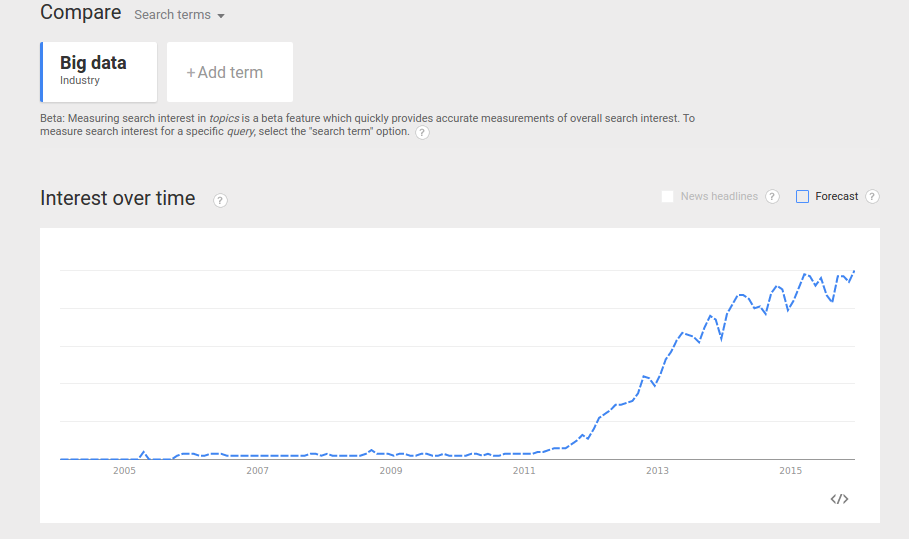
## Big data

### Определение термина

Термин Big Data появился сравнительно недавно. Google Trends показывает начало активного роста употребления словосочетания начиная с 2011 года.



Встречаются разные определения:

* Big Data – это когда данных больше, чем 100Гб (500Гб, 1ТБ, кому что нравится)
* Big Data – это такие данные, которые невозможно обрабатывать в Excel
* Big Data – это такие данные, которые невозможно обработать на одном компьютере

И даже такие:

* Вig Data – это вообще любые данные.
* Big Data не существует, ее придумали маркетологи.

Однако будем придерживаться формального определения:

Большие данные (англ. big data) — серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence. В данную серию включают средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, решениями категории NoSQL, алгоритмами MapReduce, программными каркасами и библиотеками проекта Hadoop.

В качестве определяющих характеристик для больших данных отмечают «три V»: объём (англ. volume, в смысле величины физического объёма), скорость (англ. velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (англ. variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных).

Несколько примеров того, что может быть источником данных, для которых необходимы методы работы с большими данными:

* Логи поведения пользователей в интернете
* GPS-сигналы от автомобилей для транспортной компании
* Данные, снимаемые с датчиков в большом адронном коллайдере
* Оцифрованные книги в Российской Государственной Библиотеке
* Информация о транзакциях всех клиентов банка
* Информация о всех покупках в крупной ритейл сети и т.д.

Количество источников данных стремительно растёт, а значит технологии их обработки становятся всё более востребованными.

### История

Введение термина «большие данные» относят к Клиффорду Линчу, редактору журнала Nature, подготовившему к 3 сентября 2008 года специальный номер журнала с темой «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?», в котором были собраны материалы о феномене взрывного роста объёмов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка «от количества к качеству»; термин был предложен по аналогии с расхожими в деловой англоязычной среде метафорами«большая нефть», «большая руда».

Несмотря на то, что термин вводился в академической среде, и прежде всего, разбиралась проблема роста и многообразия научных данных, начиная с 2009 года термин широко распространился в деловой прессе, а к 2010 году относят появление первых продуктов и решений, относящихся исключительно и непосредственно к проблеме обработки больших данных. К 2011 году большинство крупнейших поставщиков информационных технологий для организаций в своих деловых стратегиях используют понятие о больших данных, в том числе IBM, Oracle, Microsoft, Hewlett-Packard, EMC, а основные аналитики рынка информационных технологий посвящают концепции выделенные исследования.

В 2011 году Gartner отмечает большие данные как тренд номер два в информационно-технологической инфраструктуре (после виртуализации и как более существенный, чем энергосбережение и мониторинг). Прогнозируется, что внедрение технологий больших данных наибольшее влияние окажет на информационные технологии в производстве, здравоохранении, торговле, государственном управлении, а также в сферах и отраслях, где регистрируются индивидуальные перемещения ресурсов.

