

# Дисциплина «Проектирование баз данных»



Маркова Ирина Васильевна,  
начальник управления  
информатизации  
[markova@mit.ru](mailto:markova@mit.ru)

---



# Стратегия эвристической обработки запросов

## Улучшение логического плана запроса

Улучшению качества логических планов способны послужить многие из алгебраических законов, рассмотренных ранее, но наиболее широкое применение в оптимизаторах запросов находят следующие подходы:

- a) Продвижение операторов **выбора** «вниз» по дереву до максимально «глубокого» уровня. Если условие выбора представляет собой конъюнкцию (AND) нескольких частных условий, его можно расщепить, чтобы продвигать каждый оператор отдельно.
- b) При определенных обстоятельствах целесообразнее вначале продвинуть оператор **выбора** «вверх» по дереву выражений, и только затем – «вниз».
- c) Продвижение существующих операторов **проекции** «вниз» по дереву или добавление новых операторов, что, как и в случае с операторами выбора, требует тщательного анализа.
- d) Изъятие операторов **удаления кортежей-дубликатов** или перемещение в требуемые позиции дерева.
- e) Сочетание определенных операторов выбора с расположенными ниже по дереву операторами декартова произведения с целью замены пары операций одной операцией соединения посредством равенства (equijoin).



## Группирование ассоциативно-коммутативных операторов

- a) Традиционными синтаксическими анализаторами не создаются деревья с вершинами, обладающими неограниченно большим количеством дочерних вершин, – обычно операторы пребывают только в унарной или бинарной форме.
- b) Операторы, для которых справедливы ассоциативный и коммутативный законы, способны обладать произвольным количеством операндов.
- c) Группирование соседних вершин дерева, представляющих одноименные ассоциативно-коммутативные операторы, в единую вершину со многими дочерними вершинами (естественное соединение, объединение и пересечение).
- d) Операторы естественного и  $\Theta$ -соединения допускают возможность взаимного сочетания при выполнении следующих условий :
  - операторы естественного соединения заменены  $\Theta$ -соединениями с условиями равенства одноименных атрибутов отношений-аргументов;
  - при переходе от естественного соединения к  $\Theta$ -соединению с помощью оператора проекции удаляются дубликаты атрибутов;
  - условия операторов  $\Theta$ -соединения ассоциативны.
- e) Оператор декартова произведения, интерпретируемый как частный случай естественного соединения, может сочетаться с операторами соединения, если они представлены смежными вершинами дерева выражений.



# Анализ стоимости операций

---

При подсчете стоимости всех возможных физических планов, которые удастся построить на основе логического плана, учитывается следующая информация:

- порядок следования и группирования одноименных, ассоциативно-коммутативных операторов;
- алгоритм реализации каждого оператора логического плана;
- дополнительные операторы, необходимые для реализации физического плана, но отсутствующие в явном виде в логическом плане;
- способ передачи значений атрибутов от одного оператора другому.



# Оценка результатов промежуточных отношений

Введем обозначения:

$k_R$  – кардинальность  $R$ ;

$b_R$  – количество блоков, требуемых для хранения  $R$ ;

$a_R$  – количество различных значений атрибута  $a$  в  $R$ ;

$V_t^R$  – размер кортежа в  $R$ ;

$V_b$  – размер блока;

$k_R^b$  – коэффициент блокирования.

**Цель прогнозирования** размеров промежуточных отношений – не получение точных оценок, а упрощение выбора физического плана по принципу: минимальная стоимость – наилучший план.

Физический план выбирается таким образом, чтобы свести к минимуму примерную стоимость выполнения запроса.



## Оценка результата проекции

Проекция относится к операторам, объем результата выполнения которых вычисляется точно. Изменение объема может быть обусловлено только изменением структуры.

