Week 1
Understanding Data

Learning Objectives

- In this week, you will learn:
  - Investigate and discuss the difference between data and information
  - Explore what a database is, the various types of databases, and why they are valuable assets for decision making
  - Begin to form a basis on importance of database design
  - Explore how modern databases evolved from file systems
  - Compare and explain the differences between a traditional database system and a database management system.
  - Explore the introductory elements of the field of data mining and how this evolved from information technology.
  - Examine the various data types most often being mined along with more complex types such as dynamic data from the Web or social media.
  - Review the general approaches to the classification of data mining tasks including the use of technologies.

Data versus Information

Data	Information

• Raw facts
  • Have not yet been processed to reveal their meaning to the end user
• Building blocks of information
• **Data management**
  • Generation, storage, and retrieval of data
	• Produced by processing raw data to reveal its meaning
	• Requires context
	• Bedrock of **knowledge**
	• Should be accurate, relevant, and timely to enable good decision making

Introducing the Database

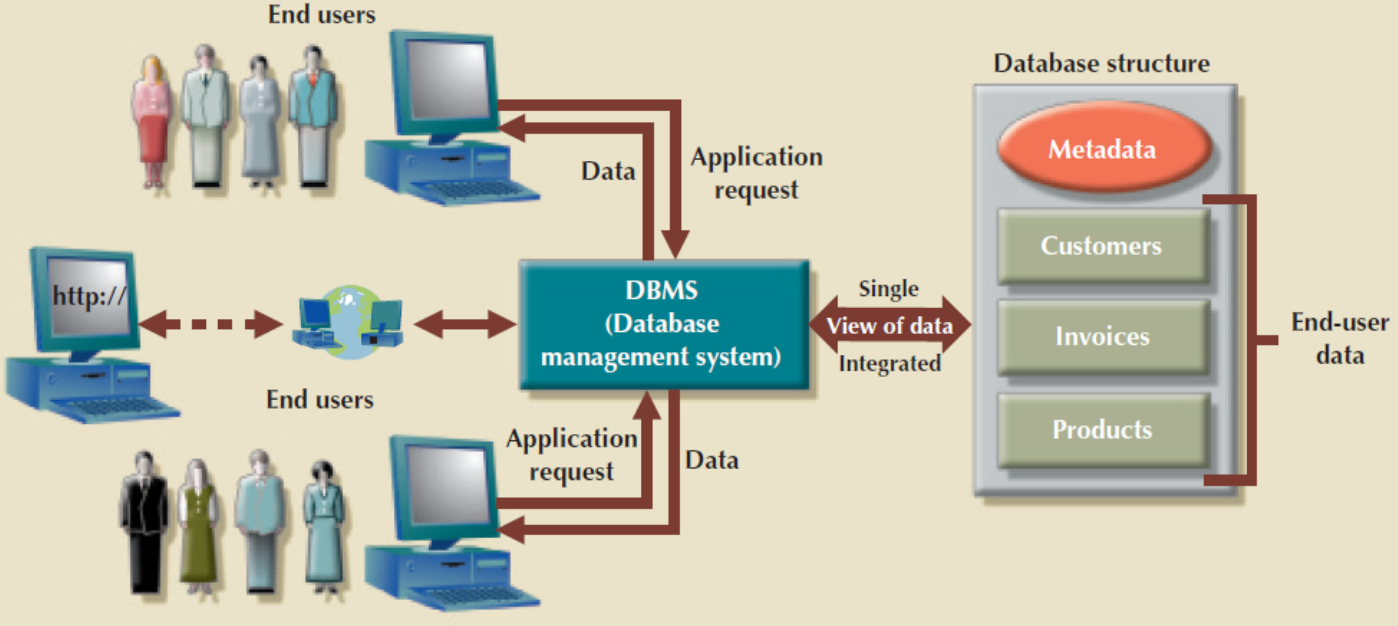
- Shared, integrated computer structure that stores a collection of:
  - End-user data - Raw facts of interest to end user
  - **Metadata**: Data about data, which the end-user data are integrated and managed
    - Describe data characteristics and relationships
- **Database management system (DBMS)**
  - Collection of programs
  - Manages the database structure
  - Controls access to data stored in the database

Role of the DBMS

- Intermediary between the user and the database
- Enables data to be shared
- Presents the end user with an integrated view of the data
- Receives and translates application requests into operations required to fulfill the requests
- Hides database's internal complexity from the application programs and users

Figure 1.3 - The DBMS Manages the Interaction between the End User and the Database

