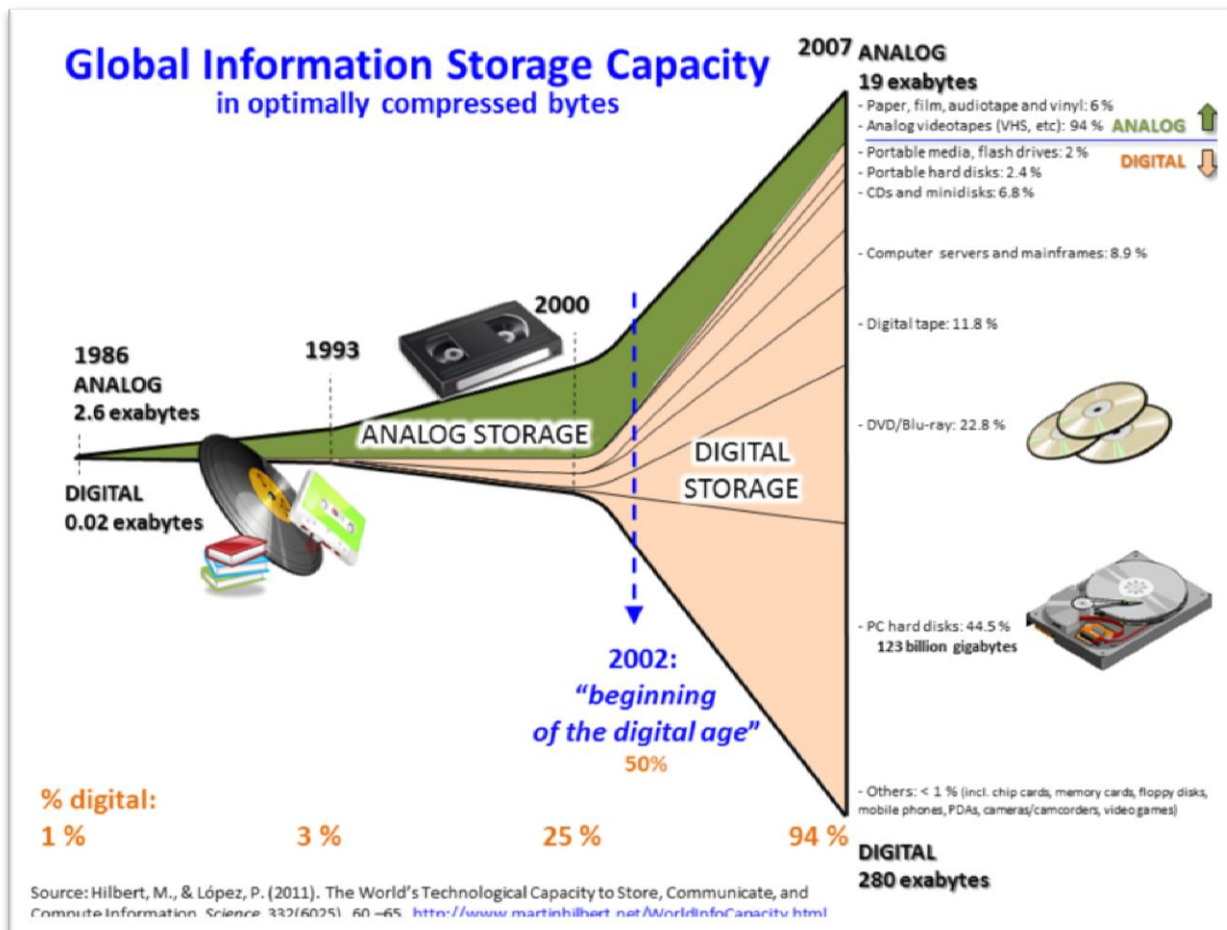


Переход от Business Intelligence к bigdata. Перспективы применения. Рассмотрение примеров существующей реализации. Предпосылки появления «больших данных»

