B.C.A. (Sem – VI)

B.C.A. - 603

Data Warehousing & Data Mining

Purushottam Singh

Unit:- 2

Data Warehousing & Mining

UNIT-2

❖ Multi dimension Data Model:-

- The Dimension Data Model was developed for Implementing Data warehouse and Data Marts.
- This is not a 3-dimensional cube, it is n-dimensional cube
- Such as a model to show star schema, snowflake schema or fact constellation schema.
- Fact constellation is multiple fact tables share dimensions tables, view as a collection of stars therefore galaxy schema or fact constellation.
- The core of the multidimensional model is the **data cube**, which consists of a large set of facts and number of dimensions.
- A data cube consists of a lattice of cuboids, each corresponding to a different degree of summarization of the given multidimensional data.
- A data cube such as sales, allows data to be modeled and viewed in multiple dimensions.
- A Multidimensional data model is typically organized around a central theme, such as sales. This theme is represented by a fact table.
- Fact table contains measures (such as dollars sold) and keys to each of the related dimension tables.
- Dimensions of the cube are the equivalent of entities in a database, e.g., how the organization wants to keep records.
- Dimension tables, such as item (item name, brand, type), or time (day, week, month, quarter, year).

Data Warehousing & Mining

❖ Schema for Multidimensional Database:-

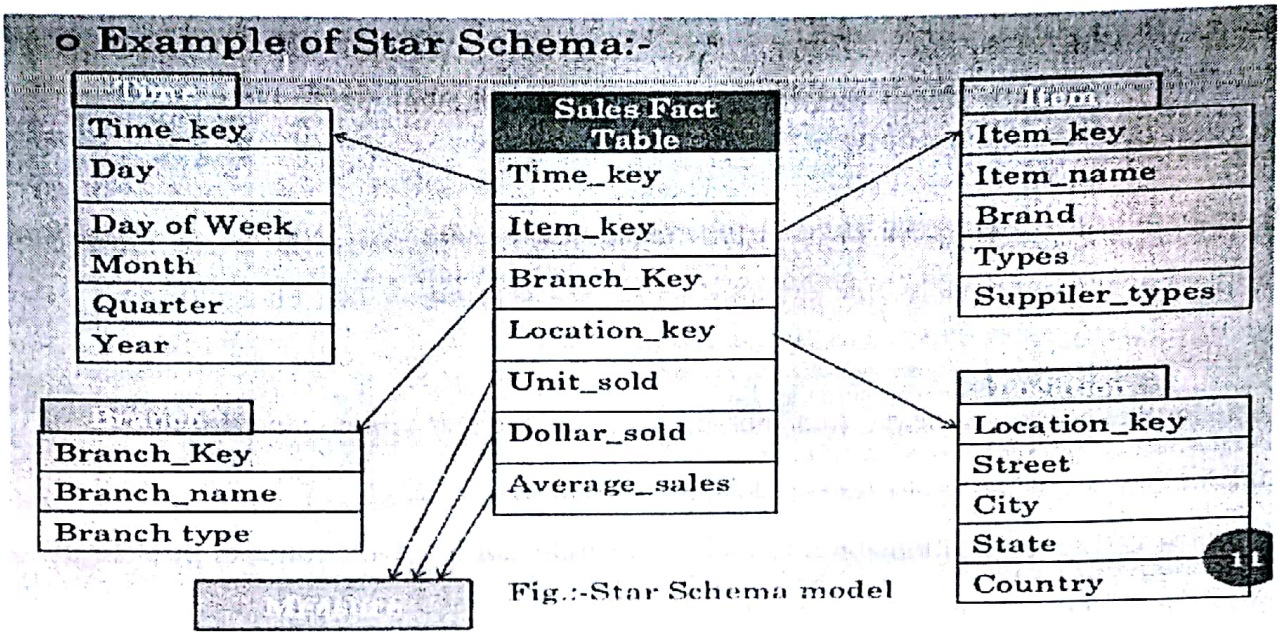


Fig.:-Star Schema model

❖ Star Schema Model:-

- It is also known as Star Join Schema.
- It is the simplest style of data warehouse schema.
- It is called a Star Schema because the entity relationship diagram of this Schema resembles a star, with points radiating from central table.
- A star query is a join between a fact table and a no. of dimension table.
- Each dimension table is joined to the fact table using primary key to foreign key join but dimension table are not joined to each other.
- A typical fact table contains key and measure.

