Chapter 11: Data Warehousing

Modern Database Management

6th Edition

Jeffrey A. Hoffer, Mary B. Prescott, Fred R. McFadden

Definition

- **Data Warehouse**:
  - A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes
  - ***Subject-oriented:*** e.g. customers, patients, students, products
  - ***Integrated:*** Consistent naming conventions, formats, encoding structures; from multiple data sources
  - ***Time-variant:*** Can study trends and changes
  - ***Nonupdatable:*** Read-only, periodically refreshed
- **Data Mart**:
  - A data warehouse that is limited in scope

Need for Data Warehousing

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

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

Table 11-1: comparison of operational and informational systems

Table 11-1 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

Table 11-2: Data Warehouse vs. Data Mart

Table 11-2 Data Warehouse Versus Data Mart

Data Warehouse	*Data Mart*
Scope	*Scope*
• Application independent	• Specific DSS application
• Centralized, possibly enterprise-wide	• Decentralized by user area
• Planned	• Organic, possibly not planned
Data	*Data*
• Historical, detailed, and summarized	• Some history, detailed, and summarized
• Lightly denormalized	• 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	• Restrictive
• Data-oriented	• Project-oriented
• Long life	• Short life
• Large	• Start small, becomes large
• Single complex structure	• Multi, semi-complex structures, together complex

Adapted from Strange (1997)

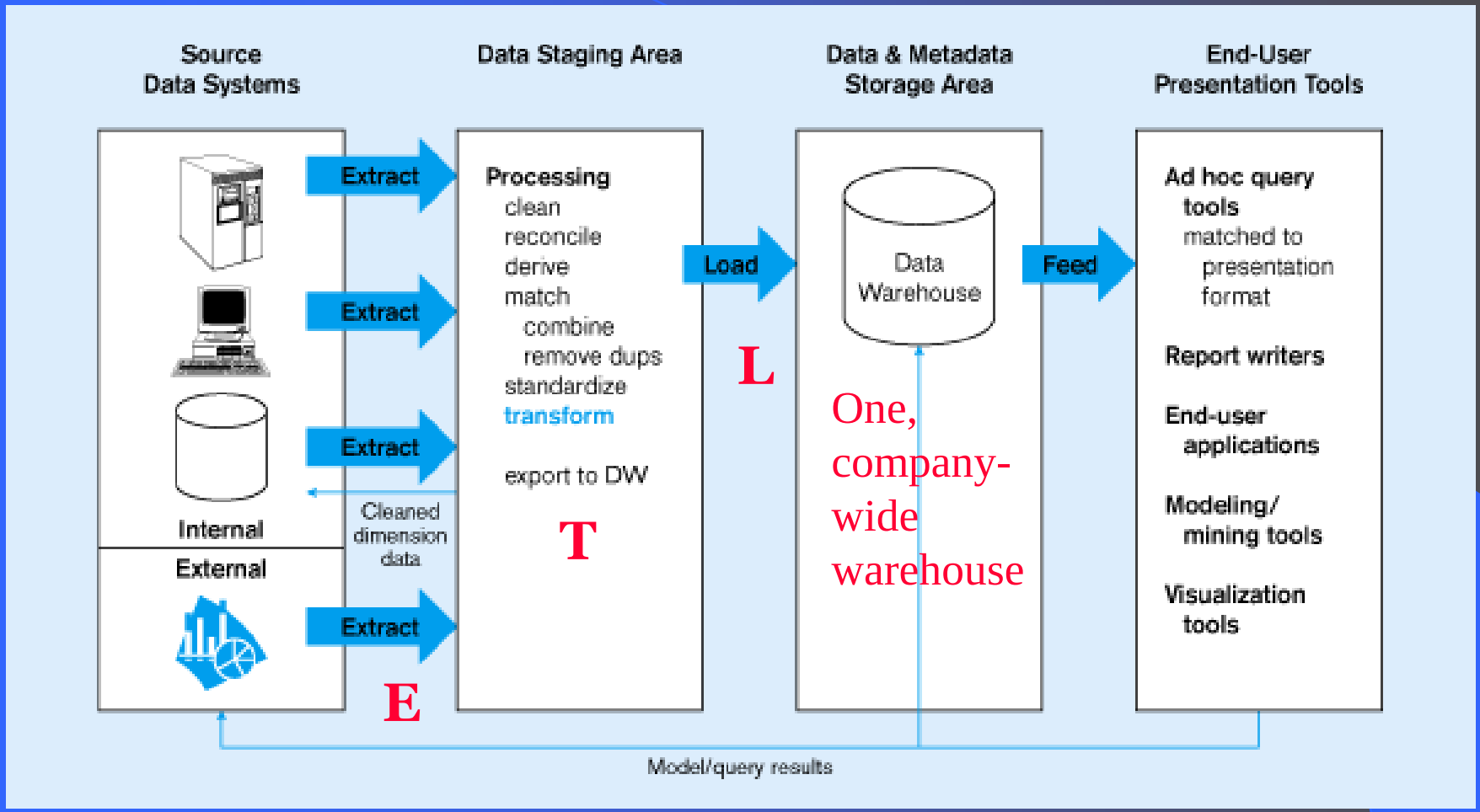
Source: adapted from Strange (1997).

Data Warehouse Architectures

- Generic Two-Level Architecture
- Independent Data Mart
- Dependent Data Mart and Operational Data Store
- Logical Data Mart and @ctive Warehouse
- Three-Layer architecture

All involve some form of ***extraction***, ***transformation*** and ***loading*** (**ETL**)

Figure 11-2: Generic two-level architecture

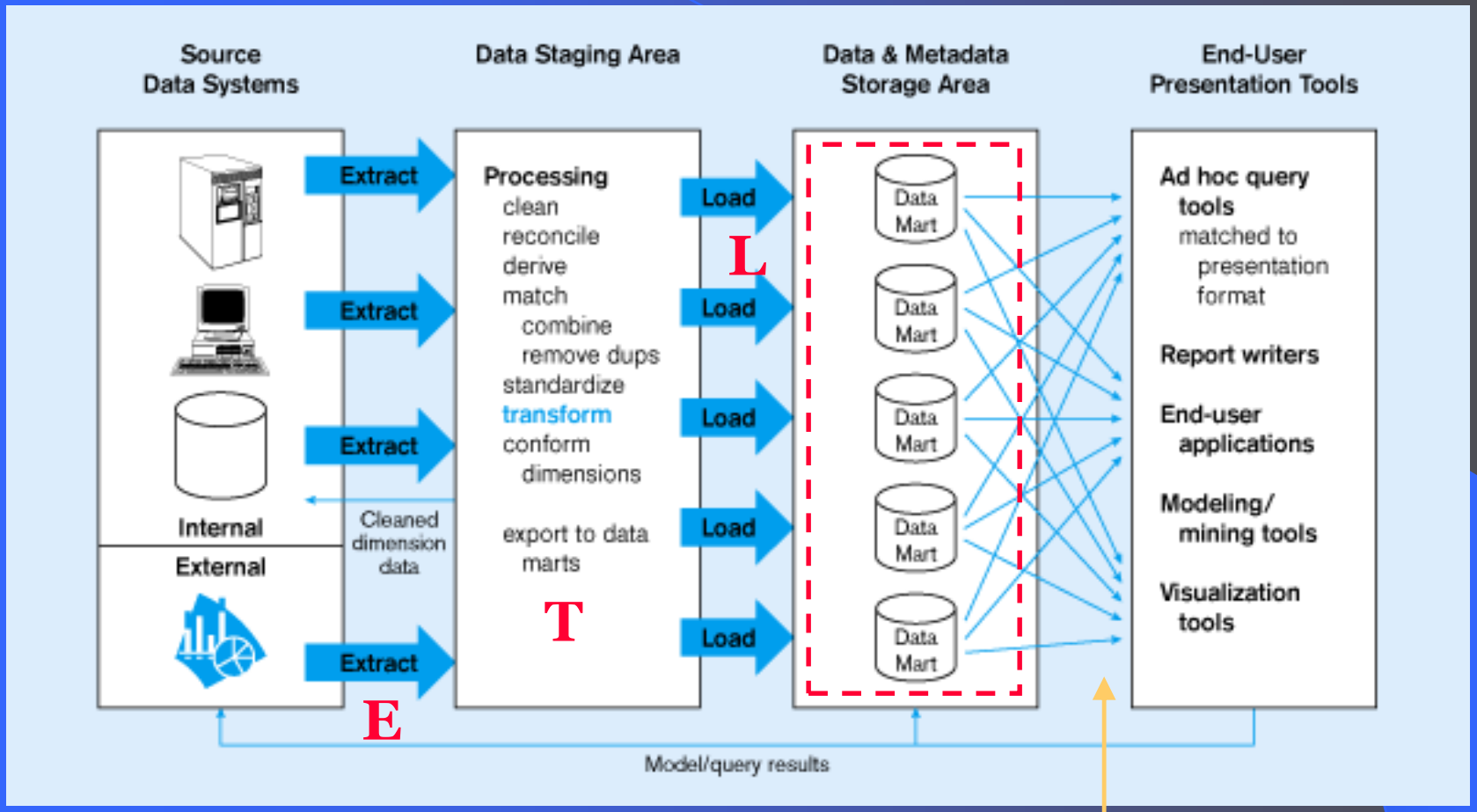


Periodic extraction ➔ data is not completely current in warehouse

Figure 11-3: Independent Data Mart

Data marts:

