

Chương 5: Chuẩn hóa cơ sở dữ liệu

Nội dung

- Giới thiệu
- Phép tách-kết nối bảo toàn (không tổn thất) thông tin
- Thuật toán kiểm tra tách không tổn thất thông tin
- Các dạng chuẩn của lược đồ quan hệ
- Dạng chuẩn Boyce codd

Phép tách-kết nối bảo toàn thông tin (lossless-join decomposition)

Cho lược đồ quan hệ Q (TENNCC, DIACHI, SANPHAM, DONGIA) có quan hệ tương ứng là r

Đặt r_1 là quan hệ có được bằng cách chiếu r lên Q_1 (TENNCC, SANPHAM, DONGIA) ,

Đặt r_2 là quan hệ có được bằng cách chiếu r lên Q_2 (TENNCC, DIACHI)

Đặt r' là quan hệ có được bằng cách kết tự nhiên giữa r_1 và r_2 qua TENNCC .

chẳng hạn:

r

TENNCC	DIACHI	SANPHAM	DONGIA
Hung	12 Nguyễn Kiệm	Gạch ống	200
Hung	12 Nguyễn Kiệm	Gạch thẻ	250
Hung	40 Nguyễn Oanh	Gạch ống	200

$r_2 = r \cdot Q_2$

TENNCC	DIACHI
Hung	12 Nguyễn Kiệm
Hung	40 Nguyễn Oanh

$r_1 = r \cdot Q_1$

TENNCC	SANPHAM	DONGIA
Hung	Gạch ống	200
Hung	Gạch thẻ	250

Phép tách-kết nối bảo toàn thông tin (lossless-join decomposition)

$$r' = r1 \bowtie r2$$

TENNCC	DIACHI	SANPHAM	DONGIA
Hung	12 Nguyễn Kiệm	Gạch ống	200
Hung	12 Nguyễn Kiệm	Gạch thẻ	250
Hung	40 Nguyễn Oanh	Gạch ống	200
Hung	40 Nguyễn Oanh	Gạch thẻ	250

Phép tách-kết nối bảo toàn thông tin (lossless-join decomposition)

Kết quả là $r \neq r'$ hay $r \neq r.Q_1 \bowtie r.Q_2$.

Với kết quả trên, ta nói phép tách $\rho(Q_1, Q_2)$ tách Q thành Q_1, Q_2 là tách-kết nối (phân rã) mất mát thông tin.

Nếu $r = r.Q_1 \bowtie r.Q_2$ ta nói phép tách $\rho(Q_1, Q_2)$ là tách-kết nối không mất mát thông tin (tách kết nối bảo toàn thông tin hay phân rã bảo toàn thông tin).

Vậy với điều kiện nào thì phép tách trở thành tách-kết nối không mất mát thông tin?

Định nghĩa phép tách Q thành 2 lược đồ con

Q là lược đồ quan hệ, Q_1, Q_2 hai lược đồ con có:

$$Q_1 \cap Q_2 = X$$

$$Q_1 \cup Q_2 = Q$$

Nói rằng lược đồ quan hệ Q được tách thành hai lược đồ con Q_1, Q_2 theo phép tách $\rho(Q_1, Q_2)$ là phép tách kết nối không mất (hay phép tách bảo toàn thông tin) nếu với r là quan hệ bất kỳ của Q ta có:

$$r = r \cdot Q_1 \overset{X}{\bowtie} r \cdot Q_2$$

Tức là r được tạo nên từ phép kết nối tự nhiên của các hình chiếu của nó trên các Q_1, Q_2

Tính chất

Nếu Q là một lược đồ quan hệ, Q_1, Q_2 là hai lược đồ quan hệ con có

$$Q_1 \cap Q_2 = X$$

$$Q_1 \cup Q_2 = Q$$

$$X \rightarrow Q_2$$

Thì

$$r = r.Q_1 \overset{X}{\bowtie} r.Q_2$$

Ví dụ

Cho $Q(SAIP)$, $Q_1 = (SA)$, $Q_2 = (SIP)$ $F = \{S \rightarrow A, SI \rightarrow P\}$. Hỏi việc tách Q thành Q_1 và Q_2 có gây ra mất mát thông tin không?

Áp dụng tính chất trên, ta có

$$Q_1 \cap Q_2 = S$$

$$Q_1 \cup Q_2 = SAIP = Q$$

$$S \rightarrow SA = Q_1$$

Theo tính chất trên, với mọi quan hệ r của Q ta luôn có $r = r.Q_1 \overset{s}{\bowtie} r.Q_2$. Suy ra phép tách trên là phép tách kết nối bảo toàn thông tin.

Thuật toán kiểm tra phép tách kết nối bảo toàn thông tin

Dữ liệu vào: lược đồ quan hệ $Q(A_1, A_2, \dots, A_n)$, tập phụ thuộc hàm F , phép tách $\rho = (Q_1, Q_2, \dots, Q_k)$.

Dữ liệu ra: kết luận phép tách ρ có phải là phép tách bảo toàn thông tin ?

- *Thiết lập bảng với $k+1$ dòng, $n+1$ cột. Cột j ứng với thuộc tính A_j ($j=1 \dots n$), hàng i ứng với lược đồ quan hệ Q_i ($i=1 \dots k$). Tại vị trí hàng i , cột j ta điền ký hiệu a_j nếu $A_j \in Q_i$, nếu không ta đặt ký hiệu b_t vào vị trí đó. (với t đầu tiên bằng 1) và sau đó tăng t lên một đơn vị.*
- *Xét lần lượt các phụ thuộc hàm trong F , áp dụng cho bảng vừa mới thành lập ở trên. Giả sử xét $(X \rightarrow Y) \in F$, chúng ta tìm những hàng giống nhau ở tất cả các thuộc tính của X , nếu thấy những hàng như vậy ta sẽ làm cho các ký hiệu của hai hàng này bằng nhau ở tất cả các thuộc tính của Y . Khi làm cho 2 ký hiệu này bằng nhau, nếu một trong hai ký hiệu là a_j thì cho ký hiệu kia trở thành a_j , nếu hai ký hiệu là b_k hoặc b_l thì có thể cho chúng trở thành b_t hoặc b_t (với $t = \min(k, l)$). Bước này được tiếp tục cho các phụ thuộc hàm còn lại của F cho đến khi không còn áp dụng được nữa.*
- *Xét bảng kết quả, nếu thấy trong bảng này có một hàng chứa toàn a_j ($j=1..n$) thì kết luận đó là phép kết nối bảo toàn thông tin, ngược lại là phép kết nối mất mát thông tin.*

Chú ý: một điều quan trọng cần phải nhớ là khi cho hai ký hiệu bằng nhau thì phải cho bằng nhau ở tất cả các xuất hiện của chúng trong bảng chứ không phải chỉ cho bằng nhau ở những ký hiệu trong phạm vi các phụ thuộc $X \rightarrow Y \in F$.

Ví dụ: Với $Q(ABCDE)$

$$Q_1 = (AD), Q_2 = (AB), Q_3 = (BE), Q_4 = (CDE), Q_5 = (AE)$$

$$F = \{A \rightarrow C, B \rightarrow C, A \rightarrow D, DE \rightarrow C, CE \rightarrow A\}$$

Kiểm tra tính bảo toàn thông tin của phép phân rã Q thành Q_1, Q_2, Q_3, Q_4, Q_5 .

