На правах рукописи

ФИЛИМОНОВ Алексей Валерьевич

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ПРОЦЕССА ПРИБЛИЖЕННОЙ ОБРАБОТКИ ЗАПРОСОВ В РЕЛЯЦИОННЫХ СИСТЕМАХ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ

Специальность 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Орёл – 2023

Работа выполнена в федеральном государственном казенном военном образовательном учреждении высшего образования «Академия Федеральной службы охраны Российской Федерации».

Научный руководитель: **Козлов Сергей Викторович,**
кандидат технических наук

Официальные оппоненты: **Алексеев Владимир Витальевич,**
доктор технических наук, профессор,
заведующий кафедрой «Информационные системы и защита информации»,
ФГБОУ ВО «Тамбовский государственный технический университет»

Саенко Игорь Борисович,
доктор технических наук, профессор,
ведущий научный сотрудник лаборатории проблем компьютерной безопасности СПИИРАН,
ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Институт системного программирования им. В. П. Иванникова Российской академии наук»
(ИСП РАН)

Защита состоится «14» апреля 2023 г. в 14:00 на заседании диссертационного совета 24.2.286.04, созданного на базе ФГБОУ ВО «Воронежский государственный технический университет», по адресу: г. Воронеж, Московский пр-т, 14, ауд. 216.

С диссертацией можно ознакомиться в научно-технической библиотеке и на сайте ФГБОУ ВО «Воронежский государственный технический университет» https://cchgeu.ru.

Автореферат разослан «20» февраля 2023 г.

Ученый секретарь
диссертационного совета Гусев Константин Юрьевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Использование современных методов машинного обучения, многофакторной обработки больших объемов данных, извлечения знаний, ориентированных на обработку сложных произвольных (ad hoc) запросов для решения задач аналитической обработки с низкой задержкой, предъявляет особые требования к архитектуре современных систем управления базами данных. На практике часто возникает ситуация, когда невозможно заранее определить подмножество отношений, участвующих в запросе. Это приводит к необходимости обработки не только отдельных запросов, но и конвейеров запросов, что существенно затрудняет решение задач оптимизации структуры хранилища. Направлением развития систем управления базами данных для решения подобных проблем является архитектура гибридных транзакционно-аналитических систем обработки данных (HTAP – hybrid transactional/analytical processing).

Одна из базовых функцией таких систем – функция приближенной обработки запросов (AQP – approximate query processing). Ее использование основано на гипотезе отсутствия необходимости получения точного результата обработки запросов для большинства приложений анализа данных. Подобный подход получения результатов обработки запросов с некоторой некритичной для пользователя погрешностью позволяет снизить временные задержки и сделать процесс обработки менее ресурсоемким. Этот подход применим, например, для запросов, результаты которых отображаются в приложениях визуализации данных в виде графиков и диаграмм. Использование методов AQP обеспечивает не только интерактивное время отклика при анализе массивных наборов данных, но и находит свое применение в обработке высокоскоростных потоков данных.

Существенный вклад в развитие исследуемой предметной области в части математического и программного обеспечения для подобных систем внесли Кузнецов С. Д., Аббасов Э. М., Григорьев Ю. А., Ухаров А. О., Chaudhuri S., Brown P. G., Haas P. J., Xu B., Tirthapura S., Busch C.

Поскольку приближенная обработка запросов широко применяется на практике, существует направление исследований, ориентированное на ее изучение в таких аспектах, как обеспечение точности результатов, получаемых в ходе обработки запроса. Одним из направлений в этой предметной области, наряду с применением методов сжимающих преобразований, является использование случайной выборки. Методы на ее основе применительно к обработке одиночных запросов демонстрируют значительное сокращение используемых ресурсов и времени отклика, обеспечивая при этом небольшую ошибку аппроксимации. Однако вопросы, связанные с использованием случайной выборки при обработке конвейера запросов, являются недостаточно исследованными.

Настоящее исследование посвящено вопросам, касающимся оптимизации пространства случайной выборки применительно

к конвейерным структурам запросов с целью получения результатов их выполнения с приемлемой точностью и за ограниченное время.

Таким образом, актуальность темы диссертационного исследования продиктована необходимостью совершенствования математического и специального программного обеспечения процесса приближенной обработки запросов в реляционных системах управления базами данных.

Тематика диссертации соответствует одному из основных научных направлений ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации» – Повышение эффективности функционирования распределенных вычислительных систем.

Целью работы является разработка средств математического и специального программного обеспечения приближенной обработки запросов в системах управления базами данных реляционного типа, обеспечивающих эффективную работу с конвейером запросов в хранилищах данных.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ актуальных исследований, посвященных проблеме повышения точности обработки конвейера запросов в реляционных системах управления базами данных на основе методов приближенной обработки.
2. Разработать математическое обеспечение для приближенной обработки конвейеров запросов в системах управления базами данных реляционного типа, обеспечивающее оптимизацию пространства случайной выборки по критерию оперативности и ресурсоемкости процесса обработки.
3. Разработать алгоритм приближенной обработки запросов с применением разработанного математического обеспечения в процессоре запросов реляционных систем управления базами данных.
4. Разработать архитектуру программного обеспечения процессора запросов системы управления базами данных, поддерживающую приближенную обработку конвейера запросов в хранилище данных с применением разработанного алгоритма.
5. Разработать элементы специального программного обеспечения процессора запросов системы управления базами данных, обеспечивающие приближенную обработку конвейера запросов, провести эксперименты по оцениванию точности результатов и задержки обработки запросов.

Объект исследования – процессор запросов в системах управления базами данных реляционного типа.

