商業智慧
Business Intelligence

資料倉儲
(Data Warehousing)

1002BI04
IM EMBA
Fri 12,13,14 (19:20-22:10) D502

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail. tku.edu.tw/myday/
2012-03-09

課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）	備註
1	101/02/17	商業智慧導論 (Introduction to Business Intelligence)	
2	101/02/24	管理決策支援系統與商業智慧 (Management Decision Support System and Business Intelligence)	
3	101/03/02	企業績效管理 (Business Performance Management)	
4	101/03/09	資料倉儲 (Data Warehousing)	
5	101/03/16	商業智慧的資料探勘 (Data Mining for Business Intelligence)	
6	101/03/24	商業智慧的資料探勘 (Data Mining for Business Intelligence)	
7	101/03/30	個案分析一 (分群分析)： Banking Segmentation (Cluster Analysis – KMeans)	
8	101/04/06	教學行政觀摩日 (--No Class--)	
9	101/04/13	個案分析二 (關連分析)： Web Site Usage Associations (Association Analysis)	

課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）	備註
10	101/04/20	期中報告 (Midterm Presentation)	
11	101/04/27	個案分析三 (決策樹、模型評估)： Enrollment Management Case Study (Decision Tree, Model Evaluation)	
12	101/05/04	個案分析四 (迴歸分析、類神經網路)：Credit Risk Case Study (Regression Analysis, Artificial Neural Network)	
13	101/05/11	文字探勘與網頁探勘 (Text and Web Mining)	
14	101/05/18	智慧系統 (Intelligent Systems)	
15	101/05/25	社會網路分析 (Social Network Analysis)	
16	101/06/01	意見分析 (Opinion Mining)	
17	101/06/08	期末報告1 (Project Presentation 2)	
18	101/06/15	期末報告2 (Project Presentation 2)	

Decision Support and Business Intelligence Systems

(9th Ed., Prentice Hall)

Chapter 8:

Data Warehousing

Learning Objectives

- Definitions and concepts of data warehouses
- Types of data warehousing architectures
- Processes used in developing and managing data warehouses
- Data warehousing operations
- Role of data warehouses in decision support
- Data integration and the extraction, transformation, and load (ETL) processes
- Data warehouse administration and security issues

Main Data Warehousing (DW) Topics

- DW definitions
- Characteristics of DW
- Data Marts
- ODS, EDW, Metadata
- DW Framework
- DW Architecture & ETL Process
- DW Development
- DW Issues

Data Warehouse Defined

- A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format

- "The data warehouse is a collection of integrated, subject-oriented databases design to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time"

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Characteristics of DW

- Subject oriented
- Integrated
- Time-variant (time series)
- Nonvolatile
- Summarized
- Not normalized
- Metadata
- Web based, relational/multi-dimensional
- Client/server
- Real-time and/or right-time (active)

Data Mart

A departmental data warehouse that stores only relevant data

- **Dependent data mart**

  A subset that is created directly from a data warehouse

- **Independent data mart**

  A small data warehouse designed for a strategic business unit or a department

Data Warehousing Definitions

- **Operational data stores (ODS)**

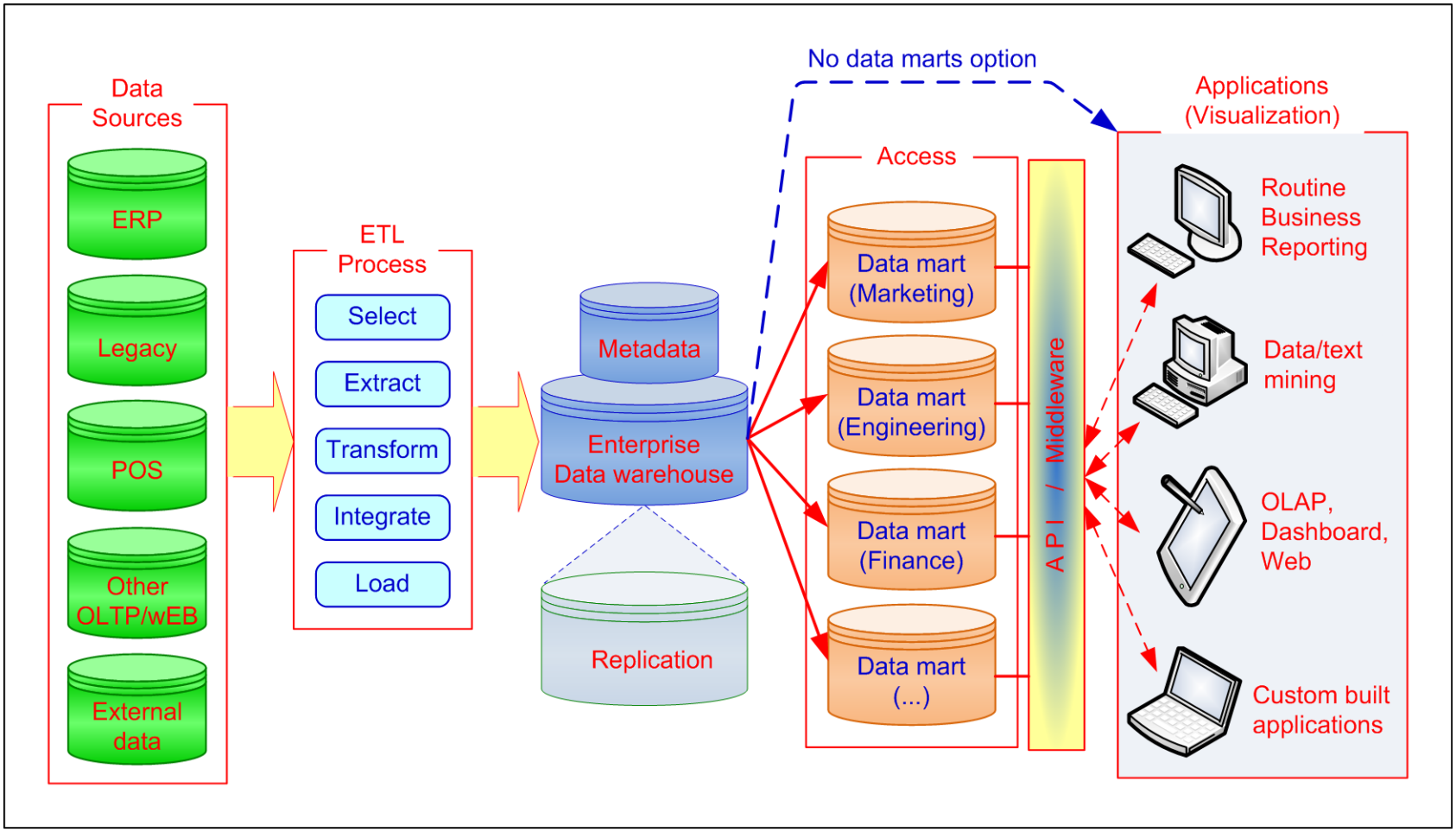
  A type of database often used as an interim area for a data warehouse
- **Oper marts**

