IT1105 – Information Systems and Technology

BIT – 1ST YEAR – SEMESTER 1
University of Colombo School of Computing

Student Manual

Lesson 3:
Organizing Data and Information
(6 Hrs)

Instructional Objectives

Students will be able to:

- Define general data management terms
- Identify advantages and disadvantages of the database approach to data management
- Identify different database models
- Identify some current database applications

3: Organizing Data and Information

3.1 Data Management

3.1.1 Introduction to DBMS

Without data and the ability to process it, an organisation would not be able to successfully complete most business activities. In order for the data to be transformed into useful information, it must be organised in a meaningful way.

3.1.2 The Hierarchy of Data

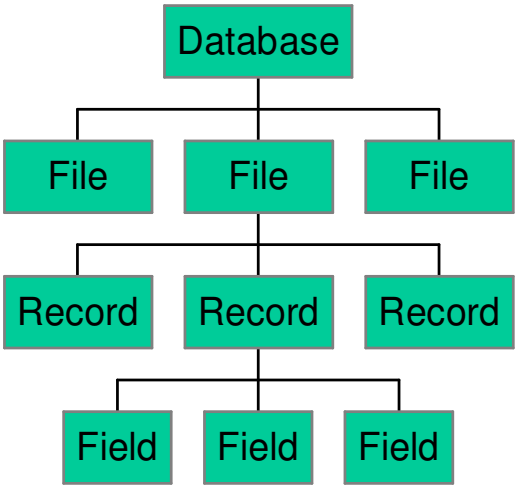

Figure 3.1: Hierarchy of Data

Data of a book is organised into characters, words, phrases, sentences, paragraphs and chapters. Similarly data in a database can be organised into fields, records and files that forms a hierarchy. Data hierarchy begins with the smallest piece of data used by computers (a bit) and progress through the hierarchy to a database.

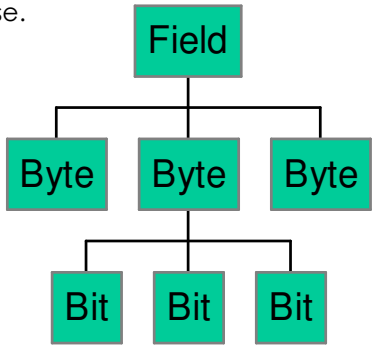

Figure 3.2: Composition of a Field

Bits can be organized into units called *bytes*. A byte is typically 8 bits. Each byte represents a **character**. Character is the basic building block of data, consisting of letters (A, B, C, …, Z, a, b, …, z), numeric digits (0, 1, 2, …, 9) or special symbols (., +, -, @, …).

Characters are put together to form a **field**. A field is typically a name (*employee name*), number (*salary*) or combination of characters (*national ID number*) that describes an aspect of a business object (e.g. an employee, a location, a vehicle) or activity (e.g. a sale).

A collection of related data fields is a **record**. An employee record is a collection of fields about an employee (i.e. *Name*, *Designation*, *Department*).

A collection of related records is a **file** (e.g. all employee records – typically known as the employee file)

Database is a collection of interrelated files (e.g. *Employee* file, *Department* file, *Payroll* file).
A **KEY** (e.g. Employee, Department and Payroll files can be linked by the Employee ID key)

3.1.3 Data Entities, Attributes, and Keys

Entities, attributes and keys are important database concepts. They are used to describe the database requirements of an application.

Entity is a generalised class of people (e.g. Employee, Student, Customer), places (e.g. City, Outlet, Warehouse) or things (e.g. Part, Item, Inventory) for which data is collected, stored and maintained. A record is an instance of an entity.

Attribute is a characteristic of an entity.
Employee name, Designation are attributes of an employee. Main purpose of an attribute is to capture the relevant characteristics of entities such as employees or customers. The specific value of an attribute, called a **data item**, can be found in the fields of the record describing an entity (e.g. *De Silva* is data item of the *name* attribute of an *employee* entity).

A **key** is a field or set of fields in a record that is used to identify the record. A **primary key** is a key that uniquely identifies the record (e.g. national ID number or employee number may be used to identify an employee uniquely).

