

ESTATÍSTICAS PARA ESTIMATIVA DE CUSTO

Sérgio Mergen

Estatísticas para estimativas de custo

- Através de estatísticas, é possível obter estimativas a respeito de quantos registros são retornados por cada operação
- Isso pode ajudar o otimizador a tomar decisões, como
 - Quais seleções fazer primeiro
 - A escolha da primeira seleção é importante, pois ela pode se valer de um índice
 - Quais junções fazer primeiro
 - Importante fazer primeiro junções que tragam menos registros
 - Quais algoritmos usar
 - O custo de cada algoritmo depende do número de registros. Sabendo esse número, dá para saber qual algoritmo tende a se sair melhor

Estatísticas para estimativa de custo

- Estatísticas usadas
 - $n(r)$: número de tuplas da relação r .
 - $v(A, r)$: cardinalidade do atributo A
 - número de valores distintos que aparecem em r para o atributo A
 - $\min(A, r)$ = menor valor em r para o atributo A
 - $\max(A, r)$ = maior valor em r para o atributo A

Estatísticas de exemplo

- Número de registros

n(movie)	=	4000
n(person)	=	50.000
n(movie_cast)	=	100.000

movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order, c_name)

idM referencia movie

idP referencia person

- Cardinalidades

v(idM, movie)	=	4000
v(idP, person)	=	50.000
v(<idM, idP>, movie_cast)	=	100.000
v(idM, movie_cast)	=	3.800
v(idP, movie_cast)	=	3.500
v(year, movie)	=	50

- Máximos e mínimos

min(year, movie)	=	1921
max(year, movie)	=	2020

ESTIMATIVAS PARA SELEÇÕES

Seletividade (probabilidade)

- Cada filtro em uma consulta é considerado como um evento independente

- Ex.

```
SELECT * FROM MOVIE  
WHERE year = 2000 AND budget > 2M
```

- Um registro ter **year = 2000** forma outro evento A
- Um registro ter **budget > 2M** forma um evento B

Seletividade (probabilidade)

- A seletividade de um filtro A está relacionada à probabilidade $P(A)$ de que o evento associado ao filtro de fato aconteça
- Considere que
 - a tabela **movie** tenha 4.000 registros
 - E existe 2% de chance de que um registro tenha year = 2000
 - $P(A) = 0.02$ (seletividade de year = 2000)
 - Ao aplicar esse filtro, a estimativa de retorno seria de 2% dos 4.000 registros
 - $4.000 * 0.02 = 80$ registros
- Diz-se que um valor mais alto de probabilidade significa que um filtro é pouco seletivo

Estimativa com filtros simples

- **WHERE tab.A = valor**
 - seletividade do filtro = $1/v(A, \text{tab})$
- Exemplo:

```
SELECT *  
FROM movie  
WHERE year = 1980
```

$$P(A) = \frac{1}{50} \quad [\text{year} = 1980]$$

$$\text{Estimativa} = 4.000 * \frac{1}{50} = 80$$

n (movie)	= 2.000
v (year, movie)	= 50

Estimativa com filtros simples

- **WHERE tab.A = valor**
 - Na falta de informações estatísticas, deve-se tentar adivinhar um valor de seletividade.
 - Ex. Seletividade = $\frac{1}{2}$.
- Exemplo:

```
SELECT *  
FROM movie  
WHERE budget = 2M
```

n (movie)	= 2.000
v (budget, movie)	= ???

$$P(A) = \frac{1}{2} \quad [\text{budget} = 2M]$$

$$\text{Estimativa} = 4.000 * \frac{1}{2} = 2000$$

Estimativa com filtros simples

- **WHERE tab.A <= v**
- Se $\min(A, \text{tab})$ e $\max(A, \text{tab})$ estiverem presentes no catálogo de estatísticas
 - se $v < \min(A, \text{tab})$
seletividade = 0
 - senão
$$\text{seletividade} = \frac{v - \min(A, \text{tab}) + 1}{\max(A, \text{tab}) - \min(A, \text{tab}) + 1}$$
- Na falta de informações estatísticas
 - seletividade = $\frac{1}{2}$.

Estimativa com filtros simples

- Exemplo:

```
SELECT *  
FROM movie  
WHERE year <= 1910
```

$P(A) = 0$ [year <= 1910]
estimativa = $4.000 * 0 = 0$

n (movie)	=	4000
min(year, movie)	=	1921
max(year, movie)	=	2020

Estimativa com filtros simples

- Exemplo:

n (movie)	=	4000
min(year, movie)	=	1921
max(year, movie)	=	2020

SELECT *
FROM movie
WHERE year <= 1950

$$P(A) = \frac{1950 - 1921 + 1}{2020 - 1921 + 1} \quad [\text{year} \leq 1950]$$

$$P(A) \approx \frac{30}{100} \approx 0,3$$

$$\text{estimativa} \approx 4000 * 0,3 \approx 1200$$

Estimativa com filtros simples

- Exemplo:

n (movie)	=	4000
min(budget, movie)	=	?
max(budget, movie)	=	?

SELECT *
FROM movie
WHERE budget < 1M

$$P(A) = \frac{1}{2}$$

$$\text{estimativa} = 4000 * \frac{1}{2} = 2000$$

Estimativa com negação

- A probabilidade de um evento A não aconteça (\bar{A}) é equivalente ao complemento da probabilidade de que ele aconteça
 - $P(\bar{A}) = 1 - P(A)$

Estimativa com negação

- Exemplo: Quantas tuplas satisfazem $\text{idM} \neq 40$?

$$P(A) = \frac{1}{4000} \quad [\text{idM} = 40]$$

$$P(\bar{A}) = 1 - \frac{1}{4000} = \frac{3999}{4000} \quad [\text{idM} \neq 40]$$

$$\textbf{Estimativa} = 4.000 * \frac{3999}{4000} = 3999$$

- Exemplo: Quantas tuplas satisfazem $\text{year} > 1950$?

$$P(A) = \frac{30}{1000} \quad [\text{year} \leq 1950]$$

$$P(\bar{A}) = 1 - \frac{30}{1000} = \frac{970}{1000} \quad [\text{year} > 1950]$$

$$\textbf{Estimativa} = 4000 * \frac{970}{1000} = 3880$$

Estimativa com filtros compostos

- **Conjunção de dois filtros:** $\sigma_{\theta 1} \wedge \sigma_{\theta 2}$
- Os registros que satisfazem múltiplos filtros conjuntivos podem ser vistos como a interseção (\cap) de eventos independentes
- Pela teoria da probabilidade, a probabilidade de intersecção de dois eventos independentes A e B é dada pela seguinte fórmula

$$P(A \cap B) = P(A) \cdot P(B)$$

Estimativa com filtros compostos

SELECT *

FROM movie

WHERE year <= 1950 **AND** budget = 1M

$P(A) = 0,3$ [year <=1950]

$P(B) = 0,5$ [budget = 1M]

$P(A \cap B) = 0,3 * 0,5 = 0,15$

Estimativa = $4000 * 0,15 = 600$

Estimativa com filtros compostos

- **Conjunção de múltiplos filtros:** $\sigma_{\theta 1} \wedge \sigma_{\theta 2} \dots \wedge \sigma_{\theta n}$
- A conjunção de múltiplos eventos (mais do que dois) é dada pelo produto de todas as probabilidades individuais

