Лекция 1. Большие данные (Big Data). Анализ больших данных

Все вокруг говорят о больших данных: что с их помощью можно анализировать бизнес-процессы, предсказывать поведение клиентов, управлять производством и даже разрабатывать искусственный интеллект. Разберемся, что это, для чего они нужны и как работают.

Сегодня любая компания, независимо от ее размера и местоположения, так или иначе имеет дело с данными. Использование информации в качестве ценного ресурса, в свою очередь, подразумевает применение специальных инструментов для анализа ключевых показателей деятельности компании. Спрос на аналитику растет пропорционально ее значимости, и уже сейчас можно определить мировые тенденции и перспективы в этом секторе.

Все больший интерес представляют собой данные, поступающие в режиме реального времени из социальных медиа, видео и фото регистраторов, электронной почты и других распределенных источников, расположенных во внешнем окружении. Основным свойством таких источников является наличие нарастающего высокоскоростного потока данных с неопределенной структурой.

За последние несколько лет человечество произвело информации больше, чем за всю историю своего существования. И рост продолжается экспоненциально. Так, согласно прогнозам консалтинговой компании IDC, к 2025-му году объем данных в мире достигнет 173 Збайт (1 Збайт = триллион Гб), и 30% сгенерированных данных будут использоваться в режиме реального времени.

Это во многом было обусловлено экспоненциальным ростом количества вычислительных средств, приложений и пользователей, участвующих в формировании глобальных потоков данных. Современная эпоха —эпоха мобильной связи, мобильного Интернета, социальных сетей, блогов, Интернета вещей привела к появлению миллиардов пользователей и миллионов приложений.

Необходимость обработки качественно новых объемов структурированных и неструктурированных данных показала, что традиционные подходы к их хранению и анализу стали неэффективными, а, следовательно, необходимы новые технологии. Аналитики рассуждают следующим образом: «Мы не знаем, нужна ли нам информация, а если нужна, то какая, до тех пор, пока не проанализируем ее». Стоимость хранения информации настолько снизилась, что появилась возможность собирать все больше данных и анализировать их, руководствуясь принципом «мы не знаем, чего мы не знаем». Например, может быть обнаружено, что площадь того или иного цвета на обложке журнала влияет на вероятность его продаж в определенном периоде.

Итак, возникла проблема построения новой вычислительной инфраструктуры, которая была бы эффективной и не очень дорогой. Ключом к построению такой инфраструктуры и стал комплекс технологий, известный в настоящее время, как Большие данные – **Big Data**.

Попытка разобраться, что собой представляют Большие данные, вооружившись только привычными понятиями и терминами аналитики данных, вряд ли увенчается успехом. Некоторые авторы рассматривают Big Data, как Data Mining с кардинально увеличенными возможностями в плане объемов хранимых и обрабатываемых данных, а также скорости доступа к ним. Другие авторы рассматривают Data Mining как небольшую составляющую Big Data. А третьи вообще не упоминают Data Mining в

контексте Big Data. И все же попробуем разобраться, что, с точки зрения специалиста по аналитике, представляет собой термин Big Data, в чем его принципиальное отличие от Data Mining, какие новые перспективы и возможности открывают Большие данные?

Приведем примеры **источников, порождающих Большие данные**.

- **Торговые сети**

Крупные розничные торговые сети регистрируют ежечасно миллионы клиентских транзакций, которые пересылаются в хранилища данных, объем которых может составлять несколько петабайт.

- **Мобильные устройства**

Более 5 миллиардов людей по всему миру говорят, обмениваются сообщениями и производят поиск в Интернет с помощью мобильных устройств.

- **Автоматические регистраторы**

Тысячи автоматических регистраторов по всему миру непрерывно фиксируют погодные условия, и передают метеорологические данные в центры их обработки.

- **Социальные сети**

Пользователи социальных сетей ежеминутно отправляют десятки миллионов сообщений.

- **Данные от измерительных устройств**
- **Журналы доступа пользователей веб-сайтов**
- **Сенсорные сети**
- **Тексты и документы из Интернета**
- **Научные данные (астрономия, геном человека, исследования атмосферы, биохимия, биология)**
- **Данные министерства обороны**
- **Медицинские наблюдения**
- **Фото- и видео-архивы**
- **Данные электронной коммерции**

Архитектуры защиты данных в общем случае должны обеспечивать защиту от потери, неотслеживаемого повреждения, вредоносных программ и злонамеренного изменения данных киберпреступниками или в результате кибервойн. Данные являются активами и все чаще используются правительствами и деловыми кругами для принятия ключевых решений, но если достоверность данных неизвестна, ценность их снижается или даже может быть утрачена, и что еще хуже - могут быть приняты плохие решения.

Термин **Big Data** был предложен Клиффордом Линчем, редактором журнала Nature, который в 2008 году выпустил отдельный номер, главной темой которого была «Как могут повлиять на будущее науки, технологии, открывающие возможности работы с большими объёмами данных?» Термин «Большие данные» был предложен по аналогии с терминами «Большая нефть», «Большая руда» и т. д.

Термин большие данные может быть причислен к данным, связанным с высочайшей изменчивостью источников данных, а также обладающим сложными взаимосвязями и трудностями изменения или удаления отдельных записей. Большие данные характеризуются гигантским объёмом, значительной скоростью поступления данных, а также многообразием самих данных. Для таких данных требуются новейшие

способы обработки, которая в дальнейшем может привести к улучшению методов принятия решений, оптимизации процессов и поиска закономерностей.

Итак, термин **Big Data** в научный и корпоративный обиход вошел в 2008 году, и он гораздо шире, чем просто очень много данных: в разном контексте под ним могут подразумеваться и данные большого объема, и технологии их обработки, а также проекты, рынок и даже компании, активно использующие эти технологии.

Выходит, что у компании есть какие-то источники данных, сами данные, оборудование и программное обеспечение для хранения и обработки этой информации. Все это вместе можно включить в определение **Big Data**.

К 2011 году понятие Big Data стало набирать популярность, в основном, в крупных корпорациях таких как Microsoft, IBM, Oracle и др.

