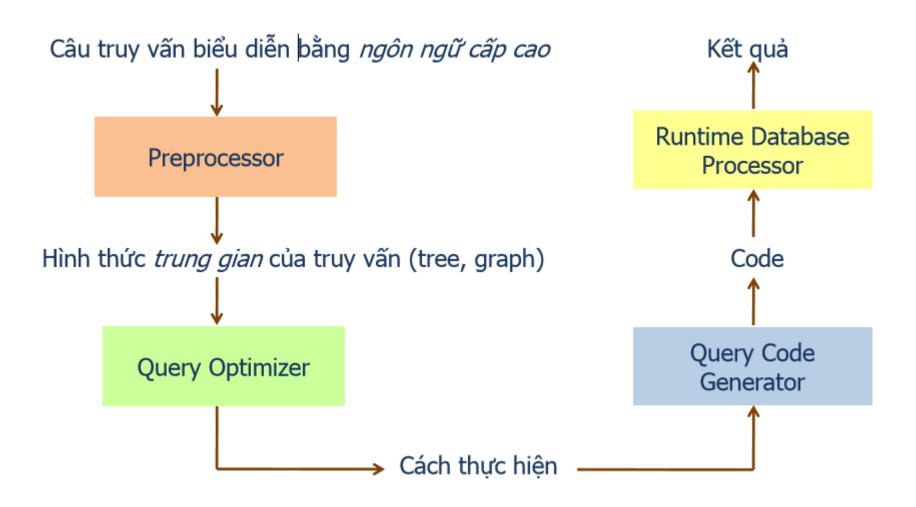
Tối ưu hóa truy vấn

Tối ưu truy vấn

- Mục tiêu
 - Hiểu quy trình thực hiện câu truy vấn
 - Xây dựng câu truy vấn một cách có hiệu quả

Nội dung

- Quy trình thực hiện câu truy vấn
- □ Tiền xử lý câu truy vấn
- Chuyển đổi câu truy vấn
- □ Tối ưu hóa câu truy vấn

