



PA152: Efektivní využívání DB  
6. Zpracování dotazů

Vlastislav Dohnal

# Vyhodnocení dotazu

## ■ Postup:

- Dotaz
- Syntaktická a sémantická kontrola
  - Strom dotazu
- Logický plán
  - Úpravy plánu
- Fyzický plán
- Vyhodnocení

# Příklad

## ■ Relace

- R(A,B,C)

- S(C,D,E)

## ■ Dotaz

- select B,D

- from R,S

- where R.C=S.C and R.A='c' and S.E=2

# Příklad

R	A	B	C
	a	1	10
	b	1	20
	c	2	10
	d	2	35
	e	3	45

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

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	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

- 1) Kartézský součin
- 2) Výběr záznamů
- 3) Projekce

$R \times S$

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
.					
.					

R × S

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 →

Výstup – výsledek dotazu

select B,D from R,S where R.C=S.C and R.A='c' and S.E=2



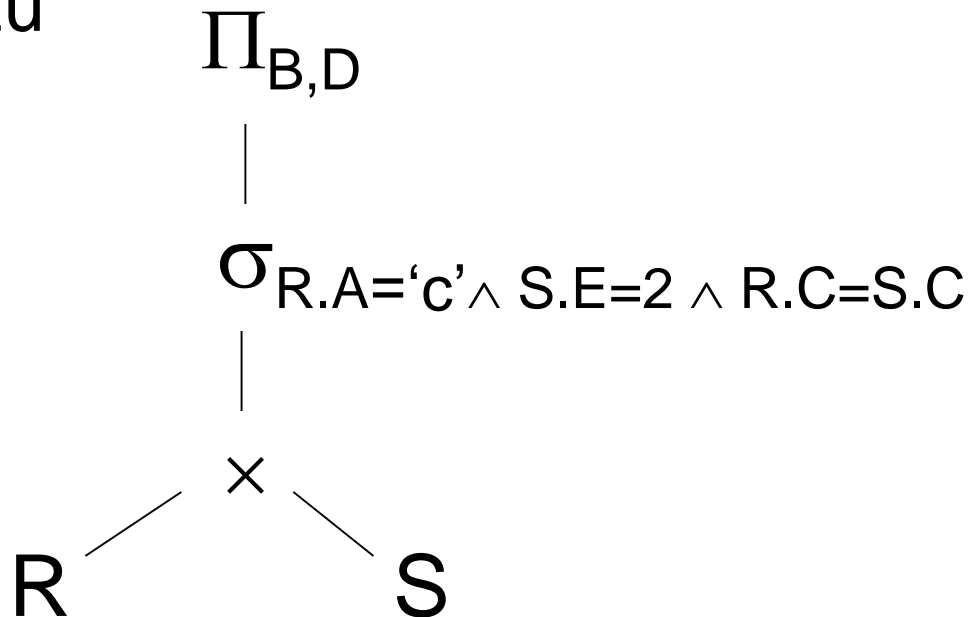
# Popis plánů provedení dotazu

- Použití relační algebry

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

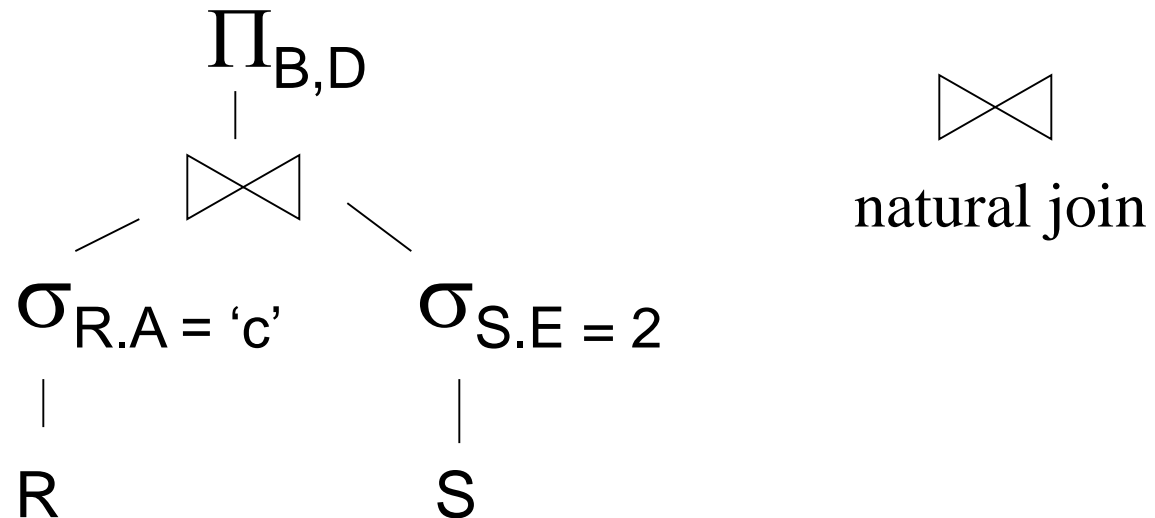
- Příklad plánu 1:

