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

Data Warehouse

Unit 1: Introduction

Unit No: 1

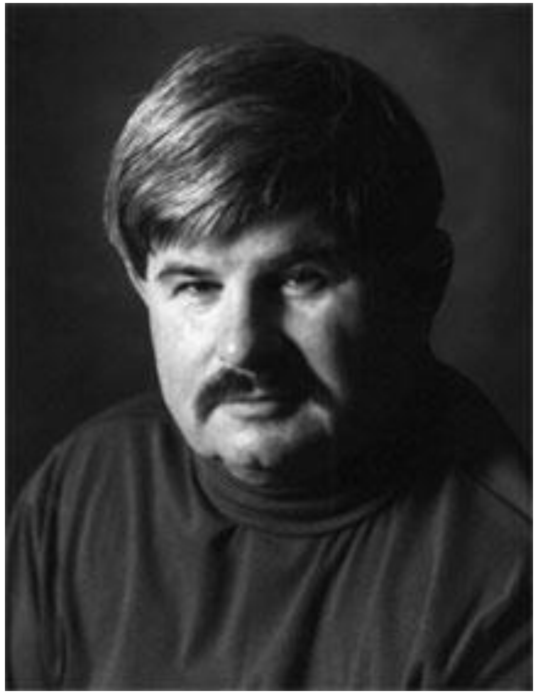
Unit name : Introduction to Data warehouse

Lecture No: 1
Overview, Definition

Inmon

Father of the data warehouse Co-creator of the Corporate Information Factory.

He has 35 years of experience in database technology management and data warehouse design.



Inmon-Cont'd

He has written about a variety of topics on the building, usage,& maintenance of the data warehouse & the Corporate Information Factory.

He has written more than 650 articles (Datamation, ComputerWorld, and Byte Magazine).

Inmon has published 45 books.

Many of books has been translated to Chinese, Dutch, French, German, Japanese, Korean, Portuguese, Russian, and Spanish.

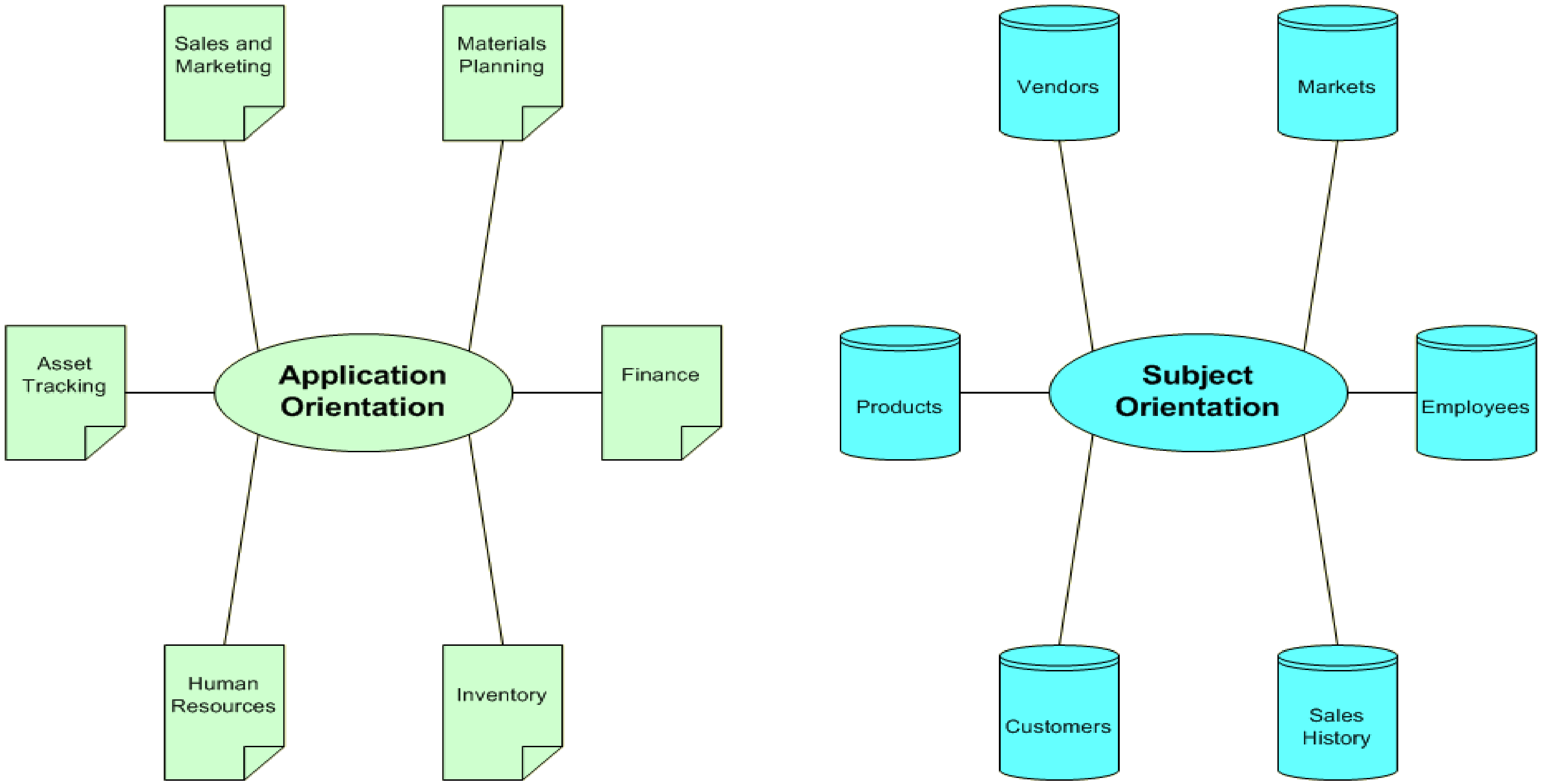
Introduction

- What is Data Warehouse?

  A data warehouse is a collection of integrated databases designed to support a DSS- Decision Support System.

- According to Inmon's (father of data warehousing) definition(Inmon,1992a,p.5):
  - It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is non-volatile and relevant to some moment in time.

A Data Warehouse is Subject Oriented



Subject Oriented

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

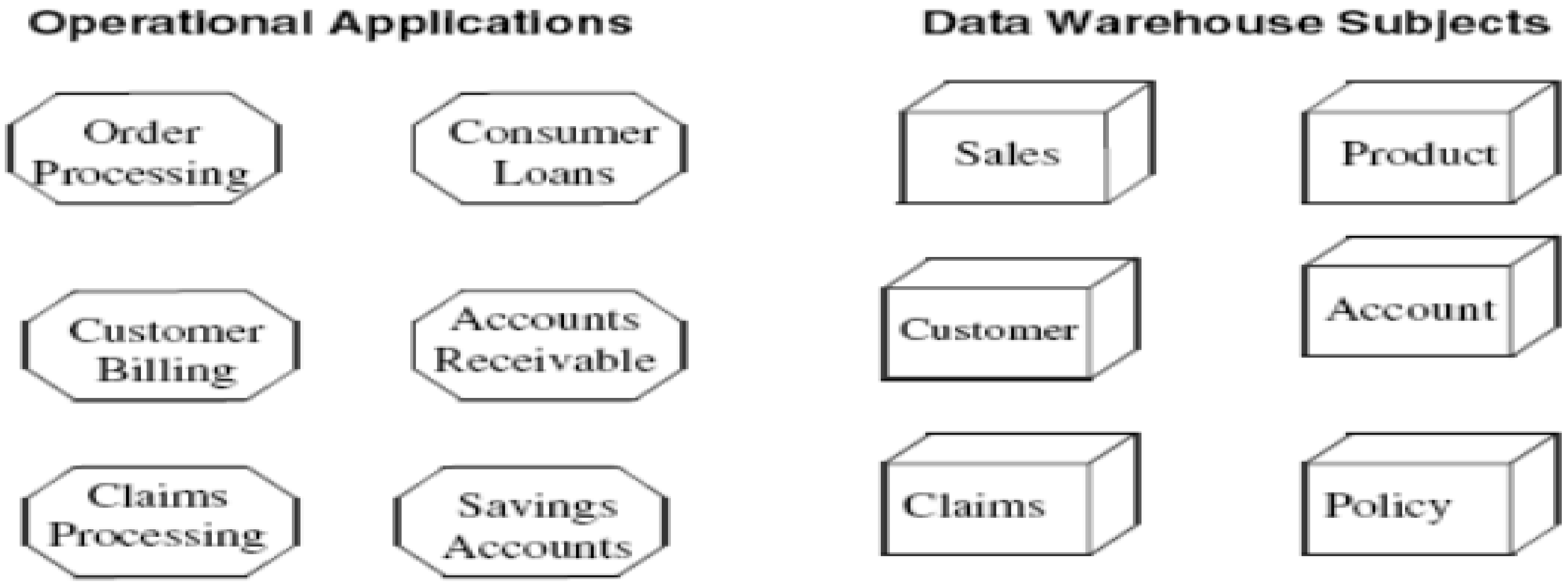


Figure 2-1 The data warehouse is subject oriented.

Introduction-Cont'd.

Where is it used?

It is used for evaluating future strategy.

It needs a successful technician:

Flexible.

Team player.

Good balance of business and technical understanding.

Introduction-Cont’d.

- The ultimate use of data warehouse is Mass Customization.

  For example, it increased Capital One’s customers from 1 million to approximately 9 millions in 8 years.

- Just like a muscle: DW increases in strength with active use.

  With each new test and product, valuable information is added to the DW, allowing the analyst to learn from the success and failure of the past.

- The key to survival:

  Is the ability to analyze, plan, and react to changing business conditions in a much more rapid fashion.

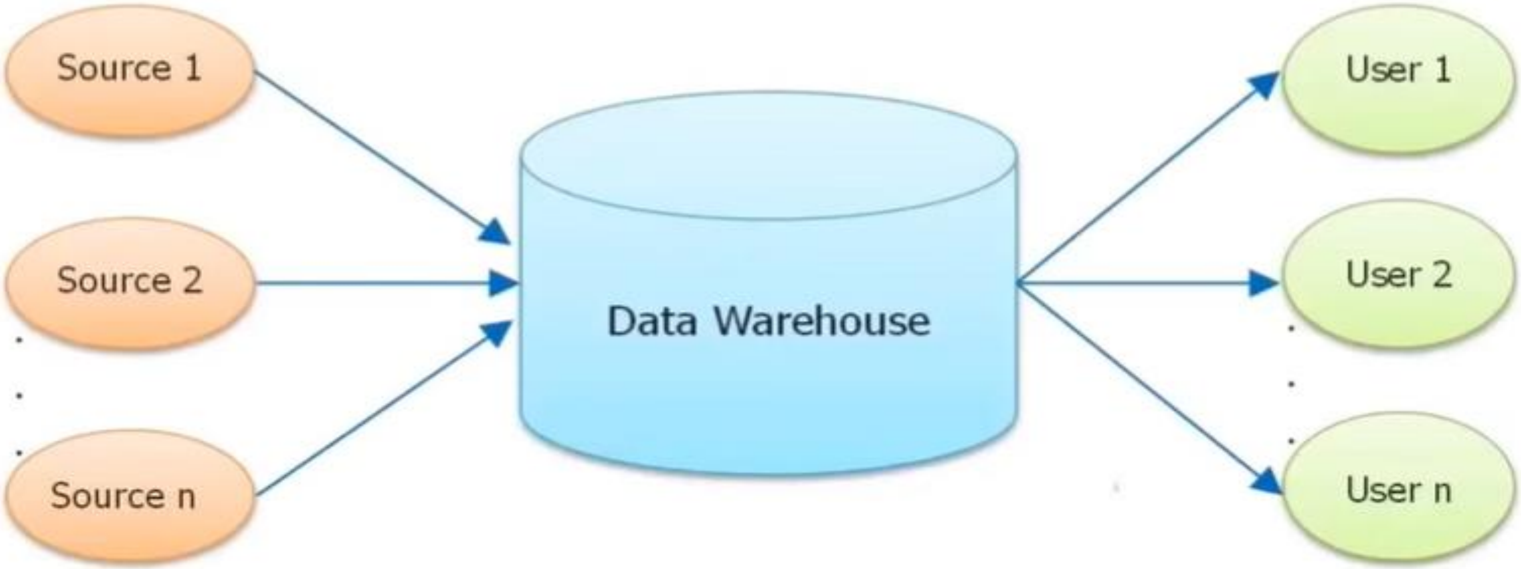
What is Data Warehouse ?

