Databases and DBMS

*What is a database?*

*Why are databases important?*

The program you are enrolled in is called *Computer Science*, but the industry that you are training for is *Information Technology*. Most of the courses you have taken so far have focused on the “technology” part of IT. In this course, the focus shifts to “Information” – what is it, what can we & do we do with it, how do we organize it, how do we use it to improve efficiency and productivity in a business. We will look at data analysis, database design, data administration and database administration in this and the database management II course. We will also learn how to use Structured Query Language (SQL) to store, retrieve and manipulate data with the Oracle database.

Think about the systems you worked on last year in your programming courses – the video store and music store systems. What sort of information did those programs produce? What information did they need in order to produce it? Think about the importance of information to those systems and how important easy and quick access to the information is. Think about the program design – IPO diagrams – Input refers to the data used to solve a problem and Output refers to the information produced as a solution to the problem. Think about class design – what are the components of a class (data & processes). Again notice the importance of the data. Data and information are the cornerstones of all modern computer systems. For this reason, the database courses are among the most important courses you will take over your three years in this program.

For the programs that you wrote in Java, you used traditional sequential files to store your data. In today’s business world, databases are the most common method of storing large amounts of data.

***Databases*** are used to ***store***, manipulate and retrieve data in nearly every type of organization including business, health care, education, government and libraries.

Database technology is routinely used by

1. individuals on personal computers,
2. workgroups accessing databases on network servers and
3. all employees using enterprise-wide distributed applications
4. everybody using a ***web application***.

We are going to begin this course with some basic concepts and definitions and then proceed to look at traditional file processing systems and some of their shortcomings that led to the development of database systems. We’ll look at the range of database applications and at the benefits, costs and risks of using database systems.

### What is a database?

An organized collection of logically related data. E.g. A salesperson may maintain a small database of customer contacts on her laptop computer that consists of a few megabytes of data. A large corporation may build a very large database consisting of several terabytes of data (a *terabyte* is a trillion bytes) on a large mainframe computer that is used for decision support applications.

### What do we mean by data?

Traditionally, *data* referred to known facts that could be recorded and stored on computer media. For example, in a salesperson’s database, the data would include facts such as customer name, address and telephone number. This definition now needs to be expanded to reflect a new reality. Databases today are used to store objects such as documents, photographic images, sound and even video segments in addition to conventional textual and numeric data. For example, the salesperson’s database might include a photo image of the customer contact. It might also include a sound recording or video clip of the most recent conversation with the customer. To reflect this reality, we use the following broadened definition:

***Data*** consists of facts, text, graphics, images, sound and video segments that have meaning in the users’ environment.

### What do we mean by organized?

* the data are structured so as to be easily stored, manipulated and retrieved by users.

### What do we mean by related?

* Data describe a domain of interest to a group of users and that the users can use the data to answer questions concerning that domain.
* For example a database for an automobile repair shop contains data identifying customers, automobiles belonging to those customers and repair histories for each of those automobiles.

### Implementing Relational Databases

* Local Area Networks**:** to facilitate sharing data and information, Local Area Networks (LAN) are used to implement databases (see fig. 1-5, p. 13 HPM)
* Intranet**:** to improve intracompany communication and decision making, an Intranet that uses Web-based Internet technology and is only accessible within the company is used
* Internet**:** to allow access to the database by customers or vendors, the Internet is used
* Data Warehouse**:** a separate database that contains historical and summarized information
* OLAP (On-line analytical processing):specialized decision support tools to query and analyze the database

### A Database Application

* An application program (or set of related programs) that is used to perform a series of activities on behalf of database users
* performs some combination of the following basic operations:

1. Create Add new data to the database
2. ReadRead current database data (often presented in a useful format on a computer screen or a printed report)
3. UpdateUpdate (or modify) current database data
4. DeleteDelete current data from the database

* Example: create class list from College database:

***Class List***

***Course Number*** *200-H01-HR* ***Course Name*** *Science I*

***Semester*** *F01* ***Teacher*** *PollyEsther Doublenitz*

***Student Number Name***

000-1112 Mahatma Coate

000-1113 Sandy Beech

# Advantages of the Database Approach

1. Program-Data Independence

Data independence is the property that the structure of a database can change without the programs that access the database having to change. With the database approach, data descriptions (i.e. file formats) are stored with the data itself in a ***repository*** in the database. This property of database systems allows an organization’s data to change and evolve (within limits) without changing the application programs that process the data.

