Database Concepts

A **relation** is a mathematical term for a table.

A relation has **attributes** which are columns or fields.

A relation has **tuples** which are rows or records.

The table name is referred to as an **entity**.

Degree is the number of attributes in a relation.

Cardinality is the number of tuples or rows.

Types of SQL Statements

There are 5 types of SQL statements, namely:

- DDL – Data Definition Language.
- DQL – Data Query Language.
- DML – Data Manipulation Language.
- DCL – Data Control Language.
- TCL – Transaction Control Language.

Note: This tutorial covers DDL and DML only.

Data Definition Language (DDL) Statements.

This type of SQL statement is used to define, change, or drop database objects such as tables.

Common DDL Statements:

CREATE - Create tables and define their columns.

ALTER - Altering tables including adding and dropping columns.

TRUNCATE - Delete data in a table but not the table itself.

DROP - Delete table.

Data Manipulation Language (DML) Statements.

This type of SQL statement is used to read and modify data in tables. These are also known as CRUD (Create, Read, Update, and Delete) operations.

Common DML Statements:

INSERT - Insert one or many more rows of data into a table.

SELECT - Reads a row or rows of data from a table.

UPDATE - Edits data in a row or rows of a table.

DELETE - Removes a row or rows of data from a table.

Normalization

When keeping records (data) one will inevitably have inconsistencies and duplicate data.

Such duplication can cause extra work when the data is updated because you must change it in multiple places (more than one place).

Normalization is the process of organizing data to reduce redundant data often by dividing large, monolithic tables into smaller multiple related tables.

There are several forms of normalization. Data professionals ought to be familiar with:

1. First Normal Form (1NF)
2. Second Normal Form (2NF)
3. Third Normal Form (3NF)

First Normal Form (1NF)

For a table to be in 1NF:

- Each row must be unique.
- Each cell must contain only a single value.

book_id	title	format	author_name
101	Lean Software Development	Paperback	Mary Poppendieck
201	Facing the Intelligence Explosion	Hardback	David Robson
301	Scala in Action	Hardback	Yehuda Kalz
401	Patterns of Software	Hardback, Paperback	Mary Poppendieck
501	Anatomy of LISP	Paperback	Eric Redmond

book_id	title	format	author_name
101	Lean Software Development	Paperback	Mary Poppendieck
201	Facing the Intelligence Explosion	Hardback	David Robson
301	Scala in Action	Hardback	Yehuda Kalz
401	Patterns of Software	Hardback, Paperback	Mary Poppendieck
501	Anatomy of LISP	Paperback	Eric Redmond

Table in 1NF after normalization

book_id	title	format	author_name
101	Lean Software Development	Paperback	Mary Poppendieck
201	Facing the Intelligence Explosion	Hardback	David Robson
301	Scala in Action	Hardback	Yehuda Kalz
401	Patterns of Software	Paperback	Mary Poppendieck
401	Patterns of Software	Hardback	Mary Poppendieck
501	Anatomy of LISP	Paperback	Eric Redmond

Second Normal Form (2NF)

For a database to be in 2NF:

- The database must be in 1NF.
- Must separate groups of values that belong to multiple rows by creating new tables.

book_id	title	format	author_name
101	Lean Software Development	Paperback	Mary Poppendieck
201	Facing the Intelligence Explosion	Hardback	David Robson
301	Scala in Action	Hardback	Yehuda Kalz
401	Patterns of Software	Paperback	Mary Poppendieck
401	Patterns of Software	Hardback	Mary Poppendieck
501	Anatomy of LISP	Paperback	Eric Redmond

book_id	title	format	author_name
101	Lean Software Development	Paperback	Mary Poppendieck
201	Facing the Intelligence Explosion	Hardback	David Robson
301	Scala in Action	Hardback	Yehuda Kalz
401	Patterns of Software	Paperback	Mary Poppendieck
401	Patterns of Software	Hardback	Mary Poppendieck
501	Anatomy of LISP	Paperback	Eric Redmond

Book ID 401 is listed 2x, hence there is some data duplication.

To meet the requirements for second normal form (2NF), and achieve just one row for book 401, split the book table so that format information of the book is separated from unrelated information, such as table author.

books

book_id (Primary key)	title	author_name
101	Lean Software Development	Mary Poppendieck
201	Facing the Intelligence Explosion	David Robson
301	Scala in Action	Yehuda Kalz
401	Patterns of Software	Mary Poppendieck
501	Anatomy of LISP	Eric Redmond

formats

book_id (Foreign key)	format
101	Paperback
201	Paperback
301	Hardback
401	Hardback
401	Paperback
501	Paperback

Note: Each resulting table is in 1NF. Also, to maintain a relationship between the tables, identify a primary key for 1 table that will be used as a foreign key in the other table.

Third Normal Form (3NF)

For a database to be in 3NF:

- The database must be in 2NF.
- Must eliminate columns that do not depend on the key.

book_id (Primary key)	title	author_name	publisher	ship_from
101	Lean Software Development	Mary Poppendieck	Tech Books	UK
201	Facing the Intelligence Explosion	David Robson	Amazing Books	US
301	Scala in Action	Yehuda Kalz	Better Tech Books	India
401	Patterns of Software	Mary Poppendieck	Publisher 101	US
501	Anatomy of LISP	Eric Redmond	Best Tech Books	Canada

Each publisher ships books from warehouses in their location, so where the book ships from depends on the publisher not the book ID. Therefore, the book table is not in 3NF, because the ships from the data do not depend on the primary key (book_id).

books

book_id (Primary key)	title	author_name	pub_id
101	Lean Software Development	Mary Poppendieck	1
201	Facing the Intelligence Explosion	David Robson	2
301	Scala in Action	Yehuda Kalz	3
401	Patterns of Software	Mary Poppendieck	4
501	Anatomy of LISP	Eric Redmond	5

publishers

pub_id (Primary key)	publisher	ship_from
1	Tech Books	UK
2	Amazing Books	US
3	Better Tech Books	India
4	Publisher 101	US
5	Best Tech Books	Canada

In [ ]: