D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Subject Name: DATA WAREHOUSING AND MINING

Unit No:1

Unit Name: INTRODUCTION TO DATA WAREHOUSING

Faculty Name : Mrs.Bhavana Chaudhari

Index

Unit :1 Unit Name: Introduction to Data Warehousing

Lecture No: 1

Introduction to Data Warehouse, Data warehouse architecture

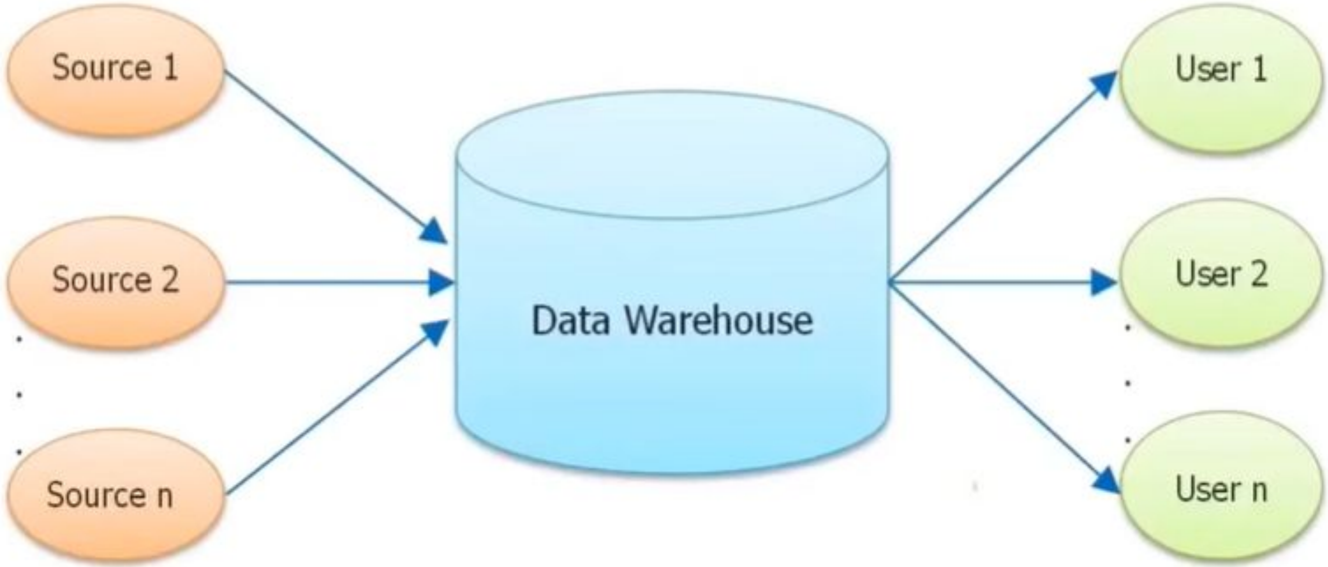
What is Data Warehouse ?

→A Data Warehouse is a central location where consolidated data from multiple locations are stored

→The end user accesses it whenever he needs some information

→Data Warehouse is not loaded every time when new data is generated

→There are timelines determined by the business as to when a Data Warehouse needs to be loaded – daily, monthly, once in a quarter etc



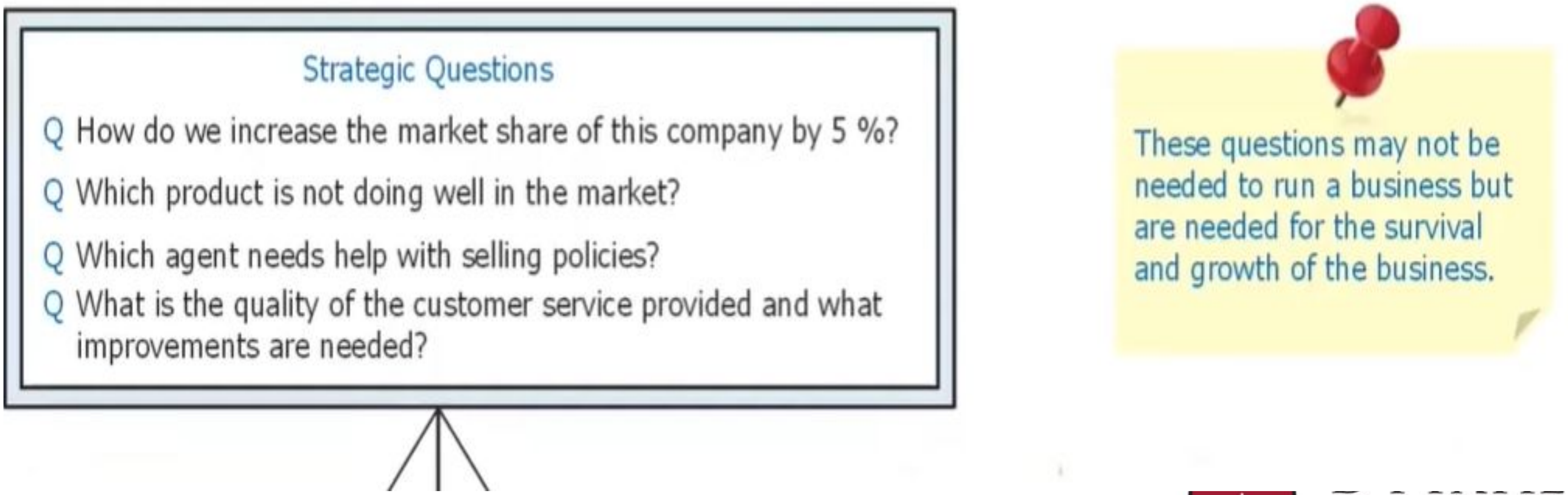
Why do we need Data Warehouse ?

→The primary reason for a Datawarehouse is, for a company to get that extra edge over its competitors

→This extra edge can be gained by taking smarter decisions

→Smarter decisions can be taken only if the executives responsible for taking such decisions have data at their disposal

→ For Example: Let's consider some strategic questions that a manager or an executive has to answer to get an extra edge over his company's competitors



Strategic Questions

Q How do we increase the market share of this company by 5 %?

Q Which product is not doing well in the market?

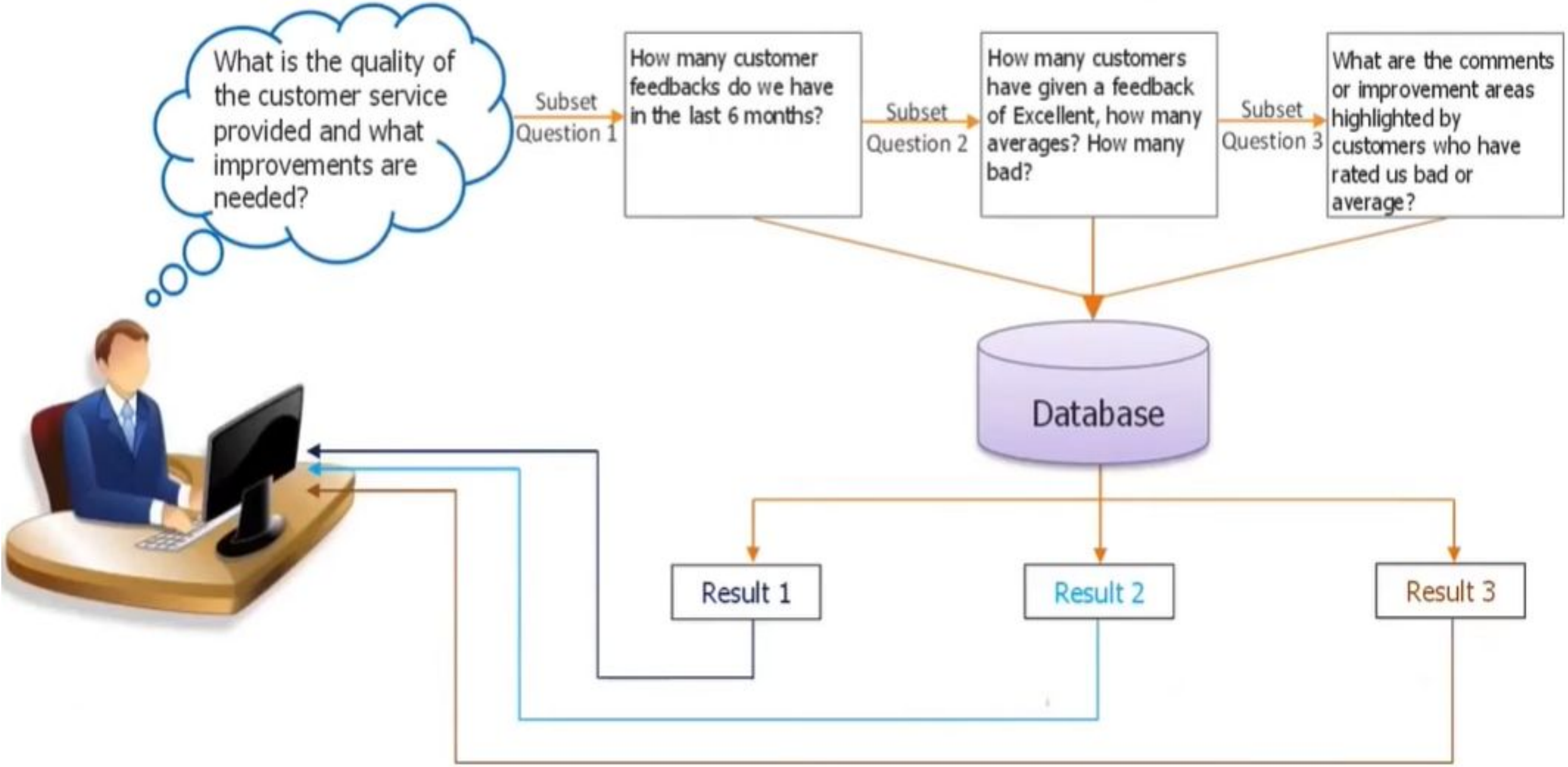
Q Which agent needs help with selling policies?

Q What is the quality of the customer service provided and what improvements are needed?

