

Data Normalization

Contents

- Learning objectives
- Introducing to some problems
- Let's Normalize

Learning objectives

- WHY breaking the table is essential?
- Have an understanding of various normal forms in a database.

Introducing to some problems

We will start with a CSV file or an excel spreadsheet. This is the most common way of storing data in various organizations. Often people organize it messy. Checkout this CSV file

Stud NO	Stud Name	Major	Course NO	Course Name	Grade	Grade Type
111	Catherine John	MBAN	580, 518, 516	TPCS INFO TECH, Data Visualization, Health and Technology	A,A,A	Excellent, Excellent, Excellent
222	Amir Khan	MDS	525,516	Web and Cloud Computing, Health and Technology	A,B	Excellent, Good
333	Gittu George	MBAN	525	Web and Cloud Computing	C	Average

This looks okay now as it's just 3 rows, but think about how it will look if we are going to store entire students in a university. Here what are the problems that come to your mind? Here are a few of them

- ***Does the order of courses in a course column matter?***

E.g., For Amir Khan, the order for course number is 525,516. Does this mean that the order in which he took classes? How many courses are a student supposed to take? We see Catherine, Amir, Gittu taking 3, 2, and 1 class, respectively. Is there a minimum number of courses that a student needs to take? Is there any maximum limit?

- ***Too many duplicates ?***

- ***What happens if I am loading incoming students to the database ?***

They won't be taking any classes until they start, which can result in lots of null values.

- ***Human errors can happen.***

Assume what happens if I accidentally entered 525a as the course number?

[Skip to main content](#)

Deleting a student can end up deleting course information. E.g., deleting Catherine can take out the course information for course numbers 580, 518.

• **Update anomaly**

We plan to change the course name for 525 from “Web and cloud computing” to “Cloud computing for Big Data”! How many places need to be updated?

• **Insertion anomaly**

We plan to launch a new course named 999 named “Big Data Systems,” and no students are currently enrolled in it. So how are we going to store this information?

Loading this data into a database table can bring in problems mentioned above, resulting in serious data integrity issues, and hence we should do data normalization.

Let’s Normalize

Normalization is the process of minimizing redundancy from a table of relations. These redundancies in relation may cause insertion, deletion, and update anomalies. Normal forms are a certain set of defined rules to “break a table smartly” to create normalized tables.

Storytime:

Ted Codd introduced the concept of normalization and first normal form in his [paper](#) in 1970, which provides theoretical foundation of designing a database. After a year, he defined 2nd and 3rd normal form in [this](#) paper. Later in 1974, Codd and Raymond F. Boyce defined the Boyce–Codd normal form (BCNF). These 4 normal forms are considered the most common or important ones; Most database systems are designed to follow the 3rd normal form or BCNF.

Later came some more restrictive normal forms (4NF,5NF,6NF), but these are mainly used for theoretical purposes and are less implemented in industries.

Note: We explain the normal forms in an easier-to-understand fashion, and they can also be explained using theoretical formulas and derivations.

First normal form (1NF)


To be in 1NF,

- All types must be atomic.❌
- No repeating groups✅

Stud NO	Stud Name	Major	Course NO	Course Name	Grade	Grade Type
111	Catherine John	MBAN	580, 518, 516	TPCS INFO TECH, Data Visualization, Health and Technology	A,A,A	Excellent, Excellent, Excellent
222	Amir Khan	MDS	525,516	Web and Cloud Computing, Health and Technology	A,B	Excellent, Good
333	Gittu George	MBAN	525	Web and Cloud Computing	C	Average

- All types must be atomic✅

[Skip to main content](#)



Stud NO	Stud FName	Stud LName	Major	Course NO 1	Course NO 2	Course NO 3	Course Name 1	Course Name 2	Course Name 3
111	Catherine	John	MBAN	580	518	516	TPCS INFO TECH	Data Visualization	Health and Technology
222	Amir	Khan	MDS	525	516		Web and Cloud Computing	Health and Technology	
333	Gittu	George	MBAN	525			Web and Cloud Computing		

- All types must be atomic. ✓
- No repeating groups ✓

Stud NO	Stud FName	Stud LName	Major	Course NO	Course Name	Grade	Grade Type
111	Catherine	John	MBAN	580	TPCS INFO TECH	A	Excellent
111	Catherine	John	MBAN	518	Data Visualization	A	Excellent
111	Catherine	John	MBAN	516	Health and Technology	A	Excellent
222	Amir	Khan	MDS	525	Web and Cloud Computing	A	Excellent
222	Amir	Khan	MDS	516	Health and Technology	B	Good
333	Gittu	George	MBAN	525	Web and Cloud Computing	C	Average

