CHAPTER 9:
DATA WAREHOUSING

Modern Database Management

12th Edition

Jeff Hoffer, Ramesh Venkataraman,
Heikki Topi

Copyright © 2016 Pearson Education, Inc.

OBJECTIVES

- Define terms
- Give reasons for information gap between information needs and availability
- List reasons for need of data warehousing
- Describe three levels of data warehouse architectures
- Describe two components of star schema
- Estimate fact table size
- Design a data mart
- Develop requirements for a data mart
- Understand future data warehousing trends

FINAL PROJECT REGISTRATION

- Link for group registration:

https://docs.google.com/spreadsheets/d/15VtCIA_GjxXpSVWFZkJrVla4yc6lMTIWHBURvBcCvBA/edit?usp=sharing

Copyright © 2016 Pearson Education, Inc.

DEFINITIONS

- **Data Warehouse**
  - A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes. Key terms are:
    - ***Subject-oriented:*** Organized on key objects such as customers, patients, students, products
    - ***Integrated:*** consistent naming conventions, formats, encoding structures; from multiple data sources
    - ***Time-variant:*** Contain time dimensions to study trends and changes
    - ***Non-updatable:*** read-only, periodically refreshed from operational systems, not from end-users
- **Data Mart**
  - A data warehouse that is limited in scope

DATA WAREHOUSING

- Is the process where organizations create and maintain data warehouse

- Extract meaning and form decision making from informational assets through these warehouses.

Copyright © 2016 Pearson Education, Inc.

NEED FOR DATA WAREHOUSING

- Integrated, company-wide view of high-quality information (from different databases)

- Separation of *operational* and *informational* systems and data (for improved performance)

Copyright © 2016 Pearson Education, Inc.

CONTENT OF A DATA WAREHOUSE

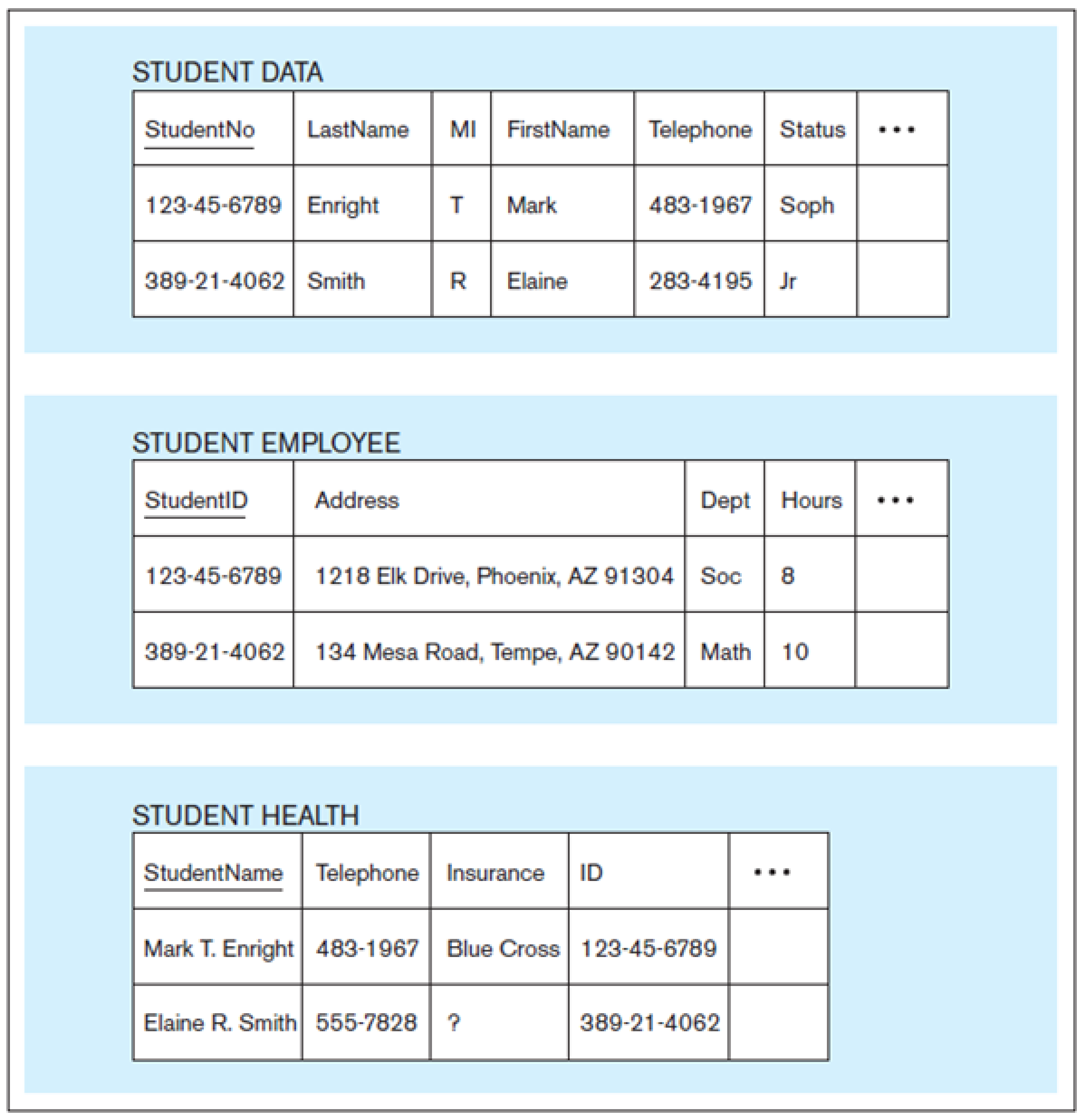
- Your data warehouse will store these types of data:
  - Historical data: Data is recorded throughout history
  - Derived data: Data is filtered and transformed to information
  - Metadata: Data that describe data and schema objects

ISSUES WITH COMPANY-WIDE VIEW (FIG 9-1)

- Inconsistent key structures: 1st and 2nd table contains number, the last contains string.
- Synonyms: StudentID and number is the same
- Free-form vs. structured fields:
  - In student health: StudentName consists of first/last name whereas in Student Data: name is broken into parts
- Inconsistent data values: Conflicts in Mr Smith phone numbers (using 1 or 2 number)
- Missing data: Insurance value is missing.

Copyright © 2016 Pearson Education, Inc.

Figure 9-1 Examples of heterogeneous data



STUDENT DATA

StudentNo	LastName	MI	FirstName	Telephone	Status	• • •
123-45-6789	Enright	T	Mark	483-1967	Soph	
389-21-4062	Smith	R	Elaine	283-4195	Jr	

STUDENT EMPLOYEE

StudentID	Address	Dept	Hours	• • •
123-45-6789	1218 Elk Drive, Phoenix, AZ 91304	Soc	8	
389-21-4062	134 Mesa Road, Tempe, AZ 90142	Math	10	

STUDENT HEALTH

StudentName	Telephone	Insurance	ID	• • •
Mark T. Enright	483-1967	Blue Cross	123-45-6789	
Elaine R. Smith	555-7828	?	389-21-4062	

Copyright © 2016 Pearson Education, Inc.

ORGANIZATIONAL TRENDS MOTIVATING DATA WAREHOUSES

- No single system of records
  - Split into several databases.
- Multiple systems not synchronized:
  - All data from separate system must be synced into additional database.
- Organizational need to analyze activities in a balanced way
  - Result must be consistent, data wh is necessary

CONT (2)

- Customer relationship management
  - To view overall picture of activity with customer across all touch points
- Supplier relationship management
  - To view overall picture of activity with supplier across all touch points, from billing, meeting, quality control, pricing and support

Copyright © 2016 Pearson Education, Inc.

SEPARATING OPERATIONAL AND INFORMATIONAL SYSTEMS

- **Operational system** – a system that is used to run a business in real time, based on current data; also called a system of record
  - For example: Reservation system, sales order processing systems, …

CONT

- **Informational system** – a system designed to support decision making based on historical point-in-time and prediction data for complex queries or data-mining applications
  - For example: sale trends analysis, human resource planning

Copyright © 2016 Pearson Education, Inc.

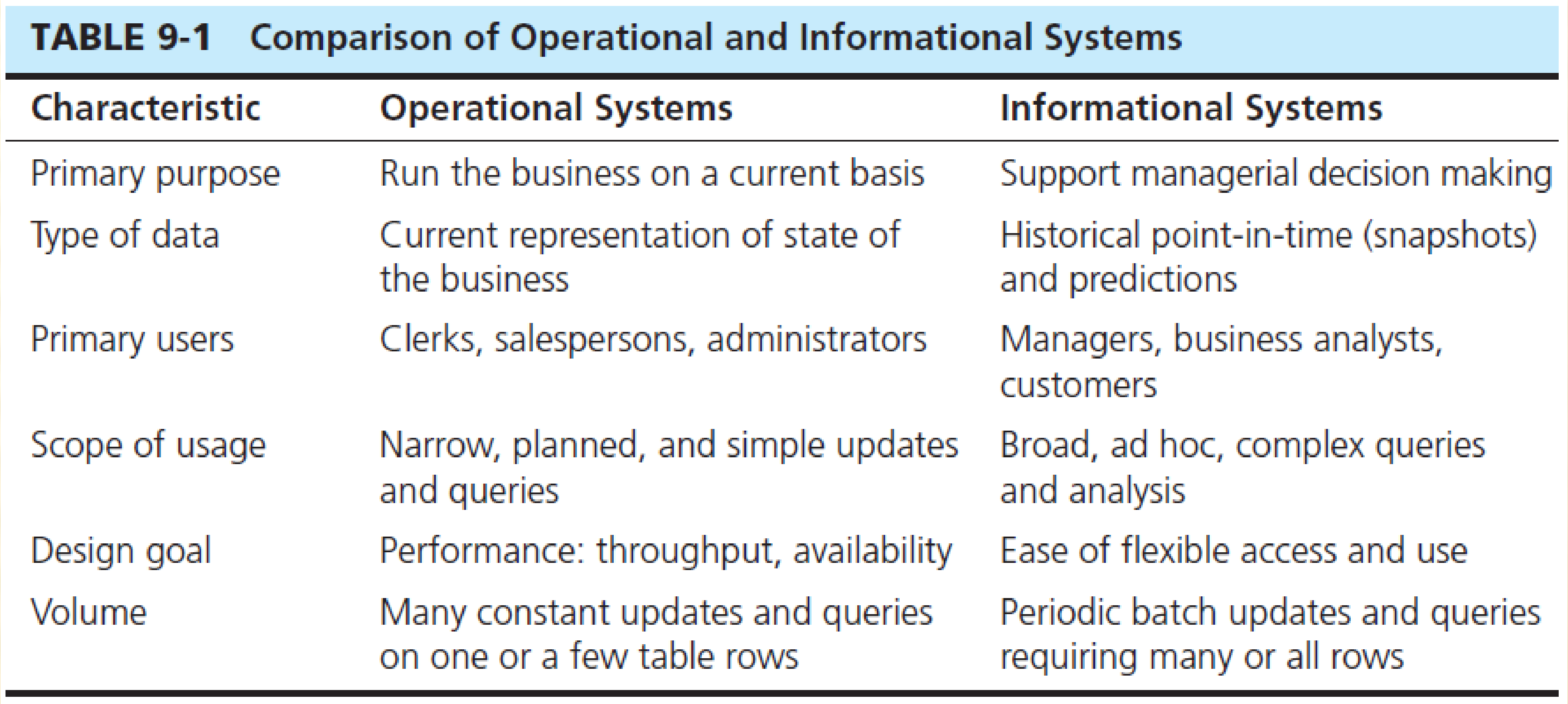


TABLE 9-1 Comparison of Operational and Informational Systems

Characteristic	Operational Systems	Informational Systems
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance: throughput, availability	Ease of flexible access and use
Volume	Many constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

THE NEED TO SEPARATE BETWEEN INFOR. AND OP. SYSTEM

- A data warehouse centralizes data that are scattered throughout disparate operational systems and makes them readily available for decision support applications.
- A properly designed data warehouse adds value to data by improving their quality and consistency.
- A separate data warehouse eliminates much of the contention for resources that results when informational applications are confounded with operational processing.

