

**МИНИСТЕРСТВО ФИНАНСОВ РЕСПУБЛИКИ ТАДЖИКИСТАН
ТАДЖИКСКИЙ ГОСУДАРСТВЕННЫЙ ФИНАНСОВО-
ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ЭКОНОМИКЕ
КАФЕДРА ИНФОРМАЦИОННОГО И МАТЕМАТИЧЕСКОГО
МОДЕЛИРОВАНИЯ**



Co-funded by the
Erasmus+ Programme
of the European Union

Анализ больших данных в финансах

Курс лекций

ДУШАНБЕ - 2021

“Утверждаю”

Декан факультета информационных
технологий в экономике

_____ к.э.н., доцент Неъматов И.У.

“ ____ ” _____ 2021г.

Составители: Гафуров П.Дж., Турахонзода Ш.Н., Сафаров С.А

The Big Data Analytics in Finance teaching materials is developed in the framework of ERASMUS+ CBHE project “Digitalization of economic as an element of sustainable development of Ukraine and Tajikistan” / DigEco 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP

This project has been funded with support from the European Commission. This document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Душанбе – 2021

ОГЛАВЛЕНИЕ

ЛЕКЦИЯ 1 БОЛЬШИЕ ДАННЫЕ В ФИНАНСАХ	4
ЛЕКЦИЯ 2. ПРОИСХОЖДЕНИЕ КОНЦЕПЦИИ БОЛЬШИХ ДАННЫХ И ИХ СОДЕРЖАНИЯ	8
ЛЕКЦИЯ 3. ФИНАНСОВЫЕ ТЕХНОЛОГИИ В ЦИФРОВОЙ ЭКОНОМИКЕ	13
ЛЕКЦИЯ 4 РОЛЬ БОЛЬШИХ ДАННЫХ В ПРИНЯТИИ ФИНАНСОВЫХ РЕШЕНИЙ.....	23
ЛЕКЦИЯ 5. СИСТЕМАИ HADOOP MAPREDUCE	69
ЛЕКЦИЯ 5 ИНТЕГРАЦИЯ ДАННЫХ И БЕЗОПАСНОСТЬ	73
ЛЕКЦИЯ 6. БАЗА ДАННЫХ NOSQL И БАЗА ДАННЫХ SQL ДЛЯ РЕШЕНИЯ ФИНАНСОВЫХ ПРОБЛЕМ	78
ЛЕКЦИЯ 7. АРХИТЕКТУРЫ СИСТЕМЫ ОБРАБОТКИ ДАННЫХ	83
ЛЕКЦИЯ 8. ПЛАТФОРМА БОЛЬШИХ ДАННЫХ ПРОТЕСТИРОВАЛА НОВУЮ ТЕХНОЛОГИЮ ЗАЩИТЫ BIG DATA.....	86
ЛЕКЦИЯ 9. ЖИЗНЕННЫЙ ЦИКЛ АНАЛИЗА БОЛЬШИХ ДАННЫХ	88
ЛЕКЦИЯ 10.БОЛЬШИЕ ДАННЫЕ КАК РЫНОК.....	99
ЛЕКЦИЯ 11. АНАЛИТИКА БОЛЬШИХ ДАННЫХ	125
ЛЕКЦИЯ 12. DATA MINING: ТЕХНОЛОГИИ И МЕТОДЫ АНАЛИЗА ДАННЫХ.....	132

ЛЕКЦИЯ 1 БОЛЬШИЕ ДАННЫЕ В ФИНАНСАХ

Феномен «большие данные» изменил структуру финансового сектора, вызвал появление новых свойств и раздвинул границы финансов. Большие данные вместе с цифрацией выступили движущей силой развития отрасли. В свою очередь, новые и повышенные требования финансов предопределили как разработку и внедрение целого ряда новых технологий учета данных, так и быстрое распространение алгоритмических решений и программных продуктов для выработки практически молниеносных решений в управлении финансово-торговыми процессами. Статистические и аналитические обзоры больших данных и уровень развития их технологий представлены в многочисленных источниках.

Так, по данным International Data Corporation, специализирующейся на рынке ИТ-технологий, цифровая вселенная удваивается каждые два года (Digitizing Intelligence, 2016, p. 3), а за последние пару лет было создано более 90% всех известных данных (Investopedia). К 2025 г. в США большие данные вместе с цифрацией будут ответственны за создание ВВП на сумму до 2,2 трлн долл. Экспоненциальный рост больших данных оказал преобразующий эффект на финансовые услуги и всю финансовую отрасль. По данным McKinsey, финансовая отрасль по масштабам развития технологий больших данных и цифрации информации находится на втором месте после информационных и коммуникационных технологий (Slideshare).

Именно большие данные и соответствующие им коммуникативные, вычислительные и программные возможности и приемы (в том числе, облачное хранение данных и децентрализованный доступ к ним, алгоритмическая торговля и блокчейновая технология) нарушили привычную модель развития и подорвали тенденции к росту масштабов финансового сектора. Новации, во многом вызванные приходом больших данных, позволили перейти к рассредоточению финансовых участников и демонополизации рынка. В сторону большей дисперсии финансовых услуг и

их инициаторов срабатывали, получившие распространение на базе организации работы с большими данными, прямые расчетно-кредитные операции без финансовых посредников, а также многочисленные финансовые стартапы, которые подрывали монополию банков и традиционных расчетно-платежных систем.

В том же направлении действовали персонализация клиента, поставка услуг и продуктов под заказ, индивидуализация обслуживания. Возможность перехода к таким формам обслуживания возникла только благодаря организации работы с возросшими объемами данных. Большие данные оказали и продолжают оказывать существенное воздействие на бизнес и быт. Наряду с общими проблемами и процессами, финансовая сфера имеет и отличительные особенности взаимодействия с данными.

В основе работы финансовых учреждений находятся стоимостные активы-данные (приход, оборот и поставка ликвидных ресурсов). Они выступают своеобразным сырьем и готовым продуктом финансового «производства». Привычная стоимостная форма финансовых данных дополняется нетрадиционными данными, которыми внутренне не присущи стоимостные свойства и денежные определенности.

Переход на большие данные существенно раздвинул границы применения последних. Потребовалась оценка всех обращающихся в финансовых учреждениях данных и их воздействия в целом и по отдельным группам на стоимостные показатели и, прежде всего, на издержки и доход финансового учреждения. Данная задача входит в микроэкономический анализ влияния больших данных на финансы. Большие данные оказали существенное воздействие на продуктовую линейку и услуги финансовых учреждений.

С одной стороны, они способствовали расширению предложения и повлияли на изменение его формы (в основном в направлении цифрации взаимодействий с клиентами), а с другой стороны, «впустили» в некогда

элитную и недоступную для посторонних сферу финансового посредничества новых участников. В целом большие данные существенно изменили общий финансовый ландшафт, а также условия стабильности и устойчивости финансовой системы.

В частности, поменяли отношения к риску, случайности и неопределенности. Кроме того, под их воздействие возникли новые риски, связанные с оборотом больших данных (прежде всего, вытекающие из обезличенности, цифрации, масштабности и разнородности информации, а также ослабления человеческого воздействия на ее движение); появились также новые участники финансовых взаимодействий (финансовые стартапы, электронные вендеры, коммуникационные системы и ИТ-организации). В результате не только существенно возросла конкуренция в отрасли, и возникли ее новые формы и виды, но и появились принципиально новые случайности и неопределенности.

В свою очередь, опора на большие данные позволила нейтрализовать и участь значительную их часть – абсорбировать через новые финансовые инструменты (сложные производные и многозвеньевые арбитражные сделки при высокоскоростной биржевой торговли). Однако постоянно возникают риски и неопределенности (прежде всего, найтианского типа), которые невозможно предвидеть и учесть. Тем самым повысился уровень неопределенности системы в целом. В результате большие данные оказали важное воздействие на макроэкономическое состояние финансов и определенность финансового рынка в системе координат между хаосом и порядком.

Отмечая большой круг проблем, связанных с переходом финансов на большие данные, мы осознаем невозможность их детализации в рамках одной статьи: некоторые из них только затронуты и нуждаются в тщательной проработке, но позиционирование их в общем контексте позволило дать характеристику важных преобразований в финансовой сфере и очертить круг

вопросов, который предстоит решать финансовой науке. Главной задачей данной статьи является определение основных направлений воздействия больших данных на финансовые учреждения и формирования под их воздействием новой эко-финансовой системы

ЛЕКЦИЯ 2. ПРОИСХОЖДЕНИЕ КОНЦЕПЦИИ БОЛЬШИХ ДАННЫХ И ИХ СОДЕРЖАНИЯ

Что такое большие данные

Если обобщить, то биг дата — это большой объем информации, который компания собирает и хранит для последующего использования. Еще когда говорят, что компания использует большие данные, часто имеют в виду не сами данные, а технологии для их обработки.

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.

Какие данные можно считать большими

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, конечно же, неправильно.

Однозначно отделить формат больших данных от обычных помогут три критерия.

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.

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 гигабайт, 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.

Данные должны быть разнородными и слабо структурированными. 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

Ответ на вопрос «Что такое big data в IT?» не так прост. Это не только сами данные, но и принципы работы с ними. Кстати, иногда эти принципы применяют и к анализу «маленьких» данных — например, можно построить модель на основе однородной информации или совсем небольшой клиентской базы.

Зачем нужна big data

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:

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.

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.

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, «Газпром нефть» [смоделировала](#) ситуацию аварийного отключения электричества, чтобы понять, почему возникает сбой автоматического перезапуска оборудования. Модель помогла обнаружить неожиданные причинно-следственные связи и устранить проблемы.

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.

Больше интересных кейсов использования big data читайте в статье [«Зачем вам большие данные: примеры использования big data в 8 отраслях»](#).

ЛЕКЦИЯ 3. ФИНАНСОВЫЕ ТЕХНОЛОГИИ В ЦИФРОВОЙ ЭКОНОМИКЕ

Понятие «финансовые технологии», или «финтех», является относительно новым. Базельский комитет по банковскому надзору под финтехом понимает «порожденные технологиями финансовые инновации, которые могут привести к созданию новых бизнес-моделей, приложений, процессов или продуктов, которые впоследствии скажутся на финансовых рынках, институтах или производстве финансовых услуг».

В качестве примеров финансовых технологий в исследовании Базельского комитета приводятся: краудфандинговые сервисы, площадки по взаимному кредитованию, онлайн-банкинг, цифровые валюты, мобильные кошельки, форекс, цифровые платформы по обмену данными, высокочастотная торговля, электронная торговля, робоэдвайзеры и пр. Понятие «финтех» также используется в отношении компаний, как правило, являющихся стартапами, которые активно используют инновационные, прорывные технологии в предоставлении финансовых услуг в условиях конкуренции с традиционными институтами.

С другой стороны, традиционные банки, страховые и управляющие компании также активно внедряют новые технологии и с этой точки зрения тоже являются участниками рынка финтеха. Бурному развитию финтеха способствовали несколько трендов. Изменения в потребительском поведении, проявляющиеся в растущем проникновении мобильных телефонов, желании делиться опытом с широким кругом людей, повышении требований к удобству пользования услугами, качеству информации и скорости ее получения.

2. Рост популярности социальных сетей, позволивший предложить новые виды финансовых услуг, основанные на обмене информацией между пользователями, например, краудфандинг, пиринговые переводы и т. д.

3. Развитие технологий обработки данных, которое привело как к появлению принципиально новых, так и к значительному улучшению существующих услуг, таких, как пиринговое кредитование, онлайн-скоринг, основанный на технологии больших данных, алгоритмическая торговля и др.

4. Повышение конкуренции на рынке. После глобального финансового кризиса 2007–2008 гг. произошло ужесточение регулирования банковского сектора, в связи с чем возник интерес потребителей к услугам, предлагаемым финтех-стартапами. Инновационные компании стали успешно конкурировать с банками, стимулируя последние к разработке инновационных продуктов.

5. Снижение стоимости финансовых услуг. Внедрение ИКТ-технологий позволило снизить издержки на обслуживание клиентов благодаря отказу от широкой сети физических офисов и переходу на электронное взаимодействие и с потребителем, и с регулятором.

6. Повышение доступности финансовых услуг за счет внедрения удаленных механизмов обслуживания и снижения порога входа на рынок.

7. Повышение прозрачности экономики и эффективности мер борьбы с отмыванием денег и финансированием терроризма. Глубокий анализ операций, сбор сведений об активности клиентов позволяют более предметно и точно противодействовать незаконной деятельности.

8. Быстрый рост использования финтех-услуг на базе смартфонов и электронных кошельков, произошедший прежде всего в Китае и Индии, где наблюдался низкий уровень проникновения традиционных банковских услуг. Происходит рост числа потребителей финтех-услуг за счет аудитории, не охваченной традиционными банковскими услугами.

9. Рост численности поколения Z. Согласно исследованиям, проведенным Facebook и MasterCard, более 90 % молодых людей в США не

доверяет традиционной банковской системе, все чаще прибегая к новым финтех-сервисам.

10. Стремительный рост инвестиционных потоков в сферу финтеха: если в 2012 г. объем финансирования финтех-стартапов составлял 8,9 млрд долл., то в 2016 г. инвестиции составили 64,0 млрд долл. В 2018 г. вложения в этот сегмент составили рекордные 120,2 млрд долл.

Влияние финансовых технологий на развитие банковской сферы

Развитие и распространение на рынке финансовых технологий создает для банков как новые возможности, так и новые риски. Банки, которые эффективно формируют партнерские отношения или поглощают перспективные финтех-компании, смогут повысить свою конкурентоспособность, защитив и даже улучшив свои рыночные позиции. Как показывает опыт, внедрение финансовых технологий в ключевые процессы традиционного банка, будь то продажи новых продуктов или сервисное обслуживание в отделениях, позволяет сократить их стоимость на 40–60 %. Улучшается обслуживание клиентов банка за счет сокращения времени, необходимого для получения продукта или услуги, количества документов и контактов клиента с банком, результатом чего результатом становится повышение лояльности клиентов и расширение клиентской базы. Так, согласно исследованию McKinsey, проведенному во Франции, розничные банки, активно развивающие дистанционные каналы обслуживания, смогли опередить традиционные кредитные учреждения по индексу потребительской лояльности в среднем на 15–60 пп.

Крупнейшие банки щедро инвестируют в цифровые технологии и извлекают выгоду из их применения благодаря эффекту масштаба. Лидеры рынка цифрового банкинга уже применяют новые подходы к организации больших массивов данных на базе единой платформы, а использование

методов углубленного анализа больших объемов данных позволяет крупнейшим банкам повышать точность кредитного скоринга, формировать индивидуальные предложения, адресованные клиентам, и эффективно распределять ресурсы.

Кроме того, крупные банки стремительно меняют формат своих отделений, сокращают их количество, оснащая оборудованием, необходимым для самостоятельного осуществления клиентами большинства сервисных операций, и сосредоточивая усилия оставшегося персонала на консультировании и продажах. Распространение цифровых технологий в финансовом секторе также связано с определенными рисками.

Согласно результатам опроса PwC среди более 500 компаний в 48 странах мира, 83 % респондентов, представляющих традиционные организации сектора финансовых услуг, полагают, что им грозит потеря части бизнеса. В случае с банками показатель оказывается еще более внушительным – 95 %.

Финтех-компании имеют шансы превратиться в агрессивных конкурентов классическим банкам и даже потеснить их позиции на рынке. Согласно прогнозам PwC, к 2020 г. финансовые технологии будут охватывать 24–28 % рынка банковского обслуживания и платежей и до 22 % рынка страхования, управления активами и управления частным капиталом. По данным Citigroup, в ближайшие 10 лет около 800 тыс. работников сферы банковских услуг потеряют работу из-за введения новых технологий.

Кроме банковских работников пострадает и сфера коммерческой недвижимости, когда банки станут повсеместно закрывать свои отделения в разных городах. По мнению McKinsey, самыми уязвимыми являются традиционные депозитарные и кредитные услуги: к 2025 г. банки могут потерять от 40 до 60% доходов от этой сферы. Отдельно выделяется сфера мобильных платежей, до 35 % которой могут забрать себе такие компании, как

Apple и Google, благодаря своим платежным сервисам, позволяющим легко и быстро оплачивать покупки в магазинах.

Перспективы развития банковского сектора в условиях внедрения современных финансовых технологий

Согласно исследованию McKinsey, в условиях дальнейшего развития финтех-сектора и распространения цифровых технологий у традиционных банков есть несколько возможных путей развития. Наиболее универсальный вариант – превращение традиционных банков в цифровые, которые предоставляют широкий спектр финансовых продуктов и услуг. Европейские банки чаще оставляют за старым банком традиционный бизнес, а для розницы создают «дочку» – цифровой банк. Другой вариант – переход от классического формата банка к финансовой экосистеме «банк-партнеры», подразумевающий выстраивание партнерских отношений с другими компаниями. Услуги, предоставляемые партнерами, должны отвечать широкому кругу потребностей клиента, что позволит владельцу экосистемы обслуживать их по принципу одного окна. Третий вариант: банки могут сосредоточиться на предоставлении базовых услуг, таких как управление бухгалтерским балансом и проведение транзакций. Это направление не принесет высокой прибыли, но может стать привлекательным благодаря меньшему уровню риска и эффекту масштаба.

Цифровые банки предоставляют широкий спектр финансовых продуктов и услуг, не имеют фронт-офисов, а для оказания услуг используют мобильные приложения и сайты. Нередко их называют онлайн-банками или директ-банками. Список банковских услуг, предлагаемых цифровыми банками, схож с услугами финтех-компаний: операции по счетам, выдача кредитов, инвестиции, работа с депозитами. Однако цифровые банки начинают использовать и новые форматы, которые зависят от экономического

и технологического развития страны, принятой законодательной базы и развитости банковской сферы. Массачусетский технологический институт описывает три волны инноваций в цифровом банкинге: фундаменталисты, цифровые гибриды и полностью цифровые банки. Фундаменталисты появились в 1970-е гг. и лишь имитировали оцифровку своих технологических процессов, ограничиваясь красивым веб-сайтом, СМС-сообщениями на телефон клиента и личным электронным кабинетом. Но основные финансовые процессы шли вполне традиционно.

Цифровые гибриды появились в 1996 г. Они используют специальную ИКТ-инфраструктуру, в которой присутствует электронный интерфейс, оптимизирующий все процессы, однако он опирается по-прежнему на централизованные хранилища информации с уязвимыми протоколами передачи данных и классический банковский бэк-офис. Полностью цифровые банки используют современные ИКТ и тесно интегрируются с жизнью современных людей, постоянно использующих мобильные устройства в интернете. Специалисты МТИ считают, что цифровой банк – это банк, обладающий современным набором возможностей

Большинство обзоров и исследований, посвященных большим данным, начинаются с концепции 3Vs. В каждой сфере данная концепция имеет определенные особенности проявления, на которые редко обращают внимание. Тем не менее именно они во многом определяют направления развития, возможности и пределы использования больших данных, а также дальнейшие перспективы той или иной отрасли.

Финансы, если не считать непосредственно сферы информационных технологий, оказались именно той отраслью, в которой большие данные смогли максимально реализовать свои функции, и для этого были свои причины. В свою очередь финансы стали крупным заказчиком новых технологий и благодаря требованиям отрасли во многом происходит

дальнейшее их развитие. Финансовые услуги связаны с аналитической обработкой большого объема данных и организации работы с неопределенностями (Ключников, Молчанова, Ключников, 2017). От правильной организации работы с большими данными во многом зависит успехи как индивидуальных, так и, тем более, институциональных инвесторов. Для этого привлекаются самые современные технологии работы с данными, а также новые информационные источники. Переход на алгоритмическую торговлю (у североамериканских бирж до 40% биржевого оборота составляет торговля акциями, 30% – облигациями (Osipovich, 2017) и 24% – валютой (Cornerly, 2014)) вызвал необходимость составления больших ценовых рядов, что потребовало моделирования и математической обработки огромных массивов информации. Во всех случаях необходимо проводить оценку данных с точки зрения их качества, учитывать их свойства, а также определенным образом их классифицировать и характеризовать.

Для этих целей вполне подходит концепция 3Vs с последующими доработками и дополнениями

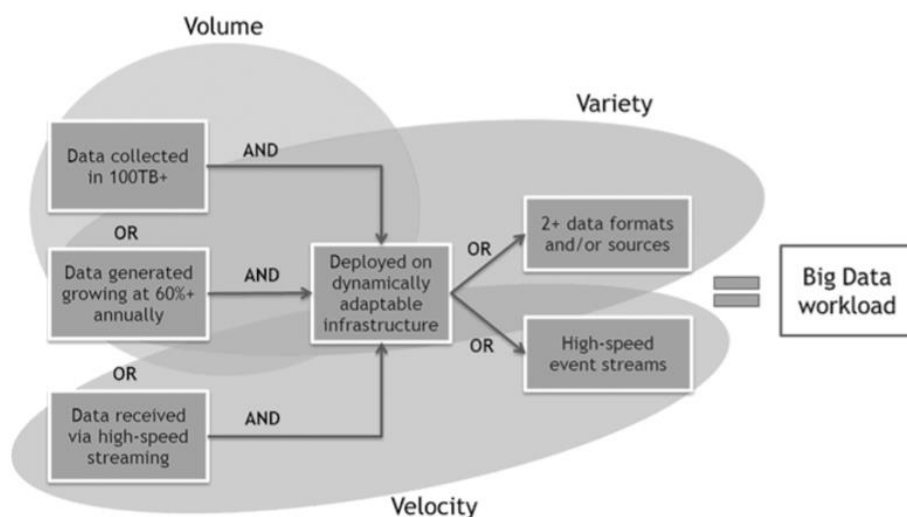


Рисунок. Критерии оценки больших данных с позиции концепции 3Vs (IDC, 2017) Концепция 3Vs (объем, разнообразие и скорость) больших данных характеризует три определяющих их свойства, а также служит для измерений

данных. Объем предназначается для количественной характеристики данных, разнообразие – характеристика числа типов данных, скорость относится к скорости их обработки.

Согласно модели 3Vs, происходит расширение всех трех свойств, а не только объема данных, включенных в систему управления. Данную модель предложил в 2001 г. Дуг Лэни (аналитик исследовательского центра Gartner Inc., специализирующегося на рынках информационных технологий). 3Vs позволяет проводить трехмерное измерение и управление данными: управление объемом (volume), разнообразием (variety) и скоростью (velocity) (рисунок). Данная модель концепция нашла применения в разнообразных сферах. Недавно данная концепция была расширена за счет включения двух дополнительных показателей Vs – изменчивости (variability), посредством которой учитывается увеличение диапазона значений (что типично для большого набора данных), и значения (value), которое оценивает состояние непосредственно финансовых данных (нефинансовых и финансовых учреждений).

Таким образом, трехмерное измерение – 3Vs, еще пару лет назад широко применяемое для характеристики больших данных в любой сфере, преобразовалось в пятимерное – 5Vs, отражающее специфику финансовой сферы второй половины текущего десятилетия. В 2017 г. IBM совместно с Hortonworks (Hortonworks) в качестве четвертого элемента Vs был добавлен показатель – veracity, который характеризует как достоверность, так и точность данных (Big Data). Посредством такого свойства в общем потоке неопределенности достигается (и при параметризации оценивается) относительная определенность.

Благодаря соответствующей аналитической и вычислительной обработки, в информационном хаосе выделяются достоверные данные, наиболее точно отражающие реальность. Данное свойство имеет

универсальное значение. Однако наиболее ценно оно в финансах. В основе всех финансовых коммуникаций лежит доверие, и когда происходят нарушения в доверии, то возрастает неустойчивость финансовой системы, возникают неплатежи, банкротства, дефолты, кризисы. Четырех- и пятимерная характеристики больших данных не включают два очень важных критерия, которые необходимы для финансового дела – это время (Ключников, Ключников, 2017) и место (Ключников, 2016). Они характеризуют соответствующие свойства и параметры финансовых данных.

С учетом времени не только составляются и рассчитываются различные временные ряды, а также выделяются финансовые циклы, но и наблюдаются временные отличия в генерации данных, устанавливаются периоды всплесков и падений, характер временных изменений различных свойств. В целом, время играет важную роль для финансовой статистики (Елисеева, Соколов, 2005). Но еще более важным является учет времени для торговли.

С этой целью выделяются критерии, которые позволяют наблюдать за торговым процессом в режиме реального времени (особые механизмы отбора и передачи информации, алгоритмы, вычислительные процедуры и программные продукты). Так, базы данных NewSQL обеспечивают масштабирование NoSQL и фокусируются только на транзакциях; они также вместе с Clustrix DB (при использовании Massively Parallel Processing, MPP) обеспечивают быструю аналитику в реальном времени, которая становится все более важной для текущей деятельности и биржевой торговли финансовыми продуктами. При этом происходит совмещение текущих данных с выводом из хранилищ с учетом времени их поступления, что необходимо для сравнений и других решения других задач.

Собственно на такой базе строятся, например, современные платформы биржевой высокочастотной торговли (ВЧТ). Поскольку интегрирование и агрегирование относятся к обязательным условиям обработки аналитики в

реальном времени, с них и начинается срочный анализ данных. Навигационные способности Clustrix DB обеспечивают масштабирование не только во времени, но и в пространстве в одном кластере, что позволяет, с одной стороны, всем узлам участвовать во всех типах запросов, с другой стороны, по мнению некоторых экспертов (Blog), ускоряет аналитику, особенно когда дело касается сложных аналитических запросов.

Последнее свойство позволяет выделять, анализировать и использовать для текущих сравнений территориальные особенности поступления данных в режиме реального времени. Такая особенность

ЛЕКЦИЯ 4 РОЛЬ БОЛЬШИХ ДАННЫХ В ПРИНЯТИИ ФИНАНСОВЫХ РЕШЕНИЙ

Сегодня одним из ключевых направлений для корпоративных лидеров становится работа с данными. Практически девять из десяти (87 %) финансовых специалистов, опрошенных в ходе настоящего исследования, полагают, что именно данные обладают потенциалом изменения самого подхода к ведению бизнеса. Другие исследования показали, что чуть больше половины (51 %) корпоративных лидеров относят большие данные и аналитику к первой десятке корпоративных приоритетов. Все это говорит о том, что после революции последних десятилетий бизнес вступает в эпоху, когда данные становятся движущей силой.

Объем и разнообразие данных, доступных для анализа, растут в геометрической прогрессии. Между тем появляются все более мощные технологии, позволяющие управлять усложненными данными и выполнять аналитику более высокого уровня. Источников данных сегодня великое множество — это и записи кол-центров, и внешние каналы, и машинные данные, и социальные сети, и многое другое.