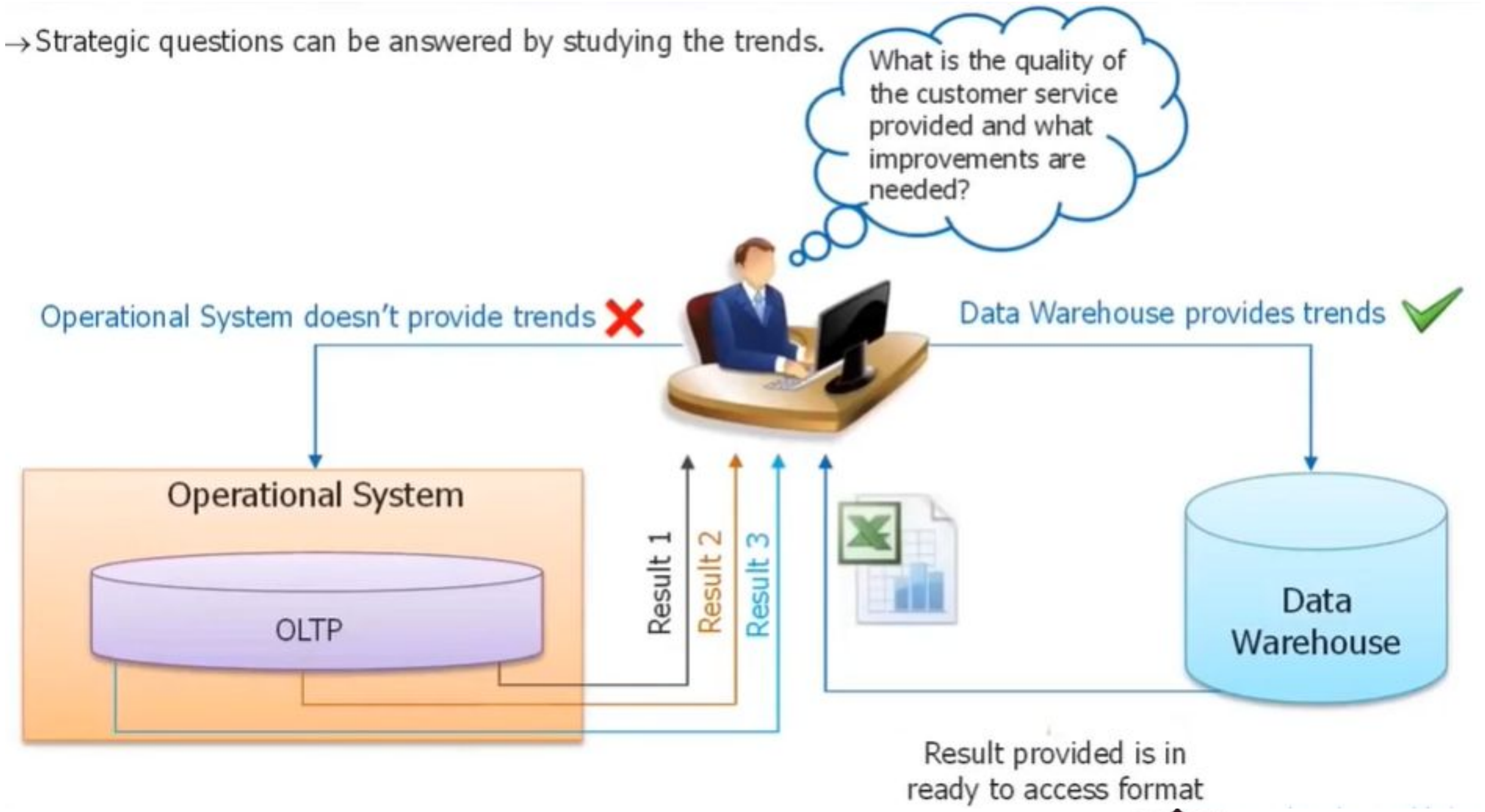
These questions may not be needed to run a business but are needed for the survival and growth of the business.

Why is Data Warehouse so important?

→ Let's consider one of the strategic question for which a manager or an executive is trying to find answer


What is the quality of the customer service provided and what improvements are needed?
Subset Question 1
How many customer feedbacks do we have in the last 6 months?
Subset Question 2
How many customers have given a feedback of Excellent, how many averages? How many bad?
Subset Question 3
What are the comments or improvement areas highlighted by customers who have rated us bad or average?
Database
Result 1
Result 2
Result 3

Why is Data Warehouse so important? Cont..



→Strategic questions can be answered by studying the trends.

What is the quality of the customer service provided and what improvements are needed?
Operational System doesn't provide trends ✗
Data Warehouse provides trends ✓
Operational System
OLTP
Result 1
Result 2
Result 3
Data Warehouse
Result provided is in ready to access format

Functional definition of Data Warehouse

The data warehouse is an informational environment that:

- Provides an integrated and total view of the enterprise
- Makes the enterprise’s current and historical information easily available for decision making
- Makes decision-support transactions possible without hindering operational systems
- Renders the organization’s information consistent
- Presents a flexible and interactive source of strategic information

Bill Inmon, considered to be the father of Data Warehousing provides the following definition:

- “A Data Warehouse is a **subject oriented**, **integrated**, **nonvolatile**, and **time variant** collection of data in support of management’s decisions.”

Features of Data Warehousing

- Sean Kelly, data warehouse practitioner, defines the data warehousing in following way. The data in data warehousing is:
  - Subject oriented
  - Integrated
  - Time stamped
  - Non-volatile

Features of Data Warehousing – subject oriented data

In the data warehouse, data is not stored by operational applications, but by business subjects.

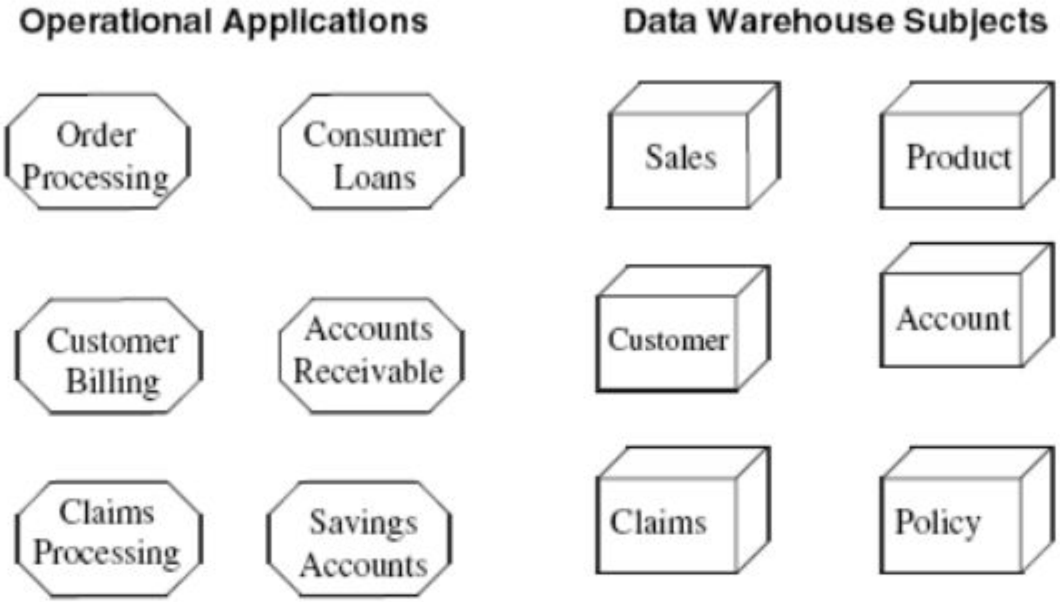


Figure 2-1 The data warehouse is subject oriented.

Features of Data Warehousing – integrated data

Data inconsistencies are removed; data from diverse operational applications is integrated.

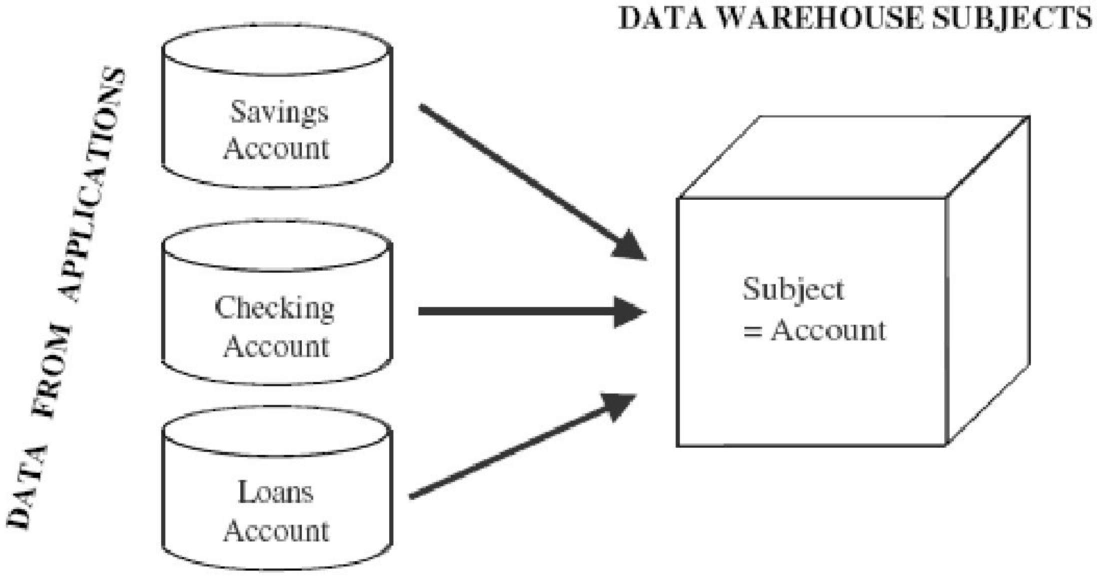


Figure 2-2 The data warehouse is integrated.

Features of Data Warehousing – integrated data

- Before the data from various disparate sources can be usefully stored in a data warehouse, you have to:
    - remove the inconsistencies;
    - standardize the various data elements;
    - make sure of the meanings of data names in each source application

- Before moving the data into the data warehouse, you have to go through a process of transformation, consolidation, and integration of the source data

- Here are some of the items that would need standardization:
    - Naming conventions
    - Codes
    - Data attributes
    - Measurements

Features of Data Warehousing – Time Variant

- For an operational system, the stored data contains the current values.

- The data in the data warehouse is meant for analysis and decision making.

