МЕТОДЫ И АЛГОРИТМЫ ПОДДЕРЖКИ ЦЕЛОСТНОСТИ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ В ПРИЛОЖЕНИЯХ КЛАССОВ OLAP И OLTP

05.13.17 – Теоретические основы информатики

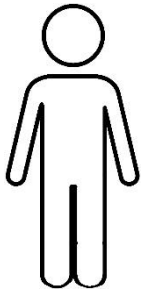
Диссертация на соискание ученой степени кандидата физико-математических наук

Владимир Сергеевич Зыкин

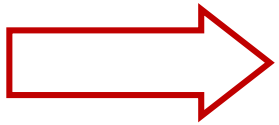
Научный руководитель:
ЦЫМБЛЕР Михаил Леонидович,
кандидат физ.-мат. наук, доцент

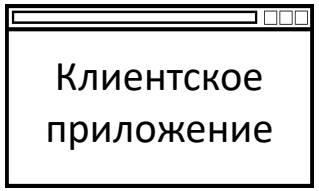
Актуальность

Внешний уровень



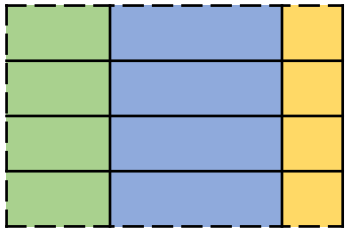
Пользователь



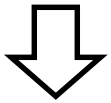


Запрет обновления записи, соответствующей нескольким кортежам

Представление

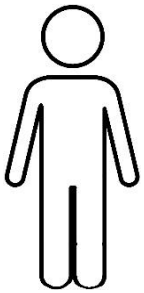


СУБД



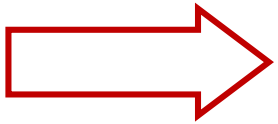
Концептуальный уровень

Некорректное соотношение кортежей с NULL значениями



Проектировщик

Избыточные зависимости включения



Экспоненциальная сложность поиска зависимостей

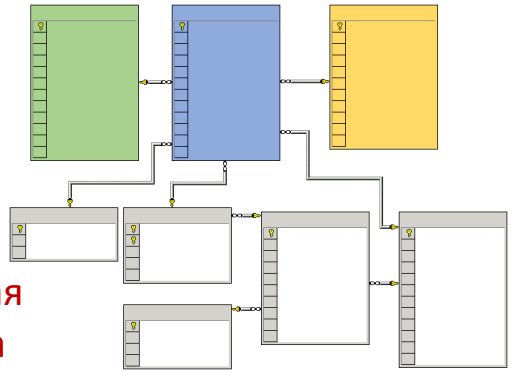


Схема БД

Цель диссертационной работы

Исследование и разработка эффективных методов и алгоритмов поддержки целостности данных на концептуальном и внешнем уровнях архитектуры реляционных баз данных для приложений классов OLAP и OLTP

Основные задачи

1. Разработать систему аксиом зависимостей включения, которая обеспечивает ссылочную целостность при наличии неопределенных значений
2. Разработать алгоритм построения неизбыточного множества зависимостей включения
3. Разработать общий подход к обновлению многотабличных представлений, обеспечивающий корректную модификацию записи в представлении, которой соответствуют несколько кортежей в хранимых отношениях баз данных
4. Реализовать предложенные методы и подходы в виде сопроцессора СУБД для приложений классов OLAP и OLTP
5. Провести вычислительные эксперименты, подтверждающие эффективность предложенных подходов

