Chương 7

Phụ thuộc hàm
và
Chuẩn hóa cơ sở dữ liệu

Nội dung trình bày

- Phụ thuộc hàm.
- Các dạng chuẩn.
- Một số thuật toán chuẩn hóa.

Vấn đề khi thiết kế CSDL quan hệ

- Làm thế nào để thiết kế CSDL quan hệ với các lược đồ quan hệ tốt nhất?
  - Chúng ta cần đặc biệt quan tâm đến mối ràng buộc giữa các dữ liệu trong quan hệ, đó chính là các phụ thuộc hàm.
- Xét ví dụ một lựa chọn chưa tốt

NHANVIEN_DUAN

MaNV	TenNV	…	MaDA	TenDA	DiaDiem	SoGio
123456789	Hung		1	Du an X	Tan Binh	8
123456789	Hung		2	Du an Y	Phu Nhuan	4
333445555	Nghia		1	Du an X	Tan Binh	4
333445555	Nghia		2	Du an Y	Phu Nhuan	4
444556666	Bao		3	Du an Z	Go Vap	8

Vấn đề khi thiết CSDL quan hệ

- Dư thừa dữ liệu: thông tin về nhân viên và dự án bị lặp lại nhiều lần.
- Dữ liệu mất tính nhất quán: là hệ quả của dư thừa dữ liệu (nếu sửa tên Hung ở bộ dữ liệu thứ nhất thành Long thì không nhất quán với bộ dữ liệu thứ 2).
- Dị thường khi thêm bộ: nếu muốn thêm dữ liệu của một nhân viên mới (chưa tham gia dự án nào) thì không được vì khóa chính gồm 2 thuộc tính MaNV, MaDA.
- Dị thường khi xóa bộ: nếu xóa bộ dữ liệu cuối cùng thì thông tin về Dự án Z cũng mất.

Chuẩn hóa lược đồ quan hệ

- Quá trình chuẩn hoá lấy một lược đồ quan hệ và thực hiện các kiểm tra để xác nhận nó có thoả mãn một dạng chuẩn nào đó hay không.
- Chuẩn hóa có thể được xem như một quá trình phân tích các lược đồ quan hệ cho trước dựa trên các phụ thuộc hàm và các khoá chính của chúng để đạt đến các tính chất mong muốn:
  - Tối thiểu sự dư thừa và
  - Tối thiểu các dị thường khi cập nhật dữ liệu.
- Các lược đồ quan hệ không thỏa mãn kiểm tra dạng chuẩn sẽ được tách ra thành các lược đồ quan hệ nhỏ hơn thỏa mãn các kiểm tra và có các tính chất mong muốn.

Dạng chuẩn 1 (1NF)

- Một quan hệ ở dạng chuẩn 1 nếu các giá trị của tất cả các thuộc tính trong quan hệ là nguyên tử (1 giá trị tại một thời điểm).

PHONGBAN

TenPB	MaPB	TrPhg	Truso
Nghien cuu	5	333445555	Tan Binh, Thu Duc
Hanh chinh	4	987654321	Go Vap

Vi phạm dạng chuẩn 1

PHONGBAN

TenPB	MaPB	TrPhg	Truso
Nghien cuu	5	333445555	Tan Binh
Nghien cuu	5	333445555	Thu Duc
Hanh chinh	4	987654321	Go Vap

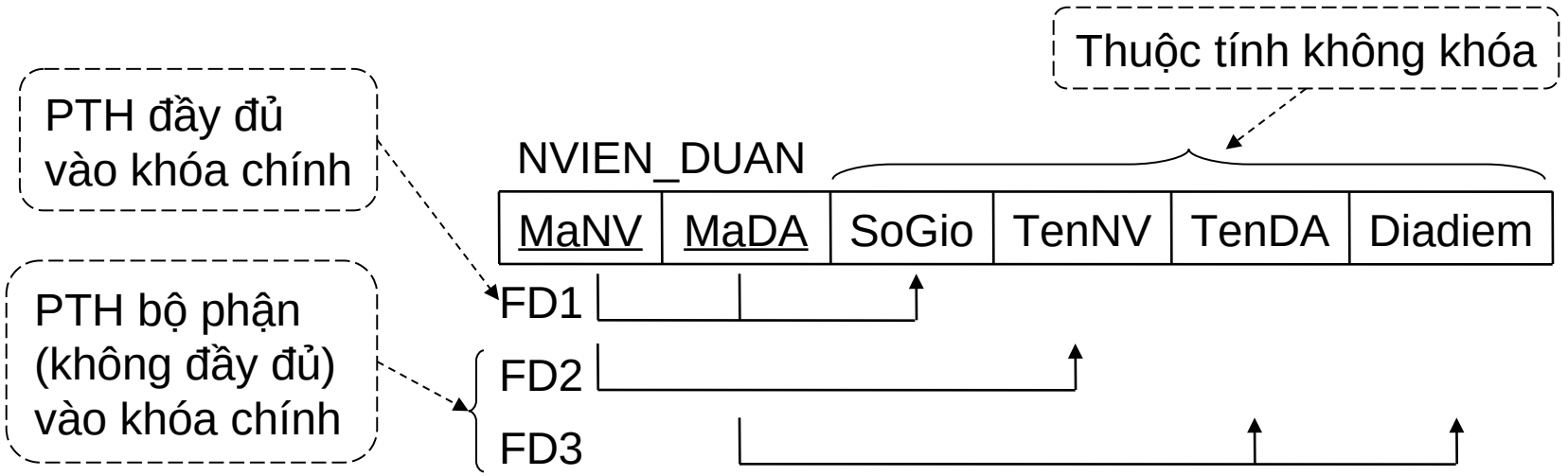
Ở dạng chuẩn 1

Dạng chuẩn 1 (1NF)

