Database Systems -

Introduction to Databases and Data Warehouses

CHAPTER 7 - Data Warehousing Concepts

Copyright (c) 2014 Pearson Education, Inc.

INTRODUCTION

- A typical organization maintains and utilizes a number of operational data sources.
  - The operational data sources include the databases and other data repositories which are used to support the organization's day-to-day operations
- A data warehouse is created within an organization as a separate data store whose primary purpose is data analysis.

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Often, the same fact can have both an operational and an analytical purpose. For example, data describing that customer X bought product Y in store Z can be stored in an operational database for business-process support purposes, such as inventory monitoring or financial transaction record keeping. That same fact can also be stored in a data warehouse where, combined with a vast number of similar facts accumulated over a period of time, it serves to reveal important trends, such as sale patterns or customer behavior.

INTRODUCTION

- Two main reasons for the creation of a data warehouse as a separate analytical database
  - The performance of operational day-to-day tasks involving data use can be severely diminished if such tasks have to compete for computing resources with analytical queries
  - It is often impossible to structure a database which can be used in an efficient manner for both operational and analytical purposes

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

INTRODUCTION

- **Operational information (transactional information)** - the information collected and used in support of day to day operational needs in businesses and other organizations
- **Analytical information** - the information collected and used in support of analytical tasks
  - Analytical information is based on operational (transactional) information

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Examples of operational information and analytical information are given on Pages 207-208.

OPERATIONAL VS. ANALYTICAL INFORMATION

Operational Data	Analytical Data
Data Makeup Differences	
Typical Time-Horizon: Days/Months	Typical Time-Horizon: Years
Detailed	Summarized (and/or Detailed)
Current	Values over time (Snapshots)
Technical Differences	
Small Amounts used in a Process	Large Amounts used in a Process
High frequency of Access	Low/Modest frequency of Access
Can be Updated	Read (and Append) Only
Non-Redundant	Redundancy not an Issue
Functional Differences	
Used by all types of employees for tactical purposes	Used by a narrower set of users for decision making
Application Oriented	Subject Oriented

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The discussion and examples of these differences is given on Pages 208-210.

Troy University Resources

Application Oriented vs. Subject Oriented– Example

An **application-oriented** database serving the Vitality Health Club Visits and Payments Application

HEALTH CLUB MEMBER

MemberID	MemberName	MemberGender	MLevelID	DateMembershipPaid
111	Joe	M	A	1/1/2013
222	Sue	F	B	1/1/2013
333	Pam	F	A	1/2/2013
...	...	...	...	...

MEMBERSHIP LEVEL

MLevelID	MLevelType	MLevelFee	MLevelDescription
A	Gold	$100	Includes the Pool Usage
B	Basic	$50	No Pool Usage

DAILY VISIT FROM NONMEMBERS

DVisitTID	DVisitLevelID	DVisitDate	DVisitorGender
11xx22	YP	1/1/2013	M
11xx23	NP	1/2/2013	M
11xx24	YP	1/2/2013	F
...	...	...	...

VISIT LEVEL

DVisitLevelID	DVisitLevelFee	DVisitLevelType
YP	$15	With Pool Usage
NP	$10	Without Pool Usage

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

This example is discussed on Pages 210-212.

Application Oriented vs. Subject Oriented– Example

A **subject-oriented** database for the analysis of the subject *revenue* in the Vitality Health Club

REVENUE

RevenueRecordID	Date	GeneratedBy	ClientGender	Pool Use Included in Purchase	Amount
1000	1/1/2013	Member	M	Yes	$100
1001	1/1/2013	Member	F	No	$50
1002	1/1/2013	Nonmember	M	Yes	$15
1003	1/2/2013	Member	F	Yes	$100
1004	1/2/2013	Nonmember	M	No	$10
1005	1/2/2013	Nonmember	F	Yes	$15
…				…	…

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

This example is discussed on Pages 210-212.

Troy University Resources

THE DATA WAREHOUSE DEFINITION

- The data warehouse is a **structured repository** of **integrated**, **subject-oriented**, **enterprise-wide**, **historical**, and **time-variant** data. The purpose of the data warehouse is the **retrieval of analytical information**. A data warehouse can store **detailed and/or summarized data**.

Copyright (c) 2014 Pearson Education, Inc.

THE DATA WAREHOUSE DEFINITION

- **Structured repository**
  - The data warehouse is a *database* containing analytically useful information
  - Any database is a *structured repository* with its structure represented in its metadata

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The data warehouse, being a database, is a structured repository. In other words, the data warehouse is not a disorganized random mass of data.

Troy University Resources

THE DATA WAREHOUSE DEFINITION

- Integrated
  - The data warehouse integrates the analytically useful data from the various operational databases (and possibly other sources)
  - *Integration* refers to this process of bringing the data from multiple data sources into a singular data warehouse.

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

In the process of bringing the data from operational databases into the data warehouse, no data is actually removed from the operational sources. Instead, the analytically useful data from various operational databases is copied and brought into the data warehouse.

