



# Chapter-2

Data warehouse Architecture



# Data Warehouse Concepts

- The basic concept of a Data Warehouse is to facilitate a single version of truth for a company for decision making and forecasting. A Data warehouse is an information system that contains historical and commutative data from single or multiple sources. Data Warehouse Concepts simplify the reporting and analysis process of organizations.
- Characteristics of Data warehouse
- Subject-Oriented
- Integrated
- Time-Variant
- Non-volatile



# Data Warehouse Architecture

- **Data Warehouse Architecture** is complex as it's an information system that contains historical and commutative data from multiple sources.
- **Three common architectures are:**
  - Data Warehouse Architecture: Basic
  - Data Warehouse Architecture: With Staging Area
  - Data Warehouse Architecture: With Staging Area and Data Marts

## Basic architecture

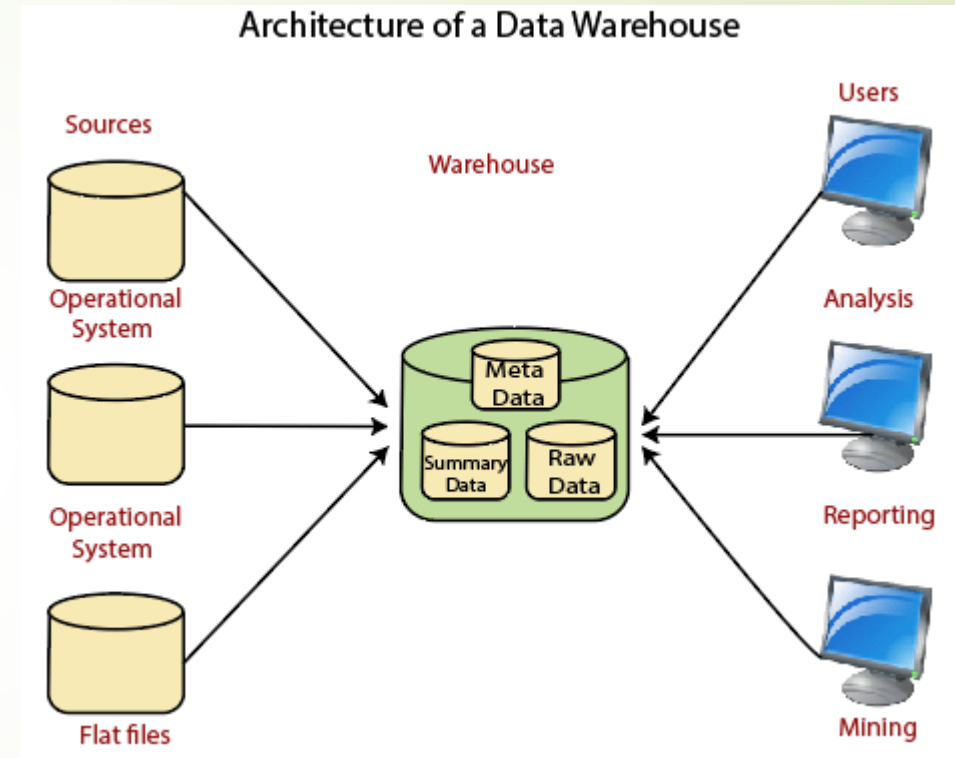
An **operational system** is a method used in data warehousing to refer to a **system** that is used to process the day-to-day transactions of an organization.

### Flat Files

A **Flat file** system is a system of files in which transactional data is stored, and every file in the system must have a different name.

### Meta Data

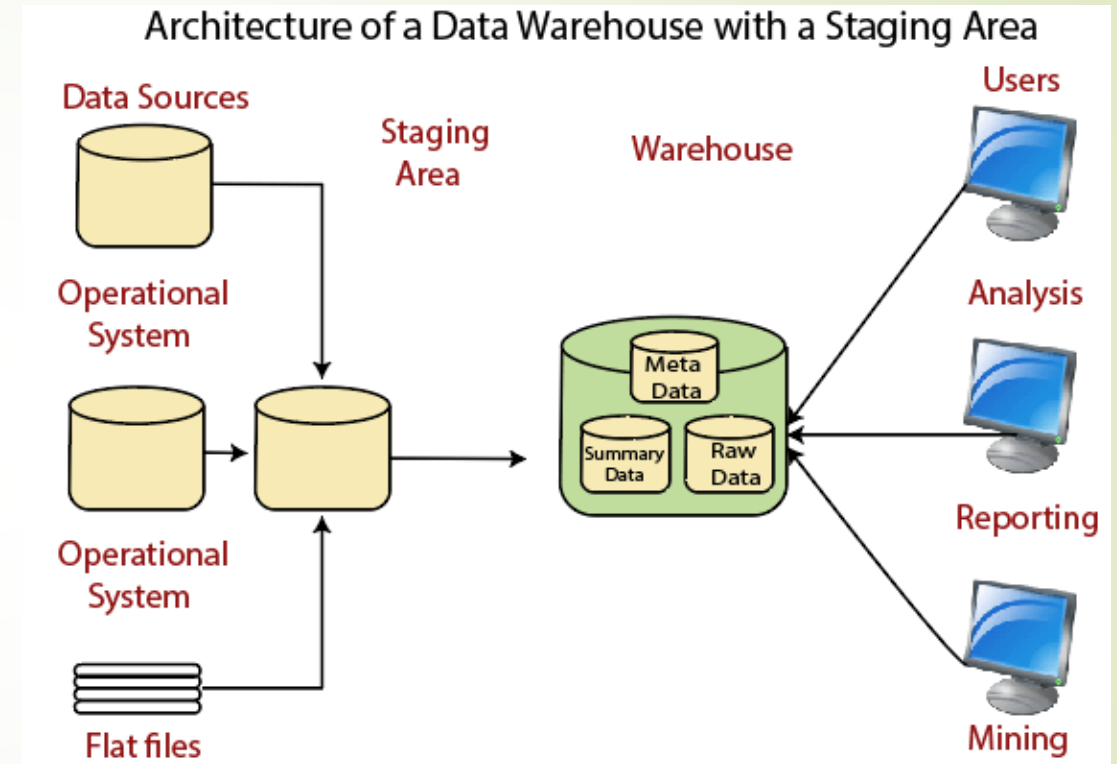
A set of data that defines and gives information about other data



# Data Warehouse Architecture: With Staging Area

We must clean and process your operational information before put it into the warehouse.

