# Тема 10: Базы данных

* 1. **Виды баз данных**

# Системы управления базами данных

# Большие данные

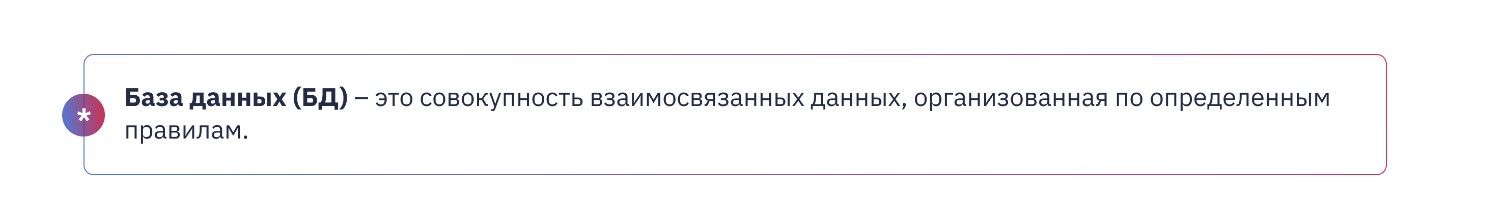
* + 1. **Виды баз данных**

В одном из прошлых разделов для иллюстрации идеи структурированных данных мы рассмотрели таблицу:

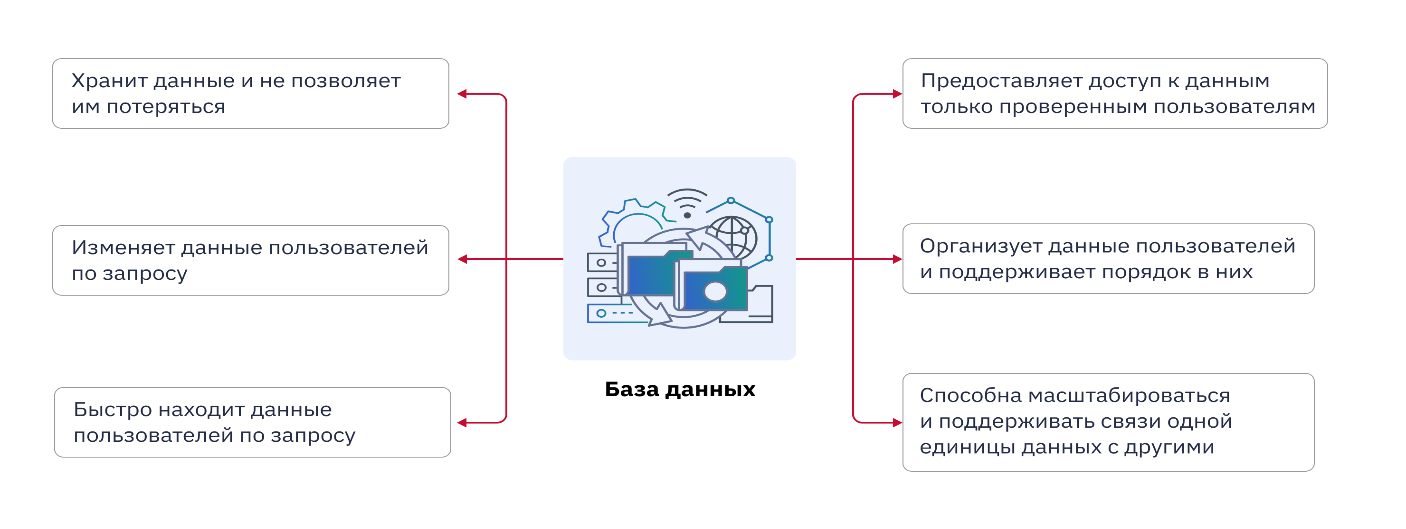


Если в такой простой таблице есть данные о паре тысяч человек, то с помощью простейших математических манипуляций можно узнать:

* демографическое распределение;
* средний и другие виды дохода;
* средние и прочие траты на кофе;
* зависимость между поколением и потреблением кофе.

Объем нового знания, которое можно получить из массива данных, зависит от целей, которые стоят перед аналитиком, и достаточности данных. Таких таблиц — однородных структур, заполненных данными, которые может прочитать машина, связанных между собой и организованных определенным способом, — может быть очень много. Все вместе они являются базой данных.  

* *В интернет-магазине, где покупатель делает заказ, данные клиента хранятся в структурированной базе данных. Затем на основе этих данных заказ комплектуется и оформляется доставка.*
* *В смартфоне все ваши фотографии, видео, переписка в мессенджерах, контакты хранятся и обрабатываются в базах данных.*
* *Порталы и электронные платформы, которые требуют регистрацию и имеют личные кабинеты, таким способом фиксируют данные пользователей. То же самое относится к соцсетям.*

Для интуитивного представления будет полезен такой образ: если у вас в квартире есть кладовка, то всё это помещение со всем содержимым может считаться базой (но не данных, а вещей или банок с огурцами).

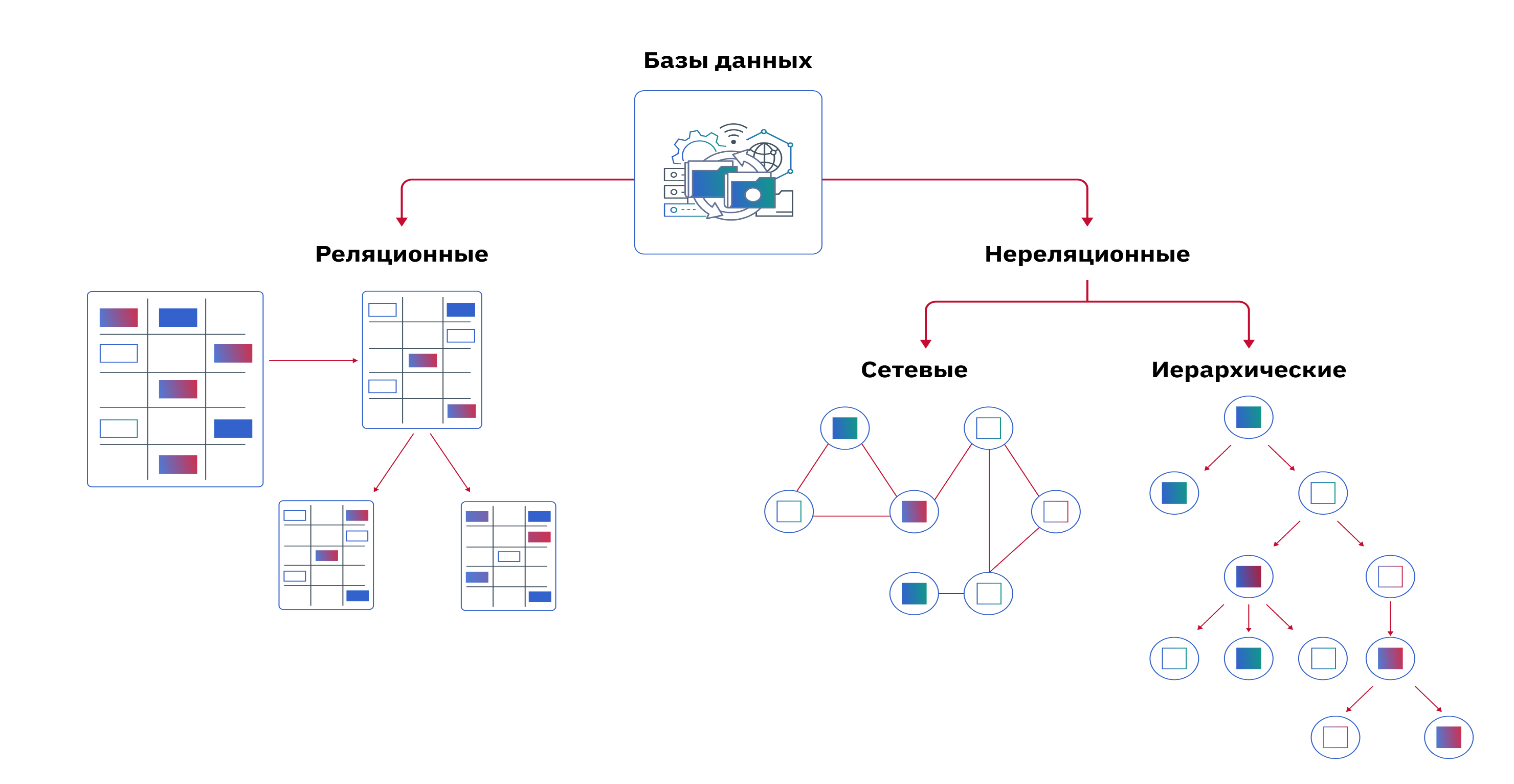
Базу можно представить как таблицу, но лишь в самом упрощенном виде. Для более сложных задач база будет представлять собой огромный склад упорядоченных коробок или даже огромный завод по фасовке данных. База данных — это обычно один отдельный файл.

