Chapter 3

DATABASE DESIGN PROCESS

3.1 OVERVIEW

Before we design a database for any organization, we have to consider many aspects to find out the practical scenario of owning a database. Few of them are discussed as follows :

Feasibility Study: This is also called preliminary investigation of the required database. It involves the area identification and selection i.e. which area or aspect is to be selected to start with. After the project is selected, it is allocated a specific fund and a proper planning is chalked out for its practical implementation. Side by side, a proper market analysis is also worked out.

Requirements Analysis: During this activity, the requirements are gathered i.e. the possible inputs for the database and the required functionality out of it. The users precisely narrate their needs of the database and the possible domain and restrictions are also chalked out.

Project Planning: A proper schedule is laid down to accomplish this activity. All the cost factors are taken into consideration i.e., the salaries of team members, their logistics involved, other trivial expenses (such as marriage gifts, insurances etc) and hardware costs.

Data Analysis: This is an important analysis aspect while designing a database. It involves the following activities :

(i) Data Flow Diagrams (DFD)
(ii) Decision Tables
(iii) Decision Trees

However, a detailed discussion on these topics is beyond the scope of this book.

3.2 DATA MODELING

Data Modeling is the process of identifying the data objects and the relationships between them.

Ingredients of Data Modeling:

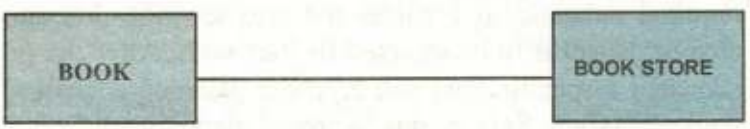
Entities/Objects : A data entity/object is anything that is participating in the system. It is always properly identifiable i.e., a TEACHER, a STUDENT, an AEROPLANE.

Attributes: *Attributes* define the objects, describe their characteristics and in some cases, make references to other objects(s) i.e., attributes for a TEACHER could be :

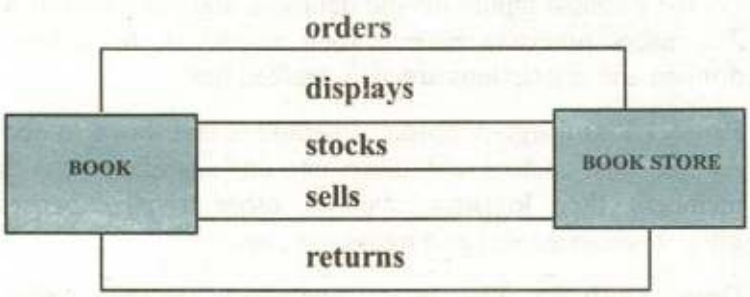
Teacher Name, Gender, Last Degree, Appointment Date, Pay Scale, Nationality, Telephone No. etc.

Relationships: The *relationship* indicates how the Entities/Objects are *Connected* or *Related* to each other.

The Data objects are related/connected to one another in different ways. Consider the data objects *BOOK* and *BOOK STORE* in the following diagrams.



(a) A basic connection between the objects



(b) Relationships between the objects

Following are the different possible and relevant relationships between them:

- A BOOK STORE **orders** BOOK(s).
- A BOOK STORE **displays** BOOK(s).
- A BOOK STORE **stocks** BOOK(s).
- A BOOK STORE **sells** BOOK(s).
- A BOOK STORE **returns** BOOK(s).

It is important to note that :

- All the relationships define the relevant *connections* between both objects.
- All the relationships are *bi-directional.*
- We have to consider only the relevant relationship (in the context of the requirement)

Cardinality:

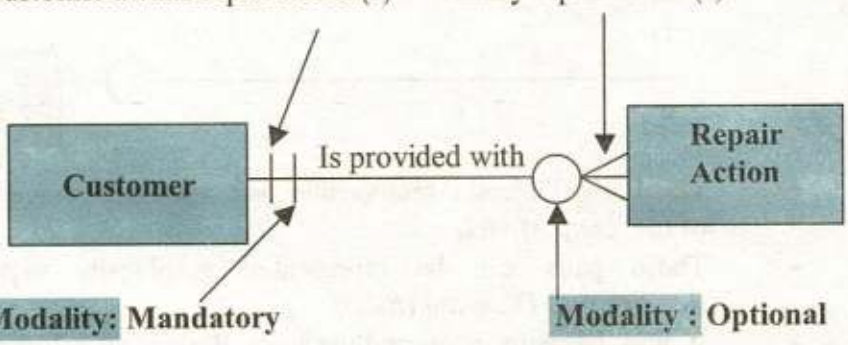
- Whether some occurrence(s) of *object*-1 are related to some occurrence(s) of *object*-2.
- It is expressed as *one* or *many* e.g.,
  - A husband can have only *one* wife and
  - A Father can have *many* children.
- The relationships can be
  - One to one
  - One to many
  - Many to many
  - Recursive
  - None

Modality:

- It defines the nature of the relationship i.e.,
  - **Optional** ..... represented by 0
  - **Mandatory** ..... represented by 1
- Consider two objects *Customer* and *Repair Action* in a Workshop environment.

Cardinality
Implies that only one
Customer awaits repair action(s)

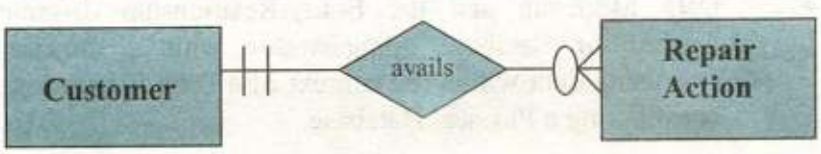
Cardinality
Implies that there may be
many repair action(s)



Implies that in order to have a repair action(s), we must have a customer

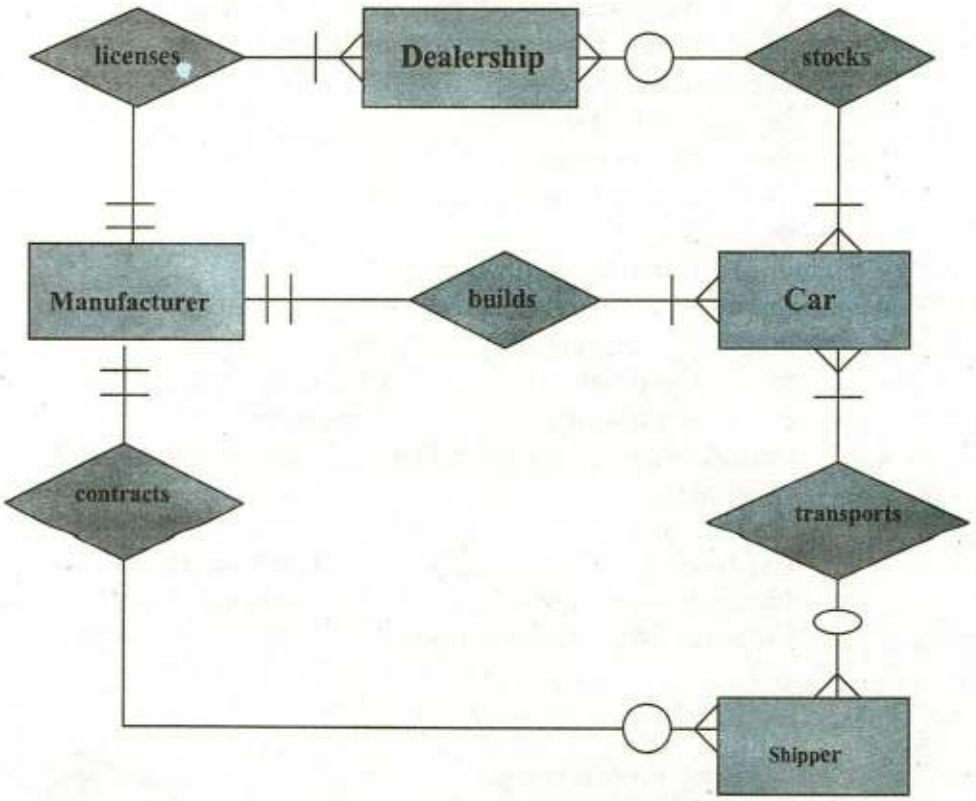
Implies that there may be a situation in which a repair action is not necessary

- A simple *Data Model* can be drawn from the above as follows:



- By connecting all the *Data Objects* alongwith their *Relationships* in the above manner, an *ERD* (Entity Relationship Diagram) is constructed.

