Database Normalization

- Jayendra Khatod

Objectives

After completing this lesson, you should be able to do the following:

- **Understand the concept of Normalization**
- **Types of Normalization**
  - **1 NF**
  - **2 NF**
  - **3 NF**
  - **BCNF**

What is Normalization?

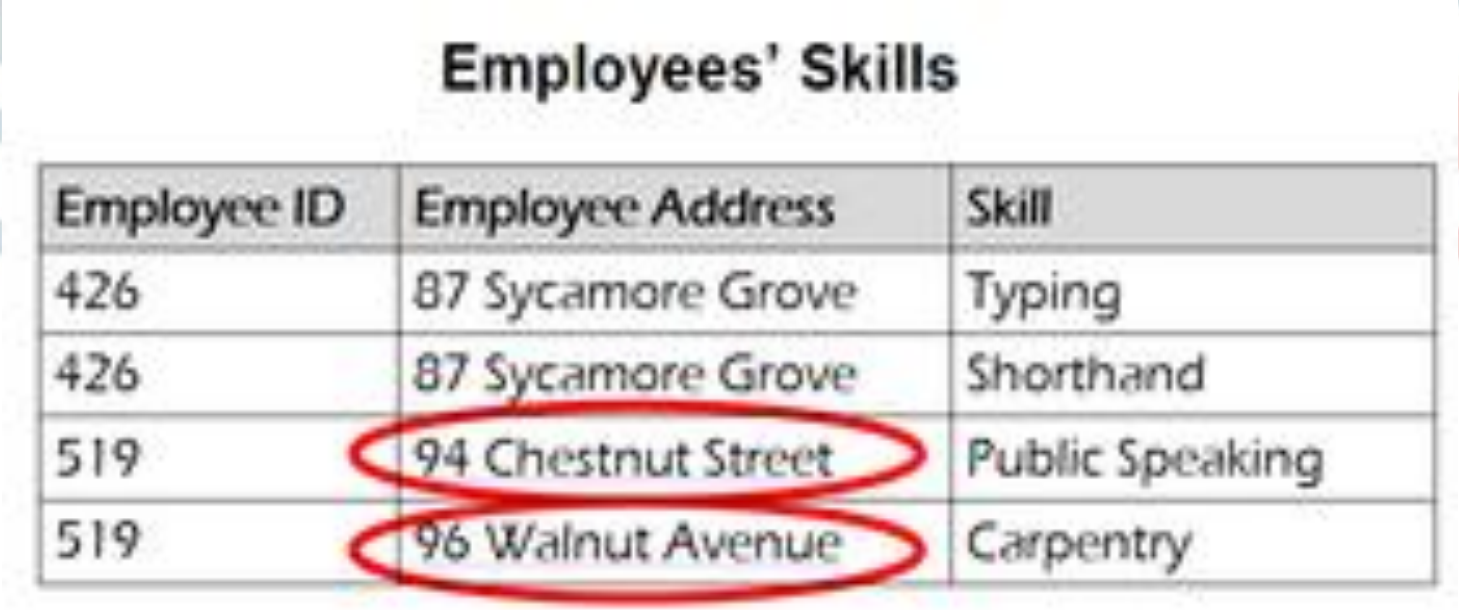
- **In the design of a relational database management system, the process of organizing data to minimize redundancy is called “Normalization”.**
- **The goal of database normalization is to decompose relations with anomalies in order to produce smaller, well-structured relations.**
- **Normalization usually involves dividing large tables into smaller (and less redundant) tables and defining relationships between them.**

History

- Edgar F. Codd, the inventor of the relational model, introduced the concept of normalization and what we now know as the First Normal Form (1NF) in 1970.
- Codd went on to define the Second Normal Form (2NF) and Third Normal Form (3NF) in 1971, and Codd and Raymond F. Boyce defined the Boyce-Codd Normal Form (BCNF) in 1974.
- Higher normal forms like 4NF, 5NF were defined by other theorists in subsequent years

What is update anomaly?