3.1.4 Traditional Approach to Data Management

From the beginning of the use of computers to perform business functions, companies have used the traditional approach to process their functions. In the traditional approach separate files have been used for each application. Today it has changed to database approach which uses a unified and integrated database for most of the transactions of the company.

Traditional Approach

Manual method of managing data is by recording them on paper (e.g. filling an employee application form) and putting them in files (e.g. employee file), which are stored using filing cabinets of the personnel division.

One of the most basic ways to manage data electronically is via computer files, because a file is a collection of related records associated with a particular application.

The **traditional approach** to data management use separate data files for each application programme (e.g. employee file for personnel application, payroll file for payroll application). For a particular application one or more files were created. As shown in figure 3.3 each application programme had a file.

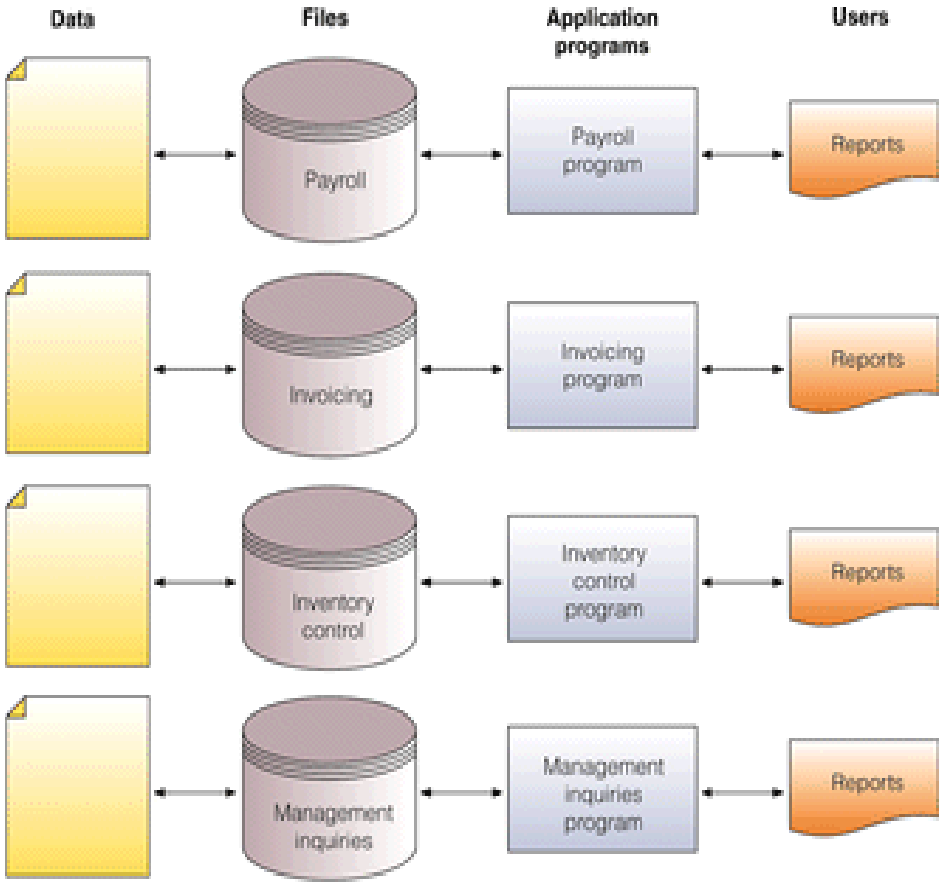

Figure 3.3: Traditional Approach to Data Management

Each division created and managed files required for their applications. Thus data which were common for several applications appeared in many files (e.g. employee name, address). This became one of the flaws of the traditional approach to data management (e.g. employee name and address appeared in employee file, payroll file, employee performance management file etc.). Duplication of data in separate files is known as **data redundancy**.

This caused problems when data had to be developed and coordinated to ensure that each file was properly updated. As this is difficult to achieve in practice lot of inconsistencies could occur among data stored in separate files.

Problems of the traditional approach are,

- Data Redundancy
  Independent data files included a lot of duplicated data; the same data (Example: Customer's name and address) was recorded and stored in several files. This data redundancy caused problems when data had to be updated, since separate file maintenance programs had to be developed and coordinated to ensure that each file was properly updated. Of course, this proved difficult in practice, so a lot of inconsistencies occurred among data stored in separate files.

