

Chương 2

Truy vấn trong CSDL phân tán

Trong chương này, chúng ta đề cập đến một số khái niệm và kết quả cơ bản liên quan đến việc truy vấn trong các CSDL phân tán.

- 2.1. Định nghĩa dữ liệu phân tán
 - 2.1.1. Mô tả dữ liệu
 - 2.1.2. Các nguyên tắc phân đoạn dữ liệu
 - 2.1.3. Các phép toán đại số quan hệ cho phép phân đoạn
- 2.2. Các phương pháp phân đoạn dữ liệu trong CSDL phân tán
 - 2.2.1. Sự phân đoạn ngang
 - 2.2.2. Sự phân đoạn dọc
 - 2.2.3. Sự phân đoạn hỗn hợp
- 2.3. Định giá các câu hỏi phân tán
 - 2.3.1. Các bước định giá câu hỏi phân tán
 - 2.3.2. Sự định vị câu hỏi phân tán
 - 2.3.3. Sự giản lược câu hỏi phân tán
- 2.4. Tối ưu hóa câu hỏi phân tán.
 - 2.4.1. Các khái niệm tối ưu hóa câu hỏi phân tán
 - 2.4.2. Tối ưu hóa nhờ trình tự kết nối
 - 2.4.3. Tối ưu hóa nhờ phép nửa kết nối

Bài tập chương 2.

2.1. ĐỊNH NGHĨA DỮ LIỆU TRONG HỆ CƠ SỞ DỮ LIỆU 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.

2.1.1. Mô tả dữ liệu

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.

Trạm tạo sinh có thể biết danh sách các trạm được phép truy cập các dữ liệu riêng của nó, nhằm thông báo cho các trạm này mọi thay đổi về mô tả dữ liệu.

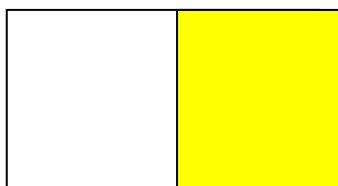
2.1.2. Các kiểu phân đoạn dữ liệu

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, đó là 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.

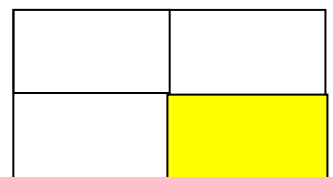
Có ba kiểu phân đoạn: *phân đoạn ngang*, *phân đoạn dọc* và *phân đoạn hỗn hợp* nhằm chia các quan hệ thành các quan hệ con, được khai báo trong lược đồ khái niệm tổng thể và được sắp đặt phù hợp trên các trạm (hình 2.1).



a). Phân đoạn ngang



b). Phân đoạn dọc



c). Phân đoạn hỗn hợp

Hình 2.1. Các kiểu phân đoạn dữ liệu

Sự phân đoạn một quan hệ được xác định bởi người quản trị CSDL, theo thiết kế của hệ thống thông tin nhằm tối ưu hóa việc cập nhật dữ liệu. Sự phân đoạn một quan hệ phải thỏa mãn các yêu cầu:

- 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. Trước khi nghiên cứu các phương pháp phân đoạn dữ liệu, ta nhắc lại các phép toán của đại số quan hệ.

2.1.3. Các phép toán đại số quan hệ cho phân đoạn dữ liệu

Việc phân đoạn một quan hệ tổng thể thành các đoạn (mảnh) và khôi phục các quan hệ tổng thể từ các đoạn được thực hiện bởi các phép toán đại số quan hệ. Đại số quan hệ là cơ sở quan trọng của một ngôn ngữ bậc cao được sử dụng để thao tác trên các quan hệ. Ngôn ngữ này bao gồm hai nhóm phép toán: các phép toán trên tập hợp (phép hợp, phép giao, phép trừ và tích Descartes) và các phép toán chỉ trên các quan hệ (phép chọn, phép chiếu, phép kết nối và phép nửa kết nối). Trước hết chúng ta đưa ra một số quy ước về ký pháp và nhắc lại các khái niệm liên quan đến các phép toán đại số quan hệ:

- Một lược đồ quan hệ R xác định trên tập thuộc tính $\Omega = \{A_1, A_2, \dots, A_n\}$, có thể ký hiệu $R(A_1, A_2, \dots, A_n)$, hay $R = \{A_1, A_2, \dots, A_n\}$. Một thể hiện của lược đồ quan hệ R là một quan hệ, được ký hiệu bằng chữ r (in thường). Quan hệ r có thể biểu diễn bằng một bảng.
- Tập các giá trị có thể của thuộc tính A_j , ký hiệu $Dom(A_j)$ và gọi là miền trị của thứ j của quan hệ R .
- Hai quan hệ là *cùng bậc* nếu chúng có cùng số thuộc tính.
- Hai quan hệ là *khả hợp* nếu chúng có cùng bậc và thuộc tính thứ j của quan hệ này có cùng miền trị với thuộc tính thứ j của quan hệ kia.
- Rõ ràng, các quan hệ xác định trên cùng một lược đồ quan hệ là các quan hệ khả hợp.

Sau đây, chúng ta nhắc lại một số định nghĩa về các phép toán đại số quan hệ và một số công thức thường được sử dụng trong đại số quan hệ.

2.1.3.1. Định nghĩa các phép toán đại số quan hệ

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

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

Như vậy, hợp của hai quan hệ r và s là một quan hệ chứa tất cả các bộ của r và s (những bộ chung chỉ xuất hiện 1 lần).

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

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

Như vậy, giao của hai quan hệ r và s là một quan hệ chỉ chứa các bộ chung của r và s .

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

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

Như vậy, hiệu của hai quan hệ r và s là một quan hệ chỉ chứa các bộ thuộc r mà không thuộc s .

4. **Tích Descartes:** Cho r là một quan hệ xác định trên tập thuộc tính $\{A_1, A_2, \dots, A_n\}$ và s là quan hệ xác định trên $\{B_1, B_2, \dots, B_m\}$. *Tích Descartes* của hai quan hệ r và s là một quan hệ được ký hiệu và xác định như sau:

$$r \times s = \{ t \mid t = (a_1, \dots, a_n, b_1, \dots, b_m), \text{ với } (a_1, \dots, a_n) \in r; (b_1, \dots, b_m) \in s \}$$

Như vậy, Tích Descartes của hai quan hệ r và s là một quan hệ mà mỗi bộ là ghép nối của một bộ trong r với một bộ trong s . Rõ ràng nếu r có p bộ và s có q bộ thì $r \times s$ chứa $p \cdot q$ bộ.

5. **Phép chiếu :** Cho r là một quan hệ trên tập thuộc tính $\Omega = \{A_1, A_2, \dots, A_n\}$, giả sử $X \subseteq \Omega$, *chiếu (projection)* của quan hệ r lên tập thuộc tính X cho kết quả là một quan hệ con của quan hệ r , được ký hiệu và xác định như sau:

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

Trong đó $t[X]$ là hạn chế của bộ t của r trên tập thuộc tính X , với $X = \{A_{i_1}, \dots, A_{i_k}\}$, thì $t[X] = (t[A_{i_1}], \dots, t[A_{i_k}])$

Như vậy, chiếu của quan hệ r trên tập thuộc tính X , là quan hệ con nhận được từ r bằng cách bỏ đi những thuộc tính (các cột) trong r mà không có mặt trong X . Với quan hệ con nhận được, nếu xuất hiện nhiều bộ giống nhau thì chỉ giữ lại một bộ trong số các bộ giống nhau.

