Why use a DBMS: Data Independence and Abstraction

- The major problem with developing applications based on files is that the application is dependent on the file structure.
- There is no program-data independence separating the application from the data it is manipulating. If the data file changes, the code that accesses the file may require changes to the application.
- One of the major advantages of databases is they provide data abstraction. Data abstraction allows the internal definition of an object to change without affecting programs that use the object through an external definition.

Course Outline: Topics

- **Part 1**: Design of Relational Databases: Weeks 1-10
  - Entity-Relationship Model (similar to UML)
  - Formal Query Languages: Rel. algebra
  - Query languages: SQL
  - Relational Schema Design and Normal Forms, Tuning
  - Overview of DBMS architecture
  - File manager, transaction processing, query processing
    - **Team based term Project**
- **Part 2**: Intro. to Databases and Analytics for Semi/Un-structured data: Weeks 11-14
  - NoSQL database models
  - Experience working with one NoSQL DB
  - Writing requirements: discussion of papers, term project report
Basic Definitions

- **Database:**
  - A collection of related data.
- **Data:**
  - Known facts that can be recorded and have an implicit meaning.
- **Mini-world:**
  - Some part of the real world about which data is stored in a database.
  - For example, student grades and transcripts at a university.
- **Database Management System (DBMS):**
  - A software package/system to facilitate the creation and maintenance of a computerized database.
- **Database System:**
  - The DBMS software together with the data itself. Sometimes, the applications are also included.

Recent Developments

- Social Networks started capturing a lot of information about people and their communications (tweets, photos, videos...)
- Facebook, Twitter, LinkedIn, ...
- All of the above constitutes data
- Search Engines: Google, Bing, Yahoo: collect their own repository of web pages for searching purposes
- New Technologies emerging to manage vast amounts of data generated on the web:
  - Big Data storage systems involving large clusters
  - NOSQL (Not Only SQL) systems
  - A large amount of data now resides on the "cloud" which means it is in huge data centers using thousands of machines.

How is your database system architected: Three-Tier Client-Server Architecture

- **Tier 1: Client (Web/mobile)**
  - User Interface (using HTML/CSS)
- **Tier 2: Application Server**
  - Business logic – written using PHP
  - Data processing logic
- **Tier 3: Database Server**
  - Data storage/management
  - Using MySQL

Data Models

- **Data Model:**
  - A set of concepts to describe the **structure** of a database,
  - the **operations** for manipulating these structures, and
  - certain **constraints** that the database should obey.

- **Data Model Structure:**
  - Constructs are used to define the database structure
  - Data types, format (table),...ex: name is a char string

- **Constraints:**
  - Constraints specify some restrictions on valid data; these constraints must be enforced at all times..ex: GWID is unique

- **Data Model Operations:**
  - These operations are used for specifying database **retrievals and updates** by referring to the constructs of the data model.
  - Ex: find student name with GWID= abcd
Categories of Data Models

- **Conceptual (high-level, semantic) data models:**
  - Provide concepts that are close to the way many users perceive data.

- **Physical (low-level, internal) data models:**
  - Provide concepts that describe details of how data is stored in the computer. These are usually specified in an ad-hoc manner through DBMS design and administration manuals.

- **Implementation (representational)/Logical data models:**
  - Provide concepts that fall between the above two, used by many commercial DBMS implementations (e.g., relational data models).

- **Self-Describing Data Models:**
  - Combine the description of data with the data values. Examples include XML, key-value stores and some NOSQL systems.

---

**Conceptual Data Model: The Entity-Relationship (ER) Model**

- Provide database design that is easy to interpret by a wide class of users.
  - Not just database/CS experts.
  - You want to provide a design to a “client” using a representation they can understand.
- What’s a natural way to provide an easy to interpret representation . . .
  - Fill in the blanks: A _______ is worth a thousand words.
- “Visual” representation of the data, how it interacts, constraints, etc.
  - And can be automatically mapped to a data model (relational).
- ER-Model is one such data model.
  - We will return to it after we cover the relational model.

