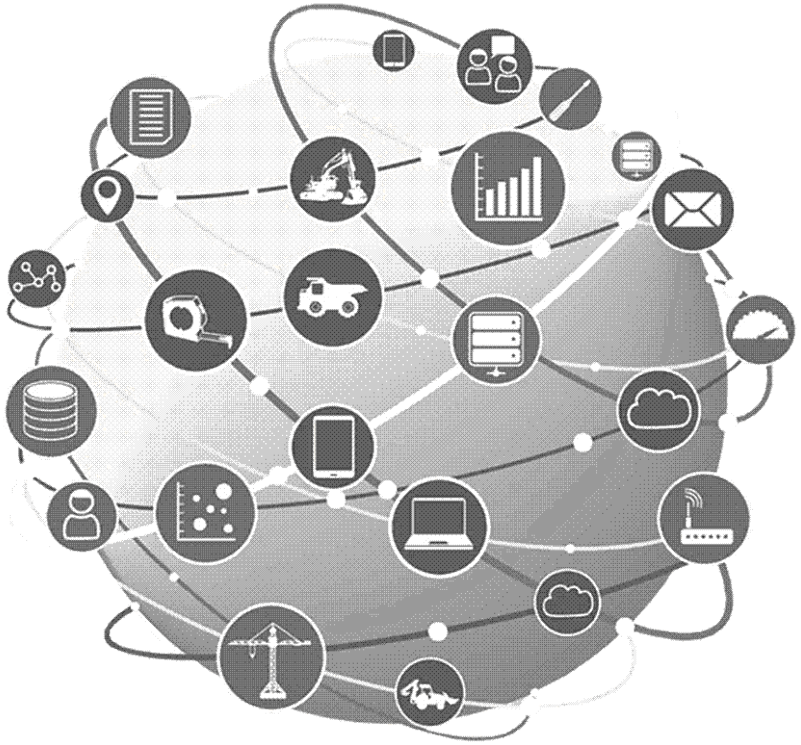
Группа компаний
NIHOL
Материалы внутренних
семинаров

BIG DATA

Большие данные



ЧТО ТАКОЕ БОЛЬШИЕ ДАННЫЕ

Big Data

ЧТО ТАКОЕ BIG DATA?

“Большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности.”
Консалтинговая компания Forrester.

Понятие “Большие Данные” подразумевает работу с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности работы, создания новых продуктов и повышения конкурентоспособности.

Big Data — это массивы информации, которые помогают принимать обоснованные решения, их ещё называют data-driven, то есть основанные на данных. Они позволяют строить прогнозные модели высокой точности.

BIG DATA – ИСТОРИЯ ТЕРМИНА

2008: Появление термина "Большие данные"

3 сентября 2008 года, вышел специальный номер научного журнала Nature, посвященный вопросу «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?».

2011: Мода на "Большие данные" расцветает

Источниками больших данных начинают рассматриваться:

- непрерывно поступающие данные с измерительных устройств,
- события от радиочастотных идентификаторов,
- потоки сообщений из социальных сетей,
- метеорологические данные,
- данные дистанционного зондирования земли,
- потоки данных о местонахождении абонентов сетей сотовой связи,
- данные с устройств аудио- и видеорегистрации

В июне 2011 года компания Gartner вводит в качестве характеристик «больших данных» три параметра (три«V») – объем (**Volume)**, скорости обмена данными **(Velocity)** и информационное разнообразие (**Variety)**.

Считается, что основной особенностью концепции больших данных является возможность получения более достоверных результатов анализа за счет обработки не репрезентативной выборки (подмножества информации), а информационного массива целиком.

BIG DATA – ИСТОРИЯ ТЕРМИНА

2013: Пик моды на Big Data

Все без исключения вендоры на рынке управления данными в это время ведут разработку технологий для менеджмента Big Data.

На январь 2013 года волна обсуждений вокруг «больших данных» превысила все мыслимые размеры. Подсчитали, что за 2012 год этот термин употреблялся около 2 млрд раз в постах, созданных около 1 млн различных авторов по всему миру. Это эквивалентно 260 постам в час, причем пик упоминаний составил 3070 упоминаний в час.

2014: Начинают развеиваться мифы о “Больших данных”

В аналитической записке осени 2014 года Gartner перечисляет ряд распространенных среди ИТ-руководителей мифов относительно Больших Данных и приводятся их опровержения. Тем не менее согласно опросу компании Assenture от 2014 года 92% компаний внедривших системы больших данных довольны результатом. Среди главных преимуществ больших данных опрошенные назвали:

- поиск новых источников дохода» (56%),
- «улучшение опыта клиентов» (51%),
- «новые продукты и услуги» (50%) и
- «приток новых клиентов и сохранение лояльности старых» (47%).

Опрошенные разошлись во мнении о том, что именно стоит считать большими данными. 65% респондентов считали, что это «большие картотеки данных», 60% были уверены, что это «продвинутая аналитика и анализ», а 50% — что это «данные инструментов визуализации»

BIG DATA – ИСТОРИЯ ТЕРМИНА

2015: Gartner исключила "Большие данные" из популярных трендов

6 октября 2015 года стало известно об исключении из отчета Gartner «Цикл зрелости технологий 2015» сведений о больших данных. Исследователи объяснили это размыванием термина — входящие в понятие «большие данные» технологии стали повседневной реальностью бизнеса и технологии входящие в состав понятия «большие данные» стали повседневным рабочим инструментом.

2016: Прогноз объединения BigData и аналитики в реальном времени

Опубликован прогноз EMC относительно развития аналитики «больших данных» на основе двухуровневой модели обработки.

Первый уровень – это «традиционная» аналитику BigData, когда большие массивы данных подвергаются анализу не в режиме реального времени.

Новый, второй уровень - возможность анализа относительно больших объемов данных в реальном времени, в основном за счет технологий аналитики в памяти (in-memory) для «аналитики на лету» с целью влияния на события, в то время, когда они происходят. Это открывало бы новые возможности для бизнеса в таких масштабах, которых раньше никто не видел.

КАК ВЕЛИКА РАЗНИЦА МЕЖДУ BI И BIG DATA?

«Бизнес-анализ является описательным процессом анализа результатов, достигнутых бизнесом в определенный период времени, между тем как скорость обработки больших данных ***позволяет сделать анализ предсказательным****, способным предлагать бизнесу рекомендации на будущее. Технологии больших данных позволяют также анализировать больше типов данных в сравнении с инструментами бизнес-аналитики, что дает возможность фокусироваться не только на структурированных хранилищах.»*

Крейг Бати, исполнительный директор по маркетингу и технологиям Fujitsu Australia

«Хотя большие данные и бизнес-аналитика имеют одинаковую цель (поиск ответов на вопрос), они отличаются друг от друга по трем аспектам.

- *Большие данные предназначены для обработки* ***более значительных объемов*** *информации, чем бизнес-аналитика*
- *Большие данные предназначены для обработки* ***более быстро получаемых и меняющихся сведений****, что означает глубокое исследование и интерактивность.*
- *Большие данные предназначены для обработки* ***неструктурированных данных****, для использования которых требуются алгоритмы для облегчения поиска тенденций, содержащихся внутри этих массивов.»*

Мэтт Слокум, O'Reilly Radar

КАК ВЕЛИКА РАЗНИЦА МЕЖДУ BI И BIG DATA?

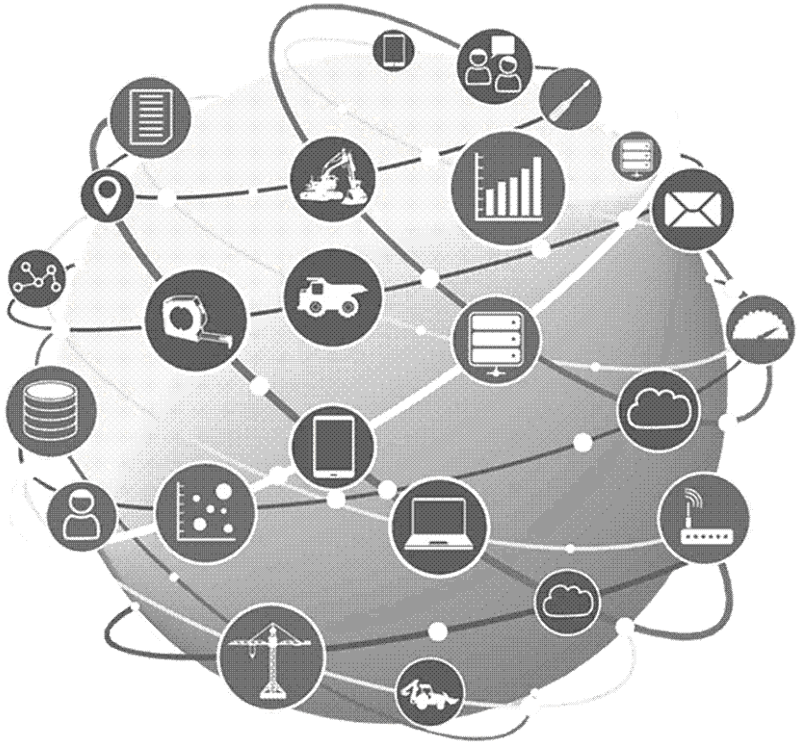
«При работе с большими данными мы подходим к информации иначе, чем при проведении бизнес-анализа.