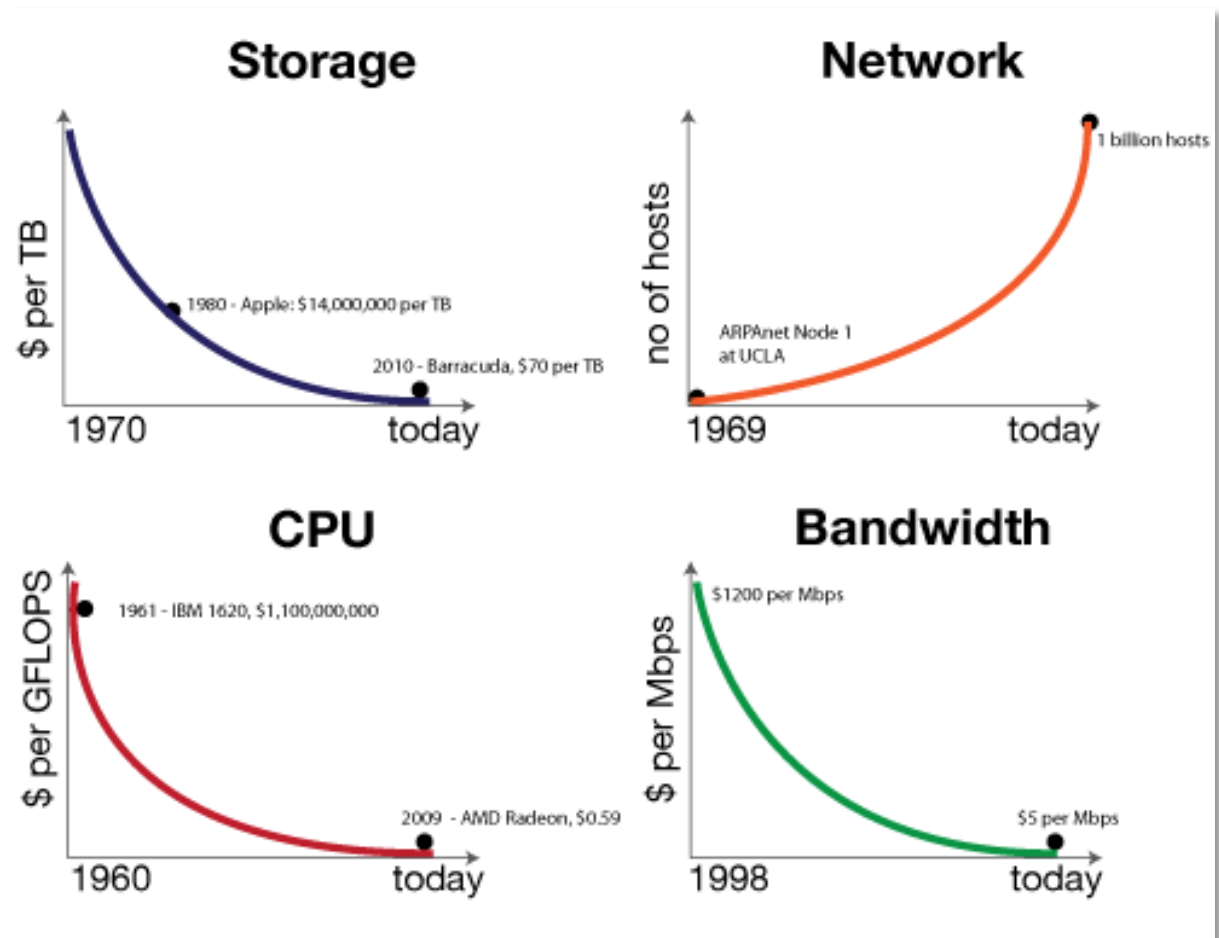


Распространение сенсоров

Увеличение пропускной способности сетей

Развитие беспроводных сетей

Развитие технологий хранения данных (удешевление)



Что же такое большие данные?

- Комбинация подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения результатов, эффективно воспринимаемых человеком
- Большие данные – это такие данные, которыми дорого управлять или из которых сложно извлечь ценность. (Майкл Франклин)

2007 год – первое упоминание термина bigdata в журнале Nature

2013 год – определение bigdata занесено в Оксфордский словарь английского языка

Отличия bigdata от «обычного» BI

	bigdata	BI
Тип данных	Неструктурированные форматы	«Колонки» и «строчки»
Объём данных	100+ терабайт	До 100 терабайт
Поток данных	Постоянный приток данных	Статический набор данных
Методы анализа	Машинное обучение	Проверка гипотез
Основная цель	Продукты	Решения

