Chuẩn hóa dữ liệu (Normalization)

Các dạng chuẩn (1NF -> BCNF)

Thuật toán phân rã thành 3NF bảo toàn phụ thuộc

Thuật toán phân rã thành BCNF có nối không mất

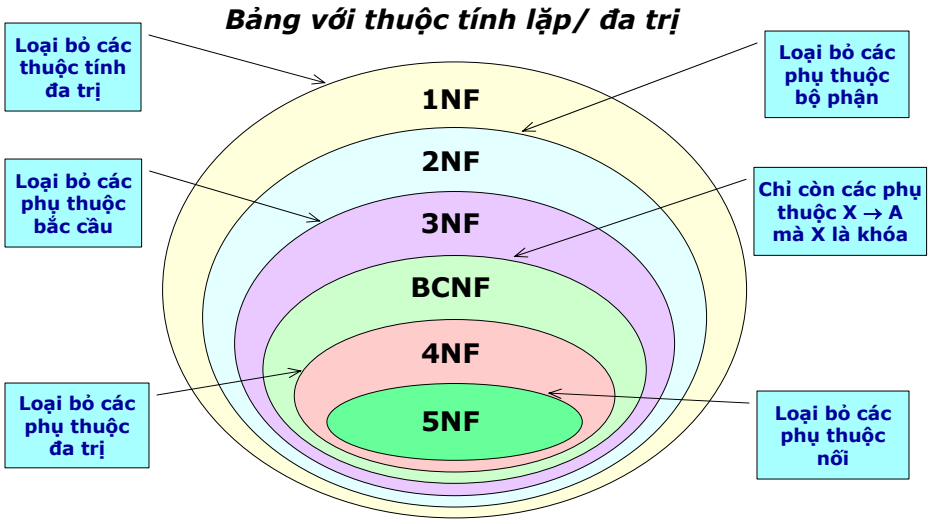
Chuẩn hóa (1)

- Chuẩn hóa là quá trình tách bảng (phân rã) thành các bảng nhỏ hơn dựa vào các phụ thuộc hàm mà không làm mất thông tin.
- Các dạng chuẩn là các chỉ dẫn để thiết kế các bảng trong CSDL.
- Mục đích của chuẩn hóa là loại bỏ các dư thừa vì
  - Yêu cầu về không gian lưu trữ
  - Tránh các lỗi khi thao tác dữ liệu (Insert, Delete, Update).
- Chú ý: chuẩn hóa làm tăng thời gian truy vấn.

Chuẩn hoá (2)

- Các dạng chuẩn sử dụng phụ thuộc hàm
- Chuẩn 3 (3NF)
  - Luôn có được phân rã có nối không mất và bảo toàn phụ thuộc
  - Cho phép dư thừa 1 chút
- Chuẩn Boyce Codd (BCNF)
  - Tránh dư thừa
  - Luôn có phân rã có nối không mất
  - Không phải luôn có được phân rã bảo toàn phụ thuộc

Các dạng chuẩn
(Normal forms)



Bảng R sau có phải là 1 quan hệ ?

MF	Tenfim	NSX	Giathue	HSX	NPP	MaKH	TenKH	Điachi	Ngaydat
P001	Hồn yêu	1996	3.2	C 20	Fafim	K001	David	Paris	1/1/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K001	David	Paris	1/2/06
						K002	Marie	Toulouse	5/1/06
						K003	John	Nice	1/2/06
P003	Hoa hậu FBI	2002	3.2	Wanner	Films	K003	John	Nice	1/2/06
P004	Taxi	2005	4.5	France	Fafim	K001	David	Paris	2/5/06
						K004	Helen	Bordeaux	1/2/06
P005	Đời cát	2003	2.5	VFC	VFC	K005	Taix	Toulouse	4/1/06

Còn bảng S sau là 1 quan hệ ?

NGUOIQL	CAPDUOI1	CAPDUOI2	CAPDUOI3	CAPDUOI4
Bob	Jim	Mary	Beth	
Mary	Mike	Jason	Carol	Mark
Jim	Alan			

Cột lặp

- Jim chỉ có một cấp dưới ?
- Nếu Mary thêm một người cấp dưới ?

