



Database Design: Logical Models: Normalization and The Relational Model

University of California, Berkeley
School of Information
IS 257: Database Management

Lecture Outline



- Normalization
- Relational Advantages and Disadvantages

Lecture Outline



- Normalization
- Relational Advantages and Disadvantages



Normalization



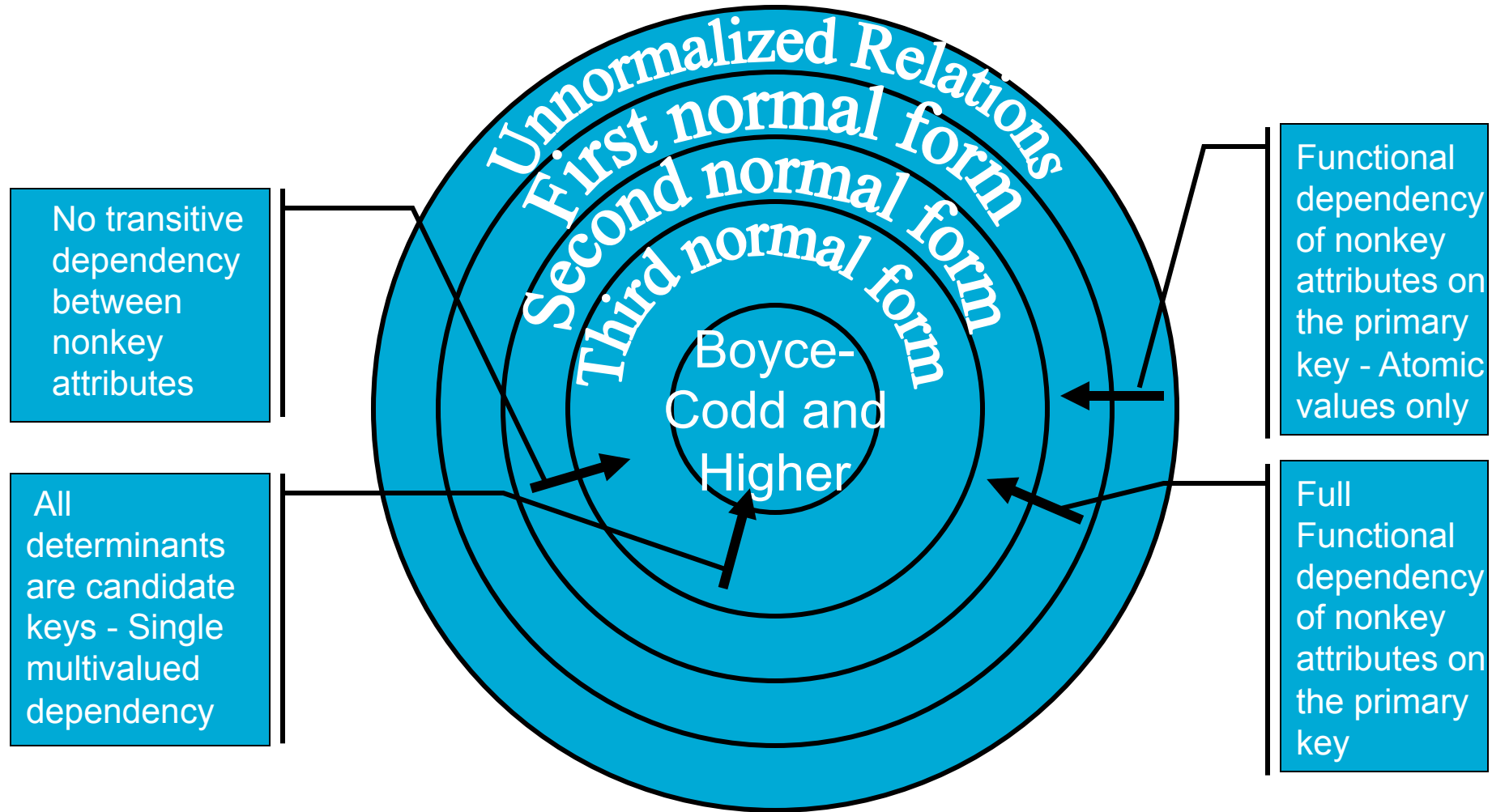
- Normalization theory is based on the observation that relations with certain properties are more effective in inserting, updating and deleting data than other sets of relations containing the same data
- Normalization is a multi-step process beginning with an “unnormalized” relation
 - Hospital example from Atre, S. *Data Base: Structured Techniques for Design, Performance, and Management.*