Copyright © 2016 Pearson Education, Inc.

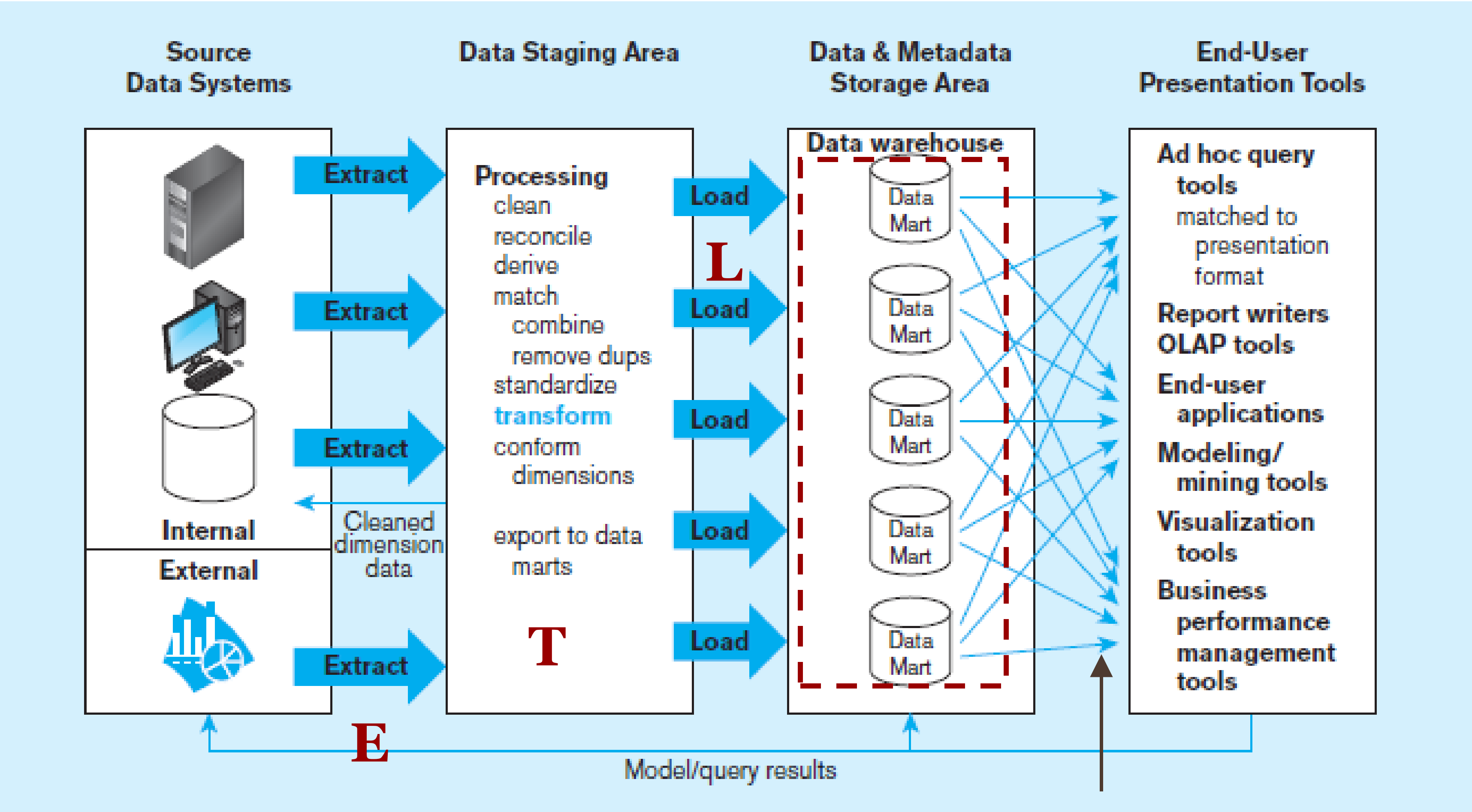
DATA WAREHOUSE ARCHITECTURES

- Independent Data Mart
- Dependent Data Mart and Operational Data Store
- Logical Data Mart and Real-Time Data Warehouse
- Three-Layer architecture

All involve some form of ***extract***, ***transform*** and ***load*** **(ETL)**

Figure 9-2 Independent data mart data warehousing architecture

Data marts:
Mini-warehouses, limited in scope



Separate ETL for each *independent* data mart

Data access complexity due to ***multiple*** data marts

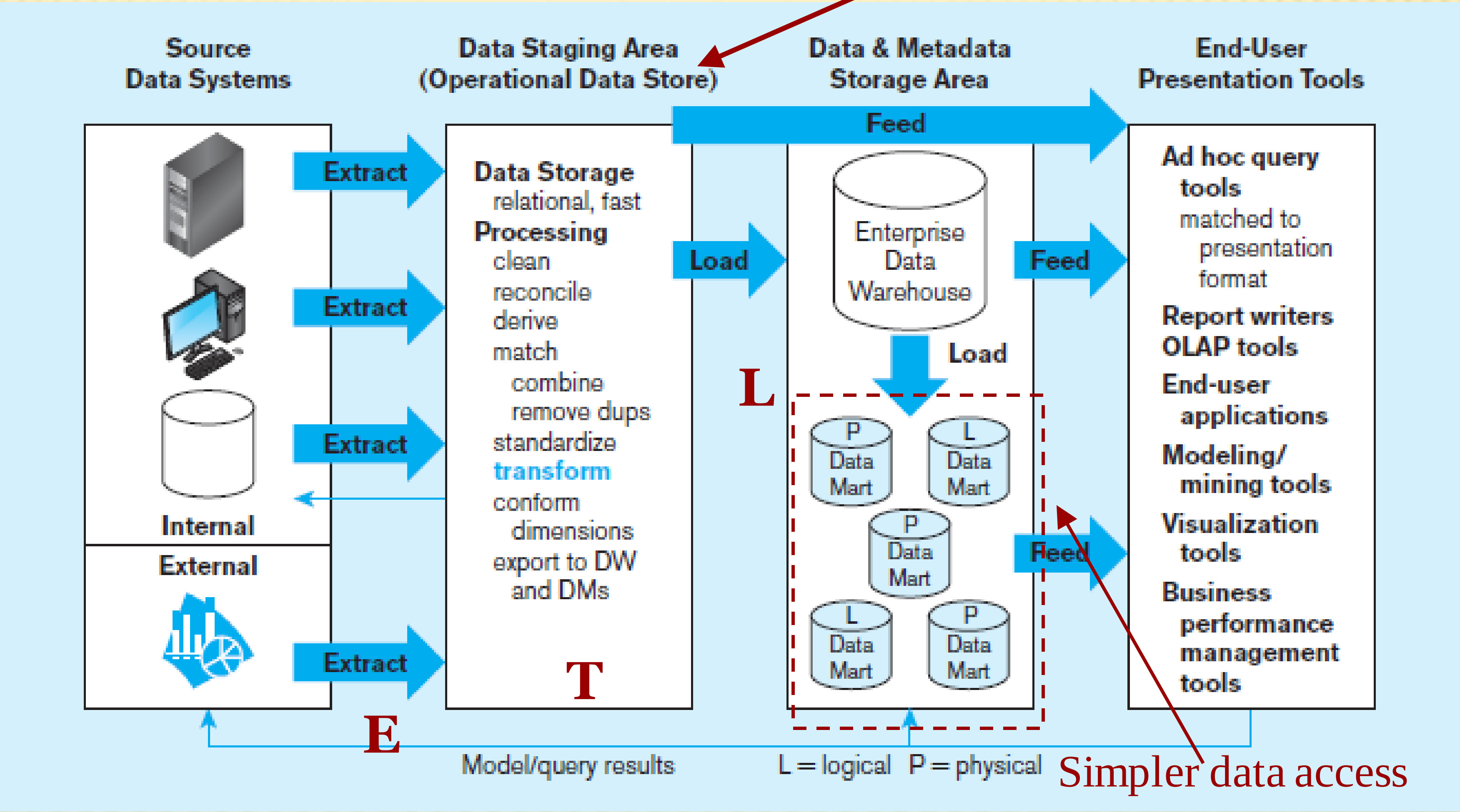
Copyright © 2016 Pearson Education, Inc.

LIMITATIONS OF INDEPENDENT DATA MARTS

- Separate ETL process for each data mart
  redundant data and processing
- Inconsistency between data marts
- Difficult to drill down for related facts between data marts, analysis is limited
- Excessive scaling costs are more applications are built since add new data mart is costly, repeat ETL process.
- High cost for obtaining consistency between marts

Figure 9-3 Dependent data mart with operational data store: a three-level architecture

ODS provides option for obtaining ***current*** data



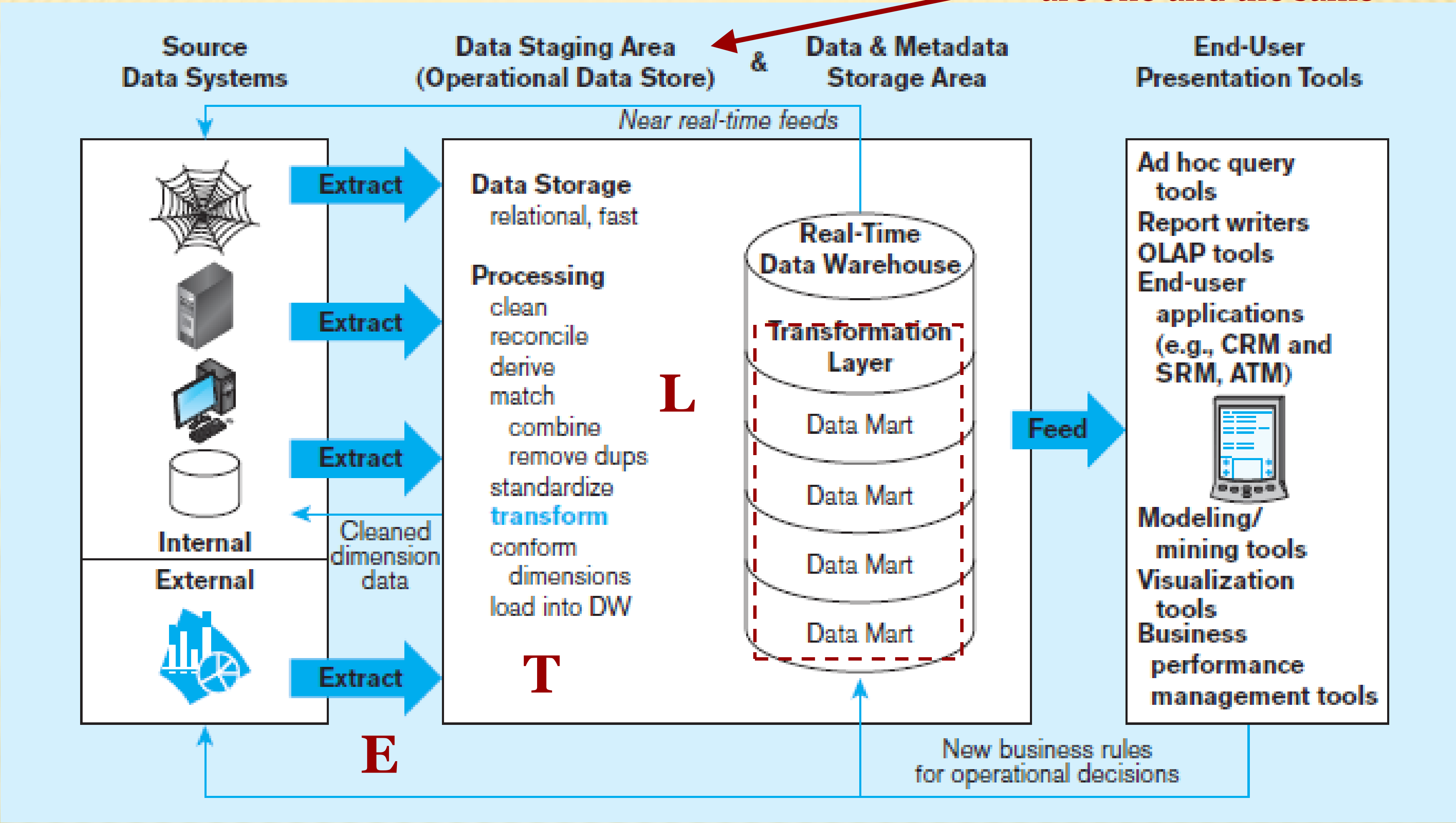
Simpler data access

Single ETL for ***enterprise data warehouse (EDW)***

Dependent data marts loaded from EDW

Figure 9-4 Logical data mart and real time warehouse architecture

ODS and **data warehouse** are one and the same



Near real-time ETL for ***Data Warehouse***

Data marts are NOT separate databases, but logical ***views*** of the data warehouse

- Easier to create new data marts

Copyright © 2016 Pearson Education, Inc.