- A data warehouse, because of the vary nature of its purpose, has to contain historical data, not just current values
  - Data is stored as snapshots over past and current periods
  - Every data structure in the data warehouse contains the time element

- The time variant nature of data in a data warehouse
  - Allows for analysis of the past
  - Relates information to the present
  - Enables forecast for the future

Features of Data Warehousing – non volatile data

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

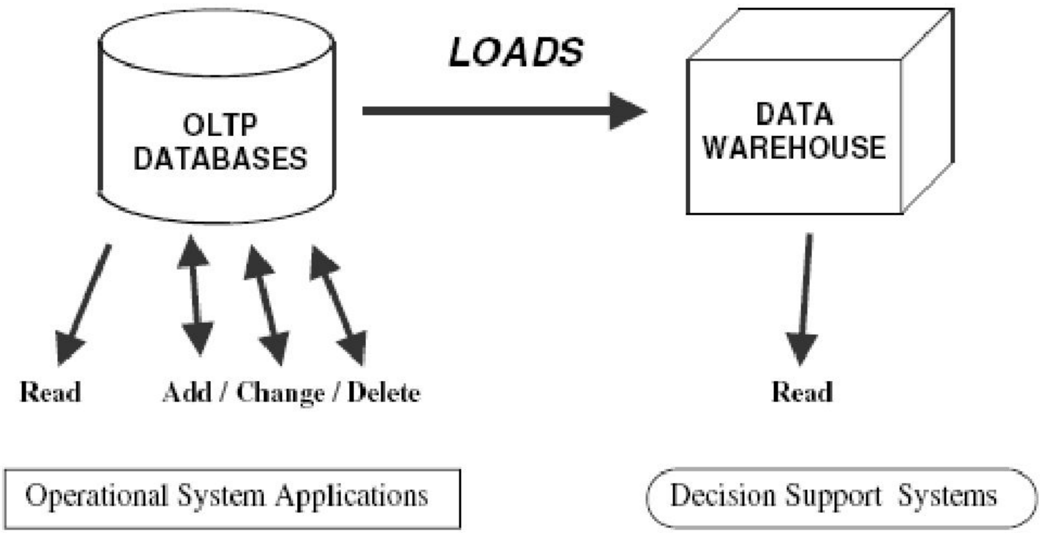
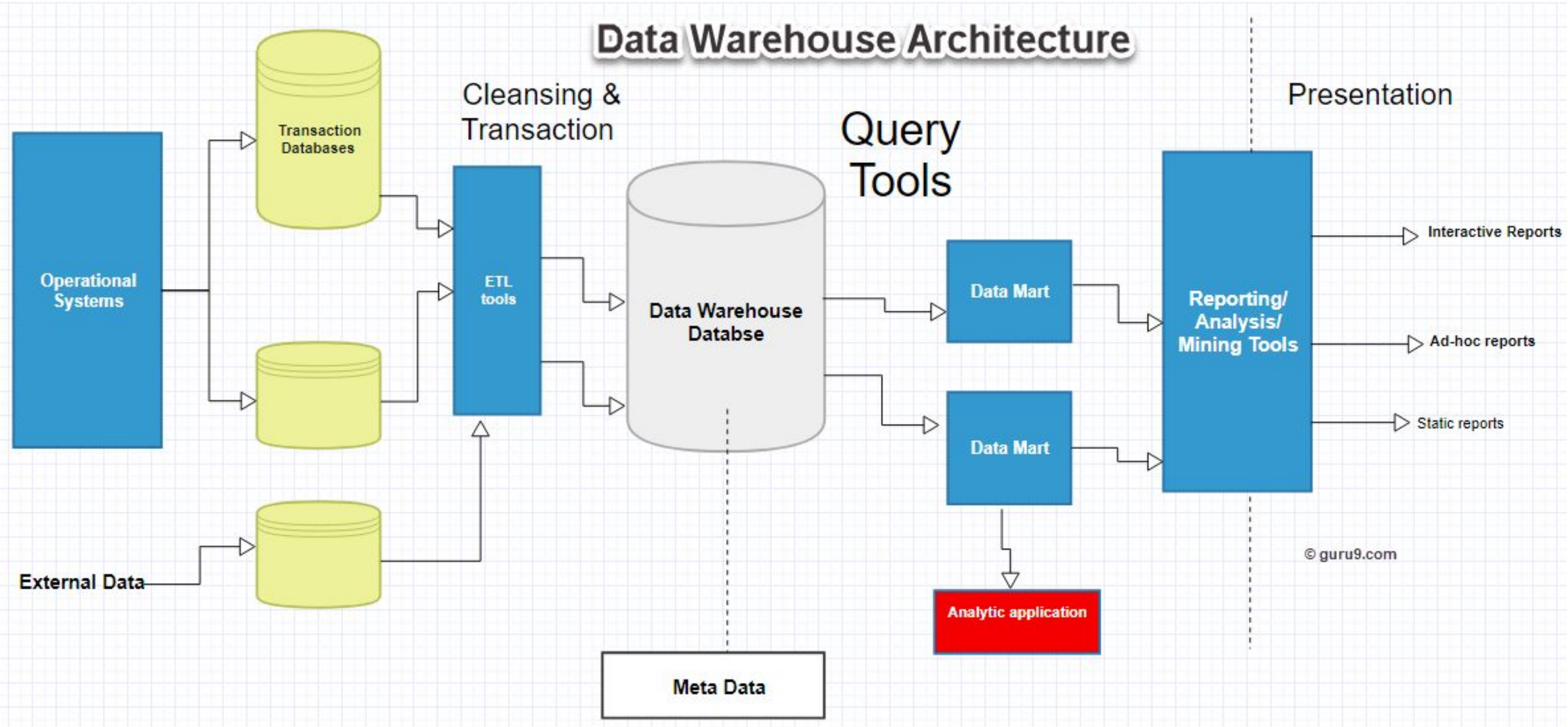


Figure 2-3 The data warehouse is nonvolatile.

Data Warehouse Architecture



Data Warehouse Architecture

There are 3 approaches for constructing Data Warehouse layers:

Single-tier architecture

The objective of a single layer is to minimize the amount of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice.

Two-tier architecture

It separates physically available sources and data warehouse.

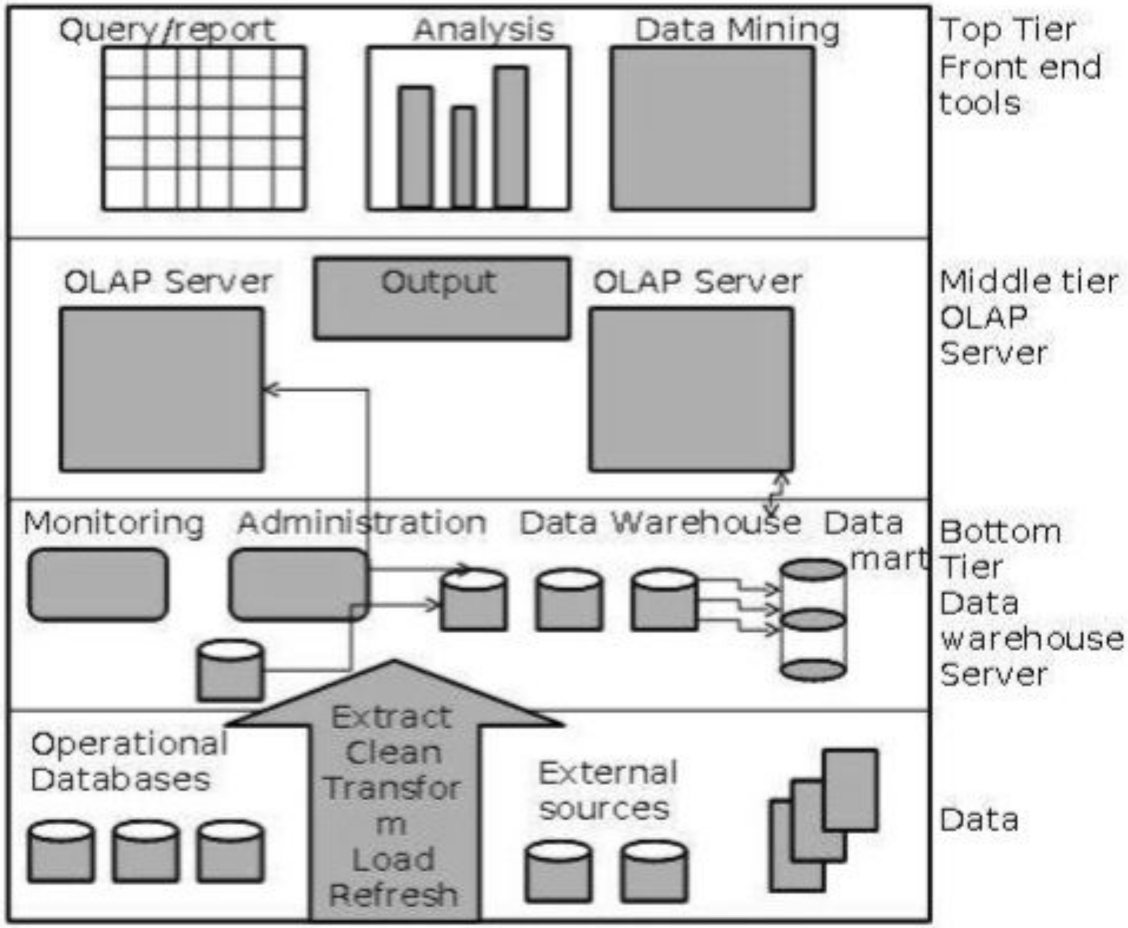
Not expandable and also not supporting a large number of end-users.

It also has connectivity problems because of network limitations.

Three-Tier Data Warehouse Architecture

This is the most widely used Architecture of Data Warehouse.

3 Tier Data Warehouse Architecture



3 Tier Data Warehouse Architecture

Generally a data warehouses adopts a three-tier architecture.

Bottom Tier − The bottom tier of the architecture is the **data warehouse database server.** It is the **relational database system**. We use the **back end tools** and utilities to feed data into the bottom tier. These back end tools and utilities perform the **Extract, Clean, Load, and refresh functions**.

Middle Tier − In the middle tier, we have the **OLAP Server** that can be implemented in either of the following ways.

- By **Relational OLAP (ROLAP)**, which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
- By **Multidimensional OLAP (MOLAP)** model, which directly implements the multidimensional data and operations.

Top-Tier − This tier is the **front-end client layer**. This layer holds the **query tools, reporting tools, analysis tools and data mining tools**.