THE DATA WAREHOUSE DEFINITION

- **Subject-oriented**
  - The term *subject-oriented* refers to the fundamental difference in the purpose of an operational database system and a data warehouse.
    - An operational database system is developed in order to support a specific business operation
    - A data warehouse is developed to analyze specific business subject areas

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

THE DATA WAREHOUSE DEFINITION

- **Enterprise-wide**
  - The term *enterprise-wide* refers to the fact that the data warehouse provides an organization-wide view of the analytically useful information it contains

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

For example, if one of the subjects of the data warehouse is cost, than all of the analytically useful data regarding cost within the operational data sources in the entire organization will be brought into the data warehouse.

THE DATA WAREHOUSE DEFINITION

- Historical
  - The term *historical* refers to the larger time horizon in the data warehouse than in the operational databases

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

For example, many traditional operational databases have a time horizon of 60 to 90 days, where it is quite common for data warehouses to contain multiple years' worth of data.

Troy University Resources

THE DATA WAREHOUSE DEFINITION

- Time variant
  - The term *time variant* refers to the fact that a data warehouse contains slices or snapshots of data from different periods of time across its time horizon
    - With the data slices, the user can create reports for various periods of time within the time horizon

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

For example, if the subject of analysis is cost and the time horizon is a number of years, we can analyze and compare the cost for the first quarter from a year ago versus the cost for the first quarter from two years ago.

THE DATA WAREHOUSE DEFINITION

- **Retrieval of analytical information**
  - A data warehouse is developed for the *retrieval of analytical information*, and it is not meant for direct data entry by the users.
    - The only functionality available to the users of the data warehouse is retrieval
    - The data in the data warehouse is not subject to changes.
    - The data in the data warehouse is referred to as *non-volatile*, *static*, or *read-only*

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The data in the data warehouse is not subject to modifications, insertions or deletions by the users.
New data in the data warehouse is periodically loaded from the operational data sources and appended to the existing data, in an automatic fashion.
The data that eventually gets older than the required time horizon is purged from the data warehouse.

THE DATA WAREHOUSE DEFINITION

- **Detailed and/or summarized data**
  - A data warehouse, depending on its purpose, may include the *detailed data* or *summary data* or *both*
  - A data warehouse that contains the data at the finest level of detail is the most powerful

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Examples of detailed and summarized data in the data warehouse are given on page 214. The discussion about the level of detail in data warehouses in various circumstances is given on Page 214.

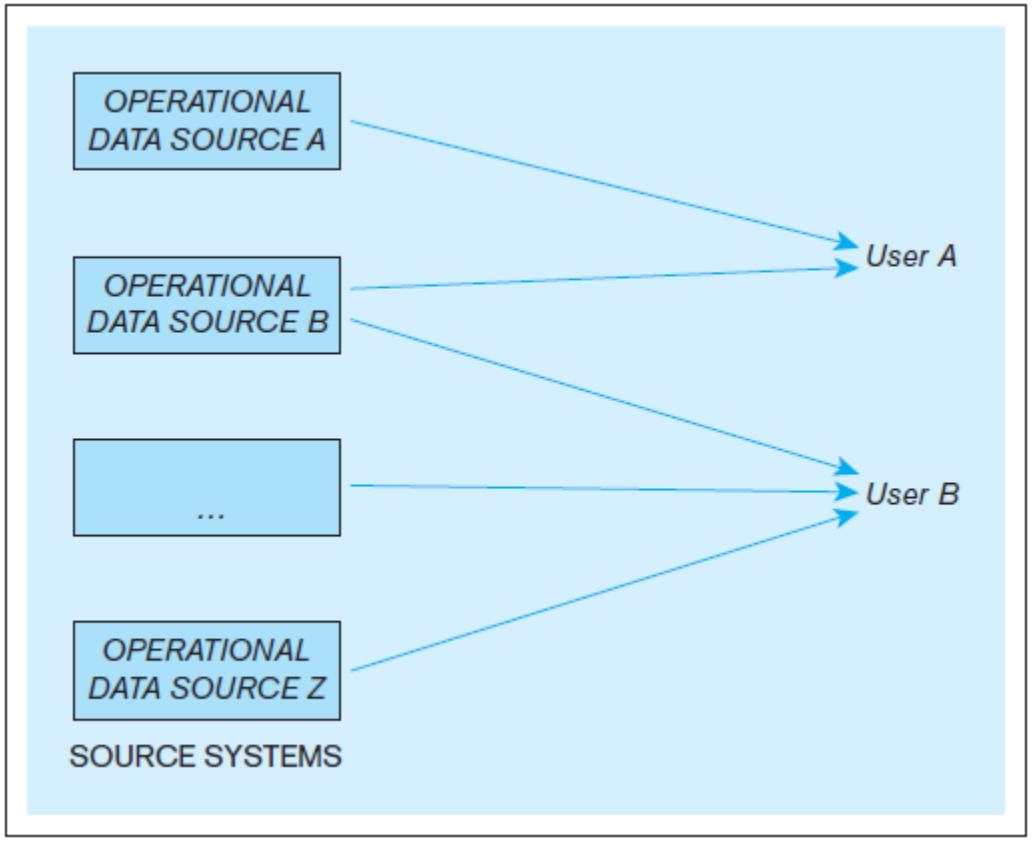
DATA WAREHOUSE COMPONENTS

- **Data warehouse components**
  - Source systems
  - Extraction-transformation-load (ETL) infrastructure
  - Data warehouse
  - Front-end applications