Работа с большими данными не похожа на обычный процесс бизнес-аналитики, где простое сложение известных значений приносит результат: например, итог сложения данных об оплаченных счетах становится объемом продаж за год.

При работе с большими данными результат получается в процессе их очистки ***путём последовательного моделирования****: сначала выдвигается гипотеза, строится статистическая, визуальная или семантическая модель, на ее основании проверяется верность выдвинутой гипотезы и затем выдвигается следующая.*

Этот процесс требует от исследователя либо интерпретации визуальных значений или составления интерактивных запросов на основе знаний, либо разработки адаптивных алгоритмов "машинного обучения", способных получить искомый результат. Причём время жизни такого алгоритма может быть довольно коротким.»

Oracle Information Architecture: An Architect's Guide to Big Data



ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ

Big Data

ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?

Массовое распространение новых технологий и принципиально новых моделей использования различных устройств и интернет-сервисов вызвало проникновения больших данных едва ли не во все сферы деятельности. В первую очередь, научно-исследовательскую деятельность, коммерческий сектор и государственное управление.

10 Февраля 2011 года в журнале Американской ассоциации содействия развитию наук «Scince» авторами Мартин Гилберт и Присцилла Лопес (Martin Hilbert and Priscila López) было опубликовано исследование «The World's Technological Capacity to Store, Communicate, and Compute Information» (Технологические возможности мира для хранения, передачи и обработки информации).

Надо отметить что подход к исследованию является довольно спорным, но тем не менее оно позволяет проиллюстрировать причины роста объема данных. В качестве примеров приведем следующие цитаты:

«Мы определили хранилище как хранение информации в течение значительного периода времени для явного последующего извлечения и оценили установленную (доступную) емкость.»

«Мы получили технологическую мощность путем умножения количества установленных технологических устройств на их соответствующие характеристики. ....В результате этих ограничений наши оценки относятся к установленной аппаратной мощности компьютеров.»

ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?

«Мы не учитываем уникальность информации, потому что очень трудно отличить действительно новую информацию от просто рекомбинированной, дублирующейся информации. Вместо этого мы предполагаем, что вся информация имеет какое-то отношение к определенному человеку.»

В следующей таблице приведены объемы хранимой информации на душу населения в мегабайтах и в предположении, что данные оптимально сжаты:

Область деятельности	Тип данных	1986	1993	2000	2007
Хранение	Цифра	4	86	2 247	42 033
	Аналог	535	2 780	6 741	2 683
Вещание	Цифра	0	0	38	196
	Аналог	241	356	482	588
Телекоммуникации	Цифра	0,03	0,16	1	27
	Аналог	0,13	0,07	0,02	0,03
Всего на душу населения в МБ (оптимальное сжатие)	Цифра	4	86	2 286	42 256
	Аналог	776	3 136	7 223	3 271

ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?

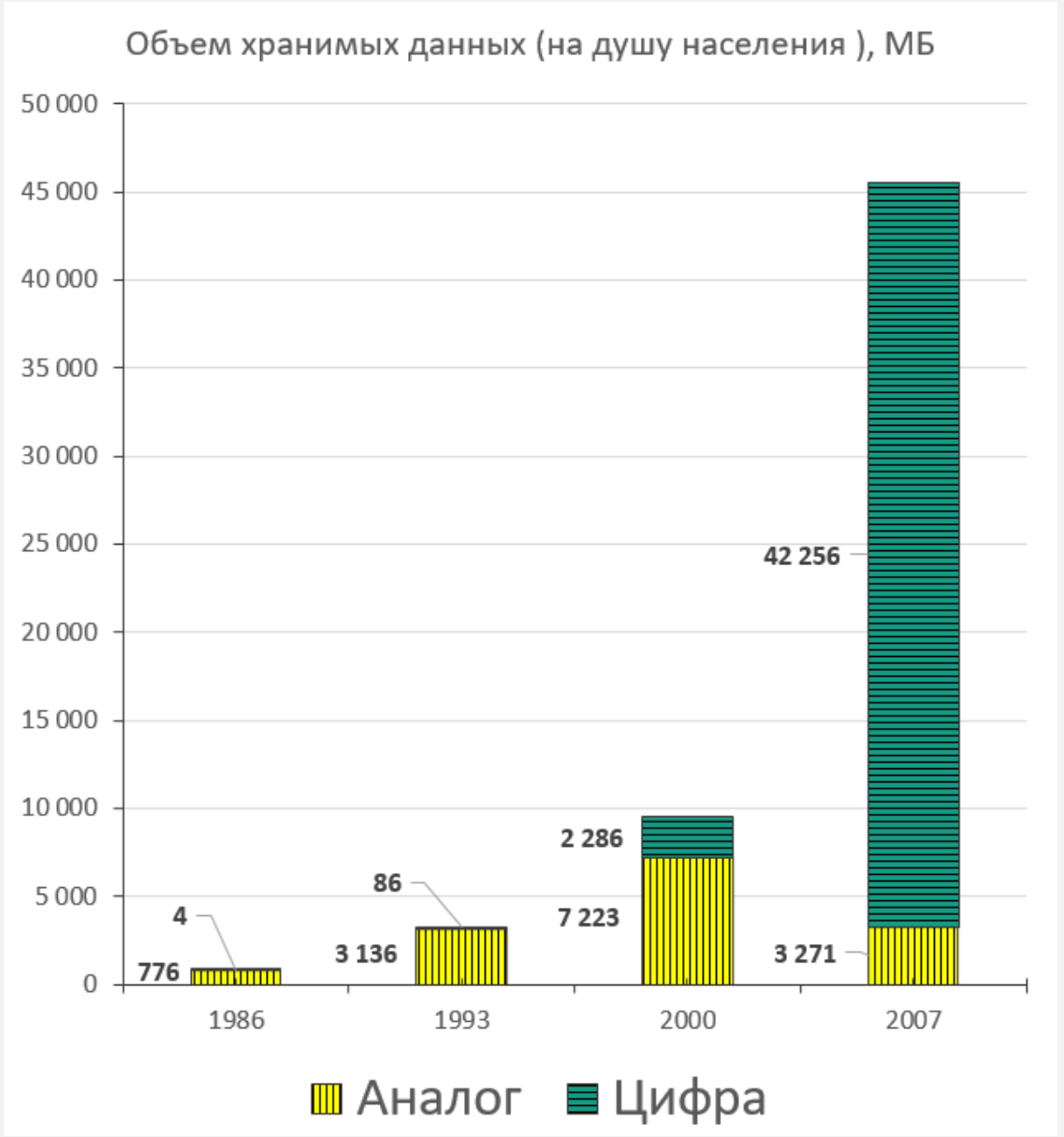
Диаграмма иллюстрирует как коррелируется рост данных хранения, вещания и коммуникаций на душу населения с переходом от с аналоговых на цифровые технологии.

Совокупный среднегодовой темп роста (Compound Annual Growth Rate – CAGR) по данным исследователей составил:

Установленный объем хранения – 22%

Объем данных вещания – 6%

Объем данных телекоммуникаций – 26%



ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?

« …теоретическая, методологическая и статистическая база наших оценок для вычислений менее надежна, чем для хранения и передачи. …не существует общепринятой теории, которая бы давала нам окончательную меру производительности компьютеров.»

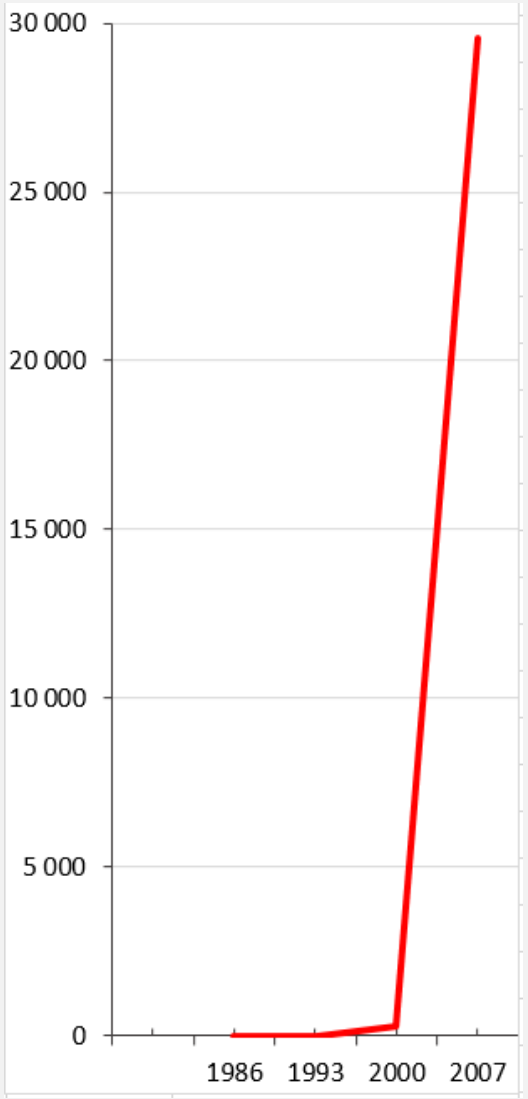
«Мы выбрали MIPS в качестве метрики производительности оборудования, которая была навязана нам реальностью доступной статистики.»