Работы по теме диссертации

Casanova M., Fagin R., Papadimitriou C. *Inclusion Dependencies and Their Interaction with Functional Dependencies*. Journal of Computer and System Sciences. **1984**. Vol 28. № 1. P. 29–59.	Аксиоматизация зависимостей включения
Levene M., Vincent M.W. *Justification for Inclusion Dependency Normal Form*. IEEE Trans. on Knowl. and Data Eng. **2000.** Vol 12, № 2. P. 281–291.	Задача совместного поиска функциональных зависимостей и зависимостей включения
Lechtenborger J. *The Impact of the Constant Complement Approach Towards View Updating*. ACM SIGMOD. **2003**. P. 49–55.	Неавтоматизированное обновление представлений
Bertossi L., Salimi B. *Causes for Query Answers from Databases: Datalog Abduction, View-updates, and Integrity Constraints*. Int. J. Approx. Reason. **2017**. Vol. 90. P. 226–252.	Использование обновляемых представлений для подготовки данных к машинному обучению
Köhler H., Link S. *Inclusion dependencies and their interaction with functional dependencies in SQL*. Journal of Computer and System Sciences. **2017**. Vol. 85. P. 104–131.	Совместная аксиоматизация функциональных зависимостей и зависимостей включения для NOT NULL ограничений
Masunaga Y., Nagata Y., Ishii T. *Making Join Views Updatable on Relational Database Systems in Theory and in Practice.* IMCOM **2019**. Vol. 935. P. 823–840.	Обновление представления с использованием триггеров PostgreSQL

Зависимости включения

R, S – отношения

X, Y – множества атрибутов

t_R, t_S – кортежи отношений R и S

Определение. *Зависимость включения* $R[X] \subseteq S[Y]$

$$\forall\, t_R \in R \quad \exists\, t_S \in S: \quad t_R[X] = t_S[Y]$$

Некорректное соотношение кортежей с NULL-значениями

Типизированные зависимости включения

Пусть R, S – отношения; X – множество атрибутов, $A \in X$

Опр. 1. Кортеж $t_R[X]$ *соответствует* кортежу $t_S[X]$ *по атрибутам* X $(t_R[X] \preccurlyeq t_S[X])$, если выполнено одно из условий:

1) $t_S[A] \neq NULL \Rightarrow t_R[A] = t_S[A] \vee t_R[A] = NULL$;
2) $t_S[A] = NULL \Rightarrow t_R[A] = NULL$.

Опр. 2. *Типизированная зависимость включения* $R[X] \subseteq S[X]$ имеет место, если

$$\forall\, t_R[X] \quad \exists t_S[X]: \quad t_R[X] \preccurlyeq t_S[X]$$

Корректное соотношение кортежей с NULL-значениями

Типизированные зависимости включения

Журнал

ID	ID датчика	ID устройства	Время	Значение
10	4	NULL	9:02	67
11	3	3	9:18	98
13	4	3	9:32	88

Датчики

ID датчика	ID устройства	Назначение
3	3	Температура
4	NULL	Влажность

Журнал[ID датчика, ID устройства] ⊆ Датчики[ID датчика, ID устройства]

Нетипизированные зависимости включения

Типизированные зависимости включения

Журнал

ID	ID датчика	ID устройства	Время	Значение
8	NULL	3	8:46	156
10	4	NULL	9:02	67
11	3	3	9:18	98
13	4	3	9:32	88

Датчики

ID датчика	ID устройства	Назначение
1	3	Давление
2	3	Расход
3	3	Температура
4	NULL	Влажность

$t_{\text{Журнал}}$[ID датчика, ID устройства] ≼ $t_{\text{Датчики}}$[ID датчика, ID устройства]

Журнал[ID датчика, ID устройства] ⊊ Датчики[ID датчика, ID устройства]

Аксиомы зависимостей включения

R, S – отношения X, Y – множества атрибутов

- **Рефлексивность:**
  $R\,[X] \subseteq R\,[X]$, X – атрибуты в R.
- **Транзитивность:**
  $R[X] \subseteq S[X] \quad \wedge \quad S[Y] \subseteq T[X] \quad \Rightarrow \quad R[X] \subseteq T[X].$
- **Проекция и перестановки:**
  $(R[X] \subseteq S[X]) \wedge (Y \subseteq X) \Rightarrow R\,[Y] \subseteq S\,[Y]$ $\forall$ перестановок в X и Y.