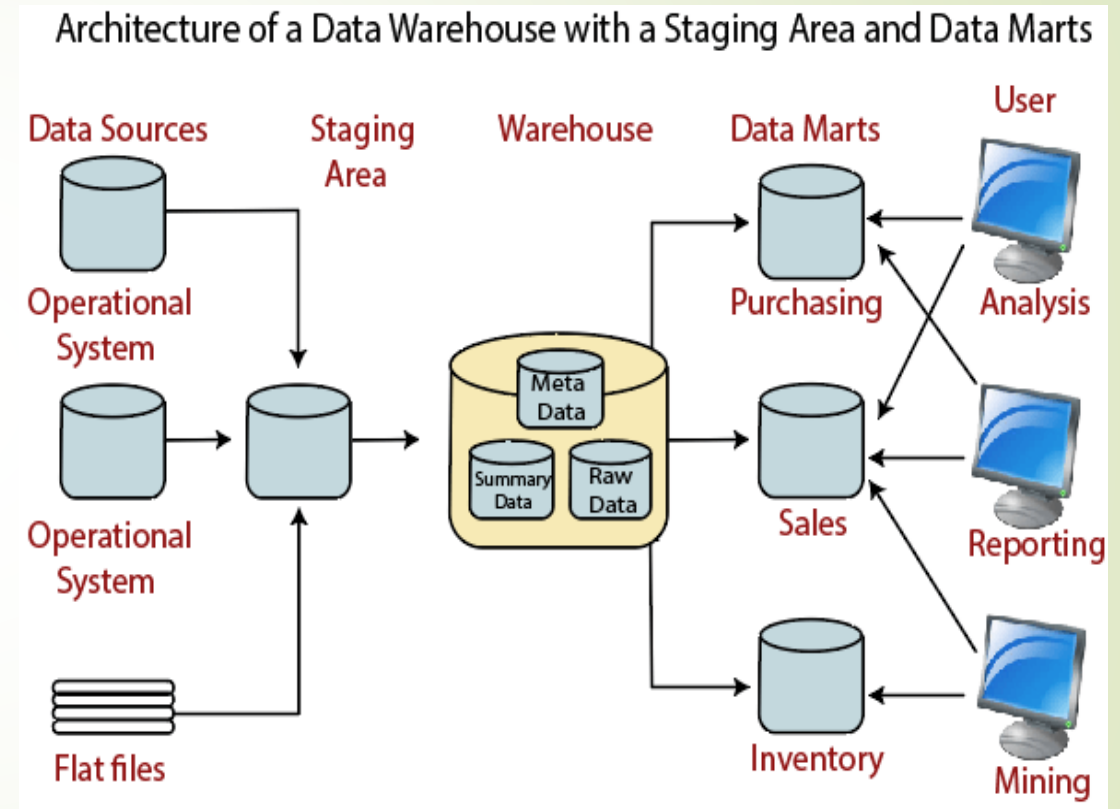
although data warehouses uses a **staging area** (A place where data is processed before entering the warehouse).



# Data Warehouse Architecture: With Staging Area and Data Marts

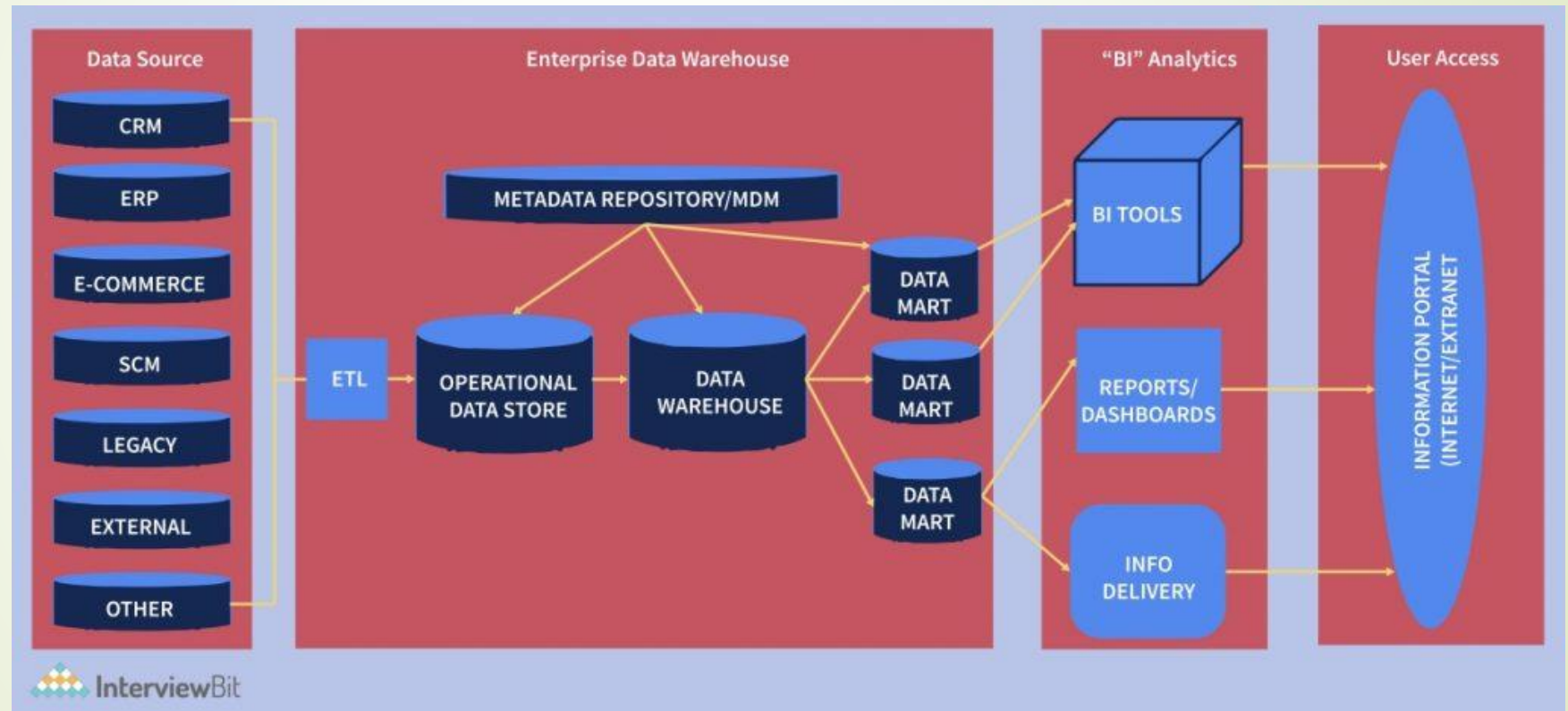
We may want to customize our warehouse's architecture for multiple groups within our organization.

We can do this by adding **data marts**.





# Data Warehouse Components



## ► **Data Warehouse Database**

- The foremost element of a Data Warehouse architecture is a database that stores all company data and makes it manageable to report.
- four database types

**Typical relational databases:** You perhaps use these row-centred databases on an everyday basis, for instance, SAP, Oracle, Microsoft SQL Server, and IBM DB2.

**Analytics databases** are specifically designed for storing data to support and handle analytics, like Greenplum and Teradata.

**Data warehouse applications** aren't exactly storage databases, but lots of dealers now provide applications that offer software for data management along with hardware for data storage. For instance, Oracle Exadata, IBM Netezza, and SAP Hana.

**Cloud-based databases** can be retrieved and hosted on the cloud so that you don't need to obtain any hardware to set up your data warehouse—for instance, Google BigQuery, Microsoft Azure SQL and Amazon Redshift.





## ➤ **Sourcing, Acquisition, Cleanup, and Transformation Tools (ETL)**

- A substantial part of the implementation effort is spent pulling data from operational systems and putting it in a format appropriate for informational applications that operate off the data warehouse.
- Erasing undesirable data from operational databases
- Transforming to common data definitions and names
- Setting defaults for missing data
- Accommodating source data definition modifications
- Anonymize data as per regulatory stipulations.
- Eliminating unwanted data in operational databases from loading into Data warehouse.
- Search and replace common names and definitions for data arriving from different sources.
- Calculating summaries and derived data
- In case of missing data, populate them with defaults.
- De-duplicated repeated data arriving from multiple datasources.

