CHUẨN HÓA DỮ LIỆU

ThS. Trần Thị Hồng Yến

yentth@uit.edu.vn

0907380471

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Đặt vấn đề

- Xét quan hệ:

  NV_DuAn(MaNV, MaDA, TenNV, TenDA, SoGio)

MaNV	MaDA	TenNV	TenDA	SoGio
01	A	Tuấn	Billing	100
01	C	Tuấn	Salary	50
02	A	Hiền	Billing	80
02	B	Hiền	Order	150
03	A	Vũ	Billing	90
03	C	Vũ	Salary	70
04	D	Hà	Sale	120

- **Dư thừa dữ liệu (Redundancy)**: Thông tin về nhân viên và dự án bị lặp lại nhiều lần. Nếu nhân viên có mã 01 tham gia 10 dự án thì thông tin về nhân viên này bị lặp lại 10 lần, tương tự đối với dự án có mã A, nếu có 1000 nhân viên tham gia thì thông tin về dự án cũng lặp lại 1000 lần.

Đặt vấn đề

- Xét quan hệ:

  NV_DuAn(MaNV, MaDA, TenNV, TenDA, SoGio)

MaNV	MaDA	TenNV	TenDA	SoGio
01	A	Tuấn	Billing	100
01	C	Tuấn	Salary	50
02	A	Hiền	Billing	80
02	B	Hiền	Order	150
03	A	Vũ	Billing	90
03	C	Vũ	Salary	70
04	D	Hà	Sale	120

- **Không nhất quán (Inconsistency)**: Là hệ quả của dư thừa dữ liệu. Giả sử sửa bản ghi thứ nhất, tên nhân viên được sửa thành Tú thì dữ liệu này lại không nhất quán với bản ghi thứ 2 (vẫn có tên là Tuấn), nên phải sửa tên nhân viên Tú cho bản ghi 2 và tất cả các bản ghi khác có tên nhân viên này.

Đặt vấn đề

- Xét quan hệ:

  NV_DuAn(MaNV, MaDA, TenNV, TenDA, SoGio)

MaNV	MaDA	TenNV	TenDA	SoGio
01	A	Tuấn	Billing	100
01	C	Tuấn	Salary	50
02	A	Hiền	Billing	80
02	B	Hiền	Order	150
03	A	Vũ	Billing	90
03	C	Vũ	Salary	70
04	D	Hà	Sale	120

- **Dị thường khi thêm bộ (Insertion anomalies)**: Nếu muốn thêm thông tin một nhân viên mới (chưa tham gia dự án nào) vào quan hệ thì không được vì khoá chính của quan hệ trên gồm 2 thuộc tính MaNV và MaDA.

Đặt vấn đề

- Xét quan hệ:

  NV_DuAn(MaNV, MaDA, TenNV, TenDA, SoGio)

MaNV	MaDA	TenNV	TenDA	SoGio
01	A	Tuấn	Billing	100
01	C	Tuấn	Salary	50
02	A	Hiền	Billing	80
02	B	Hiền	Order	150
03	A	Vũ	Billing	90
03	C	Vũ	Salary	70
04	D	Hà	Sale	120

- **Dị thường khi xoá bộ (Deletion anomalies)**: Giả sử xoá đi bản ghi cuối cùng, thì thông tin về nhân viên có mã số 04 sẽ bị xóa, kéo theo thông tin về dự án có mã là D cũng mất.

⇒ ***Nên tìm cách tách quan hệ NV_DuAn thành các quan hệ nhỏ hơn***.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Chuẩn hóa dữ liệu là gì?

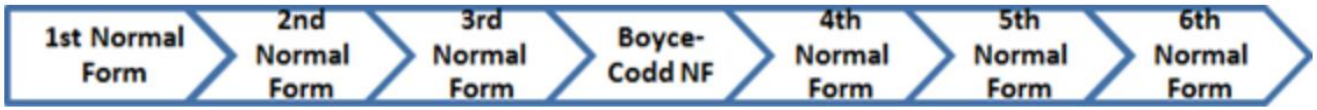
- Chuẩn hoá là quá trình tách bảng (phân rã) thành các bảng nhỏ hơn dựa vào các phụ thuộc hàm.
- Các dạng chuẩn là các chỉ dẫn để thiết kế các bảng trong CSDL.
- Mục đích của chuẩn hoá là loại bỏ các dư thừa dữ liệu và các lỗi khi thao tác dư thừa và các lỗi khi thao tác dữ liệu (Insert, Delete, Update).

