Normalization (Database Design) continued

RECAP

What is Normalization ?

- A technique for producing a set of relations with desirable properties (minimum data redundancy), given the data requirements of an enterprise.

- First developed by E.F. Codd (1972)

Normal Forms

There are several normal forms:

- 1NF
- 2NF
- 3NF
- BCNF
- 4NF
- 5NF

As normalization proceeds, relations become progressively more restricted (stronger) in format and also less vulnerable to update anomalies.

Help me Codd !

The key(1NF), the whole key(2NF), and nothing but the key (3NF) - so help me Codd !!

(Taken from Thomas Connnolly)

MATCH

1	Anomaly
2	Functional Dependency
3	A -> B
4	Prime attribute
5	Non-loss decomposition {R1, R2, ...Rn} of a relation R
6	Full FD
7	Transitive
8	Relation in 2NF
9	Relation in 3NF
10	Relation in 1NF

a	A is the dependent, B is the determinant
b	A is the determinant, B is the dependant
c	is one that is part of any primary key
d	does not have any repeating groups such as composite or mu valued attributes
e	is an inconsistent, incompatible or contradictory state of the database
f	is a many-to-one relationship between attribute set A and attribute set B
g	if 2 tuples in a relational instance agree on their X-value, the they must agree on their Y-value.
h	each value of A has associated with it exactly one value of B
i	does not have any tranistive depencies
j	is a FD of an attribute S on attribute R such that S does not depend on any proper subset of R
k	does not have any non-full dependencies
l	if there exists a FD of X -> Y and Y -> Z, then it implies that X Z also exists
m	natural join of R1, R2, ...Rn produces exactly the relation R
n	Pisceanple rules

Another Example of a bad design ?

SID	Name	Grade	Course#	Text	Major	Dept
S1	Joseph	A	CIS800	b1	CIS	CIS
S1	Joseph	B	CIS820	b2	CIS	CIS
S1	Joseph	A	CIS872	b5	CIS	CIS
S2	Alice	A	CIS800	b1	CS	MCS
S2	Alice	A	CIS872	b5	CS	MCS
S3	Tom	B	CIS800	b1	Acct	Acct
S3	Tom	B	CIS872	b5	Acct	Acct
S3	Tom	A	CIS860	b1	Acct	Acct

- Key is (SID, Course#)
- Is there any redundant data?
- Can we insert a new Course# with a new textbook?
- What should be done if 'CIS' is changed to 'MIS'?
- What would happen if we remove all CIS800 students?

Example :

An instance

Student	Course	Instructor
111	3530	Ritu
111	3110	Deb
222	3530	Fangju

Example: TEACH (Student, Course, Instructor)

- A student can take several courses and can be taught by several instructors.
- For each course, each student of that course is taught by only one teacher.
- An instructor teaches only one course.

✓ Two Candidate keys : (Student, Course) (Student, Instructor)

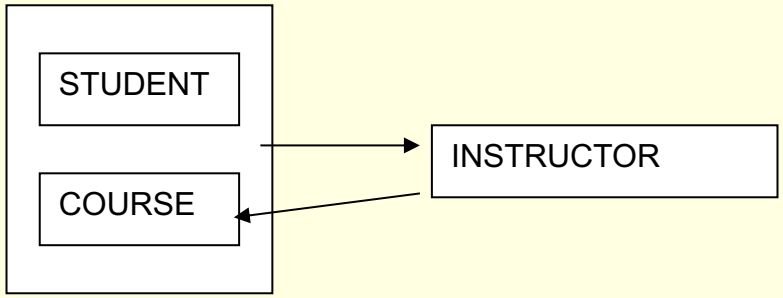
THIRD NORMAL FORM(3NF)

Relation TEACH (Student, Course, Instructor)

(STUDENT, COURSE) -> INSTRUCTOR

INSTRUCTOR -> COURSE

FD Diagram



- ✓ Is TEACH in 3NF?
- ✓ Still has Update Anomalies !!

Example :

If (Student, Course) is chosen as the Primary key, FD diagram is as given on slide 34.

If (Student, Instructor) is chosen as the Primary key ?

BOYCE/CODD NORMAL FORM

- A relation is in BCNF iff every determinant is a Candidate key.

- Update Anomalies occur in a 3NF relation R if
  - R has multiple candidate keys
  - Those candidate keys are composite and
  - The candidate keys are overlapped

- How to Normalize TEACHES in BCNF ?
  - Decompose TEACHES into

    SI (Student, Instructor) and IC (Instructor, Course)

BOYCE/CODD NORMAL FORM

- A relation is in BCNF iff every determinant is a Candidate key.

EMPLOYEE (Name, Project, Task, Office, Floor, Phone)

~~Name -> Office~~
Office -> Floor
Office -> Phone

A relation is in BCNF iff every determinant is a Candidate key.

Is EMPLOYEE in BCNF?

