Các dạng chuẩn của CSDL quan hệ

Lê Thành Văn

06-12-2023

Khoa Hệ thống thông tin quản lý

Giới thiệu

Giới thiệu

Định nghĩa

Chuẩn hóa cơ sở dữ liệu (csdl) là quá trình tách bảng dữ liệu nhằm:

- giảm thiểu việc dư thừa dữ liệu, và
- hạn chế lỗi có thể phát sinh trong quá trình thay đổi dữ liệu.

Việc chuẩn hóa csdl dựa trên các dạng chuẩn (normal form) được Edgar F. Codd đề ra trong mô hình quan hệ của mình.

Giới thiệu

Các loại lỗi

Lỗi khi thay đổi dữ liệu (thêm, sửa hoặc xóa) là dạng lỗi khi dữ liệu được thay đổi không tương thích với cấu trúc hoặc dữ liệu hiện có.

Giả sử thông tin giảng viên *cần* bao gồm mã môn học họ giảng dạy. Tuy nhiên, một giảng viên mới có thể chưa có môn, dẫn đến không thể thêm thông tin của họ.

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	?

Hình 1: *Lỗi khi thêm dữ liệu*

Một thông tin có thể xuất hiện nhiều lần trong một bảng, nên khi thay đổi có thể bị sót, dẫn đến thông tin không nhất quán trong csdl.

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

Hình 2: *Lỗi khi sửa dữ liệu*

Lỗi khi xóa dữ liệu là dạng lỗi mà khi xóa một thông tin này có thể xóa những thông tin (không cần xóa) khác.

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

Hình 3: *Lỗi khi xóa dữ liệu*

Giới thiệu

Các dạng chuẩn

Các dạng chuẩn thường gặp bao gồm (từ thấp đến cao):

- Dạng chuẩn 1 (First Normal Form, 1NF).
- Dạng chuẩn 2 (Second Normal Form, 2NF).
- Dạng chuẩn 3 (Third Normal Form, 3NF).
- Dạng chuẩn Boyce-Codd (Boyce-Codd Normal Form, BCNF, 3.5NF).

Việc đạt được một dạng chuẩn cao có nghĩa là phải thỏa mãn điều kiện của các dạng chuẩn thấp hơn nó.
Tức là, không thể đạt được dạng chuẩn 3 nếu như không đạt được dạng chuẩn 1 hoặc dạng chuẩn 2.

Một csdl quan hệ được gọi là **đã chuẩn hóa** nếu như nó đạt được dạng chuẩn 3.

Các dạng chuẩn

Chúng ta sẽ dùng dữ liệu sau đây để tìm hiểu về các dạng chuẩn cũng như quá trình chuẩn hóa csdl.

orders.xls

	A	B	C	D	E	F	G	H	I	J	K
1	**Invoice No.**	**Date**	**Cust. No.**	**Cust. Name**	**Cust. Address**	**Cust. City**	**Cust. State**	**Item ID**	**Item Descri**	**Item Qty.**	**Item Price**
2	125	9/13/2002	56	Foo, Inc.	23 Main St., Thorpleburg	Thorpleburg	TX	563	56" Blue Fre	4	$ 3.50
3								851	Spline End r	32	$ 0.25
4								652	3" Red Free	5	$ 12.00
5	126	9/14/2002	2	Freens R Us	1600 Pennsylvania Avenu	Washington	DC	563	56" Blue Fre	500	$ 3.50
6								652	3" Red Free	750	$ 12.00

Hình 4: *CSDL chưa chuẩn hóa*

Các dạng chuẩn

Dạng chuẩn 1

Dạng chuẩn 1 đạt được khi giá trị của mỗi ô là nguyên tố, tức là không thể chia nhỏ giá trị đó thành nhiều phần có ý nghĩa tương đương.