Данные в таблицах не просто хранятся, с ними работают системы управления базами данных (СУБД): добавляют данные, удаляют, меняют структуру, анализируют...

Системы управления базами данных (СУБД) — это программа для манипуляции данными в базе. Используя пример с кладовкой, можно представить СУБД так: когда семья села за стол и самого младшего отправляют в кладовку за банкой огурцов, а он приносит ее и не разбивает по дороге, - младший член семьи выполняет функции СУБД.

**Виды баз данных**

Базы данных различаются в зависимости от метода организации данных. Это могут быть реляционные и нереляционные модели

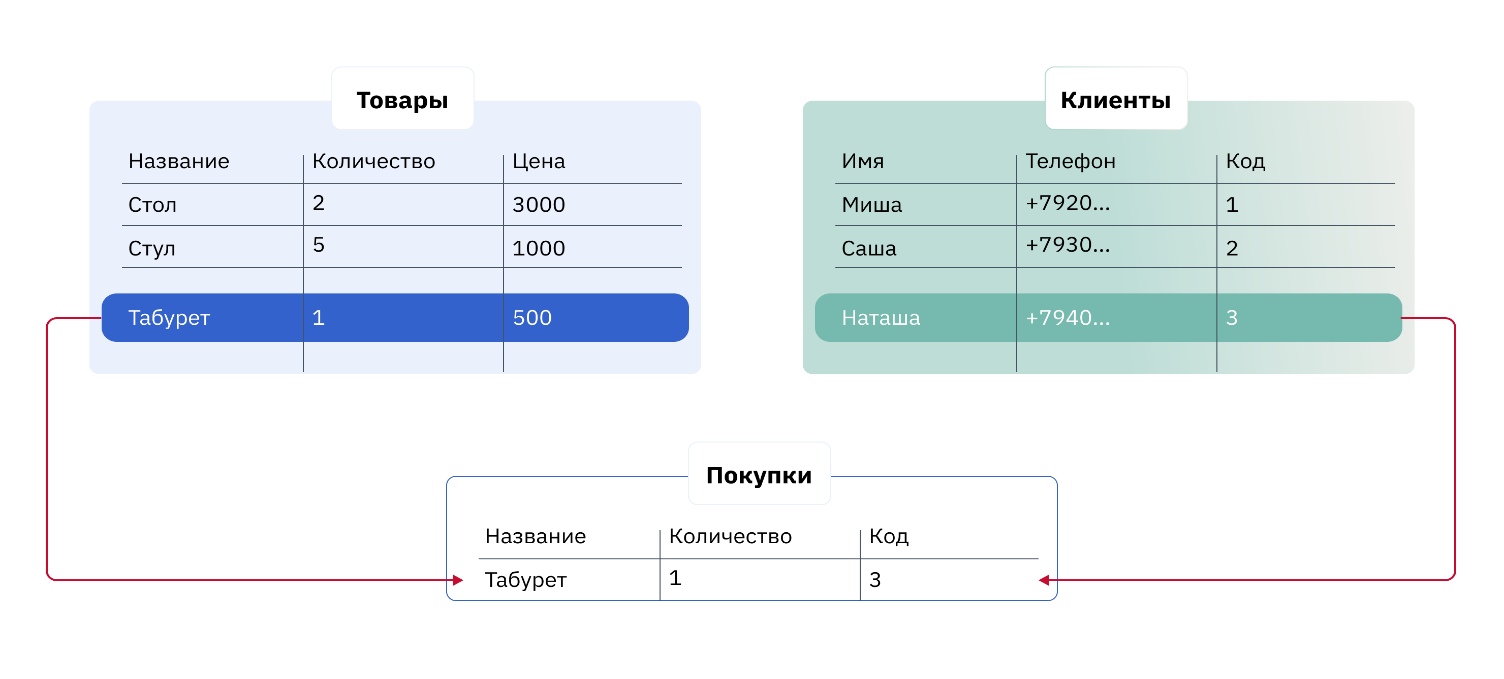


## **Реляционная база данных** https://ucarecdn.com/728a9372-60c2-4645-9632-8db49f0fc6fa/

Реляционная БД — это самая простая табличная форма, о которой мы говорили в начале статьи. Данные в ней организованы в виде таблиц, состоящих из столбцов и строк.

Табличная форма баз данных была изобретена в 1970 году в компании IBM математиком Эдгаром Коддом. Кодд доказал, что любое представление данных можно свести к совокупности двумерных таблиц особого вида, который в математике называется отношением (англ. relation) — отсюда и слово «реляционная» в переводе названия.

Отношение — это двумерная таблица с уникальным именем. Она состоит из строк (записей) и столбцов (атрибутов). Каждая строка таблицы представляет некоторый объект реального мира (например, ФИО клиента) или соотношения между объектами. Столбцы — это атрибуты (например, доход клиента) сущности, для сущности каждого типа они свои.

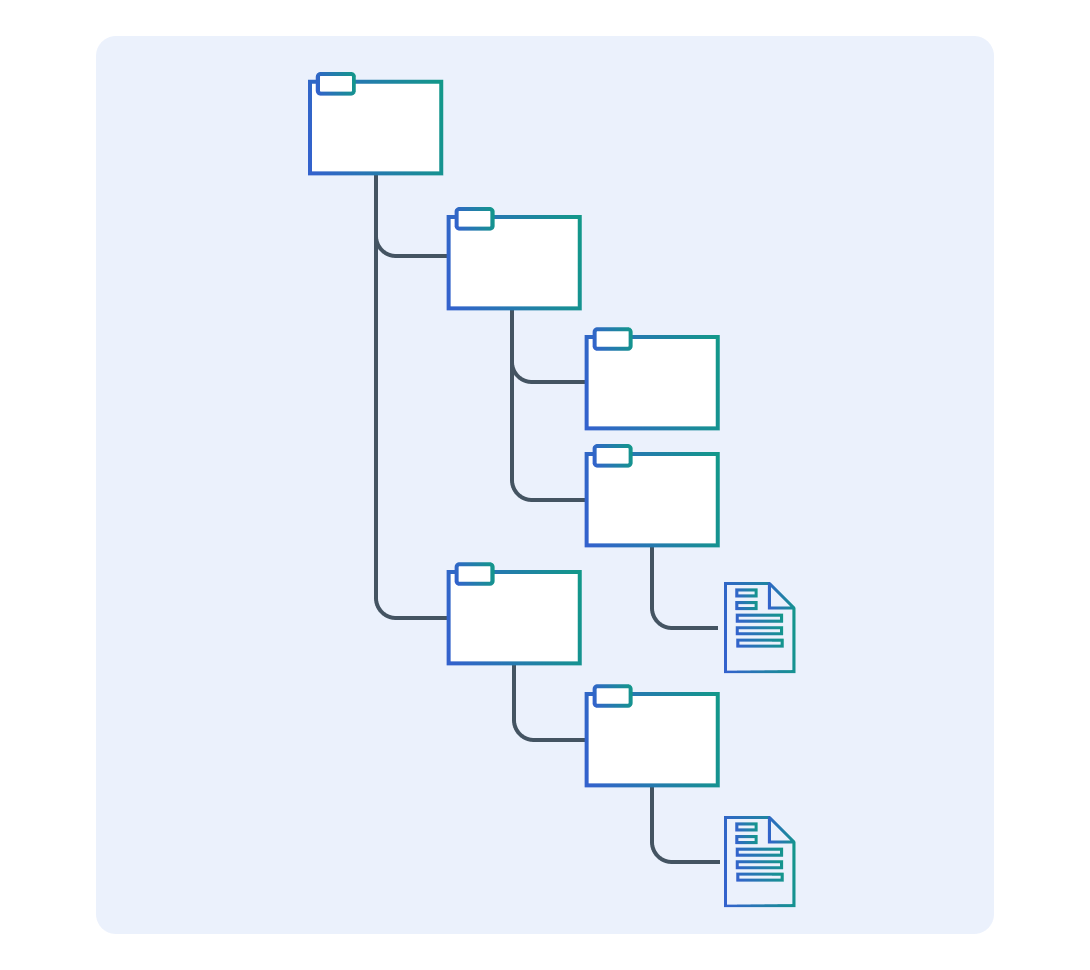
***Простая база данных магазина*****

В индустрии реляционные базы данных использовались универсально практически во всех случаях вплоть до 2000-х годов, так как являлись самым простым и эффективным видом, применимым в большом количестве случаев. Сейчас популярность набирают более сложные виды баз данных, потребляющие больше ресурсов для операций, но позволяющие решать более сложные задачи, — нереляционные модели. Дальше рассмотрим виды и примеры нереляционных баз данных, а также ситуации, в которых они используются.

## **Иерархическая база данных**

Такая база данных имеет древовидную структуру. Объекты в ней связаны друг с другом по принципу вложения одного в другое.

