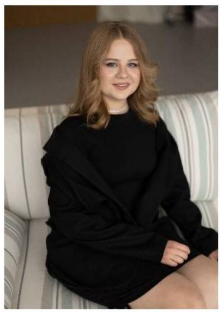
УДК 004.657:004.75-047.44

ЭФФЕКТИВНОЕ УПРАВЛЕНИЕ БОЛЬШИМИ ДАННЫМИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПЛАТФОРМ И ТЕХНОЛОГИЙ ДЛЯ ОБРАБОТКИ ИНФОРМАЦИИ

В.В. Верняховская
Заместитель декана инженерно-экономического факультета БГУИР, магистр экономических наук
verniahovskaya@bsuir.by

О.М. Раптунович
Старший преподаватель кафедры экономической информатики БГУИР, магистр
oraptunovich@gmail.com



В.А. Усова
Студентка инженерно-экономического факультета, специальности Информационные системы и технологии БГУИР
uvikamail@gmail.com

В.В. Верняховская

Окончила Белорусский государственный университет информатики и радиоэлектроники. Магистр экономических наук. Работает заместителем декана инженерно-экономического факультета БГУИР. Направления исследований: трансфер технологий, инновационная деятельность, информационные технологии в маркетинге.

О.М. Раптунович

Окончила Частый Институт Управления и Предпринимательства. Область научных интересов связана с актуальными вопросами экономики, логистики, маркетинга и информационных технологий и способами их решений.

В.А. Усова

Обучается в Белорусском государственном университете информатики и радиоэлектроники. на третьем курсе инженерно-экономического факультета. Направления исследований: информационные технологии в экономике, искусственный интеллект, машинное обучение, анализ данных.

Аннотация. В статье рассматривается эффективное использование больших данных и анализируются платформы и технологии для анализа данных. Используются различные инструменты и методы для анализа, хранения и обработки данных. Внимание уделено платформам Hadoop, Apache Spark и другим, а также технологии машинного обучения, влияющие на производительность и масштабируемость обработки данных.

Ключевые слова: большие данные, Hadoop, Apache Spark, машинное обучение, MapReduce, NoSQL базы данных, анализ данных, Kafka, Flink.

Введение. Если посмотреть на то, как изменился мир за последние десять лет, можно заметить, что данные стали неотъемлемой частью нашей жизни. По мере того, как развивались технологии объемы информации, которая поступала из разных источников, стали расти в геометрической прогрессии. В какой-то момент стало понятно, что традиционные методы хранения и анализа не справляются с таким объемом. После этого и появилась концепция «большие данные» (Big Data). На наш взгляд, важно понимать, что «большие данные» – это не про огромные размеры, а про объемы информации, которые превышают возможности обычных, традиционных систем для их обработки и хранения.

Такие данные отличаются не только внушительными размерами, но и разнообразием форматов, скоростью поступления и непостоянностью [1].

После того как стали распространены Интернета вещей, социальные сети, мобильные устройства и цифровые платформы объем создаваемых данных продолжает увеличиваться с экспоненциальной скоростью. Только за последние несколько лет было создано около 90% всех мировых данных. Если рассматривать более продолжительный период, то с 2010 года общий объем информации в мире вырос более чем в 66 раз (рисунок 1). В 2024 году данные достигли 147 зеттабайт, что почти на 22,5% больше, чем в предыдущем году. К 2025 году данные выросли до 181 зеттабайт, а к 2026 – до 221 зеттабайт [2].

Таким образом, только за один год объем произведенных данных во всем мире увеличился почти на 40 зеттабайт, что составляет рост примерно на 22,09% по сравнению с 2025 годом. На наш взгляд такая динамика наглядно подтверждает, что с каждым годом объемы данных будут только увеличиваться, и их влияние на все сферы человеческой жизни также будет только возрастать. Хранение, анализ, эффективная и правильная обработка информации станут одними из главных задач для бизнеса, науки и государства [3].

