UNIT 2

Data Extraction, Clean up and Transformation Tools

The very first step is-

- Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
- **Data pre-processing** is an important step in the data mining process.

- **Data Extraction -** Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning -** Involves finding and correcting the errors in data.
- **Data Transformation -** Involves converting the data from legacy format to warehouse format.

Extract and Load Process

- Data extraction takes data from the source systems.
- Data load takes the extracted data and loads it into the data warehouse.

Note: Before loading the data into the data warehouse, the information extracted from the external sources must be reconstructed.

A) Controlling the Process

Controlling the process involves determining when to start data extraction and the consistency check on data. Controlling process ensures that the tools, the logic modules, and the programs are executed in correct sequence and at correct time.

B) WHEN TO INITIATE EXTRACT

Data needs to be in a consistent state when it is extracted, i.e., the data warehouse should represent a single, consistent version of the information to the user.

C) Loading the Data

- After extracting the data, it is loaded into a temporary data store where it is cleaned up and made consistent.

Note: Consistency checks are executed only when all the data sources have been loaded into the temporary data store.

Clean and Transform Process

- Once the data is extracted and loaded into the temporary data store, it is time to perform Cleaning and Transforming.
- Steps involved in Cleaning and Transforming:

A) Clean and transform the loaded data into a structure

B) Partition the data

C) Aggregation

A) Clean and Transform the Loaded Data into a Structure: Cleaning and transforming the loaded data helps speed up the queries. It can be done by making the data consistent:

- ✓ within itself
- ✓ with other data within the same data source
- ✓ with the data in other source systems
- ✓ with the existing data present in the warehouse
- ✓ **Transforming** involves converting the source data into a structure.
- ✓ Structuring the data increases the query performance and decreases the operational cost.
- ✓ The data contained in a data warehouse must be transformed to support performance requirements and control the ongoing operational costs.

B) Partition the Data: It will optimize the hardware performance and simplify the management of data warehouse. Here we partition each fact table into multiple separate partitions.

C) Aggregation: Aggregation is required to speed up common queries.

Aggregation relies on the fact that most common queries will analyze a subset or an aggregation of the detailed data.

EXTRACTION, CLEANING AND TRANSFORMATION - Tasks of capturing data from source systems, cleansing and transforming it, and loading the results into a target system can be carried out either by separate products, or by a single integrated solution.

Integrated solutions can fall into one of the categories below:

- Code Generators

- Database Data Replication Tools
- Dynamic Transformation Engines

Data Warehouse DBMS

A generic 5-tier data warehouse architecture posseses five layers namely- The legacy layer, The extraction layer, The database layer, The middleware layer and the application layer.

The database layers deals with the database management system of data warehouses.

This layer is characterized by- Storage of data; processing of queries and data warehouse DBMS and the storage of metadata.

At data warehouse dbms, there exist four major systems:

1. Relational database management system (RDBMS)
2. Modified RDBMS
3. Multi-dimensional database management system

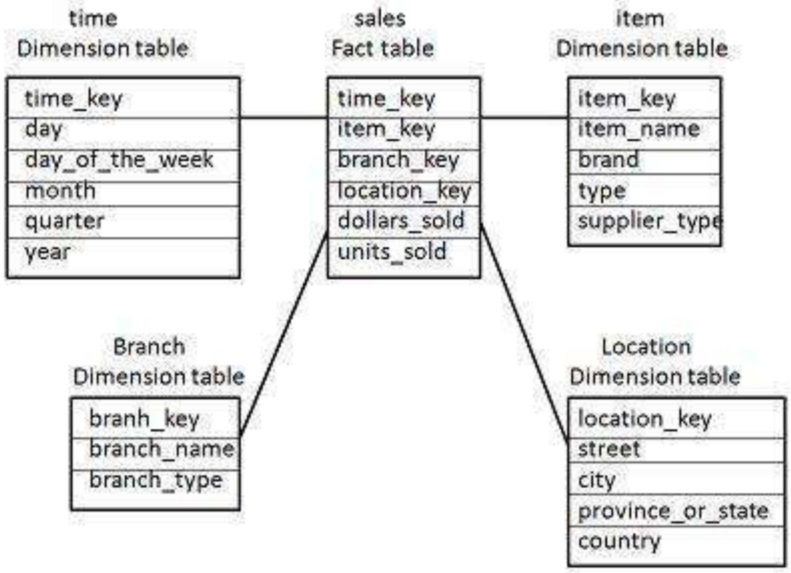
Relational database management system (RDBMS)

