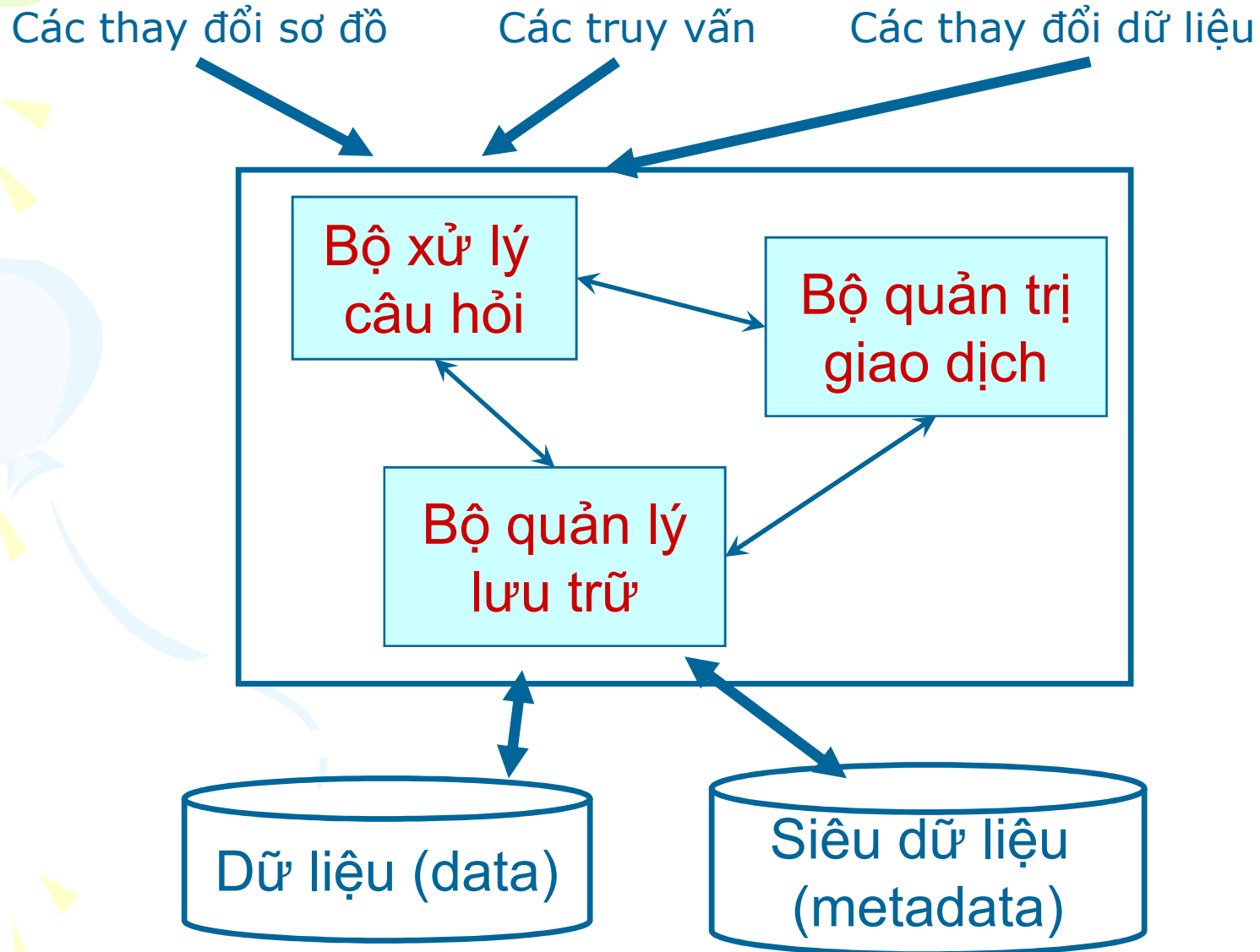


Tối ưu hóa câu truy vấn

Đặt vấn đề

Kiến trúc của một hệ quản trị CSDL



Đặt vấn đề

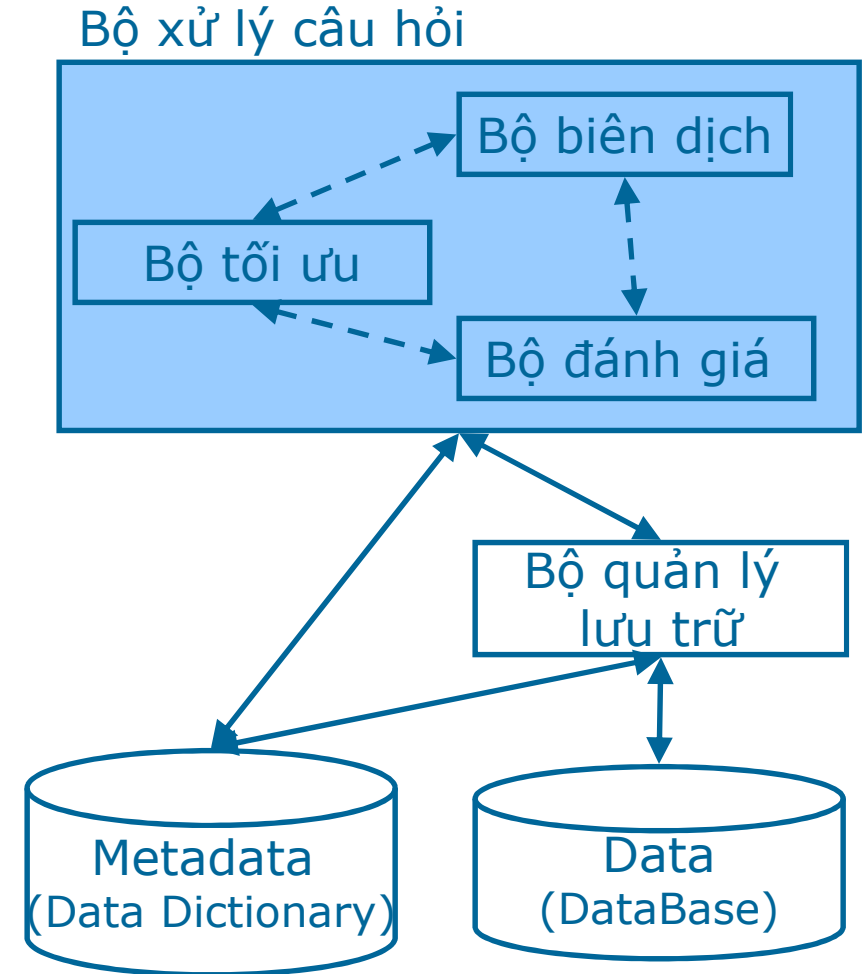
Xử lý câu hỏi

- Yêu cầu

- Nhận, phân tích yêu cầu truy vấn.
- Tìm kiếm dữ liệu trả lời cho một yêu cầu truy vấn.

- Thực hiện

- Phân tích cú pháp câu truy vấn.
- Biến đổi truy vấn ở mức cao thành các yêu cầu có thể hiểu được và xử lý được bởi máy tính.
- Lựa chọn một kế hoạch thực hiện tốt nhất để trả lời truy vấn này.