# Bottom Tier (Data Warehouse Server)

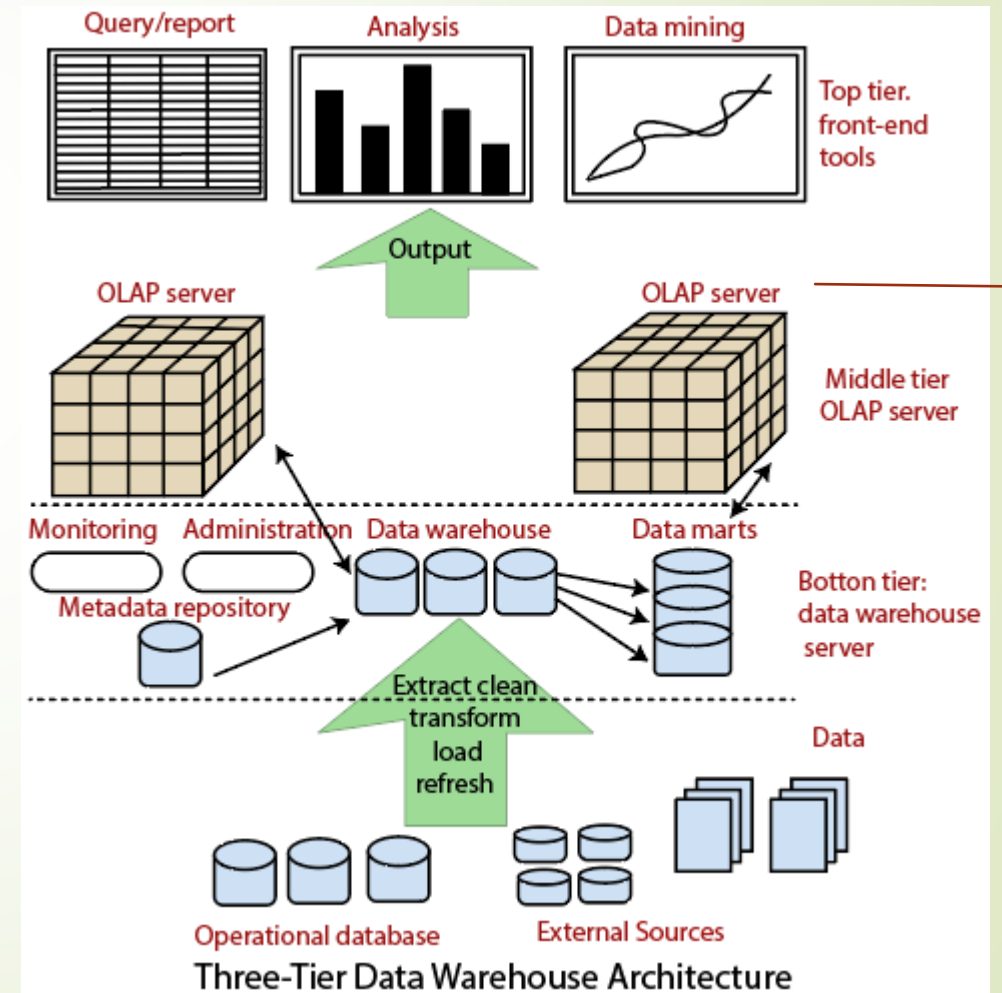
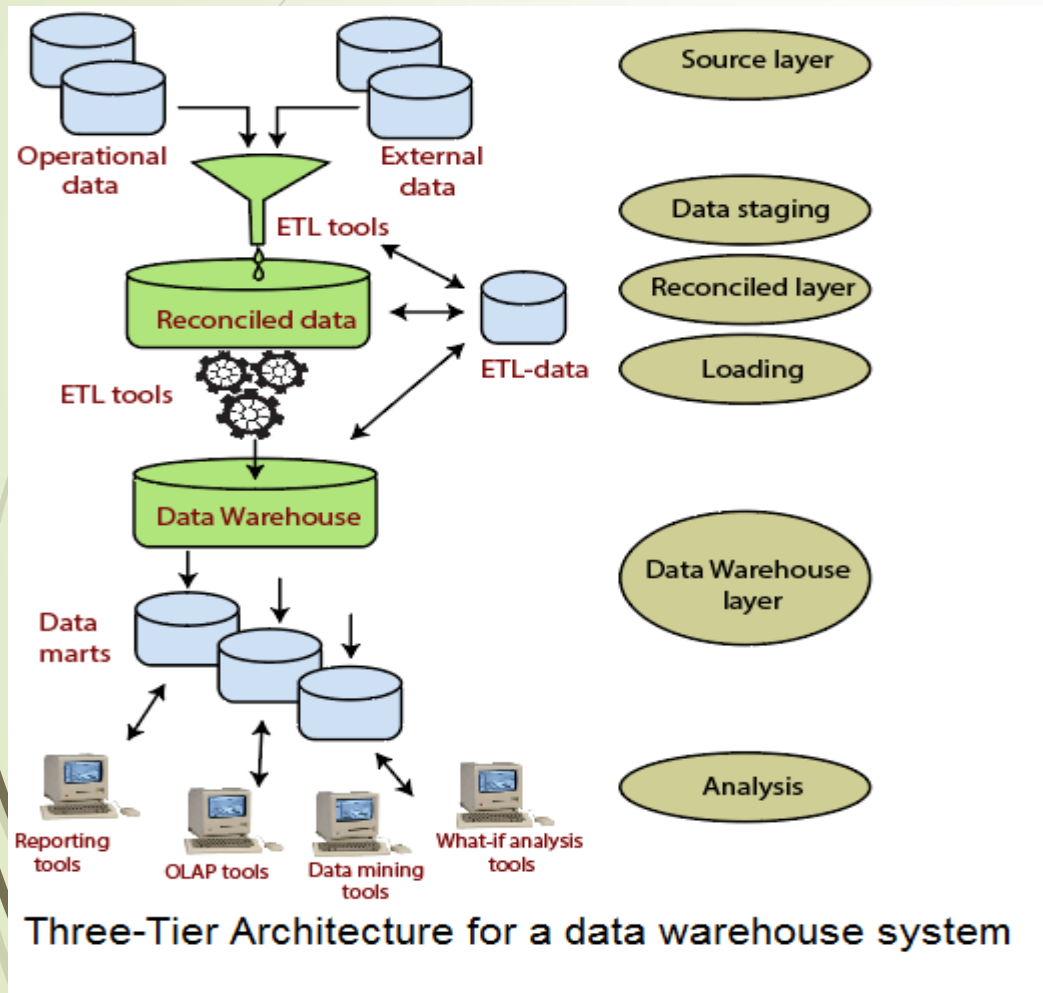
## Middle Tier (OLAP Server)

### Top Tier (Front end Tools).

- A **middle-tier** which consists of an **OLAP server** for fast querying of the data warehouse.
- The OLAP server is implemented using either
- **(1) A Relational OLAP (ROLAP) model**, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations.
- **(2) A Multidimensional OLAP (MOLAP) model**, i.e., a particular purpose server that directly implements multidimensional information and operations.

