УНИВЕРСИТЕТ ИТМО

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)

Кронверкский пр-т, д. 49,
Санкт-Петербург, Россия, 197101
Тел.: (812) 232-97-04 | Факс: (812) 232-23-07
od@itmo.ru | itmo.ru

23.01.2020 № 4-25/104

УТВЕРЖДАЮ

Проректор по научной работе федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет ИТМО», доктор техн. наук, профессор

В. О. Никифоров

«23» января 2020 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет ИТМО» на диссертационную работу ЦЫМБЛЕРА Михаила Леонидовича «Интеллектуальный анализ данных в СУБД», представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Актуальность темы диссертации

Исследование, выполненное М.Л. Цымблером в рамках диссертационной работы, связано с насущной проблемой разработки эффективных методов и подходов для интеллектуального анализа больших объемов данных. Соискателем справедливо подмечен ряд важных тенденций развития современных информационных технологий: Большие данные в результате процессов очистки преобразуются в хранилища данных на основе реляционной модели данных; в области высокопроизводительных вычислений доминируют кластерные системы с многоядерными ускорителями; СУБД с открытым кодом являются надежной альтернативой коммерческим СУБД. В силу этого востребованы параллельные реляционные СУБД на вышеуказанной аппаратной платформе, способные эффективно

обрабатывать сверхбольшие хранилища и базы данных, которые могут быть построены на основе СУБД с открытым кодом. Рассматриваемый соискателем подход связан с идеей выполнения интеллектуального анализа данных внутри параллельной СУБД. Интеграция интеллектуального анализа данных является одним из перспективных направлений развития современных СУБД, поскольку позволяет как избежать существенных накладных расходов по экспорту анализируемых данных из хранилища и импорту результатов анализа обратно в хранилище, так и использовать при анализе данных системные сервисы, заложенные в архитектуре СУБД. Ввиду вышеизложенных аргументов данных тема диссертационного исследования является актуальной.

Цель, задачи и основные результаты диссертации

Цель исследования М.Л. Цымблера состояла в разработке программной платформы интеллектуального анализа данных средствами параллельной СУБД. Соискателем были решены следующие задачи для достижения этой цели:

1. Разработаны методы и алгоритмы инкапсуляции фрагментного параллелизма в последовательную реляционную СУБД с открытым кодом. На основе предложенных решений выполнено распараллеливание свободной СУБД PostgreSQL.
2. Разработаны методы и алгоритмы для интеграции интеллектуального анализа данных в параллельную СУБД для кластерной системы с многоядерными ускорителями.
3. Разработан ряд параллельных алгоритмов решения интеллектуального анализа данных (кластеризация, поиск шаблонов, поиск похожих и аномальных подпоследовательностей временных рядов) средствами параллельной реляционной СУБД.
4. Проведены вычислительные эксперименты, демонстрирующие качественные и количественные характеристики разработанных решений.

Содержание текста диссертации

Диссертация состоит из введения, пяти глав, заключения и библиографии. Объем диссертации составляет 260 страниц, объем библиографии — 274 наименования.

Во **введении** приводится обоснование актуальности темы и степень ее разработанности; формулируются цели и задачи исследования, раскрываются новизна, теоретическая и практическая значимость полученных результатов.

Первая глава посвящена обзору методов и подходов к интеллектуальному анализу данных. Даны определения задач кластеризации данных, поиска шаблонов и анализа временных рядов, рассмотрены методы интеграции интеллектуального анализа данных в СУБД и приведен обзор работ, близких к теме диссертации.

Вторая глава представляет полученные соискателем результаты по решению задач кластеризации данных и поиска шаблонов. Разработаны два новых алгоритма кластеризации для параллельной СУБД на платформе кластерной вычислительной системы: алгоритм dbParGraph для кластеризации сверхбольших графов социальных сетей и алгоритм pgFCM для нечеткой кластеризации сверхбольших данных. Разработаны схемы базы данных этих алгоритмов и SQL-запросы, реализующие вычисления. Разработан новый параллельный алгоритм PDIC поиска частых наборов для многоядерных ускорителей. В алгоритме PDIC используется битовая карта исходных данных, которая обеспечивает векторизацию вычислений. Результаты экспериментов с реальными и синтетическими данными показывают, что разработанные алгоритмы имеют высокую масштабируемость и опережают аналоги по быстродействию.