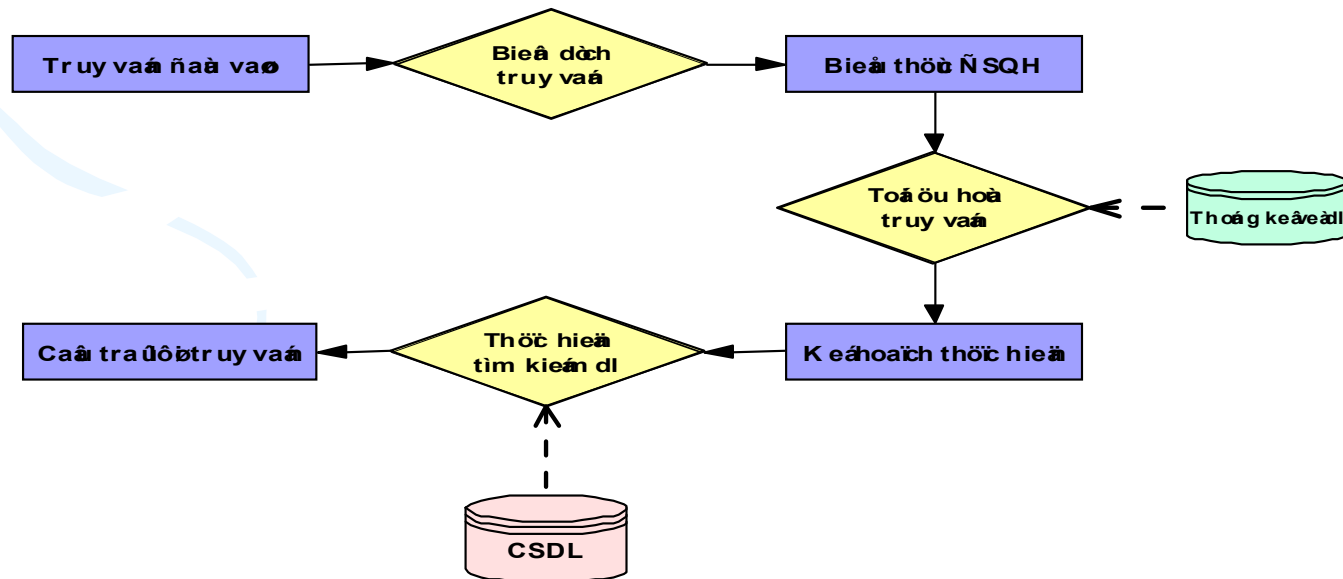


Nội dung

- Tổng quan về xử lý truy vấn
- Tối ưu hóa các biểu thức đại số quan hệ
- Cài đặt các phép toán và mô hình chi phí

Tổng quan về xử lý truy vấn

- Xử lý một truy vấn bao gồm 3 bước chính:
 - Phân tích và Biên dịch câu truy vấn
 - Tối ưu hóa câu truy vấn
 - Thực hiện đánh giá truy vấn



Phân tích và Biên dịch câu truy vấn

Trong bước này, bộ xử lý câu hỏi phải nhận, phân tích cú pháp và dịch câu truy vấn từ dạng ngôn ngữ bậc cao thành một ngôn ngữ biểu diễn dữ liệu bên trong để máy tính có thể thao tác trên đó. Dạng biểu diễn bên trong thích hợp và hỗ trợ cho bước tối ưu hóa tiếp theo là biểu diễn bằng ngôn ngữ đại số quan hệ

- Kiểm tra cú pháp câu truy vấn → xây dựng cây cú pháp
- Dịch cây cú pháp của câu truy vấn thành biểu thức đại số quan hệ → xây dựng biểu thức đại số quan hệ
- Thay thế các khung nhìn bởi biểu thức đại số quan hệ xác định khung nhìn (nếu truy vấn biểu diễn thông qua các khung nhìn)

Tối ưu hóa câu truy vấn

Tối ưu hóa câu truy vấn là quá trình lựa chọn kế hoạch thực hiện hiệu quả nhất có thể đối với câu truy vấn, có nghĩa là kế hoạch thực hiện có chi phí thấp nhất.

- Biến đổi 1 biểu thức ĐSQH đầu vào thành một biểu thức ĐSQH tương đương nhưng được thực hiện hiệu quả hơn. Bước con đầu tiên này được gọi là tối ưu hóa đại số.
- Xác định kế hoạch thực hiện hay chiến lược chi tiết đối với việc xử lý câu truy vấn, cụ thể phải lựa chọn thuật toán để thực thi mỗi phép toán, lựa chọn các chỉ dẫn đặc biệt được sử dụng với mỗi toán hạng của phép toán.

Tối ưu hóa câu truy vấn (tiếp)

- Đánh giá chi phí của mỗi kế hoạch thực hiện để lựa chọn một kế hoạch ít tốn kém nhất. Việc tính toán chi phí một cách chính xác của 1 kế hoạch là không thể nếu không thực sự đánh giá kế hoạch này. Thực tế, bộ tối ưu phải sử dụng các thông tin thống kê về CSDL được lưu trữ trong Thư mục của DBMS và các độ đo của chi phí truy vấn để ước lượng chi phí.

Thực hiện đánh giá truy vấn

Cuối cùng, từ một kế hoạch thực hiện có được do Trình tối ưu hóa cung cấp, truy vấn được đánh giá với kế hoạch này và đưa ra câu trả lời cho truy vấn đó.

➤ Đánh giá các Biểu thức ĐSQH

Trong chương này, chúng ta sẽ tập trung vào việc tối ưu hóa câu truy vấn bao gồm tối ưu biểu thức đại số quan hệ cũng như cài đặt các phép toán và mô hình chi phí để thực hiện đánh giá một truy vấn.

Đánh giá biểu thức ĐSQH

- Sau các bước phân tích & biên dịch và tối ưu hóa, một truy vấn được biểu diễn bởi một biểu thức ĐSQH. Biểu thức ĐSQH đối với các truy vấn phức tạp thường bao gồm nhiều phép toán và tác động lên nhiều quan hệ khác nhau. Việc thực hiện tính toán kết quả của biểu thức ĐSQH phức tạp hay việc đánh giá biểu thức này thường rất tốn kém thời gian và bộ nhớ.
- Có 2 hướng tiếp cận đối với việc đánh giá biểu thức ĐSQH:
 - Vật chất hóa (Materialize)
 - Đường ống (Pipeline)

Đánh giá biểu thức ĐSQH (tiếp)

- Vật chất hóa: Trong cách tiếp cận này, chúng ta lần lượt đánh giá các phép toán theo một thứ tự thích hợp. Kết quả của việc đánh giá mỗi phép toán sẽ được lưu trong một quan hệ trung gian tạm thời để sử dụng làm đầu vào cho các phép toán tiếp theo.
- Điểm bất lợi của cách tiếp cận này là việc cần thiết phải xây dựng các quan hệ trung gian tạm thời nhất là khi các quan hệ này thường phải được ghi ra đĩa (trừ khi chúng có kích thước rất nhỏ). Chi phí đối với các thao tác đọc và ghi ra đĩa là khá lớn.

Đánh giá biểu thức ĐSQH (tiếp)

- Đường ống: Chúng ta có thể cải thiện hiệu quả đánh giá truy vấn bằng cách làm giảm bớt số lượng các quan hệ trung gian tạm thời được tạo ra. Điều này có thể đạt được nhờ việc kết hợp một vài phép toán quan hệ vào một đường ống của các phép toán. Trong đường ống thì kết quả của một phép toán được chuyển trực tiếp cho phép toán tiếp theo mà không cần phải lưu lại trong quan hệ trung gian.
- Rõ ràng, cách tiếp cận thứ hai sẽ hạn chế được nhược điểm của cách tiếp cận đầu tiên, nhưng có những trường hợp, ta bắt buộc phải vật chất hóa chứ không dùng đường ống được.

