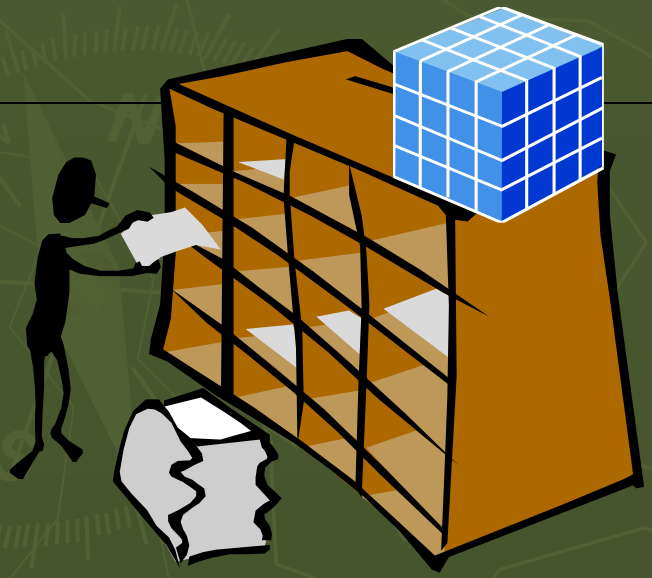
Data Warehousing

Introduction and Background



Source : From Internet, ppt from S. Sudarshan
Krithi Ramamritham

A producer wants to know....

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 ...

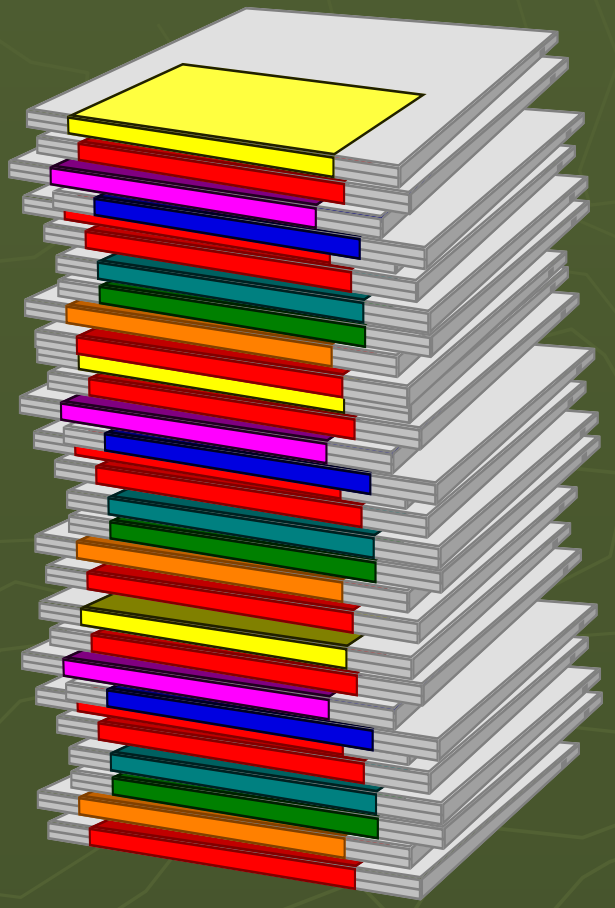
- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences

- 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.

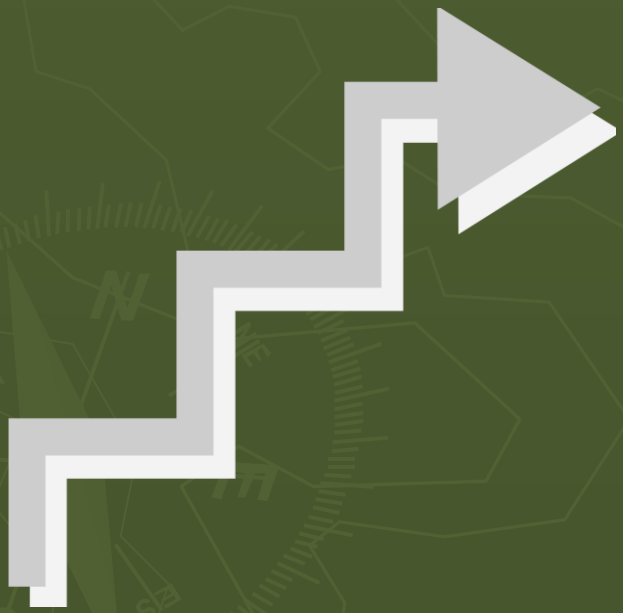
[Barry Devlin]



What are the users saying...

- Data should be integrated across the enterprise
- Summary data has a real value to the organization
- Historical data holds the key to understanding data over time
- What-if capabilities are required

What is Data Warehousing?



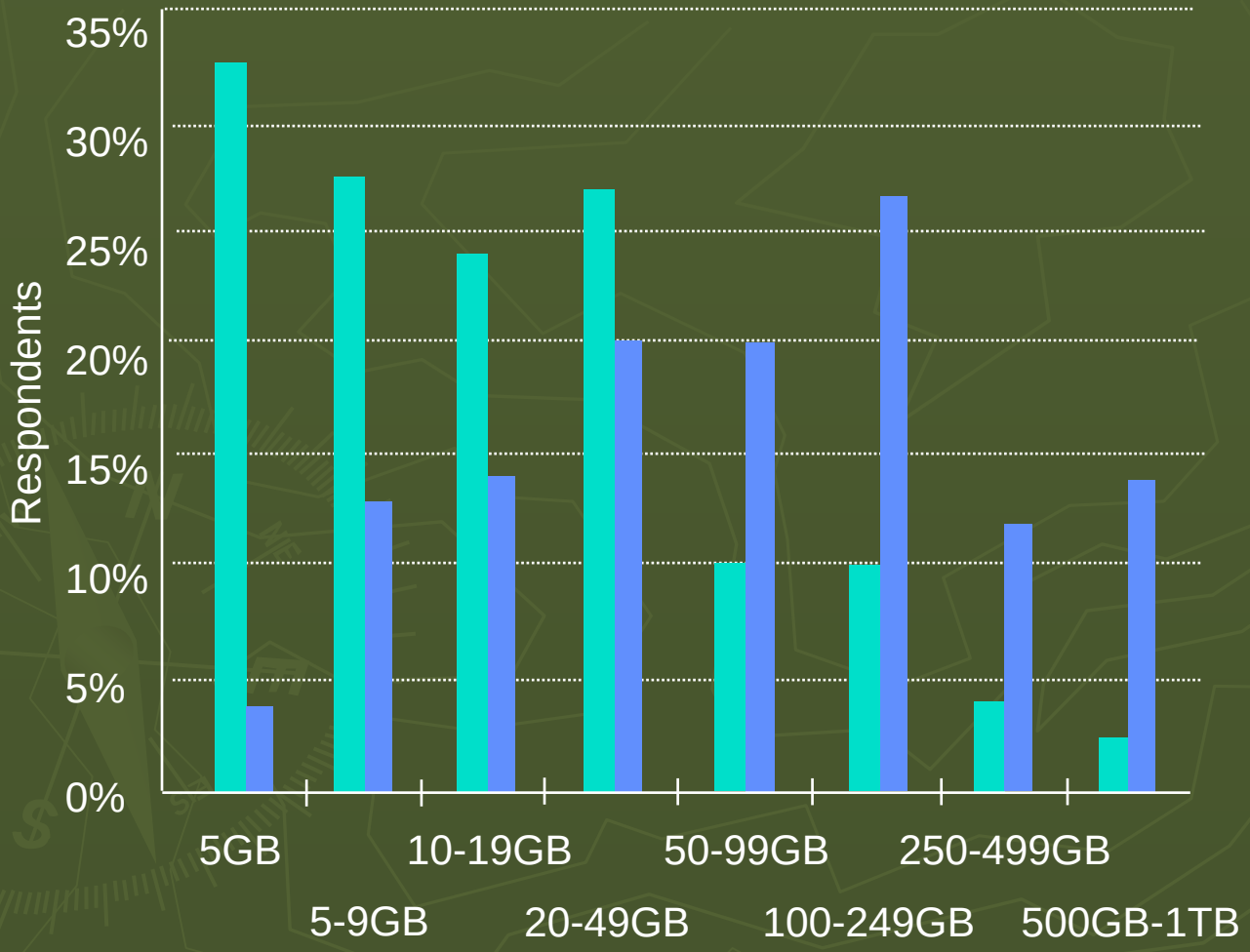
A process of transforming data into information and making it available to users in a timely enough manner to make a difference

[Forrester Research, April 1996]

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

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

The need

"Drowning in data and starving for information"

Knowledge is power, Intelligence is absolute power!

The need

POWER

INTELLIGENCE

KNOWLEDGE

INFORMATION

DATA

Historical overview

1960
Master Files & Reports

1965
Lots of Master files!

1970
Direct Access Memory & DBMS

1975
Online high performance transaction processing

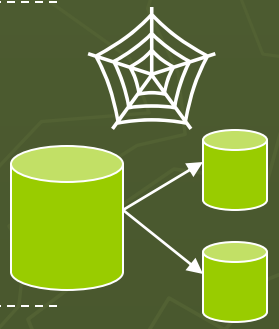
Historical overview

1980

PCs and 4GL Technology (MIS/DSS)

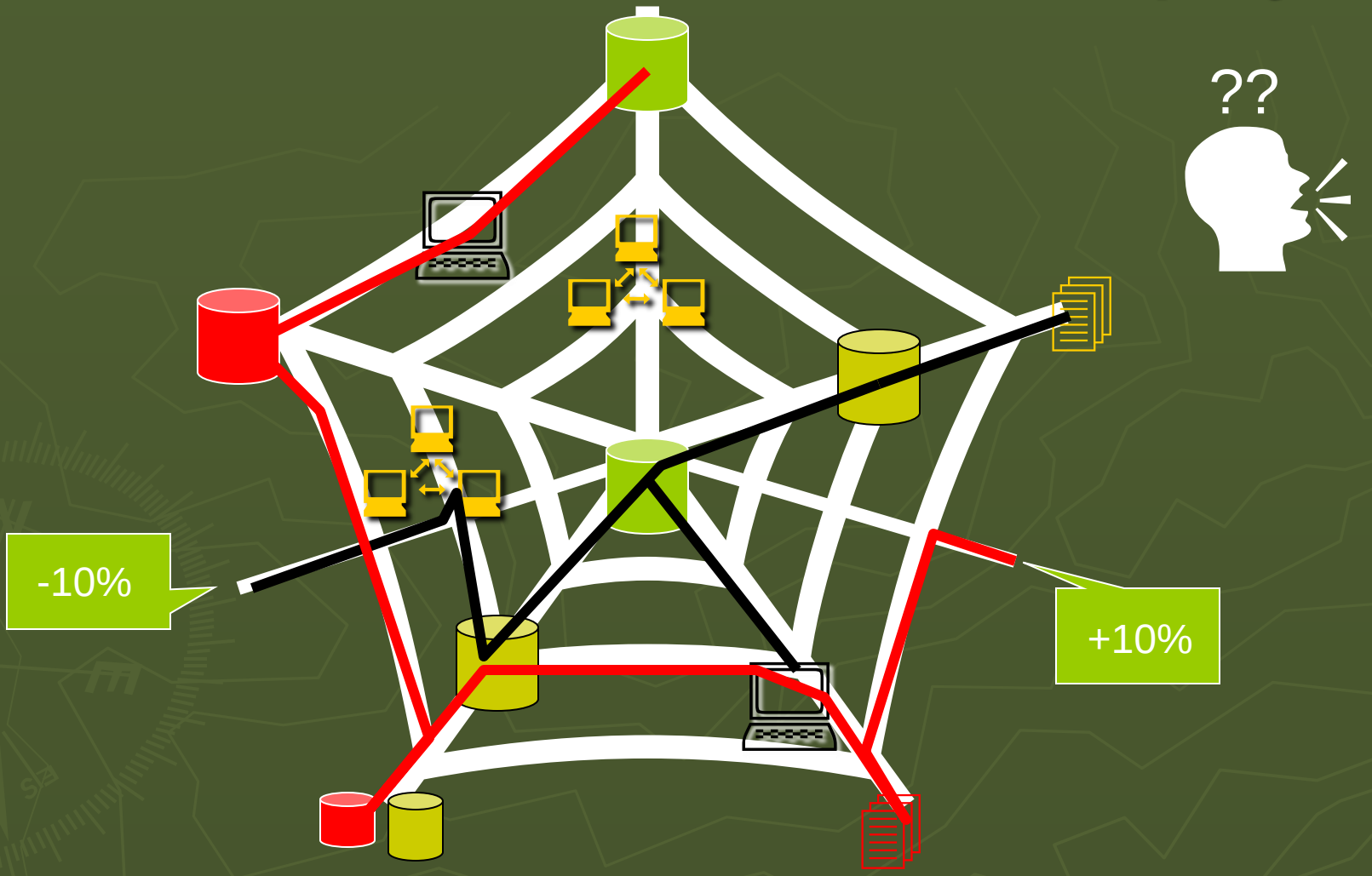
1985 & 1990

Extract programs, extract processing,
The legacy system's web



Historical overview: Crisis of Credibility

What is the financial health of our company?



