informatica

Практическое руководство по работе с большими данными

Практическое руководство для успешного запуска первого проекта по большим данным

Содержание

Подсказка. Нажмите для перехода к этому разделу

Как добиться успеха

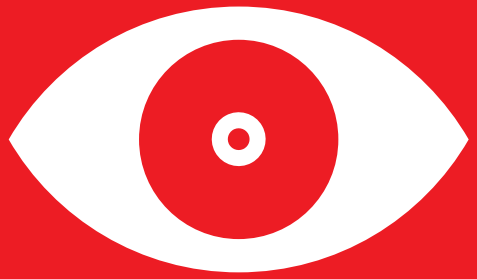
Не многие передовые технологии вызывают такое повышенное внимание, как технологии, связанные с большими данными.

И небольшое число технологий предоставляют предприятиям такой же огромный потенциал для развития, как технологии, связанные с большими данными. Когда в начале XXI века программное обеспечение вошло как неотъемлемая часть во все бизнес-процессы предприятий, стало ясно: данные изменят процессы работы.

Разумеется, слишком много разговоров о новом тренде вызывают неоправданные ожидания. В случае с большими данными мы в равной степени сталкиваемся как с обилием полезных советов, так и с дезинформацией. Новый мир бесконечных данных еще очень молод, и, к сожалению, его популяризация вызвала у многих двойственное отношение.
Цель данного руководства — внести ясность и разобраться во всех противоречивых вопросах.

Руководство создано для укрепления вашей стратегии и построения прагматического подхода. И неважно, беретесь вы за локальную, тактическую инициативу или планируете проект на уровне предприятия в целом, эта книга послужит практическим руководством в вашей работе.

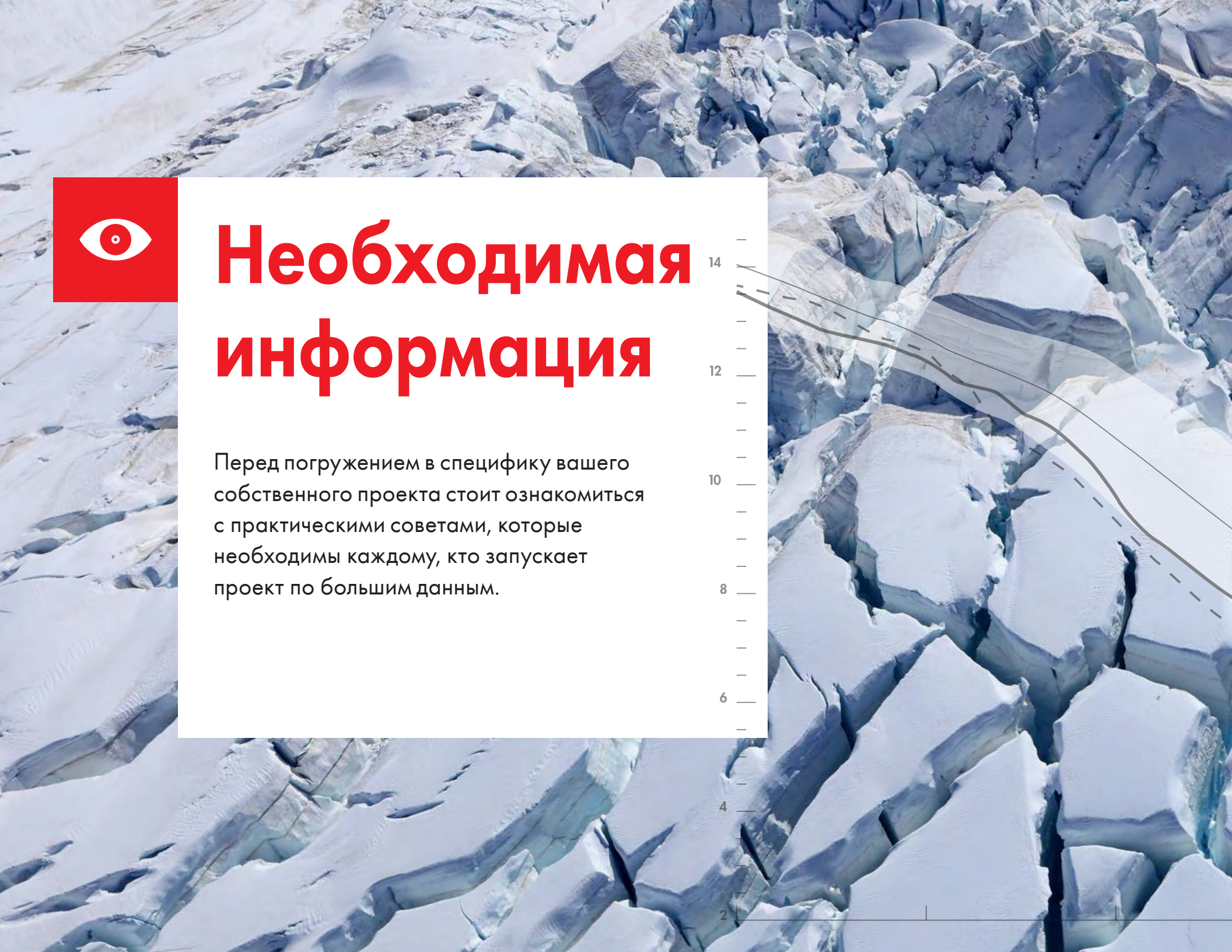
Итак, давайте начнем.



Часть A. Подготовка

Книга состоит из трех частей. В первой части мы поможем вам выбрать нужный именно для вашей организации проект.

Назад к содержанию <



Необходимая информация

Перед погружением в специфику вашего собственного проекта стоит ознакомиться с практическими советами, которые необходимы каждому, кто запускает проект по большим данным.

Необходимая информация

Почему большинство организаций начинают реализацию проектов по большим данным

Компании начинают реализацию проекта по большим данным по следующим причинам:

Компании хотят повысить качество аналитики и понимают, что для этого будет необходимо резко увеличить количество используемых данных. Обычно такие инициативы выдвигаются бизнес-подразделениями (например, отделом маркетинга).

Компании понимают, что они могут обеспечить своему продукту качественный сервис, предоставив аналитику в режиме реального времени, что поможет клиентам использовать продукт наиболее эффективно.

Компании хотят улучшить бизнес-процессы, сделать их более быстрыми и дешевыми, используя большие данные для принятия решений, касающихся конкретного бизнес-подразделения или процесса.

Компании понимают, что большие данные критически важны для каждого бизнес-подразделения, и хотят заложить фундамент подхода, ориентированного на данные.

Компании понимают, что пора начинать работать с большими данными, но не знают, с чего начать. В таком случае цель состоит в изучении больших данных и практической работе с ними.

Все это подтверждает актуальность использования больших данных. Если вы хотите внедрять проекты, которые выдержат испытание временем (и будут успешно использоваться всеми подразделениями компании), то вам необходимо четко и ясно представлять, какая из приведенных причин лучше всего отражает вашу заинтересованность.

Необходимая информация

Почему проекты по большим данным терпят неудачу

Исследование[1] показало, что 55 % проектов по большим данным не были завершены и многие проекты не достигли поставленных целей. Такая статистика довольно типична для любого нового направления в области ИТ, но следует воспользоваться уроками, которые дают нам такие проекты. Обратимся к четырем основным причинам, по которым проекты по большим данным терпят неудачу.

1

Отсутствие четко поставленной цели

Причина неудачи в большинстве случаев — нечеткая область проекта. Слишком много компаний запускает амбициозные, но вместе с этим совершенно неясные проекты без четко поставленных целей, и они терпят неудачу при жесткой расстановке приоритетов.

Реализация проекта по большим данным ради того, чтобы просто иметь проект по большим данным — верный путь к провалу. Сложность таких проектов требует четкого следования плану для получения определенных результатов. Без четко поставленных целей это невозможно.

[1] www.informationweek.com/software/information-management/vague-goals-seed-big-data-failures/d/d-id/1108384

Почему проекты по большим данным терпят неудачу

2

Несоответствие ожиданиям

Слишком много разговоров о больших данных вызывают неоправданные ожидания от проектов. Хотелось бы сказать, что результаты будут потрясающими и вы достигнете их за короткое время, но необходимо реалистично смотреть на то, что можно ожидать от проекта, сколько на него нужно будет затратить времени и какие усилия надо приложить для достижения результата.

Если от вас ждут новаторских решений и аналитических выводов, вам приходится заниматься кропотливой работой, анализируя огромное количество данных. Когда ожидания не соответствуют реальности, вам трудно адекватно оценить сроки выполнения проекта и его стоимость.

3

Превышение бюджета и нарушение сроков выполнения проекта

Большинство проектов по большим данным обычно выходит за рамки бюджета и отведенного на их выполнение времени, поскольку область больших данных абсолютно новая для многих предприятий. Как правило, такие проблемы вызывают не только неоправданные ожидания, но и отсутствие навыков в построении масштабируемой архитектуры.

Невозможно выйти за пределы «песочницы», не совершая ошибок, даже если приглашают хорошо известных дорогостоящих разработчиков Hadoop-Java для проведения широкомасштабных внедрений с кодированием вручную. В результате проекты по большим данным превращаются в научный эксперимент и так и не получают практического применения.

Необходимая информация

Почему проекты по большим данным терпят неудачу

4

Отсутствие возможности масштабирования

Очень сложно найти даже пять крупных разработчиков Hadoop Java. Когда проект расширяется и требуется 30 разработчиков Java в год, то он может потерпеть неудачу. И самое худшее заключается не в стоимости неиспользованных кластеров Hadoop, а в стоимости упущенной динамики развития и времени.

Слишком часто компании отдают предпочтение краткосрочной выгоде, а не долгосрочной стабильности. И пока мы не изменим стиль мышления, мы не сможем в полной мере осознать важность долгосрочного подхода. Для того чтобы можно было надежно управлять защищенными данными, необходимо задумываться о долгосрочных результатах проекта.

Эти четыре причины неудачи проектов по большим данным действительно встречаются очень часто. Теперь поговорим о том, как можно избежать неудачи проектов, а также о том, как внедрить долгосрочный проект.

Необходимая информация

Как создать успешный проект по большим данным

Если большинство проектов по большим данным терпят неудачу из-за недостатка ясности и вашей способности продемонстрировать их практические результаты, вам следует проработать обоснование проекта. Три полезных совета помогут обеспечить успешный запуск и стабильное развитие проекта.

1

Ставьте четкие цели и прогнозируйте реальные результаты

Если вы не уверены в цели проекта, поставьте такие же цели, как и для существующей инфраструктуры данных.

Если вашей организации требуются данные для определенных бизнес-процессов (таких, как выявление мошенничества или анализ рынка), подумайте, как использование больших данных может оптимизировать эти процессы. Вместо того чтобы полностью концентрироваться на новой проблеме, вам следует сосредоточиться на улучшении существующего процесса или проекта.

Без четкой цели и видимой ценности для бизнес-пользователей проект обречен на неудачу.

2

Задайте метрики, определяющие ценность проекта

Четко заданные метрики, привязанные к целям, избавят вас от многих проблем на пути к успеху проекта. Если вы ставите себе реальные цели, то прогресс вашего проекта будет очевиден.