Bước 1: $a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5$

	A	B	C	D	E
$Q_1(AD)$	a_1			a_4	
$Q_2(AB)$	a_1	a_2			
$Q_3(BE)$		a_2			a_5
$Q_4(CDE)$			a_3	a_4	a_5
$Q_5(AE)$	a_1				a_5

Bước 2: Điền b_1, b_2, b_3, \dots

	A	B	C	D	E
$Q_1(AD)$	a_1	b_1	b_2	a_4	b_3
$Q_2(AB)$	a_1	a_2	b_4	b_5	b_6
$Q_3(BE)$	b_7	a_2	b_8	b_9	a_5
$Q_4(CDE)$	b_{10}	b_{11}	a_3	a_4	a_5
$Q_5(AE)$	a_1	b_{12}	b_{13}	b_{14}	a_5

Sửa bảng giá trị để nó thỏa $A \rightarrow C$

Sửa b_4, b_{13} thành b_2

	A	B	C	D	E
$Q_1 (AD)$	a_1	b_1	b_2	a_4	b_3
$Q_2 (AB)$	a_1	a_2	b_2	b_5	b_6
$Q_3 (BE)$	b_7	a_2	b_8	b_9	a_5
$Q_4 (CDE)$	b_{10}	b_{11}	a_3	a_4	a_5
$Q_5 (AE)$	a_1	b_{12}	b_2	b_{14}	a_5

Sửa bảng giá trị để nó thỏa $B \rightarrow C$

Sửa b_8 thành b_2

	A	B	C	D	E
$Q_1 (AD)$	a_1	b_1	b_2	a_4	b_3
$Q_2 (AB)$	a_1	a_2	b_2	b_5	b_6
$Q_3 (BE)$	b_7	a_2	b_2	b_9	a_5
$Q_4 (CDE)$	b_{10}	b_{11}	a_3	a_4	a_5
$Q_5 (AE)$	a_1	b_{12}	b_2	b_{14}	a_5

Sửa bảng giá trị để nó thỏa $A \rightarrow D$

Sửa b_5, b_{14} thành a_4

	A	B	C	D	E
$Q_1 (AD)$	a_1	b_1	b_2	a_4	b_3
$Q_2 (AB)$	a_1	a_2	b_2	a_4	b_6
$Q_3 (BE)$	b_7	a_2	b_2	b_9	a_5
$Q_4 (CDE)$	b_{10}	b_{11}	a_3	a_4	a_5
$Q_5 (AE)$	a_1	b_{12}	b_2	a_4	a_5

Sửa bảng giá trị để nó thỏa $DE \rightarrow C$

sửa b_2 thành $a_3 \Rightarrow$ sửa tất cả b_2

thành a_3

	A	B	C	D	E
$Q_1 (AD)$	a_1	b_1	a_3	a_4	b_3
$Q_2 (AB)$	a_1	a_2	a_3	a_4	b_6
$Q_3 (BE)$	b_7	a_2	a_3	b_9	a_5
$Q_4 (CDE)$	b_{10}	b_{11}	a_3	a_4	a_5
$Q_5 (AE)$	a_1	b_{12}	a_3	a_4	a_5

Sửa bảng giá trị để nó thỏa $CE \rightarrow A$

Sửa b_7, b_{10} thành a_1 .

	A	B	C	D	E
$Q_1(AD)$	a_1	b_1	a_3	a_4	b_3
$Q_2(AB)$	a_1	a_2	a_3	a_4	b_6
$Q_3(BE)$	a_1	a_2	a_3	b_9	a_5
$Q_4(CDE)$	a_1	b_{11}	a_3	a_4	a_5
$Q_5(AE)$	a_1	b_{12}	a_3	a_4	a_5

Lần lượt xét lại các phụ thuộc hàm trong F, nếu bảng giá trị chưa thỏa phụ thuộc hàm nào thì tiếp tục làm cho nó thỏa.

Sửa bảng giá trị để nó thỏa $A \rightarrow D$

	A	B	C	D	E
$Q_1(AD)$	a_1	b_1	a_3	a_4	b_3
$Q_2(AB)$	a_1	a_2	a_3	a_4	b_6
$Q_3(BE)$	a_1	a_2	a_3	a_4	a_5
$Q_4(CDE)$	a_1	b_{11}	a_3	a_4	a_5
$Q_5(AE)$	a_1	b_{12}	a_3	a_4	a_5

Dòng thứ $Q_3(BE)$ của bảng chứa toàn giá trị a_j ($j=1..n$) nên phép phân rã trên là bảo toàn thông tin.

Các dạng chuẩn cơ sở dữ liệu

Các bất thường khi cập nhật dữ liệu

- Ví dụ với CSDL là bảng hóa đơn (INVOICE) cho khách hàng như sau:

Acme Industries INVOICE				
Customer Number: 1454837			Terms: Net 30	
Customer: W. Coyote			Ship Via: USPS	
General Delivery			Order Date: 12/01/2008	
Falling Rocks, AZ 84211				
(599) 555-9345				
<u>Product No.</u>	<u>Description</u>	<u>Quantity</u>	<u>Unit Price</u>	<u>Extended Amount</u>
SPR-2290	Super Strength Springs	2	24.00	\$48.00
STR-67	Foot Straps, leather	2	2.50	\$5.00
HLM-45	Deluxe Crash Helmet	1	67.88	\$67.88
SFR-1	Rocket, solid fuel	1	128,200.40	\$128,200.40
ELT-1	Emergency Location Transmitter	1	79.88	**FREE GIFT**
TOTAL ORDER AMOUNT:				\$128,321.28

Các bất thường khi cập nhật dữ liệu

- Bất thường khi thêm dữ liệu:

- Không thể thêm một khách hàng vào CSDL nếu khách hàng không mua một mặt hàng nào.

- Bất thường khi cập nhật dữ liệu:

- Muốn thay đổi địa chỉ của khách hàng, phải thay đổi địa chỉ trên tất cả các hóa đơn của khách hàng đó bởi vì địa chỉ của khách hàng lưu trữ dư thừa trong tất cả các hóa đơn của khách hàng.

Các bất thường khi cập nhật dữ liệu

- Bất thường khi xóa dữ liệu:
 - Nếu xóa hóa đơn cuối cùng của một khách hàng thì tất cả dữ liệu về khách hàng đó bị mất.

Các bất thường khi cập nhật dữ liệu

- Mục đích của chuẩn hóa CSDL là nhóm các thuộc tính vào các quan hệ nhằm **giảm thiểu dư thừa dữ liệu, loại bỏ các bất thường** khi cập nhật CSDL.
- > Cần có các bước chuẩn hoá dữ liệu từ một CSDL chưa chuẩn hóa sang chuẩn hóa.
- Dạng chưa chuẩn hóa (unnormalized form - **UNF**): quan hệ chưa chuẩn hóa là quan hệ chứa các bộ dữ liệu bị lặp lại giá trị.

