



*Trường Đại học Sư phạm thành phố Hồ Chí Minh*  
*Khoa Công nghệ thông tin*

**HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU**

**TỔ CHỨC KHAI THÁC  
TỐI ƯU HÓA TRUY VẤN**

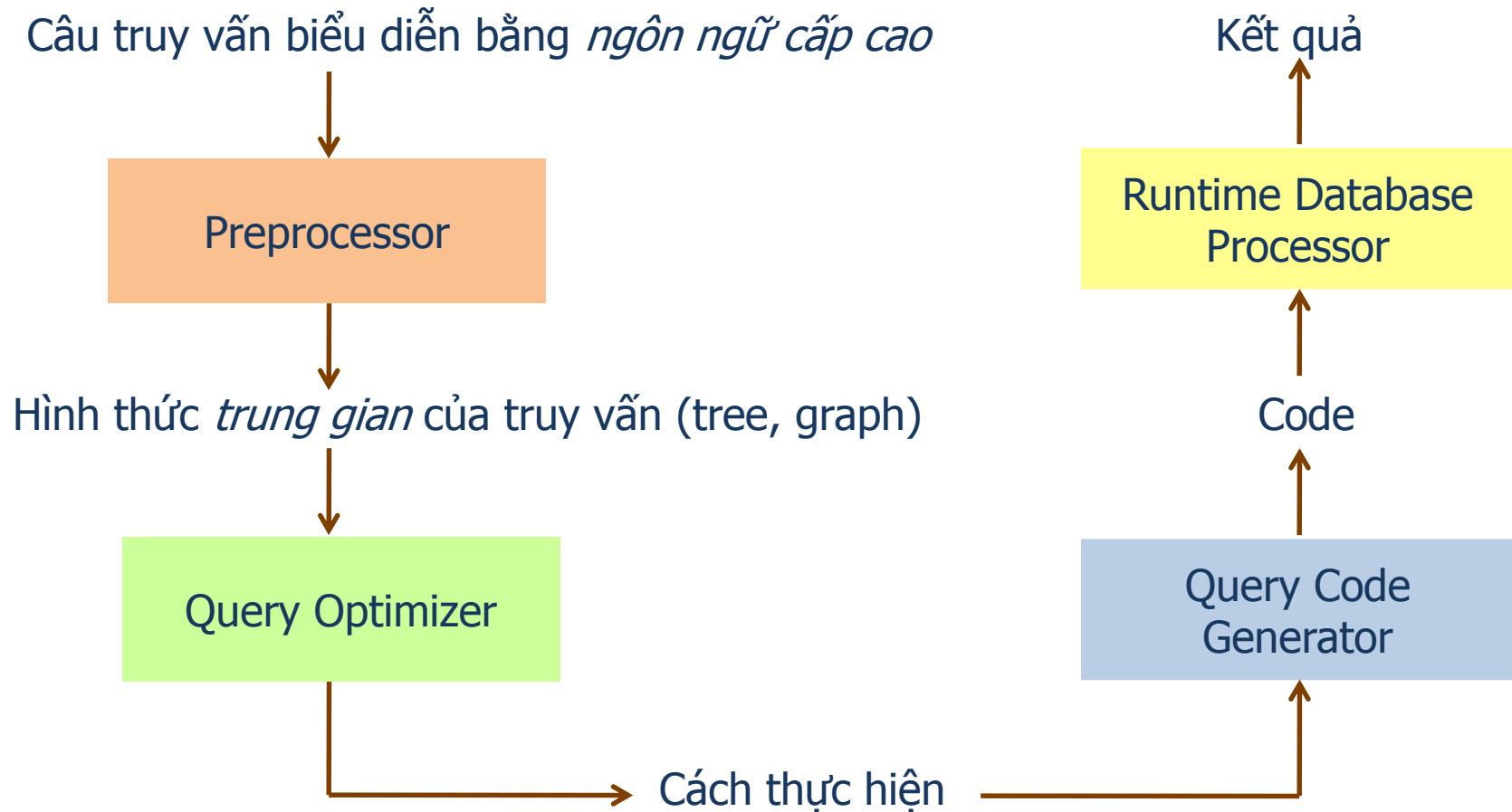
- Hiểu quy trình thực hiện các câu truy vấn
- Xây dựng những câu truy vấn một cách hiệu quả

# Tài liệu tham khảo

- [1] Ramez Elmasri, Shamkant B. Navathe, Fundamentals of Database Systems **(ch. 19), 6th Edition.**
- [2] Jeffrey D. Ullman, Jennifer Widom, Hector Garcia-Monlina, Database Systems: The complete Book **(ch. 15, ch. 16), 2001.**
- [3] Nguyễn An Tế, Nguyễn Tiến Dũng, Nguyễn Thúy Ngọc, Slide bài giảng Các hệ CSDL, 2011-2012

