DATA WAREHOUSE

FADIL INDRA SANJAYA, S.KOM., M.KOM

Data Warehouse Concept

A producer wants to know....

Which are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

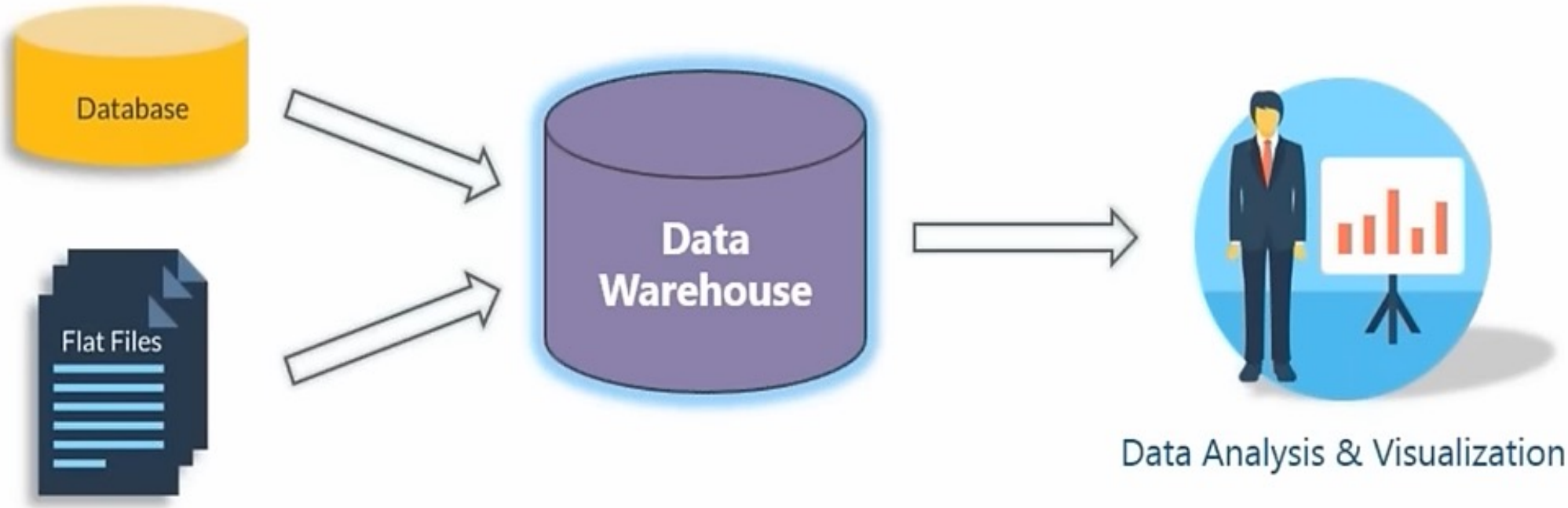
What product prom--otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

What impact will new products/services have on revenue and margins?

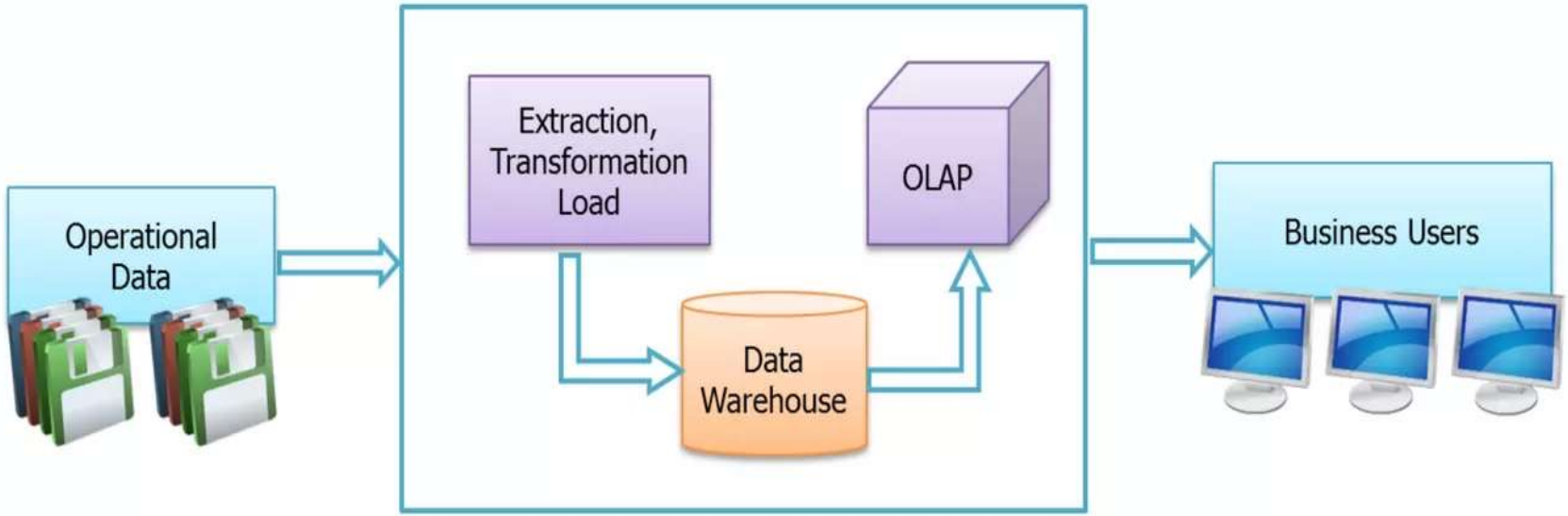
What Is A Data Warehouse?

- Data Warehouse is like a relational database designed for **analytical needs**.
- It functions on the basis of **OLAP** (Online Analytical Processing).
- It is a central location where consolidated data from multiple locations (databases) are stored.



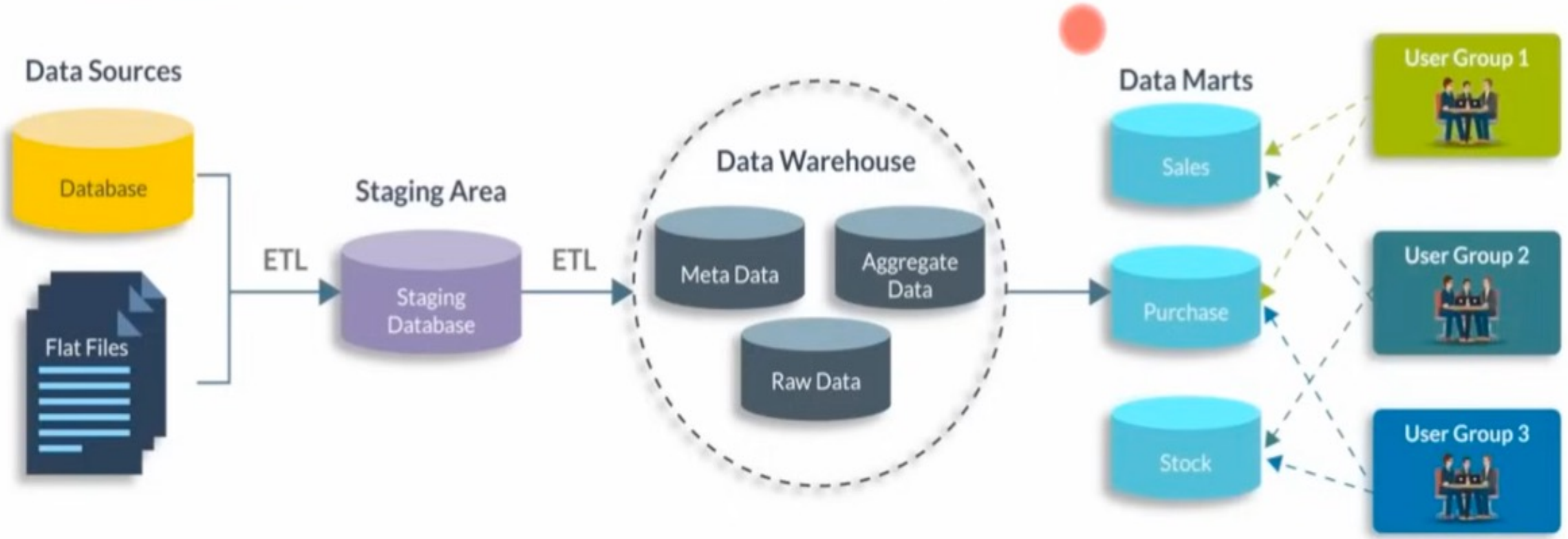
What Is A Data Warehouse?

- A central location where consolidated data from multiple locations (databases) are stored.
- DWH is maintained separately from an organization's operational database.
- End users access it whenever any information is needed.
- **Note**:- Data Warehouse is not loaded every time new data is added to database.



What Is Data Warehousing?

- Data Warehousing is the act of **organizing** & **storing** data in a way so as to make its retrieval efficient and insightful.
- It's also called as the process of transforming **data** into **information**.

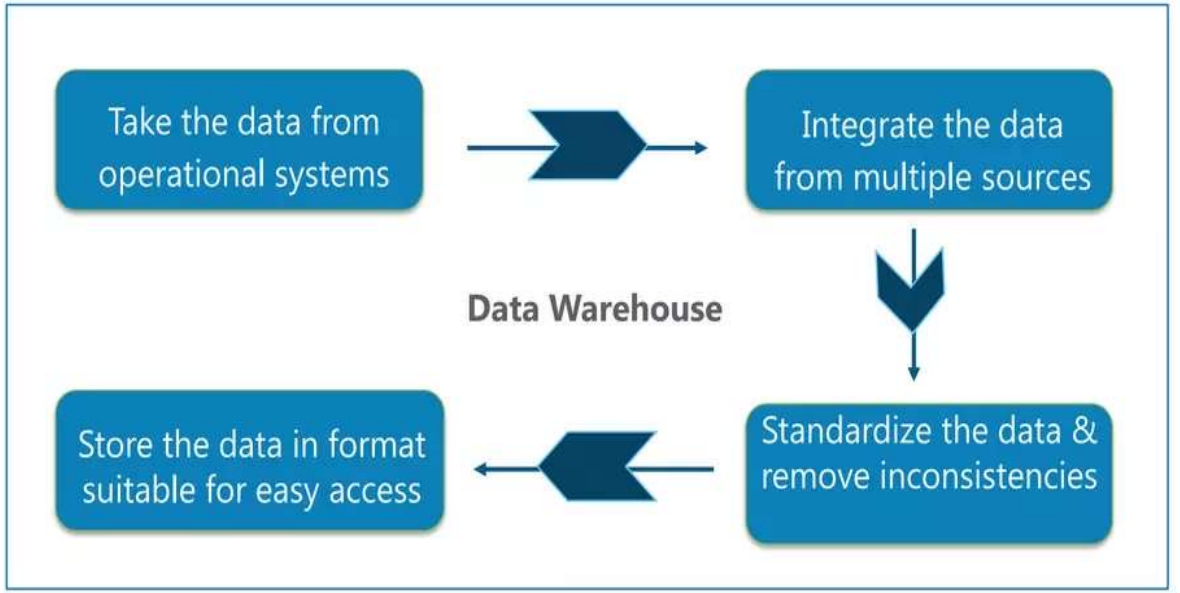


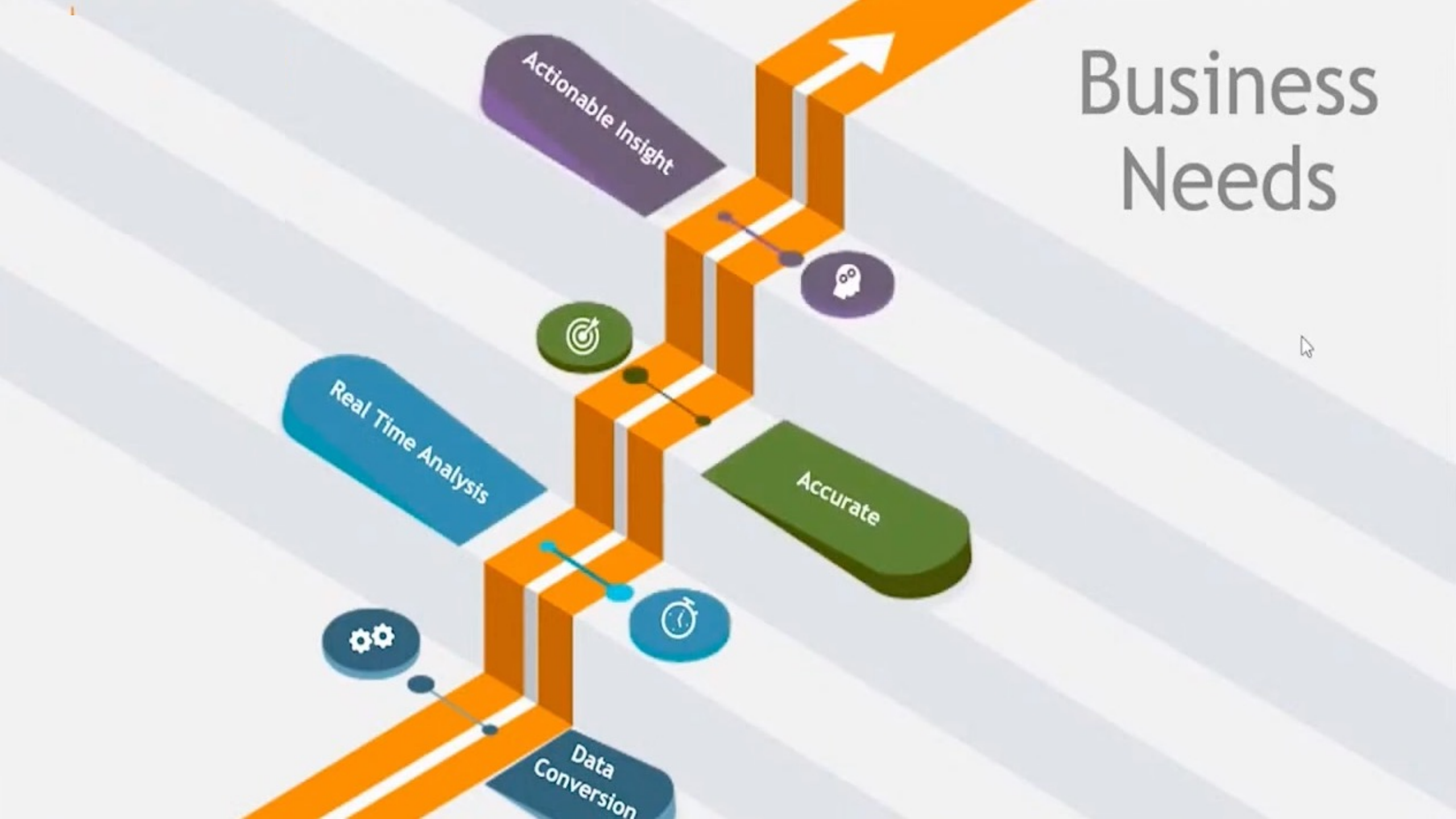
What Are The Advantages Of A Data Warehouse?

- Strategic questions can be answered by studying trends.
- Data Warehousing is faster and more accurate.
- **Note**:- Data Warehouse is not a product that a company can go and purchase, it needs to be designed & depends entirely on the company's requirement.