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B) \cdot P(C) \cdot P(D)$$

Estimativa com filtros compostos

- **Disjunção de dois filtros** $\sigma_{\theta 1} \vee \sigma_{\theta 2}$
- Os registros que satisfazem pelo menos um dos filtros podem ser vistos como a união (U) de eventos independentes
- Pela teoria da probabilidade, a probabilidade da união de dois eventos independentes A e B é dada por uma fórmula que soma probabilidades individuais e descarta as intersecções

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Estimativa com filtros compostos

SELECT *

FROM movie

WHERE year <= 1950 **OR** budget = 1M

$$P(A) = 0,3 \quad [\text{year} \leq 1950]$$

$$P(B) = 0,5 \quad [\text{budget} = 1\text{M}]$$

$$P(A \cap B) = 0,15$$

$$P(A \cup B) = 0,3 + 0,5 - 0,15 = 0,65$$

$$\text{Estimativa} \rightarrow 4000 * 0,65 \rightarrow 2600$$

Estimativa com filtros compostos

- **Disjunção de dois filtros** $\sigma_{\theta 1} \vee \sigma_{\theta 2}$
- Fórmula geral
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Em alguns casos, os filtros são mutuamente exclusivos
 - Ex. custo = 2000 OR custo = 3000
- Nesses casos, $P(A \cap B)$ é zero. Isso faz com que fórmula seja simplificada
 - $P(A \cup B) = P(A) + P(B)$

Estimativa com filtros compostos

SELECT *

FROM proj

WHERE year = 1950 **OR** year = 2000

$$P(A) = \frac{1}{50} \quad [\text{year} = 1950]$$

$$P(B) = \frac{1}{50} \quad [\text{year} = 2000]$$

$$P(A \cup B) = \frac{1}{50} + \frac{1}{50} = \frac{1}{25}$$

$$\text{Estimativa} = 4000 * \frac{1}{25} = 160$$

Estimativa com filtros compostos

- **Disjunção de múltiplos filtros** $\sigma_{\theta 1} \vee \sigma_{\theta 2} \dots \vee \sigma_{\theta n}$
- A disjunção de múltiplos eventos (mais do que dois) é dada por uma fórmula que soma ou descarta interseções
- Exemplos para três eventos
- $$P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - (P(A \cap B) + P(A \cap C) + P(B \cap C)) \\ + P(A \cap B \cap C)$$

A não ser que os eventos sejam mutuamente exclusivos, a fórmula pode se tornar complexa para um número grande de eventos

Estimativa com filtros compostos

- **Disjunção de múltiplos filtros** $\sigma_{\theta 1} \vee \sigma_{\theta 2} \dots \vee \sigma_{\theta n}$
- Outra forma de encontrar a união de eventos independentes é realizando a seguinte transformação de DeMorgan
 - $(A \vee B \vee C) = \neg (\neg A \wedge \neg B \wedge \neg C)$
- Sendo assim, nós temos
 - $P(A \cup B \cup C) = \overline{P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})}$
 - $= 1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})$

Estimativa com filtros compostos

SELECT *
FROM proj

WHERE year <= 1950 **OR** budget = 1M

o mesmo que

WHERE NOT (year > 1950 **AND** budget <> 1M)

$$P(A) = 0,3 \quad [\text{year} \leq 1950]$$

$$P(\bar{A}) = 1 - 0,3 = 0,7 \quad [\text{year} > 1950]$$

$$P(B) = 0,5 \quad [\text{ano} = 2014]$$

$$P(\bar{B}) = 1 - 0,5 = 0,5 \quad [\text{ano} \neq 2014]$$

$$P(\bar{A}) \cdot P(\bar{B}) = 0,7 * 0,5 = 0,35$$

$$P(\overline{0,45}) = 1 - 0,35 = 0,65$$

$$\text{Estimativa} = 4000 * 0,65 \rightarrow 2600$$

ESTIMATIVAS PARA JUNÇÕES

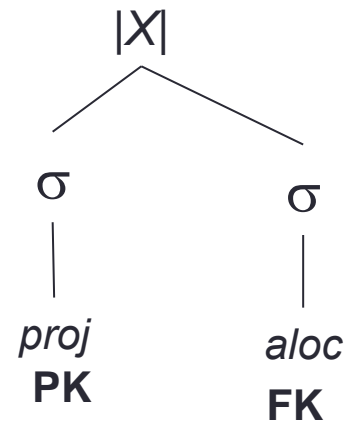
Estimativas do MySQL – Junção

- Em um plano de execução de consulta, praticamente todas operações binárias são tipos de junção
 - Assim, as estimativas referentes à junções são muito importantes
- Duas possibilidades principais
 - As colunas da junção têm relação de chave primária/estrangeira
 - Não existe critério de junção

Estimativas do MySQL – Junção

Quando as colunas da junção têm relação de chave primária/estrangeira

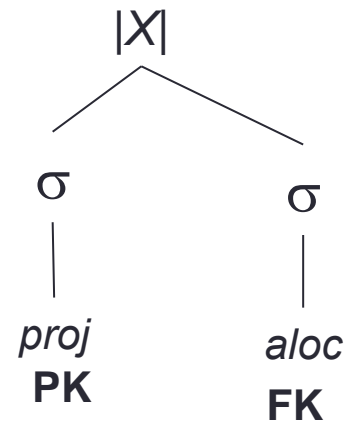
- Considerando que
 - PK: tabela onde está a chave primária
 - FK: tabela onde está a chave estrangeira:
 - $n(PK)$: número de registros que chegam de A
 - $n(FK)$: número de registros que chegam de B



Estimativas do MySQL – Junção

Quando as colunas da junção têm relação de chave primária/estrangeira

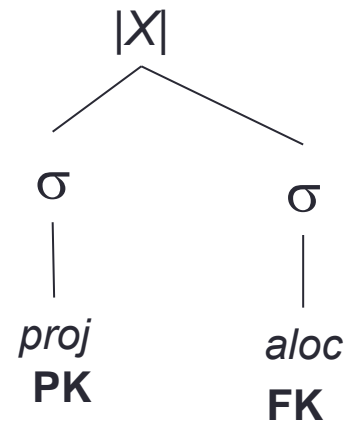
- Considerando que
 - $\text{sel}(\text{PK})$: seletividade obtida do lado PK
 - $\text{sel}(\text{FK})$: seletividade obtida do lado FK



Estimativas do MySQL – Junção

Quando as colunas da junção têm relação de chave primária/estrangeira

- O cálculo é
 - $n(\text{FK}) * \text{sel}(\text{PK})$

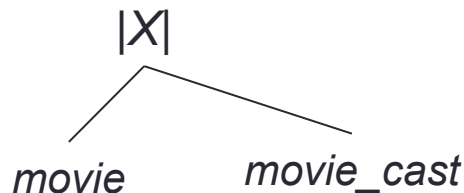


O máximo de registros gerados é equivalente ao tamanho da tabela que contém a FK

Estimativas do MySQL – Junção

- Exemplo
 - Junção por chave primária/chave estrangeira **sem filtros**

```
SELECT *  
FROM movie NATURAL JOIN  
movie_cast
```



movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

Estatísticas

n(movie) = 4000

n(movie_cast) = 10000

v(year, movie) = 50

v(order, movie_cast) = 100

Estimativas do MySQL – Junção

- Cálculo da junção
 - = $n(\text{FK}) * \text{sel}(\text{PK})$
 - = $100000 * 1$
 - = **100000**