Nội dung

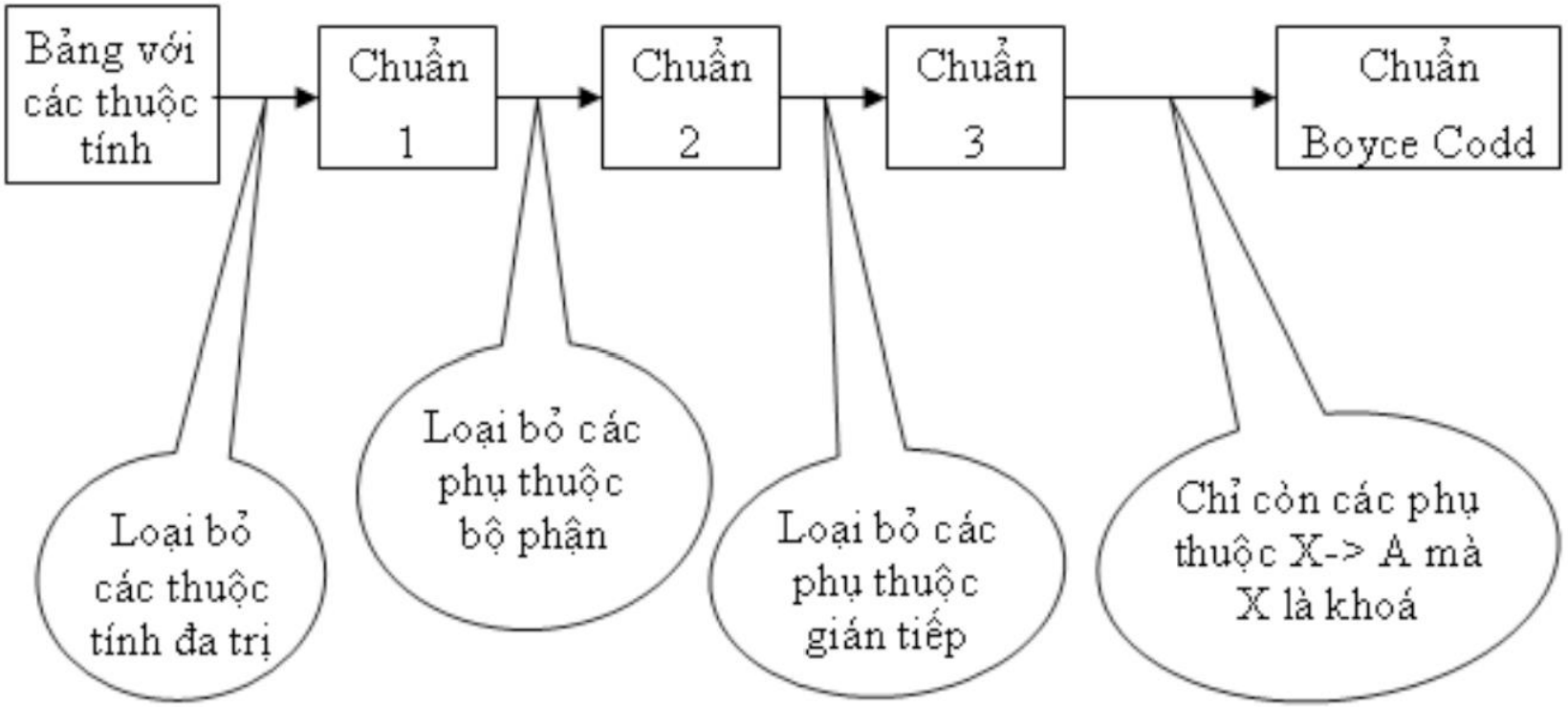
- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Các dạng chuẩn hóa (Normal Form)

- Người phát minh ra mô hình quan hệ Edgar Codd đã đề xuất lý thuyết chuẩn hóa dữ liệu với sự ra đời của dạng chuẩn 1 và ông tiếp tục mở rộng lý thuyết với dạng chuẩn 2 và dạng chuẩn 3.
- Sau đó, ông tham gia với Raymond F. Boyce để phát triển lý thuyết về dạng chuẩn Boyce-Codd.
- Lý thuyết về chuẩn hóa dữ liệu vẫn đang được phát triển thêm cho đến dạng chuẩn 6.
- Trong các ứng dụng thực tế, chuẩn hóa đạt được kết quả tốt nhất ở dạng chuẩn 3:



Các dạng chuẩn hóa (Normal Form)



Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Dạng chuẩn 1 – 1NF (First Normal Form)

- Định nghĩa:
  - Lược đồ quan hệ R ở dạng chuẩn 1 (1NF - First Normal Form) nếu mọi thuộc tính của R đều chứa các giá trị nguyên tố (atomic value), nghĩa là giá trị này **không là** một danh sách các giá trị hoặc các giá trị phức hợp (composite value).
  - Lưu ý:
    - Các thuộc tính của quan hệ R
      - Không là thuộc tính đa trị (multivalued attribute).
      - Không là thuộc tính phức hợp (composite attribute).
    - Mỗi dòng (bản ghi) phải duy nhất => Khóa chính (primary key)

Dạng chuẩn 1 – 1NF (First Normal Form)

- VD1: Xét quan hệ Customer:

Customer ID	Firstname	Surname	Telephone Number
123	Pooja	Singh	555-861-2025, 192-122-1111
456	San	Zhang	(555) 403-1659 Ext.53; 182-929-2929
789	John	Doe	555-808-9633

- Quan hệ Customer không ở dạng chuẩn 1 vì thuộc tính Telephone Number là thuộc tính đa trị.
- ⇒ ***Chỉnh sửa về dạng chuẩn 1NF***:

Customer ID	Firstname	Surname	Telephone Number
123	Pooja	Singh	555-861-2025
123	Pooja	Singh	192-122-1111
456	San	Zhang	(555) 403-1659 Ext.53
456	San	Zhang	182-929-2929
789	John	Doe	555-808-9633

Dạng chuẩn 1 – 1NF (First Normal Form)

- VD2: Xét quan hệ NV_DuAn:

MaNV	HoTen	MaPhong	TenPhong	PhanCong	
NV1	An	P1	Dự án	DA1	100
NV1	An	P1	Dự án	DA2	80
NV1	An	P1	Dự án	DA3	120
NV2	Bình	P1	Dự án	DA1	70
NV2	Bình	P1	Dự án	DA2	90
NV3	Hạnh	P2	Marketing	DA1	90

- Quan hệ NV_DuAn không ở dạng chuẩn 1 vì thuộc tính PHANCONG là thuộc tính phức hợp.

Dạng chuẩn 1 – 1NF (First Normal Form)

- ⇒ ***Chỉnh sửa về dạng chuẩn 1NF***:

MaNV	TenNV	MaPhong	TenPhong	MaDA	SoGio
NV1	An	P1	Dự án	DA1	100
NV1	An	P1	Dự án	DA2	80
NV1	An	P1	Dự án	DA3	120
NV2	Bình	P1	Dự án	DA1	70
NV2	Bình	P1	Dự án	DA2	90
NV3	Hạnh	P2	Marketing	DA1	90

- Quan hệ NV_DuAn ở dạng chuẩn 1NF vì các thuộc tính của NV_DuAn không là thuộc tính đa trị, không là thuộc tính phức hợp.