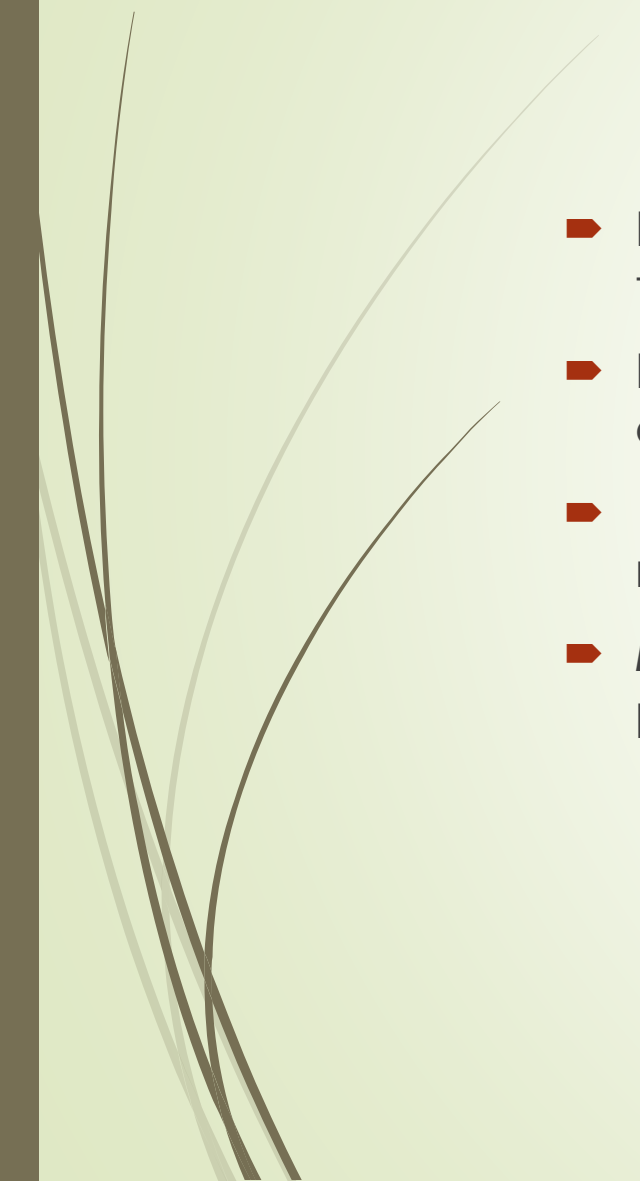


# Three-Tier Data Warehouse Architecture





# What is Facts, Measures and Dimensions

- Fact- a fact is the part of your data that indicates a specific occurrence or transaction,
  - Dimensions are the descriptive data elements that are used to categorize or classify the data.
  - **dimensions** are pieces of data that allow you to understand and index measures in your data models.
  - **Measures** can be qualitative, like a Product ID, or quantitative, like the price of a product.
- 



# Dimensional Modeling

- **Dimensional modeling** is a database design technique that supports business users to query data in a data warehouse system. The dimensional modeling is developed to be oriented to improve the query performance and ease of use.
- In dimensional modeling, there are two important concepts: facts and dimensions.

Facts are business measurements. Facts are normally but not always numeric values that could be aggregated. e.g, the number of products sold per quarter.

Dimensions are called contexts. Dimensions are business descriptors that specify the facts, for example, product name, brand, quarter, etc.





# Dimensional modelling process

- The dimensional data model is built based on star schema with a fact table at the center surrounded by a number of dimension tables. The following four-step process is commonly used in dimensional modeling design:
- Select the business process
- Declare the grain
- Identify the dimensions
- Identify the Fact



# Difference between ER Modelling and Dimensional Modelling:

## ER Modeling

- It is transaction-oriented.(OLTP)
- Entities and Relationships.
- Real-time information.
- It eliminates redundancy.
- High transaction volumes using few records at a time.
- Highly Volatile data.
- Physical and Logical Model.
- Normalization is suggested.

## Dimensional Modeling

- It is subject-oriented.(OLAP)
- Fact Tables and Dimension Tables.
- Historical information.
- It plans for redundancy.
- Low transaction volumes using many records at a time.
- Non-volatile data.
- Physical Model.
- De-Normalization is suggested.



# Data Warehousing - Schemas

- A schema is defined as a logical description of database where fact and dimension tables are joined in a logical manner. Data Warehouse is maintained in the form of Star, Snow flakes, and Fact Constellation schema.
- 

# DE normalized data

## Grocery Shop Record

Transaction_Date	Item_Name	Product_Category	ItemPrice_Per_KG	Order	Customer	Email_ID	Custemer_address
11-09-2023	Apples	Produce	\$1.99	5lbs	John Doe	john@example.com	123 Main St
11-09-2023	Milk	Dairy	\$2.49	2gallons	Jane Smith	jane@example.com	456 Elm,St
12-09-2023	Bread	Bakery	\$2.50	3loaves	John Doe	john@example.com	123 Main St
12-09-2023	Eggs	Dairy	\$1.99	1dozen	Mary Jonson	mary@example.com	789 Oak St
13-09-2023	Cereal	Breakfast	\$3.49	2boxes	Jane Smith	jane@example.com	456 Elm,St

Customer-table

Customer_id	Cutomer_Name	Customer_Email_ID	Custmer Address
123	John Doe	john@example.com	123 Main St
124	Jane Smith	jane@example.com	456 Elm,St
125	Mary Jonson	mary@example.com	789 Oak St

Item-table

Item_Id	Item Name	Item Category	Price Per Kg
1	Apples	Produce	\$1.99
2	Milk	Dairy	\$2.49
3	Bread	Bakery	\$2.50
4	Eggs	Dairy	\$1.99
5	Cereal	Breakfast	\$3.49



## Transaction-table

Transaction_Date	Item_Id	Custoem_id	Order
11-09-2023	1	123	\$1.99
12-09-2023	3	123	\$2.50
11-09-2023	2	124	\$2.49
13-09-2023	5	124	\$3.49
12-09-2023	4	125	\$1.99

## Date-table

Transaction_Date	Year	Month	Date_Type
11-09-2023	2023	September	Weekday
12-09-2023	2023	September	Weekday
11-09-2023	2023	September	Weekday
13-09-2023	2023	September	Weekday
12-09-2023	2023	September	Weekday

Item_Id	Item Name	Item Category	Price Per Kg
1	Apples	Produce	\$1.99
2	Milk	Dairy	\$2.49
3	Bread	Bakery	\$2.50
4	Eggs	Dairy	\$1.99
5	Cereal	Breakfast	\$3.49

Customer_id	Customer_Name	Customer_Email_ID	Customer Address
123	John Doe	john@example.com	123 Main St
124	Jane Smith	jane@example.com	456 Elm,St
125	Mary Jonson	mary@example.com	789 Oak St

Transaction_Date	Item_Id	Customer_id	Order
11-09-2023	1	123	\$1.99
12-09-2023	3	123	\$2.50
11-09-2023	2	124	\$2.49
13-09-2023	5	124	\$3.49
12-09-2023	4	125	\$1.99

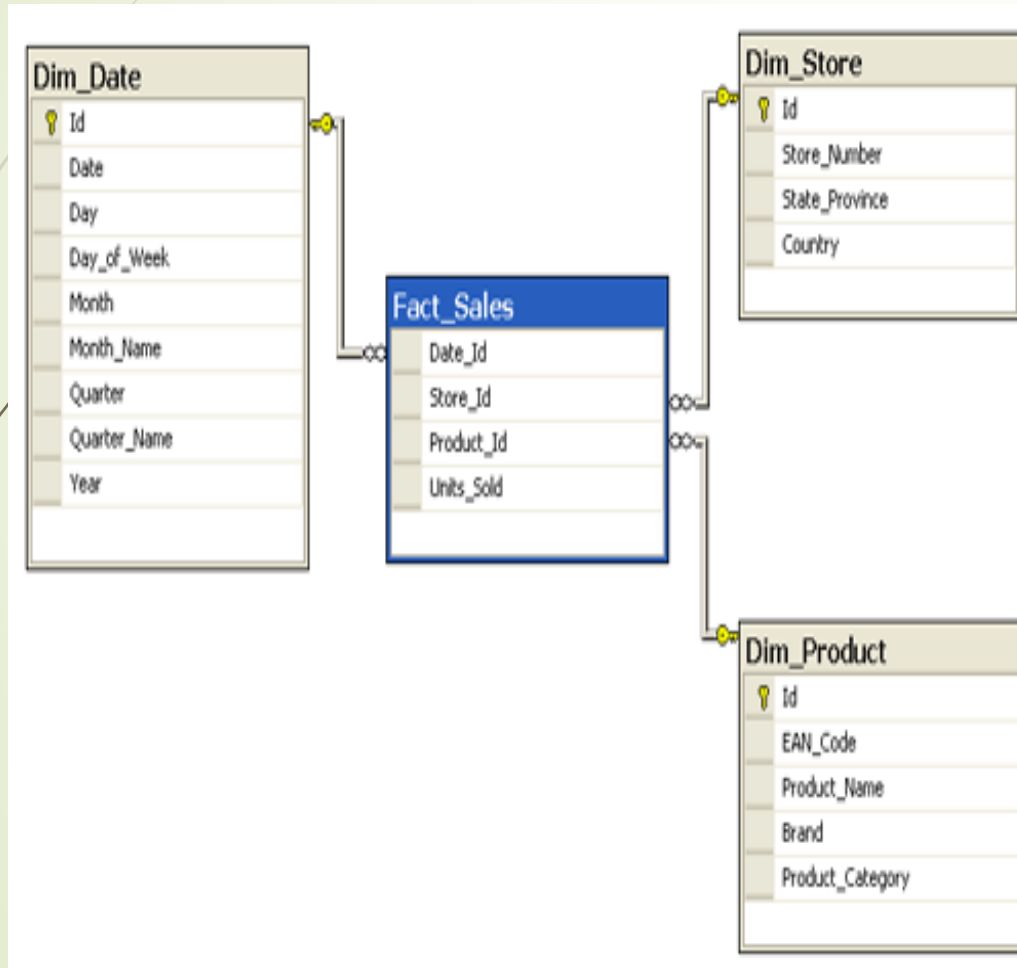
Transaction_Date	Year	Month	Date_Type
11-09-2023	2023	September	Weekday
12-09-2023	2023	September	Weekday
11-09-2023	2023	September	Weekday
13-09-2023	2023	September	Weekday
12-09-2023	2023	September	Weekday