Mini-warehouses, limited in scope

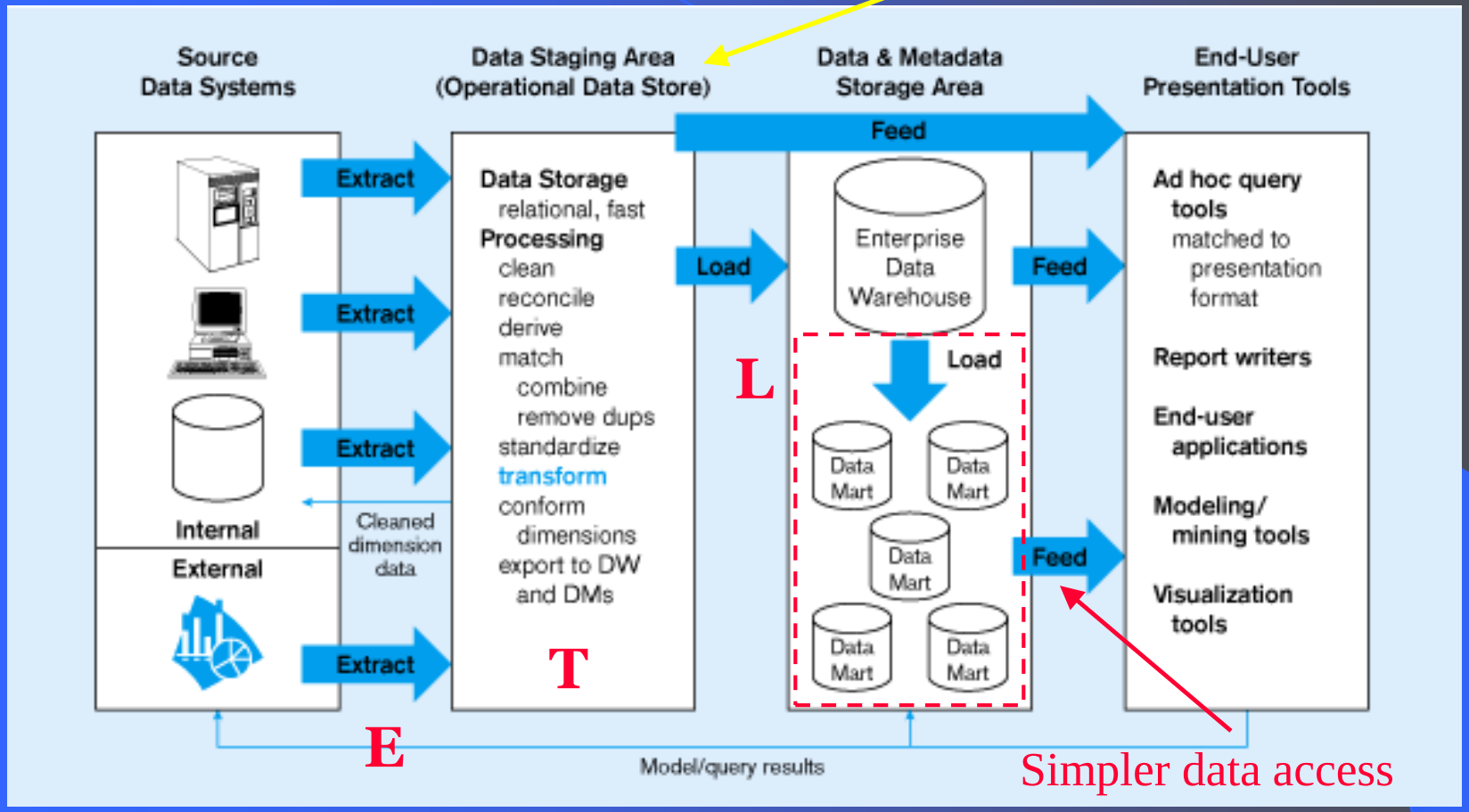


Separate ETL for each ***independent*** data mart

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

Figure 11-4:
Dependent data mart with *operational data store*

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

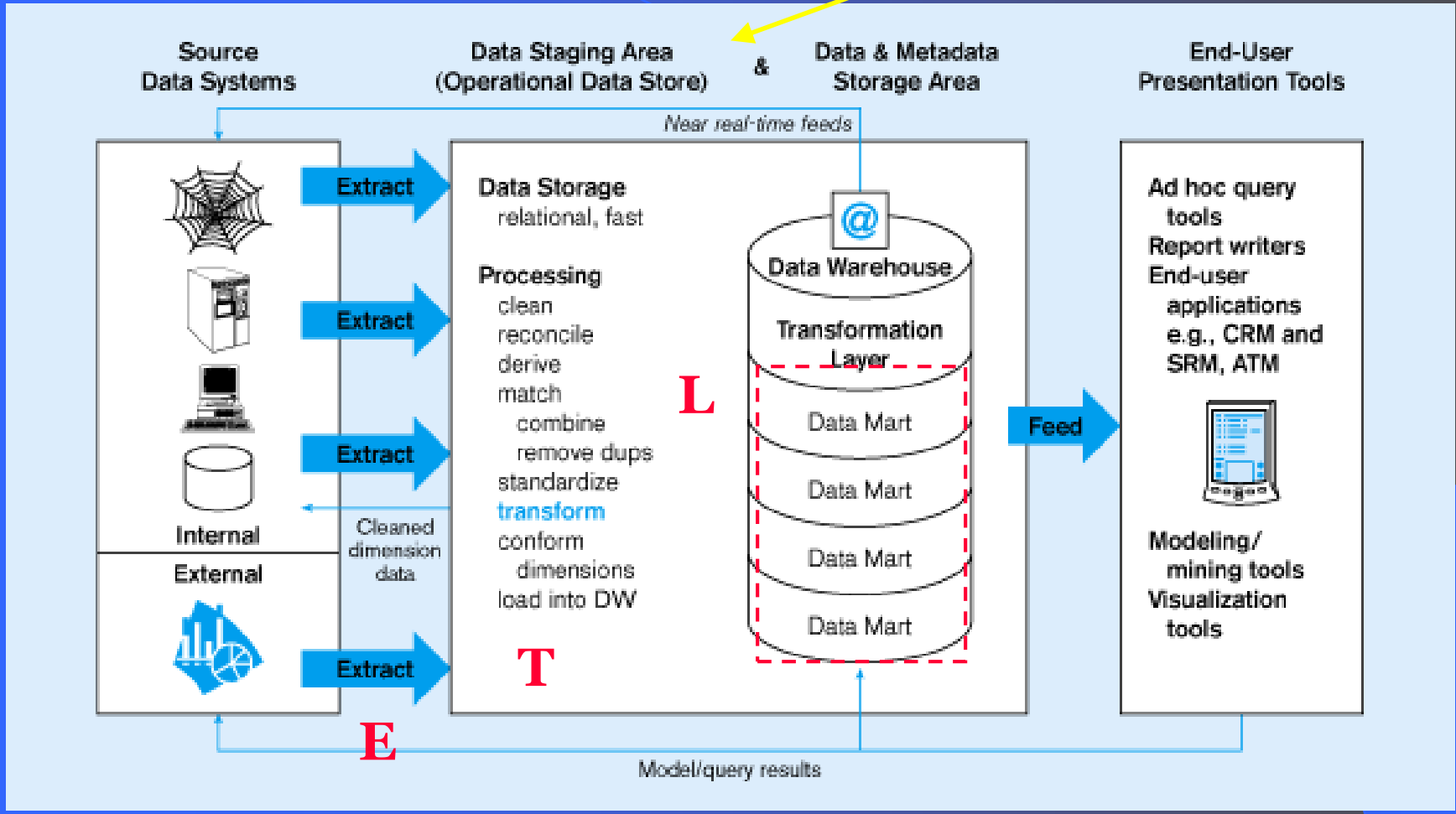


Single ETL for ***enterprise data warehouse (EDW)***

Dependent data marts loaded from EDW

Figure 11-5:
Logical data mart and @ctive data warehouse

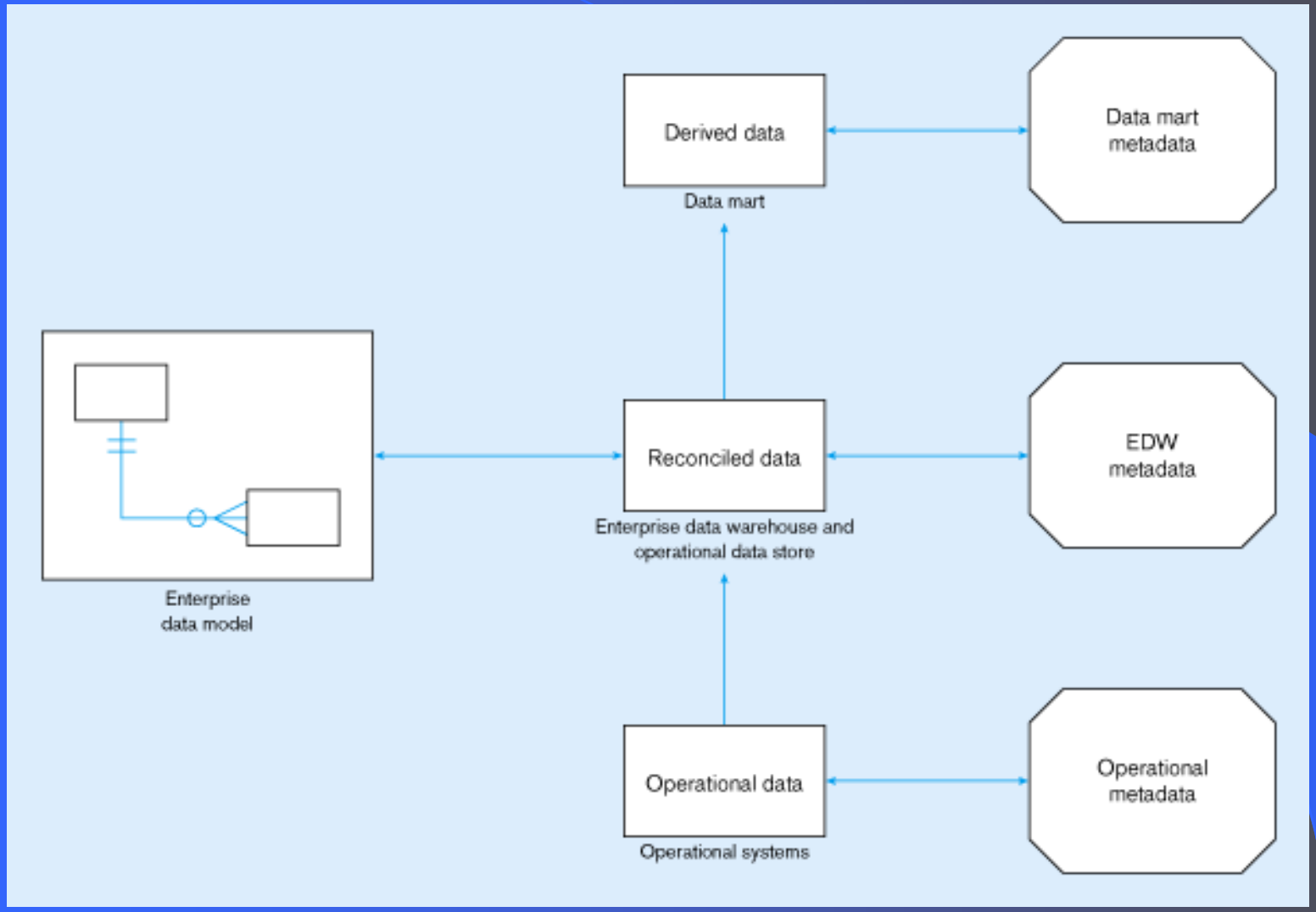
ODS and **data warehouse** are one and the same