Экспоненциальная сложность поиска зависимостей

Аксиомы типизированных зависимостей включения

- **Рефлексия**: $R[X] \subsetneq R[X]$
- **Транзитивность**: $R[X] \subsetneq S[X] \wedge S[X] \subsetneq T[X] \Rightarrow R[X] \subsetneq T[X]$
- **Проекция**: $R[Y] \subsetneq S[Y];\ X \subseteq Y \Rightarrow R[X] \subsetneq S[X]$

Дополнительные правила вывода:

$$R[X] \cap S[X] \subsetneq S[X] \cap T[X]$$
$$R[X] \cap S[X] \subsetneq R[X] \cap T[X]$$
$$R[X] \cap T[X] \subsetneq S[X] \cap T[X]$$

Полиномиальная сложность поиска зависимостей

Непротиворечивость и полнота

$\sum$ – множество зависимостей БД, σ – произвольная зависимость.

Теорема 1. Система аксиом ***непротиворечива***.
Другими словами: если σ выводима из $\sum$, тогда зависимость σ является логическим следствием $\sum$.

$$\sum \vdash \sigma \Rightarrow \sum \models \sigma$$

Теорема 2. Система аксиом ***полна***.
Другими словами: если зависимость σ является логическим следствием $\sum$, тогда σ выводима из $\sum$

$$\sum \models \sigma \Rightarrow \sum \vdash \sigma$$

Опр. 1. Зависимость σ является ***логическим следствием*** множества зависимостей $\sum$ $(\sum \models \sigma)$, если при существующем наборе $\sum$ зависимость σ всегда выполнима.

Опр. 2. Зависимость σ ***выводима*** из $\sum$ по системе аксиом $(\sum \vdash \sigma)$, если при применении правил вывода из $\sum$ и аксиом, за любое конечное число шагов появляется зависимость σ.

Теорема об избыточной типизированной зависимости включения

Типизированная зависимость включения $R_i[X] \subsetneq R_j[X]$ на схеме БД избыточна тогда и только тогда, когда $\exists p > 0$ и выполняются следующие условия:

1) $R_i[X_0] \subsetneq R_{i_1}[X_1] \subsetneq R_{i_2}[X_2] \subsetneq \cdots \subsetneq R_{i_p}[X_p] \subsetneq R_j[X]$

2) $X \subseteq X_s$,

где i_s – номера отношений

X_s – множества атрибутов, по которым устанавливаются типизированные зависимости включения

$s = 0, 1, 2, \ldots, p$

Алгоритм построения неизбыточного множества зависимостей включения

ШАГ 1. FORM_IND(OUT D)

$D = \emptyset$
for $i = 1$ **to** k
  for $\forall\, FD_l(R_i) \in R_i$
    for $j = 1$ **to** k
      if $i \neq j$ **then**
        if $FD_l(R_i) \in R_j$ **then**
          $L = L \cup Ch(R_i, R_j, V)$

$O(k^2)$

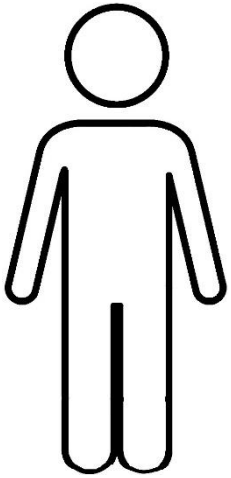
ШАГ 2. MIN-IND (IN OUT D)

for $\forall\, D(l, j, X) \in D$
  $q = 1; p(q) = l; stop =$ **FALSE**;
  while not $stop$
    for $\forall D(v, w, X_q) \in D: \{D(l, j, X) \neq D(v, w, X_q)\}$
      del_rel **= FALSE**
      if $(v \in p[1, \dots, q]) \wedge (FD_l(R_l) \subseteq R_w)$ **then**
        if $w = j$ **then**
          $L = L \setminus L(l, j, X)$
          break
        else
          if $w \notin p[1, \dots, q]$ **then**
            $q = q + 1; p(q) = w;$
            $del\_rel =$ **TRUE**
      $stop =$ **not** del_rel

