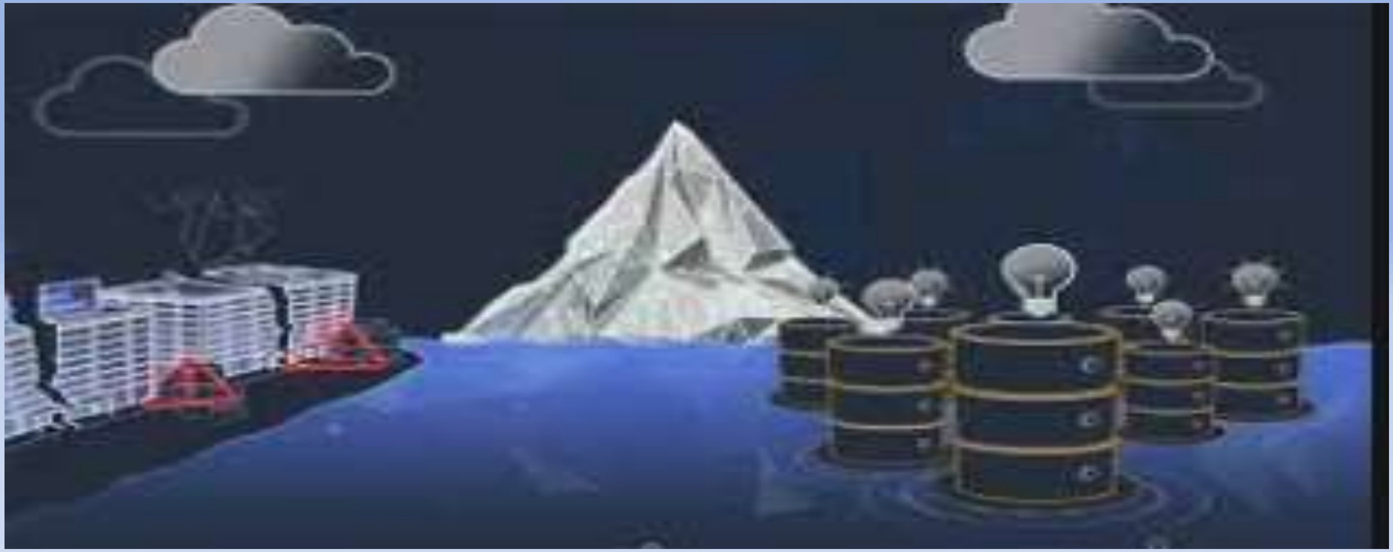
DMDW – Module-1



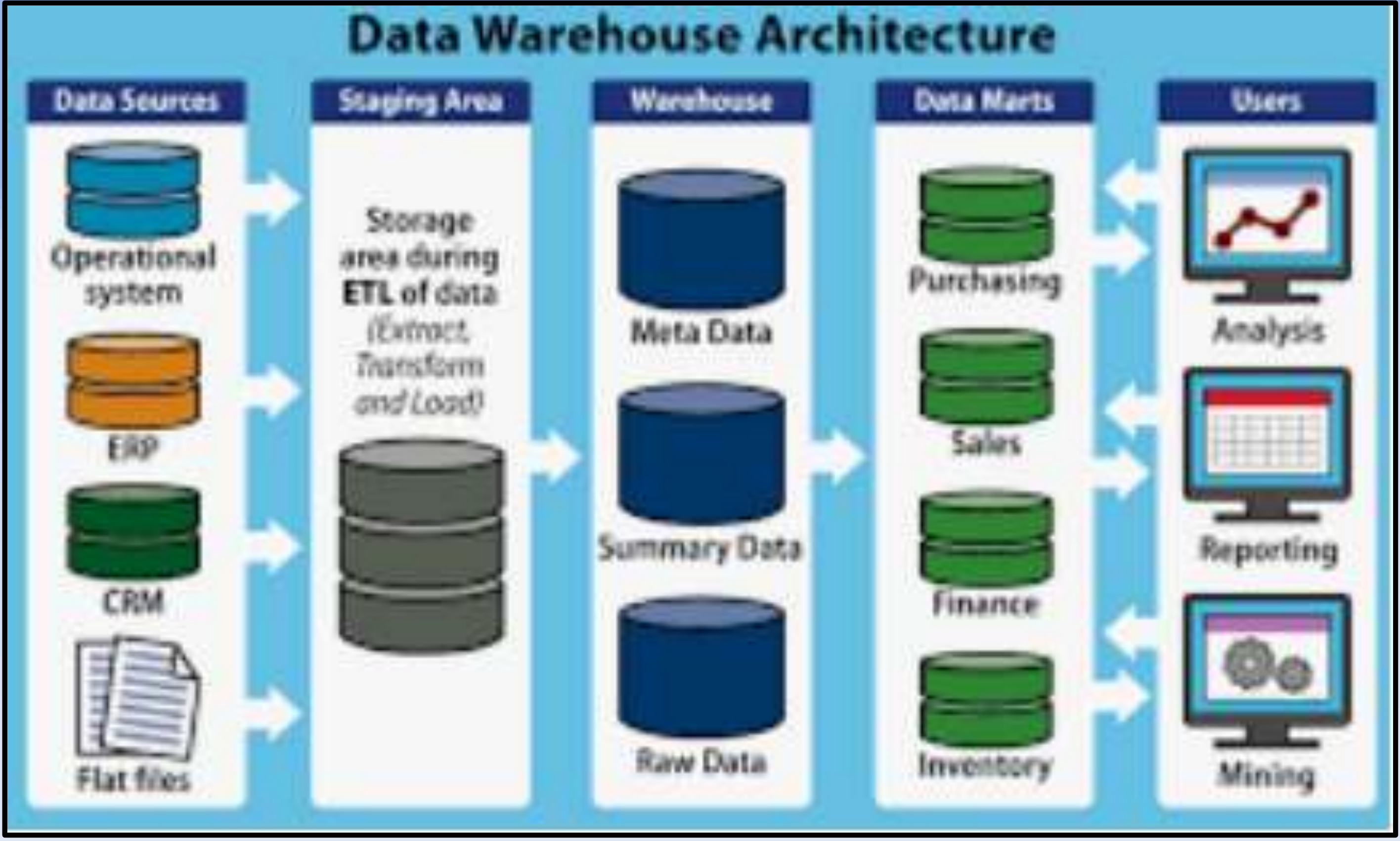
By

Dr. Pulak Sahoo

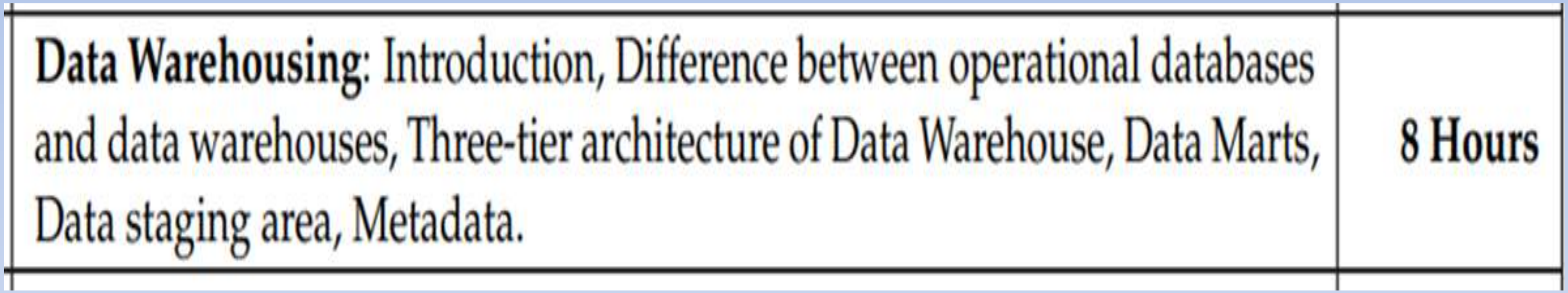
Associate Professor

Dept of CSE, SIT, BBSR

Data Warehousing

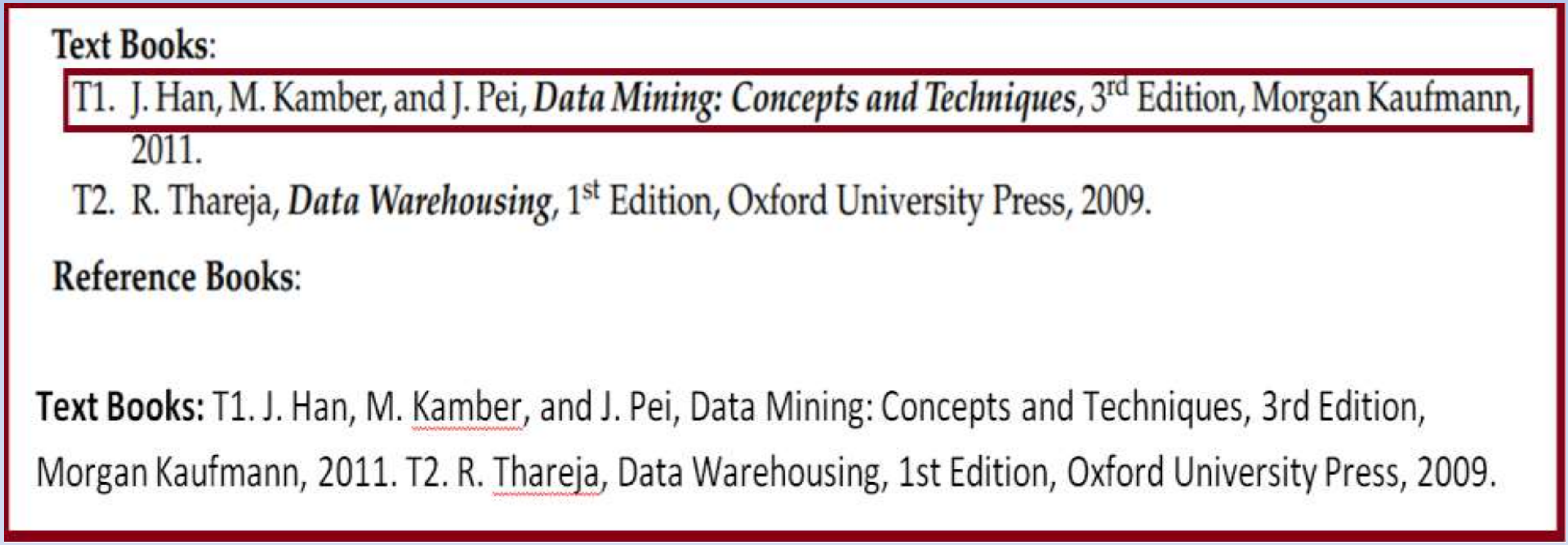


Module-1 Syllabus



Data Warehousing: Introduction, Difference between operational databases and data warehouses, Three-tier architecture of Data Warehouse, Data Marts, Data staging area, Metadata.	**8 Hours**

Books



Text Books:

T1. J. Han, M. Kamber, and J. Pei, ***Data Mining: Concepts and Techniques***, 3rd Edition, Morgan Kaufmann, 2011.

T2. R. Thareja, ***Data Warehousing***, 1st Edition, Oxford University Press, 2009.

Reference Books:

Text Books: T1. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, 2011. T2. R. Thareja, Data Warehousing, 1st Edition, Oxford University Press, 2009.

Example – Case Study-1

Village Library story

Case Study-2

Garment Chain

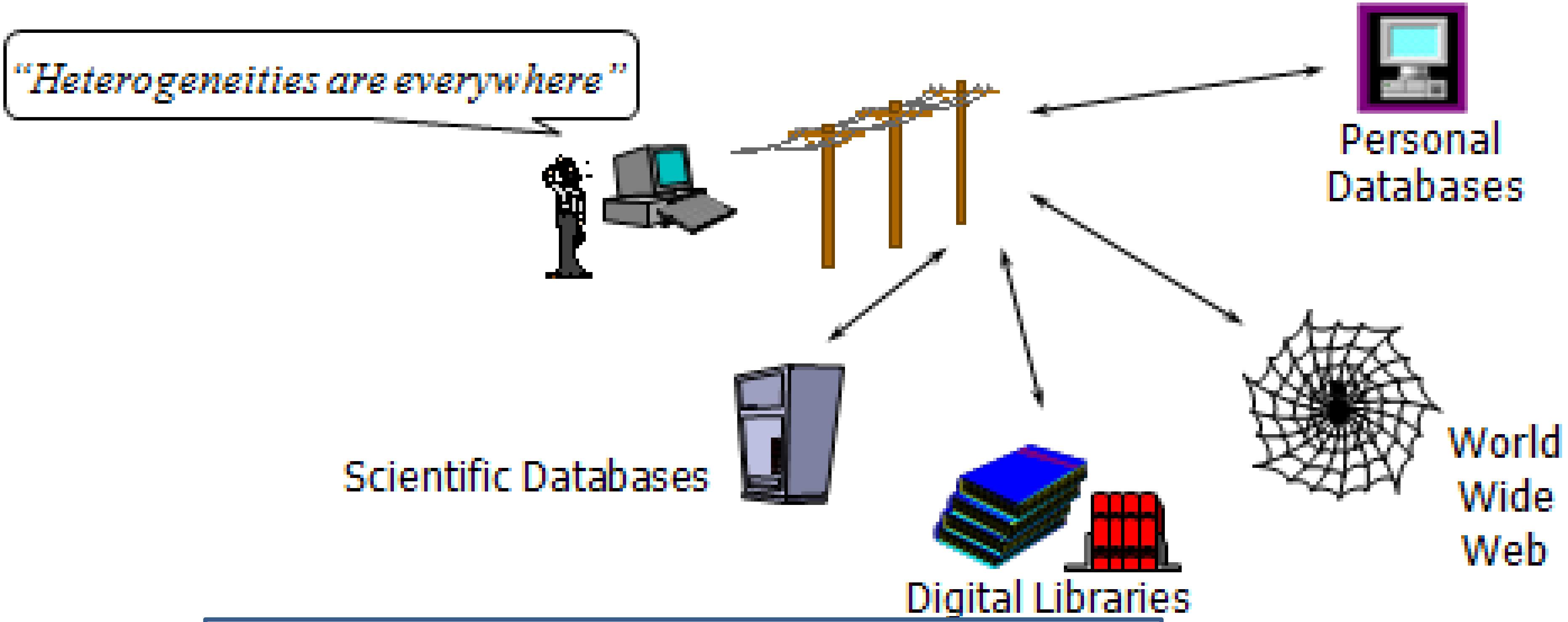
- Pallav Raj is the CEO of a large garments retail chain called JRTs.
- JRTs has approximately 100 stores spread throughout the country.
- Pallav Raj asks one of his employees to provide him:
  1) A status report on the business as he wishes to know if the company was making an overall profit or loss

  2) A detailed product report of the previous year as he wishes to know which products sold well and those that did not even have a marginal sale

Case Study: The Need for Data Warehousing

- How does the employee calculate if the company was making an overall profit or loss ???
  - Manually !!
  - Tedious task !!
- And further, how does the employee find a detailed product report of the previous year ???
  - ???

