Data Warehousing

University of California, Berkeley
School of Information

INFO 257: Database Management

Announcements

- Assignment 3 due
- Data Warehousing Lecture
- Workshop today (github & final project task)
- Questions about Assignment 4 (final project)

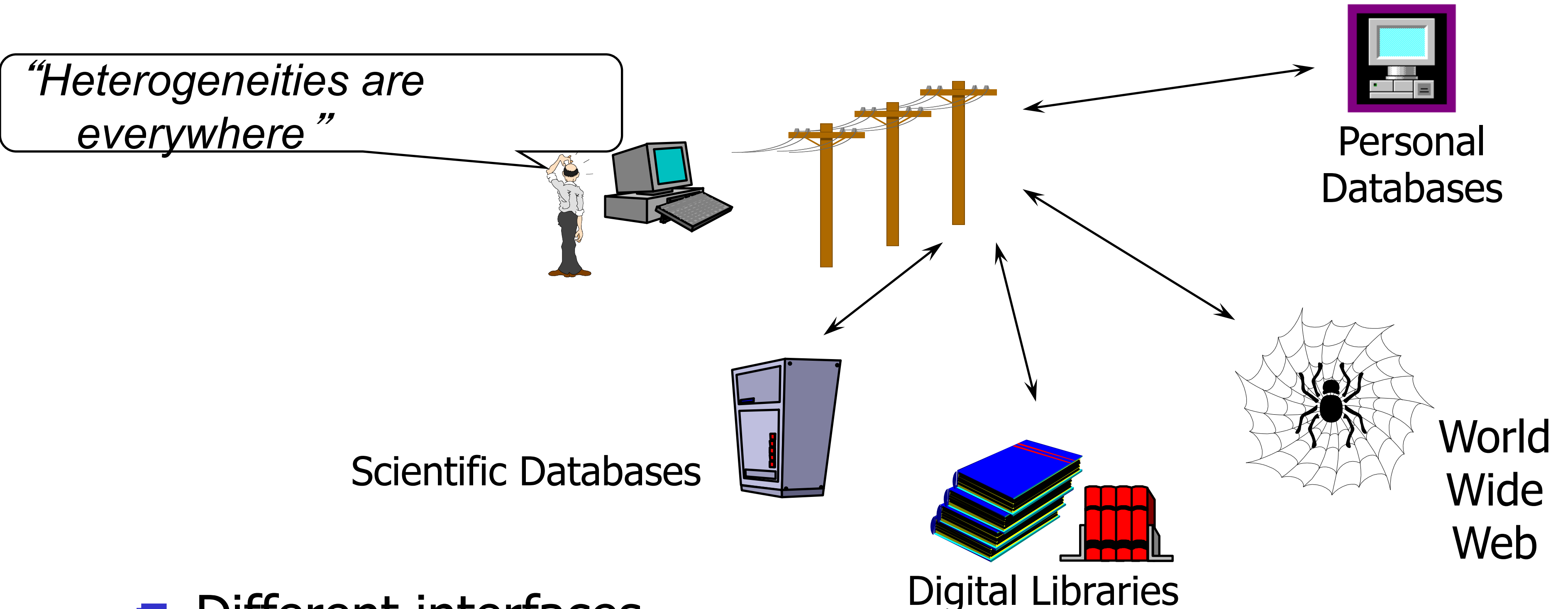
Lecture Outline

- Data Warehouses
- Introduction to Data Warehouses
- Data Warehousing
  - (Based on lecture notes from *Modern Database Management* Text (Hoffer, Ramesh, Topi); Joachim Hammer, University of Florida, and Joe Hellerstein and Mike Stonebraker of UCB)

Overview

- Data Warehouses and Merging Information Resources
- What is a Data Warehouse?
- History of Data Warehousing
- Types of Data and Their Uses
- Data Warehouse Architectures
- Data Warehousing Problems and Issues

Problem: Heterogeneous Information Sources

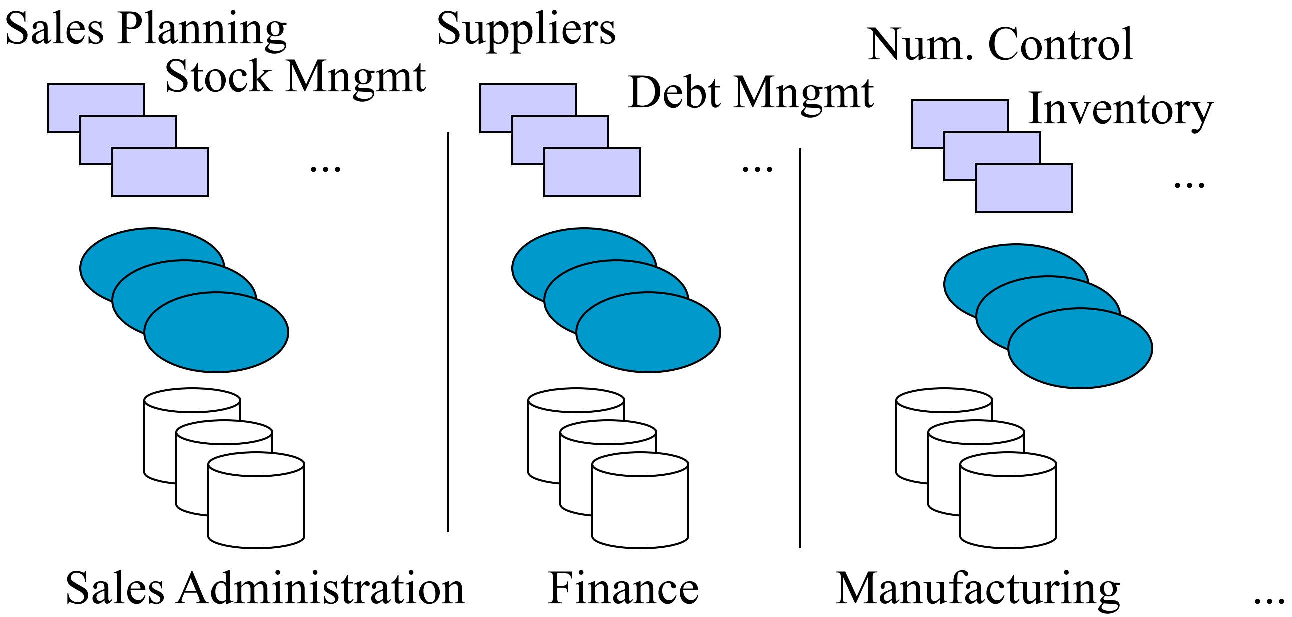


- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Slide credit: J. Hammer

Problem: Data Management in Large Enterprises

- Vertical fragmentation of informational systems (vertical stove pipes)
- Result of application (user)-driven development of operational systems



Slide credit: J. Hammer

Issues with Fragmentation

- Inconsistent key structures
- Synonyms
- Free-form vs. structured fields
- Inconsistent data values
- Missing data

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	

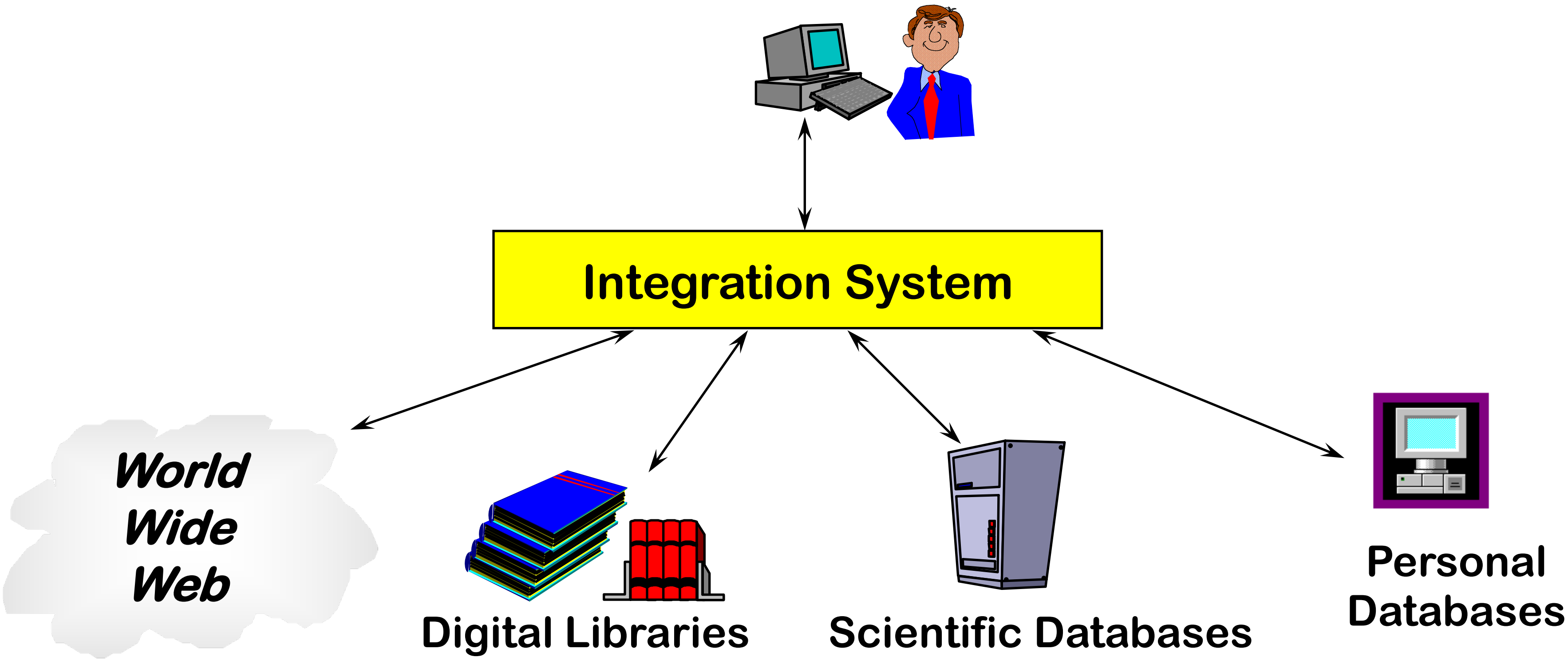
STUDENT HEALTH

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

History Leading to Data Warehousing

- Improvement in database technologies, especially relational DBMSs
- Advances in computer hardware, including mass storage and parallel architectures
- Emergence of end-user computing with powerful interfaces and tools
- Advances in middleware, enabling heterogeneous database connectivity
- Recognition of difference between operational and informational systems

Goal: Unified Access to Data

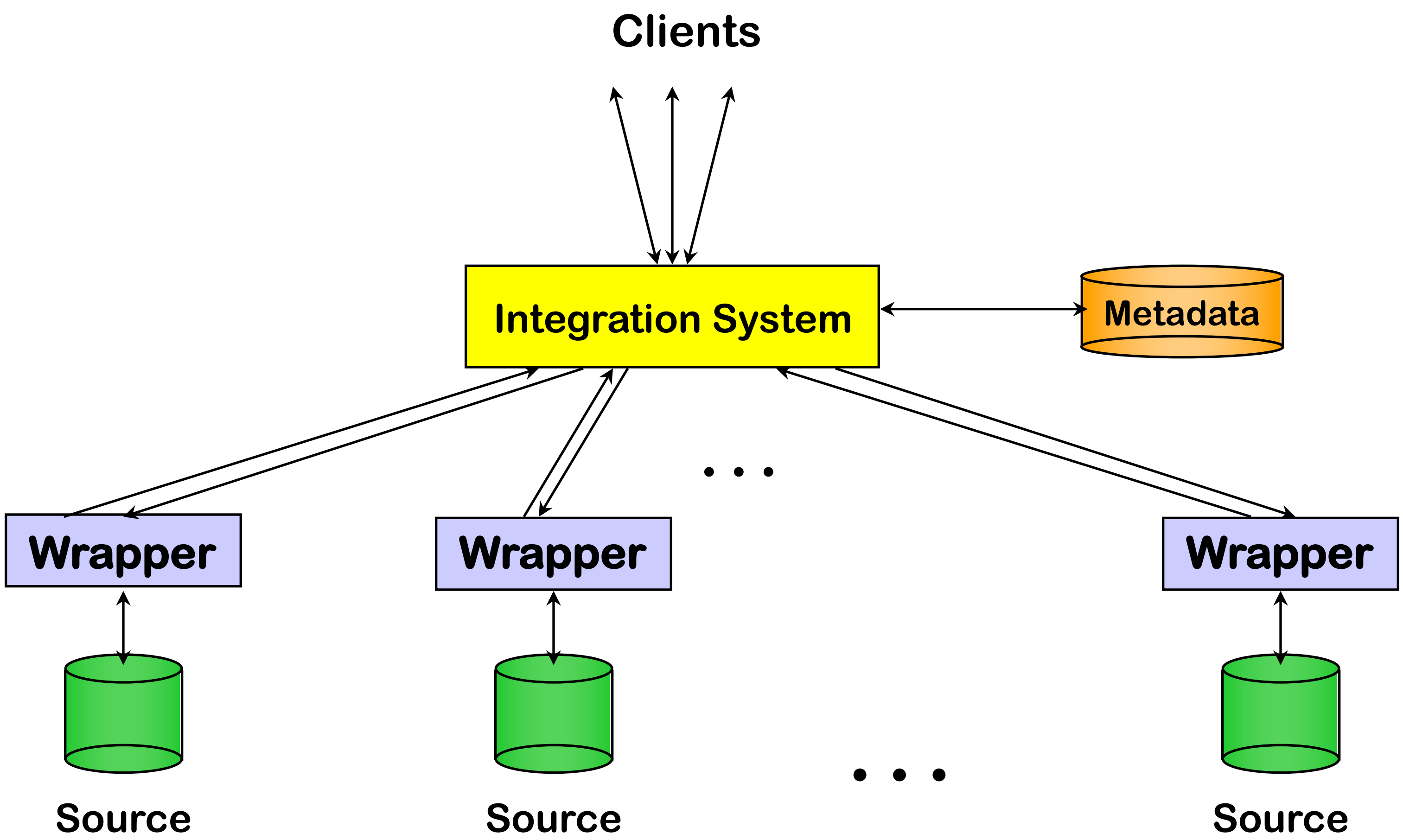


- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

Slide credit: J. Hammer

The Traditional Research Approach

- Query-driven (lazy, on-demand)