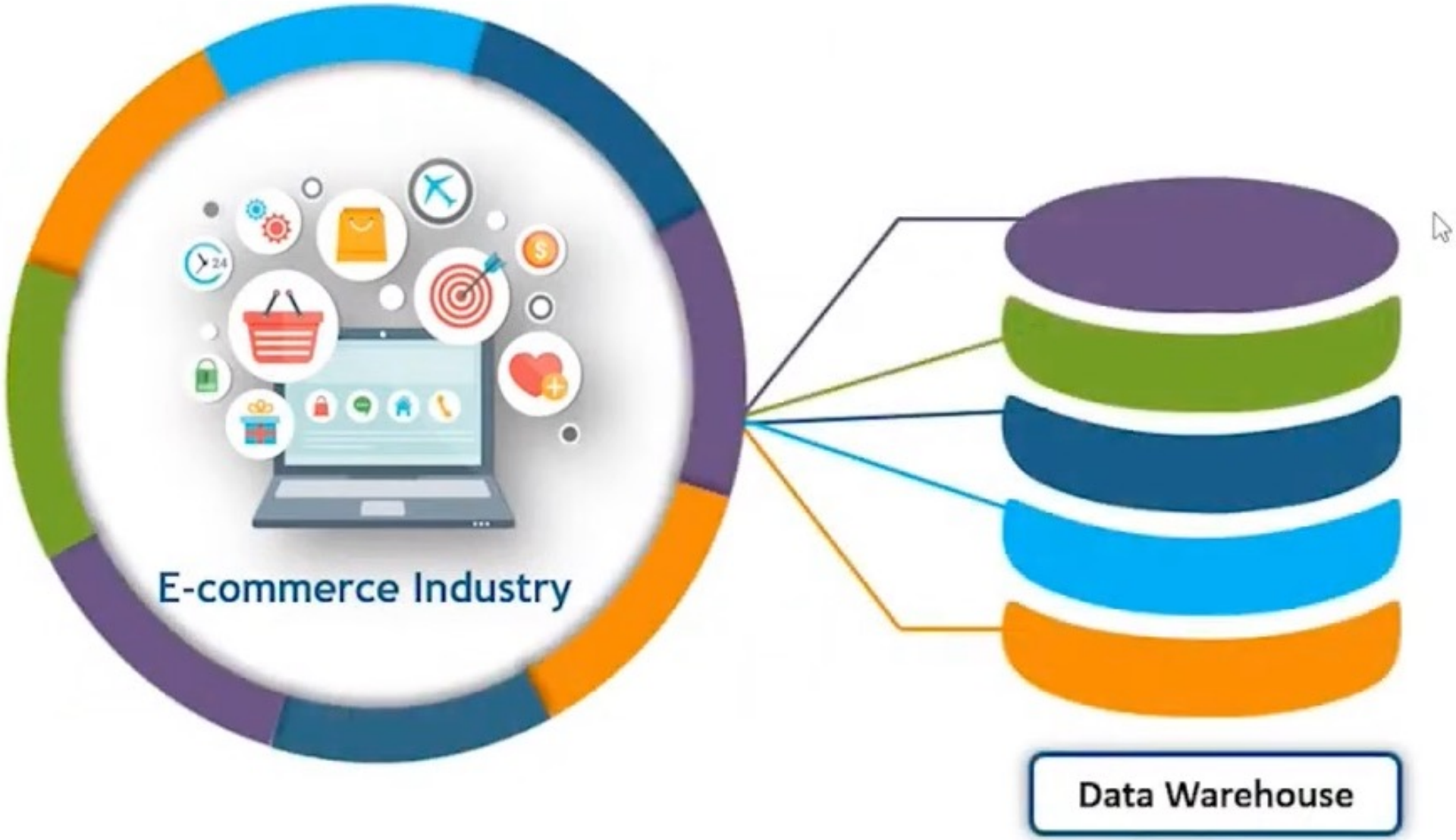
Query

Result



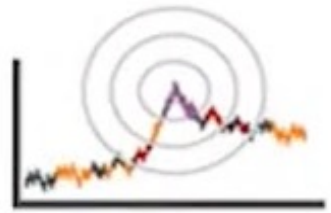

Business Needs
Actionable Insight
Real Time Analysis
Accurate
Data Conversion

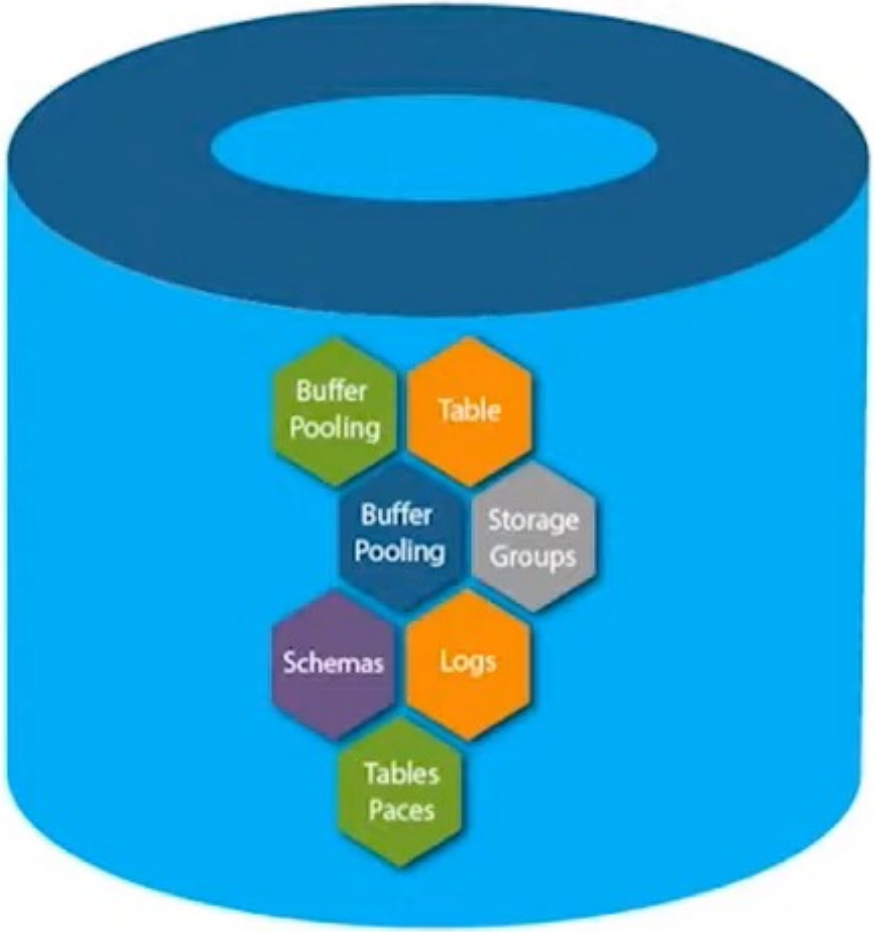

Reporting
Product Management
Promotions
Costumers Accounts Management
24/7 support
E-commerce Industry
Costumers Buy Behavior
Merchant Account
Checkout Details


E-commerce Industry
Data Warehouse

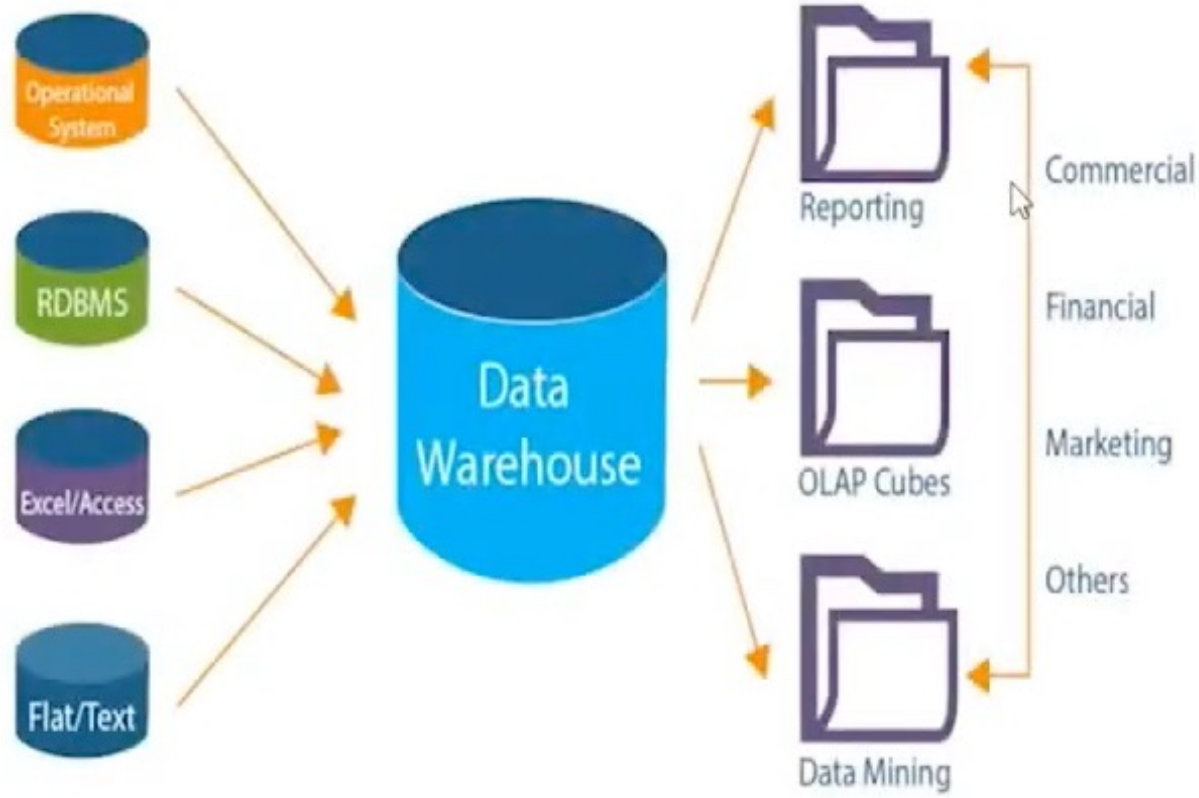
Extract – Clean –
Transform – Load

Report Generation
& Analysis

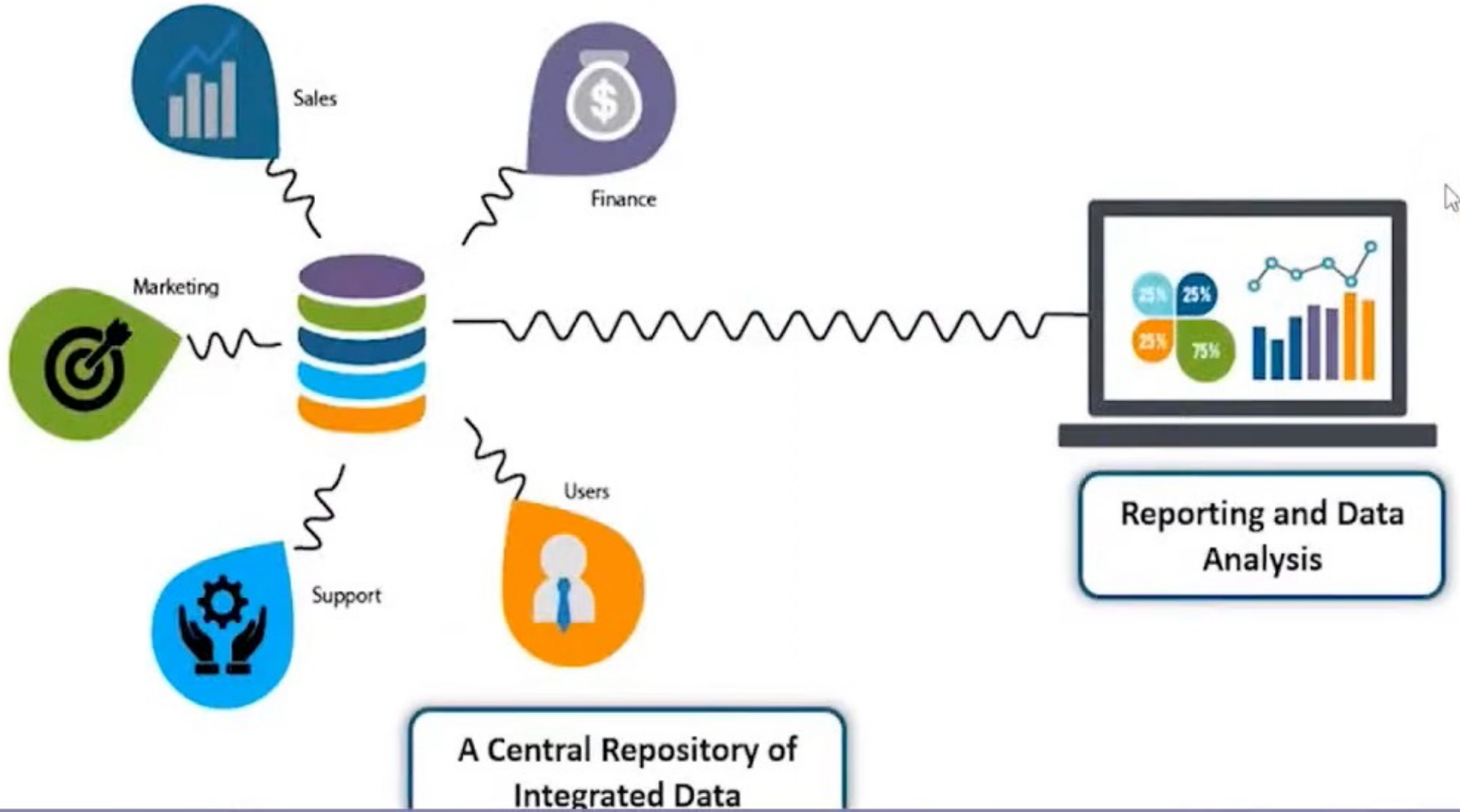




Operational Database



Data Warehouse


Sales
Finance
Marketing
Support
Users
25%
25%
25%
75%
Reporting and Data Analysis
A Central Repository of Integrated Data

Non-Volatile

- Data never removed
- Always growing

Subject-Oriented

- Gives answers to diverse questions
- Used by all functional areas

Time-Variant

- Timely data flow
- Projected data

Integrated

- Centralized
- Holds entire organization's data

Properties Of A Data Warehouse

> *"A Data Warehouse is a **subject-oriented**, **integrated**, **time-variant** and **nonvolatile** collection of data in support of management's decision-making process." -Bill Inmon, Father of Data Warehousing*

Subject-oriented

Data is categorized and stored by business subject rather than by application.

Integrated

Data on a given subject is collected from disparate sources and stored in a single place.

Time-variant

Data is stored as a series of snapshots, each representing a period of time.

Non-volatile

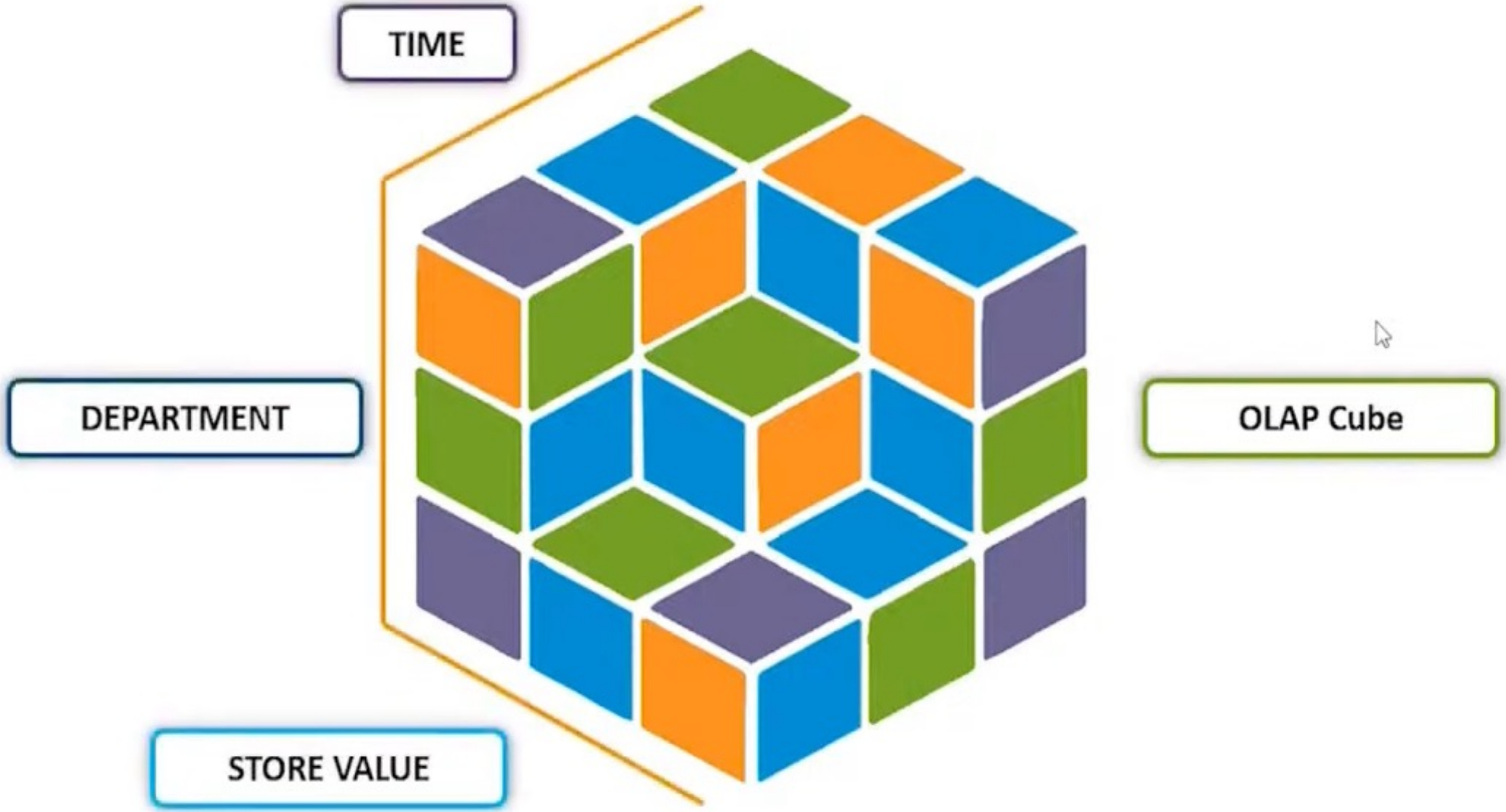
Typically data in the data warehouse is not updated or deleted.