→ Gộp các cột lặp thành một cột

- **Vấn đề gì xảy ra ???**

Dạng chuẩn 1 - 1NF

- Chuẩn 1 đảm bảo bảng đang xét là một quan hệ:
  - Giá trị của các cột là đơn trị.
  - Không chứa các cột lặp
- Cách chuyển sang dạng 1NF:
  - điền đủ dữ liệu vào các cột khác
  - Biến cột đa trị thành các dòng

MF	Tenfim	NSX	Giathue	HSX	NPP	MaKH	TenKH	Điachi	Ngaydat
P001	Hồn yêu	1996	3.2	C 20	Fafim	K001	David	Paris	1/1/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K001	David	Paris	1/2/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K002	Marie	Toulouse	5/1/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K003	John	Nice	1/2/06
P003	Hoa hậu FBI	2002	3.2	Wanner	Films	K003	John	Nice	1/2/06
P004	Taxi	2005	4.5	France	Fafim	K001	David	Paris	2/5/06
P004	Taxi	2005	4.5	France	Fafim	K004	Helen	Bordeaux	1/2/06
P005	Đời cát	2003	2.5	VFC	VFC	K005	Taix	Toulouse	4/1/06

Dạng chuẩn 2 - 2NF (1)

- Một quan hệ đạt 2NF nếu
  - Quan hệ đã đạt 1NF
  - Thuộc tính khác khóa phụ thuộc hoàn toàn vào khóa
- Một quan hệ đạt 2NF nếu thỏa mãn 1 trong các điều kiện sau
  - Khóa chính chỉ gồm 1 thuộc tính.
  - Bảng không có các thuộc tính không khóa.
  - Tất cả thuộc tính không khóa phụ thuộc hoàn toàn vào tập thuộc tính khóa chính.
- Gợi ý
  - Chỉ kiểm tra các quan hệ có đạt 2NF nếu quan hệ đó có khoá chính gồm 2 thuộc tính trở lên.
  - Để chuyển quan hệ từ dạng 1NF sang dạng 2NF, ta dùng phép chiếu.

Dạng chuẩn 2 - 2NF (2)

- Bảng R có các phụ thuộc hàm sau:
  - ① MF → Tenfim, Giathue, NSX, HSX, N.ngu, NPP
  - ② MaKH → TenKH, Diachi
  - ③ MF, MaKH → Ngaythue
  - ④ HSX → NPP
- Khóa chính: MF, MaKH.
- Các thuộc tính Tenfim, Giathue, TenKH, Diachi,...
  - là các thuộc tính không khóa
  - chỉ phụ thuộc vào một bộ phận của khóa

  → R không đạt chuẩn 2

MF	Tenfim	NSX	Giathue	HSX	NPP	MaKH	TenKH	Điachi	Ngaydat
P001	Hồn yêu	1996	3.2	C 20	Fafim	K001	David	Paris	1/1/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K001	David	Paris	1/2/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K002	Marie	Toulouse	5/1/06
P002	Đảo vô hình	2006	5.0	C 20	Fafim	K003	John	Nice	1/2/06
P003	Hoa hậu FBI	2002	3.2	Wanner	Films	K003	John	Nice	1/2/06
P004	Taxi	2005	4.5	France	Fafim	K001	David	Paris	2/5/06
P004	Taxi	2005	4.5	France	Fafim	K004	Helen	Bordeaux	1/2/06
P005	Đời cát	2003	2.5	VFC	VFC	K005	Taix	Toulouse	4/1/06

R_1 R_2 R_3

$\pi_{MaKH,TenKH,Diachi}(R)$

MaKH	TenKH	Diachi
K001	David	Paris
K002	Marie	Toulouse
K003	John	Nice
K004	Helen	Bordeaux
K005	Taix	Toulouse

$\pi_{MF,Tenfim,NSX,Giathue,HSX,NPP}(R)$

MF	Tenfim	NSX	Gia thue	HSX	NPP
P001	Hồn yêu	1996	3.2	C 20	Fafim
P002	Đảo vô hình	2006	5.0	C 20	Fafim
P003	Hoa hậu FBI	2002	3.2	Wanner	Films
P004	Taxi	2005	4.5	France	Fafim
P005	Đời cát	2003	2.5	VFC	VFC

$\pi_{MF,MaKH,Ngaydat}(R)$

MF	MaKH	Ngay dat
P001	K001	1/1/06
P002	K001	1/2/06
P002	K002	5/1/06
P002	K003	1/2/06
P003	K003	1/2/06
P004	K001	2/5/06
P004	K004	1/2/06
P005	K005	4/1/06

Các lỗi khi ở dạng 2NF

- Lược đồ R_2(MF, Tenfim, NSX, Giathue, HSX, NPP) gồm các phụ thuộc hàm:
  - MF $\rightarrow$ Tenfim, NSX, Giathue, HSX, NPP
  - HSX $\rightarrow$ NPP
- R2 đạt 2NF.
- Vấn đề
  - Có thể thêm nhà phân phối cho 1 hãng sản xuất nếu hãng đó chưa sản xuất bộ phim nào?
  - Nếu xoá thông tin 1 bộ phim có làm mất thông tin về hãng sản xuất và nhà phân phối không?
  - Khi thay đổi nhà phân phối cho 1 hãng sản xuất thì phải thực hiện thao tác đó bao nhiêu lần?