Dạng chuẩn 1 – 1NF (First Normal Form)

MaNV	TenNV	MaPhong	TenPhong	MaDA	SoGio
NV1	An	P1	Dự án	DA1	100
NV1	An	P1	Dự án	DA2	80
NV1	An	P1	Dự án	DA3	120
NV2	Bình	P1	Dự án	DA1	70
NV2	Bình	P1	Dự án	DA2	90
NV3	Hạnh	P2	Marketing	DA1	90

- Các bất thường của quan hệ ở 1NF
  - Dư thừa dữ liệu (Redundancy):
    - Thông tin về nhân viên, phòng ban, dự án bị lặp lại nhiều lần.
  - Thêm:
    - Không thể thêm thông tin của nhân viên mới có mã là NV4, tên là Phúc, thuộc phòng có mã là P2 nếu nhân viên này chưa được phân công tham gia vào dự án nào.
  - Cập nhật:
    - Không thể sửa tên của nhân viên có tên là An với tên mới là Ái vì sẽ không nhất quán với bản ghi thứ 2 và 3 (vẫn có tên là An), nên phải sửa tên nhân viên Ái cho bản ghi 2, 3 và tất cả các bản ghi khác có tên nhân viên này.
  - Xóa:
    - Nếu xoá bản ghi cuối cùng, thì thông tin về NV tên Hạnh có mã NV3 bị xóa, kéo theo thông tin về phòng Marketing có mã P2 cũng mất.
- Nguyên nhân
  - Tồn tại thuộc tính không khóa phụ thuộc hàm riêng phần vào khóa.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

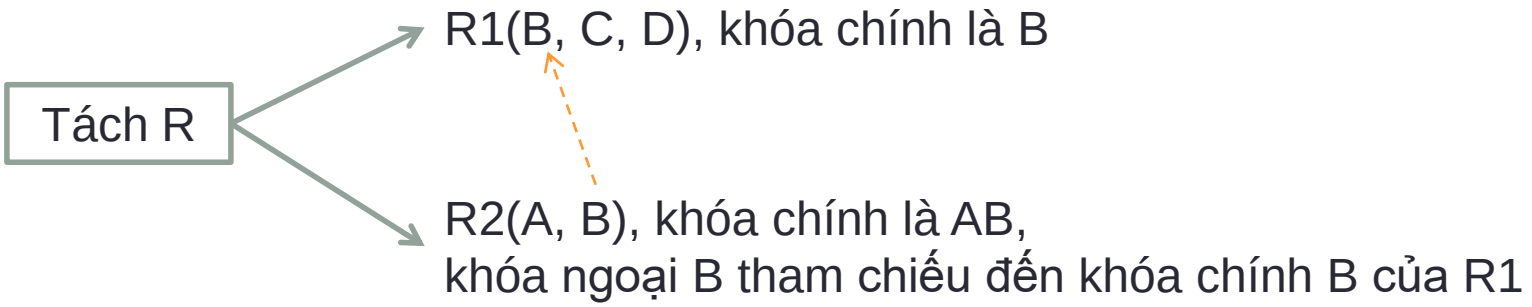
Dạng chuẩn 2 – 2NF (Second Normal Form)

- Định nghĩa:
  - Lược đồ quan hệ R ở dạng chuẩn 2 (2NF - Second Normal Form) đối với tập phụ thuộc hàm F nếu:
    - R ở dạng chuẩn 1.
    - Mọi thuộc tính không khóa đều phụ thuộc hàm đầy đủ vào mọi khóa của R.

Dạng chuẩn 2 – 2NF (Second Normal Form)

- VD3: Cho quan hệ R(A, B, C, D), khóa chính là AB và tập PTH: F={AB → C; AB → D}: đây là quan hệ đạt chuẩn 2NF.
- VD4: Cho quan hệ R(A, B, C, D), khóa chính là AB và tập PTH: F={AB → C; AB → D; B → CD}: đây là quan hệ không đạt chuẩn 2NF vì có B → CD là PTH không đầy đủ vào khóa chính AB.

⇒ ***Chỉnh sửa về dạng chuẩn 2NF***:



Dạng chuẩn 2 – 2NF (Second Normal Form)

- VD5: Xét quan hệ NV_DuAn:

MaNV	TenNV	MaPhong	TenPhong	MaDA	SoGio
NV1	An	P1	Dự án	DA1	100
NV1	An	P1	Dự án	DA2	80
NV1	An	P1	Dự án	DA3	120
NV2	Bình	P1	Dự án	DA1	70
NV2	Bình	P1	Dự án	DA2	90
NV3	Hạnh	P2	Marketing	DA1	90

Các phụ thuộc hàm:
MaNV → {TenNV, MaPhong}
MaPhong → TenPhong
{MaNV, MaDA} → SoGio

Khóa của NV_DuAn: {MaNV, MaDA}

- Lược đồ quan hệ NV_DuAn không đạt chuẩn 2NF vì thuộc tính không khóa TenNV phụ thuộc hàm riêng phần vào khóa {MaNV, MaDA} (do có PTH: MaNV → TenNV)

Dạng chuẩn 2 – 2NF (Second Normal Form)

⇒ ***Chỉnh sửa về dạng chuẩn 2NF***:

NHANVIEN

MaNV	TenNV	MaPhong	TenPhong
NV1	An	P1	Dự án
NV2	Bình	P1	Dự án
NV3	Hạnh	P2	Marketing

PHANCONG

MaNV	MaDA	SoGio
NV1	DA1	100
NV1	DA2	80
NV1	DA3	120
NV2	DA1	70
NV2	DA2	90
NV3	DA1	90

Khóa chính của NHANVIEN: **MaNV**
Khóa chính của PHANCONG: {**MaNV**, **MaDA**}
Khóa ngoại của PHANCONG: **MaNV** tham chiếu đến khóa chính của NHANVIEN: **MaNV**

- Lược đồ quan hệ R1 và R2 đều đạt chuẩn 2NF vì các thuộc tính không khóa đều phụ thuộc hàm đầy đủ vào khóa.