  An operational data mart.
- **Enterprise data warehouse (EDW**)

  A data warehouse for the enterprise.
- **Metadata**

  Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its acquisition and use

A Conceptual Framework for DW



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Generic DW Architectures

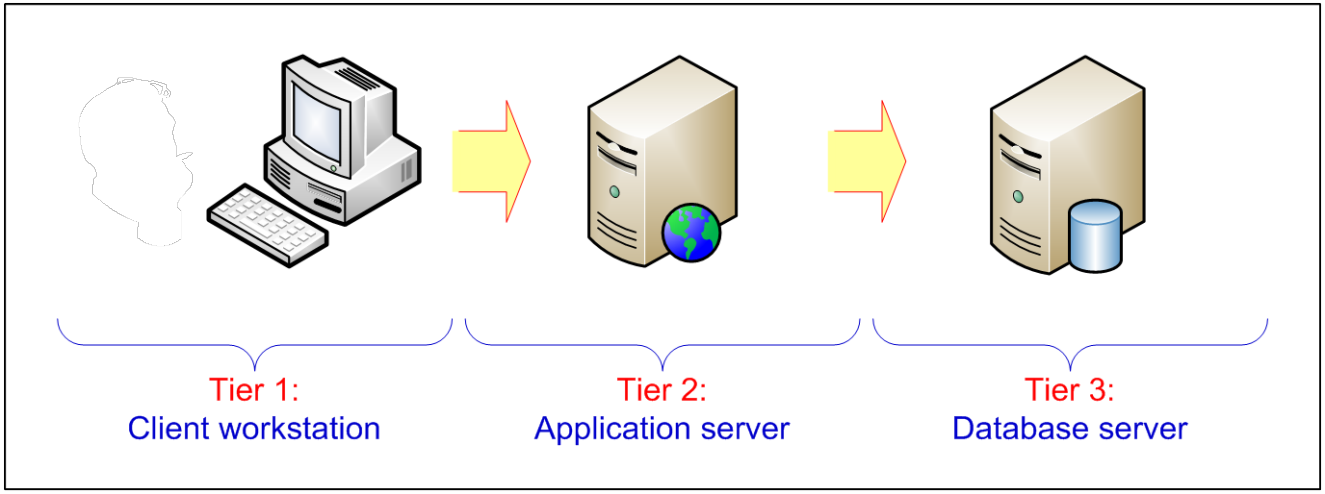
- **Three-tier architecture**
  1. Data acquisition software (back-end)
  2. The data warehouse that contains the data & software
  3. Client (front-end) software that allows users to access and analyze data from the warehouse
- **Two-tier architecture**

  First 2 tiers in three-tier architecture is combined into one

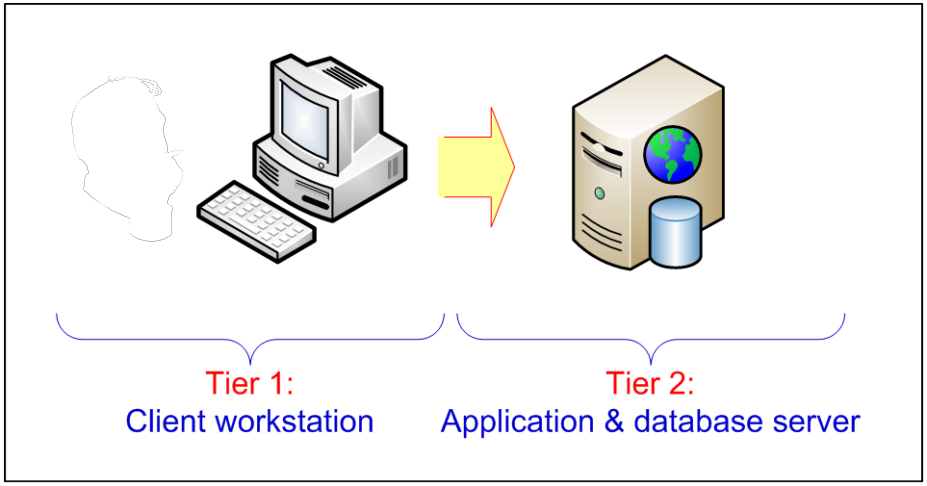
  … sometime there is only one tier?

Generic DW Architectures

3-tier architecture



2-tier architecture



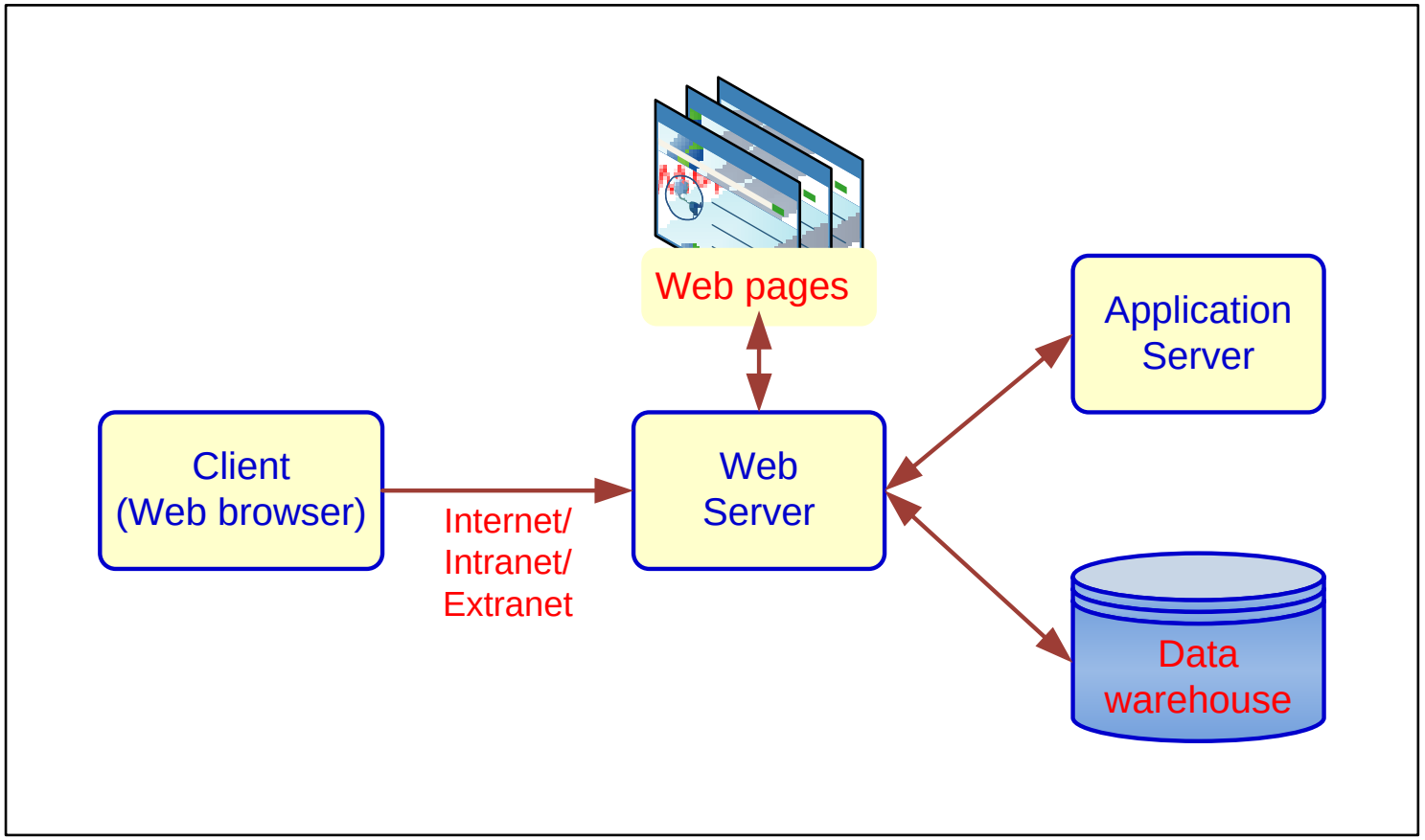
1-tier Architecture ?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