Các bất thường khi cập nhật dữ liệu

- Bảng hóa đơn (INVOICE) ở trên biểu diễn dưới dạng bảng như sau.
Bảng này ở dạng không chuẩn

INVOICE

Customer Number	Customer	Terms	Ship Via	Order Date	Product No.	Description	Quantity	Unit Price	Extended Amount
1454837	W.Coyote	Net 30	USPS	12/01/2008	SPR-2290	Super Strength Springs	2	24.00	\$48.00
1454837	W.Coyote	Net 30	USPS	12/01/2008	SRT-67	Foot Straps, leather	2	2.50	\$5.00
1454837	W.Coyote	Net 30	USPS	12/01/2008	HLM-45	Deluxe Crash Helmet	1	67.88	\$67.88
1454837	W.Coyote	Net 30	USPS	12/01/2008	SFR-1	Rocket, Solid fuel	1	128,200.4	\$128,200.4
1454837	W.Coyote	Net 30	USPS	12/01/2008	ELT-1	Emergency Location Transmitter	1	79.88	\$79.88

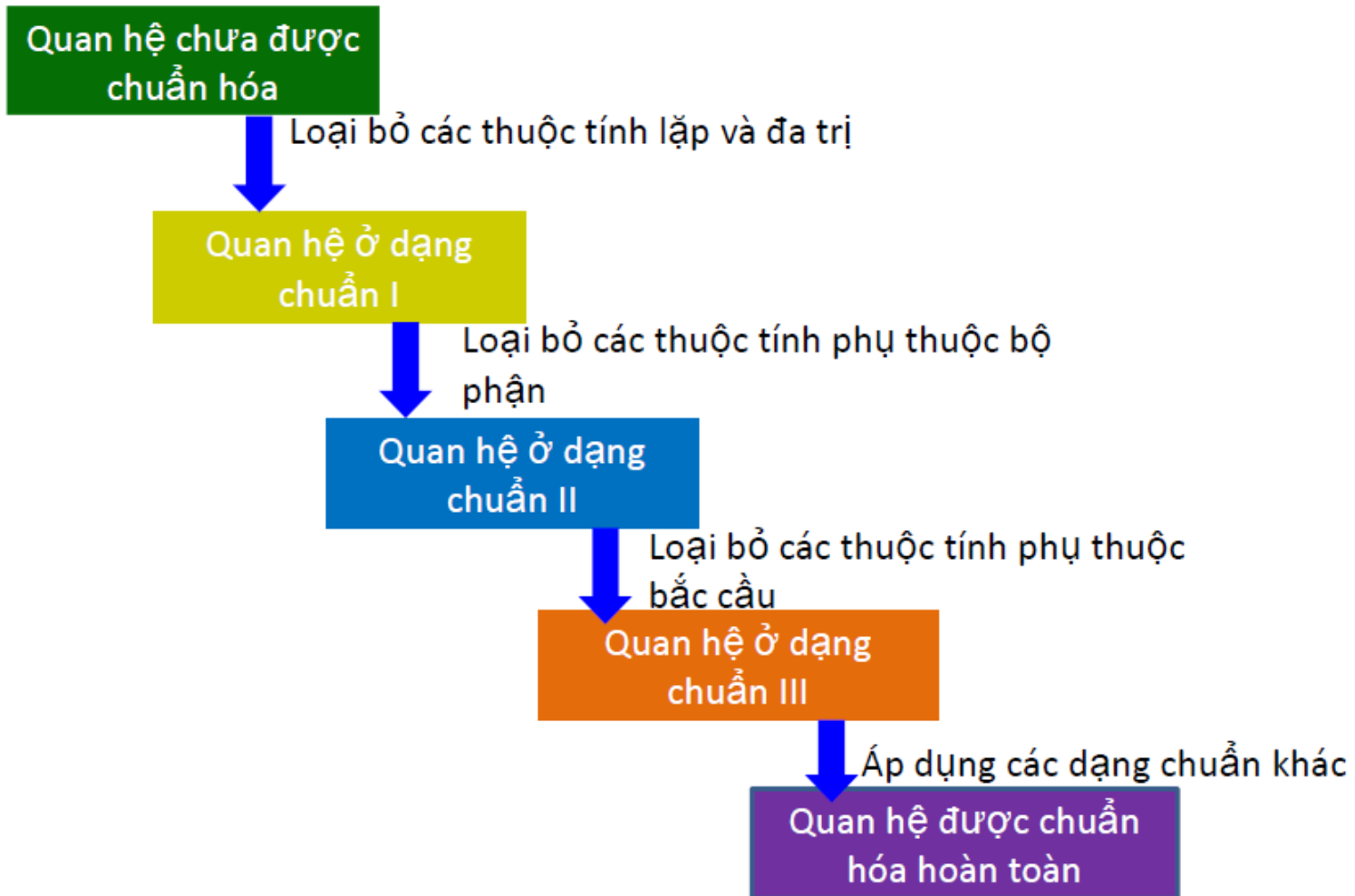
Các bất thường khi cập nhật dữ liệu

- Bảng này được chuẩn hóa bằng cách tách thành hai bảng như sau:

Customer Number	Customer	Terms	Ship Via	Order Date
1454837	W.Coyote	Net 30	USPS	12/01/2008

Customer Number	Product No.	Description	Quantity	Unit Price	Extended Amount
1454837	SPR-2290	Super Strength Springs	2	24.00	\$48.00
1454837	SRT-67	Foot Straps, leather	2	2.50	\$5.00
1454837	HLM-45	Deluxe Crash Helmet	1	67.88	\$67.88
1454837	SFR-1	Rocket, Solid fuel	1	128,200.4	\$128,200.4
1454837	ELT-1	Emergency Location Transmitter	1	79.88	\$79.88

Quy trình chuẩn hóa dữ liệu



Chuẩn hóa dữ liệu

- Chuẩn hóa là quá trình phân rã lược đồ quan hệ dựa trên một tập phụ thuộc hàm nhằm đảm các lược đồ quan hệ thoả mãn 2 tính chất:
 - Trùng lặp dữ liệu ít nhất
 - Khả năng gây ra bất thường khi cập nhật được giảm thiểu.
- Các dạng chuẩn:
 - Dạng chuẩn 1 (**1NF**)
 - Dạng chuẩn 2 (**2NF**)
 - Dạng chuẩn 3 (**3NF**)

Ví dụ dạng chưa chuẩn

■ Xét bảng dữ liệu đơn vị sau:

ĐƠN VỊ	Mã số ĐV	Tên ĐV	Mã số NQL	Địa điểm
	5	Nghiên cứu	NV002	Nam định, Hà nội, Bắc ninh
	4	Hành chính	NV014	Hà nội
	1	Lãnh đạo	NV061	Hà nội

- Ta thấy trường **Địa Điểm** của đơn vị 5 chứa **các thuộc tính đa trị** (nhiều hơn 1 giá trị) có 3 nơi là: Nam Định, Hà Nội, Bắc Ninh vì vậy bảng dữ liệu này chưa đạt dạng chuẩn hóa.

Ví dụ dạng chưa chuẩn

■ Xét bảng dữ liệu NHÂN VIÊN_DỰÁN:

NHÂNVIÊN_DỰÁN	Mã số DA	Tên DA	Tên nhân viên	Số giờ
	1	DA01	Vân	15
			Nam	20
	2	DA02	Nam	10
			Thanh	12
			Bằng	28
	3	DA03	Thanh	20
			Vân	12

- Ta thấy nhóm thuộc tính {Tên nhân viên, số giờ} bị lặp lại nhiều lần trong các dự án khác nhau. Vì vậy cơ sở dữ liệu này chưa đạt chuẩn hóa.