Supports Multidimensional Queries

Enable users to view the same data in different ways

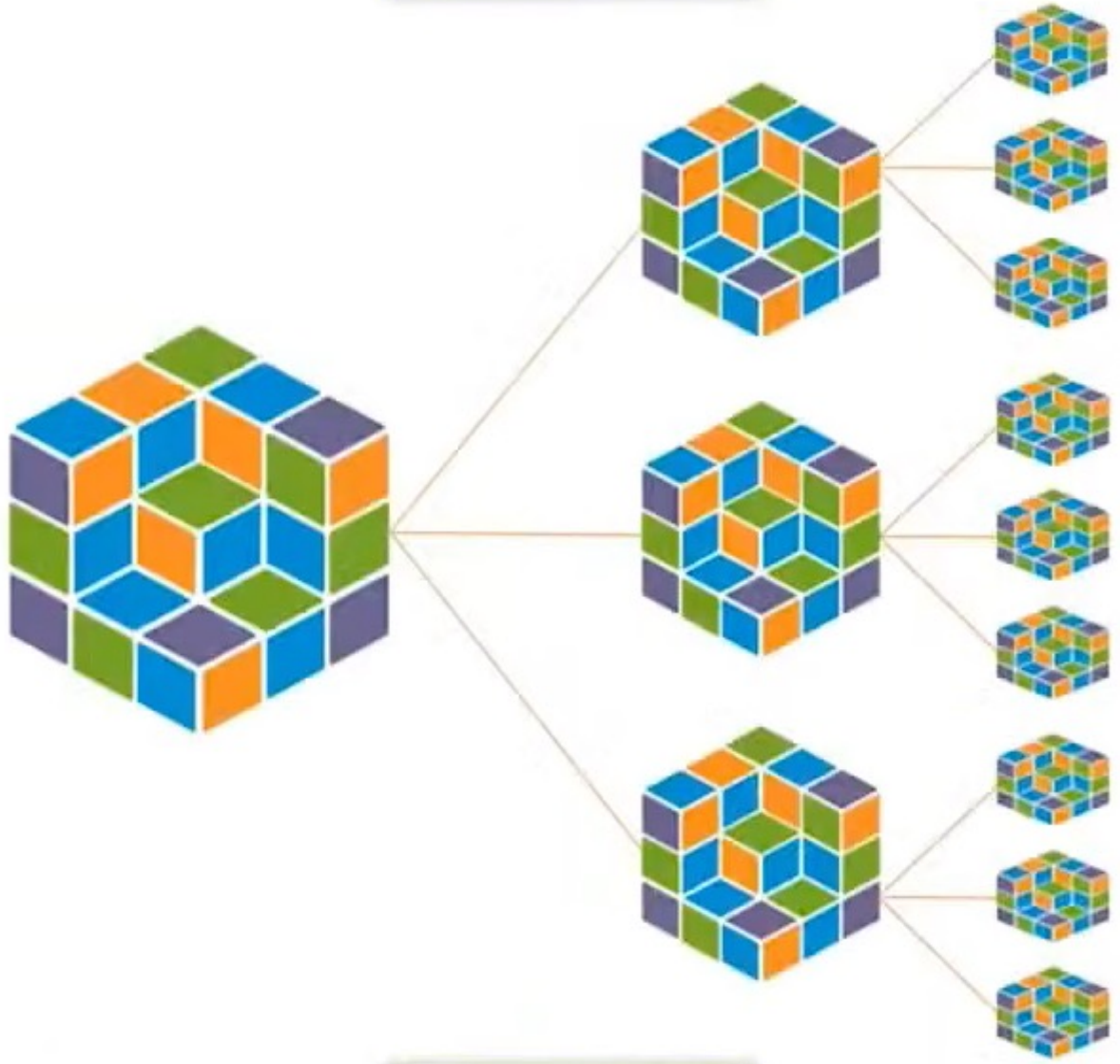
Each aspect of information represents a different dimension

Online Analytical Processing


TIME
DEPARTMENT
STORE VALUE
OLAP Cube

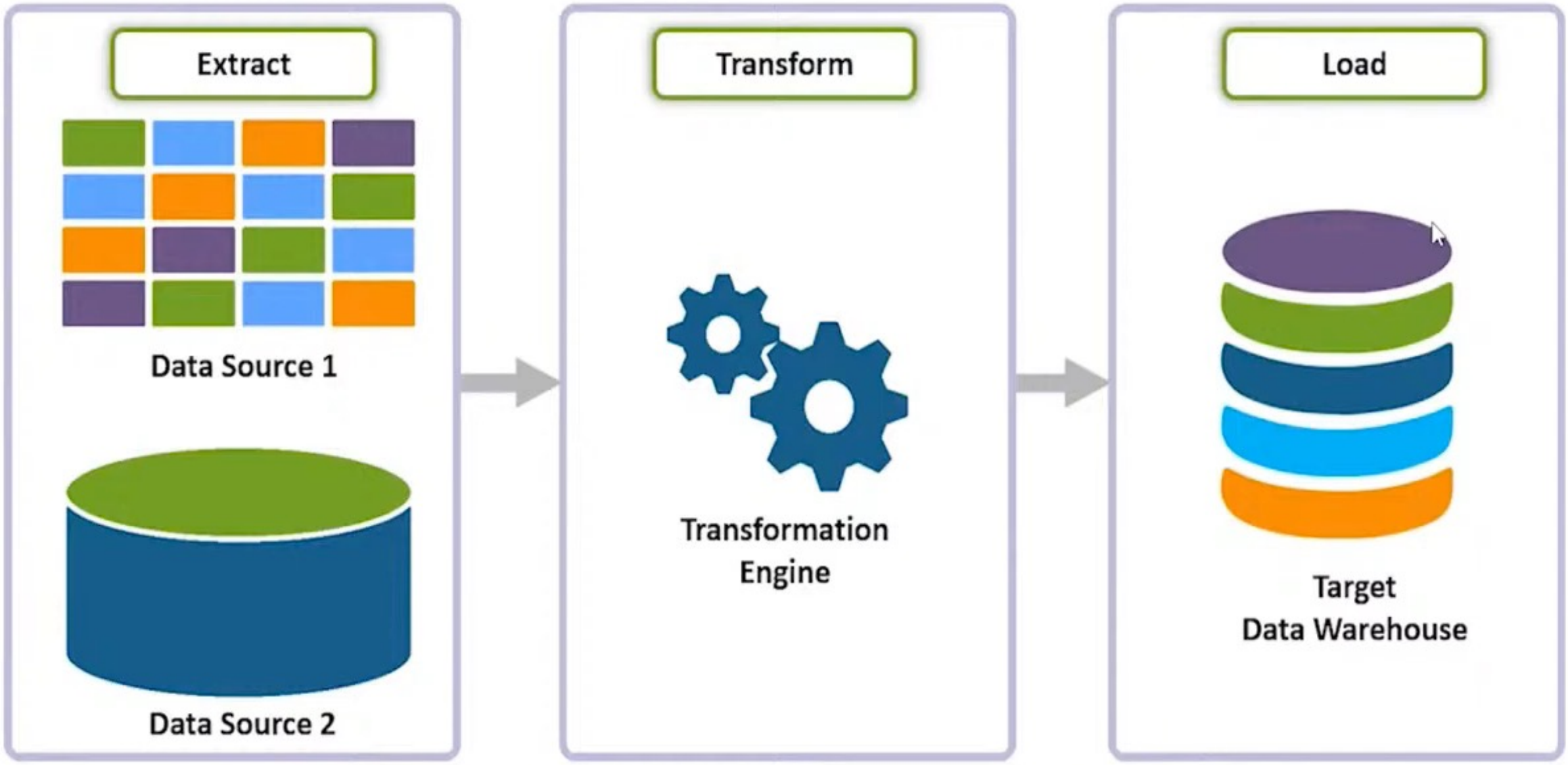
Drill Through

Drill Up

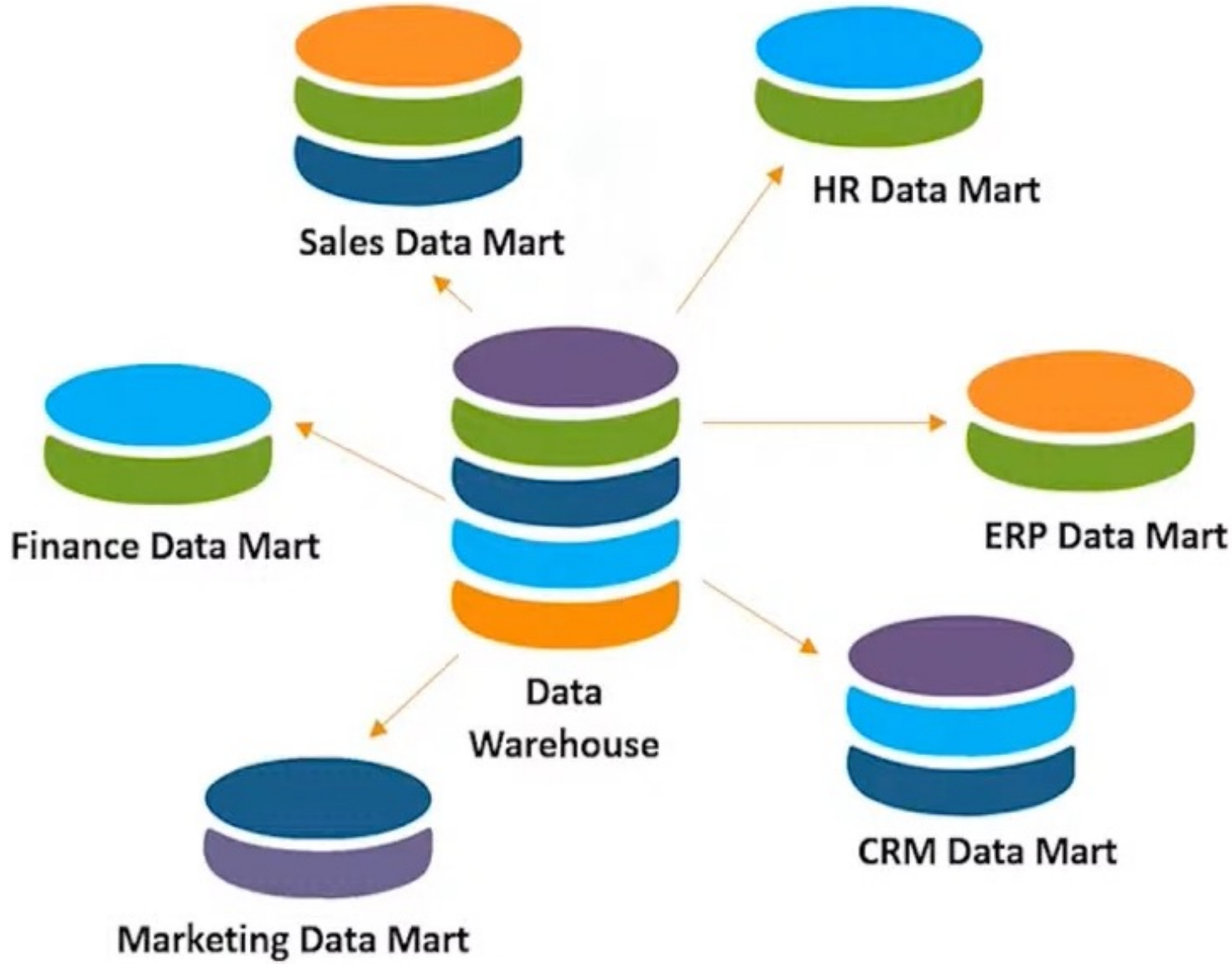


Drill Down

Drill Across


Extract
Data Source 1
Data Source 2
Transform
Transformation
Engine
Load
Target
Data Warehouse

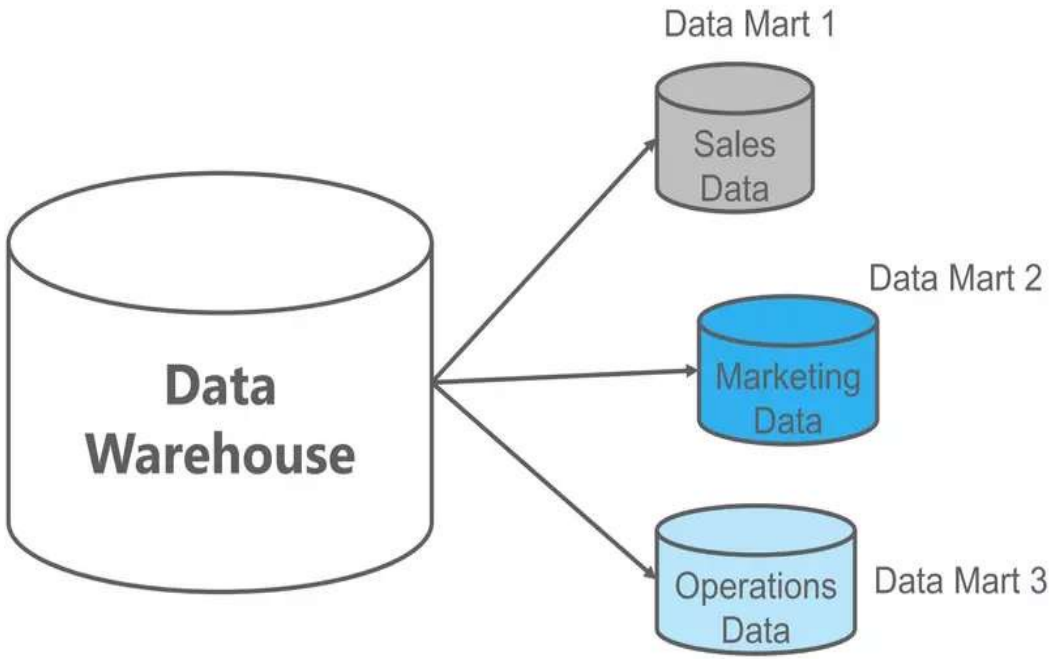
Subset of Data Warehouse



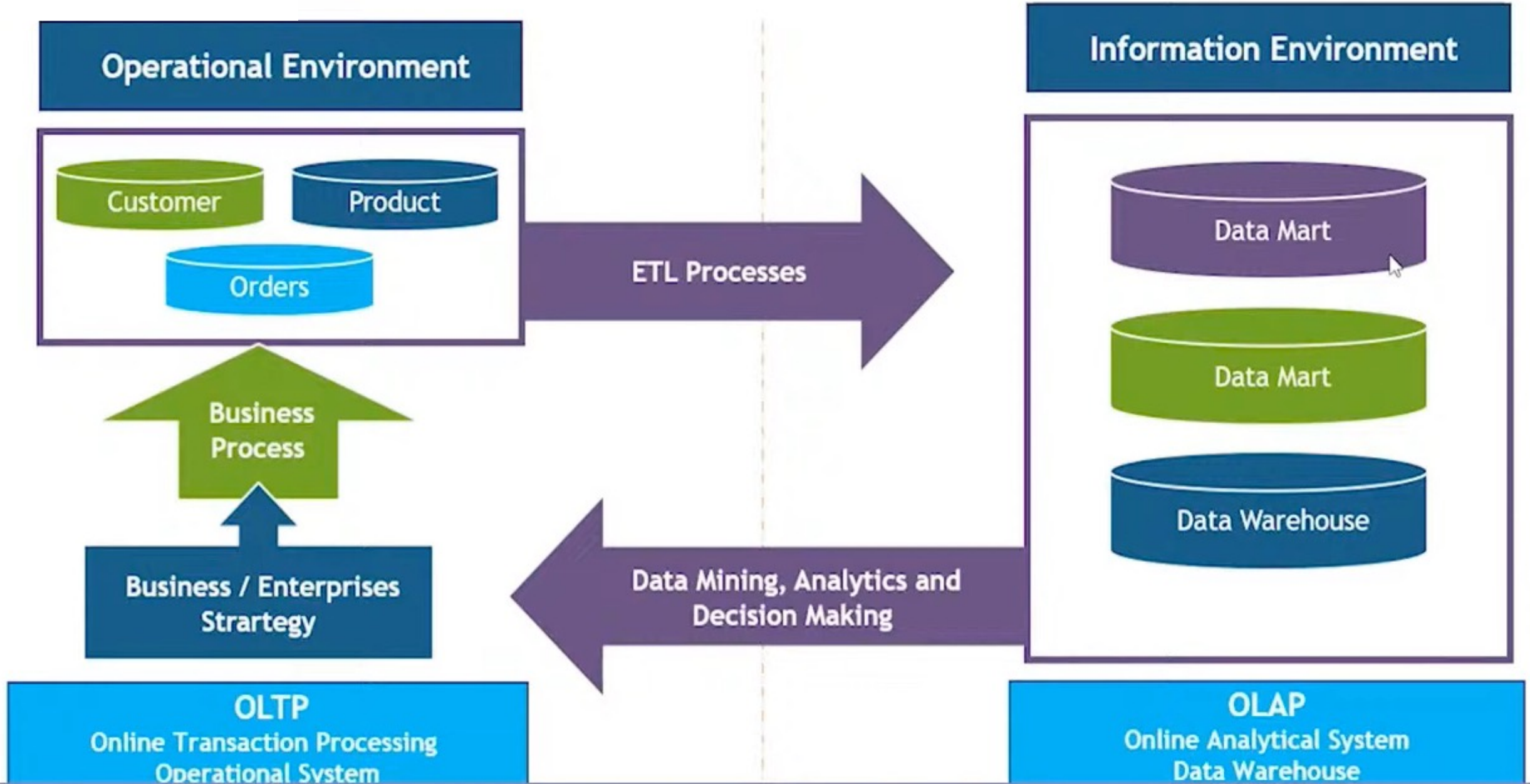
Data Mart

- Data mart is a smaller version of the Data Warehouse which deals with a single subject
- Data marts are focused on one area. Hence, they draw data from a limited number of sources
- Time taken to build Data Marts is very less compared to the time taken to build a Data Warehouse

Data Warehouse	Data Marts
Enterprise wide data	Department wide data
Multiple subject areas	Single subject area
Multiple data sources	Limited data sources
Occupies large memory	Occupies limited memory
Longer time to implement	Shorter time to implement

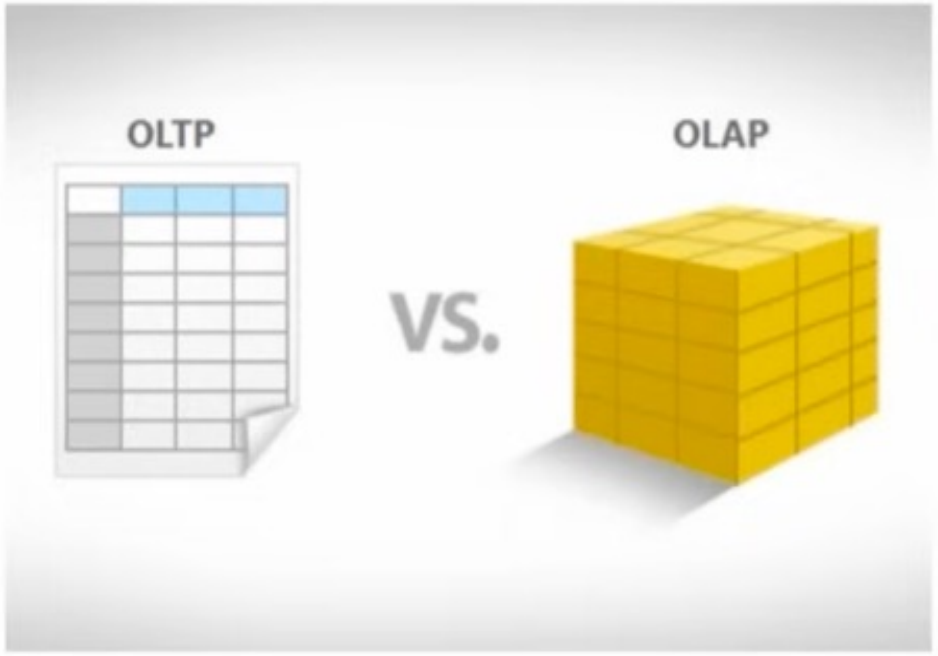


OLAP VS OLTP


Operational Environment
Customer
Product
Orders
ETL Processes
Information Environment
Data Mart
Data Mart
Data Warehouse
Business Process
Business / Enterprises Strartegy
Data Mining, Analytics and Decision Making
OLTP
Online Transaction Processing
Operational System
OLAP
Online Analytical System
Data Warehouse

OLAP (Online Analytical Processing)

- OLAP is a flexible way for you to make complicated analysis of **multidimensional data**.
- DWH is modeled on the concept of **OLAP**. DBs are modeled on the concept of OLTP (Online Transaction Processing).
- OLTP systems use data stored in the form of two-dimensional tables, with rows and columns.