Пусть имеется:

$$R(a, b, c)$$

$a, b$  - целые числа;

$c$  - строка (100 байт);

заголовок строки – 12 байт;

$$V_t^R = 4 + 4 + 100 + 12 = 120;$$

$$V_b = 1024;$$

$$k_R^b = 8;$$

$$k_R = 10000;$$

$$b_R = 1250 .$$



## Оценка результата проекции (пример)

a) Рассмотрим оператор  $S = \pi_{a+b,c}(R)$

$$V_t^S = 4 + 100 + 12 = 116;$$

$$k_R^b = 8;$$

$$k_S = 10000;$$

$$b_S = 1250.$$

Вывод: объём отношения практически не изменился.

b) Рассмотрим оператор  $U = \pi_{a,b}(R)$

$$V_t^U = 8 + 12 = 20;$$

$$k_U = 10000;$$

$$k_U^b = 50(1000 : 20 = 50);$$

$$b_U = 10000 : 50 = 200.$$

Вывод: в результате применения  $\pi$  объём результирующего отношения снизился более, чем в 6 раз.



# Оценка результата выборки

В этом случае размер отдельного кортежа сохраняется, количество кортежей уменьшается.

Оператор	Оценка	Примечание
<p>1. равенство <i>const</i> <math>S = \sigma_{a=c}(R)</math></p>	$k_S = \frac{k_R}{a_R}$	<p>Оценка точна, если значения <i>a</i> равновероятны. Эта формула остаётся наилучшей оценкой «в среднем», даже если распределение не равномерное, но все значения <i>a</i> одинаково часто упоминаются в запросах, в которых адресуется атрибут <i>a</i>. Ещё более точные оценки получаются, если СУБД обладает соответствующей статистикой и гистограммами.</p>
<p>2. условие выбора основано на неравенстве <math>S = \sigma_{a &lt; c}(R)</math></p>	$k_S \approx \frac{k_R}{3}$	<p>Эмпирическая оценка.</p>





# Оценка результата выборки (продолжение)

<p>3. <math>S = \sigma_{a \neq c}(R)</math></p>	<p>а) <math>k_S = k_R</math>          в) <math>k_S = \frac{k_R(a_R - 1)}{a_R}</math></p>	<p>Исходя из эвристики, что <math>\approx \frac{1}{a_R}</math>          часть всех кортежей R не удовлетворяют условию.</p>
<p>4. Условие F – ‘AND’          конъюнкция          (цепочка вложенных операторов, каждый из которых проверяет один конъюнкт)</p>	<p><math>k_S = k_R \times k_{(\sigma_F(R))}</math>, где</p> $k_{(\sigma_F(R))} = \begin{cases} \approx \frac{1}{3} \\ 1 \\ \approx \frac{1}{a_R} \end{cases},$	<p>Коэффициент избирательности <math>k_{(\sigma_F(R))}</math> определяется всеми частными операторами.</p>
<p>5. Условие F – ‘OR’  <math>S = \sigma_{F_1 \text{ OR } F_2}(R)</math></p>	<p><math>k_{\sigma_{F_1 \text{ OR } F_2}(R)} = n(1 - (1 - \frac{m_1}{n})(1 - \frac{m_2}{n}))</math>,          где:  <math>k_R = n</math>  <math>k_{\sigma_{F_1}(R)} = m_1</math>  <math>k_{\sigma_{F_2}(R)} = m_2</math>  <math>F_1</math> и <math>F_2</math> – независимые условия.</p>	<p><math>1 - \frac{m_1}{n}</math> – доля кортежей, не удовлетворяющих <math>F_1</math>,  <math>1 - \frac{m_2}{n}</math> – доля кортежей, не удовлетворяющих <math>F_2</math>.</p>



## Оценка результата выборки (примеры)

а) Пусть есть отношение  $R(a, b, c)$  и оператор  $S = \sigma_{a=10 \text{ and } b < 20}(R)$

$$k_R = 10000;$$

$$a_R = 50.$$

$$\text{Оценка: } k_S = \frac{k_R}{a_R \times 3} = \frac{10000}{50 \times 3} \approx 67.$$

б) Частный случай (внутреннее противоречие): пусть есть отношение  $R(a, b, c)$  и оператор  $S = \sigma_{a=10 \text{ AND } a > 20}(R)$ , тогда формально  $k_S = \frac{k_R}{a_R \times 3} = \frac{10000}{50 \times 3} \approx 67$ , однако

очевидно, что  $k_S = \emptyset$ .

На практике оптимизатор учитывает множество правил, удовлетворяющих различным частным случаям. В данном случае оптимизатор должен был применить правило сведения условия к *false*.