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

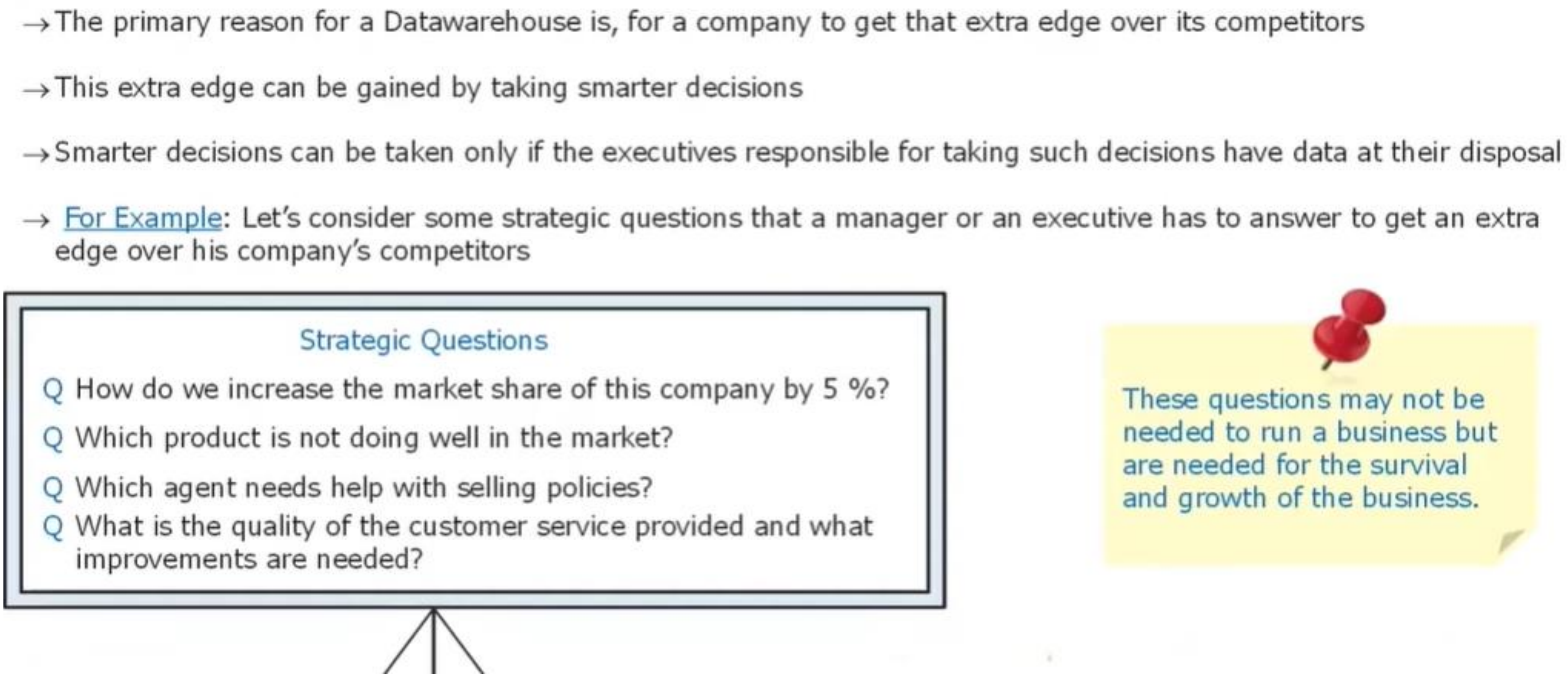
→The end user accesses it whenever he needs some information

→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?

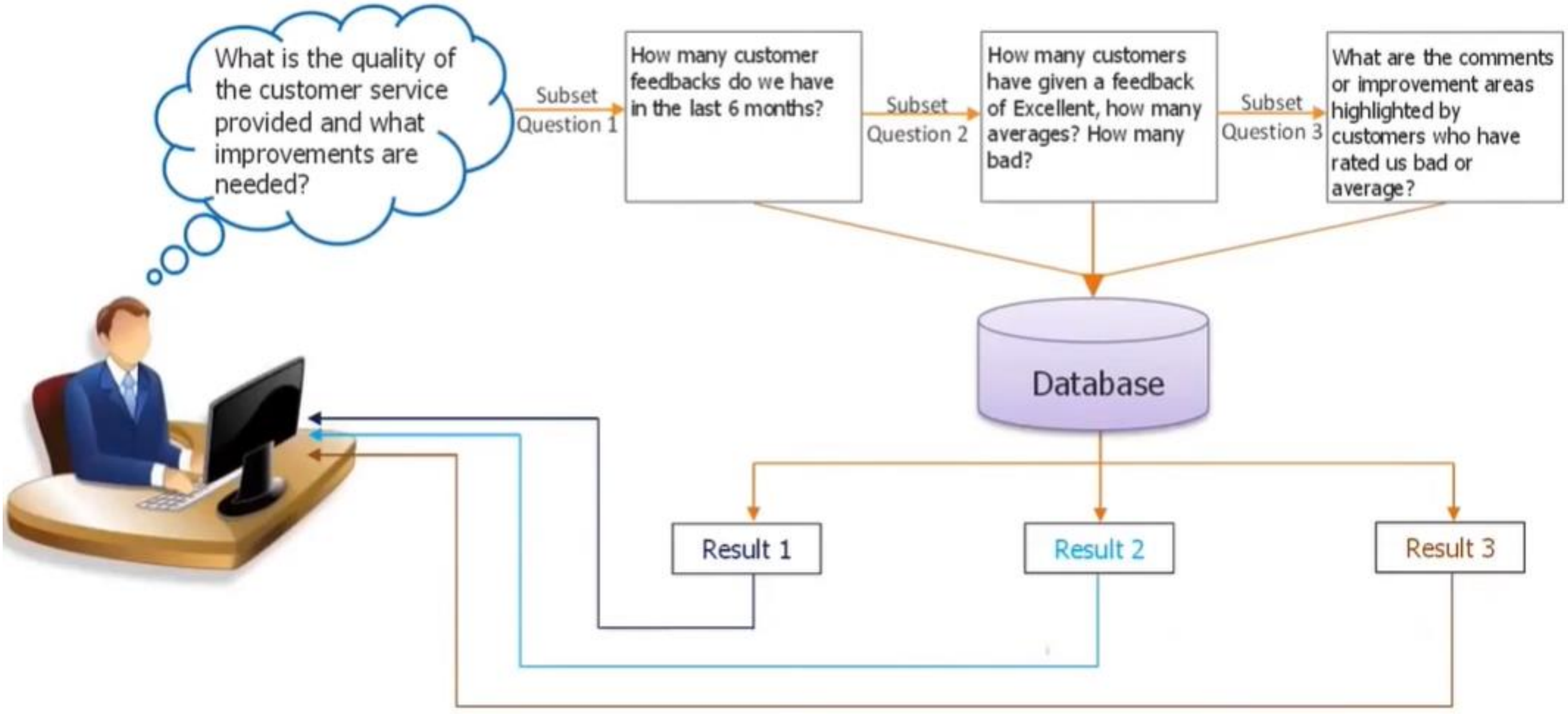
Q Which agent needs help with selling policies?

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?

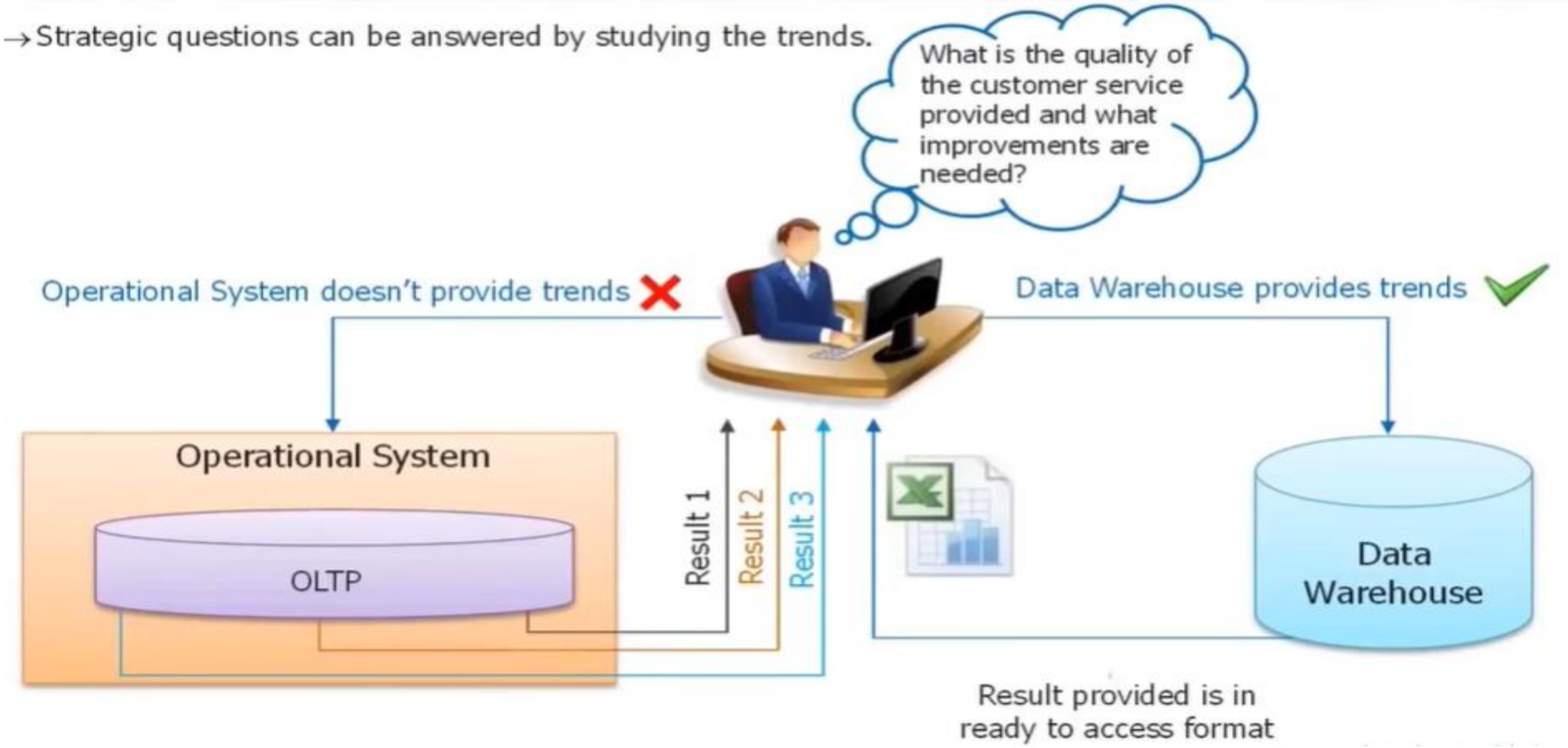
Database

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.”

Data Warehouse

- In order for data to be effective, DW must be:
  - Consistent.
  - Well integrated.
  - Well defined.
  - Time stamped.

- DW environment:
  - The data store, data mart & the metadata

The Data Store

- Its day-to-day function is to store the data for a single specific set of operational application.
- Its function is to feed the data warehouse data for the purpose of analysis.
- An operational data store (ODS) stores data for a specific application. It feeds the data warehouse a stream of desired raw data.
- Is the most common component of DW environment.
- Data store is generally subject oriented, volatile, commonly focused on customers, products, orders, policies, claims, etc…

