Управление данными
для эффективной цифровой трансформации

Олег Гиацинтов, технический директор DIS Group

План главы:

- Что такое стратегическое управление данными (Data Governance), зачем оно нужно, какие цели должно выполнять, как оно вписывается в стратегию цифровой трансформации.
- Как осуществляется Data Governance. Организационные инициативы, технологические инициативы и пр.
- Примеры организационных инициатив и рекомендации по их реализации. В частности, будут рассмотрены:
    (1) инициатива по назначению ответственных за данные, тех, кто будет описывать данные;
    (2) инициатива по определению процессов описания данных;
    (3) инициатива по определению направлений бизнеса, отделов, которые используют или заинтересованы в использовании тех или иных данных, влияют на те или иные данные;
    (4) инициативы по формированию CDOO, в частности, назначение следующих ролей: архитектор данных (правильно строит структуру каталога и бизнес-глоссария данных), эксперты по отдельным бизнес-направлениям (организовывает коммуникацию между бизнесом и ИТ), дата-стюарды (вручную заносят данные, описание данных).
- Функциональный подход к Data Governance: какие направления нужны, как они соотносятся с технологиями. В частности, будет рассмотрено управление физической, логической и концептуальной моделями данных, важность обеспечения качества данных и их защиты.
- Примеры технологий, которые необходимо использовать в рамках реализации технологических инициатив, а также рекомендации по их реализации. В частности, будут рассмотрены:
    (5) бизнес-глоссарий данных, правильная настройка (настройка ответственных за данные, ролевой модели и прочее) и использование инструментов;
    (6) каталог данных, правильная настройка и использование инструментов;
    (7) инструменты для обеспечения качества данных, правильная настройка и использование инструментов;
    (8) инструменты для защиты данных, правильная настройка и использование инструментов;
    (9) настройка совместной работы всех инструментов Data Governance.
- Практическое внедрение Data Governance: лучшие истории успеха, ошибки и результаты. В частности, будет рассмотрено построение «умного озера данных» с компонентом Data Governance, применение Data Governance в банковском секторе для оптимизации формирования отчётности.
- Заключение: ошибки, которых нужно постараться избежать. Среди таких ошибок – одноразовые акции по очистке данных.

Оглавление

1. Стратегическое управление данными, его цели и задачи

Управление данными – направление не новое. Раньше тоже нужно было понимать, какие данные есть в организации, и кто и каким образом ими пользуется, как они хранятся и в каких процессах участвуют. Но все это было не настолько важно, пока типов данных было не так много и можно было несложным путем описать их в документе или на своем портале.

Последние несколько лет появилось понимание, особенно в крупных компаниях, что изучение данных, их связей, попытки понять их ценность занимают много времени и ресурсов, что побудило заниматься этим направлением уже целенаправленно. Мировой опыт показывает, что компании, эффективно управляющие своими данными, открывают новые возможности для развития бизнеса.

Если говорить формально, то управление данными (Data Governance) – это важная корпоративная функция наряду с управлением производством, финансами, логистикой, персоналом.

Управление данными – это организационная инициатива, направленная на оптимизацию, защиту и использование информации в качестве корпоративного актива.

Основными целями для организации управления данными, с точки зрения автора, являются:

- повышение эффективности работы организации, ускорение бизнес-функций и проектов за счет повышения ценности используемых данных;
- повышение гибкости текущей деятельности компании;
- получение новых возможностей в бизнесе, развитие новых направлений деятельности;
- повышение прозрачности при работе с данными;
- снижение трудозатрат на согласование и реализацию доработок при внесении изменений;
- упрощение и повышение эффективности при формировании и анализе собираемой отчетности по подотчетным организациям;
- исполнение требований регуляторных органов в поставленные сроки.

. Управление данными максимально эффективно работает как поддерживающая функция на все направления бизнеса компании, являясь главной дисциплиной для создания повторяемых и масштабируемых политик работы с данными, процессов и стандартов для эффективного их использования. Возведение функции управления данными в главенствующую чаще всего приводит к отделению ее от основного бизнеса и невозможности нормальной совместной работы с другими направлениями.

Рассмотрим Data Governance как корпоративную функцию управления данными. Как любая другая функция организации, корпоративная функция рассматривается на стратегическом, операционном и технологическом уровнях (см. рис. 1).

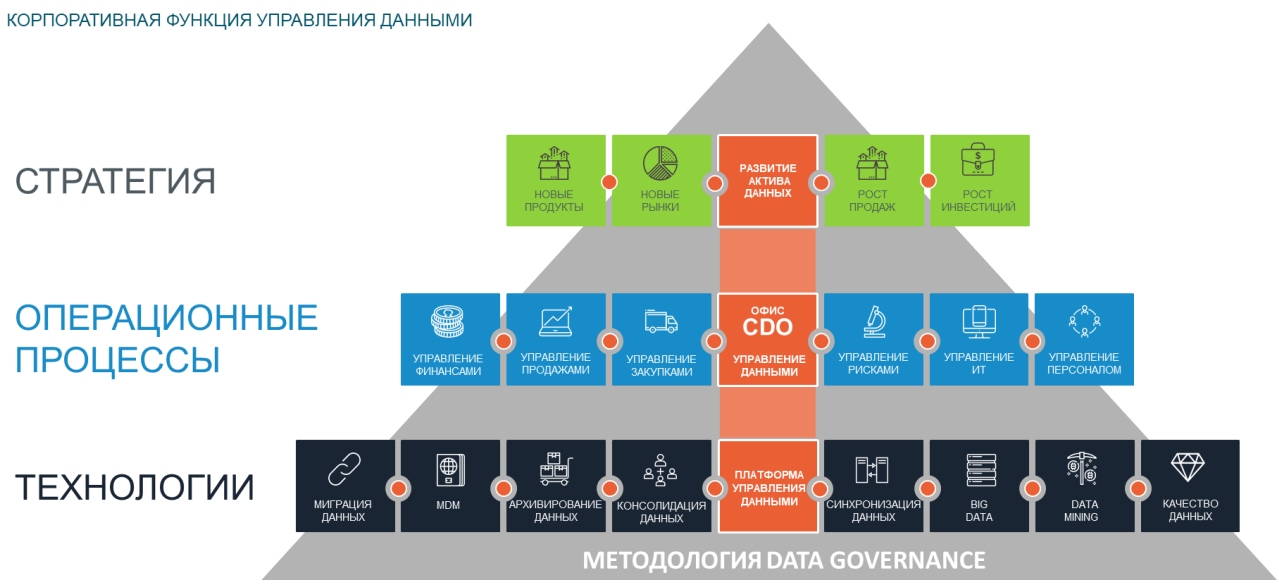

Рис. 1. Корпоративная функция управления данными

На стратегическом уровне основой функции управления данными является развитие актива данных, но не в качестве главного актива организации, а в качестве основного средства поддержания развития остальных стратегических инициатив, как рост продаж, выпуск на рынок новых продуктов, освоение новых типов бизнеса и т.д. Ключевым ответом на вызовы цифровой трансформации должна стать эффективная корпоративная стратегия управления данными для повышения качества использования и монетизации актива данных.

Операционный уровень функции представлен появлением подразделения по управлению данными (далее - офиса CDO, CDO Office, CDOO) и бизнес-процессами, которые связывают его с основными подразделениями организации. Этому уровню посвящены разделы 2-5.

Программные продукты, используемые для управления данными и формирующие технологический уровень функции, различны и многогранны. Основные типы рассмотрены в разделе 6.

2. Организационные инициативы управления данными

2.1. Общие принципы организации управления данными

Ключевым фактором успеха реализации корпоративной стратегии управления данными является наличие единого ответственного лица, лидера, объединяющего всю компанию в создании новых бизнес-моделей, основанных на активе корпоративных данных. Для этого создается отдельное подразделение, которое обычно называют Chief Data Office или Департаментом управления данными.

Руководитель департамента обычно именуется Chief Data Officer (CDO), хотя иногда его функции возлагают на руководителей цифровой трансформации (Chief Digital

Transformation Officer) или другие должности. CDO в зависимости от культуры и регламентов, сложившихся в организации, является частью бизнес-стратегии компании и главным проводником и советником для бизнес-подразделений в части работы с данными.

Нужно признать, что роль руководителя департамента управления данными состоит в выявлении проблем с данными и их решении. При этом, успешное управление данными должно быть ориентировано на бизнес, а не на ИТ. CDO возглавляет инициативы по управлению данными и дает организации возможность использовать все ресурсы данных и получать от них конкурентные преимущества. Однако CDO не только руководит инициативами. Он также должен возглавить культурные изменения, которые позволят организации иметь более стратегический подход к своим данным. Его основной задачей становится создание и поддержание процессов эффективной работы с данными. Подробнее обязанности CDO можно описать так:

- выработка стратегии управления данными в организации;
- приведение требований к данным в соответствие с имеющимися ИТ- и бизнес-потребностями;
- создание стандартов, политик и процедур по управлению данными;
- консультирование бизнес-подразделений по инициативам, существенно связанным с данными: бизнес-аналитика, big data, обеспечение качества данных и другим;
- донесение до всех заинтересованных лиц (внутри компании и вне ее) правильной информации о важности принципов управления данными;
- мониторинг использования данных при построении отчетности и в аналитике.

Можно сказать, что CDOO является «прослойкой» между бизнес-подразделениями и службой информационных технологий, переводящей запросы бизнеса на язык ИТ и обратно, а также освобождающей ИТ от «непрофильных» функций описания бизнес-задач, которые обычно ИТ выполняет. Это подразделение является драйвером повышения эффективности бизнеса, повышения качества управления и роста скорости принятия ключевых бизнес-решений в динамично меняющейся рыночной среде.

