DBMS: Normalization

Relational Database Design

SWARUP KR GHOSH

Normalization

Normalization is a step-by-step process of replacing a given relation by a successive collection of relations in order to achieve simpler and better data representation.

Objective!!

- to eliminate different anomalies that may occur due to referential integrity constraint
- to identify a suitable set of relations in database design

Criteria for decomposition in normalization!!

Loseless decomposition – Ensures no loss of information

Dependency preserving – Ensures no loss of functional dependencies.

❖ Functional Dependencies

- Let *R* be a relation schema

  $\alpha \subseteq R$ *and* $\subseteq R$

- The functional dependency

  $\alpha \rightarrow$

  holds on *R* if and only if for any legal relations *r*(R), whenever any two tuples t_1 and t_2 of *r* agree on the attributes α, they also agree on the attributes . That is,

  $t_1[\alpha] = t_2[\alpha] \Rightarrow t_1[\ ] = t_2[\ ]$

  Example: Consider *r*(α ,) with the following instance of *r.*

1	5
1	4
3	7

On this instance, $\alpha \rightarrow$ does **NOT** hold, but $\rightarrow \alpha$ does hold.

Functional Dependencies(Contd..)

K is a superkey for relation schema *R* if and only if $K \rightarrow R$

K is a candidate key for *R* if and only if

- $K \rightarrow R$, and
- for no $\alpha \subset K$, $\alpha \rightarrow R$

Functional dependencies allow us to express constraints that cannot be expressed using superkeys. Consider the schema:

bor_loan = (*customer_id, loan_number, customer_name, amount*).

We expect this functional dependency to hold:

loan_number $\rightarrow$ *amount*

but would not expect the following to hold:

amount $\rightarrow$ *customer_name*

Type Of Functional Dependencies

- **Full functional dependency** – A functional dependency of the form X→ Y is said to be full functional dependency, if any attribute is removed from X then dependency sustains no more.
- **Partial functional dependency** - A functional dependency of the form X→ Y is said to be partial functional dependency, if any attribute is removed from X then functional dependency persists.
- **Trivial functional dependency** – A functional dependency of the form X→ Y is called trivial functional dependency, if Y⊆ X.
- **Transitive functional dependency** – In a relation R, assume X→ Y and Y→ Z. In this case, the functional dependency X→ Z is known as transitive functional dependency.
- **Multi-valued dependency** – In a relation R, X, Y and Z are three different attributes (or set of attributes) such that X⊂ R, Y⊂R and Z⊂R. If for every value of X there exists a set of values for Y and Z, but the set of values for Y and Z are independent on each other, then Y and Z are multi-valued dependent on X.

Example: Functional Dependencies

- **Functional Dependency!!**
  - ✓emp-id → ename, designation.
- **Transitive Dependency!!**
  - ✓course-name→ duration
  - ✓duration → fees
  - ✓ ⇒course-name → fees
- **Multi-valued Dependency!!**
  - ✓emp-id ↠ project-no
  - ✓emp-id ↠ hobby

Anomalies due to Referential Integrity Constraints

- Repetition Anomaly:

  If a foreign key exists between two relations certain information may be repeated unnecessarily.

- Insertion Anomaly:

  This anomaly may occur during the insertion of new records into the referencing relation.

- Update Anomaly:

  This anomaly may occurs during update of existing records in both referencing and referenced relations.

- Delete Anomaly:

  During deletion of existing records from referenced relation this anomaly may occur (shifting case).

Anomalies

1. **Update Anomaly**: Employee 519 is shown as having different addresses on different records

Employees' Skills

Employee ID	Employee Address	Skill
425	87 Sycamore Grove	Typing
425	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

Resolution: Decompose the Schema

1. *Update*: (ID, Address), (ID, Skill)
2. *Insert*: (ID, Name, Hire Date), (ID, Code)
3. *Delete*: (ID, Name, Hire Date), (ID, Code)

2. **Insertion Anomaly**: Until the new faculty member, Dr. Newsome, is assigned to teach at least one course, his details cannot be recorded

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201
424	Dr. Newsome	29-Mar-2007	?

3. **Deletion Anomaly**: All information about Dr. Giddens is lost if he temporarily ceases to be assigned to any courses.