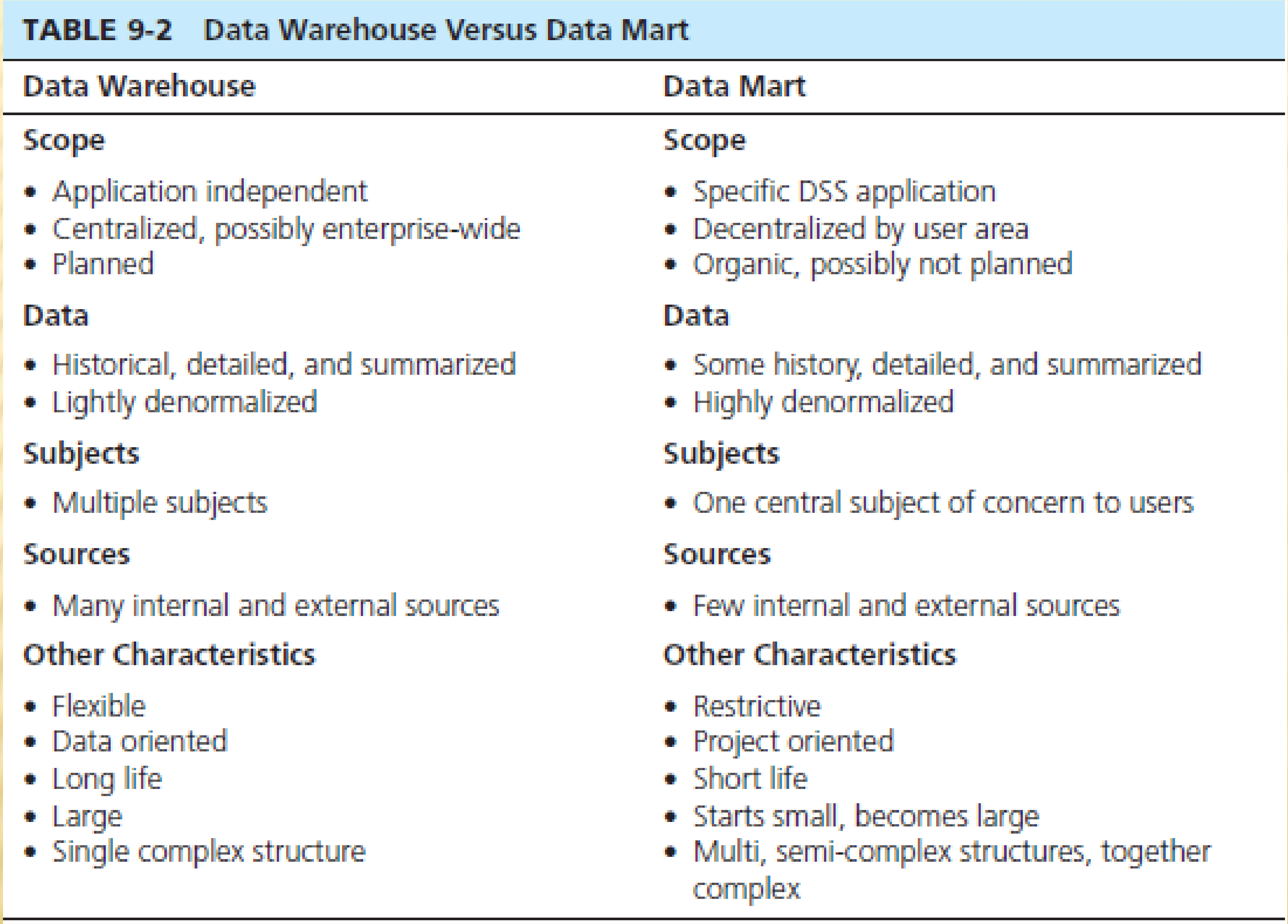
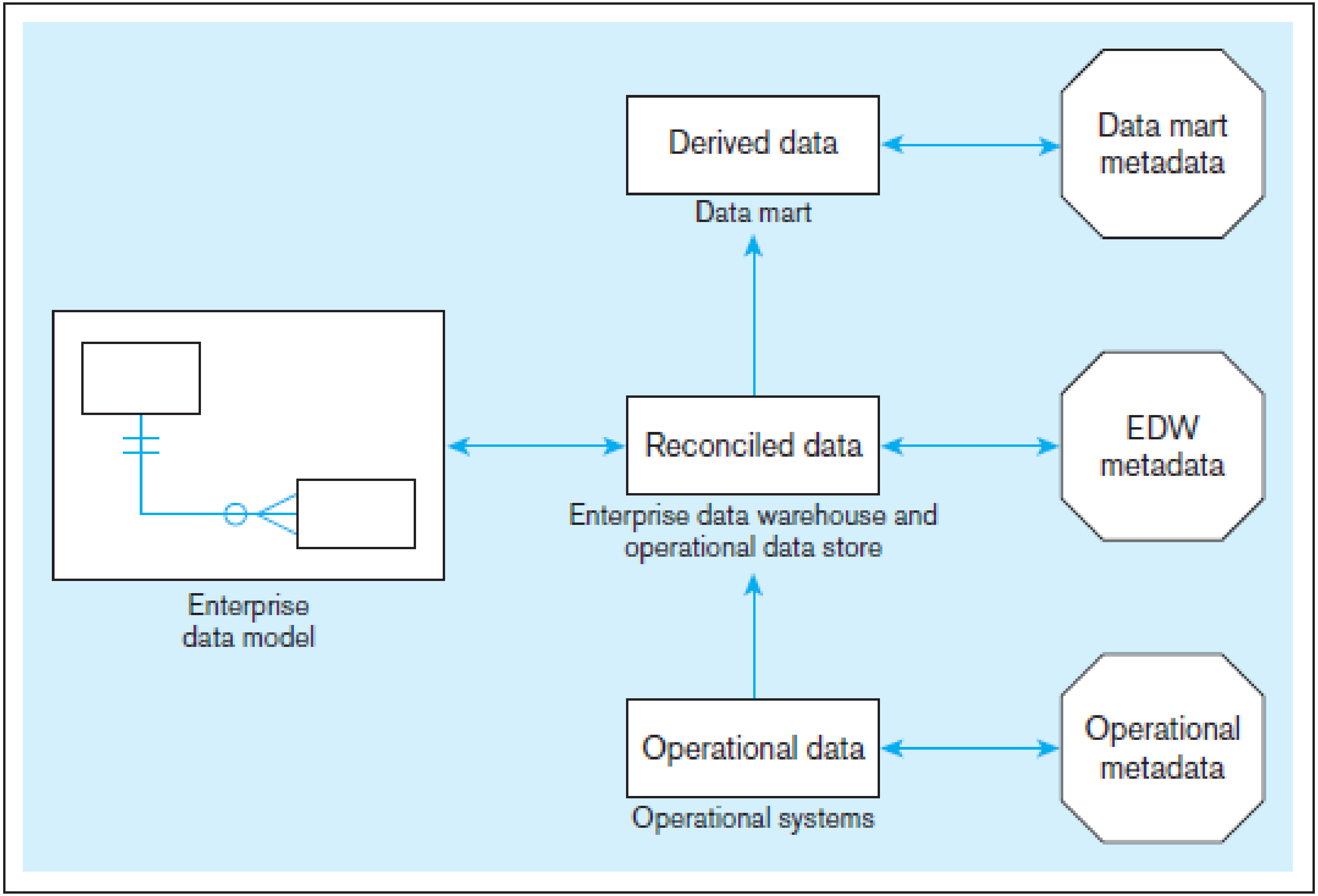


TABLE 9-2 Data Warehouse Versus Data Mart

Data Warehouse	Data Mart
Scope	**Scope**
• Application independent • Centralized, possibly enterprise-wide • Planned	• Specific DSS application • Decentralized by user area • Organic, possibly not planned
Data	**Data**
• Historical, detailed, and summarized • Lightly denormalized	• Some history, detailed, and summarized • Highly denormalized
Subjects	**Subjects**
• Multiple subjects	• One central subject of concern to users
Sources	**Sources**
• Many internal and external sources	• Few internal and external sources
Other Characteristics	**Other Characteristics**
• Flexible • Data oriented • Long life • Large • Single complex structure	• Restrictive • Project oriented • Short life • Starts small, becomes large • Multi, semi-complex structures, together complex

Figure 9-5 Three-layer data architecture for a data warehouse



DATA CHARACTERISTICS STATUS VS. EVENT DATA

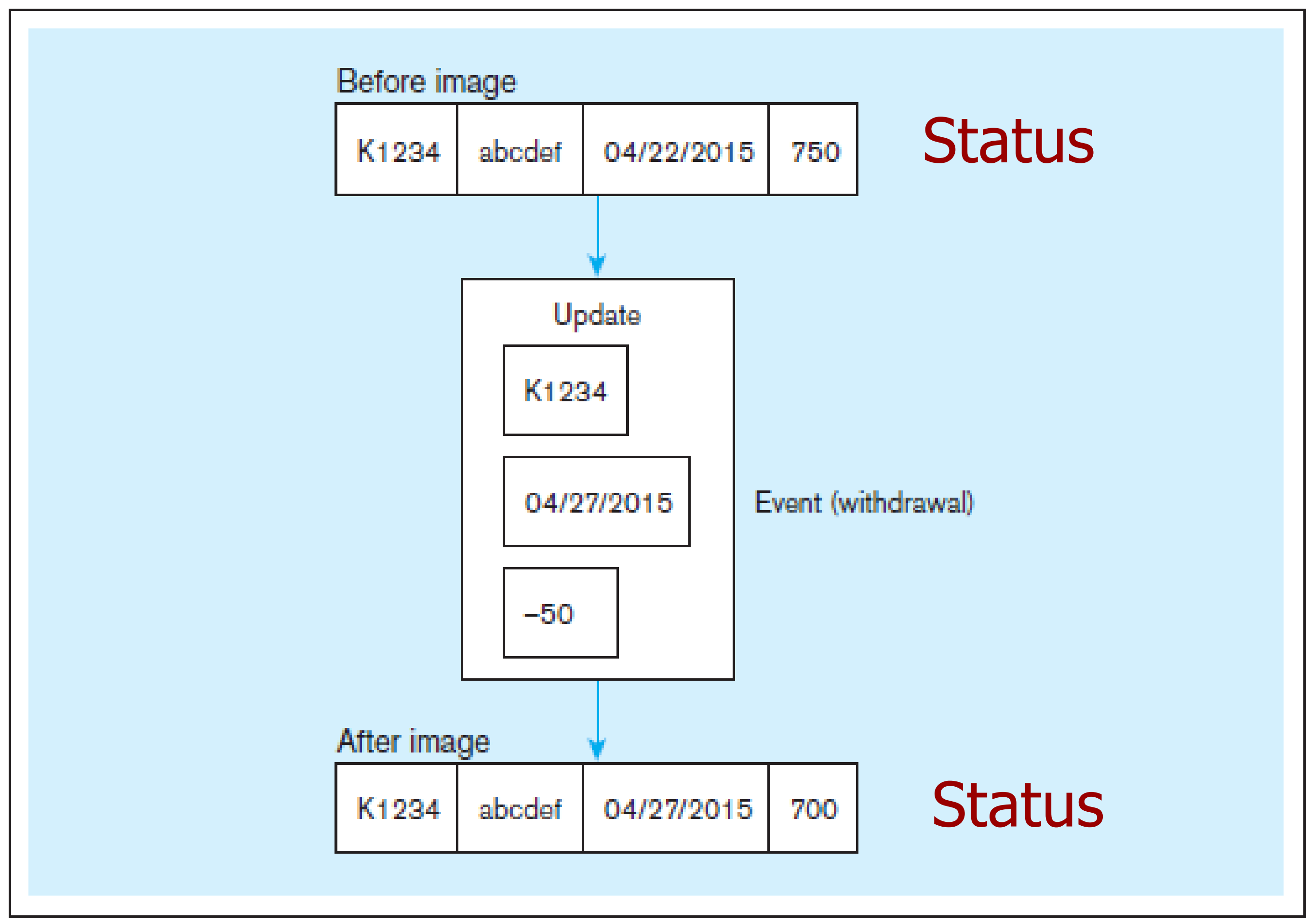


Figure 9-6 Example of DBMS log entry

Event = a database action (create/ update/ delete) that results from a transaction

Copyright © 2016 Pearson Education, Inc.

