

Попытки расчета количества информации на планете Земля

Как и физическая Вселенная, информационная Вселенная человечества разнообразна и постоянно расширяется – она обогащается ресурсами, когда кто-нибудь снимает красивые закаты или восходы на цифровую фотокамеру, когда данные собираются и анализируются от множества установленных на гигантской газовой Турбине датчиков, что делает процесс выработки электроэнергии дешевле и эффективнее; когда ученые получают и обрабатывают данные, передаваемые со скоростью 2,8 гигабайт в секунду с австралийского радиотелескопа ASKAP; когда человек делает групповое селфи на церемонии вручения Оскар и выкладывает его в сеть, где его просматривают в течение 12-часового периода 26 миллионов человек. Ниже показано данное фото с участием Брэдли Купера, Брэда Питта, Анжелины Джоли, Мерил Стрип, Джулии Робертс, Кевина Спейси, Чэннинга Татума и др.



Сегодня более чем 2 миллиарда людей и миллионы предприятий выполняют свою работу через интернет, миллиарды устройств связи отправляют и принимают данные, более одного миллиарда часов телевизионных шоу, сериалов и фильмов транслируется в интернете на основе потокового мультимедиа за один месяц только одной компанией Netflix. Вся информация, начиная от данных, собранных марсоходом Curiosity Rover на Марсе, до ваших фотографий с последнего отпуска, выложенных в Вконтакте или инстаграме с каждым днем пополняет и пополняет информационную Вселенную человечества.

В прошлой статье мы оценили масштабы таких единиц информации как Эксабайт, Зеттабайт, Йоттабайт и поэтому сейчас готовы к осознанию масштабов того количества информации, которое генерируется, хранится, передается и анализируется в современном мире. В данной статье собрано множество фактов с конкретными цифрами о нашей информационной Вселенной: как она развивалась последнее десятилетие, в каком состоянии она сейчас и что нас ждет в будущем. Начнем разбирательство с рассмотрения значимых компаний, которые вносят наиболее существенный вклад в информационную Вселенную.

Вопрос: Какая компания обрабатывает больше всего данных в мире?

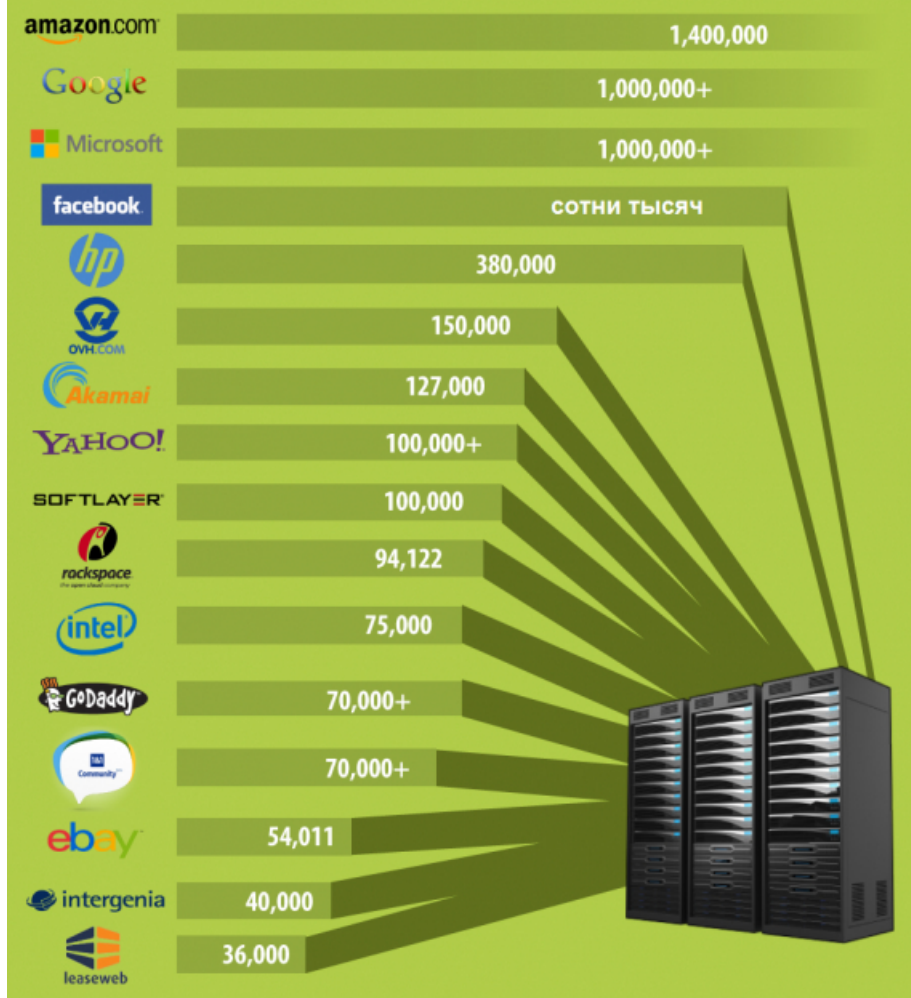
Ответ: Неудивительно, что ответ на этот вопрос – компания Google. Возможно, более удивительно, какие именно цифры сегодня стоят за этой корпорацией. Например, знаете ли вы, что Google обрабатывает 5.3 миллиарда запросов в день? Или что Google хранит более 10 эксабайт данных? Конечно Facebook, Microsoft и Amazon составляют Google серьезную конкуренцию: только Facebook имеет 2.5 миллиарда единиц контента, 2.7 миллиарда "лайков" и 300 миллионов фотографий – все это составляет более 500 терабайт данных.