---

**Example: ER Design for mini-banner:**

<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Contains information about students.</td>
</tr>
<tr>
<td>Course</td>
<td>Represents courses offered.</td>
</tr>
<tr>
<td>Professor</td>
<td>Faculty members who teach courses.</td>
</tr>
<tr>
<td>Takes</td>
<td>Relationship between students and courses.</td>
</tr>
<tr>
<td>Teaches</td>
<td>Relationship between professors and courses.</td>
</tr>
</tbody>
</table>

"Who’s taking what, and what grade do they expect?"

One picture provides info on what your system stores and models.

---

**Schemas versus Instances**

- **Database Schema:**
  - The description of a database.
  - Includes descriptions of the database structure, data types, and the constraints on the database.
- **Schema Diagram:**
  - An illustrative display of (most aspects of) a database schema.
- **Schema Construct:**
  - A component of the schema or an object within the schema, e.g., STUDENT, COURSE.
Database State/Instance

- Database State (also called instance):
  - The actual data (content) stored in a database at a particular moment in time. This includes the collection of all the data in the database.

- Initial Database State:
  - Refers to the database state when it is initially loaded into the system.

- Valid State:
  - A state that satisfies the structure and constraints of the database.

Database Schema vs. Database State

- Distinction
  - The database schema changes very infrequently. Preferably never.
  - The database state changes every time the database is updated.

Example of a Database Schema

- STUDENT
  - Name | Student_number | Class | Major

- COURSE
  - Course_name | Course_number | Credit_hours | Department

- PREREQUISITE
  - Course_number | Prerequisite_number

- SECTION
  - Section_identifier | Course_number | Semester | Year | Instructor

- GRADE_REPORT
  - Student_number | Section_identifier | Grade

Example Instance & Schema

<table>
<thead>
<tr>
<th>sid</th>
<th>name</th>
<th>sid</th>
<th>exp-grade</th>
<th>cid</th>
<th>cid</th>
<th>subj</th>
<th>sem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ross</td>
<td>1</td>
<td>A</td>
<td>550-0103</td>
<td>550-0103</td>
<td>DB</td>
<td>S13</td>
</tr>
<tr>
<td>2</td>
<td>Lee</td>
<td>1</td>
<td>A</td>
<td>700-1003</td>
<td>700-1003</td>
<td>AI</td>
<td>S13</td>
</tr>
<tr>
<td>3</td>
<td>Emily</td>
<td>3</td>
<td>C</td>
<td>500-0103</td>
<td>500-0103</td>
<td>Arch</td>
<td>F12</td>
</tr>
</tbody>
</table>

- PROFESSOR
  - fid | name   
  - fid | cid    

- Takes
  - Our focus now: relational schema – set of tables
- Can have other kinds of schemas – XML, object, ...

- COURSE
  - cid | subj | sem
  - fid | name   
  - fid | cid    

Figure 2.1 Schema diagram for the database in Figure 1.2.
History of Data Models

- Network Model
- Hierarchical Model
- Relational Model
- Object-oriented Data Models
  - Object-Relational Models
- NoSQL/ Big Data technologies
  - Schema-less designs

Next: Relational Model Concepts

- The relational Model of Data is based on the concept of a Relation
- The strength of the relational approach to data management comes from the formal foundation provided by the theory of relations
- start with review of essentials of the formal relational model
- In practice, there is a standard model based on SQL

Note: There are several important differences between the formal model and the practical model
- There are variations in SQL features provided on different DBMS systems
  - Oracle SQL, MySQL, MS-SQL,…

Relational Model Concepts

- Relation is a mathematical concept based on idea of sets
- The model was first proposed by Dr. E.F. Codd of IBM Research in 1970 in the following paper:
  - "A Relational Model for Large Shared Data Banks," Communications of the ACM, June 1970
  - Describes the data minimally and mathematically
    - A relation describes an association between data items – tuples with attributes
    - Uses standard mathematical (logical) operations over the data – relational algebra or relational calculus

- The above paper caused a major revolution in the field of database management and earned Dr. Codd the coveted ACM Turing Award
Relational Model

- Vendors: IBM, Microsoft, Oracle, MySQL etc.
- Shift from networked “records” of data to a set of tables
  - Why is this a good thing for SW developers?