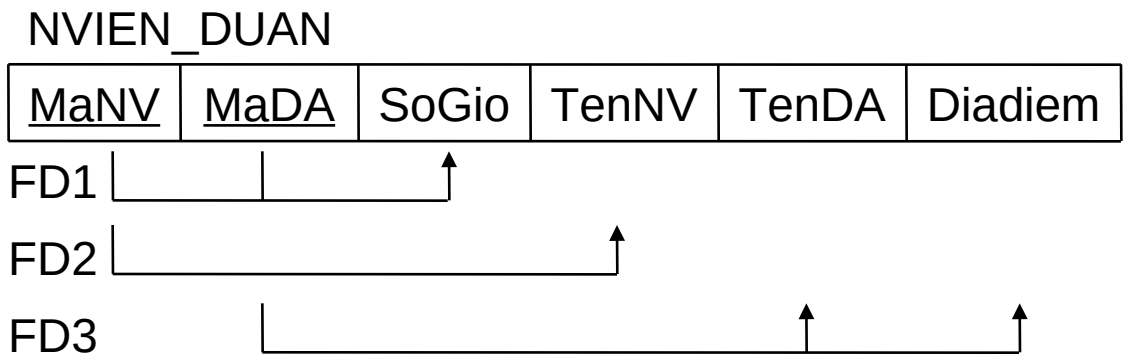
- Nhận xét
  - Mọi quan hệ đều thỏa dạng chuẩn 1 vì mô hình dữ liệu quan hệ không cho phép thuộc tính đa trị.
  - Dạng chuẩn 1 có thể tồn tại sự dư thừa dữ liệu. Do đó gây ra các dị thường về cập nhật dữ liệu.

Dạng chuẩn 2 theo khóa chính (2NF)

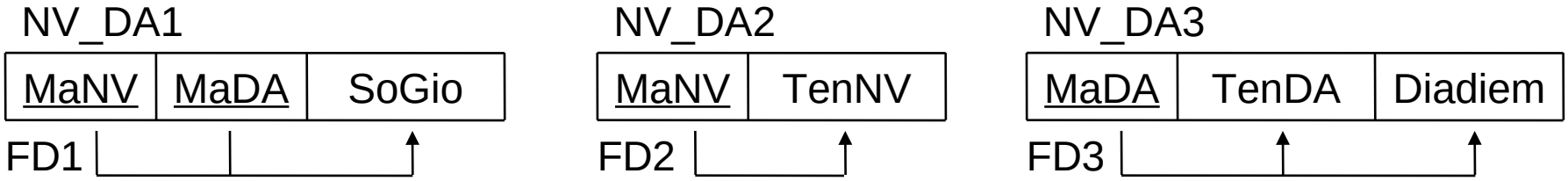
- Một quan hệ ở dạng chuẩn 2 nếu tất cả các *thuộc tính không khóa phụ thuộc hàm đầy đủ* vào khóa chính.
- Cho quan hệ R(U, Ғ), K là khóa chính của R
  - A ∈ U là *thuộc tính không khóa* nếu A ∉ K.
  - X → Y là *phụ thuộc hàm đầy đủ* nếu với thuộc tính A bất kỳ, A ∈ X thì (X - {A}) → Y không còn đúng.
  - X → Y là *phụ thuộc hàm bộ phận* nếu tồn tại một thuộc tính A ∈ X và (X - {A}) → Y vẫn đúng.



Dạng chuẩn 2 theo khóa chính (2NF)

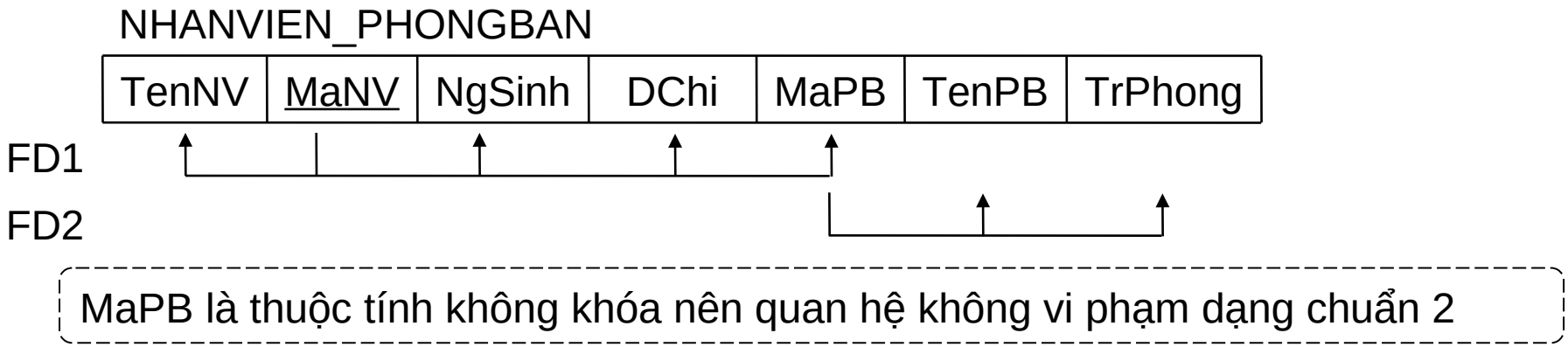


Quan hệ NVIEN_DUAN vi phạm dạng chuẩn 2



Các quan hệ NV_DA1, NV_DA2, NV_DA3 ở dạng chuẩn 2

Dạng chuẩn 2 theo khóa chính (2NF)

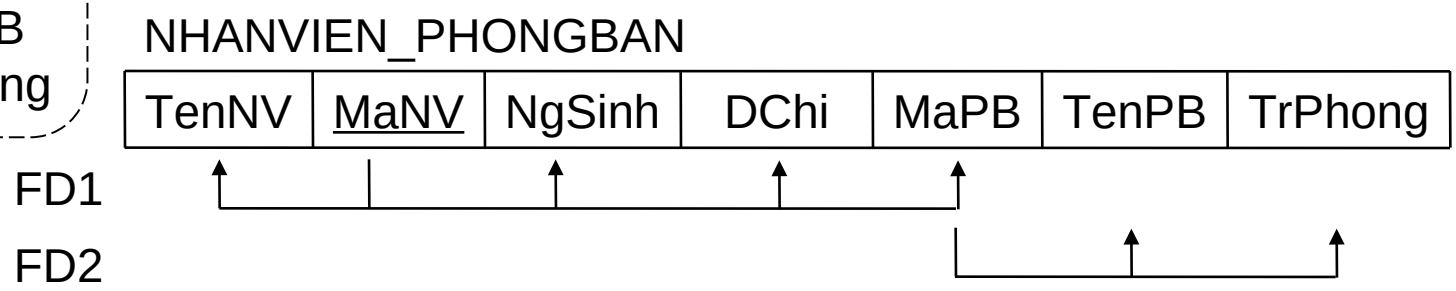


- Nhận xét
  - Mọi quan hệ ở dạng chuẩn 2 cũng ở dạng chuẩn 1.
  - Nếu quan hệ chỉ có một khóa và khóa chỉ gồm một thuộc tính thì ở dạng chuẩn 2.
  - Quan hệ ở dạng chuẩn 2 vẫn có thể tồn tại sự dư thừa dữ liệu.

