Disciplina de Gestão de Dados e de Bases de Dados

Ano Letivo 2013/2014

Data Warehousing

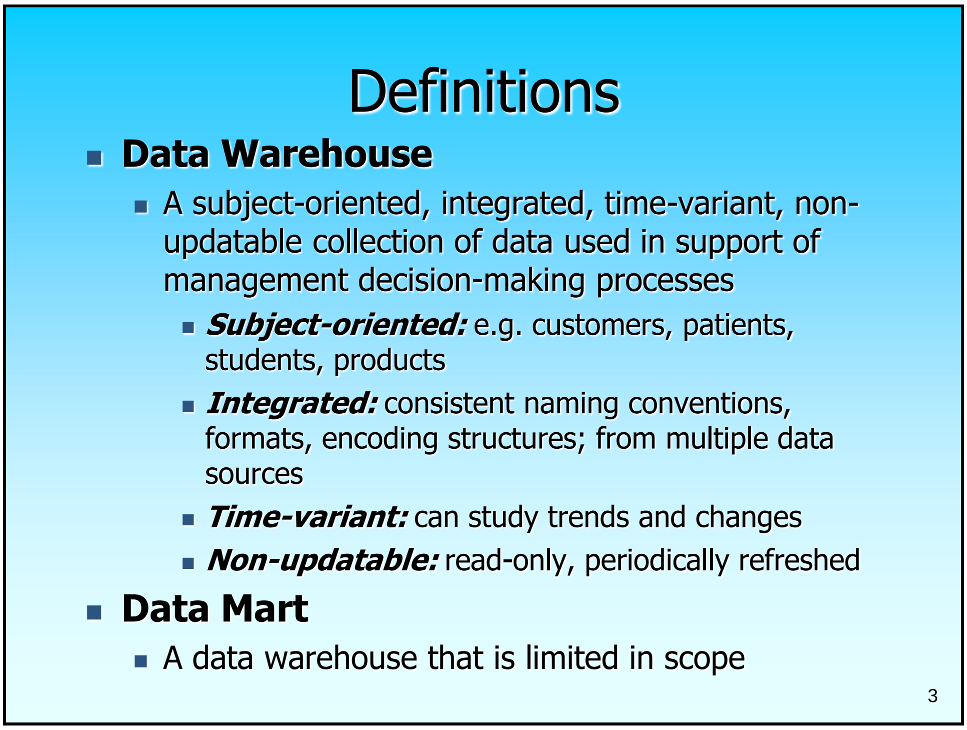
Parts of this presentation were taken from the backing material of the book

Modern Database Management, 11/E Edition, 2013

Jeffrey A. Hoffer, V. Ramesh, Heikki Topi

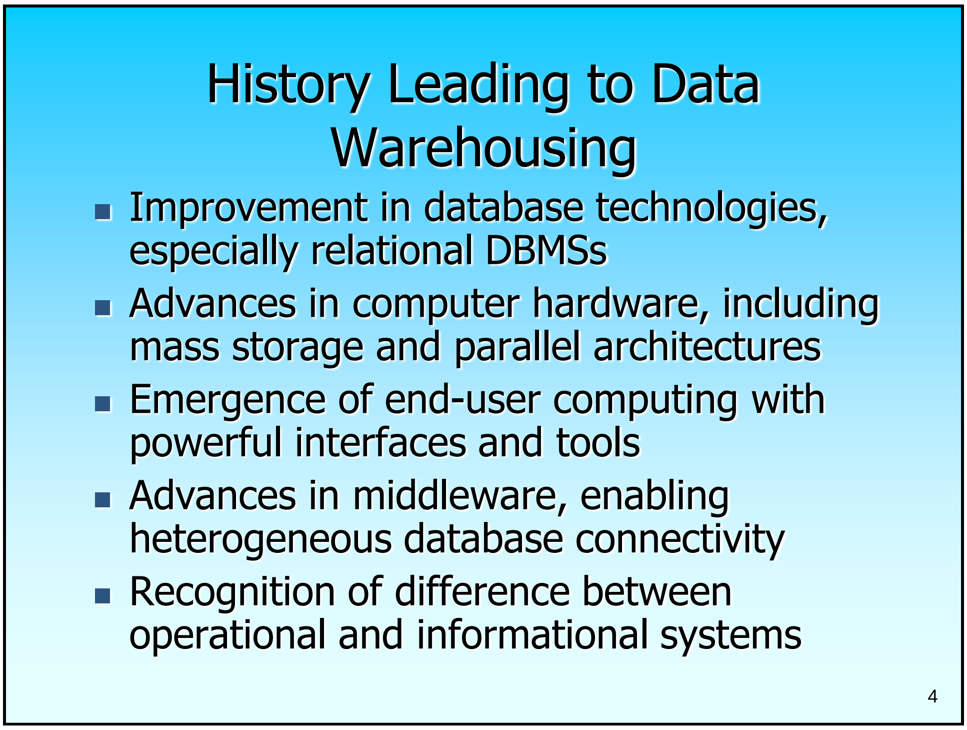
Objectives

- Define terms
- Explore reasons for information gap between information needs and availability
- Understand 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



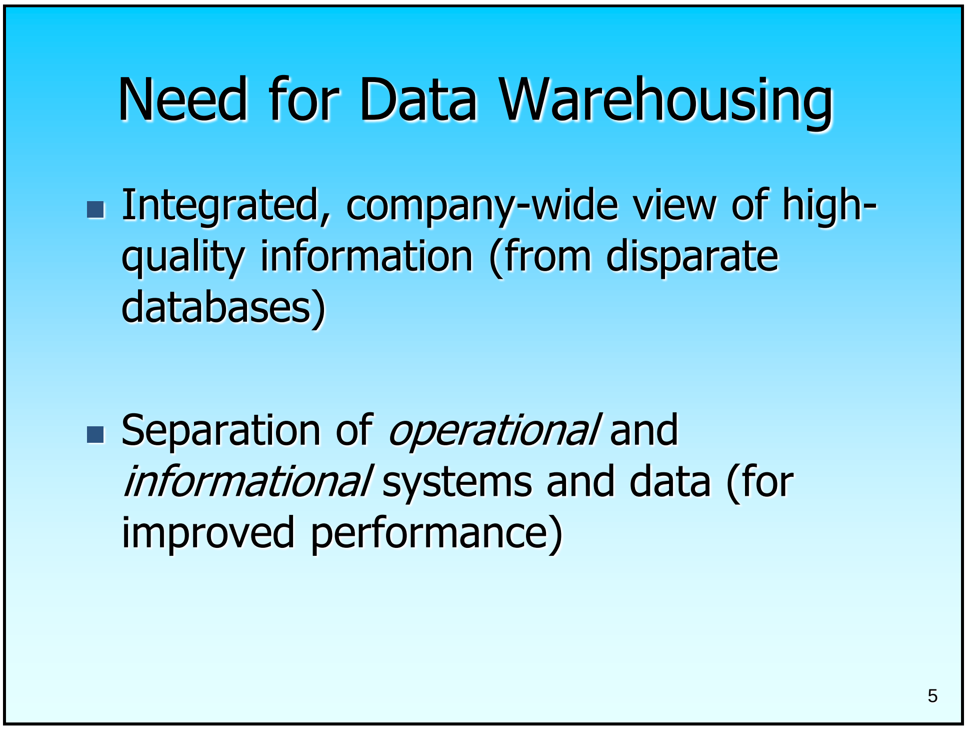
Definitions

- **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
    - ***Non-updatable:*** read-only, periodically refreshed
- **Data Mart**
  - A data warehouse that is limited in scope