Dạng chuẩn 3 – 3NF

- Một quan hệ ở 3NF nếu
  - Quan hệ đã đạt 2NF
  - Không có các phụ thuộc bắc cầu.
- **Một lược đồ quan hệ R ở 3NF nếu với mọi phụ thuộc X $\rightarrow$ A đúng trong R và A $\notin$ X thì hoặc X là khóa bao hàm hoặc A là nguyên tố.**
- A là **thuộc tính nguyên tố (thuộc tính khóa)** nếu A là thuộc một khóa dự tuyển của R, ngược lại A là phi nguyên tố.
- VD: Cho lược đồ CSZ, F = {CS $\rightarrow$ Z, Z $\rightarrow$ C}, CS và SZ đều là khóa.

  $\Rightarrow$ tất cả các thuộc tính là nguyên tố.

MF → Tenfim, NSX, Giathue, HSX, NPP
HSX → NPP

R_1

MaKH	TenKH	Diachi
K001	David	Paris
K002	Marie	Toulouse
K003	John	Nice
K004	Helen	Bordeaux
K005	Taix	Toulouse

R_2

MF	Tenfim	NSX	Gia thue	HSX	NPP
P001	Hồn yêu	1996	3.2	C 20	Fafim
P002	Đảo vô hình	2006	5.0	C 20	Fafim
P003	Hoa hậu FBI	2002	3.2	Wanner	Films
P004	Taxi	2005	4.5	France	Fafim
P005	Đời cát	2003	2.5	VFC	VFC

R_3

MF	MaKH	Ngay dat
P001	K001	1/1/06
P002	K001	1/2/06
P002	K002	5/1/06
P002	K003	1/2/06
P003	K003	1/2/06
P004	K001	2/5/06
P004	K004	1/2/06
P005	K005	4/1/06

R_1

MaKH	TenKH	Diachi
K001	David	Paris
K002	Marie	Toulouse
K003	John	Nice
K004	Helen	Bordeaux
K005	Taix	Toulouse

R_2

MF	Tenfim	...	HSX
P001	Hồn yêu		C 20
P002	Đảo vô hình		C 20
P003	Hoa hậu FBI		Wanner
P004	Taxi		France
P005	Đời cát		VFC

R_4

HSX	NPP
C 20	Fafim
Wanner	Films
France	Fafim
VFC	VFC

R_3

MF	MaKH	Ngay dat
P001	K001	1/1/06
P002	K001	1/2/06
P002	K002	5/1/06
P002	K003	1/2/06
P003	K003	1/2/06
P004	K001	2/5/06
P004	K004	1/2/06
P005	K005	4/1/06

Các lỗi khi ở dạng 3NF (1)

- Ví dụ:
  - R(MF, Tenfim, NamSX, MaBS, Tinhtrang, HSX) và các mô tả như sau:
    - Mỗi fim có thể có nhiều bản sao
    - Mỗi bản sao của một phim có tình trạng khác nhau
    - Một phim chỉ có một tên, năm sản xuất và HSX.
- Từ thông tin trên ta có các phụ thuộc hàm sau:
  - MF → Tenfim, NamSX, HSX
  - MF, MaBS → Tinhtrang
- R có khoá (MF, MaBS)

Các lỗi khi ở dạng 3NF (2)

MF	Tenfim	NSX	HSX	MBS	Tinhtrang
P001	Hồn yêu	1996	C 20	1	Mượn
P001	Hồn yêu	1996	C 20	2	Hỏng
P001	Hồn yêu	1996	C 20	3	Mượn
P002	Đảo vô hình	2006	C 20	1	Bán
P002	Đảo vô hình	2006	C 20	2	Rỗi
P003	Hoa hậu FBI	2002	Wanner	1	Rỗi

- R trên đạt 3NF.
- Lỗi phát sinh
  - Khi xoá bản sao 1 của phim P003 thì làm mất thông tin của phim này
  - Không thể thêm thông tin về 1 phim nếu phim đó chưa có bản sao nào

Dạng chuẩn Boyce-Codd - BCNF (1)

- **Một lược đồ quan hệ R ở dạng BCNF nếu với mọi phụ thuộc $X \rightarrow A$ đúng trong R và $A \notin X$ thì X là khóa bao hàm.**
- Một quan hệ ở BCNF thì cũng đạt 3NF.
- Trong thực hành các quan hệ đạt chuẩn 3NF là đủ. Tuy nhiên một quan hệ ở 3NF **không đảm bảo đã loại bỏ được tất cả các lỗi khi thao tác dữ liệu**.