И, что более важно, будут очевидны долгосрочные результаты вашего проекта. Задайте себе вопрос, как вы можете измерить результаты проекта в контексте ваших целей.

Это критически важно, поскольку поможет доказать бизнес-пользователям, что проект принесет гораздо большую пользу, чем они могут себе представить.

Необходимая информация

Как создать успешный проект по большим данным

3

Подходите к использованию инструментов и кодирования вручную стратегически

Избегайте искушения кодировать все вручную прямо в Hadoop. Помните, что цель вашего проекта не в том, чтобы построить работающую реализацию голыми руками с нуля — цель в том, чтобы ваша организация получила все преимущества от использования больших данных.

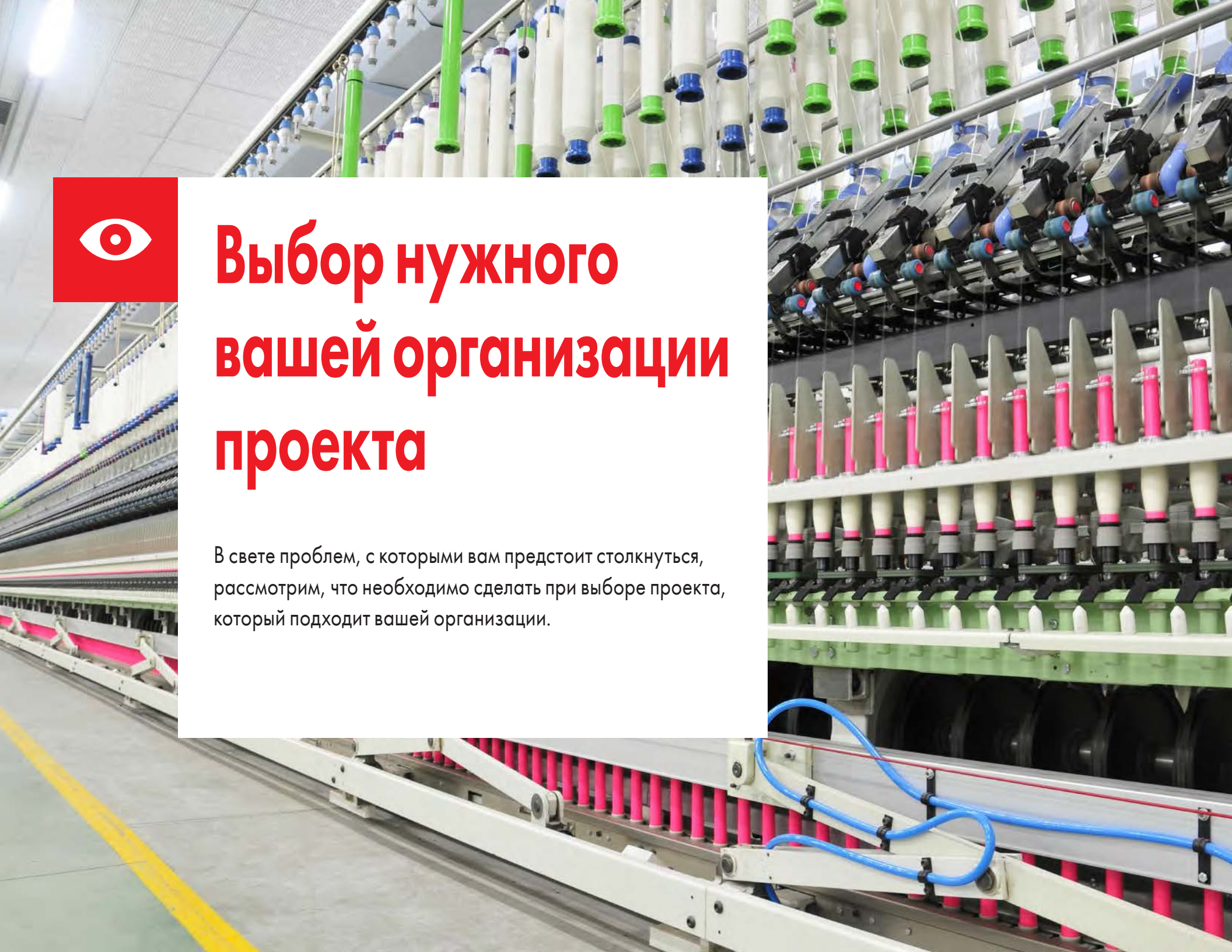
Вместо того чтобы кодировать вручную каждую интеграцию, очищать каждый набор данных и анализировать их в ручном режиме, вам следует использовать инструменты и средства автоматизации, которые помогут ускорить эти процессы.

Очень важно не растрачивать время дорогостоящих разработчиков Java на решение задач, к которым нельзя применить масштабирование или которые могут решить другие сотрудники. Ваша роль заключается в принятии стратегических решений в области развертывания ограниченных ресурсов в соответствии с поставленными целями.

Применяйте инструменты, которые могут повысить эффективность рабочей группы и которые могут использоваться специалистами по ETL, качеству данных и бизнес-аналитике, оставляя специалистам Java работу над самыми сложными задачами, для которых не существует инструментов.

Кроме того, поскольку такие технологии, как Hadoop, постоянно развиваются, надо построить уровень абстракции, защищающий от нескончаемого изменения спецификаций базовых технологий.

В первую очередь помните, что необходимые вам навыки работы встречаются редко, а инструменты доступны всегда.



Выбор нужного вашей организации проекта

В свете проблем, с которыми вам предстоит столкнуться, рассмотрим, что необходимо сделать при выборе проекта, который подходит вашей организации.

Что представляет собой проект, нужный вашей организации

Если ваша компания решилась внести изменения в бизнес-процессы и признала необходимость комплексной системы управления данными для повышения эффективности ее работы, вы можете пропустить этот раздел.

Если же вы собираетесь реализовать локализованный, тактический проект, который впоследствии может быть адаптирован для организации в целом, продолжите чтение.

Нужный вашей организации проект содержит четыре компонента, приведенных ниже.

1

Очевидная значимость

Правильно выбранный проект — это такой проект, который приносит равную пользу и ИТ, и определенному бизнес-подразделению. Это означает, что сотрудникам отдела, бизнес-подразделению или рабочей группе польза проекта должна быть очевидна.

2

Поддержка и финансирование

Для успеха проекта необходима поддержка руководителей, разделяющих вашу точку зрения. Проекты по большим данным требуют поддержки высшего руководства.

Если вы знаете, что можете создать прекрасную аналитическую систему для отдела логистики, но единственный руководитель, разделяющий вашу точку зрения, работает в маркетинговой службе, вам следует скорректировать свои планы. Если вас поддерживает отдел маркетинга, выбирайте проект в соответствии с требованиями аналитики маркетинга. Нельзя объять необъятное. Ориентируйтесь на заинтересованных в проекте бизнес-руководителей.

Что представляет собой проект, нужный вашей организации

3

Одним ударом справиться со всеми трудностями

Стратегическая важность вашего первого тактического проекта бесспорна. Вам необходимо не только доказать, что большие данные необходимы вашему бизнес-подразделению, но и убедиться, что они принесут пользу всей организации в будущем.

Подходите к выбору первого проекта стратегически. После того как вы продемонстрировали ценность больших данных, например, отделу маркетинга, вам будет легче получить поддержку отдела логистики, который, возможно, иначе не оценил бы вашу инициативу.

4

Навыки, применимые в разных областях

Итак, вам необходимо убедить в ценности вашего проекта остальные отделы организации. Поэтому вы должны быть уверены в том, что можете получить необходимые умения и навыки от реализации первого проекта. То есть, вам надо обязательно документировать весь процесс проекта, чтобы применить полученные навыки в следующем проекте. Помните, что если вы нацелены на успех, то вы нацелены и на последующие проекты.

Так что будьте готовы к тому, чтобы реализовать большее количество проектов в будущем. Это не только расширение вашего кластера. Это — вопрос расширения ваших навыков и процессов. Вам необходимо либо найти большее количество специалистов по Java/Hadoop, либо найти способ получить большую отдачу от имеющихся ресурсов.

Что надо принимать во внимание

При выборе следующего проекта также нужно учитывать, как он будет сказываться на работе организации в целом. Существуют три аспекта, играющих важную роль в оценке правильности выбора проекта по большим данным.

1

Затраты и сбои проекта

На базовом уровне стоимость вашего проекта основывается на времени и деньгах, которые будут затрачены на его реализацию. В реальности необходимо также учитывать потенциальные сбои, которые могут возникнуть при запуске проекта.

Иногда сбои являются процедурными — когда бизнес-подразделения привыкли быть хозяевами своих данных и им неудобно передавать контроль над ними централизованной системе управления данными.

В других случаях сбои связаны с технологиями и навыками — когда вам необходимо интегрировать новые технологии в существующую инфраструктуру и реорганизовать способы работы или совершенствовать навыки.

В любом случае необходимо предвидеть сбои, быть готовым к ним и минимизировать последствия или же объяснить их причину.

Что надо принимать во внимание

2

Время получения результатов от проекта

При рассмотрении различных начальных проектов вы совершенно естественно выберете те, которые принесут максимальную пользу и усовершенствование бизнесу. Следует оценить возможные результаты проекта. Когда проект начнет приносить ощутимую пользу?

И, что еще более важно, когда бизнес-пользователи это осознают? Например, вы можете внедрить систему управления мастер-данными для хранилища данных и тем самым резко повысить эффективность инструментов бизнес-аналитики.
Но важность этого шага будет оценена только после того, как бизнес-аналитики поймут, что им больше не нужно очищать финансовые данные.

3

Ресурсы и ограничения

Рассмотрев предыдущие факторы, предположим, что ресурсы находятся в вашем распоряжении. Позже мы поговорим об этом более подробно, но сейчас просто помните о том, что вы хотите, чтобы результаты проекта были более весомыми, чем вложенные в него средства.

Двойственность проекта. С одной стороны, вам необходимо добиться максимального результата проекта для бизнеса. С другой стороны, вам необходимо стратегически распоряжаться бюджетом. Конечно, вам хочется набрать команду специалистов по данным не меньше и не хуже, чем в Google, но сможете ли вы действительно это сделать? Принятие разумных решений, касающихся сотрудников и средств работы, критически важно для успеха проекта.

Тактические проекты по большим данным: Примеры

Область применения больших данных очень широка. Результаты, которых можно добиться с их помощью, впечатляют, но разнообразие применения затрудняет выбор проекта, с которого можно начать. Здесь мы приведем несколько тактических проектов по большим данным, которые реализовывали наши клиенты.

Если вы на данный момент не уверены, с какого проекта вашей организации следует начать, изучите приведенные ниже примеры, и они помогут вам с выбором.

Финансы

- Анализ рисков и портфельный анализ
- Рекомендации по инвестированию

Продажи

- Проактивное сотрудничество с клиентами
- Предоставление услуг с учетом местоположения

Медиасеть

- Отслеживание процессов в режиме реального времени
- Опции перекрестных и дополнительных продаж

Производство

- ПО «подключенных» транспортных средств
- Предупредительное техническое обслуживание

Здравоохранение

- Прогнозирование результатов лечения
- Общая стоимость медицинского обслуживания
- Подбор лекарственных препаратов

Государственный сектор

- Медицинское страхование
- Обмен
- Оптимизация налогов
- Выявление мошенничества

Тактические проекты по большим данным: Примеры

Цели наших клиентов

Посмотрите, насколько четко наши клиенты описывают свои цели. Это — та степень конкретизации, к которой следует стремиться.

