Data Warehouse & Data Mining

Lecture 12

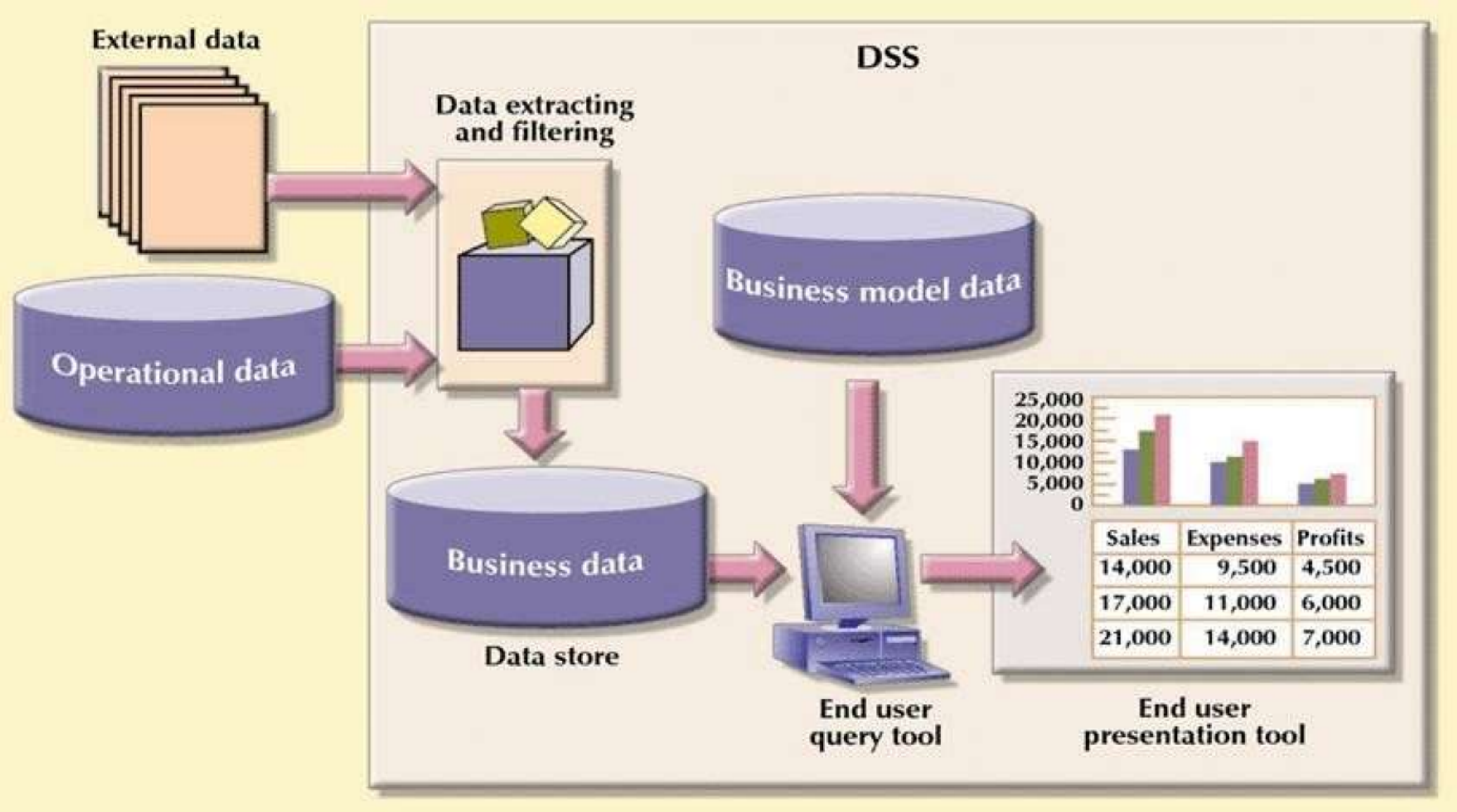
Learning Outcomes

- The need for data analysis
- What a data warehouse is, how to prepare data for one, and how to implement one
- About data analytics, data mining, and predictive analytics