The Data Mart

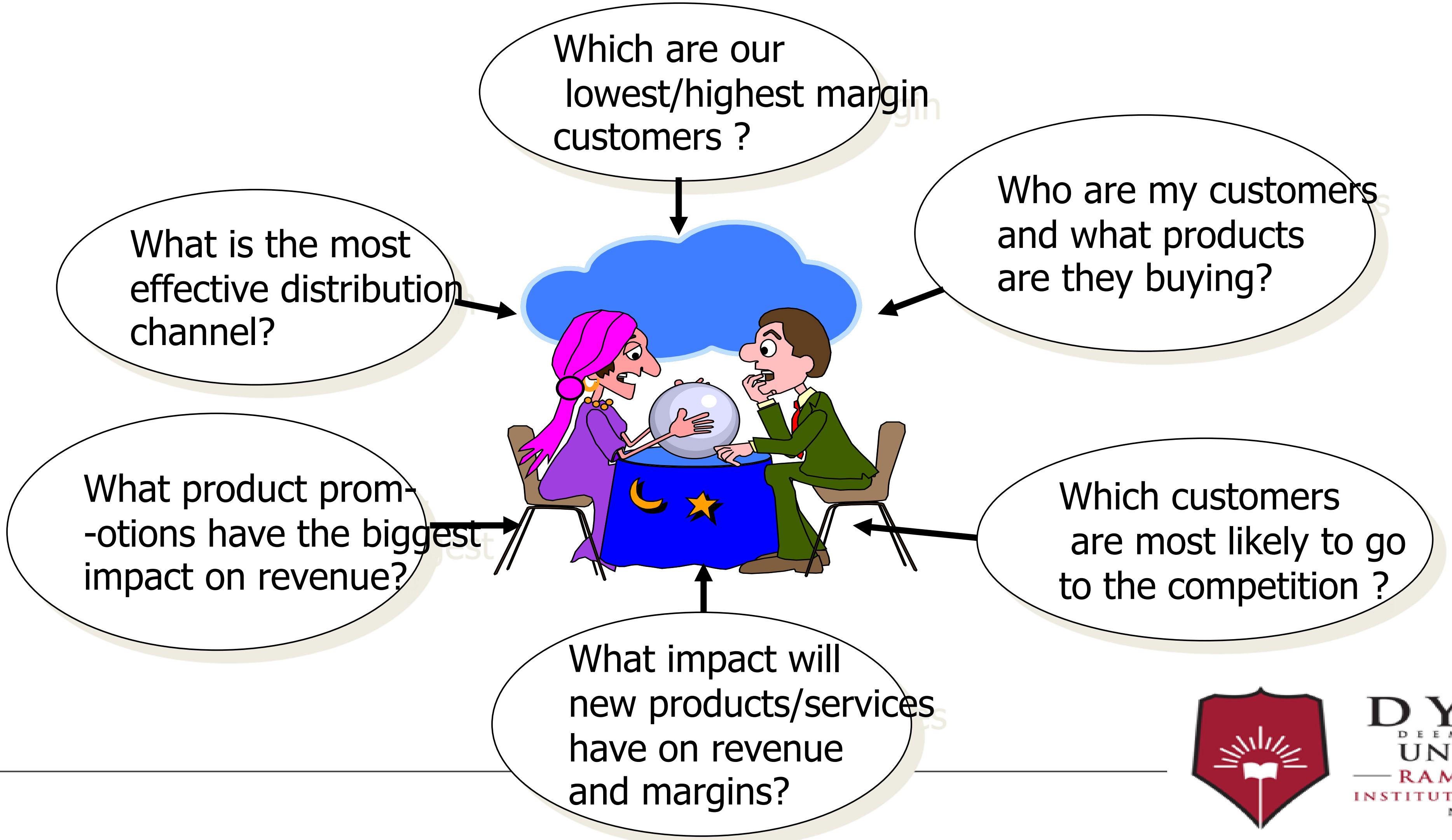
- It is lower-cost, scaled down version of the DW.

- Data Mart offer a targeted and less costly method of gaining the advantages associated with data warehousing and can be scaled up to a full DW environment over time.

Meta Data

- The name suggests some high-level technological concept, but it really is fairly simple. Metadata is "data about data".
- With the emergence of the data warehouse as a decision support structure, the metadata are considered as much a resource as the business data they describe.
- Metadata are abstractions -- they are high level data that provide concise descriptions of lower-level data.

A producer wants to know….Strategic information



Which are our lowest/highest margin customers ?

Who are my customers and what products are they buying?

What is the most effective distribution channel?

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?

Data, Data everywhere yet ...

- I can’t find the data I need

  data is scattered over the network

  many versions, subtle differences

I can’t get the data I need

need an expert to get the data

I can’t understand the data I found

available data poorly documented

I can’t use the data I found

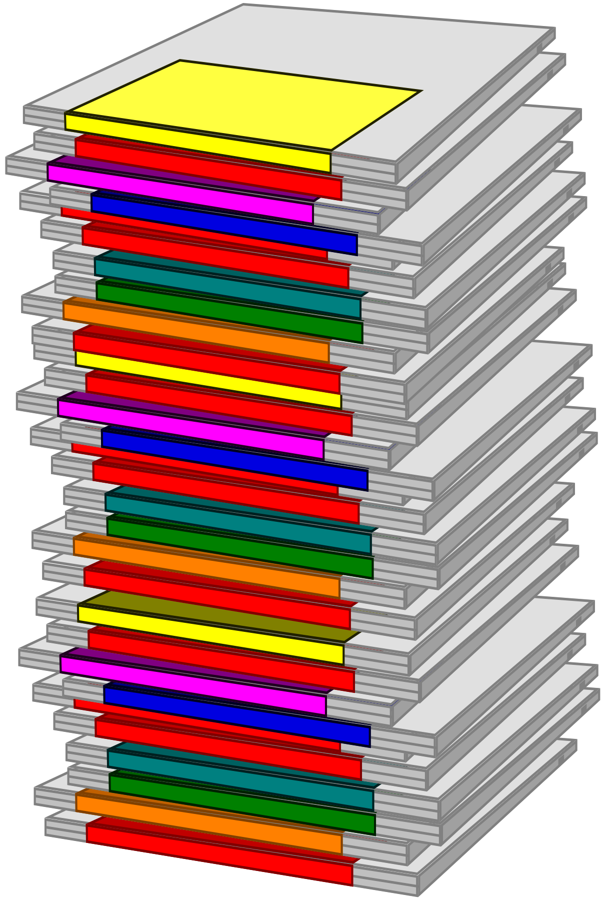
results are unexpected

data needs to be transformed from one form to other

What is a Data Warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

[Barry Devlin]

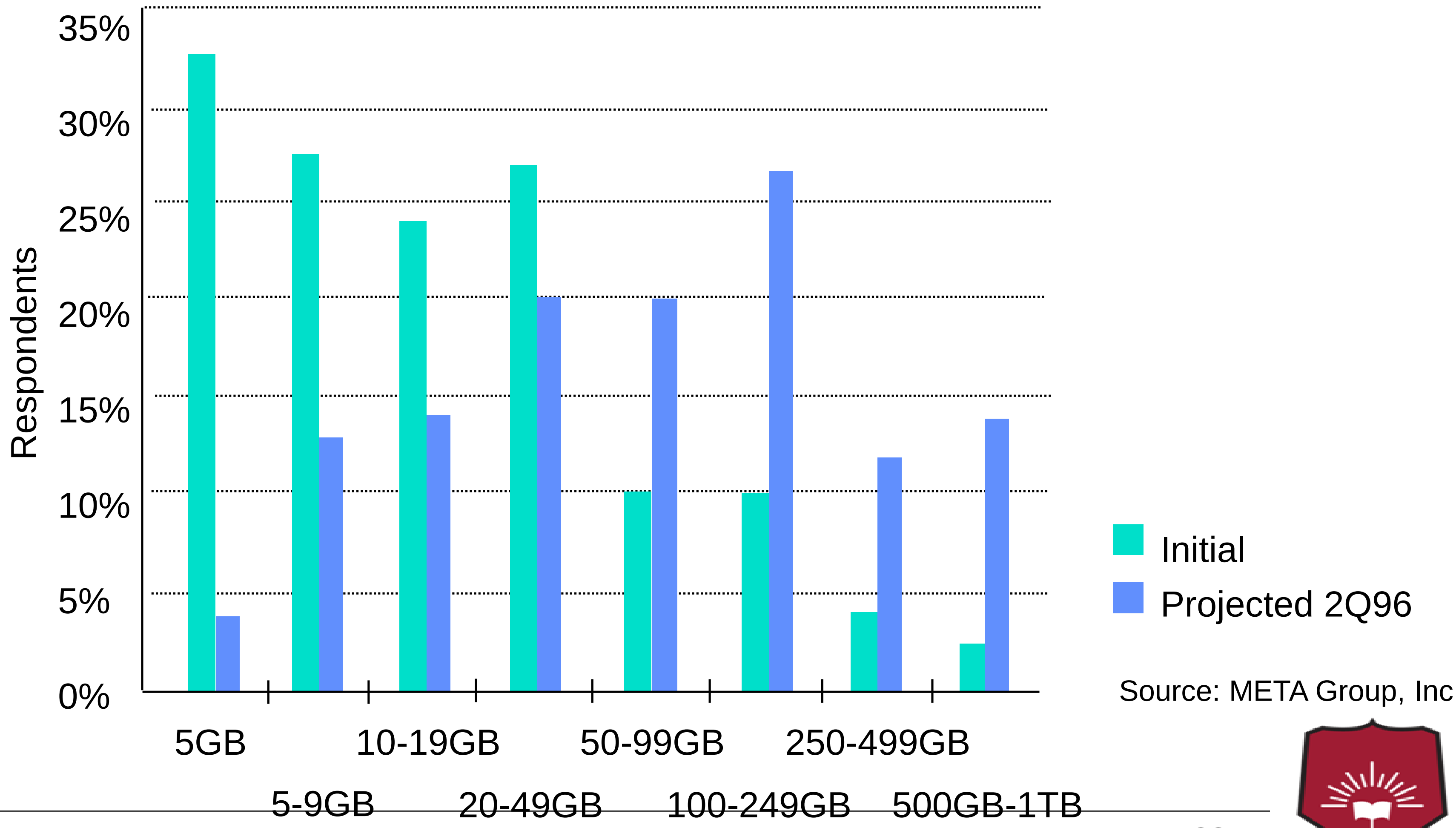


Data Warehousing -- It is a process

Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible.

A decision support database maintained separately from the organization’s operational database

Warehouses are Very Large Databases



Source: META Group, Inc.

Very Large Data Bases

Terabytes -- 10^12 bytes:	Walmart -- 24 Terabytes
Petabytes -- 10^15 bytes:	Geographic Information Systems
Exabytes -- 10^18 bytes:	National Medical Records
Zettabytes -- 10^21 bytes:	Weather images
Zottabytes -- 10^24 bytes:	Intelligence Agency Videos

Unit No: 1

Unit name : Introduction to Data warehouse

Lecture No: 2
Datawarehousing Components, Building Building a Data warehouse

DWM- Information flow mechanism

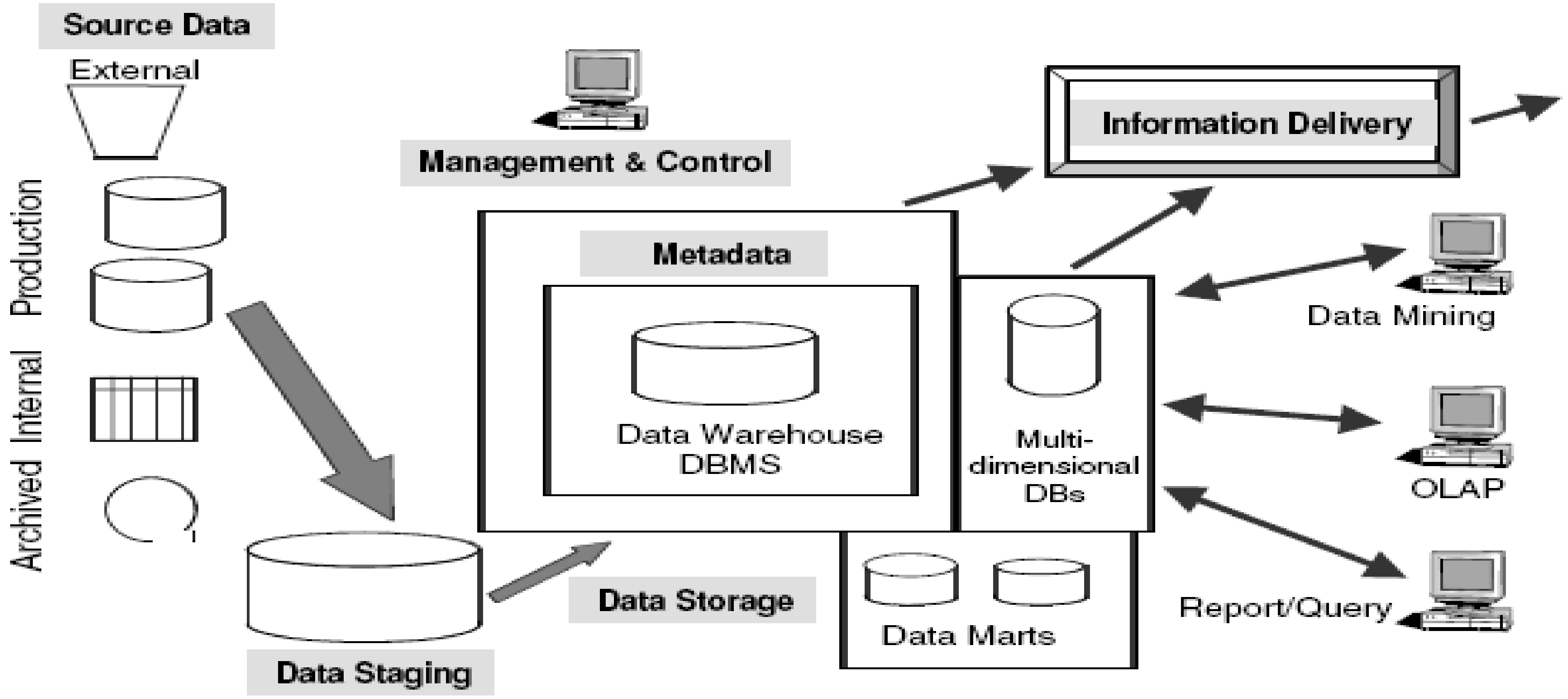


Figure 2-6 Data warehouse: building blocks or components.

