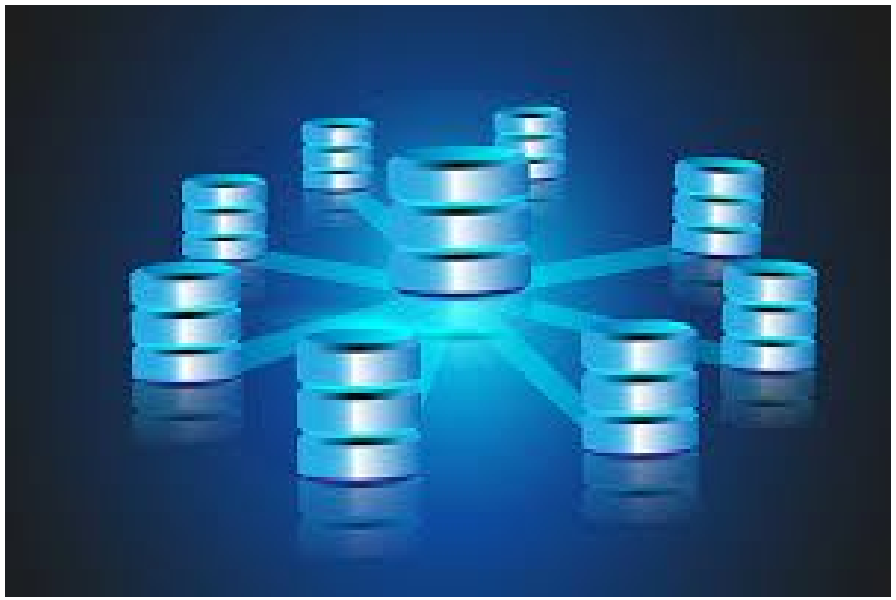
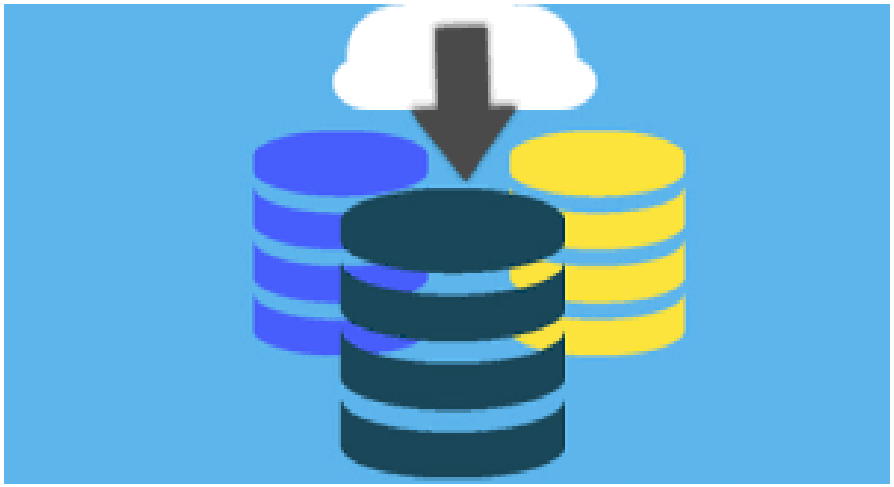


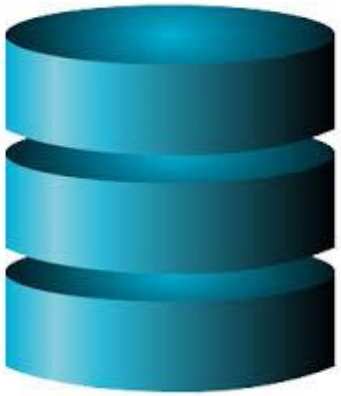


INTRODUCTION TO DATA WAREHOUSING

Type your text

V. Vaishnavee
AP–AI-DS
SNSCE

PREREQUISITE



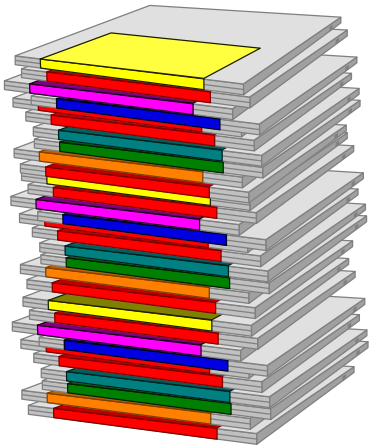
Data, Data everywhere yet ...