Problem: Heterogeneous Information Sources



"Heterogeneities are everywhere"

Personal Databases

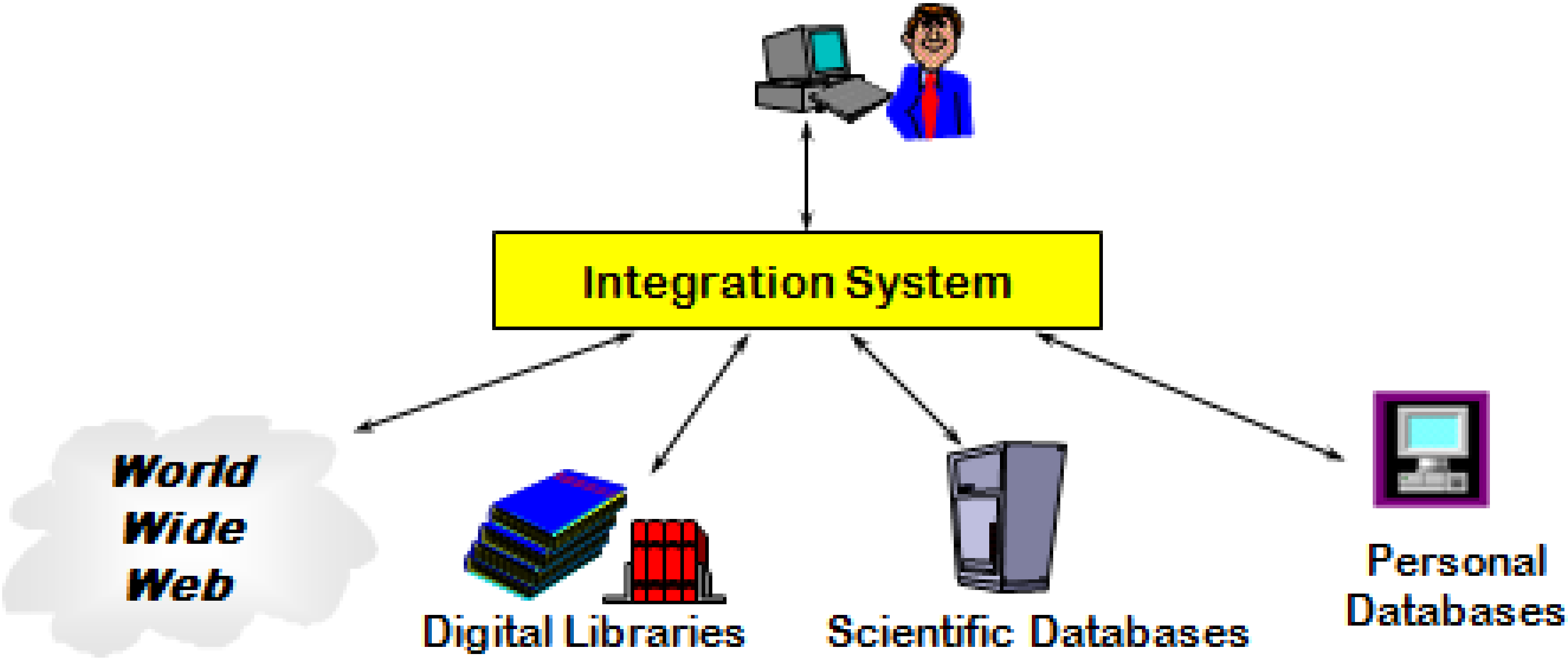
Scientific Databases

Digital Libraries

World Wide Web

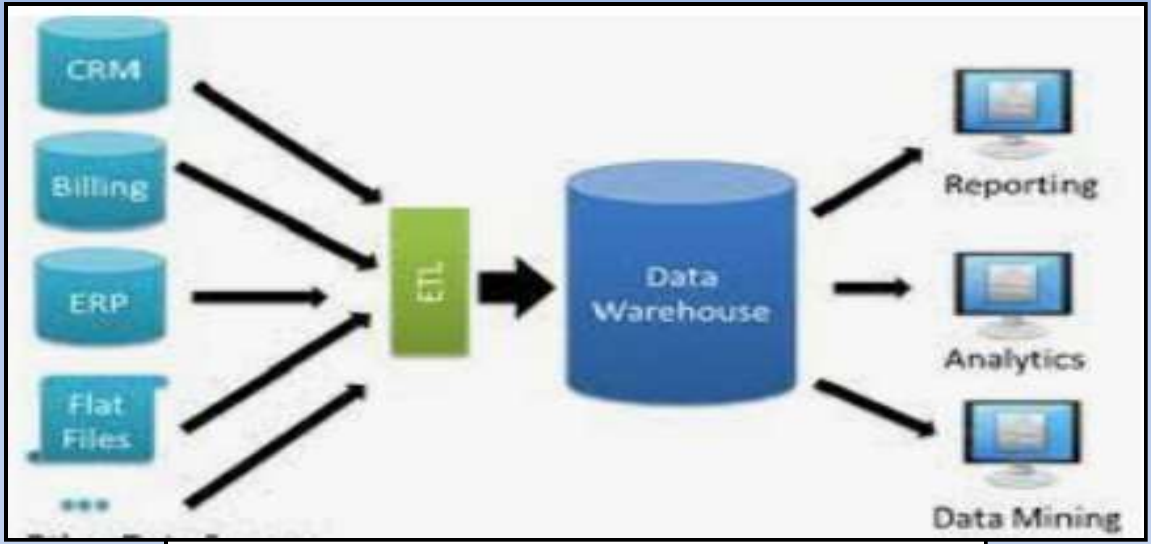
- Different interfaces
- Different data representations
- Duplicate and inconsistent information

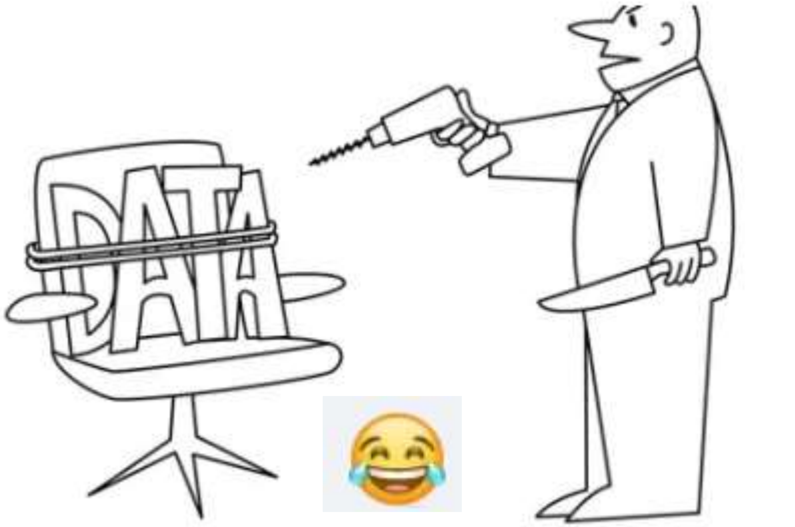
Goal: Unified Access to Data



- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

What & why **Data Warehouse??**





"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

What is a Data Warehouse?

A Practitioners Viewpoint

> "A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."
>
> -- Barry Devlin, *IBM Consultant*

William H Inmon’s definition

- Is the “**Father of Data warehouse**”

- A data warehouse is ***subject-oriented, integrated, time-variant, nonvolatile*** collection of data in support of management’s decision making process

Sean Kelly definition

Data in the **data warehouse** is :

- **Separate**
- **Available**
- **Integrated**
- **Time stamped**
- **Subject oriented**
- **Nonvolatile**
- **Accessible**

What is Data Warehouse?

- **Defined in many different ways:**
  - ➢A **decision support** database that is maintained separately from the org.'s **operational** database
  - ➢Support **info. processing** by providing a solid platform of consolidated, historical data for **analysis**
- **Data warehousing:**
  - ➢The process of **constructing** & **using data warehouses**

Data Warehouse—Subject-Oriented

Data warehouse is organized around **subjects** such as *sales, products, customers & time periods etc.*

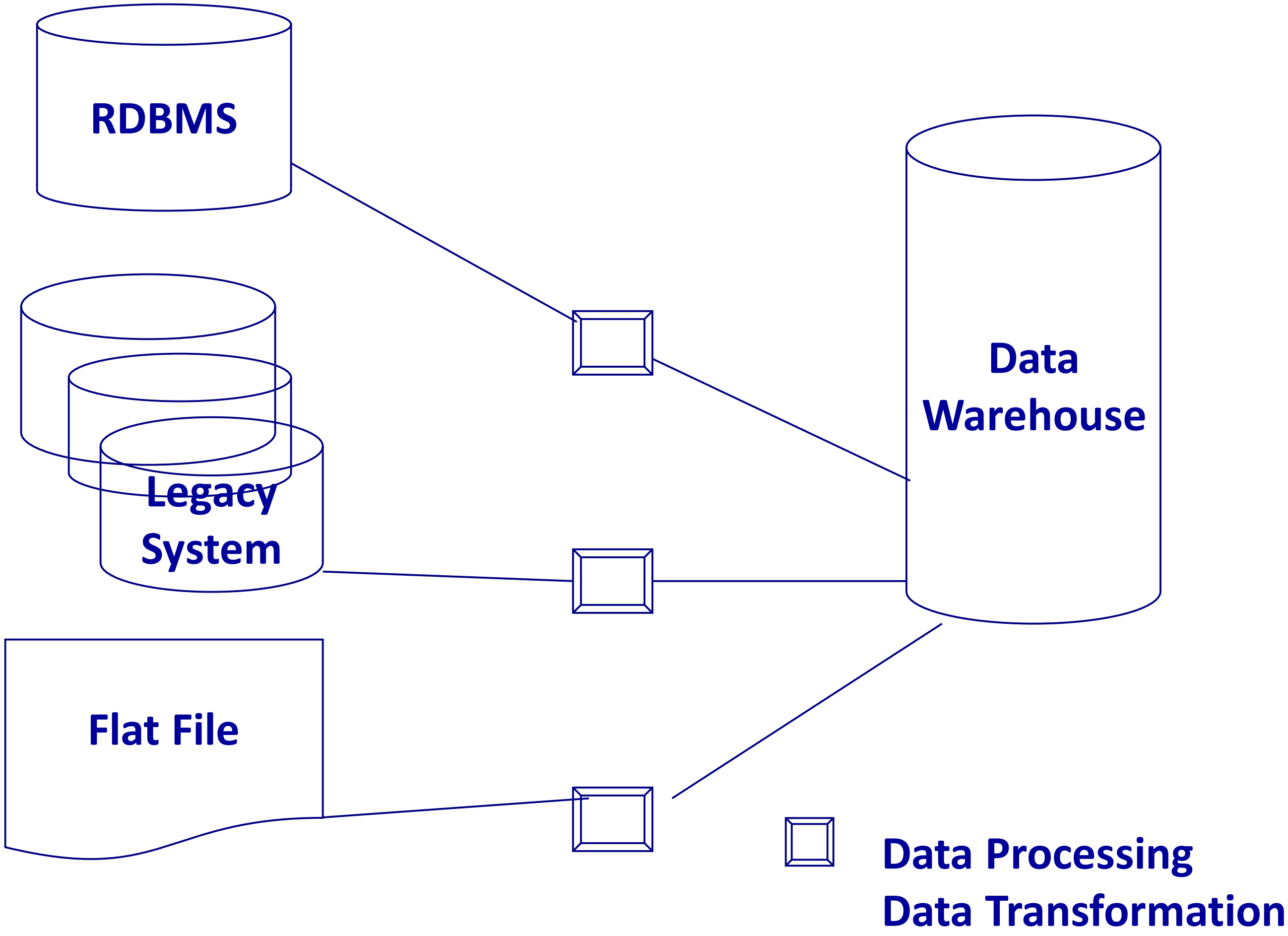
Focusing on the **modeling** & **analysis** of data for **decision making**, not for daily operations or transaction processing

Provides a **simple** & **concise view** on a **particular subject** by excluding unuseful data in the decision support process

In **operational system**, data are organized based on individual applications to support those particular operational system

Data Warehouse—Integrated

- Constructed by **integrating multiple, heterogeneous data sources** like:
  - **Relational databases, flat files, on-line transaction records**
- **Data cleaning** & **data integration** techniques are applied to
  - Ensure **consistency** in *naming conventions, encoding structures, attribute measures* etc. among **diff. data sources**
  - When data is moved to the warehouse, it is **converted**



RDBMS

Legacy System

Flat File

Data Warehouse

□ Data Processing
Data Transformation

Data Warehouse—Time Variant

- The **time horizon** for building the **data warehouse** is significantly **longer** than that of **operational systems**
  - **Operational database:** Contain current value data
  - **Data warehouse data**: provide info from a historical perspective *(Ex: past 5-10 years)*
- Every **key structure** in the **data warehouse**
  - Contains an element of time
  - But the key of operational data may not contain "time element"

- The **time-variant** nature of the **data** in a data warehouse:
  - Allows for **analysis of the past**
  - Relates info. to the **present**
  - Enables **forecasts for the future**

Data Warehouse is Nonvolatile

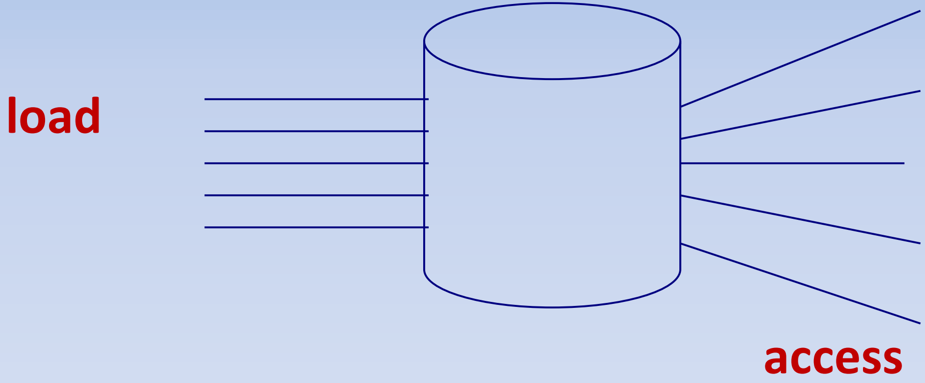
- **A physically separate storage of data transformed** from the **operational** env.

- **Operational update** of data do not occur in the **DWH env**

  - Does not require **transaction processing, recovery** & **concurrency control** mechanisms

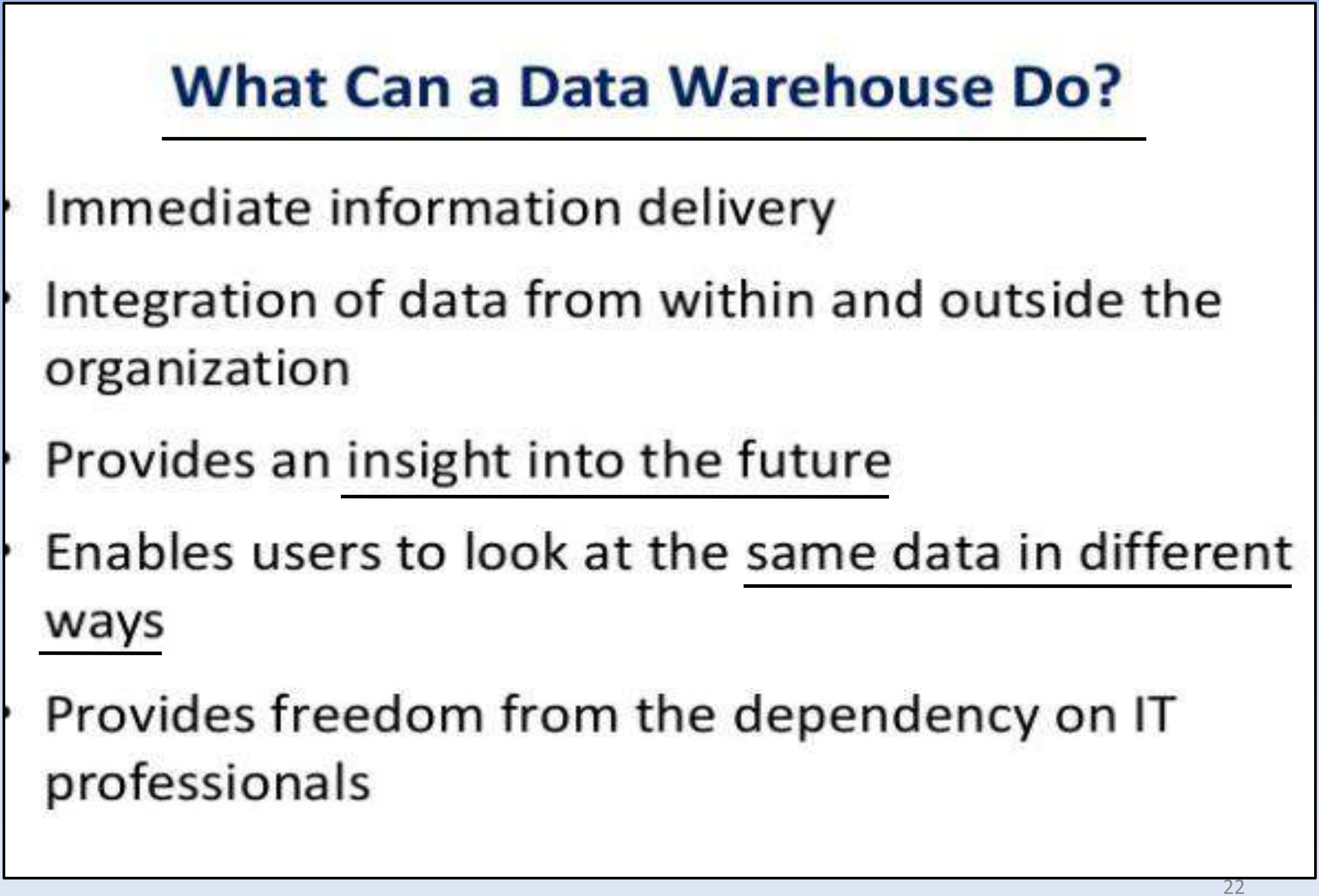
- **Data** once recorded, **cannot be updated**

- **DWH** requires **two operations** in data accessing
    - **Initial loading of data**
    - **Access of data**