The Need for Data Analysis

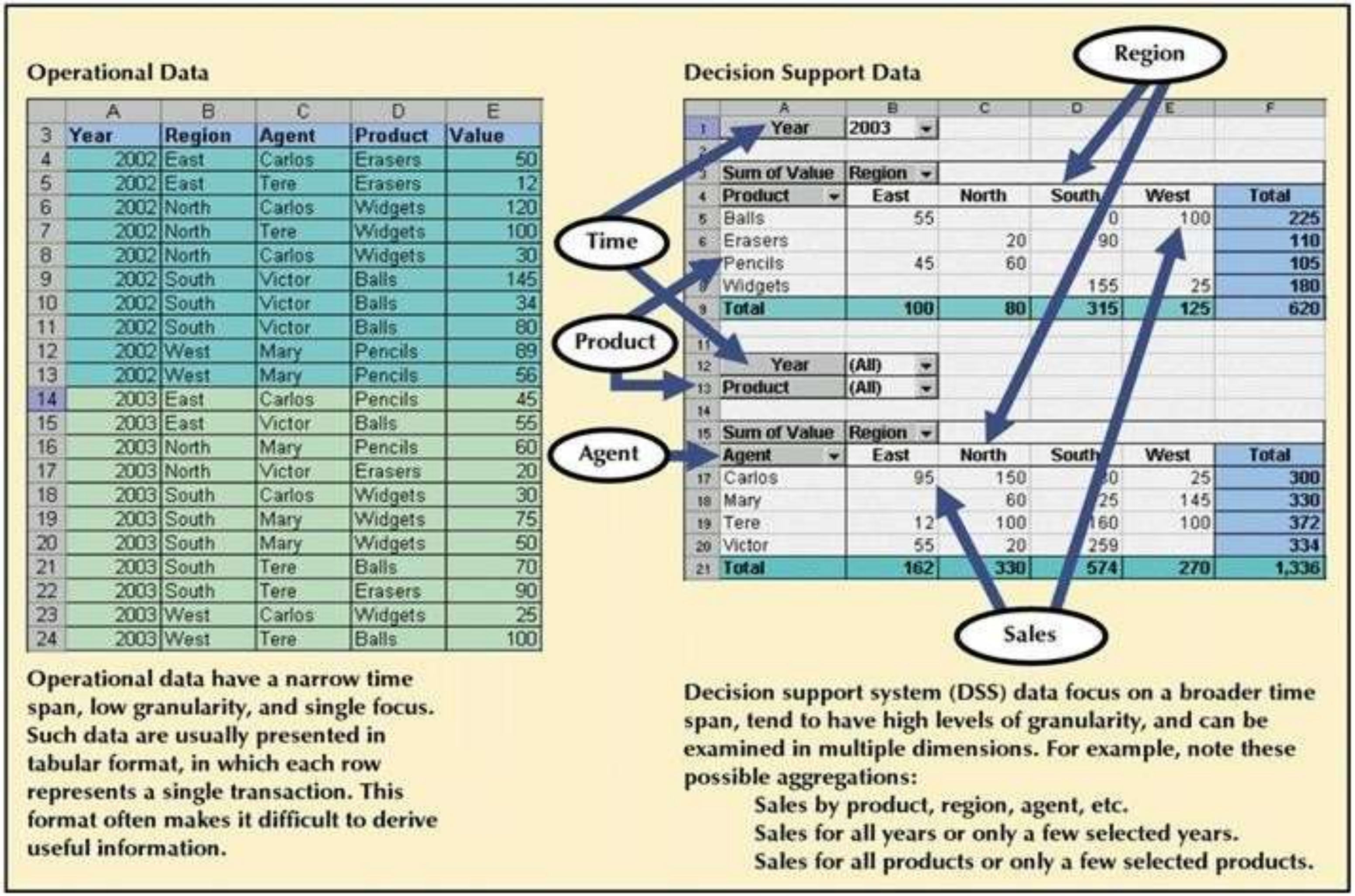
- Organizations are growing rapidly
  - Search for competitive advantage
- Managers needs to track daily transactions to evaluate how the business is performing
- Decision support system
  - Computerized tools used to extract information from data to assist managerial business decision-making

Decision Support System



Sales	Expenses	Profits
14,000	9,500	4,500
17,000	11,000	6,000
21,000	14,000	7,000

Transforming Operational Data Into Decision Support Data



Operational Data

	A	B	C	D	E
3	Year	Region	Agent	Product	Value
4	2002	East	Carlos	Erasers	50
5	2002	East	Tere	Erasers	12
6	2002	North	Carlos	Widgets	120
7	2002	North	Tere	Widgets	100
8	2002	North	Carlos	Widgets	30
9	2002	South	Victor	Balls	145
10	2002	South	Victor	Balls	34
11	2002	South	Victor	Balls	80
12	2002	West	Mary	Pencils	89
13	2002	West	Mary	Pencils	56
14	2003	East	Carlos	Pencils	45
15	2003	East	Victor	Balls	55
16	2003	North	Mary	Pencils	60
17	2003	North	Victor	Erasers	20
18	2003	South	Carlos	Widgets	30
19	2003	South	Mary	Widgets	75
20	2003	South	Mary	Widgets	50
21	2003	South	Tere	Balls	70
22	2003	South	Tere	Erasers	90
23	2003	West	Carlos	Widgets	25
24	2003	West	Tere	Balls	100

Operational data have a narrow time span, low granularity, and single focus. Such data are usually presented in tabular format, in which each row represents a single transaction. This format often makes it difficult to derive useful information.

