# IT5507 Fundamentals of Data Science

# Chapter 13
# Business Intelligence and Data Warehouses

# Learning Objectives

- After completing this chapter, you will be able to:
  - Describe the role of business intelligence in providing comprehensive business decision support
  - Describe the architecture, reporting styles, evolution, and benefits of business intelligence
  - Differentiate between operational data and decision support data
  - Identify the purpose, characteristics, and components of a data warehouse
  - Develop star and snowflake schemas for decision-making purposes
  - Describe the characteristics and capabilities of online analytical processing (OLAP)
  - Describe the role and functions of data analytics and data mining
  - Explain how SQL analytic functions are used to support data analytics
  - Define data visualization and explain how it supports business intelligence

# The Need for Data Analysis

- Organizations tend to grow and prosper as they gain a better understanding of their environment
  - Evaluate through tracking daily transactions and analyzing company data
- Organizations are always looking for a competitive advantage
  - Product development, market positioning, sales promotions, and customer service
- Companies and software vendors addressed these multilevel decision support needs by creating autonomous applications for particular groups of users
  - This more comprehensive and integrated decision support framework within organizations became known as business intelligence
  - Therefore, the term BI or Business intelligence coined as:
    - the use of data analysis tools and techniques to extract actionable insights from raw data, aiding in informed decision-making within organizations.

CENGAGE

# Business Intelligence (1 of 2)

- Comprehensive, cohesive, integrated set of tools and processes
  - Captures, collects, integrates, stores, and analyzes data
  - Generates and presents information to support business decision making
- Allows transformation
  - Data into information
  - Information into knowledge
  - Knowledge into wisdom

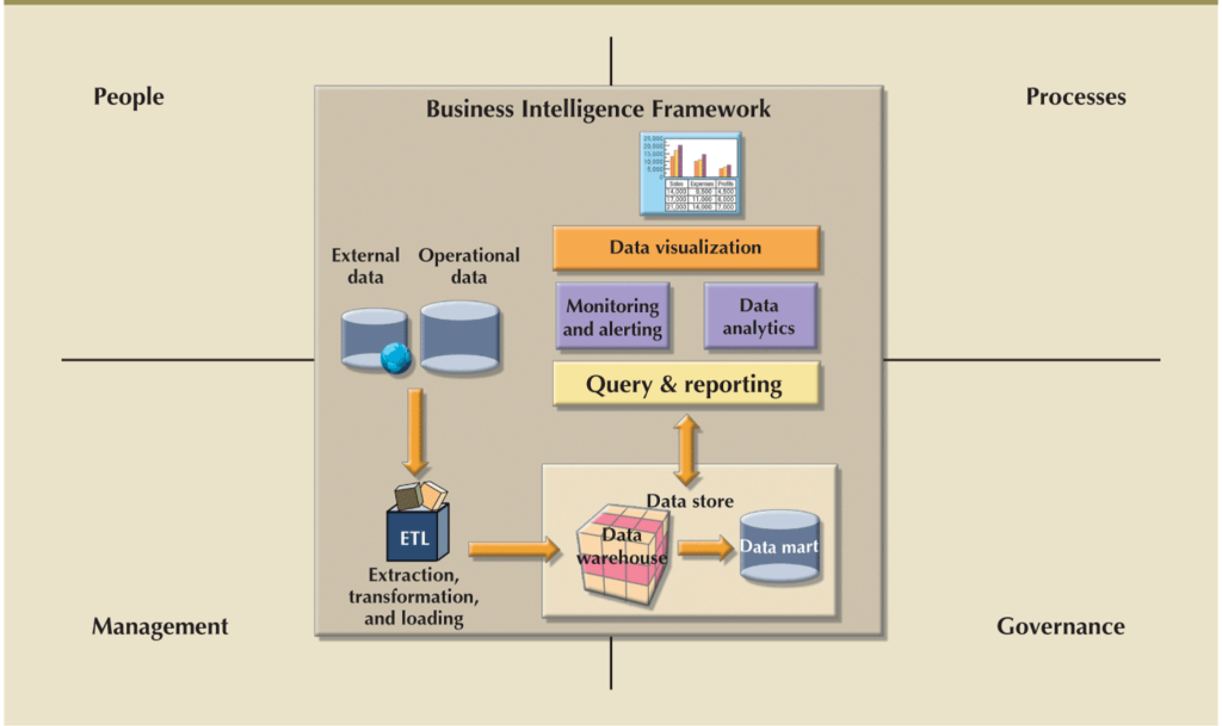
# Business Intelligence (2 of 2)

- Concepts, practices, tools and techniques to help business
  - Understand core capabilities
  - Provide snapshots of the company situation
  - Identify key opportunities to create a competitive advantage
- Provides a framework
  - Collecting and storing operational data and aggregating it into decision support data
  - Analyzing decision support data and presenting generated information to end users to support business decisions
  - Making business decisions which generate more data
  - Monitoring results to evaluate outcomes and predicting future outcomes with a high degree of accuracy

# Business Intelligence Architecture (1 of 3)

BI architecture refers to the framework and components that support the collection, analysis, and presentation of business intelligence data. It encompasses people, processes, management, and governance to ensure that data is effectively utilized to drive decision-making within an organization.

FIGURE 13.1 BUSINESS INTELLIGENCE FRAMEWORK




1. **People:** Refers to the individuals within an organization who are involved in the BI process, including analysts, managers, and decision-makers.
2. **Processes:** Encompasses the methods and workflows used to collect, analyze, and interpret data to derive actionable insights for decision-making.
3. **Management:** Involves the oversight and coordination of BI activities, including resource allocation, project prioritization, and performance monitoring.
4. **Governance:** Establishes policies, procedures, and controls to ensure the quality, integrity, and security of BI data and processes, as well as compliance with regulatory requirements.