Introduction and Background

Module 1

- The Need for Data Warehousing;
- Increasing Demand for Strategic Information; Inability of Past Decision Support System;
- Operational V/s Decisional Support System;
- Data Warehouse Defined;
- Benefits of Data Warehousing ;
- Features of a Data Warehouse;
- The Information Flow Mechanism;
- Role of Metadata; Classification of Metadata;
- Data Warehouse Architecture; Different Types of Architecture;
- Data Warehouse and Data Marts;
- Data Warehousing Design Strategies.

- Introduction to Data Warehouse, Data warehouse architecture, Data warehouse versus Data Marts,
- E-R Modeling versus Dimensional Modeling, Information Package Diagram, Data Warehouse Schemas; Star Schema, Snowflake Schema, Factless Fact Table, Fact Constellation Schema. Update to the dimension tables.
- .

Why Data Warehouse?

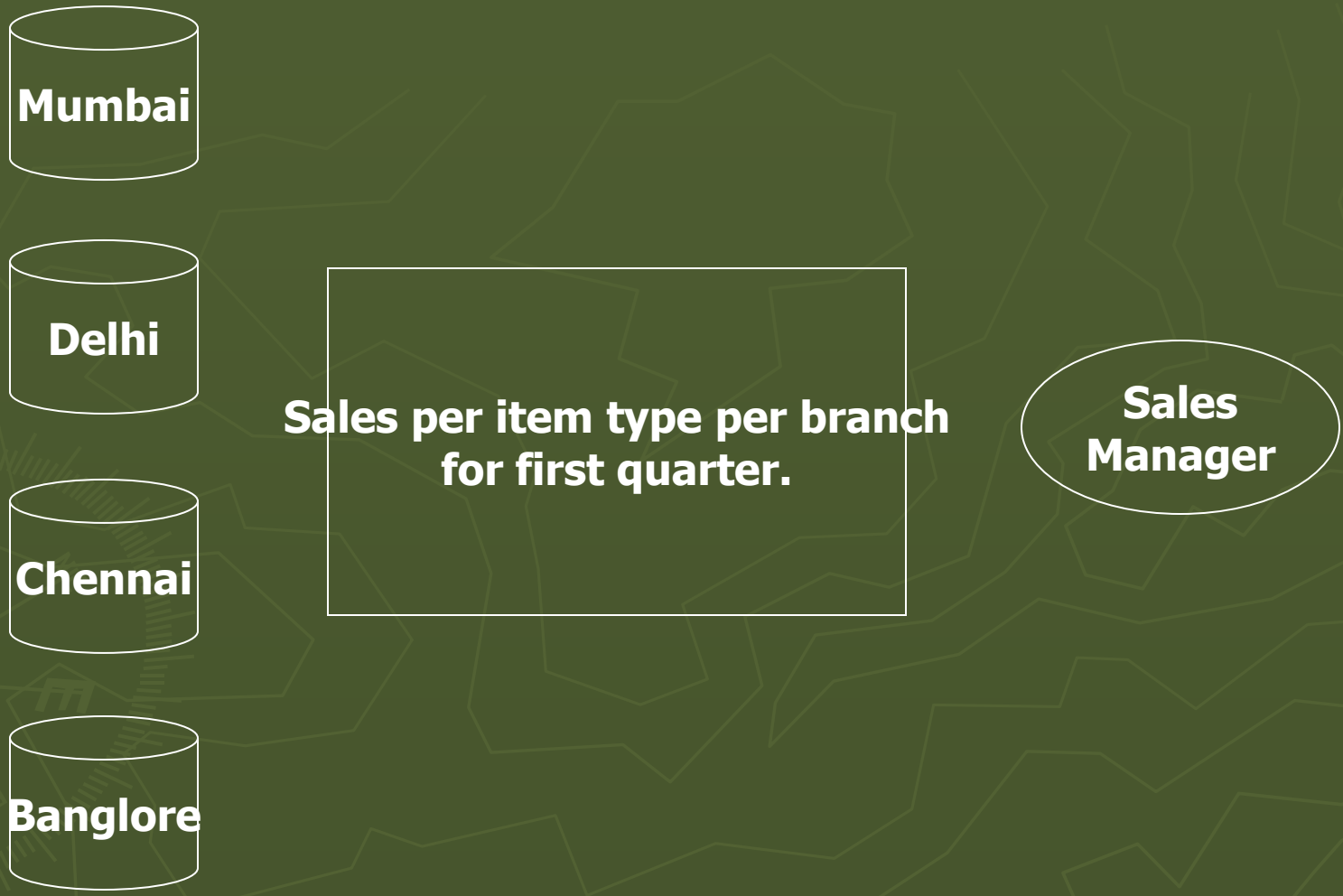
Scenario 1

ABC Pvt. Ltd is a company with branches at Mumbai, Delhi, Chennai and Bangalore.

The Sales Manager wants quarterly sales report.

Each branch has a separate operational system.

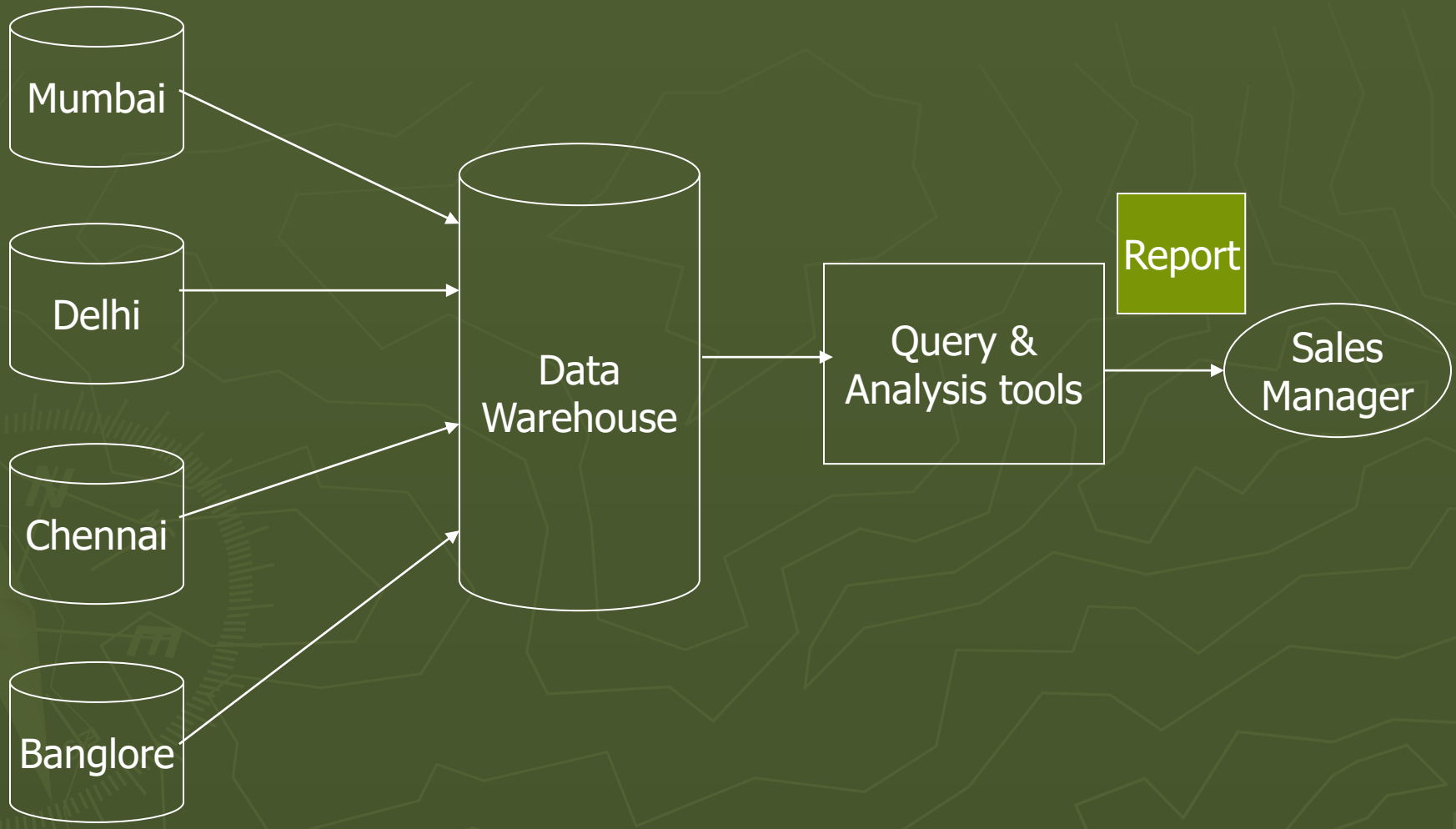
Scenario 1 : ABC Pvt Ltd.



Solution 1:ABC Pvt Ltd.

- Extract sales information from each database.
- Store the information in a common repository at a single site.

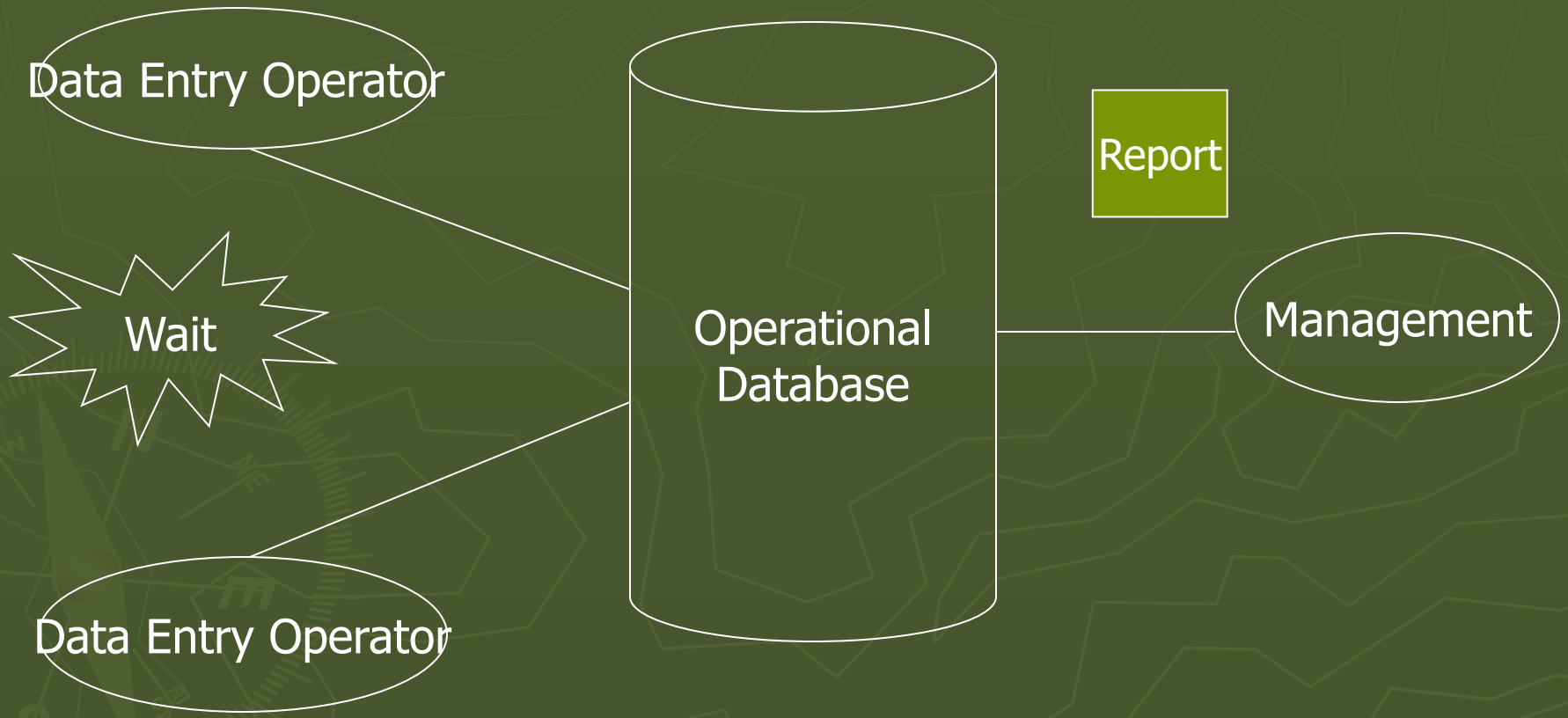
Solution 1:ABC Pvt Ltd.



Scenario 2

One Stop Shopping Super Market has huge operational database.
Whenever Executives wants some report the OLTP system becomes slow and data entry operators have to wait for some time.