Đánh giá biểu thức ĐSQH (tiếp)

- Ví dụ: Chúng ta có một biểu thức đại số quan hệ gồm 2 phép toán: kết nối và chiếu.
- Trong cách tiếp cận vật chất hóa, xuất phát từ phép toán ở mức thấp nhất là phép kết nối tự nhiên, kết quả của phép kết nối này sẽ được lưu trong một quan hệ trung gian. Sau đó, đọc từ quan hệ trung gian này để tiến hành chiếu lấy kết quả mong muốn.
- Trong cách tiếp cận đường ống, khi một bộ được sinh ra trong phép kết nối 2 quan hệ, bộ này sẽ được chuyển trực tiếp đến phép chiếu để xử lý và kết quả được ghi vào quan hệ đầu ra. Quan hệ kết quả sẽ được tạo lập một cách trực tiếp.

Tối ưu hóa các biểu thức ĐSQH

- Mục tiêu là tổ chức lại trình tự thực hiện các phép toán trong biểu thức để giảm chi phí thực hiện đánh giá biểu thức đó.
- Trong quá trình tối ưu hóa, ta biểu diễn một biểu thức ĐSQH dưới dạng một cây toán tử. Trong cây toán tử, các nút lá là các quan hệ trong biểu thức, các nút trong là các phép toán trong biểu thức
- Ví dụ : Đưa ra tên hãng cung ứng mặt hàng có mã là 'P1':

```
Select sname From S, SP Where S.sid = SP.sid And pid = 'P1'
```
- Biểu thức ĐSQH tương ứng là?
- Cây toán tử tương ứng là?

Các chiến lược tối ưu tổng quát

1. Đẩy phép chọn và phép chiếu xuống thực hiện sớm nhất có thể: vì hai phép toán này giúp làm giảm kích thước của quan hệ trước khi thực hiện các phép toán 2 ngôi
2. Nhóm dãy các phép chọn và chiếu: Sử dụng chiến lược này nếu như có một dãy các phép chọn hoặc dãy các phép chiếu trên cùng một quan hệ
3. Kết hợp phép chọn và tích Đề các thành phép kết nối: Nếu kết quả của một phép tích Đề các là đối số của 1 phép chọn có điều kiện chọn là phép so sánh giữa các thuộc tính trên 2 quan hệ tham gia tích Đề các thì ta nên kết hợp 2 phép toán thành phép kết nối.
4. Tìm các biểu thức con chung trong biểu thức đại số quan hệ để đánh giá chỉ một lần

Các chiến lược tối ưu tổng quát (tiếp)

5. Xác định các phép toán có thể được đưa vào đường ống và thực hiện đánh giá chúng theo đường ống
6. Xử lý các tệp dữ liệu trước khi tiến hành tính toán: Tạo lập chỉ dẫn hay sắp xếp tệp dữ liệu có thể góp phần làm giảm chi phí của các phép tính trung gian
7. Ước lượng chi phí và lựa chọn thứ tự thực hiện: Do với mỗi câu truy vấn có thể có nhiều cách khác nhau để thực hiện, với việc ước lượng chi phí (số phép tính, tài nguyên sử dụng, dung tích bộ nhớ, thời gian thực hiện ..) ta có thể chọn cách đánh giá biểu thức ĐSQH có chi phí nhỏ nhất.

Các phép biến đổi tương đương biểu thức ĐSQH

- Hai biểu thức ĐSQH E_1 và E_2 là tương đương nếu chúng cho cùng một kết quả khi áp dụng trên cùng một tập các quan hệ
- Trong phần này, ta có các ký hiệu dạng sau:
 - E_1, E_2, E_3, \dots là các biểu thức đại số quan hệ
 - F_1, F_2, F_3, \dots là các điều kiện chọn hoặc là các điều kiện kết nối
 - $X_1, X_2, \dots Y, Z, U_1, U_2, \dots$ là các tập thuộc tính

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

1. Quy tắc kết hợp của phép tích Đề các và kết nối

$$(E_1 \times E_2) \times E_3 \equiv E_1 \times (E_2 \times E_3)$$

$$(E_1 * E_2) * E_3 \equiv E_1 * (E_2 * E_3)$$

$$(E_1 \triangleright_{F1} \triangleleft E_2) \triangleright_{F2} E_3 \equiv E_1 \triangleright_{F1} (\triangleleft_{F2} E_2 \triangleright E_3)$$

- Quy tắc này sử dụng cho chiến lược số 7. Thứ tự thực hiện các phép kết nối hay tích Đề các là rất quan trọng vì kích thước của quan hệ trung gian có thể rất lớn nếu không cân nhắc kỹ. Lựa chọn thứ tự thực hiện các phép toán này thì tùy thuộc vào kích thước của các quan hệ tham gia phép toán và cả ngữ nghĩa của quan hệ (mỗi liên hệ)

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

- VD: $S * SP * P$ có thể được thực hiện theo 3 thứ tự như sau

1) $(S * SP) * P$

2) $(S * P) * SP$

3) $S * (SP * P)$

Xét theo ngữ nghĩa S, P không kết nối được nên (1) và (3) là tốt hơn (2). Xét về kích thước thì (3) tốt hơn (1) vì S có 4 thuộc tính còn P có 3 thuộc tính, tuy nhiên, cũng còn tùy thuộc vào lực lượng của 2 quan hệ S và P nữa

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

2. Quy tắc giao hoán trong phép tích Đề các và kết nối

$$E_1 \times E_2 \equiv E_2 \times E_1$$

$$E_1 * E_2 \equiv E_2 * E_1$$

$$E_1 \triangleright_F \triangleleft E_2 \equiv E_2 \triangleright_F \triangleleft E_1$$

3. Quy tắc đối với dãy các phép chiếu

$$\Pi_{X_1}(\Pi_{X_2} \dots \Pi_{X_n}(E) \dots) \equiv \Pi_{X_1}(E)$$

$$X_1 \subseteq X_2 \subseteq \dots \subseteq X_n$$

4. Quy tắc đối với dãy các phép chọn

$$\sigma_{F_1}(\sigma_{F_2} \dots \sigma_{F_n}(E) \dots) \equiv \sigma_{F_1 \wedge F_2 \wedge \dots \wedge F_n}(E)$$

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

5. Quy tắc giao hoán phép chọn và phép chiếu

$$\Pi_X(\sigma_F(E)) \equiv \sigma_F(\Pi_X(E))$$

Quy tắc này áp dụng khi F là điều kiện xác định được trên tập thuộc tính X . Tổng quát hơn ta có:

$$\Pi_X(\sigma_F(E)) \equiv \Pi_X(\sigma_F(\Pi_{XY}(E)))$$

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

6. Quy tắc đối với phép chọn và phép tích Đề các

- Ta ký hiệu:

- $E_1(U_1)$ có nghĩa là biểu thức E_1 xác định trên tập thuộc tính U_1
- $F_1(U_1)$ có nghĩa là điều kiện chọn F_1 xác định trên tập thuộc tính U_1
- Quy tắc biến đổi liên quan đến phép chọn và tích Đề các được phát biểu như sau:

- $\sigma_F(E_1(U_1) \times E_2(U_2))$ tương đương với:

- $\sigma_{F_1}(E_1) \times E_2$ trong trường hợp $F = F_1(U_1)$
- $\sigma_{F_1}(E_1) \times \sigma_{F_2}(E_2)$ trong trường hợp $F = F_1(U_1) \times F_2(U_2)$
- $\sigma_{F_2}(\sigma_{F_1}(E_1) \times E_2)$ trong trường hợp $F = F_1(U_1) \times F_2(U_1 U_2)$

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

7. Quy tắc đối với phép chọn và phép hợp:

$$\sigma_F(E_1 \cup E_2) \equiv \sigma_F(E_1) \cup \sigma_F(E_2)$$

8. Quy tắc đối với phép chọn và phép trừ:

$$\sigma_F(E_1 - E_2) \equiv \sigma_F(E_1) - \sigma_F(E_2)$$

Các phép biến đổi tương đương biểu thức ĐSQH (tiếp)