# Business Intelligence Architecture (2 of 3)

| Table 13.2 Basic BI Architectural Components | |
|---|---|
| **Component** | **Description** |
| ETL tools | Data extraction, transformation, and loading (ETL ) tools collect, filter, integrate, and aggregate internal and external data to be saved into a data store optimized for decision support. |
| Data store | The data store is optimized for decision support and is generally represented by a data warehouse or a data mart. The data is stored in structures that are optimized for data analysis and query speed. |
| Query and reporting | This component performs data selection and retrieval, and it is used by the data analyst to create queries that access the database and create the required reports. |
| Data visualization | This component presents data to the end user in a variety of meaningful and innovative ways. This tool helps the end user select the most appropriate presentation format, such as summary reports, maps, pie or bar graphs, mixed graphs, and static or interactive dashboards. |
| Data monitoring and alerting | This component allows real-time monitoring of business activities. The BI system will present concise information in a single integrated view. This integrated view could include specific metrics about the system performance or activities, such as number of orders placed in the last four hours, number of customer complaints by product by month, and total revenue by region. Alerts can be placed on a given metric; once the value of a metric goes below or above a certain baseline, the system will perform a given action, such as emailing shop floor managers, presenting visual alerts, or starting an application. |
| Data analytics | This component performs data analysis and data-mining tasks using the data in the data store. This tool advises the user as to which data analysis tool to select and how to build a reliable business data model. Business models are generated by special algorithms that identify and enhance the understanding of business situations and problems. |

CENGAGE

# Business Intelligence Architecture (3 of 3)

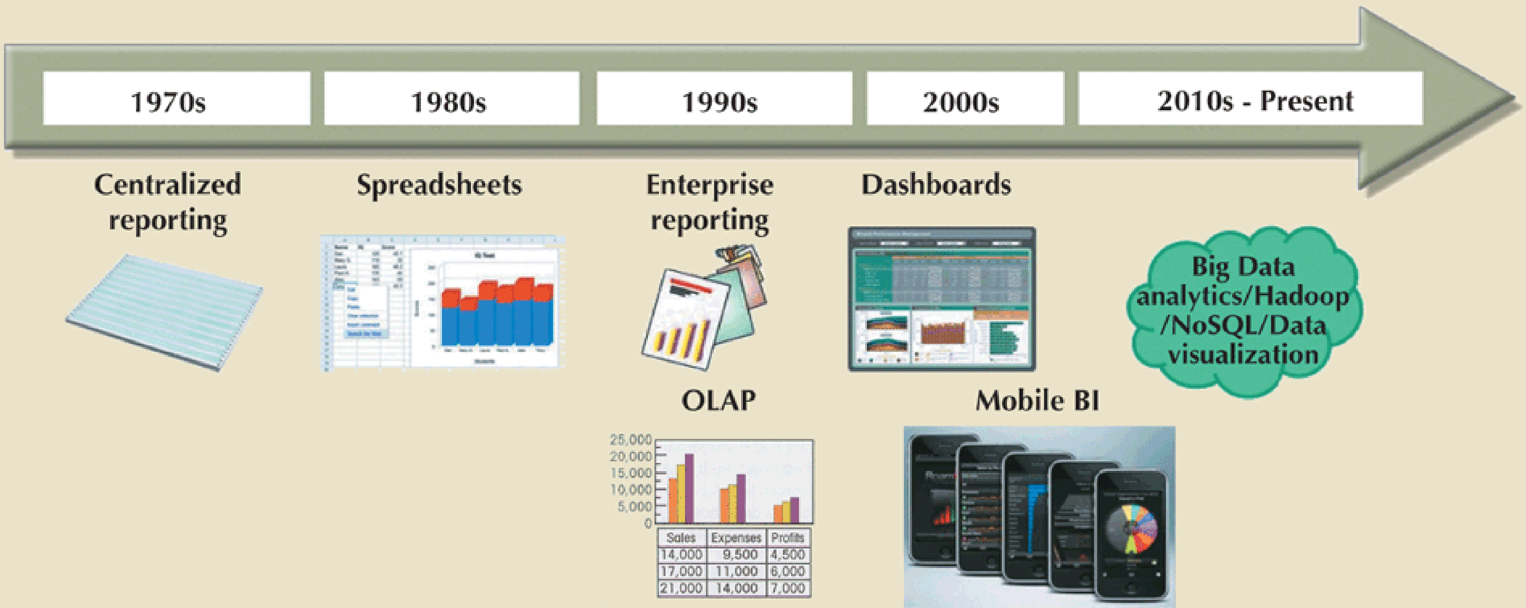
- Practices to manage data
  - Master data management (MDM): collection of concepts, techniques, and processes for identification, definition, and management of data elements
  - Governance: method of government for controlling business health and for consistent decision making
  - Key performance indicators (KPI): numeric or scale-based measurements that assess company's effectiveness in reaching its goals
    - General
    - Finance
    - Human resources
    - Education
  - Modern BI reporting styles
    - Advanced reporting
    - Monitoring and alerting
    - Advanced data analytics

# Business Intelligence Benefits

- Improved decision-making is the main goal of BI, but BI provides other benefits
  - **Integrating architecture:** BI systems integrate data from various sources into a unified platform, enabling seamless access and analysis.
  - **Common user interface for data reporting and analysis:** BI platforms offer user-friendly interfaces that allow users to access and analyze data easily, regardless of their technical expertise.
  - **Common data repository fosters single version of company data:** By centralizing data in a common repository, BI ensures that all users access the same up-to-date information, reducing discrepancies and improving decision-making accuracy.
  - **Improved organizational performance:** BI enables organizations to identify trends, patterns, and opportunities in their data, leading to better-informed decisions and ultimately improving overall performance.
- Achieving all these benefits takes a lot of human, financial, and technological resources, and time
  - BI benefits are not achieved overnight; are the result of a focused company-wide effort that could take a long time

# Business Intelligence Evolution (1 of 2)

FIGURE 13.2 EVOLUTION OF BI INFORMATION DISSEMINATION FORMATS




© Antun Hirsman/Shutterstock.com

The evolution of BI information dissemination formats has transitioned from static reports to dynamic dashboards and interactive visualizations, providing users with real-time insights and actionable intelligence.