- Крупная информационно-технологическая компания из Силиконовой долины поставила цель сократить растущие затраты на хранилище данных более чем на 10 млн долл., используя комбинацию Hadoop и традиционных технологий организации хранилища данных для снижения роста общей стоимости терабайта.

- Крупный производитель транспортных средств поставил цель снизить уровень потребления топлива автомобилями на 1 % за последующие 10 лет. Также эта компания поставила перед собой задачу снизить выбросы автомобилями углекислого газа, увеличив протяженность периода технического обслуживания на 10 % и увеличив пробег между ТО на 1 %.

- Производитель локомотивов намерен повысить среднюю скорость составов на пригородных маршрутах на 1 милю в час для того, чтобы его клиенты могли сократить издержки на 200 млн долл. в год.

- Глобальная платежная система намерена расширить интернет-сектор своего бизнеса на 30 % с помощью повышения персонализации продаж в рамках стратегии по большим данным под названием «оптимизация многоканальных продаж».

Любая команда хороших специалистов по большим данным может добиться серьезного успеха при наличии правильно поставленной цели.

Первое путешествие в мир больших данных

Если вы готовы начать работу с проектом по большим данным на уровне всего предприятия, то вам необходимо выполнить три шага, приведенных ниже.

Даже если вы планируете реализовать целый ряд тактических проектов по большим данным, вам в любом случае придется пройти эти ступени. Каждая из этих ступеней критически важна для фундаментальной целостности организации, применяющей ориентированный на данные подход. Для получения наивысшего эффекта рекомендуется выполнять эти шаги по порядку.

1

Оптимизация хранилища данных

Этот шаг подразумевает выбор для хранения и обработки данных наиболее эффективной в финансовом отношении платформы. Следует начинать с перемещения необработанных или редко используемых данных и рабочих нагрузок ETL за пределы дорогостоящего оборудования хранилища данных.

Так делают, чтобы избежать затратных обновлений хранилища данных и перейти к использованию менее дорогого оборудования и распределенных вычислительных систем, таких как Hadoop, что поможет справиться с объемом, многообразием и высокой скоростью поступления больших данных.

Выбор нужного вашей организации проекта

Первое путешествие в мир больших данных

2

Управляемое «озеро данных»

Управляемое «озеро данных» — это единое пространство для обработки всех входящих и исходящих данных. Ключевое слово здесь — «обработка». Цель обработки — превращение беспорядочной массы данных в надежную, защищенную и готовую к использованию систему информации.

Это достигается путем создания «озера данных», которое очищает, обрабатывает данные и управляет ими. Для выполнения этой непростой задачи требуется изрядная доля дальновидности — придется применять жесткие стратегические политики и процессы управления данными. Без этого ваше «озеро данных» работать не сможет.

3

Средства оперативной бизнес-аналитики в режиме реального времени

Теперь вам необходимо создать технологии (инструменты аналитики, приложения для обработки данных, клиентские интерфейсы), которые необходимы вашим сотрудникам для доступа к данным, их анализа и передачи. Приложения, созданные на этом этапе, должны быть просты в использовании, но вместе с тем предоставлять пользователям всю необходимую им информацию.

Например, для сотрудников клиентской службы таким приложением может стать интерфейс, отслеживающий деятельность клиентов по всем каналам и прогнозирующий отток определенных клиентов в последующие две недели.

Выбор нужного вашей организации проекта

Три основных шага

Как мы уже упоминали, для достижения наибольшей эффективности необходимо выполнить эти три шага по порядку.

Оптимизация хранилища данных

Сократите издержки на обслуживание инфраструктуры и укрепите архитектуру предприятия.

Управляемое «озеро данных»

Создайте единое пространство для обработки всех входящих и исходящих данных.

Аналитика в режиме реального времени

Применяйте новейшие приложения, обеспечивающие сотрудникам доступ к необходимой информации.

Выбор нужного вашей организации проекта

Первое путешествие в мир больших данных

Как наши клиенты определяют свои первоначальные цели

Проекты любого масштаба должны быть ориентированы на конкретную область. В четкой спецификации в этом случае не обязательно указывать сэкономленное время и деньги, но обязательно определить рамки и суть проекта. Ознакомьтесь со следующими примерами проектов по инфраструктуре больших данных.

- Глобальная организация, осуществляющая миллионы финансовых транзакций в сотнях стран, построила концентратор данных предприятия. Он был построен для осуществления анализа больших данных и идентификации ключевых масштабных тенденций и шаблонов во взаимодействиях с клиентами.

- Крупная информационно-технологическая компания создала облачную службу аналитики на уровне всего предприятия для ускорения вывода на рынок новых продуктов, связанных с управлением данными за счет включения новых наборов данных в аналитику всех бизнес-подразделений.

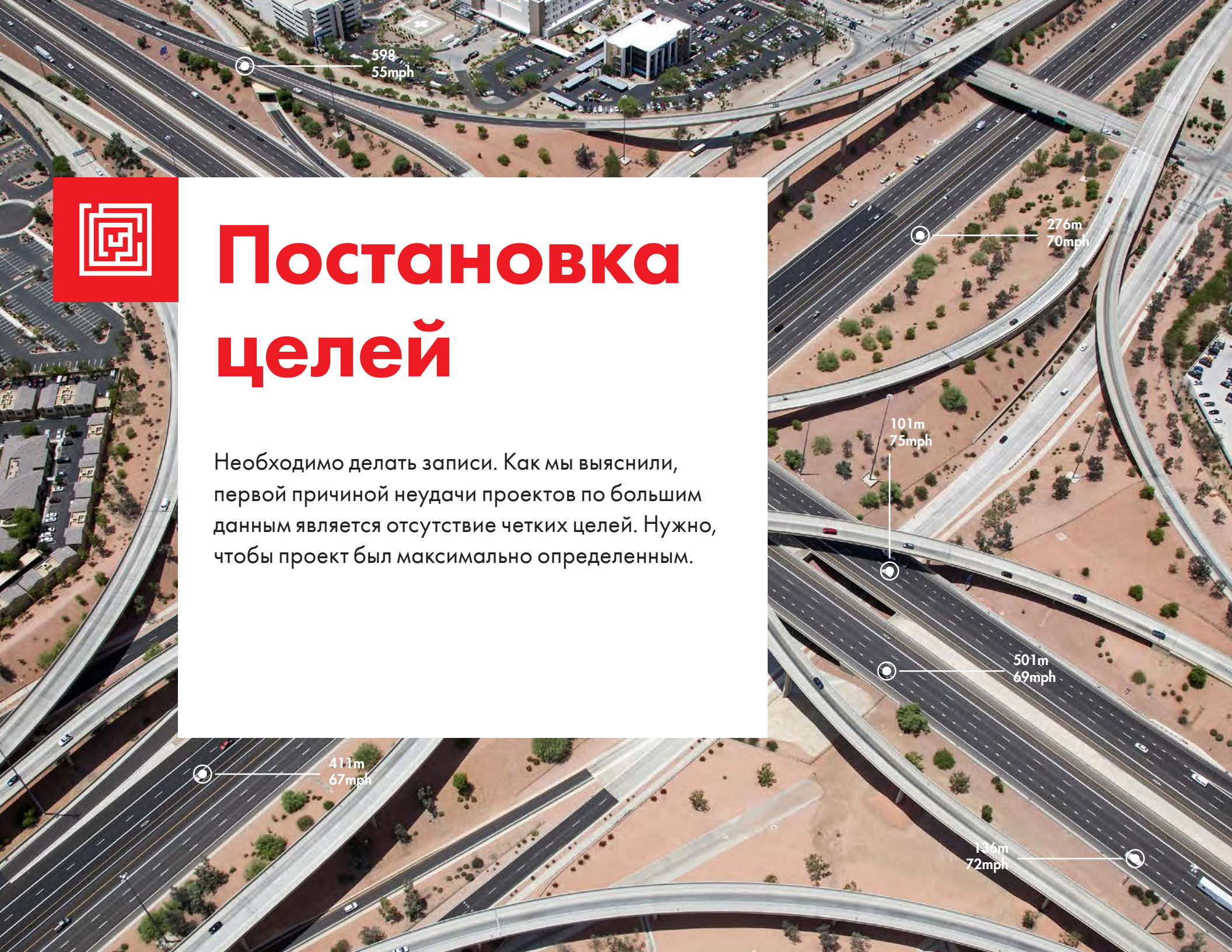
- Глобальная организация, предоставляющая услуги консалтинга в области финансов, создала логические хранилища данных для обеспечения согласованности и доступности информации на всех стандартных платформах (включая Hadoop, операционные базы данных и традиционные хранилища данных) для использования всеми сотрудниками предприятия.

Вывод. Внедрение проекта по большим данным оптимизирует бизнес-процессы, но для этого требуется надежный фундамент.

Часть Б. Стратегия работы

Теперь мы подойдем к вопросу практически и посмотрим на конкретные требования к вашему следующему (или первому) проекту по большим данным.

Назад к содержанию <



Постановка целей

Необходимо делать записи. Как мы выяснили, первой причиной неудачи проектов по большим данным является отсутствие четких целей. Нужно, чтобы проект был максимально определенным.

Бизнес-цели

Начнем с бизнес-целей, поскольку для полного одобрения проекта они должны превалировать над целями ИТ-отдела.

Ставьте цели как можно более точно. Помните, что необходимо ставить цели, приносящие измеримый результат.

Например, в случае с интерфейсом для клиентской службы, прогнозирующим отток клиентов, цель проекта не должна звучать как «повышение качества работы с клиентами».

Чем яснее цель, тем проще ее достигнуть. Пять четко сформулированных целей принесут больше пользы, чем одна размытая.

Бизнес-цели

Перечислите в порядке важности цели вашего проекта по большим данным, относящиеся к бизнесу и бизнес-пользователям. (Вы можете перечислить любое количество целей.)
Например, «Снизить отток клиентов».

Запишите минимальное и максимальное время для достижения каждой цели.
Например, «От 3 до 6 месяцев».

Теперь для каждой цели определите метрики успеха, с помощью которых можно выявить, достигнута ли цель. В идеале эти метрики должны выражаться в числах.
Например, «Снизить среднемесячный отток клиентов на X %».

Сколько времени займет реализация проекта?

Реализация вашего проекта по большим данным должна занять столько времени, сколько это необходимо для получения максимальной пользы. По опыту можно сказать, что временные рамки определяются областью охвата проекта.

Мы работали с такими клиентами, которые реализовывали тактические проекты менее чем за три месяца — другие тратили три года на разработку фундаментальных программ.

Помните, что если реализация проекта занимает много времени, то вы должны демонстрировать достигнутые результаты каждые 6 месяцев. Если вы используете динамический подход в реализации проекта, то вам легче представить различные фазы проекта в качестве отдельных, менее крупных проектов.