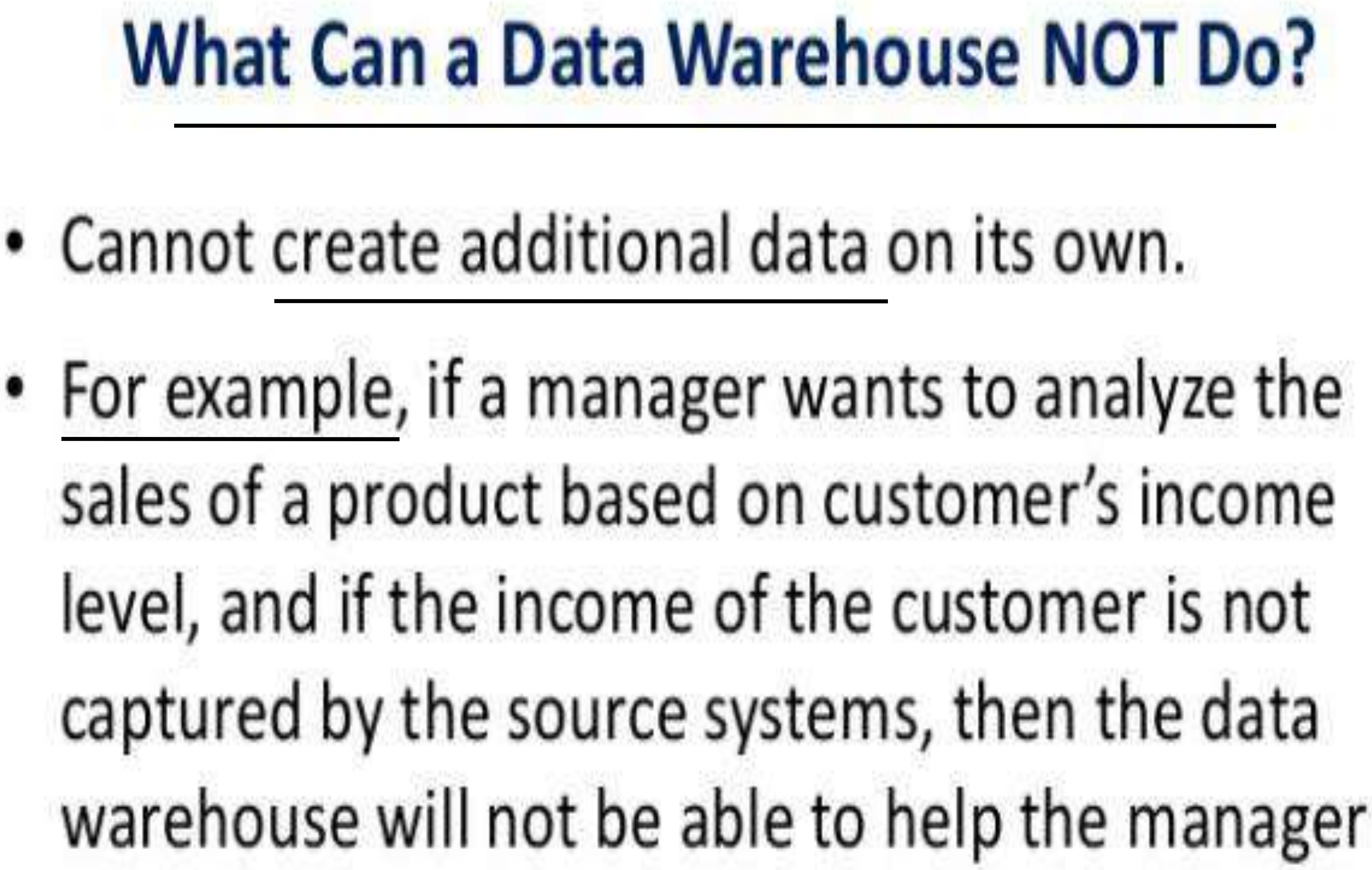
- The data in DWH is **not as volatile** as the data in oper. Database

- The **data in a DWH** is primarily for **query** & **analysis**



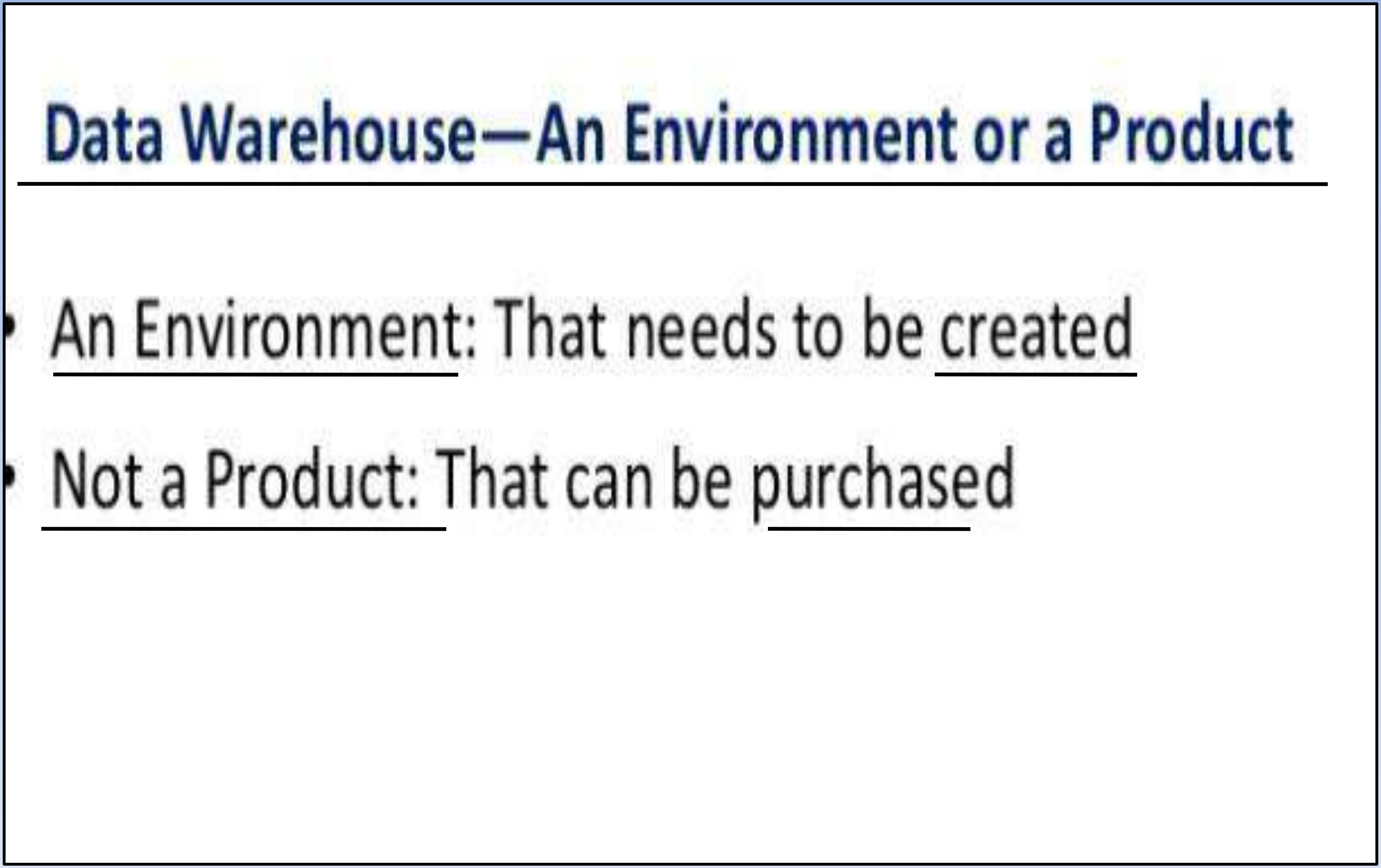
What Can a Data Warehouse Do?

- Immediate information delivery
- Integration of data from within and outside the organization
- Provides an insight into the future
- Enables users to look at the same data in different ways
- Provides freedom from the dependency on IT professionals



What Can a Data Warehouse NOT Do?

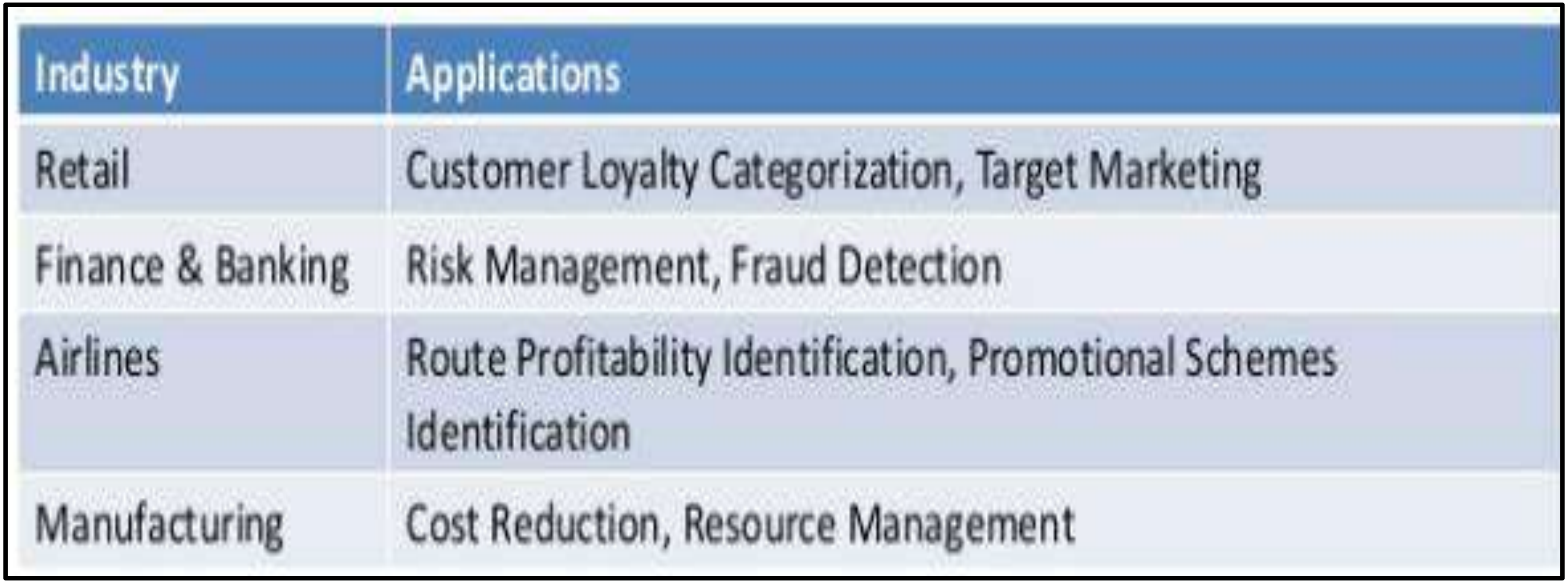
- Cannot create additional data on its own.
- For example, if a manager wants to analyze the sales of a product based on customer’s income level, and if the income of the customer is not captured by the source systems, then the data warehouse will not be able to help the manager



Data Warehouse—An Environment or a Product

- An Environment: That needs to be created
- Not a Product: That can be purchased

Applications of Data Warehouse System



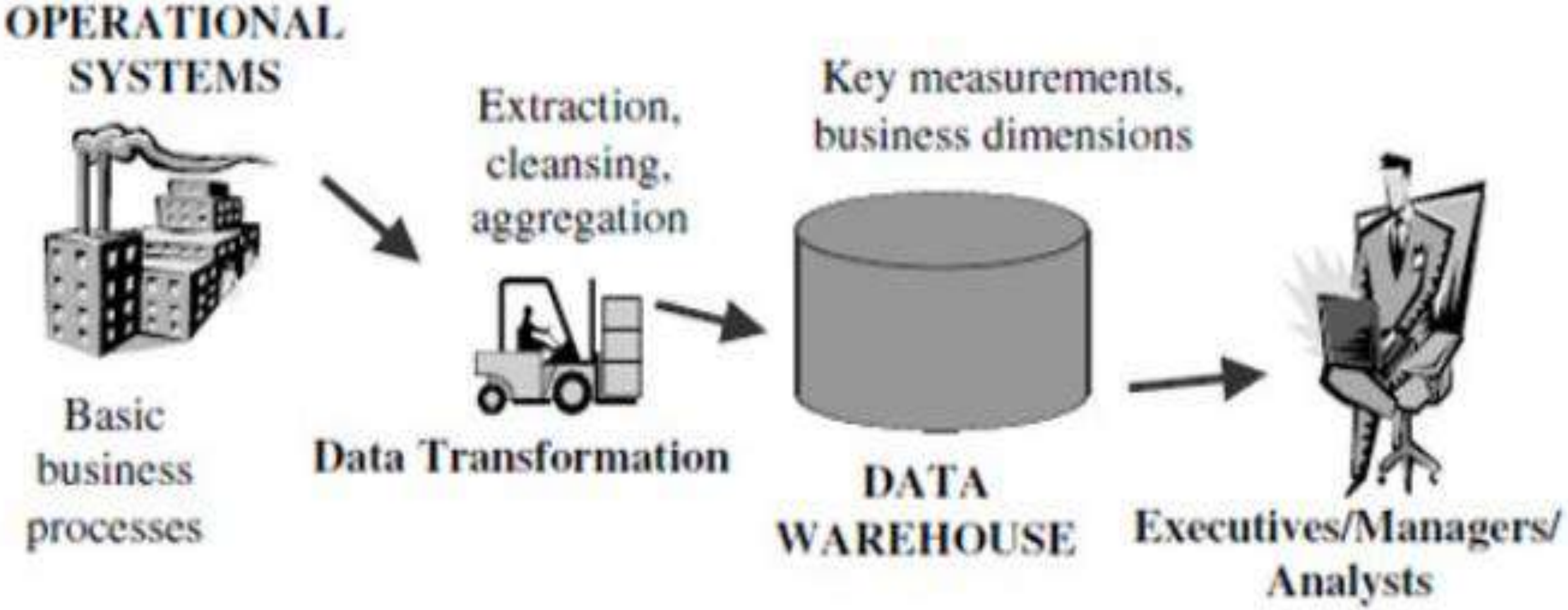
Industry	Applications
Retail	Customer Loyalty Categorization, Target Marketing
Finance & Banking	Risk Management, Fraud Detection
Airlines	Route Profitability Identification, Promotional Schemes Identification
Manufacturing	Cost Reduction, Resource Management

Government : Manpower planning, development & cost control

Other application areas include:

Insurance companies, utilities providers, healthcare providers, financial services companies, telecommunication service providers, travel, transport/tourism companies, security agencies, logistic, inventory & purchasing

Services
Banking
Insurance
Finance
Retailing
Education
Manufact -uring
Healthcare