Dạng chuẩn BCNF (2)

MF	Tenfim	NSX	HSX	MBS	Tinhtrang
P001	Hồn yêu	1996	C 20	1	Mượn
P001	Hồn yêu	1996	C 20	2	Hỏng
P001	Hồn yêu	1996	C 20	3	Mượn
P002	Đảo vô hình	2006	C 20	1	Bán
P002	Đảo vô hình	2006	C 20	2	Rỗi
P003	Hoa hậu FBI	2002	Wanner	1	Rỗi

$\pi_{MF,Tenfim,NSX,HSX}(R)$

MF	Tenfim	NSX	HSX
P001	Hồn yêu	1996	C 20
P002	Đảo vô hình	2006	C 20
P003	Hoa hậu FBI	2002	Wanner

$\pi_{MF,MaBS,Tinhtrang}(R)$

MF	MBS	Tinhtrang
P001	1	Mượn
P001	2	Hỏng
P001	3	Mượn
P002	1	Bán
P002	2	Rỗi
P003	1	Rỗi

Thuật toán phân rã thành 3NF (1)

- Dùng để phân rã R thành (R_1, R_2,...,R_n) mà
  - R_i đạt 3NF
  - Phân rã bảo toàn phụ thuộc, và
  - Phân rã có nối không mất
- Thuật toán
  - IN: Lược đồ quan hệ R, tập phụ thuộc hàm F
  - OUT: Một phân rã bảo toàn phụ thuộc của R sao cho mỗi lược đồ quan hệ đều có dạng 3NF và ứng với hình chiếu của F trên lược đồ đó.

Thuật toán phân rã thành 3NF (2)

- Thuật toán phân rã thành 3NF bảo toàn phụ thuộc

  Tính F_C

  m = 0

  IF ($\alpha \rightarrow Y \in F_c$) AND ($\alpha \rightarrow Z \in F_c$)

  { $F_C = F_C \setminus \{\alpha \rightarrow Y, \alpha \rightarrow Z\} \cup \{\alpha \rightarrow YZ\}$

  } /* Hợp các phụ thuộc có cùng vế trái*/

  FOR ($X \rightarrow Y \in F_c$) DO

  { m = m+1

  R_m = XY

  }

Thuật toán phân rã thành 3NF (3)

- **Ví dụ 1**
  - R = ABCD, F = {AB→C, C→D, BC→D, CD→B}
  - Phân rã thành 3NF
- Giải quyết
  - Tính F_C = {AB→C, C→D, C→B}
  - F_C = {AB→C, C→BD}
  - Ta có R_1 = ABC, R_2 = BCD
  - Phân rã ρ = (ABC, BCD) bảo toàn phụ thuộc

Thuật toán phân rã thành 3NF (4)

- Thuật toán cải tiến để phân rã có nối không mất

  Tính F_C

  m = 0

  IF ($\alpha \rightarrow Y \in F_c$) AND ($\alpha \rightarrow Z \in F_c$)

  { $F_C = F_C \setminus \{\alpha \rightarrow Y, \alpha \rightarrow Z\} \cup \{\alpha \rightarrow YZ\}$

  } /* Hợp các phụ thuộc có cùng vế trái*/

  FOR ($X \rightarrow Y \in F_c$) DO

  { m = m+1

  R_m = XY }

  IF (!∃ R_i với 1≤ i ≤ m chứa 1 khoá dự tuyển của R)

  {m = m+1

  R_m = 1 khoá dự tuyển bất kỳ của R }

Thuật toán phân rã thành 3NF (5)

- **Ví dụ 2**
  - R=CTHRSG, F={C→T, HR→C, HT→R, CS→G, HS→R}
  - Tìm phân rã đạt 3NF bảo toàn phụ thuộc và có nối không mất.
- Giải quyết
  - Ta có $F_C = F$
  - K = HS
  - ρ = (CT, HRC, HTR, CSG, HSR) đạt 3NF
  - Thêm lược đồ HS vào phân rã ρ; có HSR chứa khóa

  → Phân rã ρ bảo toàn phụ thuộc và có nối không mất

Thuật toán phân rã thành 3NF (6)