- Lack of Data Integration
  Having data in independent files made it difficult to provide end users with information for ad hoc requests that required accessing data stored in several different files. Special computer programs had to be written to retrieve data from each independent file. This was so difficult, time consuming and costly for some organizations that it was impossible to provide end users or management with such information. If necessary, end users had

to manually extract the required information from the various reports produced by each separate application and prepare customized reports for management.

- Data Dependence
  In file processing systems, major components of a system – the organization of files, their physical locations on storage hardware, and the application software used to access those files – depended on one another in significant ways. For example, application programs typically contained references to the specific format of the data stored in the files they used. Thus, changes in the format and structure of data and records in a file required that changes be made to all of the programs that used that file. This program maintenance effort was a major burden of file processing systems. It proved difficult to do properly, and it resulted in a lot of inconsistency in the data files.

3.1.5 Database Approach to Data Management

To overcome the problems of the traditional approach to data management the **database approach** is used. In a database approach a pool of related data is shared by multiple application programs.

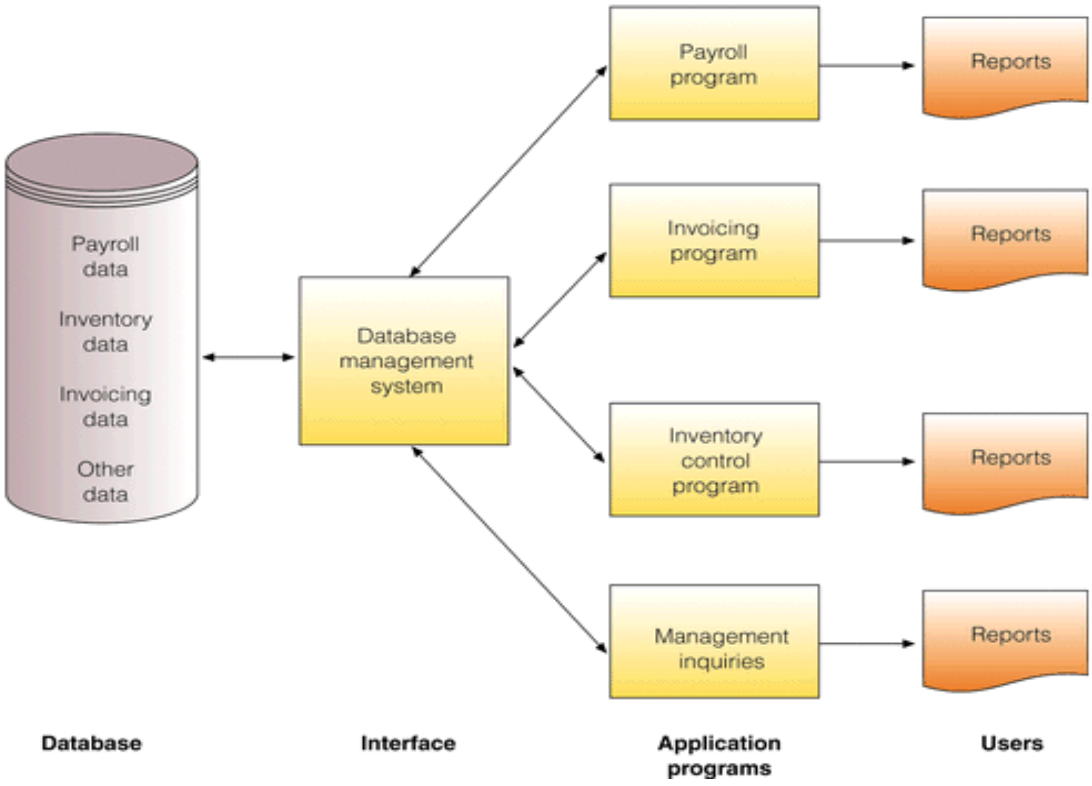

Figure 3.4: Database approach to Data Management

To use the database approach to data management, additional software called a **DataBase Management System** (DBMS) is required. The DBMS acts as a software interface between users and databases. This helps users to easily access the data in a database. Therefore, database management involves the use of database management software to control how databases are created, integrated and maintained to provide information needed by end users.