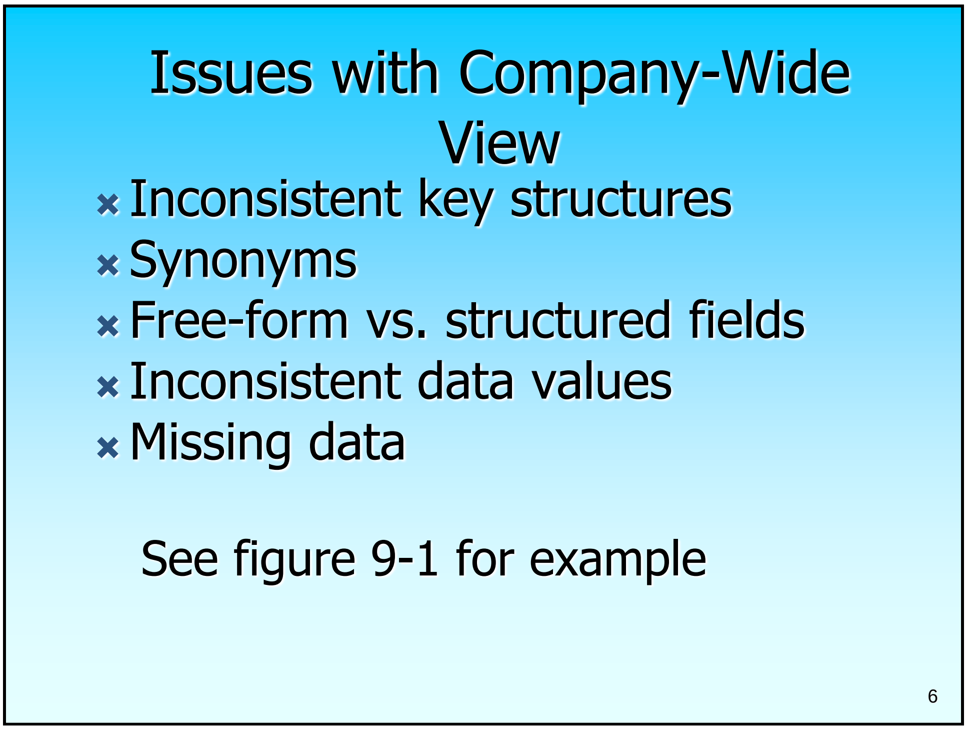
History Leading to Data Warehousing

- 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



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)

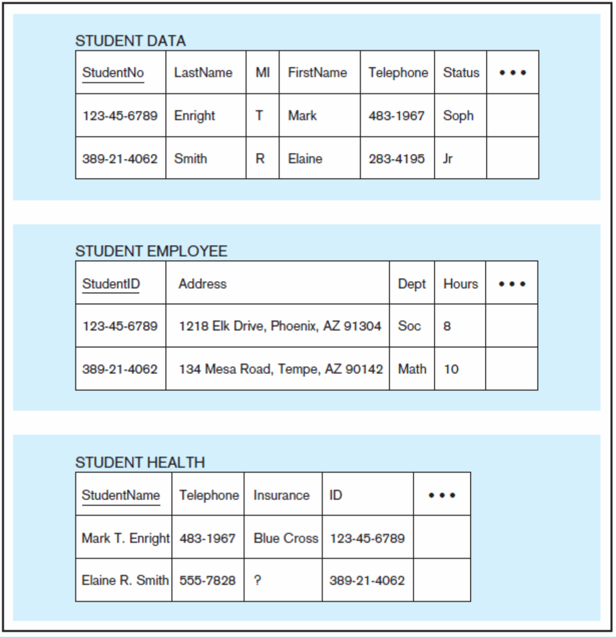


Issues with Company-Wide View

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

See figure 9-1 for example

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	

Organizational Trends Motivating Data Warehouses

- No single system of records
- Multiple systems not synchronized
- Organizational need to analyze activities in a balanced way
- Customer relationship management
- Supplier relationship management

