



HỌC VIỆN NÔNG NGHIỆP VIỆT NAM
VIETNAM NATIONAL UNIVERSITY OF AGRICULTURE

BÀI GIẢNG CƠ SỞ DỮ LIỆU 2

Chương 2: Truy vấn trong CSDL PHÂN TÁN

Hà Nội, 1 - 2018

HVN Học viện
Nông nghiệp Việt Nam

CHƯƠNG 2

Truy vấn trong CSDL phân tán

Nội dung chương gồm 3 phần:

- 2.1. Định nghĩa dữ liệu trong CSDL phân tán
- 2.2. Phân đoạn dữ liệu trong CSDL phân tán.
- 2.3. Định giá các câu hỏi phân tán.

2.1. Định nghĩa dữ liệu trong CSDL phân tán

Việc định nghĩa dữ liệu trong CSDL phân tán bao gồm hai vấn đề quan trọng, đó là :

- Sự mô tả dữ liệu
- Sự phân đoạn và sắp đặt dữ liệu trong mạng máy tính của CSDL phân tán.

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

2.1.1. Mô tả dữ liệu trong CSDL Phân tán

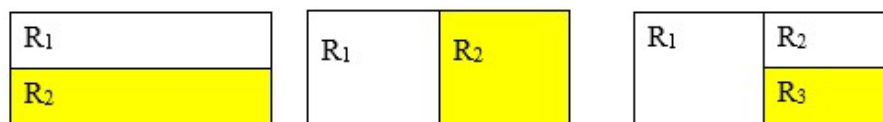
- Thao tác mô tả dữ liệu tổng thể, chẳng hạn thêm một quan hệ mới, hay sửa đổi mô tả đó phụ thuộc chủ yếu vào việc tổ chức từ điển dữ liệu.
- Với 1 từ điển DL được tổ chức tập trung, thao tác này rất đơn giản vì nó chỉ liên quan đến một trạm trung tâm lưu trữ từ điển DL.
- Với 1 từ điển DL được lưu trữ có nhân bản, thao tác mô tả phải được làm đồng bộ với tất cả các trạm có lưu các bản sao của từ điển DL.
- Với 1 từ điển được lưu trữ phân tán thì thao tác mô tả là phức tạp nhất: mô tả dữ liệu thoát đầu được đưa vào trạm cục bộ mà tại đó dữ liệu được tạo ra, trạm này gọi là trạm “tạo sinh”, sau đó trạm “tạo sinh” và các trạm khác được phép truy cập dữ liệu mới sẽ hợp tác và trao đổi thông tin về mô tả dữ liệu

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

2.1.2. Các kiểu phân đoạn dữ liệu trong CSDL Phân tán

- Sự sắp đặt dữ liệu bằng cách chia một quan hệ tổng thể thành các đơn vị logic trong lược đồ khái niệm tổng thể gọi là sự phân đoạn dữ liệu.
- Sự phân đoạn nhằm chia các quan hệ thành các đơn vị nhỏ hơn (gọi là các đoạn) để có thể sắp đặt tối ưu trong CSDL phân tán.
- Có ba kiểu phân đoạn: *phân đoạn ngang*, *phân đoạn dọc* và *phân đoạn hỗn hợp* nhằm chia các quan hệ tổng thể thành các quan hệ con và được sắp đặt trên các trạm khác nhau.

Hình 2.1. các kiểu phân đoạn của quan hệ tổng thể R



Bài giảng Cơ sở dữ liệu 2 | nvdingh@vnua.edu.vn

5

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

Các yêu cầu của sự phân đoạn dữ liệu

- Mỗi phép phân đoạn là một phân tách không mất thông tin: tức là mỗi phần tử dữ liệu của quan hệ tổng thể phải thuộc một hay nhiều đoạn của phân tách, dữ liệu không mất đi nhưng cũng không thêm dữ liệu mới.
- Khôi phục lại được quan hệ ban đầu từ các đoạn.
- Các đoạn là không trùng lặp: nhằm tránh dư thừa dữ liệu (đối với các phân đoạn ngang điều này là bắt buộc, đối với các phân đoạn dọc thì mỗi đoạn chỉ cho phép trùng lặp các thuộc tính khóa để khôi phục được quan hệ ban đầu).
- Thao tác phân đoạn và việc khôi phục lại các quan hệ được thực hiện dựa trên các phép toán của đại số quan hệ, được trình bày tóm tắt dưới đây:

Bài giảng Cơ sở dữ liệu 2 | nvdingh@vnua.edu.vn

6

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

2.1.3. Các phép toán đại số quan hệ liệu trong phân đoạn DL

1. *Hợp (union)* của hai quan hệ khả hợp r và s , là một quan hệ, được ký hiệu và xác định như sau:

$$r \cup s = \{t \mid t \in r \text{ hoặc } t \in s\}$$

2. *Giao (intersection)* của hai quan hệ khả hợp r và s , là một quan hệ, được ký hiệu và xác định như sau:

$$r \cap s = \{t \mid t \in r \text{ và } t \in s\}$$

3. *Hiệu (set difference)* của hai quan hệ khả hợp r và s , là một quan hệ, được ký hiệu và xác định như sau:

$$r - s = \{t \mid t \in r \text{ và } t \notin s\}$$

4. *Tích Descartes* của hai quan hệ r và s là một quan hệ được ký hiệu và xác định như sau:

$$r \times s = \{(t, u) \mid t \in r, u \in s\}$$

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

2.1.3. Các phép toán đại số quan hệ liệu trong phân đoạn DL (tt)

5. *Chiếu (projection)* của quan hệ r lên tập thuộc tính X là một quan hệ, được ký hiệu và xác định như sau:

$$\pi_X(r) = \{t[X] \mid t \in r\}$$

6. *Chọn (selection)* của quan hệ r với điều kiện C , là một quan hệ, được ký hiệu và xác định như sau:

$$\sigma_C(r) = \{t \in r \mid C(t) = \text{True}\}$$

7. *Kết nối (join)* của hai quan hệ r và s có tập thuộc tính chung $X_R = X_S$, là một quan hệ, được ký hiệu và xác định như sau:

$$r \bowtie s = \{(t, u) \mid t \in r, u \in s, t[X_R] = u[X_S]\}$$

8. *Nửa kết nối (semi-join)* của hai quan hệ r và s có tập thuộc tính chung $X_R = X_S$, là một quan hệ, được ký hiệu và xác định như sau:

$$r \ltimes s = \pi_{X_R}(r \bowtie s)$$



2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

Một số thí dụ cho các phép toán

□ *Thí dụ 2.1. Cho 2 quan hệ r và s:*

r	A	B	s	B	C	D
	1	2		1	2	1
	2	4		2	4	0
	3	6		2	6	1
				4	8	0

▪ *Kết nối r và s:*

$r \bowtie s$	A	B	C	D
	1	2	4	0
	1	2	6	1
	2	4	8	0

▪ *Các nửa kết nối:*

$r \ltimes s$	A	B	$s \ltimes r$	B	C	D
	1	2		2	4	0
	2	4		2	6	1
				4	8	0

➤ *Nhận xét:* Phép nửa kết nối $r \ltimes s$ là phép rút gọn quan hệ r, chỉ giữ lại những bộ của r kết nối được với s.

2.1. Định nghĩa dữ liệu trong CSDL phân tán (tt)

Một số công thức tương đương trong đại số quan hệ

- $r \times s = s \times r$; $r \bowtie s = s \bowtie r$
- $(r \times s) \times t = r \times (s \times t)$; $(r \bowtie s) \bowtie t = r \bowtie (s \bowtie t)$
- Với $X \subseteq Y$ thì: $\pi_X(\pi_Y(r)) = \pi_X(r)$
- $\sigma_{C1}(\sigma_{C2}(r)) = \sigma_{C1 \wedge C2}(r)$
- $\pi_X(\sigma_C(r)) = \sigma_C(\pi_X(r))$
- $\sigma_C(r \cup s) = \sigma_C(r) \cup \sigma_C(s)$; $\sigma_C(r \cap s) = \sigma_C(r) \cap \sigma_C(s)$
- $\pi_X(r \cup s) = \pi_X(r) \cup \pi_X(s)$; $\pi_X(r \cap s) = \pi_X(r) \cap \pi_X(s)$
- $\sigma_C(r \bowtie s) = \sigma_C(r) \bowtie \sigma_C(s)$
- $r \ltimes s = (r \ltimes s) \bowtie s = (s \ltimes r) \bowtie r$
- $r \ltimes s = r \ltimes \pi_{R \cap S}(s)$; $s \ltimes r = s \ltimes \pi_{R \cap S}(r)$.

2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.1. Sự phân đoạn ngang trực tiếp (next T4, 16/1)

□ Định nghĩa 2.1

Sự phân đoạn ngang trực tiếp là sự phân hoạch một quan hệ thành các tập con các bộ; mỗi tập con là một quan hệ con được xác định bởi một phép chọn trên quan hệ tổng thể.

□ Thí dụ 2.2. Giả sử có quan hệ R_VANG (Rượu vang):

R_VANG	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	V1	AA	1982	13.2	12
	V2	BB	1987	12.9	7
	V3	AA	1984	12.8	10
	V4	CC	1983	12.3	14

Hãy phân đoạn quan hệ trên thành các đoạn:

- Đoạn 'R_VANG1' chứa các loại rượu sản xuất trước năm 1984
- Đoạn 'R_VANG2' chứa các loại rượu còn lại?

