DWDM

Unit-I

Data:- Data is a collection of a distinct small unit of information. It can be used in a variety of forms like text, numbers, media, bytes, etc. It can be stored in pieces of paper (or) electronic memory etc.

DB:- A database is an organised collection of data so that it can be easily accessed and managed. You can organize data into tables, rows, columns and index it to make it easier to find relevant information.

DWH:- → Data warehouse is a subject oriented, integrated, time variant and non-volatile collection of data from heterogeneous sources to provide meaningful business insights.

A DW is a separate from DBMS, it stores a huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS, etc.

need for DW:- An ordinary Database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to Data warehouse.

Eg:- Social media websites:- like facebook, twitter etc. These are based on analyzing large data sets.

These sites gather data related to members, ... location etc.

* Banking - Most of the banks these days use warehous... See the spending patterns of account/card holders.

Government → Tax payments.

Difference b/w DW and operational DB :-

	DW	ODB
Basic	A dw is a repository of structured, filtered data that has already been processed for a specific purpose	These are those database where data changes frequently.
D.S	DW has de-normalized schema	It has normalized schema
	It is fast for analysis Query	It slow for analysis queries.
	It is used for OLAP.	It is used for OLT

Characteristics of DW

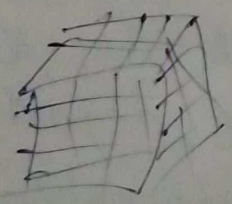
- Subject oriented → Customers, emp, Products
- Integrated → Combined → It allows for easy access and analysis of the data
- Time-Variant → data is stored with a time dimension. It allows easy access to data for specific Periods, such as last quarter (or) last year.
- Non-volatile → never updated (or) deleted only added

Benefits of DW:-

1) Better business analytics:-
2. Faster Queries:- It is designed to handle large Queries that's why it runs queries faster than the DB.
3. Improved data Quality:-
4. Histrical Insight:-

Top layer:-
Front end tools:- Users of DW use these tools
meta data → Data about the data.
Data mart → Sub part of the DW/
Sequence of the DW.

OLAP → Online Analytical process
It is used to Analyze the multi dimensional data.



Top layer:- It is communicate b/w the users and DW.

Single Tier architecture:- It is not periodically used in practice. Its purpose is to minimize the amount of data stored. only layer physically available is the source layer. In this method dws are virtual. This means the dw is implemented as a multidimensional view of operational data created by specific middleware (or) an intermidiate processing layer.

Two-tier architecture :-

The requirement for separation plays an essential role in defining the two-tier architecture for a DW system.

→ It is typically two-tier architecture to highlight a separation b/w physically available sources and DW.

4 flow stages

① Source layer :- A DW system uses a heterogeneous source data that data is stored initially to corporate relational databases.

② Data Staging :- 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.

③ DW layer :- Information is saved to one logically centralized individual repository : a DW. The DW can be directly accessed, but it can also be used as a source for creating DM, which partially replicate DW content and designed for a specific enterprise department

Meta data repositories store information on sources, access procedures, data staging, users, data mart schema and soon

Analysis :- In this layer, integrated data is efficiently and flexible accessed to issue reports, dynamically analyze information and simulate hypothetical business scenarios. Feature aggregate information navigators, complex query optimizers and customer-friendly GUIs.

Reconciled layer is that creates a standard reference data model for a whole enterprise & it is also directly used to better some operational tasks such as producing daily reports

Components of the dw:- A database serves as the foundation of your dw. Traditionally, these have been standard relational databases running on premise (or) in the cloud. But because of Big data, need for true, real-time performance, drastic reduction in the cost of RAM, in-memory db are rapidly gaining in popularity. It allows users to interact with the data in your dw. Eg:- query & reporting tools, Appl'n development tools, dm tools, and OLAP tools

ETL process:-

Extract:- The 1st stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets and flat files.

This step involves reading data from the source systems and storing it in a staging area.

(2) Transform:- The extracted data is transformed into a format this is suitable for loading into the dw. This may involve cleaning and validating the data, converting data types, combining data from multiple sources and creating new data fields.