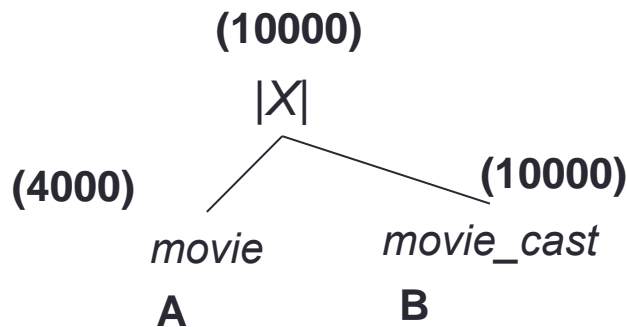
movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order, c_name)

idM referencia movie

idP referencia person



Estatísticas

$n(\text{movie}) = 4000$

$n(\text{movie_cast}) = 10000$

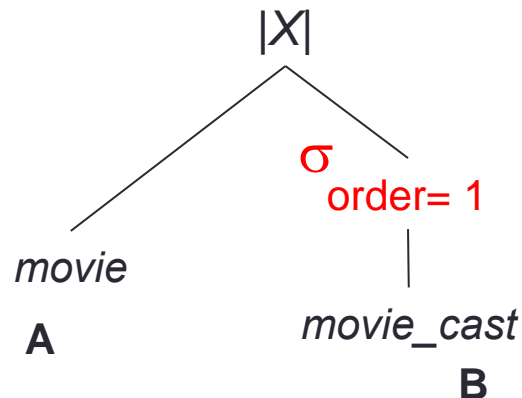
$v(\text{year, movie}) = 50$

$v(\text{order, movie_cast}) = 100$

Estimativas do MySQL – Junção

- Exemplo
 - Junção por chave primária/chave estrangeira **com filtro na tabela onde está a chave estrangeira**

```
SELECT *  
FROM movie NATURAL JOIN  
movie_cast mc  
WHERE mc.order = 1
```



movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

Estatísticas

n(movie) = 4000

n(movie_cast) = 10000

v(year, movie) = 50

v(order, movie_cast) = 100

Estimativas do MySQL – Junção

- Cálculo da junção
 - = $n(\text{FK}) * \text{sel}(\text{PK})$
 - = $100 * 1$
 - = **100**

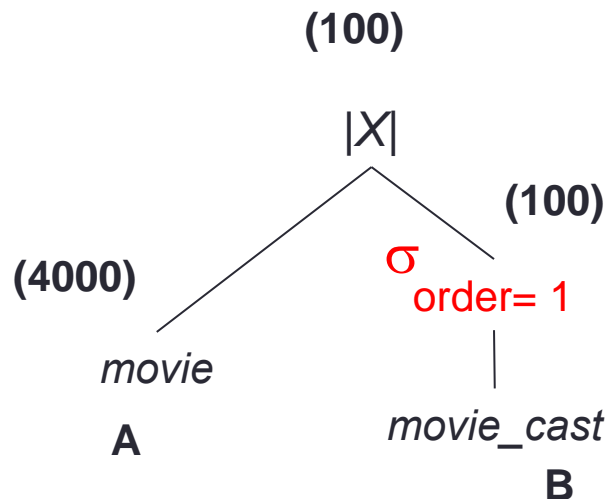
movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order, c_name)

idM referencia movie

idP referencia person



Estatísticas

$n(\text{movie}) = 4000$

$n(\text{movie_cast}) = 10000$

$v(\text{year, movie}) = 50$

$v(\text{order, movie_cast}) = 100$

Estimativas do MySQL – Junção

- Exemplo
 - Junção por chave primária/chave estrangeira **com filtro na tabela onde está a chave primária**

```
SELECT *  
FROM movie m NATURAL JOIN  
movie_cast  
WHERE m.year = 1980
```

movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

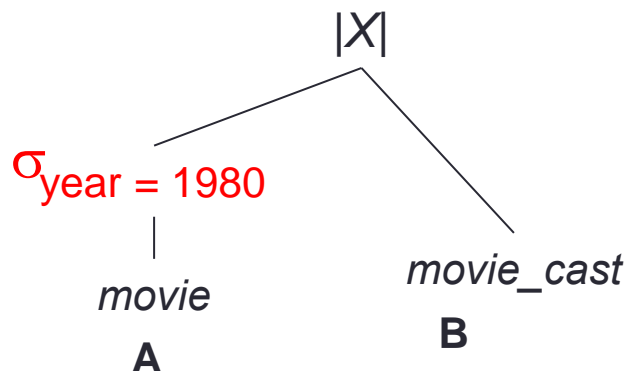
Estatísticas

n(movie) = 4000

n(movie_cast) = 10000

v(year, movie) = 50

v(order, movie_cast) = 100



Estimativas do MySQL – Junção

- Cálculo da junção
 - $= n(\text{FK}) * \text{sel}(\text{PK})$
 - $= 10000 * 1/50$
 - **= 200**

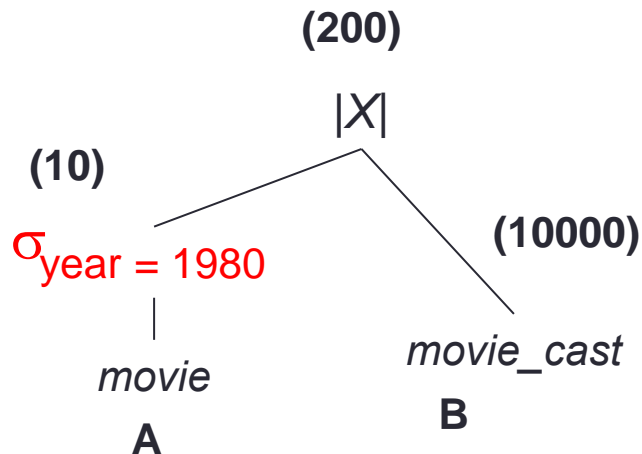
movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order, c_name)

idM referencia movie

idP referencia person



Estatísticas

$n(\text{movie}) = 4000$

$n(\text{movie_cast}) = 10000$

$v(\text{year, movie}) = 50$

$v(\text{order, movie_cast}) = 100$

Estimativas do MySQL – Junção

- Exemplo
 - Junção por chave primária/chave estrangeira **com filtro nas duas tabelas**

```
SELECT *  
FROM movie NATURAL JOIN  
movie_cast  
WHERE year= 1950 and  
order = 1
```

movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

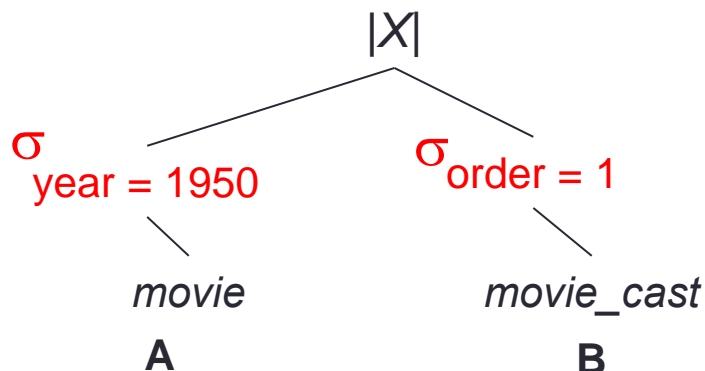
Estatísticas

n(movie) = 4000

n(movie_cast) = 10000

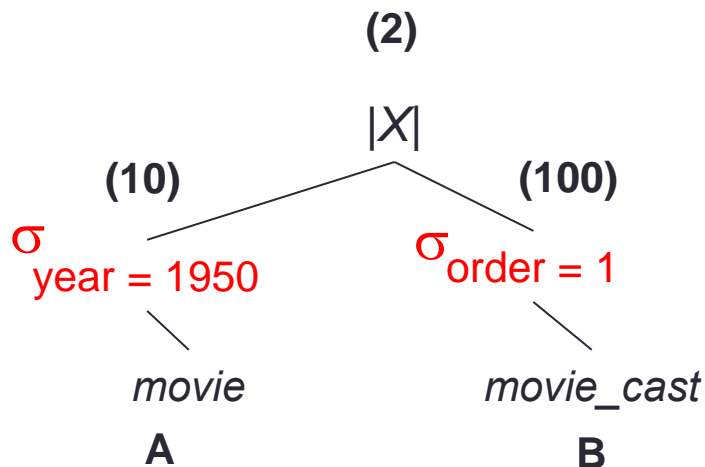
v(year, movie) = 50

v(order, movie_cast) = 100



Estimativas do MySQL – Junção

- Cálculo da junção
 - $= n(\text{FK}) * \text{sel}(\text{PK})$
 - $= 100 * 1/50$
 - $= 2$



