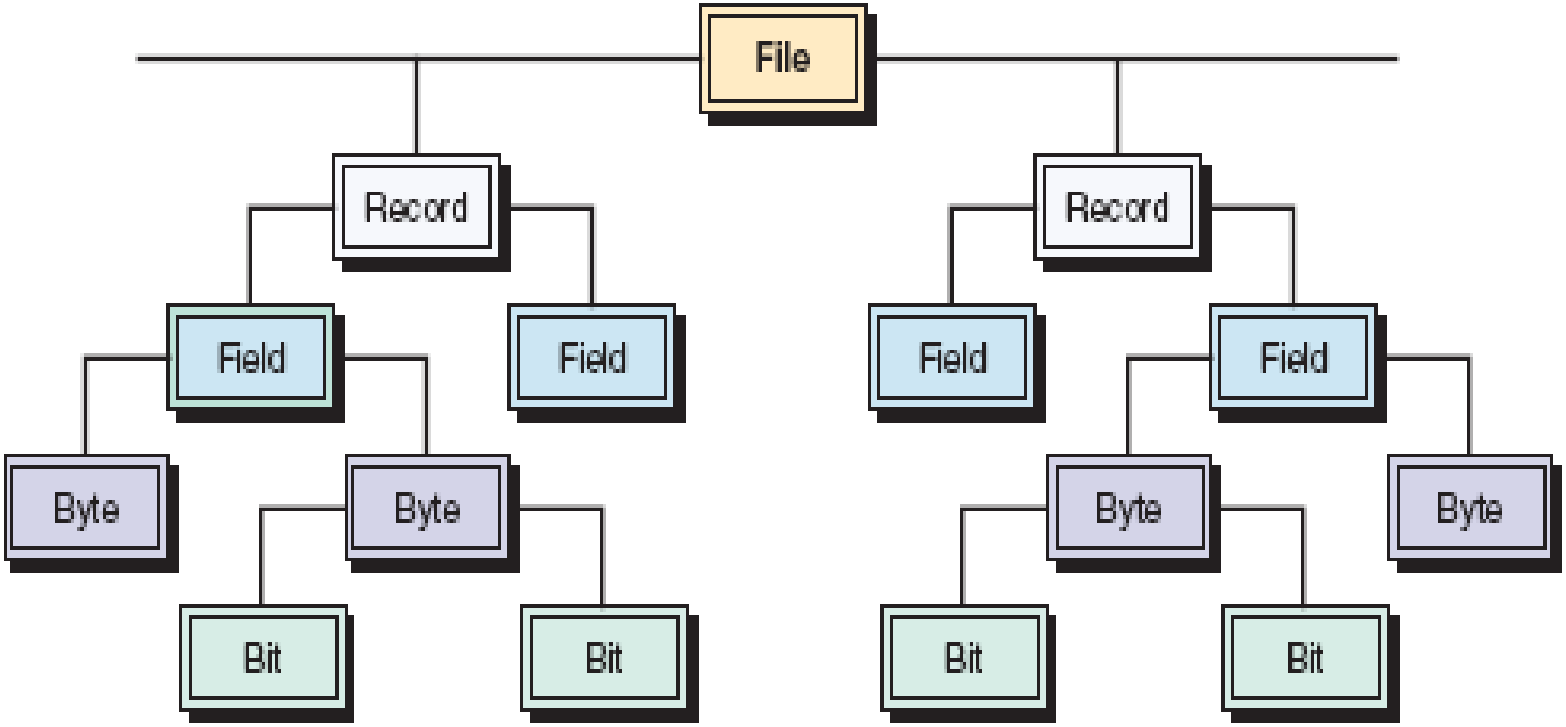
Technology Guide 3

Data and Databases

Information Technology For Management 4th Edition
Turban, McLean, Wetherbe
Lecture Slides by A. Lekacos,
Stony Brook University
John Wiley & Sons, Inc.

File Management

A computer system organizes data in a hierarchy that begins with bits, and proceeds to bytes, fields, records, files, and databases.



File Management Continued

- A bit represents the smallest unit of data a computer can process (i.e., a 0 or a 1).
- A group of eight bits, called a byte, represents a single character, which can be a letter, a number, or a symbol.
- A logical grouping of characters into a word, a group of words, or a complete number is called a field.
- A logical group of related fields, comprise a record.
- A logical group of related records is called a file.
- A logical group of related files would constitute a database.

The amount of data the average business collects and stores is doubling each year. Businesses collect data from multiple sources, including customer-relationship management and enterprise resource planning applications, online systems and suppliers & business partners.

File Management Continued

- Another way of thinking about database components is that a record describes an entity. An entity is a person, place, thing, or event on which we maintain data.
- Each characteristic or quality describing a particular entity is called an attribute (*corresponds to a field on a record*).
- Every record in a file should contain at least one field that uniquely identifies that record so that the record can be retrieved, updated, and sorted. This identifier field is called the primary key.
- Secondary keys are other fields that have some identifying information, but typically do not identify the file with complete accuracy.

File Management - Accessing Records

Records can be arranged in several ways on a storage medium, and the arrangement determines the manner in which individual records can be accessed

- Sequential file organization data records must be retrieved in the same physical sequence in which they are stored.
- Direct or random file organization, users can retrieve records in any sequence, without regard to actual physical order on the storage medium.
  - Indexed sequential access method (ISAM) uses an index of key fields to locate individual records.
  - Direct file access method uses the key field to locate the physical address of a record. This process employs a mathematical formula called a transform algorithm to translate the key field directly into the record's storage location on disk.