(3) Load:- After the data is transformed, it is loaded into the dw. This step involves creating the physical data structures and loading the data into the warehouse.

Adv:-

Improved the quality.
Better data integration
Increased data security
Improved scalability
Increased automation

Disadvantages

① High cost
② Complexity
③ Limited flexibility
④ Limited scalability
⑤ Data privacy concerns

Disadvantage:

Data modeling:-

designing Schema

Process of designing schema of detailed and Summarized information of dw.

need:- In order to improve the efficiency of a sys.
Support complex queries.

3 levels:-

1. Conceptual:- It is mainly explain the semantics of the data (meaning)
2. Logical:- It defines all the information/data in structural format. It contains such as DS, procedures, rules, rel^n so many things.
3. physical:-
It is describes how the data is presented in dw?
like tables, columns, foreign keys.

Dw models: ① Enterprise warehouse → go to TBT → implementing mainframes, super servers

It contains all information/data about the subject related to entire organisation (detailed and Summarized)

② Data mart:- It contains data specific to group of users (not entire organisation) only summarized data

↳ Independent ② dependent

③ Virtual warehouse:- It contains data copied from multiple sources during a production.
↳ Data can be easily search.

② It is based on multidimensional data model and allows the users to query on multi-dimensional data

OLAP Cube stores: It is the technology that allows users to analyze information from multiple db systems at the same time.

OLAP db's are divided into one (or) more cubes and Hypercubes these cubes are known as:

OLAP is a data structure for quick data analysis.

ETL → data is loaded onto OLAP cube.

after that different operations performed on the OLAP cubes: - analysis of the data.

It can be performed even using the Spread Sheets also.

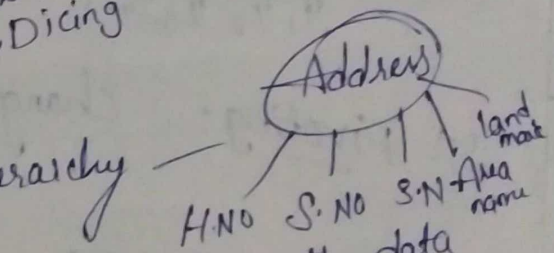
Generally, operations can performed by the Spread Sheets in two-dimensional data but DWH is multi-dimensional data.

OLAP Operations:-

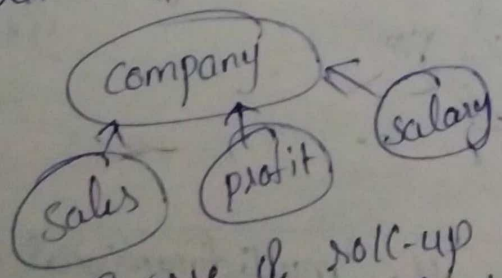
4- Operations:- ① Roll up (drill-up)
② drill-down
③ Slicing & Dicing
④ pivot

1) Roll-up (drill-up)

- It uses the concept of hierarchy — high data abstraction



② When Roll-up is performed one (or) more dimensions from the data cube removed.

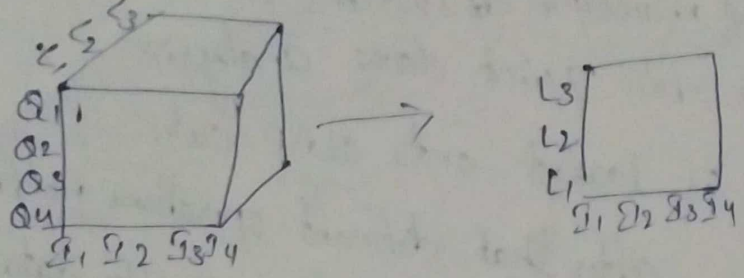


(2) Drill-down:- Reverse of roll-up operations.
data is drilled down.
low data abstraction.