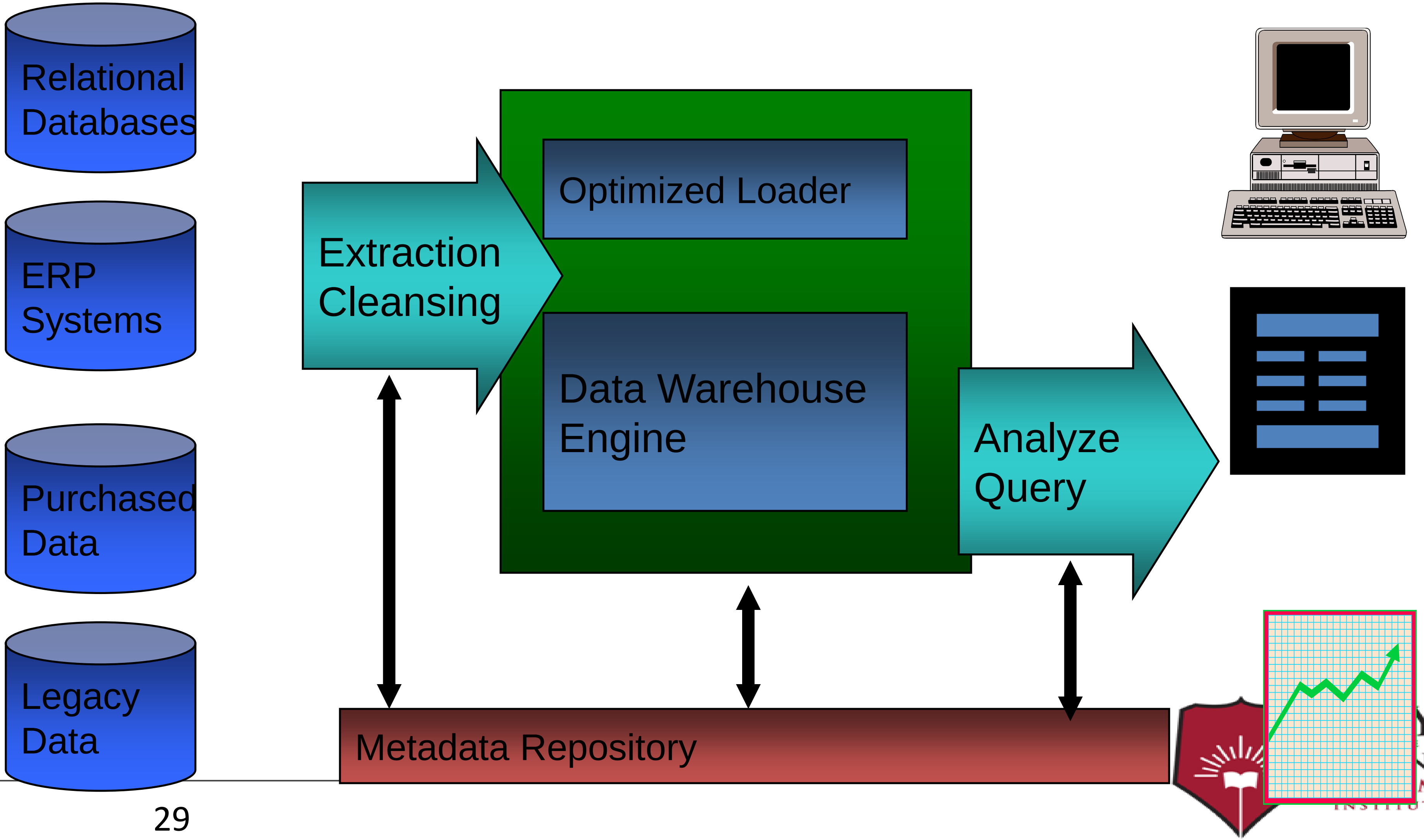
Data Staging Area

- Data Extraction – extracts data from various sources identified
- Data Transformation – transform, clean and standardize the data that has to be extracted to be stored in the data warehouse
- Data Loading – prepared data temporarily waits in data staging area from where it is loaded in the data warehouse
- **Why?** - data is pulled from many operational systems and is stored in terms of subjects and not applications
  - Staging area is required to prepare data for data warehouse
- **Advantages?** isolates raw data extracted from various sources of processed data
  - Additional security, as data warehouse users are not supposed to access staging area
  - Shares load as ‘data preparation’ and ‘data warehouse querying’ tasks are done by different systems
  - Eases development of central metadata repository which maintains documentation for involved systems

Evolution

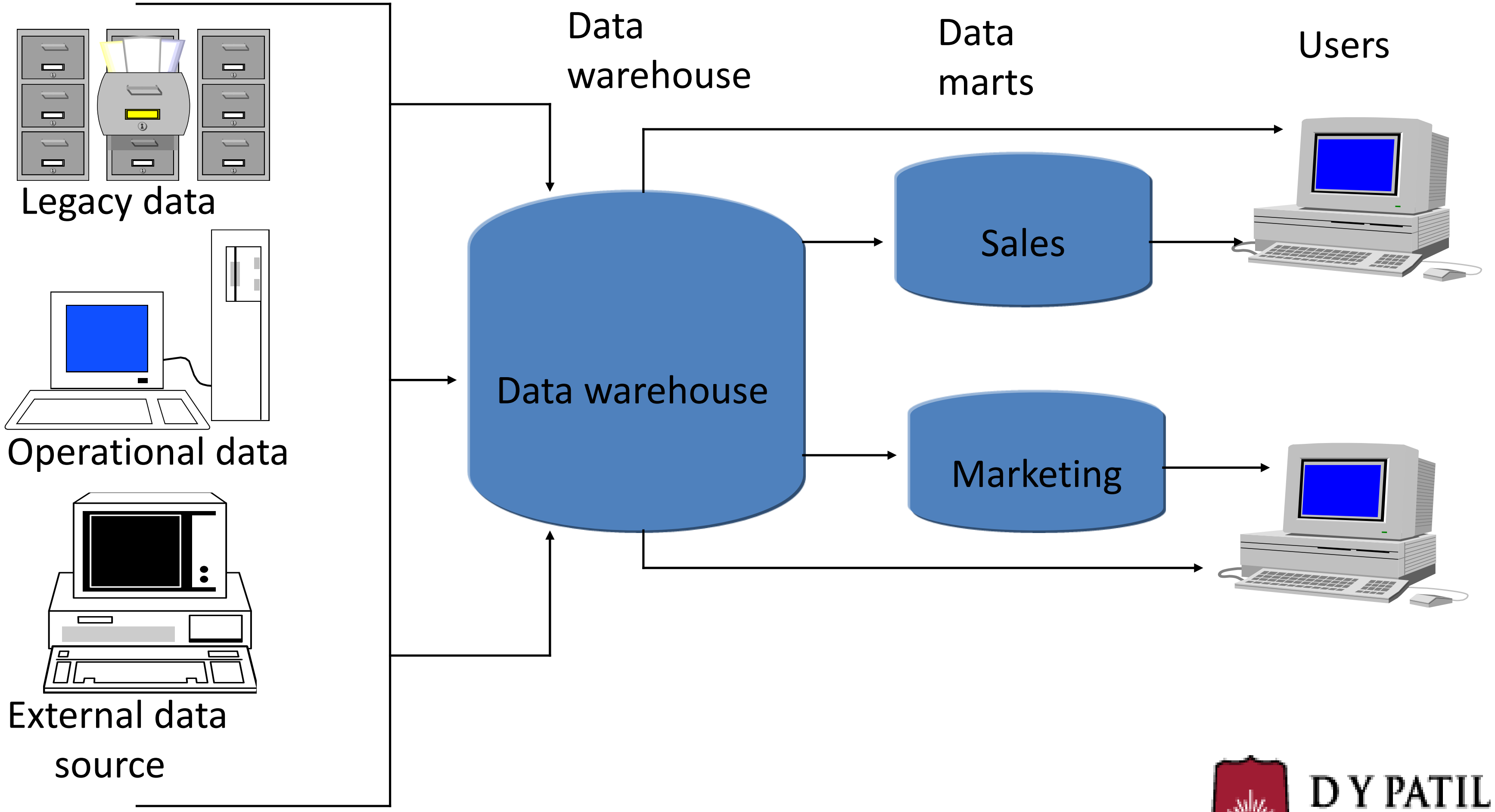
- 60’s: Batch reports
  - hard to find and analyze information
  - inflexible and expensive, reprogram every new request
- 70’s: Terminal-based DSS and EIS (executive information systems)
  - still inflexible, not integrated with desktop tools
- 80’s: Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easier to use, but only access operational databases
- 90’s: Data warehousing with integrated OLAP engines and tools

Data Warehouse Architecture



Data warehouse design strategies

- **Top-down approach**



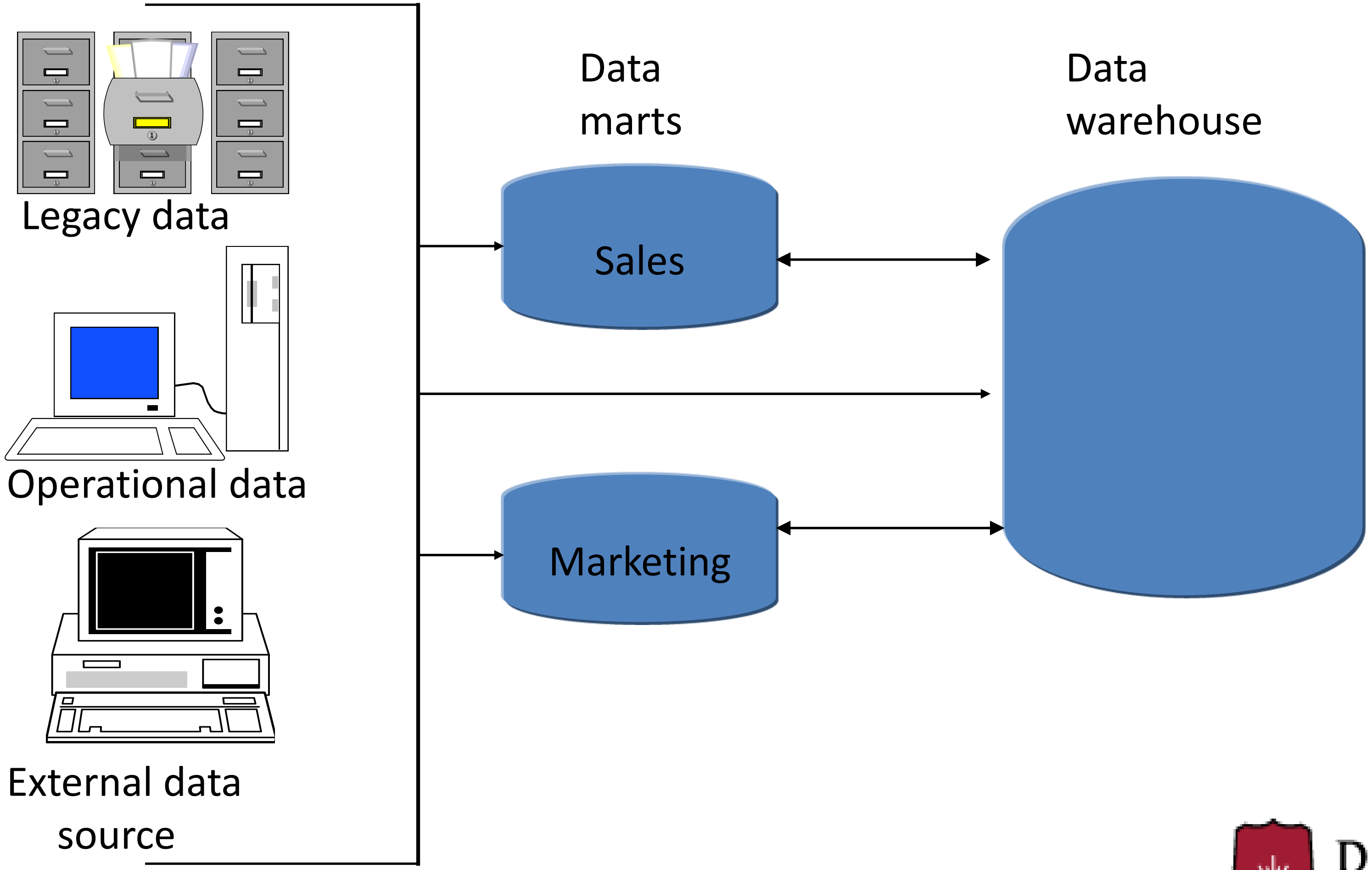
Top Down approach

- Data in data warehouse is at lowest level of granularity
- Data warehouse is created first
- Centralized data warehouse would feed dependent data marts