- **An update anomaly. Employee 519 is shown as having different addresses on different records.**

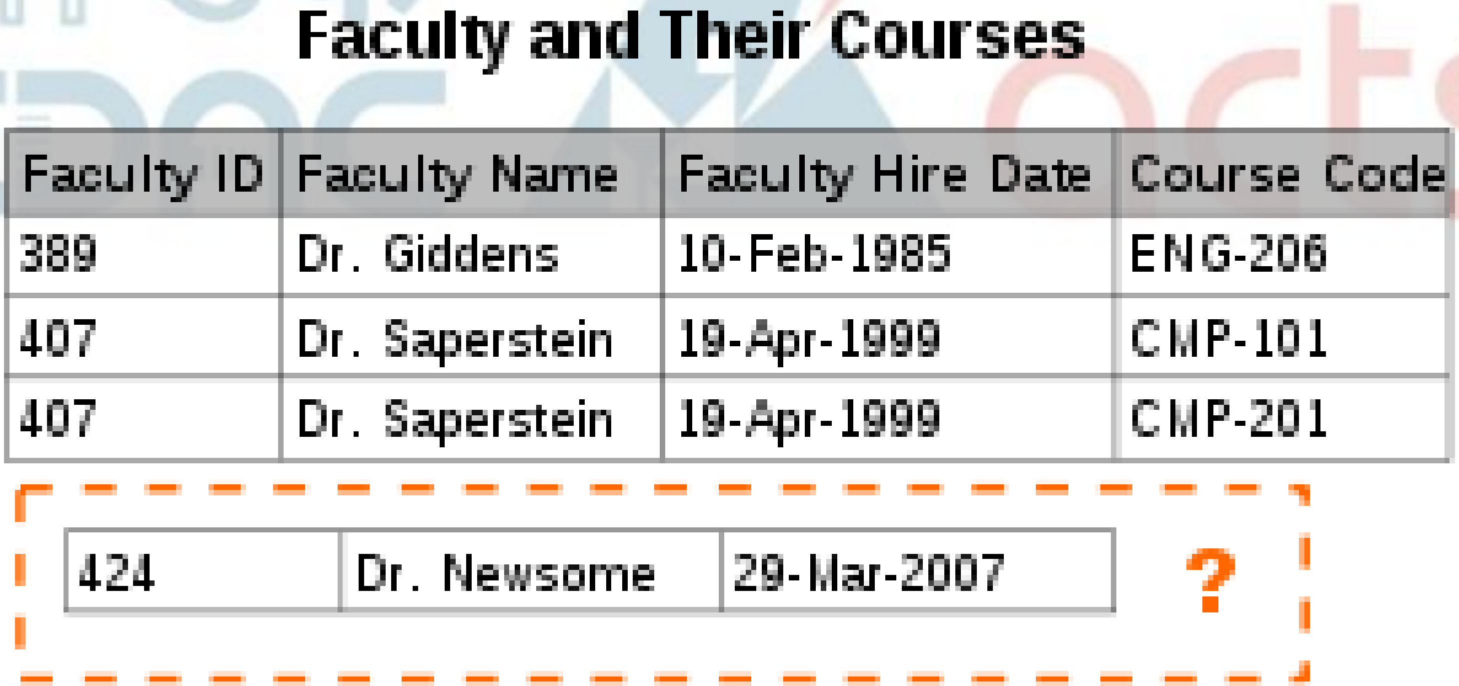


Employees' Skills

Employee ID	Employee Address	Skill
426	87 Sycamore Grove	Typing
426	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

What is Insertion anomaly?