DW Architecture Considerations

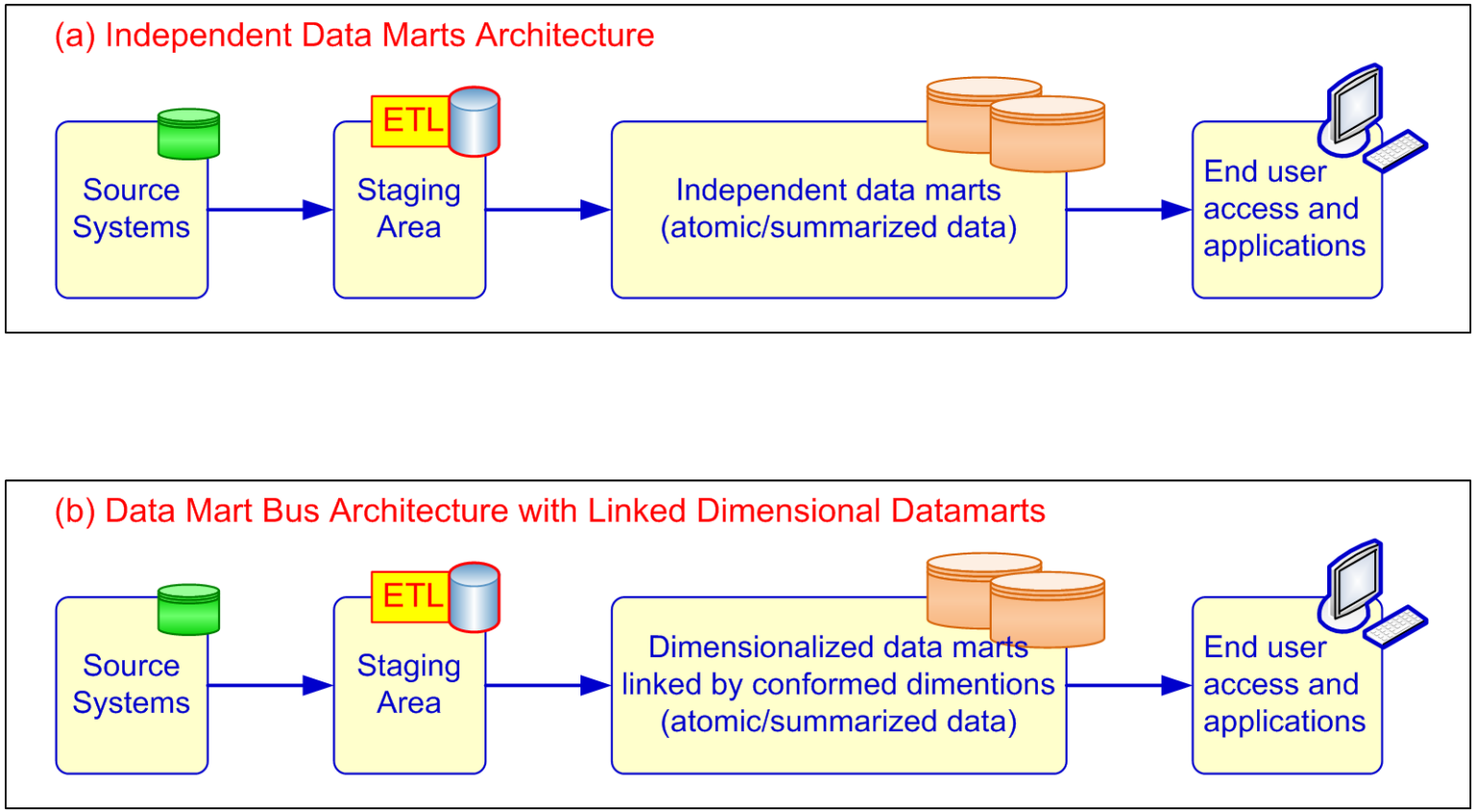
- Issues to consider when deciding which architecture to use:
  - Which database management system (DBMS) should be used?
  - Will parallel processing and/or partitioning be used?
  - Will data migration tools be used to load the data warehouse?
  - What tools will be used to support data retrieval and analysis?