❖ Advantages

- Easy to understand
- easy to define hierarchies
- low maintenance,
- very simple metadata

Data Warehousing & Mining

❖ Disadvantages:-

- Summary data in the fact table yields poorer performance for summary levels,
- huge dimension tables a problem

❖ Snowflake Schema Model:-

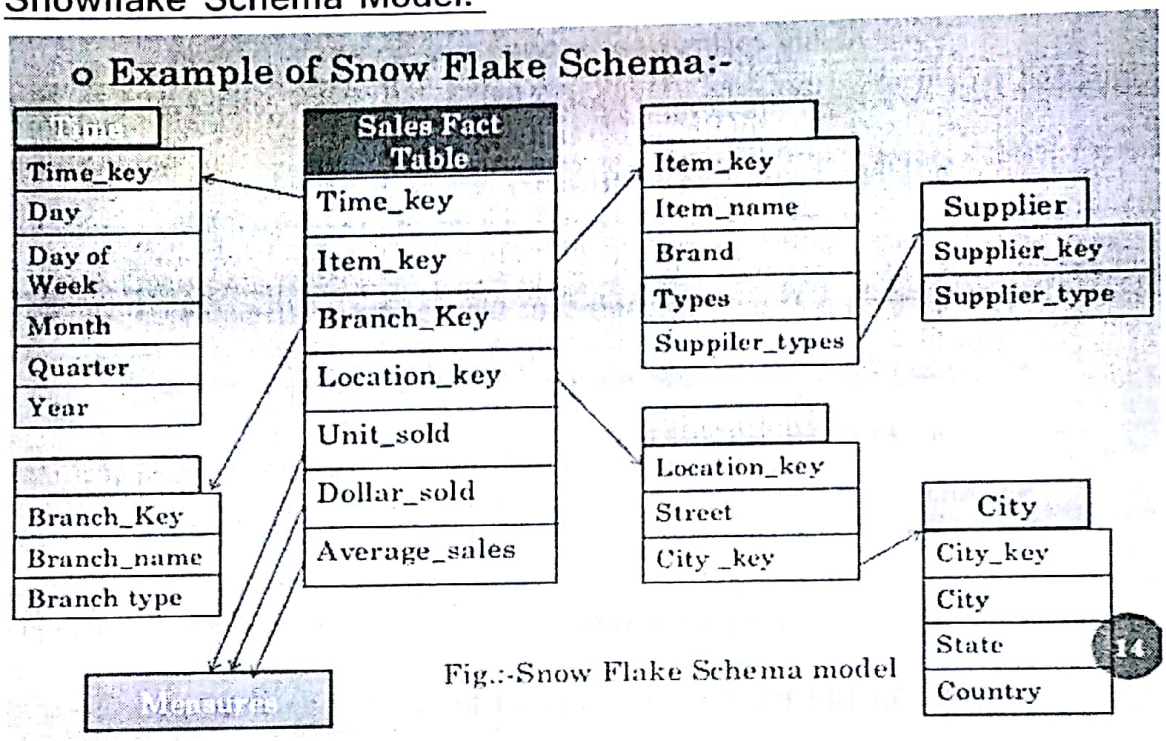


Fig.:-Snow Flake Schema model

- Snowflake schema is a type of star schema but a more complex model.
- It is slightly different from a star schema in which the dimension tables form a star schema is organized into a hierarchy by normalizing them.
- The normalization eliminates redundancy.
- The Snow Flake Schema is represented by centralized fact table which are connected to multiple dimensions.
- The Snow Flaking Effecting only affecting the dimension tables and not the fact tables.

- The result is more complex queries and reduced query performance.

Advantages:-

- Small saving in storage space
- Normalized structures are easier to update and maintain

Disadvantages:-

- Schema less intuitive and end-users are put off by the complexity
- Ability to browse through the contents difficult
- Degrade query performance because of additional joins.

❖ MDDM Base Concepts:-

- A Multi Dimensional data model is to view it as a CUBE.
- CUBE is a Data Structure that allows fast analysis of data.
- It can also be defined as the capability of manipulating and analyzing data from multiple perspectives.
- MDDM provide both a mechanism to store data and a way for business analysis.

❖ Component of MDDM:-

- The MDDM involve two types of table.

(1) Dimensions:-

- Texture attributes to analyses data.
- It is a simple primary key.

(2) Facts:-

- Numeric volume to analyze business.
- It is Compound Primary key.

❖ Geographic Interface System (GIS):-

Data Warehousing & Mining

- A GIS is an organized collection of computer, hardware, software, geographic data and personnel to efficiently capture, store, update, manipulate, analyze and display all forms of geographically referenced information.
- GIS is a computer system capable of assembling, storing, manipulating and displaying geographically referenced information. EX:- data identified according to their locations.