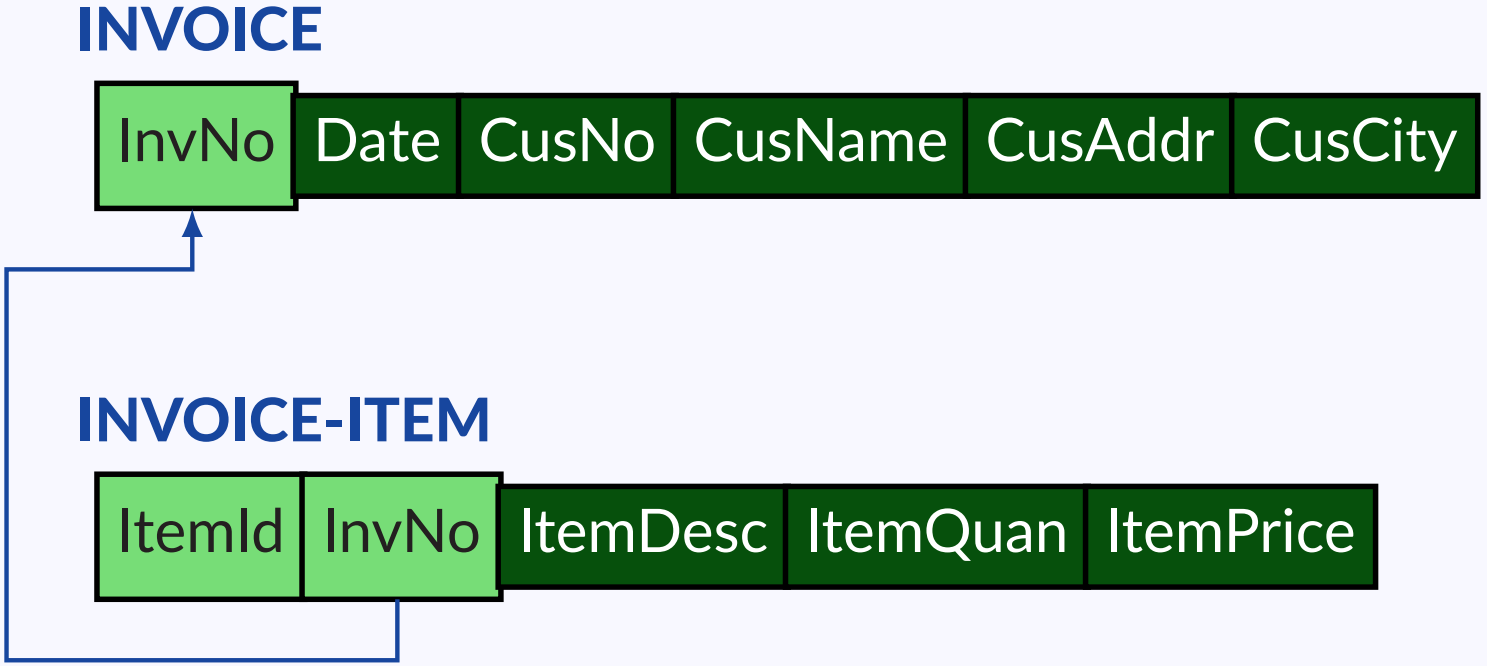
Để đưa cơ sở dữ liệu về dạng chuẩn 1, chúng ta cần tách các giá trị không nguyên tố thành từng dòng riêng biệt.

orders.xls

	A	B	C	D	E	F	G	H	I	J
1	**Invoice No.**	**Date**	**Cust. No.**	**Cust. Name**	**Cust. Address**	**Cust. City**	**Cust. State**	**Item ID**	**Item Descri**	**Item Qty.**
2	125	9/13/2002	56	Foo, Inc.	23 Main St., Thorpleburg	Thorpleburg	TX	563	56" Blue Fre	4
3	125	9/13/2002	56	Foo, Inc.	23 Main St., Thorpleburg	Thorpleburg	TX	851	Spline End r	32
4	125	9/13/2002	56	Foo, Inc.	23 Main St., Thorpleburg	Thorpleburg	TX	652	3" Red Free	5
5	126	9/14/2002	2	Freens R Us	1600 Pennsylvania Avenu	Washington	DC	563	56" Blue Fre	500
6	126	9/14/2002	2	Freens R Us	1600 Pennsylvania Avenu	Washington	DC	652	3" Red Free	750

Hình 5: *CSDL đạt dạng chuẩn 1*

Hoặc là cũng có thể tách các giá trị không nguyên tố thành bảng riêng, kèm với một khóa ngoại tham chiếu đến bảng cũ.



Các dạng chuẩn

Phụ thuộc hàm

Để hiểu về dạng chuẩn 2, chúng ta cần đến khái niệm phụ thuộc hàm.

