Chapter 1: Introduction

Slides by Silberschatz

Content and layout modified by Mike Rogers

Database Management System (DBMS)

- DBMS contains information about a particular enterprise
  - Collection of interrelated data
  - Set of programs to access the data
  - An environment that is both convenient and efficient to use

Database Management System (DBMS)

- Database Applications:
  - Banking: transactions
  - Airlines: reservations, schedules
  - Universities: registration, grades
  - Sales: customers, products, purchases
  - Online retailers: order tracking, customized recommendations
  - Manufacturing: production, inventory, orders, supply chain
  - Human resources: employee records, salaries, tax deductions
- Databases touch all aspects of our lives

University Database Example

- Application program examples
  - Add new students, instructors, and courses
  - Register students for courses, and generate class rosters
  - Assign grades to students, compute grade point averages (GPA) and generate transcripts
- In the early days, database applications were built directly on top of file systems

Purpose of Database Systems

- Drawbacks of using file systems to store data:
  - Data redundancy and inconsistency
    - Multiple file formats, duplication of information in different files
  - Difficulty in accessing data
    - Need to write a new program to carry out each new task
  - Data isolation — multiple files and formats
  - Integrity problems
    - Integrity constraints (e.g., account balance > 0) become "buried" in program code rather than being stated explicitly
    - Hard to add new constraints or change existing ones

Purpose of Database Systems (Cont.)

- Drawbacks of using file systems (cont.)
  - Atomicity of updates
    - Failures may leave database in an inconsistent state with partial updates carried out
    - Example: Transfer of funds from one account to another should either complete or not happen at all
  - Concurrent access by multiple users
    - Concurrent access needed for performance
    - Uncontrolled concurrent accesses can lead to inconsistencies
      - Example: Two people reading a balance (say 100) and updating it by withdrawing money (say 50 each) at the same time

Purpose of Database Systems (Cont.)

- Drawbacks of using file systems (cont.)
  - Security problems
    - Hard to provide user access to some, but not all, data
- Database systems offer solutions to all the above problems

Levels of Abstraction

- Physical level: describes how a record (e.g., customer) is stored.
- Logical level: describes data stored in database, and the relationships among the data.

```
type instructor = record

    ID : string;
    name : string;
    dept_name : string;
    salary : integer;

end;
```

- View level: application programs hide details of data types. Views can also hide information (such as an employee's salary) for security purposes.

Picture Description:

At the bottom is the physical level.

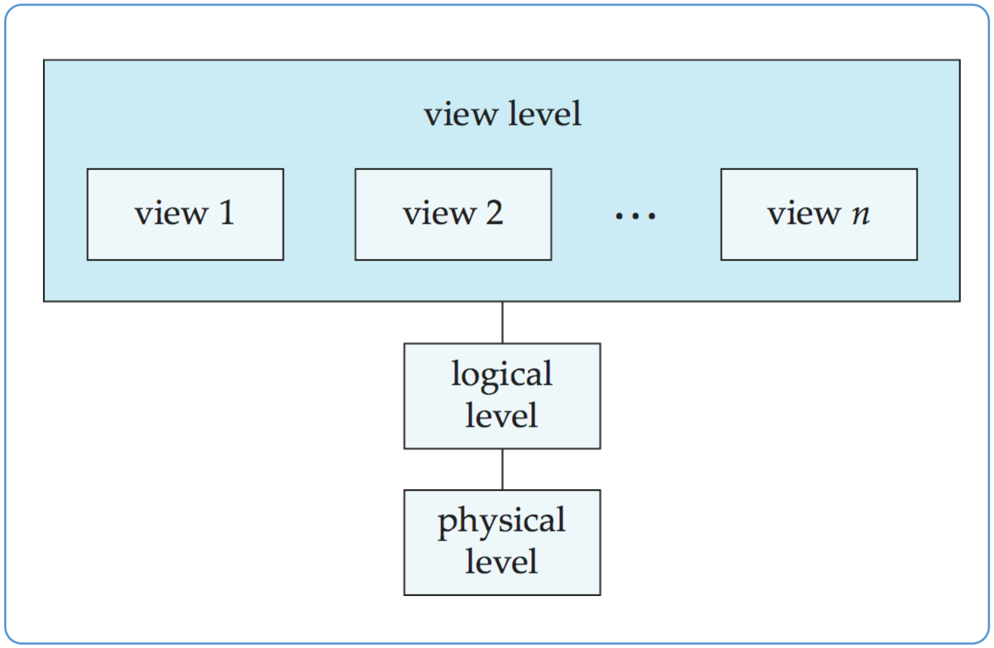
Above the physical level is the logical level.