В 2011 году исследовательская компания Garther отмечает большие данные как тренд номер два в информационно-технологической инфраструктуре после виртуализации. По прогнозам подразумевается, что внедрение технологий Big Data сильно повлияет на информационные технологии в сферах производства, здравоохранении, торговли, государственного управлении, а также в отраслях, в которых регистрируются индивидуальные перемещения ресурсов.

С 2013 года большие данные начинают преподавать в университетах в рамках вузовских программ по науке о данных, вычислительным наукам и инженерии.

Большинство программных продуктов в области Big Data являются свободными.

Итак, **большие данные** - комплексный набор методов, подходов и инструментов обработки данных колоссальных объемов.

Главной целью обработки Big Data является быстрое и эффективное использование всех видов информации в условиях непрерывного изменения и прироста в больших объёмах.

Говоря простым языком, **Big Data** представляет собой безмерный объем информации, который не может быть обработан стандартными инструментами и аппаратными средствами.

Большие данные по сравнению с обычными данными требуют иной подход к обработке. При обработке Big Data используются собственные инструменты и технологии, которые предназначены для данных со сверхбольшим объёмом информации.

На ранней стадии формирования этого понятия большие данные определялись по признаку соответствия трем «**V-характеристикам**»: **Volume** – объем, **Velocity** – скорость, **Variety** – разнообразие (Laney, 2001). Вместе с широким распространением этой концепции в организациях, стремящихся сполна реализовать потенциал колоссальных массивов слабо структурированной информации, число V-характеристик в мнемоническом правиле определения понятия больших данных удвоилось. В наши дни к ним относят данные со следующими характерными свойствами.

Volume – объем как мера количества сгенерированных и хранящихся данных. Размер данных определяет значимость и потенциал данных, а также то, могут ли они быть рассмотрены как Большие данные. Большие данные включают миллиарды полей или записей, описывающих тысячи сущностей или элементов.

Velocity – скорость регистрации/генерирования, обработки или распространения:

большие данные зачастую не только создаются, но и распространяются и даже анализируются в режиме реального времени или близком к нему.

Variety/Variability – разнообразие/вариативность формы или представления: большие данные сохраняются во всевозможных форматах, а их структура зачастую бывает несогласованной не только между наборами, но и внутри отдельно взятых наборов данных. Большие данные могут состоять из текста, изображений, аудио, видео. При сопоставлении друг с другом большие данные могут дополнять отсутствующие данные.

Viscosity – вязкость: большие данные крайне трудно поддаются как вычленению из общей массы, так и анализу и интеграции с целью практического использования.

Volatility – волатильность, как мера непостоянства: большие данные крайне переменчивы, что весьма ограничивает сроки годности полученных с их использованием результатов.

Veracity – правдоподобие по критериям проверки подлинности источника. Качество данных напрямую влияет на точность проведения анализа данных.

Главной отличительной особенностью больших данных являются колоссальные объемы занимаемой ими памяти: сегодня под большими данными по умолчанию понимают нечто свыше 100 терабайт, а то и петабайты или эксабайты данных.

Чтобы отделить большие данные от обычных, нужно ответить на вопрос: «**Big Data** – это сколько?». Таблица в Экселе на 500 000 строк – это большие данные? А если строк миллиард? Текстовый файл на тысячи слов, который весит 2 мегабайта, – это много? А распечатки графиков температуры всех метеостанций Архангельской области – много или еще недостаточно?

Тут многие скажут, что эти примеры представляют собой довольно внушительное количество информации. Действительно, с такой точки зрения, все перечисленное – большие данные. Но что вы скажете про таблицу в Экселе на миллиард строк? Это тоже большие данные – и куда побольше тех!

На интуитивном уровне специалисты, далекие от **Big Data**, привыкли называть большими данными любой объем информации, который сложно удержать в голове и/или который занимает много места. И такое интуитивное определение неправильное.

Однозначно отделить формат больших данных от обычных помогут **три критерия**.

1. Данные должны быть **цифровыми**. Книги в национальной библиотеке или стопки документов в архиве компании – это данные, и часто их много. Но термин **Big Data** означает только цифровые данные, которые хранятся на серверах.
2. Данные должны **поступать в объективно больших объемах и быстро накапливаться**. Например, база заказов интернет-магазина по продаже колясок может быть большой: 10 миллионов заказов за 20 лет, но пополняется она со скоростью 100 заказов в сутки – это не большие данные. Фильм в высоком качестве может занимать десятки гигов, но со временем его размер не растет – это тоже не **Big Data**. А вот записи показателей пары сенсоров в двигателе Боинга, поступающие в количестве несколько гигабайт в час и загружаемые на диагностический сервер производителя авиатехники – это уже **Big Data**.
3. Данные должны быть **разнородными и слабоструктурированными**. Заказы в онлайн-магазине упорядочены, из них легко извлечь дополнительные статистические параметры, например, средний чек или самые популярные товары. Поэтому эти данные

не относят к **Big Data**. Показания датчиков температуры с корпуса самолета, записанные за последние 6 месяцев, – информация, в которой есть польза, но не очень понятно, как ее извлечь. Можно, конечно, рассчитать средние значения температуры за бортом самолета за полгода, но какой в этом смысл? А если погрузиться в анализ этих данных глубоко – можно вытащить много неочевидной информации. Например, о длительности перелетов, скорости набора высоты, климатических условиях за бортом и так далее. Информация интересная и полезная, но трудноизвлекаемая, значит, это большие данные.

Этот критерий не всегда обязательный. Иногда большие объемы структурированных данных, которые постоянно пополняются, относят к формату **Big Data**, особенно если их используют для машинного обучения или выявления неочевидных закономерностей. То есть если к структурированным данным применяют методы анализа **Big Data**, можно сказать, что это они и есть.

Итак, большие данные — это трудноанализируемая цифровая информация, накапливаемая со временем и поступающая к вам солидными порциями.

Зачем нужны Big Data?