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

Example - The use of operational data sources for operational purposes in an organization

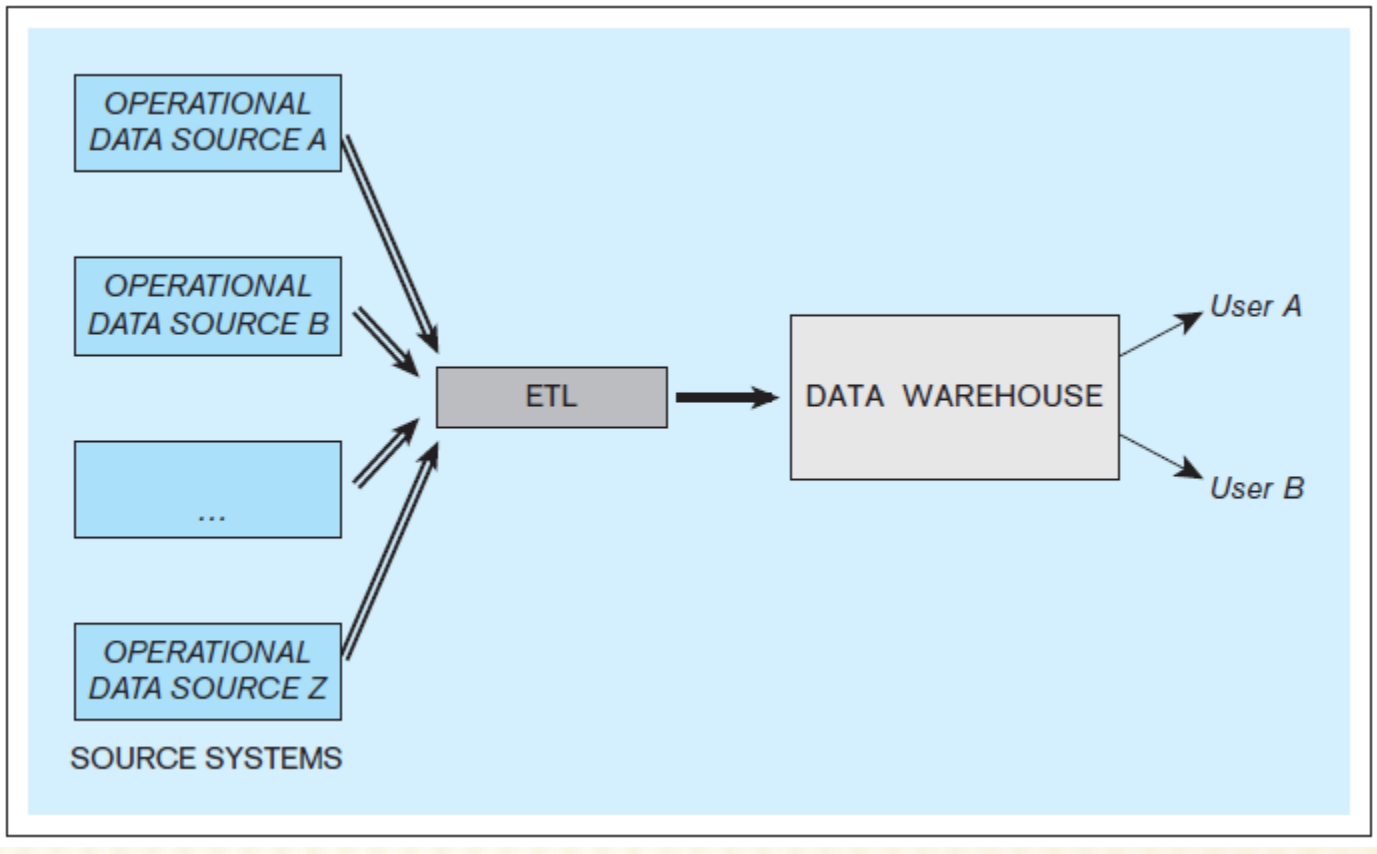


Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

This example is discussed on Pages 214-215.

Example - The core components of a data warehousing system



Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

This example is discussed on Pages 214-215.