Both of these layers have but one representation.

At the topmost level is the view level.

The view level has multiple representations, where each representation is for a particular type of user.

An architecture for a database system



Instances and Schemas

- Similar to types and variables in programming languages
- Schema – the logical structure of the database
  - Example: The database consists of information about a set of customers and accounts and the relationship between them
  - Analogous to type information of a variable in a program
  - Physical schema: database design at the physical level
  - Logical schema: database design at the logical level
- Instance – the actual content of the database at a particular point in time
  - Analogous to the value of a variable

Instances and Schemas

- Physical Data Independence – the ability to modify the physical schema without changing the logical schema
  - Applications depend on the logical schema
  - In general, the interfaces between the various levels and components should be well defined so that changes in some parts do not seriously influence others.

Data Models

- A collection of tools for describing
  - Data
  - Data relationships
  - Data semantics
  - Data constraints
- Relational model
- Entity-Relationship data model (mainly for database design)
- Object-based data models (Object-oriented and Object-relational)
- Semistructured data model (XML)
- Other older models:
  - Network model
  - Hierarchical model

Relational Model

- Details are in Chapter 2
- Example of tabular data in the relational model

ID, Name, etc. are columns.

ID	name	dept_name	salary
2222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Computer Science	75000

The table is made up of multiple Rows.

A Sample Relational Database

The *instructor* table:

ID	name	dept_name	salary
2222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
55555	Olds	History	62000
45565	Katz	Computer Science	75000

The *department* table:

dept_name	building	budget
Computer Science	Taylor	100000
Biology	Watson	90000
Finance	Painter	120000
History	Painter	50000
Physics	Watson	85000

Data Manipulation Language (DML)

- Language for accessing and manipulating the data organized by the appropriate data model
  - DML also known as query language
- Two typical classes of languages
  - Procedural – user specifies what data is required and how to get those data
  - Declarative (nonprocedural) – user specifies what data is required without specifying how to get those data
- SQL is the most widely used query language

Data Definition Language (DDL)

- Specification notation for defining the database schema
  - Example:

```
create table instructor (
    ID              char(5),
    name            varchar(20),
    dept_name       varchar(20),
    salary          numeric(8,2))
```

- DDL compiler generates a set of tables stored in a data dictionary
- Data dictionary contains metadata (i.e., data about data)
  - Database schema
  - Integrity constraints
    - Primary key (ID uniquely identifies instructors)
    - Referential integrity (references constraint in SQL)
      - e.g. dept_name value in any instructor tuple must appear in department relation
  - Authorization

SQL

- SQL: widely used non-procedural language
  - Example: Find the name of the instructor with ID equals 22222

```
select   name
 from   instructor
 where  instructor.ID = '22222'


select instructor.ID, department.dept_name
 from instructor, department
 where instructor.dept_name= department.dept_name and
       department.budget > 95000
```

- Application programs generally access databases through one of
  - Language extensions to allow embedded SQL
  - Application program interface (e.g., ODBC/JDBC) which allow SQL queries to be sent to a database
- Chapters 3, 4 and 5

Database Design

- The process of designing the general structure of the database:

- Logical Design – Deciding on the database schema. Database design requires that we find a "good" collection of relation schemas.
  - Business decision – What attributes should we record in the database?
  - Computer Science decision – What relation schemas should we have and how should the attributes be distributed among the various relation schemas?

- Physical Design – Deciding on the physical layout of the database

Database Design?

- Is there any problem with this design?

ID	name	salary	dept_name	building	budget
2222	Einstein	95000	Physics	Watson	85000
12121	Wu	90000	Finance	Painter	120000
32343	El Said	60000	History	Painter	50000
55555	Olds	62000	History	Painter	50000
45565	Katz	75000	Computer Science	Taylor	100000

Where is Biology?