Рисунок 1. Количество данных, создаваемых в мире

Чтобы работать с данными нужны специальные инструменты и платформы, которые смогут расти вместе с объемами информации и эффективно работать с ними. Для этого приходится использовать сложные методы, такие как аналитические алгоритмы и распределенные вычисления, которые смогли бы извлекать нужную и полезную информацию из необработанных данных. Важно, чтобы платформа умела работать с разными типами данных, не только с текстом и числами, но и видео, аудио и другими. На данный момент таких технологий не мало, и каждая хороша по-своему. Выбор зависит от того, с какими данными планируется работа, какая стоит задача у проекта.

Цель нашей работы – сравнить платформы и технологии для работы с большими данными, разобрать их ключевые характеристики, преимущества и недостатки и определить для каких задач каждая подходит лучше.

Что такое большие данные? Большими данными (Big Data) называют огромные массивы структурированной и неструктурированной информации, с которыми обычные, традиционные базы данных и методы обработки уже не справляются. Сюда входит и текст, и видео, и аудио, и графика – все то, что ежедневно накапливается в самых разных сферах: от финансов и медицины до соцсетей и интернета вещей.

Описать большие данные можно через модель «3V». Объем (Volume) – это сами огромные массивы информации, и с каждым днем их становится все больше и больше.

Например, посты в социальных сетях, покупки в интернете, и другие данные, которые каждый день генерируют люди могут достигать колоссальных масштабов.

Скорость (Velocity) – это про то, как данные быстро и часто поступают и их нужно обрабатывать практически мгновенно. В качестве примера можно привести поведение пользователей в реальном времени или данные с устройств интернета вещей. Разнообразие (Veriety) – информация приходит в самых разных формах: текст, видео, аудио, числа, из-за этого ее становится очень сложно анализировать.

Сейчас к этим трем характеристикам все чаще добавляют и другие. Появились модели «4V» и «5V». Сначала модель «3V» придумали, чтобы просто описать основные, главные черты больших данных, но спустя время стало понятно, что есть и другие важные аспекты. Первый – это достоверность (Veracity), то есть надежность данных. Если исходные данные неточные или противоречивые, то и выводы по ним получатся ошибочные. Второй – это ценность (Value). Большие данные сами по себе ничего не дают, важно уметь извлекать из них полезную информацию и применять ее на практике.

В некоторых случаях добавляют еще визуализацию (Visualization) и вариативность (Variability), чтобы подчеркнуть, как важно правильно представлять данные и учитывать, что они меняются со временем.

Источники больших данных самые разные. Интернет вещей (IoT) – это датчики и устройства, которые постоянно передают информацию о своем состоянии. Соцсети генерируют огромные объемы текстов, фото, видео, по которым можно изучать поведение людей. Транзакции в электронной коммерции, медицинские исследования, научные эксперименты – все это тоже требует работы с большими объемами данных, для того чтобы получить точные и надежные результаты.

Но то, что данных много и они сложные, вовсе не значит, что их нельзя обработать. Современные технологии с этим справляются. Распределенные вычисления и новые подходы к хранению, такие как Hadoop и NoSQL-базы данных, позволяют хранить информацию, которая не помещается в обычные реляционные базы. А инструменты вроде Apache Spark помогают быстро анализировать огромные массивы. Все это открывает новые возможности – от бизнеса до научных исследований [4].

В бизнесе большие данные помогают улучшить аналитику. Прогнозирование спроса, предоставление клиентам того, что им действительно нужно, улучшение обслуживания – это все становится возможным благодаря анализу данных. В медицине большие данные помогают диагностировать и лечить заболевания, а также разрабатывать новые лекарства. В науке они используются для математических моделей, изучения окружающей среды и освоения космоса. Но, несмотря на все преимущества, в работе с большими данными есть проблемы. Безопасность и конфиденциальность остаются важными и актуальными вопросами.