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences

- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

What is Data Warehouse?

- Collection of data from various sources, organized to provide useful guidance to an organizational decision makers



Warehouse:

  - A place to store large amount of products

- Data Warehouse is an important preprocessing step for Data Mining

Definition of Data Warehouse / Characteristics

- A data warehouse is a
  - subject-oriented: A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modeling and analysis of data for decision making.
  - Integrated: A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data
  - time-varying: The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view
  - non-volatile: Non-volatile means the previous data is not erased when new data is added to it.

collection of data that is used primarily in organizational decision making.

-- Bill Inmon, Building the Data Warehouse 1996

What is the need of DW in business organization?

- Decision need to be made quickly and correctly using all available data.

How Data warehouse works?

- It works as a central repository where information arrives from one or more data sources.
- Data may be:
  - Structured
  - Semi-structured
  - Unstructured data

Structured Data

- The data that has a structure and is well organized either in the form of tables or in some other way and can be easily operated is known as structured data.
- Searching and accessing information from such type of data is very easy.
- For example, data stored in the relational database in the form of tables having multiple rows and columns.
- The spreadsheet is an another good example of structured data.

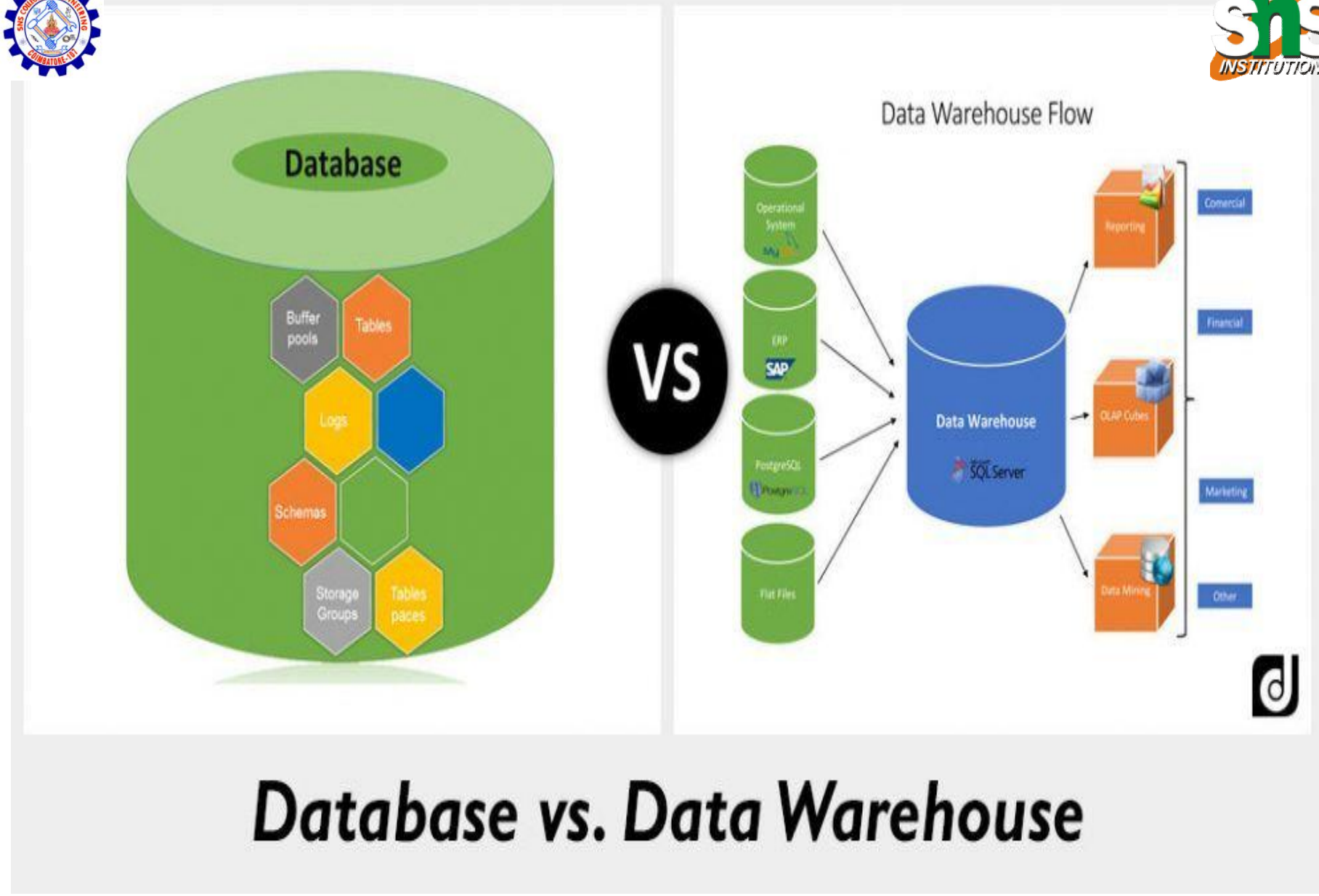
Unstructured Data & Semi-structured Data

