На правах рукописи

Ярыгина Анна Сергеевна

Методы и средства эффективного выполнения сценариев аналитической обработки данных на основе оптимизации и приближенных вычислений

05.13.17 — теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2015

Работа выполнена в Санкт-Петербургском государственном университете на кафедре информационно-аналитических систем.

Научный руководитель:	Новиков Борис Асенович доктор физико-математических наук, профессор
Официальные оппоненты:	Махортов Сергей Дмитриевич доктор физико-математических наук, доцент (ФГБОУ ВО “ВГУ”, зав. кафедрой)
	Бакин Евгений Александрович кандидат технических наук (ФГАОУ ВО ГУАП, доцент)
Ведущая организация:	МГУ имени М.В. Ломоносова

Защита диссертации состоится “17” марта 2016 года в 15-30 часов на заседании диссертационного совета Д212.232.51, созданного на базе Санкт-Петербургского государственного университета, по адресу: 198504 Санкт-Петербург, Петродворец, Университетский пр., д. 28, математико-механический факультет, ауд. 405.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб., д. 7/9 и на сайте http://spbu.ru/science/disser/soiskatelyu-uchjonoj-stepeni/dis-list/details/14/694.html.

Автореферат разослан “___”_____ 2016 года.

Ученый секретарь
диссертационного совета Д212.232.51,
доктор физико-математических наук,
профессор Демьянович Ю. К.

Общая характеристика работы

Актуальность темы исследования. Проблема эффективного анализа больших объемов данных остается актуальной на протяжении десятилетий. В последнее время в связи с ростом объема доступных данных и появлением новых источников информации, например, социальных сетей и сенсоров, эта проблема обострилась. В тоже время возможности обработки данных развиваются медленнее: это относится как к техническим возможностям оборудования и алгоритмам анализа данных, так и к средствам спецификации и программной реализации необходимых вычислений.

Многие исследования в области анализа данных концентрируются вокруг понятия больших данных (big data), которое со временем вобрало в себя широкий спектр значений: большие объемы данных, их разнообразие и качество, а также необходимость их своевременной обработки. В литературе представлено множество примеров систем анализа больших данных: SCOPE, Asterix, Hive.

Разнообразие больших данных связано с множеством типов данных, одновременно задействованных в сложном аналитическом сценарии. Информационные ресурсы, различающиеся по типу хранимых данных и характеру доступа к ним, могут быть использованы в одном процессе обработки данных. Системы анализа данных могут быть неоднородны с точки зрения модели, динамики, надежности и типа данных, а также по типу запросов на их извлечение и анализ. Необходимость совместного анализа данных, извлеченных из разнородных источников, появляется во многих приложениях и прикладных областях, в том числе при расширенном поиске, персонализации и аналитической обработке.

Высокоуровневые декларативные языки являются эффективным инструментом описания сложных аналитических сценариев, поскольку они позволяют скрыть сложность работы в неоднородной среде и организации параллельных вычислений. Системы, интегрирующие в себе разные подходы к анализу данных, как правило, используют промежуточные алгебраические языки, которые в дополнение к базовой выразительности традиционных декларативных языков запросов включают в себя их нечеткие расширения и специализированные операции для определенных классов задач: например, средства анализа естественных языков или изображений.

Скорость генерации и накопления данных, например в социальных сетях и разнообразными сенсорами, ведет к стремительному увеличению объемов анализируемых данных. Это приводит к необходимости решения вопросов, связанных со скоростью анализа. Необходимость увеличения скорости анализа больших объемов данных требует проработки на всех уровнях от аппаратного до языкового, в частности высокая эффективность может быть достигнута использованием декларативных языков запросов.

При решении задачи современного анализа данных возникает необходимость в приближенном выполнении запросов, поскольку все чаще точные вычисления невозможны или бессмысленны. Растущий спрос на обработку больших объемов

данных за ограниченное время, а также современные методы анализа данных на основе подобия вызывают необходимость в приближенных вычислениях. Например, системы Blinkdb, Sciborq поддерживают приближенное параллельное выполнение запросов в реальном времени, предоставляя пользователю статистические гарантии качества неточного результата.

Значительная часть элементов традиционной архитектуры систем выполнения декларативных запросов, включая компиляцию в промежуточный алгебраический язык, оптимизацию и интерпретацию алгебраического выражения, требует пересмотра в контексте новых моделей данных, классов систем хранения и доступа, и вычислительных архитектур.

Таким образом, современные проблемы анализа больших данных требуют создания методов и средств, обеспечивающих реализацию систем, которые позволят единообразно формулировать запросы к разнородным данным и описывать их обработку; реализуют эффективное выполнение сложных сценариев анализа данных на основе их оптимизации; и будут поддерживать приближенное выполнение запросов в реальном времени, то есть будут обеспечивать предсказуемое и контролируемое время ответа на запрос.

Целью исследования являлась разработка методов оптимизации для приближенного выполнения сценариев нечеткого анализа данных.

Для достижения цели были поставлены и решены следующие **задачи**:

- Разработать систему понятий и теоретическую модель оптимизации и контролируемого приближенного выполнения нечетких запросов;
- Разработать методы решения задачи распределения ограниченного количества вычислительных ресурсов среди операций в плане приближенного выполнения запроса;
- Предложить методы решения задачи многокритериальной оптимизации запросов, допускающих приближенное выполнение, ориентированные на специфические ограничения на количество вычислительных ресурсов и качество ответа.

Положения, выносимые на защиту:

- Предложена система понятий, составляющих теоретическую модель, формализующую оптимизацию и контролируемое приближенное выполнение декларативных запросов, на основе модели стоимости и качества операций;
- Разработана математическая модель распределения ресурсов среди операций в плане выполнения запроса и решена задача выбора оптимального распределения;
- Разработан приближенный алгоритм распределения ограниченного количества вычислительных ресурсов среди операций в плане выполнения запроса на основе моделей качества;

- Разработаны алгоритмы бикритериальной оптимизации запросов, ориентированные на специфические в контексте приближенного выполнения, ограничения на количество вычислительных ресурсов и качество ответа.

Методология и методы исследования. Объектом исследования являлась совокупность моделей, методов, инструментов оптимизации и приближенного выполнения сценариев анализа данных. Предметом исследования являлись задачи распределения ресурсов и бикритериальной оптимизации запросов, допускающих контролируемое приближенное выполнение. Методология работы основана на обобщении, индукции и дедукции, математическом моделировании, анализе и синтезе теоретического и практического материала. В работе использовались методы исследования операций; методы теории баз данных; методы теории алгоритмов; принципы построения архитектур программных систем; практика программной инженерии.

Степень разработанности темы. Следующие аспекты темы были проработаны исследовательским сообществом к моменту начала работы над темой диссертации. Алгоритмы контролируемого приближенного выполнения отдельных задач анализа данных и ограниченных классов запросов описаны во многих статьях. Методы оптимизации для точного выполнения запросов проработаны в классической теории баз данных. Модели стоимости реляционных операций используются в оптимизаторах запросов современных СУБД. Концепции качества данных исследованы главным образом для структурированных данных. Задачи многокритериальной и параметрической оптимизации исследованы в контексте точного выполнения запросов. Архитектуры, существующих систем оптимизации и приближенного исполнения сценариев анализа данных, поддерживают работу с ограниченными классами запросов и методов приближенного выполнения.

Научная новизна. Возможности контролируемого приближенного выполнения запросов в системах анализа данных проработаны нами для сценариев общего вида. В контексте приближенного выполнения запросов возникает ряд новых задач оптимизации, которые были решены в этой работе.

Расширенная модель стоимости операций, предложенная в работе, формализует связь между количеством вычислительных ресурсов и качеством результата приближенных вычислений и лежит в основе последующей оптимизации запросов, допускающих приближенное выполнение.

В работе поставлена и решена задача распределения ограниченного количества вычислительных ресурсов между операциями в плане приближенного выполнения сложного сценария анализа данных.

Разработано решение специфической бикритериальной задачи оптимизации запросов, допускающих приближенное выполнение, основанное на компактном представлении зависимости оптимального плана от ограничений на исполнение.

В рамках исследования разработана архитектура системы, которая реализует исполнение сложных сценариев анализа данных при ограничениях на вычис-

лительные ресурсы, отличительной чертой которой является ее расширяемость: алгоритмы и модели не привязаны к конкретным парадигмам анализа данных и специфическим неоднородным распределенным архитектурам.

Теоретическая и практическая значимость работы. Теоретическою ценность для дальнейших исследований представляют обзор и классификация методов и систем оптимизации и приближенного выполнения сценариев анализа данных, алгоритм распределения ресурсов на основе точного теоретического решения задачи и подход к бикритериальной оптимизации запросов, допускающих приближенное исполнение.

Разработанная теоретическая модель и предложенная архитектура системы оптимизации и приближенного выполнения запросов может быть использована для расширения систем анализа данных возможностью контролируемого учета ограничений на вычислительные ресурсы и качество результата. Это позволяет реагировать на потребность аналитиков в своевременном получении результата и в работе в реальном времени.

Разработанные теоретические модели, алгоритмы и экспериментальная среда могут быть использованы для прототипирования прикладных систем анализа больших данных в различных предметных областях, например, при финансовом мониторинге, социологическом и экономическом анализе.

Достоверность и обоснованность результатов работы подтверждается использованием строгого математического аппарата, доказательствами лемм, подтверждением теоретических положений вычислительными экспериментами.

Апробация работы. Материалы работы докладывались и обсуждались на всероссийских и международных конференциях:

- 15-ая Восточно-европейская конференция "Advances in Databases and Information Systems"(20-23 сентября 2011 г., Вена, Австрия)
- Семинар аспирантов в рамках 16-й Восточно-европейской конференции "Advances in Databases and Information Systems"(17-20 сентября 2012 г., Познань, Польша)
- 16-ая Восточно-европейская конференция "Advances in Databases and Information Systems"(17-20 сентября 2012 г., Познань, Польша)
- 10-ый Коллоквиум молодых исследователей "Spring Researchers Colloquium on Databases and Information Systems"(30-31 мая 2014 г., Великий Новгород, Россия)
- 19-ая Восточно-европейская конференция "Advances in Databases and Information Systems"(9-11 сентября 2015 г., Пуатье, Франция)

Полученные результаты прошли апробацию на научном семинаре «Проблемы современных информационно-вычислительных систем» под руководством д. ф.-м. н., проф. В. А. Васенина (25 ноября 2014 года), на семинаре Московской Секции ACM SIGMOD (26 февраля 2015 года), а также неоднократно на семинарах группы исследования методов организации информации и кафедры

информационно-аналитических систем в Санкт-Петербургском Государственном Университете.

Публикации. Все результаты диссертации опубликованы в 9 научных работах [1-8,10] и одном переводе [9]. Из них: 1 публикация [1] представлена в журнале, входящем в утвержденный приказом Минобрнауки России от 25 июля 2014 г. №793 перечень рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук; 3 статьи [2,3,9] есть в индексах Web of Science и 8 работ [2-9] опубликованы в рецензируемых зарубежных изданиях, включенных в индекс Scopus.

Все исследования, результаты которых изложены в диссертационной работе, проведены лично автором в процессе научной деятельности. Из совместных публикаций в результаты диссертационной работы включен лишь тот материал, который непосредственно принадлежит автору.

В статьях [2,3] А.С. Ярыгиной принадлежит анализ литературы, доказательство лемм, идея и реализация алгоритма, проведение вычислительных экспериментов. В статье [4] А.С. Ярыгиной принадлежит сведение общей задачи оптимизации к бикритериальной и параметрической, разработка алгоритма, проведение вычислительных экспериментов. В работе [5] Ярыгиной принадлежит детальная проработка архитектуры системы анализа данных. Б.А. Новикову в работах [2,3,4,5,7] принадлежат общие постановки задач и обоснование их актуальности, формальная модель качества. А.С. Ярыгиной в статье [6] принадлежит проработка алгебраических свойств операций и соотношений между ними; Б.А. Новикову принадлежит концептуальная модель исполнителя декларативных сценариев; Н.С. Васильевой обоснование актуальности задачи в контексте анализа больших данных. В работе [7] А.С. Ярыгиной принадлежит разработка расширенных моделей стоимости и качества для ряда операций; О.А. Долматовой принадлежит реализация моделей и проведение экспериментальной оценки. А.С. Ярыгиной в статье [10] принадлежит общая постановка задачи оптимизации запросов; Б.А. Новикову принадлежит позиционирование задачи в контексте методов исследования операций. В статье [8] А.С. Ярыгиной принадлежит сравнительный анализ методов синтеза и нормализации, реализация алгоритмов, проведение вычислительных экспериментов; Б.А. Новикову принадлежит общая постановка задачи и обоснование ее актуальности, алгебраическая систематизация методов синтеза; Н.С. Васильевой принадлежит реализация методов вычисления оценок подобия изображений.

Структура и объем диссертации. Диссертационная работа состоит из введения, 5 глав, заключения и списка литературы. Общий объем диссертации - 149 страниц. Список литературы содержит 100 названий. Рисунки и таблицы нумеруются по главам.

Содержание работы

Во **введении** сформулированы цель работы и задачи, решенные в рамках диссертационного исследования; обосновываются актуальность темы и научная новизна полученных результатов.

В **первой главе** проведен анализ работ исследовательского сообщества, посвященных точному выполнению и оптимизации декларативных запросов в традиционных системах баз данных. Особое внимание в этой главе уделено существующим подходам к работе с нечеткими запросами, а также методам приближенного выполнения сценариев анализа данных. Построена классификация подходов к оптимизации и приближенному выполнению нечетких запросов на основе сопоставления с методами, разработанными для точных декларативных запросов. Завершает главу обсуждение современных систем анализа больших данных на основе распределенных и приближенных вычислений. Проведенный анализ подходов позволил выделить основные направления дальнейшего развития методов оптимизации и приближенного выполнения сложных запросов к разнородным и распределенным источникам информации.

В **главе 2** определена теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов: понятия качества и вычислительных ресурсов, модель стоимости и качества операций. Также уточнены задачи оптимизации запросов в контексте их приближенного выполнения, и поставлена математическая задача распределения ресурсов.

В работе рассматривается алгебраический слой, который функционирует между пользовательским интерфейсом и вычислителями в послойной архитектуре системы анализа данных; введена алгебра на основе понятия нечетких множеств, которая позволяет единообразно соединять в одном аналитическом сценарии разные виды обработки разнородных данных.

Центральным понятием модели является Q-множество, представляющее результат выполнения абстрактного нечеткого запроса. Q-множество это тройка $\{(q, X, S) \in \mathbb{Q}\}$, в которой $q \in Q$ - запрос, X - базовое множество объектов, $S : X \to [0, 1]$ - функция оценки объектов из множества X по запросу q.

Над множеством $\mathbb{Q}$ можно определить n-арные алгебраические операции, составляющие пространство $\mathbb{O}$, как $o : P \times \mathbb{Q}^n \to \mathbb{Q}$, где P множество параметров операции. Операции, определенные над множеством Q-множеств, могут составлять алгебры различной выразительной силы. В работе предложен набор основных операций, являющихся расширением реляционных, например, теоретико-множественные операции, соединение и агрегирование, учитывающие оценки объектов в Q-множествах. Предложенная алгебра может быть расширена новыми алгебраическими операциями и обогащена с помощью реализации новых функций, используемых для конфигурации родовых операций алгебры.

Множество выражений алгебры над переменными V обозначим через $\mathbb{E}(V)$. Выражение алгебры, в котором нет переменных, называется определенным, и множество определенных выражений алгебры $\mathbb{E}$ строится рекурсивно. Для вы-

ражения $e \in \mathbb{E}$ $O(e)$ обозначает множество всех его подвыражений, а $o \in O(e)$ обозначает корневую операцию выражения, содержащегося в этом множестве.

В алгебре может выполняться ряд алгебраических тождеств, аналогичных алгебраическим тождествам реляционной алгебры. Замена выражения тождественным называется трансформацией. Выражения $e_1, e_2 \in \mathbb{E}$ эквивалентны, если существует конечная цепочка трансформаций (с заменой подвыражений переменными и обратно), переводящая e_1 в e_2.

Для любого выражения в алгебре можно построить множество эквивалентных ему на основе алгебраических тождеств. Запросом называется множество эквивалентных выражений в алгебре, которое может быть представлено одним из них $e \in \mathbb{E}$. Для запроса $e \in \mathbb{E}$ через $P(e) = \{e_i | e_i \equiv e\}$ обозначим множество планов выполнения запроса.

На выполнение операции могут влиять параметры, такие как конфигурация вычислителей или специфические параметры вызова алгоритма, реализующего операцию. Для операции $o \in \mathbb{O}$ пространство конфигураций $C(o)$ представляет пространство значений параметров, влияющих на ее выполнение. Операцию $o \in \mathbb{O}$ с конфигурацией $p \in C(o)$ будем обозначать через $o(p)$. Для плана выполнения запроса $e \in \mathbb{E}$ его конфигурация определяется множеством конфигураций операций его составляющих: конфигурацией плана называется функция $p : O(e) \to \bigcup_{o \in O(e)} C(o)$ такая, что $\forall o \in O(e)$ $p(o) \in C(o)$. План выполнения запроса $e \in \mathbb{E}$ с конфигурацией $p \in C(e)$ будем обозначать через $e(p)$. Пространство сконфигурированных планов выполнения запроса $\bar{e} \in \mathbb{E}$ обозначим как $\mathfrak{P}(\bar{e}) = \{(e, p) | e \in P(\bar{e}), p \in C(e)\}$.

Для операции $o \in \mathbb{O}$ с конфигурацией $p \in C(o)$ критерием эффективности называется функция $\varsigma(o(p)) : \mathbb{Q}^n \to \mathbb{R}$. Оценки эффективности планов строятся на основе оценок эффективности отдельных операций, участвующих в алгебраическом выражении. Для плана выполнения запроса $e \in \mathbb{E}$ с конфигурацией $p \in C(e)$ его эффективность по критерию ς вычисляется как функция от стоимостей отдельных операций его составляющих, то есть $\varsigma(e(p)) = F(\{\varsigma(o(p_o)) | o \in O(e), p_o = p(o)\}) \in \mathbb{R}$, где F - функция агрегирования оценок эффективности операций в плане.

Для каждого критерия оценки эффективности определен частичный порядок, формализующий относительную эффективность операций по этому критерию: для сконфигурированных планов выполнения запроса $(e_1, p_1), (e_2, p_2) \in \mathfrak{P}(e)$ $\varsigma(e_1(p_1)) \prec \varsigma(e_2(p_2))$, если $e_1(p_1)$ не менее эффективен, чем $e_2(p_2)$, по критерию ς.

Эффективность операций (и, следовательно, планов) может измеряться одновременно на основе различных критериев, таких как время вычислений, объемы ввода/вывода, процессорное время. Для операции $o \in \mathbb{O}$ с конфигурацией $p \in C(o)$ функцией стоимости называется функция $\bar{\varsigma}(o(p)) : \mathbb{Q}^n \to \mathbb{R}^m$.

На пространстве значений функций стоимости определено отношение, позволяющее сравнивать стоимость разных операций и выражений: для сконфигурированных планов выполнения запроса $(e_1, p_1), (e_2, p_2) \in \mathfrak{P}(e)$ $\bar{\varsigma}(e_1(p_1)) \prec$

$\bar{\varsigma}(e_2(p_2))$ в пространстве $\mathbb{R}^m$, тогда и только тогда, когда $\exists i \in [1 : m] : \bar{\varsigma}(e_1(p_1))_i \prec \bar{\varsigma}(e_2(p_2))_i$.

Точное вычисление оценок эффективности планов и операций невозможно, поскольку требует исполнения плана, поэтому в качестве приближения функций стоимости операций рассматриваются их модели стоимости: для операции $o \in \mathbb{O}$ с конфигурацией $p \in C(o)$ модель стоимости это функция $\bar{c}(o(p)) : \mathbb{S} \to \mathbb{R}^m$, где $\mathbb{S}$ - пространство свойств Q-множеств. Таким образом, при оценке эффективности планов вычисление функции стоимости $\bar{\varsigma}$ заменяется вычислением $\bar{c}$. Для плана выполнения запроса $e \in \mathbb{E}$ с конфигурацией $p \in C(e)$ его приближенная стоимость, вычисленная на основе моделей стоимости, обозначается через $\bar{c}(e(p)) \in \mathbb{R}^m$.

На основе понятия модели стоимости в диссертации сформулированы задачи однокритериальной, многокритериальной и параметрической оптимизации.

В этом исследовании мы рассматривали возможность приближенного выполнения запросов, в ситуациях, когда для вычисления неточного ответа тратится меньшее количество ресурсов. Для поддержки эффективной реализации, используются приближенные контролируемые алгоритмы исполнения алгебраических операций, которые позволяют влиять на количество вычислительных ресурсов и качество результата.

Для управления вычислениями применяются конфигурации алгоритмов: то есть для операции $o \in \mathbb{O}$ конфигурация $p \in C(o)$ может содержать параметры, определяющие контролируемое приближенное поведение. Мы фокусируемся на приближенном выполнении запросов и для упрощения обозначений считаем, что конфигурация операции отвечает только за ее контролируемое приближенное выполнение.

Различные типы ресурсов могут быть использованы для оценки эффективности выполнения запросов и для ограничения их приближенного выполнения. Очевидно, что оптимальный план зависит от типа ресурсов, который мы минимизируем. Однако, мы считаем, что функция стоимости оценивает скалярное значение ресурсов, отражающее его количество, но не тип. Таким образом, ресурсы это выделенный агрегированный критерий оптимизации, принимающий значения в пространстве $\mathfrak{R} \subset \mathbb{R}$. Мы предполагаем, что ресурсы могут быть распределены между операциями в плане и стоимость плана равна сумме количеств ресурсов, выделенных отдельным операциям.

Приближенное поведение операции зависит от выделенного на ее выполнение количества вычислительных ресурсов. Для операции $o \in \mathbb{O}$ существует функция отображения количества ресурсов, выделенных на выполнение операции, в параметры ее конфигурации: $m_o : \mathfrak{R} \to C(o)$. Для $o \in \mathbb{O}$ и $t \in \mathfrak{R}$ через $o(t)$ обозначим сконфигурированную операцию при данном количестве ресурсов, то есть $o(t) = o(m_o(t))$.

Несмотря на то, что понятие качества данных сложное и часто рассматривается в нескольких размерностях, мы предполагаем, что качество набора данных и Q-множества, оценивается одним числом, то есть мы применяем агрегирован-

ный критерий.

Для построения моделей стоимости/качества используются оценки относительного качества операций, принимающие значения в пространстве $\mathfrak{Q} \subset \mathbb{R}$: то есть отношения качества результирующего Q-множества к качеству аргументов операции, поскольку даже операции с неограниченным количеством ресурсов могут изменять качество данных.

Когда мы ограничиваем количество вычислительных ресурсов на исполнение операции, ее относительное качество может оцениваться через отношение абсолютного качества результата, полученного при ограничениях, к качеству достижимому при неограниченных ресурсах.

Для операции $o \in \mathbb{O}$ с конфигурацией $p \in C(o)$ бикритериальная модель стоимости это функция $\bar{c}(o(p)) : \mathbb{S} \to \mathfrak{R} \times \mathfrak{Q}$. Таким образом, в расширенной модели стоимости связаны два ключевых критерия оценки приближенных планов выполнения запросов: качество и вычислительные ресурсы. Для операции $o \in \mathbb{O}$ модель качества это пара функций $r : \mathbb{S} \times \mathfrak{R} \to \mathfrak{Q}$ и $r^{-1} : \mathbb{S} \times \mathfrak{Q} \to \mathfrak{R}$.

Задача распределения ресурсов, состоит в том, чтобы распределить ограниченное количество вычислительных ресурсов между операциями в плане выполнения запроса так, чтобы максимизировать качество ответа. Для заданного плана выполнения запроса $e \in \mathbb{E}$ и фиксированного количества ресурсов $T \in \mathfrak{R}$ множество $\bar{t}_e = \{t_o \geq 0 | o \in O(e), t_o \in \mathfrak{R}\} : \sum_{o \in O(e)} t_o \leq T$ называется распределением ресурсов, t_o определяет количество ресурсов, выделяемое $o \in O(e)$. Через $\bar{T}_e$ обозначим множество всех возможных распределений вычислительных ресурсов T, а через $\bar{R}_e = \{\bar{T}_e | T \in \mathfrak{R}\}$. Различные стратегии распределения ресурсов между операциями в плане выполнения запроса определяют качество ответа. Для плана выполнения запроса $e \in \mathbb{E}$ относительным качеством называется функция $q(e) : \bar{R}_e \to \mathfrak{Q}$.

Задача 1 *Для заданного плана выполнения запроса $e \in \mathbb{E}$ и фиксированного количества ресурсов $T \in \mathfrak{R}$ задача распределения ресурсов ставится следующим образом:* $\arg\max_{\bar{t}_e \in \bar{T}_e} q(e)(\bar{t}_e)$.

Предполагая решение задачи распределения ресурсов, можно определить функцию качества плана: для плана выполнения запроса $e \in \mathbb{E}$ определим его модель качества $Q_e : \mathfrak{R} \to \mathfrak{Q}$ так, что $Q_e(T) = \max_{\bar{t}_e \in \bar{T}_e} q(e)(\bar{t}_e)$.

В контексте приближенного выполнения мы рассматриваем два критерия оптимизации запроса: количество вычислительных ресурсов и качество результата, поэтому мы можем рассматривать задачу бикритериальной оптимизации.

Задача 2 *Задача бикритериальной оптимизации запроса $e \in \mathbb{E}$, допускающего приближенное исполнение, ставится следующим образом: найти* $\{(\bar{e}, t_{\bar{e}}) | \bar{e} \in P(e), t_{\bar{e}} \in \bar{T}_{\bar{e}} \in \bar{R}_{\bar{e}}, \forall \hat{e} \in P(e), t_{\hat{e}} \in \bar{T}_{\hat{e}}\ q(\bar{e})(t_{\bar{e}}) \geq q(\hat{e})(t_{\hat{e}})\}$.

Поскольку мы рассматриваем оптимизацию и контролируемое приближенное выполнение сценариев анализа данных, возникает специфическая постановка задачи параметрической оптимизации.

Задача 3 *Задача параметрической оптимизации запроса* $e \in \mathbb{E}$ *ставится следующим образом: найти функцию* $f : \mathfrak{R} \to P(e)$ *такую, что если* $f(T) = \bar{e}$, *то* $\forall \hat{e} \in P(e)\ Q_{\bar{e}}(T) \geq Q_{\hat{e}}(T)$, *где* $T \in \mathfrak{R}$.

Поставленные задачи бикритериальной и параметрической оптимизации запросов, допускающих приближенное исполнение, эквивалентны.

Конечная задача выбора оптимального плана для последующего исполнения может быть поставлена следующим образом: для заданного запроса $e \in \mathbb{E}$ и фиксированного количества ресурсов $T \in \mathfrak{R}$ найти $\bar{e} \in P(e)$ такой, что $\forall \hat{e} \in P(e)\ Q_{\bar{e}}(T) \geq Q_{\hat{e}}(T)$.

В **третьей главе** описано решение задачи распределения ресурсов: для плана выполнения запроса, допускающего приближенное исполнение, конструируется стратегия распределения заданного количества вычислительных ресурсов, обеспечивающая лучшее возможное качество результата.

План выполнения запроса может быть представлен деревом, в котором вершины представляют операции, а ребра соединяют операции с их аргументами в выражении плана. Деревом плана выполнения запроса $e \in \mathbb{E}$ называется граф $P = (V(P), E(P))$: любой $o \in O(e)$ соответствует вершина $l_o \in V$; если $o_1, o_2 \in O(e)$ и o_2 является аргументом o_1, то ребро $\epsilon \in E(P)$ связывает соответствующие вершины l_{o_1} и l_{o_2}. Для дерева плана P и для вершины $l \in V(P)$ $args(l) \subset V(P)$ — множество вершин-детей; для не корневой операции $parent(l)$ — родительская вершина; $\bar{l}$ — поддерево с вершиной в l.

Для вершины в дереве плана на основе модели качества соответствующей операции построим функцию качества: для $e \in \mathbb{E}$ для вершины $l \in V(P)$, соответствующей операции $o \in O(e)$ с моделью качества $r : \mathbb{S} \times \mathfrak{R} \to \mathfrak{Q}$, функцией качества называется $q(l) : \mathfrak{R} \to \mathfrak{Q}$ такая, что $q(l) = r(s)$, где $s \in \mathbb{S}$ - фиксированные статистики аргумента операции.

Для заданного дерева плана выполнения запроса P и фиксированного количества вычислительных ресурсов $T \in \mathfrak{R}$ множество $\bar{t}_P = \{t_l \geq 0 | l \in V(P), t_l \in \mathfrak{R}\} : \sum_{l \in V(P)} t_l \leq T$ называется распределением ресурсов, t_l определяет количество ресурсов, выделяемое $l \in V(P)$. Количество ресурсов, выделенное (под)дереву с корнем в l обозначим через $T_{\bar{l}} \in \mathfrak{R}$. Через $\bar{T}_P$ обозначим множество всех возможных распределений ограниченного количества вычислительных ресурсов T.

Для дерева плана выполнения запроса P и распределения ресурсов $\bar{t}_P \in \bar{T}_P$ его качество определяется следующим образом: $Q_P(\bar{t}_P) = F(\{q(m)(t_m) | m \in V(P), t_m \in \bar{t}_P\})$, где F-функция агрегирования качества в плане. Для операций с несколькими аргументами мы предполагаем, что вклад аргумента с меньшим качеством доминирует при оценке качества поддерева. Более формально, качество поддерева оценивается как произведение относительного качества корневой операции и общего (минимального) качества ее аргументов: для любого $\bar{t}_{\bar{l}} \in \bar{T}_{\bar{l}}\ Q_{\bar{l}}(\bar{t}_{\bar{l}}) = q(l)(t_l) \cdot \min_{m \in args(l)} Q_{\bar{m}}(\bar{t}_{\bar{m}})$, где $t_l \in \bar{t}_{\bar{l}}, \bar{t}_{\bar{m}} = \{t_i | i \in V(\bar{m}), t_i \in \bar{t}_{\bar{l}}\}$.

Задача 1 о распределении ресурсов между операциями в плане выполне-

ния запроса переформулирована в новых терминах дерева и его вершин: заданного дерева плана P и фиксированного количества ресурсов $T \in \mathfrak{R}$ найти $\arg\max_{\bar{t}_P \in \bar{T}_P} Q_P(\bar{t}_P)$.

Для дерева плана выполнения запроса P функцией качества называется $Q(P) : \mathfrak{R} \to \mathfrak{Q}$ такая, что $\forall T \in \mathfrak{R}\ Q(P)(T) = \max_{\bar{t}_P \in \bar{T}_P} Q_P(\bar{t}_P)$.

Мы предполагаем, что функция относительного качества вершин от выделенного объема ресурсов неубывающая непрерывная ограниченная, а также может быть аппроксимирована всюду определенной в $\mathfrak{R}$ кусочно-линейной функцией:

$$\forall l \in V(P)\ \ q(l)(t) = \begin{cases} 0, & t < t^0_{min} \\ u^i(l) + s^i(l)(t - t^i_{min}), & t^i_{min} \le t < t^i_{max} \\ u^i(l) + s^{I_l}(l)(t^{I_l}_{max} - t^{I_l}_{min}), & t^{I_l}_{max} \le t \end{cases}$$

где $i \in [0, I_l]$ количество линейных сегментов, $s^i(l)$ наклон на соответствующем сегменте, и $u^i(l) + s^i(l)(t^i_{max} - t^i_{min}) = u^{i+1}(l)$.

Необходимые условия оптимальности распределения ресурсов для ограниченных топологий дерева плана и в условиях жестких предположений о поведении функции качества доказаны в леммах. На базе точного теоретического решения, применимого к отдельным фрагментам дерева, построен алгоритм итеративного распределения ресурсов, на каждом шаге которого применение точного решения будет корректно.

После выделения минимального количества ресурсов каждой операции в плане выполнения запроса оставшееся свободное количество ресурсов может быть дополнительно распределено между вершинами в дереве, чтобы улучшить качество всего плана. Таким образом, имея некоторое распределение ресурсов $\bar{t}_P \in \bar{T}_P$ мы будем строить новое расширяющее распределение ресурсов $\bar{t'}_P \in \bar{T'}_P$ такое, что $\forall l \in V(P)\ t_l \le t'_l$, где $t_l \in \bar{t}_P, t'_l \in \bar{t'}_P$.

При распределении дополнительного количества ресурсов дерево плана снова приходит в состояние, когда каждая операция достигла некоторого текущего качества. Мы работаем с приростами ресурсов и рассматриваем на каждой итерации алгоритма только один линейный сегмент функции качества: $\forall l \in V(P)$ $q(l)(\tau) = u(l) + s(l)\tau$, где $t \in [t^i_{min},\ \ t^i_{max}]$, $t_{max} = t^i_{max} - t$, $\tau \in [0, t_{max}]$, и $u(l) = q(l)(t) = u^i(l) + s^i(l)(t - t^i_{min})$.

Мы говорим об оптимальном распределении ресурсов, имея ввиду оптимальное распределение между операциями в плане дополнительного количества ресурсов, однозначно определяющее оптимальное расширяющее распределение.

Вычислительные ресурсы должны быть выделены только тем операциям, которые оказывают влияние на качество результата и составляют критическое (под)дерево, формальное определение и алгоритм построения которого проработаны в диссертации.

Лемма 3 определяет необходимые условия распределения ресурсов между вершинами, образующими путь, в случае, если их функции качества линейны.

Лемма 3 *Пусть C - критический путь, $T \in \mathfrak{R}$ количество вычислительных ресурсов, которое необходимо распределить, функция качества каждой вершины $l \in V(C)$ - линейна, то есть $q(l)(t) = u(l) + s(l)t$, где t лежит внутри области определения функции. При оптимальном распределении ресурсов $\exists V^+(C) \subseteq V(C) : t_m > 0$ тогда и только тогда, когда $m \in V^+(C)$; и*

$$t_l = \max\left(\frac{T}{n} + \frac{1}{n}\sum_{m \in V^+(C)} \frac{u(m)}{s(m)} - \frac{u(l)}{s(l)}, 0\right),$$

где $l \in V(C), n = |V^+(C)|$.

Лемма 4 определяет оптимальное распределение ресурсов между братьями в критическом (под)дереве.

Лемма 4 *Пусть C критическое (под)дерево; $l_i \in V(C)$ корневые вершины поддеревьев-братьев с линейными функциями качества, то есть $Q(\bar{l_i})(t) = U(\bar{l_i}) + S(\bar{l_i})t$, где t лежит внутри области определения функции; $T \in \mathfrak{R}$ количество вычислительных ресурсов, которое должно быть распределено между этими поддеревьями. При оптимальном распределении ресурсов $T_{\bar{l_i}} = \frac{1/S(\bar{l_i})}{\sum_j 1/S(\bar{l_j})} T$.*

Леммы 3 и 4 определяют стратегию оптимального распределения ресурсов между вершинами в дереве определенной структуры. Поскольку леммы основаны на линейном поведении функций качества, распределение ресурсов будет происходить в несколько этапов, на каждом из которых будет распределяться некоторое количество вычислительных ресурсов. Для того, чтобы иметь возможность использовать утверждения лемм, в алгоритме строятся виртуальные гипер-вершины из определенных частей дерева плана, и при работе с каждой отдельной гипер-вершиной применяется соответствующая лемма.

Алгоритм распределения ресурсов имеет полиномиальную сложность от числа операций в плане выполнения запроса и количества линейных сегментов в функциях качества каждой из операций. Результаты экспериментов показали, что предложенный приближенный алгоритм распределения ресурсов достаточно точный и эффективный.

В **четвертой главе** представлены алгоритмы бикритериальной оптимизации запросов, допускающих контролируемое приближенное выполнение при ограничениях на время вычислений и качество ответа. Алгоритмы основаны на компактном приближенном представлении множества Парето, называемом оптимальным составным сегментом. В главе описаны адаптации различных методов оптимизации для построения оптимального составного сегмента и приведены результаты их экспериментальной оценки.

Решаемая задача 3 параметрической оптимизации при ограничениях переформулирована следующим образом: для заданного запроса $e \in \mathbb{E}$ для всех

возможных значений количества ресурса $T \in \mathfrak{R}$ найти план $\bar{e} \in P(e)$ такой, что $\forall \hat{e} \in P(e)\ Q(\bar{e})(T) \geq Q(\hat{e})(T)$.

Составным сегментом S для запроса называется неубывающая кусочно-линейная функция $\bar{Q}_S : \mathfrak{R} \to \mathfrak{Q}$, каждый интервал которой ассоциирован с планом выполнения запроса, и $\bar{Q}_S(t) = Q_e(t)$, где Q_e функция качества плана $e \in \mathbb{E}$, ассоциированного с интервалом, содержащим t. Составной сегмент S может быть представлен конечным множеством $I_S = \{\langle i, e_i \rangle : \ i = [t_l, t_u)\}$ не пересекающихся интервалов значений в пространстве вычислительных ресурсов $\{[t_l, t_u)\}$, покрывающих весь диапазон $[0, \infty)$, с ассоциированными с ними планами и их функциями качества, суженными на рассматриваемый интервал.

Составной сегмент S доминирует над T, если для всех $t \in [0, \infty)$ $\bar{Q}_S(t) \geq \bar{Q}_T(t)$. Для любых сегментов S и T доминантой называется составной сегмент $D = \mathrm{dom}(S, T)$ такой, что D доминирует над S и T и любой другой составной сегмент, доминирующий над S и T, доминирует над D.

Составной сегмент запроса, который доминирует над всеми составными сегментами запроса, называется оптимальным. Мы используем оптимальный составной сегмент в качестве компактного представления множества Парето в задаче бикритериальной оптимизации.

Разработанные алгоритмы построения оптимального составного сегмента запроса на базе известных методов оптимизации запросов основаны на перечислителе планов выполнения запроса, использующем множество доступных трансформаций. Для каждого плана, сгенерированного перечислителем, строится функция качества, представленная составным сегментом; далее эти сегменты (итеративно) сливаются в один составной сегмент, который считается оптимальным при завершении анализа. В работе рассматривалось три основных алгоритма перечисления планов выполнения запроса: полное сканирование пространства планов (full), итеративное улучшение (iterative improvement) и рекурсивное построение (recursive descent).

Экспериментальная часть работы посвящена проверке существования нетривиальных составных сегментов сценариев анализа данных и сравнению производительности предложенных методов их построения.

В **пятой главе** рассматривается архитектура системы анализа данных, в которой мы продемонстрировали работу методов, описанных в предыдущих главах, и соединили воедино несколько частей: алгебра на основе подобия, модели стоимости и качества, методы распределения ресурсов и подходы к решению задачи бикритериальной оптимизации.

Система принимает пользовательский запрос и ограничения на его выполнения, оптимизирует и исполняет на вычислителях. Контекст системы включает в себя разнородные источники данных; различные, возможно распределенные, вычислители; и модуль, в котором пользователь специфицирует конкретный сценарий-запрос на анализ данных.

Архитектура системы отвечает требованиям, сформулированным в диссертации, например, возможность работы в неоднородной среде или добавления

новых операций и исполнителей, ранее не включенных в систему.

В работе выделены и проработаны два типа интерфейсов: интерфейсы работы с внешним окружением системы, и интерфейсы расширения системы.

Предложенная система оптимизации и приближенного выполнения запросов функционирует среде, включающей в себя внешний модуль, реализующий интерфейс с пользователем; разнородные источники данных и вычислители.

Во-первых, запросы передаются в систему из некоторого внешнего модуля-клиента, реализующего интерфейс с пользователем. Клиент позволяет пользователю формулировать сценарии обработки данных и анализировать результаты их исполнения, в то время как в систему передаются задачи оптимизации и выполнения запросов на алгебраическом уровне.

Архитектура системы анализа данных строится на основе расширяемой алгебры операций над Q-множествами. В архитектуре предусмотрено расширение системы анализа данных новыми операциями или алгоритмами, что не позволяет в реализации жестко ограничить набор операций уже специфицированными.

Во-вторых, предлагаемая архитектура опирается на систему неоднородных, возможно, распределенных вычислительных модулей, в качестве которых могут выступать существующие СУБД, машины map-reduce или другие внешние программы. Система отвечает за оптимизацию и приближенное выполнение запросов на алгебраическом уровне, и передает непосредственные вычислительные задачи внешним модулям. Таким образом поиск оптимального плана приближенного выполнения запроса происходит централизованно, а непосредственное исполнение может быть распределено между вычислительными модулями.

В-третьих, в системе происходит анализ данных из внешних источников разнородных по типу и методам хранения данных, например, файлов или таблиц СУБД. Неоднородность источников данных на алгебраическом уровне скрывается работой с операциями первичной выборки, которые строят представление данных в едином формате. Таким образом, так же как и в случае работы с внешними вычислительными модулями, протокол работы с внешним источником данных определяется при реализации и регистрации в системе соответствующей алгебраической операции.

Архитектура системы исполнения запросов содержит: парсер, оптимизатор и исполнитель.

Для поддержки расширяемости алгебры используется оптимизатор, настраиваемый на новые специфические операции алгебры, которые порождают новые алгебраические тождества. Оптимизатор основан на расширяемом и настраиваемом множестве трансформаций планов (алгебраический выражений). В предложенной архитектуре системы трансформации включают в себя возможные эквивалентные перестроения планов двух типов, неразличимых с точки зрения реализации: основанные на алгебраических тождествах или на различных реализациях одной операции. Такая модель позволяет легко встраивать в систему новые операции, алгоритмы, их реализующие, и соответствующие трансформации. Для обеспечения расширяемости пространство трансформаций не должно

быть реализовано глубоко внутри оптимизатора; поэтому в архитектуре системы определены единые интерфейсы, позволяющие легко встраивать новые трансформации в систему.

Для того, чтобы иметь возможность выбрать эффективный план выполнения запроса, оптимизатор основывается на моделях стоимости и качества, определенных для всех операций, в том числе для пользовательских расширений. Архитектура системы поддерживает включение новых моделей стоимости и качества операций и настройку уже существующих.

Непосредственная оптимизация запроса, допускающего контролируемое приближенное выполнение, может быть построена на основе разных по сложности и качеству ответа моделях.

Исполнитель основан на потоковой модели исполнения запросов. Непосредственное исполнение операций происходит в вычислительных модулях, а исполнитель генерирует обращения к ним и перенаправляет потоки обрабатываемых данных между различными вычислителями на основе полученного плана выполнения запроса.

Множество операций в системе может быть расширено за счет реализации новых: функций, параметризующих родовые алгоритмы операций, например, функций предиката; родовых алгоритмов базовых операций, например, приближенных; операций, например, ранжирующего соединения.

Расширяемость алгебры реализуется в архитектуре системы с помощью библиотеки операций. Библиотека операций хранит множество алгебраических операций и все связанные с ними структуры: отображение вызова операции внутри системы в вызов ее непосредственного исполнения на вычислителе; модель стоимости; модель качества; релевантные трансформации с участием операции. Простота расширяемости обеспечивается интерфейсами включения в систему трансформаций, моделей стоимости и качества.

С точки зрения архитектуры множество алгебраических операций можно разделить на три основных класса: операции первичной выборки, унарные и бинарные. При расширении алгебры новые операции из любой из трех групп должны реализовывать предопределенный в архитектуре системы интерфейс соответствующей группы или являться конфигурацией базовой реализации. Параметры обработчиков данных и функций, конфигурирующих базовую реализацию, извлекаются из параметров вызова операции, передаваемых в виде отображения имен параметров на их значения.

Заключение

Основные результаты работы:

- Подготовлен обзор, систематизирующий существующие методы оптимизации и приближенного выполнения декларативных сценариев нечеткой аналитической обработки данных;

- Предложена система понятий, составляющих теоретическую модель, формализующую оптимизацию и контролируемое приближенное выполнение декларативных запросов, на основе модели стоимости и качества операций;
- Разработана математическая модель распределения ресурсов среди операций в плане выполнения запроса и решена задача выбора оптимального распределения;
- Разработан приближенный алгоритм распределения ограниченного количества вычислительных ресурсов среди операций в плане выполнения запроса на основе моделей качества;
- Разработаны алгоритмы бикритериальной оптимизации запросов, ориентированные на специфические в контексте приближенного выполнения, ограничения на количество вычислительных ресурсов и качество ответа;
- Реализована экспериментальная среда для анализа разработанных алгоритмов распределения ресурсов и оптимизации запросов, допускающих приближенное выполнение.

Разработанные теоретические модели, алгоритмы и экспериментальную среда рекомендуется использовать для реализации прототипов прикладных систем анализа больших данных в таких предметных областях как финансовый мониторинг, социологический и экономический анализ.

В перспективах дальнейшей разработки темы целесообразно исследовать применимость предложенных моделей и методов для других классов моделей данных, например, для графов.

Работы автора по теме диссертации

[1] **Ярыгина, А. Методы выполнения и оптимизации приближенных запросов в неоднородных системах / А. Ярыгина // Программирование.— 2013.— Vol. 39.— P. 33–44.**

[2] **Yarygina, A. Optimizing resource allocation for approximate real-time query processing / A. Yarygina, B. Novikov // Computer Science and Information Systems.— 2014.— Vol. 11.— P. 69–88.**

[3] **Yarygina, A. Optimizing the resource allocation for approximate query processing / Anna Yarygina, Boris Novikov // Advances in Databases and Information Systems / Ed. by Tadeusz Morzy, Theo Harder, Robert Wrembel.— Vol. 186 of Advances in Databases and Information Systems.— Poznan, Poland: Springer Berlin Heidelberg, 2012.— P. 297–308.**

[4] **Yarygina, A. Bi-objective optimization for approximate query evaluation / Anna Yarygina, Boris Novikov // 19th East European Conference on Advances in Databases and Information Systems and**

Associated Satellite Events (ADBIS 2015) / Ed. by Tadeusz Morzy, Patrick Valduriez, Ladjel Bellatreche et al.— Communications in Computer and Information Science (CCIS).— Springer Berlin Heidelberg, 2015.— P. 153–161.

[5] **Yarygina, A. A prototype architecture for approximate real-time query optimization and processing / Anna Yarygina, Boris Novikov // The Tenth Spring Researchers Colloquium on Databases and Information Systems 2014.— 2014.— P. 24–31.**

[6] **Novikov, B. Querying big data / Boris Novikov, Natalia Vassilieva, Anna Yarygina // Proceedings of the 13th International Conference on Computer Systems and Technologies.— CompSysTech '12.— New York, NY, USA: ACM, 2012.— P. 1–10.**

[7] **Dolmatova, O. Cost models for approximate query evaluation algorithms / Oxana Dolmatova, Anna Yarygina, Boris Novikov // Databases and Information Systems. Tenth International Baltic Conference on Databases and Information Systems. Local Proceedings, Materials of Doctoral Consortium. / Ed. by A. Caplinskas, G. Dzemyda, A. Lupeikiene, O. Vasilecas.— Vilnius: Zara, 2012.— P. 20–28.**

[8] **Yarygina, A. Processing complex similarity queries: A systematic approach / Anna Yarygina, Boris Novikov, Natalia Vassilieva // ABDIS 2011 Research Communications: Proceedings II of the 5th East-European Conference on Advances in Databases and Information Systems 20 - 23 September 2011, Vienna / Ed. by Maria Bielikova, Johann Eder, A Min Tjoa.— Austrian Computer Society, 2011.—September.— P. 212–221.**

[9] **Yarygina, A. Execution and optimization techniques for approximate queries in heterogeneous systems / A. Yarygina // Programming and Computer Software.— 2013.— Vol. 39, no. 6.— P. 309–317.**

[10] Новиков, Б. А. Задачи оптимизации запросов в распределенной среде неоднородных информационных ресурсов / Борис Асенович Новиков, Анна Сергеена Ярыгина // Математика, экономика, менеджмент: 100 лет со дня рождения Л.В. Канторовича / Ed. by Иосиф Владимирович Романовский.— Санкт-Петербургский гос. университет, 2012.—7–9 февраля.— P. 57–59.