$O(|D|^2 k^2)$

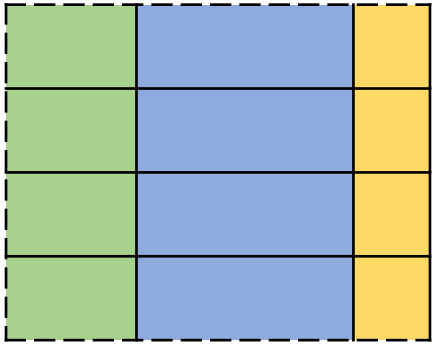
k – количество отношений
$|D|$ – количество зависимостей

Целостность многотабличных представлений



SQL оператор обновления представления

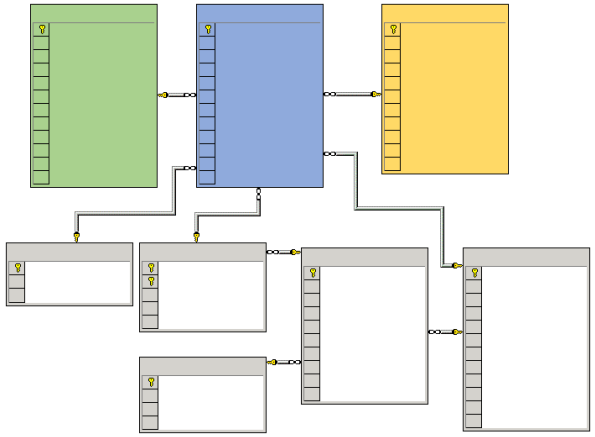
Представление



СУБД

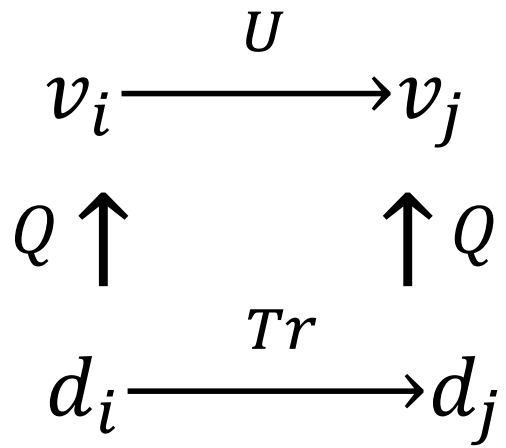
Транзакция, обновления базовых таблиц

Схема БД



Коммутативные преобразования базы данных

Транзакция Tr является ***коммутативным преобразованием*** для операции U, если и только если выполняется $U\big(Q(d)\big) = Q\big(Tr(d)\big)$



где

- d_i – исходное состояние БД;
- v_i – исходное состояние представления Q;
- v_j – состояние представления после операции обновления U;
- d_j – состояние БД после выполнения транзакции Tr;
- U – операция обновления представления, переводящая представление из состояния v_i в состояние v_j;
- Tr – транзакция, переводящая базу данных из состояния d_i в состояние d_j.

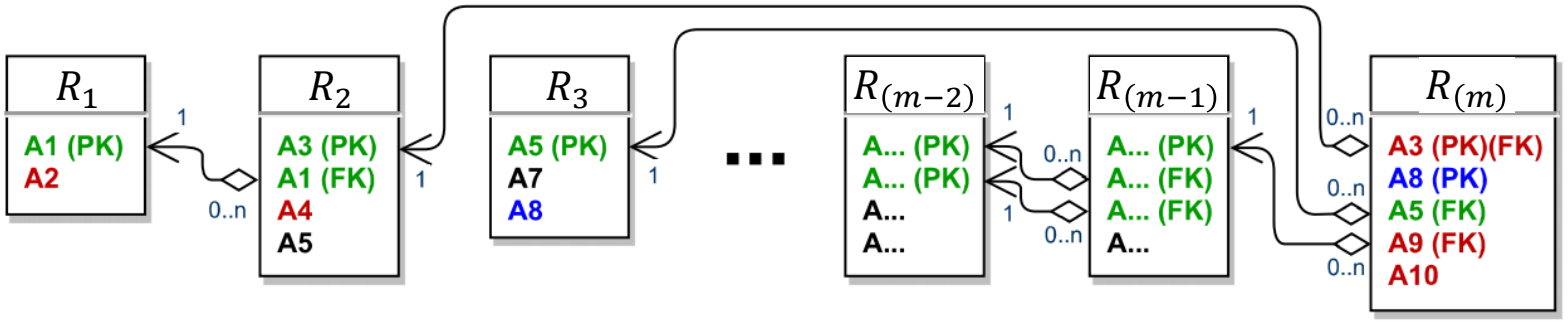
Целевое отношение. Неизбыточное множество атрибутов