Unit :1 Unit Name: Introduction to Data Warehousing

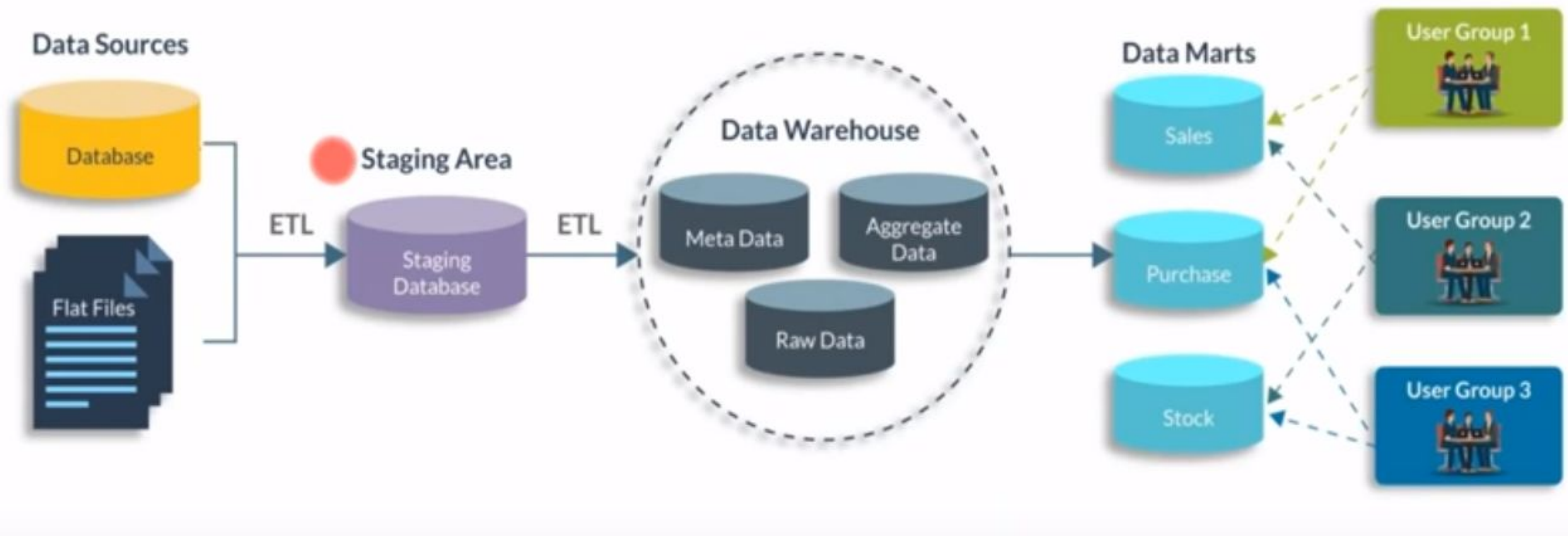
Lecture No: 2

Data warehouse versus Data Marts, E-R Modelling versus Dimensional Modelling

Data Mart

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group. For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.



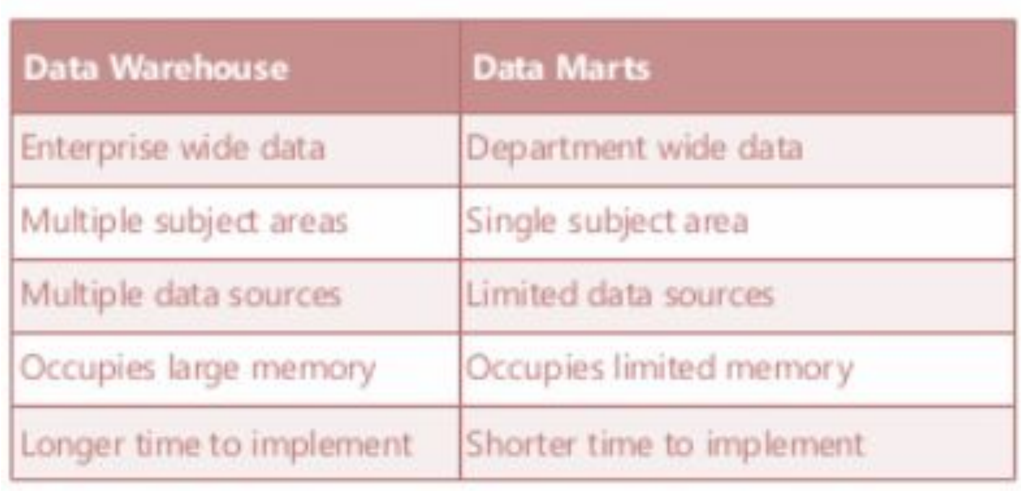
Data Mart

Points to remember about data marts –

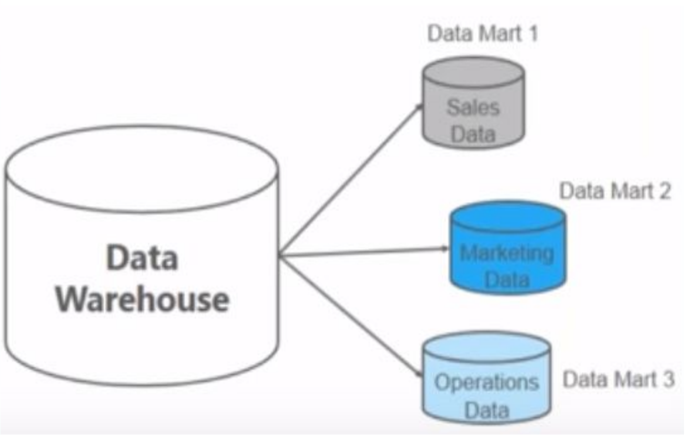
- Unix/Linux-based servers are used to implement data marts.
- They are implemented on low-cost servers.
- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data mart are flexible.

Data warehouse vs Data Mart

- Data Mart is smaller version of Data Warehouse which deals with single subject
- Data marts are focused on one area, hence they draw data from limited number of sources
- Time taken to build data mart is very less compared to DWH



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

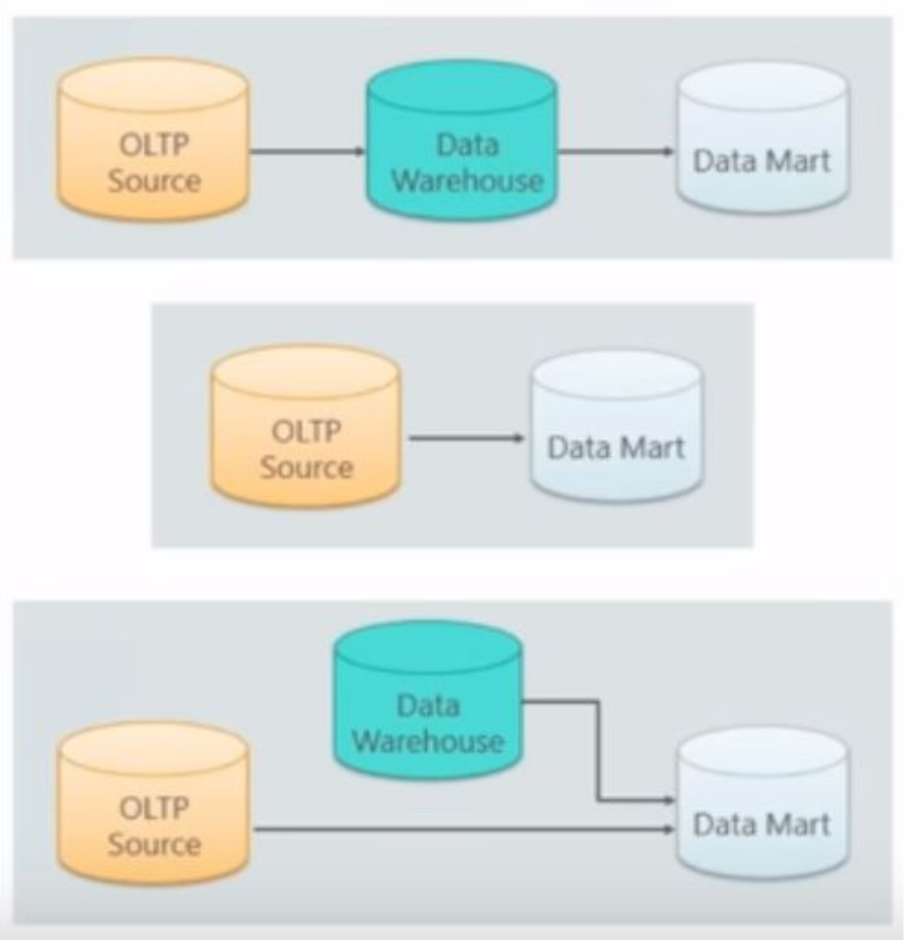


Types of Data Mart

- **Dependent Data Mart:** Data comes from OLTP source to Data Warehouse and then from data warehouse to Data Mart

- **Independent Data Mart:** Data directly received from the source system, This is suitable for small organization

- **Hybrid Data Mart:** Data fed from both OLTP source and DWH



Data Warehouse Design Approaches:Top-Down and Bottom-Up

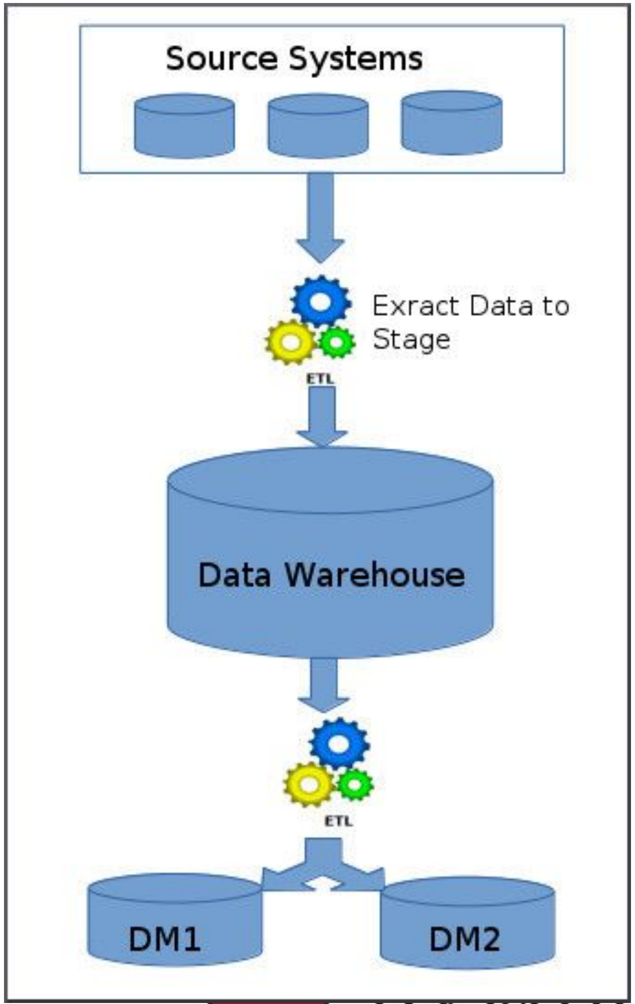
- Data Warehouse design approaches are very important aspect of building data warehouse.

- Selection of right data warehouse design could save lot of time and project cost.

- There are two different Data Warehouse Design Approaches normally followed when designing a Data Warehouse solution and based on the requirements of your project you can choose which one suits your particular scenario.