Normal Forms



- First Normal Form (1NF)
- Second Normal Form (2NF)
- Third Normal Form (3NF)
- Boyce-Codd Normal Form (BCNF)
- Fourth Normal Form (4NF)
- Fifth Normal Form (5NF)

Normalization



Unnormalized Relations



- First step in normalization is to convert the data into a two-dimensional table
- In *unnormalized* relations data may repeat within a column

Unnormalized Relation



| Patient # | Surgeon # | Surg. date | Patient Name | Patient Addr | Surgeon | Surgery | Postop drug | ug side effects |
|-----------|------------|----------------------------------|---------------|---|--------------------------------------|---|-----------------------|-----------------|
| 1111 | 145 311 | Jan 1, 1995; June 12, 1995 | John White | 15 New St. New York, NY | Beth Little Michael Diamond | Gallstone s removal; Kidney stones removal | Penicillin, none- | rash none |
| 1234 | 243 467 | Apr 5, 1994 May 10, 1995 | Mary Jones | 10 Main St. Rye, NY | Charles Field Patricia Gold | Eye Cataract removal Thrombos is removal | Tetracyclin e none | Fever none |
| 2345 | 189 | Jan 8, 1996 | Charles Brown | Dogwood Lane Harrison, NY | David Rosen | Open Heart Surgery | Cephalosp orin | none |
| 4876 | 145 | Nov 5, 1995 | Hal Kane | 55 Boston Post Road, Chester, CN | Beth Little | Cholecyst ectomy | Demicillin | none |
| 5123 | 145 | May 10, 1995 | Paul Kosher | Blind Brook Mamaronec k, NY | Beth Little | Gallstone s Removal | none | none |
| 6845 | 243 | Apr 5, 1994 Dec 15, 1984 | Ann Hood | Hilton Road Larchmont, NY | Charles Field | Eye Cornea Replacem ent Eye cataract removal | Tetracyclin e | Fever |

First Normal Form



- To move to First Normal Form a relation must contain only *atomic values* at each row and column.
 - No repeating groups
 - A column or set of columns is called a *Candidate Key* when its values can uniquely identify the row in the relation.

First Normal Form



| Patient # | Surgeon # | Surgery Date | Patient Name | Patient Addr | Surgeon Name | Surgery | Drug admin | Side Effects |
|-----------|-----------|--------------|------------------|---|--------------------|----------------------------------|-------------------|--------------|
| 1111 | 145 | 01-Jan-95 | John White | 15 New St. New York, NY | Beth Little | Gallstone s removal | Penicillin | rash |
| 1111 | 311 | 12-Jun-95 | John White | 15 New St. New York, NY | Michael Diamond | Kidney stones removal | none | none |
| 1234 | 243 | 05-Apr-94 | Mary Jones | 10 Main St. Rye, NY | Charles Field | Eye Cataract removal | Tetracyclin e | Fever |
| 1234 | 467 | 10-May-95 | Mary Jones | 10 Main St. Rye, NY | Patricia Gold | Thrombos is removal | none | none |
| 2345 | 189 | 08-Jan-96 | Charles Brown | Dogwood Lane Harrison, NY | David Rosen | Open Heart Surgery | Cephalosp orin | none |
| 4876 | 145 | 05-Nov-95 | Hal Kane | 55 Boston Post Road, Chester, CN | Beth Little | Cholecyst ectomy | Demicillin | none |
| 5123 | 145 | 10-May-95 | Paul Kosher | Blind Brook Mamaronec k, NY | Beth Little | Gallstone s Removal | none | none |
| 6845 | 243 | 05-Apr-94 | Ann Hood | Hilton Road Larchmont, NY | Charles Field | Eye Cornea Replacem ent | Tetracyclin e | Fever |
| 6845 | 243 | 15-Dec-84 | Ann Hood | Hilton Road Larchmont, NY | Charles Field | Eye cataract removal | none | none |

1NF Storage Anomalies



- **Insertion:** A new patient has not yet undergone surgery -- hence no surgeon # -- Since surgeon # is part of the key we can't insert.
- **Insertion:** If a surgeon is newly hired and hasn't operated yet -- there will be no way to include that person in the database.
- **Update:** If a patient comes in for a new procedure, and has moved, we need to change multiple address entries.
- **Deletion (type 1):** Deleting a patient record may also delete all info about a surgeon.
- **Deletion (type 2):** When there are functional dependencies (like side effects and drug) changing one item eliminates other information.