Estatísticas

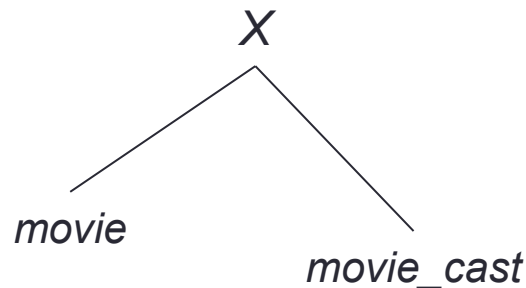
$n(\text{movie})$	=	4000
$n(\text{movie_cast})$	=	10000
$v(\text{year}, \text{movie})$	=	50
$v(\text{order}, \text{movie_cast})$	=	100

Estimativas do MySQL – Junção

- Em um plano de execução de consulta, praticamente todas operações binárias são tipos de junção
 - Assim, as estimativas referentes à junções são muito importantes
- Duas possibilidades principais
 - As colunas da junção têm relação de chave primária/estrangeira
 - Não existe critério de junção

Estimativas do MySQL – Junção

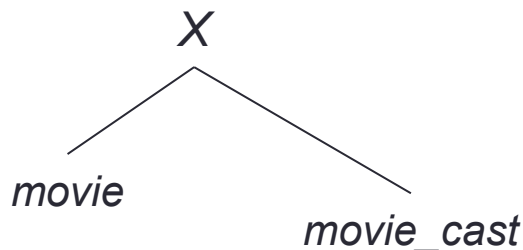
- Nesse caso, cada registro de um lado cruza com todos os registros do outro lado
- A estimativa é simplesmente o produto da quantidade de registros que chega em cada lado da junção
 - $n(\text{tab1}) * n(\text{tab2})$



Estimativas do MySQL – Junção

- Exemplo
 - Cruzamento sem nenhum critério de junção

```
SELECT *  
FROM movie CROSS JOIN  
movie_cast
```



movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

Estatísticas

n(movie) = 4000

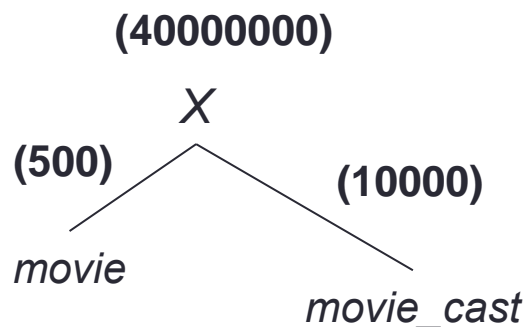
n(movie_cast) = 10000

v(year, movie) = 50

v(order, movie_cast) = 100

Estimativas do MySQL – Junção

- Cálculo da junção
 - = $n(\text{movie}) * n(\text{movie_cast})$
 - = $4000 * 10000$
 - = **40000000**



movie (idM, title, year, budget)

person (idP, name)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

Estatísticas

$n(\text{movie}) = 4000$

$n(\text{movie_cast}) = 10000$

$v(\text{year, movie}) = 50$

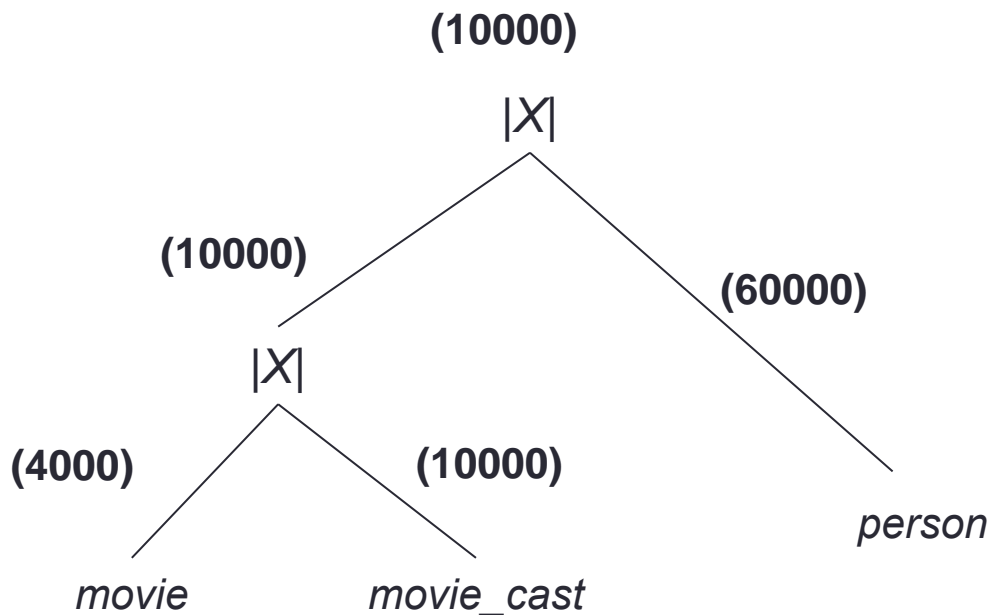
$v(\text{order, movie_cast}) = 100$

Estimativa com composição de junções

- Vimos como estimar o resultado de uma junção
 - Mas uma expressão geralmente possui várias junções
- Como estimar o resultado de uma expressão complexa?
 - Deve-se estimar o resultado das operações de nível mais baixo primeiro
 - O resultado é então usado para estimar o resultado das operações de nível mais alto

Estimativa com composição de junções

- Ex.