Часто возникает вопрос о взаимоотношениях руководителя по управлению данными и руководителя направления цифровой трансформации. Фактически, управление данными – одна из важнейших составляющих частей цифровой трансформации, не является ею полностью. Ведь помимо данных CDTO решает также вопросы применения новых аппаратных и процессных решений для развития организации.

В 2017 году было проведено исследование, в рамках которого было опрошено 108 CDO организаций из Electronic Business Group, одного из крупнейших инновационных центров Франции. Респонденты отметили следующие основные задачи CDO:

- демократизация использования данных организации;
- поиск новых бизнес-моделей, основанных на использовании актива данных (кейсов для применения);
- снижение издержек в текущих активностях организации и рост эффективности бизнес-направлений за счет управления данными.

Помимо перечисленных выше обязанностей есть еще ряд важных моментов, которые должен учитывать CDO в своей работе:

- он должен иметь собственные бюджеты на реализацию стратегических инициатив;
- для защиты своего направления должен быть выбран небольшой, но заметный на уровне организации, больной вопрос, решение которого покажет наглядно правильность выбранного подхода;
- основными проектами в начале работы департамента управления данными могут быть построение единого каталога данных, выявление критичных для бизнеса данных и защита данных.

2.2. Основные направления деятельности CDO

Деятельность департамента управления данными охватывает большое количество различных инициатив, которые обычно до его появления находились на стороне ИТ (см. рис. 2).

Рис. 2. Инициативы, покрываемые деятельностью CDO

К ним обычно относят:

- Data governance – собственно, управление данными как активом компании;
- Data ownership – процесс выявления и назначения владельцев данных;

- Data architecture – архитектура данных как дисциплина по созданию и ведению стандартов данных в системах или при взаимодействии между ними;
- Data modeling – процесс создания и ведения моделей данных;
- Data integration – интеграция данных, процессы перемещения и трансформации данных согласно требованиям пользователей;
- Database management and operations – управление хранением данных и операциями над данными в СУБД ;
- Data security and privacy – процессы предотвращения неавторизованного доступа к данным;
- Master data management – управление НСИ в части создания «единой версии правды» таких критичных для организации данных, как клиенты, продукты, материалы, счета и т.д.;
- Reference data management – процесс ведения статичных справочных данных (страны, классификации и т.д.);
- Data warehousing – процесс создания централизованного окружения для хранения и использования данных в целях отчетности и аналитики;
- Critical data elements – элементы данных, имеющих существенное влияние на регуляторную, операционную и управленческую отчетность;
- Metadata management – управление метаданными как объектами описаний данных и их характеристик – название, расположение, критичность, качество, бизнес-правила, связи с другими объектами;
- Data quality management – инициатива по управлению методами измерения и улучшения качества данных организации;
- Information lifecycle management – процесс и методология управления жизненным циклом информации от создания до удаления, включая соответствие всем внутренним и внешним требованиям;
- Content management – процесс оцифровки, сбора и классификации информации из бумажных и электронных документов.

3. Программа по развитию управления данными

Как уже упоминалось, управление данными должно помогать основным направлениям бизнеса организации, избавлять его от несвойственных, трудоемких операций по пониманию сути и качества данных. Для выстраивания этого направления можно использовать программу по развитию управления данными. Пример подобной программы приведен на рис. 3.

Рис. 3. Программа по развитию управления данными

В состав программы обычно включают следующие этапы:

- планирование программы - определение задач, возможностей и состава команды;
- оценка зрелости организации;
- разработка целевой модели;
- построение и утверждение бизнес-кейса;
- внедрение прототипа и разработка плана дальнейших внедрений по направлению.

3.1. Планирование программы

Для начала необходимо определить реальные бизнес-задачи, для которых в текущий момент времени управление данными станет эффективным инструментом для ускорения. Эта информация ляжет в основу планируемой программы и регламентов работы.

Поскольку на первом этапе сложно сказать, какой результат на самом деле окажет внедрение нового направления, нужно выделить изначальные задачи, на которых опробовать подход, не затрачивая много усилий, но результат при этом будет виден на уровне бизнес-подразделений. Эти задачи лучше всего определять на основе текущих потребностей бизнеса, в которых низкая эффективность работы с данными приводит к существенным потерям во времени и средствах.

Здесь же нужно начинать формирование команды будущего подразделения, которое возьмет на себя вопросы Data Governance, постепенно их забирая из сферы ответственности департамента ИТ.

В начале работы важно оценить, насколько подразделение, где предполагается начать решать первые задачи по управлению данными, и организация в целом готовы к

такому подходу. Наличие существующих подходов и стандартов может упростить продвижение направления в компании, развивая их без серьезной переделки бизнес-процессов, потому что именно это является крайне болезненным аспектом, приводя к неприятию со стороны бизнес-подразделений новых идей и последующему их блокированию.

Первая команда подразделения - офиса CDO (CDOO) - обязательно должна иметь в своем составе специалистов, владеющих бизнес-пониманием деятельности организации и того направления, с которого решено начать построение направления Data Governance. О составе команды подробнее информация приведена в п.4.2.

Завершение первого этапа работы программы состоит в создании предварительного плана действий – «дорожной карты», включающей в себя все последующие этапы развития направления:

- оценку зрелости компании;
- создание целевой модели;
- разработка стратегии по управлению данными;
- разработка первичных регламентов;
- выработка кейсов для прототипирования;
- разработка и защита прототипа;
- определение успешности разработанных подходов и прототипа;
- планирование дальнейшего развития.

3.2. Оценка зрелости организации в управлении данными

Возвращаясь к тому, что управление данными не является новым видом деятельности, а, скорее, тем, чему раньше меньше уделяли внимание, - сначала будет правильным оценить, в каком же состоянии находится сейчас организация в плане работы с данными.

Для оценки зрелости необходимо оценить:

- текущее состояние бизнес-процессов в области управления данными;
- наличие политик, правил, регламентов в этом направлении;
- наличие ответственных лиц, курирующих эту тему в целом по компании или по бизнес- или территориальным направлениям;
- наличие процессов мониторинга зависимости между длительностью и успешностью проектов и использованием данных в них;
- наличие программных средств по управлению данными.

На рис. 4 приведены пять основных стадий зрелости, учитывающих текущий уровень важности данной инициативы, в первую очередь, на стратегическом уровне.

Рис. 4. Стадии зрелости компании по управлению данными

Не будем рассматривать нулевую стадию, на которой подобных процессов работы с данными просто нет. Рассмотрим стадии зрелости компании в области управления данными.

Стадия 1. Реагирование.

Эта стадия характеризуется проявлением отдельных инициатив по каталогизации данных в проектах. Чаще всего, такие инициативы начинаются в качестве сопутствующих основному крупному проекту при наличии энтузиастов, вовлеченных в проект. Создаются политики и стандарты по описанию данных, которые уточняют сам проект и его данные. Однако деятельность не выходит за рамки проекта и часто не используется далее в других инициативах.

Стадия 2. Процедурность.

На основе ранее проведенных проектов, где были накоплены умения по накоплению знаний об использованных данных, начинается продвижение в компании разработанных политик и методологий с передачей знаний в другие проекты. К этому процессу подключается средний менеджмент ИТ-направления, используя повторно сделанные наработки и адаптируя их к новым проектам. Начинается оценка эффективности ИТ-проектов с применением управления данными.

Стадия 3. Структурированность.

Применяемые политики, стандарты и регламенты по описанию и повторному использованию данных и метаданных утверждаются на уровне руководства ИТ-подразделения и становятся важными для каждого нового проекта любого типа.

В рамках организации начинают формироваться центры компетенции по управлению данными. Обычно это происходит в департаменте ИТ, но для уточнения экспертизы активно используются рабочие группы с привлечением бизнес-пользователей.

Стадия 4. Инициативность.

Утвержденные стандарты и политики оформляются в виде отдельной программы управления информацией, которая учитывает интересы бизнес-подразделений и предыдущие наработки и ошибки. Программу начинают возглавлять руководители заинтересованных бизнес-подразделений вместо ИТ, которое лидировало в этом процессе ранее.

Программа применяется ко всем проектам, идущим в организации.

Стадия 5. Интегрированность.

Data Governance встраивается во все основные бизнес-процессы, а стратегия управления данными входит в состав целевой стратегии развития организации. Инициатива по отслеживанию исполнения переходит к руководству организации, поднимающему значимость направления на высший уровень. Оценка деятельности в области управления данными осуществляется как уровень влияния на бизнес компании в целом.

Понимание реальной зрелости компании дает возможность сделать следующие шаги:

- определить или уточнить бизнес-задачи;
- правильно понять требуемый состав команды и сформировать ее;
- определить регламенты и политики, а также решения руководства, необходимые для эффективной работы с данными;
- разработать план действий на ближайшую перспективу.

3.3. Разработка целевой модели

Понимая текущий уровень готовности компании, видя первые задачи и их потенциальные решения, имея поддержку от руководства, начинается построение целевой модели внедрения, включающей в себя как административные процессы, так и техническое оснащение (в первую очередь, программное обеспечение).

Для построения целевой модели на основе полученной информации о зрелости компании нужно:

- выбрать бизнес-подразделение, с задач которого будет начинаться построение подходов к управлению данными;
- найти бизнес-спонсора проекта;
- из задач подразделения сформировать первичные бизнес-кейсы и сценарии для их реализации;
- сформировать первичные регламенты, описывающие основу выбранного подхода (они будут еще меняться после решения первых задач);

- выбрать задачи для построения прототипа;
- описать планируемые результаты от реализации прототипа;
- сформировать и описать процесс мониторинга за эффективностью выбранного подхода.