Định nghĩa. Cho quan hệ $R(X, Y, ...)$, ta nói Y phụ thuộc vào X nếu như khi biết được giá trị của X thì ta xác định được **một** giá trị duy nhất của Y.

Ký hiệu: $X \to Y$, đọc là X xác định Y hoặc Y phụ thuộc vào X.

Với dữ liệu ở bảng bên, ta có thể thấy các phụ thuộc hàm sau:

- $A \to B, A \to C, \ldots$

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b1	c2	d2	e1
a3	b2	c1	d1	e1
a4	b2	c2	d2	e1
a5	b3	c3	d1	e1

Với dữ liệu ở bảng bên, ta có thể thấy các phụ thuộc hàm sau:

- $A \to B, A \to C, \ldots$
- $BC \to A, BC \to DE$

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b1	c2	d2	e1
a3	b2	c1	d1	e1
a4	b2	c2	d2	e1
a5	b3	c3	d1	e1

Với dữ liệu ở bảng bên, ta có thể thấy các phụ thuộc hàm sau:

- $A \to B, A \to C, \ldots$
- $BC \to A, BC \to DE$
- ...

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b1	c2	d2	e1
a3	b2	c1	d1	e1
a4	b2	c2	d2	e1
a5	b3	c3	d1	e1

Cho quan hệ $R(X, Y, Z)$, phụ thuộc hàm có một số tính chất:

- **Tính phản xạ**. $XY \to X$.
- **Tính tăng trưởng**. Nếu $X \to Y$ thì $XZ \to YZ$.
- **Tính bắc cầu**. Nếu $X \to Y$ và $Y \to Z$ thì $X \to Z$.

Câu hỏi. Chứng minh nếu $X \to Y$ và $X \to Z$ thì $X \to YZ$.

Chứng minh.

1. $X \to Z \Rightarrow X \to XZ$ (tính tăng trưởng).

□

Chứng minh.

1. $X \to Z \Rightarrow X \to XZ$ (tính tăng trưởng).
2. $X \to Y \Rightarrow XZ \to YZ$ (tính tăng trưởng).

□

Chứng minh.

1. $X \to Z \Rightarrow X \to XZ$ (tính tăng trưởng).
2. $X \to Y \Rightarrow XZ \to YZ$ (tính tăng trưởng).
3. Dùng tính bắt cầu từ 2 điều trên, ta có $X \to YZ$.

□

Các dạng chuẩn

Dạng chuẩn 2

Dạng chuẩn 2 đạt được khi:

- Đạt được dạng chuẩn 1.
- Không tồn tại phụ thuộc hàm một phần đối với bất kỳ khóa ứng viên có trong bảng.

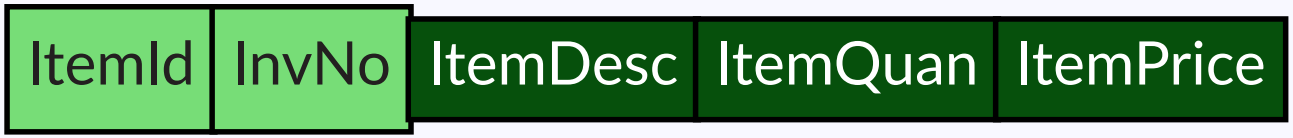
Điều kiện thứ hai có nghĩa là, với bất kỳ khóa ứng viên có nhiều hơn hai cột, giả sử là XY, thì với bất kỳ cột Z không tham gia khóa, ta **chỉ có** $XY \to Z$, mà không có $X \to Z$ hoặc $Y \to Z$.

Nhắc lại:

- **Siêu khóa** là tập hợp một hay nhiều cột có tác dụng xác định một bộ duy nhất khi biết giá trị của siêu khóa.
- **Khóa ứng viên** là siêu khóa, nhưng khi bỏ bất kỳ cột nào trong khóa ứng viên thì nó không còn là siêu khóa nữa.

Từ định nghĩa trên, có thể thấy rằng trong bảng **INVOICE-ITEM** có khóa ứng viên là {ItemId, InvNo}.

INVOICE-ITEM



Bây giờ, để tìm những cột không thỏa dạng chuẩn 2, ta cần trả lời câu hỏi *'Có thể biết được thông tin này mà không cần biết toàn bộ {ItemId, InvNo} hay không?'* cho từng cột.