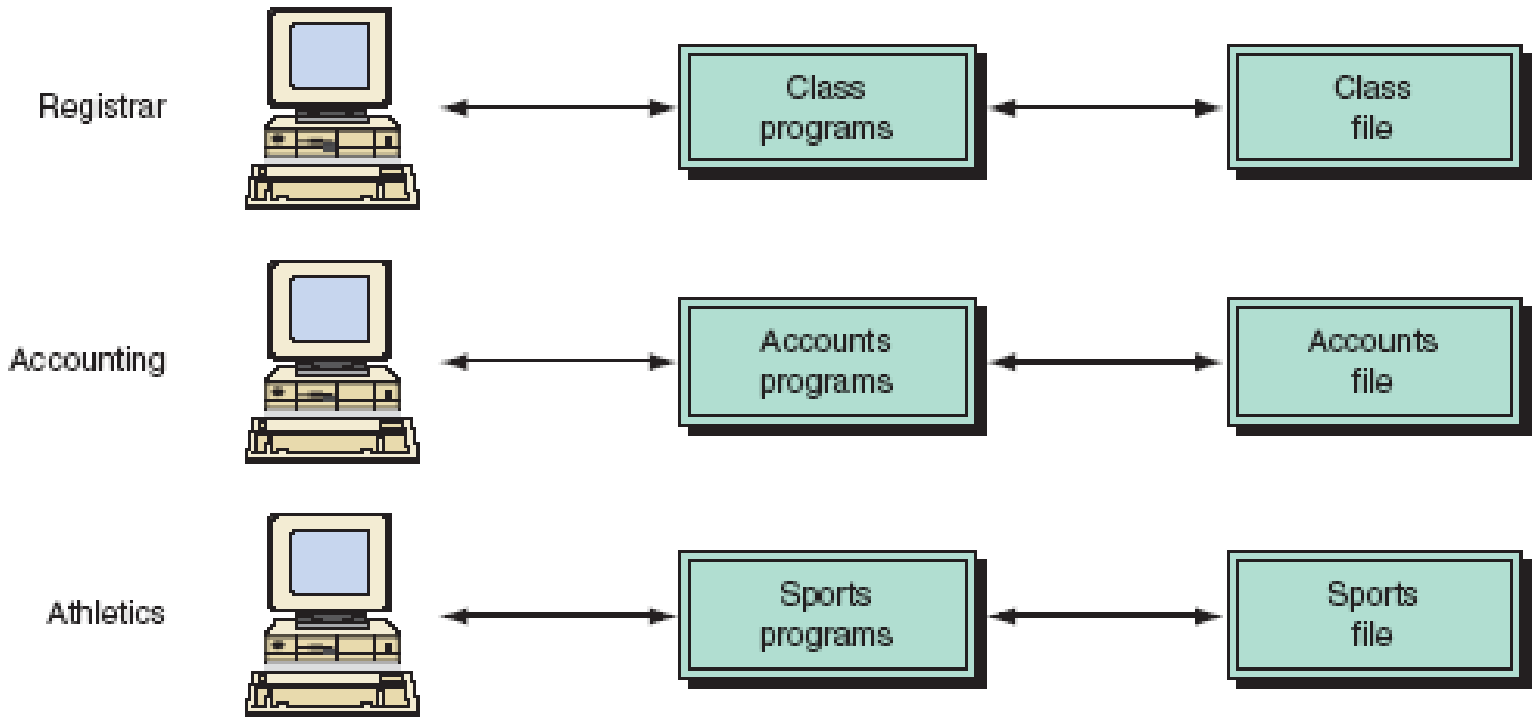
Problems in the File Environment

Organizations typically began automating one application at a time. These systems grew independently, without overall planning. Requiring its own data organized into unique data files.

- Without proper systems management other problems arose:
  - Data redundancy: as applications and their data files were created by different programmers over a period of time, the same data could be duplicated in several files.
  - Data inconsistency exist across various copies (the actual values in each file no longer agree).
  - Data isolation. Refers to the difficulty in accessing data from different applications.
  - Data integrity problems propagate more easily across multiple data files.