Эти связанные между собой потоки (обозначаемые, как правило, общим термином «большие данные») в современных организациях комбинируются в целях открытия новых источников уникальной информации и прибыли.

Чтобы воспользоваться результатами, которые дает работа с данными, компаниям придется решить целый ряд задач, поскольку неизбежно возникает потребность в новых навыках, инструментах и ином образе мышления.

Так что же несет с собой эпоха господства данных для лидеров бизнеса и, в частности, для финансистов? Какие выгоды дадут новые подходы к анализу данных? И какова здесь роль специалистов по управленческому учету? Это лишь некоторые из основных вопросов, которые освещаются в настоящем отчете. Выводы сделаны по результатам всестороннего анкетирования, опроса

более 2000 финансовых директоров и сотрудников финансовых служб, а также на основе комплексного изучения документарных материалов.

Данные становятся главным бизнес-активом.

Компании всех размеров и направлений деятельности уже обращают данные в реальную ценность многими доступными способами. Охвачены все отрасли: это и социальные сети, и интернет-компании, и фирмы, производящие продукты питания и напитки, и сети фастфуда, и банки, и промышленники, и отели, и многие другие. Научные исследования выявляют методы роста производительности за счет принятия решений, в основе которых лежат данные.

С этим согласны опрошенные финансисты, которые, опираясь на данные, ожидают получить более высокую эффективность и повышение КПЭ.

В попытках заставить данные работать на коммерческую выгоду компании сталкиваются с необходимостью роста квалификации.

В большинстве компаний всемерная адаптация к эре власти данных в бизнесе остается текущей задачей. Среди опрошенных финансистов 86 % выразили мнение, что организации испытывают трудности в стремлении делать ценные аналитические выводы из имеющихся данных, и не в последнюю очередь это связано с такими проблемами, как, например, разрозненность данных, их качество или способность работать с незнакомыми данными нефинансового характера. Чтобы помочь организациям справляться с трудностями, индустрия программного обеспечения продолжает заниматься разработкой инструментов и приложений, которые призваны облегчить процесс извлечения ценных аналитических выводов из данных.

Однако это по-прежнему остается сложной задачей, поскольку, прежде чем выбрать то или иное техническое решение, организация должна определиться с тем, каким образом она предполагает использовать данные для улучшения своей работы.

Специалисты финансовой сферы имеют все возможности для того, чтобы помочь обернуть данные в коммерческие идеи и ценности.

Более 90 % респондентов считают, что финансисты играют существенную роль в организациях, когда речь идет о получении выгод от проектов, построенных на работе с данными. Но с конкретизацией этой роли чаще всего возникают затруднения. Передовые методы анализа данных часто требуют квалификации уровня ученой степени (например, кандидата наук) в области обработки и анализа данных, что является редкостью для финансовых работников.

Но в чем действительно они могут способствовать повышению ценности — это в применении своих знаний на стыке финансов и хозяйственной деятельности, чтобы перевести аналитические выводы в коммерческое русло, используя, например, такие инструменты, как планирование, бюджетирование, прогнозирование и управление производственными показателями. «Наша работа не в том, чтобы спуститься до самого нижнего уровня данных, а в том, чтобы знать, как агрегировать результаты для последующего получения на их основе содержательного, ценного и глубокого отчета», — утверждает Джеймс Милн, старший финансовый директор финансовой службы глобальных операций Yahoo!

Задача развития потенциала данных придает как никогда высокую степень важности финансово-деловому партнерству внутри организации.

Чтобы по-настоящему раскрыть возможности больших данных, специалистам по управленческому учету необходимо более тесно сотрудничать с тремя группами ключевых заинтересованных сторон: коллегами из сферы ИТ, которые собирают большую часть данных; учеными в области обработки данных, которые способны оперировать передовыми методами анализа данных; и, наконец, бизнес-лидерами, которые могут обеспечить трансформацию новых идей в конкретные действия. Это потребует от финансистов широкого спектра управленческих навыков: открытой

коммуникации, умения вести за собой и оказывать влияние, стратегического понимания бизнеса — всего, что имеет большое значение для роли бизнес-партнера, которую, в понимании многих организаций, должны играть финансовые специалисты.

Достижение нового понимания данных для малых и средних компаний так же ценно и перспективно, как и для крупных.

Преимущества при выстраивании бизнеса, более ориентированного на обработку данных, получит любая компания, вне зависимости от ее размера. Небольшие предприятия имеют меньше ресурсов для реализации соответствующих проектов, но и зависимость от устаревших ИТ-методов у них ниже, чем у их более крупных конкурентов.

Кроме того, мелким легче двигаться вперед: они могут подключать простые в использовании приложения и услуги передачи данных, число которых день ото дня растет и которые позволяют извлекать выгоду из данных при минимальных затратах и рисках. А ключ к успеху как для малых, так и для крупных компаний — в изначальной выработке четких целей для проектов данных. Только после этого следует подбирать инструменты и сервисы, наиболее подходящие под конкретные цели.

Работая с данными, финансисты получают новые возможности карьерного роста.

Эра всевластия данных предоставляет специалистам по управленческому учету прекрасную возможность развития навыков и получения новых знаний. И действительно, 85 % респондентов считают, что повышение собственной квалификации для работы с большими данными делает более вероятными карьерный рост и востребованность на рынке труда.

Некоторые специалисты, несомненно, будут искать пути улучшения своих аналитических навыков, но именно способность играть центральную роль при трансформации потенциала данных в реальные экономические

показатели станет импульсом для тех, кто стремится занять в организации стратегические позиции.

Представленные задачи достаточно серьезны, и для их решения финансисты должны обрести профессиональные навыки в ранее незнакомых для себя областях. Впрочем, специалисты по управленческому учету однажды уже доказали свою способность адаптироваться к меняющимся требованиям, в частности развив в себе новые способности и взяв на себя новые роли, когда этого потребовала постиндустриальная эпоха. Сегодня от специалистов по управленческому учету ожидают подобного эволюционного шага, освоения новых навыков и нового мышления. Это выдвинет их на передовые позиции в деле обращения данных на пользу коммерческих выгод.

Более 90 % сертифицированных CGMA специалистов по управленческому учету подчеркивают существенную роль финансистов на пути бизнеса к получению выгод через использование данных.

Пять качеств финансовых руководителей, ориентированных на работу с данными

Руководители финансовых служб, ориентированных на обработку данных, имеют, как правило, некоторые общие черты, а именно:

Способность определять, какие точки данных полезны для понимания того, что является движущим фактором для бизнеса. Финансовые руководители, ориентированные на работу с данными, имеют понимание фундаментальных движущих факторов и количественных показателей организации, и применяют эти знания как базовый элемент при оценке новых, менее структурированных источников данных.

Ясное представление о том, что больше всего волнует заказчиков и как это отслеживать. Также они имеют четкое понимание, почему заказчики выбирают их продукты и услуги, как завоевать таких заказчиков и что помогает их удерживать.

Способность воспринимать новые формы данных и проявлять креативность, включая их в процесс принятия бизнес-решений. Помимо чисто финансовых и корпоративных данных финансовый руководитель, ориентированный на работу с данными, должен быть восприимчив к их новым видам и источникам, которые могут быть неожиданными и нетрадиционными как для него, так и для его команды.

Умение спокойно воспринимать неопределенность, в том числе осознавать, что большие данные могут и не привести к окончательному ответу. Знакомясь с новыми и, в некоторых случаях, мало проверенными источниками данных, компании начинают применять аналитические методы, которые могут и не обеспечить определенность, ожидаемую финансистами. Финансовые руководители должны принимать и внедрять новые методы и стратегии, которые способствуют совершенствованию процесса принятия решений в мире, преисполненном данными.

Умение исследовать новые способы интерпретации данных для более качественного информирования руководства. Финансовые руководители, ориентированные на работу с данными, также отлично и быстро распознают нужные результаты, часто находя новые способы визуальной передачи наиболее важных моментов и обеспечения уверенного коммерческого восприятия и сильного импульса.

Эпоха бизнеса, движимого данными

Правила бизнеса продолжают меняться под влиянием технологий. Сделав ставку на компьютеризацию рабочих мест, которая началась в конце 1980-х годов и продолжается по сей день, компании в настоящий момент пришли к эпохе господства цифровых технологий в бизнесе. Повсеместное внедрение систем планирования ресурсов предприятия, электронных точек продаж, электронной коммерции и прочих интернет-систем привело к тому, что все больше организационных данных собирается и обрабатывается в цифровом виде.

Начиная с нулевых годов нашего столетия, бизнес получил возможность использовать в своих целях новые объемы данных, которые по сей день демонстрируют экспоненциальный прирост. Здесь и развитие социальных сетей, и распространение все более вездесущих мобильных устройств, облачных сервисов, а также применение огромного количества датчиков, которые привязывают физические объекты к Интернету вещей.

Эти данные представляют собой смесь структурированных внутренних и, как правило, неструктурированных внешних данных, состоящих из голосовых записей, информации о ценах, изображений, сообщений в социальных сетях, сведений о географическом положении и многих других. Как показывает агентство Economist Intelligence Unit, в настоящее время по меньшей мере шесть из десяти компаний используют в своих целях данные контакт-центров, в том числе записи разговоров, сообщения, которыми обменивается персонал, синдицированные данные третьих лиц, например о погоде, или рыночная информация, а также данные госорганов. Чуть более половины таких компаний собирают еще данные, генерируемые компьютерными средствами. Так называемые большие данные дополняют бухгалтерскую и коммерческую информацию, которую хранит и обрабатывает компания.

Хотя исследователи данных и ИТ-специалисты дают свои определения термина «большие данные», в настоящем отчете он используется для краткого обозначения массивного увеличения объема данных, используемых сегодня в целях аккумуляции новых идей и понимания эффективности деятельности, возможностей и рисков (см. «Определяемые большие данные»).

Когда начинается обсуждение последних тенденций в сфере больших данных, внимание в основном уделяют новым формам неструктурированных данных. Специалистам по управленческому учету, безусловно, необходимо знать о том, как развивается ситуация в этой области. При этом нужно понимать, что большинство современных организаций стоят буквально у непочатого края работы со своими внутренними структурированными

данными, которые хранятся в системах управления предприятиями. Несмотря на значимость обеих областей, специалистам по управленческому учету, скорее, имеет смысл взяться в первую очередь за корпоративные данные.

По сообщению McKinsey & Company, прирост объема больших данных вызовет новую волну «инноваций, конкуренции и производительности» в деловой среде⁴. Как выразились два видных деятеля науки, «использование больших данных позволяет принимать решения на основании не интуиции, а фактических данных. По этой причине они способны совершить революцию в сфере менеджмента»⁵.

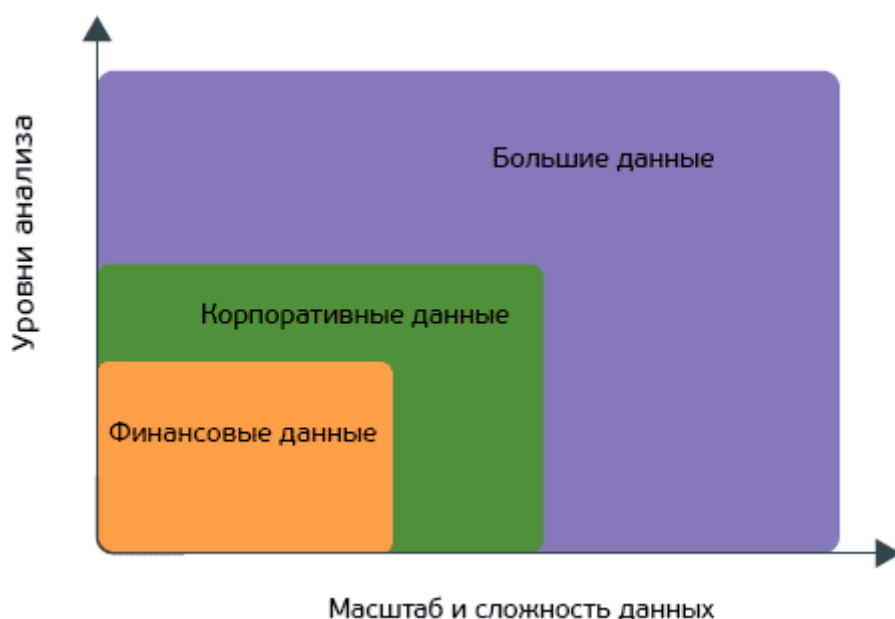


Рис. 1. Определяемые большие данные

Финансовые данные: стандартные финансовые показатели, хорошо отслеживаемые и понятные.

Корпоративные данные: то же самое плюс более широкие операционные и транзакционные данные, которые можно использовать для подкрепления аналитических выводов и в целях прогнозирования.

Большие данные: то же самое плюс новые виды внутренних и внешних данных, значительная часть которых не структурирована, но определенная часть может привести к новому пониманию эффективности деятельности, возможностей и рисков.

Одна из компаний, которая реализует эту тенденцию на практике, — Procter & Gamble (P&G). Все более глубокое и комплексное оцифровывание ими данных позволяет говорить о наличии у компании «источника конкурентного преимущества».

Это помогает улучшить все — от инновационных продуктов, анализа в реальном времени комментариев клиентов в социальных сетях до использования отслеживания изменений данных в целях оптимизации планирования магазинов розничной торговли⁶. Опыт P&G подтверждается широкими исследованиями. Исследования, проведенные в Школе менеджмента им. Слоуна показывают, что компании, принимающие решения на основании использования данных, получают 5—6%-ное повышение производительности.

Эти примеры демонстрируют, что возможности, которые несут с собой большие данные, достаточно существенны. И на этом фоне ИТ-разработчики предлагают все расширяющееся множество инструментов для обработки больших данных, начиная от новых аналитических приложений и исполнительных панелей и заканчивая системами прогнозного анализа (см. Терминологический словарь). Согласно циклу зрелости технологий Gartner, большие данные находятся на пике ожиданий.

Кривая зрелости технологий графически представляет стадии, через которые проходит технологическое новшество в процессе становления. Суть в том, что сначала срабатывает так называемый технологический триггер: возникает инновация, о ней все начинают говорить, ожидания раздуваются выше истинного потенциала, за «пиком чрезмерных ожиданий» логично приходит разочарование. Позже, как правило, в течение пяти — десяти лет, когда технология проходит практическую обкатку у пользователей и те получают осязаемые результаты, становится ясен реальный ее потенциал. Это показывает, что, прежде чем будет достигнуто полное понимание значения технологий больших данных, неминуем период разочарования.

Прогресс в технологиях наряду с более простыми и доступными инструментами анализа и визуализации данных открывает возможности для использования данных компаниями самых разных весовых категорий. Однако большинству из них предстоит хорошенько потрудиться, чтобы приспособиться к ситуации, когда управляющая роль в бизнесе отводится данным. Согласно одному из отчетов, каждая четвертая компания до сих пор не задействовала большую часть своих организационных данных, а 53 % считают, что они используют только половину своих значимых данных⁹.

Все больше внимания уделяется новым аналитическим методам и техническим средствам, цель которых — извлечение из данных новых идей. При этом лишь немногие компании обладают дополнительными навыками, необходимыми для перевода извлеченных идей в коммерческое русло. Это как раз тот момент, когда специалисты по управленческому учету могут заявить о себе.

Далее мы подробнее рассмотрим способность финансистов вносить существенный вклад в общее дело команды, работающей над новыми инициативами в сфере анализа данных. Передовые аналитические методы извлечения данных, необходимые для выявления новых корреляций и выработки алгоритмов прогнозирования поведения, — это сфера деятельности исследователей данных. Но роль, которую играют специалисты по управленческому учету в формировании финансовой отчетности, а также в процессах бюджетирования, прогнозирования и управления производственными показателями, делает необходимым их контакт с каждой гранью бизнеса.

Такого рода участие в информационном наполнении позволяет специалистам по управленческому учету прекрасно разбираться в самом бизнесе. Таким образом, они могут играть важную роль в обращении идей, полученных в результате анализа, в осязаемый коммерческий результат.

Терминологический словарь¹⁰

Последние технологические разработки меняют способы управления данными и их анализа в ведущих компаниях. Специалисты по управленческому учету должны владеть представленной ниже терминологией.

Data analytics, Анализ данных, средства анализа данных

Средства анализа данных — передовые формы анализа, которые могут быть использованы для изучения больших объемов данных и передачи идей. Они могут использоваться для выявления корреляций и выработки алгоритмов в целях прогнозирования поведения. Средства анализа структурированных или корпоративных данных хорошо известны. Они используются многими компаниями и организациями как помощь при принятии качественных бизнес-решений, а также при тестировании и проверках моделей и теорий. В настоящее время от средств анализа данных ожидают больших подвижек в отношении работы с неструктурированными данными.

Cloud computing, Облачные вычисления, облачная обработка данных

Под облачными вычислениями понимают предоставление различных сервисов, таких как программные приложения, платформы разработки, серверы, обрабатывающие мощности и хранение данных на удаленных серверах через Интернет (в отличие от локальных серверов). Как правило, за пользование облачным сервисом взимается плата, а серверные программы и инфраструктура управляются сторонним поставщиком.

Dashboards, Информационные панели

Информационные панели — пользовательский интерфейс, который организует и представляет корпоративную информацию удобочитаемым и понятным образом. Панель может использоваться для объединения ряда данных и КПЭ, как правило, визуально, для того чтобы руководители могли контролировать эффективность бизнеса.

Data mining, Извлечение данных

Набор методов, используемых для просеивания очень больших объемов данных. Извлечение данных предполагает применение методов

искусственного интеллекта и передовых статистических инструментов (например, кластерный анализ и регрессии) для выявления тенденций, закономерностей и отношений.

Data scientist, Специалист по обработке данных

Специалист по обработке данных — тот, кто выполняет статистический анализ данных и извлекает информацию из больших объемов данных, как правило, с целью выявления тенденций, показателей и других необходимых сведений, используя в работе передовые методы моделирования, статистики, анализа и математические методы.

Data visualization, Визуализация данных

Визуализация данных — общий термин, используемый для описания технологии, которая позволяет руководителям наблюдать тенденции и комбинации данных. Инструменты визуализации зачастую не ограничиваются стандартными диаграммами и графиками Excel, а используют интуитивно более понятные средства в виде циферблатов, датчиков, географических карт, графиков временных рядов, тепловых карт и пр. Программные средства визуализации позволяют вычленять комбинации данных, тенденции, корреляции, которые иначе могли бы быть пропущены.

Hadoop, платформа Apache Hadoop

Hadoop — разработанный на Java набор утилит, библиотек, поддерживающих обработку больших массивов данных в распределенной вычислительной среде. Этот проект с открытым исходным кодом, находящийся под управлением Apache Software Foundation, обычно используется для обработки огромных объемов данных с охватом тысяч серверов.

In-memory processing, Обработка данных в оперативной памяти

Обработка данных в оперативной памяти позволяет анализировать большие массивы данных значительно быстрее, чем раньше. Местом обработки является системная память, а не традиционный жесткий диск.

Internet of things, Интернет вещей

Интернет вещей — концепция вычислительной сети физических объектов, которые подключены к Интернету, за счет чего от них исходит информация или предупредительные сигналы. Подключаясь и контролируя станки, транспортные средства, оборудование, инвентарь и многое другое, организации получают в распоряжение новый источник огромного объема информации буквально о каждой грани своего бизнеса.

MapReduce

MapReduce — среда разработки программного обеспечения, позволяющего разработчикам писать программы для параллельных вычислений очень больших наборов данных в распределенных кластерах процессоров или автономных компьютеров.

Metadata, Метаданные

Метаданные — данные, которые описывают другие данные. Представляют собой ценную ссылку, которая позволяет организовать и локализовать определенные типы данных. Пример метаданных простого документа: метаданные о дате создания документа, о дате его изменения и о файле. Метаданные также используются для изображений, видео, электронных таблиц и веб-страниц.

OLAP

OLAP (англ.: online analytical processing — аналитическая обработка в реальном времени) — категория программных средств анализа данных, хранящихся в базах данных. Инструменты OLAP позволяют пользователям анализировать различные параметры многомерных данных. Например, с их помощью можно наблюдать временные ряды и анализ трендов.

Predictive analytics, Прогностическая аналитика

Прогностическая аналитика — отрасль интеллектуального анализа данных (извлечения данных), занимающаяся прогнозированием вероятностей. Используются измеряемые переменные, чтобы спрогнозировать поведение

человека либо организации. Прогностическая аналитика эффективно использует корпоративные знания организации, применяя сложные методы анализа корпоративных данных. В бизнесе прогностическая аналитика часто служит для получения сведений о поведении клиентов и выработки предложений по оптимальному распределению ресурсов для получения максимальной отдачи.

Social media, Социальные сети, Соцмедиа

Социальные сети — общее обозначение программных средств и платформ, которые позволяют группам генерировать контент и участвовать в коммуникации в одноранговых сетях. Яркие примеры таких платформ: Facebook и Twitter.

Unstructured data, Неструктурированные данные

Это данные, которые не имеют заранее определенной структуры и, как следствие, не могут легко храниться в стандартных реляционных базах данных. В качестве примеров можно привести электронную почту, текстовые документы, изображения, видео, записи кол-центров.

Использование данных в целях конкурентного преимущества

При подготовке настоящего отчета мы опросили более 2000 финансовых директоров и финансовых специалистов. У них есть ясные представления о потенциале больших объемов данных: почти девять из десяти (87 %) отметили, что в течение следующих десяти лет большие данные изменят саму модель ведения бизнеса.

На базовом уровне компании все чаще будут принимать решения на основании анализа данных — при выработке стратегий, для решения отдельных оперативных вопросов, улучшения общих производственных показателей. Помимо этого, часть предприятий найдут способы превратить данные в новый источник дохода или даже выстроят вокруг данных свои бизнес-модели.

Решать вопрос о том, как следует измениться, чтобы максимально использовать потенциал данных, придется не только организациям, но и их финансовым службам. Примерно восемь из десяти (84 %) опрошенных финансистов полагают, что большие данные и аналитика потребуют от них изменения алгоритма работы уже в ближайшие пять лет. Например, компании все чаще будут обращаться к своим финансовым службам с запросами прогноза эффективности в обозримом периоде, вместо того чтобы просто требовать данные бухгалтерского учета, которые обычно носят ретроспективный характер. При этом сегодня почти каждый третий (32 %) считает, что его организация не имеет навыков, необходимых для использования новых и растущих объемов данных. Это согласуется с прогнозами экспертов о грядущем дефиците специалистов с глубокими аналитическими навыками.

Но однажды специалисты по управленческому учету уже доказали свою способность адаптироваться к изменяющимся условиям. Сто лет назад, руководству нужна была более полная информация о затратах и процессах, поскольку с наступлением эпохи индустриализации изменилась бизнес-модель. Именно в тех условиях и зародился профессиональный управленческий учет. Сегодня новым источником идей становятся большие данные, а это означает, что процесс принятия решений в очередной раз необходимо совершенствовать. Те специалисты по управленческому учету, которые окажутся способными использовать обширные возможности, заложенные в больших данных, станут играть важную роль в процессе переформатирования бизнес-моделей.

Межотраслевые возможности

С увеличением объема доступных для обработки данных у компаний любой формы организации вне зависимости от размера возникает масса возможностей. В качестве примера недавний опрос агентства Economist Intelligence Unit, проведенный среди топ-менеджеров, показал тесную

взаимосвязь между ростом доходов и стратегическим использованием данных¹¹. Кроме того, результаты крупного исследования McKinsey продемонстрировали, что в каждом из изученных секторов процесс создания существенной ценности идет гораздо лучше, если используются данные¹². Помимо этих примеров, имеется растущий перечень компаний, которые извлекают из данных экономическую выгоду. Подробно они представлены на рис. 2, где можно наблюдать все разнообразие открывающихся возможностей.

Эти и другие примеры дают понять, что компании используют данные совершенно по-разному: большинство — для улучшения операционных показателей, но есть и те, что превращают данные в элемент своей ключевой компетенции, создавая на их основе совершенно новые продукты и услуги. В зависимости от основных задач целенаправленность исследования данных также варьируется. Некоторые компании сегодня концентрируются прежде всего на анализе структурированных данных, уже попавших в систему управления предприятием, и, очевидно, имеют возможность дальнейшего продвижения. Но, даже ограничиваясь лишь структурированными данными, компании могут восполнить пробелы, прибегнув к помощи внешних поставщиков данных, таких как, например, Experian, Dun & Bradstreet. А к услугам уже искушенных и опытных исследователей данных есть воистину непаханое поле для работы с большими данными.

Несмотря на то что большинство примеров из открытых источников относятся к крупным компаниям, небольшие тоже вовлечены в процесс. Некоторые большие организации даже создают сервисы по передаче данных, чтобы оказывать более мелким компаниям в цепочке поставок аналитические услуги. Например, P&G в настоящее время помогает небольшим ретейлерам анализировать, как улучшить эффективность работы их торговых точек¹³. Это говорит о том, что использование данных полезно с точки зрения как операционных преимуществ, так и новых возможностей для бизнеса.

Таблица ниже дает представление о том, как самые разные компании используют данные, чтобы получить операционные или стратегические преимущества. Представленная классификация не претендует на исключительную точность. Даются иллюстрирующие примеры, охватывающие все четыре сектора.