Dạng chuẩn 2 – 2NF (Second Normal Form)

NHANVIEN

MaNV	TenNV	MaPhong	TenPhong
NV1	An	P1	Dự án
NV2	Bình	P1	Dự án
NV3	Hạnh	P2	Marketing

PHANCONG

MaNV	MaDA	SoGio
NV1	DA1	100
NV1	DA2	80
NV1	DA3	120
NV2	DA1	70
NV2	DA2	90
NV3	DA1	90

- Các bất thường của quan hệ ở 2NF
  - Dư thừa dữ liệu (Redundancy):
    - Thông tin về phòng ban bị lặp lại nhiều lần.
  - Thêm:
    - Không thể thêm thông tin của phòng ban mới có mã là P3, tên là Kế hoạch, nếu phòng này chưa có nhân viên nào.
  - Cập nhật:
    - Không thể sửa tên phòng ban của bản ghi thứ 1 với tên mới là Quản lý dự án vì sẽ không nhất quán với bản ghi thứ 2 (vẫn có tên là Dự án), nên phải sửa phòng ban cho bản ghi thứ 2 và tất cả các bản ghi khác có tên phòng ban này .
  - Xóa:
    - Nếu xoá bản ghi cuối cùng, thì thông tin về NV tên Hạnh có mã NV3 bị xóa, kéo theo thông tin về phòng Marketing có mã P2 cũng mất.
- Nguyên nhân
  - Tồn tại thuộc tính không khóa phụ thuộc bắc cầu vào khóa.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

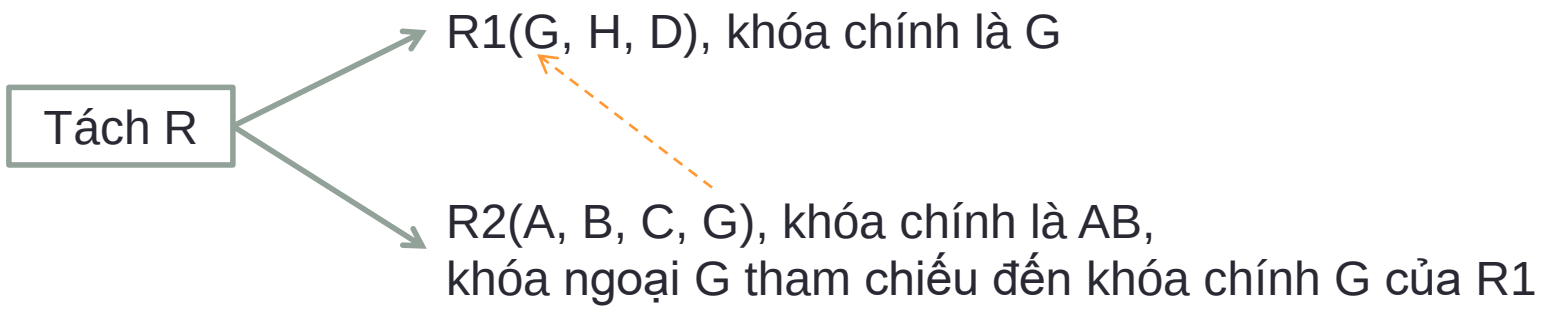
Dạng chuẩn 3 – 3NF (Third Normal Form)

- Định nghĩa:
  - Lược đồ quan hệ R ở dạng chuẩn 3 (3NF- Third Normal Form) đối với tập phụ thuộc hàm F nếu:
    - R ở dạng chuẩn 2NF.
    - Và mọi thuộc tính không khóa đều không phụ thuộc bắc cầu vào khóa của R (nghĩa là các thuộc tính không khóa phải phụ thuộc trực tiếp vào khóa chính)

Dạng chuẩn 3 – 3NF (Third Normal Form)

- VD6: Cho quan hệ R(A, B, C, D, G, H), khóa chính là AB và tập PTH: F={AB → C; AB → D, AB → GH}: đây là quan hệ đạt chuẩn 3NF.
- VD7: Cho quan hệ R(A, B, C, D, G, H), khóa chính là AB và tập PTH: F={AB → C; AB → D; AB → GH, G → DH} là quan hệ không đạt chuẩn 3NF vì có G → DH là PTH gián tiếp vào khóa chính AB.

⇒ ***Chỉnh sửa về dạng chuẩn 3NF***:



Dạng chuẩn 3 – 3NF (Third Normal Form)

- VD8: Xét quan hệ NV_DuAn:

MaNV	TenNV	MaPhong	TenPhong
NV1	An	P1	Dự án
NV1	An	P1	Dự án
NV1	An	P1	Dự án
NV2	Bình	P1	Dự án
NV2	Bình	P1	Dự án
NV3	Hạnh	P2	Marketing

- Lược đồ quan hệ R không đạt chuẩn 3NF vì thuộc tính không khóa TenPhong phụ thuộc bắc cầu vào khóa MaNV:

MaNV → MaPhong MaPhong → TenPhong

Dạng chuẩn 3 – 3NF (Third Normal Form)

⇒ ***Chỉnh sửa về dạng chuẩn 3NF***:

PHONG

MaPhong	TenPhong
P1	Dự án
P2	Marketing

NHANVIEN

MaNV	TenNV	MaPhong
NV1	An	P1
NV2	Bình	P1
NV3	Hạnh	P2

Khóa chính của PHONG: **MaPhong**
Khóa chính của NHANVIEN: **MaNV**
Khóa ngoại của NHANVIEN: **MaPhong** tham chiếu đến khóa chính của PHONG: **MaPhong**

- Lược đồ quan hệ R1 và R2 đều đạt chuẩn 3NF vì các thuộc tính không khóa đều không phụ thuộc bắc cầu vào khóa chính, mà phụ thuộc trực tiếp vào khóa chính.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

