ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

члена-корреспондента РАН, доктора физико-математических наук, профессора Воеводина Владимира Валентиновича 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 диссертационной работы М.Л. Цымблера легко объясняется жизнью, для чего достаточно проанализировать современные тенденции в сфере информационных технологий. Появляются новые парадигмы обработки данных, но несмотря на широкое распространение NoSQL-технологий, системы управления базами данных (СУБД) на основе реляционной модели остаются основным инструментом управления базами данных. Это важно и на сегодня исключительно актуально, поскольку выполнение функций интеллектуального анализа данных внутри реляционной СУБД позволяет исключить накладные расходы по экспорту данных и импорту результатов анализа, а также задействовать при анализе данных системные сервисы СУБД. Кроме того, сделанный в работе акцент на разработку решений на основе свободной СУБД, развитие отечественных СУБД и систем анализа данных способствуют обеспечению технологической независимости Российской Федерации в области системного программного обеспечения для обработки данных.

Целью диссертационной работы М.Л. Цымблера являлась разработка комплекса масштабируемых методов и параллельных алгоритмов для создания программной платформы интеллектуального анализа данных средствами свободной СУБД. Для достижения указанной цели М.Л. Цымблером были решены следующие *задачи*:

1. Разработаны методы и алгоритмы для внедрения фрагментного параллелизма в свободную последовательную реляционную СУБД. Проверка эффективности предложенных решений на СУБД PostgreSQL.

2. Разработаны методы и алгоритмы для внедрения интеллектуального анализа данных в параллельную СУБД для современных многопроцессорных платформ с многоядерными ускорителями.
3. Разработан комплекс параллельных алгоритмов решения задач кластеризации, поиска шаблонов и анализа временных рядов средствами параллельной реляционной СУБД:
   1) алгоритм кластеризации графа для параллельной СУБД на основе фрагментного параллелизма;
   2) алгоритм нечеткой кластеризации данных для параллельной СУБД на основе фрагментного параллелизма;
   3) параллельный алгоритм поиска похожих подпоследовательностей временного ряда для вычислительного кластера с многоядерными ускорителями;
   4) параллельный алгоритм поиска аномальных подпоследовательностей временного ряда для многоядерного ускорителя;
   5) параллельный алгоритм кластеризации данных для многоядерного ускорителя на основе техники медоидов;
   6) параллельный алгоритм поиска частых наборов для многоядерного ускорителя на основе битовых масок.
4. Проведены вычислительные эксперименты с синтетическими и реальными данными, подтвердившие эффективность предложенных методов и алгоритмов, а также их преимущество перед ведущими аналогами.

Обоснованность и достоверность результатов, полученных в рамках диссертационной работы, подтверждены вычислительными экспериментами, выполненными в соответствии с общепринятыми стандартами, и сравнением с существующими подходами.

Научная новизна диссертационной работы М.Л. Цымблера состоит в следующем:

1. Разработан оригинальный метод интеграции фрагментного параллелизма в последовательную свободную СУБД, применение которого не требует масштабных изменений в исходном коде. Разработаны новые параллельные алгоритмы интеллектуального анализа данных сверхбольших объемов для указанной параллельной СУБД.
2. Разработаны новые параллельные алгоритмы интеллектуального анализа данных для многоядерных ускорителей (поиск похожих и аномальных подпоследовательностей временного ряда, кластеризация данных, поиск частых наборов).
3. Разработан оригинальный метод интеграции интеллектуального анализа данных в реляционную СУБД на основе пользовательских функций, инкапсулирующих параллельные алгоритмы для современных многоядерных процессоров.

Теоретическая ценность диссертационной работы состоит в том, что в ней предложены методы и архитектурные решения, обеспечивающие инкапсуляцию параллелизма в последовательные СУБД, и разработаны параллельные алгоритмы решения ряда задач интеллектуального анализа данных для современных многоядерных ускорителей, обеспечивающие ускорение, близкое к линейному.

Практическая ценность диссертационной работы заключается в том, что предложенные методы и подходы применены к свободной СУБД PostgreSQL и разработан прототип параллельной СУБД PargreSQL. Результаты, полученные в исследовании, могут быть использованы в разработке программных систем, ориентированных на параллельную обработку и анализ больших объемов данных с использованием свободной реляционной СУБД.

Публикации и апробации. Основные результаты диссертационного исследования достаточно полно опубликованы в авторитетных рецензируемых научных изданиях, в т.ч. в журналах Перечня ВАК

(12 статей) и изданиях, индексируемых в библиографических базах данных Scopus и Web of Science (9 статей). Результаты исследования прошли апробацию на международных и всероссийских научных конференциях (17 докладов).

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.

В качестве *замечаний к диссертационной работе*, не снижающих ее общей значимости и высокого научного уровня, необходимо отметить следующее.

1. В некоторых местах диссертационной работы автор сопровождает используемые понятия эпитетами, которые не всегда четко определены и носят, скорее, эмоциональный характер. По ходу изложения они интуитивно понятны, но лучше было бы обговорить их значения. В частности, в разделе 2.1 автор пишет: “При обработке графов сверхбольших размеров…”, где по какой-то причине использует не просто слово “больших”, а “сверхбольших”.
2. Для проведения сравнения эффективности разработанных в диссертации методов интеллектуального анализа данных автор подбирает конфигурацию доступного кластера, сравнимого по производительности с теми системами, которые использовали исследователи в референсных работах. Вместе с этим, не всегда понятно, насколько такое сравнение корректно, поскольку архитектура используемых систем может быть различной, внося дополнительные коррективы в вычислительные эксперименты.
3. Весьма важный вопрос, который сильно влияет на время и эффективность обработки – это используемые структуры данных. Типичным примером являются графовые алгоритмы. В работе чаще исследуется “параллельная эффективность”, а анализу возможных

вариантов используемых структур данных уделяется гораздо меньше внимания.