Near real-time ETL for
@active Data Warehouse

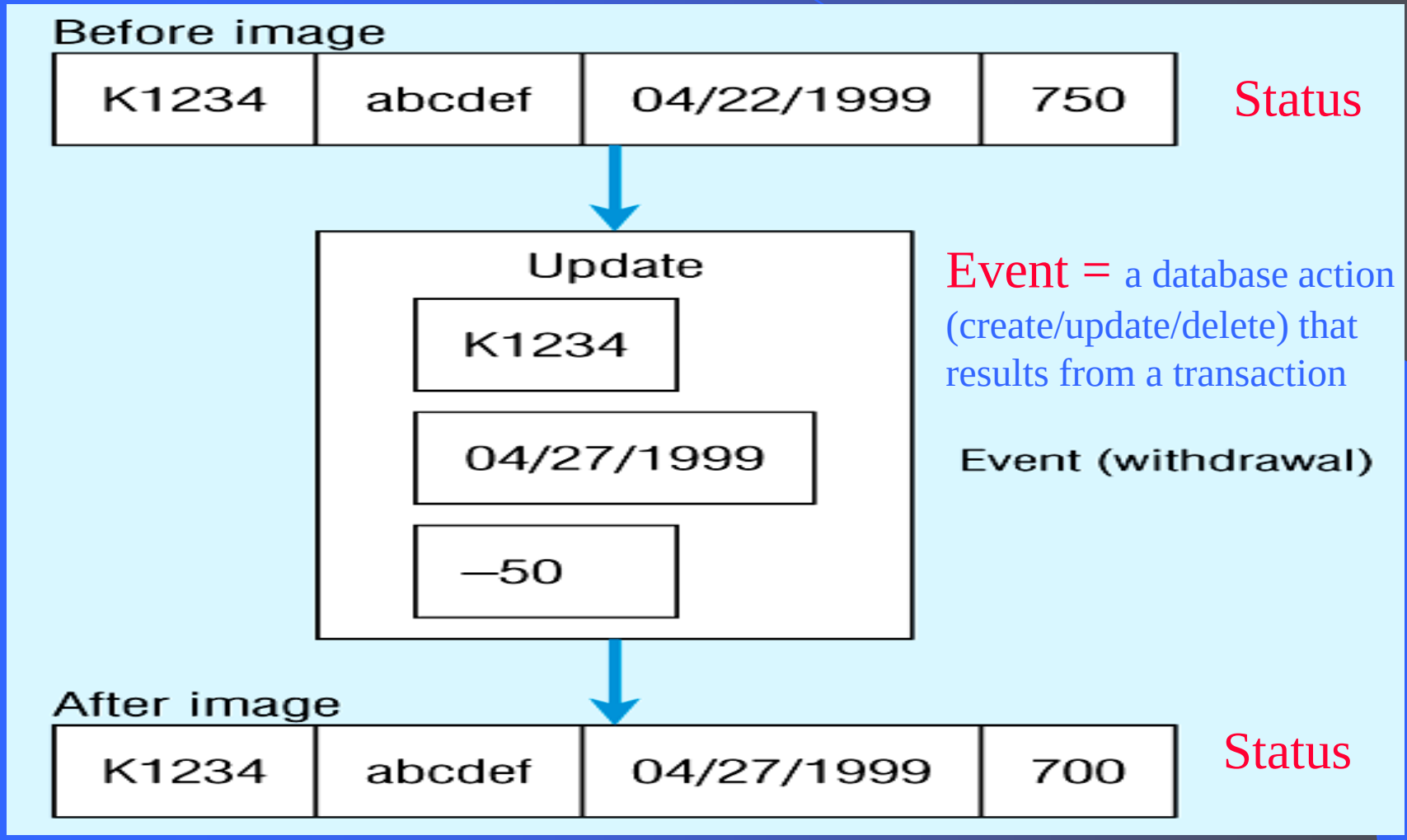
Data marts are NOT separate databases, but logical ***views*** of the data warehouse
➔ Easier to create new data marts

Figure 11-6: Three-layer architecture



Data Characteristics
Status vs. Event Data

Figure 11-7: Example of DBMS log entry

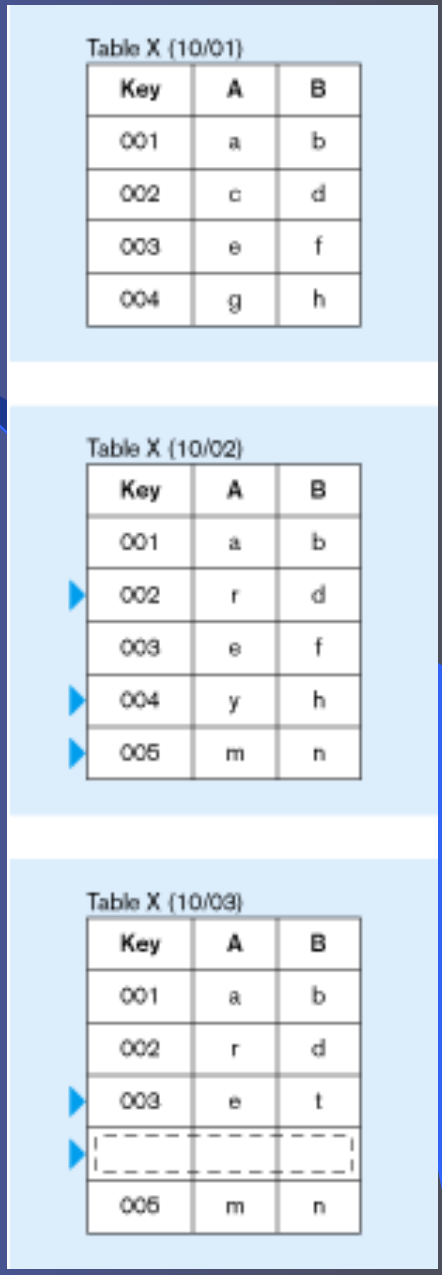


Data Characteristics

Transient vs. Periodic Data

Figure 11-8:
Transient operational data

Changes to existing records are written over previous records, thus destroying the previous data content



