Маркетинговая аналитика

Раздел 5. Big Data

5.1. Большие данные. Что делает данные большими.

Что такое большие данные (англ. big data)? На просторах интернета или в литературе можно часто встретить множество различных определений данного понятия. Из них самые популярные это:

* Big Data – это когда данных больше, чем 100Гб (500Гб, 1Тб, и т. д.)
* Big Data – это такие данные, которые невозможно обрабатывать в Excel
* Big Data – данные, которые невозможно обработать на одном компьютере.

Отчасти, все они верны. Мы же будем придерживаться следующего определения: **большие данные** – серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия для получения воспринимаемых человеком результатов. Часто, когда говорят о Big Data, упоминают правило VVV – три признака или свойства, которыми большие данные должны обладать:

* Volume – объем (данные измеряются по величине физического объема документов)
* Velocity – данные регулярно обновляются, что требует их постоянной обработки
* Variety – разнообразные данные могут иметь неоднородные форматы, быть неструктурированными или частично структурированными. Примеры: структурированные данные – таблица в Excel, частично структурированные – логи программы, неструктурированные – рентген снимок.

На самом деле, количество V можно (и нужно) увеличивать, например, данные могут обладать еще такими свойствами, как: Veracity (степень доверия к данным), Value (ценность собранных данных для бизнеса), Variability (способы, которыми данные могут быть использованы). Откуда же берется такое количество данных? Ответ на этот вопрос становится очевидным, если посмотреть на то, сколько данных генерируется в интернете каждую минуту: Snapchat – 2.1 миллиона «снапов», Google – 3.8 миллионов запросов, Facebook – ежеминутно 1 миллион людей входит в сеть, YouTube – 4.5 миллионов видео просматривается, 188 миллионов e-mails отправляется. И это огромное количество данных. Для организации работы с такими данными и был придуман термин Big Data. Чем же этот подход отличается от традиционной аналитики? Сформулируем некоторые пункты.

|  |  |
| --- | --- |
| **Традиционная аналитика** | **Big data аналитика** |
| Постепенный анализ небольших пакетов данных | Обработка сразу всего массива доступных данных |
| Редакция и сортировка перед обработкой | Данные обрабатываются в их исходном виде |
| Старт с гипотезы и её тестирование относительно данных | Поиск корреляций по всем данным до получения искомой информации |
| Данные собираются, обрабатываются, хранятся и лишь затем анализируются | Анализ и обработка больших данных в реальном времени, по мере поступления |

Теперь поговорим о технологическом и инструментальном обеспечении больших данных. При работе с большими данными, должны соблюдаться следующие принципы:

* **Горизонтальная масштабируемость.** Подразумевает под собой тот факт, что любая система, работающая с большими данными, должна быть расширяемой. Если данных стало в два раза больше, то мы увеличили количество железа в системе в два раза, и всё продолжило работать.
* **Отказоустойчивость.** Первый принцип подразумевает, что для обработки больших данных может использоваться несколько машин. Тогда если одна машина выйдет из строя, мы должны иметь возможность продолжить работать со всем объемом данных.
* **Локальность данных.** По возможности – обрабатывать данные на той же машине, что и храним. Иначе, если данные находятся физически на одном сервере, а обрабатываются на другом – расходы на передачу данных непозволительно велики.

Одной из распределенных моделей обработки данных является **MapReduce**. Предложена модель была компанией Google. Главные концепция этой модели в том, что вся обработка данных происходит в три стадии: Map, Shuffle и Reduce. Не будем углубляться в то, что происходит внутри каждой стадии, однако важное свойство – все эти стадии могут работать *параллельно,* а первая и третья стадии – ещё и *независимо*. Благодаря этому, реализуется принцип горизонтальной масштабируемости. Именно поэтому, операции, которые раньше выполняла одна машина, теперь можно выполнять на нескольких машинах одновременно. В идеале, если при увеличении количества машин в два раза, время выполнения операции снижается в два раза. Одним из главных инструментов для работы с большими данными является Hadoop. На самом деле, с Hadoop связано ещё большое количество инструментов, но мы остановимся на распределенной файловой системе Hadoop (Hadoop Distributed File System, коротко - HDFS). Ключевая проблема традиционный файловых систем следующая – файл может быть настолько большим, что не поместиться на одном жестком диске. HDFS реализует следующие принципы в своей работе:

* Данные хранятся маленькими пакетами (batch, как правило 128МБ)
* Каждый batch имеет несколько копий
* Позволяет обрабатывать данные параллельно

Примеры больших данных и пользы от них:

1. В медицине – Big Data позволяет обрабатывать данные со смартфонов и фитнес-трекеров (или других носимых устройств) пациентов, что приводит к лучшему анализу информации и помогает выявлять проблемы со здоровьем, даже если у человека нет явных симптомов.
2. Для предотвращения катастроф. Такие стихийные бедствия, как ураганы, наводнения и т. д. очень сложно предсказывать без соответствующих инструментов, поскольку количество параметров и вообще всего того, от чего зависит появление таких катастроф слишком велико. А ведь прогноз бедствия нужно делать ещё и на несколько дней вперёд. В 2012 году, благодаря технологиям Big Data было возможно предсказание урагана на восточном побережье USA.
3. В автомобилестроении. Современный автомобиль – накопитель информации. И количество способов использовать эту информацию очень велико – от улучшения производства, путем обработки клиентских жалоб, до предотвращения аварии, путем анализа дорожных условий.

## ***Литература***

1. https://habr.com/ru/company/dca/blog/267361/
2. https://www.uplab.ru/blog/big-data-technologies/
3. Андреас Вайгенд: BIG DATA. Вся технология в одной книге. – Эксмо, 2018. – 384с.