- Strom dotazu



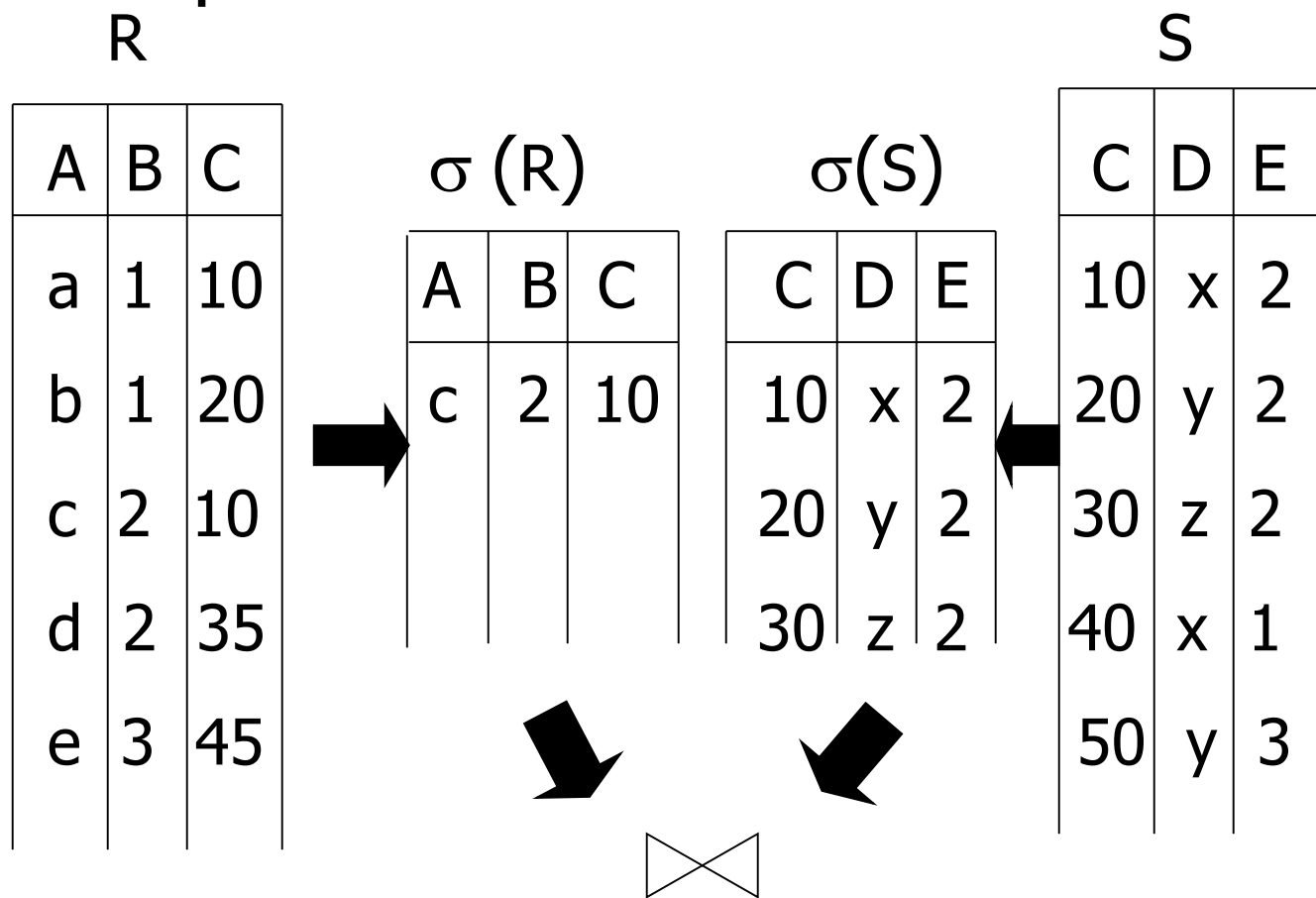
# Popis plánů provedení dotazu

## ■ Příklad plánu 2:



# Fyzický plán

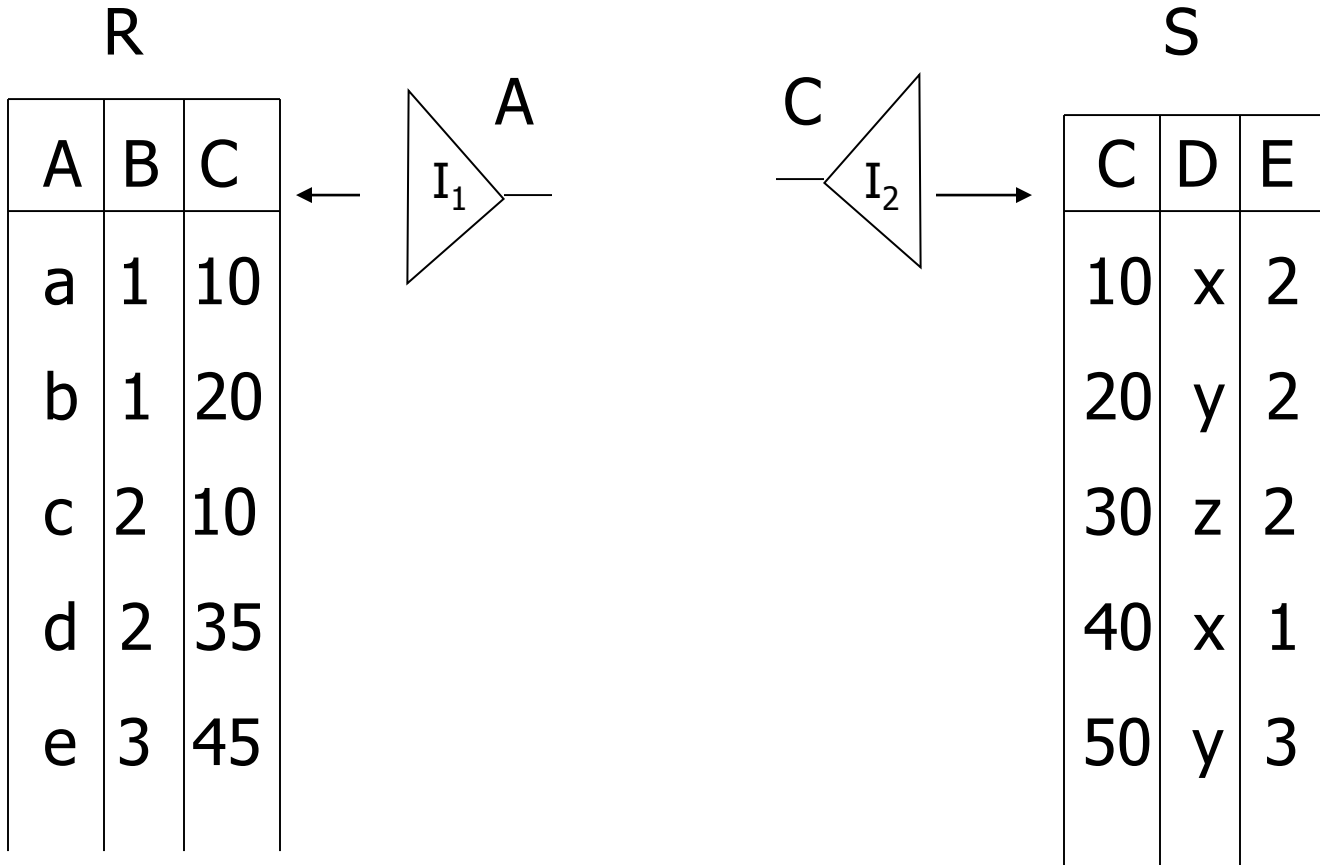
## ■ Příklad plánu 2:

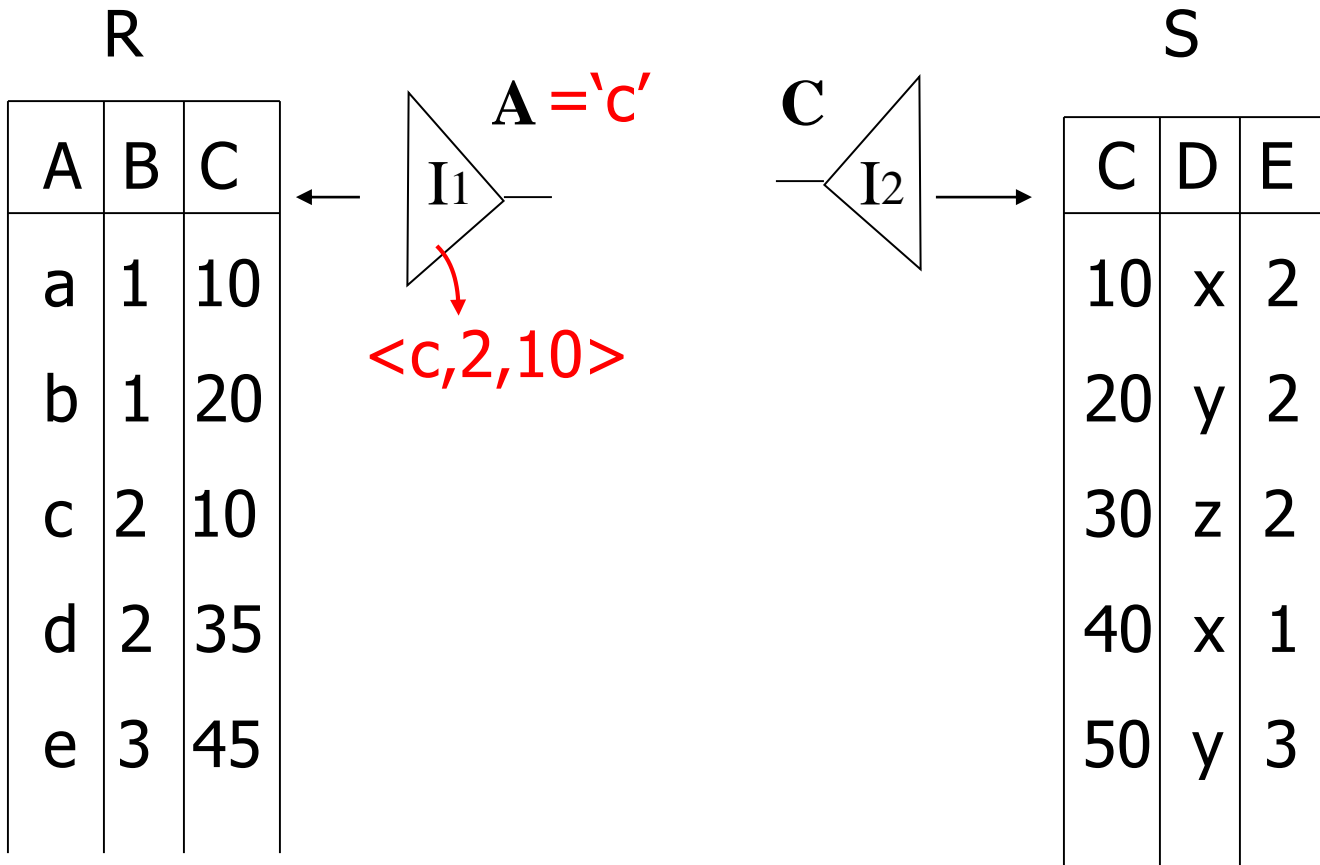


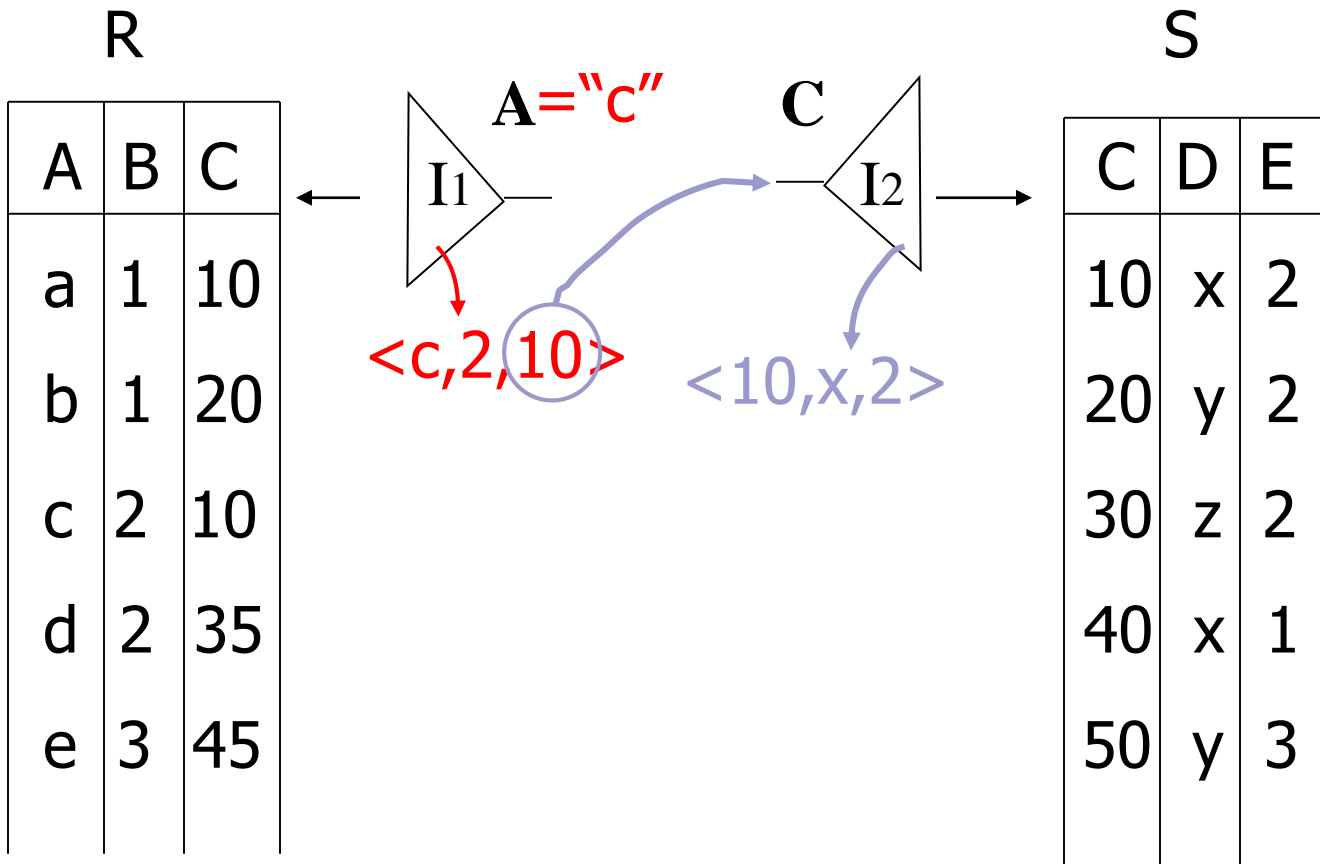
# Popis plánů provedení dotazu

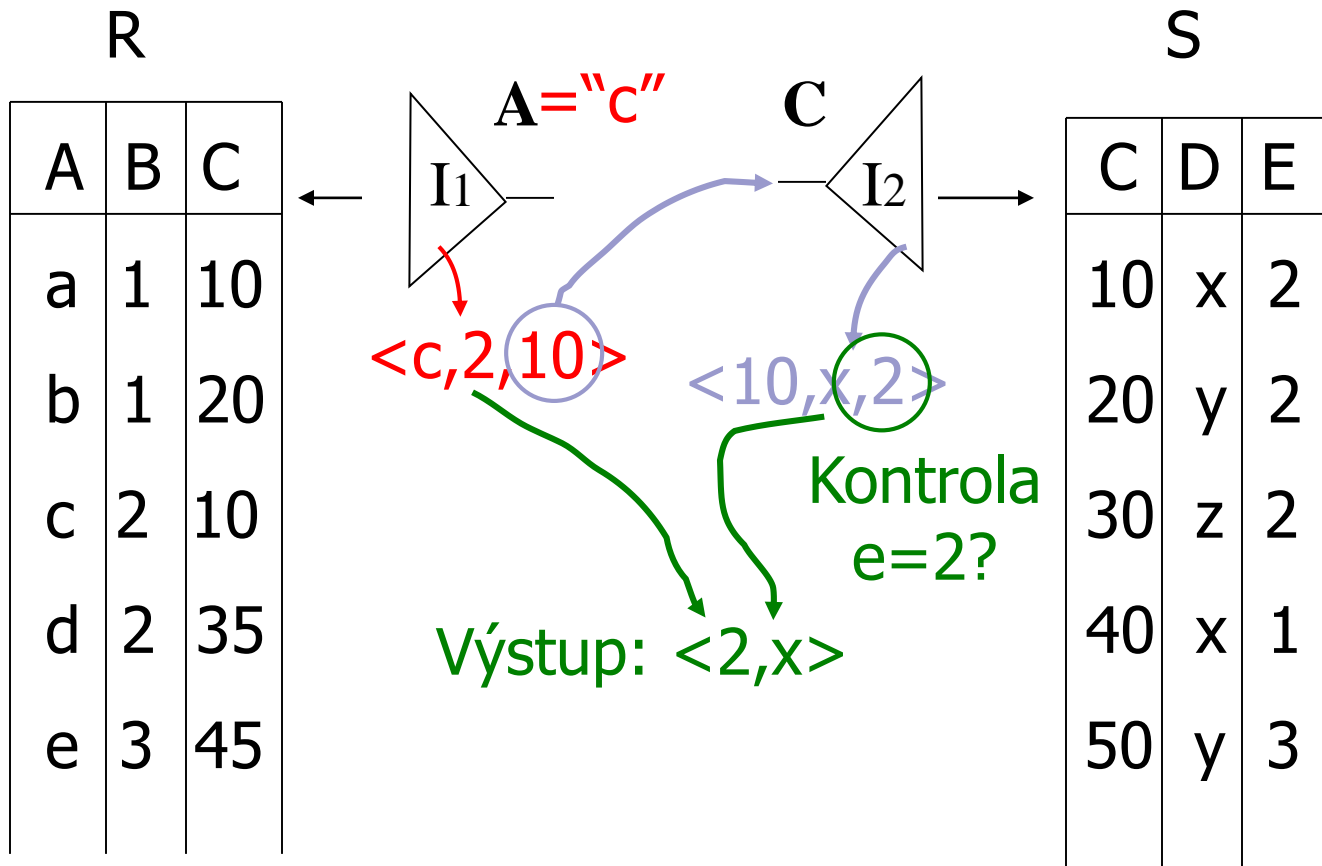
## ■ Plán 3:

- Máme indexy pro R.A a S.C
- Použijeme index R.A k nalezení záznamů R splňujících  $R.A = "c"$ 
  - 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

