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

УДК 004.62

Мацокин
Николай Павлович

Использование инструментов на основе параллельных вычислений для обработки и бизнес-аналитики больших данных в области продаж

АВТОРЕФЕРАТ
на соискание академической степени
магистра

по специальности 7-06-0611-07 – Бизнес-аналитика и цифровой маркетинг

Научный руководитель
Логинова И. П.
к.т.н., доцент

Минск 2025

Работа выполнена на кафедре экономической информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

Научный руководитель:	**ЛОГИНОВА Ирина Петровна,** доцент кафедры экономической информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», кандидат технических наук
Рецензент:	**НИКУЛЬШИН Борис Викторович,** заведующий кафедрой электронных вычислительных машин учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», кандидат технических наук, доцент

Защита диссертации состоится «23» июня 2025 г. года в $9^{\underline{00}}$ часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220005, Минск, ул. Платонова, 39, корп. 5, ауд. 209, тел. 293-89-92, E-mail: kafei@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

В современном мире объем данных, генерируемых бизнесом, неуклонно растет. Компании, работающие в сфере продаж, собирают огромные массивы информации, начиная от данных о транзакциях, поведении клиентов, до маркетинговых кампаний и логистических операций. Эти данные, в совокупности называемые большими данными (Big Data), содержат важную информацию, которая при грамотной обработке может существенно улучшить эффективность бизнеса, повысить точность прогнозирования и способствовать принятию более обоснованных решений. Однако традиционные методы обработки данных часто оказываются недостаточно эффективными для работы с такими большими объемами информации. В этом контексте использование параллельных вычислений и инструментов для работы с Big Data становится ключевым фактором успеха.

Параллельные вычисления представляют собой подход к обработке данных, при котором задача делится на несколько подзадач, которые могут выполняться одновременно на различных процессорах или вычислительных узлах. Это позволяет существенно ускорить обработку больших массивов данных и повысить эффективность аналитических процессов. Эти технологии позволяют компаниям не только справляться с большими объемами данных, но и оперативно анализировать информацию, выявляя закономерности и тренды, которые могут быть использованы для улучшения продаж и маркетинговых стратегий.

Одна из ключевых задач бизнес-аналитики в сфере продаж заключается в том, чтобы превратить собранные данные в полезные инсайты, которые могут быть использованы для оптимизации процессов. Инструменты на основе параллельных вычислений играют решающую роль в этом процессе, позволяя анализировать огромные объемы информации с минимальными затратами времени. Это включает обработку как структурированных данных, таких как транзакции или данные о клиентах, так и неструктурированных данных, таких как отзывы клиентов, посты в социальных сетях и другие источники данных, которые могут влиять на бизнес. Параллельная обработка данных дает возможность бизнесу оперативно реагировать на изменения в поведении клиентов, рынок или внутренние бизнес-процессы, что существенно повышает их конкурентоспособность.

Таким образом, использование инструментов на основе параллельных вычислений для обработки и бизнес-аналитики больших данных становится необходимым условием успеха в современной сфере продаж. Компании, которые способны эффективно управлять и анализировать большие объемы данных с использованием параллельных вычислений, получают значительные конкурентные преимущества, такие как более точные прогнозы, и оптимизация бизнес-процессов. В данной работе будут рассмотрены основные принципы работы с инструментами параллельных вычислений, их преимущества и примеры использования в области бизнес-аналитики больших данных для повышения эффективности продаж.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Применение информационных технологий в условиях современных конкурирующих мировых рынков уже десятки лет позволяет компаниям в разы повысить эффективность собственных бизнес-процессов в аналитике. Использование технологий больших данных (Big Data) также обеспечило внедрение качественно улучшенных методов хранения, обработки и использования информации как важнейшего экономического ресурса. Сама концепция Big Data известна лишь чуть больше десятка лет, но экспоненциально растущие объемы, скорость и неоднородность данных показали потребность в оптимизации процессов работы с данными. Одним из решений является применение параллельных вычислений.

Цель и задачи исследования

Целью работы является обоснование эффективности использования высокопроизводительных параллельных вычислений для анализа больших данных в области продаж.

