

Database Normalisation

Michael Jacobs

2023-07-11

What is normalisation? Why bother?

Normalisation is a process by which data structures are made as efficient as possible. . .

- Eliminating redundancy
- Minimising use of null values
- Avoiding update anomalies

1st, 2nd, and 3rd Normal Form

First Normal Form



Second Normal Form



Third Normal Form



First Normal Form (1NF)

A table is in 1NF if:

- The table stores information in rows and columns where one or more columns (called the **primary key**) uniquely identify each row.
- Each column contains atomic values, and there are not repeating groups of columns.

Non-normalised table

Non-normalised table

Name	Class	Grade	Faculty	FacultySize
William	Machine learning, Survey design	A, B	ME, ME	212, 212
Sarah	Statistics, Applied Maths	A, A+	ST, ST	133, 133
Tom	Applied Maths, Machine learning	C, C	ST, ME	133, 212

Why does this violate 1NF?

Here is converted to 1NF

Student-Class table

StudentID	ClassID	Grade	Name	Class	Faculty	FacultySize
001	ME314	A	William	Machine learning	ME	212
001	ME202	B	William	Survey design	ME	212
002	ST001	A	Sarah	Statistics	ST	133
002	ST203	A+	Sarah	Applied Maths	ST	133
003	ST203	B	Tom	Applied Maths	ST	133
003	ME314	C	Tom	Machine learning	ME	212

Primary key = {StudentID, ClassID}

Second Normal Form (2NF)

A table is in 2NF if:

- It is in 1NF
- Every **non-prime attribute** depends on **the whole primary key** (and not a subset of it).

Some useful definitions:

- A non-prime attribute is an attribute that is **not** part of the primary key.
- To say that X depends on Y means that knowing Y is sufficient to know X.

Why is this not in 2NF?

Student-Class table

StudentID	ClassID	Grade	Name	Class	Faculty	FacultySize
001	ME314	A	William	Machine learning	ME	212
001	ME202	B	William	Survey design	ME	212
002	ST001	A	Sarah	Statistics	ST	133
002	ST203	A+	Sarah	Applied Maths	ST	133
003	ST203	B	Tom	Applied Maths	ST	133
003	ME314	C	Tom	Machine learning	ME	212

Why is this not in 2NF?

Student-Class table

StudentID	ClassID	Grade	Name	Class	Faculty	FacultySize
001	ME314	A	William	Machine learning	ME	212
001	ME202	B	William	Survey design	ME	212
002	ST001	A	Sarah	Statistics	ST	133
002	ST203	A+	Sarah	Applied Maths	ST	133
003	ST203	B	Tom	Applied Maths	ST	133
003	ME314	C	Tom	Machine learning	ME	212

- Name depends on StudentID
- Class, Faculty and FacultySize depend on ClassID

Here it is converted to 2NF

Student-class table

StudentID	ClassID	Grade
001	ME314	A
001	ME202	B
002	ST001	A
002	ST203	A+
003	ST203	B
003	ME314	C

Student table

StudentID	Name
001	William
002	Sarah
003	Tom

Class table

ClassID	Class	Faculty	FacultySize
ME314	Machine learning	ME	212
ME202	Survey Design	ME	212
ST001	Statistics	ST	133
ST203	Applied Maths	ST	133

Third Normal Form (3NF)

A table is in 3NF if:

- It is in 2NF.
- No non-prime attribute is transitively dependent on prime key attribute.

NB: An attribute is *transitively* dependent on a prime key attribute if it can be identified from another non-prime attribute

Why is this not in 3NF?

Class table

ClassID	Class	Faculty	FacultySize
ME314	Machine learning	ME	212
ME202	Survey Design	ME	212
ST001	Statistics	ST	133
ST203	Applied Maths	ST	133

Why is this not in 3NF?

Class table

ClassID	Class	Faculty	FacultySize
ME314	Machine learning	ME	212
ME202	Survey Design	ME	212
ST001	Statistics	ST	133
ST203	Applied Maths	ST	133

- We can identify the FacultySize from the Faculty (but the primary key is ClassID).
- Transitive dependency: ClassID \rightarrow Faculty \rightarrow FacultySize
- Solution: Create yet another table (a Faculty table).

Here it is converted to 3NF

Student-class table

StudentID	ClassID	Grade
001	ME314	A
001	ME202	B
002	ST001	A
002	ST203	A+
003	ST203	B
003	ME314	C

Class table

ClassID	Class	Faculty
ME314	Machine learning	<i>ME</i>
ME202	Survey Design	<i>ME</i>
ST001	Statistics	<i>ST</i>
ST203	Applied Maths	<i>ST</i>

Student table

StudentID	Name
001	William
002	Sarah
003	Tom

Faculty table

Faculty	FacultySize
ME	212
ST	133

Cheat sheet

First Normal Form

- The table stores information in rows and columns where one or more columns (called the primary key) uniquely identify each row.
- Each column contains atomic values, and there are not repeating groups of columns.

Second Normal Form

- It is in 1NF
- Every non-prime attribute depends on the whole primary key (and not a subset of it).

Third Normal Form

- It is in 2NF.
- No non-prime attribute is transitively dependent on prime key attribute.