Unstructured Data

- The data that is unstructured or unorganized operating such type of data becomes difficult and requires advance tools and softwares to access information.
- For Example, images and graphics, pdf files, word document, audio, video, emails, powerpoint presentations, webpages and web contents, wikis, streaming data, location coordinates etc.

Semi-structured Data

- Semi-structured data is basically a structured data that is unorganised. Web data such JSON(JavaScript Object Notation) files,.csv files, XML and other markup languages are the examples of Semi-structured data found on the web.
- Due to unorganized information, the semi-structured is difficult to retrieve, analyze and store as compared to structured data.
- It requires software framework like Apache Hadoop to perform all this.

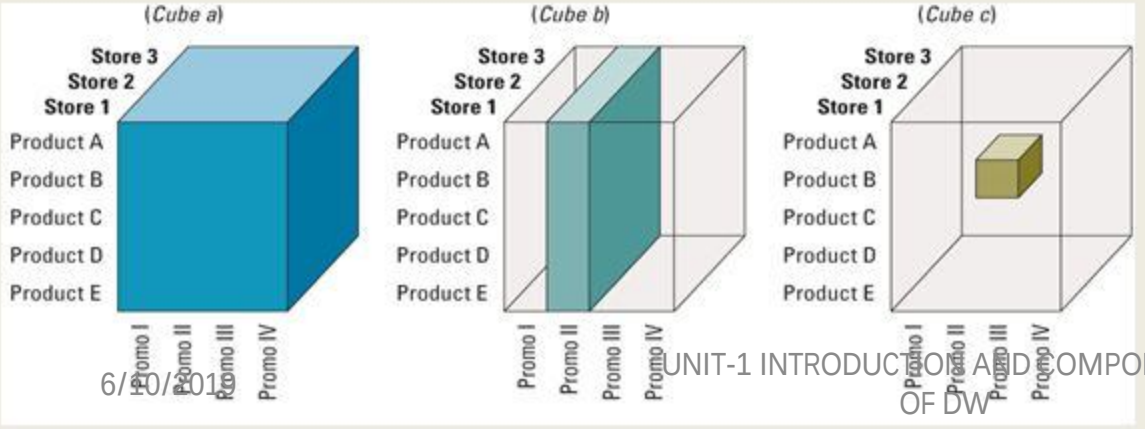


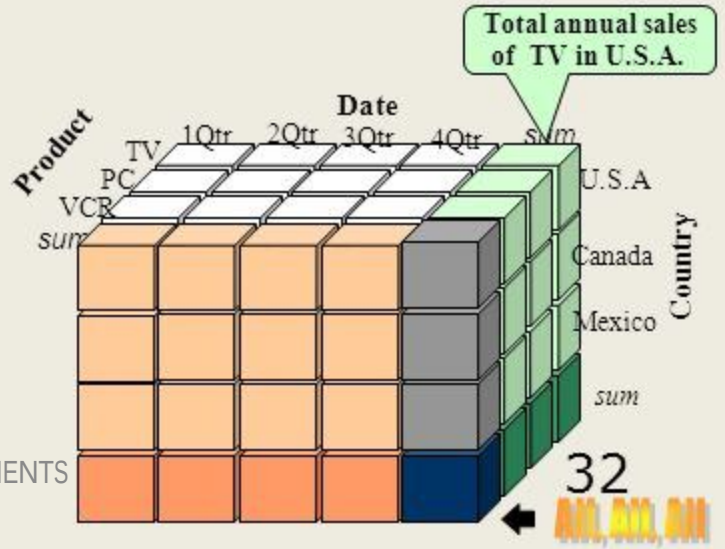
Database vs. Data Warehouse

Database	Data Warehouse
1. Used for Online Transactional Processing (OLTP) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.	1. Used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.
2. The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.	2. The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
3. Entity – Relational modeling techniques are used for RDMS database design.	3. Data – Modeling techniques are used for the Data Warehouse design
4. Optimized for write operation.	4. Optimized for read operations.
5. Performance is low for analysis queries.	5. High performance for analytical queries.
	6. Is *usually* a Database.

Database vs. Data Warehouse

- Databases contain information in a series of two-dimensional tables
- In a Data Warehouse and data mart, information is multidimensional, it contains layers of columns and rows