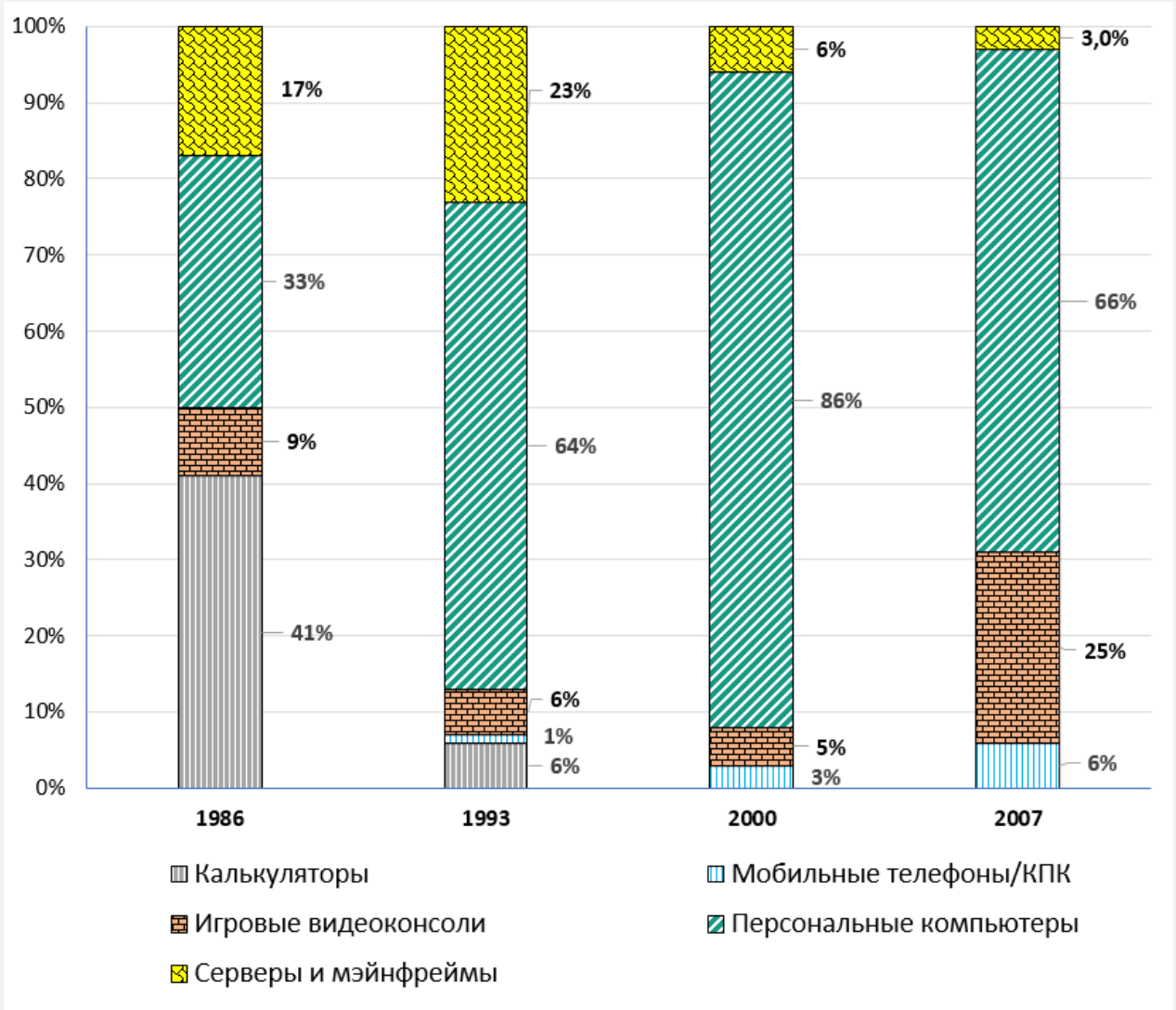
В следующей таблице приведены объемы инсталлированной мощности компьютерного оборудования на душу населения в миллионах инструкций в секунду (MIPS)

Инсталлированные вычислительные мощности (MIPS)	1986	1993	2000	2007
Калькуляторы	41%	6%		
Мобильные телефоны/КПК		1%	3%	6%
Игровые видеоконсоли	9%	6%	5%	25%
Персональные компьютеры	33%	64%	86%	66%
Серверы и мэйнфреймы	17%	23%	6%	2,7%
Суперкомпьютеры				0,3%
	100%	**100%**	**100%**	**100,0%**

«Относительно небольшая роль суперкомпьютеров (менее 0,5 %), серверов и мэйнфреймов может стать неожиданностью. Частично это можно объяснить тем, что установленная мощность, не зависит от эффективных коэффициентов использования.»

ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?





Диаграммы иллюстрируют рост установленной вычислительной мощности компьютерной техники на душу населения и изменение ее использования

ПОЧЕМУ ДАННЫЕ СТАЛИ БОЛЬШИМИ?

По оценкам, к 2025 году годовой доход мирового рынка аналитики больших данных достигнет 68,09 млрд долларов .

- Только в 2019 году глобальные соединения IoT уже сгенерировали 13,6 зеттабайт данных.
- В 2021 году во всем мире было сгенерировано 79 зеттабайт данных .
- В 2021 году 24% доходов от больших данных приходилось на программное обеспечение , 16% — на оборудование и еще 24% — на услуги .
- К 2025 году анализу потребуется более 150 зеттабайт больших данных .
- По прогнозам, к 2027 году объем использования решений для баз данных и аналитики приложений для работы с большими данными вырастет до 12 миллиардов долларов .

Источник: Big Data Statistics 2023: How Much Data is in The World?
https://firstsiteguide.com/big-data-stats/

В 2020 году каждый человек создавал около 1,7 МБ данных в секунду
Объем данных, доступных для целей аналитики, растет. В среднем обычные люди генерировали около 1,7 мегабайт информации в секунду.

Датчики, установленные на авиадвигателе, генерируют около 10 Тб за полчаса. Если все подобные данные накапливать для дальнейшей обработки, то их суммарный объем будет измеряться десятками и сотнями петабайт.

ПАРАМЕТРЫ БОЛЬШИХ ДАННЫХ

В настоящее время основные характеристики Big Data (больших данных) определяют как шесть «V»:

Volume — объём — от 150 Гб в сутки.
Velocity — скорость. Сбор и обработка данных в режиме онлайн.
Variety — разнообразие (разные объёмы и форматы, из множества разных источников)..
Veracity — достоверность. Данные собирают только из доверенных источников
Variability — изменчивость. Данные обновляются в режиме онлайн, поэтому их поток нестабилен. При анализе данных нужно учитывать все эти факторы.
Value — ценность. На их основе можно делать выводы и принимать решения.

Big Data также бывают:

- **Структурированными** — то есть уже размеченными по определённым параметрам (пример: данные медицинских показателей пациентов: температура, давление, анализы крови и ЭКГ).
- **Частично структурированными**, например файлы разного формата с записями о стихийных бедствиях в регионе за последние пять лет.
- **Неструктурированными**, например фото, музыка и сообщения пользователей соцсети

Важное отличие больших данных от обычных — распределённая структура. Управлять ими и анализировать можно с помощью множества микро сервисов

ПАРАМЕТРЫ БОЛЬШИХ ДАННЫХ

В “Big Data” данные также разделяют на персональные и обезличенные.

Персональные — это те, по которым можно безошибочно идентифицировать пользователя: например имя и фамилия, домашний адрес, номер мобильного. Им уделяется особое внимание в законодательстве, в том числе в РУз: ЗРУ-547 «О персональных данных».

