



PA152: Efektivní využívání DB
6. Zpracování dotazů

Vlastislav Dohnal

Poděkování

- Zdrojem materiálů tohoto předmětu jsou:
 - Přednášky CS245, CS345, CS345
 - Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom
 - Stanford University, California

Vyhodnocení dotazu

- Dotaz
- Strom dotazu
- Logický plán
- Úpravy
- Fyzický plán
- Vyhodnocení

Příklad

- Select B,D

From R,S

Where $R.C=S.C$ and $R.A="c"$ and $S.E=2$

Příklad

R	A	B	C	S	C	D	E
	a	1	10		10	x	2
	b	1	20		20	y	2
	c	2	10		30	z	2
	d	2	35		40	x	1
	e	3	45		50	y	3

Příklad

R	A	B	C	S	C	D	E
a	1	10	10	10	x	2	
b	1	20	20	20	y	2	
c	2	10	30	30	z	2	
d	2	35	40	40	x	1	
e	3	45	50	50	y	3	

Odpověď

B	D
2	x

Jak vyhodnotit tento dotaz?

1. způsob

- Kartézský součin
- Výběr záznamů
- Projekce

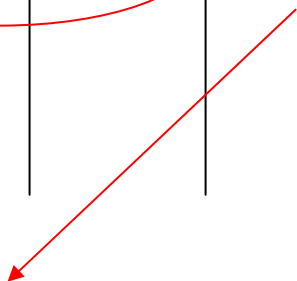
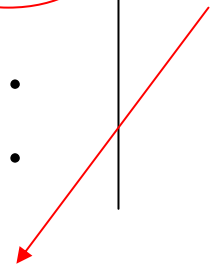
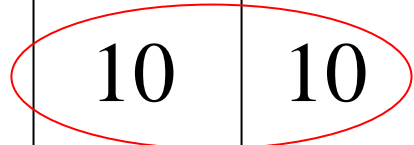
RXS

R.A	R.B	R.C	S.C	S.D	S.E
a	1	10	10	x	2
a	1	10	20	y	2
.					
.					
C	2	10	10	x	2
.					
.					

RXS

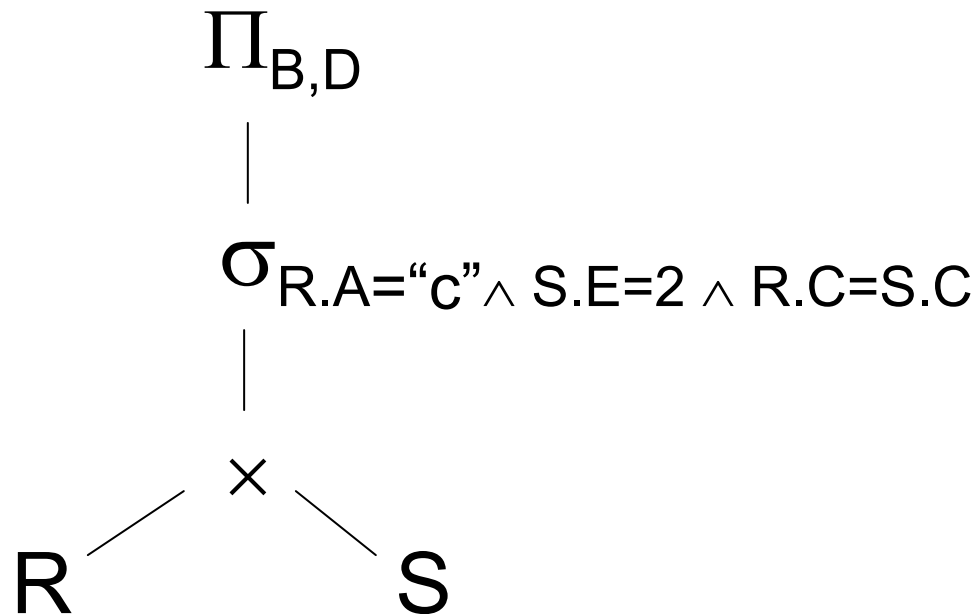
R.A	R.B	R.C	S.C	S.D	S.E
a	1	10	10	x	2
a	1	10	20	y	2
.					
.					
C	2	10	10	x	2
.					
.					

Tento záznam
vyhovuje



Popis plánů provedení dotazu

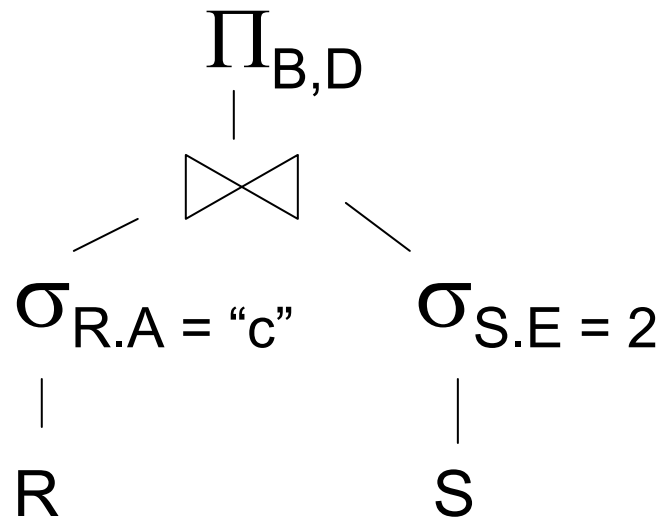
- Použití relační algebry
- Příklad plánu I:

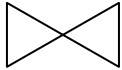


Alternativně: $\Pi_{B,D} [\sigma_{R.A="c" \wedge S.E=2 \wedge R.C=S.C} (R \times S)]$

Popis plánů provedení dotazu

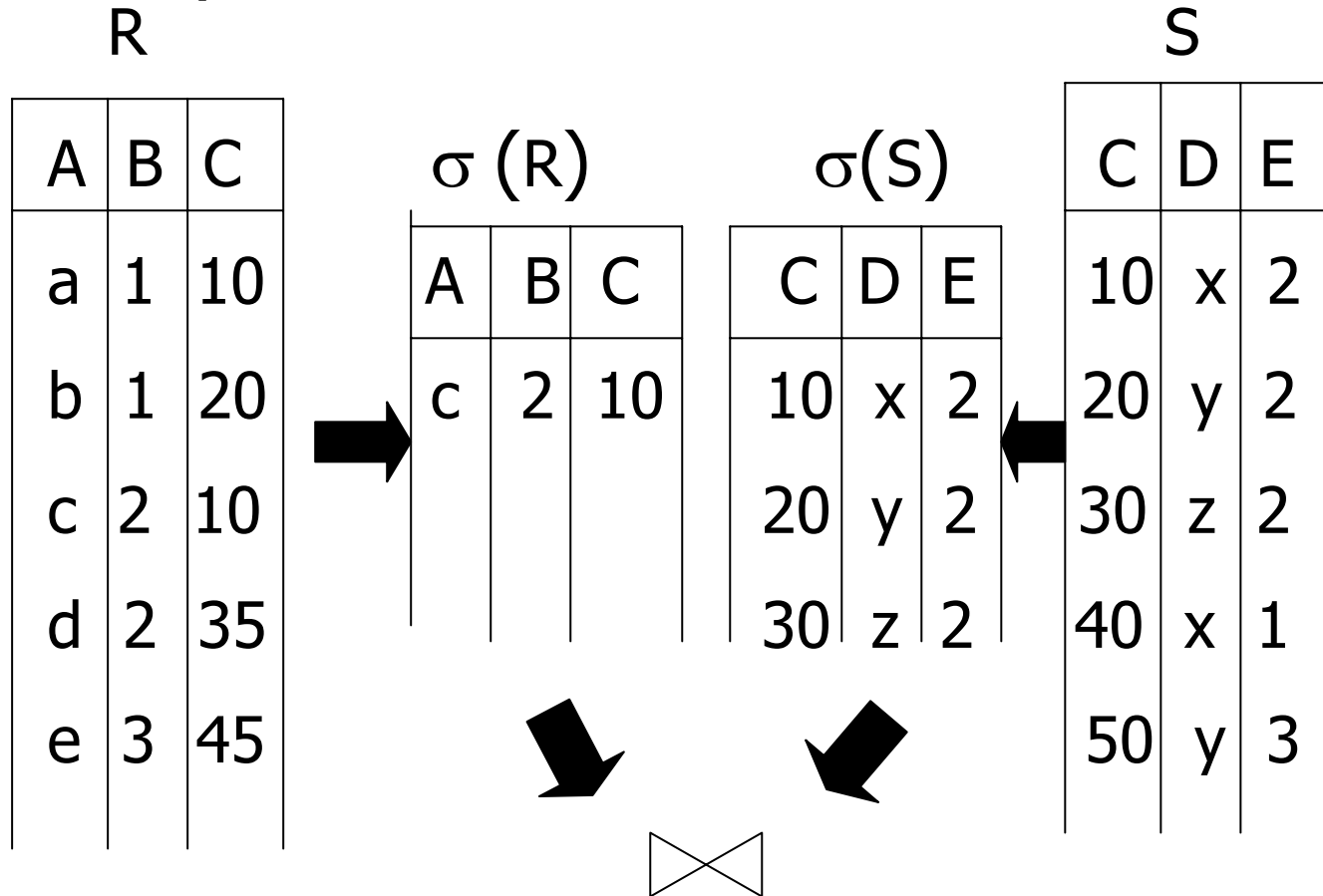
- Příklad plánu II:




natural join

Fyzický plán

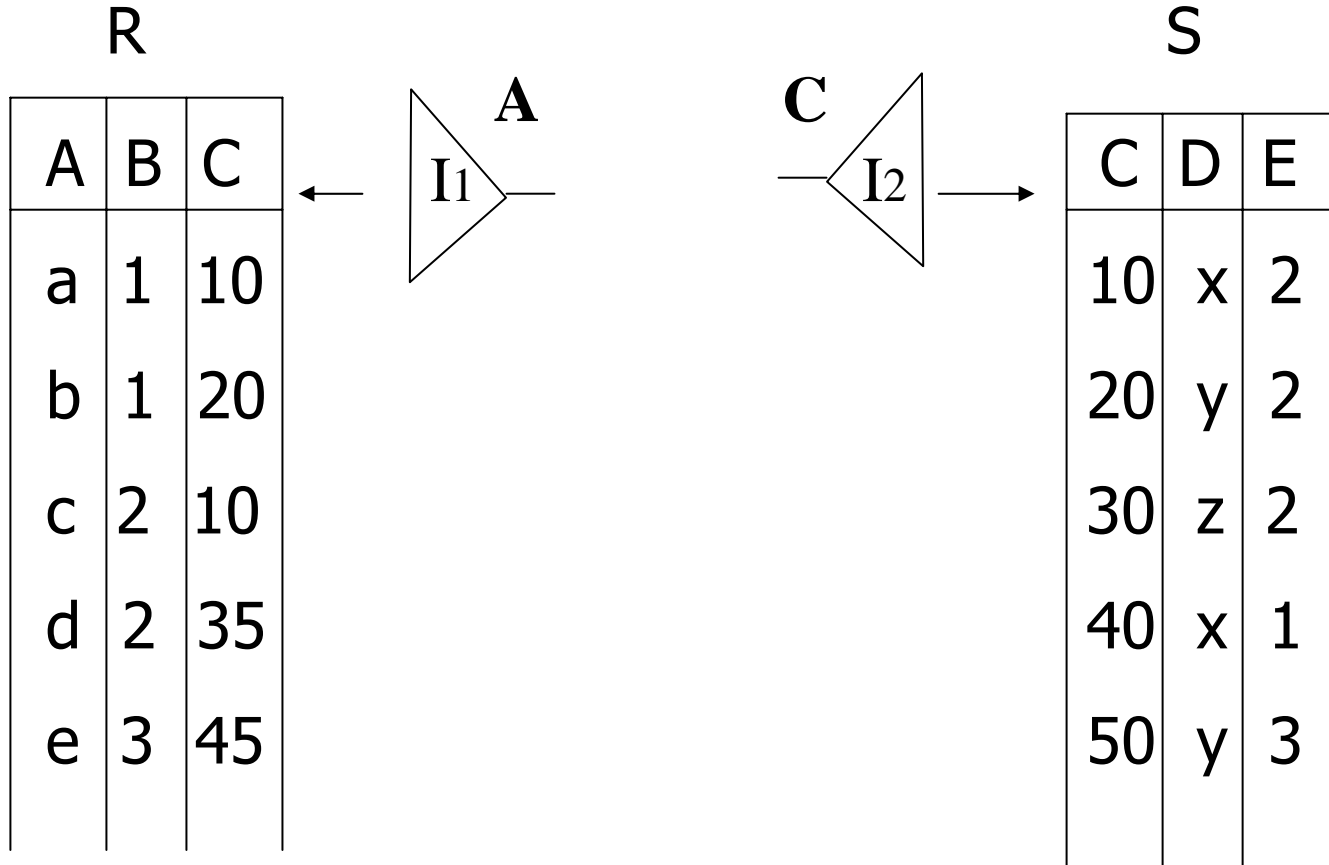
■ Příklad plánu II:

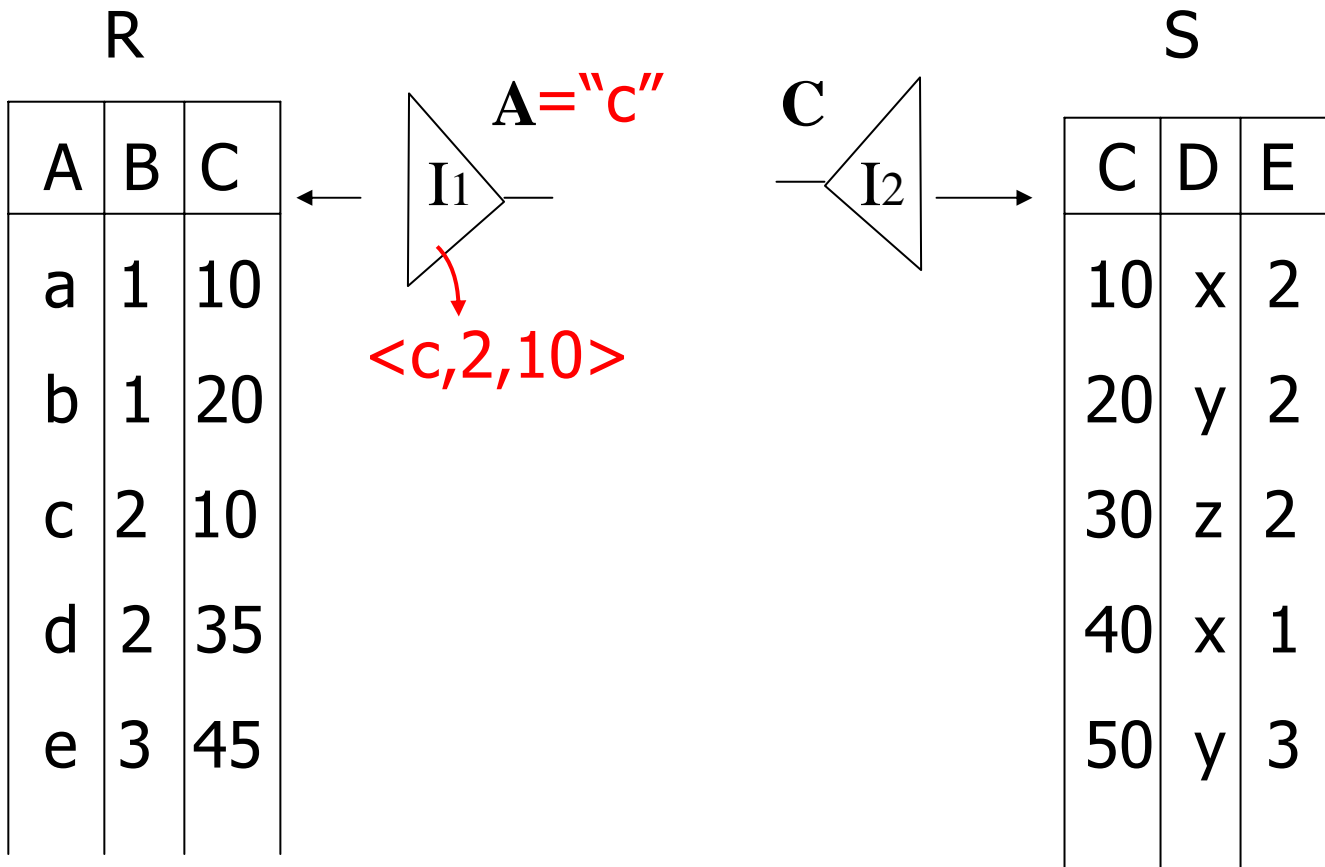


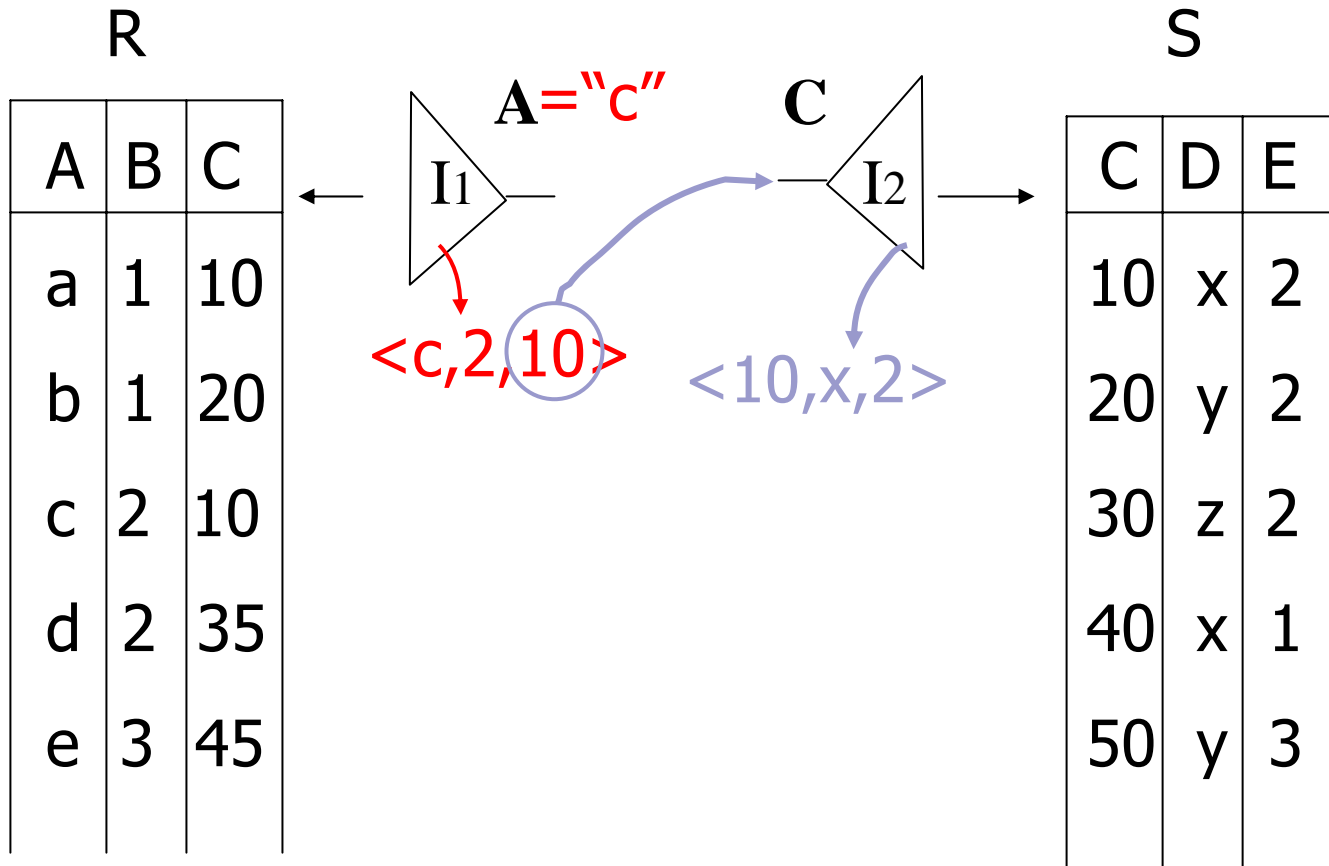
Popis plánů provedení dotazu

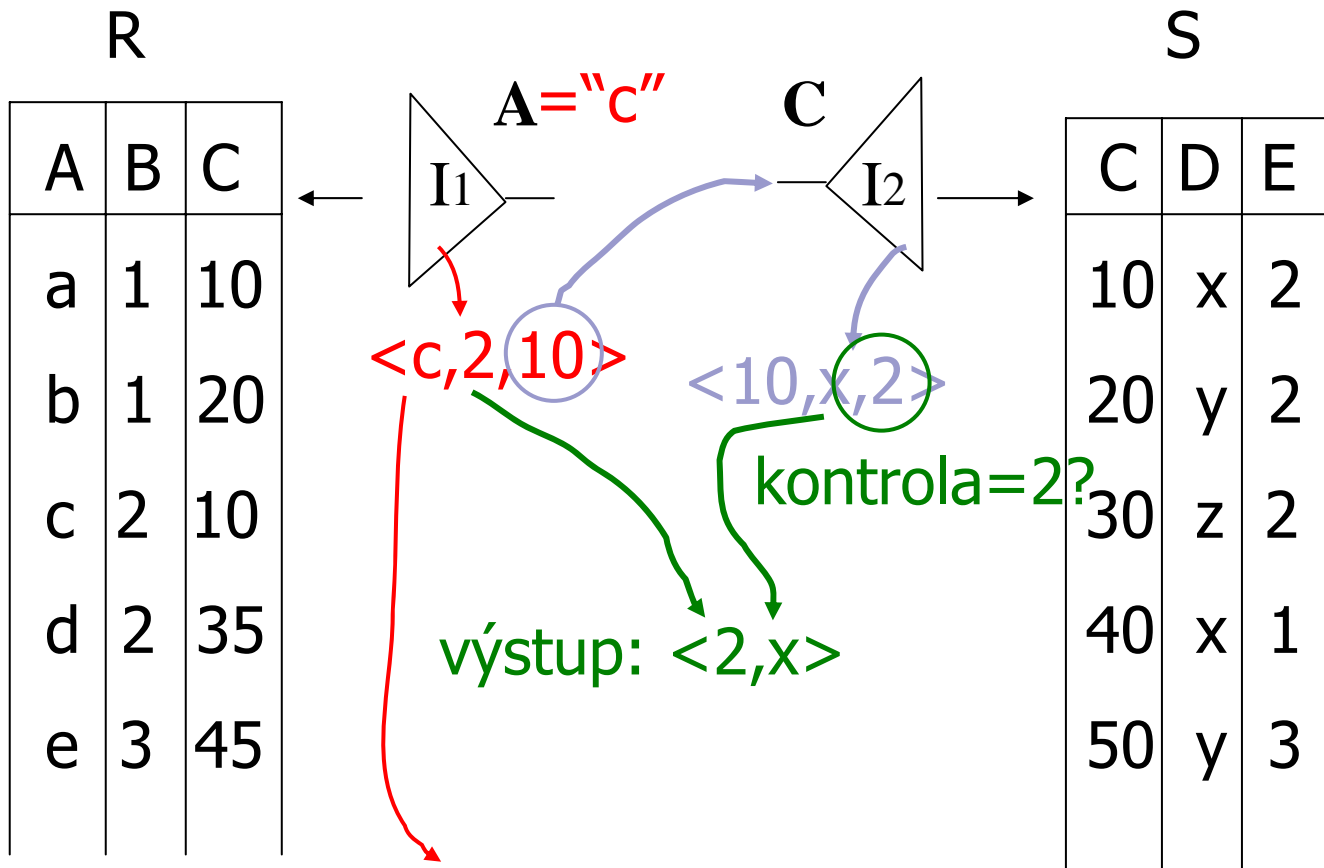
■ Plán III:

- Použití indexů pro R.A a S.C
- Použijeme index R.A k nalezení záznamů R splňující R.A = “c”
- Pro každou nalezenou hodnotu R.C použijeme index S.C pro nalezení odpovídajících záznamů
- Vypustíme záznamy S, kde S.E \neq 2
- Spojíme odpovídající záznamy R,S
- Provedeme projekci na atributy B,D