6. **Phép chọn:** *Phép chọn (selection)* là phép toán cho kết quả là một quan hệ con gồm một tập con các bộ của quan hệ r thỏa mãn điều kiện C nào đó, được ký hiệu và xác định như sau:

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

Trong đó, điều kiện C là một biểu thức boolean của các toán hạng, mỗi toán hạng là một phép so sánh giữa hai biến là hai thuộc tính, hoặc giữa một biến là thuộc tính và một hằng, C cho giá trị đúng hoặc sai đối với mỗi bộ của r , nếu đúng, bộ đó được chọn vào quan hệ mới $\sigma_C(r)$. Các phép so sánh trong C bao gồm: $=$; $<$; $\leq$; $>$; $\geq$ và $\neq$, các phép toán logic gồm $\wedge$ (và), $\vee$ (hoặc), $\neg$ (phủ định).

Như vậy, chọn của quan hệ r theo điều kiện C , ta được một quan hệ con $\sigma_C(r)$ với cùng tập thuộc tính của r , mà chỉ chứa những bộ thỏa điều kiện C .

7. **Phép kết nối (join):** Cho r là một quan hệ xác định trên tập thuộc tính $\{A_1, A_2, \dots, A_n\}$ và s là quan hệ xác định trên $\{B_1, B_2, \dots, B_m\}$. Giả sử r và s có một số thuộc tính chung (tức là $\exists i, j$ để $A_i = B_j$), khi đó kết nối của quan hệ r và quan hệ s là một quan hệ với các bộ là kết nối của các bộ trong r với các bộ trong s có giá trị các thuộc tính chung bằng nhau. *Kết nối của hai quan hệ r và s* được ký hiệu và xác định như sau:

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

Như vậy, kết nối của hai quan hệ r và s là một quan hệ có tập thuộc tính là tất cả các thuộc tính của cả r và s (với mỗi cặp thuộc tính chung cùng tên ở cả 2 quan hệ chỉ giữ lại 1 thuộc tính cho

quan hệ mới), các bộ của quan hệ mới là ghép của 2 bộ trong r và s mà có giá trị các thuộc tính chung là bằng nhau. Phép kết nối theo định nghĩa trên cũng gọi là *kết nối tự nhiên*.

8. Phép nửa kết nối (semi-join) : Cho r là một quan hệ xác định trên tập thuộc tính $\{A_1, A_2, \dots, A_n\}$ và s là quan hệ xác định trên $\{B_1, B_2, \dots, B_m\}$. Giả sử r và s có một số thuộc tính chung (tức là $\exists i, j$ để $A_i = B_j$), khi đó *nửa kết nối* của quan hệ r và quan hệ s là một quan hệ được ký hiệu và xác định như sau:

$$r \bowtie s = \Pi_R(r \bowtie s)$$

Như vậy, nửa kết nối của hai quan hệ r và s thực chất là việc loại đi những bộ của r mà không kết nối được với s.

Thí dụ 2.1. Cho các quan hệ r và s như sau:

r	A	B
	1	2
	2	4
	3	6

s	B	C	D
	1	2	1
	2	4	1
	3	6	1
	4	8	1

Tính được kết nối $r \bowtie s$:

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

Dễ dàng kiểm tra được rằng với hai quan hệ trên thì: $r \bowtie s = s \bowtie r$

Sinh viên tự tính xem như là bài tập.

Thí dụ 2.2. Với các quan hệ r và s như trong thí dụ 2.1, tính được các nửa kết nối:

$r \bowtie s$	A	B
	1	2
	2	4

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

2.1.3.2. Một số công thức trong đại số quan hệ

Trong CSDL phân tán, việc định giá câu hỏi đưa về việc tính toán các biểu thức đại số quan hệ, là các biểu thức gồm các phép toán đại số quan hệ và các toán hạng là các quan hệ.

Ký hiệu r, s, t là các quan hệ còn R, S, T là các tập thuộc tính tương ứng của chúng.

Ta có các công thức tương đương dưới đây, có thể được dùng trong khi tính toán các biểu thức đại số quan hệ.

1. Luật giao hoán đối với tích Descartes và phép kết nối:

- $r \times s = s \times r$
- $r \bowtie s = s \bowtie r$

2. Luật kết hợp đối với tích Descartes và phép kết nối:

- $(r \times s) \times t = r \times (s \times t)$
- $(r \bowtie s) \bowtie t = r \bowtie (s \bowtie t)$

3. Phép chiếu tương đương: Nếu r là một quan hệ trên tập thuộc tính $\Omega = \{A_1, A_2, \dots, A_n\}$, và giả sử $\{A_1, A_2, \dots, A_k\} \subseteq \{B_1, B_2, \dots, B_m\} \subseteq \Omega$, tức là $A_1, A_2, \dots, A_k$ phải ở trong số các thuộc tính $B_1, B_2, \dots, B_m$ thì:

$$\Pi_{A_1 A_2 \dots A_k}(\Pi_{B_1 B_2 \dots B_m}(r)) = \Pi_{A_1 A_2 \dots A_k}(r)$$

4. Phép chọn tương đương:

$$\sigma_{F_1}(\sigma_{F_2}(r)) = \sigma_{F_1 \wedge F_2}(r), \text{ trong đó } F_1 \text{ và } F_2 \text{ là các điều kiện logic nào đó.}$$

5. Phép chiếu liên kết với phép chọn:

$$\Pi_{A_1 A_2 \dots A_k}(\sigma_F(r)) = \sigma_F(\Pi_{A_1 A_2 \dots A_k}(r))$$

6. Phép chọn liên kết với phép hợp, phép giao, phép trừ:

$$\begin{aligned} \sigma_F(r \cup s) &= \sigma_F(r) \cup \sigma_F(s) \\ \sigma_F(r \cap s) &= \sigma_F(r) \cap \sigma_F(s) \\ \sigma_F(r - s) &= \sigma_F(r) - \sigma_F(s) \end{aligned}$$

7. Phép chiếu liên kết với phép hợp, phép giao, phép trừ :

$$\begin{aligned} \Pi_X(r \cup s) &= \Pi_X(r) \cup \Pi_X(s) \\ \Pi_X(r \cap s) &= \Pi_X(r) \cap \Pi_X(s) \\ \Pi_X(r - s) &= \Pi_X(r) - \Pi_X(s) \end{aligned}$$

8. Phép chọn liên kết với phép kết nối

$$\sigma_F(r \bowtie s) = \sigma_F(r) \bowtie \sigma_F(s)$$

9. Tính kết nối qua phép nửa kết nối.

$$r \bowtie s = (r \bowtie s) \bowtie s = (s \bowtie r) \bowtie r$$

10. Tính nửa kết nối qua phép kết nối và phép chiếu

$$\begin{aligned} r \bowtie s &= r \bowtie \Pi_{R \cap S}(s) \\ s \bowtie r &= s \bowtie \Pi_{R \cap S}(r), \text{ với } \Pi_{R \cap S}(s) \text{ là chiếu của } s \text{ trên tập thuộc tính chung của } r \text{ và } s, \text{ còn } \Pi_{R \cap S}(r) \text{ là chiếu của } r \text{ trên tập thuộc tính chung của } r \text{ và } s. \end{aligned}$$

2.2. CÁC PHƯƠNG PHÁP PHÂN ĐOẠN DỮ LIỆU TRONG CSDL PHÂN TÁN

Có ba kiểu phân đoạn các quan hệ trong CSDL phân tán: phân đoạn ngang (gồm phân đoạn ngang trực tiếp và phân đoạn ngang gián tiếp), phân đoạn dọc và phân đoạn hỗn hợp.

