Database Design and Normal Forms

- Database Design
- First Normal Form
- Second Normal Form
- Third Normal Form
- Boyce-Codd Normal Form
- Fourth Normal Form
- Fifth Normal Form (Projection Join Normal Form)

Pitfalls in Relational Database Design

- Given a body of data to be represented in a database, deciding the logical structure for t he data (the relations needed and their schema) constitute the database design problem.
- Relational database design requires that we find a "good" collection of relation schemas
- A bad design may lead to

Redundancy ---- Information is repeated across tuples

Because of the redundancy there may be other problems

1. **Insertion Anomalies---difficulty in representing certain information**
2. **Deletion Anomalies ---If information is repeated across tuples ,then deletion of inform ation has to be performed across all these tuples.**
3. **Update Anomalies---- If information is repeated across different tuples, then update of information has to be performed across all these tuples.**
4. **Difficulty in checking integrity constraints**

Example of Bad Design

- Consider the relation schema:

Lending-schema = (*branch-name, branch-city, assets, customer-name, loan-number, amount)*

branch-name	*branch-city*	*assets*	*customer-name*	*loan-number*	*amount*
Downtown	Brooklyn	9000000	Jones	L-17	1000
Redwood	Palo Alto	2100000	Smith	L-23	2000
Perryridge	Horseneck	1700000	Hayes	L-15	1500
Downtown	Brooklyn	9000000	Jackson	L-14	1500

- Redundant Information: Data for *branch-name, branch-city, assets* are repeated for each loan that a branc h makes
- Insertion Anomaly: Cannot store information about a branch if no loans exist .Can use null values, but th ey are difficult to handle
- Deletion Anomaly: Cancellation of a loan may lead to loss of information such as details of a branch.(if t hat is the only tuple for the branch)
- Update Anomaly: Wastes space and complicates updates, introducing possibility of inconsistency of *asse ts* value

Goal — Formalize the notion of good design

- Design Goals:
  - **Avoid redundant data and the other anomalies**
  - **Facilitate the checking of updates for violation of integrity constraints**
- Process: Normalization
  - **Decide whether a particular relation *R* is in "good" form.**
  - **In the case that a relation *R* is not in "good" form, decompose it into a set of relations {R_1, R_2, ..., R_n} such that**
    - each relation is in good form
    - the decomposition is a lossless-join decomposition
  - **i.e. Anomalies are removed from a relation R(A), by decomposing it into other relations S(B) and T(C) where B,C < A ,such that there are no anomalies in S and T.**
  - **The process is based on *functional dependencies* .Functional dependencies allow us to form alize good database design**

Normalization

- **Proposed by Codd (1972)**
- Normalization is the step by step analysis and decomposition of complex relations into simple relations based on functional dependencies to reduce redundancy and inconsistency .
- It is the reversible process of transforming an unnormalized relation into relation with progressively simple relation. Since the process is reversible no information is lost in the transformation.
- All normalized relations are in first normal form, some first normal form relations are in second normal form and some second normal form relation are also in third normal form and so on.
- The motivation behind this is that second normal form is more desirable than first normal form and so on.

Normalization (Cont.)

- Normalization Stages
  - **1NF - First normal form ------- included in the definition of a relation**
  - **2NF - Second normal form ----- defined in terms of FDs**
  - **3NF - Third normal form ------- defined in terms of FDs**
  - **BCNF – Boyce Codd Normal Form….. defined in terms of FDs**
  - **4NF - Fourth normal form…… defined using multivalued dependencies**
  - **5NF(PJNF) – Fifth normal form (Project Join normal form) …. Defined using join dependencies**
- In order to achieve one level of normal form, each previous level must be met.
- **Codd proposed 3 normal forms, the first, second and third normal form**
- **A stronger definition of 3NF - called Boyce-Codd normal form (BCNF) was proposed later**
- **Later, 4NF and 5NF were proposed**
- **The minimum, and most common, goal is to achieve 3NF.**

First Normal Form

- A relational schema R is in first normal form if the domains of all attributes of R are atomic
- A domain is atomic if its elements are considered to be indivisible units
- **Each attribute must be atomic (single value)**
  - No repeating columns within a row (composite attributes)
  - No multi-valued columns