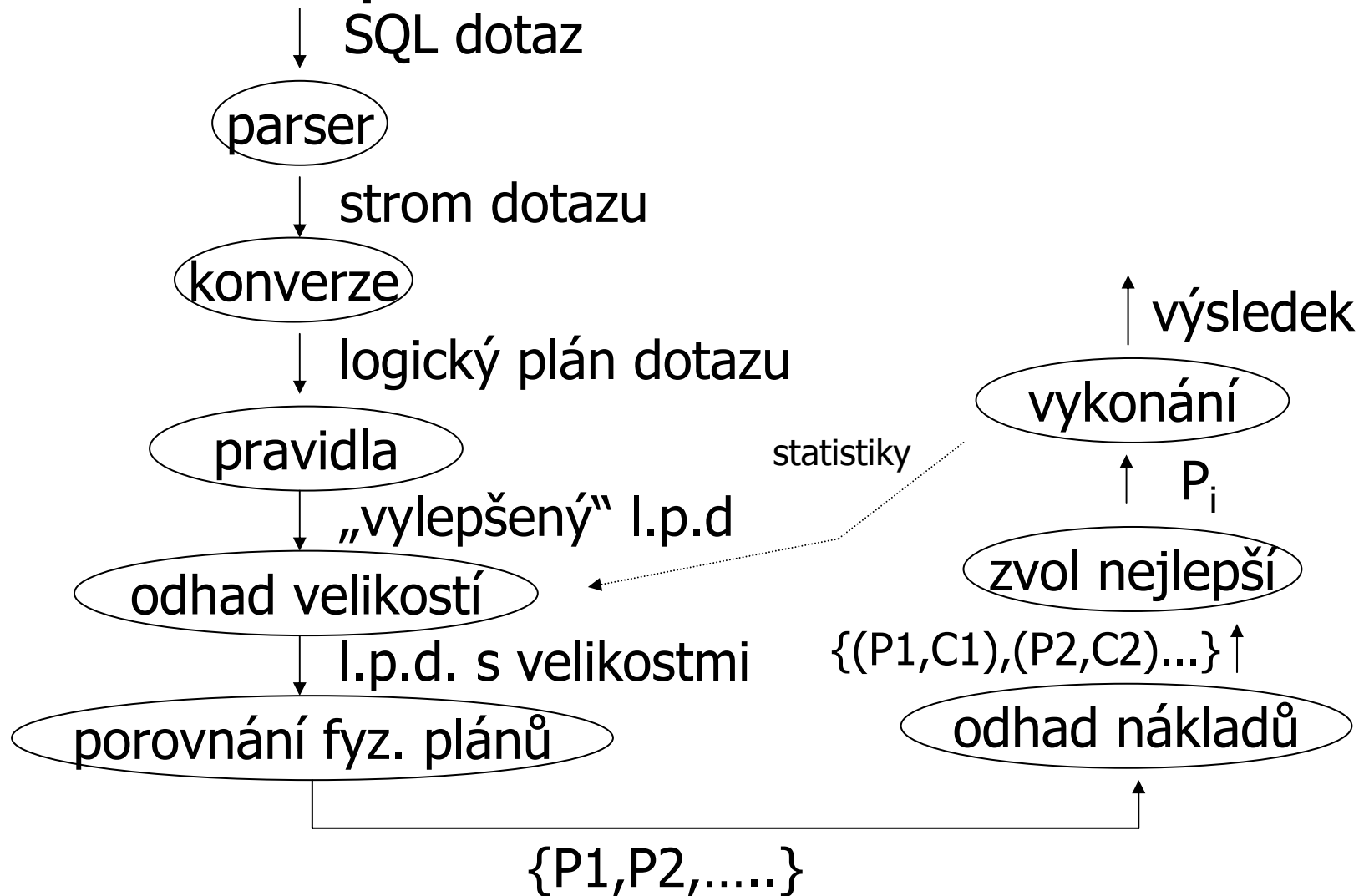






další záznam:
 $\langle c, 7, 15 \rangle$

Přehled optimalizace dotazů

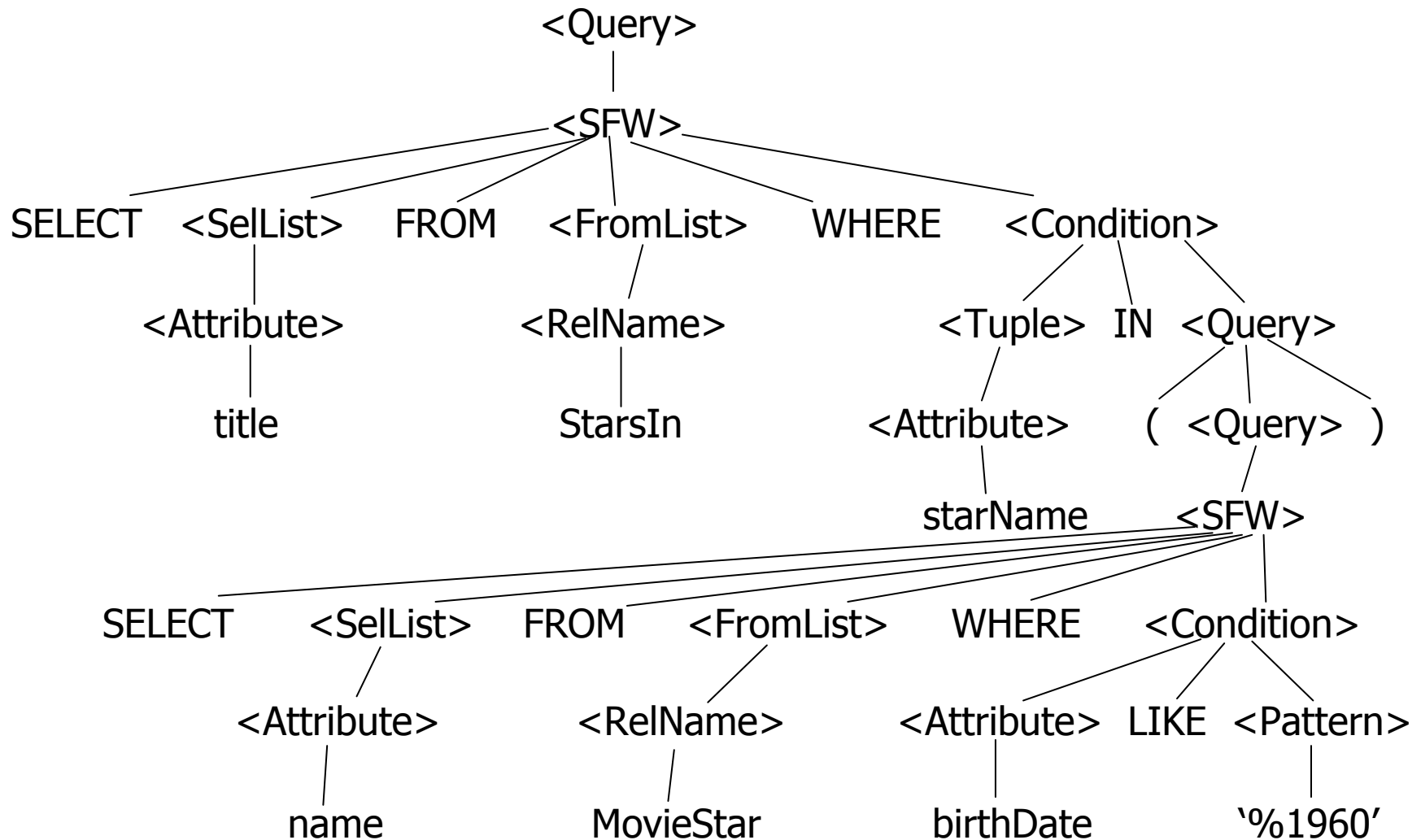


Příklad: SQL dotaz

```
SELECT title
FROM StarsIn
WHERE starName IN (
    SELECT name
    FROM MovieStar
    WHERE birthdate LIKE '%1960'
);
```

(Find the movies with stars born in 1960)

Příklad: strom dotazu



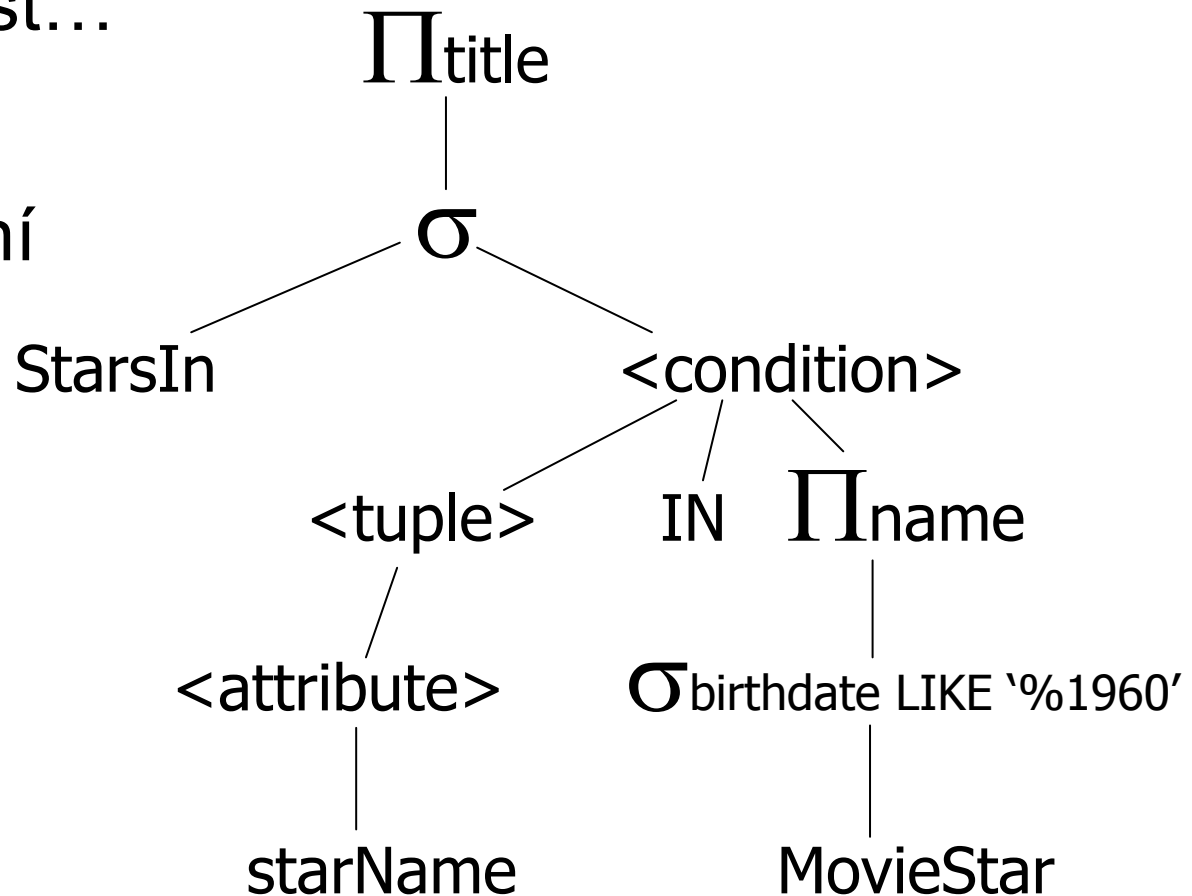
Příklad: převod do relační algebry

- Selekce má dva argumenty

- Třeba převést...

