Chapter Unit 6

Foundations of Business Intelligence: Databases and Information Management

Data vs. Information

Data is raw facts collected from environment about physical phenomena or business transactions. Data can be in any form-numerical, textual, graphical, image, sound, video etc. It has no meaning. It is input to any system in an organization. For example, data would be the marks obtained by students in different subjects.

On the other hand information is defined as refined or processed data that has been transformed into meaningful and useful form for specific users. For example, after processing the marks obtained by student it transformed into information, which is meaningful and from which we can decide which student stood first, second and so forth. Information comes from data and takes the form of table, graphs, diagrams etc.

Data hierarchy \file organization concept

A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases. A bit represents the smallest unit of data a computer can handle. A group of bits, called a byte, represents a single character, which can be a letter, a number, or another symbol. A grouping of characters into a word, a group of words, or a complete number (such as a person's name or age) is called a **field**. A group of related fields, such as the student's name, the course taken, the date, and the grade, comprises a **record**; a group of records of the same type is called a **file**.

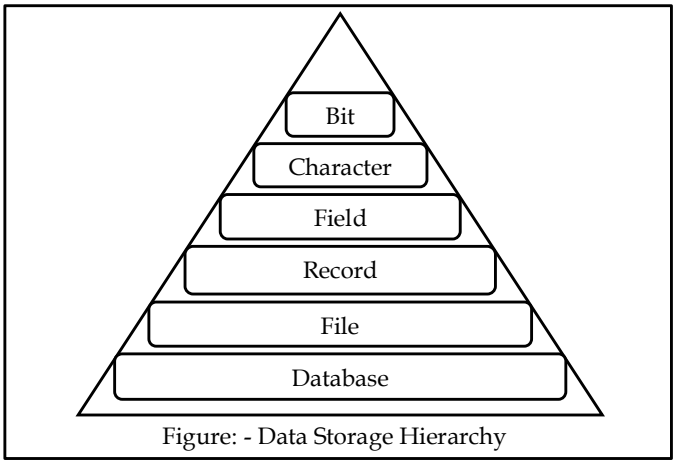
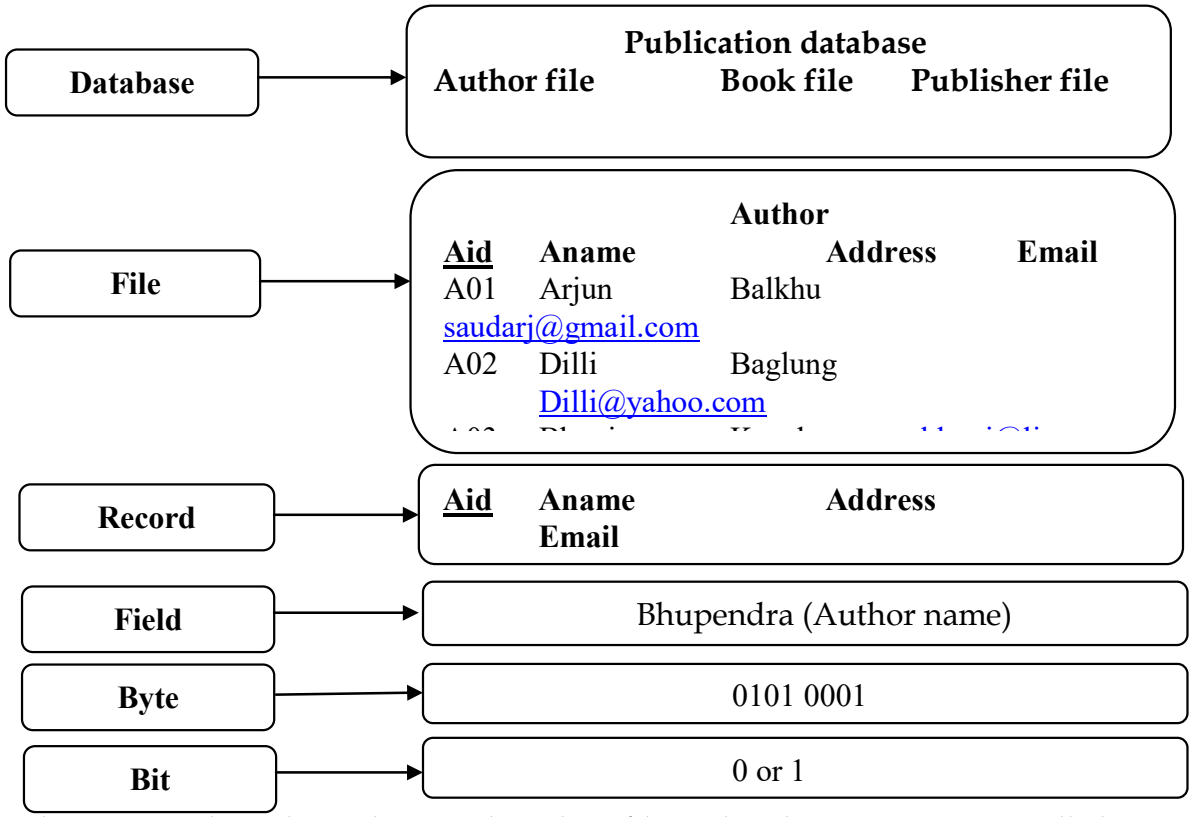