Decision Support Data

	A	B	C	D	E	F
1	Year	2003				
2						
3	Sum of Value	Region				
4	Product	East	North	South	West	Total
5	Balls	55		0	100	225
6	Erasers		20	90		110
7	Pencils	45	60			105
8	Widgets			155	25	180
9	Total	100	80	315	125	620
10						
11						
12	Year	(All)				
13	Product	(All)				
14						
15	Sum of Value	Region				
16	Agent	East	North	South	West	Total
17	Carlos	95	150	0	25	300
18	Mary		60	25	145	330
19	Tere	12	100	160	100	372
20	Victor	55	20	259		334
21	Total	162	330	574	270	1,336

Decision support system (DSS) data focus on a broader time span, tend to have high levels of granularity, and can be examined in multiple dimensions. For example, note these possible aggregations:

Sales by product, region, agent, etc.
Sales for all years or only a few selected years.
Sales for all products or only a few selected products.

Data Warehouse

- A data warehouse is an *integrated*, *subject-oriented*, *time-variant*, *non-volatile* database that provides support for decision-making
  - Integrated
    - The Data Warehouse is a centralized, consolidated database that integrates data retrieved from the entire organization.
  - Subject-Oriented
    - The Data Warehouse data is arranged and optimized to provide answers to questions coming from diverse functional areas within a company.

The Data Warehouse

- Time Variant
  - The Warehouse data represent the flow of data through time. It can even contain projected data.
- Non-Volatile
  - Once data enter the Data Warehouse, they are never removed.
  - The Data Warehouse is always growing.

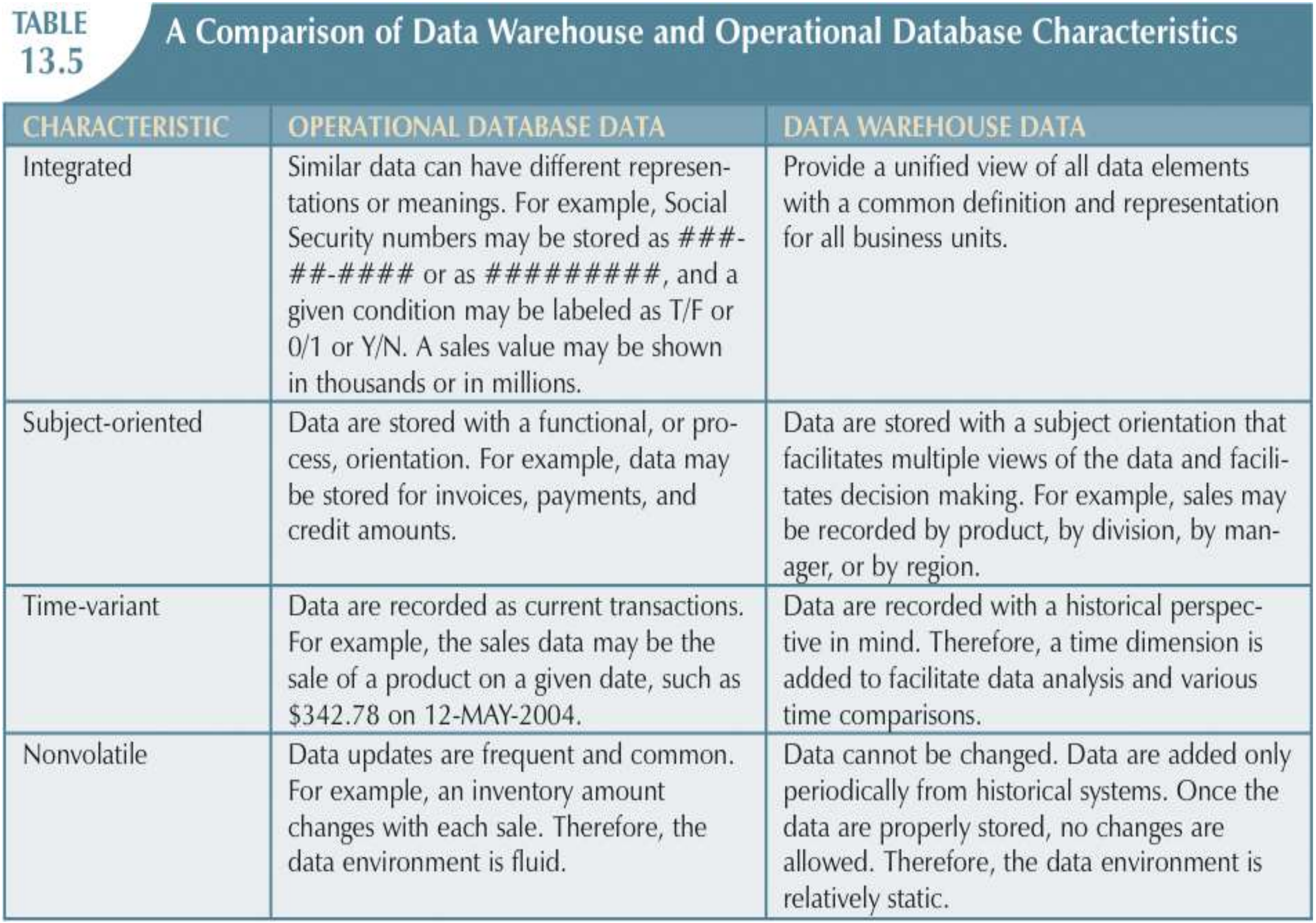
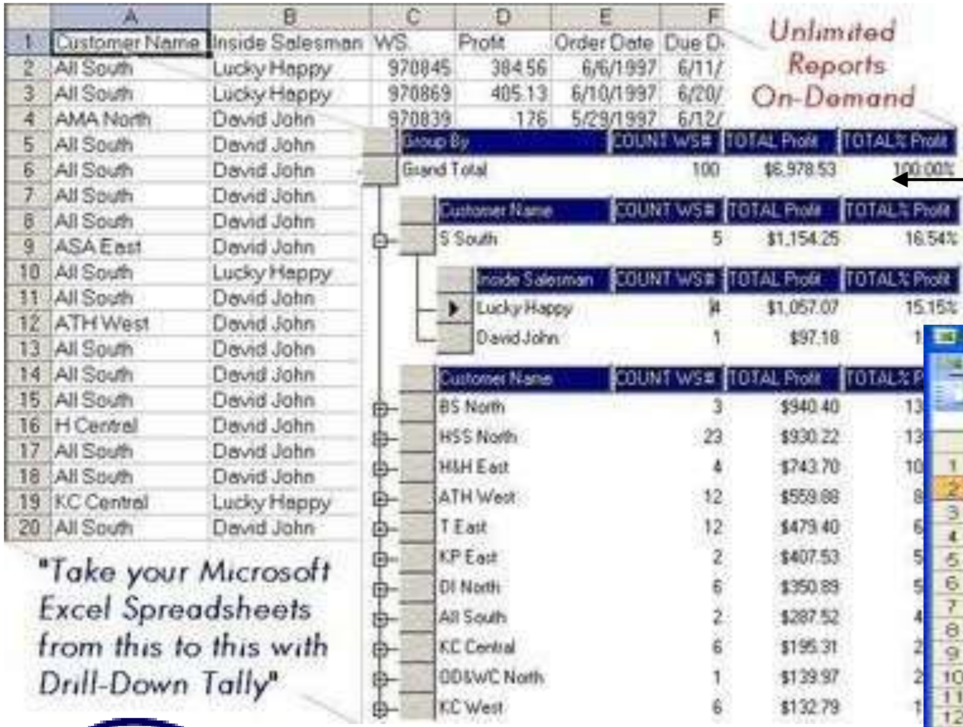