OPERATIONAL SYSTEMS

Basic business processes

Extraction, cleansing, aggregation

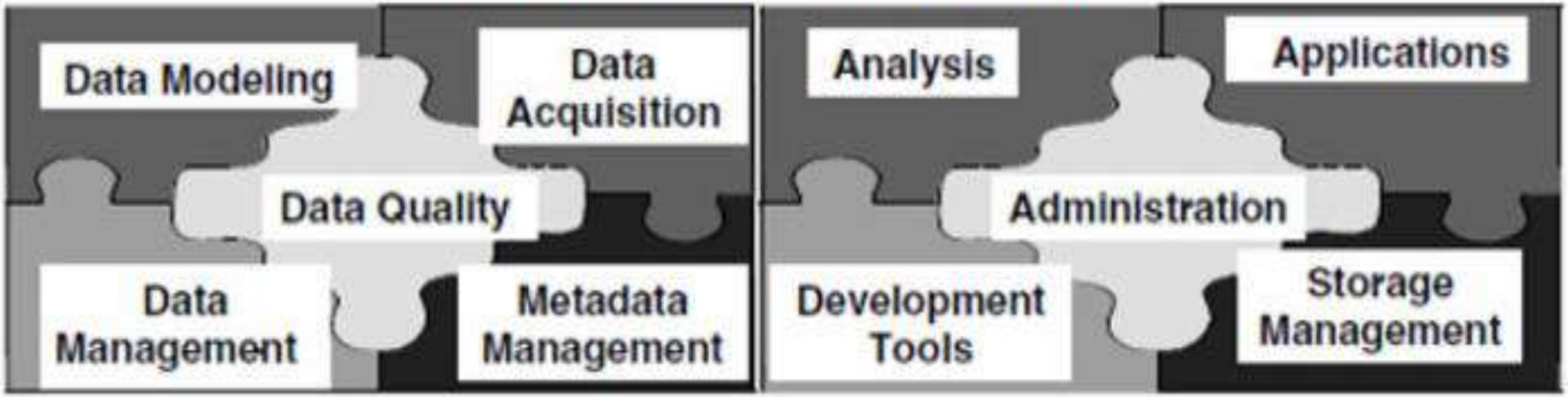
Data Transformation

Key measurements, business dimensions

DATA WAREHOUSE

Executives/Managers/ Analysts

BLEND OF TECHNOLOGIES



Data Modeling

Data Acquisition

Data Quality

Data Management

Metadata Management

Analysis

Applications

Administration

Development Tools

Storage Management

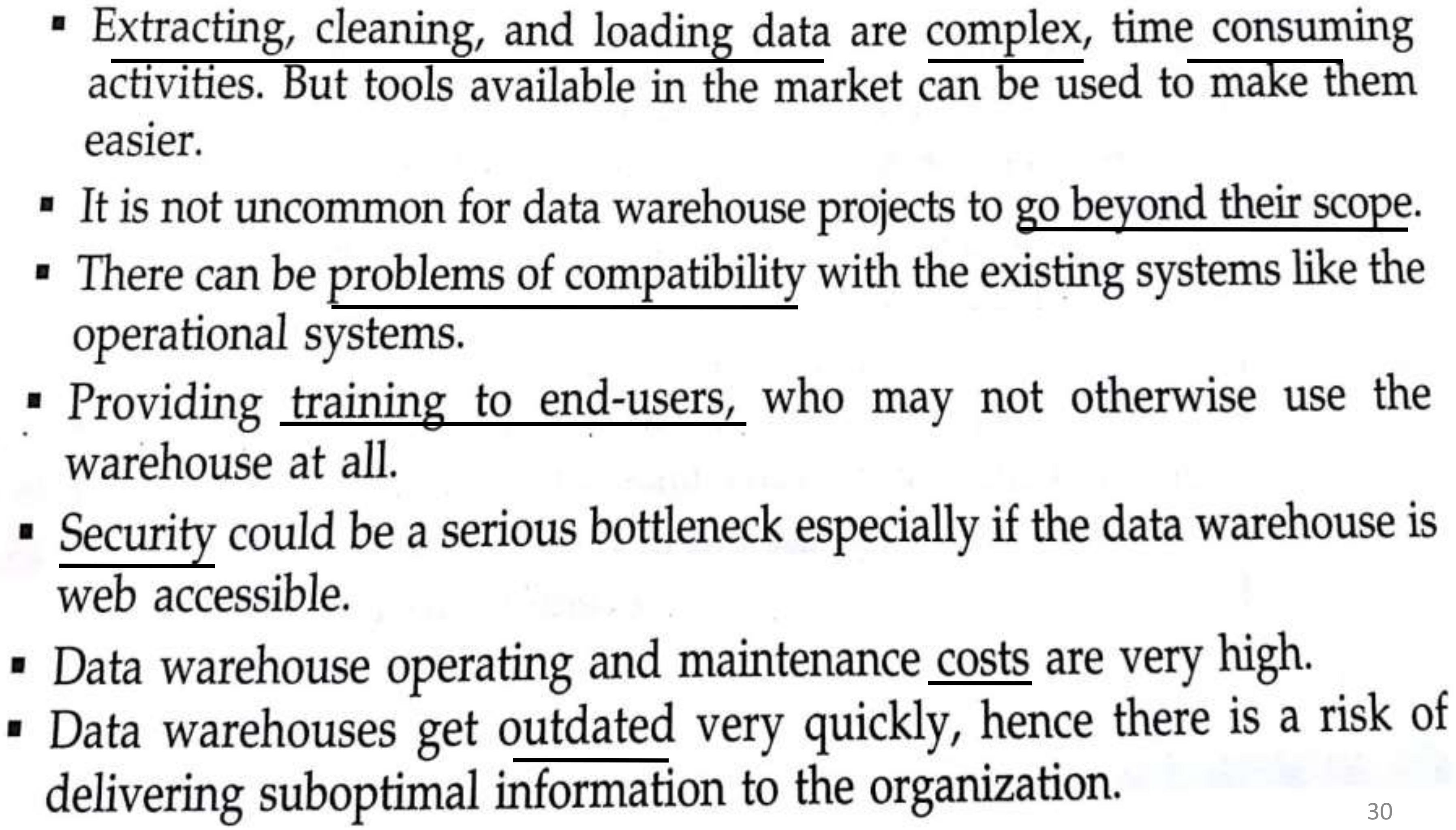
Blend of technologies

- Data acquisition
- Data management
- Metadata management
- Storage management
- Development of tools
- Data Analysis
- Data modeling

DWH Building Tasks

- Accurate identification of **business info.**
- Identification & prioritization of **subject areas**
- Selection of **hardware** /**software** components
- **Extracting** , **cleansing**, **transforming** & **validating** data
- Providing **user friendly**, **powerful tools** to users for **accessing** the data
- Giving adequate **training** to users
- Establishing **procedures** for **maintenance** & **enhancement**
- Remove the **inconsistencies**

CONCERNS IN DATA WAREHOUSING



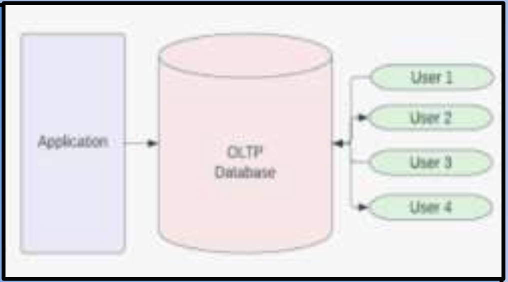
- Extracting, cleaning, and loading data are complex, time consuming activities. But tools available in the market can be used to make them easier.
- It is not uncommon for data warehouse projects to go beyond their scope.
- There can be problems of compatibility with the existing systems like the operational systems.
- Providing training to end-users, who may not otherwise use the warehouse at all.
- Security could be a serious bottleneck especially if the data warehouse is web accessible.
- Data warehouse operating and maintenance costs are very high.
- Data warehouses get outdated very quickly, hence there is a risk of delivering suboptimal information to the organization.

Data Warehouse vs. Heterogeneous DBMS

- **Heterogeneous DB integration**: A **query driven** approach:
  - Build **wrappers/mediators** on top of heterogeneous databases
  - When a query is posed to a client site, a **meta-dictionary** is used to **translate** the query into queries appropriate for individual heterogeneous sites & the results are **integrated** into a global answer set
  - **Issues**: Complex info. **filtering**, compete for **resources**

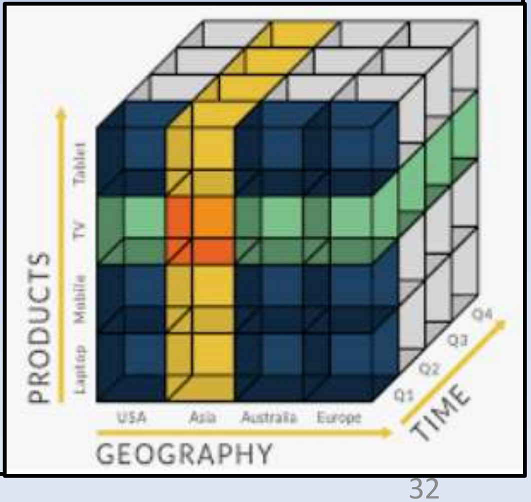
- **Data warehouse: update-driven** approach & with high performance:
  - info. from heterogeneous sources is **integrated** in advance & stored in warehouses for direct **query** & **analysis**

Data Warehouse vs. Operational DBMS



- **OLTP (on-line transaction processing)**
  - Major task of traditional relational **DBMS**
  - **Day-to-day operations**: *purchasing, inventory, banking, manufacturing, payroll, registration, accounting etc.*

- **OLAP (on-line analytical processing)**
  - Major task of **DWH system**
  - **Data analysis** & **decision making**



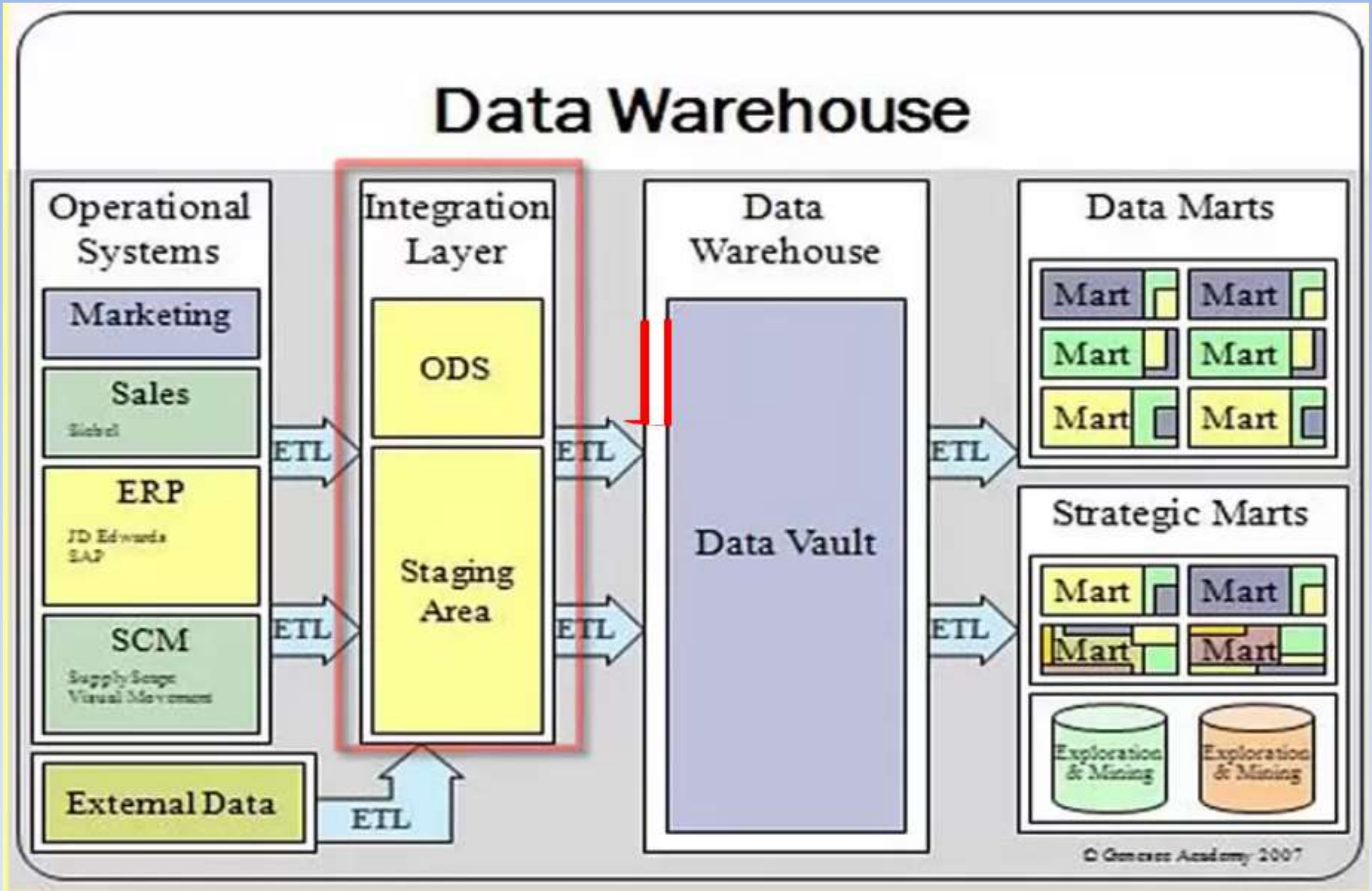
Distinct features (OLTP vs. OLAP):

- **User & system orientation**: customer vs. market
- **Data contents**: current, detailed vs. historical, consolidated
- **Database design**: ER + application vs. star + subject
- **View**: current, local vs. evolutionary, integrated
- **Access patterns**: update vs. read-only but complex queries

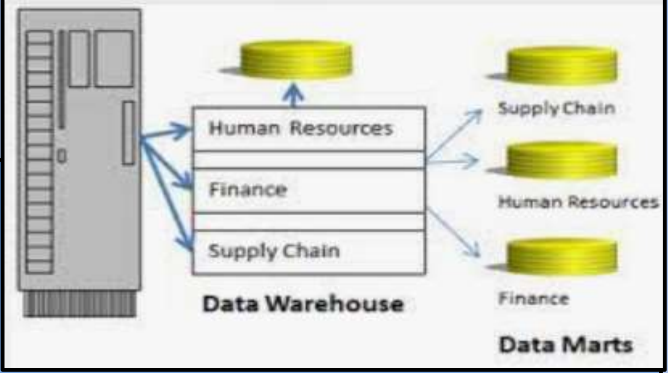
OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Data Warehouse Overview



Data Warehouse & Data Mart



What is Data Mart?

- A data mart is a **decision support system** that stores a no. of subject areas based on the needs of users in that department
- Data marts are **subset** of the enterprise DWH that are **localized** to a department & are highly aggregated & redundant
- A **subset of a DWH** that supports the requirements of a department or business function
- Data Mart is often **built** & **controlled** by a single department within an org.
- Every individual department owns the **H/W, S/W, data** & **programs** that are needed for the data mart
- The **DB design** for a data mart is done using a **star-join** structure that is optimal for the users needs within the dept

Data warehouse vs Data mart

Data Warehouse	Data Mart
▪ Corporate/Enterprise-wide scope	▪Departmental scope
▪ Low level granularity	▪High level granularity
▪ Lightly indexed	▪Highly indexed
▪ Combination of >1 business processes	▪A single business process
▪ Structure for corporate view of data	▪Structure to suit the department
▪ Takes months to years for impln	▪Takes months for impln
▪ Size varies from 100 GB to a few TB	▪Size < 100 GB
▪ Flexible query & analysis	▪Restrictive query & analysis
▪ Technology optimal for holding & managing massive amount of data	▪Technology optimal for data access & analysis

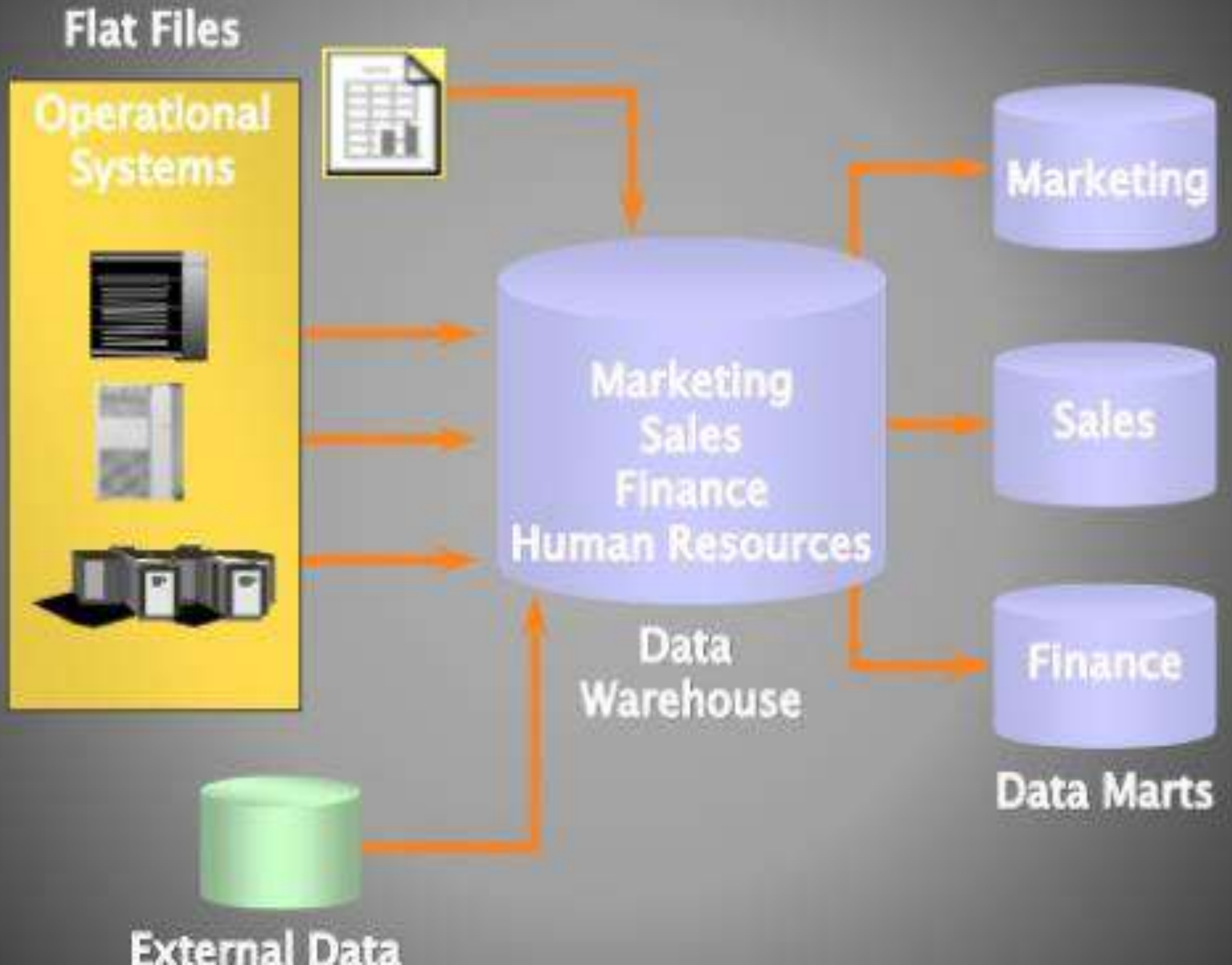
Reasons for creating Data Marts

- To enable access to the data that department needs to analyze most often
- To improve end-user response time
- To provide data in a form that matches the collective view of the data by a group of users
- Data marts use less data. So tasks like data cleansing, loading, transformation & integration are much easier & Implementing a data mart is simpler
- Requires less cost in impln in comparison with DWH