DATA CHARACTERISTICS TRANSIENT(TẠM THỜI) VS. PERIODIC (ĐỊNH KÌ) DATA

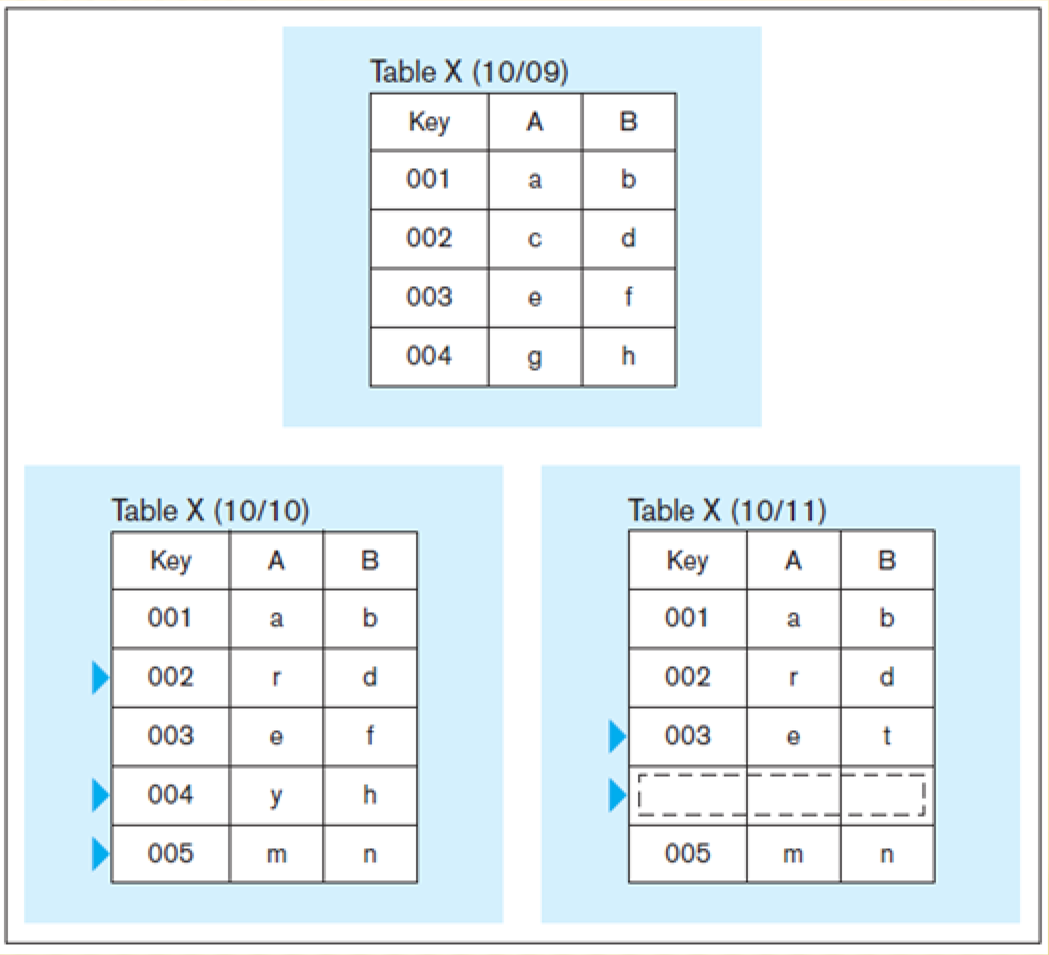


Table X (10/09)

Key	A	B
001	a	b
002	c	d
003	e	f
004	g	h

Table X (10/10)

Key	A	B
001	a	b
002	r	d
003	e	f
004	y	h
005	m	n

Table X (10/11)

Key	A	B
001	a	b
002	r	d
003	e	t
005	m	n

Figure 9-7
Transient operational data

With transient data, changes to existing records are written over previous records, thus destroying the previous data content.

DATA CHARACTERISTICS TRANSIENT VS. PERIODIC

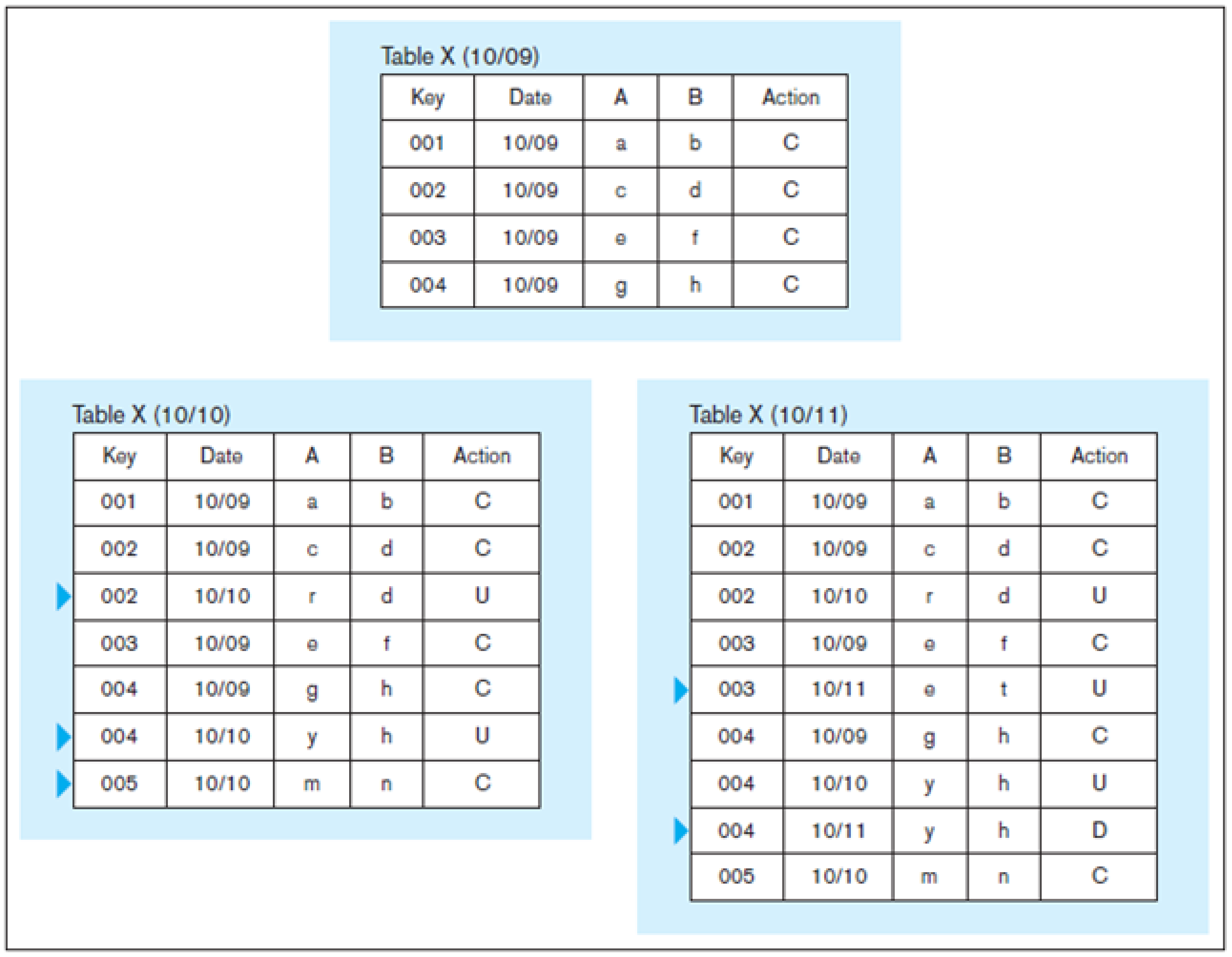


Table X (10/09)

Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
003	10/09	e	f	C
004	10/09	g	h	C

Table X (10/10)

Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
▶ 002	10/10	r	d	U
003	10/09	e	f	C
004	10/09	g	h	C
▶ 004	10/10	y	h	U
▶ 005	10/10	m	n	C

Table X (10/11)

Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
002	10/10	r	d	U
003	10/09	e	f	C
▶ 003	10/11	e	t	U
004	10/09	g	h	C
004	10/10	y	h	U
▶ 004	10/11	y	h	D
005	10/10	m	n	C

Figure 9-8 Periodic warehouse data

Periodic data are never physically altered or deleted once they have been added to the store.

Copyright © 2016 Pearson Education, Inc.

OTHER DATA WAREHOUSE CHANGES NEED TO BE ACCOMMODATED

- New descriptive attributes
- New business activity attributes
- New classes of descriptive attributes = new table
- Descriptive attributes become more refined
- Descriptive data are related to one another
- New source of data

Copyright © 2016 Pearson Education, Inc.