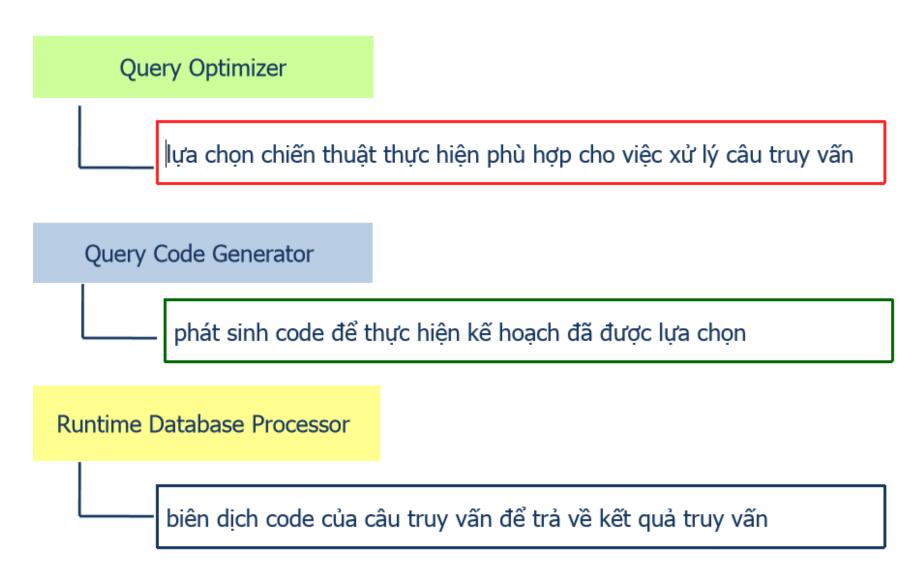


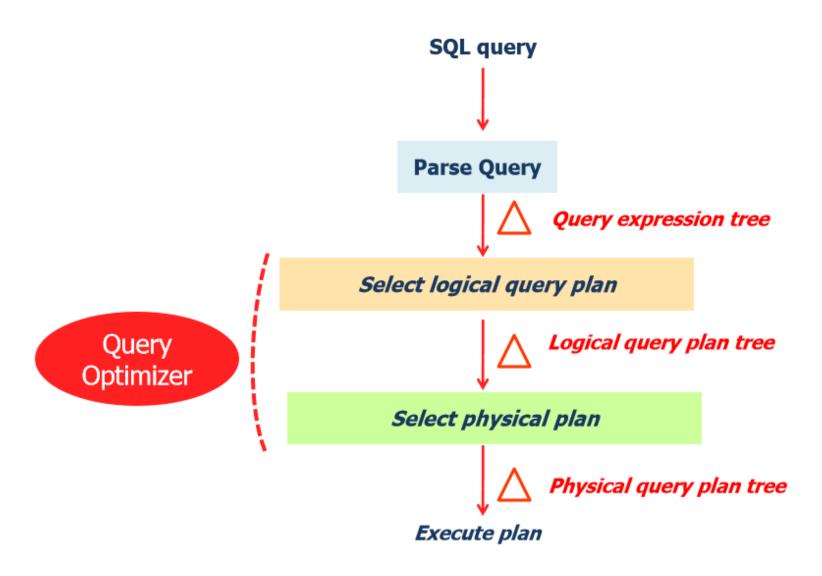


Scanning: xác định các từ khóa, tên thuộc tính, tên các quan hệ,...

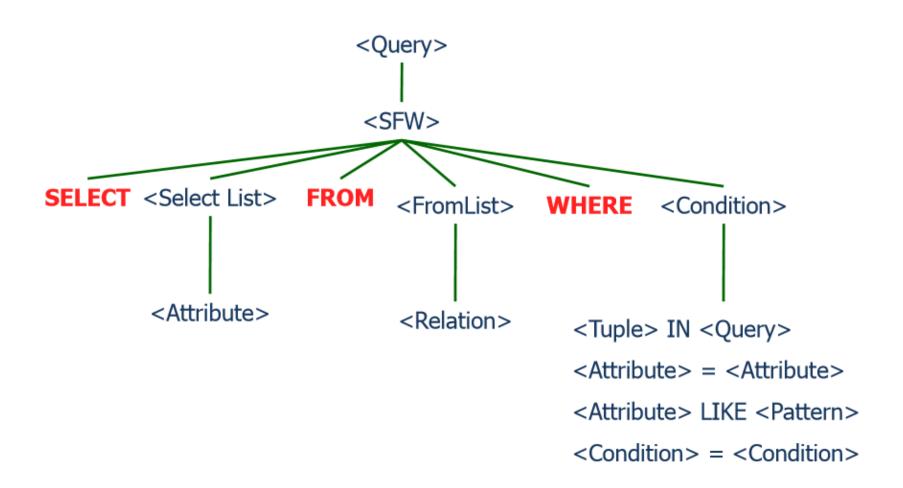
Parsing: kiểm tra cú pháp ngôn ngữ, biểu diễn Parse Tree

