

Introduction to CMPSC 430
Database Design

Spring 2016
Dr. Linda Null

Material based on slides provided by Silberchatz, Korth, and Sudarshan, course material from Juliana Freire and Claudio Silva

1

This Course...

- Is NOT
 - A tutorial on using a specific DBMS
 - A tutorial on SQL
 - A course on database implementation
- But You WILL learn
 - The foundations of database design
 - Some SQL and relational algebra
 - How to *use* databases
 - Practical benefits in using a DBMS

2

Why Study Databases?

- Databases used to be specialized applications; now they are a central component in most applications
- Shift from computation to information
- Knowledge of database concepts is essential for computer scientists
- Databases are everywhere, even when you don't see them... can you give me some examples?

3

Why Study Databases?

- Because data is valuable
 - Examples include tax records, student records, bank account records, photos, ...
 - These things must be protected, no matter what happens (disk crash, machine crash, flood, fire, etc.)
 - Must also protect from people...

4

Why Study Databases?

- Because data is typically structured
 - Tax records follow the same structure
 - Bank records follow the same structure
 - We can exploit this structure to help us
 - Store the data efficiently
 - Retrieve the data efficiently and in useful ways

5

Why Study Databases?

- Because the database field has made a number of contributions to the field of computer science
- DBs encompass most of CS (OS, language, theory, AI, logic)
- DB concepts can be applied to different problems in different areas
- DBMS s/w is highly successful as a commercial technology
- DB research is highly active and very interesting

6

Problem: Keeping An Address Book
Solution 1: A blank spiral notebook

- Entries recorded in pen, in time order
- Advantages:
 - Cheap, simple, and private
 - Reliable and space efficient
- Disadvantages
 - Hard to search, update, and share
 - Hard to add info to and limited in size
 - Hard to retrieve info from
 - Multiple entries are repeated
 - What if you lose it?

7

Solution 2: Loose Leaf Notebook

- Advantages
 - Easier to insert, update, and delete
 - Can keep sorted
- However
 - All other disadvantages of Solution 1 still apply
 - In particular, still very difficult to search by any "key" other than the one sorted on (probably name)

8

Solution 3: Text File (Text Editor)

- Advantages
 - Free format and unlimited size
 - Easily copied (can now have backup)
 - Easily shared
 - Substring searchable
 - Cleanly updatable (easy to edit)
- However...
 - What if our requirements grow?
 - This can create some unforeseen complications

9

Complication 1: The File Gets Very Large

- Problem
 - Searching gets very slow and imprecise
 - Searching for "Les Street" can yield "Charles Streeter"
- Solution
 - Structure data into "fields"
 - Add indexes on fields commonly searched

Database Concepts:

- Schema
- Record organization
- Indexes

Complication 2: Data Redundancy

- Why do we have redundancy?
 - Large families, frequent moves
 - Might forget to update addresses of some family members
 - Want a single point of update and to save space
 - Notion of residence as a separate entity is an important concept (e.g., 1 Xmas card per family)
- Solution
 - Separate residences from names: 2 files, one for people and one for residence
 - How to we associate a residence with a person?

Database Concepts:

- Consistency
- Normalization
- Foreign Keys

Complication 3: Multiple Associations of Persons and Residences

- What does this mean?
 - People can own, rent, manage, visit residences
 - May want constraints on the number of residences per person
- Examples
 - Many-to-one (single family), many-to-many (rich people with vacation homes), one-to-many (builders)

Database Concepts:

- Relationships
- Cardinality Constraints

Complication 4: Need to Add Information for New Purposes

- Examples
 - Xmas cards sent and received
 - Post office gives big discount for using zip+4
- Requirements
 - Adding fields and/or new tables

Database Concept:
• Schema Evolution

Complication 5: Doing Ad Hoc Analysis and Retrieval

- Example
 - “Who have we sent cards to for each of the past 5 years, but received 2 or fewer cards in return?”
- Requirements
 - Language for expressing analysis and retrieval
 - Implementation that performs analysis and retrieval correctly and efficiently