Troy University Resources

DATA WAREHOUSE COMPONENTS

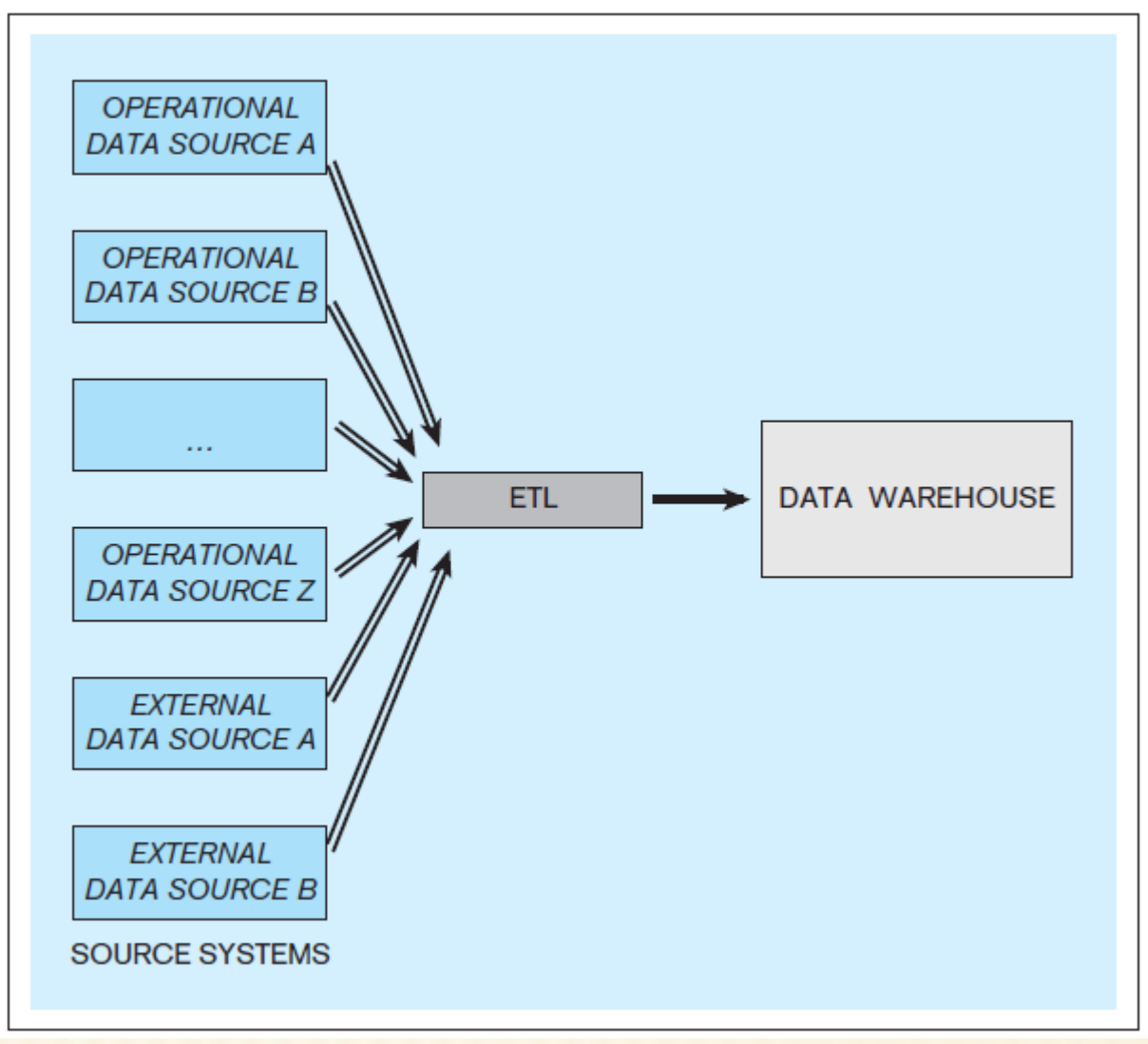
- **Source systems**
  - In the context of data warehousing, *source systems* are operational databases and other operational data repositories (in other words, any sets of data used for operational purposes) that provide analytically useful information for the data warehouse's subjects of analysis
  - Every operational data store that is used as a source system for the data warehouse has two purposes:
    - The original operational purpose
    - As a source system for the data warehouse
  - Source systems can include *external data sources*

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

External data sources examples include market research data, census data, stock market data, weather data.

Example - A data warehouse with internal and external source systems



Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

This example is discussed on Page 215.

Troy University Resources

DATA WAREHOUSE COMPONENTS

- **Data warehouse**
  - The *data warehouse* is sometimes referred to as the *target system*, to indicate the fact that it is a destination for the data from the source systems
  - A typical data warehouse periodically retrieves selected analytically useful data from the operational data sources

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

In so-called “active” data warehouses, the retrieval of data from operational data sources is continuous.

DATA WAREHOUSE COMPONENTS

- **ETL infrastructure**
  - The infrastructure that facilitates the retrieval of data from operational databases into the data warehouses
  - ETL includes the following tasks:
    - *Extracting* analytically useful data from the operational data sources
    - *Transforming* such data so that it conforms to the structure of the subject-oriented target data warehouse model (while ensuring the quality of the transformed data)
    - *Loading* the transformed and quality assured data into the target data warehouse

Copyright (c) 2014 Pearson Education, Inc.