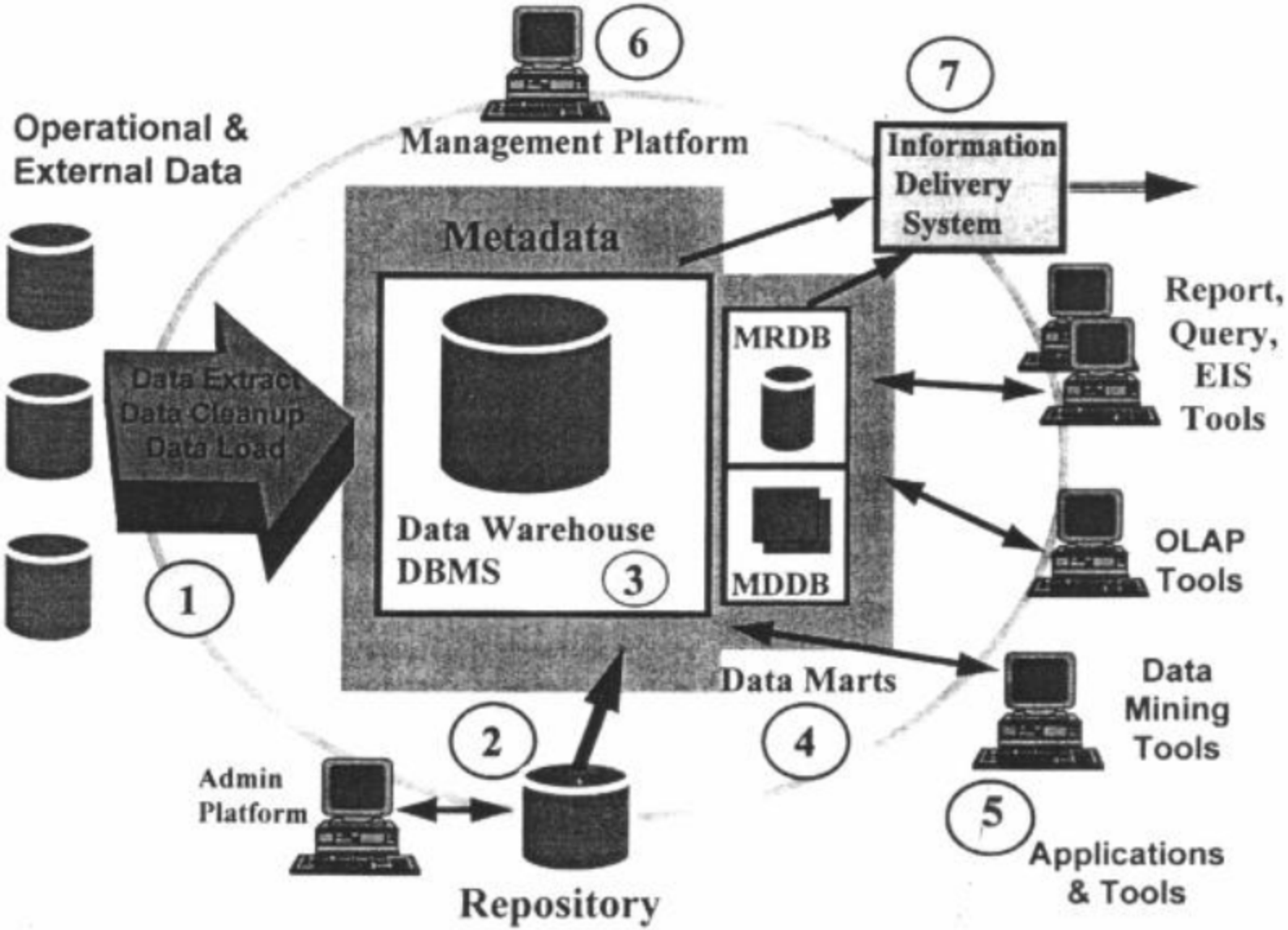


DATA WAREHOUSE ARCHITECTURE *OR* COMPONENTS OF DW

COMPONENTS OF DW

1. Data sourcing, cleanup, transformation, and migration tools
2. Metadata repository
3. Warehouse/database technology
4. Data marts
5. Data query, reporting, analysis, and mining tools
6. Data warehouse administration and management
7. Information delivery system

- Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data. The central repository information is surrounded by number of key components designed to make the environment is functional, manageable and accessible.


6
Management Platform
7
Information Delivery System
Operational & External Data
Metadata
Data Extract
Data Cleanup
Data Load
MRDB
MDDB
Report, Query, EIS Tools
Data Warehouse DBMS
3
1
OLAP Tools
Data Marts
4
Data Mining Tools
2
Admin Platform
5
Applications & Tools
Repository

1. Data warehouse database

- The data source for data warehouse is coming from operational applications.
- The data entered into the data warehouse transformed into an integrated structure and format.
- The transformation process involves conversion, summarization, filteration.
- The data warehouse must be capable of holding and managing large volumes of data as well as different structure of data structures over the time.
- **Data warehouse database** is the central part of the data warehousing environment. This is the item number 2 in the above arch. diagram.
- **Data warehouse database** is implemented based on RDBMS technology.

2. Sourcing, Acquisition, Clean up, and Transformation Tools

- This is item number 1 in the above arch diagram.
- Also called as Extract, Transform and Load (ETL) Tools
- They perform conversions, summarization, key changes, structural changes and condensation.