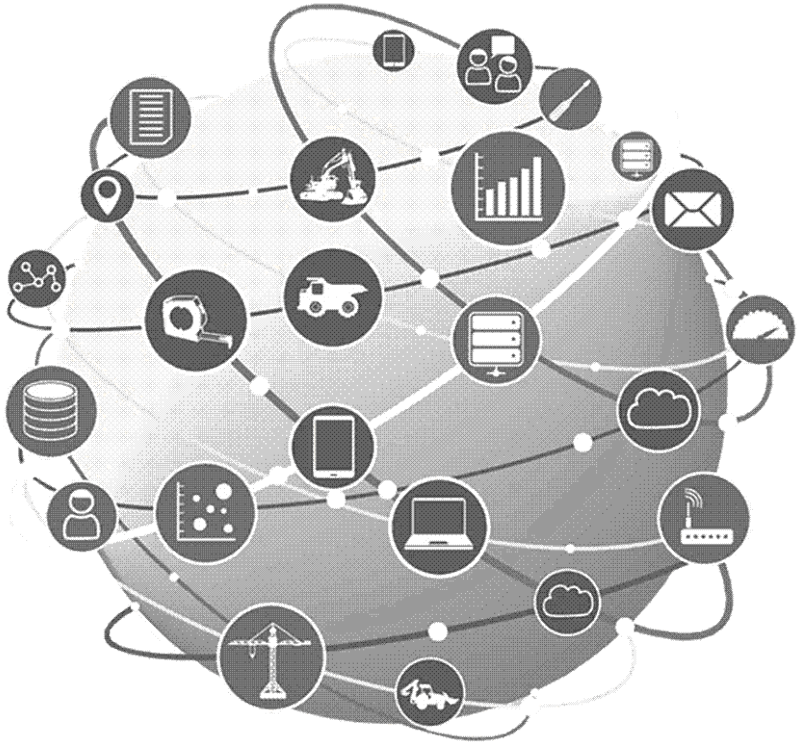
К **обезличенным** данным относят всё остальное: геолокация или список покупок без привязки к конкретному человеку и его номеру телефона и т.п..

Например в Китае действует более 200 разных законов и правил, которые призваны защищать персональную информацию — в том числе ту, что автоматически собирают приложения для смартфонов. Из-за этого в стране заблокировано большинство зарубежных соцсетей и международных онлайн-сервисов, однако эти данные доступны госорганам.

Интересно другое: если собрать обезличенные данные из разных источников и сопоставить их, тоже можно получить довольно исчерпывающую информацию.

В частности, данные такси и службы доставки помогают понять, где человек живёт и работает, какой у него примерный доход и что он предпочитает покупать. На этом и зарабатывают главные держатели больших данных в мире.

В США главные потребители и держатели Big Data — крупные корпорации: Apple, Google, Facebook, Amazon. Однако государство с каждым годом всё больше ограничивает их деятельность и монополию на сбор и хранение данных.



КАК РАБОТАЮТ ТЕХНОЛОГИИ BIG DATA

Большие данные

ГДЕ ПРИМЕНЯЕТСЯ АНАЛИТИКА BIG DATA

Транспорт. Используя больших данных о маршрутах и скорости машин навигаторы предлагают самый короткий путь с учётом пробок.

Мобильная связь и интернет. Большие данные используются для прогнозирования нагрузки на сети в каждой зоне и планирования развертывания новых базовых станций

Маркетинг. К примеру, Amazon использует систему рекомендаций товаров, которая обучена с помощью больших данных и приносит сервису до 35% от всей выручки.

Производство. Большие данные помогают организовать работу сотрудников так, чтобы снизить риски аварий и несчастных случаев в цехах.

Финансы. На основе данных обо всех случаях мошенничества банки могут создать наиболее безопасные сервисы для онлайн-платежей.

Интернет вещей. Сенсорные датчики, камеры наблюдения, системы управления беспилотными автомобилями.

Наука. Результаты исследований, опросы и показания приборов помогают выявлять неочевидные закономерности и совершать новые открытия в разных областях науки.

Государственное управление. Большие данные в виде статистики помогают лучше распределять ресурсы и реагировать на актуальные для людей проблемы.

Искусственный интеллект и роботы. Датасеты с реальными диалогами позволяют обучать голосовых и чат-ботов, которые заменяют сотрудников техподдержки или кол-центра.

МЕТОДИКИ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Методики анализа массивов данных, как правило заимствованы из статистики и информатики. Вот только некоторые из них:

- **Association rule learning** - выявления взаимосвязей между величинами в больших массивах
- **Classification** - предсказание поведения потребителей рынков (покупки, отток, объем и проч.)
- **Cluster analysis** - классификация объектов по группам за счет выявления наперед не известных общих признаков.
- **Crowdsourcing** - сбор данных из большого количества источников.
- **Data fusion and data integration** - анализ комментариев пользователей соцсетей и их сопоставление с продажами в режиме реального времени.
- **Data mining** – определение целевой аудитории продукта или услуги, предсказание поведенческой модели потребителей.
- **Ensemble learning** - использование предикативных моделей для повышается качество прогнозов.
- **Genetic algorithms** - работа с возможными решениями как с `хромосомами`, которые могут комбинироваться и мутировать.
- **Machine learning** - создание алгоритмов самообучения на основе анализа эмпирических данных (`искусственный интеллект`) .
- **Natural language processing (NLP)** - методики распознавания естественного языка человека.
- **Regression** - выявления закономерности между изменением зависимой переменной и одной или несколькими независимыми. Используется в data mining.
- **Визуализация** - Методы графического представления результатов анализа больших данных в виде диаграмм или анимированных изображений

КТО РАБОТАЕТ С BIG DATA

Специалистов, чья работа связана с Big Data, можно поделить на три большие группы:

1. **Инфраструктурные профессии:** сотрудники обеспечивающие технический сбор и хранение данных, дата-инженеры и разработчики ЦОДов.
2. **Аналитические профессии:** аналитики данных, маркетологи. Их задача — обработать большие данные, чтобы сделать сервис более удобным для пользователей.
3. **Специалисты по ИИ и машинному обучению:** используют Big Data, чтобы обучать нейросети и создавать роботизированные сервисы на их основе.

Профессии в сфере анализа данных:

Data engineer. Строит системы, которыми пользуются аналитики данных и data scientist. Он разворачивает хранилища, настраивает системы очистки и анализа данных, выдаёт аналитикам данные по их запросу и следит, чтобы всё работало нормально.

Data scientist. В основном работают с такими методами анализа, как модели, нейронные сети и визуализация.

Аналитик данных. Аналитик использует уже собранные данные, анализирует их с помощью специальных технологий, таких как статистический анализ и другие математические методы и предоставляет отчёт. На основе этого отчёта менеджеры и руководители принимают бизнес-решения. Но это не обязательное правило: аналитики часто работают с визуализацией, data scientist — со статистикой. Зависит от задач, которые определяет бизнес.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

1. **Сбор.** Большие данные собирают из разных источников. Вот некоторые из них:

Социальные — всё, что публикуют и делают пользователи в соцсетях, онлайн-сервисах и приложениях. Сюда относят фото, видео, аудио, сообщения в мессенджерах, геолокации и хештеги.

Статистические — все данные от госорганов и исследовательских компаний о людях, животных, транспортных средствах, товарах и услугах, политических и экономических явлениях.

Медицинские — данные из электронных карт о медицинских показаниях, анализах, аппаратной диагностике, вакцинациях, историях болезней.

Машинные — записи с камер наблюдения, видеорегистраторов, систем управления и умных устройств.

Транзакционные — данные о платежах и переводах через банки и другие финансовые сервисы.

В процессе сбора данные проходят очистку, или **Data Cleaning**. На этом этапе, с помощью специальных программ, данные находят, отбирают и фильтруют, проверяя на точность и соответствие заданным параметрам. Специалисты по Data Cleaning размечают массивы данных так, чтобы алгоритмам было проще находить нужные сегменты информации в ответ на запросы пользователей.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

2. Хранение - Большими данные хранят и обрабатывают с помощью облачных серверов и распределённых вычислительных мощностей. Для хранения больших данных используют:

DWH (data warehouse) — единое хранилище для всех данных, на основе которых компания формирует отчёты и принимает решения. Файлы в них сгруппированы по областям применения и расположены по хронологии. Данные в DWH них поступают по принципу **ETL** (**E**xtract, **T**ransform, **L**oad): сначала извлекаются, затем трансформируются, а потом загружаются в едином формате.

Data Lake — озёра данных, которые не имеют единого формата и чёткой структуры. Порядок действий здесь такой: извлечение, загрузка в базу и трансформация в формат, который подходит для текущих задач. Озеро данных напоминает виртуальный диск, где хранятся тексты, фото и PDF, а база данных — это таблица, где все они перечислены.

СУБД — **С**истемы **У**правления **Б**азами **Д**анных, бывают реляционными или не реляционными. Для работы с большими данными чаще используются первые — данные в них организованы в виде таблиц, которые связаны между собой ключами, а для запросов используют специальный язык — SQL. Это позволяет быстро сформировать отчёт по нескольким параметрам сразу, поскольку все они расположены в соседних ячейках.