1. Planned Data Redundancy

The design goal with the database approach is that previously separate (and redundant) data files are integrated into a single, logical structure. Each primary fact is recorded (ideally) in only one place in the database. For example, if the College systems were converted to a centralized database system, the student address would only be stored in one location. The database approach does not eliminate redundancy entirely, but it enables the designer to control the type and amount of redundancy. For example, the program code is repeated in the Student table to establish the relationship between a student and the program he/she is registered in.

1. Improved Data Consistency

If data is stored in only one location, there is no opportunity for inconsistency. Also updating data values is greatly simplified when each value is stored in one place only.

1. Improved Data Sharing

A database is designed as a shared corporate resource. Authorized users are granted permission to use the database, and each user (or group of users) is provided one or more user views to facilitate this use. A ***user view*** is a logical description of some portion of the database that is required by a user to perform some task.

If a new application is needed, it can use the existing data in the database - a new file is not necessary.

1. Increased Productivity of Application Development

Database applications can be much more rapidly developed than conventional file applications because:

1. The programmer can concentrate on the specific functions required for the new application, without having to worry about file design or low-level implementation details. (i.e. The programmer does not have to worry about file management programs.)

2. The database management system provides a number of high-level productivity tools such as forms and report generators and high-level languages that automate some of the activities of database design and implementation.

3. Significant improvement in programmer productivity, sometimes estimated to be as high a 60 percent is currently being realized through the use of Web services (HTTP/S), based on the use of standard Internet protocols and a universally accepted data format (XML).

1. Enforcement of Standards
   * one individual or group should be responsible for maintaining the database structure
   * this group is usually called the Database Administration group or the Database Administrator in the case of an individual
   * the DBA balances the conflicting demands of different groups within the organization and therefore provides a database that is useful to the whole company
   * The database administration function enforces single-point authority and responsibility for establishing and enforcing data standards including naming conventions, data quality standards and uniform procedures for accessing, update, and protecting data.
2. Improved Data Quality

A recent report by The Data Warehousing Institute (TDWI) estimated that data quality problems currently cost U.S. businesses some $600 billion each year.

Two ways of improving data quality are:

1. Specifying integrity constraints that are enforced by the DBMS. A constraint is a rule that cannot be violated by dataset users.

*Examples*:

* a rule that a Canadian postal code must be 6 characters long beginning with a letter and alternating letters and digits for the next 5 characters.
* a rule that an employee who is allocated a PC must exist in the database

2. Cleaning up or "scrubbing" operational data before they are placed in the data warehouse.

*Example*:

* Do you ever receive multiple copies of a catalogue? The company that sends you three copies of each of its mailing pieces could realize significant postage and printing savings if its data were scrubbed.

1. Improved Data Accessibility and Responsiveness

End-users without any programming experience or ability can directly retrieve and display data using SQL and QBE.

1. Reduced Program Maintenance

Because of data independence and reduced redundancy, program maintenance becomes much simpler. Changes to the data format need only be done in the database once for each field.

1. Increased Security

Modern databases can restrict access to individual attributes of entities thereby ensuring that only authorized individuals have access to the corporation's data.

## Costs and Risks of Database Processing

1. New Specialized Personnel

* DBMS's are very complex programs that require many choices to be made in designing the database – incorrect decisions can impact greatly on use, flexibility and speed of the database
* Require individuals to design and implement, provide database administrations services and manage a staff of new people
* Because of rapid changes in technology retraining must be provided on a regular basis

1. Installation and Management Cost and Complexity

* A DBMS is a very large and complex program that has a high initial cost, requires a staff of trained personnel to install and operate, requires megabytes of disk space to store and lots of memory to run and has substantial annual maintenance and support costs.
* may require upgrades to the hardware and data communications systems
* substantial training is normally required on an ongoing basis to keep up with new releases and upgrades

1. Conversion Costs

* older legacy systems which are based on file processing and/or older database technology can be prohibitively expensive to convert to modern database technology – in terms of dollars, time and organizational commitment.
* data warehouses are one way to continue to use older systems while at the same time exploiting modern database technology and techniques.

1. Need for Explicit Backup and Recovery

* Because all the users are now using the same file (database), failure of the database creates problems for all users, not just those using the particular data affected
* The process of recovering from a failure is much more difficult because of the complexity of the database and because of the number of users accessing the data
* A shared corporate database must be accurate and available at all times

1. Organizational Conflict

* a shared database requires a consensus on data definition and ownership as well as responsibilities for accurate data maintenance
* handling conflict issues over these decisions requires organizational commitment to the database approach, organizationally astute database administrators and a sound evolutionary approach to database development.