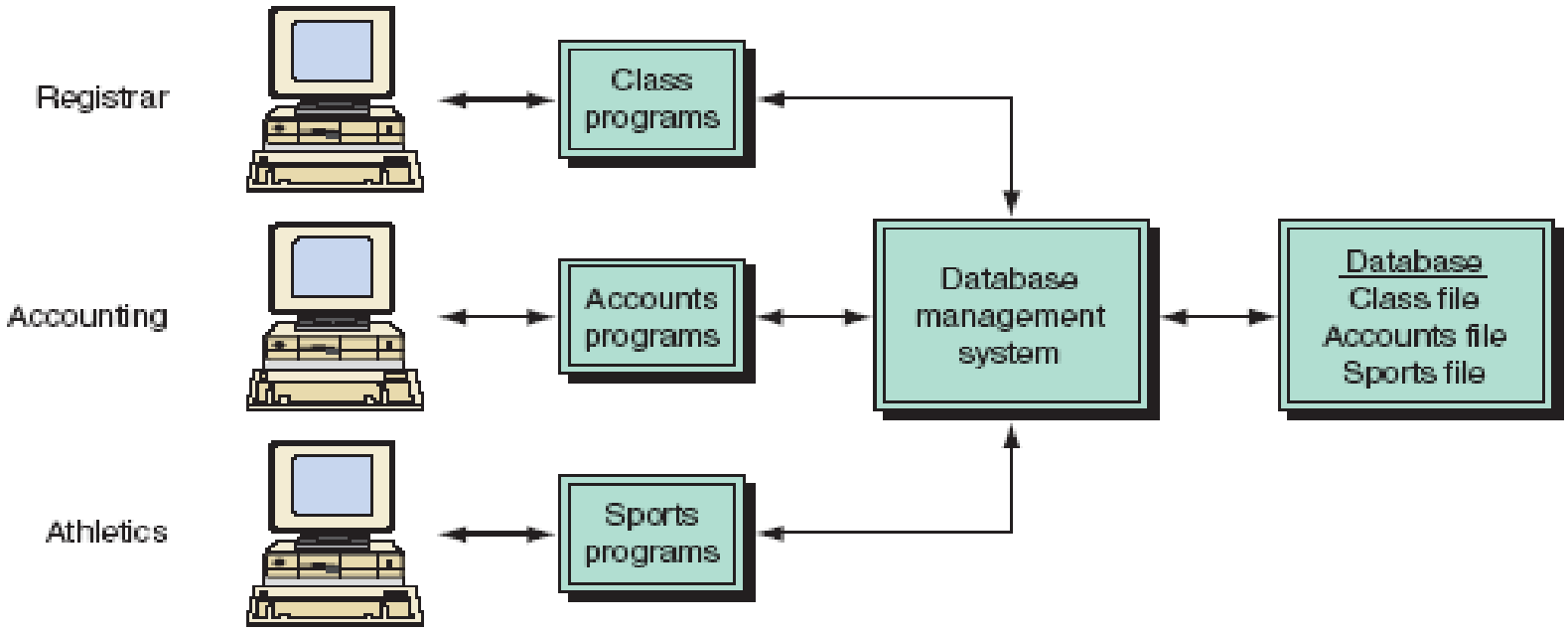
Problems in the File Environment

Storing data in data files that are tightly linked to their applications eventually led to organizations having hundreds of applications and data files, with no one knowing what the applications did or what data they required. There was no central listing of data files , data elements or definitions of the data.



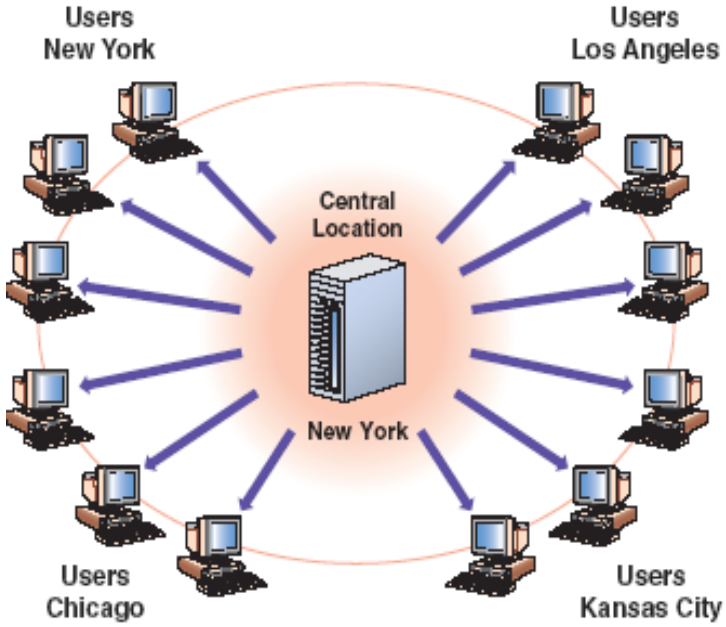
Databases

A database is an organized logical grouping of related files. In a database, data are integrated and related so that one set of software programs provides access to all the data, minimizing the problems associated with data file environments (data redundancy, data isolation, data inconsistency and data sharing).



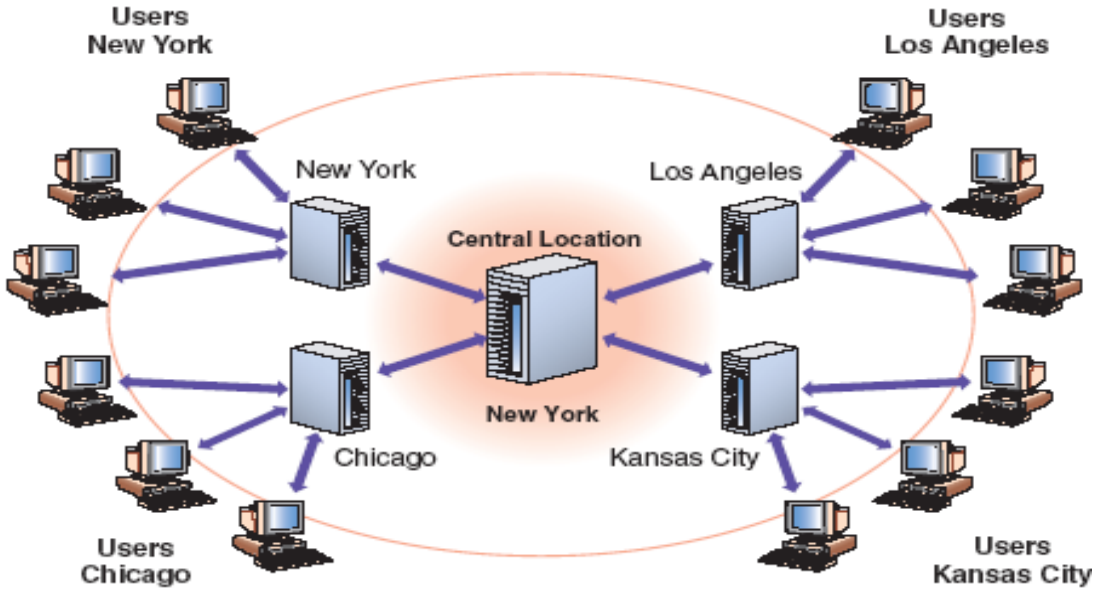
Databases - Centralized

A centralized database has all the related files in one physical location. Centralized database files on large, mainframe computers were the main database platform for decades, primarily because of the enormous capital and operating costs of other alternatives. Not only do centralized databases save the expenses associated with multiple computers, but they also provide database administrators with the ability to work on a database as a whole at one location.