## Оценка результата выборки (продолжение)

с) Пусть есть отношение  $R(a, b)$  и оператор  $S = \sigma_{a=10 \text{ OR } b < 20}(R)$

$$k_R = 10000;$$

$$a_R = 50.$$

$$\text{Оценка: } k_S = \frac{k_R}{a_R} + \frac{k_R}{3} \approx 200 + 3333 \approx 3533.$$

Если принять во внимание, что условия  $a = 10$  и  $b < 20$  – независимы, то

$$k_S = n(1 - (1 - \frac{m_1}{n})(1 - \frac{m_2}{n})) = 10000(1 - (1 - \frac{200}{10000})(1 - \frac{3333}{10000})) \approx 3466$$

В данном случае они мало различаются и не способны повлиять на решение о предпочтении той или иной оценки.



## Оценка результата соединения

Рассмотрим естественное соединение, все остальные варианты соединения могут трактоваться в соответствии со следующими правилами:

- a) размер итогового отношения для соединения на основе равенства после изменения имен атрибутов вычисляется так же, как и в случае естественного соединения;
- b) размер итогового отношения для  $\oplus$ -соединения оценивается как операция декартового соединения с последующей выборкой при выполнении следующих условий:
  - количество кортежей в итоговом отношении равно произведению кортежей-операндов;
  - количество кортежей, удовлетворяющих условию равенства, можно оценить, используя приемы для прогнозирования результатов естественного соединения;
  - условия с неравенствами двух атрибутов (вида  $R.a < S.b$ ) следует трактовать как неравенства вида  $R.a < c$  (эвристика в данном случае: коэффициент избирательности  $\approx \frac{1}{3}$  – для «сложного» условия и  $\approx \frac{1}{2}$  – для простого).



## Оценка результата соединения (продолжение)

Пусть необходимо выполнить естественное соединение  $R(X, Y) \bowtie S(Y, Z)$ .

Возможные варианты связи значений  $R.Y$  и  $S.Y$ :

- $\{R.Y\} \bowtie \{S.Y\} = \emptyset$ , соответственно  $k_{R \bowtie S} = 0$ ;
- $S.Y$  – первичный ключ,  $R.Y$  – внешний ключ, соответственно каждый кортеж  $R$  соединяется с единственным кортежем  $S$ :  $k_{R \bowtie S} = k_R$ ;
- большинство кортежей  $R$  и  $S$  могут иметь одинаковые значения  $Y$ , тогда  $k_{R \bowtie S} \approx k_R \cdot k_S$ .



# Оценка результата соединения (допущения)

## Упрощающие допущения:

1. принадлежность одного множества значений совпадающего атрибутов другому:

если  $Y$  общий атрибут отношений и  $S$ , тогда значения  $Y$  в каждом отношении выбираются в порядке их следования в списке  $(y_1, y_2, y_3, \dots)$ , как следствие, если  $Y_R \leq Y_S$ , то каждое значение атрибута  $R.Y$  будет присутствовать в  $S.Y$ , т.е.  $\{R.Y\} \subseteq \{S.Y\}$ .

2. сохранность множества значений несовпадающих атрибутов:

при соединении отношения  $R$  с другим отношением  $S$  множество значений атрибута  $A$ , не являющегося общим, не сокращается, т.е.  $A \subset r(R)$  и  $A \not\subset r(S)$ , то  $\{A_{(R \bowtie S)}\} = \{A_R\}$  (порядок соединения не важен  $\{A_{(S \bowtie R)}\} = \{A_R\}$ ).



## Оценка результата соединения (один общий атрибут)

Принимая во внимание приведенные допущения, оценим размер  $R(X, Y) \bowtie S(Y, Z)$ .