1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn

# 1. Quy trình thực hiện câu truy vấn



# 1. Quy trình thực hiện câu truy vấn (tt.)

Preprocessor

**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

# 1. Quy trình thực hiện câu truy vấn (tt.)

Query Optimizer

lựa chọn chiến thuật thực hiện phù hợp cho việc xử lý câu truy vấn

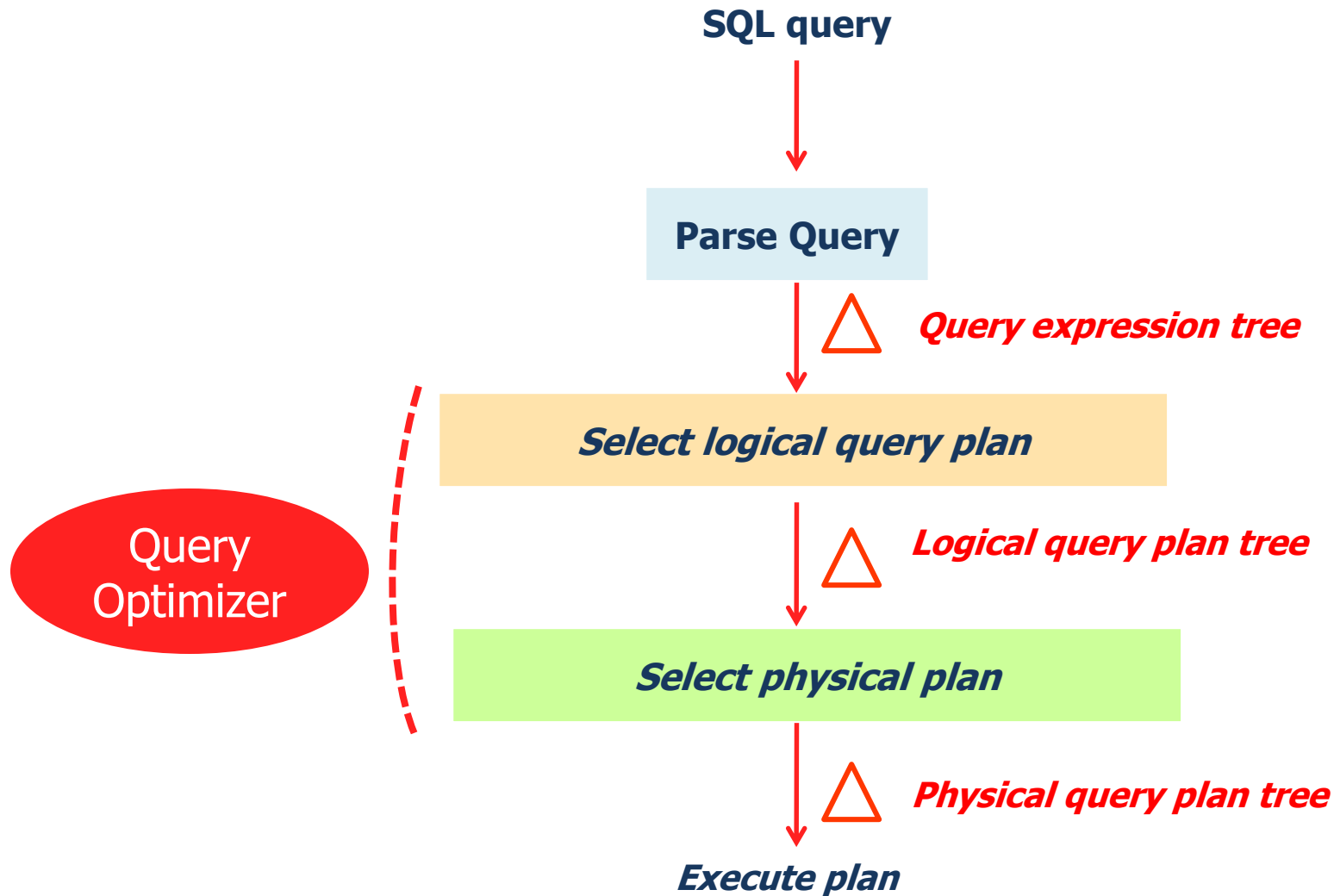
Query Code Generator

phát sinh code để thực hiện kế hoạch đã được lựa chọn

Runtime Database Processor

biên dịch code của câu truy vấn để trả về kết quả truy vấn

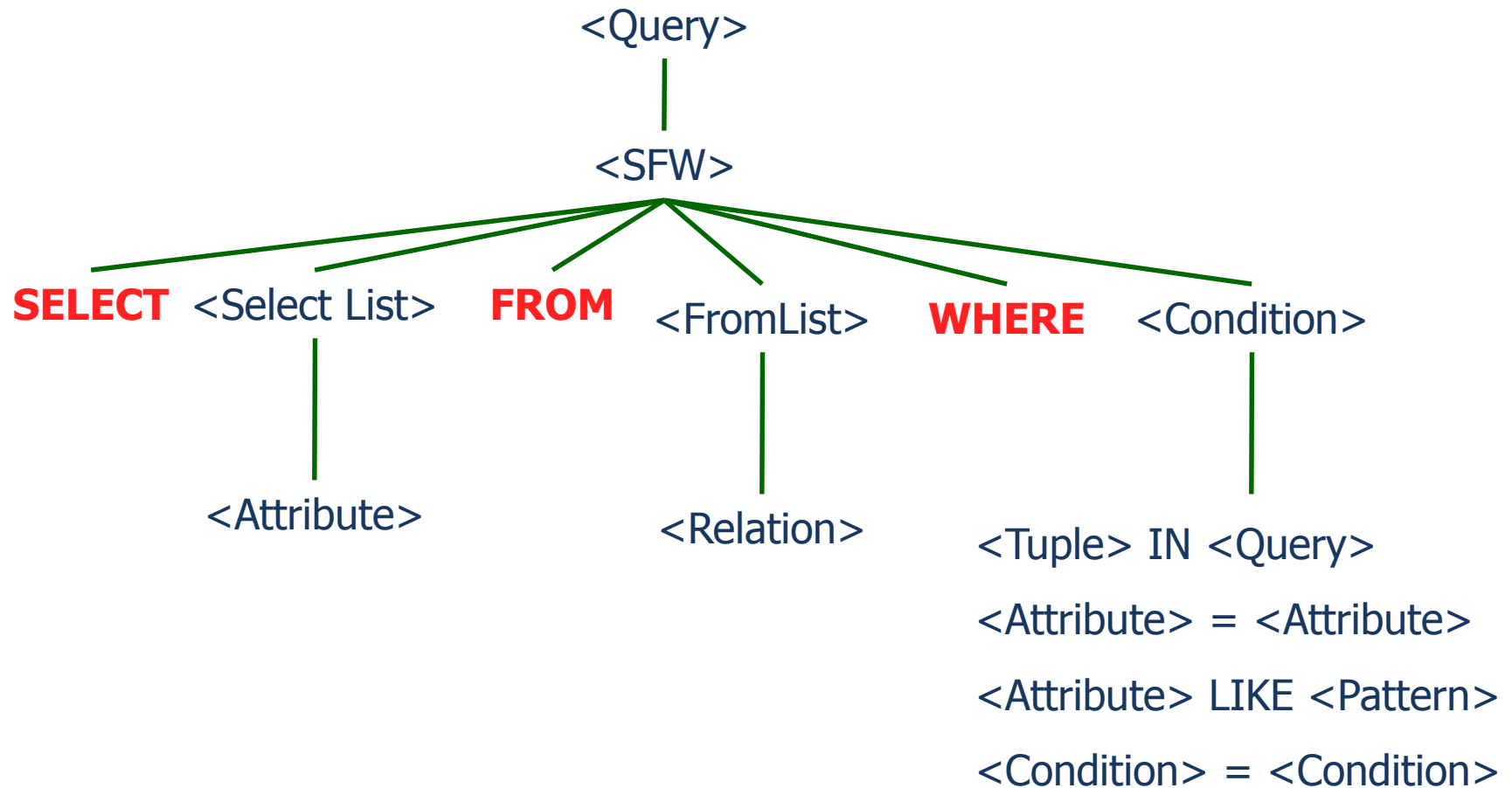
# 1. Quy trình thực hiện câu truy vấn (tt.)