Ключевые свойства больших данных

Volume - ...да, размер имеет значение

Variety - ...у данных нет постоянной структуры

Velocity - ...они очень быстро генерируются

Value - ...у них есть реальная бизнес-ценность

Veracity - ...они достоверны (!= «белый шум»)

Vicarious - ...получены «извне» нашей Компании

BI – формирование знаний, помощь в принятии решений

bigdata – формирование новых **продуктов** на основе располагаемой информации

Проблема выбора

Разумеется, при принятии решения - применять или не применять подходы, предлагаемые «большими данными» в стратегических планах развития бизнеса – необходимо в первую очередь решить для самих себя, какая именно политика предпочтительнее: выйти вперёд по сравнению с конкурентами, вооружившись новыми знаниями и технологиями (раннее большинство) или дождаться пока все концепции и идеи станут хорошо зарекомендовавшими себя инструментами и вступить в гонку наравне с остальными (позднее большинство), потому что все уже используют «большие данные».

Конечно, сейчас, всё ещё на относительно ранней стадии развития, многие решения и идеи по-прежнему базируются на парадигме «классической» аналитики. Любой крупный бизнес, а телекоммуникационный – в особенности, располагает значительным, во многом – избыточным объёмом информации, ежедневно и ежеминутно закладываемой в недра OSS/BSS-систем. И уже из этих данных, находящихся в распоряжении оператора, можно почерпнуть сведения и знания, которые позволят поднять эффективность бизнеса на новый уровень:

- сформировать портреты типовых абонентов, в зависимости от их предпочтений в выборе услуг, информации, предоставленной при заключении договора, и предпочитаемых способов связи с различными службами;

- определить признаки лояльности и выявить причины, влияющие на лояльность абонентов, решая задачу оптимизации и управления оттоком пользователей оператора связи;
- раскрыть и описать фактические правила использования сети, с точки зрения распределения и изменения нагрузки на неё, с возможностью выдавать распоряжения и рекомендации по реконфигурации сети в наиболее узких местах;
- управлять эффективностью цепочки взаимодействия с поставщиками оборудования, собирая и обрабатывая информацию об уже установленном на сети оборудовании, оценивая и учитывая время наработки на отказ того или иного сетевого элемента, анализируя качество конкретной партии и уровень поставщика в целом;
- и многое, многое другое.

Более того, эти примеры не являются какой-то фантастической абстракцией, они не просто «теоретически осуществимы», они уже существуют здесь и сейчас. И при этом они реализуются при помощи «классической» аналитики, решений класса Business Intelligence или DSS (Decision Support System).

Здесь мы снова возвращаемся к вопросу – стоит ли в этом случае переключаться на другую идеологию анализа, если и существующая может удовлетворить ожидания топ-менеджеров? Нужно ли искать «добра от добра»? И имеет ли смысл объединять решения «классической» аналитики с новыми методами, предлагаемыми для обработки и анализа «больших данных»?

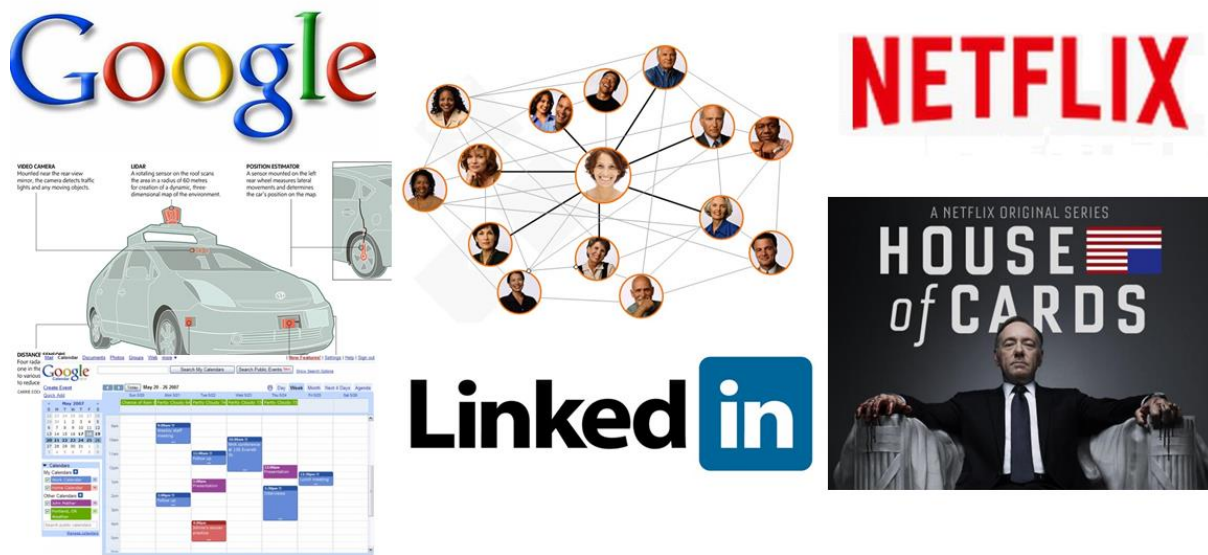
Взгляд в будущее?