## Conventional Files versus Databases: Pros and Cons

|  |  |
| --- | --- |
| **Conventional Files** | **Databases** |
| Easy to design and implement (based on a single application) | Complex to design - need to adhere to rigorous design principles. Need DBMS to provide interface between programmer and data |
| Fast Processing Speed | Slow - additional processing overhead (often requires purchase of larger computer) |
| Redundancy - duplicate input, duplicate storage | Little or no redundancy |
| Integrity problems due to redundant data | Integrity problems due to shared data |
| Application programs dependent on the structure of the data (DATA DIVISION in COBOL) | Application programs independent of the structure of the data |
|  | Complexity of backup and recovery |
| Inflexibility - can't easily accommodate new needs (e.g. file reorganization necessary | More flexible formats - allows us to use the data in ways not originally specified by the end-users |

# Components of the Database Environment



* Computer-aided software engineering (CASE) tools– used to design databases and application programs
* Repository– a centralized knowledge base of all data definitions, data relationships, screen and report formats and other system components – contains the metadata for the database
* Database Management System (DBMS)– commercial software (and occasionally, hardware and firmware) system used to define, create, maintain and provide controlled access to the database and the repository
* Database– an organized collection of logically related data, usually designed to meet the information needs of multiple users in an organization
* Application Programs– Computer programs that are used to create and maintain the database and provide information to users
* User Interface– Languages, menus, and other facilities by which users interact with various system components, such as CASE tools, application programs, the DBMS and the repository.
* Data administrators– persons who are responsible for the overall information resources of an organization – use CASE tools to improve the productivity of database planning and design
* System developers– persons such as systems analysts and programmers who design new application programs – often use CASE tools for system requirement analysis and program design.
* End users– Persons throughout the organization who add, delete, and modify data in the database and who request or receive information from it. All user interactions with the database must be routed through the DBMS>
* The DBMS operational environment is an integrated system of hardware, software and people that is designed to facilitate the store, retrieval and control of the information resource and to improve the productivity of the organization

# The Range of Database Applications

* 5 categories: personal databases, workgroup databases, department databases, enterprise databases and web-enabled databases

## Personal Databases

* support one user
* reside on PC’s and mobile devices
* *example*: Customer contact lists for salespersons
* widely used because they can improve personal productivity
* ***risk***: the data cannot easily be shared with other users – if data are of interest to one person, they probably are (or will soon become) of interest to others as well.
* ***Key decisions in developing***:

1. Should the application be purchased from an outside vendor or developed within the organization?
2. If the database application is developed internally, should it be developed by the end user or by a professional with the information systems (IS) department?
3. What data are required by the user and how should the database be designed?
4. What commercial database management system (DBMS) product should be used for the application?
5. How should data in the personal database be synchronized with data in other databases?
6. Who is responsible for the accuracy of the data in the personal database?

## Workgroup Databases

* ***a workgroup*** 
  + a relatively small team of people who collaborate on the same project or application or on a group of similar projects or applications.
  + typically fewer than 25 people (e.g. a class or a team within a class)
* a workgroup database is designed to support the collaborative efforts of such a team
* each member of the workgroup has a desktop computer and the computers are linked by means of a local area network (LAN)
* the database is stored on a central device called the database server, which is connected to the network
* each member of the workgroup has access to the shared data
* different types of group members (e.g. developer or project manager) may have different user views of this shared database



* New problems: data security and data integrity with concurrent user data updating
* ***Key decisions in developing***:
* Same as for personal databases plus

1. How can the design of the database be optimized for a variety of group members’ information requirements?
2. How can the various members use the database concurrently without compromising the integrity of the database?
3. Which database processing operations should be performed at a workstation and which should occur on the server?

