

Database Engineering

Dr. Sambit Bakshi

NIT Rourkela

Course Overview

Subject Code: CS2008

Subject Title: Database Engineering (L-T-P:3-0-0, Credit: 3)

Evaluation:

- **Mid term examination:** 30
- **End term examination:** 50
- **Teacher's assessment:** 20 (Class Test-I: 05, Class Test-II: 05, Assignment: 10)

Course Overview

Syllabus:

- **Introduction to database systems:** Data Independence, Data Models, Levels of abstraction, structure of DBMS
- **Data models and query languages:** Relational Model, Relational Languages, Query Languages: Relational Algebra, Relational Calculus, SQL, QUEL, QBE, Integrity constraints, Aggregate operators, Embedded and Dynamic SQL.
- **Database design:** E-R Model, Functional dependencies, Decomposition, Normalization, Multivalued dependencies.
- **Concurrency control and recovery:** Transaction, Schedules, Lock based concurrency, Lock management, Concurrency control without locking, Crash recovery- log, check pointing, media recoveries.
- **File Organization:** Storage, Buffer management, Disk Management, File organization techniques, Indexing.
- **Query Processing and Optimization:** Query processing on various operations, Translating SQL queries, Estimating the cost.
- **Advanced topics:** Database Security, Distributed databases design, Object Oriented database design & its implementation, Introduction to recent advances in database technology.

Reading References

Essential reading:

- J.D. Ullman, *Principles of Database Systems*, Computer Science Press, 1982.
- A. Silberschatz, H.F. Korth, and A. Sudarshan, *Database System Concepts*, McGraw Hill, 5th ed, 2006.

Supplementary Reading:

- B. Desai, *An Introduction to Database System*, Galgotia, 1997.
- C.J. Date: *An Introduction to Database Systems*, Addison Wesley, 1995.
- R. Elmasri, S. Navathe, S.B. Navathe, and R. Sunderraman, *Fundamentals of Database Systems*, Addison Wesley, 1994.
- R.R. Krishnan, *Database Management Systems*, McGraw Hill, 2007.

Outline

- 1 Introduction
- 2 Database Application
- 3 Drawbacks of Conventional File-Processing System
- 4 Characteristics of Database Approach
- 5 Advantage and Disadvantage of DBMS
- 6 Classification of DBMS Users
- 7 DBMS Facilities
- 8 Data Abstraction and its Levels
- 9 System Architecture
- 10 Two and Three-Tier Architecture

Introduction

- **Data:** Known facts that can be recorded
- **Database:** Collection of interrelated and persistent data
- **Database Management System (DBMS):** Collection of a set of programs to define, store, manipulate, control, and retrieve data
- **Objective of DBMS:** To provide a **convenient** and **effective** method of operation on database
- **Database System (DBS):** Consolidation of database and DBMS

Types of Databases

- **Traditional Database:** Data is text / number
(e.g. Name or DOB of a student)
- **Multimedia Database:** Data is image / video / speech
(e.g. Fingerprint data of students kept in NITR, Aadhaar database related to human biometrics, Youtube video database, GIS data from satellites)

What is data warehouse?

- A huge and special type of database that needs specialized techniques for retrieval
- May store temporal data spanning across a timeline

What is data mining?

- Process of discovering anomalies, patterns and correlations within large data sets involving methods at the intersection of machine learning, statistics, and database systems
- Two steps in 'knowledge discovery in databases' (KDD): (I) raw analysis and visualization to summarize the past data, (II) analysis to predict future anomalies, patterns and correlations (through data mining)

Database Applications

- Banking
- Airlines Reservation
- Credit Card Transaction
- Telecommunication
- Finance and Sales
- On-line Retailers
- Manufacturing and Human Resources
- Student / Employee Management Systems

Drawbacks of Conventional File-Processing System - I

- **Atomicity Problems:** Data might not be stored in atomic nature, creating trouble in fetching proper data while query is triggered (e.g. someone can mention country within address field, which cannot be retrieved later by querying).
- **Data Redundancy and Inconsistency:** Same data may be present in different files (causing more space cost), and may lead to inconsistent data when one instance of a data is modified, but its other instances in other files are not modified.
- **Integrity Problems:** Data stored may sometimes require a natural constraint (e.g. AGE cannot be a negative integer) or business-specific constraints (e.g., GRADE of a student can be E / A / B / C / D / F / I , and no other character) to be applied on them. Managing these integrity constraints on creation and during each modification of data is not possible through file-processing system.
- **Data Isolation:** Isolating a desired data from more than one file is challenging in file-processing system.

Drawbacks of Conventional File-Processing System - II

- **Difficulty in Accessing Data:** In file-processing system, users are not abstracted from the physical access of data. Hence users should know the location of files to retrieve data. In course of time, if file structure changes, the users face trouble.
- **Concurrent-access Anomalies:** In desired scenario of multiple users accessing/altering same data, file-processing system may fail. Locks cannot be created between user access causing the problem.
- **Security Problems:** All users are not designated to access all data. There is a logical grant / revoke of 'right of access' by business policy makers, which a database should adhere to. In file-processing system, there is no such scope.