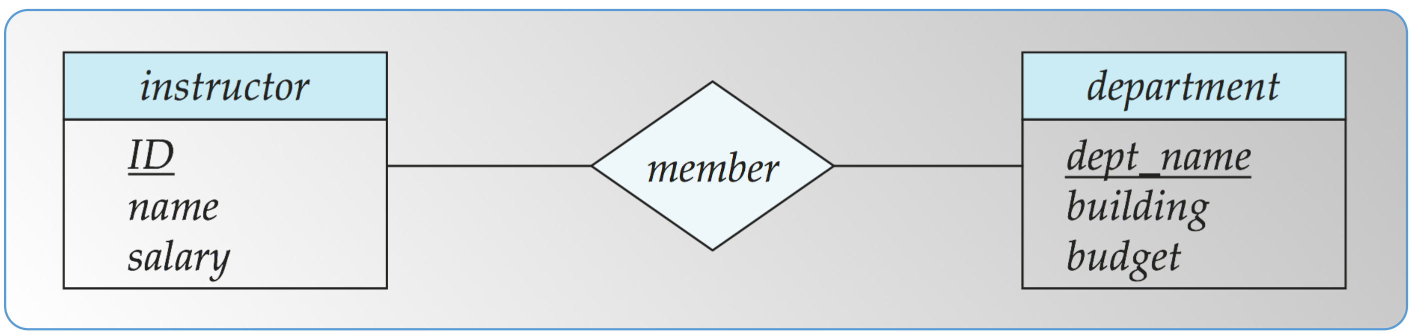
How many times is the info for the history department repeated?

Design Approaches

- Normalization Theory (covered in Chapter 8)
  - Formalize what designs are bad, and test for them
- Entity Relationship Model (covered in Chapter 7)
  - Models an enterprise as a collection of entities and relationships
    - Entity: a "thing" or "object" in the enterprise that is distinguishable from other objects
      - Described by a set of attributes
    - Relationship: an association among several entities
  - Represented diagrammatically by an entity-relationship diagram:

The Entity-Relationship Model

- Models an enterprise as a collection of entities and relationships
  - Entity: a "thing" or "object" in the enterprise that is distinguishable from other objects
    - Described by a set of attributes
  - Relationship: an association among several entities
- Represented diagrammatically by an entity-relationship diagram:



What happened to dept_name of instructor and student?

Object-Relational Data Models

Probably Won't Include Object-Relational Chapter in this Course

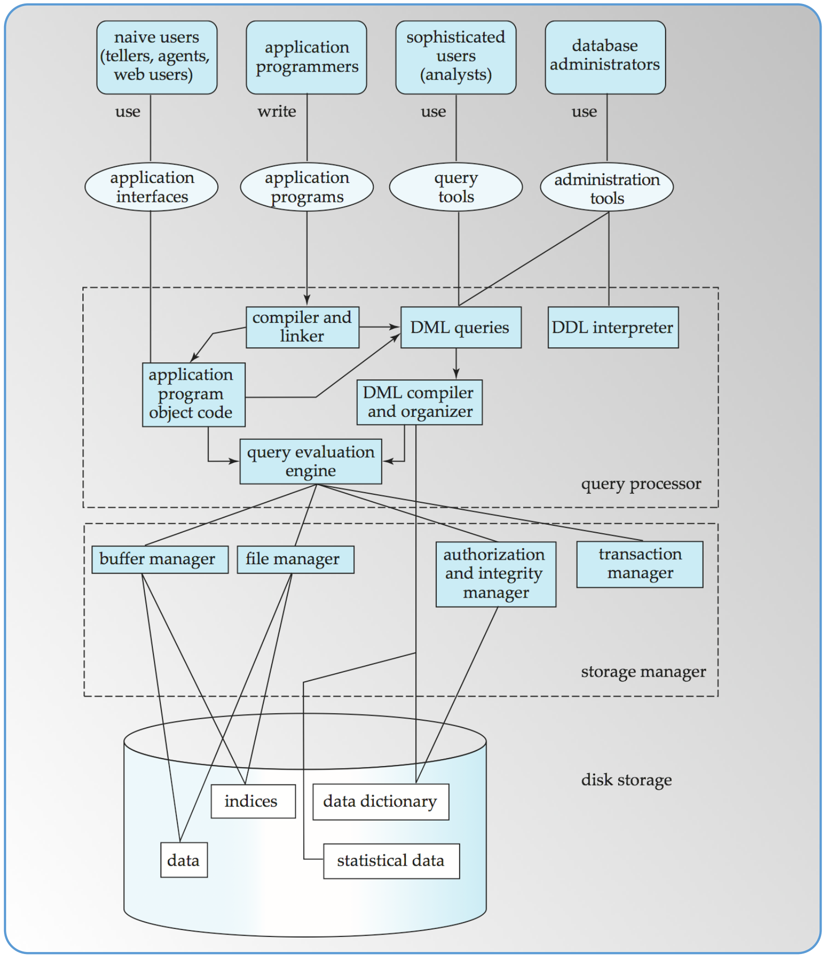
- Relational model: flat, "atomic" values
- Object Relational Data Models
  - Extend the relational data model by including object orientation and constructs to deal with added data types.
  - Allow attributes of tuples to have complex types, including non-atomic values such as nested relations.
  - Preserve relational foundations, in particular the declarative access to data, while extending modeling power.
  - Provide upward compatibility with existing relational languages.