Example ---Unnormalized Relation

Patient #	Surgeon #	Surg. date	Patient Name	Patient Addr	Surgeon	Surgery	Postop drug	ug side effec
1111	145 311	Jan 1, 1995; June 12, 1995	John White	15 New St. New York, NY	Beth Little Michael Diamond	Gallstones 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 Thrombosis removal	Tetracycline none	Fever none
2345	189	Jan 8, 1996	Charles Brown	Dogwood Lane Harrison, NY	David Rosen	Open Heart Surgery	Cephalosporin	none
4876	145	Nov 5, 1995	Hal Kane	55 Boston Post Road, Chester, CN	Beth Little	Cholecystectomy	Demicillin	none
5123	145	May 10, 1995	Paul Kosher	Blind Brook Mamaroneck, NY	Beth Little	Gallstones Removal	none	none
6845	243	Apr 5, 1994 Dec 15, 1984	Ann Hood	Hilton Road Larchmont, NY	Charles Field	Eye Cornea Replacement Eye cataract removal	Tetracycline	Fever

To move to First Normal Form a relation must contain only atomic values

Patient_ Surgery_Details

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	Gallstones 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	Tetracycline	Fever
1234	467	10-May-95	Mary Jones	10 Main St. Rye, NY	Patricia Gold	Thrombosis removal	none	none
2345	189	08-Jan-96	Charles Brown	Dogwood Lane Harrison, NY	David Rosen	Open Heart Surgery	Cephalosporin	none
4876	145	05-Nov-95	Hal Kane	55 Boston Post Road, Chester, CN	Beth Little	Cholecystectomy	Demicillin	none
5123	145	10-May-95	Paul Kosher	Blind Brook Mamaroneck, NY	Beth Little	Gallstones Removal	none	none
6845	243	05-Apr-94	Ann Hood	Hilton Road Larchmont, NY	Charles Field	Eye Cornea Replacement	Tetracycline	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 if it is in 1NF and when every non key attribute is fully functionally dependent on any key of R.
- That is, every nonkey attribute needs the full primary key for unique identification
- Full functional dependency:

  An FD X → A for which there is no proper subset Y of X such that Y → A (A is said to be fully functionally dependent on X)

Example

- In the example the primary key is

  Patient#, surgeon#, surgery date
- The FDs in the example are:

  Patient#, surgeon#, surgery date → Patient Name

  Patient#, surgeon#, surgery date → Patient Addr

  Patient#, surgeon#, surgery date → Surgeon Name

  Patient#, surgeon#, surgery date → Surgery

  Patient#, surgeon#, surgery date → Drug admin

  Patient#, surgeon#, surgery date → Side Effects

  Patient# → Patient Name

  Patient# → Patient Addr

  surgeon# → Surgeon Name

  Drug admin → Side Effects

Example (Cont.)

- The example relation is not in second normal form because the attributes Patient Name ,Patient Addr ,Surgeon Name are not fully functionally dependent on the primary key. (They are partially dependent on the primary key.)
- The solution is to decompose the relation so that the resultant relations are in 2NF there by removing the storage anomalies noted earlier.
- The decomposition :

  Patient(Patient#, Patient Name ,Patient Addr)

  Surgeon(Surgeon#, Surgeon Name)

  Surgery_details (Patient#, Surgeon#, Surgery Date, Surgery, Drug Admin ,Side Effects)

Example (cont.) Second Normal Form

- Patient

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,
5123	Paul Kosher	Blind Brook Mamaroneck, NY
6845	Ann Hood	Hilton Road Larchmont, NY

Example (cont.) Second Normal Form

Surgeon

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

Surgery_details

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 effe ct unless it is given to a patient
- Deletion: If John White receives some other drug because of the penicillin ra sh, and a new drug and side effect are entered, we lose the information that p enicillin can cause a rash
- Update: If drug side effects change (a new formula) we have to update multip le occurrences of side effects

Third Normal Form

- A relation is said to be in Third Normal Form if it is in 2NFand every non key attribute is no n transitively dependent on any key of the relation.
- In the surgery_details relation the key is (Patient#,surgeon #,Surgery Date)
- FDs :1.Patient#, surgeon#, surgery date → Surgery

  2.Patient#, surgeon#, surgery date → Drug admin

  3.Patient#, surgeon#, surgery date → Side Effects

  4.Drug admin → Side Effects

The third FD is transitive FD because of the FDs 2 and 4.