Data Characteristics

Transient vs. Periodic Data

Figure 11-9:
Periodic warehouse data

Data are never physically altered or deleted once they have been added to the store

Table X (10/01)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
003	10/01	e	f	C
004	10/01	g	h	C

Table X (10/02)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
002	10/02	r	d	U
003	10/01	e	f	C
004	10/01	g	h	C
004	10/02	y	h	U
005	10/02	m	n	C

Table X (10/03)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
002	10/02	r	d	U
003	10/01	e	f	C
003	10/03	e	f	U
004	10/01	g	h	C
004	10/02	y	h	U
004	10/03	y	h	D
005	10/02	m	n	C

Data Reconciliation

Typical operational data is:

- Transient – not historical
- Not normalized (perhaps due to denormalization for performance)
- Restricted in scope – not comprehensive
- Sometimes poor quality – inconsistencies and errors

After ETL, data should be:

- Detailed – not summarized yet
- Historical – periodic
- Normalized – 3rd normal form or higher
- Comprehensive – enterprise-wide perspective
- Quality controlled – accurate with full integrity

The ETL Process

Capture

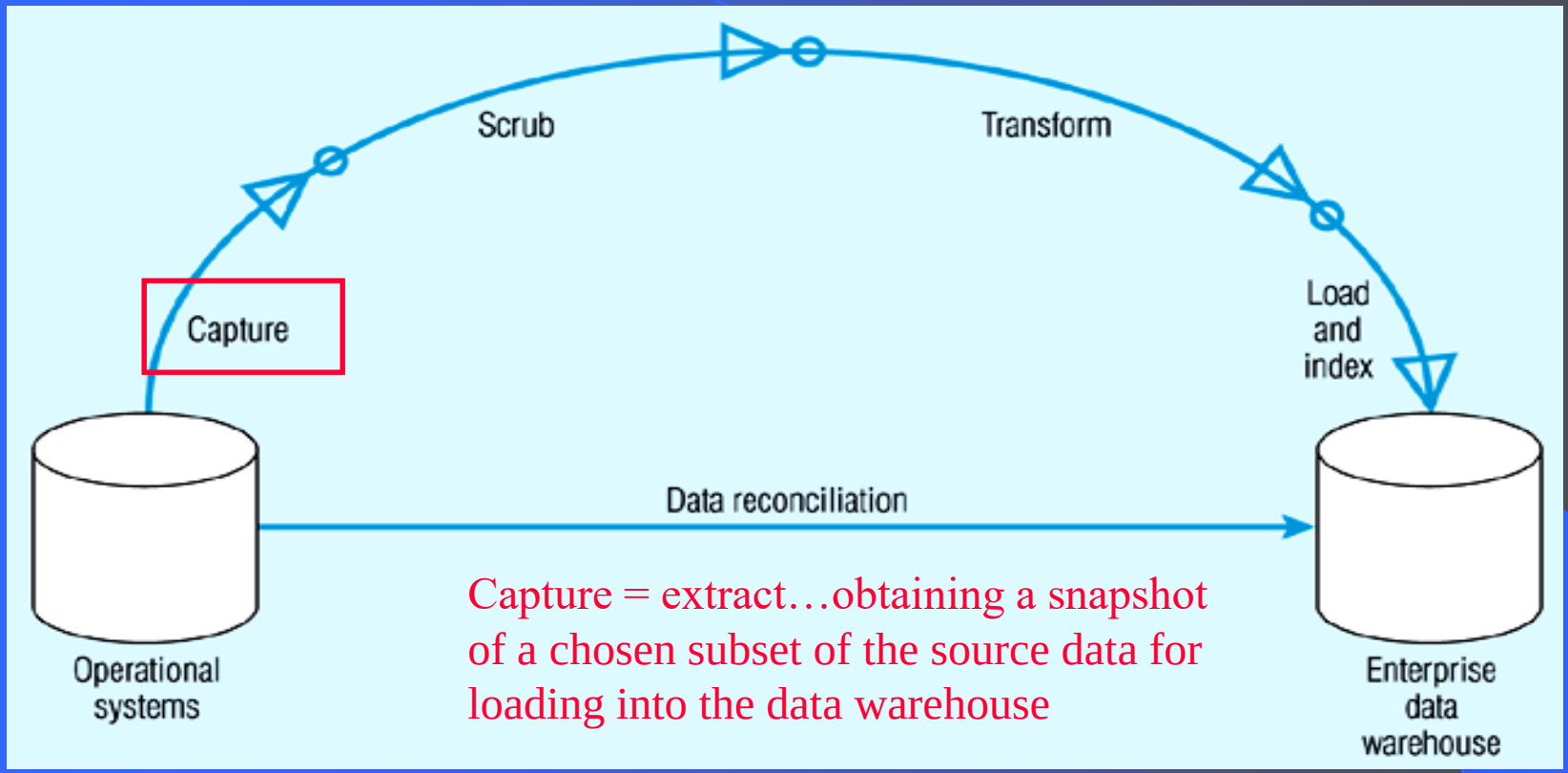
Scrub or data cleansing

Transform

Load and Index

ETL = Extract, transform, and load

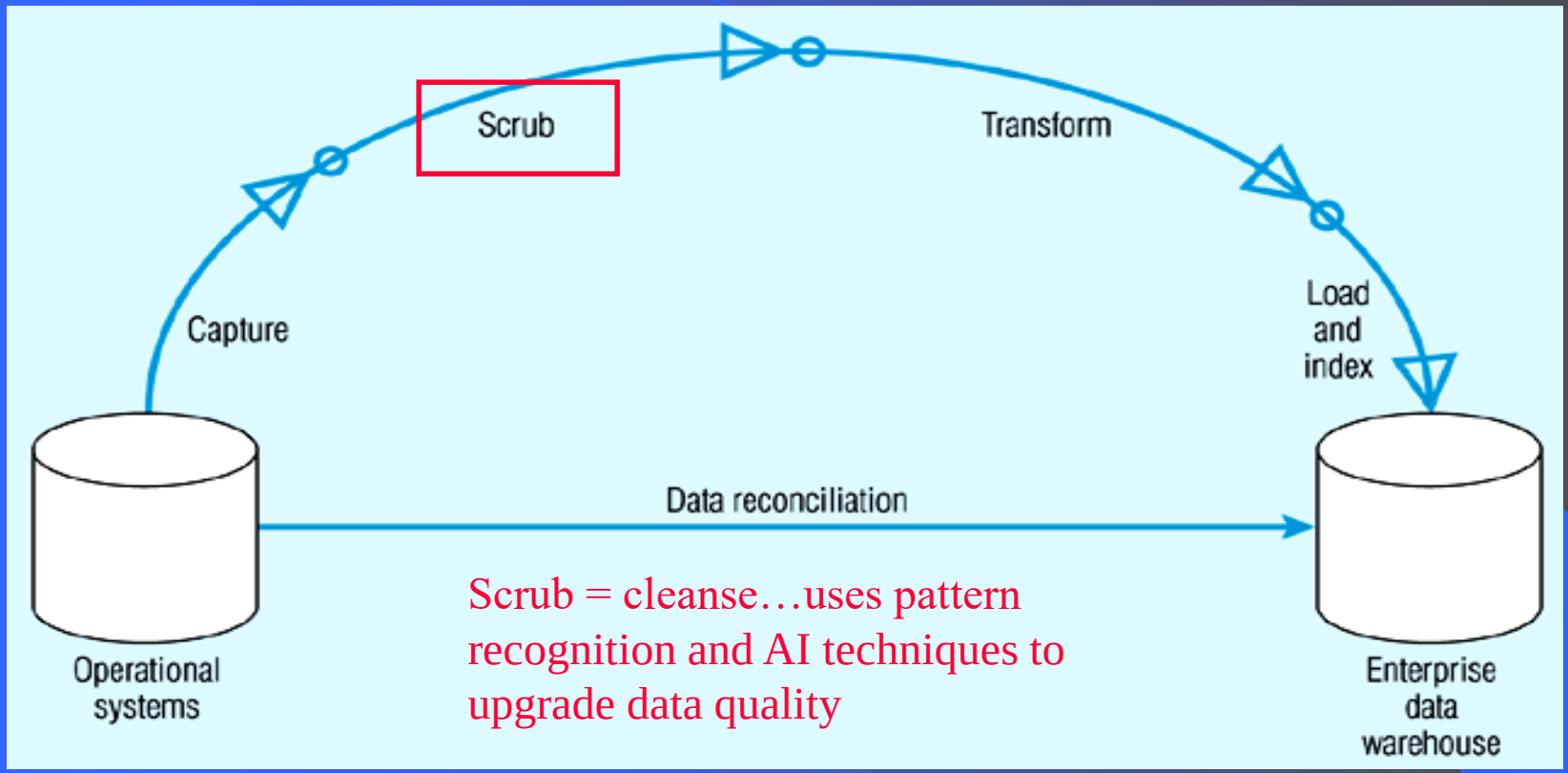
Figure 11-10: Steps in data reconciliation



Static extract = capturing a snapshot of the source data at a point in time

Incremental extract = capturing changes that have occurred since the last static extract

