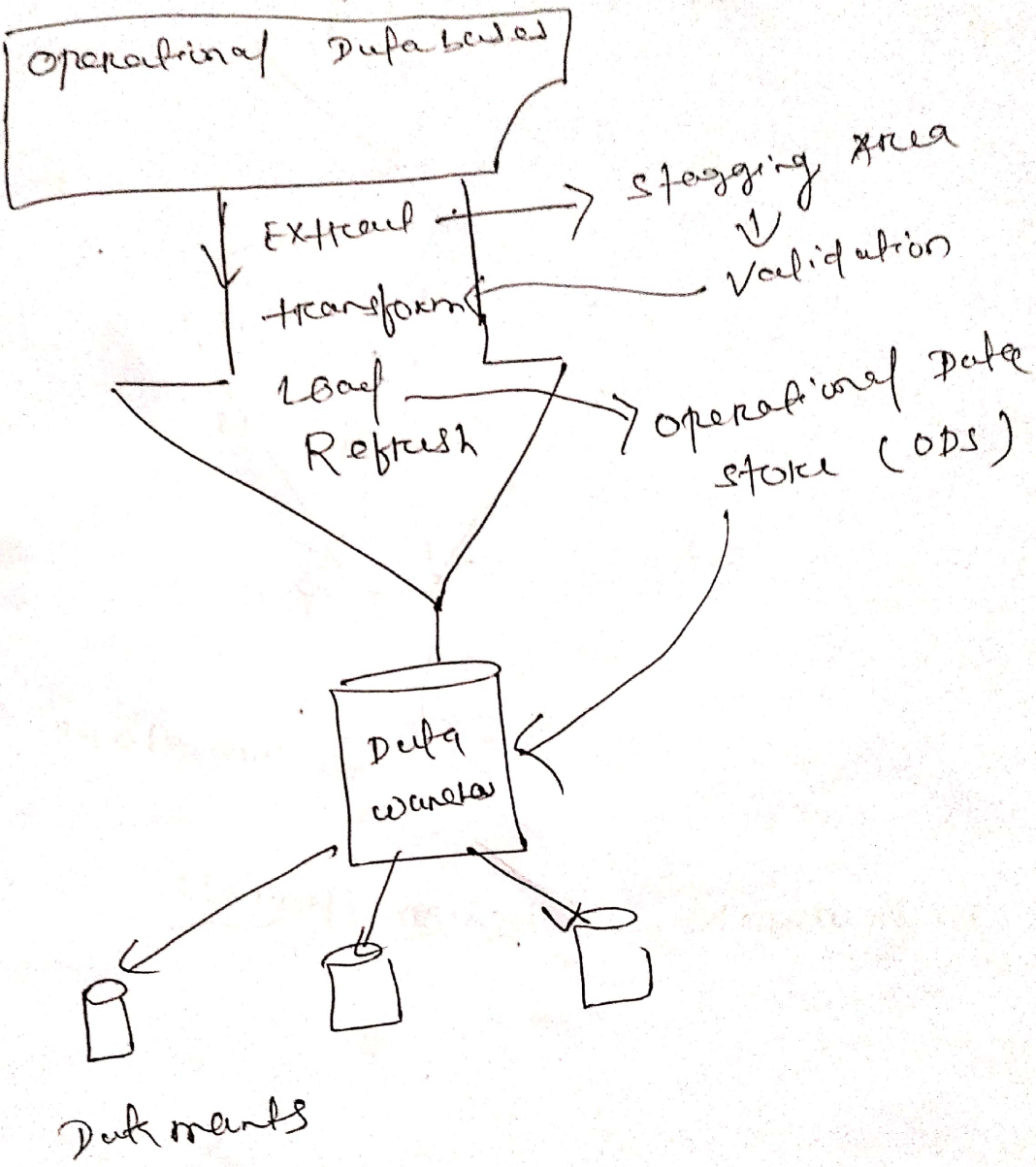
→ Processing time is more.

→ OLAP DB doesn't get frequently modified. Hence, data integrity is not affected.