В не реляционных СУБД данные преобразуются не в связанные друг с другом таблицы, а хранятся по другой, заранее заданной схеме. Это позволяет быстро помещать и извлекать нужную информацию из хранилища, а также запускать высоконагруженные приложения (примером не реляционной СУБД является DynamoDB - СУБД типа NoSQL от Amazon)

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

3. Обработка - Информацию большого объёма с помощью обычных инструментов обработать будет сложно: на это уйдёт слишком много времени. Для этих задач применяют особое ПО, которое работает по технологии **MapReduce**.

Сначала алгоритм отбирает данные по заданным параметрам, затем распределяет между отдельными узлами, серверами или компьютерами, а потом они одновременно обрабатывают эти сегменты данных, параллельно друг с другом.

4. Анализ - Чтобы применять большие данные в работе, необходимо анализировать их по самым разным параметрам. В этом помогают:

- SQL — язык запросов, который применяют при работе с реляционными СУБД.
- Нейросети, натренированные с помощью машинного обучения так, чтобы за секунды обработать тонны информации и представить точные данные для самых сложных задач.

Чтобы извлекать нужные сегменты информации и преобразовывать их в понятные отчёты и графики, используют специальные аналитические сервисы на базе **Business Intelligence (BI)**.

Например, **Power BI Microsoft** — сервис бизнес-аналитики, который собирает данные из CRM, Excel-таблиц и других источников, а затем представляет их в виде интерактивных отчётов.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

Что такое очистка данных?

Чаще всего большие данные собирают данные из множества разных источников. Информация получается разнородной, часто содержит ошибки, пустые значения, дубли и другие артефакты, которые помешают анализу. Это нормально — данные всегда поступают «грязными», такова их природа.

Чтобы убрать эти помехи, существует специальный процесс — очистка данных, который ещё называют **data cleaning** или **scrubbing**. Задача очистки данных — избавиться от большинства ошибок с помощью специальных инструментов и алгоритмов, сделать будущий анализ более точным.

Теоретически данные можно начать анализировать и без очистки. Однако на практике это может привести к проблемам — не соответствующим реальности графикам и отчётам или не сбывающимся прогнозам. Поэтому грамотная работа с данными подразумевает их обязательную предварительную очистку.

Очистка данных — это ещё и важный этап машинного обучения. Если модель обучить на неочищенных данных, она просто не будет нормально работать и станет выдавать совершенно неадекватные результаты.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

От чего надо чистить большие данные?

В хранилищах данных есть две сущности — записи и признаки. **Записи** — это строки таблицы, то есть объекты, попавшие в хранилище. **Признаки** — переменные в конкретных ячейках, то есть определённые значения, соответствующие какой-то записи. Ошибки могут возникать в обеих сущностях, но они разные.

Проблемы с записями:

Дублирование - Один и тот же объект попадает в таблицу дважды. Например, приходит из двух разных источников либо продублировался из-за ошибочной операции записи. Это нарушает статистику, так как удваивает значение, которое должно быть одинарным.

Не уникальные значения - Иногда конкретной записи должен соответствовать уникальный параметр. Например, у двух разных людей не может быть одинакового номера страхового полиса. Если в двух записях номер один и тот же — это ошибка. В результате обработка таких данных может дать сбой и не завершиться.

Противоречивые записи - В этом случае объект один, записей о нём несколько, а данные в этих записях не одинаковые. Например, указаны разные номера телефонов. Эта ошибка возникает при сборе данных из нескольких источников и требует корректировки — объединения или удаления. Иначе алгоритм либо воспримет объекты как разные, либо не сможет обработать их корректно как один.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

ПРОБЛЕМЫ С ПРИЗНАКАМИ

Отсутствующие значения - Какого-то значения, например номера телефона или адреса клиента, в таблице просто нет — его или не ввели, или потеряли при переносе.

Недопустимые значения - Значение есть, но неправильное. Эта ошибка часто возникает, если не настроить фильтры на сборе данных.

Орфографические ошибки и опечатки - Слово введено неправильно, это искажает статистику и не позволяет фильтровать значения.

Аномальные значения - Например, какой-то товар зачастую продаётся в количестве 10 штук в сутки, и вдруг в самый обычный день продажи резко перевалили за 1000. Такой всплеск, скорее всего, — результат ошибки, и его можно считать отсутствующим значением, потому что реальное не было записано.

Многозначность - Ошибка возникает, если для обозначения одной сущности используют синонимы. Например, где-то товар подписан как «картофель», а где-то как «картошка».

Ошибки типов, форматов и кодировок - Например, дата в одном источнике в формате 05.05.2001, а в другом 5 мая 2001 г..

Шум - Обычно возникает при обработке аналоговой информации — например показания датчиков температуры, звуки или видео. Часть данных это просто помехи, которые не имеют значения. Перед анализом данных нужно провести очистку от шума специальными методами.

Существуют и другие ошибки, которые встречаются реже или описываются гораздо сложнее.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

КАК ЧИСТЯТ ДАННЫЕ

Универсальных решений для очистки данных от всех ошибок не существует. Как правило, этот процесс — комбинация разных методов очистки данных, которые вместе позволяют последовательно уменьшить количество дублей, опечаток и других артефактов.

Методы очистки:

Удалять записи с ошибками по какому-то критерию — к примеру, оставлять последнюю и стирать все более старые (устранение для дублей или противоречивых данных).

Исправлять данные статистически - если удаление ведет к неправильному анализу, корректировать данные используя ожидаемое значение (эффективен когда данных много)

Сравнивать записи и выбирать подходящее значение – вместо аномального значения (например не уникальный номер паспорта) посмотреть на другие строки и применить значение, которое встречается чаще всего.

Применять словарь для исправления опечаток - заранее собрать все самые частые ошибки и опечатки в текстовых полях и применить словарь к данным для автоматической замены несоответствия.

КАК РАБОТАЕТ ТЕХНОЛОГИЯ BIG DATA

Способы Data Cleaning

Методов очистки существует много, в зависимости от типов ошибок, а вот способов — всего три:

1. Полностью автоматизированный, с помощью инструментов, которые уже встроены в хранилище данных. Как правило, в инструментах для хранения больших данных есть готовые наборы, которые позволяют решать простые проблемы с данными.

2. С помощью скриптов. Их пишет аналитик данных, обычно на Python. Скрипты исправляют ошибки, характерные конкретно для его хранилищ данных.

3. Ручной, когда аналитик исправляет данные вручную. Этот метод используют редко и, как правило, объединяют с другими. Например, скрипты помогают найти строки с нетипичными ошибками, которые не получится исправить автоматически. После этого аналитик уже исправит их вручную, сверив данные с базами.

Как правило, эти способы очистки данных аналитики применяют вместе, в зависимости от ситуации.

Аналитик данных должен подходить к каждому случаю индивидуально - проверить как появились аномалии, и выдвинуть гипотезы о причинах искажений. Из рассуждений и анализа вариантов рождаются методы исправления ошибок.

ОЗЁРА ДАННЫХ И ХРАНИЛИЩА ДАННЫХ

Для работы с большими объёмами данных используют хранилища данных (Data Warehouse) и озёра данных (Data Lake).

Разберем, в чём отличия понятий хранилище и озера данных на упрощенном примере:

Имеем торговую компанию, у которой много данных из разных источников: управления производством, обычные и интернет-магазины, кассовые аппараты, результаты опросов клиентов, CRM-система, записи камер видеонаблюдения и т.д.. Все эти данные нужно где-то хранить, чтобы использовать для анализа.

Использование одной базы данных (как реляционной таки и не реляционной) для этого будет не лучшим выбором по двум причинам:

- База данных плохо предназначена для работы с большими объёмами данных, поступающими регулярно и непрерывным потоком. База с ростом нагрузки начнет медленнее читать и записывать информацию, а также не сможет равномерно расширяться, или автоматически масштабироваться, если данных станет слишком много.
- Если База данных обслуживает работу нескольких основных систем бизнеса, (например, кроме перечисленных выше систем обслуживает еще и аналитику и корпоративный сайт), она может перегрузиться, из-за чего сайт начнёт тормозить или сломается. Поэтому базу для бизнеса и для аналитики нужно разделять.

ОЗЁРА ДАННЫХ И ХРАНИЛИЩА ДАННЫХ

ХРАНИЛИЩЕ ДАННЫХ (DATA WAREHOUSE).

Построено на основе распределённых баз данных — как классических, так и специальных, например типа ClickHouse. Хранилище данных содержит уже отсортированную, преобразованную и структурированную информацию. Данные из хранилища можно сразу использовать в анализе. Помещать информацию в хранилище занимает больше времени, потому что её нужно предварительно обработать и структурировать. Из-за структуры данные в хранилище занимают больше места и требуют более сложного обслуживания, поэтому само хранилище обходится дороже, чем озеро данных.