Validating: kiểm tra ngữ nghĩa: quan hệ, thuộc tính, kiểu dữ liệu





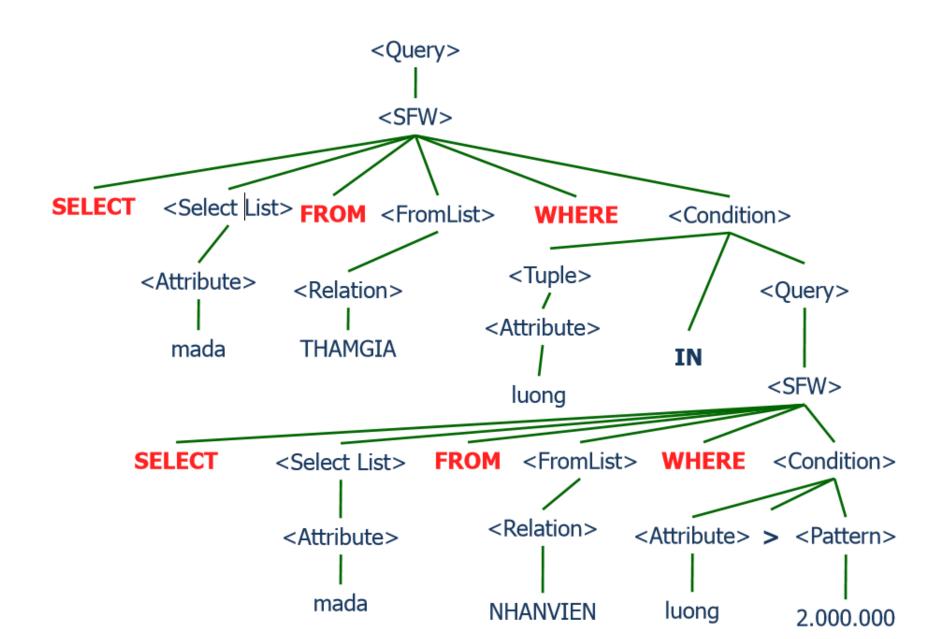
Tiền xử lý câu truy vấn



Ví dụ

```
NHANVIEN (manv, tennv, ngaysinh, phai, luong)
THAMGIA (mada, manv, ngaybatdau, ngayketthuc)
Liệt kê mã đề án mà nhân viên tham gia có lương >2.000.000
SELECT mada
FROM THAMGIA
WHERE many IN (
       SELECT many
       FROM NHANVIEN
       WHERE luong > '2.000.000'
```

Vẽ cây phân tích cú pháp (query expression tree)



Chuyển đổi câu truy vấn

- Chuyển đổi từ SQL sang đại số quan hệ
- Các quy tắc biến đổi tương đương

Chuyển đổi câu truy vấn

- Query block: khối truy vấn đơn vị SELECT FROM WHERE GROUP BY HAVING dùng để chuyển sang ĐSQH
- Truy vấn lồng: tách khối lệnh ghép thành các khối truy vấn đơn vị

Chuyển đổi câu truy vấn

SELECT honv, tennv

FROM NHANVIEN

WHERE luong> (SELECT MAX (luong)

FROM NHANVIEN

WHERE maphong = 5);

SELECThonv, tennvFROMNHANVIENWHEREluong > C $\Pi_{honv, tennv}$ ($\sigma_{luong>C}$ (NHANVIEN))

SELECT MAX (luong)
FROM NHANVIEN
WHERE maphong = 5

 $\mathcal{F}_{MAX | luong} (\sigma_{maphong=5} (NHANVIEN))$

Quy tắc 1: Xử lý các toán tử AND trong điều kiện

$$\sigma_{c1ANDc2...ANDcn}(R) \equiv \sigma_{c1}(\sigma_{c2}(...\sigma_{cn}(R))...)$$

■ Ví dụ

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

```
omaphong = 'KT' AND phai = 'NAM' (NHANVIEN)

≡

maphong = 'KT' ( omaphong = 'KT' ( omaphong = 'NAM' (NHANVIEN))
```

Quy tắc 2: Thay đổi thứ tự của các phép chọn

$$\sigma_{c1}(\sigma_{c2}(R)) \equiv \sigma_{c2}(\sigma_{c1}(R))$$

■ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

Quy tắc 3: Xử lý các phép chiếu

$$\pi_{}(\pi_{}(...\pi_{}(R))...) \equiv \pi_{}(R)$$

■ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

 $\pi_{\text{manv, honv, tennv}}(\pi_{\text{manv, honv, tennv, ngaysinh}}(\text{NHANVIEN}))$

π many, hony, tenny (NHANVIEN)

Quy tắc 4: Thay đổi thứ tự các phép chọn và phép chiếu

$$\pi_{A1,A2,...,An}\left(\sigma_{c}\left(R\right)\right) \equiv \sigma_{c}\left(\pi_{A1,A2,...,An}\left(R\right)\right)$$

Nếu như c ⊂ [A1...An]

■ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

```
\pi_{\text{manv, honv, tennv, phai}} (\sigma_{\text{phai} = \text{'NAM'}} (NHANVIEN))

\equiv
\sigma_{\text{phai} = \text{'NAM'}}(\pi_{\text{manv, honv, tennv, phai}} (NHANVIEN))
```

Quy tắc 5: Tính giao hoán của phép kết và tích Descartes

$$(R \bowtie_{C} S) = (S \bowtie_{C} R) \qquad (R \times S) = (S \times R)$$

■ Ví dụ:

NHANVIEN (many, honv, tennv, ngaysinh, phai, luong, maphong)

PHONGBAN (maphong, tenphong, maql)

(NHANVIEN ⋈ PHONGBAN)

NV.maphong= PB.maphong

(PHONGBAN M NHANVIEN)

NV.maphong= PB.maphong

Quy tắc 6: Thay đổi thứ tự giữa phép chọn và phép kết

$$\sigma_c(R \bowtie S) \equiv (\sigma_c(R)) \bowtie S$$

Nếu như c $\subset R$ (hay c $\subset S$)

■ Ví dụ:

NHANVIEN (many, honv, tennv, ngaysinh, phai, luong, maphong)

PHONGBAN (maphong, tenphong, maql)