Представьте молодого врача. К нему пришел пациент, который жалуется на одышку, боли в груди и периодическую изжогу. Врач убедился, что давление и показания сердечного ритма пациента в норме и ничего подозрительного у него прежде не замечалось. Отметил полноту пациента. Поскольку такие симптомы типичны для людей с избыточным весом, врач заверили пациента, что все в порядке, и посоветовал найти время для упражнений.

Очень часто к неверному диагнозу при сердечно-сосудистых заболеваниях приводит такое суждение. У пациентов в этом состоянии проявляются симптомы, которые схожи с симптомами ожирения, и врачи прекращают диагностику, которая могла бы обнаружить более серьезное заболевание. Мы – люди, и наши суждения обусловлены ограниченным субъективным опытом и несовершенными знаниями. Это ухудшает процесс принятия решения и, как в случае с неопытным врачом, удерживает от дальнейших проверок, которые могли бы привести к более точным выводам.

Не ограничиваясь суждением одного индивида, современные методы позволяют задействовать для принятия лучшего решения информацию из разных источников. Например, можно было бы свериться со статистикой по пациентам с такими симптомами и обнаружить диагнозы, о которых молодой врач даже не подумал.

С современным вычислением и передовыми алгоритмами мы можем:

- обнаружить скрытые тенденции в больших наборах данных;
- воспользоваться этими тенденциями для прогнозирования;
- вычислить вероятность любого возможного исхода;
- получить точные результаты быстро.

Когда в любом IT-проекте начинают работать с данными, в первую очередь анализируют наиболее очевидные, значимые и понятные показатели. Так, если речь идет об онлайн-торговле, сначала смотрят на средние чеки заказов, топ продаж и объемы складских запасов. Когда речь идет о самолетах – смотрят скорость, высоту, расход топлива.

Сбор и анализ очевидных метрик позволяет вносить в систему простые и понятные корректировки. Такие улучшения практически сразу дают ощутимый результат. Это называется «сбор фруктов с нижних веток дерева».

По мере эволюции системы инженеры прорабатывают все видимые узкие места в проекте. После этого начинается стагнация продукта: для поиска новых путей развития нужно лезть выше, чтобы собрать плоды с более высоких веток. Инженеры и аналитики начинают собирать и анализировать косвенные данные, напрямую не связанные с основными метриками проектов.

Например, в онлайн-торговле можно собирать со страниц магазина данные о перемещении курсора (или пальца) по экрану. Или собирать данные с большого числа сенсоров самолета, например: число оборотов двигателя, состав топливно-воздушной смеси, забортную температуру и температуру выхлопа. Или анализировать слова в комментариях клиентов в соцсетях для оценки их лояльности.

Это означает, что технологии **Big Data** чаще всего нужны тогда, когда требуется более глубокий анализ процессов.

Такие данные напрямую не связаны с основными метриками IT-системы и бизнеса, но при правильном анализе могут рассказать много интересного о возможных точках оптимизации в проекте. Работа с такими данными – как поиск нефти. Нужно пробовать разные места, применять различные стратегии поиска и извлечения скрытых ресурсов, спрятанных в данных. Далеко не все попытки будут успешны, но в итоге находки могут принести массу выгоды.

Большие данные в основном помогают решать четыре задачи:

1. **Анализировать текущее положение дел и оптимизировать бизнес-процессы**. С помощью больших данных можно понять, какие товары предпочитают покупатели, оптимально ли работают станки на производстве, нет ли проблем с поставками товаров. Обычно для этого ищут закономерности в данных, строят графики и диаграммы, формируют отчеты.

   Например, с помощью больших данных компания Intel обнаружила, что делает много лишних тестов при производстве процессоров. Они проанализировали данные, отказались от лишних тестов и сэкономили около 30 миллиардов долларов.
2. **Делать прогнозы**. Данные о прошлом помогают сделать выводы о будущем. Например, примерно прикинуть продажи в новом году или предсказать поломку оборудования до того, как оно действительно сломается. Чем больше данных, тем точнее предсказания.

   Например, логистическая компания ПЭК запустила Центр управления перевозками с использованием **Big Data**. В итоге они стали прогнозировать загрузку складов – предсказывать, когда склады будут заполнены, а когда пусты. Это помогло планировать маршруты транспорта и избегать простоев.
3. **Строить модели**. На основе больших данных можно собрать компьютерную модель магазина, оборудования или нефтяной скважины. Потом с этой моделью можно экспериментировать: что-то в ней изменять, отслеживать разные показатели, ускорять или замедлять разные процессы для их анализа.

   Например, «Газпромнефть» смоделировала ситуацию аварийного отключения электричества, чтобы понять, почему возникает сбой автоматического перезапуска оборудования. Модель помогла обнаружить неожиданные причинно-следственные связи и устранить проблемы.
4. **Автоматизировать рутину**. На больших данных учатся автоматические программы,

которые умеют выполнять определенные задачи, например, сортировать документы или общаться в чатах. Это могут быть как примитивные алгоритмы, так и искусственный интеллект: голосовые помощники или нейросети.

Так, компания Stafory разработала робота-рекрутера Веру. Этот робот выполняет простую рекрутерскую работу: распознает голос, сортирует резюме, задает простые вопросы и принимает ответы. В итоге рекрутерам-людям остаются только более сложные и творческие задачи – реальные собеседования и окончательный отбор кандидатов.

Технологии работы с большими данными

Рассмотрим в общих чертах как работают системы анализа больших данных и какие инструменты нужны для их работы.

Упрощенно работа с **Big Data** происходит по следующей схеме: информацию собирают из разных источников → данные помещают на хранение в базы и хранилища → данные обрабатывают и анализируют → обработанные данные выводят с помощью средств визуализации или используют для машинного обучения.

Для технологий, которые работают с большими данными, базовым принципом считают горизонтальную масштабируемость, то есть возможность обрабатывать данные сразу на множестве узлов (серверов, компьютеров). Если обрабатывать такой массив информации на одном узле, это займет слишком много времени.

