

Chương 6

Tối ưu hóa câu truy vấn

Thời lượng: 3 tiết



Nội dung chi tiết



- Giới thiệu
- Bộ biên dịch câu truy vấn (query compiler)
- Phân tích cú pháp
 - Cây phân tích (parse tree)
- Chuyển cây phân tích sang ĐSQH
 - Câu truy vấn đơn giản
 - Câu truy vấn lồng - lồng tương quan
- Quy tắc tối ưu cây truy vấn
- Tối ưu hóa câu truy vấn
 - **Giải thuật Heuristic**
 - Ước lượng chi phí

Giới thiệu



- R(A, B, C)
- S(C, D, E)

```
SELECT  B, D
FROM    R, S
WHERE   R.A='c' AND S.E=2 AND R.C=S.C
```

3

Giới thiệu (tt)



- Câu truy vấn được thực hiện như thế nào?

R	A	B	C
	a	1	10
	b	1	10
	c	2	10
	d	2	10
	e	3	10

S	C	D	E
	10	x	2
	20	y	2
	30	z	2
	40	x	1
	50	y	3

Kết quả

B	D
2	x

4

Giới thiệu (tt)



- Cách 1
 - Tích cartesian
 - Phép chọn (selection)
 - Phép chiếu (projection)

$$\Pi_{B,D} [\sigma_{R.A=c' \wedge S.E=2 \wedge R.C = S.C} (R \times S)]$$

5

Giới thiệu (tt)



RxS

A	B	C	C	D	E
a	1	10	10	x	2
a	1	10	20	y	2
⋮					
c	2	10	10	x	2
c	2	10	20	y	2
c	2	10	30	z	2
⋮					

6

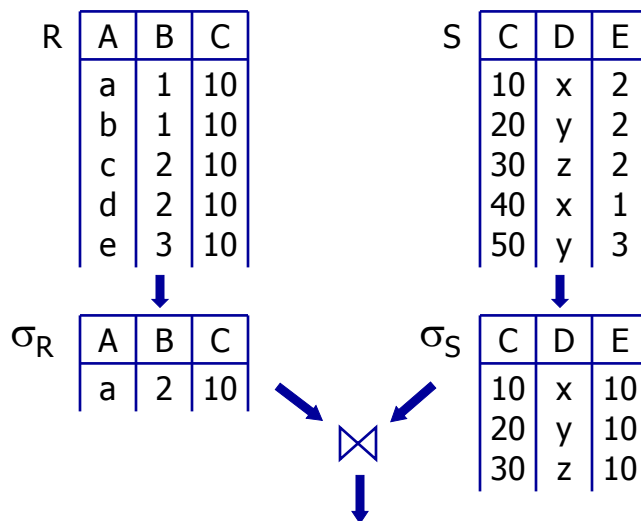
Giới thiệu (tt)

- Cách 2
 - Phép chọn (selection)
 - Phép kết (natural join)
 - Phép chiếu (projection)

$$\Pi_{B,D} [\sigma_{R.A='c'}(R) \bowtie \sigma_{S.E=2}(S)]$$

7

Giới thiệu (tt)



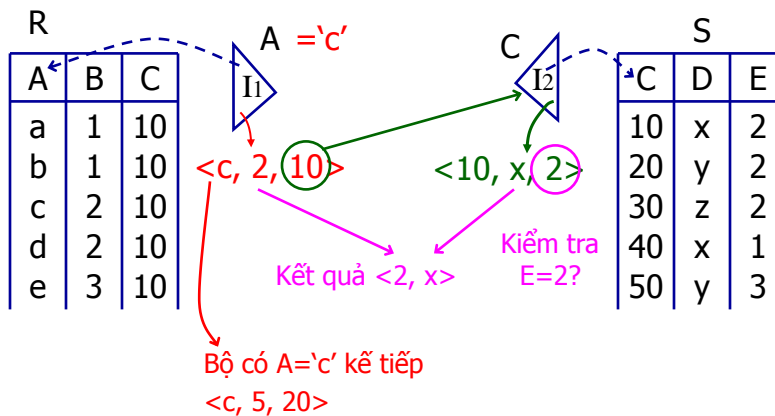
8

Giới thiệu (tt)

- Cách 3 - sử dụng chỉ mục trên R.A và S.C
 - Tìm các bộ trong R thỏa $R.A = 'c'$
 - Với mỗi bộ tìm thấy, tìm tiếp các bộ trong S thỏa $R.C = S.C$
 - Bỏ đi những bộ $S.E \neq 2$
 - Kết các bộ phù hợp của R và S
 - Chiếu trên thuộc tính B và D

9

Giới thiệu (tt)



10

Giới thiệu (tt)