- Advantages:
    - An enterprise view of data
    - Inherently architected, not union of disparate data marts
    - Single, central storage of data about the content
    - Centralized rules and control
    - May see quick results if implemented with iterations
- Disadvantages:
    - Takes longer to build even with an iterative method
    - High exposure to risk of failure
    - Needs high level of cross-functional skills
    - High outlay with out proof of concept

Data warehouse design strategies - Bottom – up approach

- **Bottom – up approach**



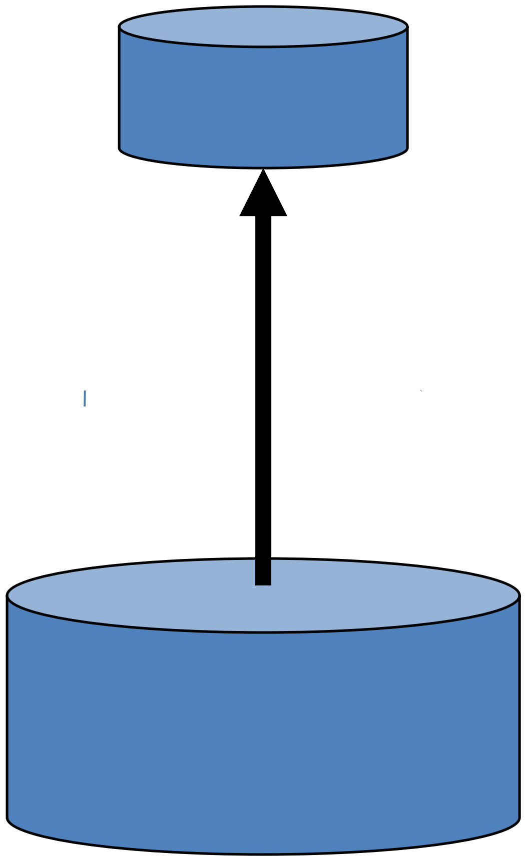
Bottom up approach

- Data marts are created first to provide analytical and reporting capabilities based on dimensional data model
- Data marts contain data at lowest level of granularity
- These data marts are joined or unioned by conforming the dimensions
- Advantages
  - Faster and easier implementation of manageable pieces
  - Favorable return on investment and proof of concept
  - Less risk of failure
  - Inherently incremental; can schedule important data marts first
  - Allows project team to learn and grow
- Disadvantages
  - Each data mart has its own narrow view of data
  - Permeates redundant data in every data mart
  - Perpetuates inconsistent and irreconcilable data
  - Proliferates unmanageable interfaces

Data Warehouse for Decision Support & OLAP

- Putting Information technology to help the knowledge worker make faster and better decisions
  - Which of my customers are most likely to go to the competition?
  - What product promotions have the biggest impact on revenue?
  - How did the share price of software companies correlate with profits over last 10 years?

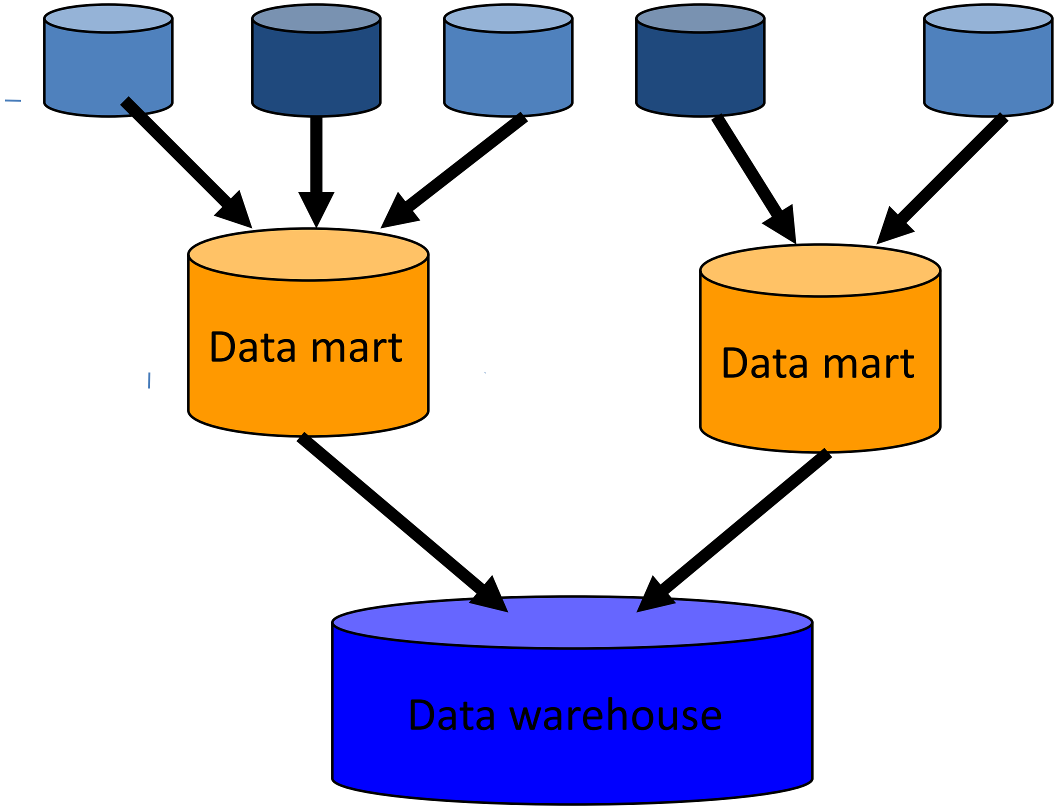
Data warehouse and Data marts



- OLAP
- Data Mart
- Lightly summarized
- Departmentally structured

- Organizationally structured
- Atomic
- Detailed Data Warehouse

Data warehouse and Data marts



Information Delivery

Different ways to access data warehouse –

- Queries
- Reports
- Analysis
- Applications
- OLAP
- Data Mining

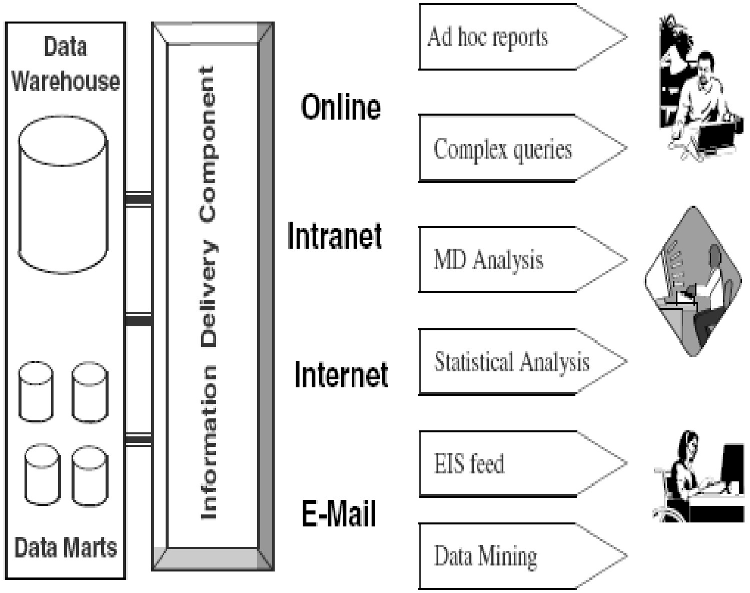
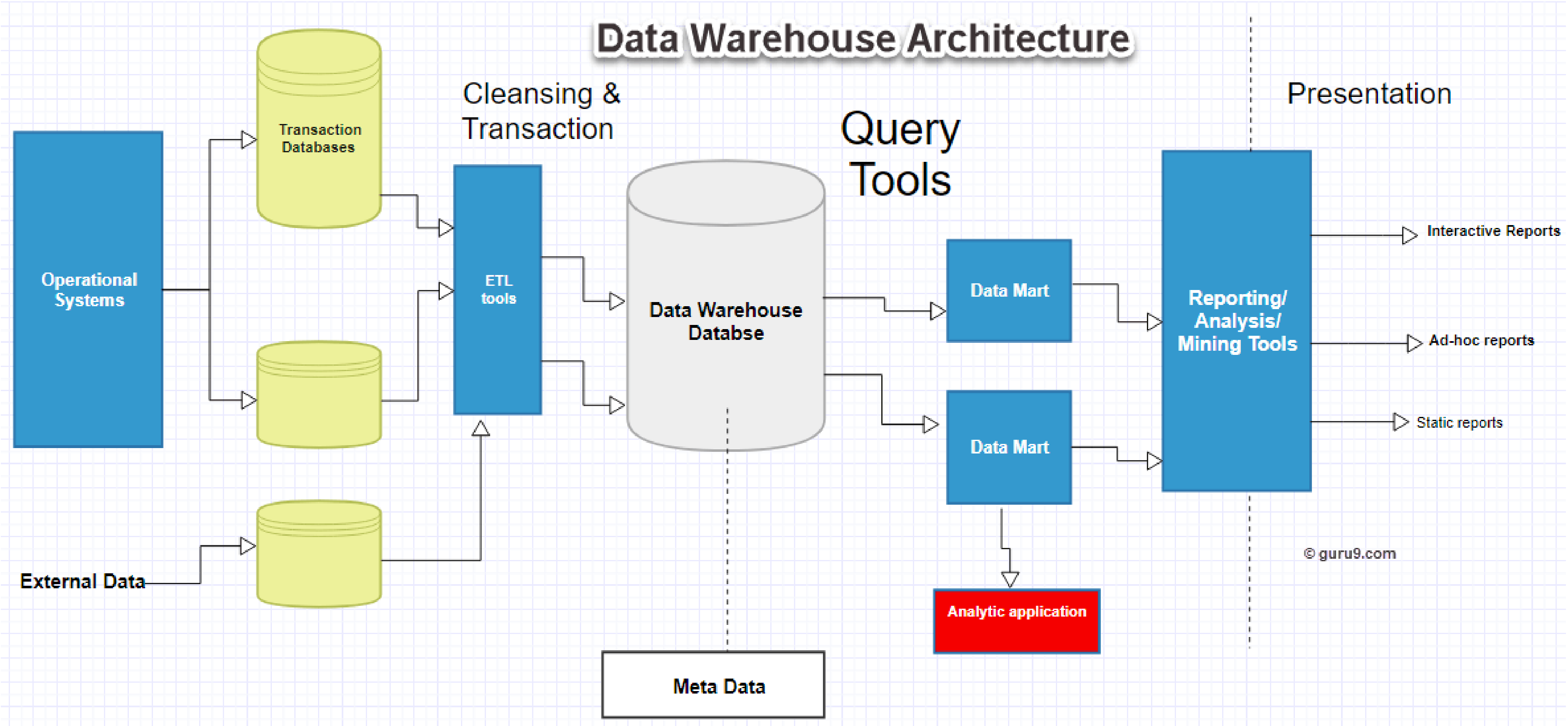


Figure 2-8 Information delivery component.

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.

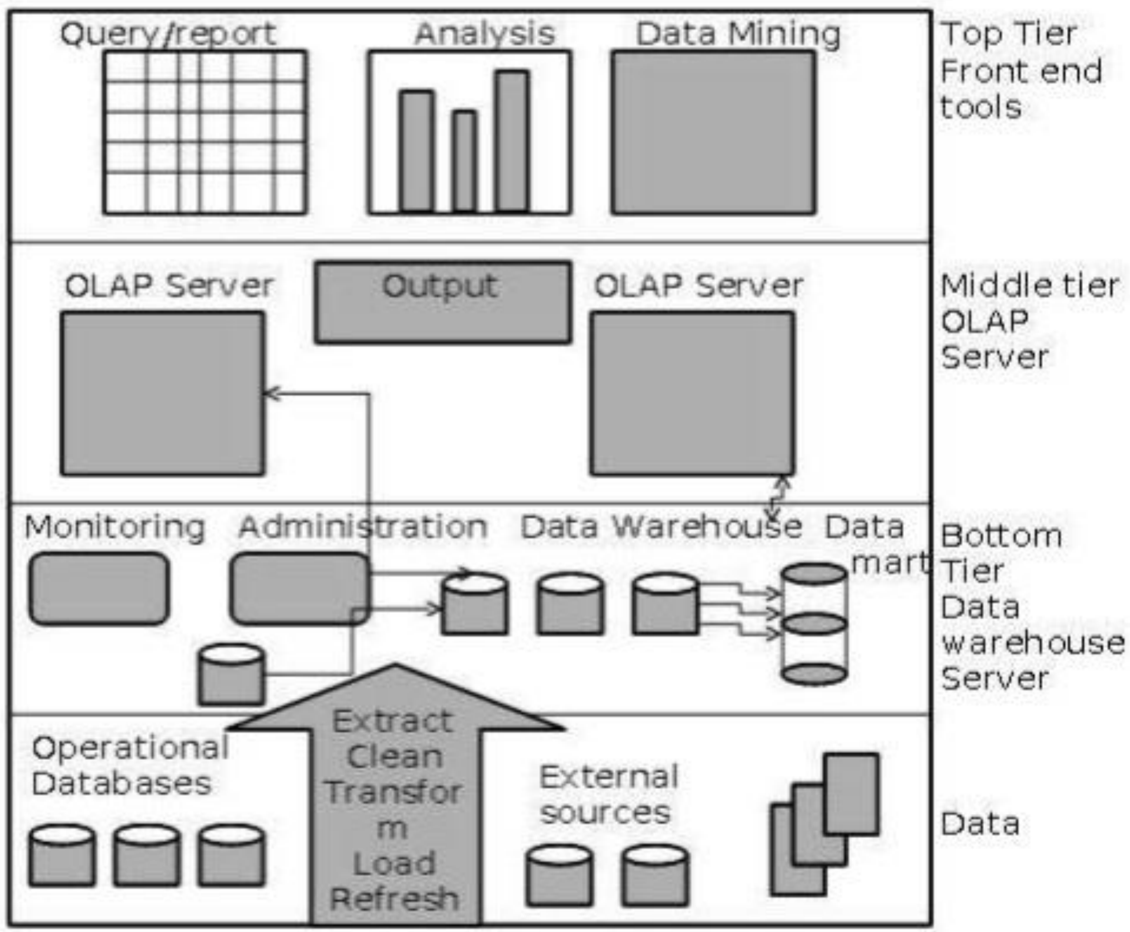
Not expandable and also not supporting a large number of end-users.