Итак, к **основным технологиям для работы с большими данными** относят:

MapReduce. Это модель распределенных вычислений, разработанная Google. Ее суть в том, что обработка больших объемов информации происходит на большом количестве серверов (узлов), которые образуют кластер. На каждом сервере производятся одинаковые элементарные задания по обработке, потом все результаты обработки сводят вместе. Если копнуть чуть глубже, мы увидим, что в основе технологии лежат две процедуры функционального программирования. Первая – **map**, она применяет нужную функцию к каждому элементу данных. Вторая — **reduce**, она объединяет результаты работы. Такой подход позволяет быстрее обрабатывать большие данные.

NoSQL – термин расшифровывается как Not Only SQL, «не только **SQL**». Это подход к реализации систем управления базами данных. В общих чертах – особенность в том, что для хранения информации в базах данных **NoSQL** не требуется заранее заданная схема данных. Это значит, что любые данные можно легко помещать в хранилище и быстро извлекать оттуда. Когда у вас большое количество разнородных данных, именно это и нужно.

Hadoop – инструмент для разработки решений, которые работают по модели **MapReduce**. По сути, это конструктор, из которого можно создавать хранилища данных под потребности бизнеса. Технология лежит в основе многих облачных решений для обработки больших данных. Например, сервис для анализа **Big Data** от Mail.ru **Cloud Solutions** построен на базе Hadoop, Spark и ClickHouse.

R. Язык программирования для работы с графикой и статистической обработки данных. Стандарт для создания аналитических и статистических программ, без которых по определению невозможен анализ **Big Data**.

Сегодня **Python** стал первым предпочтительным языком, особенно когда речь идет о больших данных. Его простой синтаксис упрощает выполнение различных задач, и

именно поэтому он приобрел популярность за последние несколько лет. Будучи языком программирования с открытым исходным кодом, Python также был построен с обширными наборами библиотек, которые идеально подходят для специалистов по обработке данных, и это позволяет им выполнять практически любую задачу без каких-либо проблем.

Еще аналитики часто используют языки Scala, Java.

Рынок Big Data

С географической точки зрения по результатам 2019 г. наиболее крупным стал рынок США с объёмом доходов 100 млрд $. Второе и третье место по объёму заняли Япония (9,6 млрд) и Великобритания (9,2 млрд). Также в пятёрку крупнейших рынков вошли КНР (8,6 млрд) и Германия (7,9 млрд).

Вплоть до 2025 года лидерство на рынке будет удерживать Северная Америка, в частности США.

В основном такой рост вызван повышением интереса к интернету вещей IoT (всевозможные датчики, измерительные приборы для расхода воды, оборудование на роботизированных заводах, умный транспорт – все это порождает огромное количество информации, которая передается от машины к машине, а потом подвергается исследованию и обработке людьми). Сейчас к интернету вещей подключено 30,73 млрд устройств, а к 2025 году их будет 75,44 млрд. Кроме того, уже сейчас без больших данных компании не выдерживают конкуренцию с теми, кто использует Big Data, так как не могут обеспечивать достаточный уровень клиентского сервиса.

При этом нельзя сказать, что большие данные просто тренд, которому компании следуют бездумно. Опросы показывают, что Big Data помогают бизнесу на 8% увеличить прибыль и на 10% снизить расходы.

Давайте внесем конкретику, где же используются большие данные. Согласно обзору компании Деловой профиль, можно представить следующую диаграмму использования больших данных в различных индустриях России. Направление **Big Data** более активно развивается в компаниях, которые накопили большие пласты структурированной и неструктурированной информации: финансовая сфера, телекоммуникации, интернет-коммерция, ритейл. Особенно большой выигрыш от больших данных в России получат отрасли добычи полезных ископаемых, торговли, ремонта и строительства.

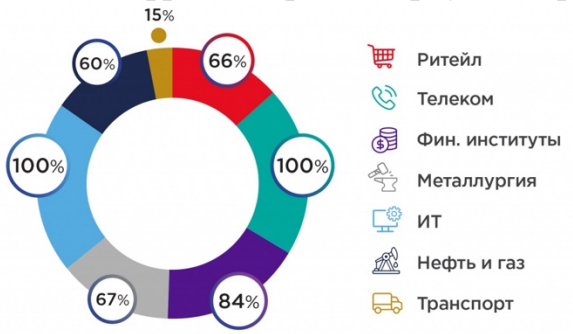
Телеком-операторы России работают с большим объёмом данных о своих пользователях. Они применяют технологии **Big Data** к целому ряду направлений: прогнозирование оттока абонентов, прогнозирование жалоб, планирование мероприятий по удержанию клиентов, предотвращение мошеннических финансовых операций и т.д.

Наиболее низкий показатель у транспортной отрасли, который равен 15% означает, что только 15% компаний используют технологии больших данных в своей работе. Конечно, это говорит о перспективах роста. Что касается ИТ и Телекома, то большие данные являются их хлебом.

О каких суммах идет речь?

Объем глобального мирового рынка big data в 2022 году 275 млрд долларов. Российский рынок пока занимает незначительную долю в мировом предложении и потреблении информационных технологий. На текущий момент российский рынок

примерно равен 10-30 млрд, ожидают что в 2024 году рост составит в 10 раз и рынок будет равен 300. Конечно такие цифры говорят в первую очередь о перспективах.



Однако ещё в 2018 году было принято немало решений и реализовано достаточное количество законодательных инициатив, способствующих развитию отечественного рынка Big Data.

Сегодня лидерами по внедрению технологий в российских компаниях являются такие инструменты цифровизации, как роботизированная автоматизация бизнес-процессов, использование чат-ботов, инструментов анализа больших данных и предиктивной аналитики.

Основные тенденции рынка в России развивают следующие компании, создав осенью 2018 года Ассоциацию больших данных (АБД).

Основная цель Ассоциации — создание условий для развития технологий и продуктов в сфере больших данных в России.

Ассоциация занимается созданием единых принципов и стандартов обработки, хранения, передачи и использования больших данных.

«Яндекс», VK, «Сбербанк», «Газпромбанк», «Тинькофф Банк», «МегаФон», «Ростелеком», oneFactor, QIWI, билайн, «МТС», Фонд «Сколково», Аналитический центр при Правительстве РФ, «Банк ВТБ», Центр стратегических разработок (ЦСР), «Магнит».