# Business Intelligence Evolution (2 of 2)

- Decision support system (DSS) is an arrangement of computerized tools used to assist managerial decision making
- A Decision Support System (DSS) is a computer-based tool that helps users make decisions by providing relevant data, analysis, and modelling tools.
  - Typically has a much narrower focus and reach than a BI solution
- BI information technology has evolved from centralized reporting styles to the current, mobile BI and Big Data analytics style in the span of just a few years
  - The rate of technological change is not slowing down; technology advancements are accelerating the adoption of BI to new levels

*An example of a Decision Support System (DSS) is a sales forecasting tool used by a retail company. This system collects data on past sales, market trends, and other relevant factors to generate forecasts that assist managers in making decisions about inventory management, marketing strategies, and resource allocation.*

# Business Intelligence Technology Trends

- Several technological advances are driving the growth of business intelligence technologies
  - **Data storage improvements:** Enhancements in data storage technologies, such as faster access times and increased storage capacity, facilitate more efficient data handling.
  - **Business intelligence appliances:** Dedicated hardware systems designed to streamline BI processes, offering optimized performance for data analysis and reporting tasks.
  - **Business intelligence as a service:** Cloud-based BI solutions that provide on-demand access to analytics tools and capabilities, enabling organizations to leverage BI functionalities without extensive infrastructure investments.
  - **Big Data analytics:** Techniques and technologies for processing and analyzing large volumes of diverse data types to extract valuable insights and inform decision-making.
  - **Personal analytics:** Tools and applications that enable individuals to track and analyze personal data, such as health metrics or productivity metrics, to gain insights and make informed decisions.

# Decision Support Data

- Although BI is used at the strategic and tactical managerial levels within organizations, its effectiveness depends on the quality of data gathered at the operational level
  - Operational data is seldom well suited to decision-support tasks

The operational level in an organization refers to the day-to-day activities and transactions that occur within the business. This includes activities such as sales transactions, inventory management, customer interactions, and production processes. Operational data is generated from these activities and is typically used for immediate tasks and processes within the organization.

For example, in a retail store, the operational level involves activities like processing customer orders, managing inventory levels, and tracking sales transactions at the checkout counter. Data collected from these activities, such as sales figures, inventory levels, and customer purchase history, constitutes operational data.

# Operational Data versus Decision Support Data (1 of 3)

- **Operational data** and **decision support data** serve different purposes within an organization. **Operational data** primarily captures daily business transactions, providing a detailed record of activities such as sales, inventory movements, and customer interactions. On the other hand, **decision support data** transforms this operational data into strategic and tactical insights that help guide business decisions.

There are three main differences between decision support data and operational data:

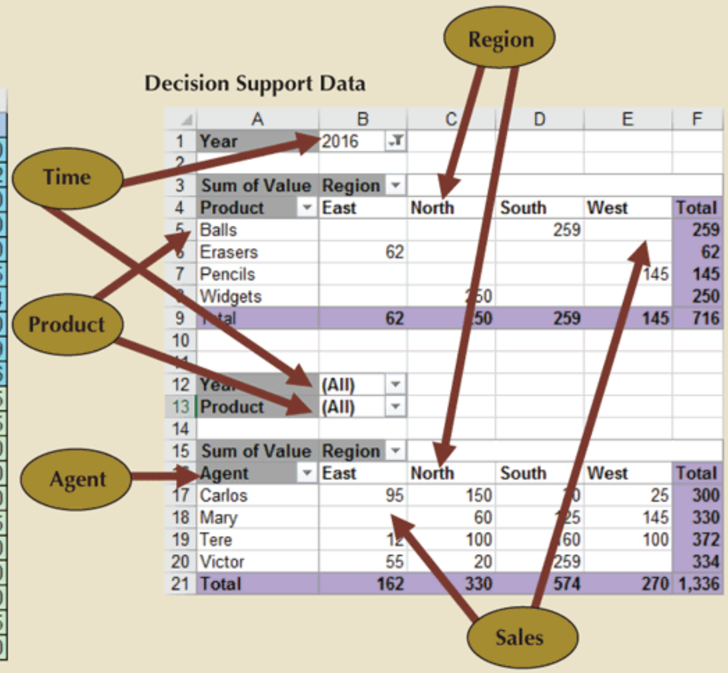
- **Time Span:** Operational data focuses on the present and recent past, capturing real-time or near-real-time transactions. Decision support data often spans longer time periods, including historical data that provides context and trends for analysis.
- **Granularity:** Operational data is typically detailed and transactional, capturing individual events or transactions at a low level of aggregation. Decision support data is often aggregated or summarized to a higher level, providing a broader perspective for analysis.
- **Dimensionality:** Operational data usually deals with a limited number of dimensions or attributes relevant to specific business transactions. Decision support data may involve more complex multidimensional analysis, incorporating additional dimensions such as time, geography, or product categories to provide deeper insights.

FIGURE 13.3 TRANSFORMING OPERATIONAL DATA INTO DECISION SUPPORT DATA

**Operational Data**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Year | Region | Agent | Product | Value |
| 2 | 2016 | East | Carlos | Erasers | 50 |
| 3 | 2016 | East | Tere | Erasers | 12 |
| 4 | 2016 | North | Carlos | Widgets | 120 |
| 5 | 2016 | North | Tere | Widgets | 100 |
| 6 | 2016 | North | Carlos | Widgets | 30 |
| 7 | 2016 | South | Victor | Balls | 145 |
| 8 | 2016 | South | Victor | Balls | 34 |
| 9 | 2016 | South | Victor | Balls | 80 |
| 10 | 2016 | West | Mary | Pencils | 89 |
| 11 | 2016 | West | Mary | Pencils | 56 |
| 12 | 2017 | East | Carlos | Pencils | 45 |
| 13 | 2017 | East | Victor | Balls | 55 |
| 14 | 2017 | North | Mary | Pencils | 60 |
| 15 | 2017 | North | Victor | Erasers | 20 |
| 16 | 2017 | South | Carlos | Widgets | 30 |
| 17 | 2017 | South | Mary | Widgets | 75 |
| 18 | 2017 | South | Mary | Widgets | 50 |
| 19 | 2017 | South | Tere | Balls | 70 |
| 20 | 2017 | South | Tere | Erasers | 90 |
| 21 | 2017 | West | Carlos | Widgets | 25 |
| 22 | 2017 | West | Tere | Balls | 100 |

**Decision Support Data**




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Year | 2016 | | | | |
| 2 | | | | | | |
| 3 | Sum of Value | Region | | | | |
| 4 | Product | East | North | South | West | Total |
| 5 | Balls | | | 259 | | 259 |
| 6 | Erasers | 62 | | | | 62 |
| 7 | Pencils | | | | 145 | 145 |
| 8 | Widgets | | 250 | | | 250 |
| 9 | Total | 62 | 250 | 259 | 145 | 716 |
| 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 | 30 | 25 | 300 |
| 18 | Mary | | 60 | 125 | 145 | 330 |
| 19 | Tere | 12 | 100 | 160 | 100 | 372 |
| 20 | Victor | 55 | 20 | 259 | | 334 |
| 21 | Total | 162 | 330 | 574 | 270 | 1,336 |

