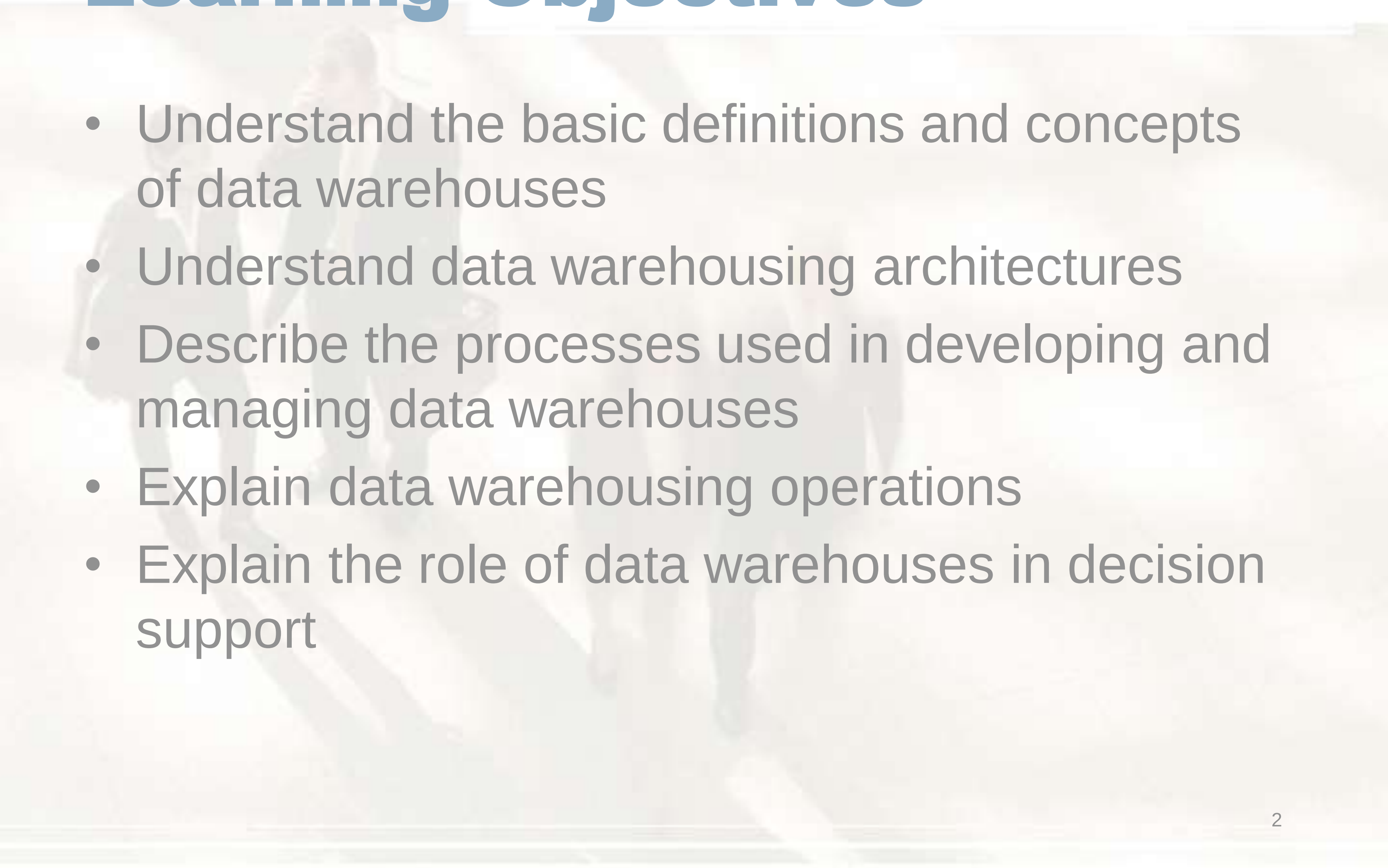
Chapter 5

DATA WAREHOUSING

8th Edition

Learning Objectives



- Understand the basic definitions and concepts of data warehouses
- Understand data warehousing architectures
- Describe the processes used in developing and managing data warehouses
- Explain data warehousing operations
- Explain the role of data warehouses in decision support

Learning Objectives

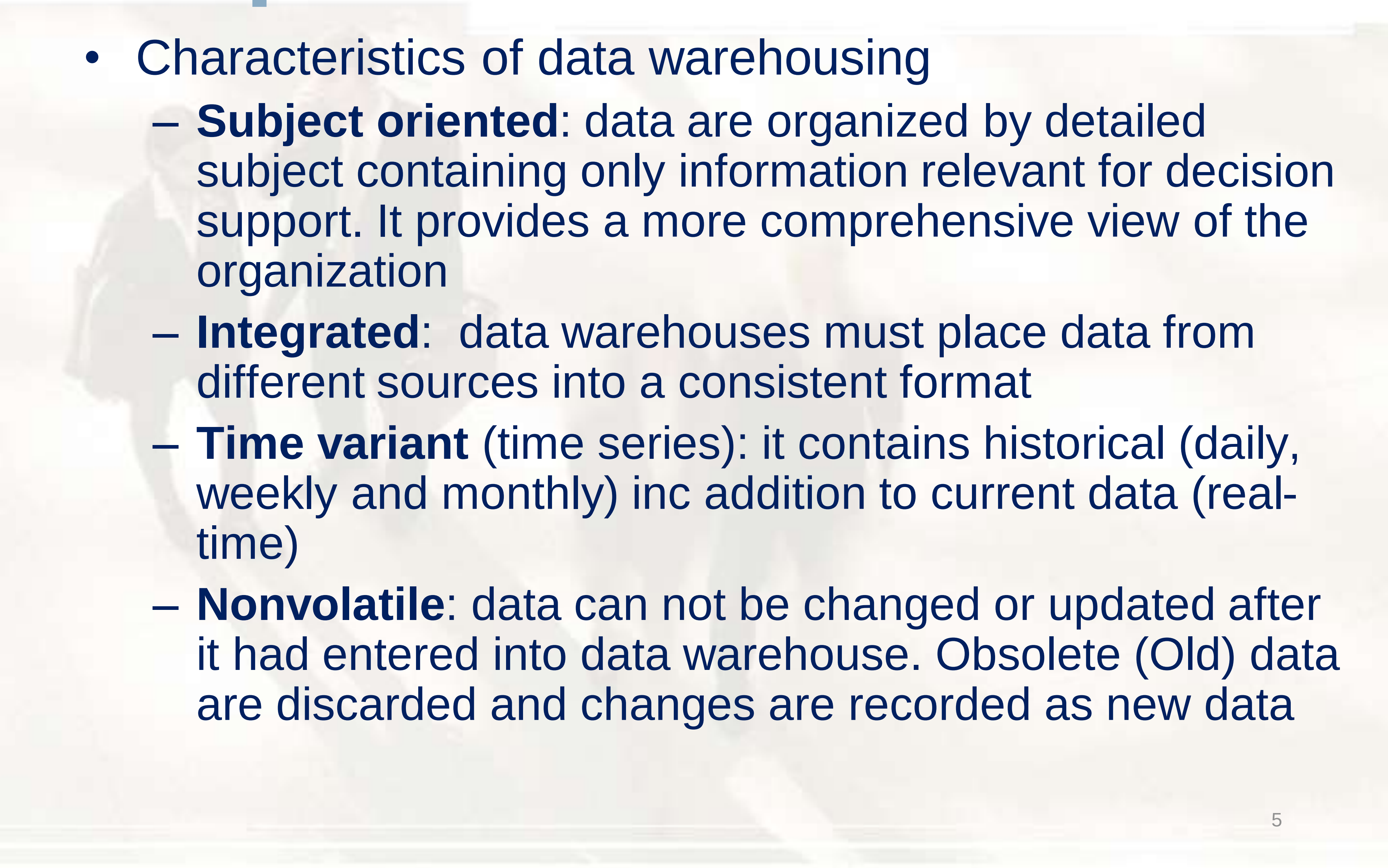
- Explain data integration and the extraction, transformation, and load (ETL) processes
- Describe real-time (active) data warehousing
- Understand data warehouse administration and security issues

(5.2) Data Warehousing Definitions and Concepts

- Using real-time data warehouse in conjunction with DSS and BI tools is an important way to conduct business processes.
- With real-data warehouse an organization can view the current state of its business and identify problems, which is the first step toward solving problems.
- **Data warehouse**

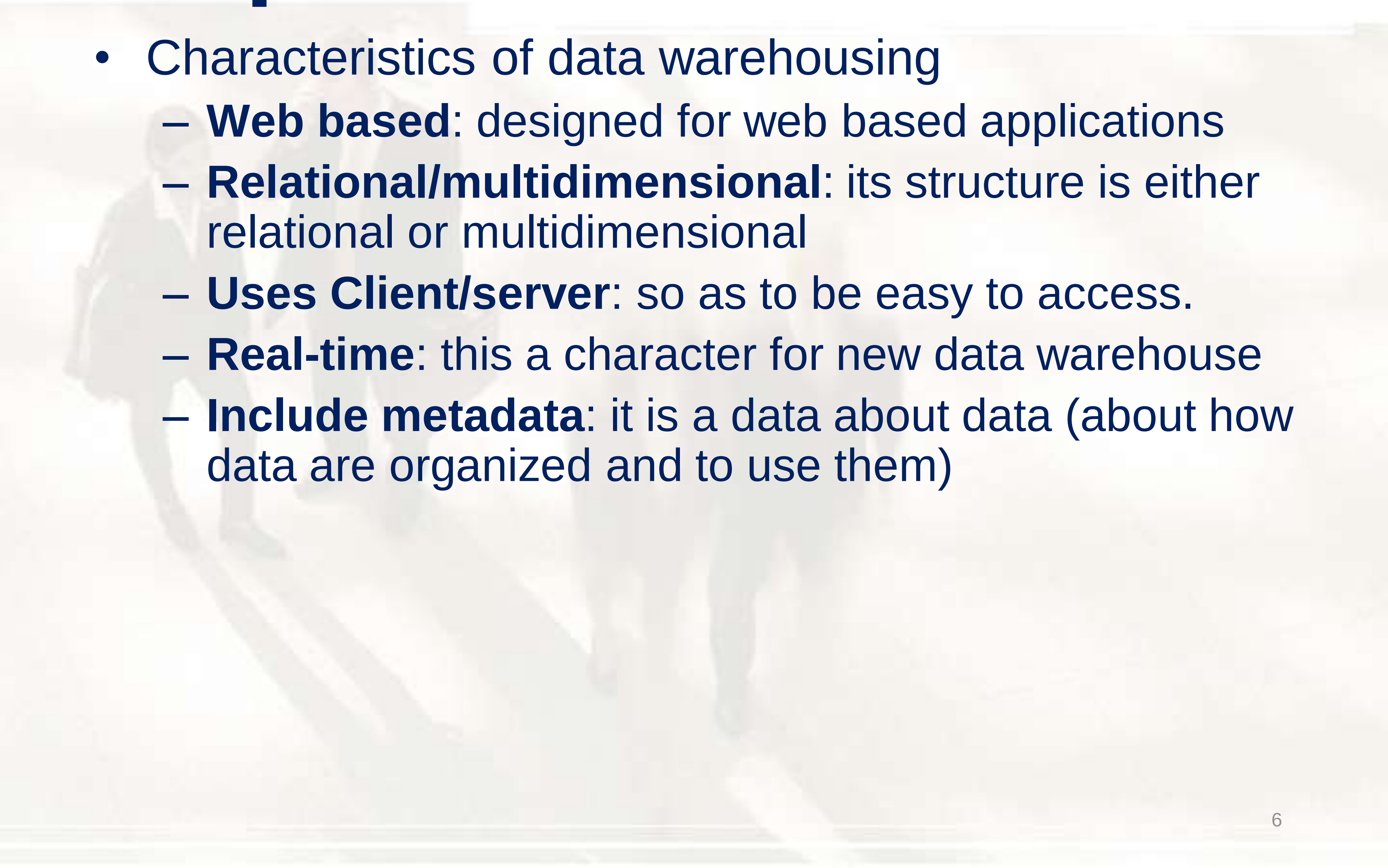
  A physical repository where relational data (current and historical) are specially organized to provide enterprise-wide, cleansed data in a standardized format

(5.2) Data Warehousing Definitions and Concepts



- Characteristics of data warehousing
  - **Subject oriented**: data are organized by detailed subject containing only information relevant for decision support. It provides a more comprehensive view of the organization
  - **Integrated**: data warehouses must place data from different sources into a consistent format
  - **Time variant** (time series): it contains historical (daily, weekly and monthly) inc addition to current data (real-time)
  - **Nonvolatile**: data can not be changed or updated after it had entered into data warehouse. Obsolete (Old) data are discarded and changes are recorded as new data

(5.2) Data Warehousing Definitions and Concepts



- Characteristics of data warehousing
  - **Web based**: designed for web based applications
  - **Relational/multidimensional**: its structure is either relational or multidimensional
  - **Uses Client/server**: so as to be easy to access.
  - **Real-time**: this a character for new data warehouse
  - **Include metadata**: it is a data about data (about how data are organized and to use them)

(5.2) Data Warehousing Definitions and Concepts

- **Data mart**

  A departmental data warehouse that stores only relevant data (usually smaller that warehouse)

- **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 (SBU) or a department and its source is not the EDW (Enterprise Data Warehouse)

(5.2) Data Warehousing Definitions and Concepts

- **Operational data stores (ODS)**

  A type of database often used as an interim (temporal) area for a data warehouse, especially for customer information files

- **Oper marts**

  An operational data mart. An oper mart is a small-scale data mart typically used by a single department or functional area in an organization when they need to analyze operational data

(5.2) Data Warehousing Definitions and Concepts

- **Enterprise data warehouse (EDW)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse that is used across the enterprise for decision support