Аналитика больших данных

Аналитика больших данных (**Big Data Analytics, BDA**) становится одной из самых востребованных задач в современном бизнесе. Сейчас аналитика больших данных используется в более чем 70% компаний по всему миру. При том, что в 2015 году этот показатель составлял всего лишь 17%. **Big Data** активнее всего используется компаниями, которые работают в сфере телекоммуникаций и финансовых услуг. Затем

идут компании, которые специализируются на технологиях в здравоохранении. Минимальное использование аналитики **Big Data** в образовательных компаниях: в большинстве случаев представители этой сферы заявляли о намерении использовать технологии в ближайшем будущем.

К решениям **аналитики больших данных** эксперты относят инструменты и приложения, которые позволяют собирать структурированные и неструктурированные данные, управлять ими, организовывать, анализировать, обеспечивать доступ и передавать. Аналитика больших данных включает анализ крупных, сложных и часто неструктурированных наборов данных, позволяющий выявлять ценную информацию, с точностью определять тенденции, прогнозировать производственные показатели и оптимизировать расходы.

Большие данные могут быть классифицированы в соответствии с несколькими главными компонентами.

Что нужно для работы с Big Data

Чтобы работать с большими данными, придется учесть несколько моментов:

- **Готовьте много места**. Если вернуться к определению **Big Data**, то видно, что данных будет немало, значит, нужно быть готовыми где-то их хранить. Также информация может поступать с высокой скоростью, поэтому заранее смотрите, чтобы ширины входного канала и скорости дисков хватало для обработки входящего потока байтов.
- **Готовьте больше серверов**. Данные нужно не только хранить, но и как-то обрабатывать. Из-за больших объемов вам, скорее всего, придется разбивать информацию на порции и обрабатывать их параллельно на разных машинах, то есть использовать упомянутые выше технологии **MapReduce**. Для этого придется заранее озаботиться достаточным количеством железа для вычислений.
- **Готовьте правильные инструменты**. IT-специалисты много лет занимаются поиском крупиц золота в горах разнообразных больших данных. Для их расчетов создано много надежных, классных и быстрых инструментов, например: Hadoop, Spark и другие. Познакомьтесь с основными продуктами на рынке и выберите, что подойдет вам.

Подготовка инфраструктуры занимает много времени, поэтому лучше переложить ее на плечи профессиональных администраторов и присмотреться к облачным решениям по обработке **Big Data**.

Четыре основных типа данных

В настоящее время все существующие данные делятся на:

1. Структурированные
2. Полуструктированные
3. Квазиструктурированные
4. Неструктурированные

К структурированным данным относятся данные, определяющие конкретную предметную область. Такие данные упорядочены специальным образом и организованы таким образом, чтобы над такими данными можно было выполнить анализ. Обычно такие данные хранятся в виде таблиц в реляционных базах данных. Почти все алгоритмы машинного обучения и Data Mining работают со структурированными данными.

К полуструктурированным данным относятся данные, которые не соответствуют чёткой структуре таблиц и отношений в реляционных базах данных, однако такие данные содержат специальные теги и иные маркёры, позволяющие отделить семантические элементы. Такие данные принадлежат одному классу, но при этом могут иметь разные атрибуты. Xml документы являются простейшем примером полуструктурированных данных.

К неструктурированными данным относятся данные, которые не имеют определённой формы, могут включать в себя видео, аудио файлы, свободный текст, информацию, поступающую из социальных сетей. На сегодняшний день 80% информации входит в группу неструктурированной. Такую информацию необходимо комплексно анализировать, для упрощения ее дальнейшей обработки.

Квазиструктурированные данные - в которых можно выделить некую структуру, однако структура эта заранее целиком или частично неизвестна, либо может меняться с течением времени. При поступлении однотипных документов от различных источников (предприятий) информационное содержимое документов идентично, но оформление и даже формат данных могут кардинально отличаться. Например: гидродинамические исследования нефтяных и газовых скважин производятся несколькими различными организациями. Отчет о выполненных исследованиях по предприятиям формируется как в MS Word, так и в MS Excel. Оформление и порядок следования информации в документах различен, хотя информационное наполнение документов идентично. Анализировать такие документы — затруднительная и порой невыполнимая задача.

- Данные, структура которых даже априорно неизвестна. В этом случае формирование структуры каталога данных и его заполнение происходит одновременно по мере поступления информации. Примером может служить набор, произвольным образом связанных документов в различных форматах с произвольными описаниями каждого документа.
- Структура данных известна частично. Известны основные описания базовых сущностей или классов системы; вместе с тем имеется необходимость хранения и обработки заранее неопределенной дополнительной информации. Так, информация о покупателях или продавцах, может сопровождаться набором примечаний о каждой сделке, контактными телефонами, дополнительным условиями и т.д.
- Структура данных известна и четко определена, но может меняться с течением времени. Простейшим примером может служить набор реквизитов счета-фактуры или реквизитов организации, которые меняются с изменением законодательства.

Задачи, решаемые Big Data

- Управление документами и доступом
- Выявление мошенничества
- Системы гарантирования доходов
- Анализ текучести клиентов
- Мониторинг интеллектуальных счётчиков
- Мониторинг оборудования
- Оптимизация ценовой политики
- Оптимизация автомобильного движения
- Анализ социальных сетей

- Анализ поведения клиентов
- Оптимизация ИТ инфраструктуры
- Прогноз погоды
- Разведка природных ресурсов
- Управление контентом для предоставления в судебной практике
- Аналитика в медицинских исследованиях
- Управление гарантийными обязательствами
- Анализ рекламных компаний
- Исследования в области естественных наук

Примеры использования Big Data в отраслях

Анализ больших данных помогает прогнозировать поведение клиентов, повышать продажи, выявлять мошенников и предотвращать аварии на производстве. Расскажем, как можно использовать большие данные в разных сферах и покажем кейсы **Big data** от реальных компаний.

Big data в промышленности: предсказание аварий и оптимизация производства

Предиктивная аналитика. Сейчас на производстве часто внедряют IoT-системы: устанавливают датчики на оборудовании и в помещениях, а потом анализируют собранные ими данные. Эти данные и есть **Big data**, их можно использовать для мониторинга состояния оборудования, моделирования производственных процессов, выявления и предотвращения сбоев.