2.2. Phân đoạn dữ liệu trong CSDL phân tán

□ Giải Thí dụ 2.2 (tt):

Phân đoạn ngang trực tiếp quan hệ R_VANG thành các đoạn:

R_VANG	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	V1	AA	1982	13.2	12
	V2	BB	1987	12.9	7
	V3	AA	1984	12.8	10
	V4	CC	1983	12.3	14

- $R_VANG1 = \sigma_{N.HIEU < 1984} (R_VANG)$:

Kết quả:

R_VANG1	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	V1	AA	1982	13.2	12
	V4	CC	1983	12.3	14

- $R_VANG2 = \sigma_{N.HIEU \geq 1984} (R_VANG)$:

Kết quả:

R_VANG2	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	V2	BB	1987	12.9	7
	V3	AA	1984	12.8	10

- Biểu thức khôi phục: $R_VANG = R_VANG1 \cup R_VANG2$

➤ **Nhận xét:** Sự phân đoạn quan hệ R_VANG như trên thỏa mãn điều kiện: không mất thông tin, không trùng lặp và khôi phục lại được

2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.2. Sự phân đoạn ngang gián tiếp

□ Định nghĩa 2.2

Sự phân đoạn ngang gián tiếp là sự phân hoạch một quan hệ thành các tập con các bộ; mỗi tập con là một quan hệ con được xác định bởi một phép nối kết nối của quan hệ tổng thể với mỗi đoạn của một quan hệ khác.

□ Thí dụ 2.3. Giả sử có quan hệ T_THU (Tiêu thụ)

T_THU	MAKH	MARV	NGAY	D_DIEM	S_LUONG
	B1	V1	12/4/15	HANOI	5
	B2	V2	25/9/15	HUE	2
	B3	V3	12/4/15	HANOI	7

Hãy phân đoạn quan hệ trên thành các đoạn ngang:

- T_THU1: các loại rượu sản xuất trước năm 1984 được tiêu thụ
- T_THU2: các loại rượu sản xuất từ năm 1984 được tiêu thụ



2.2. Phân đoạn dữ liệu trong CSDL phân tán

□ Giải Thí dụ 2.3 (tt):

- Tính $T_THU1 = T_THU \bowtie R_VANG1$

T_THU	MAKH	MARV	NGAY	D_DIEM	S_LUONG	R_VANG1	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	B1	V1	12/4/15	HANOI	5		V1	AA	1982	13.2	12
	B2	V2	25/9/15	HUE	2		V4	CC	1983	12.3	14
	B3	V3	12/4/15	HANOI	7						

➤ Kết quả

T_THU1	MAKH	MARV	NGAY	D_DIEM	S_LUONG
	B1	V1	12/4/15	HANOI	5

- Tính $T_THU2 = T_THU \bowtie R_VANG2$

T_THU	MAKH	MARV	NGAY	D_DIEM	S_LUONG	R_VANG2	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	B1	V1	12/4/15	HANOI	5		V2	BB	1987	12.9	7
	B2	V2	25/9/15	HUE	2		V3	AA	1984	12.8	10
	B3	V3	12/4/15	HANOI	7						

➤ Kết quả:

T_THU2	MAKH	MARV	NGAY	D_DIEM	S_LUONG
	B2	V2	25/9/15	HUE	2
	B3	V3	12/4/15	HANOI	7



2.2. Phân đoạn dữ liệu trong CSDL phân tán

□ **Nhận xét cho Thí dụ 2.3 (tt):**

- Sự phân đoạn của quan hệ T_THU như trên sẽ là một phân tách không mất thông tin, không trùng lặp và khôi phục lại được
- Biểu thức khôi phục là: $T_THU = T_THU1 \cup T_THU2$
- Sự phân đoạn ngang gián tiếp như trên sẽ cho phép tăng hiệu năng của phép kết nối các quan hệ R_VANG với T_THU:
 - Bởi vì nếu bố trí trên cùng một trạm các cặp đoạn (R_VANG1 , T_THU1) và (R_VANG2 , T_THU2), thì tại một thời điểm có thể thực hiện song song hai phép kết nối R_VANG 1 với T_THU1 và R_VANG2 với T_THU2, cho phép tăng hiệu năng xử lý dữ liệu



2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.3. Sự phân đoạn dọc

□ **Định nghĩa 2.3**

Sự phân đoạn dọc là sự phân hoạch một quan hệ thành các tập các bộ con; mỗi tập là một quan hệ con được xác định bởi một phép chiếu cho quan hệ tổng thể.

- Để khôi phục được quan hệ ban đầu từ các đoạn dọc, mỗi đoạn dọc cần chứa khóa của quan hệ ban đầu, vì vậy phải chấp nhận phá vỡ tính không trùng lặp trong phân đoạn dọc. Như vậy các phân đoạn dọc đảm bảo không mất thông tin và khôi phục được quan hệ ban đầu.
- Để khôi phục quan hệ ban đầu, ta sử dụng phép kết nối các đoạn dọc theo sự bằng nhau của các thuộc tính chung (khóa).