ОЗЕРО ДАННЫХ (DATA LAKE)

Разрозненные, неструктурированные данные разных форматов разместить в хранилище не получится. Чтобы решить эту проблему, придумали Data Lake, в переводе с английского «озеро данных». Это инструмент, который позволяет хранить любые данные: csv, xml, json, parquet, jpg, png, mov, mp3, pdf и другие. В него можно загружать таблицы, у которых нет чёткой структуры, то есть периодически меняется количество и названия колонок и строк. Все эти данные можно загружать в озеро без обработки, то есть практически мгновенно.

Озера данных Предназначены для хранения данных любых типов. Перед аналитикой их нужно обязательно найти, очистить и структурировать, то есть поместить в озеро просто, а извлекать — сложнее. Из-за отсутствия структуры и простого обслуживания озеро данных обходится дешевле, чем хранилище.

ОЗЁРА ДАННЫХ И ХРАНИЛИЩА ДАННЫХ

КАК УСТРОЕНО ОЗЕРО ДАННЫХ

Озеро представляет собой файловое хранилище на нескольких серверах, в котором лежат данные. Как правило данные распределены между серверами, чтобы хранилище можно было быстро масштабировать — подключить новые серверы для расширения места.

К серверам настраивают подключение разных источников данных, доступных компании. Каналы поставки данных называют пайплайнами (англ. **Pipeline** – “трубопровод”), а всю схему подключения — **ETL**-процессом. Обычно всё настроено так, чтобы данные загружались автоматически.

Хотя **Data Lake** и неструктурированное, порядок в нём всё-таки должен быть, иначе спустя время накопится огромное количество данных, в которых невозможно будет разобраться. Поэтому перед добавлением в озеро данные размечают и запоминают, откуда и в каком формате они поступили.

В итоге внутри озера данных хранятся не только сами объекты, но и метаданные, то есть информация об объектах. Это облегчает поиск, извлечение и анализ данных в будущем.

Data Lake полезны всем компаниям, которые планируют анализировать большие данные любой области, например ретейла, IT, промышленности или логистики.

Само по себе озеро данных бесполезно, потому что это просто хранилище. Чтобы с ним работать, нужны инструменты для очистки, структурирования, извлечения и анализа данных, и специалисты для работы с этими инструментами.

ОЗЁРА ДАННЫХ И ХРАНИЛИЩА ДАННЫХ

НЕДОСТАТКИ ОЗЁР ДАННЫХ

Потеря качества данных. Озеро имеет склонность становиться «болотом» — накапливать данные, которые плохо размечены и никому не нужны. Это может привести к тому, что **Data Lake** просто больше нельзя будет использовать для аналитики

Техническая сложность. Создание озера данных — непростая задача. Нужна инфраструктура: серверы, каналы связи, объёмы дискового пространства, инженеры.

Дополнительные затраты на извлечение данных. Помещать данные в озеро можно почти мгновенно, а для извлечения часто нужны сложные инструменты поиска и очистки, которые придётся настраивать отдельно. В этом плане озеро уступает хранилищу данных, в котором всё хранится по заранее проработанной структуре.

Хранение лишнего. Данные в озеро часто поступают бесконтрольно. Из-за этого в нём может быть много дублей и файлов, которые вообще не нужны ни для какой аналитики. Из-за этого озеро может разрастись и потреблять слишком много ресурсов бизнеса.

*Аналитики Gartner ввели понятие **темных данных (Dark data)**, определив их как информационные активы, которые компания собирает, обрабатывает и хранит на постоянной основе, но обычно не использует. При этом стоимость хранения и обеспечение безопасности этих данных больше, чем их ценность.*
В промышленном контексте темные данные могут включать информацию, собранную датчиками и телематическими устройствами. По мнению компании IBM 90% данных с датчиков и прочих устройств никогда не используются.

ETL-ПРОЦЕСС

КАК РАБОТАЕТ ETL-ПРОЦЕСС

Аббревиатура **ETL** (**E**xtract, **T**ransform, **L**oad) содержит всего три этапа обработки данных - Извлечение, Трансформация, Загрузка. На практике процесс обычно можно разделяется на шесть шагов:

1. **Подключение к источнику** – подключение к системе, из которой будут выгружать данные. Это делают с помощью специальных приложений, например Apache Airflow.
2. **Выгрузка данных из источника** – с использованием SQL-запросов и обращений к API внешнего источника (например получение данных из CRM, получения файлов или почты)
3. **Первичная очистка данных** – например очистка от тестовой информации или от дублей
4. **Маппинг (mapping) данных** – связывание между собой данных полученных из нескольких разных источников, не связанных между собой.
5. **Агрегация данных** — соединение данных в итоговой таблице. При агрегации получится расчёт нужных данных
6. **Загрузка данных в систему-приёмник**. Если система-приёмник — это база, то данные загружаются в таблицы. Если же аналитик работает с API, файловым хранилищем или другими сервисами, то загрузка должна проходить по заданным правилам.

ПРЕИМУЩЕСТВА ТЕХНОЛОГИИ BIG DATA

Большие данные — драйвер мировой экономики. Они помогают:

Работать с большими объёмами информации - например, базы данных миллионов пользователей социальных сетей: у каждого из них сотни сообщений, фотографий, музыки и видео.

Строить более точные прогнозы и принимать более взвешенные решения - например, планировать рекламную кампанию, опираясь на информацию соцсети о миллионах пользователей и цифровом следе каждого из них — браузер, настройки, посещаемые сайты. В результате, например, показывать рекламу запчастей только тем, кто посещает автомобильные сайты.

Мгновенно реагировать на сбои и уязвимости - благодаря доступу к большим данным обо всех действиях пользователей, банки или платёжные сервисы могут сразу отследить подозрительные действия и остановить мошенников.

Строить долгосрочные стратегии - если у компании есть данные о продажах, прибыли и убытках за несколько лет, их анализ поможет планировать инвестиции, работу с персоналом и ассортиментом.

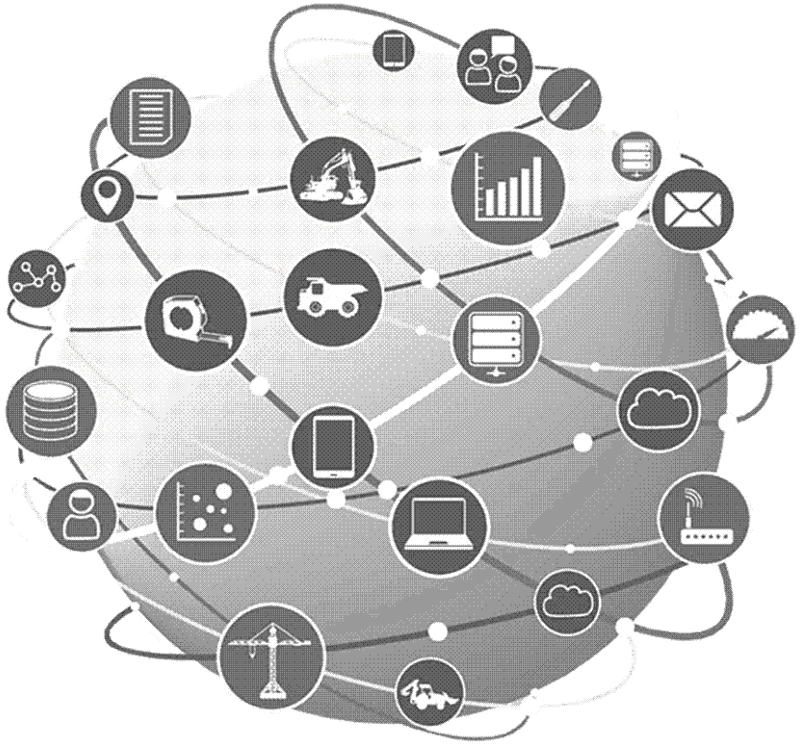
Исправлять ошибки и улучшать продукт - предположим, специалисты службы доставки заметили, что в вечернее время люди часто отменяют доставку через 30 минут после заказа. Это значит, что клиенты не хотят долго ждать и предпочитают сходить в магазин сами. Проблему можно решить, увеличив число курьеров в эти часы.

МИНУСЫ ТЕХНОЛОГИИ BIG DATA

Трудности с масштабированием – Например: на начальном этапе, сервисом стартапа пользуются 10 тыс. человек. После рекламной кампании приходит 1 млн новых пользователей. Не каждая система для хранения и обработки данных справится с таким резким притоком. Решить проблему помогают специальные облачные хранилища, которые можно масштабировать в любой момент.

Высокие риски - большие данные повышают требования к безопасности. Например, если взломают базу данных крупных банков, миллионы клиентов лишатся денег. Чтобы этого избежать, компании-владельцы Big Data используют распределённый доступ: у разных групп сотрудников разный уровень доступа и только к определённым сегментам баз данных. Кроме того, данные шифруют и структурируют на каждом уровне.

