Higher Normalization

by Sophia

WHAT'S COVERED

In this lesson, you will explore the higher levels of normalization. These are known as the Boyce-Codd normal form, fourth normal form (4NF), and fifth normal form (5NF). It does so in four parts. Specifically, this lesson covers:

1. Introduction

With transactional databases, the third normal form (3NF) is generally the furthest you would take normalization. It does not make sense to go beyond it in most situations. There are certain cases, though, where additional normalization could be useful.

2. Boyce-Codd Normal Form

The **Boyce-Codd normal form (BCNF)** is a higher level of database normalization designed to eliminate certain anomalies that may still exist in databases normalized up to the third normal form (3NF). The BCNF is named after Raymond F. Boyce and Edgar F. Codd, who contributed independently to its development.

A table must satisfy two conditions in order to achieve BCNF:

- Determinants (groups of attributes that determine another attribute uniquely) must be **candidate keys**.
- The non-key attributes must be fully functionally dependent on the primary key as a whole, without any partial dependencies.

KEY CONCEPT

The concept of candidate keys helps in the process of designing a well-structured and normalized database schema.

A candidate key is a set of one or more attributes whose values can be used to uniquely identify a row in a table. For a set of attributes to be considered a candidate key, it must satisfy two properties:

- Uniqueness: Each combination of values in the candidate key must be unique across all the rows in the table. No two rows can have the same values for the candidate key.
- Irreducibility (Minimality): No proper subset of the candidate key should have the uniqueness property. In other words, if any attribute is removed from the candidate key, it should no longer be able to uniquely identify each row.

In a table, there can be multiple candidate keys. The primary key is the candidate key that is chosen to be the main method of uniquely identifying records in the table. Other candidate keys are referred to as **alternate keys**.

Candidate keys matter because none of them should be functionally dependent on another one. A candidate key depends on a part, but only part, of the attribute when there are non-trivial functional dependencies.

A table that meets these conditions will remain free of insertion, update, and deletion anomalies and will not contain any redundancy based on functional dependencies.

In order to achieve BCNF, large tables may need to be broken down into multiple smaller ones, and they may need to be linked together. As BCNF provides a higher level of normalization than 3NF, it may lead to more complex data models, which may affect query performance. This requires careful consideration of the trade-off between data integrity and performance.

Boyce-Codd Normal Form

The Boyce-Codd normal form, also known as BCNF, is viewed as 3.5NF. BCNF requires that the database design first fulfills the requirements of the third normal form (3NF) but also has every determinant in a table as a candidate key.

Candidate Key

A set of one or more attributes that can uniquely identify each entry in a table.

Alternate Key

A candidate key that has not been selected as the primary key or part of the composite key.

3. Fourth Normal Form

A **fourth normal form (4NF)** builds upon the principles of the third normal form (3NF) in order to normalize databases at a higher level. Databases normalized up to 3NF can have certain multivalued dependencies. By eliminating multivalued dependencies between non-key attributes, 4NF eliminates data redundancy and anomalies.

To achieve 4NF, non-key attribute should be functionally dependent on a set of attributes not part of the candidate key.

It means that all non-key attributes must be functionally dependent on the whole candidate key, and there should be no multivalued dependencies. A table with multivalued dependencies indicates redundant information, which 4NF aims to eliminate.

4NF may require splitting the table into multiple smaller tables to eliminate multivalued dependencies. In addition to making a database design more efficient and concise, normalization promotes data integrity. The data manipulation process is also made more efficient by reducing anomalies.

The 4NF approach ensures that the database schema is free from certain types of redundancy, and that each table contains only essential and non-redundant information, making the database easier to manage, scalable, and maintain over time.

TERM TO KNOW

Fourth Normal Form (4NF)

A database design that satisfies all of the properties of the third normal form (3NF) and Boyce-Codd normal form (BCNF), and additionally should not have any multivalued dependencies.

4. Fifth Normal Form

Fifth normal form (5NF), also known as project-join normal form (PJNF), is the highest level of database normalization. It addresses certain join dependencies in databases normalized to 4NF by extending the principles of 4NF. With 5NF, undesirable join dependencies between non-key attributes are eliminated, which reduces data redundancy.

