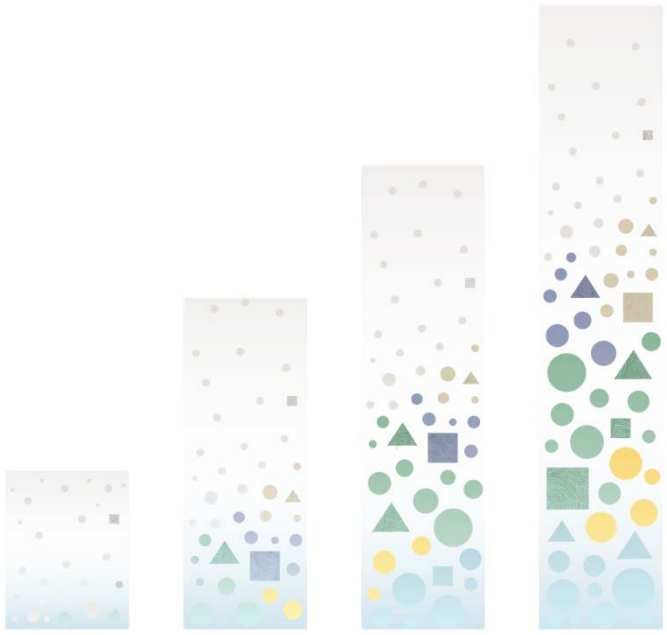
Основы анализа больших данных

Лекция 1. Введение. Определение больших данных, области в которых используется анализ больших данных

Введение

Область больших данных развилась от дисциплины статистического анализа до современных передовых технологий платформ данных.

В этой обзорной лекции мы опишем историю развития больших данных, какие проблемы возникают на этом пути, и как организации используют платформы данных, чтобы получить от данных больше пользы, чем когда-либо прежде. Рассмотрим текущее состояние и перспективы развития технологий больших данных, познакомимся с основными методами анализа, технологиями и инструментами для анализа больших данных, одним из которых является программная среда и язык R - язык программирования для статистической обработки данных и работы с графикой.



Определение больших данных

Большие данные — это концепция, описывающая наборы разнообразных данных, поступающих с более высокой скоростью, объем которых постоянно растет. Простыми словами, большие данные — более крупные и сложные наборы данных, особенно из новых источников данных. Размер этих наборов данных настолько велик, что традиционные инструменты для обработки не могут с ними справиться. Однако эти большие данные можно использовать для решения бизнес-задач, которые раньше не могли быть решены. Т.о., они определяются тремя V: большой объем, высокая скорость поступления, разнообразие.

Основные свойства больших данных

Три V больших данных:

- Объем (**Volume**)
- Скорость (**Velocity**)
- Разнообразие (**Variety**)

Ценность больших данных, их достоверность и изменчивость

Еще 3 свойства сформировались за последние несколько лет:

- **ценность (value)**
- **достоверность (veracity)**
- **изменчивость (variability)**

Данные имеют внутренне присущую им ценность (или значимость). Однако чтобы они приносили пользу, эту ценность необходимо раскрыть. Не менее важно и то, насколько достоверны большие данные и насколько можно на них полагаться.

Главные источники больших данных:

- интернет вещей (IoT) и подключенные к нему устройства;
- соцсети, блоги и СМИ;
- данные компаний
- показания приборов
- статистика городов и государств
- медицинские данные.

Сегодня большие данные стали разновидностью капитала.

История больших данных

Корни понятия больших данных берут начало в «бизнес-аналитике» (business intelligence), термине, который компания IBM ввела в 1958 году и определила его как «*способность воспринимать взаимосвязи представленных фактов таким образом, чтобы направлять действия к желаемой цели*» (IBM, 1958)

Сам термин «большие данные» предложил редактор журнала Nature Клиффорд Линч в спецвыпуске 2008 года*,описывая взрывной рост потоков информации. К большим данным Линч отнес любые массивы неоднородных данных более 150 Гб в сутки, однако единого критерия до сих пор не существует.

* Nature, Volume 455 Issue 7209, 4 September 2008

Новые подходы к большим данным

Примерно в 2005 году мы вступили в эпоху Web 2.0, когда компании начали осознавать, сколько данных пользователи генерируют через социальные сети и другие онлайн-сервисы. Требовался новый подход.