- The data transformation is required so that the information can by used by decision support tools.
- The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts, and SQL DDL code etc., to move the data into data warehouse from multiple operational systems.

2. Sourcing, Acquisition, Clean up, and Transformation Tools

- **The functionalities of these tools are listed below:**
  - To remove unwanted data from operational db
  - Converting to common data names and attributes
  - Calculating summaries and derived data
  - Establishing defaults for missing data
- *Issues to be considered while data sourcing, cleanup, extract and transformation:*
- Database heterogeneity. DBMSs are very different in data models, data access language, data navigation, operations, concurrency, integrity, recovery etc.
- Data heterogeneity. This is the difference in the way data is defined and used in different models – homonyms, synonyms, unit compatibility (U.S. vs metric), different attributes for the same entity and different ways of modeling the same fact.

3. Meta data

- It is data about data. It is used for maintaining, managing and using the data warehouse.
- It is classified into two:

***1.Technical Meta data*:** It contains information about data warehouse data used by warehouse designer, administrator to carry out development and management tasks. It includes,

- Information about data stores
- Transformation descriptions. That is mapping methods from operational db to warehouse db
- Warehouse Object and data structure definitions for target data
- The rules used to perform clean up, and data enhancement
- Data mapping operations
- Access authorization, backup history, archive history, info delivery history, data acquisition history, data access etc.