Все принятые решения обязательно требуется утверждать на уровне руководства и бизнес-спонсора для повышения уровня легитимности этой новой инициативы.

3.4. Построение и утверждение бизнес-кейса

Вся разработанная целевая модель должна стать основой для главного документа направления – стратегии по управлению данными, определяющей управление данными в организации как бизнес-функцию с описанием ее целей, задач, планируемых результатов и подходов для достижения поставленных целей.

Понимая первичные задачи, для которых можно будет применить подходы управления данными, вырабатываются задачи для прототипа решения. Проект с построением прототипа предлагаемых процессов и решений на выбранных задачах дает возможность понять и изменить выбранный подход, а также утвердить его у руководства. Он же влияет на выбор последующих бизнес-кейсов для работы и построение конвейера проектов по монетизации данных.

Для каждого выбранного бизнес-кейса просчитывается планируемый финансовый результат и устанавливается мониторинг реальных результатов по ходу проекта. Крайне важно согласовать критерии успешности прототипа до начала его реализации. Все критерии должны быть «осязаемыми», а не эфемерными. Они должны оцениваться в цифрах, показывающих процент ускорения существующих процессов при полном соответствии функций используемого программного обеспечения поставленным требованиям. Самый лучший подход – возможность измерить улучшения в денежном выражении, отталкиваясь от текущих затрат на подобные задачи без использования подходов Data Governance.

Хорошо понимая, что одна единственная задача с применением нового подхода может не дать положительного результата, обычно в прототипе решается от трех до пяти взаимосвязанных задач, на которых совместно можно достичь успеха с бОльшей долей вероятности.

3.5. Задачи для построения прототипа

Понимание текущей ситуации в компании с уровнем зрелости хорошо дает понять набор последующих действий, которые необходимо предпринять, чтобы достичь реальной выгоды от существующих данных и понять реальную потребность в новых. И в этом основную ценность приобретает построение прототипа, результаты которого определят, как дальше организация будет идти по пути построения процесса «Знай Свои Данные» (Know Your Data) в организационном и технологическом аспектах. Набор действий для организационного развития организации по управлению данными приведен на развиваться в направлении эффективной работы с данными и насколько выбранный подход приживется в ней.

Главное условие выбора задач для реализации прототипа – не пытаться объять необъятное. Выбор может оказаться крайне сложным, чтобы он продемонстрировал возможности подхода к управлению данными наилучшим образом.

Приведу несколько рекомендаций по этому вопросу.

1. Задача должна быть довольно важной, проблематичной для компании, чтобы ее решение было заметно.
2. Демонстрировать нужно подход к решению задачи и результаты, а не программное обеспечение, которое помогает ее реализовать.
3. Эффект от применения может быть косвенным в снижении затрат на решение проблемы по сравнению с текущими затратами или в уменьшении сроков.
4. В каждой крупной организации могут быть десятки тысяч сущностей данных, описывающих ее деятельность. Начинать стоит с 50-100.
5. Обязательно наличие поддержки со стороны руководства для оперативного прохождения любого рода административных барьеров, которые могут не дать развиться инициативе.

3.6. Примеры задач для реализации прототипа

Здесь я бы хотел выделить ряд практических задач, которые могут послужить основой для внедрения системы управления данными. Такие задачи наиболее часто встречаются в проектах, показывая реальную пользу.

1. **Ускорение аналитических исследований.**
   Часто аналитики, выявляющие какие-либо зависимости и аномалии данных, вынуждены тратить бОльшую часть своего времени на изучение, что за данные для них доступны, откуда они взялись, какого они качества и как их использовать. Необходимо ускорить этот процесс, снять с аналитиков вопросы поиска данных и их классификации и оставив им построение гипотез и их проверки.
   Аналитические задачи могут быть из любой, важной для организации, области – продажи, ремонты, логистика и т.д.
2. **Устранение непрозрачности в отчетности.**
   Эта задача является довольно сложной, если пытаться решить ее сразу и целиком. Ее суть состоит в том, чтобы выявить наличие дубликатов отчетов и разных путей доставки данных до них. Выявив суть проблемы, нужно постепенно, понемногу начинать выводить ненужные процессы обработки данных и структуры хранения из эксплуатации, оставляя правильные с точки зрения бизнес-логики.
3. **Построение сервисов по заказу данных.**
   Суть задачи состоит в построении единого интерфейса для бизнес-пользователей по запросу и получению данных, давая изначально возможность им рассмотреть, какие данные могут быть доступны и какого они качества.
4. **Аудит состояния хранилища или озера данных.**
   Данная задача состоит в анализе используемости существующих отчетов и структур (постоянных или временных), которые заполняются данными в течение длительного времени. Чаще всего, после значительных изменений в системах разные таблицы детальных слоев и витрины, а также их поля оставляют «на

всякий случай». Выявление и последующее удаление структур и процессов их наполнения существенно снизит затраты на инфраструктуру.
Эта задача должна выполняться регулярно.

5. **Сведение идентичных справочников.**
Задача возникает из-за многократного использования разными подразделениями одних и тех же по сути, но разных по наполнению справочников. Необходимо выявить причину разночтений и доказать необходимость сведения справочников в один или добиться построения процессов реклассификации справочников, чтобы при использовании их значений всегда можно было получить однозначное соответствие значениям в связанных справочниках.

Довольно часто в рамках построения прототипа по управлению данными приходит понимание, что есть необходимость в создании или реорганизации решений по отчетности и аналитике. Вопрос ставится так, что построение отчетности – это и есть Data Governance. Но это не так. Управление данными является одним из инструментов для улучшения ситуации с отчетностью и аналитикой, ускоряет его, но не подменяет.

С другой стороны, часто высказывается мнение, что при отсутствии грамотно выстроенной отчетности нет смысла заниматься управлением данными. И это не так, потому что как раз задачу формирования отчетности или какой-либо аналитики для одного из подразделений хорошо использовать в качестве прототипа.

3.7. Планирование дальнейших действий

Сделав прототип, оценив результаты и сделав работу над ошибками, можно будет начать построение масштабной программы управления данными в организации в целом с тиражированием полученного подхода по следующим задачам.

На финальном этапе реализации программы по развитию управления данными необходимо закрепить положительные результаты выбранного подхода и реализованного прототипа. Необходимо вынести на руководящий орган – комитет, совет по данным – защиту проекта по реализации прототипа вместе с обоснованием правильности и эффективности методов управления данными, найденных ошибках и предлагаемых изменениях, планах дальнейшей работы

Утверждение результатов дает «зеленый свет» дальнейшим инициативам по использованию подходов Data Governance для новых бизнес-кейсов. Тиражирование наработок в компании приводит к постепенному появлению «конвейера» проектов, ускоряющих различные бизнес-функции – конвейеру монетизации данных.

Как раз в этот момент завершается формирование целевого подразделения – департамента управления данными, а также выстраивается основная линия его работы – процесс Know Your Data.

4. Команда CDO

4.1. Местоположение подразделения по управлению данными

Часто возникает вопрос – где в штатном расписании должно находиться подразделение, ответственное за управление данными. Давайте рассмотрим несколько вариантов.

При подчинении подразделения по управлению данными финансам или одному из бизнес-подразделений иногда происходит перекос деятельности CDOO в сторону именно этого направления, то есть, основные задачи устанавливает руководитель бизнес-направления, а совместная работа с ИТ может снизиться до минимальных размеров.

В случае, когда внутри ИТ создаётся офис работы с данными, часто не возникает никаких изменений в работе всей организации, потому что не возникает «прослойка» между бизнесом и ИТ, как и новые виды взаимодействия между ними.

Выделение CDOO в качестве отдельной бизнес-единицы с подчинением высшему руководству организации решает задачи ускорения межведомственного взаимодействия, изменения существующих и появления новых бизнес-процессов. Однако иногда такой подход возводит управление данными в ранг отдельного бизнес-направления, что негативно влияет на решение задач для основных подразделений организации.

Поэтому всегда важны личные качества руководителя CDOO. При этом, на первом этапе работы крайне важна поддержка сверху в лице руководства компании. Создание новой бизнес-функции – это всегда нелегкая задача, которая, с точки зрения автора главы, может идти только от лица высшего эшелона власти на предприятии.

Крайне важным является наличие управляющего органа выше CDO на уровне руководства организации. Появление любого нового направления редко воспринимается в других департаментах положительно. Часто возникает проблема в блокировании доступа к необходимым данным и выстраивании ненужных барьеров для защиты «своей территории». Наличие влиятельного центра принятия решений в виде комитета по управлению данными с возможностью сломать выстроенные преграды на уровне среднего менеджмента решает эти вопросы. Из опыта видно, что при отсутствии такой поддержки деятельность CDOO довольно быстро становится отделенной от всего происходящего в компании и постепенно отмирает.

4.2. Организационная структура департамента управления данными

Команда, которую у себя собирает CDO, должна быть нацелена на результат. Общую структуру ролей в команде хорошо иллюстрирует рис. 5.

Рис. 5. Организационная структура для управления данными в организации

В подразделение по управлению данными входит ряд позиций. Есть много разных мнений на этот счет, но нужно остановиться на нескольких должностях, которые, с точки зрения автора, обязательно должны входить в состав такого подразделения. Все эти люди важны для эффективного функционирования CDOO.