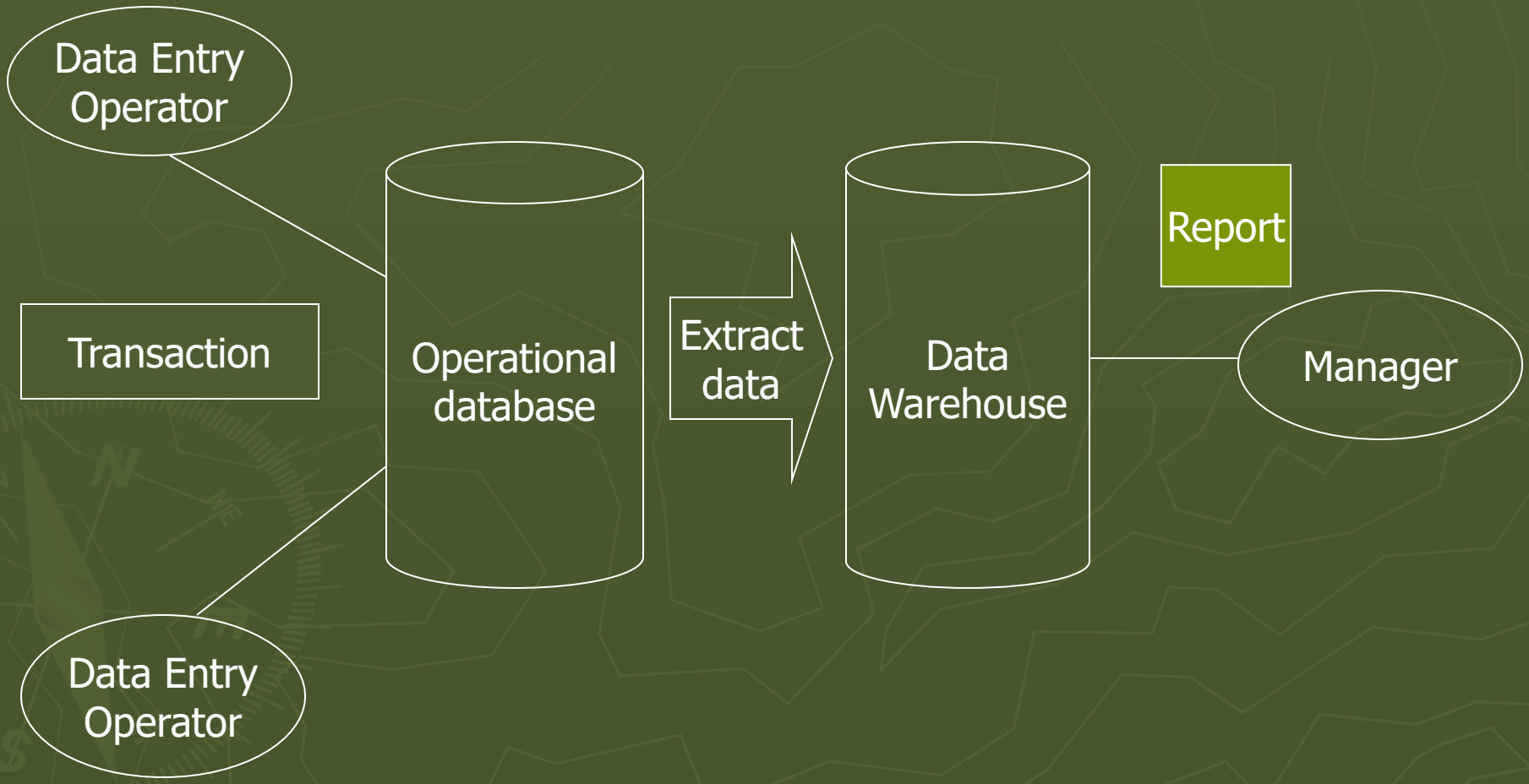
Scenario 2 : One Stop Shopping



Solution 2

- Extract data needed for analysis from operational database.
- Store it in another system, the data warehouse.
- Refresh warehouse at regular intervals so that it contains up to date information for analysis.
- Warehouse will contain data with historical perspective.

Solution 2


Data Entry Operator
Transaction
Data Entry Operator
Operational database
Extract data
Data Warehouse
Report
Manager

Scenario 3

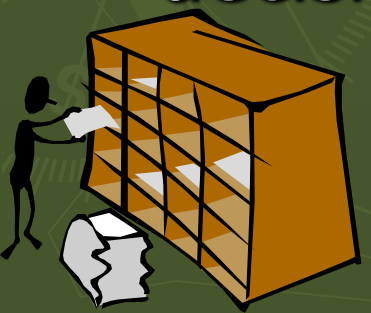
Cakes & Cookies is a small, new company. The chairman of this company wants his company to grow. He needs information so that he can make correct decisions.

Solution 3

- Improve the quality of data before loading it into the warehouse.
- Perform data cleaning and transformation before loading the data.
- Use query analysis tools to support adhoc queries.

Why a Data Warehouse (DWH)?

- Data recording and storage is growing.
- History is excellent predictor of the future.
- Gives total view of the organization.
- Intelligent decision-support is required for decision-making.



Reason-1: Why a Data Warehouse?

- Data Sets are growing.

How Much Data is that?

Reason-1: Why a Data Warehouse?

- Size of Data Sets are going up ↑.
- Cost of data storage is coming down ↓.
  - The amount of data average business collects and stores is **doubling every year**
  - Total hardware and software cost to store and manage 1 Mbyte of data
    - 1990: ~ $15
    - 2002: ~ ¢15 (Down 100 times)
    - By 2007: < ¢1 (Down 150 times)

Reason-1: Why a Data Warehouse?

- A Few Examples
  - WalMart: 24 TB
  - France Telecom: ~ 100 TB
  - CERN: Up to 20 PB by 2006
  - Stanford Linear Accelerator Center (SLAC): 500TB

Caution!

A Warehouse of Data
is NOT a
Data Warehouse

Caution!

Size
is NOT
Everything

Reason-2: Why a Data Warehouse?

- Businesses demand Intelligence (BI).
  - Complex questions from integrated data.
  - "Intelligent Enterprise"

Reason-2: Why a Data Warehouse?

DBMS Approach

List of all items that were sold last month?

List of all items purchased by Tariq Majeed?

The total sales of the last month grouped by branch?

How many sales transactions occurred during the month of January?

Reason-2: Why a Data Warehouse?

Intelligent Enterprise

Which items sell together? Which items to stock?

Where and how to place the items? What discounts to offer?

How best to target customers to increase sales at a branch?

Which customers are most likely to respond to my next promotional campaign, and why?

Reason-3: Why a Data Warehouse?

- Businesses want much more…
  - What happened?
  - Why it happened?
  - What will happen?
  - What is happening?
  - What do you want to happen?

Stages of Data Warehouse

What is a Data Warehouse?

A complete repository of historical corporate data extracted from transaction systems that is available for ad-hoc access by knowledge workers.

What is a Data Warehouse?

- Complete repository
- History
- Transaction System
- Ad-Hoc access
- Knowledge workers

What is a Data Warehouse?

Transaction System

- Management Information System (MIS)
- Could be typed sheets (NOT transaction system)

Ad-Hoc access

- Dose not have a certain access pattern.
- Queries not known in advance.
- Difficult to write SQL in advance.

Knowledge workers

- Typically NOT IT literate (Executives, Analysts, Managers).
- NOT clerical workers.
- Decision makers.

What is a Data Warehouse ?

It is a blend of many technologies, the basic concept being:

- Take all data from different operational systems.
- If necessary, add relevant data from industry.
- Transform all data and bring into a uniform format.
- Integrate all data as a single entity.

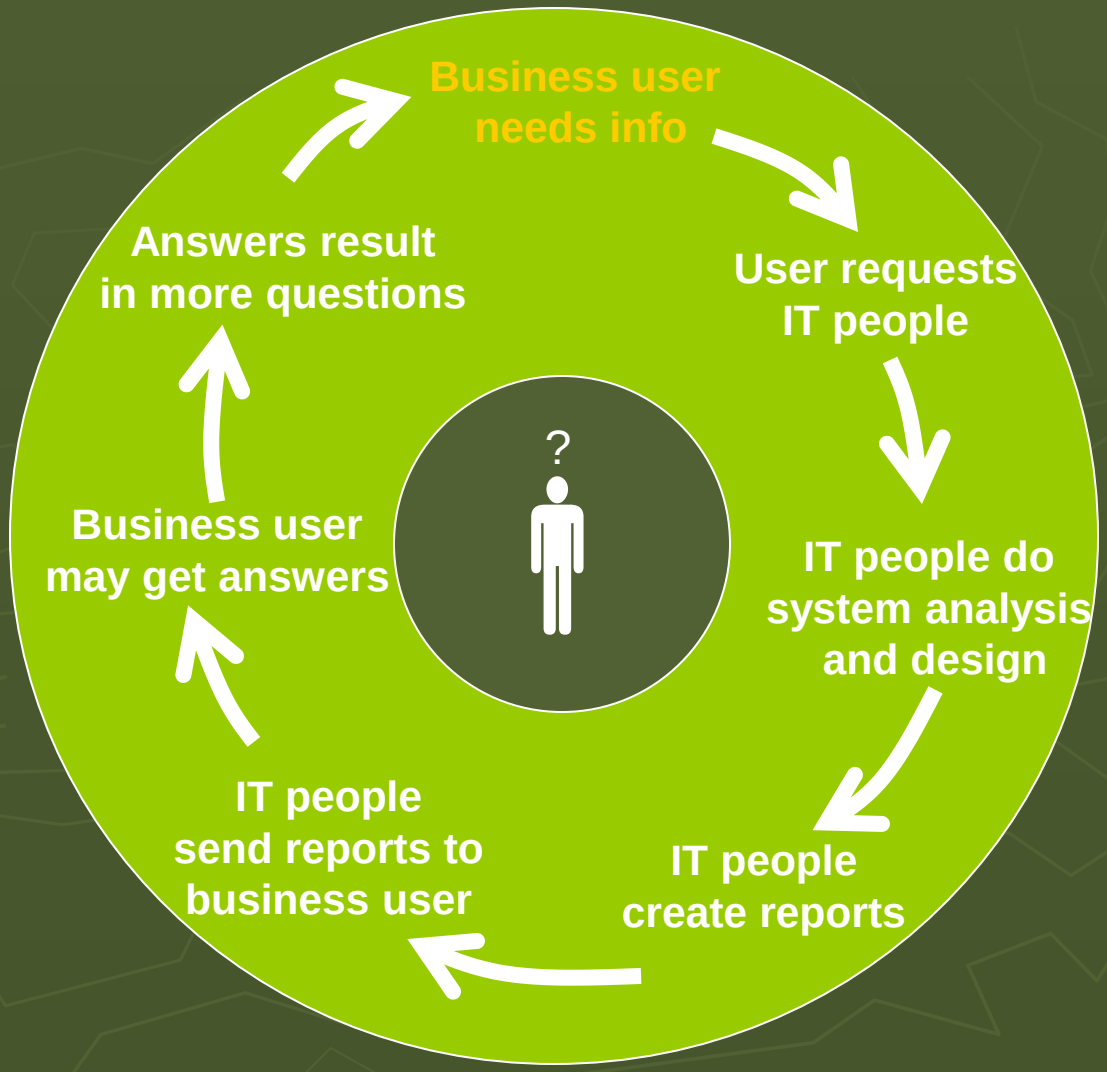
What is a Data Warehouse ? (Cont…)

It is a blend of many technologies, the basic concept being:

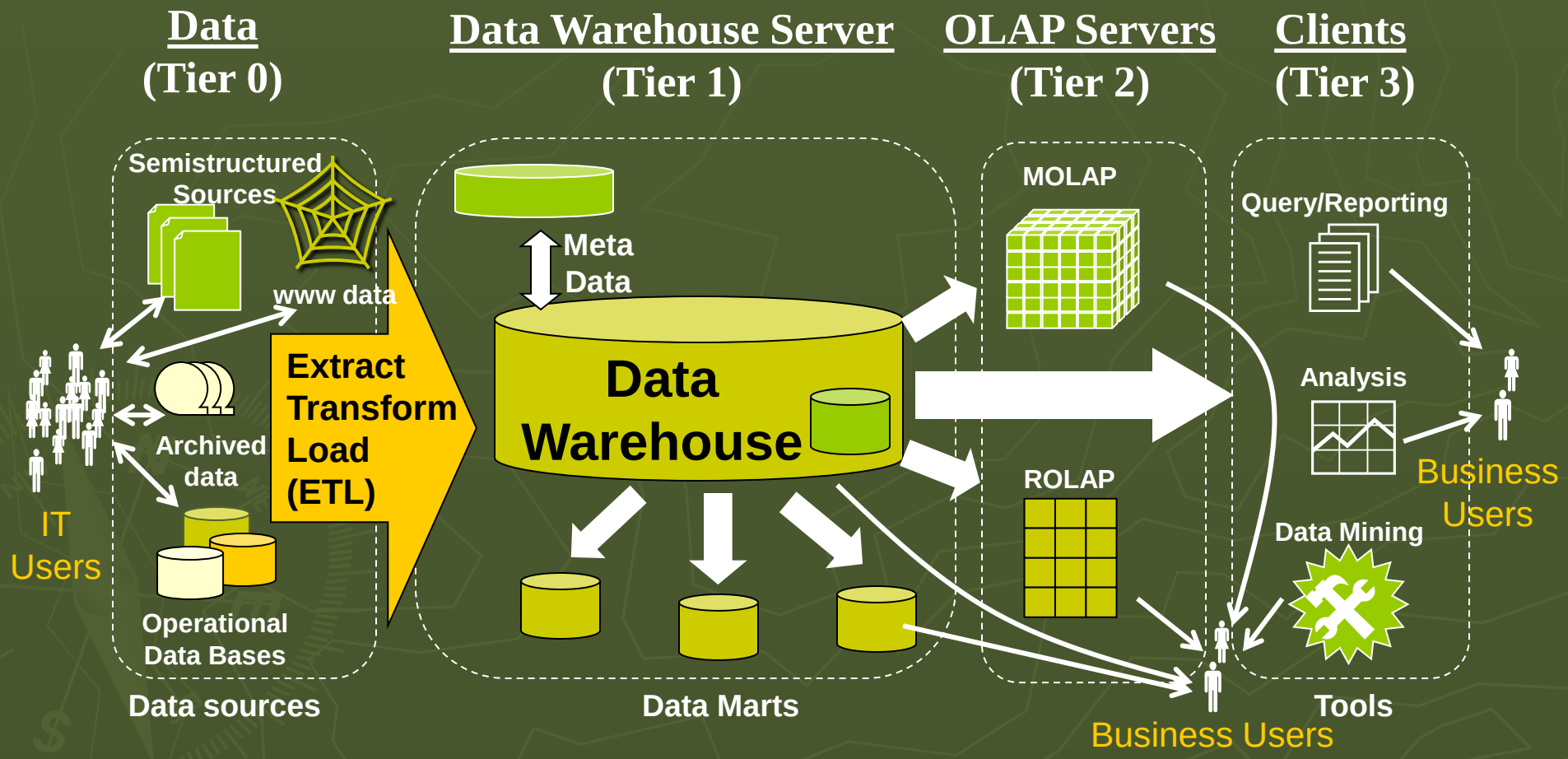
- Store data in a format supporting easy access for decision support.
- Create performance enhancing indices.
- Implement performance enhancement joins.
- Run ad-hoc queries with low selectivity.