В третьей главе описаны новые параллельные алгоритмы решения задач поиска похожих подпоследовательностей и поиска аномальных подпоследовательностей во временном ряде. Алгоритм PBM выполняет поиск похожих подпоследовательностей во временном ряде на кластерной системе с вычислительными узлами на базе многоядерных ускорителей. В

алгоритме PBM предлагается ряд индексных структур для хранения данных в оперативной памяти, которые обеспечиваю эффективную векторизацию вычислений, и используется предвычисление нижних границ схожести подпоследовательностей временного ряда с поисковым запросом. Параллельный алгоритм MDD предназначен для нахождения аномальной подпоследовательности временного ряда на многоядерном ускорителе. Предложены индексные структуры данных, обеспечивающие векторизацию вычислений, и использующиеся для проверки подпоследовательностей-потенциальных аномалий. Проведены эксперименты, в которых разработанные алгоритмы показывают высокую масштабируемость и опережают аналоги по быстродействию.

Четвертая глава описывает оригинальный подход к внедрению интеллектуального анализа данных в реляционную СУБД. Определяемая пользователем SQL-функция становится в СУБД оберткой параллельного алгоритма интеллектуального анализа данных для многоядерных ускорителей, реализованного на языке C. Предложен новый параллельный алгоритм PPAM для многоядерного ускорителя, выполняющий кластеризацию данных с шумами и выбросами. Алгоритм PPAM предвычисляет матрицу евклидовых расстояний использует тайлинг циклов. Проведены эксперименты, в которых PPAM показал высокую масштабируемость и опередил аналоги по качеству кластеризации при работе с зашумленными данными.

Пятая глава описывает разработанные соискателем методы распараллеливания свободной реляционной СУБД на основе изменения ее исходных кодов. Представлен прототип параллельной СУБД PargreSQL для вычислительных систем с кластерной архитектурой, полученный как результат распараллеливания свободной СУБД PostgreSQL. Приведены результаты экспериментов, в которых PargreSQL демонстрирует высокую масштабируемость.

В **заключении** резюмируются итоги исследования, представляются отличия данной работы от ранее выполненных родственных работ других авторов, рассматриваются направления дальнейших исследований в данной области.

Обоснованность и достоверность полученных результатов

Обоснованность и достоверность полученных результатов подтверждается вычислительными экспериментами, а также сравнением полученных результатов с существующими подходами.

Научная новизна работы

Научная новизна диссертационной работы состоит в следующем:

1. Разработан новый метод инкапсуляции фрагментного параллелизма в последовательную свободную СУБД, применение которого не требует масштабных изменений в исходном коде. Разработаны новые параллельные алгоритмы для интеллектуального анализа сверхбольших данных в указанной параллельной СУБД.
2. Разработаны параллельные алгоритмы интеллектуального анализа данных для многоядерных ускорителей: поиск похожих и аномальных подпоследовательностей временного ряда, кластеризация данных с шумами и выбросами, поиск частых наборов.
3. Разработан метод интеграции интеллектуального анализа данных в реляционную СУБД, предполагающий инкапсуляцию параллельных алгоритмов анализа данных для многоядерных процессоров.

Теоретическая и практическая ценность работы

Теоретическая значимость исследования М.Л. Цымблера состоит в следующем. Предложенные соискателем методы и подходы обеспечивают интеграцию в последовательные реляционные СУБД параллельных обработки и интеллектуального анализа данных. Разработанные соискателем параллельные алгоритмы для многоядерных ускорителей обеспечивают решение задач интеллектуального анализа данных с ускорением, близким к линейному. Практическая значимость исследования М.Л. Цымблера состоит

в том, что на основе применения предложенных методов и подходов к свободной СУБД PostgreSQL разработан прототип параллельной СУБД PargreSQL.

Публикации и апробации

Основные результаты, полученные М.Л. Цымблером в рамках диссертационного исследования, достаточно полно опубликованы в авторитетных рецензируемых научных изданиях: 12 статей в журналах Перечня ВАК и 9 статей в изданиях баз данных Scopus и Web of Science. По результатам исследования соискателем сделано 17 докладов на международных и всероссийских научных конференциях.