Figure: - Data Storage Hierarchy

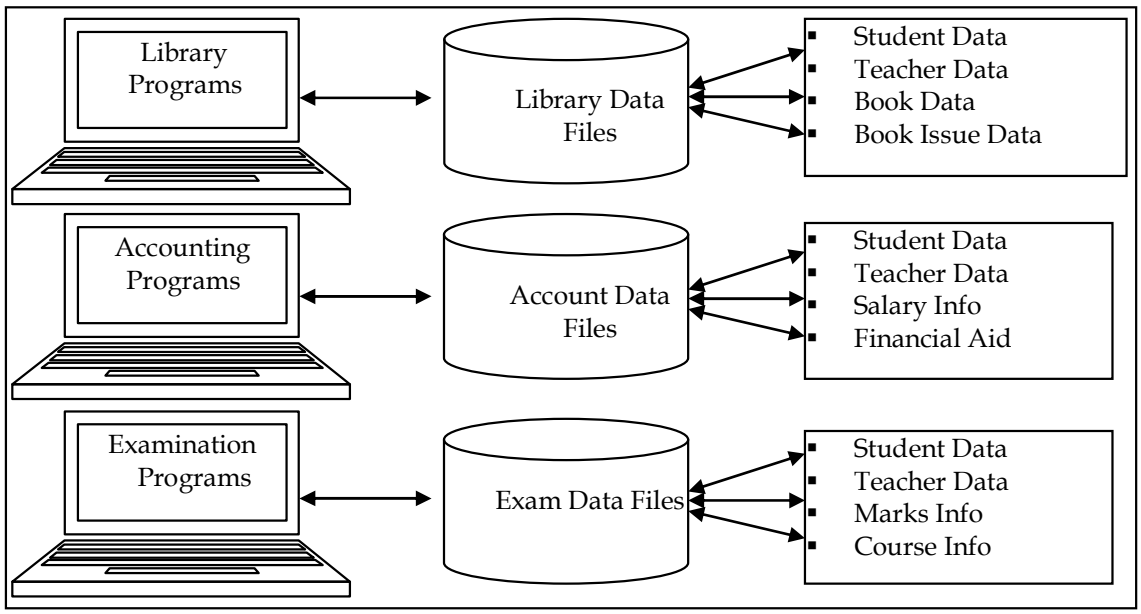
Organizing Data in a traditional file environment

File management systems (FMS) are also called flat file systems. It stores data in a plain text file. A flat file is a file that contains records, and in which each record is specified in a single line. Fields from each record may simply have a fixed width with padding, or may be delimited by whitespace, tabs, commas or other characters. Extra formatting may be needed to avoid delimiter collision. There are no structural relationships and the data are "flat" as in a sheet of paper. For example, the records in Figure below could constitute an author file. A group of related files makes up a **database (E.g. files author, book, and publisher etc makes up a database publication)**. The author file illustrated in Figure below could be grouped with files on book and publisher to create a publication database.



In this approach each application has data files related to it containing all the data records needed by the application. Thus, an organization has to develop number of application programs each with an associated application-specific data files. For example, in a college MIS, the library programs, accounting

programs, and examination programs would have their own data files as shown in figure below:



Problems with the traditional file environment

The key problems in the traditional file environment are listed below:

- Data Redundancy
- Data Inconsistency
- Data Isolation
- Difficulties in accessing data
- Integrity problem
- Atomicity problem
- Concurrent access anomalies
- Security problems

Data redundancy: Data redundancy means duplication of same data or data files in different places. Flat file systems are suffered from the problem of high data redundancy. For example, record (such as student id, name, level, program, section etc) of a student may appear in library data files as well as examination data files. This redundancy leads to higher storage and access cost. On the other hand database management systems can greatly reduce the problem of data redundancy. Note that DBMS cannot remove data redundancy problem completely.

Data inconsistency: Data inconsistency is side effect of data redundancy. Data is said to be inconsistent if various copies of the same data may no longer agree. Data inconstancy occurs if changed data is reflected in data files in one place but not elsewhere in the system. For example, if library data file contains cell number of a student as 9841567843 but examination data files stores 9851167895 as cell number of the student then we can say that data is inconsistent. Flat file systems may suffer from the problem of data inconsistency. But database systems can remove the problem of data inconstancy by automatically propagating data updates done in one file in the database in other data files.

Data isolation: Because data are scattered in various files, and files may be in different formats, writing new application programs to retrieve the appropriate data is difficult in flat file systems. For example, one data file may contain data separated of comma and another data file may contain data separated by white space.