Давайте немного пофантазируем. Как мы прекрасно знаем, «большими» данные становятся не только из-за своего размера. В исходной концепции, ключевых характеристик у больших данных было 3 - **volume, variety и velocity** (тот самый размер, вариативность и скорость их добавления в наши хранилища, т.е. появления в нашем распоряжении). Однако в настоящее время, по мере развития представления о том, не только ЧТО такое «большие данные», но и КАК их использовать в реальной жизни, можно расширить исходный набор характеристик ещё как минимум тремя – **value, veracity, vicarious**. А именно их бизнес-ценностью для компании, их достоверностью (нет смысла собирать и хранить «белый шум», надеясь, что в далёком будущем что-то из того, что собрано сейчас, неожиданно приобретёт новый смысл и вышеупомянутую ценность для бизнеса) и их чужеродность, косвенность, по крайней мере относительно тех данных, которые были сгенерированы непосредственно внутри бизнеса.

Представьте себе, какие совершенно новые перспективы откроются перед оператором связи, если он, например:

- получит доступ к информации о всех транзакциях по всем банковским картам своих пользователей? (и тогда можно будет строить более точные поведенческие модели своих абонентов, формировать точечные маркетинговые акции для ключевых сегментов)
- присоединит к своей системе распределения нагрузки модуль анализа новостных лент крупных агрегаторов новостей? (и тогда появится возможность управления сетью с проактивной балансировкой нагрузки как между различными участками сети, так и между различными наложенными услугами на сеть)
- настроит анализ предпочтений по заказываемому контенту, внедрив у себя технологию DPI? (что позволит создавать целевые предложения, сопровождающие вывод на рынок новых продуктов)
- наладит постоянную обработку в режиме реального времени геоинформационных данных, поступающих от абонентского оборудования? (и определит предпочтительные места новых точек взаимодействия с клиентами).

И ведь эти примеры – это, если можно так выразиться, «фантазии о настоящем». Уже сейчас компания Google расширяет записи в календарях своих пользователей информацией о предстоящих авиарейсах, анализируя электронные билеты, полученные ими в свою почту. Уже в наши дни LinkedIn предлагает своим подписчикам возможность находить новые контакты, используя внешние источники информации. Уже 2 года назад компания Netflix заранее просчитала все составляющие успеха до непосредственного запуска в производство своего знаменитого сериала, обработав предпочтения по заказываемому контенту среди всех своих абонентов.



Безусловно, игнорировать происходящие изменения, без ущерба для темпов развития бизнеса – не получится. В работе с данными, в поиске и анализе информации, возможно, как ни в какой другой сфере реализуется принцип, что если вы стоите на месте – вы откатываетесь назад. Однако и бросаться, сломя голову, в омут «больших данных», не представляя – какую именно ценность и значимость они могут принести непосредственно в конкретной ситуации конкретного бизнеса – не менее ошибочно, поскольку гарантированно принесёт только убытки и разочарование в самом подходе. К новым условиям, к новым правилам существования стремиться необходимо, но правильнее всего делать избегая политики «большого скачка», постепенно вводя новые методики и технологии в свою повседневную практику.