С 2013 года большие данные как академический предмет изучаются в появившихся вузовских программах по науке о данных и вычислительным наукам, и инженерии.

### Принципы работы с большими данными

Исходя из определения Big Data, можно сформулировать основные принципы работы с такими данными:

* Горизонтальная масштабируемость. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжило работать.
* Отказоустойчивость. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин (Facebook — 1400, Ebay - 532, Spotify — 120, Adobe - 30). Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий.
* Локальность данных. В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.

Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы разработки средств разработки данных.

### Методы анализа

Методы и техники анализа, применимые к большим данным:

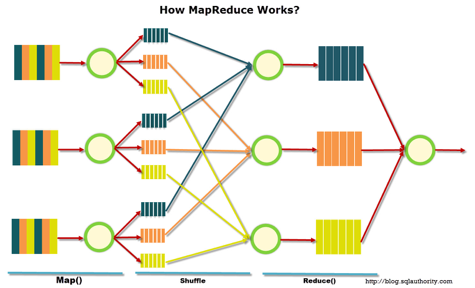
* методы класса Data Mining: обучение ассоциативным правилам (англ. association rule learning), классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ;
* краудсорсинг — категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения;
* смешение и интеграция данных (англ. data fusion and integration) — набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа, в качестве примеров таких техник, составляющих этот класс методов приводятся цифровая обработка сигналов и обработка естественного языка (включая тональный анализ);
* машинное обучение, включая обучение с учителем и без учителя, а также Ensemble learning — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. constituent models, ср. со статистическим ансамблем в статистической механике);
* искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;
* распознавание образов;
* прогнозная аналитика;
* имитационное моделирование;
* пространственный анализ (англ. Spatial analysis) — класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
* статистический анализ, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов;
* визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

### Технологии

Наиболее часто указывают в качестве базового принципа обработки больших данных в SN-архитектуру (англ. Shared Nothing Architecture), обеспечивающую массивно-параллельную обработку, масштабируемую без деградации на сотни и тысячи узлов обработки. При этом, кроме рассматриваемых большинством аналитиков технологий NoSQL, MapReduce, Hadoop, R, иногда включается в контекст применимости для обработки больших данных также технологии Business Intelligence и реляционные системы управления базами данных с поддержкой языка SQL.

Рассмотрим наиболее популярные технологии подробней.

#### MapReduce

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах.  


MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 стадии:

* Стадия Map. На этой стадии данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи. Функция map() примененная к одной входной записи и выдаёт множество пар ключ-значение. Множество – т.е. может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение. Что будет находится в ключе и в значении – решать пользователю, но ключ – очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce.
* Стадия Shuffle. Проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти корзины послужат входом для reduce.
* Стадия Reduce. Каждая «корзина» со значениями, сформированная на стадии shuffle, попадает на вход функции reduce(). Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины». Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

Несколько дополнительных фактов про MapReduce:

* Все запуски функции map работают независимо и могут работать параллельно, в том числе на разных машинах кластера.
* Все запуски функции reduce работают независимо и могут работать параллельно, в том числе на разных машинах кластера.
* Shuffle внутри себя представляет параллельную сортировку, поэтому также может работать на разных машинах кластера. Пункты 1-3 позволяют выполнить принцип горизонтальной масштабируемости.
* Функция map, как правило, применяется на той же машине, на которой хранятся данные – это позволяет снизить передачу данных по сети (принцип локальности данных).
* MapReduce – это всегда полное сканирование данных, никаких индексов нет. Это означает, что MapReduce плохо применим, когда ответ требуется очень быстро.

#### Hadoop

Парадигму MapReduce предложила компания Google в 2004 году в своей статье MapReduce: Simplified Data Processing on Large Clusters. Поскольку предложенная статья содержала описание парадигмы, но реализация отсутствовала – несколько программистов из Yahoo предложили свою реализацию в рамках работ над web-краулером nutch. С этого и началась история Hadoop.  
Изначально Hadoop был, в первую очередь, инструментом для хранения данных и запуска MapReduce-задач, сейчас же Hadoop представляет собой большой стек технологий, так или иначе связанных с обработкой больших данных (не только при помощи MapReduce).  
Основными (core) компонентами Hadoop являются:

* Hadoop Distributed File System (HDFS) – распределённая файловая система, позволяющая хранить информацию практически неограниченного объёма.
* Hadoop YARN – фреймворк для управления ресурсами кластера и менеджмента задач, в том числе включает фреймворк MapReduce.
* Hadoop common.