Ví dụ dạng chưa chuẩn

■ Xét bảng dữ liệu Invoice INVOICE

Customer Number	Customer	Terms	Ship Via	Order Date	Product No.	Description	Quantity	Unit Price	Extended Amount
1454837	W.Coyote	Net 30	USPS	12/01/2008	SPR-2290	Super Strength Springs	2	24.00	\$48.00
1454837	W.Coyote	Net 30	USPS	12/01/2008	SRT-67	Foot Straps, leather	2	2.50	\$5.00
1454837	W.Coyote	Net 30	USPS	12/01/2008	HLM-45	Deluxe Crash Helmet	1	67.88	\$67.88
1454837	W.Coyote	Net 30	USPS	12/01/2008	SFR-1	Rocket, Solid fuel	1	128,200.4	\$128,200.4
1454837	W.Coyote	Net 30	USPS	12/01/2008	ELT-1	Emergency Location Transmitter	1	79.88	\$79.88

Đây là 1 hóa đơn của 1 khách hàng, Nếu có thêm 1 hóa đơn khác thì các thuộc tính {Product No, Description, Quantity, Unit Price, Extended Mount} sẽ lặp lại trong các hóa đơn khác vì vậy bảng này chưa đạt dạng chuẩn hóa.

Dạng chuẩn 1 (1 NF)

- Khái niệm: Lược đồ quan hệ R được gọi là 1NF nếu và chỉ nếu tất cả các thuộc tính của R thoả mãn cả 3 điều kiện sau:
 - Là nguyên tố
 - Giá trị của các thuộc tính trên các bộ là đơn trị, không chứa nhóm lặp.
 - Không có một thuộc tính nào có giá trị mà có thể tính toán được từ một số thuộc tính khác.

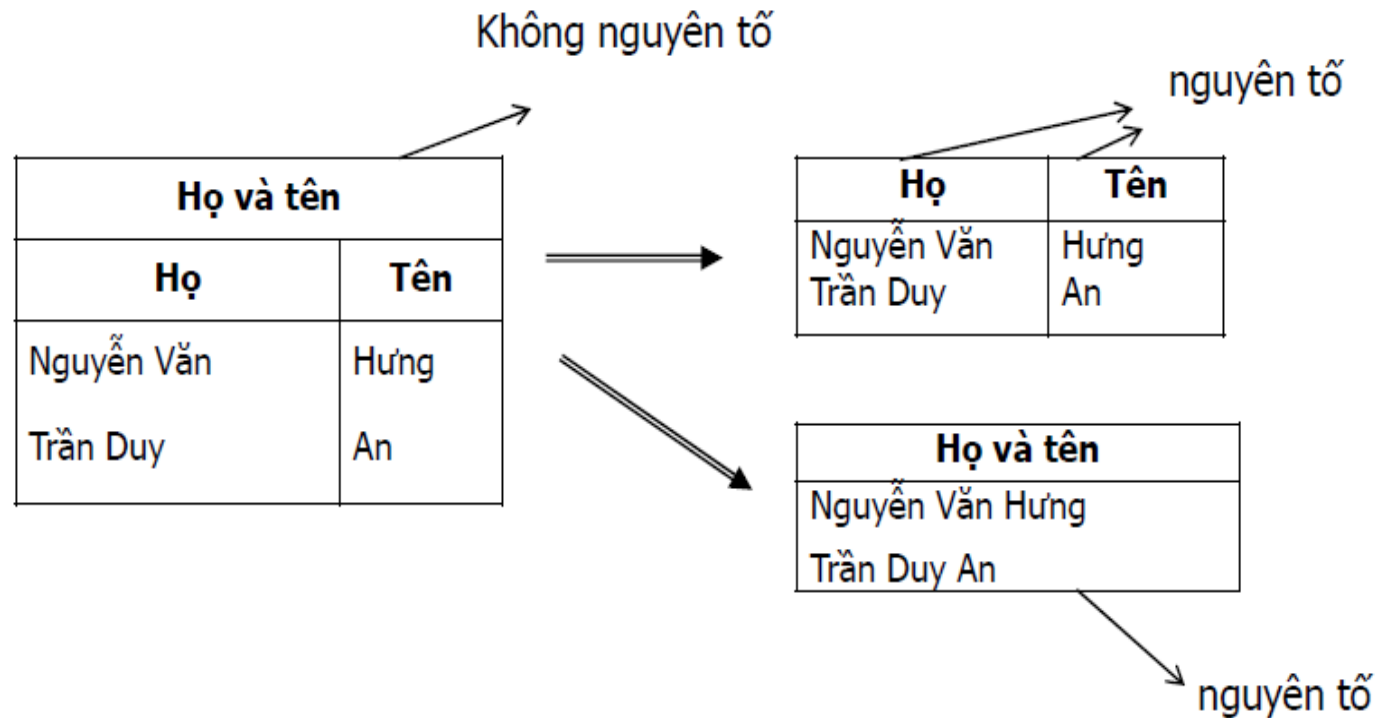
Dạng chuẩn 1 (1 NF)

- Ví dụ 1: Một lược đồ quan hệ đạt chuẩn 1NF

<u>MSKH</u>	TÊNKH	TP	PVC	<u>MSMH</u>	TÊNMH	ĐG	SL
S1	An	HCM	01	P1	Táo	650	300
S1	An	HCM	01	P2	Chanh	500	200
S1	An	HCM	01	P3	Táo	450	400
S2	Hòa	HN	02	P1	Chanh	650	100
S2	Hoà	HN	02	P3	Táo	450	300
S3	Thanh	NT	03	P2	Chanh	500	200
S4	Trang	NT	03	P2	Chanh	500	210

Dạng chuẩn 1 (1 NF)

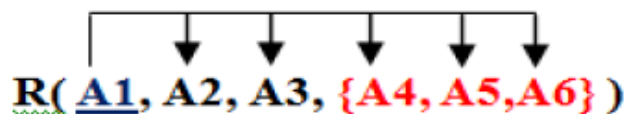
- Ví dụ 2: Quan hệ sau được chuẩn hóa về dạng chuẩn 1



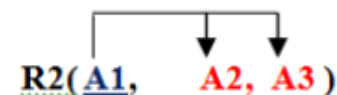
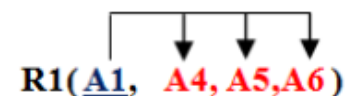
Dạng chuẩn 1 (1 NF)

- Nguyên tắc chung: Loại bỏ thuộc tính lặp hoặc đa trị
- Các bước thực hiện:
 - Tách nhóm thuộc tính lặp / đa trị sang một bảng mới.
 - Khóa của bảng mới là Khóa của bảng ban đầu và khóa nhóm lặp
 - Bảng còn lại là bảng gồm có khóa và các thuộc tính còn lại.

- Ví dụ: cho quan hệ



- Với A1 là khóa chính của quan hệ, {A4,A5,A5} là thuộc tính lặp.
- Quan hệ R sẽ được tách thành 2 quan hệ R1, R2



Dạng chuẩn 1 (1 NF)

■ Cho bảng ĐƠN VỊ

ĐƠN VỊ	Mã số ĐV	Tên ĐV	Mã số NQL	Địa điểm
	5	Nghiên cứu	NV002	Nam định, Hà Nội, Bắc Ninh
	4	Hành chính	NV014	Hà Nội
	1	Lãnh đạo	NV061	Hà Nội

- Ta có quan hệ sau **R**(Mã số ĐV, Tên ĐV, Mã số NQL, {Địa điểm})
- Thuộc tính đa trị là {Địa điểm} quan hệ này được tách thành:
- **R1**(Mã số ĐV, Địa điểm)
- **R2**(Mã số ĐV, Tên ĐV, Mã số NQL)