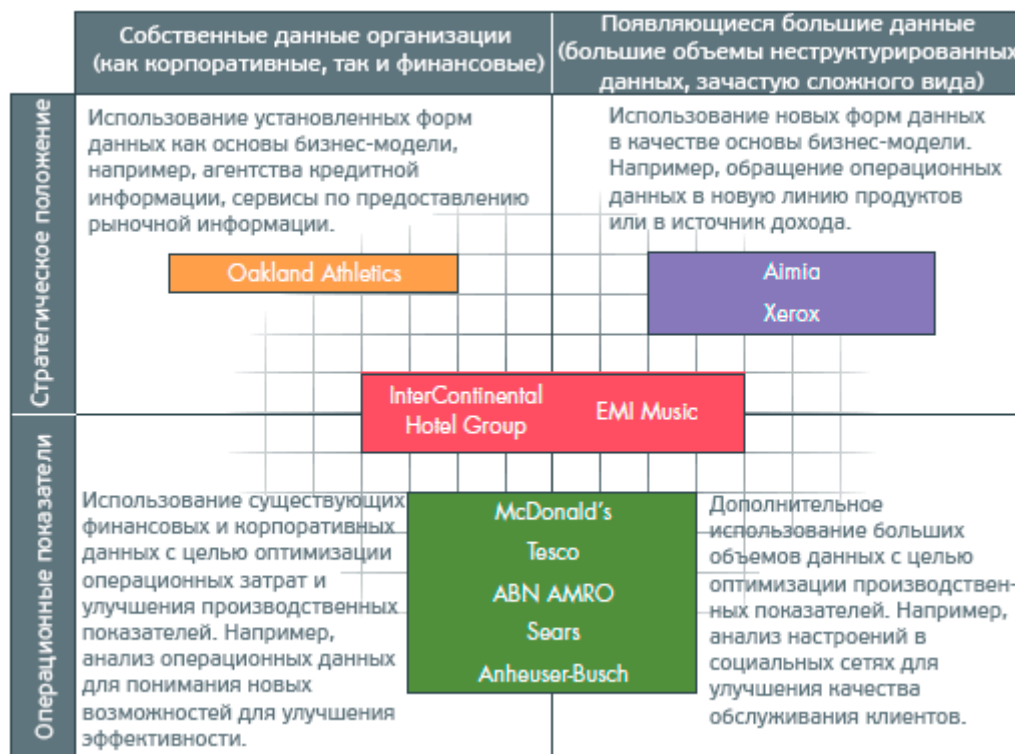


Рис. 2. Четыре основных вида использования данных [14](#)

Компан ия	Сектор	Действия по использованию данных	Полученные преимущества
Anheuser-Busch	Напитки	Отслеживание корпоративных и больших данных, охватывающих розничные точки и разные модели продаж с целью точного позиционирования пивных брендов,	Повышение качества мерчандайзинга в магазинах в целях роста продаж

		мерчандайзинга в отдельных магазинах и районах	
Sears	Розничная торговля	Создана служба анализа расширенных данных, чтобы сократить время разработки промоакции с восьми недель до одной и одновременно повысить эффективность кампаний, которые стали носить гораздо более целенаправленный характер. Извлечение огромных объемов данных о клиентах, продукции и рекламных данных.	Улучшение показателей продаж и эффективности маркетинговой деятельности
ABN AMRO	Розничные банковские услуги	Исследование целого ряда новых источников данных с целью лучшего понимания клиентов и моделей продаж банка, например более точное сегментирование потенциально выгодных клиентов и создание для них персональных предложений.	Повышение качества обслуживания клиентов и эффективности продаж
Tesco	Розничная торговля	Обработка транзакционных данных	Улучшение показателей

		ККМ, привязанных к картам лояльности клиентов, для лучшего понимания закономерностей совершения покупок. Более целенаправленные предложения, например подгузники и пиво новоиспеченным родителям — ввиду того, что с рождением ребенка они не смогут часто посещать пабы.	продаж и эффективности маркетинговой деятельности
McDonald's	Фастфуд	Анализ больших объемов данных с целью комплексной оптимизации — от выбора блюд из меню, затрат на выдачу заказов до уровня запасов и комплектации кадров. Например, можно ускорить продвижение очереди, прогнозируя вероятные заказы.	Повышение общей операционной эффективности
InterContinental Hotel Group	Гостиничный бизнес	Создание платформы данных, которая использует широкий круг внутренних и внешних данных для улучшения качества	Повышение операционной эффективности и качества принимаемых

		операционной отчетности и обслуживания клиентов, прогнозирование моделей покупок гостей, упреждающее выявление рыночных тенденций, что помогает лучше направлять инвестиции.	стратегических решений
EMI Music	Медиа и развлечения	Использование больших данных при отборе перспективных артистов, задействование различных источников данных, например о скачивании в Интернете, упоминание на телевидении и радио, объемы пиратского копирования и т. д. Использование полученных данных для поиска новых имен с гарантированными суперхитами.	Улучшение показателей продаж, эффективности маркетинговой деятельности и принятия стратегических решений
Oakland Athletics	Профессиональный бейсбол	История показана в художественном фильме Moneyball, когда испытывающая трудности бейсбольная команда использует аналитические	Улучшение качества стратегических решений

		данные, чтобы найти ценных игроков по низкой цене и улучшить положение команды в лиге	
Aimia	Программа лояльности	В партнерстве с Yahoo! предлагают клиентам баллы лояльности в обмен на рекламный контент, содержание которого определяется в привязке к истории поиска; контент становится гораздо более целенаправленным, появляется совершенно новое бизнес-предложение.	Создание новых продуктов и услуг на базе данных
Херох	Управление документами и печать документов	Создание в городах сервисов по обработке и передаче данных на базе уже существующих печатных сервисов и сервисов по управлению документами. Например, разработка динамической системы ценообразования для парковки в Лос-Анджелесе на базе существующих сервисов по печати талонов,	Создание новых продуктов и услуг на базе данных

		позволяющей диверсифицировать бизнес.	
--	--	--	--

Большие данные — малый бизнес

Большие данные и аналитика — это многочисленные возможности также и для малых и средних предприятий (МСП). Более мелкие компании, как правило, располагают ограниченными ресурсами и более скромным бюджетом, однако у них все же есть кое-какие преимущества: часто более гибкая ИТ-инфраструктура, которая испытывает меньше проблем, связанных с использованием устаревших систем или разрозненных баз данных, и способность быстро приспособиться к новым методам работы.

Результаты исследования показывают, что, несмотря на склонность небольших компаний полагаться при принятии решений скорее на интуицию, чем на данные, они все же понимают возможности, которые несут большие данные, и маловероятно, что их будет сдерживать фактор цены. В эту эпоху господства данных малые и средние предприятия, игнорирующие ценность данных при принятии решений, рискуют проиграть тем, кто использует данные для улучшения производственных показателей и извлекает из данных новые идеи.

Используя специализированное программное обеспечение или сервисы, малые и средние предприятия могут объединять свои корпоративные данные с внешними данными и таким образом обнаруживать новые идеи. Несмотря на то что традиционные инструменты, предлагаемые крупными поставщиками, могут быть не по карману многим МСП, для анализа данных им вполне подходят онлайн- и облачные сервисы, среди которых и Google Analytics, и Kaggle, и Swipely. Эти инструменты дают МСП доступ к мощным средствам, которые помогут им осмыслить эффективность их деятельности с новых позиций. Крупные производители также разрабатывают аналитические платформы для компаний среднего бизнеса, в том числе облачные приложения.

Это может начинаться на самых базовых уровнях: когда в свое время оператору парков развлечений Morey's Piers (США) потребовалось сопоставить популярность различных аттракционов, там просто собирали билеты в мешки и взвешивали их¹⁵. «Это было весьма бесхитростное и прямое решение», — отмечает финансовый директор фирмы Джим Блейк, дипломированный бухгалтер, имеющий статус CGMA. Сегодня же электронные считывающие устройства собирают точные сведения о посещении аттракционов, на основании которых можно сделать вывод о популярности аттракциона либо о необходимости выведения его из эксплуатации. Чтобы закрепить результат и двигаться дальше, Блейк привлек фирму, специализирующуюся на анализе внешних данных, с целью проведения более глубокого анализа продаж билетов через Интернет и определения самых продаваемых аттракционов. «Размышляя о том, куда нам двигаться, я думаю стратегически, — объясняет Блейк. — Речь здесь не идет о ближайшей перспективе, на неделю. Мне нужен прогноз на три, четыре, пять лет с настоящего момента».

Сеть магазинов спортивных товаров для женщин Sweaty Betty (Великобритания) привлекла для детального изучения своих клиентов компанию Experian, чьи специалисты проанализировали необходимые геодемографические данные и представили рекомендации по месторасположению новых торговых точек.

Недалеко от Бостона находится ресторан Farmstead Table, в нем подают блюда, приготовленные из местных продуктов. Ресторан привлек для анализа своих посетителей и транзакций компанию Swipely, чтобы найти решения, которые позволили бы удивлять и привлекать гостей каким-то новым, изысканным способ. Например, выяснив в ходе анализа, что часть постоянных клиентов всегда заказывают свежеевыловленного лосося, ресторан стал заранее сообщать им о поступлении рыбы нового улова. В результате повысились продажи, как и лояльность.

Некоторые малые предприятия даже разрабатывают новые услуги по передаче данных на базе этих программных средств. Один из таких примеров — агентство Powerhouse Factories, оказывающее услуги по созданию брендов. Они взяли инструменты по обработке данных от аналитической платформы Tableau, исключили привязку к Excel и добавили средство визуализации, которое помогает клиентам в самых разных видах применения.

Используя его, можно, например, понять, как задержки при обслуживании на кассе влияют на общие показатели. Платформа помогла упростить процесс взаимодействия с клиентами, поскольку все данные стали размещаться в одном месте, доступном для каждого. При этом обеспечивается единое и последовательное представление необходимой корпоративной информации.

Более предприимчивые МСП пользуются даже ресурсами, которыми обладают передовые исследователи данных, не имея ни бюджета, ни технических возможностей разместить такие ресурсы¹⁹. Так, в Jetpac (это компания-разработчик мобильного приложения, которое составляет путеводители из общедоступных фотографий пользователей) решили разработать алгоритм автоматизации процесса поиска лучших фотографий. Они обратились к платформе Kaggle, которая сопоставила относящиеся к данным запросы с ведущими исследователями данных со всего мира.

Через три недели — и всего за 5000 долларов — задача была решена. Гораздо дешевле и быстрее, чем если бы Jetpac делали все самостоятельно.

В конечном итоге такие инструменты расширяют спектр доступных для МСП возможностей в применении к анализу данных. Необходимо подчеркнуть, что использование инструментов анализа данных не накладывает на МСП неподъемных финансовых обязательств перед поставщиками ИТ-услуг.

Кейс. Секреты успеха от DUNNHUMBY

Dunnhumby, ведущая компания по изучению потребителей и потребительского поведения, использует передовые средства анализа данных, чтобы помочь своим заказчикам понять собственных покупателей. Компания заявила о себе в середине 1990-х годов, проведя анализ поведения постоянных покупателей Tesco — крупнейшей британской сети супермаркетов. Для рассмотрения клиентской базы ретейлера с совершенно новых позиций были использованы данные карт Tesco Clubcard.

В компании, с которой сотрудничают многие ведущие мировые бренды, считают, что сегодня малые предприятия также в состоянии извлечь пользу из качественного анализа данных. Но, прежде чем начинать какой-либо аналитический проект, компании должны получить представление о том, какие информационные активы уже имеются в их распоряжении. «Нужно начинать с ваших собственных данных, — говорит исполнительный директор по глобальным данным Мэтью Кейлок. — Покупать информацию о потребителях у третьих лиц имеет смысл, когда вы ищете потенциальных заказчиков, это может детализировать ваши собственные наработки, но не будет полезно для выстраивания собственной системы».

По этой причине Кейлок рекомендует изначально оценить, какие источники данных уже есть в наличии (например, контактные данные покупателей, информация об оргструктуре, история контактов, в том числе история покупки продуктов и услуг, и т. д.). Также важно попробовать найти тех, кто владеет данными внутри организации, и изучить, каким образом через обмен источниками данных можно достичь быстрых результатов.

«Компаниям нужно не только стараться максимально использовать свои источники данных и создавать ценность именно на их основе, но и передавать знания о покупателях по всей товарно-сбытовой цепочке», — советует Кейлок. Компании должны сегментировать своих покупателей и проводить регулярную переоценку таких данных, чтобы понимать тенденции и закономерности клиентской базы. Это нужно, например, для выявления

моментов снижения покупательского интереса и поиска соответствующих коррекционных мер, а также для поощрения желательного поведения. «По мере упорядочения данных получаемые из них идеи, как правило, ставят под сомнение уже сложившиеся корпоративные нормы, и вы узнаете многое из того, чего до этого не знали», — добавляет Кейлок.

И, наконец, Кейлок рекомендует компаниям воплощать аналитические проекты поэтапно и выстраивать реализацию последующих этапов с учетом результатов предыдущих, то есть закреплять успех. Он замечает, что «масштабный проект по анализу данных, после реализации которого ожидается получение какой-либо ценности — это всегда высоко рискованное мероприятие!».

В чем потенциал данных для финансовой сферы

Финансовые специалисты приходят к более полному пониманию новых возможностей, которые можно получить, анализируя данные. Когда им задавали вопрос, где, по их мнению, организация могла бы получить наибольшую выгоду от улучшения качества данных и более качественной аналитики, самый популярный ответ относился непосредственно к итоговым показателям организации: максимальная выгода заключается в обретении способности определять возможности сокращения издержек или повышения эффективности.

Это согласуется с результатами других исследований. К примеру, по оценке McKinsey & Company, если использовать большие данные для управления эффективностью, годовая экономия только в секторе здравоохранения США достигнет 300 млрд долларов^{[21](#)}.

Помимо этого, финансовые специалисты также видят явные возможности для извлечения выгоды вследствие разработки и мониторинга новых или усовершенствованных КПЭ. Например, финансовая служба компании Unilever создала информационную панель данных, которая использует весьма разнообразный набор источников — от социальных сетей до маркетинговых

агентств, чтобы сформировать комплекс ключевых показателей, которые будут актуальны в глобальном масштабе, последовательны и практически осуществимы и которые можно задействовать при составлении отчета о финансовых результатах и в работе с денежными потоками.

Артур Тан, старший директор по бизнес-планированию Worldwide Studios (группа разработчиков видеоигр для Sony PlayStation) сообщает, что изучение новых источников данных помогло компании разработать более действенные показатели игровой эффективности — от среднего за день количества геймеров и коэффициентов удержания до процентов монетизации. «Вероятнее всего, вы не сможете извлечь такие результаты из своих финансовых систем, используя традиционные источники данных. Большие данные — вот, что вам необходимо», — поясняет он.

По мнению финансистов, на третьем месте после определения возможностей и разработки КПЭ стоит польза от применения данных для прогнозирования на основе движущих факторов или базирование финансовых прогнозов на операционных факторах (см. рис. 3). Такое прогнозирование требует исключительно качественных и актуальных данных²³, что объясняет, почему многие респонденты видят в этом столь явную потенциальную выгоду.

Приоритеты использования данных (в порядке убывания):



Source: CGMA data survey 2013²⁵

Приоритеты МСП в сравнении с крупными компаниями

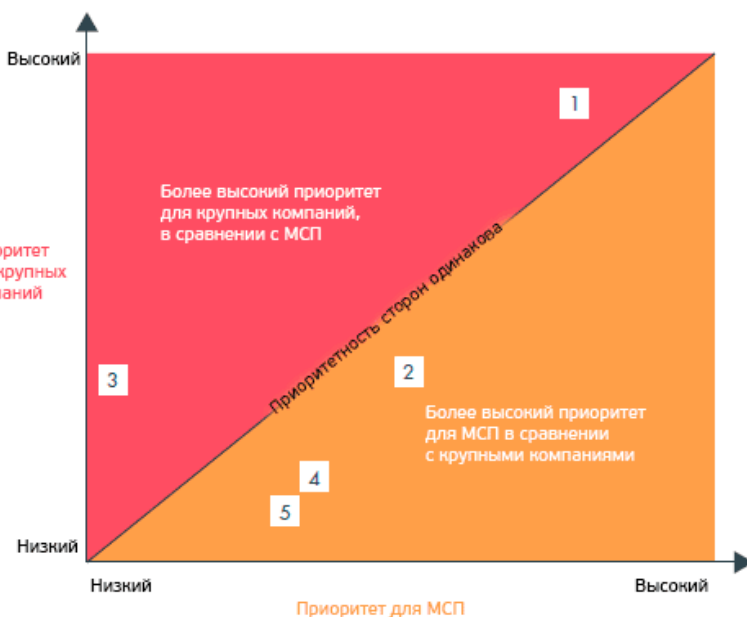


Рис. 3. Первая пятерка направлений бизнеса, которые могли бы выиграть от более высокого качества данных и более совершенной аналитики

Но на этом преимущества не заканчиваются. Выгода еще и в том, что можно повысить отдачу от принятия решений и стратегического планирования. Вот как объясняет это Джеймс Милн, старший финансовый директор финансовой службы глобальных операций Yahoo!: «Вместо того чтобы полагаться на некогда традиционные ежемесячные отчеты, теперь мы в первую очередь смотрим на ежедневные показатели, которые говорят нам о том, как наши пользователи ведут себя в самых разных аспектах. Это позволяет нам быть более гибкими в принятии решений».

Вооружившись знаниями, полученными в результате анализа данных, компании могут улучшить основные процессы и скорректировать бизнес-модели. Часто это приводит к серьезной экономии средств или к новым возможностям для бизнеса.

В United Parcel Service использовали анализ и оптимизацию маршрутов доставки и режимов вождения и, как следствие, в 2011 году сократили потребление топлива на 8,4 миллиона галлонов²⁵. Другой пример: в Visa проанализировали транзакционные данные наряду с целым рядом новых источников информации, таких как геолокация, и им удалось снизить годовые убытки от мошенничества на миллиарды долларов²⁶.

Конечно, данные можно обратить в доход. В Sony Games данные по продажам и игровому процессу теперь используются для стимулирования продаж, способствуя более глубокому пониманию игрового опыта геймеров (см. CASE STUDY «Использование данных с целью повышения прибыльности игр в Sony»). Один из самых ярких примеров этого — Rolls-Royce, где получили миллиарды долларов прибыли от обслуживания, благодаря тому, что стали продавать не просто реактивные двигатели, а летные часы.

Теперь мы в первую очередь смотрим на ежедневные показатели, которые говорят нам о том, как ведут себя наши пользователи... Это позволяет нам быть более гибкими в принятии решений.

James Miln, Yahoo!

Кейс. Использование данных с целью повышения прибыльности игр в SONY

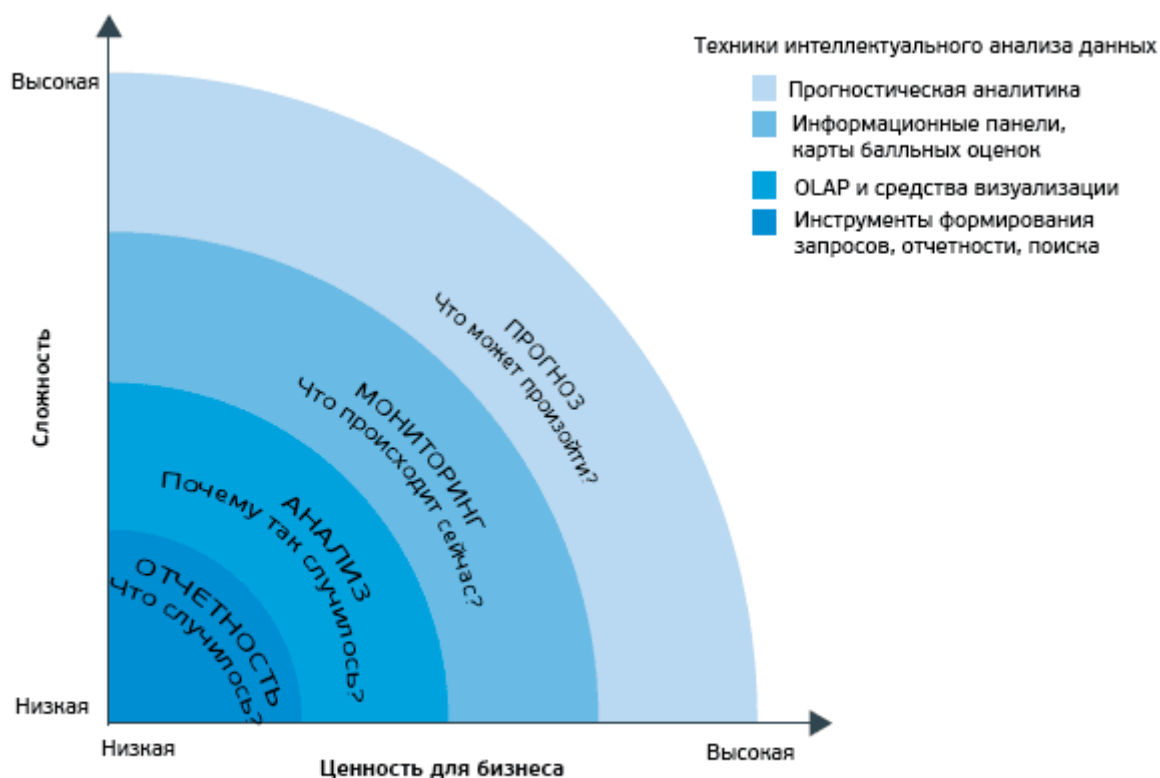
Sony PlayStation — это крупнейший мировой бренд. Многие из выпущенных для этой платформы игр подолгу остаются бестселлерами. Это жизненно важно для бизнеса, потому что консоли продаются с минимальной наценкой, а основная прибыль поступает от игр. Игры — лидеры продаж приносят высокую прибыль, отсюда следует, что бизнесу чрезвычайно важно определить, что же станет бестселлером. Это ключевой вызов для Артура Тана, который контролирует показатели и отвечает за финансовое планирование и результаты интеллектуального анализа данных со стороны подразделения Sony, обеспечивающего разработку новых игр.

И если разработчики продуктов под маркой Sony используют данные для повышения уровня игр, то основная задача Тана — использовать данные для улучшения показателей бизнеса в целом. Это начинается с элементарных КПЭ — от прибылей и убытков по каждой игре и до общего объема продаж и скачиваний. Помимо этого, в Sony стали исследовать дополнительные данные: сколько времени люди тратят на игры, как часто они вводят учетные данные, что они делают в игре — это лишь несколько примеров.

Определив тенденции в целых комплексах поведенческих данных, компания вводит новые параметры и изменения в настройках, чтобы игра стала более привлекательной, либо пересматривает ценовую модель, чтобы повысить спрос. Один из последних примеров: данные показали, что продукт не принесет ожидаемой пользовательской конверсии при значительных расходах, и на основании данных было принято решение свернуть проект, в результате чего были сэкономлены значительные средства.

Предстоящие вызовы

Стремительный рост масштабов данных, которыми сегодня располагают предприятия, ведет и к расширению спектра идей, которые могут быть генерированы с использованием разнообразных аналитических методов. Необходим анализ самого широкого диапазона — от базовой отчетности и вплоть до прогностической аналитики (см. рис. 4).



Source: The Data Warehousing Institute²⁸

Рис. 4. Различные уровни анализа данных

Кроме того, по мере разработки дополнительных инструментов и сервисов ширится диапазон доступных видов анализа. Консалтинговые компании предлагают аналитические услуги в таких сферах, как оборотный капитал, клиентские базы, управление персоналом, маркетинг, профилактическое обслуживание активов, эффективность работы

предприятия, современное планирование, противодействие мошенничеству, анализ рисков и многое другое.

Учитывая размах возможностей, организациям следует начинать с определения ключевых вопросов, ответы на которые им действительно необходимы. В частности, прежде чем углубляться в мир неструктурированного анализа данных, следует в полной мере понять свою бизнес-модель и иметь представление о собственных нематериальных активах, а также о структуре, качестве и источниках данных.

Джо Пеппард, профессор Европейской школы менеджмента и технологий, отмечает: «Предприятия могут использовать данные двумя способами: либо в качестве подспорья в принятии решений, либо в качестве источника новых знаний для себя. Открытие новых знаний, как правило, начинается с постановки вопросов о данных, и зачастую это приводит к тому, что мы обнаруживаем нечто "неизвестное неизвестное".

Таким образом, первостепенное значение имеет и качество наших запросов, и качество используемых данных. О необходимости некоторых решений организации могут узнать, анализируя отчеты и данные информационных панелей, которые еще ранее были определены как вспомогательные средства. Обычно они создаются методами интеллектуального анализа данных и программных генераторов отчетов. О необходимости других решений можно узнать лишь в процессе обнаружения знаний».

Предприятия могут использовать данные двумя способами: либо в качестве подспорья в принятии решений, либо в качестве источника новых знаний.

Джо Пеппард, профессор, European school of Management and Technology

После вычленения проблемы организации должны идентифицировать данные, необходимые для ответа на поставленные ими вопросы. Серьезной самостоятельной задачей является приведение данных к виду, который

позволит проводить анализ. В нашем исследовании наиболее слабым звеном, препятствующим агрегации данных, 62 % респондентов уверенно назвали разрозненность корпоративных данных.



Источник: Исследование данных CGMA, 2013 год.²⁹

Рис. 5. Где вашей организации не хватает навыков, чтобы ухватить и извлечь ценную информацию из данных?

Еще одной проблемой является качество данных. Почти половина финансистов именно в этом видят слабость своих организаций. Кроме того, около двух третей считают, что для того, чтобы качественно информировать ответственных за принятие решений лиц и обеспечить лучшую производительность, сначала следует добиться более высокой точности и достоверности данных. Здесь важно не переусердствовать и найти правильный баланс, так как бухгалтеры склонны преувеличивать значимость целостности данных, поскольку рассматривают ситуацию через призму официальных отчетов. А при исследовании данных необходимо скорее найти модели или взаимосвязи, которые позволят понять направленность, чем выдержать предельную точность.

Сопутствующей задачей при обращении с такой чувствительной материей, как данные, является обеспечение специалистами по

управленческому учету защиты данных и соблюдение стандартов управления. Это отдельная важная тема, выходящая за рамки настоящего отчета, однако совершенно очевидно, что все более глубокое погружение в океан данных требует от специалистов по управленческому учету знаний о потенциальных рисках и осведомленности в этических вопросах, возникающих при обращении с конфиденциальной информацией.

Ярослав Чрупек, менеджер по глобальным данным в British American Tobacco, отмечает, что для максимальной отдачи от инвестиций в технологии обработки больших данных решающее значение имеют надежность информационной инфраструктуры организации и процедуры по контролю соответствия нормам и правилам. «Инвестиции в аналитику больших данных могут быть значительными в сравнении с аналогичными расходами на построение надежной системы управления информацией, что гарантирует надлежащее управление стандартами, имеющими отношение к данным, и постоянное повышение квалификации соответствующих групп специалистов в области использования аналитических инструментов и средств интеллектуального анализа данных».

Кейс. Связь финансов и ИТ в компании BRITISH AMERICAN TOBACCO

Один из крупнейших в мире производителей сигарет, компания British American Tobacco (BAT) весьма заинтересована в получении преимуществ, которые дают стратегии, выстроенные с учетом анализа данных. Ярослав Чрупек — менеджер по глобальным данным в BAT, в чьи обязанности входит наблюдение за данными и контроль управления процессов с позиции как информационных технологий, так и финансов. «В British American Tobacco есть специалисты, которые прекрасно разбираются в финансовых вопросах и обладают знаниями и опытом в сфере ИТ. Такое сочетание позволяет им понимать, что конкретно необходимо бизнесу и каковы перспективы достижения этого при помощи данных».

ВАТ добилась значительного прогресса за счет внедрения единой платформы данных, которая позволила изменить операционную модель компании и обрести высокую операционную гибкость. Например, теперь усовершенствование новой интегрированной системы в целях соответствия меняющимся потребностям бизнеса стало намного проще и дешевле, чем раньше, когда это делалось на нескольких платформах.