How is it Different?

- Fundamentally different



Putting the pieces together



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.

Misconception about data Mart

- A data mart is not warehouse.
- A data mart is not just a small data warehouse
- A collection of data mart is not a data warehouse
- A data warehouse is not DSS

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.

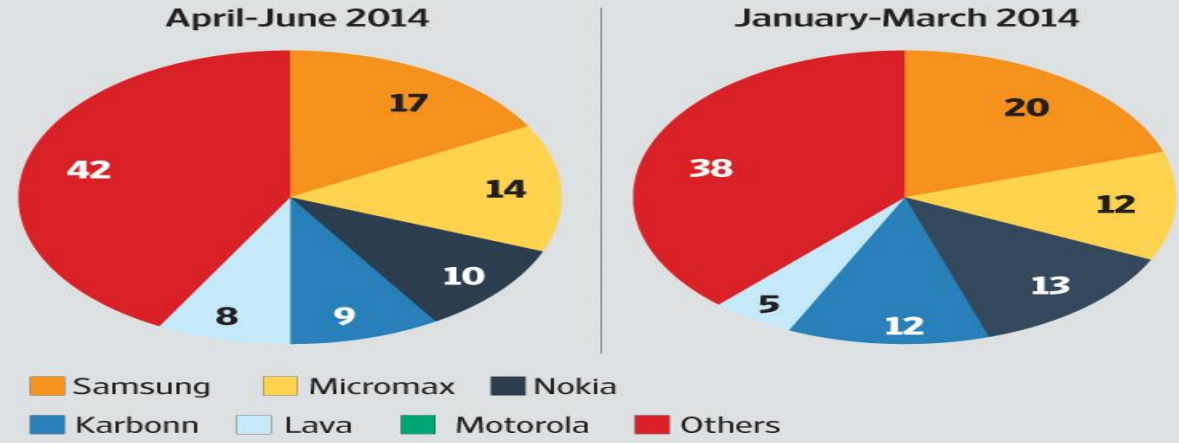
Characteristics of Data Warehouse

- **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.

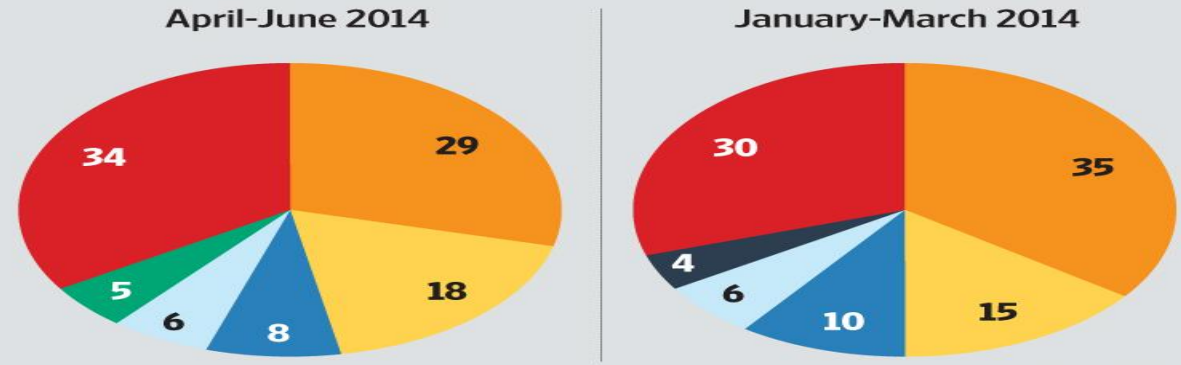
SAMSUNG STILL NO. 1: IDC REPORT

Samsung is still No.1 in India, both in the overall handset market and the rapidly growing smartphone segment, but faces a "real possibility" of losing its position to home-grown brands such as Micromax, according to a new report released on Monday. Research firm International Data Corp. (IDC) said in a report that Samsung stood first with 17% market share in the June quarter of 2014 followed by Micromax with 14%, and Nokia with 10%.

OVERALL MOBILE PHONE MARKET SHARE (%)



SMARTPHONE MARKET SHARE (%)

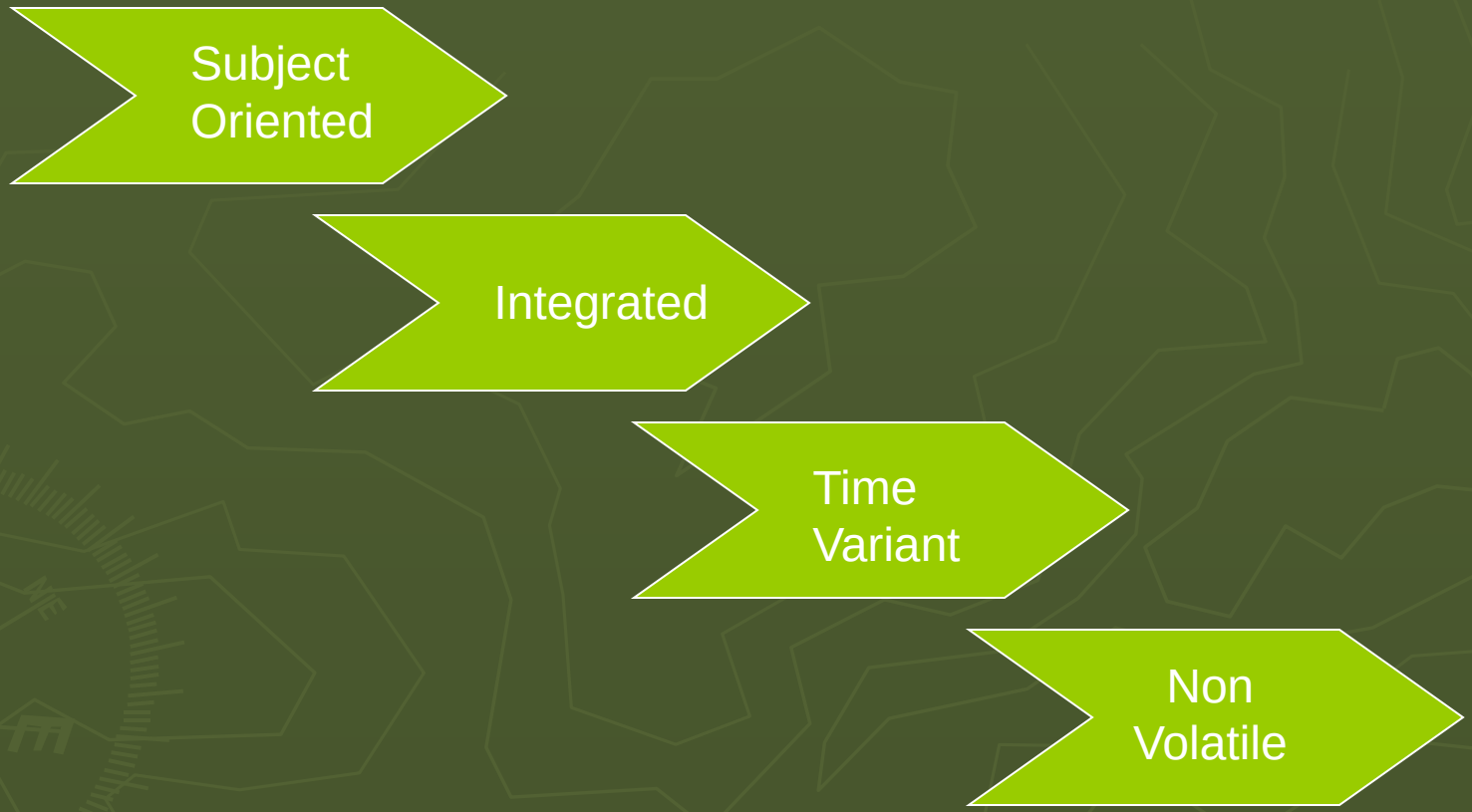


Source: IDC

Operational Data

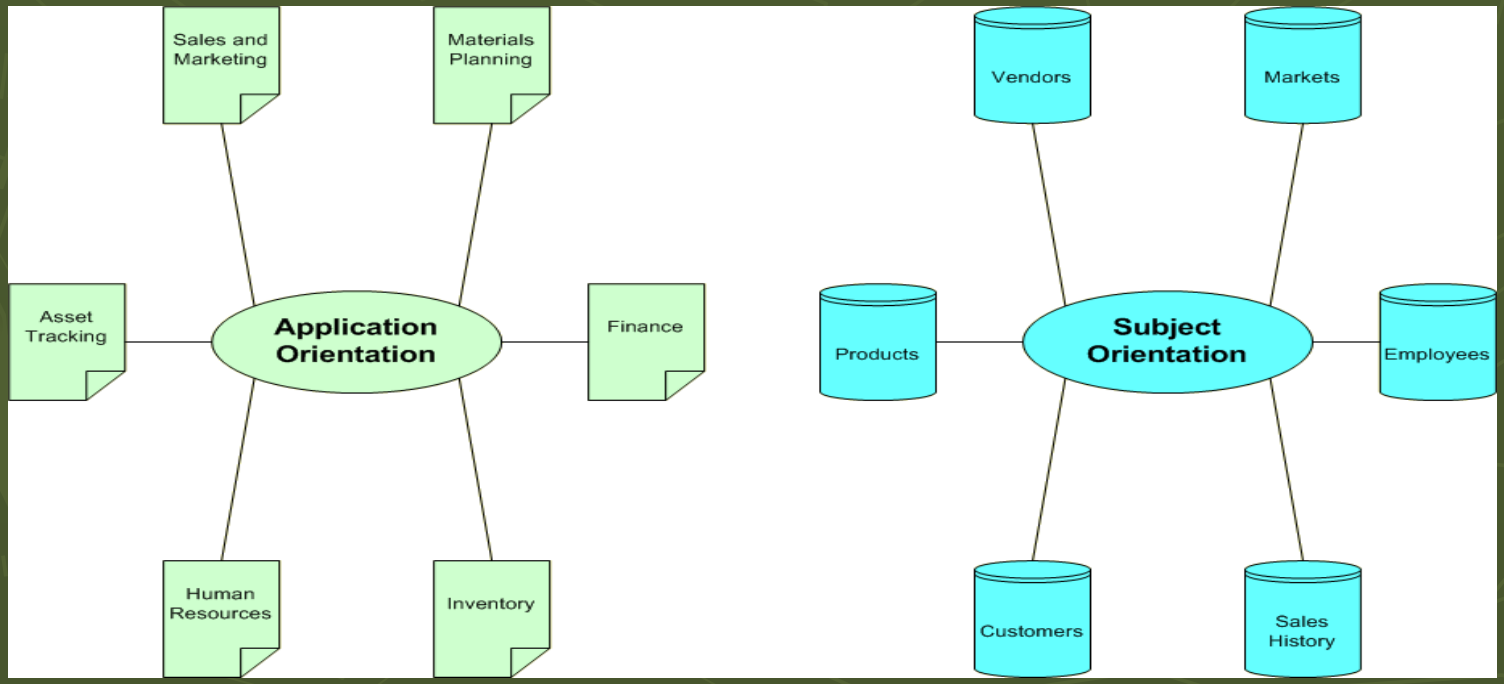
Informational data

Characteristics of Data Warehouse



Subject Orientation

Application Environment	Data warehouse Environment
Design activities must be equally focused on both process and database design	DW world is primarily void of process design and tends to focus exclusively on issues of data modeling and database design