Bảng ĐƠN VỊ_ĐỊA ĐIỂM

<u>Mã số ĐV</u>	<u>Địa điểm</u>
5	Nam định
5	Hà nội
5	Bắc Ninh
4	Hà nội
1	Hà nội

Bảng ĐƠN VỊ

<u>Mã số ĐV</u>	Tên ĐV	Mã số NQL
5	Nghiên cứu	NV002
4	Hành chính	NV014
1	Lãnh đạo	NV061

Dạng chuẩn 1 (1 NF)

- Cho bảng NHÂNVIÊN_DỰÁN

NHÂNVIÊN_DỰÁN	Mã số DA	Tên DA	Tên nhân viên	Số giờ
	1	DA01	Vân	15
			Nam	20
	2	DA02	Nam	10
			Thanh	12
			Bằng	28
	3	DA03	Thanh	20
			Vân	12

- Quan hệ này là **R(Mã số DA, Tên DA, {Tên nhân viên, Số giờ})**
- Thuộc tính lặp là {Tên nhân viên, Số giờ} quan hệ được tách thành:
- R1(Mã số DV, Tên nhân viên, Số giờ)**
- R2(Mã số DV, Tên DA)**

NHÂNVIÊN_DỰÁN

<u>Mã số DA</u>	<u>Tên nhân viên</u>	<u>Số giờ</u>
1	Vân	15
1	Nam	20
2	Nam	10
2	Thanh	12
2	Bằng	28
3	Thanh	20
3	Vân	12

DỰÁN

<u>Mã số DA</u>	<u>Tên DA</u>
1	DA01
2	DA02
3	DA03

Dạng chuẩn 1 (1 NF)

- Cho bảng NHÂNVIÊN_DỰÁN

NHÂNVIÊN_DỰÁN	Mã số DA	Tên DA	Tên nhân viên	Số giờ
	1	DA01	Vân	15
			Nam	20
	2	DA02	Nam	10
			Thanh	12
			Bằng	28
	3	DA03	Thanh	20
			Vân	12

- Quan hệ này là **R(Mã số DA, Tên DA, {Tên nhân viên, Số giờ})**
- Thuộc tính lặp là {Tên nhân viên, Số giờ} quan hệ được tách thành:
- R1(Mã số DV, Tên nhân viên, Số giờ)**
- R2(Mã số DV, Tên DA)**

NHÂNVIÊN_DỰÁN

<u>Mã số DA</u>	<u>Tên nhân viên</u>	<u>Số giờ</u>
1	Vân	15
1	Nam	20
2	Nam	10
2	Thanh	12
2	Bằng	28
3	Thanh	20
3	Vân	12

DỰÁN

<u>Mã số DA</u>	<u>Tên DA</u>
1	DA01
2	DA02
3	DA03

Dạng chuẩn 1 (1 NF)

- Xét bảng INVOICE được viết thành quan hệ sau (thêm Invoice Number làm thuộc tính khóa cho bảng Invoice vì mỗi hóa đơn có Số hóa đơn là Khóa)
- R(Invoice Number (PK), Customer Number, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone, Terms, Ship Via, Order Date, {Product Number, Product Description, Quantity, Unit Price, Extended Amount, Total Order Amount })
- Thuộc tính lặp được bôi đỏ và đặt trong dấu { } quan hệ được tách
- **R1-INVOICE** (Invoice Number (PK), Product Number, Product Description, Quantity, Unit Price, Extended Amount, Total Order Amount)
- **R2-INVOICE LINE ITEM** (Invoice Number (PK), Customer Number, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone, Terms, Ship Via, Order Date)

Dạng chuẩn 1 (1 NF)

- Kết quả bảng được tách thành hai bảng như sau

Customer Number	Customer	Terms	Ship Via	Order Date
1454837	W.Coyote	Net 30	USPS	12/01/2008

Product No.	Customer Number	Product No.	Description	Quantity	Unit Price	Extended Amount
SPR-2290	1454837	SPR-2290	Super Strength Springs	2	24.00	\$48.00
SRT-67	1454837	SRT-67	Foot Straps, leather	2	2.50	\$5.00
HLM-45	1454837	HLM-45	Deluxe Crash Helmet	1	67.88	\$67.88
SFR-1	1454837	SFR-1	Rocket, Solid fuel	1	128,200.4	\$128,200.4
ELT-1	1454837	ELT-1	Emergency Location Transmitter	1	79.88	\$79.88

Dạng chuẩn 2 (2 NF)

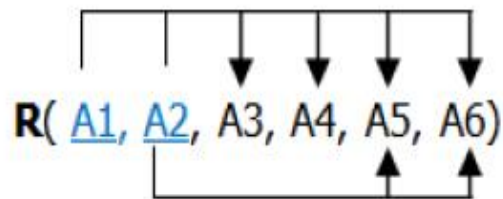
- Quan hệ R được gọi là đạt dạng chuẩn 2 nếu đã là **1NF** và tất cả các thuộc tính không khóa đều **phụ thuộc hàm đầy đủ** vào khóa.
 - Thuộc tính không khóa: là thuộc tính không tham gia vào bất kỳ khoá nào.

Dạng chuẩn 2 (2 NF)

- Quy tắc chuẩn hóa từ **1NF - 2NF**:

- Bước 1: Loại bỏ các thuộc tính không khoá phụ thuộc vào một bộ phận khoá chính và tách thành ra một bảng riêng, khoá chính của bảng là bộ phận khoá mà chúng phụ thuộc vào.
- Bước 2: Các thuộc tính còn lại lập thành một quan hệ, khóa chính của nó là khóa chính ban đầu.

Quy tắc chuẩn hóa từ 1NF - 2 NF



Khóa chính là bộ $\{A_1, A_2\}$ (tức là từ $\{A_1, A_2\} \rightarrow \{A_3, A_4, A_5, A_6\}$)

Có phụ thuộc hàm phụ thuộc vào một phần của khóa $\{A_2\} \rightarrow \{A_5, A_6\}$

Quan hệ được tách thành hai quan hệ sau

$R_1(A_2, A_5, A_6)$

- Tách các thuộc tính phụ thuộc vào 1 phần của khóa

$R_2(A_1, A_2, A_3, A_4)$

- Bảng ban đầu còn lại

Ví dụ 1: Tách chuẩn từ 1NF - 2 NF

- Ví dụ: cho quan hệ **R**(MSKH, TÊNKH, TP, PVC, MSMH, TÊNMH, ĐG, SL) và tập phụ thuộc hàm F được định nghĩa trên R:

- Tập phụ thuộc hàm:

$F = \{$
 MSKH \rightarrow TÊNKH, TP
 MSMH \rightarrow TÊNMH, ĐG
 MSKH, MSMH \rightarrow SL
 TP \rightarrow PVC $\}$

<u>MSKH</u>	TÊNKH	TP	PVC	<u>MSMH</u>	TÊNMH	ĐG	SL
S1	An	HCM	01	P1	Táo	650	300
S1	An	HCM	01	P2	Chanh	500	200
S1	An	HCM	01	P3	Cam	450	400
S2	Hòa	HN	02	P1	Táo	650	100
S2	Hoà	HN	02	P3	Cam	450	300
S3	Thanh	NT	03	P2	Chanh	500	200
S4	Trang	NT	03	P2	Chanh	500	210

- Quan hệ này được tách thành các quan hệ sau
- R1(MSKH, TÊNKH, TP, PVC)
- R2(MSKH, MSMH, TÊNMH, ĐG, SL)
- Trong R2 chứa phụ thuộc hàm {MSMH} \rightarrow {TÊNMH, ĐG} tách thành
- R21(MSMH, TÊNMH, ĐG)
- R22(MSKH, MSMH, SL)