Entity-Relation Diagram (An Example)

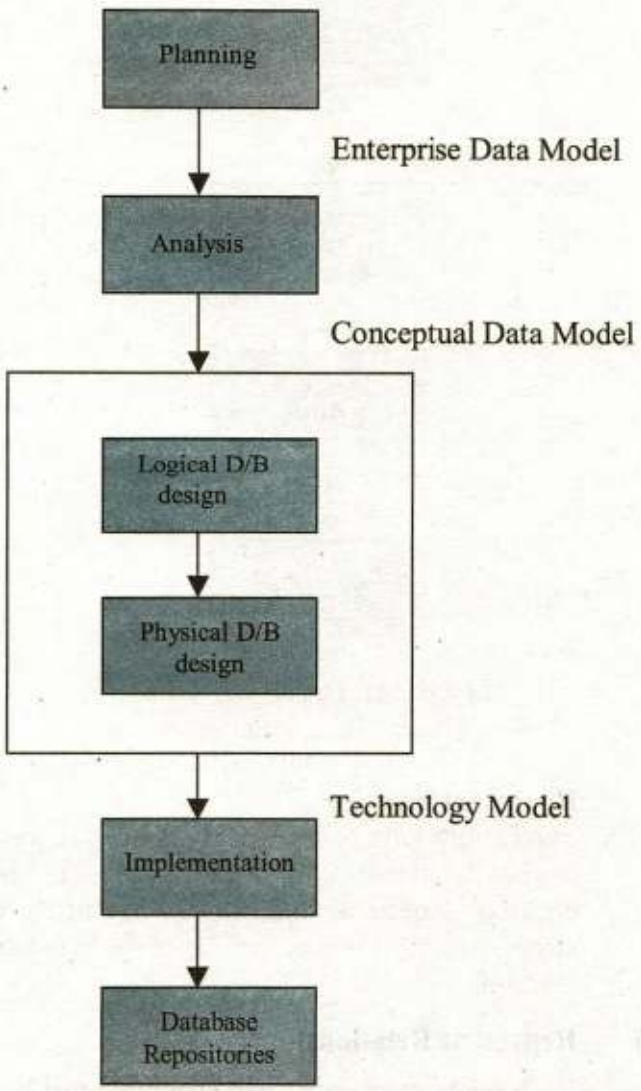


- The Entity/Object – relationship pair discussed above is the objective of the *Data Model*.
- These pairs can be represented graphically using the *Entity-Relationship Diagram* (*ERD*).
- It was basically proposed/used for design of a Relational Database System and now is being adopted for other Database types also.
- A set of primary components are identified for the *ERD*: Data objects, Attributes, Relationships, Cardinality and Modality.
- The primary objective of the *ERD* is to represent *Entities/Objects* and their *relationships / association*.
- Data Modeling and the Entity-Relationship diagram provide the Analyst or database administrator with a concise notation for examining data within the context of a Data Processing Application or constructing a Physical Database.

3.3 DATABASE DESIGN

The major objective of Database design is to map the conceptual data model to an implementation model that a particular DBMS can process with performance that is acceptable to all users throughout the organization. In today's compatible economy, database users require information that is complete and up-to-date and they expect to be able to access this information quickly and easily.

Following is the database development process.



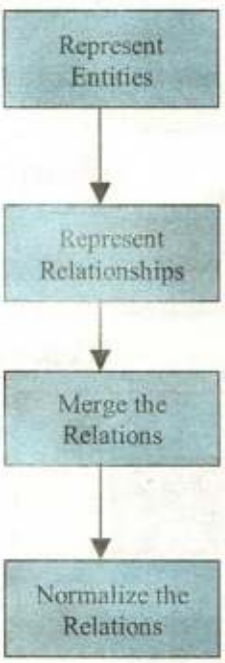
Database design can be divided into the following two phases:

3.3.1 Conceptual (Logical) Database Design:

The process of mapping the conceptual data models (from analysis) to structures that are specific to the target DBMS. If the target environment is a relational DBMS, then the conceptual data models are mapped to normalized relations.