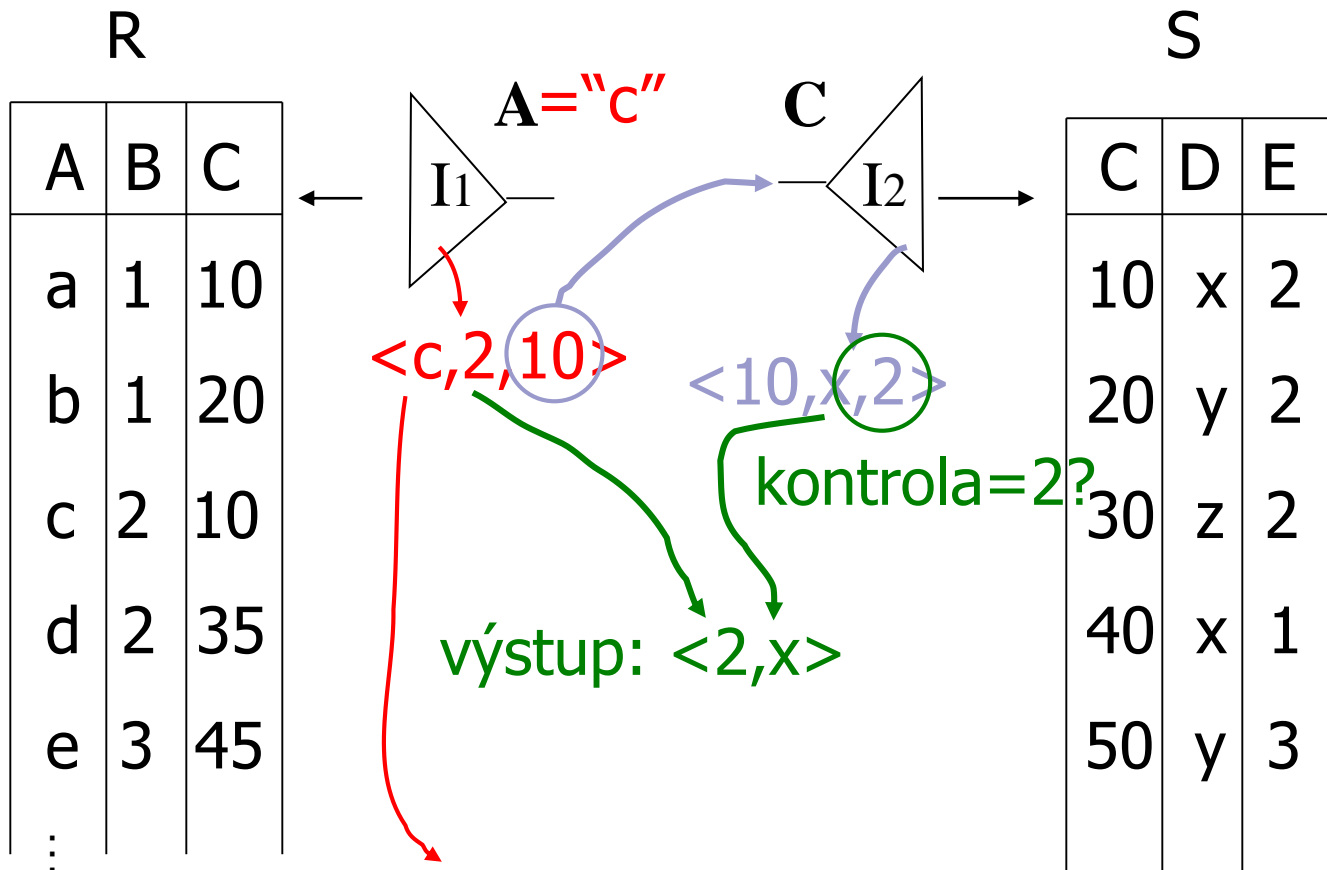




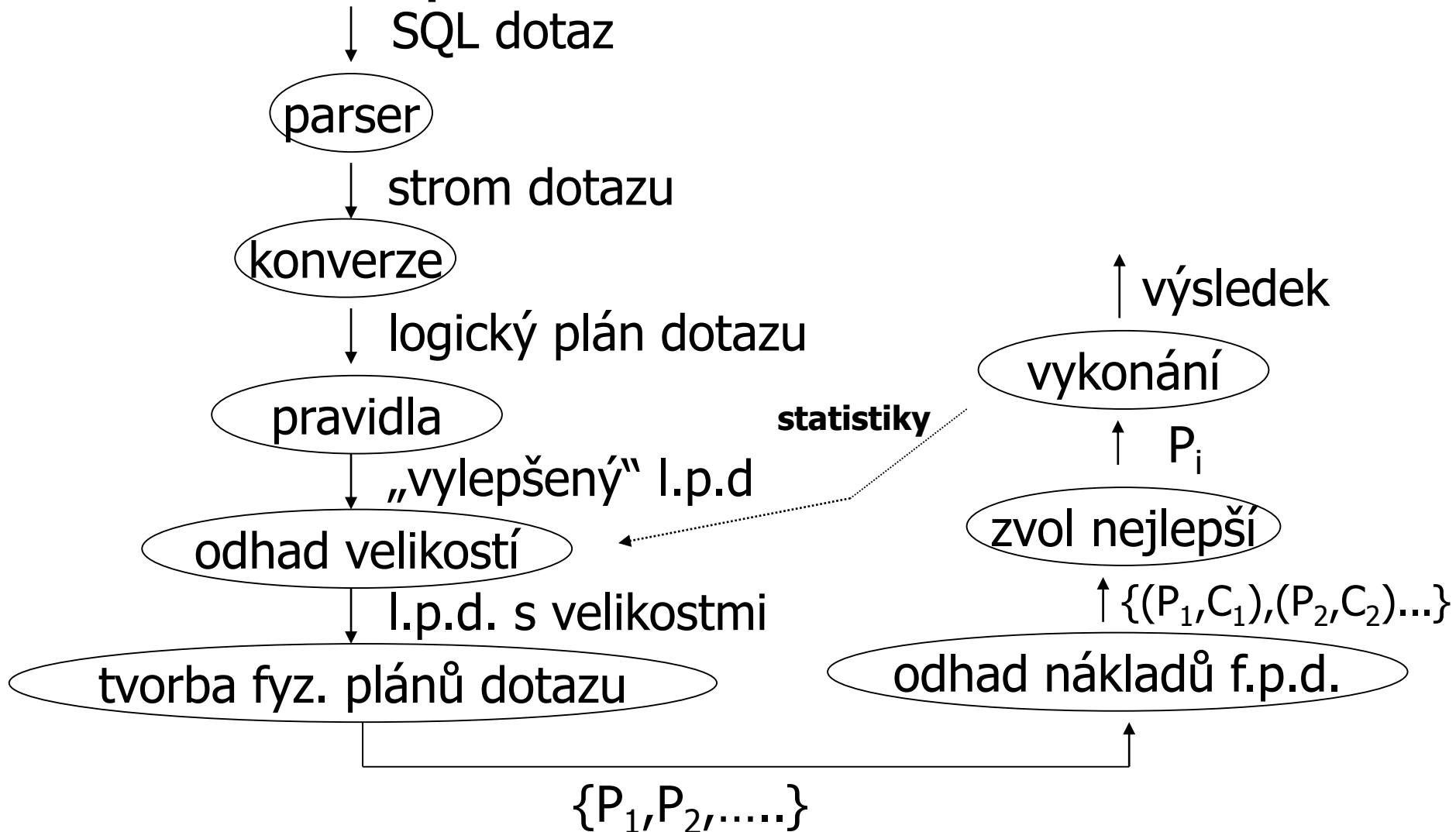








# Přehled optimalizace dotazů



# Příklad: SQL dotaz

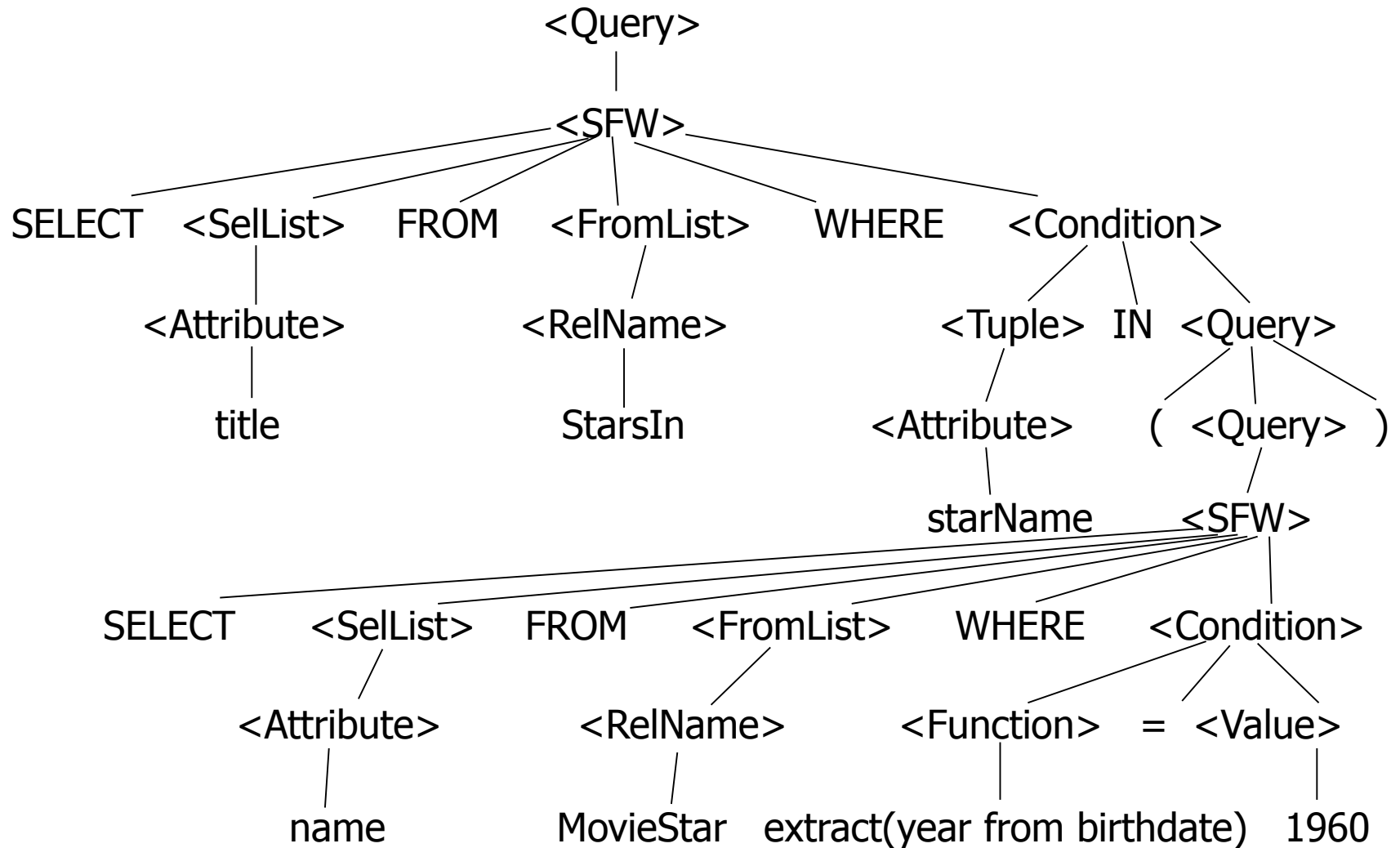
## ■ Relace

- StarsIn(title, year, starName)
- MovieStar(name, birthdate)

## ■ Dotaz

- Najdi filmy, ve kterých hrají herci narození v roce 1960:
- ```
SELECT title
FROM StarsIn
WHERE starName IN (
    SELECT name
    FROM MovieStar
    WHERE extract(year from birthdate) = 1960
);
```

# Příklad: strom dotazu



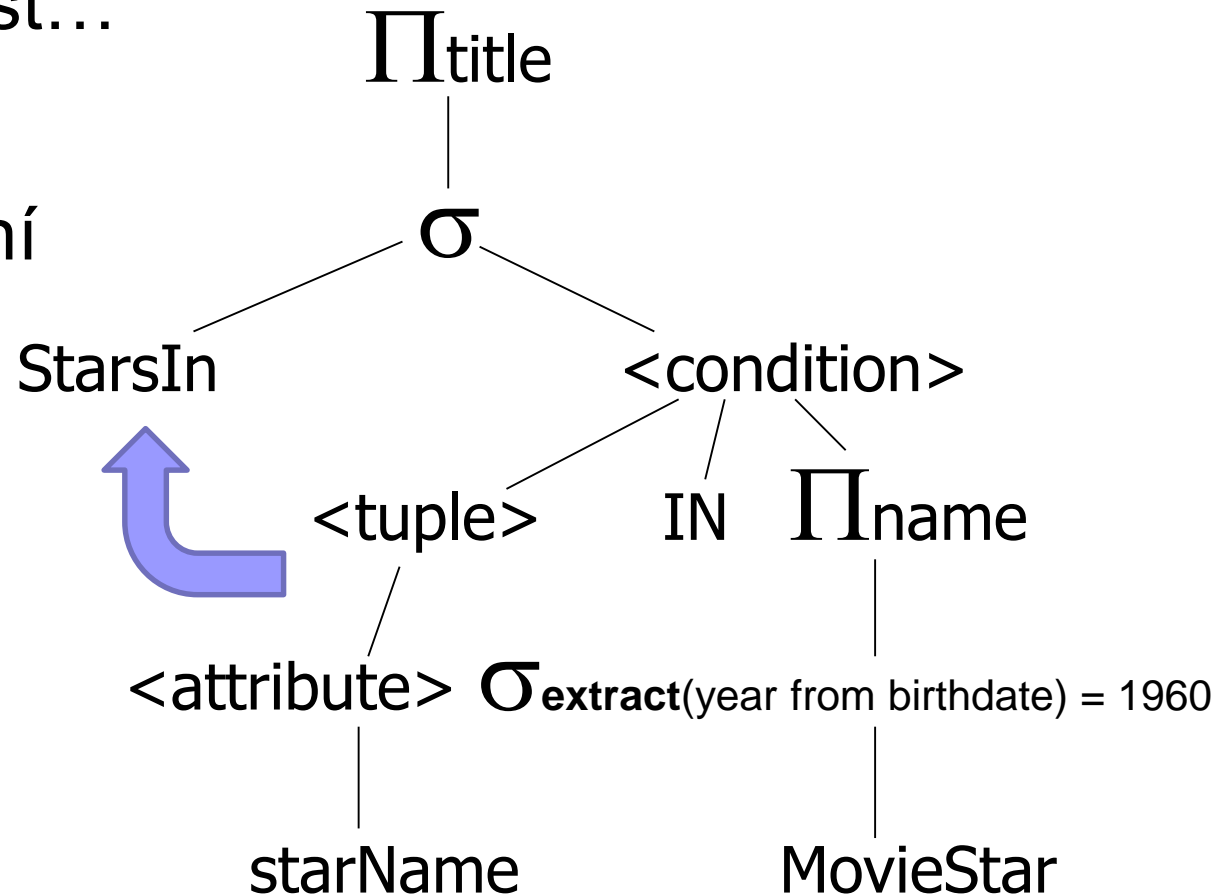
# Příklad: převod do relační algebry

- Selekce má dva argumenty

- Třeba převést...