Получается, большие данные – это не просто модное слово, тренд, а важный инструмент, который меняет подход к принятию решений. С их помощью повседневная жизнь становится удобнее и открываются новые возможности в разных сферах.

Платформы и технологии для обработки больших данных. Один из ключевых подходов к работе с большими данными – распределенные вычисления. Суть подхода проста: задачи делятся на части и распределяются между множества узлов – это и позволяет эффективно обрабатывать гигантские массивы информации. На такой основе создаются мощные системы для хранения и анализа данных. Один из инструментов – Apache Hadoop. Это целая платформа для распределенного хранения и обработки данных. В ее основе лежит технология MapReduce, которая выполняет задачи параллельно на разных узлах по схеме «главный – подчиненные». В качестве главного выступает сервер JobTracker, раздающий задания подчиненным узлам кластера и контролирующий их выполнение (рисунок 2) [5].

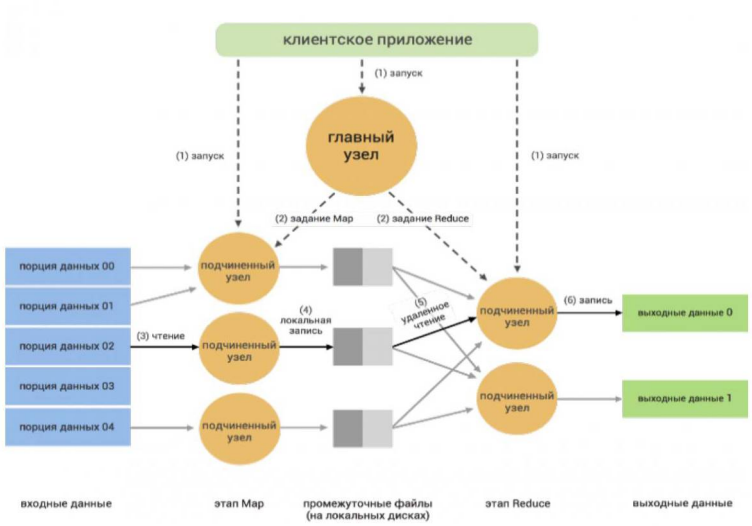


Рисунок 2. Архитектура Hadoop

Это открывает возможности для обработки петабайтов информации. Hadoop использует распределенную файловую систему HDFS (Hadeer Distributed File System), которая эффективно справляется с хранением больших объемов данных на различных машинах.

Но у MapReduce есть свои недостатки, например он не может быстро обрабатывать данные в реальном времени. Чтобы решить данную проблему разработали платформу Apache Spark, она более гибкая и быстрая для обработки больших данных.

В отличие от классического обработчика ядра Apache Hadoop с двухуровневой концепцией MapReduce на базе дискового хранилища, Spark использует специализированные примитивы для рекуррентной обработки в оперативной памяти. Благодаря этому многие вычислительные задачи реализуются в Смарк значительно быстрее. Сравнение Apache Hadeer и Spark представлено на рисунке 3.

Hadoop MapReduce	Apache Spark
Fast	100x faster than MapReduce
Batch Processing	Real-time Processing
Stores Data on Disk	Stores Data in Memory
Written in Java	Written in Scala

Рисунок 3. Сравнение Apache Hadoop и Spark

Стоит отметить, что Spark Streaming, в отличие от, например, Apache Storm, Flink или Samza, не обрабатывает потоки Big Data целиком. Для этого существует микропакетный подход. В нем поток данных разбивается на небольшие пакеты временных интервалов. Обычные реляционные базы данных, которым нужны жесткие схемы данных и которые не могут эффективно работать с огромными объемами информации не всегда подходят для хранения и обработки больших данных. В этой связи широко распространены NoSOL базы данных, которые предназначены для обработки неструктурированных и полуструктурированных данных, таких как текст, изображения и видео [6].

