*Хашагульгова Хяди Умаровна*

Ингушский государственный университет, Республика Ингушетия, Россия

khashagulgova777@icloud.com

Научный руководитель: *Кокурхаева Радима Магомет-Башировна*

Ингушский государственный университет, Республика Ингушетия, Россия

radima75@bk.ru

**«Большие данные» (Big Data): сбор, анализ и применение данных для принятия решений в эпоху цифровой трансформации**

**Аннотация:** Статья посвящена изучению феномена "больших данных" (Big Data) в контексте цифровой трансформации. Рассмотрены ключевые характеристики "больших данных" (5V), их источники и методы сбора, а также методы анализа и инструменты, используемые для обработки больших объемов информации. Особое внимание уделено применению "больших данных" для принятия решений в различных сферах, включая бизнес, здравоохранение, государственное управление и научные исследования. В заключение подчеркнута важность учета этических вопросов и необходимости ответственного подхода к использованию "больших данных" для достижения устойчивого развития и прогресса общества.

**Ключевые слова:** Большие данные (BigData), цифровизация, сбор данных, анализ данных, машинное обучение, искусственный интеллект, визуализация данных, облачные технологии, принятие решений на основе данных, бизнес-аналитика, здравоохранение, государственное управление, наука, этика данных, конфиденциальность, цифровое неравенство.

В современном мире, который стремительно переходит в эпоху цифровизации, данные стали не просто информацией, а ключевым ресурсом, определяющим конкурентоспособность, инновационный потенциал и способность адаптироваться к изменениям. Лавина цифровых данных, генерируемая ежедневно и ежесекундно, получила название "большие данные" (Big Data). Этот феномен, характеризующийся огромными объемами, высокой скоростью поступления, разнообразием форматов и не всегда достоверными источниками, открывает беспрецедентные возможности для бизнеса, науки, правительства и общества в целом. Однако, чтобы извлечь максимальную пользу из этих данных, необходимо не только уметь их собирать, но и разрабатывать эффективные методы анализа и применять полученные результаты для принятия обоснованных, взвешенных и своевременных решений.

"Большие данные" (Big Data) – это термин, обозначающий массивы данных, настолько огромные, сложные и быстро меняющиеся, что их анализ с использованием традиционных инструментов и методов обработки становится крайне затруднительным или практически невозможным. Эти данные генерируются из самых разнообразных источников и имеют отличительные характеристики, которые обычно обозначают как "5V", хотя некоторые исследователи добавляют и другие "V":

Огромное количество данных, исчисляемое терабайтами, петабайтами, эксабайтами и даже зеттабайтами. Этот объем постоянно растет и требует мощных вычислительных ресурсов для хранения, обработки и анализа.

Данные генерируются и обновляются с высокой скоростью, часто в режиме реального времени. Это требует быстродействующих систем сбора, обработки и анализа, способных реагировать на изменения и динамику поступления данных.

Данные поступают в самых разных форматах – структурированные (таблицы баз данных, CSV-файлы), неструктурированные (тексты, изображения, видео, аудио, данные социальных сетей) и полуструктурированные (логи, XML, JSON). Работа с таким разнообразием требует гибких и адаптивных инструментов.

Данные могут быть неполными, неточными, противоречивыми или содержать ошибки. Это требует тщательной проверки качества данных и применения методов очистки и нормализации.

Сами по себе данные не имеют ценности, но правильно проанализированные и интерпретированные, они могут стать источником ценной информации, знаний и инсайтов, которые могут быть использованы для принятия стратегических и оперативных решений.

В условиях цифровизации практически все действия и взаимодействия оставляют цифровой след, формируя беспрецедентные массивы данных. Источники "больших данных" становятся все более разнообразными и многочисленными:

Платформы, такие как Facebook, Twitter, Instagram, TikTok и LinkedIn, генерируют огромные объемы данных о пользователях, их предпочтениях, мнениях, интересах и взаимодействиях.