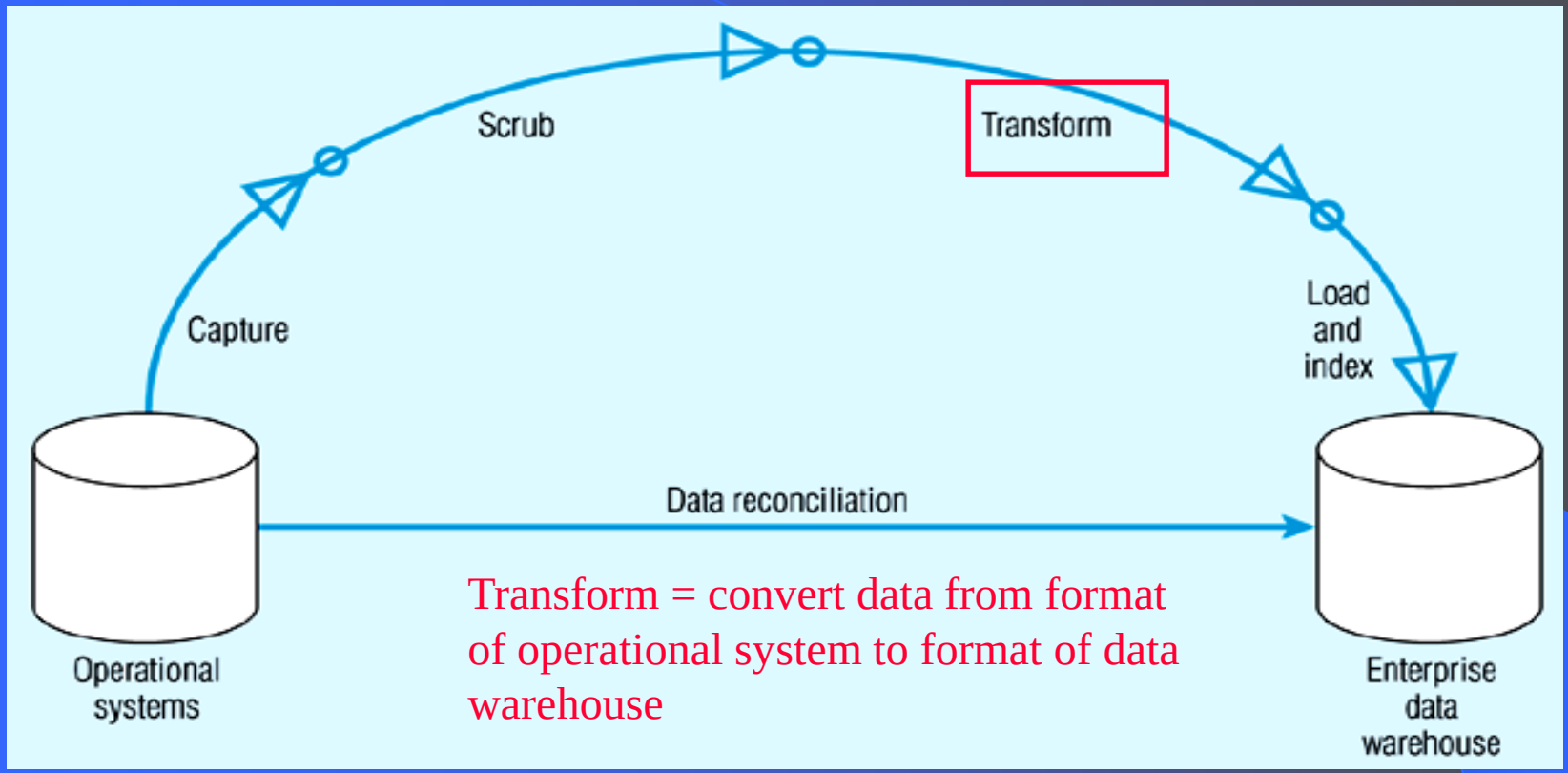
Figure 11-10: Steps in data reconciliation (continued)



Fixing errors: misspellings, erroneous dates, incorrect field usage, mismatched addresses, missing data, duplicate data, inconsistencies

Also: decoding, reformatting, time stamping, conversion, key generation, merging, error detection/logging, locating missing data

Figure 11-10: Steps in data reconciliation (continued)



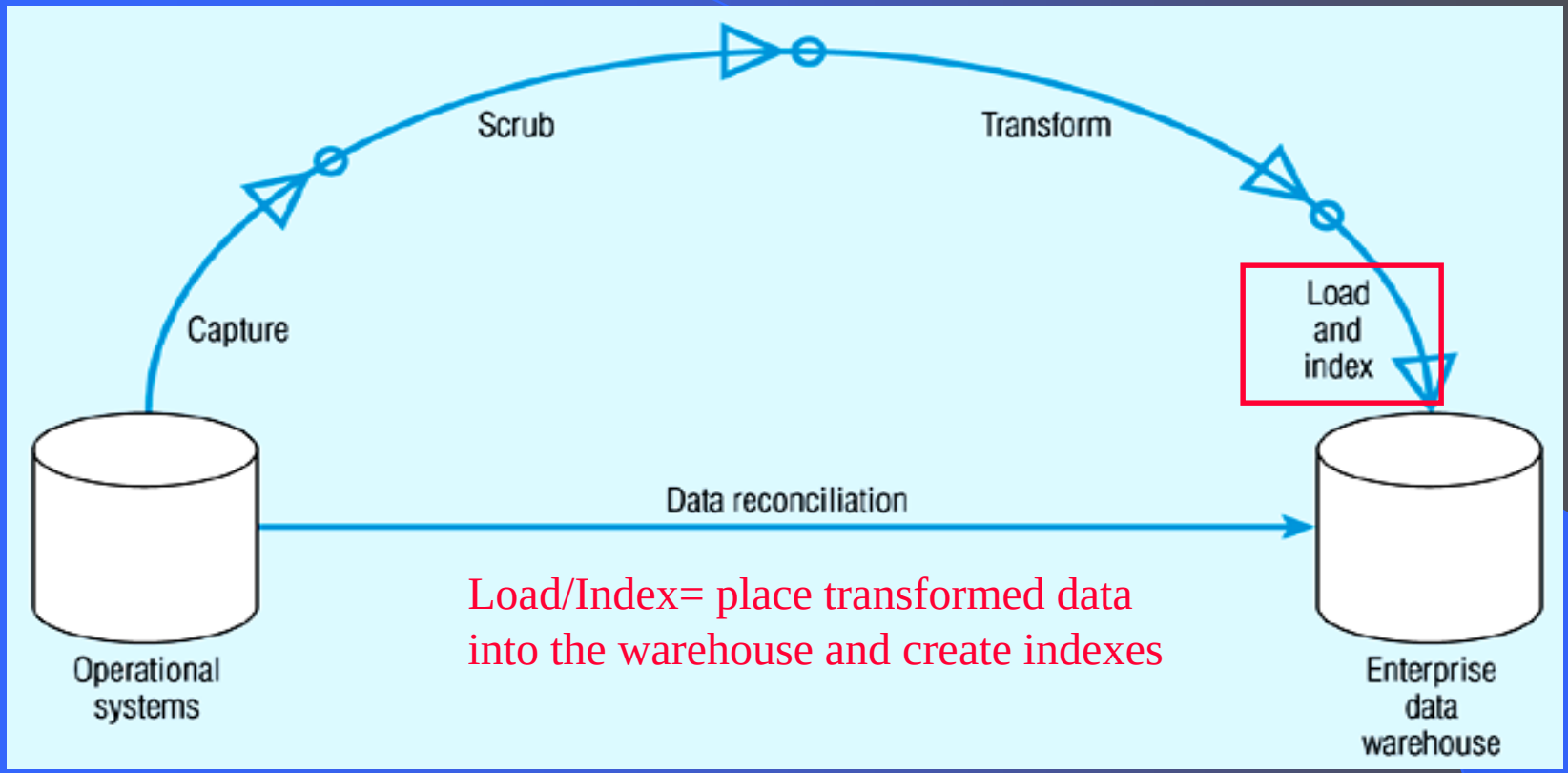
Record-level:

Selection – data partitioning
Joining – data combining
Aggregation – data summarization

Field-level:

single-field – from one field to one field
multi-field – from many fields to one, or one field to many

Figure 11-10: Steps in data reconciliation (continued)