Поставленная цель работы определяет **следующие основные задачи:**

1. Исследовать рынок электронной коммерции, концепции больших данных и параллельных технологий.
2. Изучить существующие методы и средства аналитики в области продаж на основе больших данных, а также технологии, предназначенные для создания приложений для обработки больших данных с использованием параллельных вычислений.
3. Разработать и провести тестирование программного средства для оценки эффективности использования параллельных технологий при анализе больших данных в области продаж.

Область исследования

Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры) ОСВО 7-06-0611-07-2023 специальности 7-06-0611-07 Бизнес-аналитика и цифровой маркетинг.

Теоретическая и методологическая основа исследования

В основу диссертации легли работы белорусских и зарубежных ученых в области анализа больших данных и технологий параллельных вычислений, а также анализ текущей экономической ситуации электронной коммерции в Республике Беларусь.

Информационная база исследования сформирована на основе литературы, открытой информации, технических нормативно-правовых актов, сведений из электронных ресурсов, а также материалов научных конференций и семинаров.

Научная новизна, теоретическая и практическая значимость

Научная новизна и значимость полученных результатов работы заключается в разработке методики оценки использования средств на основе параллельных технологий для анализа больших данных в области продаж.

Теоретическая значимость работы заключается в детальном описании процессов анализа больших данных и использования параллельных технологий.

Практическая значимость диссертации состоит в разработанном программном средстве оценке эффективности использования параллельных вычислений больших данных, позволяющая оценить приоритет той или иной технологии, а также ускорить процесс анализа больших данных.

Основные положения, выносимые на защиту

1. Обзор рынка электронной коммерции в Республики Беларусь, актуальность применения IT-технологий в сфере продаж и использование больших данных в бизнес-аналитике.

2. Методы, модели и средства бизнес-анализа больших данных, использование параллельных технологий для анализа больших данных с применением языка Python и его модулей.

3. Разработанное приложение для оценки использования параллельных технологий для анализа больших данных, результаты тестирования и обоснование использование конкретной технологии под конкретные задачи.

Апробация диссертации и информация об использовании ее результатов

Результаты исследований, вошедшие в диссертацию, докладывались и обсуждались на 61-ой научно-технической конференции аспирантов, магистрантов и студентов БГУИР (г. Минск, Беларусь, 2025 год).

Отдельные положения диссертации могут быть использованы при преподавании дисциплин «Технологии параллельного программирования».

Публикации

Общий объем публикаций по теме диссертации составляет 3 страницы.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, трех глав с краткими выводами по каждой главе, заключения, библиографического списка и приложений.

В первой главе проведен обзор современного состояния рынка электронной коммерции Республики Беларусь, а также рассмотрена используемость IT-технологий в белорусском бизнесе. Были пояснены концепции больших данных и технологий параллельных вычислений.

Во второй главе рассмотрены методы, модели и технологии бизнес-аналитики, технологии параллельных вычислений, а также приведено обоснование выбора языка Python и его модулей в качестве основы для создания приложения с целью исследования влияния параллельных вычислений на бизнес-аналитику больших данных.

В третьей главе описана разработка приложения на языке Python с целью оценки эффективности работы модулей с использованием технологий параллельных вычислений в сравнении с оригинальным ПО. Была подтверждена адекватность выбора модулей, также было доказано полезность их применение в отношении анализа больших данных.

В приложении представлен листинг кода и графический материал диссертации.

Общий объем диссертационной работы составляет 66 страницы. Из них 13 иллюстраций на 12 страницах, 4 таблицы на 4 страницах, библиографический список из 38 наименований на 3 страницах, список собственных публикаций соискателя из 1 наименования на 1 странице, 3 приложения на 19 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе анализируются роли IT в бизнес-аналитике: рассматриваются роли BI-платформы, автоматизации процессов и внедрение ИИ. Делается акцент на важности поведенческого подхода к анализу информации и роли человека в интерпретации аналитических инструментов. Это подводит к пониманию того, что технологии — лишь часть решения, и важно учитывать социальные и когнитивные аспекты.

Затем освещается текущее состояние электронной торговли в Беларуси. Представлены данные по объёмам, темпам роста и каналам продвижения, с акцентом на общий объём получаемых в цифровой деятельности данных. В параграфе о Big Data объясняется основы концепции, приводятся примеры использования в экономике, государстве и медицине. Раздел о параллельных вычислениях подробно описывает виды параллелизма (по данным, задачам, потокам), а также ключевые проблемы (гонка данных, взаимоблокировка и т.д.) и примеры использования (смартфоны, блокчейн, суперкомпьютеры).