Advantages Of OLAP Over OLTP

1. Opens up new views of looking at data.
2. Supports filtering/ sorting of data.
3. Data can be refined.

OLAP vs OLTP

Purpose

OLAP is for complex data analysis and reporting, enabling users to gain insights from large volumes of historical data.

OLAP vs OLTP

Purpose

OLTP is designed for real-time transaction processing, handling day-to-day operations such as inserting, updating, and deleting individual records

OLAP vs OLTP

Data Structures

Data Analysis For Everyone

OLAP organizes data into dimensions and measures to facilitate multidimensional analysis and drill-down capabilities.

OLAP vs OLTP

Data Structures

Data Analysis For Everyone

OLTP uses a normalized data model with tables and relationships, aiming for efficient transactional processing.

OLAP vs OLTP

Data Volume

OLAP deals with large volumes of historical data, typically containing years of data.

OLAP vs OLTP

Data Volume

OLTP deals with relatively smaller volumes of data, usually representing real-time transactions happening within a shorter timeframe.

OLAP vs OLTP

Response Time

OLAP allows for longer response times since it deals with complex queries and large data sets.

OLAP vs OLTP

Response Time

OLTP requires fast response times to support real-time transaction processing.

OLAP vs OLTP

Data Modification

OLAP is read-only or minimally updated. Data is loaded into OLAP cubes periodically to update the analytical database with new information.

OLAP vs OLTP

Data Modification

OLTP involves frequent data modification, including insertions, updates, and deletions.

OLAP vs OLTP

Data Backup and Recovery

OLAP data is usually derived from the OLTP system, and backup and recovery are less critical.

OLAP vs OLTP

Data Backup and Recovery

OLTP data is critical, and backup and recovery processes are essential. Regular backups are taken to ensure data integrity and provide the ability to restore the system in case of failures.

Information Systems:- OLTP (DB) vs. OLAP (DWH)

Relational Database (OLTP)	Analytical Data Warehouse (OLAP)
Contains current data	Contains historical data
Useful in running the business	Useful in analyzing the business
Based on Entity Relationship Model	Based on Star, Snowflake and Fact Constellation Schema
Provides primitive and highly detailed data	Provides summarized and consolidated data
Used for writing data into the database	Used for reading data from the data warehouse
Database size ranges from 100 MB to 1 GB	Data Warehouse size ranges from 100 GB to 1 TB
Fast; provides high performance	Highly flexible; but not fast
Number of records accessed is in tens	Number of records accessed is in millions
Ex: All bank transactions made by a customer	**Ex**: Bank transactions made by a customer at a particular time.

Information Systems:- OLTP (DB) vs. OLAP (DWH)

OLTP Examples:

1. A supermarket server which records every single product purchased at that market.
2. A bank server which records every time a transaction is made for a particular account.
3. A railway reservation server which records the transactions of a passenger.

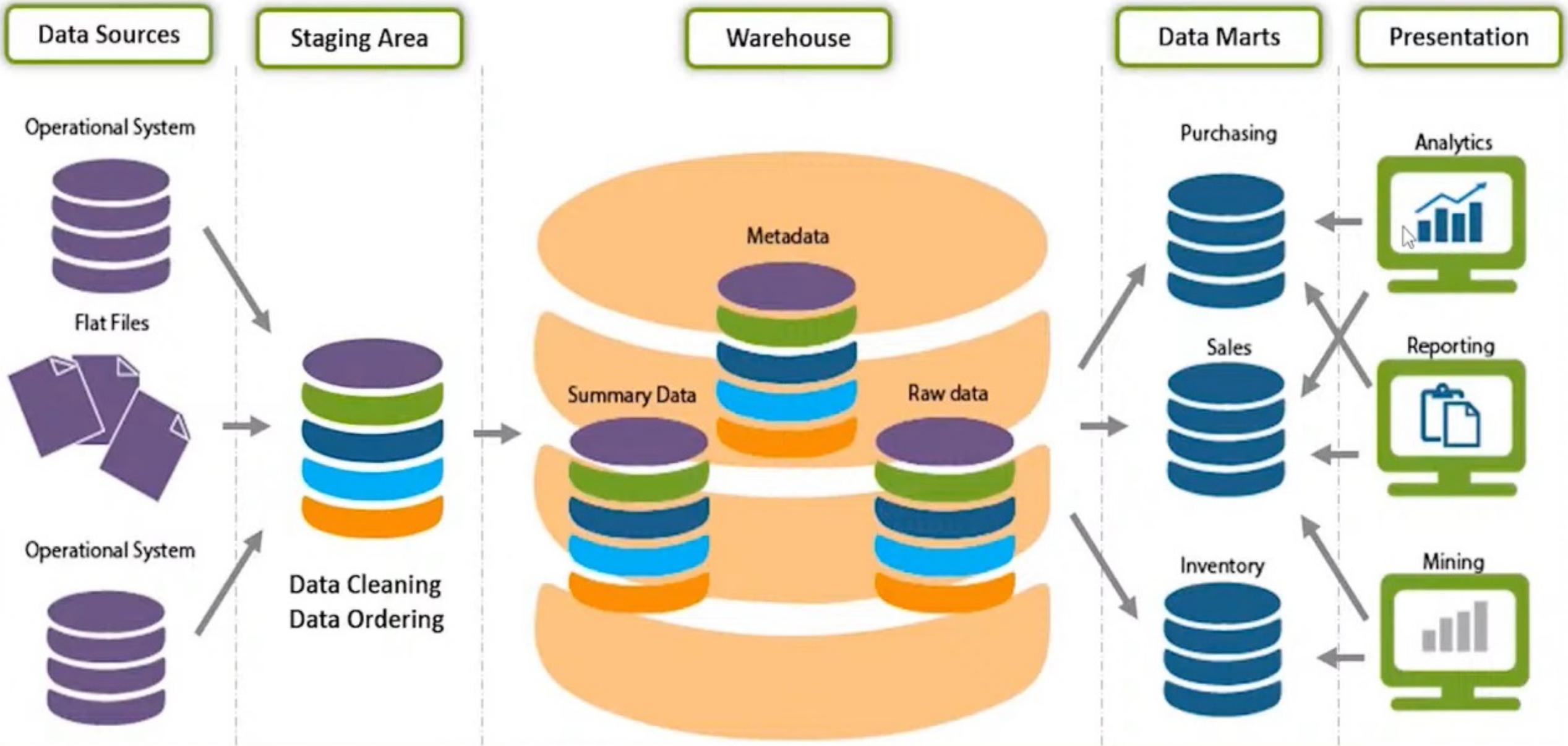
OLAP Examples:

1. Bank Manager wants to know how many customers are utilizing the ATM of his branch. Based on this he may take a call whether to continue with the ATM or relocate it.
2. An insurance company wants to know the number of policies each agent has sold. This will help in better performance management of agents.

DATA WAREHOUSE ARCHITECTURE

An Interface Design from Operational Systems

The Individual Data Warehouse Design


Data Sources
Staging Area
Warehouse
Data Marts
Presentation
Operational System
Flat Files
Operational System
Data Cleaning
Data Ordering
Metadata
Summary Data
Raw data
Purchasing
Sales
Inventory
Analytics
Reporting
Mining

ETL (EXTRACT, TRANFORM, LOAD)

ETL → Extract, Transform & Load

ETL is the process of extracting the data from various sources, transforming this data to meet your requirement and then loading it into a target data warehouse.

WHY ETL?

- Companies need a way to analyze their data for critical business decisions.
- Transactional Database can’t answer complex business questions.
- A data warehouse provide a common data repository
- ETL provide a method of moving data from various source into a data warehouse.

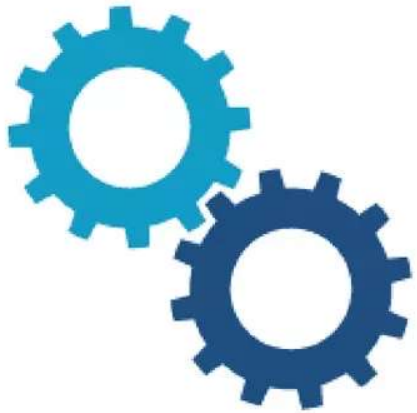
Extraction

Gathering the data

- Raw data that was written directly into the disk.
- Data written to flat files or relational tables from structured source systems.
- Data can be read multiple times, if needed.

Cleansing the data

- Eliminate duplicates or fragmented data.
- Exclude unwanted / unneeded information.



Transform

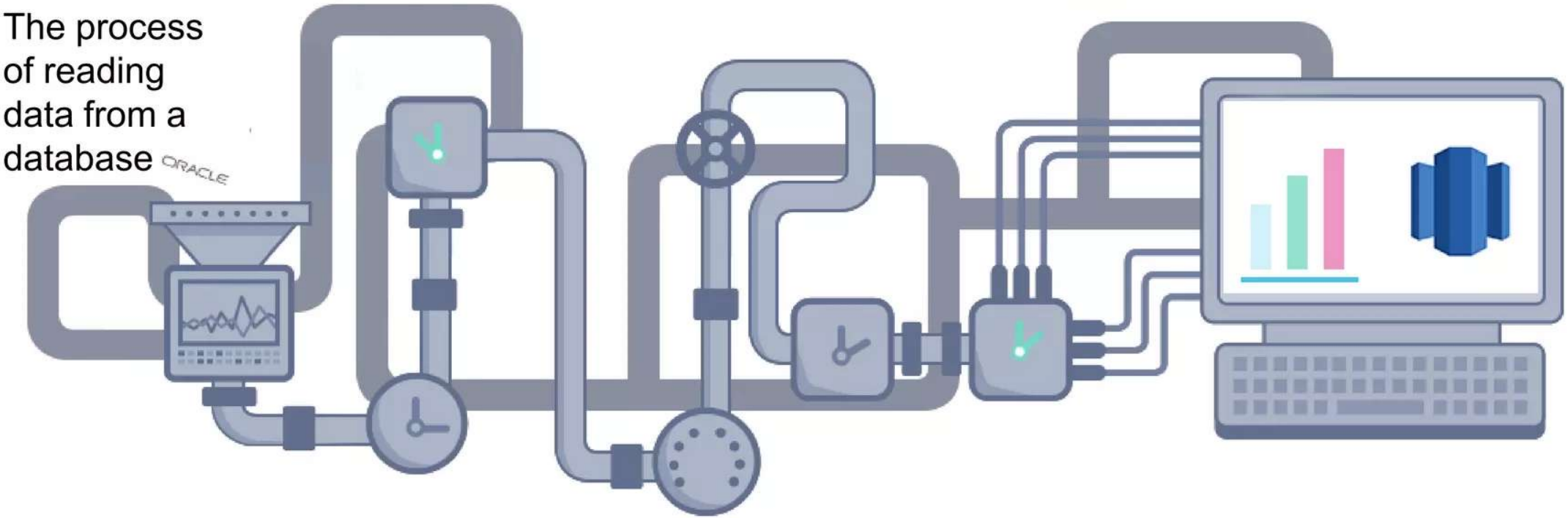
- Preparing the data to be housed in the data warehouse.
- Converting the extracted data:
  - Using rules and lookup tables
  - Combining data
  - Verification/Validity checks
  - Standardization

Load

- Storing the transformed data in the data warehouse.
- Batch/Real-time processing
- Can follow star schema and snowflake schema

Extraction Transform Load Process

The process of reading data from a database



The process of writing data into the target database

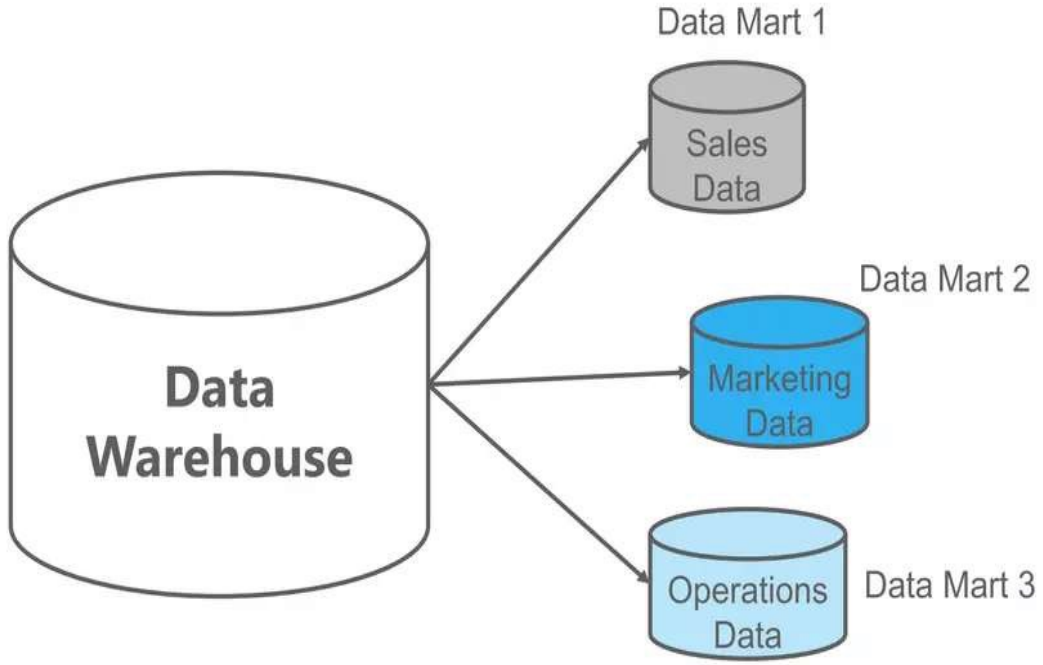
The process of converting data from one form to another

DATA MART

Data Mart

- Data mart is a smaller version of the Data Warehouse which deals with a single subject
- Data marts are focused on one area. Hence, they draw data from a limited number of sources
- Time taken to build Data Marts is very less compared to the time taken to build a Data Warehouse

Data Warehouse	Data Marts
Enterprise wide data	Department wide data
Multiple subject areas	Single subject area
Multiple data sources	Limited data sources
Occupies large memory	Occupies limited memory
Longer time to implement	Shorter time to implement

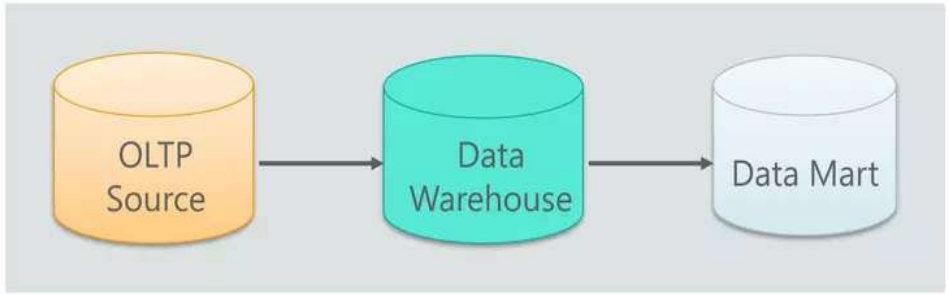


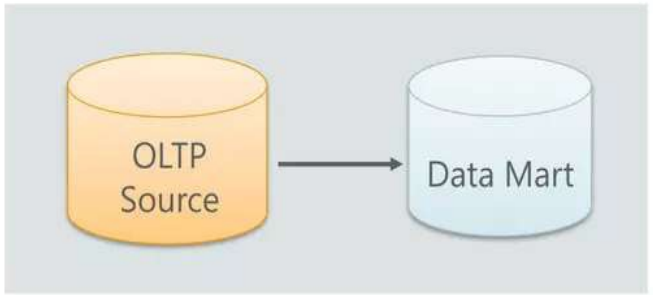
Types Of Data Mart

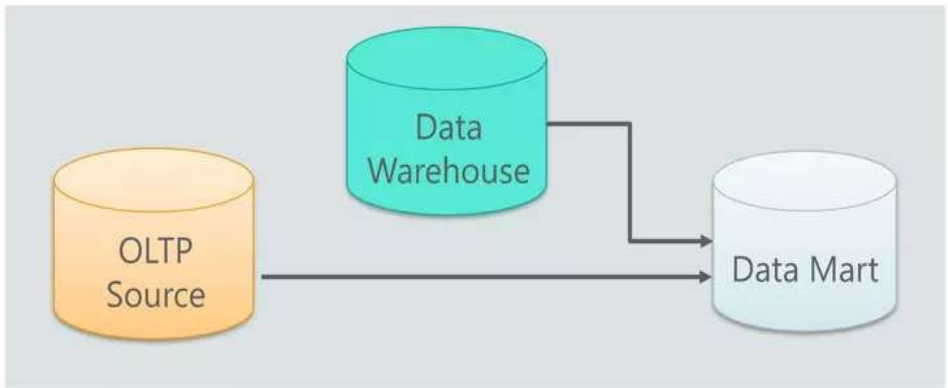
1. **Dependent Data Mart**
   - The data is first extracted from the OLTP systems and then populated in the central DWH
   - From the DWH, the data travels to the Data Mart

2. **Independent Data Mart**
   - The data is directly received from the source system
   - This is suitable for small organizations or smaller groups within an organization