 Quy tắc 6: Phân phối giữa phép chọn và phép kết

$$\sigma_c(R \bowtie S) \equiv (\sigma_{c_1}(R)) \bowtie (\sigma_{c_2}(S))$$

Nếu c = c1 and c2, $(c1 \in R \text{ và } c2 \in S)$

□ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

PHONGBAN (maphong, tenphong, maql)

 Quy tắc 7: Phân phối giữa phép chiếu và phép kết

$$L = \{A_1, ..., A_N, B_1, ..., B_M\}; R(A_1, ..., A_N); S(B_1, ..., B_M) V \acute{o}i c \subset L$$

□ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

PHONGBAN (maphong, tenphong, maql)

$$\pi_{\text{manv,tennv,maphong,tenphong}}$$
 (NHANVIEN PHONGBAN)

NV.maphong=PB.maphong

 $(\pi_{\text{manv, honv, maphong}}(\text{NHANVIEN})) \bowtie (\pi_{\text{tenphong, maphong}}(\text{PHONGBAN}))$

Quy tắc 7: Phân phối giữa phép chiếu và phép kết

$$\prod_{L} (R \bowtie_{C} S) = (\prod_{A_{1}, A_{2}, A_{3}, \dots A_{N}, A_{N+1} A_{N+2} \dots A_{N+K}} (R)) \bowtie_{C} (\prod_{B_{1}, B_{2}, B_{3}, \dots B_{M} B_{M+1} B_{M+2} \dots B_{M+P}} (S))$$

Với c $\not\subset$ L, R(A₁,...,A_N, A_{N+1},... A_{N+K}) S(B₁,...,B_M, B_{M+1},...,B_{M+P})

■ Ví dụ:

NHANVIEN (manv, honv, tennv, ngaysinh, phai, luong, maphong)

PHONGBAN (maphong, tenphong, maql)

NV.maphong=PB.maphong

 $(\pi_{\text{manv, tennv, maphong}}(\text{NHANVIEN})) (\pi_{\text{tennv,maphong}}(\text{PHONGBAN}))$

Quy tắc 8: Giao hoán của phép hội và phép giao

$$R \cup S \equiv S \cup R$$
$$R \cap S \equiv S \cap R$$

Quy tắc 9: Kết hợp giữa phép kết, tích Descartes, hội và giao

$$(R \theta S) \theta T = R \theta (S \theta T)$$

Quy tắc 10: Phân phối của phép chọn đối với các phép toán

$$\sigma_c(R \theta S) = (\sigma_c(R)) \theta (\sigma_c(S))$$

Nếu θ là 1 trong các phép toán \cap , \cup , -

Quy tắc 11: Phân phối của phép chiếu đối với các phép toán

Nếu θ là 1 trong các phép toán \cap , \cup , -

$$\prod_{L} (R \theta S) = (\prod_{L} (R)) \theta (\prod_{L} (S))$$

- □ Quy tắc 12: Chuyển các phép (σ, x) thành phép kết
- Luật De Morgan

```
c = NOT (c1 AND c2) = NOT (c1) OR NOT (c2)
```

c = NOT (c1 OR c2) = NOT (c1) AND NOT (c2)

Tối ưu hóa câu truy vấn

- ☐ Giải thuật Heuristic
- Uớc lượng chi phí
 - Hàm chi phí cho Select
 - Hàm chi phí cho Join

Giải thuật Heuristic

- Áp dụng QT 1, tách các phép chọn liên kiện thành 1 dãy các phép chọn.
- 2. Áp dụng QT 2,4,6 và 10, để đẩy phép chọn xuống càng sâu càng tốt
- 3. Áp dụng **QT 9** để tái tổ chức cây cú pháp sao cho phép chọn được thực hiện có lợi nhất (chọn ít nhất)→heuristic
- 4. Phối hợp tích Decartes với các phép chiếu thích hợp theo sau
- 5. Áp dụng **QT 3, 4, 7 và 11** để đẩy phép chiếu xuống càng sâu càng tốt (có thể phát sinh phép chiếu mới)
- 6. Tập trung các phép chọn
- 7. Áp dụng QT3 để loại những phép chiếu vô ích

Giải thuật Heuristic

Liệt kê họ tên NHANVIEN sinh sau năm 1960 và làm dự án 'ABC'

Ngôn ngữ SQL

SELECT honv, tennv

FROM NHANVIEN NV, DEAN DA, THAMGIA TG

WHERE mada='ABC' AND NV.manv=TG.manv AND

DA.mada=TG.mada AND ngaysinh> '31-12-1960'