Среди популярных NoSOL баз данных можно выделить следующие. Cassandra – масштабируемость и высокая доступность. Apache Cassandra – это распределенная

NoSQL база данных, предназначенная для хранения больших объемов данных с высокой доступностью и масштабируемостью. Она была разработана для того, чтобы решать проблемы, возникающие при работе с огромными массивами данных в реальном времени. Cassandra выбирают компании, которым нужно, чтобы система продолжала работать, даже если несколько узлов базы данных вышли из строя. Основной особенностью Cassandra является архитектура peer-to-peer, в которой все узлы в кластере равны. Тут нет единой точки отказа, данные дублируются на нескольких узлах, поэтому они остаются доступными даже при сбоях. А если копии хранятся на серверах в разных географических точках, надежность становится еще выше. Cassandra используют там, где доступность критически важна, а именно в соцсетях, финансах, интернете вещей.

Еще плюсом выделяют гибкость в работе с данными. Cassandra использует модель колоночных семей, что позволяет удобно хранить и обрабатывать информацию, которая не укладывается в жесткие структуры или постоянно меняется. Система хорошо работает с огромными объемами данных разного формата и масштабируется по мере роста записей.

И главное – это то, что Cassandra выдерживает высокие нагрузки и отвечает на вопросы с минимальной задержкой, поэтому ее так любят компании, которым нужно обрабатывать большие данные в реальном времени [7].

MongoDB – одна из самых популярных NoSQL-баз. Она хранит данные в виде документов. В отличие от Cassandra, в ней используется формат похожий на JSON, поэтому MongoDB отлично подходит для работы с неструктурированными или не полностью структурированными данными, например, данные из соцсетей, веб-приложений или мобильных сервисов. Каждый документ может содержать разные типы данных: массивы, вложенные объекты – что угодно. Это дает разработчикам свободу: не нужно жестко задавать структуру заранее, ее можно менять по ходу дела, если меняются требования бизнеса или логика работы с данными.

MongoDB умеет ускорять запросы с помощью индексов, также поддерживает агрегацию, с помощью которой можно легко посчитать сумму, среднее или другие статистические показатели. Еще в ней есть встроенные инструменты для сложных запросов, поэтому для аналитики она подходит хорошо. Если данных становится слишком много, то можно добавить новые узлы в кластер, база масштабируется горизонтально без лишних сложностей. Чаще всего MongoDB используют в веб-разработке, мобильных приложениях, интернет-магазинах и задачах аналитики. Она прекрасно подходит для приложений, требующих быстрого доступа к данным, например, для ведения каталогов товаров или хранения пользовательских данных и предпочтений.

Apache HBase – это распределенная база данных, которая хорошо масштабируется. Она построена на основе Google Bigtable и предназначена для хранения и обработки больших объемов структурированной информации. HBase используется там, где важна высокая скорость записи и чтения, а также когда нужно работать с большими таблицами. Она хорошо подходит для хранения данных, которые удобно представляются в табличном виде, например, аналитические платформы, где хранятся большие наборы данных и требуется быстрая обработка.

Также как и Cassandra, HBase поддерживает репликацию, которая дает отказоустойчивость и надежный доступ к данным, но в отличие от MongoDB, где данные хранятся в виде документов, HBase использует табличную структуру и формат «ключ – значение», это позволяет эффективно работать с огромными объемами.

Одно из главных достоинств HBase – способность обрабатывать данные с очень высокой скоростью записи, поэтому ее часто выбирают для приложений, которые работают в реальном времени, например, системы мониторинга, веб-аналитики, обработки транзакций.

HBase не требует строгой системы, в ней можно хранить данные с разной структурой, система гибко масштабируется по мере роста объемов. Правда, в отличие от Cassandra и MongoDB, HBase заточена на работу с большими данными в кластерном режиме.