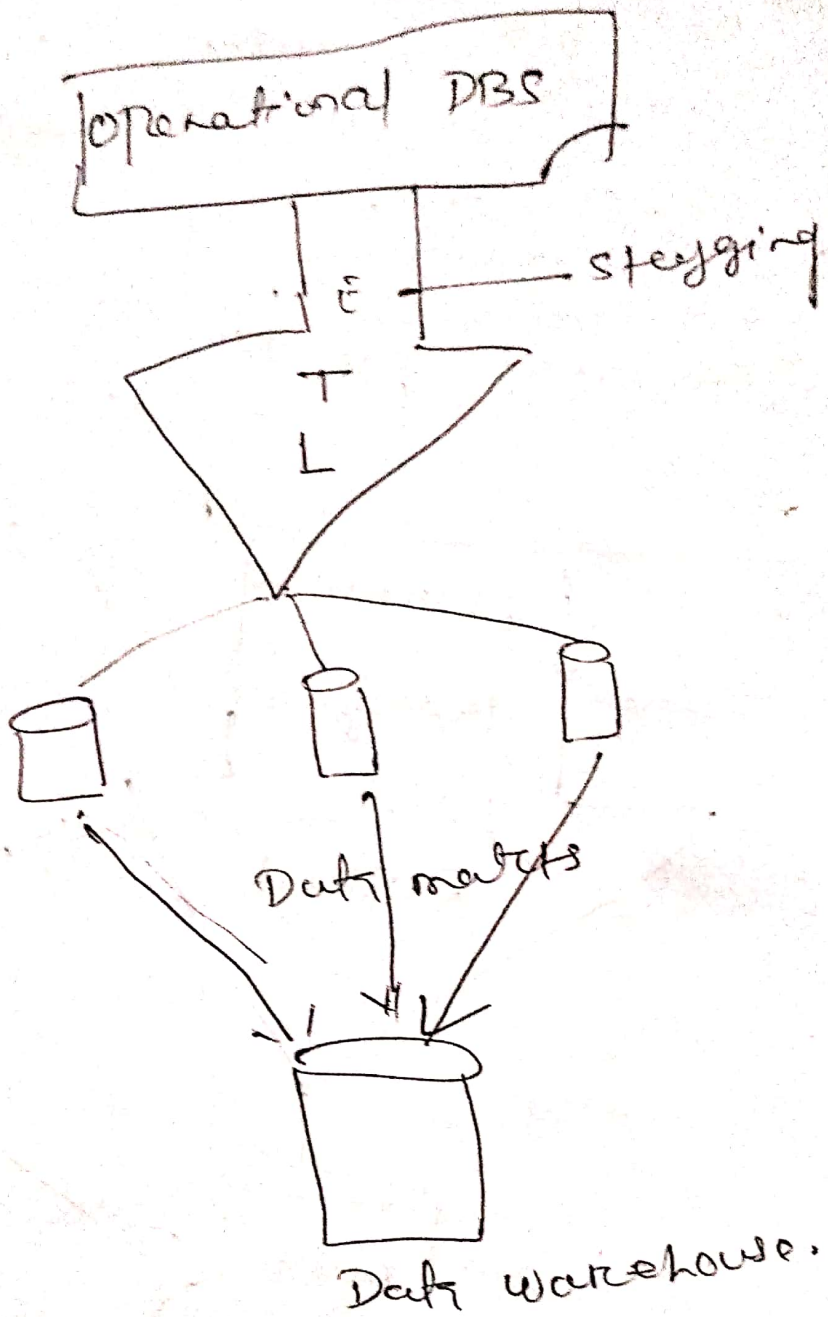
→ Extract data for analyzing that helps in decision making.

Approaches for building DW

1) Top down approach



2) Button-up Approach


Operational DBS
E
T
L
staging area
Data marts
Data warehouse.

Top - down Approach

→ It is considered as data driven approach.
→ first data is collected & integrated.
→ After integration data marts are created.

Steps

→ Raw data are collected from external sources.
→ store the data into staging area.
→ Apply ETL operation.
→ Then load into Datawarehouse.
→ Then we can create data marts which stores the information of subject wise.

Advantages

→ Developing new data mart from Datawarehouse is very easy.
→ Data marts provides consistent dimensional view of data marts.

Disadvantages

→ we can't change the dependental data when requi-red.
→ The cost of implementation is high.

Bottom - up Approach

→ first data is extracted from external sources.
→ Then store into staging area and load into Data marts.

→ Here, first data marts are created.
→ Now, data marts are integrated into Data warehouse.
→ Then apply data mining techniques.