9. Quy tắc đối với phép chiếu và tích Đề các:

$$\Pi_X(E_1(U_1) \times E_2(U_2)) \equiv \Pi_Y(E_1) \times \Pi_Z(E_2)$$

$$X = YZ, Y \subset U_1, Z \subset U_2$$

10. Quy tắc đối với phép chiếu và phép hợp:

$$\Pi_X(E_1 \cup E_2) \equiv \Pi_X(E_1) \cup \Pi_X(E_2)$$

Ví dụ minh họa

Cho CSDL gồm các quan hệ:

S (sid, sname, size, city)

P (pid, pname, colour, weight, city)

SP (sid, pid, quantity)

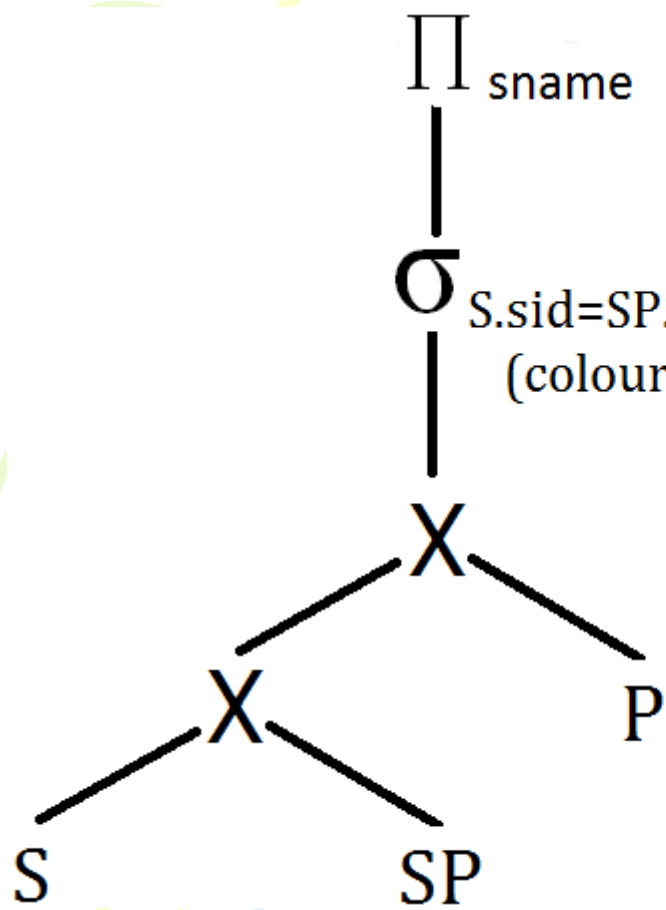
- Tìm tên hãng cung ứng ít nhất một mặt hàng màu đỏ hoặc màu xanh

SELECT sname FROM S, P, SP

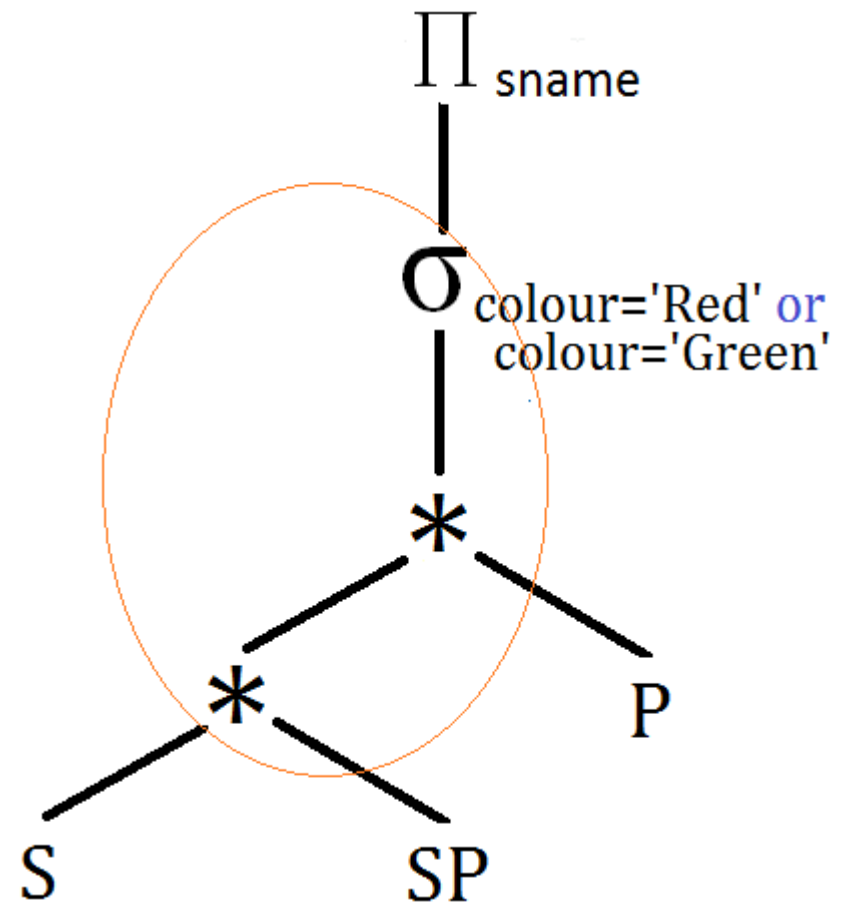
WHERE S.sid = SP.sid AND P.pid = SP.pid AND (colour = 'Red' OR colour = 'Green');

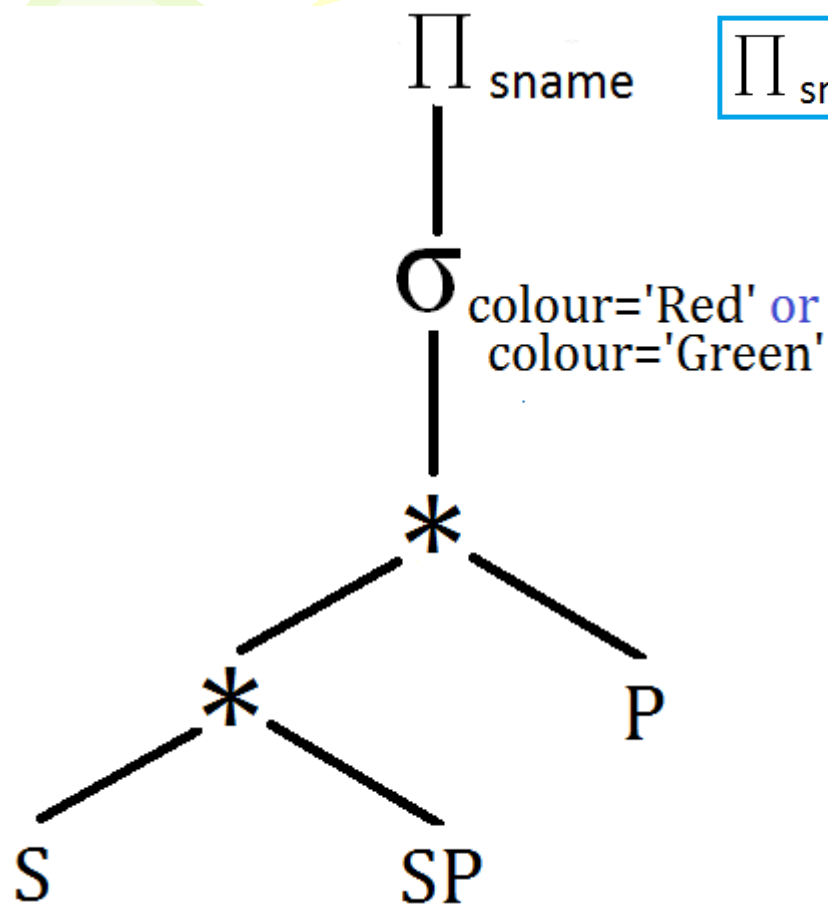
- Biểu thức đại số quan hệ tương đương với câu truy vấn trên là:

$$\Pi_{sname} (\sigma_{S.sid=SP.sid \wedge P.pid=SP.pid \wedge (colour='Red' \vee colour='Green')} (S \times SP \times P))$$