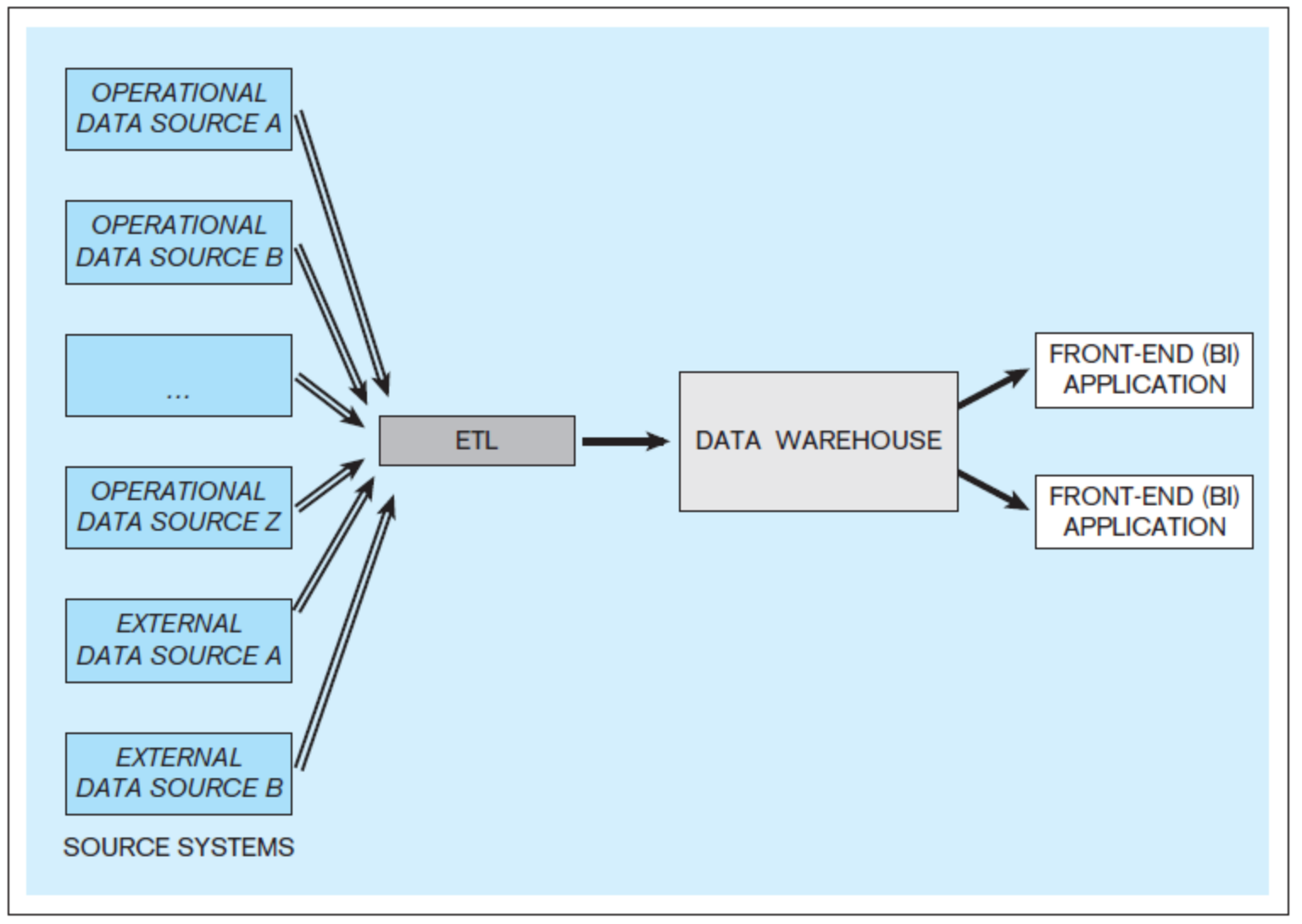
Troy University Resources

DATA WAREHOUSE COMPONENTS

- **Data warehouse front-end (BI) applications**
  - Used to provide access to the data warehouse for users who are engaging in indirect use

Copyright (c) 2014 Pearson Education, Inc.

Example - A data warehouse with front-end applications



Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

DATA MARTS

- Data mart
  - A data store based on the same principles as a data warehouse, but with a more limited scope

	DATA WAREHOUSE	DATA MART
Subjects	*Multiple*	*Single*
Data Sources	*Many*	*Fewer*
Typical Size	*Very big (routinely terabytes of data and larger)*	*Not as big*
Implementation Time	*Relatively long (months, years)*	*Not as long*
Focus	*Organization-wide*	*Often narrower than organization-wide*

Copyright (c) 2014 Pearson Education, Inc.

DATA MARTS

- **Independent data mart**
  - Stand-alone data mart, created in the same fashion as the data warehouse
  - Independent data mart has its own source systems and ETL infrastructure