Name	Project	Task	Office	Floor	Phone
Bill	100X	T1	400	4	1400
Bill	100X	T2	400	4	1400
Bill	200Y	T1	400	4	1400
Bill	200Y	T2	400	4	1400
Sue	100X	T33	442	4	1442
Sue	200Y	T33	442	4	1442
Sue	300Z	T33	442	4	1442
Ed	100X	T2	588	5	1588

EMPLOYEE (Name, Project, Task, Office, Floor, Phone)

~~Name -> Office~~
Office -> Floor
Office -> Phone

R1

Name	Project	Task
Bill	100X	T1
Bill	100X	T2
Bill	200Y	T1
Bill	200Y	T2
Sue	100X	T33
Sue	200Y	T33
Sue	300Z	T33
Ed	100X	T2

R2

Name	Office
Bill	400
Sue	442
Ed	588

R3

Office	Floor	Phone
400	4	1400
442	4	1442
588	5	1588

Sue is given another task T34

MULTIVALUED DEPENDENCY(MVD)

- Although BCNF removes any anomalies due to FDs, there is another type of dependency called MVD which causes data redundancy.
- The possible existence of MVDs in a relation is due to 1NF which disallows an attribute in a tuple from having a set of values.
- An MVD exists in a relation R when there are two independent 1:N relationships in R

Example :MVD

SCH (Student, Courses, Hobbies)

Student	Courses	Hobbies
Jin	DBMS	Tennis
	OS	Cooking
Habib	C++	Reading
	Networking	Surfing
	CA	

- ✓ Is SCH Normalized ?

Example : MVD

- ✓ Convert SCH to 1NF

Student	Courses	Hobbies
Jin	DBMS	Tennis
Jin	DBMS	Cooking
Jin	OS	Tennis
Jin	OS	Cooking
Habib	C++	Reading
Habib	C++	Surfing
Habib	Networking	Reading
Habib	Networking	Surfing
Habib	CA	Reading
Habib	CA	Surfing

Two 1:N relationships in SCH
Student:Courses
Student:Hobbies

Definition : MVD

- An MVD represents a dependency between attributes A,B and C such that for each A, there is a set of values for B and a set of values for C AND the set of values for B and C are independent of each other.
- Represented as $A \rightarrow\rightarrow B|C$
- MVD in SCH : Student $\rightarrow\rightarrow$Courses|Hobbies

Higher Normal Forms : 4NF

Definition : A relation is in 4NF if it is in BCNF and it has no MVDs OR all its MVDs are also FDs.

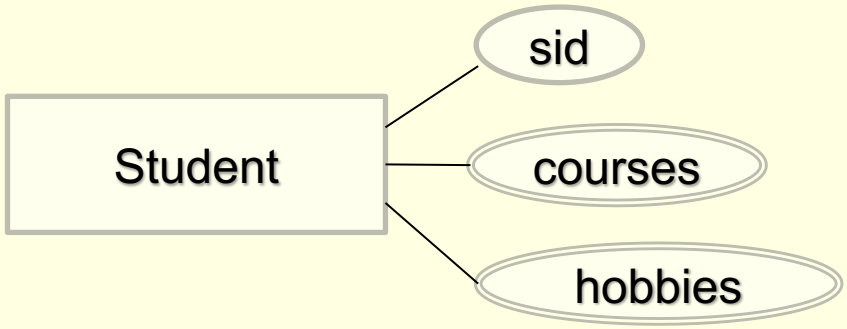
- ✓ Decomposition : same as done with other normal forms (break into smaller relations eliminating the MVDs)
- ✓ Decompose SCH into

  SC(Student, Courses) and SH (Student, Hobbies)

Note : Relations mapped from ER-Model need no transformation to 4NF !!

Higher Normal Forms : 4NF

Note : Relations mapped from ER-Model need no transformation to 4NF !!



Examples :MVD

- Do the following relations have MVDs ?

- ✓ PERSON(SIN, PhoneNos, KidsSIN)
- ✓ Bookstore(Course, Instructor, textbook)
- ✓ FACULTY(UoGuelphId, coursesTaught, dependentSIN)
- ✓ STUINFO(StudentNo, Height, ShoeSize)

4NF

Is

STUINFO(StudentNo, Height, ShoeSize)

in 4NF - yes, because its MVDs are also FD!

Example :

Assume that an employee can have multiple assignments and can also be involved in multiple service organizations.

Suppose employee 10123 does volunteer work for the Red Cross and United Way. In addition, the same employee might be assigned to work on 3 projects 1,5 and 12.

- ✓ Is the following table emp_ser_assn a correct representation of the above requirement ?
- ✓ What is the Primary key ?

EMP_NUM	EMP_SERVICE	EMP_ASSIGN
10123	Red Cross	1
10123	United Way	5
10123		12

Example – Continued

- ✓ Represent the relation emp_ser_assn correctly ?
- ✓ Does it have anomalies ? Why ?
- ✓ Decompose emp_ser_assn into 4NF relations

Denormalization

- Normalization is one of many database design goals
- But Is it always the best possible design ?

Why Denormalize ?

- Normalization is the process of putting one fact in one appropriate place. This optimizes updates at the expense of retrievals.

- When a fact is stored in only one place, retrieving many different but related facts usually requires going to many different places. This tends to slow the retrieval process.

- Updating is quicker, however, because the fact you're updating exists in only one place.

Why Denormalize ?

- A relational normalized database imposes a heavy access load over physical storage of data even if it is well tuned for high performance.

- A normalized design will often store different but related pieces of information in separate logical tables (called relations).
  - If these relations are stored physically as separate disk files, completing a database query that draws information from several relations (a *join operation*) can be slow.

Denormalization

- **Denormalization**
  - is the process of attempting to optimize the performance of a database by adding redundant data or by grouping data.
  - technique to move from higher to lower normal forms of database modeling in order to speed up database access.

An Example of Denormalization

Consider a relation :

Contact (Name, Street, Zip, City, Province)

Name	Street	Zip	City	Province
John	401 Sunset Av.	N91 Q23	Windsor	ON
Harry	402 Sunset Av.	N4T 3R5	Windsor	ON
Bill	123 First St.	N4Y 7Y8	London	ON

PK of Contact ?

Is Contact in 3NF ?

Example of Denormalization(contd)

Was it worth the decomposition ?