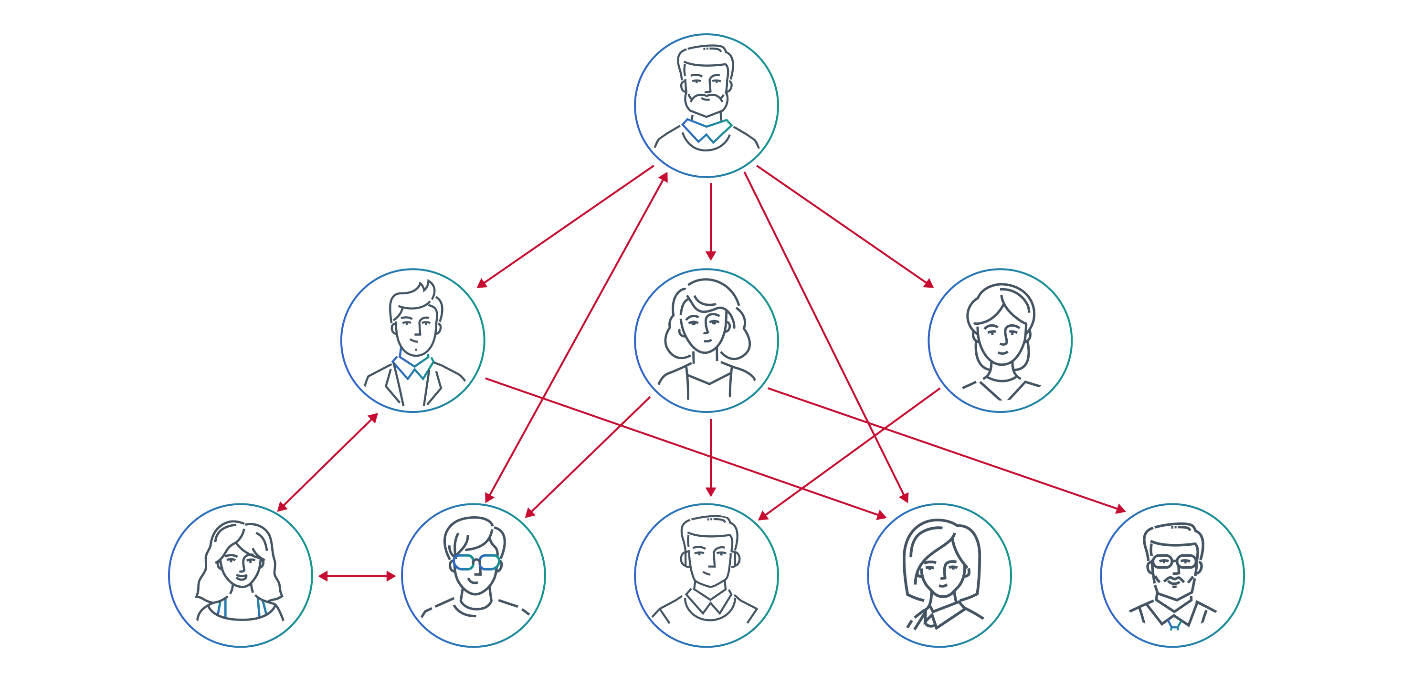
***Файловая система с директориями, поддиректориями и документами***

**

## **Сетевая база данных**

Это разновидность иерархической модели, которая хранит много связей между множеством объектов. Данные в ней представлены в виде графов.

*База данных соцсети: у вас может быть много «друзей», и у каждого из них тоже множество «друзей».*

**

Другие модели баз данных используют для более специфических задач.

**Распределенная база данных**может иметь в своем составе много фрагментов, каждый из которых хранится на разных устройствах. Обычно такая модель используется для обработки больших объемов данных, которые невозможно разместить на одном компьютере и поэтому приходится распределять по нескольким серверам.

Базы данных постоянно развиваются, появляются новые модели: с открытым исходным кодом, на облачных платформах, с многомодельной архитектурой.

https://ucarecdn.com/67ad49b8-a8d6-4c88-8325-60a09ee3e7a4/

Системы управления базами данных (СУБД) уже не просто хранят, ищут и выдают данные по запросу. Они должны:

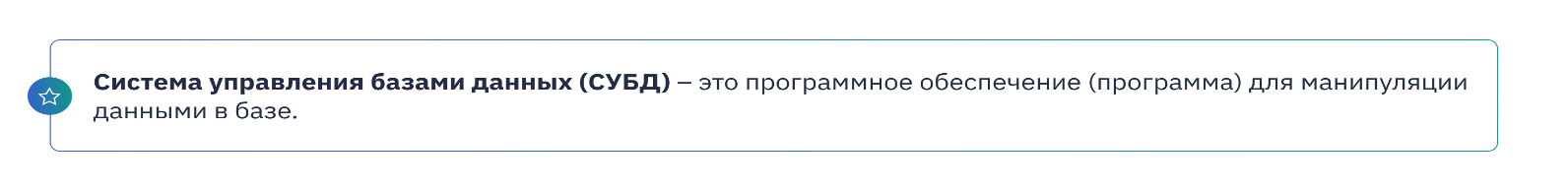
* оперативно обслуживать, настраивать и мониторить данные, постоянно растущие за счет датчиков, различных устройств и приборов;
* защищать данные от хакеров и утечек, но в то же время делать их легко доступными для пользователей;
* давать доступ к данным в режиме реального времени, чтобы бизнес и государство использовали их (а не просто накапливали).

# 2. Системы управления базами данных

Какая бы полезная для организации информация ни хранилась на ее серверах, она должна соответствовать определенным критериям. Если данные заносятся бессистемно в разные файлы и хранятся в разных форматах, то польза от затраченных на их сбор усилий будет весьма невелика.

Для ведения бухгалтерского учета в небольшом магазине вполне хватит нескольких Excel-файлов и мощности персонального компьютера стандартной модели. Поддержание порядка в данных такого объема потребует лишь некоторого уровня личной дисциплины владельца и знания базовых правил обработки информации (например, периодически сохранять таблицу и не записывать файлы поверх друг друга).

Однако когда речь идет о более крупной организации, в которой данные создаются и изменяются большим количеством устройств, поддерживать их в упорядоченном виде становится радикально сложнее. В этом случае необходимо обеспечить возможность внесения правок в один и тот же файл несколькими пользователями одновременно, регулярное резервное копирование всех файлов, обеспечение разных уровней прав пользователей. Такая система требует детальной настройки и регулярных изменений, а алгоритмы поиска информации потребляют все больше мощности. Обработка стандартного запроса приведет среднестатистический компьютер в состояние синего экрана от нагрузки. В таких случаях необходим набор программ, которые будут управлять организацией и хранением данных в базе данных, — система управления базами данных (СУБД).

Используем еще раз аналогию с кладовкой из прошлой статьи. Когда ваш склад состоит из небольшой комнаты, а единственный работник склада — это вы сами, проблем с организацией работы не возникает. Однако если склад занимает тысячи квадратных метров, на нем одновременно хранятся миллионы единиц товаров, работает сотня сотрудников и десяток единиц техники, вам потребуются правила организации работы. Это и есть СУБД.  


Основные функции системы управления базами данных:

* создание новых баз данных как сущностей;
* заполнение и редактирование баз данных;
* сортировка данных и поиск информации в базах данных;
* вывод информации из базы данных;
* защита баз данных через резервное копирование и восстановление.

В организации, оперирующей значимыми объемами данных, для поддержания всей этой работы часто существует отдельный сотрудник — администратор баз данных. Его задачи: разработка правил сохранения новых данных, обеспечение непротиворечивости структуры баз данных, администрирование системы (например, наделение пользователей правами) и обслуживание инфраструктуры. Так, каждый рядовой сотрудник должен иметь возможность вносить информацию, аналитики с разными уровнями доступа к информации — формировать новые выборки анализируемых данных. Подробно о профессии администратора баз данных можно прочитать по [ссылке](https://infostart.ru/1c/articles/1498927/).

## **Механизм работы СУБД**

СУБД принимает запросы прикладных программ и инструктирует операционную систему для передачи соответствующей информации. То есть система управления базами данных выступает как посредник (интерфейс) между базой данных, программами и пользователями.

### **Упрощенная схема работы СУБД** https://ucarecdn.com/7fe3a3f4-8b82-41f8-b6c8-258dec3f6ff9/

СУБД «общается» с базой данных на языке запросов. Запрос содержит ключевые слова, которые сформируют выборку. Самый популярный язык для работы с реляционными базами данных — SQL (Structured Query Language, «язык структурированных запросов»). SQL — один из самых простых и интуитивно понятных языков программирования, его можно в идеале освоить буквально за несколько дней. Первые успехи вы совершите уже в течение первого часа-двух. Попробовать свои силы в написании структурированных запросов на SQL можно, используя [тренажер](https://sql-academy.org/ru).