Также существует большое количество проектов, непосредственно связанных с Hadoop, но не входящих в Hadoop core:

* Hive – инструмент для SQL-like запросов над большими данными (превращает SQL-запросы в серию MapReduce–задач);
* Pig – язык программирования для анализа данных на высоком уровне. Одна строчка кода на этом языке может превратиться в последовательность MapReduce-задач;
* Hbase – колоночная база данных, реализующая парадигму BigTable;
* Cassandra – высокопроизводительная распределенная key-value база данных;
* ZooKeeper – сервис для распределённого хранения конфигурации и синхронизации изменений этой конфигурации;
* Mahout – библиотека и движок машинного обучения на больших данных.

Отдельно хотелось бы отметить проект Apache Spark, который представляет собой движок для распределённой обработки данных. Apache Spark обычно использует компоненты Hadoop, такие как HDFS и YARN для своей работы, при этом сам в последнее время стал популярнее, чем Hadoop.

### Big Data в действии

Зрелость технологий Больших Данных в организации определяется составом задач, для которых они используются, что, в свою очередь, сильно зависит от внутренней среды организации. Выделяются пять уровней зрелости для бизнес-модели организации с точки зрения интегрированных групп задач, которые решаются на основе технологий Больших Данных: мониторинг бизнеса, анализ бизнеса, оптимизация бизнеса, монетизация данных, трансформация бизнеса. Интерес представляют высокие уровни зрелости. Так, уровень монетизации данных предполагает, что организация не только обладает высоким цифровым потенциалом для того, чтобы собирать и обрабатывать данные, но и способна создавать на этой основе новый цифровой продукт или услугу, опираясь на аналитику Больших Данных. Развитая аналитика Больших Данных создает базис для разработки стратегий развития предприятия, направленных на существенную трансформацию бизнес-моделей с точки зрения новых продуктов, услуг и рынков.

Наиболее важные перемены связаны с изменением управления данными в организации.

Учитывая разнообразие структуры Больших Данных, их содержания, характеристик источников, предполагаемого использования, необходимо опираться на применение существенно большего, по сравнению с традиционным, спектра инструментов и методов, а также на многоплатформность программных и аппаратных решений и тщательный подход к формированию персонала, ответственного за управление.

Особую сложность представляет управление потоковыми данными, поступающими в реальном времени. Задачи обработки в этом случае направлены на выявление в потоке значимой для решения задач информации и событий, сопоставление с существующей информацией и ее сохранение. Инструменты для обработки сложных событий (Complex Event Processing, CEP) позволяют настраиваться на особенности определенных потоковых данных, проводить корреляцию данных из многих источников, обеспечивая более высокую отдачу от их использования.

Требования к перечню необходимых компетенций аналитика в сфере Больших Данных еще только формируются. В общем виде компетенции специалиста по исследованию данных можно представить в виде диаграммы Венна.



Предполагается, что аналитические навыки включают знание как традиционной аналитики, которая объясняет, что происходит и почему, так и аналитики предсказательной и предписывающей. Последняя направлена не только на поддержку решений, но и на автоматизацию принятия решений.

Бизнес-навыки предполагают умение правильно формулировать вопросы к аналитике для решения бизнес-задач и производить отбор критериев для оценки результатов, а также знание существующих ограничений, методов принятия решений, способов обеспечения прозрачности применяемых механизмов и методов. ИТ-навыки включают знание технологий сбора, хранения и обработки данных. Важными являются компетенции в области высокопроизводительных вычислений и доставки данных.

Компетенции и навыки в области исследования данных связаны с пониманием свойств данных, их моделей, соответствующих аналитических методов, способов представления и интерпретации данных. В области пересечения находятся интегрированные компетенции, позволяющие решать задачи бизнеса на основе аналитики Больших Данных. Это компетенции, обеспечивающие способность находить нужные данные, понимать их, осуществлять управление ими в организации, находить на основе их исследования новые возможности и решения для организации.