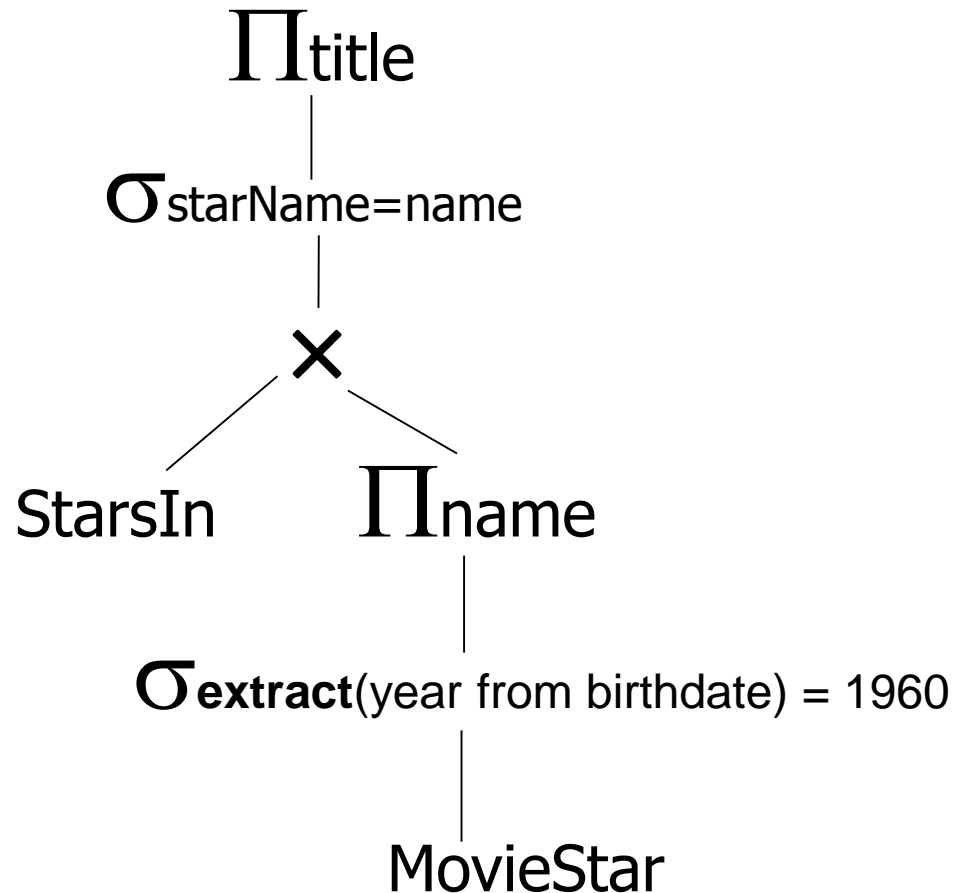
- Operátor IN

- Tj. odstranění vnořených dotazů



# Příklad: logický plán dotazu

- Operátor IN nahrazen součinem



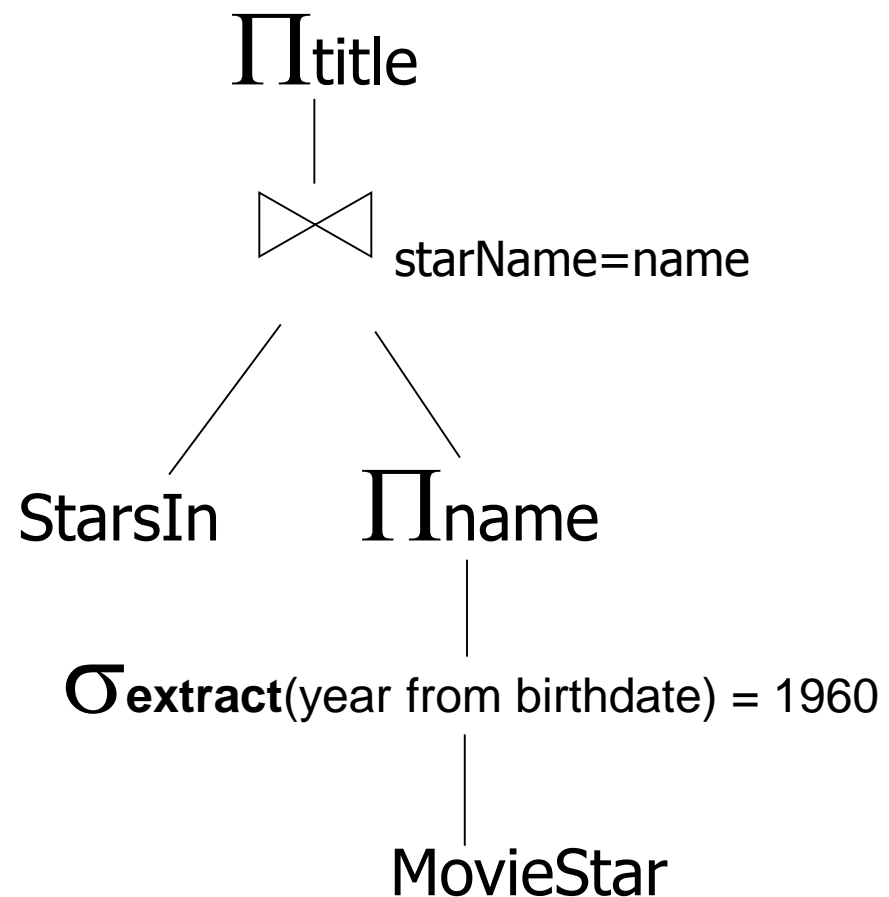
# Příklad: vylepšení logického plánu

## ■ Nahrazení součinu a selekce

- Provedení spojení

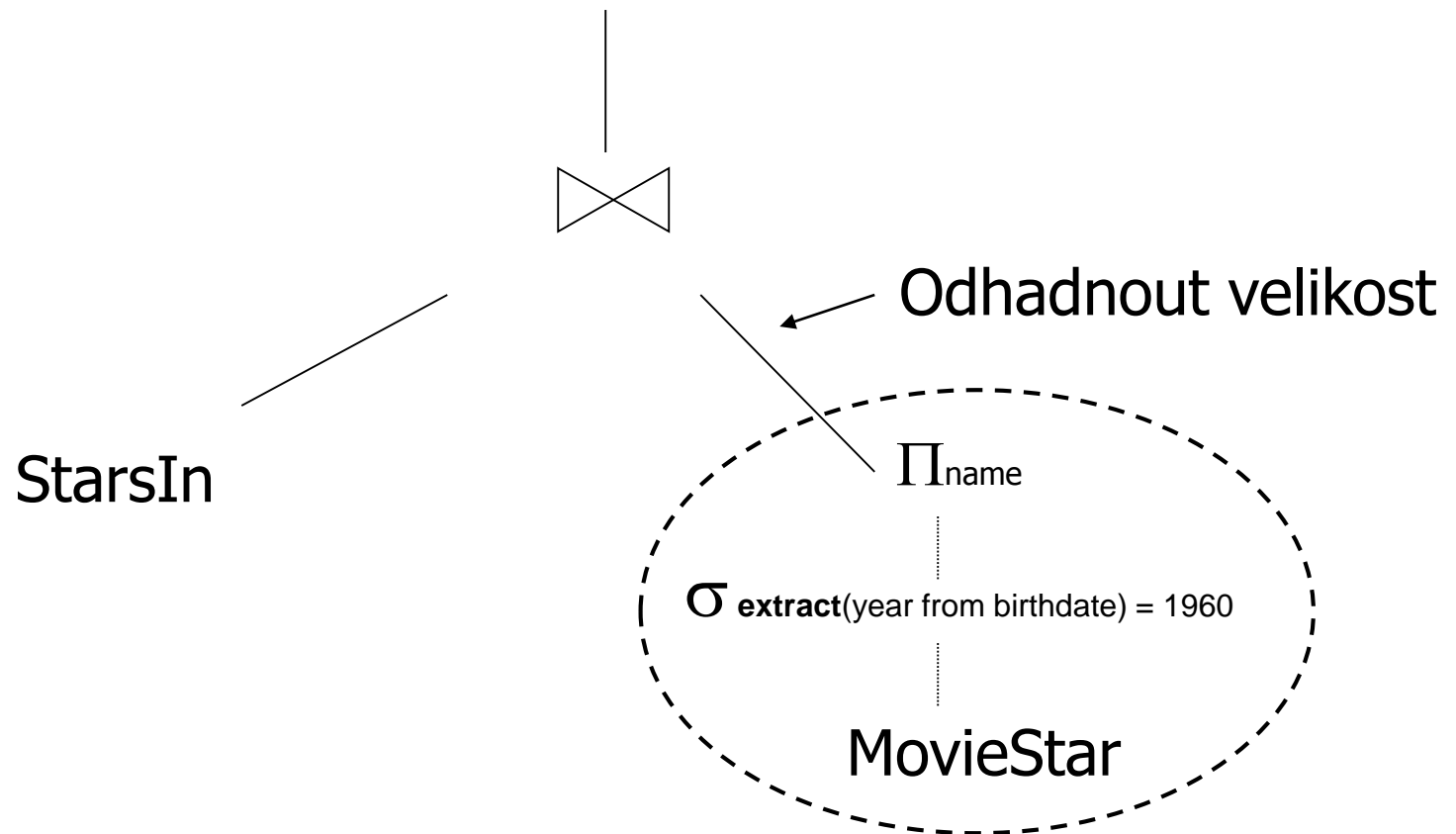
## ■ Další možnost

- Posunout projekci k relaci *StarsIn*?