Second normal form (2NF)

To be in 2NF

- Must be in first normal form ✓
- Any non-key attributes fully depend on the candidate key ✗

Stud NO	Stud FName	Stud LName	Major	Course NO	Course Name	Grade	Grade Type
111	Catherine	John	MBAN	580	TPCS INFO TECH	A	Excellent
111	Catherine	John	MBAN	518	Data Visualization	A	Excellent
111	Catherine	John	MBAN	516	Health and Technology	A	Excellent
222	Amir	Khan	MDS	525	Web and Cloud Computing	A	Excellent
222	Amir	Khan	MDS	516	Health and Technology	B	Good
333	Gittu	George	MBAN	525	Web and Cloud Computing	C	Average

- Must be in first normal form ✓
- Any non-key attributes fully depend on the candidate key ✓

[Skip to main content](#)

Stud NO	Stud FName	Stud LName	Major
111	Catherine	John	MBAN
222	Amir	Khan	MDS
333	Gittu	George	MBAN

Stud NO	Course NO	Grade	Grade Type
111	580	A	Excellent
111	518	A	Excellent
111	516	A	Excellent
222	525	A	Excellent
222	516	B	Good
333	525	C	Average

Course NO	Course Name
580	TPCS INFO TECH
518	Data Visualization
516	Health and Technology
525	Web and Cloud Computing

Third normal form

To be in 3NF

- Must be in second normal form ✓
- There should be no transitive dependency ✗

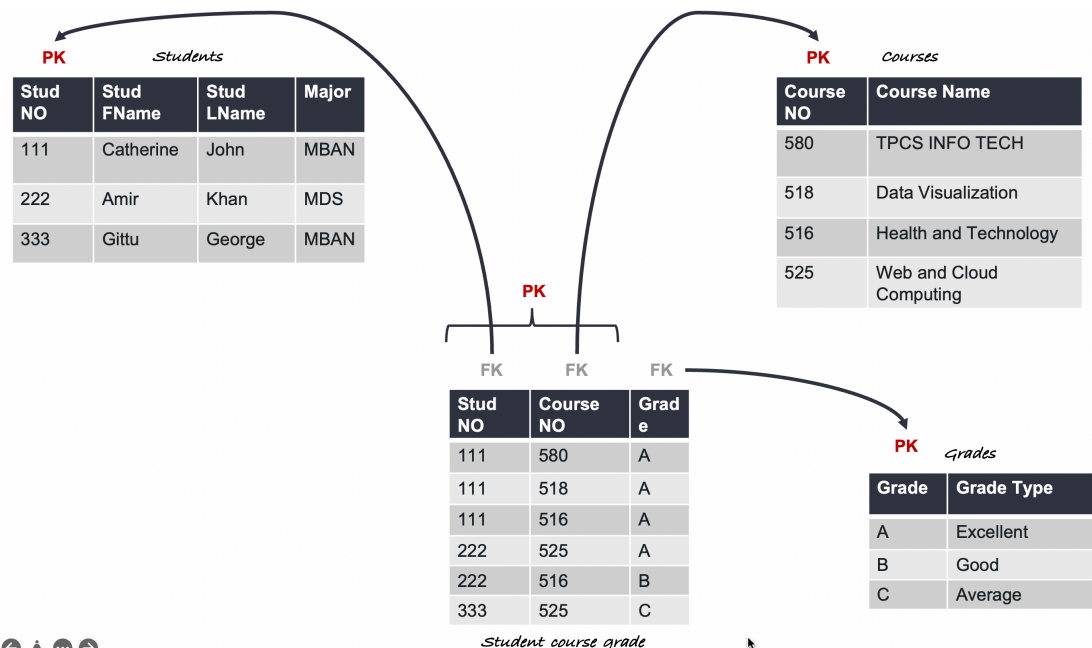
Stud NO	Stud FName	Stud LName	Major
111	Catherine	John	MBAN
222	Amir	Khan	MDS
333	Gittu	George	MBAN

Stud NO	Course NO	Grade	Grade Type
111	580	A	Excellent
111	518	A	Excellent
111	516	A	Excellent
222	525	A	Excellent
222	516	B	Good
333	525	C	Average

Course NO	Course Name
580	TPCS INFO TECH
518	Data Visualization
516	Health and Technology
525	Web and Cloud Computing



- Must be in second normal form ✓
- There should be no transitive dependency ✓



Now we have finished normalizing the table, let's create this table's in our database.