В схеме базы данных $,\ldots,R_m$ отношение R_m назовем **целевым**, если выполняются следующие условия:

1) $R_1,\ldots,R_m$ упорядочены по внешним ключам
2) отсутствуют отношения, ссылающиеся на R_m

Неизбыточное множество атрибутов X_i отношения R_i включает в себя каждый атрибут, для которого выполнено одно из условий:

1) атрибут входит в заголовок представления Q
2) атрибут входит во внешний ключ отношения R_i
3) атрибут задан с помощью предиката, ограничивающего значение



Атрибут заголовка Атрибут внешний ключ Атрибут с ограничением

Теорема о коммутативности операции удаления записи

Пусть

$R_1, \dots, R_m$ – схема базы данных, R_m – целевое отношение,

X_i – неизбыточное множество атрибутов,

Q – представление, t – удаляемая запись.

Тогда операция **DELETE** *коммутативна*.

$$\mathbf{DELETE}\ (\boldsymbol{Q}, \boldsymbol{t}) = \boldsymbol{R_m} \setminus \boldsymbol{\pi_{\langle R_m \rangle}(T_D)}$$

$$T_D = \pi_{X_0}\left(\sigma_{c_{del}}\left(\bowtie_{i=1}^{m} \pi_{X_i}(R_i)\right)\right)$$

$$c_{del} = F \wedge (X_0 \equiv \langle t \rangle)$$

Теорема о коммутативности операции вставки записи

Пусть

$R_1, \ldots, R_m$ – схема базы данных, R_m – целевое отношение,

X_i – неизбыточное множество атрибутов,

Q – представление, t – вставляемая запись.

Тогда операция **INSERT** *коммутативна*.

$$\mathbf{INSERT}(\boldsymbol{Q}, \boldsymbol{t}) = \boldsymbol{R_m} \cup \boldsymbol{T_I}$$

$$T_I = \pi_{X_m}\left(\sigma_F(T_I' \cup t)\right)$$

$$T_I' = \pi_Y\left(\sigma_{c_{ins}}\left(\bowtie_{i=1}^{m-1}\pi_{X_i}(R_i)\right)\right)$$

$$Z = X_0 \cap \cup_{i=1}^{m-1} X_i$$

$$Y = (\langle R_m \rangle \cup \langle F \rangle \cup X_0) \cap \cup_{i=1}^{m-1} X_i$$

