Chapter 6: Normalization

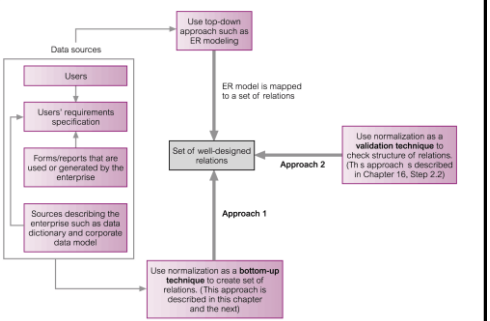
Introduction
Data Redundancy and Update Anomalies
Functional Dependencies
The Process of Normalization
Advantages and Disadvantages

Introduction

"a technique for producing a set of suitable relations that support the data requirements of an enterprise."

- developed by E.F.Codd (1972).
- 1974 - BCNF (Boyce-Codd NF) introduced by R.Boyce and E.F.Codd.
- based on :-
  a) primary key (or candidate key-BCNF)
  b) functional dependencies.
- the higher the form of normalization, the narrower the format is going to be and less vulnerable to interrupt the operation.

How Normalization Supports Database Design



Data Redundancy and Update Anomalies

- Major **aim** of relational database design is ***to group attributes into relations to minimize data redundancy.***
- Problems associated with data redundancy are illustrated by comparing the Staff and Branch relations with the StaffBranch relation.

Staff

staffNo	sName	position	salary	branchNo
SL21	John White	Manager	30000	B005
SG37	Ann Beech	Assistant	12000	B003
SG14	David Ford	Supervisor	18000	B003
SA9	Mary Howe	Assistant	9000	B007
SG5	Susan Brand	Manager	24000	B003
SL41	Julie Lee	Assistant	9000	B005

Branch

branchNo	bAddress
B005	22 Deer Rd, London
B007	16 Argyll St, Aberdeen
B003	163 Main St, Glasgow

Staff Branch

staffNo	sName	position	salary	branchNo	bAddress
SL21	John White	Manager	30000	B005	22 Deer Rd, London
SG37	Ann Beech	Assistant	12000	B003	163 Main St, Glasgow
SG14	David Ford	Supervisor	18000	B003	163 Main St, Glasgow
SA9	Mary Howe	Assistant	9000	B007	16 Argyll St, Aberdeen
SG5	Susan Brand	Manager	24000	B003	163 Main St, Glasgow
SL41	Julie Lee	Assistant	9000	B005	22 Deer Rd, London

- StaffBranch relation **has redundant data**; ***the details of a branch are repeated for every member of staff.***
- In contrast, ***the branch information appears only once for each branch in the Branch relation and only the branch number (branchNo) is repeated in the Staff relation, to represent where each member of staff is located.***
- Relations that contain redundant information may potentially suffer from update anomalies.

- Types of update anomalies include
  - **Insertion**
  - **Deletion**
  - **Modification**

STUDGRADING (Stud_No, Stud_Name, Major, Sub_Code, Sub_Name, Credit, Lect_No, Lect_Name, Grade)

Insertion Anomalies

"Inability to represent certain information"

- comes in two types :-

a) ***to insert a new record where part of the information has already exist in the table.***

example :- *to insert new record of student taking CS subject.*

{Stud_No : P3222, Stud_Name : Lisa, Major : Comp Sci, Sub_Code : CS002, Sub_Name : Comp Syst, Credit : 3, Lect_No : A123, Lect_Name : Rina}

b) ***to insert new subject that currently has no student into the table.***

example :- *there is a new subject to be introduced.*

{Sub_Code : CS003, Sub_Name : JAVA, Credit : 3}

Deletion Anomalies

"Loss of Useful Information."

- especially to the data that has one record only.
- example : *if the record of student P2115 wants to be deleted, so the other information about the Info Syst subject IS100 is lost.*
- happens when we stored unrelated information together in one table.

Modification Anomalies

"Updating record that involved more than one tuple."

- example:- *suppose the lecturer for CS001 changes from Jason to Ally but not all of the applicable tuples are updated to reflect the change.*
- the database now contains conflicting information about ***who is the lecturer for this subject***.
- if the updating is not done to all the tuples, this will lead to data inconsistency.

Functional Dependencies (FD)

"describe the relationship between attributes in a relation."

- based on FD analysis.
- involves in 1:M relationship from one attribute to a set of other attributes in a relation.
- **example** :-

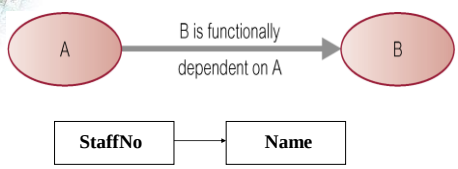
if A and B are attributes of relation R, B is functionally dependent on A, if each value of A in R associated with exactly one value of B in R.

R (A , B)

A ⟶ B

A determines B or B functionally dependent on A

- Diagrammatic representation:-



- Property of the meaning or semantics of the attributes in a relation.
- FD is determines by the attributes semantics based on the organization requirements.