A Web-based DW Architecture



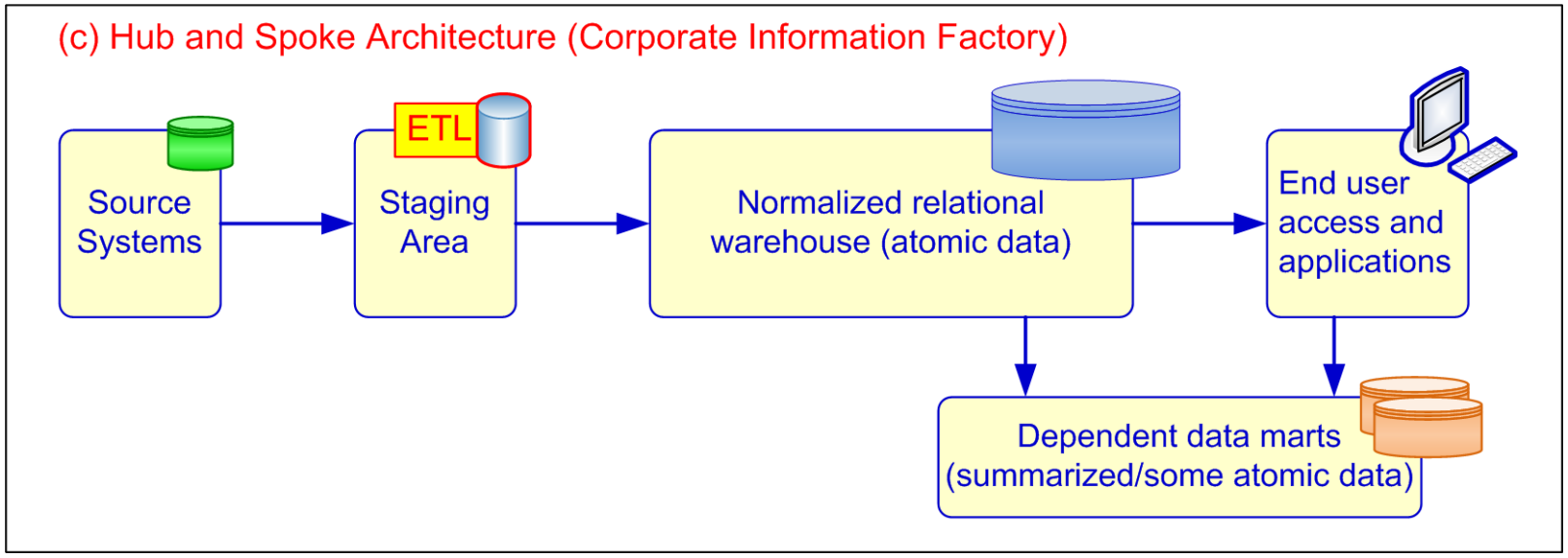
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

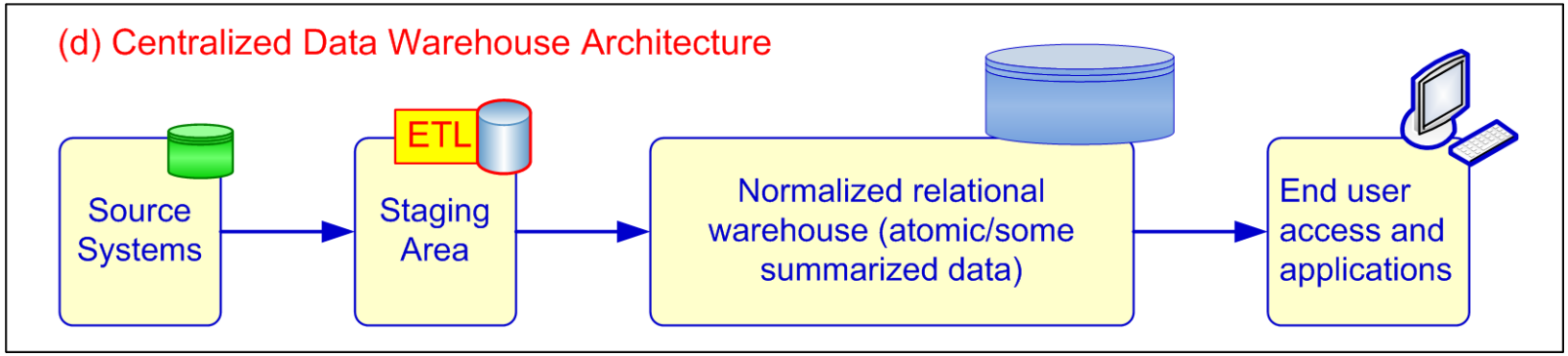
Alternative DW Architectures



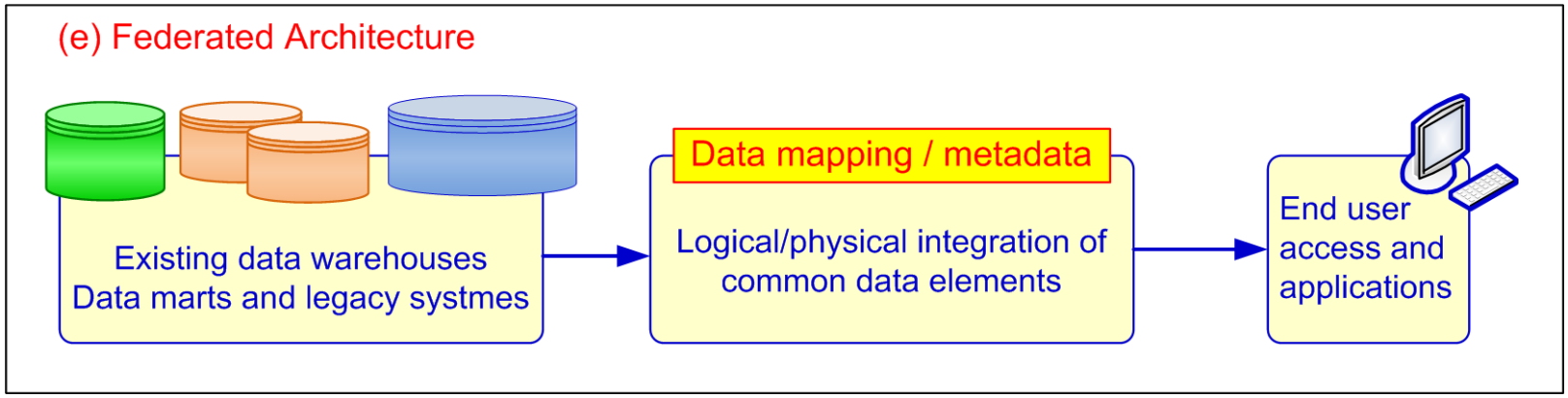
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Alternative DW Architectures





Alternative DW Architectures



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Which Architecture is the Best?