Refresh mode: bulk rewriting of target data at periodic intervals

Update mode: only changes in source data are written to data warehouse

Figure 11-11: Single-field transformation

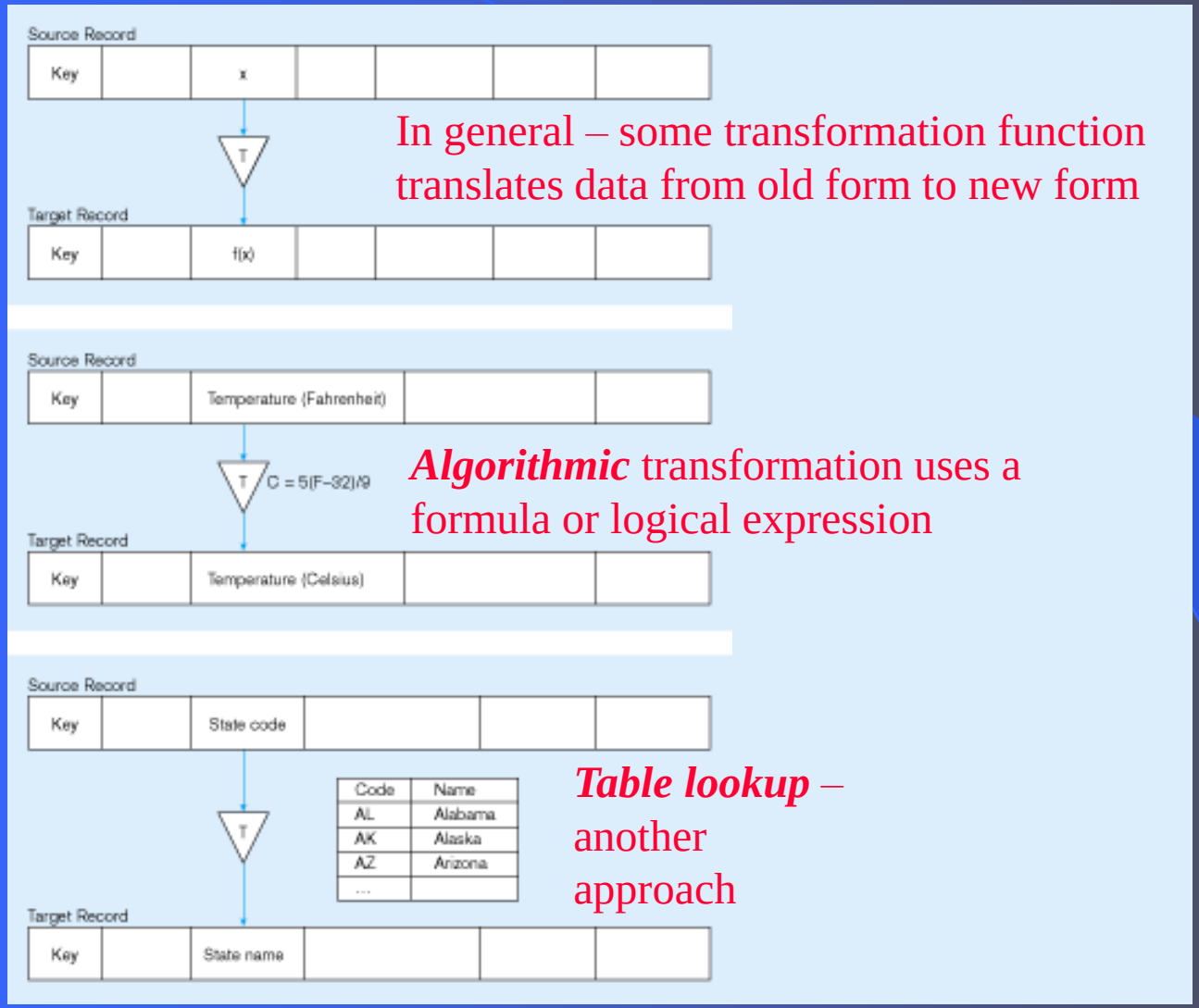
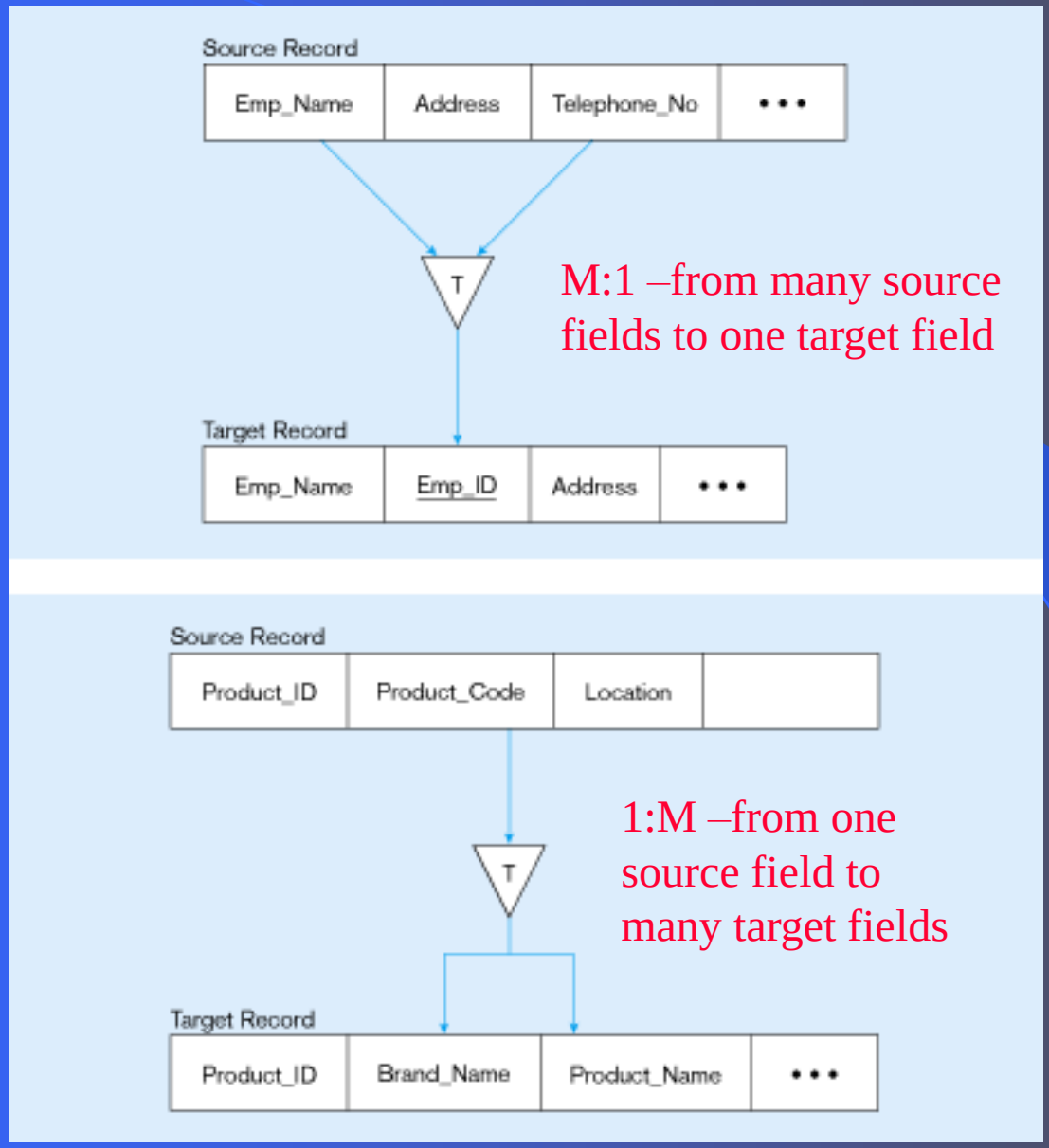


Figure 11-12: Multifield transformation



Derived Data

Objectives

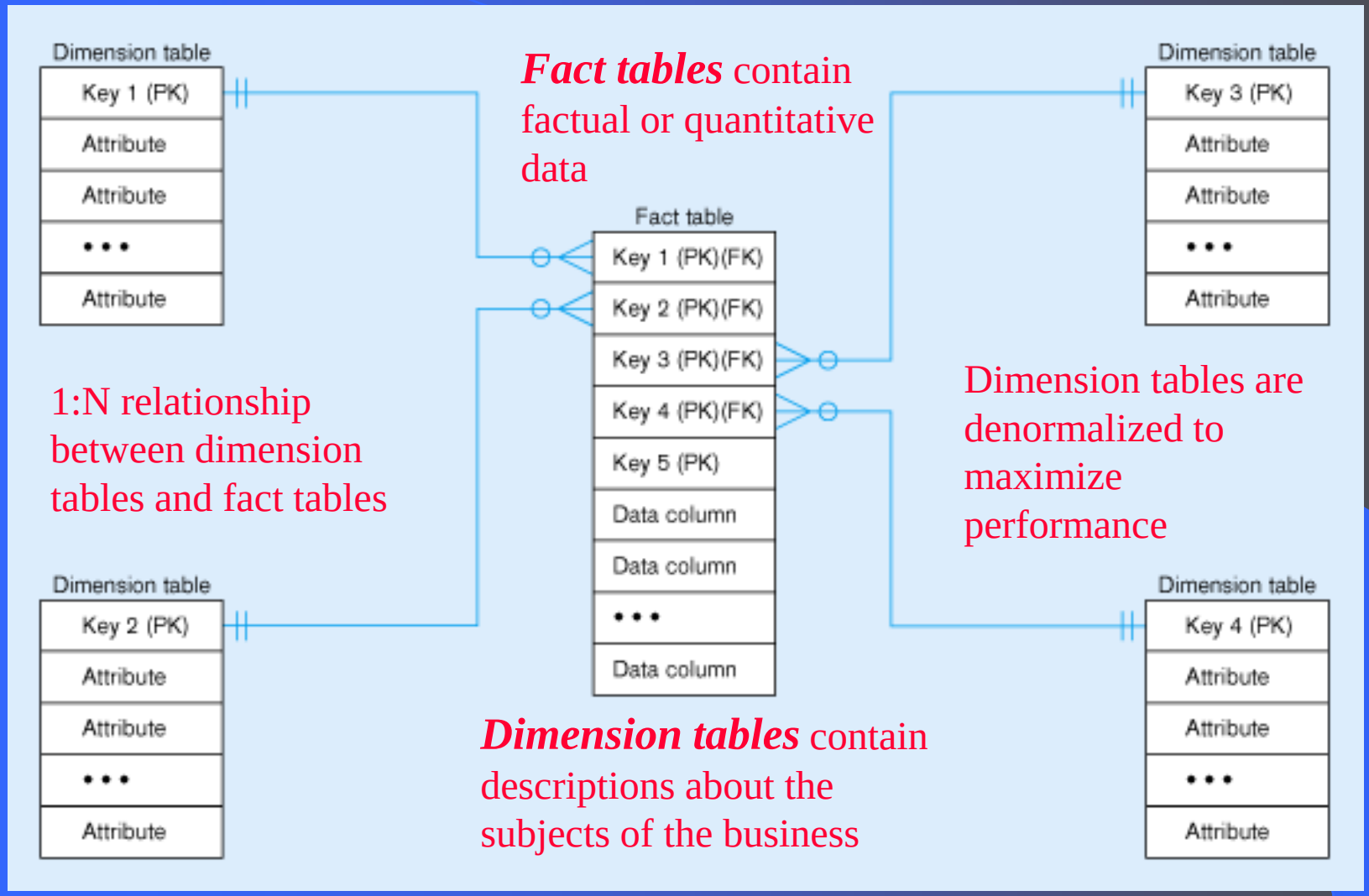
- Ease of use for decision support applications
- Fast response to predefined user queries
- Customized data for particular target audiences
- Ad-hoc query support
- Data mining capabilities

