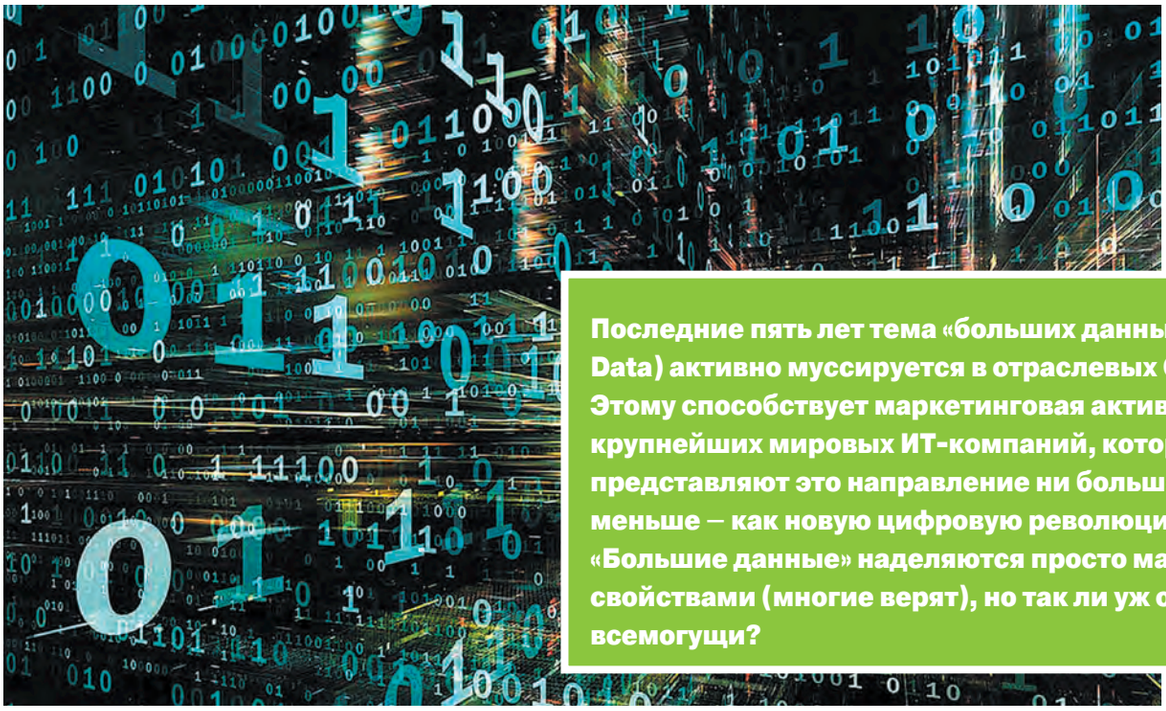


Последние пять лет тема «больших данных» (Big Data) активно муссируется в отраслевых СМИ. Этому способствует маркетинговая активность крупнейших мировых ИТ-компаний, которые представляют это направление ни больше ни меньше — как новую цифровую революцию. «Большие данные» наделяются просто магическими свойствами (многие верят), но так ли уж они всемогущи?

Большие данные — большие надежды

Любое производство так или иначе формирует большое количество отходов, утилизация которых часто превращается в прибыльный бизнес, а иногда даже формирует совершенно новые отрасли промышленности. Еще древние шумеры, добывая ценный для них битум, старались избавиться от вредной горючей примеси, которую они называли «нафта». Гораздо позже при производстве керосина «ненужный» бензин попросту сжигали в специальных ямах. Роль отходов долго отводилась и мазуту, и природному газу, и еще многим другим полезным сегодня веществам и элементам (например, платине).

В условиях «цифровой экономики» многие компании не «производят» ничего кроме данных, большая часть из которых — это «цифровые отходы» (часто, для важности, именуемые «архивом»). И тут как бы сама собой напрашивается аналогия с промышленностью и возникает вопрос: нельзя ли извлечь дополнительную пользу из вроде бы ненужных сведений? Ведь их так много, наверняка там есть что-то ценное. Примерно из таких соображений и родилась концепция, которая со временем получила название «больших данных» (Big Data). Для того чтобы сделать аналогию с промышленностью еще более очевидной и придать информации вес категорий реального сектора, данные часто стали именовать таким эпитетом, как «новая (или цифровая) нефть».

