

СЕКЦИЯ 1. СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ, ОБРАБОТКА ИНФОРМАЦИИ

DOI: 10.58168/CISMP2024_13-17

УДК 004.9

ОБРАБОТКА БОЛЬШИХ ДАННЫХ

Д.А. Алексеева

ФГБОУ ВО «Воронежский государственный лесотехнический университет
имени Г.Ф. Морозова»

Аннотация. В данной работе был проведен комплексный анализ выбранной темы. Была изучена соответствующая литература на тему обработки больших данных.

Ключевые слова: Big Data, обработка больших данных.

BIG DATA PROCESSING

D.A. Alexeeva

Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. In this work, a comprehensive analysis of the selected topic was carried out. The relevant literature on the topic of big data processing was studied.

Keywords: Big Data, big data processing.

В эпоху цифровой трансформации объемы данных, генерируемых ежедневно, стремительно растут, создавая новые возможности для бизнеса, науки и общества в целом. Обработка больших данных (BigData) стала ключевым аспектом современного анализа информации, позволяя извлекать ценные инсайты из массивов данных, которые ранее считались слишком объемными или сложными для традиционных методов обработки. Эта статья исследует основные концепции и технологии, стоящие за обработкой больших данных, их применение в различных отраслях и влияние на принятие решений.

Мы рассмотрим не только технические аспекты, такие как алгоритмы и инструменты, но и этические вопросы, связанные с использованием данных, что делает тему особенно актуальной в условиях современного мира.

Определение BigData. К большим данным можно отнести информацию, объем которой может превышать сотни терабайтов и петабайтов, причем такая информация обновляется на регулярной основе. Иногда в понятие «большие данные» включают способы и методики их обработки.

Когда речь идет о терминологии, «BigData» охватывает не только сами данные, но и методы их обработки, возможности их дальнейшего использования и способы поиска конкретной информации в огромных объемах. Эти вопросы остаются актуальными и важными для систем, которые на протяжении многих лет собирали и хранили разнообразные данные.

Ключевые аспекты, связанные с обработкой больших данных, включают:

1. **Объем (Volume):** Огромные объемы данных, требующие особые технологии для обработки и хранения.

2. **Скорость (Velocity):** Быстрота, с которой данные создаются и обрабатываются. Особенно важным это является для систем, которые работают в реальном времени.

3. **Разнообразие (Variety):** Разнообразие типов данных (структурированные, неструктурированные, полуструктурированные), различные форматы.

4. **Достоверность (Veracity):** Надежность и точность данных, что является критичным для принятия обоснованных решений.

5. **Ценность (Value):** Способность извлекать полезные инсайты из данных, а это является основной целью обработки больших данных.

Сбор и хранение больших данных перед обработкой

Ресурсами, выдающими большие данные, могут являться:

1. Интернет

2. Устройства, собирающие информацию (например, GPS-сигналы автомобилей и т.д.)

3. Корпоративные источники

Саму операцию и совокупность методик по сбору больших данных называют DataMining. Формат данных может быть разным — текст, видео, таблицы, SAS.

Процесс сбора и обработки больших данных можно разбить на основные этапы:

1. Формулирование задачи для аналитической программы.

2. Сбор данных с их подготовкой.

3. Выбор алгоритма для анализа данных.

4. Обучение программы выбранному алгоритму и дальнейший анализ выявленных закономерностей.

В основном необработанные данные хранятся в «озере данных» (DataLake). Обработка больших баз данных требует серьезные вычислительные мощности, а облако, которое является месторасположением «озера», значительно снижает стоимость работ. По этой причине множество компаний выбирают облачные хранилища.

Облачные решения предлагают несколько преимуществ по сравнению с собственными дата-центрами. Поскольку сложно предсказать будущую нагрузку на инфраструктуру, приобретение оборудования становится нецелесообразным. Оборудование, купленное на случай увеличения потребностей, может просто простаивать, что ведет к убыткам. В то же время, если мощности окажутся недостаточными, это негативно скажется на работе системы.

Облако же не имеет ограничений по объему хранимых данных, что делает его экономически выгодным для компаний с быстро растущими нагрузками и для бизнеса, занимающегося тестированием различных гипотез.

Преимущества технологии BigData

Большие данные – драйвер мировой экономики. Они помогают:

1. Работать с большими объёмами информации.
2. Принимать более взвешенные решения и строить точные прогнозы.
3. Мгновенно реагировать на сбои и уязвимости: банки и платежные

системы могут практически сразу отследить подозрительные действия и предотвратить мошенничество благодаря доступу ко всем данным о действиях пользователей.

4. Разрабатывать долгосрочные планы действий. Если компания располагает информацией о продажах, прибыли и убытках за несколько лет, то её анализ позволит эффективно планировать инвестиции, управление персоналом и ассортиментом.

5. Устранять недочёты и совершенствовать продукт.

Однако существуют и недостатки:

1. Сложности с расширением.