1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn

## 2. Tiền xử lý câu truy vấn



# Quy trình thực hiện câu truy vấn (tt.)

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 manv IN (
```

```
    SELECT manv
```

```
    FROM NHANVIEN
```

```
    WHERE luong >'2.000.000')
```

**Vẽ cây phân tích cú pháp (query expression tree)**



1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
  - 3.1 Chuyển đổi từ SQL sang ĐSQH
  - 3.2 Các quy tắc biến đổi tương đương trong ĐSQH
4. Tối ưu hóa câu truy vấn

## 3.1 Chuyển đổi từ SQL sang ĐSQH

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

### 3.1 Chuyển đổi từ SQL sang ĐSQH (tt.)

SELECT	honv, tennv	
FROM	NHANVIEN	
WHERE	luong > (SELECT	MAX (luong)
	FROM	NHANVIEN
	WHERE	maphong = 5);

**SELECT**  
**FROM**  
**WHERE**

honv, tennv  
NHANVIEN  
luong > C

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

**SELECT**  
**FROM**  
**WHERE**

MAX (luong)  
NHANVIEN  
maphong = 5

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

1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
  - 3.1 Chuyển đổi từ SQL sang ĐSQH
  - 3.2 Các quy tắc biến đổi tương đương trong ĐSQH
4. Tối ưu hóa câu truy vấn



## 3.2 Các quy tắc biến đổi – QT 1

QT1: Xử lý các toán tử AND trong điều kiện

$$\sigma_{c1 \text{ AND } c2 \dots \text{ AND } cn}(R) \equiv \sigma_{c1}(\sigma_{c2}(\dots \sigma_{cn}(R)) \dots)$$

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

$\sigma_{\text{maphong} = \text{'KT' AND phai = \text{'NAM'}}$  (**NHANVIEN**)

$\equiv$

$\sigma_{\text{maphong} = \text{'KT'}}$  ( $\sigma_{\text{phai} = \text{'NAM'}}$  (**NHANVIEN**))

## 3.2 Các quy tắc biến đổi – QT 2

QT2: Thay đổi thứ tự của các phép chọn

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

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

$$\sigma_{\text{maphong} = \text{'KT'}}(\sigma_{\text{phai} = \text{'NAM'}}(\text{NHANVIEN}))$$

$\equiv$

$$\sigma_{\text{phai} = \text{'NAM'}}(\sigma_{\text{maphong} = \text{'KT'}}(\text{NHANVIEN}))$$

## 3.2 Các quy tắc biến đổi – QT 3

QT3: Xử lý các phép chiếu

$$\pi_{\langle DS1 \rangle} \left( \pi_{\langle DS2 \rangle} \left( \dots \pi_{\langle DSn \rangle} (R) \right) \dots \right) \equiv \pi_{\langle DS1 \rangle} (R)$$

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

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

$\equiv$

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

## 3.2 Các quy tắc biến đổi – QT 4

QT4: Thay đổi thứ tự các phép chọn và phép chiếu

$$\pi_{A1,A2,\dots,A_n}(\sigma_c(R)) \equiv \sigma_c(\pi_{A1,A2,\dots,A_n}(R))$$

*Nếu như  $c \subset [A1...An]$*

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

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

$\equiv$

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

