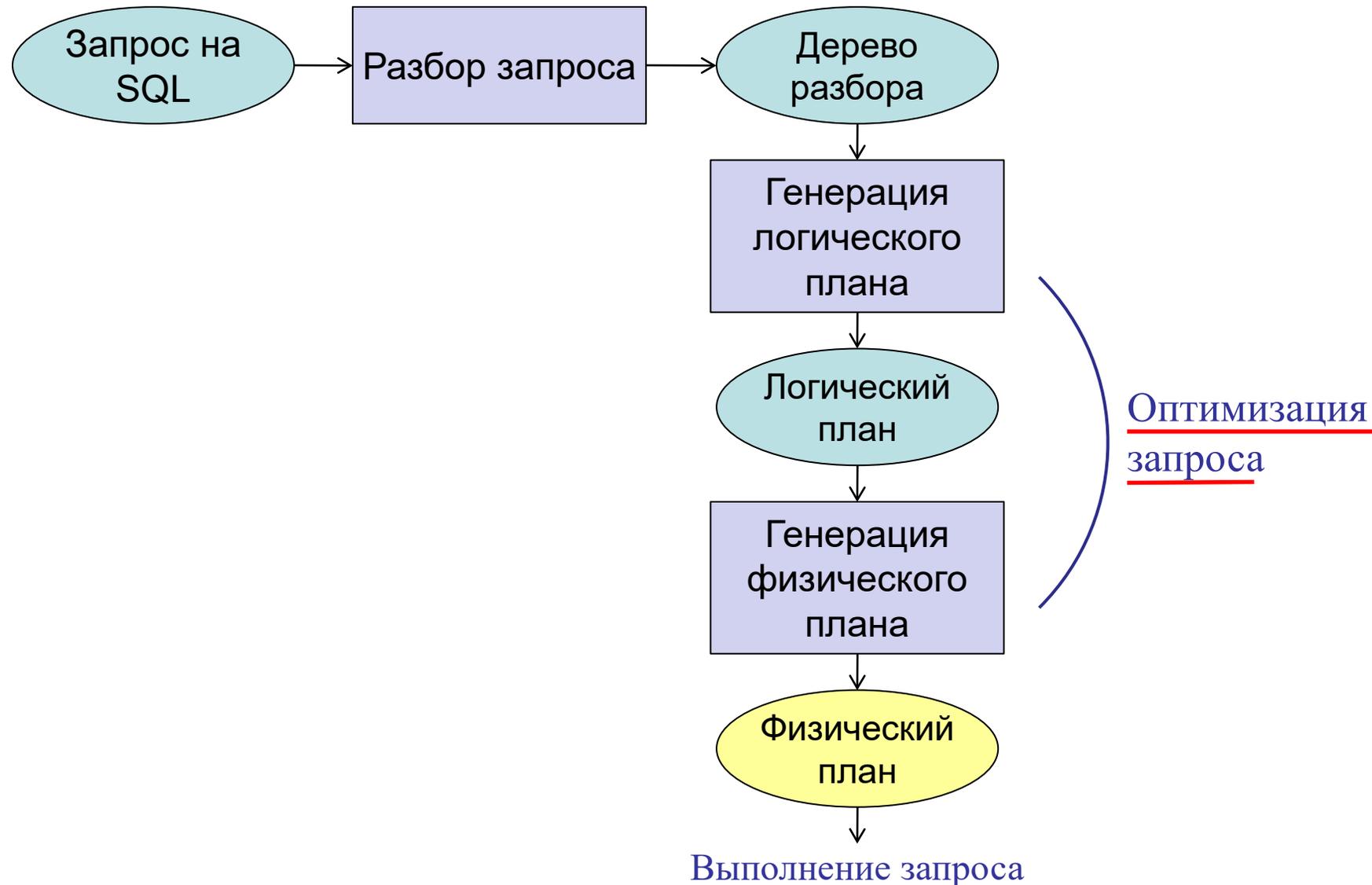


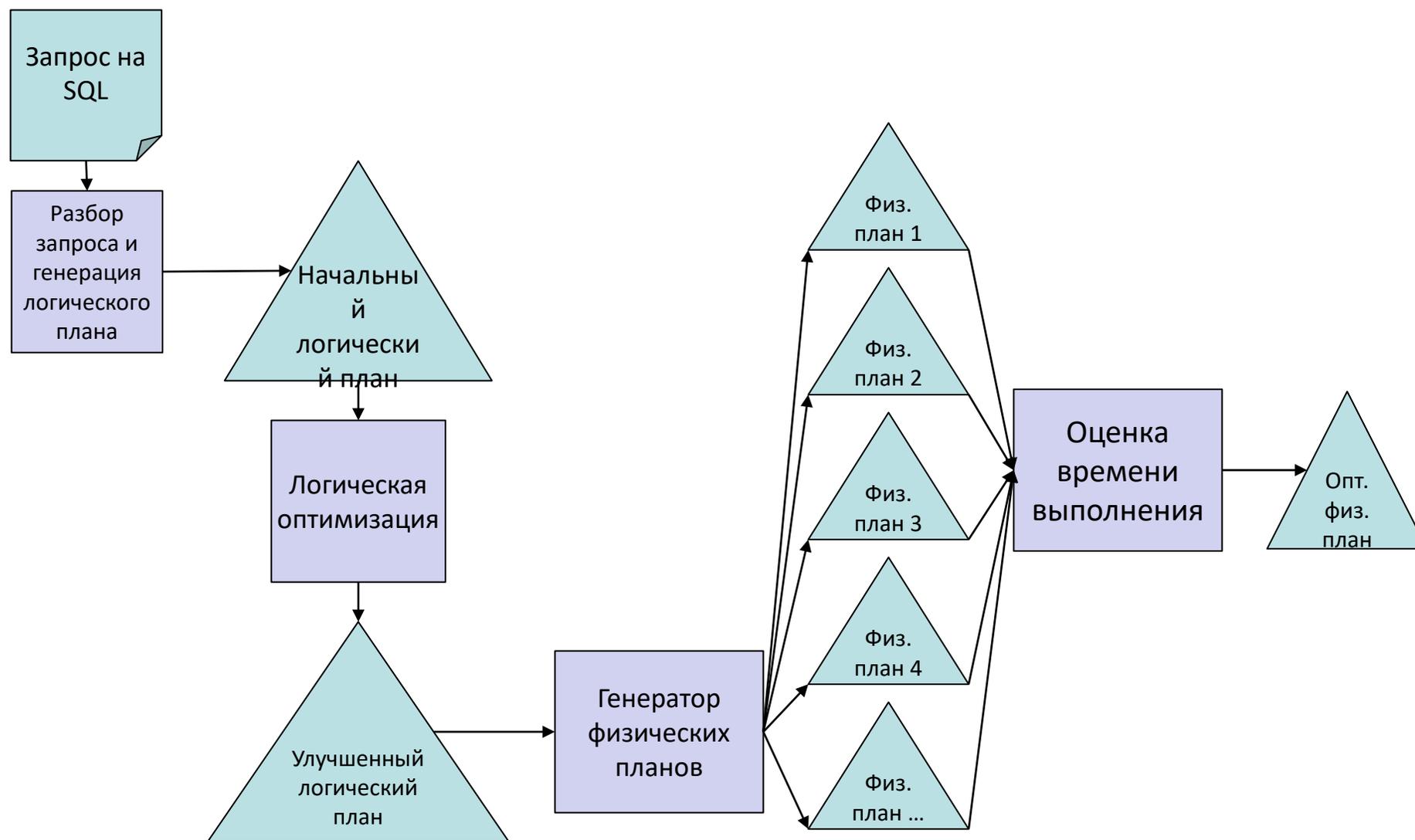
Методы и системы обработки больших данных