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201

Utility of Normalization

- To avoid unnecessary redundancy.
- To control those anomaly which occurs due to referential constraints.
- To get better data representation.

Decomposition

- Partition on relation
- Lossless Decomposition:

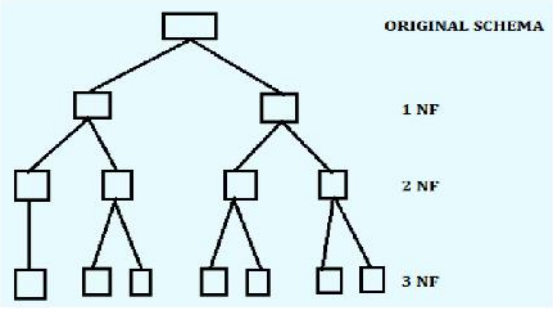
Ensure no loss of information.

- Dependency Preservation:

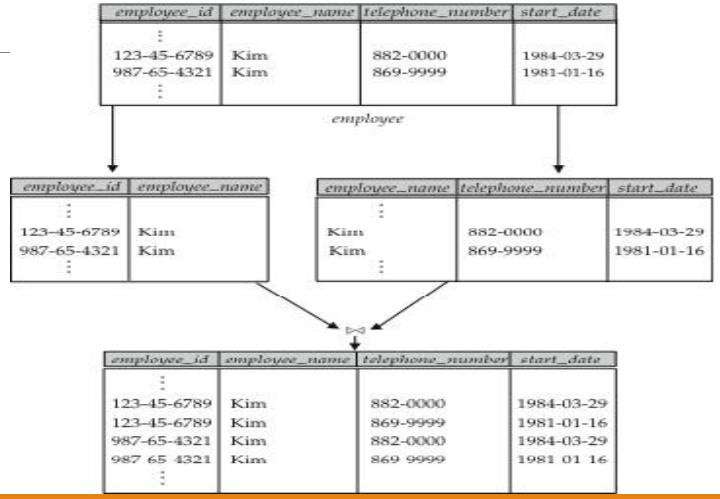
Ensure no loss of functional dependency.

Desirable Properties of Decomposition

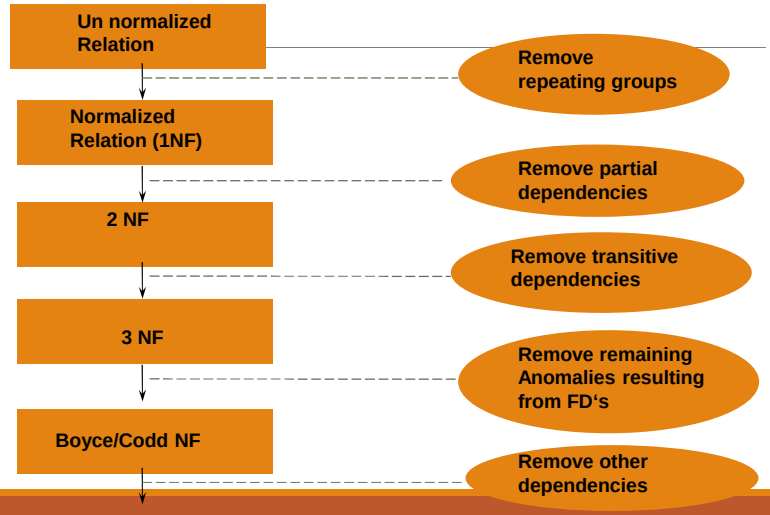
- Lossless Join Decomposition Property
  - It should be possible to reconstruct the original table
- Dependency Preserving Property
  - No functional dependency (or other constraints should get violated)



A Lossy Decomposition



Type of Normalization



13

First Normal Form (1 NF)

A relation R is said to be in the first normal form (1NF) if and only if all the attributes of the relation R are atomic in nature i.e., there does not exist any repeating group.