## 3.2 Các quy tắc biến đổi – QT 5

QT5: Tính giao hoán của phép kết và tích Descartes

$$(R \bowtie_c S) = (S \bowtie_c R) \quad (R \times S) = (S \times R)$$

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

**PHONGBAN** (maphong, tenphong, maql)

(NHANVIEN  $\bowtie$  PHONGBAN)

$\equiv$

(PHONGBAN  $\bowtie$  NHANVIEN)

NV.maphong= PB.maphong

NV.maphong= PB.maphong

## 3.2 Các quy tắc biến đổi – QT 6a

QT6a: 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  $c \subset S$ )*

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

**PHONGBAN** (maphong, tenphong, maql)

$$\sigma_{\text{phai} = \text{'NAM'}}(\text{NHANVIEN} \bowtie \text{PHONGBAN})$$

$\equiv$

$$(\sigma_{\text{phai} = \text{'NAM'}}(\text{NHANVIEN})) \bowtie \text{PHONGBAN}$$

## 3.2 Các quy tắc biến đổi – QT 6b

QT6b: Phân phối giữa phép chọn và phép kết

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

*Nếu  $c = c_1$  and  $c_2$ , ( $c_1 \in R$  và  $c_2 \in S$ )*

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

**PHONGBAN** (maphong, tenphong, maql)

$$\sigma_{\text{phai} = \text{'NAM'} \text{ AND } \text{tenphong} = \text{'Kế toán'}}(\text{NHANVIEN} \bowtie \text{PHONGBAN})$$

$\equiv$

$$(\sigma_{\text{phai} = \text{'NAM'}}(\text{NHANVIEN})) \bowtie (\sigma_{\text{tenphong} = \text{'Kế toán'}}(\text{PHONGBAN}))$$

## 3.2 Các quy tắc biến đổi – QT 7a

QT7a: Phân phối giữa phép chiếu và phép kết

$$\prod_L (R \bowtie_C S) \equiv (\prod_{A_1, A_2, A_3, \dots, A_N} (R)) \bowtie_C (\prod_{B_1, B_2, B_3, \dots, B_M} (S))$$

$L = \{A_1, \dots, A_N, B_1, \dots, B_M\}; R(A_1, \dots, A_N); S(B_1, \dots, B_M)$  Với  $c \subset L$

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

**PHONGBAN** (maphong, tenphong, maql)

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

$\equiv$

NV.maphong=PB.maphong

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

NV.maphong=PB.maphong



## 3.2 Các quy tắc biến đổi – QT 7b

QT7b: Phân phối giữa phép chiếu và phép kết

$$\Pi_L(R \bowtie_C S) \equiv (\Pi_{A_1, A_2, A_3, \dots, A_N, A_{N+1} A_{N+2} \dots A_{N+K}}(R)) \bowtie_C (\Pi_{B_1, B_2, B_3, \dots, B_M, B_{M+1} B_{M+2} \dots B_{M+P}}(S))$$

Với  $c \notin L$ ,  $R(A_1, \dots, A_N, A_{N+1}, \dots, A_{N+K})$   $S(B_1, \dots, B_M, B_{M+1}, \dots, B_{M+P})$

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

**PHONGBAN** (maphong, tenphong, maql)

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

NV.maphong=PB.maphong

≡

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

NV.maphong=PB.maphong

## 3.2 Các quy tắc biến đổi – QT 8

QT8: Giao hoán của phép hội và phép giao

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

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