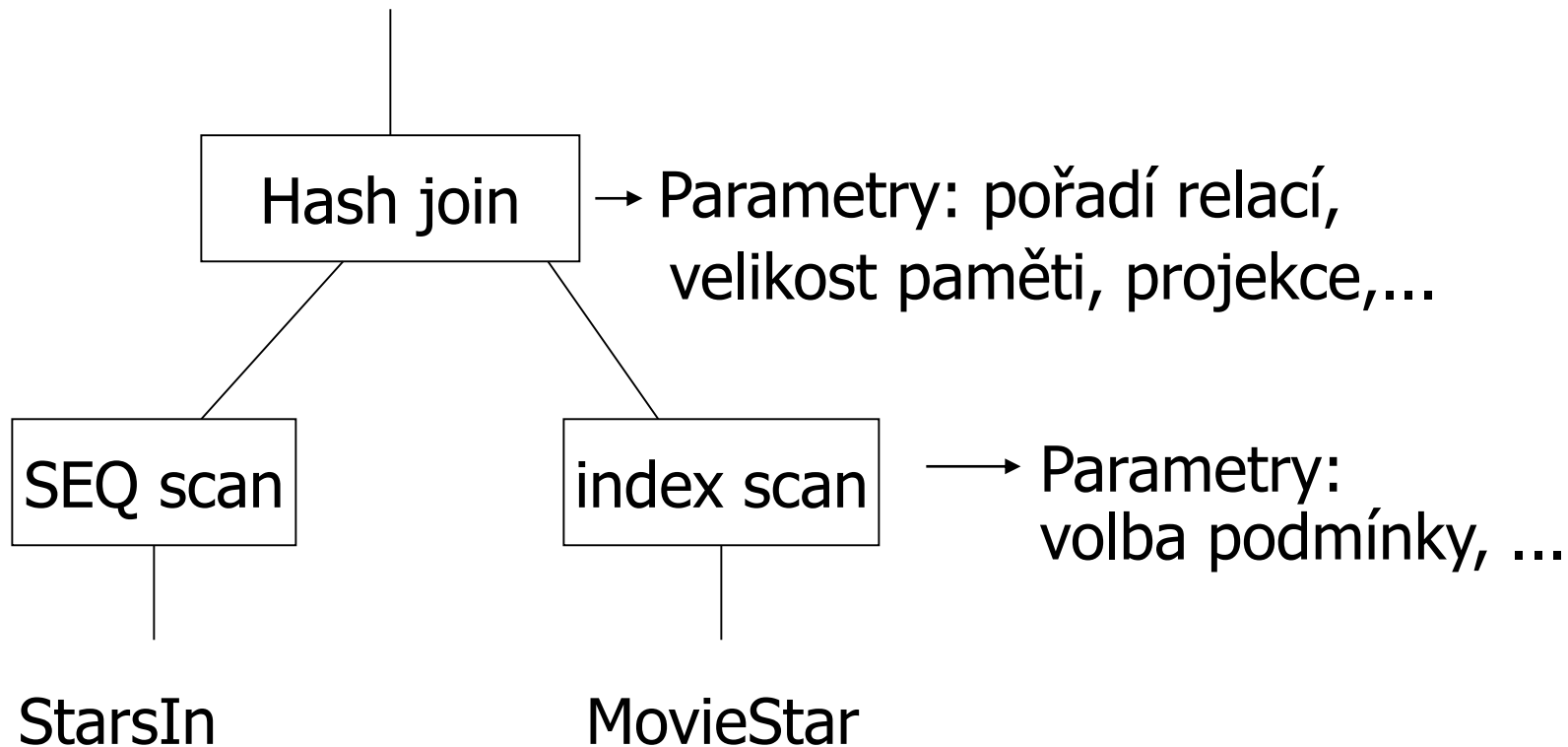
# Příklad: odhad velikostí výsledků

- Před generováním fyzických plánů
- Ovlivňují odhad ceny provedení

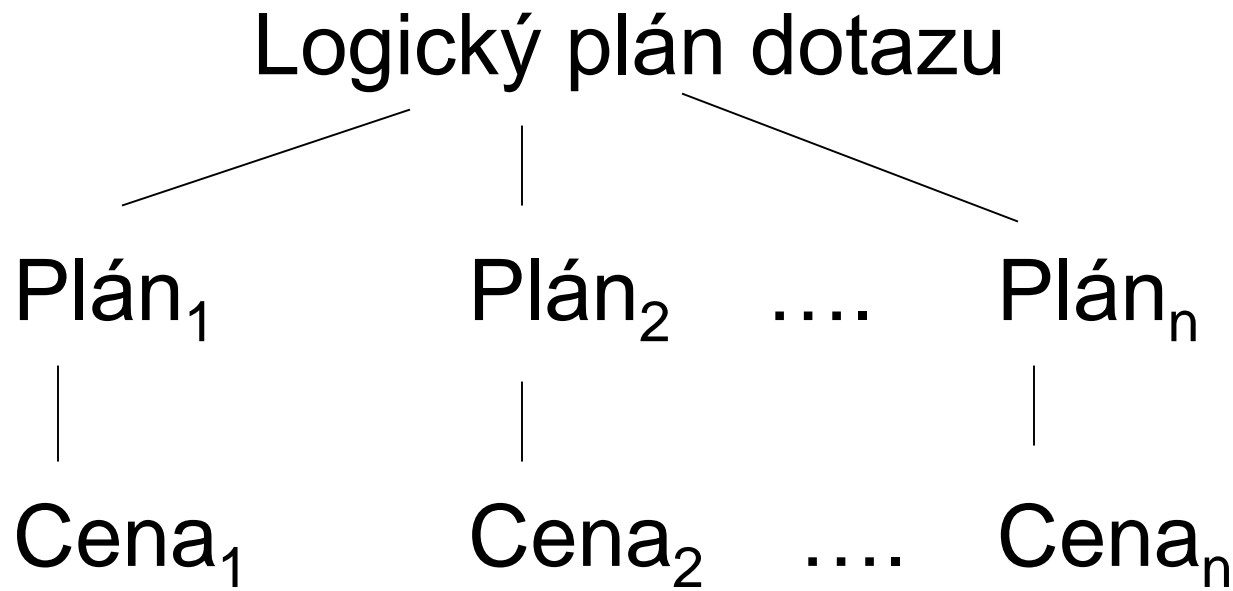




# Příklad: jeden fyzický plán



# 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?

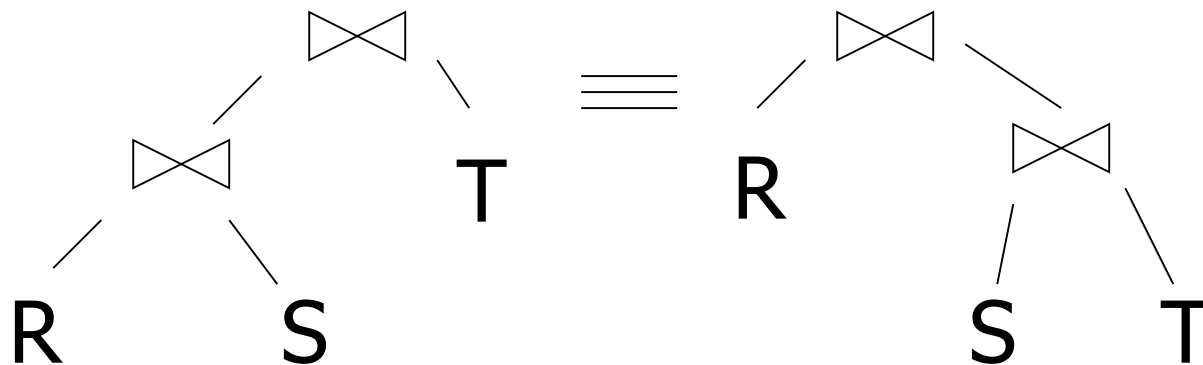
# Transformační pravidla

## ■ 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

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

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

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

# Problém duplicit

## ■ Množiny nebo multimnožiny?

- Relace jsou multimnožiny

## ■ Příklad

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

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

## ■ $R \cup S = ?$

## ■ Možnosti $\cup$

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

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

## ■ $R \cap S = ?$

- MIN:  $R \cap S = \{b, b, c\}$       v SQL: INTERSECT ALL



# Možnost SUM: sjednocení relací

## ■ Sjednocení dvou relací

□  $R \cup S$

v SQL: UNION ALL

## ■ Příklad

□ Poslanci(id, rok, partaj, jméno, ...)

□ Senátoři(id, rok, partaj, jméno, ...)

□  $R = \pi_{\text{rok,partaj}}(\text{Senátoři})$

$S = \pi_{\text{rok,partaj}}(\text{Poslanci})$

| rok  | partaj |
|------|--------|
| 1997 | ODS    |
| 2003 | ČSSD   |
| 2007 | SZ     |

| rok  | partaj |
|------|--------|
| 1997 | ODS    |
| 1998 | KDU    |
| 1996 | ČSSD   |

# Možnost MAX: rozklad selekce

## ■ Rozklad selekce:

$$\sigma_{p_1 \vee p_2}(R) = \sigma_{p_1}(R) \cup \sigma_{p_2}(R)$$

## ■ Příklad: $R = \{a, a, b, b, b, c\}$

□  $p_1$  splňují a,b;  $p_2$  splňují b,c

$$\sigma_{p_1 \vee p_2}(R) = \{a, a, b, b, b, c\}$$

$$\sigma_{p_1}(R) = \{a, a, b, b, b\}$$

$$\sigma_{p_2}(R) = \{b, b, b, c\}$$

$$\sigma_{p_1}(R) \cup_{\max} \sigma_{p_2}(R) = \{a, a, b, b, b, c\}$$

# Volba správné možnosti

## ■ Pragmatické rozhodnutí

- Použití “SUM” pro sjednocení multimnožin
- „MAX“ pro rozdělení podmínky „*nebo*“ (v)

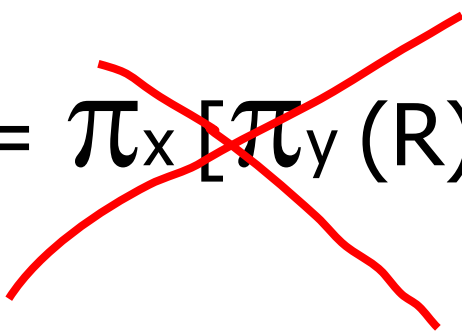
- Některá pravidla nemůžeme pro multimnožiny použít

# 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  $R$  i  $S$

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

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

# Transformační pravidla

- Kombinace selekce a přirozeného spojení
  - Další pravidla lze odvodit

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

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

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

$$\left[ (\sigma_p (R)) \bowtie S \right] \cup_{\max} \left[ R \bowtie (\sigma_q (S)) \right]$$

# Transformační pravidla

- Kombinace selekce a přirozeného spojení
  - 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 selekce a přirozeného spojení

- Příklad odvození pravidla

- Necht'

n = výraz obsahující pouze atributy  
společné R i S

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



# Transformační pravidla

- Kombinace projekce a selekce

- Necht'

$x$  = podmnožina atributů  $R$

$z$  = atributy použité ve výrazu  $P$   
(podmnožina  $R$ )

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

# Transformační pravidla

- Kombinace projekce a přirozeného spojení
- Necht'
  - $x$  = podmnožina atributů  $R$
  - $y$  = podmnožina atributů  $S$
  - $z$  = průnik atributů  $R, S$

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

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

# Transformační pravidla

- Kombinace navíc se selekcí ( $\pi$ ,  $\sigma$ ,  $\bowtie$ )

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

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

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

# Transformační pravidla

- Kombinace projekce, selekce a kartézského součinu

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

# Transformační pravidla

## ■ Kombinace selekce a sjednocení

$$\sigma_p(R \cup_{\text{sum}} S) = \sigma_p(R) \cup_{\text{sum}} \sigma_p(S)$$

## ■ 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

□ Musí  $P$  něco splňovat?

# Vhodné transformace

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

$$\sigma_{p_1 \vee p_2} (R) \rightarrow \sigma_{p_1}(R) \cup_{\max} \sigma_{p_2}(R)$$

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

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

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

# Vhodné transformace

- Projekce co nejdříve

- Příklad:

- $R(A,B,C,D,E,F,G,H,I,J)$     výsledek={E}

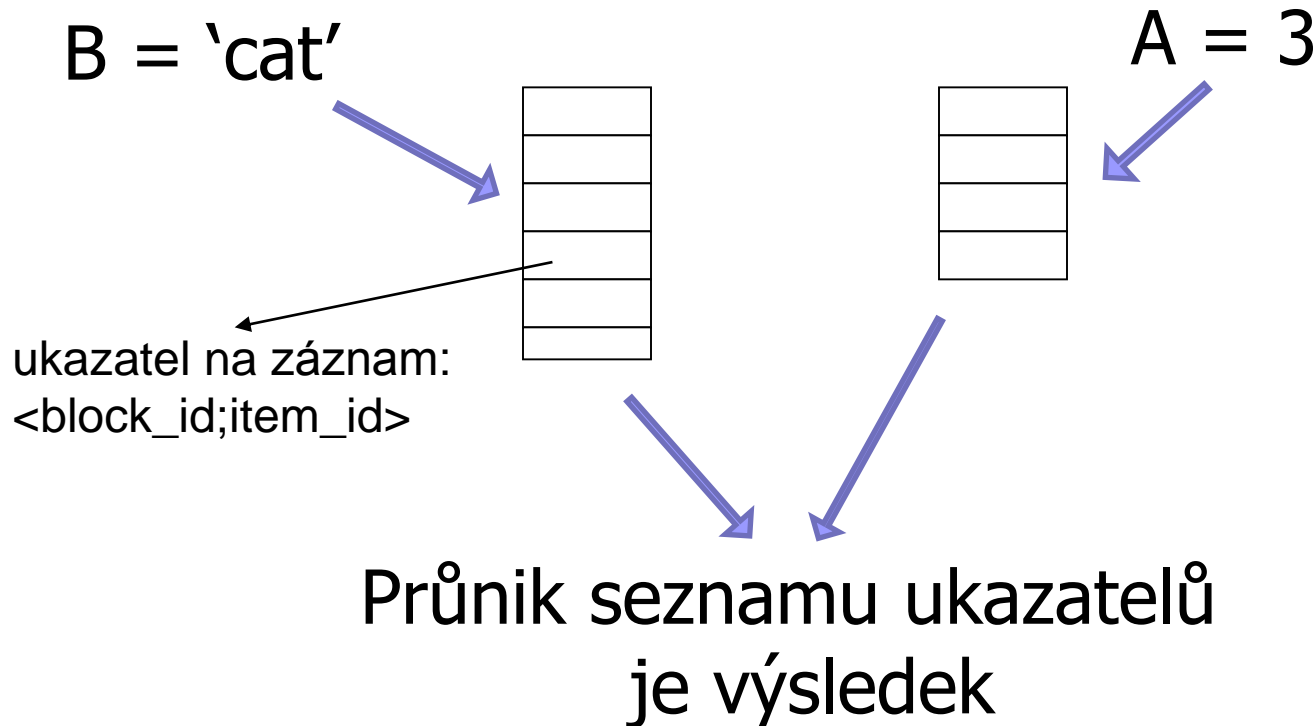
- Filtr P:  $(A=3) \wedge (B=\text{"cat"})$