Вопрос: какая компания имеет наибольшее количество серверов?

Ответ: После того, как дан ответ на предыдущий вопрос, логично подумать опять про Google. Но на самом деле ответ – Amazon. В компании размещают 1 000 000 000 гигабайт данных на более чем 1 400 000 серверах. Предполагается, что Google и Microsoft имеют около 1 000 000 серверов каждый, но, ни одна из компаний не оглашает точные цифры.

Количество серверов у корпораций



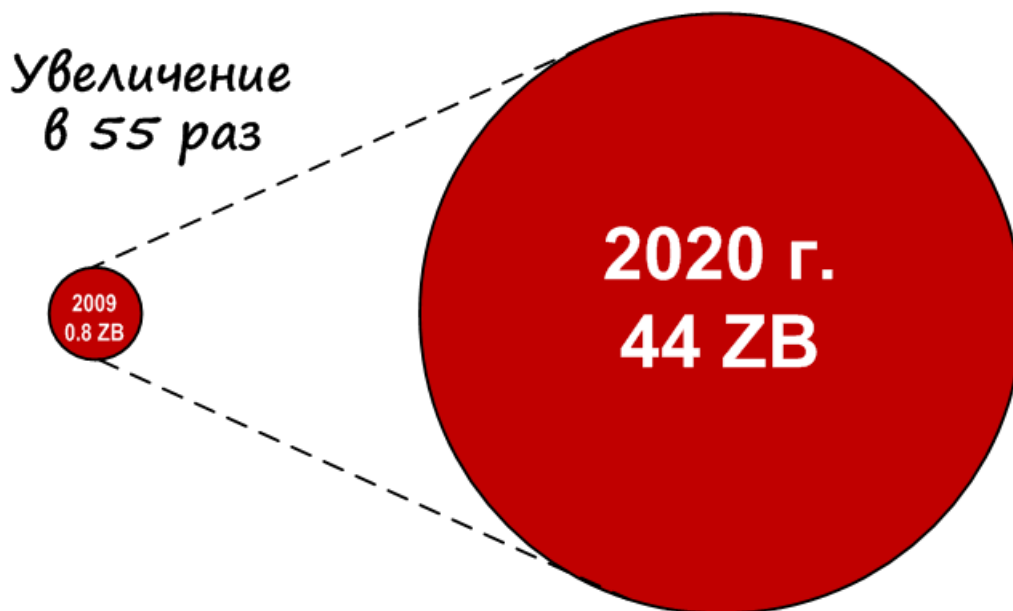
Вопрос: Какой процент данных хранится в цифровом формате?

Ответ: представьте каждую книгу в каждой библиотеке, каждой школе, каждом доме во всем мире – это много книг. Тем не менее, все эти книги вместе взятые составляют максимум 6 процентов от общей суммы всех человеческих данных. В 2007 году было подсчитано, что только 6% всех данных поступают из книг и из других печатных материалов – остальные 94% данных представляются уже в цифровом формате.

Вопрос: Что ждет нас в будущем?

Ответ: заглядывая в будущее, эксперты прогнозируют, что к 2020 году будет существовать 44 зеттабайт данных. Три года назад, по некоторым оценкам, вся всемирная паутина содержала около 500 эксабайт – это 5 миллиардов гигабайт, и только половина одного зеттабайта!