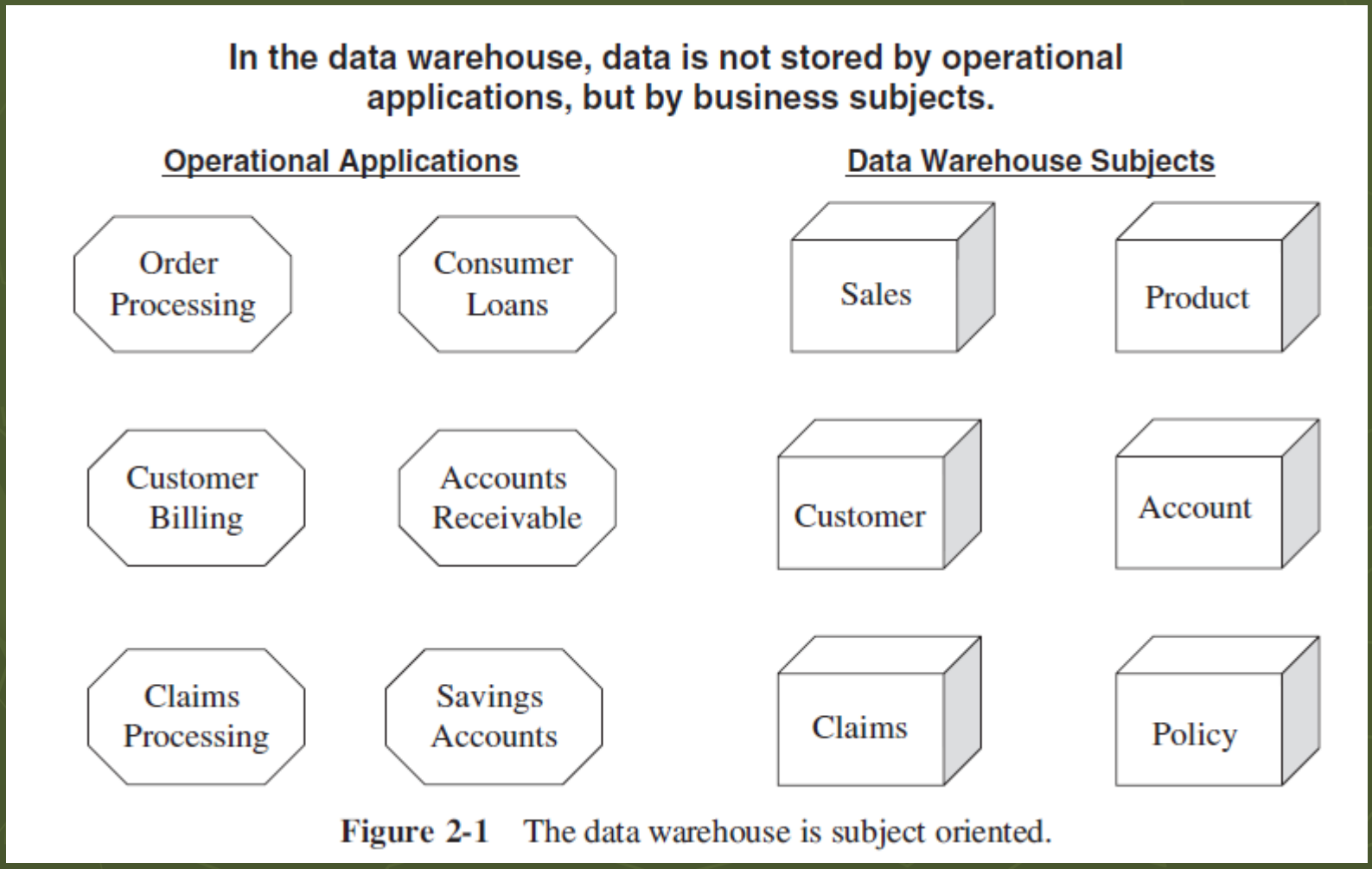


Figure 2-1 The data warehouse is subject oriented.

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

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)

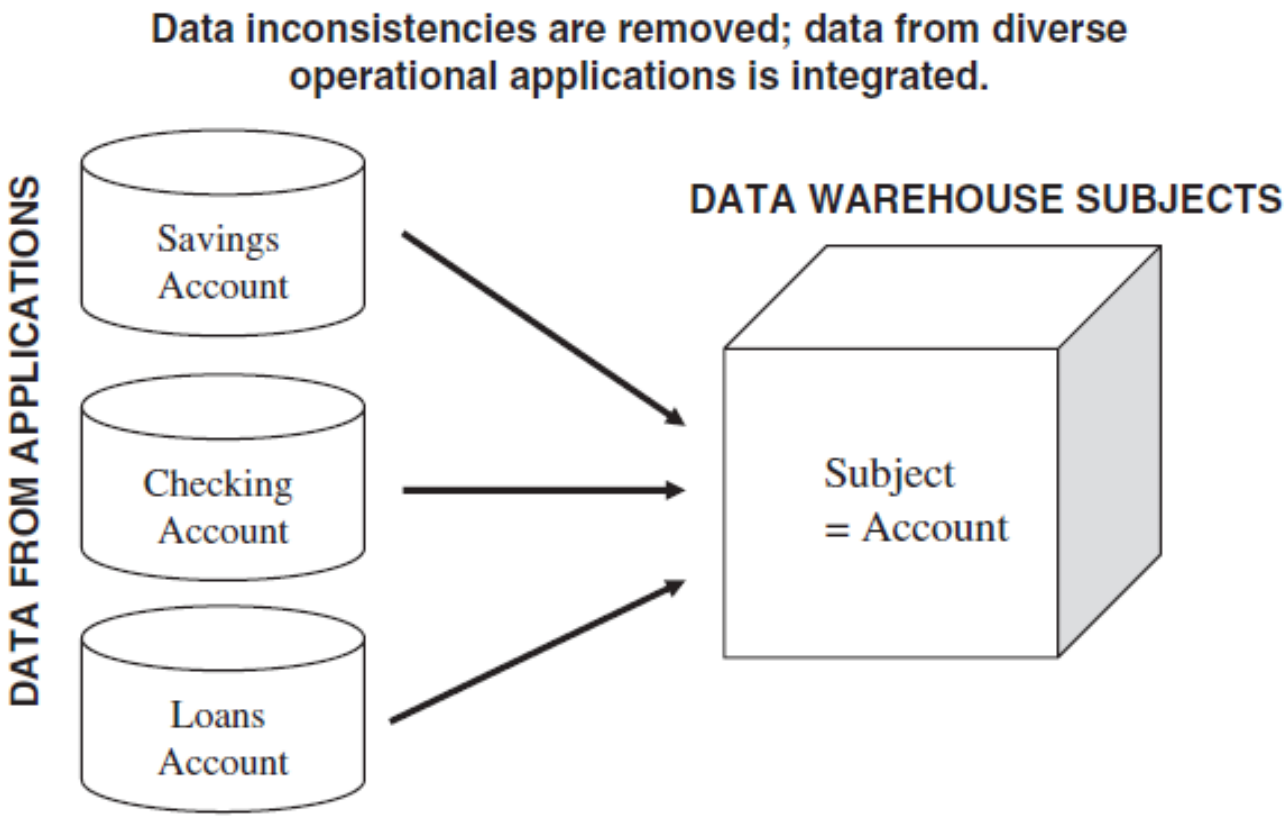


Figure 2-2 The data warehouse is integrated.

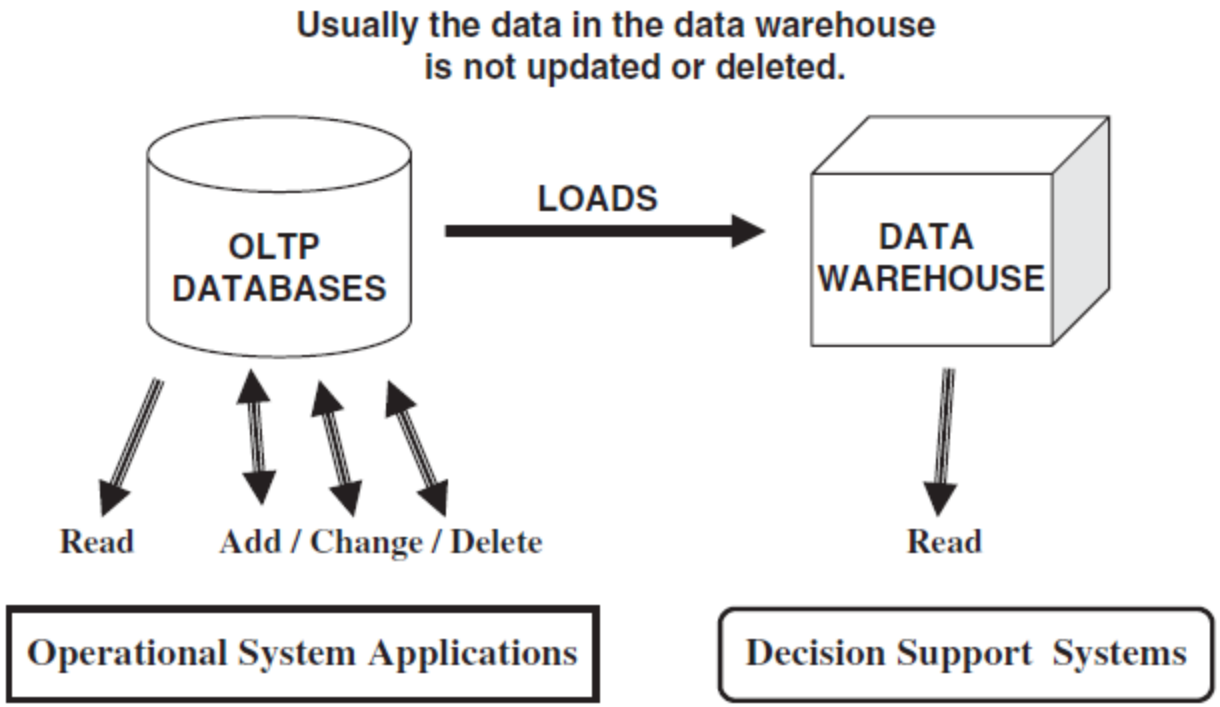


Figure 2-3 The data warehouse is nonvolatile.

Nonvolatility

Application	DW
The design issues must focus on data integrity and update anomalies. Complex processes must be coded to ensure that the data update processes allow for high integrity of the final product.	Such issues are no concern to in a DW environment because data update is never performed.
Data is placed in normalized form to ensure a minimal redundancy (totals that could be calculated would never be stored)	Designers find it useful to store many of such calculations or summarizations.
The technologies necessary to support issues of transaction and data recovery, roll back, and detection and remedy of deadlock are quite complex.	Relative simplicity in technology

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

Daily Detail	Monthly Summary	Quarterly Summary
Account	Account	Account
Activity Date	Month	Quarter
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present. Many data warehouses have at least dual levels of granularity.

Figure 2-4 Data granularity.

DATA WAREHOUSE	DATA MART
◆ Corporate/Enterprise-wide	◆ Departmental
◆ Union of all data marts	◆ A single business process
◆ Data received from staging area	◆ STARjoin (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.

What is a Data Warehouse Architecture

- Primarily based on the business processes of a business enterprise
- Conceptualization of how the data warehouse is built

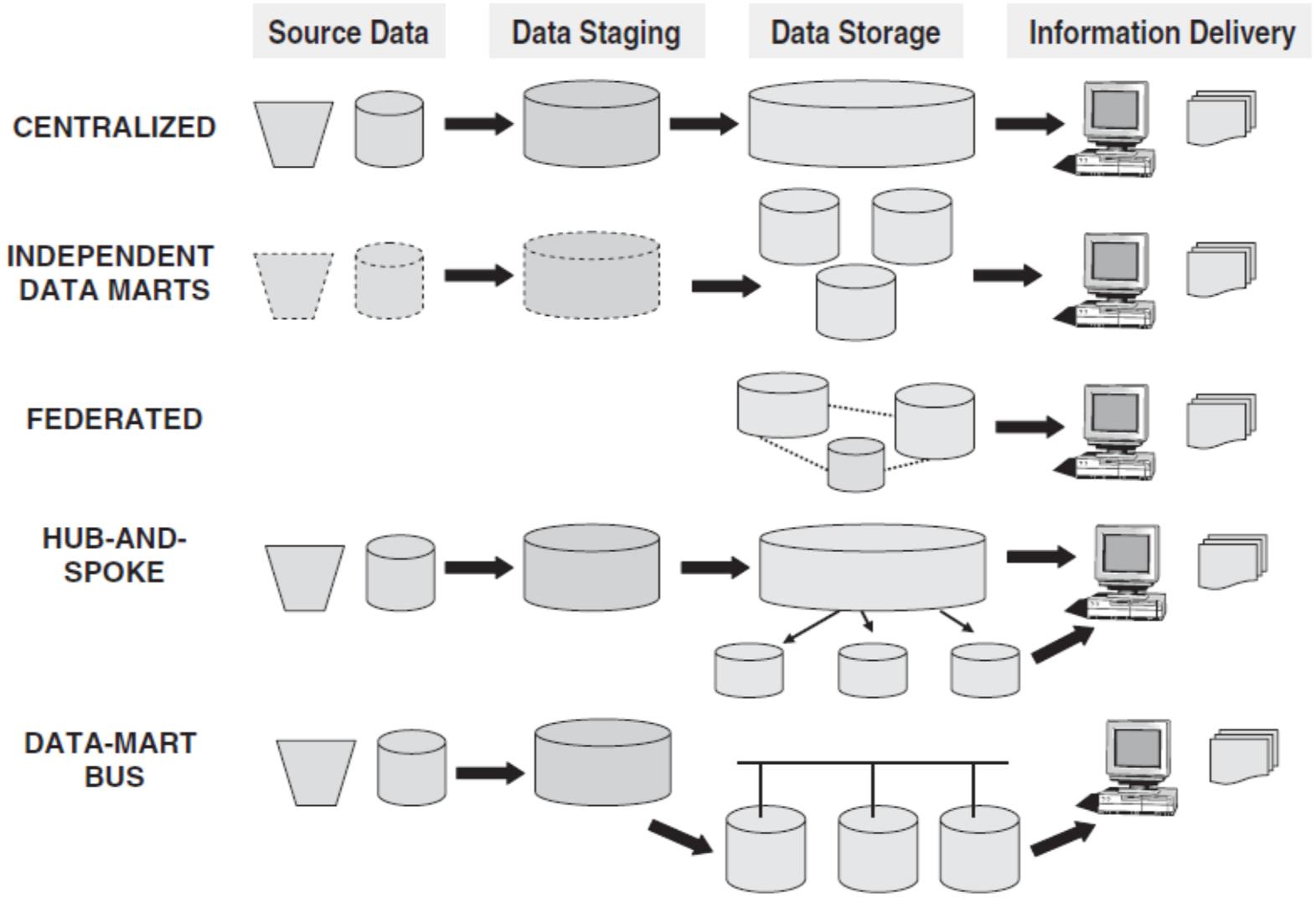


Figure 2-6 Data warehouse architectural types.