Why Did It Take So Many Years to Implement Relational Databases?

- Codd’s original work: 1969-70
- Earliest relational database research: ~1976
- Commercial Relational DBMSs: ~mid 1980s
- Widespread deployment” mid-1990’s
- Why the gap? Top 10 reasons…
  1. “You could do the same thing in other ways”
  2. “Nobody wants to write math formulas”
  3. “Why would I turn my data into tables?”
  4. “It won’t perform well”
  …
- What do you think?

Relational Model Definitions

- A relation is a table with columns and rows.
- An attribute is a named column of a relation.
- A tuple is a row of a relation.
- A domain is a set of allowable values for one or more attributes.
- The degree of a relation is the number of attributes it contains.
- The cardinality of a relation is the number of tuples it contains.
- A relational database is a collection of normalized relations with distinct relation names.

Example of a Relation

<table>
<thead>
<tr>
<th>Relation Name</th>
<th>Attributes</th>
<th>Tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDENT</td>
<td>Name, Sex, Home_phone, Address, Office_phone, Age, Gpa</td>
<td>Benjamin Bayer, Chung-ya Ken, Dick Davidson, Rohan Panchal, Barbara Benson</td>
</tr>
</tbody>
</table>

The attributes and tuples of a relation STUDENT.

Degree = 7, Cardinality = 5
Relational Model: Formal Definition

- Formally, a table is a relation over K sets (domains)
  - \( R \subseteq A_1 \times A_2 \times \ldots \times A_K \)
    Subset of the cartesian product of the K domains
  - Tuple = \((t_1, t_2, \ldots, t_K)\), where \( t_i \) is an element from domain/set \( D_i \)

- A database is a collection of relations
- Theoretically: a relation is a set of tuples; no tuple can occur more than once
  - Real systems may allow duplicates for efficiency or other reasons – we’ll ignore this for now

Relation Schemas and Instances

- A relation schema is a definition of a single relation.
- A relational database schema is a set of relation schemas (modeling a particular domain).

- A relation instance denoted \( r(R) \) over a relation schema \( R(A_1, A_2, \ldots, A_n) \) is subset of the Cartesian product of the domains of all attributes in the relation schema \( r(R) \subseteq \text{dom}(A_1) \times \text{dom}(A_2) \times \ldots \times \text{dom}(A_n) \)
  - i.e., a set of n-tuples \(<d_1, d_2, \ldots, d_n>\) where each \( d_i \) is an element of \( \text{dom}(A_i) \) or is null.
  - A value of null represents a missing or unknown value.

Definition Summary

<table>
<thead>
<tr>
<th>Informal Terms</th>
<th>Formal Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table</td>
<td>Relation</td>
</tr>
<tr>
<td>Column Header</td>
<td>Attribute</td>
</tr>
<tr>
<td>All possible Column Values</td>
<td>Domain</td>
</tr>
<tr>
<td>Row</td>
<td>Tuple</td>
</tr>
<tr>
<td>Table Definition</td>
<td>Schema of a Relation</td>
</tr>
<tr>
<td>Populated Table</td>
<td>State of the Relation</td>
</tr>
</tbody>
</table>
**Properties and constraints of the data**

- Data represented as a relation/table
- Schema specifies the attributes/columns and their type

- Any other properties we need to define to capture the "application"

- Need to capture the 'business' rules:
  - How do we uniquely identify a student?
  - Can their letter grade be any alphabet?
  - Can a student take a course that is not in the course schedule/bulletin?
  - ...