➔ Characteristics

- Detailed (mostly periodic) data
- Aggregate (for summary)
- Distributed (to departmental servers)

Most common data model = **star schema**
(also called "dimensional model")

Figure 11-13: Components of a **star schema**



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

Figure 11-14: Star schema example

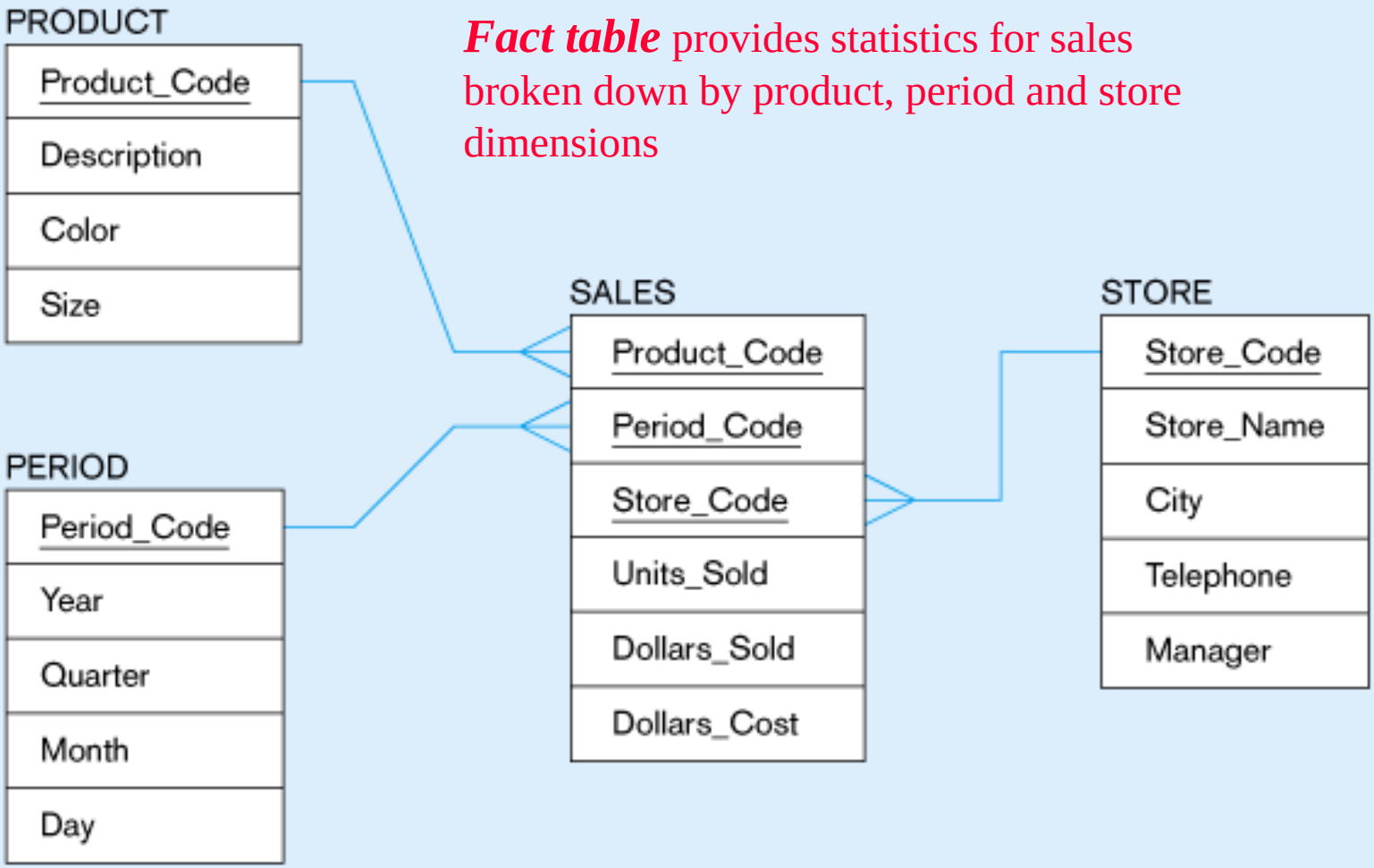
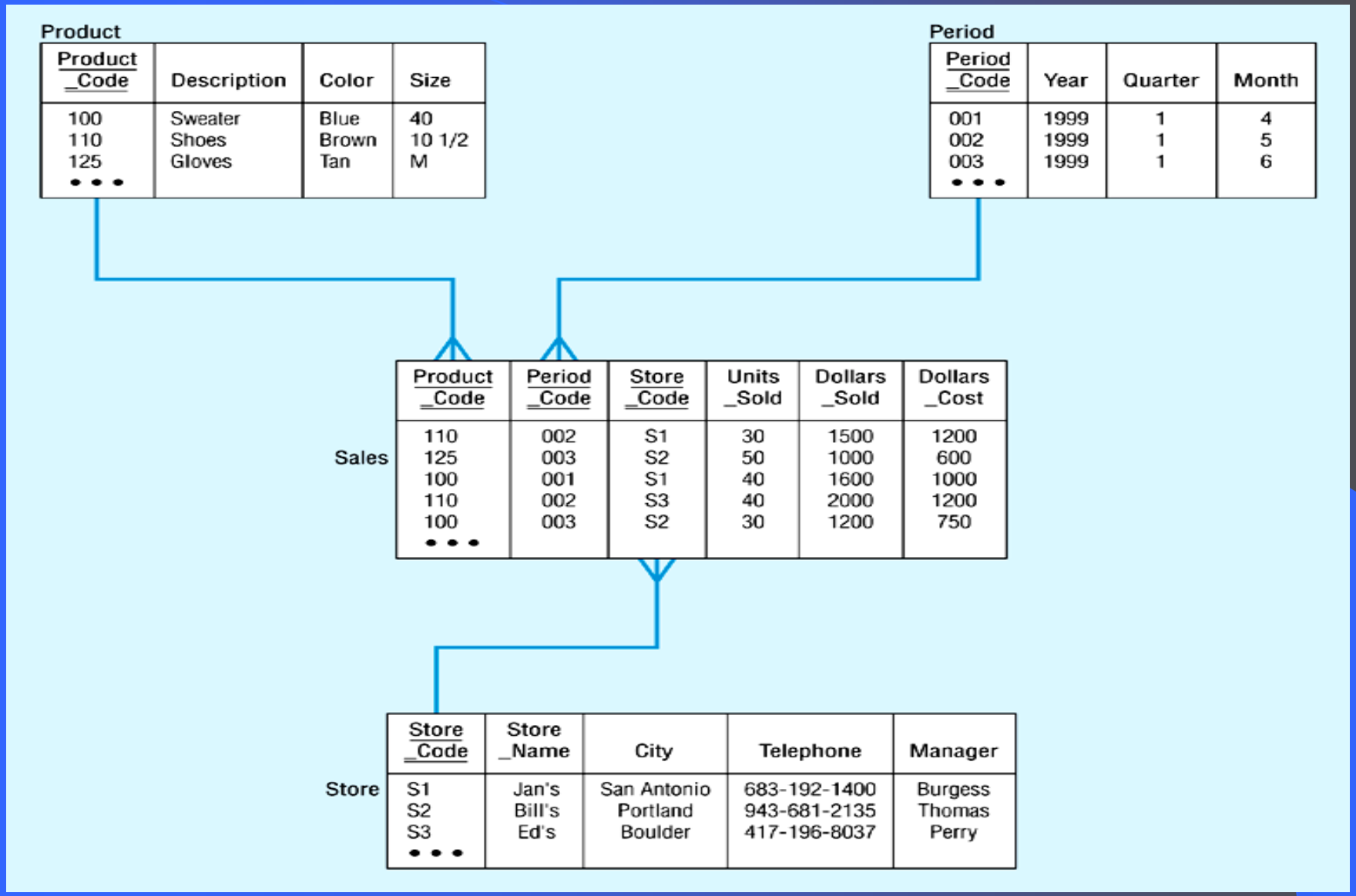