Slide credit: J. Hammer

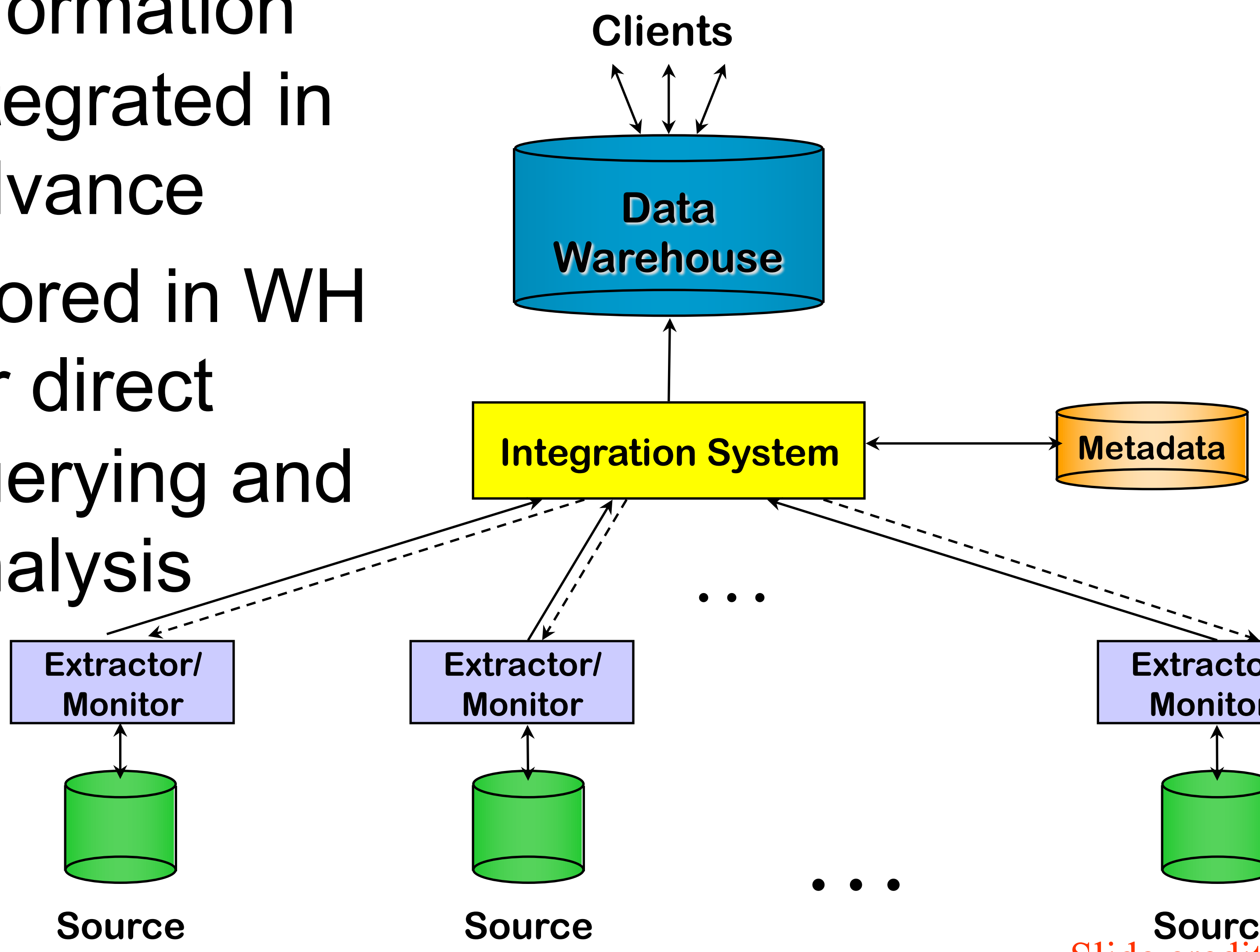
Disadvantages of Query-Driven Approach

- Delay in query processing
  - Slow or unavailable information sources
  - Complex filtering and integration
- Inefficient and potentially expensive for frequent queries
- Competes with local processing at sources
- Hasn’t caught on in industry

Slide credit: J. Hammer

The Warehousing Approach

- Information integrated in advance
- Stored in WH for direct querying and analysis



Slide credit: J. Hammer

Advantages of Warehousing Approach

- High query performance
  - But not necessarily most current information
- Doesn't interfere with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources
- Information copied at warehouse
  - Can modify, annotate, summarize, restructure, etc.
  - Can store historical information
  - Security, no auditing
- **Has** caught on in industry

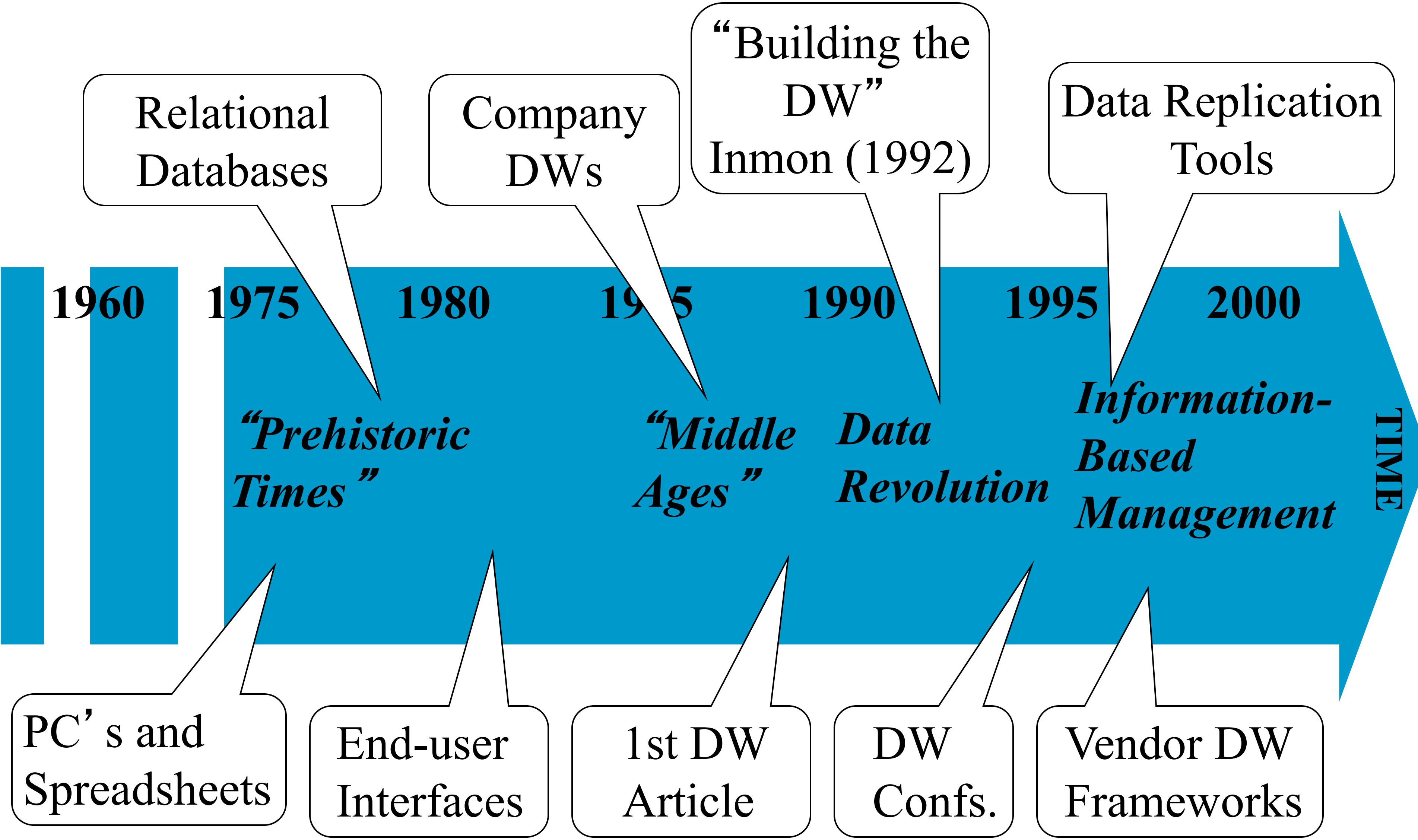
Slide credit: J. Hammer

Not Either-Or Decision

- Query-driven approach still better for
  - Rapidly changing information
  - Rapidly changing information sources
  - Truly vast amounts of data from large numbers of sources
  - Clients with unpredictable needs

Slide credit: J. Hammer

Data Warehouse Evolution



Slide credit: J. Hammer

What is a Data Warehouse?

"A Data Warehouse is a
- ***subject-oriented,***
- ***integrated,***
- ***time-variant,***
- ***non-volatile***

collection of data used in support of management decision making processes."

-- Inmon & Hackathorn, 1994: viz. Hoffer, Chap 11

DW Definition…

- Subject-Oriented:
  - The data warehouse is organized around the key subjects (or high-level entities) of the enterprise. Major subjects include
    - Customers
    - Patients
    - Students
    - Products
    - Etc.

DW Definition…

- Integrated
  - The data housed in the data warehouse are defined using consistent
    - Naming conventions
    - Formats
    - Encoding Structures
    - Related Characteristics

DW Definition…

- Time-variant
  - The data in the warehouse contain a time dimension so that they may be used as a historical record of the business

DW Definition…

- Non-volatile
  - Data in the data warehouse are loaded and refreshed from operational systems, but cannot be updated by end-users

What is a Data Warehouse? A Practitioners Viewpoint

- "A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."
- -- Barry Devlin, IBM Consultant

Slide credit: J. Hammer

A Data Warehouse is...

- Stored collection of diverse data
  - A solution to data integration problem
  - Single repository of information
- Subject-oriented
  - Organized by subject, not by application
  - Used for analysis, data mining, etc.
- Optimized differently from transaction-oriented db
- User interface aimed at executive decision makers and analysts

... Cont' d

- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at WalMart
  - Complete client histories at insurance firm
  - Stockbroker financial information and portfolios

Slide credit: J. Hammer

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
- **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

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

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

Warehouse is a Specialized DB

Standard (Operational) DB

- Mostly updates
- Many small transactions
- Mb - Gb of data
- Current snapshot
- Index/hash on p.k.
- Raw data
- Thousands of users (e.g., clerical users)

Warehouse (Informational)

- Mostly reads
- Queries are long and complex
- Gb - Tb of data
- History
- Lots of scans
- Summarized, reconciled data
- Hundreds of users (e.g., decision-makers, analysts)

Slide credit: J. Hammer

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

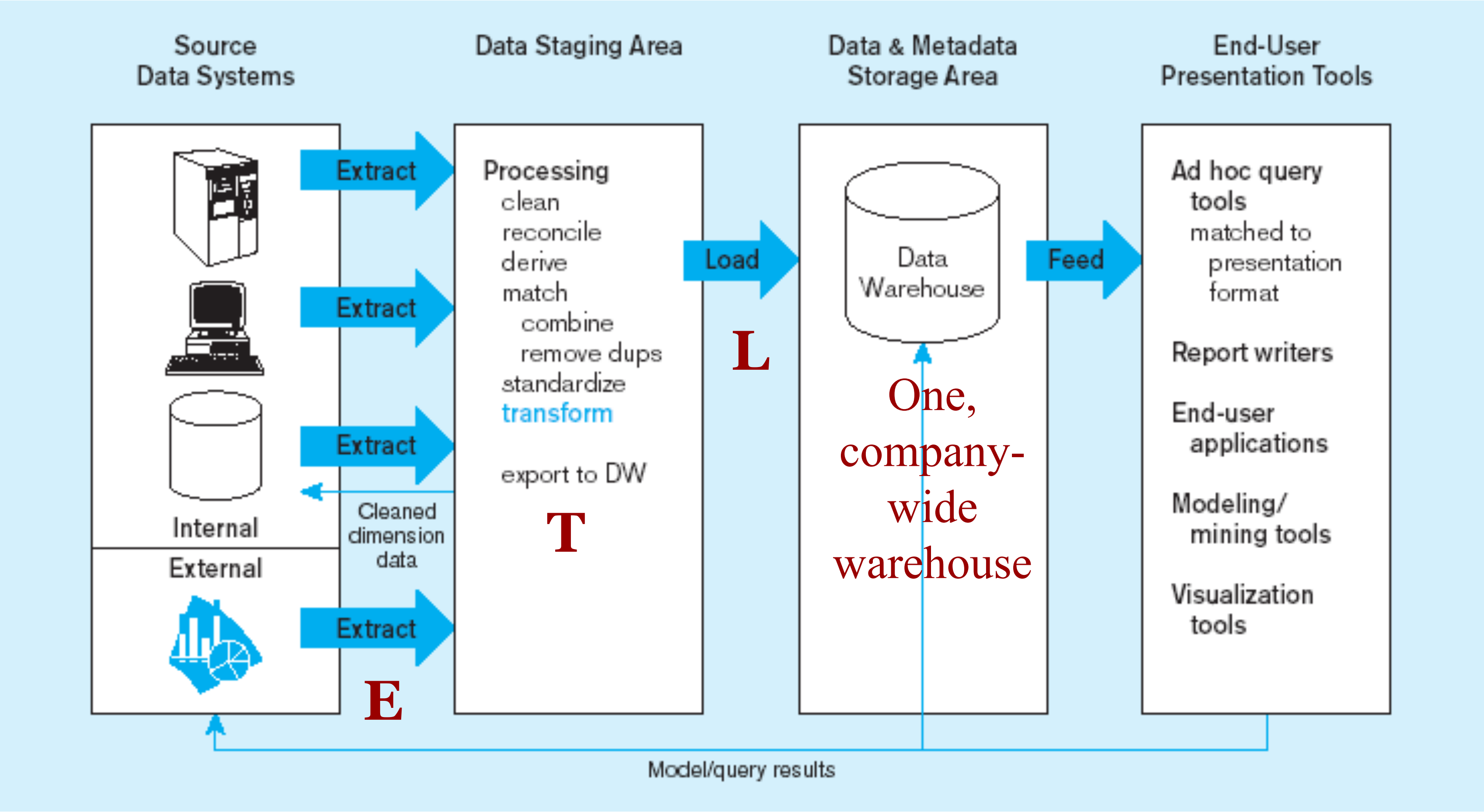
Adapted from Strange (1997)