Dạng chuẩn 3 theo khóa chính (3NF)

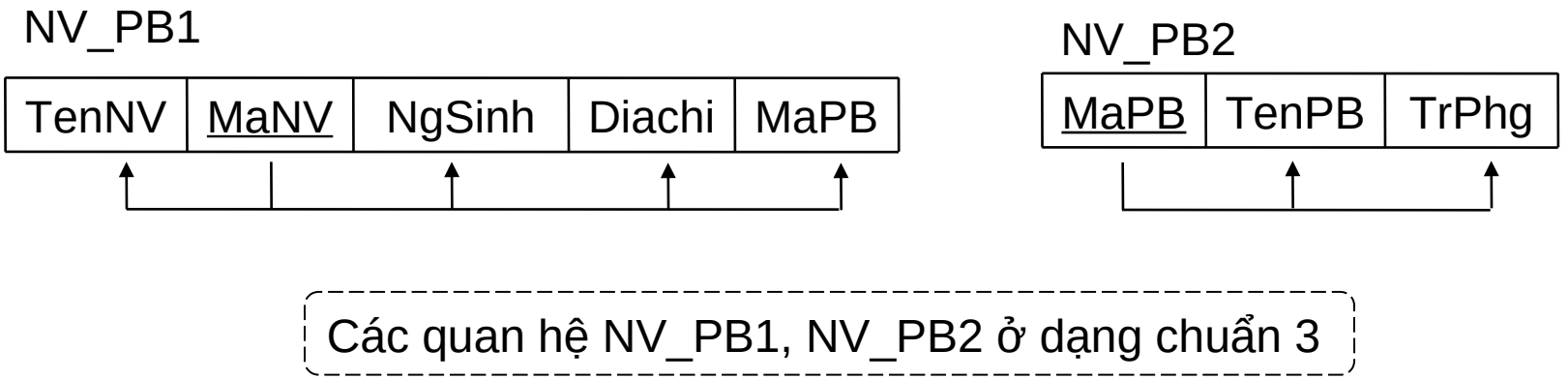
- Một quan hệ ở dạng chuẩn 3 nếu quan hệ ở dạng chuẩn 2 và *không có thuộc tính không khóa nào phụ thuộc hàm bắt* cầu vào khóa chính.
- Cho quan hệ R(U, Ғ)
  - X → Y là *phụ thuộc hàm bắt cầu* nếu tồn tại tập thuộc tính Z ⊆ U, Z không là khóa và cũng không là tập con của một khóa nào mà cả X → Z và Z → Y đều đúng.

Gây ra các PTH bắt cầu vào khóa chính:
MaNV → TenPB
MaNV → TrPhong



Quan hệ NHANVIEN_PHONGBAN vi phạm dạng chuẩn 3

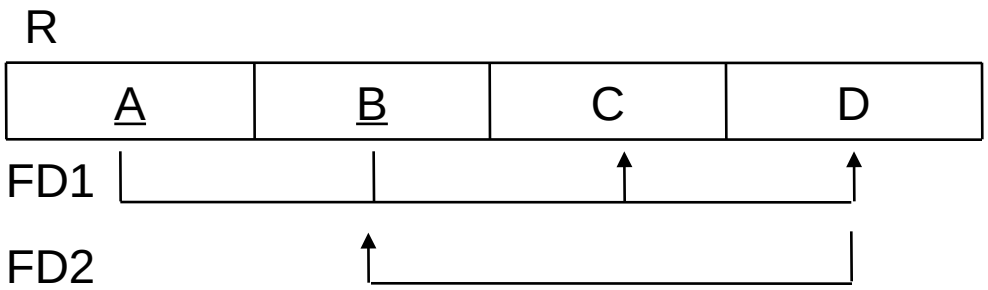
Dạng chuẩn 3 theo khóa chính (3NF)



- Nhận xét
  - Quan hệ ở dạng chuẩn 3 cũng ở dạng chuẩn 2.
  - Một cơ sở dữ liệu tốt thì các quan hệ tối thiểu phải ở dạng chuẩn 3.
  - Quan hệ ở dạng chuẩn 3 vẫn có thể tồn tại sự dư thừa dữ liệu nếu có hai khóa dự tuyển giao nhau hoặc có thuộc tính không khóa xác định thuộc tính khóa.

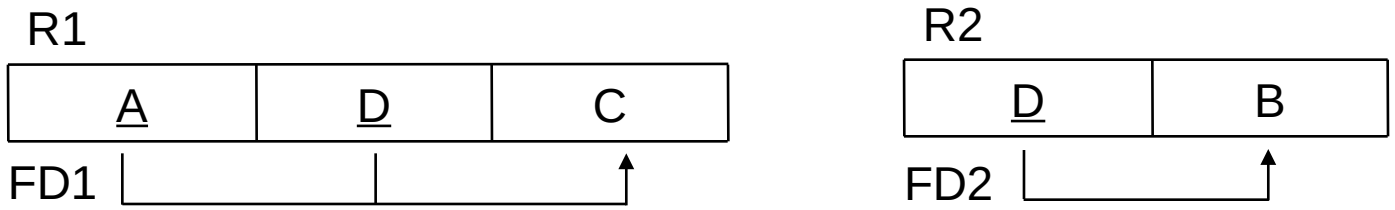
Dạng chuẩn Boyce-Codd theo khóa chính (BCNF)

- Một quan hệ là ở dạng chuẩn Boyce-Codd nếu quan hệ ở dạng chuẩn 3 và không có các thuộc tính khóa phụ thuộc hàm vào thuộc tính không khóa.
- Nhận xét:
  - Thuộc tính khóa phụ thuộc hàm vào thuộc tính không khóa có thể gây ra sự dư thừa dữ liệu.