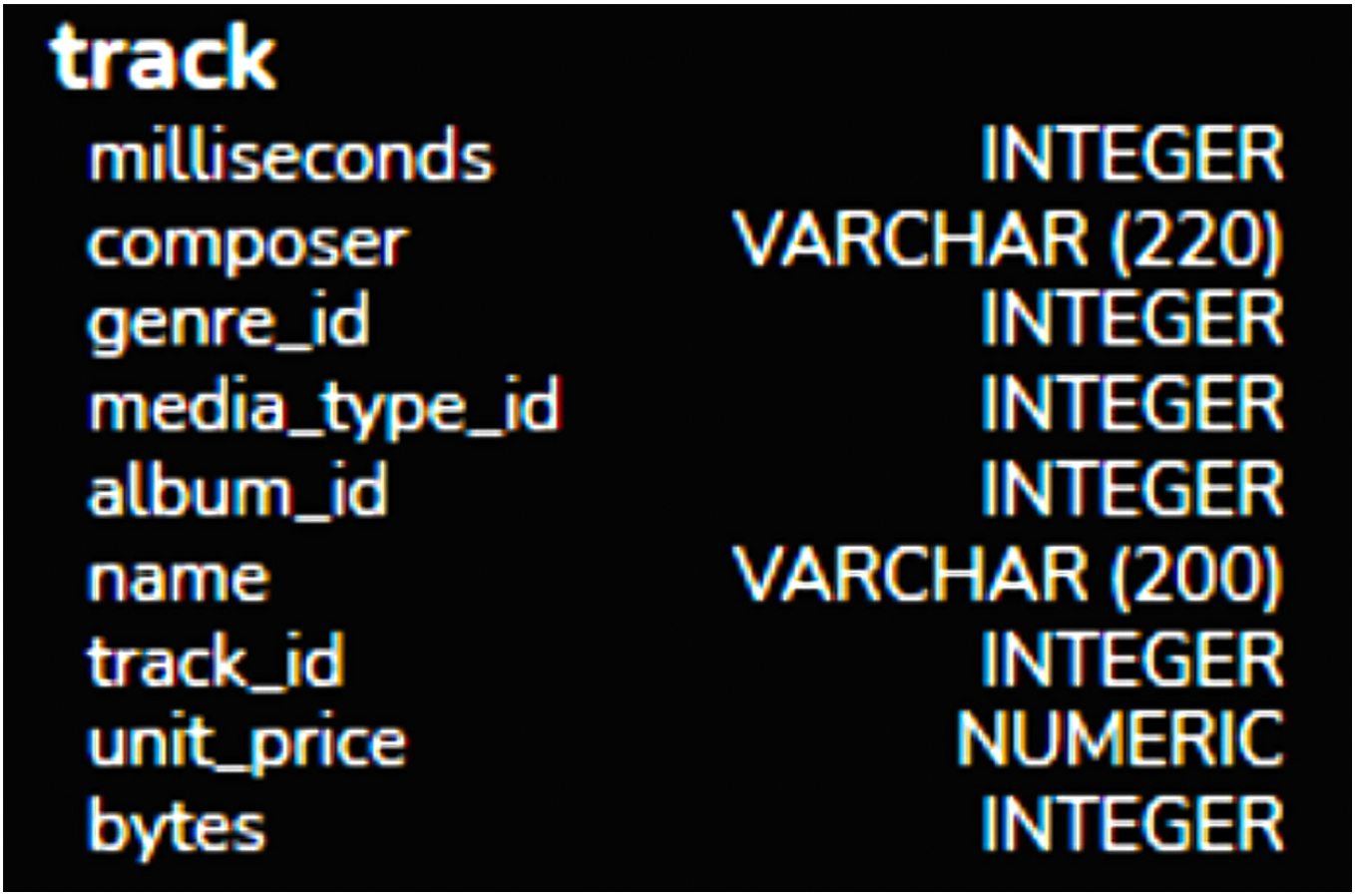
To achieve 5NF, the candidate keys must imply each join dependency in the table.

In simpler terms, this means that any non-trivial join dependency between non-key attributes in the table is a logical consequence of the candidate keys. When join dependencies cannot be inferred from candidate keys, the table is not in the 5NF standard, and further normalization is required.

In 5NF, certain types of redundancy can be eliminated by removing denormalized information from a database design. By optimizing each table, it ensures data integrity and the logical and efficient storage of information.

To achieve 5NF, it may be necessary to break down tables into smaller ones and introduce relationships between them. While 5NF provides high data integrity and eliminates some types of redundancy, it can result in more complex data models and queries. The value of normalizing to 5NF will depend on your specific requirements and the complexity of your database design. Query performance is also affected by data design decisions, including those relating to data integrity.

It would be similar to taking the track table shown below and splitting up each of the combined items:



You would have a separate table for track_id and genre_id, track_id and album_id, track_id and media_type_id, genre_id and media_type_id, genre_id and album_id, and lastly, media_type_id and album_id. In other words, every combination would be set up as its own table to avoid redundancy. However, you can easily see why this would be impractical in a real-world setting.

TERM TO KNOW

Fifth Normal Form (5NF)

The fifth normal form (5NF) is mostly conceptual. A relation between tables is in 5NF if it is in the fourth normal form (4NF) and contains no join dependencies.

SUMMARY

In this lesson, you learned in the **introduction** that the third normal form (3NF) is generally the furthest you would take normalization; however, there are some exceptions. The **Boyce-Codd normal form**, also known as BCNF, is viewed as 3.5NF. This database design first fulfills the requirements of the third normal form (3NF) but also has every determinant in a table as a candidate key. You learned that the **fourth normal form** (4NF) adds the concept that there should not be any multivalued dependencies. And the **fifth normal form** (5NF) is mostly conceptual and contains no join dependencies. Remember, higher levels of normalization like the Boyce-Codd normal form, fourth normal form (4NF), and fifth normal form (5NF) are generally not used in real-world databases.

Source: THIS TUTORIAL WAS AUTHORED BY DR. VINCENT TRAN, PHD (2020) AND Faithe Wempen (2024) FOR SOPHIA LEARNING. PLEASE SEE OUR TERMS OF USE.

TERMS TO KNOW

Alternate Key

A candidate key that has not been selected as the primary key or part of the composite key.

Boyce-Codd Normal Form

The Boyce-Codd normal form, also known as BCNF, is viewed as 3.5NF. BCNF requires that the database design first fulfills the requirements of the third normal form (3NF) but also has every determinant in a table as a candidate key.

Candidate Key

A set of one or more attributes that can uniquely identify each entry in a table.

Fifth Normal Form (5NF)

The fifth normal form (5NF) is mostly conceptual. A relation between tables is in 5NF if it is in the fourth normal form (4NF) and contains no join dependencies.

Fourth Normal Form (4NF)

A database design that satisfies all of the properties of the third normal form (3NF) and Boyce-Codd normal form (BCNF), and additionally should not have any multivalued dependencies.