Ngôn ngữ ĐSQH

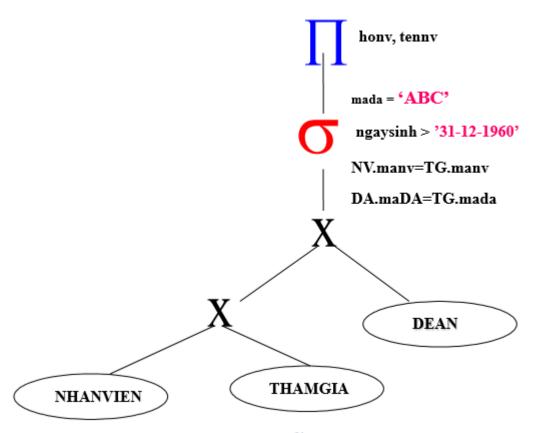
```
honv, tennv mada = 'ABC' ngaysinh > '31-12-1960' \

NV.manv=TG.manv \ DA.mada=TG.mada

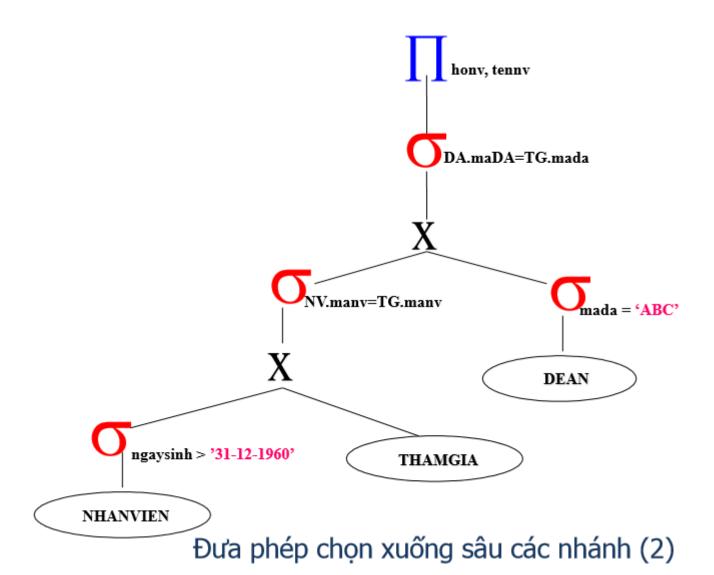
(NHANVIEN x DEAN x THAMGIA))
```