По словам Чрупека, данные могут стать бесценным инструментом, с помощью которого специалисты по управленческому учету достигнут одной из своих главных целей — максимально обширного и глубокого понимания бизнеса, в котором они работают. При этом он советует финансовым службам сфокусироваться на изучении потенциала данных, которыми уже располагает организация, прежде чем принимать решение об инвестировании значительных средств в новейшие инструменты анализа больших данных. «Мощности для работы с большими данными довольно дорого обходятся, а использовать их могут только специалисты по работе с реальными данными. Такой же большой эффект можно получить, используя другие базовые средства, например инструменты интеллектуального анализа, которые сможет применять каждый средний пользователь в организации», — заключает Чрупек.

Обращение идеи в импульс развития

Наше исследование показывает, что финансисты уже сотрудничают с другими службами в деле извлечения из данных ценности и идей. Многие отмечают, что в некоторых областях финансовые службы уже берут на себя ведущую роль. Часто задачи распределяются между несколькими внутренними командами.

Например, по словам Гвинет Джиттингс, вице-президента по глобальной сегментной и управленческой отчетности American Express, анализом корпоративных данных занимаются и финансовая, и другие службы. Специалисты по управленческому учету компании активно участвуют в

анализе процессов и производительности, целью чего является как повышение эффективности внутри организации, так и улучшение результатов бизнес-активности, а специалисты по бухгалтерской отчетности больше заняты анализом результатов и разбивкой данных на конкретные компоненты. Прогностический анализ, который выполняют многие группы поддержки финансовых решений, также находится в зоне ответственности специализированных отделов и специалистов по обработке данных, не состоящих в финансовых службах.

Это не единичный пример. Среди опрошенных 37 % — единственная самая большая группа — сообщили, что в их компаниях финансовые службы играют ведущую роль во всех мероприятиях по анализу данных, когда это связано с финансами, и вспомогательную — если анализ данных необходим для чего-то другого. Это согласуется с сообщениями опрошенных о том, что при достаточно небольшом вкладе финансистов в рамках углубленной аналитики они выходят на первые роли, когда речь заходит о преобразовании полученных в результате анализа идей в ощутимый коммерческий результат. Фактически внимание почти каждого четвертого (23 %) преимущественно сосредоточено на представлении отчетности по данным о производительности, а аналитический компонент отводится другим службам. Несмотря на разные подходы, опрошенные эксперты в целом согласились, что для результативного участия в процессе освоения данных финансовым службам просто необходимо быть частью команды внутри организации, в которую входят специалисты разных служб.



Рис. 6. Какую роль в извлечении идей из данных в вашей организации играет финансовая служба?

У финансистов есть все возможности поставить правильные вопросы на самых первых этапах проектов по анализу данных, с другой стороны, они могут обеспечить использование «добытых» идей при принятии решений и превращения их в фактор, влияющий на эффективность бизнеса. В сущности, они могут выполнять функцию моста между специалистами-аналитиками, которые проводят анализ основных данных, советом директоров и топ-менеджментом, которым необходимо иметь в распоряжении конкретные идеи, извлеченные из данных. Как выразился Джеймс Милн, старший директор финансовой службы глобальных операций Yahoo!, «финансовая служба — связующее звено между данными и финансовыми результатами использования этих данных». Кроме того, это дает финансовым службам ключевую возможность играть более активную роль при выявлении новых возможностей для бизнеса.

Партнерство — вот что исключительно важно

Повышение роли специалистов по управленческому учету в извлечении коммерческой ценности из данных предполагает, что им необходимо более тесное сотрудничество с тремя группами заинтересованных сторон. Во-первых, жизненно важен плотный контакт с ИТ-специалистами, поскольку без них не обойтись при сборе и извлечении данных из внутренних ИТ-систем. Во-вторых, сегодня многие компании привлекают для работы профессиональных специалистов по обработке данных, которым, в свою очередь, необходима помощь финансистов при формулировании правильных запросов, а также интерпретации идей и закономерностей, вытекающих из аналитического процесса. И, наконец, финансисты должны иметь тесные связи с руководителями структурных подразделений организации, чтобы обеспечить сбор и интерпретацию данных по всем правилам и гарантировать таким образом последовательное представление данных, необходимое для использования их при принятии бизнес-решений.

Джеймс Милн из Yahoo! подчеркивает роль, которую должна играть здесь финансовая служба: «Грамотные финансовые специалисты задают правильные вопросы о бизнесе». По крайней мере они могут проконтролировать, чтобы анализ данных проводился в соответствии с потребностями бизнеса. Как выразился Ярослав Чрупек из ВАТ, «скорее всего, выделить информацию, которая необходима именно для бизнеса, будет более трудной задачей для ИТ-специалиста, а не для того, кто занимается финансами в организации». Зачастую финансисты становятся частью группы, которая занимается основными данными, что является основой сочетания разнообразных навыков и, значит, залогом надежного анализа.

В рамках таких отношений финансисты могут обеспечивать контроль качества, чтобы гарантировать правильное использование и интерпретацию данных более широкой группой. «Сегодня организации рассчитывают, что финансовые структуры создадут вокруг информации атмосферу строгости и

доверия, — говорит Дэвид Экссон, управляющий директор по управленческому консультированию Accenture. — От них ждут качественной проверки и аналитической целостности. Например, в процессе работы с данными организации получили много корреляций, но весьма разрозненного характера. Нам нужно понять, что на самом деле имеет смысл, и установить причинно-следственные связи. Специалистам по управленческому учету необходимо проверить данные и понять, какие из них важны».

Во-вторых, финансовая служба должна более тесно контактировать с руководителями структурных подразделений и брать на себя инициативу в деле преобразования идей в осязаемые результаты — при включении их в КПЭ, отчетность и бюджеты, с тем чтобы они стали вполне реальны и осуществимы практически. Известный специалист по повышению эффективности бизнеса Бернард Марр считает: «Не нужно превращать финансовых специалистов в исследователей данных. Их присутствие необходимо, чтобы помочь вам интерпретировать данные. Они также могут побудить организацию искать другие наборы данных, которые позволят-таки ответить на важные для бизнеса вопросы».

Это наиболее ценный вклад, которые они могут привнести, однако сегодня совсем немного компаний внедрило у себя подобный порядок. «Финансовые службы, безусловно, играют стратегическую роль при трансформации идей в цифры, которые представляются руководству», — уверены Дэвид Гире и Малколм Уилкинсон из аналитической лаборатории Deloitte. Как показано далее, еще один немаловажный фактор — способность финансовых структур находить новые способы сообщения и визуализации корпоративных данных.

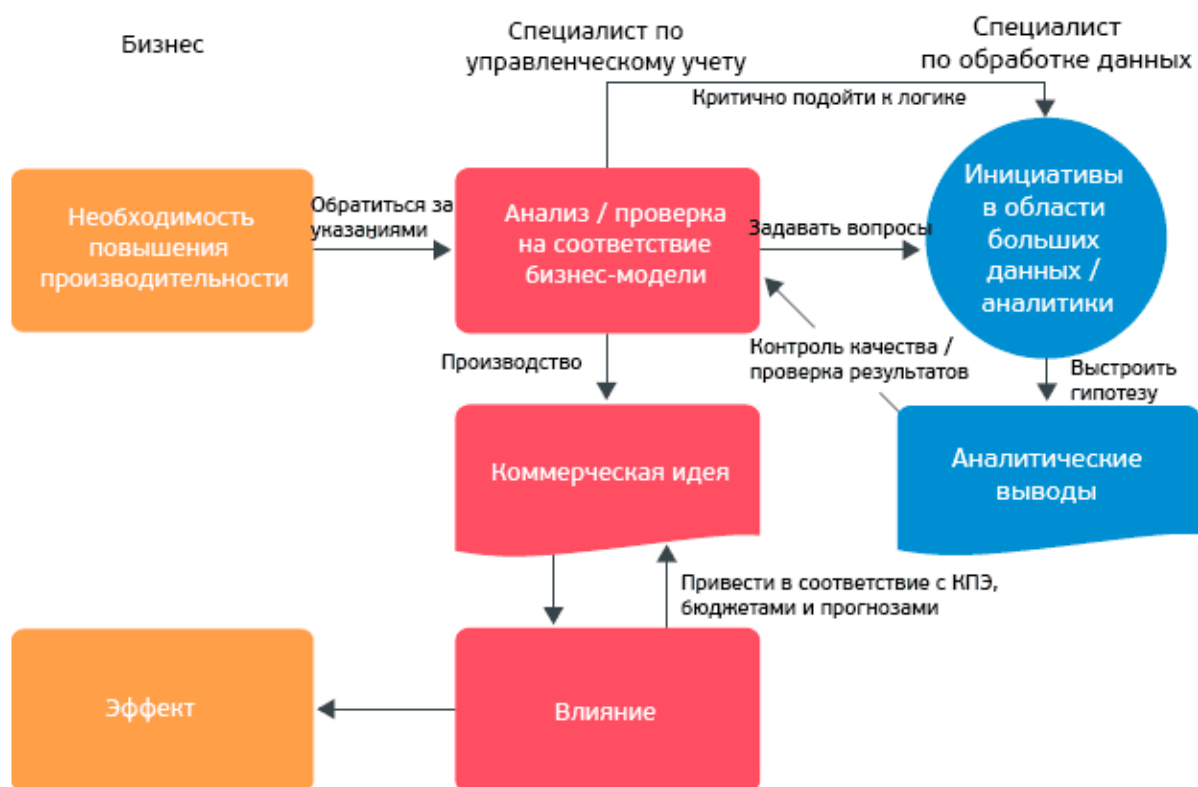


Рис. 7. Каким образом специалисты по управленческому учету связывают бизнес и анализ данных

Но, чтобы добиться успеха на этом поприще, им требуется умение ориентироваться в стратегических вопросах и понимание механизмов функционирования организации. «Специалисты по управленческому учету должны быть достаточно проницательны, чтобы охватить ситуацию целиком и критически оценить представленные результаты. Но это возможно, только если они понимают бизнес-модель и динамику рынка», — считает Ярослав Чрупек из ВАТ.

Сегодня организации рассчитывают на то, что команды финансистов создадут вокруг информации атмосферу строгости и доверия.

Дэвид Экссон, Accenture

Кейс. Еще больше сближая финансы с бизнесом в YANOO!

Многие компании сектора высоких технологий становятся первыми, кто внедряет у себя методы работы, которые впоследствии распространяются и на

весь деловой мир. Так, финансовая служба Yahoo! уже давно четкое представляет данные, которые получает посредством своих ключевых продуктов, таких как Yahoo! Mail, и которые дают объективное понимание того, что же является движущей силой получения прибыли. По мнению Джеймса Милна, старшего финансового директора, возглавляющего службу глобальных финансовых операций Yahoo!, для компании это жизненно необходимо. «В финансовом подразделении должны понимать, кто наши пользователи, чем они занимаются и что делает их привлекательными для рекламодателей. Мы тесно сотрудничаем с группами по обработке данных — даже на уровне элементарных дефиниций типа «что такое пользователь» и «что значит придавать пользователю ценность».

Такой подход становится еще более важным по мере того, как оцифровывается все больше информации. Значительное влияние на продукты Yahoo! оказывают модели отношений между брендами и пользователями в социальных сетях. «Еще совсем недавно мы опирались на ежемесячные отчеты. Теперь мы находимся в ситуации, когда в режиме реального времени нам доступно данных больше, чем когда-либо. Это означает, что мы способны ускоряться», — говорит Милн.

Сделать данные жизнеспособными: коммуникация и визуализация

Финансовая служба играет ключевую роль на этапе преобразования идей, полученных в результате анализа, в реальный коммерческий результат. Однако для этого необходимы коммуникационные навыки и умение преподнести информацию. Это представляет определенные проблемы: 63 % респондентов отметили, что им сложно быть убедительными при передаче своего видения результатов данных.

Ключевым моментом является способность умело обращаться с данными и доходчиво представлять их аудитории. Начать можно с навыка

проникновения в суть и создания перекрестных ссылок уже существующих данных, таких как показатели выручки или расходов в отчете о прибылях и убытках. «Финансистам традиционно отведена роль анализа данных, которые уже заранее определены и выведены, но в современном мире этого уже не достаточно», — подчеркивает Мартин Фицпатрик, вице-президент компании QlikTech по финансам, Азиатско-Тихоокеанский и Австралийский регион (QlikTech — разработчик программной платформы интеллектуального анализа данных QlikView).

Специалисты финансовой сферы также проявляют все больше интереса к возможности представления результатов своей работы с использованием методов визуализации данных. В настоящее время существует целый ряд техник, которые позволяют представлять данные более наглядным способом, и это реальная польза при работе с массивными базовыми наборами данных. Уверенно применяя такие инструменты, финансисты смогут более результативно участвовать в процессе трансформации идей, полученных в результате анализа, в коммерческий результат.

Визуализация может быть достаточно простой, например графическое изображение ключевых бизнес-показателей на цифровой карте вместо скучных длинных списков. Таким образом, реципиент получает моментальную и мощную визуальную отсылку к проблемным точкам. Как отмечает специалист по повышению эффективности бизнеса Бернард Марр, «финансисты должны понимать, что большинство людей ненавидят цифры — наша коммуникация всегда строилась на историях и картинках».

Поиск лучших способов представления и визуализации данных имеет большое значение для любого бизнеса, но особенно полезным он будет для небольших организаций, многие из которых не могут позволить себе содержать специальные ресурсы, например штатных бизнес-аналитиков, целенаправленно занимающихся оценкой комплекса данных. Отвечая на вопрос, на что более всего полагается организация при принятии

стратегических решений — на интуицию и опыт или на данные и анализ, — 65 % респондентов из МСП ответили, что на первое. В крупных организациях на интуицию и опыт полагаются 47 %, что тоже немало.

Кейс. Инструменты GOOGLE для МСП

Анализируя большие объемы корпоративных данных, компания Google значительно повысила свой доход, и не в последнюю очередь за счет определения для себя новых прибыльных ниш, таких как МСП. Прежде в центре внимания Google были крупные рекламодатели с большими рекламными бюджетами. Но, по словам Рэя Мерфи, руководителя службы продаж аналитических и системных решений для глобального привлечения МСП в Google, они уже прекрасно разбирались в том, каковы движущие факторы для крупного бизнеса и гораздо меньше понимали рекламодателей из сектора МСП.

Благодаря аналитике в Google смогли ответить на целый ряд вопросов в отношении таких предприятий. Например, была поставлена задача оценить, что именно побуждает рекламодателей из МСП вкладывать больше средств в рекламу. Успешное постижение этого дало компании существенный годовой рост, к тому же это направление деятельности стало одним из наиболее быстро развивающихся в компании. Анализ данных также позволил оптимизировать внутреннюю эффективность: например, здесь полагали, что ключевым фактором, влияющим на оборот, генерируемый службами продаж, является трудовой стаж их сотрудников, но фактически оказалось, что более молодое поколение делало больше продаж, и это позволило компании провести оптимизацию кадрового состава.

Эта ситуация показывает, как финансовая служба может извлечь коммерческий эффект из данных. «Во многих компаниях специалисты по управленческому учету уже приносят значительную ценность, добавляя ценное содержание в такие сферы, как отчетность и бизнес-планирование, — объясняет Мерфи. — К настоящему времени финансовые специалисты уже

выросли до уровня, когда они помогают организации определять области возможного высокого роста. Теперь они уделяют больше времени представлению стратегического видения, решению проблем и направлению специалистов по продажам в нужную сторону».

Будущее специалистов по управленческому учету

Большие данные поднимают вопрос о роли финансовой службы в будущем. Возможно, бухгалтеры сойдут с дистанции как специалисты, занимающиеся формированием финансовой отчетности по прошлым результатам. Но, как вариант, они могут воспользоваться шансом, чтобы стать экспертами по большим данным — источником информации для принятия решений, и таким образом изменить саму модель бизнеса.

Финансовой службе предстоит выйти за рамки традиционных показателей бизнеса и увидеть в больших данных коммерческий потенциал. Помимо базовых финансовых и корпоративных данных, с которыми финансисты уже работают, им необходимо заняться нетрадиционными источниками данных, поскольку это принесет более глубокое понимание эффективности коммерческой деятельности организации.

Специалисты по управленческому учету могут извлечь большую пользу из ситуации, когда аналитика больших данных стала занимать важное место в бизнесе, и при этом внести в ее развитие свой вклад. Они уже знают, как работать с данными, как функционирует бизнес изнутри, они располагают всеми возможностями, чтобы помочь преобразовать идеи, полученные в результате анализа, в осязаемый коммерческий результат. Участие специалистов по управленческому учету в разработке продукта повышает доверие к конечному продукту. Успеха добьются те из них, кто адаптируется к работе с большими данными и сделает шаг к активному деловому партнерству с остальными структурами в своих организациях.

Это не означает, что финансистам предписано стать специалистами по работе с техническими данными. Напротив, специалист по управленческому

учету сможет быть более полезен в сотрудничестве со специалистом по обработке данных в целях наделения новых данных коммерческим смыслом. Дэвид Бёрджесс и Густаво Пенас из департамента финансовых операций компании Shell видят это так:

«Специалисты по управленческому учету должны знать, какие сегменты данных важны и какие идеи можно из них извлечь. Им необязательно глубоко осмысливать структуры баз данных или самостоятельно выполнять аналитику. Они просто должны разбираться в ее результатах и понимать, какую ценность эти результаты могут иметь для бизнеса».

Для финансовых специалистов переориентация на бизнес, в основе которого лежит работа с данными, имеет далеко идущие последствия. Компаниям необходимо, чтобы их финансовые службы взяли на себя ключевую роль и работали в одной упряжке со специалистами по анализу с целью извлечения из корпоративных данных содержательных коммерческих идей и, главное, чтобы финансисты помогали всей управленческой команде интерпретировать данные и направлять их потенциал на принятие более взвешенных стратегических решений.

Это неизбежно будет наталкивать финансистов на новые, незнакомые им области, а также потребует от них новых навыков и нового образа мышления. Тот, кто успешно преодолевает переход к изменившимся условиям, получит возможность выйти на первые позиции в области, которая в ближайшие годы будет становиться все более важным источником конкурентных преимуществ для компаний самых разных видов и размеров.

Финансовый персонал должен разбираться в результатах аналитики и понимать, какую ценность эти результаты могут иметь для бизнеса

Дэвид Бёрджесс и Густаво Пенас, Shell

Пять шагов к созданию бизнеса, ориентированного на данные

Шаг 1. Убедитесь, что понимаете, какие новые данные имеют отношение к вашей бизнес-модели и конкурентной позиции

Составьте бизнес-модель и обозначьте нематериальные активы вашего бизнеса. В частности, определите сегменты основных источников дохода (по клиентам, каналам, продукции), а также расходы на каждый сегмент (например, на логистику, собственно процесс, рекламу и т. д.).

Идентифицируйте данные, которые необходимо описать, и осмыслите определяющие факторы источников доходов и расходов.

Обдумайте то, что вам нужно понять еще лучше для повышения эффективности бизнеса в целом.

Шаг 2. Оцените, какие начинания в области данных уже реализуются в вашем бизнесе

Проверьте, какие платформы по обработке данных и инициативы уже работают в организации, какие данные уже собраны и (или) анализируются.

Оцените скорость и уровень, с которыми вы можете выполнять прогноз на основе движущих факторов или расчет параметров эффективности бизнеса. Это вполне возможно при наличии лишь традиционных корпоративных данных, безо всяких средств углубленной аналитики.

Изучите, какие внешние источники данных потенциально доступны для рассмотрения.

Шаг 3. Выявите возможности быстрого достижения результатов или незначительной коррекции концептуальных проектов

Соберите команду увлеченных квалифицированных специалистов разных направлений (ИТ, анализ, финансы, операции), которую поддержит специалист высокого класса по коммерческим вопросам.

Работая с небольшим сегментом доступных данных, наглядно покажите, как извлекать идеи и какую ценность это может иметь для бизнеса.

Шаг 4. Реализуйте настоящий проект обработки данных с целью формирования соответствующей стратегии

Приступите к выполнению полномасштабного проекта по сбору и анализу данных, примените полученные идеи.

Определите совокупность применяемых методов, навыки и структуру, необходимые для успешной реализации стратегии.

Составьте экономическое обоснование.

Шаг 5. Используйте эту инициативу как фундамент развития культуры данных

Убедитесь в том, что данные рассматриваются как актив бизнеса. Чтобы обеспечить хорошее качество данных внутри организации, необходима совместная, комплексная работа всех подразделений, а также приверженность общему делу всех работников компании.

Ставьте под сомнение все, что «само собой разумеется», нормой должно стать доказывание любого мнения. Помогайте проводить комплексную проверку для подтверждения всех связанных требований.

Поощряйте инновационные подходы применительно к данным, например, испытание новых источников данных для изучения альтернативных идей.

Спокойно относитесь к неудачам: если выясняется, что ранее принятая точка зрения неверна, убедитесь, что у тот, кто эмоционально отстаивал эту позицию, не оказывает сопротивления новой идее.

Помните, что данные — это зачастую чувствительный и ценный актив, и в этой связи важно соблюдать конфиденциальность и следовать самым высоким стандартам деловой этики и управления.

ЛЕКЦИЯ 5. СИСТЕМА И HADOOP MAPREDUCE

Hadoop – это свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Эта основополагающая технология хранения и обработки больших данных ([Big Data](#)) является проектом верхнего уровня фонда Apache Software Foundation.

ИЗ ЧЕГО СОСТОИТ HADOOP: КОНЦЕПТУАЛЬНАЯ АРХИТЕКТУРА

Изначально проект разработан на Java в рамках вычислительной парадигмы [MapReduce](#), когда приложение разделяется на большое количество одинаковых элементарных заданий, которые выполняются на распределённых компьютерах (узлах) кластера и сводятся в единый результат [1].

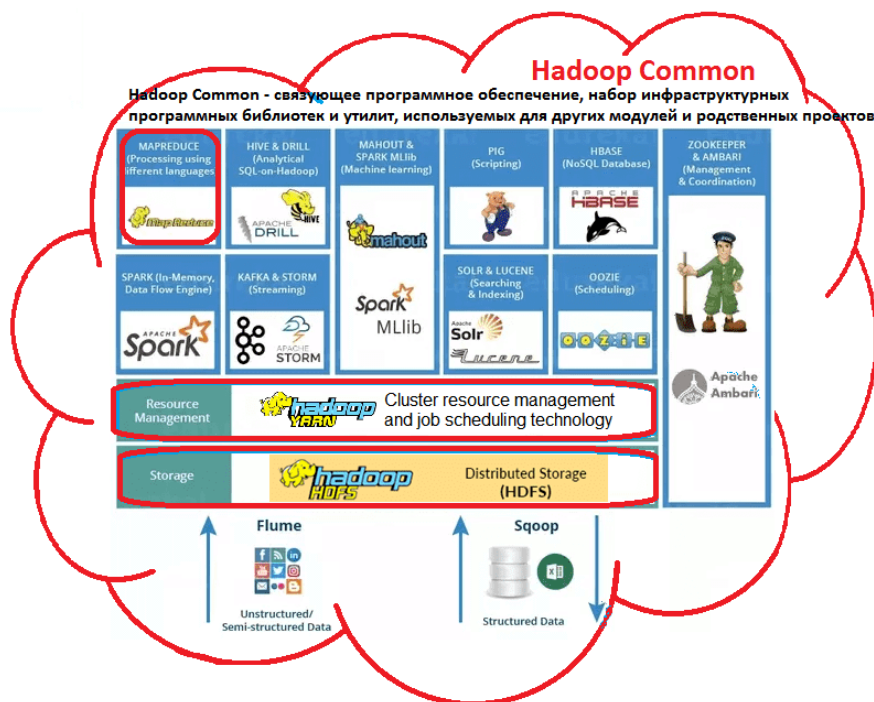
Проект состоит из основных 4-х модулей:

Hadoop Common – набор инфраструктурных программных библиотек и утилит, которые используются в других решениях и родственных проектах, в частности, для управления распределёнными файлами и создания необходимой инфраструктуры [1];

[HDFS](#) – распределённая файловая система, Hadoop Distributed File System – технология хранения файлов на различных серверах данных (узлах, DataNodes), адреса которых находятся на специальном сервере имен (мастере, NameNode) [2]. За счет дублирования (репликации) информационных блоков, HDFS обеспечивает надежное хранение файлов больших размеров, поблочное распределённых между узлами вычислительного кластера [1];

[YARN](#) – система планирования заданий и управления кластером (Yet Another Resource Negotiator), которую также называют [MapReduce 2.0 \(MRv2\)](#) – набор системных программ (демонов), обеспечивающих совместное использование, масштабирование и надежность работы распределённых приложений [3]. Фактически, YARN является интерфейсом между аппаратными ресурсами кластера и приложениями, использующих его мощности для вычислений и обработки данных [1];

Hadoop MapReduce – платформа программирования и выполнения распределённых MapReduce-вычислений, с использованием большого количества компьютеров (узлов, nodes), образующих кластер.



Архитектурная концепция экосистемы Hadoop

Сегодня вокруг Hadoop существует целая экосистема связанных проектов и технологий, которые используются для интеллектуального анализа больших данных ([Data Mining](#)), в том числе с помощью машинного обучения ([Machine Learning](#)) [2].

КАК ПОЯВИЛСЯ ХАДУП: ИСТОРИЯ РАЗРАБОТКИ И РАЗВИТИЯ

Технология хадуп появилась почти 15 лет назад и постоянно развивается. Далее показаны основные вехи ее истории:

2005 – публикация сотрудников Google Джеффри Дина и Санжая Гемавата о вычислительной концепции MapReduce сподвигла Дуга Каттинга на инициацию проекта. Разработку в режиме частичной занятости вели Дуг Каттинг и Майк Кафарелла, чтобы построить программную инфраструктуру распределённых вычислений для свободной программной поисковой машины на Java. Свое название проект получил в честь игрушечного слонёнка ребёнка

основателя [1]. Именно поэтому хадуп неформально называют “железный слон” и изображают его в виде этого животного.

2006 – корпорация Yahoo пригласила Каттинга возглавить специально выделенную команду разработки инфраструктуры распределённых вычислений, благодаря чему Hadoop выделился в отдельный проект [1].

2008 – Yahoo запустила кластерную поисковую машину на 10 тысяч процессорных ядер под управлением Hadoop, который становится проектом верхнего уровня системы проектов Apache Software Foundation. Достигнут мировой рекорд производительности в сортировке данных: за 209 секунд кластер из 910 узлов обработал 1 Тбайт информации. После этого технологию внедряют Last.fm, Facebook, The New York Times, облачные сервисы Amazon EC2.