Student Data file
S01, Suman, BBA,Mahendranagar
S02, Binu, B.Sc. CSIT, Kathmandu
S03, Ronit, B.E. Computer, Chitwan

(a)

Book Data file
B05 DBMS SaudArjun Kanchanpur
B06 java. SharmaDilli Baglung
B07 MIS BhattaJagadish Kanchanpur

(b)

Figure: *Data files (a) Data separated by comma (b) Data separated by white space*

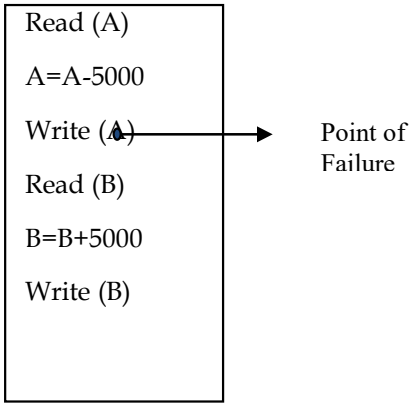
Database management systems provide shared access to centrally stored data therefore it is easy for application programs to retrieve required data from centralized database. Application programs do not need to bother about format of stored data.

Difficulty in accessing data: File processing systems do not allow required data to be retrieved in efficient and convenient way**.** For example, assume we already have program to generate the list of books on the basis of subject. Now, if we need to generate the list of books on the basis of author name, either we need to extract the data from book data files manually or we should request the programmer to write a program to retrieve required data from the book data file. Both of the alternatives are not satisfactory.

Integrity problems: Integrity means correctness of data before and after execution of a transaction. Integrity constraints are condition applied to the data. For example, if maximum salary in an organization is 150,000 then we have the integrity constraint "salary ≤ 150,000". Integrity constraints are important to maintain correctness of data. It plays vital to prevent users from doing mistakes. For example, if user mistakenly types 200,000 in place of 20,000 while transferring salary of an employee in his/her account, specified integrity constrain is violated and hence the system tell the user about the mistake.
Unfortunately, flat file systems do not allow us to specify integrity constraint and hence it is difficult to maintain correctness of data. On the other hand, database management systems allow us to specify integrity constraints on data therefore relatively it is easy to maintain correctness of data

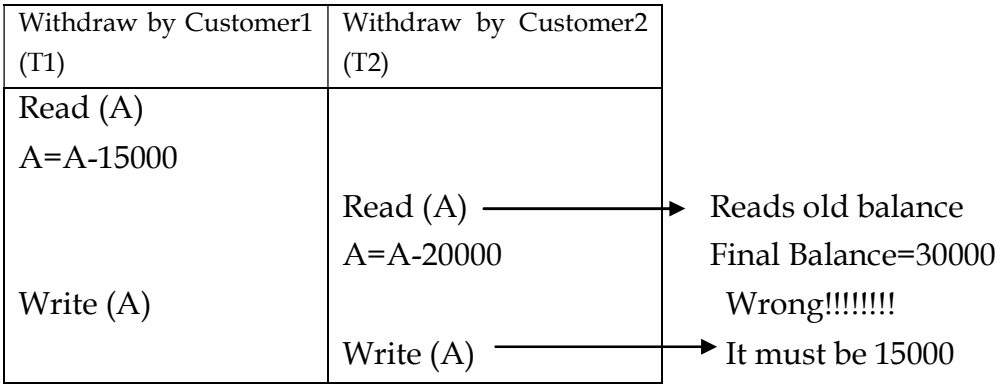
.

Atomicity problems: Execution of transactions must be atomic. This means transactions must execute at its entirety or not at all. If execution of transaction is not atomic, it leaves database in incorrect sate. Consider the example of transaction that transfers 5000 rupees from account A to account B.
If the execution of transaction is failed at the point specified in above figure, it causes 5000 rupees to be deducted from account A without depositing it in account B. File processing system do not guarantee atomic execution of transactions and hence this type of problems may occur in databases. But database systems guarantees atomicity of execution of transaction and hence above mentioned problem can be eliminated.



Concurrent-access anomalies: Concurrent updates to same data by different transactions at the same time may result in inconsistent data. Consider bank account 'A', containing 50,000 Rs. If two customers withdraw funds say 15,000

and 20,000 respectively from account A at about the same time, the result of the concurrent executions may leave the account in an incorrect (or inconsistent) state, if the programs executing on behalf of each withdrawal read the old balance as below. Flat file systems do not supports execution of concurrent transactions and hence may suffer from the problem mentioned below. But, database systems support concurrent execution of transactions on the same data without resulting into inconstant data.



Security problems: In database system we may create different user accounts and provide different authorization to different users. Thus we are able to hide certain information from some users. For example, in a banking system, payroll personnel need to see only that part of the database that has information about the various bank employees. They do not need access to information about customer accounts. This type of restriction is essential for security purpose. But, file processing system do not allow us to create user accounts thus all users have equal access to the data. Due to this it is difficult to maintain security of flat file systems. Besides this file processing systems do not have any provisions for periodic backup of data and recovery from data loss which are provided by database management systems.

The database approach to data management

Database technology cuts through many of the problems of traditional file organization. A more rigorous definition of a **database** is a collection of data organized to serve many applications efficiently by centralizing the data and

controlling redundant data. Rather than storing data in separate files for each application, data are stored so as to appear to users as being stored in only one location.

Database Management Systems (DBMS)

A database is an organized collection of logically related data that contains information relevant to an enterprise. The database is also called the repository or container for a collection of data files. For example, **university database** maintains information about students, courses and grades in university.