- **Metadata**

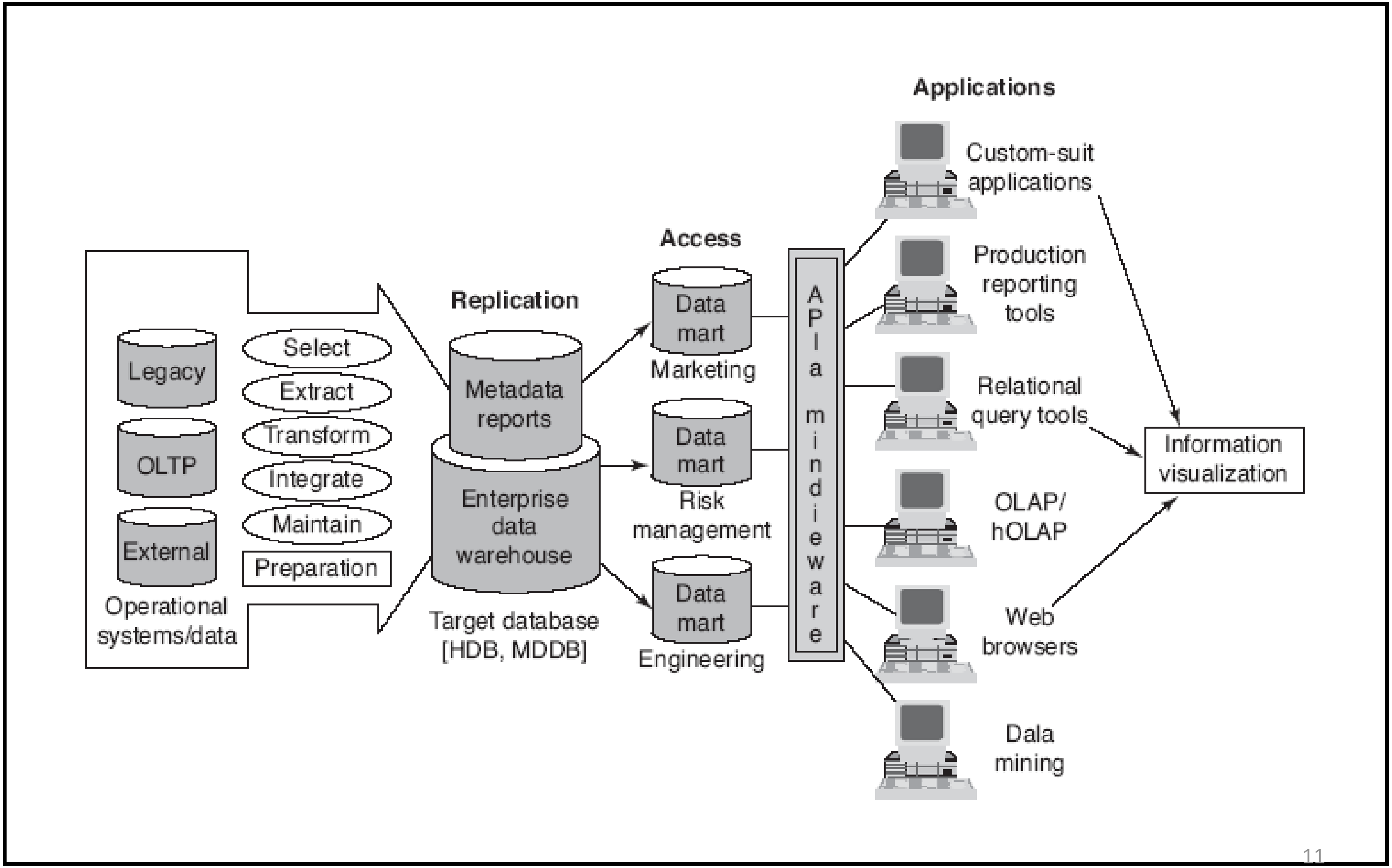
  Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use

(5.3) Data Warehousing Process Overview

- Organizations continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems
- The number of users needing to access the information continues to increase as a result of improved reliability and availability of network access, especially the Internet

(5.3) Data Warehousing Process Overview

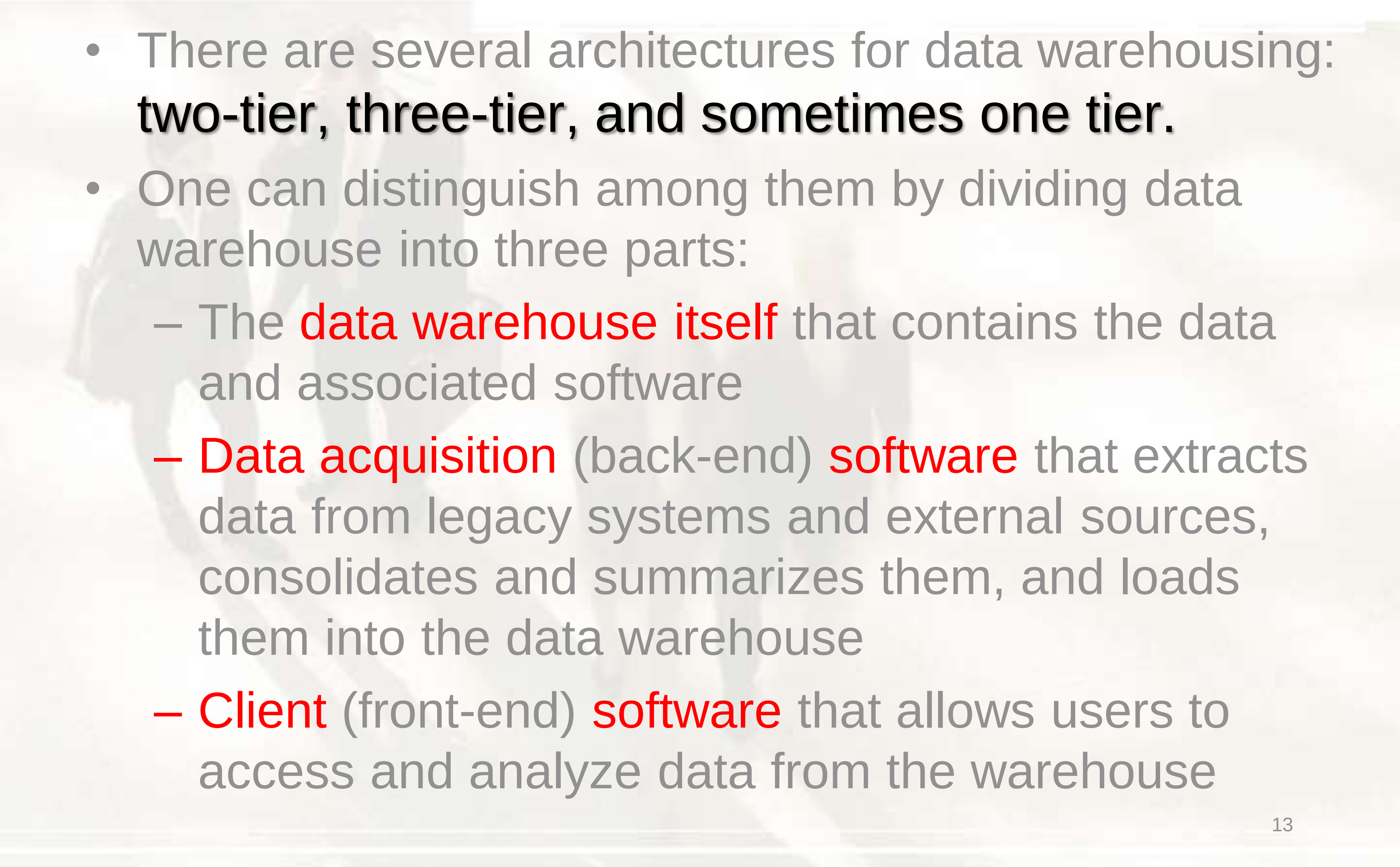
FIGURE 5.1 Data Warehouse Framework and Views



(5.3) Data Warehousing Process Overview

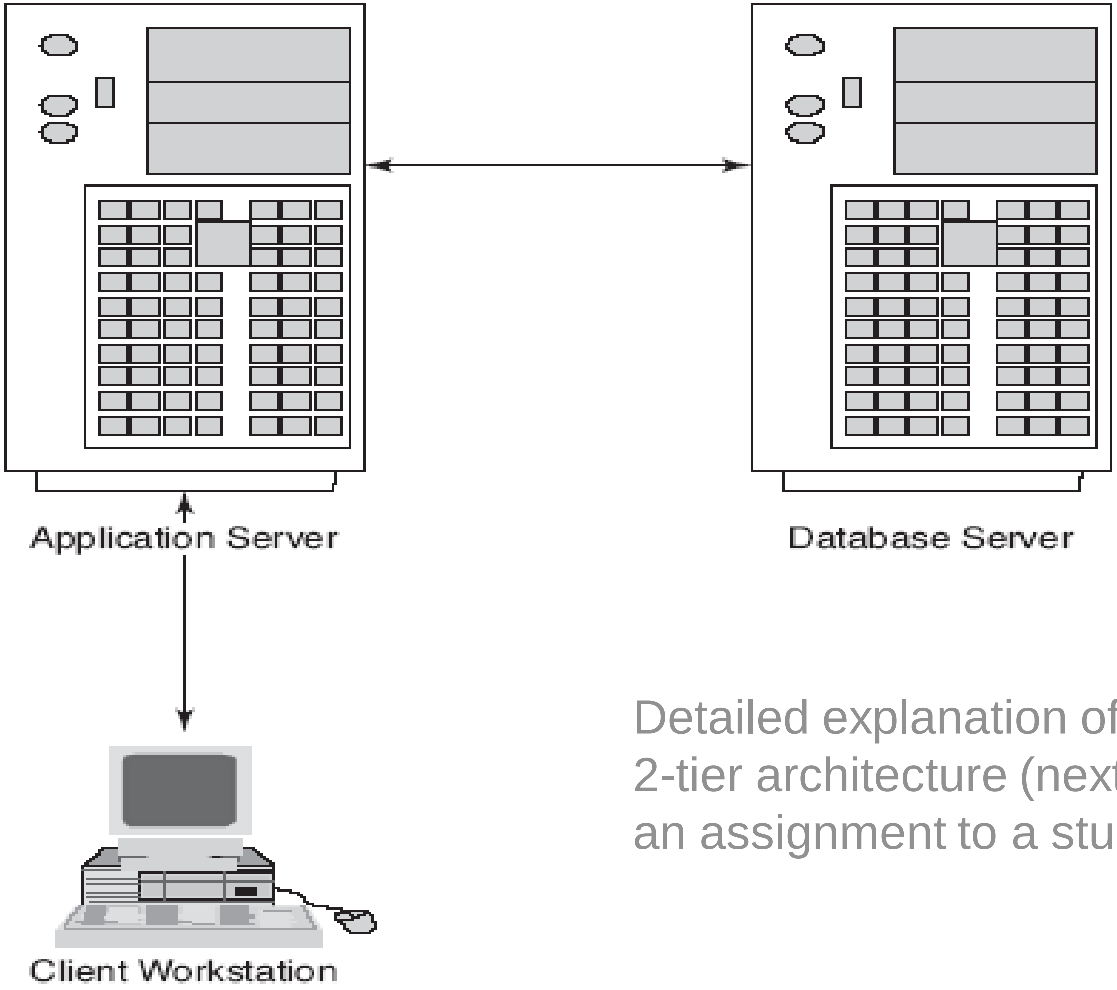
- The major components of a data warehousing process
  - **Data sources**: internal, external (data provider), OLAP, ERP, Web data
  - **Data extraction**: using custom-written or commercial software called (ETL)
  - **Data loading**: loaded into a staging area to be transformed and cleansed, then loaded into the warehouse
  - **Comprehensive database**: It is the EDW to support all decision analysis
  - **Metadata**: to ease indexing and search
  - **Middleware tools**: to enable access to DW. It includes data mining tools, OLAP, reporting tools, and data visualization tools.

(5.4) Data Warehousing Architectures



- There are several architectures for data warehousing: **two-tier, three-tier, and sometimes one tier.**
- One can distinguish among them by dividing data warehouse into three parts:
  - The data warehouse itself that contains the data and associated software
  - Data acquisition (back-end) software that extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
  - Client (front-end) software that allows users to access and analyze data from the warehouse

(5.4) Data Warehousing Architectures



Detailed explanation of 3-tier and 2-tier architecture (next slide) is an assignment to a student

FIGURE 5.2 Architecture of a Three-Tier Data Warehouse

(5.4) Data Warehousing Architectures

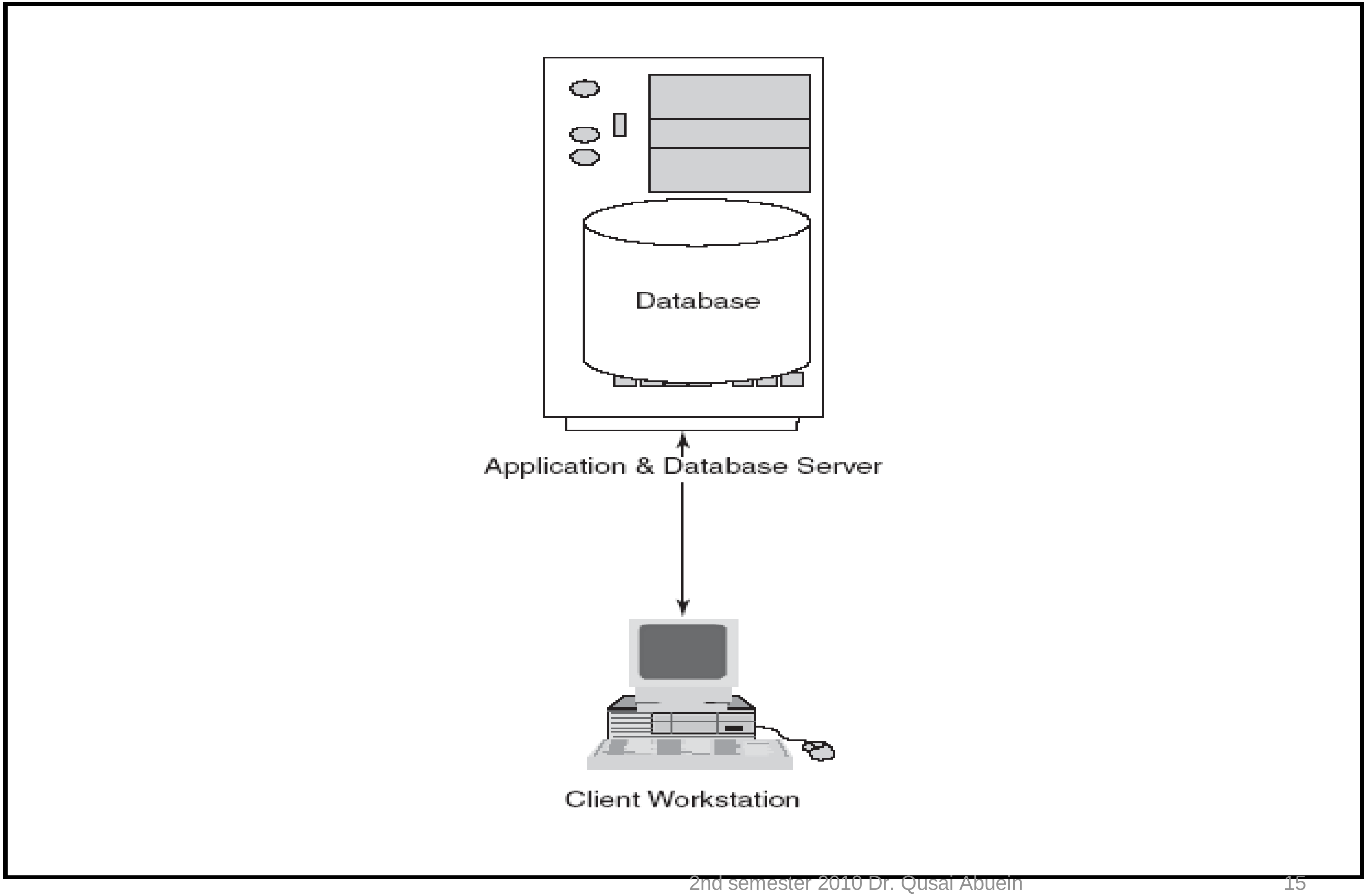
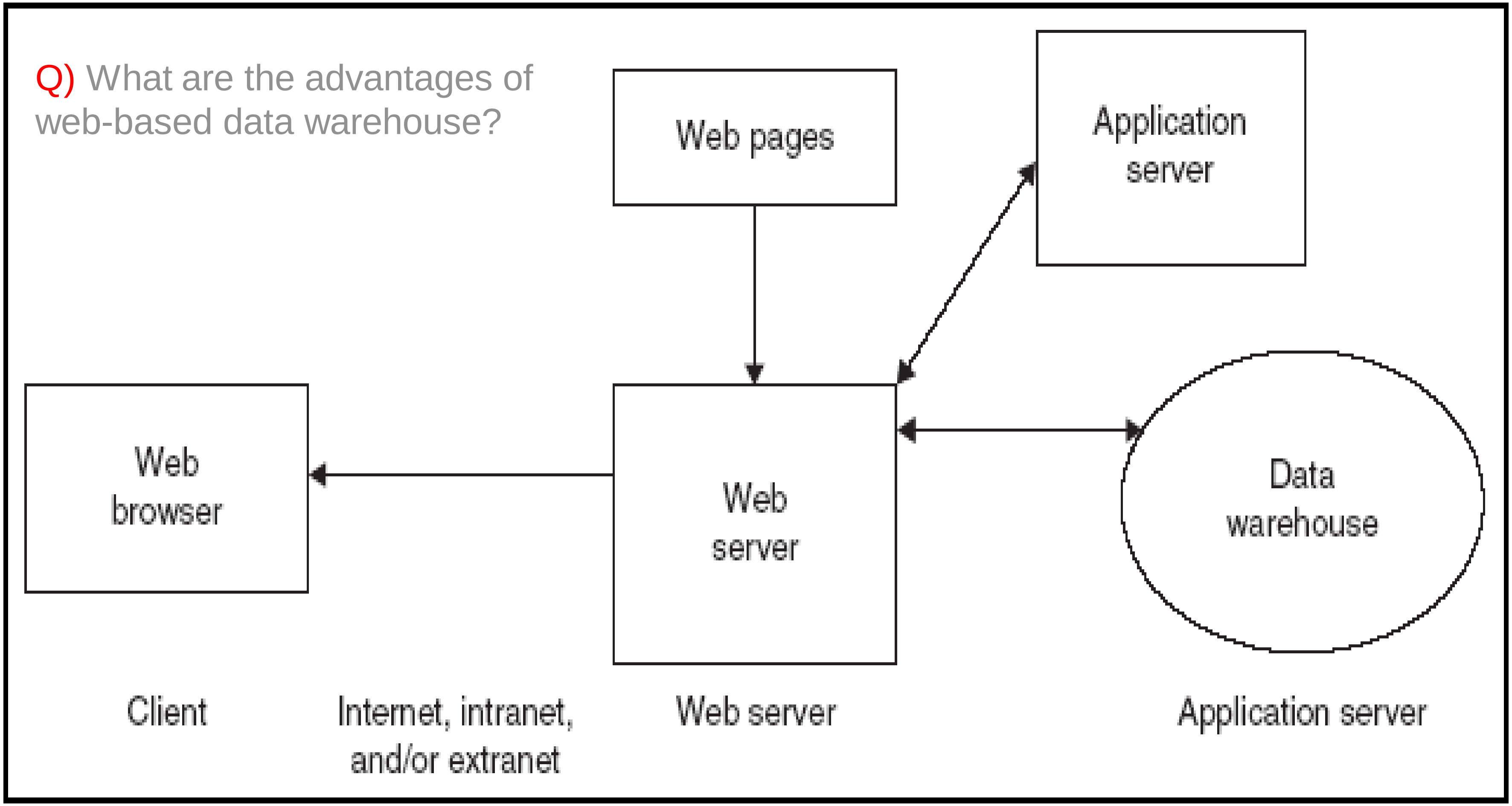