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 No: 1

Unit name : Introduction to Data warehouse

Lecture No: 3
Data Warehouse Features

Characteristics of Data Warehouse

- **Subject oriented.** Data are organized based on how the users refer to them.
- **Integrated**. All inconsistencies regarding naming convention and value representations are removed.
- **Nonvolatile**. Data are stored in read-only format and do not change over time.
- **Time variant**. Data are not current but normally time series.

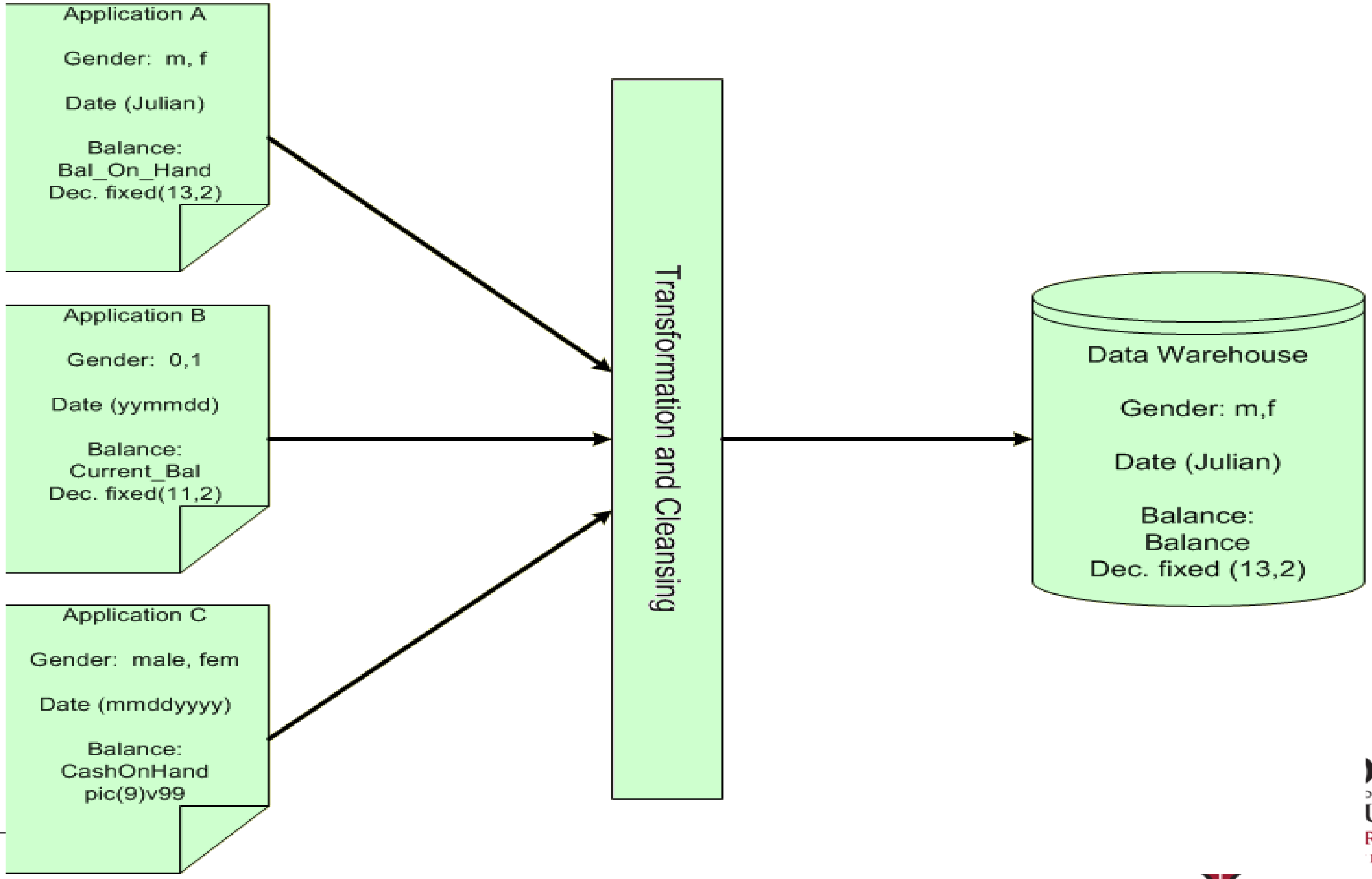
Characteristics of Data Warehouse

- **Summarized** Operational data are mapped into a decision-usable format
- **Large volume**. Time series data sets are normally quite large.
- **Not normalized**. DW data can be, and often are, redundant.
- **Metadata**. Data about data are stored.
- **Data sources**. Data come from internal and external unintegrated operational systems.

Data Integrated

- **Integration** –consistency naming conventions and measurement attributers, accuracy, and common aggregation.
- Establishment of a common unit of measure for all synonymous data elements from dissimilar database.
- The data must be stored in the DW in an integrated, globally acceptable manner

Data Integrated



Features of Data Warehousing – integrated data

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

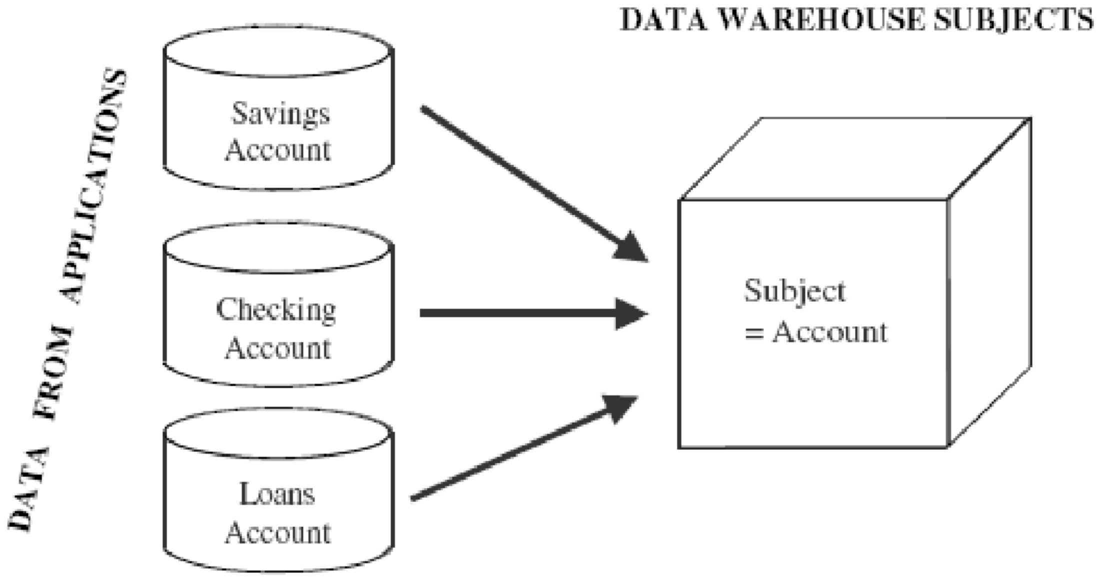


Figure 2-2 The data warehouse is integrated.

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

Time Variant

- In an operational application system, the expectation is that all data within the database are accurate as of the moment of access. In the DW data are simply assumed to be accurate as of some moment in time and not necessarily right now.
- One of the places where DW data display time variance is in the structure of the record key. Every primary key contained within the DW must contain, either implicitly or explicitly an element of time(day, week, month, etc)

Time Variant

- Every piece of data contained within the warehouse must be associated with a particular point in time if any useful analysis is to be conducted with it.
- Another aspect of time variance in DW data is that, once recorded, data within the warehouse cannot be updated or changed.

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

Nonvolatility

- Typical activities such as deletes, inserts, and changes that are performed in an operational application environment are completely nonexistent in a DW environment.
- Only two data operations are ever performed in the DW: data loading and data access .

Features of Data Warehousing – non volatile data

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

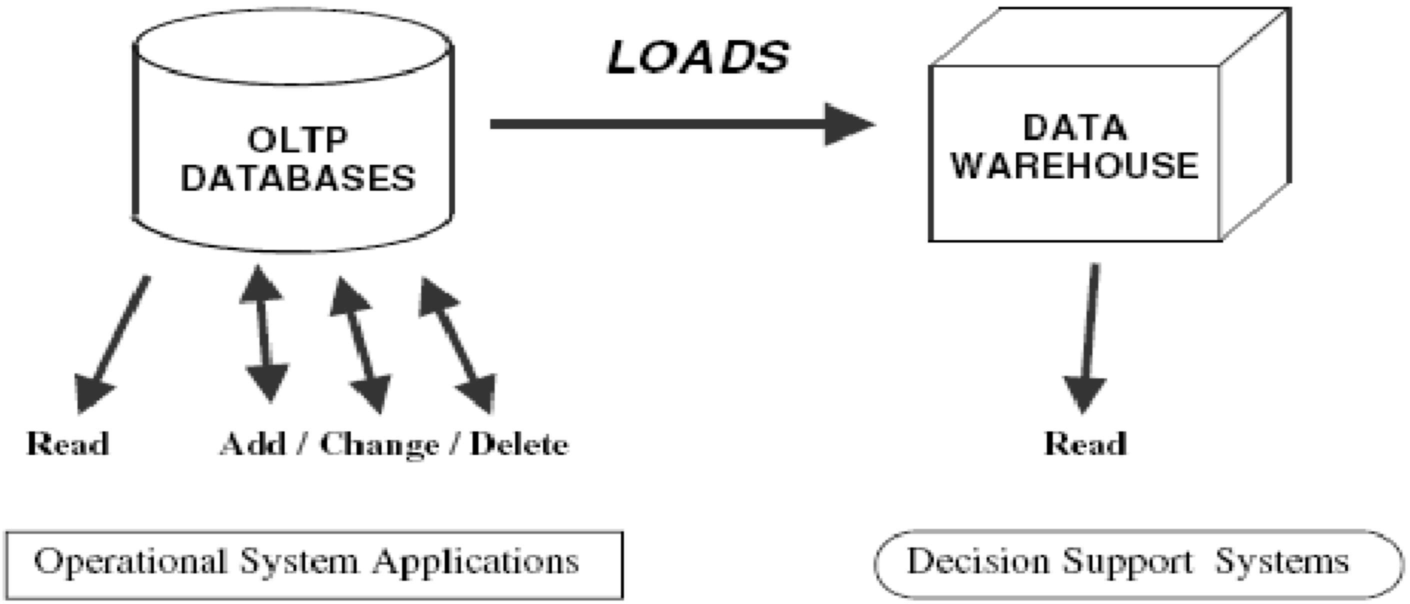


Figure 2-3 The data warehouse is nonvolatile.

The Metadata

- The name suggests some high-level technological concept, but it really is fairly simple. Metadata is “data about data”.
- With the emergence of the data warehouse as a decision support structure, the metadata are considered as much a resource as the business data they describe.
- Metadata are abstractions -- they are high level data that provide concise descriptions of lower-level data.