Во второй главе анализируются ключевые методы и технологии, применяемые в бизнес-аналитике. Рассматриваются алгоритмы кластеризации, такие как k-средних, Apriori и DBSCAN, а также модель MapReduce — как фундамент распределённых вычислений. Приводятся примеры её практического применения (включая модификации вроде ReduceJoin) и оцениваются достоинства и ограничения.

Отдельное внимание уделяется облачным платформам, включая Google BigQuery, Weka и другие. Описываются особенности их архитектур, форматы работы с данными, возможности расширения и недостатки. Также детально рассматриваются Python и его библиотеки для бизнес-аналитики (Pandas, Matplotlib) с использованием параллельных вычислений (Dask, Modin и Swifter), с разъяснением, как они работают. Глава закладывает теоретическую основу для практической части, обосновывая выбор Python и его модулей для демонстрации использования параллельных вычислений.

В третьей главе описывается экспериментальное исследование, в котором разрабатывается приложения проверки эффективности работы модулей Dask, Modin и Swifter в сравнении с оригинальным Pandas. С разработкой также проводится тестирование с использованием реального датасета больших данных. Рассматриваются типовые операции: чтение, преобразование строк, группировка, фильтрация и нормализация. Были обоснованы и объяснена общая эффективность использования технология и нюансы связанные с изменением размерности данных.

Результаты показали, что технологии параллельных вычислений в действительности могут помочь в ускорении обработки больших данных. Однако с увеличением объема данных эффективность модулей менялась, и ни одна не показала универсального преимущества. Делается вывод, что выбор модуля должен зависеть от типа задачи и технических ресурсов. Эксперимент подтверждает: параллельные вычисления действительно повышают производительность анализа больших данных в несколько раз.

ЗАКЛЮЧЕНИЕ

Были достигнуты следующие научные результаты:

1. Показана актуальность применения больших данных и параллельных вычислений в бизнес-аналитике, особенно в рамках Республики Беларусь. Современные предприятия, особенно в сфере электронной коммерции, сталкиваются с возрастающим объёмом данных, а потому использование параллельных вычислений позволяет эффективно обрабатывать большие данные и обеспечивает оперативное принятие решений на основе аналитики в реальном времени.

2. Выделены ключевые технологии, методы и инструменты анализа параллельных вычислений и анализа больших данных. Исследование охватило современные инструменты обработки данных, включая Python и его модули, MapReduce, облачные платформы и алгоритмы кластеризации. Особое внимание уделено модулям Dask, Modin и Swifter, имеющие различные преимущества в зависимости от типа выполняемых задач.

3. Выполнена разработка приложения для экспериментального сравнения производительности модулей для параллельного выполнения при анализе данных. На основании результатов работы приложения и тестов было обосновано, что ранее упомянутые модули действительно повышают производительность, сокращая время выполнения операций оригинального Pandas.

4. Было доказано, что при большом объёме данных может возникать падение скорости выполнения операций, которые должны ускоряться с применением правильного модуля, использующий технологии параллельных вычислений. Это подчёркивает важность выбора подходящего инструмента под конкретные типы операций и технические условия.

Полученные результаты формируют теоретическую и практическую базу для разработки ПО компьютерных систем для решения задач оценки эффективности параллельных технологий с применением компьютеров общего назначения, функционирующих в режиме реального времени. Они могут быть использованы для модернизации и дальнейшего развития существующих систем на языке Python. Также написанная и описанная в приложении программа может быть использована для проверки результатов использования параллельных технологий на Python не только для модулей на основе Pandas, но и для других тех, что используют технологии параллельные вычислений при правильной модификации и использовании.

Некоторые результаты научных изысканий при написании диссертационной работы были внедрены в учебный процесс на кафедре проектирования информационно-компьютерных систем учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» в учебный курс «Технологии параллельного программирования».

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Статьи в сборниках научных трудов

1. Мацокин, Н. П. Python-библиотеки для реализации параллельных вычислений при решении задач линейной алгебры. / Н. П. Мацокин, М. П. Мацокин // Актуальные вопросы экономики и информационных технологий: материалы 61-й Научной конференции аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники». – Минск: БГУИР, 2025. – с. 103-105.