DBMS thực hiện cách nào

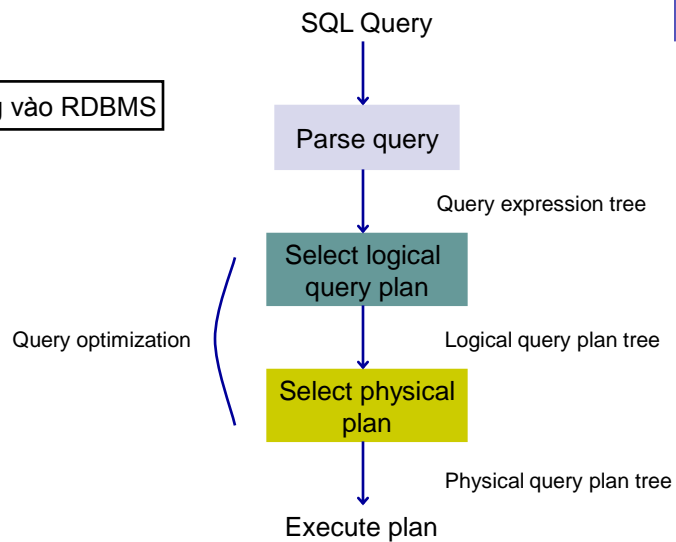


11

Bộ biên dịch

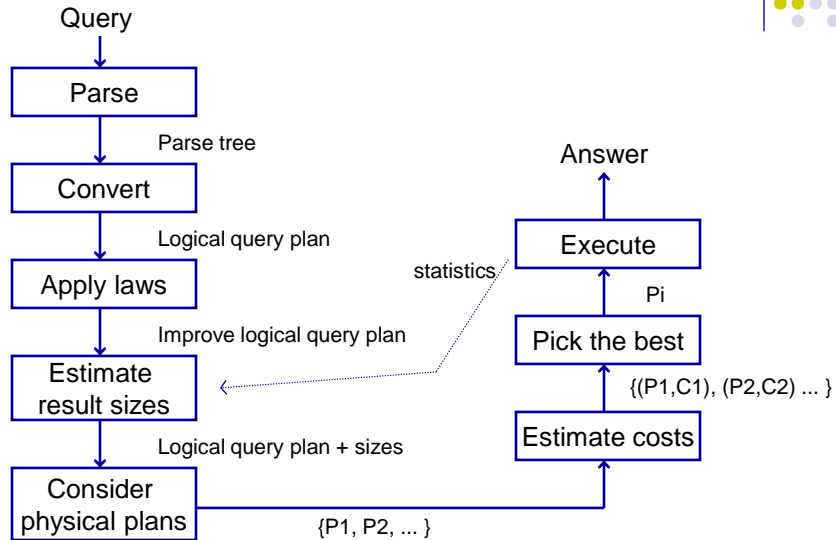


Tập trung vào RDBMS



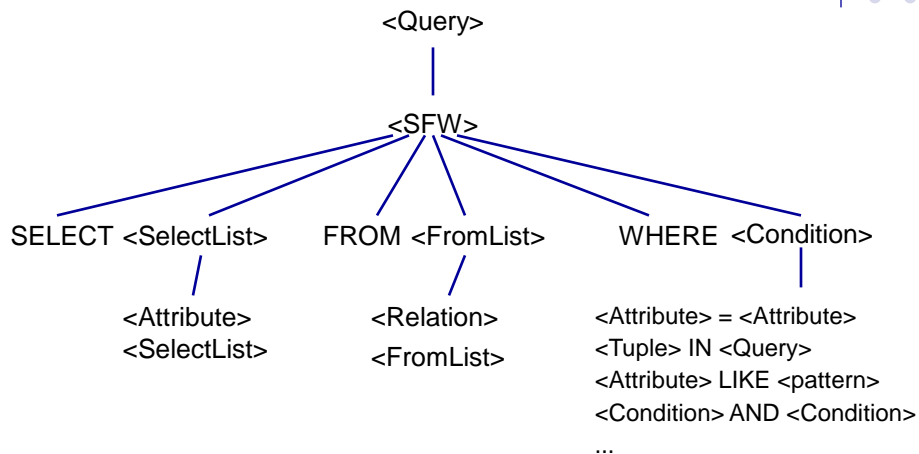
12

Quá trình biên dịch



13

Cây phân tích



14

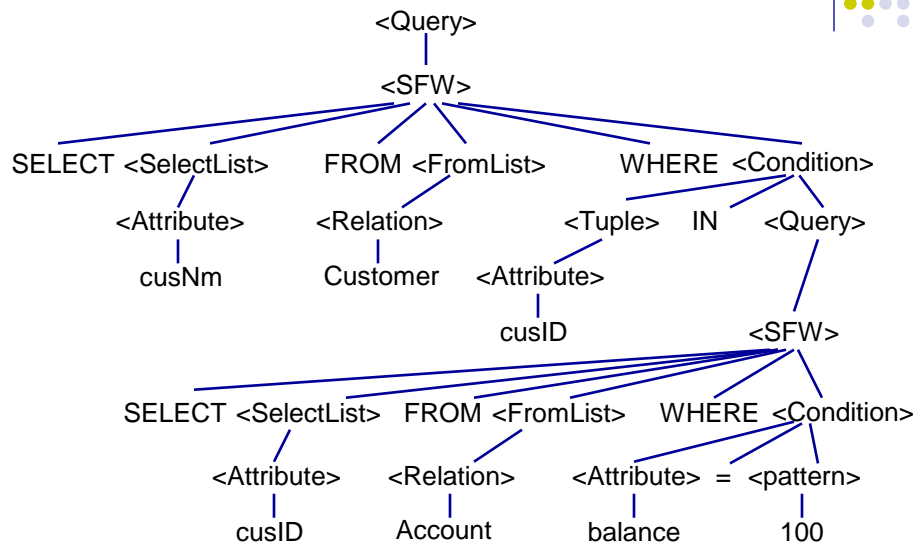
Ví dụ 1

- Customer(cusID, cusNm, cusStreet, cusCity)
- Account(acclID, cusID, balance)

```
SELECT cusNm
FROM Customer
WHERE cusID IN (
  SELECT cusID
  FROM Account
  WHERE balance > 100)
```

15

Ví dụ 1 (tt)



16

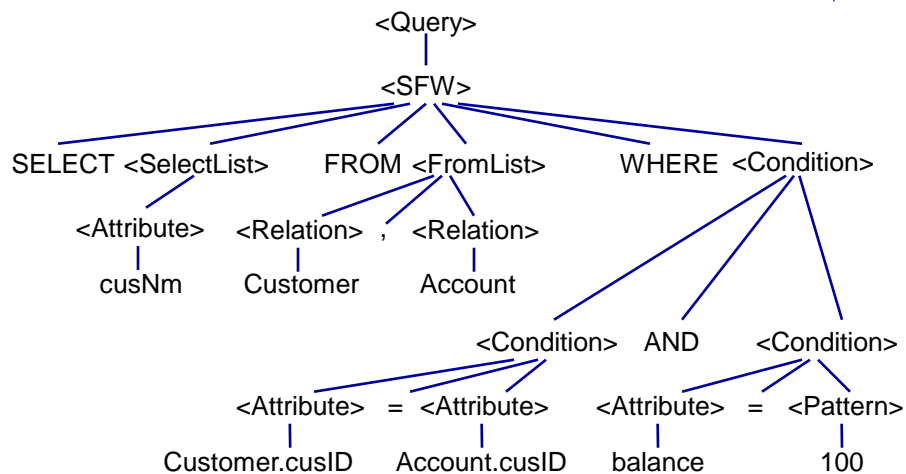
Ví dụ 2

- Customer(cusID, cusNm, cusStreet, cusCity)
- Account(acclD, cusID, balance)

```
SELECT cusNm
FROM Customer, Account
WHERE Customer.cusID = Account.cusID
AND balance = 100
```

17

Ví dụ 2 (tt)



18

Nhận xét



- Giới hạn
 - GROUP BY
 - HAVING
 - ORDER BY
 - DISTINCT
- Aggregation function (Max, Min, Count, Sum, Avg)
- Alias name