TABLE 13.5 A Comparison of Data Warehouse and Operational Database Characteristics

CHARACTERISTIC	OPERATIONAL DATABASE DATA	DATA WAREHOUSE DATA
Integrated	Similar data can have different representations or meanings. For example, Social Security numbers may be stored as ###-##-#### or as #########, and a given condition may be labeled as T/F or 0/1 or Y/N. A sales value may be shown in thousands or in millions.	Provide a unified view of all data elements with a common definition and representation for all business units.
Subject-oriented	Data are stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, and credit amounts.	Data are stored with a subject orientation that facilitates multiple views of the data and facilitates decision making. For example, sales may be recorded by product, by division, by manager, or by region.
Time-variant	Data are recorded as current transactions. For example, the sales data may be the sale of a product on a given date, such as $342.78 on 12-MAY-2004.	Data are recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.
Nonvolatile	Data updates are frequent and common. For example, an inventory amount changes with each sale. Therefore, the data environment is fluid.	Data cannot be changed. Data are added only periodically from historical systems. Once the data are properly stored, no changes are allowed. Therefore, the data environment is relatively static.

Role of Data Warehouse

- Focal point for decision support systems
- Support decision making by allowing users to :
  - *drill-down for a more detailed information,*
  - *roll-up to view summarized information,*
  - *slice and dice a dimension for selection of a specific item of interest and*
  - *pivot to re-orientate the view of multidimensional data.*



	A	B	C	D	E	F
1	Customer Name	Inside Salesman	WS	Profit	Order Date	Due D
2	All South	Lucky Happy	970845	384.56	6/6/1997	6/11/
3	All South	Lucky Happy	970869	405.13	6/10/1997	6/20/
4	AMA North	David John	970839	176	5/29/1997	6/12/
5	All South	David John				
6	All South	David John				
7	All South	David John				
8	All South	David John				
9	ASA East	David John				
10	All South	Lucky Happy				
11	All South	David John				
12	ATH West	David John				
13	All South	David John				
14	All South	David John				
15	All South	David John				
16	H Central	David John				
17	All South	David John				
18	All South	David John				
19	KC Central	Lucky Happy				
20	All South	David John				