Có thể thấy được rằng:

- Cột ItemQuan phụ thuộc vào {ItemId, InvNo}.

Có thể thấy được rằng:

- Cột ItemQuan phụ thuộc vào {ItemId, InvNo}.
- Cột ItemDesc và ItemPrice chỉ phụ thuộc vào ItemId.

Có thể thấy được rằng:

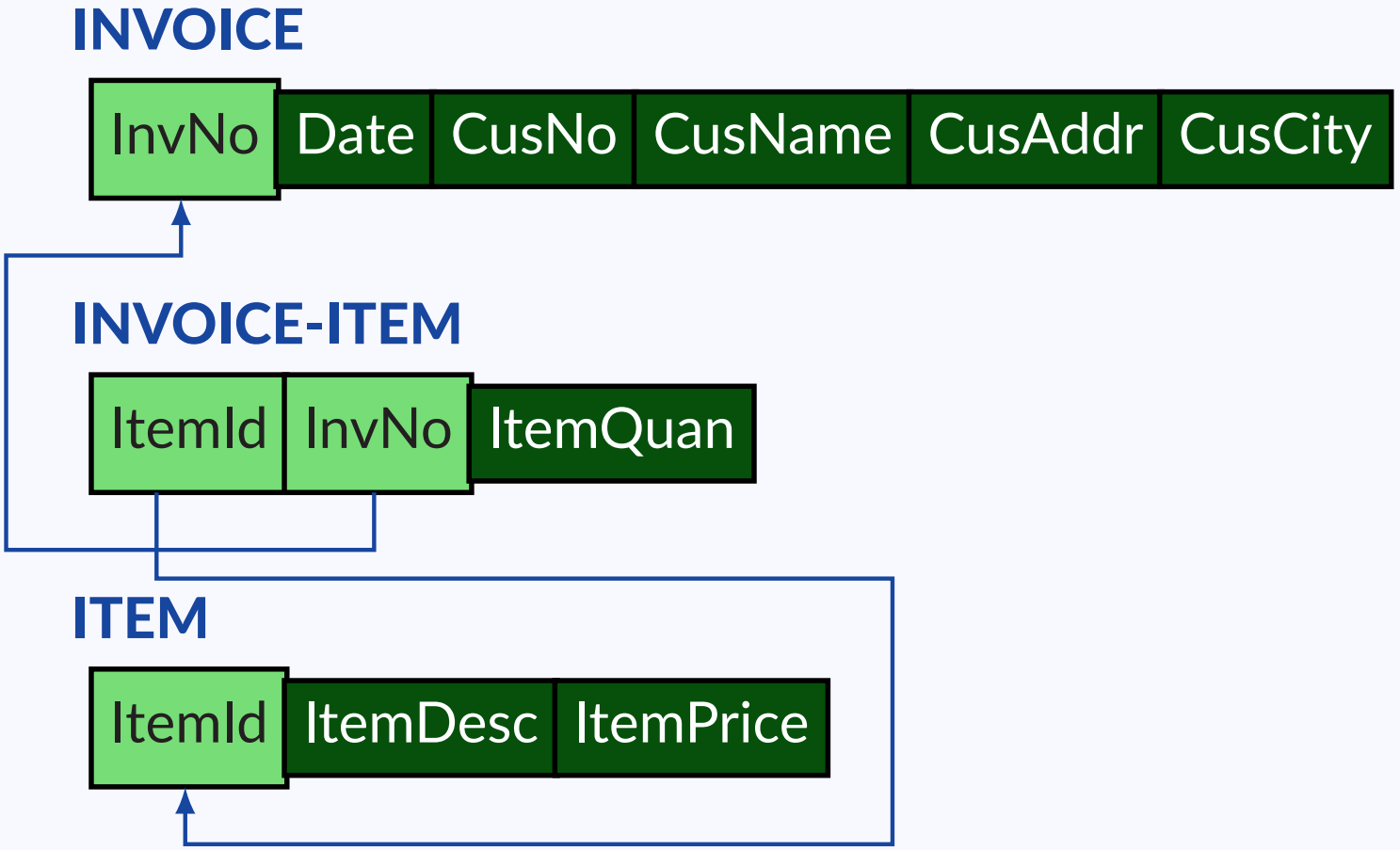
- Cột ItemQuan phụ thuộc vào {ItemId, InvNo}.
- Cột ItemDesc và ItemPrice chỉ phụ thuộc vào ItemId.