- The relation is in 2NF.

Surgery_details

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

Third Normal Form (Cont.)

- The example relation is not in third normal form because of the transitive functional dependency.
- The solution is to decompose the relation so that the resultant relations are in 3NF there by removing the storage anomalies noted earlier.
- The decomposition :

  Surgery_description (Patient#, Surgeon#, Surgery Date, Surgery, Drug Admin)

  Drug (Drug Admin, Side Effects)

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

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 recieves some other drug as a result of the rash from penicillin, but t he information on penicillin and rash is maintained
- Update: The side effects for each drug appear only once

Boyce-Codd Normal Form (BCNF)

- 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 att ributes)**
  - **There is more than one candidate key in the relation, and**
  - **The keys are not disjoint, that is, some attributes in the keys are commo n**

Boyce-Codd Normal Form (BCNF) (Cont.)

- E.g. Consider the schema

  gradeInfo (rollNo,Adhar,course,grade)
- Suppose the following FDs hold:

  1)rollNo,course → grade

  2) Adhar, course ↛ grade

  3)rollNo → Adhar

  4) Adhar → rollNo
- Keys: (rollNo, course) and (Adhar,course)
- The relation is in 3 NF. But it has storage anomalies such as Adhar is stored redundantly alo ng with every course being done by the student.

Boyce-Codd Normal Form (BCNF) (Cont.)

- Definition : Relation schema R is in BCNF if for every nontrivial FD X ↛ A, X is a *superk ey* of R.
- In gradeInfo, FDs 3,4 are nontrivial but LHS is not a superkey

  So, gradeInfo is not in BCNF
- Decompose :gradeInfo (rollNo,course,grade)

  studInfo (rollNo,name)
- Redundancy allowed by 3NF is disallowed by BCNF
- BCNF is stricter than 3NF

Fourth Normal Form

- Any relation is in Fourth Normal Form if it is in BCNF *and* any multivalued dependencies a re trivial
- Eliminate non-trivial multivalued dependencies by projecting into simpler tables
- So in the example relation service there is a nontrivial MVD and it suffers redundant inform ation and and the associated updation anomaly for example to add the information that flight 108 uses a new plane type it is necessary add 3 new tuples.
- The solution is to decompose the relation as shown below .

Fourth Normal Form- Example

- Example: Relation *Service*

Flight	Day	Plane-Type
106	Monday	747
106	Thursday	747
106	Monday	1011
106	Thursday	1011
204	Wednesday	707
204	Wednesday	727

Fourth Normal Form-Example (Cont.)

- The relation service can be decomposed losslessly onto Service-Day(FLIGHT, DAY) and Service-Type(FLIGHT, PLANE-TYPE) as shown below and thereby removing redundancy and associated anomalies.

- Service-Day

Flight	Day
106	Monday
106	Thursday
204	Wednesday

Service-Type

Flight	Plane-Type
106	747
106	1011
204	707
204	727

Fifth Normal Form (Projection Join Normal Form)

- Any relation r is in Fifth Normal Form whenever there exits a non trivial JD * (R1,R2,…Rn) in r every Ri is a super key of R.
- SPJ is not in 5 NF because there exists non trivial JD

  * (SP, PJ, JS) and SP, PJ and JS are not super keys in R.
- The projections SP, PJ and JS are in 5NF since they do not involve JDs.

S#	P#	J#
S1	P1	J2
S1	P2	J1
S2	P1	J1
S1	P1	J1

S#	P#
S1	P1
S1	P2
S2	P1

P#	J#
P1	J2
P2	J1
P1	J1

S#	J#
S1	J2
S1	J1
S2	J1