Unlimited Reports On-Demand

Group By	COUNT WS#	TOTAL Profit	TOTAL% Profit
Grand Total	100	$6,978.53	100.00%

Customer Name	COUNT WS#	TOTAL Profit	TOTAL% Profit
S South	5	$1,154.25	16.54%

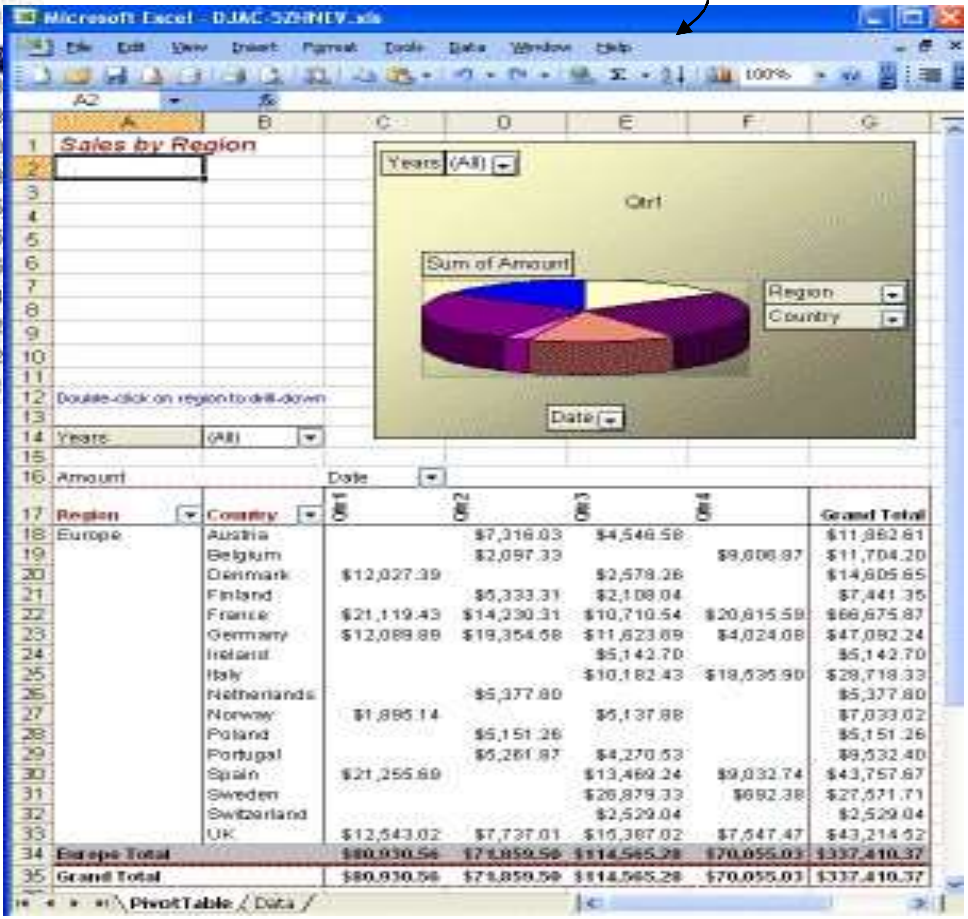
Inside Salesman	COUNT WS#	TOTAL Profit	TOTAL% Profit
Lucky Happy	4	$1,057.07	15.15%
David John	1	$97.18	1

Customer Name	COUNT WS#	TOTAL Profit	TOTAL% P
BS North	3	$940.40	13
HSS North	23	$930.22	13
H&H East	4	$743.70	10
ATH West	12	$559.88	8
T East	12	$479.40	6
KP East	2	$407.53	5
DI North	6	$350.89	5
All South	2	$287.52	4
KC Central	6	$195.31	2
DD&WC North	1	$139.97	2
KC West	6	$132.79	1

"Take your Microsoft Excel Spreadsheets from this to this with Drill-Down Tally"

Drill-down and Roll-up

Pivot table



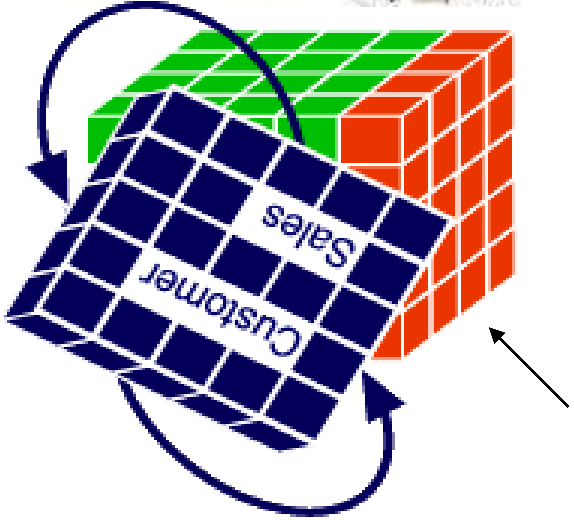
Sales by Region

Double-click on region to drill-down

Years (All)