Ví dụ 1: Tách chuẩn từ 1NF - 2 NF

- Tách thành các quan hệ sau

R ₁				ĐH		
<u>MSKH</u>	TÊNKH	TP	PVC	<u>MSKH</u>	<u>MSMH</u>	SL
S1	An	HCM	01	S1	P1	300
S2	Hoà	HN	02	S1	P2	200
S3	Thanh	NT	03	S1	P3	400
S4	Trang	NT	03	S2	P1	100
				S2	P3	300
				S3	P2	200
				S4	P2	210

$F_{R1} = \{MSKH \rightarrow TÊNKH, TP$
 $TP \rightarrow PVC\}$

MH

<u>MSMH</u>	TÊNMH	ĐG
P1	Táo	650
P2	Cam	500
P3	Chanh	450

$F_{MH} = \{MSMH \rightarrow TÊNMH, ĐG\}$

$F_{ĐH} = \{MSKH, MSMH \rightarrow SL\}$

Ví dụ 2: Tách chuẩn từ 1NF - 2 NF

NHÂNVIÊN_DỰÁN(Mã sốNV, Mã sốDA, Số giờ, Họ tênNV, TênDA, Địa điểmDA)

với các phụ thuộc hàm:

{Mã sốNV, Mã sốDA} → {Số giờ}

{Mã sốNV} → {Họ tênNV}

{Mã sốDA} → {TênDA, Địa điểmDA}

<u>Mã sốNV</u>	<u>Mã sốDA</u>	Số giờ	Họ tênNV	TênDA	Địa điểmDA
NV01	DA01	10	Nguyễn Lan	Dự án 1	Hà Nội
NV02	DA02	20	Lê Trung	Dự án 2	Hà Nam
NV01	DA02	15	Nguyễn Lan	Dự án 2	Hà Nam
NV02	DA01	30	Lê Trung	Dự án 1	Hà Nội
NV03	DA01	12	Đinh Trung	Dự án 1	Hà Nội
NV03	DA03	15	Đinh Trung	Dự án 3	Hải Phòng
NV02	DA03	26	Lê Trung	Dự án 3	Hải Phòng
NV04	DA02	16	Hoài Nam	Dự án 2	Hà Nam

- Quan hệ này được tách thành các quan hệ sau

R1 (Mã sốNV, Họ tênNV)

R2 (Mã sốNV, Mã sốDA, TênDA, Địa điểmDA, Số giờ)

Trong quan hệ R2 chứa phụ thuộc hàm {Mã sốDA} → {TênDA, Địa điểmDA} tách thành

R21(Mã sốDA, TênDA, Địa điểmDA)

R22(Mã sốNV, Mã sốDA, Số giờ)

Ví dụ 2: Tách chuẩn từ 1NF - 2 NF

- Kết quả tách ta được 3 bảng sau

NHÂNVIÊN

<u>Mã số NV</u>	Họ tên NV
NV01	Nguyễn Lan
NV02	Lê Trung
NV03	Đinh Trung
NV04	Hoài Nam

DỰ ÁN

<u>Mã số DA</u>	Tên DA	Địa điểm DA
DA01	Dự án 1	Hà Nội
DA02	Dự án 2	Hà Nam
DA03	Dự án 3	Hải Phòng

NHÂNVIÊN_DỰ ÁN

<u>Mã số NV</u>	<u>Mã số DA</u>	Số giờ
NV01	DA01	10
NV02	DA02	20
NV01	DA02	15
NV02	DA01	30
NV03	DA01	12
NV03	DA03	15
NV02	DA03	26
NV04	DA02	16

Ví dụ 3: Tách chuẩn từ 1NF - 2 NF

Xét Quan hệ INVOICE LINE ITEM:

(Invoice Number (PK), Product Number (PK), Product Description, Quantity, Unit Price, Extended Amount)

- Ta thấy có phụ thuộc hàm vào một phần của khóa như sau

{Product Number} -> {Product Description, Unit Price}

- Quan hệ này được tách thành 2 quan hệ sau:
- PRODUCT (Product Number, Product Description, Unit Price)
- INVOICE LINE ITEM (Invoice Number (PK), Product Number (PK), Quantity, ExtAmount)

Dạng chuẩn 3 (3NF)

- Cho quan hệ R , F là tập phụ thuộc hàm định nghĩa trên R .
- R được gọi là đạt dạng chuẩn 3 nếu thoả mãn các điều kiện sau:
 - R phải đạt dạng chuẩn 2NF
 - Mọi thuộc tính không khoá không phụ thuộc bắc cầu vào khóa (tức là tất cả các thuộc tính phải được suy ra trực tiếp từ khóa).

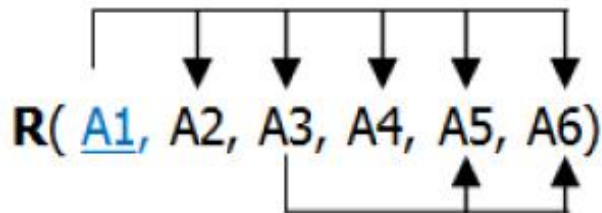
Ví dụ:

- Ví dụ: quan hệ $R(\underline{\text{MSKH}}, \text{TÊNKH}, \text{TP}, \text{PVC})$, $F_R =$
 $\{\underline{\text{MSKH}}\} \rightarrow \{\text{TÊNKH}, \text{TP}, \text{PVC}\}$ MSKH là khóa chính
 $\{\text{TP}\} \rightarrow \text{PVC}$
- chưa đạt chuẩn NF3 vì có thuộc tính PVC được suy ra từ thuộc tính TP (không phải là khóa).

Quy tắc chuẩn hóa về 3NF

- Bước 1: Loại bỏ các thuộc tính phụ thuộc bắc cầu ra khỏi quan hệ và tách chúng thành một quan hệ riêng có khoá chính là thuộc tính bắc cầu.
- Bước 2: Các thuộc tính còn lại lập thành một quan hệ có khóa chính là quan hệ ban đầu.

Quy tắc chuẩn hóa về 3NF



Khóa chính là $\{A1\}$ (tức là từ $\{A1\} \rightarrow \{A2, A3, A4, A5, A6\}$)

Có phụ thuộc hàm phụ thuộc bắc cầu $\{A3\} \rightarrow \{A5, A6\}$

Quan hệ được tách thành hai quan hệ sau

$R1(A3, A5, A6)$ - Tách các thuộc tính phụ thuộc bắc cầu

$R2(A1, A2, A3, A4)$ - Bảng ban đầu còn lại

Ví dụ 1: Chuẩn hóa về 3NF

- Ví dụ: Quan hệ R(MSKH, TÊNKH, TP, PVC) được tách thành 2 bảng
- R-VC (TP, PVC)
- R-KHÁCHHÀNG (MSKH, TÊNKH, TP)

KH			VC	
<u>MSKH</u>	TÊNKH	TP	<u>TP</u>	PVC
S1	An	HCM	HCM	01
S2	Hà	HN	HN	02
S3	Bình	NT	NT	03
S4	Trang	NT		

$F_{KH} = \{MSKH \rightarrow TÊNKH, TP\}$

$F_{VC} = \{TP \rightarrow PVC\}$

MH		
<u>MSMH</u>	TÊNMH	ĐG
P1	Táo	650
P2	Cam	500
P3	Chanh	450

$F_{MH} = \{MSMH \rightarrow TÊNMH, ĐG\}$

ĐH		
<u>MSKH</u>	<u>MSMH</u>	SL
S1	P1	300
S1	P2	200
S1	P3	400
S2	P1	100
S2	P3	300
S3	P2	200
S4	P2	210

$F_{ĐH} = \{MSKH, MSMH \rightarrow SL\}$

Ví dụ 2: Chuẩn hóa về 3NF

Cho quan hệ sau:

NHÂNVIÊN_ĐƠNVI(Mã sốNV, Họ tênNV, Ngày sinh, Địa chỉ, Mã sốĐV, TênĐV, Mã sốNQL)

Có chứa các phụ thuộc hàm dưới đây:

{Mã sốNV} → {Họ tênNV, Ngày sinh, Địa chỉ, Mã sốĐV, TênĐV, Mã sốNQL} Mã sốNV là khóa chính của bảng

{Mã sốĐV} → {TênĐV, Mã sốNQL}

Chứa phụ thuộc hàm bắc cầu vào thuộc tính không phải khóa

<u>Mã sốNV</u>	Họ tênNV	Ngày sinh	Địa chỉ	Mã sốĐV	TênĐV	Mã sốNQL
NV01	Tùng	1/1/1983	16 Thụy Khuê HN	DV01	Phòng tin học	NV02
NV02	Cúc	2/2/1980	10 Hàng Đào HN	DV01	Phòng tin học	NV02
NV03	Trúc	3/3/1983	Đà Nẵng	DV02	Hành chính	NV05
NV04	Mai	5/5/1982	Hà Giang	DV02	Hành chính	NV05
NV05	Sen	6/6/1986	Hải Phòng	DV02	Hành chính	NV05
NV06	Hồng	6/8/1984	Hà Nam	DV03	Kinh doanh	NV06

- Quan hệ này được tách thành các quan hệ sau

ĐƠNVI (Mã sốĐV, TênĐV, Mã sốNQL)

NHÂNVIÊN (Mã sốNV, Họ tênNV, Ngày sinh, Địa chỉ, Mã sốĐV)

Ví dụ 2: Chuẩn hóa về 3NF

ĐƠN VỊ

<u>Mã số DV</u>	Tên DV	Mã số NQL
DV01	Phòng tin học	NV02
DV02	Hành chính	NV05
DV03	Kinh doanh	NV06

NHÂN VIÊN

<u>Mã số NV</u>	Họ tên NV	Ngày sinh	Địa chỉ	Mã số DV
NV01	Tùng	1/1/1983	16 Thụy Khuê HN	DV01
NV02	Cúc	2/2/1980	10 Hàng Đào HN	DV01
NV03	Trúc	3/3/1983	Đà Nẵng	DV02
NV04	Mai	5/5/1982	Hà Giang	DV02
NV05	Sen	6/6/1986	Hải Phòng	DV02
NV06	Hồng	6/8/1984	Hà Nam	DV03

Ví dụ 3: Chuẩn hóa về 3NF

- Xét quan hệ **INVOICE** đã được tách như sau:
 - **INVOICE**: **Invoice Number** (PK), Customer Number, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone, Terms, Ship Via, Order Date, Total Order Amount
- Ta thấy có chứa thuộc tính bắc cầu sau
- {Customer Number} -> {Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone}
- Bảng này được tách thành hai bảng sau:
- **CUSTOMER** (**Customer Number**, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone)
- **INVOICE** (**Invoice Number** (PK), Customer Number, Terms, Ship Via, Order Date, Total Order Amount)

Ví dụ 3: Chuẩn hóa về 3NF

■ Cách làm:

- Chuẩn hóa quan hệ INVOICE về dạng chuẩn 1NF
- Tiếp tục chuẩn hóa về dạng 2NF
- Chuẩn hóa về dạng 3NF

Chuẩn hóa về 1NF

- Sau khi chuyển sang dạng chuẩn 1NF, CSDL sẽ gồm hai quan hệ như sau:
 - **Quan hệ INVOICE:** Invoice Number (PK), Customer Number, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone, Terms, Ship Via, Order Date, Total Order Amount
 - **Quan hệ INVOICE LINE ITEM:** Invoice Number (PK), Product Number (PK), Product Description, Quantity, Unit Price, Extended Amount

Chuẩn hóa về 2NF

- Ví dụ 2: Quan hệ INVOICE được phân rã về dạng 2NF như sau:

- **INVOICE**: Invoice Number (PK), Customer Number, Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone, Terms, Ship Via, Order Date, Total Order Amount
- **INVOICE LINE ITEM**: Invoice Number (PK), Product Number (PK), Quantity, Sale Unit Price, Extended Amount
- **PRODUCT**: Product Number (PK), Product Description, List Unit Price

Chuẩn hóa về 3NF

- Quan hệ INVOICE được phân rã về dạng 3NF như sau:
 - **INVOICE**: Invoice Number (PK), Customer Number, Terms, Ship Via, Order Date
 - **INVOICE LINE ITEM**: Invoice Number (PK), Product Number (PK), Quantity, Sale Unit Price
 - **PRODUCT**: Product Number (PK), Product Description, List Unit Price
 - **CUSTOMER** : Customer Number (PK), Customer Name, Customer Address, Customer City, Customer State, Customer Zip Code, Customer Phone

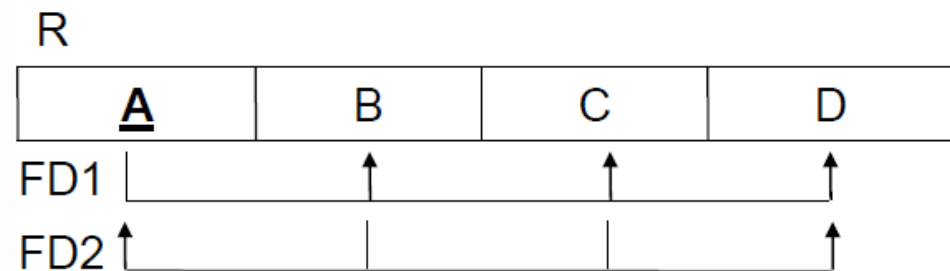
Dạng chuẩn Boyce-Codd

- ❑ Một lược đồ quan hệ R được gọi là ở dạng chuẩn Boyce-Codd (BCNF) nếu nó
 - Thỏa mãn dạng 3NF
 - $\forall X \rightarrow Y \in F^+$ thì X là siêu khóa (chứa khóa của quan hệ) hoặc $Y \subset X$. Nói cách khác, quan hệ đạt R sẽ không đạt BCNF nếu tồn tại phụ thuộc hàm mà vế trái không phải là khóa

❑ Ví dụ 1

f1: $A \rightarrow BCD$

f2: $BC \rightarrow AD$



Dạng chuẩn Boyce-Codd

□ Ví dụ 2:

Cho quan hệ R (A1,A2,A3,A4,A5)

Với các phụ thuộc hàm:

$$f_1: A1, A2 \rightarrow A3, A4, A5$$

$$f_2: A4 \rightarrow A2$$

- R đạt dạng chuẩn 3NF (Không có thuộc tính không khóa phụ thuộc bắc cầu vào khóa)
- R chưa đạt BCNF vì $f_2: A4 \rightarrow A2$ có vế trái không phải siêu khóa

Dạng chuẩn Boyce-Codd

□ Ví dụ 3: Cho quan hệ
SV_MH_GV(MaSV, MONHOC, GIANGVIEN)

Với các phụ thuộc hàm:

$f_1: \text{MaSV} \rightarrow \text{MonHoc}, \text{GiangVien}$

$f_2: \text{MonHoc} \rightarrow \text{Giangvien}$

- R chưa đạt BCNF vì f_2 có vế trái không phải siêu khóa

Dạng chuẩn Boyce-Codd

❑ Nếu một lược đồ quan hệ không thỏa mãn điều kiện BCNF, thủ tục chuẩn hóa bao gồm:

- Loại bỏ các thuộc tính khóa phụ thuộc hàm vào thuộc tính không khóa ra khỏi quan hệ
- tách chúng thành một quan hệ riêng có khoá chính là thuộc tính không khóa gây ra phụ thuộc.