2010 – корпорация Google предоставила Apache Software Foundation права на использование технологии MapReduce. Hadoop позиционируется как ключевая технология обработки и хранения больших данных (Big Data). Начала формироваться Hadoop-экосистема: возникли продукты [Avro](#), [HBase](#), [Hive](#), Pig, [Zookeeper](#), облегчающие операции управления данными и распределёнными приложениями, а также анализ информации [1].

2011 – получение ежегодной инновационной награды медиагруппы Guardian за универсальный подход к хранению и обработке распределённых данных («швейцарский нож XXI века») [1].

2013 – появление модуля YARN в релизе Hadoop 2.0 значительно расширяет парадигму MapReduce, повышая надёжность и масштабируемость распределённых систем [3].

ГДЕ И ЗАЧЕМ ИСПОЛЬЗУЕТСЯ HADOOP

Выделяют несколько областей применения технологии [4]:

быстрая обработка графических данных, например, газета New York Times с помощью хадуп и Web-сервиса Amazon Elastic Compute Cloud (EC2) всего за 36 часов преобразовала 4 терабайта изображений (TIFF-картинки размером в 405 КБ, SGML-статьи размером в 3.3 МБ и XML-файлы размером в 405 КБ) в PNG-формат размером по 800 КБ.



технологии Big Data

Все подробности о распределенной обработке данных, администрировании и применении Hadoop для проектов Big Data и Machine Learning на наших компьютерных курсах обучения разных групп пользователей, от «чайников» до профессионалов – хадуп для инженеров, администраторов и аналитиков больших данных в Москве:

HDDE: Hadoop для инженеров данных

ЛЕКЦИЯ 5 ИНТЕГРАЦИЯ ДАННЫХ И БЕЗОПАСНОСТЬ

Сегодня интеграция является отправной точкой проекта. Для локального объекта это интеграция подсистем безопасности (видеонаблюдение, СКУД, ОПС/ОТС, защита периметра и т.д.) в общую информационную среду. Для распределенных объектов – интеграция комплексов систем на верхнем уровне, на уровне единого центра мониторинга. Например, системы безопасности нескольких аэропортов или НПЗ с центром мониторинга в управляющей компании.

Для многих крупных предприятий полезность такого решения не вызывает сомнений. Объекты географически удалены друг от друга, условия и обстановка на каждом из них определяются множеством факторов, но при этом задача одна – обеспечить прозрачность, управляемость и высокий уровень защищенности всех предприятий. Как этого добиться? Путем интеграции систем локальных объектов в единое комплексное решение. Данные идеи заложены и в ряд отраслевых законодательных требований к обеспечению безопасности предприятий.

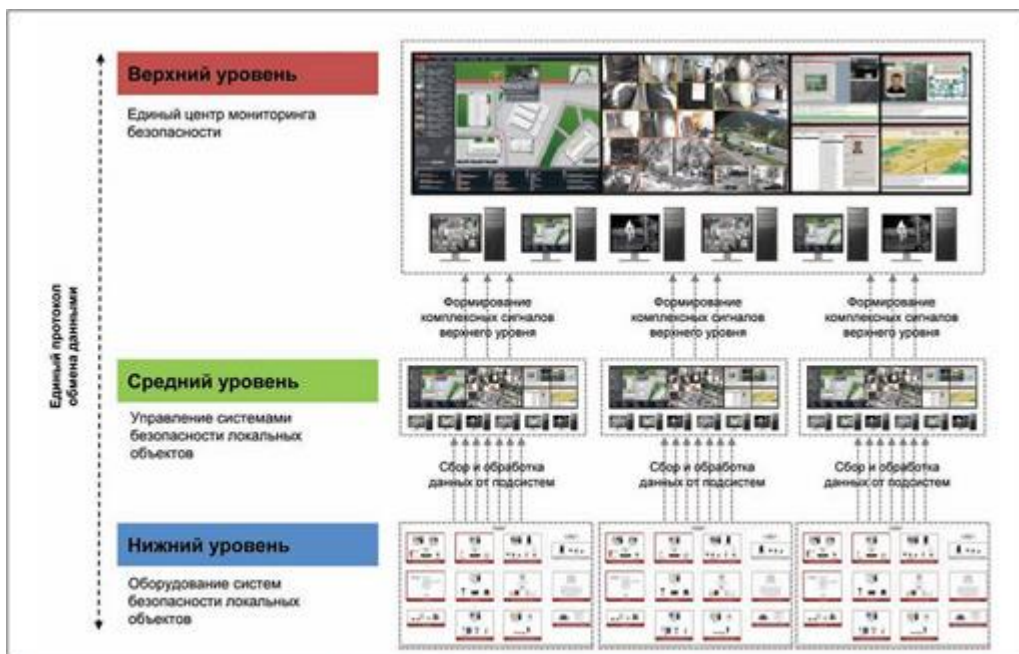


Схема комплекса систем безопасности с единым центром мониторинга

Нередко эту задачу пытаются решить посредством установки большого количества специального оборудования. Однако даже самые современные

технологии оказываются неэффективными и слишком затратными, если применять их бессистемно. В действительности эффективность конечного продукта и перспектива его развития зависят от трех простых вопросов:

1. Какую стратегию интеграции выбрать?
2. Какой должна быть архитектура комплекса?
3. Как построить работу с информацией?

В большинстве случаев на различных объектах установлены системы различных производителей, управляемые разным программным обеспечением. Это самая распространенная ситуация. Проекты выполнялись в разное время, часто независимо друг от друга. Соответственно и техническая политика в каждом проекте своя. Поэтому вариантов интеграции – два:

1. либо централизованно переходить на продукты и решения одного производителя;
2. либо проектировать гетерогенный комплекс, то есть определять единый стандарт информационного обмена, который должны будут поддерживать все интегрируемые системы.

Решение на одном бренде

Первый вариант очень удобен для любого производителя и разработчика: есть стандарт – определенное решение или продукт, которое внедряется на каждом предприятии. Сложности возникают у заказчика. Во-первых, что делать с теми системами, которые уже работают на локальных объектах и дают хороший эффект на своем уровне? Во-вторых, сколько будет стоить такой масштабный переход на продукт одного производителя? В-третьих, не поставит ли это в зависимость от конкретного производителя или разработчика? Ответы будут не в пользу такого подхода.

Гетерогенный комплекс

Второй вариант – построение гетерогенного комплекса с единым стандартом обмена данными – предполагает более приемлемую для заказчика

стратегию. Он не перечеркивает существующие наработки, менее требователен с точки зрения затрат и оставляет пользователю свободу выбора – как применяемых технических решений, так и подрядчиков.

Стандарт определяет протокол, который должны поддерживать системы всех локальных объектов, а также передаваемые данные и правила их передачи. Протокол может быть задан заказчиком или взят из существующих на рынке (таких протоколов сейчас достаточно). Большинство "дружелюбных" производителей и разработчиков идут на разработку дополнительных модулей интеграции своих систем в соответствии с требованиями принятого единого протокола.

Такой вариант имеет и еще одно преимущество. Применение различных технологий, которые интегрируются на основе общего протокола, открывает заказчику доступ к большей функциональности и гибкости для решения различных специфических задач на разных объектах, что само по себе гораздо удобнее, чем работа в рамках одного продукта (см. схему).

Мы ведем речь о группе объектов, каждый из которых является самостоятельным предприятием в части основных производственных процессов, но подчиненным в части управления ключевыми политиками и показателями. В данном случае архитектура комплекса систем безопасности должна быть многоуровневой и предполагать распределение задач и информационных потоков между уровнями.

Количество уровней определяется масштабом комплекса. Но общая логика такова:

- нижний уровень – уровень конкретных подсистем;
- средний уровень – уровень управления комплексами локальных объектов (ЦПУ или диспетчерские центры локальных объектов);

- верхний уровень – уровень единого центра мониторинга. Верхний уровень не привязан к нижним и обладает своей логикой и сервисной моделью, необходимой для обработки и демонстрации поступающих данных.

Пристальное внимание к нюансам архитектуры позволит избежать многих проблем, связанных с взаимозависимостью элементов системы друг от друга, алгоритмами обработки данных, распределением информационных потоков и т.д.

Как работать с информацией?

Даже небольшая локальная система безопасности генерирует огромный поток информации. В комплексе, где таких систем несколько, необходима грамотная политика распределения и обработки данных.

Во-первых, это требования к уровням обработки информации в соответствии с иерархией объектов (например, центральный офис – территориальное отделение – локальный объект).

Во-вторых, это четко ограниченный объем данных для каждого уровня и правила обмена данными между уровнями системы.



В-третьих, это алгоритмы обработки данных, которые различаются в зависимости от уровня. Если на локальном объекте фиксируется каждое событие и параметр, то для верхнего уровня (центр мониторинга) эта информация будет излишней. Там необходимо оценивать ситуацию целиком, управлять глобальными параметрами комплекса, поэтому особое значение приобретает не полнота, а информативность данных. Становятся

необходимыми комплексные индикаторы (например, состояние защищенности, уровни тревог и т.д.), которые формируются в результате сочетания различных параметров локальных объектов.

Таким образом, на каждом уровне комплекса по своим правилам обрабатывается свой объем данных, что в конечном итоге позволяет добиться соответствующего качества информации и своевременного ее получения сотрудниками и руководством служб безопасности.

Несмотря на то что были затронуты далеко не все нюансы построения крупных распределенных систем, мы обозначили ключевые принципы, без которых получить хороший результат вряд ли возможно. Это:

- принцип гетерогенности – интеграция разнородных систем на основе единого стандарта обмена данными;
- принцип многоуровневости – позволяет грамотно структурировать функциональность и задачи внутри комплекса;
- принцип распределенности – необходим для организации системы обработки, передачи и визуализации данных.

Результатом грамотного проектирования комплекса в соответствии с обозначенными принципами будет надежное и эффективное решение, которое обеспечит пользователей всех уровней необходимым инструментарием и информацией.

Данный подход применим в любой отрасли, и в настоящее время реализуется нашей компанией на нескольких объектах гражданской авиации. Проектирование такого комплекса требует в том числе разработки подробной модели защиты всех объектов и подбора соответствующих технологий и платформ. Но это уже тема для другого обсуждения.

ЛЕКЦИЯ 6. БАЗА ДАННЫХ NOSQL И БАЗА ДАННЫХ SQL ДЛЯ РЕШЕНИЯ ФИНАНСОВЫХ ПРОБЛЕМ

Что такое реляционные и нереляционные базы данных

Реляционная база данных (SQL) — база, где данные хранятся в формате таблиц, они строго структурированы и связаны друг с другом. В таблице есть строки и столбцы, каждая строка представляет отдельную запись, а столбец — поле с назначенным ей типом данных. В каждой ячейке информация записана по шаблону.

Нереляционная база данных (NoSQL) — хранит данные без четких связей друг с другом и четкой структуры. Вместо структурированных таблиц внутри базы находится множество разнородных документов, в том числе изображения, видео и даже публикации в социальных сетях. В отличие от реляционных БД, NoSQL базы данных не поддерживают запросы SQL.

Реляционные базы данных, или базы данных SQL

Особенности. Основная особенность — надежность и неизменяемость данных, низкий риск потери информации. При обновлении данных их целостность гарантирована, они заменяются в одной таблице.

Реляционные базы данных, в отличие от нереляционных, соответствуют ACID — это требования к транзакционным системам. Соответствие им гарантирует сохранность данных и предсказуемость работы базы данных:

Atomicity, или атомарность — ни одна транзакция не будет зафиксирована в системе частично.

Consistency, или непротиворечивость — фиксируются только допустимые результаты транзакций.

Isolation, или изолированность — на результат транзакции не влияют транзакции, проходящие параллельно ей.

Durability, или долговечность — изменения в базе данных сохраняются несмотря на сбои или действия пользователей.

При работе с такими СУБД надо учитывать, что любые изменения в объектах нужно отражать в структуре таблиц, физическая структура данных не соответствует объектной модели приложения.

Реляционные БД идеальны для работы со структурированными данными, структура которых не подвержена частым изменениям.

Так выглядит хранение данных в реляционной базе, по сути, это просто таблица:

Клиент	Средний чек	Число покупок за период
1	1000	10
2	1500	5
3	800	6

Масштабируемость. Вертикальная, то есть при росте нагрузки растет производительность сервера. Если в базу поступает большой объем данных, рано или поздно наступит порог вертикального масштабирования — сервер не сможет далее увеличивать производительность. Тогда понадобится горизонтальное масштабирование — параллельная обработка данных в кластере серверов.

В больших распределенных системах это может привести к тому, что общая производительность системы упадет, так как нужно поддерживать согласованность данных в нескольких узлах. Это не значит, что СУБД на SQL не подходят для больших проектов — они поддерживают кластеризацию, просто нужно приложить усилия, чтобы настроить систему. Либо

использовать базы данных в облаке — там можно получить уже настроенные и надежно работающие кластеры в несколько кликов.

Самые известные SQL-базы данных

MySQL — одна из самых популярных open source реляционных баз данных. Подходит небольшим и средним проектам, которым нужен недорогой и надежный инструмент работы с данными. Поддерживает множество типов таблиц, есть огромное количество плагинов и расширений, облегчающих работу с системой.

Отличается простой установкой, может быть интегрирована с другими СУБД, также интеграция с MySQL есть в любой CMS, фреймворке, языке программирования. Среди минусов — не все задачи выполняет автоматически, если что-то нужное не включено в функционал, придется потратить время на доработку, нет встроенной поддержки OLAP.

[MySQL доступна как облачный сервис](#) — в облаке не нужно тратить много времени на развертывание и конфигурацию СУБД. MySQL server стоит выбрать на старте бизнеса, чтобы тестировать гипотезы с минимальными затратами или для небольших проектов как транзакционную базу данных общего назначения.

PostgreSQL — вторая по популярности open source SQL СУБД. У нее много встроенных функций и дополнений, в том числе для масштабирования в кластер и шардинга таблиц. Подходит, если важна сохранность данных, предполагается их сложная структура. Позволяет работать со структурированными данными, но поддерживает JSON/BSON, что дает некоторую гибкость в схеме данных.

Отличается стабильностью, ее практически невозможно вывести из строя или что-то сломать в таблицах.

Из минусов — сложность конфигурации требует от пользователей некоторого опыта. Также скорость работы может падать во время проведения пакетных операций или при запросах на чтение.

[PostgreSQL также можно развернуть в облаке](#) — в отличие от MySQL, она подходит для крупных и масштабных проектов. Кроме того, ее стоит выбрать, если недопустимы ошибки в данных или есть особые требования к базе данных, например поддержка геоданных. Различные расширения PostgreSQL позволяют реализовать многие специализированные запросы.

Подробнее о разнице между PostgreSQL и MySQL мы [писали в отдельной статье](#).

Нереляционные базы данных, или базы данных NoSQL

Особенности. В отличие от реляционных, в нереляционных базах данных схема данных является динамической и может меняться в любой момент времени. К данным сложнее получить доступ, то есть найти внутри базы что-то нужное — с таблицей это просто, достаточно знать координаты ячейки. Зато такие СУБД отличаются производительностью и скоростью. Физические объекты в NoSQL обычно можно хранить прямо в том виде, в котором с ними потом работает приложение.

Базы данных NoSQL подходят для хранения больших объемов неструктурированной информации, а также хороши для быстрой разработки и тестирования гипотез.

В них можно хранить данные любого типа и добавлять новые в процессе работы.

Масштабируемость. NoSQL базы имеют распределенную архитектуру, поэтому хорошо масштабируются горизонтально и отличаются высокой производительностью. Технологии NoSQL могут автоматически распределять данные по разным серверам. Это повышает скорость чтения данных в распределенной среде.

Четыре вида нереляционных баз данных

Документоориентированные базы данных — данные хранятся в коллекциях документов, обычно с использованием форматов JSON, XML или BSON. Одна запись может содержать столько данных, сколько нужно, в

любом типе данных (или типах) — ограничений нет. Внутри одного документа есть внутренняя структура, однако, она может отличаться от одного документа к другому. Также документы можно вкладывать друг в друга.

То есть вместо столбцов и строк мы описываем все данные в одном документе. Если нам нужно было бы добавить новые данные в таблицу реляционной базы данных, пришлось бы изменять ее схему данных. В случае с документами нужно только добавить в них дополнительные пары ключ-значение.

Пример такой базы данных: MongoDB.

Вот так будет выглядеть хранение данных в отдельных документах вместо таблицы со столбцами и строками:

ЛЕКЦИЯ 7. АРХИТЕКТУРЫ СИСТЕМЫ ОБРАБОТКИ ДАННЫХ

Архитектура файл-сервер

В архитектуре файл-сервер базы данных хранятся на сервере, клиент обращается к серверу с файловыми командами, а механизм управления всеми информационными ресурсами находится на компьютере клиента (рис. 9.1).

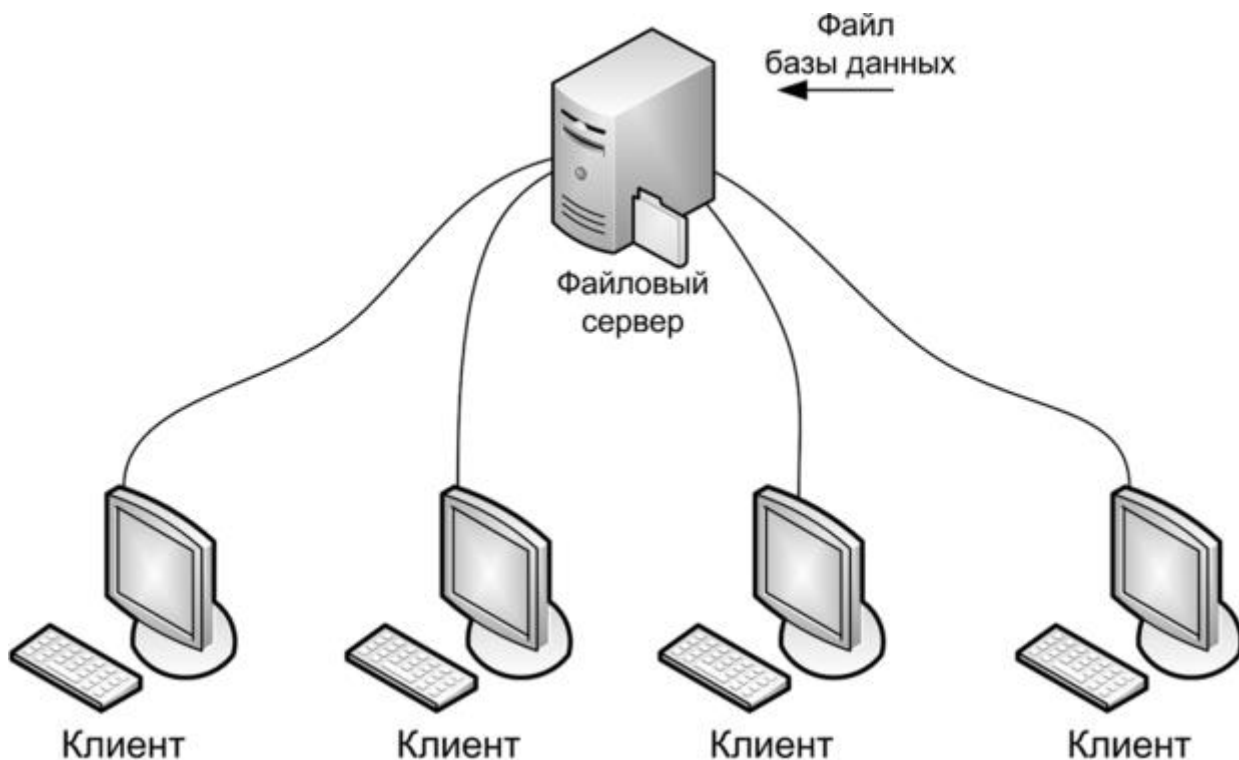


Рис. 9.1. Архитектура «файл-сервер»

Файл-серверные базы данных могут быть доступны многим клиентам через сеть. Для каждого клиента во время работы создается локальная копия данных, с которой он манипулирует. При этом возникают проблемы, связанные с возможным одновременным доступом нескольких пользователей к одной и той же информации. Эти проблемы решаются разработчиками приложений баз данных (каждый раз при обращении к данным проверяется их доступность).

Архитектура файл-сервер обладает значительными недостатками. Одним из них является непроизводительная загрузка сети. При каждом запросе

клиента данные в его локальной копии полностью обновляются из базы данных на сервере. Даже если запрос относится всего к одной записи, обновляются все записи базы данных. Если записей в базе данных много, то даже при небольшом числе клиентов сеть будет загружена очень основательно, что серьезно скажется на скорости выполнения запросов.

В результате циркуляции в сети больших объемов избыточной информации резко возрастает нагрузка на сеть, что приводит к соответствующему снижению ее быстродействия и производительности информационной системы в целом. Значительный сетевой трафик особенно сказывается при организации удаленного доступа к базам данных на файл-сервере через низкоскоростные каналы связи.

Другой недостаток связан с тем, что забота о целостности данных при такой организации работы возлагается на программы клиентов.

Одним из традиционных средств, на основе которых создаются файл-серверные системы, являются локальные СУБД. При их использовании задача обеспечения целостности данных возлагается на клиентские приложения, что приводит к их усложнению. Если они недостаточно тщательно продуманы, в базу данных легко занести ошибки, которые могут отразиться на всех пользователях.

Кроме того, в файл-серверной архитектуре изменения, сделанные в базе данных одним пользователем, не видны другим пользователям. Пока один пользователь редактирует какую-либо запись, она заблокирована для других клиентов. Возникает необходимость синхронизации работы отдельных пользователей, связанная с блокировкой записей.

И еще один недостаток — управление базой данных осуществляется с разных компьютеров, поэтому в значительной степени затруднена организация контроля доступа, соблюдения конфиденциальности, что также усложняет поддержку целостности базы данных.

Тем не менее файл-серверная архитектура привлекает своей простотой и доступностью. Поэтому файл-серверные информационные системы до сих пор представляют интерес для малых рабочих групп. Такие системы имеют достаточно обширную область применения. Прежде всего это государственные (муниципальные) учреждения, сфера образования, сфера обслуживания, малый и средний бизнес.

Специфика решаемых там задач заключается в том, что объемы данных не являются чрезвычайно большими, обновления производятся не слишком часто, организация территориально обычно расположена в одном небольшом здании, количество пользователей колеблется от одного до десяти человек. В подобных условиях использование настольных СУБД для управления информационными системами вполне оправданно.

Одними из первых СУБД были так называемые dBase-совместимые программные системы, разработанные разными фирмами: dBase III — PLUS (фирма Achton-Tate), Clipper (фирма Nantucket Inc.), FoxPro (фирма Fox Software), FoxBase+ (фирма Fox Software), Visual

FoxPro (фирма Microsoft). Одно время достаточно широко использовалась СУБД PARADOX (фирма Borland International). В последние годы очень широкое распространение получила система управления базами данных Microsoft Access, которая входит в целый ряд версий пакета Microsoft Office (фирма Microsoft).

Для крупных организаций ситуация принципиально меняется. Там использование файл-серверных технологий является неприемлемым по описанным выше причинам. Поэтому используются более дорогостоящие серверные СУБД.

ЛЕКЦИЯ 8. ПЛАТФОРМА БОЛЬШИХ ДАННЫХ ПРОТЕСТИРОВАЛА НОВУЮ ТЕХНОЛОГИЮ ЗАЩИТЫ BIG DATA

Специалисты совместного предприятия ВТБ и Ростелекома — «Платформы больших данных» — протестировали новую технологию совместных конфиденциальных вычислений (Multi-partial computations, MPC).

Технология позволяет компаниям совместно работать с массивами данных для улучшения качества сервисов, повышения эффективности бизнеса и решения других задач.

Особенность технологии MPC в отсутствии обмена исходными данными между компаниями: это дает возможность пользователям сервиса реализовать подход Data Fusion — он заключается в объединении данных, слиянии или переносе алгоритмов из одной области машинного обучения в другую, а также в синергии процессов в машинном обучении.

Технология Multi-partial computations была протестирована при участии компании-разработчика сервисов фильтрации и анализа трафика Ubiс. Совместные конфиденциальные вычисления позволяют создавать модели машинного обучения, которые построены на объединенных данных различных компаний. Каждая из сторон разделяет информацию и передает только тот объём защищённых данных, на которых можно строить модели. При этом информацию для получения первичных данных расшифровать невозможно.

Такая технология позволяет разрабатывать сервисы на базе различных источников и поможет обогатить компетенции и экспертизу различных направлений, от розничной торговли до страхового бизнеса.

«Безопасность данных клиентов — приоритет «Платформы больших данных», новая технология позволяет вывести работу с Big Data и

защищенность данных на новый уровень. Благодаря МРС, мы можем работать с объединенными массивами информации для совершенствования продуктов, сервисов и бизнес-процессов. Объединение уникальных массивов обезличенных данных, технологий и компетенций, а также применение подхода Data Fusion, позволит «Платформе больших данных» достичь лидерства на рынке Big Data», — отметил старший вице-президент, руководитель департамента анализа данных и моделирования ВТБ **Максим Коновалихин**.

«Объединение данных разных компаний дает существенный синергетический эффект. Но на пути интеграции данных множество препятствий — процессуальных, технологических и нормативных. Технология МРС позволяет оптимально и безопасно преодолевать эти ограничения. Она увеличивает гибкость в развитии продуктов, формирует новый масштаб анализируемой среды, повышает точность вычислений и, соответственно, на выходе дает более качественный результат для заказчика. Все то, что ранее было доступно только при высоких ресурсных затратах, а иногда и вовсе закрыто, теперь будет внедряться в бизнес-процессы максимально эффективно и быстро», — сказал **Александр Айвазов**, вице-президент по развитию бизнеса «Ростелекома».

«Наша ключевая задача – помочь самым разным компаниям внедрить у себя передовые технологии работы с большими данными. Технология МРС позволяет просто и безопасно обогатить внутреннюю экспертизу самыми актуальными рыночными тенденциями и при этом можно будет сохранить свои сведения закрытыми для остальных. Для многих компаний, которые ранее только думали о том, как построить платформы обогащения данных, это упиралось в огромные инвестиции и прогнозирование дальнейшего использования этой инфраструктуры. Решение на базе технологии МРС поможет решить эти вопросы», – считает **Алексей Каштанов**, генеральный директор «Платформы больших данных».