3.1.6 Advantages/disadvantages of Database Approach

The facilities offered by DBMS vary. However, a good DBMS should provide the following advantages.

Advantages of the Database Approach

- Data and program independence - both the database and the user program can be altered independently of each other.
- Ability to share data and non redundancy of data - enables applications to share an integrated database containing all the data needed by the applications and this eliminates data redundancies.
- Integrity – helps to maintain the integrity of data. Inconsistencies between two entries representing the same `fact' give an example of lack of integrity (caused by redundancy in the database).
- Centralized control - With central control of the database, the Database Administrator (DBA) can ensure that standards are followed in the representation of data.
- Security - Having control over the database the DBA can ensure that access to the database is through proper channels and can define the access rights of any user to any data items or defined subset of the database. The security system must prevent corruption of the existing data either accidently or maliciously.

  *DBA - a person responsible for the installing, configuring, upgrading, administrating, monitoring and maintaining of databases in an organization

Disadvantages of the Database Approach

- Costly
  - Specialized DBMS software
  - Specialized DBMS administrators and operators
- Increased vulnerability
  - Single point of failure
  - Targets for attacks

3.1.7 Types of DBMS

Data model is a tool that the database designers use to show the logical relationships among data. When data modeling done at a level of entire organization it is known as enterprise data modeling.

Based on the type of data modeling used different DBMS exist. They are hierarchical, network, relational and Object-Oriented models. Based on the no. of users too DBMS types are identified. They are single user (e.g. MS Access) and multi-user (e.g. Oracle) DBMS.

Let's consider the hierarchical, network, relational and Object-Oriented database models.

Hierarchical Databases

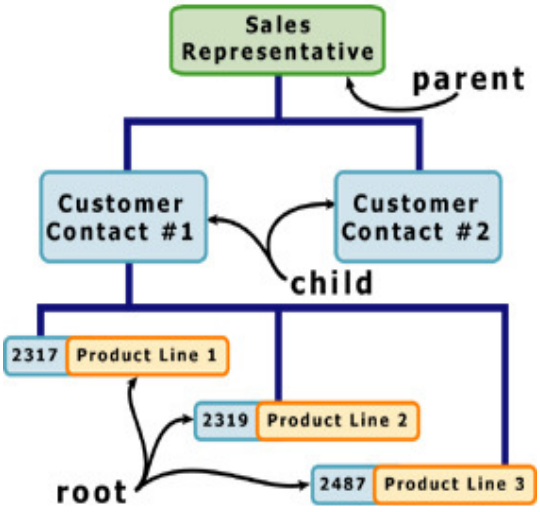

Figure 3.5: Hierarchical Database

It is one of the oldest methods of organizing and storing data, and used by few organizations.
Related fields or records are grouped together so that there are higher-level records and lower-level records, just like the parents in a family tree sit above the subordinated children. Furthermore, each child can also be a parent with children underneath it.

The parent record at the top of the pyramid is called the **root record**. A child record always has only one parent record to which it is linked, just like in a normal family tree. In contrast, a parent record may have more than one child record linked to it. Hierarchical databases work by moving from the top down. A record search is conducted by starting at the top and working down through the tree from parent to child until the appropriate child record is found.

The advantage of hierarchical databases is that they can be accessed and updated rapidly because the tree-like structure and the relationships between records are defined in advance. The disadvantage of this type of database structure is that each child in the tree may have only one parent, and relationships or linkages between children are not permitted, even if they make sense from a logical standpoint. Hierarchical databases are so rigid in their design that adding a new field or record requires that the entire database be redefined.

Network Databases

Figure 3.6: Network Database

Network databases are similar to hierarchical databases by also having a hierarchical structure. There are a few key differences, however. Instead of looking like an upside-down tree, a network database looks more like a cobweb or interconnected network of records. In network databases, children are called **members** and parents are called **owners**. The most important difference is that each child or member can have more than one parent (or owner).

Like hierarchical databases, network databases are principally used on mainframe computers. Since more connections can be made between different types of data, network databases are considered more flexible. However, two limitations must be considered when using this kind of database. Similar to hierarchical databases, network databases must be defined in advance. There is also a limit to the number of connections that can be made between records.