19

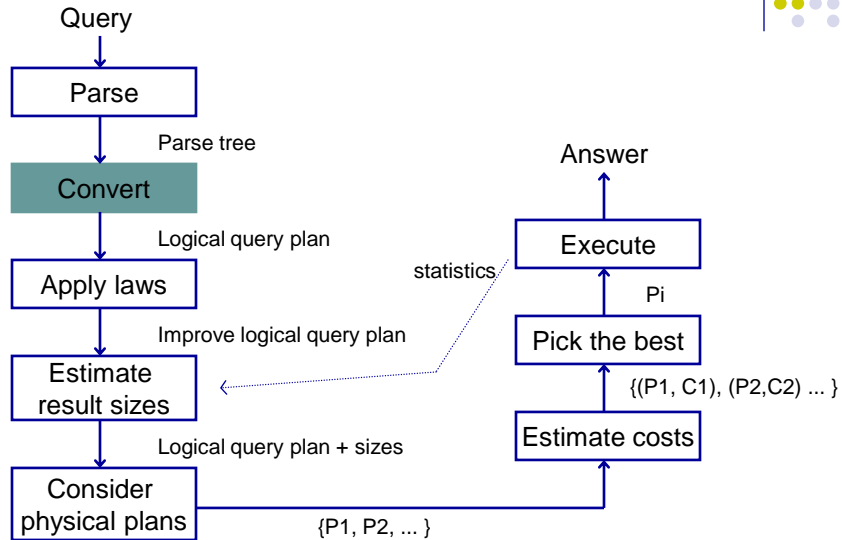
Tiền xử lý (preprocessing)



- Kiểm tra ngữ nghĩa
 - Quan hệ
 - Thuộc tính
 - Select
 - From
 - Kiểu dữ liệu
 - Where

20

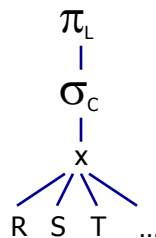
Quá trình biên dịch



21

Biến đổi sang ĐSQH

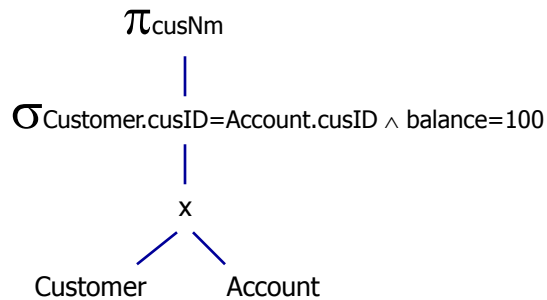
- Truy vấn đơn
 - Xét câu trúc <SFW>
 - Thay thế <FromList> thành các biến quan hệ
 - Sử dụng phép tích cartesian cho các biến quan hệ
 - Thay thế <Condition> thành phép chọn σ_C
 - Thay thế <SelectList> thành phép chiếu π_L



Cây truy vấn

22

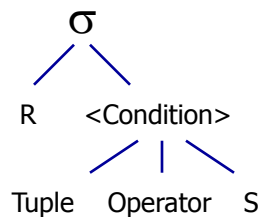
Xét ví dụ 2



23

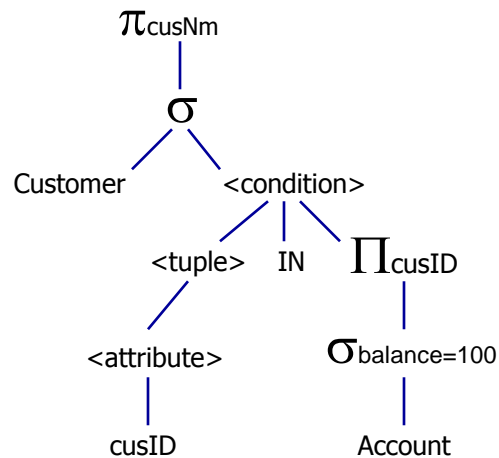
Biến đổi sang ĐSQH (tt)

- Truy vấn lồng
 - Tồn tại câu truy vấn con S trong <Condition>
 - Áp dụng qui tắc <SFW> cho truy vấn con
 - Phép chọn 2 biến (two-argument selection)
 - Nút là phép chọn không có tham số
 - Nhánh con trái là biến quan hệ R
 - Nhánh con phải là <condition> áp dụng cho mỗi bộ trong R



24

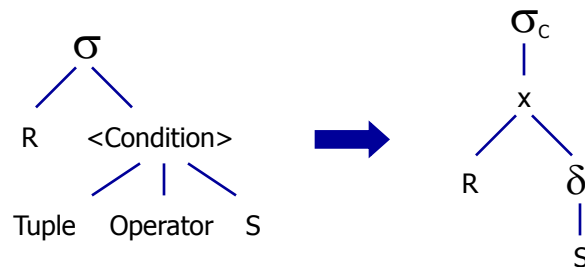
Xét ví dụ 1



25

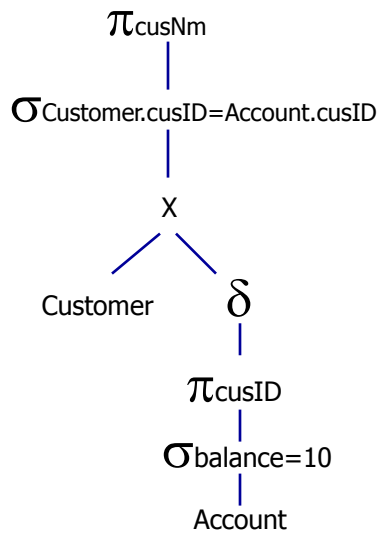
Biến đổi sang ĐSQH (tt)

- Truy vấn lồng
 - Biến đổi phép chọn 2 biến
 - Thay thế <Condition> bằng 1 cây có gốc là S
 - Nếu S có các bộ trùng nhau thì phải lược bỏ bớt bộ trùng nhau đi
 - Sử dụng phép δ
 - Thay thế phép chọn 2 biến thành σ_C
 - σ_C là kết quả của phép cartesian của R và S



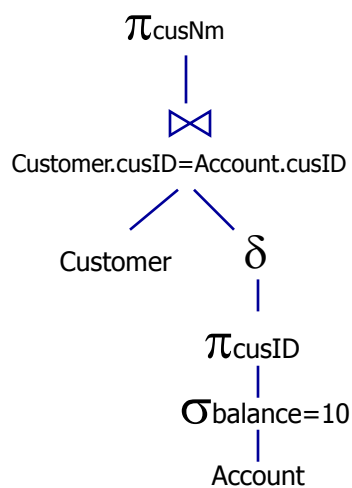
26

Xét ví dụ 1 (tt)



27

Xét ví dụ 1 (tt)