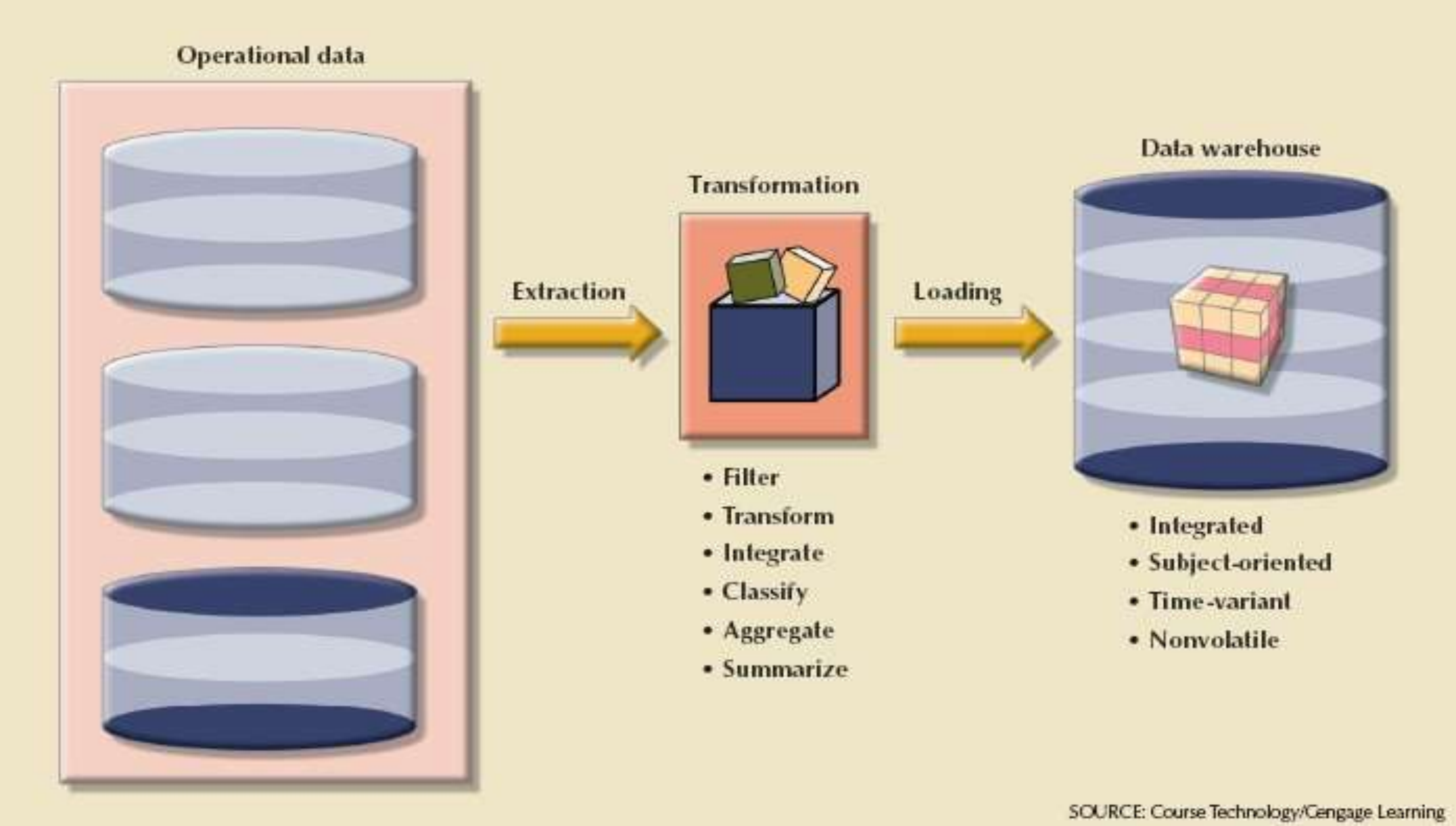
Amount Date

Region	Country	Qtr1	Qtr2	Qtr3	Qtr4	Grand Total
Europe	Austria		$7,316.03	$4,546.58		$11,862.61
	Belgium		$2,097.33		$9,606.87	$11,704.20
	Denmark	$12,027.39		$2,578.26		$14,605.65
	Finland		$5,333.31	$2,108.04		$7,441.35
	France	$21,119.43	$14,230.31	$10,710.54	$20,615.59	$66,675.87
	Germany	$12,089.89	$19,354.58	$11,623.69	$4,024.08	$47,092.24
	Ireland			$5,142.70		$5,142.70
	Italy			$10,182.43	$18,535.90	$28,718.33
	Netherlands		$5,377.80			$5,377.80
	Norway	$1,895.14		$5,137.88		$7,033.02
	Poland		$5,151.26			$5,151.26
	Portugal		$5,261.87	$4,270.53		$9,532.40
	Spain	$21,255.89		$13,469.24	$9,032.74	$43,757.87
	Sweden			$26,879.33	$692.38	$27,571.71
	Switzerland			$2,529.04		$2,529.04
	UK	$12,543.02	$7,737.01	$15,387.02	$7,547.47	$43,214.52
Europe Total		$80,930.56	$71,859.50	$114,565.28	$70,055.03	$337,410.37
Grand Total		$80,930.56	$71,859.50	$114,565.28	$70,055.03	$337,410.37



Slice & dice

The ETL Process



SOURCE: Course Technology/Cengage Learning

Twelve Rules That Define a Data Warehouse

- Data warehouse and operational environments are **separated**
- Data warehouse data are **integrated**
- Data warehouse contains **historical data** over long time
- Data warehouse data are snapshot **data captured at given point in time**
- Data warehouse data are **subject-oriented**

Twelve Rules That Define a Data Warehouse (cont'd.)