РЭЗЮМЭ

Мацокін Мікалай Паўлавіч

Выкарыстанне інструментаў на аснове паралельных вылічэнняў для апрацоўкі і бізнес-аналітыкі вялікіх даных у вобласці продажаў

Ключавыя словы: вялікія даныя, паралельныя вылічэнні.

Мэта працы: ацэнка эфектыўнасці павышэння прадукцыйнасці аналізу вялікіх даных у галіне продажаў пры выкарыстанні тэхналогій паралельнага вылічэння.

Атрыманыя вынікі і іх навізна: выкананы аналіз бягучага стану рынку ІТ-прадукцыі у Беларусі. Было выяўлена, што ў цяперашні час ІТ-тэхналогіі шырока выкарыстоўваюцца ў сферы электроннай камерцыі, часцей за ўсё ў сферы анлайн-продажаў. Таксама быў праведзены агляд Big Data (вялікіх даных) і бягучых тэхналогій паралельнага вылічэнняў; распрацавана метад і ПЗ з мэтай ацэнкі эфектыўнасці розных сродкаў для аналізу вялікіх даных на аснове паралельных вылічэнняў на мове Python; у выніку распрацоўкі і тэсціравання выяўлены моцныя і слабыя бакі розных модуляў для паралельных вылічэнняў; пацверджана істотнае паскарэнне працэсаў апрацоўкі і бізнес-аналітыкі пры выкарыстанні паралельных вылічэнняў.

Ступень выкарыстання: вынікі ўкаранены ў навучальны працэс на кафедры праектавання інфармацыйна-камп'ютэрных сістэм ўстановы адукацыі «Беларускі дзяржаўны універсітэт інфарматыкі і радыеэлектронікі» ў навучальны курс «Тэхналогіі паралельнага праграмавання».

Вобласць ужывання: бізнес-аналіз вялікіх даных, тэхналогіі паралельных вылічэнняў.

РЕЗЮМЕ

Мацокин Николай Павлович

Использование инструментов на основе параллельных вычислений для обработки и бизнес-аналитики больших данных в области продаж

***Ключевые слова*:** большие данные, параллельные вычисления.

***Цель работы*:** оценка эффективности повышения производительности анализа больших данных в области продаж при использовании технологий параллельного вычисления.

***Полученные результаты и их новизна*:** выполнен анализ текущего состояния рынка IT-продукции в Беларуси. Было выявлено, что в настоящее время IT-технологии широко используются в сфере электронной коммерции, чаще всего в сфере онлайн-продаж. Также был проведён обзор Big Data (больших данных) и текущих технологий параллельного вычислений; разработано метод и ПО с целью оценки эффективности различных средств для анализа больших данных на основе параллельных вычислений на языке Python; в результате разработки и тестирования выявлены сильные и слабые стороны различных модулей для параллельных вычислений; подтверждено существенное ускорение процессов обработки и бизнес-аналитики при использовании параллельных вычислений.

***Степень использования*:** результаты внедрены в учебный процесс на кафедре экономической информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» в учебный курс «Технологии параллельного программирования».

***Область применения*:** бизнес-анализ больших данных, технологии параллельных вычислений.

SUMMARY

Matsokin Nickolay Pavlovich

Using Parallel Computing Tools for Big Data Processing and Business Analytics in Sales

***Keywords*:** Big Data, parallel computing.

***The object of study*:** To evaluate the effectiveness of improving the performance of Big Data analysis in sales using parallel computing technologies.

***The results and novelty*:** an analysis of the current state of the IT products market in Belarus. It was revealed that at present IT technologies are widely used in the field of e-commerce, most often in the field of online sales. A review of Big Data and current parallel computing technologies was also conducted; a method and software for assessing the effectiveness of various tools for analyzing Big Data based on parallel computing in the Python language were developed; as a result of development and testing, the strengths and weaknesses of various modules for parallel computing were identified; a significant acceleration of processing and business analytics was confirmed when using parallel computing.

***Degree of use*:** the results implemented in the educational process at the department of design information and computer systems educational institution «Belarusian State University of Informatics and Radioelectronics» in the training course «Parallel programming technologies».

Sphere of application: Big Data business analysis, parallel computing technologies.