Data Warehouse Architectures

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

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

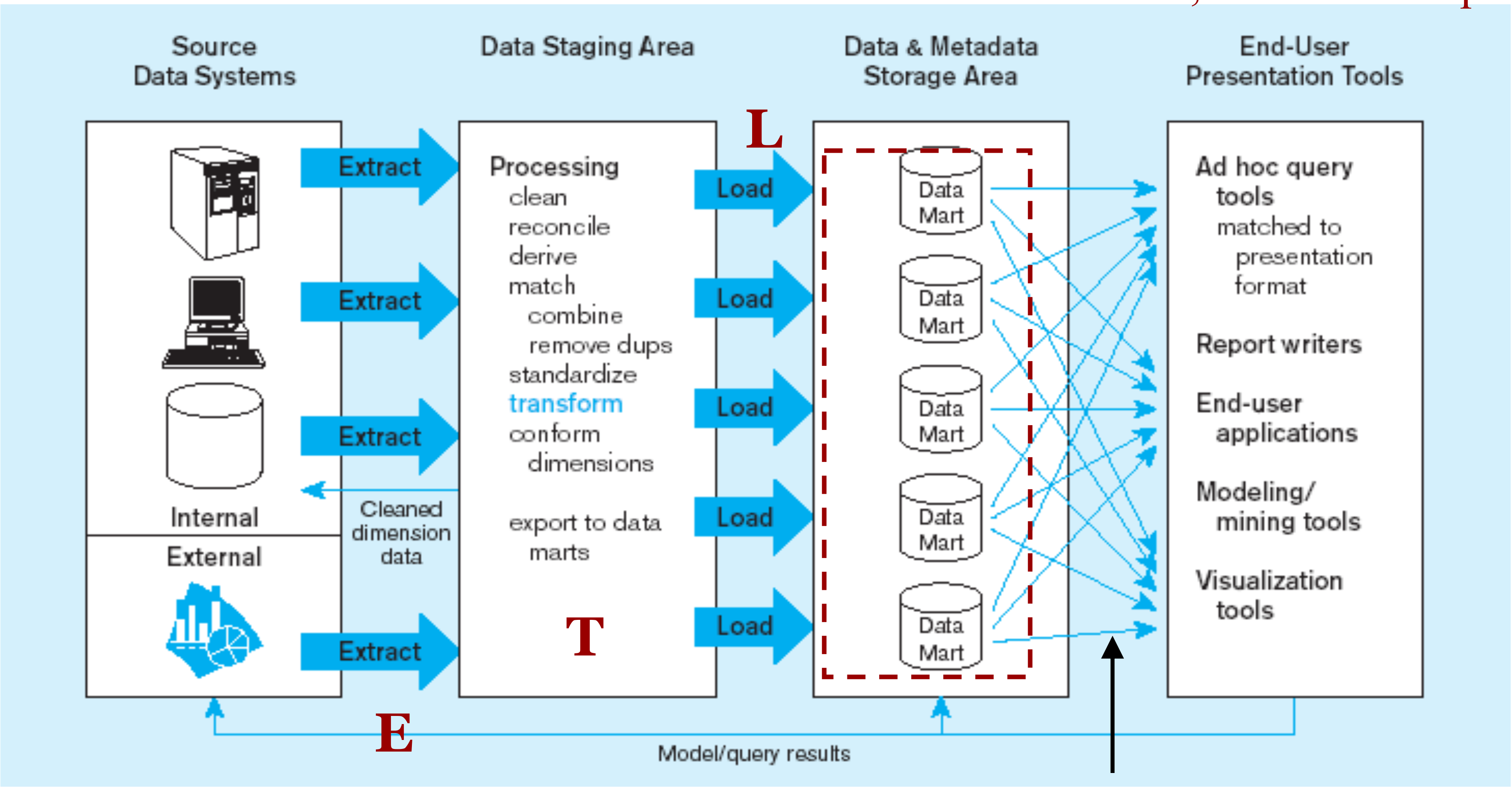
Generic two-level data warehousing architecture