Region Table

Platform

# Introduction to star schema

(Star schema example)



A Star schema contains a fact table and multiple dimension tables.

Each dimension is represented with only one-dimension table.

They are not normalized.

The Dimension table contains a set of attributes.



# Dimension tables & Fact table

A dimension table consists of columns that represent dimensions that provide the context needed for studying the facts. A dimension table typically stores characters that describe facts.

- The fact table is at the core of the star schema—the fact table stores facts or measures of interests. Normally facts are numbers that can be aggregated, summarized, or rolled up.
- The fact table contains surrogate keys as a part of its primary key. Those keys are the foreign key of the dimension tables.



# Factless Fact Table

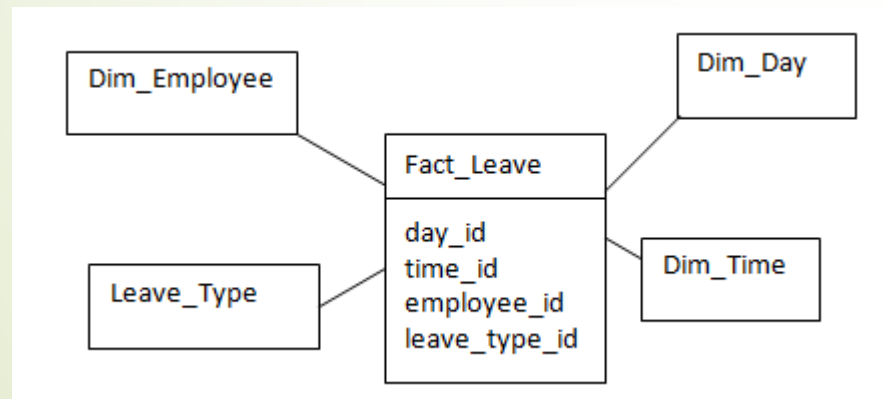
- The factless fact table is a fact table that does not contain any facts. There are two kinds of factless fact tables:
- Factless fact table describes events or activities.
- Factless fact table describes a condition, eligibility, or coverage.
- Both kinds of factless fact tables play a very important role in your dimensional model design.
- **Factless fact table for event or activity**
- When designing a dimensional model, you often find that you want to track events or activities that occur in your business process but you can't find measures to track. In these situations, you can create a transaction-grained fact table that has no facts to describe that events or activities. Even though there are no facts storing in the fact table, the event can be counted to produce very meaningful process measurements.



# Factless fact table for event or activity example

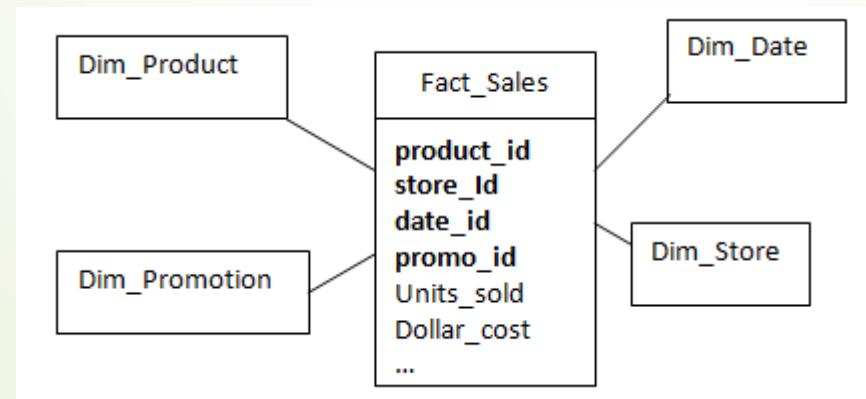
For example, you may want to track employee leaves. How often and why your employee leaves are very important for you to plan your daily activities and resources.

At the center of the diagram below is the FACT\_LEAVE table that has no facts at all. However, the FACT\_LEAVE table is used to measure employee leave the event when it occurs.




# Factless fact table for condition, eligibility, or coverage

- A factless fact table can be also used in these situations:
- Tracking salesperson assigned to each prospect or customer
- Logging the eligibility of employees for a compensation program
- Capturing the promotion campaigns that are active at specific times such as holidays.

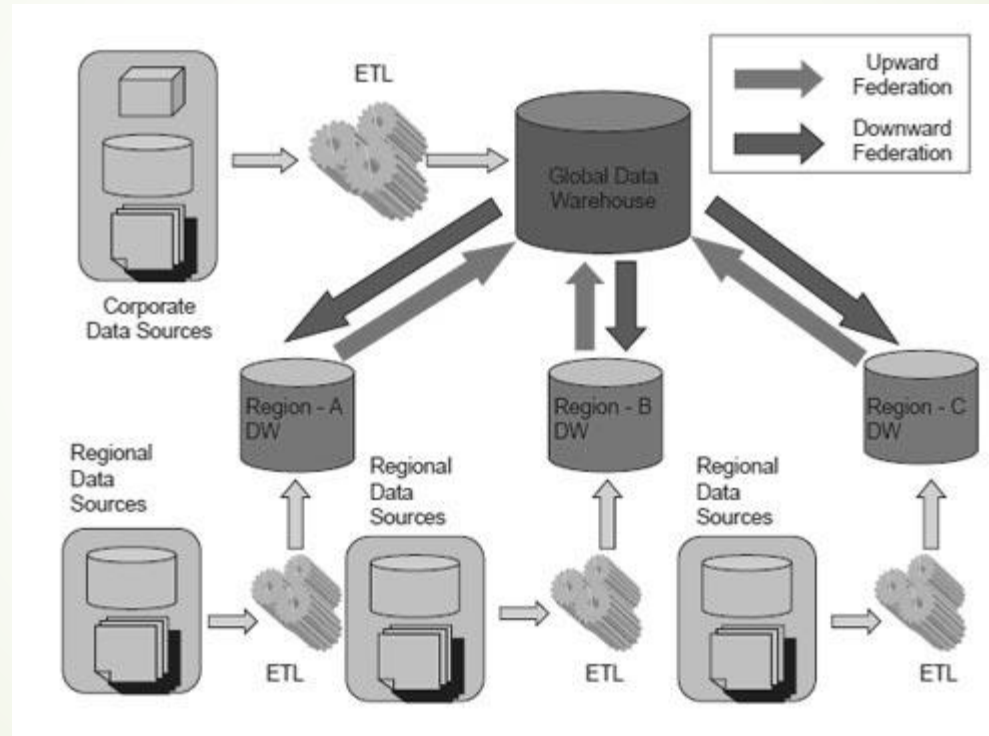




# Introduction to federated data warehouse

- corporate usually has a set of heterogeneous system landscape that contains transaction systems and business intelligence tools which provide analytical capabilities for each individual department needs.
  - Each department views a business model from its own perspective.
  - A federated data warehouse is a practical approach to achieving the “single version of the truth” across the organization. The federated data warehouse is used to integrate key business measures and dimensions. The foundations of the federated data warehouse are the common business model and common staging area.
- 