The Relational Database Model
The relational model describes data using a standard tabular format. In a database structured according to the relational model, all data elements are placed in two dimensional tables, called relations which are the logical equivalent of files. The tables in relational databases organize data in rows and columns, simplifying data access and manipulation. It is easier for managers to understand the relational model than other database models.

In the relational model, each row of a table represents a data entity, with the columns of the table representing attributes. Each attribute can take on only certain values. The allowable values for these attributes are called the domain. The domain for a particular attribute indicates what values can be placed in each of the columns of the relational table.
The relational database model is widely used.

Manipulating Data

Data table 1: Project table

Project number	Description	Dept. number
155	Payroll	257
498	Widgets	632
226	Sales Manual	598

Data table 2: Department table

Dept. number	Dept. name	Manager SSN
257	Accounting	005-10-6321
632	Manufacturing	549-77-1001
598	Marketing	098-40-1370

Data table 3: Manager table

SSN	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-1997	257
549-77-1001	Buckley	Bill	02-17-1979	632
098-40-1370	Fiske	Steven	01-05-1985	598

Figure 3.7: Link between Databases

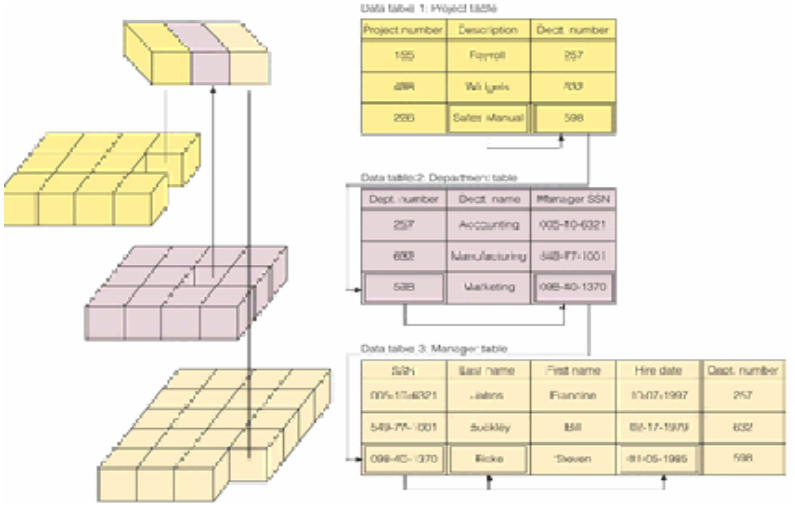

Figure 3.8: Linking Databases (supporting inquiries)

Once data has been placed into a relational database, users can make inquiries & analyze data. To manipulate relational databases a set of relational operators have been defined. Basic data manipulations using relational operators include selecting, projecting & joining.

Selecting involves eliminating rows according to certain criteria.
Projecting involves eliminating columns in a table.
Joining involves combining two or more tables.