Databases - Distributed

A distributed database has complete copies of a database, or portions of a database, in more than one location. There are two types of distributed databases: A replicated database has complete copies of the entire database in many locations, primarily to alleviate the single-point-of-failure problems of a centralized database as well as to increase user access responsiveness. A partitioned database is subdivided, so that each location has a portion of the entire database thus enhancing local response.



Database Management System (DBMS)

The program (or group of programs) that provides access to a database is known as a database management system (DBMS). The DBMS acts as an interface between application programs and physical data files while providing users with tools to add, delete, maintain, display, print, search, select, sort, and update data.

TABLE T-3.1 Advantages and Capabilities of a DBMS

- Access and availability of information can be increased.
- Data access, utilization, security, and manipulation can be simplified.
- Data inconsistency and redundancy is reduced.
- Program development and maintenance costs can be dramatically reduced.
- Captures/extracts data for inclusion in databases.
- Quickly updates (adds, deletes, edits, changes) data records and files.
- Interrelates data from different sources.
- Quickly retrieves data from a database for queries and reports.
- Provides comprehensive data security (protection from unauthorized access, recovery capabilities, etc.).
- Handles personal and unofficial data so that users can experiment with alternative solutions based on their own judgment.
- Performs complex retrieval and data manipulation tasks based on queries.
- Tracks usage of data.
- Flexibility of information systems can be improved by allowing rapid and inexpensive ad hoc queries of very large pools of information.
- Application-data dependence can be reduced by separating the logical view of data from its physical structure and location.

DBMS Languages

A DBMS contains four major components: the data model, the data definition language, the data manipulation language, and the data dictionary.

- The data model defines the way data are conceptually structured.
- The data definition language (DDL) is the language used by programmers to specify the types of information and structure of the database. The schema is the logical description of the entire database and the listing of all the data items and the relationships among them. A subschema is the specific set of data from the database that is required by each application.
- Data manipulation language (DML) is used with a 3rd or 4th generation languages to manipulate the data in the database. Structured query language (SQL) is the most popular relational database language, combining both DML and DDL features.
- The data dictionary stores definitions of data elements and data characteristics such as usage, physical representation, ownership, authorization, and security. A data element represents a field.

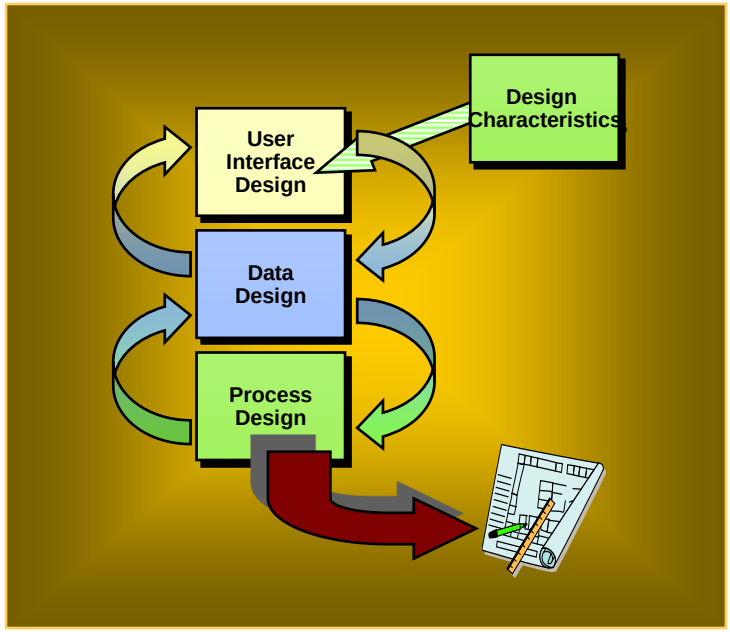
DBMS Benefits

Database management systems provide many advantages to the organization:

- Improved strategic use of corporate data
- Reduced complexity of the organization's information systems environment
- Reduced data redundancy and inconsistency
- Enhanced data integrity
- Application-data independence
- Improved security
- Reduced application development and maintenance costs
- Improved flexibility of information systems
- Increased access and availability of data and information