❖ Principle of GIS:-

- Database Management and Update:- data security, data integrity, data storage retrieval and data maintenance abilities.
- Geographic Analysis:- The collected information is analyzed and interpreted qualitatively and quantitatively.
- Preparing Result: - one of the most exciting aspects of GIS technology is the variety of different ways in which the information can be presented.

❖ Components of GIS:-

- Hardware:- computer system, scanner, printer, plotter, Flat Board.
- Software:- GIS software in use are MapInfo, ARC/Info, AutoCAD Map, etc. The software available can be said to be application specific.
- Data:- Geographic data and related tabular data can be collected in house or purchased from a commercial data provider.

❖ Advantages:-

- GIS give the accurate data.
- Better Predictions and Analysis.

❖ DisAdvantages:-

- Expensive software
- Integration with traditional map is difficult.

❖ Relational Database :-

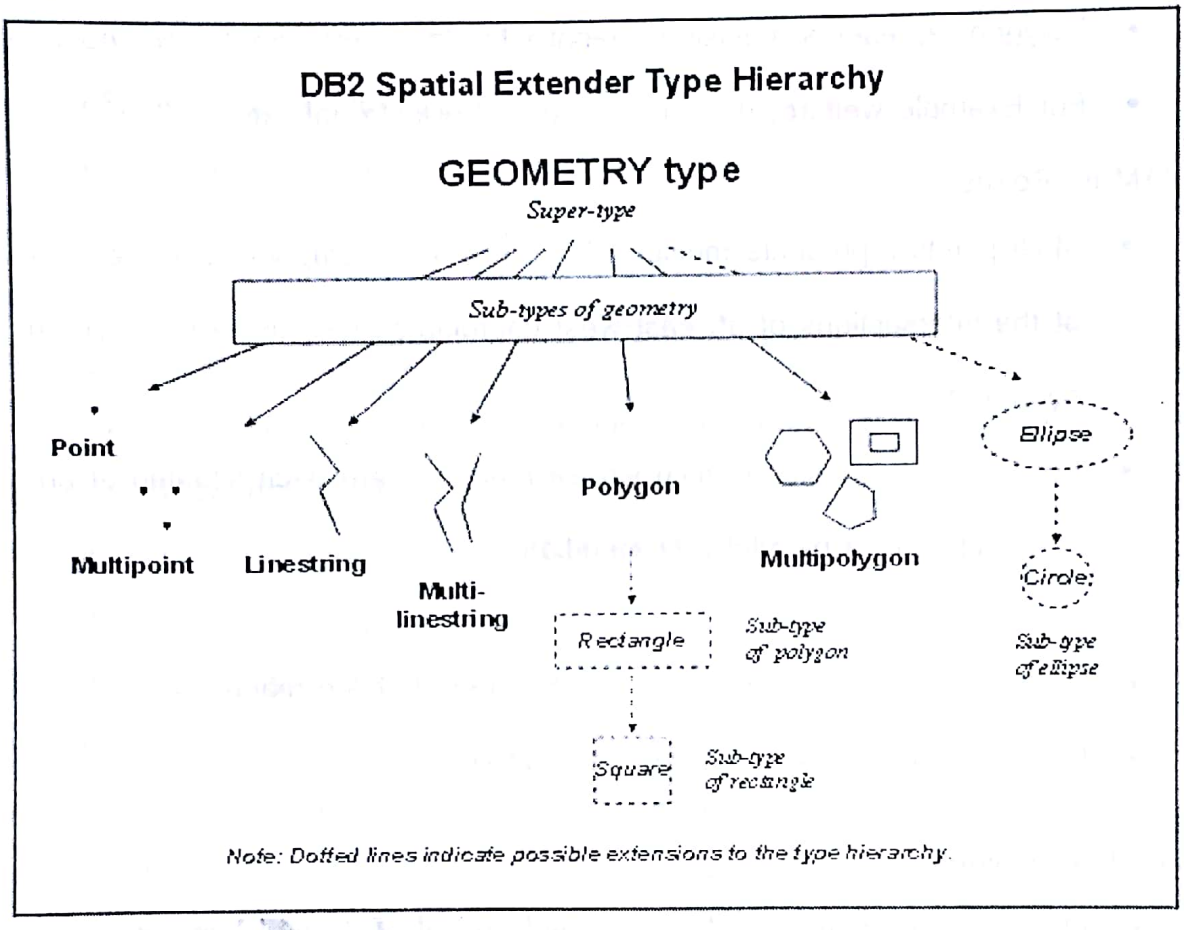
- A relational database is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute value.
- A semantic data model such as an entity-relationship (ER) data model is often constructed for relational database.
- An ER data model represents the database as a set of entities and their relationship.
- Relational data can be accessed by database queries written in a relational query language, Ex.:-SQL or with the assistance of graphical user interface.
- A given Query is transformed into a set of relational operations, such as join, selection and projection,and is then optimized for efficient processing.