## **Свойства СУБД**

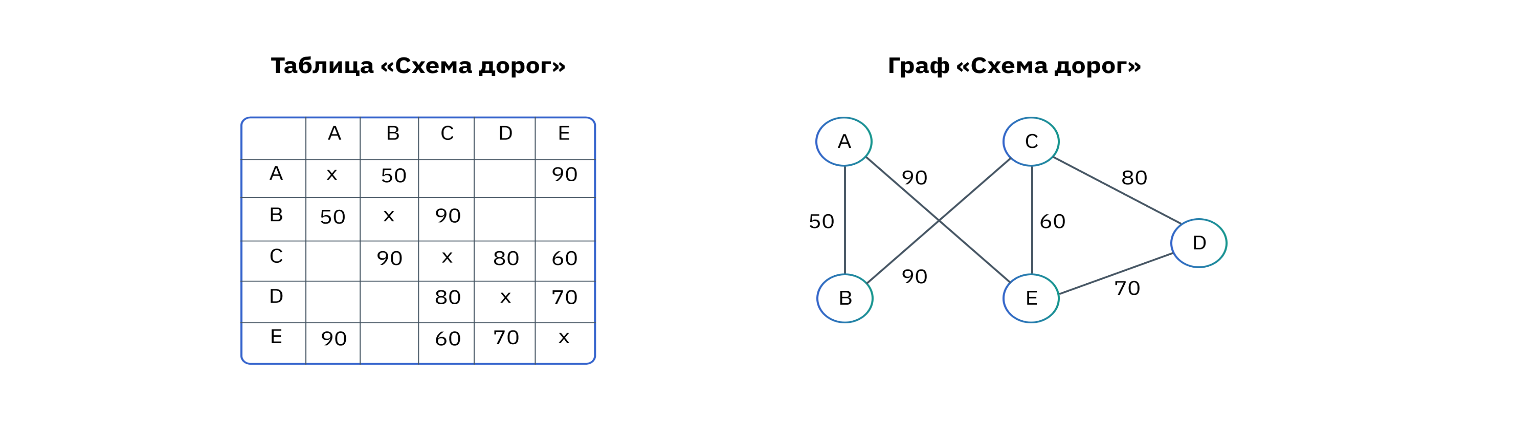
Реляционные СУБД просты, удобны и предсказуемы, а их рынок - один из самых стабильных в ИТ-отрасли. Реляционные базы остаются самым востребованным инструментом в очень разных отраслях. По данным [DB-Engines](https://db-engines.com/en/ranking_categories) за февраль 2021 года, мировая доля реляционных СУБД составляет 74% от всех используемых.

В начале 2000-х получили распространение нереляционные базы. Помимо решения новых задач (например, анализа социальных графов, распространившихся из-за появления соцсетей), в них разработчиками был сделан упор на исправление главных недостатков реляционных баз — проблем с гибкостью, низкой производительностью и масштабируемостью.  
  


Масштабируемость — способность системы выдерживать увеличение рабочей нагрузки при добавлении ресурсов. Реляционную базу данных (то есть набор таблиц) трудно масштабировать горизонтально, то есть распределять таблицы по разным устройствам (серверам). При достижении определенного потолка в объеме данных построение запросов и связывание таблиц начинает занимать слишком много времени.

Поэтому растущую базу приходится помещать на более мощный и дорогой сервер, то есть масштабировать вертикально. Но возможности даже самой мощной машины ограничены, поэтому реляционные базы плохо приспособлены для хранения действительно больших данных.  
  
Из-за модели организации данных в реляционной базе она имеет объемную структуру. Так, если в одном файле нужно отобразить наличие дороги и расстояние между всеми  1117 городами России, то такая таблица должна иметь

1117 строк и 1117 столбцов.



При этом на практике каждый город обычно соединен дорогами не более чем с 3–5 другими городами. То есть каждая строка, обозначающая город, будет иметь только 3–5 заполненных значений, а остальные 1110 ячеек будут пустыми. Такая таблица будет состоять из пустых ячеек более чем на 99%. Однако алгоритмы перебирают все ячейки таблицы, в том числе пустые, поэтому на обработку любого запроса (например, «покажи длину дороги между Москвой и Петербургом») компьютер потратит огромное количество излишней энергии и времени, перебирая все 1,2 миллиона ячеек таблицы.

Скорость обработки запроса зависит от числа таблиц, к которым запрос обращается. Представьте себе, что таких таблиц 100, 200, 1000, — тогда СУБД будет выдавать результат крайне медленно, а код запроса будет очень громоздким.

Из этого примера видно, что обработка графовых данных будет происходить в сотни раз более эффективно с использованием нереляционной СУБД (то есть той, где данные организованы не в форме почти полностью пустой таблицы, а иным способом).

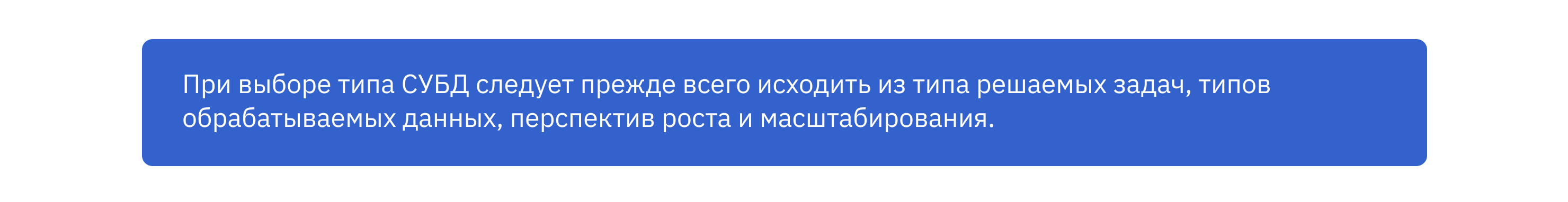
Одним из инструментов при проектировании баз данных является нормализация данных — устранение избыточности информации, при которой каждый факт должен храниться только в одном месте. Основным принципом нормализации является разбиение таблиц на более мелкие, которые будут быстрее добавлять, удалять и редактировать данные. Это означает, что увеличивается количество таблиц, нужно выполнять индексацию и устанавливать по ключам связи между таблицами. Отметим, что существует также обратное понятие — денормализация, то есть объединение в одну таблицу. Существуют ситуации, когда работа с одной громоздкой таблицей будет предпочтительнее, чем с огромным количеством малых.

## **Наиболее популярные современные СУБД**

* **С открытым исходным кодом и бесплатные:**  
  MySQL и PostgreSQL
* **Коммерческие:**  
  СУБД Oracle Database, Microsoft SQL Server
* **NoSQL:**  
  MongoDB, Redis и Cassandra

### **NoSQL**

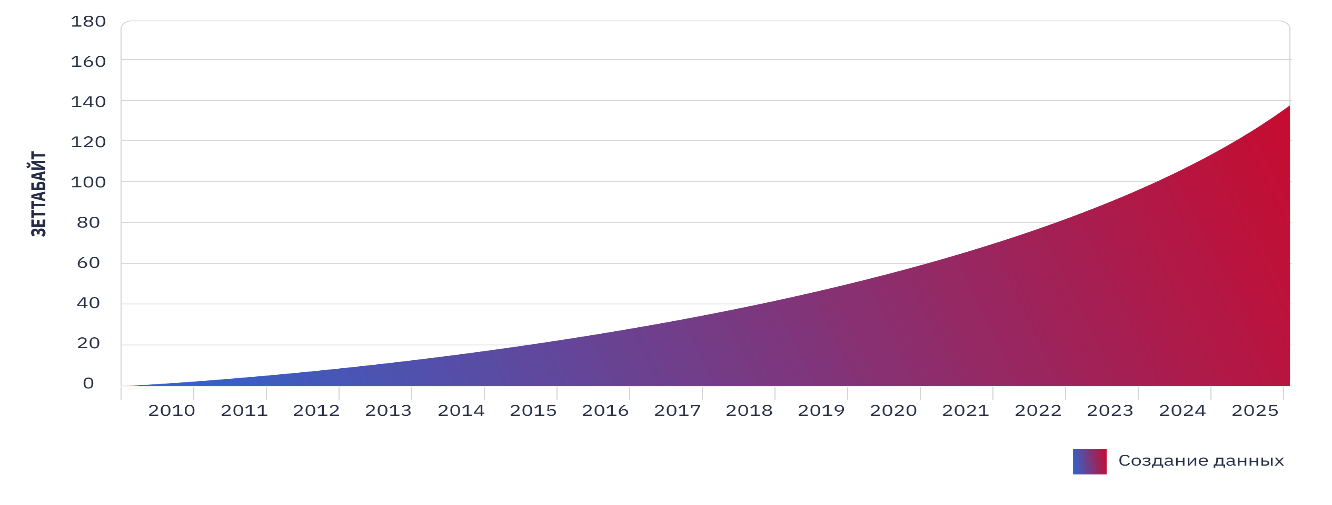
NoSQL (от англ. not only SQL — «не только SQL») - это обозначение широкого класса разнородных систем управления базами данных, появившихся в конце 2000-х — начале 2010-х годов и существенно отличающихся от традиционных реляционных. Эти СУБД настроены на конкретные модели данных и шаблоны доступа. В них смягчены жесткие требования к непротиворечивости данных, присущие реляционным СУБД. Производительность достигается за счет горизонтального масштабирования с использованием распределенного аппаратного обеспечения, без добавления дорогостоящих серверов.