FIGURE 1.3 THE DBMS MANAGES THE INTERACTION BETWEEN THE END USER AND THE DATABASE



Advantages of the DBMS

- Better data integration and less data inconsistency
  - **Data inconsistency**: Different versions of the same data appear in different places
- Increased end-user productivity
- Improved:
  - Data sharing
  - Data security
  - Data access
  - Decision making
- **Data quality**: Accuracy, validity, and timeliness of data

Types of Databases

- **Single-user database**: Supports one user at a time
  - **Desktop database**: Runs on PC
- **Multiuser database**: Supports multiple users at the same time
  - **Workgroup databases**: Supports a small number of users or a specific department
  - **Enterprise database**: Supports many users across many departments

Types of Databases

- **Centralized database**: Data is located at a single site
- **Distributed database**: Data is distributed across different sites
- **Cloud database**: Created and maintained using cloud data services that provide defined performance measures for the database

Types of Databases

- **General-purpose databases**: Contains a wide variety of data used in multiple disciplines
- **Discipline-specific databases**: Contains data focused on specific subject areas
- **Operational database**: Designed to support a company's day-to-day operations

Types of Databases

- **Analytical database**: Stores historical data and business metrics used exclusively for tactical or strategic decision making
  - **Data warehouse**: Stores data in a format optimized for decision support
  - **Online analytical processing (OLAP)**
    - Tools for retrieving, processing, and modeling data from the data warehouse
- **Business intelligence**: Captures and processes business data to generate information that support decision making

Types of Databases

- **Unstructured data**: It exists in their original state
- **Structured data**: It results from formatting
  - Structure is applied based on type of processing to be performed
- **Semistructured data**: Processed to some extent
- **Extensible Markup Language (XML)**
  - Represents data elements in textual format

Database Design

- Focuses on the design of the database structure that will be used to store and manage end-user data
- Well-designed database
  - Facilitates data management
  - Generates accurate and valuable information
- Poorly designed database causes difficult-to-trace errors