Но вы не должны угадывать, сколько времени займет реализация. Рассчитайте необходимое время на основе своего опыта и опыта тех, кто уже занимался реализацией подобных проектов. Вы можете обратиться за помощью к нам.

Цели ИТ

Теперь посмотрим на цели ИТ, относящиеся к вашему проекту.

Помните о том, что если ваш проект направлен только на повышение скорости и качества работы ИТ-отдела, то будет очень сложно продать его бизнес-пользователям. Именно поэтому цели ИТ должны быть тесно связаны с целями, которыми уже заинтересовались ваши бизнес-пользователи.

Перечислите в порядке важности цели вашего проекта по большим данным, относящиеся к ИТ. (Вы можете перечислить любое количество целей.)

Например, отладка процессов сбора, очистки, хранения и управления объединенными данными о клиентах, кредитных картах, социальных графах и показателях оттока клиентов в режиме реального времени.

Развивайте сотрудничество

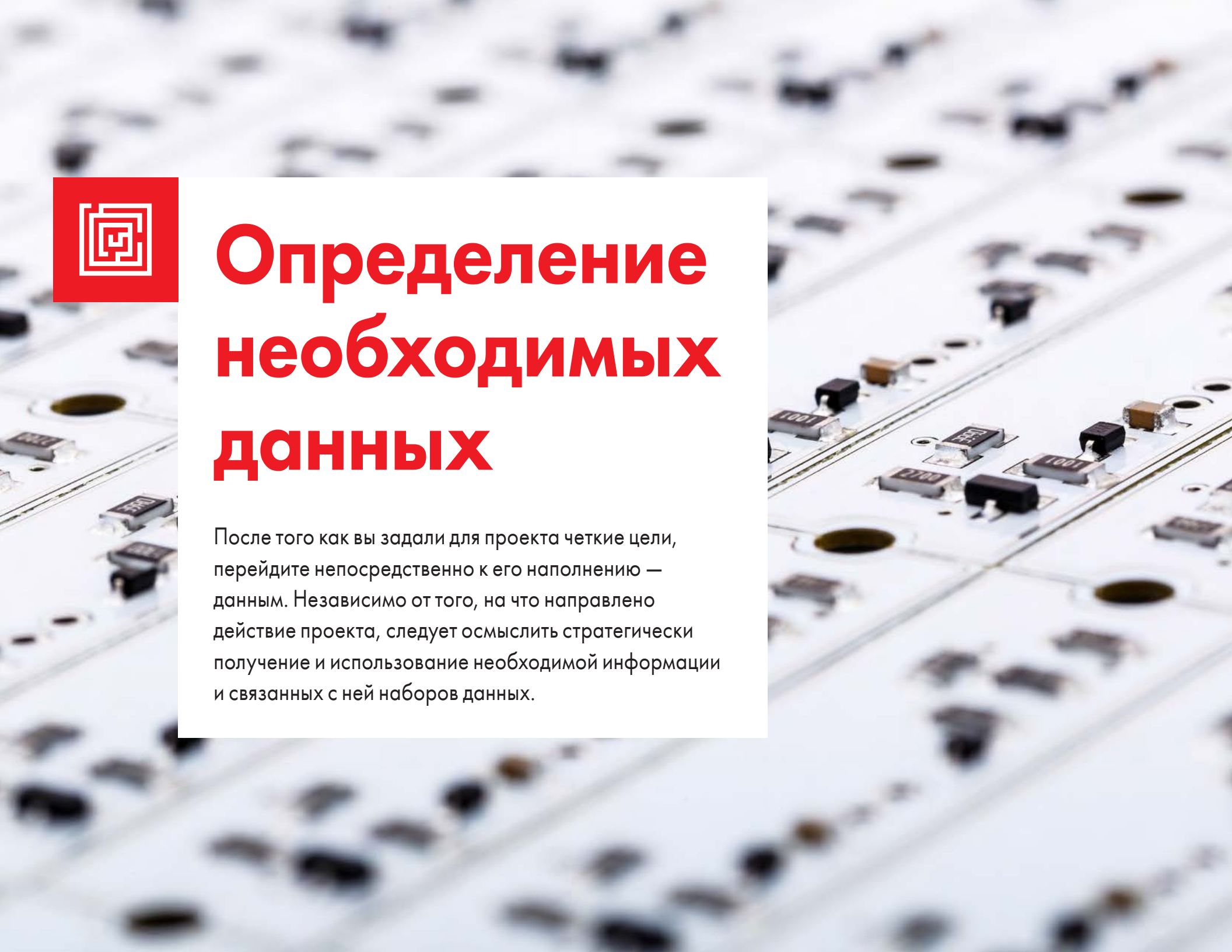
Это руководство было создано для того, чтобы вы могли запустить проект по большим данным независимо от того, где вы работаете — в бизнес- или ИТ-подразделении. В любом случае тщательно обдумайте свои цели. Если вам необходима помощь в постановке целей, проконсультируйтесь с партнерами, имеющими необходимый опыт.

Успех проекта без стратегического сотрудничества невозможен.

Цели ИТ

Запишите минимальное и максимальное время для достижения каждой цели.
Например, «2 или 4 месяца».

Теперь для каждой цели определите метрики успеха, с помощью которых можно выявить, достигнута ли цель. В идеале эти метрики должны выражаться в числах, например: «X % точных прогнозов по оттоку клиентов».



Определение необходимых данных

После того как вы задали для проекта четкие цели, перейдите непосредственно к его наполнению — данным. Независимо от того, на что направлено действие проекта, следует осмыслить стратегически получение и использование необходимой информации и связанных с ней наборов данных.

Какие данные вам необходимы?

Во-первых, вернемся к основной цели вашего проекта и информации, которую с его помощью вы предоставите организации. Ответьте на приведенные ниже вопросы максимально точно.

Чтобы проект достиг поставленных бизнес-целей, что бизнес-пользователи должны знать о нем для вынесения обоснованного решения?
Например, какие из наиболее ценных клиентов собираются прекратить сотрудничество с компанией и почему.

Какие данные могут быть использованы для получения этой информации?
Например, данные об истории операций с клиентом, данные по отзывам, частоте операций, о несостоявшихся телефонных разговорах, частоте отказов, качестве клиентского сервиса.

Какие данные вам необходимы?

Какие исходные системы содержат эти наборы данных?
Например, записи клиентской службы, эксплуатационные характеристики продукта, база данных об операциях клиентов, управление мастер-данными о клиентах.

Существуют ли другие данные, кроме вышеупомянутых, которые могут обеспечить контекст или повысить ценность вашей аналитики?
Например, данные опросов клиентской службы, анализа конкурентов, метеорологические, социальные данные.

Какие данные вам необходимы?

Какие наборы данных, к которым пока нет доступа, могут содержать дополнительные контекстные данные?

Например, сторонние социальные данные, сторонние данные рынка, метеорологические данные.

Поиск темных данных

Базы данных, к которым вы не имеете доступа, находятся не только за пределами вашей организации. Компания Gartner установила, что большинство предприятий используют только 15 % данных, находящихся в пределах организации[2]. Appfluent, компания, осуществляющая статистический анализ использования хранилищ данных, выяснила, что, как правило, от 30 % до 70 % данных хранилища находится в «спящем» режиме.

Остальные скрыты в труднодоступных, дорогих в эксплуатации и разрозненных источниках и архивах данных. К сожалению, вам приходится платить за хранение всех этих данных.

При поиске нужных данных следует начать с тех, которые уже существуют в базах данных организации.

[2] Сайт Gartner: www.gartner.com/technology/topics/big-data.jsp

Пять основных аспектов обработки данных

После того как вы определили необходимые для проекта данные, становится понятно, с какими трудностями вам придется столкнуться при реализации проекта по большим данным. В частности, существует пять ключевых факторов, которые следует учесть перед дальнейшим продвижением проекта, поскольку они определяют действия для каждого набора данных, включая набор больших данных.

1

Большие объемы данных

Вам нужно быть готовым к работе с огромными объемами необходимых данных. Классифицируйте ваши данные на основе их ценности (клиентские транзакции), использования (частота доступа), размера (гигабайты, терабайты), сложности (машинные данные, реляционные данные, видеофайлы), прав доступа (только специалисты по данным или все бизнес-пользователи).

Тщательная систематизация данных поможет понять, как обработать весь объем. Оцените свои возможности для хранения и обработки данных и найдите наиболее эффективные (в том числе и в финансовом отношении) средства их масштабирования.

Пять основных аспектов обработки данных

2

Разнообразие данных

Наиболее проблемный аспект больших данных — это многообразие форматов и структур, которые нужно будет свести воедино для анализа. Вам будет необходимо объединить огромное число источников новых типов и структур данных (социальные данные, сенсоры, видео) с уже используемыми источниками (реляционные, устаревшие мейнфреймы).

Кодирование вручную при каждой интеграции займет все время и ресурсы, которыми вы располагаете. Оптимально используйте все доступные вам инструменты интеграции и управления качеством данных для повышения эффективности и ускорения процесса.

3

Регулируйте скорость процесса

Поступающие в реальном времени данные вместе с уже имеющимися данными обычно способствуют повышению эффективности средств аналитики. Итак, некоторые из необходимых вам данных будут ценными только в том случае, если они непрерывно поступают в систему.

Большинство средств аналитики, работающих в режиме реального времени, используют непрерывно поступающие данные — часто из различных источников и в различных форматах. Используйте в вашем проекте технологии потоковой аналитики и логическую инфраструктуру для того, чтобы обрабатывать полный объем данных.

Пять основных аспектов обработки данных

4

Обеспечьте достоверность данных

Ценность вашего анализа будет сведена к нулю, если используемые данные не будут достоверны. Чем больше данных используют для анализа, тем более важно поддерживать высокий уровень качества данных.

Надо хорошо знать цели, чтобы использовать данные в соответствии с ними. Если специалисты по данным намерены найти шаблоны в объединенных данных о клиентах, необходимая подготовка данных будет минимальной.

С другой стороны, данные финансовых отчетов и цепи поставок должны быть тщательно отобраны, очищены и проверены на точность и соответствие нормативам.

Создайте категории, основанные на необходимом объеме подготовки данных — от необработанных данных до тщательно подготовленного хранилища, содержащего очищенные, надежные и соответствующие нормативам данные.

5

Соответствие нормативам

Разнообразные наборы данных, с которыми вы будете работать, обладают различными условиями и требованиями безопасности. Вам нужно будет решить, как анонимизировать каждый набор данных в соответствии с политикой безопасности компании.

Огромные массивы данных могут содержаться в сотнях хранилищ по всему предприятию. Выясните, где находятся конфиденциальные данные, и защитите их в источнике путем шифрования и последующего контроля доступа.