2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.3. Sự phân đoạn dọc

❑ **Thí dụ 2.4.** Trở lại quan hệ R_VANG

R_VANG	MARV	V_NHO	N_HIEU	D_RUOU	GIA
	V1	AA	1982	13.2	12
	V2	BB	1987	12.9	7
	V3	AA	1984	12.8	10
	V4	CC	1983	12.3	14

▪ Hãy phân đoạn quan hệ trên thành các đoạn dọc nhờ các phép chiếu:

$VANG1 = \pi_{MARV, V_NHO, N_HIEU}(R_VANG)$

VANG1	MARV	V_NHO	N_HIEU
	V1	AA	1982
	V2	BB	1987
	V3	AA	1984
	V4	CC	1983

$VANG2 = \pi_{MARV, D_RUOU, GIA}(R_VANG)$

VANG2	MARV	D_RUOU	GIA
	V1	13.2	12
	V2	12.9	7
	V3	12.8	10
	V4	12.3	14

▪ Sự khôi phục quan hệ ban đầu nhờ phép kết nối:

$$R_VANG = VANG1 \bowtie VANG2$$



2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.4. Sự phân đoạn hỗn hợp

❑ **Định nghĩa 2.4**

Sự phân đoạn hỗn hợp là sự phân hoạch một quan hệ thành các tập con các bộ con; các bộ con được xác định bởi sự phân đoạn dọc, còn các tập con được xác định bởi sự phân đoạn ngang. Mỗi tập con này là một quan hệ con của quan hệ tổng thể.

- Như vậy sự phân đoạn hỗn hợp là một quan hệ vừa được phân đoạn ngang vừa được phân đoạn dọc. Trong mỗi đoạn ngang có thể lại được phân đoạn dọc, hoặc trong mỗi đoạn dọc lại được phân đoạn ngang.
- Các phân đoạn hỗn hợp đảm bảo không mất thông tin và khôi phục được quan hệ ban đầu.
- Để khôi phục quan hệ ban đầu, ta bắt đầu từ các quan hệ nhỏ nhất trước để khôi phục các đoạn dọc hoặc các đoạn ngang, sau đó khôi phục quan hệ ban đầu.



2.2. Phân đoạn dữ liệu trong CSDL phân tán

2.2.4. Sự phân đoạn hỗn hợp

Thí dụ 2.5.

- Cho quan hệ R_VANG:
- Hãy phân đoạn quan hệ trên thành các đoạn:

R_VANG	MARV	V_NHO	N_HIEU	D_RUOU	GIA
V1	AA	1982	13.2	12	
V2	BB	1987	12.9	7	
V3	AA	1984	12.8	10	
V4	CC	1983	12.3	14	

- $VANG1^* = \pi_{MARV, D_RUOU, GIA}(R_VANG)$
- $VANG2^* = \sigma_{N.HIEU < 1984}(\pi_{MARV, V.NHO, N.HIEU}(R_VANG))$
- $VANG3^* = \sigma_{N.HIEU \geq 1984}(\pi_{MARV, V.NHO, N.HIEU}(R_VANG))$

VANG1*	MARV	D_RUOU	GIA
V1		13.2	12
V2		12.9	7
V3		12.8	10
V4		12.3	14

VANG2*	MARV	V_NHO	N_HIEU
V1	AA	1982	
V4	CC	1983	

VANG3*	MARV	V_NHO	N_HIEU
V2	BB	1987	
V3	AA	1984	

- Công thức khôi phục quan hệ ban đầu

$$R_VANG = VANG1^* \bowtie (VANG2^* \cup VANG3^*)$$



2.3. Định giá các câu hỏi phân tán

2.3.1. Các bước định giá câu hỏi phân tán

Một số khái niệm:

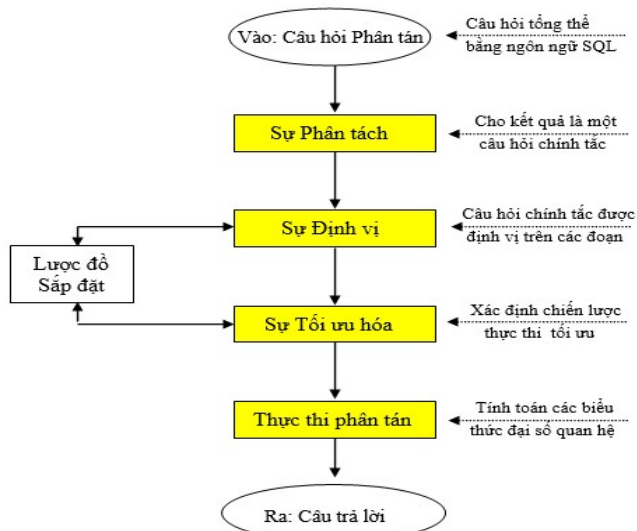
- ❑ **Câu hỏi phân tán**: Các câu hỏi phân tán thường được đưa ra ở mức các khung nhìn và được diễn tả phù hợp với các dữ liệu được mô tả bởi lược đồ tổng thể của CSDL phân tán.
- ❑ **Sự định giá một câu hỏi phân tán**: là việc biến đổi câu hỏi phân tán thành các câu hỏi có thể trả lời được bằng việc thực hiện các xử lý cục bộ trên các trạm (gọi là các phương án thực thi phân tán).
- ❑ **Các giai đoạn định giá câu hỏi trong CSDL phân tán**: quá trình định giá câu hỏi phân tán gồm 4 giai đoạn:
 - Giai đoạn 1: Sự phân tách câu hỏi.
 - Giai đoạn 2: Sự định vị câu hỏi
 - Giai đoạn 3: Sự tối ưu hóa câu hỏi
 - Giai đoạn 4: Sự thực thi phân tán (thực hiện các phép toán trên các trạm)