2.2.1 Sự phân đoạn ngang trực tiếp

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 2.3: 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

- *Giải thích:*
 - MARV: mã loại rượu vang
 - V_NHO: Vùng nho (vùng nguyên liệu, chẳng hạn Vang Bordeaux, Vang Đà Lạt...)
 - N_HIEU: Niên hiệu (năm sản xuất)
 - D_RUOU: Độ rượu (nồng độ cồn của rượu vang).
 - GIA: Giá tiền 1 chai rượu

Ta có thể phân hoạch quan hệ R_VANG thành các đoạn con bằng phép phân đoạn ngang trực tiếp nhờ các phép chọn sau:

$$R_VANG1 = \sigma_{N.HIEU < 1984}(R_VANG), \text{ và}$$

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

Kết quả nhận được các đoạn VANG_1 và VANG_2, là các quan hệ con như sau:

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

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

Sự phân đoạn của quan hệ R_VANG như trên sẽ là một phân tách không mất thông tin, không trùng lặp và khôi phục lại được.

- Thật vậy, vì $R_VANG1 \cap R_VANG2 = \emptyset$ nên các đoạn con của quan hệ R_VANG là không trùng lặp
- Sự khôi phục có thể được thực hiện nhờ phép toán:

$$R_VANG = R_VANG1 \cup R_VANG2.$$

Chú ý rằng việc lựa chọn các điều kiện của phép chọn là quan trọng, vì chẳng hạn nếu ta chọn điều kiện trong phép chọn xác định R_VANG2 là 'N_HIEU $\geq$ 1983' thì khi đó tính không trùng lặp sẽ bị vi phạm vì bộ thứ tư của quan hệ R_VANG sẽ thuộc về cả hai đoạn.

Nhận xét: Sự phân đoạn ngang sẽ tạo thuận lợi trong khai thác CSDL, chẳng hạn nếu tìm kiếm các loại rượu vang có niên hiệu 1982 thì chỉ cần tham chiếu đến đoạn R_VANG1, số các thao tác tìm kiếm sẽ giảm đi rất nhiều.

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 nửa kết nối của quan hệ tổng thể với mỗi đoạn của một quan hệ khác.

Thí dụ 2.4: 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

- *Giải thích:*
 - MAKH: mã khách hàng
 - MARV: mã loại rượu vang
 - NGÀY: ngày/ tháng/ năm mua hàng
 - D_DIEM: địa điểm mua hàng
 - S_LUONG: số chai rượu mua

Ta có thể phân hoạch quan hệ T_THU thành các đoạn con bằng phép phân đoạn ngang gián tiếp nhờ các phép nửa kết nối với các đoạn của quan hệ R_VANG như sau:

$$T_THU1 = T_THU \bowtie R_VANG1$$

$$T_THU2 = T_THU \bowtie R_VANG2$$

Thực hiện phép nửa kết nối T_THU $\bowtie$ R_VANG1, nhận được quan hệ con T_THU1:

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

Kết quả phép nửa kết nối:

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

Thực hiện phép nửa kết nối T_THU $\bowtie$ R_VANG2, nhận được quan hệ con T_THU2:

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

Kết quả phép nửa kết nối:

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

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.

- Thật vậy, vì $T_THU1 \cap T_THU2 = \emptyset$ nên các đoạn con của quan hệ T_THU là không trùng lặp
- Sự khôi phục có thể được thực hiện nhờ phép toán: $T_THU = T_THU1 \cup T_THU2$.

Nhận xét : 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, nếu ta bố trí trên cùng một trạm các cặp đoạn (R_VANG , T_THU1) và (R_VANG , T_THU2) thì có thể thực hiện song song hai phép kết nối $R_VANG1 \bowtie T_THU1$ và $R_VANG2 \bowtie T_THU2$, tăng hiệu năng xử lý dữ liệu.

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 này là một quan hệ con được xác định bởi một phép chiếu cho quan hệ tổng thể.

Để khôi phục lại quan hệ tổng thể từ các đoạn dọc, cần thêm vào mỗi đoạn dọc một hay nhiều thuộc tính chung làm khóa. Khi đó, việc khôi phục có được nhờ vào phép kết nối các đoạn con theo sự bằng nhau của các thuộc tính chung. Do việc thêm vào các thuộc tính chung này mà quy tắc không trùng lặp bị vi phạm, nhưng phải chấp nhận sự vi phạm này để đảm bảo sự khôi phục lại của quan hệ tổng thể.

Thí dụ 2.5. 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

Ta phân hoạch quan hệ này thành hai đoạn dọc nhờ các phép chiếu trên quan hệ R_VANG.

- Với phép chiếu: $VANG1 = \Pi_{MARV, V_NHO, N_HIEU}(R_VANG)$, ta được quan hệ:

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

- Với phép chiếu $VANG2 = \Pi_{MARV, D_RUOU, GIA}(R_VANG)$, ta được quan hệ:

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 R_VANG nhờ phép kết nối: $R_VANG = VANG1 \bowtie VANG2$

Các bạn sinh viên tự kiểm tra lại biểu thức khôi phục này.

Nhận xét : Sự phân đoạn dọc như trên làm thuận lợi cho việc xử lý các câu hỏi có liên quan đến các thuộc tính được dùng để định nghĩa các đoạn. Chẳng hạn, để tìm các vùng nho và niên hiệu của các loại rượu vang, ta chỉ cần hướng tới đoạn VANG1.

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

Phân đoạn hỗn hợp cho phép tối ưu hóa các câu hỏi chọn-chiều hay kết nối-chiều.

Thí dụ 2.6. Trở lại quan hệ R_VANG. Ta áp dụng phép phân đoạn hỗn hợp quan hệ này thành các đoạn con như sau:

- $VANG1^* = \Pi_{MARV, D_RUOU, GIA}(R_VANG)$
- $VANG2^* = \sigma_{N_HIEU < 1984}(\Pi_{MARV, V_NHO, N_HIEU}(R_VANG))$
- $VANG3^* = \sigma_{N_HIEU \geq 1984}(\Pi_{MARV, V_NHO, N_HIEU}(R_VANG))$

Quan hệ tổng thể ban đầu:

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

Thực hiện các biểu thức đại số quan hệ trên, ta nhận được các quan hệ kết quả:

VANG1				VANG1* = VANG2		
	MARV	V_NHO	N_HIEU	MARV	Đ_RUOU	GIA
VANG2*	V1	AA	1982	V1	13.2	12
	V4	CC	1983	V4	12.3	14
VANG3*	V2	BB	1987	V2	12.9	7
	V3	AA	1984	V3	12.8	10

- Sự khôi phục quan hệ ban đầu, trước hết phải bắt đầu từ các đoạn bên trong:

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

2.3 ĐỊNH GIÁ CÁC CÂU HỎI PHÂN TÁN

2.3.1 Các bước định giá câu hỏi phân tán

Trước tiên ta sẽ làm rõ một số khái niệm:

- Câu hỏi phân tán là câu hỏi được diễn tả phù hợp với các dữ liệu được mô tả bởi lược đồ tổng thể của CSDL phân tán (Định nghĩa 1.3). Các câu hỏi phân tán thường được đưa ra ở mức lược đồ ngoài tổng thể là các khung nhìn.
- Định giá một câu hỏi phân tán là quá trình 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 một phương án thực thi phân tán) để cho các kết quả phù hợp với yêu cầu của câu hỏi ban đầu.

2.3.1.1 Định giá câu hỏi trong CSDL tập trung:

Trong CSDL tập trung thì việc định giá câu hỏi được chia làm 3 giai đoạn và thực thi trên một CSDL duy nhất.

- Giai đoạn 1: Sự phân tách.

Giai đoạn này chuyển câu hỏi được diễn đạt bởi ngôn ngữ bậc cao (chẳng hạn SQL hay QUEL) thành 1 câu hỏi được diễn đạt trong 1 biểu thức của đại số quan hệ. Biểu thức này được gọi là câu hỏi chính tắc. Kết quả của giai đoạn này là một biểu thức hỏi biểu diễn bằng đại số quan hệ, được gọi là *câu hỏi chính tắc*.

- Giai đoạn 2: Sự tối ưu hóa.