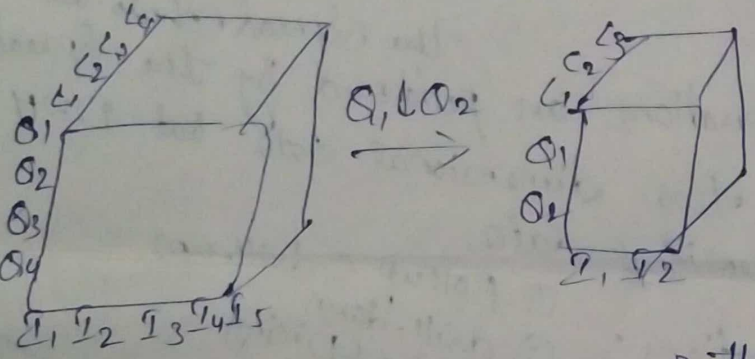
3. Slicing and dicing:

It is the selection technique.

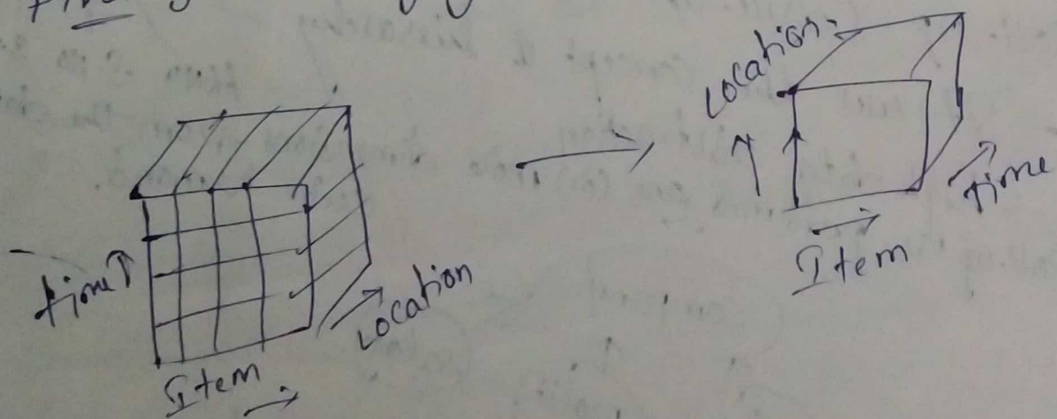
Data cube is sliced and information is divided and new cub is formed.



Dicing:- more than one dimension one is considered.



4) Pivoting: Changing the axes of the data (Rotation).



OLAP based on the multidimensional data model.

OLAP implement the multidimensional analysis of business information.

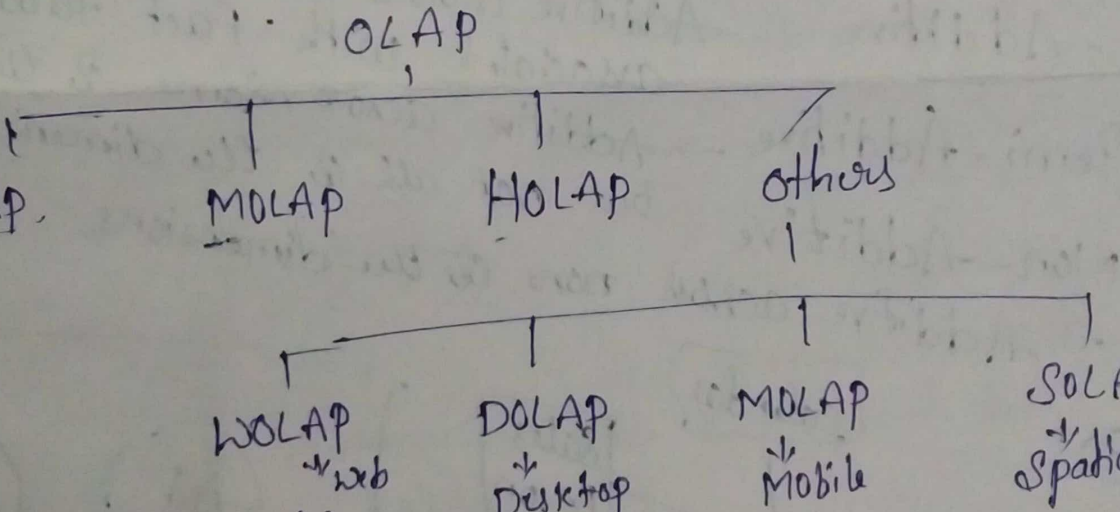
OLAP support the capability to complex estimation treat analysis and sophisticated data modeling.