Các giai đoạn định giá câu hỏi phân tán trình bày trong sơ đồ sau:



2.3. Định giá các câu hỏi phân tán

2.3.1. Các bước định giá câu hỏi phân tán (tt)



2.3. Định giá các câu hỏi phân tán

2.3.2. Sự định vị câu hỏi phân tán

- Sự định vị một câu hỏi phân tán bao gồm 2 giai đoạn:

1. Diễn đạt câu hỏi chính tắc trên các đoạn:

Đó là việc biểu diễn các quan hệ tổng thể trong câu hỏi chính tắc qua các đoạn của nó, nhờ các biểu thức khôi phục cho các quan hệ tổng thể

2. Giảm lược câu hỏi:

Sử dụng các quy tắc tương đương trong ĐS quan hệ để rút gọn câu hỏi đã được diễn đạt trên các đoạn, bằng việc thay đổi thứ tự tính toán trên các quan hệ, loại bỏ các phép toán vô ích.

2.3. Định giá các câu hỏi phân tán

2.3.2. Sự định vị câu hỏi phân tán (tt)

▪ Một số quy tắc giản lược cho các biểu thức $\mathcal{DS}$ quan hệ:

1. Các phép toán một ngôi (chiếu và chọn) càng đẩy xuống gần các đoạn (tại các trạm) càng tốt, vì sẽ làm giảm kích thước các quan hệ trước khi thực hiện các phép toán 2 ngôi.
2. Một phép chọn trên một đoạn ngang với điều kiện của phép chọn mâu thuẫn với điều kiện của sự phân đoạn ngang sẽ là vô ích (vì sẽ cho kết quả là rỗng.)
3. Một phép chọn trên một đoạn ngang với điều kiện của phép chọn trùng với điều kiện của sự phân đoạn ngang sẽ là vô ích (vì sẽ cho kết quả bằng đoạn đó).
4. Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu (trừ thuộc tính chung cho sự khôi phục) đều không thuộc về đoạn đó sẽ là vô ích (vì sẽ cho kết quả là rỗng).

2.3. Định giá các câu hỏi phân tán

❑ **Thí dụ 2.6.** Xét các quan hệ R_VANG và T_THU đã cho trong 2.2

- Giả sử 2 quan hệ R_VANG và T_THU đã được phân đoạn như sau:

VANG1	MARV	V_NHO	N_HIEU
	V1	AA	1982
	V2	BB	1987
	V3	AA	1984
	V4	CC	1983

VANG2	MARV	D_RUOU	GIA
	V1	13.2	12
	V2	12.9	7
	V3	12.8	10
	V4	12.3	14

- $VANG1 = \Pi_{MARV, V_NHO, N_HIEU}(R_VANG)$
- $VANG2 = \Pi_{MARV, D_RUOU, GIA}(R_VANG)$
- $T_THU1 = \sigma_{D_DIEM = HANOI}(T_THU)$
- $T_THU2 = \sigma_{D_DIEM \neq HANOI}(T_THU)$

T_THU1	MAKH	MARV	NGAY	D_DIEM	S_LUONG
	B1	V1	12/4	HANOI	5
	B3	V3	12/4	HANOI	7

T_THU2	MAKH	MARV	NGAY	D_DIEM	S_LUONG
	B2	V2	25/9	HUE	2

- Xét câu hỏi phân tán:

”Tìm các vùng nho của các loại rượu vang được tiêu thụ tại HANOI” (*)

Phân tích câu hỏi phân tán (*) ta nhận được câu hỏi chính tắc dưới dạng một biểu thức đại số quan hệ:

$$\Pi_{V_NHO}(\sigma_{D_DIEM = HANOI}(R_VANG \bowtie T_THU)) \quad (1)$$

Ta sẽ tiến hành diễn đạt câu hỏi chính tắc (1), và giản lược câu hỏi này.