- Based on standard, normalized relational table
- Known technology, many supporting applications, portable
- Standard query interface(SQL)
- Supports easy summation and calculations
- Can support very large databases.
- Can be slow when processing complex queries
- Established suppliers

Modified RDBMS

- Uses star-join schema based data structure.
- Expanded SQL, good for business queries.
- Provides a more readily understandable interfaces
- Specially designed for quick access and fast calculations
- Highly indexed
- Often used in data marts
- Good market reputation
- Demands good knowledge of users information need.

Example of star join schema in modified RDBMS



Multi-dimensional database management system

- Uses a meta-cube as standard data structure
- Data stored as an array with any number of possible dimension
- Optimized for OLAP applications
- Often only compatible with proprietary systems
- No branch standards
- Demands new competence
- Can have bad performance with large data volumes

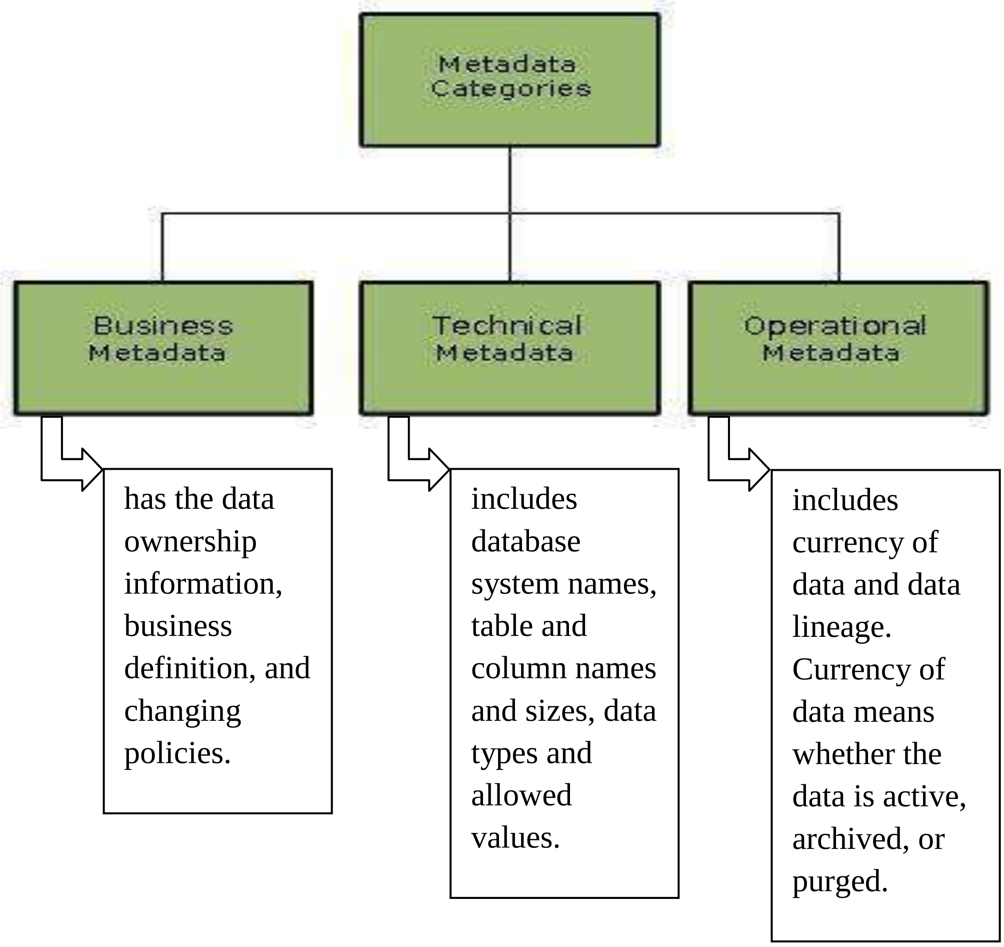
Data Warehouse Meta-Data

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata.

Metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Categories of Metadata



has the data ownership information, business definition, and changing policies.

includes database system names, table and column names and sizes, data types and allowed values.

includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged.