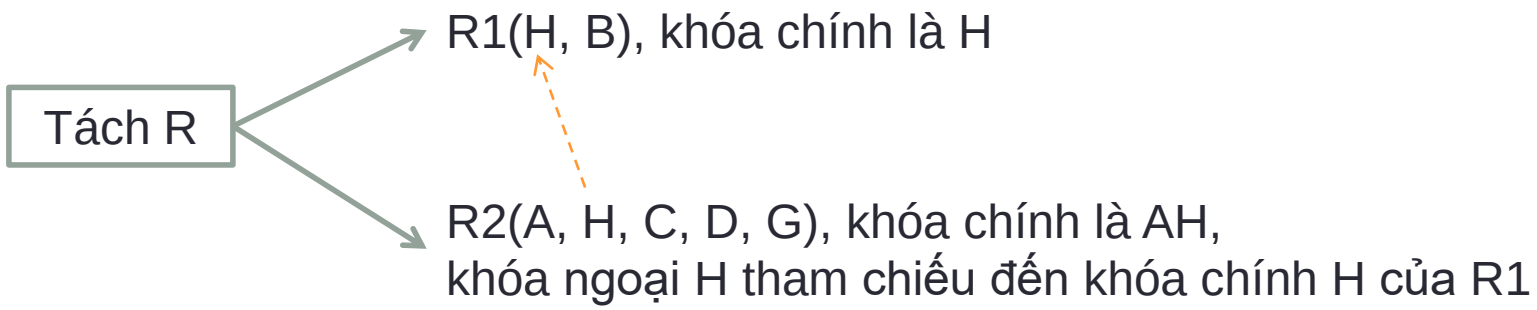
Dạng chuẩn BCNF (Boyce - Codd Normal Form)

- Định nghĩa:
  - Lược đồ quan hệ R ở dạng chuẩn Boyce-Codd (BCNF) đối với tập phụ thuộc hàm F nếu:
    - R ở dạng chuẩn 3.
    - Mọi phụ thuộc hàm $X \to A$ với $A \notin X$ thì X là một siêu khóa của R. (nghĩa là không có thuộc tính khóa mà phụ thuộc hàm vào thuộc tính không khóa.)
  - Nếu lược đồ quan hệ R ở dạng chuẩn Boyce-Codd thì R cũng ở dạng chuẩn 3.

Dạng chuẩn BCNF (Boyce - Codd Normal Form)

- VD9: Cho quan hệ R(A, B, C, D, G, H), khóa chính là AB và tập PTH: F={AB → C; AB → D; AB → GH}: đây là quan hệ đạt chuẩn BCNF.
- VD10: Cho quan hệ R(A, B, C, D, G, H), khóa chính là AB và tập PTH: F={AB → C; AB → D; AB → GH; H → B} là quan hệ không đạt chuẩn BCNF vì có thuộc tính khóa B phụ thuộc hàm vào thuộc tính không khóa H (H → B).

⇒ ***Chỉnh sửa về dạng chuẩn BCNF***:



Dạng chuẩn BCNF (Boyce - Codd Normal Form)

- VD11: Xét quan hệ NV_DuAn:

TenNV	TenDA	TruongNhom
An	Dự án 1	Tuấn
An	Dự án 2	Tú
Bình	Dự án 1	Tuấn
Bình	Dự án 2	Tú
Hạnh	Dự án 1	Tuấn
Phúc	Dự án 3	Minh

- Khóa là {TenNV, TruongNhom}
- Các phụ thuộc hàm:
  - TruongNhom $\rightarrow$ TenDA
  - {TenNV, TenDA} $\rightarrow$ TruongNhom
- Lược đồ quan hệ NV_DuAn không đạt chuẩn BCNF vì thuộc tính khóa TruongNhom phụ thuộc hàm vào thuộc tính không khóa TenDA.

Dạng chuẩn BCNF (Boyce - Codd Normal Form)

⇒ ***Chỉnh sửa về dạng chuẩn BCNF***:

QLY_NHOM

TruongNhom	TenDA
Tuấn	Dự án 1
Tú	Dự án 2
Minh	Dự án 3

QLY_NV

TenNV	TruongNhom
An	Tuấn
An	Tú
Bình	Tuấn
Bình	Tú
Hạnh	Tuấn
Phúc	Minh

Khóa chính của QLY_NHOM: **TruongNhom**
Khóa chính của QLY_NV: {**TenNV, TruongNhom**}
Khóa ngoại của QLY_NV : **TruongNhom** tham chiếu đến khóa chính của QLY_NHOM: **TruongNhom**

- Lược đồ quan hệ QLY_NHOM và QLY_NV đều đạt chuẩn BCNF vì không có thuộc tính khóa mà phụ thuộc hàm vào thuộc tính không khóa.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Các dạng chuẩn 4NF, 5NF và 6NF

- Dạng chuẩn 4NF (Four Normal Form):
  - Không có cá thể bản CSDL nào chứa 2 hoặc nhiều dữ liệu độc lập và đa trị mô tả thực thể có liên quan.
- Dạng chuẩn 5NF (Five Normal Form):
  - Quan hệ đạt chuẩn 4NF và không thể được phân tách thành bất kỳ quan hệ nhỏ hơn mà không mất mát dữ liệu.
- Dạng chuẩn 6NF (Six Normal Form):
  - Không được chuẩn hóa và đang phát triển.

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Chuẩn hóa dữ liệu

- Quy trình chuẩn hóa:

Dạng chuẩn 1
(loại bỏ thuộc tính phức hợp, đa trị, lặp, có thể tính toán được)

Dạng chuẩn 2
(loại bỏ các PTH riêng phần)

Dạng chuẩn 3
(loại bỏ các PTH bắc cầu)

Các dạng chuẩn khác (BC, …)

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 1:
  - Nguyên tắc chung: loại bỏ thuộc tính đa trị, phức hợp hoặc lặp.
    - Thuộc tính đa trị: tách thành nhiều dòng.
    - Thuộc tính phức hợp: tách thành nhiều thuộc tính khác nhau.
    - Thuộc tính lặp: tách ra thành một bảng mới.