Автореферат и текст диссертации

Текст диссертации М.Л. Цымблера характеризуется ясным научным стилем изложения материала, высоким уровнем математической культуры и оформлен в соответствии с требованиями Минобрнауки. Автореферат диссертации в полной мере отражает содержание диссертации.

Соответствие содержания диссертации паспорту научной специальности

Содержание и полученные результаты диссертации соответствуют паспорту специальности 05.13.11 по следующим областям:

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.

Рекомендации по использованию результатов диссертации

Результаты диссертации могут быть использованы в разработке программных комплексов, обеспечивающих обработку и интеллектуальный анализ больших объемов реляционных данных в параллельной СУБД для кластерных систем с многоядерными ускорителями.

Замечания по диссертации

В качестве замечаний к диссертации необходимо отметить следующее:

1) Недостаточно полно выполнен аналитический обзор современных решений в данной области. Из приведенных в диссертации сведений не ясно, зачем (кроме решения задачи импортозамещения) потребовалась разработка собственной версии MPP-подобного решения, хотя на рынке существуют такие продукты, как greenplum (на данный момент это СУБД с открытым кодом и лицензией Apache 2.0) и отечественный ClickHouse. Кроме того, в анализе не отражена общемировая тенденция развития нереляционных СУБД, как наиболее известных продуктов для работы с большими данными (HBase, Cassandra, Apache Kudu, и др.), которые и могли бы стать основным потребителем предложенных алгоритмов.

2) Сравнение разработанных алгоритмов с аналогами проведено косвенными методами. Диссертант не проводил реализацию самих алгоритмов-аналогов в реляционной СУБД, а использовал оценки времени, приведенные в публикациях, масштабируемые в соответствии с пиковой производительностью вычислительной системы. Как следствие, такой подход не позволяет однозначно определить источник ускорения алгоритмов, предложенных диссертантом. Не ясно, является ли ускорение алгоритмическим, или основной эффект связан с "перемещением вычислений к данным" в рамках самой СУБД?

3) Для ряда алгоритмов (например, кластеризации) отсутствует сравнение с алгоритмами-аналогами по качеству полученного решения. В этой ситуации уменьшение времени выполнения алгоритма не может считаться обоснованным показателем превосходства.

4) Диссертант акцентирует изложение актуальности именно на реляционных СУБД и важности их использования. Однако при изложении приведенных в диссертации алгоритмах и для представления данных самих примеров реляционная модель ***не*** используется (т.е. нет отношений типа foreign key и т.п.).

Заключение

Диссертационная работа М.Л. Цымблера представляет собой законченную научно-квалификационную работу, которая в полной мере соответствует требованиям Положения о порядке присуждения ученых степеней, в том числе п. 9. М.Л. Цымблером разработан комплекс новых параллельных алгоритмов для интеллектуального анализа данных в рамках параллельной СУБД на платформе кластерных вычислительных систем с многоядерными ускорителями, а также предложил метод инкапсуляции параллелизма в последовательные свободные СУБД, что в совокупности можно квалифицировать как научное достижение, и соискатель заслуживает присуждения ученой степени доктора 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 заседании научного семинара Научно-исследовательского института наукоемких компьютерных технологий Университета ИТМО, протокол №1 от 17.01.2020.

Директор научно-исследовательского
института наукоемких компьютерных
технологий Университета ИТМО,
д.т.н., профессор А.В. Бухановский

Наименование организации, предоставившей отзыв:
федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»
Адрес организации: 197101, Санкт-Петербург, Кронверкский пр., дом 49
Телефон: +7 (812) 232-97-04
Email: od@mail.ifmo.ru
WWW: http://www.ifmo.ru/