3. **Hybrid Data Mart**
   - The data is fed both from OLTP systems as well as the Data Warehouse

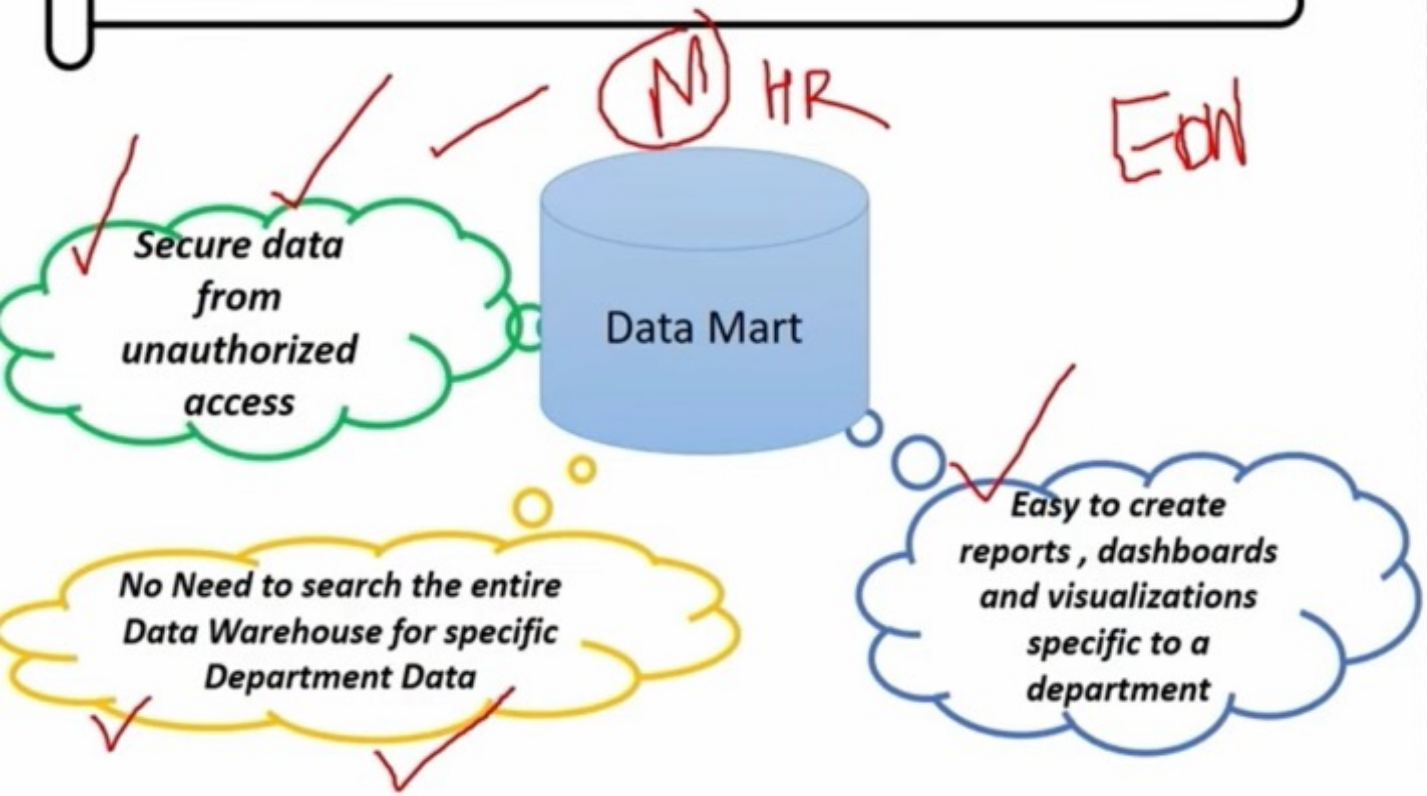


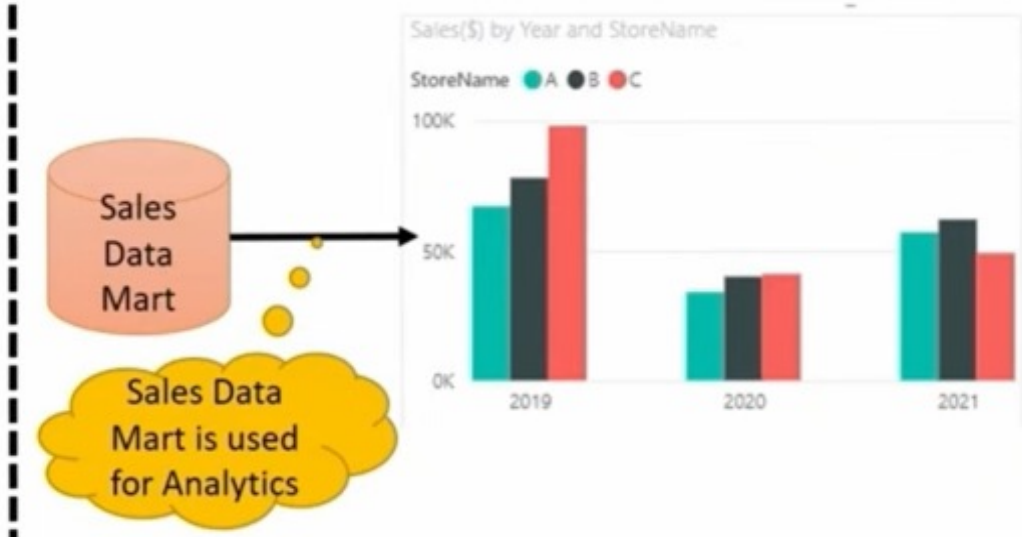


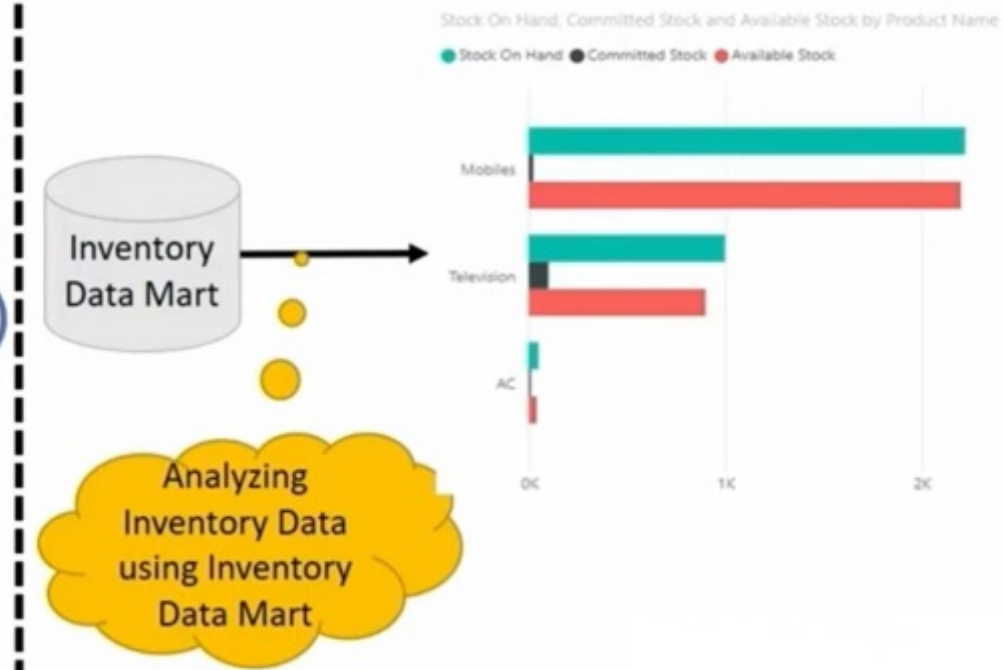


Data Marts

Data Mart : It is a Subset of Data Warehouse focused towards specific Line of Business or built for specific group of users.





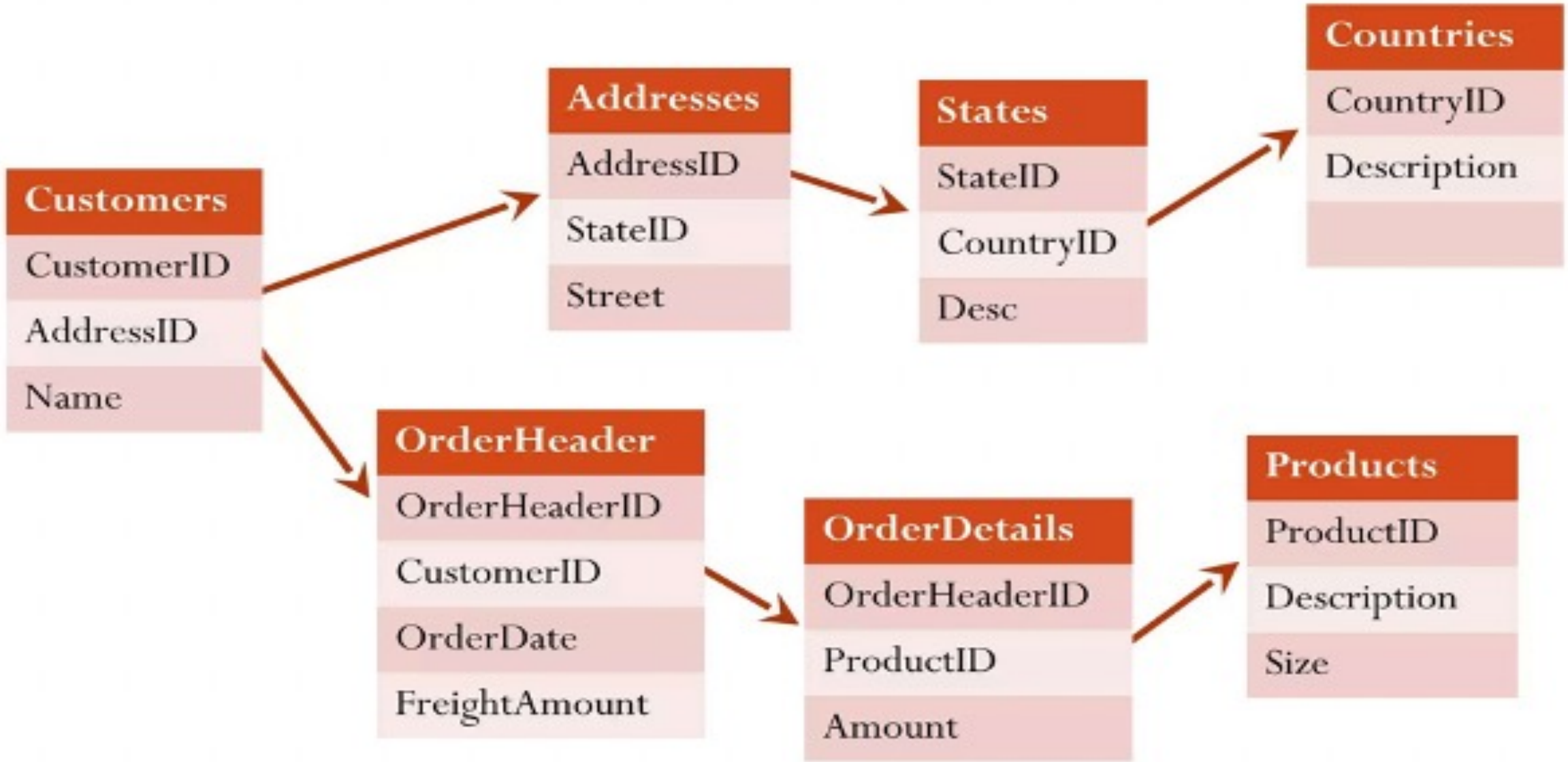


FACT VS DIMENSIONS

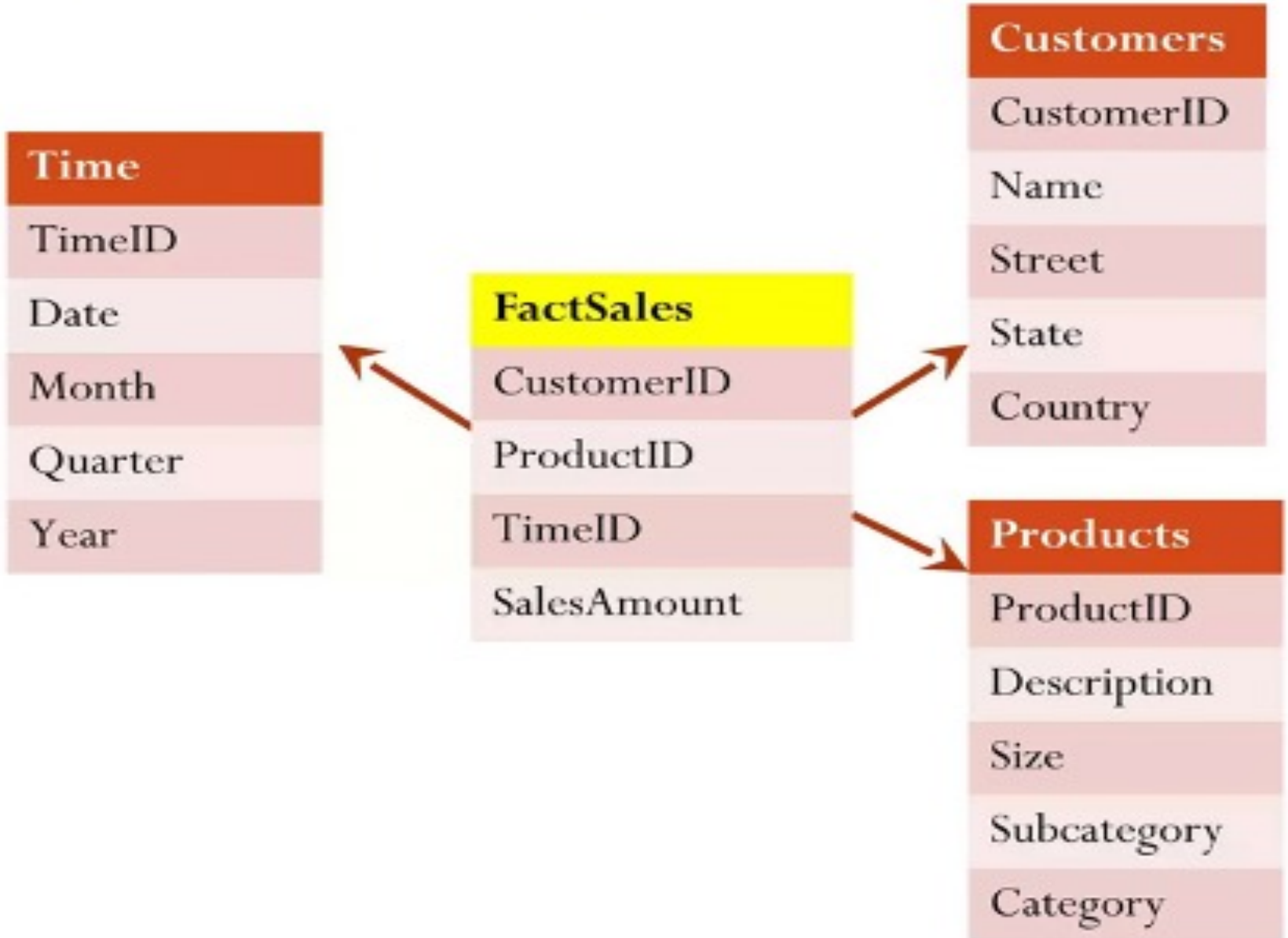
Dimensional Modeling

- **Dimensional modeling** (DM) names a set of techniques and concepts used in data warehouse design.
- Dimensional modeling is one of the methods of data modeling, that help us store the data in such a way that it is relatively easy to retrieve the data from the database.
- Dimensional modeling always uses the concepts of facts (measures), and dimensions (context).

A Transactional Database



A Dimensional Model



Facts & Dimensions

- There are two main types of objects in a dimensional model
  - **Facts** are quantitative measures that we wish to analyse and report on.
  - **Dimensions** contain textual descriptors of the business. They provide *context* for the facts.

Dimensions

- The tables that describe the dimensions involved are called **Dimension tables**.
- Dividing a Data Warehouse project into dimensions provides structured information for analysis & reporting.