Periodic extraction ➔ data is not completely current in warehouse

Independent data mart data warehousing architecture

Data marts:
Mini-warehouses, limited in scope

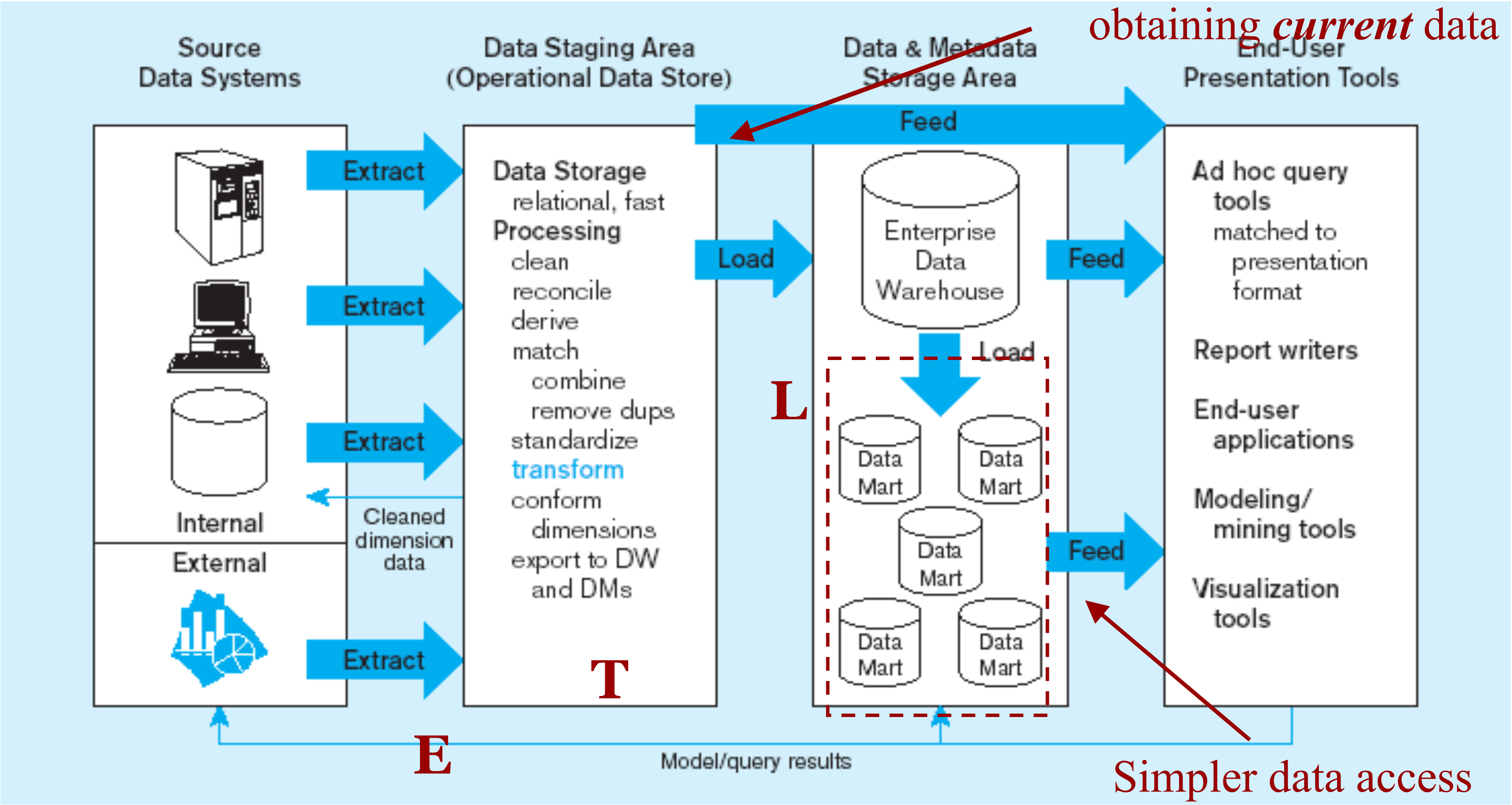


Separate ETL for each *independent* data mart

Data access complexity due to *multiple* data marts

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

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

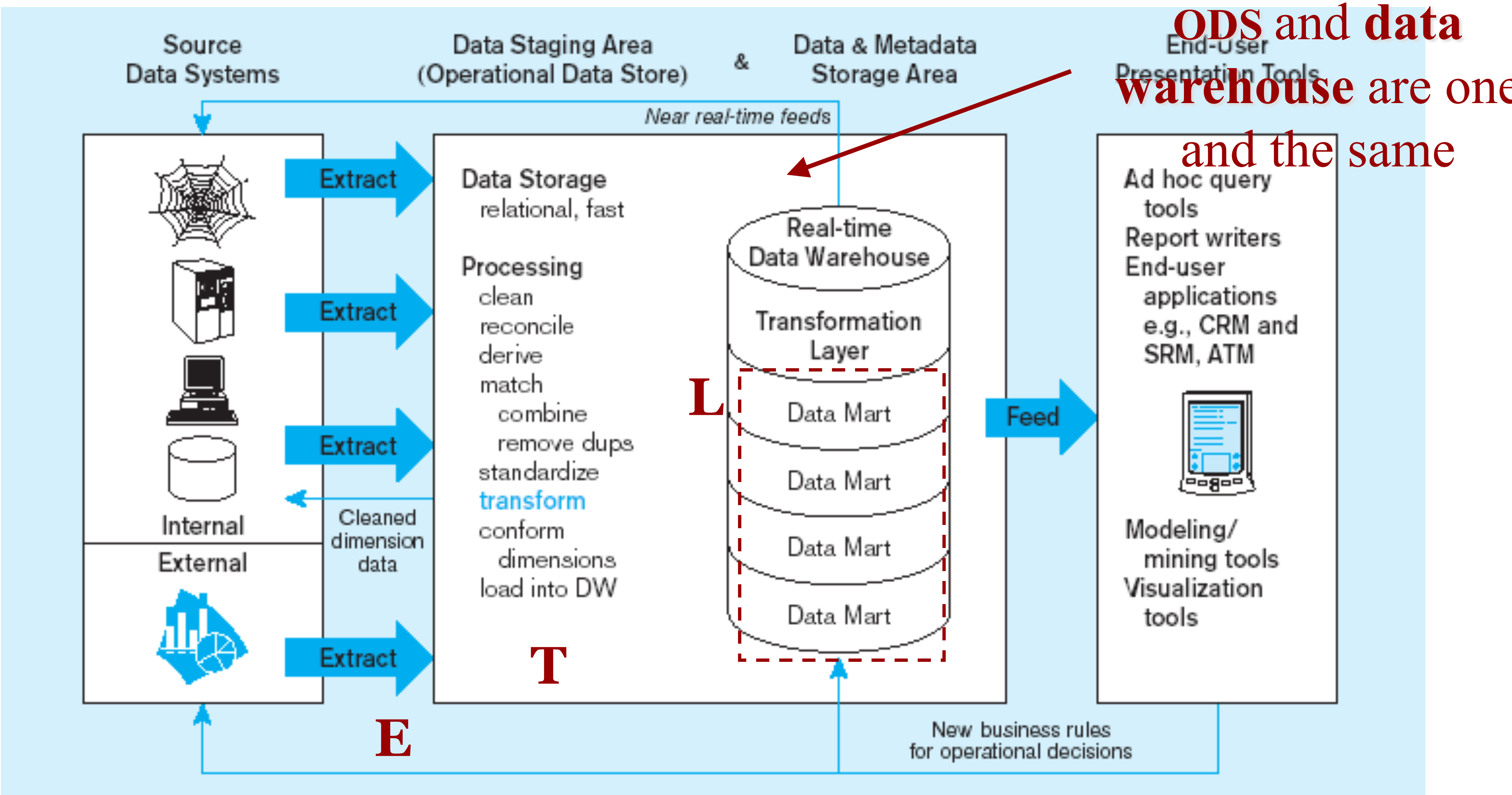


Simpler data access

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

Dependent data marts loaded from EDW

Logical data mart and real time warehouse architecture

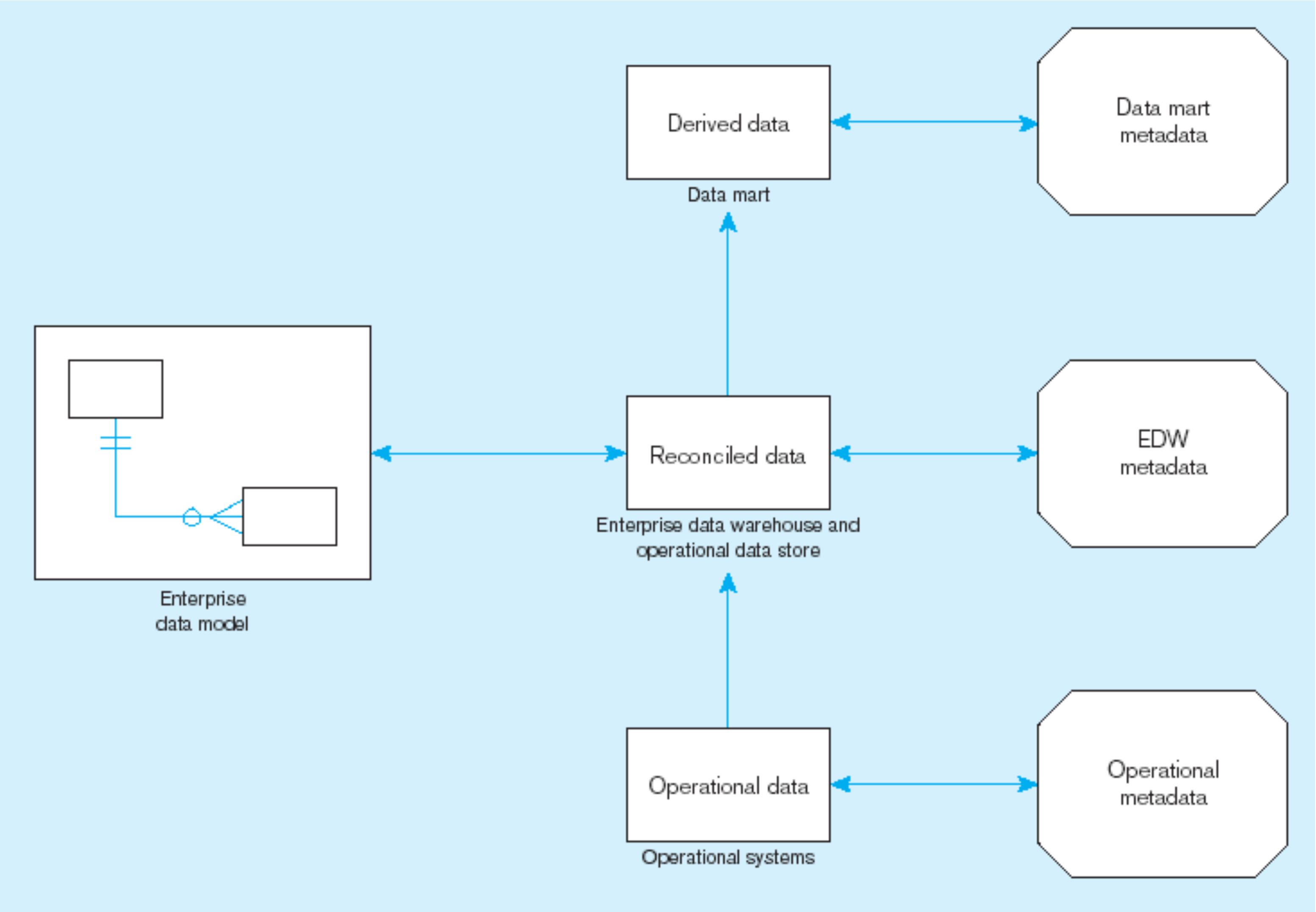


Near real-time ETL for ***Data Warehouse***

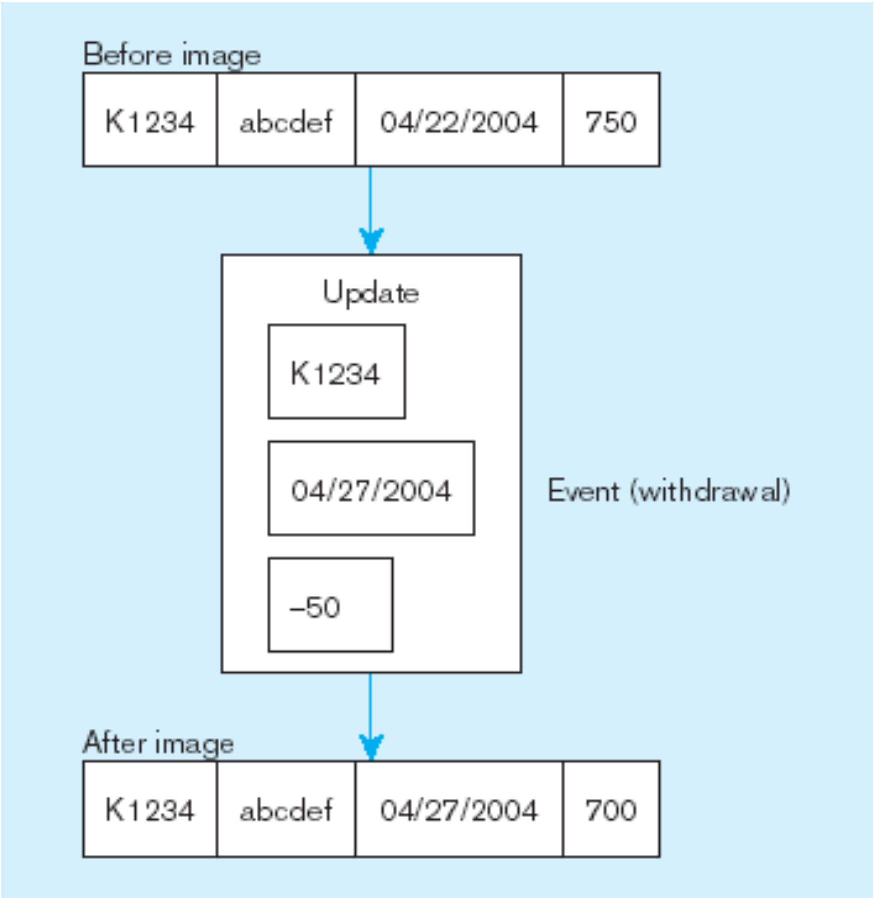
Data marts are NOT separate databases, but logical *views* of the data warehouse

➔ Easier to create new data marts

Three-layer data architecture for a data warehouse



Data Characteristics
Status vs. Event Data



Status

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

Status

Data Characteristics
Transient vs. Periodic Data

Table X (10/05)

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

Table X (10/06)

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

Table X (10/07)

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

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

Data Characteristics
Transient vs. Periodic Data

Table X (10/05)

Key	Date	A	B	Action
001	10/03	a	b	C
002	10/03	c	d	C
003	10/03	e	f	C
004	10/03	g	h	C

Table X (10/06)

Key	Date	A	B	Action
001	10/05	a	b	C
002	10/05	c	d	C
002	10/06	r	d	U
003	10/05	e	f	C
004	10/05	g	h	C
004	10/06	y	h	U
005	10/06	m	n	C

Table X (10/07)

Key	Date	A	B	Action
001	10/05	a	b	C
002	10/05	c	d	C
002	10/06	r	d	U
003	10/05	e	f	C
003	10/07	e	t	U
004	10/05	g	h	C
004	10/06	y	h	U
004	10/07	y	h	D
005	10/06	m	n	C

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

Other Data Warehouse Changes

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

The Reconciled Data Layer

- 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
  - Timely–data should be current enough to assist decision-making
  - Quality controlled–accurate with full integrity

Types of Data

- Business Data - *represents meaning*
  - Real-time data (ultimate source of all business data)
  - Reconciled data
  - Derived data
- Metadata - *describes meaning*
  - Build-time metadata
  - Control metadata
  - Usage metadata
- Data as a product* - *intrinsic meaning*
  - Produced and stored for its own intrinsic value
  - e.g., the contents of a text-book

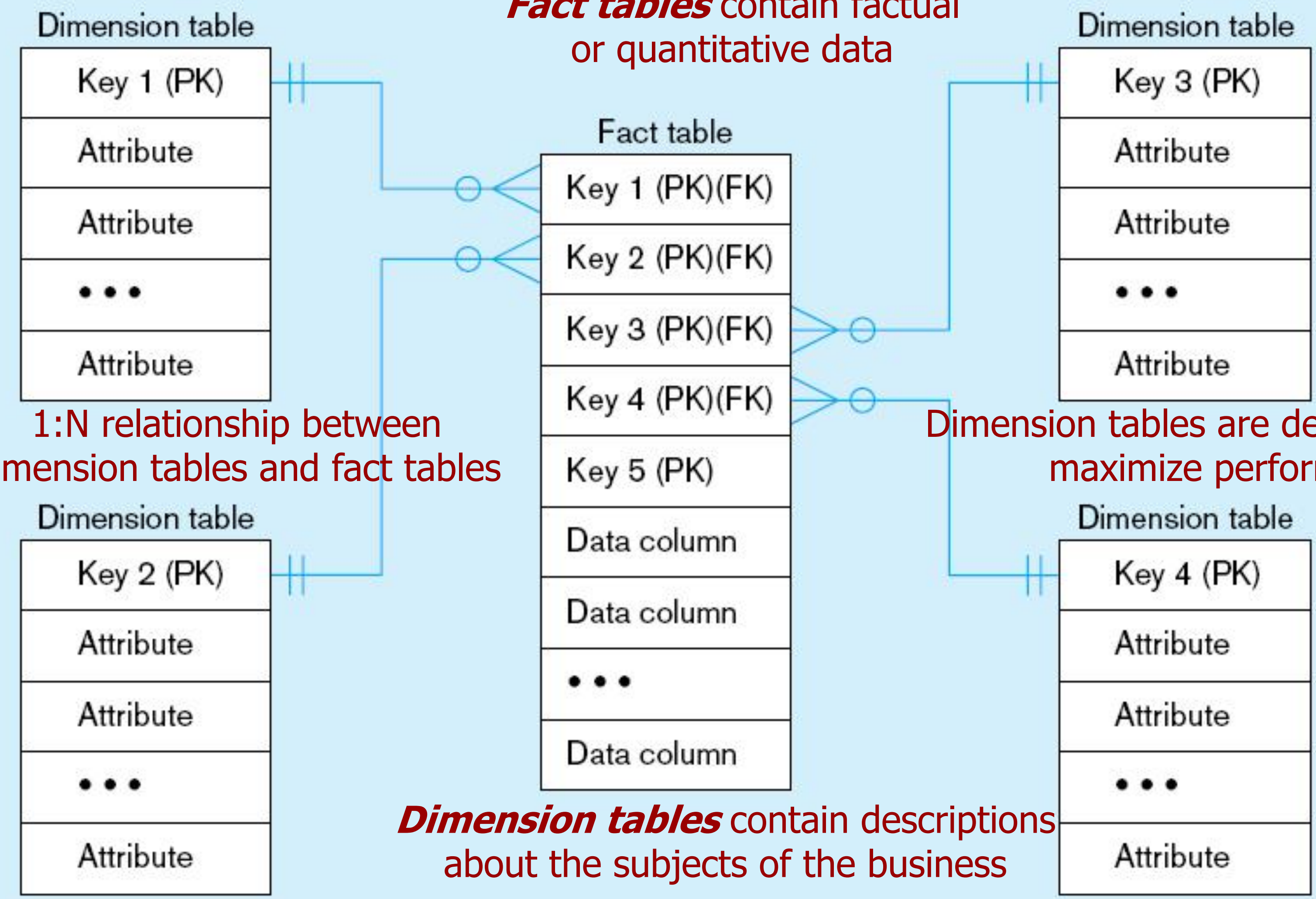
Slide credit: J. Hammer

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**)

Components of a **star schema**



Fact tables contain factual or quantitative data

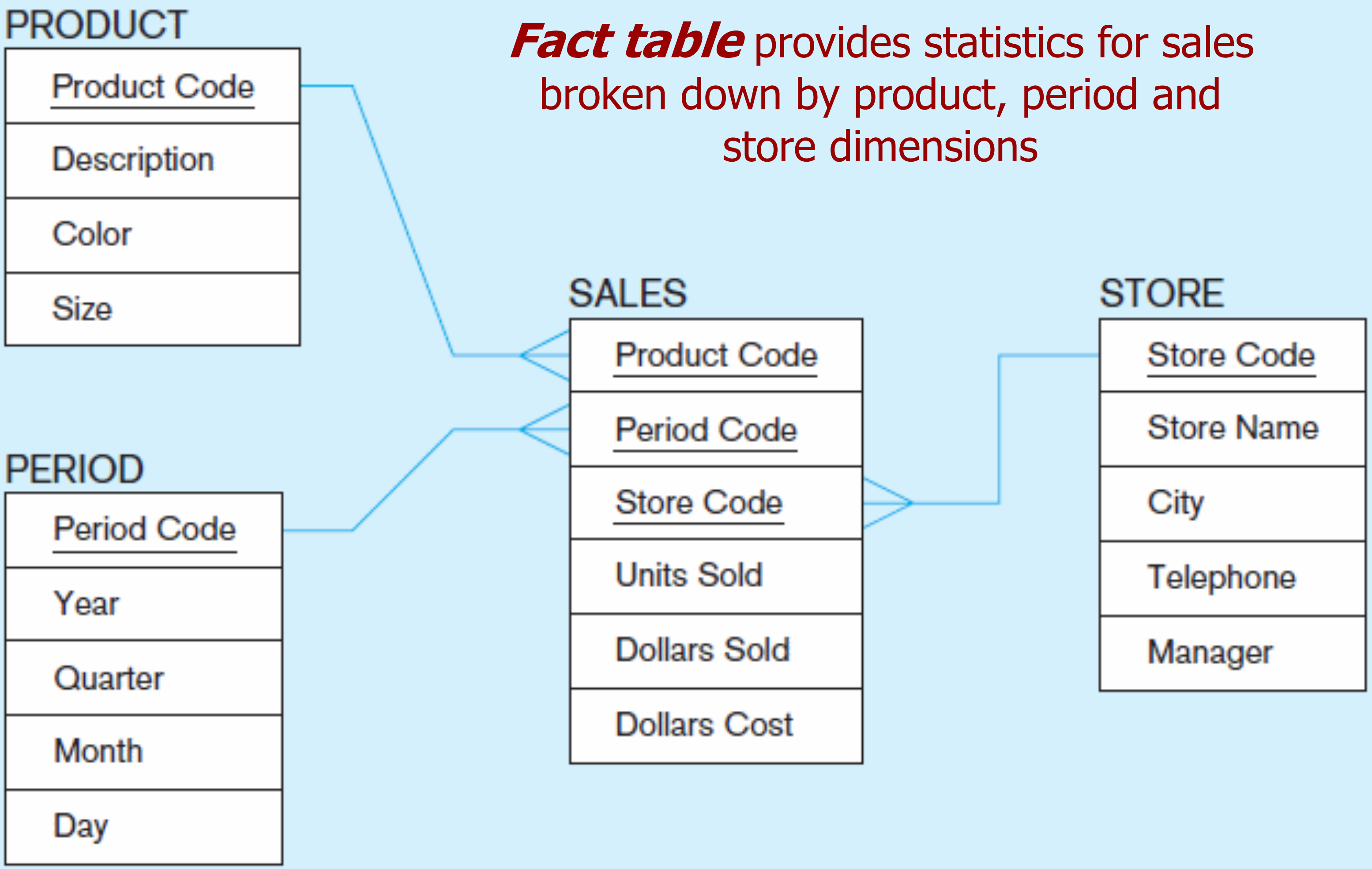
1:N relationship between dimension tables and fact tables

Dimension tables are denormalized to maximize performance

Dimension tables contain descriptions about the subjects of the business

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

Star schema example



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

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
• • •				

Surrogate 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

Grain of the Fact Table

- 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
  - In Web-based commerce, finest granularity is a click

Duration of the Database

- Natural duration–13 months or 5 quarters
- Financial institutions may need longer duration
- Older data is more difficult to source and cleanse