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

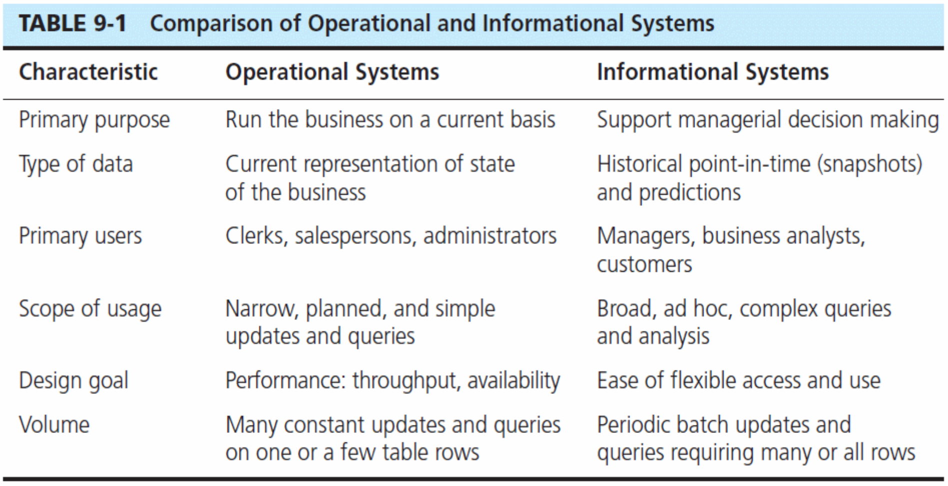


TABLE 9-1 Comparison of Operational and Informational Systems

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

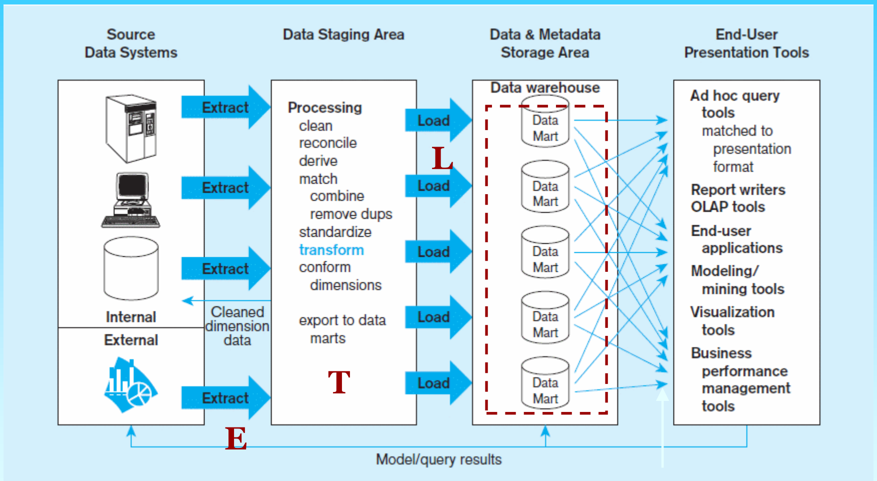
Data Warehouse Architectures

- 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

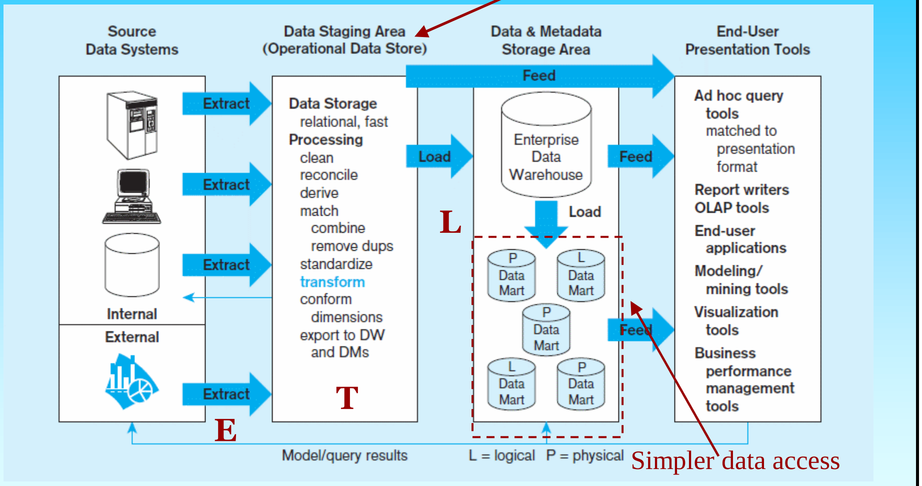


Separate ETL for each ***independent*** data mart

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

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

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



Source Data Systems

Data Staging Area (Operational Data Store)

Data & Metadata Storage Area

End-User Presentation Tools

Internal

External

Extract

Extract

Extract

Extract

E

Data Storage
relational, fast
Processing
clean
reconcile
derive
match
combine
remove dups
standardize
transform
conform
dimensions
export to DW
and DMs

T

Load

L

Feed

Enterprise Data Warehouse

Load

P Data Mart

L Data Mart

P Data Mart

L Data Mart

P Data Mart

Feed

Feed

Ad hoc query tools matched to presentation format
Report writers
OLAP tools
End-user applications
Modeling/ mining tools
Visualization tools
Business performance management tools

Model/query results

L = logical P = physical

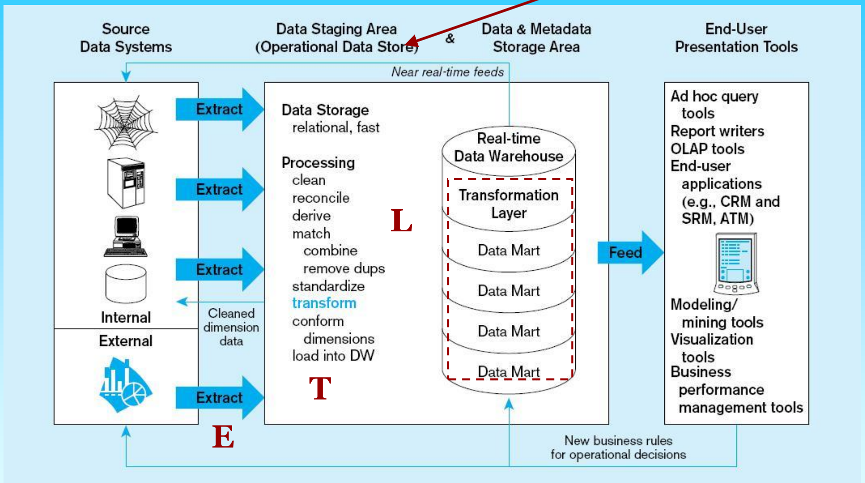
Simpler data access

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



Source Data Systems

Data Staging Area (Operational Data Store) & Data & Metadata Storage Area

End-User Presentation Tools

Near real-time feeds

Internal

External

Extract

Extract

Extract

Extract

E

Cleaned dimension data

Data Storage
relational, fast
Processing
clean
reconcile
derive
match
combine
remove dups
standardize
transform
conform
dimensions
load into DW

T

L

Real-time Data Warehouse

Transformation Layer

Data Mart

Data Mart

Data Mart

Data Mart

Feed

Ad hoc query tools
Report writers
OLAP tools
End-user applications (e.g., CRM and SRM, ATM)
Modeling/ mining tools
Visualization tools
Business performance management tools

New business rules for operational decisions

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

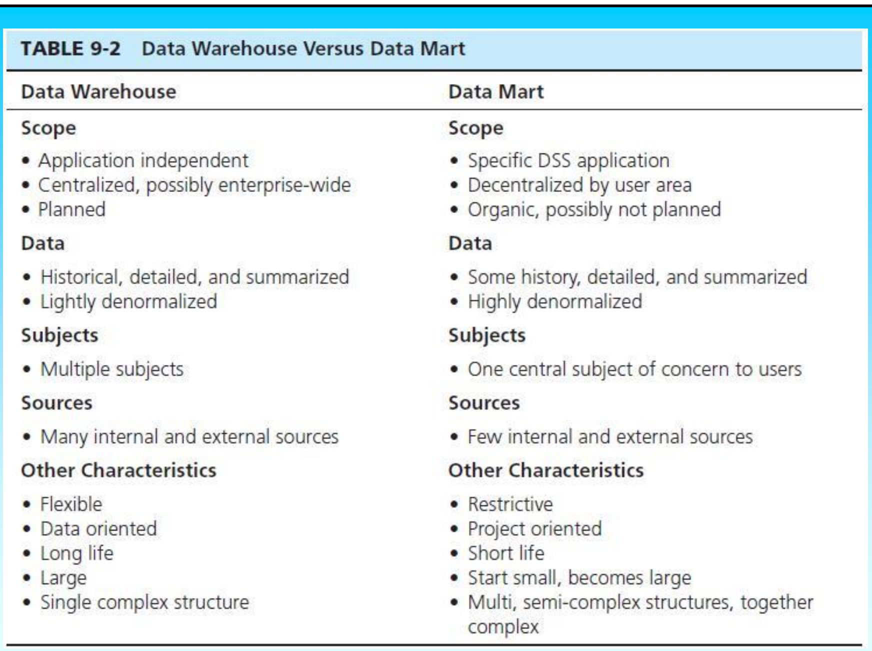
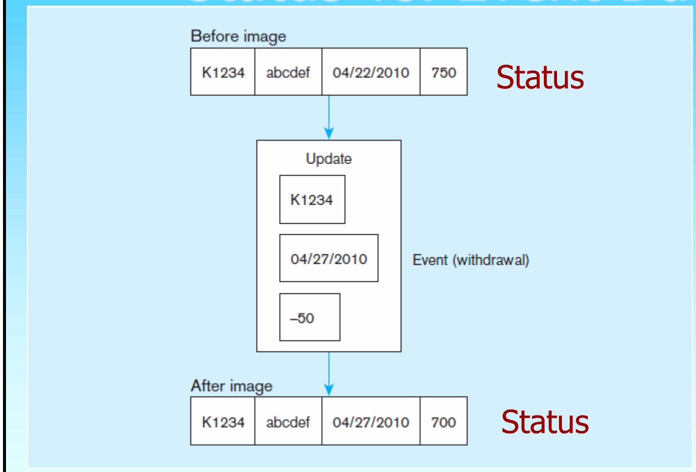


TABLE 9-2 Data Warehouse Versus Data Mart

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 • Start small, becomes large • Multi, semi-complex structures, together complex