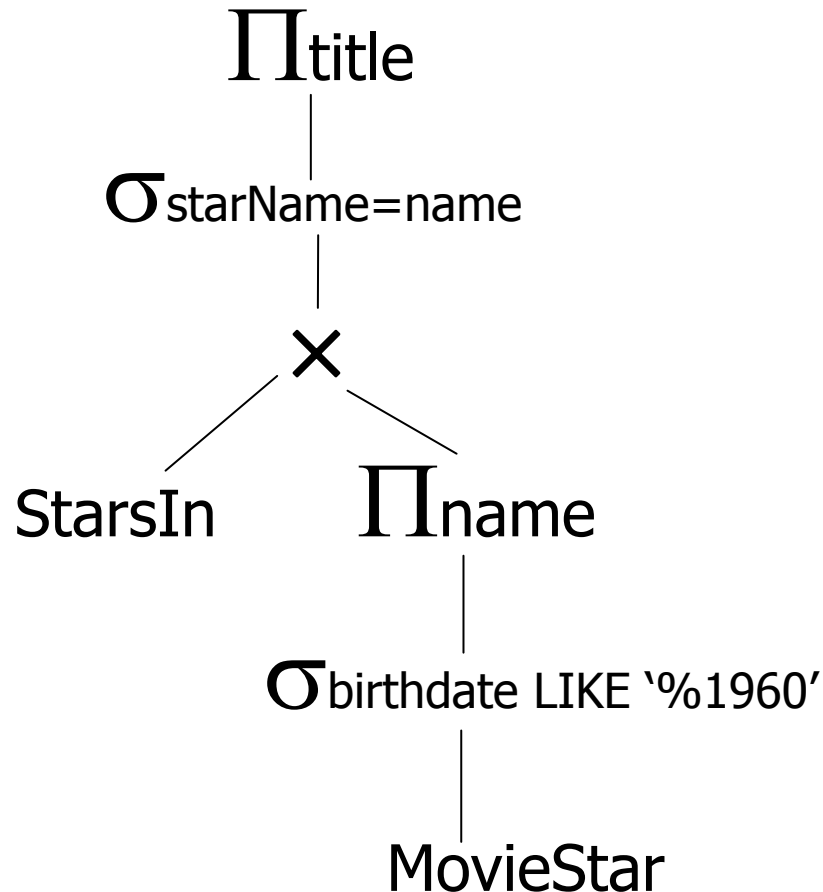
- Operátor IN

- Tj. odstranění vnořených dotazů



Příklad: logický plán dotazu

- Operátor IN nahrazen součinem



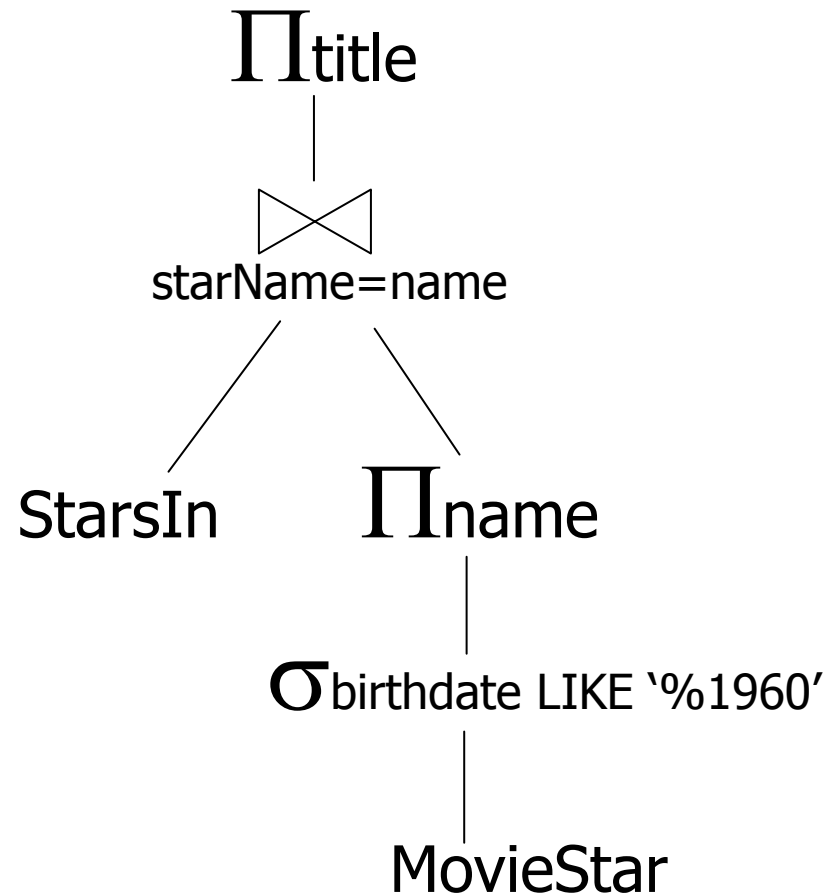
Příklad: Vylepšení logického plánu

■ Nahrazení součinu a selekce

- Provedení spojení

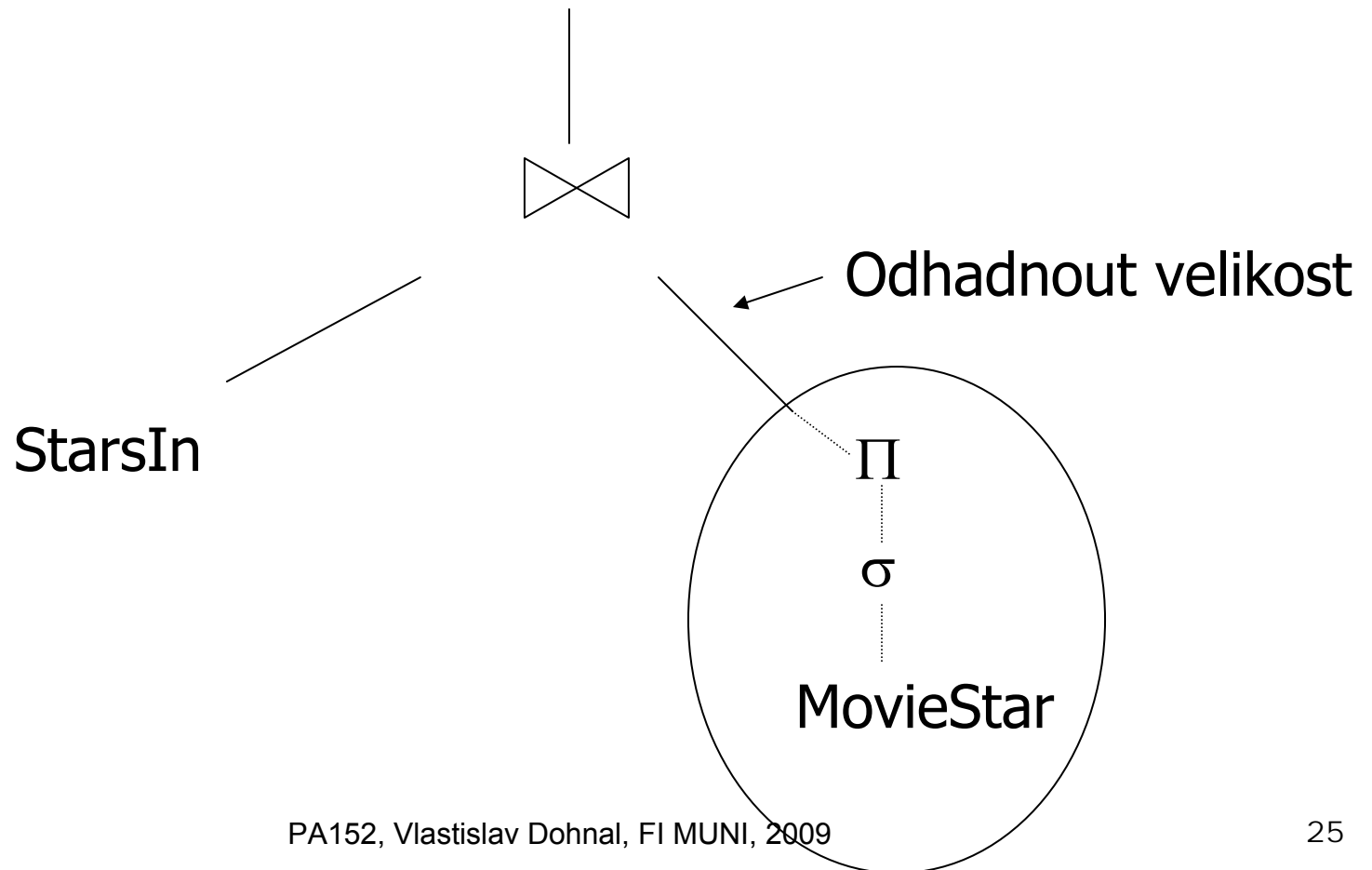
■ Další možnost

- Posunou projekci k relaci StarsIn?