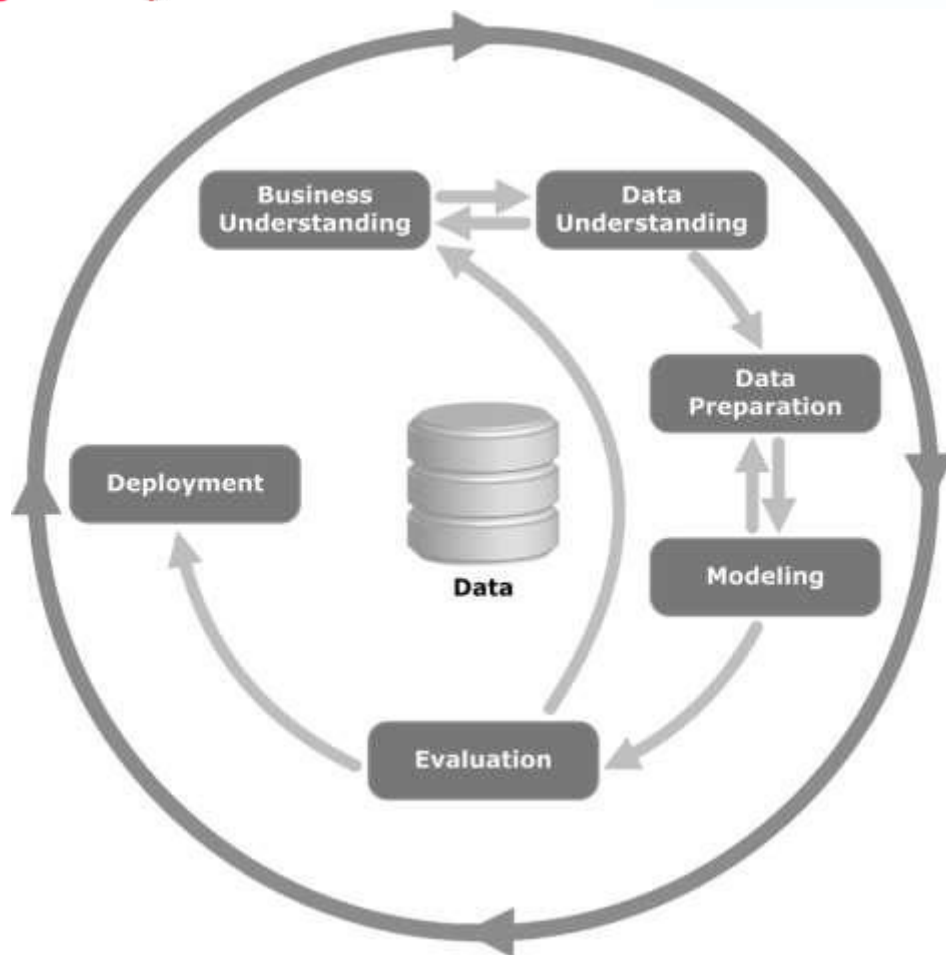
ЛЕКЦИЯ 9. ЖИЗНЕННЫЙ ЦИКЛ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Чтобы обеспечить основу для организации работы, необходимой для организации, и предоставить четкое представление о больших данных, полезно рассматривать ее как цикл с различными этапами. Это ни в коем случае не линейно, то есть все этапы связаны друг с другом. Этот цикл имеет поверхностное сходство с более традиционным циклом интеллектуального анализа данных, как описано в методологии CRISP .

Методология CRISP-DM

Методология CRISP-DM , обозначающая межотраслевой стандартный процесс интеллектуального анализа данных, представляет собой цикл, описывающий часто используемые подходы, используемые экспертами по интеллектуальному анализу данных для решения проблем в традиционном интеллектуальном анализе данных BI. Он все еще используется в традиционных командах интеллектуального анализа данных.

Посмотрите на следующую иллюстрацию. Он показывает основные этапы цикла, описанные методологией CRISP-DM, и их взаимосвязь.



CRISP-DM был задуман в 1996 году, а в следующем году он был реализован как проект Европейского Союза в рамках инициативы финансирования ESPRIT. Проект возглавляли пять компаний: SPSS, Teradata, Daimler AG, NCR Corporation и OHRA (страховая компания). Проект был наконец включен в SPSS. Методология чрезвычайно детально ориентирована на то, как должен быть определен проект интеллектуального анализа данных.

Давайте теперь узнаем немного больше о каждом из этапов жизненного цикла CRISP-DM —

Понимание бизнеса — на этом начальном этапе основное внимание уделяется пониманию целей и требований проекта с точки зрения бизнеса, а затем преобразованию этих знаний в определение проблемы интеллектуального анализа данных. Предварительный план предназначен для достижения целей. Можно использовать модель принятия решений, особенно

модель, построенную с использованием модели принятия решений и стандарта обозначений.

Понимание данных. Фаза понимания данных начинается с первоначального сбора данных и продолжается действиями, чтобы ознакомиться с данными, выявить проблемы с качеством данных, обнаружить первое понимание данных или обнаружить интересные подмножества, чтобы сформировать гипотезы для скрытого Информация.

Подготовка данных. Этап подготовки данных охватывает все действия по построению окончательного набора данных (данных, которые будут вводиться в инструмент (ы) моделирования) из исходных необработанных данных. Задачи подготовки данных, вероятно, будут выполняться несколько раз, а не в каком-либо предписанном порядке. Задачи включают выбор таблиц, записей и атрибутов, а также преобразование и очистку данных для инструментов моделирования.

Моделирование — на этом этапе выбираются и применяются различные методы моделирования, а их параметры калибруются до оптимальных значений. Как правило, существует несколько методов для одного и того же типа проблемы интеллектуального анализа данных. Некоторые методы предъявляют особые требования к форме данных. Поэтому часто требуется вернуться к этапу подготовки данных.

Оценка — на этом этапе проекта вы создали модель (или модели), которая, по-видимому, имеет высокое качество с точки зрения анализа данных. Прежде чем приступить к окончательному развертыванию модели, важно тщательно оценить модель и рассмотреть шаги, выполненные для построения модели, чтобы убедиться, что она правильно достигает бизнес-целей.

Основная цель — определить, существует ли какая-то важная бизнес-проблема, которая не была в достаточной мере рассмотрена. В конце этого

этапа должно быть принято решение об использовании результатов интеллектуального анализа данных.

Развертывание — создание модели, как правило, не является концом проекта. Даже если целью модели является повышение знаний о данных, полученные знания необходимо будет организовать и представить так, чтобы это было полезно для клиента.

В зависимости от требований фаза развертывания может быть такой же простой, как создание отчета, или такой сложной, как реализация повторяющейся оценки данных (например, распределение сегментов) или процесса интеллектуального анализа данных.

Понимание бизнеса — на этом начальном этапе основное внимание уделяется пониманию целей и требований проекта с точки зрения бизнеса, а затем преобразованию этих знаний в определение проблемы интеллектуального анализа данных. Предварительный план предназначен для достижения целей. Можно использовать модель принятия решений, особенно модель, построенную с использованием модели принятия решений и стандарта обозначений.

Понимание данных. Фаза понимания данных начинается с первоначального сбора данных и продолжается действиями, чтобы ознакомиться с данными, выявить проблемы с качеством данных, обнаружить первое понимание данных или обнаружить интересные подмножества, чтобы сформировать гипотезы для скрытого Информация.

Подготовка данных. Этап подготовки данных охватывает все действия по построению окончательного набора данных (данных, которые будут вводиться в инструмент (ы) моделирования) из исходных необработанных данных. Задачи подготовки данных, вероятно, будут выполняться несколько раз, а не в каком-либо предписанном порядке. Задачи включают выбор таблиц, записей и атрибутов, а также преобразование и очистку данных для инструментов моделирования.

Моделирование — на этом этапе выбираются и применяются различные методы моделирования, а их параметры калибруются до оптимальных значений. Как правило, существует несколько методов для одного и того же типа проблемы интеллектуального анализа данных. Некоторые методы предъявляют особые требования к форме данных. Поэтому часто требуется вернуться к этапу подготовки данных.

Оценка — на этом этапе проекта вы создали модель (или модели), которая, по-видимому, имеет высокое качество с точки зрения анализа данных. Прежде чем приступить к окончательному развертыванию модели, важно тщательно оценить модель и рассмотреть шаги, выполненные для построения модели, чтобы убедиться, что она правильно достигает бизнес-целей.

Основная цель — определить, существует ли какая-то важная бизнес-проблема, которая не была в достаточной мере рассмотрена. В конце этого этапа должно быть принято решение об использовании результатов интеллектуального анализа данных.

Развертывание — создание модели, как правило, не является концом проекта. Даже если целью модели является повышение знаний о данных, полученные знания необходимо будет организовать и представить так, чтобы это было полезно для клиента.

В зависимости от требований фаза развертывания может быть такой же простой, как создание отчета, или такой сложной, как реализация повторяющейся оценки данных (например, распределение сегментов) или процесса интеллектуального анализа данных.

Во многих случаях именно заказчик, а не аналитик данных, будет выполнять этапы развертывания. Даже если аналитик развертывает модель, для клиента важно заранее понять действия, которые необходимо будет выполнить для фактического использования созданных моделей.

Методология SEMMA

SEMMA — это еще одна методология, разработанная SAS для моделирования интеллектуального анализа данных. Он обозначает S достаточно, E xplore, M odify, M odel и A sses. Вот краткое описание его этапов —

Выборка — процесс начинается с выборки данных, например, с выбора набора данных для моделирования. Набор данных должен быть достаточно большим, чтобы содержать достаточную информацию для извлечения, но достаточно маленьким, чтобы его можно было эффективно использовать. Этот этап также касается разделения данных.

Исследовать — Эта фаза охватывает понимание данных путем обнаружения ожидаемых и непредвиденных связей между переменными, а также отклонений с помощью визуализации данных.

Modify — фаза Modify содержит методы для выбора, создания и преобразования переменных при подготовке к моделированию данных.

Модель. На этапе модели основное внимание уделяется применению различных методов моделирования (интеллектуального анализа данных) к подготовленным переменным с целью создания моделей, которые могут обеспечить желаемый результат.

Оценка — Оценка результатов моделирования показывает надежность и полезность созданных моделей.

Выборка — процесс начинается с выборки данных, например, с выбора набора данных для моделирования. Набор данных должен быть достаточно большим, чтобы содержать достаточную информацию для извлечения, но достаточно маленьким, чтобы его можно было эффективно использовать. Этот этап также касается разделения данных.

Исследовать — Эта фаза охватывает понимание данных путем обнаружения ожидаемых и непредвиденных связей между переменными, а также отклонений с помощью визуализации данных.

Modify — фаза Modify содержит методы для выбора, создания и преобразования переменных при подготовке к моделированию данных.

Модель. На этапе модели основное внимание уделяется применению различных методов моделирования (интеллектуального анализа данных) к подготовленным переменным с целью создания моделей, которые могут обеспечить желаемый результат.

Оценка — Оценка результатов моделирования показывает надежность и полезность созданных моделей.

Основное различие между CRISM-DM и SEMMA заключается в том, что SEMMA фокусируется на аспекте моделирования, тогда как CRISP-DM придает большее значение этапам цикла, предшествующим моделированию, таким как понимание бизнес-проблемы, подлежащей решению, понимание и предварительная обработка данных, подлежащих обработке. используется в качестве входных данных, например, алгоритмы машинного обучения.

Жизненный цикл больших данных

В сегодняшнем контексте больших данных предыдущие подходы являются либо неполными, либо неоптимальными. Например, методология SEMMA полностью игнорирует сбор данных и предварительную обработку различных источников данных. Эти этапы обычно составляют большую часть работы в успешном проекте больших данных.

Большой цикл анализа данных может быть описан следующим этапом —

Определение бизнес-проблемы

Исследование

Оценка человеческих ресурсов

Получение данных

Data Munging

Хранилище данных

Исследовательский анализ данных

Подготовка данных для моделирования и оценки

моделирование

Реализация

В этом разделе мы расскажем о каждом из этих этапов жизненного цикла больших данных.

Определение бизнес-проблемы

Это обычное явление в традиционном жизненном цикле ВІ и аналитики больших данных. Обычно это нетривиальный этап проекта больших данных, чтобы определить проблему и правильно оценить, какой потенциальный выигрыш он может принести для организации. Кажется очевидным упомянуть об этом, но необходимо оценить, каковы ожидаемые выгоды и затраты по проекту.

Исследование

Проанализируйте, что другие компании сделали в такой же ситуации. Это включает в себя поиск решений, которые являются разумными для вашей компании, даже если это включает в себя адаптацию других решений к ресурсам и требованиям, которые есть у вашей компании. На этом этапе следует определить методологию будущих этапов.

Оценка человеческих ресурсов

Как только проблема будет определена, целесообразно продолжить анализ того, сможет ли текущий персонал успешно завершить проект. Традиционные команды ВІ могут быть не в состоянии предоставить оптимальное решение для всех этапов, поэтому следует рассмотреть это перед началом проекта, если есть необходимость передать часть проекта на аутсорсинг или нанять больше людей.

Получение данных

Этот раздел является ключевым в жизненном цикле больших данных; он определяет, какой тип профилей потребуется для доставки итогового продукта данных. Сбор данных является нетривиальным шагом процесса; обычно он включает сбор неструктурированных данных из разных источников. В

качестве примера можно привести написание сканера для получения отзывов с веб-сайта. Это включает в себя работу с текстом, возможно, на разных языках, обычно требующих значительного времени для завершения.

Data Munging

Когда данные извлекаются, например, из Интернета, они должны храниться в удобном для использования формате. Чтобы продолжить с примерами обзоров, давайте предположим, что данные извлекаются с разных сайтов, каждый из которых по-разному отображает данные.

Предположим, что один источник данных дает обзоры с точки зрения рейтинга в звездах, поэтому его можно считать отображением для переменной отклика $y \in \{1, 2, 3, 4, 5\}$. Другой источник данных дает обзоры с использованием системы двух стрелок, одна для голосования «за», а другая для голосования «за». Это подразумевало бы переменную ответа вида $y \in \{\text{положительный, отрицательный}\}$.

Чтобы объединить оба источника данных, необходимо принять решение, чтобы сделать эти два представления ответов эквивалентными. Это может включать преобразование первого отклика представления источника данных во вторую форму, считая одну звезду отрицательной и пять звезд положительной. Этот процесс часто требует больших временных затрат для обеспечения хорошего качества.

Хранилище данных

Как только данные обработаны, их иногда необходимо сохранить в базе данных. Технологии больших данных предлагают множество альтернатив в этом отношении. Наиболее распространенной альтернативой является использование файловой системы Hadoop для хранения, которая предоставляет пользователям ограниченную версию SQL, известную как язык запросов HIVE. Это позволяет выполнять большинство задач аналитики аналогично тому, как это делается в традиционных хранилищах данных BI, с

точки зрения пользователя. Другими вариантами хранения, которые следует учитывать, являются MongoDB, Redis и SPARK.

Эта стадия цикла связана со знаниями человеческих ресурсов с точки зрения их способности реализовывать различные архитектуры. Модифицированные версии традиционных хранилищ данных все еще используются в крупномасштабных приложениях. Например, teradata и IBM предлагают базы данных SQL, которые могут обрабатывать терабайты данных; Решения с открытым исходным кодом, такие как postgresSQL и MySQL, все еще используются для крупномасштабных приложений.

Несмотря на различия в работе различных хранилищ в фоновом режиме, на стороне клиента большинство решений предоставляют SQL API. Следовательно, хорошее понимание SQL по-прежнему является ключевым навыком для анализа больших данных.

Этот этап априори представляется наиболее важной темой, на практике это не так. Это даже не существенная стадия. Можно реализовать решение для больших данных, которое будет работать с данными в реальном времени, поэтому в этом случае нам нужно только собрать данные для разработки модели, а затем реализовать ее в режиме реального времени. Таким образом, не было бы необходимости формально хранить данные вообще.

Исследовательский анализ данных

После того, как данные были очищены и сохранены таким образом, что из них можно извлечь информацию, фаза исследования данных является обязательной. Целью этого этапа является понимание данных, обычно это делается с помощью статистических методов, а также построения графиков данных. Это хороший этап для оценки того, имеет ли определение проблемы смысл или выполнимость.

Подготовка данных для моделирования и оценки

Этот этап включает изменение предварительно очищенных данных, полученных ранее, и использование статистической предварительной

обработки для определения пропущенных значений, обнаружения выбросов, нормализации, выделения признаков и выбора признаков.

моделирование

На предыдущем этапе должно было быть подготовлено несколько наборов данных для обучения и тестирования, например, прогнозирующая модель. Этот этап включает в себя попытки различных моделей и с нетерпением ждем решения проблемы бизнеса под рукой. На практике обычно желательно, чтобы модель дала некоторое представление о бизнесе. И, наконец, выбирается лучшая модель или комбинация моделей, которая оценивает ее производительность в неактивном наборе данных.

Реализация

На данном этапе разработанный продукт данных внедряется в конвейер данных компании. Это включает в себя настройку схемы проверки во время работы продукта данных, чтобы отслеживать его производительность. Например, в случае реализации прогнозирующей модели, этот этап будет включать применение модели к новым данным и, как только ответ будет получен, оцените модель.

ЛЕКЦИЯ 10. БОЛЬШИЕ ДАННЫЕ КАК РЫНОК

По данным Data Age Report человечество сформировало порядка 51 зеттабайта информации в 2020 году. К 2025 году объем этих данных вырастет до 175 зеттабайтов ежегодно

ResearchAndMarkets: Рынок аналитики больших данных будет расти в среднем на 11,9% с 2020 по 2028 год

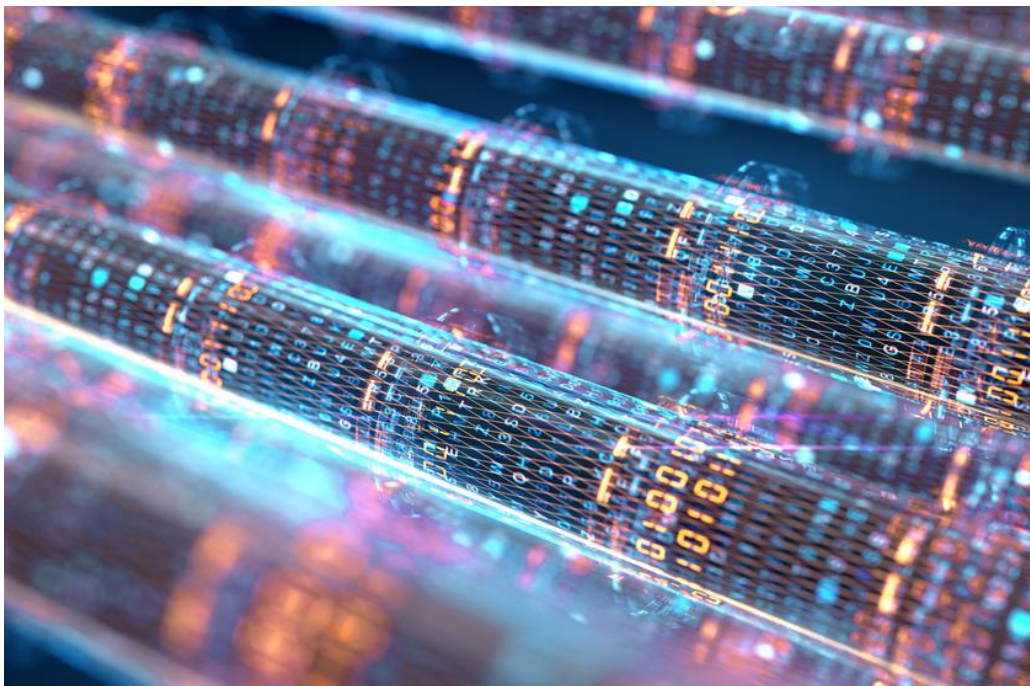
В сентябре 2020 г. агентство ResearchAndMarkets опубликовало отчет о глобальном рынке аналитики [больших данных](#). Согласно опубликованной информации, мировой рынок аналитики больших данных оценивается в \$41,85 млрд по итогам 2019 года. По прогнозам аналитиков, он вырастет до \$115,13 млрд, при средней динамике в 11,9% в течение прогнозируемого периода с 2020 по 2028 год.

Как отмечается в отчете, аналитика больших данных может быть названа сердцем цифрового мира, на основании анализа и преобразования данных в информацию, которая предоставляет ценные идеи для бизнеса. [Облачные](#) платформы являются основными для приложений анализа больших данных. При этом для проведения анализа больших данных крупные организации в основном используют гибридную облачную платформу, в то время как публичные облака преобладают среди малых и средних организаций.

Евросоюз создаёт единый рынок данных, чтобы европейские ИТ-компании стали лучше американских и китайских

В конце января 2020 года стало известно, что [Европейский союз](#) хочет создать единый рынок данных, который должен противостоять доминированию таких технологических гигантов, как [Facebook](#), [Google](#) и [Amazon](#).

Это предложение подчеркивает желание ЕС сломить господство технологических компаний США в области работы с огромными массивами данных и обойти китайских конкурентов.



Еврокомиссия разрабатывает единые правила и стандарты обмена данными между странами Евросоюза

В настоящее время очень малое число крупных технологических компаний хранит значительную часть мировых данных. Это серьезное препятствие для развития бизнеса на основе данных, для появления и роста новых компаний и введения инноваций, в том числе в Европе, но при правильном подходе перед нами могут раскрыться огромные возможности, — говорится в предложении Европейской комиссии.

В документе, который занимает 25 страниц, также подчеркивается срочность выполнения предстоящей задачи.

Конкуренты, такие как Китай и США, уже внедряют инновации и представляют свои концепции доступа к данным и их использования по всему миру, — предупреждают члены комиссии.



Они предполагают, что ЕС способен самостоятельно извлечь выгоду из огромного объема промышленных и профессиональных данных с целью продвижения технологических инноваций в Европе. Цель состоит в том, чтобы «создать единое европейское пространство данных, подлинный единый рынок данных».

Меры по достижению этой цели включают ряд новых правил, касающихся трансграничного использования данных, функциональной совместимости данных и стандартов, связанных с производством, контролем изменения климата, автомобильной промышленностью, здравоохранением, финансовыми услугами, сельским хозяйством и энергетикой.

Комиссия предлагает исключить правила, которые препятствуют обмену данными, и, возможно, ввести новые, которые помешают крупным онлайн-платформам навязывать условия для доступа к данным. [\[1\]](#)

2019

Frost & Sullivan: Рынок аналитики больших данных будет расти на 35,9% ежегодно

Аналитика [больших данных](#) (Big Data Analytics, BDA) становится одной из самых востребованных задач в современном бизнесе. По оценкам компании [Frost & Sullivan](#) в 2021 году общий объем мирового рынка аналитики больших данных увеличится по сравнению с показателем 2016 года более чем в 2,5 раза и составит \$67,2 млрд, при ежегодных темпах роста (CAGR) на уровне 35,9%. При этом крупнейшими сегментами рынка станут производственный сектор, [финансы](#), [здравоохранение](#), охрана окружающей среды (ООС) и розничная торговля, сообщили [TAdviser](#) в Frost & Sullivan 28 января 2019 года.

[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. В

производственном сегменте и других промышленных отраслях аналитики фиксируют повышенный спрос на BDA: увеличение объема инвестиций в аналитику больших данных здесь обусловлено необходимостью увеличения производительности предприятий и оптимизации ресурсов.



Развитие рынка аналитики больших данных в производственном секторе во многом определяется общей тенденцией [«Индустрии 4.0»](#), характерной для создания умных производств, — отметил [Денис Степанов](#), менеджер по развитию бизнеса Frost & Sullivan в [России](#). — В сегментах розничной торговли, [здравоохранения](#), банковского дела и финансов BDA применяется в течение последних нескольких лет. В ближайшем будущем широкое применение получит развитие направления персонализированных услуг. Спрос на аналитику больших данных в сфере ООС будет высоким, особенно в развивающихся странах Азиатско-Тихоокеанского региона.

По мнению представителей Frost & Sullivan, применение BDA позволяет обеспечить глубокое понимание клиентских потребностей, что особенно

важно для смарт-банкинга. В финансовом сегменте аналитика больших данных используется для персонализации сервисов, прогноза/профилактики оттока клиентов, обнаружения мошенничества и т.д. Например, [MoneyGram International](#), международная компания, предоставляющая платежные услуги, внедрила решение [IBM InfoSphere Identity Insight](#) для контроля финансов и предотвращения мошеннических действий, связанных с переводом средств.

Также аналитика больших данных все чаще используется частными компаниями и правительственными организациями для оценки экологических рисков, оптимизации использования ресурсов и обеспечения соблюдения экологических норм. Global Forest Watch 2.0 — один из проектов по обеспечению сохранности лесов, реализуемый World Resource Institute в сотрудничестве с [Google Inc.](#) С момента начала проекта и внедрения решений BDA в 2013 году темпы обезлесения Амазонии снизились на 80% по сравнению с 2004 годом.

По словам Дениса Степанова, к 2025 году также значительно увеличится область использования и количество проектов, внедряющих аналитику больших данных. В частности, BDA будет использоваться в таких областях, как управление рисками в реальном времени, [блокчейн](#)-аналитика и удаленный мониторинг анализируемых объектов, считает эксперт.

2018

Рынок ПО для работы с большими данными и бизнес-аналитики находится на взлёте

В 2018 году объем глобального рынка [программного обеспечения](#) для работы с большими данными и бизнес-аналитики ([Big Data](#) and Analytics, BDA) достиг \$60,66 млрд, что на 12,3% больше, чем годом ранее. Об этом свидетельствуют данные аналитиков [IDC](#).

К решениям BDA 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.

Vendor	Revenue (\$M)			Market Share (%)			Growth (%)	
	2016	2017	2018	2016	2017	2018	2016-2017	2017-2018
Oracle	6,956.7	7,420.0	7,993.2	14.2	13.7	13.2	6.7	7.7
Microsoft	4,771.2	5,755.0	7,112.8	9.8	10.7	11.7	20.6	23.6
SAP	5,484.7	5,882.7	6,318.5	11.2	10.9	10.4	7.3	7.4

Лидеры рынка программного обеспечения для работы с большими данными и бизнес-аналитики

В IDC делят рынок [ПО](#) для работы с большими данными и бизнес-анализа на три основных сегмента:

решения для управления эффективностью [бизнес-процессов](#) на основе анализа данных;

инструменты для бизнес-аналитики;

платформы для интеграции и управления данными.

Крупнейшим производителем BDA-программ названа [Oracle](#), которая в 2018 году заработала на этом рынке почти \$8 млрд, что соответствует доли в размере 13,2%. На втором месте находится [Microsoft](#) с 11,7%-м показателем присутствия. При этом доходы корпорации увеличились на 23,6%, что позволило Microsoft приблизиться к Oracle. Лидирующую тройку замкнула [SAP](#) (10,4%).