Нефть, конечно же, не имеет ничего общего с информацией, но это сравнение простое и понятное, а значит, имеет все шансы прочно укрепиться в массовом сознании. Есть лишь одна проблема — нефть продать легко, она нужна всем, а ценность данных очевидна далеко не всегда. Но по мере осознания их стоимости в мире начал активно формироваться новый сектор бизнеса — продажа данных огромного количества пользователей, которые получены, например, из социальных сетей или других массовых источников и могут использоваться в качестве репрезентативной выборки для анализа и исследования предпочтений аудитории.

Big Data — начало

Мир тонет в огромных массивах ненужной информации. Уже сегодня люди, пытающиеся оценить объем мировых данных, оперируют такими величинами, как эксабайты и зеттабайты. Год от года ситуация усугубляется, глобальный поток данных нарастает лавинообразно. При этом очевидно, что 99,999...% из них, скорее всего, никогда не будут использованы повторно, к тому же данные эти в общем случае не структурированы. Говорят, что сегодня человек, находясь в цивилизованной стране, получает за неделю больше информации, чем его предок двести лет назад мог узнать за всю жизнь. Возможно, так оно и есть, но если мы будем рассматривать информацию

в контексте полезности, то условный предок окажется в выигрыше, поскольку он не был перегружен лишним информационным шумом. Зато современный человек не сможет овладеть навыками своего предка ни за неделю, ни за год, хотя данных он получит в разы больше. Да, можно сказать, что мир усложнился, но ведь и системы, регулирующие его существование, тоже стали более совершенными (вам, например, уже не надо помнить наизусть номера телефонов друзей или адреса веб-сайтов, и это самое простое из огромного числа вспомогательных технологий). Так что основная масса современных данных — это мусор.

Но может быть хотя бы из небольшой части этой огромной совокупности можно извлечь что-то полезное? Примерно с 2012 года все крупнейшие мировые производители пытаются убедить рынок в том, что это вполне реально, а «большие данные» таят много скрытых возможностей, для раскрытия которых только и нужно, что купить фирменный набор инновационных решений. Иногда даже встречаются успешные проекты. Но — обо всем по порядку.

Как это часто бывает, удачный термин или название, придуманное для одного явления, необъяснимым образом начинает использоваться для именования совершенно иных вещей. Это и произошло с понятием Big Data, первый зафиксированный случай упоминания которого относится к 1997 году. Произошло это на одной из конференций IEEE. Но тогда речь шла лишь о нехватке емкости основной памяти жесткого диска для выполнения задач виртуализации.

ТЕРМИНОЛОГИЯ BIG DATA

Big Data («большие данные») – основной термин, подразумевающий большие массивы неструктурированных данных (в ряде случаев также технологии и методы их обработки). При этом вопрос о том, какие объемы считать «большими», все еще четко не определен. В общем случае имеются в виду показатели, которые в разы превосходят некое среднее значение по ИТ-отрасли в каждый момент времени.

Data Lake («озеро данных») – информационный массив, содержащий большое количество исходных неструктурированных данных.

Data mining («добыча информации») – совокупность методов анализа данных, используемых с целью выявления полезных закономерностей.

Data science («наука о данных») – направление в науке и околонаучных кругах, изучающее вопросы, связанные с анализом, обработкой и представлением цифровой информации в воспринимаемой форме.

Зато настоящую популярность термин обрел более чем через десять лет после статьи в журнале Nature, где Клиффорд Линч, главный редактор издания, рассуждал о взрывоподобном росте научных данных, а также проблемах и перспективах, которые связаны с этим процессом.