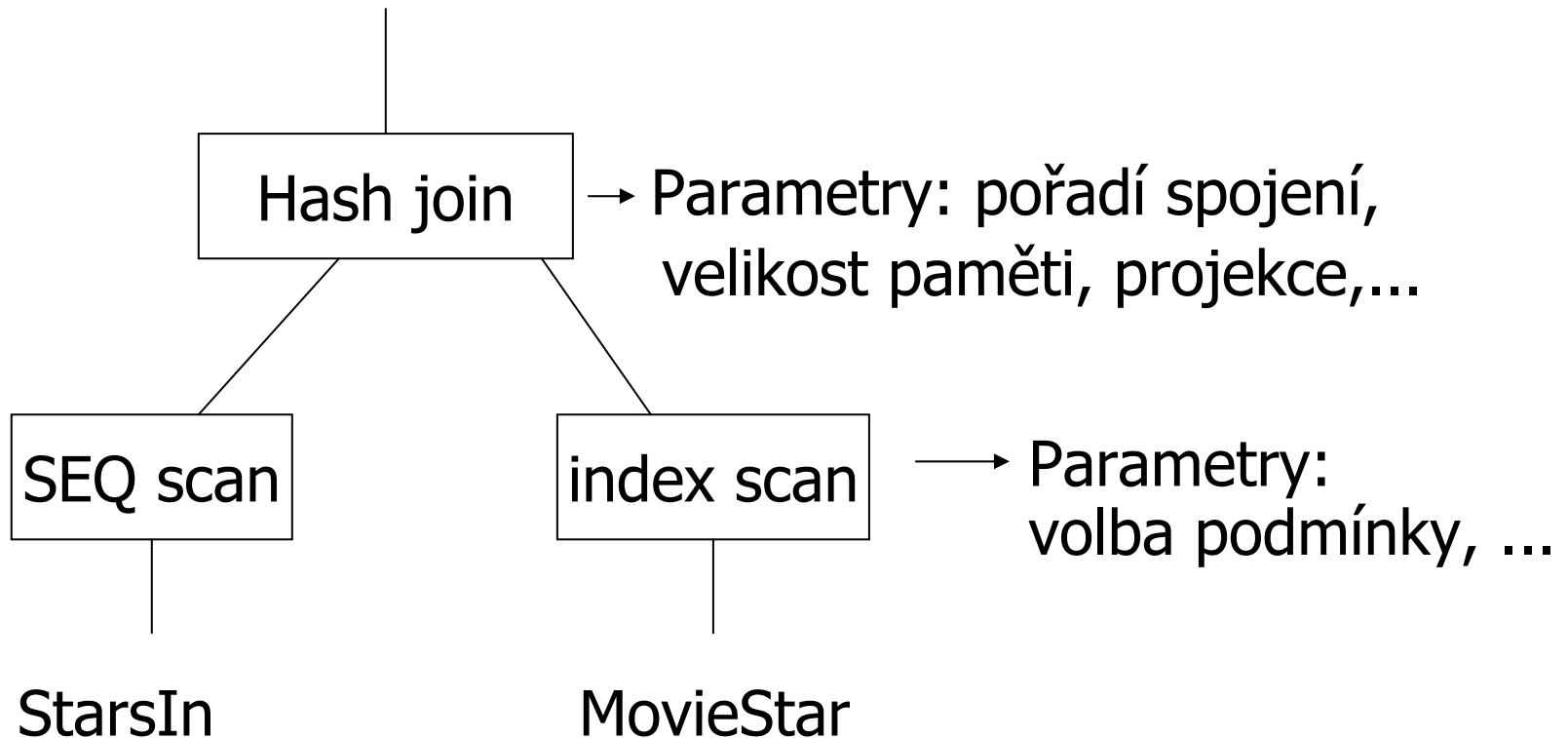


Příklad: odhad velikostí výsledků

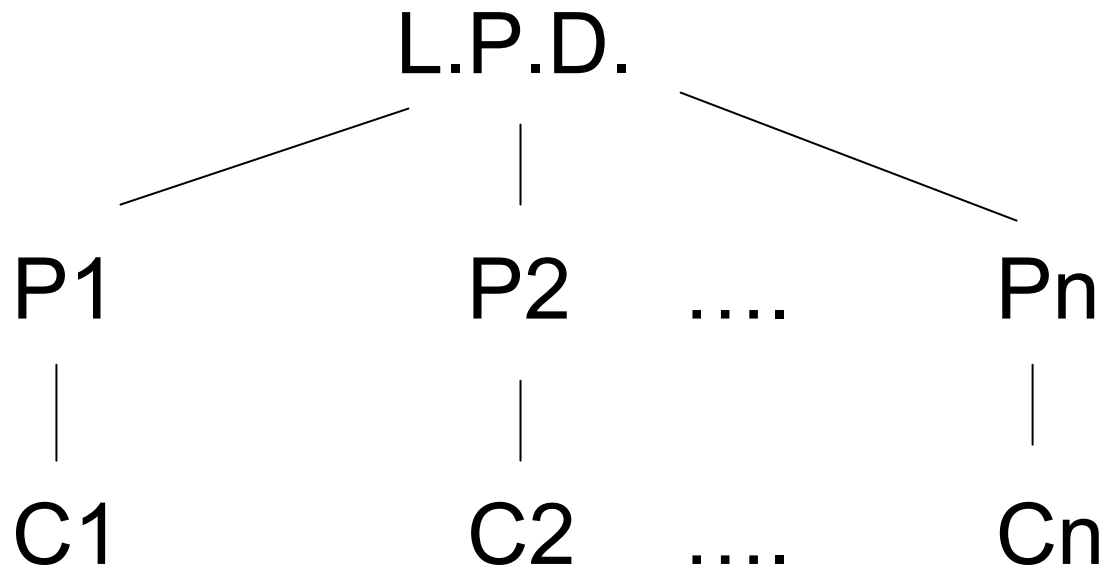
- Před generováním fyzických plánů
- Ovlivňují odhad ceny provedení



Příklad: jeden fyzický plán



Příklad: ohodnocení plánů cenou



Vyber plán s nejnižší cenou!

Optimalizace dotazu

- Úroveň relační algebry
- Úroveň podrobného plánu dotazu
 - Odhad ceny
 - Bez indexů
 - S indexy
 - Vytvoření a porovnání plánů

Optimalizace relační algebry

■ Transformační pravidla

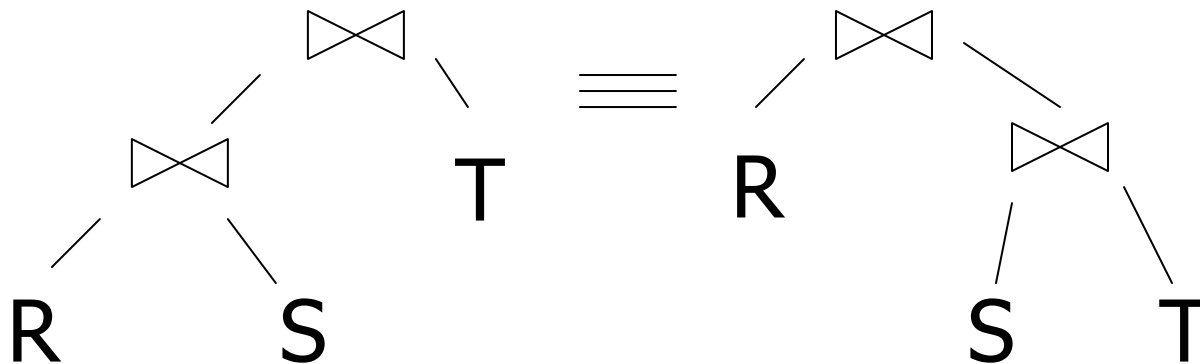
- Musí zajistit ekvivalenci výsledků
- Jaké transformace jsou vhodné?

Transformační pravidla

■ Přirozené spojení

- Protože jsou všechny atributy zachovány, není pořadí důležité

■ Příklad: $R \bowtie S = S \bowtie R$
 $(R \bowtie S) \bowtie T = R \bowtie (S \bowtie T)$



Transformační pravidla

- Stejně pro kartézský součin a sjednocení

$$R \times S = S \times R$$

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

$$R \cup S = S \cup R$$

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

Transformační pravidla

■ Selekcce

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

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

Problém duplicit

- Množiny nebo multimnožiny?

- Relace jsou multimnožiny

- Příklad

- $R = \{a, a, b, b, b, c\}$

- $S = \{b, b, c, c, d\}$

- $R \cup S = ?$

- Možnost 1: SUM

- $R \cup S = \{a, a, b, b, b, b, b, c, c, c, d\}$

- Možnost 2: MAX

- $R \cup S = \{a, a, b, b, b, c, c, d\}$

Možnost 1: SUM

- Sjednocení relací

- $R \cup S$

Možnost 2: MAX

■ Rozklad selekce:

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

■ Příklad: $R = \{a, a, b, b, b, c\}$

□ P1 splňují a,b; P2 splňují b,c

$$\sigma_{p1 \vee p2}(R) = \{a, a, b, b, b, c\}$$

$$\sigma_{p1}(R) = \{a, a, b, b, b\}$$

$$\sigma_{p2}(R) = \{b, b, b, c\}$$

$$\sigma_{p1}(R) \cup \sigma_{p2}(R) = \{a, a, b, b, b, c\}$$

Volba správné možnosti

■ Pragmatické rozhodnutí

- Použití “SUM” pro sjednocení multimnožin
- Některá pravidla nemůžeme pro multimnožiny použít

Transformační pravidla

■ Značení:

- X = množina atributů
- Y = množina atributů
- $XY = X \cup Y$

■ Projekce

$$\pi_{xy}(R) = \pi_x[\pi_y(R)]$$

Transformační pravidla

- Kombinace selekce a přirozeného spojení

- Necht'

p = výraz obsahující pouze atributy R

q = výraz obsahující pouze atributy S

m = výraz obsahující atributy R a S

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

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

Transformační pravidla

- Kombinace selekce a přirozeného spojení
 - Další pravidla lze odvodit

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

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

$$\sigma_{p \vee q} (R \bowtie S) =$$

$$\left[(\sigma_p (R)) \bowtie S \right] \cup \left[R \bowtie (\sigma_q (S)) \right]$$

Příklad odvození pravidla

$$\sigma_{p \wedge q} (R \bowtie S) =$$

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

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

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

Transformační pravidla

- Kombinace projekce a selekce

- Necht'

x = podmnožina atributů R

z = atributy použité ve výrazu P
(podmnožina R)

$$\pi_x[\sigma_p(R)] = \{\sigma_p[\pi_x(R)]\}$$

Transformační pravidla

- Kombinace projekce a selekce

- Necht'

x = podmnožina atributů R

z = atributy použité ve výrazu P
(podmnožina R)

$$\pi_x[\sigma_p(R)] = \pi_x \left\{ \sigma_p \left[\overset{\pi_{xz}}{\cancel{\pi_x}}(R) \right] \right\}$$

Transformační pravidla

- Kombinace projekce a přirozeného spojení

- Necht'

x = podmnožina atributů R

y = podmnožina atributů S

z = průnik atributů R, S

$$\pi_{xy} (R \bowtie S) =$$

$$\pi_{xy} \{ [\pi_{xz} (R)] \bowtie [\pi_{yz} (S)] \}$$

Transformační pravidla

- Kombinace navíc se selekcí

$$\pi_{xy} \{ \sigma_p (R \bowtie S) \} =$$

$$\pi_{xy} \{ \sigma_p [\pi_{xz'} (R) \bowtie \pi_{yz'} (S)] \}$$

$$z' = z \cup \{ \text{atributy použité v } P \}$$

Transformační pravidla

- Kombinace selekce a sjednocení

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

- Kombinace selekce a rozdílu

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

- Selekcí je možné aplikovat i na S

- Může být vhodné pro zmenšení relace před provedením rozdílu

- Otázka: Musí P něco splňovat?