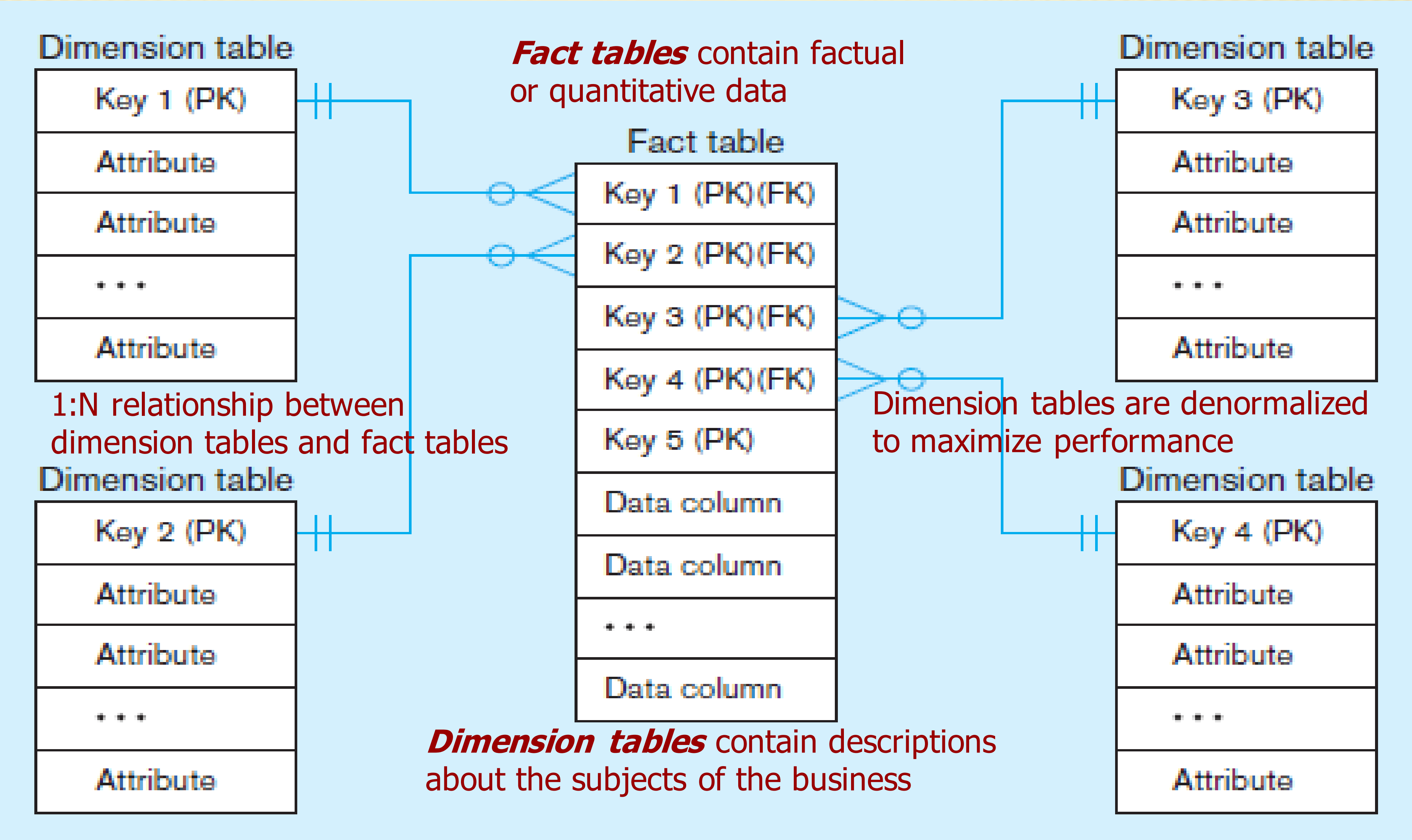
DERIVED DATA

- Objectives
  - Ease of use for decision support applications
  - Fast response to predefined user queries
  - Customized data for particular target audiences
  - Ad-hoc query support
  - Data mining capabilities
- Characteristics
  - Detailed (mostly periodic) data
  - Aggregate (for summary)
  - Distributed (to departmental servers)

Most common data model = **dimensional model** (usually implemented as a **star schema**)

Copyright © 2016 Pearson Education, Inc.

Figure 9-9 Components of a **star schema**



Fact tables contain factual or quantitative data

Dimension tables contain descriptions about the subjects of the business

1:N relationship between dimension tables and fact tables

Dimension tables are denormalized to maximize performance

Dimension table
Key 1 (PK)
Attribute
Attribute
• • •
Attribute

Dimension table
Key 2 (PK)
Attribute
Attribute
• • •
Attribute

Fact table
Key 1 (PK)(FK)
Key 2 (PK)(FK)
Key 3 (PK)(FK)
Key 4 (PK)(FK)
Key 5 (PK)
Data column
Data column
• • •
Data column

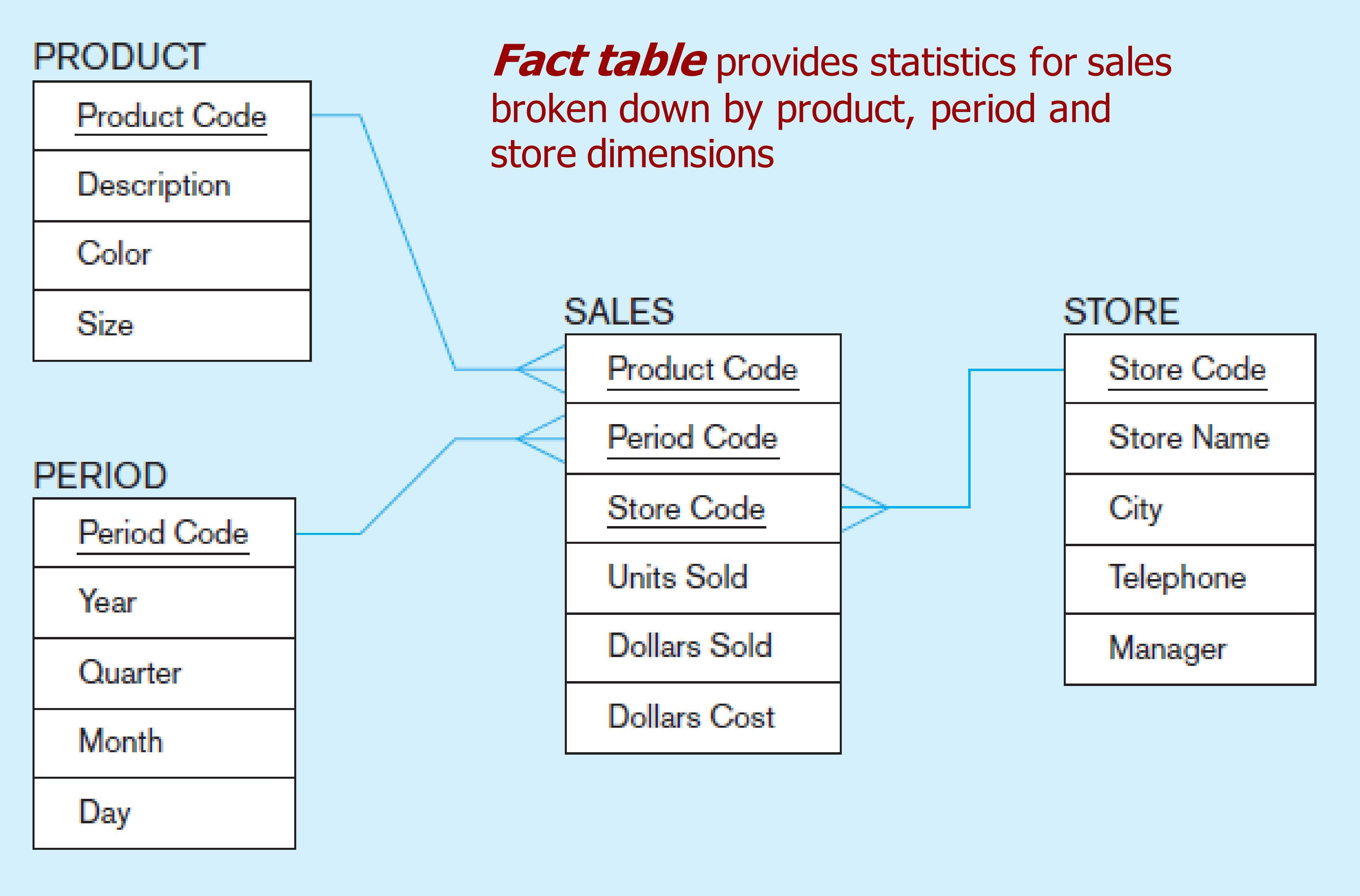
Dimension table
Key 3 (PK)
Attribute
Attribute
• • •
Attribute

Dimension table
Key 4 (PK)
Attribute
Attribute
• • •
Attribute

Excellent for ad-hoc queries, but bad for online transaction processing

Copyright © 2016 Pearson Education, Inc.

Figure 9-10 Star schema example (data recored daily)



PRODUCT

Product Code

Description

Color

Size

PERIOD

Period Code

Year

Quarter

Month

Day

SALES

Product Code

Period Code

Store Code

Units Sold

Dollars Sold

Dollars Cost

STORE

Store Code

Store Name

City

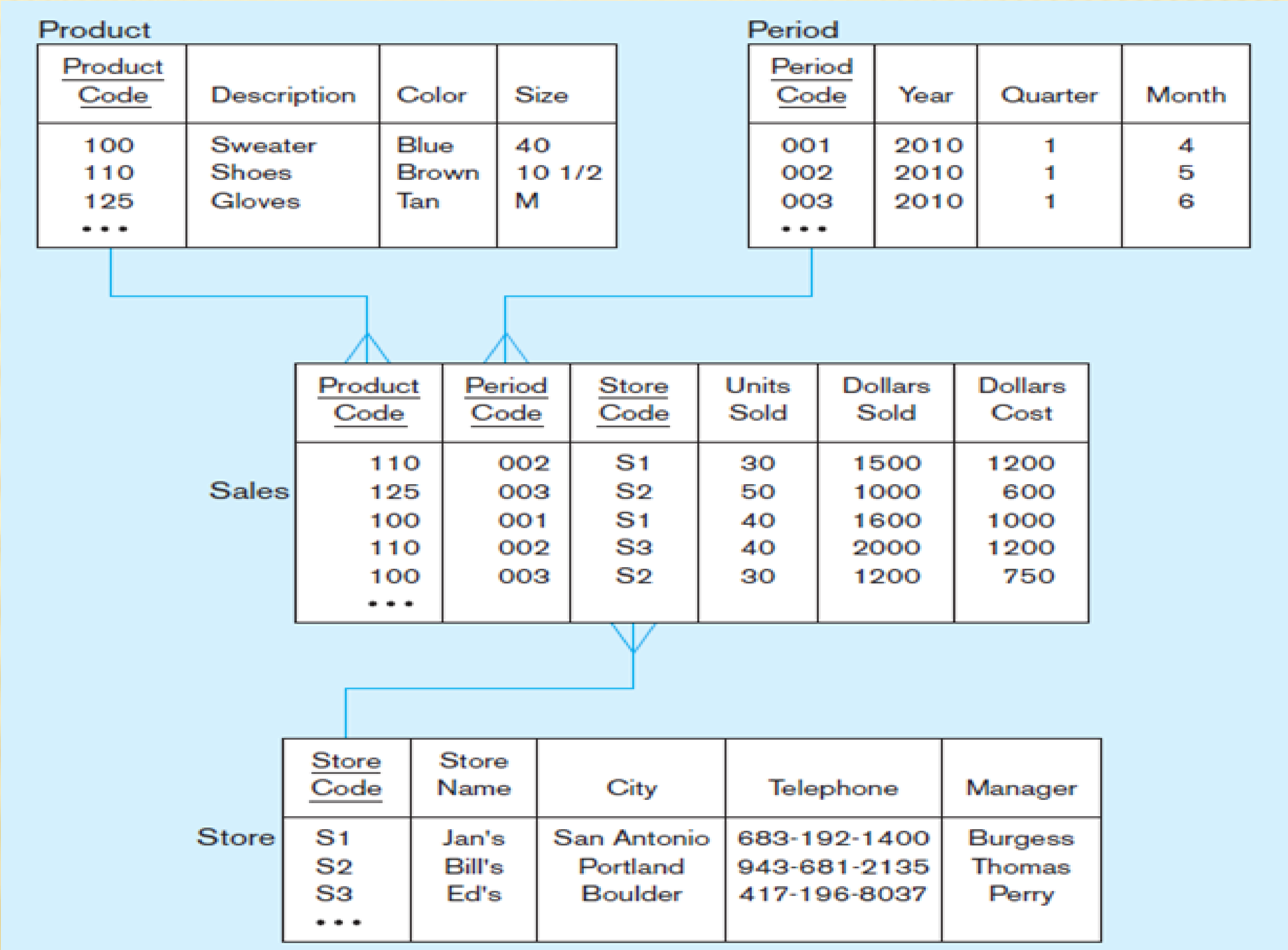
Telephone

Manager

Fact table provides statistics for sales broken down by product, period and store dimensions

Copyright © 2016 Pearson Education, Inc.