- Concept of Constraints

---

**CONSTRANTS**

Constraints determine which values are permissible and which are not in the database. Three main types:

- **Inherent or Implicit Constraints:**
  - These are based on the data model itself. (E.g., relational model does not allow a list as a value for any attribute)

- **Schema-based or Explicit Constraints – Integrity Constraints:** are rules or restrictions that apply to the database and limit the data values it may store
  - They are expressed in the schema by using the facilities provided by the model.

- **Application based or semantic constraints:** These are beyond the expressive power of the model and must be specified and enforced by the application programs.

---

**Integrity Constraints (ICs)**

- **IC:** condition that must be true for any instance of the database; e.g., domain constraints.
  - ICs are specified when schema is defined.
  - ICs are checked when relations are modified.
  - A legal instance of a relation is one that satisfies all specified ICs.
  - DBMS should not allow illegal instances.
  - Why is this useful

- If the DBMS checks ICs, stored data is more faithful to real-world meaning.
- Avoids data entry errors, too!
- Think of the constraints as the business rules derived from the application

---

**Where do ICs Come From?**

- ICs are based upon the semantics of the real-world enterprise that is being described in the database relations.
  - Need to carefully analyze the application before reaching a conclusion on the Integrity Constraints!

- We can check a database instance to see if an IC is violated, but we can NEVER infer that an IC is true by looking at an instance.
  - An IC is a statement about all possible instances!
  - From example, we know name is not a key, but the assertion that sid is a key is given to us.

- Key and foreign key ICs are the most common; more general ICs supported too.
  - Shall return to these after we cover DML aspect of SQL
Definition: Relational Model Integrity Constraints

- Integrity rules are used to insure the data is accurate.
- Types of constraints:
  - Domain constraint - Every value for an attribute must be an element of the attribute's domain or be null.
    null represents a value that is currently unknown or not applicable.
  - Entity integrity constraint - In a base relation, no attribute of a primary key can be null.
  - Key constraint – every relation must have a key
  - Referential integrity constraint - If a foreign key exists in a relation, then the foreign key value must match a primary key value of a tuple in the referenced relation or be null.

Key Constraints

- **Superkey** of R:
  - Is a set of attributes SK of R with the following condition:
    - No two tuples in any valid relation state r(R) will have the same value for SK
    - That is, for any distinct tuples t1 and t2 in r(R), t1[SK] ≠ t2[SK]
    - This condition must hold in any valid state r(R)

- **Key** of R:
  - A "minimal" superkey
  - That is, a key is a superkey K such that removal of any attribute from K results in a set of attributes that is not a superkey (does not possess the superkey uniqueness property)
  - A Key is a Superkey but not vice versa

Primary Key Constraints

- Every relation must have a key
- A set of fields is a key for a relation if:
  1. No two distinct tuples can have same values in all key fields, and
  2. This is not true for any subset of the key.
- Part 2 false? A superkey.
- If there’s >1 key for a relation, one of the keys is chosen (by DBA) to be the primary key.
- E.g., what is a key for Students relation?
- sid is a key for Students. (What about name?) The set {sid, gpa} is a superkey.

Key Constraints (continued)

- Example: Consider the CAR relation schema:
  - CAR(State, Reg#, SerialNo, Make, Model, Year)
  - CAR has two keys:
    - Key1 = {State, Reg#}
    - Key2 = {SerialNo}
  - Both are also superkeys of CAR
  - {SerialNo, Make} is a superkey but not a key.
- In general:
  - Any key is a superkey (but not vice versa)
  - Any set of attributes that includes a key is a superkey
  - A minimal superkey is also a key
  - Can have many candidate keys, one of them chosen as primary key
Key Constraints (continued)
- If a relation has several candidate keys, one is chosen arbitrarily to be the primary key.
  - The primary key attributes are underlined.
- Example: Consider the CAR relation schema:
  - CAR(State, Reg#, SerialNo, Make, Model, Year)
  - We chose SerialNo as the primary key
- The primary key value is used to uniquely identify each tuple in a relation
  - Provides the tuple identity
- Also used to reference the tuple from another tuple
  - General rule: Choose as primary key the smallest of the candidate keys (in terms of size)
  - Not always applicable – choice is sometimes subjective

Referential Integrity
- A constraint involving two relations
  - The previous constraints involve a single relation.
- Used to specify a relationship among tuples in two relations:
  - The referencing relation and the referenced relation.
- Examples:
  - Student enrolled in a course
  - Employee working on a project: in Works_On table

Foreign Keys in SQL
- Only students listed in the Students relation should be allowed to enroll for courses.