Architecture is the proper arrangement of the components.

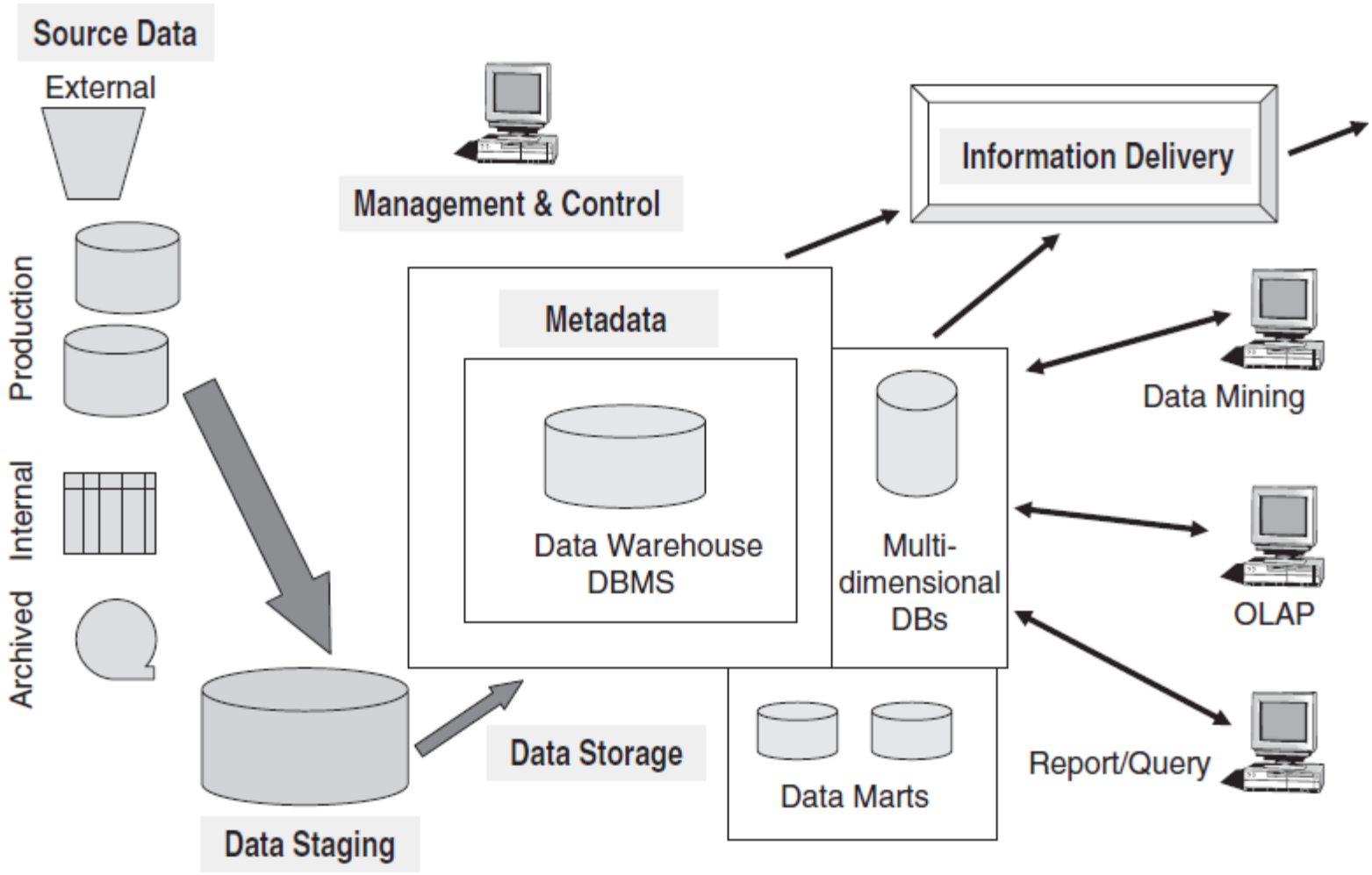


Figure 2-7 Data warehouse: building blocks or components.

Components

- Major components
  - Source data component
  - Data staging component
  - Information delivery component
  - Metadata component
  - Management and control component

1. Source Data Components

- Source data can be grouped into 4 components
  - Production data
    - Comes from operational systems of enterprise
    - Some segments are selected from it
    - Narrow scope, e.g. order details
  - Internal data
    - Private datasheet, documents, customer profiles etc.
    - E.g. Customer profiles for specific offering
    - Special strategies to transform 'it' to DW (text document)
  - Archived data
    - Old data is archived
    - DW have snapshots of historical data
  - External data
    - Executives depend upon external sources
    - E.g. market data of competitors, car rental require new manufacturing. Define conversion

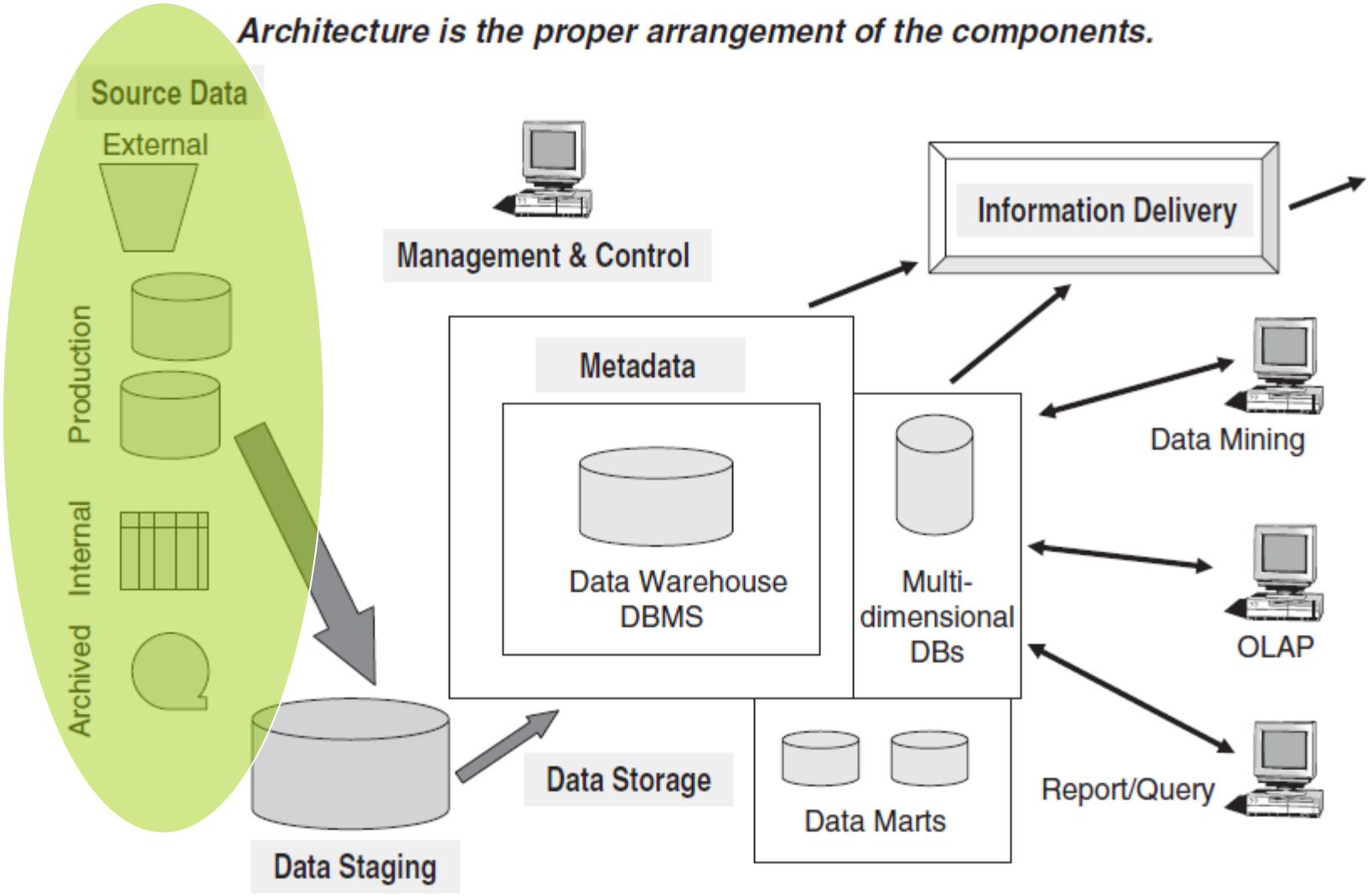


Figure 2-7 Data warehouse: building blocks or components.

2. Data Staging Components

- After data is extracted, data is to be prepared
- Data extracted from sources needs to be changed, converted and made ready in suitable format
- Three major functions to make data ready
  - Extract
  - Transform
  - Load
- Staging area provides a place and area with a set of functions to
  - Clean
  - Change
  - Combine
  - Convert

Architecture is the proper arrangement of the components.

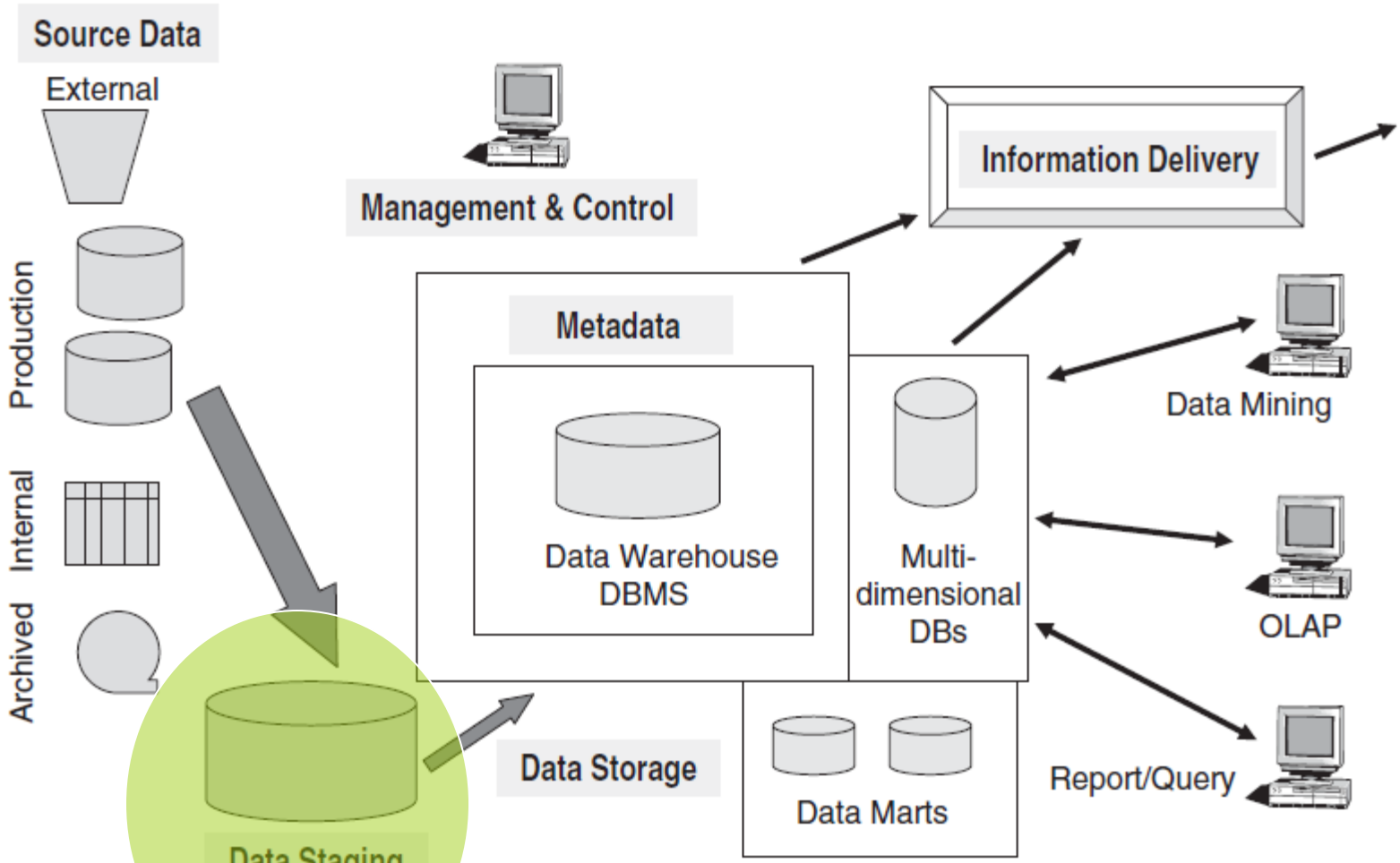


Figure 2-7 Data warehouse: building blocks or components.