Following diagram presents an overview of logical design process.

Conceptual Data Model
(E-R Diagram)



LOGICAL DATA MODEL
(Normalized Relations)

(i) **Represent Entities:**

Each entity type in the E-R diagram is represented as a relation in the Relational View or Data Model. The identifier of the entity type becomes the Primary key of the relation, and other attributes of the entity type become non-key attributes of the relation.

(ii) **Represent Relationships:**

Each relationship in an E-R diagram must be represented in the relational model. It depends on its nature. For example, in some cases, we represent a relationship by making the primary

key of one relation a foreign key of another relation. In other cases, we create a separate relation to represent a relationship.

(iii) Merge the Relations:

In some cases, there may be redundant relations (that is, two or more relations that describe the same entity type). They must be merged to remove the redundancy. This process is also known as View Integration. Suppose we have one relation as:
EMPLOYEE1(EMPNO , NAME,ADDRESS,PHONE)

And another relation as:
EMPLOYEE2(EMPNO , ENAME, EMP-ADDR, EMP-JOB-CODE, EMP-DOB)

Since the two relations have the same primary key (EMPNO) they describe the same entity and may be merged into one relation. The result of merging the relations is the following relation.

EMPLOYEE(EMPNO , NAME,ADDRESS,PHONE,EMP-JOB-CODE,EMP-DOB)

(iv) Normalize the Relations:

The relations that are created in step (i) and (ii) may have unnecessary redundancy and may be subject to anomalies (or errors) when they are updated. Normalization is the process that refines the relations to avoid these problems. (A detailed discussion is given in chapter 4)

3.3.2 Physical Database Design

It is the last stage of the database design process. The major objective of physical database design is to implement the database as a set of stored records, files, indexes and other data structures that will provide adequate performance and ensure database integrity, security and recoverability.

There are three major inputs to Physical database design.

(i) **Logical database structures** (developed during logical database design) i.e., the Normalized Relations.

(ii) **User processing requirements** i.e. size and frequency of use of the data-Base , response time, security, backup, recovery etc.

(iii) **Characteristics** of the DBMS and other components of the computer Operating environment.

Components of Physical Database Design:

(i) **Data Volume and Usage Analysis**:

To estimate the size or volume and the usage patterns of the database. Estimates of database size are used to select Physical storage devices and estimate the costs of storage. Estimates of usage paths or patterns are used to select the file organization and access methods, to plan for the use of indexes and to plan a strategy for data distribution.

(ii) **Data Distribution Strategy**:

Many organizations today have distributed computing networks. For these organizations, a significant problem in physical database design is deciding at which nodes (or sites) in the network to physically locate the data.

Basic data Distribution Strategies.

a. **Centralized:** All data are located at a single site. It is fairly easy to do but it has at least three disadvantages.
   - data are not readily accessible at remote sites. Data communication costs may be high.
   - The database system fails totally when the central system fails.

b. **Partitioned:** The database is divided into partitions (fragments). Each partition is assigned to a particular site. Major advantage of this is that data is moved closer to local users and so is more access-able.

c. **Replicated:** Full copy of database is assigned to more than one site in the network. This approach maximizes local access but creates update problems, since each database change must be reliably processed and synchronized at all of the sites.

d. **Hybrid:** In this strategy, the database is partitioned into critical and non-critical fragments. Non-critical fragments are stored at only one site, while critical fragments are stored at multiple sites.

(iii) **File Organization:**

A technique for physically arranging the records of a file on secondary storage devices. For selecting a file organization, the system designer must recognize several constraints, including the physical characteristics of the secondary storage devices, available operating systems and file management software, and user needs for storing and accessing data. Following is the criteria for selecting file organizations.

- Fast access for retrieval.