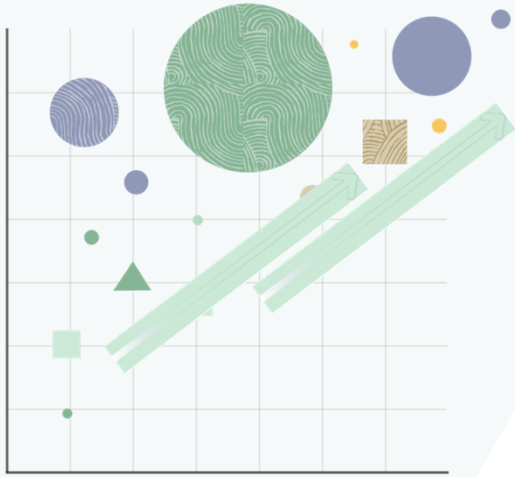
Google опубликовал статью о **MapReduce** — модели программирования, определяющей систему обработки больших наборов данных. Yahoo подключилась к проекту, и в 2008 был создан **Hadoop**.

Комбинация **Hadoop + MapReduce** позволила реализовать сценарии использования больших данных, которые ускорили развитие цифровой экономики. Ранее такие варианты использования были невозможны или стоили слишком дорого.

Spark, механизм обработки данных с открытым исходным кодом для больших наборов данных, стал популярным, поскольку он обеспечивал скорость вычислений, масштабируемость и программируемость для больших данных, особенно с приложениями для потоковой передачи данных, графических данных, машинного обучения (**ML**) и искусственного интеллекта (**AI**).

Большие данные ускорили развитие цифровой экономики,

позволив реализовать сценарии использования, которые раньше было невозможно или слишком затратно реализовать.



Сложности при использовании больших данных

- Большие данные занимают много места
- Данные необходимо использовать, чтобы они приносили выгоду
- Необходимо успевать за развитием больших данных
- Правильный баланс между экономией и наличием кластерной инфраструктуры.
- Стоимость хранения
- Уязвимость и утечка данных

Появление публичных облаков (public clouds) в 2010 году обещало решить эти проблемы. Благодаря гибким вычислительным возможностям и надежному недорогому хранилищу они стали привлекательным вариантом для создания собственных кластеров данных.

Озера данных

Джеймс Диксон (тех. директор Pentaho) придумал термин «**озеро данных**». Вместо создания изолированных хранилищ данных озеро данных должно было стать единым хранилищем всей информации компании.

Озера данных могут быть построены с использованием технологий Hadoop или с помощью объектного хранилища и служб управляемых данных, предоставляемых поставщиком облачных услуг. Делегируя работу по инфраструктуре и управлению приложениями поставщику облачных услуг, компании могут сократить объем IT-работы, связанной с выполнением задач по работе с большими данными, и сосредоточиться на управлении данными.

Ниже приведены некоторые инструменты обработки данных, которые многие поставщики облачных услуг предлагают своим пользователям:

Хранилище объектов
Позволяет организациям хранить любые типы данных в их собственном формате — это идеально подходит для создания современных приложений, требующих масштабируемости и гибкости.

Интеграция данных
Простые в использовании инструменты, которые подключаются к общедоступным и частным источникам данных, таким как базы данных и приложения, и надежно передают и синхронизируют данные с хранилищами данных в озере данных.

Подготовка данных
Визуальные инструменты для создания преобразований данных между источником и целью

Каталог данных
Инвентаризация общекорпоративных информационных ресурсов, помогающая осуществлять поиск, исследование и управление данными в озере данных.