4. Автор в работе приводит много сравнений эффективности разработанных им методов с существующими алгоритмами и реализациями. Однако далеко не всегда по тексту диссертационной работы понятно, являются ли выбранные для сравнения методы и данные “лучшими в мире“ или они “одни из множества существующих”. Интуитивно понятно, что сравнение ведется с лучшими имеющимися примерами, но хорошо бы это оговорить явно.
5. В работе нет четкого описания модели кластерной вычислительной системы, на которую ориентируется автор в своих исследованиях. На стр.9 автор говорит про ориентацию на многопроцессорные системы с многоядерными ускорителями, а на стр.144 упоминаются лишь “многоядерные процессоры”. Должны ли в состав кластеров входить ускорители? Важна ли автору иерархия подсистемы памяти, развитая на современных параллельных платформах? Описанию модели кластерной вычислительной системы следовало бы в начале уделить больше внимания.
6. В экспериментальном исследовании разработанных алгоритмов кластеризации dbParGraph и pgFCM (глава 2 диссертации) автор ограничивается сравнением быстродействия алгоритмов с аналогами, но не исследует масштабируемость, что было бы желательно в контексте данной работы.

Заключение. Считаю, что диссертационная работа М.Л. Цымблера «Интеллектуальный анализ данных в СУБД» представляет собой законченную научно-квалификационную работу, в которой предложены и обоснованы подходы к формированию программной платформы интеллектуального анализа данных средствами свободной СУБД на основе комплекса новых параллельных алгоритмов интеллектуального анализа данных, ориентированных на свободные СУБД на кластерных

вычислительных системах с многоядерными ускорителями, что можно квалифицировать как научное достижение. Диссертационная работа М.Л. Цымблера в полной мере отвечает требованиям Положения о порядке присуждения ученых степеней, включая п. 9, а ее автор заслуживает присуждения ученой степени доктора физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Официальный оппонент:

Владимир Валентинович Воеводин
15 января 2020 г.

Член-корреспондент РАН, доктор физ.-мат. наук, профессор,
директор Научно-исследовательского вычислительного центра Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова», Москва
Адрес организации: 119234, Москва, Ленинские Горы, дом 1, стр. 4
Телефон: +7 (495) 939-5424
Email: voevodin@parallel.ru

Подпись Вл.В.Воеводина заверяю,
Ученый секретарь НИВЦ МГУ В.В.Суворов

СВЕДЕНИЯ ОБ ОФИЦИАЛЬНОМ ОППОНЕНТЕ

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

Фамилия, имя, отчество	Воеводин Владимир Валентинович
Ученая степень (с указанием номера и шифра специальности)	Доктор физико-математических наук, 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей
Ученое звание	член-корреспондент РАН, профессор
Организация основного места работы	Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова»
Ведомственная принадлежность	Правительство Российской Федерации
Занимаемая должность	директор Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова
Почтовый адрес	119234, Москва, Ленинские Горы, д. 1, стр. 4
Телефон	+7 (495) 939-5424
Адрес электронной почты	voevodin@parallel.ru

Список основных публикаций по теме диссертации в рецензируемых научных изданиях

1. Chernyavskiy A.Y., Voevodin, Vad.V., Voevodin, Vl.V. Parallel Computational Structure of Noisy Quantum Circuits Simulation. 2018. Lobachevskii Journal of Mathematics, vol. 39, no. 4, pp. 494-502.
2. Afanasyev I., Voevodin V. The Comparison of Large-Scale Graph Processing Algorithms Implementation Methods for Intel KNL and NVIDIA GPU. 3rd Russian Supercomputing Days Conference, RuSCDays 2017, Moscow, Russian Federation, 25-26 September 2017. Communications in Computer and Information Science, vol. 793, 2017, pp. 80-94.
3. Voevodin Vl., Voevodin Vad., Shaikhislamov D., Nikitenko D. Data mining method for anomaly detection in the supercomputer task flow. 2nd International Conference on Numerical Computations: Theory and

Algorithms, NUMTA 2016, Pizzo Calabro, Italy, 19-25 June 2016. AIP Conference Proceedings, vol. 1776, art. no. 090015.

4. Никитенко Д.А., Воеводин В.В., Жуматий С.А. OctoShell: система для администрирования больших суперкомпьютерных комплексов // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2016. Т. 5. № 3. С. 76-95.
5. Nikitenko D.A., Voevodin Vl.V., Teplov A.M., Zhumatiy S.A., Voevodin Vad.V., Stefanov K.S., Shvets P.A. Supercomputer application integral characteristics analysis for the whole queued job collection of large-scale HPC systems // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2016. Т. 5. № 4. С. 32-45.
6. Воеводин Вл.В. Открытая энциклопедия свойств алгоритмов AlgoWiki: от мобильных платформ до экзафлопсных суперкомпьютерных систем // Вычислительные методы и программирование: новые вычислительные технологии. 2015. Т. 16. № 1. С. 99-111.
7. Stefanov K., Voevodin Vl. Distributed modular monitoring (DiMMon) approach to supercomputer monitoring. Proceedings - IEEE International Conference on Cluster Computing, ICCC 2015, Chicago, United States, 8-11 September 2015. vol. 2015, pp. 502-503.

/ Воеводин В.В. /

Подпись Вл.В.Воеводина заверяю,
Ученый секретарь НИВЦ МГУ В.В.Суворов