Пусть  $Y_R \leq Y_S$ , тогда каждый кортеж из  $R$  может быть соединен с определенным кортежем отношения  $S$  с вероятностью  $1/Y_S$ , прогнозируемое число кортежей  $S$ , способных к соединению с каждым кортежем  $R$   $k_S/Y_S$ , т.к. кортежей в  $R$  —  $k_R$ , то

$$k_{R \bowtie S} = \frac{k_R \cdot k_S}{Y_S},$$

Если же  $Y_S < Y_R$ , то в силу симметрии:

$$k_{R \bowtie S} = \frac{k_S \cdot k_R}{Y_R}.$$

Если учитывать оба случая в совокупности, то в качестве знаменателя следует выбрать наибольшее значение из  $Y_R$  и  $Y_S$ :

$$k_{R \bowtie S} = \frac{k_R \cdot k_S}{\max(Y_R, Y_S)}.$$



## Оценка результата соединения с одним общим атрибутом (пример)

Пусть имеются отношения  $R(a,b)$ ,  $S(b,c)$ ,  $U(c,d)$

	$R(a,b)$	$S(b,c)$	$U(c,d)$
Кардинальное число	1 000	2 000	5 000
Число различных значений $b$	20	50	
Число различных значений $c$		100	500





## Оценка результата соединения с одним общим атрибутом (вариант 1)

1. Пусть необходимо вычислить  $R \bowtie S \bowtie U$ .

Сгруппируем  $(R \bowtie S) \bowtie U$  и получим:

$$k_{(R \bowtie S)} = \frac{k_R \cdot k_S}{\max(b_R, b_S)} = \frac{1000 \cdot 2000}{50} = 40000,$$

В соответствии с предположением о сохранности несовпадающих атрибутов  $c_{R \bowtie S} = c_S = 100$ :

$$k_{(R \bowtie S) \bowtie U} = \frac{k_{R \bowtie S} \cdot k_U}{\max(c_{R \bowtie S}, c_U)} = \frac{40000 \cdot 5000}{500} = 400000.$$



## Оценка результата соединения с одним общим атрибутом (вариант 2)

2. Пусть необходимо вычислить  $R \bowtie S \bowtie U$ .  
Сгруппируем теперь  $R \bowtie (S \bowtie U)$  и получим:

$$k_{(S \bowtie U)} = \frac{k_S \cdot k_U}{\max(c_S, c_U)} = \frac{2000 \cdot 5000}{500} = 20000,$$

$$k_{R \bowtie (S \bowtie U)} = \frac{k_R \cdot k_{S \bowtie U}}{\max(b_R, b_{S \bowtie U})} = \frac{1000 \cdot 20000}{50} = 400000.$$

Вывод: результат не зависит от порядка соединения.



## Естественное соединение отношений с несколькими общими атрибутами

Рассмотрим случай, когда  $Y$  – не один атрибут, а несколько атрибутов, подлежащих естественному соединению  $R(X, Y) \bowtie S(Y, Z)$ .

Пусть имеется оператор  $R(x, y_1, y_2) \bowtie S(y_1, y_2, z)$ .

Вероятность совпадения кортежей в  $R$  и  $S$  по  $y_1$ , если  $y_{1_R} \geq y_{1_S}$  равна  $1/y_{1_R}$ . В силу симметрии, если  $y_{1_R} < y_{1_S}$ , то вероятность равна  $1/y_{1_S}$ .

В общем случае, вероятность того, что кортежи  $R$  и  $S$  согласуются в атрибуте  $y_1$  оценивается следующим образом:

$$1 / \max(y_{1_R}, y_{1_S}).$$

Аналогично, для  $y_2$ :

$$1 / \max(y_{2_R}, y_{2_S})$$



## Естественное соединение отношений с несколькими общими атрибутами (продолжение)

Так как значения  $y_1$  и  $y_2$  независимы, вероятность одновременного равенства – это произведение двух указанных выше дробей. Поэтому с учетом общего количества различных пар кортежей  $R$  и  $S$ , прогнозируемое число пар кортежей, совпадающих одновременно в компонентах  $y_1$  и  $y_2$ , равно:

$$k_{R \bowtie S} = \frac{k_R \cdot k_S}{\max(y_{1_R}, y_{1_S}) \cdot \max(y_{2_R}, y_{2_S})}$$

При произвольном количестве общих атрибутов в отношениях-операндах справедливо:

$$k_{R \bowtie S} = \frac{k_R \cdot k_S}{\prod_i \max(y_{i_R}, y_{i_S})},$$

где  $1 < i < n$  – количество общих атрибутов.