Sự tối ưu ở đây được hiểu là nhằm làm cực tiểu một hàm chi phí nào đó (chẳng hạn về thời gian tính toán, chi phí truyền dữ liệu...). Giai đoạn này sẽ biến đổi câu hỏi chính tắc nhận được sau giai đoạn phân tách, thành các câu hỏi tương đương ngắn gọn hơn, và xác định một trình tự tính toán tối ưu để thực hiện các phép toán trong biểu thức đại số quan hệ, cũng như xác định các thuật toán tối ưu để truy cập, truyền tải dữ liệu. Kết quả của giai đoạn này là đưa ra một biểu thức đại số quan hệ tương đương với câu hỏi chính tắc, nhưng tối ưu hóa việc định giá câu hỏi.

- Giai đoạn 3: Thực thi (thực hiện các phép toán trên CSDL tập trung)

Giai đoạn này sẽ thực hiện các phép toán của câu hỏi đã tối ưu hóa để nhận được kết quả.

2.3.1.2 Định giá câu hỏi trong CSDL phân tán:

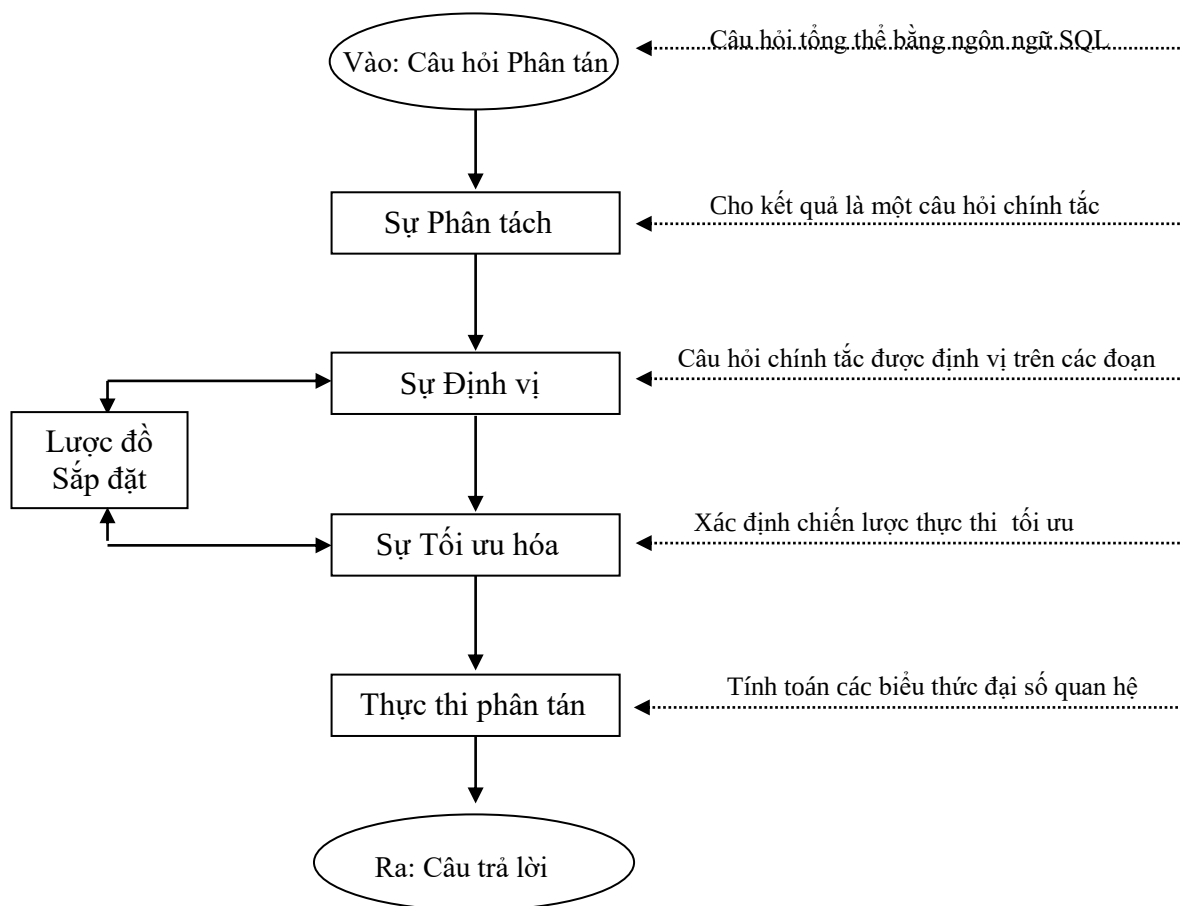
Đối với CSDL phân tán thì việc định giá câu hỏi cũng tương tự như trên, nhưng sau giai đoạn phân tách phải thêm một giai đoạn định vị, để biến đổi câu hỏi chính tắc (câu hỏi phân tán) thành một câu hỏi chính tắc tương đương được diễn đạt trên các đoạn, dựa trên các thông tin của lược đồ sắp đặt. Như vậy, việc định giá câu hỏi phân tán sẽ gồm 4 giai đoạn:

Giai đoạn 1: Sự phân tách

Giai đoạn 2: Sự định vị

Giai đoạn 3: Sự tối ưu hóa

Giai đoạn 4: Thực thi phân tán (thực hiện các phép toán trên các trạm)



Hình 2.1. Các bước định giá một câu hỏi phân tán

Giai đoạn phân tách (giai đoạn 1) được thực hiện hoàn toàn giống như việc định giá các câu hỏi tròn CSDL tập trung, tức là từ câu hỏi ban đầu, cần xác định một biểu thức hỏi biểu diễn bằng đại số quan hệ.

Giai đoạn thực thi phân tán cũng giống như trong CSDL tập trung, chỉ khác là phải tính toán các biểu thức đại số quan hệ trên các trạm (các CSDL cục bộ).

Dưới đây, chúng ta chỉ tập trung phân tích 2 giai đoạn định vị và tối ưu hóa vì đây là hai giai đoạn cần có những xử lý đặc biệt đối với CSDL phân tán.

2.3.2 Định vị câu hỏi phân tán

Từ câu hỏi tổng thể được cho ở ngôn ngữ bậc cao (chẳng hạn SQL), sau giai đoạn phân tách ta nhận được một câu hỏi chính tắc (ta cũng gọi câu hỏi chính tắc này là câu hỏi phân tán), và giai đoạn tiếp theo là định vị câu hỏi, đó là 1 quá trình gồm 2 bước: diễn đạt câu hỏi chính tắc trên các đoạn và sau đó là giản lược câu hỏi này.

Thông tin cần thiết cho sự định vị câu hỏi phân tán được lưu trữ trong lược đồ sắp đặt.