❑ Ví dụ trên: R (A1, A2, A3, A4, A5)

Với các phụ thuộc hàm:

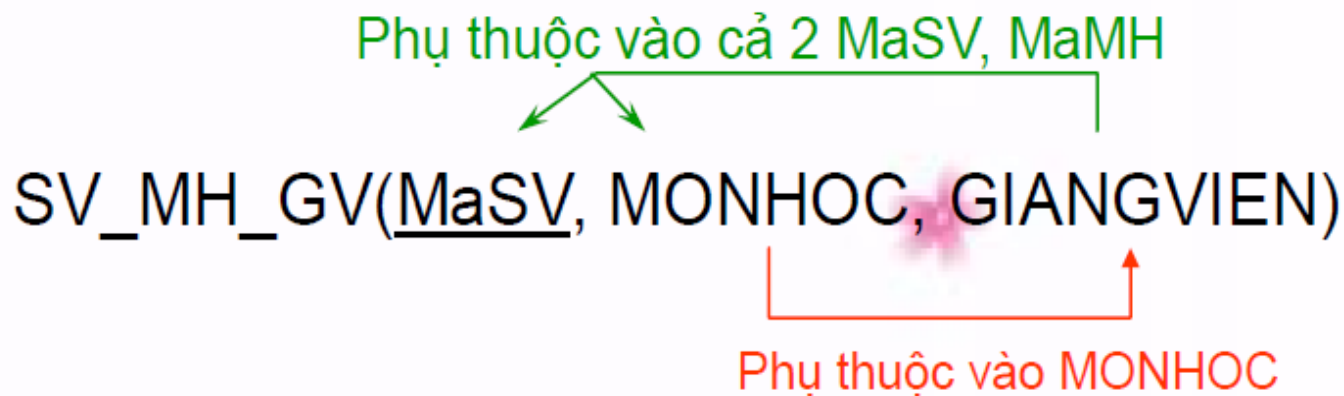
- $A1, A2 \rightarrow A3, A4, A5$
- $A4 \rightarrow A2$

❑ lược đồ được tách ra như sau:

- $R1(\underline{A4}, A2)$
- $R2(\underline{A1}, \underline{A4}, A3, A5)$

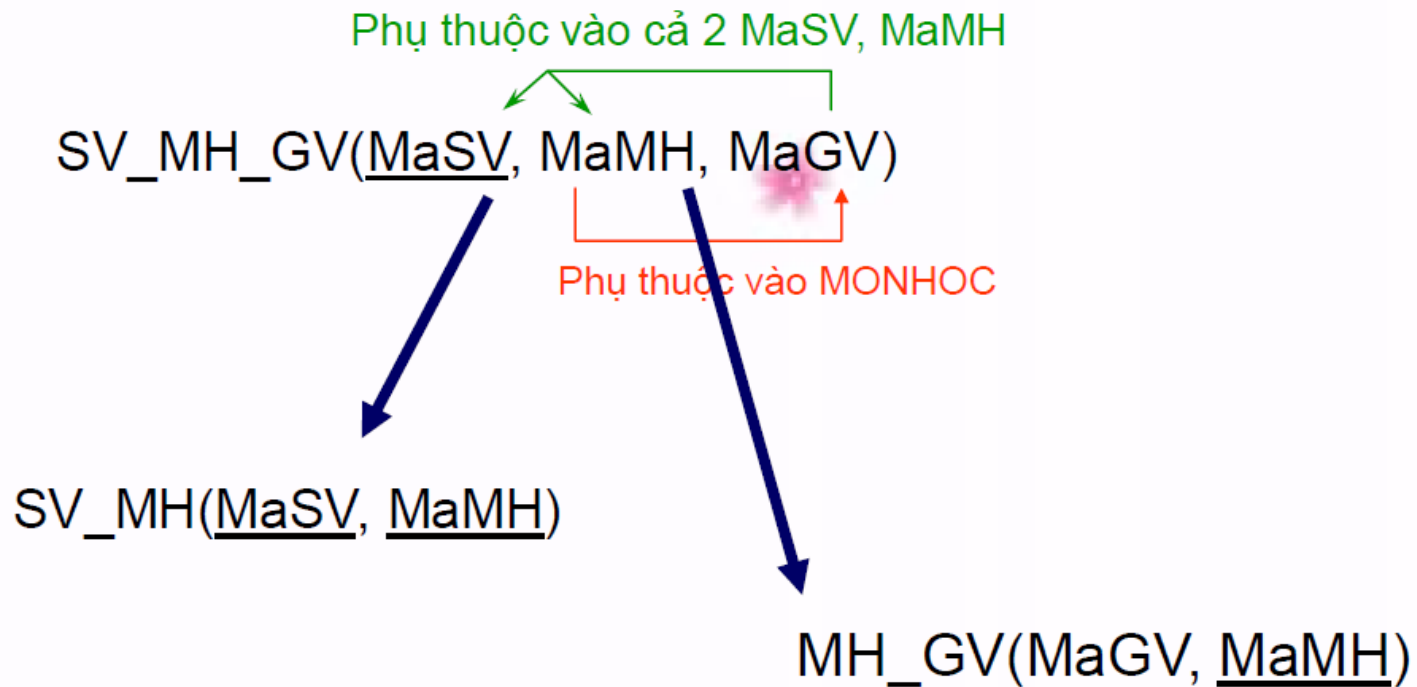
Dạng chuẩn Boyce-Codd

■ Ví dụ



Dạng chuẩn Boyce-Codd

■ Ví dụ



Khử chuẩn (Denormalization)

- Quá trình chuẩn hóa làm tăng thêm số lượng các bảng và các liên kết.
- Khi người sử dụng CSDL gặp vấn đề về hiệu suất mà không thể được giải quyết được bằng các phương tiện khác (chẳng hạn như điều chỉnh các cơ sở dữ liệu hoặc nâng cấp phần cứng) -> Cần phải khử chuẩn.
- Các bước khử chuẩn:
 - Kết hợp lại các quan hệ đã chia để thỏa mãn các quy tắc chuẩn hóa
 - Lưu trữ dữ liệu dư thừa trong các bảng
 - Lưu trữ dữ liệu tóm tắt trong các bảng

Tổng kết

- Các thao tác trên dữ liệu (thêm, sửa, xóa) trên CSDL chưa chuẩn hóa có thể gây ra các bất thường trên CSDL.
- Chuẩn hóa CSDL để đảm bảo dữ liệu không dư thừa và loại bỏ các bất thường trên CSDL
- 3 bước chuẩn hóa thông thường:
 - Chuẩn hóa về dạng 1NF: Loại bỏ dữ liệu dư thừa
 - Chuẩn hóa về dạng 2NF: Loại bỏ các phụ thuộc hàm bộ phận
 - Chuẩn hóa về dạng 3NF: Loại bỏ các phụ thuộc hàm bắc cầu