Постепенный переход

Можно сказать, что наиболее выгодной стратегией трансформации «классической» аналитики в аналитику «больших данных», выглядит постепенный переход на новый уровень, расширение возможностей существующих в компании BI-инструментов за счёт новых идей, предлагаемых «старшим братом». Например, переход на хранение имеющейся в распоряжении бизнеса накопленной информации из классических СУБД в системы, поддерживающие набор технологий, применяемых для больших данных, сам по себе может принести заметную экономическую выгоду. За счёт увеличения скорости аналитической обработки данных можно реализовать даже на существующих программных BI-решениях обеспечить более гибкий подход к процессу анализа. Если в «классической» BI-системе аналитик действовал последовательно – собирал подходящий набор данных, формировал модель, сравнивал предположения с фактами, предлагал решение, - то теперь количество таких итераций, выполняемых в сопоставимое время, можно значительно увеличить, при этом добиваясь более высокой точности выводов.

При этом ждать чудес заранее – совершенно ошибочно. Практически все ИТ-концепции и подходы, придуманные и воплощённые в жизнь за последние десятилетия, проходили и продолжают проходить общие стадии в своём развитии, которые очень хорошо иллюстрирует

Gartner Hype Cycle: завышенные ожидания – разочарование – последовательное улучшение и выход на уровень реальной продуктивности. Конечно, не все идеи пережили вторую стадию, но те, кто смог пройти через переходный период – по-прежнему с нами. По оценке ИТ-рынка от Gartner за 2014 год, концепция bigdata как раз перешагнула через пик «завышенных ожиданий» («Наконец-то! Ещё немного и у нас будет магическая кнопка «Сделать всё хорошо!», к тому же она ещё и нажиматься будет сама») и передвинулась на вторую стадию. Почему так произошло и насколько радужными остаются перспективы? Давайте разберёмся.

Отрицательная мифология

Наступление стадии разочарования в любой концепции практически неизбежно. Каждое новшество на пути своего развития обрастает различными мифами и легендами. Что же именно ожидалось первоначально от bigdata и почему эти ожидания не оправдываются?

Миф №1. Bigdata = анализ данных. *Можно собрать в общее хранилище любые доступные нам данные, выполнить автоматически над ними какие-то магические пассы и заклинания – и на выходе получить готовый набор логически взаимосвязанных фактов, выводов и предложений по необходимым действиям.*

К сожалению, Bigdata – не философский камень, позволяющий превращать любой мусор в золото. Без изначальной постановки цели, исходного понимания – на какой именно вопрос требуется получить ответ, компания окажется в ситуации, что даже обладая «золотыми» данными, которые будут заложены в bigdata-инструменты, в результате обработки получатся «мусорные» знания, применить которые окажется совершенно невозможно.

Миф №2. Bigdata = лёгкие деньги. *Мы потратим гораздо меньшие средства (например, по сравнению с внедрением технологий хранения и обработки данных, требующихся для классического Business Intelligence) на развёртывание bigdata-решения, применим распространённые open-source-технологии и в результате получим очень быструю отдачу в виде новых продуктов и идей, которые нам подскажет выполненный анализ данных.*

Эта идея тоже оказалась в определённой степени ошибочной. Несмотря на действительно более низкую удельную стоимость владения технологиями bigdata (особенно в области непосредственного хранения данных), в целом практически половина компаний, решившихся на внедрение этих технологий, столкнулась со сложностями в самих принципах измерения уровня возврата инвестиций. Да, в качестве результата можно получить информацию о ранее неизвестных корреляциях в поведении клиентов, увидеть перспективную бизнес-идею, но – эти результаты не приносят финансовой выгоды сами по себе, для их реализации и внедрения будут требоваться отдельные инвестиции. В результате, бизнес приходит к пониманию, что на самом деле – вложения в bigdata представляют собой, за редким исключением, очень длинные деньги, что отпугивает вчерашних адептов.