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

DATA MARTS

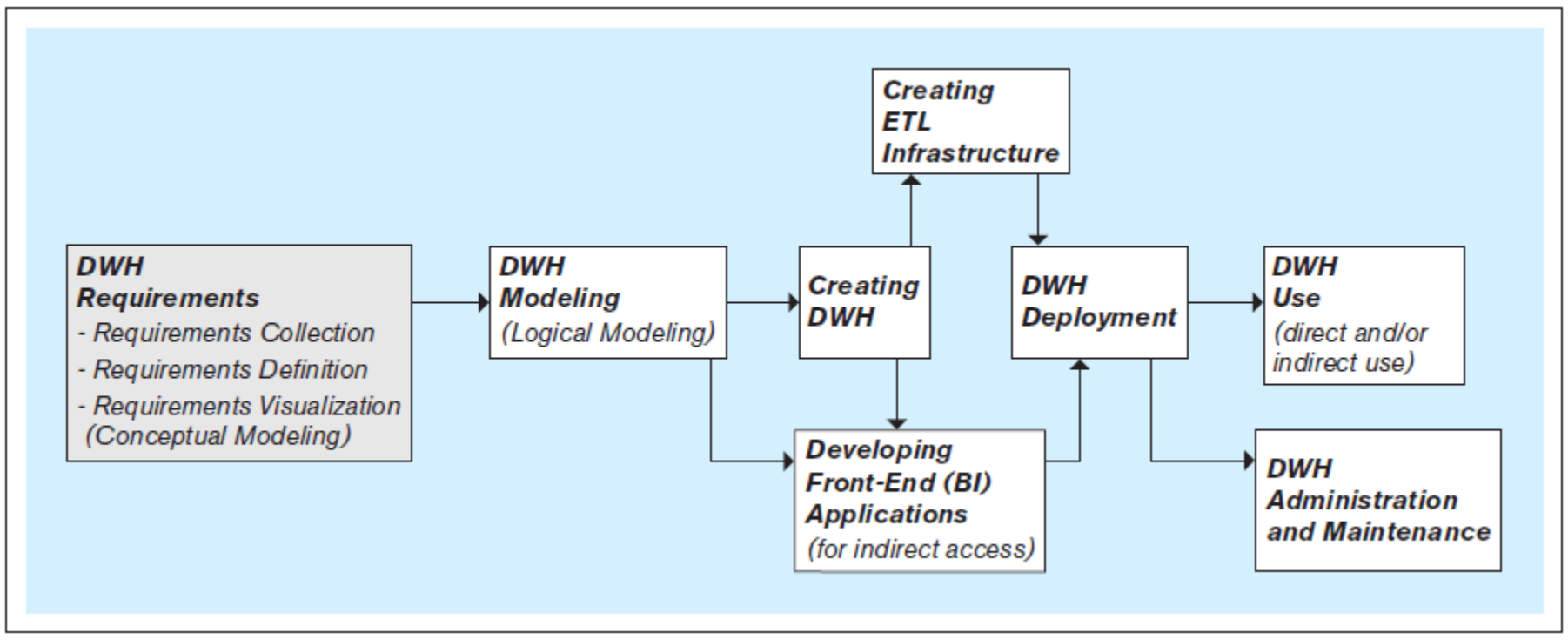
- Dependent data mart
  - Does not have its own source systems
  - The data comes from the data warehouse

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Dependent data marts provide users with a subset of the data from the data warehouse, in cases when users or applications do not want, need, or are not allowed to have access to all the data in the entire data warehouse.

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES



Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Requirements collection, definition, and visualization** - results in the requirements specifying the desired capabilities and functionalities of the future data warehouse
  - The requirements are based on the analytical needs that can be met by the data in the internal data source systems and available external data sources
  - The requirements are collected through interviewing various stakeholders of the data warehouse
  - In addition to interviews, additional methods for eliciting requirements from the stakeholders can be used

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The first and most critical step in the development of the data warehouse.

The requirements collection process aspires to analytically take advantage of all data available for consideration, but it cannot be based on data that is not available or does not exist.

Additional methods for eliciting requirements include focus groups, questionnaires, surveys, and observations of existing analytical practices to determine what users really do with data and what data they actually use and need

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Requirements collection, definition, and visualization**
  - The **collected** requirements should be clearly **defined** and stated in a written document, and then **visualized** as a conceptual data model

Copyright (c) 2014 Pearson Education, Inc.