FIGURE 5.3 Architecture of a Two-Tier Data Warehouse

(5.4) Data Warehousing Architectures

FIGURE 5.4 Architecture of Web-Based Data Warehousing

Q) What are the advantages of web-based data warehouse?

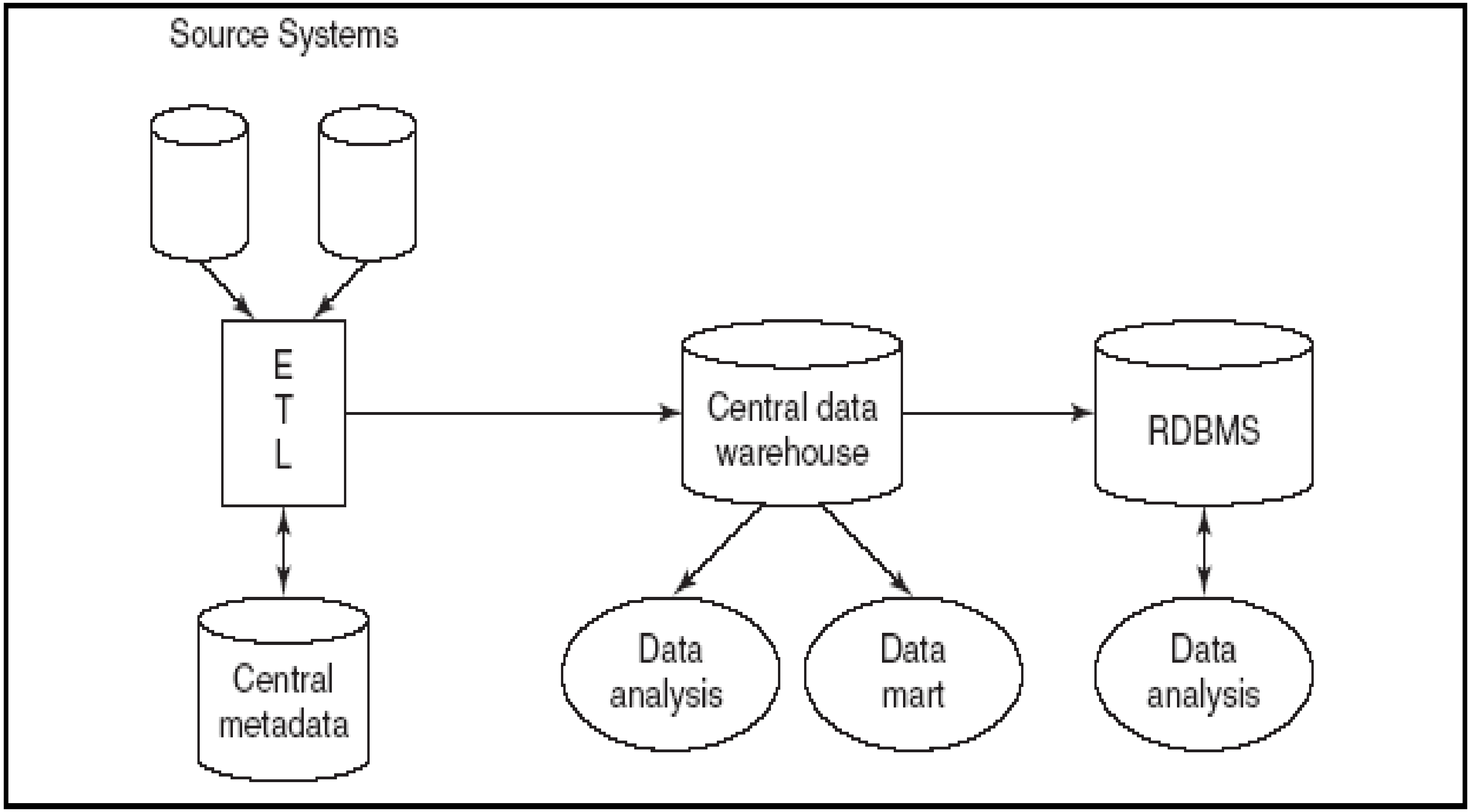


(5.4) Data Warehousing Architectures

- 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?*

(5.4) Data Warehousing Architectures

FIGURE 5.5 Alternative Data Warehouse Architectures



5.5a Enterprise Data Warehousing Architecture

(5.4) Data Warehousing Architectures

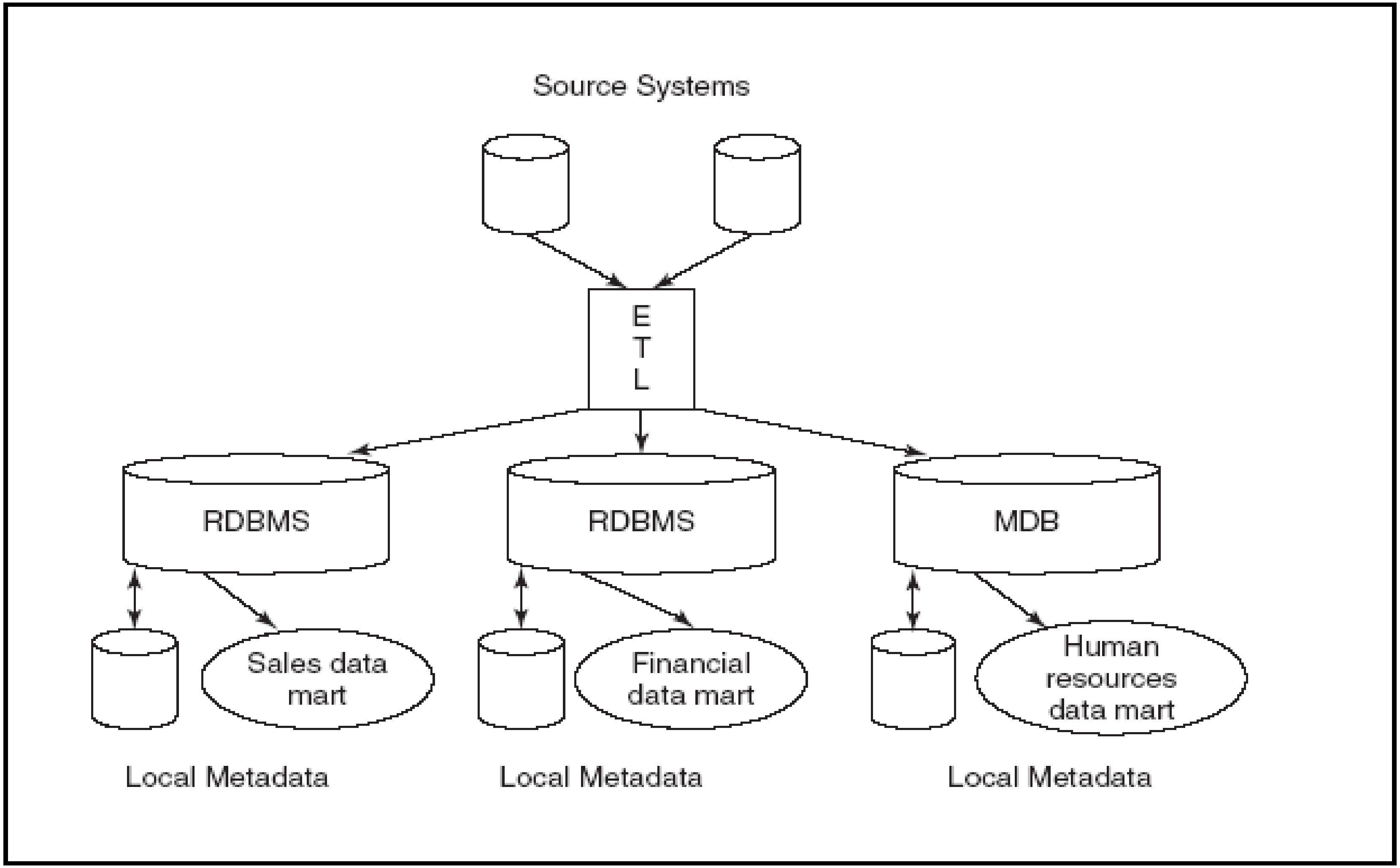


FIGURE 5.5b Data Mart Architecture

(5.4) Data Warehousing Architectures

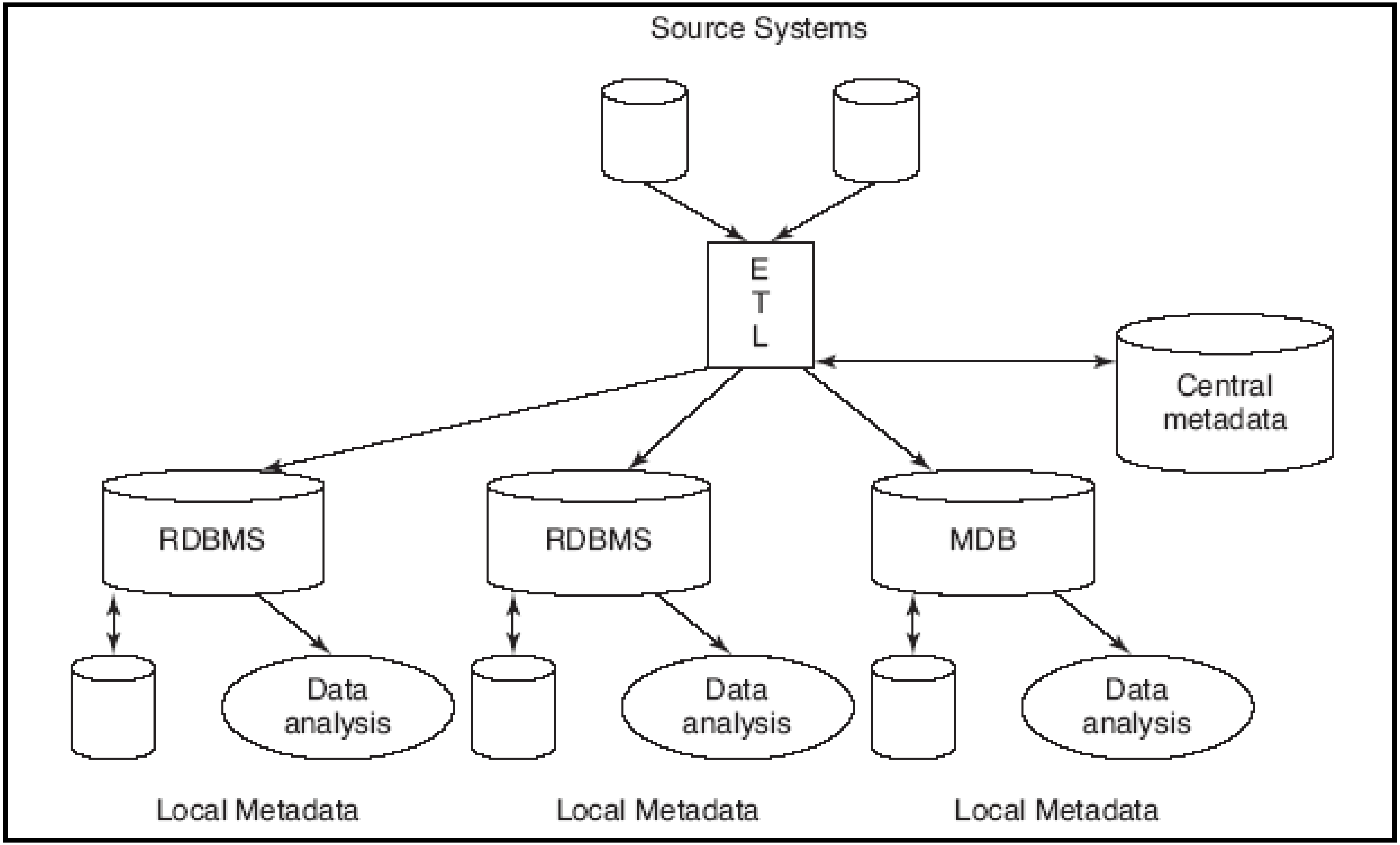


FIGURE 5.5c Hub-and-Spoke Data Mart Architecture

(5.4) Data Warehousing Architectures

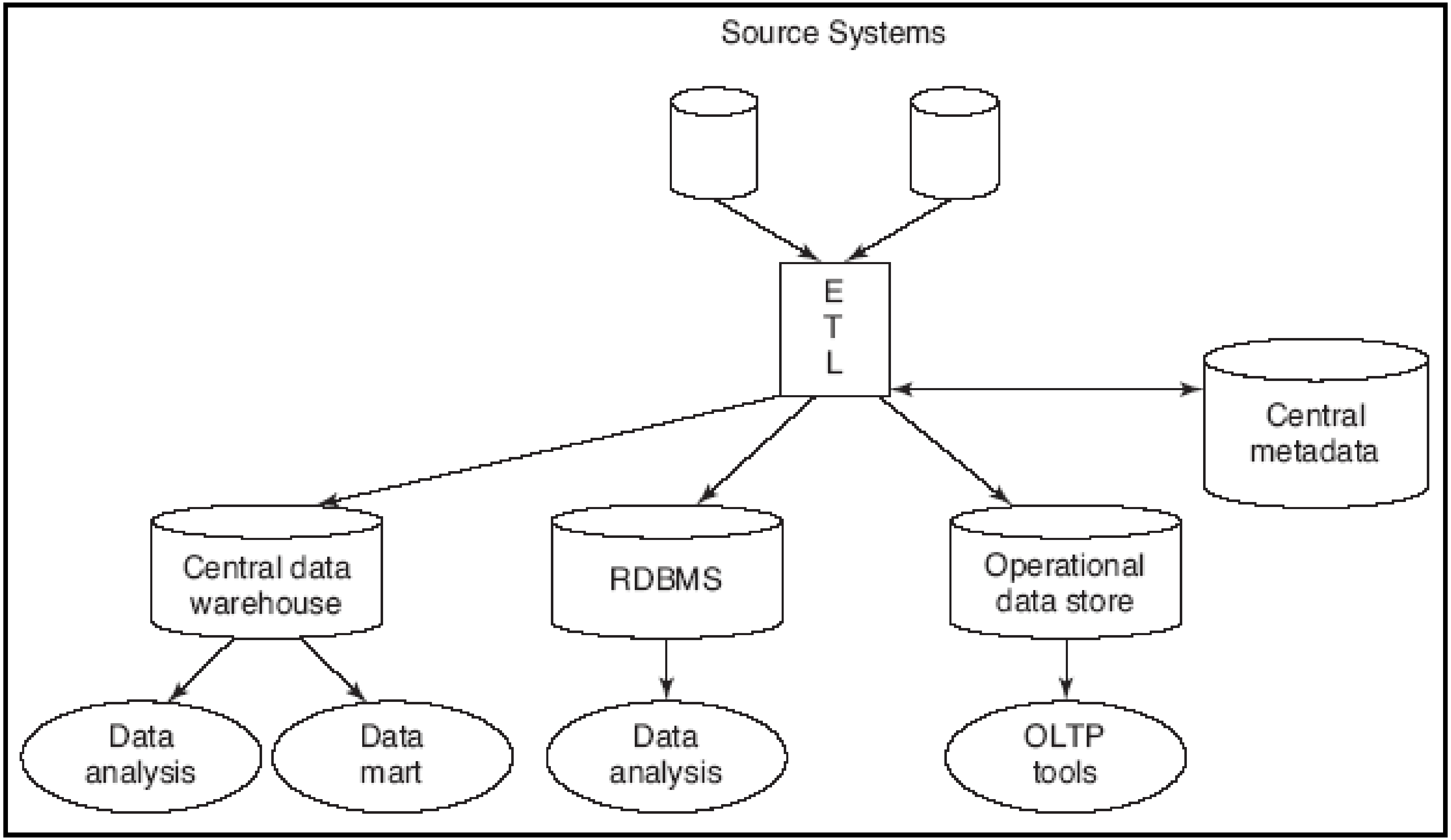


FIGURE 5.5d Enterprise Warehouse and Operational Data Store

(5.4) Data Warehousing Architectures

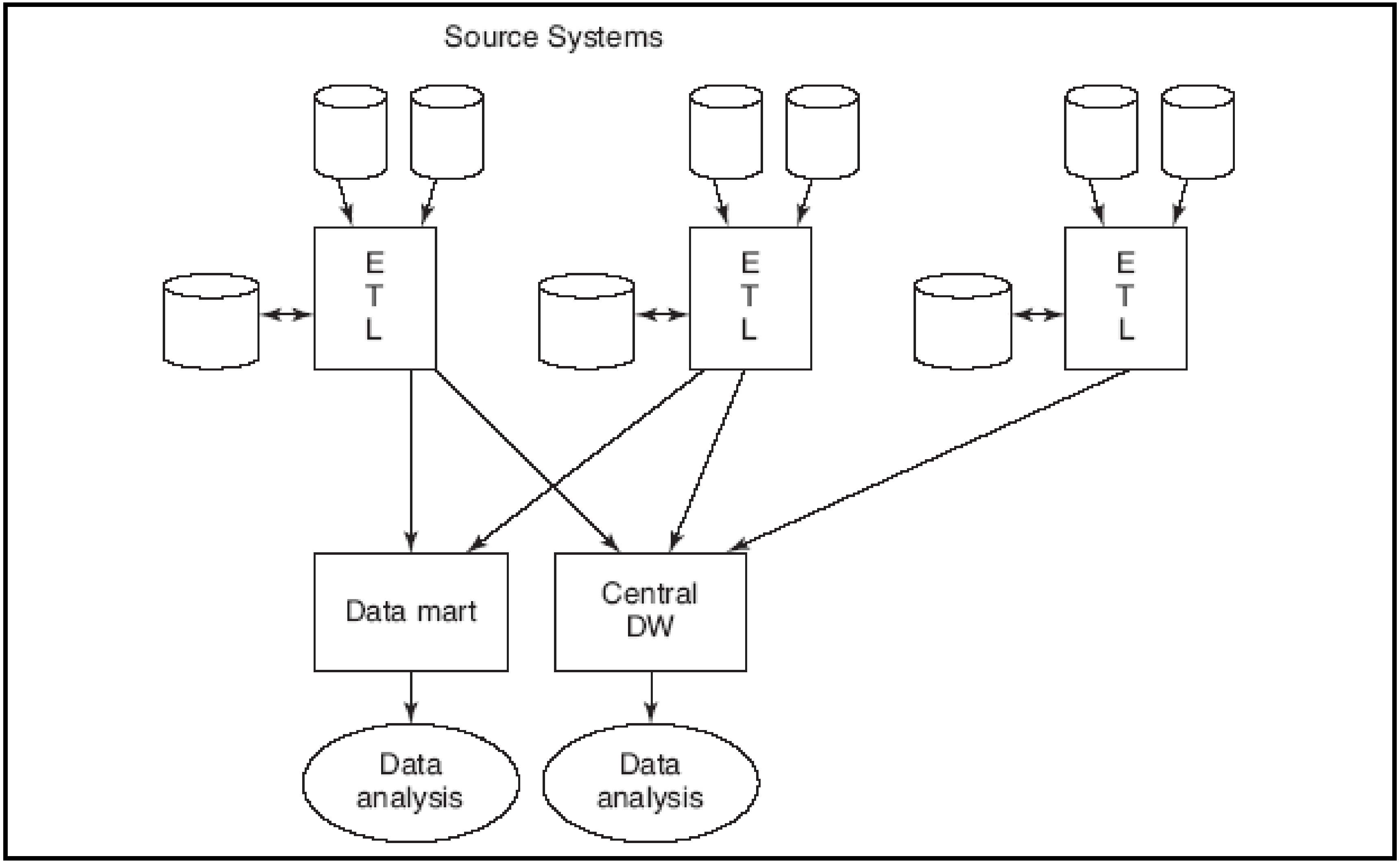
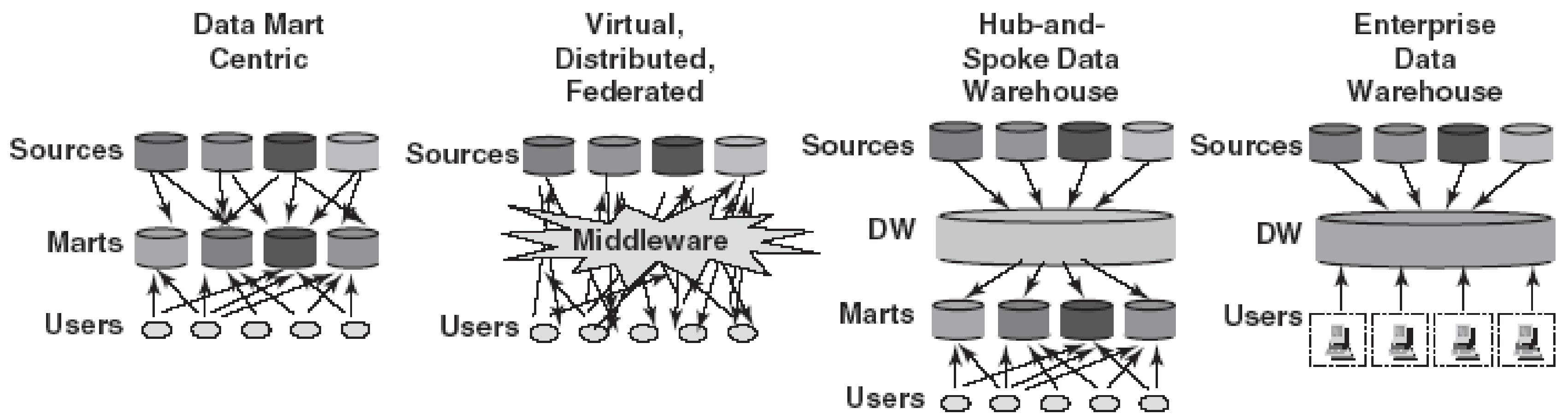


FIGURE 5.5e Distributed Data Warehouse Architecture

(5.4) Data Warehousing Architectures



	Independent Data Marts	Leave Data Where it Lies	Dependent Data Marts	Centralized Integrated Data With Direct Access