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 1:
  - VD12:

MASV	HOTEN	DIACHI	MAMON	TENMON	DIEM
A01	Lê Na	12 Thái Hà	M01M02	CSDLAnh	89
A02	Trần An	56 Mã Mây	M01	CSDL	8
A03	Hà Nam	24 Cầu Gỗ	M01M02M03	CSDLAnhToán 1	689

  - Thuộc tính đa trị → tách thành nhiều dòng:

MASV	HOTEN	DIACHI	MAMON	TENMON	TENGV	PHONG	DIEM
A01	Lê Na	12 Thái Hà	M01	CSDL	Mai	P401	8
A01	Lê Na	12 Thái Hà	M02	Anh	Hương	P405	9
A02	Trần An	56 Mã Mây	M01	CSDL	Mai	P401	8
A03	Hà Nam	24 Cầu Gỗ	M01	CSDL	Mai	P401	6
A03	Hà Nam	24 Cầu Gỗ	M02	Anh	Hương	P405	8
A03	Hà Nam	24 Cầu Gỗ	M03	Toán 1	Hoa	P406	9

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 1:
  - VD13:

sname	city	product	
		name	price
Blake	London	Nut	100
		Bolt	120
Smith	Paris	Screw	75

  - Thuộc tính phức hợp → tách thành nhiều thuộc tính:

sname	city	item	price
Blake	London	Nut	100
Blake	London	Bolt	120
Smith	Paris	Screw	75

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 1:
  - VD14:

MsDV	TenDV	MsNQL	DiaDiem
1	Nghiên cứu	002	Nam Định
2	Hành chính	014	Hà Nội
3	Lãnh đạo	061	Hà Nội
4	Nhân sự	134	Hà Nội

  - Thuộc tính lặp → tách ra một bảng mới:

MsDV	TenDV	MsNQL	MsDiaDiem
1	Nghiên cứu	002	01
2	Hành chính	014	02
3	Lãnh đạo	061	02
4	Nhân sự	134	02

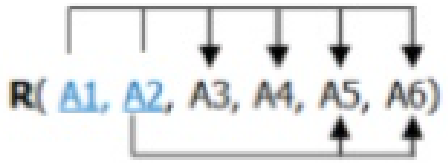
MsDiaDiem	TenDiaDiem
01	Nam Định
02	Hà Nội

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 2:
  - Loại bỏ các thuộc tính không khóa phụ thuộc vào một bộ phận của khóa chính (P) -> tách thành một quan hệ mới, với khóa chính P.
  - Các thuộc tính còn lại lập thành một quan hệ, với khóa chính là khóa chính ban đầu.

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 2:



  - Khóa chính: A1, A2
  - Phụ thuộc hàm bộ phận A2→ {A5, A6}
  - Tách thành hai quan hệ:

+ R1(A2, A5, A6)

+ R2(A1, A2, A3, A4)

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 2:
  - VD15: Cho quan hệ R (MsKH, TenKH, TP, MsMH, TenMH, DG, SL)

MsKH	TenKH	TP	MsMH	TenMH	DG	SL
S1	An	TPHCM	P1	Táo	70	300
S1	An	TPHCM	P2	Cam	50	200
S1	An	TPHCM	P3	Dâu	120	400
S2	Hòa	HN	P1	Táo	70	120
S2	Hòa	HN	P3	Dâu	120	200
S3	Thanh	NT	P2	Cam	50	100
S4	Trang	NT	P2	Cam	50	50

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 2:
  - Cho quan hệ R (MsKH, TenKH, TP, MsMH, TenMH, DG, SL)
  - Tập phụ thuộc hàm:

    F={ MsKH → {TenKH, TP};

    MsMH → {TenMH, DG};

    {MsKH, MsMH} → SL}

  - Tách thành hai quan hệ:

    + R1(MsKH, TenKH, TP)

    + R2(MsKH, MsMH, TenMH, DG, SL)

  - Tách R2 thành hai quan hệ:

    + R21(MsMH, TenMH, DG)

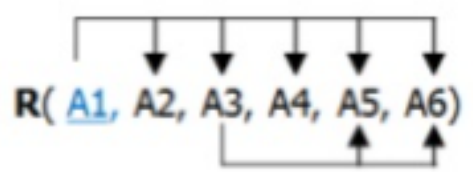
    + R22(MsKH, MsMH, SL)

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:
  - Loại bỏ các thuộc tính phụ thuộc bắc cầu/gián tiếp ra khỏi quan hệ -> tách thành một quan hệ riêng; khóa chính của bảng mới là thuộc tính bắc cầu.
  - Các thuộc tính còn lại lập thành một quan hệ, khóa chính của nó là khóa chính ban đầu.

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:



  - Khóa chính: A1
  - Phụ thuộc hàm bắc cầu: A3→ {A5, A6}
  - Tách thành hai quan hệ:

+ R1(A3, A5, A6)

+ R2(A1, A2, A3, A4)

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:
  - VD16: Cho quan hệ R (MsKH, TenKH, TP, MsMH, TenMH, DG, SL)

MsKH	TenKH	TP
S1	An	TPHCM
S2	Hòa	HN
S3	Thanh	NT
S4	Trang	NT

MsMH	TenMH	DG
P1	Táo	70
P2	Cam	50
P3	Dâu	120

MsKH	MsMH	SL
S1	P1	300
S1	P2	200
S1	P3	400
S2	P1	120
S2	P3	200
S3	P2	100
S4	P2	50

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:
  - VD17: Cho quan hệ R (R (MsKH, TenKH, TP, PVC)
  - F={ MsKH → TenKH, TP; TP → PVC}

MsKH	TenKH	TP	PVC
S1	An	TPHCM	01
S2	Hòa	HN	02
S3	Thanh	NT	03
S4	Trang	NT	03

  - Tách thành hai quan hệ:

+ R1(MsKH, TenKH, TP)

+ R2(TP, PVC)

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:
  - VD17: Cho quan hệ R (R (MsKH, TenKH, TP, PVC)
  - F={ MsKH → TenKH, TP; TP → PVC}