В NoSQL нет таких понятий, как строки, столбцы, таблицы и их соединения. Данные в нереляционных базах хранятся как объекты с произвольными атрибутами: это могут быть пары «ключ—значение», документы в формате [JSON](https://ru.wikipedia.org/wiki/JSON), графы.

# Большие данные

Автором термина «большие данные» принято считать редактора американского научного журнала Nature Клиффорда Линча, подготовившего в сентябре 2008 года специальный выпуск на тему «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?». В этом номере были собраны материалы о феномене взрывного роста объемов и разнообразия обрабатываемых данных и технологических перспективах в случае вероятного скачка «от количества к качеству».

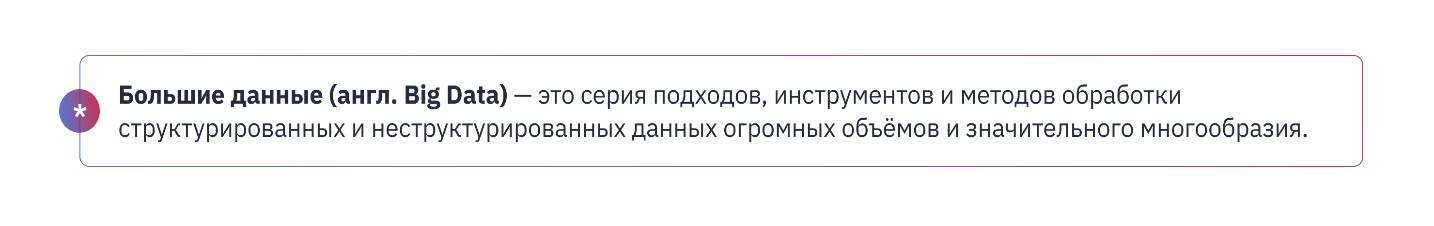
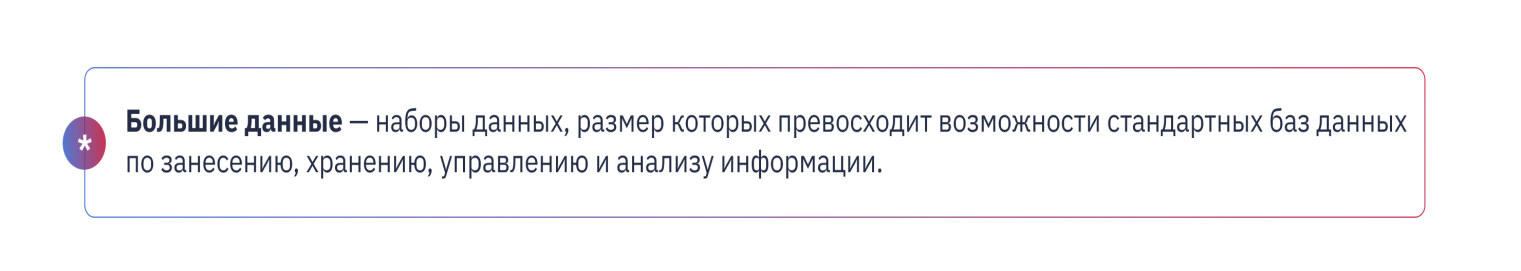
### **Динамика объема генерируемых в мире данных**



Изначально в выпуске журнала 2008 года речь шла исключительно об обработке научных данных: как ученым успевать анализировать всё растущие объемы поступающей информации. Однако обсуждаемое явление и термин, который его описывал, привлекли огромный интерес в общественном сознании, и поэтому быстро были перенесены и в сферы информационных технологий и бизнеса. Актуальность явления косвенно подтверждает и тот факт, что уже в 2013 году «большие данные» появились в названиях университетских программ, в первую очередь по вычислительным наукам.

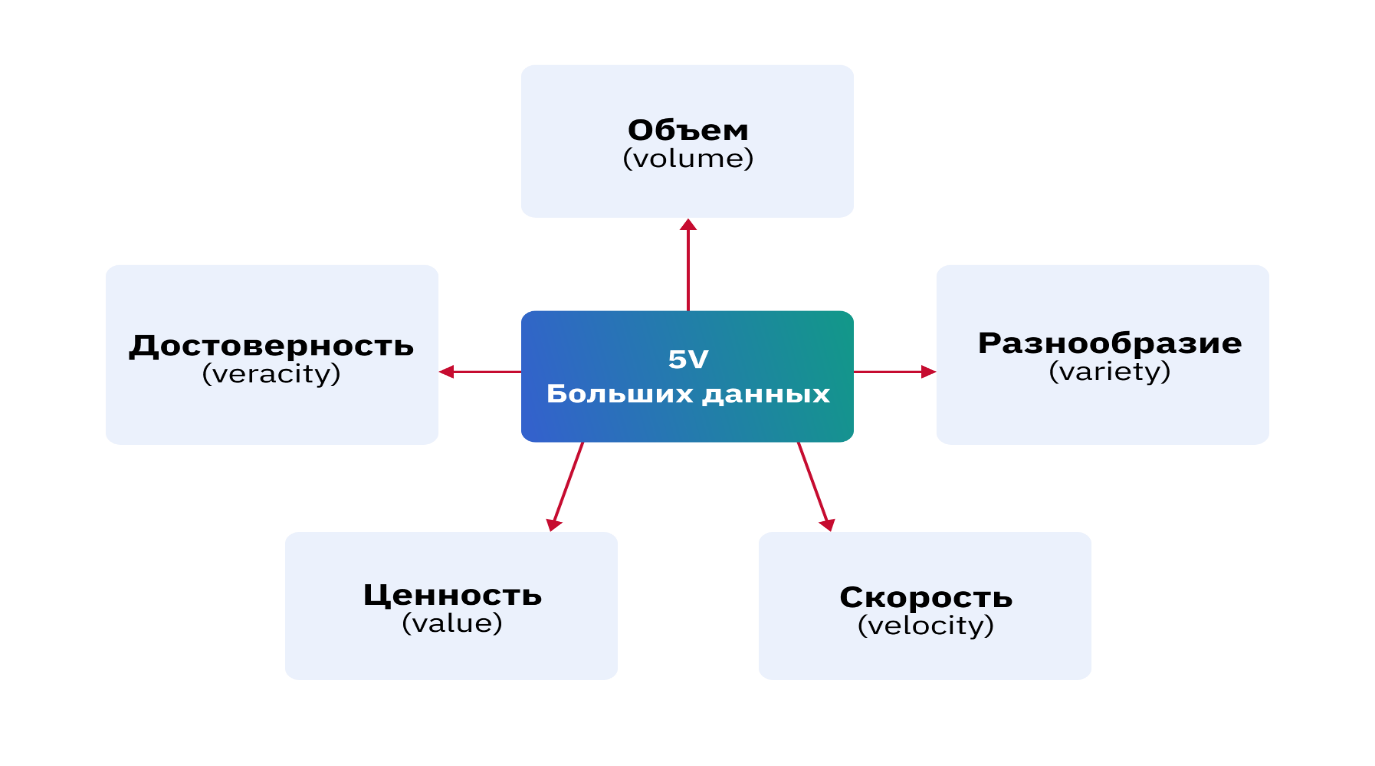
## **Понятие больших данных**

Термин «большие данные» в зависимости от контекста может использоваться в нескольких смыслах. Наиболее строгим и научно корректным можно считать следующее определение:

Подчеркнем, что в этом определении речь идет именно о наборе математических методов, о технологиях обработки, а не о самих данных, над которыми производятся операции. Попытка формализовать описание именно данных приводит к другому определению:

Очевидно, что с помощью этого определения будет сложно однозначно отделить большие данные от «малых». Так, например, является ли большими данными файл размером 100 гигабайт? Часто используется следующий неформальный критерий: большими данными являются те, которые невозможно обработать на стандартном персональном компьютере. Но и в этом случае граница, отделяющая большие данные от «малых», может колебаться между десятками гигабайт и несколькими терабайтами.

Одним из методов определить, какие данные являются большими, является их соответствие трем либо пяти критериям. Каждый из этих критериев на английском начинается на букву V, из-за чего закрепилось название «правило 3V» или «правило 5V».