Customer

Customer ID	First Name	Surname	Telephone Number
123	Pooja	Singh	555-861-2025, 192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53, 182-929-2929
789	John	Doe	555-808-9633

Telephone is Multivalued

Customer

Customer ID	First Name	Surname	Telephone Number1	Telephone Number2
123	Pooja	Singh	555-861-2025	192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53	182-929-2929
789	John	Doe	555-808-9633	

11/11/2020 14

First Normal Form (1 NF)

Customer

Customer ID	First Name	Surname	Telephone Number
123	Pooja	Singh	555-861-2025
123	Pooja	Singh	192-122-1111
456	San	Zhang	182-929-2929
456	San	Zhang	(555) 403-1659 Ext. 53
789	John	Doe	555-808-9633

Customer Name

Customer ID	First Name	Surname
123	Pooja	Singh
456	San	Zhang
789	John	Doe

Customer Telephone Number

Customer ID	Telephone Number
123	555-861-2025
123	192-122-1111
456	(555) 403-1659 Ext. 53
456	182 929 2929
789	555-808-9633

Second Normal Form (2 NF)

A relation R is said to be in the second normal form (2NF) if and only if :

1. It is in 1 NF.
2. No partial dependency exists between non-key attributes and key attributes.

STUDENT-COURSE-RESULT TABLE

Student_ID	Student Name	Date Of Birth	Course-ID	Course Name	Duration in Days	Date of Exam	Marks	Grade
101	Davis	11/4/96	M2	Math	7	11/6/16	82	E
101	Davis	11/4/96	H4	History	4	22/6/16	79	A
102	Danny	6/11/97	M2	Math	7	11/6/16	62	B
102	Danny	6/11/97	P3	Phys	13	14/6/16	68	B
103	Susan	31/8/98	M1	MIS	10	20/6/16	85	E
103	Susan	31/8/98	E1	ECO	14	18/6/16	72	A
104	Sandra	25/9/97	M1	MIS	10	20/6/16	59	C
104	Sandra	25/9/97	S1	Stat	12	16/6/16	49	D

Second Normal Form (2 NF)

STUDENT TABLE

Student ID	Student Name	Date Of Birth
101	Davis	11/4/96
102	Danny	6/11/97
103	Susan	31/8/98
104	Sandra	25/9/97

COURSE TABLE

Course ID	Course Name	Duration in Days
M2	Math	7
H4	History	4
P3	Phys	13
M1	MIS	10
E1	ECO	14
S1	Stat	12

RESULT TABLE

Student ID	Course ID	Marks	Grade
101	M2	82	E
101	H4	79	A
102	M2	62	B
102	P3	68	B
103	M1	85	E
103	E1	72	A
104	M1	59	C
104	S1	49	D

EXAM_DATE TABLE

Course ID	Date of Exam
M2	11/6/16
H4	22/6/16
P3	14/6/16
M1	20/6/16
E1	18/6/16
S1	16/6/16

Third Normal Form (3 NF)

A relation R is said to be in the third normal form (3NF) if and only if :

1. It is in 2 NF.
2. No transitive dependency exists between non-key attributes and key attributes.

MARKS TABLE

Student_ID	Course-ID	Marks
101	M2	82
101	H4	79
102	M2	62
102	P3	68
103	M1	85
103	E1	72
104	M1	59
104	S1	49

MARKS_GRADE TABLE

Upper Bound	Lower Bound	Grade
100	90	O
89	80	E
79	70	A
69	60	B
59	50	C
49	40	D

Boyce Codd Normal Form (BCNF)

A relation R is said to be in the Boyce Codd Normal Form(BCNF) if and only if all the determinants are candidate keys.

RESULT TABLE

Student ID	Email-ID	Course ID	Marks
101	Davis@gmail.com	M2	82
101	Davis@gmail.com	H4	79
102	Danny@gmail.com	M2	62
102	Danny@gmail.com	P3	68
103	Susan@gmail.com	M1	85
103	Susan@gmail.com	E1	72
104	Sandra@gmail.com	M1	59
104	Sandra@gmail.com	S1	49

MARKS TABLE

Student_ID	Course-ID	Marks
101	M2	82
101	H4	79
102	M2	62
102	P3	68
103	M1	85
103	E1	72
104	M1	59
104	S1	49

STUDENT TABLE

Student ID	Email-ID
101	Davis@gmail.com
102	Danny@gmail.com
103	Susan@gmail.com
104	Sandra@gmail.com

Multivalued Dependency

In a relation R (A, B, C), consider the following assumption:

- For each value of A there are number of values B.
- For each value of A there are number of values C.
- B and C are independent to each other.
- B and C are multivalued dependent to A

Definition: In a relation R, three different attributes A, B, C such that A, B and C are subset of R. If for every value of A there exists a set of values of B and C but B and C are independent to each other then B and C are multivalued dependent on A.