Volba vhodné transformace

$$\sigma_{p_1 \wedge p_2} (R) \rightarrow \sigma_{p_1} [\sigma_{p_2} (R)]$$

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

$$R \bowtie S \rightarrow S \bowtie R$$

$$\pi_x [\sigma_p (R)] \rightarrow \pi_x \{ \sigma_p [\pi_{xz} (R)] \}$$

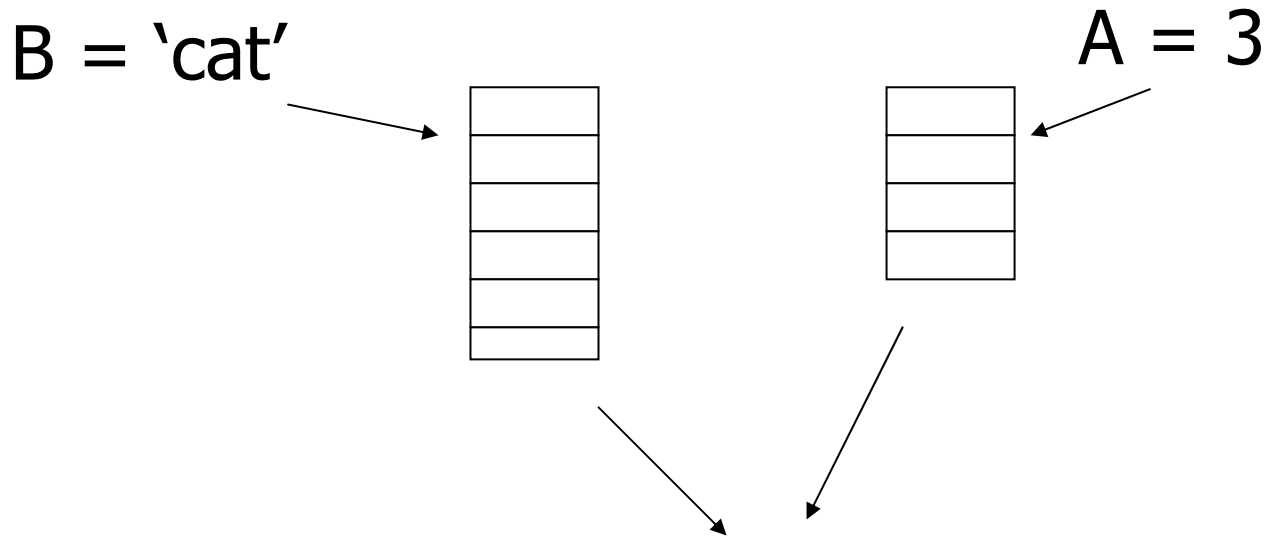
Volba transformace

- Projekce co nejdříve
- Příklad:
 - $R(A,B,C,D,E)$ výsledek= $\{E\}$
 - Filtr P: $(A=3)$ and $(B=\text{"cat"})$

$$\pi_x \{ \sigma_p (R) \} \quad \text{vs.} \quad \pi_E \{ \sigma_p \{ \pi_{ABE}(R) \} \}$$

Volba transformace (2)

- Máme indexy
 - Pro A i pro B



Průnik seznamu ukazatelů
dává výsledek přímo

Volba transformace (3)

- Obecná pravidla:
 - Bez transformací neuděláme chybu
 - Většinou výhodné
 - Selektce nejbliže relacím
- Eliminace společných podvýrazů
- Eliminace duplicit

Zpracování dotazu: přehled

- Úroveň relační algebry
 - Transformační pravidla
 - Volba vhodných pravidel
- Úroveň podrobného plánu dotazu
 - Odhad ceny
 - Vytvoření a porovnání plánů

Odhad ceny plánu dotazu

1. Odhad velikosti výsledku operace
2. Odhad počtu V/V operací

Odhad velikosti výsledku operace

■ Statistiky pro relaci R

- $T(R)$ – počet záznamů
- $S(R)$ – velikost záznamu v bajtech
- $B(R)$ – počet obsazených bloků
- $V(R, A)$ – počet unikátních hodnot atributu A

■ Pro správné odhady

- Statistiky musí být aktuální

Příklad statistik

■ Relace R

- Atribut A – řetězec, 20 bajtů
- Atribut B – celé číslo, 4 bajty
- Atribut C – datum, 8 bajtů
- Atribut D – řetězec, 5 bajtů

A	B	C	D
cat	1	10.2.98	a
cat	1	20.3.98	b
dog	1	30.4.98	a
dog	1	14.6.98	c
bat	1	15.6.98	d

■ Statistiky

- $T(R) = 5$
- $V(R,A) = 3$
- $V(R,C) = 5$
- $S(R) = 37$
- $V(R,B) = 1$
- $V(R,D) = 4$

Odhad velikosti

- Kartézský součin $W = R_1 \times R_2$
 - $T(W) = T(R_1) \cdot T(R_2)$
 - $S(W) = S(R_1) + S(R_2)$

Odhad velikosti

■ Selekcce $W = \sigma_{Z=val}(R)$

□ $S(W) = S(R)$

□ $T(W) = ?$

■ $W = \sigma_{A='cat'}(R)$

$$T(W) = \frac{T(R)}{V(R,A)} = 5/3$$

A	B	C	D
cat	1	10.2.98	a
cat	1	20.3.98	b
dog	1	30.4.98	a
dog	1	14.6.98	c
bat	1	15.6.98	d

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

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

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

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

Odhad velikosti

- Předpoklad předchozího odhadu

- Rovnoměrné rozložení hodnot mezi hodnotami v $R!$

- Tj. $f(\text{val}) = T(R) / V(R,Z)$

- Alternativní předpoklad

- Rovnoměrné rozložení hodnot v celé doméně

- Počet hodnot v doméně označujeme $DOM(R,Z)$

Odhad velikosti: příklad

■ Selekce $W = \sigma_{Z=val}(R)$

□ $T(W) = ?$

■ Odvození

□ $W = \sigma_{C=val}(R)$

■ $T(W) = (1/10)1 + (1/10)1 + \dots$
 $= 5/10 = 0,5$

□ $W = \sigma_{B=val}(R)$

■ $T(W) = (1/10)5$

□ $W = \sigma_{A=val}(R)$

■ $T(W) = 0,5$

A	B	C	D
cat	1	10.2.98	a
cat	1	20.3.98	b
dog	1	30.4.98	a
dog	1	14.6.98	c
bat	1	15.6.98	d

$V(R,A)=3$

$DOM(R,A)=10$

$V(R,B)=1$

$DOM(R,B)=10$

$V(R,C)=5$

$DOM(R,C)=10$

$V(R,D)=4$

$DOM(R,D)=10$

Odhad velikosti

■ Selekce $W = \sigma_{Z=val}(R)$

□ Původní návrh

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

□ Alternativní návrh

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

A	B	C	D
cat	1	10.2.98	a
cat	1	20.3.98	b
dog	1	30.4.98	a
dog	1	14.6.98	c
bat	1	15.6.98	d

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

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

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

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

$$DOM(R,A)=10$$

$$DOM(R,B)=10$$

$$DOM(R,C)=10$$

$$DOM(R,D)=10$$

Odhad velikosti

- Selekcce $W = \sigma_{Z \geq \text{val}}(R)$

- Návrh 1

- $T(W) = T(R) / 2$

- Návrh 2

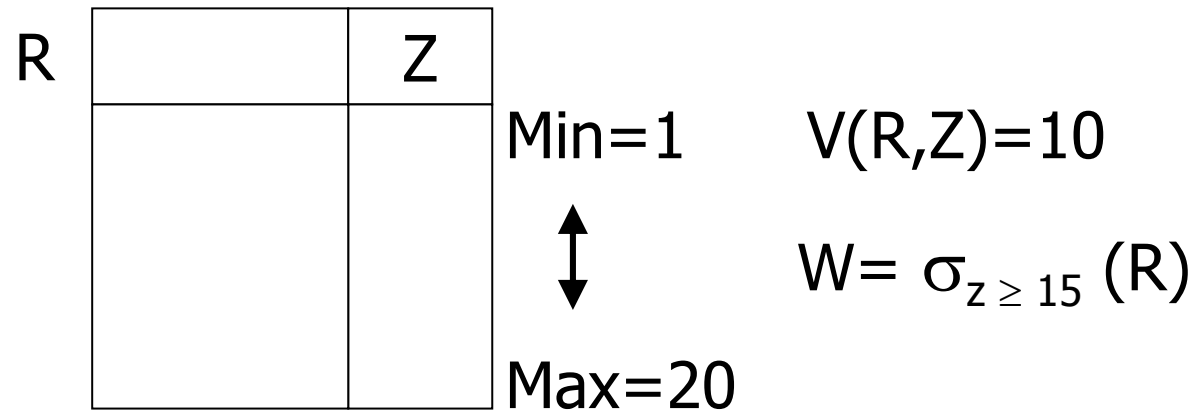
- $T(W) = T(R) / 3$

- Návrh 3

- Podle velikosti rozsahu

Odhad velikosti

- Selekcce – podle velikosti rozsahu



- Vypočítej podíl hodnot (unikátních)

$$f = \frac{20-15+1}{20-1+1} = \frac{6}{20}$$

□ $T(W) = f \cdot T(R)$

Odhad velikosti

■ Přirozené spojení $W = R_1 \bowtie R_2$

□ Značení

- X – atributy R_1
- Y – atributy R_2

■ Příklad 1

- $X \cap Y = \emptyset$
- Stejně jako $R_1 \times R_2$

■ Příklad 2

- $X \cap Y = A$
- Viz dále...