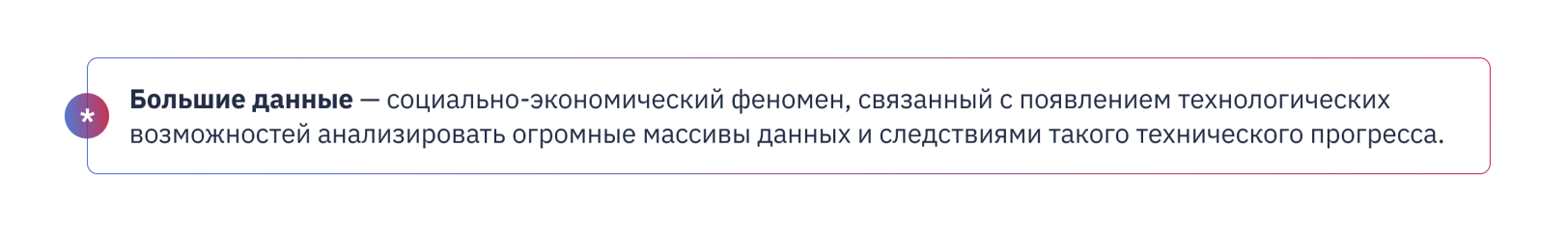
  
  
Данные растут в геометрической прогрессии. При этом организациям требуется анализ всего множества данных, а не только фрагмента или выборки.

  
  
Разнородные неструктурированные данные невозможно уместить в реляционной базе данных, их сложнее хранить и анализировать.  
  
  
  
Данные изменяются с высокой скоростью, на первый план выходит скорость их чтения с носителей, передачи и обработки.



Хранить и обрабатывать большие данные дорого. Вложения в работу с ними должны быть оправданы, то есть данные должны приносить организации пользу.  
   
  
  
Не все данные могут быть абсолютно достоверными, но их «чистота» должна быть достаточна для эффективного анализа.

https://ucarecdn.com/a82d8b15-347a-44ea-bab4-c28d60c1c0d8/

Часто термин используется и в более широком, абстрактном смысле:

Большие данные используются бизнесом и государством для получения информации о потребностях, ресурсах и возможностях клиентов, принятия решений об изменениях, разработке и внедрении новых продуктов.

Источниками больших данных могут быть клики в смартфоне, поисковые запросы, электронные приборы, которые нас окружают. Большие данные могут происходить из внутренней информации предприятий и организаций, часто являются результатами научных наблюдений (например, в сфере медицины). Классическими источниками больших данных являются интернет вещей и социальные медиа. Этот факт объясняет актуальность явления сейчас, но не 20 лет назад, когда массового интернета вещей и социальных сетей еще не было.

Распространение явления больших данных стало возможным в последние 15—20 лет из-за роста вычислительной мощности компьютеров и падения цены обработки информации. Многие процессы, которые описываются большими данными, происходили и раньше, но фиксировать и обрабатывать данные о них было экономически необоснованно.

## **Использование больших данных**

Существует огромное количество способов применения больших данных, и регулярно изобретаются новые.

Так, поисковые системы и онлайн-магазины накопили огромный объем данных о поведении каждого своего клиента в интернете. Они знают, чем человек интересуется в последнее время, и могут использовать эти данные, чтобы показывать рекламу, соответствующую его интересам. Поисковику известно, как человек перемещается по городу, в какие магазины заходит, что покупает, что ищет в интернете, сколько минут проводит в социальных сетях. Сбор и анализ большого количества «точечных» данных позволяет делать выводы о запросах и предпочтениях каждого отдельного потребителя. Используя эти новые знания, становится возможно создавать индивидуальные предложения для каждого отдельного потребителя.  
  


*В начале 2010-х годов огромный резонанс в СМИ получила «система прогнозирования беременности» американского интернет-магазина Target.*

*В одно из отделений сети в Миннеаполисе обратился крайне раздраженный мужчина средних лет. Его 16-летняя дочь получила от магазина по почте несколько комплектов рекламы товаров для беременных, что вызвало возмущение у мужчины консервативных взглядов. Менеджер магазина искренне попросил прощения и убедил гневного отца, что произошла ошибка. Однако через несколько недель мужчина обратился в тот же магазин с извинениями: оказалось, что его дочь действительно была беременна. Но в этот раз он хотел знать, как интернет-магазин узнал о беременности раньше, чем дочь рассказала кому бы то ни было.*

*Проведенное расследование показало, что набор буклетов был отправлен по почте автоматически на основе выгрузки из ИТ-системы онлайн-магазина.*

*Каждый человек имеет устоявшиеся шаблоны потребительского поведения. Интернет-магазин фиксирует многие детали поведения (на товары каких категорий кликает клиент, сколько времени изучает страницу каждого из них), и на основе анализа этих данных изменяет тот список товаров, что будет показываться конкретному клиенту. Если ИТ-система магазина будет отображать рекламу, актуальную конкретно для этого клиента, то он сделает заказ с большей вероятностью. И магазин получит больше прибыли. То есть онлайн-магазин имеет прямые мотивы собирать о клиентах максимум информации.*

*Если шаблон потребительского поведения клиента в какой-то момент изменяется, то система начнет относиться к нему соответствующе, то есть предлагать товары, типичные для клиентов другой, новой категории. Так, если вместо покупки яркой косметики и туфель на каблуках девушка начинает регулярно заказывать одежду мягких тонов и наборы витаминов, система сделает вывод, что покупательница перешла из одной категории (условно «бунтующие подростки») в другую («молодые мамы»), - и отобразит иной набор рекламных предложений.*

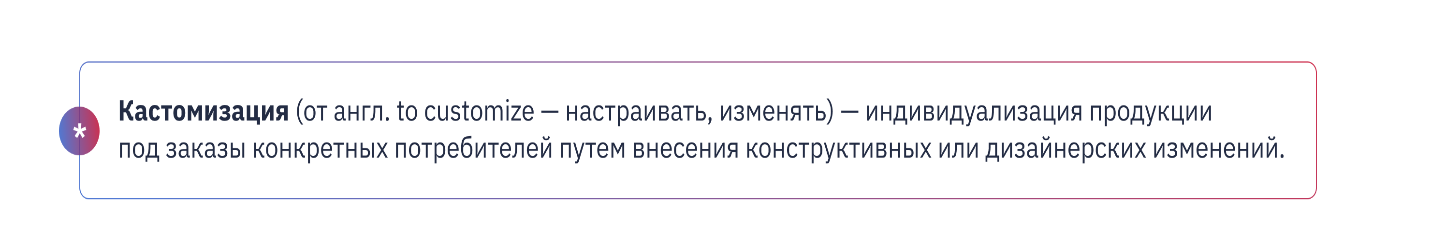
*В данном случае расследование показало, что причиной предложения послужил следующий факт: большая доля беременных женщин в какой-то момент покупала определенный бренд лосьона с кокосовым маслом. Из этого система сделала вывод, что если девушка покупает этот лосьон, то есть большая вероятность ее беременности. Значит, показывая такому клиенту рекламу других товаров, актуальных для беременных (например, детской одежды), магазин увеличит вероятность оформления заказа.*

*Как оказалось, магазин не использовал никаких персональных данных девушки и, следовательно, не нарушил закона. Самым ярким результатом этого прецедента стала именно его публичность: он привлек широкое внимание общества к феномену использования больших данных.*

*Оригинал статьи доступен на*[*сайте New York Times*](https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all)*, а пересказ истории на русском языке - по*[*ссылке*](https://habr.com/ru/post/147284/)*.*

Важно отметить, что для того чтобы установить связь между покупкой лосьона и беременностью, исследователям компании Target потребовалось проанализировать данные о поведении сотен тысяч клиентов, перебрать огромное количество комбинаций совместных покупок. Связь лосьона и беременности не была единственным практическим выводом такого исследования, но в силу публичности стала наиболее ярким из них. Очевидно, что при исследовании шаблонов потребительского поведения существует значительная погрешность (то есть не все, кто увидит рекламу детской одежды, действительно будут беременны). Однако если система обслуживает сотни тысяч клиентов и опирается на сотни шаблонов вроде «лосьон — беременность», то итоговая выдача рекламы будет в среднем довольно точной.

Именно так, извлекая из данных новое знание и используя его для каких-то действий (для отображения таргетированной рекламы), бизнес получает пользу от больших данных. И, важно отметить, для коммерческих компаний дополнительная прибыль от использования технологии должна превышать все расходы на сбор, анализ данных и эксплуатацию системы.

