Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК

Малич
Константин Васильевич

Методы и алгоритмы обработки больших объёмов данных при проектировании высоконагруженных систем

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

подпись магистранта

Научный руководитель
Куликов С. С.
к.т.н., доцент

подпись научного руководителя

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

Под большими данными понимается широкое разнообразие массивов данных, которые не могут быть надлежащим образом обработаны традиционными приложениями из-за своего огромного объема или сложного состава. Сложность анализа больших данных заключается в специфике их сбора, курирования, разделения, хранения, передачи, визуализации и сохранении конфиденциальности информации. Под анализом больших данных часто понимается применение прогнозной аналитики или других передовых методов с целью извлечения из множества данных определенной полезной информации. Точность при анализе больших данных помогает принимать более рациональные решения. В свою очередь, принятие наилучших решений позволяет увеличить производственную эффективность, сократить расходы и снизить риски.

Анализ больших данных может применяться в таких областях, как отслеживание конъюнктуры рынка, предотвращение распространения эпидемий и борьба с преступностью. Ученые, руководители крупнейших компаний, работники сферы масс-медиа и рекламы, а также правительственные органы часто сталкиваются с трудностями при анализе массивов данных огромных объемов в таких областях, как поиск в сети интернет, информационные технологии в сфере бизнеса и финансов и т.п. Работа ученых, особенно метеорологов, медиков, изучающих геномы, исследователей, работающих в области изучения средств коммуникации, 1 физиков, создающих сложные симуляторы, а также биологов и экологов часто ограничивается возможностями обработки огромных массивов данных.

На основании вышеизложенного можно выделить актуальную проблему обработки больших объемов данных. Для такой обработки необходимо проектировать архитектуру систем таким образом, чтобы она могла справиться с нарастающей нагрузкой. Также для обработки данных необходимо разработать алгоритмы, позволяющие эффективно использовать вычислительные мощности для обработки таких данных.

Диссертационная работа посвящена методам и алгоритмам обработки больших объемов данных. Возможность использования таких алгоритмов позволит эффективно работать с нарастающими объемами данных.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является анализ существующих методов и алгоритмы обработки больших объемов данных. Выделение достоинств и недостатков каждого из них. Обзов подходов к проектированию архитектуры высоконагруженных систем. Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать наиболее популярные методы и алгоритмы обработки больших объемов данных.
2. Рассмотреть наиболее популярные архитектурные решения, применяемые при проектировании высоконагруженных систем.
3. Реализовать различные алгоритмы обработки больших объёмов данных.
4. Провести нагрузочное тестирование и сравнить результаты для различных алгоритмов.

Объектом исследования являются системы, работающие с большими объемами данных.

Предметом исследования являются методы и алгоритмы обработки больших объемов данных. Подходы к проектированию архитектуры высоконагруженных систем.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность построения архитектуры систем таким образом, чтобы избежать сбоев в работе ввиду обработки больших объемов данных. Возможность проектирования таких систем является решением большого количества задач, с которыми сталкиваются существующие программные продукты.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «Разработка моделей, методов, алгоритмов, повышающих показатели проектирования, внедрения и эксплуатации программных средств для перспективных платформ обработки информации, решения интеллектуальных задач, работы с большими массивами данных и внедрение в современные обучающие комплексы» (ГБ № 16-2004, № ГР 20163588, научный руководитель НИР – Н. В. Лапицкая).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя С.С. Куликова, заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

По теме диссертации опубликовано 1 печатная работа в сборнике материалов IX международной научно-методической конференции.

По теме диссертации опубликовано 1 печатная работа в сборнике материалов 54-й научно-технической конференции аспирантов, магистрантов и студентов БГУИР.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения. Вторая глава посвящена разработке архитектуры ПО и алгоритмов для высоконагруженных систем. В третьей главе предложены алгоритмы, позволяющие обрабатывать большие данные. В четвертой главе проведено нагрузочное тестирование разработанных алгоритмов, а также сравнительный анализ результатов тестирования.