Second Normal Form



- A relation is said to be in Second Normal Form when every nonkey attribute is **fully functionally dependent** on the primary key.
 - That is, every nonkey attribute needs the full primary key for unique identification
- *This is typically accomplished by projecting (think splitting) the relations into simpler relations with simpler keys*

Second Normal Form



| Patient # | Patient Name | Patient Address |
|-----------|---------------|---|
| 1111 | John White | 15 New St. New York, NY |
| 1234 | Mary Jones | 10 Main St. Rye, NY |
| 2345 | Charles Brown | Dogwood Lane Harrison, NY |
| 4876 | Hal Kane | 55 Boston Post Road, Chester, Blind Brook |
| 5123 | Paul Kosher | Mamaroneck, NY |
| 6845 | Ann Hood | Hilton Road Larchmont, NY |



Second Normal Form



| Surgeon # | Surgeon Name |
|-----------|-----------------|
| 145 | Beth Little |
| 189 | David Rosen |
| 243 | Charles Field |
| 311 | Michael Diamond |
| 467 | Patricia Gold |



Second Normal Form



| Patient # | Surgeon # | Surgery Date | Surgery | Drug Admin | Side Effects |
|-----------|-----------|--------------|------------------------|---------------|--------------|
| 1111 | 145 | 01-Jan-95 | Gallstones removal | Penicillin | rash |
| 1111 | 311 | 12-Jun-95 | stones removal | none | none |
| 1234 | 243 | 05-Apr-94 | Eye Cataract removal | Tetracycline | Fever |
| 1234 | 467 | 10-May-95 | Thrombosis removal | none | none |
| 2345 | 189 | 08-Jan-96 | Open Heart Surgery | Cephalosporin | none |
| 4876 | 145 | 05-Nov-95 | Cholecystectomy | Demicillin | none |
| 5123 | 145 | 10-May-95 | Gallstones Removal | none | none |
| 6845 | 243 | 15-Dec-84 | Eye cataract removal | none | none |
| 6845 | 243 | 05-Apr-94 | Eye Cornea Replacement | Tetracycline | Fever |

1NF Storage Anomalies Removed



- **Insertion:** Can now enter new patients without surgery.
- **Insertion:** Can now enter Surgeons who haven't operated.
- **Deletion (type 1):** If Charles Brown dies the corresponding tuples from Patient and Surgery tables can be deleted without losing information on David Rosen.
- **Update:** If John White comes in for third time, and has moved, we only need to change the Patient table

2NF Storage Anomalies



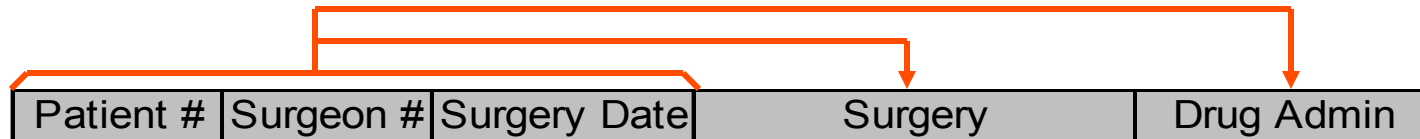
- **Insertion**: Cannot enter the fact that a particular drug has a particular side effect unless it is given to a patient.
- **Deletion**: If John White receives some other drug because of the penicillin rash, and a new drug and side effect are entered, we lose the information that penicillin can cause a rash
- **Update**: If drug side effects change (a new formula) we have to update multiple occurrences of side effects.

Third Normal Form




- A relation is said to be in Third Normal Form if there is no transitive functional dependency between nonkey attributes
 - When one nonkey attribute can be determined with one or more nonkey attributes there is said to be a transitive functional dependency.
- The side effect column in the Surgery table is determined by the drug administered
 - Side effect is transitively functionally dependent on drug so Surgery is not 3NF

Third Normal Form

| Patient # | Surgeon # | Surgery Date | Surgery | Drug Admin |
|-----------|-----------|--------------|------------------------|---------------|
| 1111 | 145 | 01-Jan-95 | Gallstones removal | Penicillin |
| 1111 | 311 | 12-Jun-95 | Kidney stones removal | none |
| 1234 | 243 | 05-Apr-94 | Eye Cataract removal | Tetracycline |
| 1234 | 467 | 10-May-95 | Thrombosis removal | none |
| 2345 | 189 | 08-Jan-96 | Open Heart Surgery | Cephalosporin |
| 4876 | 145 | 05-Nov-95 | Cholecystectomy | Demicillin |
| 5123 | 145 | 10-May-95 | Gallstones Removal | none |
| 6845 | 243 | 15-Dec-84 | Eye cataract removal | none |
| 6845 | 243 | 05-Apr-94 | Eye Cornea Replacement | Tetracycline |