1. Стюард данных. Для каждого вида деятельности нужны «рабочие руки». И стюарды данных – это основные специалисты CDOO, которые осуществляют ведение модели по требованиям архитектора данных с наполнением от экспертов. Они же осуществляют поиск информации, если бизнес-пользователь по разным причинам не делает это сам. Кроме того, при наличии каталога метаданных появляется задача сравнения требований от бизнес-пользователей с реализованными разработками, например, при формировании показателей отчетов или данных для анализа. Иногда стюарды располагаются также в функциональных подразделениях, имея косвенное управление со стороны подразделения по управлению данными.
2. Архитектор данных. На его плечах лежит вся ответственность за построение моделей данных (концептуальной и логической), за те типы связей, которыми пользуются стюарды данных при наполнении моделей, за построение иерархии в глоссарии и за все те дополнительные типы информации, которые нужны бизнес-пользователям для понимания собственных данных. Его главный документ для работы – соглашение о моделировании, описывающее основные аспекты построения моделей данных, связей между ними, ответственности и многое другое. Кстати, обычно архитектор данных присутствует в единственном числе.
3. Эксперт по данным (или ключевой бизнес-пользователь данных). Одной из целей работы CDOO является упрощение взаимодействия между бизнес-подразделениями и ИТ, ведь часто причиной задержек проектов является именно непонимание или недопонимание между специалистами разных

подразделений. Задача эксперта – понять запрос от бизнес-пользователя и донести до стюардов и архитектора, как нужная информация должна быть описана в глоссарии, какая бизнес-логика должна быть применена и какие еще требования могут применяться. Эксперты (в английском варианте – Subject Matter Expert) часто работают не в подразделении по управлению данными, а по своим основным подразделениям, но при этом могут иметь функциональное подчинение руководителю департамента управления данными. Такой подход, по мнению автора, очень эффективен, так как эксперт постоянно подпитывается информацией из бизнес-подразделения, находясь в его составе, и делится ей с сотрудниками CDOO, которые далее транслируют ее по компании в виде описаний в системах.

4. Офицер по качеству данных. Эта должность может находиться и в CDOO, и в ИТ. Важен набор его функций. В связи с разрастанием числа проверок качества данных в организации рано или поздно начинается дублирование таких правил, что приводит постепенно к дополнительным затратам на инфраструктуру. Задачи офицера – выстраивание единой политики по качеству, ведение реестра правил проверки и обеспечения качества данных, оптимизация существующих правил и правильная постановка задач ИТ-специалистам по созданию новых.

5. Деятельность CDOO

5.1. Взаимодействие бизнес-подразделений и CDOO

Поговорим о процессах. Как же меняется обычное взаимодействие между бизнес-подразделениями и ИТ при появлении офиса CDO? И зачем вообще нужно что-то менять?

Общаясь с крупными организациями, я для себя выяснил, что даже такой запрос от бизнес-подразделения в сторону ИТ, как поменять или добавить два-три поля в отчете растягивается иногда на месяцы или годы. Когда же работу закончат, то нет никаких гарантий, что она кому-то уже нужна именно в таком виде. Требования и правила меняются сейчас быстрее, чем ИТ-служба с текущими ресурсами и загрузкой успевает их реализовать.

Например, в компании, где нет еще CDOO (или людей с такими функциями) бизнес-подразделение формирует свой запрос в сторону ИТ-службы о необходимости создать новую выгрузку, отчет, что-то еще подобного рода или внести изменения в существующие разработки. В хорошем случае бизнес еще вместе с запросом направляет и бизнес-требования, хотя обычно ИТ-специалисты пишут его сами. Понятно, что ИТ, имея несколько другой круг общения, обычно не все сразу понимают в поставленных требованиях.

ИТ-специалисты пишут техническое задание. Оно уже более понятно техническим специалистам, но менее понятно бизнес-пользователям. Процесс согласования может затянуться из-за непонимания бизнесом написанного или невозможности реализовать желаемое.

Когда разработка завершена, на UAT (тестировании бизнес-пользователями) начинают вылезать многочисленные проблемы, связанные с тем самым непониманием, которое заложили еще при написании технического задания.

Конечно, так происходит не всегда. Но, согласитесь, достаточно часто.

Тот же процесс при наличии CDOO выглядит совершенно иначе. На вид (см. рис. 6) он выглядит более сложным, чем просто взаимодействие двух подразделений. На деле такой процесс экономит до 40% времени по оценкам директоров по данным крупных российских и западных компаний.

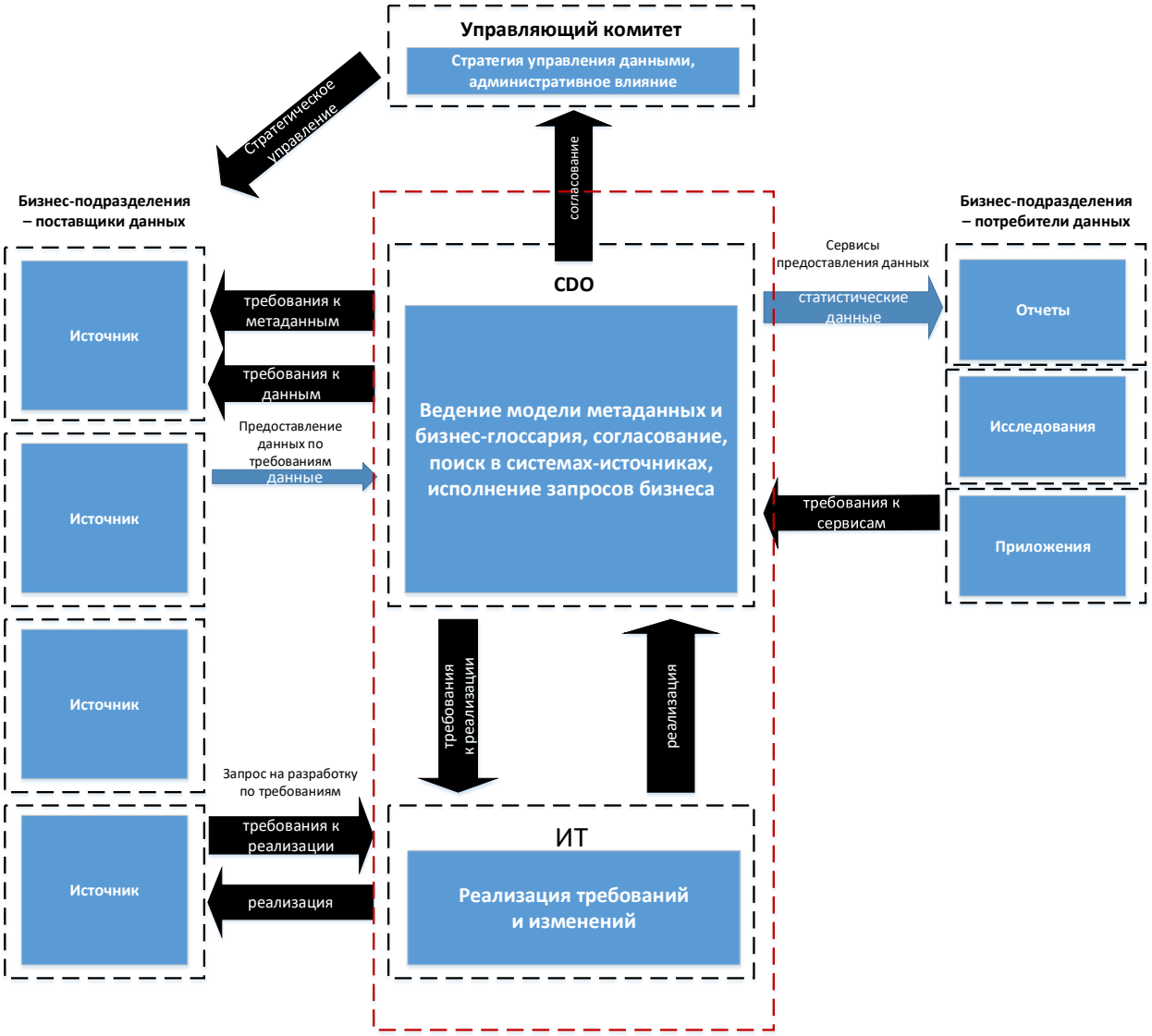

Рис. 6. Пример процесса взаимодействия бизнес- и ИТ-подразделений при наличии CDO

Бизнес-специалист обращается в офис CDO с вопросом о предоставлении данных, изменении или создании отчета, выгрузки, витрины. Формат обращения может быть использован обычный – через электронную почту или внутреннюю систему взаимодействия – или путем заказа данных в бизнес-глоссарии. Подобные функции, которые обычно называются Data Marketplace, присутствуют с недавнего времени в основных системах ведения терминологии и логических моделей данных.

Далее всю работу ведут сотрудники департамента управления данными. Эксперты по данным, являясь специалистами по своему направлению, понимают запрос, его суть и текущую реализацию. Они же проверяют наличие более ранних наработок по запросам, которые уже были заданы другими подразделениями (см. рис. 7).