Evolution of File System Data Processing

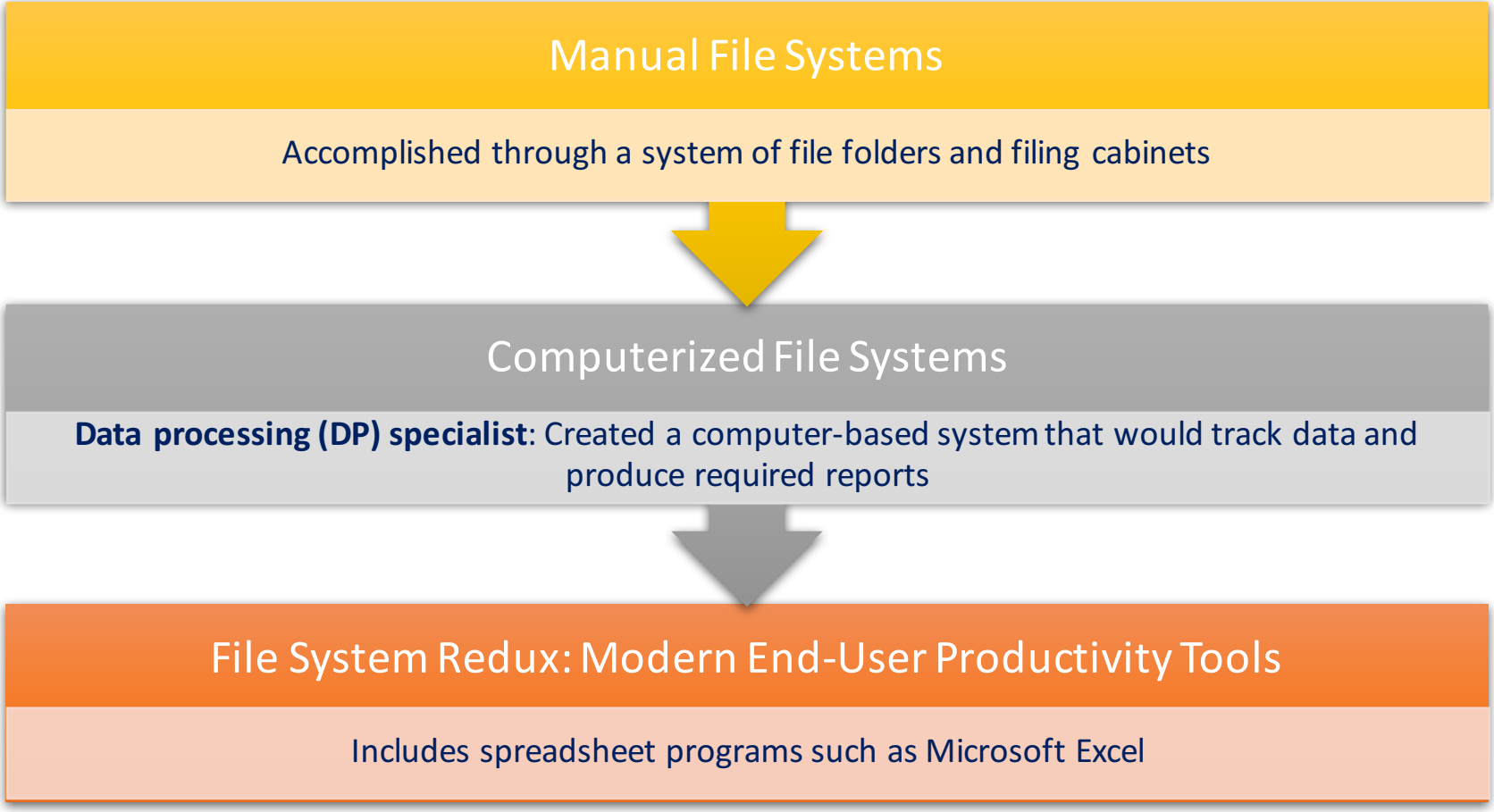


Table 1.2 - Basic File Terminology

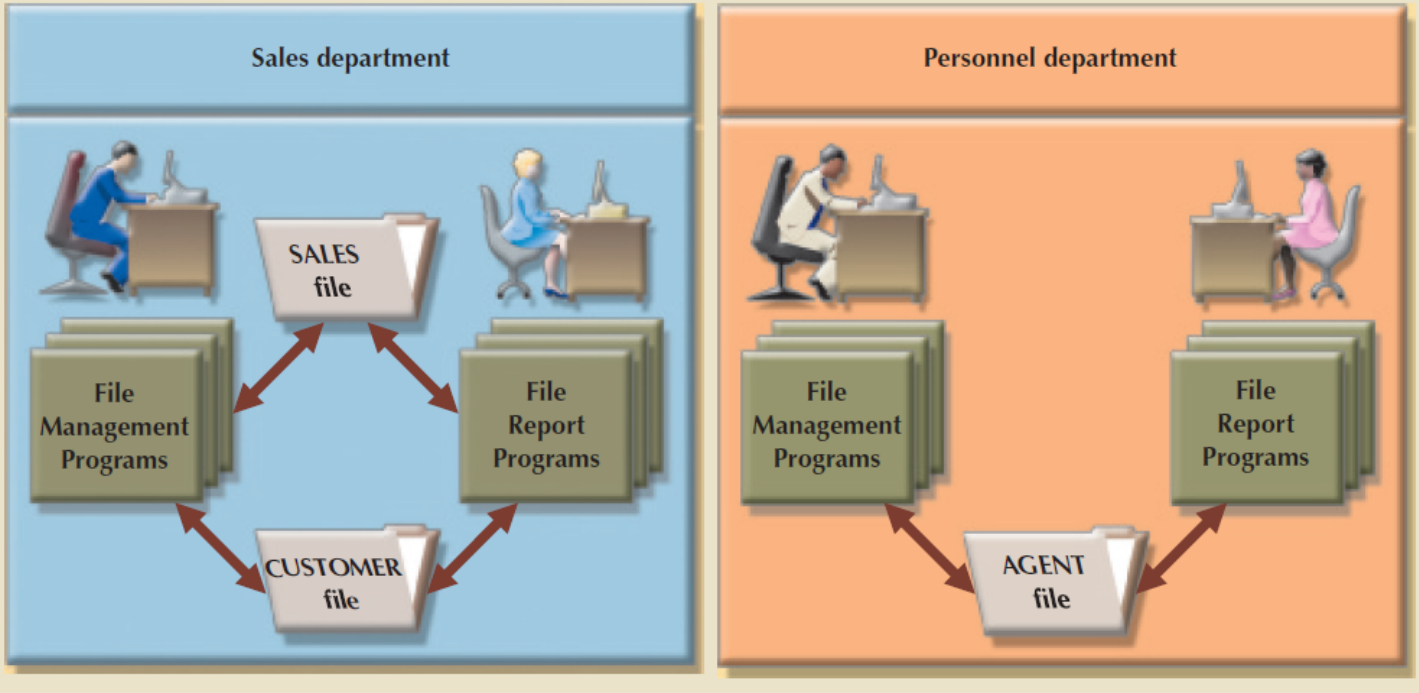
TABLE 1.2

BASIC FILE TERMINOLOGY

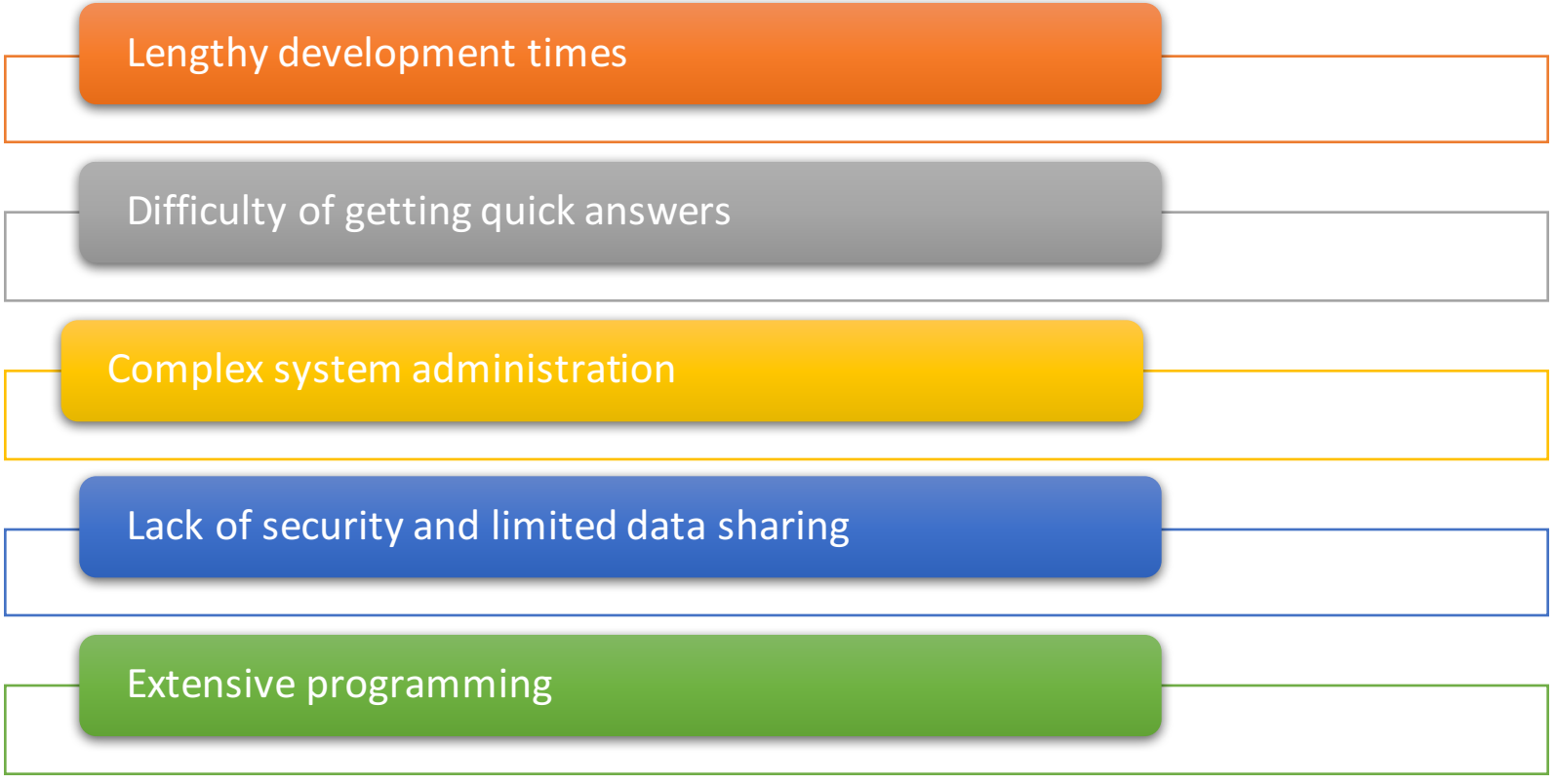
TERM	DEFINITION
Data	Raw facts, such as a telephone number, a birth date, a customer name, and a year-to-date (YTD) sales value. Data have little meaning unless they have been organized in some logical manner.
Field	A character or group of characters (alphabetic or numeric) that has a specific meaning. A field is used to define and store data.
Record	A logically connected set of one or more fields that describes a person, place, or thing. For exam-ple, the fields that constitute a record for a customer might consist of the customer's name, address, phone num-ber, date of birth, credit limit, and unpaid balance.
File	A collection of related records. For example, a file might contain data about the students currently en-rolled at Gigantic University.

