Больши́е да́нные (англ. *big data*, ['bɪg 'deɪtə]) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence **Технологии больших данных жизненно необходимы современным банкам. Управлять активами, оценивать риски, сохранять и наращивать клиентскую базу — ключевые потребности кредитных организаций нельзя будет удовлетворить, не научившись пользоваться инструментами big data.** По данным консалтинговой компании Alacer, крупнейшие банки США накопили уже 1 эксабайт (10^{18} байт) информации. Такой объем данных содержится, например, в 275 млрд аудиозаписей песен в формате mp3.

Безопасность и противодействие отмыванию денег

инструментами big data.

По данным консалтинговой компании Alacer, крупнейшие банки США накопили уже 1 эксабайт (10^{18} байт) информации. Такой объем данных содержится, например, в 275 млрд аудиозаписей песен в формате mp3.

Не использовать данные и возможности, которые они таят, означало бы отказаться от дальнейшего развития. И банки их активно используют.

Мы рассмотрели пять основных сфер банковской деятельности, которые меняются с наступлением эры больших данных.

Безопасность и противодействие отмыванию денег

С помощью систем обработки данных банк знает о потребительском поведении своих клиентов.

Допустим, клиент N, женатый и с двумя детьми, имеет недвижимость в городе и автомобиль, стабильный личный доход 100 тыс. рублей в месяц, держит накопительный счет и имеет кредитную линию. Банк из истории трансакций знает,

сколько N тратит в месяц на еду и одежду, поездки, содержание автомобиля, оплату коммунальных счетов, на развлечения и прочее.

В один прекрасный день N снимает большую часть наличности, закрывает счета и покупает билет в один конец в европейскую страну. Или переводит деньги на другой счет, блокирует карты и перестает проводить привычные трансакции.

Из такого поведения можно сделать два вывода: либо N бросил семью с двумя детьми и пустился в бега, либо доступ к его картам и мобильному банку получили злоумышленники. По статистике, более вероятно второе.

Система безопасности банка, исходя из анализа поведения клиента, тут же подает сигнал тревоги. Специалисты могут принять меры — заморозить трансакции и связаться с N, чтобы выяснить, все ли в порядке.

Если система выявляет аномальное поведение — резкий рост покупательской активности, перевод непривычных сумм на другие счета, вывод средств, — это становится сигналом тревоги. Предупредительные меры можно принять еще до того, как клиент обнаружит кражу кредитной карты и взлом онлайн-банка.

Банк также может сравнивать поведение одного клиента с поведением других, сопоставимых по уровню доходов. Искусственный интеллект со временем составит портрет типичного потребителя для каждой группы клиентов. Исходя из этого шаблона, система сможет предсказывать дальнейшее поведение потребителей и выявлять факторы риска.

Управление рисками

Это одна из самых благодатных сфер применения больших данных в банковском деле. Управление любым видом рисков — операционных, рыночных, кредитных, правовых — зависит от полноты и объективности информации, которую получают риск-менеджеры. Инструменты big data помогут нарисовать всеобъемлющую картину на любом уровне, будь то благонадежность заемщика или экономическая ситуация в отдельном регионе страны.

Инвестиционная стратегия банка тоже зависит от оценки рисков в конкретной отрасли и регионе. Благо сегодня уже достаточно много эффективных инструментов, основанных на анализе big data, предназначенных для работы на рынке ценных бумаг.

При этом рынки все активнее обмениваются информацией и становятся взаимозависимыми в своих движениях вверх-вниз. В периоды высокой волатильности банки теперь могут не только быстро подстраиваться под ситуацию, но и предвидеть ее.

Обслуживание клиентов

Для клиента важно, чтобы банк обслуживал его быстро, качественно и внимательно. При этом клиент не терпит никаких проблем и сбоев. Если они случаются, то должны решаться быстро и желательно без его участия. Компания McKinsey провела опрос американских банков, согласно которому 76% из них используют Big Data для привлечения клиентов, построения лучшего взаимодействия и поддержки лояльности. В то же время, по данным Alacer, 50% клиентов традиционных банков регулярно подумывают о том, чтобы сменить банк. Комплекс действий специалиста по работе с клиентами, называемый customer service, по сути сводится к выстраиванию эффективного диалога. А лучший собеседник, как известно, тот, кто умеет слушать. Банк, применяющий инструменты анализа клиентских данных, — это собеседник, который многое знает о человеке еще до начала разговора, понимает его проблемы и знает, как их решить. Идеальный собеседник, с которым хочется общаться больше.