## Department Databases

* ***department***:
  + is a functional unit within an organization (e.g. Human Resources, marketing, manufacturing, accounting, Computer Science departments)
  + Larger than a workgroup (typically between 25 and 100 persons)
  + Responsible for a more diverse range of functions
* Department databases
  + support the various functions and activities of a department
  + the most common of the five types of databases
* ***example***: a personnel database that is designed to track data concerning employees, jobs, skills and job assignments.
  + Users can query the database to obtain answers to questions such as the following:
  1. For a particular job classification (such as Software Engineer), what job opportunities exist in the company at the present time?
  2. For that same job classification, what skill (or skills) are required?
  3. What skills are possessed by a particular employee? Conversely, which employees possess a particular skill (such as Java programming)?
  4. Which employees have held a particular job assignment? Conversely, what is the job history for a particular employee?
  5. Which employees are supervised by a particular manager?
* ***Key decisions in developing***:
* Same as for personal and workgroup databases plus

1. How can the database and its environment be designed to produce adequate performance, given the large number of users and user transactions?
2. How can adequate security be provided to protect against unauthorized disclosure or distribution of sensitive data?
3. What database and application development tools should be used in this complex environment?
4. Do other departments maintain the same type of data, and if so, how can data redundancy and consistency of data and metadata best be managed?
5. Are the users of the database geographically dispersed or the size of the database so great that data must be stored on several computer systems, thus creating a *distributed database*?

## Enterprise Databases

* one whose scope is the entire organization or enterprise (or at least, many different departments)
* intended to support organization-wide operations and decision making
* an organization may have several enterprise databases, so such a database is not inclusive of all organizational data
* has resulted in two major developments:
  1. Enterprise resource planning systems (ERP)
  2. Data warehousing implementations

#### Enterprise Resource Planning (ERP) Systems

* + integrates all functions of the enterprise, such as manufacturing, sales, finance, marketing, inventory, accounting and human resources
  + software applications that provide the data necessary for the enterprise to examine and manage its activities
  + evolved from the materials requirements planning (MRP) and manufacturing resources planning (MRP-II)
  + MRP systems scheduled the raw materials, component and subassembly requirements for manufacturing processes
  + MRP-II systems also scheduled shop floors and product distribution

#### Data Warehouses

* + an integrated decision support database whose content is derived from the various operational databases, including personal, workgroup and department databases
  + provide users with the opportunity to work with historical data to identify patterns and trends and answers to strategic business questions
  + ***Example****:* Consider a large health care organization that operates a group of medical centers including hospitals, clinics and nursing homes. Each center has a separate database, which is adequate to support most functions. A single consolidated view of the entire organization allows the organization to see the total activities with a single patient or supplier. Operating efficiencies can be achieved for example, by central ordering of supplies for all medical units and scheduling staff and services across all units. An ERP system makes these approaches possible. Corporate decision making, dealing with external suppliers and reporting to various agencies require compiled historical data and information 🡺 a data warehouse



* ***Key decisions in developing***:

1. How should the data be distributed among the various locations in the corporate structure?
2. How can the organization develop and maintain standards concerning data names, definitions, formats and related issues?
3. What actions must be taken in order to successfully integrate numerous systems, including legacy data from earlier systems that are desired for analysis.

## Web-enabled Databases

* Use the Internet to provide improved customer information and service
* *Examples*:
  + Customers configure and order their latest personal computer directly from the computer manufacturer
  + Airline tickets are ordered via the WWW
  + Information about open positions and company activities is readily available within many companies
* The easy connection to multiple platforms allows companies to reorganize their operation and develop new applications faster and at lower cost
* A standard interface allows users to be productive with less training and to require less support
* When a database is web-enabled, the Web browser interface allows users to ask unique and specific questions and receive answers based on current information
* Support BtoC (business to customer) and BtoB (business to business) types of applications
* **Extranets** used for BtoB applications
* An **extranet** uses Internet technology, but access is restricted to business suppliers and customers with whom an agreement has been reached about legitimate access and use of each others’ data and information.
* An **Intranet** uses Internet protocols to establish access to company data and information that is limited to the organization

Questions:

1. What technology (Web services, middleware, protocols, etc.) should be used to link Web applications to client databases?
2. What special measures are required to protect the security and privacy of data in this environment?
3. How should an organization manage the mountains of data that are generated through Internet transactions?
4. How can an organization maintain data quality when so much data are generated outside the organization?

### Summary of Database Applications

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Database** | **Typical number of users** | **Typical Architecture** | **Typical database size** |
| Personal | 1 | Desktop/laptop computer, PDA | Megabytes |
| Workgroup | 5-25 | Client/server (two-tier) | Megabytes – gigabytes |
| Department | 25-100 | Client/server (three-tier) | Gigabytes |
| Enterprise | >100 | Client/server (distributed or parallel server) | Gigabytes – terabytes |
| Internet | >1000 | Web server and application servers | Megabytes – gigabytes |