Normalization and Other Data Modeling Methods

There are many paths to the top of the mountain but the view is always the same

Chinese proverb
Normalization

- An alternative database design tool to data modeling
- A theoretical foundation for the relational model
- Application of a series of rules that gradually improve the design
Functional dependency

- A relationship between attributes in an entity
  - One or more attributes determine the value of another attribute
- An identifier functionally determines all the attributes of an entity
  - stock code $\rightarrow$ firm name, stock price, stock quantity, stock dividend
  - If we know stock code we know the value of firm name, etc.

Multivalued dependency

- Formulae
  - (stock dividend, stock price) $\rightarrow$ yield
Full functional dependency

Yield is fully functionally dependent on stock dividend and stock price because both of these attributes are required to determine the value of yield

\[(\text{stock dividend, stock price}) \rightarrow \text{yield}\]

Determinant

- An attribute that fully functionally determines another attribute
  - e.g., stock code determines stock PE
Multidetermination

A given value can determine multiple values

- A multidetermines B
- A $\rightarrow \rightarrow$ B
- e.g., Department multidetermines course

Multivalued dependency means functional dependencies are multivalued
Attribute relationships

One-to-one

 Emblem: A value of an attribute determines the value of another attribute and vice versa

- A \rightarrow B and B \rightarrow A

- e.g.,
  - CH \rightarrow Switzerland
  - Switzerland \rightarrow CH
Attribute relationships

One-to-many

A value of one attribute determines the value of another attribute but **not** vice versa

- country name → currency unit
- currency unit not → country name
Attribute relationships

Many-to-many

- Neither attribute determines the other
- A not $\rightarrow$ B
- B not $\rightarrow$ A
  - country name not $\rightarrow$ language
  - language not $\rightarrow$ country name
    - French and Flemish is spoken in Belgium
    - French is spoken in many countries
Normal forms

- A classification of relations
- Stacked like a set of Russian dolls
  - Innermost is first normal form
First normal form (1NF)

- All rows must have the same number of columns
- Single valued attributes only
Second normal form (2NF)

- Violated when a nonkey column is a fact about part of the primary key
- A column is not fully functionally dependent on the primary key

- `customer-credit` in this case

<table>
<thead>
<tr>
<th>order</th>
<th>itemno</th>
<th>customerid</th>
<th>quantity</th>
<th>customer-credit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>57</td>
<td>25</td>
<td>OK</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>679</td>
<td>3</td>
<td>POOR</td>
</tr>
</tbody>
</table>

![Entity relationship diagram](image)
Third normal form (3NF)

- Violated when a nonkey column is a fact about another nonkey column
- A column is not fully functionally dependent on the primary key

- exchange rate in this case

<table>
<thead>
<tr>
<th>stock</th>
<th>stock code</th>
<th>nation</th>
<th>exchange rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>stock code</td>
<td>nation</td>
<td>exchange rate</td>
<td></td>
</tr>
<tr>
<td>MG</td>
<td>USA</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>AUS</td>
<td>0.46</td>
<td></td>
</tr>
</tbody>
</table>
Fourth normal form (4NF)

A row should not contain two or more multivalued independent facts
Fifth normal form (5NF)

- A table can be reconstructed from other tables
- There exists some rule that enables a relation to be inferred
- Base case
  - Consultants provide skills to one or more firms and firms can use many consultants; a consultant has many skills and a skill can be used by many firms; and a firm can have a need for many skills and the same skill can be required by many firms.
The rule

If a consultant has a certain skill (e.g., database) and has a contract with the firm that requires that skill (e.g., IBM), then the consultant advises the firm on that skill (i.e., he advises IBM on database)
Data modeling and normalization

- Data modeling is often an easier path to good database design.
- A high-fidelity data model will be of high normal form.
- 5NF is likely to create the most problems.
  - Check for special rules.
Goal

Learn to think like a data modeler

Different dialects and greater precision (e.g., cardinality) come easily once the basics are mastered
Key points

- Normalization is one approach to data modeling
- There are multiple representations for data model
- Learning to model is difficult
- Learning to represent a model is easy