Vì vậy, bảng này không đạt dạng chuẩn 2.

Việc cần làm là tách những cột phụ thuộc một phần cùng với cột xác định của chúng thành một quan hệ mới.

Việc cần làm là tách những cột phụ thuộc một phần cùng với cột xác định của chúng thành một quan hệ mới.
Cột xác định trong bảng cũ sẽ thành khóa ngoại tham chiếu đến bảng vừa tách.

Áp dụng quy tắc trên, ta được csdl mới như sau:



Ngoài ra, có thể thấy trong bảng **INVOICE**, bên cạnh việc cột InvNo có thể xác định được CusName, CusAddr, CusCity, các cột đó cũng phụ thuộc vào CusNo.

Từ đó, ta có phụ thuộc hàm bắc cầu InvNo → CusNo → {CusName, CusAddr, CusCity}.
Trong dạng chuẩn 3, chúng ta sẽ làm việc với những phụ thuộc hàm bắc cầu này.

Các dạng chuẩn

Dạng chuẩn 3

Dạng chuẩn 3 đạt được khi:

Dạng chuẩn 3 đạt được khi:

1. Đạt được dạng chuẩn 2.

Dạng chuẩn 3 đạt được khi:

1. Đạt được dạng chuẩn 2.
2. Không tồn tại phụ thuộc hàm bắt cầu.

Tương tự cách đưa về dạng chuẩn 2, để đưa về dạng chuẩn 3:

Tương tự cách đưa về dạng chuẩn 2, để đưa về dạng chuẩn 3:

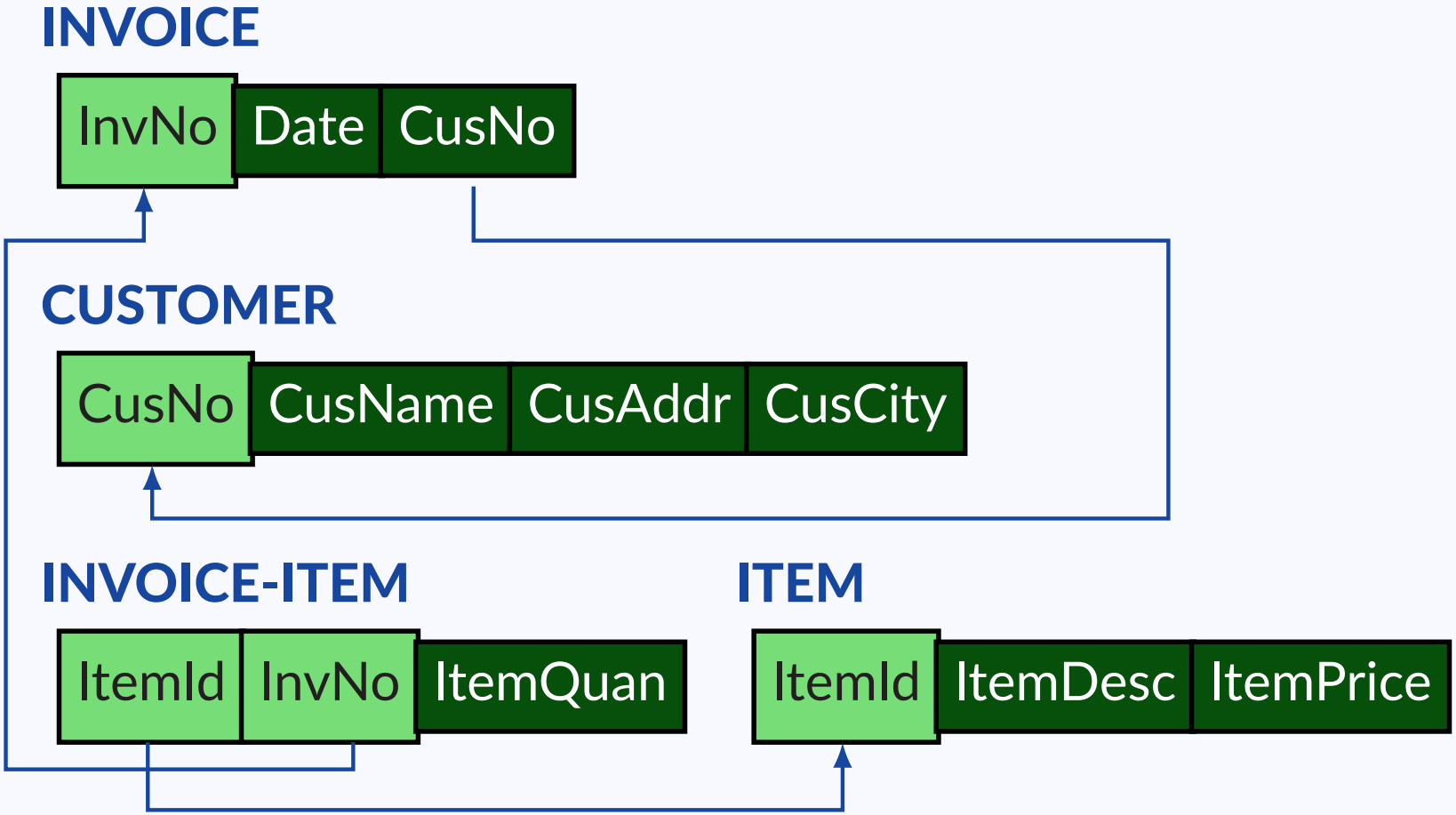
1. Tách các cột phụ thuộc bắc cầu cùng cột (không phải khóa) xác định chúng thành bảng riêng.