28

Ví dụ 3

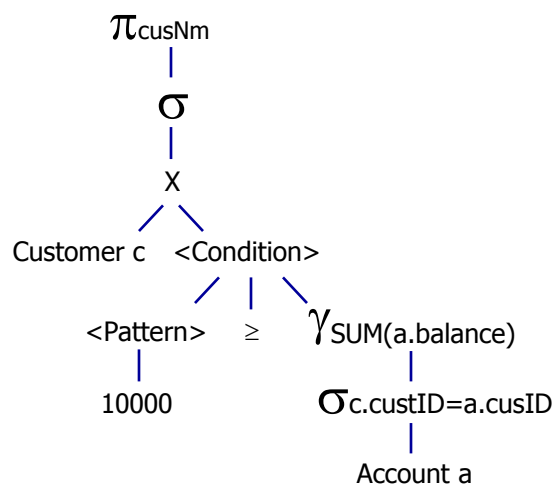
- Customer(cusID, cusNm, cusStreet, cusCity)
- Account(acclD, cusID, balance)

```
SELECT c.cusNm
FROM Customer c
WHERE 10000 >= (
    SELECT SUM(a.balance)
    FROM Account a
    WHERE a.cusID=c.cusID)
```

Truy vấn lồng tương quan

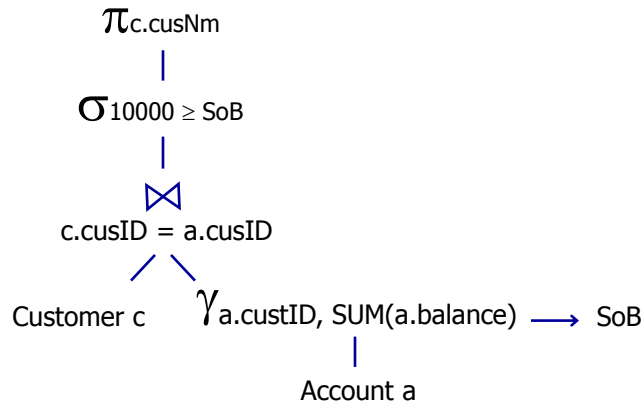
29

Ví dụ 3 (tt)



30

Ví dụ 3 (tt)



31

Qui tắc: Kết tự nhiên, tích cartesian, hội



$$\begin{aligned}
 R \bowtie S &= S \bowtie R \\
 (R \bowtie S) \bowtie T &= R \bowtie (S \bowtie T)
 \end{aligned}$$

$$\begin{aligned}
 R \times S &= S \times R \\
 (R \times S) \times T &= R \times (S \times T)
 \end{aligned}$$

$$\begin{aligned}
 R \cup S &= S \cup R \\
 R \cup (S \cup T) &= (R \cup S) \cup T
 \end{aligned}$$

32

Qui tắc: Phép chọn σ



- Cho
 - p là vị từ chỉ có các thuộc tính của R
 - q là vị từ chỉ có các thuộc tính của S
 - m là vị từ có các thuộc tính của R và S

Pushing selections

$$\sigma_{p1 \wedge p2}(R) = \sigma_{p1} [\sigma_{p2}(R)]$$

$$\sigma_{p1 \vee p2}(R) = [\sigma_{p1}(R)] \cup [\sigma_{p2}(R)]$$

Quan hệ R là tập hợp
 \cup_S là phép hội trên tập hợp

33

Qui tắc: σ, \bowtie



$$\sigma_p(R \bowtie S) = [\sigma_p(R)] \bowtie S$$

$$\sigma_q(R \bowtie S) = R \bowtie [\sigma_q(S)]$$

34

Qui tắc: σ, \bowtie (tt)



$$\sigma_{p \wedge q}(R \bowtie S) = [\sigma_p(R)] \bowtie [\sigma_q(S)]$$

$$\sigma_{p \wedge q \wedge m}(R \bowtie S) = \sigma_m[\sigma_p(R) \bowtie \sigma_q(S)]$$

$$\sigma_{p \vee q}(R \bowtie S) = [\sigma_p(R) \bowtie S] \cup [R \bowtie \sigma_q(S)]$$

35

Qui tắc: σ, \cup và $\sigma, -$



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

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

36

Qui tắc: Phép chiếu π



- Cho
 - X = tập thuộc tính con của R
 - Y = tập thuộc tính con của R
- Ta có
 - $XY = X \cup Y$

$$\pi_{XY}(R) = \pi_X[\cancel{\pi_Y(R)}]$$

37

Qui tắc: π, \bowtie



- Cho
 - X = tập thuộc tính con của R
 - Y = tập thuộc tính con của S
 - Z = tập giao thuộc tính của R và S

Pushing projections

$$\pi_{XY}(R \bowtie S) = \pi_{XY}[\pi_{XZ}(R) \bowtie \pi_{YZ}(S)]$$

Except intersection and difference

38

Qui tắc: σ, π



- Cho
 - X = tập thuộc tính con của R
 - Z = tập thuộc tính con của R xuất hiện trong vị từ p

$$\pi_X[\sigma_p(R)] = \pi_X\{\sigma_p[\pi_{XZ}(R)]\}$$

39

Qui tắc: σ, π, \bowtie



- Cho
 - X = tập thuộc tính con của R
 - Y = tập thuộc tính con của S
 - Z = tập giao thuộc tính của R và S
 - $Z' = Z \cup \{\text{các thuộc tính xuất hiện trong vị từ } p\}$

$$\pi_{XY}[\sigma_p(R \bowtie S)] =$$

$$\pi_{XY}\{\sigma_p[\pi_{XZ'}(R) \bowtie \pi_{YZ'}(S)]\}$$

40

Nhận xét: σ , π



- Ví dụ
 - $R(A, B, C, D, E)$
 - $X=\{E\}$
 - $p: A=3 \wedge B='a'$

$$\pi_X [\sigma_p (R)]$$

Chọn trước
tốt hơn???



$$\pi_E \{ \sigma_p [\pi_{ABE}(R)] \}$$

Chiếu trước
tốt hơn???