A **Database Management System (DBMS)** is the set of programs that is used to store, retrieve and manipulate the data in convenient and efficient way. Main goal of database management system (DBMS) is to hide underlying complexities of data management from users and provide easy interface to them. Some common examples of the DBMS software are Oracle, Sybase, Microsoft SQL Server, DB2, MySQL, Postgres, Dbase, Ms-Access etc. Database management system that maintains relationship between multiple data files is called **Relational Database Management system** (**RDBMS**).

Student

S-ID	*Name*	*Address*	*Program-ID*
S-12	Pawan	Joshi	C002
S-14	Yamman	Karki	C021
S-51	Abin	Saud	C321
S-11	Binak	Singh	C112

Program-ID	*Program-Name*
C002	BBA
C021	B. Sc CSIT
C112	BIM
C321	B. ed.

Fi

g: F Foreign Keys **oreign key**

A **database system** consists of database, d ***Program*** gement system, and application programs. tware that uses DBMS for data management is called database system. Thus MS word is just an application program but it is not databas Primary Keys e it does not use DBMS for the purpose of managing data. On the other hand a library management information system can be database system if it uses DBMS for the purpose managing database. As mentioned above database is a repository for a collection of computerized data files. Users of database system can perform a variety of operation on such file. For example, in a database management system, the library system, accounting management system, and examination system

Relationships

programs would have a common database. This database based approach to data processing is shown in fig below:

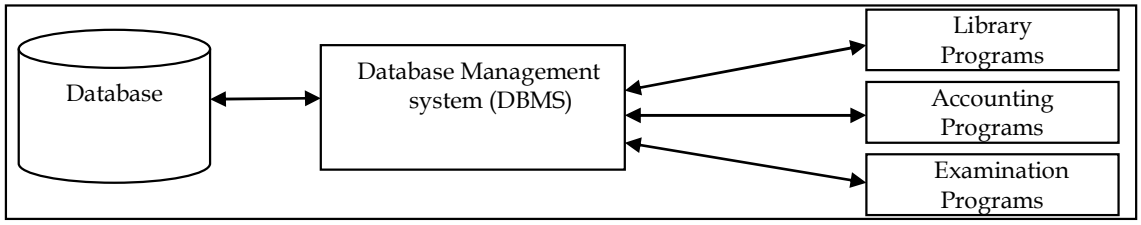


Figure: Database system approach to data processing

Capabilities of database Management systems

A DBMS includes capabilities and tools for organizing, managing, and accessing the data in the database. The most important are its data definition language, data dictionary, and data manipulation language.

DBMS have a **data definition** capability to specify the structure of the content of the database. It would be used to create database tables and to define the characteristics of the fields in each table. This information about the database would be documented in a data dictionary. A **data dictionary** is an automated or manual file that stores definitions of data elements and their characteristics.

The key capabilities of database management systems are listed below:

- Querying and reporting
- Maintaining complex relationship among data
- Provide backup and recovery
- Data availability
- Maintaining data integrity
- Minimize data redundancy
- Improve data security
- Handling concurrent access anomalies

Querying and reporting

The database contains the huge amount of data. Querying helps to filter the data and present only what the user requires. The most popular type of query language is SQL. It uses English like structured syntax for creating queries.

Now after filtering data from the database through query languages it is equally necessary to present the data in an appropriate structure. DBMS have the

capabilities to generate reports on the user-desired data based on user-desired structure. Crystal report is a popular report generator for large corporate DBMS.

Maintaining complex relationship among data

A database is the collection of interrelated data. The DBMS has the capabilities of creating a link between the data in various tables. This capability helps to retrieve complete information in a timely fashion as well as efficient update of data. In DBMS this type of relation can be created by using the concept of primary key and foreign key i.e. using referential integrity constraint concept.

Provide backup and recovery

DBMS has the capability to create a backup of the data. This backup can be used to recover the lost data when accidental loss of data occurs.

Data availability

One of the principle advantages is that the same business data can be made available to different employees anytime anywhere. DBMS enables multi-user access to information that is available remotely and 24 hours a day, 7 days a week.

Maintaining data integrity

Data accuracy, consistency and relevant data is a sign of data integrity. With the help of DBMS, companies can foster the integrity of their business records because changes to the data only have to be made in one place.

Minimize data redundancy

DBMS has the capabilities to keep the relationship among different tables due to which there is less chances of data redundancy. Information in DBMS is kept concise and appears just once. This capability reduces data redundancy.

Improve data security

DBMS allows user authentication through password. Authorized users can access only the data for which they are granted privilege. Thus it provides a system of permissions to restrict user access to certain data resources.

Handling concurrent access anomalies

Database management system prevents from accessing the data concurrently. It is achieved through isolation, which guarantees that other operations cannot access data that are being processed or modified during the transaction until it is completed.

How a DBMS Solves the Problems of the Traditional File Environment?

