

Формализация информации и Big Data

<http://vikchas.ru>

Лекция 2 «Большие данные и перспективы их развития с использованием БФМ ИИ»

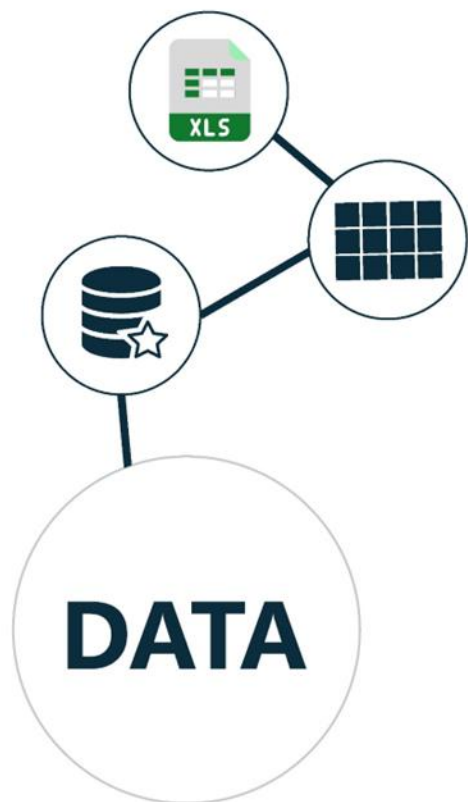
Часовских Виктор Петрович

д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2026

Большие данные и их характеристики



Объем данных Сбербанк

ОБЪЕМ ДАННЫХ
СБЕРБАНКА
> 244 ПБ



≈



ФОТОГРАФИЙ
В VK
290 000 000 000



за 1
секунду



200 000
ТРАНЗАКЦИЙ

×3,27



~ **61 000**

ПРОСМОТРОВ
РОЛИКОВ В YOUTUBE

ДАННЫЕ: Большие массивы цифровых данных различной структуры

Несколько важных факторов

1 Используются только оцифрованные данные

2 Основные источники больших данных

- Логи поведения пользователей в интернете
- Профили и клиентский контент в социальных сетях
- GPS-сигналы от различных устройств
- Данные, снимаемые с датчиков
- Данные фото и видеофиксации
- Оцифрованные книги
- Информация о транзакциях всех клиентов банка
- Информация о всех покупках в крупной ритейл сети
- События от радиочастотных идентификаторов
- ГЕО-локация абонентов сетей сотовой связи
- ...

3 Владельцы данных

- Корпорации
- Физические лица
- Государство

Свойства данных

- Большие объемы
- Разнообразие
- Высокая скорость генерации

Требования
к технологиям

Способы получения данных

- Покупка
- Сбор из открытых источников
- Партнерства
- Собственная генерация

- Потребности в партнерствах
- Государственном регулировании
- соглашениях с клиентами Банка



СЕГОДНЯ КЛЮЧЕВУЮ РОЛЬ В ДОСТУПНОСТИ ВНЕШНИХ ДАННЫХ ИГРАЮТ СОГЛАШЕНИЯ С КЛИЕНТОМ, ДОГОВОРЕННОСТИ С КОРПОРАЦИЯМИ И РЕГУЛИРОВАНИЕ СО СТОРОНЫ ГОСУДАРСТВА

5 ключевых элементов экосистемы больших данных



Базовая инфраструктура

выполняет две основные функции: сбор и передача данных

Оборудование для сбора данных

Любое оборудование,
способное записывать данные



Смартфоны



IoT-датчики
(сетей объектов Интернета вещей)



Камеры



"Умная" бытовая техника

И т.д.

Сеть

Средства передачи данных с оборудования,
осуществляющего их сбор, в цифровую
инфраструктуру



Проводные локальные сети



Мобильные сети



Wi-Fi и Bluetooth



Протоколы беспроводной связи для
Интернета вещей (LoRaWAN, ZigBee, NB-IoT)

И т.д.

Большие фундаментальные модели ИИ существенно расширяют возможности анализа больших данных.

Работа с неструктурированными данными

Раньше большая часть полезной информации оставалась в текстах, письмах, документах, изображениях, аудио и видео.

LLM и мультимодальные модели позволяют: анализировать договоры; обрабатывать обращения клиентов; классифицировать отзывы; извлекать сведения из отчётов; распознавать изображения; анализировать звонки; искать смысловые связи в документах.

Это делает неструктурированные данные полноценным источником аналитики.