## 3.2 Các quy tắc biến đổi – QT 9

QT9: Kết hợp giữa phép kết, tích Descartes, hội và giao

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

***Trong đó  $\theta$  là 1 trong các phép toán  $\bowtie$ ,  $\times$ ,  $\cap$ ,  $\cup$***

## 3.2 Các quy tắc biến đổi – QT 10

QT 10: Phân phối của phép chọn đối với các phép toán

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

***Nếu  $\theta$  là 1 trong các phép toán  $\cap, \cup, -$***

## 3.2 Các quy tắc biến đổi – QT 11

QT 11: Phân phối của phép chiếu đối với các phép toán

***Nếu  $\theta$  là 1 trong các phép toán  $\cap, \cup, -$***

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

## 3.2 Các quy tắc biến đổi – QT 12

QT 12: Chuyển các phép  $(\sigma, \times)$  thành phép kết

$$\sigma_c(R \times S) = R \bowtie_c S$$

Luật De Morgan

$c \equiv \text{NOT } (c1 \text{ AND } c2) \equiv \text{NOT } (c1) \text{ OR } \text{NOT } (c2)$

$c \equiv \text{NOT } (c1 \text{ OR } c2) \equiv \text{NOT } (c1) \text{ AND } \text{NOT } (c2)$

1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn
  - 4.1 Giải thuật Heuristic
  - 4.2 Ước lượng chi phí

# 4.1 Giải thuật heuristic

1. 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. Phôi 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



## 4.1 Giải thuật heuristic (tt.)

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

### Ngôn ngữ SQL

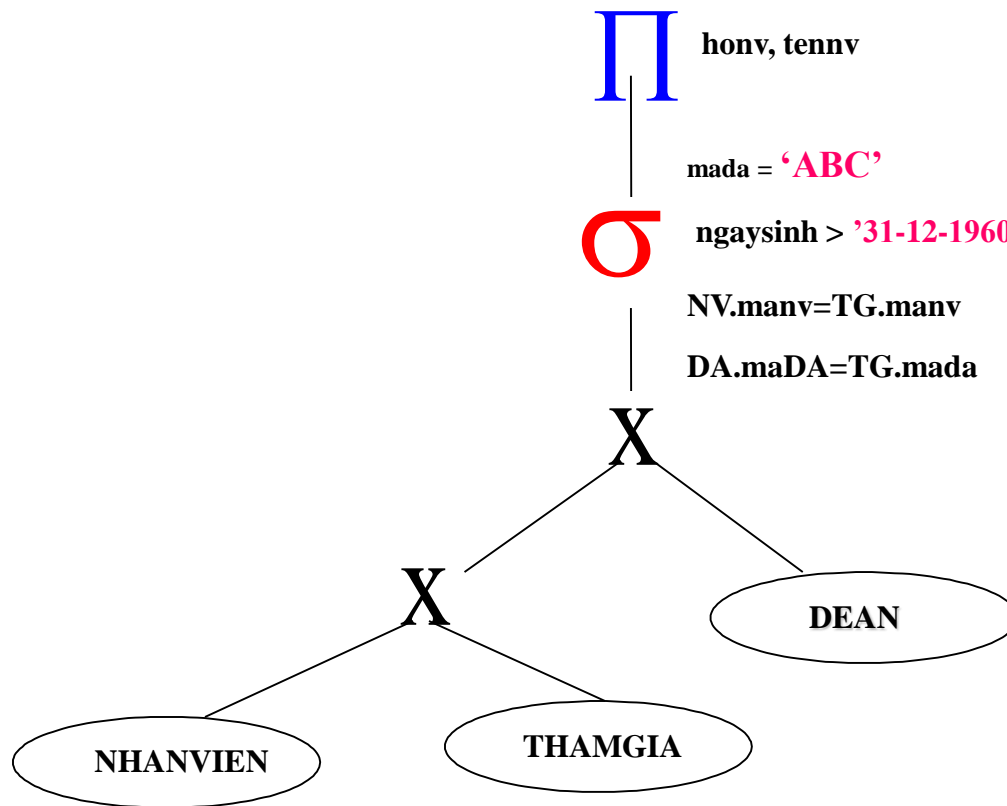
```
SELECT honv, tennv  
FROM NHANVIEN NV, DEAN DA, THAMGIA TG  
WHERE mada='ABC' AND NV.manv=TG.manv AND  
DA.mada=TG.mada AND ngaysinh > '31-12-1960'
```

### Ngôn ngữ ĐSQH

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

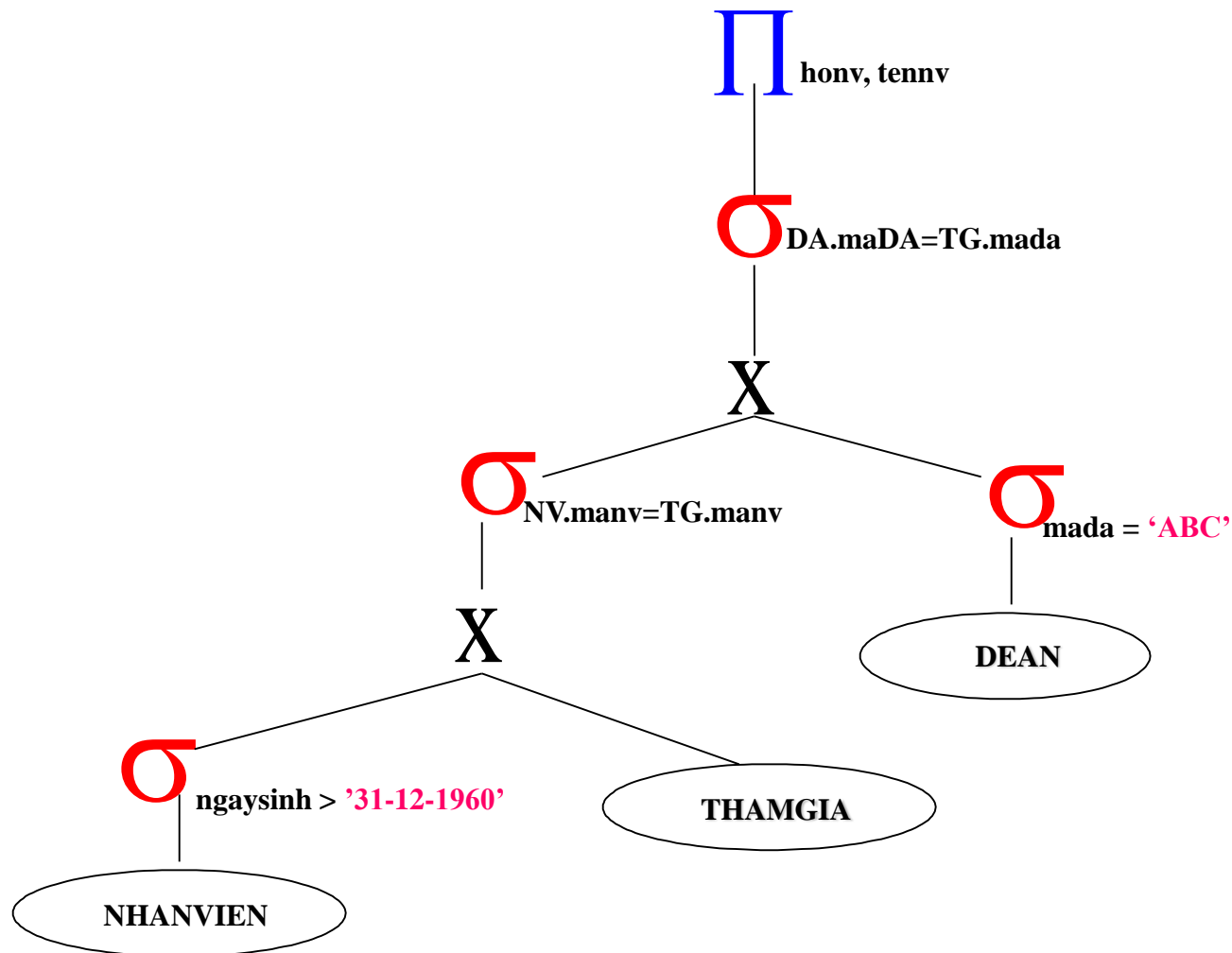
## 4.1 Giải thuật heuristic (tt.)

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



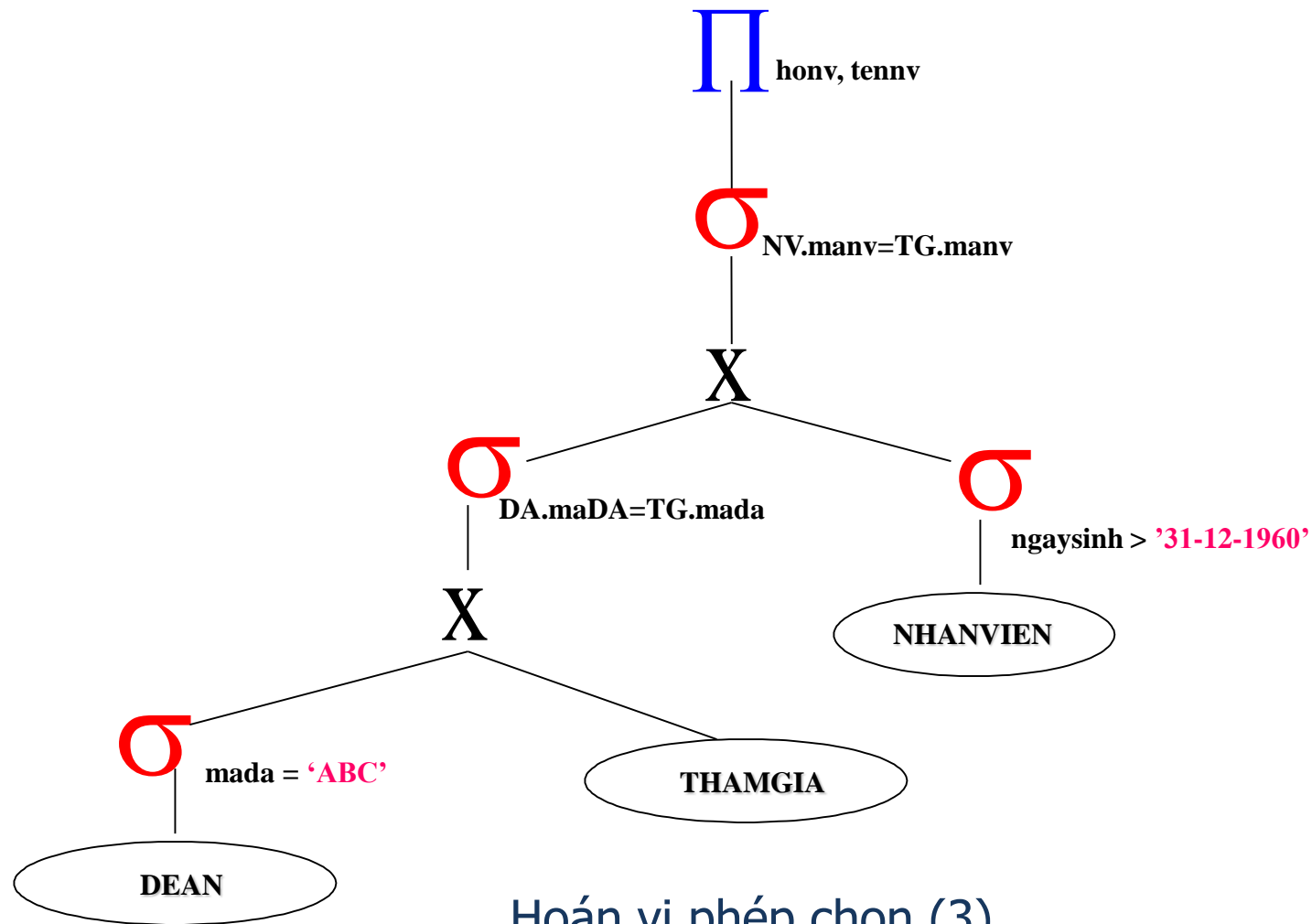
Cây biểu diễn biểu thức truy vấn (1)

## 4.1 Giải thuật heuristic (tt.)

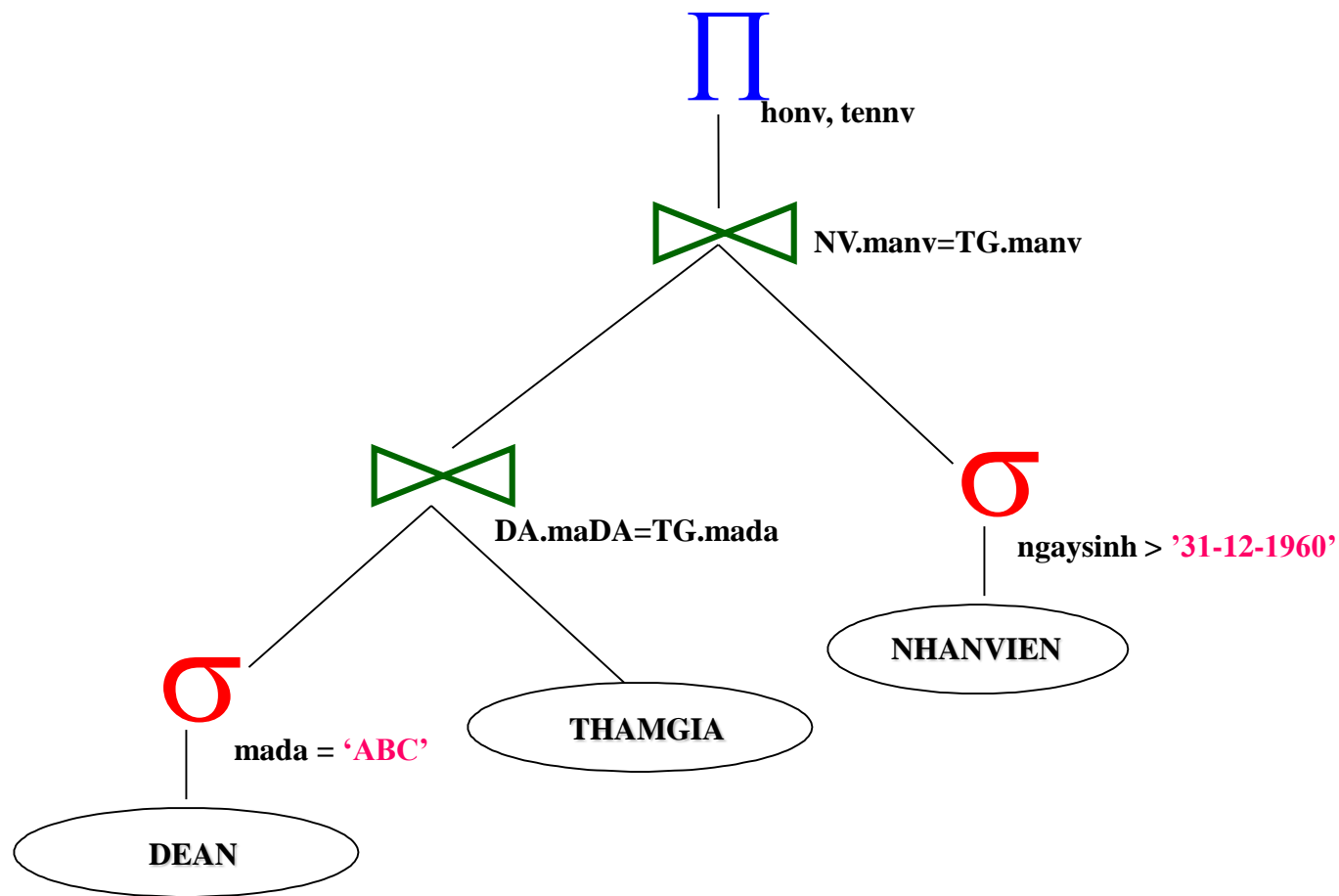


Đưa phép chọn xuống sâu các nhánh (2)

## 4.1 Giải thuật heuristic (tt.)

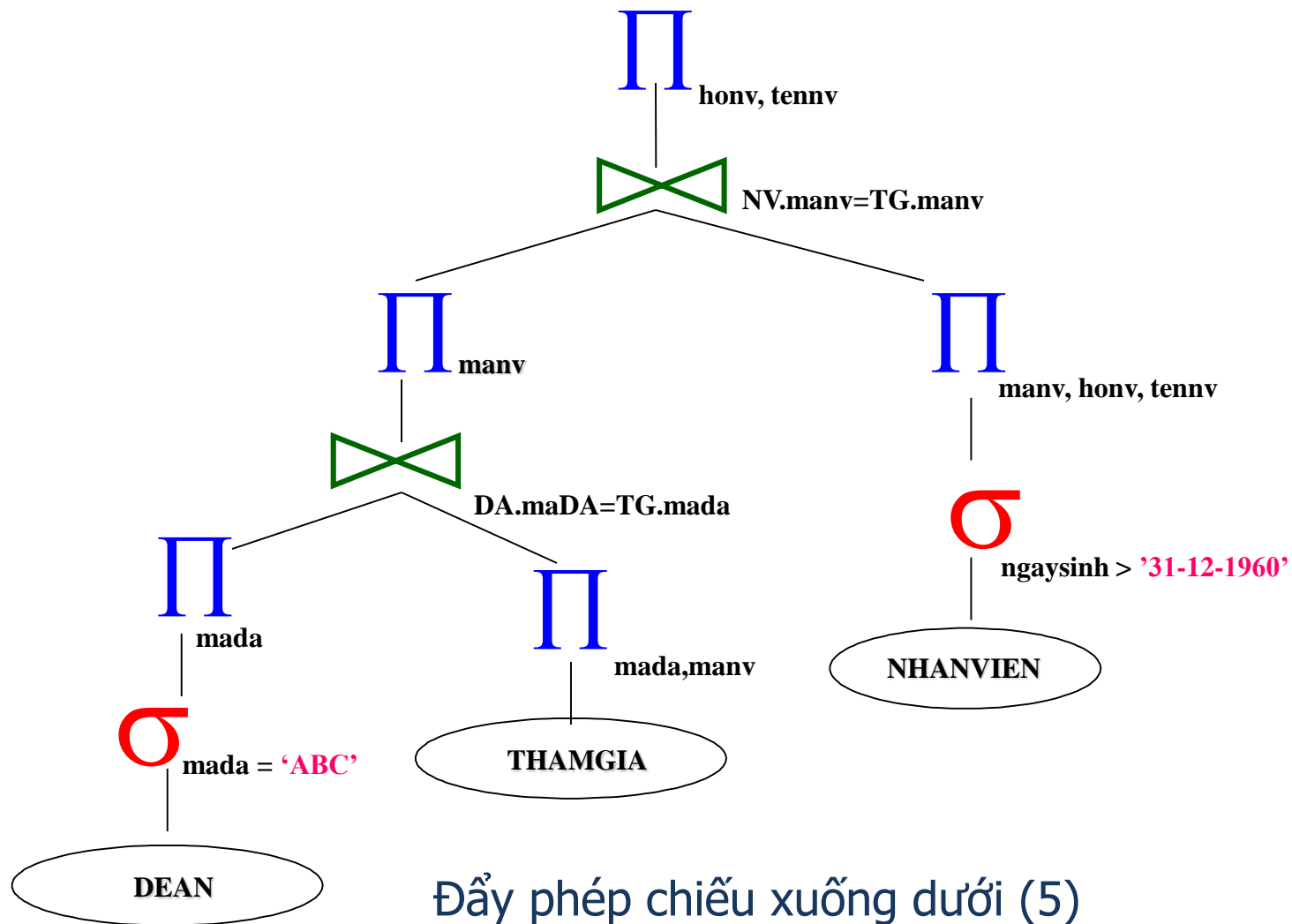


## 4.1 Giải thuật heuristic (tt.)



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

## 4.1 Giải thuật heuristic (tt.)



## 4.1 Giải thuật heuristic (tt.)

### Ngôn ngữ ĐSQH

$$\begin{aligned} & \Pi_{\text{honv,tennv}} \left( \left( \pi_{\text{mada}} \left( \sigma_{\text{mada} = \text{'ABC'}} (\text{DEAN}) \right) \right) \bowtie \left( \pi_{\text{mada, manv}} (\text{THAMGIA}) \right) \right. \\ & \qquad \qquad \qquad \text{DA.mada=TG.mada} \\ & \left. \bowtie \left( \pi_{\text{manv, honv, tennv}} \left( \sigma_{\text{ngaysinh} > \text{'31-12-1960'}} (\text{NHANVIEN}) \right) \right) \right) \\ & \qquad \qquad \qquad \text{NV.manv=TG.manv} \end{aligned}$$

### 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
```