Потоковая передача данных
Позволяет организациям обрабатывать данные в режиме реального времени, обеспечивая отказоустойчивые операции обработки потоков, такие как фильтры, объединения, сопоставления, агрегирование и другие преобразования

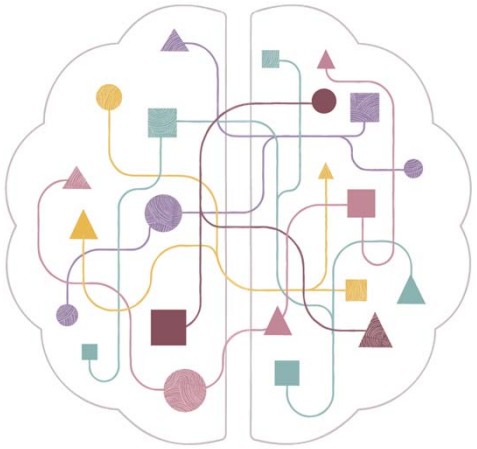
Управление данными
Hadoop, Spark, базы данных и инструменты запросов, которые помогают организациям управлять данными во всех хранилищах озера данных.

Аналитика
Инструменты, помогающие организациям понимать и обнаруживать тенденции в своих данных и использовать их для принятия решений.

Используя эти инструменты, компании могут создавать озера данных для своих неструктурированных данных в небольших масштабах и постоянно расширять их новыми типами данных, источниками данных и приложениями для извлечения пользы из данных.

Методы анализа больших данных

1. Описательная аналитика (descriptive analytics)
2. Прогнозная или предикативная аналитика (predictive analytics)
3. Предписательная аналитика (prescriptive analytics)
4. Диагностическая аналитика (diagnostic analytics)



Инструменты и технологии для анализа больших данных

- Специальное ПО: NoSQL, MapReduce, Hadoop, R
- Data mining
- ИИ и нейросети
- Визуализация аналитических данных
- Смешение и интеграция данных
- Статистический анализ
- Имитационное моделирование.

Примеры использования больших данных

Большие данные можно применять в самых различных сферах деятельности — от взаимодействия с заказчиками до аналитики. Вот лишь несколько сценариев практического использования – разработка продуктов, предиктивное управление обслуживанием, взаимодействие с заказчиками, обнаружение несанкционированного доступа, машинное обучение, операционная эффективность, внедрение инноваций.
Существуют сотни способов, с помощью которых большие данные могут дать бизнесу конкурентное преимущество. Здесь мы рассмотрим лишь несколько примеров того, как отрасли используют большие данные.

Финансовые услуги

. Благодаря большим данным финансовые учреждения могут

- Выявлять закономерности, указывающие на мошенничество, и оптимизировать нормативную отчетность.
- Достигать лучшего понимания тенденций рынка и потребностей клиентов, что может улучшить процесс принятия решений о новых продуктах и услугах.

Здравоохранение

Благодаря большим данным специалисты здравоохранения могут

- Идентифицировать гены заболеваний и биомаркеры, чтобы помочь пациентам точно определить проблемы со здоровьем, с которыми они могут столкнуться в будущем.
- Обеспечивать лучшее лечение и улучшать качество медицинской помощи без увеличения затрат.
- Обнаруживать потенциальное страховое мошенничество, отмечая определенные паттерны поведения для дальнейшего изучения.

Производство

Благодаря большим данным производители могут

- Прогнозировать отказы оборудования
- Оценивать производственные процессы, активно реагировать на отзывы клиентов и предугадывать будущие потребности.
- Лучше понимать поток производственных линий и определить причину задержек

Розничная торговля

Благодаря большим данным ритейлеры могут

- Прогнозировать потребительский спрос и запускать новые продукты
- Определять самых лояльных клиентов компании и таргетированно делать им специальные предложения.
- Использовать технологии прогнозирования, чтобы поддерживать наличие на полках товаров и избегать сбоев в цепочке поставок.

Телекоммуникации

Большие данные позволяют телекоммуникационным компаниям

- Определять области с избыточной пропускной способностью и при необходимости перенаправлять полосу пропускания.
- Прогнозировать общую удовлетворенность клиентов, анализируя уже имеющиеся у них данные о качестве и удобстве обслуживания.
- Улучшать понимание поведения клиентов для разработки новых продуктов и функций.

Заключение

Большие данные — это не столько отдельные фрагменты данных, сколько извлечение из них ценности и смысла. Сегодня данные поступают из большего количества источников, в большем количестве форматов и быстрее, чем когда-либо прежде. Имея правильную платформу данных, организации могут извлечь выгоду из своих данных и быстрее принимать решения на основе данных.