Естественно-языковой интерфейс к данным

Фундаментальные модели позволяют обращаться к данным обычным языком.

Например, пользователь может спросить:

«Покажи продажи по регионам за последний квартал и объясни, почему в Сибири снизился спрос».

ИИ может: понять вопрос; сформировать SQL-запрос; получить данные; построить график; объяснить результат простыми словами.

Это снижает барьер входа в аналитику и делает большие данные доступными не только программистам и аналитикам.

Автоматизация очистки и подготовки данных

Большие данные часто содержат ошибки: дубли; пропуски; разные форматы; опечатки; противоречивые записи; шум.

ИИ-модели могут помогать: находить аномалии; сопоставлять похожие записи; стандартизировать названия; заполнять пропуски; объяснять ошибки; предлагать правила очистки.

Улучшение поиска и извлечения знаний

Традиционный поиск часто работает по ключевым словам. Фундаментальные модели позволяют выполнять **семантический поиск**, то есть поиск по смыслу.

Например, запрос:

«документы о задержке поставок» может найти тексты, где нет слова «задержка», но есть фразы: «сроки доставки были перенесены»; «отгрузка не выполнена вовремя»; «логистический сбой».

Это особенно важно для корпоративных архивов, юридических документов, медицины и научных баз знаний.

Генерация аналитических выводов

ИИ может не только находить данные, но и формулировать выводы: почему изменилась метрика; какие факторы повлияли на результат; какие риски существуют; какие действия можно предпринять; какие сценарии развития возможны.

Например: «Рост оттока клиентов связан с увеличением времени ответа службы поддержки и снижением доступности товара в двух регионах»

Объединение разных типов данных

Фундаментальные модели развивают мультимодальный анализ.

Они могут объединять: текстовые документы; таблицы; изображения; видео; аудио; данные датчиков. Например, в промышленности ИИ может одновременно анализировать: показания оборудования; текстовые журналы ремонтов; фотографии дефектов; сообщения операторов.

Это позволяет точнее прогнозировать поломки и повышать эффективность производства.

Архитектура современных решений

Типичная система больших данных с использованием фундаментальной модели ИИ может включать:

Источники данных базы данных; документы; датчики; интернет-источники.

Хранилище данных

Обработка данных

Индексирование и поиск

Перспективные направления развития

Интеллектуальные аналитические платформы

Будут развиваться платформы, где пользователь сможет взаимодействовать с данными через диалог.

Автономные ИИ-агенты для работы с данными

ИИ-агенты смогут самостоятельно выполнять цепочки действий: находить нужные данные; проверять их качество; строить модели; запускать расчёты; готовить отчёты; уведомлять ответственных сотрудников.

Персонализированная аналитика

Развитие мультимодальной аналитики

Будет расти значение анализа нескольких типов данных одновременно.

Примеры: медицина: снимки + история болезни + анализы + жалобы пациента; промышленность: датчики + видео + журналы ремонтов; образование: тексты работ + поведение в LMS + устные ответы; транспорт: камеры + GPS + погодные данные.

Основные риски и проблемы

Несмотря на перспективы, существуют серьёзные вызовы.

Конфиденциальность данных

Большие данные часто содержат персональную или коммерческую информацию. Нужно обеспечить: шифрование; разграничение доступа; анонимизацию; соответствие законодательству; контроль использования данных моделью.

Качество данных

Если данные неполные или ошибочные, ИИ может сделать неправильные выводы.

Галлюцинации моделей

LLM может создавать правдоподобные, но неверные ответы.

Высокая стоимость

Работа с большими данными и фундаментальными моделями требует: вычислительных ресурсов; хранилищ; специалистов; инфраструктуры; мониторинга.

Безопасность

Возможны атаки: утечка данных через ответы модели; подмена документов; манипуляция входными данными; несанкционированный доступ.

Требования к успешному внедрению

Для эффективного применения больших фундаментальных моделей в больших данных нужны: качественная архитектура данных; единые справочники и метаданные; управление доступом; чистка и нормализация данных; векторный и семантический поиск; мониторинг качества ответов; защита персональных данных; обучение сотрудников; понятные бизнес-сценарии; оценка экономического эффекта.

Будущее больших данных и ИИ

В ближайшие годы большие данные и фундаментальные модели ИИ будут развиваться совместно.