XML: Extensible Markup Language

Probably Won't Include XML Chapter in this Course

- Defined by the WWW Consortium (W3C)
- Originally intended as a document markup language not a database language
- The ability to specify new tags, and to create nested tag structures made XML a great way to exchange data, not just documents
- XML has become the basis for all new generation data interchange formats.
- A wide variety of tools is available for parsing, browsing and querying XML documents/data

Database System Internals

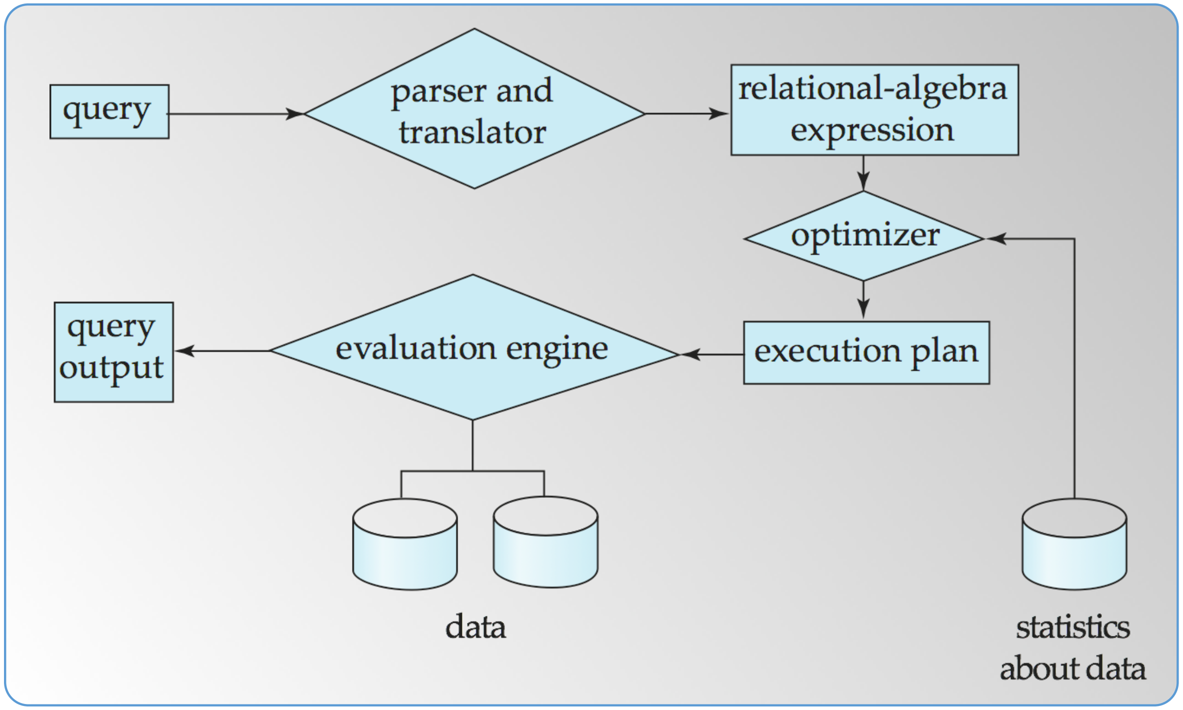


Storage Management

- Storage manager is a program module that provides the interface between the low-level data stored in the database and the application programs and queries submitted to the system.
- The storage manager is responsible to the following tasks:
  - Interaction with the file manager
  - Efficient storing, retrieving and updating of data
- Issues:
  - Storage access
  - File organization
  - Indexing and hashing

Query Processing

1. Parsing and translation
2. Optimization
3. Evaluation



Query Processing (Cont.)