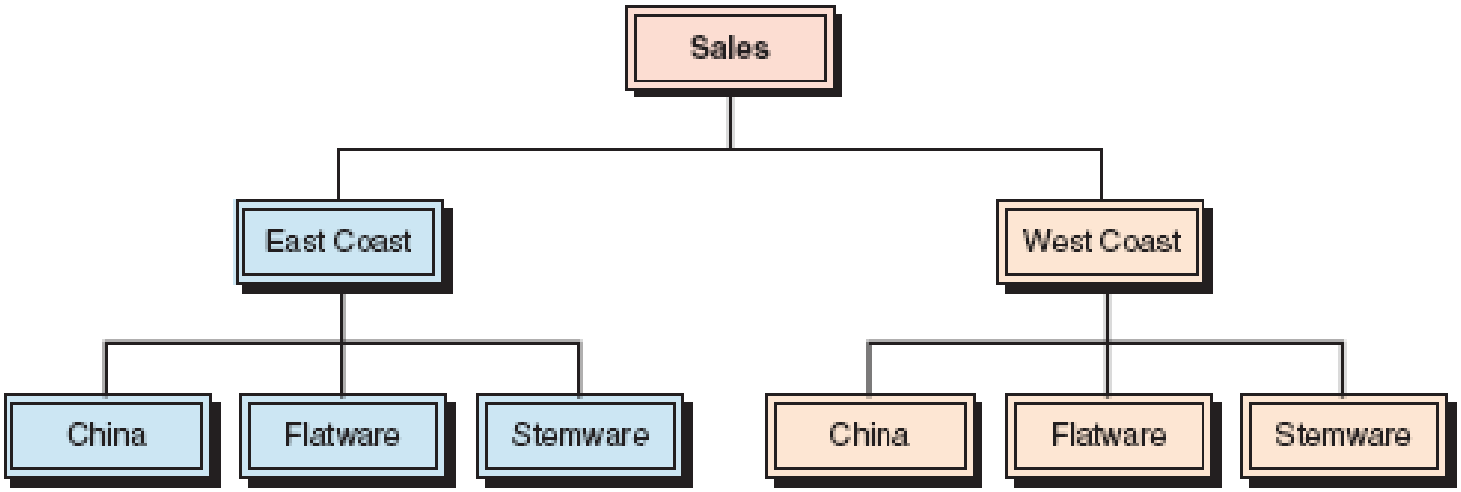
Data Organization

There are many ways to structure the data organizations need. The three basic models for logically structuring databases are: hierarchical, network, and relational. Four additional models are emerging: multidimensional, object-oriented, small-footprint, and hypermedia. Using these various models, database designers can build logical or conceptual views of data that can then be physically implemented into virtually any database.



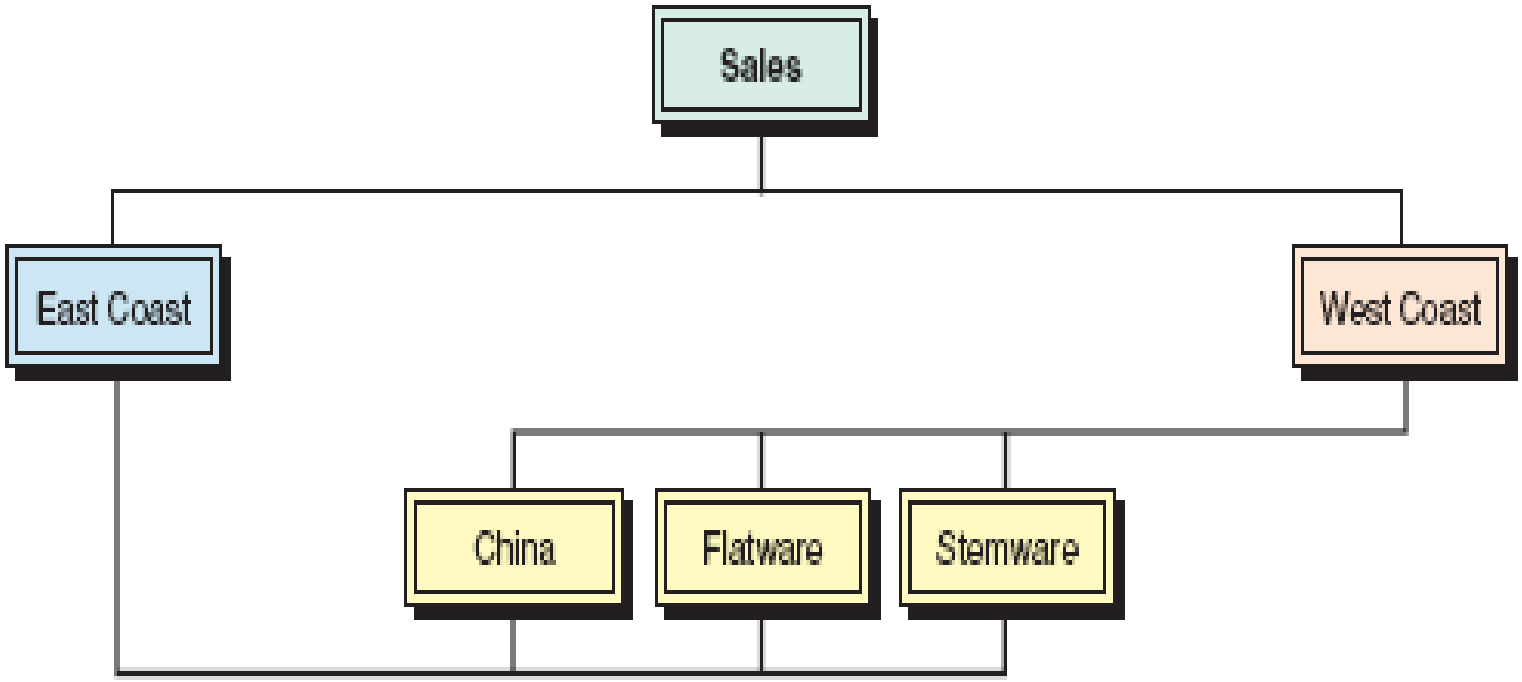
Data Organization Hierarchical structure

The hierarchical structure was developed because hierarchical relationships are commonly found in traditional business organizations and processes. This mode relates data by rigidly structuring data into an inverted "tree" in which records contain two elements: A master field, often called a key, which identifies the ordering of the records and A variable number of subordinate fields that defines the rest of the data within a record.