Third Normal Form



| Drug Admin | Side Effects |
|---------------|--------------|
| Cephalosporin | none |
| Demicillin | none |
| none | none |
| Penicillin | rash |
| Tetracycline | Fever |

2NF Storage Anomalies Removed



- **Insertion**: We can now enter the fact that a particular drug has a particular side effect in the Drug relation.
- **Deletion**: If John White receives some other drug as a result of the rash from penicillin, but the information on penicillin and rash is maintained.
- **Update**: The side effects for each drug appear only once.

Normalization Checkpoint




- After completing 0 to 3, all nonkeys will be dependent on the primary key, the whole primary key, and nothing but the primary key (“so help you Codd!”)

Boyce-Codd Normal Form




- Most 3NF relations are also BCNF relations.
- A 3NF relation is NOT in BCNF if:
 - Candidate keys in the relation are composite keys (they are not single attributes)
 - There is more than one candidate key in the relation, and
 - The keys are not disjoint, that is, some attributes in the keys are common

Most 3NF Relations are also BCNF – Is this one?



| Patient # | Patient Name | Patient Address |
|-----------|---------------|---|
| 1111 | John White | 15 New St. New York, NY |
| 1234 | Mary Jones | 10 Main St. Rye, NY |
| 2345 | Charles Brown | Dogwood Lane Harrison, NY |
| 4876 | Hal Kane | 55 Boston Post Road, Chester, Blind Brook |
| 5123 | Paul Kosher | Mamaroneck, NY |
| | | Hilton Road |



BCNF Relations



| Patient # | Patient Name |
|-----------|------------------|
| 1111 | John White |
| 1234 | Mary Jones |
| 2345 | Charles Brown |
| 4876 | Hal Kane |
| 5123 | Paul Kosher |
| 6645 | Alfred E. Newman |

| Patient # | Patient Address |
|-----------|---|
| 1111 | 15 New St. New York, NY |
| 1234 | 10 Main St. Rye, NY |
| 2345 | Dogwood Lane Harrison, NY |
| 4876 | 55 Boston Post Road, Chester, Blind Brook |
| 5123 | Mamaroneck, NY |
| 6645 | Hilton Road |

Fourth Normal Form



- Any relation is in Fourth Normal Form if it is BCNF *and any multivalued dependencies are trivial*
- Eliminate non-trivial multivalued dependencies by projecting into simpler tables

Fourth Normal Form Example



| Restaurant | Pizza Variety | Delivery Area |
|---------------|---------------|---------------|
| Zoppo's Pizza | Thick Crust | Berkeley |
| Zoppo's Pizza | Thick Crust | Albany |
| Zoppo's Pizza | Thick Crust | Oakland |
| Zoppo's Pizza | Stuffed Crust | Berkeley |
| Zoppo's Pizza | Stuffed Crust | Albany |
| Zoppo's Pizza | Stuffed Crust | Oakland |
| Domino's | Thin Crust | Oakland |
| Domino's | Stuffed Crust | Oakland |
| Xtreme Pizza | Thick Crust | Berkeley |
| Xtreme Pizza | Thick Crust | Albany |
| Xtreme Pizza | Thin Crust | Berkeley |
| Xtreme Pizza | Thin Crust | Albany |
| ... | | |

Fourth Normal Form Example



- Each row indicates that a particular restaurant can deliver a particular kind of pizza to a particular city.
- There are NO non-key attributes because the only key is (Restaurant, Pizza Variety, Delivery Area).
- But, if we assume that the Pizza Varieties for a given Restaurant are the same regardless of the delivery area, then it is NOT in fourth normal form.