Миллиарды "умных" устройств, включая смартфоны, умные часы, датчики, сенсоры, подключенные к интернету, постоянно собирают и передают данные о самых различных процессах и явлениях.

Данные о покупках, платежах, банковских операциях, посещениях веб-сайтов, поисковых запросах, просмотре видео и других действиях в сети являются ценным источником информации о поведении и предпочтениях потребителей.

Современное промышленное оборудование, оснащенное датчиками и сенсорами, генерирует данные о своей работе, состоянии, производительности и эффективности.

Электронные медицинские карты, результаты анализов, обследований, генетические данные и данные научных исследований являются важным источником информации для развития медицины и улучшения здравоохранения.

Данные о местоположении, маршрутах, перемещениях людей и объектов, полученные с помощью GPS, ГЛОНАСС и других систем, предоставляют ценную информацию для управления транспортом, логистикой и городским планированием.

Данные, полученные в результате научных экспериментов, наблюдений, моделирований и других исследовательских работ, являются основой для развития науки и технологий.

Сбор "больших данных" осуществляется с использованием различных методов:

•  Веб-скрейпинг (Web scraping): Автоматизированный сбор данных с веб-сайтов, осуществляемый с помощью специальных программ и ботов.

•  API (Application Programming Interface): Получение доступа к данным через программные интерфейсы приложений, которые предоставляют стандартизированный способ обмена информацией.

•  Сбор данных с датчиков и сенсоров: Получение данных с устройств IoT, которые могут отслеживать температуру, влажность, движение, освещенность, давление и другие параметры.

•  Сбор данных из баз данных: Получение данных из структурированных баз данных, где данные хранятся в виде таблиц.

•  Сбор данных из логов: Получение данных из файлов журналов, которые содержат записи о событиях, происходящих в компьютерных системах.

Собранные данные требуют тщательного анализа и обработки, чтобы преобразовать их в ценную информацию и знания. Для этого используются различные методы и инструменты:

•  Визуализация данных: Представление данных в графическом виде (графики, диаграммы, карты, дашборды) для облегчения их понимания, выявления закономерностей, тенденций и аномалий.

•  Машинное обучение (Machine Learning): Использование алгоритмов для обучения компьютеров на основе данных, что позволяет им делать прогнозы, принимать решения, классифицировать объекты и выявлять закономерности без прямого программирования.

•  Искусственный интеллект (Artificial Intelligence): Создание систем, способных имитировать когнитивные функции человека, такие как обучение, рассуждение, распознавание образов, понимание естественного языка и принятие решений.

•  Статистический анализ: Использование статистических методов для выявления закономерностей, связей, зависимостей и трендов в данных.

•  Обработка естественного языка (Natural Language Processing, NLP): Использование алгоритмов для понимания и обработки текстовой информации, включая анализ тональности текста, распознавание именованных сущностей и машинный перевод.

•  Облачные технологии (Cloud Computing): Использование облачных платформ для хранения, обработки, анализа и визуализации данных, что обеспечивает масштабируемость, гибкость и экономическую эффективность.

•  Платформы для работы с большими данными: Использование специализированных платформ и инструментов, таких как Hadoop, Spark, TensorFlow и PyTorch, для обработки, анализа и машинного обучения на больших массивах данных.

Проанализированные "большие данные" могут использоваться для принятия обоснованных решений в самых разных сферах:

 •  Бизнес и маркетинг:

•  Персонализация маркетинга: Анализ данных о потребителях для создания индивидуализированных рекламных кампаний и предложений.

• Прогнозирование спроса: Анализ данных о продажах, потребительских предпочтениях и рыночных тенденциях для точного прогнозирования спроса на товары и услуги.

•  Оптимизация ценообразования: Анализ данных о затратах, ценах конкурентов и эластичности спроса для установления оптимальных цен.