Types of data mart

1. **Dependent data mart**
2. **Independent data mart**

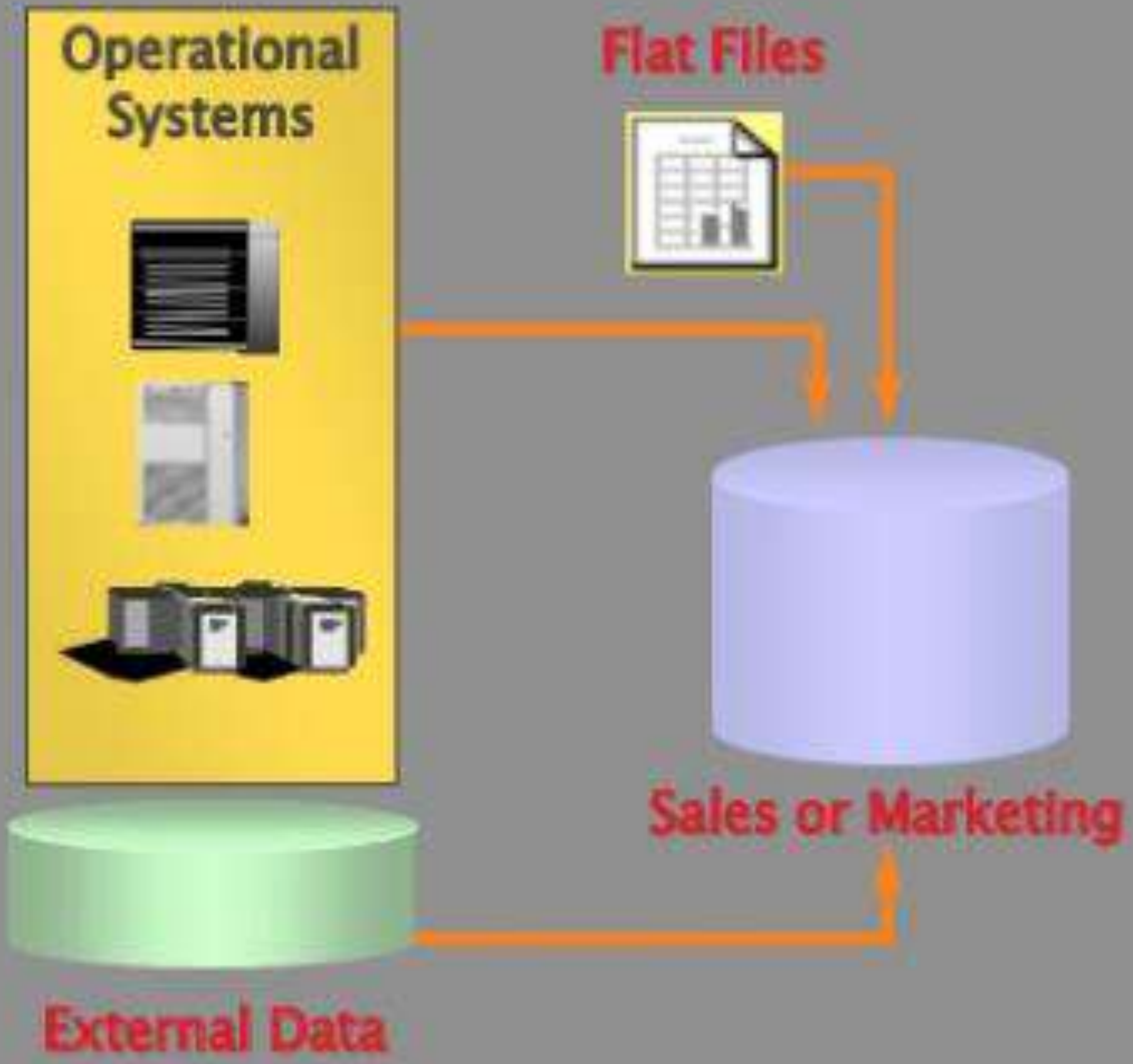
Dependent Data Mart



Dependent Data Mart

- A dependent data mart is one that takes the data feed from DWH
- All dependent data marts are fed from the same source as the DWH
- Dependent data marts are created with a subset of info in the DWH
- These data marts are easier to use because they have only the info that the specific user group within that particular dept. needs
- These data marts are architecturally more sound & stable than the independent data marts

Independent Data Mart



Independent Data Mart

- Independent data mart is one that depends upon applications env. for its data source
- Each independent data mart is fed separately by the operational systems applications

Disadvantages :

- Several source systems need to be handled to get the data content
- Additional effort & time is needed to clean, transform & integrate the data
- There are additional complexities involved for flexibility, reliability & maintenance of data
- There are problems in maintaining data consistency

Advantages of a data mart

- **Cost** is low
- **Implementation time** is short, often < 90 days
- They are controlled locally rather than centrally
- They contain less info than DWH & hence have more rapid response & are more easily understood
- They allow a business unit to build its own decision support systems without relying on others

Limitations of a data mart

- Performance degradation occurs as the size of the data mart increases
- Administration of multiple data marts becomes difficult
- Problems in building & implementing multiple data marts arise

Building Data Marts

There are two main approaches:

1. **Top-down approach**
2. **Bottom-up approach**

1. In the **top-down approach**, the **DWH project team** looks at the larger picture of the org. & builds a huge DWH first that will feed the individual data marts

2. In the **bottom-up approach**, the DWH project team caters to the requirements of individual dept. & builds data marts first that will feed data to the corporate wide DWH

    - In this approach, the departmental data marts are created first

Top-Down Approach: Advantages

- A truly corporate effort, an enterprise view of data
- Single, central storage of data about the content
- Centralized rules & control
- May see quick results if implemented with iterations

Disadvantages

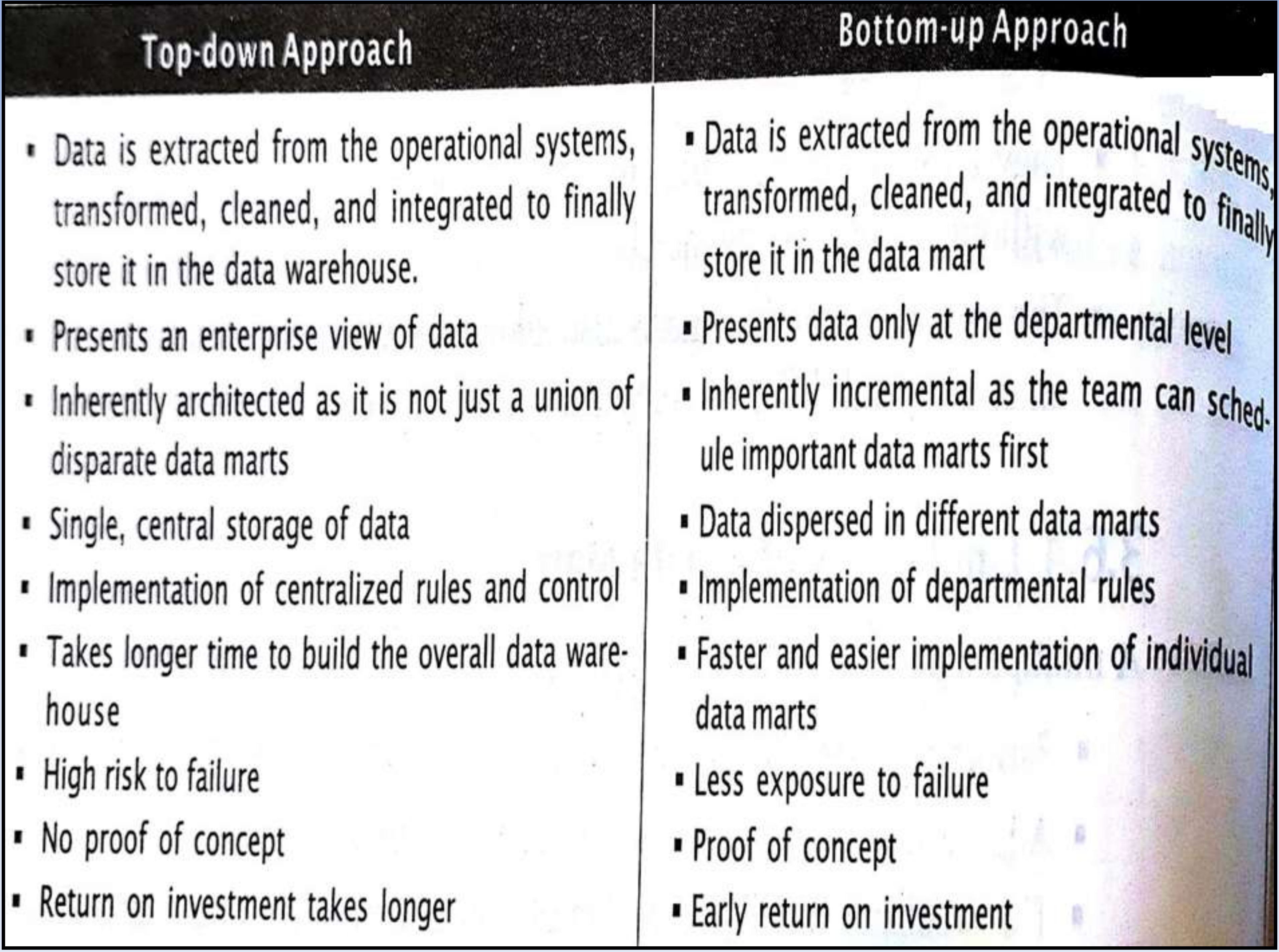
- Take longer to build even with an iterative method
- High exposer & risk to failure
- Needs high level of cross-functional skills

Bottom-Up Approach: Advantages

- Faster & easier implementation of manageable pieces
- Favorable return on investment & proof of concept
- Less risk of failure
- **Inherently incremental**: can schedule important data marts first
- Allows project team to learn & grow

Disadvantages

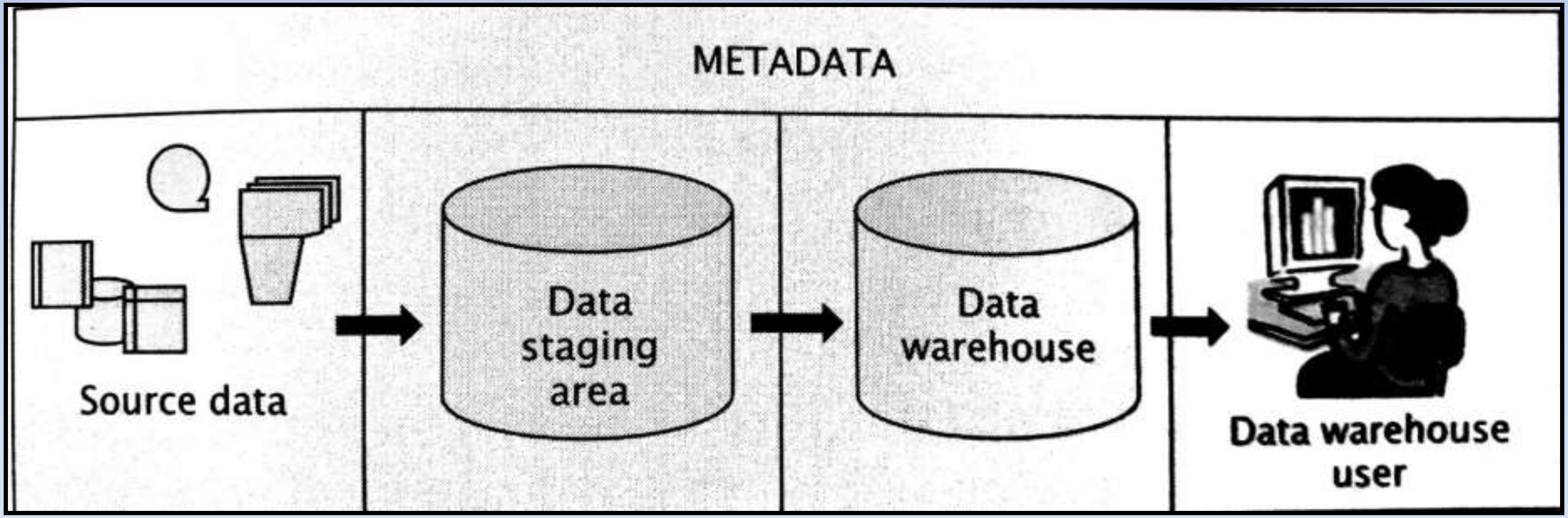
- Each data mart has its own narrow view of data
- Redundant data in every data mart
- Inconsistent & irreconcilable data
- Unmanageable interfaces



Top-down Approach	Bottom-up Approach
Data is extracted from the operational systems, transformed, cleaned, and integrated to finally store it in the data warehouse.	Data is extracted from the operational systems, transformed, cleaned, and integrated to finally store it in the data mart
Presents an enterprise view of data	Presents data only at the departmental level
Inherently architected as it is not just a union of disparate data marts	Inherently incremental as the team can schedule important data marts first
Single, central storage of data	Data dispersed in different data marts
Implementation of centralized rules and control	Implementation of departmental rules
Takes longer time to build the overall data warehouse	Faster and easier implementation of individual data marts
High risk to failure	Less exposure to failure
No proof of concept	Proof of concept
Return on investment takes longer	Early return on investment

info. flow Mechanism

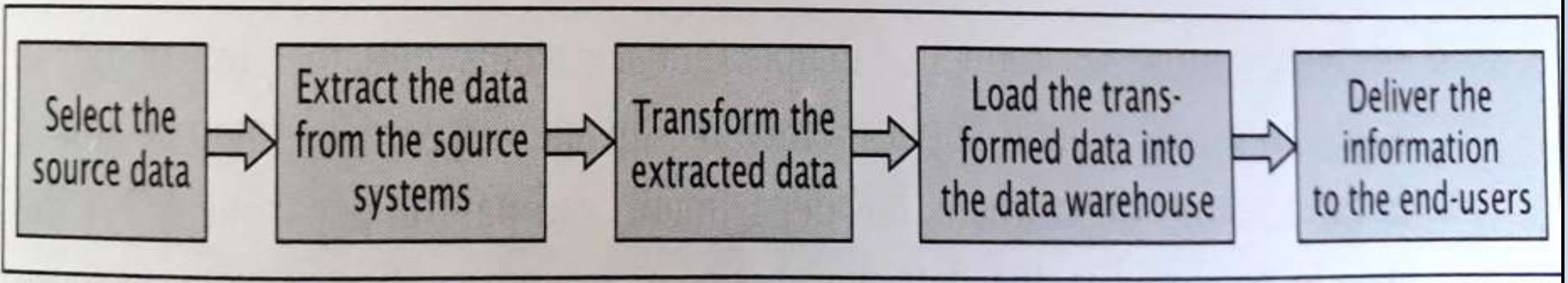
- How the **huge mountains of data** that exist in the **source system** get delivered to the **DWH users**????