OLAP Applications are used in: ① Finance and Accounting
② Sales and marketing
③ production

OLAP Characteristics:-

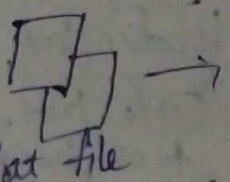
1. useful in analyzing the business.
2. It focus on Information out.
3. contains historical data.
4. No. of users is in hundreads.
5. no. of record accessed is in millions.
6. Database size is from 100 GB to 1TB.
7. Highly flexible.

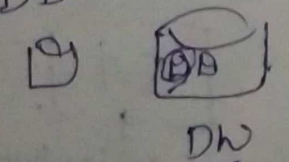
OLAP ~~Applic~~ Server ~~Architecture~~:-



ROLAP: (Relational)

- extension of RDBMS. → will be represented in tables.
- capable of large amount of data analysis.
- can store and analyse changing amount of data also





2. MOLAP :- (multidimensional)

- Limited amount of data.
- Arrays are used.
- Fast information retrieval.
- Complex calculations.

* It is not ... extend of RDBMS

1GB
↓
2GB

3. HOLAP (Hybrid) :-

ROLAP + MOLAP.

Here, Detailed data and Fast information retrieval.

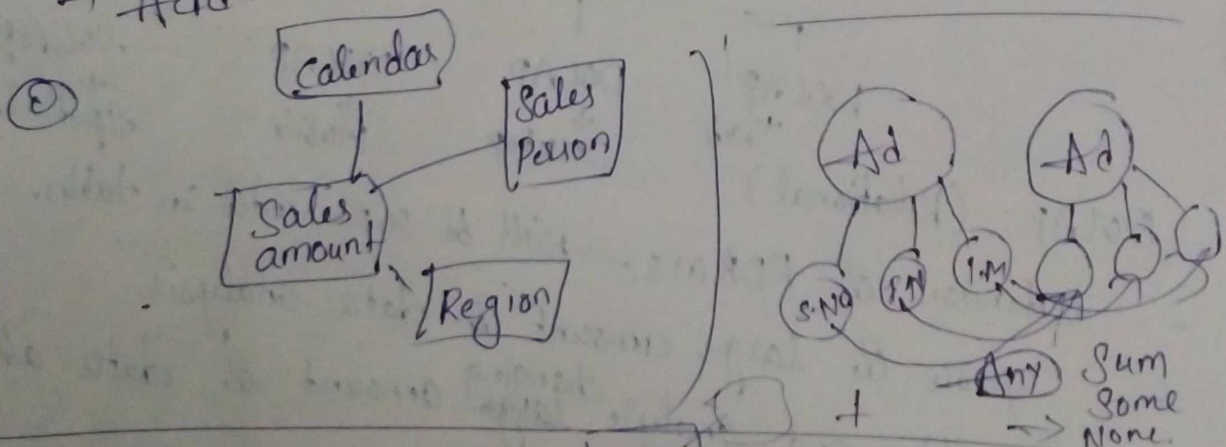
It is large amount of data.

Types of measures in fact table :-

Additive :- Additive across all the dimensions associated with fact table.

Semi-Additive → Additive across some of the dimensions but not all of the dimensions.

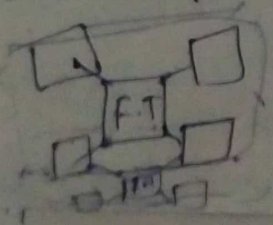
Non-Additive → Additive across none of the dimensions.

