ОТЗЫВ

официального оппонента, кандидата физико-математических наук Пана Константина Сергеевича на диссертационную работу **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 тематики представленной работы обусловлена тем, что решение задачи эффективного использованием колоночного представления информации для параллельной обработки запросов на кластерных вычислительных системах с многоядерными ускорителями имеет существенное значение для обработки сверхбольших баз данных. В последнее время большой интерес у исследователей вызывают системы баз данных с хранением данных по столбцам. Колоночное представление в отличие от традиционного строкового представления оказывается намного более эффективным при выполнении запросов класса OLAP. Недостатком колоночного представления является низкая эффективность при выполнении строково-ориентированных операций, таких, например, как добавление или удаление кортежей. Вследствие этого колоночные СУБД могут проигрывать по производительности строковым при выполнении запросов класса OLTP. Данное диссертационное исследование предлагает решение по интеграции преимуществ столбцовой модели хранения данных в строковые СУБД, что делает работу, безусловно, актуальной.

Работа Ивановой Е.В. 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.

Первая глава диссертации посвящена описанию тенденций развития аппаратного обеспечения и обзору научных исследований в области современных технологий баз данных. Анализируются публикации, наиболее близ-

ко относящихся к теме диссертации. На основе проведенного анализа делается *обоснованный вывод о том, что перспективным решением является создание колоночного сопроцессора КСОП, совместимого с реляционной СУБД и использующего такие проектные решения, как аппаратно-независимую реализацию, массивно-параллельную и многоядерную обработку, материализационную модель обработки данных, гибридную модель хранения данных и использование оперативной памяти в качестве основного места хранения данных.*

Во второй главе описываются формальная доменно-колоночная модель представления данных, оригинальный способ фрагментации колоночных индексов и методы декомпозиции реляционных операций на основе использования фрагментированных колоночных индексов. В соответствии с этим *является обоснованным вывод о том, что доменно-колоночная модель и распределенные колоночные индексы позволяют выполнить декомпозицию всех основных реляционных операций на подоперации, выполнение которых не требует обменов данными.*

Третья глава посвящена разработке программной системы «Колоночный СОПроцессор (КСОП)» для кластерных вычислительных систем, реализующей доменно-колоночную модель представления данных и методы декомпозиции реляционных операций. Результаты третьей главы позволяют сделать *обоснованный вывод о том, что теоретические разработки диссертационной работы могут быть воплощены в программную систему для обработки сверхбольших баз данных на кластерной вычислительной системе, оснащенной многоядерными ускорителями.*

В четвертой главе описывается методика проведения экспериментов, приводятся методы генерации синтетической базы данных, разработанной на основе теста TPC-H. Приводятся результаты вычислительных экспериментов. В качестве СУБД, взаимодействующей с КСОП, использовалась реляционная СУБД с открытыми кодами PostgreSQL. Результаты, полученные в этой главе, позволяют сделать *обоснованный вывод о том, что подходы и методы параллельного выполнения запросов класса OLAP, разработанные на базе доменно-колоночной модели, демонстрируют хорошую масштабируемость (до*

нескольких сотен процессорных узлов и десятков тысяч процессорных ядер) для запросов с большой селективностью, которые являются типичными для хранилищ данных.

Таким образом, все научные положения, выводы и рекомендации, сформулированные в диссертации, полностью **обоснованы**.

Достоверность полученных результатов обеспечивается теоремами, снабженными детальными доказательствами, и подкрепляется результатами вычислительных экспериментов.

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.

Представленные в работе методы и алгоритмы разработаны лично Е.В. Ивановой. Основные результаты достаточно полно опубликованы в рецензируемых научных изданиях, входящих в перечень ВАК. Автореферат правильно отражает содержание диссертации.

Диссертационная работа Е.В. Ивановой характеризуется высоким научным уровнем, превосходной математической культурой и ясным изложением материала.

В качестве **замечаний** к работе, не снижающих ее общего высокого уровня, необходимо отметить следующее:

1. В правой части равенства (15), на странице 43, аргументом функции фрагментации служит отношение, хотя функция фрагментации, в отличие от реляционных операций, определена на множестве кортежей, а не отношений.
2. Введенная автором на странице 37 операция разыменования «&» также используется в автореферате, но в автореферате ее смысл не поясняется.

3. Для избежания коллизий в случае хеш-индексов в работе предлагается использовать инъективную функцию. Однако она может быть плохо применима на практике, поскольку потребует большой разрядности — длина чисел будет прямо пропорциональна количеству индексируемых хэш-индексом атрибутов.
4. Возможно, стоит явно упоминать, что во всех случаях декомпозиции, где используется соединение индексов, предполагается, что индексы фрагментированы одинаково.

Диссертация Е.В. Ивановой представляет собой законченную научно-квалификационную работу, в которой **решена задача эффективного применения колоночного представления данных для параллельной обработки сверхбольших баз данных на кластерных вычислительных системах, которая имеет существенное значение в области сверхбольших баз данных**. Диссертационная работа в полной мере отвечает требованиям Положения о порядке присуждения ученых степеней, а ее автор заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Разработчик ООО «ПОСТГРЕС
ПРОФЕССИОНАЛЬНЫЙ РАЗРАБОТКА»,
кандидат физико-математических наук _____ К.С. Пан

7 декабря 2015 г.

E-mail: kvapen@gmail.com
Тел.: (495) 150-06-91
Адрес организации: 117036, г. Москва, ул. Дмитрия Ульянова, д. 7А.

Подпись К.С. Пана заверяю:

Генеральный директор
ООО „ПпП Разработка"
Коротков А.Е. 07.12.15