2.3.2.1 Diễn đạt câu hỏi chính tắc trên các đoạn.

Câu hỏi chính tắc được diễn đạt trên các đoạn, là một biểu thức đại số quan hệ tương đương với câu hỏi chính tắc ban đầu, nhận được bằng việc thay thế mỗi quan hệ tổng thể trong

câu hỏi phân tán bằng sự khôi phục quan hệ đó từ các đoạn của nó bằng một phép toán đại số quan hệ xác định sự khôi phục. (đôi khi gọi phép toán này là *yêu cầu khôi phục*)

Thí dụ 2.7. Giả sử có các quan hệ R_VANG và T_THU trong các thí dụ trên đây. Hãy xét câu hỏi phân tán:

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

Rõ ràng câu hỏi này viện dẫn tới các quan hệ R_VANG và T_THU đã biết, và giả sử các quan hệ này đã được phân đoạn:

$$VANG1 = \Pi_{MARV, V.NHO, N.HIEU} (R_VANG),$$

$$VANG2 = \Pi_{MARV, D.RUOU, GIA} (R_VANG),$$

$$T_THU1 = \sigma_{D.DIEM = HANOI} (T_THU),$$

$$T_THU2 = \sigma_{D.DIEM \neq HANOI} (T_THU).$$

Câu hỏi chính tắc được xác định như sau:

$$\Pi_{V.NHO} (\sigma_{D.DIEM = HANOI} (R_VANG \bowtie T_THU)) \quad (1)$$

Do phép chiếu chỉ liên quan đến quan hệ R_VANG, còn phép chọn chỉ liên quan đến quan hệ T_THU, cho nên biểu thức (1) có thể viết lại là:

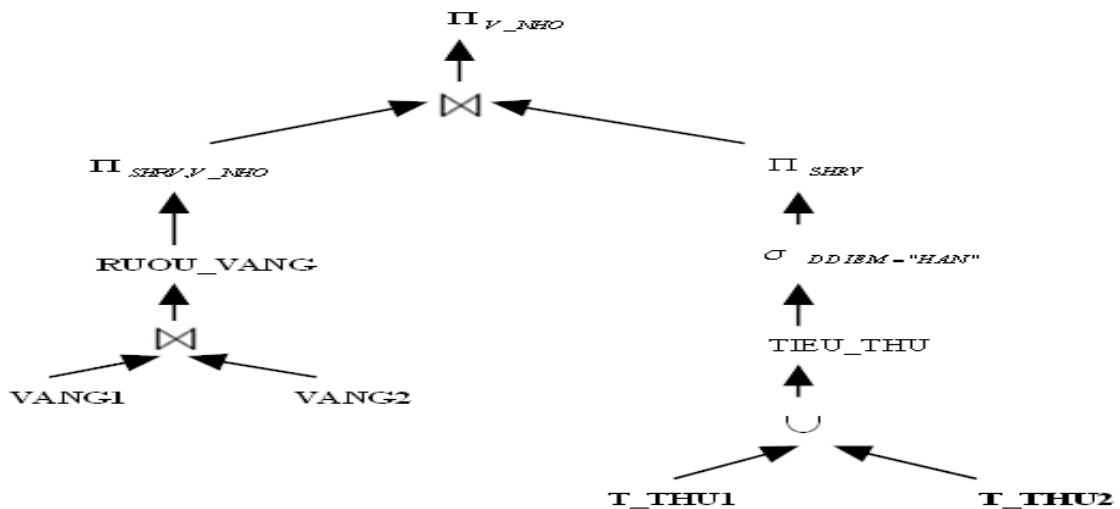
$$\Pi_{V.NHO} \{ [\Pi_{MARV, V.NHO} (R_VANG)] \bowtie [\Pi_{MARV} (\sigma_{D.DIEM = HANOI} (T_THU))] \} \quad (2)$$

Biểu thức (2) là câu hỏi chính tắc nhận được từ câu hỏi phân tán (*) sau giai đoạn phân tách. Để diễn đạt câu hỏi này, ta chỉ cần khôi phục các quan hệ tổng thể từ các đoạn của nó, tức là trong (2) ta thay: $R_VANG := VANG1 \bowtie VANG2$ và thay $T_THU := T_THU1 \cup T_THU2$.

Khi đó ta có câu hỏi chính tắc được diễn đạt trên các đoạn là:

$$\Pi_{V.NHO} \{ [\Pi_{MARV, V.NHO} (VANG1 \bowtie VANG2)] \bowtie [\Pi_{MARV} (\sigma_{D.DIEM = HANOI} (T_THU1 \cup T_THU2))] \} \quad (3)$$

Biểu thức (3) là câu hỏi chính tắc đã được diễn đạt trên các đoạn, có thể được biểu diễn bởi cây đại số quan hệ sau (cây T1)



Hình 2.2. Cây đại số quan hệ T1 của câu hỏi chính tắc (3).

2.3.2.2 Giảm lược câu hỏi:

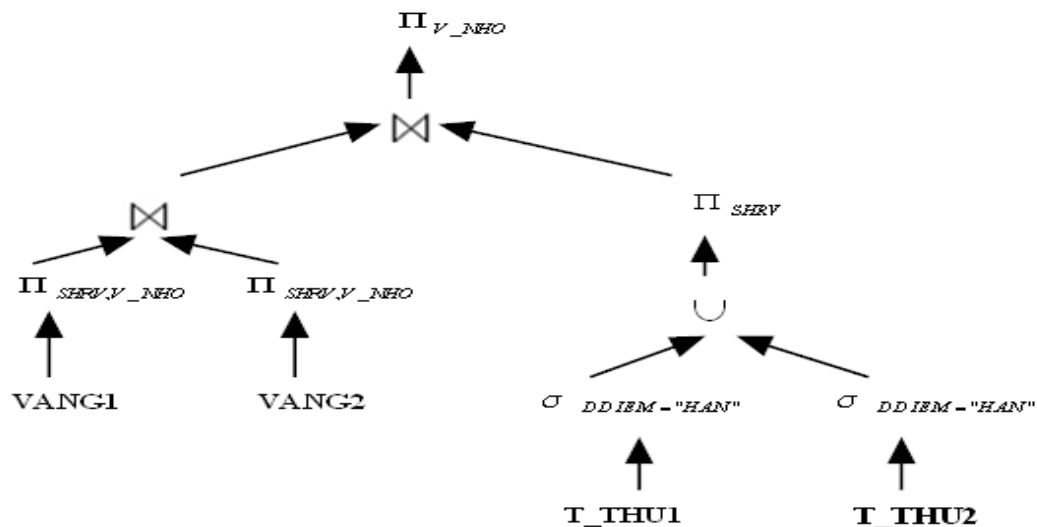
Câu hỏi chính tắc sau khi định vị là một biểu thức đại số quan hệ, có thể thực hiện biểu thức này trên các trạm. Tuy nhiên, nói chung caauhoir chính tắc sau khi định vị có thể chứa đựng các phép toán vô ích và có thể giảm lược nhờ các quy tắc trong đại số quan hệ, từ đó có thể cấu trúc lại cây đại số quan hệ, để được các biểu thức tương đương nhưng đơn giản hơn.

Các quy tắc giảm lược cho các biểu thức đại số 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 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 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 cho kết quả là rỗng).

Các quy tắc trên dựa trên tính giao hoán của phép chiếu với phép kết nối, tính giao hoán của phép chọn và phép hợp..đã trình bày ở phần trên,

Chẳng hạn sau khi thực hiện quy tắc giảm lược 1 cho câu hỏi (3): đưa các phép chiếu và chọn xuống càng thấp càng tốt, ta có cây đại số quan hệ T2 tương đương với T1:

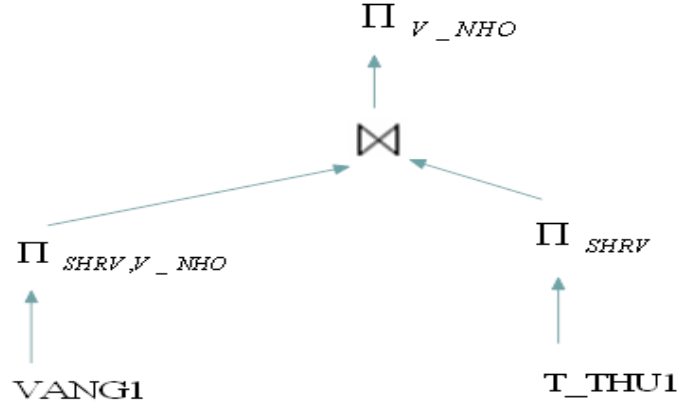


Hình 2.3. Cây đại số quan hệ T2 tương đương cây T1(của câu hỏi chính tắc (3)).