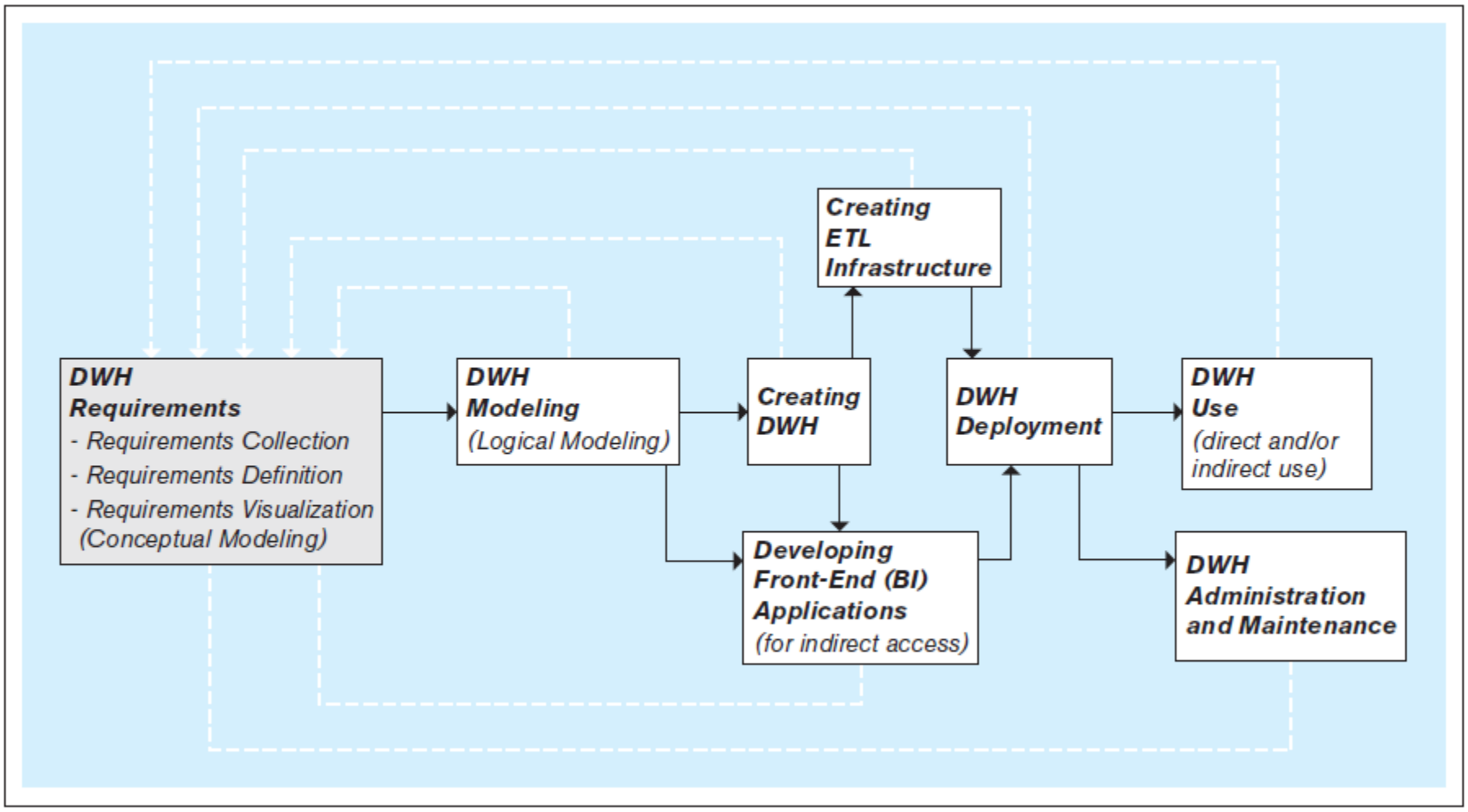
SPEAKER NOTES

Conceptual data warehouse modeling techniques include ER modeling (covered earlier in Chapter 2) and dimensional modeling (covered later in Chapter 8).

Troy University Resources

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

Iterative nature of the data warehouse requirements collection, definition, and visualization process



Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The discussion about iterative nature of the data warehouse requirements collection, definition, and visualization process is given on Pages 218-219.

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Data warehouse modeling** (**logical data warehouse modeling**) - creation of the data warehouse data model that is implementable by the DBMS software

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Creating the data warehouse** - using a DBMS to implement the data warehouse data model as an actual data warehouse
  - Typically, data warehouses are implemented using a relational DBMS (RDBMS) software

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Regular relational DBMS packages, such as Microsoft SQL Server or Oracle, can be used for implementing both operational databases and data warehouses. However, other relational DBMS packages, such as Teradata, that are specialized for processing large amounts of data typically found in data warehouses are also available for implementation of data warehouses.

Creating data marts can be accomplished in the same fashion as creating data warehouses by using relational DBMS packages. In some cases, data marts can also be implemented as so-called “cubes,” using a data management technology different from relational DBMS as described in Chapter 9.

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- Creating ETL infrastructure
  - Creating necessary procedures and code for:
    - Automatic extraction of relevant data from the operational data sources
    - Transformation of the extracted data, so that its quality is assured and its structure conforms to the structure of the modeled and implemented data warehouse
    - The seamless load of the transformed data into the data warehouse
  - Due to the amount of details that have to be considered, creating ETL infrastructure is often the most time- and resource-consuming part of the data warehouse development process

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

ETL infrastructure has to account for and reconcile all of the differences in the metadata and the data between the operational sources and the target data warehouses. In many cases, organizations have multiple separate operational sources with overlapping information. In such cases, the process of creating the ETL infrastructure involves deciding how to bring in such information without creating misleading duplicates (i.e., how to bring in all the useful information while avoiding the uniqueness data quality problem).

Troy University Resources

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Developing front-end (BI) applications** - designing and creating applications for indirect use by the end-users
  - Front-end applications are included in most data warehousing systems and are often referred to as *business intelligence (BI) applications*
  - Front-end applications contain interfaces (such as forms and reports) accessible via a navigation mechanism (such as a menu)

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The design and creation of data warehouse front-end applications can take place in parallel with data warehouse creation. For example, the look and feel of the front-end applications, as well as the number and functionalities of particular components (e.g., forms and reports), can be determined before the data warehouse is implemented. This can be done based on the data warehouse model and the requirements specifying the capabilities and functionalities of the system needed by the end users. The actual creation of the front-end application involves connecting it to the implemented data warehouse, which can only be done once the data warehouse is implemented.