Figure 1.8 - A Simple File System

FIGURE 1.8 A SIMPLE FILE SYSTEM



Problems with File System Data Processing



Structural and Data Dependence

- **Structural dependence**: Access to a file is dependent on its own structure
  - All file system programs are modified to conform to a new file structure
- **Structural independence**: File structure is changed without affecting the application's ability to access the data

Structural and Data Dependence

- Data dependence
  - Data access changes when data storage characteristics change
- Data independence
  - Data storage characteristics is changed without affecting the program's ability to access the data
- Practical significance of data dependence is difference between logical and physical format

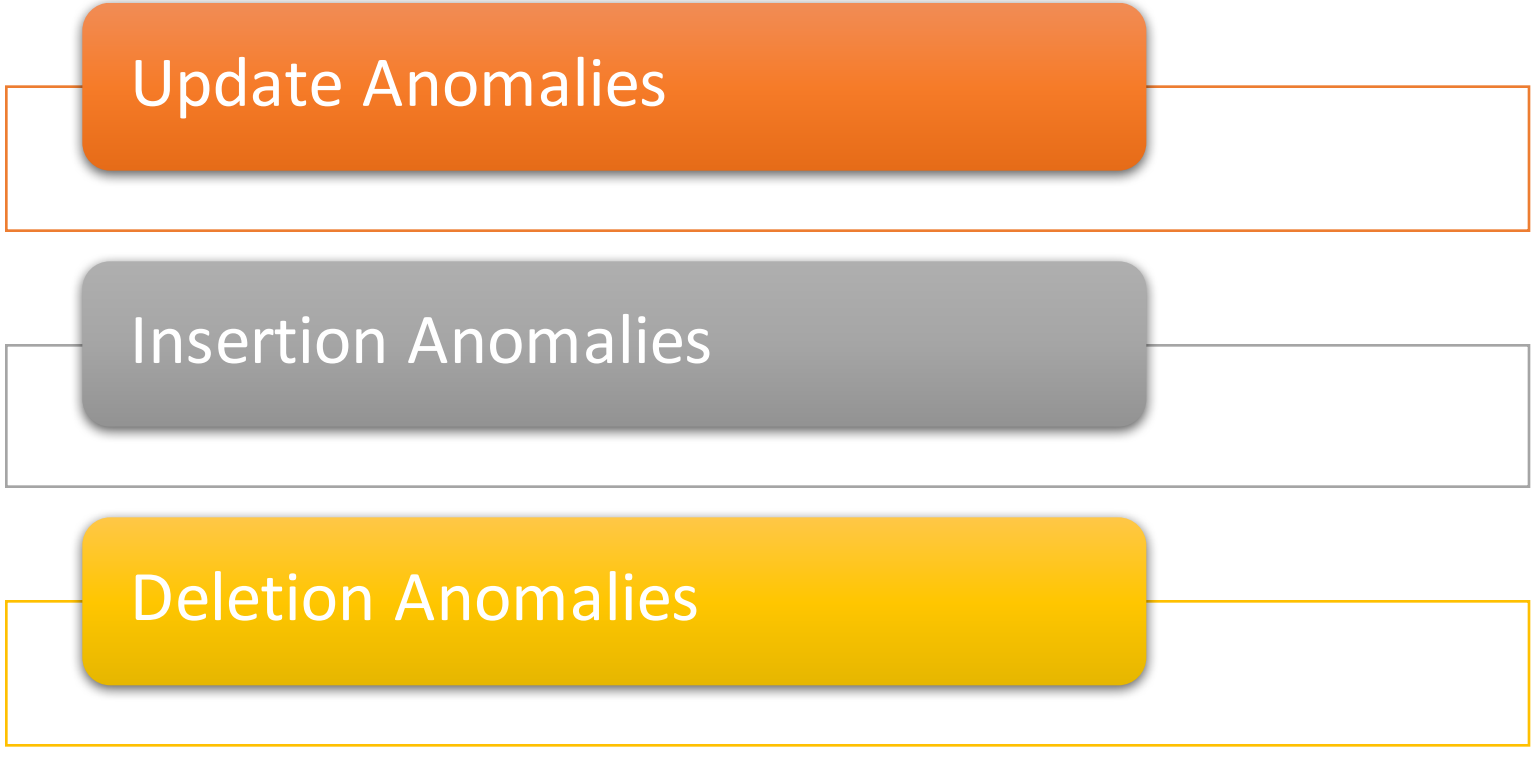
Data Redundancy

- Unnecessarily storing same data at different places
- **Islands of information**: Scattered data locations
  - Increases the probability of having different versions of the same data