❖ DB2 Spatial Extender:-

- Use DB2 Spatial Extender to generate and analyze spatial information about geographic features, and to store and manage the data on which this information is based.
- It provides special functions and indexes for querying and manipulating that data
- using something like Structured Query Language (SQL).
- A geographic feature is anything in the real world that has an identifiable location, or anything that could be imagined as existing at an identifiable location.

Data Warehousing & Mining

- In DB2 Spatial Extender, a geographic feature can be represented by one or more data items; for example, the data items in a row of a table. (A data item is the value or values that occupy the cell of a relational table.)
- A spatial database gives you both a storage tool and an analysis tool.



❖ Types Of DB2 Spatial Extender:-

(1) Point:-

- Point represents discrete features that are perceived as occupying the locus where an east-west coordinates line (such as a parallel) intersects a north-south coordinate line (such as meridian).
- For example suppose that the notation on a large scale map show that each city on the map is located at the intersection of a parallel and a meridian. A point could represent each city.

(2) Line string:-

- Line string represents linear geographic features.
- For Example streets, canals and pipelines.

(3)Polygon:-

- Polygon represents multisided geographic features.
- For Example welfare, districts, forest and wildlife habitats.

(4)Multi Points:-

- Multi points represents multipart feature whose components are each located at the intersections of an east-west coordinates line and a north-south coordinate line.
- For Example an island chain whose members are each situated at an intersection of a parallel and meridian.

(5)Multiline string:-

- Multiline string represents multipart features that are made up
- For Example river system, highway system.

(6)Multi polygon:-

- Multi polygon represents features made up of multisided units or components.
- For Example the collective farmlands in a specific region, or system of lakes.

❖ Difference Between OLAP V/S OLTP:-

- OLAP is Online Analytical Processing.
- OLTP is Online Transaction Processing.
- OLAP is customer-oriented.
- OLTP is market oriented.

Data Warehousing & Mining

- The Online Analytical processing is used for data analysis by clients, IT professionals, and clerks.
- The Online Transaction Processing is used for analysis of the data by executives and managers.
- OLAP is stores History data for analysis.
- OLTP is a stores current data.
- OLAP is stores descriptive data.
- OLTP is stores typically coded data.
- OLAP is a fully denormalized.
- OLTP is a fully normalized.
- OLAP is use for Application in Management Information System (MIS),Decision support System(DSS).
- OLTP is use for Application in ERP, CRM.
- OLAP is access simplified data model.
- OLTP is access Complex Data model.
- OLAP is focus on multiple record accesses to use.
- OLTP is focus on single record access to use.
- OLAP Transaction recovery is not necessary.
- OLTP Transaction recovery is necessary.
- OLAP is Long database transactions.
- OLTP is Short database transactions.