Figure 9-11 Star schema with sample data



Product

Product Code	Description	Color	Size
100	Sweater	Blue	40
110	Shoes	Brown	10 1/2
125	Gloves	Tan	M
• • •			

Period

Period Code	Year	Quarter	Month
001	2010	1	4
002	2010	1	5
003	2010	1	6
• • •			

Sales

Product Code	Period Code	Store Code	Units Sold	Dollars Sold	Dollars Cost
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750
• • •					

Store

Store Code	Store Name	City	Telephone	Manager
S1	Jan's	San Antonio	683-192-1400	Burgess
S2	Bill's	Portland	943-681-2135	Thomas
S3	Ed's	Boulder	417-196-8037	Perry
• • •				

Copyright © 2016 Pearson Education, Inc.

SURROGATE (REPRESENTATIVE) KEYS

- Dimension table keys should be ***surrogate*** (non-intelligent and non-business related), because:
  - Business keys may change over time
  - Helps keep track of nonkey attribute values for a given production key
  - Surrogate keys are simpler and shorter
  - Surrogate keys can be same length and format for all key

Copyright © 2016 Pearson Education, Inc.

GRAIN OF THE FACT TABLE

- Granularity of Fact Table: level of detail in the fact table
  - Transactional grain–finest level
  - Aggregated grain–more summarized
  - Finer grains better ***market basket analysis*** capability
  - Finer grain more dimension tables, more rows in fact table

DURATION OF THE DATABASE

- Amount of history to be kept on database
- Natural duration–13 months or 5 quarters
- Financial institutions may need longer duration
- Older data is more difficult to source and cleanse

Copyright © 2016 Pearson Education, Inc.

SIZE OF FACT TABLE

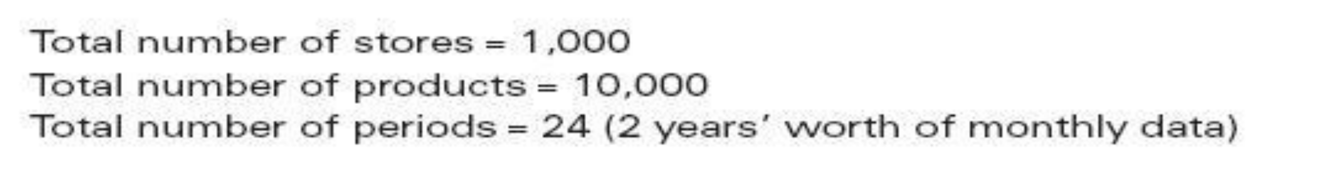
- Depends on the number of dimensions and the grain of the fact table

- Number of rows = product of number of possible values for each dimension associated with the fact table

Copyright © 2016 Pearson Education, Inc.

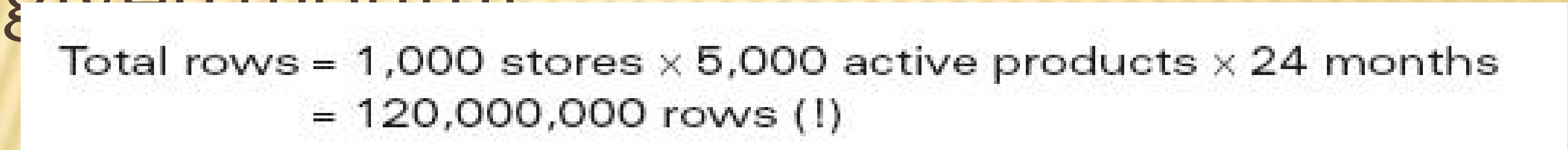
SIZE OF FACT TABLE

- Example: Assume the following for Figure 9-11:



Total number of stores = 1,000
Total number of products = 10,000
Total number of periods = 24 (2 years' worth of monthly data)

- Total rows calculated as follows (assuming only half the products record sales for a given month):

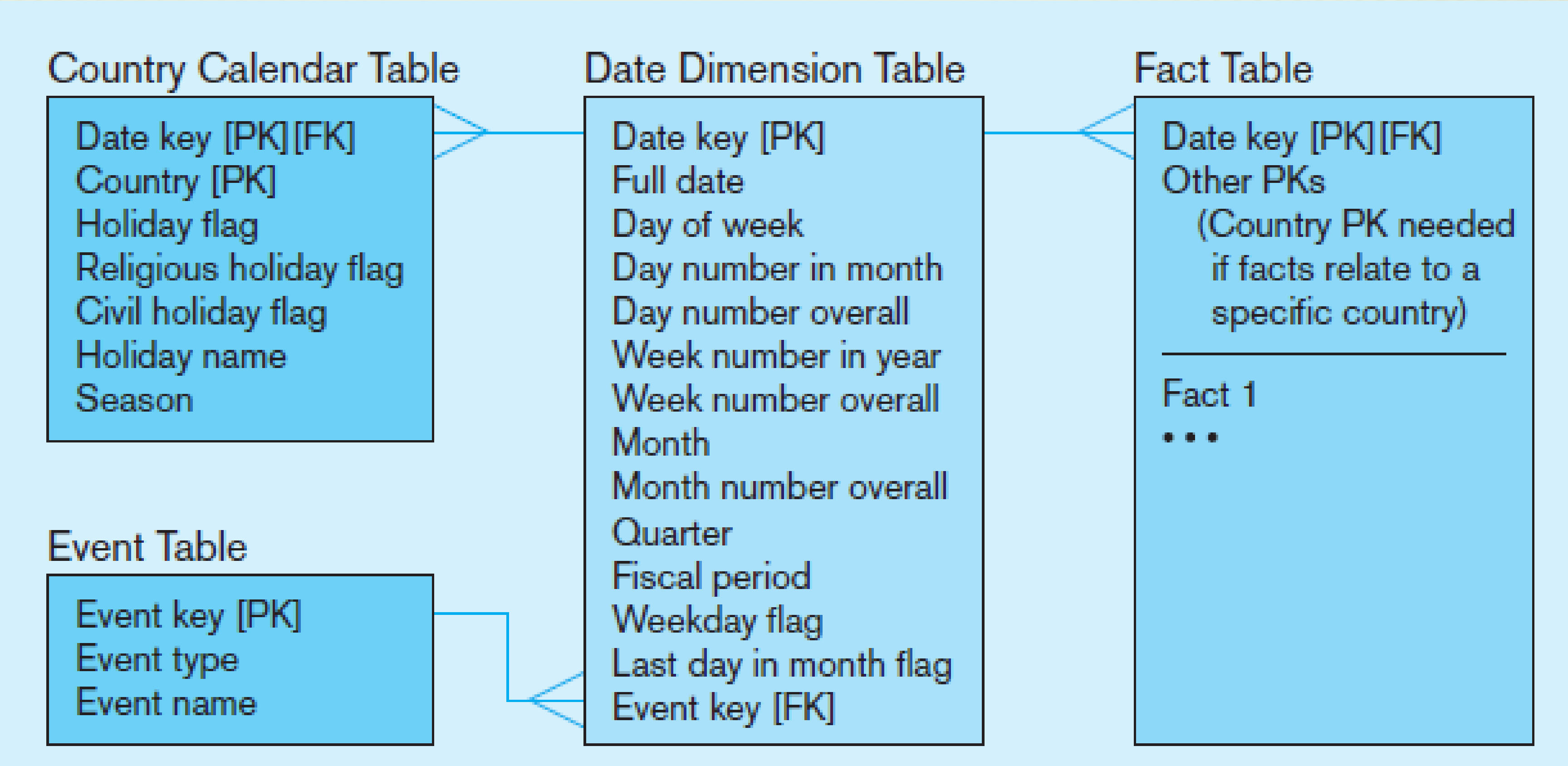


Total rows = 1,000 stores × 5,000 active products × 24 months
= 120,000,000 rows (!)

- If fact table contains 6 fields, each of 4 bytes

=> 120.000 krows * 6 * 4 = 2.88GB of data.

Figure 9-12 Modeling dates and time



Country Calendar Table

Date key [PK][FK]
Country [PK]
Holiday flag
Religious holiday flag
Civil holiday flag
Holiday name
Season

Date Dimension Table

Date key [PK]
Full date
Day of week
Day number in month
Day number overall
Week number in year
Week number overall
Month
Month number overall
Quarter
Fiscal period
Weekday flag
Last day in month flag
Event key [FK]

Fact Table

Date key [PK][FK]
Other PKs
(Country PK needed if facts relate to a specific country)

Fact 1
• • •

Event Table

Event key [PK]
Event type
Event name

Fact tables contain time-period data

☐ Date dimensions are important

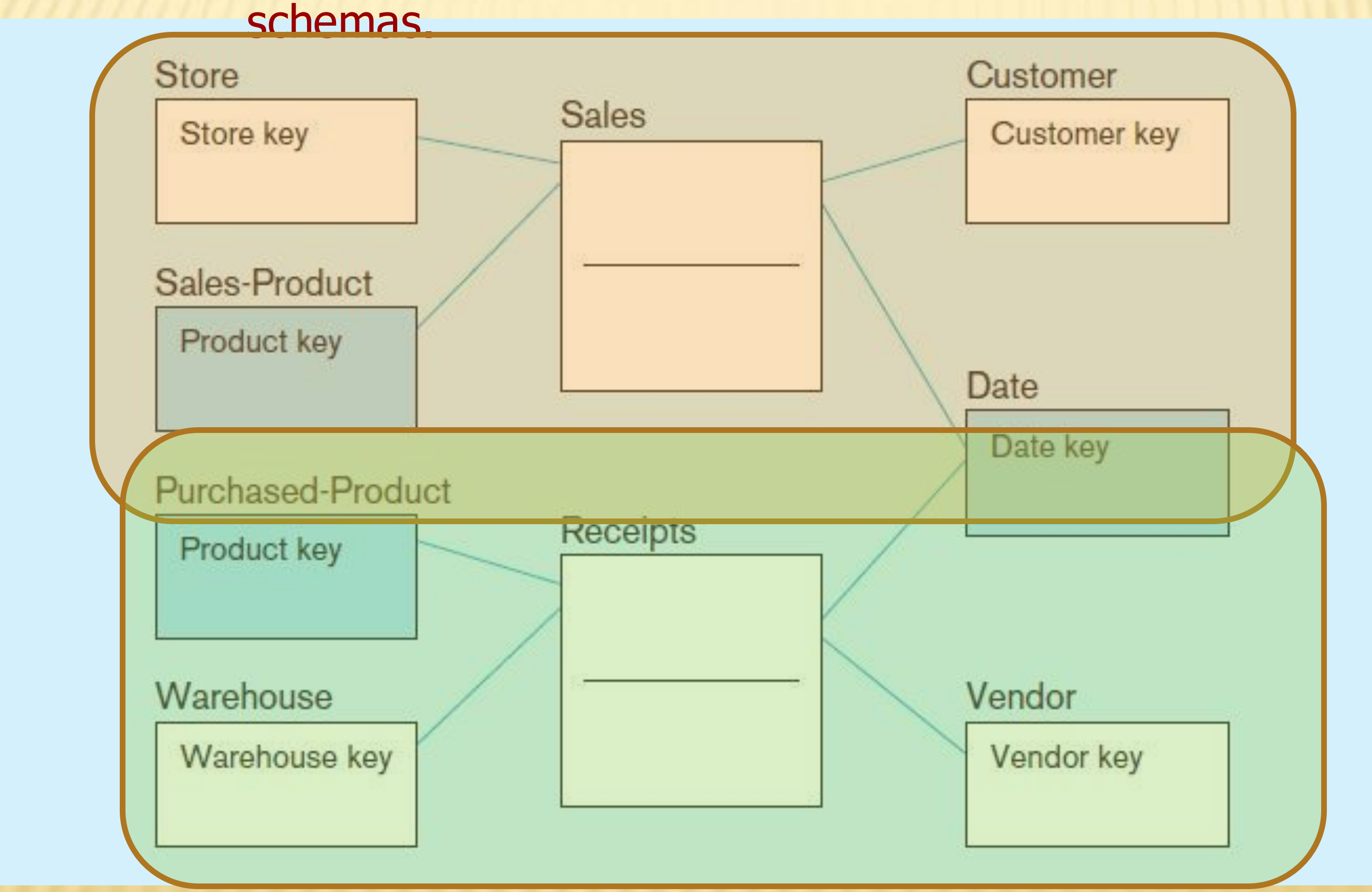
Copyright © 2016 Pearson Education, Inc.