- semantics → ***how the attributes relate to one another and specify the FD between attributes.***
- when the FD is present, the dependency is specified as a constraint between the attributes.
- example :-

STAFF-PROJECT(StaffNo, ProjectNo, Cost)

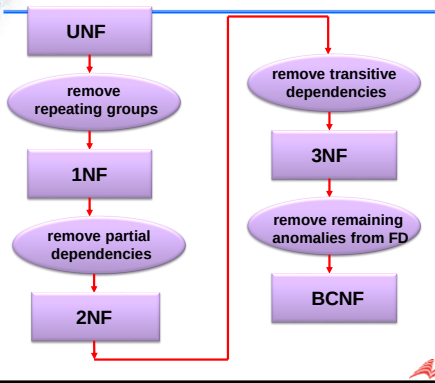
StaffNo ⟶ ProjectNo, Cost
ProjectNo⟶ Cost

- ***determinant*** *--> attribute or group of attributes on the left-hand side of the arrow.*
- ***candidate key*** *--> an attribute or combination of attributes in a relation that has a unique value for each record.*
- ***primary key*** *--> an important key and all non-key attributes FD on primary key.*

STUDGRADING Table

Stud_No	Stud_Name	Major	Sub_Code	Sub_Name	Credit	Lect_No	Lect_Name	Grade
P1050	Jerry	Comp Sci	CS001	Comp Appli	3	A121	Jason	B
P1050	Jerry	Comp Sci	CS002	Comp Syst	3	A123	Rina	A
P2115	Ahmad	Comp Sci	IS100	Info Syst	3	A120	Johan	A
P2115	Ahmad	Comp Sci	CS001	Comp Appli	3	A121	Jason	B
P2020	Sarah	Comp Sci	CS002	Comp Syst	3	A123	Rina	B

Process of Normalization



Unnormalized Form (UNF)

" a table that contains one or more repeating groups."

Repeating group = **(Sub_Code, Sub_Name, Credit, Lect_No, Lect_Name, Grade)**

- Identify the repeating group(s) in the unnormalized table which repeats for the key attribute(s).

First Normal Form (1NF)

"a relation in which the intersection of each row and column contains one and only one value."

⇨ two common approaches to remove repeating groups :-

a) ***remove repeating groups by entering appropriate data in the empty columns of rows containing the repeating data.***

b) ***placing the repeating data along with a copy of the original key attribute(s) into a separate relation.***

Student (Stud_No, Stud_Name, Major)
Stud_Subject (Stud_No, Sub_Code, Sub_Name, Credit, Lect_No, Lect_Name, Grade)

STUDGRADING Table

Stud_No	Stud_Name	Major	Sub_Code	Sub_Name	Credit	Lect_No	Lect_Name	Grade
P1050	Jerry	Comp Sci	CS001	Comp Appli	3	A121	Hamid	B
			CS002	Comp Syst	3	A123	Rina	A
P2115	Ahmad	Comp Sci	IS100	Info Syst	3	A120	Johan	A
			CS001	Comp Appli	3	A121	Hamid	B
P2020	Sarah	Comp Sci	CS002	Comp Syst	3	A123	Rina	B

Second Normal Form (2NF)

"a relation that is in 1NF and every non-primary-key attribute is fully FD on the primary key."

⇨ based on the concept of full functional dependency.

⇨ **full functional dependency** :-

if A and B are attributes of a relation, B is fully functionally dependent on A if B is functionally dependent on A but not on any proper subset of A.

⇨ fully FD A ⟶ B :-

if removal of any attribute from A results in the dependency not being sustained any more.

⇨ **partial** A⟶ B :-

if there is some attribute that can be removed from A and the dependency still holds.

⇨ applies to relations with composite keys, that is, relations with a primary key composed of two or more attributes.

⇨ 1NF --> 2NF – *involves the removal of partial dependencies by placing them in a new relation along with the determinant.*

Stud_No ⟶ Stud_Name, Major

Sub_Code ⟶ Sub_Name, Credit, Lect_No, Lect_Name

Stud_No, Sub_Code ⟶ Grade

⇨ in a form of relational database schema :-

Student (Stud_No, Stud_Name, Major)
Grading (Stud_No, Sub_Code, Grade)
Sub_Lect (Sub_Code, Sub_Name, Credit, Lect_No, Lect_Name)

Third Normal Form(3NF)

"a relation that is in 1NF and 2NF and in which no non-primary –key attribute is transitively dependent on the primary key."

⇨ **transitive dependency** :-

a condition where A, B and C are attributes of a relation such that A ⟶ B and B ⟶ C, then C is transitively dependent on A via B (provided that A is not FD on B or C).

Sub_Code ⟶ Sub_Name, Credit, Lect_No

Lect_No ⟶ Lect_Name

- in a form of relational database schema :-

Student (Stud_No, Stud_Name, Major)
Grading (Stud_No, Sub_Code, Grade)
Subject (Sub_Code, Sub_Name, Credit, Lect_No)
Lecturer (Lect_No, Lect_Name)

Advantages and Disadvantages

- **advantages** :-

a) can prevent from anomalies that cause the deletion, insertion and modification data.
b) normalized table is easier to understand.
c) can achieve appropriate logical design that can give space for additional attributes, entities set and new relations.
d) reduce redundancy of data and anomalies in database by ensuring the design is complete and consistent.

- **disadvantages** :-

a) for a larger database that involves many attributes, the normalization process will be difficult and takes time.
b) it changes the real-world data representation that happens in daily life.
c) if anomalies exist, part of the information might be lost and leads to data inconsistency in future.