- High throughput for processing transactions.
- Efficient use of storage space.
- Protection from failure or data loss.
- Minimizing need for re-organization.
- Accommodating growth.
- Security from un-authorized use.

(iv) Indexes:

An index is a table that is used to determine the location of rows in a table (or tables) that satisfy some condition. They may be created on primary key, secondary key, foreign key etc.

(v) Integrity Constraints:

Database integrity refers to the correctness and consistency of data. It is another form of database protection. While it is related to security and precision, it has some broader implications. Security involves protecting the data from unauthorized operations, while integrity is concerned with the quality of data itself. Integrity is usually expressed in terms of certain constraints which are the consistency rules that the database is not permitted to violate. A few of them are discussed in chapter 4.

3.4 IMPLEMENTATION

In database implementation phase, the builder or the database administrator normally requires a server computer which will be linked with hundreds and thousands of computer users who would want to share and interact with the server (database).

For this purpose, the dba might need the services of network administrators to connect the users with the server. The users are normally given the authorizations / permissions defined by their respective managers so that they can perform the authorized tasks while using the database facilities.

In distributed computing environment, the database servers and users might be thousands of kilometers apart, so a lot of expensive telecommunication links are required to perform the designated tasks. NADRA and CRICKINFO are some of the typical examples of this type of databases.

Exercise 3c

1. Fill in the blanks:

(i) During _____ phase, the project requirements are gathered and identified.

(ii) DFD stands for _____.

(iii) The process of identifying data objects and relationship between them is called _____.

(iv) The number of occurrences of participating entities in a relationship is determined by the _____ ratio.

(v) Modality determines whether the participation of an entity in a relationship is _____ or optional.

(vi) ERD stands for _____.

(vii) In ERD model, a(n)_____ is represented by a rectangular box.

(viii) In _____ database systems, all the data is stored at a single site.

(ix) In _____ database multiple copies of the same data are stored at different sites on the network.

(x) In distributed databases, the data is _____ among various sites.

2. Select the correct option:

(i) Which of the following keys does not hold uniqueness property:

a) candidate key b) foreign key

c) primary key d) secondary key

(ii) An entity related to itself in an ERD model refers to:

a) recursive relationship b) one-to-many relationship

c) many-to-many relationship d) one-to-one relationship

(iii) Database development process involve mapping of conceptual data model into:

a) Object oriented data model b) Network data model

c) Implementation model d) Hierarchical data model

(iv) In ERD model, the relationship between two entities is represented by a:
a) diamond symbol b) rectangular box
c) oval symbol d) line

(v) In hybrid distribution which kind of fragments are stored at only one site:
a) critical fragments b) non-critical fragments
c) critical and non-critical fragments d) only large fragments

3. Write T for true and F for false statement.

(i) In one-to-one relationship only one instance of each entity can participate in the relationship.

(ii) The optional modality is represented by 1.

(iii) One-to-many is a uni-directional relationship.

(iv) In ERD model, a condition is mentioned in a diamond symbol.

(v) ERD is a physical data model.

(vi) In hybrid distribution the database is partitioned in critical and non-critical fragments.

(vii) In distributed databases, the consistency refers to availability of same data at all sites of the network.

(viii) Indexing maximizes the time required to search a piece of information from a database.

(ix) Analysis is less important activity than coding, so minimum time should be spent over analyzing the system.

(x) Relationship defines the logical connection between entities.

4. Describe different steps involved in analysis stage while designing a database.

5. Explain the following with the help of figures:
(i) Entity/Object
(ii) Attribute
(iii) Relationship
(iv) Cardinality
(v) Modality

6. Draw and explain ER diagram for the system of getting admission in your college.

7. Explain the following:
   (i) Physical data model
   (ii) Conceptual data model

8. What are the components of a logical data model?

9. What elements combined , produce the physical database design? Explain

10. Define and explain the following terms:
    (i) Data distribution strategy
    (ii) File Organization

11. Define the term Analysis. Briefly discuss the following terms:
    (i) Feasibility study
    (ii) Requirement Analysis
    (iii) Project Planning
    (iv) Data Analysis

12. Briefly explain the database design process with the help of a diagram.