Высокие затраты - Большие данные требуют больших вычислительных мощностей, более дорогих сервисов для хранения и обработки. Для обучения нейросетей нужны огромные датасеты, которые есть только у очень крупных корпораций и часто недоступны для свободного пользования. Для работы с большими данными нужно привлекать специалистов: аналитиков данных, DWH-аналитиков, специалистов по BI.



BIG DATA - ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ

Большие данные

ФАРМАКОЛОГИЯ

Фармпроизводители (производители лекарств) стремятся получить доступ к медицинским данным пациентов и наперегонки заключают сделки с технологическими компаниями, сведущими в области анализа больших данных (Big Data) - инструмента, который открывает новые возможности для понимания того, как работают лекарства в реальной жизни

Изучение реальных свидетельств позволяет фармпроизводителям доказывать полезность и ценности своих лекарств. Наиболее активно подобные исследования ведутся в области онкологии, заболеваний сердца и респираторных расстройств.

Отследить, как действуют лекарства в повседневной клинической практике, всегда было непросто, но с цифровизацией медицины, внедрением носимой электроники и социальных сетей появляются новые возможности для сбора данных. Медицина становится более персонализированной, поэтому информация об опыте реальных пациентов, которые представляют собой более широкую выборку, чем достаточно узкий круг людей, участвующих в традиционных клинических испытаниях, приобретает особенную важность.

У всех крупнейших фармацевтических компаний мира имеются подразделения, которые занимаются сбором и использованием реальных данных о различных заболеваниях, а некоторые производители уже провели научные исследования с использованием полученной информации для более глубокого понимания механизмов действия своих лекарств. Среди таковых - исследования в области диабета компаний AstraZeneca и Sanofi, совместные изыскания Pfizer и Bristol-Myers Squibb в сфере профилактики инсульта, а также проект Takeda Pharmaceutical, посвященный заболеваниям кишечника.

ЭЛЕКТРОННАЯ КОММЕРЦИЯ

Информация о клиенте — вот за что не жалко отдать обе половины царства. И Big Data дает ответы на многие животрепещущие вопросы о заказчике: что он купил и что хотел бы купить, что ему понравилось, а что нет, когда он совершал покупки, как расплатился. И даже больше: персональные данные (адрес, пол, возраст), интересы (какие сайты посетил, кто в друзьях), активность (когда выходит в Интернет, что там ищет, какие отзывы оставляет) и многое другое.

Анализ такой информации — это шанс понять, нравится ли бренд покупателям. Готовы они покупать еще и еще или их следует немного «подтолкнуть» скидками и другими бонусами? Ответы на эти вопросы помогут создать идеального клиента. Того, который всегда готов купить товар по любой цене, активен в сообществах в социальных сетях, заинтересован в развитии бренда и рассказывает всем о понравившейся продукции

Хорошо это или плохо, но обслуживать клиентов, основываясь на их личных предпочтениях, сегодня можно только с помощью Big Data. Большие корпорации нанимают целые команды разработчиков, которые изучают их бизнес и создают уникальные приложения. Представители малого и среднего бизнеса используют более общие готовые решения. Но у всех цель одна — дать клиенту то, что он хочет, помогая тем самым e-commerce расти, развиваться и процветать.

ГРАЖДАНСКАЯ АВИАЦИЯ

Технологии Big Data применяются для выполнения ряда задач в сфере гражданской авиации. В первую очередь это ремонт и техническое обслуживание, обеспечение экономии топлива, создание цифровых двойников, оптимизация операционной деятельности (включая прогнозирование задержек рейсов), формирование персональных предложений для пассажиров и т.д.

Одним из направлений применения технологии больших данных в гражданской авиации является поддержание летной годности воздушных судов - техническое обслуживание (ТО) и ремонт. Техническое обслуживание и ремонт значительно опережают все прочие сферы по важности. С появлением прогнозного моделирования (predictive modelling) стала возможной замена деталей, которые на базе анализа определены как требующие замены, до того, как они вышли из строя, а именно во время плановых работ по ремонту и ТО.

С темой предиктивных (проактивных) ремонтов также тесно связано использование так называемых "цифровых двойников" ("digital twins"). Цифровые двойники — это виртуальные реплики физических активов, способные демонстрировать инженерам на земле работу двигателя, в то время как самолёт находится в воздухе. Чтобы сделать это возможным, на этапе проектирования и производства двигателя определяются и устанавливаются тысячи точек сбора данных. Затем они используются для создания цифровой модели, которая отслеживает и контролирует двигатель в режиме реального времени, обеспечивая необходимую информацию на протяжении всего его жизненного цикла, например, температуру, давление и расход воздуха.

ГОРОДСКАЯ БЕЗОПАСНОСТЬ

Государственные структуры могут использовать потенциал больших данных для обеспечения безопасности граждан. Как пример - инфраструктурные проекты, связанные с установкой камер видеонаблюдения. Однако данные, поступающие с камер – это только один канал данных. Качественное повышение уровня безопасности требует перехода к про-активной деятельности, позволяющей спрогнозировать преступление и заранее спланировать распределение ресурсов для его предотвращения. Это возможно при анализе исторических данных о прецедентах для построения профилей риска – условий, при которых воспроизводится то или иное событие или совершается преступление. Построение таких профилей возможно с помощью моделирования зависимости между набором характеристик, описывающих объект, и исследуемым явлением.

Например, в Лондоне пожарная служба использует социально-демографические данные для оценки и профилактики пожарных рисков. Такие показатели жителей как возраст, образование, доход, тип занятости, тип жилья и другие позволяют построить предиктивную модель, повышающую качество оценок риска пожара по районам города и спланировать географически оптимальное размещение ресурсов для ликвидации уже наступившего пожара.

Применение аналитики в области городской безопасности позволяет повысить эффективность уже существующих процессов – инспекций, патрулирования и других. Процесс управления совершенствуется непосредственно в организации, которая осуществляет управление. Решение на поверхности: данные уже есть, их нужно только добавить в анализируемый массив. Для этого не требуется дополнительных вложений в инфраструктуру.

ГЕОЛОКАЦИЯ

Глобальная база местоположения всех точек доступа Wi-Fi

Производители персональных портативных устройств (смартфонов, планшетов) уже давно научились использовать широкое распространение Wi-Fi в своих целях как вспомогательный инструмент в помощь сервису геолокации для определения местоположения устройства.

Изначально служба геолокации смартфонов применяет для определения местоположения модуль GPS. Если позиция успешно вычислена, устройство сканирует Wi-Fi-эфир и отправляет через тот же Интернет данные о географическом положении близлежащих точек доступа Wi-Fi, которые собираются в общую базу данных производителя системы геолокации операционной системы (ОС):

- для смартфонов Android — в базу данных Google;
- для смартфонов iPhone — в базу данных Apple2 (2).

Эта информация используется как приложениями Google и Apple, так и другими, установленными на смартфоне (фитнес-трекерами и др.).

Во многих зданиях и внутри метрополитена полноценно функционируют корпоративные и общественные Wi-Fi-сети. Системы Wi-Fi-позиционирования используются владельцами торговый центров, аэропортов, стадионов и метрополитенов для сбора и анализа.

Производители Wi-Fi-инфраструктуры и провайдеры Wi-Fi-услуг уже давно научились позиционировать персональные Wi-Fi-устройства в реальном времени с точностью вплоть до 1 м.

БАНКОВСКАЯ ДЕЯТЕЛЬНОСТЬ

Финансовые институты все чаще используют в своей деятельности большие данные и методы их обработки. Основными трендами в использовании больших данных являются:

Оценка кредитоспособности клиентов – традиционная область для использования больших данных. Позволяет повысить качество кредитного скоринга новых клиентов, предлагать новые услуги для клиентов, например мгновенное кредитование или выдачу заранее одобренных кредитов новым клиентам.

Маркетинг и взаимодействие с клиентами – позволяют отслеживать различные аспекты поведения клиентов и получать сведения об их предпочтениях и направлять актуальные предложения по продуктам банка, которые смогут заинтересовать клиента

Управление активами - анализ больших данных с целью оптимизации доходности своих портфелей активов

Предотвращение операционных рисков, в том числе кибер-рисков и рисков нарушения положений о противодействии отмыванию денег и финансированию терроризма. Анализ больших данных позволяет выявлять случаи мошенничества и обеспечивать кибербезопасность (анализ киберугроз, выявление подозрительной деятельности, которая может нарушить работу внутренних систем банка).