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Data warehouse deployment** - releasing the data warehouse and its front-end (BI) applications for use by the end users

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

Typically, prior to this step, the initial load populating the created data warehouse with an initial set of data from the operational data sources via the ETL infrastructure is executed.

Troy University Resources

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Data warehouse use** - the retrieval of the data in the data warehouse
  - Indirect use
    - Via the front-end (BI) applications
  - Direct use
    - Via the DBMS
    - Via the OLAP (BI) tools

Copyright (c) 2014 Pearson Education, Inc.

STEPS IN THE DEVELOPMENT OF DATA WAREHOUSES

- **Data warehouse administration and maintenance** - performing activities that support the data warehouse end user, including dealing with technical issues, such as:
  - Providing security for the information contained in the data warehouse
  - Ensuring sufficient hard-drive space for the data warehouse content
  - Implementing the backup and recovery procedures

Copyright (c) 2014 Pearson Education, Inc.

Troy University Resources

THE NEXT VERSION OF THE DATA WAREHOUSE

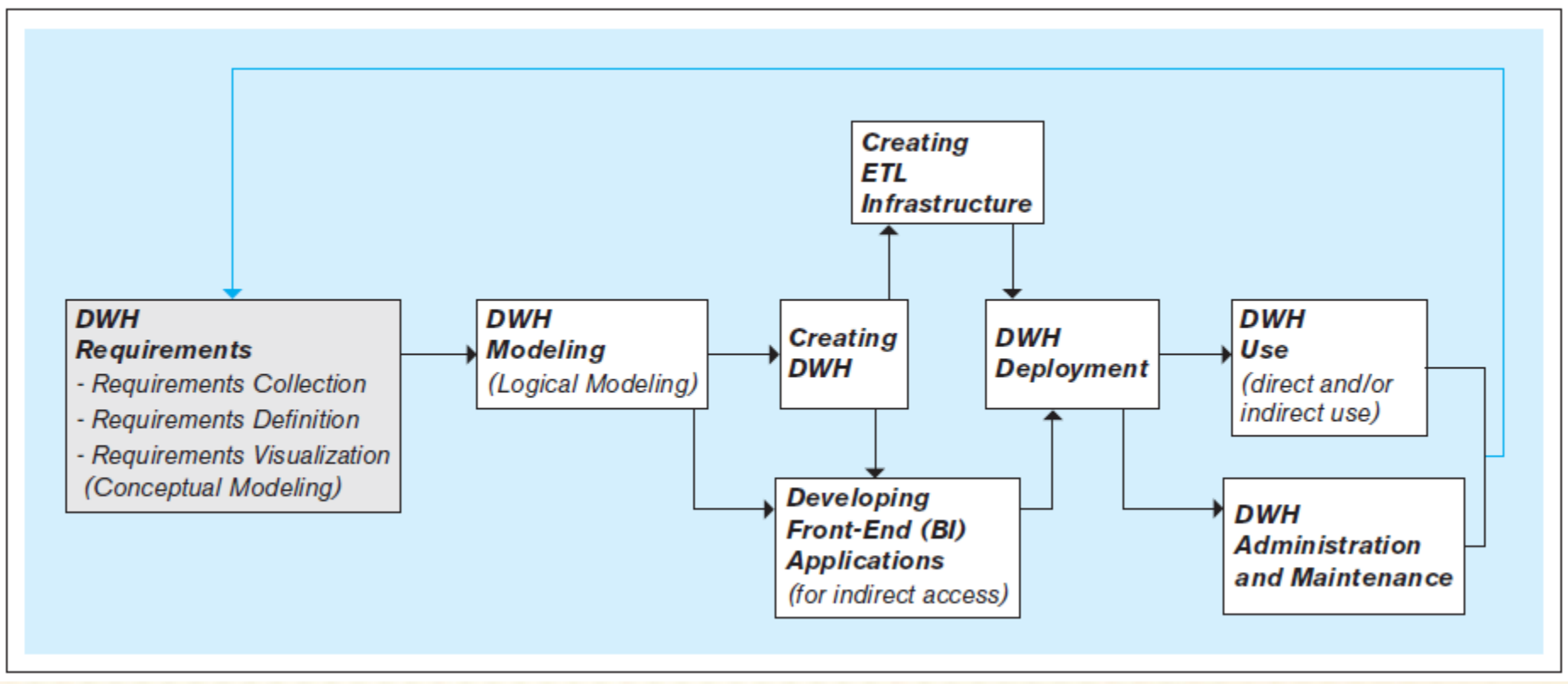
- The new version of the data warehouse follows the same development steps as the initial version

Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

In most cases, after a certain period of use, the need for modifications and expansion of the existing data warehouse becomes apparent, and the development of a new version of the existing data warehouse system is initiated.

THE NEXT VERSION OF THE DATA WAREHOUSE



Copyright (c) 2014 Pearson Education, Inc.

SPEAKER NOTES

The difference between developing the initial and subsequent versions of the data warehouse is discussed on Page 221.

Troy University Resources