3. Data Storage Components

- Separate repository
- Data structured for efficient processing
- Redundancy is increased
- Updated after specific periods
- Only read-only

Architecture is the proper arrangement of the components.

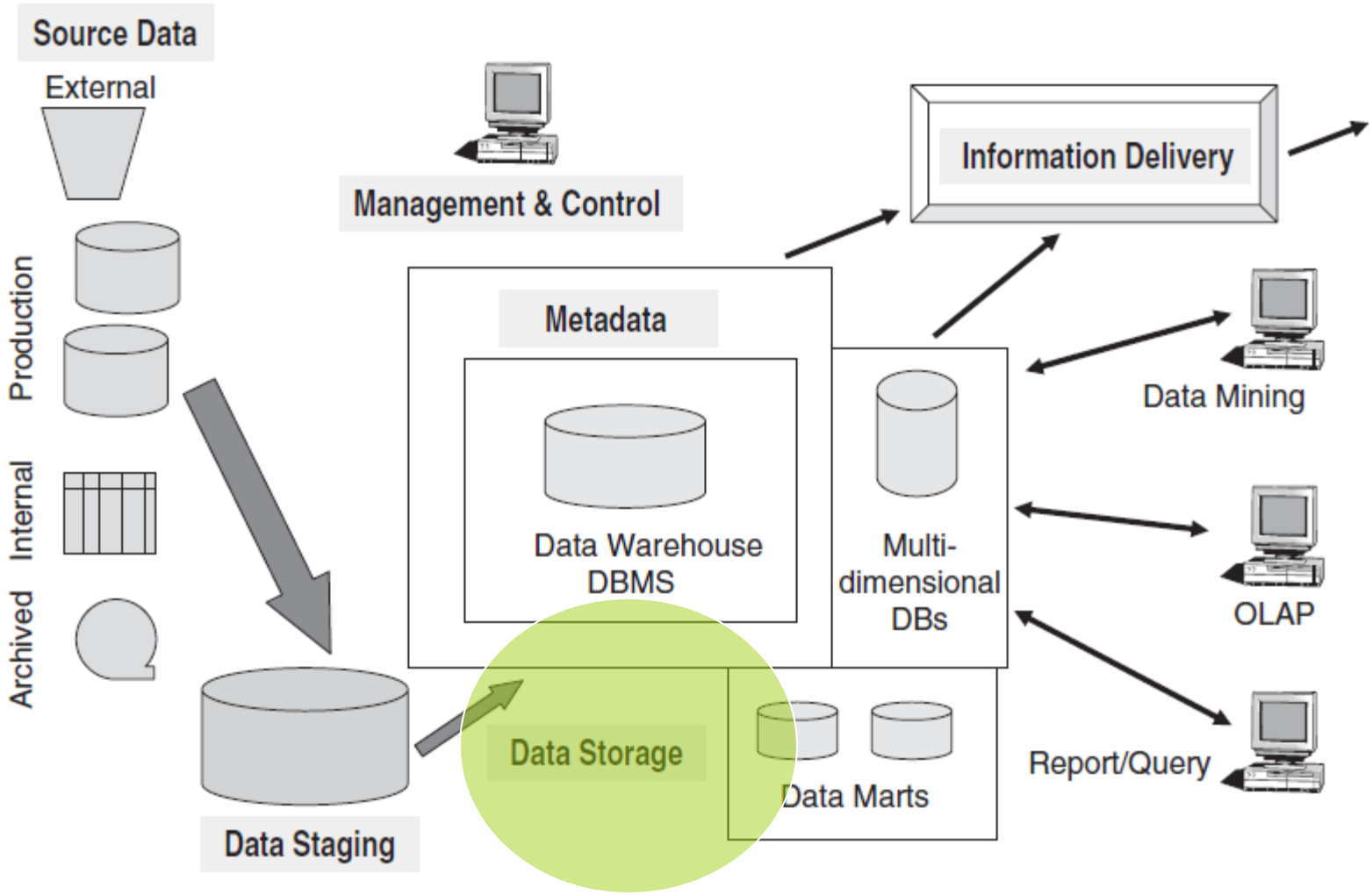


Figure 2-7 Data warehouse: building blocks or components.

4. Information Delivery Component

- Authentication issues
- Active monitoring services
  - Performance, DBA note selected aggregates to change storage
  - User performance
  - Aggregate awareness
  - E.g. mining, OLAP etc

Architecture is the proper arrangement of the components.

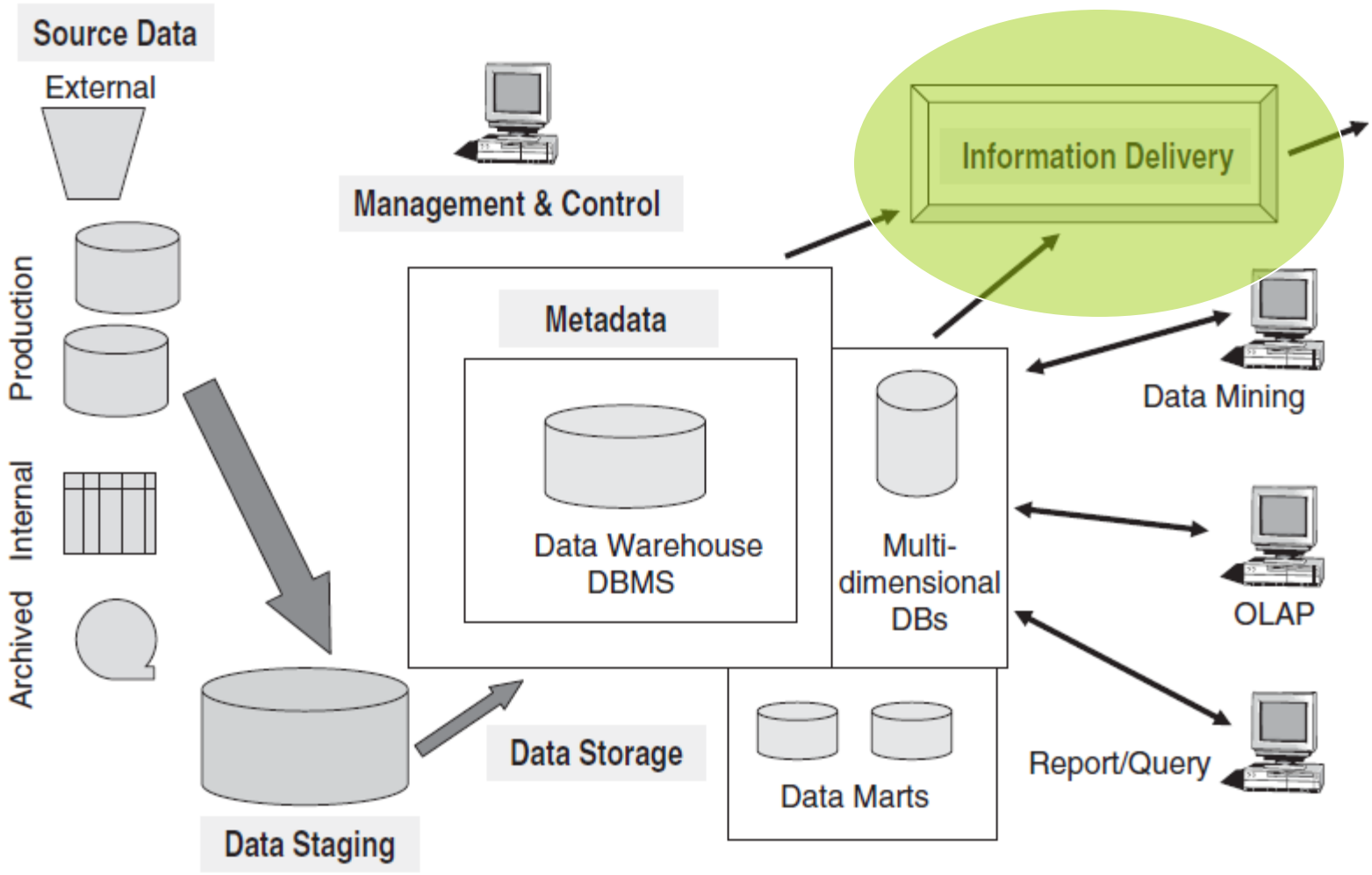


Figure 2-7 Data warehouse: building blocks or components.

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.

Architecture is the proper arrangement of the components.

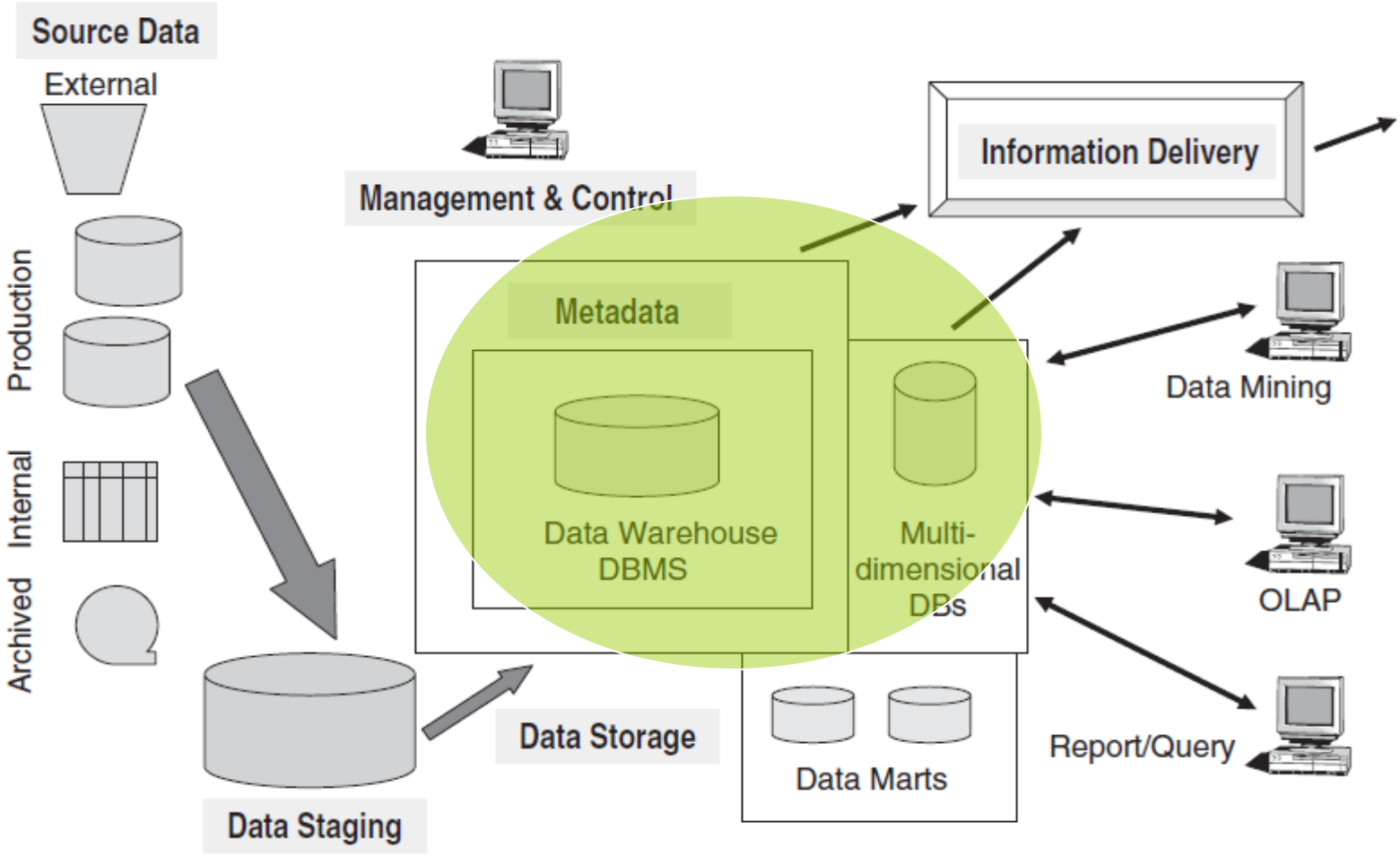


Figure 2-7 Data warehouse: building blocks or components.

Metadata

- Information about data ware house ,location, Structure
- Information regarding refreshment of warehouse cleanup
- Information regarding security authentication and usage statistics
- Information regarding characteristics of components

Significance Of meta data

- It act as glue that connect all part of data warehouse.
- It provide information about the contain and structure to the developers
- It open the door to the end users and makes the contents recognizable in their own terms.

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.