MsKH	TenKH	TP	PVC
S1	An	TPHCM	01
S2	Hòa	HN	02
S3	Thanh	NT	03
S4	Trang	NT	03

Chuẩn hóa dữ liệu

- Đưa về dạng chuẩn 3:

MsKH	TenKH	TP	PVC
S1	An	TPHCM	01
S2	Hòa	HN	02
S3	Thanh	NT	03
S4	Trang	NT	03

MsKH	TenKH	TP
S1	An	TPHCM
S2	Hòa	HN
S3	Thanh	NT
S4	Trang	NT

TP	PVC
TPHCM	01
HN	02
NT	03

Nội dung

- Đặt vấn đề
- Chuẩn hóa dữ liệu là gì?
- Các dạng chuẩn hóa:
  - Dạng chuẩn 1
  - Dạng chuẩn 2
  - Dạng chuẩn 3
  - Dạng chuẩn Boyce – Codd
  - Các dạng chuẩn 4NF, 5NF và 6NF
- Chuẩn hóa dữ liệu
- Bài tập ứng dụng

Bài tập ứng dụng

1. Cho quan hệ **NhanVien_DuAn** (MsNV, MsDA, SoGio, TenNV, TenDA, DiaDiemDA)

   F={ MsNV, MsDA → SoGio;

   MsNV → TenNV;

   MsDA → TenDA, DiaDiemDA}

   Hãy chuẩn hóa quan hệ **NhanVien_DuAn** trên.

Gợi ý:

- Quan hệ **NhanVien_DuAn** không đạt chuẩn 2 vì có PTH riêng phần: *MsNV → TenNV*

  => Tách thành 2 quan hệ mới:

  NhanVien(MsNV, TenNV) => đạt chuẩn 3

  PhanCong_DuAn(MsNV, MsDA, SoGio, TenDA, DiaDiemDA)

- Quan hệ PhanCong_DuAn không đạt chuẩn 2 vì có PTH riêng phần *MsDA → TenDA, DiaDiemDA*

  => Tách thành 2 quan hệ mới:

  DuAn(MsDA, TenDA, DiaDiemDA) => đạt chuẩn 3

  PhanCong(MsNV, MsDA, SoGio) => đạt chuẩn 3

Bài tập ứng dụng

2. Cho quan hệ **NhanVien_DonVi** (MsNV, TenNV, NgaySinh, DiaChi, MsDonVi, TenDonVi, MsNQL)

   F={MsNV → TenNV, NgaySinh, DiaChi, MsDonVi;

   MsDonVi → TenDonVi, MsNQL}

   Hãy chuẩn hóa quan hệ **NhanVien_DonVi** trên.

Gợi ý:

- Quan hệ **NhanVien_DonVi** không đạt chuẩn 3 vì có PTH bắc cầu/gián tiếp: MsNV → *MsDonVi; MsDonVi → TenDonVi*

  => Tách thành 2 quan hệ mới:

  - **DonVi(MsDonVi, TenDonVi, MsNQL) => đạt chuẩn 3**
  - **NhanVien(MsNV, TenNV, NgaySinh, DiaChi, MsDonVi) => đạt chuẩn 3**

Bài tập ứng dụng

3. Cho quan hệ R(ABCDEFGHIJ)

   F={ A, B → C, D, E, F, G, H, I, J;

   A → E, F, G, H, I, J;

   F → I, J }

   Hãy chuẩn hóa quan hệ R trên.

Gợi ý:

- Quan hệ **R** không đạt chuẩn 2 vì có PTH riêng phần A → E, F, G, H, I, J
- => Tách thành 2 quan hệ mới:

  R1(A, E, F, G, H, I, J)

  R2(A, B, C, D) => đạt chuẩn 3
- QH **R1** không đạt chuẩn 3 vì có PTH bắc cầu: F → I, J
- => Tách thành 2 quan hệ mới:

  R11(F, I, J) => đạt chuẩn 3

  R12(A, E, F, G, H) => đạt chuẩn 3

Bài tập ứng dụng

4. **Cho quan hệ R(XYZTV)**

   F={ XY → Z; Y → T; Z → V}

   a. **Tìm khóa của R.**

   b. **Hãy chuẩn hóa quan hệ R trên.**

   c. **CMR: cấu trúc CSDL sau khi chuẩn hóa vẫn bảo toàn thông tin.**

Gợi ý:

a. Tìm khóa của R:

$N = U - \bigcup_{\forall f \in F} right(f)$ = {XYZTV} - {ZTV} = {XY}

$N_F^+ = \{XY\}_F^+$ = {XYZVT} = U => Vậy R chỉ có 1 khóa là {XY}

b. Hãy chuẩn hóa quan hệ R trên:

R không đạt chuẩn 2 vì có PTH riêng phần *Y → T*

=> Tách thành 2 quan hệ mới:

R1(YT) => đạt chuẩn 3

R2(XYZV)

R2 không đạt chuẩn 3 vì có PTH bắc cầu/gián tiếp *XY → Z; Z → V*

=> Tách thành 2 quan hệ mới:

R21(ZV) => đạt chuẩn 3

R22(XYZ) => đạt chuẩn 3

Bài tập ứng dụng

4. **Cho quan hệ R(XYZTV)**

   F={ XY → Z; Y → T; Z → V}

- b. Hãy chuẩn hóa quan hệ R trên:

Gợi ý:

Vậy R được chuẩn hóa thành 3 quan hệ sau:

- R1(YT)
- R2(ZV)
- R3(XYZ)

Thuật toán kiểm tra phép phân rã không mất mát thông tin

Thuật toán kiểm tra phép phân rã không mất mát thông tin

Input

Lược đồ quan hệ R={A1, A2, . . . , An}
Tập các phụ thuộc hàm F
Phép tách ρ(R1, R2, . . . , Rk)

OutputKết luận phép tách ρ không mất mát thông tin.