Quan hệ R vi phạm dạng chuẩn Boyce-Codd do PTH D → B

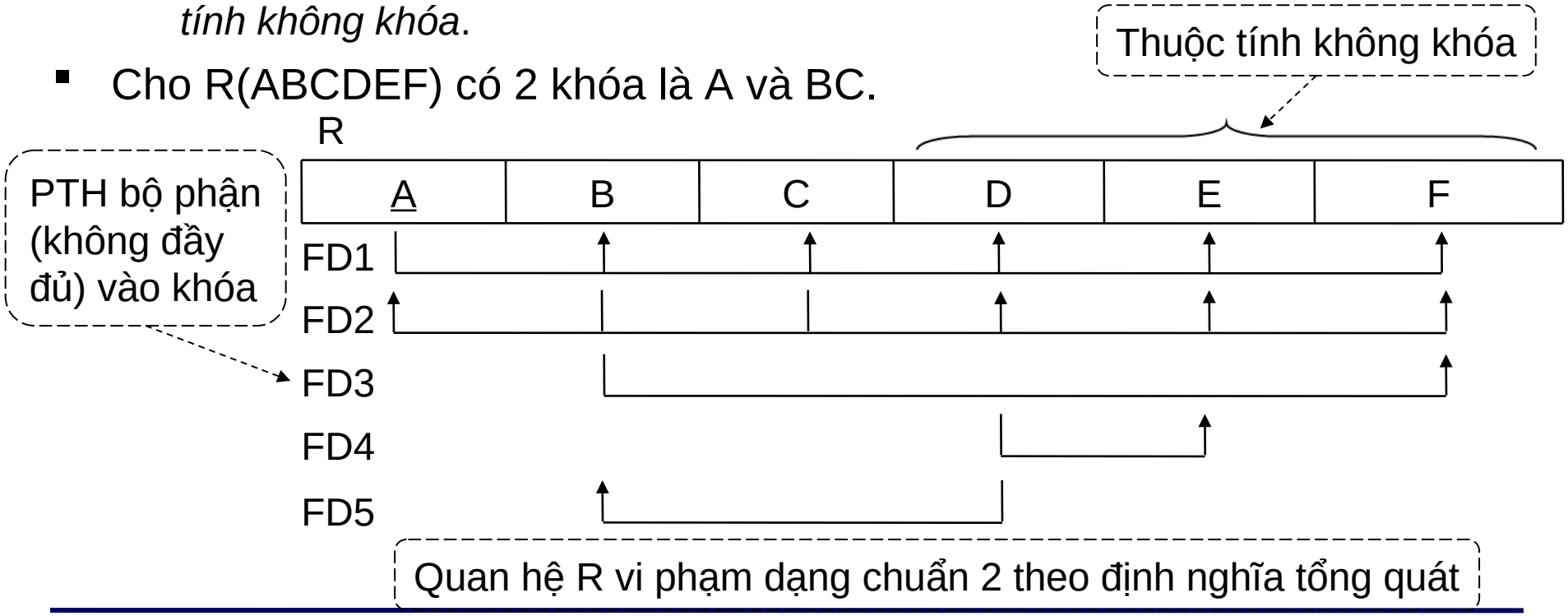
Dạng chuẩn Boyce-Codd theo khóa chính (BCNF)



Các quan hệ R1, R2 ở dạng chuẩn Boyce-Codd

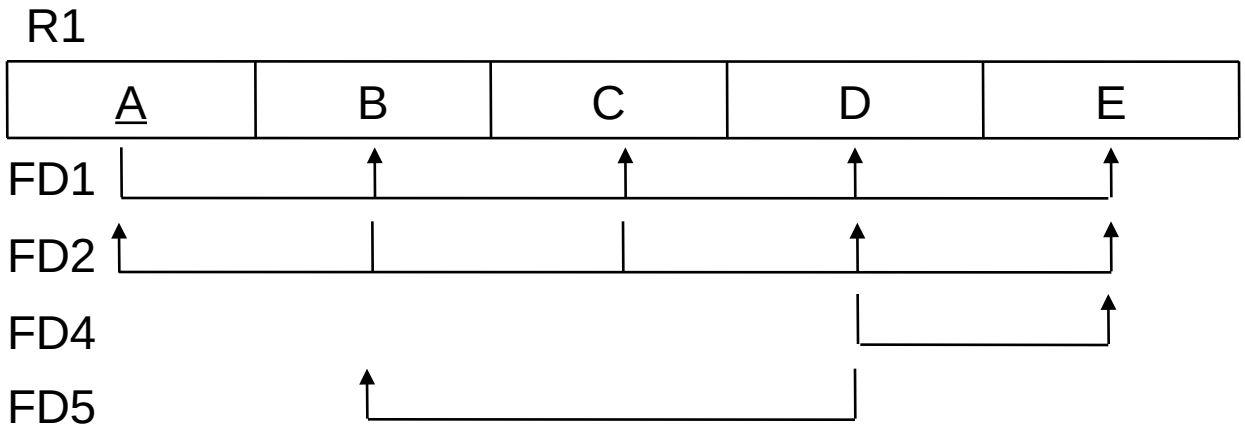
Dạng chuẩn 2 tổng quát

- Một quan hệ ở dạng chuẩn 2 nếu tất cả các *thuộc tính không khóa phụ thuộc hàm đầy đủ* vào *tất cả các khóa*.
- Cho quan hệ R(U, Ғ), định nghĩa lại khái niệm thuộc tính khóa cho trường hợp tổng quát
  - A ∈ U là *thuộc tính khóa* nếu A thuộc một khóa nào đó. Ngược lai A là *thuộc tính không khóa*.
- Cho R(ABCDEF) có 2 khóa là A và BC.