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

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

- Sales by product, region, agent, and so on
- Sales for all years or only a few selected years
- Sales for all products or only a few selected products

# Operational Data versus Decision Support Data (3 of 3)

| Table 13.5 Contrasting Operational and Decision Support Data Characteristics | | |
|---|---|---|
| **Characteristic** | **Operational Data** | **Decision Support Data** |
| Data currency | Current operations<br>Real-time data | Historic data<br>Snapshot of company data<br>Time component (week/month/year) |
| Granularity | Atomic-detailed data | Summarized data |
| Summarization level | Low; some aggregate yields | High; many aggregation levels |
| Data model | Highly normalized<br>Mostly relational DBMSs | Non-normalized<br>Complex structures<br>Some relational, but mostly multidimensional DBMSs |
| Transaction type | Mostly updates | Mostly query |
| Transaction volumes | High-update volumes | Periodic loads and summary calculations |
| Transaction speed | Updates are critical | Retrievals are critical |
| Query activity | Low to medium | High |
| Query scope | Narrow range | Broad range |
| Query complexity | Simple to medium | Very complex |
| Data volumes | Hundreds of gigabytes | Terabytes to petabytes |

CENGAGE

# Decision Support Database Requirements

- Database schema
  - Must support complex, non-normalized data representations
  - Data must be aggregated and summarized
  - Queries must be able to extract multidimensional time slices
- Data extraction and filtering
  - Allow batch and scheduled data extraction
  - Support different data sources and check for inconsistent data or data validation rules
  - Encourage advanced integration, aggregation, and classification
- Database size
  - Very large databases (VLDBs)
  - Advanced storage technologies
  - Multiple-processor technologies

# The Data Warehouse (1 of 3)

A data warehouse is a centralized repository that integrates data from various sources and is designed to support decision-making processes. Its components include:

1. **Integrated:** Data from different sources is consolidated and standardized to ensure consistency and accuracy across the warehouse.
2. **Subject-oriented:** The data is organized around key subject areas or topics relevant to the organization's business processes, making it easier to analyze and understand.
3. **Time-variant:** Data in the warehouse includes historical information, allowing users to analyze trends and patterns over time.
4. **Nonvolatile:** Once data is stored in the warehouse, it is not typically updated or changed. Instead, new data is added to the warehouse, preserving a historical record of changes and transactions.

These components collectively enable organizations to perform sophisticated analysis and reporting, leading to more informed decision-making.

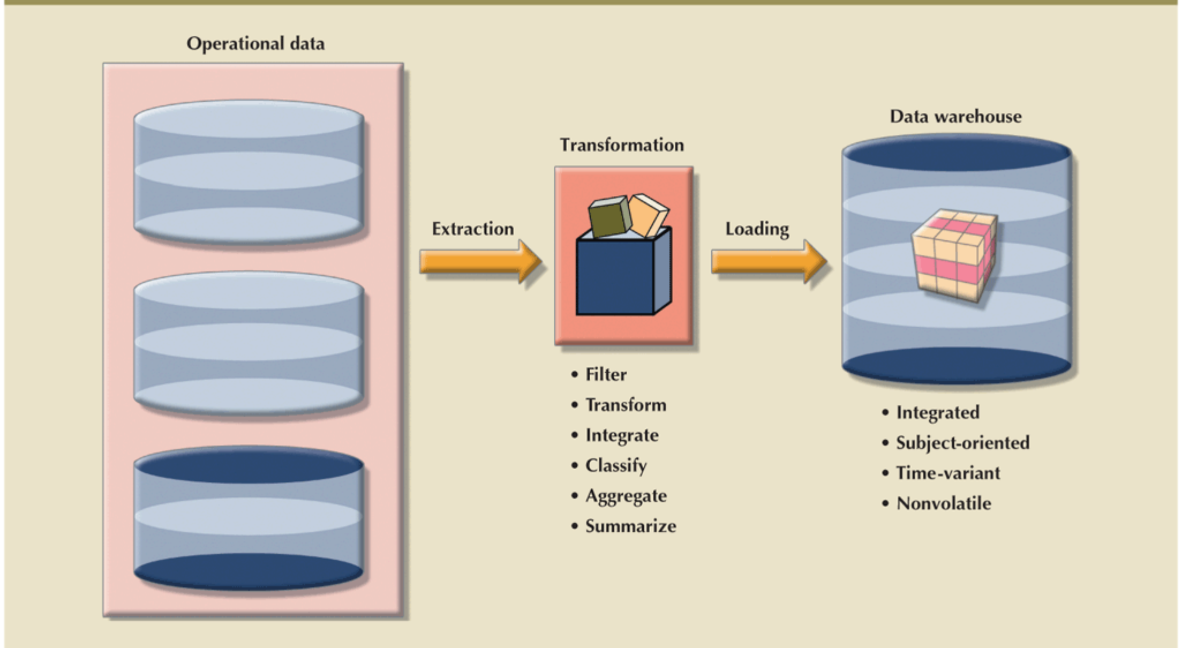
# The Data Warehouse (2 of 3)

| **Table 13.8 Characteristics of Data Warehouse Data and Operational Database Data** | | |
|---|---|---|
| **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 is stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, and credit amounts. | Data is stored with a subject orientation that facilitates multiple views of the data and decision making. For example, sales may be recorded by product, division, manager, or region. |
| Time-variant | Data is 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-2016. | Data is 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 is added only periodically from historical systems. Once the data is properly stored, no changes are allowed. Therefore, the data environment is relatively static. |

CENGAGE

# The Data Warehouse (3 of 3)

FIGURE 13.4 THE ETL PROCESS




The ETL (Extract, Transform, Load) process is integral to the functioning of a data warehouse. It begins with the extraction of data from various sources, which is then transformed to ensure consistency, quality, and compatibility with the data warehouse schema. This involves tasks like data cleansing, validation, enrichment, and aggregation. Finally, the transformed data is loaded into the data warehouse, where it is organized and stored for analysis and reporting. The ETL process ensures that the data warehouse contains accurate, consistent, and usable data to support decision-making processes.