Миф №3. Bigdata – это Plug & Play. *Самой концепции уже 8 лет, методы и технологии за прошедшее время уже устоялись, рынок предлагает значительный диапазон конкретных решений, мы всегда можем выбрать что-то, что нам точно подойдёт и без осложнений внедрить это у себя.*

Технически – это действительно так. Но, чтобы приносить пользу, инструмент должен не просто быть внедрён – он должен работать. А для работы будут требоваться данные. Для которых нужно описать модель, адаптировав её под выбранное решение. И в среднем четверть компаний (согласно отчётам IDC), использующих у себя bigdata-решения, сталкиваются со сложностью разработки такой модели, а также с трудностями принципиального определения – какие именно данные требуется фиксировать и хранить, а какие можно игнорировать. А кроме самих данных, инструменту нужны пользователи, которыми должны выступать аналитики. По данным того же отчёта IDC, в котором перечисляются причины, тормозящие повсеместное проникновение bigdata-концепции –

примерно 10% компаний испытывают проблемы с подбором соответствующего персонала. А ещё 10% компаний подтверждают, что испытывают трудности в том, чтобы из всего многообразия, существующего на рынке, выбрать оптимальную технологию, наиболее точно подходящую для решения интересующих задач.

Миф №4. Bigdata – тот же BI, только больше. *У нас уже есть продукт класса BI, а поскольку bigdata-решение – это тоже аналитический продукт, значит, у нас не будет никаких сложностей, чтобы его применить с пользой.*

Наверное, это самый устойчивый миф. Есть даже поговорка, что «BI-продукт есть уже почти у всех, а поскольку разработчикам программного обеспечения не хочется терять прибыли, было придумано новое название, чтобы то же самое продать второй раз». Однако в действительности, цели и задачи BI и bigdata – во многом сильно отличаются. BI – в первую очередь, система, направленная на поддержку принятия решений, работающая в чётко очерченных, задаваемых пользователем, рамках, работающая на «ретроспективных» данных и позволяющая найти чёткие доказательства «почему было именно так». Bigdata-концепция нацелена на поиск и предложения новых идей, открытия потенциальных продуктов, основанных на информации, автоматизируя области, которые оказываются чрезмерно ресурсоёмкими для проведения анализа классическими способами, и основанная в своей идеологии, фактически, на «живых», on-line данных. В результате следования последнему мифу, компания может оказаться просто неготовой к получению выгоды от использования bigdata.

Слагаемые успеха

Стратегия

Как ни банально это прозвучит, но для того, чтобы отправиться в успешное плавание по океану больших данных, необходимо чётко представлять конечную точку маршрута, ту цель, достижение которой для компании оправдывает применение bigdata-концепции. Распыляя усилия сразу на все доступные области, собирая и пуская в ход все виды данных, которыми располагает или до которых может дотянуться, обрабатывая внешние источники, компания, тоже можно натолкнуться на неожиданное озарение, которое принесёт прибыль, но это будет случайным исключением. Технологии обработки больших данных уже сейчас дают большой толчок совершенно различным сферам бизнеса, таким как:

- медицина (удалённый и персональный мониторинг, соответствие назначенных методов лечения и фактических результатов и т.д.);
- автомобильная промышленность (сбор и анализ информации с сотен датчиков и электронных систем, устанавливаемых на современных автомобилях, взаимодействие с государственными службами, с распределёнными сетями для обеспечения дополнительной безопасности водителей и т.п.);
- телекоммуникации (анализ причин и перераспределение нагрузки в сетях передачи данных, оценка и контроль лояльности абонентов, формирование таргетированных предложений, основанных на определении интересов конкретного пользователя и т.д.);
- и многих, многих других.

Разумеется, нельзя забывать о том, что внедрение технологий больших данных повлечёт изменения и внутри отдельно взятой организации, перестройку её сложившихся бизнес-процессов:

- для подразделений маркетинга (расширение базы для анализа, появление ранее недоступной информации, которая может помочь объяснить причины изменений в настроениях клиентов и предложить новые способы воздействия на них);

- для продавцов (использование данных со смартфонов и уже упомянутых выше датчиков с автомобилями, которые помогут определить эффективность конкретного продавца);
- для складских и логистических подразделений (комбинирование результатов GPS- и RFID-мониторинга, позволяющее более точно планировать перемещение оборудования, распределение его по складам и точкам продаж);
- отделов по работе с персоналом (определение лояльности сотрудников, на основе анализа данных, получаемых из социальных сетей);
- и так далее.