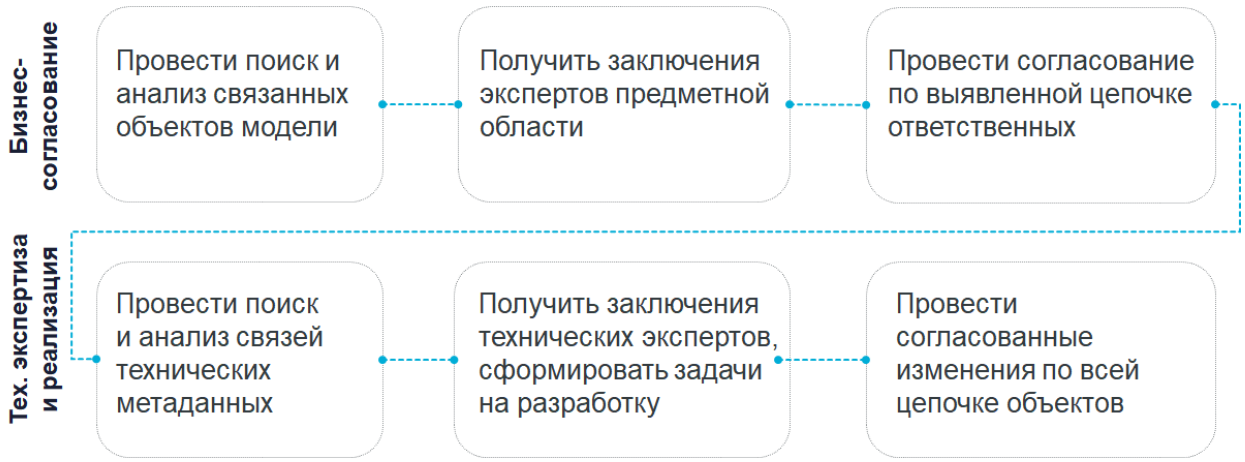

Рис. 7. Процесс согласования в CDOO

В случае отсутствия разработанного решения подключается стюард данных, который ищет возможные реализованные решения в каталоге метаданных и бизнес-глоссарии, а также ответственных лиц (владельцев данных, экспертов и т.д.). Фактически, стюард формирует техническое задание на основе знаний эксперта, согласовывает возможность доступа к источникам данных через своего руководителя и владельцев данных и направляет в службу ИТ детальное задание на реализацию.

Специалисты ИТ, в свою очередь, обязаны подобрать правильный вариант исполнения и выполнить описанную задачу. Таким образом, ИТ постепенно теряет функции системного и бизнес-анализа, передавая их в CDOO. Результаты работы также принимает департамент управления данными, ведь его эксперты могут оценить как техническую реализацию, так и соответствие решения поставленной бизнес-задаче.

Если процесс согласования предоставления данных с владельцем данных затягивается или не приводит к положительному результату, руководитель CDOO передает запрос на решение проблемы на уровень Комитета по данным. Собираясь с определенной частотой, Комитет является основным органом, который способен своей властью постановить, стоит ли предоставлять данные и в каком виде тому или иному бизнес-подразделению. Полномочия Комитета по данным должны быть выше уровня отдельных подразделений.

Сдачу работ бизнес-подразделению производят уже специалисты CDOO.

Существенным обстоятельством является необходимость анализа всего набора связей для изменяемых стюардом данных. Особенно это важно при существенном изменении выгрузки или отчета, а также при удалении устаревших объектов. Обязательно нужно проверять, какие бизнес-процессы, проекты, рабочие группы пользуются данными из запроса, чтобы не нарушить их нормальный ход работы. Иногда возникает ситуация, что, например, изменение законодательства или решение руководства приводит к необходимости создать новую отчетность и отказаться от использовавшейся ранее. При этом весьма вероятно, что данными из существующих объектов пользуются и другие подразделения, и удаление старых отчетов просто приведет к остановке их работы и усложнению взаимоотношений между подразделениями.

Выстраивание работы специалистов CDOO – это построение нового процесса внутри организации, который можно назвать «Знай Свои Данные» (Know Your Data). А успешная работа всего подразделения создает конвейер сервисов по монетизации данных.

5.2. Процесс Know Your Data («Знай Свои Данные»)

Выше мы рассмотрели процессы взаимодействия департамента управления данными с другими подразделениями. Теперь надо взглянуть на внутренний процесс в самом CDO – процесс Know Your Data (или по-русски – «Знай Свои Данные»).

Процесс состоит в слаженной работе сотрудников подразделения для выполнения запросов от бизнес-пользователей. На рис. 8 показан общий вид процесса и его фаз.

Рис. 8. Процесс Know Your Data

Рассмотрим его более детально.

1. Поиск данных. Важнейшая деталь в процессе, чтобы найти не только сами объекты метаданных, которые, например, указал бизнес-пользователь в своем запросе, но и все связанные с ним элементы, включая другие объекты моделей данных разных уровней, заинтересованных лиц, рабочих групп и подразделений, политик, регламентов и многое другое, что влияет на эти данные или на что влияют они.
2. Интеграция данных. Этот этап является более техническим и касается реализации поставки или подготовки запрошенных данных или предоставления среды и доступов к источникам нужных данных для использования бизнесом или аналитиками. На данном этапе обязательно нужно иметь возможность проверки реализованных задач с помощью каталога метаданных и обеспечения разработчиков нужной для них бизнес- и технической информацией о данных.
3. Качество данных. При разработке нового процесса предоставления данных нужно учитывать требования к их качеству. И нельзя забывать о проверках качества существующих интеграционных процессов.
4. Обновление каталога метаданных. Для успешной работы офиса CDO представление физической модели существующих разработок и процессов должно соответствовать реальности. Поэтому работа каталога метаданных должна обязательно выполняться периодически (лучше ежедневно) в автоматическом режиме. И очень критичным требованием является проверка всех изменений, которые произошли в модели с момента предыдущей загрузки метаданных. Такие изменения могут повлиять на логическую и концептуальные модели, внося в них новые связи и удаляя исчезнувшие. Один из основных инструментов для архитектора данных.
5. Определение бизнес-глоссария. Глоссарий не может существовать отдельно. Работы по его ведению не должны быть дополнительными, иначе он станет дополнительной нагрузкой, и от решения рано или поздно откажутся.

Необходимо построить работу так, чтобы ни один запрос не мог пройти вне глоссария, заставляя вносить требуемые изменения в концептуальной и логической моделях до начала реализации. Бизнес-глоссарий всегда будет являться тем решением, в которое могут зайти любые пользователи организации.

6. Определение бизнес-владельцев. Для каждого термина, каждого правила, каждой системы должны быть назначены владельцы и другие стейкхолдеры. Процесс их выявления и согласования должен быть прописан в регламенте компании.
7. Публикация и использование. Все деятельность процесса Know Your Data направлена на максимально быстрое предоставление нужных данных нужного качества. Когда данные нашли, описали, выполнили реализацию, проверили качество, проверили все связи и изменения в моделях данных, указали ответственных, тогда настает последний этап – опубликовать готовые наработки и дать к ним доступ пользователям.

Процесс Know Your Data циклический. Постоянные запросы и изменения от бизнеса приводят к поддержанию моделей всех уровней в актуальном состоянии, нужном качестве и с готовыми доступами к отчетам, результатам анализа, наработкам.

5.3. Назначение владельцев данных

О такой щекотливой теме, как выявление и назначение владельцев данных и любых других заинтересованных лиц, поговорим отдельно.

Для каждого нового или существующего объекта данных требуется владелец и ответственные лица, потому что за несоответствие данных требованиям или предоставление какой-то новой информации должен кто-то отвечать.

Все привыкли к понятию «владелец системы», то есть подразделению, которое пользуется данными, содержащимися в этой системе. И это же понятие распространяют обычно на сами данные.

Но это чаще всего не так. Простой пример. Операторы в банке заносят паспортные данные в системы банка. Значит ли это, что владельцем данных является операционный блок. Нет. Если в данных будет ошибка, и клиент напишет жалобу, то ее придется рассматривать подразделению, ответственному за соответствие требований регулятора (обычно оно называется подразделением комплаенса).

Важно понимать следующее при выявлении и назначении владельцев данных.

1. Владельцем данных является не подразделение, а конкретный человек, руководящий направлением.
2. Владелец данных выставляет требования к данным и отвечает за них. В случае с приведенным примером владельцем данных будет именно руководитель комплаенса как главный ответственный за требования к заполнению паспортных данных клиента. А вот сотрудники операционного блока являются пользователями этих данных.

3. Владелец данных может отвечать и за сущность, и за отдельный атрибут, а иногда – за часть атрибута.
4. Владельцев данных может быть несколько. Все зависит от требований к данным. Если требования разных владельцев различаются, нужно разделять терминологические понятия на несколько, чтобы не было противоречий.
5. Владелец данных может делегировать свои полномочия своему подчиненному. Но у того в таком случае должно быть право принятия решения по изменению требований к данным.

Обычно в крупных организациях процесс назначения владельцев данных происходит болезненно с попыткой снять с себя ответственность в сторону пользователей. В процесс согласования обязательно вовлекается комитет по данным. Как я писал раньше, в состав комитета должны входить руководители высочайшего уровня для решения подобных вопросов без промедлений.

Вот еще пример в том же банке. Кто является владельцем данных о дате рождения клиента? Это паспортные данные, и вы правы – это директор (руководитель, начальник) по комплаенсу. А вот кто является владельцем данных о возрасте клиента? Этот показатель к паспортным данным отношения не имеет, хотя и рассчитывается с учетом даты рождения. А кто к нему ставит требования? Скорее всего, это будут подразделения по управлению рисками или маркетинг. Или оба. Зависит от конкретной ситуации. И владельцев данных может быть в таком случае больше одного.

5.4. Конвейер монетизации данных

Естественно, работа CDOO вплетена в работу всех остальных подразделений компании. Она не просто состоит в том, чтобы отвечать на запросы, но в том, чтобы повышать ценность данных как эффективного актива для развития компании, развивать монетизацию данных.

Монетизация данных редко работает напрямую, принося доход от продажи данных. Это обычно называется «внешней монетизацией». Таких примеров довольно много – взаимодействие между, например, телеком-оператором и службой такси.

Чаще всего, основной эффект монетизации данных состоит в ускорении работы подразделений, более быстрому появлению новых продуктов (уменьшение показателя time-to-market), снижении затрат на поддержание существующих решений и продуктов. Это «внутренняя монетизация».

Запросы от бизнеса порождают фактически «конвейер монетизации», давая эффект от каждой задачи. Эффект может быть незначительным в рамках конкретного случая, если рассматривать его «под лупой», но все вместе они приносят ощутимый результат. Именно поэтому крайне важно стартовать с небольшой, но показательной задачи, чтобы и подход к работе, и результат были заметны.