movie (idM, title, year, budget)

person (idP, name, year)

movie_cast (idM, idP, order,
c_name)

idM referencia movie

idP referencia person

Estatísticas

n(movie) = 4000

n(movie_cast) = 10000

n(person) = 60000

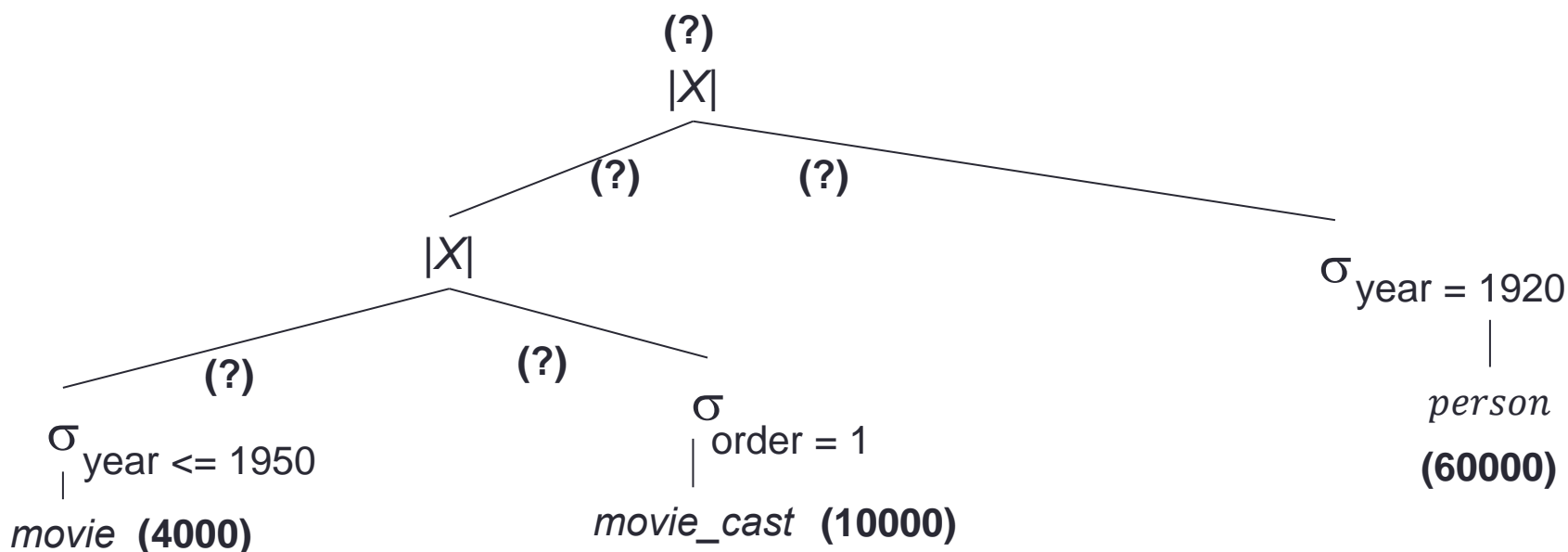
v(year, movie) = 50

v(year, person) = 30

v(order, movie_cast) = 100

Estimativa com composição de junções

- Como estimar nesse caso?

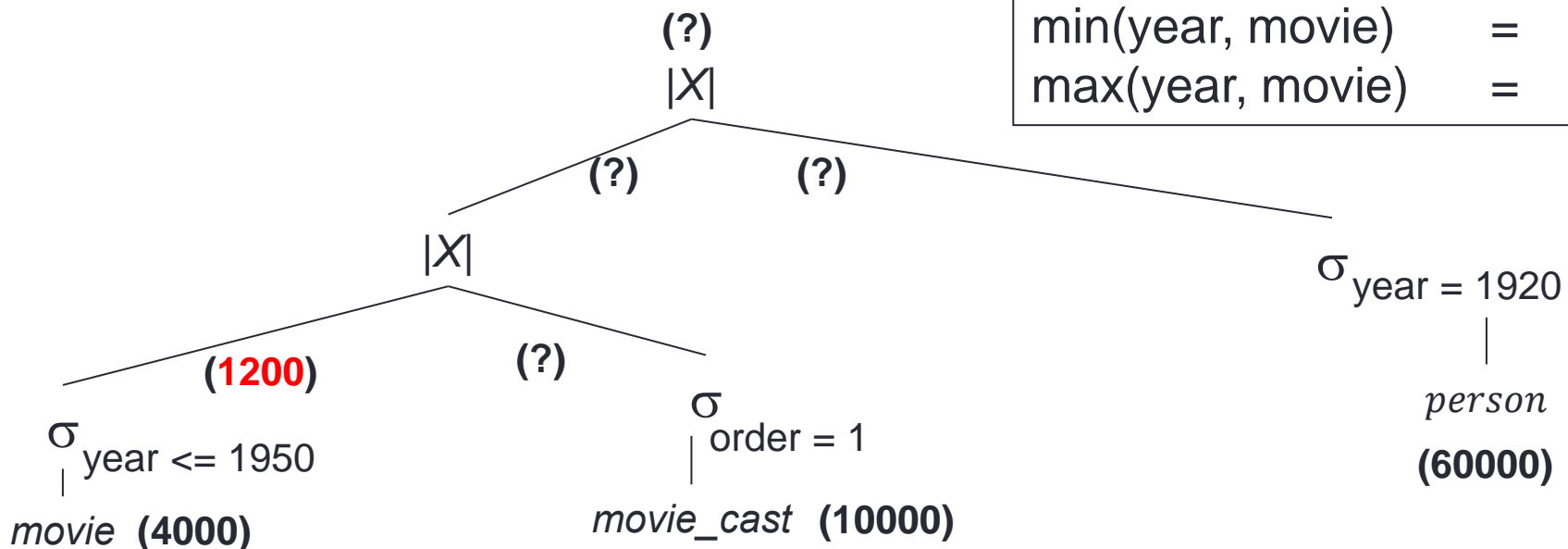


Estimativa com composição de junções

- $\text{year} \leq 1950$
 - $= n(\text{movie}) * (v - \min(\text{year}) + 1) / (\max(\text{year}) - \min(\text{year}) + 1)$
 - $= 4000 * (1950 - 1921 + 1) / (2020 - 1921 + 1)$
 - $= 4000 * 30 / 100$
 - **= 1200**

Estatísticas

$n(\text{movie})$	=	4000
$\min(\text{year}, \text{movie})$	=	1921
$\max(\text{year}, \text{movie})$	=	2020



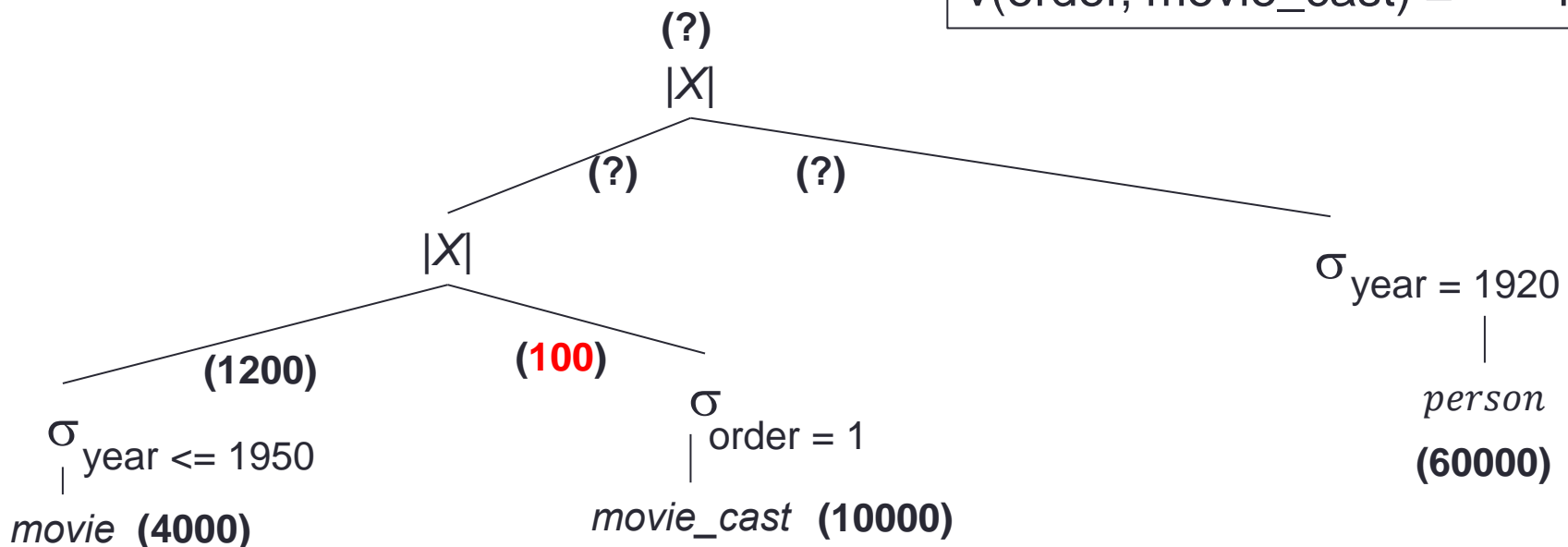
Estimativa com composição de junções

- $order=1$
 - $= n(movie_cast) * 1/v(order, movie_cast)$
 - $= 10000 * 1/100$
 - $= \mathbf{100}$