Общий объем работы составляет 67 страниц, из которых основного текста – 45 страниц, 29 рисунков на 10 страницах, список использованных источников из 30 наименований на 3 страницах и 2 приложения на 9 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** проведен анализ методов и алгоритмов обработки больших объемов данных. Приведены различные подходы и технологии, используемые для обработки больших объемов данных.

В настоящее время большие данные являются важным источником информации. К большим данным можно отнести любую информацию с любых информационных источников. Для того чтобы использовать большие данные необходимо их собрать, затем структурировать для последующей обработки.

Для сбора данных используется широкий спектр различных подходов, затрагивающий различные области технологий. Основным способом сбора данных является Data Mining.

Вторая глава посвящена особенностям проектирования архитектуры ПО, работающего с большими данными.

Для всех систем работы с Большими Данными характерны четыре общих требования, которые нужно учитывать при проектировании и которые в совокупности вынуждают существенно отклониться от архитектуры

традиционных бизнес-систем, имеющих ограничения на рост объема данных и функциональности.

Данная глава содержит описание архитектурных подходов, применяемых в проектировании высоконагруженных систем. Также в данной главе представлены архитектурные решения, позволяющие эффективно работать с большими данными.

В **третьей главе** предложены реализации различных алгоритмов, позволяющих обрабатывать большие объемы данных. Предложнные алгоритмы были разработаны на языке C# с использованием стандартных библиотек и возможностей языка. Был сделан вывод о том, что оптимальным способом обработки больших объемов данных является распалаллеливание вычислений по потокам. Однако при этом стоит помнить о проблемах синхронизации доступа к общему хранинщу.

В **четвертой главе** представлены результаты нагрузочного тестирования разработанных алгоритмов. Классы Parallel, и PLINQ могут эффективно распределять задачи и порождать новые потоки выполнения на всех доступных ядрах процессора. Изменение библиотеки или даже различных перегрузок одной и той же функции может помочь добиться более высокой производительности и загрузки ЦП. Для синхронизации доступа к общему источнику данных лучше использовать ConcurrentDictionary.

Также был проведен сравнительный анализ разработанных алгоритмов и выделены наиболее эффективные алгоритмы для решения поставленной задачи.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Рассмотрены и систематизированы методы и алгоритмы обработки больших объемов информации. По каждому из рассматриваемых методов дана характеристика и возможность применимости для решения различных практических задач, связанных с работой с большими данными.
2. Рассмотрены различные подходы к проектированию архитектуры высоконагруженных систем. Рассмотрены основные проблемные места и типовые задачи, с которые решаются при проектировании таких систем.
3. Реализованы различные алгоритмы обработки больших объемов данных на языке C# с использованием различных подходов и шаблонов для работы с большими данными. Написанный код использует только стандартные классы и библиотеки языка C#.
4. Проведено нагрузочное тестирование разработанных алгоритмов. На основании тестирования были предложены 3 реализации, продемонстрировавшие наилучшие результаты тестирования. Данные алгоритмы могут быть использованы для задач обработки больших объемов данных.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки ПО компьютерных систем для решения задач проектирования систем, работающих с большими данными. Они могут быть использованы как для изменения существующих модулей систем, так и для разработки новых.

2. Разработанные методы и алгоритмы могут применяться в различных типах решений.

3. Реезультаты проектирования архитектур высоконагруженных систем могут быть использованы системными архитекторами и разработчиками, которые занимаются проектированием систем, которые впоследствии будут работать с большими данными.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Малич, К.В. DevOps инженеры как перспективное направление обучения / К.В. Малич, С.С. Куликов // Высшее техническое образование: проблемы и пути развития: материалы IX Междунар. науч.-метод. конф. БГУИР, Минск, 1-2 нояб. 2018. – Минск, 2018. – С. 289-291.

2.Малич, К. В. Балансирование нагрузки в web приложениях / К. В. Малич, С.С. Куликов (науч. рук.) // Компьютерные системы и сети: материалы 54-й научной конференции аспирантов, магистрантов и студентов, Минск, 23 – 27 апреля 2018 г. / БГУИР. – Минск, 2018. – С. 89 – 90.