Tiếp tục áp dụng các quy tắc giảm lược:

- Do $\Pi_{SHRV,V_NHO}(VANG2) = \phi \Rightarrow$ Ta loại bỏ cây con tương ứng của nó
- Do $\sigma_{DDIEM \neq "HANOI"}(T_THU2) = \phi \Rightarrow$ Ta loại bỏ cây con tương ứng của nó.

Cuối cùng, sau các phép giảm lược, ta nhận được cây đại số quan hệ T3 của câu hỏi phân tán đã được định vị và giảm lược là:



Hình 2.4. Cây đại số quan hệ T3 của câu hỏi chính tắc đã định vị và giản lược.

Từ cây T3 (tương đương T2 và T1), ta xác định được câu hỏi chính tắc đã được định vị và giản lược:

$$\Pi_{V_NHO} \{ \Pi_{MARV, VNHO} (VANG1) \bowtie \Pi_{MARV} (T..THU1) \} \quad (4)$$

2.3.3 Tối ưu hóa câu hỏi phân tán

Sau giai đoạn định vị, câu hỏi phân tán (là một biểu thức đại số quan hệ, hay còn gọi là câu hỏi chính tắc) đã được diễn đạt trên các đoạn, tức là biểu thức đại số quan hệ chỉ chứa các quan hệ con là các đoạn của các quan hệ tổng thể được đặt trên các trạm, vì vậy việc tính các biểu thức đại số quan hệ này là có thể thực hiện được từ các CSDL cục bộ. Tuy nhiên, có nhiều cách tính tương đương cho mỗi biểu thức đại số quan hệ, mỗi cách tính đó gọi là một chiến lược thực thi câu hỏi phân tán.

Khái niệm: **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.

Với mỗi câu hỏi phân tán, sẽ có rất nhiều chiến lược thực thi câu hỏi, bài toán chọn một chiến lược tối ưu là bài toán thuộc lớp NPC (NP- đầy đủ), nói chung rất khó để chọn được một chiến lược tối ưu hóa đúng tuyệt đối. Cho nên mục tiêu của tối ưu hóa câu hỏi phân tán là lựa chọn một chiến lược gần với tối ưu hóa, để tránh các chiến lược dẫn đến một chi phí quá lớn.

Để đơn giản, ta chọn hàm chi phí là thời gian thực thi câu hỏi, và cũng hạn chế các phép toán trong câu hỏi chỉ là ‘phép chiếu (P)’, ‘phép chọn (S)’ và phép ‘kết nối (J)’. Các biểu thức đại số quan hệ như vậy thường gọi là ‘biểu thức P-S-J’.

Giả sử rằng thời gian tính toán với các quan hệ trên cùng một trạm là không đáng kể. (Với các phép chiếu và chọn, là các phép toán ‘1 ngôi’, chỉ thực hiện trên 1 quan hệ con (một đoạn) trên 1 trạm, nên thời gian là không đáng kể). Như vậy thời gian tính toán chủ yếu lại là thời gian truyền dữ liệu giữa các trạm để thực hiện các phép kết nối. Thời gian truyền dữ liệu giữa các trạm tỷ lệ với kích thước các quan hệ (tính bằng số phần tử dữ liệu trong mỗi quan hệ) tham gia vào phép kết nối. Vì vậy từ nay ta gọi chi phí cần tối ưu là chi phí truyền tải dữ liệu giữa các trạm, chi phí này tỷ lệ với số phần tử dữ liệu cần truyền tải giữa các trạm.

Với những giả thiết và phân tích trên đây, việc tối ưu hóa câu hỏi phân tán sẽ là tối ưu hóa việc tính các kết nối trong quá trình thực thi phân tán. Giả sử rằng cần tính kết nối của hai quan hệ (kết nối của nhiều quan hệ cũng phải tính lần lượt kết nối của từng cặp quan hệ), mỗi quan hệ có thể ở mỗi trạm khác nhau và trước khi tính ta phải chuyển 1 quan hệ về trạm của quan hệ kia để tính. Như vậy chiến lược tối ưu câu hỏi phân tán là một chiến lược thực thi câu hỏi sao cho chi phí truyền dữ liệu giữa các trạm là nhỏ nhất.

Ta sẽ lần lượt xét 2 vấn đề cần thực hiện để giảm chi phí truyền dữ liệu:

1. Xác định một trình tự kết nối tối ưu của nhiều quan hệ
2. Rút gọn kích thước các quan hệ trước khi kết nối.

2.3.3.1 Xác định trình tự kết nối tối ưu.

Giả sử ta phải tính kết nối của các quan hệ r_1, r_2, r_3 và r_4 bằng một dãy các kết nối từng cặp. Trước tiên ta có thể tính $r_1 \bowtie r_2$ để nhận được một quan hệ, giả sử là có năm trăm bộ. Sau đó ta kết nối quan hệ trên với r_3 và nhận được một quan hệ mới với một ngàn bộ, cuối cùng ta kết nối quan hệ gồm một ngàn bộ này với r_4 và nhận được kết quả cuối cùng chỉ có 100 bộ. Trình tự kết nối trên đây được xác định bởi biểu thức: $((r_1 \bowtie r_2) \bowtie r_3) \bowtie r_4$ (*)

Rõ ràng trình tự kết nối này là chưa tốt, vì có các kết quả trung gian có số bộ lớn hơn số bộ ở kết quả cuối cùng. Điều này sẽ lãng phí công sức để tính toán, lưu trữ và truyền tải các quan hệ rất lớn trong quá trình tính toán, mà kết quả lại là một quan hệ có số bộ nhỏ hơn nhiều.

Với 4 quan hệ trên có thể có nhiều trình tự kết nối khác, chẳng hạn $(r_1 \bowtie r_2) \bowtie (r_3 \bowtie r_4)$, và v.v... Do phép kết nối có tính giao hoán, nên dù kết nối theo trình tự nào thì việc kết nối 4 quan hệ trên vẫn cho kết quả là một quan hệ duy nhất với 100 bộ, vậy vấn đề đặt ra là có thể tìm một trình tự kết nối tối ưu, trong đó mọi quan hệ kết quả trung gian có số bộ không lớn hơn số bộ ở kết quả cuối cùng.

Định nghĩa 2.5

Một trình tự kết nối tối ưu cho các quan hệ $r_1, r_2, \dots, r_k$ 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.

Việc xác định một trình tự kết nối tối ưu là khá phức tạp, chúng ta chưa trình bày trong giáo trình này, mà chỉ minh họa vấn đề này bằng một thí dụ.

Thí dụ 2.7. Giả sử cần tính kết nối : $r_1 \bowtie r_2 \bowtie r_3$, với các quan hệ được cho như sau:

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

Ta có thể tính dãy kết nối trên theo nhiều thứ tự, chẳng hạn: $(r_1 \bowtie r_2) \bowtie r_3$, $r_1 \bowtie (r_2 \bowtie r_3)$, $(r_1 \bowtie r_3) \bowtie r_2$, ... có tất cả $3! = 6$ trình tự kết nối. Do phép kết nối có tính giao hoán, nên mọi trình tự kết nối đều cho kết quả như nhau, vấn đề là ta cần chọn một trình tự kết nối tối ưu, tức là trình tự kết nối sao cho số bộ ở mọi kết quả trung gian không lớn hơn số bộ ở kết quả cuối cùng.

Với trình tự 1: ta thực hiện phép kết nối trên theo thứ tự: $(r_1 \bowtie r_2) \bowtie r_3$,

- *bước 1*: tính $r_1 \bowtie r_2$, cho kết quả kết nối của 2 quan hệ (trong goặc đơn):

$r_1 \bowtie r_2$	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

Ta nhận được kết quả kết quả trung gian là một quan hệ có 10 bộ.