The ETL Process

Steps of the process of transformation of data into info. are :

- **Select the source data**
- **Extract the data from the source systems**
- **Transform the extracted data**
- **Load the transformed data into the DWH**
- **Deliver the info. to the end-users**

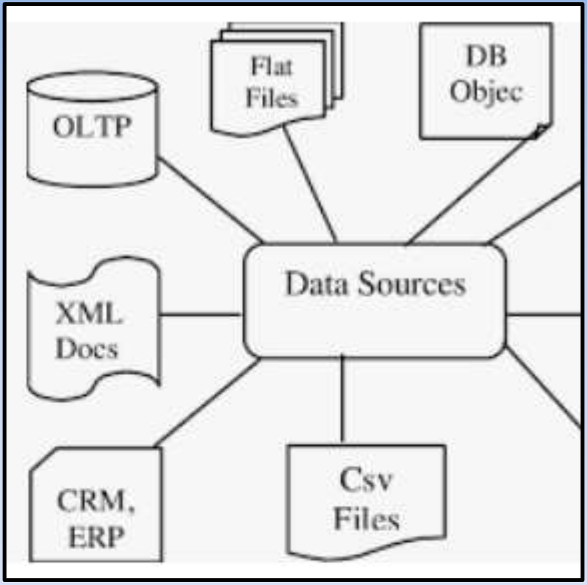


Select the source data → Extract the data from the source systems → Transform the extracted data → Load the transformed data into the data warehouse → Deliver the information to the end-users

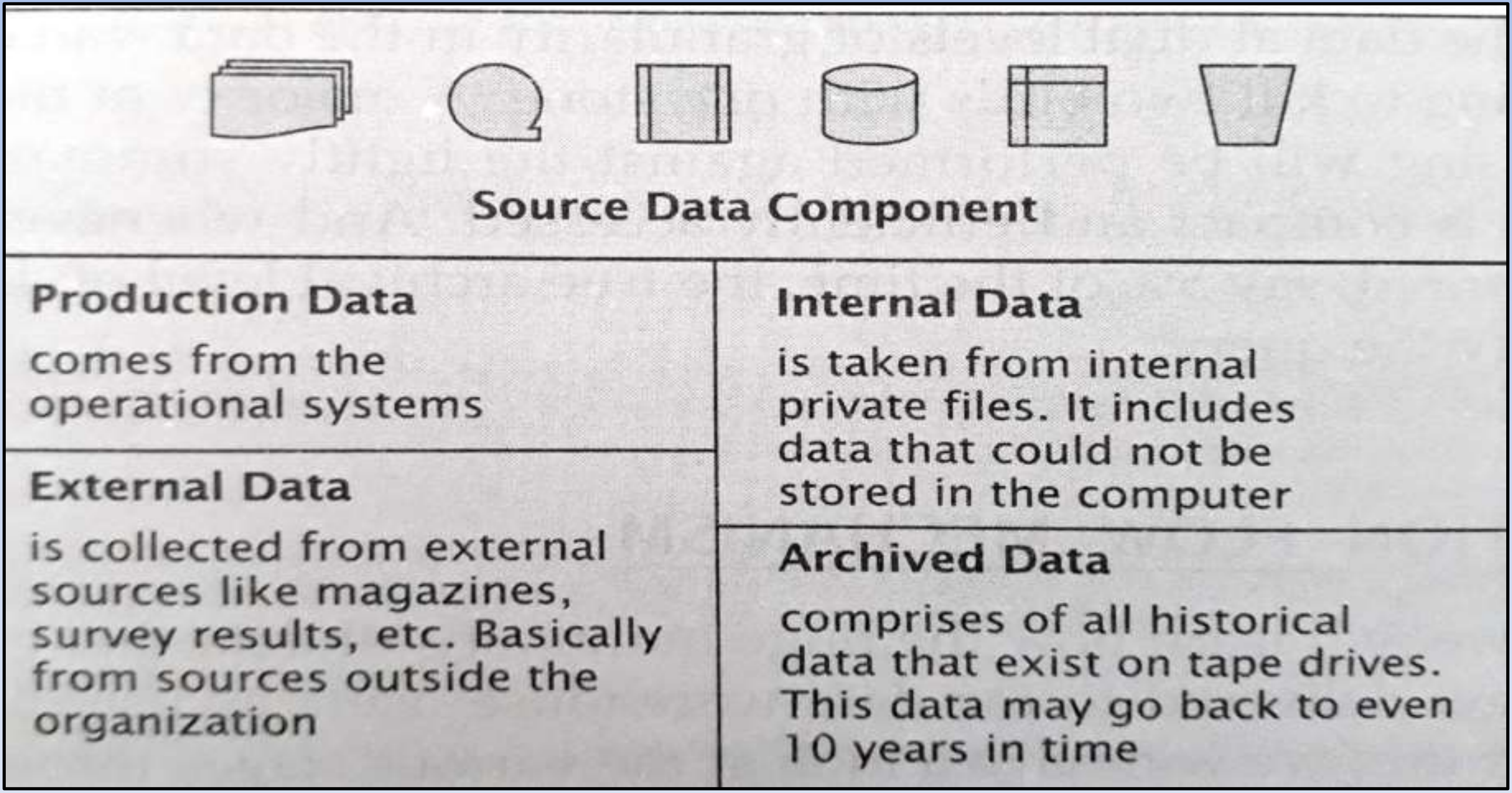
Figure 2.9 The ETL process

Select the source data

- **Source data coming into the DWH is grouped into 4 broad categories :**
- **Production Data**
- **Internal Data**
- **External Data**
- **Archived Data**



Categories of source data



Source Data Component

Production Data

comes from the operational systems

External Data

is collected from external sources like magazines, survey results, etc. Basically from sources outside the organization

Internal Data

is taken from internal private files. It includes data that could not be stored in the computer

Archived Data

comprises of all historical data that exist on tape drives. This data may go back to even 10 years in time

Extract data from the source systems

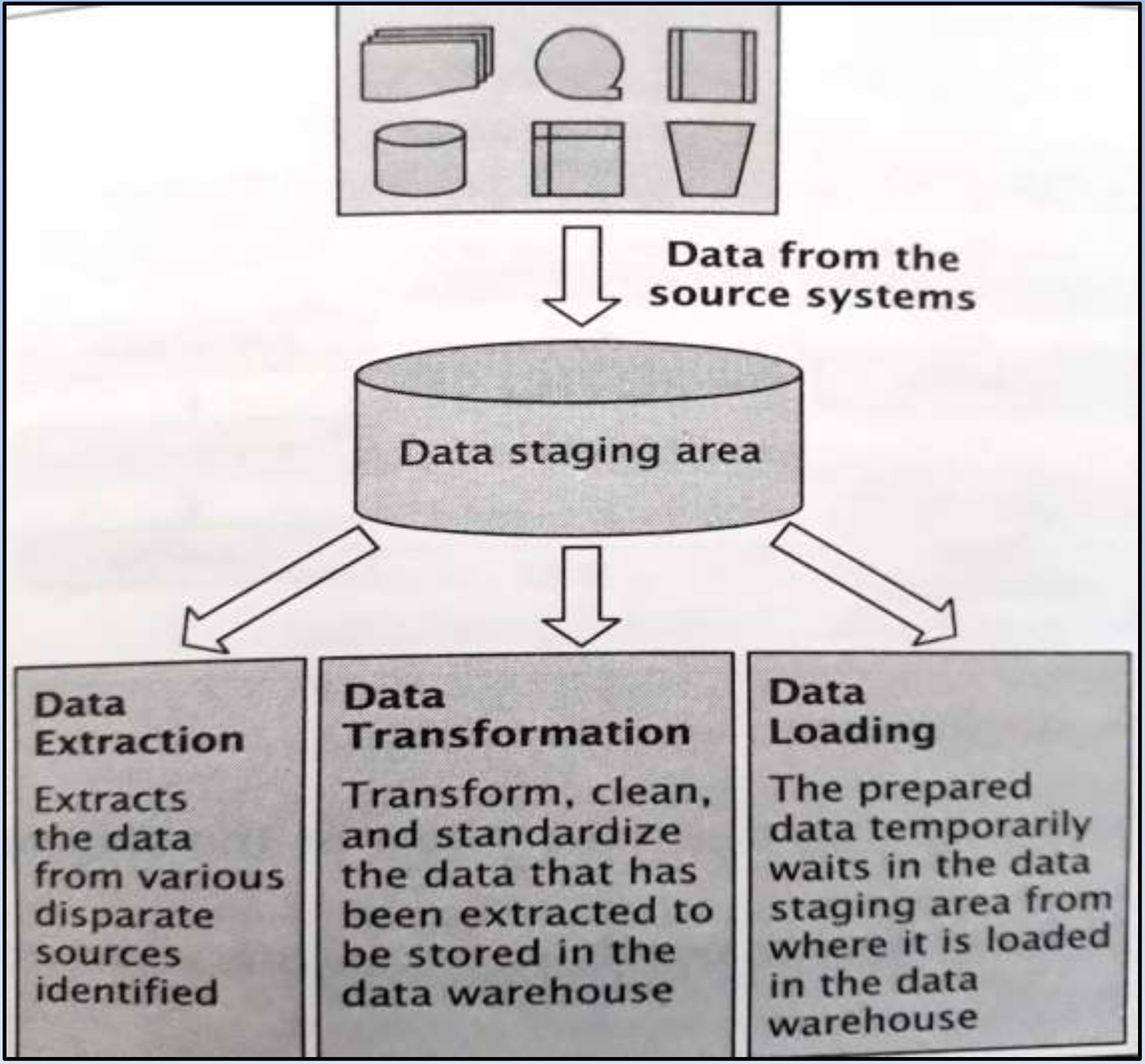
- **The data extraction process has to deal with multiple data sources**

- Since the source data are inconsistent, erroneous & stored in multiple formats, the extracted data is temporarily stored in the **data staging area** where all **data cleansing** & **transformations** are performed

Data Extraction process performs the following functions:

- Identify the sources of data
- Finalize the filters to be applied to every source system to extract the data
- Produce automatic extract files from the operational systems
- Generate intermediary files to store selected data to be merged later
- Reformat input from outside sources
- Reformat & standardize the input from departmental data files, databases & spreadsheets
- Produce common application code for data extraction
- Resolve inconsistencies for common data that will be extracted from multiple source systems

Data Staging Area



Data Extraction

Extracts the data from various disparate sources identified

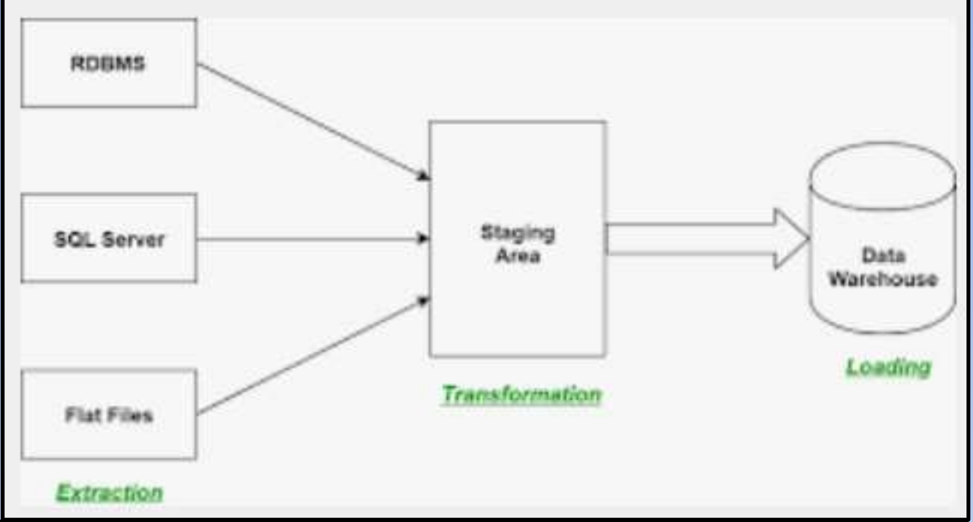
Data Transformation

Transform, clean, and standardize the data that has been extracted to be stored in the data warehouse

Data Loading

The prepared data temporarily waits in the data staging area from where it is loaded in the data warehouse

Data Staging Area



- **Data staging area** is the place where all the extracted data are temporarily stored & prepared for loading into the **DWH**

- It is the area where the extracted files are examined, business rules are reviewed, the data transformation functions are performed, data is sorted/merged, inconsistencies are resolved & the data are cleaned

Why Data Staging area?

- This approach **isolates** the **raw data** extracted from a no. of sources from the **processed data**

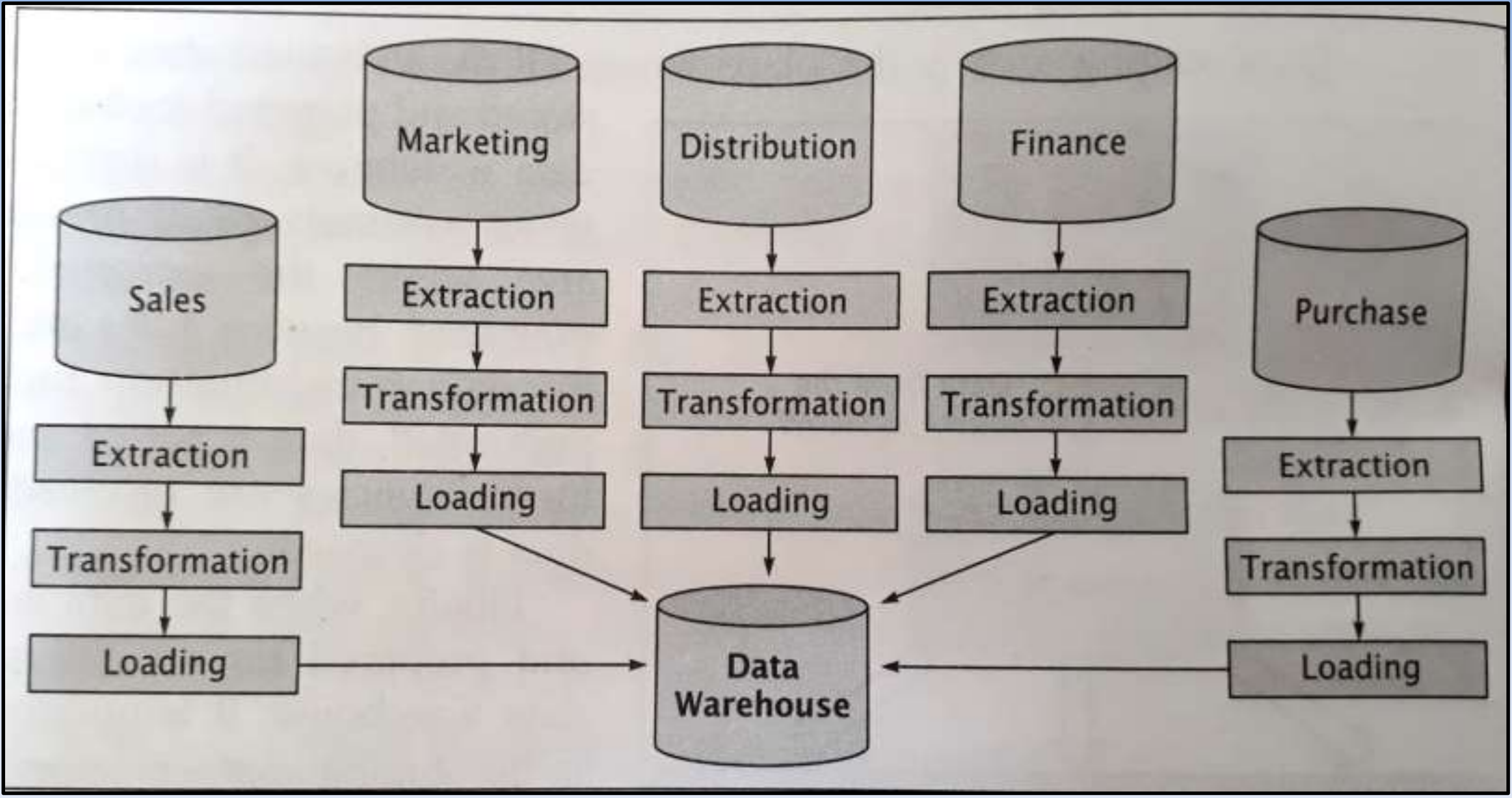
- As the **DWH** users are not supposed to access the staging area, it offers additional **security** & **process quality**

- It helps in sharing the load as ‘**data preparation**’ tasks & ‘**DWH querying tasks**’ are handled by separate systems

- It eases the development of **central metadata repository** which maintains documentation for all involved systems

Data preprocessing at the staging area

- **The main issue is:** in a DWH, you pull in the data from many source systems **& store it based on subjects** not by applications
- Data in a **DWH** is **subject-oriented** & cuts across applications
- A **separate staging area is compulsory** for preparing data for the **DWH**



Types of raw data processing that take place at the staging area:

- **Standardization of data**: data is transformed into a standard format
- **Sorting** of records
- **Comparing** & **merging** of records that belong to the same object but are derived from different sources
- **Aggregation** & **summarization** of data
- **Filling missing values** with default values
- **Converting data** according to technological platform used by the **DWH server**