Pros	• Easy to Build Organizationally • Easy to Build Technically	• No need for ETL • No need for separate platform	• Allows easier customization of user interfaces & reports	Business Enterprise view Design consistency & data quality Data reusability
Cons	• Business Enterprise view unavailable • Redundant data costs • High ETL costs • High App costs • High DBA and operational costs	• Only viable for low volume • Meta data issues • Network bandwidth and join complexity issues • Workload typically placed on workstation	• Business Enterprise view challenging • Redundant data costs • High DBA and operational costs • Data latency	Requires corporate leadership and vision

FIGURE 5.6 Alternative Architectures for Data Warehousing Efforts

(5.4) Data Warehousing Architectures

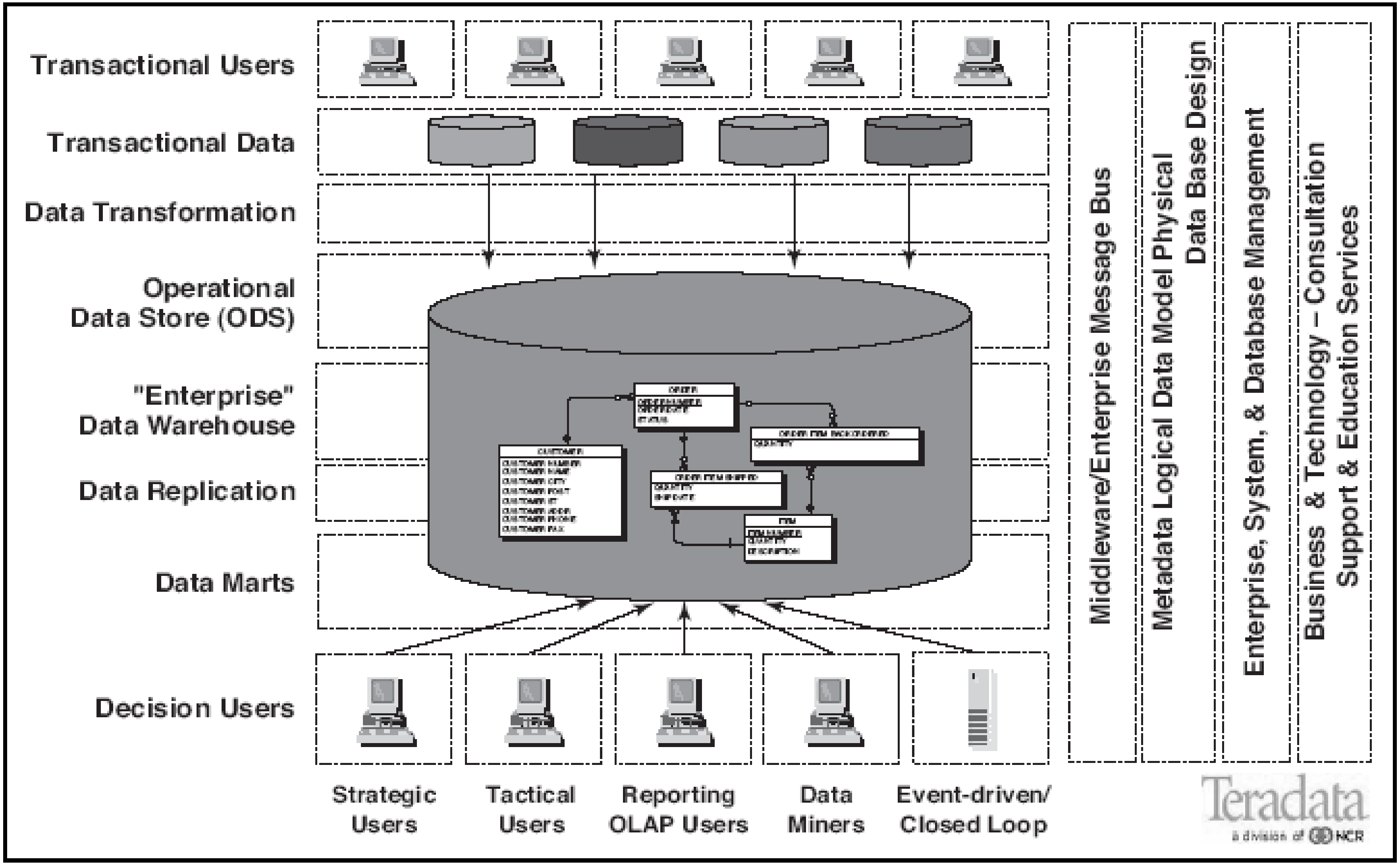

Transactional Users
Transactional Data
Data Transformation
Operational Data Store (ODS)
"Enterprise" Data Warehouse
Data Replication
Data Marts
Decision Users
Strategic Users
Tactical Users
Reporting OLAP Users
Data Miners
Event-driven/ Closed Loop
Middleware/Enterprise Message Bus
Metadata Logical Data Model Physical Data Base Design
Enterprise, System, & Database Management
Business & Technology – Consultation Support & Education Services
Teradata
a division of NCR