Database Concepts:
• Query Languages
• Query Optimization and Execution

Complication 6: Want to Organize the Data Differently for Some Users

- Examples
 - Other family members want to see the names and residences together
 - You don't want your kids to see your business entries
- Solution
 - Use stored queries as “windows” into the database
 - Data not selected by query is “not there”

Database Concepts:
• Joins
• Views
• Security

Complication 7: Required Existence of Associated Data

- **Examples**
 - Can't send a Xmas card to someone without an address
 - Names are not unique unless qualified by residence: the Tim Wahls living at 123 Main Street
- **Solutions**
 - Refuse to insert a name unless it is associated with an address
 - Refuse to delete an address if it is associated with a name
 - Or, tolerate multiple non-unique names...

Database Concepts:

- *Referential Integrity*
- *Consistency*

Complication 8: Want Programmed Access to Data

- **What does this mean?**
 - Want to write a Java or C++ program to search, display, and/or update entries
- **Solution**
 - Use data organization to define corresponding data types
 - Use access library to open, retrieve, update data

Database Concepts:

- *Database Schemas*
- *API*
- *Embedded Querying*

Complication 9: Multiple Updates on All or None Basis

- **Examples**
 - Two households merge
 - Requires changing residences of several people
 - What if computer crashes halfway through?
- **Solution**
 - Present illusion that all updates are done simultaneously
 - Implanted by commit or rollback of entire piece of work

Database Concepts:

- *Transactions*
- *Atomicity*

Complication 10: Your Computer Crashes (again...)

- Will your data still be there?
 - Uncorrupted?
 - In what state, given that a transaction was in progress?
- Solution
 - Make sure the old data are safely accessible until the latest *commit*

Database Concepts:

- *Data Durability*
- *Recovery*

Complication 11: Two Computers in Your Home

- How can data be shared?
 - USB key? Leads to version problems...
 - Let's assume we do figure out how to share it
 - What if one user is merging households and another is splitting them up? What do the results even mean?
- Common policy
 - Transactions are atomic
 - They appear to run one after the other, in some order

Database Concepts:

- *Transaction Isolation*
- *Concurrency Control*
- *Transaction Serializability*

Complication 12: A Home Computer and a Business Computer

- Is there one database or two?
 - Need speed and reliability at each site
 - Logically, have one database for maintenance and one for querying
 - Two database communicate
- Solutions
 - Personal data on home computer
 - Business data on business computer
 - Common logical view

Database Concepts:

- *Distributed Databases*
- *Data Partitioning*
- *Data Replication*

Complication 13: Want to Add Family Photos, Sound, and Videos

- Requirements
 - Ability to capture, store, and play new media
 - Logical integration into existing data
 - Querying: all photos of Tim Wahls fishing...

Database Concepts:

- *Multimedia Data*
- *Query by Content*

Complication 14: Your Uncle Sonny Gets the Genealogy Bug

- Vision
 - All family members pool their databases over the Internet
 - Together, all genealogy relationships can be recorded
- But
 - Aunt Sara is paranoid: will not reveal birthdates
 - You don't want business associates in genealogy database
 - Everyone wants complete control over safety of their own data
 - People use different formats for records and different name abbreviations for entries

Database Concepts:

- *Federated Databases*
- *Data Integration*

Complication 15: You Become President

- Of a university, a large organization, the USA
 - Your address list grows to hundreds of thousands or more
 - You realize it contains useful information *in the large*
- Examples
 - Which are the top 10 zip codes on the list?
 - Which zip codes have addresses that are most likely to send cards to you when you send card so them?
 - Which of those zip codes are in states that had less than 5% difference in Republican/Democratic presidential votes in 2012?

Database Concepts:

- *Data Mining*
- *Online Analytical Processing*