Dạng chuẩn 3 tổng quát

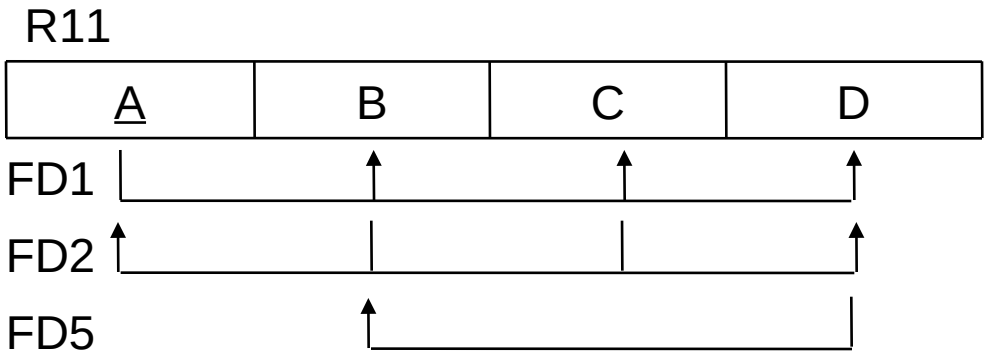
- Một quan hệ ở dạng chuẩn 3 nếu nếu một phụ thuộc hàm X → A đúng thì
  - X là một siêu khóa, hoặc
  - A là thuộc tính khóa.
- Cho R1(ABCDE) có 2 khóa là A và BC.



Quan hệ R1 vi phạm dạng chuẩn 3 theo định nghĩa tổng quát vì PTH D → E trong đó:
- D không là siêu khóa, và
- E không là thuộc tính khóa

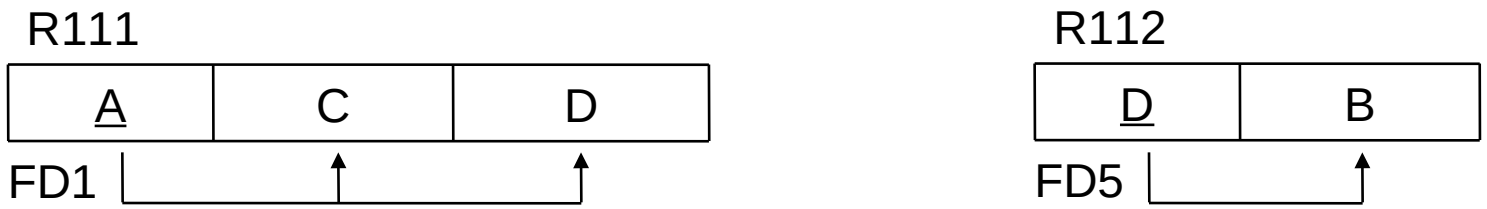
Dạng chuẩn Boyce - Codd tổng quát

- Một quan hệ là ở dạng chuẩn Boyce-Codd nếu một phụ thuộc hàm X → Y đúng thì X là một siêu khóa.
- Cho R11(ABCD) có 2 khóa là A và BC.



Lược đồ R11 thuộc dạng chuẩn 3,nhưng không thuộc dạng chuẩn BC

Dạng chuẩn Boyce - Codd tổng quát



Các quan hệ R111, R112 ở dạng chuẩn Boyce-Codd

- Nhận xét
  - Quan hệ ở dạng chuẩn Boyce-Codde cũng ở dạng chuẩn 3.

Nội dung trình bày

- Phụ thuộc hàm.
- Các dạng chuẩn.
- Một số thuật toán chuẩn hóa.

Cách tiếp cận thiết kế CSDL

- Thiết kế trên – xuống (Top – Down design) hay được sử dụng trong thiết kế ứng dụng cơ sở dữ liệu thương mại
  - Thiết kế lược đồ mức khái niệm bằng mô hình dữ liệu cấp cao (ER).
  - Ánh xạ lược đồ khái niệm vào một tập hợp các quan hệ.
  - Các quan hệ được kiểm tra, phân tích dựa trên các phụ thuộc hàm và khóa chính đã xác định theo các dạng chuẩn với khóa chính để loại bỏ các phụ thuộc hàm bộ phận và bắt cầu.
  - Việc phân tích cũng có thể được thực hiện trong giai đoạn thiết kế mức khái niệm.

Cách tiếp cận thiết kế CSDL

- Thiết kế dưới – lên (Bottom – Up design) là một phương án thiết kế lược đồ cơ sở dữ liệu một cách chặt chẽ:
  - Xây dựng một quan hệ phổ quát chứa tất cả các thuộc tính của cơ sở dữ liệu
  - Xác định tất cả các phụ thuộc hàm giữa các thuộc tính dựa trên các quy tắc dữ liệu.
  - Áp dụng các thuật toán chuẩn hóa để tổng hợp các lược đồ quan hệ. Mỗi lược đồ quan hệ riêng rẽ ở dạng chuẩn 3NF hoặc BCNF hoặc ở dạng chuẩn cao hơn.