Cassandra, MongoDB и HBase – три мощные NoSQL-базы, у каждой свои сильные стороны и особенности, которые подходят для разных типов задач. Какую из них выбрать зависит от конкретного проекта, объемов данных и того, что нужно делать с приложением.

С ростом скорости генерации данных все важнее становится обработка в реальном времени. Платформы, которые это умеют, дают серьезное преимущество в бизнес-аналитике, мониторинге и управлении.

Apache Kalfka и Apache Flink представляют собой две платформы, которые часто применяются вместе для обработки данных в потоковом режиме. Kalfka является распространенной платформой для передачи потоковых данных. Ее характеристики – это высокая пропускная способность, масштабируемость и гарантии доставки сообщений. Платформа поддерживает хранение данных в заданный промежуток времени, что обеспечивает возможность их повторного анализа. Kalfka разделяет данные на темы, что позволяет улучшить отказоустойчивость и параллельную обработку.

Apache Flink ориентирован на обработку потоковых данных в реальном времени. Он берет данные из потоков и позволяет делать с ними сложные операции: аналитику, обработку событий, обучение моделей. Flink умеет управлять состоянием вычислений, работать с временными окнами и обрабатывать как бесконечные потоки, так и пакетные данные. Отметим, что события обрабатываются строго в том порядке, в котором поступили.

Вместе Kalfka и Flink работают просто: первый доставляет данные, второй обрабатывает. Такая пара хорошо подходит для мониторинга, аналитики в реальном времени и задач, связанных с машинным обучением. Вместе они дают высокую производительность, гибкость и возможность строить масштабируемые системы обработки потоков [8].

Технологии анализа данных и их применение. За последние 10 лет технологии анализа данных приобрели высокую значимость, потому что стали важной частью многих сфер (бизнес, наука, здравоохранение, государственное управление). Современные инструменты позволяют находить и получать полезную информацию из больших объемов данных, это открывает возможности для более обоснованных решений и оптимизации процессов.

Машинное обучение (Machine learning, ML) – это область искусственного интеллекта, в которой компьютер учится улучшать свои результаты на основе данных, без прописывания правил. Алгоритмы выявляют закономерности в примерах и потом применяют их к новым данным, которые раньше не встречались [9].

В зависимости от способа обучения выделяют три подхода. Обучение с учителем, при котором алгоритм обучается на заранее подготовленных данных, то есть на тех, где для каждого примера есть ответы. Обучение без учителя работает без заранее подготовленных данных, алгоритм сам ищет скрытые структуры, для этого группирует объекты (кластеризация), выявляет связи между ними. Этот подход применяется в рекомендательных системах. Обучение с подкреплением строится на взаимодействии с окружающей средой. Модель делает действие, получает положительную или отрицательную оценку и постепенно учится выбирать действия, которые приносили бы максимальную положительную оценку. Такой подход чаще используется в автономных системах.

Одним из популярных направлений машинного обучения в последнее время стало глубокое обучение. Он основан на многослойных нейронных сетях. Такие модели хорошо работают с большими объемами данных и сложными закономерностями.

Анализ данных шире чем машинное обучение. Он включает в себя и статистические методы и подходы на основе ИИ. Задачей анализа является выявление паттернов, трендов и связи данных. Информация для анализа данных может поступать из разных источников, например, базы данных, файлы, социальные сети, веб-страницы. Перед анализом данные очищают: убирают пропуски, устраняют выбросы, нормализуют значения – от этого зависит качество следующих результатов. Далее применяют статистические методы (проверка гипотез, расчет средних и отклонений) или алгоритмы машинного обучения. Визуализируем результаты с помощью графиков, диаграмм, интерактивных панелей, которые помогают увидеть данные с новой стороны и делают результаты более наглядными. На основе полученных данных и прогнозов принимаются обоснованные решения.

Машинное обучение и анализ данных продолжают развиваться. Сочетание мощных алгоритмов с большими объемами информации позволяет строить более точные прогнозы, оптимизировать процессы и создавать новые продукты и услуги.