Data Organization Network structure

The network database model creates relationships among data through a linked-list structure in which subordinated records (called *members*) can be linked to more than one parent.



Data Organization Relational structure

Most business data, especially accounting and financial data, have traditionally been organized into tables of columns and rows. The relational database model is based on this simple concept of tables in order to capitalize on characteristics of rows and columns of data, which is consistent with real-world business situations.

Division

Code	*Name*
01	Stemware

Title

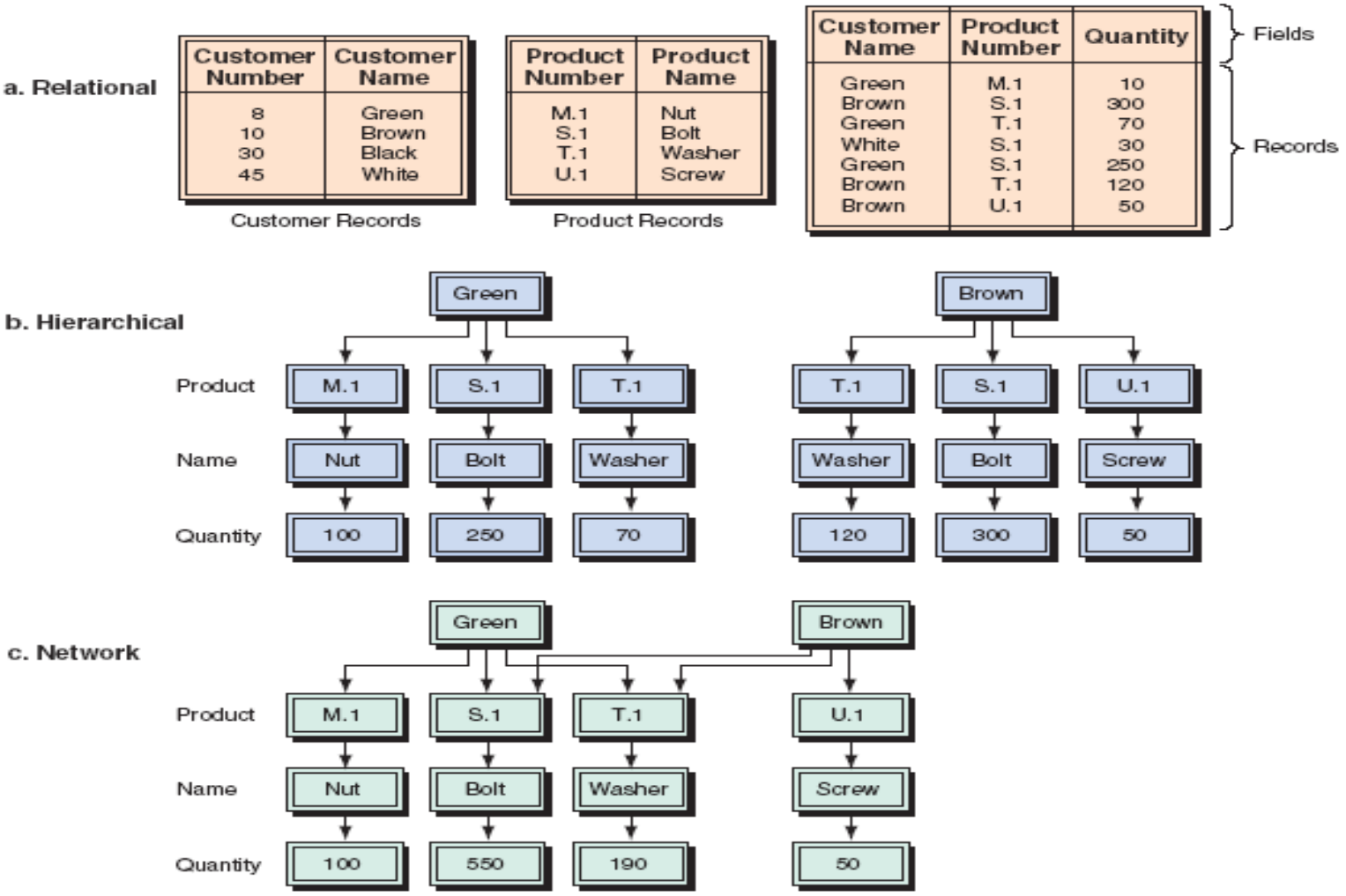
Code	*Description*
01	Director

Employee

Name	*Title Code*	*Division Code*	*Age*
Smith, A.	01	01	42

In a relational database, the tables are called **relations,** and the model is based on the mathematical theory of sets and relations. In this model, each row of data is equivalent to a *record,* and each column of data is equivalent to a *field*. In the relational model terminology, a row is called a **tuple,** and a column is called an **attribute.**

Data Organization Continued



Data Organization Continued

TABLE T-3.3 Advantages and Disadvantages of Logical Data Models

Model	Advantages	Disadvantages
Hierarchical database	Searching is fast and efficient.	Access to data is predefined by exclusively hierarchical relationships, predetermined by administrator. Limited search/query flexibility. Not all data are naturally hierarchical.
Network database	Many more relationships can be defined. There is greater speed and efficiency than with relational database models.	This is the most complicated model to design, implement, and maintain. Greater query flexibility than with hierarchical model, but less than with relational mode.
Relational database	Conceptual simplicity; there are no predefined relationships among data. High flexibility in ad-hoc querying. New data and records can be added easily.	Processing efficiency and speed are lower. Data redundancy is common, requiring additional maintenance.