Data Redundancy Implications

- Poor data security
- Data inconsistency
- Increased likelihood of data-entry errors when complex entries are made in different files
- **Data anomaly**: Develops when not all of the required changes in the redundant data are made successfully

Types of Data Anomaly



Database Systems

- Logically related data stored in a single logical data repository
  - Physically distributed among multiple storage facilities
- DBMS eliminates most of file system's problems
- Current generation DBMS software:
  - Stores data structures, relationships between structures, and access paths
  - Defines, stores, and manages all access paths and components

Figure 1.9 - Contrasting Database and File Systems

FIGURE 1.9 CONTRASTING DATABASE AND FILE SYSTEMS

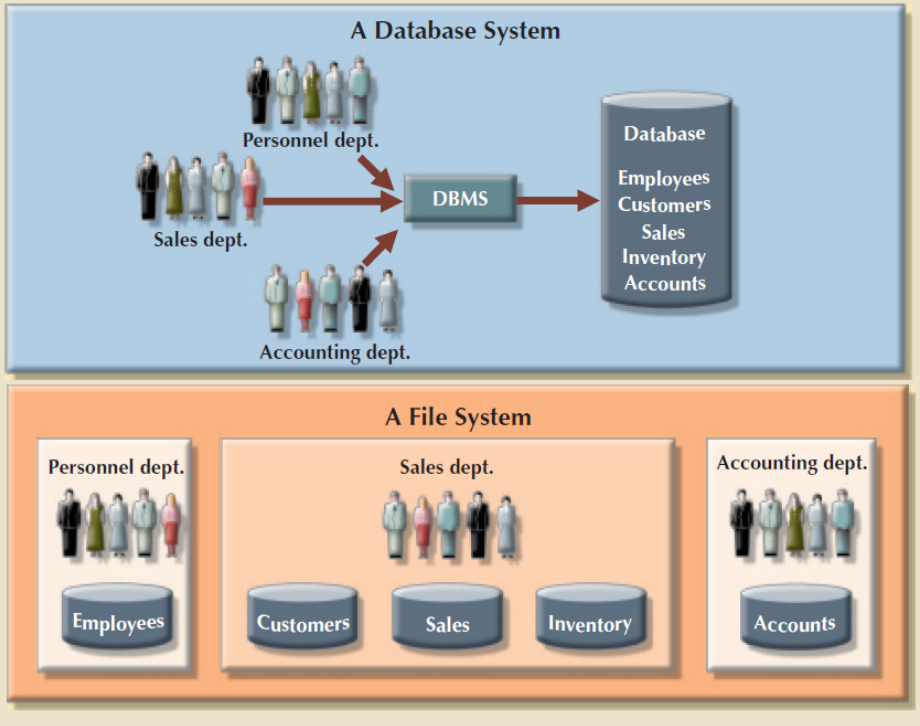
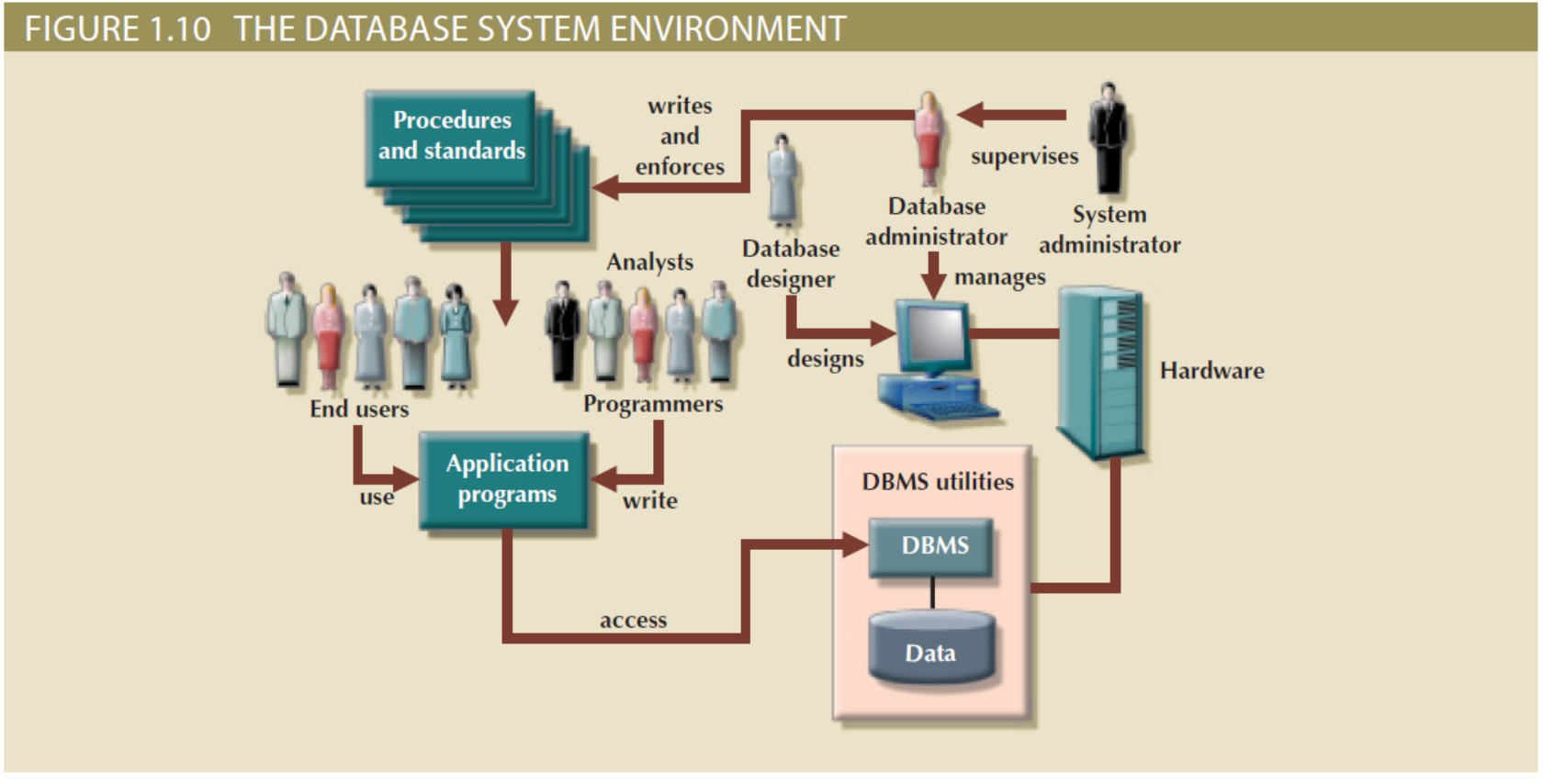