Size of Fact Table

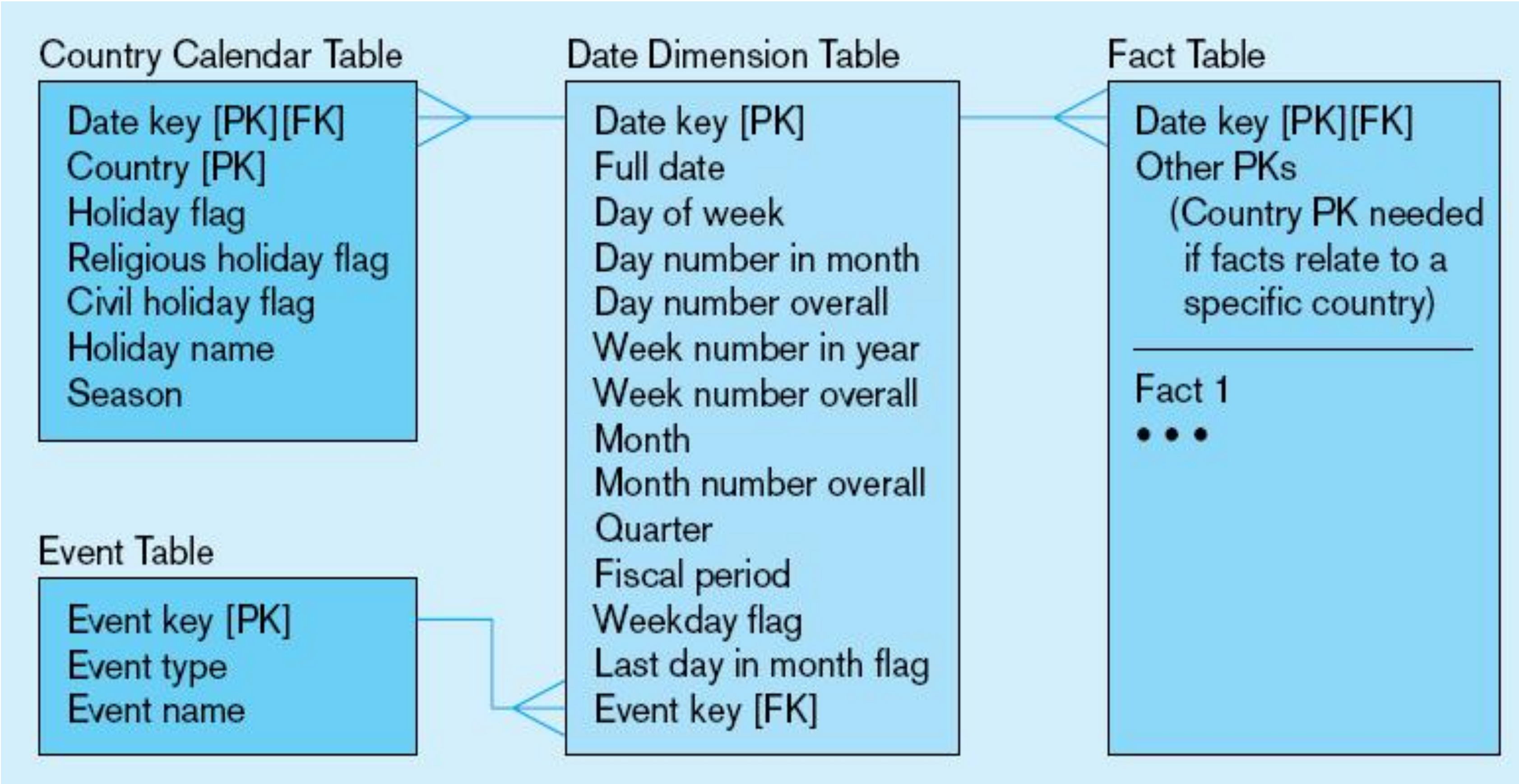
- Depends on the number of dimensions and the grain of the fact table
- Number of rows = product of number of possible values for each dimension associated with the fact table
- Example: assume the following:

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 (!)

Modeling Dates



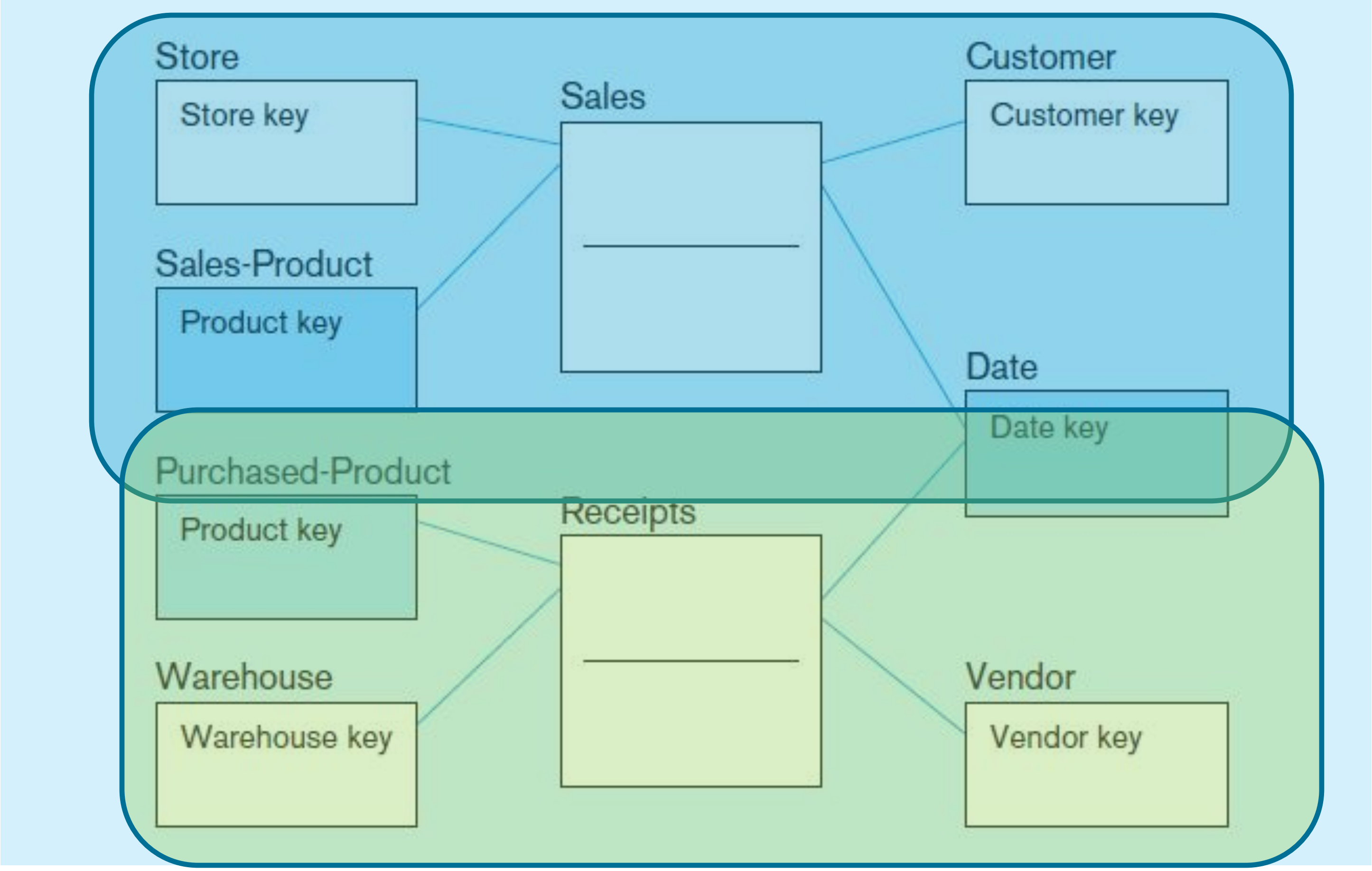
Fact tables contain time-period data
➔ Date dimensions are important

Variations of the Star Schema

- Multiple Facts Tables
  - Can improve performance
  - Often used to store facts for different combinations of dimensions
  - Conformed 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

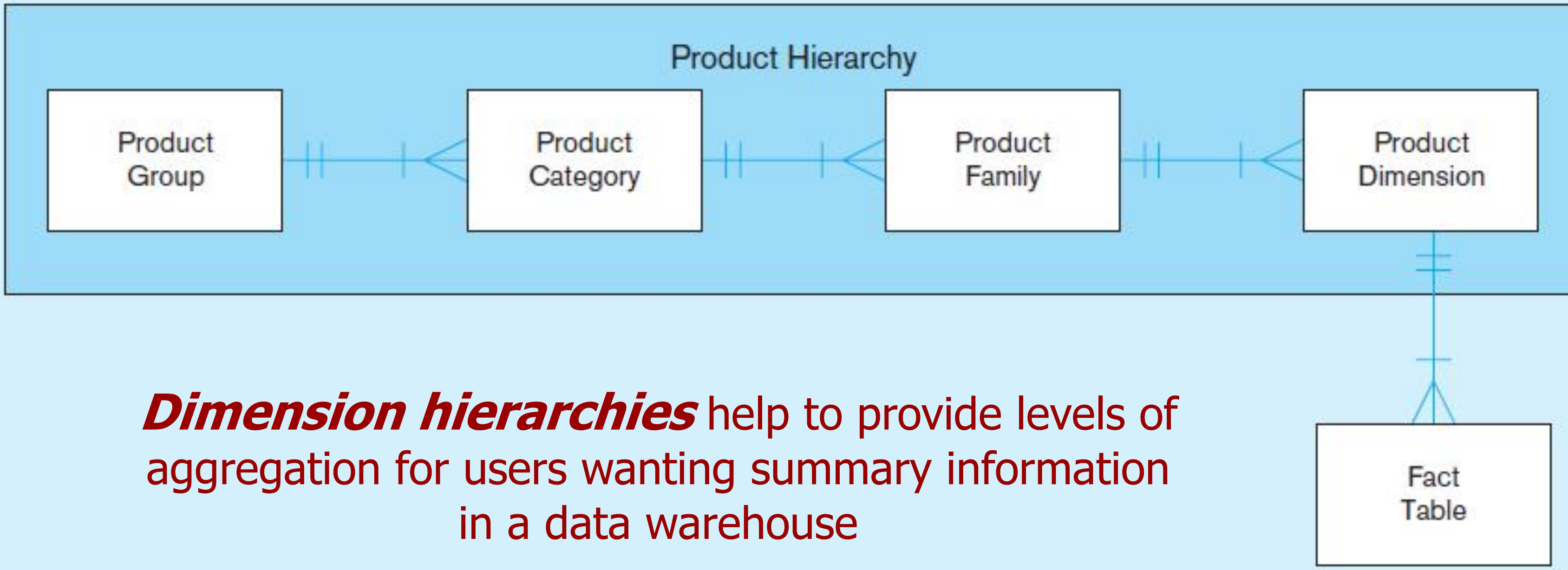
Conformed dimensions

Two fact tables ➔ two (connected) start schemas.



Conformed dimension
Associated with multiple fact tables

Fixed product hierarchy

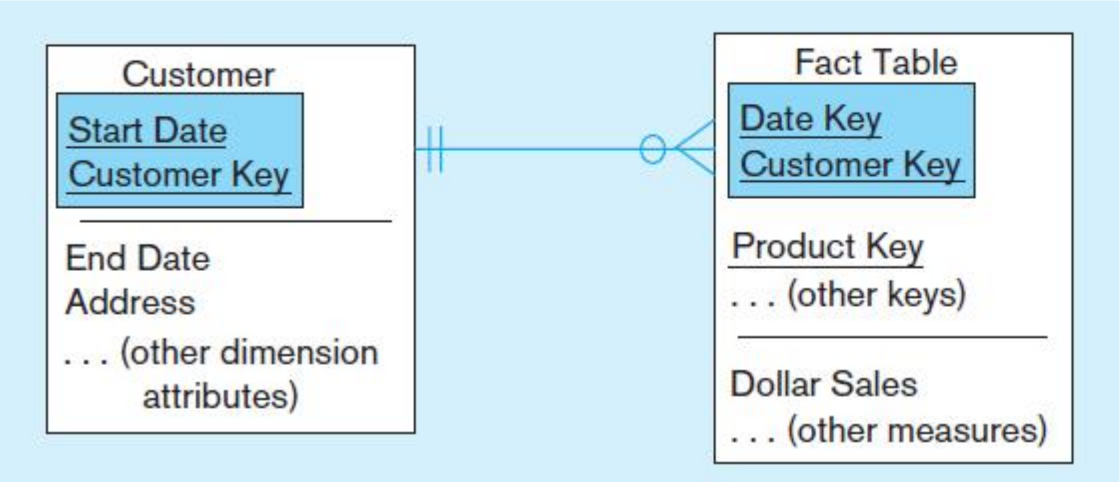


Dimension hierarchies help to provide levels of aggregation for users wanting summary information in a data warehouse

Slowly Changing Dimensions (SCD)

- How to maintain knowledge of the past
- Kimble's approaches:
  - Type 1: just replace old data with new (lose historical data)
  - Type 3: for each changing attribute, create a current value field and several old-valued fields (multivalued)
  - Type 2: create a new dimension table row each time the dimension object changes, with all dimension characteristics at the time of change. Most common approach.

Example of Type 2 SCD Customer dimension table



The dimension table contains several records for the same customer. The specific customer record to use depends on the key and the date of the fact, which should be between start and end dates of the SCD customer record.