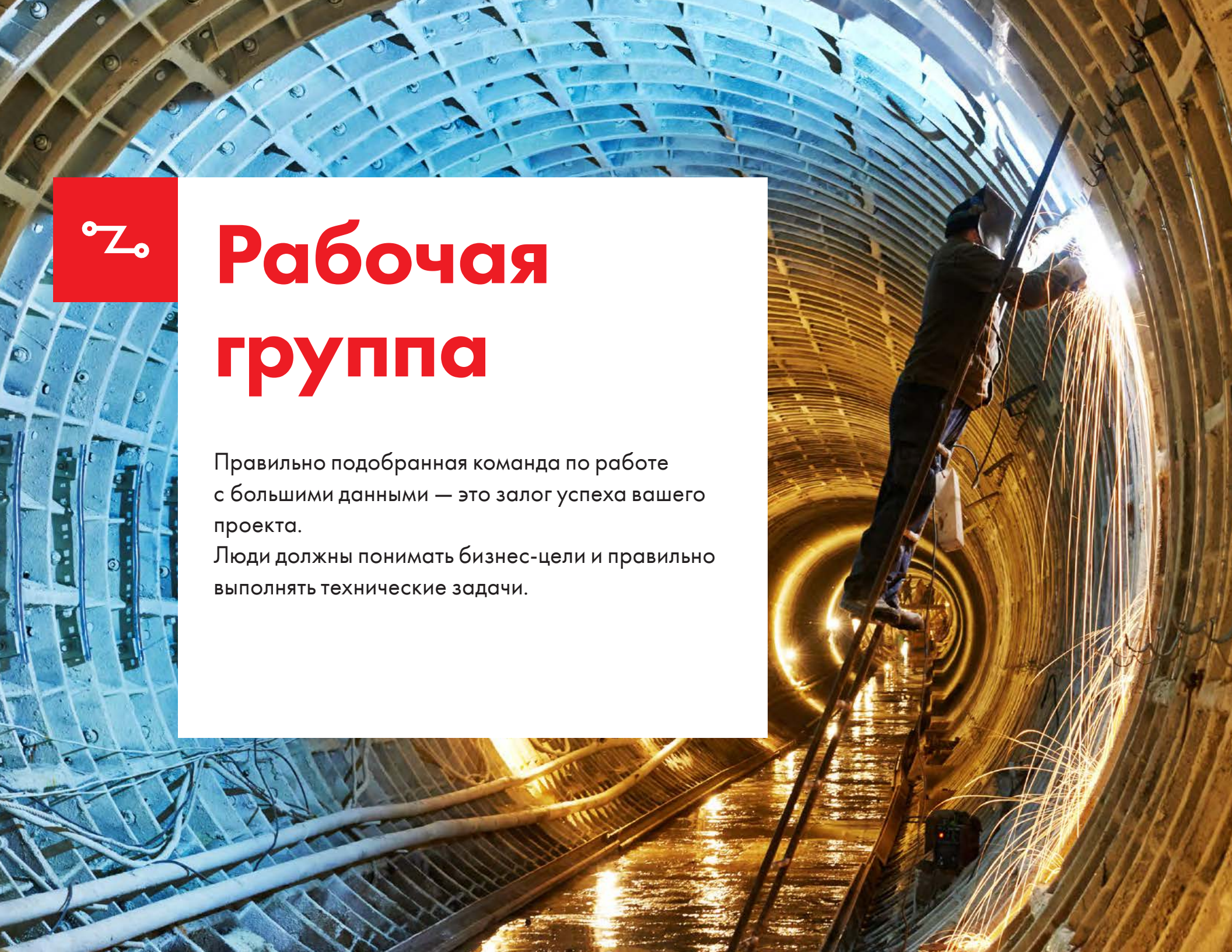
Помимо обеспечения безопасности конфиденциальных данных и их архивации, необходимо маскировать их по заранее заданным правилам каждый раз при их переносе или входе в среду разработки или тестирования.

Учитывайте эти пять аспектов при работе с каждым набором данных, и вы будете подготовлены к реальной работе с большими данными.

Часть B. Канал передачи больших данных

Традиционные средства бизнес-аналитики и методы хранения данных не могут быть масштабированы таким образом, чтобы соответствовать требованиям проектов по большим данным. Рассмотрим, как вы можете расширить рабочую группу, процессы и инфраструктуру.

Назад к содержанию <



Рабочая группа

Правильно подобранная команда по работе с большими данными — это залог успеха вашего проекта.
Люди должны понимать бизнес-цели и правильно выполнять технические задачи.

Пять советов по построению эффективной рабочей группы

Большинство организаций недооценивает уровень навыков, необходимых для успешного использования таких новых технологий, как Hadoop.

Распределенными системами данных достаточно сложно управлять. Для того чтобы проект был реализован успешно, вам будет необходимо привнести в него множество новых умений и навыков — от навыков работы с Java, использующихся для разработки в Hadoop, до совершенно новых навыков в области науки о данных, и вам придется нанимать новых специалистов.[3]

Когда вы приступите к созданию рабочей группы, обязательно примените следующие советы к стратегии по набору кадров.

[3] "Hadoop, Python, and NoSQL lead the pack for big data jobs", InfoWorld, May 5, 2014: www.infoworld.com/t/it-jobs/hadoop-python-and-nosql-lead-the-pack-big-data-jobs-241884)

Пять советов по построению эффективной рабочей группы

1

Знания и навыки приглашенных вами специалистов — ценное преимущество

Одна из грубейших ошибок — приглашать специалистов по данным и количественному анализу и предлагать им выполнять несоответствующие их квалификации задачи. Когда наиболее квалифицированные сотрудники тратят свое время на кодирование интеграций и очистку данных вручную, вы не только разочаровываете их ожидания — вы теряете преимущество от использования редких профессиональных навыков.

Задействуйте этих специалистов для выполнения таких задач, которые действительно требуют их квалификации. Не стоит терять лучших сотрудников или тратить их время впустую. Лучше автоматизировать подобные операции.

2

Подойдите к созданию рабочей группы стратегически

Если все получится, то проект будет масштабирован и вам будет доступно больше ресурсов. Планируйте ваши действия стратегически, чтобы в определенное время не потерять скорости масштабирования необходимых процессов из-за отсутствия достаточного количества сотрудников с необходимыми вам навыками — даже в Силиконовой долине.

Если проект будет расширен, какие специалисты вам потребуются и в какой момент? Например, специалистов в области данных (data scientists) гораздо сложнее найти и обучить, чем разработчиков.[4]

Сбалансированность рабочей группы критически важна. Вам надо найти людей с уникальным сочетанием опыта в управлении данными и энтузиазма в изучении новых инструментов. В дополнение к этому нужно иметь примерно одинаковое количество сотрудников с техническими навыками и предметных специалистов для построения необходимых моделей.

4. "Big Data's High-Priests of Algorithms," Wall Street Journal, August 8, 2014: http://online.wsj.com/articles/academic-researchers-find-lucrative-work-as-big-data-scientists-1407543088

Пять советов по построению эффективной рабочей группы

3

Определите цели проекта заранее и объясните их своим сотрудникам

Одна из частых ошибок, которые допускают компании при приеме на работу новых сотрудников, заключается в том, что новым сотрудникам не сообщают истинные цели проекта. С первого собеседования и до ввода непосредственно в проект сотруднику должно быть предельно ясно, какой проект вы планируете для бизнес-пользователей. Пользуйтесь поддержкой руководителей, чтобы продвигать проект, и сообщайте им не только о проблемах, но и об успехах проекта.

Без полного и четкого понимания ценности проекта для бизнеса новые сотрудники будут думать только об ИТ-целях проекта.

4

С увеличением рабочей группы возрастает сложность управления

Если новая технология может быть развернута, внедрена и интегрирована в систему и она начинает работать, то новым сотрудникам сначала необходимо ознакомиться с целями и задачами проекта. Руководителю проекта придется уделять больше времени руководству командой новых сотрудников.

Сплоченность коллектива очень важна при работе над проектом. Вам следует подумать, как интегрировать новых сотрудников в рабочие процессы. Может быть, вы не сможете обучить их определенным навыкам, но в ваших силах помочь им быстро влиться в команду.

Пять советов по построению эффективной рабочей группы

5

Ваша рабочая группа должна развиваться

Появляются все новые технологии по работе с большими данными. А уже существующие стремительно развиваются. Сейчас наступило время развития для тех компаний, которые осмелятся взять на вооружение самые новые технологии. Но при этом нельзя забывать о существующей в этой области конкуренции.

Ваши сотрудники должны развивать свои навыки так же быстро, как изменяется мир вокруг них. Однако ничто так не мотивирует лучших работников, как стремление быть «впереди планеты всей». Надо организовывать обучение сотрудников и обсуждать с ними проблемы. Это необходимо для развития их навыков — и, как следствие, ваших возможностей.

Необходимо мыслить стратегически

Вопрос, который вам нужно решить — как осуществлять интеграцию, вручную или автоматизированными инструментами.

Кодирование вручную предоставляет вам полный контроль над тем, что вы создаете. Иногда кодирование вручную неоценимо и необходимо, например в случаях, когда требуется написание сложного сценария для извлечения метаданных не существовавшим ранее способом.

Применение автоматических инструментов, с другой стороны, предоставляет гораздо большую маневренность и способность многократного повторения одного и того же процесса. Для таких задач, как интеграция данных и управление качеством данных, требуется применение автоматических инструментов, поскольку оно избавляет высококвалифицированных специалистов от необходимости выполнять работу вручную.

Реалистично оценивайте ваши ресурсы. Если вы не можете создать такую же крупную и талантливую рабочую группу, как в Google, не тратьте свои и без того ограниченные ресурсы на попытки это сделать.

Рабочая группа

Организация структуры управления данными

Когда вы основательно подойдете к реализации проектов по большим данным, нужно будет создать рабочую структуру для управления данными. Если ваш проект по большим данным нацелен на интересы одного подразделения, вы можете создать небольшой совет по вопросам управления данными для того, чтобы справляться с возможными проблемами в будущем.

Ваш совет по вопросам управления данными будет представлять собой группу руководителей, задача которых состоит в наблюдении за процессом управления данными на предприятии. Также эта группа должна включать в себя специалистов по управлению данными — специалистов от подразделений, отвечающих за обработку данных, поступающих от конкретных бизнес-подразделений.

(Наши клиенты распределяют роли управляющих данными в зависимости от домена данных. Это означает, что один сотрудник отвечает за данные о продуктах, другой — за данные о клиентах и так далее.)

Рабочая группа

Организация структуры управления данными

Вам следует стремиться к созданию процессов, которые позволят совету по вопросам управления данными успешно функционировать. Нужно избегать бюрократии в работе совета, чтобы каждый вовлеченный сотрудник следовал общим целям и соблюдал временные рамки.

Совет по вопросам управления данными должен обладать пятью характеристиками.

1

Кросс-функциональность

Совет по управлению данными, объединяющий различных людей с одинаковыми обязанностями, не будет эффективным. Цель состоит в том, чтобы создать команду, где будут охвачены цели и потребности каждого бизнес-подразделения, интересы которых представляет проект по большим данным.

2

Коммуникабельность

Без коммуникации между всеми участниками совета ваш проект обречен на неудачу из-за бюрократии и недопонимания. Такие случаи встречаются очень часто. Следите за тем, чтобы все разногласия разрешались успешно и вовремя.

Организация структуры управления данными

3

Эффективность

Кросс-функциональность процесса не должна препятствовать вашему проекту. Наоборот, это придаст вашему проекту дополнительную маневренность. Поэтому автоматизируйте все возможные процессы и применяйте инструменты совместной работы, чтобы сохранять сферу общения открытой и эффективной.