- *bước 2*: tính $(r_1 \bowtie r_2) \bowtie r_3$, cho kết quả kết nối của 3 quan hệ (kết quả cuối cùng):

$(r_1 \bowtie r_2) \bowtie r_3$	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

Theo trình tự kết nối này, ta có số bộ ở kết quả trung gian là 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.

Với trình tự 2: ta thực hiện phép kết nối trên theo thứ tự: $r_1 \bowtie (r_2 \bowtie r_3)$,

- *bước 1*: tính $r_2 \bowtie r_3$ cho kết quả kết nối của 2 quan hệ (trong goặc đơn):

$r_2 \bowtie r_3$	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

Đây là kết quả của kết nối $r_2 \bowtie r_3$, cho một quan hệ trung gian gồm 4 bộ.

- bước 2 : tính $r_1 \bowtie (r_2 \bowtie r_3)$: cho kết quả kết nối của 3 quan hệ (kết quả cuối cùng):

$r_1 \bowtie (r_2 \bowtie r_3)$	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

Theo trình tự kết nối này, ta có số bộ ở kết quả trung gian là 4 bộ, số bộ ở kết quả cuối cùng là 6 bộ, vậy số bộ ở kết quả trung gian không lớn hơn số bộ ở kết quả cuối cùng. Vậy đây là trình tự kết nối tối ưu.

2.3.3.2 Rút gọn các quan hệ nhờ phép nửa kết nối.

Một phép kết nối có thể được thực hiện bằng một tổ hợp các phép nửa kết nối và kết nối. Áp dụng các công thức tương đương trong đại số quan hệ, ta có

- $r \bowtie s = (r \bowtie s) \bowtie s$ (hoặc $r \bowtie s = (s \bowtie r) \bowtie r$),

trong đó có thể tính các nửa kết nối qua phép kết nối và phép chiếu:

- $r \bowtie s = r \bowtie \Pi_{R \cap S}(s)$ (hoặc $s \bowtie r = s \bowtie \Pi_{R \cap S}(r)$),

với $\Pi_{R \cap S}(s)$ là chiếu của quan hệ s trên tập thuộc tính chung của r và s, còn $\Pi_{R \cap S}(r)$ là chiếu của quan hệ r trên tập thuộc tính chung của r và s.

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. Ta có thể tính $r \bowtie s$ bằng 2 cách:

Cách 1. Kết nối trực tiếp: 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.

Chi phí cho phép kết nối sẽ tỷ lệ với số phần tử cần truyền tải, để đơn giản ta coi chi phí bằng số phần tử này.

- Vậy chi phí theo cách kết nối trực tiếp là :

$$\text{cost}_{tt}(r, s) = \min \{T(r), T(s)\} \quad (1)$$

Cách 2. Chiến lược nửa kết nối: Áp dụng tính kết nối qua phép nửa kết nối (còn gọi là kết nối theo chiến lược nửa kết nối), giả sử ta áp dụng công thức $r \bowtie s = (r \bowtie s) \bowtie s$, ta thực hiện các bước sau:

1. Tính $\Pi_{R \cap S}(s)$ tại trạm chứa quan hệ s, không phải truyền dữ liệu, chi phí = 0.
2. Chuyển $\Pi_{R \cap S}(s)$ sang trạm chứa r, chi phí là $T'(s)$, số phần tử dữ liệu của $\Pi_{R \cap S}(s)$.
3. Tính $r \bowtie s = r \bowtie \Pi_{R \cap S}(s)$ tại trạm chứa r, chi phí = 0.
4. Chuyển $r \bowtie s$ vừa tính được sang trạm chứa s, chi phí là $T''(r)$, số phần tử của $r \bowtie s$.
5. Tính kết nối tại vị trí của s theo công thức: $r \bowtie s = (r \bowtie s) \bowtie s$, chi phí = 0.

Trong 5 bước trên, chi phí truyền dữ liệu mà ta quan tâm chỉ phát sinh ở bước 2 và bước 4, vậy tổng chi phí truyền dữ liệu theo chiến lược nửa kết nối là $T'(s) + T''(r)$ (hoặc $T'(r) + T''(s)$, nếu áp dụng công thức $r \triangleright \triangleleft s = (s \triangleright \triangleleft r) \triangleright \triangleleft r$).

- Vậy ta có chi phí theo chiến lược nửa kết nối là:

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

Vậy chiến lược nửa kết nối sẽ tốt hơn kết nối trực tiếp, nếu có $(2) < (1)$, tức là:

$$\min \{T'(s) + T''(r); T'(r) + T''(s)\} < \min \{T(r), T(s)\} \quad (3)$$

Những điều kiện làm cho bất đẳng thức (3) thỏa mãn là:

- (i) r và s có kích thước gần như nhau,

(ii) $T'(s)$ và $T''(s)$ có kích thước nhỏ hơn nhiều so với $T(s)$ (tương tự: $T'(r)$ và $T''(r)$ có kích thước nhỏ hơn nhiều so với $T(r)$).

Những điều kiện trên trong thực tế thường được thỏa mãn, chẳng hạn, do các quan hệ thường có rất ít các thuộc tính chung, do đó $T'(s)$ thường rất nhỏ so với $T(s)$, mặt khác quan hệ s thường chỉ có một số ít bộ kết nối được với r , cho nên $T''(s)$ cũng rất nhỏ so với $T(s)$

Vì vậy có thể khẳng định rằng chiến lược nửa kết nối thường có chi phí nhỏ hơn nhiều so với kết nối trực tiếp.

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

Hãy 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

a/. Kết nối trực tiếp:

Do $T(r) < T(s)$ nên ta chuyển r sang trạm chứa s và 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

Chi phí truyền dữ liệu $\text{cost}_{\text{tt}}(r, s) = \min \{T(s), T(r)\} = 15$.

b/. Sử dụng chiến lược nửa kết nối qua các bước:

- Tính $\Pi_{R \cap S}(s)$ tại trạm chứa s :

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

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

3. Tính $r \triangleright < s = r \triangleright < \Pi_{R \cap S}(s)$ tại trạm chứa r:

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

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

5. Tính kết nối tại trạm chứa s theo công thức $r \triangleright < s = (r \triangleright < s) \triangleright < s$, nhận được kết quả :

$r \triangleright < s$	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 theo nửa kết nối: $\text{cost}_{\text{nkn}}(r, s) = T'(s) + T''(r) = 8$

Nhận xét: Việc dùng chiến lược nửa kết nối sẽ làm tăng số phép toán so với các kết nối trực tiếp, tuy nhiên lại giảm đáng kể kích thước các quan hệ tham gia vào các phép kết nối. Chính vì vậy chi phí theo cách nửa kết nối thường nhỏ hơn nhiều so với chi phí theo cách kết nối trực tiếp.

Chú ý: Nếu kích thước quan hệ s nhỏ hơn kích thước quan hệ r, tức là $T(s) < T(r)$ thì ta cần chuyển s sang trạm chứa r. Khi đó ta sử dụng các công thức : $r \triangleright < s = (s \triangleright < r) \triangleright < r$ và $s \triangleright < r = s \triangleright < \Pi_{R \cap S}(r)$. Các bước thực hiện trong chiến lược nửa kết nối hoàn toàn tương tự như đã trình bày trên đây, chỉ cần đổi vai trò r và s.