VARIATIONS OF THE STAR SCHEMA

- Multiple Facts Tables
  - Can improve performance
  - Often used to store facts for different combinations of dimensions
  - Conformed dimensions
- Factless Facts Tables
  - No nonkey data, but foreign keys for associated dimensions
  - Used for:
    - Tracking events
    - Inventory coverage

Figure 9-13 Conformed dimensions

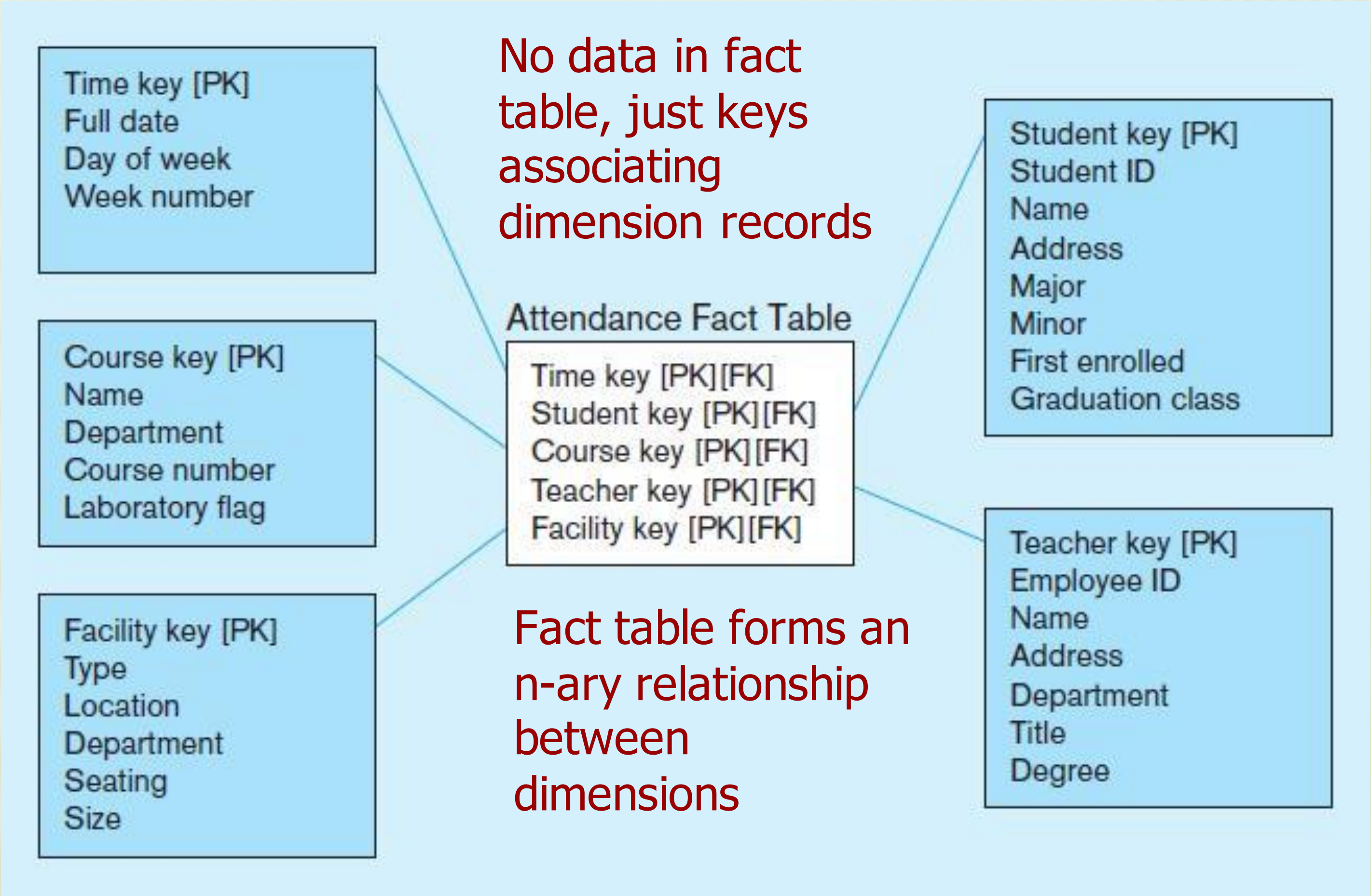
Two fact tables □ two (connected) star schemas.



Conformed dimension
Associated with multiple fact tables, here, date & product key

Copyright © 2016 Pearson Education, Inc.

Figure 9-14a Factless fact table showing occurrence of an event



Time key [PK]
Full date
Day of week
Week number

Course key [PK]
Name
Department
Course number
Laboratory flag

Facility key [PK]
Type
Location
Department
Seating
Size

No data in fact table, just keys associating dimension records

Attendance Fact Table

Time key [PK][FK]
Student key [PK][FK]
Course key [PK][FK]
Teacher key [PK][FK]
Facility key [PK][FK]

Fact table forms an n-ary relationship between dimensions

Student key [PK]
Student ID
Name
Address
Major
Minor
First enrolled
Graduation class

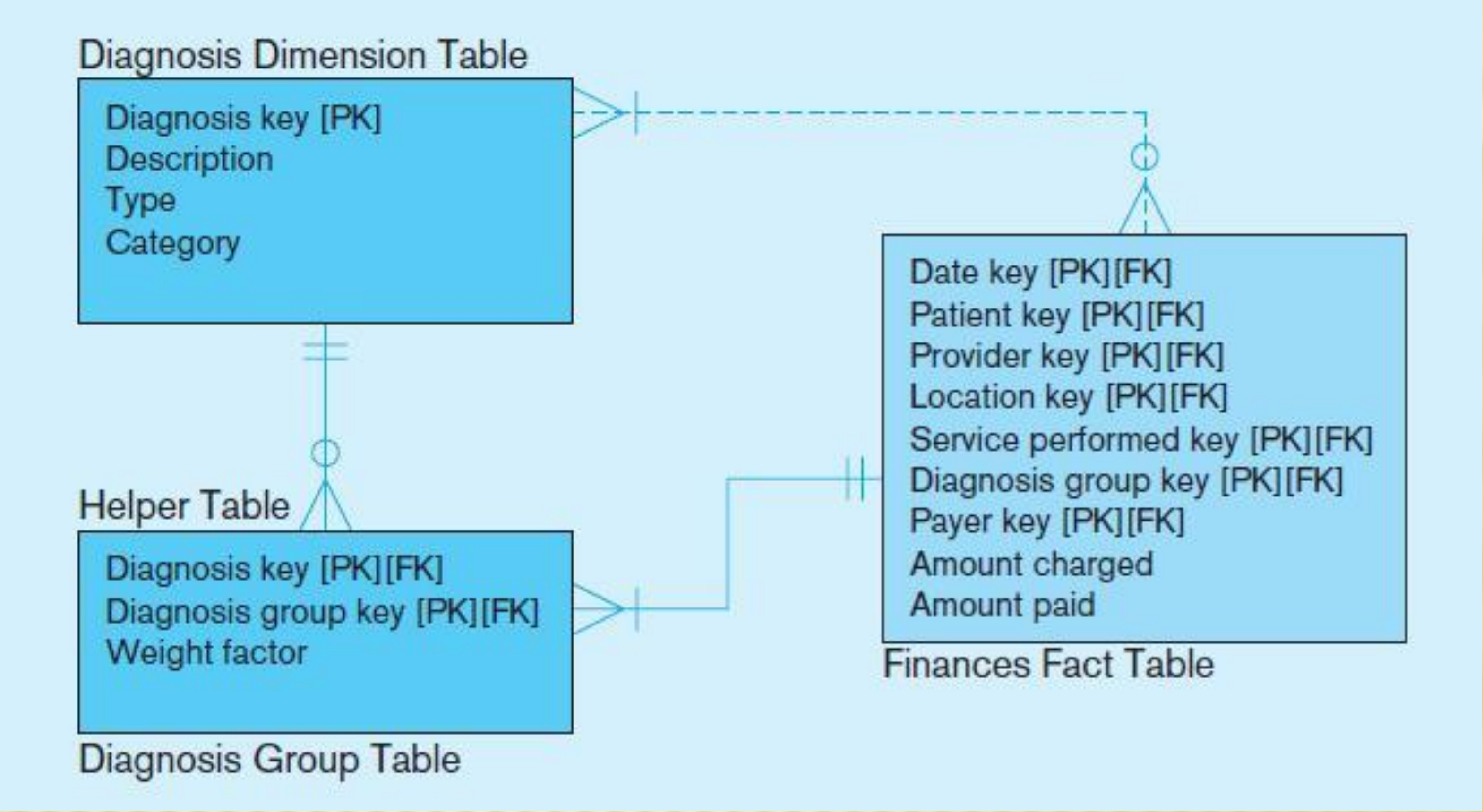
Teacher key [PK]
Employee ID
Name
Address
Department
Title
Degree

NORMALIZING DIMENSION TABLES

- Multivalued Dimensions
  - Facts qualified by a set of values for the same business subject
  - Normalization involves creating a table for an associative entity between dimensions
- Hierarchies
  - Sometimes a dimension forms a natural, fixed depth hierarchy
  - Design options
    - Include all information for each level in a single denormalized table
    - Normalize the dimension into a nested set of 1:M table relationships

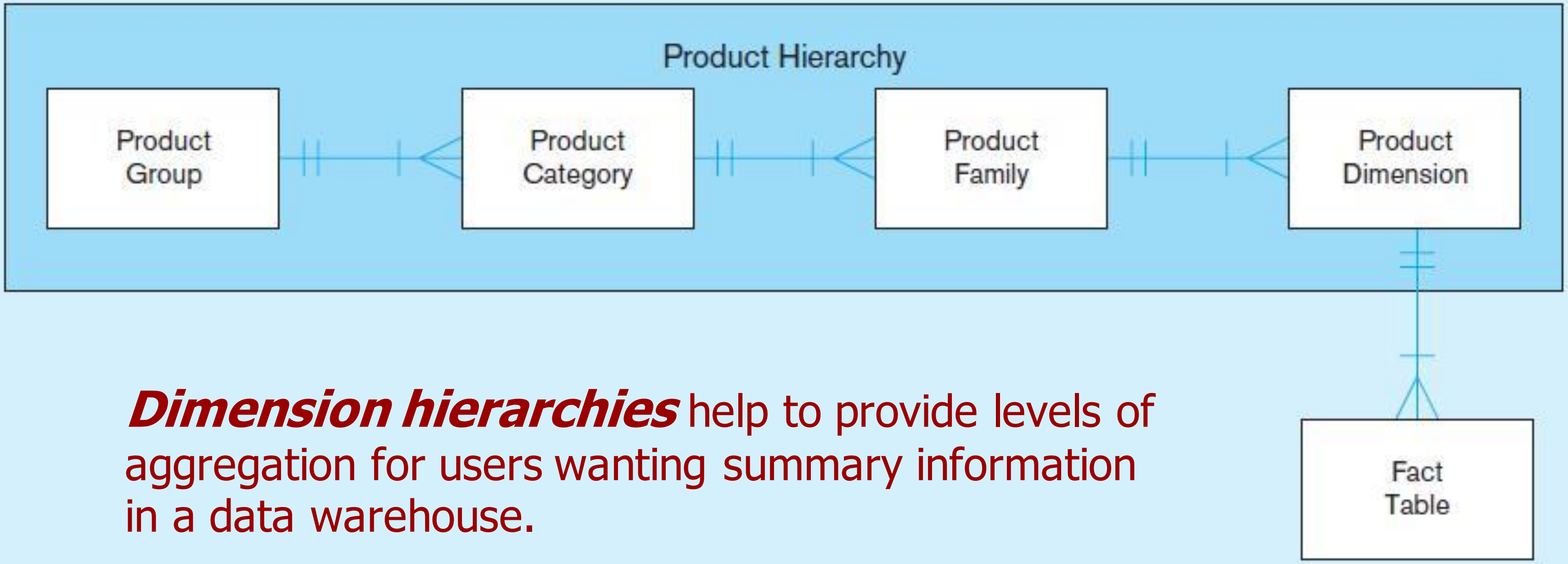
Copyright © 2016 Pearson Education, Inc.