# Data Marts

- Small, single-subject data warehouse subset
  - Provides decision support to a small group of people
- Benefits over data warehouses
  - Lower cost and shorter implementation time
  - Technologically advanced
  - Inevitable “people issues”

  - For example, a retail company might create a data mart specifically for analyzing sales performance in its online division. This data mart would contain relevant data such as customer transactions, website interactions, and product inventory for the online segment of the business. Managers and analysts responsible for the online sales channel could then use this data mart to track key metrics, identify trends, and make informed decisions to optimize online sales strategies.

CENGAGE

# Twelve Rules That Define a Data Warehouse

| Table 13.9 | Twelve Rules for a Data Warehouse |
| --- | --- |
| **Rule No.** | **Description** |
| 1 | The data warehouse and operational environments are separated. |
| 2 | The data warehouse data is integrated. |
| 3 | The data warehouse contains historical data over a long time. |
| 4 | The data warehouse data is snapshot data captured at a given point in time. |
| 5 | The data warehouse data is subject oriented. |
| 6 | The data warehouse data is mainly read-only with periodic batch updates from operational data. No online updates are allowed. |
| 7 | The data warehouse development life cycle differs from classical systems development. Data warehouse development is data-driven; the classical approach is process-driven. |
| 8 | The data warehouse contains data with several levels of detail: current detail data, old detail data, lightly summarized data, and highly summarized data. |
| 9 | The data warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities at a time. |
| 10 | The data warehouse environment has a system that traces data sources, transformations, and storage. |
| 11 | The data warehouse's metadata is a critical component of this environment. The metadata identifies and defines all data elements. The metadata provides the source, transformation, integration, storage, usage, relationships, and history of each data element. |
| 12 | The data warehouse contains a chargeback mechanism for resource usage that enforces optimal use of the data by end users. |

CENGAGE

# Star Schemas (1 of 5)

- Data-modeling technique
  - Maps multidimensional decision support data into a relational database
  - Creates the near equivalent of multidimensional database schema from existing relational database
  - Yields an easily implemented model for multidimensional data analysis

# Star Schemas (2 of 5)

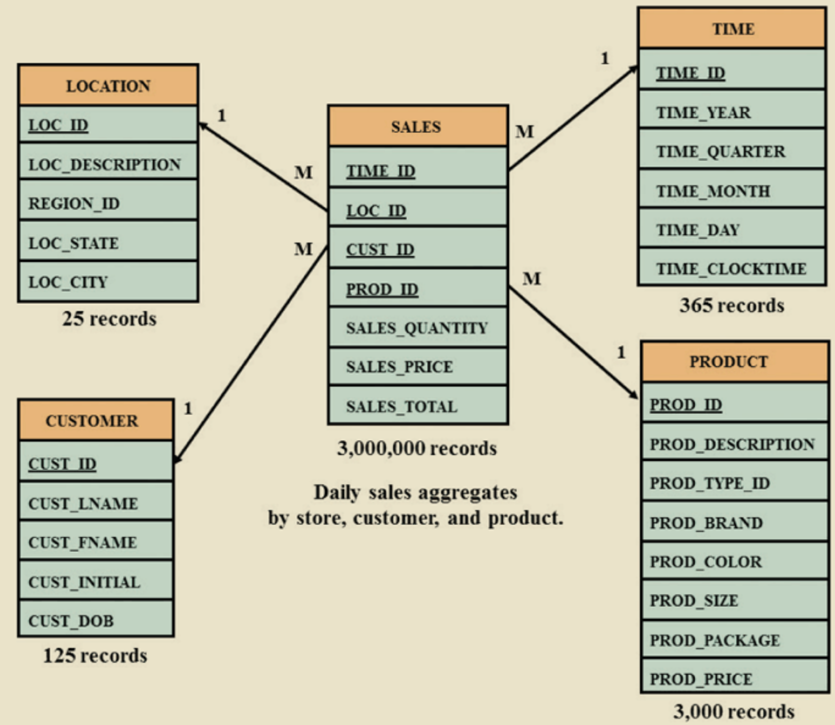
- Basic star schema components
  - Facts: numeric values that represent a specific business aspect
  - Dimensions: qualifying characteristics that provide additional perspectives to a given fact
  - Attributes: used to search, filter, and classify facts
    - Slice and dice: ability to focus on slices of the data cube for more detailed analysis
  - Attribute hierarchies: provide a top-down data organization
    - Aggregation and drill-down/roll-up data analysis

# Star Schemas (3 of 5)

- Star schema representation
  - Facts and dimensions represented by physical tables in data warehouse database
  - Many-to-one (M:1) relationship between fact table and each dimension table
- Fact and dimension tables
  - Related by foreign keys
  - Subject to primary and foreign key constraints
  - Primary key of a fact table
    - Composite primary key because the fact table is related to many dimension tables
    - Always formed by combining the foreign keys pointing to the related dimension tables

# Star Schemas (4 of 5)

FIGURE 13.10 STAR SCHEMA FOR SALES




The star schema for sales involves a central "Sales" entity surrounded by dimension entities like "Time," "Product," "Location," and "Customer." This design allows for efficient analysis of sales data with the "Sales" entity at the centre for easy querying and reporting.

# Star Schemas (5 of 5)

- Performance-improving techniques for the star schema
  - Normalizing dimensional tables
    - Snowflake schema: dimension tables can have their own dimension tables
  - Maintaining multiple fact tables to represent different aggregation levels
    - Save processor cycles at run time, thereby speeding up data analysis
  - Denormalizing fact tables
    - Improves data access performance and saves data storage space
  - Partitioning and replicating tables
    - Partitioning: splits tables into subsets of rows or columns and places them close to the client computer
    - Replication: makes copy of table and places it in a different location
    - Periodicity: provides information about the time span of the data stored in the table

# Online Analytical Processing (OLAP)

- Online analytical processing (OLAP) is a BI style whose systems share three main characteristics
  - Multidimensional data analysis techniques :allow users to analyze data from multiple perspectives or dimensions.
  - Advanced database support: ensures efficient storage, retrieval, and processing of large volumes of data for analysis.
  - Easy-to-use end-user interfaces: provide intuitive tools for querying, exploring, and visualizing data, enhancing user productivity and decision-making.