Transform the Extracted data

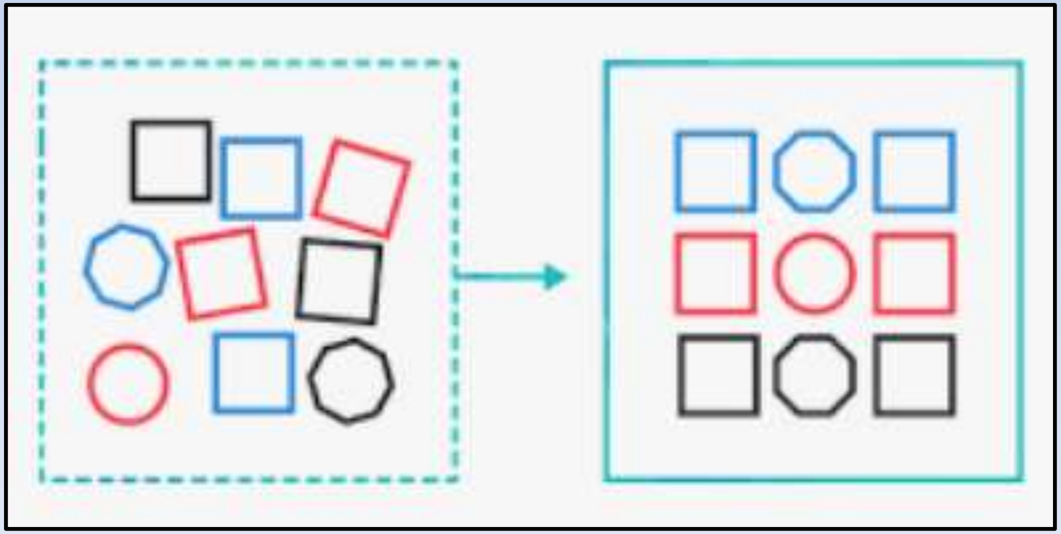
- The data extracted from the source system **can't be stored directly** in the DWH

- Before moving the extracted data into the DWH, various types of **data transformation** have to be performed

- Since this data come from **several dissimilar source systems**, there is a need to transform the data according to a **standard format**

- Ensured that the data do **not violate any business rule**

- Tasks performed as a part of data transformation
  - **Data cleaning:** includes misspelling corrections, conflicts resolution between data elements, providing default values for missing data & elimination of duplicates
  - **Standardization of data elements:** the data types & field lengths of same data elements from various sources are standardized
  - **Semantic standardization:** synonyms & homonyms are resolved



Load Transformed data into Data Warehouse

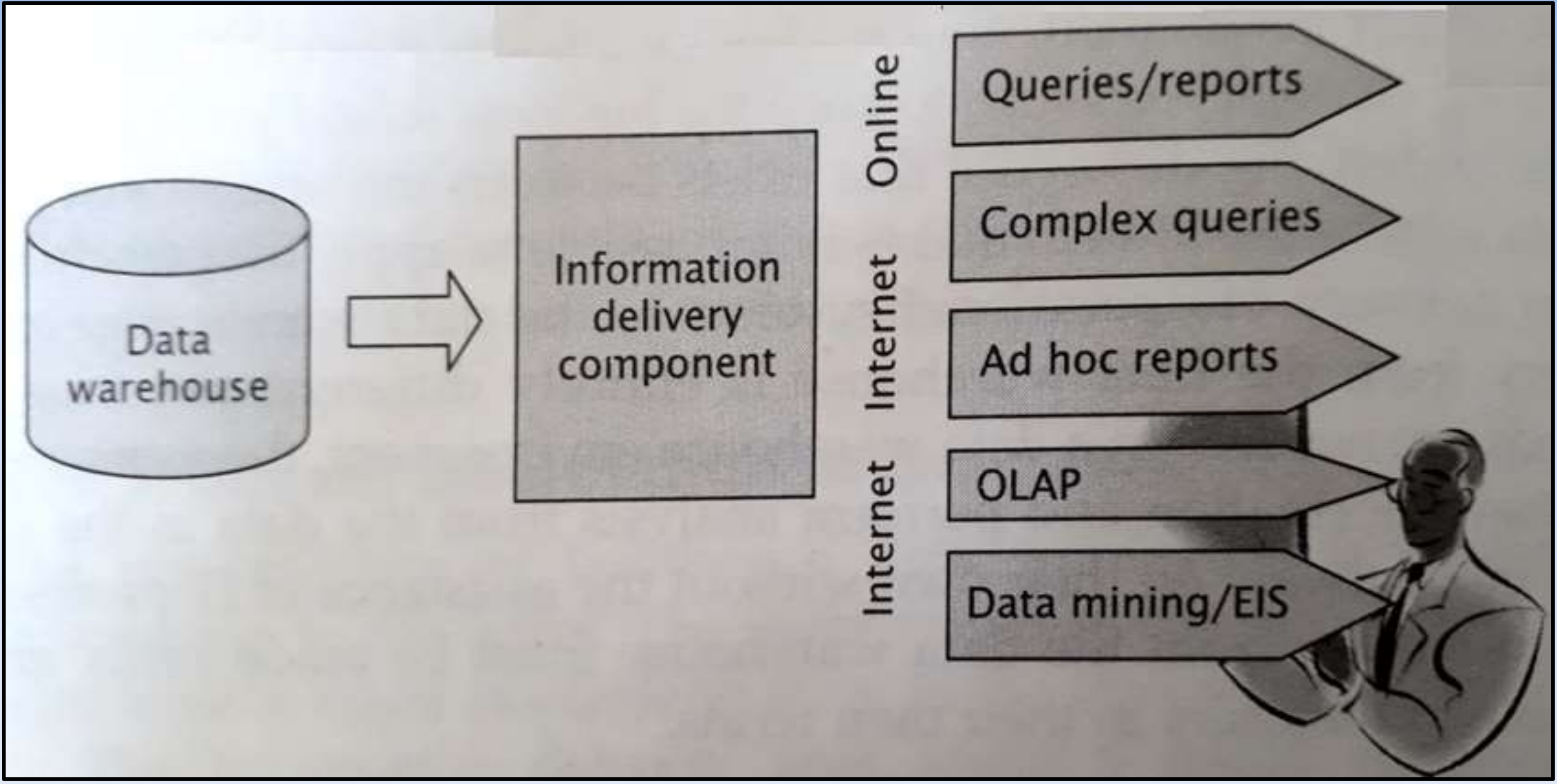
- Once the data transformation ends & we have a collection of integrated, cleaned, standardized & summarized data,
  - **Data is ready to be loaded into the DWH**

- **Two categories of tasks form the data loading :**
  - When the design & construction of DWH is completed for the 1st time, the **initial loading** of the data into the **DWH** is done. The initial load **moves large volumes of data** consuming a lot of time
  - Once the initial loading is over, the DWH is **constantly updated** to add new records

Deliver info. to end-users

- The **info. delivery system** is responsible for distributing the data stored in the warehouse to its end-users

- To satisfy the informational needs of a wide range of users, the info. delivery component includes different methods of info. delivery



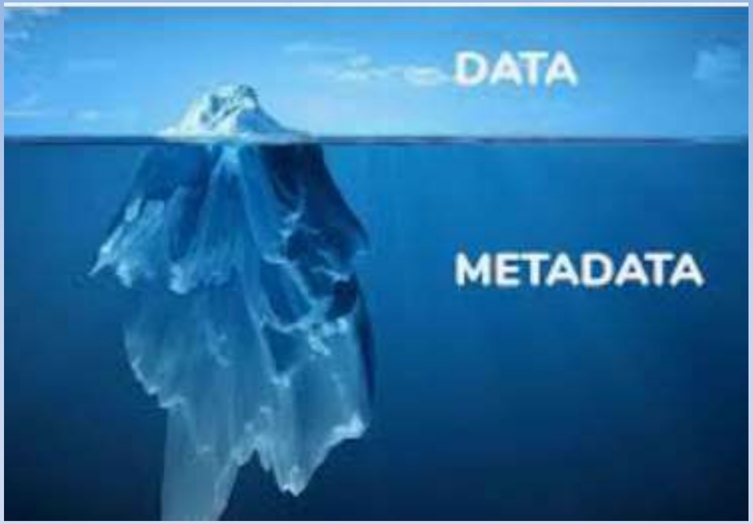
METADATA

- A **metadata** gives the description of the **entity** & other details explaining the **syntax** & **semantics** of the **data elements**
- **Metadata** describes all aspects of the data in the DWH precisely to help both the users & the developers of the DWH

A typical metadata contains info. about the following:

i. **Structure of data** from **programmer's** perspective

ii. **Structure of data** from **end user's** perspective



i. **Source systems** that feed the DWH

ii. **Transformation process** applied on the source system data

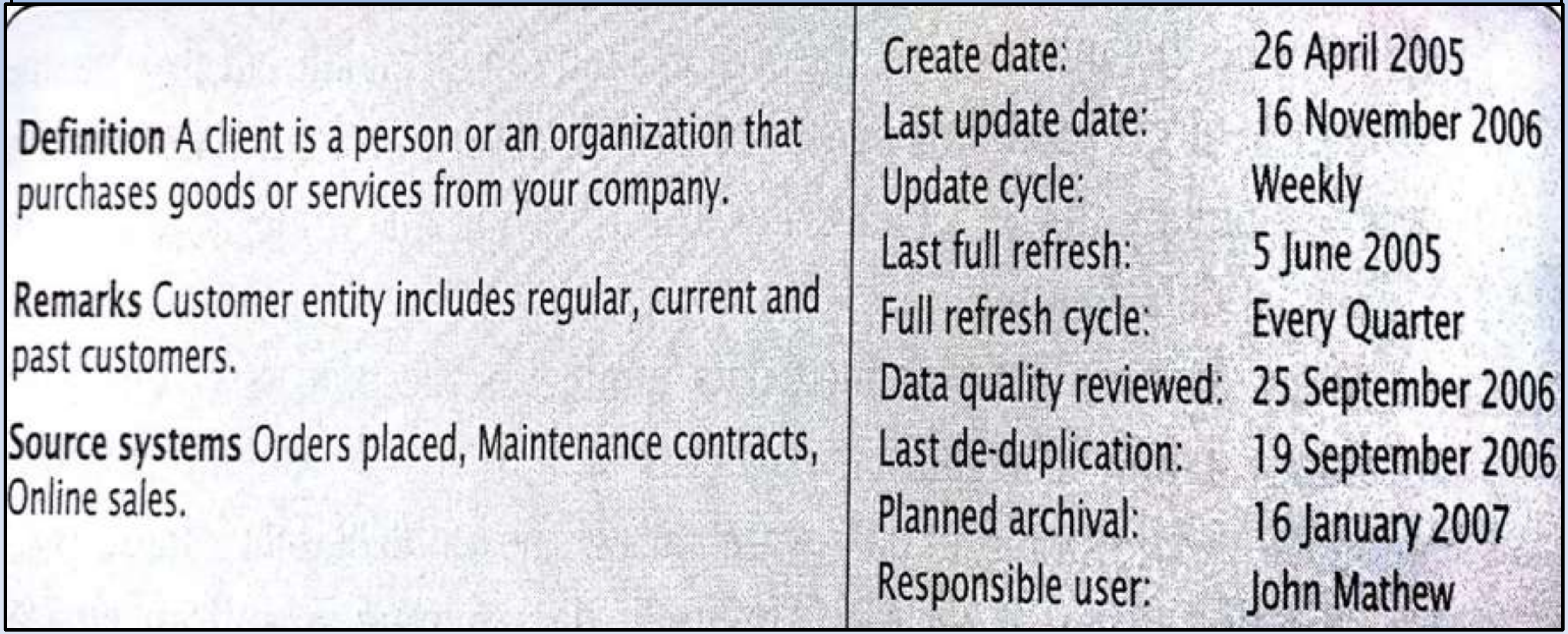
iii. **History** of **data extraction process**

iv. **Data model**

An Example of Metadata

- If the user wants to know about **customer entity** in the DWH, then he will search for this info. in the **metadata repository**

- A sample look of how the details are stored about the **customer entity** is given in the next slide



Definition A client is a person or an organization that purchases goods or services from your company.

Remarks Customer entity includes regular, current and past customers.

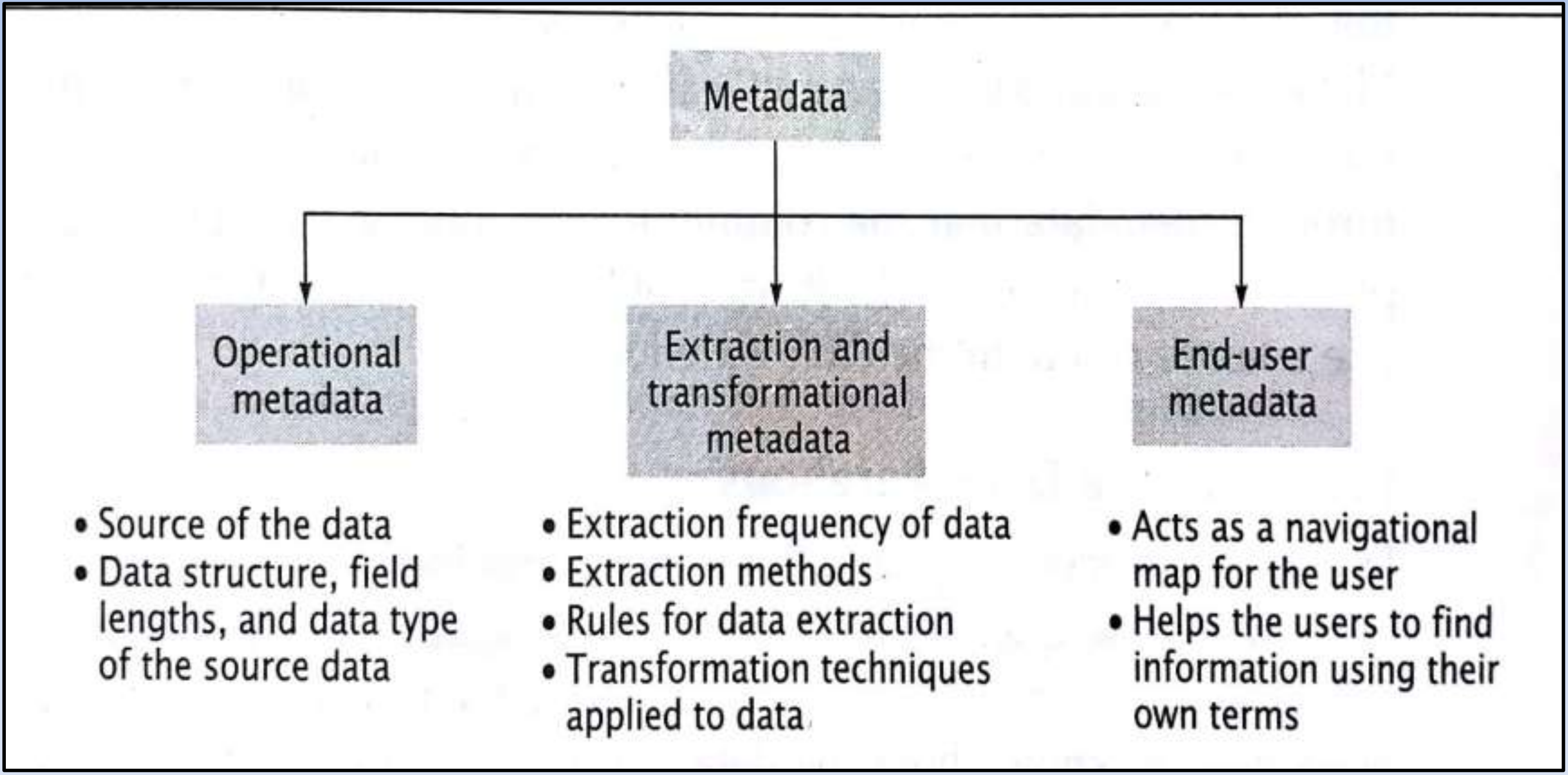
Source systems Orders placed, Maintenance contracts, Online sales.