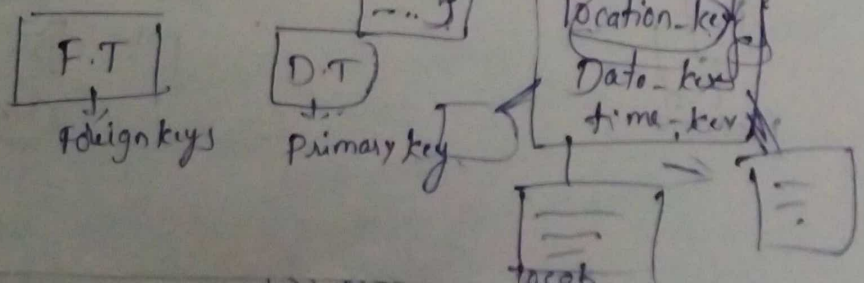


Star Schema →

Snowflake →

Fast Constellation :-



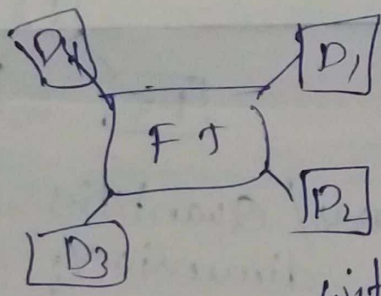


Schema :- It is a logical description of the entire database. It includes the name and description of records of all record types. (Will be represent in structural format) → collection of database objects.

Fact tables and dimension tables are joined logically.

① Star Schema.
② Snowflake "
③ fact constellation.

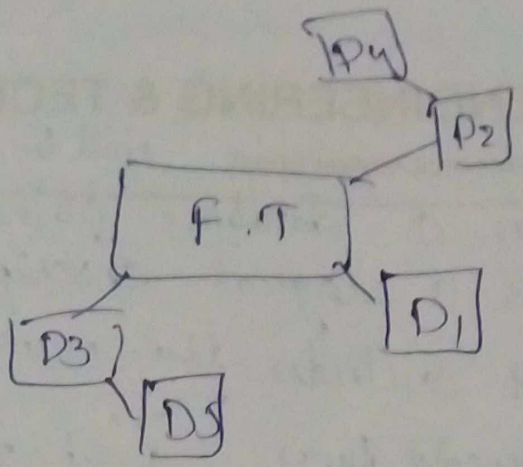
① 2 important components - FT & DT
only 1 fact table & many dimensional tables.
→ The → primary keys
will contains foreign keys
→ It is every dimensional table.



① foreign keys.
Each foreign key associated with dimensional table.
Each dimensional table ~~has~~ will have the primary key.

Here, fact table can be normalized → removing the redundency in the data
dimension table can't be.

② It is a type of Star schema. where dimensional tables are also normalized.
How → By existing dimensional table you will be deriving the new D.T.
→ It will manages size of the f.T.



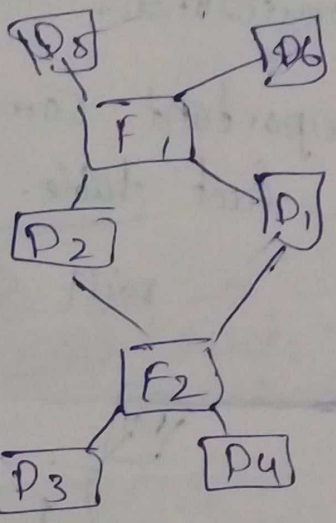
D2, D1, D3 is the existing ones.
D5 is derived out of D3.
D4 is derived out of D2

3) Fact constellation Schema:-

It is Group of Star Schemas.

- Complicated design.
- Hard to understand and implement.

Many fact table & DT.



Fact:- numerical measures/quantities by which can analyze relationship b/w dimensions.

Dimensions:- collection of logically related attributes used to modelling the data.

Fact table:- relations b/w multi dimensional data.

Dimension table:-
Tables related to each dimension and helps in describing dimension further.

Types of