41

Nhận xét: σ , π (tt)



- Bình thường
 - Chiếu trước
- Nhưng
 - Giả sử A và B được cài đặt chỉ mục (index)
 - Physical query plan dùng chỉ mục để chọn ra những bộ có $A=3$ và $B='a'$ trước
 - Nếu thực hiện chiếu trước $\pi_{AB}(R)$ thì chỉ mục trên A và B là vô ích
 - Chọn trước

→ Thông thường chọn trước tốt hơn

42

Qui tắc: \times, \bowtie



$$\sigma_C(R \bowtie S) = R \bowtie_C S$$

$$R \times S = \pi_L[\sigma_C(R \times S)]$$

43

Qui tắc: δ



$$\delta(R \bowtie S) = \delta(R) \bowtie \delta(S)$$

$$\delta(R \times S) = \delta(R) \times \delta(S)$$

$$\delta[\sigma_C(R)] = \sigma_C[\delta(R)]$$

$$\begin{aligned} \delta(R \cap_B S) &= \delta(R) \cap_B S = R \cap_B \delta(S) \\ &= \delta(R) \cap_B \delta(S) \end{aligned}$$

Except: \cup_B, \neg_B, π

44

Qui tắc: γ



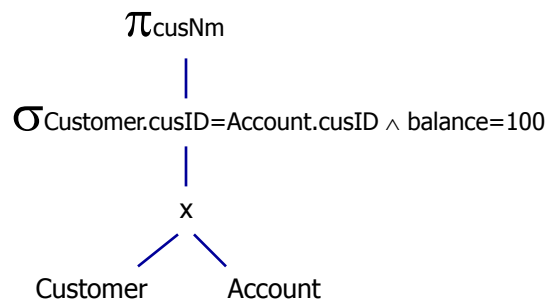
- Cho
 - X = tập thuộc tính trong R được gom nhóm
 - $Y = X \cup \{\text{một số thuộc tính khác của } R\}$

$$\delta[\gamma_X(R)] = \gamma_X(R)$$

$$\gamma_X(R) = \gamma_X[\pi_Y(R)]$$

45

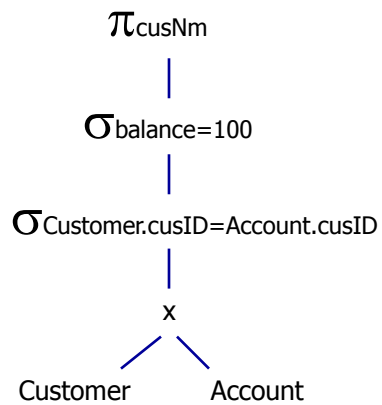
Xét ví dụ 2



46

Xét ví dụ 2

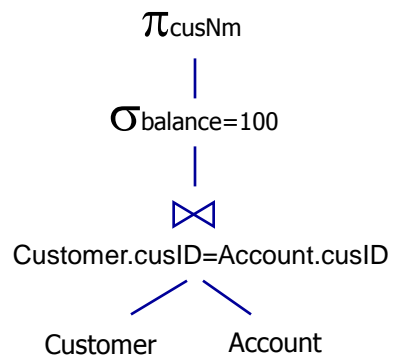
Qui tắc σ



47

Xét ví dụ 2 (tt)

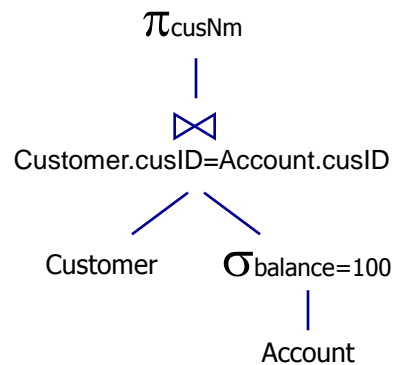
Qui tắc σ, \bowtie



48

Xét ví dụ 2 (tt)

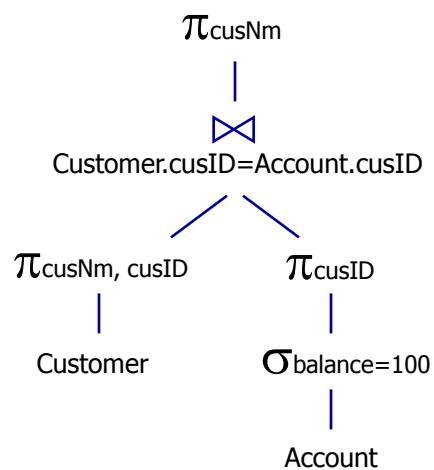
Pushing σ



49

Xét ví dụ 2 (tt)

Pushing π



50

Tóm tắt: Các quy tắc biến đổi tương đương



QT1: Xử lý các toán tử AND trong điều kiện

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

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

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

\equiv

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

51

Tóm tắt: Các quy tắc biến đổi tương đương



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

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

52

Tóm tắt: Các quy tắc biến đổi tương đương



QT3: Xử lý các phép chiếu

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

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

$$\begin{aligned} & \pi_{\text{manv, honv, tennv}} (\pi_{\text{manv, honv, tennv, ngaysinh}} (\text{NHANVIEN})) \\ & \equiv \\ & \pi_{\text{manv, honv, tennv}} (\text{NHANVIEN}) \end{aligned}$$

53

Tóm tắt: Các quy tắc biến đổi tương đương



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

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

Nếu như $c \subset [A1 \dots An]$

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

$$\begin{aligned} & \pi_{\text{manv, honv, tennv, phai}} (\sigma_{\text{phai} = \text{'NAM'}} (\text{NHANVIEN})) \\ & \equiv \\ & \sigma_{\text{phai} = \text{'NAM'}} (\pi_{\text{manv, honv, tennv, phai}} (\text{NHANVIEN})) \end{aligned}$$

54

Tóm tắt: Các quy tắc biến đổi tương đương



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)

$$\begin{aligned} & (NHANVIEN \bowtie PHONGBAN) \\ \equiv & \quad NV.maphong = PB.maphong \\ & (PHONGBAN \bowtie NHANVIEN) \\ & \quad NV.maphong = PB.maphong \end{aligned}$$

55

Tóm tắt: Các quy tắc biến đổi tương đương



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)

$$\begin{aligned} & \sigma_{\text{phai} = \text{'NAM'}}(NHANVIEN \bowtie PHONGBAN) \\ \equiv & \quad (\sigma_{\text{phai} = \text{'NAM'}}(NHANVIEN)) \bowtie PHONGBAN \end{aligned}$$

56

Tóm tắt: Các quy tắc biến đổi tương đương



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 $c = c_1 \text{ and } c_2$, ($c_1 \in R$ và $c_2 \in S$)

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

PHONGBAN (maphong, tenphong, maql)

$$\begin{aligned} & \sigma_{\text{phai} = \text{'NAM'} \text{ AND } \text{tenphong} = \text{'Kế toán'}}(\text{NHANVIEN} \bowtie \text{PHONGBAN}) \\ & \equiv \\ & (\sigma_{\text{phai} = \text{'NAM'}}(\text{NHANVIEN})) \bowtie (\sigma_{\text{tenphong} = \text{'Kế toán'}}(\text{PHONGBAN})) \end{aligned}$$

57

Tóm tắt: Các quy tắc biến đổi tương đương



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)

$$\begin{aligned} & \pi_{\text{manv}, \text{tennv}, \text{maphong}, \text{tenphong}}(\text{NHANVIEN} \bowtie \text{PHONGBAN}) \\ & \equiv \text{NV.maphong=PB.maphong} \\ & (\pi_{\text{manv}, \text{honv}, \text{maphong}}(\text{NHANVIEN})) \bowtie (\pi_{\text{tenphong}, \text{maphong}}(\text{PHONGBAN})) \\ & \text{NV.maphong=PB.maphong} \end{aligned}$$