Если посмотреть на конвейер монетизации, то видно, что он по своей структуре мало отличается от производственного процесса (см. рис. 9). Только все указанные действия производятся не с деталями, а с данными.

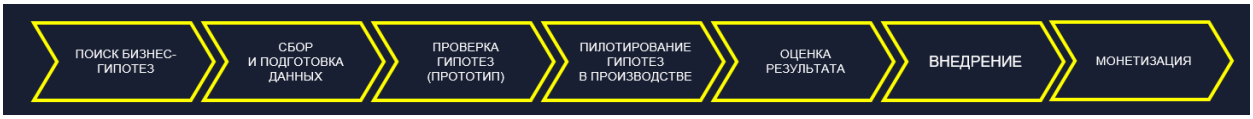

Рис. 9. Конвейер монетизации данных

Поиск бизнес-гипотез развивается по двум сценариям:

- реактивному – как результат реакции на возникающие события – запросы пользователей, изменение законодательства;
- проактивному – предположение выстраивается на основе экспертного мнения, приводя к новой гипотезе, которая даст улучшение в бизнесе.

А дальше выполняется классический процесс понимания сути бизнес-гипотезы, наличия данных и ответственных, уровней качества, связей и много другого, что уже было описано выше. Используем парадигму «AS-IS». Рассмотрению подходов «AS-IS» и «TO-BE» посвящен отдельный пункт в третьем разделе.

На основе выстроенного понимания формируется новая концептуальная и логическая модель (или меняется существующая) метаданных. Выстраиваем видение «TO-BE». Эти оформленные идеи передаются на реализацию, которая в конце концов приводит к появлению набора уже данных – прототипа будущего решения, который готов для проверки гипотезы.

Оценив прототип, проверив гипотезу и убедившись в ее дееспособности, можно начинать применять ее сначала в одном из направлений бизнеса, а затем тиражировать в другие.

На каждом этапе обязателен мониторинг по срокам и затраченным ресурсам, чтобы сравнить затем с реальными затратами до внедрения. Полученная разница даст первое понимание по реальной монетизации.

На рис. 10 приведено подробное описание по этапам работы конвейера монетизации данных. Процессы поиска и заказа данных, оценки их качества и возможности их использования приведены в пп.6.2-6.4, посвященных каталогу метаданных, бизнес-глоссарию и средствам обеспечения качества данных.

Рис. 10. Детальное описание этапов конвейера монетизации данных

Нельзя отчаиваться, если гипотеза не принесла планируемого эффекта. По статистике больше 80% гипотез являются несостоятельными. Но 20% «выстреливших» гипотез покроют постепенно затраты по всем остальным направлениям.

6. Технологии управления данными

6.1. Общий взгляд на технологии управления данными

Теперь давайте разберемся, с помощью каких технологий можно достичь эффекта в управлении данными. Есть решения, которые считаются основными для таких проектов, а есть поддерживающие. Все зависит от степени заинтересованности бизнеса не только в понимании данных, но и в их качестве и защите.

На рис. 11 приведены технологии, которые закрывают основные потребности в части Data Governance:

- единый каталог бизнес-терминов (бизнес-глоссарий);
- каталог метаданных;
- средства проверки и обеспечения качества данных;
- средства выявления критичных данных и их защиты.

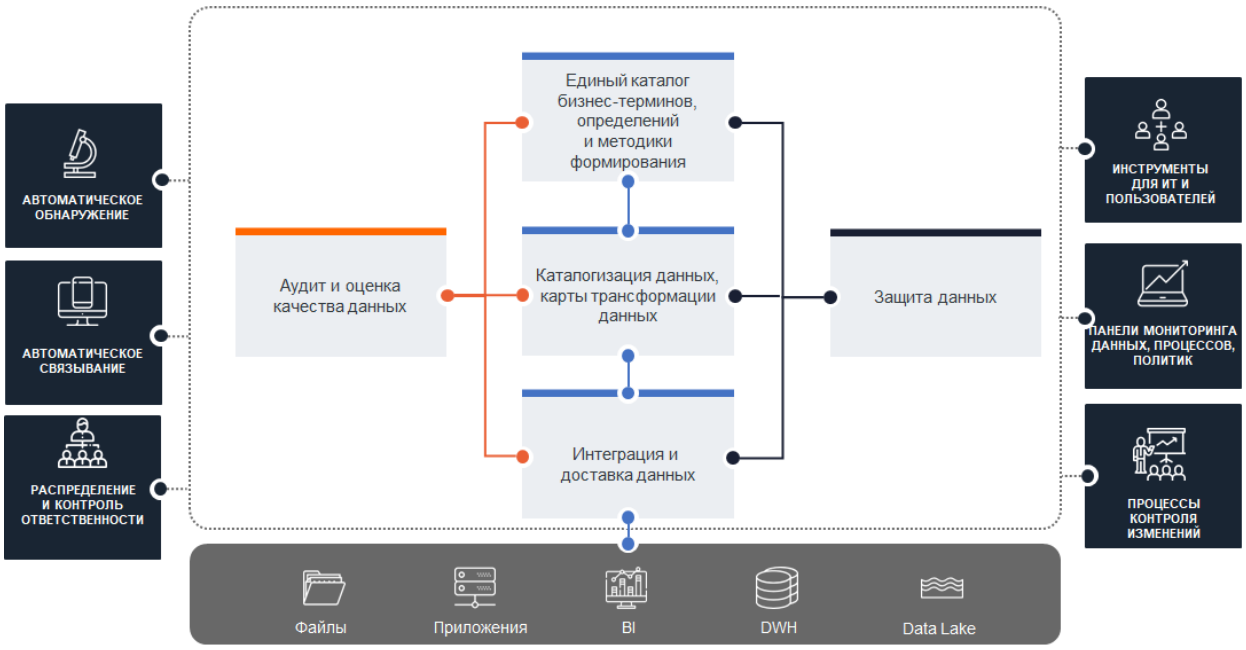

Рис. 11. Технологии Data Governance

6.2. Бизнес-глоссарий

6.2.1. Функции бизнес-глоссария

Бизнес-глоссарий как средство учета терминологии, используемой в компании, применяется очень давно. Традиционно именно ведение иерархии терминов считалось основой для понимания ценности данных и их влияния на бизнес. Но для целей управления данными этих функций явно недостаточно.

Основными функциями современного бизнес-глоссария считаются следующие:

- ведение концептуальной модели данных в парадигмах «AS-IS» и «TO-BE»;
- ведение логической модели данных в парадигмах «AS-IS» и «TO-BE»;
- поиск объектов любых категорий метаданных;
- ведение списков и ролей заинтересованных лиц (стейкхолдеров) для каждого объекта описания данных;
- ведение дополнительных категорий метаданных, ответственных за описание влияющих на данные бизнес- и технических аспектов (бизнес-процессы, рабочие группы и многое другое);
- ведение жизненного цикла данных;
- ведение запросов на изменение моделей данных;
- построение связей с физической моделью данных в каталоге метаданных;
- отслеживание и визуализация текущего уровня качества данных, описываемых логической и концептуальной моделями;
- оформление заказа на поставку данных (функция супермаркета данных).

6.2.2. Поисковые механизмы

Специалист CDOO, работающий с бизнес-глоссарием, должен в кратчайшие сроки найти описания искомых данных, просмотреть их владельцев, бизнес-логику создания, логику реализации и уровень качества данных согласно запросу. Поэтому каждый бизнес-

глоссарий должен иметь очень мощную поисковую машину (включая, например, «нечеткую логику» поиска) и хорошую визуализацию связей.

Стандартные возможности поиска бизнес-глоссария обычно в себя включают возможность найти объекты метаданных в одной категории метаданных с различными условиями расширения или уменьшения выборки. Крайне важно, мне кажется, чтобы для ускорения работы специалиста выбранный бизнес-глоссарий обеспечивал:

- выборку объектов метаданных одновременно из нескольких категорий метаданных (например, термины и наборы данных, системы и связанные с ними владельцы систем, регламенты и связанные атрибуты);
- возможность поиска по части текста с подстановкой;
- механизмы поиска не только классическими механизмами сравнения, а еще и с применением современных механизмов искусственного интеллекта или «нечеткой логики»;
- возможность обращения к объектам, к которым недавно было обращение;
- возможность сохранять поисковые запросы и делиться ими или их результатами с другими пользователями.

Последний приведенный пункт вносит существенное ускорение в работу специалистов разных ролей, позволяя им одновременно совместно работать над одним проектом. Не нужно тратить свое время на изобретение новых запросов для поиска, если это кто-то уже сделал. Но, естественно, такие сохраненные запросы должны быть грамотно описаны и структурированы. И это опять к вопросу регламентации работы и к функциям архитектора данных.

Еще один пример. Замечено, что в ряде компаний считают, что обычные пользователи из бизнес-подразделений будут постоянно использовать глоссарий для поиска и изменения объектов метаданных. А вот как раз пользователи часто не подозревают об этом и не планируют появления дополнительной нагрузки в своей работе. Все это выходит наружу при первых демонстрациях нового подхода по управлению данными для бизнес-пользователей. И негатив с их стороны, скорее всего, неизбежен. Для таких случаев в новом поколении бизнес-глоссариев предусмотрены функции небольшого фонового приложения, которое запускается, когда пользователь осуществляет поиск данных, помогая найти также и похожие термины из глоссария. Такое приложение может быть отдельным или встроено во внутренние интранет-порталы или системы управления знаниями.