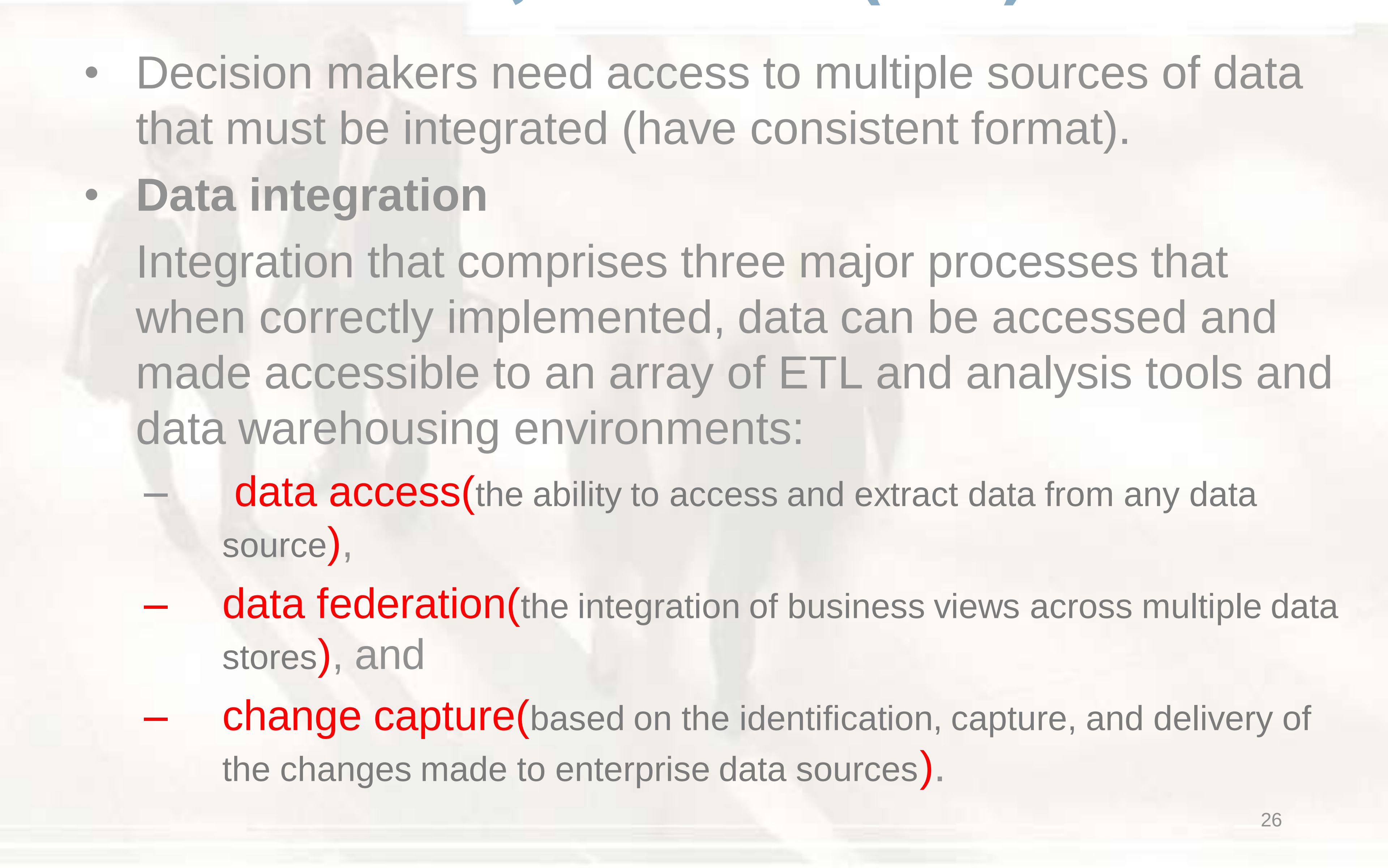
FIGURE 5.7 Teradata Corp.'s Enterprise Data Warehouse

(5.4) 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

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



- Decision makers need access to multiple sources of data that must be integrated (have consistent format).
- **Data integration**

  Integration that comprises three major processes that when correctly implemented, data can be accessed and made accessible to an array of ETL and analysis tools and data warehousing environments:

  - data access(the ability to access and extract data from any data source),
  - data federation(the integration of business views across multiple data stores), and
  - change capture(based on the identification, capture, and delivery of the changes made to enterprise data sources).

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

- SAS Institute have developed strong data integration tools
- Oracle business intelligence suite assists in integrating data as well
- A major purpose of data warehouse is to integrate data from multiple sources.
- Various technologies enable data integration:
  - Enterprise application integration (EAI)
  - Service-oriented architecture (SOA)
  - Enterprise information integration (EII)
  - Extraction, transformation, and load (ETL)

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

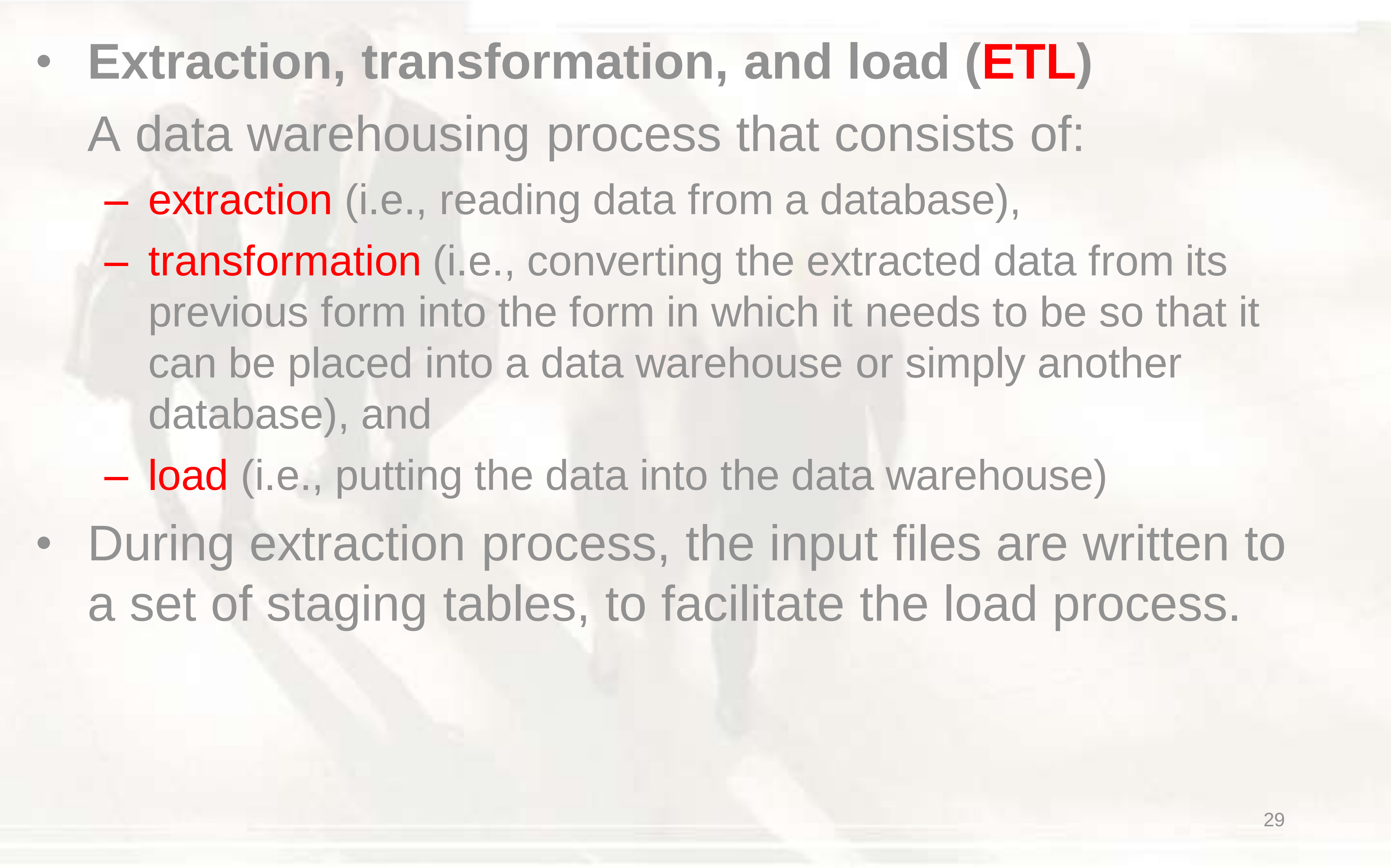
- **Enterprise application integration (EAI)**

  A technology that provides a vehicle for pushing data from source systems into a data warehouse. It focuses on sharing functionality (rather than data)

- **Enterprise information integration (EII)**

  An evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases

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



- **Extraction, transformation, and load (ETL)**

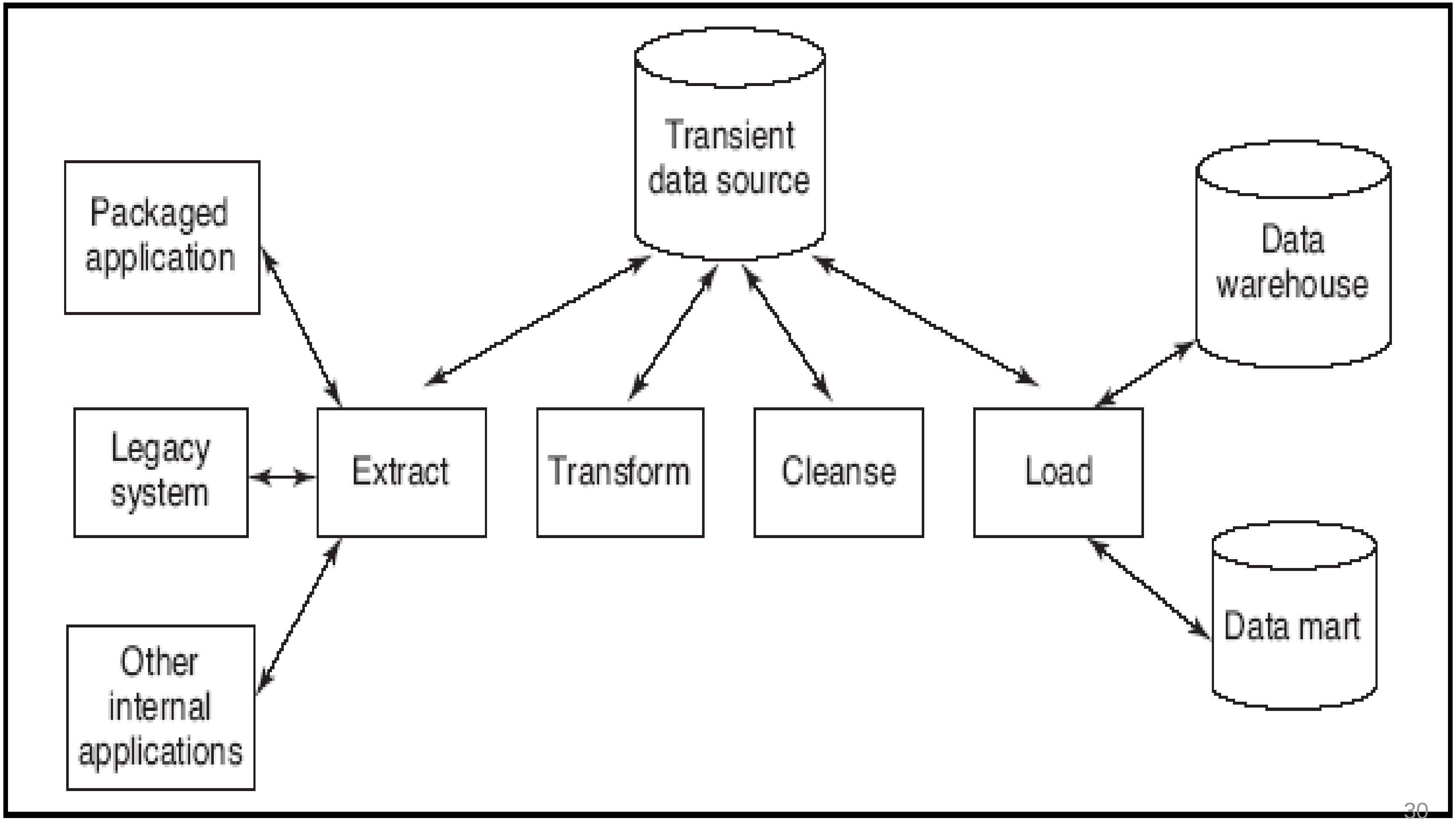
  A data warehousing process that consists of:

  - extraction (i.e., reading data from a database),
  - transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and
  - load (i.e., putting the data into the data warehouse)

- During extraction process, the input files are written to a set of staging tables, to facilitate the load process.

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

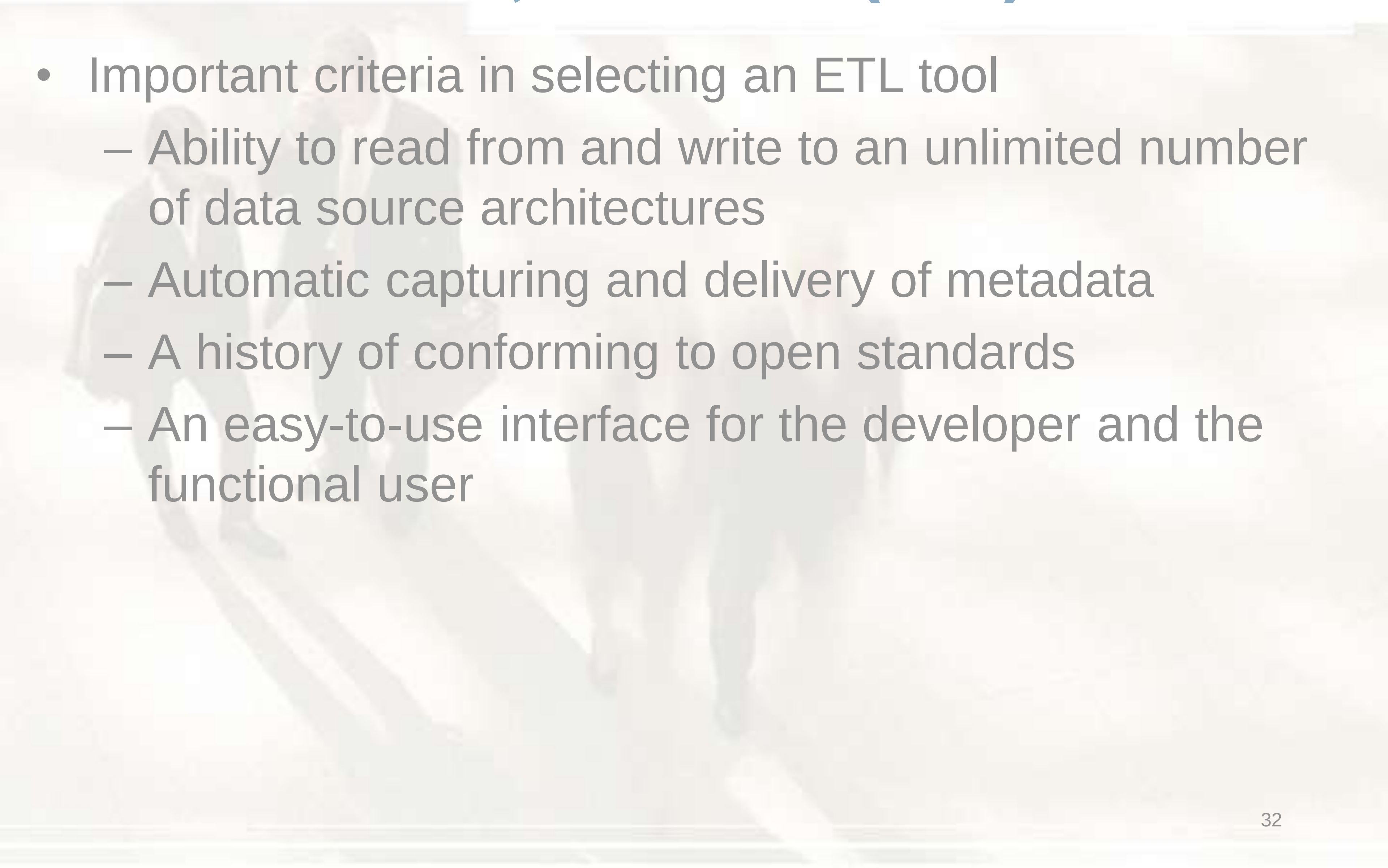
FIGURE 5.8 The ETL Process



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

- Issues affect whether an organization will purchase data transformation tools or build the transformation process itself:
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
  - It is difficult to measure how the IT organization is doing until it has learned to use the data transformation tools