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

# Vhodné transformace

- Máme indexy
  - Pro A i pro B

$$\begin{aligned}\sigma_{(A=3) \wedge (B=\text{"cat"})}(R) \\ = \sigma_{(A=3)}(R) \cap \sigma_{(B=\text{"cat"})}(R)\end{aligned}$$



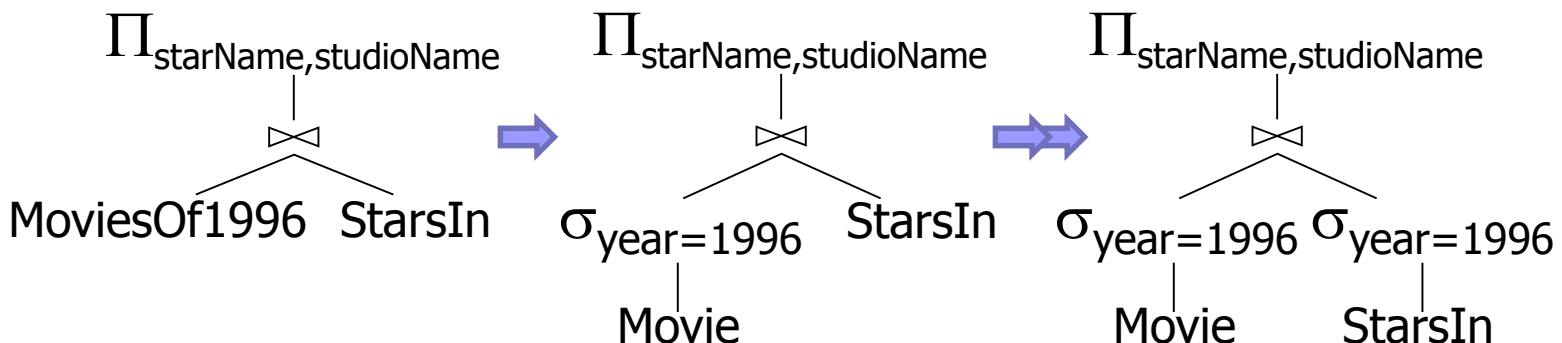


# Vhodné transformace

- Obecná pravidla:
  - Bez transformací neuděláme chybu
  - Většinou výhodné
    - Selekcce nejbližze relacím
    - Projekce nejbližze relacím
- Eliminace společných podvýrazů
- Eliminace duplicit

# Vhodné transformace: příklad

- Přesun selekce co nejbliže relacím → zdánlivě ok
  - Ale: Nejdříve vhodné přesunout co nejdále a pak nejbliže
- Příklad:
  - Relace:  $StarsIn(\underline{title}, \underline{year}, \underline{starName})$   
 $Movie(\underline{title}, \underline{year}, \text{studioName})$
  - Pohled: create view *MoviesOf1996* as  
select \* from *Movie* where *year* = 1996;
- Dotaz: select *starName*, *studioName*  
from *MoviesOf1996* natural join *StarsIn*;



# Zpracování dotazu: přehled

- Úroveň relační algebry
  - Transformační pravidla
  - Použití doporučených 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

## ■ Statistiky pro relaci R

- $T(R)$  – počet záznamů

- $S(R)$  – velikost záznamu v bajtech

  - $S(R,A)$  – velikost atributu (hodnoty) v bajtech

- $B(R)$  – počet obsazených bloků

- $V(R, A)$  – počet unikátních hodnot atributu A

## ■ Pro správné odhady

- Aktuální statistiky nutné!

# Příklad statistik

## ■ Relace R

□ Atribut A – řetězec, max. 20 B

■  $S(R,A) = 3$  ← průměrná délka

□ Atribut B – celé číslo, 4 B

□ Atribut C – datum, 8 B

□ Atribut D – řetězec, 5 B

■  $S(R,D) = 1$

## ■ Statistiky

□  $T(R) = 5$

$S(R) = 20$

□  $V(R,A) = 3$

$V(R,B) = 1$

□  $V(R,C) = 5$

$V(R,D) = 4$

| 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 |

# Odhad velikosti výsledku

- 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 výsledku

## ■ Selekce $W = \sigma_{Z=val}(R)$

□  $S(W) = S(R)$

□  $T(W) = ?$

| 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 |

## ■ $W = \sigma_{A='cat'}(R)$

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

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

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

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

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

## ■ $W_2 = \sigma_{B=2}(R)$

$$T(W_2) = ?$$



# Odhad velikosti výsledku

- Předpoklad předchozího odhadu
  - Rovnoměrné rozložení hodnot mezi hodnotami v  $R!$ 
    - $f(\text{val}) = 1 / V(R,Z)$
    - $T(\sigma_{Z=\text{val}}(R)) = T(R) \cdot f(\text{val})$
- Alternativní předpoklad
  - Rovnoměrné rozložení hodnot v celé doméně
    - Počet hodnot v doméně označujeme  $\text{DOM}(R,Z)$
    - $f(\text{val}) = 1 / \text{DOM}(R,Z)$

# Odhad velikosti výsledku: příklad

## ■ Selekcce $W = \sigma_{Z=val}(R)$

□  $T(W) = ?$

- Podle  $DOM(R,*)$

## ■ Odvození

□  $W = \sigma_{C=val}(R)$

- $T(W) = f(val) \cdot T(R)$   
 $= 1/10 * 5 = 0,5$

□  $W = \sigma_{B=val}(R)$

- $T(W) = (1/10)*5$

□  $W = \sigma_{A=val}(R)$

- $T(W) = 0,5$

| 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 |

$V(R,A)=3$

$V(R,B)=1$

$V(R,C)=5$

$V(R,D)=4$

$DOM(R,A)=10$

$DOM(R,B)=10$

$DOM(R,C)=10$

$DOM(R,D)=10$

# Odhad velikosti výsledku

## ■ 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

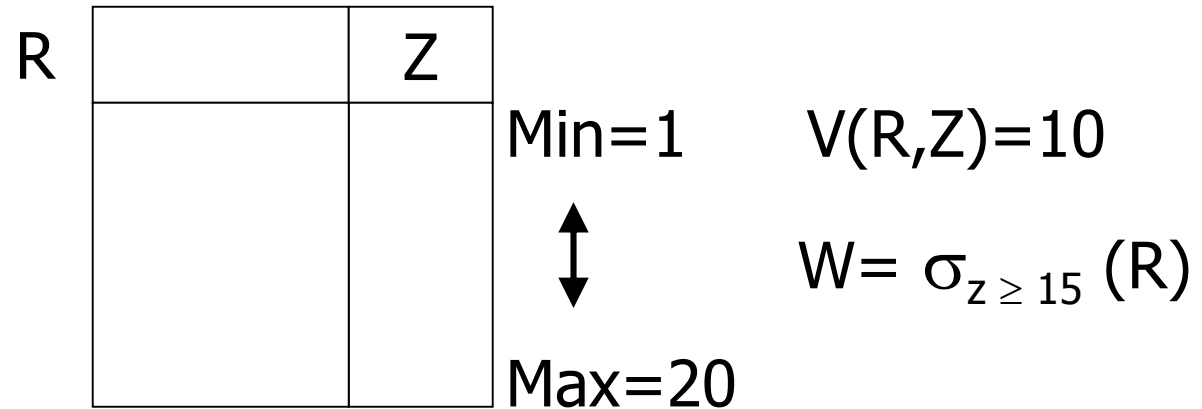
- $T(W) = T(R) / 3$

- 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

- Selektce  $W = \sigma_{Z \neq \text{val}}(R)$

- $T(W) = T(R) \cdot (1 - f(\text{val})) = T(R) \cdot (1 - 1/V(R,Z))$   
 $= T(R) - \frac{T(R)}{V(R,Z)}$