6.2.3. Ведение концептуальной и логической моделей данных

Современные бизнес-глоссарии сочетают в себе инструменты ведения концептуальной и логической моделей данных, оставляя функции управления физической модели каталогу метаданных. Все объекты глоссария, содержащие в себе описание и связи с точки зрения работы компании, создают концептуальную модель данных подразделения или организации, а описание структур хранения и перемещения данных с логикой их обработки – логическую модель.

Обычно типы объектов, хранимых инструментом по ведению бизнес-глоссария, называют категориями метаданных. Каждая из них описывает данные со своей уникальной стороны (см. рис. 12).

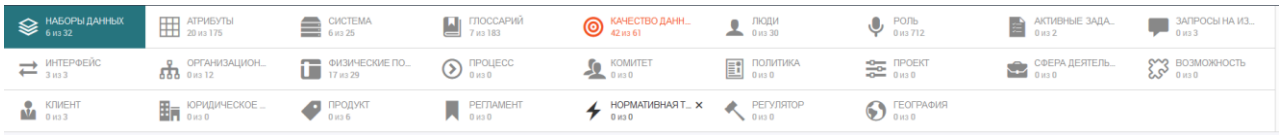

Рис. 12. Примеры категорий метаданных в бизнес-глоссарии

Чтобы сотрудники CDOO могли точно оценить степень влияния данных в каждом подразделении и организации в целом, нужно обеспечить наличие следующей информации:

- термины, присущие деятельности направления, их иерархию (см. рис. 13), синонимы, дубликаты и связи (см. рис. 14);
- бизнес-правила и бизнес-логику формирования терминов, особенно, в части показателей;
- наборы значений, форматов, масок и т.п., которые присущи данным, их жизненный цикл и статус;
- правила проверки качества данных и реальные уровни качества, которые можно соотнести с объектом данных на техническом или бизнес-уровне;
- всех заинтересованных лиц (стейкхолдеров) и подразделений, связанных с управлением или владением данными – владельцы и кандидаты на эту позицию, сотрудники с делегированными правами, эксперты, аналитики, стюарды, операторы и любые другие лица, которые занимаются данными, с любой точки зрения, в компании;
- связанные с данными бизнес-процессы компании, проекты, рабочие группы, их квалификация, внешние организации, влияющие на представление данных или являющиеся их потребителями;
- требования регуляторов, документы, внутренние и внешние распоряжения, регламенты, влияющие на вид данных, периодичность их создания и изменения, на их качество;
- связанные с данными системы, наборы данных, межсистемные интерфейсы, атрибуты, логика трансформации – все технические аспекты работы с данными, которые формируют логическую модель данных (см. рис. 15);
- связи логической и концептуальной моделей с физическим представлением данных в реальных системах;
- критичность для бизнеса и наличие средств защиты для данных, описываемых метаданными, доступность или конфиденциальность объектов;
- любые другие требования, которые характерны для конкретного бизнес-департамента или компании.

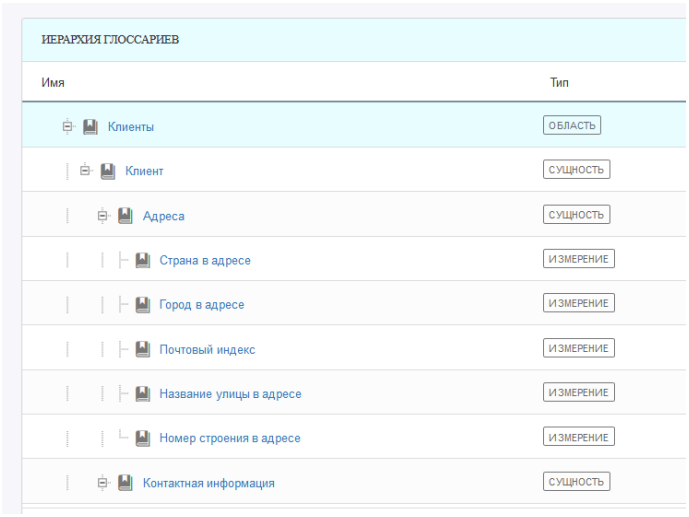

Рис. 13. Пример иерархии терминов – концептуальная модель данных в бизнес-глоссарии

Рис. 14. Пример связей между терминами – концептуальная модель данных в бизнес-глоссарии

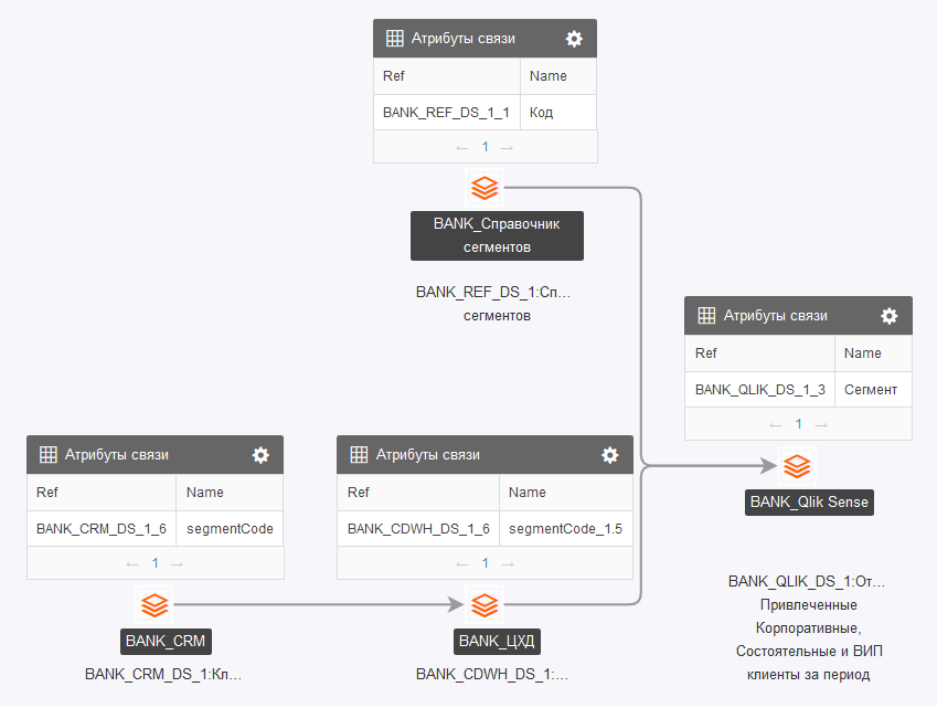

Рис. 15. Пример логических связей между наборами данных (датасетами) с указанием атрибутов

Часто возникает вопрос, что есть необходимость расширить число категорий метаданных для покрытия всех возможных комбинаций описаний данных, которые встречаются в работе специалистов CDOO. Как и в любом другом деле, во всем нужна мера. В противном случае через какое-то время будет невозможно разобраться в этих категориях, и сложность ведения такого глоссария приведет к его ненужности.

Обычно все конструируемые комбинации объектов можно описать в виде термина, указав его тип. Чаще всего объект «термин» в бизнес-глоссарии описывает следующие виды метаданных (см. рис. 16):

- сущность – термин, описывающий суть работы бизнес-подразделения (например, Клиент, Продукт, Цена, Скидка и т.д.);
- домен или область терминов – область, описываемая иерархией терминов, и верхний (корневой) объект этой иерархии;
- показатель – обычно расчетный объект в каком-либо отчете или выгрузке для анализа, а также в составе результатов анализа или расчета;
- измерение – сопутствующий объект, описывающий временные или другие справочные значения;
- справочник или измерение – объект, описывающий наличие списка значений, требуемых для описания других терминов;
- отчет – концептуальное представление, объединяющее расчетные показатели;
- правило – чаще всего применяется в качестве указания на наличие повторно используемой бизнес-логики.

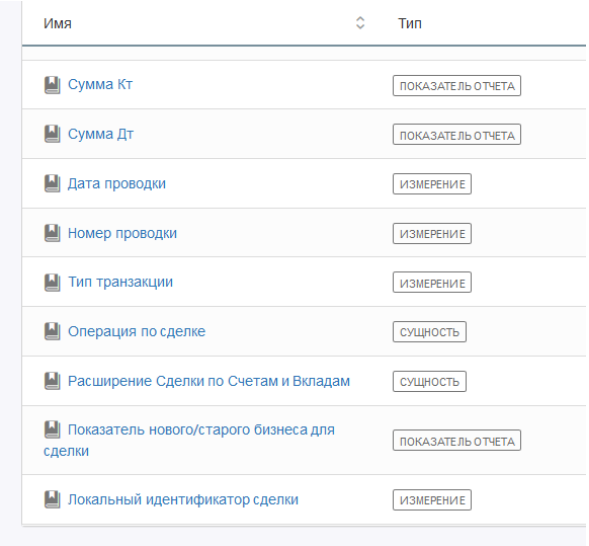

Рис. 16. Пример терминов разных типов в бизнес-глоссарии

Естественно, может быть много других вариантов, но, в большинстве своем, они покрываются описаниями терминов разных типов. Архитектор данных отвечает за состав и регламент используемых типов терминов, других категорий метаданных и типы связей между ними. А выбранное программное средство должно давать ему все возможности для конструирования моделей.

Еще одной важной задачей для архитектора данных является поддержание уникальности терминологии, описываемой в бизнес-глоссарии. Ни для кого не секрет, что одним и тем же словом можно описать самые различные вещи. Теперь посмотрим на любую крупную организацию – каждое ее бизнес-подразделение имеет свою терминологию, которая внешне похожа на термины соседнего департамента, но значит зачастую несколько другое.