EX.

EMP_Id	Project_Id	Hobby
01	01	Reading
01	02	Swimming
02	02	Swimming
02	04	Singing

EMP_Id →→ Hobby

EMP_Id →→ Project_Id

Fourth Normal Form (4 NF)

A relation in **Fourth normal form (4NF)** contains no non-trivial multi-valued dependency.

In order to achieve 4NF, all non-trivial multi-valued dependencies are to be converted in trivial multi-valued dependencies in the given relation.

Multivalued Dependencies (MVDs):

Let *R* be a relation schema and let $\alpha \subseteq R$ and $\beta \subseteq R$. The *multivalued dependency*

$$\alpha \; ì \; ì \; \beta$$

holds on *R* if in any legal relation *r(R),* for all pairs for tuples t_1 and t_2 in *r* such that $t_1[\alpha] = t_2[\alpha]$, there exist tuples t_3 and t_4 in *r* such that:

$$t_1[\alpha] = t_2[\alpha] = t_3[\alpha] = t_4[\alpha]$$
$$t_3[\beta] = t_1[\beta]$$

Fourth Normal Form (4 NF)

EMP Table

Emp ID	Project ID	Hobby
101	P01	Singing
101	P01	Reading
101	P02	Singing
101	P02	Reading
102	P03	Singing
102	P04	Reading
102	P04	Singing
102	P03	Reading

EmpID ì ì ProjectID
EmpID ì ì Hobby

This is not in 4NF

E1 Table

Emp ID	Project ID
101	P01
101	P02
102	P03
102	P04

E2 Table

Emp ID	Hobby
101	Singing
101	Reading
102	Singing
102	Reading

Join Dependency

In a relation R (A, B, C), consider the following assumption:

- B, C → A
- A, C → B
- A, B → C
- Then join dependency (JD) exists

- Definition: A join dependency (JD), denoted by JD(R1,R2, … Rn), specified on relation schema R, specifies a constraint on the states r of R.
- Natural join (R1(r),R2(r), … Rn(r)) = r
- Join dependency, multiway decomposition, results the fifth normal form (5NF)

Join Dependency & 5NF

- A MVD is a special case of a JD with n=2.
- JD(R1,R2) ➔ MVD(R1 ∩ R2) ->> R1 – R2
- MVD(R1 ∩ R2) ->> R2 – R1

- A JD is trivial if any of Ri is R.

- The 5NF is also called project-join normal form (PJNF).

Definition: A relation schema is in 5NF or project-join normal form(PJNF) w.r.t a set of F of functional, multivalued and join dependencies if, for every join dependency JD(R1,R2, … , Rn) in closure of F, every Ri is a super key of R.

Fifth Normal Form (5 NF)

A Relation R is in 5NF iff every join dependency in R is implied by the candidate keys of R.

Company Table		
Agent	Company	Product
A1	PQR	Nut
A1	PQR	Bolt
A1	XYZ	Nut
A1	XYZ	Bolt
A2	PQR	Nut

R1	
Agent	Company
A1	PQR
A1	XYZ
A2	PQR

R2	
Company	Product
PQR	Nut
PQR	Bolt
XYZ	Nut
XYZ	Bolt

R3	
Agent	Product
A1	Nut
A1	Bolt
A2	Nut

R1, R2 and R3 are in 5NF

This is not in 5 NF

Go to: https://www.youtube.com/watch?v=mbj3HSK28Kk

Note: Join decomposition is a further generalization of Multivalued dependencies. If the join of R1 and R2 over C is equal to relation R, then we can say that a join dependency (JD) exists. Where R1 and R2 are the decompositions R1(A, B, C) and R2(C, D) of a given relations R (A, B, C, D).

Thank You