- Bill Inmon versus Ralph Kimball
- Enterprise DW versus Data Marts approach

	Independent Data Marts	Bus Architecture	Hub-and-Spoke Architecture	Centralized Architecture (No Dependent Data Marts)	Federated Architecture
Information Quality	4.42	5.16	5.35	5.23	4.73
System Quality	4.59	5.60	5.56	5.41	4.69
Individual Impacts	5.08	5.80	5.62	5.64	5.15
Organizational Impacts	4.66	5.34	5.24	5.30	4.77

Empirical study by Ariyachandra and Watson (2006)

Data Warehousing Architectures

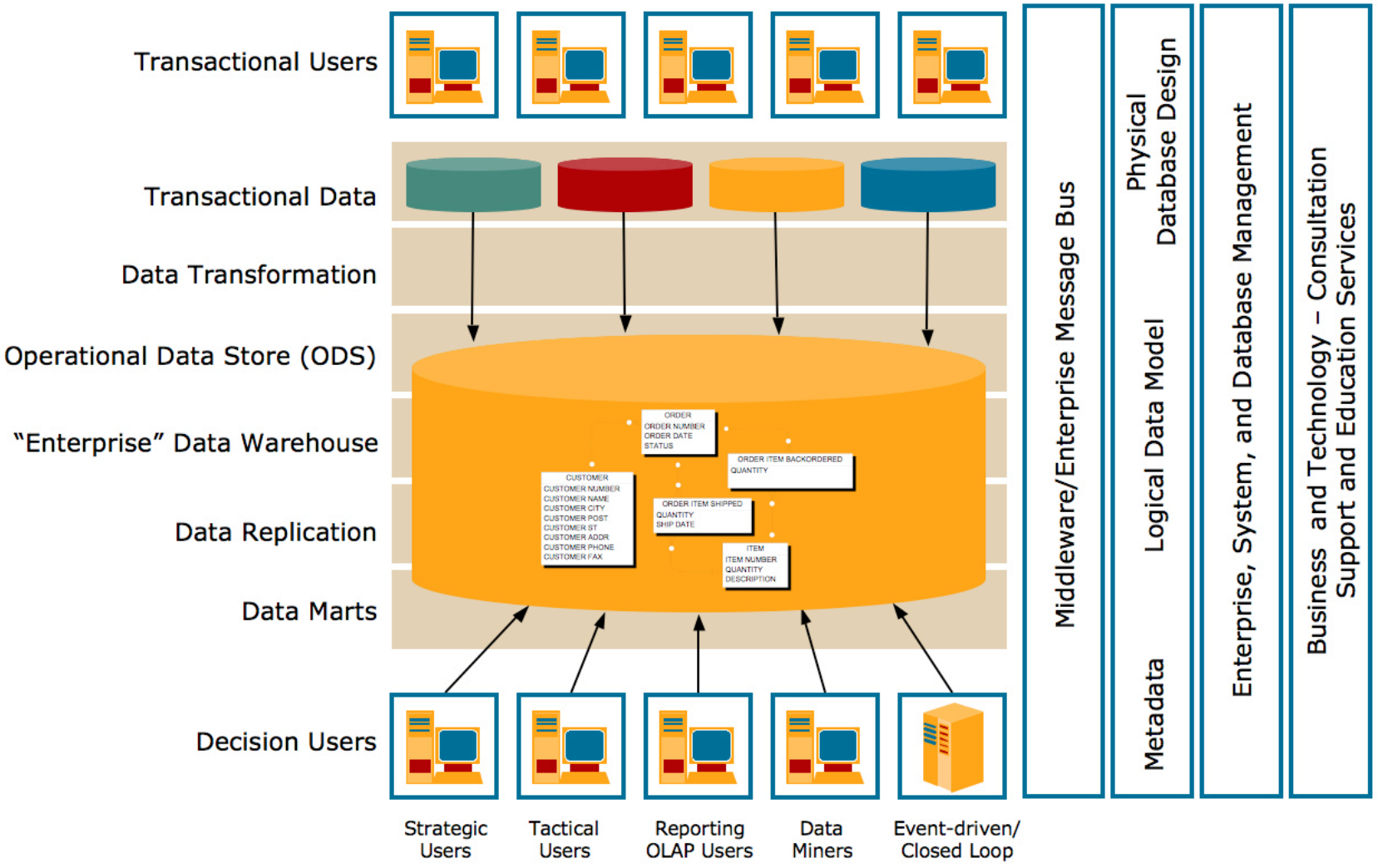
Ten factors **that** potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Enterprise Data Warehouse

(by Teradata Corporation)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Integration and the Extraction, Transformation, and Load (ETL) Process

- **Data integration**

  Integration that comprises three major processes: data access, data federation, and change capture.
- **Enterprise application integration (EAI)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse
- **Enterprise information integration (EII)**

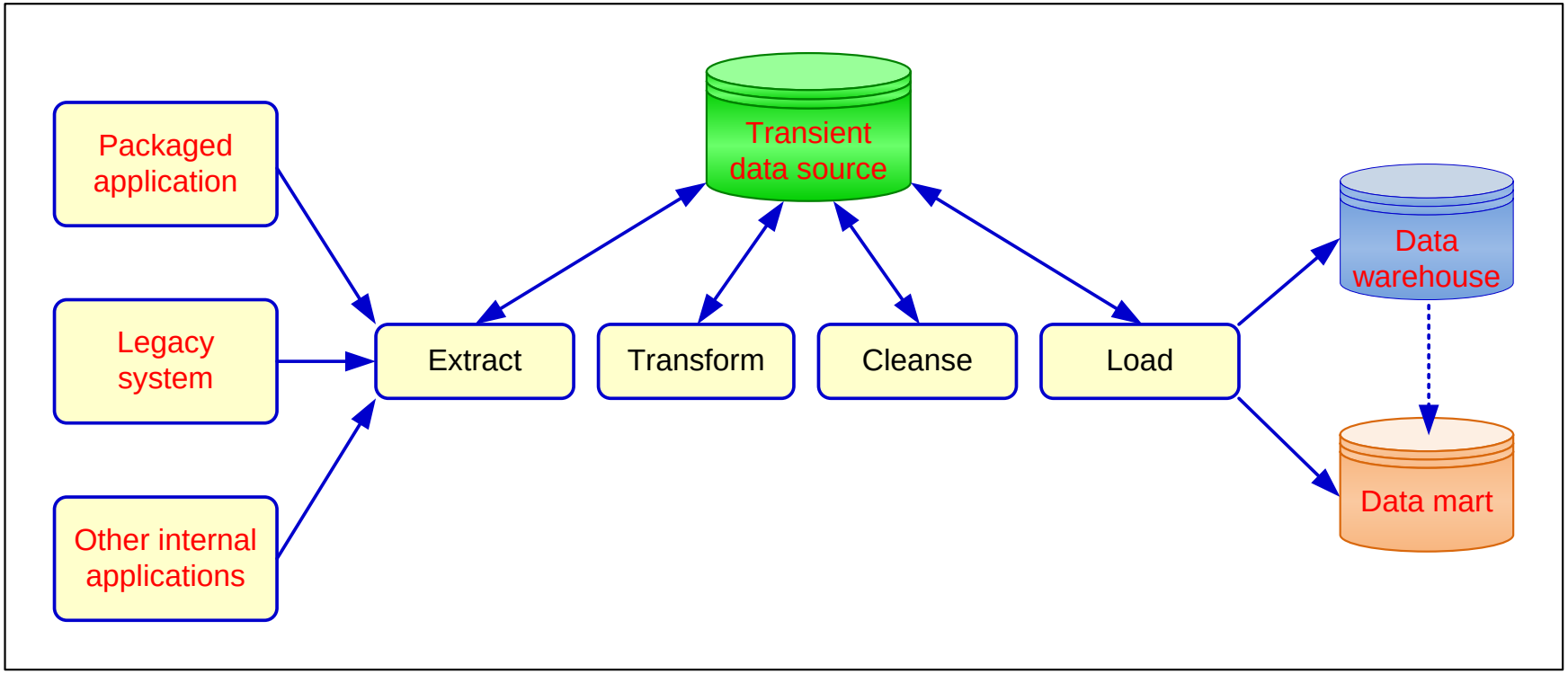
  An evolving tool space that promises real-time data integration from a variety of sources
- **Service-oriented architecture (SOA)**