Сегодня клиентские данные включают не только внутренние банковские сведения о состоянии счета и истории трансакций, но и внешнюю информацию. Как человек ведет себя в соцсетях. Что ищет в Google. Что покупает в интернет-магазинах (и на что ему не хватает денег). С кем переписывается по e-mail и какую рассылку получает. Куда отправляется на праздники и в отпуск. Чем больше банк знает о своем клиенте, тем более персональным будет клиентское обслуживание.

Комплекс действий специалиста по работе с клиентами, называемый customer service, по сути сводится к выстраиванию эффективного диалога. А лучший собеседник, как известно, тот, кто умеет слушать. Банк, применяющий инструменты анализа клиентских данных, — это собеседник, который многое знает о человеке еще до начала разговора, понимает его проблемы и знает, как их решить. Идеальный собеседник, с которым хочется общаться больше.

Сегодня клиентские данные включают не только внутренние банковские сведения о состоянии счета и истории трансакций, но и внешнюю информацию. Как человек ведет себя в соцсетях. Что ищет в Google. Что покупает в интернет-магазинах (и на что ему не хватает денег). С кем переписывается по e-mail и какую рассылку

получает. Куда отправляется на праздники и в отпуск. Чем больше банк знает о своем клиенте, тем более персональным будет клиентское обслуживание.

Инвестиционные консультации

Чтобы давать правильные советы, финансовый консультант должен знать о возможностях/потребностях своего клиента и хорошо разбираться в ситуации на рынках. С помощью инструментов big data банки могут и то и другое.

Банк может быть в курсе таких событий в жизни клиента, как свадьба, рождение ребенка, поступление в университет, переход на новую работу, смена интересов, будь то новое увлечение или решение отправиться в кругосветное путешествие.

Исходя из полученных сведений, для каждого человека формируется предложение. Например, клиенту предлагают начать откладывать на образование, когда его ребенку исполняется 10 лет. А если семья планирует купить дом, банк может заранее предложить им ипотеку для молодых.

С другой стороны, банки активно используют алгоритмы для интеллектуального анализа ситуации на фондовых рынках. Инструменты big data помогают извлекать выгоду как в краткосрочной перспективе, так и в долгосрочных вложениях. Для этого анализируют массивы самых разных неструктурированных данных — от погоды до тональности местных новостей в разных частях света, от уровня безработицы до настроений в соцсетях.

В США крупнейшие инвестиционные банки одними из первых оценили пользу от анализа больших данных. Еще бы, ведь когда речь заходит о повышении прибыльности игры на рынке, алгоритмы big data просто незаменимы. Более того, банки сами стали триггером для роста индустрии анализа данных. Известен случай, когда уволившийся сотрудник одного инвестиционного банка меньше чем за полгода создал популярную программу на основе big data для торговли на бирже.

Принципы работы с большими данными

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

1. **Горизонтальная масштабируемость**. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжило работать.

2. **Отказоустойчивость**. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин (по этойссылке можно посмотреть размеры кластера в разных организациях). Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий.

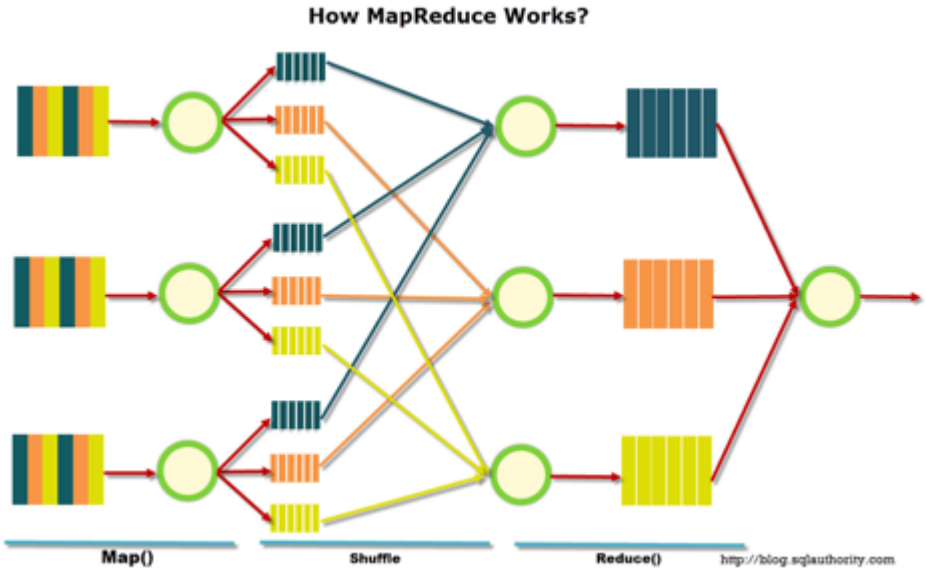
3. **Локальность данных.** В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.

Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы разработки средств разработки данных. Один из самых классических методов я разберу в сегодняшней статье.

MapReduce

Про MapReduce на хабре уже писали (раз, два, три), но раз уж цикл статей претендует на системное изложение вопросов Big Data – без MapReduce в первой статье не обойтись J

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах. MapReduce неплохо иллюстрируется следующей картинкой (взято по ссылке):



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

1. **Стадия Map**. На этой стадии данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи.

Функция map() примененная к одной входной записи и выдаёт множество пар ключ-значение. Множество – т.е. может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение. Что будет находится в ключе и в значении – решать пользователю, но ключ – очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce.

2. **Стадия Shuffle**. Проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти корзины послужат входом для reduce.

3. **Стадия Reduce**. Каждая «корзина» со значениями, сформированная на стадии shuffle, попадает на вход функции reduce().

Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины». Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

Несколько дополнительных фактов про MapReduce:

1) Все запуски функции **map** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.

2) Все запуски функции **reduce** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.

3) Shuffle внутри себя представляет параллельную сортировку, поэтому также может работать на разных машинах кластера. **Пункты 1-3 позволяют выполнить принцип горизонтальной масштабируемости.**

4) Функция map, как правило, применяется на той же машине, на которой хранятся данные – это позволяет снизить передачу данных по сети (принцип локальности данных).

5) MapReduce – это всегда полное сканирование данных, никаких индексов нет. Это означает, что MapReduce плохо применим, когда ответ требуется очень быстро.

Большие данные в бизнесе и маркетинге

Стратегии развития бизнеса, маркетинговые мероприятия, реклама основаны на анализе и работе с имеющимися данными. Большие массивы позволяют «перелопатить» гигантские объемы данных и соответственно максимально точно скорректировать направление развития бренда, продукта, услуги.

Например, аукцион RTB в контекстной рекламе работают с big data, что позволяет эффективно рекламировать коммерческие предложения выделенной целевой аудитории, а не всем подряд.

Какие выгоды для бизнеса:

- Создание проектов, которые с высокой вероятностью станут востребованными у пользователей, покупателей.

- Изучение и анализ требований клиентов с существующим сервисом компании. На основе выкладки корректируется работа обслуживающего персонала.
- Выявление лояльности и неудовлетворенности клиентской базы за счет анализа разнообразной информации из блогов, соцсетей и других источников.
- Привлечение и удержание целевой аудитории благодаря аналитической работе с большими массивами информации.

Технологии используют в прогнозировании популярности продуктов, например, с помощью сервиса Google Trends и Яндекс. Вордстат (для России и СНГ).

Например, Master Card используют большие данные для предотвращения мошеннических операций со счетами клиентов. Так удается ежегодно спасти от кражи более 3 млрд. долларов США.

В игровой сфере big data позволяет проанализировать поведение игроков, выявить предпочтения активной аудитории и на основе этого прогнозировать уровень интереса к игре.

Сегодня бизнес знает о своих клиентах больше, чем мы сами знаем о себе – поэтому рекламные кампании Coca-Cola и других корпораций имеют оглушительный успех.

Не столько «большие», сколько «умные»

Несмотря на явные преимущества, которые дает бигдата-аналитика в банковском бизнесе, сегодня большинство российских кредитных организаций используют во благо лишь ничтожную долю хранящейся у них информации. Только 30 крупнейших банков России заявляют, что внедрили и применяют соответствующие системы.

При этом речь идет уже не столько о больших данных (вскоре вся информация будет подпадать под этот термин), сколько об «умных» данных — тех сведениях, которые потенциально монетизируемы. Их использование станет таким же обязательным условием выживания банка, как достаточность капитала и сбалансированная кредитная политика.

Самые новые технические решения для банков в области больших данных будут представлены в этом году 15 сентября в Москве на Big Data Conference. Это дискуссионное и презентационное мероприятие проводится с 2014 года. Здесь встречаются создатели технологий, представители бизнеса и ученые.

Профессионалы банковской отрасли смогут воочию увидеть, как работают решения на основе big data, и пообщаться с лидерами области Data Science.

Программа конференции разделена на три ключевых трека: бизнес-кейсы, технологические решения и научный семинар. Наиболее интересными для бизнеса будут первые два трека, где рассматриваются уже действующие решения, их применение в практике, а также перспективные технологии, которые уже завтра будут работать в банковской сфере.