2. Высокие риски. Большие объёмы данных повышают требования к безопасности. Если, например, злоумышленники получают доступ к базе данных крупного банка, то миллионы клиентов могут потерять свои средства. Чтобы предотвратить подобные ситуации, компании, работающие с BigData, применяют распределённый доступ: у разных групп сотрудников разный уровень доступа только к определённым сегментам баз данных. Кроме того, данные шифруются и структурируются на каждом уровне.

3. Большие расходы. Работа с большими данными требует значительных вычислительных ресурсов и более дорогостоящих сервисов для хранения и обработки информации. Например, для обучения нейросетей необходимы обширные наборы данных, которые есть только у крупных корпораций и часто недоступны для свободного использования. Для работы с большими данными требуется привлекать квалифицированных специалистов: аналитиков данных, экспертов по хранилищам данных (DWH), специалистов по бизнес-аналитике (BI).

Сферы применения аналитики больших данных

Bigdata активно используются в бизнесе, государственном секторе и промышленности. Вот некоторые области, где большие данные находят наиболее широкое применение:

1. Транспорт. Навигаторы, используя большие данные о маршрутах и скорости движения транспортных средств, предлагают кратчайший путь с учётом пробок.

2. Мобильная связь и интернет. Операторы мобильной связи применяют большие данные для прогнозирования нагрузки на сеть в разных зонах и определения необходимости установки дополнительных базовых станций, включая станции 5G там, где они наиболее востребованы.

3. Медицина. Анализ больших данных позволяет прогнозировать развитие эпидемий и своевременно увеличивать производство вакцин.

4. Маркетинг.
5. Подбор персонала. Используя информацию из миллионов профилей успешных кандидатов, можно разработать алгоритм, который будет автоматически отбирать наиболее подходящих соискателей и отправлять им приглашения на собеседования.
6. Производство. Анализ больших объёмов данных помогает оптимизировать работу сотрудников и снизить риски аварий и несчастных случаев на производстве.
7. Финансовые технологии. На основе информации о случаях мошенничества банки могут создать более безопасные сервисы для онлайн-платежей.
8. Образование. Благодаря большим объёмам данных можно составлять персонализированные рекомендации вузов для абитуриентов и профессий для выпускников.
9. Интернет вещей. Большие данные используются для улучшения работы умных устройств, сенсорных датчиков, камер наблюдения и систем управления беспилотными автомобилями, что делает жизнь людей проще и удобнее.
10. Наука. Научные исследования, опросы и показания приборов позволяют обнаруживать неочевидные связи и делать новые открытия в различных областях науки.
11. Государственное управление. Государственные органы и международные организации используют большие данные в виде статистических сведений для более эффективного распределения ресурсов и реагирования на проблемы, которые действительно важны для людей.
12. Искусственный интеллект и робототехника. Компании обучают голосовых и чат-ботов, заменяющих сотрудников техподдержки или колл-центра, используя датасеты с реальными диалогами.

В статье мы рассмотрели тему обработки больших данных. Выяснили, в чём заключается значимость BigData, но также и какие сложности и риски могут возникнуть при работе с большими данными. Изучили сферы, в которых применяется аналитика больших данных.

Список литературы

1. Агафонов А. Ю. Обработка и хранение больших данных. Прикладные аспекты // Современные информационные технологии и ИТ-образование. 2017. № 4. С. 35–40.
2. Васильев А. Н., Тархов Д. А. Методы обработки больших данных в информационных системах // Информационные системы и технологии в моделировании и управлении. 2020. № 1 (13). С. 6–15.
3. Наместников А. М., Чугунов А. В. Технологии обработки больших объёмов данных // Труды СПИИРАН. 2018. № 5 (60). С. 96–112.
4. Сухомлин В. А. Введение в анализ больших данных // Прикладная информатика. 2013. № 3 (45). С. 11–25.

5. Иваньо, Я.М. Применение больших данных для планирования производства продовольственной продукции в условиях неопределенности / Я.М. Иваньо, П.Г. Асалханов, Н.В. Бендик // Моделирование систем и процессов. – 2021. – Т. 14, № 2. – С. 13-20. – DOI: 10.12737/2219-0767-2021-14-2-13-20.

References

1. Agafonov A. Yu. Processing and storing big data. Applied aspects // Modern information technologies and IT education. 2017. No. 4. P. 35–40.

2. Vasiliev A. N., Tarkhov D. A. Methods of processing big data in information systems // Information systems and technologies in modeling and management. 2020. No. 1 (13). P. 6–15.

3. Namestnikov A. M., Chugunov A. V. Technologies for processing large volumes of data // Proceedings of SPIIRAS. 2018. No. 5 (60). P. 96–112.

4. Sukhomlin V. A. Introduction to big data analysis // Applied informatics. 2013. No. 3 (45). P. 11–25.

5. Ivano, Ya.M. The use of big data for planning food production in conditions of uncertainty / Ya.M. Ivano, P.G. Asalkhanov, N.V. Bendik // Modeling of systems and processes. - 2021. – Vol. 14, No. 2. – pp. 13-20. – DOI: 10.12737/2219-0767-2021-14-2-13-20.