Оптимизация отчетности и других процессов - большие данные применяются для оптимизации и повышения операционной эффективности в деятельности финансовых институтов, в частности для совершения операций, взаимодействия с регуляторами и повышения эффективности работы сотрудников

И ЕЩЁ О РИСКАХ …

Большие данные

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

Использование технологий больших данных финансовыми институтами имеет значительные преимущества, однако также несет в себе существенные риски

1. Методологические риски

При использовании больших данных в риск менеджменте и оптимизации операционной деятельности возникают **методологические риски**, в том числе **риски, связанные с качеством данных**. Методологии анализа больших данных пока находятся в процессе развития. До сих пор не ясно, как лучшим образом выбрать данные, их обработать и агрегировать. Также есть вопросы относительно аналитических инструментов, которые требуются для того, чтобы интегрировать результаты анализа больших данных с информацией, получаемой из традиционных источников.

Следует отметить, что использование внешних данных, связанных с распознаванием текста, анализом связей с помощью обработки неструктурированной информации из СМИ, социальных сетей и других источников, требует новых автоматизированных подходов к управлению качеством данных для обнаружения искажений фактов или дезинформации (fact checking).

Для повышения уровня полноты и качества больших внешних данных важно, чтобы используемые данные имели происхождение из независимых друг от друга источников.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

2. Модельные риски

Еще одной смежной проблемой является **модельный риск** при построении моделей на основе больших данных – использование ошибочных исходных данных или допущений, применение модели для целей иных, чем предполагалось при ее разработке, а также ошибки при разработке самой модели.

В этой связи важен **вопрос квалификации** сотрудников, работающих с большими данными.

Неполное понимание функционирования различных форм аналитики больших данных и качества данных или моделей может привести к реализации модельного риска. Высший менеджмент, в свою очередь, должен обладать достаточными знаниями для понимания результатов моделирования и применимости моделей к задачам, для которых они используются в финансовой организации.

Реализация модельного риска может привести к таким неблагоприятным последствиям, как систематическая некорректная оценка риска при использовании больших данных в риск менеджменте. Может быть неправильно оценена платежеспособность клиентов при проведении кредитного скоринга, и данные ошибки могут носить не единичный, а массовый характер, что приведет к накоплению кредитных рисков в банковском секторе.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

3. Защита персональных данных

При использовании больших данных для клиентов финансовых институтов возникают риски, связанные с защитой персональных данных.

Большие данные подразумевают сбор различной информации о клиентах. С одной стороны, это помогает повысить качество финансовых услуг за счет кастомизации. С другой стороны, увеличиваются **риски ненадлежащего использования персональных данных** и необходимость их защиты.

Наиболее чувствительной является финансовая информация и информация о состоянии здоровья потребителей (которая может использоваться в первую очередь страховыми компаниями). Любое потенциальное ненадлежащее использование больших данных и недостаточная информационная безопасность могут подорвать доверие потребителей в долгосрочной перспективе.

При этом зачастую люди сами делятся о себе значимой информацией, в том числе в социальных сетях, не до конца понимая важность той информации, которой они делятся, и спектра тех задач, для которых используется или может использоваться информация личного характера.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

4. Риск ценовой дискриминации

Большие данные позволяют более точно оценить потребности каждого потребителя индивидуально, но также позволяют финансовым институтам оценить готовность каждого конкретного клиента заплатить за услугу определенную стоимость, что потенциально может привести к возникновению ценовой дискриминации.

То есть фактически одна и та же услуга может стоить по-разному для клиентов в зависимости от их платежеспособности или других обстоятельств.

Данная ситуация приводит к снижению излишка потребителя, особенно если клиент нуждается в быстром предоставлении услуги (например, кредита или страхового полиса), он не провел заранее анализ цен на рынке или не имеет склонности к такому анализу, исходя из истории его предыдущих покупок в Интернете.

На основе анализа IP адресов устройств, с которых покупатели выходят в Интернет, данных геолокации компании продавцы могут отследить, живет ли покупатель в богатом или бедном районе, и в зависимости от этого предлагать разную цену.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

5. Риск неценовой дискриминации

Одним из следствий недостатка транспарентности и интерпретируемости методов обработки больших данных с использованием искусственного интеллекта являются возможные проявления дискриминации (неценовой) по расовому, национальному, религиозному, гендерному и прочим признакам.

В силу недостаточной транспарентности указанных методов (в отличие от традиционных статистических моделей) дискриминация со стороны финансовой организации. может быть неумышленной, поскольку разным группам людей свойственны разные особенности поведения, в том числе в Интернете, социальных сетях, в процессе интернет покупок и так далее.

Недостаточно качественная модель может на основе этих особенностей выдать результат, свидетельствующий о более низком уровне кредитоспособности заемщика по сравнению с его реальной кредитоспособностью в результате недоучета особенностей поведения разных групп населения. Таким образом, модели, основанные на больших данных, могут быть неточными и необъективными по отношению к разным группам заемщиков.

При этом неценовая дискриминация теоретически может носить и умышленный характер со стороны финансовой организации опять же в силу недостаточной транспарентности используемых методов при работе с большими данными для регуляторов и аудиторов.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

6. Риск влияния на конкуренцию

Неоднозначным является вопрос влияния использования больших данных на конкуренцию.

С одной стороны, использование больших данных крупными финансовыми институтами, которые изначально располагают значительным объемом информации о своих клиентах и обладают достаточными ресурсами для внедрения новых технологий, **дает таким компаниям конкурентные преимущества** по сравнению с более мелкими игроками на рынке.

Большие базы данных позволяют наилучшим образом использовать элементы машинного обучения и искусственного интеллекта для скоринговой оценки заемщика, оценки риска страхователя или для предложения кастомизированной финансовой услуги, в то время как отсутствие большого объема данных часто делает эти технологии неэффективными в применении.

С другой стороны, использование больших данных **может стимулировать** конкуренцию за счет выхода на рынок финансовых услуг финтех и бигтехкомпаний, которые зачастую используют технологии обработки больших данных, что в целом создает конкуренцию традиционным финансовым институтам.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

7. Риск использования сторонних поставщиков услуг

Серьезными рисками для финансовой стабильности являются риски сторонних поставщиков больших данных, услуг по их обработке и появление, в связи с этим, новых системных рисков.

Поскольку далеко не у всех финансовых организаций есть ресурсы и компетенции для работы с большими данными силами внутренних подразделений, вовлечение в эту деятельность новых, сторонних игроков (поставщиков внешней информации, разработчиков моделей анализа больших данных, облачных сервисов) приводит к усложнению взаимосвязей на финансовых рынках.

Финансовые институты могут не иметь возможности в полной мере оценивать эти риски и управлять ими, так как данные риски выходят за рамки их организационной структуры.

Масштабы новых взаимосвязей могут увеличить сложность финансовой системы и создать новые каналы для распространения системных рисков.

Если доля нескольких крупных сторонних поставщиков на отдельных сегментах финансового рынка будет высока, то нарушения в операционной деятельности таких компаний могут привести к широкомасштабным сбоям в других частях финансовой системы или экономике в целом.

РИСКИ ИСПОЛЬЗОВАНИЯ BIG DATA

8. Риск попадания услуг в «серую» зону

Поставщики услуг в области работы с большими данными или новые типы компаний, в том числе поставщики больших данных и моделей на их основе, могут выпасть из полноценного внимания финансовых регуляторов. В результате могут появляться «серые» области финансовой деятельности, не подлежащие регулированию. Особую опасность данная ситуация будет представлять в случае, если компании, выходящие за пределы регуляторного периметра, станут системно значимыми игроками на рынке.

9. Риск отставания во внедрении Big Data

Еще одним вызовом является повышение ряда рисков для финансовых институтов, запаздывающих с внедрением технологий больших данных. Ряд финансовых институтов успешно используют большие данные для мониторинга и предотвращения реализации операционных рисков, в том числе кибер рисков, и рисков нарушения положений о противодействии отмыванию денег и финансированию терроризма.

Существует риск того, что мошенники обратят внимание на финансовые институты, которые отстают во внедрении Big Data (например малые и средние финансовые институты, у которых может не хватать ресурсов для внедрения новых технологий).

Благодарю за внимание!

Ронжин В.В.

Для компании Nihol
2023

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

- “Большие данные (Big Data)“. www.tadviser.ru. 2017
- “Воздушная математика (Big Data в мире ГА)“. www.tadviser.ru. 2019
- “Большие данные в государственном секторе“. www.tadviser.ru. 2019
- “Большие данные (Big Data) мировой рынок“. www.tadviser.ru. 2022
- “Big Data от А до Я”. Части 1-4. www.habr.com. 2015-2016
- “Wi-Fi следит за тобой, или Wi-Fi как система мониторинга”. www.habr.com. 2016
- Банк России “Использование больших данных в финансовом секторе и риски финансовой стабильности.” Доклад для общественных консультаций. Москва 2021
- The World's Technological Capacity to Store, Communicate, and Compute Information Martin Hilbert, *et al*. *Science* 332, 60 (2011); DOI: 10.1126/science.1200970