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 User Interface

- 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

Figure 9-19
Example of drill-down

Starting with summary data, users can obtain details for particular cells

a) Summary report

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

b) Drill-down with color attribute 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 Warehousing: Two Distinct Issues

- (1) How to get information into warehouse
  - “Data warehousing”
- (2) What to do with data once it’s in warehouse
  - “Warehouse DBMS”
- Both rich research areas
- Industry has focused on (2)

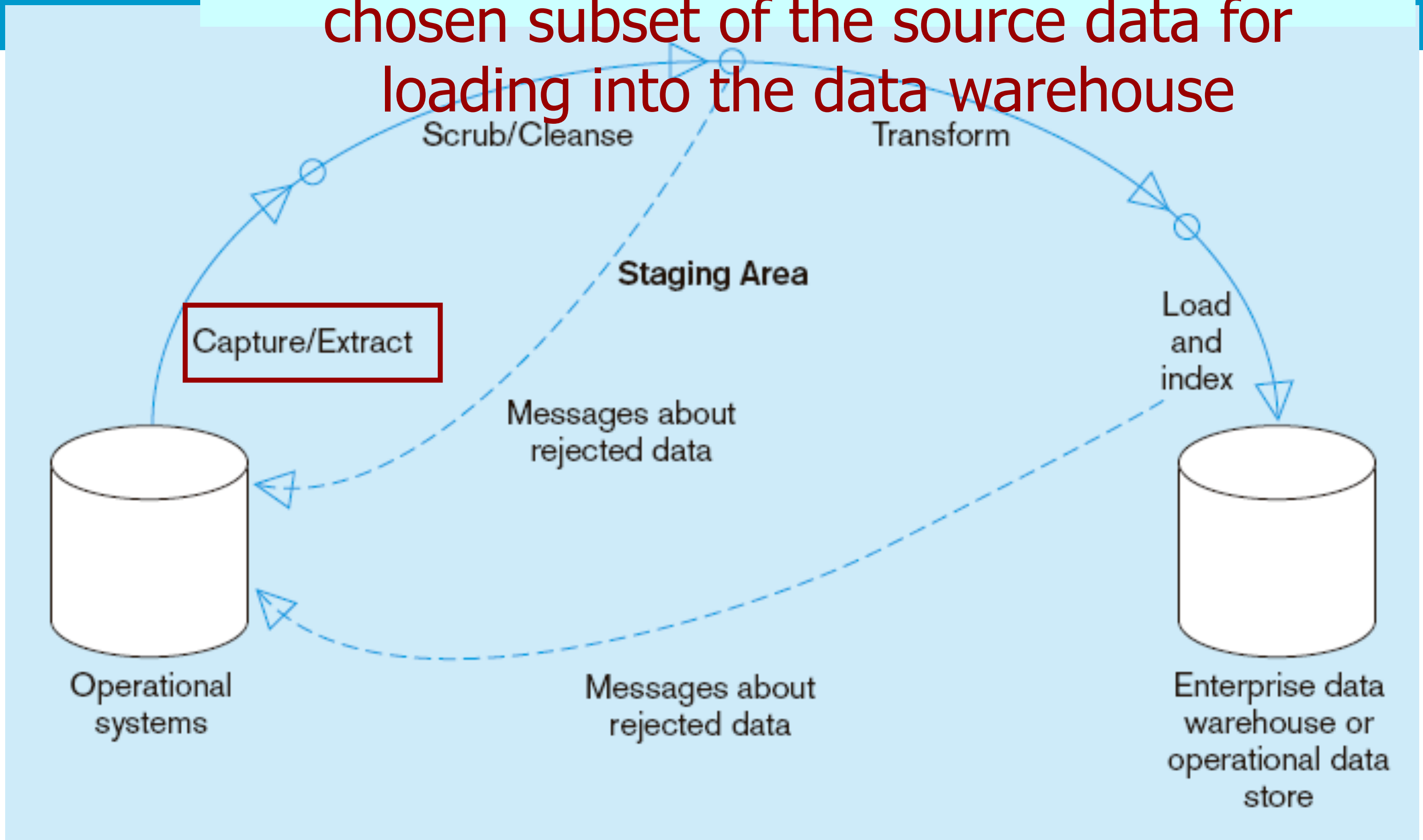
Slide credit: J. Hammer

The ETL Process

- Capture/Extract
- Scrub or data cleansing
- Transform
- Load and Index

ETL = Extract, transform, and load

Capture/Extract…obtaining a snapshot of a chosen subset of the source data for loading into the data warehouse



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

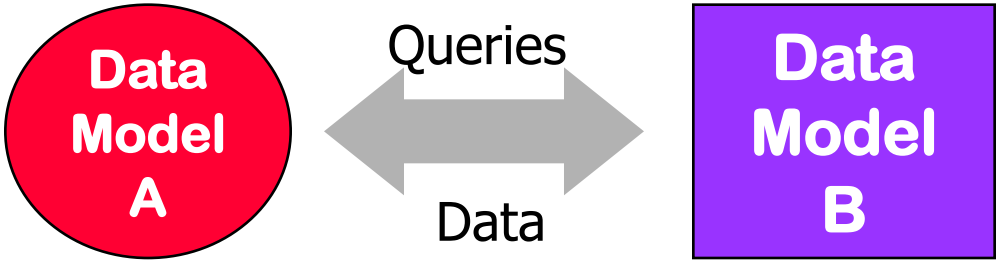
Data Extraction

- Source types
  - Relational, flat file, WWW, etc.
- How to get data out?
  - Replication tool
  - Dump file
  - Create report
  - ODBC or third-party “wrappers”

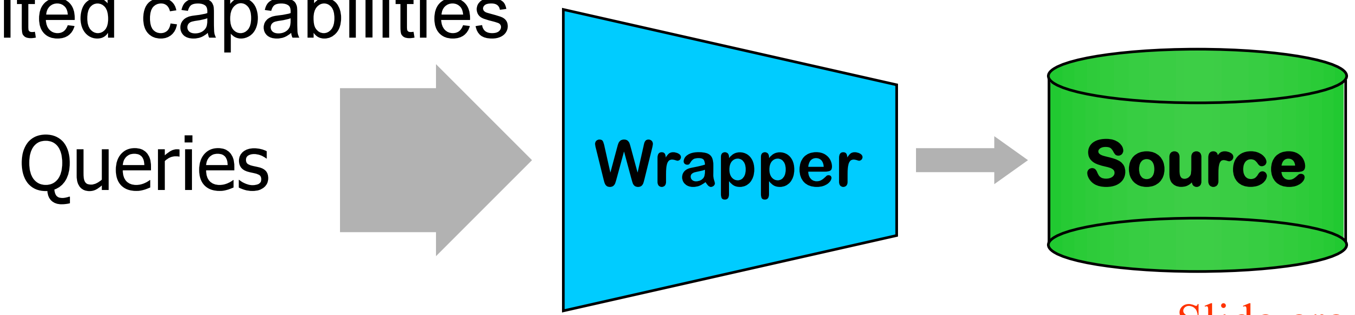
Slide credit: J. Hammer

Wrapper

- Converts data and queries from one data model to another



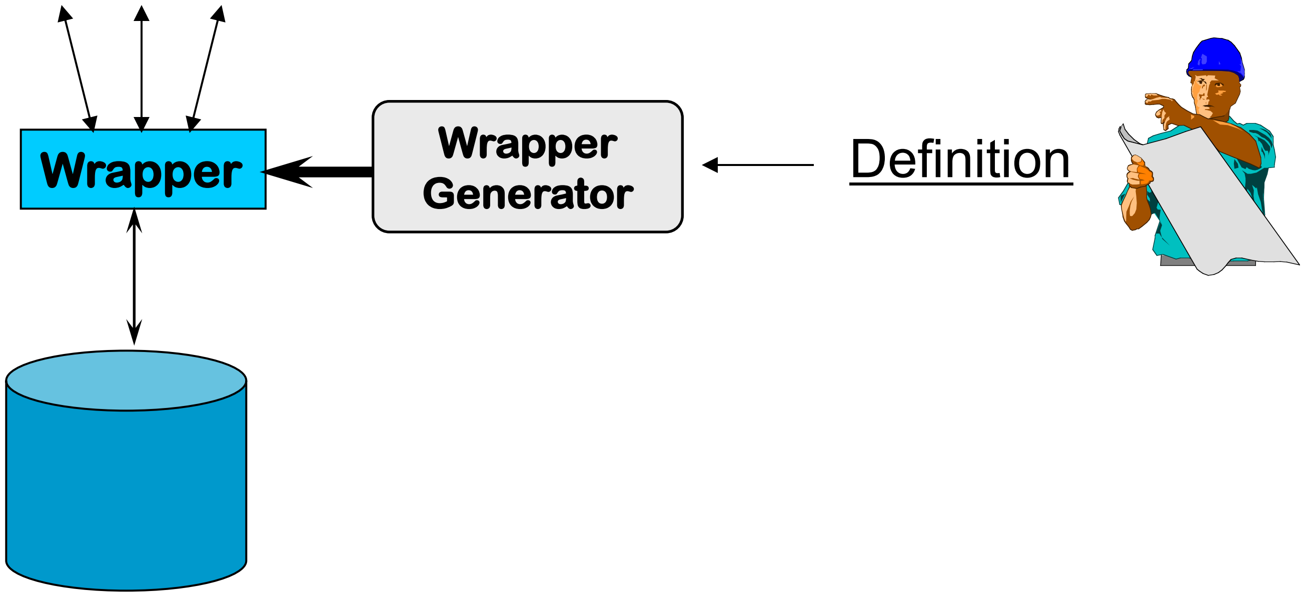
- Extends query capabilities for sources with limited capabilities



Slide credit: J. Hammer

Wrapper Generation

- Solution 1: Hard code for each source
- Solution 2: Automatic wrapper generation



Slide credit: J. Hammer

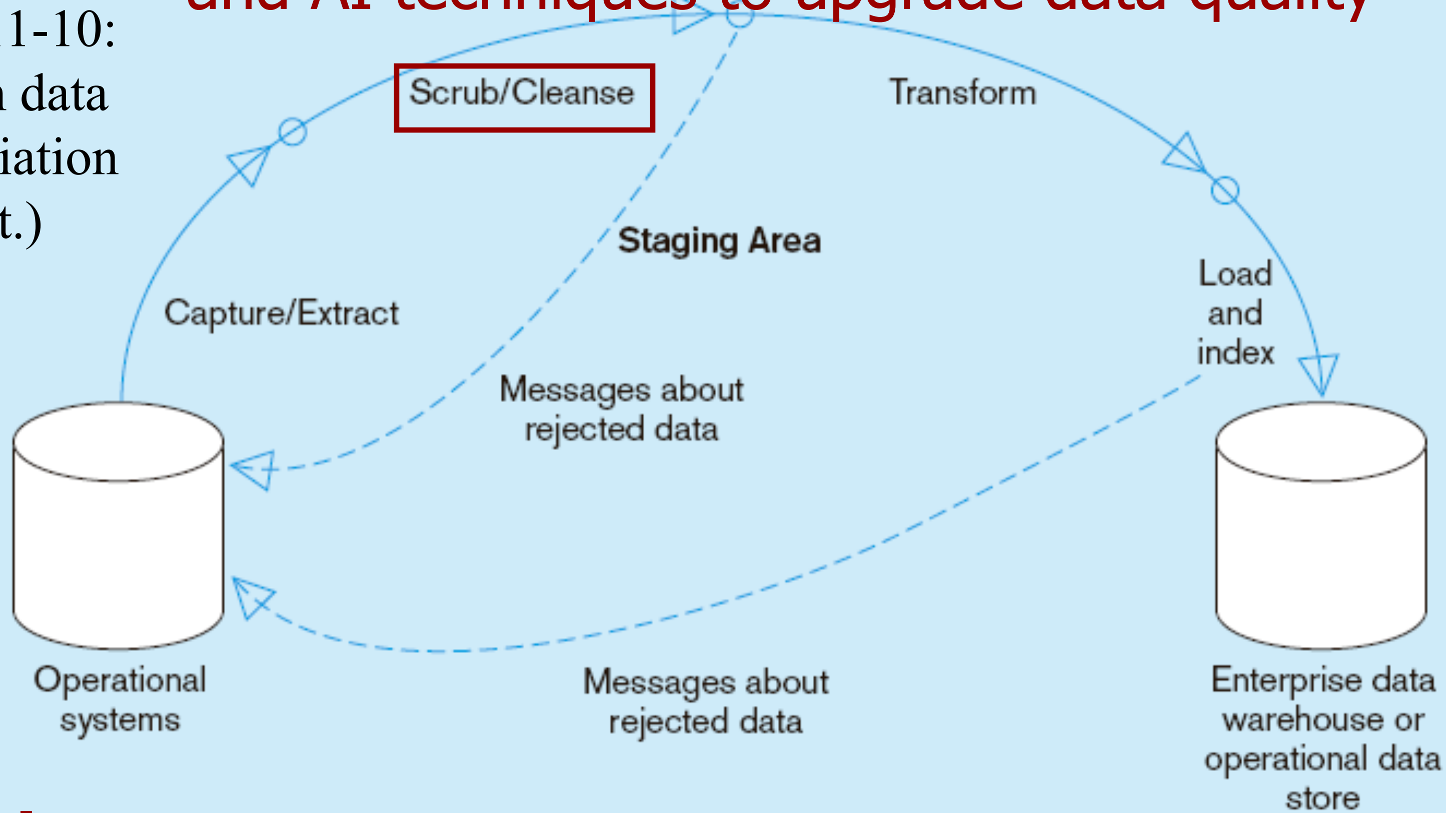
Monitors

- Goal: Detect changes of interest and propagate to integrator
- How?
  - Triggers
  - Replication server
  - Log sniffer
  - Compare query results
  - Compare snapshots/dumps

Slide credit: J. Hammer

Scrub/Cleanse…uses pattern recognition and AI techniques to upgrade data quality

Figure 11-10: Steps in data reconciliation (cont.)