  - An example of OLAP is a financial reporting system that enables users to analyze financial data from different perspectives such as departments, cost centers, and time periods to assess performance, identify outliers, and make informed financial decisions.

# Multidimensional Data Analysis Techniques

- Multidimensional data analysis techniques :allow users to analyze data from multiple perspectives or dimensions.

- Data are processed and viewed as part of a multidimensional structure
  - Particularly attractive to business decision makers who tend to view business data as being related to other business data
- Augmented advanced functions
  - Data presentation
  - Data aggregation, consolidation, and classification
  - Computational
  - Data-modeling

# Advanced Database Support

- Advanced database support: ensures efficient storage, retrieval, and processing of large volumes of data for analysis.
- OLAP tools must have the following features to deliver efficient decision support:
  - Access to many different kinds of DBMSs, flat files, and internal and external data sources
  - Access to aggregated data warehouse data and operational database detail data
  - Advanced data navigation features
  - Rapid and consistent query response times
  - Ability to map end-user requests
  - Support for very large databases

*Aggregated data refers to data that has been summarized or combined from individual-level data into a higher-level view, typically to provide insights or analysis at a broader level.

- An example of aggregated data would be monthly sales figures for a retail store, where individual daily sales transactions are summarized into total sales for each month.

# Easy-to-Use End-User Interface

- Easy-to-use end-user interfaces: provide intuitive tools for querying, exploring, and visualizing data, enhancing user productivity and decision-making.

- When proper implementation leads to simple navigation and accelerated decision-making or data analysis
  - Advanced OLAP features are more useful when access is kept simple
  - Many interface features are borrowed from previous generations of data analysis tools

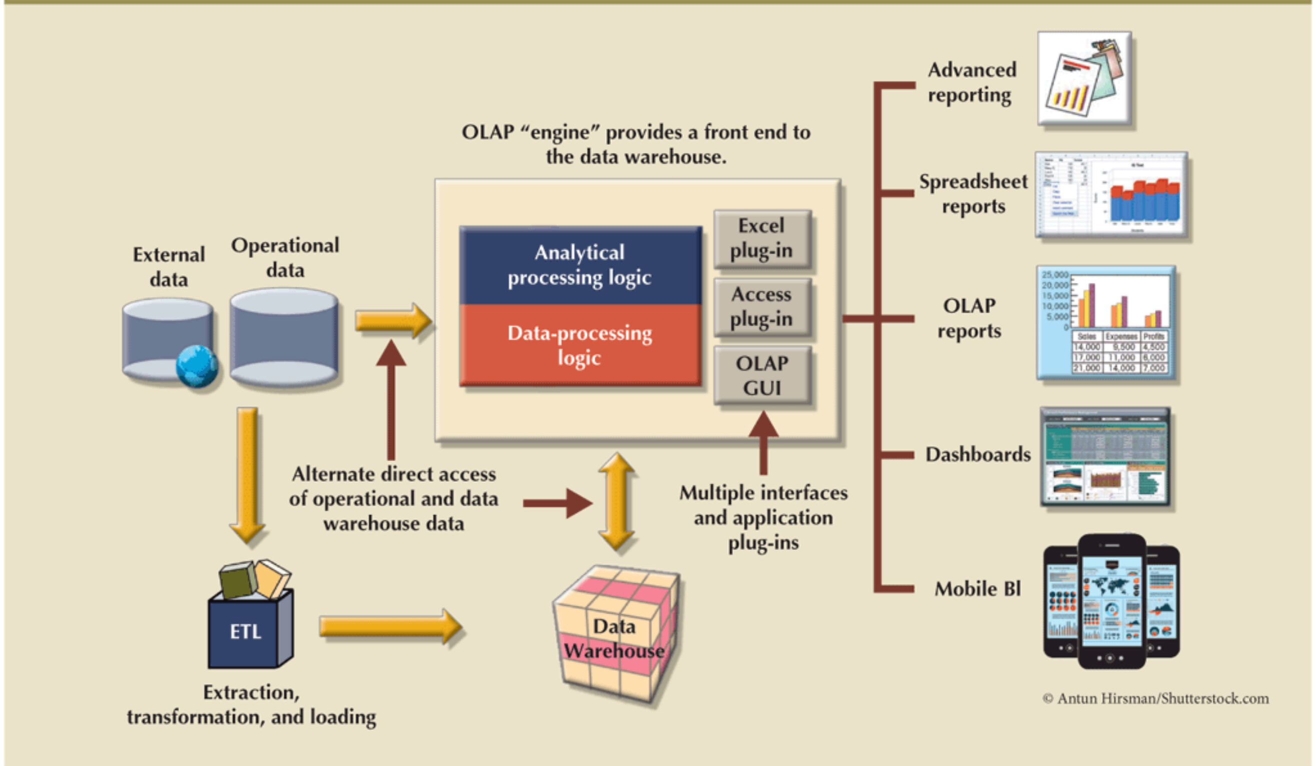
# OLAP Architecture (1 of 2)

- Designed to meet ease-of-use requirements while keeping the system flexible
- Main architectural components
  - Graphical user interface (GUI)
  - Analytical processing logic
  - Data-processing logic

  - The user interacts with the system through a Graphical User Interface (GUI), which then applies analytical processing logic to manipulate data according to user queries or commands. This processed data is then handled by the data-processing logic, ensuring that the system remains flexible and user-friendly.

# OLAP Architecture (2 of 2)

FIGURE 13.16 OLAP ARCHITECTURE




OLAP architecture comprises three main components: a Graphical User Interface (GUI) for user interaction, analytical processing logic to handle data manipulation based on user queries, and data-processing logic to ensure flexibility and user-friendliness. Through this architecture, users can easily navigate and analyze multidimensional data, enabling efficient decision-making and data exploration.