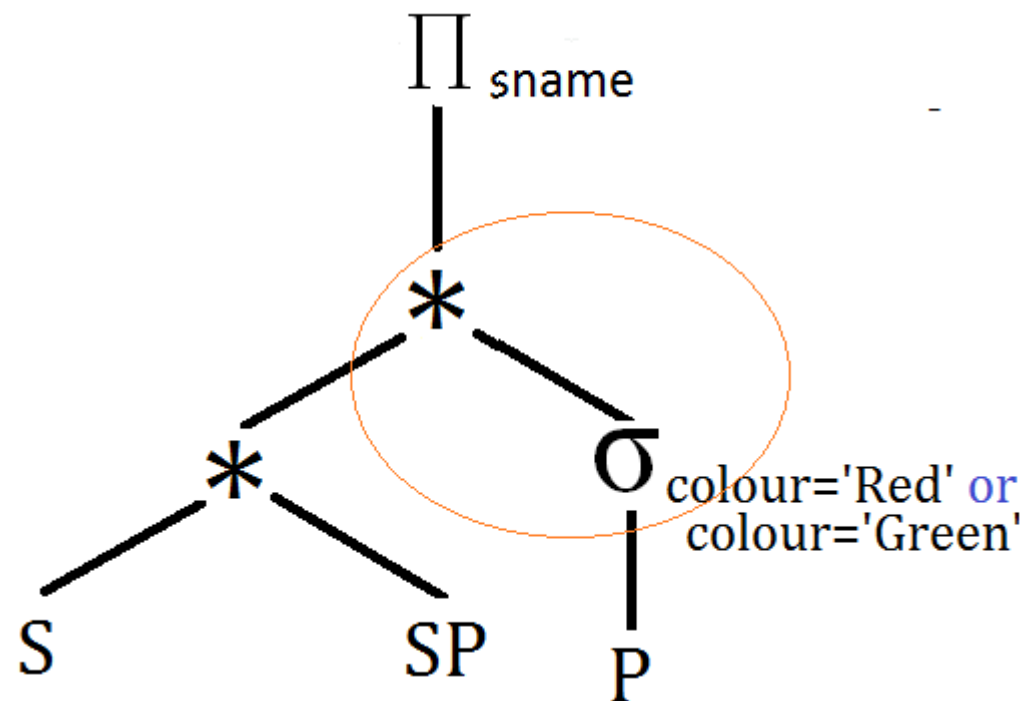


$\Pi_{\text{sname}} \sigma_{S.\text{sid}=\text{SP}.\text{sid} \text{ and } \text{SP}.\text{pid}=\text{P}.\text{pid} \text{ and } (\text{colour}=\text{'Red'} \text{ or } \text{colour}=\text{'Green'})} (S \times \text{SP} \times P)$





$$\Pi_{\text{sname}} \sigma_{\text{colour}='Red' \text{ or } \text{colour}='Green'} (S * SP * P)$$



$$\Pi_{\text{sname}} (S * SP * \sigma_{\text{colour}='Red' \text{ or } \text{colour}='Green'} (P))$$

Π_{sname}

*

*

S

SP

P

$\sigma_{\text{colour}=\text{'Red'} \text{ or } \text{'Green'}}$

$\Pi_{\text{sname}} (S * SP * \sigma_{\text{colour}=\text{'Red'} \text{ or } \text{'Green'}} (P))$

Π_{sname}

$\Pi_{\{\text{sname}, \text{pid}\} \cup \{\text{pid}\}}$

*

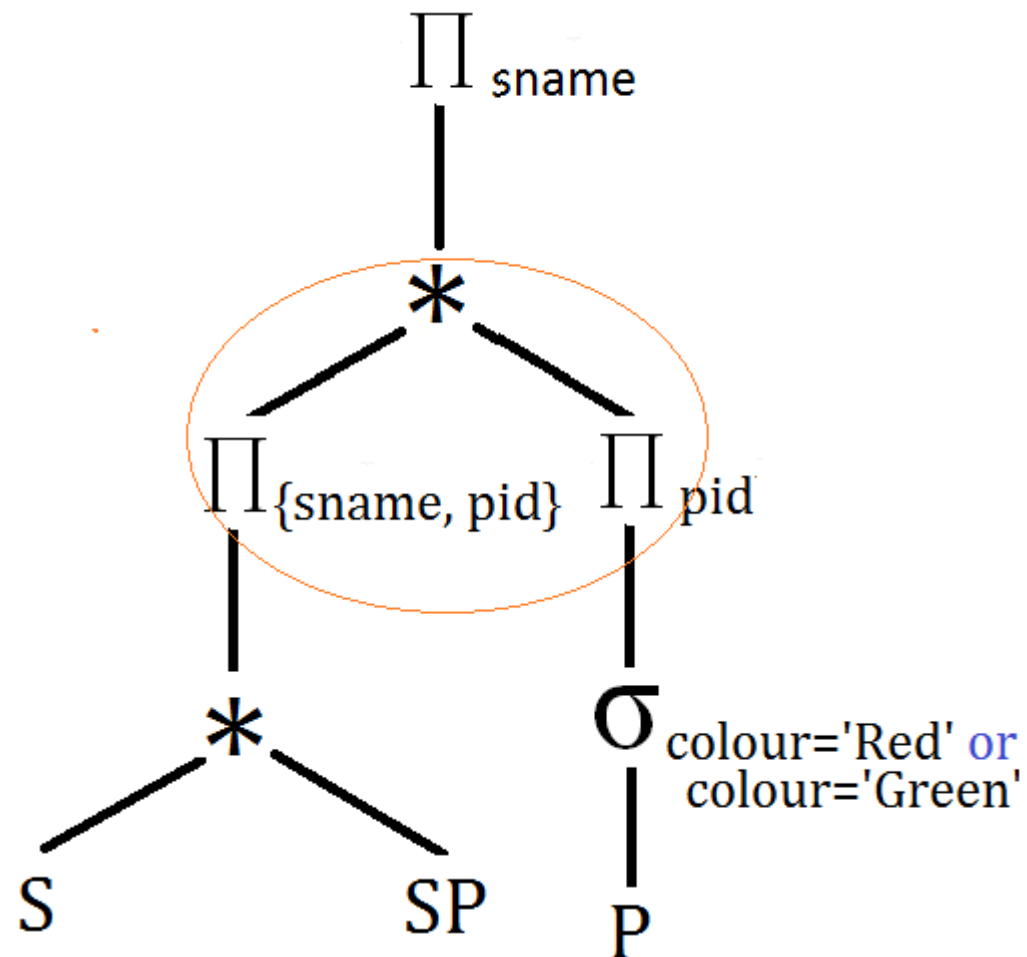
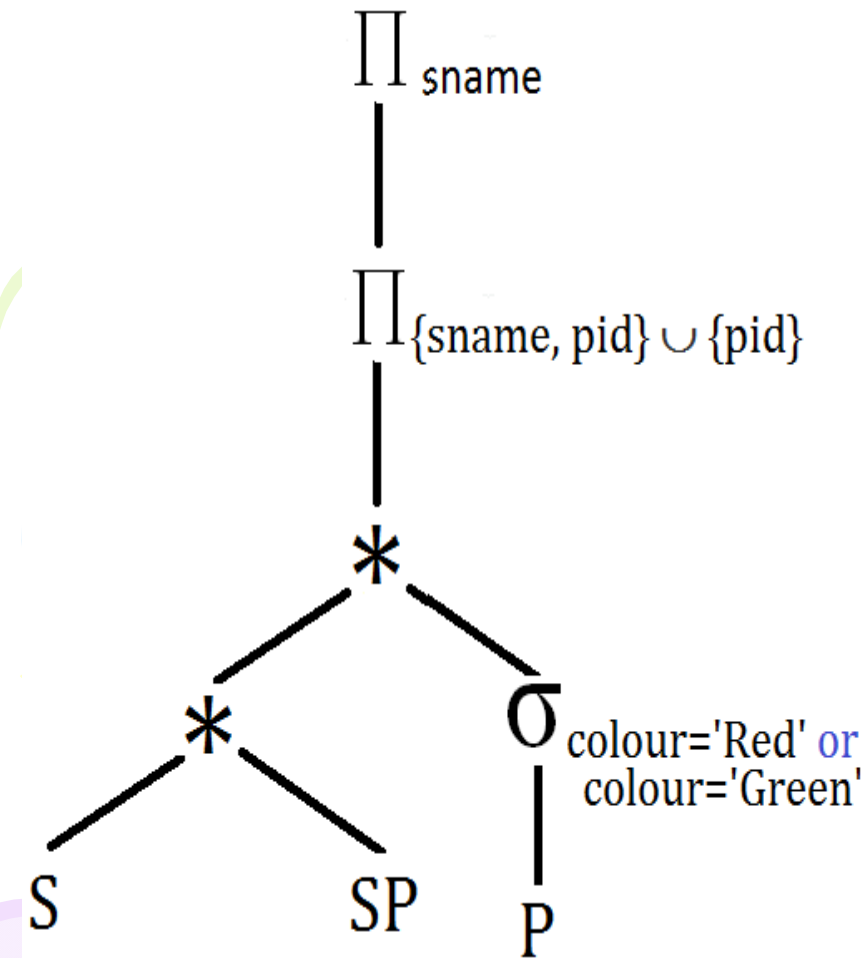
*