  A new way of integrating information systems

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, transformation, and load (ETL) process



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

ETL

- Issues affecting the purchase of and ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

Benefits of DW

- Direct benefits of a data warehouse
  - Allows end users to perform extensive analysis
  - Allows a consolidated view of corporate data
  - Better and more timely information
  - Enhanced system performance
  - Simplification of data access
- Indirect benefits of data warehouse
  - Enhance business knowledge
  - Present competitive advantage
  - Enhance customer service and satisfaction
  - Facilitate decision making
  - Help in reforming business processes

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Warehouse Development

- Data warehouse development approaches
  - Inmon Model: EDW approach (top-down)
  - Kimball Model: Data mart approach (bottom-up)
  - Which model is best?
    - There is no one-size-fits-all strategy to DW
  - One alternative is the hosted warehouse
- Data warehouse structure:
  - The Star Schema vs. Relational
- Real-time data warehousing?

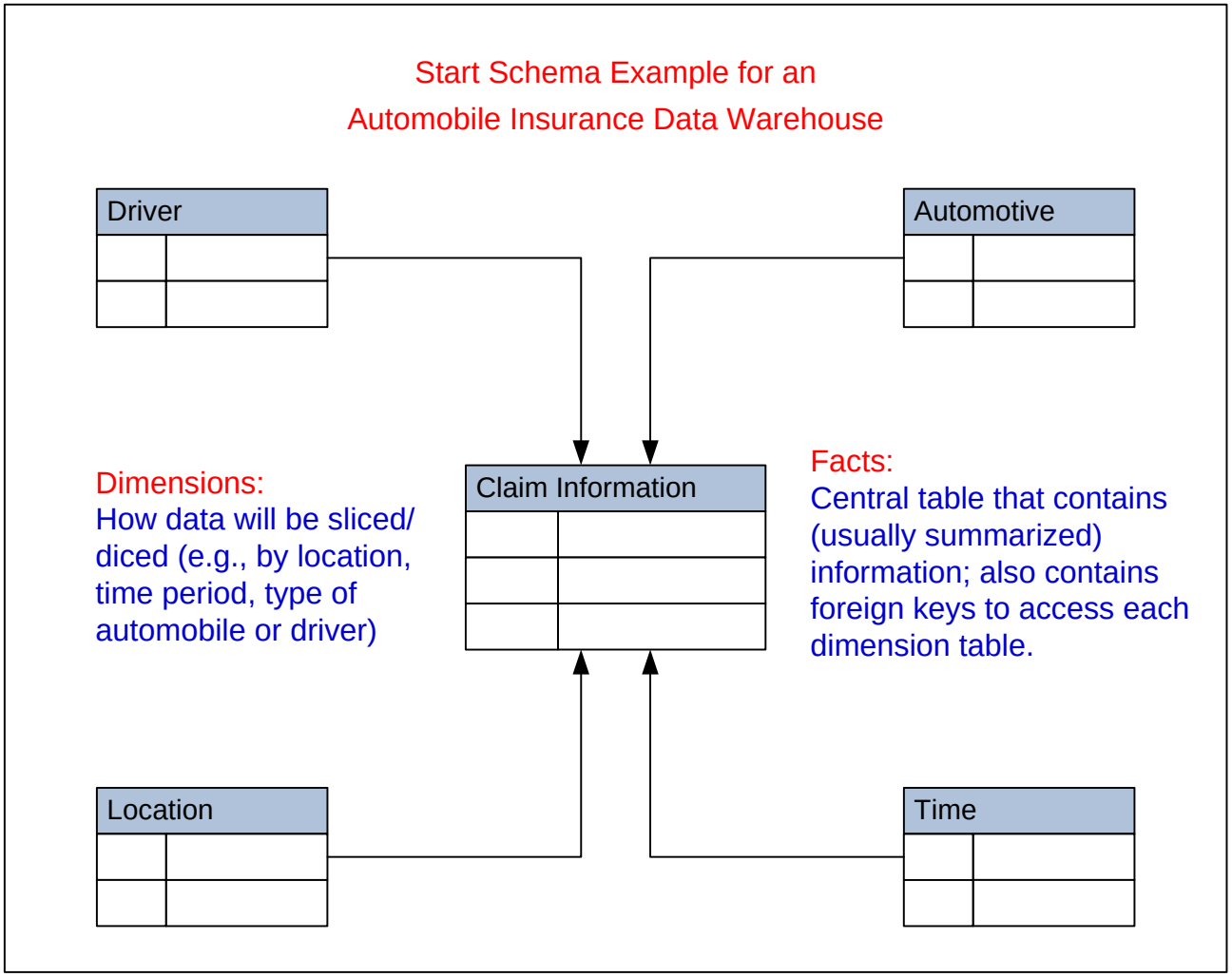
DW Development Approaches

(Kimball Approach) (Inmon Approach)

Effort	Data Mart Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	$10,000 to $100,000+	$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High
Update frequency	Hourly, daily, weekly	Weekly, monthly
Technology		
Hardware	Workstations and departmental servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

DW Structure: Star Schema (a.k.a. Dimensional Modeling)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Dimensional Modeling