A DBMS reduces data redundancy and inconsistency by minimizing isolated files in which the same data are repeated. The DBMS may not enable the organization to eliminate data redundancy entirely, but it can help control redundancy. Even if the organization maintains some redundant data, using a DBMS eliminates data inconsistency because the DBMS can help the organization ensure that every occurrence of redundant data has the same values. The DBMS uncouples programs and data, enabling data to stand on their own. Access and availability of information will be increased and program development and maintenance costs reduced because users and programmers can perform ad hoc queries of data in the database. The DBMS enables the organization to centrally manage data, their use, and security.

Designing Database

To create a database, you must understand the relationships among the data, the type of data that will be maintained in the database, how the data will be used, and how the organization will need to change to manage data from a company-wide perspective. Designing a database requires an understanding of the business functions you want to model. It also requires an understanding of the database concepts and features that you want to use to represent those business functions. Make sure that you accurately design the database to model the business, because it can be time-consuming to significantly change the design of

a database after you implement it. A well-designed database also performs better.

The database requires both a conceptual design and a physical design. The conceptual, or logical, design of a database is an abstract model of the database from a business perspective, whereas the physical design shows how the database is actually arranged on direct-access storage devices.

The conceptual database design describes how the data elements in the database are to be grouped. The design process identifies relationships among data elements and the most efficient way of grouping data elements together to meet business information requirements. The process also identifies redundant data elements and the groupings of data elements required for specific application programs. Groups of data are organized, refined, and streamlined until an overall logical view of the relationships among all the data in the database emerges.

The conceptual database design deals with two important concepts:

- ☞ Normalization and
- ☞ Entity relationship diagram

1) Normalization

"The process of decomposing unsatisfactory "bad" relations by breaking up their attributes into smaller relations is called normalization"

While designing a database out of an entity–relationship model, the main problem existing in that raw database is redundancy. Redundancy is storing the same data item in more one place. A redundancy creates several problems like the following:

- Extra storage space: storing the same data in many places takes large amount of disk pace.
- Entering same data more than once during data insertion.
- Deleting data from more than one place during deletion.
- Modifying data in more than one place.
- Anomalies may occur in the database if insertion, deletion, modification etc are no done properly. It creates inconsistency and unreliability in the database.

To solve this problem, the raw database needs to be normalized. This is a step by step process of removing different kinds of redundancy and anomaly at each

step. At each step a specific rule is followed to remove specific kind of impurity in order to give the database a slim and clean look. The process of reducing data redundancy and removing database modification anomaly in a relational database is called normalization.

In brief the process of creating small, stable, yet flexible and adaptive data structures from complex groups of data is called **normalization**.

Example: Let's take a relation that is in un-normalized form as,

Student

Sid	Sname	Address	Phone_No
1	Bishnu	Kalanki	9849145464, 9813335467
2	Ramhari	Balkhu	9841882345, 099392844
3	Geeta	Kirtipur	9848334898,
4	Dipika	Pokhara	9849283847
5	Monika	Ratopool	9840084732, 9803267499

Since in this relation multi-valued attribute exist thus this relation is not in normalized form. Now converting this relation into normal form by decomposing this relation into two relations as,

Student

Sid	Sname	Address
1	Bishnu	Kalanki
2	Ramhari	Balkhu
3	Geeta	Kirtipur
4	Dipika	Pokhara
5	Monika	Ratopool

Phone

Sid	Phone_No
1	9849145464
1	9813335467
2	9841882345
2	099392844
3	9848334898
4	9849283847
5	9840084732
5	9803267499

Fig: Relations in Normalized form

Example 2: **Employee-Department**

Emp-Id	Emp-Name	Emp-Salary	Dept-No	Dept-Name
1	Bhupi	40000	D1	BBA
1	Bhupi	40000	D2	CSIT
2	Bindu	30000	D3	BBS

3	Arjun	60000	D1	CSIT

In the above relation {Emp-Id, Dept-No} is the primary key. Emp-Name, Emp-Salary and Dept-Name all depend upon {Emp-Id, Dept-No}. Again Emp-Id→Emp-Name, Emp-Id→Emp-Salary and Dept-No→Dept-Name, thus there also occur partial dependency. Due to which this relation is not in 2 NF.

Now converting this relation into 2 NF by decomposing this relation into three relations as,