Data Characteristics
Status vs. Event Data



Status

Status

Figure 9-6
Example of DBMS log entry

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

DATA CHARACTERISTICS STATUS VS. EVENT DATA

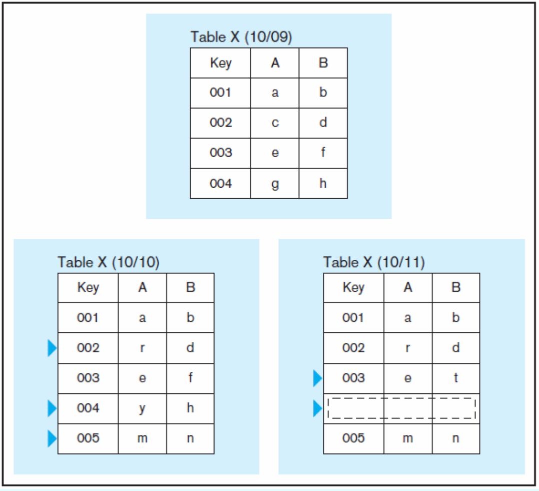


Table X (10/09)

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 STATUS VS. EVENT DATA

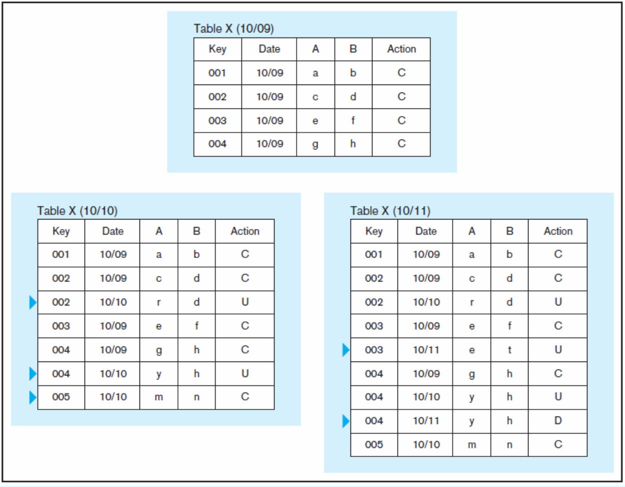


Table X (10/09)

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)

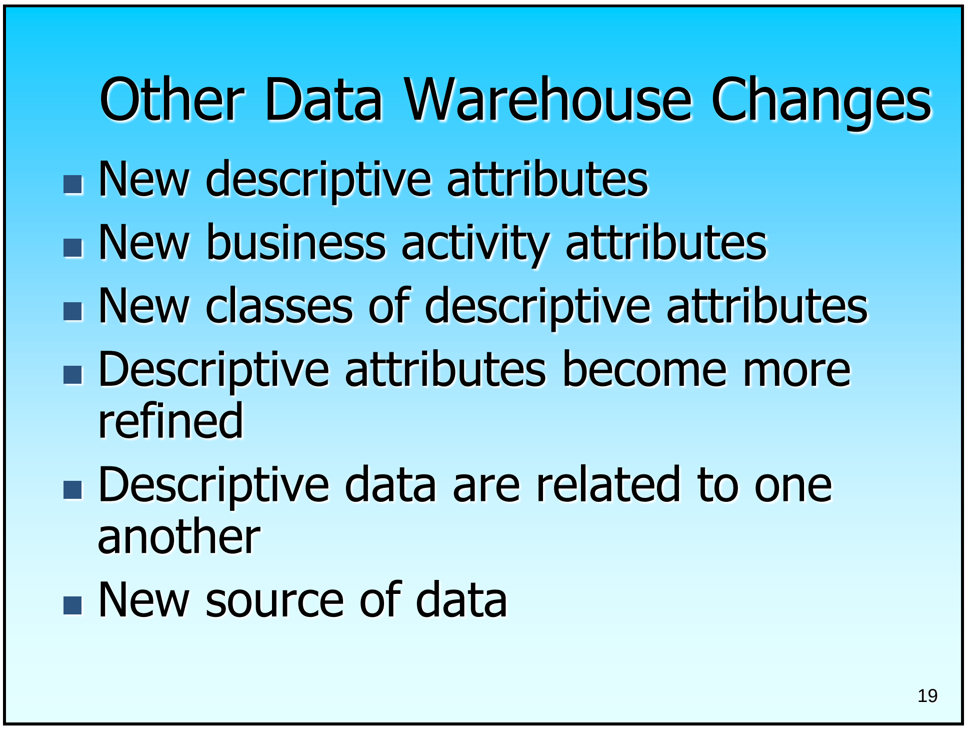
Key	Date	A	B	Action
001	10/09	a	b	C
002	10/09	c	d	C
▶ 002	10/10	r	d	U
003	10/09	e	f	C
004	10/09	g	h	C
▶ 004	10/10	y	h	U
▶ 005	10/10	m	n	C

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

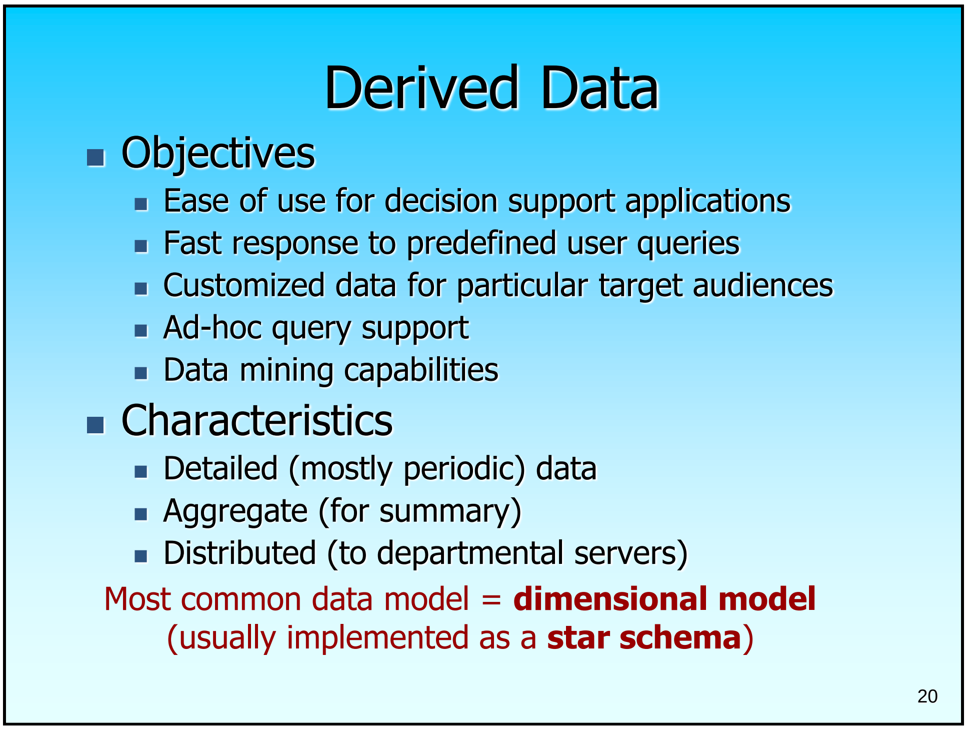
Figure 9-8 Periodic warehouse data

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



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

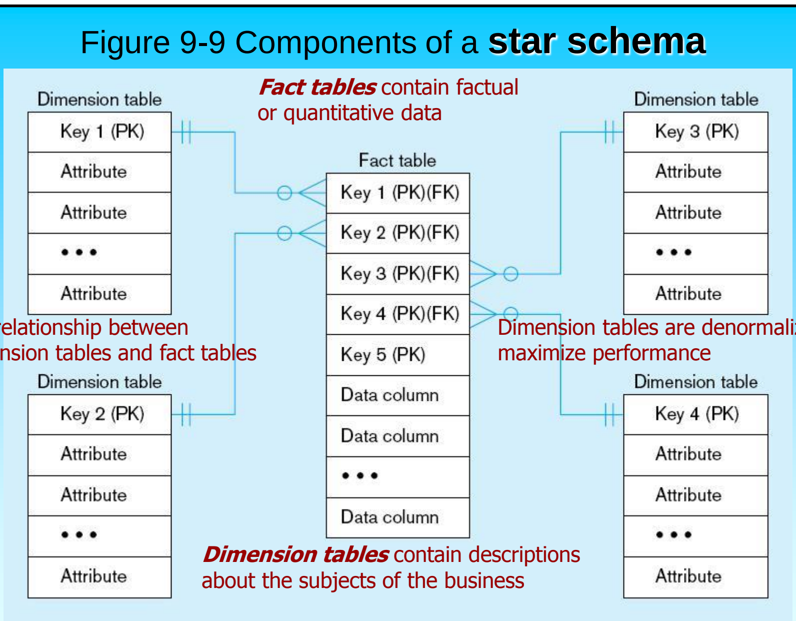


Derived Data

- 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**)