2.3. Định giá các câu hỏi phân tán

Thí dụ 2.6.(tt). Câu hỏi ch. tắc (1): $\Pi_{V_NHO}(\sigma_{D_DIEM = HANOI}(R_VANG \bowtie T_THU))$

- Thay các quan hệ R_VANG và T_THU bởi các biểu thức khôi phục của nó, ta nhận được câu hỏi chính tắc diễn đạt trên các đoạn là:

$$\Pi_{V_NHO}\{\sigma_{D_DIEM = HANOI}[(VANG1 \bowtie VANG2) \bowtie (T_THU1 \cup T_THU2)]\} \quad (2)$$

- Áp dụng các công thức tương đương trong ĐSQH cho (2), ta có:

$$\Pi_{V_NHO}\{[\Pi_{MARV, VNHO}(VANG1 \bowtie VANG2)] \bowtie [\Pi_{MARV}(\sigma_{D_DIEM = HANOI}(T_THU1 \cup T_THU2))]\} \quad (3)$$

- Áp dụng quy tắc giản lược 1: đưa các phép chiếu và phép chọn về các đoạn:

$$\Rightarrow \Pi_{V_NHO}\{(\Pi_{MARV, VNHO} VANG1 \bowtie \Pi_{MARV, VNHO} VANG2) \bowtie [\Pi_{MARV}(\sigma_{D_DIEM = HANOI} T_THU1) \cup \Pi_{MARV}(\sigma_{D_DIEM = HANOI} T_THU2)]\}$$

- Áp dụng các quy tắc giản lược 2, 3:

- Loại bỏ các kết quả rỗng: $\Pi_{MARV, VNHO}(VANG2) = \emptyset$ $\sigma_{D_DIEM = HANOI}(T_THU2) = \emptyset$
- Loại bỏ phép chọn vô ích: $\sigma_{D_DIEM = HANOI}(T_THU1) = T_THU1$

Ta nhận được câu hỏi chính tắc đã giản lược và định vị trên các đoạn:

$$\Pi_{V_NHO}\{\Pi_{MARV, VNHO}(VANG1) \bowtie \Pi_{MARV}(T_THU1)\} \quad (4)$$

2.3. Định giá các câu hỏi phân tán

Thí dụ 2.6 (tt). Như vậy, câu hỏi : "Tìm các vùng nho của các loại rượu vang được tiêu thụ tại HANOI", được định giá bởi việc thực hiện biểu thức đại số quan hệ:

$$\Pi_{V_NHO}\{\Pi_{MARV, VNHO}(VANG1) \bowtie \Pi_{MARV}(T_THU1)\} \quad (4)$$

Thực hiện (4) trên các đoạn VANG1 và T_THU1.

VANG1	MARV	V_NHO		T_THU1	MARV	
	V1	AA			V1	
	V2	BB			V3	
	V3	AA				
	V4	CC				

➤ Các bước thực hiện biểu thức (4)

- Bước 1: thực hiện phép chiếu 1 trên VANG1 → kết quả là (P1)
- Bước 2: thực hiện phép chiếu 2 trên T_THU1 → quan hệ kết quả là (P2)
- Bước 3: Kết nối hai quan hệ kết quả (P1) và (P2) nhận được quan hệ:
- Thực hiện phép chiếu Π_{V_NHO} trên quan hệ cuối cùng:
- Câu trả lời: các loại rượu vang được tiêu thụ tại HANOI có vùng nho 'AA'

V_NHO
AA
AA

2.3. Định giá các câu hỏi phân tán

2.3.3. Tối ưu hóa câu hỏi phân tán

- Sau giai đoạn định vị, câu hỏi phân tán đã được diễn đạt trên các đoạn, đó là một biểu thức ĐSQH với các quan hệ con đặt trên các trạm.

Một số khái niệm

- Có nhiều cách để tính được giá trị biểu thức ĐSQH giữa các đoạn, mỗi cách tính được gọi là một **chiến lược thực thi câu hỏi phân tán**.
- Tối ưu hóa câu hỏi phân tán**: là việc xác định một chiến lược thực thi câu hỏi phân tán, làm cực tiểu hóa một hàm chi phí nào đó.
- Có nhiều cách chọn hàm chi phí, thường chọn là **chi phí truyền dữ liệu giữa các trạm**, chi phí này liên hệ tỷ lệ với số phần tử dữ liệu cần truyền tải giữa các trạm.
- Như vậy tối ưu hóa câu hỏi phân tán là **làm cực tiểu số lượng phần tử dữ liệu cần truyền tải** giữa các trạm
- Chú ý rằng chỉ các phép toán 2 ngôi (hợp, kết nối...) mới cần truyền DL.



2.3. Định giá các câu hỏi phân tán

2.3.3. Tối ưu hóa câu hỏi phân tán (tt)

- Các biểu thức của câu hỏi trong CSDLPT thường gặp các phép toán Chiều (P), Chon (S) và kết nối (J), và được gọi là **các biểu thức PSJ**.
- Chỉ có phép toán kết nối (J) là phép toán 2 ngôi, khi thực hiện thì phải đặt cả 2 quan hệ tại cùng 1 trạm, do đó phát sinh chi phí truyền dữ liệu để chuyển một quan hệ ở trạm này về trạm khác.
- Như vậy, việc tối ưu chi phí truyền dữ liệu chủ yếu thực hiện khi kết nối các quan hệ
- Hai cách tiếp cận để giảm chi phí truyền dữ liệu

1. Xác định một trình tự kết nối tối ưu

2. Rút gọn kích thước các quan hệ trước khi kết nối



2.3. Định giá các câu hỏi phân tán

2.3.3. Tối ưu hóa câu hỏi phân tán (tt)

2.3.3.1. Xác định một trình tự kết nối tối ưu

□ Định nghĩa 2.5.