Fourth Normal Form Example



- The table features two non-trivial multivalued dependencies on the **Restaurant** attribute (which is not a superkey)
- These are:
 - Restaurant ->> Pizza Variety
 - Restaurant ->> Delivery Area
- This leads to redundancy in the table (e.g., we are told three times that Zoppo's has Thick Crust)

Fourth Normal Form Example



- If Zoppo's Pizza starts producing Cheese Crust pizzas then we will need to add multiple rows, one for each of Zoppo's delivery areas
 - And there's nothing to stop us from doing this incorrectly by *not including each delivery area*
- To eliminate these anomalies, the facts about varieties offered can be put in a different table from the facts about delivery areas
- This gives us two tables that are both in 4NF

Fourth Normal Form Example



| Restaurant | Pizza Variety |
|---------------|---------------|
| Zoppo's Pizza | Thick Crust |
| Zoppo's Pizza | Stuffed Crust |
| Domino's | Thin Crust |
| Domino's | Stuffed Crust |
| Xtreme Pizza | Thick Crust |
| Xtreme Pizza | Thin Crust |

| Restaurant | Delivery Area |
|---------------|---------------|
| Zoppo's Pizza | Berkeley |
| Zoppo's Pizza | Albany |
| Zoppo's Pizza | Oakland |
| Domino's | Oakland |
| Xtreme Pizza | Berkeley |
| Xtreme Pizza | Albany |

Fourth Normal Form Example



- But, suppose that the pizza varieties offered by a restaurant sometimes **did** legitimately vary from one delivery area to another, the original three-column table would satisfy 4NF

Fifth Normal Form



- A relation is in 5NF if every join dependency in the relation is implied by the keys of the relation
- *And* if it cannot have a lossless decomposition into any number of smaller tables
- *Implies that relations that have been decomposed in previous NF can be recombined via natural joins to recreate the original NF relations*

Normalization



- Normalization is performed to reduce or eliminate Insertion, Deletion or Update anomalies.
- However, a completely normalized database **may not be the most efficient or effective implementation.**
- “Denormalization” is sometimes used to improve efficiency.

Normalizing to death



- Normalization splits database information across multiple tables.
- To retrieve complete information from a normalized database, the JOIN operation must be used.
- JOIN tends to be expensive in terms of processing time, and very large joins are very expensive.

Denormalization

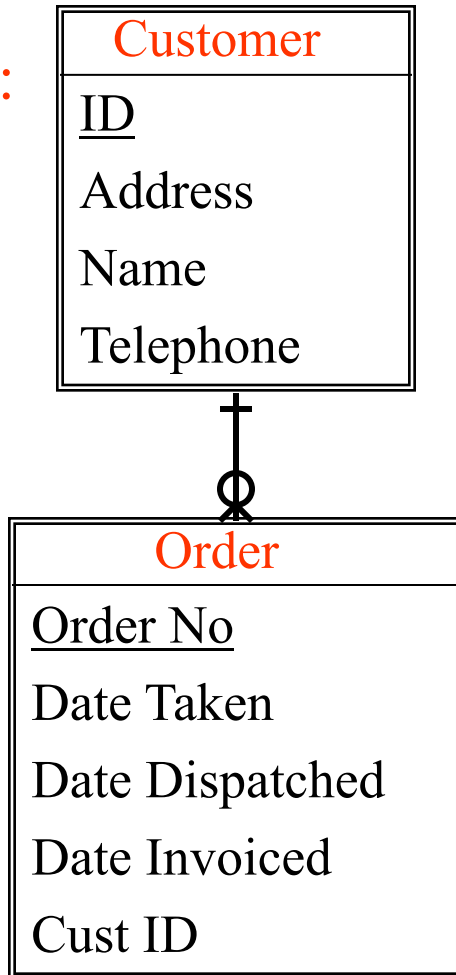


- Usually driven by the need to improve query speed
- Query speed is improved at the expense of more complex or problematic DML (Data manipulation language) for updates, deletions and insertions.

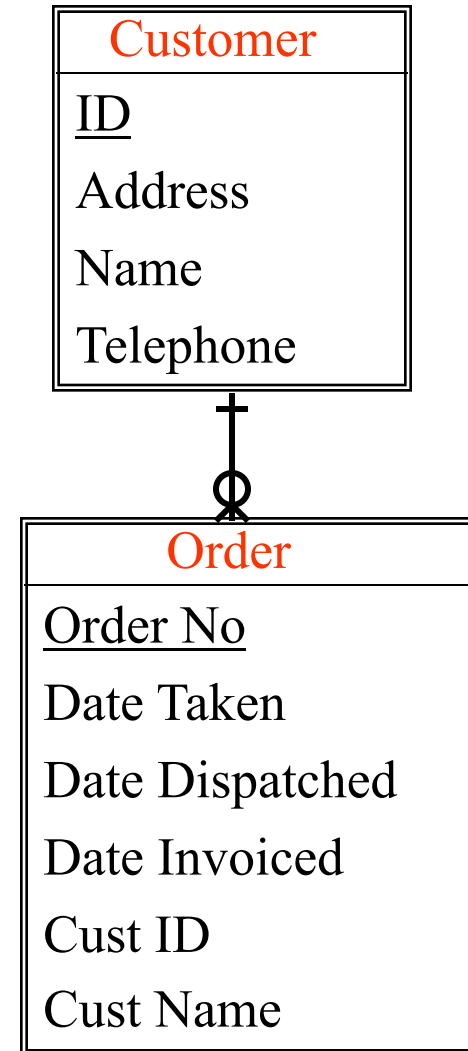
Downward Denormalization



Before:



After:



Upward Denormalization



| Order |
|-----------------|
| <u>Order No</u> |
| Date Taken |
| Date Dispatched |
| Date Invoiced |
| Cust ID |
| Cust Name |



| Order Item |
|-------------|
| Order No |
| Item No |
| Item Price |
| Num Ordered |

| Order |
|-----------------|
| <u>Order No</u> |
| Date Taken |
| Date Dispatched |
| Date Invoiced |
| Cust ID |
| Cust Name |
| Order Price |



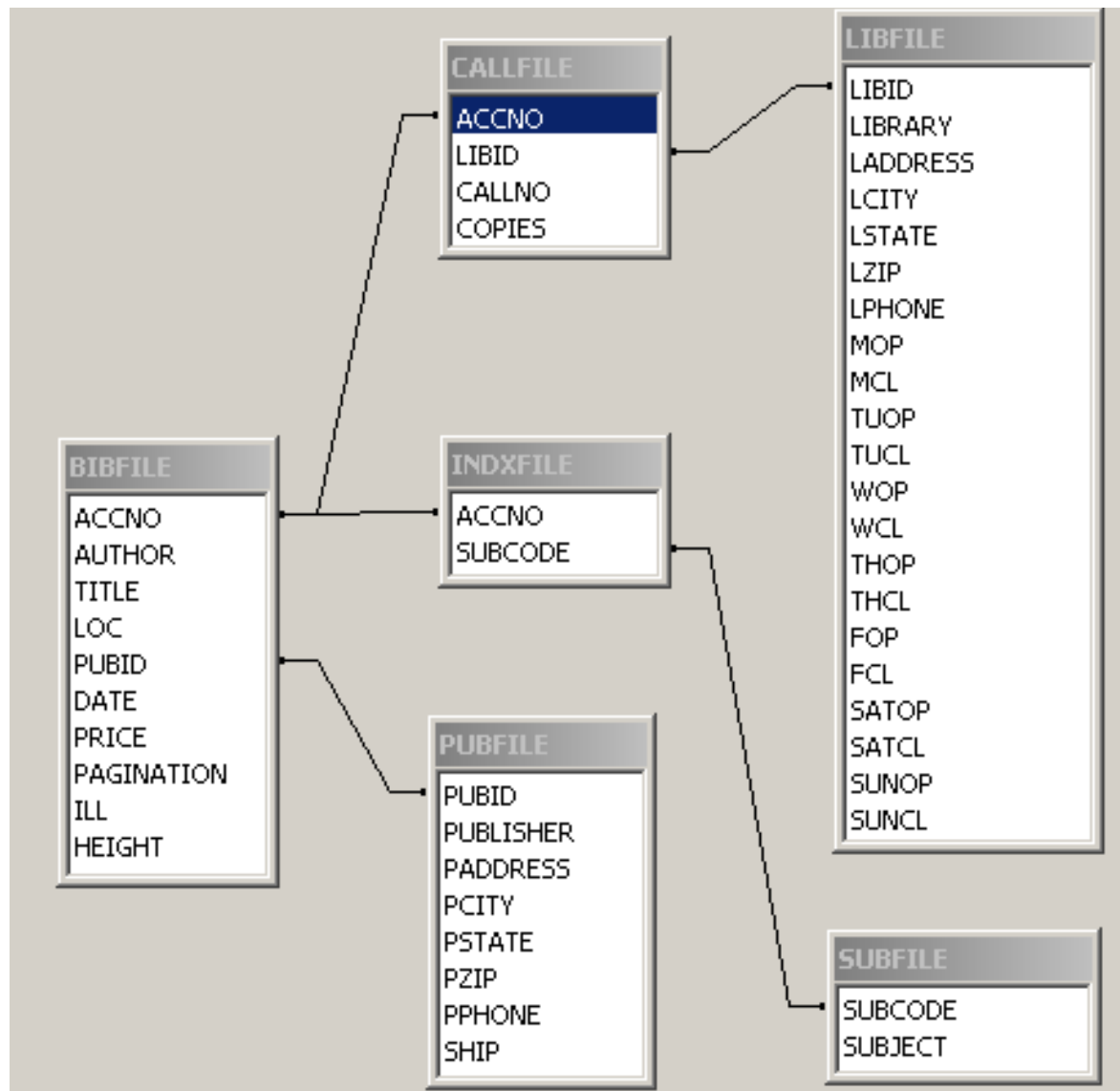
| Order Item |
|-------------|
| Order No |
| Item No |
| Item Price |
| Num Ordered |