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



- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data source 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

(5.6) Data Warehouse Development

- Data warehouse is very important for an organization because it comprises and influences many departments.
- It provides several benefits:
  - Direct benefits
  - Indirect benefits
- Direct benefits of a data warehouse
  - Allows end users to perform extensive analysis
  - Allows a consolidated view of corporate data (single version of the truth)
  - Better and more timely information
  - Enhanced system performance
  - Simplification of data access

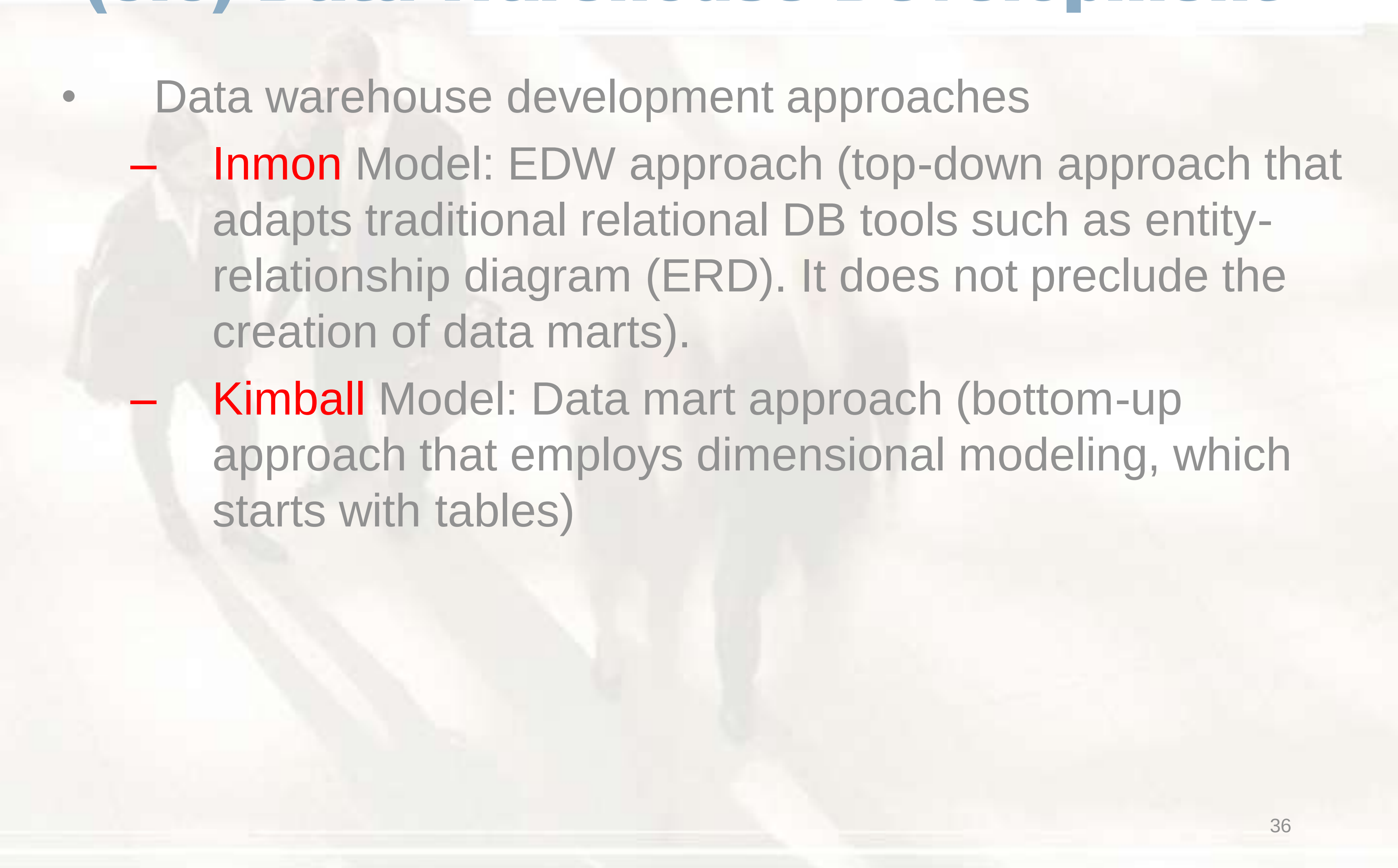
(5.6) Data Warehouse Development

- Indirect benefits result from end users using these direct benefits
  - Enhance business knowledge
  - Present competitive advantage
  - Enhance customer service and satisfaction
  - Facilitate decision making
  - Help in reforming business processes

(5.6) Data Warehouse Development

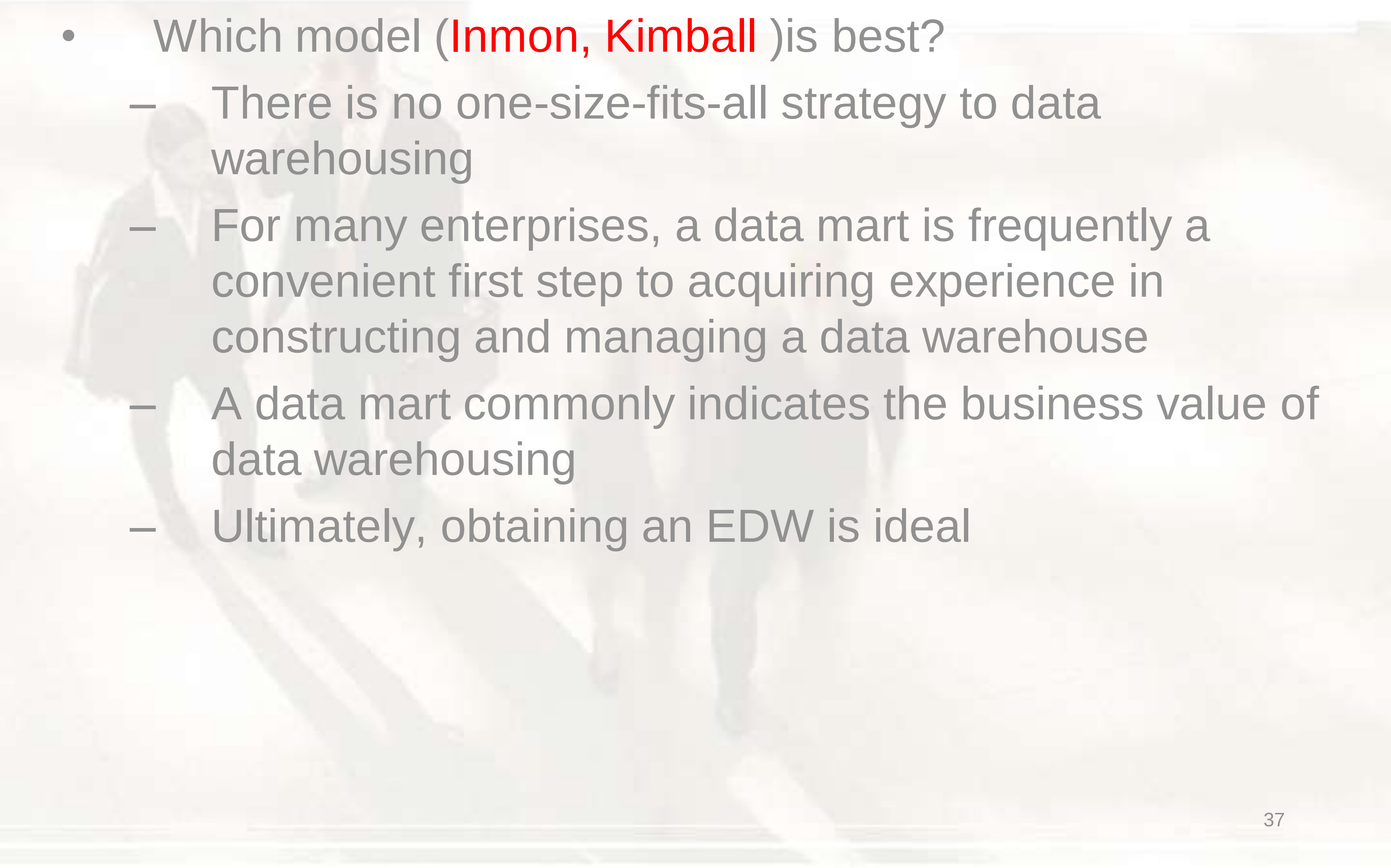
- Data warehouse vendors
  - Six guidelines to considered when developing a vendor list:
    1. Financial strength
    2. ERP linkages
    3. Qualified consultants
    4. Market share
    5. Industry experience
    6. Established partnerships

(5.6) Data Warehouse Development



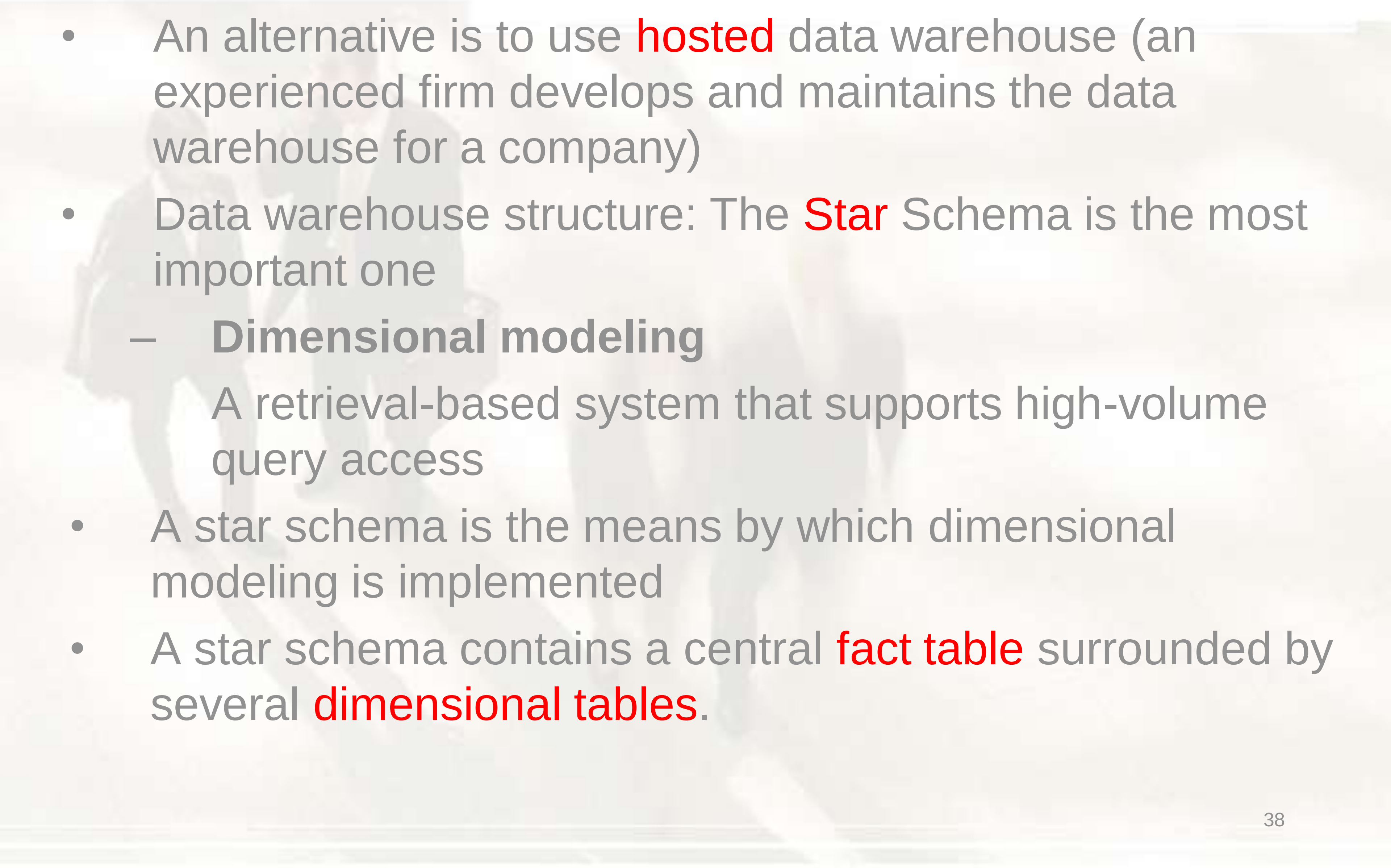
- Data warehouse development approaches
  - Inmon Model: EDW approach (top-down approach that adapts traditional relational DB tools such as entity-relationship diagram (ERD). It does not preclude the creation of data marts).
  - Kimball Model: Data mart approach (bottom-up approach that employs dimensional modeling, which starts with tables)

(5.6) Data Warehouse Development



- Which model (Inmon, Kimball)is best?
  - There is no one-size-fits-all strategy to data warehousing
  - For many enterprises, a data mart is frequently a convenient first step to acquiring experience in constructing and managing a data warehouse
  - A data mart commonly indicates the business value of data warehousing
  - Ultimately, obtaining an EDW is ideal