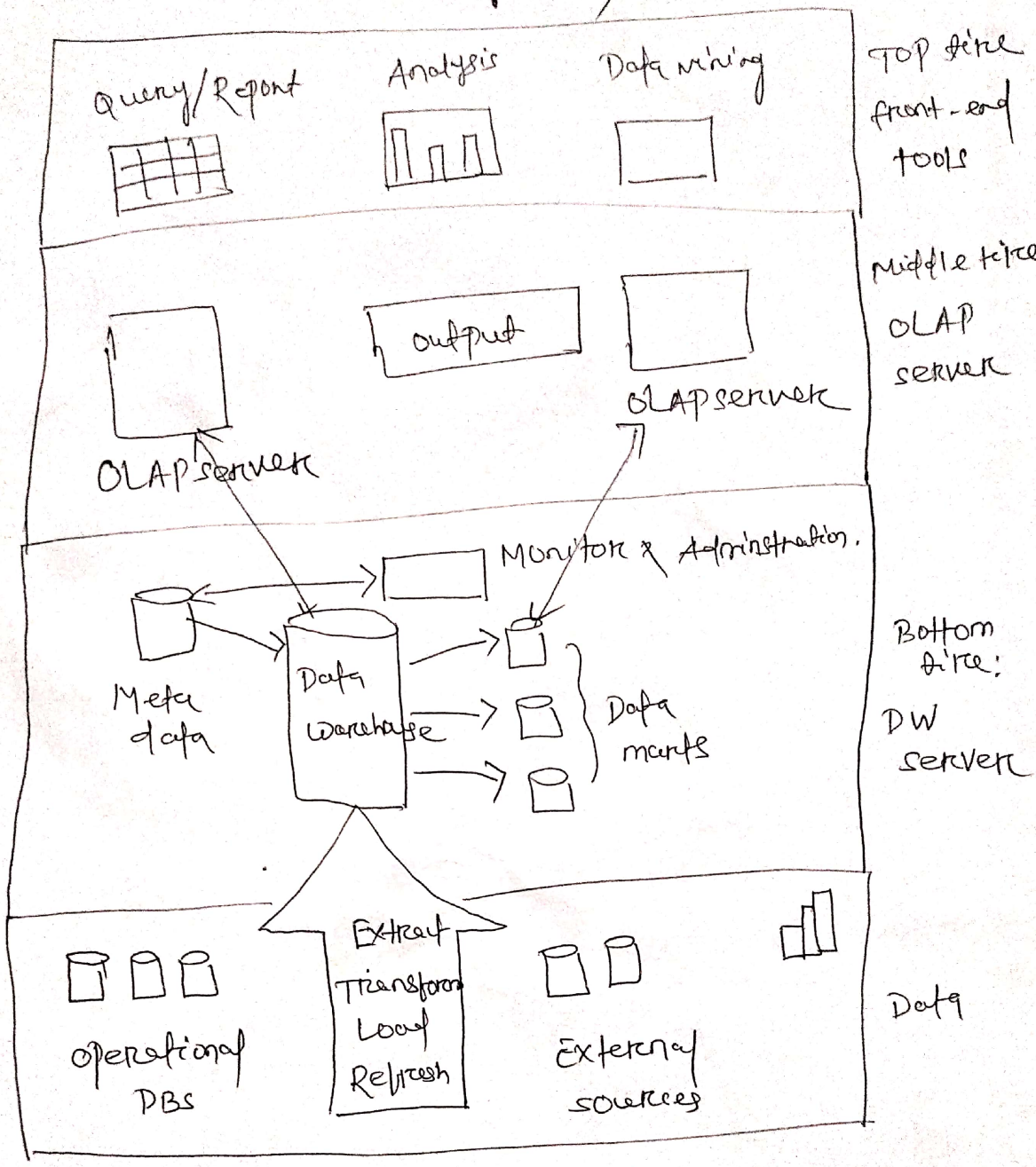
Advantages

→ Reports are generated quickly from the data marts.
→ Data warehouse can be extended by adding more number of data marts.
→ Cost & time taken in designing this model is low.

Disadvantages

→ The data marts are not consistent, so the dimensional view of data are inconsistent.

Datawarehouse Architecture (3-tire Architecture of DW)



Bottom tire:-

→ The database of datawarehouse servers are the bottom tire.

→ It is usually a relational Database systems.

→ Here, Data is cleaned, transformed & loaded into this layer.

Middle layer:-

→ This layer contains OLAP servers which is implemented using ROLAP or MOLAP.

→ This layer presents an abstract view to the user.

→ This is an interface between end-user and the data warehouse.

TOP-tier:-

→ This layer is for end users.

→ It contains tools / APIs that can help to retrive data from DW.

→ The tools avilable at this layer are

- Query & Reporting tools.
- Analysis tools
- Data mining tools

→ Basically it is a front-end client layer.