E-commerce Company								
Customer			Product			Date		
ID	Name	Address	ID	Name	Type	Order date	Shipment date	Delivery date

← Subject

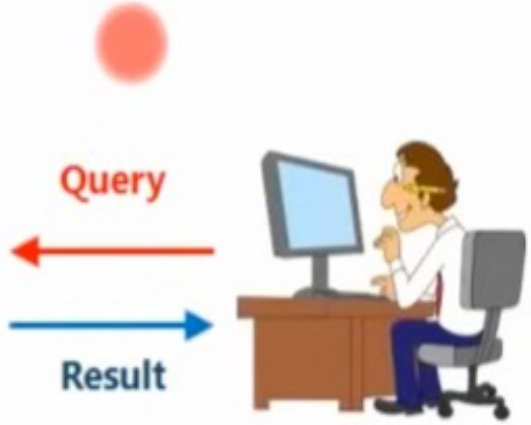
← Dimensions

← Attributes

Dimensions

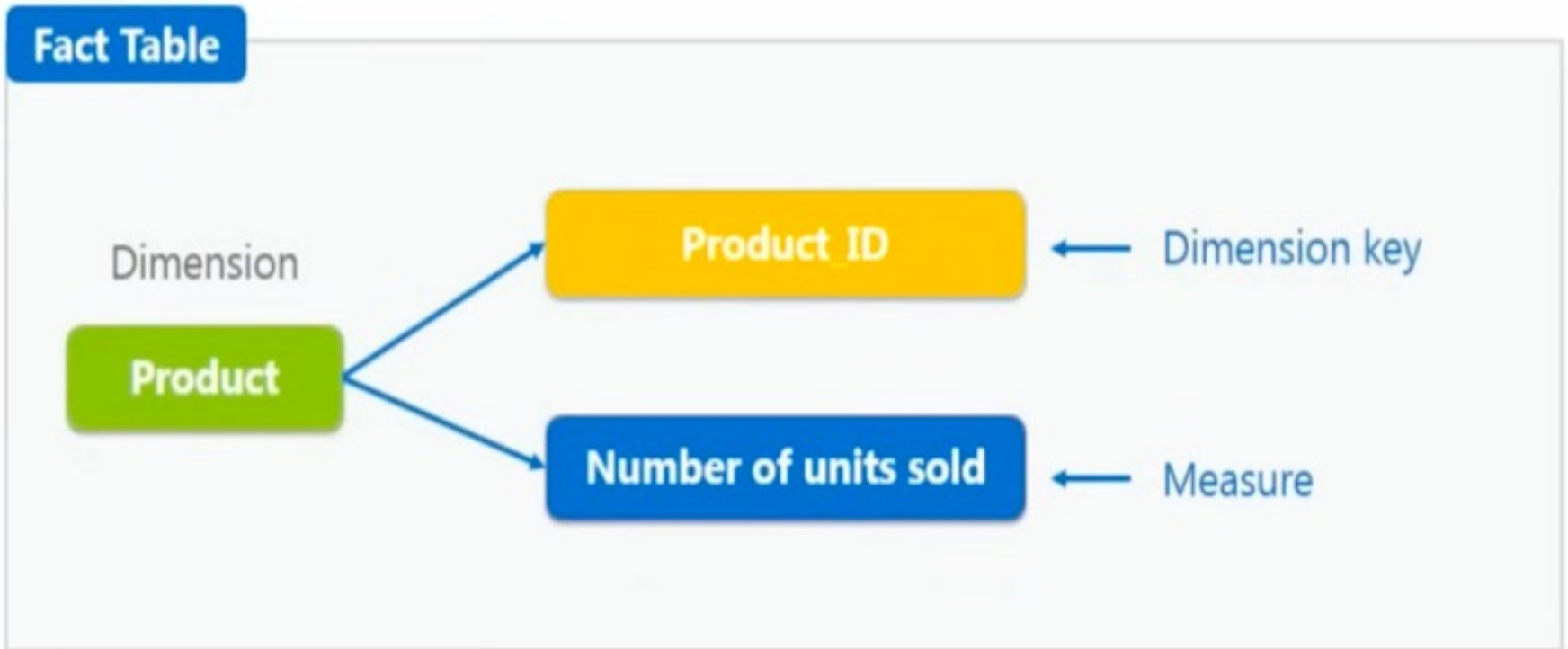
- End users fire queries on these dimension tables which contain descriptive information.

E-commerce Company								
Customer			Product			Date		
ID	Name	Address	ID	Name	Type	Order date	Shipment date	Delivery date
1	Rita	ABC	001	CD	1A	1/06/14	3/06/14	5/06/14
2	John	XYZ	002	AC	2B	6/06/14	9/06/14	11/06/14
3	Paul	PQR	003	TV	3C	10/06/14	14/06/14	16/06/14



Facts & Measures

- A fact is a measure that can be summed, averaged or manipulated.
- A Fact table contains 2 kinds of data – a **dimension key** and a **measure.**
- Every Dimension table is linked to a Fact table.



Schemas

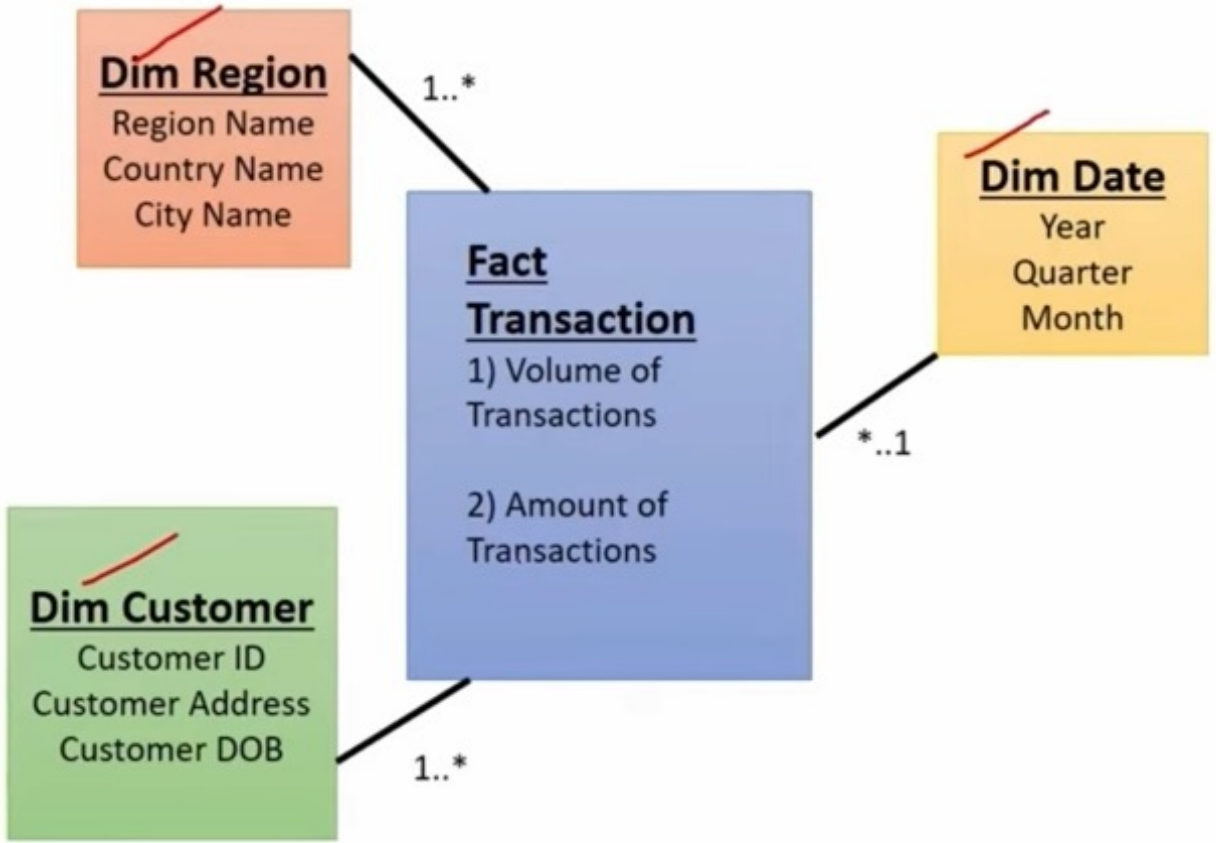
- A schema gives the logical description of the entire data base.
- It gives details about the constraints placed on the tables, key values present & how the key values are linked between the different tables.
- A database uses relational model, while a data warehouse uses **Star**, **Snowflake** and **Fact Constellation** schema.

Employee				
ID	First Name	Last Name	Age	Dept_ID
1234	Rita	Joe	25	0674
4321	John	Smith	35	0825
5678	Paul	Brady	45	0752
7890	Rose	Michael	65	0825

Linked

Department	
Dept_ID	Dept_Name
0674	Sales
0752	HR
0825	Production

Dimensions and Facts in Data Warehouse(Scenario)



Real World scenario

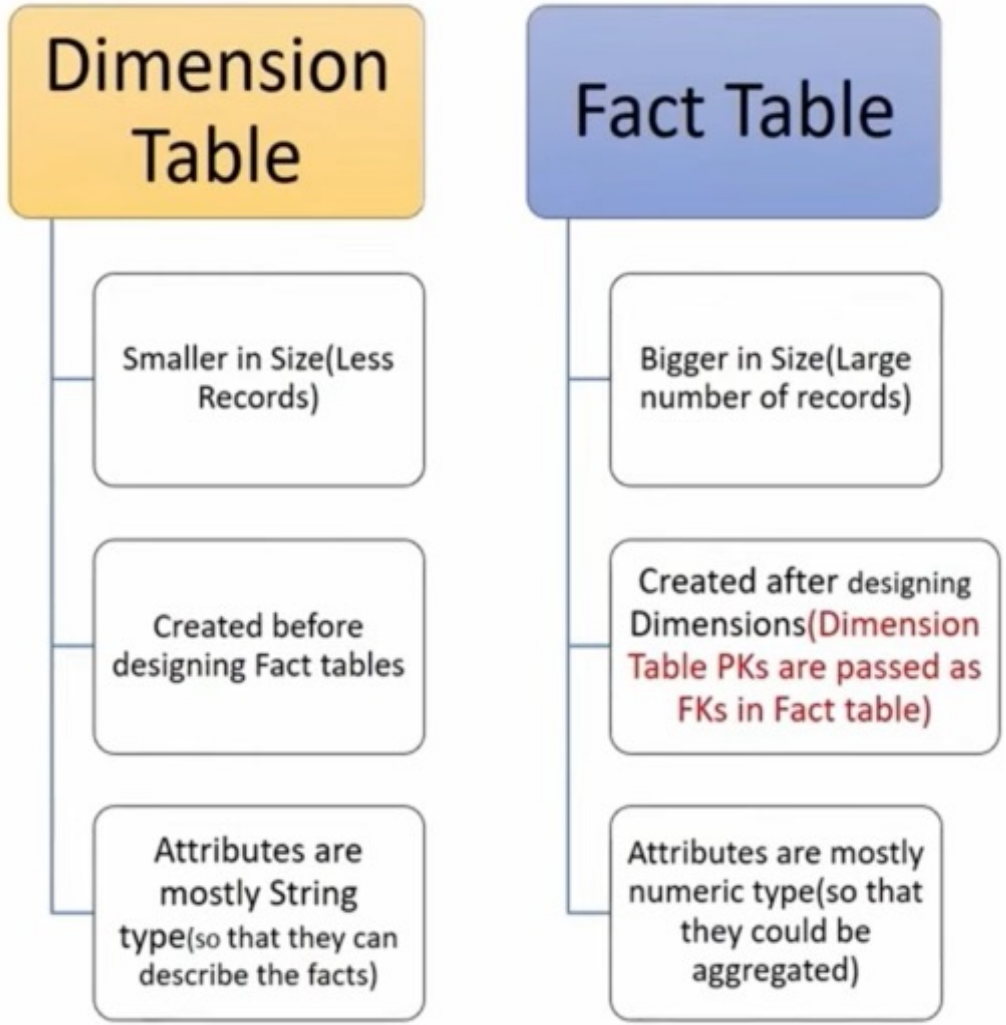
We can analyze Fact : volume of digital transactions made through payment gateways for a

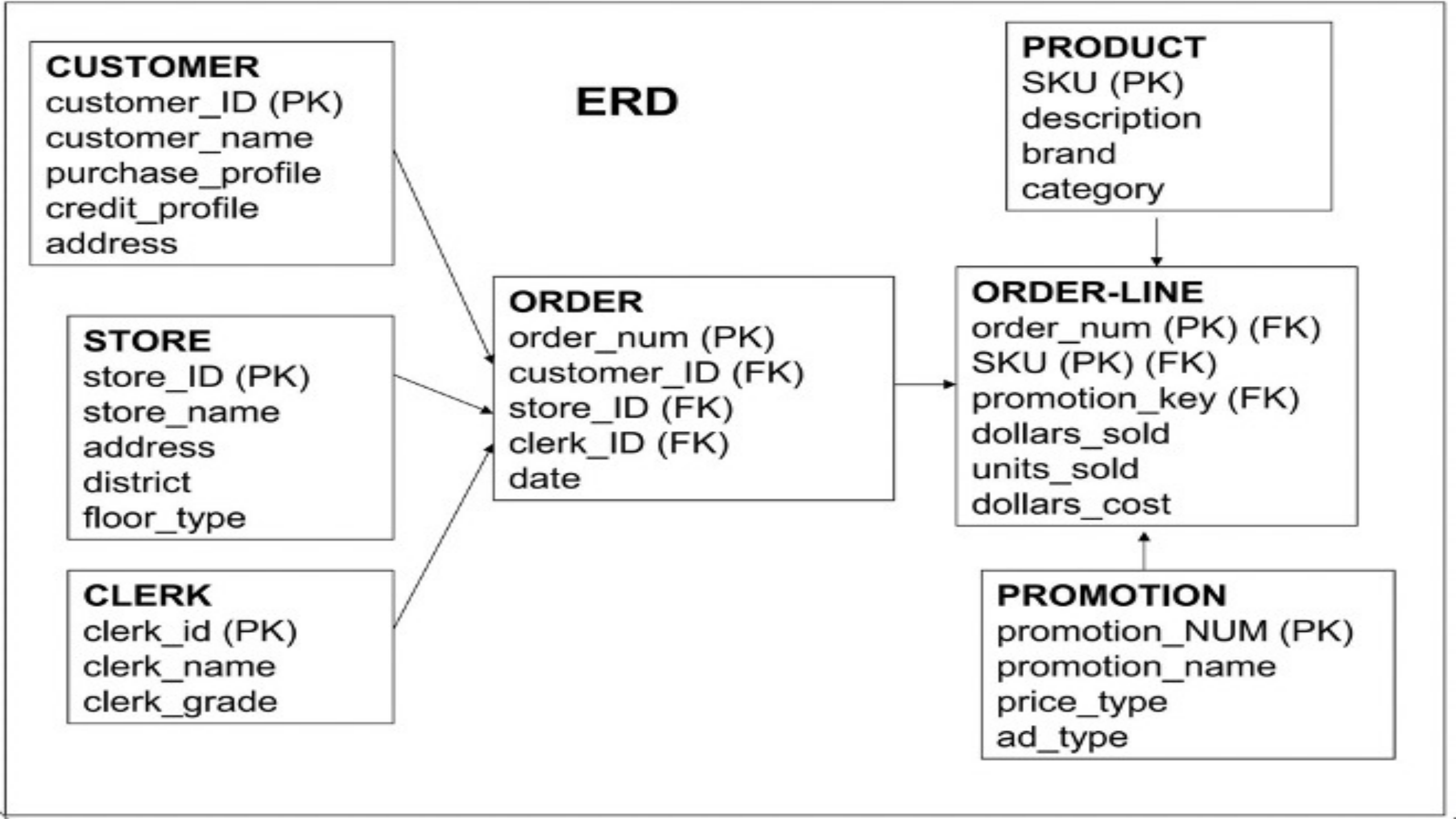
Region and answer the question like number of transactions completed in Asia region etc.

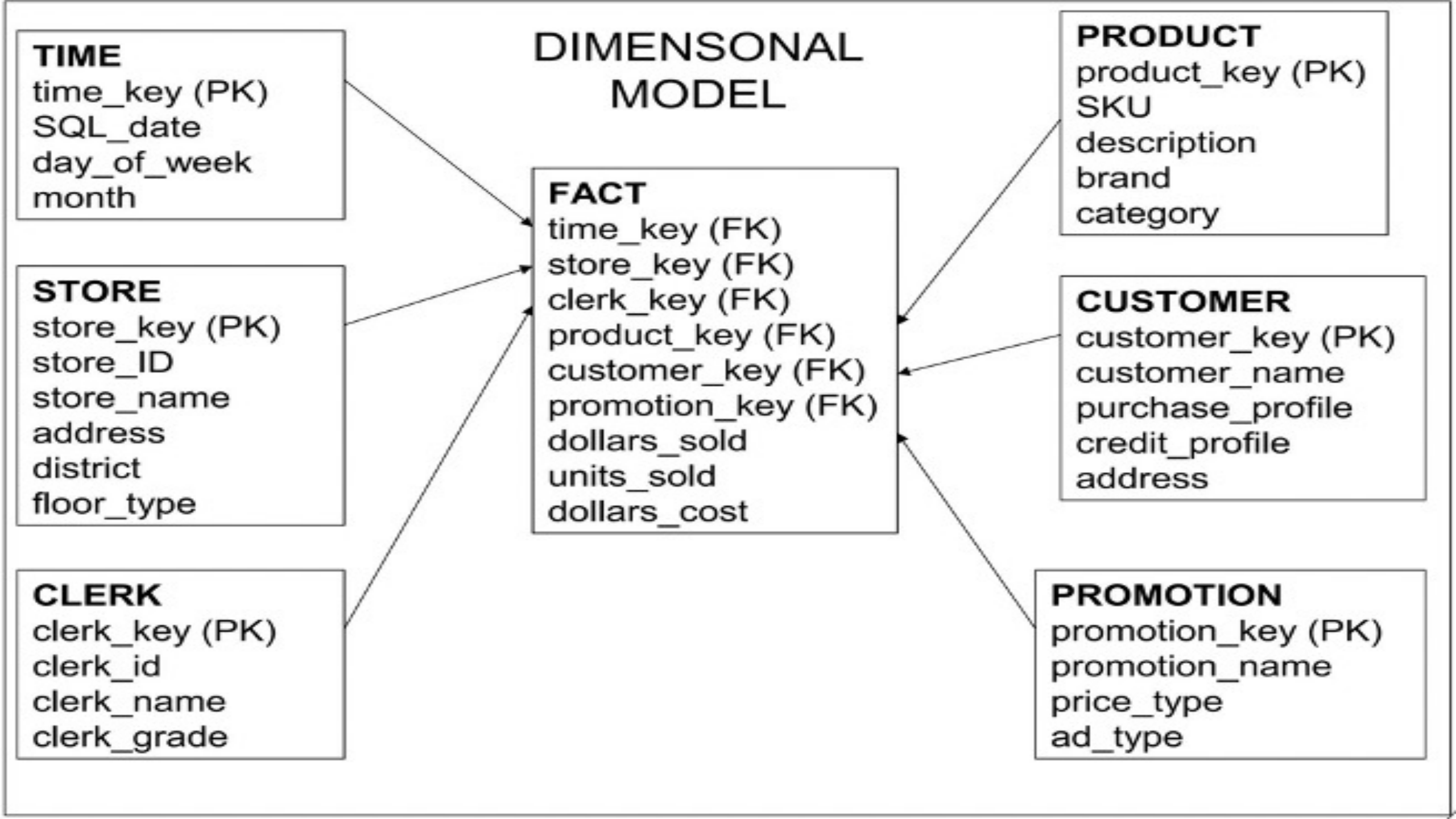
Or Customers who are doing maximum transactions

Or in which Month or Quarter maximum transactions are completed

Dimensions and Facts in Data Warehouse




ERD
CUSTOMER
customer_ID (PK)
customer_name
purchase_profile
credit_profile
address
STORE
store_ID (PK)
store_name
address
district
floor_type
CLERK
clerk_id (PK)
clerk_name
clerk_grade
ORDER
order_num (PK)
customer_ID (FK)
store_ID (FK)
clerk_ID (FK)
date
PRODUCT
SKU (PK)
description
brand
category
ORDER-LINE
order_num (PK) (FK)
SKU (PK) (FK)
promotion_key (FK)
dollars_sold
units_sold
dollars_cost
PROMOTION
promotion_NUM (PK)
promotion_name
price_type
ad_type


DIMENSONAL MODEL
TIME
time_key (PK)
SQL_date
day_of_week
month
STORE
store_key (PK)
store_ID
store_name
address
district
floor_type
CLERK
clerk_key (PK)
clerk_id
clerk_name
clerk_grade
FACT
time_key (FK)
store_key (FK)
clerk_key (FK)
product_key (FK)
customer_key (FK)
promotion_key (FK)
dollars_sold
units_sold
dollars_cost
PRODUCT
product_key (PK)
SKU
description
brand
category
CUSTOMER
customer_key (PK)
customer_name
purchase_profile
credit_profile
address
PROMOTION
promotion_key (PK)
promotion_name
price_type
ad_type

Fact Tables

- A fact table stores quantitative information for analysis and is often denormalized.
- Contains two or more foreign keys.
- Tend to have huge numbers of records.
- Useful facts tend to be numeric and additive.
- A fact table holds the data to be analyzed, and a dimension table stores data about the ways in which the data in the fact table can be analyzed. Thus, the fact table consists of two types of columns. The foreign keys column allows joins with dimension tables, and the measures columns contain the data that is being analyzed.