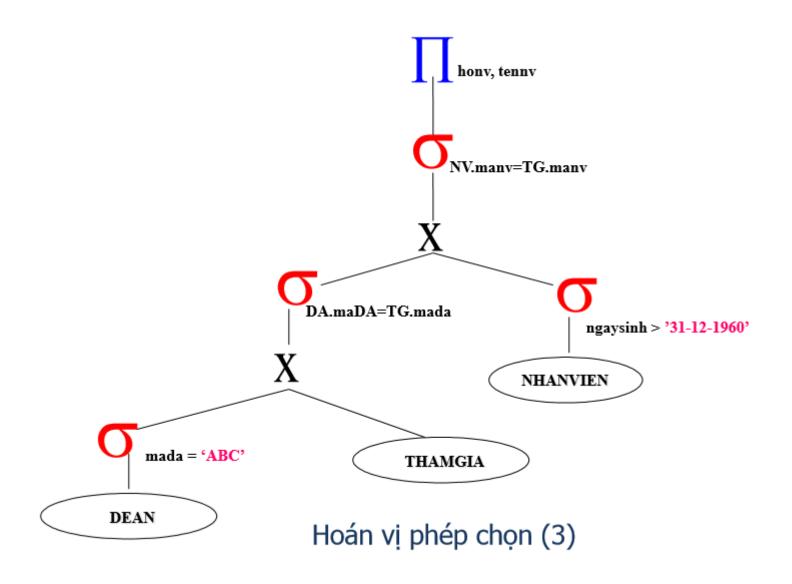
Giải thuật Heuristic

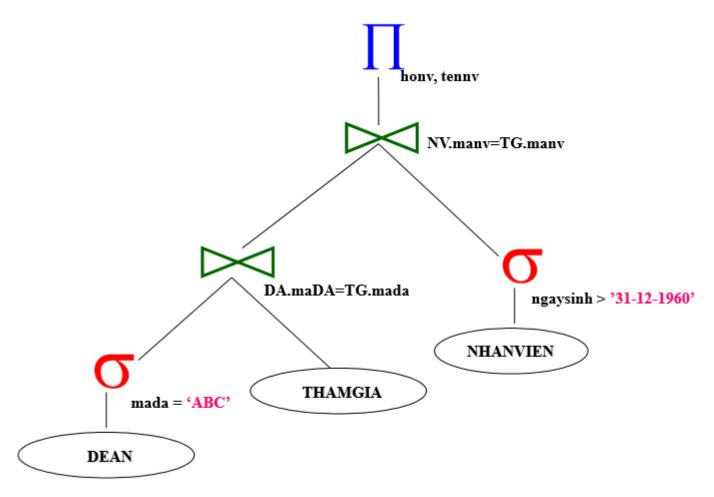
 $\Pi_{\text{honv, tennv}}$ ($\sigma_{\text{mada}} = \text{`ABC'} \land \text{ngaysinh} > \text{`31-12-1960'} \land \text{NV.manv} = \text{TG.manv} \land \text{DA.mada} = \text{TG.mada}$ (NHANVIEN x DEAN x THAMGIA))



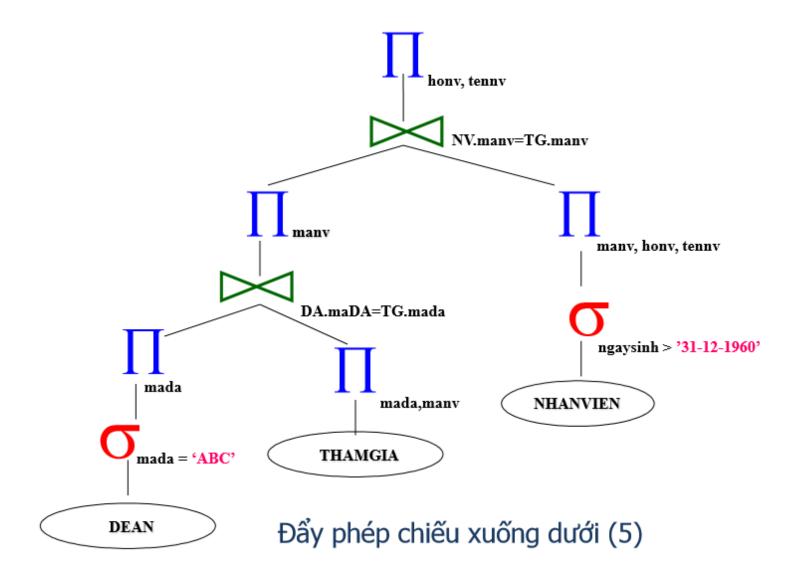
Cây biểu diễn biểu thức truy vấn (1)







Thay thế các phép tích Descartes và phép chọn bằng phép kết (4)



Ngôn ngữ ĐSQH

```
\Pi_{\text{honv,tennv}}((\pi_{\text{mada}}(\sigma_{\text{mada}} = \text{'ABC'}(\text{DEAN})))) \bowtie (\pi_{\text{mada, manv}}(\text{THAMGIA}))
DA.mada = TG.mada
\bowtie_{NV.manv = TG.manv} (\pi_{\text{manv,honv,tennv}}(\sigma_{\text{ngaysinh}} > \text{'31-12-1960'}(\text{NHANVIEN}))))
```

Ngôn ngữ SQL

```
SELECT honv, tennv
FROM

(SELECT mada FROM DEAN

WHERE mada = 'ABC') AS DA INNER JOIN

(SELECT mada, manv FROM THAMGIA) AS TG

ON DA.mada=TG.mada INNER JOIN

(SELECT manv, honv, tennv FROM NHANVIEN WHERE ngaysinh> '31-12-1960' ) NV

ON NV.manv=TG.manv
```

Ước lượng chi phí

So sánh chi phí giữa những cách thực hiện câu truy vấn: chọn cách có chi phí thấp nhất

Chi phí lưu trữ thứ cấp

Chi phí lưu trữ

Chi phí tính toán

Chi phí sử dụng bộ nhớ

Chi phí truyền thông

Ước lượng chi phí

·Các tham số về kích thước file

NHANVIEN

| manv | tenv | phai | hsl |
|------|-------|------|-----|
| NV01 | An | Nam | 1.5 |
| NV02 | Bình | Nam | 1.5 |
| NV03 | Dung | Nữ | 3 |
| NV04 | Duyên | Nữ | 2.5 |

manv: char (20)

tennv: nvarchar (50)

phai: nvarchar (10)

hsl (hệ số lương): double

■Số mẩu tin của bảng (tuples): T(R)