Employee

E_Id	***E_Name***	***E_Salary***
1	Bhupi	40000
2	Bindu	30000
3	Arjun	60000

Emp-Dept

E_Id	***D_No***
1	*D1*
1	*D2*
2	*D3*
3	*D1*

Department

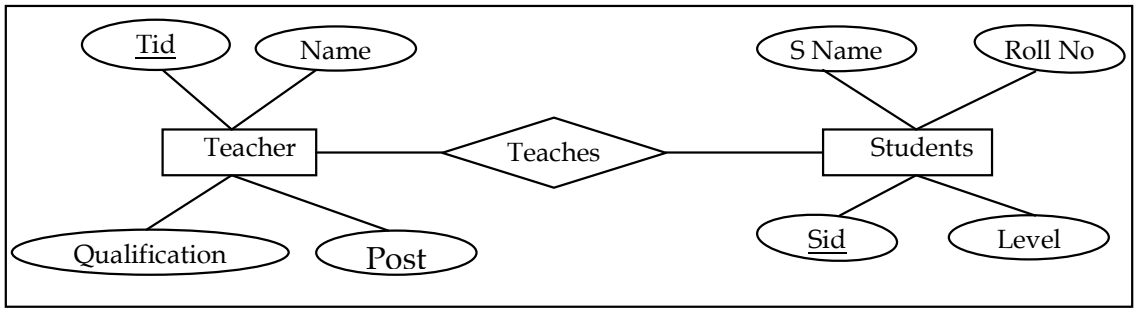
D_No	***D_Name***
D1	BBA
D2	CSIT
D3	BBS

Fig: Relations in 2 NF

2) Entity relationship diagram

An **E-R diagram** is a specialized graphical tool that demonstrates the interrelationships among various entities of a database. It is used to represent the overall logical structure of the database. While designing E-R diagrams, the emphasis is on the schema of the database and not on the instances. This is because the schema of the database is changed rarely; however, the instances in the entity and relationship sets change frequently. Thus, E-R diagrams are more useful in designing the database. E-R diagram focuses high level database design and hides low level details of database representation therefore it can be used to communicate with users of the system while collecting information.

Example:



Using databases to improve business performance and decision making

In a large company, with large databases or large systems for separate functions, such as manufacturing, sales, and accounting, special capabilities and tools are required for analyzing vast quantities of data and for accessing data from multiple systems. These capabilities include data warehousing, data mining, and tools for accessing internal databases through the Web.

Data Warehouse

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision-making process.

a. **Subject-Oriented**: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

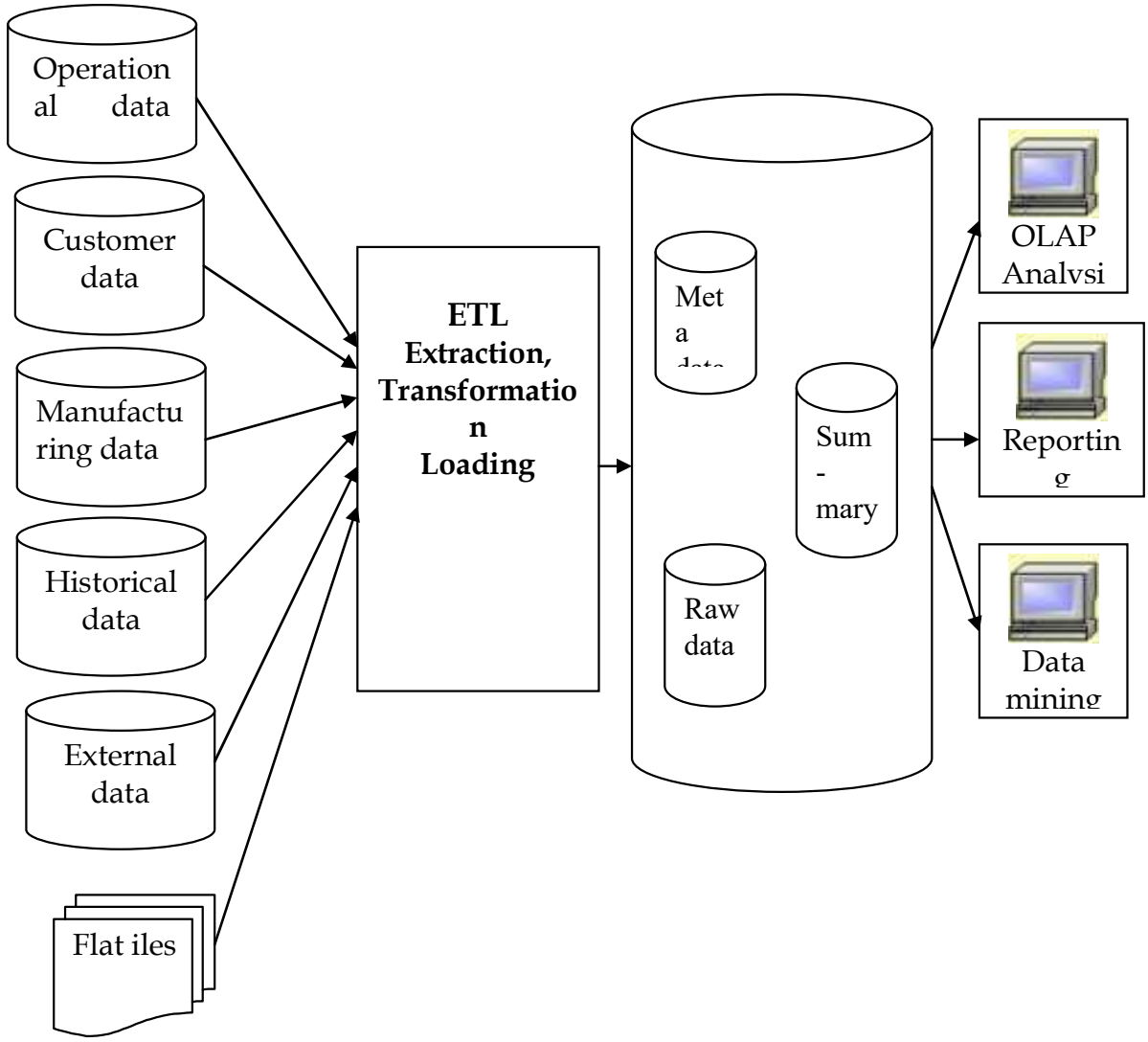
b. **Integrated**: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