Difference between DBMS and Data Warehouse

Database System	Data Warehouse
It supports operational processes.	It supports analysis and performance reporting.
Capture and maintain the data.	Explore the data.
Current data.	Multiple years of history.
Data is balanced within the scope of this one system.	Data must be integrated and balanced from multiple system.
Data is updated when transaction occurs.	Data is updated on scheduled processes.
Data verification occurs when entry is done.	Data verification occurs after the fact.
100 MB to GB.	100 GB to TB.
ER based.	Star/Snowflake.
Application oriented.	Subject oriented.
Primitive and highly detailed.	Summarized and consolidated.
Flat relational.	Multidimensional.

Unit No: 1

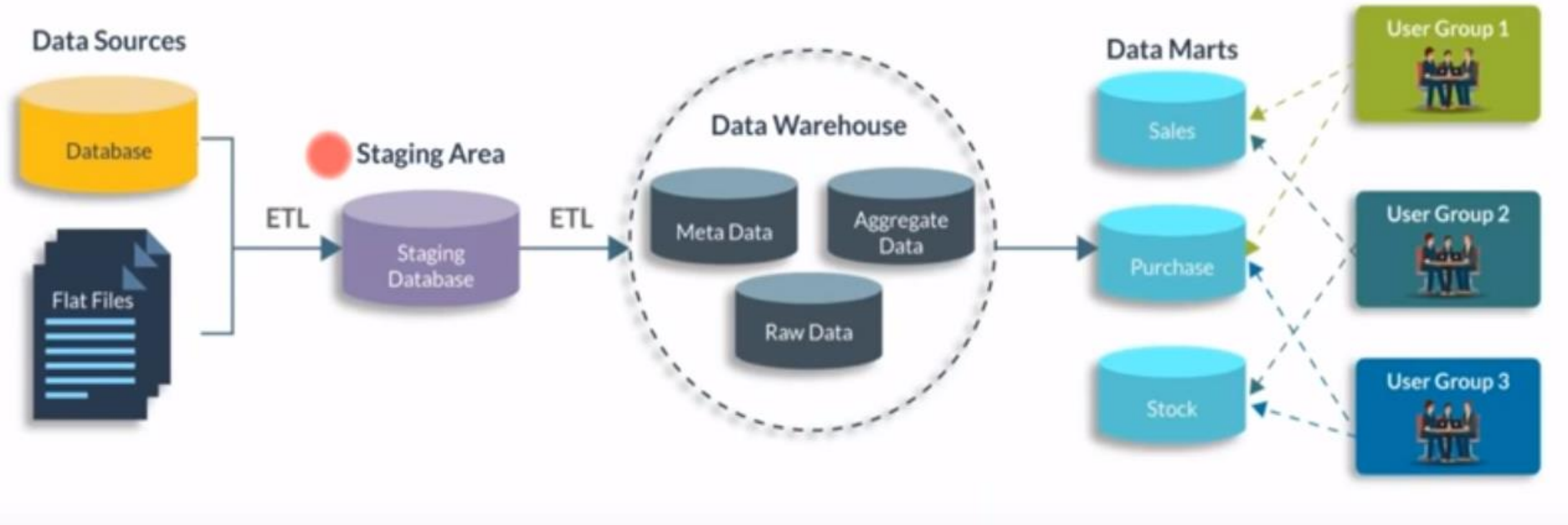
Unit name : Introduction to Data warehouse

Lecture No: 4
Data Warehouse V/S Data Mart

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.



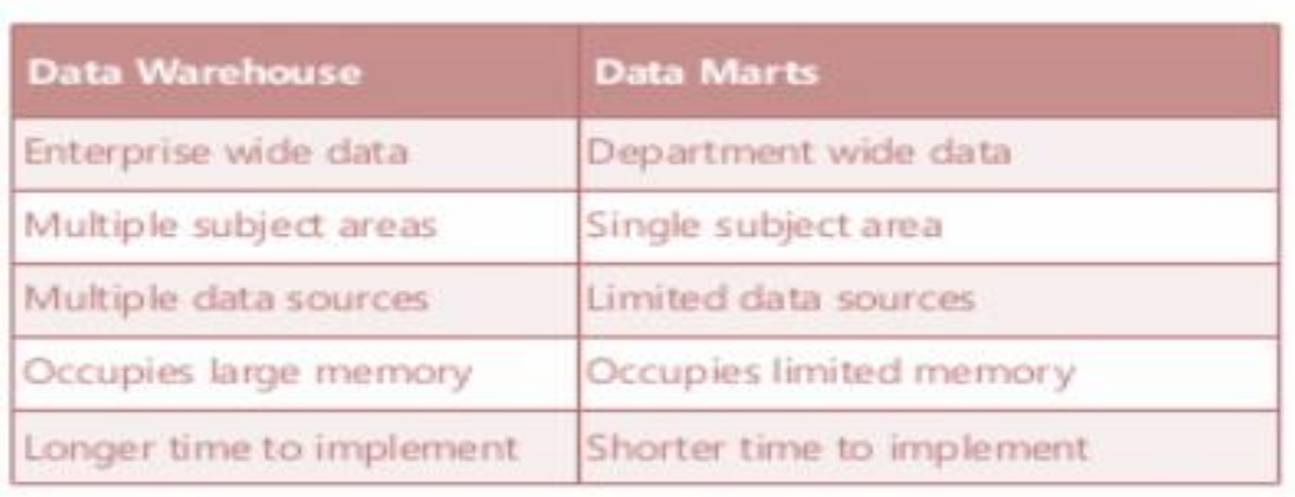
Data Mart

Points to remember about data marts –

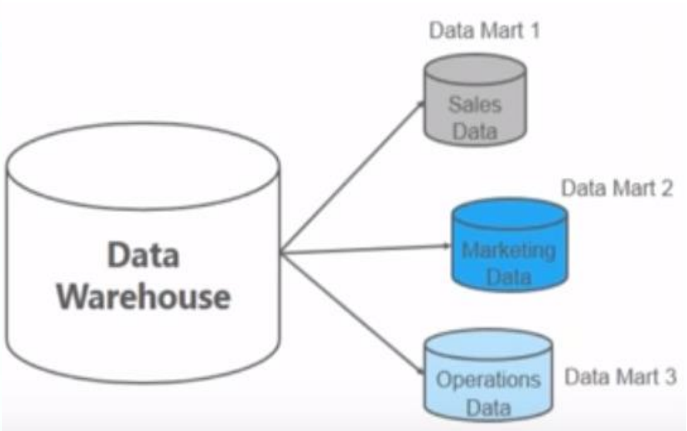
- Unix/Linux-based servers are used to implement data marts.
- They are implemented on low-cost servers.
- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
- Data marts are small in size.
- Data marts are customized by department.
-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

Data warehouse vs Data Mart

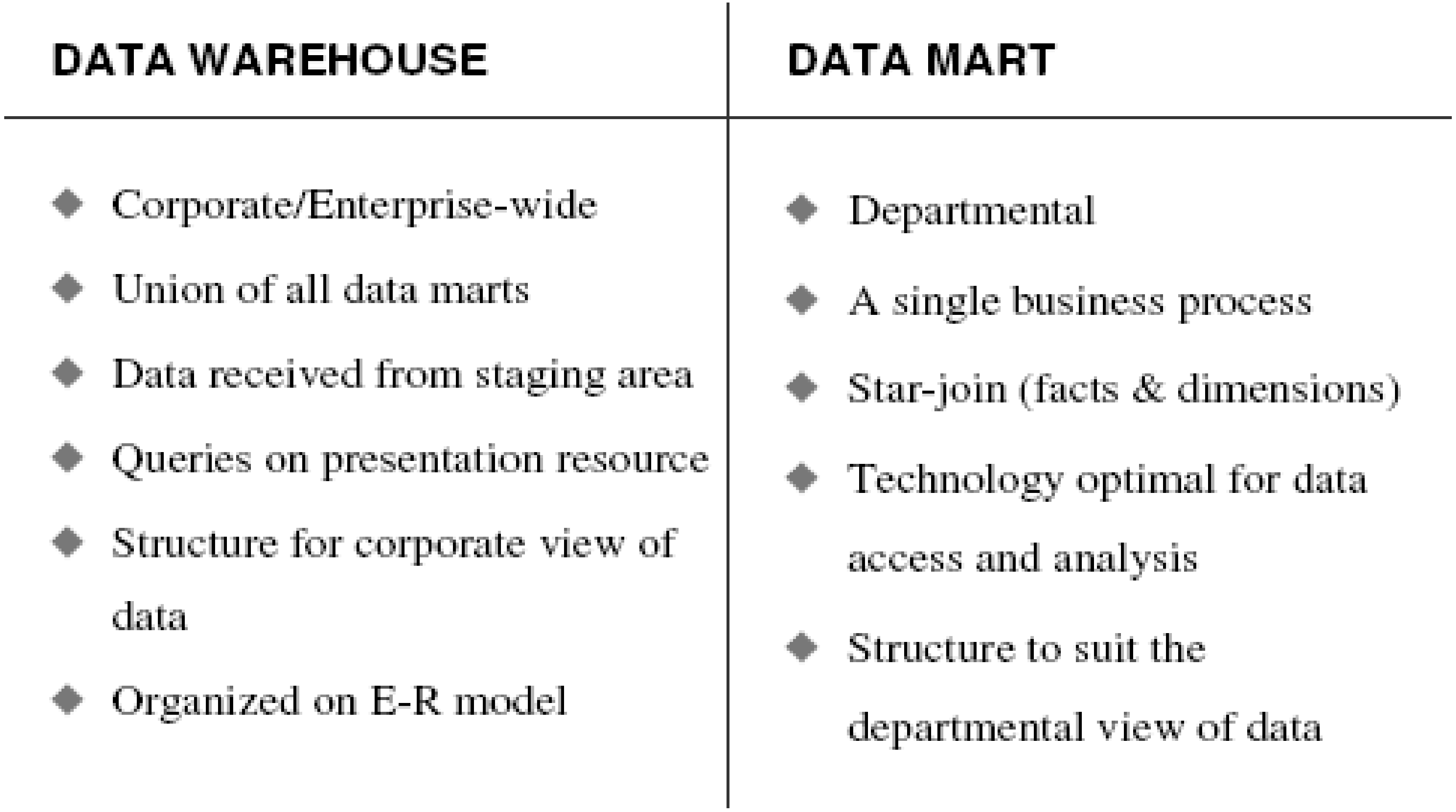
- Data Mart is smaller version of Data Warehouse which deals with single subject
- Data marts are focused on one area, hence they draw data from limited number of sources
- Time taken to build data mart is very less compared to DWH



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



Data warehouse and Data marts



DATA WAREHOUSE	DATA MART
Corporate/Enterprise-wide	Departmental
Union of all data marts	A single business process
Data received from staging area	Star-join (facts & dimensions)
Queries on presentation resource	Technology optimal for data access and analysis
Structure for corporate view of data	Structure to suit the departmental view of data
Organized on E-R model	

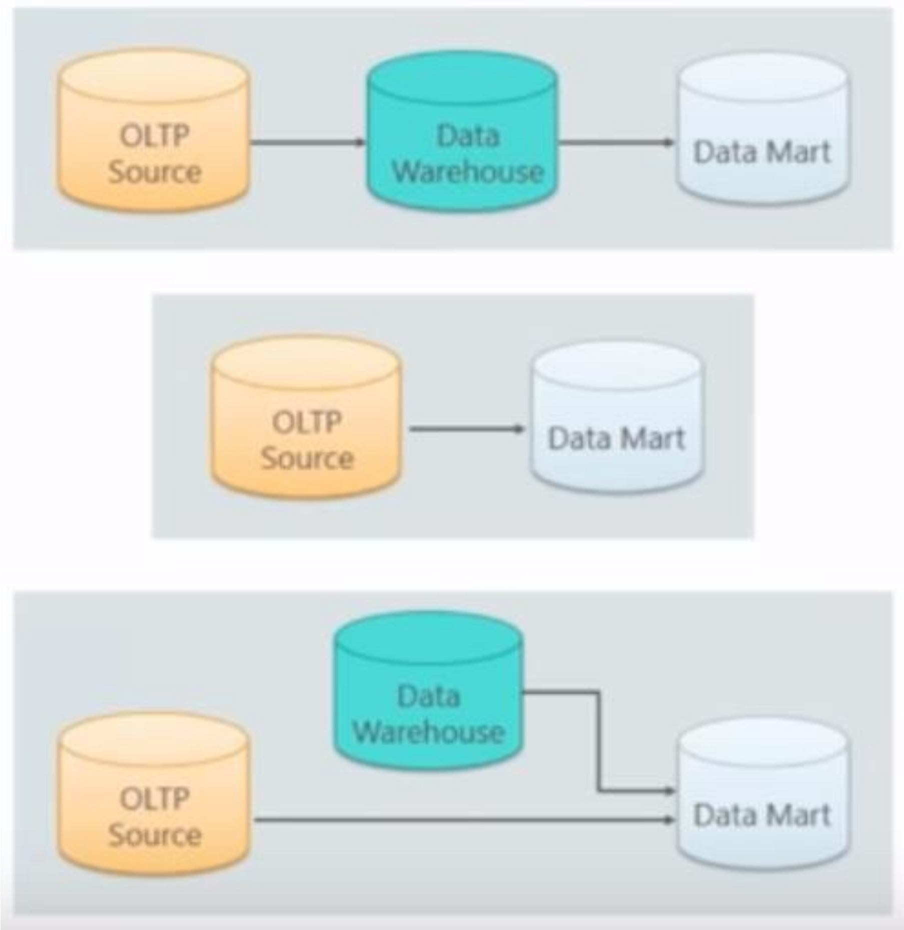
Figure 2-5 Data warehouse versus data mart.