Example :Fact Table

- Suppose that a company sells products to customers. Every sale is a fact that happens, and the fact table is used to record these facts. For example:
-

Time ID	Product ID	Customer ID	Unit Sold
4	17	2	1
8	21	3	2
8	4	1	1
5	20	2	5
3	4	4	7

Dimension Table

- A dimension table stores attributes, or dimensions, that describe the objects in a fact table.
- A data warehouse organizes descriptive attributes as columns in dimension tables.
- For Example: A customer dimension's attributes could include first and last name, birth date, gender, etc., or a website dimension would include site name and URL attributes.
- A dimension table has a primary key column that uniquely identifies each dimension record (row).

Example:Dimension Table

Customer ID	Name	Gender	Income	Education	Region
1	Brian Edge	M	2	3	4
2	Fred Smith	M	3	5	1
3	Sally Jones	F	1	7	3

- The dimension table is associated with a fact table using this PRIMARY key.
- It is not uncommon for a dimension table to have 50 to 100 attributes;
- Dimension tables tend to have fewer rows than fact tables

- Dimension tables are referenced by fact tables using keys.
- When creating a dimension table in a data warehouse, a system-generated key is used to uniquely identify a row in the dimension. This key is also known as a surrogate key.
- The surrogate key is used as the primary key in the dimension table.
- The surrogate key is placed in the fact table and a foreign key is defined between the two tables. When the data is joined, it does so just as any other join within the database.

- describe the “who, what, where, when, how, and why” associated with the event.

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

Facts and Dimensions

Criteria	Fact Attributes	Dimension Attributes
Purpose	Measurements for reporting or analysis	Constraints or qualifiers for the measurements
Data type	Additive or semi-additive quantitative data	Textual, descriptive
Size	Larger number of records	Smaller number of records
Reporting use	Main report contents	Row or report headers
Examples	Measurements for sales	About time, people, departments, objects, geographic units

SCHEMAS MODEL OF DW

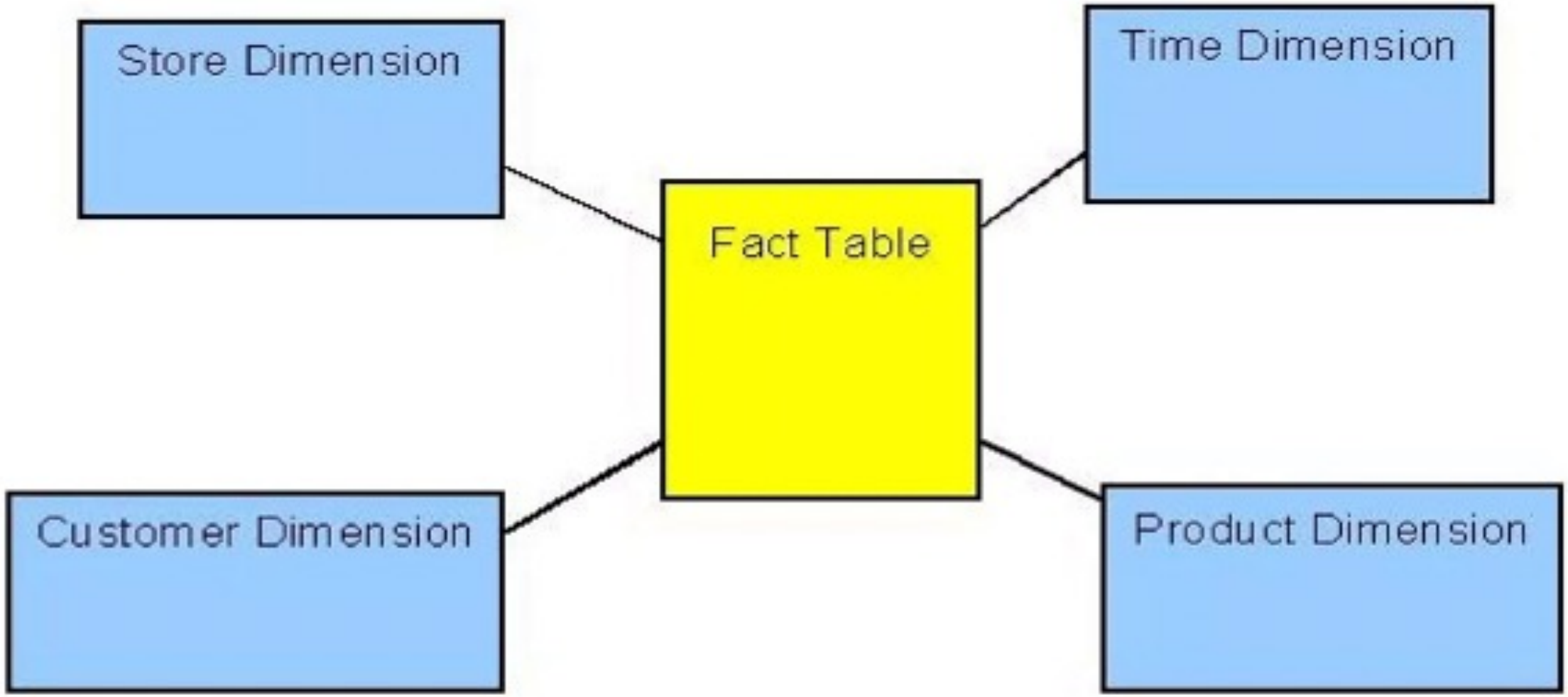
Conceptual Modeling of Data Warehouses

- Star schema
- Snowflake schema
- Fact constellations

Star schema

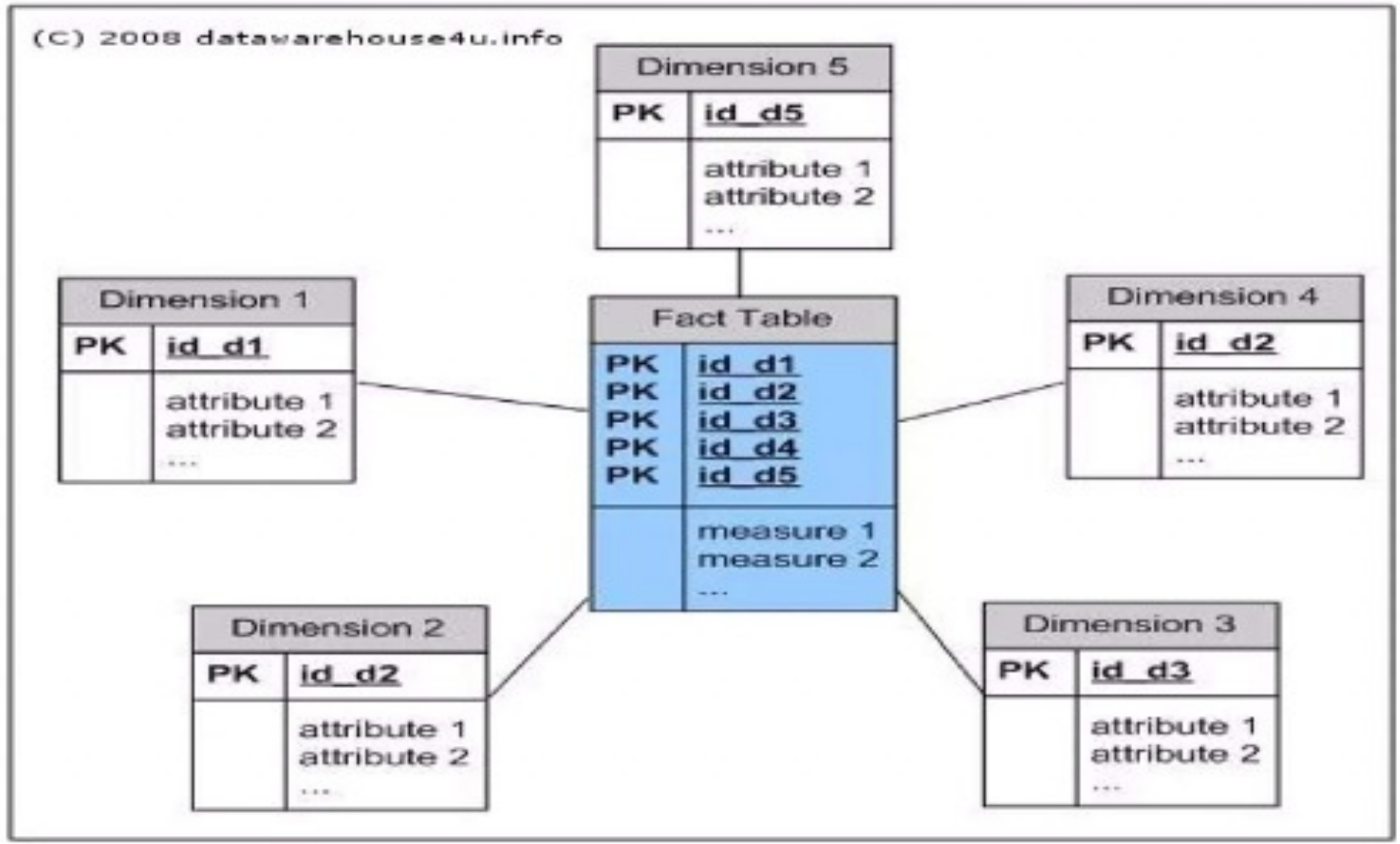
- The star schema architecture is the simplest data warehouse schema.
- It is called a star schema because the diagram resembles a star, with points radiating from a center.
- The center of the star consists of fact table and the points of the star are the dimension tables.
- Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized.

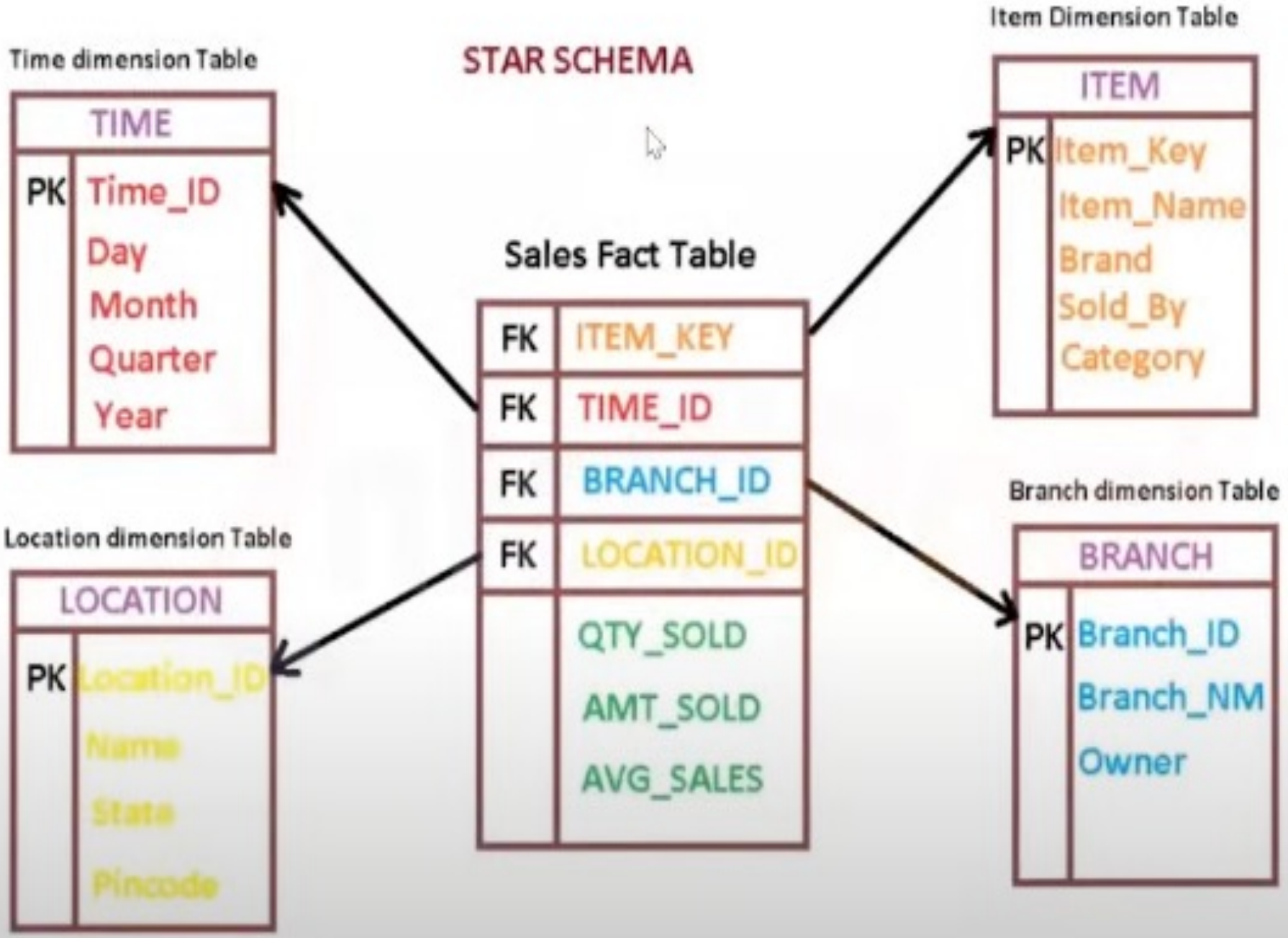
Star schema



http://www.1keydata.com/datawarehousing/star-schema.html

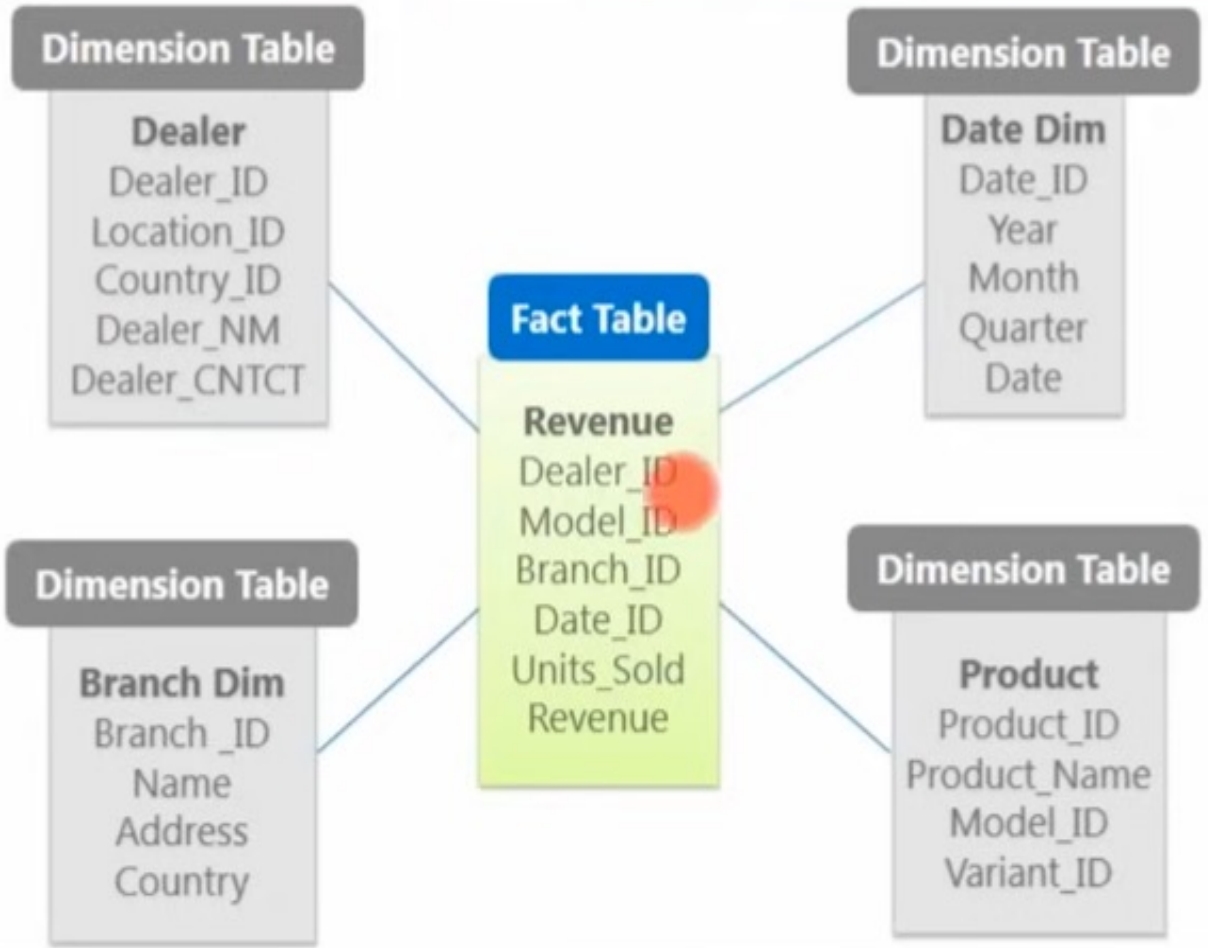
Star schema


(C) 2008 datawarehouse4u.info
Dimension 5
PK
id_d5
attribute 1
attribute 2
...
Dimension 1
PK
id_d1
attribute 1
attribute 2
...
Fact Table
PK
id_d1
PK
id_d2
PK
id_d3
PK
id_d4
PK
id_d5
measure 1
measure 2
...
Dimension 4
PK
id_d2
attribute 1
attribute 2
...
Dimension 2
PK
id_d2
attribute 1
attribute 2
...
Dimension 3
PK
id_d3
attribute 1
attribute 2
...


