Формализация информации и Big Data

http//vikchas.ru

Тема 2. Big Date

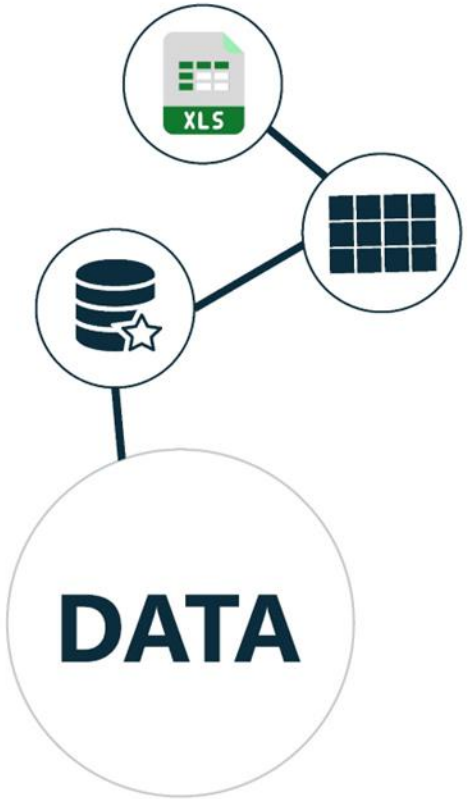
Лекция 2 «Большие данные и перспективы их развития»

Часовских Виктор Петрович
д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2024

Большие данные
и их характеристики



Объем данных Сбербанка

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

Большие данные

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

Личная активность/ операции, движение транспортных средств, производст-венные процессы и т.д.

Генерация данных

Вертикальные решения и услуги

- Вертикальные решения
- Бизнес-услуги и продажа данных

Решения для конечного пользователя: приложения и услуги, направленные на решение конкретных проблем

Технологические инструменты

- Базовая аналитика
- Углубленная аналитика и ИИ

Программное обеспечение, предназначенное для выполнения общих задач, таких как анализ данных, ИИ (искусственный интеллект), машинное обучение (ML)

Цифровая инфраструктура

- Платформа для интеграции больших данных
- Хранение и обработка

Цифровые платформы, обеспечивающие поток данных, их хранение и вычисление

Базовая инфраструктура

- Оборудование для сбора данных
- Сеть

Оборудование для сбора данных и сеть для передачи 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.

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 статистическое 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
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

Благодарю за внимание!