Metadata Repository

Metadata repository is an integral part of a data warehouse system. It has the following metadata:

- **Definition of data warehouse** - It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
- **Business metadata** - It contains has the data ownership information, business definition, and changing policies.
- **Operational Metadata** - It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

- **Data for mapping from operational environment to data warehouse** - It includes the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Data warehouse Administration & Management Tools:

It covers a wide area of field such as:

- Data warehouse and business strategy
- Development life cycle
- Data warehouse Team
- Process Management
- Security Management

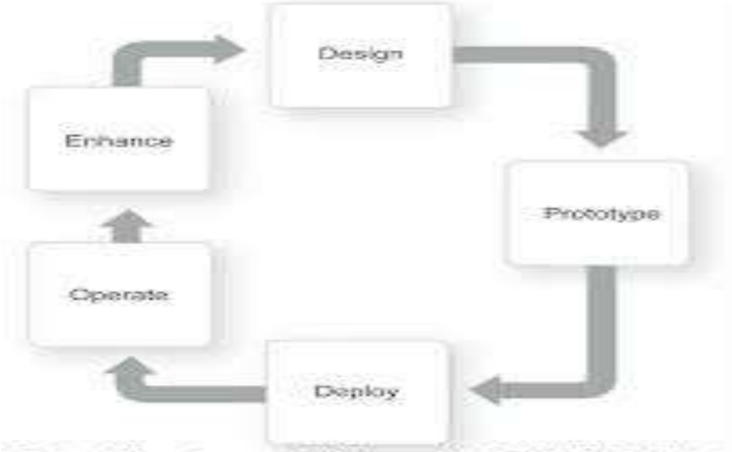
A) Data warehouse & Business Strategy

- "Data is a blood for any organization"
- Business specific objectives:-

1. Identifying environmental factors
2. Formulating Strategic plans
3. Determining specific objectives

B) Development Life cycle

Data warehouse System development phase are similar to the phase of SDLC.



Data Warehouse Lifecycle (DWLC) Model
Figure 1

C) Data warehouse Team

- Building a data warehouse is a large system development process.
- Participants range from DWA(Data warehouse administrator) to business analyst.
- Team should be supposed to lead the organization.
- Team Members: DWA, Managers , project managers , executive sponsors, Business analyst, System architect etc.

D) Process Management

- Process management is not required for every field.
- But it is mainly required for:

❑ Process schedule

❑ Task initiation

❑ Process map definition

E) Security Management

- The goal of database security is the protection of data from accidental or intentional threats to its integrity and access.
- A method to protect data from threats is encryption and decryption.
- Confidential and sensitive data can be store in a seperate tables where only authorized users can access.

Management Tools

For the various types of metadata and day-to-day operations of the data ware house, the administration and management tools must be capable of supporting the following task:

- Monitoring data sources from multiple sources.
- Data quality and integrity check.
- Managing and updating metadata.
- Monitoring database performance to ensure efficient query response time and resource utilization.

Administration Tools

The procedures in this guide refer to and sometimes require the following products, tools, and utilities to achieve your goals with your data warehouse:

- Oracle universal installer
- Oracle enterprise manager

- Oracle warehouse Builder
- Oracle Tuning pack

Challenges & Problems

- Data warehouse management and administration is facing a major problem and challenge for data security and unstructured data.
- For which big data is coming into picture.

Operational system

An **operational system** is a term used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization. These systems are designed in a manner that processing of day-to-day transactions is performed efficiently and the integrity of the transactional data is preserved.