# The architecture of federated data warehouse









## ***Inside a Dimension Table***

---

- **Dimension table key:** Uniquely identify each row. Use surrogate key (integer).
- **Table is wide:** A table may have many attributes (columns).
- **Textual attributes.** Descriptive attributes in string format. No numerical values for calculation.
- **Attributes not directly related:** E.g., product color and product package size. No transitive dependency.
- **Not normalized** (star schemar).
- **Drilling down and rolling up** along a dimension.
- **One or more hierarchy** within a dimension.
- **Fewer number of records.**



# Inside Dimensional Table

- Dimension tables are used to describe dimensions; they contain dimension keys, values and attributes.
- Dimension tables are typically small, ranging from a few to several thousand rows. Occasionally dimensions can grow fairly large, however. For example, a large credit card company could have a customer dimension with millions of rows. Dividing a data warehouse project into dimensions, provides structured information for reporting purpose.
- When you create a dimension, you logically create a structure for your projects. This dimension table can be utilized across for reports and it's about re-usability. If there are any changes to be made, it is evident that only a particular table will get affected. When a company wants to create a report, they can read the data from the dimension table since the table consists of necessary information.



# Inside Fact Table,

- A transaction fact table captures detailed information about individual business transactions or events. It records every occurrence at the most granular level, providing a comprehensive view of operational data.
- Fact tables and entities aggregate *measures*, or the numerical data of a business. To measure data in a fact table or entity, all of the measures in a fact table or entity must be of the same grain.



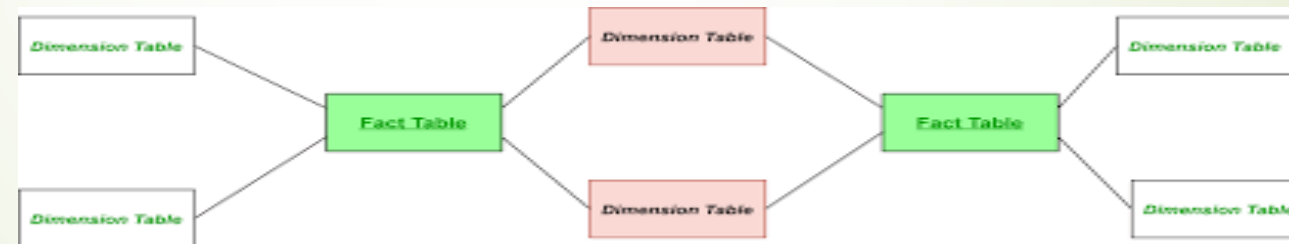
# Granularity



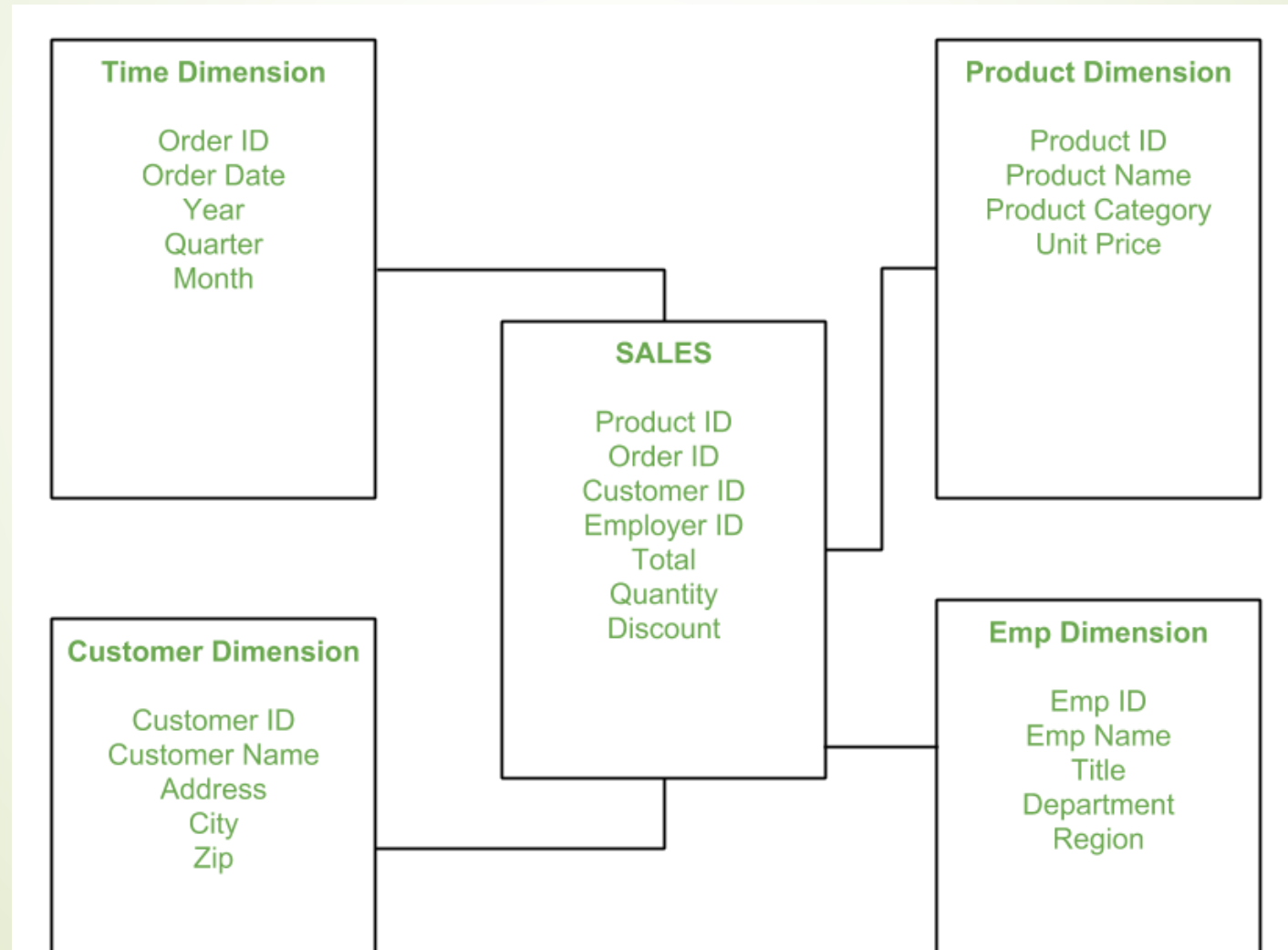
- ▶ **Granular data** is detailed data, or the lowest level that data can be in a target set. It refers to the size that data fields are divided into, in short how detail-oriented a single field is.
- ▶ Data granularity is the lowest level of detail that's available within a data collection. Information that's present in one single line or field within a database or data warehouse has coarse granularity, as it doesn't have any subdivisions

# Fact Constellation

- **Fact Constellation** is a schema for representing multidimensional model. It is a collection of multiple fact tables having some common dimension tables. It can be viewed as a collection of several star schemas and hence, also known as *Galaxy schema*. It is one of the widely used schema for Data warehouse designing and it is much more complex than star and snowflake schema.