# Relational OLAP

- Relational online analytical processing (ROLAP)
  - Provides OLAP functionality using relational databases and familiar relational tools to store and analyze multidimensional data
- Extensions added to traditional RDBMS technology
  - Multidimensional data schema support within the RDBMS
  - Data access language and query performance optimized for multidimensional data
  - Support for very large databases (VLDBs)

  - *ROLAP, or Relational Online Analytical Processing, is a type of OLAP that operates directly with relational databases. Unlike traditional OLAP systems that store data in multidimensional cubes, ROLAP relies on relational database management systems (RDBMS) to perform data aggregation and analysis. ROLAP systems leverage SQL queries to access and manipulate data stored in relational tables, providing users with the flexibility to work with large datasets and complex queries while maintaining compatibility with existing database infrastructure.*

# Multidimensional OLAP

- Multidimensional online analytical processing (MOLAP)
  - Extends OLAP functionality to multidimensional database management systems (MDBMSs)
    - MDBMS uses proprietary techniques store data in matrix-like *n*-dimensional arrays
    - End users visualize stored data as a  three dimensional data cube
      - Grow to *n* number dimensions, thus becoming hypercubes
      - Held in memory in a cube cache to speed access
  - Sparsity: measures density of data held in the data cube

  - *MOLAP, or Multidimensional Online Analytical Processing, is a type of OLAP that stores data in a multidimensional database format. Unlike ROLAP, which directly queries relational databases, MOLAP systems organize data into cubes for efficient analysis. These cubes precalculate and store aggregated data along multiple dimensions, enabling fast query performance and interactive analysis. MOLAP systems are well-suited for scenarios where performance and scalability are critical, providing users with powerful tools for exploring and visualizing data across various dimensions.*

# Relational versus Multidimensional OLAP

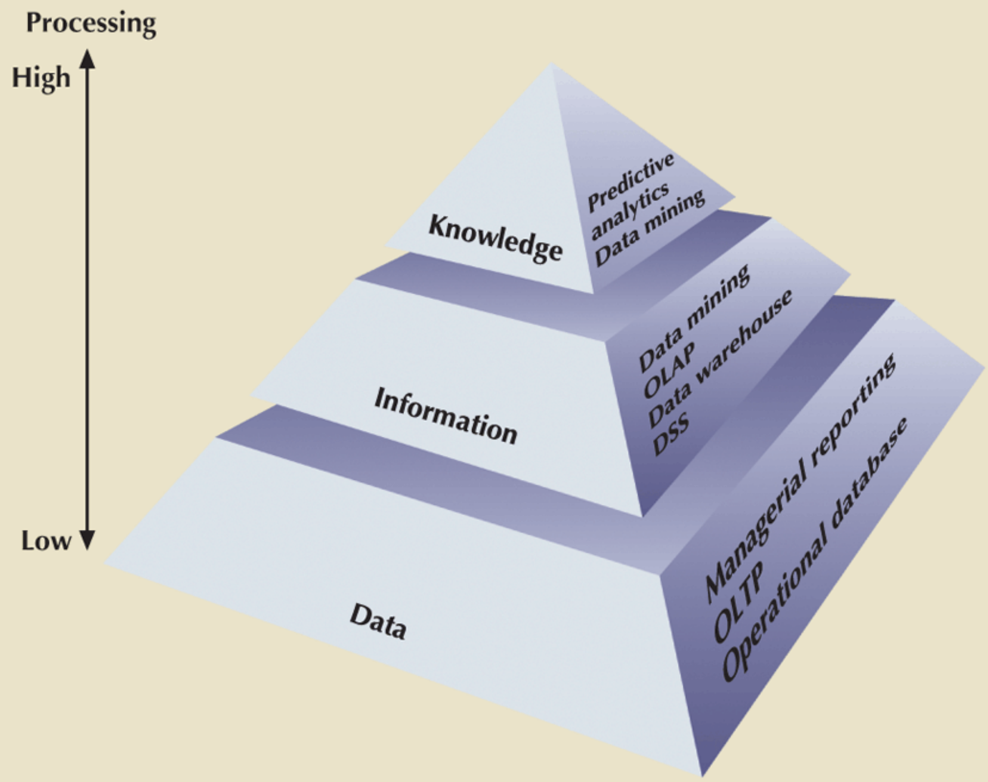
| Table 13.12 Relational vs. Multidimensional OLAP | | |
|---|---|---|
| Characteristic | ROLAP | MOLAP |
| Schema | Uses star schema<br>Additional dimensions can be added dynamically | Uses data cubes<br>Multidimensional arrays, row stores, column stores<br>Additional dimensions require re-creation of the data cube |
| Database size | Medium to large | Large |
| Architecture | Client/server<br>Standards-based | Client/server<br>Open or proprietary, depending on vendor |
| Access | Supports ad hoc requests<br>Unlimited dimensions | Limited to predefined dimensions<br>Proprietary access languages |
| Speed | Good with small data sets; average for medium-sized to large data sets | Faster for large data sets with predefined dimensions |

# Data Analytics (1 of 4)

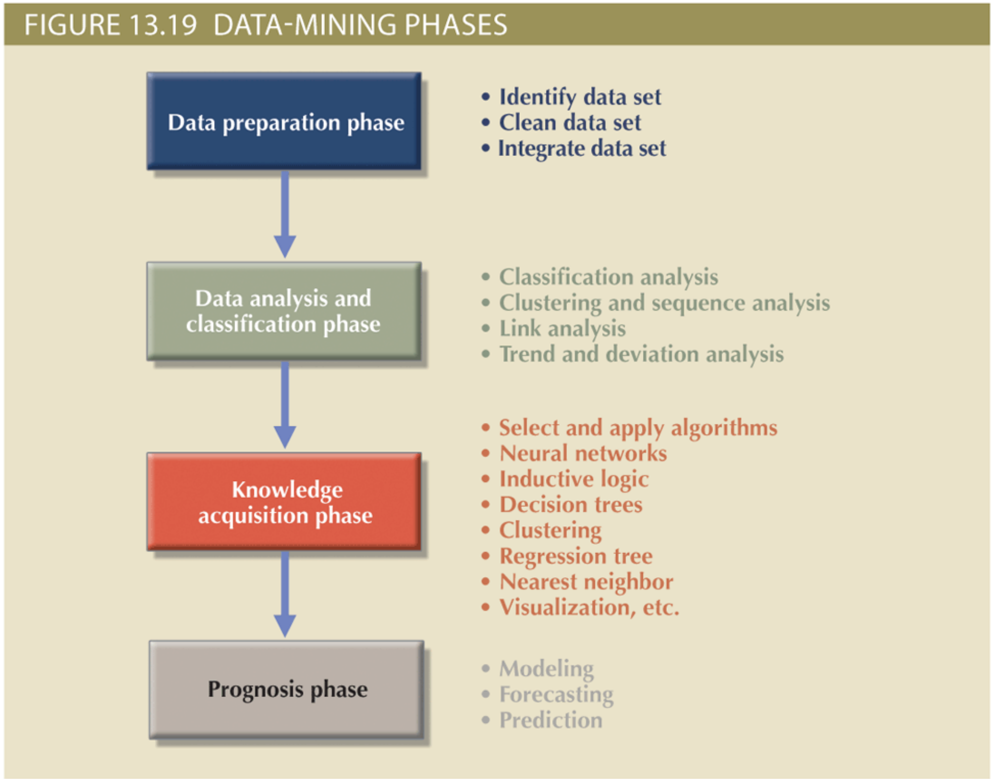
- Subset of business intelligence (BI) functionality that encompasses a wide range of mathematical, statistical, and modeling techniques with the purpose of extracting knowledge from data
  - Explanatory analytics: focuses on discovering and explaining data characteristics and relationships based on existing data
  - Predictive analytics: focuses on predicting future data outcomes with a high degree of accuracy
- Data mining focuses on the discovery and explanation stages of knowledge acquisition
  - Analyzing massive amounts of data to uncover hidden trends, patterns, and relationships; to form computer models to simulate and explain the findings; and to use such models to support business decision making