Operational vs. Informational Systems

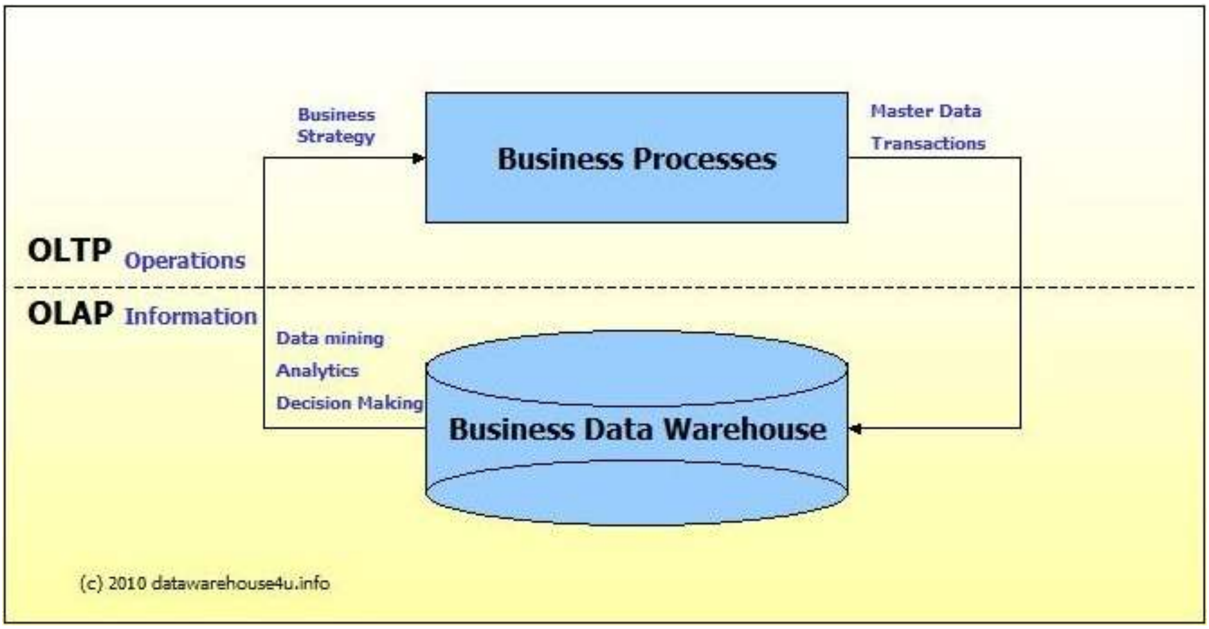
- Perhaps the most important concept that has come out of the Data Warehouse movement is the recognition that there are two fundamentally different types of information systems in all organizations: operational systems and informational systems. "Operational systems" are just what their name implies; they are the systems that help us run the enterprise operation day-to-day. These are the backbone systems of any enterprise, our "order entry', "inventory", "manufacturing", "payroll" and "accounting" systems. Because of their importance to the organization, operational systems were almost always the first parts of the enterprise to be computerized. Over the years, these operational systems have been extended and rewritten, enhanced and maintained to the point that they are completely integrated into the organization. Indeed, most large organizations around the world today couldn't operate without their operational systems and the data that these systems maintain.

- On the other hand, there are other functions that go on within the enterprise that have to do with planning, forecasting and managing the organization. These functions are also critical to the survival of the organization, especially in our current fast-paced world. Functions like "marketing planning", "engineering planning" and "financial analysis" also require information systems to support them. But these functions are different from operational ones, and the types of systems and information required are also different. The knowledge-based functions are informational systems.

- "Informational systems" have to do with analyzing data and making decisions, often major decisions, about how the enterprise will operate, now and in the future. And not only do informational systems have a different focus from operational ones, they often have a different scope. Where operational data needs are normally focused upon a single area, informational data needs often span a number of different areas and need large amounts of related operational data.

- In the last few years, Data Warehousing has grown rapidly from a set of related ideas into an architecture for data delivery for enterprise end-user computing.

OLTP vs. OLAP



<u>OLTP (On-line Transaction Processing)</u> is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE).

The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second.

In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

<u>OLAP (On-line Analytical Processing)</u> is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations.

For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques.

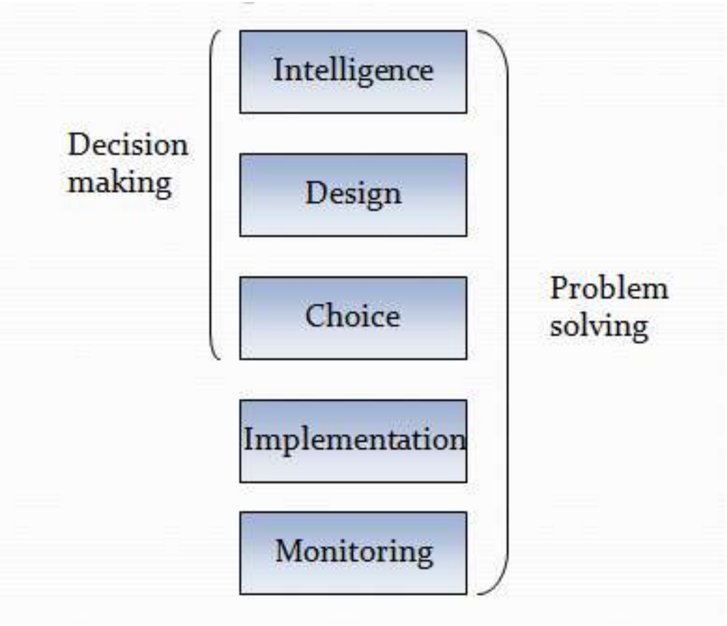
In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

