2.6. Database Normalization and Information Theory (More)

Print to PDF

The previous section introduced two database theoretic concepts, keys and functional dependencies, as information theoretic equalities. We're going to make this a little more formal.

Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalization in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.

2.6.1. 1NF (First Normal Form)

1NF is the most basic stipulation on a database: (1) each table cell should contain a single value, each record needs to be unique. Uniqueness basically means that each table (without further decomposition) is not compressible. Let R be a table over columns $X_1, X_2, \ldots, X_k$.

$$H(X_1, \ldots, X_k) = \log |R|$$

2.6.2. 2NF

2NF stipulates that the each table not only contains unique records but also rows have consistent ids. 2NF: (1) must be in 1NF, and (2) there exists a single column key that uniquely identifies each row. Let R be a table over columns $X_1, X_2, \ldots, X_k$.

$$\exists i : (1, \ldots, k) s.t. H(X_i) = \log |R|$$

2.6.3. 3NF/BCNF

A transitive functional dependency is when changing a non-key column, might cause any of the other non-key columns to change. 3NF stipulates that every non-trivial functional dependency begins with a key or a subset of one. It turns out that certain types of still anomalies can arise when there are multiple keys. BCNF (Boyce-Codd Normal Form) is an extension of 3NF to avoid such cases.

Essentially the logic can be broken down into, consider a functional dependency $X_j \rightarrow X_i$:

$$H(X_i | X_j) = 0 and H(X_j) = \log |R|$$

2.6.4. 4NF and Multivalued Dependencies

It turns out that there can be more complicated forms of redundancy. Let's look at this example shamelessly taken from Wikipedia.

Course	Book	Lecturer
AHA	Silberschatz	John D
AHA	Nederpelt	John D
AHA	Silberschatz	William M
AHA	Nederpelt	William M
AHA	Silberschatz	Christian G
AHA	Nederpelt	Christian G
OSO	Silberschatz	John D
OSO	Silberschatz	William M

This is a table of university courses, the books recommended for the course, and the lecturers who will be teaching the course. Since there are multiple recommended books for some courses there is redundancy. However, this redundancy is not a functional dependency (notice there is no FD between Course and Lecturer and Book).

This particular structure is called a multivalued dependency (MVD). The above condition can be expressed as follows: if we denote by (x, y, z) the tuple having values for $\alpha, \beta, R - \alpha - \beta$ collectively denoted by a tupl;e (x, y, z), then whenever the tuples (a, b, c) and (a, d, e) exist in R, the tuples (a, b, e) and (a, d, c) should also exist in R.

Tables in 4NF are in BCNF but have no non-trivial MVDs. Essentially, an MVD formalizes a type of independence, and this condition is way easier to see with information theory:

$$H(\alpha|\backslash\mathrm{Beta}, R - \alpha - \beta) = H(\alpha)$$

By Sanjay Krishnan