Types of Metadata

- Operational Metadata
- Extraction and transformation Metadata
- Enduser Metadata
- Semantic metadata

General Metadata Issues

General metadata issues associated with Data Warehouse use:

- What tables, attributes and keys does the DW contain?
- Where did each set of data come from?
- What transformations were applied with cleansing?
- How have the metadata changed over time?
- How often do the data get reloaded?
- Are there so many data elements that you need to be careful what you ask for?

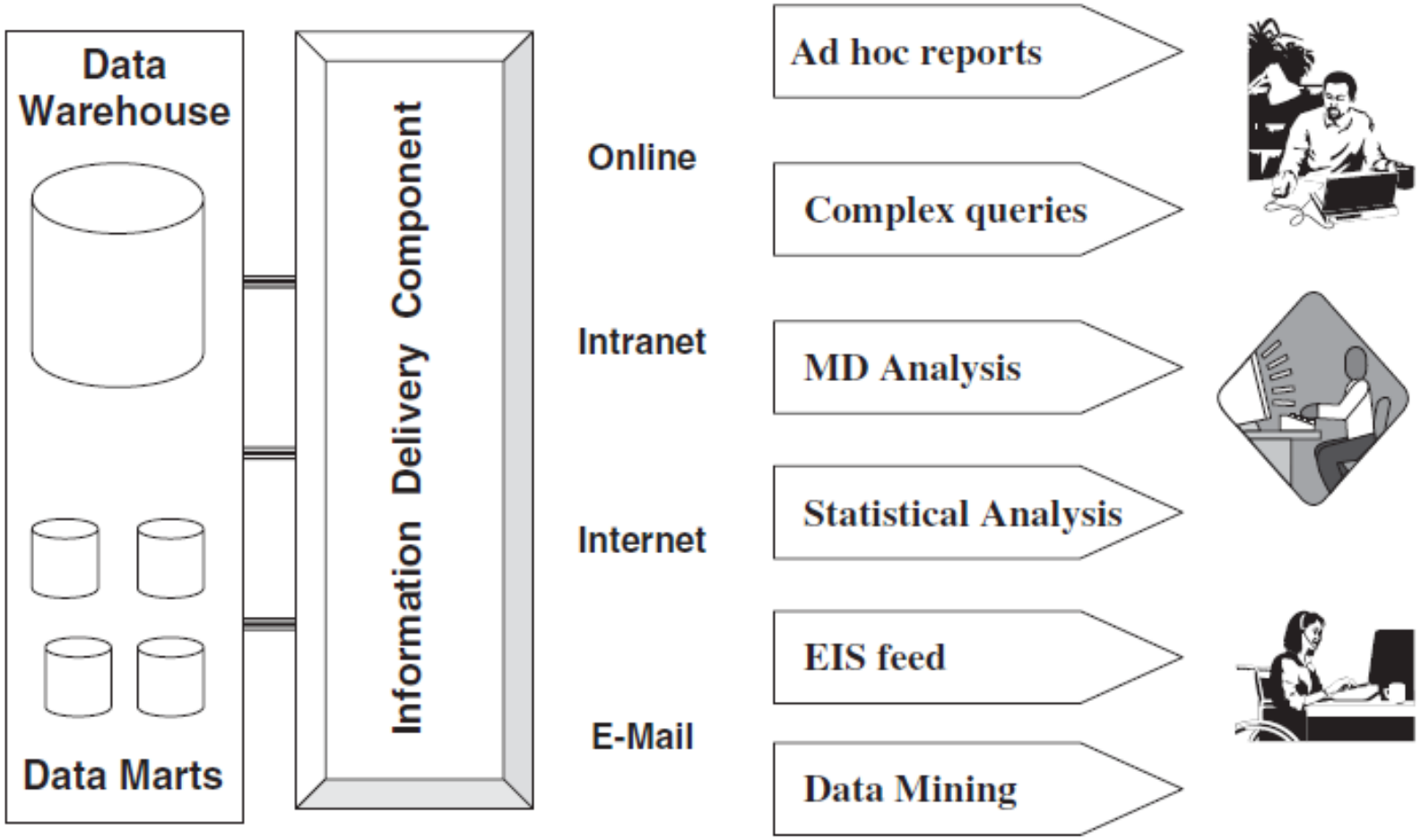


Figure 2-9 Information delivery component.

1. Match the columns:

1. nonvolatile data	A. roadmap for users
2. dual data granularity	B. subject-oriented
3. dependent data mart	C. knowledge discovery
4. disparate data	D. private spreadsheets
5. decision support	E. application flavor
6. data staging	F. because of multiple sources
7. data mining	G. details and summary
8. metadata	H. read-only
9. operational systems	I. workbench for data integration
10. internal data	J. data from main data warehouse

2. A data warehouse is subject-oriented. What would be the major critical business subjects for the following companies?

   a. an international manufacturing company

   b. a local community bank

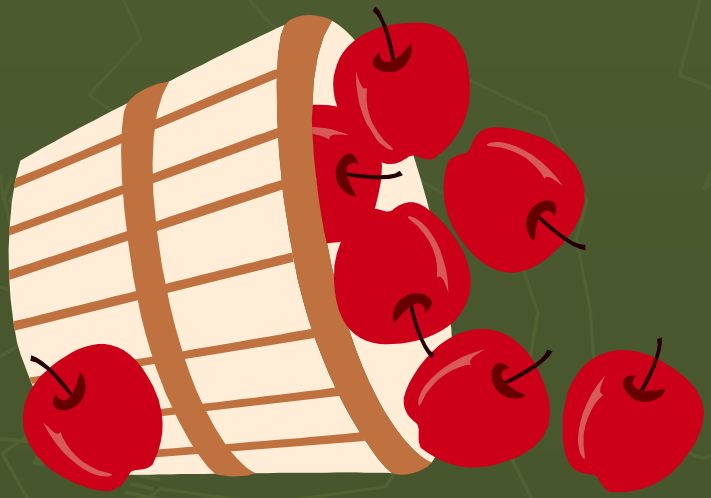
   c. a domestic hotel chain

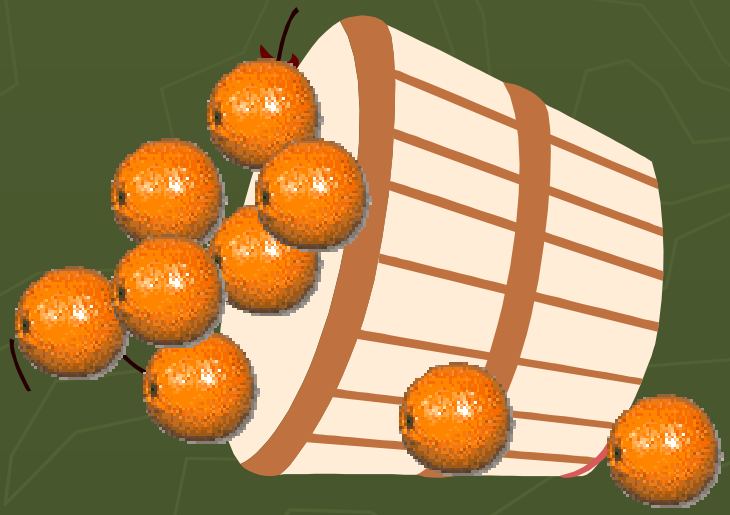
How much history?

Economic value of data
Vs.
Storage cost

Data Warehouse a complete repository of data?

So, what's different?





Data Warehouse Vs. OLTP

OLTP (On Line Transaction Processing)

Select tx_date, balance from tx_table
Where account_ID = 23876;

Data Warehouse Vs. OLTP

OLTP	DWH
Primary key used	Primary key NOT used
No concept of Primary Index	Primary index used
Few rows returned	Many rows returned
May use a single table	Uses multiple tables
High selectivity of query	Low selectivity of query
Indexing on primary key (unique)	Indexing on primary index (non-unique)

Data Warehouse Vs. OLTP

OLTP: OnLine Transaction Processing (MIS or Database System)

	Data Warehouse	OLTP
Scope	* Application –Neutral * Single source of "truth" * Evolves over time * How to improve business	* Application specific * Multiple databases with repetition * Off the shelf application * Runs the business
Data Perspective	* Historical, detailed data * Some summary * Lightly denormalized	* Operational data * No summary * Fully normalized
Queries	* Hardly uses PK * Number of results returned in thousands	* Based on PK * Number of results returned in hundreds
Time factor	* Minutes to hours * Typical availability 6x12	* Sub seconds to seconds * Typical availability 24x7

To summarize ...

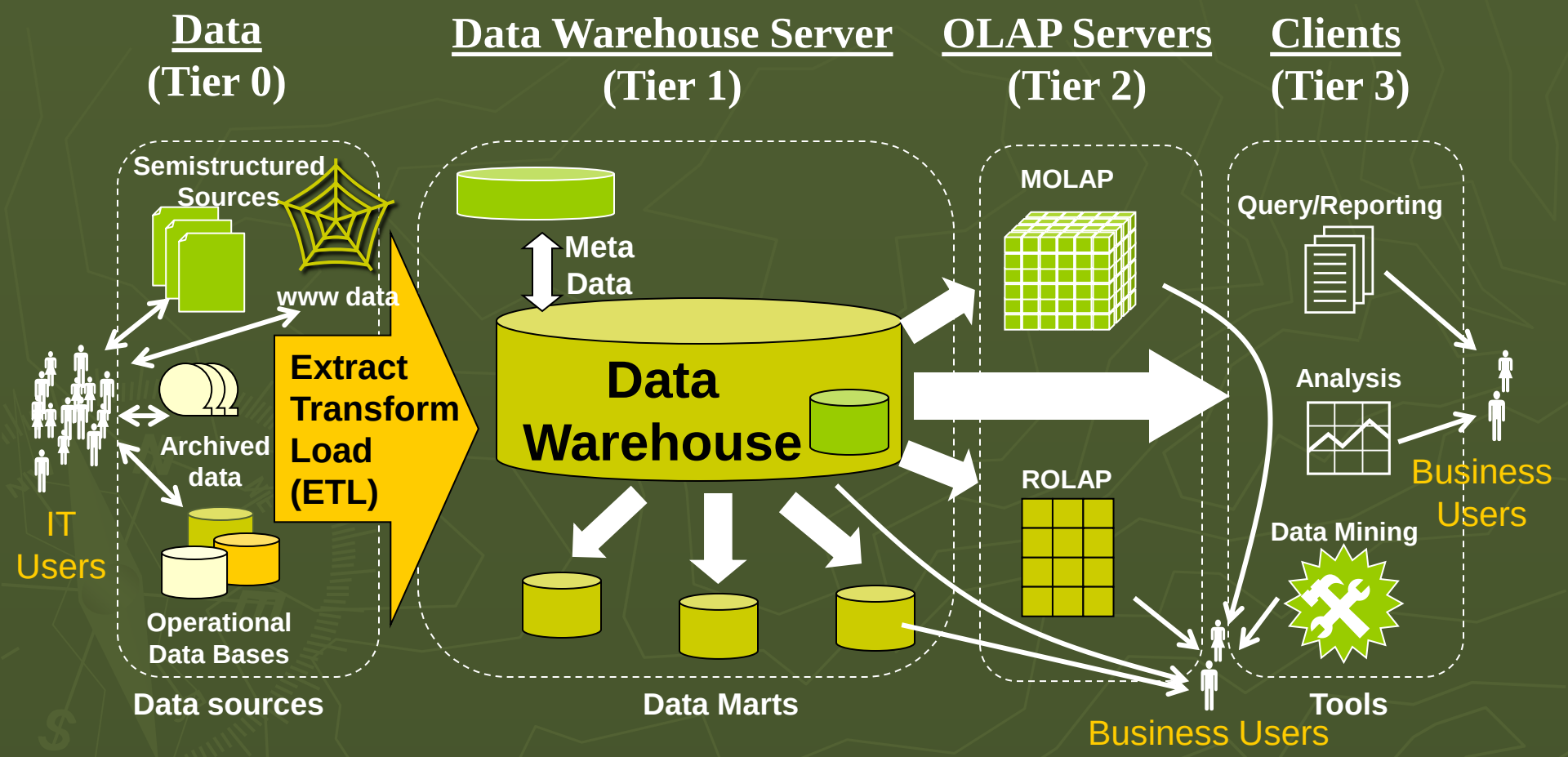
- OLTP Systems are used to *"run"* a business

- The Data Warehouse helps to *"optimize"* the business

Why Now?

- Data is being produced
- ERP provides clean data
- The computing power is available
- The computing power is affordable
- The competitive pressures are strong
- Commercial products are available

Putting the pieces together



Important questions

- What are the different characteristics of a DW?[2]
- Write detailed notes on DW architecture?[5]
- Explain the role of metadata in a DW?[7]
- What is a DW?Explain the 3 tier architecture of DW with a block diagram
- Differences between DW and DM?
- Define DW? Explain what is the need for developing a DW and hence explain its architecture?
- Differentiate between the top-bottom and bottom-top approaches for building a DW? Explain the advantages and disadvantages[2]
- Explain the characteristics of data present in the DW?

Books and Chapter number to refer for Module 1

- Data warehousing "Fundamentals for IT Professionals" – Paulraj Ponniah, Second Edition
- Module 1-Part 1-Overview and Concepts
- Part 2-DW the building block
- Part 3 Architectural types
- Part 3-chapter 8-infrastructure
- Part 3-chapter 9-significant role of metadata