- **An insertion anomaly. Until the new faculty member, Dr. Newsome, is assigned to teach at least one course, his details cannot be recorded**



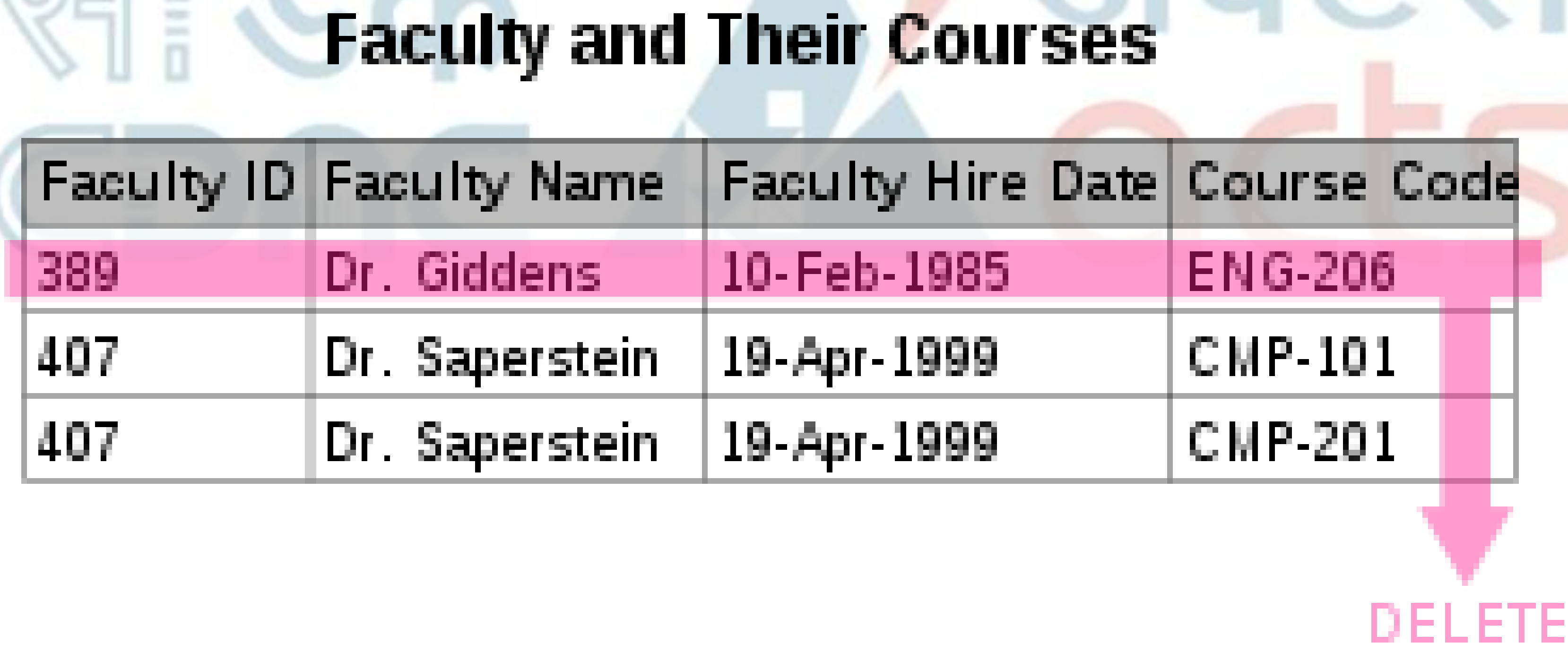
Faculty and Their Courses

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	?

What is Deletion anomaly?

- **A deletion anomaly. All information about Dr. Giddens is lost when 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

Objectives of Normalization

- A basic objective of the first normal form defined by Codd in 1970 was to permit data to be queried and manipulated using a “universal data sub-language”

  SQL is an example of such a data sub-language
- Get rid of the modification anomalies
- Minimize redesign when extending the database structure
- Make the data model more informative to users

1NF

- **A table is in 1NF if and only if it is "isomorphic to some relation", which means, specifically, that it satisfies the following five conditions:**
  - **There's no top-to-bottom ordering to the rows.**
  - **There's no left-to-right ordering to the columns.**
  - **There are no duplicate rows.**
  - **Every row-and-column intersection contains exactly one value from the applicable domain (i.e. No repeating groups).**

1NF (continued..)