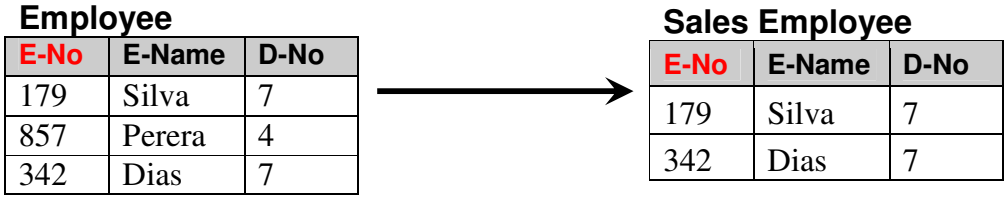

Figure 3.9: Selecting Operation

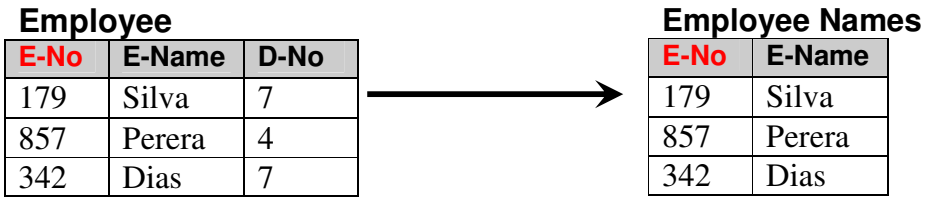

Figure 3.10: Projection Operation

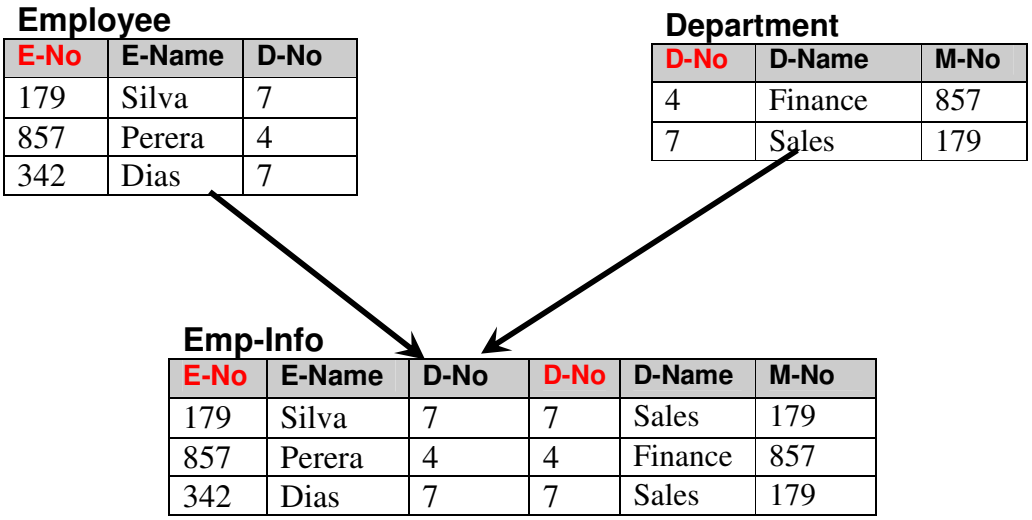

Figure 3.11: Join Operation

As long as the table share at least one common data attribute, the tables in a relational database can be linked to provide useful information and reports. One of the primary advantages of a relational database is that it allows tables to be linked. It is easier to control, more flexible, more intuitive than other approaches because it organizes data in tables. The ability to link relational tables also allows users to relate data in new ways without having to redefine complex relationships.

Object Oriented Database Model

Hierarchical and network databases are all designed to handle structured data; that is, data that fits nicely into fields, rows, and columns. They are useful for handling small snippets of information such as names, addresses, zip codes, product numbers, and any kind of statistic or number you

can think of. On the other hand, an object-oriented database can be used to store data from a variety of media sources, such as photographs and text, and produce work, as output, in a multimedia format.
Object-oriented databases use small, reusable chunks of software called objects. The objects themselves are stored in the object-oriented database. Each object consists of two elements: 1) a piece of data (e.g., sound, video, text, or graphics), and 2) the instructions, or software programs called methods, for what to do with the data. Part two of this definition requires a little more explanation. The instructions contained within the object are used to do something with the data in the object. For example, test scores would be within the object as would the instructions for calculating average test score.

Object-oriented databases have two disadvantages. First, they are more costly to develop. Second, most organizations are reluctant to abandon or convert from those databases that they have already invested money in developing and implementing. However, the benefits to object-oriented databases are compelling. The ability to mix and match reusable objects provides incredible multimedia capability. Healthcare organizations, for example, can store, track, and recall CAT scans, X-rays, electrocardiograms and many other forms of crucial data.

3.2 Database Management Systems and Applications

Database Management Systems (DBMS) is a collection of programs that manages a databases structure and control access to the data stored in the database. It is a software which facilitates the process of defining, storing, manipulating and sharing database among various users and applications.

As stated below, **Database Applications** enable organisations to generate information useful for decision making. Furthermore, with the help of databases organisations try to improve their efficiency as well as achieve competitive.
For e.g. databases support organisations to carry out data mining and business intelligence which will help to identify customer preferences effectively.

3.2.1 Popular Database Management Systems

Typical users of the DBMS are Database Administrator, database designers, end users, systems analysts and application programmers.
DBMS performs several important factors which are discussed below.

Like other software products, there are a number of commercial database systems (e.g. SQL Server, DB2, Oracle, Informix) and open-source (PostgreSQL and MySQL).
mySQL : This is the most popular open-source database management system.