BÀI TẬP CHƯƠNG 2

- Mục đích và yêu cầu của sự phân đoạn trong CSDL phân tán. Trình bày các phương pháp phân đoạn và các biểu thức khôi phục.
- Định nghĩa các phép kết nối, nửa kết nối của 2 quan hệ, cho thí dụ minh họa.. Việc thay thế phép kết nối trực tiếp bởi các các phép kết nối và nửa kết nối (theo chiến lược nửa kết nối) gồm những bước nào?
- Việc định giá các câu hỏi phân tán gồm những giai đoạn nào. Việc tối ưu hóa câu hỏi nhằm mục đích gì, và có thể được thực hiện ở các giai đoạn nào?
- Cho hai quan hệ r và s như các bảng dưới đây

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

s	B	C	D	E
4	4	6	7	1
5	5	7	7	2
4	4	8	9	3
5	5	9	9	4

- Tính các kết nối $r \bowtie s; s \bowtie r$
- Tính các nửa kết nối: $r \ltimes s; s \ltimes r$

- Cho 3 quan hệ:

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

r ₂	C	D	E
5	5	7	2
5	5	8	4
5	5	9	6
6	6	8	8

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

- Tính kết nối 3 quan hệ trên theo trình tự: $(r_1 \bowtie r_2) \bowtie r_3$
- Tính kết nối 3 quan hệ trên theo trình tự: $(r_1 \bowtie r_3) \bowtie r_2$.
- Trình tự kết nối nào tốt hơn, tại sao?

- Cho 2 quan hệ r và s như các bảng dưới đây. Tính kết nối 2 quan hệ bằng phương pháp kết nối trực tiếp và áp dụng chiến lược nửa kết nối. So sánh chi phí của 2 phương pháp:

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

s	B	C	D	E
5	5	6	7	1
4	4	7	7	1
5	5	8	9	2
4	4	9	9	2

- Bài tập trắc nghiệm: Trong các câu hỏi dưới đây, hãy chọn chỉ một phương án trả lời đúng nhất (tô đầy hình tròn ứng với mỗi phương án đã chọn):

7.1 Phát biểu nào dưới đây là đúng nhất về sự phân đoạn của CSDL phân tán

- (A) Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu không thuộc về đoạn đó, sẽ cho kết quả là chính đoạn đó.
- (B) Một phép chọn trên một đoạn ngang có điều kiện của phép chọn trùng với điều kiện của sự phân đoạn sẽ cho kết quả là rỗng.
- (C) Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu trùng với các thuộc tính của đoạn, sẽ cho kết quả là rỗng.
- (D) Tất cả các câu trả lời A , B và C đều sai.

7.2 Phát biểu nào dưới đây là đúng nhất về sự phân đoạn của CSDL phân tán.

- (A) Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu đều thuộc về đoạn đó, sẽ cho kết quả là chính đoạn đó.
- (B) Một phép chọn trên một đoạn ngang có điều kiện của phép chọn trùng với điều kiện của sự phân đoạn sẽ cho kết quả là chính đoạn đó
- (C) Tất cả các câu trả lời A và B đều sai.
- (D) Tất cả các câu trả lời A và B đều đúng.

7.3 Phát biểu nào dưới đây là đúng nhất về sự phân đoạn của CSDL phân tán

- (A) Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu không thuộc về đoạn đó, sẽ cho kết quả là chính đoạn đó.
- (B) Một phép chọn trên một đoạn ngang có điều kiện của phép chọn trùng với điều kiện của sự phân đoạn sẽ cho kết quả là chính đoạn đó
- (C) Một phép chiếu của một đoạn dọc mà tất cả các thuộc tính chiếu trùng với các thuộc tính của đoạn, sẽ cho kết quả là rỗng.
- (D) Tất cả các câu trả lời A , B và C đều sai.

7.4 Trong CSDL phân tán, sự phân đoạn ngang gián tiếp một quan hệ tổng thể được thực hiện bởi :

- (A) Phép chiếu trên quan hệ tổng thể.
- (B) Phép chọn trên quan hệ tổng thể.
- (C) Phép nối kết nối quan hệ tổng thể với các đoạn con của một quan hệ khác.
- (D) Cả phép chiếu và phép chọn trên quan hệ tổng thể.

7.5 Trong CSDL phân tán, sự phân đoạn dọc một quan hệ tổng thể được thực hiện bởi :

- (A) Phép chiếu trên quan hệ tổng thể.
- (B) Phép chọn trên quan hệ tổng thể.
- (C) Phép nối kết nối quan hệ tổng thể với các đoạn con của một quan hệ khác.
- (D) Cả phép chiếu và phép chọn trên quan hệ tổng thể.

7.6 Trong CSDL phân tán, sự khôi phục một quan hệ tổng thể từ các phân đoạn ngang được thực hiện bởi :

- (A) Phép kết nối các đoạn con.
- (B) Phép hợp các đoạn con.
- (C) Tất cả các câu trả lời A và B đều sai.
- (D) Tất cả các câu trả lời A và B đều đúng.

7.7 Trong CSDL phân tán, sự khôi phục một quan hệ tổng thể từ các phân đoạn dọc được thực hiện bởi :

- (A) Phép kết nối các đoạn con.
- (B) Phép hợp các đoạn con.
- (C) Tất cả các câu trả lời A và B đều sai.
- (D) Tất cả các câu trả lời A và B đều đúng.

- 7.8 Giả sử cần kết nối 3 quan hệ P, Q, R được đặt trên 3 trạm khác nhau:
- ☐ (A) Việc thay đổi thứ tự kết nối các quan hệ sẽ không làm thay đổi chi phí truyền dữ liệu
 - ☐ (B) Việc thay đổi thứ tự kết nối các quan hệ sẽ không làm thay đổi kết quả của phép kết nối
 - ☐ (C) Tất cả các câu trả lời A và B đều sai
 - ☐ (D) Tất cả các câu trả lời A và B đều đúng.
- 7.9 Thay thế phép kết nối bằng tổ hợp các phép nửa kết nối và kết nối (theo chiến lược nửa kết nối):
- ☐ (A) Làm tăng kích thước các quan hệ trước khi kết nối
 - ☐ (B) Làm giảm số phép toán
 - ☐ (C) Tất cả các câu trả lời A và B đều đúng
 - ☐ (D) Tất cả các câu trả lời A và B đều sai.
- 7.10 Thay thế phép kết nối bằng tổ hợp các phép nửa kết nối và kết nối (theo chiến lược nửa kết nối):
- ☐ (A) Làm giảm kích thước các quan hệ trước khi kết nối
 - ☐ (B) Làm tăng số phép toán
 - ☐ (C) Tất cả các câu trả lời A và B đều đúng
 - ☐ (D) Tất cả các câu trả lời A và B đều sai.
- 7.11 Cho R và S là 2 quan hệ có các thuộc tính chung và các thuộc tính riêng. Khi đó:
- ☐ (A) $R \bowtie S$ có số bộ nhiều hơn số bộ của mỗi quan hệ
 - ☐ (B) $R \bowtie S$ có số bộ ít hơn số bộ của mỗi quan hệ
 - ☐ (C) $R \bowtie S$ có số thuộc tính nhiều hơn số thuộc tính của mỗi quan hệ R hoặc S
 - ☐ (D) $R \bowtie S$ có số thuộc tính ít hơn số thuộc tính của mỗi quan hệ R hoặc S.
- 7.12 Cho R và S là các quan hệ, khi đó:
- ☐ (A) $R \bowtie S = (R \bowtie S) \bowtie S$
 - ☐ (B) $R \bowtie S = (S \bowtie \bigcap_{R \cap S} R) \bowtie R$
 - ☐ (C) Tất cả các câu trả lời A và B đều sai
 - ☐ (D) Tất cả các câu trả lời A và B đều đúng.
- 7.13 Cho R là quan hệ không kết nối hoàn toàn được với quan hệ S, khi đó :
- ☐ (A) Nửa kết nối $R \bowtie S$ sẽ loại bỏ một số hàng của S..
 - ☐ (B) Nửa kết nối $R \bowtie S$ sẽ loại bỏ một số hàng của R
 - ☐ (C) Nửa kết nối $R \bowtie S$ sẽ làm tăng số thuộc tính của S.
 - ☐ (D) Nửa kết nối $R \bowtie S$ sẽ làm tăng số thuộc tính của R.