Figure 1.10 - The Database System Environment



DBMS Functions

Data dictionary management

- **Data dictionary**: Stores definitions of the data elements and their relationships

Data storage management

- **Performance tuning**: Ensures efficient performance of the database in terms of storage and access speed

Data transformation and presentation

- Transforms entered data to conform to required data structures

Security management

- Enforces user security and data privacy

DBMS Functions

Multiuser access control

- Sophisticated algorithms ensure that multiple users can access the database concurrently without compromising its integrity

Backup and recovery management

- Enables recovery of the database after a failure

Data integrity management

- Minimizes redundancy and maximizes consistency

DBMS Functions

Database access languages and application programming interfaces

- **Query language**: Lets the user specify what must be done without having to specify how
- **Structured Query Language (SQL)**: De facto query language and data access standard supported by the majority of DBMS vendors

Database communication interfaces

- Accept end-user requests via multiple, different network environments

Disadvantages of Database Systems

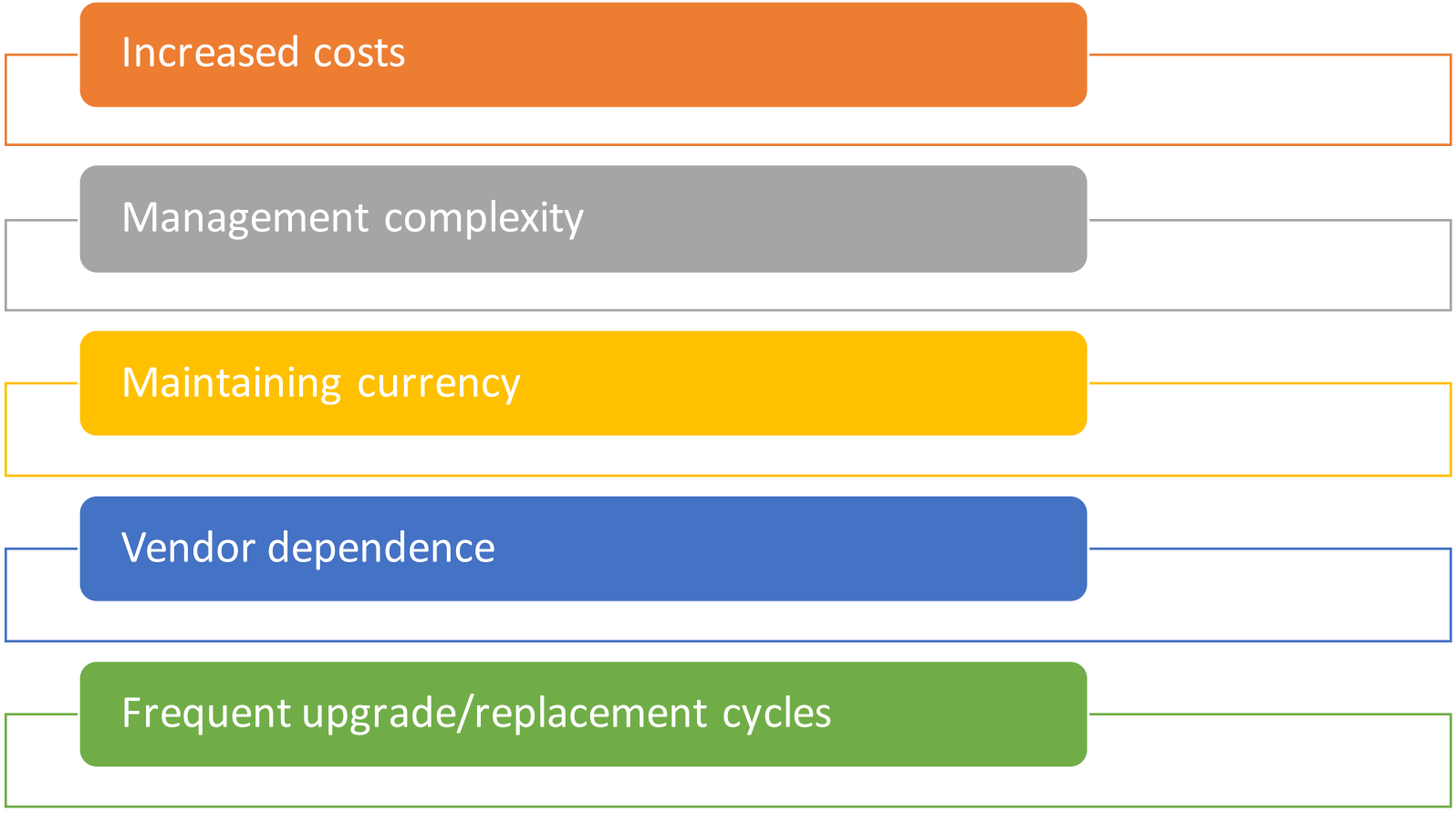


Table 1.3 - Database Career Opportunities

TABLE 1.3

DATABASE CAREER OPPORTUNITIES

JOB TITLE	DESCRIPTION	SAMPLE SKILLS REQUIRED
Database Developer	Create and maintain database-based applications	Programming, database fundamentals, SQL
Database Designer	Design and maintain databases	Systems design, database design, SQL
Database Administrator	Manage and maintain DBMS and databases	Database fundamentals, SQL, vendor courses
Database Analyst	Develop databases for decision support reporting	SQL, query optimization, data warehouses
Database Architect	Design and implementation of database environments (conceptual, logical, and physical)	DBMS fundamentals, data modeling, SQL, hardware knowledge, etc.
Database Consultant	Help companies leverage database technologies to improve business processes and achieve specific goals	Database fundamentals, data modeling, database design, SQL, DBMS, hardware, vendor-specific technologies, etc.
Database Security Officer	Implement security policies for data administration	DBMS fundamentals, database administration, SQL, data security technologies, etc.
Cloud Computing Data Architect	Design and implement the infrastructure for next-generation cloud database systems	Internet technologies, cloud storage technologies, data security, performance tuning, large databases, etc.