Cho lược đồ CSDL sau:

- SVIEN (MASV, TENS<sub>V</sub>, NAM, MAKHOA)
- HPHAN (MAHP, MAMH, HOCKY, NAM, GV)
- KQUA (MASV, MAHP, DIEM)

Câu 1 (2 điểm):

Dùng thuật toán heuristic tối ưu hóa việc thực hiện câu truy vấn

*«Liệt kê Tên sinh viên, Mã môn học, và điểm số của sinh viên ở học kì 1, năm 2013»*

$\Pi_{TENS_V, MAMH, DIEM}($

$\sigma_{HOCKY = 1 \wedge NAM = 2013 \wedge \text{đk dây quan hệ}}$

$(SVIEN \times KQUA \times HPHAN))$



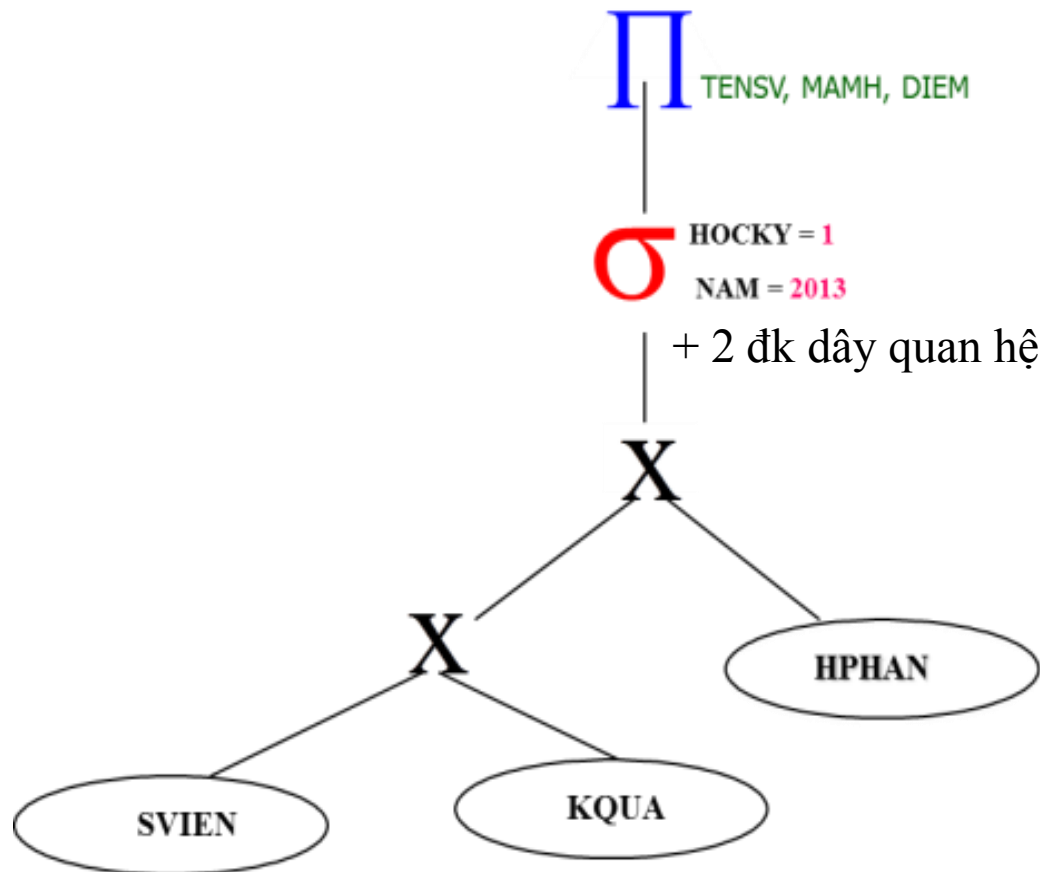
$\Pi_{\text{TENSV, MAMH, DIEM}}($

$\sigma_{\text{HOCKY} = 1 \wedge \text{NAM} = 2013 \wedge \text{đk dây quan hệ}}$