6. Оценка стоимости операций

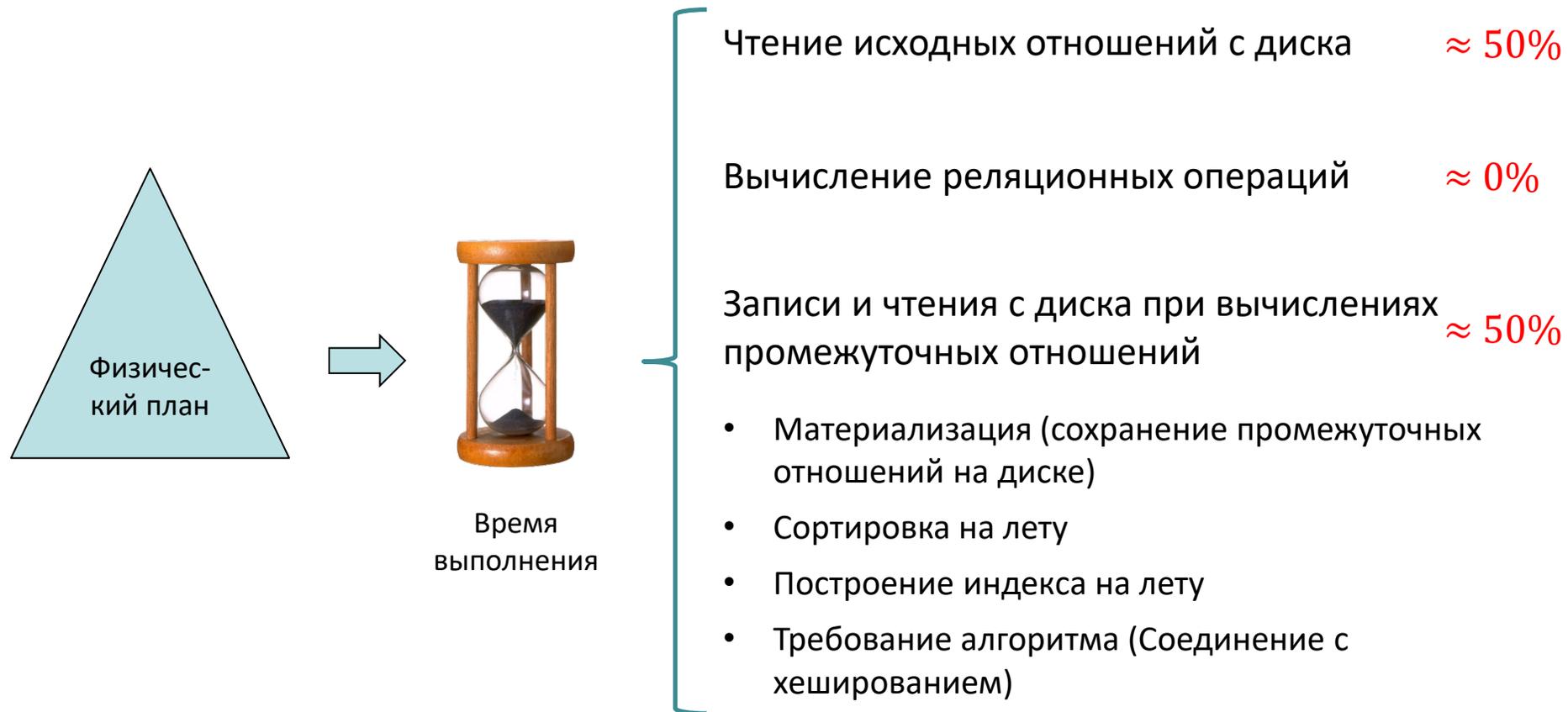
Общая схема обработки запроса



Получение «оптимального» физического плана



Временные затраты на выполнение запроса



Основные идеи

- Время выполнения физического плана пропорционально количеству обменов с диском =>
- План, который делает меньше обменов будет более быстрым
- Зная размеры промежуточных отношений путем анализа алгоритмов можно посчитать количество обменов с диском для всех операций плана =>
- Необходимо для каждой операции оценить размеры промежуточных отношений исходя из размеров исходных отношений
- Для оценки размеров промежуточных отношений достаточно реляционной алгебры

Система обозначений

- $B(R)$ – количество блоков для хранения всех кортежей R
- $T(R)$ – количество кортежей отношения R
- $V(R, A)$ – количество различных значений атрибута A в отношении R

Пример

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

$$T(R) = ?$$

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

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

Пример

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

$$T(R) = 6$$

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

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

Пример

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

$$T(R) = 6$$

$$V(R, B) = 2$$

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

Пример

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

$$T(R) = 6$$

$$V(R, B) = 2$$

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

Оценка проекции

$$S = \pi_{X,Y}(R)$$

- Количество кортежей остается неизменным
- Изменяется размер кортежа и количество блоков

$$T(S) = T(R)$$

$$B(S) \leq B(R)$$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

Длина записи = ?

Записей в блоке = ?

$B(R) = ?$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

Длина записи = $12 + 4 \times 2 + 100 = 120$

Записей в блоке = ?

$B(R) = ?$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = ?$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$$S = \pi_{A,C}(R)$$

$$T(S) = ?$$

Длина записи = ?

Записей в блоке = ?

$$B(S) = ?$$

$$S = \pi_{A,B}(R)$$

$$T(S) = ?$$

Длина записи = ?

Записей в блоке = ?

$$B(S) = ?$$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024 - 24}{120} \right\rfloor = 8$$

$$B(R) = 10000 / 8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$$S = \pi_{A,C}(R)$$

$$T(S) = ?$$

Длина записи = ?

Записей в блоке = ?

$$B(S) = ?$$

$$S = \pi_{A,B}(R)$$

$$T(S) = ?$$

Длина записи = ?

Записей в блоке = ?

$$B(S) = ?$$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = ?

$B(S) = ?$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = ?$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = 10000/8 = 1250$

$S = \pi_{A,B}(R)$

$T(S) = ?$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = 10000/8 = 1250$

$S = \pi_{A,B}(R)$

$T(S) = 10000$

Длина записи = ?