Можно ожидать: перехода от отчётной аналитики к диалоговой; роста роли мультимодальных данных; появления корпоративных ИИ-ассистентов; автоматизации рутинной аналитики; развития отраслевых моделей; увеличения значения RAG и векторных баз; усиления требований к безопасности и регулированию; широкого применения ИИ-агентов; роста спроса на специалистов по данным и ИИ.

Главная перспектива заключается в том, что большие данные перестанут быть просто массивами информации. С помощью фундаментальных моделей ИИ они станут основой для интеллектуальных систем, которые могут не только хранить и обрабатывать данные, но и понимать контекст, объяснять выводы, предлагать решения и помогать человеку в сложных задачах.

Цифровая инфраструктура обеспечивает прием, хранение и обработку данных

Платформы для интеграции больших данных

Инструменты для получения,
организации и управления данными



Инструменты получения данных

Получение данных,
в т.ч. в режиме реального времени



Инструменты интеграции и целостности данных

Управление несколькими источниками данных



Управление данными

Базы данных, хранилища данных и среды

Хранение и обработка

Ресурсы для хранения и обработки данных
посредством оборудования и/или облачных
технологий



Облачная среда

Виртуальные аппаратные средства



ИТ оборудование

Компьютерное оборудование

Технологические инструменты позволяют выполнять общие задачи, такие как анализ данных, машинное обучение и работа ИИ

Базовая аналитика

Инструменты для извлечения данных
и выполнения простых расчетов



Поисковые системы



Инструменты запроса данных
и отчетности

Углубленная аналитика и ИИ

Инструменты, позволяющие использовать
сложные научные методы, такие как статисти-
ческое моделирование и машинное обучение



Инструменты для построения
статистических моделей



Инструменты для использования анализа
местоположения и контента на основе
машинного обучения (ML)



Платформы с поддержкой ИИ,
объединяющие различные инструменты
углубленной аналитики

Различные вертикальные приложения для решения конкретных задач могут быть разработаны самостоятельно или приобретены на рынке

Вертикальные решения

Решения, нацеленные на улучшение одной или нескольких функций в организации



Решения для отдельных функций

Управление цепочками поставок, аналитика данных в сфере персонала, аналитика операционной деятельности



Межфункциональные решения

Планирование ресурсов, управление рисками, нормативно-правовое соответствие (комплаенс-контроль)

Бизнес-услуги

Сервисы поддержки вертикальных решений для больших данных



Поддержка внедрения

Консалтинг и аутсорсинг процессов с использованием больших данных



Продажа данных

Продажа обработанных данных на рынке

Большие данные как объект управления

Данные

Большие массивы цифровых структурированных и неструктурированных данных

Таблицы, текст, изображение, голос, видео

Технологии

Возможность хранить и обрабатывать практически неограниченные объемы данных любой структуры

Существенное снижение стоимости хранения и обработки данных

Hadoop, Spark..

Аналитика и машинное обучение

Выявление скрытых зависимостей на основе анализа всего объема данных.

Новое качество результатов машинного обучения

Исследователи данных открывают новые закономерности и возможности для бизнеса

Люди

специалисты в области больших данных, потребители больших данных

Data scientists, Data engineers, аналитики

№ пп	СУБД	HADOOP (Хэдуп) мощная open-source платформа для распределённой обработки и хранения больших данных.
1.	Традиционные базы данных на основе строк и столбцов, в основном используемые для хранения, обработки и извлечения данных.	Программное обеспечение с открытым исходным кодом, используемое для хранения данных и одновременного запуска приложений или процессов.
2.	В этой в основном обрабатываются структурированные данные.	В ней обрабатываются как структурированные, так и неструктурированные данные.
3.	Она лучше всего подходит для среды OLTP.	Она лучше всего подходит для работы с большими данными.
4.	Она менее масштабируема, чем Hadoop.	Она обладает высокой масштабируемостью.
5.	В СУБД требуется нормализация данных.	В Hadoop нормализация данных не требуется.

6.	В ней хранятся преобразованные и агрегированные данные.	В ней хранится огромный объем данных.
7.	У нее нет задержки в ответе.	У нее есть некоторая задержка в ответе.
8.	Схема данных СУБД имеет статический тип.	Схема данных Hadoop имеет динамический тип.
9.	Доступна высокая целостность данных.	Доступная целостность данных ниже, чем у СУБД.
10.	Стоимость указана за лицензионное программное обеспечение.	Бесплатно, поскольку это программное обеспечение с открытым исходным кодом.



PostgreSQL

СУБД

ADABAS

СУБД

Благодарю за внимание!