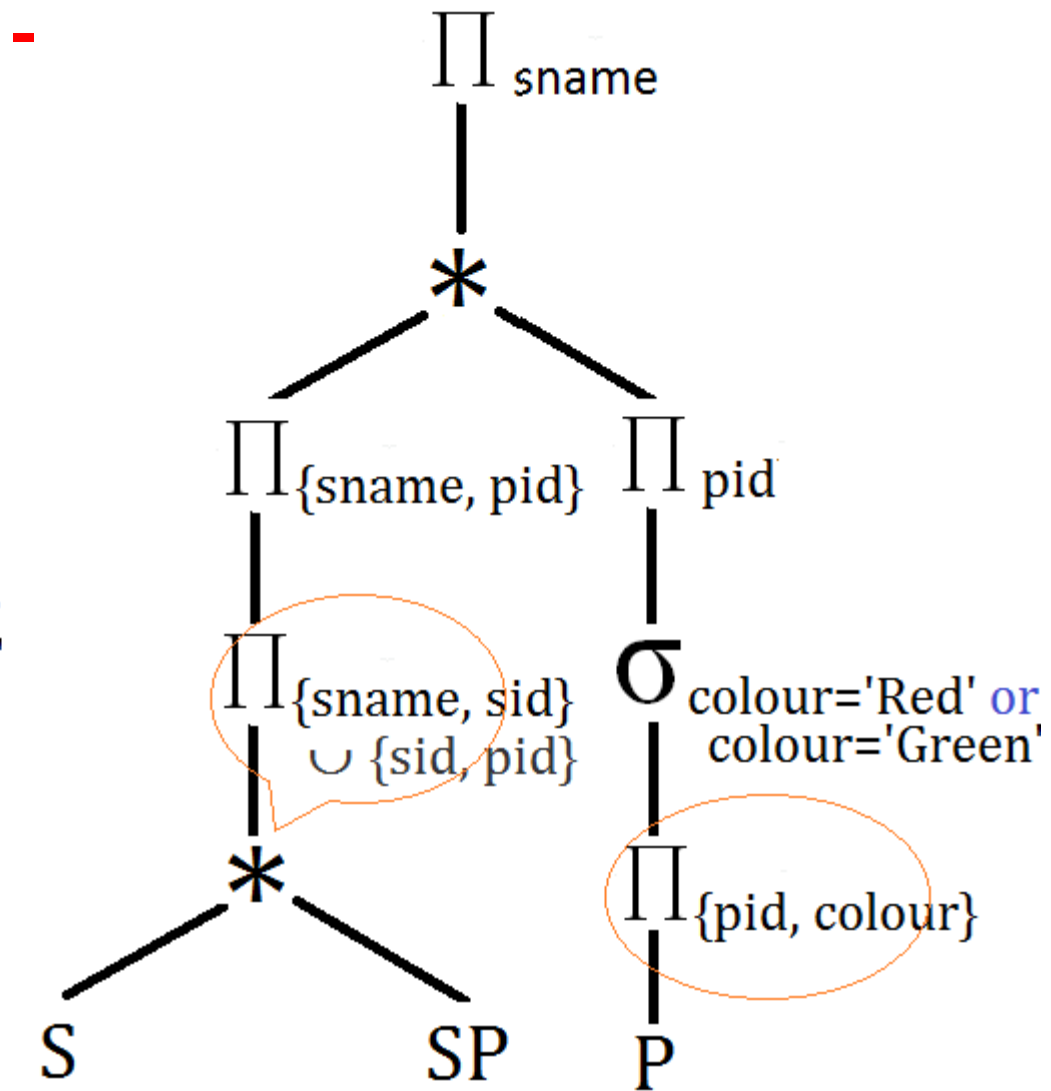
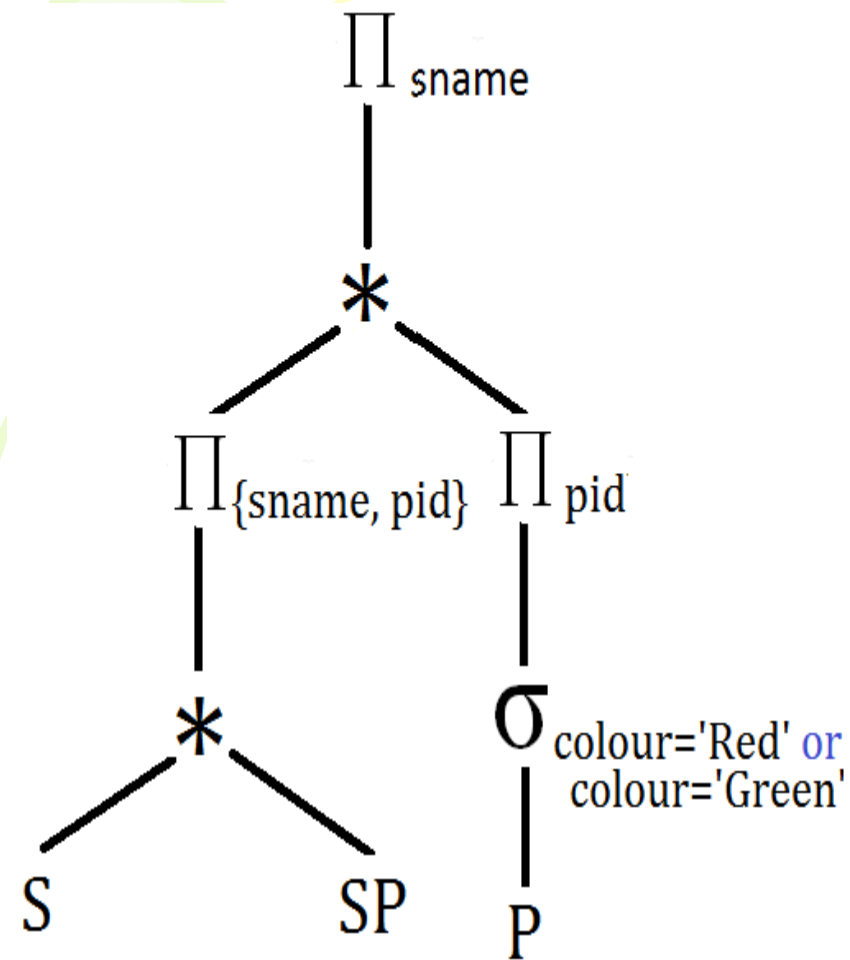
S

