**Большие данные**

**Преимущества, которые предоставляет Big Data:**

1. Сбор данных из разных источников.
2. Улучшение бизнес-процессов через аналитику в реальном времени.
3. Хранение огромного объема данных.
4. Инсайты. Big Data более проницательна к скрытой информации при помощи структурированных и полуструктурированных данных.
5. Большие данные помогают уменьшать риск и принимать умные решения благодаря подходящей риск-аналитике

**Примеры Big Data**

**Нью-Йоркская Фондовая Биржа** ежедневно генерирует *1 терабайт* данных о торгах за прошедшую сессию.

**Социальные медиа**: статистика показывает, что в базы данных Facebook ежедневно загружается *500 терабайт* новых данных, генерируются в основном из-за загрузок фото и видео на серверы социальной сети, обмена сообщениями, комментариями под постами и так далее.

**Реактивный двигатель** генерирует *10 терабайт* данных каждые 30 минут во время полета. Так как ежедневно совершаются тысячи перелетов, то объем данных достигает петабайты.

**Классификация Big Data**

Формы больших данных:

* Структурированная
* Неструктурированная
* Полуструктурированная

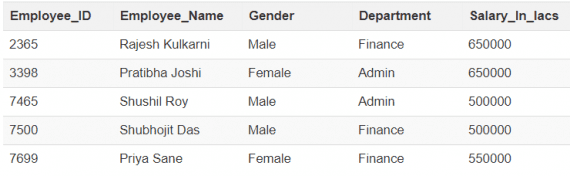
Структурированная форма

Данные, которые могут храниться, быть доступными и обработанными в форме с фиксированным форматом называются структурированными. За продолжительное время компьютерные науки достигли больших успехов в совершенствовании техник для работы с этим типом данных (где формат известен заранее) и научились извлекать пользу. Однако уже сегодня наблюдаются проблемы, связанные с ростом объемов до размеров, измеряемых в диапазоне нескольких зеттабайтов.

***1 зеттабайт соответствует миллиарду терабайт***

Глядя на эти числа, нетрудно убедиться в правдивости термина Big Data и трудностях сопряженных с обработкой и хранением таких данных.

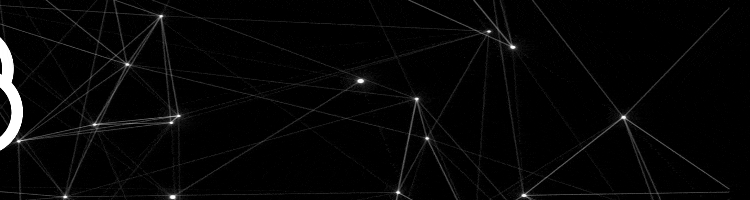
Данные, хранящиеся в реляционной базе — структурированы и имеют вид ,например, таблицы сотрудников компании



Неструктурированная форма

Данные неизвестной структуры классифицируются как неструктурированные. В дополнении к большим размерам, такая форма характеризуется рядом сложностей для обработки и извлечении полезной информации. Типичный пример неструктурированных данных — гетерогенный источник, содержащий комбинацию простых текстовых файлов, картинок и видео. Сегодня организации имеют доступ к большому объему сырых или неструктурированных данных, но не знают как извлечь из них пользу.

Примером такой категории Big Data является результат Гугл поиска:

[[](https://t.me/+duC5xkID7FhmNDYy)](https://t.me/+duC5xkID7FhmNDYy" \t "_blank)

Полуструктурированная форма

Эта категория содержит обе описанные выше, поэтому полуструктурированные данные обладают некоторой формой, но в действительности не определяются с помощью таблиц в реляционных базах. Пример этой категории — персональные данные, представленные в XML файле.

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>

<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>

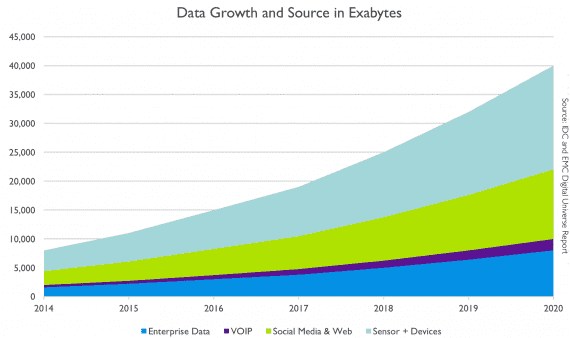
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

**Характеристики Big Data**

Рост Big Data со временем:



Синим цветом представлены структурированные данные (Enterprise data), которые сохраняются в реляционных базах. Другими цветами — неструктурированные данные из разных источников (IP-телефония, девайсы и сенсоры, социальные сети и веб-приложения).

В соответствии с Gartner, большие данные различаются по объему, скорости генерации, разнообразию и изменчивости. Рассмотрим эти характеристики подробнее.

1. **Объем**. Сам по себе термин Big Data связан с большим размером. Размер данных — важнейший показатель при определении возможной извлекаемой ценности. Ежедневно 6 миллионов людей используют цифровые медиа, что по предварительным оценкам генерирует 2.5 квинтиллиона байт данных. Поэтому объем — первая для рассмотрения характеристика.
2. **Разнообразие** — следующий аспект. Он ссылается на гетерогенные источники и природу данных, которые могут быть как структурированными, так и неструктурированными. Раньше электронные таблицы и базы данных были единственными источниками информации, рассматриваемыми в большинстве приложений. Сегодня же данные в форме электронных писем, фото, видео, PDF файлов, аудио тоже рассматриваются в аналитических приложениях. Такое разнообразие неструктурированных данных приводит к проблемам в хранении, добыче и анализе: 27% компаний не уверены, что работают с подходящими данными.
3. **Скорость генерации**. То, насколько быстро данные накапливаются и обрабатываются для удовлетворения требований, определяет потенциал. Скорость определяет быстроту притока  информации из источников — бизнес процессов, логов приложений, сайтов социальных сетей и медиа, сенсоров, мобильных устройств. Поток данных огромен и непрерывен во времени.
4. **Изменчивость** описывает непостоянство данных в некоторые моменты времени, которое усложняет обработку и управление. Так, например, большая часть данных неструктурирована по своей природе.

Big Data аналитика: в чем польза больших данных

**Продвижение товаров и услуг**: доступ к данным из поисковиков и сайтов, таких как Facebook и Twitter, позволяет предприятиям точнее разрабатывать маркетинговые стратегии.

**Улучшение сервиса для покупателей**: традиционные системы обратной связи с покупателями заменяются на новые, в которых Big Data и обработка естественного языка применяется для чтения и оценки отзыва покупателя.

**Расчет риска**, связанного с выпуском нового продукта или услуги.

**Операционная эффективность**: большие данные структурируют, чтобы быстрее извлекать нужную информацию и оперативно выдавать точный результат. Такое объединение технологий Big Data и хранилищ помогает организациям оптимизировать работу с редко используемой информацией.