$$c_{ins} = F' \wedge (Z \equiv \langle \pi_Z(t) \rangle)$$

Теорема о коммутативности операции обновления

Пусть

$R_1, \ldots, R_m$ – схема базы данных, R_m – целевое отношение,

X_i – неизбыточное множество атрибутов,

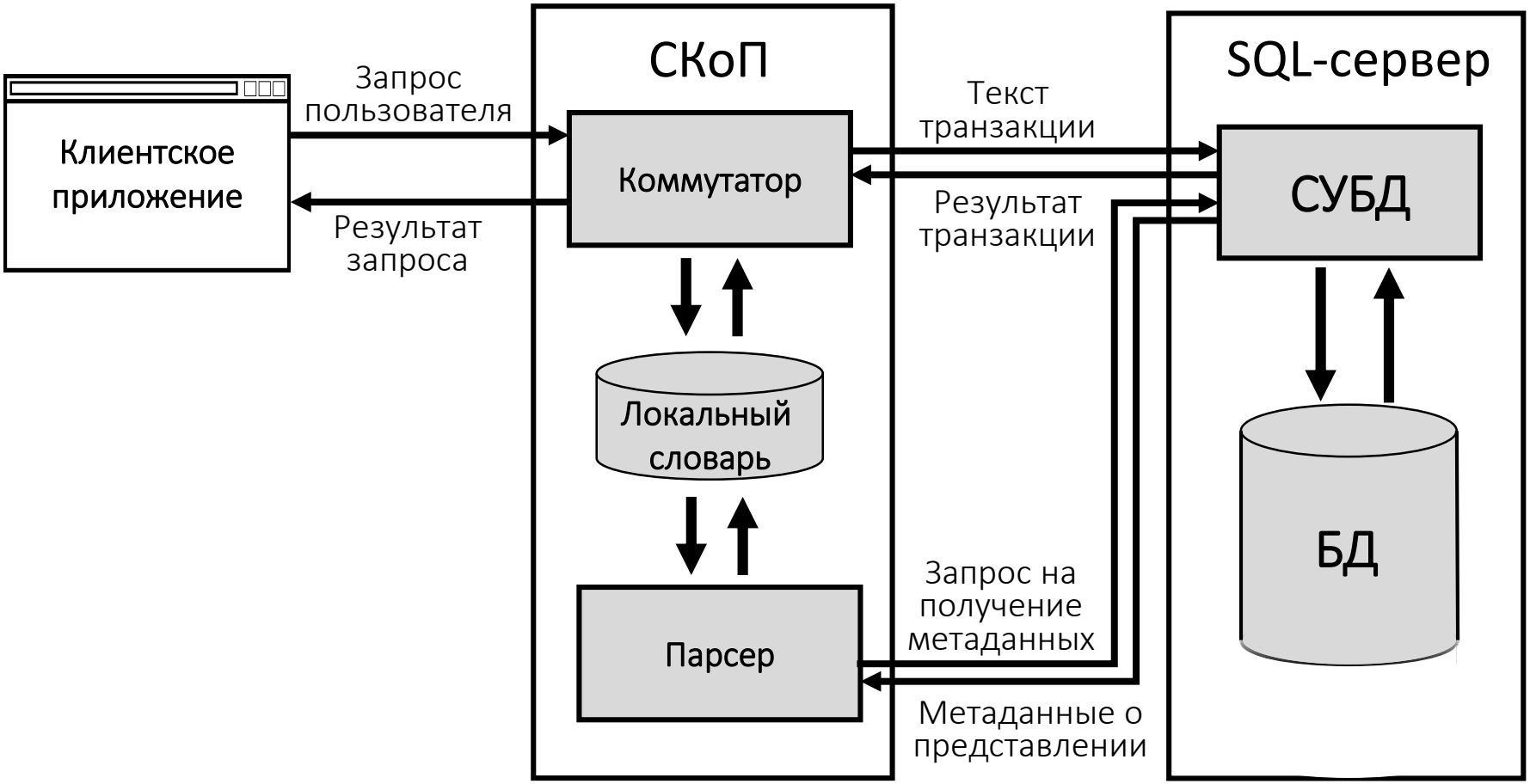
Q – представление,

t – старая запись, t' – новая запись.

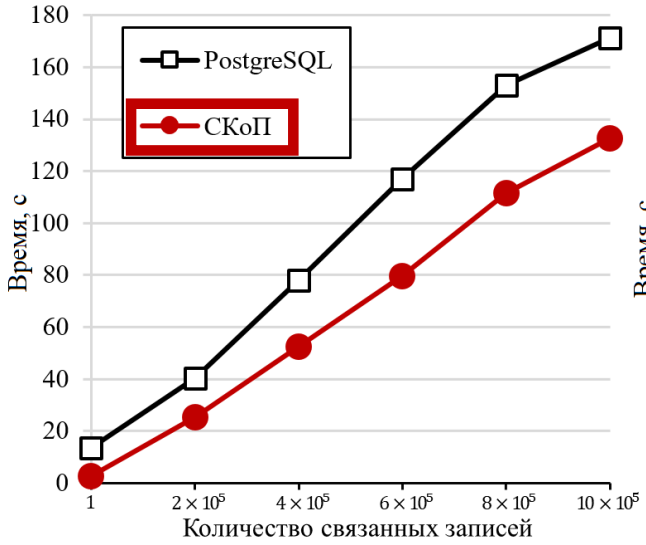
Тогда операция **UPDATE** *коммутативна*.