В настоящее время объемы цифровых данных удваиваются каждые два года. Международная исследовательская и консалтинговая компания IDC, занимающаяся изучением мирового рынка информационных технологий и телекоммуникаций, сначала прогнозировала, что с 2009 г до 2020 г. объем мировых данных увеличится в 44 раза, потом, что в 50 раз, теперь уже значится цифра 55 раз. Каждый год IDC с учетом анализа новых данных перестраивает кривую роста вверх, как правило, на несколько зеттабайт, по последнему отчету с 2009 г объем данных за год с 0,8 зеттабайт вырастет до 44 зеттабайт в 2020 г.



Картинка отлично показывает, насколько велик 44-кратный рост. Однако она в действительности не передает, насколько много 44 Зеттабайта на самом деле, и в каком объеме данных мы будем плавать (или тонуть) в 2020 году.

Число байт в 44 зеттабайтах является слишком большим даже для записи его в Microsoft Excel (Число настолько велико, что при вычислениях Excel урежет последние семь цифр точности). Предполагая, что закон Мура позволит нам удвоить емкость смартфонов в три раза до 2020 года, потребуется 188 978,561,024 (188+ триллион) смартфонов для хранения 44 ZB. Это число соответствует 25 смартфонам на человека для всего населения планеты. Представьте себе задачу управления данными, распределенными по множеству смартфонов.

Facebook в настоящее время хранит 300 Петабайт (PB) данных (это число растет на 600 ТБ в день, а по некоторым другим данным даже больше). Довольно внушительная цифра. Однако это намного, намного меньше, чем даже один Зеттабайт, который вмещает в себя 3 тысячи 495 хранилищ данных Facebook.

Сегодня Facebook имеет 1.3 миллиарда активных пользователей в месяц (стоит отметить, что при этом Facebook содержит около 81 миллиона фейковых профилей). Если бы к 2020 г. в Facebook зарегистрировались все жители на Земле – 7.5 миллиарда человек и данные продолжали бы увеличиваться ежегодно в 10 раз, то хранилище данных Facebook все равно содержало бы только 1/60 часть от 1 ZB:



Незаурядные попытки расчета количества информации человечества

В 2003 году Марк Либерман рассчитал требования к хранилищу всей когда-либо произносимой человеческой речи. Его расчеты показали примерно 42 зеттабайта, если бы оцифровка звука осуществлялась с параметрами 16 кГц/16-бит (что, на мой взгляд, для речи жирно). Он сделал это, чтобы проверить следующий факт: "все слова, произнесенные когда-либо людьми, несут около 5 эксабайт данных".

Расчеты были банально простыми:

Допустим в 1800 году на планете жил 1 миллиард человек, в 1900 году – 1,6 миллиарда человек, а в 2000 году – 6,1 миллиарда человек. Предположим, что 10 миллиардов человек прожили в среднем 50 лет, говоря в среднем по 2 часа в день на протяжении всей своей жизни. Это

$10 * 10^9 * 50 * 365 * 2 * 60 * 60 = 1.314 * 10^{18}$ секунд.

Если оцифровать одноканальный звук со скоростью 32 КБ в секунду, то получим

$1.314 * 10^{18} * 3.2 * 10^4 = 4.208 * 10^{22}$ байт или 42 зеттабайта

После полученных результатов Либерман предположил, что возможно, авторы приведенной выше фразы имели в виду "все слова, написанные когда-либо людьми".

Для тех, кому все эти биты и байты кажутся немного абстрактными в 2015 году студенты Лестерского университета попытались посчитать размер Интернета в физических терминах.

Сначала они подсчитали, что в индексированной поисковой сети насчитывается примерно 47 миллиардов страниц. Между тем, чтобы узнать, сколько печатных страниц будет у каждой веб-страницы из этих 47 миллиардов, нужно было посетить репрезентативную выборку веб-сайтов и попытаться распечатать каждый. Чтобы получить справедливое среднее значение с 5-процентной погрешностью и 95-процентным доверительным интервалом пришлось бы протестировать 385 случайных сайтов с помощью программы Random Website Machine. Цитата студента-исследователя Лестерского университета: "Это забавная игра, вы можете играть в нее у себя дома, если вам больше абсолютно нечем заняться: нажмите кнопку "случайный сайт"; нажмите CTRL+A и CTRL+P; и запишите полученное число страниц в таблицу".