3. Meta data

2.Business Meta data: It contains information that gives information stored in data warehouse to users. It includes,

- Subject areas, and information object type including queries, reports, images, video, audio clips etc.
- Internet home pages
- Information related to info delivery system
- Data warehouse operational info such as ownerships, audit trails etc.,

- Meta data helps the users to understand content and find the data. Meta data are stored in a separate data stores which is known as informational directory or Meta data repository which helps to integrate, maintain and view the contents of the data warehouse.

3. Meta data

- **The following lists the characteristics of info directory/ Meta data:**
  - It is the gateway to the data warehouse environment
  - It supports easy distribution and replication of content for high performance and availability
  - It should act as a launch platform for end user to access data and analysis tools
  - It should support the sharing of info
  - It should support scheduling options for request
  - It should support and provide interface to other applications
  - It should support end user monitoring of the status of the data warehouse environment

4. Access tools

- Its purpose is to provide information to business users for decision making.
- User interact with DW using front end tools. There are five categories:
  - Data query and reporting tools
  - Application development tools
  - Executive information system tools (EIS)
  - OLAP tools
  - Data mining tools

4. Access tools

4. 1. Query and reporting tools are used to generate query and report.

Two category:

- ❖ reporting tools
- ❖ managed query tools

Reporting tools types:

- ✓Production reporting tool used to generate regular operational reports like calculating and printing paychecks
- ✓Desktop report writer are inexpensive desktop tools designed for end users.

Managed Query tools:

- ✓ used to generate SQL query.
- ✓ It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement.
- ✓ This tool is a preferred choice of users to perform segment identification, demographic analysis, territory management and preparation of customer mailing lists etc.

4. Access tools

4.2. Application development tools:

- This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all db systems.
- Application development platforms integrate well with popular OLAP tools, and can access all major DB systems including Oracle, Sybase, Informix.
- Examples of application development environments include Visual Basic from Microsoft, PowerBuilder from PowerSoft.

4. Access tools

4.3. OLAP Tools:

- are used to analyze the data in multi dimensional and complex views.
- Business applications for these tools : product performance, profitability, effectiveness of sales program or marketing campaign
- To enable multidimensional properties it uses MDDB (Multi Dimensional Data Base) and MRDB (Multi Relational Data Base)

4.4. Data mining tools:

- are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.
- DM used to perform segmentation (group customer records for custom-tailored marketing), classification (assignment of input data to a predefined class, discovery), association (discovery of cross-sales opportunities), preferencing (determining preference of customer's majority)

5. Data marts

- Inexpensive alternate to DW, requires less time and money to build. -> Independent Data Mart
- Data Mart – meaning – different things to different people.

- Data store that is subsidiary to a DW of integrated data.
- set of denormalized, summarized, aggregated data

→ Dependent data marts, because their content is sourced from DW

- It is directed at a partition of data, that is created for the use of dedicated group of users. Or Departmental subsets that focus on selected subjects.
- Data mart is used in the following situation:
  - Extremely urgent user requirement
  - The absence of a budget for a full scale data warehouse strategy
  - The decentralization of business needs
  - The attraction of easy to use tools and mind sized project
- Independent Data mart presents two problems:

  1. Scalability: A small data mart can grow quickly in multi dimensions. So that while designing it, the organization has to pay more attention on system scalability, consistency and manageability issues
  2. Data integration

6. Data warehouse administration and management

- The management of data warehouse includes,
  - Security and priority management
  - Monitoring updates from multiple sources
  - Data quality checks
  - Managing and updating meta data
  - Auditing and reporting data warehouse usage and status
  - Purging data
  - Replicating, sub setting and distributing data
  - Backup and recovery
  - Data warehouse storage management which includes capacity planning, hierarchical storage management and purging of aged data etc.,

7.Information delivery system

- It is used to enable the process of subscribing for data warehouse information.
- Delivery to one or more destinations according to specified scheduling algorithm.
- In other words, It distributes warehouse-stored data and other information objects to other data warehouses and end-user products like spreadsheet and local databases


THE END