58

Tóm tắt: Các quy tắc biến đổi tương đương



QT7b: Phân phối giữa phép chiếu và phép kết

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

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

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

PHONGBAN (maphong, tenphong, maql)

$$\begin{aligned} & \pi_{\text{manv, tennv, tenphong}} (\text{NHANVIEN} \bowtie \text{PHONGBAN}) \\ & \equiv \\ & (\pi_{\text{manv, tennv, maphong}} (\text{NHANVIEN})) \bowtie (\pi_{\text{tennv, maphong}} (\text{PHONGBAN})) \end{aligned}$$

NV.maphong=PB.maphong

59

Tóm tắt: Các quy tắc biến đổi tương đương



QT8: Giao hoán của phép hội và phép giao

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

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

60

Tóm tắt: Các quy tắc biến đổi tương đương



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

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

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

61

Tóm tắt: Các quy tắc biến đổi tương đương



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

62

Tóm tắt: Các quy tắc biến đổi tương đương



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

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

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

63

Tóm tắt: Các quy tắc biến đổi tương đương



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

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

Luật De Morgan

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

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

64

Tối ưu hóa: Giải thuật heuristic



1. Áp dụng QT 1, tách các phép chọn liên tiếp 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.

65

Ví dụ: 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

```

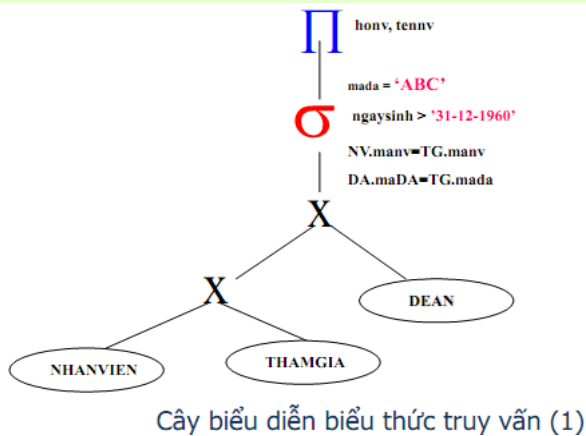
$$\Pi_{honv, tennv} (\sigma_{mada = 'ABC' \wedge ngaysinh > '31-12-1960' \wedge NV.manv=TG.manv \wedge DA.mada=TG.mada} (NHANVIEN \times DEAN \times THAMGIA))$$

```

66

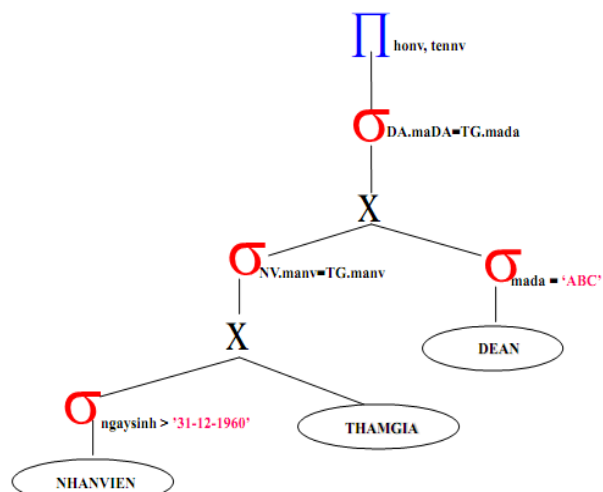
Ví dụ: Giải thuật heuristic

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



67

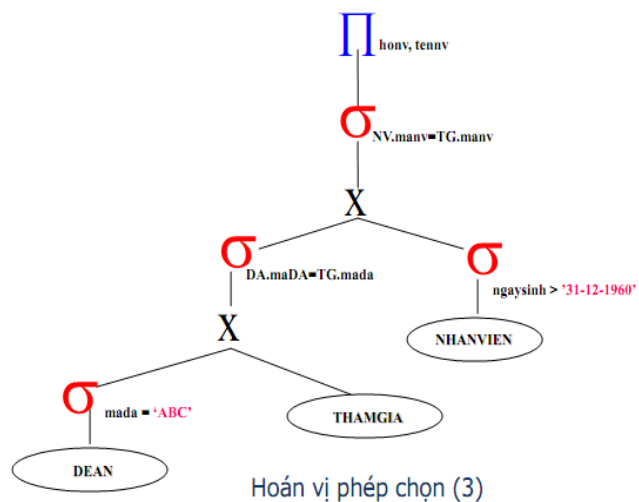
Ví dụ: Giải thuật heuristic



Đưa phép chọn xuống sâu các nhánh (2)