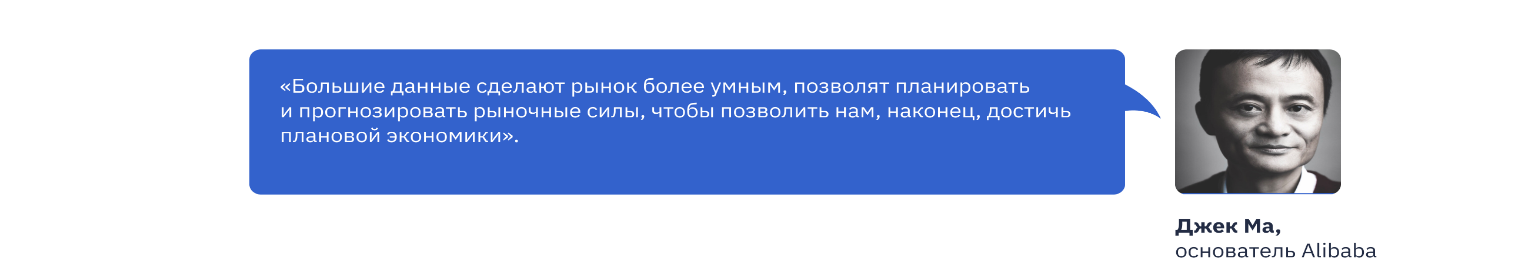


Так, использование технологий больших данных сделало возможным даже создание индивидуальных медицинских препаратов (с индивидуально рассчитанной дозировкой активных веществ) для пациентов на основе данных их анализов.

## **Сочетание плановой и рыночной экономики**

«Революция в больших данных может возродить плановую экономику» —  именно так называлась самая популярная статья в Financial Times, одной из самых авторитетных деловых газет в мире, в сентябре 2017 года.

По словам Джека Ма, основателя Alibaba, большие данные вместе с искусственным интеллектом углубляют наше понимание окружающего мира в реальном времени. Так он сказал на [экономической конференции](http://www.globaltimes.cn/content/1051715.shtml) в прошлом году:



Цифровые платформы являются мощным инструментом централизации и управления информацией. Большие данные позволяют совместить макроэкономическую координацию (уровень целых отраслей экономики) с микроэкономической (поведение отдельных людей и компаний). Платформы способны собирать огромные объёмы информации почти мгновенно и и одновременно с этим отслеживать индивидуальные предпочтения. На основе этих данных возможно и принятие стратегических управленческих решений, в том числе автоматическое, что позволяет повысить качество решений.

# Примеры использования больших данных

Большие данные хранятся и используются государством и бизнесом для получения информации о потребностях и возможностях клиентов, принятия решений об изменениях, разработке и внедрении новых продуктов. Американская консалтинговая компания Forrester предлагает следующую формулировку: «Большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности».

Рассмотрим несколько сценариев практического использования больших данных.

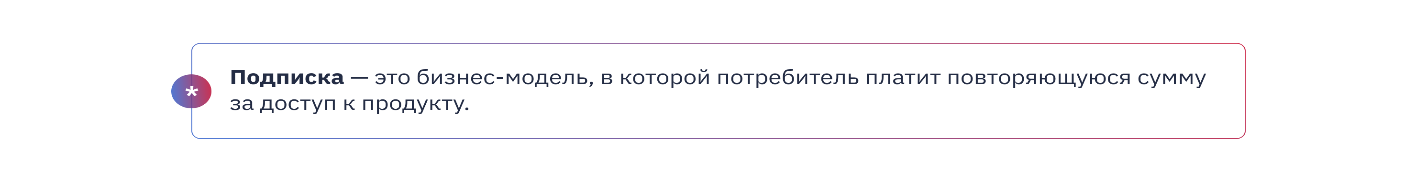
## **Предиктивное обслуживание**

В промышленности механические устройства регулярно дают сбои под действием различных физических факторов или износа. Однако если с помощью различных датчиков отслеживать состояние деталей оборудования, то замену «подозрительных» деталей можно обеспечить еще до того, как они сломаются (англ. predict — предсказывать). Таким образом можно значительно продлить срок службы оборудования и тем самым повысить экономическую эффективность производства.  
  


*Компания Rolls-Royce является не только производителем сверхдорогих автомобилей, но и одним из мировых лидеров в производстве авиадвигателей. С начала 2010-х годов она начала устанавливать на свои двигатели множество датчиков высоты, давления, трения, которые подробно фиксируют взаимодействие деталей двигателя и влияние внешней среды. Примерный объем данных, собираемых ими на одном двигателе, — около 1 терабайт в час. Это сравнимо с объемом жесткого диска на персональном компьютере новой модели.*

*По прибытии самолета после каждого рейса эти данные передаются в вычислительный центр компании, где проходят анализ системой предсказательной аналитики. Сбор и анализ настолько подробной статистики позволяют замечать даже мелкие отклонения от желаемого режима работы каждой детали и прогнозировать сбои в работе двигателя заранее.****В итоге двигатели Rolls-Royce ремонтируются или заменяются до того, как они сломались.***

*Для авиакомпаний это крайне удобно, так как обеспечивается максимальное использование имеющейся техники без простоев и накладок. После перехода на промышленную эксплуатацию такой системы Rolls-Royce стал использовать новую подписную бизнес-модель. Теперь компания не просто продает авиадвигатель, а заключает контракт на его корректную работу на срок Х лет, то есть берет на себя не только поставку товара, но и его обслуживание. Очевидно, что компания, обладающая огромными объемами данных о любом конкретном двигателе, сможет при необходимости починить его гораздо быстрее, чем любой другой сервис. И такой ремонт будет стоить дешевле.*

**

*В такой подписной бизнес-модели, основанной на использовании анализа данных, производитель продукта (например, авиадвигателя) сам несет все издержки за время поломок и поэтому максимально заинтересован в том, чтобы продукт работал максимально корректно.*

Пример Rolls-Royce не является уникальным или даже необычным. Так, по подписной модели работает компания ThyssenKrupp, мировой лидер на рынке лифтов и эскалаторов: клиент платит за услугу «Бесперебойная работа лифта в течение года со временем надежной работы не менее 98%». На новых моделях стандартных лифтов для жилого дома для этого устанавливается около 300 датчиков интернета вещей. Как и в случае с авиадвигателями, специфика бизнес-модели предполагает, что компания заинтересована внедрить систему максимально быстро (чтобы быстрее начать получать платежи за подписку) и обеспечивать режим работы без поломок (чтобы избежать штрафов).

## **Принятие управленческих решений с использованием результатов анализа данных**

Использование анализа данных, чтобы получить весомые обоснования для принятия управленческих решений, может дать огромные преимущества даже в сферах, имеющих мало общего с ИТ.

Не является секретом, что мобильная связь является источником очень подробных данных о нас, о наших потребностях, интересах и привычках. Мобильные операторы имеют либо могут узнать из поведения пользователей следующие данные:

* присутствие абонента в зоне действия базовой станции обслуживания;
* модель телефона;
* продолжительность разговоров, интернет-сессий;
* пол и возраст абонента;
* примерный доход и затраты;
* информация о поездках в другие регионы и за границу.

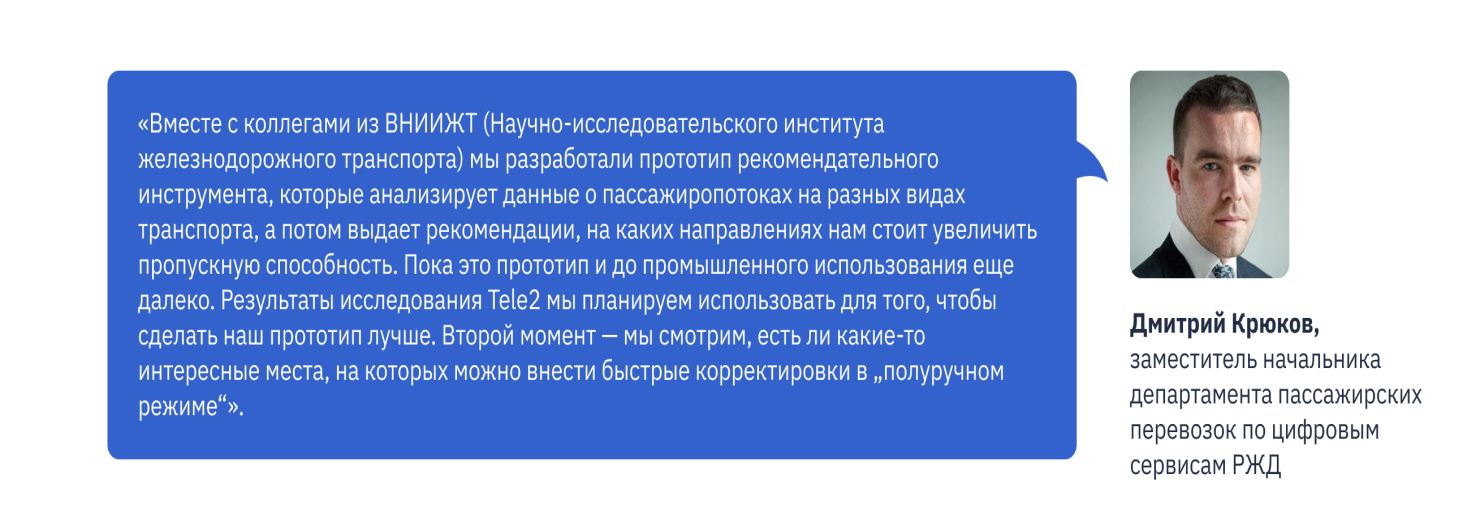
Объектом изучения обычно служат не персональные данные какого-либо конкретного абонента, а большие массивы обезличенных данных, усредненных по одному из признаков. Так, для проведения исследования сотрудников компании — мобильного оператора будет интересовать не конкретный москвич, Петр Иванов, 37 лет, который ездит на выходные в Подмосковье и скачивает на китайском смартфоне модели 2017 года около 5 Гб данных в месяц. Исследователя данных обычно интересует какая-либо закономерность поведения большого количества абонентов, схожих по какому-то признаку. Например, для разработки нового тарифа ему может быть актуально узнать, какая доля абонентов ездит на выходные за город или сколько абонентов определенной возрастной категории скачивает от 5 до 10 Гб трафика в месяц.  
  