Create date:	26 April 2005
Last update date:	16 November 2006
Update cycle:	Weekly
Last full refresh:	5 June 2005
Full refresh cycle:	Every Quarter
Data quality reviewed:	25 September 2006
Last de-duplication:	19 September 2006
Planned archival:	16 January 2007
Responsible user:	John Mathew

Role of Metadata

- **Metadata** in the **DWH** is similar to the **data dictionary** in a **DBMS**
- The **metadata** stores **data about the data** in the **DWH**
- It is used for building, maintaining, managing & using the **DWH**
- It is the key for providing users & developers with a **road map** to the info. in the warehouse
- **The three main functions performed by metadata :**
  1. **Connects** the different parts of the **DWH**
  2. Provides info. about the **contents** of the data & its **underlying structure**
  3. Enables end-users to **search** for desired data in their own terms

Classification of Metadata



Metadata

Operational metadata

- Source of the data
- Data structure, field lengths, and data type of the source data

Extraction and transformational metadata

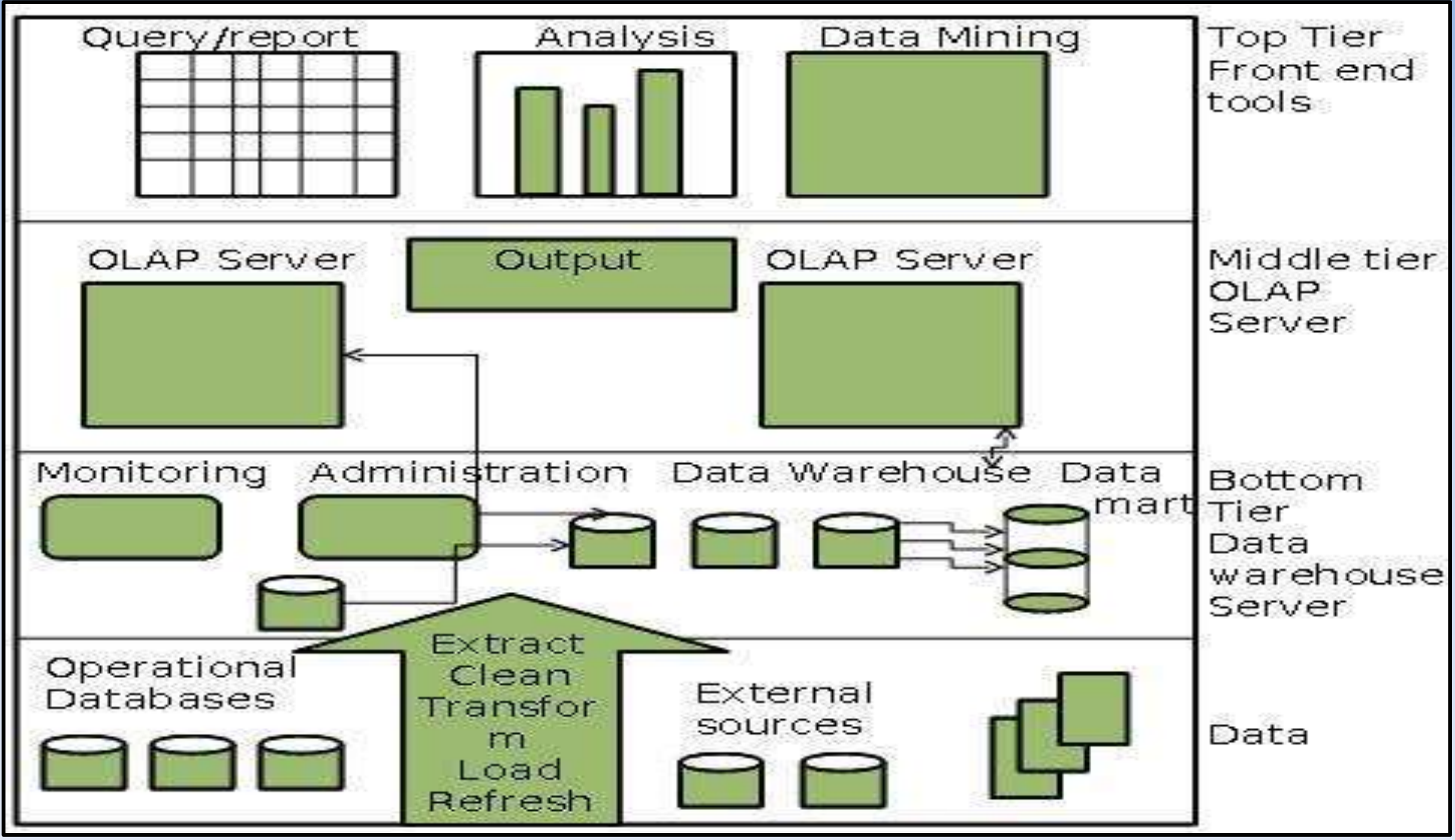
- Extraction frequency of data
- Extraction methods
- Rules for data extraction
- Transformation techniques applied to data

End-user metadata

- Acts as a navigational map for the user
- Helps the users to find information using their own terms

Data Warehousing Architecture

- **DWH architecture** is a way of representing the **overall structure of the data, processing & presentation** that exists for end-user computing within the org.
- **Three tier architecture :**



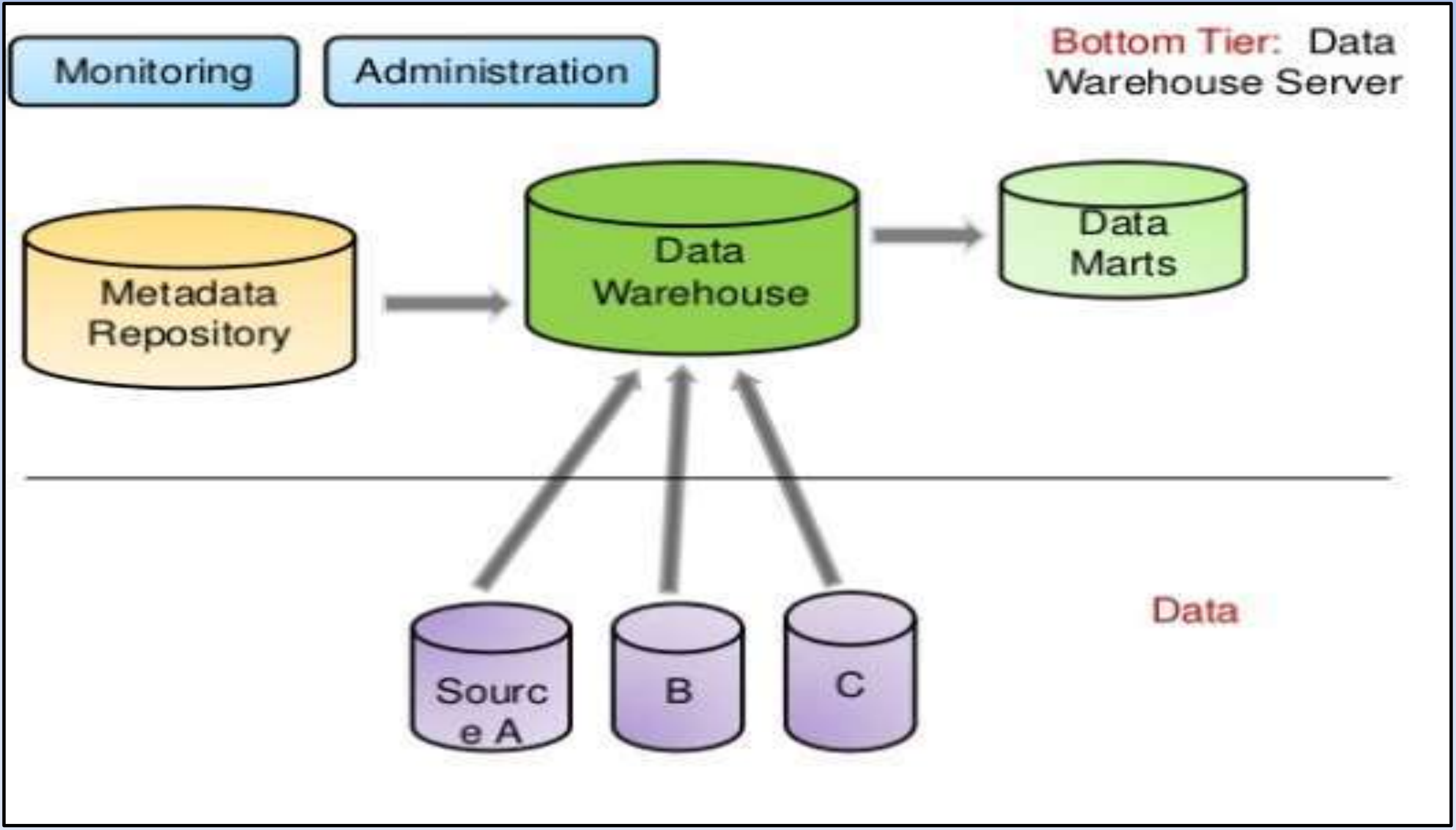
Bottom tier: Data Warehouse Server

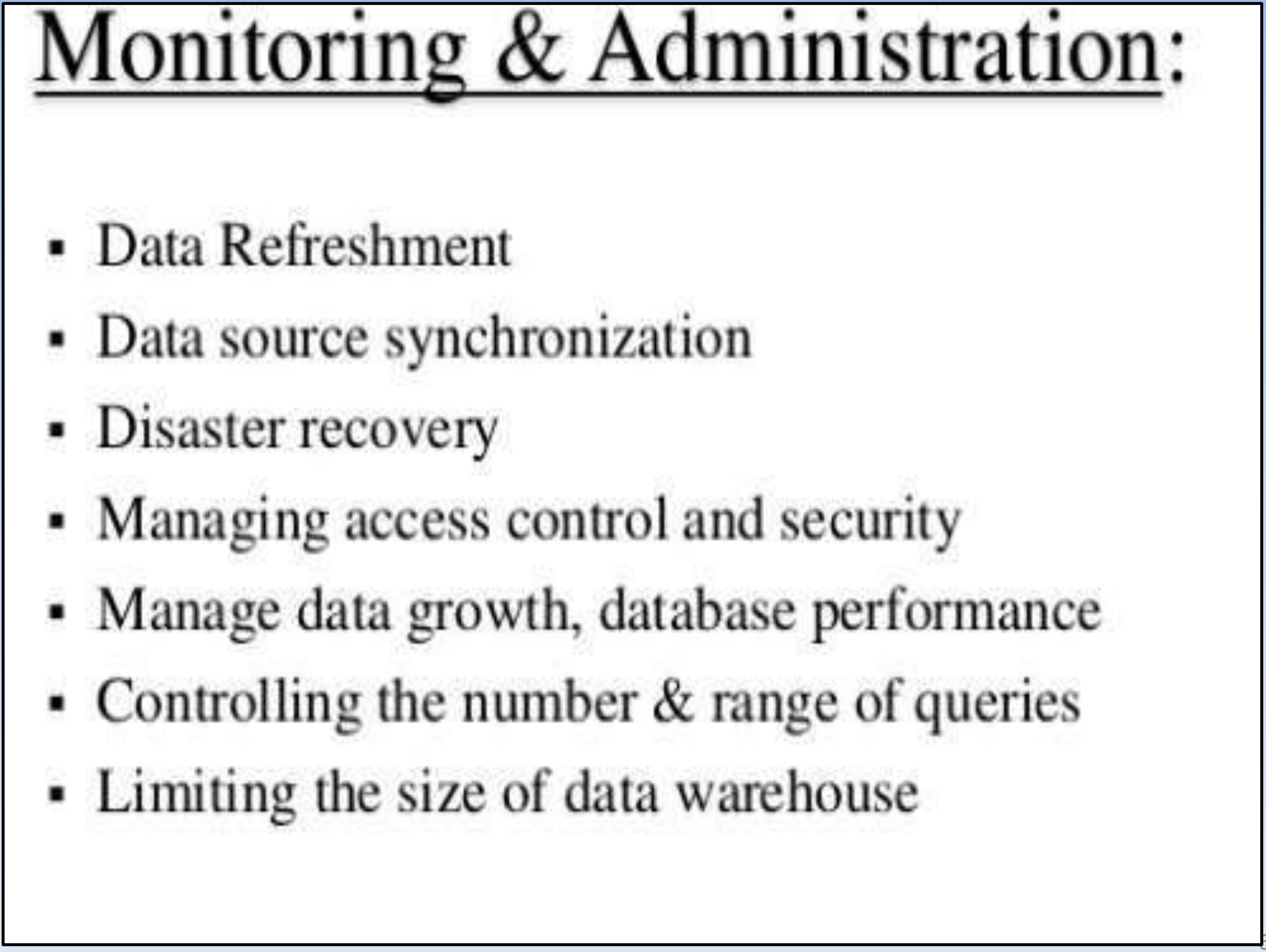
- **Bottom tier** is a ***warehouse database server*** that is mostly a RDBMS
- **Back-end tools** & **utilities** are used to **feed data** into the **bottom tier** from operational databases or other external sources
- These **tools** perform **data extraction, cleaning** & **transformation** as well as **load** & **refresh** functions to update the DWH
- This tier also contains a **metadata repository** which stores info. about the DWH & its contents

Bottom tier contains

- **DWH**
- **Meta data repository**
- **Data marts**
- **Monitoring & administration**

Bottom tier


Monitoring
Administration
Bottom Tier: Data Warehouse Server
Metadata Repository
Data Warehouse
Data Marts
Sourc e A
B
C
Data

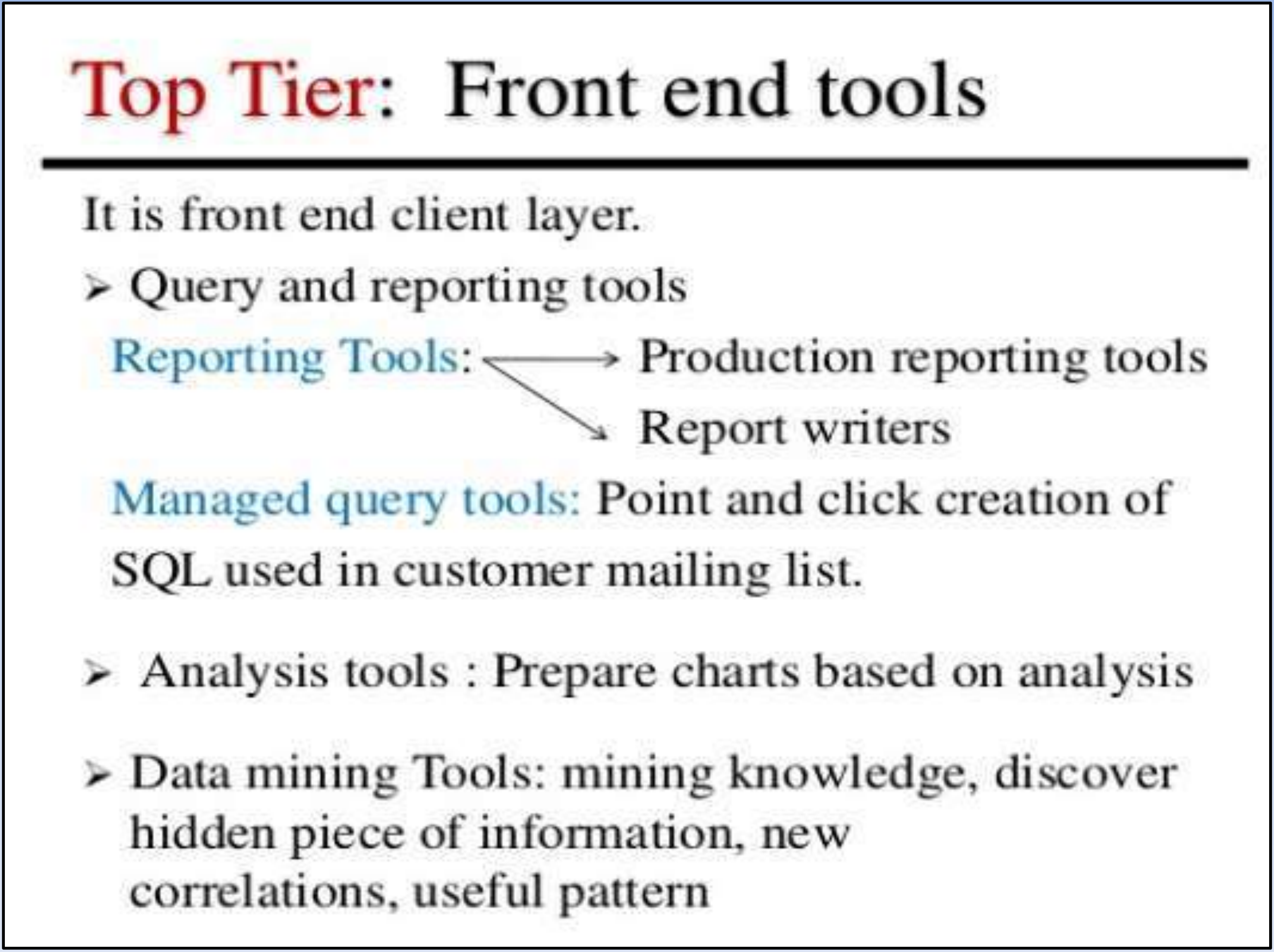


Monitoring & Administration:

- Data Refreshment
- Data source synchronization
- Disaster recovery
- Managing access control and security
- Manage data growth, database performance
- Controlling the number & range of queries
- Limiting the size of data warehouse

Middle tier : OLAP Server

- The middle tier is an **OLAP Server**
- It is implemented using a **ROLAP** or a **MOLAP** model
- A ***Relational OLAP (ROLAP)*** model is an extended RDBMS that maps operations on multidimensional data to standard relational operations
- A ***Multidimensional OLAP (MOLAP)*** model is a special-purpose server that directly implements multi dimensional data & operations



Top Tier: Front end tools

It is front end client layer.

- Query and reporting tools

Reporting Tools: → Production reporting tools

→ Report writers

Managed query tools: Point and click creation of SQL used in customer mailing list.

- Analysis tools : Prepare charts based on analysis
- Data mining Tools: mining knowledge, discover hidden piece of information, new correlations, useful pattern

THAT'S
ALL FOLKS