Một trình tự kết nối tối ưu cho các quan hệ là một thứ tự thực hiện các phép kết nối sao cho mọi kết quả trung gian có số bộ không lớn hơn số bộ ở kết quả cuối cùng.

□ **Thí dụ 2.7.** Xem xét kết nối của 3 quan hệ r_1, r_2, r_3 đặt trên các trạm cách xa nhau:

r_1	A	B	C
1	3	5	
1	4	5	
2	3	5	
2	4	6	

r_2	C	D	E
5	7	10	
5	8	10	
5	9	11	
6	8	11	

r_3	B	C	D	F
3	5	7	1	
4	5	8	2	
3	5	9	3	
4	6	8	4	

- Do phép kết nối có tính giao hoán, ta có thể tính dãy kết nối $r_1 \bowtie r_2 \bowtie r_3$ theo nhiều trình tự khác nhau nhưng kết quả là duy nhất.
- Ta xem xét 2 trình tự khác nhau: $(r_1 \bowtie r_2) \bowtie r_3$ và $r_1 \bowtie (r_2 \bowtie r_3)$



2.3. Định giá các câu hỏi phân tán

Thí dụ 2.7 (tt).

□ Với trình tự 1: kết nối $(r_1 \bowtie r_2) \bowtie r_3$ được thực hiện qua 2 bước:

- Bước 1.** tính kết nối $(r_1 \bowtie r_2)$ được kết quả như hình bên.
 - Ta nhận được kết quả trung gian với 10 bộ
- Bước 2.** kết nối kết quả trung gian với r_3 để được kết quả cuối cùng
 - Ta nhận được kết cuối cùng $(r_1 \bowtie r_2) \bowtie r_3$ với chỉ 6 bộ.
- Nhận xét: Theo trình tự kết nối này, kết quả trung gian có 10 bộ, nhiều hơn số bộ ở kết quả cuối cùng là 6 bộ, vậy đây **không phải là trình tự kết nối tối ưu**.

$r_1 \bowtie r_2$	A	B	C	D	E
1	3	5	7	10	
1	3	5	8	10	
1	3	5	9	10	
1	4	5	7	10	
1	4	5	8	10	
1	4	5	9	10	
2	3	5	7	10	
2	3	5	8	10	
2	3	5	9	10	
2	4	6	8	11	

$(r_1 \bowtie r_2) \bowtie r_3$	A	B	C	D	E	F
1	3	5	7	10	1	
1	3	5	9	10	3	
1	4	5	8	10	2	
2	3	5	7	10	1	
2	3	5	9	10	3	
2	4	6	8	11	4	



2.3. Định giá các câu hỏi phân tán

Thí dụ 2.7 (tt).

□ Với trình tự 2: kết nối $r_1 \triangleright \triangleleft (r_2 \triangleright \triangleleft r_3)$ được thực hiện qua 2 bước:

- **Bước 1.** tính kết nối $(r_2 \triangleright \triangleleft r_3)$ được kết quả như hình bên.

$r_2 \triangleright \triangleleft r_3$	B	C	D	E	F
	3	5	7	10	1
	4	5	8	10	2
	3	5	9	11	3
	4	6	8	11	4

- Ta nhận được kết quả trung gian với chỉ 4 bộ

- **Bước 2.** kết nối kết quả trung gian với r_1 để được kết quả cuối cùng

- Ta nhận được kết cuối cùng $r_1 \triangleright \triangleleft (r_2 \triangleright \triangleleft r_3)$ với 6 bộ.

$(r_1 \triangleright \triangleleft r_2) \triangleright \triangleleft r_3$	A	B	C	D	E	F
	1	3	5	7	10	1
	1	3	5	9	10	3
	1	4	5	8	10	2
	2	3	5	7	10	1
	2	3	5	9	10	3
	2	4	6	8	11	4

- Nhận xét: Theo trình tự kết nối này, kết quả trung gian có 4 bộ, không lớn hơn số bộ ở kết quả cuối cùng (6 bộ), vậy đây là một **trình tự kết nối tối ưu**.

2.3. Định giá các câu hỏi phân tán

2.3.3. Tối ưu hóa câu hỏi phân tán (tt)

2.3.3.2. Rút gọn kích thước các quan hệ trước khi kết nối

Việc sử dụng các phép nửa kết nối để rút gọn các quan hệ trước khi kết nối gọi là **chiến lược nửa kết nối**, gồm các bước sau:

1. Tính $\Pi_{R \cap S}(s)$ tại trạm chứa quan hệ s. không phải truyền dữ liệu.
2. Chuyển $\Pi_{R \cap S}(s)$ sang trạm chứa r, chi phí là $T'(s)$. $T'(s)$ là số phần tử dữ liệu của $\Pi_{R \cap S}(s)$.
3. Tính $r \triangleright \triangleleft s = r \triangleright \triangleleft \Pi_{R \cap S}(s)$ tại trạm chứa r, không phải truyền dữ liệu.
4. Chuyển $r \triangleright \triangleleft s$ vừa tính được sang trạm chứa s, chi phí là $T''(r)$, $T''(r)$ là số phần tử của $r \triangleright \triangleleft s$.
5. Tính kết nối tại vị trí của s theo công thức: $r \triangleright \triangleleft s = (r \triangleright \triangleleft s) \triangleright \triangleleft s$.

