LESSON PREVIEW

Databases

Whether you realize it or not, unless you are living completely off the grid, you are interacting with databases from the time you wake up in the morning until you set down your last digital device at night. Wake up and check the weather report? It comes from a database that uses your location to pull *your* weather report from the weather everywhere in the world. Get directions to the coffee shop where you're meeting a study partner? All that comes from a database. Want to order a new smartphone case? The search engine you use has all its information stored in a database, and the site you select to place your order has all its inventory and customer accounts in more databases.

No matter what your future employment goals are, it is certain you will need to use databases. That is what this module is about.

After completing this section, you will be able to:

- Explain the purpose of database management systems (DBMS) and how they solve file management issues.
- Explain DBMS tables, records, and fields and how they are related with primary and foreign keys.
- Explain the use of SQL to generate queries, forms, and reports.
- Explain the use of data normalization and ERDs.
- Explain other types of databases, including NoSQL and Cloud databases.
- Describe the concept of Big Data, including the four *V*s.
- Describe and explain the uses of BI, data warehouses, and data marts.
- Explain how in-memory computing and analytic platforms, including OLAP, assist businesses.
- Describe the processes and uses of data mining and web mining.
- Explain how information policies are used and administered.
- Describe the role of a database administrator and data quality audits.
- Describe the process of data scrubbing.

NOTES

Database and DBMS Overview

A **database** is a collection of data organized in a manner that allows a computer to quickly search for and retrieve information. It is a collection of tables and the relationships between those tables.

Today's companies and organizations use a **database management system (DBMS)**. A DBMS is a computer program that is used to create, process, and administer a database.

NOTES

Relational Databases: How DBMS Solves File Management Issues

Relational databases organize data into tables, which hold information about the objects represented in the database. This information is stored in **rows** called **records** or **objects**, and in **columns** called **fields**. When a field is common to two or more tables, it can be used to create a relationship. In this way data can be accessed without reorganizing the tables.

A DBMS helps to overcome issues associated with traditional file management systems, such as:

- **Data redundancy**: DBMS can eliminate duplication of data, reduce storage and storage fees, and create more efficient access times.
- **Data inconsistency**: With a DBMS, a single change to data in one place ensures consistency everywhere in the database.
- **Data security**: A DBMS allows for various users to have specific access to only what they need. This protects data and users from identity theft, data leaks, and the misuse of data.

A relational database system is known as an **RDBMS** (relational database management system).

NOTES

Tables, Keys, and Referential Integrity

A group of tables, which are defined by data in rows and columns, creates a database. Each database must have a table with one column that contains a **unique identifier**, called the primary key.

A **primary key** is something that is unique to each item in each row. For example, in a university database, the ID number of each student would be the primary key since only one student can have a specific ID number. By using the primary key of any particular student, all the different information about that student in all the database tables can be accessed.

A **foreign key** is a field that is linked to another table's primary key field in a relationship between two tables.

Referential integrity refers to the accuracy and consistency of data within a table relationship. It requires that whenever a foreign key value is used it must reference a valid, existing primary key in the parent table.

NOTES

Data Definition and Data Dictionaries

In the context of a database, **data** refer to all the single items stored in the database.

A **data dictionary** is a file or set of files containing a database's **metadata**. Metadata are data that describe the data. The data dictionary is a crucial component of a relational database, but typically only database administrators interact with the data dictionary.

NOTES

Fields, Records, Forms, and Reports

A **field** is a column in a table that represents a characteristic of something or someone.

A **record** is a row in a table that includes a collection of fields.

Forms are used to control how data are entered into a database. Forms structure data input to ensure data integrity. Users enter data into blank areas of the form.

Reports offer a way to view, format, and summarize the information in a database. Reports can be used to display or distribute a summary of data or archive snapshots of the data. Reports can also be used to provide details about individual records and to create labels.

NOTES

Using SQL for Queries, Forms, and Reports

SQL (pronounced “ess-que-el”or “sequel”—both are acceptable) is the standard programming database language used for human interface and communication with relational databases. Users generate code **(SQL statements)** to answer questions against the database.

SQL statements can perform a variety of database tasks, such as the retrieval of data, updating, adding to and deleting data, and generating forms and reports. When SQL statements are generated to get information from a database, this is called a query. SQL statements can also be used to generate a report about the results of a query.

A **query** is a request for information from one or more tables in a database. Queries allow a user to find detailed and exact data quickly by filtering information in the database.