•  Управление отношениями с клиентами (CRM): Анализ данных о взаимодействии с клиентами для повышения уровня их удовлетворенности и лояльности.

 •  Здравоохранение:

  •  Ранняя диагностика заболеваний: Анализ медицинских данных (генетические данные, результаты анализов, изображения) для раннего выявления заболеваний и назначения своевременного лечения.

  •  Разработка новых лекарств и методов лечения: Анализ данных о заболеваниях, клинических испытаниях и научных исследованиях для разработки новых лекарств и методов лечения.

  •  Персонализированная медицина: Анализ данных о пациентах для разработки индивидуальных планов лечения, учитывающих особенности организма, генетику и образ жизни.

 •  Государственное управление:

   •  Управление городским транспортом: Анализ данных о транспортных потоках, пробках, авариях для оптимизации работы общественного транспорта и управления дорожным движением.

  •  Управление ресурсами: Анализ данных о потреблении энергии, воды, газа и других ресурсов для оптимизации их распределения и сокращения потерь.

  •  Общественная безопасность: Анализ данных о преступлениях, социальных конфликтах и террористических угрозах для разработки программ профилактики и повышения уровня безопасности.

 •  Наука и исследования:

   •  Астрономия: Анализ данных, полученных с помощью телескопов и космических зондов, для изучения Вселенной.

  •  Геномика: Анализ геномов для изучения болезней и разработки методов лечения.

  •  Климатология: Анализ климатических данных для прогнозирования изменений климата и разработки стратегий адаптации.

Работа с "большими данными" сопряжена с рядом вызовов и этических вопросов, которые требуют особого внимания:

•  Защита персональных данных: Сбор, обработка и анализ персональных данных требует строгого соблюдения законодательства о защите персональных данных и обеспечения безопасности и конфиденциальности информации.

•  Предвзятость алгоритмов: Алгоритмы машинного обучения могут воспроизводить и усиливать существующие предрассудки и дискриминацию, если они обучаются на предвзятых данных.

•  Прозрачность и объяснимость: Необходимо обеспечить прозрачность работы алгоритмов и понимание того, как они принимают решения, чтобы избежать необъективности и дискриминации.

•  Ответственность и подотчетность: Необходимо определить ответственность за решения, принятые на основе анализа "больших данных", особенно в критических областях, таких как медицина, транспорт и финансы.

•  Цифровое неравенство: Неравный доступ к цифровым технологиям и образованию может усугубить социальное неравенство и лишить некоторые группы населения возможностей, которые предоставляют "большие данные".

"Большие данные" – это мощный инструмент, который может революционизировать различные сферы нашей жизни. Однако для того, чтобы в полной мере использовать их потенциал, необходимо не только уметь собирать, анализировать и интерпретировать данные, но и подходить к их применению с пониманием этических вопросов, социальной ответственности и необходимостью защиты персональных данных. Только в этом случае "большие данные" станут двигателем прогресса, благополучия и устойчивого развития для всего человечества.

**Список использованной литературы**

1.Натан Марц, Большие данные. Принципы ипрактика построения масштабируемых системобработки данных в реальном времени / Натан Марц,Джеймс Уоррен, – М.: Вильямс. -2016. -292 c.

2. Анализ больших данных: Spark и Hadoop [Электронный ресурс]. - URL: https://coincase.ru/blog/47715/ (дата обращения: 20к.06.2020).

3. Базы данных [Электронный ресурс]. – URL: <https://siblec.ru/in->formatika-i-vychislitelnaya-tekhnika/bazy-dannykh (дата обращения: 18.06.2020).

4. Большие данные (Big Data) [Электронный ресурс]. - URL:

https://www.tadviser.ru/index.php/Статья:Большие\_данные\_(Big\_Data)

(дата обращения: 21.06.2020).

5. Большие данные [Электронный ресурс]. - URL: https://yan-

dex.ru/profi/courses2019/big\_data (дата обращения: 21.06.2020).