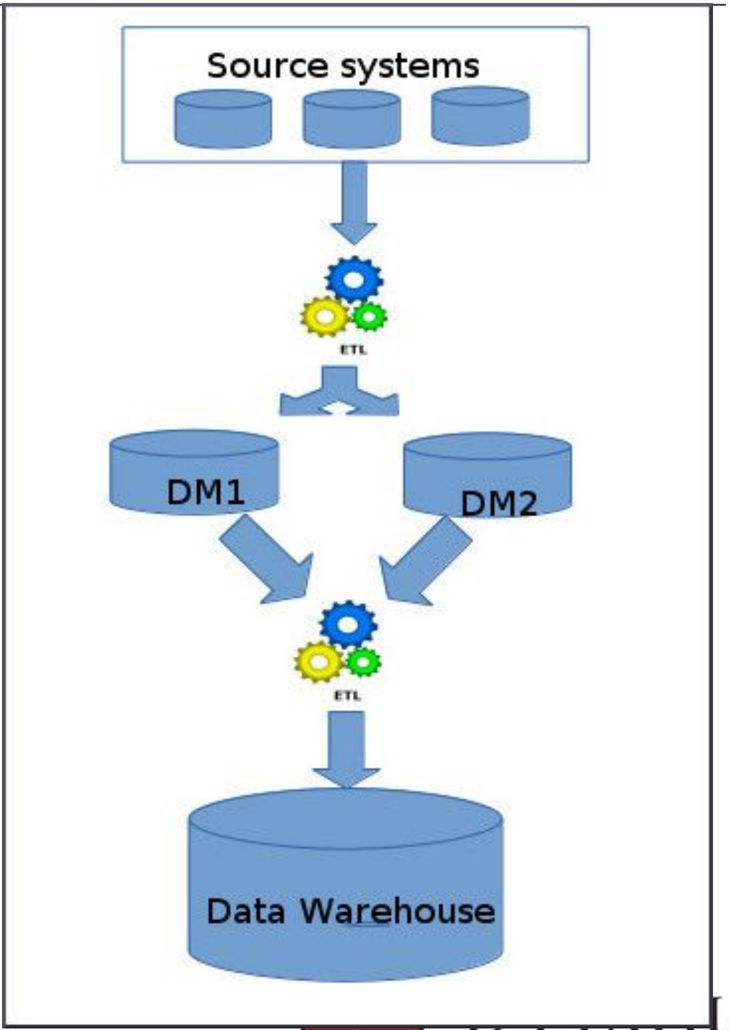
Top-Down Approach for Data warehouse Design

- In the top-down approach, the data warehouse is designed first and then data mart are built
- Below are the steps that are involved in top-down approach:
- Data is extracted from the various source systems using ETL tools, it is validated and pushed to the data warehouse.
- You will apply various aggregation, summerization techniques on extracted data from data warehouse and loaded back to the data warehouse
- Once the aggregation and summerization is completed, various data marts extract that data and apply the some more transformation to make the data structure as defined by the data marts.
- This is bill inmons methodology



Bottom-up Approach for Data warehouse Design

- Ralph Kimball proposed data warehouse design approach is called dimensional modelling or the Kimball methodology.
- This methodology follows the bottom-up approach
- As per this method, data marts are first created to provide the reporting and analytics capability for specific business process
- Later with these data marts, enterprise data warehouse is created



Dimensional Modelling

What is Dimensional Model?

- A dimensional model **is a data structure technique optimized for Data warehousing** tools.
- The concept of Dimensional Modelling was **developed by Ralph Kimball** and is comprised of "**fact**" and "**dimension**" tables.
- A Dimensional model is **designed to read, summarize, analyze numeric information** like values, balances, counts, weights, etc. in a data warehouse.
- In contrast, **relational models are optimized for addition, updating and deletion of data** in a real-time Online Transaction System.
- ER modeling is for reducing redundancy of data, where as dimensional model arranges data in such a way that it is easier to retrieve information and generate reports
- These dimensional and relational models have their unique way of data storage that has specific advantages.
- Dimensional models are used in data warehouse systems and not a good fit for relational systems

Elements of Dimensional Data Model

Fact

Facts are the measurements/metrics or facts from your business process. For a Sales business process, a measurement would be quarterly sales number

Dimension

Dimension provides the context surrounding a business process event. In simple terms, they give who, what, where of a fact. In the Sales business process, for the fact quarterly sales number, dimensions would be

Who – Customer Names

Where – Location

What – Product Name

In other words, a dimension is a window to view information in the facts.

Elements of Dimensional Data Model

Attributes

The Attributes are the various characteristics of the dimension in dimensional data modeling.

In the Location dimension, the attributes can be
State

Country

Zipcode etc.

Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes

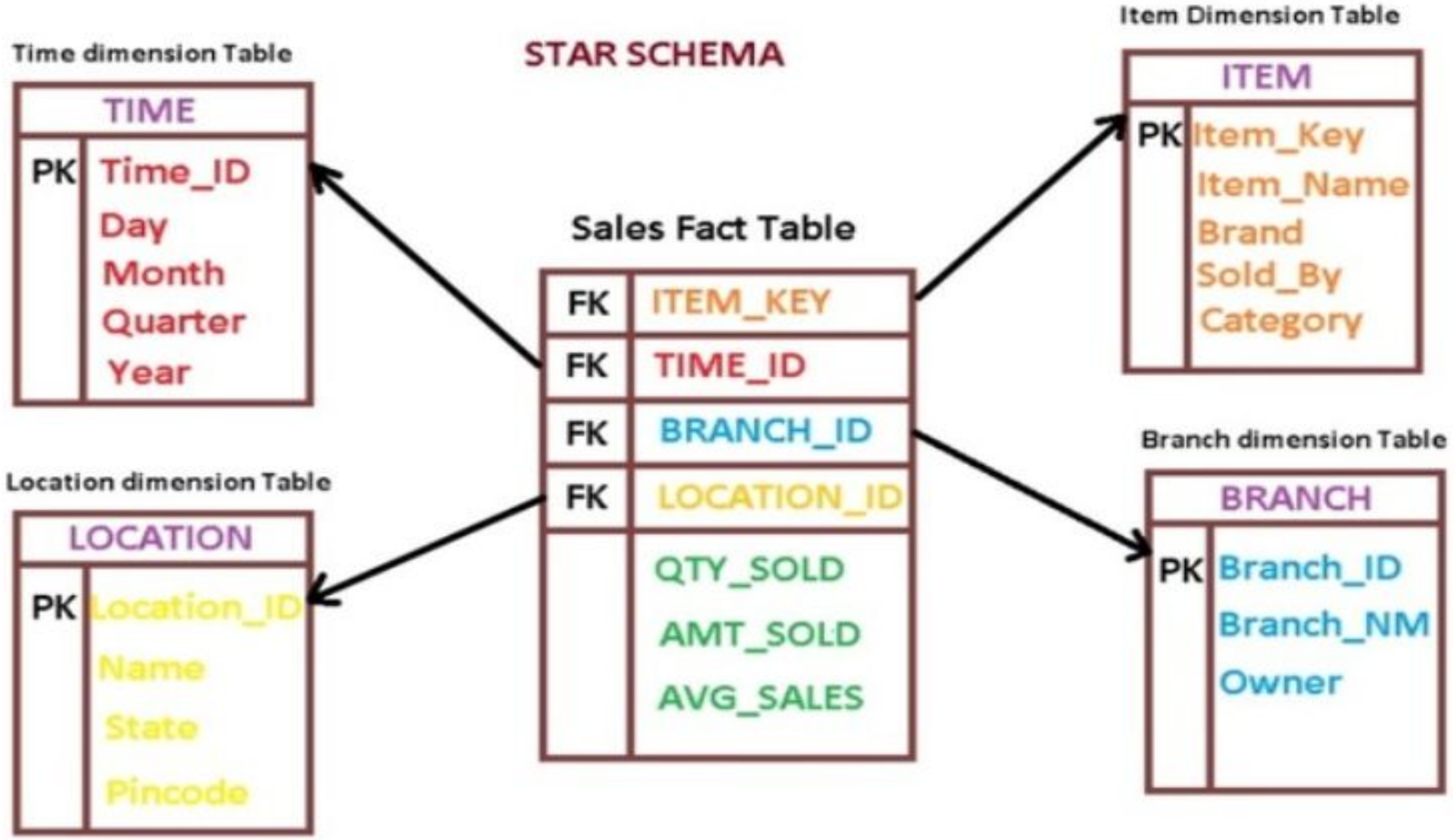
What is Fact Table?

- A Fact table stores quantified data to measure the business performance.
- It is a measure that can be summed, averaged or manipulated.
- Fact table is a table surrounded by the dimension tables in the Star Schema of Data Warehouse.

The Fact table consists of two types of column:

- **A Dimension key (foreign key)** – A foreign key that joins with dimension tables
- **A Measure** – where data is analyzed
- A dimension table is a table in a star schema of a data warehouse.
- A dimension table stores attributes, or dimensions, that describe the objects in a fact table.

Fact and Dimensional Table



STAR SCHEMA

Time dimension Table

TIME	
PK	Time_ID
	Day
	Month
	Quarter
	Year

Sales Fact Table

FK	ITEM_KEY
FK	TIME_ID
FK	BRANCH_ID
FK	LOCATION_ID
	QTY_SOLD AMT_SOLD AVG_SALES

Item Dimension Table

ITEM	
PK	Item_Key
	Item_Name
	Brand
	Sold_By
	Category

Location dimension Table

LOCATION	
PK	Location_ID
	Name
	State
	Pincode

Branch dimension Table

BRANCH	
PK	Branch_ID
	Branch_NM
	Owner

ER Modelling vs Dimensional Modelling

ER Modeling	Dimensional Modeling
Data Stored in RDBMS	Data Stored in RDBMS or Multidimensional databases
Tables are unit of storage	Cubes are the unit of storage
Data is normalized and used for OLTP	Data is de normalized and used for data warehouse and data marts
Several tables and chain of relationship between them	Few facts tables are connected to several dimension tables
Volatile(frequent updates)	Non volatile
Time variant	Time invariant
Detailed level of transaction data	Summary of bulky transaction data (Aggregations and measures) are used in business decisions
SQL is used to manipulate the data	SQL or MDX are used to manipulate the data
Normal reports	Interactive reports, user friendly, drag and drop MD OLAP reports

Unit :1 Unit Name: Introduction to Data Warehousing

Lecture No: 3
Information Package Diagram, Data Warehouse Schemas; Star Schema

DEFINING THE BUSINESS REQUIREMENTS

- In several ways, building a data warehouse is very different from building an operational system.

- This becomes notable especially in the requirements gathering phase.

- Because of this difference, the traditional methods of collecting requirements that work well for operational systems cannot be applied to data warehouses.

Dimensional Nature of Business Data

- In data warehousing system, the users are generally unable to define their requirements clearly.
- Users cannot define precisely what information they really want from the data warehouse, nor can they express how they would like to use the information or process it.
- Managers think of the business in terms of business dimensions.
- If your users of the data warehouse think in terms of business dimensions for decision making, you should also think of business dimensions while collecting requirements.