Estatísticas

$n(movie_cast) = 10000$

$v(order, movie_cast) = 100$

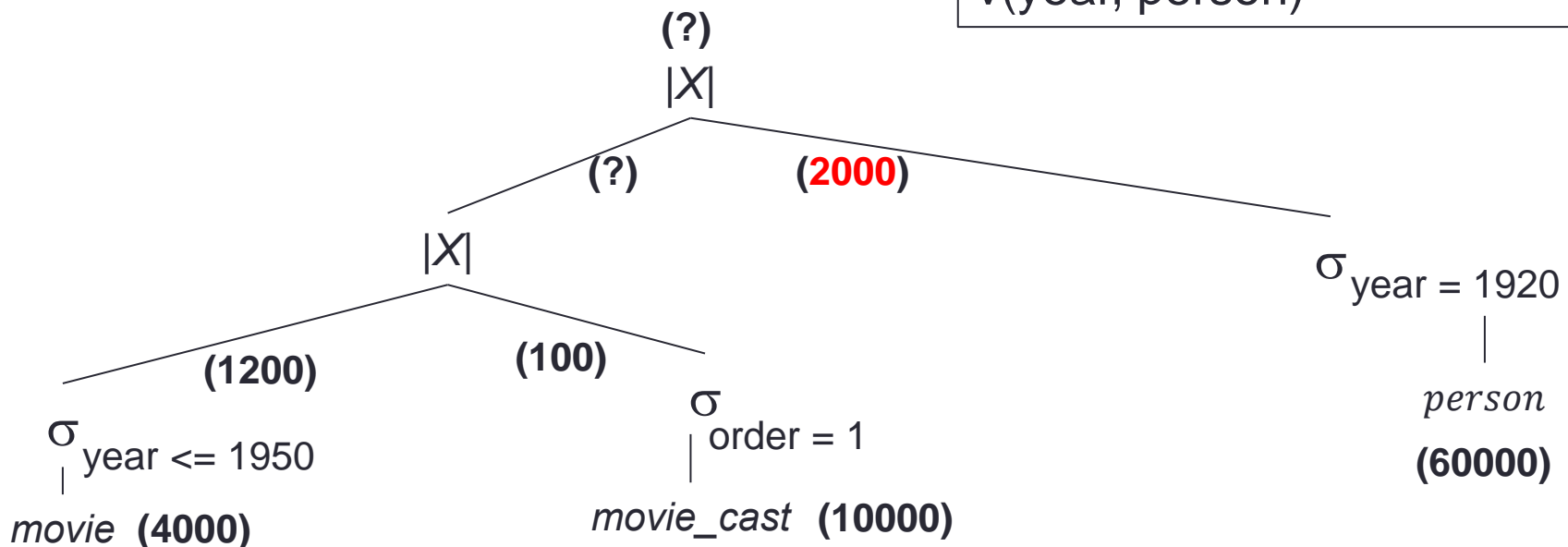


Estimativa com composição de junções

- $\text{year} = 1920$
 - $= n(\text{person}) * 1 / v(\text{year}, \text{person})$
 - $= 60000 * 1/30$
 - $= 2000$