И конечно, одним из возможных результатов внедрения bigdata-концепции может стать открытие совершенно новых рынков, ранее недоступных для отрасли. Например, телекоммуникационные компании, которые буквально «сидят» на огромных объёмах различных данных, генерируемых самими абонентами, могут использовать получаемую в BSS/OSS-комплексе информацию не только с целью оптимизации внутренних бизнес-процессов, но и для формирования внешних предложений, превращаясь в «фабрику данных». Информация о типах звонков (локальные/международные), о местонахождении (в пределах страны/в роуминге), о частоте и качестве оплаты услуг – формирует вполне законченный портрет абонента, который будет интересен не только самому телеком-оператору, но и внешним компаниям, также взаимодействующим с этим абонентом. И, надо отметить, что в текущем году два мобильных оператора из «большой тройки» реализовали пилотные проекты по сотрудничеству с рядом банков, предлагая им обезличенный скоринговый балл по запрашиваемым абонентам, который в дальнейшем используется при оценке платёжеспособности в процессе одобрения кредитов.

Таким образом, перспективы применения bigdata-концепции по-прежнему остаются поистине безграничными. Важно лишь выбрать наиболее выгодное направление для приложения усилий.

IT-ландшафт

Следующим по важности, после определения направления применения bigdata-технологий, является модернизация IT-ландшафта компании. Концепция больших данных для эффективного своего использования требует наличия сильной IT-инфраструктуры. И здесь речь идёт не о том, чтобы компания просто располагала мощными серверами и сверхпрофессиональными сотрудниками. Мы говорим о том, что успешное внедрение bigdata зависит от качественной архитектуры ИТ. Ни для кого не секрет, что в крупных компаниях, например, телекоммуникационного сектора, занятых в последние годы регулярными процессами слияния и поглощения, накопилось значительное количество дублирующих друг друга информационных систем, а связь между различными сегментами информационного поля может просто отсутствовать в силу разобщённости различных бизнес-направлений. Скажем, абонент, потребляющий услуги и мобильной, и фиксированной связи от одного провайдера услуг, может быть заведён одновременно в несколько различных учётных систем. В этой ситуации разработка непротиворечивой модели данных, которая будет в дальнейшем использоваться и в процессе обработки больших данных, может оказаться не просто сложной – а практически невыполнимой задачей.

Поэтому для того, чтобы bigdata-решения действительно показали свою эффективность, они должны быть размещены в прозрачном информационном поле. Соответственно, выравнивание IT-структуры, создание корпоративной шины данных, выделенных справочных классификаторов, настройка и оптимизация сквозных бизнес-процессов – все эти сопутствующие проекты должны быть предусмотрены и реализованы для того, чтобы работа с большими данными в действительности принесла пользу. Конечно, это увеличивает трудоёмкость перехода на эти технологии, но, в любом случае, здесь есть как минимум одна хорошая новость – даже если в последствии компания в силу каких-либо причин разочаруется

в идеологии bigdata, порядок в ИТ-архитектуре уже будет наведён и положительный эффект от этого будет распространяться на все области деятельности организации.

Люди

И последним, но от этого не менее важным, слагаемым возможного успеха, являются люди. Причём, как те, кто должен будет провести «подготовительную» работу, описанную в предыдущем разделе, организовав и успешно завершив проекты по трансформации ИТ-ландшафта, так и те, кто будет заниматься непосредственно работой с bigdata-инструментами, анализируя данные в поисках новых открытий. И если, возвращаясь к отчёту IDC, компании, внедряющие у себя bigdata-решения, в трети случаев сетуют на то, что крайне сложно найти сотрудников с хорошими навыками управления проектами, то со сложностями в поиске людей, которые будут исполнять роль аналитиков данных, сталкивается уже 2/3 из опрошенных.

Так кто же сможет стать хорошим аналитиком данных? Какими навыками должен обладать такой сотрудник?

Программирование – умение писать код и понимание архитектуры стека bigdata-технологий. По-прежнему довольно редко можно найти источники данных, информацию из которых можно напрямую передавать на анализ. Нужно выполнить значительный объём подготовительных работ для того, чтобы полученные данные принесли пользу.

Научная деятельность – умение формировать решения на основе доказательств, стремление открывать что-то новое. Без таких навыков, сотрудник, занятый обработкой результатов, полученных в ходе применения bigdata-инструментов, может просто «пройти мимо» чего-то действительно ценного.

Опыт ведения переговоров – навык убеждения оппонентов и взаимодействия с окружающими. Чаще всего недостаточно просто найти какую-то свежую идею, самое сложное – убедить остальных в том, что это идея стоящая и действительно способна принести пользу компании.