Предмет исследования – процесс приближенной обработки конвейера запросов в процессоре запросов систем управления базами данных реляционного типа.

Методы исследования. В ходе решения задач, поставленных в диссертации, использовались методы теории баз данных, теории системного анализа, теории множеств, теории вероятностей, математической статистики и планирования экспериментов, методы нелинейной оптимизации.

Тематика работы соответствует следующим пунктам паспорта специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей (технические науки): п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем»; п. 4 «Интеллектуальные системы машинного обучения, управления базами данных и знаний, инструментальные средства разработки цифровых продуктов».

Научная новизна работы. В диссертации получены следующие результаты, характеризующиеся научной новизной:

− математическое обеспечение приближенной обработки конвейера запросов в реляционных системах управления базами данных, отличающееся применением коэффициента вариации как метрики качества выборки, обеспечивающее получение равноточных оценок случайных величин из различных кластеров;

− алгоритм приближенной обработки конвейера запросов, отличающийся от известного наличием двухпроходной процедуры расчета требуемых статистических характеристик в цепочках запросов, обеспечивающий получение результатов выполнения конвейера запросов с оптимальным распределением пространства выборки;

− архитектура процессора запросов системы управления базами данных реляционного типа, отличающаяся наличием механизмов кэширования планов и результатов выполнения промежуточных запросов из конвейера, тем самым обеспечивающая снижение ресурсоемкости повторной обработки частных запросов в других конвейерах;

− структура программного обеспечения приближенной обработки конвейера запросов, отличающаяся реализацией взаимодействия с модулем сбора статистики системы управления базами данных и механизма кэширования и обеспечивающая повышение точности результатов выполненных запросов в заданных временных ограничениях.

Положения, выносимые на защиту:

− математическое обеспечение приближенной обработки запросов в реляционных системах управления базами данных обеспечивает повышение точности результатов, получаемых при приближенной обработке запросов;

− алгоритм приближенной обработки конвейера запросов процессором реляционной системы управления базами данных реализует разработанное в исследовании математическое обеспечение для оптимизации пространства выборки и обеспечивает получение результатов выполнения конвейера запросов с оптимальным распределением пространства выборки;

− архитектура процессора запросов системы управления базами данных реляционного типа обеспечивает снижение ресурсоемкости повторной обработки частных запросов в других конвейерах;

– структура программного обеспечения приближенной обработки конвейера запросов обеспечивает повышение точности результатов выполненных запросов в заданных временных ограничениях.

Теоретическая значимость исследования состоит в разработке математического обеспечения процесса приближенной обработки запросов в реляционных системах управления базами данных, обеспечивающего оптимизацию пространства выборки конвейера запросов.

Практическая значимость исследования заключается в разработке средств математического и программного обеспечения приближенной обработки запросов в системах управления базами данных реляционного типа на основе теоретико-вероятностного подхода, обеспечивающих снижение задержки обработки запросов при заданном уровне точности результата обработки. На элементы разработанных программных средств получено свидетельство о государственной регистрации программы для электронной вычислительной машины в реестре Федеральной службы по интеллектуальной собственности.

Результаты внедрения. Основные положения и результаты диссертационной работы реализованы в виде специального программного модуля для прототипа процессора запросов реляционной системы управления базами данных ИСП РАН в рамках выполнения исследований по совершенствованию программного обеспечения, реализующего гибридную транзакционно-аналитическую обработку данных. Теоретические результаты работы используются в учебном процессе ФГКВОУ ВО «Академия ФСО России» в рамках дисциплины «Системы баз данных».

Апробация результатов диссертационного исследования. Основные положения диссертационной работы докладывались и обсуждались на следующих конференциях: Modern informatization problems in simulation and social technologies MIP-2020’ES. Proceedings of the XXV-th International Open Science Conference (Yelm, WA, USA, 2020); Всероссийская межведомственная научно-техническая конференция по теоретическим и прикладным проблемам развития и совершенствования автоматизированных систем управления и связи «НАУКА и АСУ – 2021» (Москва, Зеленоград, 2021); Modern informatization problems in simulation and social technologies MIP-2022’AS. Proceedings of the XXVII-th International Open Science Conference (Yelm, WA, USA, 2022).

Публикации. По результатам диссертационного исследования опубликованы 9 печатных работ, в том числе 3 статьи в изданиях, рекомендованных ВАК РФ, свидетельство о государственной регистрации программы, а также статья в издании, индексируемом в Scopus. В работах, опубликованных в соавторстве и приведенных в конце автореферата, лично автором получены следующие результаты: [2] – математическое обеспечение приближенной обработки запросов; [3] – алгоритм приближенной обработки запросов и конвейера запросов в реляционной системе управления базами данных; [4] – архитектура процессора запросов с механизмом приближенной

обработки в реляционной системе управления базами данных; [5, 6] – реализация специального программного обеспечения приближенной обработки конвейера запросов к хранилищам данных.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и приложений. Работа изложена на 96 страницах машинописного текста, содержит 35 рисунков, 3 таблицы и список литературы, включающий 104 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, сформулированы цель и задачи исследования, представлены основные научные результаты, а также приведены сведения об их апробации и внедрении.

В первой главе проведен анализ состояния проблемы повышения точности результатов, получаемых с применением механизмов приближенной обработки запросов.

Был исследован подход, представленный в исследованиях Григорьева Ю. А. и Ухарова А. О., решающий задачу приближенной обработки запросов применением алгоритмов сжатия с потерями над данными, извлеченными из OLTP-системы, и загрузкой сжатого представления в OLAP-хранилище. Наряду с широкими преимуществами, данный подход содержит некоторые недостатки. Во-первых, необходимая точность получаемых данных определяется на этапе применения сжимающего преобразования. Во-вторых, выполнение аналитических запросов над данными с учетом вновь поступившей в хранилище данных информации требует выполнения процедуры «извлечение–сжатие–загрузка», что приводит к повышению задержки актуализации данных в аналитическом хранилище.

Принципиально иным подходом является реализация механизмов приближенной обработки запросов непосредственно в OLTP-системах. В этом случае в основе лежит применение случайной выборки. Наиболее простым часто используемым методом является формирование равномерной выборки. Однако такой метод имеет определенный недостаток, так как количество выбранных элементов из группы определяется мощностью множества этой группы, то группа небольшого размера может быть недостаточно представительной. Очевидно, это приводит к большим значениям ошибки для некоторых групп элементов, что подтверждается данными исследований.

Таким образом, рассмотрены подходы к реализации приближенной обработки запросов в системах управления базами данных. На основе проведенного анализа сформулирована цель диссертационного исследования: делается формальная постановка задачи исследования и определяются частные задачи достижения цели.

Вторая глава посвящена разработке математического обеспечения приближенной обработки запросов, в частности, конвейеров запросов,

в реляционных системах управления базами данных, созданного с целью повышения точности результатов их выполнения.

В основе разработанного математического обеспечения лежит метод равномерной случайной выборки, который не лишен недостатков, в частности он не обеспечивает достаточной точности результатов на реальных наборах данных. Для устранения указанного недостатка в математическом обеспечении предполагается, во-первых, переход к стратифицированной случайной выборке, а во-вторых, выбор новый метрики качества выборки на основе коэффициента вариации.

В исследовании предложены способы агрегирования коэффициента вариации на основе нормы ℓ_2, который в дальнейшем может быть распространен для нормы ℓ_∞ (максимума) коэффициента вариации. Дополнительно в процесс приближенной обработки запросов включен механизм приоритетов, за счет которого повышается точность оценок для заданных пользователем групп. В исследовании этот механизм реализован с помощью присвоения весов для каждой из групп. Веса определены в положительном подмножестве действительных чисел для каждого из элементов множества результатов $w=\{w_i\}$ и нормированы так, что $\sum w_i = 1$. Таким образом, большее значение w_i выражает более высокие требования к точности для значения y_i. С учетом этого выражение для определения величины ошибки принимает следующий вид:

$$\varepsilon(y,w) = \ell_2(\mathrm{C},w) = \sqrt{\sum_{i=1}^{r} w_i \cdot \left(\mathrm{CV}\left[y_i\right]\right)^2}\ . \qquad (1)$$

Многопараметрический характер вариантов запросов с группированием не позволяет сформировать их обобщенную модель. В силу этого целесообразно рассматривать классы запросов с различным значением числа функций агрегирования и операторов группирования. Базовым вариантом запроса с группированием при этом является запрос с одной функцией агрегирования и одиночным группированием. Обобщенное представление такого запроса на языке SQL представлено на рисунке 1.

```
SELECT a1,a2,...,an-1,AGG(an)
FROM T
GROUP BY (a1,a2,...,an-1);
```

Рисунок 1 – Обобщенное представление запроса с единственной функцией агрегирования и одиночным группированием

В рассматриваемом запросе страты будут однозначно соответствовать получаемым в результате группам. При этом важным допущением метода стратифицированной выборки является то, что вероятности выбора записи

из таблицы не обязательно равны в отдельных стратах, но равны для записей внутри одной страты.

Далее представлено выражение количественного соотношения между объемами выборки в различных стратах с учетом указанного характера данных.

Пусть даны множество переменных $s_1, s_2, \ldots, s_k$, определяющее объем выборки данных для каждой страты, и константы M, $\alpha_1, \alpha_2, \ldots, \alpha_k$, определяющие соответственно общий суммарный объем выборки и значения $\alpha_i = \frac{w_i \sigma_i^2}{\mu_i^2}$, определяющие статистический характер данных в каждой страте на основе предложенной метрики качества выборки.

Определены функции $f(s) = \sum_{i=1}^{k} \frac{\alpha_i}{s_i}$ – ошибки оценки значения функции агрегирования и $g(s) = \sum_{i=1}^{k} s_i - M$ – величина объема выборки для всего запроса в целом.

Постановка оптимизационной задачи при этом выражается как минимизация функции $f(s)$ в ограничениях функции $g(s)$:

$$\begin{gathered} f(s) = \sum_{i=1}^{k} \frac{\alpha_i}{s_i} \to \min \\ \sum_{i=1}^{k} s_i \le M \end{gathered} \tag{2}$$

Для решения этой задачи предложено применить метод множителей Лагранжа, что позволило получить следующее выражение:

$$s_i = M \cdot \frac{\sqrt{\alpha_i}}{\sum_{j=1}^{k} \sqrt{\alpha_j}}. \tag{3}$$

Тогда для запросов с единственной функцией агрегирования и одиночным оператором группирования с учетом вектора весов w и общего объема выборки M можно найти оптимальное распределение объемов выборки s для каждой группы $i = 1 \ldots r$ по выражению

$$s_i = M \frac{\sqrt{w_i}\sigma_i / \mu_i}{\sum_{j=1}^{r} \frac{\sqrt{w_j}\sigma_j}{\mu_j}}. \tag{4}$$

Аналогичное решение представляется для запросов с множественной агрегацией и одиночным оператором группирования. Обобщенное представление такого типа запросов представлено на рисунке 2.

SELECT $a_1,\ldots,a_{n-1},\mathrm{AGG}_1(a_n),\ldots,\mathrm{AGG}_t(a_{n+t})$

FROM T

GROUP BY $\left(a_1,\ldots,a_{n-1}\right)$;

Рисунок 2 – Обобщенное представление запроса с несколькими функциями агрегирования и одиночным группированием

Пусть запрос имеет t агрегируемых значений и r групп. Тогда агрегируемые значения для каждой группы возможно представить в виде матрицы размером $t \times r$. Введены следующие обозначения: $\mu_{i,j}$, $\sigma_{i,j}$ – соответственно среднее значение и стандартное отклонение в *i*-й группе для *j*-го значения; n_i – размер *i*-й группы; s_i – объем выборки, полученной из *i*-й группы; $y_{i,j}$ – оценка случайной величины $\mu_{i,j}$, полученная по выборке.

Тогда решение задачи оптимального распределения размеров выборки s для каждой из r групп при заданных весах w и общем размере выборки M выражается следующим равенством:

$$s_i = M \frac{\sqrt{\alpha_i}}{\sum_{j=1}^{r} \sqrt{\alpha_j}}. \quad (5)$$

где $\alpha_i = \sum_{j=1}^{t} \frac{w_{i,j}\sigma_{i,j}^2}{\mu_{i,j}^2}$ – коэффициент выборки *i*-го кластера для *j*-го значения агрегирования.

Разработанное математическое обеспечение приближенной обработки запросов позволяет произвести оптимизацию распределения пространства выборки между стратами, что приведет к повышению точности результатов при выполнении запросов на наборах данных страты, в которых обладают статистической неоднородностью. На основе разработанного математического обеспечения предлагаются разработка алгоритмов процессора запросов в реляционной системе управления базами данных и проведение экспериментальных исследований.

В третьей главе представлены разработанные в ходе исследования алгоритмы процессора запросов реляционной системы управления базами данных для реализации механизма приближенной обработки запросов и конвейера запросов. Данные алгоритмы используют разработанное

математическое обеспечение, решающее задачу оптимизации распределения выборки в запросах.

Разработанное математическое обеспечение может быть применено при синтезе алгоритма приближенной обработки запросов, функционирующего в пространстве процессора системы управления базами данных. На рисунке 3 представлена схема алгоритма для частного случая запросов с единственной функцией агрегирования.

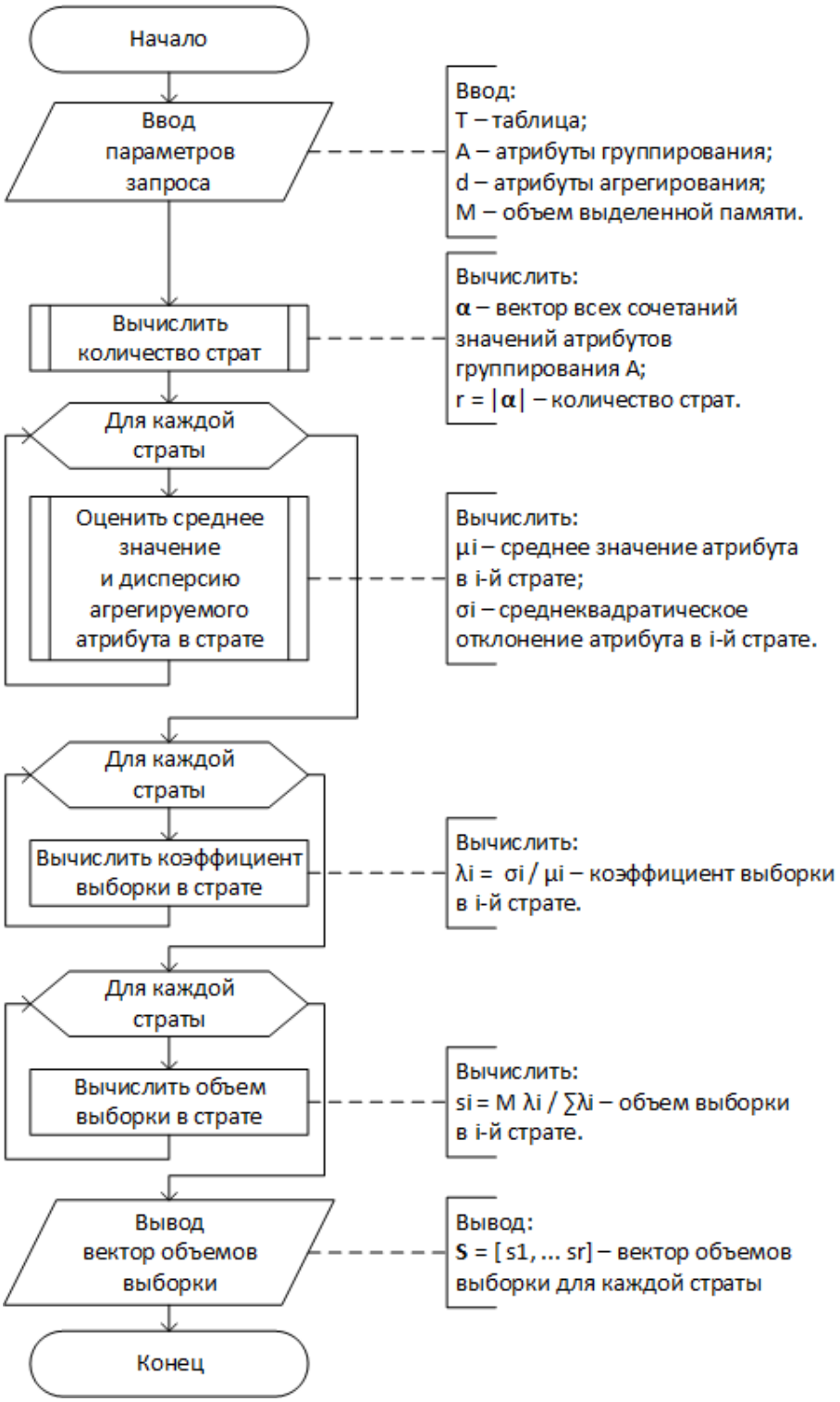

Рисунок 3 – Структурная схема алгоритма приближенной обработки запроса

Данный алгоритм включает в себя предварительную процедуру определения страт, для каждой из которых в последующем итерационно вычисляется коэффициент выборки на основе оценок статистических характеристик кластера и определяется объем выборки исходя из заданного объема выборки.

В современных хранилищах данных существует тенденция не только к увеличению объема данных, но и к повышению сложности задач аналитической обработки этих данных. Для решения таких задач существует устоявшийся подход, который основан на переходе от одного комплексного запроса к последовательности более простых локальных запросов. Такая последовательность может быть описана моделью конвейера запросов (query pipeline). На рисунке 4 представлен пример такой модели для конвейера, состоящего из пяти запросов к четырем таблицам.

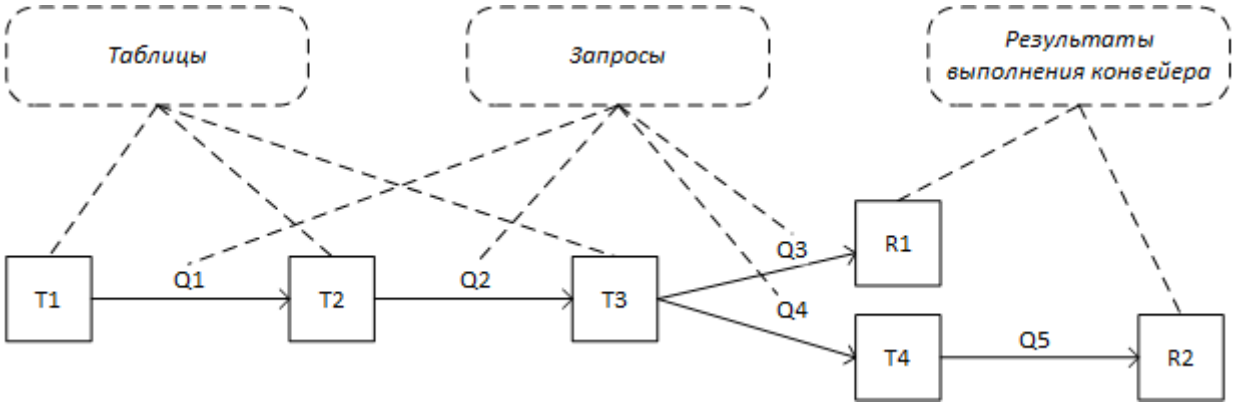

Рисунок 4 – Пример конвейера запросов в форме ориентированного ациклического графа

Конвейер запросов – это упорядоченный по выполнению набор запросов и таблиц. Модель конвейера представляется направленным ациклическим графом, в котором таблицы представлены узлами, а запросы – ребрами. Результаты каждого частного запроса материализуются в виде таблицы, которая может обслуживать другие запросы в потоке. Таким образом, сложный аналитический запрос разбивается на множество подзадач с промежуточными результатами. Такая модель предоставляет следующие преимущества: модульность, оптимизация обработки полного потока запроса, масштабируемость. На рисунке 5 представлен алгоритм, позволяющий применить приближенную обработку конвейера запросов с учетом изложенных выше особенностей.

Приближенная обработка конвейера запросов предполагает получение результата с оцененным доверительным интервалом. В таком случае процесс необходимо адаптировать так, чтобы необходимые данные для вычисления доверительного интервала проходили через весь конвейер. Для этой цели представлена двухпроходная схема алгоритма. На первом этапе алгоритма выполняется обратный проход по конвейеру запросов (от приемника

к источнику). На каждой итерации этой фазы алгоритм вычисляет информацию, необходимую для получения всех последующих оценок и их доверительных интервалов. На втором этапе выполняется прямой проход, в ходе которого полученные оценки вносятся непосредственно в запросы конвейера и в схемы таблиц с промежуточными результатами.

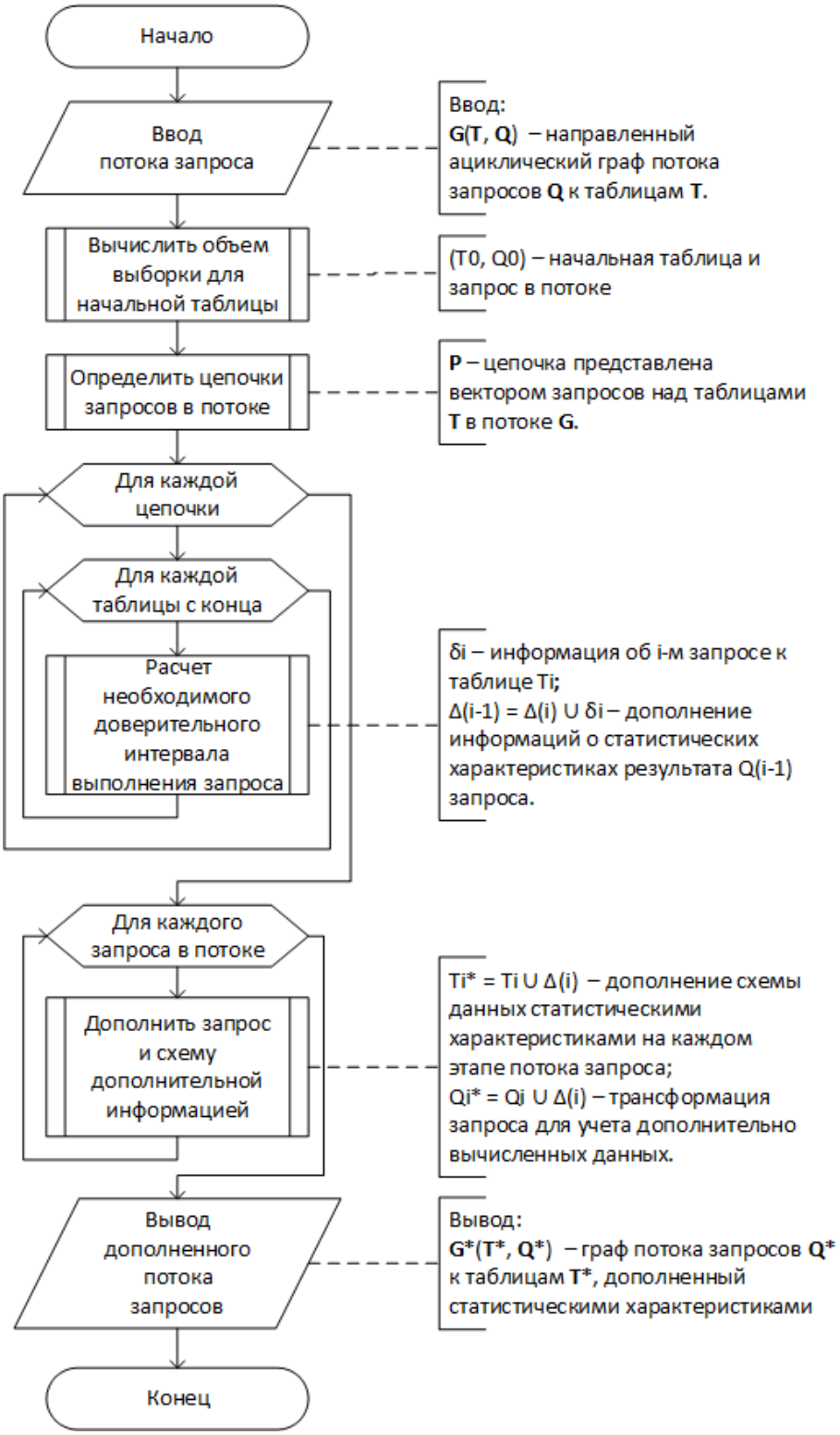

Рисунок 5 – Структурная схема алгоритма приближенной обработки конвейера запросов

Результатом этого алгоритма является направленный ациклический граф конвейера запросов, топологически эквивалентный исходному, узлы и ребра которого дополнены информацией для выполнения приближенной обработки каждого подзапроса. Это позволит обеспечить результат выполнения требуемыми статистическими гарантиями за счет применения алгоритма приближенной обработки для каждого частного запроса в конвейере.

Разработанные алгоритмы применены при синтезе архитектуры и структуры специального программного обеспечения процессора запросов системы управления базами данных, что, таким образом обеспечивает функцию приближенной обработки запросов в хранилищах данных, построенных на базе реляционных систем управления базами данных. Такой подход находится в плоскости существующей задачи разработки современных систем гибридной транзакционно-аналитической обработки запросов.

В четвертой главе представлены предлагаемая архитектура процессора запросов системы управления базами данных реляционного типа, схема взаимодействия элементов программного обеспечения и результаты сравнительных экспериментов, которые подтверждают пригодность полученных решений.

Для реализации разработанных алгоритмов и экспериментальной проверки их адекватности была выбрана система управления базами данных PostgreSQL, поскольку она имеет открытый исходный код и необходимую для эксперимента расширяемость.

Для реализации механизмов приближенной обработки запросов в существующие процедуры были внесены следующие модификации:

− в процедурах лексического и синтаксического анализа внесены функции разбора лексических конструкций, которые описывают заданные параметры приближенного запроса в теле SQL-запроса. Данные функции обслуживаются библиотеками синтаксического анализа Flex и Bison;

− в процедуре первичной оптимизации внедрены функции, определяющие кластеры результатов запроса на основе атрибутов группирования;

− алгоритмы оценки кардинальности и селективности расширены с учетом необходимости вычислений средних значений и стандартных ошибок для каждого кластера. С этой целью система управления базами данных PostgreSQL расширена механизмами, которые кэшируют данными по используемым предикатам группирования и значения оценок для атрибутов агрегирования;

− в ходе оптимизации плана выполнения запроса используется модель стоимостной оптимизации системы управления базами данных PostgreSQL. После этого план выполнения запросов дополняется информацией об объеме выборки для каждого кластера с учетом заданного значения времени обработки запроса. При этом значение времени обработки за вычетом уже затраченного на этом этапе времени отображается в предельную стоимость

исполнения запроса, что позволит применить приближенную обработку на уровне методов доступа к данным;

– процедура исполнения запросов расширена механизмом целенаправленной выборки записей, которые принадлежат необходимому кластеру, что обеспечивает равномерную и достаточную выборку записей внутри кластера. Дополненные механизмом приближенной обработки данные размещаются в объекте, хранящем состояние выполняемого запроса.

Представленная архитектура позволяет в полной мере реализовать механизм приближенной обработки запросов в процессоре реляционной системы управления базами данных и получить результат с определенным доверительным интервалом и уровнем достоверности, не превышая заданный пользователем бюджет времени.

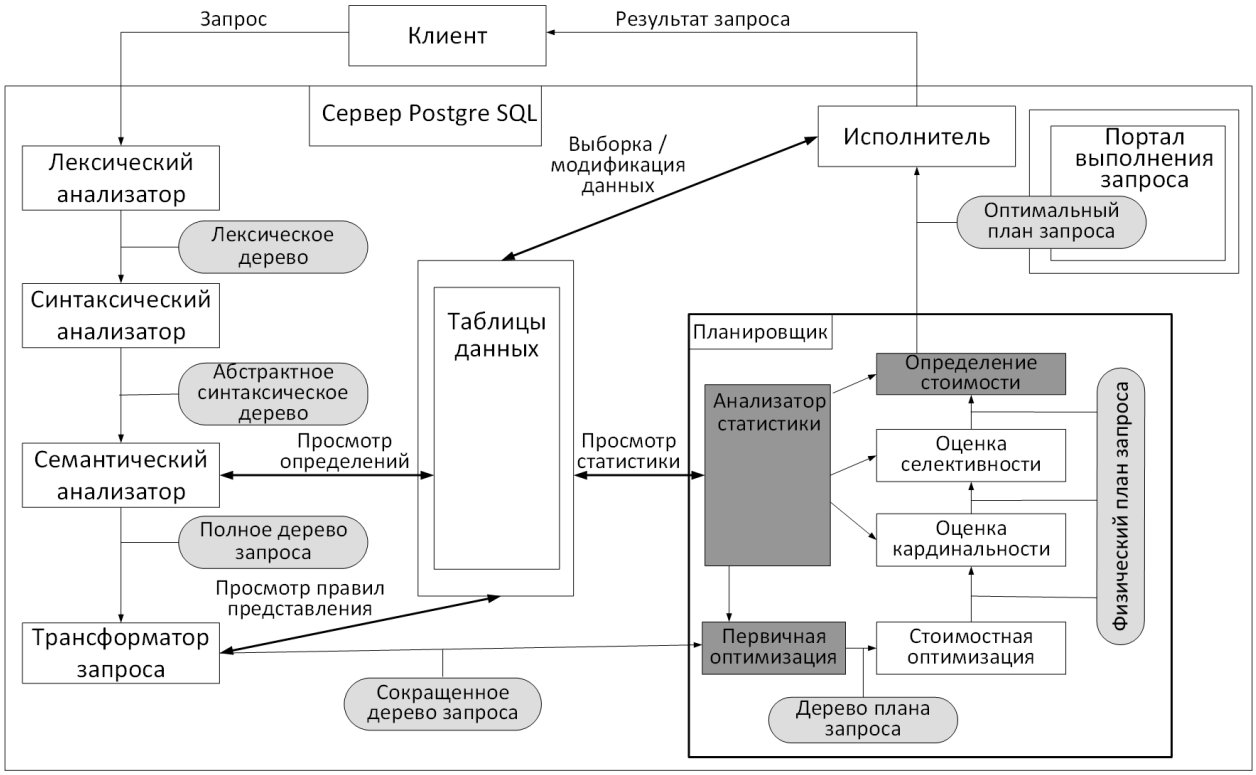

Рисунок 6 – Архитектура программной системы процессора приближенной обработки конвейера запросов

На рисунке 6 представлена архитектура специального программного обеспечения, которое предоставляет функции приближенной обработки конвейера запросов на основе модифицированного процессора запросов реляционной системы управления базами данными PostgreSQL.

Структура программного обеспечения приближенной обработки конвейера запросов предусматривает взаимодействие с системой управления базами данных в части компонента сбора статистики, что необходимо для оценки статистических параметров, используемых для оптимизации распределения пространства выборки. Программное обеспечение применяет механизм кэширования промежуточных результатов частных запросов

в конвейере, что позволяет производить уточняющую выборку при повторном выполнении запроса и, таким образом, повысить точность результата с учетом вновь поступивших данных или снизить время обработки конвейера запросов.

Программное обеспечение приближенной обработки конвейера данных может быть размещено в пространстве модуля, функционирующего в рамках выполнения протокола расширенных запросов. Архитектура системы предполагает работу с последовательностью запросов, которые взаимосвязаны по результатам выполнения частных запросов из обрабатываемого конвейера.

Для сравнительного эксперимента было выбрано существующее расширение AQP++ для системы управления базами данных PostgreSQL, которое обеспечивает приближенную обработку запросов на основе метода равномерной случайной выборки.

В эксперименте были получены численные значения средней процентной ошибки для существующего специального программного обеспечения AQP++ и предлагаемого, которое условно названо Opt-AQP. Сравнительный график представлен на рисунке 7. Анализ результатов показывает снижение дисперсии средней процентной ошибки на 10,84 % при использовании оптимизированного пространства выборки на основе разработанной метрики качества выборки.

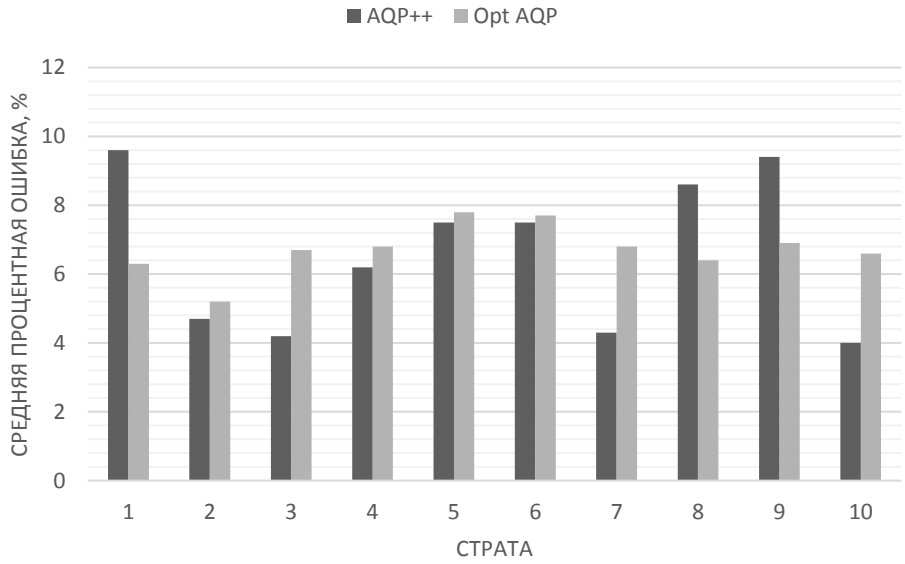

Рисунок 7 – Сравнительный график средней процентной ошибки по стратам результата запроса при использовании расширения AQP++ и предлагаемого программного обеспечения Opt-AQP

Таким образом, разработана архитектура процессора запросов реляционной системы управления базами данных и конвейера запросов, проведен сравнительный эксперимент. На реализованные элементы специального программного обеспечения получено свидетельство о регистрации программы для ЭВМ в реестре ФИПС.

ЗАКЛЮЧЕНИЕ

В процессе выполнения диссертационного исследования получены следующие основные результаты:

1. Проведен анализ состояния проблемы повышения точности приближенной обработки запросов в реляционных системах управления базами данных.
2. Разработано математическое обеспечение для приближенной обработки запросов в системах управления базами данных реляционного типа.
3. Разработан алгоритм приближенной обработки запросов в процессоре запросов реляционных систем управления базами данных.
4. Разработана архитектура программного обеспечения процессора запросов системы управления базами данных, поддерживающая приближенную обработку конвейера запросов в хранилищах данных.
5. Разработаны элементы специального программного обеспечения процессора запросов системы управления базами данных, обеспечивающие приближенную обработку конвейера запросов, и проведены численные эксперименты по оцениванию точности результатов и задержки обработки запросов на основе представленных решений.

Основные результаты диссертации опубликованы в следующих работах:

Публикации в изданиях списка ВАК:

1. С. В. Козлов Подходы к приближенной обработке аналитических запросов в реляционных системах управления базами данных / С. В. Козлов, А. А. Невров, И. П. Латышев, А. В. Филимонов // I-methods. – 2021. – № 4 (13). Режим доступа : http://intech-spc.com/wp-content/uploads/archive/2021/4/7-kozlov.pdf.

2. Д. Д. Громей Оптимизация распределения пространства выборки для запросов с группированием в процессе их приближенной обработки / Д. Д. Громей, С. В. Козлов, А. В. Филимонов // Системы управления и информационные технологии. – 2022. – № 3 (89). – С. 48–54.

3. Филимонов, А. В. Разработка алгоритма приближенной обработки конвейера запросов в реляционной системе управления базами данных / А. В. Филимонов // Моделирование, оптимизация и информационные технологии. – 2022. – № 10 (3). Режим доступа : https://moitvivt.ru/ru/journal/article?id=1242.

Публикации в изданиях, индексируемых в Scopus:

4. Morkovin S. V. Comparative analysis of metrics for estimating the introduced distortions in images when injection digital watermarks in the frequency spectrum / S. V. Morkovin, A. V. Filimonov, M. S. Rykshin, O. L. Tsvetkova // AIP Conference Proceedings. – 2021. – № 2402 (1) : 050013.

Свидетельства о государственной регистрации программ и патенты:

5. Свидетельство о государственной регистрации программы для ЭВМ № 2022618111. ProjectZero Server / С. В. Шекшуев, Д. Л. Жусов, М. А. Сазонов, А. В. Филимонов; заявители и правообладатели: С. В. Шекшуев, Д. Л. Жусов, М. А. Сазонов, А. В. Филимонов; опубл. 28.04.2022.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2023611412. Программа оценки эффективности сложных программно-аппаратных комплексов / А. С. Белов, С. В. Козлов, М. М. Добрышин, А. В. Филимонов, С. В. Шекшуев, Ю. А. Кирикова; заявитель и правообладатель: Федеральное государственное казенное военное образовательное учреждение высшего образования «Академия Федеральной службы охраны Российской Федерации»; опубл. 19.01.2023.

Статьи и материалы конференций:

7. Filimonov, A. V. An approach to the synthesis of persistent data structures in relative database management systems / S. V. Kozlov, A. V. Filimonov // Modern informatization problems in simulation and social technologies MIP-2020’ES. Proceedings of the XXV-th International Open Science Conference, Yelm, WA, USA. – January 2020. – pp. 67–70.

8. Филимонов, А. В. Подходы к приближенной обработке запросов в реляционных системах управления базами данных / С. В. Козлов, А. В. Филимонов // Всероссийская межведомственная научно-техническая конференция по теоретическим и прикладным проблемам развития и совершенствования автоматизированных систем управления и связи «НАУКА и АСУ – 2021». – Москва. – 2021. – С. 38–39.

9. Filimonov, A. V. Online aggregation methods in solving the problem of approximate query processing in database management systems / A. V. Filimonov // Modern informatization problems in simulation and social technologies MIP-2022’AS. Proceedings of the XXVII-th International Open Science Conference, Yelm, WA, USA. – January 2022. – pp. 209–213.

Подписано в печать 10.02.2023. Формат 29,7×42/4. Усл. печ. л. 1,0. Зак. № __. Тир. 80 экз.
Размножено в тип. Академии ФСО России. 302015, г. Орёл, ул. Приборостроительная, 35.