Kích thước 1 mẩu tin: S(R)

Tổng số block chứa tất cả các bộ: b

Số mẩu tin của 1 block: bfr

■Số giá trị khác nhau của thuộc tính A

(kích thước của miền giá trị): V(R,A)

| R | Α | В | С | D |
|---|---|---|----|---|
| | Х | 1 | 10 | а |
| | Х | 1 | 20 | b |
| | у | 1 | 30 | а |
| | у | 1 | 40 | C |
| | z | 1 | 50 | d |

A: chuỗi 20 bytes

B: số nguyên 4 bytes

C: ngày 8 bytes

D: chuỗi 68 bytes

1 block = 1024 bytes (block header: 24 bytes)

$$T(R) = 5$$
 $V(R, A) = 3$ $V(R, B) = 1$ $S(R) = 100*$ $V(R, C) = 5$ $V(R, D) = 4$ $V(R) = 1$

Cho quan hệ R(a,b,c)

Trong đó: a,b integer 4 Bytes

c string 100 Bytes

Header mỗi bộ là 12 Bytes

1 Block 1024 Bytes

Block Header 24 Bytes

Số mẫu tin của bảng T(R)= 10 000

Tính số mẫu tin trong 1 block?

Số Block cần thiết để lưu trữ 10 000 mẫu tin?

Kích thước file tối thiểu chứa được số mẫu tin trên?

Cho quan hệ R(a,b,c)

Trong đó: a,b integer 4 Bytes

c string 100 Bytes

Header mỗi bộ là 12 Bytes

1 Block 1024 Bytes

Block Header 24 Bytes

Số mẫu tin của bảng T(R)= 10 000

Tính số mẫu tin trong 1 block?

 $btr = 1000/120 \approx 8$

Số Block cần thiết để lưu trữ 10 000 mẫu tin?

B(R)= 10 000/8= 1250

Kích thước file tối thiểu chứa được số mẫu tin trên? (1250*1024) Bytes

Cho quan hệ R(a,b,c)

Trong đó: a,b integer 4 Bytes

c string 100 Bytes

Header mỗi bộ là 12 Bytes

1 Block 1024 Bytes

Block Header 24 Bytes

Số mẫu tin của bảng T(R)= 10 000

Nếu S= $\prod_{a+b,c}(R)$ Tính B(R) (gợi ý: 1 Tuple 116 Bytes)

Nếu U= $\prod_{a,b}(R)$ Tính B(R)

 \square Ước lượng W= $\sigma_{A=x}(R)$ (đối với điều kiện =)

$$T(W) = \frac{T(R)}{V(R,A)}$$

 \square Ước lượng W= $\sigma_{A>x}(R)$ (đối với điều kiện >, >=, <, <=)

Cách 1

$$T(W) = \frac{T(R)}{2}$$

Cách 2

$$T(W) = \frac{T(R)}{3}$$

■ Ví dụ

Cho R (A, B, C), tính chi phí S=
$$\sigma_{A=10 \ \land \ B<20}$$
 (R)
Với T(R)=10.000; V(R,A) = 50
Ta có:

$$T(W) = \frac{T(R)}{V(R,A)} = \frac{10000}{50 \times 3}$$

- Hàm tính chi phí cho Select theo phương pháp tìm kiếm P_i: S_i
- Chi phí truy cập block tính theo hàm S_i: C_{Si}

S1. Tìm kiếm tuyến tính

- Duyệt từng mẩu tin, và kiểm tra giá trị thuộc tính của mẩu tin đó có thỏa
 mãn điều kiện chọn (không nhất thiết là điều kiện =) hay không
- Độ phức tạp: O(n)
- Đối với thuộc tính không khóa

$$C_{S1a} = b$$

Đối với điều kiện =, thuộc tính khóa

$$C_{S1b} = (b/2)$$

 $_{\circ}$ đặc biệt, nếu không tìm thấy mẩu tin nào $C_{\rm S1b}$ = b

S2. Tìm kiếm nhị phân

- Nếu điều kiện chọn (=) trên thuộc tính có sắp xếp thứ tự thì việc tìm kiếm
 nhị phân hiệu quả hơn tìm kiếm tuyến tính
- Độ phức tạp: O(log₂n)

$$C_{S2} = log_2b + [(s/bfr)] - 1$$

s: số mầu tin thỏa mãn điều kiện = trên thuộc tính A_k

Đặc biệt đối với điều kiện = trên thuộc tính khóa (hay UNIQUE)

$$C_{S2} = log_2b$$

■ Ví dụ: Cho lược đồ quan hệ

Nhanvien (manv, honv, tennv, ngaysinh, gioitinh, luong, maphong)