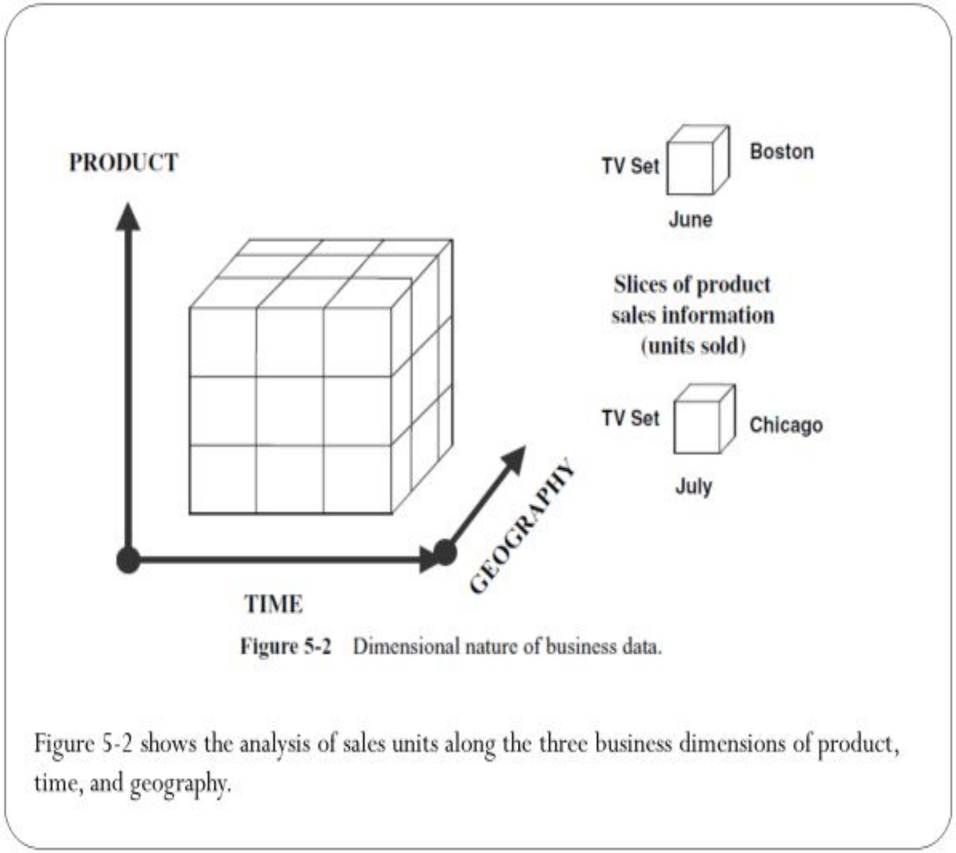


Figure 5-2 Dimensional nature of business data.

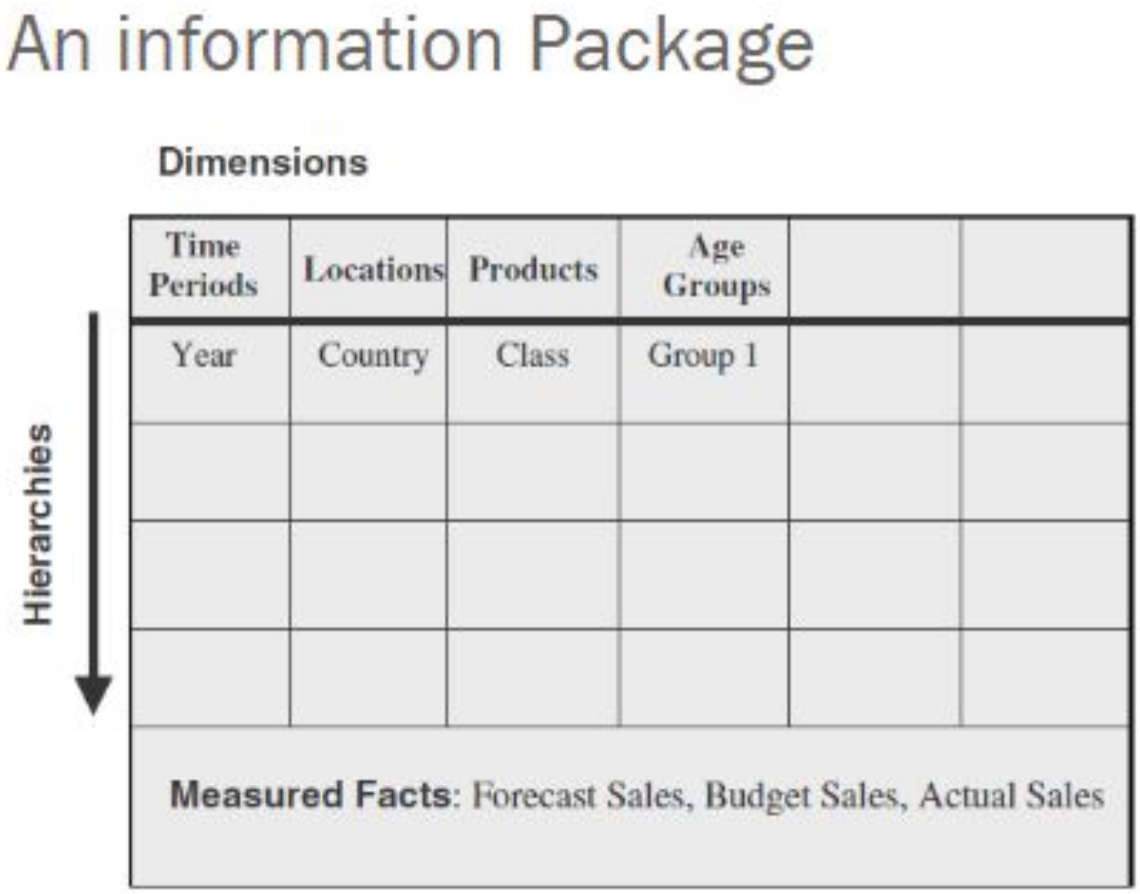
Figure 5-2 shows the analysis of sales units along the three business dimensions of product, time, and geography.

Information Package Diagram

Information Packages –

- Novel idea for determining and recording information requirements for a data warehouse.
- Determining requirements for a data warehouse is based on business dimensions
- The relevant dimension and measurements in that dimension are captured and kept in a data warehouse
- This creates an information package for a specific subject

Information Package Diagram



An information Package

Dimensions

Hierarchies

Time Periods	Locations	Products	Age Groups		
Year	Country	Class	Group 1		

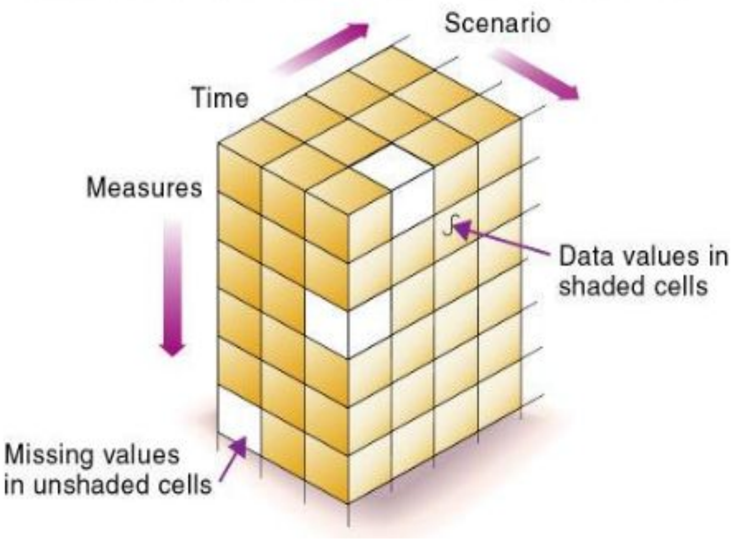
Measured Facts: Forecast Sales, Budget Sales, Actual Sales

Information Package Diagram

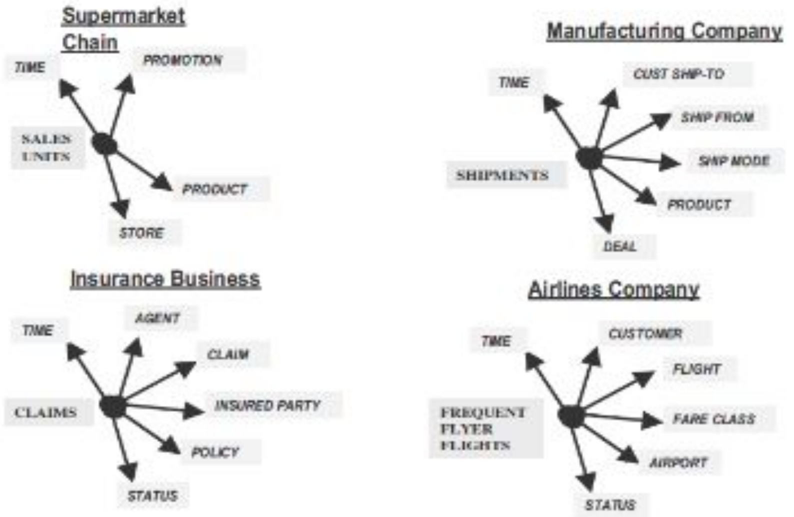
Business dimensions

- In requirements collection phase, the end users can provide the measurements which are important to that department.
- They can also give insights of combining the various pieces of information for strategic decision making.
- Managers think of business in terms of business dimensions
- The managers try to evaluate business in different dimensions.

Data block for Caffeine Free Cola->New York



Examples of business dimensions



Example :information package for analyzing sales for a certain business

- The subject here is sales.
- The measured facts or the measurements that are of interest for analysis are shown in the bottom section of the package diagram. In this case, the measurements are **actual sales, forecast sales, and budget sales.**
- The business dimensions along which these measurements are to be analyzed are shown at the top of diagram as column headings.
- In our example, these dimensions are **time, location, product, and demographic age group.** Each of these business dimensions contains a hierarchy or levels.
- For example, the time dimension has the hierarchy going from year down to the level of individual day. The other intermediary levels in the time dimension could be quarter, month, and week.
- These levels or hierarchical components are shown in the information package diagram. The subject here is sales.

IPD enables you to….

- Define the common subject areas
- Design key business metrics
- Decide how data must be presented
- Determine how users will aggregate or roll up
- Decide the data quantity for user analysis or query
- Decide how data will be accessed
- Establish data granularity
- Estimate data warehouse size
- Determine the frequency for data refreshing
- Ascertain how information must be packaged

Dimension Hierarchies/Categories

- When a user analyzes the measurements along a business dimension, the user usually would like to see the numbers first in summary and then at various levels of detail.
- What the user does here is to traverse the hierarchical levels of a business dimension for getting the details at various levels.
- The hierarchy of the time dimension consists of the levels of **year, quarter, and month**.
- The dimension hierarchies are the paths for **drilling down** or **rolling up** in our analysis.
- Within each major business dimension there are categories of data elements that can also be useful for analysis.
- Hierarchies and categories are included in the information packages for each dimension.