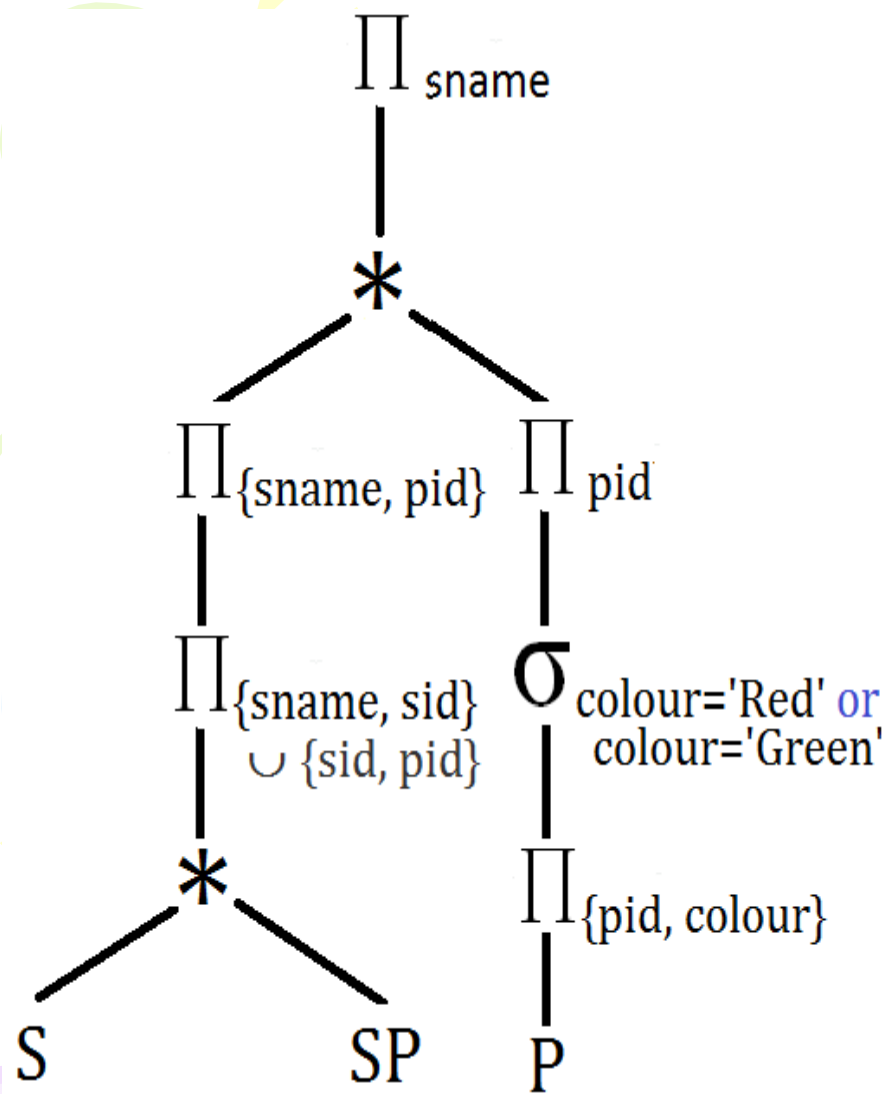
SP

P

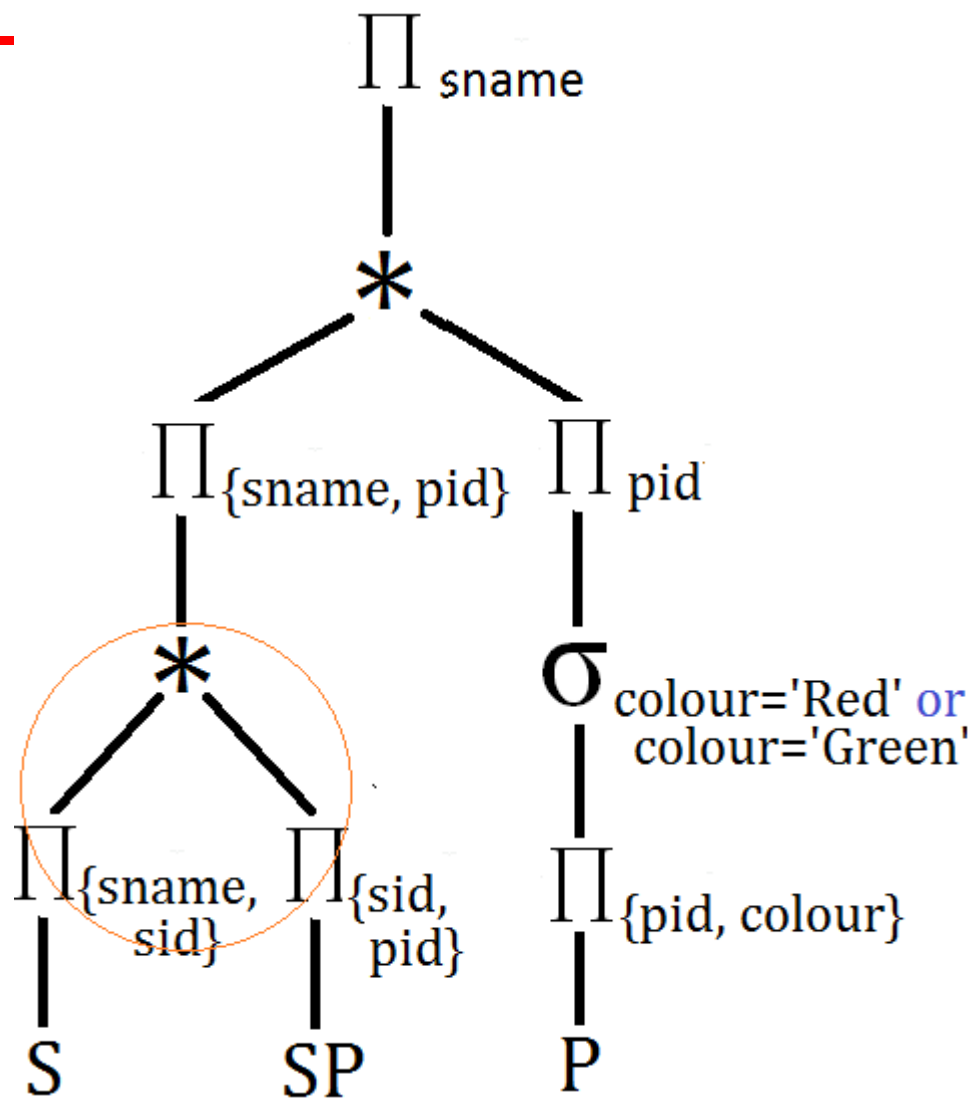
$\sigma_{\text{colour}=\text{'Red'} \text{ or } \text{'Green'}}$

$$\Pi_{\text{sname}} \Pi_{\{\text{sname}, \text{pid}\}} (S * SP * \sigma_{\text{colour}='Red' \text{ or } \text{colour}='Green'} (P))$$