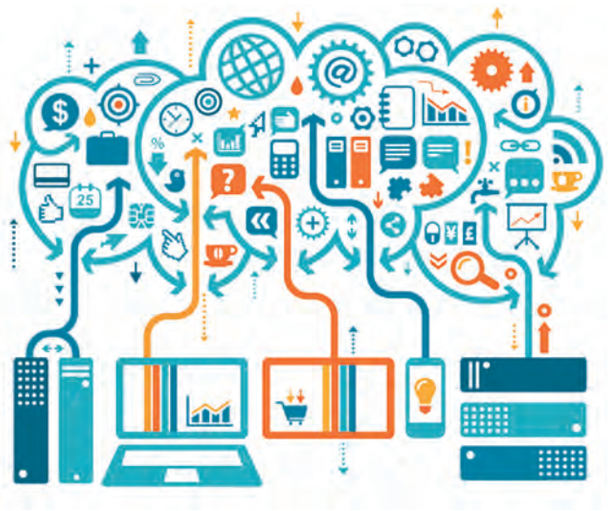
Хотя само понятие «больших данных» в статье было определено весьма абстрактно, термин пришелся по душе многим коммерческим компаниям, которые занимались разработкой и продажей комплексных систем, предназначенных для работы с большими массивами информации. А благодаря СМИ словосочетание Big Data стало известно, наверное, всем, кто имеет доступ к Интернету, хотя на самом деле четкого описания термина до сих пор нет. Имеются лишь более-менее определенные трактовки, которые с той или иной долей консенсуса приняты в ИТ-сообществе. При этом все имеющиеся разногласия возникают лишь вокруг понятия «большие», о том, что такое «данные», споров идет гораздо меньше.

Поначалу были попытки определить конкретные объемы, назывались цифры 100 ГБ в день и более, однако впоследствии от четких рамок фактически отказались, зато термин оброс дополнительными понятиями, основные из которых приведены во ***Врезке***.

Причем тут Google?

Но есть ли еще что-то, что отличает «большие данные» от «обычных», помимо условного объема? В принципе, да. Это технологии обработки. И здесь надо немного заглянуть в историю развития Интернета, точнее поисковых машин. В конце 90-х годов число веб-сайтов, каждый из которых содержал множество страниц, исчислялось десятками миллионов. Для успешного поиска их все надо было проиндексировать. Уже тогда было понятно, что традиционными методами обработки данных такой объем чрезвычайно трудно «переварить».

Каждая поисковая система решала эту задачу по-своему. Так, Alta Vista (детище DEC) использовала кластеры мини-компьютеров VAX, Excite полагался на мощные Unix-серверы Sun Microsystems, а Yahoo поначалу вообще использовал труд специально обученных людей, которые каждый день в прямом смысле слова просматривали и индексировали веб-страницы (сейчас трудно поверить, но пока Интернет был не таким большим, этот подход был достаточно эффективным).



Однако самый прогрессивный подход использовала компания Google. Опуская множество подробностей и выделяя главное, можно сказать, что ее поисковая система прекрасно масштабировалась и была готова к восприятию практически любого объема данных («больших данных»). В результате, когда количество веб-страниц стало стремительно расти, технологии, предложенные Google, довольно быстро превратили компанию в доминирующего игрока на рынке. Даже сейчас, когда в Интернете свыше 2 млрд сайтов, данная поисковая система не испытывает проблем. Собственно, подходы, предложенные Google, и легли, в том или ином виде, в основу алгоритмов работы с «большими данными» во всех сферах, даже за пределами Интернета.

Базовыми составляющими методологии стали три компонента: GFS, Map Reduce и BigTable. Был и еще один революционный момент — в качестве аппаратной основы Google использовала не дорогие серверы, как ее конкуренты, а большой массив обычных дешевых ПК. Нюанс в том, что создателям поисковика удалось сформировать из этого массива суперкластер, работающий как единая вычислительная система. Чтобы он мог эффективно функционировать, в 2000 году пришлось разработать специальную файловую систему GFS (Google File System), ее детали держатся в секрете, но общие принципы построения были опубликованы в 2003 году. Главное, она обеспечивает когерентность (общее использование) памяти для всех вычислительных узлов, входящих в систему. Технология MapReduce позволяет распределять большую задачу между огромным количеством узлов (можно использовать сотни, тысячи серверов или сколько будет необходимо), а BigTable — это специальная не реляционная БД, оптимизированная для работы с большими объемами информации.

Главным итогом всей этой истории явился технологический прорыв, навсегда изменивший лицо ИТ-отрасли. Теперь масштабные вычисления стали достаточно дешевыми, чтобы превратиться в массовое явление. Задачи, которые раньше требовали очень мощных и дорогих серверов, теперь стало возможно решать на базе массивов недорогих систем. Более того, доступными стали задачи

столь огромные по своему масштабу, что ни один единичный сервер не смог бы их обработать. Отсюда до «больших данных» было уже рукой подать. Собственно, супермасштабные задачи были и раньше, но отсутствовал подход, позволяющий их эффективно обработать.