Студенты подсчитали, что потребуется 2 процента тропических лесов Амазонки, чтобы сделать бумагу, на которой можно будет распечатать все текстовые данные сети интернет. Исследователи оценили, что средняя веб-страница содержит данных на 6,5 печатных страниц. При таком предположении для печати всех текстовых данных интернета потребуется 305,5 миллиарда листов (это 212 миллионов копий произведения "Война и мир" Льва Толстого или 75 миллионов копий всей серии книг о Гарри Поттере).

Конечно, оценивать огромный объем данных, размещенный в сети Интернет в текстовом виде нерационально. Согласно исследованиям Cisco, в 2015 году IP-трафик по передаче видео составил 8000 петабайт в месяц, по сравнению с около 3000 петабайт по загрузке web страниц и электронной почты. В целом, по итогам года видео составляло большую часть интернет трафика – 34 000 петабайт, второе место занял общий доступ к файлам – 14 000 петабайт.

Тезисная подборка других интересных фактов:

- Сегодня в цифровой Вселенной насчитывается 2.7 зеттабайт данных.
- Компания Akamai (поставщик услуг для акселерации веб-сайтов, провайдер платформ доставки контента и приложений) анализирует 75 миллионов событий в день, чтобы улучшить целевую рекламу.
- По всему миру сегодня более 5 миллиардов человек звонят, пишут друг другу в мессенджерах и просматривают веб-страницы на мобильных телефонах.
- На расшифровку генома человека первоначально было потрачено 10 лет. Сегодня эту же работу по обработке информации можно было выполнить за одну неделю.
- В 2008 году Google обрабатывал 20 000 терабайт данных в день.
- Самая большая база данных AT&T может похвастаться названиями, включая самый большой объем данных в одной уникальной базе данных (312 терабайт) и второе по величине количество строк в уникальной базе данных (1,9 триллиона), которая включает в себя обширные записи вызовов AT&T.
- Пользователи YouTube загружают 48 часов нового видео каждую минуту, каждый день загружается видео на 10 терабайт, что эквивалентно 405 Blu-Ray дискам.
- Каждую минуту создается 571 новых сайтов.
- В 2006 году совокупное пространство всех жестких дисков компьютера в мире оценивалось примерно в 160 эксабайт.
- В 1986 году технологический потенциал сетей телерадиовещания во всем мире составлял 0.432 зеттабайта оптимально сжатой информации, в 1993 году эта же цифра составила 0.715 зеттабайта, в 2000 году – 1,2 зеттабайта, в 2007 – 1,9 зеттабайта, что является информационным эквивалентом того, как если бы каждый человек на Земле получал в день 174 газеты.
- По состоянию на 2009 год, вся Всемирная паутина содержала около 0.5 зеттабайт данных.
- Компания Seagate Technology сообщила о продаже только в 2011 году жестких дисков общей емкостью 330 эксабайт.
- К 2025 году, согласно исследованию IDC по заказу Seagate, количество информации вырастет до 163 зеттабайт.
- Стандарт формата размещения [таблиц разделов](#) на физическом [жестком диске](#) допускает максимальный размер диска и раздела 7,02 зеттабайт или 5,946 зебибайт при использовании секторов размером 512 байт.
- Файловая система ZFS (Zettabyte File System), которая поддерживает большие объемы данных, обеспечивает максимальный объем хранилища около 256 квадриллионов зеттабайт.
- Человеческий мозг способен хранить 10 терабайт информации.

- За месяц по WhatsApp отправляют 1 641 600 000 000 сообщений.
- По электронной почте за месяц отправляют 8 078 400 000 000 писем



Генерировать информацию это дело несложное, а вот обработать ее и проанализировать уже гораздо сложнее. В 2013 году только 22% информации в цифровой Вселенной потенциально можно было проанализировать, т. е. эта часть содержала хоть какие-то метаданные (чаще всего мы мало что знаем о данных, если только они каким-то образом не охарактеризованы или не помечены); менее 5% из этих данных были фактически проанализированы. Согласно некоторым оценкам к 2020 году этот полезный процент может вырасти до 35% за счет роста объема данных, получаемых от встраиваемых информационных систем. Если задуматься над приведенными цифрами, то в скором времени будет не только проблема большого количества мусора на нашей планете, но также появится проблема большого количества перерабатываемого информационного мусора.

Напоследок, представляем вашему вниманию интересную картинку: "что происходит при помощи Интернета за 1 минуту?"

2017



2018



Составители:
 @LoriLewis
 @OfficiallyChadd