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

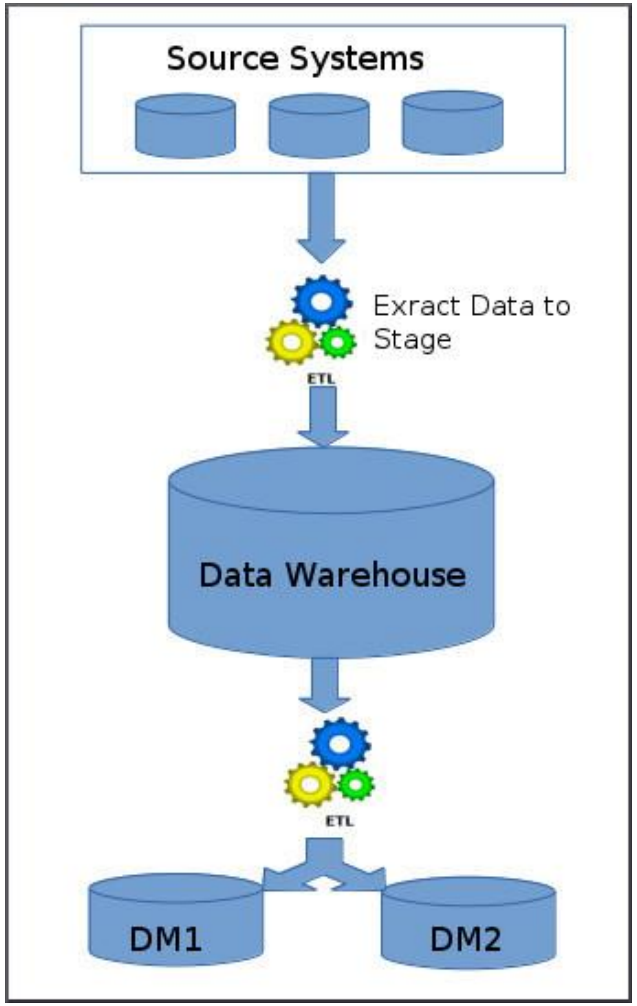
- Data Warehouse design approaches are very important aspect of building data warehouse.

- 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.

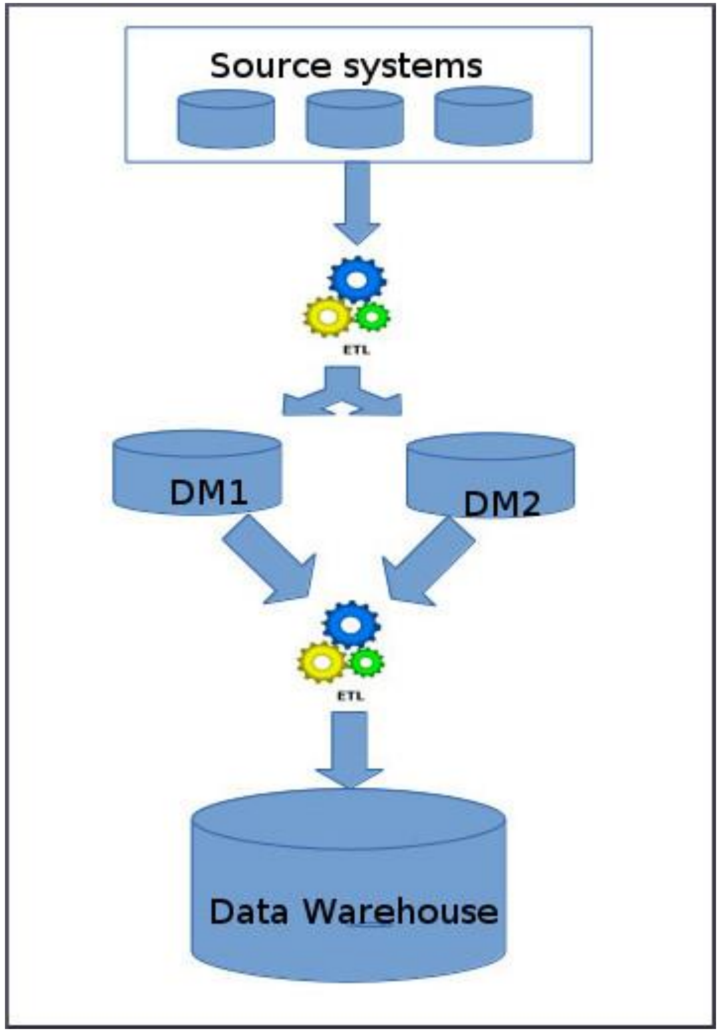
Top-Down Approach for Data warehouse Design

- 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



The Meta Data

- Last component of DW environments.

- It is information that is kept about the warehouse rather than information kept within the warehouse.

- Legacy systems generally don’t keep a record of characteristics of the data (such as what pieces of data exist and where they are located).

- The metadata is simply data about data.

The Metadata

For example, a line in a sales database may contain: 4056 KJ596 223.45

This is mostly meaningless until we consult the metadata that tells us it was store number 4056, product KJ596 and sales of $223.45

The metadata are essential ingredients in the transformation of raw data into knowledge. They are the “keys” that allow us to handle the raw data.

Metadata

- Metadata similar to data dictionary in DBMS
- It stores data about data in data warehouse
- Used for building, maintaining, managing and using the data warehouse

Metadata contains information about

- Structure of data from programmer's perspective
- Structure of data from end-user's perspective
- Source systems that feed data warehouse
- Transformation process that was applied before the data was passed to data warehouse
- Data model
- History of data extraction process

Why Metadata is important?

Users to compose and run the query can have several important questions:

- Are there any predefined queries I can look at?
- What are the various elements of data in the warehouse?
- Is there information about unit sales and unit costs by product?
- How can I browse and see what is available?
- From where did they get the data for the warehouse? From which source systems?
- How did they merge the data from the telephone orders system and the mail orders system?
- How old is the data in the warehouse?
- When was the last time fresh data was brought in?
- Are there any summaries by month and product?

Dimensional Model v/s ER Model

- ER diagram is complex diagram – represent multiple processes
- Single ER diagram can be broken into multiple dimensional model diagrams

- Tables in dimensional model are in de-normalized form, where as in ER diagram, main aim is to remove redundancy by normalizing the tables

- ER model is designed to express microscopic relationships between data elements, where as key idea behind dimensional model is to capture business measures.

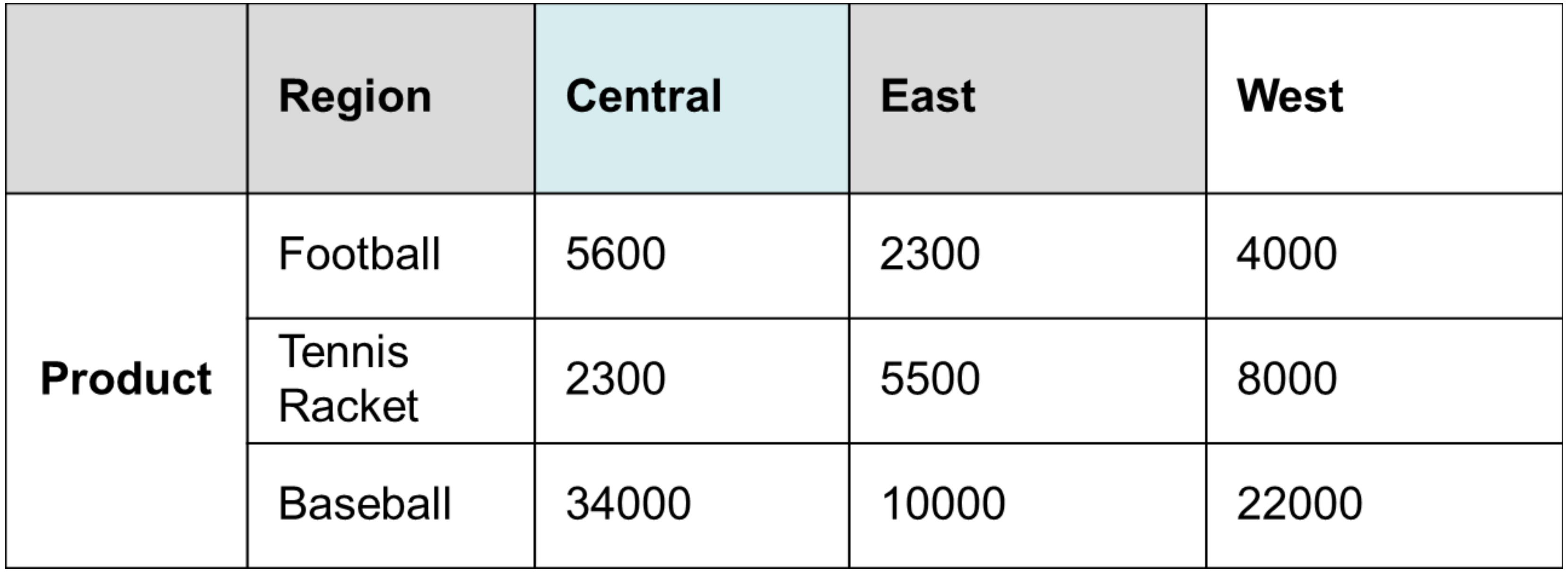
- A dimensional model is designed to answer queries on overall business process to reveal trends – how managers think of their business
  - ER model is suited to answer queries at transaction level

In-Short

- A Data Warehouse is a collection of integrated subject-oriented databases designed to support a DSS.
  - Each unit of data is non-volatile and relevant to some moment in time.
- An operational data store (ODS) stores data for a specific application. It feeds the data warehouse a stream of desired raw data.
- A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users (rather than the entire firm).
- The metadata is information that is kept about the warehouse.

Dimension – example

- Cube dimension



	Region	Central	East	West
Product	Football	5600	2300	4000
	Tennis Racket	2300	5500	8000
	Baseball	34000	10000	22000

Dimensional Modeling

- Dimensional Modeling is a logical design technique used in data warehouses(DWH)

- Dimensional Modeling is a design technique for databases intended to support end-user queries in a DWH

- It is oriented around understandability as opposed to database administration

- A de-normalized relational model
  - Made up of tables with attributes
  - Relationships defined by keys and foreign keys

- Organized for understandability and ease of reporting rather than update

Dimensional Modeling

- Dimensional Modeling consists of two types of tables:
  - Dimension tables
  - Fact tables
- It is important to identify the *granularity* in dimensional modeling – level of detail stored in the table
- For example

1. A table such as date_A (year, **quarter**) has a granularity at the quarter level but does not have information for individual days or months.

2. A table such as date_B (year, quarter, **month**) table has granularity at the month level, but does not contain information at the day level.

- The more detail there is, the lower the level of granularity.
- The less detail there is, the higher the level of granularity.

Dimensional Model v/s ER Model

Dimensional Model	ER Model
Captures critical measures, views along dimensions	Removes data redundancy, ensures data consistency
A view of data from business processing	A view of data from data processing
It contains only physical model	It contains both logical and physical model
It process de-normalized data	It process normalized data
Data: It uses historical data	Data: It uses current data
User: Using only top management	Use: More than 1000
Size: GB to TB	Size: MB to TB
Process: De-normalization	Process: Normalization
Data Storage: Non-Volatile	Data Storage: Volatile

“A real-time enterprise without real-time business intelligence is a real fast, dumb organization.”

Stephen Brobst
Chief Technology Office
Teradata

Interesting Facts

- Harrah’s Entertainment’s Data Warehouse holds 30 terabytes, or 30 trillion bytes of data, roughly three times the number of printed characters in the Library of Congress

- Casinos, retailers, airlines, and banks are piling up data so vast, it would have been unthinkable years ago; result from the curse of cheap storage

Interesting Facts

- Storage Shipments as of 2004: 22 exabytes or 22 million trillion bytes of hard disk space, double the amount in 2002.

- Equivalent to 4x’s the space needed to store every word ever spoken by every human being who has ever lived.

- Should double again in 2006

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

THANK YOU