c. **Time-Variant**: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

d. **Non-volatile**: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

A **data warehouse** is a repository of current and historical data of an organization that are organized to facilitate reporting and analysis. The data originate in many core operational transaction systems, such as systems for sales, customer accounts, and manufacturing, and may include data from Web site transactions. The data warehouse consolidates and standardizes information from different operational databases so that the information can be used across

the enterprise for management analysis and decision making. Figure below illustrates how a data warehouse works. The data warehouse makes the data available for anyone to access as needed, but it cannot be altered. A data warehouse system also provides a range of ad hoc and standardized query tools, analytical tools, and graphical reporting facilities. Many firms use intranet portals to make the data warehouse information widely available throughout the firm.



Data warehouse

How does a data warehouse differ from a database?

There are a number of fundamental differences which separate a data warehouse from a database. The biggest difference between them is that most database place an emphasis on a single application, and this application will generally be one that is based on transaction. If the data is analyzed, it will be done within a single domain. In contrast, data warehouses deal with multiple domains simultaneously.

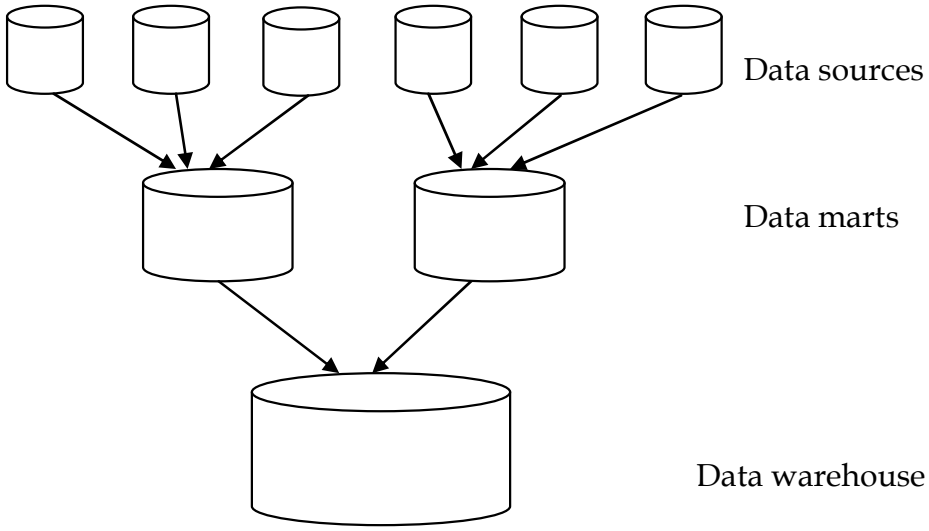
Because data warehouse deals with multiple subject areas, the data warehouse finds connections between them. This allows the data warehouse to show how the company is performing as a whole, rather than in individual areas.

Another powerful aspect of data warehouse is their ability to support the analysis of trends. They are not volatile, and the information stored in them doesn't change as much as it would in a common database. Some of the major differences between them are listed below:

Database	Data Warehouse
1. In database tables and joins of different tables are complex since they are normalized for RDBMS. This is done to reduce redundant data and to save storage space. 2. Entity Relational modeling techniques are used for RDBMS database design. 3. Performance is low for analysis queries. 4. Database is the place where the data is taken as a base and managed to get available fast and efficient access. 5. Optimized for write operation. 6. Used for Online Transaction Processing (OLTP) but can be used for other purpose such as data warehousing. This records the data from the user for history.	1. In data warehouse tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries. 2. Data modeling techniques are used for Data Warehouse design. 3. High performance for analytical queries 4. Data warehouse is the place where the application data is managed for analysis and reporting purpose. 5. Optimized for read operations. 6. Used for Online Analytical Processing (OLAP). This reads the historical data for the users for business decision.

Data Marts

Data mart is a database that contains a subset of data present in a data warehouse. We can divide a data warehouse into data marts after the data warehouse has been created. A **data mart** is a subset of a data warehouse in which a summarized or highly focused portion of the organization's data is placed in a separate database for a specific population of users. For example, a company might develop marketing and sales data marts to deal with customer information. A data mart typically focuses on a single subject area or line of business, so it usually can be constructed more rapidly and at lower cost than an enterprise-wide data warehouse.



Reasons for creating a data mart

- Creates collective view by a group of users
- Easy access to frequently needed data
- Ease of creation
- Improves end-user response time
- Lower cost than implementing a full data warehouse
- Potential users are more clearly defined than in a full data warehouse
- Contains only business essential data and is less cluttered

Tools for business Intelligence

Once data have been captured and organized in data warehouses and data marts, they are available for further analysis using tools for business intelligence. Business intelligence tools enable users to analyze data to see new patterns, relationships, and insights that are useful for guiding decision making. Principal tools for business intelligence include software for database querying and reporting, tools for multidimensional data analysis (online analytical processing), and tools for data mining.