Проблемы и вызовы при работе с большими данными. Когда мы работаем с большими данными, появляется возможность для анализа и принятия решений, но еще с этим появляются проблемы, которые приходится решать, чтобы использование данных было эффективным.

Одна из ключевых проблем – это защита чувствительной информации, такой как личные данные пользователей, коммерческая тайна и другие сведения, неправомерное использование которых может иметь серьезные последствия. Чтобы минимизировать риски применяют шифрование, механизмы контроля доступа и другие методы защиты, но с ростом объемов собираемых данных обеспечивать их безопасность становится труднее.

Объем сам по себе не гарантирует ценности данных, их полезность определяется точностью, полнотой и актуальностью. На практике обычно встречаются ошибки в записях, пропущенные значения, дублирование, противоречивая или устаревшая информация. Некачественные данные дают ошибочные выводы и делают ненадежными модели, построенные на их основе, именно поэтому предобработка становится обязательным этапом. Но даже при использовании отработанных методов с ростом количества данных гарантировать отличное качество данных сложно.

Чем больше данных, тем больше нужно вычислительных мощностей, чтобы их обработать. Если сервера слабые или каналы передачи данных узкие, то сложные задачи выполнятся долго. В некоторых случаях время очень важно, например, когда нужно мониторить потоки данных в реальном времени, обновлять прогнозные модели без задержек. Справиться с этим помогают распределенные системы. Они разбивают работу на части и раскладывают их по разным узлам. Это немного улучшает производительность, но не решает проблему полностью. Данных становится все больше, задачи усложняются, и даже хорошо настроенная распределенная инфраструктура рано или поздно достигает своего предела.

Будущее технологий больших данных. Дальнейшее развитие технологий больших данных связано с ростом объемов информации, генерируемой людьми, устройствами и автоматизированными системами. Эффективное извлечение значимых сведений из этих массивов становится ключевым фактором для бизнеса, научных исследований и управления системами.

Одно из основных направлений – это интеграция больших данных с методами искусственного интеллекта и машинного обучения. Такое совмещение позволяет повысить точность анализа и прогнозирования, а также сократить время принятия решений. На сегодняшний день предсказательная аналитика уже применяется для оптимизации бизнес-процессов, прогнозирования спроса и диагностики заболеваний. В дальнейшем подобные решения будут внедряться в различные сферы жизни, например, в городское управление, где анализ данных в реальном времени с использованием ИИ может использоваться для

регулирования транспортных потоков, контроля энергопотребления и мониторинга экологической обстановки.

Сейчас все больше внимания уделяют тому, как и где хранятся данные. Старые методы работы с большими массивами уже не справляются с потребностями новых типов данных, потому что появляется много «тяжелых» форматов: видео, аудио, изображение. Приходится искать новые подходы, облака и распределенные системы вроде тех, что работают на облачных платформах, помогают гибко расширять мощности под такие нагрузки и делают хранение более надежным.

В ближайшее время начнут серьезнее относиться к вопросам этики и безопасности данных. Чем больше информации собираем, тем выше риск, что она утечет или ее будут использовать не по назначению. Поэтому для компаний, которые работают с данными, на первый план выйдут не просто технические решения, а соблюдение законов, защита конфиденциальности пользователей и внятные стандарты безопасности. Технологии, такие как блокчейн, тоже могут сыграть ключевую роль в обеспечении прозрачности и безопасности обработки данных.

В целом, все идет к тому, что технологии больших данных будут становиться только интереснее, так как каждый год появляются новые инструменты, которые позволяют лучше понимать мир вокруг и принимать более обоснованные решения. Такие технологии будут продолжать развиваться, будут менять не только способы ведения бизнеса, но и повседневную жизнь людей, делая ее удобнее и безопаснее.