Лидером в сегменте инструментов для бизнес-аналитики названа Microsoft, на которую пришлось 10,6% продаж такого программного обеспечения в 2018 году. Следом за Microsoft расположилась SAP, записавшая в актив 8,6% сегмента, на третьей строчке расположилась [IBM](#) (7,8%).

Если в категории решений для бизнес-аналитики нет ярко выраженного лидера, то среди производителей [софта](#) для продвинутой и предиктивной аналитики доминирует [SAS](#). Доля компании в этом сегменте по итогам 2018 года оказалась равной 27,7% против 30,1% в 2017-м. На второй позиции разместилась IBM с 12,5%-ой долей, на третьей — [MathWorks](#) (11%).

[[Image:Снимок_экрана_2019-10-24_в_15.49.44.png|840px|thumb|Категории рынка BDA]

В исследовании также указан сектор программных платформ на базе [искусственного интеллекта](#). Здесь нет равных IBM, на которую пришлось 9,3% такого программного обеспечения по итогам 2018 года. Далее идет плотная группа преследователей, состоящая из следующих компаний: SAS, Microsoft, [Palantir](#), [Google](#) и [Amazon Web Services \(AWS\)](#). Их доли в сегменте составляют около 2-3%.

SAS отдали первое место по объему реализации ПО для интеграции данных с долей в размере 17,8% в 2018 году. Входящие в топ-3 компании [Informatica](#) и IBM получили доли на уровне 16,8% и 15,6%. Еще одним крупным производителем таких решений является Microsoft, а остальные компании сильно отстают от лидеров.

Первое место по продажам ПО для анализа взаимоотношений компаний с клиентами делят [Adobe](#) и [Salesforce](#), которые заработали на таких продуктах примерно по \$1,2 млрд в 2018 году. В тройку лидеров ещё попала Oracle.

Как сообщается в исследовании IDC, цифровая трансформация стала ключевым драйвером роста расходов на BDA благодаря инициативам руководства, что приводит к глубокой оценке существующих бизнес-практик и требований к лучшему, более быстрому и более полному доступу к данным, аналитике и пониманию.

Компании перестраивают архитектуру для удовлетворения этих потребностей и инвестируют в современные технологии, которые позволят им

внедрять инновации и оставаться конкурентоспособными. Решения BDA лежат в основе многих из этих вложений.

Worldwide Artificial Intelligence Software Platforms Revenue by Vendor, 2016-2018 (\$M)

Vendor	Revenue (\$M)			Share of Total Market (%)			Growth (%)	
	2016	2017	2018	2016	2017	2018	2016-2017	2017-2018
IBM	161.0	190.0	240.6	10.0	9.2	9.3	18.0	26.6
SAS	10.0	43.5	89.0	0.6	2.1	3.4	335.0	104.6
Microsoft	12.7	36.9	88.5	0.8	1.8	3.4	190.6	139.8
Palantir	63.4	71.5	75.2	3.9	3.5	2.9	12.8	5.2
Google	25.8	36.2	58.5	1.6	1.8	2.3	40.3	61.6
Amazon Web Services	11.0	24.9	51.3	0.7	1.2	2.0	126.4	106.0
Digital Reasoning	28.7	38.1	49.8	1.8	1.9	1.9	32.8	30.7
IPsoft	19.5	26.8	41.1	1.2	1.3	1.6	37.4	53.4
Wipro	19.6	29.4	40.6	1.2	1.4	1.6	50.0	38.1
IFLYTEK	6.7	13.0	26.1	0.4	0.6	1.0	94.0	100.8
Subtotal	358.4	510.3	760.7	22.2	24.8	29.3	42.4	49.1
Other	1,259.5	1,543.8	1,838.8	77.8	75.2	70.7	22.6	19.1
Total	1,617.9	2,054.1	2,599.5	100.0	100.0	100.0	27.0	26.6

Лидеры рынка ПО с искусственным интеллектом

Все больше компаний начинают понимать необходимость более решительных действий в плане BDA-инициатив. Прежде активность в данной сфере, в основном, заключалась в подготовке к внедрению технологий обработки [Big Data](#) и бизнес-анализа. В силу этого значительная часть расходов приходилась на аппаратное обеспечение и консалтинговые услуги. Однако в 2018 году ситуация медленно начала меняться по мере того, как компании переходят от стадии подготовки к фактическому внедрению BDA. Это, как ожидается, приведет к росту затрат на платформы, инструменты и приложения, которые обеспечивают реальные доходы.

Больше всего в технологии обработки Big Data и бизнес-аналитики инвестируют представители таких отраслей, как банковский сектор, дискретное и непрерывное производство, профессиональные услуги и федеральное правительство.

В исследовании за 2018 год не называются крупнейшие рынка сбыта решений BDA. Аналитики IDC выпускали прогноз, согласно которому на рынке бизнес-аналитики лидируют [США](#), на которых придется около \$100 млрд выручки по итогам 2019 года. В последующие годы [американский](#) рынок будет занимать больше половины инвестиций в BDA.

Вторым по величине регионом продаж для работы с большими данными аналитики называют [Японию](#), где расходы в 2019-м будут измеряться в объеме \$9,6 млрд. В тройку лидеров войдет [Британия](#) (\$9,2 млрд), а следом расположатся [Китай](#) (\$8,6 млрд) и [Германия](#) (\$7,9 млрд).

Согласно прогнозу аналитиков [Frost & Sullivan](#), объем мирового рынка инструментов для анализа больших данных будут расти примерно на 35,9% ежегодно и достигнут \$67,2 млрд в 2021-м, что более чем в 2,5 раза превышает показатель 2016 года. Чаще всего технологии BDA будут использоваться в таких областях как, как производство, [финансы](#), [здравоохранение](#), охрана окружающей среды и розничная торговля. [\[2\]](#)

Китай вышел в лидеры по объему создаваемых данных

В феврале 2019 года стало известно о лидерстве [Китая](#) с точки зрения объема генерируемых данных. В последующие годы отрыв Поднебесной будет только увеличиваться.

Согласно исследованию, проведенному аналитиками [IDC](#) совместно с [Seagate](#), в 2018 году китайские компании и потребители произвели 7,6 Збайт данных, тогда как в [США](#) показатель составил 6,9 Збайт. Прогнозируется, что в 2025 году в этих странах будет сгенерированы данные в объеме 48,6 и 30,6 Збайт соответственно.

Что касается глобальных показателей, то в 2018 году на свет появились данные на 33 Збайт, а к 2025-му их будет в мире уже 175 Збайт. Наибольший рост объемов выпускаемых данных покажут развлекательные платформы, системы [видеонаблюдения](#), подключаемое к интернету оборудование и инструменты для повышения производительности.



Если США в 2025 должны создать 30,6 зеттабайт, то у Китая к в то же время будет на 18 зеттабайт больше

В аналитическом отчете данные называются «сердцем цифрового мира» и «самым ценным нематериальным активом компаний, который может обеспечить конкурентное преимущество в цифровой трансформации». Бизнес использует данные, для того чтобы выходить на новые рынки, улучшать обслуживание клиентов, создавать дополнительные потоки доходов, продавая эту информацию. Аналитики добавляют, что информация все чаще рассматривается как валюта, на которую опирается мировая экономика.



Данные лежат в основе цифрового мира, и мы все больше превращаемся в информационную экономику,
— говорится в докладе.



По прогнозам экспертов, в [Китае](#) скорость генерирования данных будет примерно на 3% выше, чем в мире в целом. КНР оторвется от других стран благодаря своему огромному населению, а также развитию новых технологий, таких как [интернет вещей](#) и сети [5G](#).

Авторы исследования также обращают внимание на то, что использование личной информации все чаще вызывает обеспокоенность, поскольку данные могут применяться для построения моделей поведения или реагирования.[\[3\]](#)

2017

Изменения в тройке лидеров: Microsoft опередила SAP

В сентябре 2018 года аналитическая компания [International Data Corporation \(IDC\)](#) представила некоторые результаты исследования мирового рынка [программного обеспечения](#) для работы с большими данными ([Big Data](#)) и аналитических инструментов. По итогам 2017 года продажи такого [софта](#) в глобальном масштабе достигли почти \$45 млрд против \$40,8 млрд годом ранее и \$37,5 млрд в 2015 году. На десятку крупнейших производителей пришлось больше 70% доходов на рассматриваемом рынке.

Речь идет о [ПО](#), которое, как поясняют в [IDC](#), позволяет собирать, управлять, организовывать, анализировать, обеспечивать доступ и распространять структурированные и неструктурированные данные.

Рынок программных решений Big Data и аналитики поделен специалистами на следующие сегменты:

приложения для управления производительностью и анализа;

[бизнес-аналитика](#) и аналитические инструменты;

ПО для управления анализом данных и интеграции платформ.

Каждый из этих категорий имеет подсегменты, которые приведены на иллюстрации ниже:

Разделение рынка ПО для работы с Big Data и аналитических инструментов на сегменты и подсегменты, данные IDC

Крупнейшим производителем ПО для работы с Big Data и анализа информации в IDC называют компанию [Oracle](#), доля которой на рынке в 2017 году составила 16,5%. В тройку лидеров также попали [Microsoft](#) (12,8%), которая смогла опередить компанию [SAP](#) (12,7%), занимавшую вторую позицию в 2016 году.

Крупнейшие производители ПО для работы с Big Data и аналитических инструментов, данные IDC за 2015-2017 гг.

Следом в рейтинге крупнейших разработчиков расположились

[IBM](#) (доля по итогам 2017 года — 10,6%);

SAS (6,6%);

[Salesforce](#) (2,8%);

[Amazon Web Services](#) (2,4%);

[Teradata](#) (2,1%);

[Adobe](#) (1,9%).

С точки зрения продаж ПО для бизнес-аналитики (с учетом аналитических инструментов) в 2016 году Microsoft вышла на первое место, а в 2017-м — осталась на нем. SAP, которая лидировала в 2015 году, скатилась на вторую позицию. Третью строчку занимает IBM, четвертую — SAS, пятую — Oracle.[\[4\]](#)

IDC: Прогноз объема рынка

Согласно обновленному полугодовому прогнозу [IDC](#) (март 2017 года), в 2017 году мировой доход на рынке [больших данных](#) и [бизнес-аналитики](#) (BDA) достигнет 150,8 млрд. долл., увеличившись на 12,4% относительно прошлого года. Ожидается, что коммерческие закупки оборудования, [программного обеспечения](#) и услуг, связанных с BDA, будут

демонстрировать среднегодовой темп роста (в сложных процентах, CAGR) 11,9% в период до 2020 года, когда доход составит более 210 млрд. долл.

Сегменты с самым быстрым ростом расходов на большие данные и аналитику: CAGR за период 2015-2020 гг. Источник: IDC

Самый значительный объем инвестиций в технологии BDA в 2017 году ожидается в банковской сфере, в сегментах дискретного и непрерывного производства, в федеральных/центральных органах [государственной власти](#), а также в сфере профессиональных услуг. В совокупности, эти пять сегментов потратят 72,4 млрд. долл. на BDA в этом году. Они останутся крупнейшими сегментами расходов и в 2020 году, когда их общие инвестиции составят 101,5 млрд. долл. Самые высокие темпы роста расходов на BDA будут демонстрировать банковский сектор (CAGR 13,3%) и [здоровоохранение](#), страхование, сегмент операций с ценными бумагами и услуг управления инвестициями, а также отрасль [телекоммуникационных](#) услуг (каждый — с CAGR 12,8%). Следует отметить, однако, что лишь три сегмента из рассмотренных в прогнозе IDC продемонстрируют двузначный CAGR в период 2015-2020 гг.

Лидером инвестиций в технологии BDA будут сегменты ИТ- и бизнес-услуг, которые вместе составят более половины всех расходов на средства обработки больших данных и бизнес-анализа в 2017 году и в течение прогнозного периода. Расходы на услуги продемонстрируют также самые высокие темпы роста, с пятилетним CAGR 14,4%. Инвестиции в [ПО](#) вырастут до более чем 70 млрд. долл. в 2020 году, основная их часть будет приходиться на обработку опросов конечных пользователей, инструменты отчетности и анализа и средства управления хранилищами данных. Самые высокие темпы роста будут демонстрировать нереляционные хранилища аналитических данных и [когнитивные](#) платформы ПО (CAGR 38,6% и 23,3% соответственно), поскольку организации будут расширять использование

больших данных и аналитики. Закупки серверов и [СХД](#), связанные с технологиями BDA, будут расти на 9,0% ежегодно (CAGR) и достигнут 29,6 млрд. долл. в 2020 году.

Что касается размера компаний-заказчиков, то очень крупные организации (со штатом более 1 тыс. чел.) принесут более 60% всего дохода от BDA на протяжении прогнозного периода, и IDC ожидает, что в 2018 году эта группа в целом превысит планку 100 млрд. долл. расходов. Сегмент мелкого и среднего бизнеса также будет активно закупать средства BDA, составив почти четверть мирового дохода от компаний, насчитывающих менее 500 чел. персонала.

Что касается картины по регионам, то [США](#) будут крупнейшим рынком средств обработки больших данных и бизнес-анализа, и в 2017 году расходы здесь достигнут 78,8 млрд. долл. Вторым по величине расходов будет регион Западной [Европы](#) (34,1 млрд. долл. в 2017 году), а следом идет Азиатско-Тихоокеанский регион без [Японии](#) (13,6 млрд. долл.). Самый высокий темп роста расходов на BDA будет наблюдаться в Латинской Америке и APeJ, с пятилетним CAGR 16,2% и 14,4% соответственно.

2016

IDC: Объем рынка в \$130 млрд

Согласно предварительным данным аналитиков [IDC](#), в 2016 году объем мирового рынка [программного обеспечения](#), оборудования и сервисов в области бизнес-аналитики и работы с [большими данными \(Big Data\)](#) составил \$130,1 млрд, что на 11,3% больше, чем годом ранее.

Большая часть указанных расходов — около 13,1% или \$17 млрд — пришлось на банковский сектор (см. диаграмму ниже). Второе и третье место заняли дискретное и непрерывное производство (11,9% и 8,4% соответственно). Доля инвестиций от государственных органов и компаний, оказывающих разного рода профессиональные услуги, измерялась примерно по 7,5%.



Распределение расходов на покупку решений в области бизнес-аналитики и работы с большими данными, предварительные данные IDC за 2016 г.

Ожидается, что в 2015-2020 годах наиболее быстрорастущие продажи решений для Big Data и бизнес-аналитики будут в банковском секторе. Кроме того, высокие темпы роста будут наблюдаться на рынках телекоммуникаций, коммунальных услуг, страхования и транспортных перевозок.

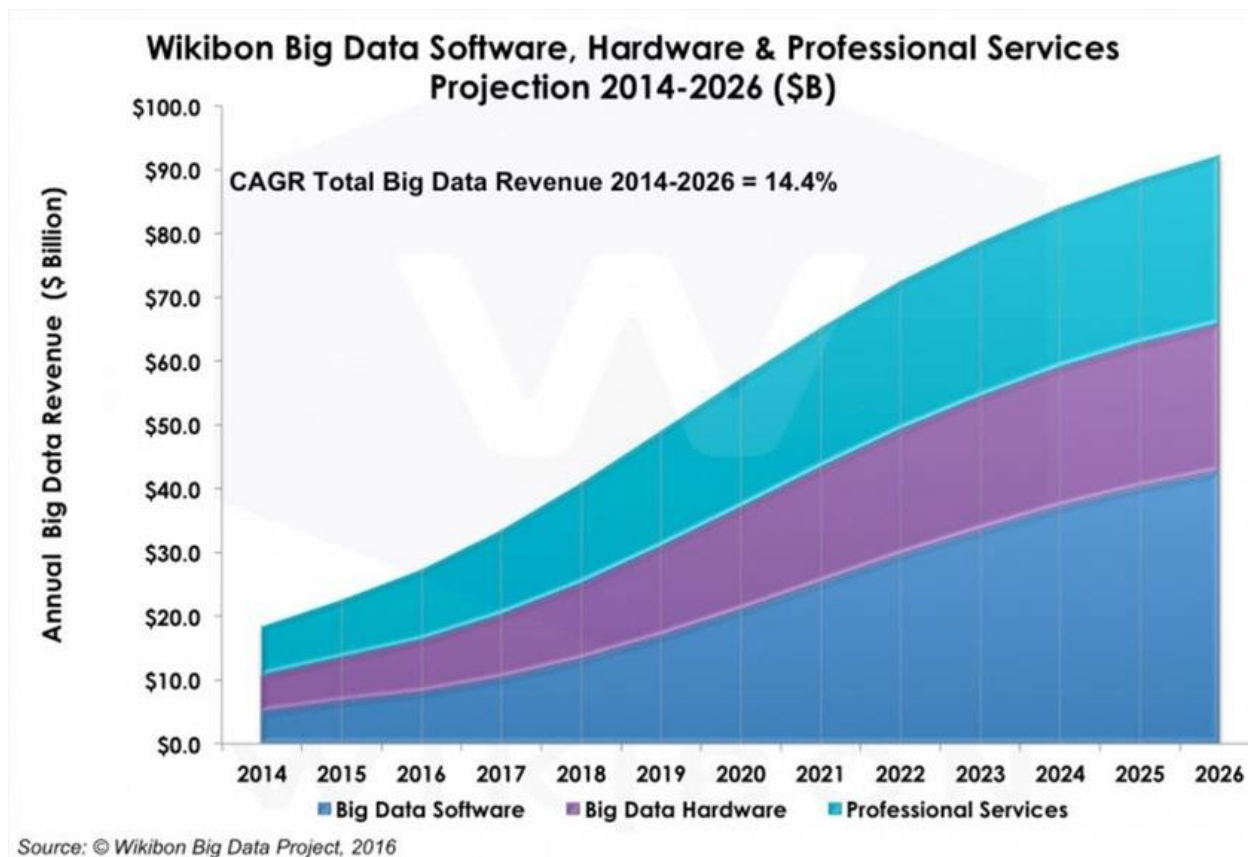
Что касается сегментов, то самые высокие продажи продемонстрируют ИТ- и бизнес-сервисы, на долю которых придется больше половины выручки на рынке технологий Big Data и бизнес-аналитики в период с 2015 по 2020 годы. Сегмент услуг будет расти на 14,9%, сегмент оборудования — на 8,7%. Последний достигнет \$29,9 млрд в 2020 году против \$60 млрд у программного сегмента, прогнозируют исследователи.

По словам вице-президента подразделения IDC Analytics and Information Management Дэна Вессета (Dan Vesset), спрос на технологии и сервисы в сфере работы с большими данными и бизнес-аналитикой стимулируют такие факторы, как увеличение объемов данных, развитие новых технологий и культурный сдвиг в сторону принятия решений, связанных с информацией. Эксперты ожидают, что с 2015 по 2020 годы объем рынка будет расти в

среднем на 11,7% ежегодно и превысит \$203 млрд к концу этого отрезка времени. [5]

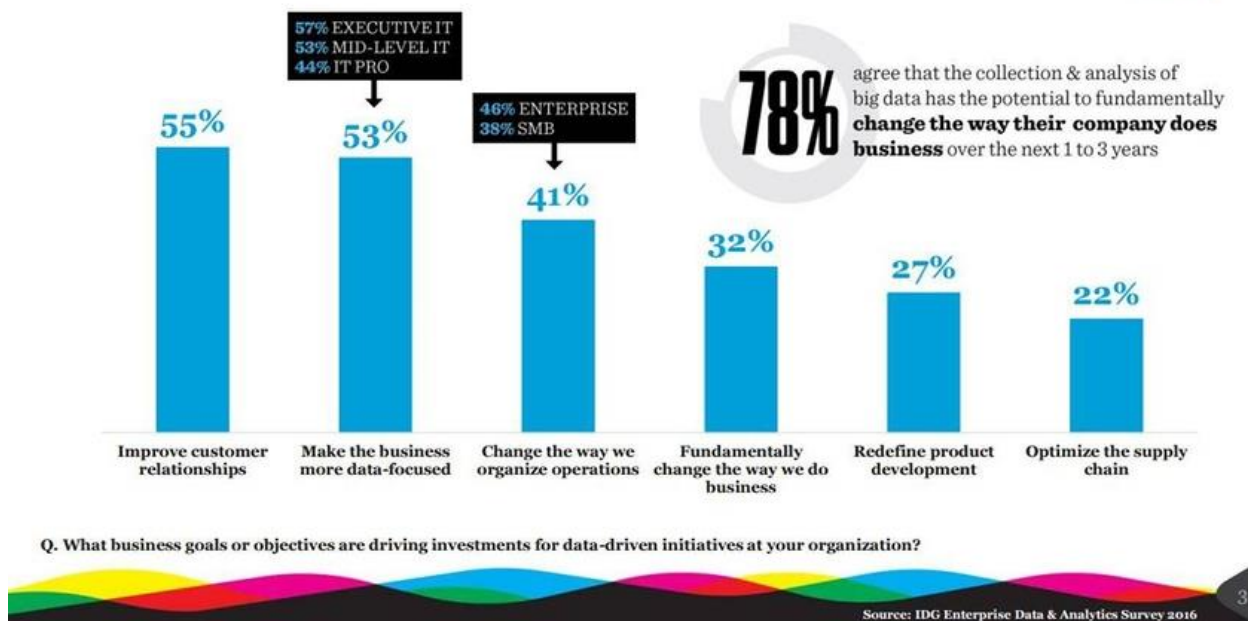
Данные Wikibon Big data Project

Прогноз роста рынка больших данных, 2011–2020 (\$ млрд), Wikibon



Данные IDG

Data Helps Customer Focused Organizations



2014

IDC: Объем рынка решения Big Data достигнет \$41 млрд к 2018 году

Агентство IDC опубликовало в сентябре 2014 года исследование, согласно которому объем рынка технологий и услуг в сфере больших данных будет ежегодно расти примерно на 26,4% и к 2018 г. достигнет \$41,5 млрд. Несмотря на снижение скорости роста — в прошлом пятилетнем прогнозе IDC предрекал рынку рост ежегодно на 27% — этот сегмент по-прежнему будет расти в 6 раз быстрее всего ИТ-рынка. Исследователи отмечают, что 2014 г. на оборудование пришлось 48,2% всего рынка больших данных. Эксперты также ожидают, что активные поглощения на этом рынке продолжатся.

Переход технологий больших данных на более высокий уровень развития подтверждают и другие аналитические агентства. В августе компания Gartner опубликовала последнюю версию «Цикла зрелости технологий». В обновленном графике большие данные покинули пик чрезмерных ожиданий, который занимали в прошлом году, и перешли в стадию избавления от иллюзий (Trough of Disillusionment). Финальной стадии зрелости, по оценкам экспертов, большие данные достигнут через 5–10 лет. А пока на пике

появляется анализ данных ([Data Science](#)), объединяющий весь комплекс подходов работы с большими данными.

Несмотря на то, что больше всего инвестиций в эти технологии по-прежнему приходится на [США](#), остальные мировые игроки склонные проявлять больше активности, чем раньше. Так, по данным IDC, в 2014 г. на страны [Европы](#), Ближнего востока, Африки и Азиатско-тихоокеанского региона (без учета [Японии](#)) пришлось около 45% рынка оборудования, [ПО](#) и услуг в сфере больших данных.

Согласно объявленному в январе 2013 года прогнозу [IDC](#), мировой рынок технологий Big data будет расти в среднем на 31,7% в год и достигнет к 2016 году \$23,8 млрд. Аналитики отмечают бурный тем роста соответствующих технологий и сервисов, а также активное формирование подsegmentов этого рынка, в том числе появление новых. На протяжении ближайших четырех лет эти высокие темпы прироста рынок сохранит, полагают эксперты.

Рост отдельных segmentов мирового рынка Big data будет варьироваться при этом от 21,1% для сервисов до 53,4% для систем хранения. Высокий темп роста в segmentе систем хранения больших данных связан с особенностями их использования в настоящий момент.

Недостаток специалистов на рынке с компетенцией в области аналитики и управления большими данными приведет к тому, что все больше покупателей будут выбирать облачные решения, а также [аутсорсинг](#) аналитических сервисов, не только непосредственно технологий.

Решения Big data для обработки данных, сгенерированных машинами, начнут набирать обороты ближе к концу прогнозного периода. Таковы основные тренды, ожидаемые IDC на этом рынке.

2013: IDC рисует невероятные перспективы развития больших данных

Согласно обновленному в январе 2013 года прогнозу [IDC](#), мировой рынок технологий Big data будет расти в среднем на 31,7% в год и достигнет к 2016

году \$23,8 млрд. Аналитики отмечают бурный тем роста соответствующих технологий и сервисов, а также активное формирование подsegmentов этого рынка, в том числе появление новых. На протяжении ближайших четырех лет эти высокие темпы прироста рынок сохранит, полагают эксперты.

Сегмент визуальных инструментов анализа данных будет расти в 2,5 раза быстрее, чем остальная часть рынка средств бизнес-анализа. К 2018 году инвестиции в системы визуализации данных, удобные для самостоятельного применения всеми пользователями, станут обязательными для всех предприятий.

В следующие пять лет затраты на облачные решения [больших данных](#) и аналитики будут расти втрое быстрее, чем на локальные решения. Обязательными станут гибридные облачные системы.

Сохранится нехватка квалифицированного персонала. Только в [США](#) в 2018 году будет 181 тыс. рабочих мест для специалистов по углубленному анализу и в пять раз больше — для работников с навыками в области управления данными и их интерпретации.

К 2017 году основой стратегии в области Больших Данных и аналитики станет единая платформа данных, на которой работают средства управления информацией, анализа и поиска.

В 2015 году ускорится рост применения приложений, выполняющих сложную и прогнозную аналитику, в том числе с применением технологий [машинного обучения](#). Сегмент этих приложений будет расти на 65% быстрее, чем рынок приложений, не имеющих функций прогнозирования.

70% крупных организаций уже приобретают внешние данные, а к 2019 году этот показатель вырастет до 100%. Параллельно больше организаций начнут извлекать доход из своих данных, продавая их или предоставляя какой-то ценный контент на их основе.

Внедрение технологий для непрерывного анализа потоков событий в 2015 году ускорится в связи с распространением технологий [Интернета вещей](#), а

среднегодовой темп роста рынка Интернета вещей на протяжении следующих пяти лет будет составлять 30%.

Сегмент платформ управления принятием решений до 2020 года будет увеличиваться ежегодно на 60% — в связи с потребностями в систематизации процесса принятия решений и сохранении знаний о нем.