-

DSS (Decision Support System)

- A **DSS** is a computer-based information system that supports business or organizational decision-making activities.
- A **DSS** is a collection of integrated software applications and hardware that form the backbone of an organization's decision making process and help to make decisions, which may be rapidly changing and not easily specified in advance.

Decision Making as a Component of Problem Solving



DSS Characteristics :

- **Facilitation :** DSS facilitate and support specific decision-making activities and/or decision processes.
- **Interaction :** DSS are computer-based systems designed for interactive use by decision makers or staff users who control the sequence of interaction and the operations performed.
- **Ancillary :** DSS can support decision makers at any level in an organization. They are NOT intended to replace decision makers.
- **Repeated Use :** DSS are intended for repeated use. A specific DSS may be used routinely or used as needed for ad hoc decision support tasks.
- **Identifiable :** DSS may be independent systems that collect or replicate data from other information systems OR subsystems of a larger, more integrated information system.

- **Task-oriented :** DSS provide specific capabilities that support one or more tasks related to decision-making, including: intelligence and data analysis; identification and design of alternatives; choice among alternatives; and decision implementation.

- **Decision Impact :** DSS are intended to improve the accuracy, timeliness, quality and overall effectiveness of a specific decision or a set of related decisions.

- **Supports individual and group decision making :** It provides a single platform that allows all users to access the same information and access the same version of truth, while providing autonomy to individual users and development groups to design reporting content locally.

- **Comprehensive Data Access :** It allows users to access data from different sources concurrently, leaving organizations the freedom to choose the data warehouse that best suits their unique requirements and preferences.

DSS Objectives :

1. Increase the effectiveness of the manager's decision-making process.

2. Supports the manager in the decision-making process but does not replace it.

3. Improve the director effectiveness of decision making.

DSS Components :

DSS components may be classified as:

- **Inputs :** Factors, numbers, and characteristics to analyze.

- **User Knowledge and Expertise :** Inputs requiring manual analysis by the user.

- **Outputs :** Transformed data from which DSS "decisions" are generated.

- **Decisions :** Results generated by the DSS based on user criteria.

DSS Advantages:

1. Time savings

2. Enhance effectiveness

3. Improve interpersonal communication

4. Competitive advantage

5. Cost reduction

6. Increase decision maker satisfaction

7. Promote learning
8. Improves personal efficiency

DSS Disadvantages:

1. Monetary cost.
2. Overemphasize decision making.
3. Assumption of relevance.
4. Transfer of power.
5. Unanticipated effects.
6. Obscuring responsibility.
7. False belief in objectivity.
8. Status reduction.
9. Information overload.

DSS Applications:

- Medical diagnosis.
- Business and Management.
- Agricultural production.
- Forest management.

Types of Distributed Data Warehouse

✓ Business is distributed geographically or over multiple, differing product lines.

✓ The data warehouse environment will hold a lot of data, and the volume of data will be distributed over multiple processors.

✓ The data warehouse environment grows up in a uncoordinated manner-first one data warehouse appears, then another.

Concerns for each type of Data Warehouse

✓ The Local Data Warehouses

✓ The Global Data Warehouses

- ✓ Intersection of Global and Local data
- ✓ Redundancy
- ✓ Access of Local and Global Data
- ✓ The Technologically Distributed Data Warehouse
- ✓ The Independently Evolving Distributed Data Warehouse

The Nature of the Development Efforts

- ✓ Four possible meaning to “multiple groups building the data warehouse”
- ✓ Completely un-integrated lines of business each with their own data warehouse
- ✓ The same data warehouse but with distributed parts
- ✓ Different levels of data within the same data warehouse
- ✓ Different non-distributed parts of the detailed level of the data warehouse

Distributed Data Warehouse Development

- ✓ Coordinating Development across Distributed Locations
- ✓ The Corporate Data Model – Distributed
- ✓ Meta data in the Distributed Warehouse