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

New approaches for Data Cleansing

- It is generally been found that 70-90 percent of the time and effort in large data management and analysis tasks is taken up with data cleansing
- New tool “Data Wrangler” from Stanford and Berkeley CS folks
- http://vis.stanford.edu/wrangler/

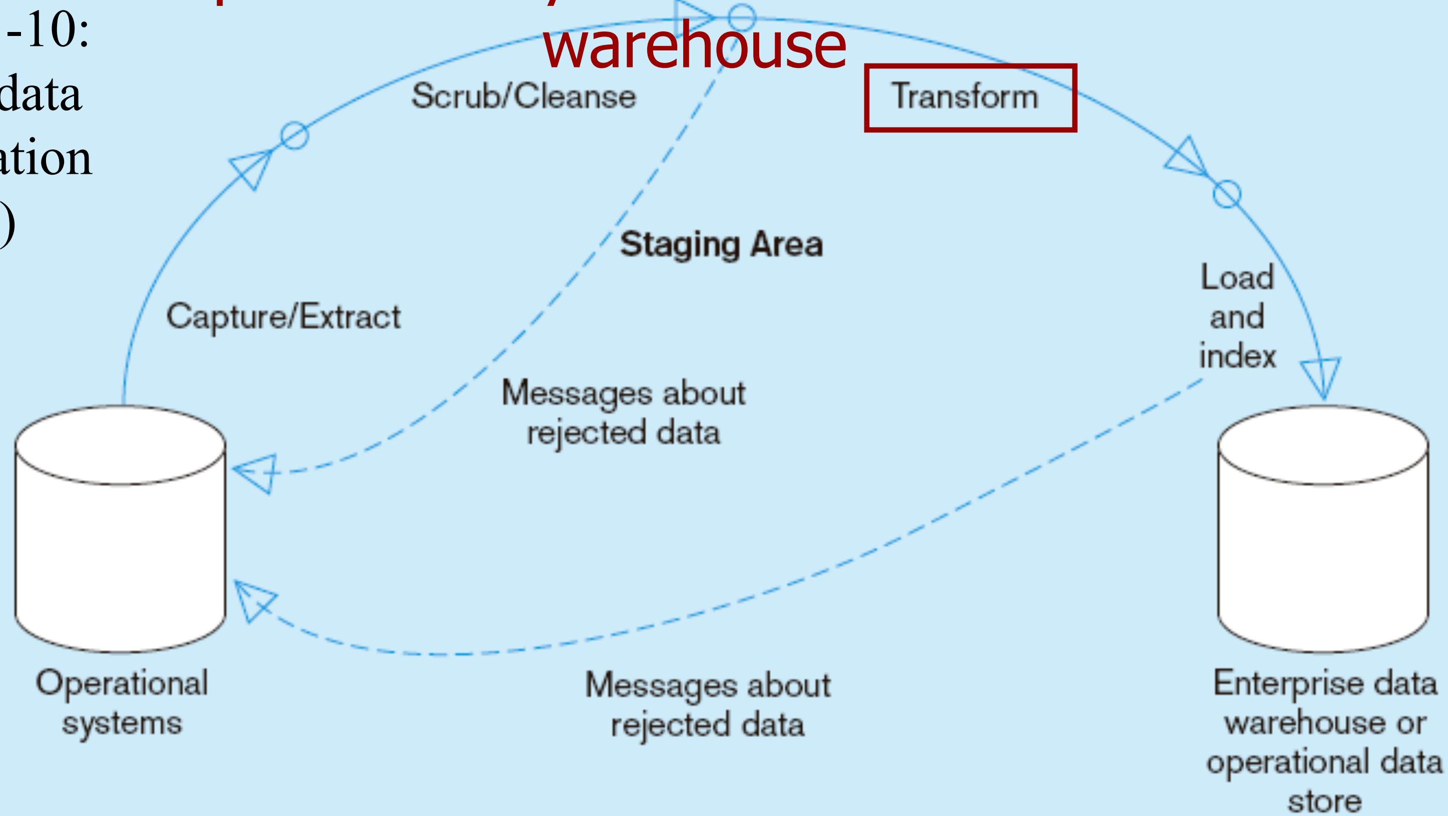
Data Cleansing

- Find (& remove) duplicate tuples
  - e.g., Jane Doe vs. Jane Q. Doe
- Detect inconsistent, wrong data
  - Attribute values that don't match
- Patch missing, unreadable data
- Notify sources of errors found

Slide credit: J. Hammer

Transform = convert data from format of operational system to format of data warehouse

Figure 11-10: Steps in data reconciliation (cont.)



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

Data Transformations

- Convert data to uniform format
  - Byte ordering, string termination
  - Internal layout
- Remove, add & reorder attributes
  - Add key
  - Add data to get history
- Sort tuples

Slide credit: J. Hammer

Load/Index= place transformed data into the warehouse and create indexes

Figure 11-10: Steps in data reconciliation (cont.)

Refresh mode: bulk rewriting of target data at periodic intervals

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

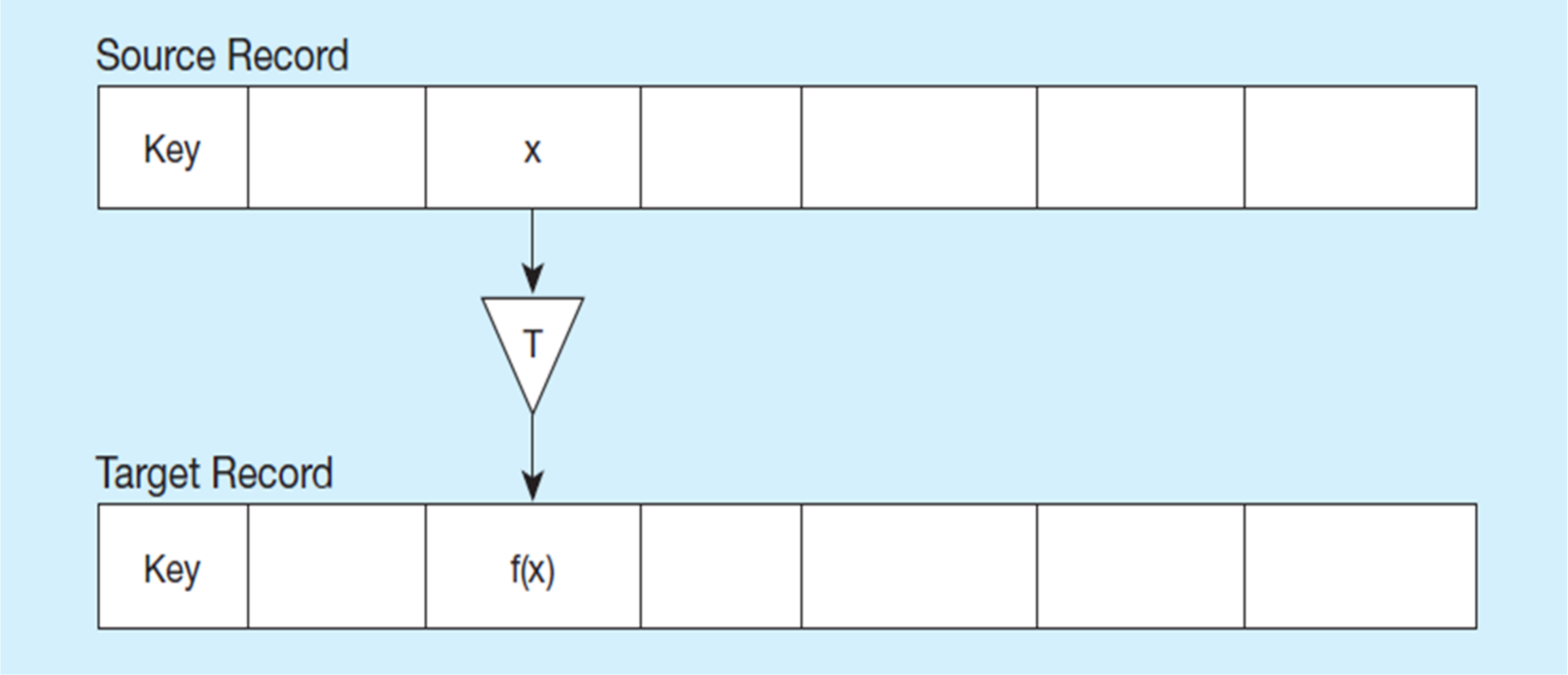
Data Integration

- Receive data (changes) from multiple wrappers/monitors and integrate into warehouse
- Rule-based
- Actions
  – Resolve inconsistencies
  – Eliminate duplicates
  – Integrate into warehouse (may not be empty)
  – Summarize data
  – Fetch more data from sources (wh updates)
  – etc.

Slide credit: J. Hammer

Single-Field Transformations (1 of 3)

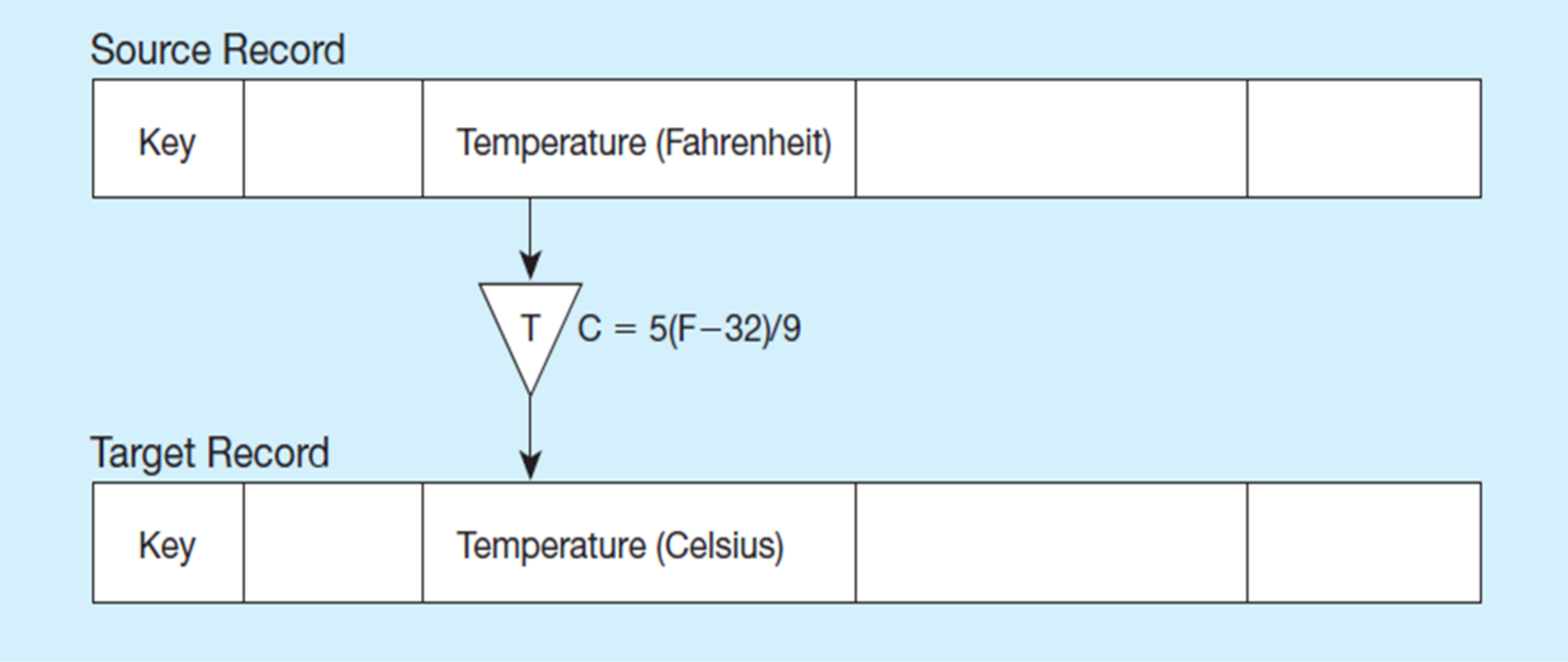
a) Basic representation



Single-Field Transformations (2 of 3)

b) Algorithmic

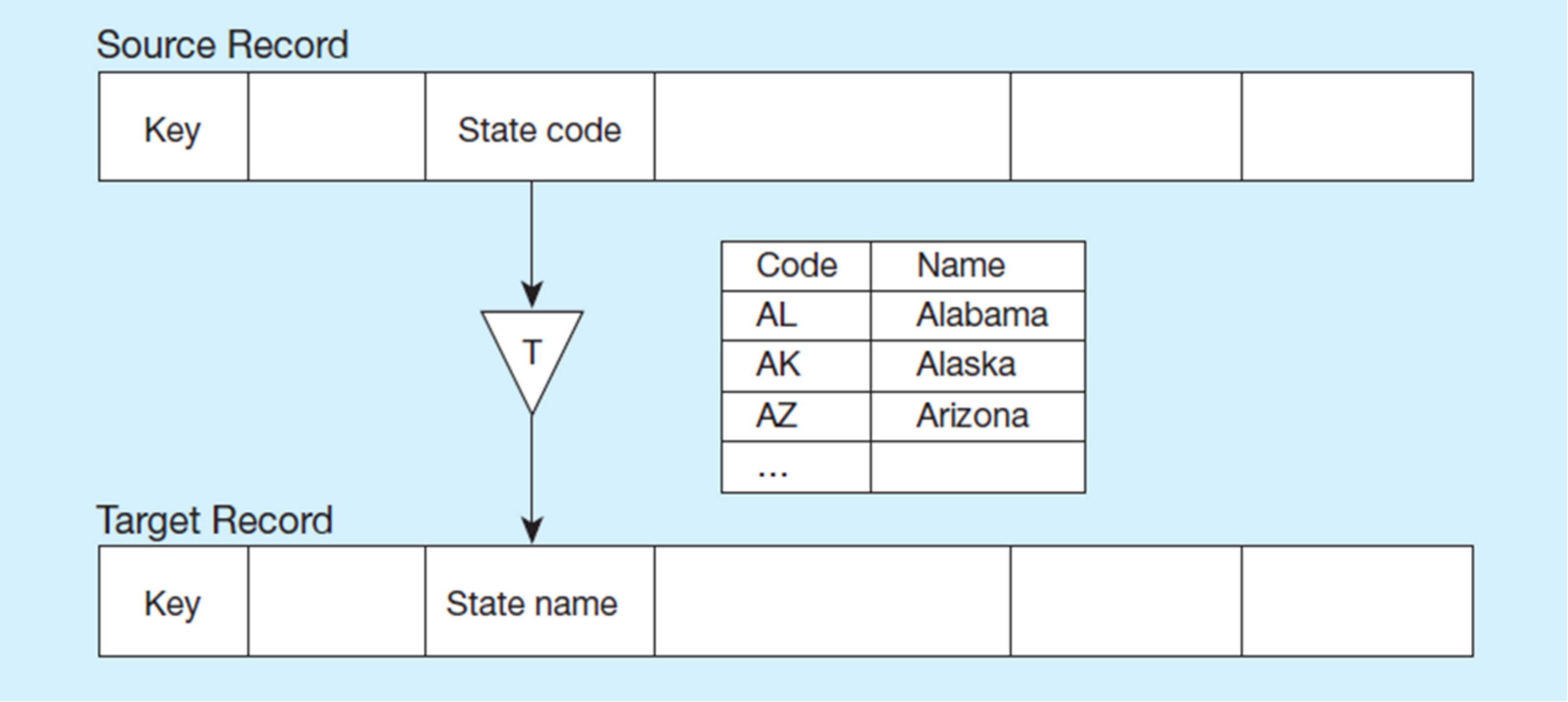
Uses a formula or logical expression



Single-Field Transformations (3 of 3)