Phongban (maphong, tenphong, ngaythanhlap, maql)

□ Tính chi phí cho câu truy vấn sau

Truy vấn: σ_{maphong>5} ∧ _{manv='NV05'} (Nhanvien)

Biết r_{NV} = 10.000 mẩu tin, b_{NV} =2000 blocks

Truy vấn: σ_{maphong>5 ∧ manv='NV05'} (Nhanvien)

Đối với điều kiện maphong>5

$$C_{S1a} = b = 2000$$

Đối với điều kiện manv='NV05'

$$C_{S1a} = b/2 = 1000$$

$$C_{S2} = log_2b = log_22000 = log_22.10^3 = 1 + 3log_210 \approx 11$$

→ Vậy chọn điều kiện manv='NV05' để thực hiện trước

R1 (A, B, C); R2 (A, D)
$$W = R1 \bowtie R2$$

TH1: $V(R1,A) \leq V(R2,A)$

$$T(W) = T(R1)x \frac{T(R2)}{V(R2, A)}$$

TH2:
$$V(R2,A) \leq V(R1,A)$$

$$T(W) = T(R2)x \frac{T(R1)}{V(R1, A)}$$

Tổng quát

$$T(W) = \frac{T(R1)x T(R2)}{\max(V(R1, A), V(R2, A))}$$

Số lượng giá trị của thuộc tính không tham gia phép kết

```
V (W, A) = min {V (R1, A), V(R2, A)}
V (W, B) = V (R1, B)
V (W, C) = V (R1, C)
V (W, D) = V (R2, D)
```

Hàm chi phí cho Join [Elmasri+2003]

- •Hàm tính chi phí cho Join theo phương pháp tìm kiếm P_i: J_i
- Chi phí truy cập block tính theo hàm J_i: C_{ii}

Lưu ý: hàm tính chi phí chỉ dựa trên số block chuyển từ memory đến đĩa (chưa đề cập thời gian tính toán, chi phí lưu trữ và các yếu tố khác)

Độ chọn lọc của phép kết (js)

$$js = |(R \bowtie_C S)| / |R \times S| = |(R \bowtie_C S)| / (|R| * |S|)$$

0 <= js <= 1

Kích thước của tập kết quả sau khi thực hiện phép kết

$$| (R \bowtie_C S) | = js * |R| * |S|$$

- R.A=S.B
 - Nếu A là khóa của R thì
 - Nếu B là khóa của S thì

$$| (R \bowtie _{C} S) | <= |S|, js <= 1/|R|$$

$$| (R \bowtie _{C}S) | <= |R|, js <= 1/|S|$$

- J1. Phép kết lồng nhau
 - Giả sử R là vòng lặp ngoài

$$C_{J1} = b_R + (b_R * b_S) + ((js * |R| * |S|)/bfr_{RS})$$

Ví dụ: Tính chi phí cho phép kết sau

Truy vấn: NHANVIEN NV.maphong=PB.maphong PHONGBAN

Biết: r_{Nhanvien}=10000, r_{Phongban}=125, b_{Nhanvien}=2000, b_{Phongban}=13, brf_{NV_PB=4}

- js = 1/|PHONGBAN| = 1/125
- Sử dụng J1 với NHANVIEN là vòng lặp ngoài

$$C_{J1} = b_{NV} + (b_{NV} * b_{PB}) + ((js * r_{NV} * r_{PB})/brf_{NV_{PB}})$$

=2000 + (2000*13) + ((1/125 * 10000 * 125)/4) =30500

Sử dụng J1 với PHONGBAN là vòng lặp ngoài

$$C_{J1} = b_{PB} + (b_{PB} * b_{NV}) + ((js * r_{NV} * r_{PB})/brf_{NV_PB})$$

=13+ (13*2000) + ((1/125 * 10000 * 125)/4) =28513

Vậy sử dụng PHONGBAN là vòng lặp ngoài