Figure 9-15 Multivalued dimension



Helper table is an associative entity that implements a M:N relationship between dimension and fact.

Figure 9-16 Fixed product hierarchy



Dimension hierarchies help to provide levels of aggregation for users wanting summary information in a data warehouse.

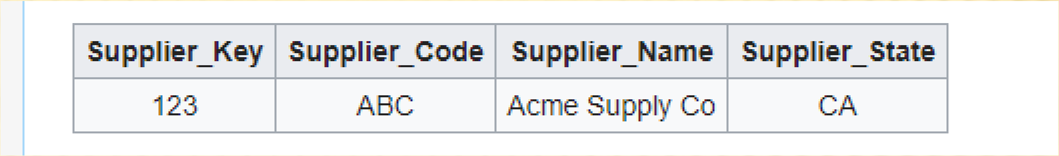
Dimension tables are normalized into several related tables

SLOWLY CHANGING DIMENSIONS (SCD)

- How to maintain knowledge of the past
- Kimball’s approaches:
  - Type 1: just replace old data with new (lose historical data)
  - Type 2: for each changing attribute, create a current value field and several old-valued fields (multivalued)
  - Type 4: create a new dimension table row each time the dimension object changes, with all dimension characteristics at the time of change. Most common approach

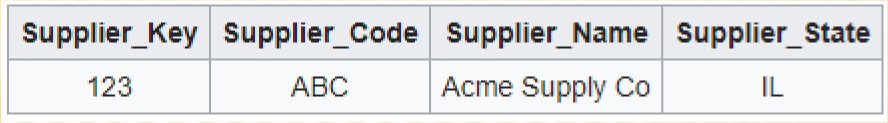
TYPE 1

- Consider [Supplier] table



Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

- When apply Type 1

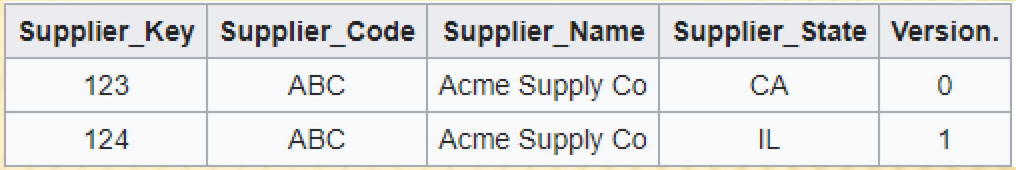


Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

- No history changes tracking

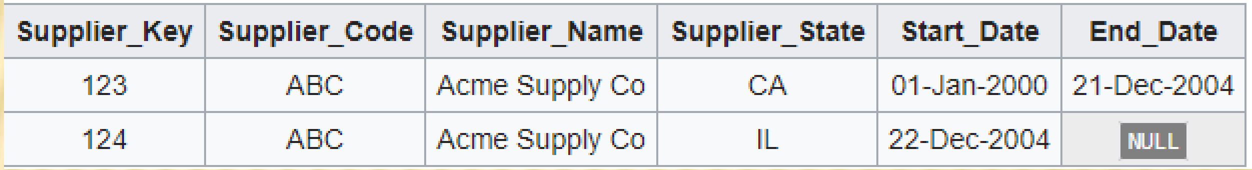
TYPE 2

- Add new version column



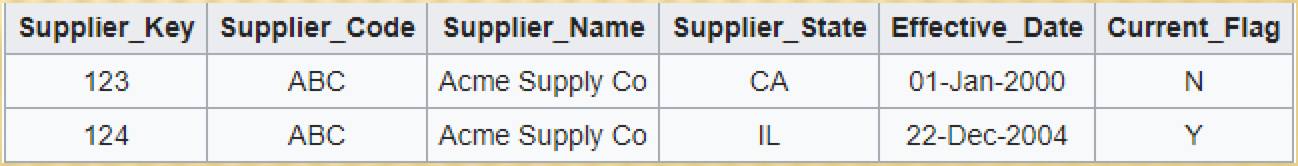
Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Version.
123	ABC	Acme Supply Co	CA	0
124	ABC	Acme Supply Co	IL	1

- Or add start/end date column’



Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
123	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004
124	ABC	Acme Supply Co	IL	22-Dec-2004	NULL

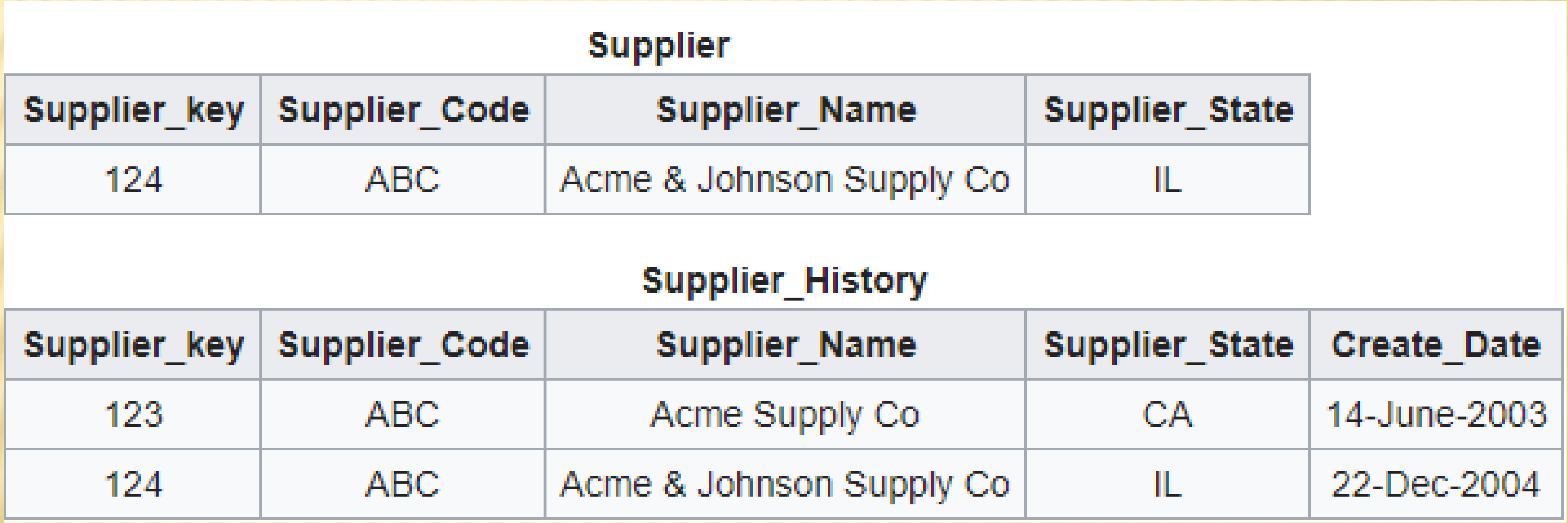
- Or add date with flag (Y: current version)



Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Effective_Date	Current_Flag
123	ABC	Acme Supply Co	CA	01-Jan-2000	N
124	ABC	Acme Supply Co	IL	22-Dec-2004	Y

TYPE 4:

- Add new history log table



Supplier

Supplier_key	Supplier_Code	Supplier_Name	Supplier_State
124	ABC	Acme & Johnson Supply Co	IL

Supplier_History

Supplier_key	Supplier_Code	Supplier_Name	Supplier_State	Create_Date
123	ABC	Acme Supply Co	CA	14-June-2003
124	ABC	Acme & Johnson Supply Co	IL	22-Dec-2004

Copyright © 2016 Pearson Education, Inc.

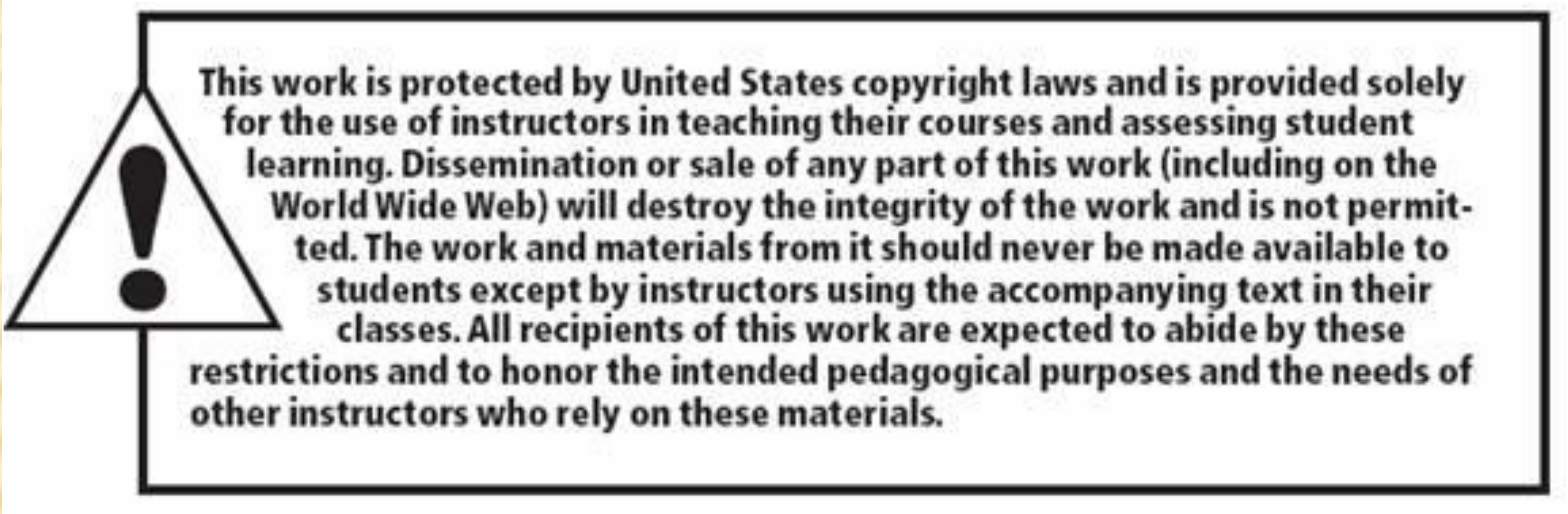
10 ESSENTIAL RULES FOR DIMENSIONAL MODELING

- Use atomic facts
- Create single-process fact tables
- Include a date dimension for each fact table
- Enforce consistent grain
- Disallow null keys in fact tables
- Honor hierarchies
- Decode dimension tables
- Use surrogate keys
- Conform dimensions
- Balance requirements with actual data

THE FUTURE OF DATA WAREHOUSING: INTEGRATION WITH BIG DATA AND ANALYTICS

- Issue of Big Data (huge volume, often unstructured)
- Speed of processing
  - Design/purchase storage, database, and networking aspects in tandem
  - Use in-memory databases (RAM instead of disk)
  - Add analytics capabilities closer to the original data sources instead of separate data warehouses
- Cost of Data Storage
  - Move data warehouse to the cloud
  - Use Columnar databases for storage optimization
- Unstructured Data
  - NoSql “Not only SQL”
  - Hadoop

Copyright © 2016 Pearson Education, Inc.



This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.