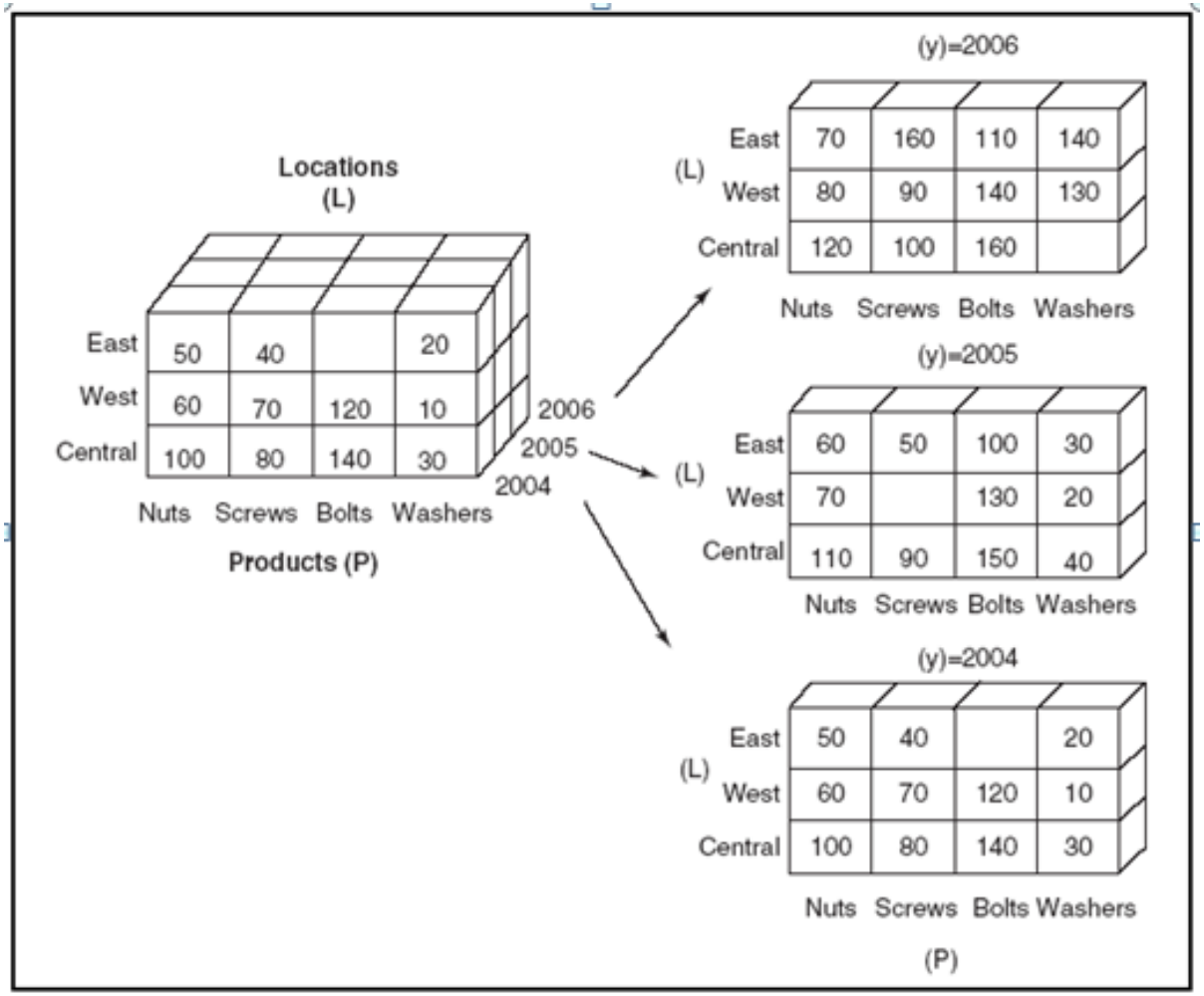
Data cube

A two-dimensional, three-dimensional, or higher-dimensional object in which each dimension of the data represents a measure of interest

-Grain

-Drill-down

-Slicing



Best Practices for Implementing DW

- The project must fit with corporate strategy
- There must be complete buy-in to the project
- It is important to manage user expectations
- The data warehouse must be built incrementally
- Adaptability must be built in from the start
- The project must be managed by both IT and business professionals (a business–supplier relationship must be developed)
- Only load data that have been cleansed/high quality
- Do not overlook training requirements
- Be politically aware.

Risks in Implementing DW

- No mission or objective
- Quality of source data unknown
- Skills not in place
- Inadequate budget
- Lack of supporting software
- Source data not understood
- Weak sponsor
- Users not computer literate
- Political problems or turf wars
- Unrealistic user expectations

(Continued …)

Risks in Implementing DW – Cont.

- Architectural and design risks
- Scope creep and changing requirements
- Vendors out of control
- Multiple platforms
- Key people leaving the project
- Loss of the sponsor
- Too much new technology
- Having to fix an operational system
- Geographically distributed environment
- Team geography and language culture

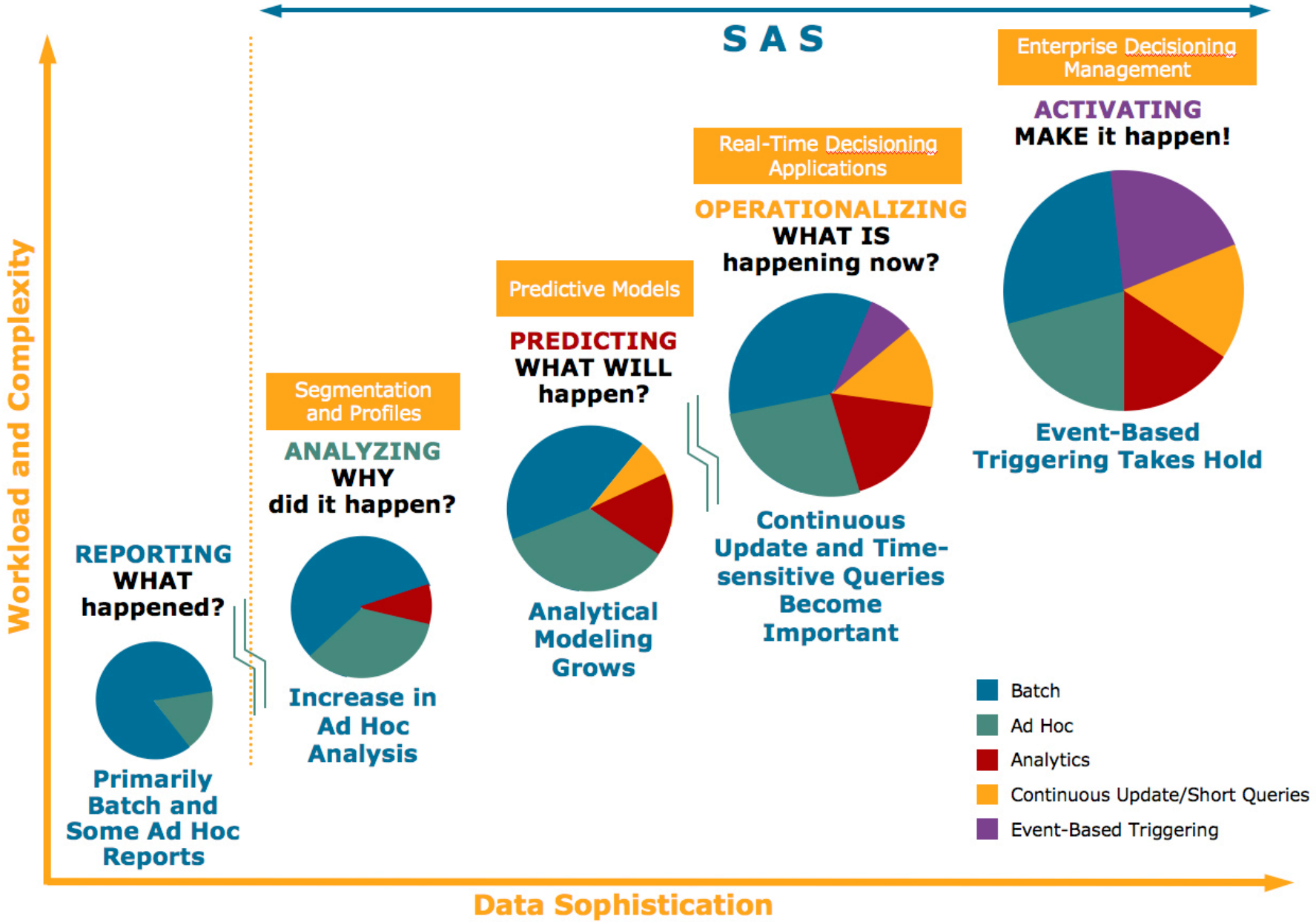
Things to Avoid for Successful Implementation of DW