Online analytical processing (OLAP): Multidimensional data analysis

OLAP supports multidimensional data analysis, enabling users to view the same data in different ways using multiple dimensions (data cube). Multidimensional data models are designed expressly to support data analyses. The goal of multidimensional data models is to support analysis in a simple and faster way by executives, managers and business professionals. These people are not interested in the overall architecture.

Suppose your company sells five different products—Laptops, Computers, TVs, Camera and Mobiles—in the East, West, North and Central regions. If you wanted to ask a fairly straightforward question, such as how many Computers were sold in the last week, you could easily find the answer by using sales database. But what if you wanted to know how many Computers sold in each of your sales regions and compare actual results with projected sales, then the querying becomes complicated. In such a case OLAP is used.

Each aspect of information—product, pricing, cost, region, or time period—represents a different dimension. So, a product manager could use a multidimensional data analysis tool to learn how many Computers were sold in the East reason in this week, how that compares with the previous week, and how it compares with the sales forecast. OLAP enables users to obtain online answers to ad hoc questions such as these in a fairly rapid amount of time, even when the data are stored in very large databases, such as sales figures for multiple years.

Time	Product	Location	Sales
2072-01-01	Computer	East region	30
2072-01-01	Laptop	West region	20
2072-01-01	Camera	Central region	50
2072-01-07	Mobile	East region	11

2072-01-07	TV	North region	23
2072-01-15	Computer	West region	54
2072-01-15	Laptop	Central region	09
2072-01-25	Laptop	East region	32
2072-01-25	TV	West region	19

Fig: Tabular representation

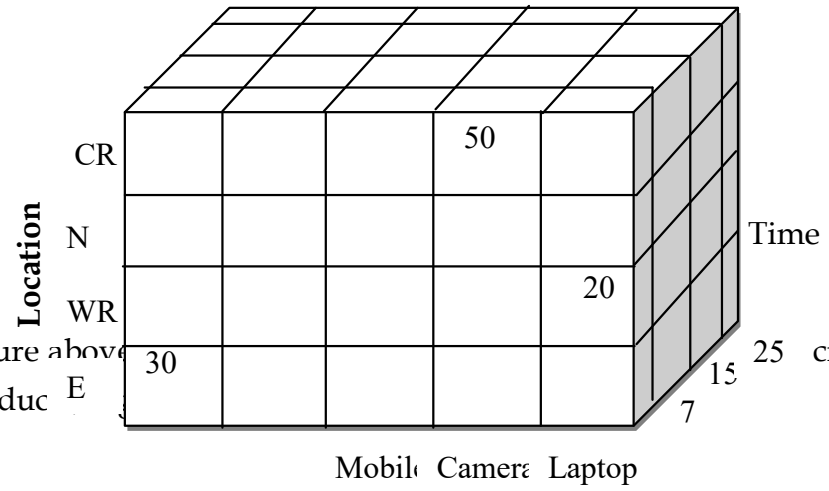


Figure above created to represent product

Data mining

Data mining refers to extracting or "mining" knowledge, interesting information or patterns from large amount of data. Data mining is a process of discovering interesting knowledge from large amounts of data stored either, in database, data warehouse, or other information repositories.

It is the semi-automatic process of extracting and identifying patterns from stored data. A data mining application, or data mining tool, is typically a software interface which interacts with a large database containing customer or other important data. Data mining is widely used by companies and public bodies for such uses as marketing, detection of fraudulent activity etc. That is, data mining deals with "knowledge discovery in databases. There are a wide variety of data mining applications available, particularly for business uses, such as Customer Relationship Management (CRM). These applications enable marketing managers to understand the behaviors of their customers and also to predict the potential behavior of prospective clients.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their business.

.

Functions of data mining

The types of information obtainable from data mining include associations, sequences, classifications, clusters, and forecasts.

- **Association:** Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. For instance, books that tends to be bought together. If a customer buys a book, an online bookstore may suggest other associated books. If a person buys a camera, the system may suggest accessories that tend to be bought along with cameras.
- **Prediction:** The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. For instance, when a person applies for a credit card, the credit-card company wants to predict if the person is a good credit risk. The prediction is to be based on known attributes of the person, such as age, income, debts, and past debt repayment history.
- **Classification:** Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.
- **Clustering:** Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts

objects in each class, while in the classification techniques, objects are assigned into predefined classes. For example in a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name.

Text mining

Text mining is the discovery of patterns and relationships from large sets of unstructured data—the kind of data we generate in e-mails, phone conversations, blog postings, online customer surveys, and tweets.

Web mining

The discovery and analysis of useful patterns and information from the World Wide Web or simply web is called web mining. Web mining is the application of data mining technique to find interesting and potentially useful knowledge from web data. So web mining is the application of data mining technique to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites etc.

Businesses might turn to Web mining to help them understand customer behavior, evaluate the effectiveness of a particular Web site, or quantify the success of a marketing campaign. For instance, marketers use Google Trends and Google Insights for Search services, which track the popularity of various words and phrases used in Google search queries, to learn what people are interested in and what they are interested in buying.