Odhad velikosti: přirozené spojení

$R_1 \bowtie R_2$

R_1	A	B	C

R_2	A	D

■ Předpoklad

□ $V(R_1, A) \leq V(R_2, A)$

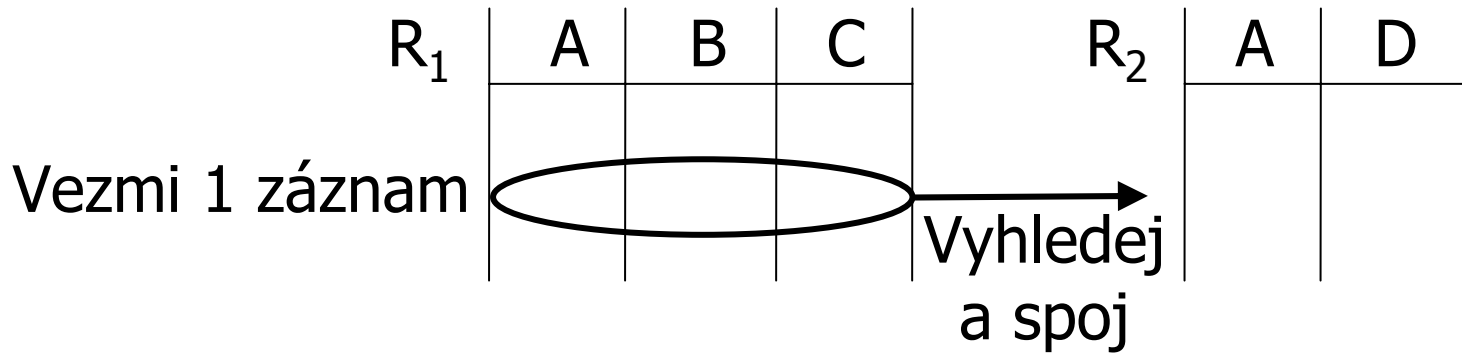
→ každá hodnota A z R_1 je i v R_2

□ $V(R_2, A) \leq V(R_1, A)$

→ každá hodnota A z R_2 je i v R_1

Odhad velikosti: přirozené spojení

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



- 1 záznam se spojí s $T(R_2) / V(R_2, A)$ záznamy
- Výsledek:

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

Odhad velikosti: přirozené spojení

■ Shrnutí obou variant

□ $V(R_1, A) \leq V(R_2, A)$

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

□ $V(R_2, A) \leq V(R_1, A)$

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

■ Rozdíl je pouze v děliteli

Odhad velikosti: přirozené spojení

■ Obecný závěr

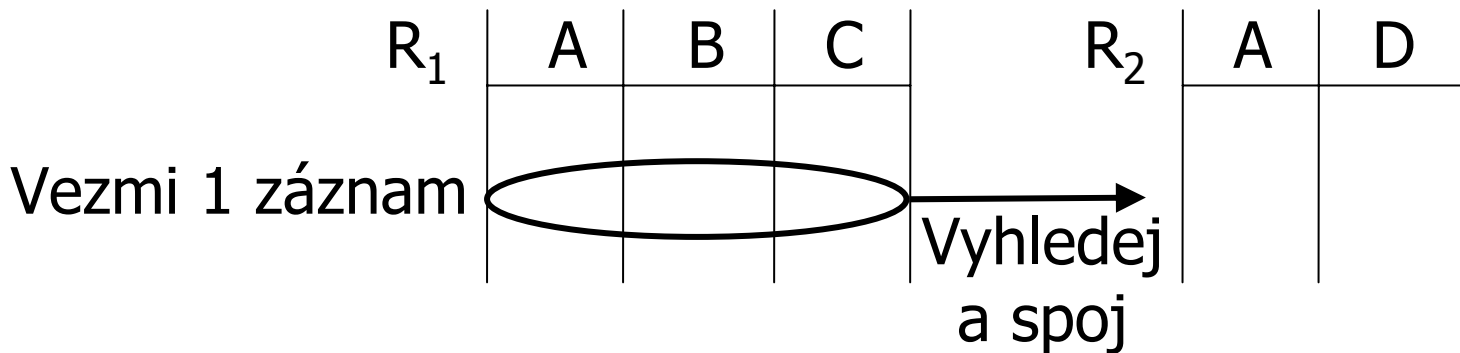
$$\square W = R_1 \bowtie R_2$$

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

Odhad velikosti: přirozené spojení

■ Alternativní definice

- Rovnoměrné rozložení v doméně



- 1 záznam se spojí s $T(R_2)/\text{DOM}(R_2, A)$ záznamy

- Výsledek:

$$T(W) = \frac{T(R_1) \cdot T(R_2)}{\text{DOM}(R_2, A)} = \frac{T(R_1) \cdot T(R_2)}{\text{DOM}(R_1, A)}$$

předpokládáme stejné

Odhad velikosti: přirozené spojení

- $W = R_1 \bowtie R_2$

- Velikost záznamu

- $S(W) = S(R_1) + S(R_2) - S(A)$

- Platí pro všechny varianty

Odhad velikosti: ostatní operace

■ Projekce

- $\Pi_{AB}(R), T(W)=T(R), S(W)=S(AB)$

■ Selekcce

- $\sigma_{A=a \vee B=b}(R), S(W)=S(R),$ necht' $n=T(R)$

- $T(W) = n \cdot (1 - (1-m_1/n) \cdot (1-m_2/n))$

- $m_1 = T(R) / V(R,A) \qquad m_2 = T(R) / V(R,B)$

■ Sjednocení, průnik, rozdíl

- $\cup, \cap, -$

- $T(W)$ – počítá se průměrná velikost

Odhad velikosti

- Pro složitější výrazy jsou třeba ostatní statistiky
- Příklad

$$\square W = [\underbrace{\sigma_{A=a}(R_1)}] \bowtie R_2$$

označme jako U

- $T(U) = T(R_1) / V(R_1, A)$ $S(U) = S(R_1)$
 - Pro výsledek potřebujeme i $V(U, *)$!

Odhad počtu hodnot

■ Odhady $V(U,*)$

- $U = \sigma_{A=a}(R_1)$

- Předpokládejme, že $R_1(A,B,C,D)$

Odhad počtu hodnot: příklad

■ Relace R_1

■ $U = \sigma_{A=a}(R_1)$

A	B	C	D
cat	1	10.2.98	a
cat	1	20.3.98	b
dog	1	30.4.98	a
dog	1	14.6.98	c
bat	1	15.6.98	d

$V(R,A)=3$

$V(R,B)=1$

$V(R,C)=5$

$V(R,D)=4$

■ Výsledek

□ $V(U,A) = 1$

□ $V(U,B) = 1$

□ $V(U,C) = 1 \dots (T(R_1) / V(R_1,A))$

□ $V(U,D) = 1 \dots (T(R_1) / V(R_1,A))$

Odhad počtu hodnot: praxe

■ Obvyklé řešení

- $U = \sigma_{A=a}(R_1)$

- $V(U, A) = 1$

- $V(U, *) = V(R, *)$

Odhad počtu hodnot: spojení

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

- Výsledek:

- $V(U,A) = \min\{ V(R_1, A), V(R_2, A) \}$

- $V(U,B) = V(R_1, B)$

- $V(U,C) = V(R_2, C)$

Odhad počtu hodnot: spojení

■ Příklad

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

$$\square T(R_1) = 1000 \quad V(R_1,A)=50 \quad V(R_1,B)=100$$

$$\square T(R_2) = 2000 \quad V(R_2,B)=200 \quad V(R_2,C)=300$$

$$\square T(R_3) = 3000 \quad V(R_3,C)=90 \quad V(R_3,D)=500$$

Odhad počtu hodnot: spojení

■ Mezivýsledek

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

□ Výsledek:

- $T(U) = T(R_1) \cdot T(R_2) / \max\{ V(R_1,B), V(R_2,B) \} =$
 $= 1000 \cdot 2000 / 200$
- $V(U,A) = 50$
- $V(U,B) = 100$
- $V(U,C) = 300$

Odhad počtu hodnot: spojení

- Celkový výsledek

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

- Výsledek:

- $T(Z) = 10\,000 \cdot 3\,000 / 300$

- $V(Z,A) = 50$

- $V(Z,B) = 100$

- $V(Z,C) = 90$

- $V(Z,D) = 500$

Odhad počtu hodnot: histogram

- Histogram hodnot atributu

- Zpřesnění odhadů

- Počet různých hodnot

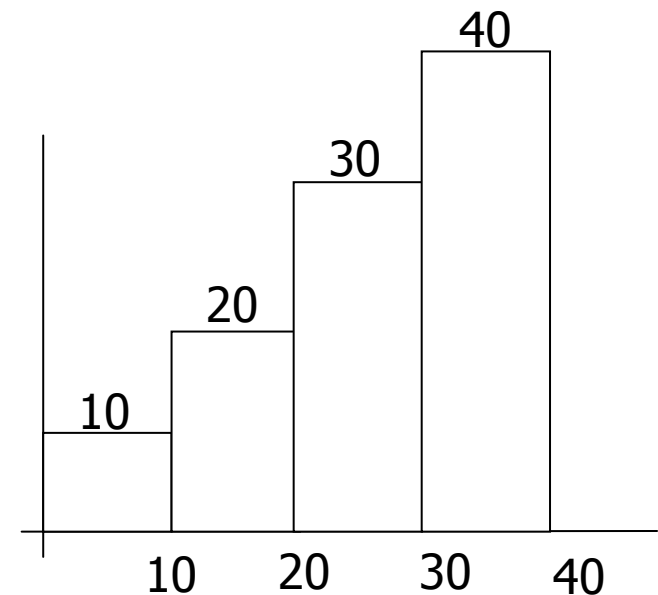
- Málo → pro každou počet

- Hodně → segmentace

- Stejné intervaly

- Percentily

- Pouze pro nejfrekventovanější, ostatní dohromady



Shrnutí

- Odhad velikosti výsledků je „umění“
- Nezapomeňte:
 - Pro korektní odhad potřebujeme korektní statistiky
 - nutnost udržovat tabulky při modifikacích
 - Jaké jsou náklady takové údržby?

Aktualizace statistik

- Statisticky se příliš nemění
 - v krátkém časovém úseku
- I nepřesné statistiky mohou být užitečné
- Okamžitá aktualizace statistik
 - Může být úzkým místem – statistiky jsou velmi často používány
- → Neaktualizuj příliš často

Aktualizace statistik

- Prováděno periodicky
 - Po uplynutí určitého času
 - Po určitém počtu změn
- Pomalé pro $V(R,A)$
 - Zejména pokud se počítají histogramy
 - → Počítáno na vzorku dat
 - Pokud je většina hodnot různých → $V(R,A) \approx T(R)$
 - Pokud je málo různých hodnot → pravděpodobně jsme většinu ze všech viděli

Odhad ceny plánu dotazu: přehled

- Odhad velikosti výsledku operace
 - Již probráno
- Odhad počtu V/V operací
 - Další přednáška
- Vytvoření a porovnání plánů