Trong 5 bước trên, chi phí truyền dữ liệu chỉ phát sinh ở bước 2 và bước 4.

2.3. Định giá các câu hỏi phân tán

Tính toán chi phí giữa kết nối trực tiếp và chiến lược kết nối:

Giả sử ta cần kết nối 2 quan hệ r và s là các quan hệ đặt ở các trạm khác nhau, mỗi quan hệ có tương ứng $T(r)$ và $T(s)$ phần tử dữ liệu.

- **Kết nối trực tiếp:** Do phép kết nối là phép toán 2 ngôi, ta cần chuyển r sang trạm chứa s hoặc chuyển s sang trạm chứa r (tùy theo $T_r < T_s$, hoặc ngược lại) sau đó thực hiện phép kết nối tại trạm chứa cả r và s . Để đơn giản ta coi chi phí bằng số phần tử cần truyền tải.

➤ Vậy chi phí theo cách kết nối trực tiếp ($cost_{tt}$) là :

$$cost_{tt}(r, s) = \min \{T(r), T(s)\} \quad (1)$$

- **Chiến lược nửa kết nối:** Trong 5 bước của chiến lược này chỉ có bước 2 và bước 4 phát sinh chi phí.

➤ Vậy ta có chi phí theo chiến lược nửa kết nối ($cost_{nkn}$) là:

$$cost_{nkn}(r, s) = \min \{T'(s) + T''(r); T'(r) + T''(s)\} \quad (2)$$

Chi phí theo chiến lược nửa kết nối nói chung luôn nhỏ hơn kết nối trực tiếp.

2.3. Định giá các câu hỏi phân tán

- ❑ **Thí dụ 2.8.** Trong CSDL phân tán, có các quan hệ r, s đặt tại hai trạm ở cách xa nhau.

r	A	B	E
7	4	1	
8	5	2	
7	6	3	
9	7	4	
8	8	5	

s	B	C	D	F
4	6	7	1	
5	6	7	2	
4	11	9	3	
5	11	9	4	

- Tính kết nối của r với s theo cách trực tiếp và theo chiến lược nửa kết nối.
- Đánh giá chi phí theo mỗi cách kết nối.

- ❑ **Giải a/.** Kết nối trực tiếp: Do $T(r) < T(s)$ nên ta chuyển r sang trạm chứa s để thực hiện kết nối:

$r \triangleright \triangleleft s$	A	B	C	D	E	F
7	4	6	7	1	1	
8	5	6	7	2	2	
7	4	11	9	1	3	
8	5	11	3	2	4	

- Kết quả kết nối:

- Chi phí truyền dữ liệu theo cách kết nối trực tiếp:

$$cost_{tt}(r, s) = \min \{T(s), T(r)\} = 15$$

2.3. Định giá các câu hỏi phân tán

Thí dụ 2.8. (tt)

□ Giải b/. Tính kết nối theo chiến lược nửa kết nối.

1. Tính $\Pi_{R \cap S}(s)$ tại trạm chứa s, nhận được kết quả:

$\Pi_{R \cap S}(s)$	B
	4
	5

2. Chuyển $\Pi_{R \cap S}(s)$ sang trạm chứa r, chi phí ở bước này là $T'(s) = 2$.

3. Tính $r \triangleright < s = r \triangleright < \Pi_{R \cap S}(s)$ tại trạm chứa r, kết quả của phép nửa kết nối là:

$r \triangleright < s$	A	B	E
	7	4	1
	8	5	2

4. Chuyển $r \triangleright < s$ sang trạm chứa s, chi phí ở bước này là $T''(r) = 6$

5. Tính kết nối tại trạm chứa s theo

$$\text{công thức } r \triangleright < s = (r \triangleright < s) \triangleright < s$$

kết quả của phép kết nối là:

$r \triangleright < s$	A	B	C	D	E	F
	7	4	6	7	1	1
	8	5	6	7	2	2
	7	4	11	9	1	3
	8	5	11	3	2	4

▪ Chi phí truyền dữ liệu theo chiến lược nửa kết nối:

$$\text{cost}_{\text{nkn}}(r, s) = T'(s) + T''(s) = 8$$



- [Bài tập chương 2:](#)
- https://fita.vnua.edu.vn/wp-content/uploads/2013/06/BaiTap_Ch.2.pdf
- Yêu cầu làm tất cả bài tập vào giấy A4.
(Có thể được sử dụng làm tài liệu tham khảo khi thi)