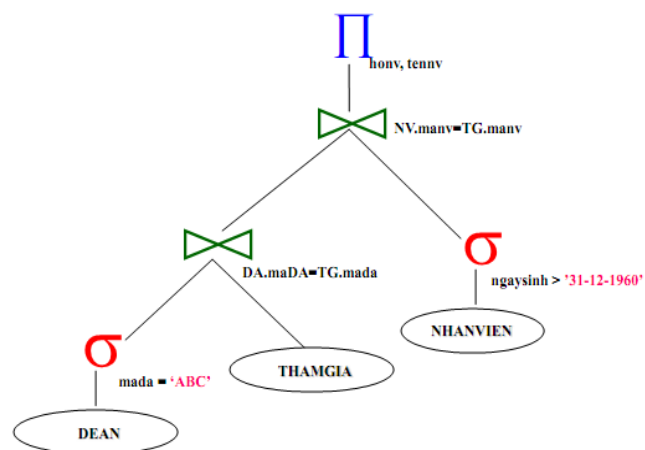
68

Ví dụ: Giải thuật heuristic



69

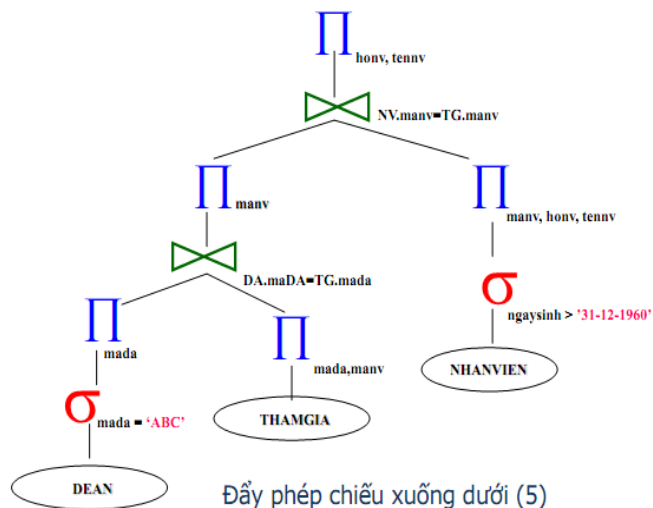
Ví dụ: Giải thuật heuristic



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

70

Ví dụ: Giải thuật heuristic



71

Ví dụ: Giải thuật heuristic

Ngôn ngữ ĐSQH

$$\pi_{\text{honv, tennv}}((\pi_{\text{mada}}(\sigma_{\text{mada='ABC'}}(\text{DEAN}))) \bowtie (\pi_{\text{mada, manv}}(\text{THAMGIA}))) \bowtie (\pi_{\text{manv, honv, tennv}}(\sigma_{\text{ngaysinh > '31-12-1960'}}(\text{NHANVIEN})))$$