- Alternative ways of evaluating a given query
  - Equivalent expressions
  - Different algorithms for each operation
- Cost difference between a good and a bad way of evaluating a query can be enormous
- Need to estimate the cost of operations
  - Depends critically on statistical information about relations which the database must maintain
  - Need to estimate statistics for intermediate results to compute cost of complex expressions

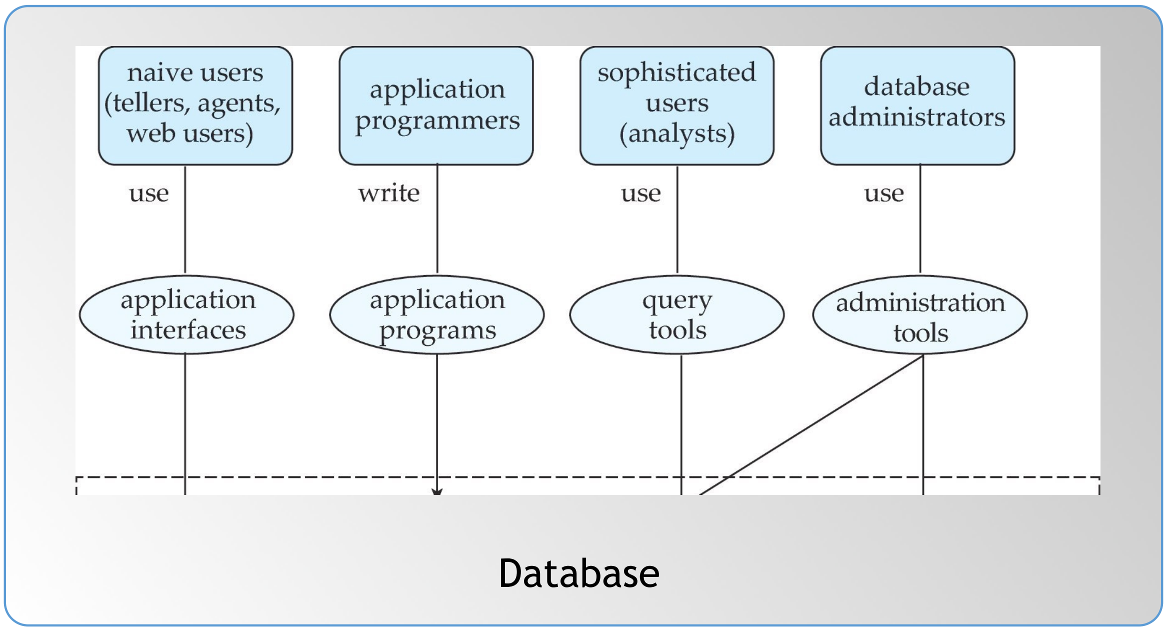
Transaction Management

- What if the system fails?
- What if more than one user is concurrently updating the same data?
- A transaction is a collection of operations that performs a single logical function in a database application
- Transaction-management component ensures that the database remains in a consistent (correct) state despite system failures (e.g., power failures and operating system crashes) and transaction failures.
- Concurrency-control manager controls the interaction among the concurrent transactions, to ensure the consistency of the database.

Database Architecture

- The architecture of a database systems is greatly influenced by the underlying computer system on which the database is running:
- Centralized
- Client-server
- Parallel (multi-processor)
- Distributed

Database Users and Administrators



History of Database Systems

- 1950s and early 1960s:
  - Data processing using magnetic tapes for storage
    - Tapes provided only sequential access
  - Punched cards for input
- Late 1960s and 1970s:
  - Hard disks allowed direct access to data
  - Network and hierarchical data models in widespread use
  - Ted Codd defines the relational data model
    - Would win the ACM Turing Award for this work
    - IBM Research begins System R prototype
    - UC Berkeley begins Ingres prototype
  - High-performance (for the era) transaction processing

History (cont.)

- 1980s:
  - Research relational prototypes evolve into commercial systems
    - SQL becomes industrial standard
  - Parallel and distributed database systems
  - Object-oriented database systems
- 1990s:
  - Large decision support and data-mining applications
  - Large multi-terabyte data warehouses
  - Emergence of Web commerce

History (cont.)

- Early 2000s:
  - XML and XQuery standards
  - Automated database administration
- Later 2000s:
  - Giant data storage systems
    - Google BigTable, Yahoo PNuts, Amazon, ..