# Data Analytics (2 of 4)

FIGURE 13.18 EXTRACTING KNOWLEDGE FROM DATA



# Data Analytics (3 of 4)




In data mining, the prognosis phase refers to the stage where predictive models are developed based on historical data. This phase involves training machine learning algorithms on the available data to predict future outcomes or trends. The goal is to build models that can accurately forecast future events or behaviors based on the patterns discovered during the data exploration and modeling phases.

# Data Analytics (4 of 4)

- Predictive analytics focuses on creating actionable models to predict future behaviors and events
  - Employs mathematical and statistical algorithms, neural networks, artificial intelligence, and other advanced modeling tools to create actionable predictive models based on available data
  - Used in areas such as customer relationships, customer service, customer retention, fraud detection, targeted marketing, and optimized pricing

# SQL Analytic Functions (1 of 2)

- The ROLLUP extension
  - Used with GROUP BY clause to generate aggregates by different dimensions
  - Enables subtotal for each column listed except for the last one, which gets a grand total
- The CUBE extension
  - Used with GROUP BY clause to generate aggregates by the listed columns
  - Enables you to get a subtotal for each column listed in the expression, in addition to a grand total for the last column listed

*For example, consider a table containing sales data with columns for product, region, and sales amount. Using the ROLLUP extension, you can generate subtotals for each product and region combination as well as a grand total for all regions and products combined. Similarly, with the CUBE extension, you can obtain subtotals for each product, region, and sales amount combination along with a grand total for all possible combinations.*

# SQL Analytic Functions (2 of 2)

- Materialized views
  - Dynamic table that contains SQL query command to generate rows and stores the actual rows
    - Created the first time query is run and summary rows are stored in the table
    - Automatically updated when base tables are updated
    - Requires specified privileges

    Materialized views are precomputed tables that store the results of a SQL query, allowing for quick access to aggregated or computed data. They are created based on a query's results and are automatically updated when underlying data changes. This provides improved query performance and reduces the need to repeatedly execute complex queries.

    Imagine you have a list of daily sales transactions stored in a table. Each transaction includes the product sold and the quantity purchased. Now, you want to create a summary of total sales for each product. Instead of running a complex query every time you need this information, you can create a materialized view called product_sales_summary. This view would contain pre-calculated totals for each product, making it easy to query and analyze total sales without recalculating the sums each time. So, the materialized view product_sales_summary provides a snapshot of the total sales for each product, making it more efficient to retrieve this information whenever needed.

# Data Visualization (1 of 2)

- Process of abstracting data to provide a visual data representation that enhances the user's ability to comprehend the meaning of the data
  - Goal is to allow the user to quickly and efficiently see the data's big picture by identifying trends, patterns, and relationships
- The need for data visualization
  - Ability to zoom in and out, drill down and up, filter, etc. is one of the many advantages of current breed of data visualization tools
  - Makes it easier to understand data—in particular, large amounts of data
- The science of data visualization
  - Roots in cognitive sciences: how the human brain receives, interprets, organizes, and processes information
    - Pattern recognition
    - Spatial awareness
    - Aesthetics

# Data Visualization (2 of 2)

- Understanding the data
  - The same data can be presented in multiple ways
- In In qualitative data, nominal and ordinal are two types of categorical variables used to classify observations or responses:
- Nominal: Nominal variables represent categories or labels that do not have a natural order or ranking. Examples include gender, ethnicity, or marital status. Nominal variables only indicate differences between categories, not the magnitude of those differences.
- Ordinal: Ordinal variables represent categories with a natural order or ranking. While the categories have a relative position, the intervals between them may not be equal or precisely quantifiable. Examples include ratings (e.g., low, medium, high), Likert scale responses, or educational levels. Ordinal variables allow for comparisons in terms of relative magnitude or order, but the exact differences between categories may not be uniform or well-defined.

# Summary (1 of 2)

- Business intelligence (BI) is a term for a comprehensive, cohesive, and integrated set of applications used to capture, collect, integrate, store, and analyze data with the purpose of generating and presenting information to support business decision making
- Decision support systems (DSSs) refer to an arrangement of computerized tools used to assist managerial decision making within a business
- Operational data is not well suited for decision support
- The data warehouse is an integrated, subject-oriented, time-variant, nonvolatile collection of data that provides support for decision making
- The star schema is a data-modeling technique used to map multidimensional decision support data into a relational database for advanced data analysis

# Summary (2 of 2)

- Online analytical processing (OLAP) refers to an advanced data analysis environment that supports decision making, business modeling, and operations research
- Data analytics is a subset of BI functionality that provides advanced data analysis tools to extract knowledge from business data
- Data mining automates the analysis of operational data to find previously unknown data characteristics, relationships, dependencies, and trends
- SQL has been enhanced with analytic functions that support OLAP-type processing and data generation
- Data visualization provides visual representations of data that enhance the user's ability to comprehend the meaning of the data