- **Ví dụ 3**
  - F={MF→Tenfim,NSX,Giathue,HSX,NPP; HSX→NPP}
  - Tìm phân rã đạt 3NF bảo toàn phụ thuộc và có nối không mất.
- Giải quyết
  - F_C = {MF→Tenfim,NSX,Giathue,HSX; HSX→NPP}
  - K = (MF, HSX)
  - Phân rã gồm R_1 = (MF, Tenfim, NSX, Giathue, HSX) và R_2 = (HSX,NPP) đạt 3NF

Phân rã thành BCNF (1)

- Bổ đề:

1. Mỗi lược đồ có 2 thuộc tính đều có dạng BCNF.
2. Nếu R không có dạng BCNF thì ta có thể tìm được các thuộc tính A và B trong R sao cho (R - AB) → A đúng (có thể (R - AB) → B cũng đúng). Điều ngược lại chưa chắc đã đúng.

Thuật toán phân rã thành BCNF (1)

- Dùng để phân rã R thành (R_1, R_2,...,R_n) mà
  - R_i đạt BCNF
  - Phân rã có nối không mất
- Thuật toán:
  - IN: Lược đồ quan hệ R, tập phụ thuộc hàm F .
  - OUT: Một phân rã của R có nối không mất, sao cho mỗi lược đồ quan hệ trong phân rã có dạng BCNF ứng với hình chiếu của F trên lược đồ đó.

Thuật toán phân rã thành BCNF (2)

- Phương pháp: Phân rã lược đồ R thành 2 lược đồ:
  1. Lược đồ 1: có tập các thuộc tính XA, có dạng BCNF và phụ thuộc $X \rightarrow A$ đúng.
  2. Lược đồ 2: R – A
  - Tiếp tục phân rã lược đồ R - A theo các bước 1, 2 cho đến khi
    - không thể phân rã được nữa
    - lược đồ chỉ còn 2 thuộc tính

Thuật toán phân rã thành BCNF (3)

Chương trình chính:

```
BEGIN
   Z := R;
   REPEAT
      Phân rã Z thành Z-A và XA  /*gọi TT phân rã D*/
      Thêm XA vào phân rã;
      Z := Z – A;
   UNTIL (không thể phân rã Z)
   Thêm Z vào phân rã
END.
```


Thuật toán phân rã thành BCNF (4)

- **Thủ tục phân rã (D)**

```
BEGIN
   IF Z không chứa A, B sao cho A ∈ (Z - AB)+ THEN
      return Z có dạng BCNF và không phân rã được
   ELSE
      BEGIN
        Tìm một cặp A và B;  Y := Z;
        WHILE (Y chứa A và B sao cho (Y - AB) → A) DO
          Y:=Y – B;
        return phân rã Z-A và Y /*Y là XA trong CT chính*/
      END;
END;
```

Thuật toán phân rã thành BCNF (5)

- Cho lược đồ quan hệ CTHRSG và tập phụ thuộc F = (C→T, HR→C, HT→R, CS→G, HS→R)
- Áp dụng thuật toán:
  - Y=CTHRSG, A=C, B=T, C ∈ $(HRSG)^+$ ⇒ Y=CHRSG
  - Y=CHRSG, A=R, B=C, R ∈ $(HSG)^+$ ⇒ Y=HRSG
  - Y=HRSG, A=R, B=G, R ∈ $(HS)^+$ ⇒ Y=HRS
  - Không phân rã được nữa, **HRS** là một lược đồ trong phân rã.
  - Z=CTHRSG – R=CTHSG
  - Tiếp tục với Y=CTHSG

Thuật toán phân rã thành BCNF (6)

  - Y=CTHSG, A=T, B=H, T ∈$(CSG)^+$ ⇒ Y=CTSG
  - Y=CTSG, A=T, B=S, T ∈$(CG)^+$ ⇒ Y=CTG
  - Y=CTG, A=T, B=G, T ∈$(C)^+$ ⇒ Y=CT
  - CT không phân rã được nữa, **CT** là một lược đồ trong phân rã.
  - Z = CHSG
  - Tiếp tục với Y = CHSG
  - …
- Cuối cùng được phân rã (HRS, CT, CSG, CHS)
- **Chú ý:** Thứ tự chọn cặp A, B khác có thể thu được một phân rã khác, ρ = (CHS, THS, HSG, HSR).

Một vài chú ý về BCNF

- Với cùng lược đồ, có thể có nhiều phân rã khác nhau đạt BCNF
- Thuật toán chỉ sinh ra 1 trong các phân rã đó
- Phân rã BCNF có thể bảo toàn phụ thuộc
- Thứ tự chọn phụ thuộc hàm trong thuật toán có thể sinh ra phân rã khác nhau