3.2.2 Linking the Company Database to the Internet

Today customers, suppliers and company employees must be able to access corporate database through the internet, intranet and extranet to meet various business needs.
Example:

1. When a customer is going to buy a book through internet he is accessing a database to find the book information, author, price, etc.
2. With the help of the databases the suppliers can check the raw materials and the current production schedule to determine when & how much of their products must be delivered to support *just-in-time* inventory management
3. Employees Of a company working from abroad may want to access the internal databases through the Internet or the intranet to make important decisions.

Developing a seamless integration of traditional databases with the internet is often called a **semantic web.** The semantic web is about taking the relational database & webbing it. It allows

accessing and manipulating a number of traditional databases at the same time through the internet.

Instead of the internet, organizations are gaining access to databases through networks to get good prices and reliable services. However, linking company databases to external network such as the Internet can be potentially dangerous due to issues related to security. For example a competitor or any other hacker may gain access to these databases.

3.2.3 Data Mining Applications

Data warehouses and Data Mining
The raw data necessary to make sound business decisions is stored in a variety of locations and formats.

Using data warehouses and data mining data can be used to support decision making.

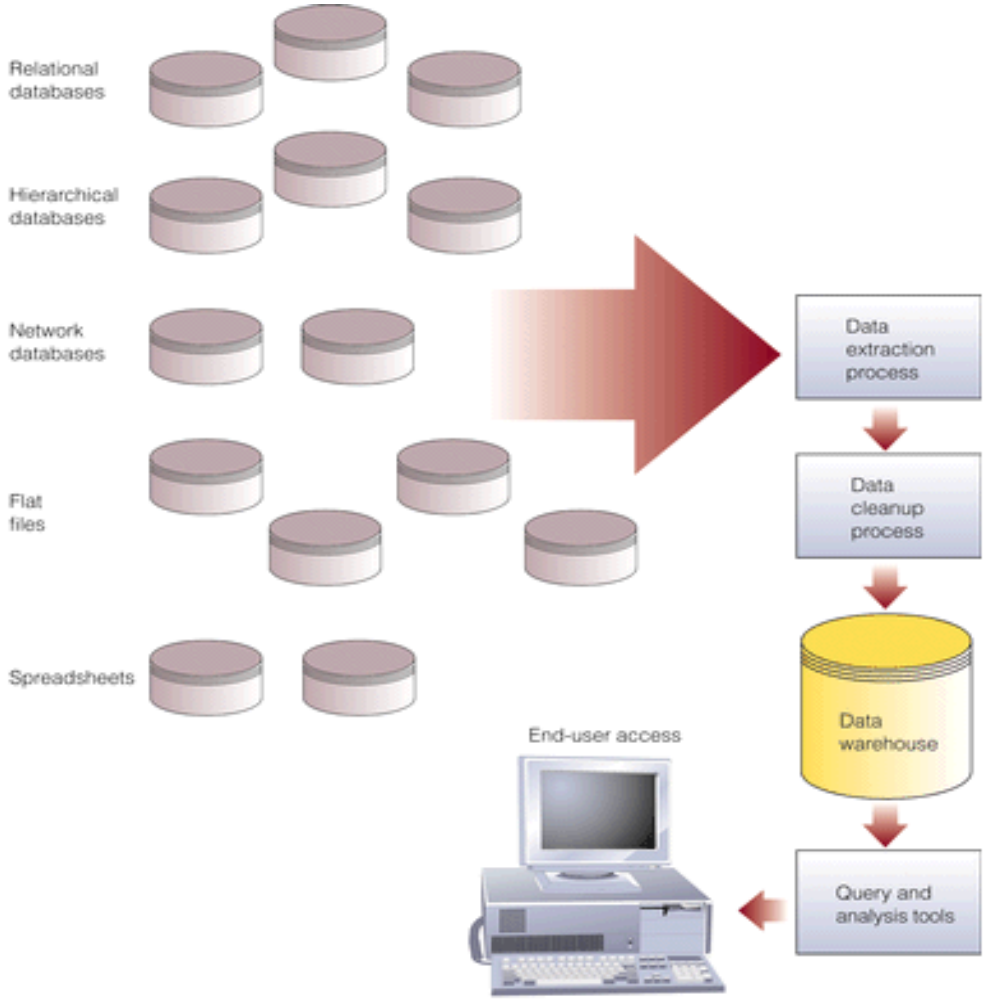

Figure 3.12: Data Warehouses

A data warehouse stores data that have been extracted from the various operational, external and other databases of an organisation. It is a central source of the data that have been cleaned, transformed, catalogued so that they can be used by managers and other business professionals for data mining, online analytical processing and other form of