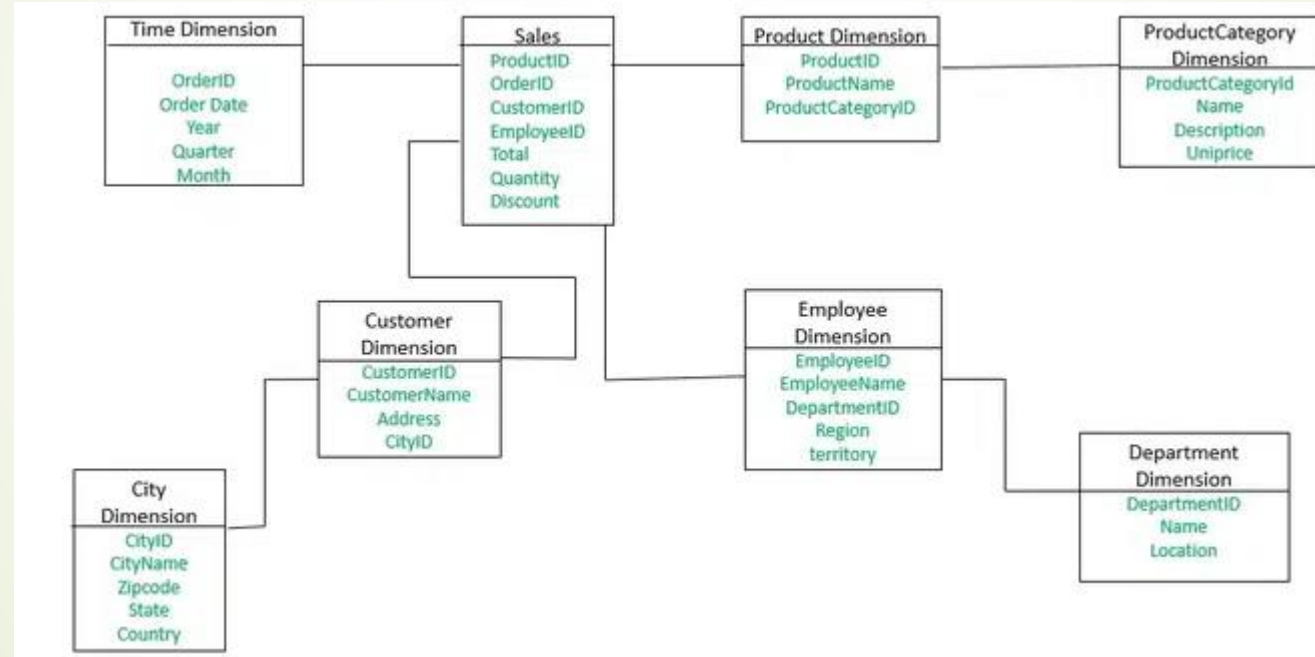


# Star Schema in Data Warehouse modeling



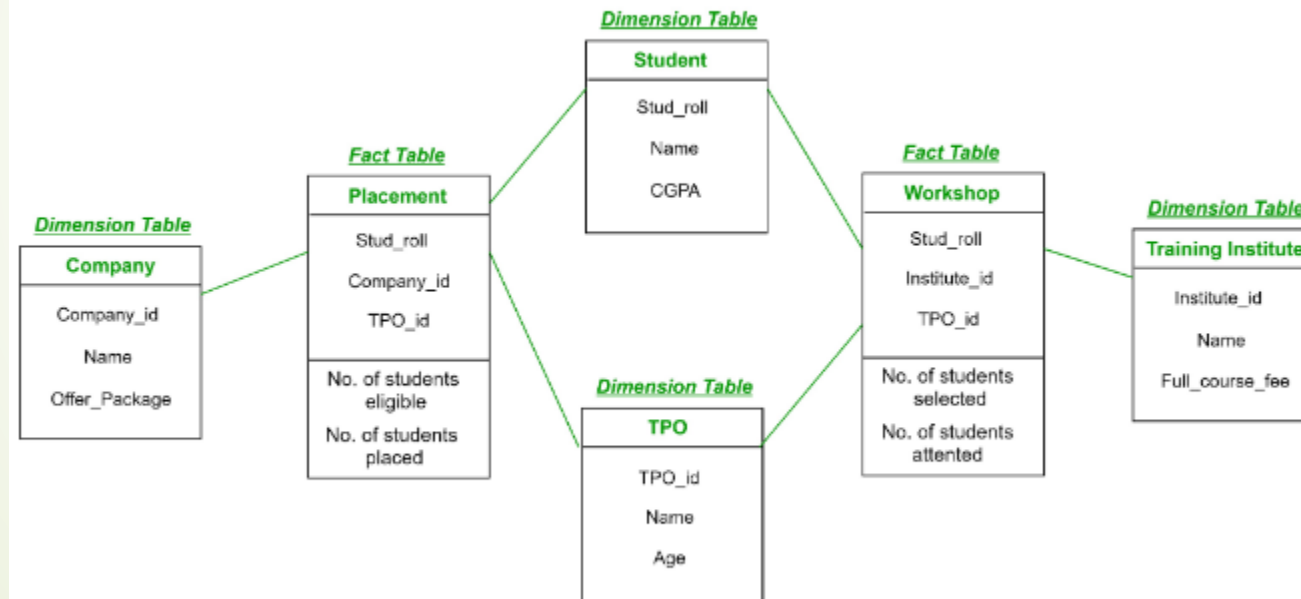


# Snowflake Schema in Data Warehouse Model



# Fact Constellation in Data Warehouse modelling

Example:



# What is Metadata?

- Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book.
- 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.