*Мобильный оператор Tele2 совместно с «Российскими железными дорогами» реализовал проект по геоаналитике в транспортной отрасли. Специалисты по работе с данными исследовали пассажиропоток между 81 парой российских городов, чтобы узнать, сколько людей путешествует на поездах, самолетах, а также по трассам — на машинах и автобусах. Для исследования использовались следующие данные:*

* *обезличенные данные абонентов, предоставленные Tele2;*
* *собственные данные о пассажирах РЖД;*
* *информация о количестве авиа- и автобусных рейсов, скачанная в открытом доступе;*
* *данные о прочих видах перевозок, полученные от авиакомпании «Победа» и сервиса Blablacar.*

*Целью РЖД было получить алгоритм, который позволял бы на основе данных принимать решения о запуске новых поездов по тем или иным направлениям.*

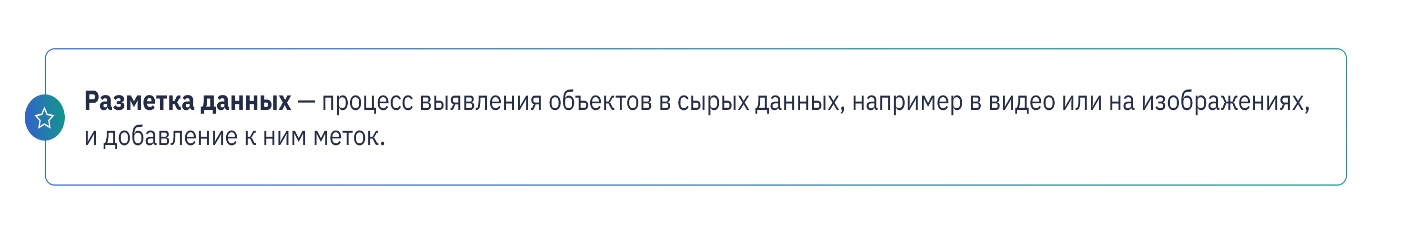
*Так, выяснилось, что общий пассажиропоток между Москвой и Ярославлем, согласно полученным данным, в 10 раз превышает тот, что помещается в текущие три-четыре быстрых поезда и несколько транзитных, более медленных. Эта информация уже является достаточной для того, чтобы реагировать на ситуацию и принимать меры.*

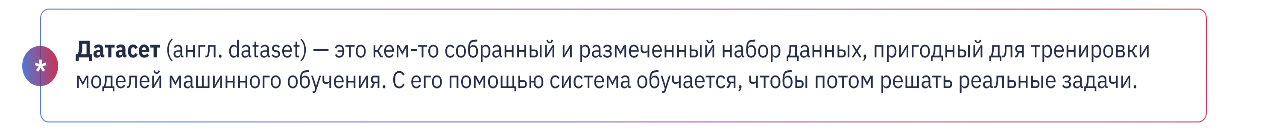


**Наборы данных (датасеты)**

Чтобы обучить программу (алгоритм машинного обучения) распознавать людей или машины на фотографии, ей нужно «скормить» по крайней мере несколько тысяч примеров, на которых будут отмечены люди и машины. После этого алгоритм сможет анализировать фото и обнаруживать объекты самостоятельно, и часто будет делать это даже лучше человека (то есть с меньшим количеством ошибок).

Первая проблема, с которой сталкиваются исследователи данных при решении практических задач, — получение набора размеченных данных необходимого размера и качества. Существенно облегчают задачу датасеты, или готовые наборы данных.







В коммерческих проектах по анализу данных формирование и разметка датасета обычно является наиболее длительным и дорогостоящим этапом проекта, так как требует огромного объема ручного труда. Так, для решения задачи обнаружения людей на фото потребуется, чтобы живой человек на десятках тысяч фото очертил квадраты, где присутствуют лица. Именно поэтому перед тем, как браться за формирование нового датасета, стоит проверить все возможные источники, где можно найти готовый набор данных.

Некоторые наборы данных по различным тематикам уже доступны в открытых ресурсах. Их можно использовать в своих научных, образовательных и бизнес-проектах.

Рассмотрим варианты, где можно искать датасеты.

Такие конкурсы, хакатоны, чемпионаты по анализу данных и машинному обучению обычно проводятся в онлайн-режиме, а результаты работ участников размещаются в открытом доступе.

**На международном уровне:**

* [Kaggle](https://www.kaggle.com/c/titanic) – платформа корпорации Google. Правила соревнований по машинному обучению и анализу данных устанавливают организаторы – такие компании, как Google, Intel, Mercedes-Benz и другие. Помимо этого, Kaggle хранит датасеты от этих постановщиков задач, которые обычно можно скачать и свободно использовать.
* [KDD CUP](https://www.kdd.org/kdd-cup) – кубок Ассоциации вычислительной техники (ACM) по машинному обучению и анализу данных. На этом портале команды со всего мира с помощью данных решают значимые для общества проблемы, например предсказывают загрязнение воздуха на основе данных экологического мониторинга.

**В России:**

* [Yandex Cup](https://yandex.ru/cup/) — открытый онлайн-чемпионат для разработчиков, среди его направлений есть и машинное обучение. Для каждой задачи соревнования команда организаторов готовит новый датасет, который раньше нигде не использовался. Кстати, участвуя в соревновании, можно испытать себя, получить денежные призы и даже трудоустроиться в «Яндекс».
* «[Яндекс.Толока](https://toloka.ai/ru/datasets)» — крупнейший источник данных, размеченных тысячами пользователей «Яндекса». Каждый день они делают сотни заданий и производят миллионы оценок.
* [Boosters.pro](https://mlbootcamp.ru/en/main/) – российская и восточноевропейская платформа для чемпионатов по анализу данных. Крупные компании предоставляют исходные данные для решения реальной бизнес-задачи, выявляют и награждают победителей.



Существует несколько наборов данных, которые по разным причинам стали популярными в какой-то момент истории, привлекли к себе внимание исследователей, поэтому регулярно дополнялись и улучшались. Сейчас они представляют собой огромные наборы данных отличного качества, которые можно использовать для тренировки алгоритмов.

[«Титаник»](https://www.kaggle.com/c/titanic) – сведения о пассажирах знаменитого корабля, попавшего в катастрофу. Этот датасет используется для обучения работе с предсказательными моделями.

[«Ирисы Фишера»](https://en.wikipedia.org/wiki/Iris_flower_data_set) – сведения о физических характеристиках отдельных цветков, собранных британским исследователем Рональдом Фишером. Датасет очень удобно использовать в алгоритмах классификации и кластеризации (разделения на группы).

[ImageNet](https://image-net.org/) – набор содержит сотни тысяч размеченных изображений и очень удобен при изучении и создании алгоритмов распознавания образов.

Существуют и специальные подборки датасетов:

* [Датасеты по машинному обучению из «Википедии»](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research).
* [Подборка лучших датасетов для машинного обучения и интеллектуального анализа данных](https://pub.towardsai.net/the-50-best-public-datasets-for-machine-learning-d80e9f030279).
* [25 открытых датасетов для глубокого обучения.](https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/)
* [Свободные датасеты для любых проектов по работе с данными.](https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/)



Регулярными поставщиками информации являются государственные органы и научные сообщества. Так, с начала 2000-х годов во всем мире развивается концепция электронного правительства (e-Government). В ее рамках многие страны, в числе которых и Россия, создали специальные сайты **открытых государственных данных (ОГД)**.

В России хорошим примером является платформа [Инфраструктура научно-исследовательских данных](https://data-in.ru/), благодаря которой исследователи могут получать доступ к данным, а представители органов государственной власти — ставить прикладные задачи и предоставлять для их решения наборы данных.

В свободном доступе или по платной подписке можно получить также наборы данных по событиям и ситуациям в мире, в том числе от международных организаций. Пример – сервис [Knoema](https://knoema.com/).

Также всегда можно воспользоваться [поисковым инструментом от Google](https://datasetsearch.research.google.com/), ориентированным специально на датасеты.