Tương tự cách đưa về dạng chuẩn 2, để đưa về dạng chuẩn 3:

1. Tách các cột phụ thuộc bắc cầu cùng cột (không phải khóa) xác định chúng thành bảng riêng.
2. Thêm khóa ngoại cho bảng được tách tham chiếu đến bảng vừa tạo.

Áp dụng quy tắc trên, ta được csdl mới như sau:

Áp dụng quy tắc trên, ta được csdl mới như sau:



Trong một số trường hợp, dạng chuẩn 3 là chưa đủ để tránh khỏi lỗi trong quá trình thay đổi dữ liệu.

Cùng xem ví dụ sau:

- Một sinh viên có thể có nhiều chuyên ngành.
- Với mỗi chuyên ngành của mình, sinh viên có đúng một người hướng dẫn.
- Mỗi chuyên ngành có nhiều người hướng dẫn.
- Mỗi người hướng dẫn chỉ hướng dẫn một chuyên ngành.
- Mỗi người hướng dẫn có thể hướng dẫn nhiều sinh viên.

HƯỚNG DẪN

mã sinh viên	chuyên ngành	người hướng dẫn

Chúng ta có các phụ thuộc hàm:

1. *{mã sinh viên, chuyên ngành} → người hướng dẫn.*
2. *{mã sinh viên, người hướng dẫn} → chuyên ngành.*
3. *người hướng dẫn → chuyên ngành.*

Bảng này tuy đạt dạng chuẩn 3, nhưng vẫn có những lỗi xảy ra khi thay đổi dữ liệu:

Bảng này tuy đạt dạng chuẩn 3, nhưng vẫn có những lỗi xảy ra khi thay đổi dữ liệu:

- **Thêm dữ liệu**: Thêm người hướng dẫn mới cần có sinh viên.

Bảng này tuy đạt dạng chuẩn 3, nhưng vẫn có những lỗi xảy ra khi thay đổi dữ liệu:

- **Thêm dữ liệu**: Thêm người hướng dẫn mới cần có sinh viên.
- **Cập nhật dữ liệu**: Có thể không nhất quán.

Bảng này tuy đạt dạng chuẩn 3, nhưng vẫn có những lỗi xảy ra khi thay đổi dữ liệu:

- **Thêm dữ liệu**: Thêm người hướng dẫn mới cần có sinh viên.
- **Cập nhật dữ liệu**: Có thể không nhất quán.
- **Xóa dữ liệu**: Xóa thông tin sinh viên có thể làm mất thông tin người hướng dẫn.

Để giải quyết, chúng ta cần tách bảng để đạt được dạng chuẩn Boyce-Codd.

Các dạng chuẩn

Dạng chuẩn Boyce-Codd

Dạng chuẩn Boyce-Codd đạt được khi:

Dạng chuẩn Boyce-Codd đạt được khi:

1. Đạt được dạng chuẩn 3.