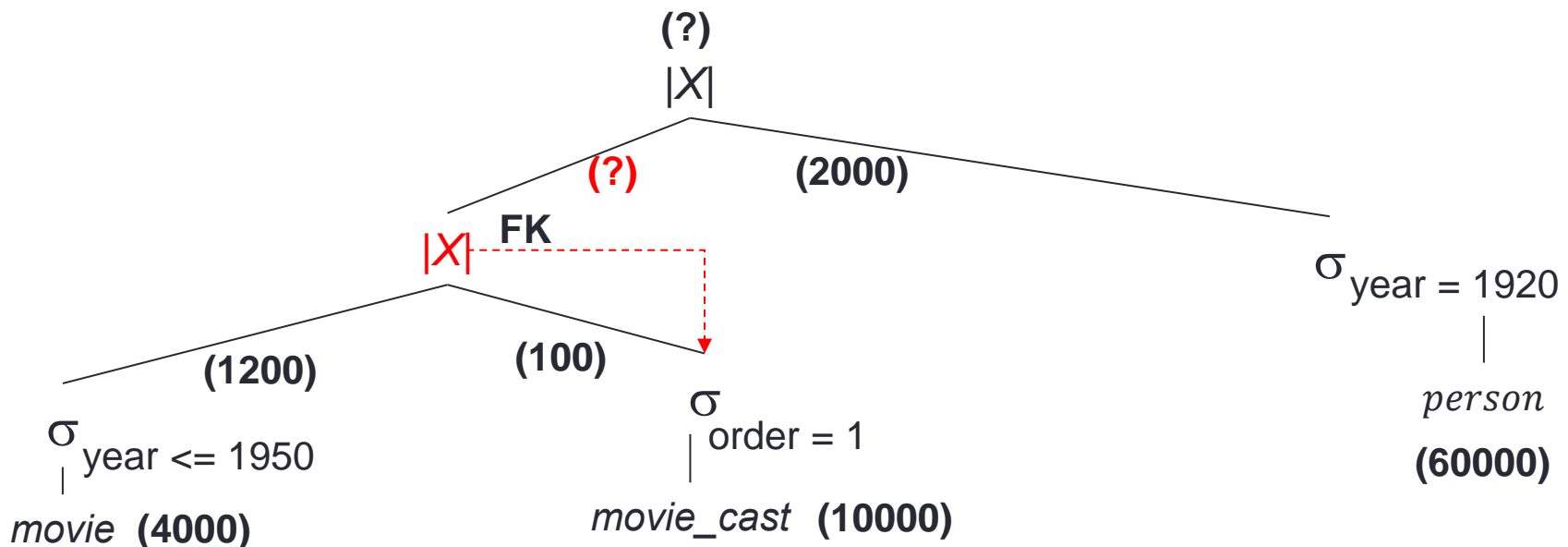
Estatísticas

$n(\text{person})$	$= 60000$
$v(\text{year}, \text{person})$	$= 30$



Estimativa com composição de junções

- Nesta primeira junção
 - a chave estrangeira está do lado direito

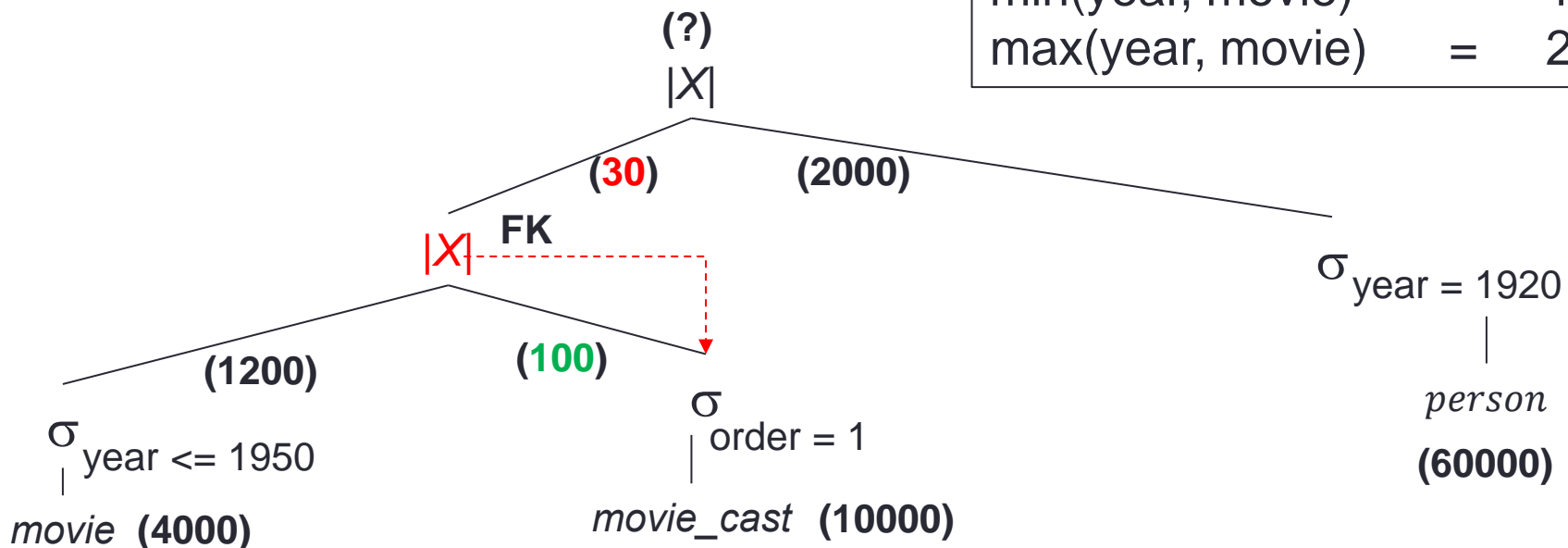


Estimativa com composição de junções

- primeira junção
 - $= n(\text{FK}) * \text{sel}(\text{PK})$
 - $= n(\text{FK}) * \text{sel}(\text{year} \leq 1950)$
 - $= 100 * (v - \min(\text{year}) + 1) / (\max(\text{year}) - \min(\text{year}) + 1)$
 - $= 100 * (1950 - 1921 + 1) / (2020 - 1921 - 1)$
 - $= 100 * 30 / 100$
 - $= 30$

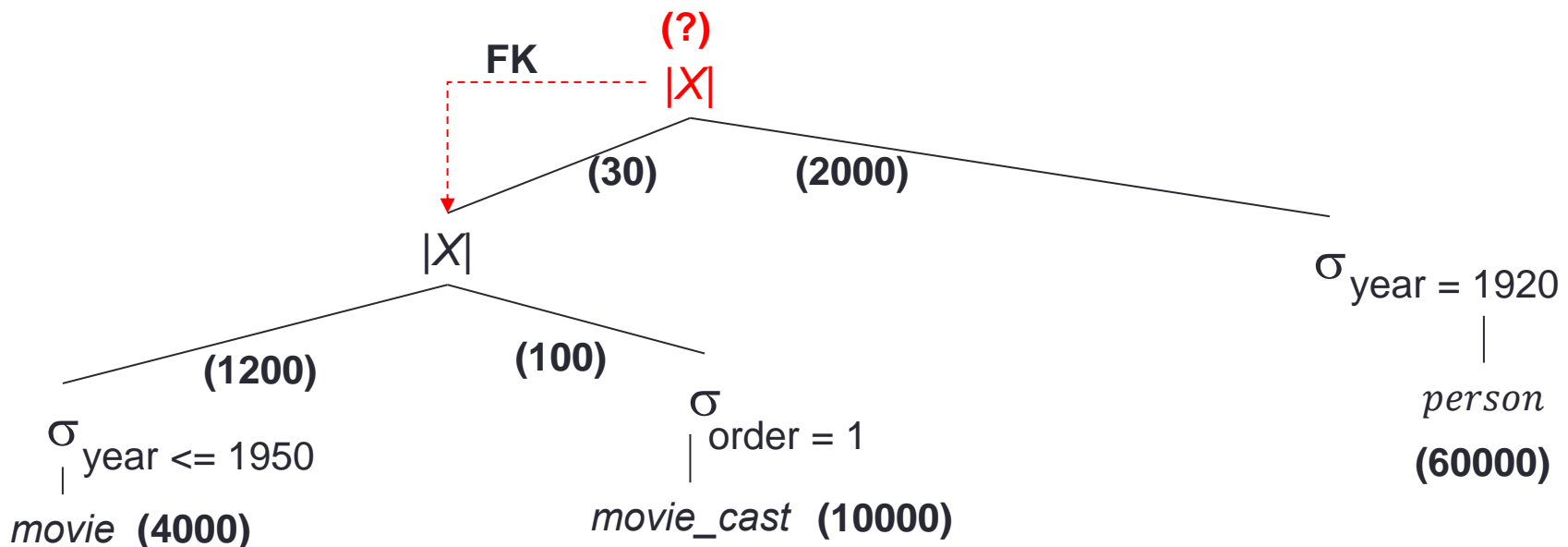
Estatísticas

$\min(\text{year}, \text{movie})$	=	1921
$\max(\text{year}, \text{movie})$	=	2020



Estimativa com composição de junções

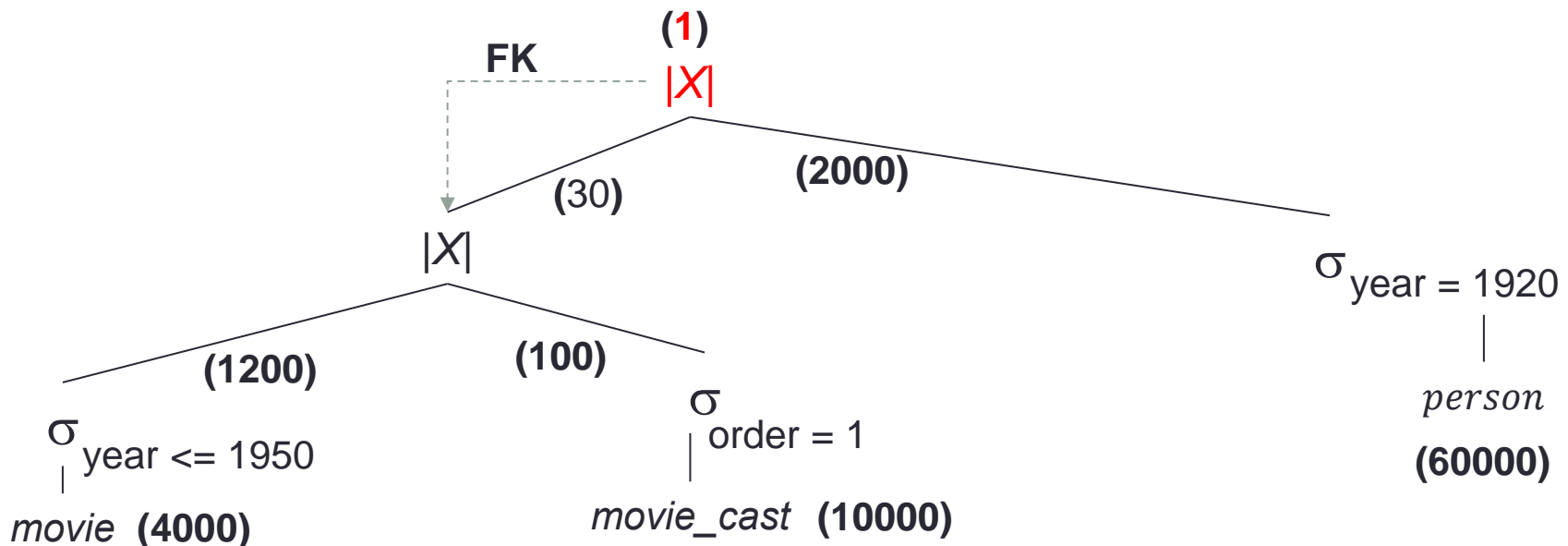
- Na segunda junção
 - a chave estrangeira está do lado esquerdo