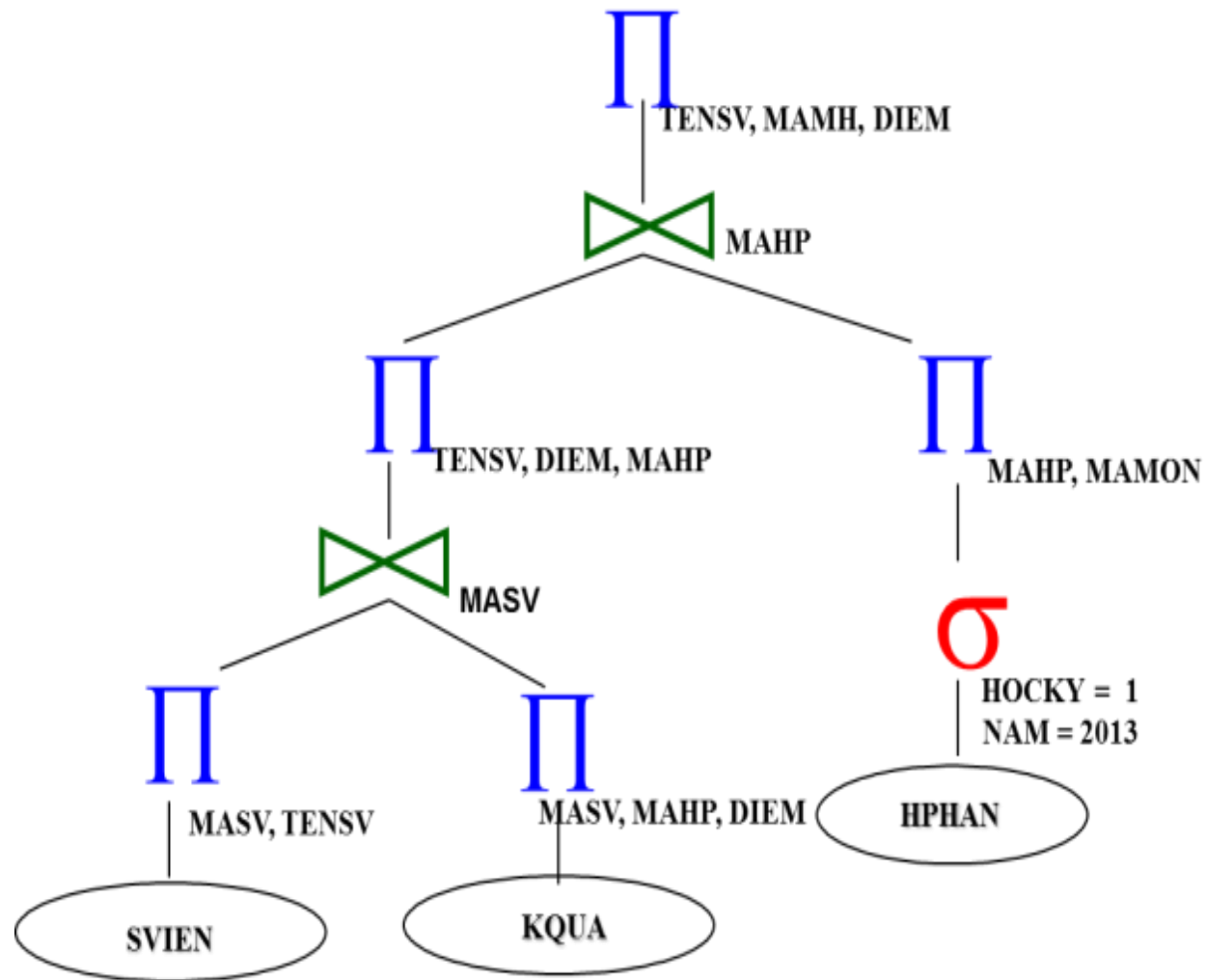
$(\text{SVIEN} \times \text{KQUA} \times \text{HPHAN}))$

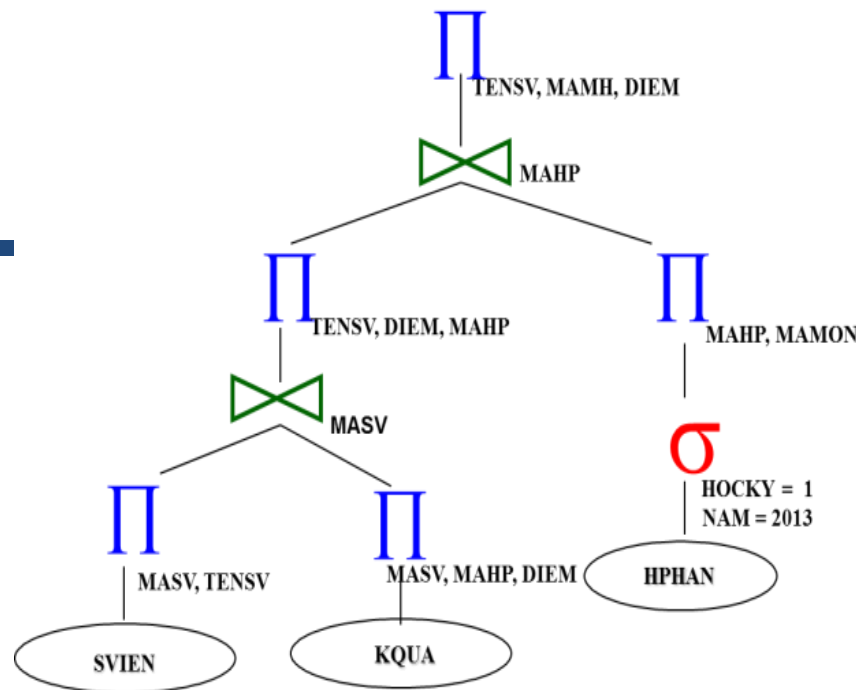
$\Pi_{\text{TENSV, MAMH, DIEM}}($

$\sigma_{\text{HOCKY} = 1 \wedge \text{NAM} = 2013}(\text{SVIEN} \times \text{KQUA} \times \text{HPHAN}))$



Cây biểu diễn biểu thức truy vấn (1)





```

SELECT A. TENS, B. DIEM, C. MAMH
FROM ( SELECT TENS,DIEM,MAHP
      FROM (SELECT MASV, TENS FROM SVIEN) A
      JOIN (SELECT MASV, MAHP, DIEM FROM KQUA) B
      ON A.MASV=B.MASV )) C
JOIN (SELECT MAHP, MAMH FROM HPHAN
      WHERE HOCKY=1 AND NAM=2013) D
ON C.MAHP=D.MAHP

```

Cho lược đồ CSDL sau:

CAUTHU(MACT, TENCT, *MACLB*, NGAYSINH)

CLB(MACLB, TENCLB, SOLUONGCT)

Viết câu truy vấn ‘Liệt kê Mã cầu thủ (MACT), tên cầu thủ (TENCT), ngày sinh (NGAYSINH) thuộc câu lạc bộ có số lượng cầu thủ (SOLUONGCT) nhỏ hơn 12 cầu thủ’.

Tối ưu hóa việc thực hiện câu truy vấn trên dùng thuật toán heuristic.