Например, у «Газпромнефти» сбоил автоматический перезапуск насосов после аварийного отключения электричества. Разобраться в проблеме помогли большие данные. Аналитики собрали 200 миллионов записей с контроллеров систем управления, проанализировали их, смоделировали события и выявили неожиданные причинно-следственные связи. В итоге сбои удалось прекратить.

Снижение стоимости продукции и оптимизация производства. Если собрать много данных о работе станков, проценте брака и каждом этапе производства, а потом их проанализировать, можно понять:

- при каких условиях чаще всего происходит брак;
- на какие этапы производства тратится больше всего времени и почему;
- какие тесты продукции малополезны и не дают новой информации;
- как можно оптимизировать и ускорить работу на отдельных этапах;
- как использовать меньше расходных материалов.

Все это помогает уменьшить издержки и снизить стоимость производства, а значит, повысить прибыль.

Например, компания **Intel** производит процессоры. Перед поставкой в магазин каждый процессор должен пройти примерно 19 000 тестов – это долго и дорого. Анализ данных всего производственного процесса помог понять, какие тесты избыточные. В итоге на них удалось сэкономить около 30 миллионов долларов.

Поиск новых месторождений. При добыче природных ресурсов месторождения часто приходится искать почти вслепую. Однако с помощью анализа больших данных можно обнаруживать закономерности, изучать состояние почв, наличие подземных пустот, температуру пород – и таким образом эффективно искать перспективные месторождения, сравнивая новые участки с уже известными аналогами.

Так, ООО НПЦ «Геостра» занимается обработкой и интерпретацией больших объемов данных, полученных в ходе поиска нефтяных месторождений. В качестве пилотного проекта на облачную платформу Mail.ru Cloud Solutions перенесли геофизическое ПО для обработки высокоплотных геофизических данных. Проект оказался успешным: облачные вычислительные мощности справились с поставленной задачей.

Big data в логистике: **планирование грузоперевозок и оптимизация маршрутов.**

Планирование грузоперевозок. В логистике на перевозку товаров влияет много разных факторов: загрузка складов, пробки на дорогах, состояние парка машин, расположение автозаправок. Если собрать все эти факторы вместе, сопоставить их и проанализировать, можно эффективнее планировать маршруты и время доставки, чтобы избежать простоев транспорта.

Например, компания ПЭК запустила Центр управления перевозками на базе **Big Data**. Это помогло им прогнозировать загрузку 189 складов по всей России на месяц вперед и планировать маршруты грузового транспорта.

Сокращение времени доставки. Учет разных факторов перевозки товаров помогает не только планировать грузоперевозки, но и сокращать время доставки: выбирать самые короткие маршруты, избегать пробок и трудных участков пути, экономить бензин.

Например, в логистике есть «проблема последней мили» – она стоит примерно 28% от общей стоимости доставки. Так происходит, поскольку водителю приходится заезжать во дворы, искать парковку, останавливаться и разворачиваться.

В службе доставки DHL решили это исправить. Они стали анализировать «последние мили» и оптимизировать маршруты, собирая данные с дорог и GPS. В итоге им удалось сократить время на доставку и снизить расход топлива.

РЖД внедрили технологии Big Data совместно с компанией SAP. Данные технологии помогли сократить срок подготовки отчетности в 43,5 раза (с 14,5 часов до 20 минут), повысить точность распределения затрат в 40 раз. Также Big Data были внедрены в процессы планирования и тарифного регулирования. Всего компаний используется более 300 систем на базе решений SAP, задействовано 4 дата-центра, а количество пользователей составило 220 000.

Big data в ритейле: персональные предложения и оптимизация выкладки товаров

Повышение продаж. Информация о поведении клиентов в магазине или на сайте – это большие данные. На их основе можно предполагать, что именно люди будут покупать, и использовать это для повышения продаж:

- предлагать подходящие сопутствующие товары во время покупок;
- устраивать акции и скидки на товары, актуальные в это время для большинства клиентов;
- рассылать персональные скидки и предложения, например, предлагать молодым мамам скидки на детские товары.

Например, онлайн-ритейлер Amazon использует большие данные для системы рекомендаций товаров. Их система основана на машинном обучении – она учитывает

поведение других покупателей, ваши предыдущие покупки, время года и десятки других факторов.

В итоге 35% всех продаж в Amazon генерируют рекомендации, а 86% пользователей сервиса утверждают, что рекомендации влияют на их решения о покупке.

А в сети супермаркетов «Лента» работает система лояльности – анализируются данные о покупках клиента, после чего ему предлагают персональные скидки. Например, система может понять, что вы сели на диету, и предлагать вам скидки на диетические продукты.

Оптимизация выкладки товаров. Для расположения товаров на полках тоже можно использовать большие данные: анализировать предпочтения покупателей, информацию об ассортименте, форму и цвет упаковки, чтобы повысить продажи.

В сети магазинов «Карусель» внедрили систему умной выкладки товаров. Она собирает данные, анализирует их и потом подсказывает продавцам, как сформировать витрину. Судя по отчетам, выкладка эффективна – продажи удалось увеличить на 5–10% в зависимости от товарной категории.

Big data в финансах: оценка платежеспособности и повышение качества сервиса

Оценка платежеспособности. Банкам важно выдавать кредиты только тем, кто точно сможет их вернуть, чтобы не понести убытки. Анализ больших данных помогает анализировать платежеспособность клиентов и оценивать риски.

Например, Mastercard работает не только как платежная система – она собирает данные, которые помогают выявлять неплатежеспособных контрагентов, не возвращающих кредиты. Mastercard предупреждает финансовые организации, что с этими организациями не стоит вести дела.

Улучшение клиентского сервиса. Big data в банках также используют для того, чтобы делать клиентам персонализированные предложения. Это как в интернет-магазинах, только в качестве «рекомендуемых товаров» выступают банковские продукты и услуги.