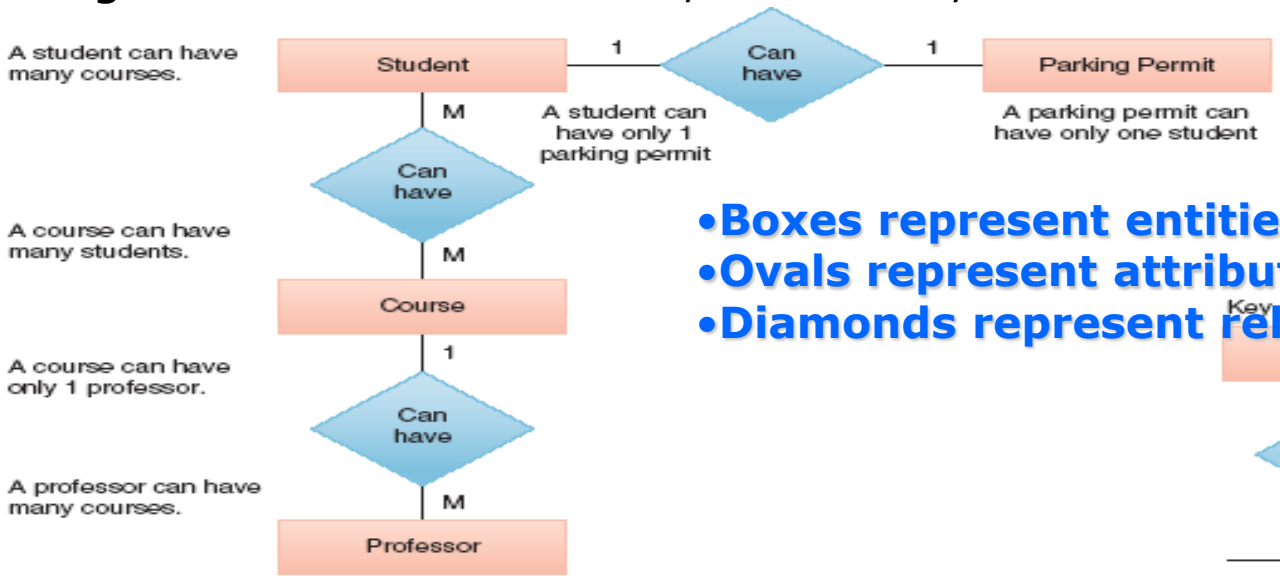
Creating Databases

To create a database, designers must develop a conceptual design and a physical design. The conceptual design of a database is an abstract model of the database from the user or business perspective. The physical design shows how the database is actually arranged on direct access storage devices. To produce optimal database design, entity-relationship modeling and normalization are employed.

- The design process identifies relationships among data elements
- The most efficient way of grouping data elements together to meet information requirements.
- It then identifies redundant data elements
- Then the groupings of data elements for specific applications.
- This process is continued until an overall logical view of the relationships among all of the data elements in the database appears.

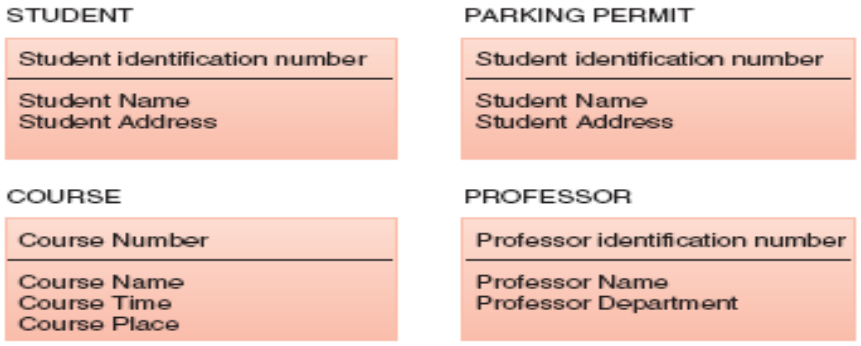
Creating Databases E-R Diagrams

ER diagrams consist of entities, attributes, and relationships.



- **Boxes represent entities,**
- **Ovals represent attributes**
- **Diamonds represent relationships**

(a)



Creating Databases E-R Diagrams

Entities are associated with one another in relationships, which can include many entities. The number of entities in a relationship is the degree of the relationship. Relationships of degree 2 are common and are called binary relationships.

- There are three types of binary relationships.
  - In a 1:1 (one-to-one) relationship, a single-entity instance of one type is related to a single-entity instance of another type.
  - The second type of relationship, 1:M (one-to-many) For example a professor can have many courses, but each course can have only one professor.
  - The third type of relationship, M:M (many-to-many), for example a student can have many courses, and a course can have many students.

Creating Databases Normalization

In order to use a relational database model effectively, complex groupings of data must be streamlined to eliminate redundant data elements and awkward many-to-many relationships. The process of creating small, stable data structures from complex groups of data is called normalization.

- Eliminate redundancy caused by fields repeated within a file or record
- Eliminate attributes that do not directly describe the entity
- Eliminate fields that can be derived from other fields .
- Avoid update anomalies (i.e., errors from inserting, deleting, and modifying records).
- Represent accurately the item being modeled.
- Simplify maintenance and information retrieval.

Emerging Database Models

Many of today's applications require database capabilities that can store, retrieve, and process diverse media and not just text and numbers. Full-motion video, voice, photos, and drawings cannot be handled effectively or efficiently by either hierarchical, network, or relational databases.