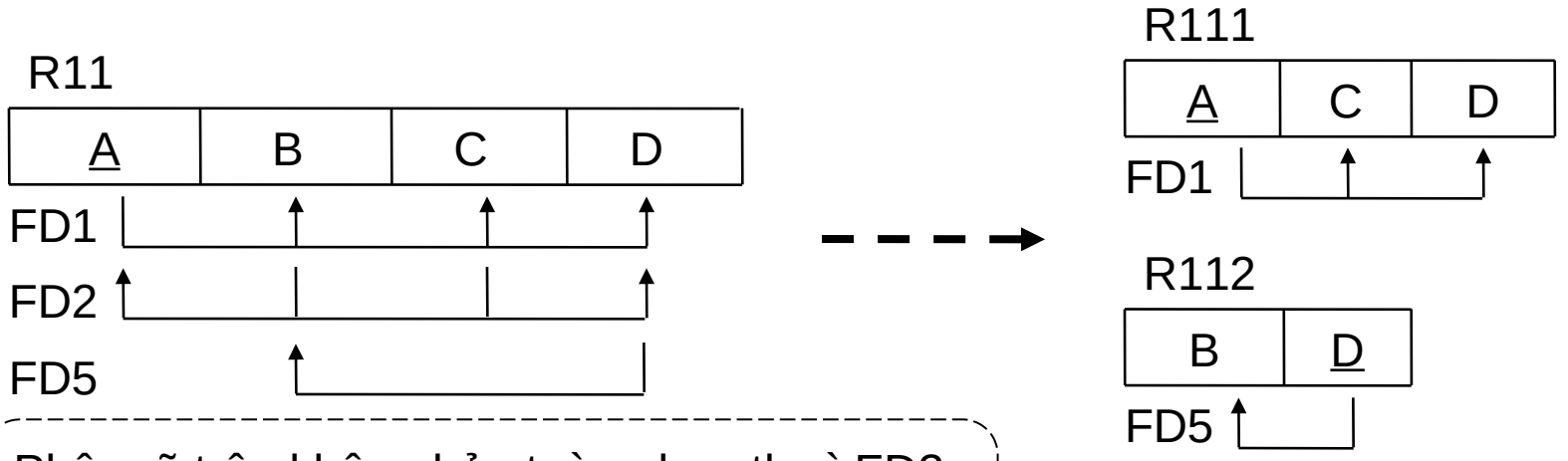
Phân rã quan hệ

Cho lược đồ quan hệ phổ quát R với tập hợp tất cả các thuộc tính của cơ sở dữ liệu $\{A_1, \ldots, A_n\}$ và tập phụ thuộc hàm F xác định trên các thuộc tính của R.

- Phân rã quan hệ là quá trình sử dụng các thuật toán chuẩn hóa để tách quan hệ phổ quát R thành một tập hợp các quan hệ $D = \{R_1, \ldots, R_m\}$. D được gọi là một phân rã của R.
- Một phân rã phải đảm bảo:
  - Bảo toàn thuộc tính nghĩa là mỗi thuộc tính của R phải xuất hiện trong ít nhất một R_i.
  - Các quan hệ R_i phải ở dạng chuẩn 3NF hoặc BCNF.

Phân rã bảo toàn phụ thuộc hàm

- Cho quan hệ R(U) và tập pth F. Giả sử D = {R_1, …, R_m} là một phân rã của R.
  - Phép chiếu của F trên R_i là $\pi_{Ri}(F) = \{X \rightarrow Y \in F^+ : X \cup Y \subset R_i\}$.
  - D được gọi là phân rã bảo toàn phụ thuộc hàm đối với F nếu $(\pi_{R1}(F) \cup \ldots \cup \pi_{Rm}(F))^+ = F^+$. Nghĩa là các pth trong F hoặc xuất hiện trực tiếp trong các quan hệ R_i hoặc được suy diễn từ các pth trong R_i.



Phân rã trên không bảo toàn phụ pth vì FD2 không xuất hiện trong R111, R112 và Không được suy diễn từ các pth trong R111, R112

Phân rã bảo toàn phụ thuộc hàm

R11	A	B	C	D
	1	α	β	2
	2	β	γ	3
	3	α	δ	2

R111	A	C	D
	1	β	2
	2	γ	3
	3	δ	2
	4	β	4

R112	D	B
	2	α
	3	β
	4	α

A	B	C	D
1	α	β	2
...	...	...	...
4	α	β	4

Thêm bộ (4, β, 4) vào R111 và (4, α) vào R112
thì trạng thái csdl sẽ không thỏa PTH FD2

Thuật toán phân rã bảo toàn PTH

- Định lý 7.1

  Luôn tồn tại một phân rã bảo toàn pth D = {R_1, …, R_m} đối với tập phụ thuộc hàm F của một quan hệ phổ quát R sao cho các R_i ở dạng chuẩn 3NF.

- Thuật toán 7.4
  - Nhập: R(U), U = {A_1, …, A_n} và tập pth F.
  - Xuất: Phân rã D = {R_1, …, R_m}, R_i ở dạng chuẩn 3NF.
  - *B1*:

    Tìm phủ tối thiểu G của F.
  - *B2*:

    Với mỗi $X \rightarrow A_i \in G$, xây dựng quan hệ $R_i(U_i)$, $U_i = X \cup \{A_j\}$. Khóa chính của R_i là X.

Thuật toán phân rã bảo toàn PTH

- *B3*:

  Giả sử xong B2 ta có các quan hệ R_1, …, R_m. Nếu ($U_1 \cup \ldots \cup U_m$) $\neq$ U thì xây dựng thêm lược đồ $R_{m+1}(U_{m+1})$, $U_{m+1} = U - (U_1 \cup \ldots \cup U_m)$. Khóa chính của R_{m+1} là U_{m+1}.

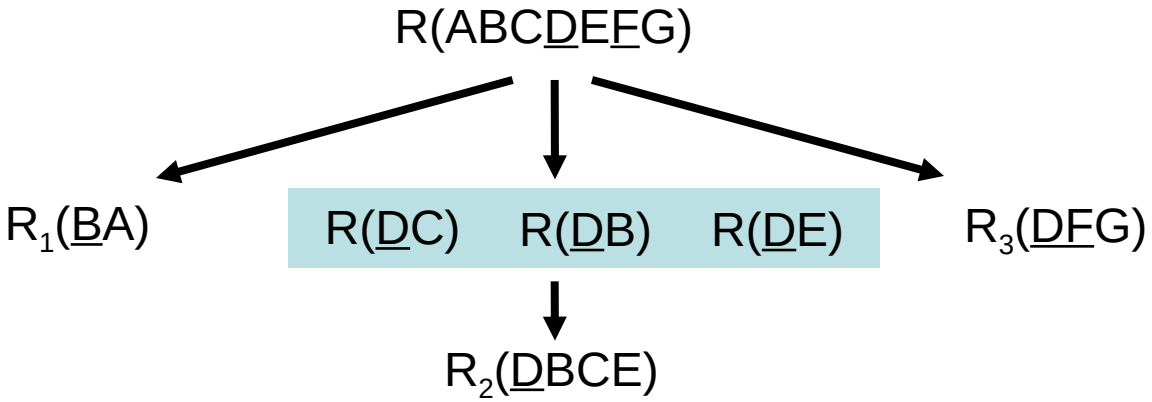
- *B4*:

  Phân rã D là tập hợp các quan hệ R_i.