Что [не]могут «большие данные»

Спектр задач, решаемых с помощью «больших данных», сегодня чрезвычайно широк. Инструменты Big Data используются повсеместно — от решения научных задач на Большом адронном коллайдере до предсказания эффективности маркетинговых компаний. В Сети без труда можно отыскать множество историй успеха на эту тему. Приведем здесь лишь несколько примеров.

О положительных результатах использования «больших данных» заявляет, например, Procter & Gamble. Компания считает, что они помогают ей в процессе прогнозирования деятельности. Target (входящий в тройку мировых лидеров розничной торговли) сообщает, что анализ «больших данных» о покупателях помог увеличить прибыль на 15–30%. Министерство труда ФРГ использует технологии Big Data для анализа заявок на выдачу пособий по безработице, благодаря такому подходу удалось уменьшить издержки на 10 млрд евро. В ряде штатов США полиция использует большие массивы данных для предсказания наиболее вероятных мест совершения преступлений (говорят, достаточно эффективно).

Но наиболее активно Big Data используют банки, операторы связи и представители рекламного бизнеса. Например, Chase Bank — один из крупнейших представителей финансового сектора США, активно работающий в сфере ипотечного кредитования, использует инструментарий «больших данных» для эффективного предсказания банкротства физлиц, получивших ипотеку. Кроме того, банк просчитывает также и вероятность того, кто из заемщиков готов погасить кредит досрочно (таким клиентам предлагаются дополнительные продукты или их просто перепродают другим банкам). Из операторов можно назвать Telenor, который благодаря «большим данным» смог сократить отток клиентов на 36%.

Проблема всех этих воодушевляющих примеров состоит в том, что ни в одном случае невозможно четко определить, что стало решающим фактором, который привел к успеху, и какова истинная роль «больших данных». Так, сокращение издержек Министерства труда ФРГ могло быть вызвано введением более жесткой политики удовлетворения заявок, повышение прибыли Target — совокупным

успехом общей маркетинговой активности, а отток клиентов Telenor мог прекратиться по естественным причинам (к определенному моменту ушли все, кто хотел это сделать). Но вот с чем трудно спорить, так это с эффективностью применения «больших данных» в банковском секторе, где многофакторный анализ клиентов позволяет создавать новые прибыльные продукты.

Однако критиков у всего, что связано с Big Data, не меньше, чем сторонников. О провальных проектах распространяться не принято, но по некоторым сведениям, до 2/3 всех внедрений оканчиваются неудачей, а в тех проектах, которые все же удалось довести до завершения, почти во всех случаях наблюдаются проблемы или же результаты внедрения неочевидны. Так происходит по многим причинам, например, из-за нехватки специалистов и достаточно «сырого» состояния многих продуктов, рассчитанных для работы с «большими данными». Но главная проблема состоит в правильной постановке задачи, а это, во многих случаях, — целое искусство, от которого зависит успех проекта. В общем виде проблему можно сформулировать примерно так: информации настолько много, что на ее основе можно получить любой ответ, соответствующий заданным характеристикам, при этом невозможно определить, является ли этот ответ верным или оптимальным.

Но когда задача четко поставлена и сформулирована, решение можно найти традиционными методами. В этом контексте стоит еще раз вспомнить Google, точнее самый известный провал Big Data, связанный с компанией, который произошел в 2013 году. Тогда специализированный сервис Google Flu Trends не смог предсказать эпидемию гриппа и допустил ошибку на 140% (проект был запущен в 2008 году и преподносился в качестве эффективного инструмента предсказания вспышек заболевания в том или ином регионе). Этот факт не свидетельствует против «больших данных» (несколько лет алгоритм работал замечательно), но лишь говорит о том, что любую модель необходимо постоянно уточнять, пересматривать, а наличие большой и актуальной выборки — лишь один из факторов успеха.