Заключение. Большие данные прочно вошли в современную информационную среду. Благодаря таким технологиям как Hadoop, NoSQL-базы и другие, организации теперь могут работать с объемами информации, которые раньше казались неподъемными. Но дело не только в том, чтобы хранить и обрабатывать это данные, подключение методов машинного обучения и искусственного интеллекта позволяет капать глубже и принимать решения не «на глазок», а на основе того, что реально видно из анализа. При этом как и в любой сложной системе, есть свои подводные камни. Безопасность, качество данных, производительность – все это остается вызовом, который не пропадает, а только усиливается со временем. Перспективы использования больших данные связаны с дальнейшем развитием технологий таких, как облачные вычисления, квантовые вычисления и автоматизацию процессов обработки данных.

Список литературы

[1] Тесленко И.Б., Губернаторов А.М., Дигилина О.Б., Крылов В.Е. BIG DATA=Большие данные: учебное пособие. Владимир: ВлГУ, 2021. – 123 с.

[2] Сколько данных создается каждый день? [Электронный ресурс]. – Режим доступа: https://inclient.ru/data-create-stats/. – Дата доступа: 01.03.2026

[3] Статистика больших данных за 2026 год [Электронный ресурс]. – Режим доступа: https://www.demandsage.com/big-data-statistics/. – Дата доступа: 01.03.2026

[4] Обработка потоковых данных [Электронный ресурс]. – Режим доступа: https://datareview.info/article/obrabotka-potokovyx-dannyx-storm-spark-i-samza/. – Дата доступа: 01.03.2026

[5] Обработка больших данных: первые шаги в понимании Hadoop MapReduce и Spark [Электронный ресурс]. – Режим доступа: https://devby.io/news/luxoft-big-data#google_vignette. – Дата доступа: 01.03.2026

[6] Модели данных в NoSQL [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/otus/articles/760226/. – Дата доступа: 01.03.2026

[7] Модель данных Apache Cassandra [Электронный ресурс]. – Режим доступа: https://bigdataschool.ru/wiki/cassandra#. – Дата доступа: 01.03.2026

[8] Построение архитектуры данных реального времени с помощью Apache Kafka, Flink и Druid [Электронный ресурс]. – Режим доступа: https://clck.ru/3GSU89. – Дата доступа: 01.03.2026

[9] В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. Теория и практика машинного обучения: учебное пособие. Ульяновск : УлГТУ, 2017. – 290 с.

Авторский вклад

Верняховская Вероника Владимировна – сформулировала задачу исследования, разработала структуру статьи, провела анализ платформ обработки больших данных, участвовала в формировании теоретической и методологической базы.

Раптунович Ольга Михайловна – выполнила сравнительный анализ технологий обработки и хранения данных, описала архитектуру и принципы работы рассматриваемых систем, подготовила иллюстрации и участвовала в написании основной части статьи.

Усова Виктория Александровна – исследовала методы анализа данных и машинного обучения, встроила полученные результаты в общую концепцию работы, сформулировала выводы и заключение, выполнила редактирование и подготовку статьи к публикации.

EFFECTIVE MANAGEMENT OF BIG DATA: A COMPARATIVE ANALYSIS OF PLATFORMS AND TECHNOLOGIES FOR INFORMATION PROCESSING

V.V. Verniahovskaya
Master of economics, Deputy Dean of the faculty of engineering and Economics at BSUIR

O.M. Raptunovich
Senior Lecturer at the Department of Economic Informatics BSUIR, master

V.A. Usova
Student of BSUIR, Faculty of Engineering and Economics, specialty Information Systems and Technologies

Abstract. The article examines the effective use of big data and analyzes platforms and technologies for data analysis. Various tools and methods are used for data analysis, storage, and processing. Attention is paid to the platforms Hadoop, Apache Spark and others, as well as machine learning technologies that affect the performance and scalability of data processing.

Keywords: big data, Hadoop, Apache Spark, machine learning, MapReduce, NoSQL databases, data analysis, Kafka, Flink.