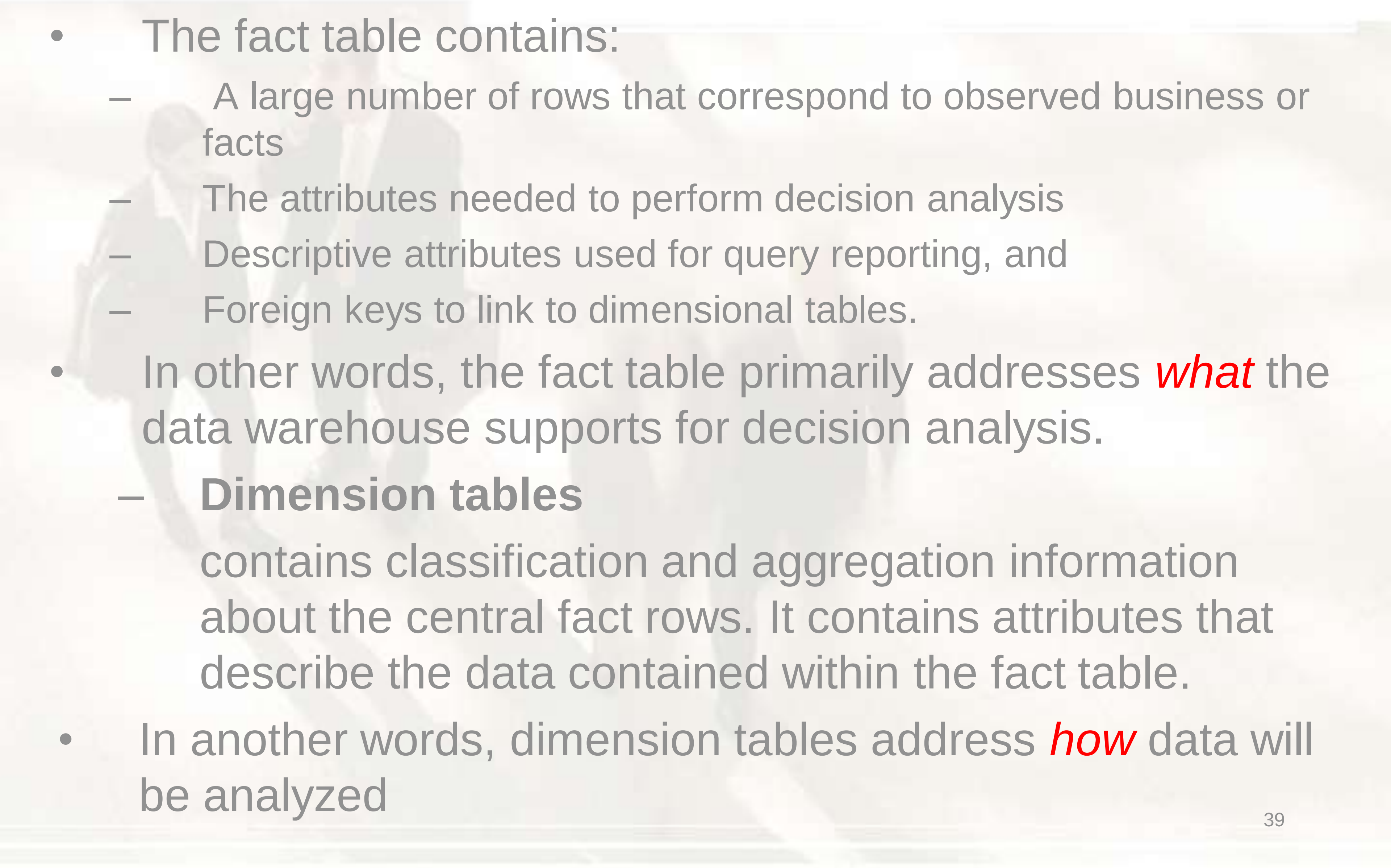
(5.6) Data Warehouse Development



- An alternative is to use hosted data warehouse (an experienced firm develops and maintains the data warehouse for a company)
- Data warehouse structure: The Star Schema is the most important one
  - **Dimensional modeling**

    A retrieval-based system that supports high-volume query access
- A star schema is the means by which dimensional modeling is implemented
- A star schema contains a central fact table surrounded by several dimensional tables.

(5.6) Data Warehouse Development



- The fact table contains:
  - A large number of rows that correspond to observed business or facts
  - The attributes needed to perform decision analysis
  - Descriptive attributes used for query reporting, and
  - Foreign keys to link to dimensional tables.
- In other words, the fact table primarily addresses *what* the data warehouse supports for decision analysis.
  - **Dimension tables**

    contains classification and aggregation information about the central fact rows. It contains attributes that describe the data contained within the fact table.
- In another words, dimension tables address *how* data will be analyzed

(5.6) Data Warehouse Development

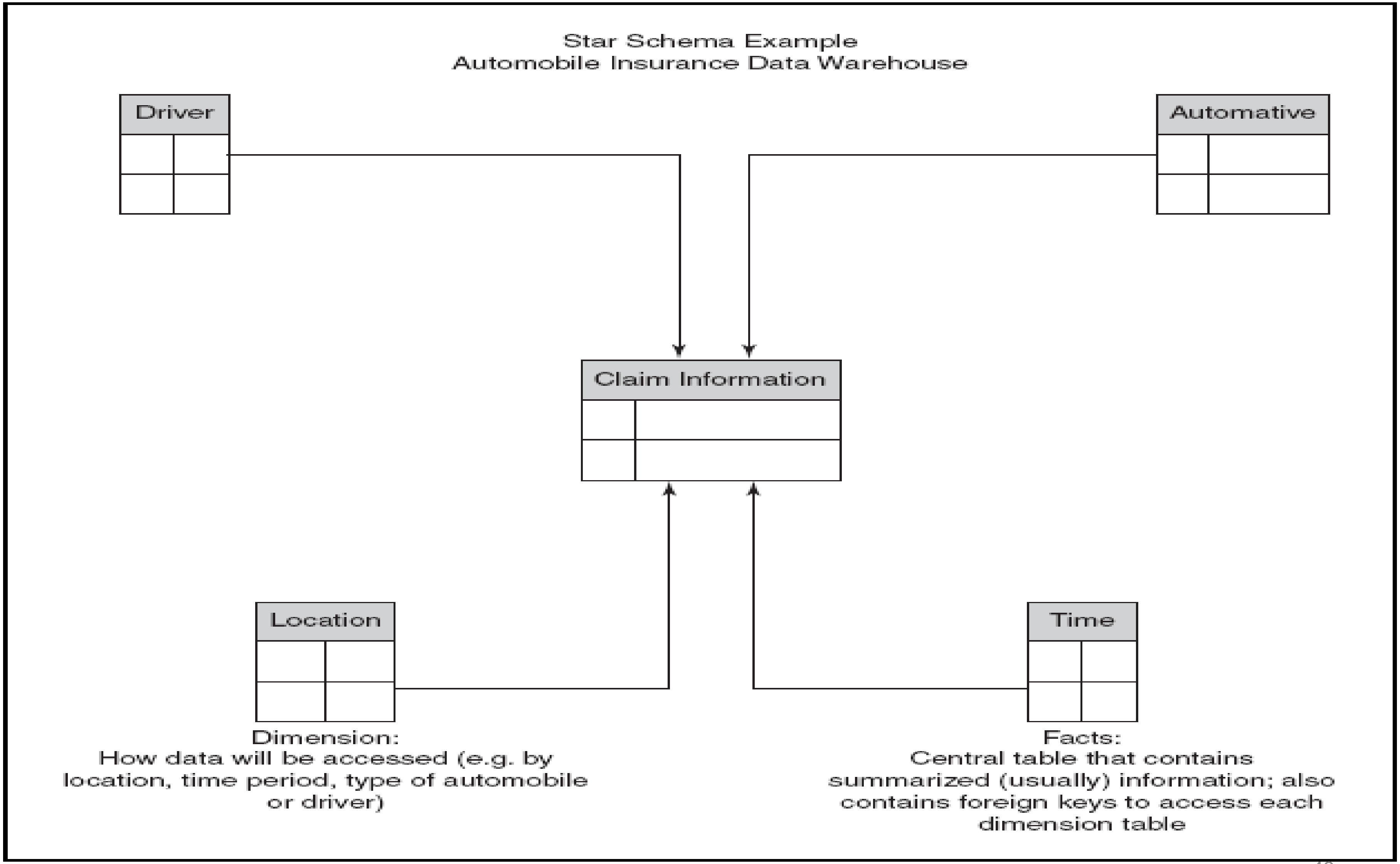

Star Schema Example
Automobile Insurance Data Warehouse
Driver
Automative
Claim Information
Location
Time
Dimension:
How data will be accessed (e.g. by location, time period, type of automobile or driver)
Facts:
Central table that contains summarized (usually) information; also contains foreign keys to access each dimension table

FIGURE 5.9 Star Schema

(5.6) Data Warehouse Development

- **Grain of a data warehouse is:**

  A definition of the highest level of detail that is supported in a data warehouse

  The grain indicates whether the DW is highly summarized or also includes detailed transaction data.
- High grain means no detail requests
- **Drill-down analysis is:**

  The process of probing beyond a summarized value to investigate each of the detail transactions that comprise the summary (we reach this because the grain is high)
- Low grain means more data being stored in DW.
- More detail → affects performance →makes response time longer

(5.6) Data Warehouse Development

- Data warehousing implementation issues
  - Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods
  - There are many facets to the project lifecycle, and no single person can be an expert in each area

(5.6) Data Warehouse Development

Eleven major tasks that could be performed in parallel for successful implementation of a data warehouse (Solomon, 2005) :

1. Establishment of service-level agreements and data-refresh requirements
2. Identification of data sources and their governance policies
3. Data quality planning
4. Data model design
5. ETL tool selection
6. Relational database software and platform selection
7. Data transport
8. Data conversion
9. Reconciliation process
10. Purge and archive planning
11. End-user support

(5.6) Data Warehouse Development

- Some best practices for implementing a data warehouse (Weir, 2002):
  - Project must fit with corporate strategy and business objectives
  - There must be complete buy-in to the project by executives, managers, and users
  - It is important to manage user expectations about the completed project
  - The data warehouse must be built incrementally
  - Build in adaptability

(5.6) Data Warehouse Development

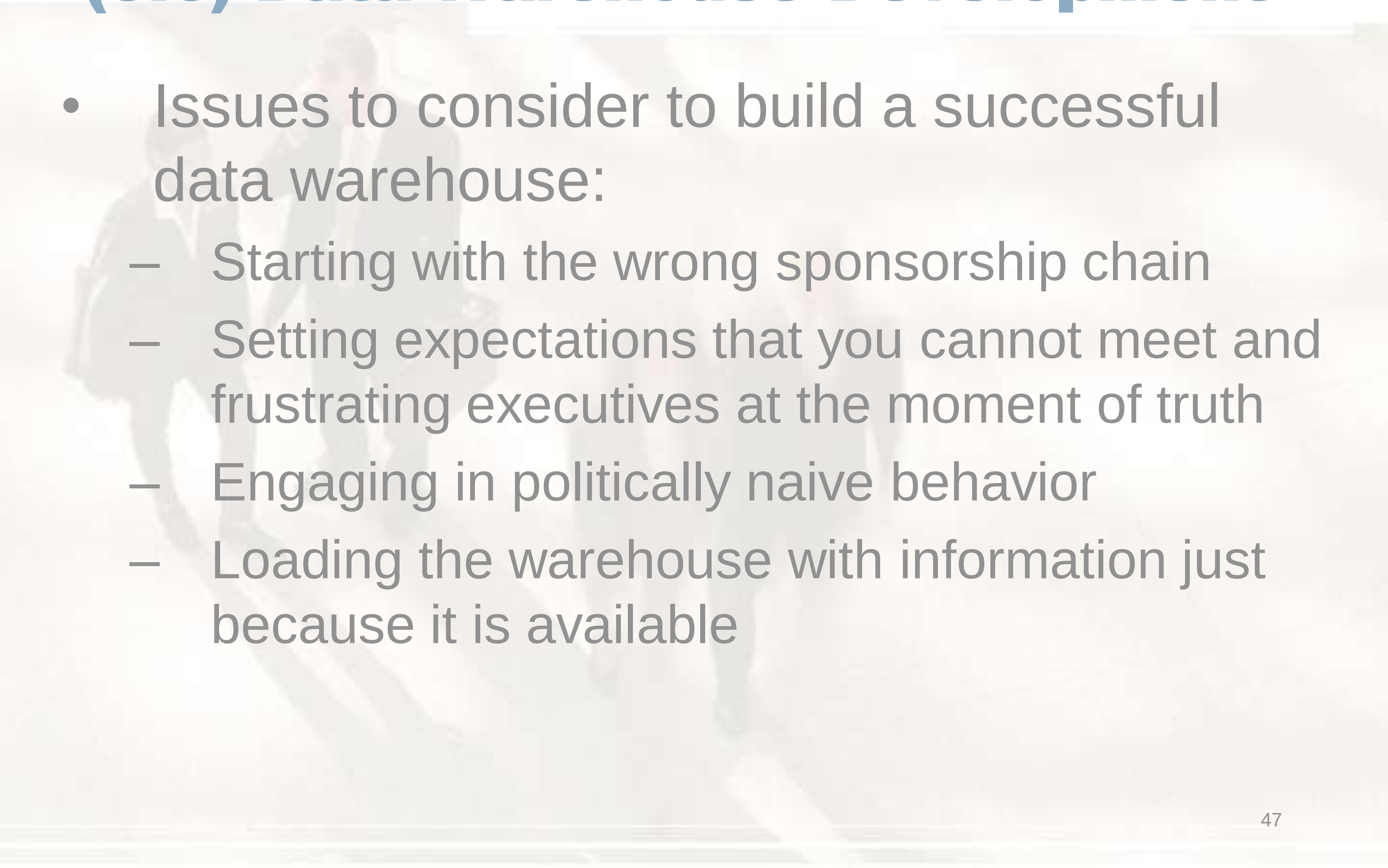
- Some best practices for implementing a data warehouse (Weir, 2002):
    - The project must be managed by both IT and business professionals
    - Develop a business/supplier relationship
    - Only load data that have been cleansed and are of a quality understood by the organization
    - Do not overlook training requirements
    - Be politically aware

(5.6) Data Warehouse Development

- Failure factors in data warehouse projects:
  - Cultural issues being ignored
  - Inappropriate architecture
  - Unclear business objectives
  - Missing information
  - Unrealistic expectations
  - Low levels of data summarization
  - Low data quality

(5.6) Data Warehouse Development

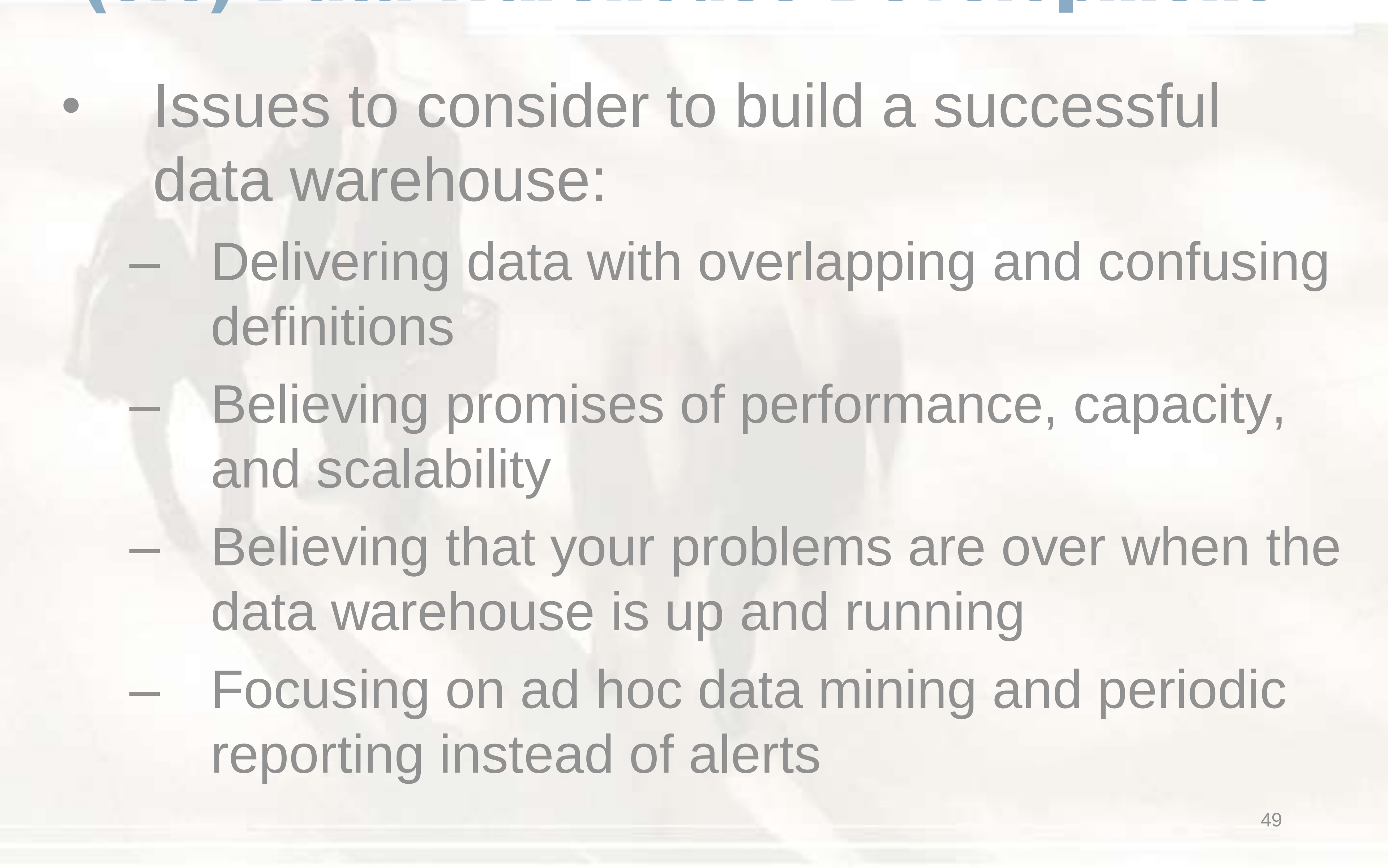
- Issues to consider to build a successful data warehouse:
  - Starting with the wrong sponsorship chain
  - Setting expectations that you cannot meet and frustrating executives at the moment of truth
  - Engaging in politically naive behavior
  - Loading the warehouse with information just because it is available