Рынок систем анализа медиаданных — видео, аудио и изображений — за 2015 год вырастет как минимум втрое, на долю соответствующих инструментов будет приходиться большая часть всех инвестиций в технологии Больших Данных и аналитики.

2012

IDC: Прогноз роста расходов на Big Data до \$34 млрд уже в 2013 году

В следующие 8 лет количество данных в мире достигнет 40 зеттабайт, что эквивалентно 5200 гигабайт (Гб) на каждого жителя планеты, по данным исследования [IDC Digital Universe](#), опубликованного в декабре 2012 года. 40 зеттабайт эквивалентно 40 триллионам Гб, что в 57 раз больше, чем количество песчинок на пляжах на всей поверхности Земли. По прогнозам, количество данных на планете будет удваиваться каждые два года вплоть до 2020 года.

Рынок big data

▲ **31,7%**

Среднегодовой рост
рынка технологий big
data

★ **\$23 800 млн**

Достигнет рынок big
data в 2016 году по
версии IDC

★ **\$34 000 млн**

В такую сумму
оценивают в Gartner
расходы на big data в
2013 году

На технологии работы с большими данными в 2013 году в мире будет потрачено порядка 34 млрд долл., а к 2015 году в этом секторе будет создано

4,4 млн рабочих мест, но самым сложным вопросом все равно остается то, какие результаты при этом предполагается получить. Даже [банки](#) не занимаются анализом больших объемов данных в производственных масштабах, поскольку основная их часть хорошо организована и хранится на мэйнфреймах. Дальше всего в этом направлении продвинулись сети розничной торговли. Они с успехом обрабатывают данные средств радиочастотной идентификации ([RFID](#)), систем снабжения и покупательских карт. Во многих других отраслях только начинают искать способы монетизации имеющихся данных — но далеко не все данные являются ресурсом, и на анализ некоторых из них тратятся лишние деньги, указывает аналитик.

IDC: Объем мирового рынка Big Data в 2012 году - \$8 млрд

Аналитики [IDC](#) подсчитали, что объем мирового рынка Big Data с учетом затрат на создание инфраструктуры, в том числе с использованием технологии Cloud, затрат на очистку, обработку данных и так далее, составил в 2012 году порядка 8 миллиардов [долларов](#). В мире быстрее всего будет развиваться доля, приходящаяся на [системы хранения данных](#), затем, «облачные» технологии (IaaS), решения «под ключ». Технологии Big Data могут принести пользу в [медицине](#), например, для анализа мониторинга здоровья пациентов, накопления знаний о человеческом теле. Эти технологии вполне можно применить для улучшения жизни города (они внедрены в [Лондоне](#) и Барселоне для планирования транспортной инфраструктуры).

Между тем, мировые расходы на big data в 2012 году должны были составить \$28 млрд, а в 2013 году – уже \$34 млрд, согласно прогнозу [\[6\]](#), опубликованному [Gartner](#) в октябре 2012 года. Большая часть этих средств будет затрачена на адаптацию традиционных решений к работе с большими данными, а на разработку нового функционала под bid data – только \$4,3 млрд в 2012 году.

Наибольшее влияние big data аналитики отмечают в таких областях как аналитика социальных сетей и контент-аналитика: на них придется 45% расходов как в 2012, так и в 2013 году. Другие технологические области, на которые оказывают существенное влияние большие данные, это инфраструктура приложений, промежуточное [ПО](#), ПО хранения информации, системы управления БД, интеграция данных, [BI](#) и [SCM](#).

Марк Байер (Mark Beyer), вице-президент Gartner, заявил, что в 2011 году большие данные стали новым драйвером расходов на ИТ практически во всех категориях. По его словам, к 2018-2020 годам, тем не менее, требование о поддержке big data станет абсолютно рутинным для всех традиционных вендоров промышленных ИТ решений.

По словам эксперта, явление big data оказывает «распределенное» влияние на рынок [информационных технологий](#), так что в определенный момент времени поддержка [больших данных](#) станет не более чем отраслевым стандартом, и «большие данные» снова станут «просто данными». Те технологии, которые не смогут соответствовать новой информационной реальности, станут морально устаревшими. Причем организации, сопротивляющиеся этим изменениям, будут испытывать на себе тяжелые экономические последствия.

IDC: Ноябрьский прогноз роста рынка в 2012 году на 50% до \$7 млрд

Мировой объем рынка [Big Data](#) по итогам 2012 года вырастет почти на 50% до \$7 млрд с \$4,8 млрд [долларов](#) в 2011 году. Об этом 15 ноября 2012 года заявил менеджер по исследованиям [IDC Россия Александр Прохоров](#). При этом в 2011 году распределение рынка Big Data на сегменты было следующим: 41% пришелся на услуги, 30% - на [программное обеспечение](#), 14% - на серверные технологии, 12% - на технологии [хранения данных](#), 3% - на сетевые технологии.

Большая часть рынка Big Data (но не вся) принадлежит рынку бизнес аналитики, который по прогнозам [IDC](#) в 2012-м году составит около \$100

млрд. По данным опроса IDC, 75% организаций в мире с числом сотрудников более 500 человек планируют инвестировать в решения бизнес-анализа данных в ближайшие 12 месяцев.

Согласно прогнозам IDC - рынок связанный с технологией Big Data вырастет до \$16,9 млрд в 2015 году. То есть продемонстрирует совокупный среднегодовой темп роста (CAGR) равный 39,4%, или будет расти примерно в семь раз быстрее, чем ИКТ-рынок в целом.

РАС прогнозирует рост рынка в 7 раз за 6 лет

Мировой рынок решений для обработки [больших данных \(Big Data\)](#) вырастет почти в 7 раз до 20 млрд евро в 2016 году по сравнению с 3 млрд евро в 2010 году, к таким выводам пришли аналитики [РАС](#) в исследовании, опубликованном в конце сентября 2012 года.

Согласно данным агентства, ежегодный рост рынка составит около 38% с 2010 года по 2016 год. Популярность [Big Data](#) объясняется аналитиками тем, что сегмент позволяет пользоваться расширенным функционалом, не требуя при этом дополнительных специализированных навыков, в то время, как подобный ему, сегмент системной интеграции (SI), уже требует иной уровень подхода к работе.

"Мы видим большой интерес к Big Data со стороны многих компаний, работающих с большим объемом неструктурированной информации, как то, журналы, аналитические компании", - говорит главный аналитик направления Software [РАС](#) Филипп Карнелли. "При этом, есть несколько факторов, которые пока сдерживают рост рынка".

По мнению аналитика, Big Data может существенно облегчить и улучшить работу потребительски-ориентированных компаний в различных областях: телекоммуникации, [банки](#), розница. Эти компании смогут собирать и обрабатывать данные из социальных сетей и иных источников, связанных с их бизнесом.

Филипп Карнелли отмечает, что вместе с ростом спроса на Big Data, общие затраты компаний на развитие подобных проектов снижаются за счет того, что компетентных специалистов становится больше. "Существует нечто вроде компромисса между дешевыми инструментами, требующими последующей дорогой отладки, и более надежными и простыми в использовании инструментами, которые стоят больше, но в целом, затраты на развертывание проектов снижаются", - сказал он.

Ряд компаний практически одновременно объявили о скором выходе коробочного варианта Big Data ([Microsoft](#), [IBM](#), [Teradata](#), [Oracle](#), [SAP](#)). Такие инструменты, по мнению [PAC](#), будут способствовать развитию рынка [Big Data](#). "Их деятельность помогает "одобрить" проекты Big Data даже для более консервативных или менее технически продвинутых компаний, предоставляя новые варианты", - продолжил Карнелли. "Еще более важно, что появился новый класс предложений - "данные-как-услуга", от [Google](#) и других. Это делает выход на рынок Big Data гораздо более простым".

IDC: Рынок технологий и сервисов для обработки больших данных вырастет до \$16,9 в 2015 г

Согласно прогнозу [IDC](#) (Worldwide Big Data technology and services forecast), вышедшему в марте 2012 года, рынок технологий и сервисов для обработки [больших данных](#) вырастет с \$3,2 млрд в 2010 году до \$16,9 млрд в 2015 году. Это соответствует среднегодовому темпу роста (CAGR) на уровне 40%, что примерно в 7 раз больше, чем среднегодовой темп роста всего ИКТ рынка в целом.

IDC указывает, что работа с Big Data станет еще одной must-have компетенцией в 2012 году. По данным IDC, в 2012 году будет сгенерировано около 2,43 Зеттабайт цифрового контента, что на 48% больше, чем в 2011 году. Более 90% этой информации будут представлять собой неструктурированные данные (изображения, видео, mp3 файлы и файлы социальных медиа), то есть данные, которые очень сложно поддаются анализу.

Исследование, проведенное LogLogic, показало, что 49% ИТ-руководителей уверены в готовности своих компаний справиться с Большими Данными. При этом 38%, напротив, признали, что даже не имеют подного представления о сути этого явления.

ЛЕКЦИЯ 11. АНАЛИТИКА БОЛЬШИХ ДАННЫХ

Аналитика больших данных – это процесс анализа больших и сложных источников данных для выявления тенденций, моделей поведения клиентов и рыночных предпочтений, что помогает принимать более эффективные бизнес-решения. Сложность анализа больших данных требует новых аналитических инструментов, таких как прогнозирующая аналитика, машинное обучение, потоковая аналитика, и такие методы, как анализ в базе данных и в кластере.

Большие данные обычно определяются четырьмя V:

Объем: большие объемы данных

Разнообразие: много разных форм данных, неструктурированных и структурированных

Скорость: частота входящих данных

Правдивость: достоверность данных

Помимо огромного объема данных, сложность собираемых данных создает проблемы в управлении данными, их интеграции и анализе. Но компании, которые объединяют неструктурированные источники данных, такие как контент социальных сетей, с существующими структурированными данными, такими как транзакции, могут добавлять контекст и генерировать новые, а зачастую и более богатые идеи.

Кроме того, большие данные описывают повышенную скорость входящих данных, поступающих из разрастающихся источников, таких как датчики, мобильные устройства, веб-потоки кликов и транзакции, что приводит к необходимости аналитики в реальном времени. Организации, которые могут извлечь выгоду из того, что происходит сейчас, чтобы предотвратить отказ оборудования, рекомендовать предмет для покупки,

выявить мошенничество с кредитными картами и многое другое, быстро становятся лидерами в своих отраслях.

Наконец, большие данные относятся к степени точности данных, точности и достоверности. Это не означает, что все данные должны быть тщательно отобранными и чистыми, потому что более сложные источники данных, такие как социальные сети, могут привести к новому пониманию с определенным анализом. Но важно, чтобы организации знали качество, точность и достоверность данных, используемых для формирования понимания и принятия решений.

Однако вместо того, чтобы полностью согласиться с битвой, в прошлом году TIBCO Software выпустила TIBCO Spotfire X, чтобы восстановить утраченные позиции. Этот выпуск не только добавил возможности, такие как поддержка потоковой передачи данных, которые уже были доступны в Tableau, но также позволил TIBCO Software предоставить значительно более быструю версию своего программного обеспечения, которая включает в себя возможности на естественном языке и некоторые возможности искусственного интеллекта (AI). Кроме того, TIBCO Software ранее в этом году еще раз продемонстрировала свою готовность инвестировать в эту область, приобретя SnappyData, поставщика фабрики данных, специально оптимизированной для интеграции различных типов данных.

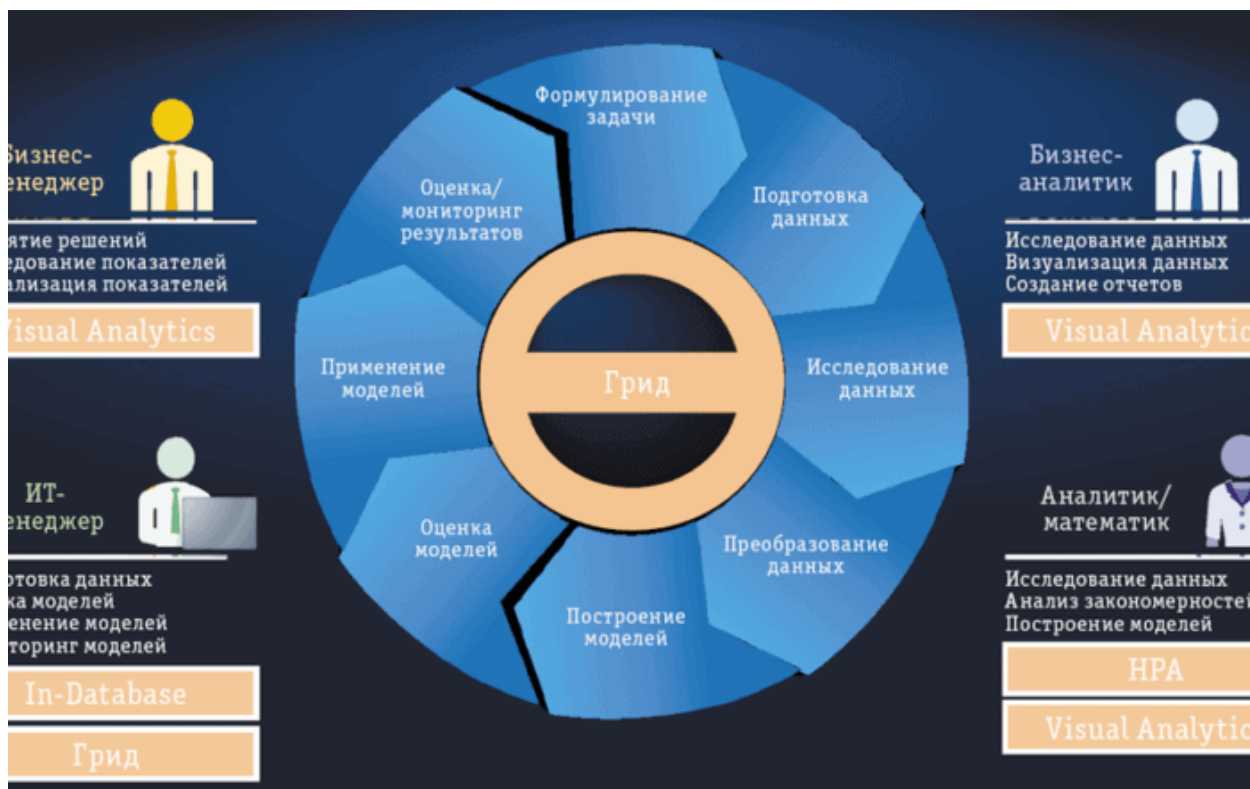
Влияние аналитики больших данных

В основе аналитики лежит превращение данных в пронизательные действия, которые повышают ценность для организации. Но рост структурированных и неструктурированных данных, также известных как большие данные, радикально изменил функцию аналитики.

В то время как большие данные расширили возможности, доступные для бизнеса, они также создали больше проблем для сбора, хранения и доступа к информации.

Этот революционный сдвиг предъявляет значительные новые требования к хранению данных и ставит новые задачи для аналитического программного обеспечения. Это также создает мощные возможности для обнаружения и реализации новых стратегий для развития конкурентного преимущества.

Реализация этих возможностей требует двух вещей: технологического потенциала для сбора и хранения больших данных, а также новых инструментов для превращения данных в понимание и, в конечном итоге, в ценность.



Получение ценности от больших данных

Аналитика больших данных может объединять данные в состоянии покоя (традиционные структурированные данные) с данными в движении (данные в

реальном времени), чтобы выявлять возможности и использовать их в данный момент.

Большие данные уже стали реальностью для большинства компаний, но объем и огромная сложность больших данных могут показаться ошеломляющими. Компании должны работать над тем, чтобы осмыслить и создать возможности как для данных в состоянии покоя, так и для данных в движении.

Большие данные представляют собой расширяющийся набор новых мощных возможностей. Хотя точно неизвестно, как именно большие данные будут использоваться через год, три или пять лет, необходимость превратить их в конкурентное преимущество означает, что ни одна компания не может позволить себе ждать, чтобы принять меры.

Получение выгоды от больших инвестиций в данные требует способности эффективно использовать данные. Поиск фрагментов информации, которые повышают рентабельность инвестиций для организации, похож на поиск иголки в стоге сена, поэтому многие компании сообщают о низком коэффициенте окупаемости инвестиций в большие данные.

Чтобы реализовать потенциал эпохи больших данных и минимизировать ее риски, предприятиям требуется унифицированная архитектура данных, а также программное обеспечение для анализа и визуализации данных.

Преимущества решения для анализа больших данных

Видимость в неизвестность

Надежная платформа для анализа больших данных позволяет пользователям обнаруживать невидимые тенденции и закономерности в больших и сложных наборах данных, которые способствуют более быстрой идентификации стратегических возможностей и угроз.

Единый взгляд на бизнес

Благодаря унифицированной архитектуре данных компаниям предоставляется обширное, согласованное и всеобъемлющее окно данных, повышающее эффективность принятия решений и позволяющее пользователям работать с самой точной и своевременной информацией.

Самое быстрое время для действий

Аналитика больших данных повышает производительность принятия решений, позволяя любому сотруднику компании предвидеть ситуации и возможности, задавать актуальные и своевременные вопросы и получать ответы, которые ведут к решительным действиям.

Основные возможности аналитики больших данных

Расширенные статистические и машинного обучения расчеты

Инструменты обнаружения [наук о](#) данных и статистические вычисления берут большие объемы исторических данных и используют их для получения новых знаний и поиска закономерностей. Машинное обучение помогает создавать и обучать мощные алгоритмы, которые могут улучшить бизнес-процессы и повысить ценность бизнеса.

Потоковая аналитика

Вы можете автоматизировать действия в режиме реального времени, применяя аналитические и прогнозные модели к оперативным данным. Используя среду визуальной разработки для быстрого создания и

развертывания потоковых приложений, вы можете позволить операционным системам оценивать данные, отправлять оповещения и быстро выполнять действия для своевременного принятия решений в соответствии с контекстом.

Визуализация данных

Для визуализации больших данных вам нужна простая статистика и встроенные коннекторы данных, которые позволяют быстро импортировать данные в интуитивно понятные панели мониторинга. Это позволит вам предоставить своим бизнес-пользователям возможность анализировать большие источники данных, принимать решения, основанные на реальных данных, и постоянно использовать информационные панели, отвечающие потребностям бизнеса.

Виртуализация данных

Виртуализация данных обеспечивает современный уровень данных, который позволяет пользователям получать доступ, объединять, преобразовывать и доставлять наборы данных с невероятной скоростью и экономичностью. Благодаря виртуализации данных пользователи получают не только актуальные, но и самые последние данные, которые легко найти, использовать и понять.

Управление активами данных

Управление данными обеспечивает постоянный доступ, доставку, управление и безопасность данных в соответствии с требованиями организации, используя такие инструменты, как управление основными данными, виртуализация данных, каталог данных, а также подготовка и обработка данных самообслуживания.

Обнаружение данных самообслуживания

Большое аналитическое решение для данных позволяет пользователям во всей организации исследовать данные и получать ответы без необходимости специализированного, углубленного моделирования данных. Это уменьшает

зависимость от ИТ и выделенных ресурсов бизнес-аналитики (BI) и значительно ускоряет процесс принятия решений.

Общие источники данных для аналитики больших данных

- Платформы больших данных
- Транзакционные данные клиента
- IoT / данные датчика
- Социальные медиа
- Интернет / онлайн данные
- Данные мобильного устройства
- Носимые устройства
- Исторические данные и данные в реальном времени
- Торговая точка (PoS)
- Гео-место
- Текстовые данные

Каковы некоторые из лучших вариантов использования?

- 360-градусный обзор клиентов
- Предотвращение мошенничества
- Разведка безопасности
- Оптимизация цен
- Операционная эффективность
- Эффективность цепочки поставок
- Рекомендации двигателей
- Анализ социальных сетей и ответ
- Профилактическое / профилактическое обслуживание
- IoT

ЛЕКЦИЯ 12. DATA MINING: ТЕХНОЛОГИИ И МЕТОДЫ АНАЛИЗА ДАННЫХ

Основной задачей любого аналитика является генерация гипотез, основанных на «внешних знаниях»: данных компании, рынка и другой полезной информации. Сегодня извлечение нужной информации из огромной «кучи» хранилищ данных называют интеллектуальным анализом данных – Data Mining, методы анализа данных которого позволяют принимать стратегически важные решения.

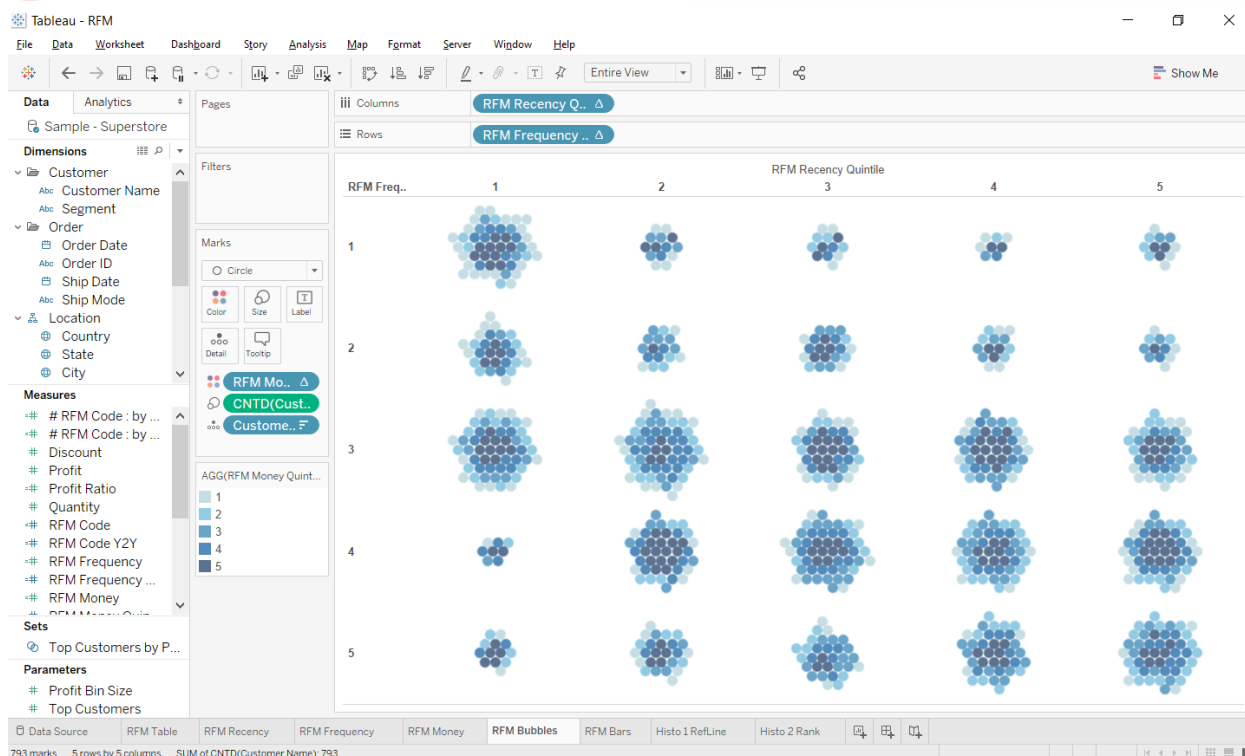
Data Mining благодаря широте возможностей, которые он открывает, нашел широкое применение в науке: его используют как отличный метод исследования. Однако в бизнесе он играет не меньшую роль: то, что помогает науке, двигать человечество по пути к прогрессу, позволяет бизнесу увеличивать прибыль и количество лояльных клиентов. Методы анализа данных Data Mining в бизнесе позволяют:

- сегментировать клиентов,
- прогнозировать продажи,
- проводить аналитику складских запасов,
- принимать решения об индивидуальных скидках для клиентов,
- привлекать новых клиентов.

Как используют Data Mining в компании Mail.ru?

Методы анализа данных в Data Mining

Визуализация Data Mining в аналитической программе Tableau:



Сегодня известны статистические и кибернетические методы анализа данных в Data Mining. Первые базируются на уже накопленных знаниях и данных, вторые – на разных математических подходах.

Статистические методы анализа данных Data Mining включают в себя: анализ исходных данных, многомерный статистический анализ, анализ связей и анализ временных рядов. Кибернетические методы Data Mining объединяют подходы, основанные на математике и применении искусственного интеллекта. Вот некоторые методы интеллектуального анализа данных:

- кластеризация – поиск и объединение похожих структур и объектов (слово «кластер» в переводе означает скопление или гроздь), не помогает делать выводы, лишь находит и объединяет объекты с общими свойствами;

- алгоритм k-средних (k-means, быстрый кластерный анализ) – помогает определить гипотезы относительно количества кластеров, значение k может зависеть от ранее проведенных исследований, предположений или даже интуиции,

— байесовские сети – графические структуры, представляющие вероятностные отношения между огромным массивом переменных, служат для создания вероятностного вывода на основе этих переменных,

— искусственные нейронные сети – очень популярная тема в последнее время, но прежде чем воспользоваться нейронной сетью, аналитик ее должен «обучить» (от того, насколько правильно, верно и точно будет обучена сеть, зависит ее эффективность в решении тех или иных задач).

Визуализация анализа данных в Data Mining

Представление информации, полученной любыми методами анализа данных в Data Mining, в визуальном формате улучшает ее понимание в несколько раз. Если говорить о том, что проведение некоторых методов анализа – технически сложный процесс, то его результаты могут и должны быть понятны практически любому пользователю информации. В этом может быть действенна только визуализация.

Создать визуализацию в Data Mining помогают профессиональные аналитические программы, созданные специально для визуализации данных и работы с ними специалистов нетехнического профиля.

Одной из таких аналитических программ является Tableau. Продукт предлагает простой и быстрый способ сделать бизнес-аналитику визуально приятной, красивой и интерактивной. Необходимые отчеты отделов в программе создаются за секунды, а освоить Tableau может специалист без IT-знаний, так как интерфейс программы достаточно прост и интуитивно понятен.

Мы, компания АНАЛИТИКА ПЛЮС, с 2012 года помогаем нашим клиентам работать с данными – находить полезные инсайты и использовать эту информацию для увеличения прибыли компании.

За это время мы разработали и внедрили решения для различных отраслей и направлений бизнеса:

— анализ продаж,

- прогнозирование эффективности промо-акций,
- отслеживание воронки продаж по конкретной кампании с показателями конверсии на каждом этапе,
- сегментация по различным методам: ABC, RFM и т.д.,
- ключевые показатели интернет-маркетинга,
- анализ товара на складах,
- аналитика для отдела кадров (карточки сотрудников, обучение, анализ KPI и т.д.),
- анализ финансовых показателей
- и многое другое.