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



Fact tables contain factual or quantitative data

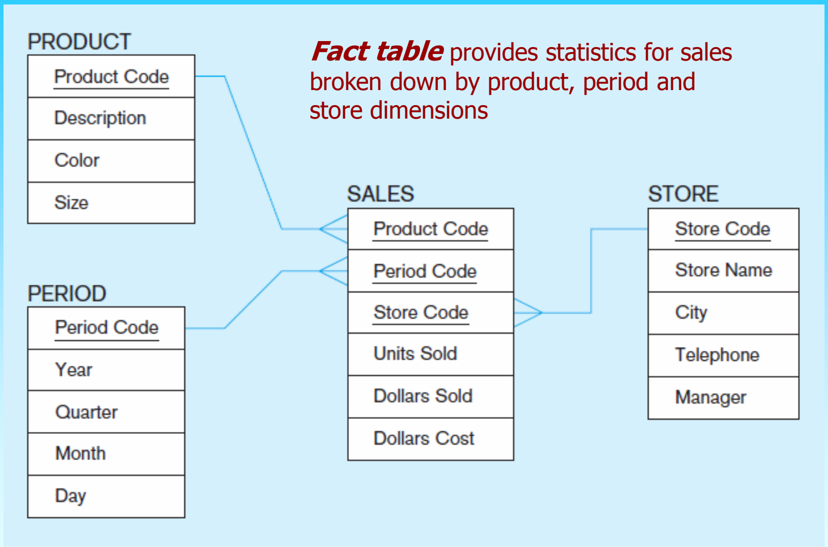
1:N relationship between dimension tables and fact tables

Dimension tables are denormalized to maximize performance

Dimension tables contain descriptions about the subjects of the business

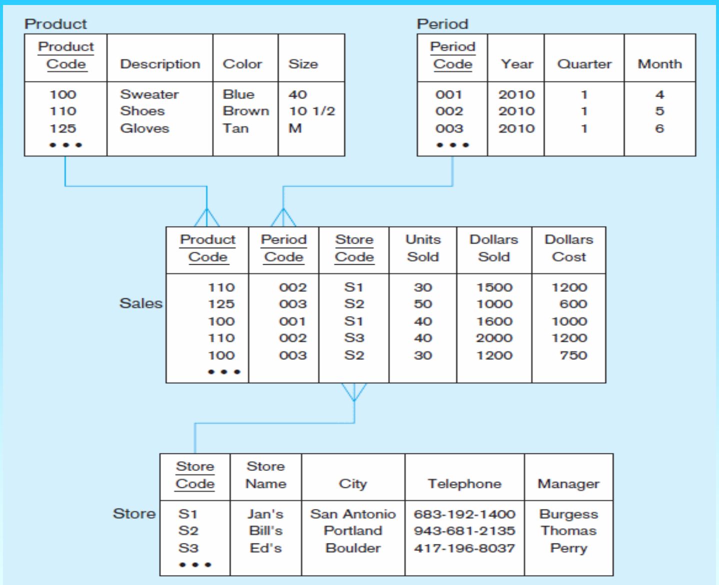
Excellent for ad-hoc queries, but bad for online transaction processing

Figure 9-10 Star schema example



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

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

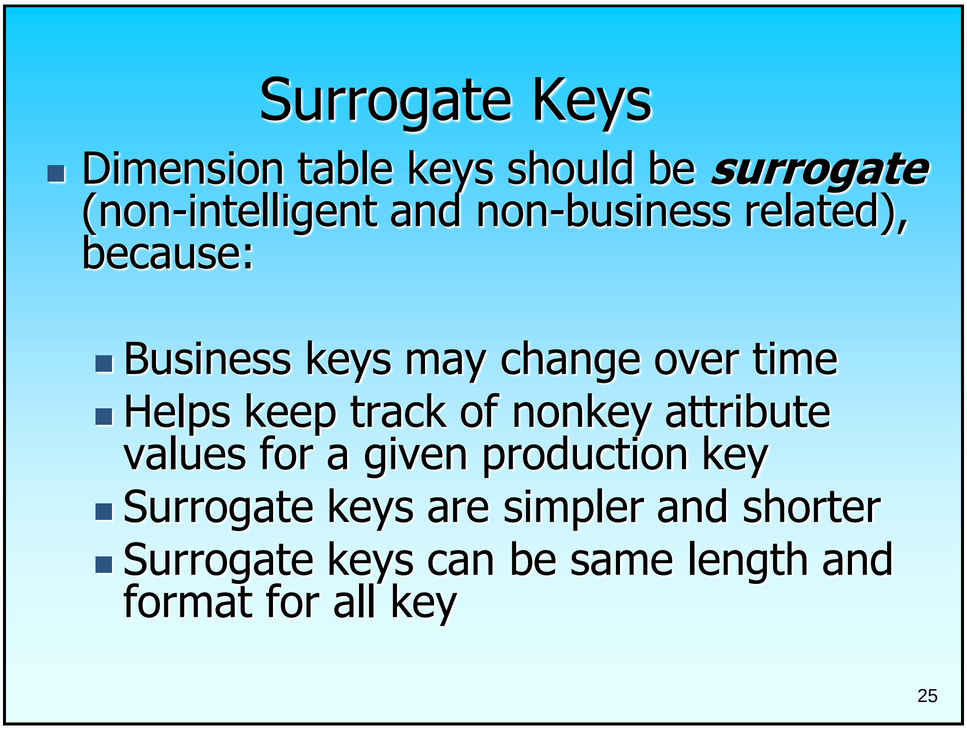
Product Code	Period Code	Store Code	Units Sold	Dollars Sold	Dollars Cost
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750
• • •					