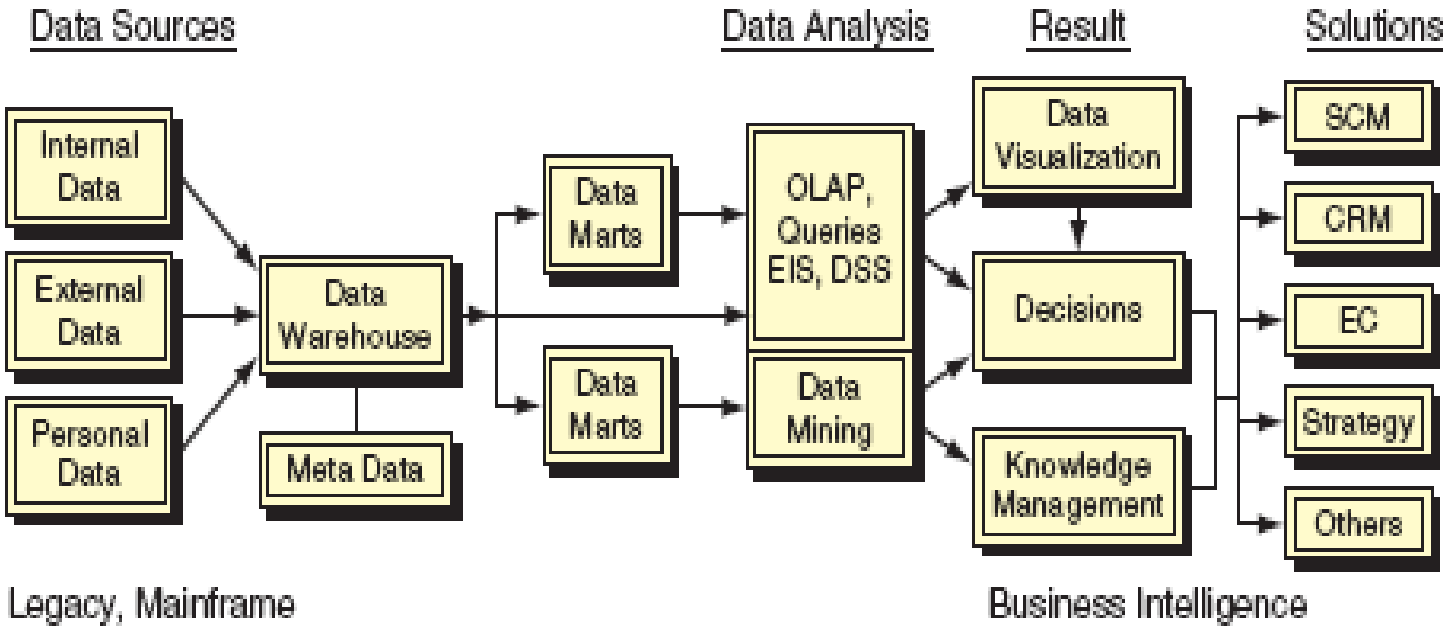
- Multidimensional database. This database enables end users to quickly retrieve and present complex data that involve many dimensions.
- Deductive databases support knowledge-based applications that require deductive reasoning for searches.
- Object-oriented databases. In order to work in an object-oriented environment, it is necessary to use OO programming and OO databases.

Emerging Database Models continued

- Multimedia and hypermedia databases. These are analogous to contemporary databases for textual and numeric data; however, they have been tailored to meet the special requirements of dealing with different types of media.
- Small-footprint databases enable organizations to put certain types of data in the field where workers with portable machines can access information. These databases have replication mechanisms that take into account the occasionally connected nature of laptops and handhelds.
- Hypermedia database model stores chunks of information in the form of nodes connected by links established by the user. The nodes can contain text, graphics, sound, full-motion video, or executable computer programs.

Data Warehouses

A data warehouse is an additional database that is designed to support DSSs, EISs, online analytical processing (OLAP), and other end-user activities, such as report generation, queries, & graphical presentation. A data mart is smaller, less expensive, and more focused than a large-scale data warehouse. Data marts can be a substitution for a data warehouse or they can be used in addition to it.



IP-Based Storage - SANS and NAS

Storage connected to servers over IP (*Internet protocol*) networks, also known as IP storage, enables servers to connect to SCSI storage devices and treat them as if they were directly attached to the server, regardless of the location. IP storage is a transport mechanism that seeks to solve the problem of sending storage data over a regular network in the block format it prefers rather than the file format generally used.

TABLE T-3.4 Pros and Cons of SANs and NAS

	Pros	Cons
SANs	Off-loads storage traffic from existing network	Expensive—requires new sub-network
	Flexible design improves reliability	Manages data in blocks, not files, so it requires specialized software
	Equipment is designed to be highly scalable	Requires fiber channel networking skills
NAS	Uses existing network infrastructure	Slower—network protocols are not streamlined for storage
	Manages data as files	Loads already burdened network with storage data, including backup
	Easy to install and use	Doesn't scale up easily

Data Storage Infrastructure

Direct Access File System (DAFS) protocol is one of the important technologies in data center storage infrastructure that will enable databases, Web servers, e-mail backends, and a host of other server-resident applications to achieve performance levels that are simply unattainable in the pre-DAFS world.

Storage resource management (SRM) and storage virtualization are pieces of software that help manage storage as a *whole entity* rather than the disparate bits of technology. It works much like network management devices on corporate networks.

Technology Guide 3