Characteristics of Database Approach

- Self-describing nature of DB system
- Insulation between program, data, and data abstraction
- Support of multiple views of data
- Sharing of data and Multiple user Transaction processing

Advantage and Disadvantage of DBMS

Advantage

- Data duplication eliminated
- Controlled redundancy
- Restricting unauthorized access
- Providing backup and recovery
- Providing Multiple User Interface
- Enforcing Security and integrity constraints
- Centralized Control and Data quality enhanced

Disadvantage

- Problems associated with centralization
- Cost of software/hardware and migration
- Complexity of backup and recovery.

Classification of DBMS Users

- **Naive Users:** Users who need not be aware of the presence of the database system or any other system supporting their usage.
- **Online Users:** These are users who may communicate with the database directly via an online terminal or indirectly via a user interface and application program.
- **Application Programmers:** Professional programmers who are responsible for developing application program or user interface utilized by the naive and online users.
- **Database Administrator (DBA):** Centralized control of the database is exerted by a person or group of persons. This person or group is referred to as the Database Administrator (DBA). They are responsible for creating, modifying, and maintaining its three levels.

DBMS Facilities

DDL (Data Definition Language)

- Used for defining the conceptual scheme
- Used to specify additional properties of the data
- provides facilities to specify constraints

DML (Data Manipulation Language)

- Enables users to access or manipulate data as organized by the appropriate data model.
- Types of Access are:
 - Retrieval of information
 - Insertion of new information
 - Deletion of information
 - Modification of Information

Apart from these, there exists **DCL (Data Control Language)** and **TCL (Transaction Control Language)**

Data Abstraction and its levels

It generally refers to the suppression of details of data organization and storage, and the highlighting of the essential features.

- **External Schema:**

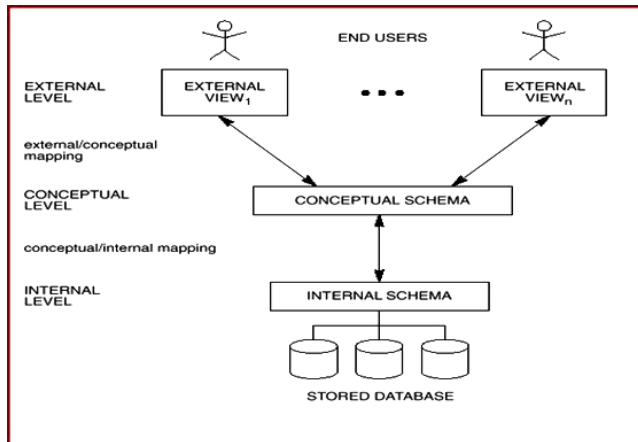
- Allow data access to be customized at the level of individual users or groups of users
- Each external schema consists of a collection of one or more views and relations from the conceptual schema.
- Any given database has exactly one conceptual schema and one physical, but it may have several external schema

- **Conceptual Schema:** Also called as Logical Schema. It describes data in terms of data model

- **Physical / Internal Schema:** describes how a record is stored. Specified additional storage details

The model was proposed by ANSI-SPARC (American National Standards Institute, Standards Planning And Requirements Committee) in 1975

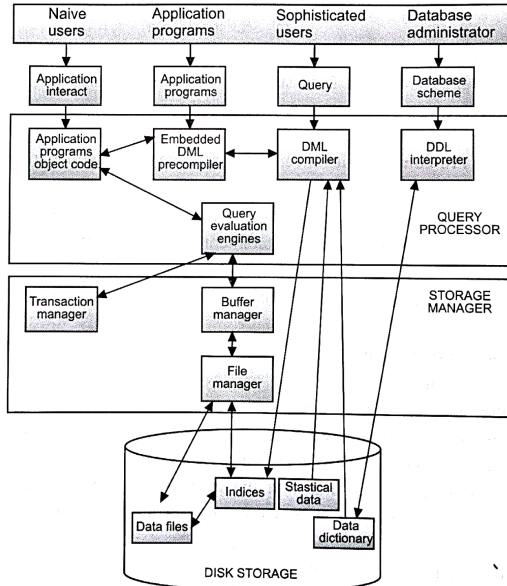
Data Abstraction and its levels



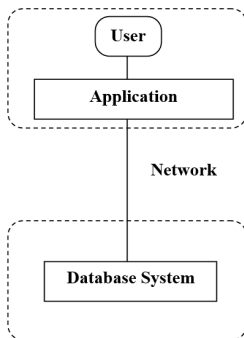
Structure of a DBMS

- **DDL Compiler:** It converts the data definition statements into a set of tables.
- **Data Manager:** It is central software component of DBMS. Its function is to convert operation in the user's queries coming directly via the query processor.
- **File Manager:** Responsible for the structure of files and managing the file space rests with file manager.
- **Disk Manager:** It is part of OS of the host computer and physical input and output operation. It transfers the block or page requested by the file manager.
- **Query Processor:** It is used to interpret the online user's query and convert it into an efficient series of operation in a form capable of being sent to data manager.

System Architecture



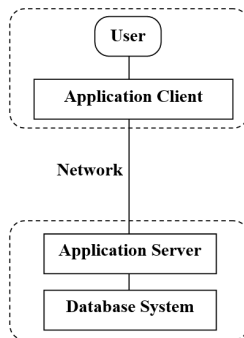
Two and Three-Tier Architecture



(a) Two-tier Architecture

client

Server



(a) Three-tier Architecture