Using RDBMS to help normalize



- Example database: Cookie
- Database of books, libraries, publisher and holding information for a shared (union) catalog

Cookie relationships



Cookie BIBFILE relation



| ACCNO | AUTHOR | TITLE | LOC | PUBID | DATE | PRICE | PAGINATIO | ILL | HEIGHT |
|-------|-----------------------|-----------------------|-----------------|-------|------|---------|---------------|--------|--------|
| A003 | AMERICAN LIBRARY ASS | ALA BULLETIN | CHICAGO | 04 | | \$3.00 | 63 V. | ILL. | 26 |
| T082 | ANDERSON, THEODORE | THE TEACHING OF MC | PARIS | 53 | 1955 | \$10.95 | 294 P. | | 22 |
| C024 | AXT, RICHARD G. | COLLEGE SELF STUD | BOULDER, CO. | 51 | 1960 | \$7.00 | X, 300 P. | GRAPHS | 28 |
| B006 | BALDERSTON, FREDERIC | MANAGING TODAYS U | SAN FRANCISCO | 27 | 1975 | \$6.00 | XVI, 307 P. | | 24 |
| B007 | BARZUN, JACQUES | TEACHER IN AMERICA | GARDEN CITY | 18 | 1954 | \$7.00 | 280 P. | | 18 |
| B005 | BARZUN, JACQUES | THE AMERICAN UNVE | NEW YORK | 24 | 1970 | \$5.00 | XII, 319 P. | | 20 |
| B008 | BARZUN, JACQUES | THE HOUSE OF INTEL | NEW YORK | 24 | 1961 | \$8.00 | VIII, 271 P. | | 21 |
| B010 | BELL, DANIEL | THE COMING OF POS | NEW YORK | 09 | 1976 | \$10.00 | XXVII, 507 P. | | 21 |
| B009 | BENSON, CHARLES S. | IMPLEMENTING THE LI | SAN FRANCISCO | 27 | 1974 | \$9.00 | XVII, 147 P. | | 24 |
| B012 | BERG, MAR | EDUCATION AND JOBS | BOSTON | 10 | 1971 | \$12.00 | XX, 200 P. | | 21 |
| B011 | BERSI, ROBERT M. | RESTRUCTURING THE | WASHINGTON, D.C | 03 | 1973 | \$11.00 | IV, 160P. | | 23 |
| B014 | BEVERIDGE, WILLIAM I. | THE ART OF SCIENTIF | NEW YORK | 58 | 1957 | \$14.00 | XIV, 239 P. | | 18 |
| B013 | BIRD, CAROLINE | THE CASE AGAINST C | NEW YORK | 08 | 1975 | \$13.00 | XII, 308 P. | | 18 |
| B016 | BISSELL, CLAUDE T. | THE STRENGTH OF TH | TORONTO | 57 | 1968 | \$14.00 | VII, 251 P. | | 21 |
| B017 | BLAIR, GLENN MYERS | EDUCATIONAL PSYCH | NEW YORK | 30 | 1962 | \$11.00 | 678 P. | | 24 |
| F047 | BLAKE, ELIAS, JR. | THE FUTURE OF THE | CAMBRIDGE, MA. | 02 | 1971 | \$14.25 | VIII, PP. 539 | | 23 |
| B116 | BOLAND, R.J. | CRITICAL ISSUES IN IN | CHICHESTER, ENG | 63 | 1987 | \$30.95 | XV, 394 P. | ILL. | 24 |
| S102 | BROWN, SANBORN C., E | SCIENTIFIC MANPOWE | CAMBRIDGE, MAS | 29 | 1971 | \$4.00 | X, 180 P. | | 26 |
| B118 | BUCKLAND, MICHAEL K. | LIBRARY SERVICES IN | ELMSFORD, NY | 70 | 1983 | \$12.00 | XII, 201 P. | ILL. | 23 |
| B018 | BUDIG, GENE A. | ACADEMIC QUICKSAN | LINCOLN, NEBRAS | 37 | 1973 | \$13.00 | 74 P. | | 23 |
| C031 | CALIFORNIA. DEPT. OF | LAW IN THE SCHOOL | MONTCLAIR, N.J. | 35 | 1974 | \$0.50 | IV, 87 P. | | 21 |
| C032 | CAMPBELL, MARGARET | WHY WOULD A GIRL | OLD WESTBURY, M | 48 | 1973 | \$1.50 | V, 114 P. | | 24 |
| C034 | CARNEGIE COMMISSION | A DIGEST OF REPORT | NEW YORK | 30 | 1974 | \$3.50 | 399 P. | | 24 |