- Data warehouse data are mainly **read-only**
  - Periodic batch updates from operational data
  - No online updates allowed
- Data warehouse development life cycle **differs** from classical systems development
- Data warehouse contains data with **several levels of detail**:
  - Current detail data, old detail data, lightly summarized data, and highly summarized data

Twelve Rules That Define a Data Warehouse (cont'd.)

- Read-only transactions to **very large data sets**
- Data warehouse environment **traces data sources, transformations, and storage**
- Data warehouse's **metadata are critical** component of this environment
- Data warehouse contains **chargeback mechanism** for resource usage
  - Enforces optimal use of data by end users – users are charged when involving data warehouse processing.

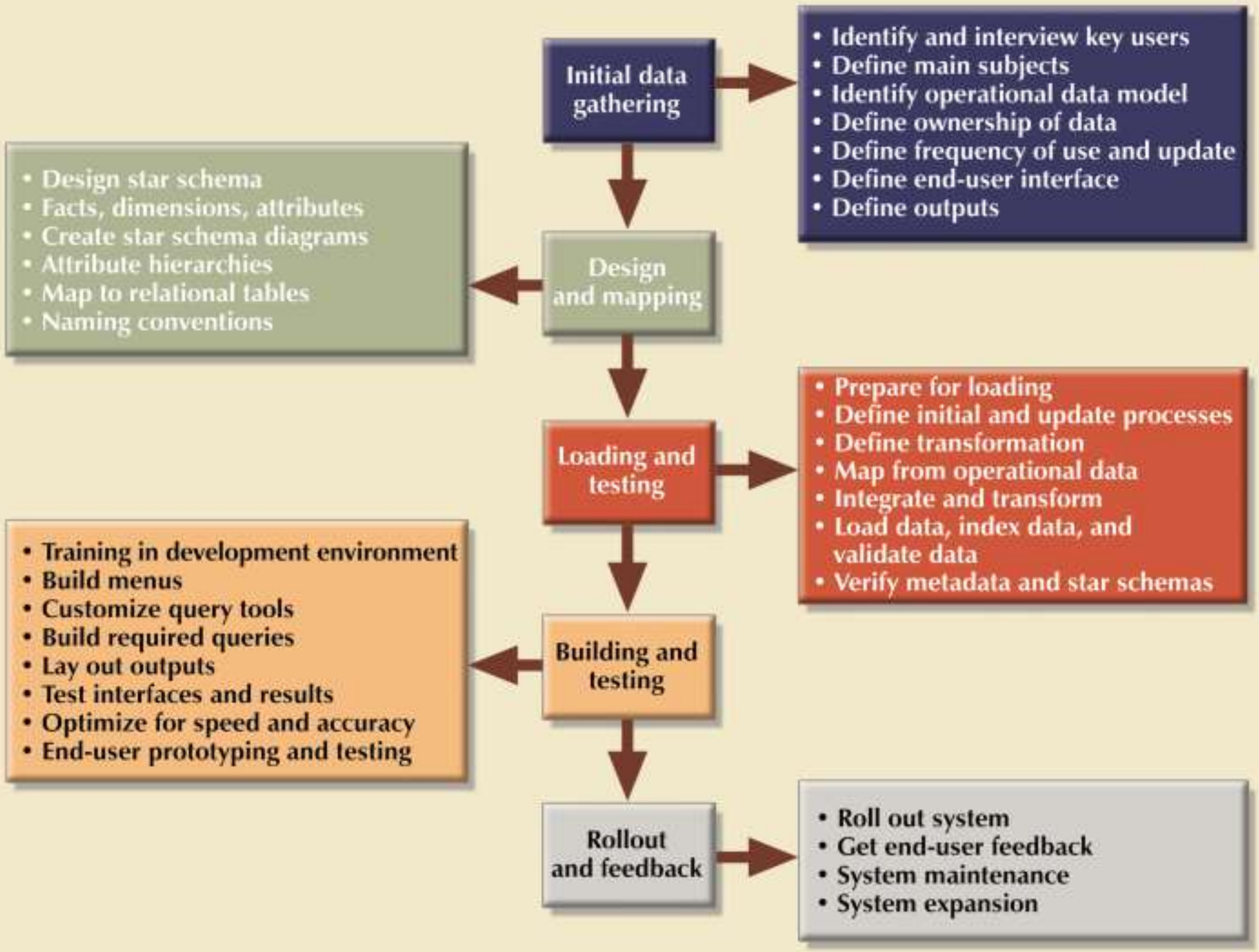
Implementing a Data Warehouse

- Numerous constraints, including:
  - Available funding
  - Management's view of role played by an IS department
    - Extent and depth of information requirements
  - Corporate culture
- No single formula can describe perfect data warehouse development

The Data Warehouse as an Active Decision Support Framework

- Data warehouse:
  - *Is not a static database*
  - *Is a dynamic framework for decision support that is always a work in progress*