Dimensional Data Modeling

- Dimensional Data Modeling is one of the data modeling techniques used in data warehouse design.
- Goal: Improve the data retrieval
- The concept of Dimensional Modeling was developed by Ralph Kimball which is comprised of facts and dimension tables
- Since the main goal of this modeling is to improve the data retrieval so it is optimized for SELECT OPERATION
- The advantage of using this model is that we can store data in such a way that it is easier to store and retrieve the data once stored in a data warehouse.
- Dimensional model is the data model used by many OLAP systems.

Dimensional Data Modeling

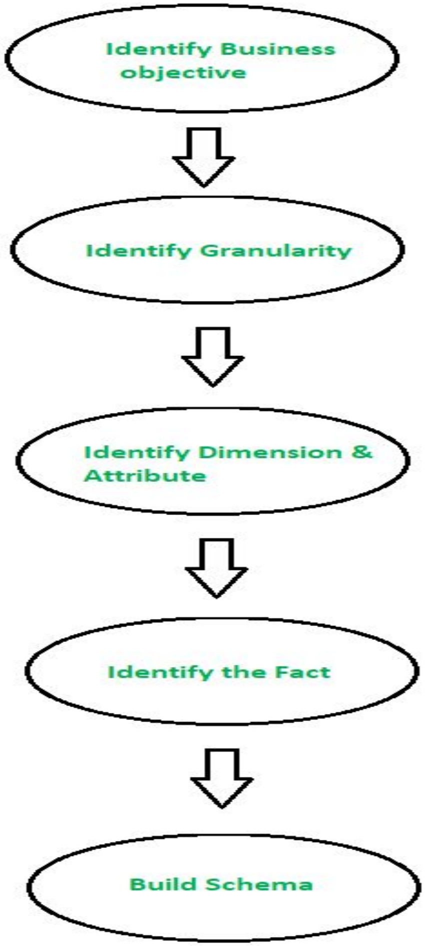
Steps to Create Dimensional Data Modeling:

Step-1: Identifying the business objective –

The first step is to identify the business objective. Sales, HR, Marketing, etc. are some examples as per the need of the organization.
Since it is the most important step of Data Modelling the selection of business objective also depends on the quality of data available for that process.

Step-2: Identifying Granularity –

Granularity is the lowest level of information stored in the table.
The level of detail for business problem and its solution is described by Grain.



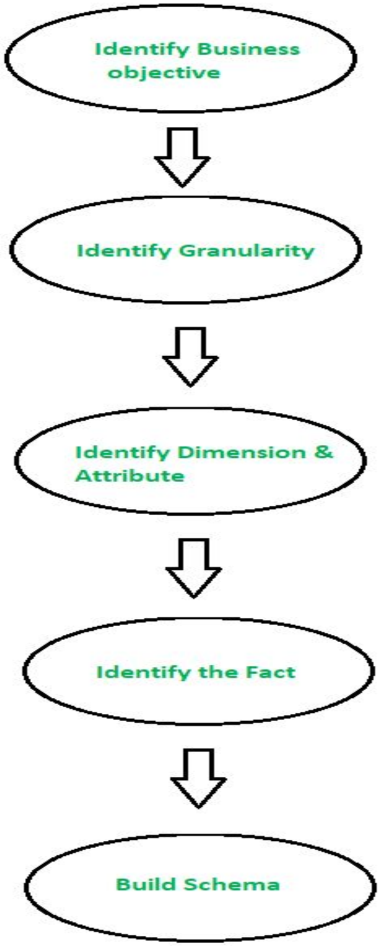
Dimensional Data Modeling

Step-3: Identifying Dimensions and its Attributes –
Dimensions are objects or things like table. Dimensions categorize and describe data warehouse facts and measures in a way that support meaningful answers to business questions.

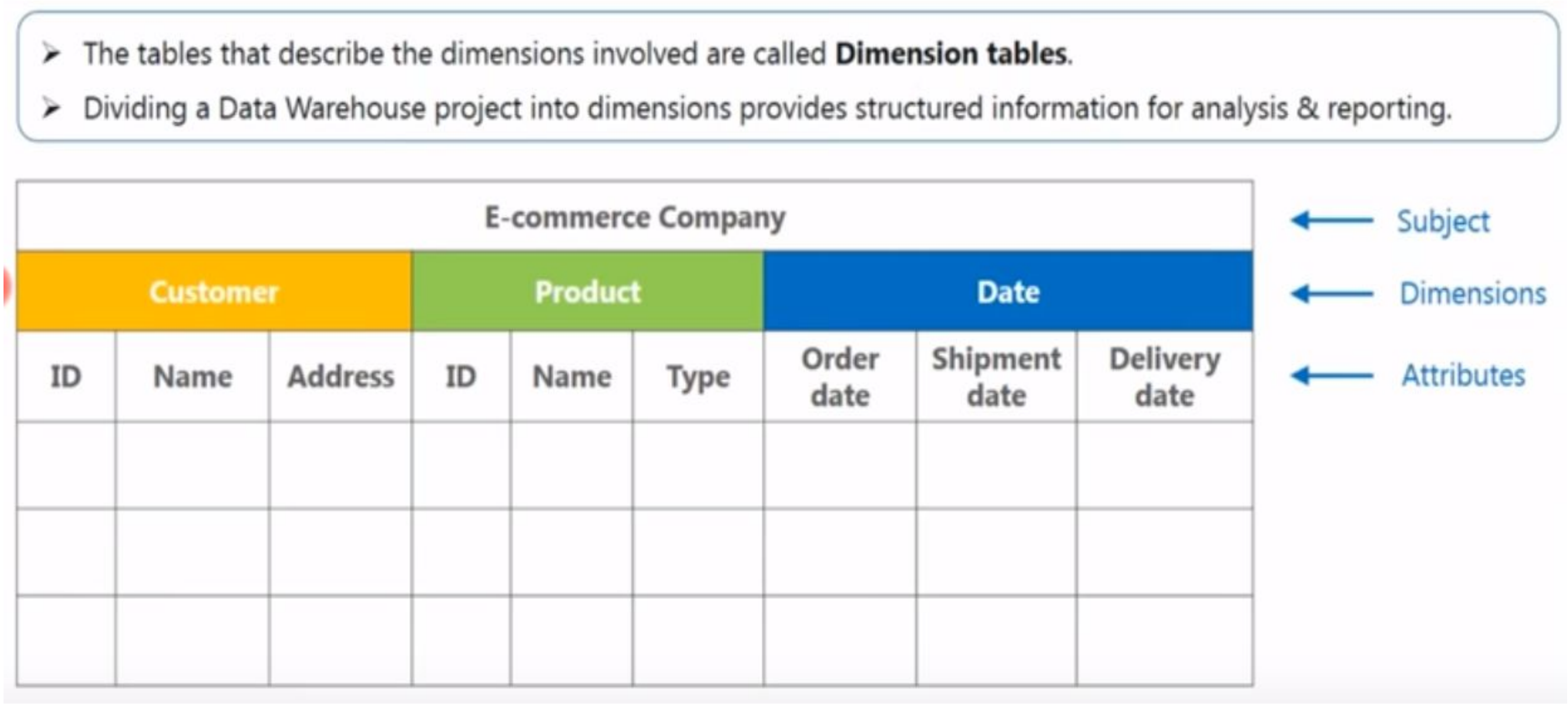
A data warehouse organizes descriptive attributes as columns in dimension tables. For Example, the data dimension may contain data like a year, month and weekday.

Step-4: Identifying the Fact –
The measurable data is hold by the fact table. Most of the fact table rows are numerical values like price or cost per unit, etc.

Step-5: Building of Schema –
We implement the Dimension Model in this step. A schema is a database structure.
Popular schemes: Star Schema, Snowflake Schema, Fact constellation scheme



Dimensions

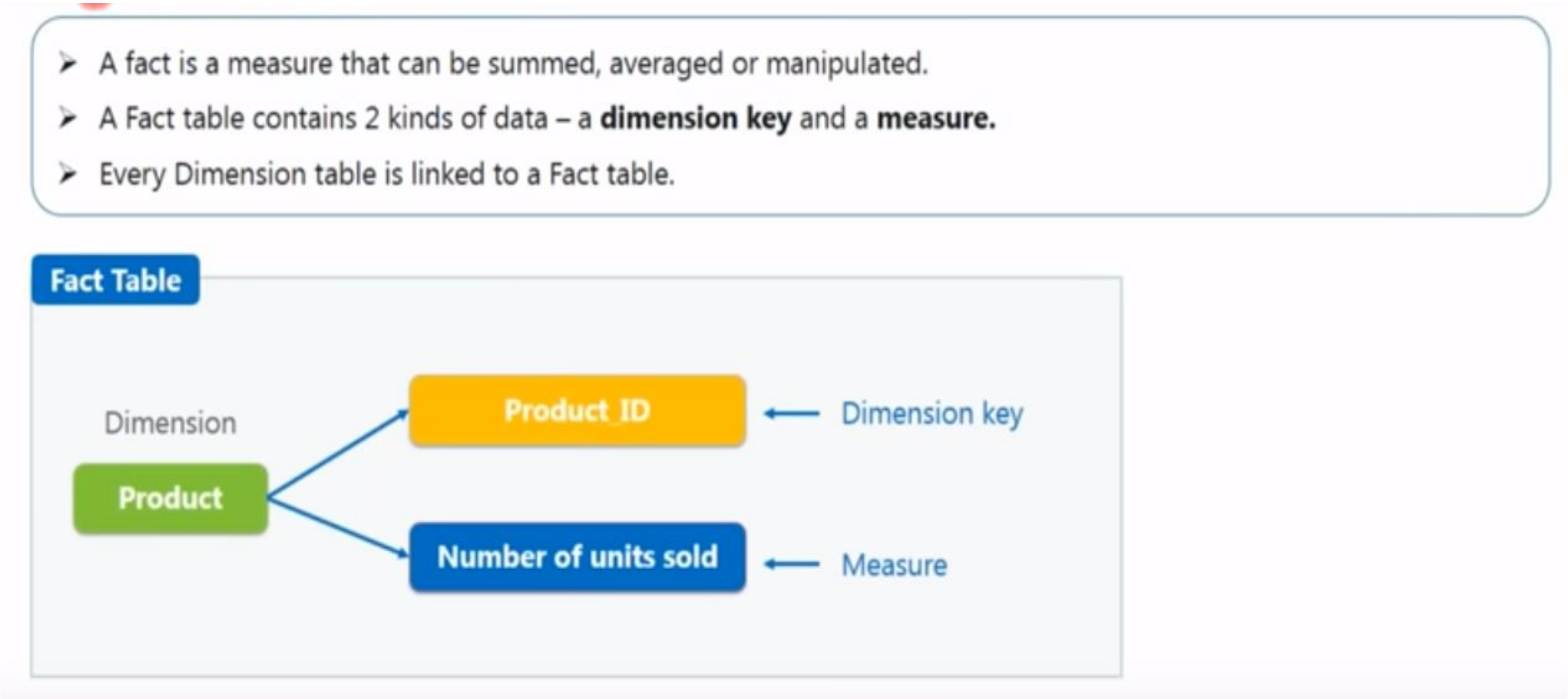


- The tables that describe the dimensions involved are called **Dimension tables**.
- Dividing a Data Warehouse project into dimensions provides structured information for analysis & reporting.