- Starting with the wrong sponsorship chain
- Setting expectations that you cannot meet
- Engaging in politically naive behavior
- Loading the warehouse with information just because it is available
- Believing that data warehousing database design is the same as transactional DB design
- Choosing a data warehouse manager who is technology oriented rather than user oriented

Real-time DW (a.k.a. Active Data Warehousing)

- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  - Push vs. Pull (of data)
- Concerns about real-time BI
  - Not all data should be updated continuously
  - Mismatch of reports generated minutes apart
  - May be cost prohibitive
  - May also be infeasible

Evolution of DSS & DW



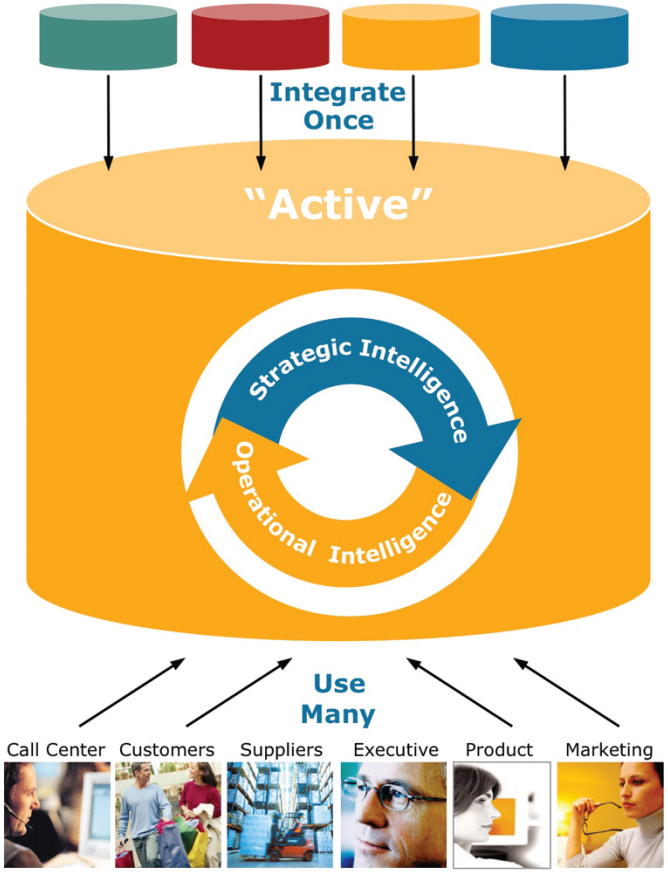
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Active Data Warehousing
(by Teradata Corporation)

Active Access
Front-Line operational decisions or services supported by near-real-time (NRT) access; Service Level Agreements of 5 seconds or less

Active Load
Intra-day data acquisition; Mini-batch to NRT trickle data feeds measured in minutes or seconds

Active Events
Proactive monitoring of business activity initiating intelligent actions based on rules and context; to systems or users supporting an operational business process



Active Workload Management
Dynamically manage system resources for optimum performance and resource utilization supporting a mixed-workload environment

Active Enterprise Integration
Integration into the Enterprise Architecture for delivery of intelligent decisioning services

Active Availability
Business Continuity to support the requirements of the business (up to 7X24)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Comparing Traditional and Active DW

Traditional Data Warehouse Environment	Active Data Warehouse Environment
Strategic decisions only	Strategic and tactical decisions
Results sometimes hard to measure	Results measured with operations
Daily, weekly, monthly data currency acceptable; summaries often appropriate	Only comprehensive detailed data available within minutes is acceptable
Moderate user concurrency	High number (1,000 or more) of users accessing and querying the system simultaneously
Highly restrictive reporting used to confirm or check existing processes and patterns; often uses predeveloped summary tables or data marts	Flexible ad hoc reporting, as well as machine-assisted modeling (e.g., data mining) to discover new hypotheses and relationships
Power users, knowledge workers, internal users	Operational staffs, call centers, external users

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Warehouse Administration

- Due to its huge size and its intrinsic nature, a DW requires especially strong monitoring in order to sustain its efficiency, productivity and security.
- The successful administration and management of a data warehouse entails skills and proficiency that go past what is required of a traditional database administrator.
  - Requires expertise in high-performance software, hardware, and networking technologies

DW Scalability and Security

- Scalability
  - The main issues pertaining to scalability:
    - The amount of data in the warehouse
    - How quickly the warehouse is expected to grow
    - The number of concurrent users
    - The complexity of user queries
  - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse
- Security
  - Emphasis on security and privacy

Summary

- Definitions and concepts of data warehouses
- Types of data warehousing architectures
- Processes used in developing and managing data warehouses
- Data warehousing operations
- Role of data warehouses in decision support
- Data integration and the extraction, transformation, and load (ETL) processes
- Data warehouse administration and security issues