<table>
<thead>
<tr>
<th>Enrolled</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>sid</td>
<td>cid</td>
</tr>
<tr>
<td>53666</td>
<td>Jazz101</td>
</tr>
<tr>
<td>53666</td>
<td>Reggae203</td>
</tr>
<tr>
<td>53650</td>
<td>Topology112</td>
</tr>
<tr>
<td>53666</td>
<td>History105</td>
</tr>
<tr>
<td>sid</td>
<td>name</td>
</tr>
<tr>
<td>53666</td>
<td>Jones</td>
</tr>
<tr>
<td>53688</td>
<td>Smith</td>
</tr>
<tr>
<td>53650</td>
<td>Smith</td>
</tr>
</tbody>
</table>

Foreign Keys, Referential Integrity
- Foreign key: Set of fields in one relation that is used to refer to a tuple in another relation. (Must correspond to primary key of the second relation.)
  - Like a ‘logical pointer’.
- In Enrolled table – sid is a student, what can we say about the students table?
  - Enrolled(sid: string, cid: string, grade: string)
- sid is a foreign key referring to Students:
  - The student with this ID MUST exist in the Students table
  - If all foreign key constraints are enforced, referential integrity is achieved, i.e., no dangling references.
Referential Integrity (or foreign key) Constraint – more general definition

- Statement of the constraint
  - The value in the foreign key column (or columns) FK of the referencing relation R1 can be either:
    1. a value of an existing primary key value of a corresponding primary key PK in the referenced relation R2, or
    2. a null.
- In case (2), the FK in R1 should not be a part of its own primary key.

Other Types of Constraints

- Semantic Integrity Constraints:
  - based on application semantics and cannot be expressed by the model per se
  - Example:
    - “the max. no. of hours per employee for all projects he or she works on is 40 hrs per week”;
    - Grade cannot be any alphabet; …
  - A constraint specification language may have to be used to express these
  - SQL-99 allows CREATE TRIGGER and CREATE ASSERTION to express some of these semantic constraints
  - Keys, Permissibility of Null values, Candidate Keys (Unique in SQL), Foreign Keys, Referential Integrity etc. are expressed by the CREATE TABLE statement in SQL.

Relational Database Schema

- Relational Database Schema:
  - A set S of relation schemas that belong to the same database.
  - S is the name of the whole database schema
  - S = {R1, R2, ..., Rn} and a set IC of integrity constraints.
  - R1, R2, ..., Rn are the names of the individual relation schemas within the database S
  - Following slide shows a COMPANY database schema with 6 relation schemas

 COMPANY Database Schema

<table>
<thead>
<tr>
<th>EMPLOYEE</th>
<th>DEPARTMENT</th>
<th>DEPT_LOCATIONS</th>
<th>PROJECT</th>
<th>WORKS_ON</th>
<th>DEPENDENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Dnumber</td>
<td>Name</td>
<td>Name</td>
<td>Date</td>
<td>Name</td>
</tr>
<tr>
<td>Ssn</td>
<td></td>
<td></td>
<td>Phonenumber</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td>Location</td>
<td>Hours</td>
<td></td>
</tr>
<tr>
<td>Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Super_ssn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dno</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.5
Schema diagram for the COMPANY relational database schema.
Slide 5-44
Populated database state
- Whenever the database is changed, a new state arises
- Basic operations for changing the database:
  - INSERT a new tuple in a relation
  - DELETE an existing tuple from a relation
  - MODIFY an attribute of an existing tuple
- Next slide shows an example state for the COMPANY database schema
- The update operations must keep the database in a consistent state – i.e. all instances must satisfy integrity constraints
- Updates may propagate to cause other updates automatically. This may be necessary to maintain integrity constraints.