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

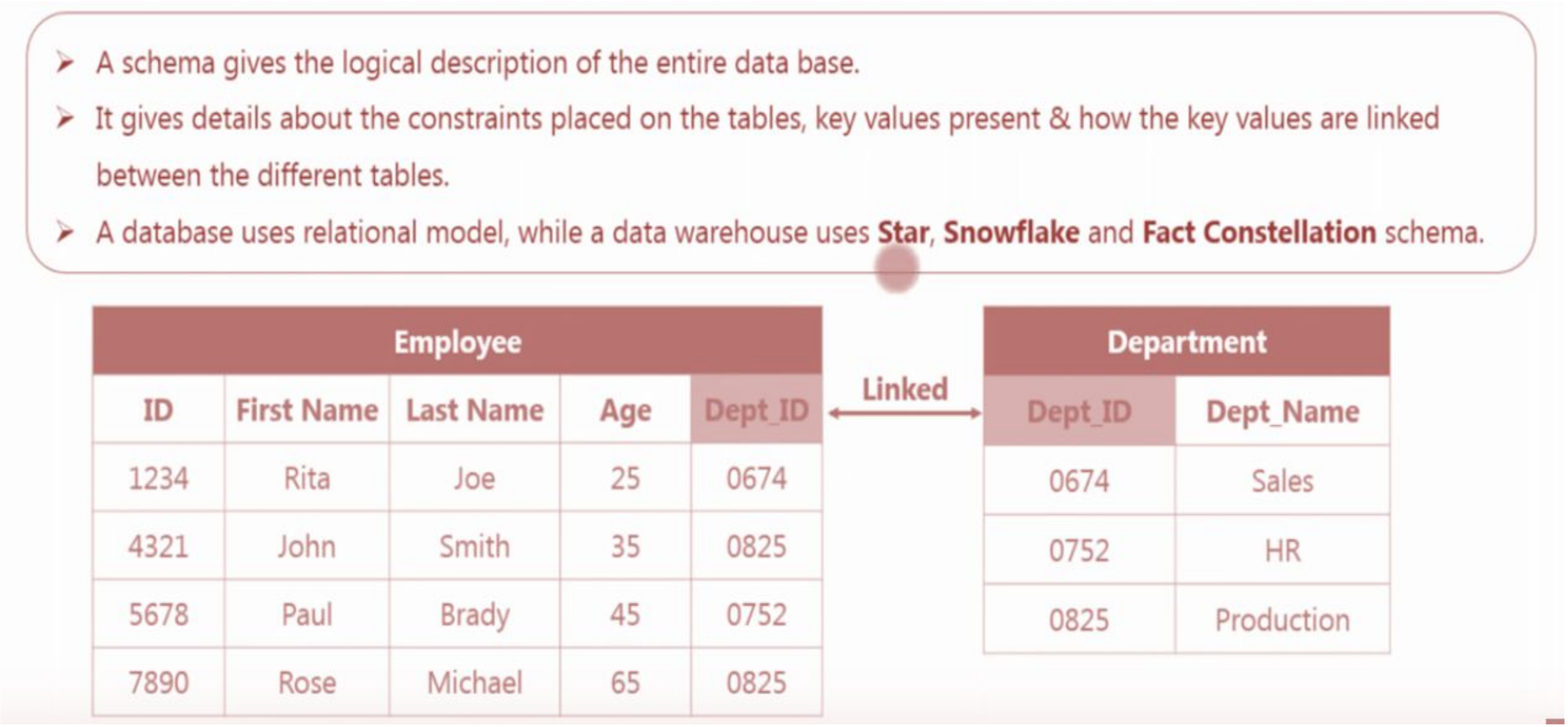
End Users fires a queries on these tables which contains descriptive information

Facts and 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.

Schema



- 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

Star Schema in Data Warehouse modeling

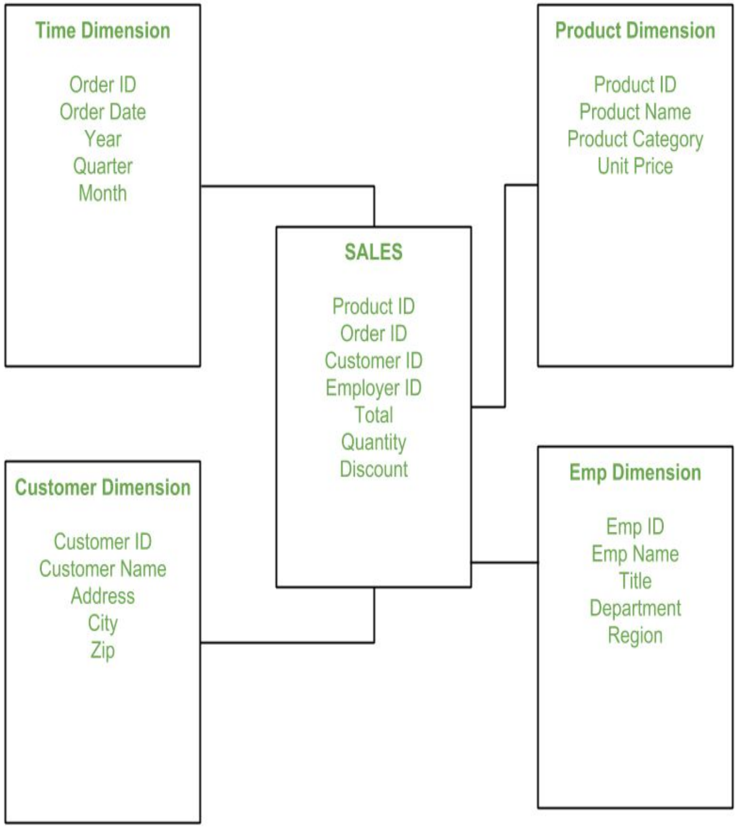
Star schema is the fundamental schema among the data mart schema and it is simplest.

This schema is widely used to develop or build a data warehouse and dimensional data marts.

It includes one or more fact tables indexing any number of dimensional tables.

The star schema is a necessary case of the snowflake schema. It is also efficient for handling basic queries.

It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star's points.



Star Schema in Data Warehouse modeling

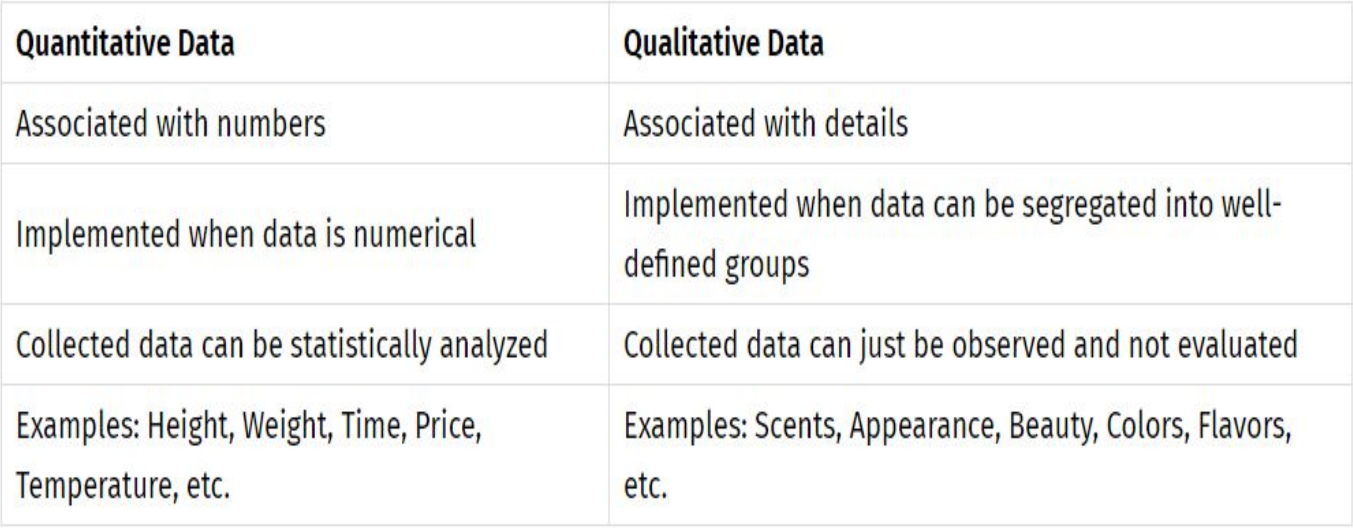
In the above demonstration,

- **SALES** is a **fact table** having attributes i.e. (**Product ID, Order ID, Customer ID, Employer ID, Total, Quantity, Discount**) which references to the dimension tables(first 4) and next 3 are measures.
- **Employee dimension** table contains the attributes: **Emp ID, Emp Name, Title, Department and Region.**
- **Product dimension** table contains the attributes: **Product ID, Product Name, Product Category, Unit Price.**
- **Customer dimension** table contains the attributes: **Customer ID, Customer Name, Address, City, Zip.**
- **Time dimension** table contains the attributes: **Order ID, Order Date, Year, Quarter, Month.**

Star Schema in Data Warehouse modeling

Model of Star Schema –
In Star Schema, Business process data, that holds the **quantitative data about a business is distributed in fact tables**, and **dimensions which are descriptive characteristics related to fact data**.

Sales price, sale quantity, distance, speed, weight, and weight measurements are few examples of fact data in star schema.



Quantitative Data	Qualitative Data
Associated with numbers	Associated with details
Implemented when data is numerical	Implemented when data can be segregated into well-defined groups
Collected data can be statistically analyzed	Collected data can just be observed and not evaluated
Examples: Height, Weight, Time, Price, Temperature, etc.	Examples: Scents, Appearance, Beauty, Colors, Flavors, etc.

Star Schema in Data Warehouse modeling

Advantages of Star Schema –

Simpler Queries:
Join logic of star schema is quite cinch in compare to other join logic which are needed to fetch data from a transactional schema that is highly normalized.

Simplified Business Reporting Logic:
In compared to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as as-of reporting and period-over-period.

Feeding Cubes:
Star schema is widely used by all OLAP systems to design OLAP cubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.

Star Schema in Data Warehouse modeling

Disadvantages of Star Schema –

- **Data integrity is not enforced well** since in a highly de-normalized schema state.
- **Not flexible** in terms if analytical needs as a normalized data model.
- **Star schemas don’t reinforce many-to-many relationships within business entities** – at least not frequently.

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Thank You