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
• • •				

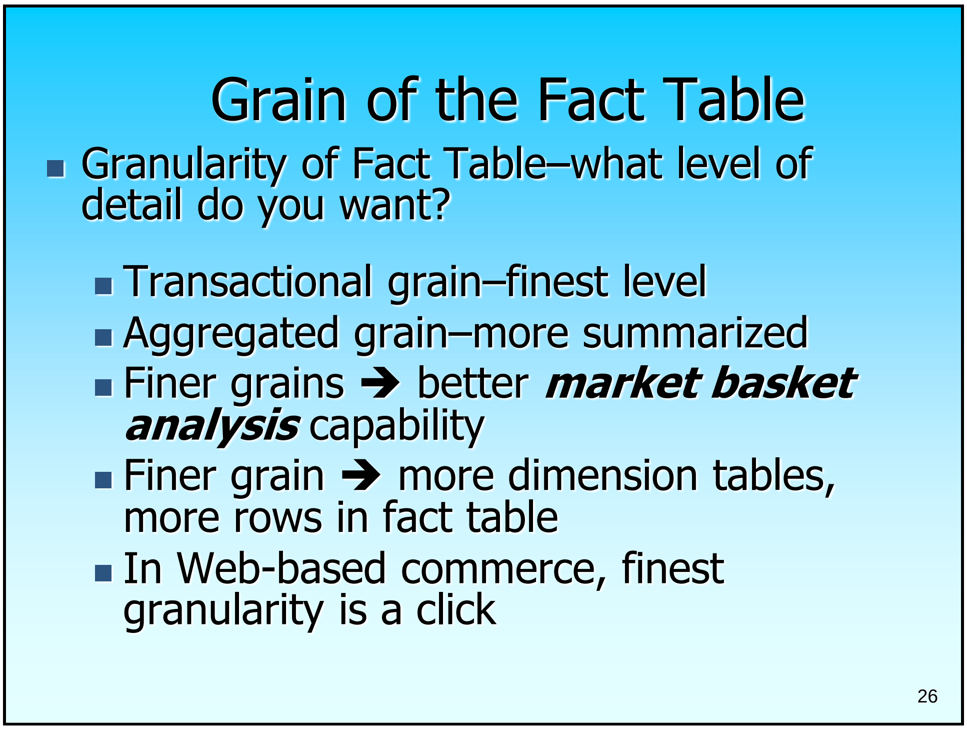
Surrogate Dimension Keys

- 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 keys



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

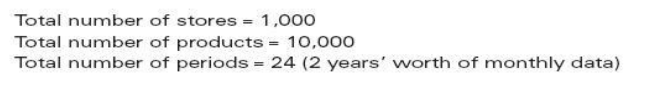
- Granularity of Fact Table–what level of detail do you want?
  - Transactional grain–finest level
  - Aggregated grain–more summarized
  - Finer grains → better ***market basket analysis*** capability
  - Finer grain → more dimension tables, more rows in fact table
  - In Web-based commerce, finest 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

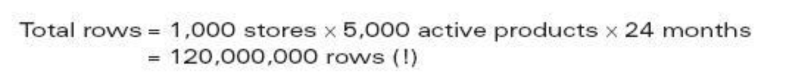
Size of Fact Table

- 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 for Figure 9-11:



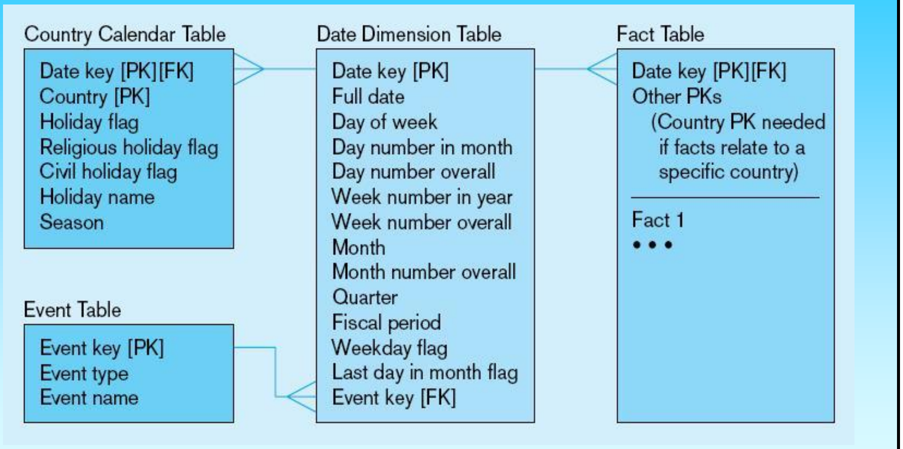
Total number of stores = 1,000
Total number of products = 10,000
Total number of periods = 24 (2 years' worth of monthly data)

- 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 (!)

Figure 9-12 Modeling dates



Country Calendar Table

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

Event Table

Event key [PK]
Event type
Event name

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
• • •

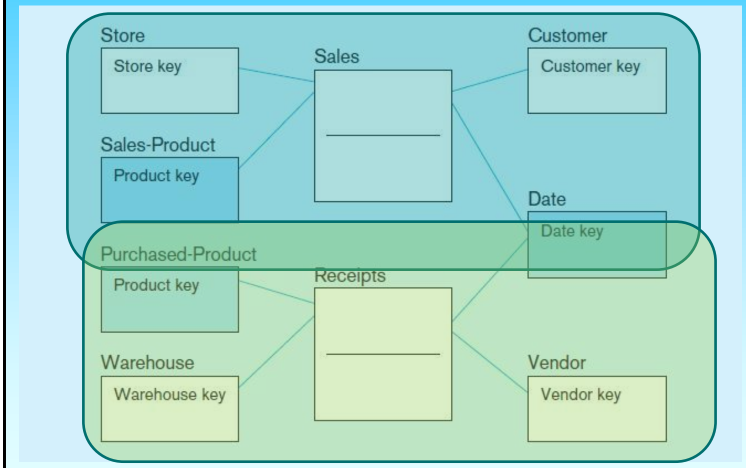
Fact tables contain time-period data
➔ Date dimensions are important

Variations of the Star Schema

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



Store
Store key

Sales

Customer
Customer key

Sales-Product
Product key

Date
Date key

Purchased-Product
Product key

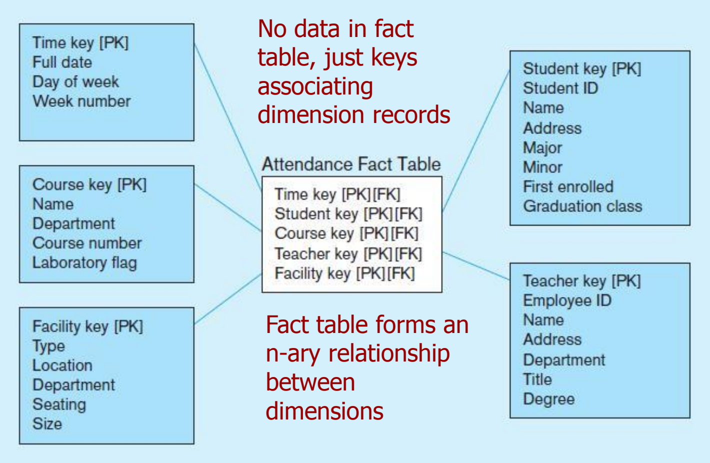
Receipts

Warehouse
Warehouse key

Vendor
Vendor key

Conformed dimension
Associated with multiple fact tables

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

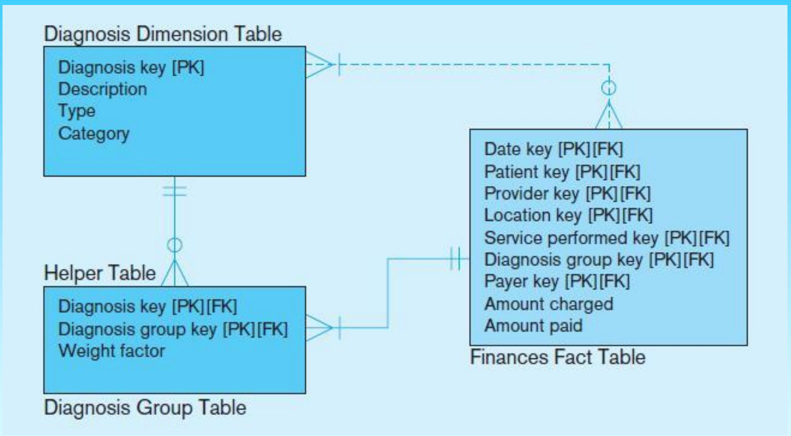
Student key [PK]
Student ID
Name
Address
Major
Minor
First enrolled
Graduation class