4

Следование целям проекта

Донесите основные цели проекта до всех сотрудников и сделайте так, чтобы каждый стремился к достижению этих целей. Наличие общих целей помогает процессу.

5

Централизация

Наибольшие проблемы в совете по вопросам управления данными появляются при постановке целей одного из бизнес-подразделений выше целей остальных. Убедитесь в том, что ваши решения в перспективе принесут пользу всем бизнес-подразделениям, даже если сейчас они выгодны лишь для одного из них.

Навыки, которые вам нужны, и навыки, которыми вы обладаете

Необходимо делать записи. Сейчас, когда вы видите все возможные препятствия, с которыми можно столкнуться, и возможности, которые может открыть перед проектом функциональная рабочая группа, давайте рассмотрим, что она должна собой представлять.

На следующей странице приведен список должностей сотрудников, работающих с большими данными, аналогичный тому, который используют наши клиенты. Основываясь на данных о сотрудниках, которые работают в компании, и времени, которое планируется затратить на реализацию проекта, (как указано в начале раздела на странице 24), рассчитайте количество сотрудников для найма.

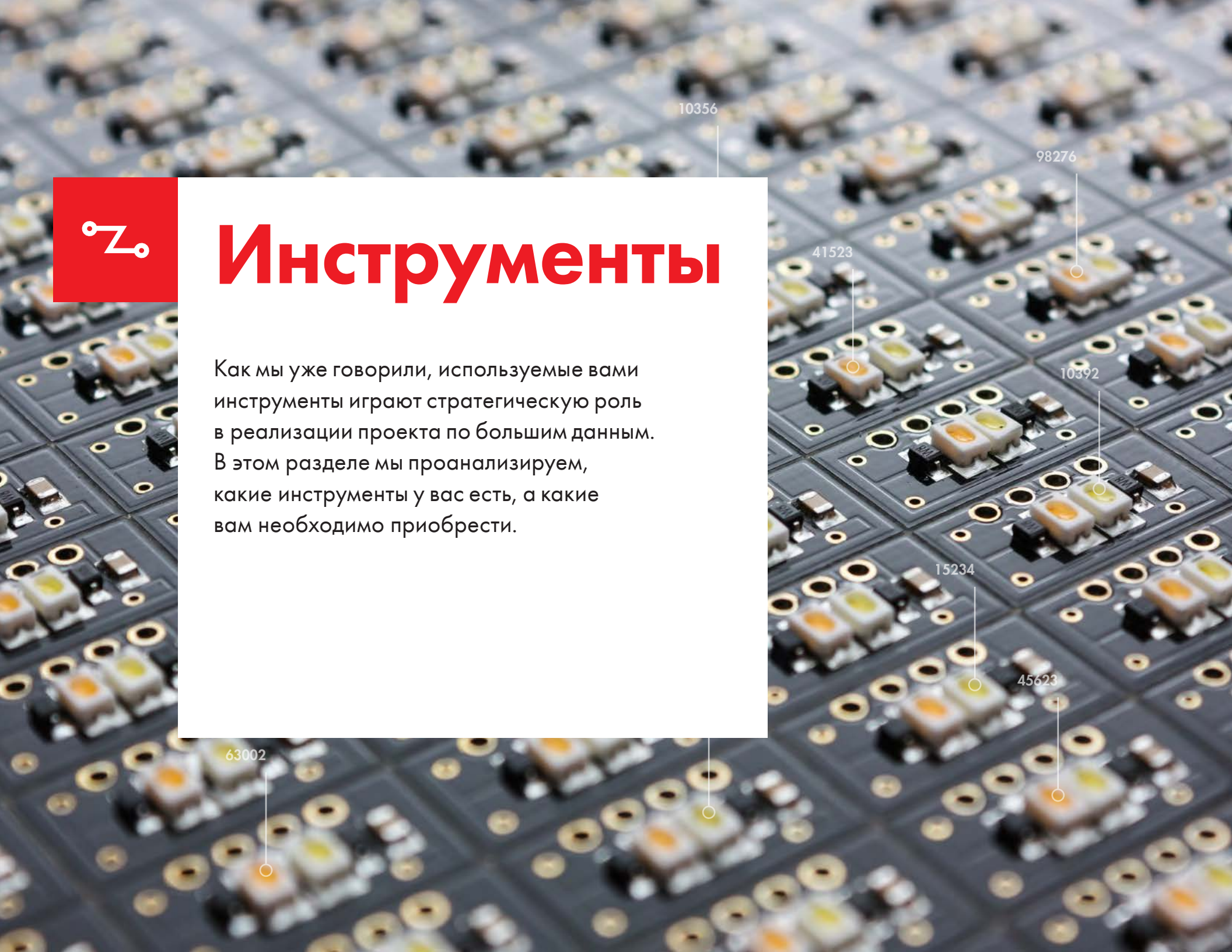
Рабочая группа

Должность	Может ли кто-либо занять эту должность сейчас?	Необходимо нанять на эту должность нового сотрудника	Основываясь на количестве времени, мне необходимо нанять X сотрудников
Специалист по данным	✓ или ✕	✓ или ✕	
Эксперт по доменам			
Бизнес-аналитик			
Аналитик данных			
Инженер данных			
Администратор баз данных			
Архитектор предприятия			
Архитектор решений для бизнеса			
Архитектор данных			
Управляющий данными			
Разработчик ETL (решений по интеграции данных)			
Разработчик приложений			
Разработчик панели управления			
Разработчик модели данных			
Другое			
Другое			
Другое			
Другое			
Другое			

Необходимость в комплексном мышлении

При подборе новых сотрудников не ограничивайтесь поиском людей только с нужными навыками. Поиск людей только с нужными навыками уже сам по себе является трудной задачей. Подбирайте сотрудников, способных следовать бизнес-целям и имеющих необходимые технические навыки.

Очень важно, чтобы приходящие в проект по большим данным сотрудники были способны понять реалии бизнеса и выполнить сложные задачи по изучению данных. Людей с таким комплексным мышлением достаточно сложно найти. Комплексному мышлению нужно обучать, и оно этого стоит.



Инструменты

Как мы уже говорили, используемые вами инструменты играют стратегическую роль в реализации проекта по большим данным. В этом разделе мы проанализируем, какие инструменты у вас есть, а какие вам необходимо приобрести.

Общие сведения об инструментах для работы с большими данными

Основываясь на собственном опыте, мы приводим список основных инструментов для создания архитектуры, необходимой для проектов по большим данным (архитектуру мы обсудим более подробно позже). Набор технологий, необходимых для реализации вашего конкретного проекта, должен определяться целями этого проекта и имеющимися ресурсами.

Ознакомьтесь со списком и поставьте ☒ напротив наиболее важных для вашего проекта инструментов.

Получение данных

Процесс получения необходимых данных должен соответствовать потребностям проекта, быть эффективным и последовательным.

- ○ **Пакетная загрузка.**
  Есть ли у вас доступ ко всем необходимым типам данных и можете ли вы эффективно увеличить производительность пакетной загрузки в ваши хранилища данных?

- ○ **Сбор измененных данных**
  Можете ли вы отслеживать изменения данных в исходных системах, не затрагивая при этом сами системы?

- ○ **Потоковая передача данных**
  Можете ли вы собрать надежные данные в режиме реального времени и передать их в хранилища в потоковом режиме?

- ○ **Архивация**
  Можете ли вы архивировать и сжимать данные, которые используются очень редко, сохраняя при этом простоту доступа к архивированным данным в случае необходимости?

Общие сведения об инструментах для работы с большими данными

Ознакомьтесь со списком и поставьте ☒ напротив наиболее важных для вашего проекта инструментов.

Управление данными

Все политики, процессы и практические методы, необходимые для управления эффективностью, точностью, надежностью и доступностью ваших данных.

- ○ **Интеграция данных**
  Можете ли вы подготовить и объединить различные структуры и источники в один целостный набор данных для анализа?

- ○ **Качество данных**
  Можете ли вы надежно очистить ваши данные, устранить дублирование и ошибки?

- ○ **Безопасность данных**
  Можете ли вы найти и защитить данные, находящиеся во всех хранилищах, путем задания правил использования, предоставления доступа и разрешений?

- ○ **Система виртуальных данных**
  Можете ли вы создать слой абстракции для ваших данных, который мягко соединит обработку данных и окружение развертывания?

- ○ **Управление мастер-данными**
  Обладаете ли вы консолидированной, полной и единственно верной версией данных для различных доменов?

- ○ **Распределенная система данных**
  Можете ли вы использовать технологии, подобные Hadoop, для эффективного масштабирования и обработки данных?

- ○ **Хранилище данных**
  Обладаете ли вы такой технологией хранилища данных, которая способна удовлетворить требования к производительности, функциональности и масштабированию, необходимые для анализа больших данных и интеграции с инфраструктурами Hadoop?

Общие сведения об инструментах для работы с большими данными

Ознакомьтесь со списком и поставьте ☒ напротив наиболее важных для вашего проекта инструментов.

Доставка данных

Процесс отправки имеющихся данных в работающие с ними системы и приложения.

- ○ **Пакетная загрузка**
  Можете ли вы эффективно масштабировать пакетную загрузку данных в различные источники, средства аналитики и серверные операционные системы?

- ○ **Потоковая передача данных в реальном времени**
  Можете ли вы обеспечить передачу данных в потоковом режиме в различные приложения, средства аналитики и серверные системы?

- ○ **Концентратор для интеграции данных**
  Можете ли вы обеспечить доступность данных, используя модель «публикация и подписка», чтобы избежать двухточечных интеграций?

- ○ **Виртуализация данных**
  Можете ли вы доставить данные из систем без их перегрузки?

- ○ **Обработка событий**
  Можете ли вы обнаружить угрозы, возможности и другие ключевые события в режиме реального времени, проанализировать их и принять соответствующие меры?

Общие сведения об инструментах для работы с большими данными

Ознакомьтесь со списком и поставьте ☒ напротив наиболее важных для вашего проекта инструментов.

Аналитика

Инструменты и процессы, превращающие необработанные данные в аналитические выводы, шаблоны, прогнозы и вычисления для домена, который вы анализируете.