Так, Альфа-Банк собирает данные обо всех своих клиентах. Затем с помощью анализа и сегментации делит их на группы. Например, клиент раз в неделю покупает подгузники или детские смеси, значит, скорее всего, у него есть ребенок. И можно предложить кредит или бонусную программу на детские товары.

Big data в HR: наем сотрудников и предупреждение увольнений

Наем сотрудников. На начальном этапе найма сотрудников часто требуется отсеять тех, кто мало заинтересован в работе или совсем для нее не подходит. Эту задачу можно решать с помощью больших данных: собирать информацию о кандидатах и резюме, выявлять в них закономерности, использовать эти данные для разработки скриптов или обучения роботов и нейросетей.

Например, компания Stafory разработала робота-рекрутера Веру. Он умеет сортировать резюме, обзванивать сотрудников, распознавать голос и выделять наиболее заинтересованных кандидатов. Компании уже используют робота для подбора персонала. В частности, Вера помогла PepsiCo заполнить 10% от необходимых вакансий.

Оптимизация HR-стратегии. Компании часто анализируют поведение клиентов,

и по тем же принципам можно анализировать поведение сотрудников: отслеживать эффективность их работы, переработки, признаки усталости или выгорания.

В Google есть отдел People Analytics, который анализирует большие данные, связанные с поведением сотрудников. У них есть несколько успешных кейсов применения **Big Data**:

1. Еще в 2002 году в компании проанализировали работу тысяч менеджеров и создали «8 стратегий поведения менеджеров Google». Сейчас стратегии регулярно дополняют и используют при найме и обучении сотрудников.
2. Аналитики постоянно отслеживают поведение и состояние сотрудников: сколько они зарабатывают, часто ли задерживаются на работе, насколько эффективны. На основе этого принимают решение о дополнительных выплатах или продлении отпусков.
3. Специальные алгоритмы предупреждают, что конкретный сотрудник в ближайшее время захочет уволиться. Это помогает менеджерам вовремя среагировать.

Big data в медицине: прогноз заболеваний и сбор данных о пациентах

В медицинской сфере большие данные в перспективе можно использовать для диагностики и лечения, большинство интересных проектов пока находятся на стадии разработки или тестирования, но есть и уже реализованные.

Прогнозирование заболеваний. Если собрать достаточно данных о пациентах, можно делать предположения о том, чем они больны сейчас или могут заболеть в ближайшее время.

Так, в детской больнице Торонто внедрили проект Artemis. Больничная система собирает и анализирует данные по новорожденным — она каждую секунду анализирует 1260 показателей. На основе этих данных система может предсказать нестабильное состояние ребенка, чтобы ему смогли вовремя помочь.

Ведение базы пациентов. У многих пациентов длинная история болезни, которая часто хранится в разных больницах и у разных врачей. Чтобы увидеть полную картину, нужно собрать данные в единую базу. С помощью технологий **Big Data** можно не только организовать такую базу, но и настроить в ней удобный поиск и аналитику.

Например, в Массачусетской больнице общего профиля создали систему QPID, которая собирает электронные данные о пациентах и быстро предоставляет нужную информацию: и пациентам, и врачам. К примеру, пациент может посмотреть информацию по своей болезни: анализы, диагнозы, снимки, назначенные лекарства. А врач может увидеть информацию о хронических заболеваниях и прошлом лечении.

Big data в образовании: помощь в выборе курсов и предотвращение отчислений

Помощь в выборе курсов. В образовании проекты **Big Data** помогают студентам с профориентацией: анализируют их способности и помогают выбрать направление обучения и будущую профессию.

Так, в американском университете Остин Пии разработали рекомендательную систему подбора курсов. Она собирает данные об успеваемости, находит «похожих» студентов, и на основе этого подбирает курсы для конкретного человека. Предсказания устраивают студентов в 90% случаев.

Предотвращение отчислений. В США из университетов отчисляются 400 тысяч студентов в год. Чтобы решить эту проблему, в Университете Содружества

Виргинии проанализировали данные об отчислениях и построили алгоритм, который выявляет студентов в группе риска.

Система оповещает, когда студент становится проблемным. И тогда с ним работают индивидуально, например, предлагают перевод на другой курс или помощь репетитора. По итогам семестра число студентов, закончивших курс, увеличилось на 16%.

Big data в маркетинге: повышение прибыли и привлечение клиентов

Создание коммерчески успешных продуктов. Большие данные о поведении клиентов помогут предсказывать спрос и позволяют до вывода продукта на рынок понять, будет ли он успешным.

Например, такие технологии использует Netflix. Этой платформой для просмотра фильмов и сериалов пользуются более 150 миллионов человек. В компании анализируют поведение клиентов: какие сериалы они смотрят, какие бросают, какие моменты перематывают. Это помогает лучше понимать психологию зрителей и грамотно рекомендовать им новые сериалы.

Еще Netflix анализирует поведение зрителей, чтобы снимать успешные сериалы и эффективно их продвигать. Например, перед созданием «Карточного домика» в компании проанализировали 30 миллионов сценариев, 4 миллиона зрительских оценок и 3 миллиона поисковых запросов.

Таргетированная реклама и снижение стоимости привлечения клиента. Big data помогает лучше настраивать целевые аудитории и показывать таргетированную рекламу более точечно.

Например, ритейлер Ozon использует большие данные для таргетированной рекламы и рекомендации товаров. Для этого на сайте и в мобильном приложении собирают логи пользователей — фиксируют всё, что они просмотрели, пролистали, на что кликнули. На основе данных составляют прогноз: планирует ли пользователь покупку, товар какой категории, скорее всего, его заинтересует. Релевантные товары показывают в таргетированной рекламе.

Также в Ozon тестировали полки рекомендаций для различных товаров. Пользователей разделили на две группы: для первой рекомендации вручную составили эксперты, для второй — собрали автоматически на основе данных логов. В итоге во второй группе продажи оказались в 10 раз ваше.

В компании Nestle Purina начали использовать платформу для сбора данных о клиентах. Они проанализировали поведение покупателей и выделили в отдельную категорию людей, которые недавно искали в интернете щенков. С помощью таргетированной рекламы в Facebook этим клиентам показывали товары для щенков. Благодаря такому подходу конверсия выросла на 300%, а стоимость привлечения клиента снизилась на 90%.