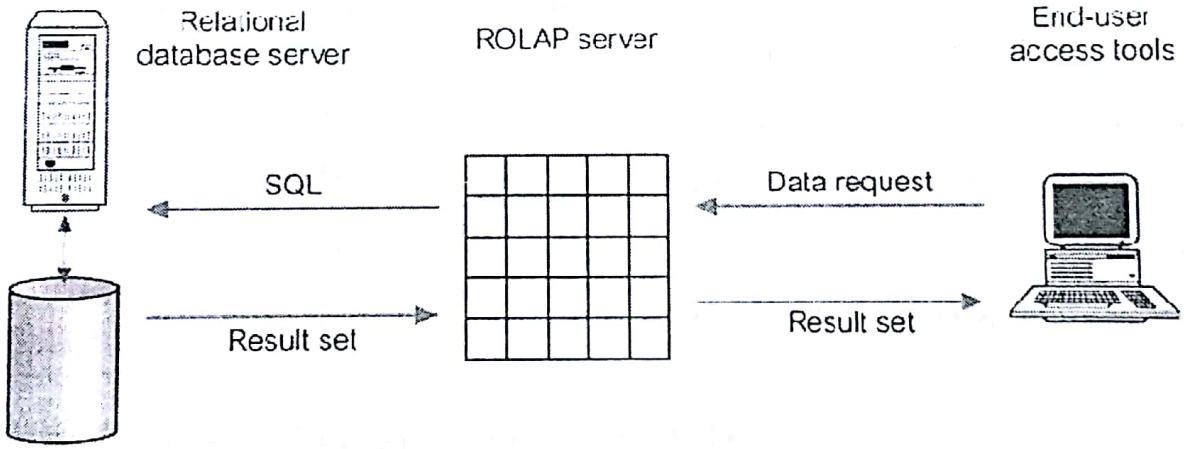
❖ Difference between ROLAP V/S MOLAP:-

- ROLAP means Relational Online Analytical Processing.
- MOLAP means Multidimensional Online Analytical Processing.
- ROLAP is a simple Design.

- MOLAP is a complex Design.
- ROLAP in Database size is medium to large.
- MOLAP in Database size is small to medium.
- ROLAP is Flexibility is very high.
- MOLAP is Flexibility is very Low.
- ROLAP is support client/server Architecture.
- MOLAP is support client/server Architecture.
- ROLAP is Access to support Ad-hoc request.
- MOLAP is Access Limited to pre-defined dimensions.
- ROLAP Resources is high.
- MOLAP Resources is very high.

❖ Types of OLAP Server:-

- Relational OLAP(ROLAP)



- The Relational OLAP servers are placed between relational back-end server and client front-end tools.
- Data access language and Query performance are optimized for multidimensional data

Data Warehousing & Mining

- Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces.
- It is support for very large database.
- Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services.

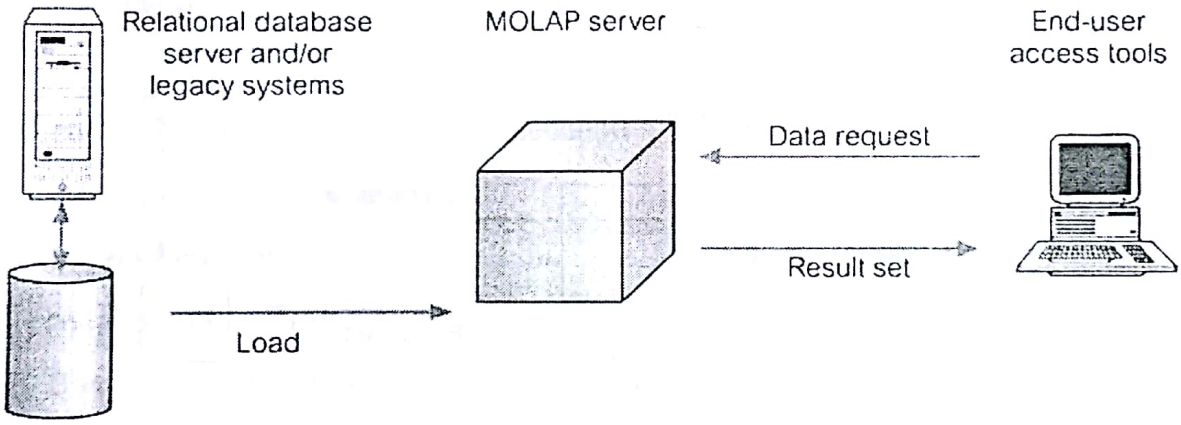
❖ Advantages:-

- The ROLAP servers are highly scalable.
- They can be easily used with the existing RDBMS.

❖ Disadvantages:

- Poor query performance.

❖ Multidimensional OLAP (MOLAP)



- Multidimensional OLAP (MOLAP) uses the array-based multidimensional storage engines for multidimensional views of data.
- The MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
- MOLAP Server uses the two level of data storage representation to handle dense and sparse data sets.