- **Suppose a novice designer defines a customer table which looks like this:**

Customer			
Customer ID	First Name	Surname	Telephone Number
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659
789	Maria	Fernandez	555-808-9633

1NF (continued..)

- The designer then becomes aware of a requirement to record multiple telephone numbers for some customers.

Customer			
Customer ID	First Name	Surname	Telephone Number
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659 555-776-4100
789	Maria	Fernandez	555-808-9633

1NF (continued..)

- **Repeating groups across columns:**
  - **The designer might attempt to get around this restriction by defining multiple Telephone Number columns:**

Customer					
Customer ID	First Name	Surname	Tel. No. 1	Tel. No. 2	Tel. No. 3
123	Robert	Ingram	555-861-2025		
456	Jane	Wright	555-403-1659	555-776-4100	555-403-1659
789	Maria	Fernandez	555-808-9633		

1NF (continued..)

- **Repeating groups within columns**
  - **The designer might, alternatively, retain the single Telephone Number column but alter its domain, making it a string of sufficient length to accommodate multiple telephone numbers:**

Customer			
Customer ID	**First Name**	**Surname**	**Telephone Numbers**
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659, 555-776-4100
789	Maria	Fernandez	555-808-9633

1NF (continued..)

- **A design that complies with 1NF**
  - **A design that is unambiguously in 1NF makes use of two tables: a Customer Name table and a Customer Telephone Number table.**
  - **Repeating groups of telephone numbers do not occur in this design**

Customer Name		
<u>Customer ID</u>	**First Name**	**Surname**
123	Robert	Ingram
456	Jane	Wright
789	Maria	Fernandez

1NF (continued..)

Customer Telephone Number	
Customer ID	<u>Telephone Number</u>
123	555-861-2025
456	555-403-1659
456	555-776-4100
789	555-808-9633

2NF

- **A table that is in first normal form (1NF) must meet additional criteria if it is to qualify for second normal form.**
- **Specifically: a 1NF table is in 2NF if and only if, given any candidate key K and any attribute A that is not a constituent of a candidate key, A depends upon the whole of K rather than just a part of it**
- **In slightly more formal terms: a 1NF table is in 2NF if and only if all its non-prime attributes are functionally dependent on the whole of every candidate key.**

2NF (Continued..)

- **Consider a table describing employees' skills:**

Employees' Skills		
Employee	**Skill**	**Current Work Location**
Jones	Typing	114 Main Street
Jones	Shorthand	114 Main Street
Jones	Whittling	114 Main Street
Bravo	Light Cleaning	73 Industrial Way
Ellis	Alchemy	73 Industrial Way
Ellis	Flying	73 Industrial Way
Harrison	Light Cleaning	73 Industrial Way

2NF (Continued..)

- **Neither {Employee} nor {Skill} is a candidate key for the table**

- **Only the composite key {Employee, Skill} qualifies as a candidate key for the table.**

- **The remaining attribute, Current Work Location, is dependent on only part of the candidate key, namely Employee. Therefore the table is not in 2NF.**

2NF (Continued..)

- **A 2NF alternative to this design would represent the same information in two tables: an "Employees" table with candidate key {Employee}, and an "Employees' Skills" table with candidate key {Employee, Skill}:**

Employees	
Employee	**Current Work Location**
Jones	114 Main Street
Bravo	73 Industrial Way
Ellis	73 Industrial Way
Harrison	73 Industrial Way

2NF (Continued..)

Employees' Skills	
Employee	**Skill**
Jones	Typing
Jones	Shorthand
Jones	Whittling
Bravo	Light Cleaning
Ellis	Alchemy
Ellis	Flying
Harrison	Light Cleaning

- **Neither of these tables can suffer from update anomalies**

3NF

- **Codd's definition states that a table is in 3NF if and only if both of the following conditions hold:**
  - **The relation R (table) is in second normal form (2NF)**
  - **Every non-prime attribute of R is directly dependent on every candidate key of R.**

3NF (Continued..)