Cho lược đồ CSDL sau:

CAUTHU(MACT, TENCT, MACLB, NGAYSINH)

CLB(MACLB, TENCLB, SOLUONGCT)

**Câu 2 ( điểm):**

Viết câu truy vấn ‘Liệt kê Mã cầu thủ (MACT), tên cầu thủ (TENCT), ngày sinh (NGAYSINH) thuộc câu lạc bộ có số lượng cầu thủ (SOLUONGCT) nhỏ hơn 12 cầu thủ’. Tối ưu hóa việc thực hiện câu truy vấn trên dùng thuật toán heuristic.

```
SELECT A.MACT, A.TENCT, A.NGAYSINH
FROM (SELECT MACT,TENCT, NGAYSINH, MACLB
      FROM CAUTHU) A
JOIN
      (SELECT MACLB,SOLUONGCT
      FROM CLB WHERE SOLUONGCT<12) B
ON A.MACLB=B.MACLB
```

1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn
  - 4.1 Giải thuật Heuristic
  - 4.2 Ước lượng chi phí

## 4.2 Ước lượng chi phí (tt.)

- 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



## 4.2 Ước lượng chi phí (tt.)

### • 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 mẫu tin của bảng (tuples):  **$T(R)$**
- Kích thước 1 mẫu tin:  **$S(R)$**
- Tổng số block chứa tất cả các bộ:  **$b$**
- Số mẫu tin của 1 block:  **$bfr$**
- Số giá trị khác nhau của thuộc tính  $A$  (kích thước của miền giá trị):  **$V(R,A)$**

R	A	B	C	D
	x	1	10	a
	x	1	20	b
	y	1	30	a
	y	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)

$$T(R) = 5$$

$$S(R) = 100^*$$

$$B(R) = 1$$

$$V(R, A) = 3$$

$$V(R, C) = 5$$

$$V(R, B) = 1$$

$$V(R, D) = 4$$

# Bài tập ví dụ:

- 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?

Tính kích thước file tối thiểu chứa được số mẫu tin trên?

# Bài tập ví dụ:

- 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?

$$bfr = 1000/120 \approx 8$$

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?     $(1250 \times 1024)$  Bytes

# Bài tập ví dụ:

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

1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn
  - 4.1 Giải thuật Heuristic
  - 4.2 Ước lượng chi phí
    - 4.2.1 Hàm chi phí cho Select
    - 4.2.2 Hàm chi phí cho Join

## 4.2.1 Hàm chi phí cho Select

[Ullman + 2001]

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

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

## 4.2.1 Hàm chi phí cho Select

[Ullman + 2001]

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

Cách 1

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

Cách 2

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



## 4.2.1 Hàm chi phí cho Select

- Ví dụ

Cho R (A, B, C), tính chi phí  $S = \sigma_{A=10 \wedge B < 20} (R)$

Với  $T(R) = 10.000$ ;  $V(R, A) = 50$

Ta có:

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

## 4.2.1 Hàm chi phí cho Select (tt.)

*[Elmasri+2003]*

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

## 4.2.1 Hàm chi phí cho Select (tt.)

- S1. Tìm kiếm tuyến tính

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

## 4.2.1 Hàm chi phí cho Select (tt.)

- S1. Tìm kiếm tuyến tính (tt.)
  - Đối với thuộc tính không khóa

$$C_{S1a} = b$$

- Đối với điều kiện =, thuộc tính khóa

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

- đặc biệt, nếu không tìm thấy mẫu tin nào  $C_{S1b} = b$

## 4.2.1 Hàm chi phí cho Select (tt.)

- 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:  $O(\log_2 n)$

## 4.2.1 Hàm chi phí cho Select (tt.)

- S2. Tìm kiếm nhị phân (tt.)

$$C_{S2} = \log_2 b + [(s/bfr)] - 1$$

*s: số mẫu tin thỏa mãn điều kiện = trên thuộc tính  $A_k$*

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

$$C_{S2} = \log_2 b$$

## 4.2.1 Hàm chi phí cho Select (tt.)

- **Ví dụ:** Cho lược đồ quan hệ

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

**Phongban** (maphong, tenphong, ngaythanhlap, maql)

**Câu hỏi:** Tính chi phí cho câu truy vấn sau

Truy vấn:  $\sigma_{\text{maphong} > 5} \wedge \text{manv} = \text{'NV05'}$  (**Nhanvien**)

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

## 4.2.1 Hàm chi phí cho Select (tt.)

- Truy vấn:  $\sigma_{\text{maphong} > 5 \wedge \text{manv} = \text{'NV05'}} (\text{Nhanvien})$

- Đối với điều kiện ***maphong > 5***

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

- Đối với điều kiện ***manv = 'NV05'***

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

$$C_{S2} = \log_2 b = \log_2 2000 = \log_2 2 \cdot 10^3 = 1 + 3 \log_2 10 \approx 11$$

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



1. Quy trình thực hiện câu truy vấn của DBMS
2. Tiền xử lý câu truy vấn
3. Chuyển đổi câu truy vấn
4. Tối ưu hóa câu truy vấn
  - 4.1 Giải thuật Heuristic
  - 4.2 Ước lượng chi phí
    - 4.2.1 Hàm chi phí cho Select
    - 4.2.2 Hàm chi phí cho Join

## 4.2.2 Hàm chi phí cho Join

[Ullman,+ 2001]

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

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

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

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

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

## 4.2.2 Hàm chi phí cho Join (tt.)

[Ullman + 2001]

- Tổng quát

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

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

$$V(W, A) = \min \{V(R1, A), V(R2, A)\}$$

$$V(W, B) = V(R1, B)$$

$$V(W, C) = V(R1, C)$$

$$V(W, D) = V(R2, D)$$

## 4.2.2 Hàm chi phí cho Join (tt.)

*[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  $J_i$  :  $C_{ji}$

**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)

## 4.2.2 Hàm chi phí cho Join (tt.)

- Độ chọn lọc của phép kết (js)

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

$0 \leq js \leq 1$

- 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

$$| (R \bowtie_c S) | \leq |S|, js \leq 1/|R|$$

$$| (R \bowtie_c S) | \leq |R|, js \leq 1/|S|$$

## 4.2.2 Hàm chi phí cho Join (tt.)

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

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

## 4.2.2 Hàm chi phí cho Join (tt.)

- Ví dụ: Tính chi phí cho phép kết sau

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

Biết:  $r_{\text{Nhanvien}}=10000$ ,  $r_{\text{Phongban}}=125$ ,  $b_{\text{Nhanvien}}=2000$ ,  $b_{\text{Phongban}}=13$ ,  
 $\text{brf}_{\text{NV\_PB}}=4$

## 4.2.2 Hàm chi phí cho Join (tt.)

- $js = 1/|PHONGBAN| = 1/125$

- Sử dụng J1 với NHANVIEN là vòng lặp ngoài

$$\begin{aligned}C_{J1} &= b_{NV} + (b_{NV} * b_{PB}) + ((js * r_{NV} * r_{PB}) / brf_{NV\_PB}) \\ &= 2000 + (2000 * 13) + ((1/125 * 10000 * 125) / 4) = 30500\end{aligned}$$

- Sử dụng J1 với PHONGBAN là vòng lặp ngoài

$$\begin{aligned}C_{J1} &= b_{PB} + (b_{PB} * b_{NV}) + ((js * r_{NV} * r_{PB}) / brf_{NV\_PB}) \\ &= 13 + (13 * 2000) + ((1/125 * 10000 * 125) / 4) = 28513\end{aligned}$$

**Vậy sử dụng PHONGBAN là vòng lặp ngoài**