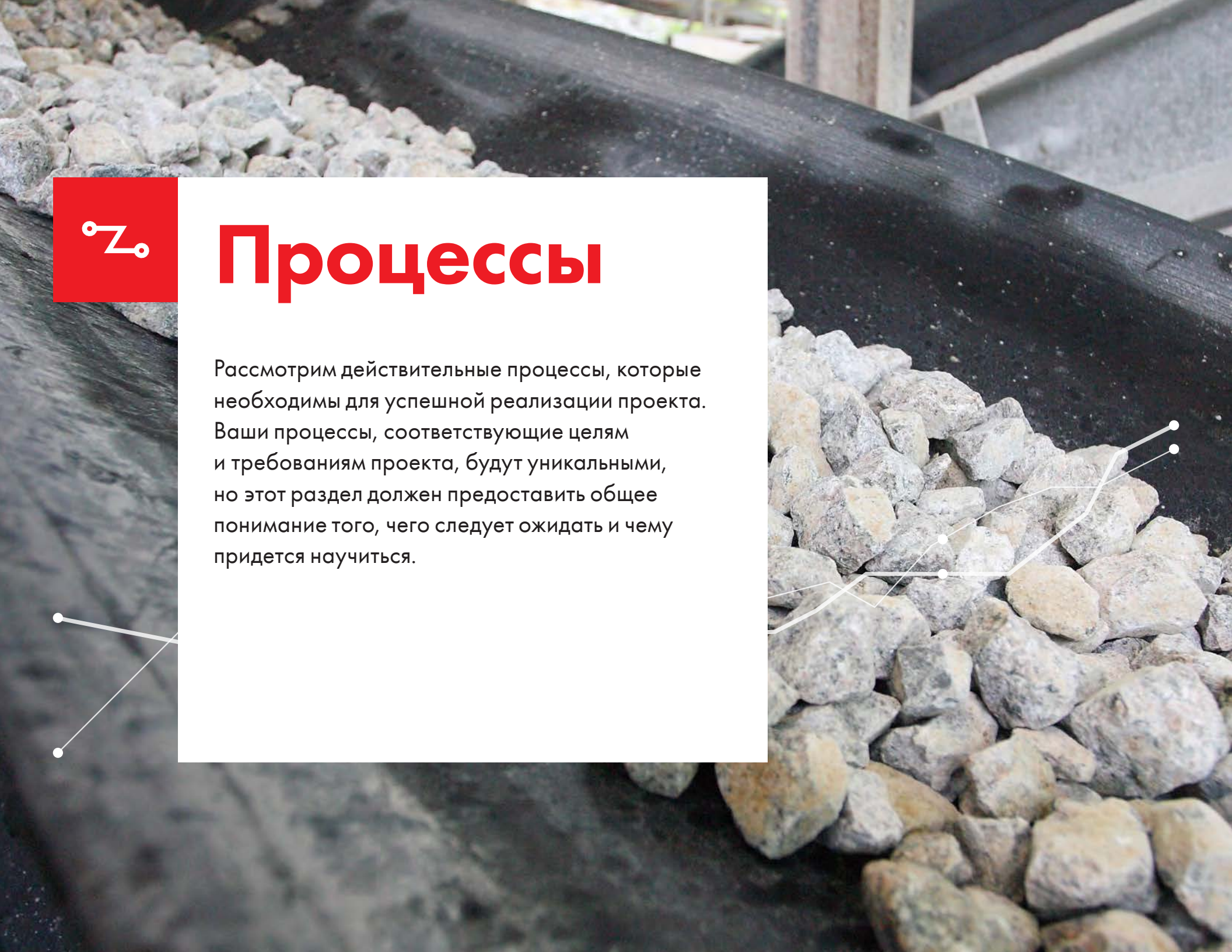
○ **Визуализация**
Можете ли вы представить свои данные и аналитические выводы в доступной для понимания форме?

○ **Средства расширенной аналитики**
Применяете ли вы к своим наборам данных передовые алгоритмы аналитики для проведения сложных вычислений?

○ **Машинное обучение**
Можете ли вы применить сложные алгоритмы машинного обучения для поиска шаблонов и прогнозирования на соответствующем уровне?

Некоторые из этих инструментов и технологий особенно важны. К их числу относятся интеграция данных, управление качеством данных и управление мастер-данными. Они настолько фундаментальны, что изменять их не имеет смысла. Количество времени и ресурсов, необходимых для построения этих технологий своими руками, не оправдывают себя при выполнении вашего проекта по большим данным.

Вспомните цели проекта — они не включали построение всего своими руками.



Процессы

Рассмотрим действительные процессы, которые необходимы для успешной реализации проекта. Ваши процессы, соответствующие целям и требованиям проекта, будут уникальными, но этот раздел должен предоставить общее понимание того, чего следует ожидать и чему придется научиться.

Восемь этапов работы с большими данными

Основываясь на собственном опыте, мы можем предложить ряд эффективных методик, которые станут прочной основой для проектов по большим данным. Они позволят вам реализовать ожидания, учиться на ошибках и постоянно совершенствовать рабочие процессы. Как уже было сказано, подход к проекту целиком зависит от вас.

В любом случае восемь этапов работы над проектом станут основополагающими для любого канала передачи больших данных. В процессе работы убедитесь, что ваша команда и вы лично создаете эффективные процессы для каждого этапа.

1

Ввод и вывод данных

Первой задачей станет получение всех необходимых данных. В некоторых случаях это означает прием передаваемых в потоковом режиме данных, в других — извлечение данных из базы. Создайте повторяемые, управляемые процессы, чтобы эти данные затем распределялись по хранилищам в соответствии с предполагаемыми условиями использования.

2

Интеграция данных

Наиболее сложная задача в работе с большими данными связана с многообразием структур и форматов данных. Для успешного проведения анализа данных необходимо создать процесс, позволяющий интегрировать и нормализовать все эти данные. Объем обработки данных вручную при этом должен быть минимальным.

Восемь этапов работы с большими данными

3

Очистка данных

Для того чтобы результаты анализа были достоверными, вам необходимо очистить данные: удалить ошибки, дублированные, неточные и неполные данные. Этот процесс избавит наиболее ценных аналитиков и специалистов от рутинной работы.

4

Обработка данных

Одним из способов обеспечения надежного источника чистых и согласованных данных является создание процесса их обработки. Целью этого будет формирование полного набора консолидированных данных, распределенных по доменам (например, продукты, клиенты и т. п.) и обогащенных анализом больших данных. Такие данные могут быть использованы во всех ваших системах.

5

Защита данных

Теперь вам необходимо создать два основных процесса. Первый заключается в задании правил безопасности и шаблонов действий, применяемых к каждому набору данных. Второй — в обнаружении конфиденциальных данных и их маскировании, постоянном или динамическом, для обеспечения соответствия правилам безопасности.

Восемь этапов работы с большими данными

6

Анализ данных

Процесс анализа данных зависит от вашего аналитика, применяемых инструментов и предъявляемых требований. Если вы хотите оптимизировать этот процесс и сделать его более быстрым, дешевым и эффективным по времени, критически необходимо его постоянно совершенствовать.

7

Анализ потребностей бизнеса

Этот шаг очень важен, но вместе с тем его часто упускают из поля зрения. Создайте четко организованный процесс для анализа потребностей бизнеса, даже если вы анализируете свои данные. Если вы потеряете связь с потребностями бизнеса, вы рискуете изолировать свой проект и минимизировать его пользу для бизнеса.

8

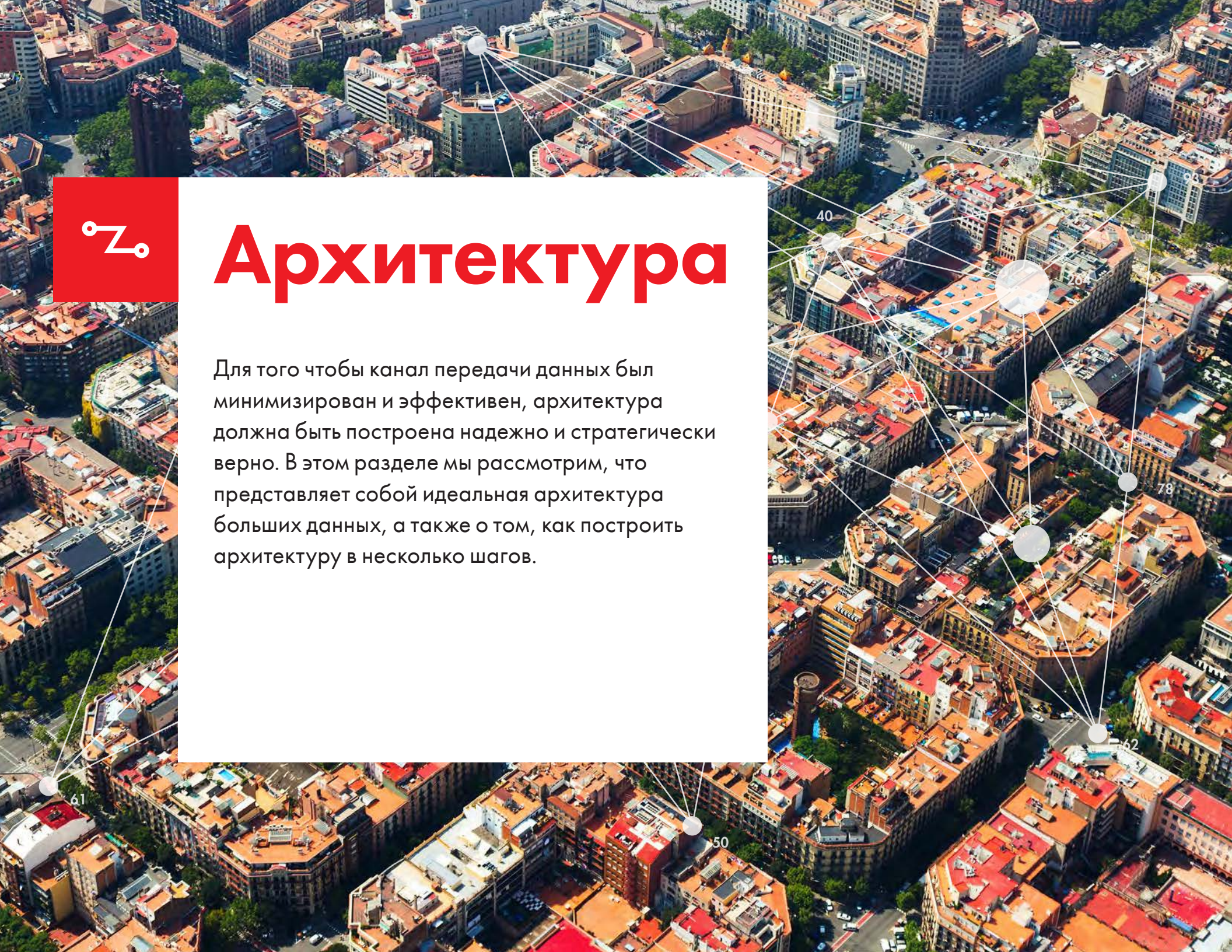
Применение полученных знаний

Как мы уже говорили раньше, польза, приносимая вашим проектом по большим данным для бизнеса, должна быть ощутимой. Создайте автоматизированные каналы для передачи результатов работы вашего проекта тем бизнес-пользователям, которые больше всего в них нуждаются. Например, данные о клиентах, которые с наибольшей вероятностью компания может потерять, должны быть доступны сотрудникам клиентской службы через панель управления. Также обязательно создайте канал обратной связи с бизнес-пользователями.

Необходимость документирования

Освойте эти восемь этапов, и ваш проект по большим данным будет развиваться в верном направлении. Ваша цель — создать четко организованные, повторяемые, масштабируемые, постоянно совершенствующиеся процессы. Документирование этих процессов и постоянное их совершенствование особенно важны для вашей рабочей группы.

Навыки, возможности и уроки, извлеченные из текущего проекта, должны быть легко применимы к другим.



Архитектура

Для того чтобы канал передачи данных был минимизирован и эффективен, архитектура должна быть построена надежно и стратегически верно. В этом разделе мы рассмотрим, что представляет собой идеальная архитектура больших данных, а также о том, как построить архитектуру в несколько шагов.