Отметим, что в сфере массовых коммуникаций использование Big Data очень популярно для создания персонализированной рекламы, которая, как считается, более эффективна по сравнению с традиционным подходом (более детально мы рассмотрим этот вопрос в следующем разделе). Справедливость этого утверждения оставим на совести маркетологов, но наверняка каждый сталкивался с тем, что после поиска какого-то товара, скажем, через Google, целевые сообщения еще долго продолжают вас преследовать, даже если вы уже совершили покупку или однозначно решили от нее отказаться. Более того, даже если какой-то запрос был сделан просто из любопытства (например, в целях повышения эрудиции вы захотели узнать, что такое оверлок или, скажем, дарбука), скорее всего эти категории товара будут длительное время выдаваться вам в виде целевых рекламных сообщений. В общем, технология еще далека от совершенства, но кое-кто пытается представить все, что связано с Big Data в совсем уж фантастическом свете.

«Большие данные» для Президента

В контексте вышесказанного интересна, например, недавняя история с предвыборной компанией Президента США Дональда Трампа, в ходе которой якобы именно «большие данные» (в совокупности с другими инструментами) сыграли очень значительную роль. Напомним, речь идет о статье, которая появилась в швейцарском журнале Das Magazin в конце 2016 года, где, в частности, речь шла о том, что огромный массив информации («больших данных»), полученный из социальных сетей и проанализированный с применением математических и психологических методов, позволил составить очень точные психологические портреты пользователей и использовать т.н. микротаргетирование целевых сообщений.

Иными словами, якобы к каждому потенциальному избирателю был найден индивидуальный подход, позволяющий в итоге манипулировать общественным мнением. Именно с данной целью предвыборным штабом Трампа и была нанята компания Cambridge Analytica. Все это было преподнесено буквально как революция в коммуникационной и рекламной отрасли. Шутка ли, по словам авторов статьи, эффективность политической рекламы в социальных сетях выросла на 1400% по сравнению с традиционными методами агитации. Этот «феноменальный результат» настолько впечатлил непрофильные мировые СМИ, что они наперебой бросились тиражировать материал, даже не разобравшись в сути происходящего, в результате «большие данные» стали восприниматься чуть ли не как инструмент оруеловского «Большого брата». Мол, все мы теперь под колпаком. Но более детальный анализ публикации показал, что все не совсем так, как могло показаться на первый взгляд, а высокие показатели эффективности — всего лишь результат некорректной оценки и статистических манипуляций.

Начнем с того, что упомянутый Das Magazin всего лишь легкое воскресное приложение с очень широким кругом тем к более серьезной газете Tages Anzeiger, а сфера ИТ вообще-то не характерна ни для того, ни для другого издания. Но самый важный момент связан с тем, что упомянутый рост конверсии в 1400% — это ничем не подтвержденная цифра, буквально взятая с потолка. Ее озвучили всего два источника, которые нельзя считать независимыми — руководитель Cambridge Analytica Александр Никс и сотрудник Стэнфордского университета Михал Косински (соавтор метода, положенного в основу работы системы).

Нюанс заключается в том, что аналитики могут подсчитать лишь количество доставленных таргетированных сообщений, но никак не их эффективность (насколько больше людей действительно проголосовали за нужного кандидата). На поведение избирателей, особенно в США, влияют десятки факторов, и какой из них сыграл решающую роль, пока понять невозможно, и этому не помогут даже самые современные технологии. Но почему же предвыборный штаб Трампа последовательно заключил с Cambridge Analytica целую серию контрактов

на общую сумму около $15 млн? Возможно, методы компании были признаны эффективными, но вполне вероятно, свою роль сыграл и тот факт, что в совет директоров Cambridge Analytica входит Стив Бэннон — один из руководителей избирательной компании нынешнего президента США.

К тому же штаб Хиллари Клинтон — основного конкурента Дональда Трампа — тоже активно использовал методы глубокой социальной аналитики, «большие данные» и таргетированную агитацию, для чего было задействовано более шестидесяти высококлассных специалистов в области прикладной математики, статистики и других наук. Правда, оценки эффективности использования подобных методов оказались в разы ниже, чем упомянутые 1400%.

