

УТВЕРЖДАЮ

Проректор — начальник
Управления научной политики и
организации научных
исследований Московского госу-
дарственного университета имени
М.В. Ломоносова,
профессор, д.ф.-м.н.
А.А. Федянин



А.А. Федянин
« 1 » декабря 2015 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» на диссертацию Ивановой Елены Владимировны «Методы параллельной обработки сверхбольших баз данных с использованием распределенных колоночных индексов», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Актуальность темы

Диссертационная работа Е.В. Ивановой посвящена разработке и исследованию эффективных методов параллельной обработки сверхбольших баз данных с использованием колоночного представления информации, ориентированных на кластерные вычислительные системы, оснащенные многоядерными ускорителями, и допускающих интеграцию с реляционными СУБД. Одним из наиболее важных классов приложений, связанным с обработкой сверхбольших баз данных, являются хранилища данных, для которых характерны запросы типа OLAP. Исследования показали, что для таких приложений выгодно использовать колоночную модель представления данных, позволяющую получить на порядок лучшую производительность по сравнению с традиционными системами баз данных, использующими строчную модель представления данных. Одной из важных нерешенных задач остается задача разработки методов параллельной обработки сверхбольших баз данных, сочетающих преимущества реляционной модели с колоночным представлением информации.

Предложенный в диссертации подход на основе распределенных колоночных индексов позволяет разработать колоночный сопроцессор баз данных для кластерной вычислительной системы с многоядерными ускорителями, позволяющий во взаимодействии с реляционной СУБД получить почти линейное ускорение при выполнении ресурсоемких реляционных операций. Поэтому работа, безусловно, является *актуальной*.

Содержание работы

Диссертация состоит из введения, четырех глав, заключения и библиографии.

Во *введении* обоснована актуальность темы, описана степень разработки темы, сформулирована цель, поставлены задачи диссертационного исследования. Дана характеристика научной новизны предлагаемой работы. Изложены основные положения, выносимые на защиту. Рассмотрены вопросы теоретической и практической значимости проведенных исследований. Перечислены методы исследований, приведена степень достоверности и апробации работы, приведены публикации по теме диссертации.

Первая глава работы носит обзорный характер. В ней рассматриваются тенденции развития аппаратного обеспечения и дается обзор научных исследований в области современных технологий баз данных. Описываются методы обработки баз данных на вычислительных системах с многоядерными ускорителями и колоночная модель хранения данных. Приводится анализ публикаций, наиболее близко относящихся к теме диссертации.

Вторая глава посвящена построению формальной доменно-колоночной модели представления данных. Вводятся колоночные индексы. Описывается оригинальный способ фрагментации колоночных индексов, названный доменно-интервальной фрагментацией. Вводится понятие транзитивной фрагментации одного колоночного индекса относительно другого для атрибутов, принадлежащих одному и тому же отношению. Рассматриваются методы декомпозиции реляционных операций на основе использования фрагментированных колоночных индексов. Для всех методов декомпозиции, приведенных во второй главе, даны математические доказательства их корректности.

В *третьей главе* описывается процесс проектирования и реализации программной системы «Колоночный сопроцессор КСОП» для кластерных вычислительных систем, представляющей собой резидентную параллельную программу, взаимодействующую с реляционной СУБД.

В *четвертой главе* приводятся результаты вычислительных экспериментов по исследованию эффективности разработанных в диссертации моделей, методов и алгоритмов обработки сверхбольших баз данных с использованием распределенных колоночных индексов.

В *заключении* приводятся основные результаты проведенных исследований.

Обоснованность и достоверность результатов и выводов диссертации

Все утверждения, связанные с декомпозицией реляционных операций, сформулированы в виде теорем и снабжены строгими доказательствами. Теоретические построения подтверждены вычислительными экспериментами.

Научная новизна

Научная новизна работы заключается в разработке автором оригинальной доменно-колоночной модели представления данных, на базе которой введены колоночные индексы с доменно-интервальной фрагментацией, и выполнением на ее основе декомпозиции основных операций реляционной алгебры. По сравнению с ранее известными методами параллельной обработки больших объемов данных предложенный подход позволяет сочетать эффективность колоночной модели хранения данных с возможностью использования мощных механизмов оптимизации запросов, разработанных для реляционной модели.

Соответствие содержания диссертации специальности 05.13.11

Содержание и результаты работы соответствуют паспорту специальности 05.13.11 по следующим областям исследований:

4. Системы управления базами данных и знаний.
8. Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования.

Теоретическая и практическая значимость

Теоретическая ценность работы состоит в том, что в ней дано формальное описание методов параллельной обработки сверхбольших баз данных с использованием распределенных колоночных индексов, включающее в себя доменно-колоночную модель представления данных. *Практическая ценность* работы заключается в том, что на базе предложенных методов и алгоритмов разработан колоночный сопроцессор для кластерной вычислительной системы с многоядерными ускорителями, позволяющий во взаимодействии с СУБД PostgreSQL получить линейное ускорение при выполнении ресурсоемких реляционных операций

Рекомендации по использованию результатов диссертации

Представленные в диссертационной работе подходы к обработке сверхбольших баз данных могут быть использованы для создания колоночных сопроцессоров для коммерческих и свободно распространяемых реляционных СУБД, а также для разработки специального учебного курса в рамках магистратуры по образовательному направлению 02.04.02 «Фундаментальная информатика и информационные технологии».

Оформление текстов диссертации и автореферата

Оформление диссертации соответствует требованиям, установленным Минобрнауки России. Автореферат в полной мере отражает содержание диссертации и позволяет составить достаточно полное представление о ней. Список литературы достаточно представительен по рассматриваемой в диссертации теме.

Апробации и публикации результатов диссертации

Результаты диссертационной работы прошли апробацию на международных научных конференциях и были опубликованы в рецензируемых научных изданиях, рекомендованных ВАК. Количество публикаций в рецензируемых научных изданиях соответствует требованиям Положения о порядке присуждения ученых степеней.

Замечания по диссертационной работе

В качестве замечаний к работе, не снижающих ее общего высокого уровня, можно отметить следующее.

1. Во введении в качестве подтверждения актуальности темы работы упоминается проблематика сверхбольших данных – видимо, имеется в виду направление Big Data. Далее приводится ссылка на М.Стоунбрейкера, предлагающего использовать для их обработки технологии СУБД. Но в самой работе не приводится ни обоснования этого тезиса, ни упоминания иных способов работы со сверхбольшими данными, основанных не на реляционных моделях. В этом контексте упоминание сверхбольших данных в целом выглядит несколько незавершенным.
2. В главе 1 больше трех страниц уделено описанию архитектуры и основных принципов работы ускорителя Intel Xeon Phi. Далее в тексте указывается, что КСОП создан с помощью технологий MPI и OpenMP и не имеет специальных оптимизаций для Phi. Более того, в главе 4 при сравнении вычислительных систем на основе Xeon X5680 и Xeon Phi 7120 выигрывает как раз первая система. Не очень понятно, почему тогда архитектуре Phi уделено столь много внимания.
3. В той же главе 1 подробно описываются GPU-ориентированные СУБД и достаточно кратко – Phi-ориентированные. В обзорной части явно не хватает обоснований или некоторого вывода, почему автор отказался от использования ускорения с помощью графических процессоров и предпочел ускорители типа Phi.
4. В главе 4 приведены результаты экспериментов с модельными данными в СУБД PostgreSQL без использования и с использованием КСОП. С нашей точки зрения, сравнение стоило бы дополнить результатами испытаний еще хотя бы одной сторонней СУБД, например, из числа упомянутых в обзорной части, с целью демонстрации преимущества решения, предложенного автором работы.

В целом считаем, что диссертация Ивановой Е.В. соответствует критериям, установленным Положением о порядке присуждения ученых степеней, включая пункт 9 Положения, и является самостоятельным и завершенным научным исследованием, содержащим решение задачи эффективного использования колоночного представления информации для параллельной обработки запросов на кластерных вычислительных системах, имеющей существенное значение для теории и практике системного программирования, а Иванова Е.В. заслуживает присуждения ей ученой степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертация и отзыв обсуждены и одобрены на объединенном заседании лабораторий НИВЦ МГУ: “Параллельных информационных технологий”, “Вычислительных систем и прикладных технологий программирования”, “Автоматизации программных вычислительных комплексов”, протокол №3 от 12 ноября 2015 г.

Заместитель директора НИВЦ МГУ
член-корреспондент РАН,



Вл.В. Воеводин

Наименование организации, предоставившей отзыв: Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова»

Почтовый адрес: Россия, 119991, г. Москва, Ленинские горы, д. 1.

Телефон: +7 (495) 939-51-66, +7(495) 939-5424

Email: voevodin@parallel.ru

Web: <http://srcc.www.msu.ru>