## Естественное соединение отношений с несколькими общими атрибутами (пример)

Пусть имеются отношения  $R(a,b,c)$ ,  $S(d,e,f)$ , обладающие следующими статистическими характеристиками:

	$R(a,b,c)$	$S(d,e,f)$
Кардинальное число	1 000	2 000
Число различных значений $b$	20	
Число различных значений $d$		50
Число различных значений $c$	100	
Число различных значений $e$		50



## Естественное соединение отношений с несколькими общими атрибутами (пример)

Пусть необходимо вычислить  $(R(a, b, c) \bowtie_{R.b=S.d \wedge R.c=S.e} S(d, e, f))$ .

$$k_{R \bowtie S} = \frac{k_R \cdot k_S}{\max(y_{1_R}, y_{1_S}) \cdot \max(y_{2_R}, y_{2_S})} = \frac{1000 \cdot 2000}{50 \cdot 100} = \frac{2000}{5} = 400.$$

Выводы, справедливые для естественного соединения, остаются в силе для любых разновидностей соединения по равенству.



## Соединение нескольких отношений

Рассмотрим общий случай естественного соединения:

$$S = R_1 \bowtie R_2 \bowtie \dots \bowtie R_n.$$

Пусть  $A \subset r_i$ ,  $1 \leq i \leq k$  и  $a_{R_i} : v_1 \leq v_2 \leq v_3 \leq \dots \leq v_k$ . Допустим, что в каждом из  $k$  отношений выбрано по одному кортежу  $t_i$ .

Какова вероятность того, что все эти кортежи совпадут в атрибуте  $A$ ?



## Соединение нескольких отношений (продолжение)

Рассмотрим кортеж  $t_1$  (выбран из  $R$  с минимальным количеством  $v_1$  различных значений атрибута). В соответствии с допущением о принадлежности одного множества значений совпадающего атрибута другому каждое из  $v_1$  значений можно найти в компонентах  $A$  кортежей всех других  $k - 1$  отношений  $(\{v_1\} \subset \{v_2\} \subset \{v_3\} \subset \dots \subset \{v_k\})$ . Кортеж  $t_i$  совпадает с  $t_1$  в атрибуте  $A$  с вероятностью  $p_i = \frac{1}{v_1}$ . Это утверждение верно для всех  $i = 2, 3, \dots, k$ .

Вероятность того, что все кортежи совпадут в атрибуте  $A$  :

$$P = \frac{1}{v_2 \cdot v_3 \cdot \dots \cdot v_k}.$$

Размер итогового отношения, возвращаемого оператором соединения с произвольным числом аргументов:

$$k_S = \frac{k_1 \cdot k_2 \cdot \dots \cdot k_R}{v_2 \cdot v_3 \cdot \dots \cdot v_k},$$

для каждого атрибута  $A$ , присутствующего, как минимум, в двух отношениях.





## Соединение нескольких отношений (пример)

Пусть имеются отношения  $R(a,b,c)$ ,  $S(b,c,d)$  и  $U(b,e)$ , обладающие следующими статистическими характеристиками:

	$R(a,b,c)$	$S(b,c,d)$	$U(b,e)$
Кардинальное число	1 000	2 000	5 000
Число различных значений $a$	100		
Число различных значений $b$	20	50	200
Число различных значений $c$	200	100	
Число различных значений $d$		400	
Число различных значений $e$			500



## Пример (продолжение)

Пусть необходимо вычислить  $R(a,b,c) \bowtie S(b,c,d) \bowtie U(b,e)$ .

Оценка размера итогового отношения будет иметь вид:

$$k_{R \bowtie S \bowtie U}^b = \frac{k_R \cdot k_S \cdot k_U}{v_S \cdot v_U} \text{ и } k_{R \bowtie S \bowtie U}^c = \frac{k_R \cdot k_S \cdot k_U}{v_R},$$

Таким образом,

$$k_{R \bowtie S \bowtie U} = \frac{k_R \cdot k_S \cdot k_U}{v_S \cdot v_U \cdot v_R} = \frac{1000 \cdot 2000 \cdot 5000}{(50 \cdot 200) \cdot 200} = 5000.$$

Независимо от группирования и упорядочения аргументов выражения естественного соединения  $n$  отношений, правило, применяемое к каждому частному соединению отдельно, дает тот же результат, что и в случае применения к соединению всех  $n$  отношений.