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = 10000/8 = 1250$

$S = \pi_{A,B}(R)$

$T(S) = 10000$

Длина записи = $4 \times 2 + 12 = 20$

Записей в блоке = ?

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = 10000/8 = 1250$

$S = \pi_{A,B}(R)$

$T(S) = 10000$

Длина записи = $4 \times 2 + 12 = 20$

Записей в блоке = $\left\lfloor \frac{1024-24}{20} \right\rfloor = 50$

$B(S) = ?$

Пример вычисления размера результирующего отношения для проекции

$R(A, B, C)$

A, B – целые числа по 4 байта

C – строка 100 байт

Заголовок записи – 12 байт

Размер блока – 1024 байт

Заголовок блока – 24 байт

$T(R) = 10000$

$$\text{Длина записи} = 12 + 4 \times 2 + 100 = 120$$

$$\text{Записей в блоке} = \left\lfloor \frac{1024-24}{120} \right\rfloor = 8$$

$$B(R) = 10000/8 = 1250$$

$B(R)$ – количество блоков для хранения всех кортежей R

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$S = \pi_{A,C}(R)$

$T(S) = 10000$

Длина записи = $12 + 4 + 100 = 116$

Записей в блоке = $\left\lfloor \frac{1024-24}{116} \right\rfloor = 8$

$B(S) = 10000/8 = 1250$

$S = \pi_{A,B}(R)$

$T(S) = 10000$

Длина записи = $4 \times 2 + 12 = 20$

Записей в блоке = $\left\lfloor \frac{1024-24}{20} \right\rfloor = 50$

$B(S) = 10000/50 = 200$

Оценка выборки для «равно»

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = ?$$

$$T(\sigma_{C=9}(R)) = ?$$

$$T(\sigma_{B=x}(R)) = ?$$

$$T(\sigma_{C=y}(R)) = ?$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

Оценка выборки для «равно»

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = 3$$

$$T(\sigma_{C=9}(R)) = ?$$

$$T(\sigma_{B=x}(R)) = ?$$

$$T(\sigma_{C=y}(R)) = ?$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

Оценка выборки для «равно»

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = 3$$

$$T(\sigma_{C=9}(R)) = 2$$

$$T(\sigma_{B=x}(R)) = ?$$

$$T(\sigma_{C=y}(R)) = ?$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

Оценка выборки для «равно»

 R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = 3$$

$$T(\sigma_{C=9}(R)) = 2$$

$$T(\sigma_{B=x}(R)) \approx T(R)/V(R, B)$$

$$T(\sigma_{C=y}(R)) = ?$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

Оценка выборки для «равно»

R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = 3$$

$$T(\sigma_{C=9}(R)) = 2$$

$$T(\sigma_{B=x}(R)) \approx T(R)/V(R, B)$$

$$T(\sigma_{C=y}(R)) \approx T(R)/V(R, C)$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

Оценка выборки для «равно»

R

A^*	B	C
1	3	4
2	7	9
3	3	2
4	7	4
5	7	2
6	3	9

$$T(R) = 6$$

$$V(R, B) = 2$$

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

$$T(\sigma_{B=7}(R)) = 3$$

$$T(\sigma_{C=9}(R)) = 2$$

$$T(\sigma_{B=x}(R)) \approx T(R)/V(R, B)$$

$$T(\sigma_{C=y}(R)) \approx T(R)/V(R, C)$$

$T(R)$ – количество кортежей отношения R

$V(R, B)$ – количество различных значений атрибута B в отношении R

$$T(\sigma_{B=x}(R)) \approx T(R)/V(R, B) = 3$$

$\sigma_{B=7}(R)$

A^*	B	C
2	7	9
4	7	4
5	7	2

$$T(\sigma_{C=y}(R)) \approx T(R)/V(R, C) = 2$$

$\sigma_{C=9}(R)$

A^*	B	C
2	7	9
6	3	9

Оценка размера выборки

1) $\sigma_{A=x}(R)$

2) $\sigma_{A \neq x}(R)$

3) $\sigma_{A < x}(R)$

4) $\sigma_{A > x}(R)$

Оценка выборки $\sigma_{A=x}(R)$

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$$T(\sigma_{A=x}(R)) \approx T(R) / V(R, A)$$

Указанная оценка является точной для распределения

Оценка выборки $\sigma_{A=x}(R)$

$T(R)$ – количество кортежей отношения R

$V(R, A)$ – количество различных значений атрибута A в отношении R

$$T(\sigma_{A=x}(R)) \approx T(R) / V(R, A)$$

Указанная оценка является точной для равномерного распределения

При неравномерном распределении для получения более точной оценки необходимо использовать гистограммы (см. лекцию «7. Статистические характеристики данных»)