Ví dụ phân rã bảo toàn PTH

- Cho R(ABCDEFG), Ғ = {B → A, D → C, D → EB, DF → G}
- Tìm phân rã bảo toàn pth của R
  - *Tìm phủ tối thiểu*

    G = {B → A, D → C, D → B, D → E, DF → G}.
  - *Cây phân rã*



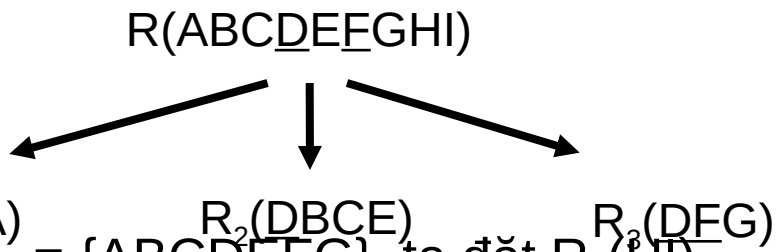
  - *Phân rã bảo toàn pth của R*

    Xuất D = {R_1, R_2, R_3}.

Ví dụ phân rã bảo toàn PTH

- Cho R(ABCDEFGHI), Ғ = {B → A, D → C, D → EB, DF → G}
- Tìm một phân rã bảo toàn pth của R
  - *Tìm phủ tối thiểu*

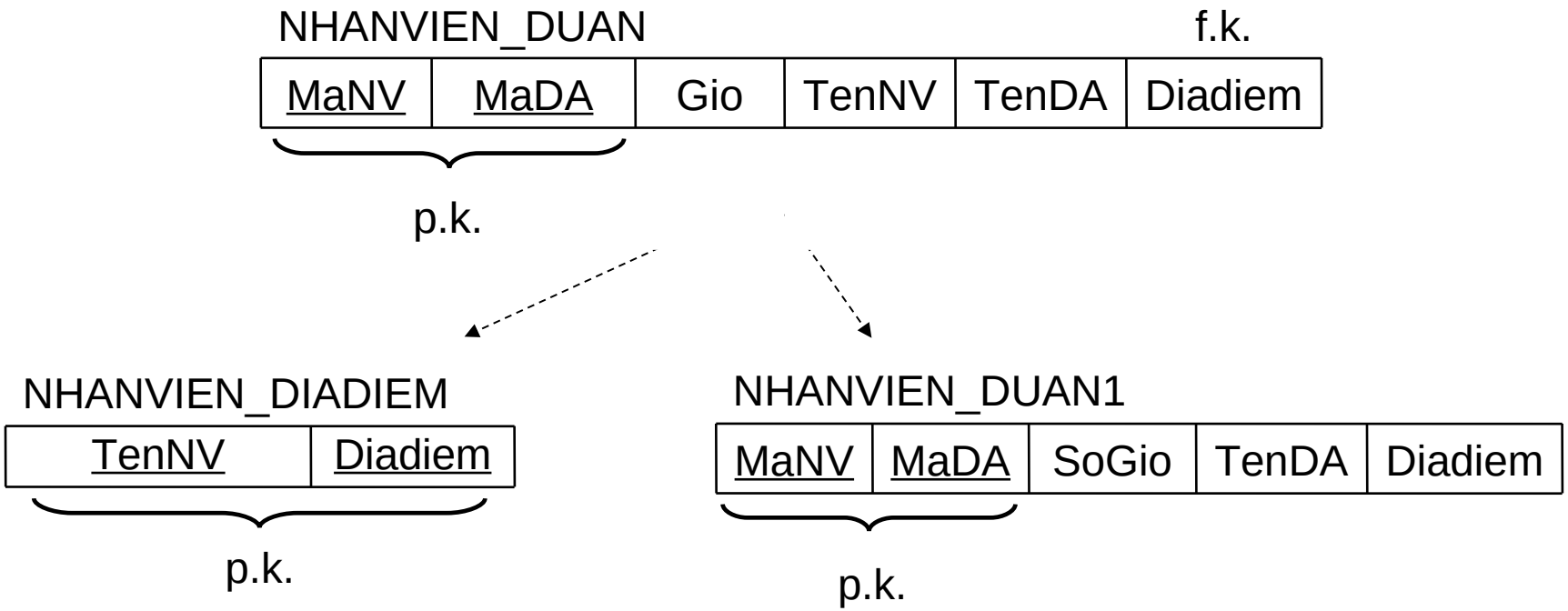
    G = {B → A, D → C, D → B, D → E, DF → G}.
  - *Cây phân rã*



  - Vì $U_1 \cup U_2 \cup U_3$ = {ABCDEFG}, ta đặt R_4(HI).
  - *Phân rã bảo toàn pth của R*

    D = {R_1, R_2, R_3, R_4}.

Phân rã không mất thông tin



NHANVIEN_DUAN

MaNV	MaDA	Gio	TenNV	TenDA	Diadiem
123456789	1	32.5	Hung	San pham X	Tan Binh
123456789	2	7.5	Hung	San pham Y	Thu Duc
333445555	2	10	Nghia	San pham Y	Thu Duc

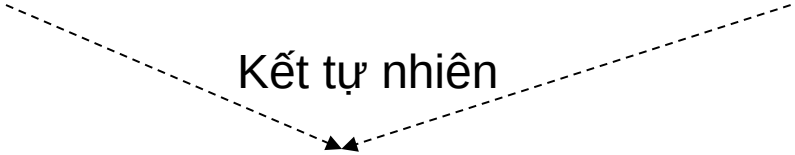
Phân rã không mất thông tin

NHANVIEN_DIADIEM

TenNV	Diadiem
Hung	Tan Binh
Hung	Thu Duc
Nghia	Thu Duc

NHANVIEN_DUAN1

MaNV	MaDA	SoGio	TenDA	Diadiem
123456789	1	32.5	San pham X	Tan Binh
123456789	2	7.5	San pham Y	Thu Duc
333445555	2	10	San pham Y	Thu Duc