Dạng chuẩn Boyce-Codd đạt được khi:

1. Đạt được dạng chuẩn 3.
2. Với bất kỳ phụ thuộc hàm $X \to Y$ thì X phải là siêu khóa.

Có thể thấy được rằng, ví dụ **HƯỚNG DẪN** ở trên không đạt dạng chuẩn Boyce-Codd do có phụ thuộc hàm *người hướng dẫn* $\rightarrow$ *chuyên ngành*.

Vi phạm dạng chuẩn Boyce-Codd thường xuất hiện khi một bản có nhiều khóa ứng viên và các khóa này có một hoặc nhiều cột giống nhau.

Để đưa về dạng chuẩn Boyce-Codd:

Để đưa về dạng chuẩn Boyce-Codd:

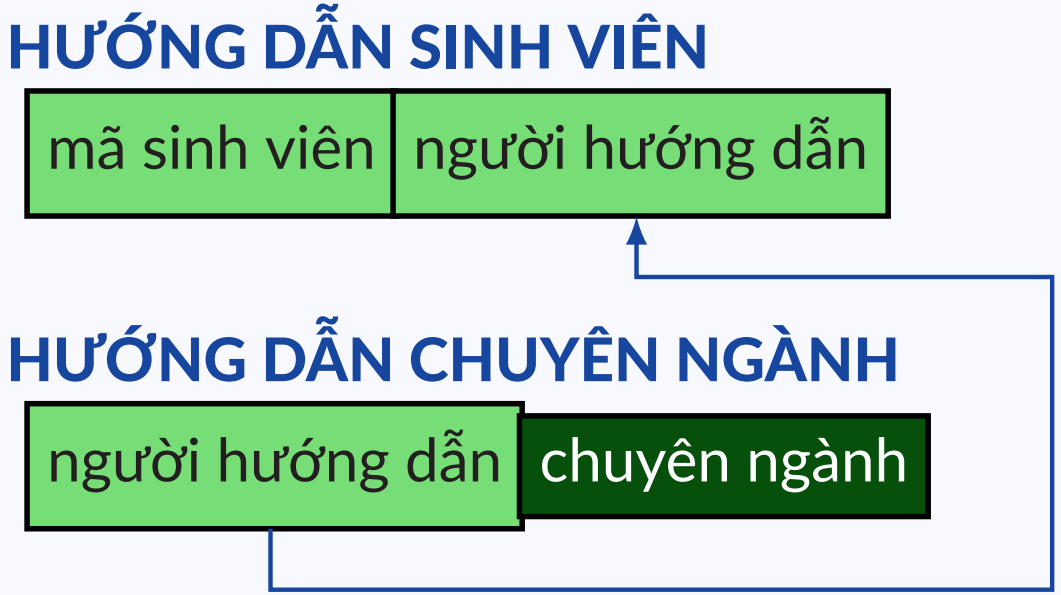
1. Tách các phụ thuộc hàm vi phạm dạng chuẩn Boyce-Codd thành bảng riêng.

Để đưa về dạng chuẩn Boyce-Codd:

1. Tách các phụ thuộc hàm vi phạm dạng chuẩn Boyce-Codd thành bảng riêng.
2. Thêm khóa ngoại cho bảng được tách tham chiếu đến bảng vừa tạo.

Áp dụng quy trình trên, ta được

Áp dụng quy trình trên, ta được



Quy trình chuẩn hóa

Như có thể thấy, quy trình chuẩn hóa có thể tóm gọn như sau:

Như có thể thấy, quy trình chuẩn hóa có thể tóm gọn như sau:

1. Xác dịnh các cột không nguyên tố để đưa về dạng chuẩn 1.

Như có thể thấy, quy trình chuẩn hóa có thể tóm gọn như sau:

1. Xác dịnh các cột không nguyên tố để đưa về dạng chuẩn 1.
2. Xác định các phụ thuộc hàm.

Như có thể thấy, quy trình chuẩn hóa có thể tóm gọn như sau:

1. Xác dịnh các cột không nguyên tố để đưa về dạng chuẩn 1.
2. Xác định các phụ thuộc hàm.
3. Tách bảng dựa trên phụ thuộc hàm.