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

Figure 9-15 Multivalued dimension



Diagnosis Dimension Table

Diagnosis key [PK]
Description
Type
Category

Helper Table

Diagnosis key [PK][FK]
Diagnosis group key [PK][FK]
Weight factor

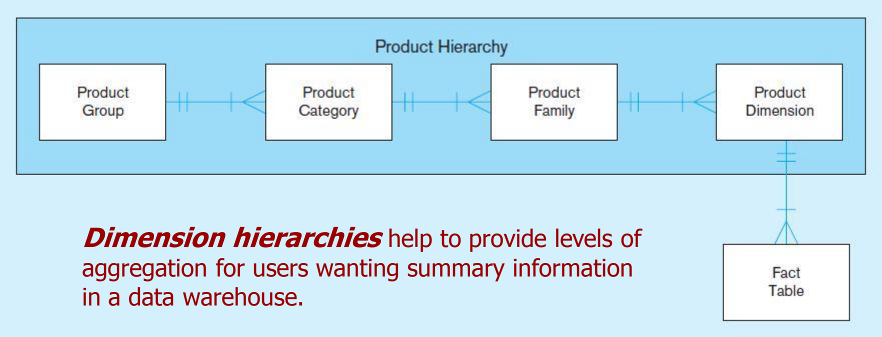
Diagnosis Group Table

Date key [PK][FK]
Patient key [PK][FK]
Provider key [PK][FK]
Location key [PK][FK]
Service performed key [PK][FK]
Diagnosis group key [PK][FK]
Payer key [PK][FK]
Amount charged
Amount paid

Finances Fact Table

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

Figure 9-16 Fixed product hierarchy

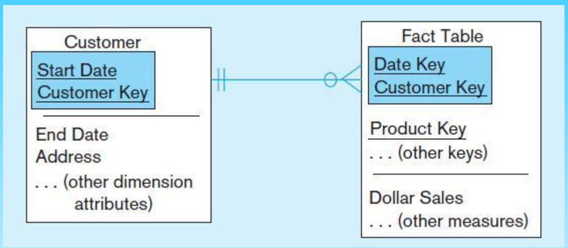


Dimension hierarchies help to provide levels of aggregation for users wanting summary information in a data warehouse.

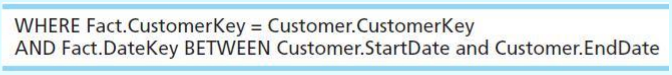
Slowly Changing Dimensions (SCD)

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

Figure 9-18 Example of Type 2 SCD Customer dimension table



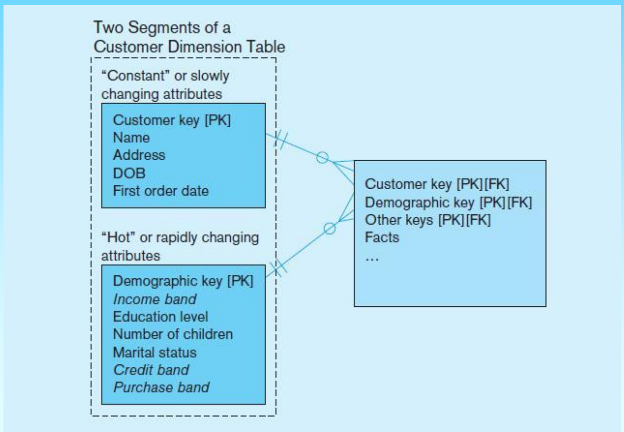
The dimension table contains several records for the same customer. The specific customer record to use depends on the key and the date of the fact, which should be between start and end dates of the SCD customer record.

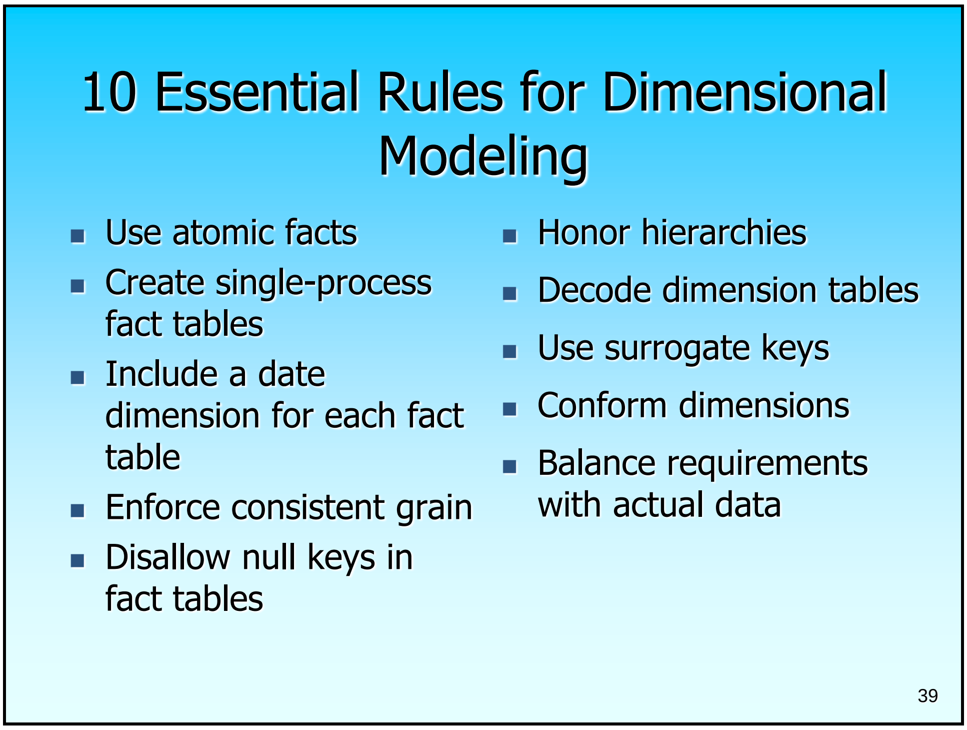


```
WHERE Fact.CustomerKey = Customer.CustomerKey
AND Fact.DateKey BETWEEN Customer.StartDate and Customer.EndDate
```