Figure 11-15: Star schema with sample data



Issues Regarding Star Schema

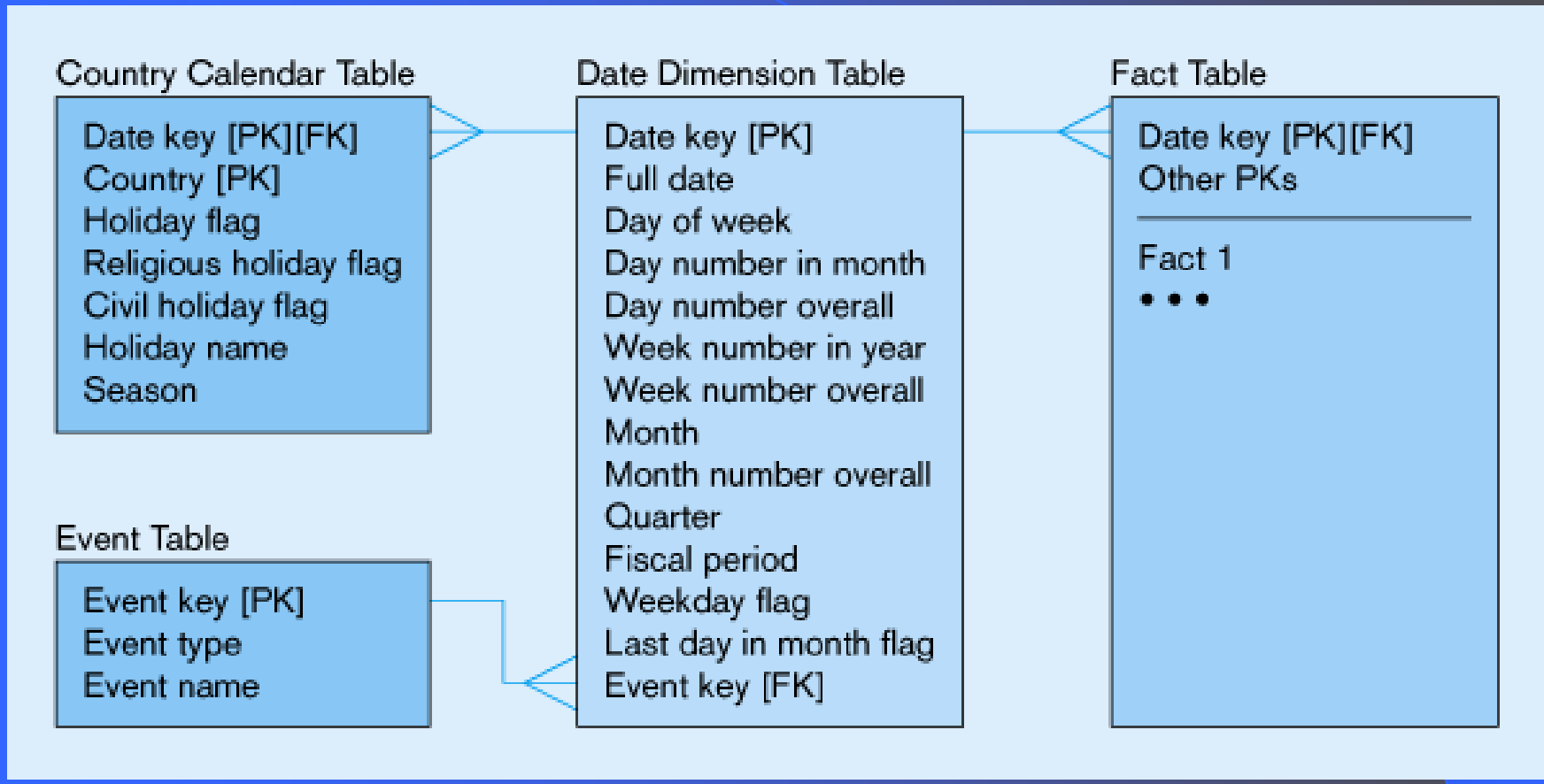
Dimension table keys must be ***surrogate*** (non-intelligent and non-business related), because:

- Keys may change over time
- Length/format consistency

Granularity of Fact Table – what level of detail do you want?

- 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

Figure 11-16: Modeling dates



Fact tables contain time-period data

➔ Date dimensions are important

The User Interface Metadata (data catalog)

- Identify subjects of the data mart
- Identify dimensions and facts
- Indicate how data is derived from enterprise data warehouses, including derivation rules
- Indicate how data is derived from operational data store, including derivation rules
- Identify available reports and predefined queries
- Identify data analysis techniques (e.g. drill-down)
- Identify responsible people

On-Line Analytical Processing (OLAP)

The use of a set of graphical tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques

Relational OLAP (ROLAP)

- Traditional relational representation

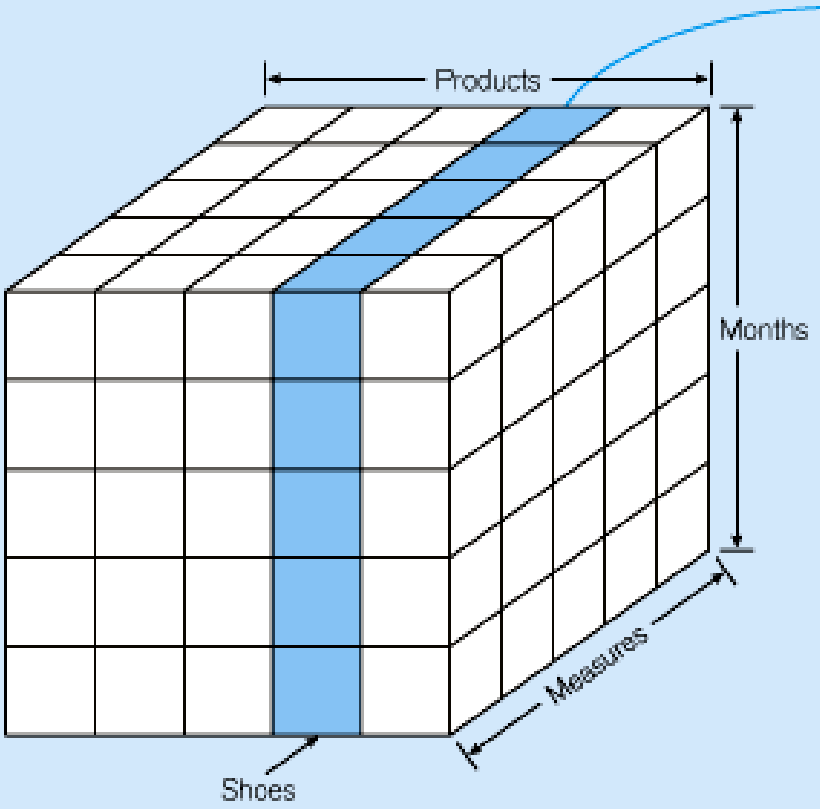
Multidimensional OLAP (MOLAP)

- **Cube** structure

OLAP Operations

- ***Cube slicing*** – come up with 2-D view of data
- ***Drill-down*** – going from summary to more detailed views

Figure 11-22: Slicing a data cube



	Measure		
	Units	Revenue	Cost
January	250	1564	1020
February	200	1275	875
March	350	1800	1275
April	400	1935	1500
May	485	2000	1560

Product: Shoes

Figure 11-23: Example of drill-down

Summary report

Brand	Package size	Sales
SofTowel	2-pack	$75
SofTowel	3-pack	$100
SofTowel	6-pack	$50

Drill-down with color added

Brand	Package size	Color	Sales
SofTowel	2-pack	White	$30
SofTowel	2-pack	Yellow	$25
SofTowel	2-pack	Pink	$20
SofTowel	3-pack	White	$50
SofTowel	3-pack	Green	$25
SofTowel	3-pack	Yellow	$25
SofTowel	6-pack	White	$30
SofTowel	6-pack	Yellow	$20

Data Mining and Visualization

Knowledge discovery using a blend of statistical, AI, and computer graphics techniques

Goals:

- Explain observed events or conditions
- Confirm hypotheses
- Explore data for new or unexpected relationships

Techniques

- Case-based reasoning
- Rule discovery
- Signal processing
- Neural nets
- Fractals

Data visualization – representing data in graphical/multimedia formats for analysis