Estimativa com composição de junções

- segunda junção
 - $= n(\text{FK}) * \text{sel}(\text{PK})$
 - $= n(\text{FK}) * \text{sel}(\text{year} = 1920)$
 - $= 30 * 1/v(\text{year}, \text{person})$
 - $= 30 * 1/30$
 - $= 1$

Estatísticas

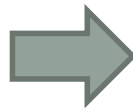
$$v(\text{year, person}) = 30$$


ESTIMATIVAS PARA OUTRAS OPERAÇÕES

Estimativa para operações de conjunto

- Para intersecção de seleções sobre a mesma relação
 - Reescrever usando **conjunção** de filtros
 - Encontrar a estimativa para a expressão reescrita
 - Ex. $\sigma_{\theta_1}(r) \cap \sigma_{\theta_2}(r) \rightarrow \sigma_{\theta_1 \wedge \theta_2}(r)$

```
SELECT title
FROM movie
WHERE year <= 1950
INTERSECT
SELECT title
FROM movie
WHERE budget = 1M
```

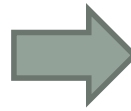


```
SELECT title
FROM movie
WHERE year <= 1950
AND      budget = 1M
```

Estimativa para operações de conjunto

- Exemplo

```
SELECT title  
FROM movie  
WHERE year <= 1950  
INTERSECT  
SELECT title  
FROM movie  
WHERE budget = 1M
```



```
SELECT title  
FROM movie  
WHERE year <= 1950  
AND      budget = 1M
```

n (movie)	=	4000
min(year, movie)	=	1921
max(year, movie)	=	2020

$P(A) = 0,3$ [year <= 1950]

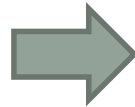
$P(B) = 0,5$ [budget = 1M]

Estimativa = $4000 * 0,3 * 0,5 = 600$

Estimativa para operações de conjunto

- Para união de seleções sobre a mesma relação
 - Reescrever usando **disjunção** de filtros
 - Encontrar a estimativa para a expressão reescrita
 - Ex. $\sigma_{\theta_1}(r) \cup \sigma_{\theta_2}(r) = \sigma_{\theta_1 \vee \theta_2}(r)$

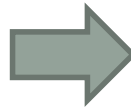
```
SELECT title
FROM movie
WHERE year <= 1950
UNION
SELECT title
FROM movie
WHERE budget = 1M
```



```
SELECT title
FROM movie
WHERE year <= 1950
OR      budget = 1M
```


Estimativa para operações de conjunto

```
SELECT title  
FROM movie  
WHERE year <= 1950  
UNION  
SELECT title  
FROM movie  
WHERE budget = 1M
```



```
SELECT title  
FROM movie  
WHERE year <= 1950  
OR      budget = 1M
```

n (movie)	=	4000
min(year, movie)	=	1921
max(year,movie)	=	2020

$P(A) = 0,3$ [year <= 1950]

$P(B) = 0,5$ [budget = 1M]

Estimativa = $4000 * (1 - (0,7 * 0,5)) = 2600$

Estimativa para operações de conjunto

- Para operações em relações diferentes
 - $r \cup s$
 - tamanho de r + tamanho de s .
 - Estimativa pouco precisa, mas pelo menos estabelece um teto

- Exemplo:

n (movie)	=	4000
n (movie_cast)	=	10000

```
SELECT title
FROM movie
UNION
SELECT c_name
FROM movie_cast
```

Estimativa = 14000

Estimativa para operações de conjunto

- Para operações em relações diferentes
 - $r \cap s$
 - *Min* (tamanho de r, tamanho de s).
 - Estimativa pouco precisa, mas pelo menos estabelece um teto

- Exemplo:

n (movie)	=	4000
n (movie_cast)	=	10000

```
SELECT title
FROM movie
INTERSECT
SELECT c_name
FROM movie_cast
```

Estimativa = 4000

Estimativa para operações de conjunto

- Para operações em relações diferentes
 - $r - s$
 - r .
 - Estimativa pouco precisa, mas pelo menos estabelece um teto

- Exemplo:

n (movie)	=	4000
n (movie_cast)	=	10000

```
SELECT title
FROM movie
EXCEPT
SELECT c_name
FROM movie_cast
```

Estimativa = 4000

Estimativa para projeção

- Projeção: $\Pi_A(r)$
 - Estimativa é $v(A, r)$
- Exemplo:

SELECT DISTINCT year
FROM movie

Estimativa = 50

n (movie)	=	4000
v (year, movie)	=	50

Estimativa para agregação

- Agregação: $A \mathbf{G}_F(r)$
 - Estimativa é $v(A, r)$

- Exemplo:

```
SELECT year, count(*)  
FROM movie  
GROUP BY year
```

Estimativa = 50

n (movie)	=	4000
v (year, movie)	=	50

Demais estimativas

- Pode-se obter estimativas para outras operações
 - Junções externas
 - Número de valores distintos em seleções com vários atributos ou filtros
 - ...
- Na maior parte dos casos, sua obtenção é bastante intuitiva
- Ex. Como estimar o tamanho do resultado nesse caso?

```
SELECT DISTINCT year  
FROM movie  
WHERE year <= 2000
```

- A análise desse e de outros casos fica como exercício

Atividade Individual

- Com base no banco de dados de movie
 - Descubra as estatísticas referentes a cada tabela
 - Estime a quantidade de registros que cada consulta deverá retornar
 - Usando as estatísticas descobertas
 - Compare a estimativa com a quantidade que é realmente retornada para cada consulta
 - Algumas estimativas devem ser pouco precisas. Indique o motivo.

Atividade Individual

- Complete com as estatísticas descobertas

N (movie)	
N (person)	
N (movie_cast)	
V(year,movie)	
Min(year, movie)	
Max(year, movie)	
V(cast_order, movie_cast)	

Atividade Individual

- Consultas a serem executadas:

1. **SELECT * FROM movie CROSS JOIN person**

2. **SELECT * FROM movie NATURAL JOIN movie_cast
WHERE cast_order = 1 AND release_year > 2000**

3. **SELECT * FROM movie JOIN movie_cast USING (movie_id)
JOIN person USING (person_id)
WHERE cast_order = 1 AND (release_year = 1990 OR release_year = 2000)**

4. **SELECT title FROM movie UNION ALL
SELECT character_name FROM movie_cast**

5. **SELECT title FROM movie EXCEPT
SELECT character_name FROM movie_cast**