NOTES

Data Normalization and Entity Relationship Diagrams

Normalization is an organized approach to breaking down and/or simplifying tables to eliminate data redundancy and undesirable data characteristics.

An **entity relationship diagram (ERD)** is a method used to structurally represent a database design via the use of diagrams. An ERD involves the use of different symbols and connectors that help to visualize two different types of information: the entities within the system and the interrelationships among these entities.

NOTES

Other Types of Databases

Today, the volume of data is so large and is changing so rapidly with the increased use of web services and social media that companies use nonrelational databases. These **NoSQL databases** are designed to handle huge data sets across many platforms and can use data pulled from continuously changing data, such as that generated by social media and web apps.

Cloud databases are built and accessed via a Cloud platform. They support both SQL and NoSQL databases and are accessed through the web or through a vendor-provided API (Application Programming Interface). Users can host these databases without having to buy and maintain dedicated hardware since this is provided by the vendor.

NOTES

Uses of DBMS in Today's Business Environment

Business data are constantly changing and evolving, creating challenges for organizations that seek to capitalize on the data. DBMS allow for the capture, processing, management, and reporting of an organization's data and are used for the following:

- Sorting and searching: This makes it easier for customers and database administrators to find information in a database.
- Relationship identification: DBMS allow users to identify relationships within the data. Identification of data relationships aids in informed decision making.
- Data storage: DBMS allow for the storage and retrieval of critical data.

NOTES

Big Data and the Four *V*s

Big Data includes all the analysis tools and processes related to applying and managing large volumes of data. It allows organizations to use analytics to help uncover predictive behavior. The four *V*s are the common characteristics of Big Data.

- **Volume** refers to the scale of data; it is estimated that 2.3 trillion gigabytes of information are created each day.
- **Variety** refers to the different forms of data. Data come from many sources, including social media platforms, email, photos, videos, and POS interactions.
- **Veracity** means the data are meaningful, true, and useful. Poor-quality data are estimated to cost the U.S. economy over $3.3 trillion per year.
- **Velocity** refers to the need of Big Data to analyze data as they are changing with unprecedented speed. Every minute on Facebook there are 510,000 comments posted, 293,000 statuses updated, and 136,000 photos uploaded.

NOTES

Business Intelligence, Data Warehouses, and Data Marts

Business intelligence (BI) includes the technologies, computer applications, and procedures for the collection, analysis, and presentation of business information to help support decision making. BI systems provide a picture of historic, current, and future views of operations using information stored in data warehouses, data marts, in-memory computing, and other analytic platforms.

A **data warehouse** is a repository of data and information that organizations analyze to make informed business and operational decisions. Data scientists, key decision makers, and data specialists use BI tools, SQL, and other analytics applications to interpret the data. Data warehouses allow businesses to gauge the performance of an enterprise over time and compare different time periods to make more informed business decisions.

A **data mart** is a subsection of a data warehouse that is designed and built specifically for individual departments or business functions. There are three types of data marts: dependent, independent, and hybrid. Data marts collect and measure data from specific operational areas of a business and are used by individual departments or groups. They track inventories, purchase transactions, and follow the supply chain to analyze what data a user needs.

NOTES

In-Memory Computing and Analytic Platforms

In-memory computing uses middleware software to help store data across a group of different computers. It allows for more RAM to be used than what is offered by a single user's operating system and provides for parallel processing of data. It helps to increase the amount and speed at which data can be ingested and analyzed.

Data analytics refers to the analysis of raw data to help make conclusions about the information contained in the data. Organizations use data analytics tools to extract and transform data from large data sets into useful information. Analytic platforms use processes and algorithms to convert raw data into usable information.

Analytic platforms assist large data-driven companies to analyze and interpret data. Various software providers have developed high-speed platforms that are used with both relational and nonrelational database technologies to provide information such as customer analytics, sales and marketing analytics, social media analytics, cybersecurity, and facilities data.

NOTES

OLAP: How It Supports Multidimensional Data Analysis

Online analytical processing (OLAP) is included in many BI software applications and is used for report creation and analysis, complex calculations, forecasting, budgeting, planning, and predictive analysis. OLAP stores information in multidimensional structures using **data cubes**.

Most OLAP tools are built on basic analytical operations that produce reports, including:

- **Consolidation** (also call roll-up): allows the analyst to summarize data.
- **Drill-down**: allows analysts to dig deeper into the data.
- **Slicing**: refers to the analysis of one level of information.
- **Dicing**: refers to the analysis of data from multiple dimensions.