Big data в госструктурах

В России технологии Больших Данных также стали осваивать такие государственные органы, как Пенсионный Фонд, Федеральная Налоговая Служба и Фонда обязательного медицинского страхования. Потенциал реализации проектов с использованием Big Data большой, данные технологии могли бы помочь в улучшении качества сервисов, и, как следствие, уровня жизни населения.

Перспективы Big Data

Сейчас почти все крупные компании используют большие данные: собирают их, анализируют, применяют в связке с другими технологиями. Они занимают около 60% рынка. Но по оценкам экспертов, сегмент средних и малых предприятий до 2027 года тоже будет расти, со временем большие данные перестанут быть прерогативой больших компаний.

Для работы с большими данными необязательно строить свою инфраструктуру и тратить деньги на ее обслуживание. Можно арендовать готовое решение в облаке. Например, платформа Mail.ru Cloud Solutions позволяет хранить, обрабатывать и анализировать данные, используя в том числе машинное обучение и инструменты визуализации. При этом вы платите только за используемые вычислительные мощности.

Большие данные мало просто собрать, их нужно как-то использовать, например, чтобы строить прогнозы развития бизнеса или проверять маркетинговые гипотезы. А для использования данные требуется структурировать и анализировать.

Что же такое **аналитика данных** – это процесс преобразования первичных данных в знания и полезную информацию, которую можно использовать для принятия решений, что позволит оптимизировать деятельность: увеличить доход, сократить затраты или достичь других важных результатов.

Технологии для аналитики больших данных:

1. **Предиктивная аналитика** – предсказание будущего на основе собранных данных.

2. **Краудсорсинг** – ручной анализ силами большого количества людей.

3. **Смешение и интеграция данных** – приведение данных из разных источников к одному виду, уточнение и дополнение данных.

4. **Распознавание образов** – отнесение исходных данных к определенному классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы данных.

5. **Машинное обучение и нейронные сети** – создание программ, которые умеют анализировать и принимать решения, выстраивая логические связи.

6. **Имитационное моделирование** – построение моделей на основе больших данных, которые помогают провести эксперимент в компьютерной реальности, без влияния на реальное положение вещей.

7. **Статистический анализ** – подсчет данных по формулам и выявление в них тенденций, сходств и закономерностей.

8. **Пространственный анализ** – набор алгоритмов (функций), обеспечивающих анализ местоположения (размещения), связей и иных пространственных отношений пространственных объектов, включая анализ зон видимости/невидимости, анализ соседства, анализ сетей, создание и обработку цифровых моделей рельефа и т.д.

9. **Data Mining** – технология добычи новой значимой информации из большого объема данных.

10. **Визуализация** – представление больших данных и результатов их анализа

в виде удобных графиков и схем, понятных человеку.

Приведем жизненный пример того, как и при каких условиях необходимо производить анализ данных.

Представим, для аналитика стоит задача оценить продажи ряда магазинов (куда поставляется товар) и каждый магазин ведет свой учет проданных товаров. Реальность такова, что формы учета будут заполнены разными способами, то есть у них будет разная структура и разный формат хранения (некоторая форма таблиц).

Решим эту задачу способом, который первый приходит в голову, то есть предоставим простейший способ решения данной задачи. Пусть у нас есть N таблиц и нам нужно их собрать вместе в одну таблицу. Напишем N скриптов, которые преобразуют эти таблицы, и один сборщик, который собирает их вместе. У такого подхода есть положительные и отрицательные стороны. К минусам можно отнести: необходимо поддерживать N скриптов одновременно (где N в порядках тысяч); при изменении структуры отчетов магазинов во времени (например, в магазине появился новый сотрудник) необходимо искать и переписывать отдельные скрипты; при появлении нового магазина, необходимо писать новый скрипт; при изменении формата отчетности необходимо вносить изменения во все скрипты; сложная отладка и поддержка, так как магазины не уведомляют об изменении структуры и не следуют никаким спецификациям.

Немного усложним задачу и приблизим ее к реалиям современных компаний. Итак, производитель товаров на самом деле работает не напрямую с магазинами, а через некоторых посредников. Посредники посещают магазины и непосредственно своими действиями пытаются стимулировать продажи. Соответственно, они являются материально заинтересованными лицами и информацию, которую они выдают, приходится перепроверять.

На основании вышесказанного давайте переформулируем задачу. Пусть у нас есть N магазинов и K дистрибьюторов. Можем ли мы агрегировать данные магазинов и сравнить их с результатами дистрибьюторов? При этом у всех данные имеют разную структуру и формат.

Эта задача обладает несколькими основными характеристиками. Входные данные представляют собой огромное количество разнообразных форматов, к которым добавляются отчеты дистрибьюторов. Как правило, задача характеризуется очень низким качеством данных, в том числе дублированием, несогласованностью и ошибками. На основе полученных результатов и сравнения данных отдел по закупкам принимает решения о том, сколько, что, кому и по какой цене отгружать. То есть решение этой задачи непосредственно влияет на финансовые показатели компании, что безусловно важно.

Существуют также другие варианты решения подобной задачи. Например, самописное решение. Компании производителю будет необходимо нанять специалиста не по профилю компании и критичное ПО будет зависеть от данного специалиста. Если он уйдет, то компания будет вынуждена срочно искать замену, которая сможет поддерживать ПО, и качество будет напрямую зависеть от нанятого специалиста.

Также всегда есть возможность закупить ПО у третьей стороны. Отрицательной стороной такого подхода являются цена, качество и время интеграции. Как правило, цена и время интеграции слишком высоки для среднего производителя и в том числе требует существенных временных затрат сотрудников. Выбор поставщика также не

тривиален.

Есть вариант применить SaaA решения, но методология еще нова для рынка и многие компании скептически относятся к подобным сервисам.

Процесс анализа на уровне компании в общем случае выглядит следующим образом: данные консолидируются, определенным образом трансформируются (агрегируются) и загружаются в систему для анализа.