STAR SCHEMA
Time dimension Table
TIME
PK
Time_ID
Day
Month
Quarter
Year
Item Dimension Table
ITEM
PK
Item_Key
Item_Name
Brand
Sold_By
Category
Sales Fact Table
FK
ITEM_KEY
FK
TIME_ID
FK
BRANCH_ID
FK
LOCATION_ID
QTY_SOLD
AMT_SOLD
AVG_SALES
Location dimension Table
LOCATION
PK
Location_ID
Name
State
Pincode
Branch dimension Table
BRANCH
PK
Branch_ID
Branch_NM
Owner

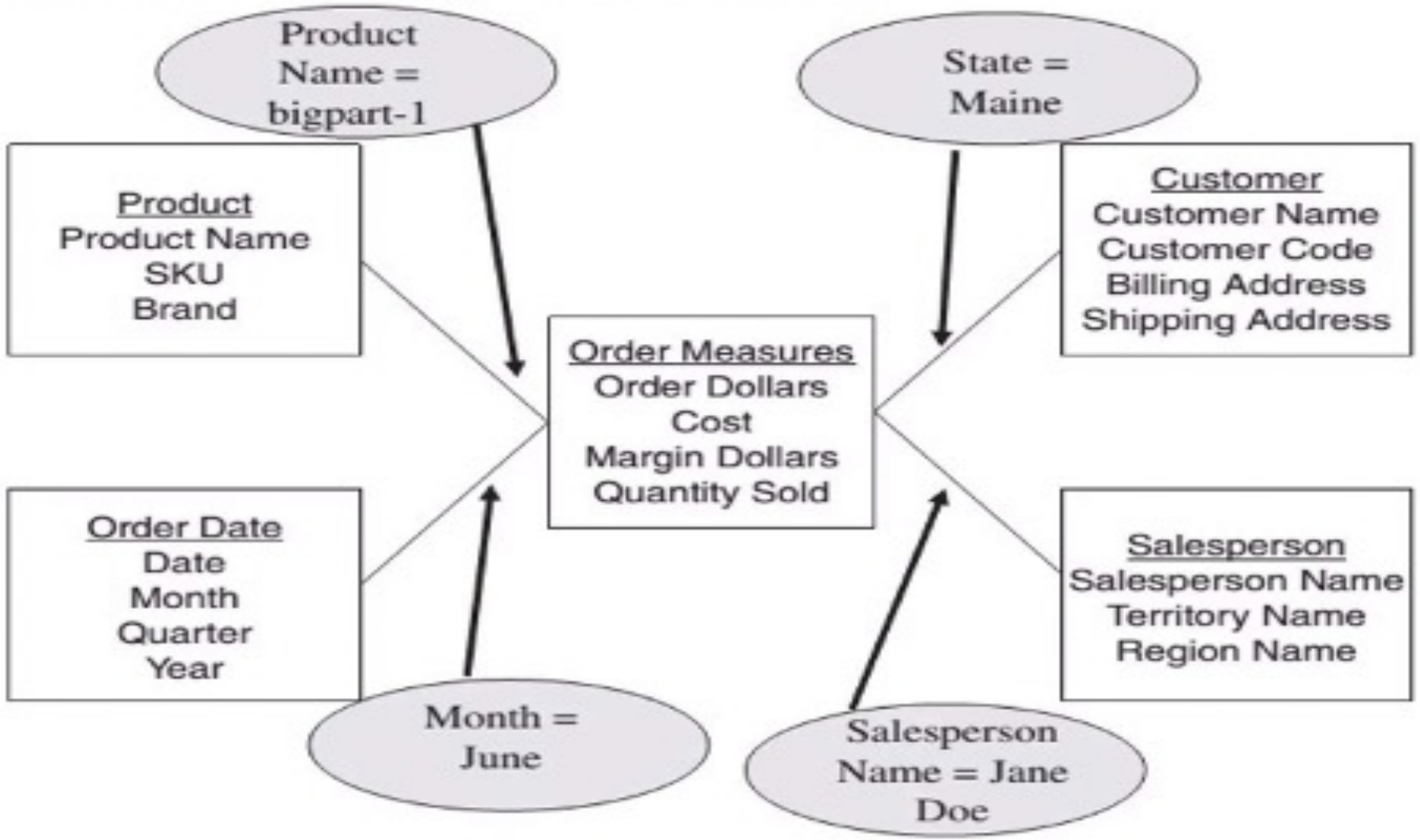
Types Of Schemas:- Star Schema

- Each dimension in a star schema is represented with a **one-dimension table** which contains a set of attributes.
- **Fact table** is at the center. which contains keys to every dimension table & attributes like: *units sold* and *revenue*.



Querying Star Schema

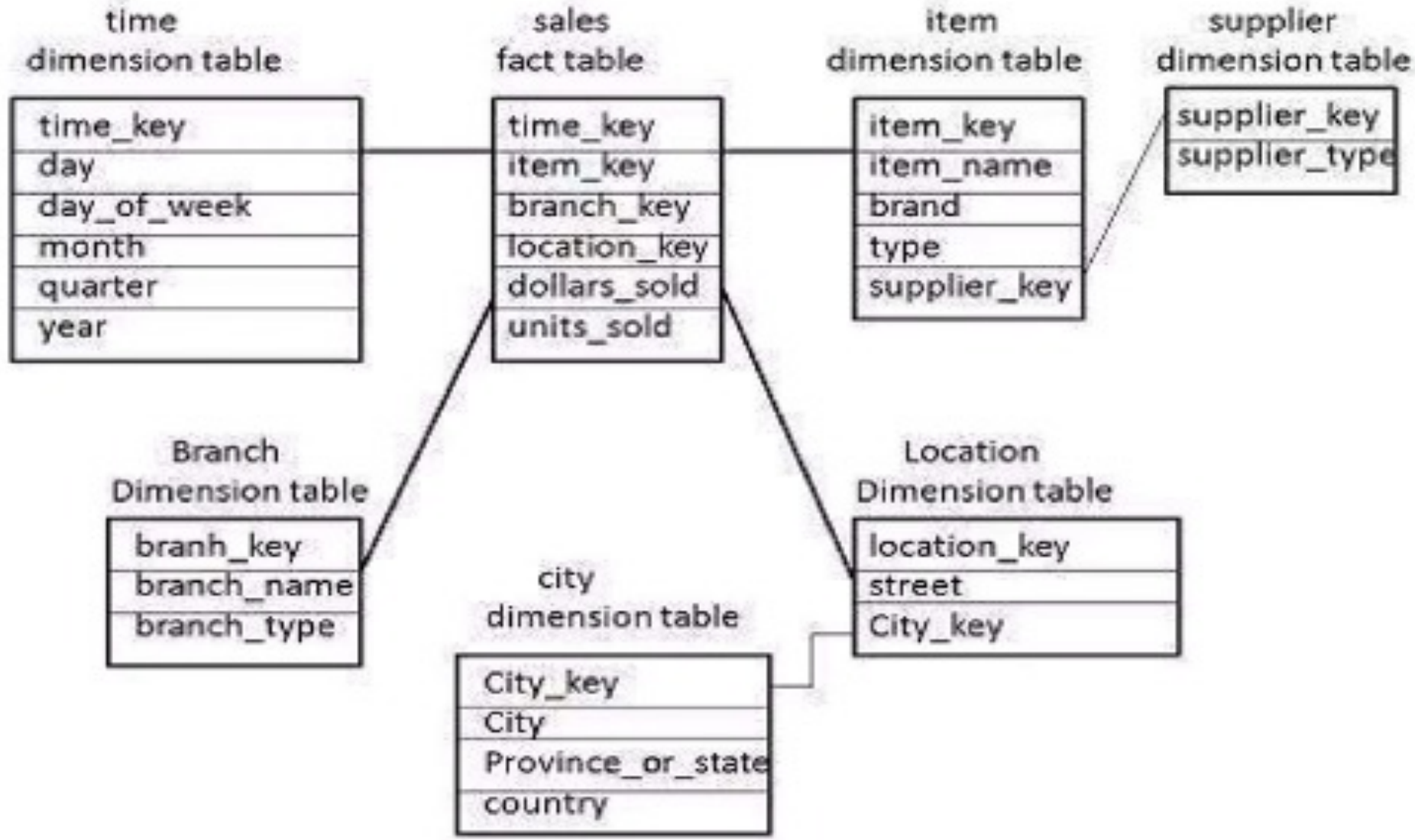
If marketing department wants the quantity sold and order dollars for productbigpart-1, relating to customers in the state ofMaine, obtained by salesperson Jane Doe, during the month of June.



Snowflake Schema

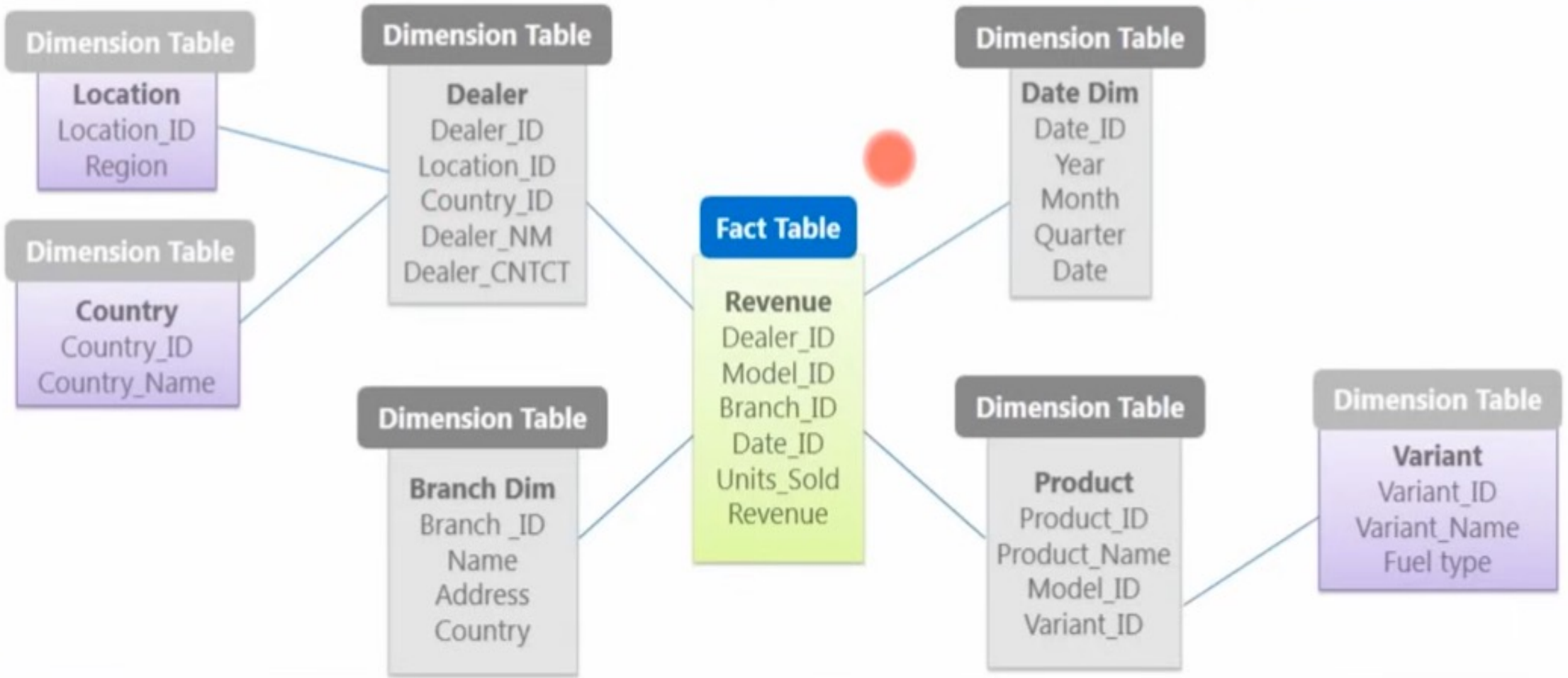
- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

Snowflake Schema



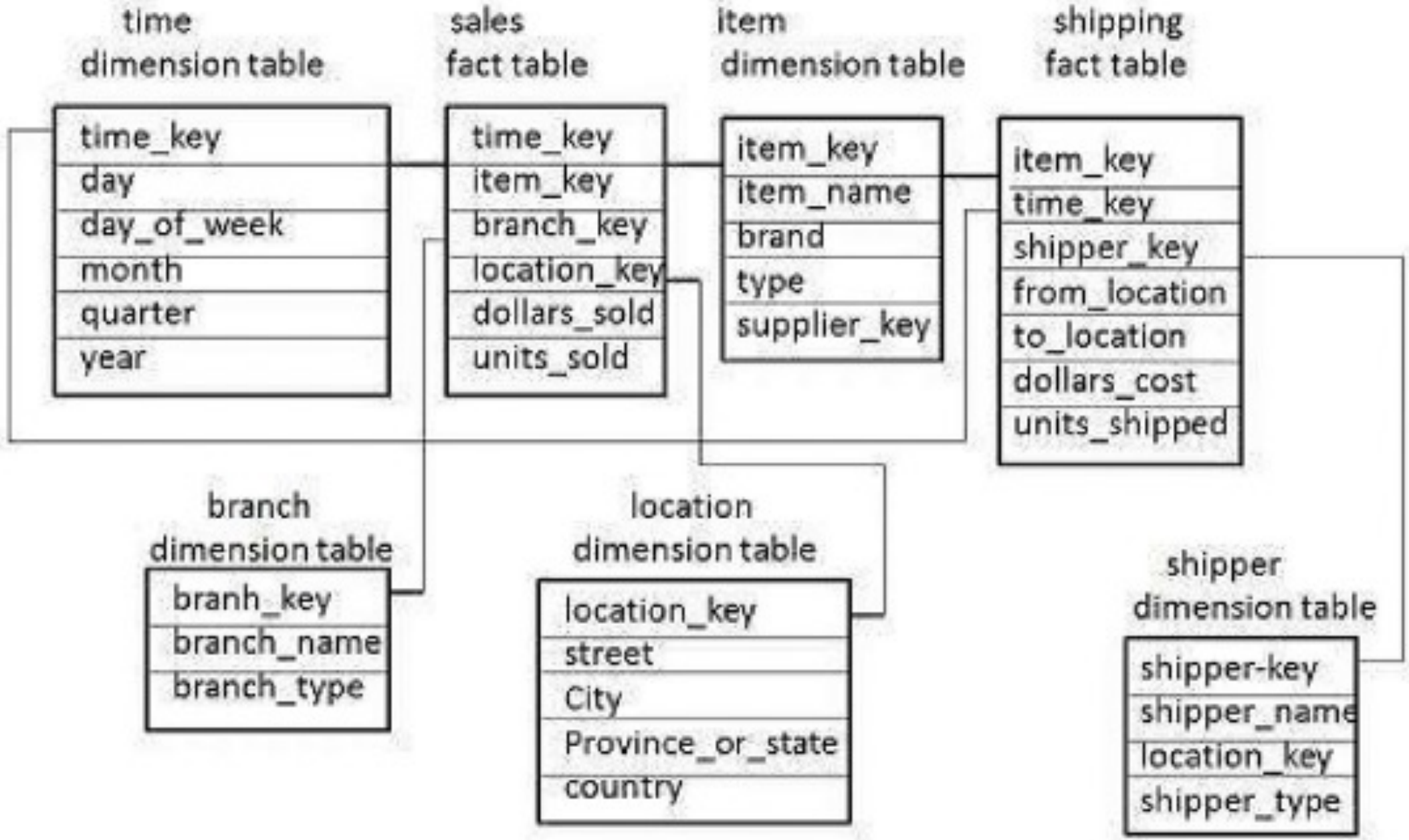
Types Of Schemas:- Snowflake Schema

- Dimension tables in the **Snowflake schema** are **normalized**. (Split into additional tables).
- **Dealer** dimension table is split into **Location** & **Country**. Product dimension table is split into **Product** & **Variant**.



Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.


time
dimension table
time_key
day
day_of_week
month
quarter
year
sales
fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
dimension table
item_key
item_name
brand
type
supplier_key
shipping
fact table
item_key
time_key
shipper_key
from_location
to_location
dollars_cost
units_shipped
branch
dimension table
branh_key
branch_name
branch_type
location
dimension table
location_key
street
City
Province_or_state
country
shipper
dimension table
shipper-key
shipper_name
location_key
shipper_type


thank you