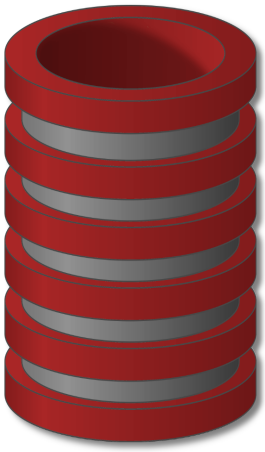
Announcements

- Assignments:
  - HW#3: due Wednesday!
- Midterm:
  - Wed 2/19
    - Recursion (Datalog, PostgreSQL)
    - XML: DTDs, XPath, XQuery
    - Relation Design Theory (BCNF, 4NF)
    - OLAP
- Today:
  - Closing Remarks on Design Theory
  - Online Analytical Processing (OLAP)



Relational Design Theory

Shortcomings of BCNF/4NF

Boyce-Codd Normal Form

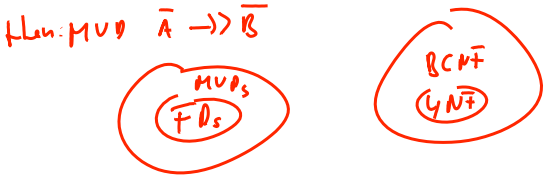
Relation R with FDs is in BCNF if:

For each A → B, A is a key

Fourth Normal Form

Relation R with MVDs is in 4NF if:

For each nontrivial A ↠ B, A is a key



Example: College application info.

Apply(SSN, cName, date, major)

Can apply to each college once for one major (1)

Colleges have non-overlapping application dates (2)

FDs: (1) SSN, cName → date, major (2) date → cName

Keys: {SSN, cName}

BCNF: No → decompose (2)

A1(date, cName)

A2(date, SSN, major)

separated

3NF: $\bar{A} \to \bar{B}$ then $\bar{A}$ is a key OR $\bar{B} \subseteq \bar{K}$ for a key $\bar{K}$

Good design? ?! To check FD (1), need to compute $A_1 \bowtie A_2$

BCNF is not dependency preserving

3NF is -"- -"-

Example #2

Student(SSN, HSname, GPA, priority)

Multiple HS okay, priority determined from GPA

FDs: SSN (1)→ GPA , GPA (2)→ priority , SSN (3)→ priority

(3) Keys: {SSN, HSname}

BCNF: No

S1(SSN, priority) ✓
[S2(SSN, HSname, GPA) ✗]
(1) S21(SSN, GPA) ✓
S22(SSN, HSname) ✓

BCNF ✓

Good design? ☹

Alternate design (3) with rule

[S1'(SSN, GPA, priority) ✗]
S2'(SSN, HSName) ✓
(2) S11'(GPA, priority)
S12'(SSN, GPA)

☹

Boyce-Codd Normal Form

Relation R with FDs is in BCNF if:

For each A → B, A is a key

Fourth Normal Form

Relation R with MVDs is in 4NF if:

For each nontrivial A ↠ B, A is a key

FOR BCNF, 4NF

After decomposition, no guarantee dependencies can be checked on decomposed relations

(in 3 NF : OK, i.e., dependency preserving)

Example #3

Scores(SSN, sName, SAT, ACT)

"Denormalized" relation

Multiple SATs and ACTs allowed

All queries return name + composite score for SSN

FDs + keys: SSN → sName ; Key : {SSN, SAT, ACT}

MVDs: SSN, sName →→ SAT | SSN, sName →→ ACT

4NF: No. S1(SSN, sName, SAT) | S2(SSN, sName, ACT)

S11(SSN, sName) S12(SSN, SAT) S21(SSN, sName) S22(SSN, ACT) 4 NF

$S_{11} \bowtie S_{12} \bowtie S_{22}$ ☹

for such query loads, keep denormalized relation (or use materialized view)

Example #4

FD

College(cName, state)
CollegeSize(cName, enrollment)
CollegeScores(cName, avgSAT)
CollegeGrades(cName, avgGPA)
...

BCNF/4NF? Yes

Good Design? ☺

"Too decomposed"

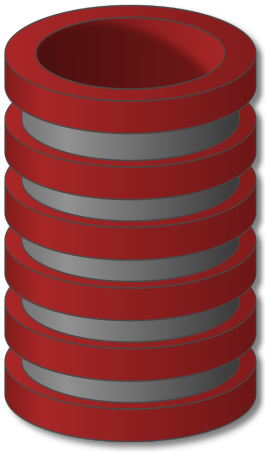
Shortcomings of BCNF/4NF

Designing a database schema

- Usually many designs possible
- Some are (much) better than others!
- How do we choose?

❖ Very nice theory for relational database design

- Normal forms - "good" relations
- Design by decomposition
- Usually intuitive and works well
- Some shortcomings
  - Dependency enforcement ~ need joins
  - Query workload ~ denormalization, materialized views
  - Over-decomposition ~ compose some



On-Line Analytical Processing (OLAP)

Introduction

Two broad types of database activity

- OLTP – Online Transaction Processing
  - Short transactions
  - Simple queries
  - Touch small portions of data
  - Frequent updates

traditional

- OLAP – Online Analytical Processing
  - Long transactions
  - Complex queries
  - Touch large portions of the data
  - Infrequent updates

decision support

More terminology

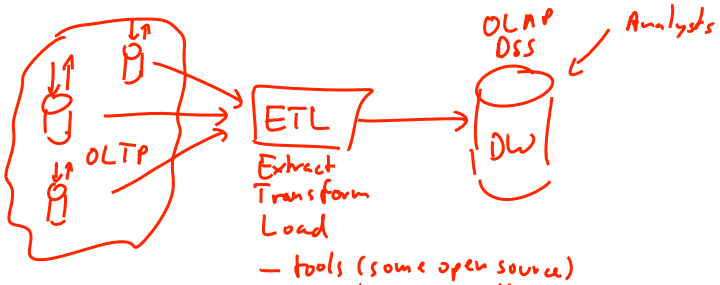
1995 · "OLAP-SQL" → SQL
1999 : Skyline-queries
1999 : Schema → BCNF/4 NF

- Data warehousing

  Bring data from operational (OLTP) sources into a single "warehouse" for (OLAP) analysis

- Decision support system (DSS)

  Infrastructure for data analysis
  E.g., data warehouse tuned for OLAP



– tools (some open source)
– scripts (Perl, Python, embedded SQL, PgSQL)