Các bước của thuật toán

Bước 1

Thiết lập một bảng với n cột (tương ứng với n thuộc tính) và k dòng (tương ứng với k quan hệ), trong đó cột thứ j ứng với thuộc tính Aj, dòng thứ i ứng với lược đồ Ri.
Tại dòng i và cột j, ta điền ký hiệu aj nếu thuộc tinh Aj∈Ri.Ngược lại ta điền ký hiệu bij.

Bước 2

Xét các phụ thuộc hàm trong F và áp dụng cho bảng trên.
Giả sử ta có phụ thuộc hàm X→Y∈F, xét các dòng có giá trị bằng nhau trên thuộc tính X **thì làm bằng**các giá trị của chúng trên Y. Ngược lại làm bằng chúng bằng ký hiệu bij. Tiếp tục áp dụng các pth cho bảng (kể cả việc lặp lại các phụ thuộc hàm đã áp dụng) cho tới khi không còn áp dụng được nữa.

Bước 3

Xem xét bảng kết quả. Nếu xuất hiện một dòng chứa toàn giá trị **a1, a2 ,…,an** thì kết luận phép tách ρ không mất mát thông tin.

Bài tập ứng dụng

4. **Cho quan hệ R(XYZTV)**

F={ XY→ Z; Y → T; Z → V}

c. CMR: cấu trúc CSDL sau khi chuẩn hóa vẫn bảo toàn thông tin.

Gợi ý:

Ma trận khởi tạo:

	1	2	3	4	5
	X	**Y**	**Z**	**T**	**V**
R1(YT)	b11	a2	b13	a4	b15
R2(ZV)	b21	b22	a3	b24	a5
R3(XYZ)	a1	a2	a3	b34	b35

Ma trận sau khi xét:

	1	2	3	4	5
	X	**Y**	**Z**	**T**	**V**
R1(YT)	b11	a2	b13	a4	b15
R2(ZV)	b21	b22	a3	b24	a5
R3(XYZ)	a1	a2	a3	a4	a5

Vậy cấu trúc dữ liệu sau khi chuẩn hóa ở câu b vẫn bảo toàn thông tin.

Bài tập ứng dụng

5. Cho quan hệ R và tập PTH:

   F={ customer_id (A) → customer_name (B), phone (C), address (D)
   product_id (E) → description (F), unit_price (G)
   order_id (H) → customer_id (A), order_date (I)
   order_id (H), product_id (E) → quantity (J) }

   a. Tìm khóa của R.
   b. Hãy chuẩn hóa quan hệ R trên.
   c. CMR: cấu trúc CSDL sau khi chuẩn hóa vẫn bảo toàn thông tin.

Bài tập ứng dụng

5. Cho quan hệ R chưa được chuẩn hóa dưới đây:

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

$E \rightarrow \{F, G\};$

$H \rightarrow \{A, I\};$

$\{H, E\} \rightarrow J \}$

Gợi ý:

a. Tìm khóa của R:

$N = U - \bigcup_{\forall f \in F} right(f)$ = {ABCDEFGHIJ} - {ABCDFGIJ} = {EH}

$N_F^+ = \{EH\}_F^+$ = {EHJFGAIBCD} = U

=> {EH} là khóa duy nhất của R.

Bài tập ứng dụng

5. Cho quan hệ R chưa được chuẩn hóa dưới đây:

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

$E \rightarrow \{F, G\};$

$H \rightarrow \{A, I\};$

$\{H, E\} \rightarrow J \}$

Gợi ý:

b. Chuẩn hóa quan hệ **R(EHABCDFGIJ):**

R chưa đạt chuẩn 3 vì có PTH bắc cầu/gián tiếp: $A \rightarrow \{B, C, D\}$

=> Tách thành 2 quan hệ mới:

R1(ABCD) => đạt chuẩn 3

R2(EHAFGIJ)

R2 không đạt chuẩn 2 vì có PTH riêng phần: $E \rightarrow \{F, G\}$

=> Tách thành 2 quan hệ mới:

R21(EFG) => đạt chuẩn 3

R22(EHAIJ)

R22 không đạt chuẩn 2 vì có PTH riêng phần: $H \rightarrow \{A, I\}$

=> Tách thành 2 quan hệ mới:

R221(HAI) => đạt chuẩn 3

R222(EHJ) => đạt chuẩn 3

Vậy R được chuẩn hóa thành 4 quan hệ sau:

R1(ABCD)
R2(EFG)
R3(HAI)
R4(EHJ)

Bài tập ứng dụng

5. Cho quan hệ R chưa được chuẩn hóa dưới đây:

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

$E \rightarrow \{F, G\};$

$H \rightarrow \{A, I\};$

$\{H, E\} \rightarrow J \}$

Gợi ý:

- c. CMR: cấu trúc CSDL sau khi chuẩn hóa vẫn bảo toàn thông tin.

Ma trận khởi tạo:

	1	2	3	4	5	6	7	8	9	10
	E	**H**	**A**	**B**	**C**	**D**	**F**	**G**	**I**	**J**
R1(ABCD)	b11	b12	a3	a4	a5	a6	b17	b18	b19	b110
R2(EFG)	a1	b22	b23	b24	b25	b26	a7	a8	b29	b210
R3(HAI)	b31	a2	a3	b34	b35	b36	b37	b38	a9	b39
R4(EHJ)	a1	a2	b43	b44	b45	b46	b47	b48	b49	a10

Ma trận sau khi xét:

	1	2	3	4	5	6	7	8	9	10
	E	**H**	**A**	**B**	**C**	**D**	**F**	**G**	**I**	**J**
R1(ABCD)	b11	b12	a3	a4	a5	a6	b17	b18	b19	b110
R2(EFG)	a1	b22	b23	b24	b25	b26	a7	a8	b29	b210
R3(HAI)	b31	a2	a3	a4	a5	a6	b37	b38	a9	b39
R4(EHJ)	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10

Vậy cấu trúc dữ liệu sau khi chuẩn hóa ở câu b vẫn bảo toàn thông tin.