MaNV	MaDA	Gio	TenDA	Diadiem	TenNV
123456789	1	32.5	San pham X	Tan Binh	Hung
123456789	2	7.5	San pham Y	Thu Duc	Hung
123456789	2	7.5	San pham Y	Thu Duc	Nghia
333445555	2	10	San pham Y	Thu Duc	Hung
333445555	2	10	San pham Y	Thu Duc	Nghia

Phân rã không mất thông tin

- Cho quan hệ R(U) và tập pth F. Giả sử D = {R_1, …, R_m} là một phân rã của R.
  - D được gọi là phân rã không mất thông tin đối với F nếu với mọi trạng thái r ∈ R thì (π_{R1}(r) * … * π_{Rm}(r)) = r.
  - π_{Ri}(r) là phép chiếu trạng thái r lên R_i.
- Định lý 7.2

  Phân rã D = {$R_1(U_1)$, $R_2(U_2)$} của R(U) không mất thông tin đối với tập PTH F nếu và chỉ nếu:
  - $(U_1 \cap U_2) \rightarrow (U_1 - U_2) \in F^+$, hoặc
  - $(U_1 \cap U_2) \rightarrow (U_2 - U_1) \in F^+$.
- Định lý 7.3

  Nếu phân rã D = {R_1, …, R_m} của R không mất thông tin đối với F và phân rã D_i = {Q_1, …, Q_k} của R_i không mất thông tin đối với π_{Ri}(F) thì D' = {R_1, …, R_{i-1}, Q_1, …, Q_k, R_{i+1}, …, R_m} của R cũng không mất thông tin.

Thuật toán phân rã không mất thông tin

- Thuật toán 7.5
  - Nhập: R(U), U = {A_1, …, A_n} và tập PTH F.
  - Xuất: D = {R_1, …, R_m}, R_i ở dạng chuẩn BCNF.
  - *B1*:

    D = {R};
  - *B2*:

    -Nếu có lược đồ $Q(U_Q) \in D$ không ở dạng chuẩn BC thì

    + Tìm $X \rightarrow Y \in \pi_Q(F)$ làm Q vi phạm điều kiện BC.

    + $D = (D - \{Q\}) \cup Q_1(U_{Q1}) \cup Q_2(U_{Q2})$ với $U_{Q1} = U_Q - Y$ và $U_{Q2} = X \cup Y$.

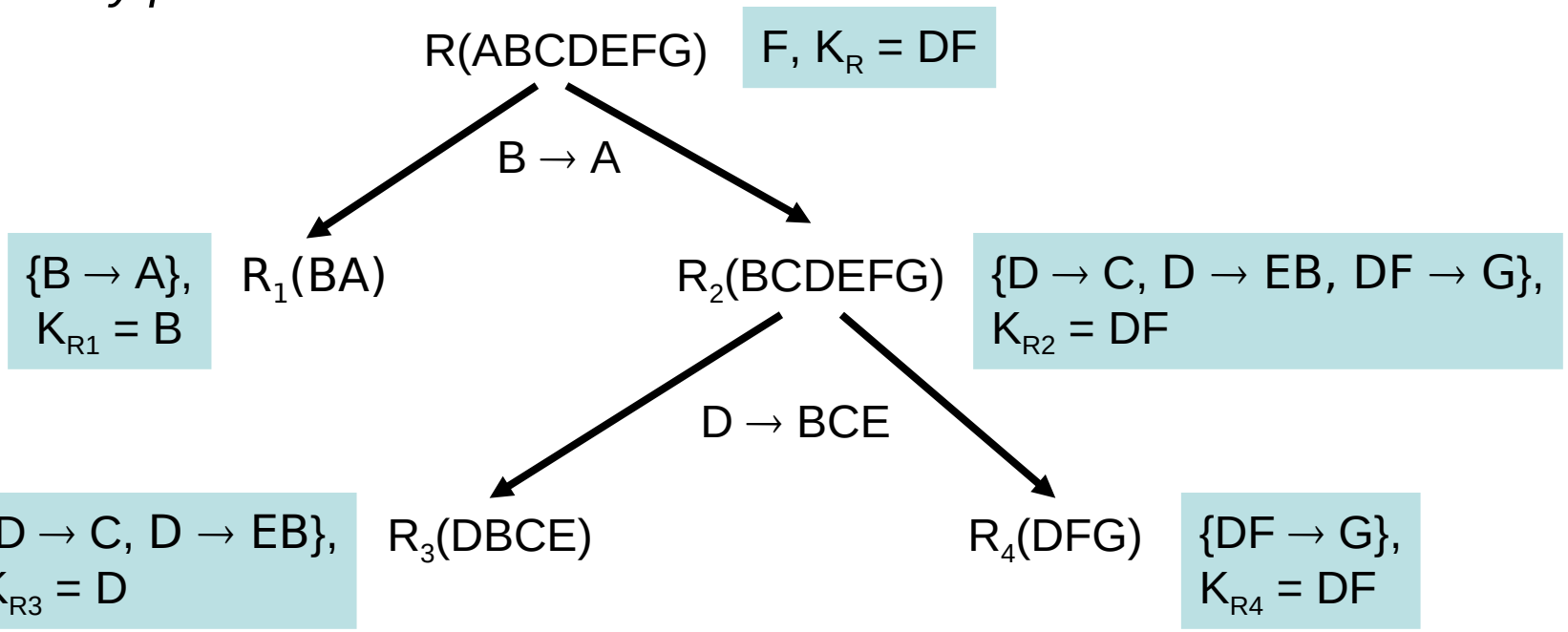
    + Quay lại B2.

    -Ngược lại, chuyển sang B3.
  - *B3*:

    D là phân rã không mất thông tin của R

Ví dụ phân rã không mất thông tin

- Cho R(ABCDEFG), F = {B → A, D → C, D → EB, DF → G}
- Tìm một phân rã không mất thông tin của R
  - *Cây phân rã*



  - *Phân rã không mất thông tin của R*

    D = {R_1, R_3, R_4}

Ví dụ phân rã không mất thông tin