FIGURE 13.21 Data warehouse design and implementation road map



Initial data gathering

- Identify and interview key users
- Define main subjects
- Identify operational data model
- Define ownership of data
- Define frequency of use and update
- Define end-user interface
- Define outputs

Design and mapping

- Design star schema
- Facts, dimensions, attributes
- Create star schema diagrams
- Attribute hierarchies
- Map to relational tables
- Naming conventions

Loading and testing

- Prepare for loading
- Define initial and update processes
- Define transformation
- Map from operational data
- Integrate and transform
- Load data, index data, and validate data
- Verify metadata and star schemas

Building and testing

- Training in development environment
- Build menus
- Customize query tools
- Build required queries
- Lay out outputs
- Test interfaces and results
- Optimize for speed and accuracy
- End-user prototyping and testing

Rollout and feedback

- Roll out system
- Get end-user feedback
- System maintenance
- System expansion

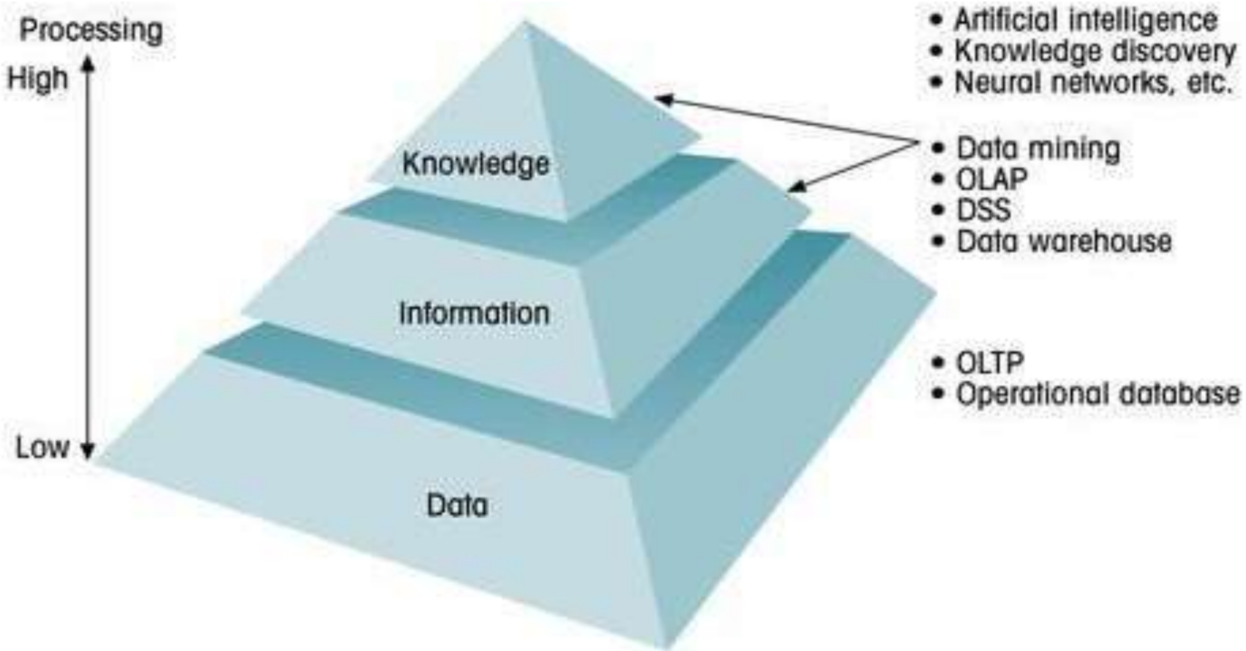
Data Mart

- A data mart is a **small**, **single-subject data warehouse** subset that provides decision support to a small group of people.
- Data Marts can serve as a test vehicle for companies exploring the potential benefits of Data Warehouses.
- Data Marts - address local or departmental problems
- Data Warehouse - involves a company-wide effort to support decision-making at all levels in the organization.

Data Mining

- Data mining tools **automatically** search the data for anomalies and identify possible relationships, thereby identifying problems that have not yet been identified by the end user.
- Requires minimal end-user intervention

Data Mining (cont')



Data-mining tools use advanced techniques from knowledge discovery, artificial intelligence, and other fields to obtain "knowledge" and apply it to business needs. Knowledge is then used to make predictions of events or forecasts of values such as sales returns, etc. Several OLAP tools have integrated at least some of these data-mining features in their products.

FIGURE 13.22 EXTRACTION OF KNOWLEDGE FROM DATA

Phases in Data Mining

1. ***Data Preparation***
   - Identify, collect and consolidating data for analysis
2. ***Data Analysis***
   - Identify and explore data with the goal of discovering useful information
3. ***Knowledge Acquisition***
   - Select the appropriate modeling or knowledge acquisition algorithms.
   - Examples: neural networks, decision trees, etc.
4. ***Prognosis***
   - Predict future behavior and forecast business outcomes using the data mining findings.