## Data Warehouse Metadata

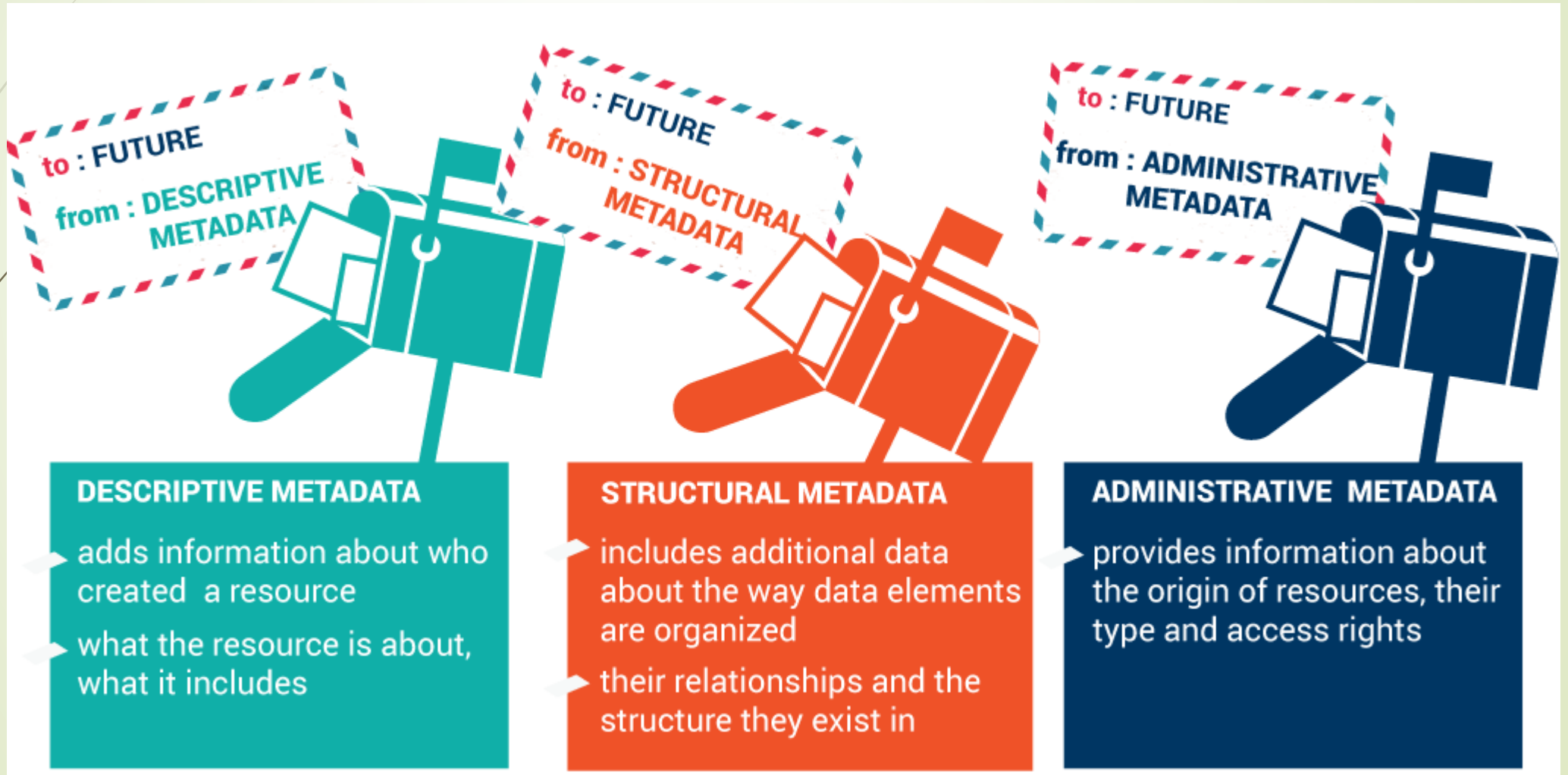




# Categories of Metadata

- **Operational Metadata** :- data for the data warehouse comes from various operational systems of the enterprise. These source systems include different data structures. The data elements selected for the data warehouse have various fields lengths and data types.
- **Extraction and Transformation Metadata**:- Extraction and transformation metadata include data about the removal of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction. Also, this category of metadata contains information about all the data transformation that takes place in the data staging area.
- **End-User Metadata** ;- The end-user metadata is the navigational map of the data warehouses. It enables the end-users to find data from the data warehouses. The end-user metadata allows the end-users to use their business terminology and look for the information in those ways in which they usually think of the business.

# The types of metadata





# The types of metadata


- Metadata is categorized by the purpose it serves in the business environment. The four types of metadata are
- Descriptive:-this kind of metadata represents the descriptive knowledge of any data. For example, it may include title, subject, genre, author name, etc. it helps in understanding the data in the context of discovery and identification.
  - Unique identifiers (such as an ISBN)
  - Physical attributes (such as file dimensions or Pantone colors)
  - Bibliographic attributes (such as the author or creator, title, and keywords)
- Structural:- this kind of metadata represents the structure of the data or file. It shows information about how the resource data or file is organized. An example of this kind of data can be the table of contents of any article or book and the name of the components of the human body.
  - Page numbers
  - Sections
  - Chapters
  - Indexes
  - Table of contents
- administrative :- this type of metadata includes information about details of the data or file like when it was created, rights, permissions, technical characteristics like size, file type, data type, etc. it is often created automatically when we enter the data in the database. They are helpful in managing the content.
  - **Technical Metadata** – Information necessary for decoding and rendering files
  - **Preservation Metadata** – Information necessary for the long-term management and archiving of digital assets
  - **Rights Metadata** – Information pertaining to intellectual property and usage rights



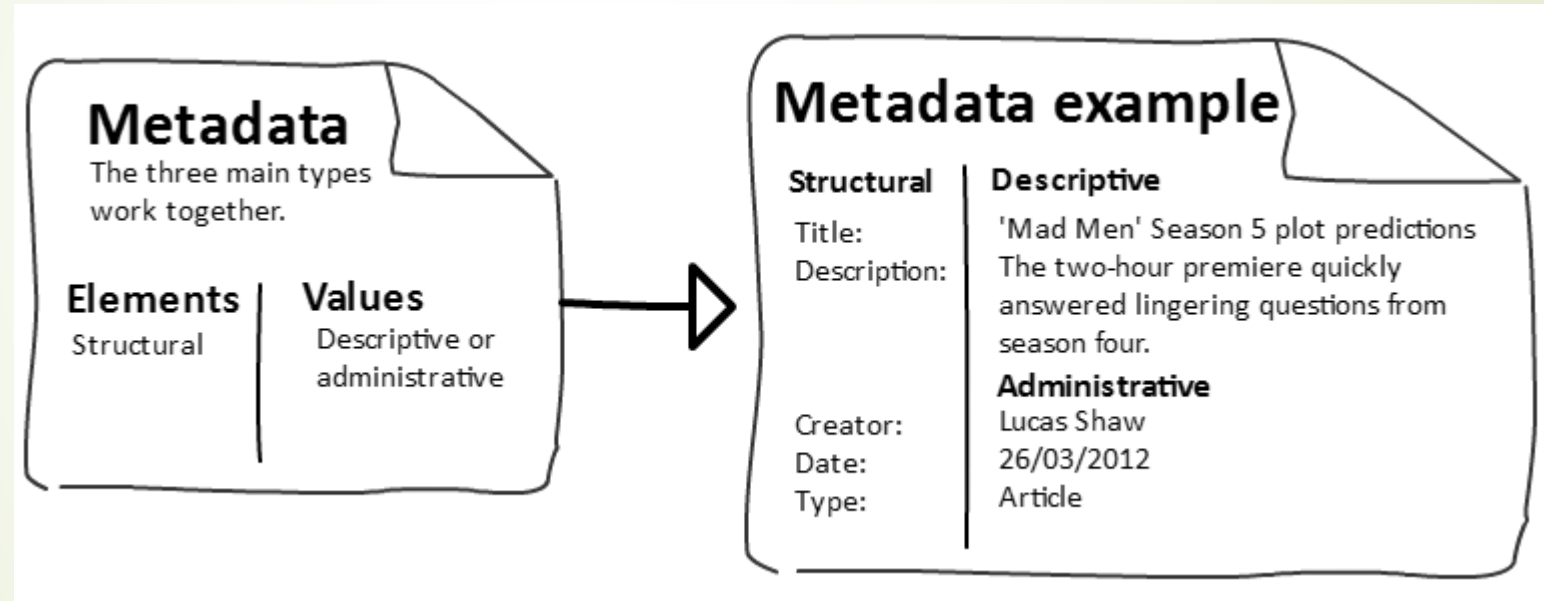


# Classification of Metadata

- Data Acquisition
  - Data Extraction
  - Data Transformation
  - Data Cleaning
  - Data Integration
  - Data Staging
  - Data Storage
    - Data Loading
    - Data Archiving
    - Data management
- 

- 
- 
- Information Delivery
    - Report Generation
    - Query Processing
    - Complex Analysis

# Diagram





# Tools for metadata Management

- Metadata Repository
- The metadata itself is housed in and controlled by the metadata repository. The software of metadata repository management can be used to map the source data to the target database, integrate and transform the data, generate code for data transformation, and to move data to the warehouse.
- **Benefits of Metadata Repository**
- It provides a set of tools for enterprise-wide metadata management.
- It eliminates and reduces inconsistency, redundancy, and underutilization.
- It improves organization control, simplifies management, and accounting of information assets.
- It increases coordination, understanding, identification, and utilization of information assets.
- It enforces CASE development standards with the ability to share and reuse metadata.

- 
- 
- Meta Extraction Tools
  - Data Modeling
  - ETL
  - BI Reporting