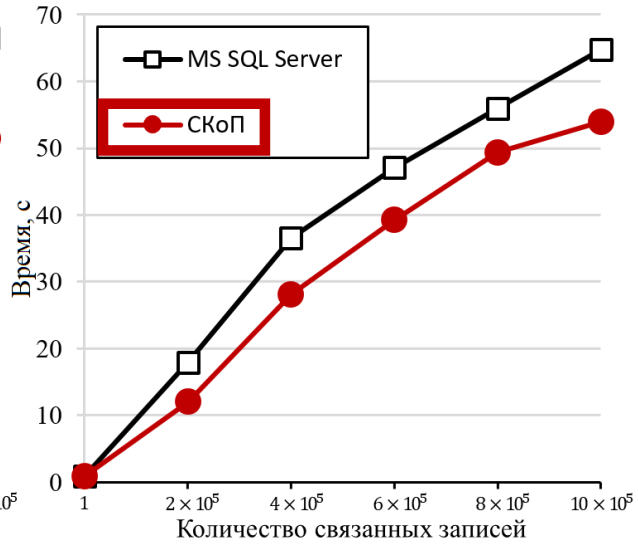
$$\mathbf{UPDATE}\ (Q, t, t') = \mathbf{DELETE}\ (Q, t);\ \mathbf{INSERT}(Q, t')$$

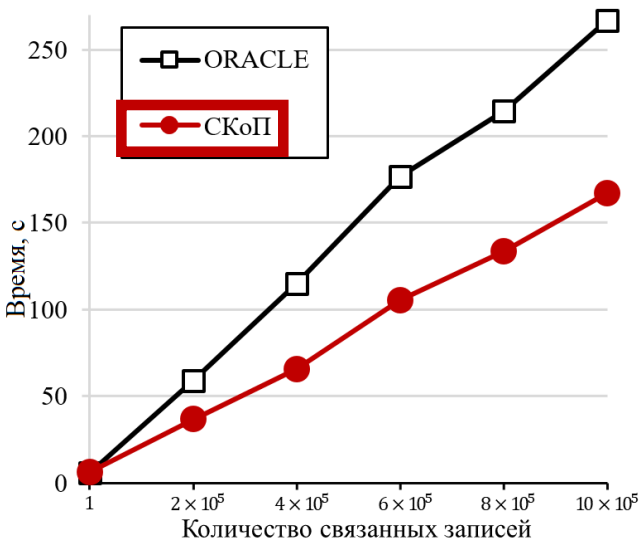
Сопроцессор коммутативных преобразований (СКоП)



Эксперименты (приложения класса OLAP)