Оценка выборки $\sigma_{A \neq x}(R)$

Если A – первичный ключ, то

$$T(\sigma_{A \neq x}(R)) \approx ?$$

В общем случае:

$$T(\sigma_{A \neq x}(R)) \approx ?$$

Оценка выборки $\sigma_{A \neq x}(R)$

Если A – первичный ключ, то

$$T(\sigma_{A \neq x}(R)) \approx T(R) - 1$$

В общем случае:

$$T(\sigma_{A \neq x}(R)) \approx ?$$

Оценка выборки $\sigma_{A \neq x}(R)$

Если A – первичный ключ, то

$$T(\sigma_{A \neq x}(R)) \approx T(R) - 1$$

В общем случае:

$$T(\sigma_{A \neq x}(R)) \approx T(R) - T(\sigma_{A = x}(R)) = T(R) - T(R)/V(R, A)$$

Оценка выборки $\sigma_{A < x}(R)$

Пусть известны величины $T(R)$ и $V(R, A)$.

Тогда $T(\sigma_{A < x}(R)) \approx ?$

Оценка выборки $\sigma_{A < x}(R)$

Пусть известны величины $T(R)$ и $V(R, A)$.

Тогда $T(\sigma_{A < x}(R)) \approx T(R)/3$

Почему делим на 3?

Оценка выборки $\sigma_{A < x} (R)$

Пусть известны величины $T(R)$ и $V(R, A)$.

Тогда $T(\sigma_{A < x} (R)) \approx T(R)/3$

Необходимо назначить начальника низшего звена:

«Выбрать сотрудников в возрасте до 31 года»

$\sigma_{Age < 31} (Staff)$ ← **выборка меньшей части**

Для “>” будет 2/3?

Оценка выборки $\sigma_{A < x} (R)$

Пусть известны величины $T(R)$ и $V(R, A)$.

Тогда $T(\sigma_{A < x} (R)) \approx T(R)/3$

Необходимо назначить начальника низшего звена:

«Выбрать сотрудников в возрасте до 31 года»

$\sigma_{Age < 31} (Staff)$ ← **выборка меньшей части**

Для “>” будет 2/3?

Оценка выборки $\sigma_{A < x} (R)$

Пусть известны величины $T(R)$ и $V(R, A)$.

Тогда $T(\sigma_{A < x} (R)) \approx T(R)/3$

Необходимо сократить штат компании за счет пенсионеров:

«Выбрать сотрудников в возрасте старше 64 лет»

$\sigma_{Age > 64} (Staff)$ ← **выборка меньшей части**

Для “>” будет 1/3

Оценка прямого произведения

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

Естественное соединение по одному атрибуту

$$T(R \underset{(A)}{\bowtie} S) \approx \frac{T(R)T(S)}{\max(V(R, A), V(S, A))}$$

Указанная оценка является точной для соединения по первичному и внешнему ключу:

$$R(A^*, B); S(C, A^\#)$$

В этом случае: $T(R \underset{(A)}{\bowtie} S) = ?$

Естественное соединение по одному атрибуту

$$T(R \underset{(A)}{\bowtie} S) \approx \frac{T(R)T(S)}{\max(V(R, A), V(S, A))}$$

Указанная оценка является точной для соединения по первичному и внешнему ключу:

$$R(A^*, B); S(C, A^\#)$$

В этом случае: $T(R \underset{(A)}{\bowtie} S) = T(S)$

Пример для естественного соединения

$$R(A, B), \quad T(R) = 1000, \quad V(R, B) = 20$$

$$S(B, C), \quad T(S) = 2000, \quad V(S, B) = 50$$

$$T(R \bowtie S) \approx \frac{1000 \cdot 2000}{\max(20, 50)} = \frac{2000000}{50} = 40000$$

Естественное соединение по двум атрибутам

$$T(R \bowtie_{(A,B)} S) \approx \frac{T(R)T(S)}{\max(V(R, A), V(S, A)) * \max(V(R, B), V(S, B))}$$

Оценка для комбинации операций

$$R(A, B); S(B, C)$$

$$\begin{aligned} T(\sigma_{A=3}(R \bowtie S)) &\approx \left(\frac{T(R)T(S)}{\max(V(R, B), V(S, B))} \right) / V(R, A) \\ &= \frac{T(R)T(S)}{V(R, A) \cdot \max(V(R, B), V(S, B))} \end{aligned}$$

Конец лекции 6