NOTES

Data Mining

Data mining refers to the methods used to search large data stores and use the data to uncover patterns and trends.

Data mining is executed using mathematical algorithms that segment the data and evaluate the likelihood of future events. It is a multidisciplinary skill that uses machine learning, statistics, and artificial intelligence (AI) and database technology.

The four key properties of data mining include:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

Data mining assists businesses to uncover information that may be hidden in data sets, but it does not offer specific reasons why this information may be valuable. Interpretation of results is usually executed by managers and data analysts. Data mining yields probabilities, not exact answers.

NOTES

Web Mining

Web mining uses the principles of data mining to uncover and extract information from websites, social media sites, e-commerce platforms, and web services. Organizations uses web mining to gain a better understanding of consumer behavior, website efficacy, and usage patterns and to analyze how web searches are used.

Three techniques commonly used to gain information from the web are:

- **Web content mining (WCM)**: extraction of information from web pages and documents, including text, images, videos, and other interactives
- **Web structure mining (WSM)**: analysis of hyperlinks, nodes, and related web pages
- **Web usage mining (WUM)** (also called log mining): analysis of web access logs (i.e., the when, how, and frequency that websites are accessed)

Businesses and organizations use web mining to improve a user's web experience and, therefore, improve their outcomes. This leads to improved website visibility, usability, and accessibility.

Accessing databases via the web is valuable in today's world because most users can easily access and use web browsers and digital devices.

NOTES

Policies, Administration, and Data Governance

Database policies refer to how information in a database should be handled and disseminated. Ensuring data security and structure is critical. These policies specify the rules used in database design, who has access to the data, how the data are collected and maintained, and where information and data are distributed.

Data administration is responsible for the policies and procedures that are used to manage an organization's data. These data administrators ensure that policies and procedures covering organizational data are created and followed.

Data governance (DG) includes the personnel, processes, and technology needed to oversee and secure an organization's data and data assets. Data governance policies ensure that an organization's data are valid, understandable, complete, and accessible. Areas covered in digital governance include data architecture, data quality, data modeling, data warehousing and BI, and data security.

The goals of data governance include risk mitigation, rules for data use, compliance to requirements, cost reduction, and high-quality data. Data governance should determine:

- Where/who: the goals and objectives of the organization
- What: the aspects of the business that governance should cover
- How: the technical aspects and specifications

NOTES

The Database Administrator

A **database administrator (DBA)** is a technically specific role that is usually part of an organization's IT department. Database administrators monitor and troubleshoot the database to ensure it is functional and available when needed. Specific DBA tasks include database security, tuning, backup, and the creation of queries and reports that are used to assist business decisions. The responsibilities of a database administrator include:

- Development: setting database requirements, establishing a data model, design, hardware and software selection, and design evaluation
- Operation: supervising rights, security management, training, DBMS management, and performance tracking/monitoring
- Backup and recovery: establishing a backup and recovery plan, monitoring backup procedures, and recovery management
- Adaptation: setting up a system to make changes to the DBMS and managing configuration changes

NOTES

Data Quality Audits

Poor quality or invalid data pose many risks. To ensure that an organization's data are of the highest quality, a **data quality audit** can be used. A data quality audit uses statistical analysis to examine data, variabilities, and outcomes against test data. Executing a data quality audit can reduce the risk of data inconsistency, reduce data storage costs, and make recommendations for data improvement.

The process of analyzing data includes:

- Quality assessment: analyzes the quality of source data and includes data profiling, which is the process of investigating data abnormalities and redundancies
- Data design: involves the creation of quality processes used to manage data
- Quality transformation: incorporates correction maps that correct issues in source data
- Quality monitoring: examines data over a given amount of time to ensure data rules are being followed and that data are valid

NOTES

Data Scrubbing for Data Integrity

Data scrubbing (or data cleansing) refers to detecting errors in data sets and removing or correcting these mistakes to ensure data validity. Specialized data-scrubbing software is commonly used to analyze, correct errors in the data, and integrate the data to make the data compliant with the organization's policies and procedures.

Data scrubbing can resolve and fix such errors as duplicate database records, misspellings, incorrect names and addresses, and syntax issues. Common elements of a data scrubbing plan include:

- Identifying bad data and data sources: Data are checked to ensure all information matches with rules and guidelines.
- Data cleaning: Bad data are removed or replaced with clean data.
- Updating: Data sets are brought up to date.

NOTES

Design element: Vector abstract geometric background; -strizh-/Getty Images