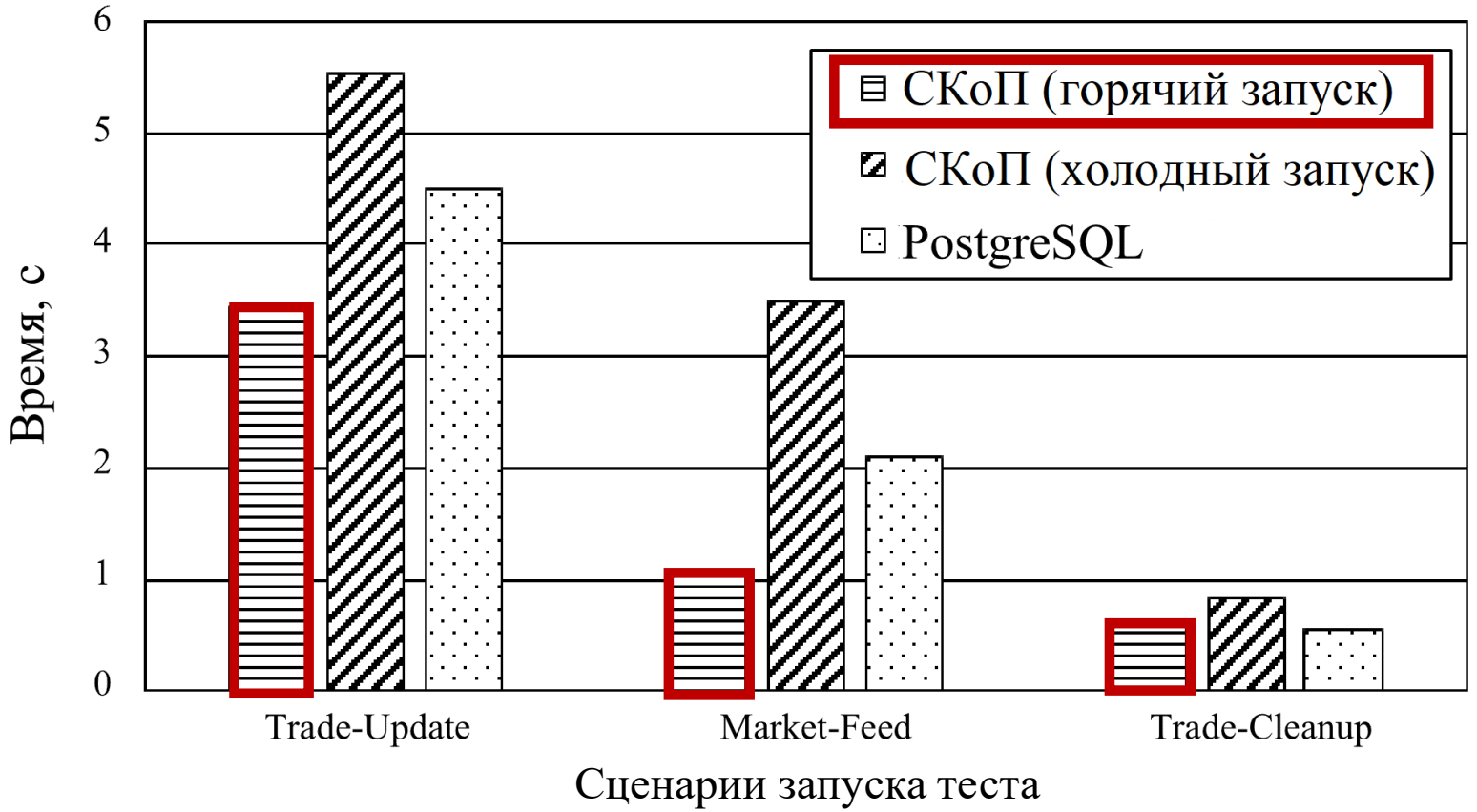


Ускорение СКоП по сравнению с аналогами

PostgreSQL: на **35%** **MS SQL Server**: на **17%** **ORACLE**: на **30%**

База данных теста **TPC-H**:
Количество таблиц: **8**
Количество записей: **8.7 млн.**

Эксперименты (приложения класса OLTP)



База данных теста **TPC-E**:

Количество таблиц: **33**

Количество записей: **585.2 млн.**

Основные результаты, выносимые на защиту

1. Предложена система аксиом типизированных зависимостей включения с неопределенными значениями в реляционных БД и доказана ее полнота и непротиворечивость
2. Разработан алгоритм построения неизбыточного множества типизированных зависимостей включения в реляционных БД, доказана его корректность и получена оценка вычислительной сложности
3. Сформулированы и доказаны теоремы о коммутативных преобразованиях для обновления многотабличных представлений в реляционных БД. Разработана архитектура сопроцессора коммутативных преобразований СУБД и реализован сопроцессор СУБД PostgreSQL
4. Проведены вычислительные эксперименты, подтверждающие эффективность предложенных подходов