When do Integrity Constraints get triggered...

Update Operations on Relations
- In case of integrity violation, several actions can be taken:
  - Cancel the operation that causes the violation (RESTRICT or REJECT option)
  - Perform the operation but inform the user of the violation
  - Trigger additional updates so the violation is corrected (CASCADE option, SET NULL option)
  - Execute a user-specified error-correction routine
Possible violations for each operation

- **INSERT** may violate any of the constraints:
  - Domain constraint:
    if one of the attribute values provided for the new tuple is not of the specified attribute domain
  - Key constraint:
    if the value of a key attribute in the new tuple already exists in another tuple in the relation
  - Referential integrity:
    if a foreign key value in the new tuple references a primary key value that does not exist in the referenced relation
  - Entity integrity:
    if the primary key value is null in the new tuple

Possible violations for each operation

- **DELETE** may violate only referential integrity:
  - If the primary key value of the tuple being deleted is referenced from other tuples in the database
    Can be remedied by several actions: RESTRICT, CASCADE, SET NULL (see Chapter 6 for more details)
    - RESTRICT option: reject the deletion
    - CASCADE option: propagate the new primary key value into the foreign keys of the referencing tuples
    - SET NULL option: set the foreign keys of the referencing tuples to NULL
  - One of the above options must be specified during database design for each foreign key constraint

Possible violations for each operation

- **UPDATE** may violate domain constraint and NOT NULL constraint on an attribute being modified
  - Any of the other constraints may also be violated, depending on the attribute being updated:
    - Updating the primary key (PK):
      Similar to a DELETE followed by an INSERT
      Need to specify similar options to DELETE
    - Updating a foreign key (FK):
      May violate referential integrity
    - Updating an ordinary attribute (neither PK nor FK):
      Can only violate domain constraints

Relational Query Languages

- A major strength of the relational model: supports simple, powerful querying of data.
  Queries can be written intuitively, and the DBMS is responsible for efficient evaluation.
  - The key: precise semantics for relational queries.
  - Allows the optimizer to extensively re-order operations, and still ensure that the answer does not change.
Relational Query Languages

*Query languages:*
- Allow specification of schemas and constraints
- Allow manipulation and retrieval of data from a database.
*Relational model supports simple, powerful QLs:*
  - Strong formal foundation based on logic.
  - Allows for much optimization.
*Query Languages ≠ programming languages!*
  - QLs not expected to be “Turing complete”.
  - QLs not intended to be used for complex calculations.
  - QLs support easy, efficient access to large data sets.

Formal Query Languages

*Formal query languages are defined as mathematical operators over the set*
- advantage of a formal language?
- Relational algebra, Relational calculus are examples
*Procedural vs Non-procedural languages*
  - Procedural: what data to fetch from DB and how/where to get the data
  - Non-procedural: what data to fetch from DB
    - System/DBMS needs to figure out the “how”
  - Can have a mix in practice
    - Relational algebra: procedural language
    - Relational calculus: non-procedural (declarative)

SQL: Structured Query Language

*The standard language for relational data*
- Invented by folks at IBM, esp. Don Chamberlin
- Actually not a great language…
- Beat a more elegant competing standard, QUEL, from Berkeley
<sep>Separated into a DML & DDL
<sep>SQL DML component based on relational algebra & calculus

SQL

*components*
- Data definition (DDL) – to define schema/tables
  - Define Schema
  - Define Constraints
- Manipulation/query (DML) – for queries
- Transaction control – to specify a transaction
- Index – to specify storage and indexing schemes
- Authorization- for access control/security
  - We will cover the DDL and query part of SQL first
  - Shall return to the other components after we cover those topics
The Big Picture: SQL to Algebra to Query Plan to Web Page

SELECT *
FROM STUDENT, Takes, COURSE
WHERE STUDENT.sid = Takes.sID
AND Takes.cID = cid