- **An example of a 2NF table that fails to meet the requirements of 3NF is:**

Tournament Winners			
Tournament	**Year**	**Winner**	**Winner Date of Birth**
Indiana Invitational	1998	Al Fredrickson	21 July 1975
Cleveland Open	1999	Bob Albertson	28 September 1968
Des Moines Masters	1999	Al Fredrickson	21 July 1975

3NF (Continued..)

- **Because each row in the table needs to tell us who won a particular Tournament in a particular Year, the composite key {Tournament, Year} is a minimal set of attributes guaranteed to uniquely identify a row. That is, {Tournament, Year} is a candidate key for the table.**
- **The breach of 3NF occurs because the non-prime attribute Winner Date of Birth is transitively dependent on the candidate key {Tournament, Year} via the non-prime attribute Winner.**
- **Since this table can contain same winner for different years and tournaments, it is subject to update anomalies.**

3NF (Continued..)

- **A 3NF alternative to this design would represent the same information in two tables:**

Tournament Winners		
Tournament	Year	Winner
Indiana Invitational	1998	Al Fredrickson
Cleveland Open	1999	Bob Albertson
Des Moines Masters	1999	Al Fredrickson
Indiana Invitational	1999	Chip Masterson

3NF (Continued..)

Player Dates of Birth	
Player	Date of Birth
Chip Masterson	14 March 1977
Al Fredrickson	21 July 1975
Bob Albertson	28 September 1968

Update anomalies cannot occur in these tables, which are both in 3NF.

Boyce–Codd Normal Form

- **Boyce–Codd normal form is a slightly stronger version of the third normal form.**
- **BCNF was developed in 1974 by Raymond F. Boyce and Edgar F. Codd to address certain types of anomaly not dealt with by 3NF as originally defined.**
- **Only in rare cases does a 3NF table not meet the requirements of BCNF**
- **A 3NF table which does not have multiple overlapping candidate keys is guaranteed to be in BCNF.**

Boyce–Codd Normal Form (Continued..)

- **An example of a 3NF table that does not meet BCNF is:**

Today's Court Bookings			
Court	Start Time	End Time	Rate Type
1	09:30	10:30	SAVER
1	11:00	12:00	SAVER
1	14:00	15:30	STANDARD
2	10:00	11:30	PREMIUM-B
2	11:30	13:30	PREMIUM-B
2	15:00	16:30	PREMIUM-A

Boyce–Codd Normal Form (Continued..)

- Each row in the table represents a court booking at a tennis club that has one hard court (Court 1) and one grass court (Court 2)
- A booking is defined by its Court and the period for which the Court is reserved
- Additionally, each booking has a Rate Type associated with it. There are four distinct rate types:
  - SAVER, for Court 1 bookings made by members
  - STANDARD, for Court 1 bookings made by non-members
  - PREMIUM-A, for Court 2 bookings made by members
  - PREMIUM-B, for Court 2 bookings made by non-members

Boyce–Codd Normal Form (Continued..)

- **The design can be amended so that it meets BCNF:**

Rate Types		
Rate Type	**Court**	**Member Flag**
SAVER	1	Yes
STANDARD	1	No
PREMIUM-A	2	Yes
PREMIUM-B	2	No

Boyce–Codd Normal Form (Continued..)

Today's Bookings		
Rate Type	Start Time	End Time
SAVER	09:30	10:30
SAVER	11:00	12:00
STANDARD	14:00	15:30
PREMIUM-B	10:00	11:30
PREMIUM-B	11:30	13:30
PREMIUM-A	15:00	16:30

Having one Rate Type associated with two different Courts is now impossible, so the anomaly affecting the original table has been eliminated.

Summary

Normal form	Brief definition
First normal form (1NF)	Table faithfully represents a relation and has no *repeating groups*
Second normal form (2NF)	No non-prime attribute in the table is functionally dependent on a proper subset of a candidate key
Third normal form (3NF)	Every non-prime attribute is non-transitively dependent on every candidate key in the table
Boyce–Codd normal form (BCNF)	Every non-trivial functional dependency in the table is a dependency on a super key

Thank You !