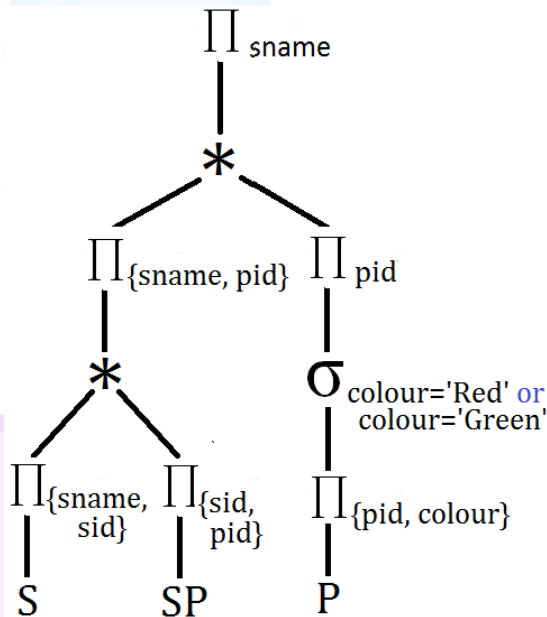



-



$$\Pi_{\text{sname}} \left(\Pi_{\{\text{sname}, \text{pid}\}} \left(\Pi_{\{\text{sname}, \text{sid}\}}(S) * \Pi_{\{\text{sid}, \text{pid}\}}(SP) \right) * \right. \\ \left. * \Pi_{\text{pid}} \left(\sigma_{\text{colour}='Red' \text{ or } \text{colour}='Green'} \left(\Pi_{\{\text{pid}, \text{colour}\}}(P) \right) \right) \right)$$

SELECT sname FROM
 (SELECT sname, pid FROM
 (SELECT sname, sid FROM S) AS S1,
 (SELECT sid, pid FROM SP) AS SP1
 WHERE S1.sid = SP1.sid) AS SSP1,
 (SELECT pid FROM
 (SELECT pid, colour FROM P) AS P1
 WHERE P1.colour = 'Red' or P1.colour = 'Green') AS P2
 WHERE SSP1.pid = P2.pid





```
SELECT sname FROM S WHERE sid IN  
  (SELECT sid FROM SP WHERE pid IN  
    (SELECT pid FROM P WHERE  
      colour = 'Red' OR colour = 'Green' ) )
```

Các thuật toán thực hiện các phép toán đại số

- Các phép toán cơ bản
- Các phép toán 1 lần duyệt:
 - Scan
 - Select
 - Project
- Các phép toán nhiều lần duyệt :
 - Join
 - Aggregation, union, etc.

Các thuật toán thực hiện các phép toán đại số

- Các phép toán 1 lần duyệt : Duyệt một quan hệ.
 - Duyệt tuần tự: đọc tuần tự các khối của quan hệ
 - Duyệt theo index : tìm các bộ theo thứ tự được đánh chỉ dẫn trong index

Các thuật toán thực hiện các phép toán đại số

- Nested-loop JOIN

```
For each tuple tr in r {  
    for each tuple ts in s {  
        if (tr and ts satisfy the join  
condition) {  
            add tuple tr x ts to the result  
set  
        }  
    }  
}
```

- Không có index
- Độ phức tạp tính toán : $O(n^2)$

Các thuật toán thực hiện các phép toán đại số

- Single-loop JOIN (Index-based)

```
for each tuple tr in R {  
    search for tuple ts in s thought index {  
        if ts.exist() {  
            add tuple tr x ts to the result set  
        }  
    }  
}
```

- Sử dụng index trên thuộc tính kết nối
- Độ phức tạp tính toán: $O(n \log m)$

Các thuật toán thực hiện các phép toán đại số

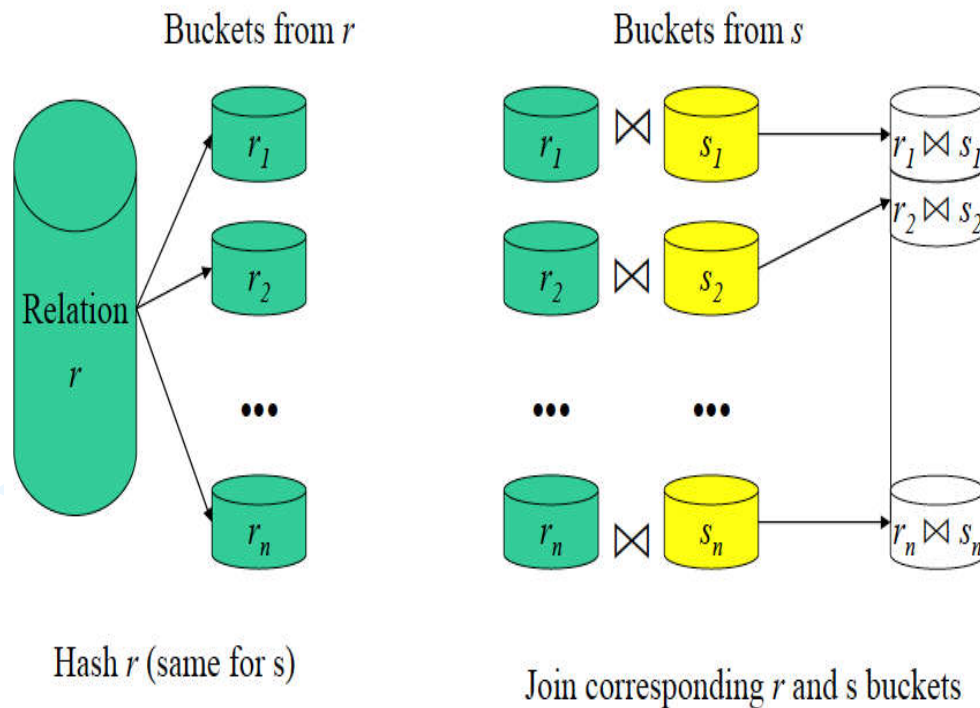
- **Sort-merge JOIN**

- Dữ liệu được sắp theo thuộc tính kết nối:
Merge & join sorted files, đọc tuần tự từng khối
- Sử dụng 2 con trỏ tệp
 - While bộ tại $R <$ bộ tại S , lấy bộ tiếp theo của R (và ngược lại)
 - While các bộ phù hợp, output tất cả các cặp có thể
- Rất hiệu quả nếu dữ liệu đã được sắp.

Các thuật toán thực hiện các phép toán đại số

• Partition-hash JOIN

- Thực hiện bằng 2 quan hệ trên thuộc tính kết nối
- Kết nối các cụm tương ứng





Mô hình chi phí

• Các nhân tố ước lượng chi phí

➤ Thông tin thư mục bao gồm:

- Số bộ trong quan hệ
- Kích thước của một bộ
- Số bộ của quan hệ mà một khối có thể chứa
- Số khối (block) cần thiết để chứa được tất cả các bộ của quan hệ
- Các thông tin về cơ chế truy nhập, chỉ dẫn trên quan hệ

➤ Các độ đo của chi phí truy vấn: Chi phí cho việc thực hiện một truy vấn được đo bởi chi phí sử dụng tài nguyên như việc truy cập đĩa, thời gian CPU dùng để thực hiện một truy vấn.

Do vậy, chúng ta chỉ có thể ước lượng chi phí đ/v mỗi kế hoạch thực hiện truy vấn để tìm ra kế hoạch tốt nhất có thể.