DA.mada=TG.mada

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
```

72

Ước lượng chi phí



- Ước lượng kích thước cây truy vấn
 - Quan hệ
 - Các phép toán
- Ước lượng số lần truy xuất IOs
 - Số blocks được đọc hoặc ghi để thực hiện cây truy vấn

73

Ước lượng kích thước



- Thống kê quan hệ R
 - $T(R)$: số bộ trong R
 - $S(R)$: tổng số byte của 1 bộ trong R
 - $B(R)$: tổng số block chứa tất cả các bộ của R
 - $V(R, A)$: số giá trị khác nhau mà thuộc tính A trong R có thể có

74

Ví dụ

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)

$$T(R) = 5$$

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

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

$$S(R) = 100$$

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

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

$$B(R) = 1$$

75

Ước lượng: $W = R_1 \times R_2$

$$S(W) = S(T_1) + S(T_2)$$

$$T(W) = T(R_1) \times T(T_2)$$

76

Ước lượng: $W = \sigma_{Z = \text{val}}(R)$



$$S(W) = S(R)$$

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

Số bộ trung bình thỏa điều kiện $Z = \text{val}$

77

Ước lượng: $W = \sigma_{Z \geq \text{val}}(R)$



$$T(W) = ???$$

- Cách 1

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

- Cách 2

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

78

Ví dụ



- Cho
 - $R(A, B, C)$
 - $T(R) = 10000$
 - $V(R, A) = 50$
- Ước lượng kích thước biểu thức

$$S = \sigma_{A=10 \wedge B < 20}(R)$$

$$T(S) = \frac{T(R)}{V(R, A) \times 3} = \frac{10000}{50 \times 3} = 67$$

79

Ví dụ (tt)



- Ước lượng kích thước biểu thức

$$S = \sigma_{A=10 \vee B < 20}(R)$$

- Giả sử
 - n là $T(R)$
 - m_1 là số bộ thỏa $A=10$ trong R
 - m_2 là số bộ thỏa $B < 20$ trong R

$$T(S) = n(1 - (1 - \frac{m_1}{n})(1 - \frac{m_2}{n}))$$

80

Ước lượng: $W = R_1 \bowtie R_2$



- Cho
 - X = tập thuộc tính của R_1
 - Y = tập thuộc tính của R_2
- Xét trường hợp $X \cap Y = \emptyset$

$$T(W) = ?$$

Tương tự $R_1 \times R_2$

81

Ước lượng: $W = R_1 \bowtie R_2$ (tt)



- Xét trường hợp $X \cap Y = A$

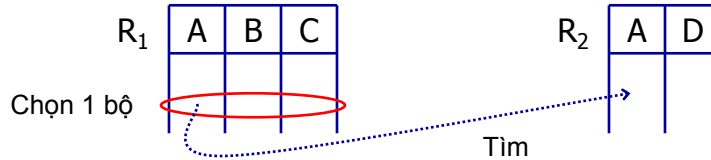
R_1	A	B	C

R_2	A	D

- Giả sử
 - $V(R_1, A) \leq V(R_2, A)$
 - Mọi giá trị của A trong R_1 thì có trong R_2
 - $V(R_2, A) \leq V(R_1, A)$
 - Mọi giá trị của A có trong R_2 thì có trong R_1

82

Ước lượng: $W = R_1 \bowtie R_2$ (tt)



1 bộ trong R_1 sẽ thỏa với $\frac{T(R_2)}{V(R_2, A)}$ bộ trong R_2

$$T(W) = T(R_1) \times \frac{T(R_2)}{V(R_2, A)}$$

83

Ước lượng: $W = R_1 \bowtie R_2$ (tt)



- $V(R_1, A) \leq V(R_2, A)$

$$T(W) = T(R_1) \times \frac{T(R_2)}{V(R_2, A)}$$

- $V(R_2, A) \leq V(R_1, A)$

$$T(W) = T(R_2) \times \frac{T(R_1)}{V(R_1, A)}$$

- Tổng quát

$$T(W) = \frac{T(R_1) T(R_2)}{\max\{V(R_1, A), V(R_2, A)\}}$$

84

Ước lượng: $W = R_1 \bowtie R_2$ (tt)



- Xét trường hợp $X \cap Y = A$

 R_1

A	B	C

 R_2

A	D

- $W(A, B, C, D)$
 - Các thuộc tính không tham gia vào phép kết thì số lượng các giá trị vẫn giữ nguyên
 - $V(W, A) = \min \{V(R_1, A), V(R_2, A)\}$
 - $V(W, B) = V(R_1, B)$
 - $V(W, C) = V(R_1, C)$
 - $V(W, D) = V(R_2, D)$

85

Ví dụ



$$Z = R_1(A, B) \bowtie R_2(B, C) \bowtie R_3(C, D)$$

R_1	R_2	R_3
$T(R_1) = 1000$	$T(R_2) = 2000$	$T(R_3) = 3000$
$V(R_1, A) = 50$	$V(R_2, B) = 200$	$V(R_3, C) = 90$
$V(R_1, B) = 100$	$V(R_2, C) = 300$	$V(R_3, D) = 500$

86

Ví dụ (tt)



$$U = R_1(A, B) \bowtie R_2(B, C)$$

$$T(U) = \frac{1000 \times 2000}{200}$$

$$V(U, A) = 50$$

$$V(U, B) = 100$$

$$V(U, C) = 300$$

87

Ví dụ (tt)



$$Z = U \bowtie R_3(C, D)$$

$$T(Z) = \frac{1000 \times 2000 \times 3000}{200 \times 300}$$

$$V(Z, A) = 50$$

$$V(Z, B) = 100$$

$$V(Z, C) = 90$$

$$V(Z, D) = 500$$

88

Nhận xét



- Phép chiếu
 - Phép tích
- } Ước lượng chính xác
- Phép chọn
 - Phép kết
- } Ước lượng tương đối hợp lý
- số lượng bộ của các quan hệ tương đối lớn
- giá trị của các thuộc tính phân bố đồng đều
- Phép toán khác
 - Hội
 - Giao
 - Trừ
 - Lược bỏ trùng lặp
 - Gom nhóm

89

Ước lượng: $W = R_1 \cup R_2$



- R_1 và R_2 là bag

$$T(W) = T(R_1) + T(R_2)$$

- R_1 và R_2 là set

$$T'(W) = T(R_1) + T(R_2)$$

$$T''(W) \leq T(R_1) + T(R_2)$$

$$\rightarrow T(W) = \frac{T'(W) + T''(W)}{2}$$

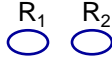
90

Ước lượng: $W = R_1 \cap R_2$



- Cách 1

- TH1: $T'(W)=0$



- TH2: $T''(W) = T(R_1)$ hoặc $T''(W) = T(R_2)$



$$\rightarrow T(W) = \frac{T'(W) + T''(W)}{2}$$

- Cách 2

- Trường hợp đặc biệt của phép kết tự nhiên

- Chỉ áp dụng cho \cap_s

$$T(W) = \frac{T(R_1) T(R_2)}{\max\{V(R_1, Z), V(R_2, Z)\}}$$

91

Ước lượng: $W = R_1 - R_2$



- TH1: $T(W) = T(R_1)$

- TH2: $T(W) = T(R_1) - T(R_2)$

$$\rightarrow T(W) = T(R_1) - \frac{1}{2} T(R_2)$$

92

Ước lượng: $W = \delta(R)$



- TH1: $T(W) = 1$
 - Nếu trong R không có bộ nào thì $T(W)=0$
- TH2: $T(W) = T(R)$
 - $R(a_1, a_2, \dots, a_n)$
 - Số bộ phân biệt tối đa của R là tích các $V(R, a_i), i=1..n$

$$\rightarrow T(W) = \min\left\{\frac{1}{2} T(R_1), \text{tích các } V(R, a_i)\right\}$$

93

Ước lượng: $W = \gamma(R)$

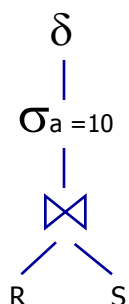


- $\gamma_L(R)$
 - Số lượng bộ trong W và cũng là số lượng nhóm
- TH1: $T(W) = 1$
- TH2: $T(W) = T(R)$
 - $R(a_1, a_2, \dots, a_n)$
 - Số lượng nhóm tối đa là tích các $V(R, a_i), i=1..n$

$$\rightarrow T(W) = \min\left\{\frac{1}{2} T(R_1), \text{tích các } V(R, a_i)\right\}$$

94

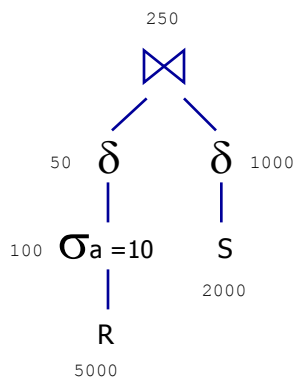
Ví dụ



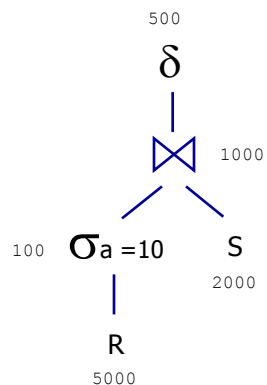
- $R(a, b)$
 - $T(R)=5000$
 - $V(R, a)=50$
 - $V(R, b)=100$
- $S(b, c)$
 - $T(S)=2000$
 - $V(S, b)=200$
 - $V(S, c)=100$

95

Ví dụ (tt)



(1)



(2)

96

Ví dụ (tt)



- Cộng kích thước sau khi thực hiện các phép toán, ngoại trừ
 - Các nút lá
 - Nút gốc
- (1): $100+50+1000=1150$
- (2): $100+1000=1100$
- Phép lược bỏ trùng lặp thực hiện sau thì tốt hơn

97

Ước lượng số lần truy xuất IOs



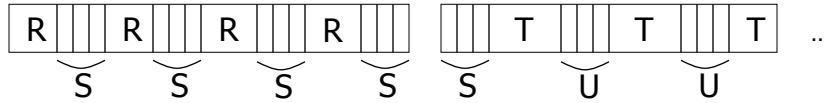
- Các tham số thống kê
 - $B(R)$: tổng số block chứa tất cả các bộ của R
 - $f(R)$: số bộ tối đa trong mỗi block
 - M: số block trống trên bộ nhớ
- Quan tâm
 - Quan hệ R có được gom thành cụm không (clustered)?
 - Thuộc tính trong các phép toán có chỉ mục không (index)?
 - Chỉ mục có gom cụm không (clustering index)?
 - Kết quả cần được sắp thứ tự không?

98

Clustering



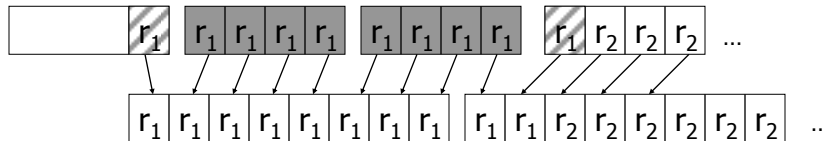
- Clustered-file



- Clustered relation



- Clustering index



99

Ví dụ



- $R_1 \bowtie R_2$

- $T(R_1) = 10000$
- $T(R_2) = 5000$
- $S(R_1) = S(R_2) = 1/10$ block
- $M=101$ blocks

- Số block được đọc (bỏ qua việc ghi) để thực hiện phép kết tự nhiên trên là bao nhiêu?

100

Nhận xét



- Ước lượng số lần truy xuất IOs không là cách tốt nhất
 - Bỏ qua chi phí CPU
 - Bỏ qua tham số thời gian
 - Xét trường hợp M đủ hoặc thiếu

101

Bài tập



Hãy trình bày tiến trình tối ưu hóa câu truy vấn sau:

```
SELECT  CTPX.MSMH, TENHG,SL
FROM    KH, PX, CTPX,MH
WHERE    PX.SP = CTPX.SP
           AND PX.MSKH = KH.MSKH
           AND CTPX.MSMH = MH.MSMH
           AND TENKH="Cty ABD"
           AND  SL > 100
```

102