- Obvyklé řešení pro  $V(R,Z) \approx T(R)$

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

# 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

- $X \cap Y = Z$
- 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  $R_2$

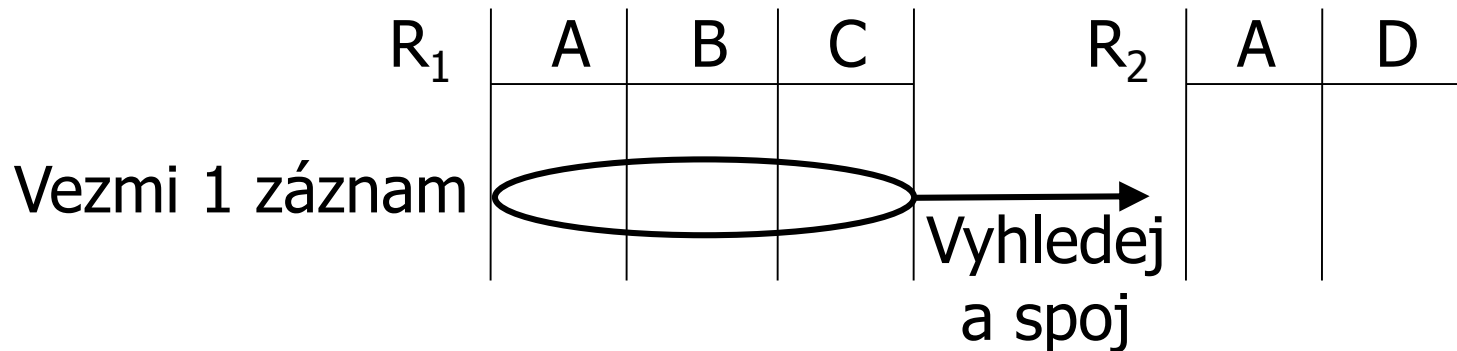
□  $V(R_1, A) \geq V(R_2, A)$

→ 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

- Opět předpoklad rovnoměrného rozložení

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

# Odhad velikosti: přirozené spojení

- Shrnutí obou variant

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

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

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

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

- Rozdíl je pouze ve jmenovateli

# Odhad velikosti: přirozené spojení

## ■ Obecný závěr

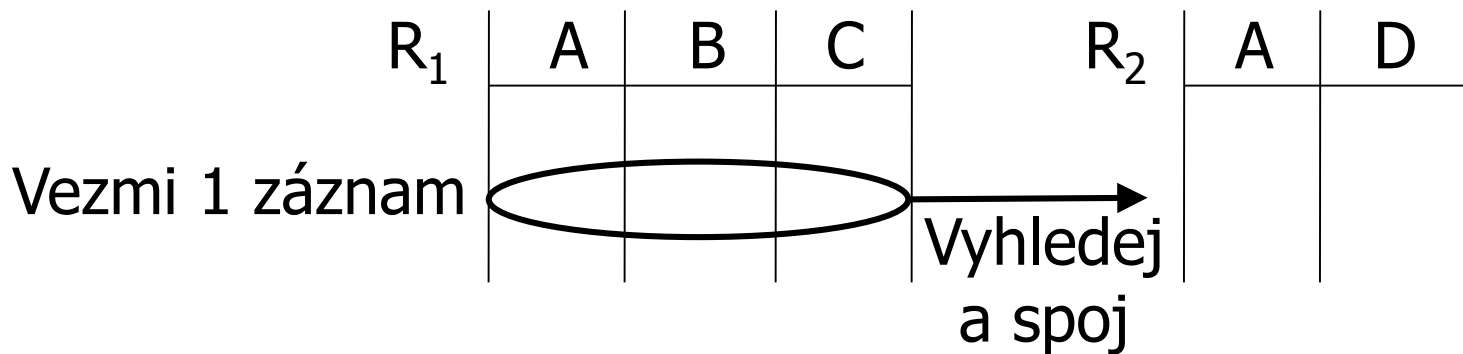
$$\square W = R_1 \bowtie R_2$$

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

# 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

- $W = R_1 \bowtie R_2$ 
  - $R_1(X), R_2(Y), X \cap Y = Z$
- Velikost záznamu
  - $S(W) = S(R_1) + S(R_2) - S(R_1, Z)$
  - Platí pro všechny varianty
- Počet záznamů, když  $Z$  má více atributů?
  - Předpokládáme, že jsou nezávislé.

$$T(W) = \frac{T(R_1) \cdot T(R_2)}{\max\{V(R_1, A_1), V(R_2, A_1)\} \cdot \max\{V(R_1, A_2), V(R_2, A_2)\}}$$

# Odhad velikosti: projekce, selekce

## ■ Projekce $\Pi_{AB}(R)$

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

- $S(W) = S(R, AB)$

## ■ Selekce $\sigma_{A=a \vee B=b}(R)$

- $S(W) = S(R)$ , necht'  $n = T(R)$

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

- $m_1 = T(R) / V(R, A)$

- $m_2 = T(R) / V(R, B)$

# Odhad velikosti: množinové operace

## ■ Sjednocení, průnik, rozdíl

□  $\cup, \cap, -$

- $T(W)$  – choose average size

□ Např.

- $T(R \cup S) = T(R) + T(S)$  ... if  $\cup$  means UNION ALL here
- $T(R \cup S) = [ \max\{T(R), T(S)\}, T(R) + T(S) ]$ 
  - So use:  $T(R \cup S) = \text{avg}\{ \max\{T(R), T(S)\}, T(R) + T(S) \}$ 
    - If *set union* is evaluated
- $T(R - S) = T(R) - \frac{1}{2} T(S)$
- $T(R \cap S) = \text{avg}\{ 0, \min\{T(R), T(S)\} \}$

# 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 odhady pro W 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 $R_1$

## ■ $U = \sigma_{A=a}(R_1)$

□  $T(U) = T(R_1) / V(R_1, A)$

| 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

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

- $V(U, A) = 1$

- $V(U, K) = T(U)$

- $K =$  primární klíč relace  $R_1$

- $V(U, *) = V(R, *)$  resp.  $V(U, *) = T(U)$

## ■ Výsledně lze využít původní $V(R, *)$

- $V(U, *) = \min \{ V(R, *), T(U) \}$

# Odhad počtu hodnot: spojení

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

- Výsledek:

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

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

- Resp.  $\min\{ V(R_1,B), T(U) \}$

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

# 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:

- $T(U) = T(R_1) \cdot T(R_2) / \max\{ V(R_1,B), V(R_2,B) \} =$   
 $= 1000 \cdot 2000 / 200 = 10\,000$
- $V(U,A) = 50$
- $V(U,B) = 100$
- $V(U,C) = 300$

# Odhad počtu hodnot: spojení

## ■ Celkový výsledek

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

$$\quad \blacksquare U(A,B,C)$$

## $\square$ Výsledek:

$$\blacksquare T(Z) = 10\,000 \cdot 3\,000 / 300 = 100\,000$$

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

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

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

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

# Odhad počtu hodnot: histogram

## ■ Histogram hodnot atributu

- Místo  $V(R,A)$  a  $DOM(R,A)$
- Zpřesnění odhadů

## ■ Počet různých hodnot

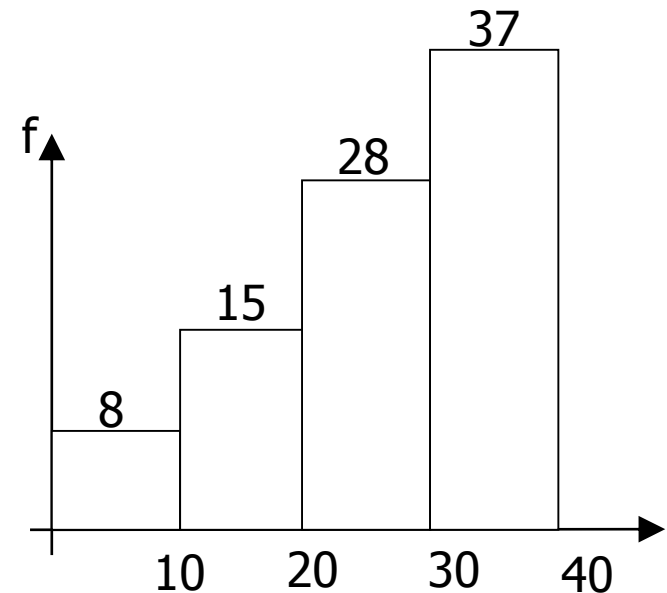
- Málo → pro každou počet
- Hodně → kvantizace

### ■ Stejně intervaly

### ■ Percentily

### ■ Pouze pro nejfrekventovanější

- ostatní dohromady (tj. výsledně rovnoměrně)





# Příklad statistik PostgreSQL

- Připojte se k fakultní DB PostgreSQL
  - Návod viz první přednáška
- Ve schématu *xdohnal* jsou tabulky
  - *predmet, skupina, hotel*
    - Statistiky jak na relacích, tak i attributech.
  - Významy jednotlivých polí
    - <http://www.postgresql.org/docs/9.2/interactive/view-pg-stats.html>

# Příklad statistik PostgreSQL

## ■ Tabulka hotel

| Statistic               | Value                        |
|-------------------------|------------------------------|
| Sequential Scans        | 4                            |
| Sequential Tuples Read  | 500                          |
| Index Scans             | 1                            |
| Index Tuples Fetched    | 500                          |
| Tuples Inserted         | 500                          |
| Tuples Updated          | 0                            |
| Tuples Deleted          | 0                            |
| Tuples HOT Updated      | 0                            |
| Live Tuples             | 500                          |
| Dead Tuples             | 0                            |
| Heap Blocks Read        | 5                            |
| Heap Blocks Hit         | 514                          |
| Index Blocks Read       | 4                            |
| Index Blocks Hit        | 599                          |
| Toast Blocks Read       |                              |
| Toast Blocks Hit        |                              |
| Toast Index Blocks Read |                              |
| Toast Index Blocks Hit  |                              |
| Last Vacuum             |                              |
| Last Autovacuum         |                              |
| Last Analyze            |                              |
| Last Autoanalyze        | 2010-04-15 13:52:03.54614+02 |
| Table Size              | 40 kB                        |
| Toast Table Size        | none                         |
| Indexes Size            | 32 kB                        |

# Příklad statistik PostgreSQL

## ■ Atribut hotel.id








| Statistic               | Value                                      |
|-------------------------|--------------------------------------------|
| Null Fraction           | 0                                          |
| Average Width           | 4                                          |
| Distinct Values         | -1                                         |
| Most Common Values      |                                            |
| Most Common Frequencies |                                            |
| Histogram Bounds        | {1,50,100,150,200,250,300,350,400,450,500} |
| Correlation             | 1                                          |

## ■ Atribut hotel.name








| Statistic               | Value                                                                                                       |
|-------------------------|-------------------------------------------------------------------------------------------------------------|
| Null Fraction           | 0                                                                                                           |
| Average Width           | 9                                                                                                           |
| Distinct Values         | -1                                                                                                          |
| Most Common Values      |                                                                                                             |
| Most Common Frequencies |                                                                                                             |
| Histogram Bounds        | {street1,street143,street189,street233,street279,street323,street369,street413,street459,street53,street99} |
| Correlation             | -0.117997                                                                                                   |

# Příklad statistik PostgreSQL

## ■ Atribut hotel.state

| Properties                                                                                                | Statistics | Dependencies                                                                          | Dependents |
|-----------------------------------------------------------------------------------------------------------|------------|---------------------------------------------------------------------------------------|------------|
| Statistic                                                                                                 |            | Value                                                                                 |            |
|  Null Fraction           |            | 0                                                                                     |            |
|  Average Width           |            | 7                                                                                     |            |
|  Distinct Values         |            | 50                                                                                    |            |
|  Most Common Values      |            | {state32,state8,state14,state36,state42,state48,state6,state16,state30,state47}       |            |
|  Most Common Frequencies |            | {0.038,0.03,0.028,0.028,0.028,0.028,0.028,0.026,0.026,0.026}                          |            |
|  Histogram Bounds        |            | {state1,state12,state18,state21,state25,state29,state34,state4,state44,state5,state9} |            |
|  Correlation             |            | -0.00743129                                                                           |            |

## ■ Atribut hotel.distance\_to\_center

| Properties                                                                                                  | Statistics | Dependencies                                                  | Dependents |
|-------------------------------------------------------------------------------------------------------------|------------|---------------------------------------------------------------|------------|
| Statistic                                                                                                   |            | Value                                                         |            |
|  Null Fraction            |            | 0                                                             |            |
|  Average Width           |            | 4                                                             |            |
|  Distinct Values         |            | 10                                                            |            |
|  Most Common Values      |            | {6,7,10,3,9,8,2,1,4,5}                                        |            |
|  Most Common Frequencies |            | {0.108,0.108,0.108,0.106,0.102,0.098,0.096,0.094,0.092,0.088} |            |
|  Histogram Bounds        |            |                                                               |            |
|  Correlation             |            | 0.102588                                                      |            |

# 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

- Statistiky 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