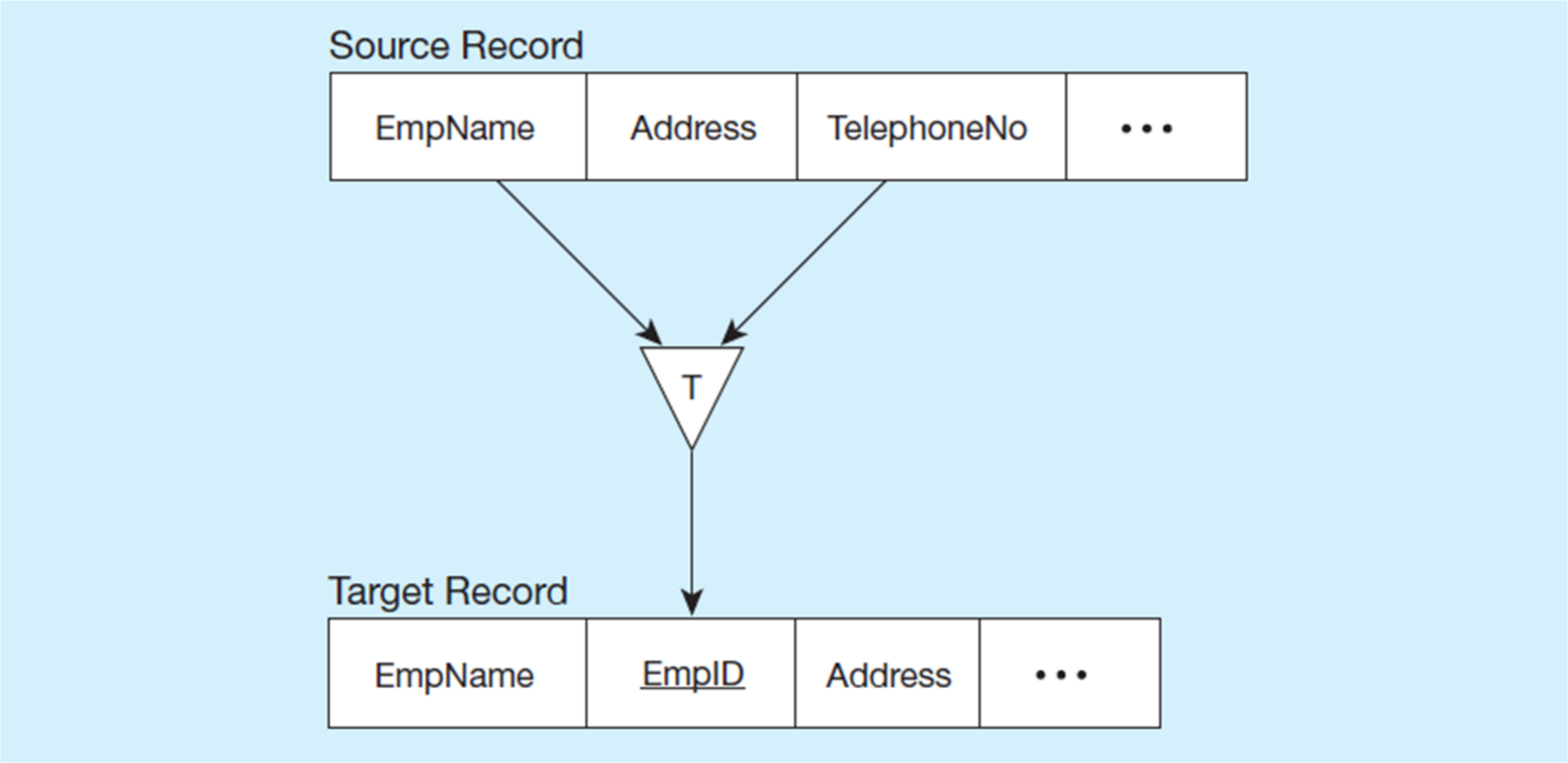
c) Table lookup

Uses a separate table keyed by source record code



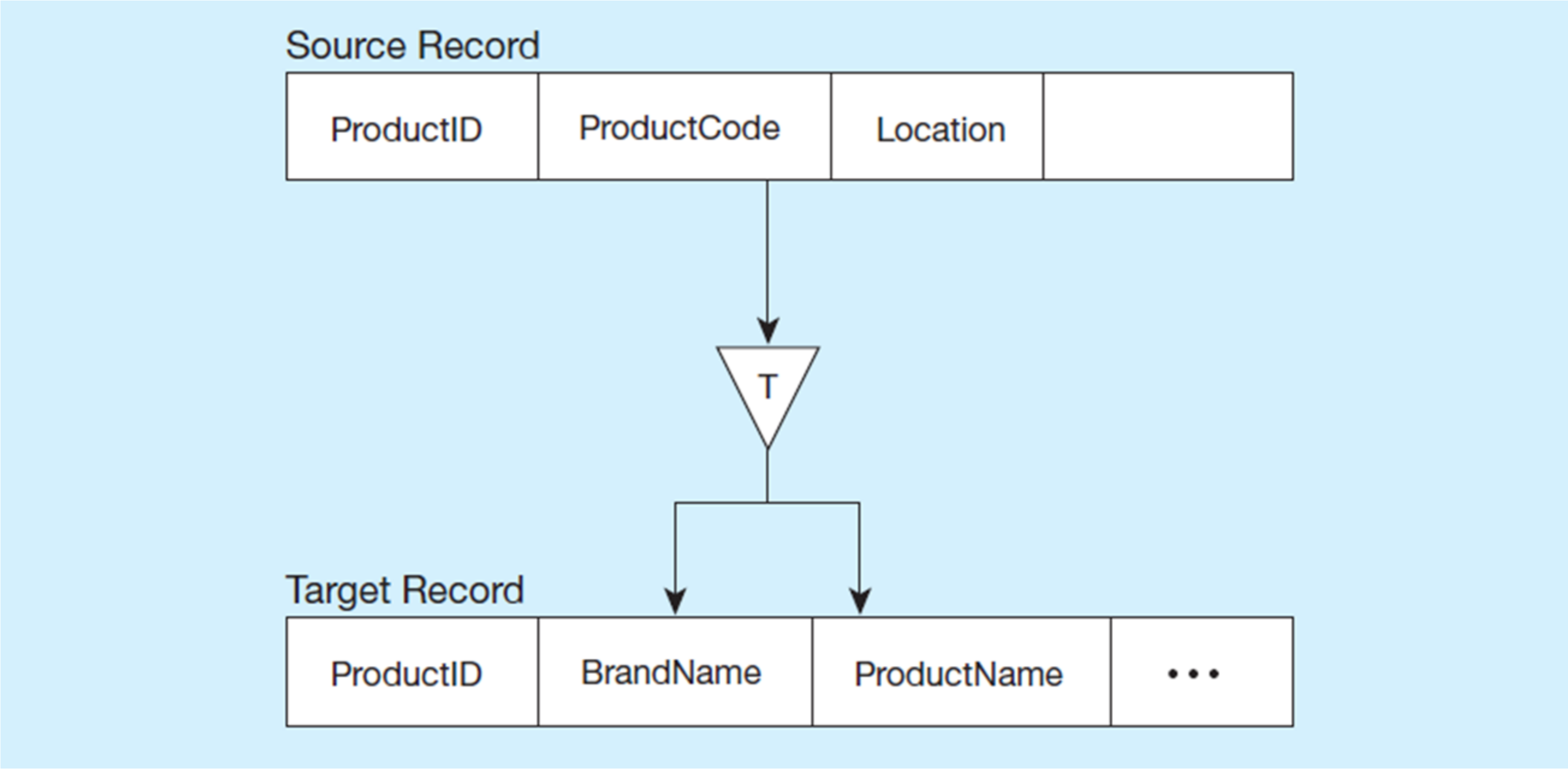
Multifield Transformations (1 of 2)

a) Many sources to one target



Multifield Transformations (2 of 2)

b) One source to many targets



Warehouse Maintenance

- Warehouse data ≈ materialized view
  - Initial loading
  - View maintenance
- View maintenance

Slide credit: J. Hammer

Differs from Conventional View Maintenance...

- Warehouses may be highly aggregated and summarized
- Warehouse views may be over history of base data
- Process large batch updates
- Schema may evolve

Slide credit: J. Hammer

Differs from Conventional View Maintenance…

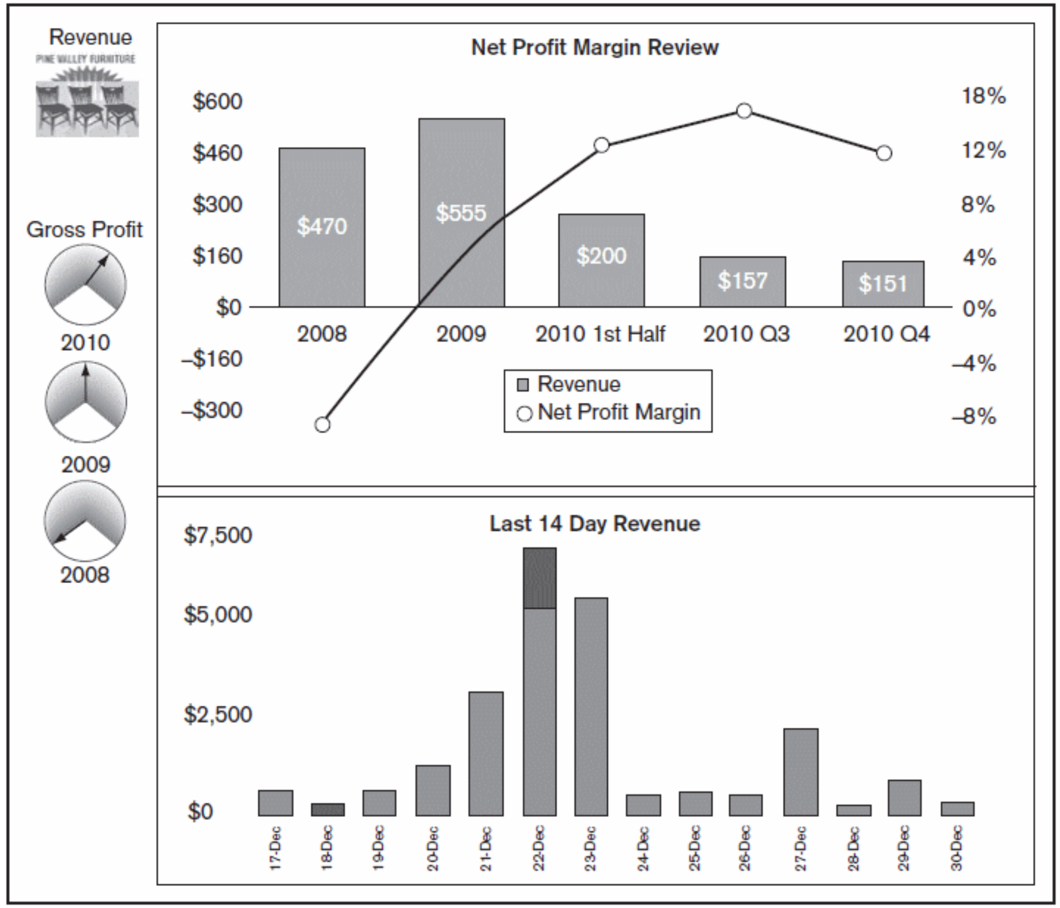
- Base data doesn't participate in view maintenance
  - Simply reports changes
  - Loosely coupled
  - Absence of locking, global transactions
  - May not be queriable

Slide credit: J. Hammer

Business Performance Mgmt (BPM)

Figure 9-22
Sample Dashboard

BPM systems allow managers to measure, monitor, and manage key activities and processes to achieve organizational goals. Dashboards are often used to provide an information system in support of BPM.



Charts like these are examples of **data visualization**, the representation of data in graphical and multimedia formats for human analysis.

Data Mining

- 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

TABLE 9-4 Data-Mining Techniques

Technique	Function
Regression	Test or discover relationships from historical data
Decision tree induction	Test or discover if . . . then rules for decision propensity
Clustering and signal processing	Discover subgroups or segments
Affinity	Discover strong mutual relationships
Sequence association	Discover cycles of events and behaviors
Case-based reasoning	Derive rules from real-world case examples
Rule discovery	Search for patterns and correlations in large data sets
Fractals	Compress large databases without losing information
Neural nets	Develop predictive models based on principles modeled after the human brain

Additional Research Issues

- Historical views of non-historical data
- Expiring outdated information
- Crash recovery
- Addition and removal of information sources
  - Schema evolution

Slide credit: J. Hammer

Warehousing and Industry

- Data Warehousing is big business
  - $2 billion in 1995
  - $3.5 billion in early 1997
  - Predicted: $8 billion in 1998 [Metagroup]
- Wal-Mart said to have the largest warehouse
  - 1000-CPU, 583 Terabyte, Teradata system (InformationWeek, Jan 9, 2006)
  - "Half a Petabyte" in warehouse (Ziff Davis Internet, October 13, 2004)
  - 1 billion rows of data or more are updated *every day* (InformationWeek, Jan 9, 2006)
  - Reported to be 2.5 Petabytes in 2008
    - http://gigaom.com/2013/03/27/why-apple-ebay-and-walmart-have-some-of-the-biggest-data-warehouses-youve-ever-seen

Those are small change today…

- Some databases are larger, however…
  - eBay: has two Teradata systems. Its primary data warehouse is 9.2 petabyes; its “singularity system” that stores web clicks and other “big” data is more than 40 petabytes. It includes a single table that’s 1 trillion rows. (2013)
    - http://gigaom.com/2013/03/27/why-apple-ebay-and-walmart-have-some-of-the-biggest-data-warehouses-youve-ever-seen
  - Apple: “Multiple Petabytes” in 2013
  - Yahoo! for web user behavioral analysis, storing two petabytes and claimed to be the largest data warehouse using a heavily modified version of PostgreSQL (Wikipedia 2012)

Largest Data Warehouses Today

The Guinness World Record Largest Data Warehouse was creating in the SAP/Intel shared lab in Santa Clara, California. The data warehouse is 12.1PB of data running on 25 HP ProLiant DL580 G7 servers with Intel processors on a Red Hat® Enterprise Linux® 6.4 X86-64 operating system using SAP HANA and SAP IQ 16 with BMMsoft Federated EDMT® 9. The server environment is connected to a SAN comprised of 20 NetApp E5460 storage arrays through HP 8 Gb/s Fibre switches.

http://global.sap.com/news-reader/index.epx?category=ALL&articleID=22468

More Information on DW

- Agosta, Lou, The Essential Guide to Data Warehousing. Prentise Hall PTR, 1999.
- Devlin, Barry, Data Warehouse, from Architecture to Implementation. Addison-Wesley, 1997.
- Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.
- Widom, J., "Research Problems in Data Warehousing." Proc. of the 4th Intl. CIKM Conf., 1995.
- Chaudhuri, S., Dayal, U., "An Overview of Data Warehousing and OLAP Technology." ACM SIGMOD Record, March 1997.