business analysis, market research and decision support. A data warehouse can also be viewed as a database for historical data from different functions within a company.

The structure of data warehouse is easier for end users to navigate, understand and query against unlike the relational databases it is primarily designed to handle lots of transactions. Data warehouse enable queries that cast across different segments of a company's operation.

Example: Production data could be compared against inventory data even if they were originally stored in different tables with different structures. Data warehousing is an efficient way to manage and report on data collected from a variety of sources, which are non uniform and scattered throughout a company.

Data Mining
Data mining is a major use of data warehouse databases. It is an information-analysis tool that involves the automated discovery of hidden patterns and trends in historical business activity. Data mining's objective is to extract patterns, trends and rules from data warehouses to evaluate (i.e. predict or score) proposed business strategies, which in turn will improve competitiveness, improve profits and transform business processes. With the help of data mining it is possible to improve customer retention, campaign management and customer segmentation analysis.

3.2.4 Business Intelligence

Databases can be used for the purpose of business intelligence closely linked to data mining. Business Intelligence is the process of gathering enough of the right information in a timely manner and usable form and analyzing it so that it can have a positive impact on business strategy, tactics or operations. Business Intelligence turns data into valuable information and distributes it throughout an enterprise. This information is used by the companies to improve strategic discussions about which markets to enter, how to select and manage key customer relationships, how to improve sales promotions etc. Business Intelligence tools (applications) can be found in different categories such as Business planning, Customer Relationship Management (CRM), Management Information Systems (MIS) etc.

Online Analytical Processing (OLAP)
Online analytical processing allows users to explore data from a number of different perspectives. OLAP involves analysing complex relationships among thousands or even millions of data items stored in data marts, data warehouses, and other multidimensional databases to discover patterns, trends and exceptional conditions.

3.2.5 Important factors when Selecting a Database Management System

The database administrator often selects the best database management system for an organization. The process begins by analyzing database needs and characteristics. The information needs of the organization affect the type of data that is collected and the type of database management system that is used.

The important features that have to be considered when selecting a Database Management System are as follows.

- Database size

Database size depends on the number of records or files in the database. The size determines the overall storage requirement for the database.
To maintain good performance and to reduce costs companies are trimming the size of their databases.

- Number of concurrent users

Number of simultaneous users that can access the contents of the database is also an important factor. A database that is used by a large workgroup must be able to support

number of concurrent users. If it cannot, then the efficiency of the user requests will be lowered. To provide flexibility to the database, highly scalable DBMS is preferred by the companies. Scalability describes how well a database performs as the size of the database or the number of concurrent users increase.

- Performance

How fast the database is able to update records can be the most important performance criterion for some organizations.
Example: Credit and airline companies must have database systems that can immediately update customer records and check credit or make a plane reservation in seconds not minutes. However payroll applications can be processed once a week or less frequently and do not require immediate processing. When an application demands immediacy, it also demands rapid recovery facilities in the event that the computer system shuts down temporarily. Other performance considerations include the number of concurrent users that can be supported and the amount of memory that is required to execute the database management program.

- Integration

A key aspect of any database is its' ability to be integrated with other applications and databases. A key determinant here is what operating systems it can run under – such as Linux, UNIX or Windows. Some companies use several databases for different applications at different locations.

- Features

The features of the database management system can also make a big difference. Most database programs come with security procedures, privacy protection and a variety of tools.

- The vendor

The size, reputation and financial stability of the vendor is also an important aspect. Some organisations would rely on vendor support to handle operational aspects of the system.

- Cost

Cost of a database system varies from few thousands to millions of rupees based on the number of users and functionalities. In addition to the initial cost of the database package, annual or monthly maintenance or operating costs should be considered.