(5.6) Data Warehouse Development

- Issues to consider to build a successful data warehouse:
  - Believing that data warehousing database design is the same as transactional database design
  - Choosing a data warehouse manager who is technology oriented rather than user oriented
  - Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video

(5.6) Data Warehouse Development

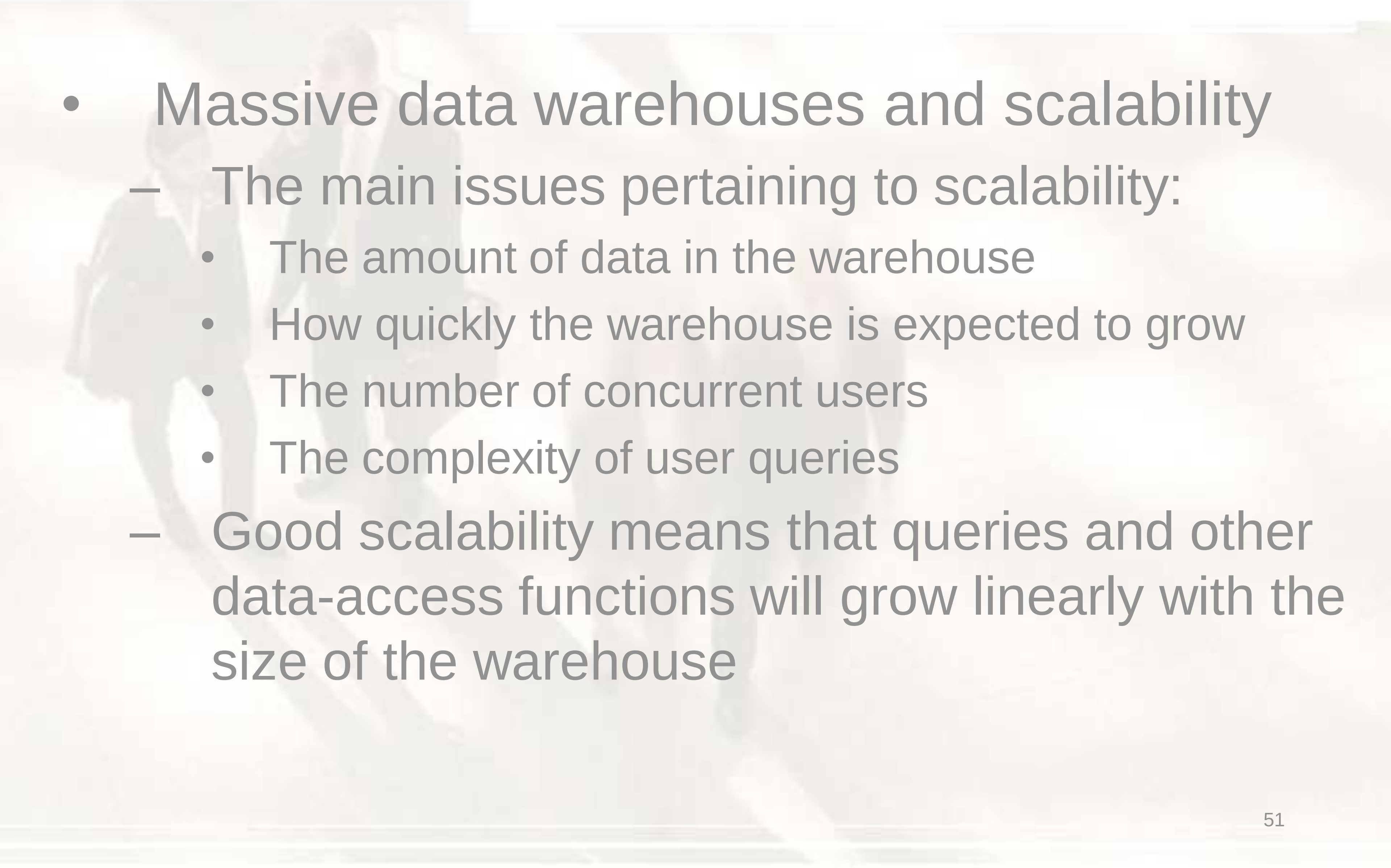


- Issues to consider to build a successful data warehouse:
  - Delivering data with overlapping and confusing definitions
  - Believing promises of performance, capacity, and scalability
  - Believing that your problems are over when the data warehouse is up and running
  - Focusing on ad hoc data mining and periodic reporting instead of alerts

(5.6) Data Warehouse Development

- Implementation factors that can be categorized into three criteria
  - Organizational issues
  - Project issues
  - Technical issues
- User participation in the development of data and access modeling is a critical success factor in data warehouse development

(5.6) Data Warehouse Development



- Massive data warehouses and 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

(5.7) Real-Time Data Warehousing

- **Real-time (active) data warehousing**

  The process of loading and providing data via a data warehouse as they become available
- RDW or ADW helps ,making fast and consistent decisions.

(5.7) Real-Time Data Warehousing

- Levels of data warehouses (evolution):

1. Reports what happened
2. Some analysis occurs
3. Provides prediction capabilities,
4. Operationalization
5. Becomes capable of making events happen (such as creating sales and making campaigns, or identify opportunities)

(5.7) Real-Time Data Warehousing

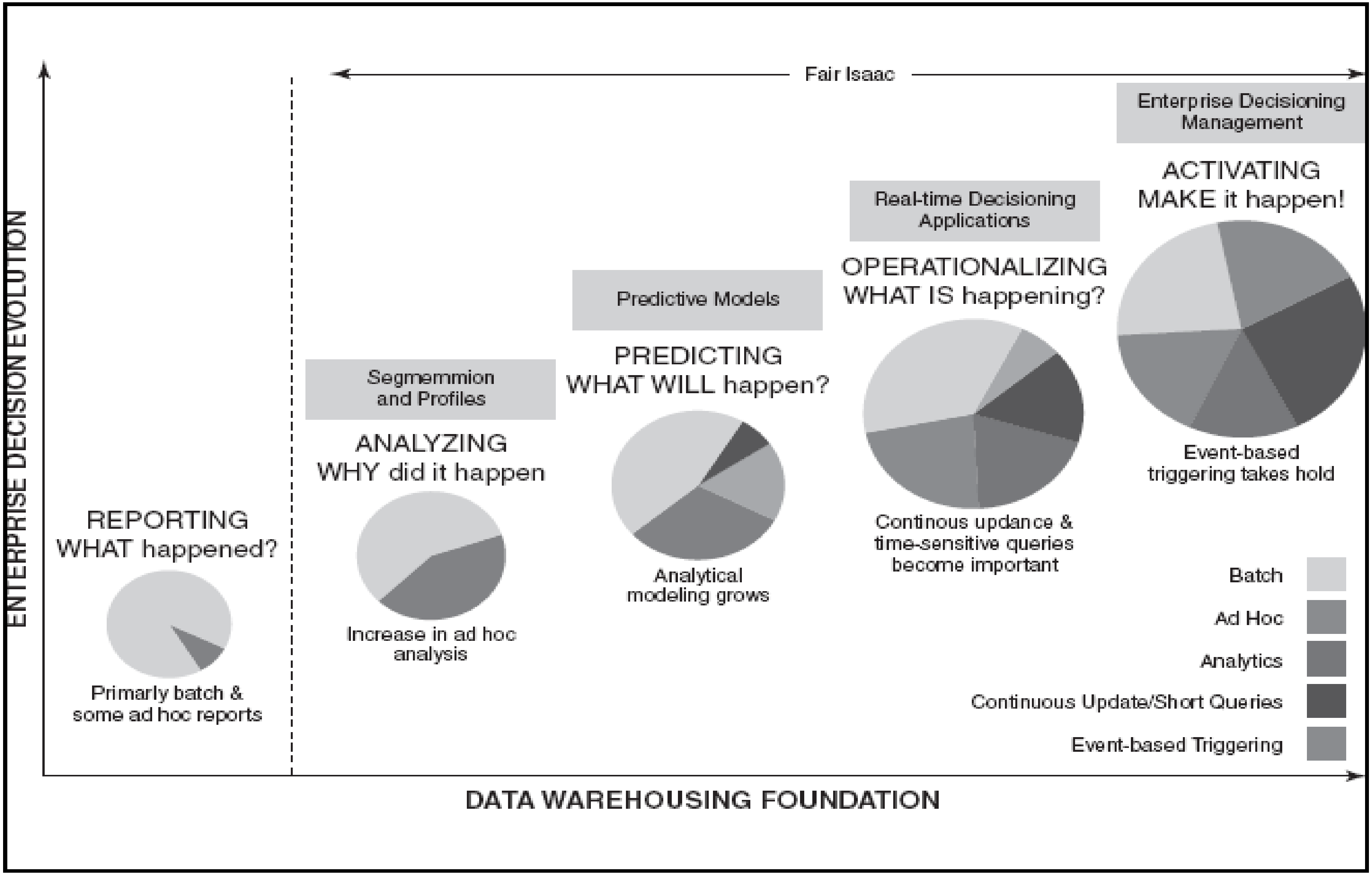
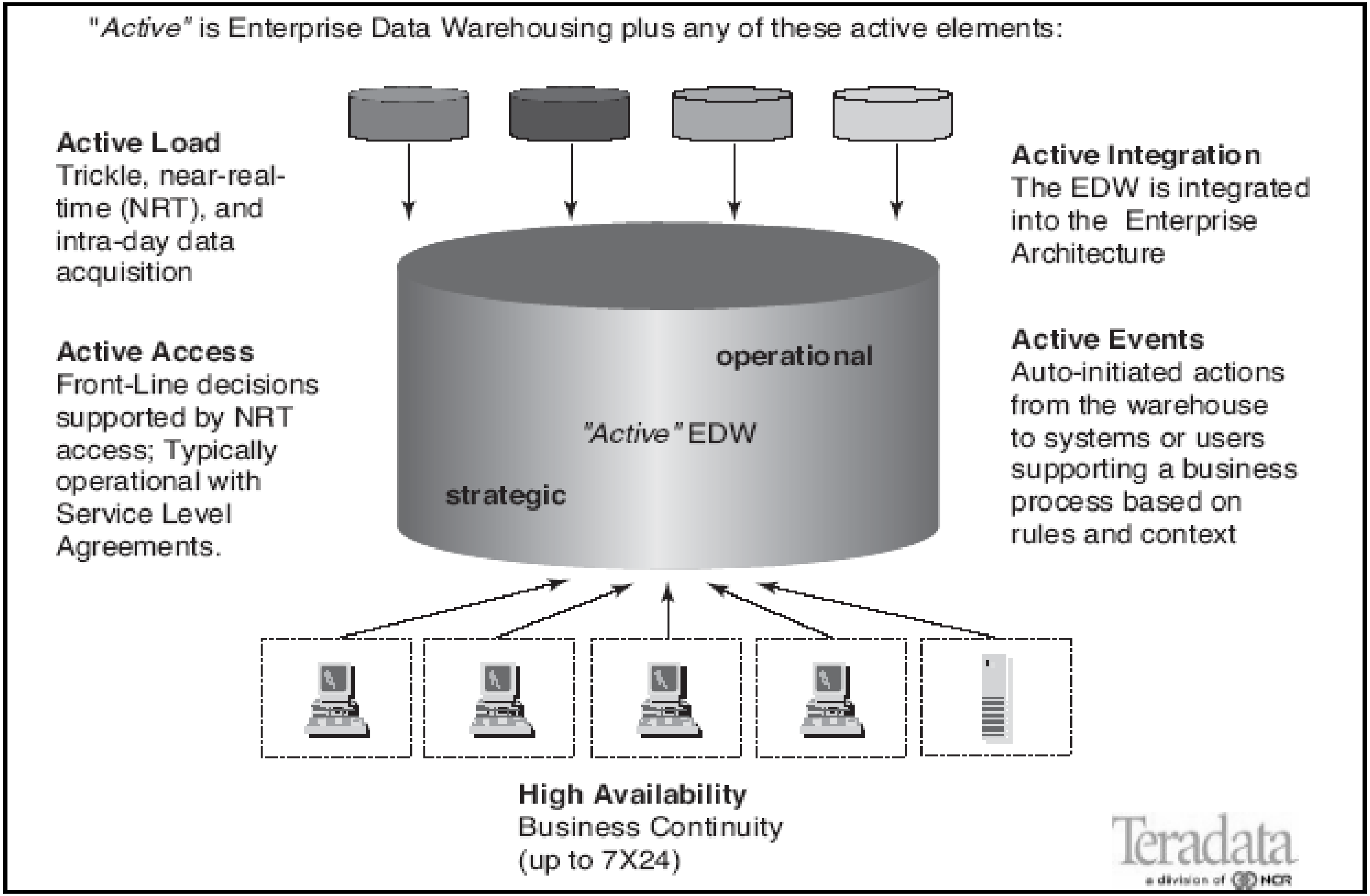


FIGURE 5.10 Enterprise Decision Evolution

(5.7) Real-Time Data Warehousing



"*Active*" is Enterprise Data Warehousing plus any of these active elements:

Active Load
Trickle, near-real-time (NRT), and intra-day data acquisition

Active Access
Front-Line decisions supported by NRT access; Typically operational with Service Level Agreements.

Active Integration
The EDW is integrated into the Enterprise Architecture

Active Events
Auto-initiated actions from the warehouse to systems or users supporting a business process based on rules and context

High Availability
Business Continuity
(up to 7X24)

FIGURE 5.11 The Teradata Active EDW

(5.7) Real-Time Data Warehousing

- The need for real-time data
  - A business often cannot afford to wait a whole day for its operational data to load into the data warehouse for analysis
  - Provides incremental real-time data showing every state change and almost analogous patterns over time
  - Maintaining metadata in sync is possible
  - Less costly to develop, maintain, and secure one huge data warehouse so that data are centralized for BI/BA tools
  - An EAI with real-time data collection can reduce or eliminate the nightly batch processes