Но вернемся к кампании Трампа и подходам Cambridge Analytica. Основным способом получения психологического портрета потенциального избирателя здесь был усовершенствованный метод OCEAN — достаточно спорная (если судить по доступным результатам ее применения) концепция, которая якобы позволяет описать тип личности на основе пяти параметров: доброжелательность, добросовестность, невротизм, открытость и экстраверсия. Исследования в этой сфере ведутся с 80-х годов прошлого столетия и представляют собой всего лишь одно из множества направлений изучения личности, без явного преимущества над остальными подходами. Активную научную работу в данной сфере осуществляет и упомянутый Михаил Косински, чьи наработки были использованы Cambridge Analytica. OCEAN — достаточно спорная концепция, если судить по доступным результатам ее применения, однако новизна метода, использованного в кампании Трампа, заключалась как раз в привлечении «больших данных», под которыми главным образом подразумевались индивидуальные предпочтения миллионов участников соцсетей. На основе данных о поставленных «лайках» пользователи разбивались на таргет-группы и для них формировались целевые рекламно-агитационные сообщения. Но повторим, конкретные цифры эффективности применения данного метода в кампании Трампа не подтверждены и, более того, принципиально не проверяемы. Помогли ли «большие данные» повысить эффективность президентской кампании? Скорее всего, да. Превратились ли они в решающий фактор? Очевидно, нет.

Во всей этой истории примечательно то, что она как нельзя лучше отражает сегодняшнее состояние в сфере «больших данных», где имеется теоретически обоснованный математический аппарат и доступны вычислительные ресурсы для обработки гигантских массивов информации, но в то же время полученные результаты нуждаются в правильной интерпретации, иначе их ценность будет весьма спорной.

Глобальная экосистема

Как бы то ни было, но сегодня продукты для работы с большими данными разрабатывают десятки мировых корпораций (и, очевидно, тысячи региональных компаний): IBM, Microsoft, SAS, SAP, Oracle, HPE, Dell EMC — самые известные имена в этом сегменте. Кратко рассмотрим особенности наиболее популярных решений.

Так, компания ***SAP*** работает на рынке Big Data с 2007 года, сегодня в портфель решений входят несколько основных продуктов: аналитические СУБД Hana и IQ, комплексное решение SAP Hana, ПО для прогнозной аналитики KXEN и ряд других решений. Собственных аппаратных систем SAP не выпускает, предпочитая сотрудничать в этом направлении с профильными компаниями. В свою очередь, ***IBM*** предлагает как оборудование для работы с «большими данными» (PureData, Watson), так и программные продукты (СУБД DB2, InfoSphere, ПО для бизнес-аналитики Cognos, SPSS и др.). Широкий набор технологий предлагает и ***Oracle***. В числе наиболее известных решений — аналитические СУБД Database, MySQL и Essbase, СУБД в оперативной памяти Oracle TimesTen и Event Processing, комплексные системы Big Data Appliance, Exadata и Exalytics. В качестве аппаратной основы Oracle использует собственные разработки. ***Microsoft*** делает упор на облачные технологии Azure, предлагая для обработки «больших данных» такие сервисы, как Stream Analytics, Data Factory и Machine Learning. В портфолио ***HPE*** имеются такие инструменты, как облачная платформа Heven, СУБД Vertica, ПО Autonomy (для анализа информации из соцсетей) и др. Множество решений для «больших данных» предлагается также и компанией ***Dell EMC***, у которой имеются как облачные продукты, так и программно-аппаратные комплексы, специализированные СУБД (например, GemFire) и множество иных разработок. Ключевыми игроками на рынке «больших данных» являются также Cloudera, Amazon и, конечно же, Google.

Упомянутые компании предлагают комплексные инфраструктурные решения, но помимо них есть также масса специализированных поставщиков. Одни, например, разрабатывают алгоритмы, другие концентрируют усилия на сборе релевантных данных с целью продажи, есть, конечно же, интеграторы, консалтинговые компании. В общем «большие данные» сегодня — это сам по себе большой бизнес, который, пока, к сожалению, обходит Украину стороной.

Так что же такое «большие данные» — новая реальность цифровой эпохи, открывающая большие возможности, или просто удачный маркетинговый трюк. Очевидно — и то и другое. Потенциал концепции огромен, но ожидания клиентов, распаленные с помощью рекламных кампаний, пока еще слишком завышены. Возможности и сферы применимости Big Data очень часто переоценивают, а сложности реализации эффективного инструментария, напротив, преуменьшают. Но в целом при должном подходе, наличии опыта внедрений и четко поставленных задач «большие данные» способны открыть новые, невиданные доселе возможности во многих сферах жизни современного общества.

Игорь КИРИЛЛОВ, СиБ