Unit 7 Normalization

A database allows for a structured approach to data storage. Databases are the foundation of most modern services. Data normalization techniques can help guide developers on how best to structure databases for best effect.

Recap

What is a database composed of?

A Database management system is the software that allows someone to interact with the data base.

A database store:

1. Raw values
2. Information describing the data format (which provides meaning)

A database stores objects and their attributes as well as the relationships between the objects. These components can be described as:

1. Entity's
2. Attributes
3. Relationships

Keys

A **primary key** is an attribute that can be used to uniquely identify every record in a table.

A **composite key** is a unique key which is composed of multiple attributes.

A **foreign key** is attribute which references a record in another table by a shared attribute.

Data Normalization

Data normalization aims to reduce data redundancy and maximize data integrity. Normalization involves examining the attributes of different tables.

Normalization can be used to group attributes and create a good set of entity's and their relationships.

Characteristics of a good set of entity relations:

- Minimal number of attributes per entity
- Attributes are grouped with close logical relationships
- Minimal redundancy
- Only primary keys should be repeated (as foreign keys)

Anomaly's

Redundant information produces a number of harmful anomaly's when the user interacts with the database.

Insertion Anomaly's

- Unrelated information is required for record entry
- NULL values must be entered if redundant information is not available.

Deletion Anomaly's

- Bad 'coupling' of information in the same row means that deleting one type of information can cause another type of information to be deleted.

Update Anomaly's

- Multiple rows need to be updated to change one piece of information
- It can be impossible to distinguish between entries with the same value as 'entity's' do not have unique ids.

Problems with data redundancy

Data redundancy occurs when a piece of information is stored in multiple places. Data redundancy can create the need for an excessive number of operations (I.e. SQL queries) to be required for a simple task. As it is harder to update all instances of a piece of data across many places rather than a single source, there is a greater chance of inconstancies being introduced. I like the term concept of having a 'single source of truth' for information in a system.

Normalization Usages

The technique of normalization can be used to help design databases or to validate the structure of existing databases.

Dependency's

Functional dependency

If one column is dependent on another column then then this is a functional dependency.

For example, a person's name is dependent on the persons id (their unique identifier).

Transitive Dependency

If an attribute is not directly dependent on a column it can transitively dependent if it is dependent on a column which is itself functionally dependent on that column. A column that is transitively dependent on another column is also functionally dependent on that column.

The Normal Forms

1st Normal form

Every cell should only contain one value. If a cell only contains one value it is called an atomic value. For example a cell should not contain a comma separated list of values.

2nd Normal from

All attributes should be fully functionally dependent on the primary key (can be a composite key).

3rd normal form

No attributes should be only transitively dependent on the primary key.