ОТЗЫВ

официального оппонента, доктора технических наук Кузнецова Сергея Дмитриевича 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 темы диссертационной работы Е.В. Ивановой определяется тем, что в настоящее время перспективным является направление, связанное с параллельной обработкой сверхбольших баз данных. Для обработки больших данных необходимы высокопроизводительные вычислительные системы. В этом сегменте вычислительной техники сегодня доминируют системы с кластерной архитектурой, узлы которых оснащены многоядерными ускорителями. Недавние исследования показывают, что кластерные вычислительные системы могут эффективно использоваться для хранения и обработки сверхбольших баз данных. Однако в этой области остается целый ряд нерешенных масштабных научных задач, в первую очередь связанных с проблемой больших данных. В последние годы основным способом наращивания производительности процессоров является увеличение количества ядер. Сегодня GPU (Graphic Processing Units) и Intel MIC (Many Integrated Cores) значительно опережают традиционные процессоры в производительности по арифметическим операциям и пропускной способности памяти, позволяя использовать сотни процессорных ядер для выполнения десятков тысяч потоков. Последние исследования показывают, что многоядерные ускорители могут эффективно использоваться для обработки запросов к базам данных, хранящимся в оперативной памяти. 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. Недостатком колоночной модели представления данных является то, что в колоночных СУБД затруднено применение техники эффективной оптимизации SQL-запросов, хорошо зарекомендовавшей себя в реляционных СУБД. Кроме этого, колоночные СУБД значительно уступают строковым по производительности на запросах класса OLTP. Актуальной является задача разработки новых эффективных методов параллельной обработки сверхбольших баз данных в оперативной памяти на кластерных вычислительных системах, оснащенных многоядерными ускорителями, которые позволили бы совместить преимущества реляционной модели с колоночным представлением данных.

Целью диссертационного исследования Е.В. Ивановой является разработка и исследование 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.

Для достижения поставленной цели автором были *решены следующие задачи*:

1. На основе колоночной модели хранения информации разработаны вспомогательные структуры данных (колоночные индексы), позволяющие ускорить выполнение ресурсоемких реляционных операций.
2. Разработаны методы фрагментации (распределения) колоночных индексов, минимизирующие обмены данными между вычислительными узлами при выполнении реляционных операций.
3. На основе использования распределенных колоночных индексов разработаны методы декомпозиции основных реляционных операций, позволяющие организовать параллельное выполнение запросов без массовых пересылок данных между вычислительными узлами.
4. Реализованы предложенные модели и методы в виде колоночного сопроцессора СУБД, работающего на кластерных вычислительных системах с многоядерными ускорителями Intel Xeon Phi.

5. Проведены вычислительные эксперименты, подтверждающие эффективность предложенных подходов.

Достоверность полученных результатов подтверждена утверждениями, сформулированными в виде теорем, снабженных строгими доказательствами. Вычислительные эксперименты, проведенные в соответствии с общепринятыми стандартами, подтверждают теоретические построения.

Научная новизна проведенного исследования заключается в том, что Ивановой Е.В. разработана оригинальная доменно-колоночная модель представления данных, на основе этой модели введены колоночные индексы с доменно-интервальной фрагментацией и выполнена декомпозиция основных операций реляционной алгебры.

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. Результаты, полученные в диссертационной работе, могут быть использованы для создания колоночных сопроцессоров для других коммерческих и свободно распространяемых реляционных СУБД.

В качестве замечаний можно отметить следующие моменты.

1. В целом первые главы работы кажутся мне несколько перегруженными формализмами. В области баз данных более распространена тенденция выставлять на передний план идейную сторону работы, а формализмы оставлять тем, кто нуждается в доказательствах обычно очевидных утверждений.
2. Хотелось бы понять, насколько предлагаемая общая архитектура СУБД завязана на особенности конкретного используемого GPU. Каковы общие требования к аппаратной среде предполагаемой СУБД? Ведь странно ориентироваться на уникальную аппаратуру.
3. При описании экспериментов в конце главы 4 «Использование КСОП при выполнении SQL-запросов» следовало бы более четко специфици-

ровать условия экспериментов. Известно, что в стандартной PostgreSQL не используется колоночная модель хранения таблиц. Как в этом случае применялся колоночный индекс? Если же система хранения в PostreSQL была переделана (как, например, Greenplum), то об этом стоило бы сказать особо.

Указанные замечания не снижают общей значимости исследования, выполненного Е.В. Ивановой.

Опубликованные автором печатные работы с достаточной полнотой отражают основные результаты диссертации. Результаты исследования прошли апробацию на международных научных конференциях.

Текст автореферата соответствует содержанию диссертации.

Диссертационное исследование Е.В. Ивановой является завершенной научно-квалификационной работой, в которой предложено решение задачи эффективной параллельной обработки OLAP запросов к сверхбольшим базам данных на вычислительных кластерах, оснащенных многоядерными ускорителями, которая имеет существенное значение в области технологий баз данных. Считаю, что диссертация Е.В. Ивановой в полной мере удовлетворяет всем требованиям ВАК, предъявляемым к кандидатским диссертациям, а Е.В. Иванова заслуживает присвоения ей ученой степени кандидата физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Главный научный сотрудник
ФГБУН «Институт системного
программирования РАН»
доктор тех. наук, профессор С.Д. Кузнецов

7 декабря 2015 г.

E-mail: kuzloc@ispras.ru
Тел.: +7 (495) 912-56-59 (доб. 412)
Адрес организации: 109004, Москва, ул. Александра Солженицына, 25

Подпись С.Д. Кузнецова заверяю:
ученый секретарь ИСП РАН, к.ф.-м.н. Пакулин Н.В.