Figure 9-19 Dimension segmentation

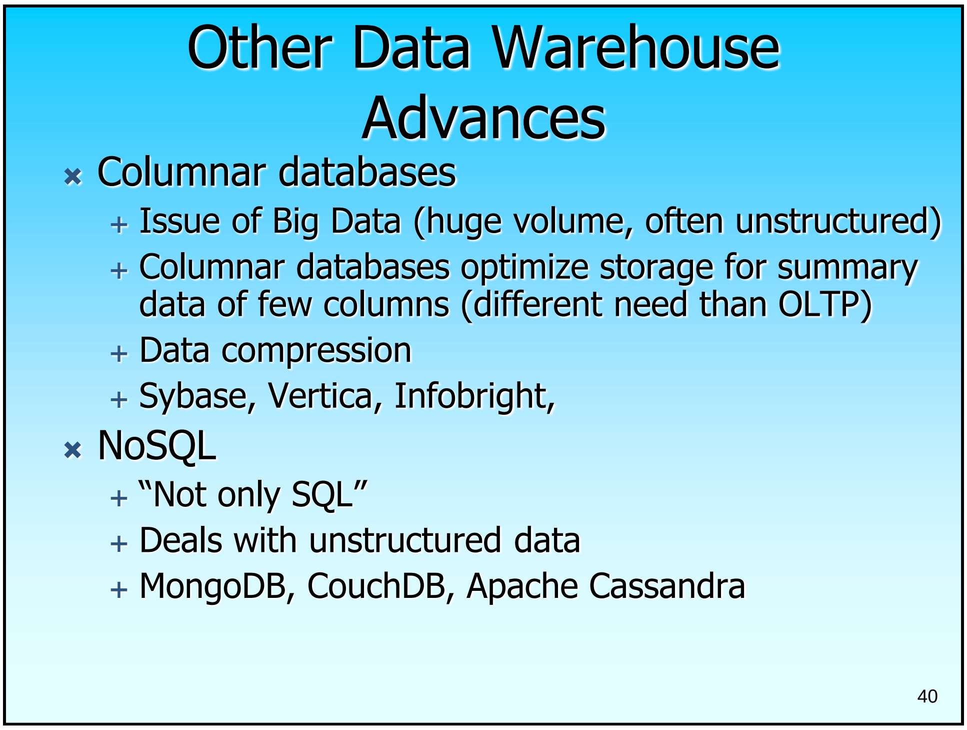
For rapidly changing attributes (hot attributes), Type 2 SCD approach creates too many rows and too much redundant data. Use segmentation instead.





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



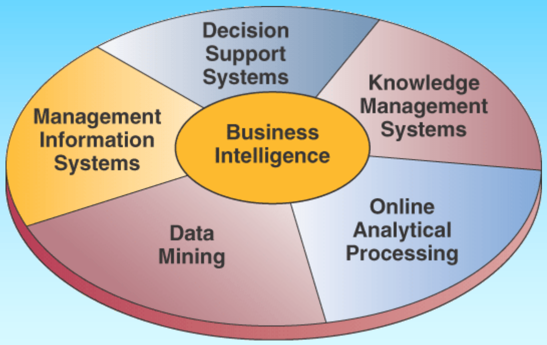
Other Data Warehouse Advances

- Columnar databases
  - Issue of Big Data (huge volume, often unstructured)
  - Columnar databases optimize storage for summary data of few columns (different need than OLTP)
  - Data compression
  - Sybase, Vertica, Infobright,
- NoSQL
  - “Not only SQL”
  - Deals with unstructured data
  - MongoDB, CouchDB, Apache Cassandra

The User Interface
Metadata (data catalog)

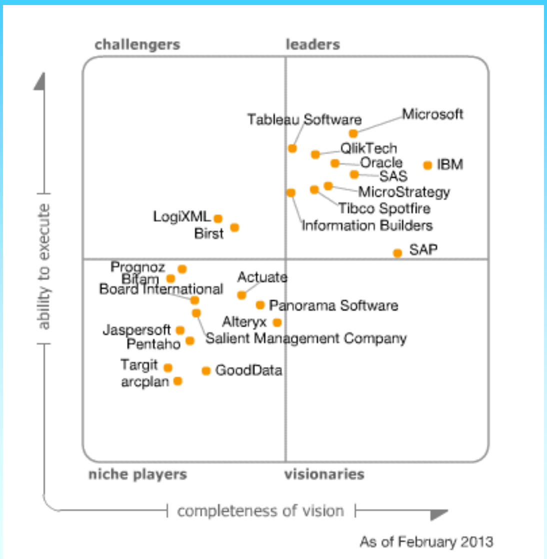
- Identify subjects of the data mart
- Identify dimensions and facts
- Indicate how data is derived from enterprise data warehouses, including derivation rules
- Indicate how data is derived from operational data store, including derivation rules
- Identify available reports and predefined queries
-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

Business Intelligence (BI)


Decision Support Systems
Knowledge Management Systems
Management Information Systems
Business Intelligence
Online Analytical Processing
Data Mining

O´Brien, J. A., Marakas, G. M. (2007). *Introduction to Information Systems.* McGraw-Hill.

Magic Quadrant for Business Intelligence and Analytics Platforms **Gartner 13 February 2013**

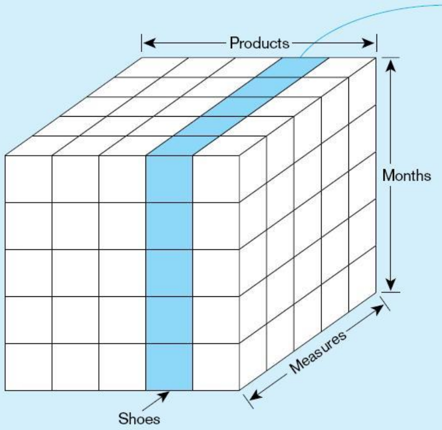


http://www.gartner.com/technology/reprints.do?id=1-1DZLPF2&ct=130207&st=sb

Online Analytical Processing (OLAP) Tools

- 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 9-21 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 9-22 Example of drill-down

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

Summary report

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

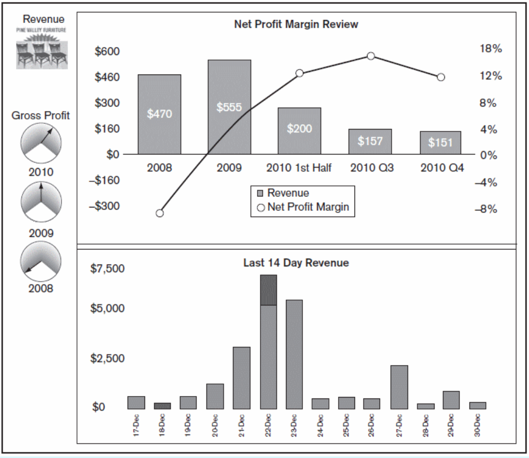
Drill-down with color added

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

Business Performance Mgmt (BPM)

Figure 9-25
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

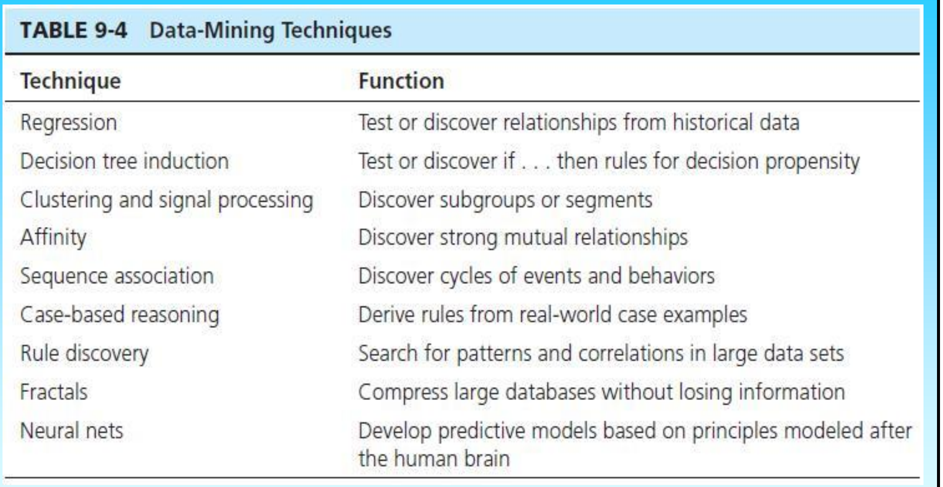


TABLE 9-4 Data-Mining Techniques

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

Teradata University

http://www.teradatauniversitynetwork.com/tun/

Pass: UnifiedDataArchitecture