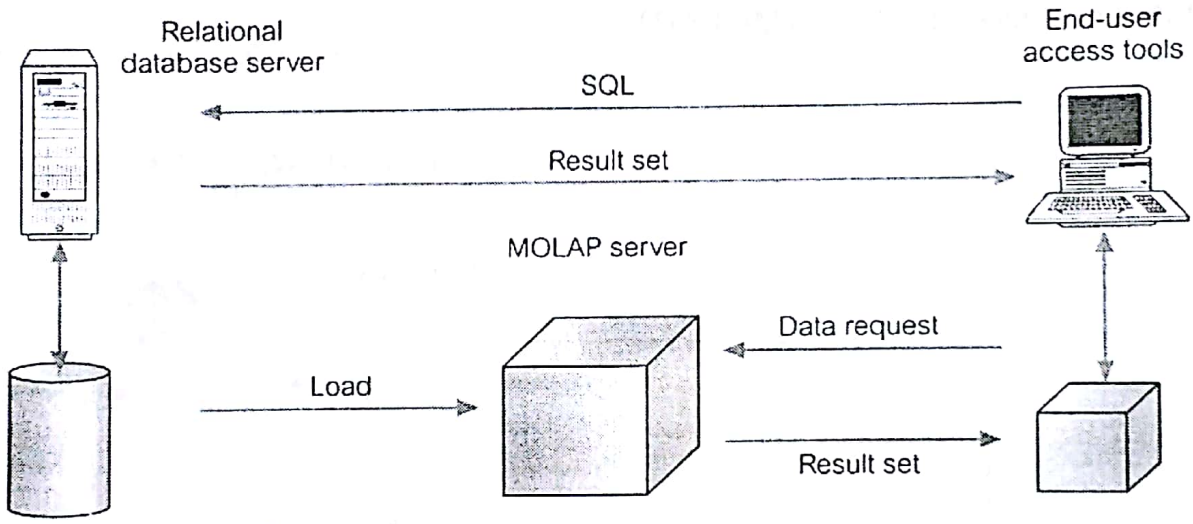
❖ Advantages:-

- Easy to use therefore MOLAP is best suitable for inexperienced user.
- Fast indexing to pre-computed summarized data.
- Data stored in arrays (n-dimensional array).
- Direct access to array data structure.
- Excellent indexing properties.

❖ Disadvantages:-

- MOLAP are not capable of containing detailed data.
- DBMS facility is weak.

❖ Hybrid OLAP (HOLAP)



- The hybrid OLAP technique combination of ROLAP and MOLAP both.
- It has both the higher scalability of ROLAP and faster computation of MOLAP.
- HOLAP server allows storing the large data volumes of detail data. the aggregations are stored separated in MOLAP store.
- The Microsoft SQL Server 7.0 Service supports a hybrid olap server.

❖ Advantages:-

- Good scalability.

- Quick Data Processing.

❖ Specialized SQL servers

- specialized SQL servers provides advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

❖ **Metadata:-** The term "meta" comes from a Greek word that denotes something of a higher or more fundamental nature. Metadata, then, is data about other data.

"The term refers to any data used to aid the identification, description and location of networked electronic resources".

❖ Types of Metadata:-

1. Structural Metadata - used to describe the structure of computer systems such as tables, columns and indexes.
2. Guide Metadata - used to help humans find specific items and is usually expressed as a set of keywords in a natural language.
3. Back room metadata - are used for Extract, transform, load functions to get OLTP data into a data warehouse.
4. Front room metadata - used to label screens and create reports
5. OLAP metadata: The descriptions and structures of Dimensions, Cubes, Measures (Metrics), Hierarchies, Levels, Drill Paths
6. Reporting metadata: The descriptions and structures of Reports, Charts, Queries, Datasets, Filters, Variables, Expressions

7. **Data Mining metadata:** The descriptions and structures of Datasets, Algorithms, Queries Business Intelligence metadata can be used to understand how corporate financial reports reported to Wall Street are calculated, how the revenue, expense and profit are aggregated from individual sales transactions stored in the data warehouse. A good understanding of Business Intelligence metadata is required to solve complex problems such as compliance with corporate governance standards.

8. **Descriptive metadata:** aids the discovery and identification of the resource. Elements may include: title, creator, abstract, subject keywords, and identifiers.

9. **Administrative metadata:** provides information to help manage a resource. It can include: provenance data recording how and when an item was created; technical data such as the format of the resource; rights information associated with the item; other data to aid the preservation of an item.

❖ Source of Metadata:-

- Source of Metadata directly represents metadata for an enterprise information system and captures exactly where and how the data is maintained.
- Source Metadata sounds similar to technical metadata, but Source Metadata can contain both technical and business metadata.
- The Source of Metadata into the terminology and domain of different application.
- When you model Source Metadata, you are modeling the data that your enterprise information systems contain.
- When you model the Source Metadata within your enterprise information systems, you capture some detailed information, including:
    - Identification of data type
    - Storage formats
    - Constraints

- Source-specific locations and names

- The Source Metadata captures this detailed technical metadata to provide a map of the data, the location of the data, and how you access it.
- This collection of Source Metadata comprises a direct mapping of the information sources within your enterprise.