Normalisation

Dr Peadar Grant

November 26, 2024

Contents

1 Required reading

1. Wikipedia article on database normalisation.
2. Textbook [Connolly and Begg, 2015] Chapter 14 (Normalisation).
3. Original paper by Codd.

2 Normalisation

Connolly and Begg [2015] defines normalisation as:

> A technique for producing a set of relations with desirable properties, giving the data requirements of an enterprise.

2.1 Desireable properties

- Minimise the number of attributes
- Attributes with close logical relationship (functionally dependent) should be in same relation (table)
- Minimal redundancy: **data represented once**:
  - except for enabling joins on primary keys / foreign keys.

2.2 Benefits

- Minimal redundancy means updates affect fewer attributes and relations meaning reduced number of operations.
- Reduction in disk storage.
- Avoidance of **anomalies** / inconsistencies.

3 Anomalies

Redundant data causes anomalies where data manipulation operations are not applied to all instances of the data.

This is often seen when ad-hoc solutions (e.g. Excel) are used to store data that should be stored in a relational or other database.

3.1 Update anomaly

An update anomaly arises where an update does not modify all instances of specific data, .

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

Figure 1: Update anomaly (Wikipedia)

3.2 Insertion anomaly

An insertion anomaly arises where an insert causes either nulls to be required to create a dummy record, . An insert that causes inconsistency with existing data also is identified as an insertion anomaly.

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	?

Figure 2: Insertion anomaly (Wikipedia)

3.3 Deletion anomaly

A deletion anomaly arises where data is lost about one entity when another entity is deleted,
.

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

DELETE

Figure 3: Deletion anomaly (Wikipedia)

4 Functional dependency

Definition:

> If A and B are attributes of a relation R, B functionally depends on A if every value of A is associated with one value of B.

Note:

- Can say A functionally determines B.
- Would say A is the determinant in the relationship

4.1 Full / partial functional dependency

Let A and B be attributes of a relation R:

Full functional dependency is where B functionally depends on A, but not any proper subset of A.

Partial functional dependency exists where B functionally depends on A and on one or more proper subsets of A.

4.2 Characteristics

When normalising, functional dependencies should:

1. Have 1:1 relationship between determinant and the functional dependent.
2. Hold for all time.
3. Be full (not partial), i.e. have minimum number of attributes necessary

4.3 Transitive dependency

Let A, B, C be attributes of a relation R. Assume that A isn't functionally dependent on B or C.

If $A \Rightarrow B$ and $B \Rightarrow C$ then C is transitively dependent on A.

5 Primary key

A primary key must be UNIQUE and NOT NULL.

5.1 Simple

Simple primary key is one that involves only ONE column.

Any column directly defined as a `PRIMARY KEY` will be a simple primary key.

5.2 Compound

A compound key is a primary key composed of more than 1 column.

5.3 Candidate keys

In practice there may be a number of possible columns and groups thereof that could be used as a primary key.

Each column or combination of columns is termed a candidate key.

6 Normal forms

6.1 First Normal Form (1NF)

The first normal form (1NF) requires **atomic values** (i.e. no repeating groups or multi-valued attributes) for each column in each row of a table.

To bring a table to 1NF we must eliminate:

1. Non-atomic values such as lists (e.g. tags separated by a comma).
2. Multiple columns, e.g. `tag1`, `tag2`.

Approaches:

1. Flattening table: duplicate rows
    - Vulnerable to anomalies
2. Separate relations to represent repeating data

- Most flexible form

6.2 Second Normal Form (2NF)

2NF eliminates partial dependencies on the primary key (or any other candidate key).

2NF requires that:

- The relation is in 1NF
- every non-candidate-key attribute is fully functionally dependent on any candidate key.

A 1NF relation will be automatically 2NF if its primary key is a simple (non-compound key).

6.3 Third Normal Form (3NF)

3NF eliminates transitive dependencies on the primary key (or any other candidate key).
3NF requires that:

- The relation is in both 1NF and 2NF
- No non-candidate-key attribute is transitively dependent on any candidate key

A relation in 2NF can be transformed to 3NF by removing transitive dependencies.

References

Thomas M Connolly and Carolyn E Begg. *Database Systems: A Practical Approach to Design, Implementation and Management*. Pearson, 6th edition, 2015.