СВЕДЕНИЯ О ЛИЦАХ, УТВЕРДИВШИХ И ПОДГОТОВИВШИХ ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию Цымблера М.Л. «Интеллектуальный анализ данных в СУБД», представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Наименование организации	федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»	
Ведомственная принадлежность	Министерства науки и высшего образования Российской Федерации	
Почтовый адрес, телефон, адрес электронной почты, адрес официального сайта в сети «Интернет».	197101, Санкт-Петербург, Кронверкский пр., дом 49 +7 (812) 232-97-04 od@mail.ifmo.ru http://www.ifmo.ru/	
Сведения о лице, утвердившем отзыв	**ФИО**	Никифоров Владимир Олегович
	Ученая степень (с указанием шифра специальности)	Д.т.н., 05.13.01
	Должность	Проректор по научной работе
Сведения о лице, подготовившем отзыв	**ФИО**	Бухановский Александр Валерьевич
	Ученая степень (с указанием шифра специальности)	Д.т.н., 05.11.16
	Должность	Директор научно-исследовательского института наукоемких компьютерных технологий Университета ИТМО

Список основных работ сотрудников ведущей организации по теме диссертации в рецензируемых научных изданиях

1. Абухай Т.М., Ковальчук С.В., Балахонцева М.А., Бухановский А.В. Моделирование, анализ и прогнозирование процессов оказания кардиологической помощи в стационаре // Известия высших учебных заведений. Приборостроение. 2018. Т. 61. № 8. С. 730-733.
2. Baimuratov I.R., Zhukova N.A. An approach to clustering models estimation. 22nd Conference of Open Innovations Association, FRUCT 2018, Jyvaskyla, Finland; 15-18 May 2018. vol. 2018-May, art. no. 8468286, pp. 19-24.
3. Bochenina K., Kesarev S., Boukhanovsky A. Scalable parallel simulation of dynamical processes on large stochastic Kronecker graphs. Future Generation Computer Systems, 2018. vol. 78, pp. 502-515.
4. Boukhanovsky A.V., Krzhizhanovskaya V.V., Bubak M. Urgent computing for decision support in critical situations. Future Generation Computer Systems, 2018. vol. 79, pp. 111-113.
5. Kovalchuk S.V., Krotov E., Smirnov P.A., Nasonov D.A., Yakovlev A.N. Distributed data-driven platform for urgent decision making in cardiological ambulance control. Future Generation Computer Systems. 2018. vol. 79, pp. 144-154
6. Severiukhina O., Bochenina K., Kesarev S., Boukhanovsky A. Parallel data-driven modeling of information spread in social networks. 18th International Conference on Computational Science, ICCS 2018, Wuxi, China, 11-13 June 2018. 2018. Lecture Notes in Computer Science, vol. 10860, pp. 247-259.
7. Kotenko I., Saenko I., Kushnerevich A. Parallel big data processing system for security monitoring in internet of things networks. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2017. vol. 8, no. 4, pp. 60-74.
8. Presbitero A., Quax R., Krzhizhanovskaya V., Sloot P. Anomaly Detection in Clinical Data of Patients Undergoing Heart Surgery. Procedia Computer Science. 2017. vol. 108, pp. 99-108.
9. Bochenina K., Kesarev S. A parallel algorithm for modeling of dynamical processes on large stochastic Kronecker graphs. Procedia Computer Science. 2016. vol. 80, pp. 2413-2417.
10. Khoruzhnikov S.E., Grudinin V.A., Sadov O.L., Shevel A.Y., Kairkanov A.B. Preliminary study of Big Data transfer over computer network // Компьютерные исследования и моделирование. 2015. Т. 7. № 3. С. 421-427.
11. Khoruzhnikov S.E., Grudinin V.A., Sadov O.L., Shevel A.Y. Kairkanov A.B. Transfer of Large Volume Data over Internet with Parallel Data Links and

SDN. Advances in Swarm and Computational Intelligence, 6th International Conference, ICSI 2015, in conjunction with the Second BRICS Congress, CCI 2015, Beijing, China, June 25-28, 2015, Proceedings. 2015. Lecture Notes in Computer Science, vol. 9142, pp. 463-471.

12. Иванов С.В., Бухановский А.В. Анализ неопределенности предсказательного моделирования сложных систем: усвоение данных и ансамблевые технологии // Известия высших учебных заведений. Приборостроение. 2013. Т. 56. № 12. С. 66-68.

Ученый секретарь Ученого совета
Университета ИТМО, к.т.н. М.Я. Марусина