Первые шаги. «Песочница»

Когда вы начинаете создание архитектуры проекта по большим данным, наиболее логичным началом будет создание «песочницы», в которой вы можете использовать тестовые данные для проверки работоспособности вашей архитектуры. Примите во внимание несколько советов.

Начните с малого

При создании четко обозначенной, полностью контролируемой «песочницы» вы сможете последовательно прийти к наиболее успешному построению архитектуры. Когда вы начнете построение, документируйте все знания, которые вы получите на каждом этапе.

Размер имеет значение

Ключевая разница между «песочницей» и настоящей реализацией проекта заключается в том, что производственная среда будет намного масштабней. Она потребует автоматизированных процессов получения, объединения, очистки данных и распространения выходных данных. Потребуется более жесткая структура, надежные компоненты и процессы, устойчивые в условиях постоянно изменяющейся производственной среды.

Первые шаги. «Песочница»

Маскируйте данные перед тестированием

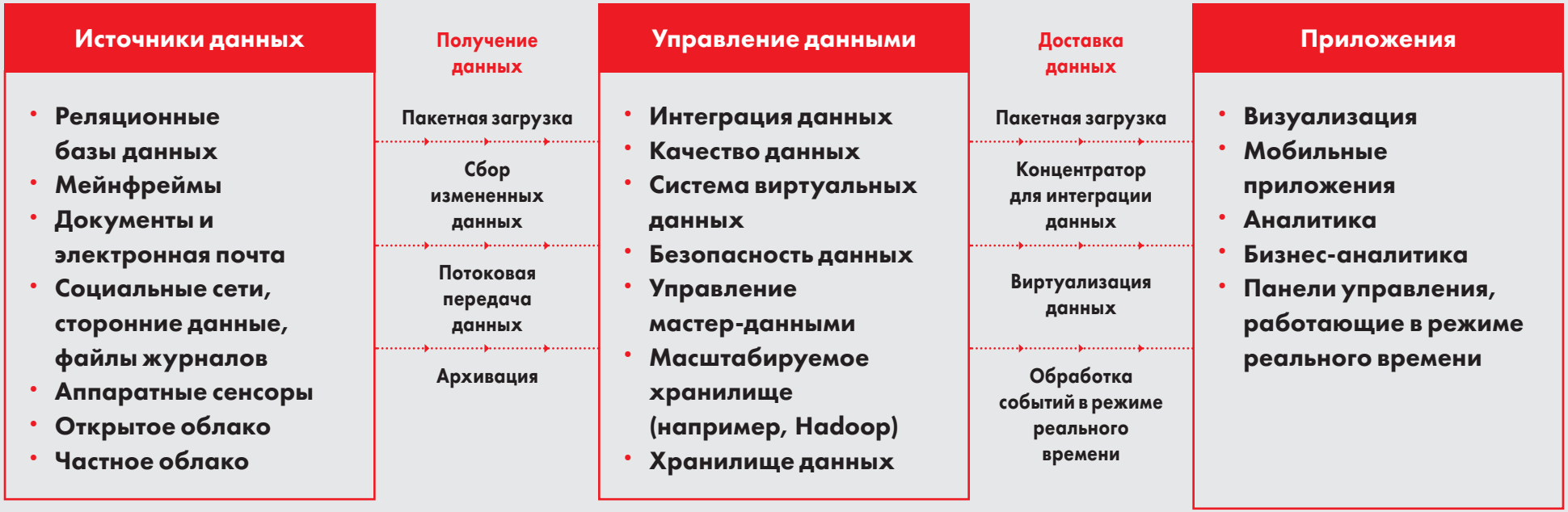
В качестве тестовых данных обычно используют вариант «живых» производственных данных организации, поскольку их форматы и структуры должны отражать заданные условия. К сожалению, отсутствие должного уровня маскирования таких данных может привести к тому, что конфиденциальные данные окажутся незащищенными в небезопасной тестовой среде.

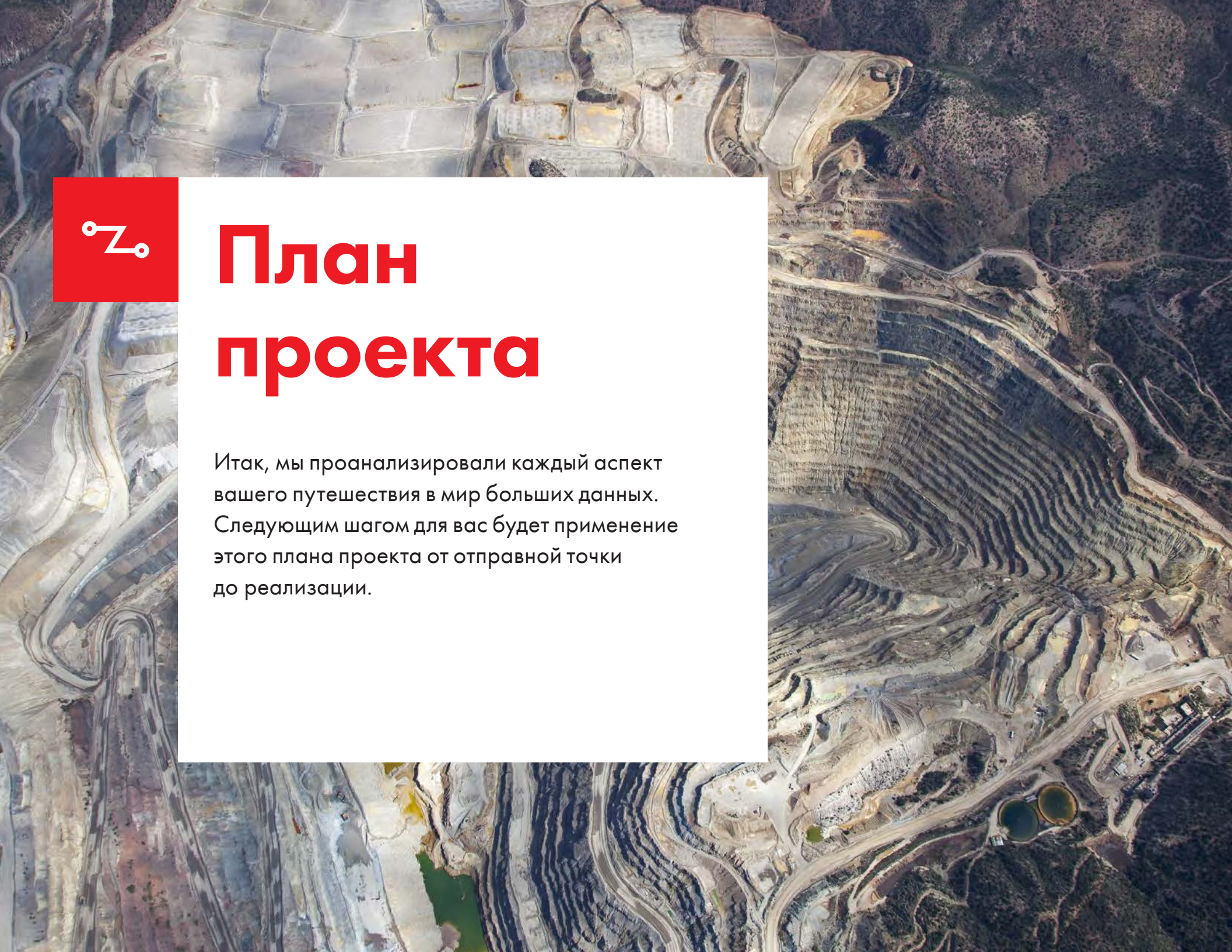
Как избежать ошибок в кодировании

Одна из наиболее распространенных причин задержек и увеличения бюджета проектов по большим данным заключается в том, что ошибки кодирования вручную, сделанные в «песочнице», проявляются при построении архитектуры. Если важные элементы архитектуры были кодированы вручную, вам, в конце концов, придется вычищать значительную часть кода, чтобы он удовлетворял требованиям уровня производственной среды. В качестве альтернативы вы можете использовать инструменты автоматизации и повышения производительности для того, чтобы избежать исправления огромного количества ошибок.

Идеальная архитектура больших данных

Приведенная ниже схема отображает модель, по которой мы рекомендуем создавать архитектуру больших данных и процессов.





План проекта

Итак, мы проанализировали каждый аспект вашего путешествия в мир больших данных. Следующим шагом для вас будет применение этого плана проекта от отправной точки до реализации.

План проекта

Используйте шаблон плана проекта для документирования деталей и разделов вашего проекта по большим данным. Далее используйте подробно составленное предложение для предоставления его другим подразделениям организации и получения поддержки. Также он может пригодиться при переговорах с внешними партнерами.

Этап 1. Стратегия	Этап 2. Данные
Определите цели для бизнеса и ИТ	**Определите, какая информация вам необходима**
Определите метрики успеха	**Определите данные и источники их получения**

План проекта

Этап 3. Канал передачи данных

Сотрудники

- Оценка необходимых навыков
- Оценка имеющихся навыков

Процесс

- Доступ к данным
- Интеграция данных
- Очистка данных
- Обработка данных
- Обеспечение безопасности данных
- Анализ данных
- Анализ потребностей бизнеса

Инструменты

- Распределенные вычисления (например, Hadoop)
- Качество данных
- Интеграция данных
- Управление мастер-данными
- Маскирование данных
- Визуализация
- Потоковая аналитика
- Аналитика
- Машинное обучение

Этап 4. Применение полученных знаний

Разработайте панели управления

Автоматизируйте процессы доставки данных

Настройте обратную связь

Начальный этап

Используйте всю информацию, приведенную в этом руководстве, для того, чтобы реализовать потенциал больших данных в вашей организации. Каким бы ни был размер проекта на данный момент, мы уверены в том, что вы прекрасно подготовлены ко встрече с трудностями, которые несет реализация проекта.

Помните, что необходимо мыслить стратегически при планировании ресурсов для проекта и концентрироваться на развитии процессов и навыков, которые будет достаточно легко перенести в другие проекты, масштабировать и постоянно улучшать. Если при реализации этого проекта вы придерживаетесь долгосрочного подхода, то сможете улучшить качество анализа и принятия решений в вашей организации в будущем на длительное время.

Мы надеемся, что вам запомнится работа над вашим первым проектом по большим данным. Ошибки, которые вы неизбежно допустите, и рабочая группа, которую вам предстоит создать — это части большого проекта, который имеет неоспоримое стратегическое значение для вашей компании.

Избежав множества проблем и следуя целям проекта, вы можете навсегда изменить к лучшему работу вашей организации.

Изменения будут колоссальными.

Следующие шаги

Готовы ли вы применить полученные знания?

Если вы разрабатываете продукты корпорации Informatica, то вы можете применить свои навыки в разработке Hadoop. Наши демо-версии программ по большим данным, соединители и службы помогут вам в разработке вашего проекта.

informatica

О корпорации Informatica

Корпорация Informatica — ведущий мировой независимый поставщик программного обеспечения по управлению данными. Во всем мире организации получают помощь от корпорации Informatica по устранению наиболее распространенных ошибок в управлении данными для успешной реализации масштабируемых и повторяемых проектов по большим данным.

Обратитесь в корпорацию Informatica

IN18-1014-2730