Аналитика – опыт статистического и визуального анализа, знания в сфере машинного обучения и анализе неструктурированных данных (текст, изображения, видео и т.д.). Инструменты, предлагаемые для работы с большими данными, не являются общеупотребительными, необходимые практические навыки использования таких же или схожих методов анализа.

Экспертиза в бизнесе – знание, как работает бизнес, на котором специализируется компания, понимание того, как компания обеспечивает себе прибыль. Подобный профильный опыт позволит сосредоточиться на анализе информации, которая сможет оказать максимальное влияние на области, наиболее специфичные и важные для конкретного бизнеса, тем самым увеличив отдачу от самого проекта использования больших данных.

Разумеется, вероятность обнаружить аналитика, который полностью бы удовлетворял всем перечисленным характеристикам, внутри собственной компании или на открытом рынке, бесконечно мала. К тому же, в настоящий момент, по-прежнему отсутствуют выделенные направления в высших учебных заведениях, которые бы выпускали всесторонне подготовленных специалистов в области анализа данных. Тем не менее, в ряде Университетов начали уделять этой области более пристальное внимание и вводят в образовательную программу специфические дисциплины. И, конечно, в качестве стандартной практики, многие ведущие разработчики программного обеспечения, предлагающие собственные продукты для работы с большими данными (такие как IBM, Oracle, Google), организуют отдельные курсы, позволяющие слушателям изучить особенности как общей методологии bigdata, так и особенности применения конкретных технологий.

Поэтому в настоящий момент наиболее правильным сценарием будет подбирать команды единомышленников, где все участники смогут дополнять друг друга различными навыками. И здесь самым важным для достижения успеха является правильное управление. Когда

менеджмент компании не просто заинтересован в достижении быстрых результатов, но полностью понимает всю глубину и проблематику bigdata-концепции и готов регулярно и своевременно оказывать помощь в решении различных возникающих организационных задач.

Для чего ещё можно применять?

Доступ к информации о всех транзакциях по всем банковским картам своих пользователей

- ➔ Более точные поведенческие модели абонентов, формировать точечные маркетинговые акции для ключевых сегментов

Модуль анализа новостных лент крупных агрегаторов новостей

- ➔ Возможность управления сетью с проактивной балансировкой нагрузки как между различными участками сети, так и между различными наложенными услугами на сеть

Анализ предпочтений по заказываемому контенту , внедрив у себя технологию DPI

- ➔ Целевые предложения, сопровождающие вывод на рынок новых продуктов

Постоянная real-time обработка геоинформационных данных, поступающих от абонентского оборудования

- ➔ Предпочтительные места новых точек взаимодействия с клиентами



Кластер	Примеры кейсов
Измерение опыта на уровне абонента	Измерение качества сервиса на уровне каждого абонента (скорость интернета, качество голосовых соединений), привязка к геолокации и опросным данным NPS. Прогнозирование NPS для всех абонентов на основе собранных опросных данных. Мониторинг пути клиента по всем точкам контакта и интегрированное измерение опыта. Мониторинг качества сервиса по всем точкам геолокации абонента
Релевантные предложения	Релевантные предложения услуг: для роуминга при прибытии в аэропорт. После обращения в офис SMS о возможностях «личного кабинета» для решения подобного вопроса как более удобная альтернатива
Антиспам	Определение и блокирование спама на уровне каждого абонента.
Антифрод	Распознавание и защита абонента от денежных махинаций третьих сторон и вирусных активностей
Контакт центр	Релевантный IVR; после обращения в колл-центр SMS о возможностях «личного кабинета» для решения подобного вопроса как более удобная альтернатива
Развитие сети	Развитие сети 3G/LTE, где это больше всего требуется: определение областей, где находятся клиенты, пользующиеся мобильным интернетом наиболее часто
Дополнительные сервисы	В реальном времени информирование клиента о близлежащих POI, затруднениях на обычном маршруте, актуальных партнерских сервисах
Розничная сеть	Учитывая геолокацию абонентов перенаправлять их в менее загруженные офисы, учитывать расположение абонентов при планировании открытия новых офисов, планирование развития розничной сети учитывая геолокацию абонентов