How to Normalize?



- Currently no way to have multiple authors for a given book, and there is duplicate data spread over the BIBFILE table
- Can we use the DBMS to help us normalize?
- It is possible (but takes a bit more SQL knowledge than has been hinted at so far)
 - We will return to this problem later
 - But CONCEPTUALLY...

Using RDBMS to Normalize



Create a new table for Authors that includes author name and an automatically incrementing id number (for primary key)

Populate the table using the unique author names (which get assigned id numbers) by extracting them from the BIBFILE...

```
CREATE TABLE AUTHORS (AU_ID INT AUTO_INCREMENT PRIMARY KEY)  
AS SELECT DISTINCT (Author) from BIBFILE;
```

Create a new table containing a author_id and an ACCNO

Populate the new table by matching the Authors and BIBFILE names...

```
CREATE TABLE AU_BIB (AU_ID INT, ACCNO INT) AS SELECT AUTHORS.AU_ID,  
BIBFILE.ACCNO FROM AUTHORS, BIBFILE WHERE AUTHORS.Author = BIBFILE.Author;
```

Drop the Author name column from BIBFILE

```
ALTER TABLE BIBFILE DROP COLUMN Author
```

Lecture Outline



- Review
 - Logical Model for the Diveshop database
- Normalization
- **Relational Advantages and Disadvantages**

Advantages of RDBMS



- Relational Database Management Systems (RDBMS)
- Possible to design complex data storage and retrieval systems with ease (and without conventional programming).
- Support for **ACID** transactions
 - Atomic
 - Consistent
 - Independent
 - Durable

Advantages of RDBMS



- Support for *very large* databases
- Automatic optimization of searching (when possible)
- RDBMS have a simple view of the database that conforms to much of the data used in business
- Standard query language (SQL)

Disadvantages of RDBMS



- Until recently, no real support for complex objects such as documents, video, images, spatial or time-series data. (ORDBMS add -- or make available support for these)
- Often poor support for storage of complex objects from OOP languages (Disassembling the car to park it in the garage)
- Usually no efficient and effective *integrated* support for things like text searching within fields (MySQL now does have simple keyword searching with index support, but no ranking)

Effectiveness and Efficiency Issues for DBMS



- Our primary focus has been, and will continue to be, on the relational model
- Any column in a relational database can be searched for values
- To improve efficiency indexes using storage structures such as BTrees and Hashing are used
- But many useful functions are not indexable and require complete scans of the the database

Example: Text Fields



- In conventional RDBMS, when a text field is indexed, only exact matching of the text field contents (or Greater-than and Less-than).
 - Can search for individual words using pattern matching, but a full scan is required.
- Text searching is still done best (and fastest) by specialized text search programs (Search Engines) that we will look at more later

Assignment 2



Let's talk Assignment 2 and DB ideas.



Assignment 2b



- Due Friday March 9th
 - Personal Database Project Design
 - **Note: decide groups by February 23th**
-
- The following information should be turned in for the preliminary design of your personal database project.
 1. A written description of the data you will be using for the database, and what uses you might expect the database to have. (2-4 pages)
 2. A preliminary data dictionary for the entities and attributes and format of the data elements of the database. You should have *at least 5 entities with some logical connections between them*. The data dictionary consists of all of the attributes that you have identified for each entity, along with indication of whether the attribute is a primary key (or part of a primary key), and what format the data will be (e.g.: text, decimal number, integer, etc.)
 3. Produce an entity-relationship diagram of the database *OR* a UML diagram.
 - These will be preliminary design specifications, so do not feel that you must follow everything that you describe here in the final database design.
 - The report should be in PDF format

Discussion of Projects



- Anyone have any ideas for projects for this class?
- Introduce yourself and give a DB idea