Вот пример. В одной из крупных российских страховых компаний при смене руководства был поставлен вопрос о количестве клиентов. В течение полугода было получено восемь вариантов расчетов, разных подходов и ни один из них не совпал с любым другим. Вопрос остался без ответа. Почему так произошло? Каждый департамент понимает под словом «Клиент» что-то свое. Для одного подразделения счет идет по действующим контрактам, для другого – по выплатам в случае наступления страхового случая, а для путешественников, которые заключают контракт на время перелета или короткого пребывания в другом месте, вообще нет ясности, как их учитывать в общем расчете.

Вот и получается, что термин «Клиент», если его указать для приведенного примера в иерархии терминологии для каждого из таких подразделений, будет иметь различный бизнес-смысл. То есть, в глоссарии на уровне организации начнут появляться дубликаты. А теперь представьте себе стюарда данных, который должен ответить на запрос о дополнении какого-то отчета информацией о клиенте. И какого же из «Клиентов»-дубликатов ему нужно использовать для правильного описания задачи?

Именно поэтому важной особенностью правильного ведения корпоративного бизнес-глоссария является его уникальность. И это одна из основных задач архитектора

данных. Тогда появятся, для примера, термины «Клиент страхования автомобилей», «Клиент страхования жизни» и так далее.

В списке основных функций бизнес-глоссария упомянуты две основные парадигмы ведения концептуальной и логической моделей – «AS-IS» и «TO-BE». Они отражают подходы подразделений по управлению данными.

Подход «AS-IS» является чисто инвентаризационным, показывающим состояние уже созданных бизнес- и технических структур, что удобно при поиске любых существующих в организации данных. Однако такой подход всегда несколько запаздывает по сравнению с реальностью и несет в себе риски упущения изменений, произошедших с данными, особенно в том случае, если ИТ-служба и CDOO не синхронизируют информацию о планах обновления или существенного изменения систем.

«TO-BE» является противоположностью первой парадигме, предполагая ведение проектирования моделей данных на основе задач и ранее описанных объектов метаданных. Отсутствие какой-либо требуемой информации в модели приводит к необходимости описания всех дополнительных изменений для продвижения к основной цели. Из положительного, при правильном проектировании еще до этапа реализации будут проработаны все особенности реализации бизнес-требований, а после завершения работы ИТ-специалистами у стюарда данных появится возможность сравнения полученного решения с поставленными задачами (при наличии каталога метаданных).

Обычно применяют оба подхода, описывая основную часть концептуальной модели бизнеса и связанную с ней логическую модель данных, следуя за изменениями согласно требованиям бизнеса. Постепенно будет достигнута и полная инвентаризация всех объектов метаданных, но это не должно становиться основной целью, а жить вместе с организацией.

6.2.4. Жизненный цикл метаданных

Ранее упоминалось несколько раз о понятии «жизненный цикл метаданных». Сразу стоит отметить, что к жизненному циклу самих данных это имеет несущественное отношение.

Любое описание данных также проживает свою жизнь, и задача архитектора данных состоит в выстраивании основных вех для каждой категории метаданных. Проще всего рассмотреть жизненный цикл на терминах глоссария. Обычно выделяют такой набор стадий жизни метаданных:

- предположение;
- черновик;
- активный;
- кандидат на изменение;
- кандидат на удаление;
- удален.

Стадии жизненного цикла часто диктуются отделом методологии, если такой есть в организации. И, если опыт ведения концептуальной модели уже есть, то он легко должен переноситься на инструмент ведения бизнес-глоссария, не меняя сложившиеся правила.

Все стадии жизненного активно задействованы в запросах на изменение, которые являются необъемлемой частью процесса согласования любого изменения или проектирования. Важно, чтобы в процесс согласования был внесен обязательно элемент коллаборации, то есть совместной работы. Это может быть сделано посредством чатов или других средств коммуникации, привязанных к запросу и объектам метаданных, которые участвуют в планируемом изменении. В этом процессе обязательно участвуют все стейкхолдеры, имеющие отношение как к самому проекту, так и к связанным с ними другими объектами метаданных. Бизнес-глоссарий как программное средство должен иметь такие функции, чтобы пользователи не использовали несколько приложений, замедляя свою работу.

6.2.5. Супермаркет данных

Теперь поговорим о том, как с помощью бизнес-глоссария ускорить сам процесс запроса со стороны пользователей на поставку или изменение данных в виде отчета или выборки. Самый обычный способ – это электронная почта или другие корпоративные средства коммуникации. Но дать привязку к конкретным объектам такой способ не может. Сейчас в ряде средств управления метаданных появилась функция супермаркета данных или Data Marketplace. Ее особенность состоит в том, что пользователь может просмотреть доступные варианты данных – например, сделанные ранее кем-то отчеты, выгрузки данных, витрины… И тут же уделить внимание тому, что из себя эти данные представляют с точки зрения бизнес-ценности и управления ими. При необходимости бизнес-пользователь может тут же произвести заказ на доступ к этим данным или на предоставление других данных, которые присутствуют в логической модели в бизнес-глоссарии. В этом случае весь процесс обсуждения, согласования, внесения изменений и получения доступа осуществляется в одной системе.

Функции бизнес-глоссария как средства управления данными могут и будут расширяться в сторону облегчения работы бизнес-пользователей. Нужно следить за тенденциями развития, так как направление Data Governance крайне востребовано.

6.3. Каталог метаданных

6.3.1. Основные функции каталога метаданных

Следующим уровнем управления данных является каталог метаданных, хотя иногда его вместе с глоссарием называют каталогом данных. Я буду придерживаться первого названия – оно точнее отражает суть его работы.

Каталог метаданных представляет собой инструмент для реального понимания технической стороны деятельности организации. Можно назвать работу каталога «инвентаризацией», потому что он показывает текущее представление всех реализованных технических решений. К основным задачам каталога можно отнести следующие:

- сканирование физических метаданных систем хранения данных;

- сканирование физических метаданных систем интеграции, трансформации, передачи данных;
- сканирование систем проверки качества данных;
- построение визуальных связей зависимостей метаданных – data lineage;
- проведение анализа зависимостей (impact analysis);
- профилирование данных;
- выявление доменов данных на основе правил или искусственного интеллекта;
- автоматизированная привязка бизнес-терминологии к объектам физических метаданных;
- поиск среди объектов метаданных;
- сертификация объектов метаданных;
- поиск объектов, имеющих наборы данных, аналогичные другим объектам.

6.3.2. Задачи каталога метаданных

Использование каталога метаданных эффективно для следующих задач:

- проверка соответствия технической реализации поставленным бизнес-требованиям – стандартная задача для специалистов CDOO;
- инвентаризация и техническое описание систем с привязкой к глоссарию;
- изучение новых подключенных источников метаданных (например, при слиянии компаний или изучении системной архитектуры дочерних организаций);
- исследование функционирующих хранилищ или озер данных для вычистки неиспользуемых объектов - таблиц, витрин, отчетов – и удаления их самих и процессов их наполнения для сокращения временных затрат и на инфраструктуру.

Пользователями каталога метаданных являются технические специалисты. Бизнес-пользователи обычно не работают с такой системой, чтобы не вникать в особенности разработки каждой системы и межсистемной интеграции.

6.3.3. Системы-источники для каталога метаданных

Каждый каталог должен уметь работать, как минимум, с основными видами систем, установленными в большинстве организаций, для чтения их метаданных. К таким системам обычно относят:

- основные реляционные и нереляционные СУБД;
- ERP-системы;
- CRM-системы;
- BI-средства и системы аналитики;
- системы ведения хранилищ и витрин данных;
- основные производственные системы;
- средства интеграции данных, такие как ETL/ELT;
- скрипты SQL различных вариаций;
- сервисы передачи данных.

Часто многие системы-источники не требуется обрабатывать на уровне объектов, которыми система оперирует. Достаточно провести сканирование базы данных, которой система пользуется и где сохраняет свои данные.

Однако есть системы, для которых система хранения данных устроена настолько сложно, что с результатами работы каталога на уровне СУБД не даст никаких результатов. В этом случае каталог метаданных должен уметь отрабатывать взаимодействие с системой на уровне ее объектов, функций, сервисов и т.д.

А что же делать в том случае, если системы написаны непосредственно для компании или встречаются только на определенной территории или в единичном количестве? Для таких случаев в каталоге обычно предусмотрен специальный механизм создания отдельного соединения с системой, который предусматривает, что из системы есть возможность любым способом получить метаданные, затем распознать и преобразовать их в единый формат и провести сканирование средствами каталога. Такие виды подключения предполагают дополнительную работу по настройке, но зато снимают ограничение по составу систем для выявления состояния данных.

Результат сканирования представляет собой список объектов хранения или обработки данных, специфичных для каждого типа систем:

- для СУБД – таблицы, представления, их поля и форматы, процедуры и SQL-запросы;
- для ERP-систем – объекты хранения, объекты и процедуры обработки данных и т.д.;
- для BI – отчеты разных типов, агрегаты, сборки, поля и запросы;
- для систем интеграции данных – маппинги и процессы обработки данных.

Обычно сканирование метаданных устанавливается по расписанию на определенное время. Чаще всего это процесс проводят по ночам, когда процесс внесения изменений в системы и разработки уже завершен.

6.3.4. Состав объектов на уровне физической модели

После чтения метаданных каталог должен предоставлять возможность поиска среди найденных объектов. Аналогично бизнес-глоссарию, каталог должен уметь осуществлять поиск среди десятков, а иногда и сотен тысяч объектов, чтобы показать наиболее подходящие результаты пользователю. Поэтому поисковые возможности каталога метаданных не должны ограничиваться только стандартными возможностями по наименованию, но и иметь возможности подстановки текста, учитывать семантические особенности языка и дополнительные особенности, увеличивающие шанс найти нужный объект. К ним могут относиться признаки сертификации объектов экспертом, оценки пользователей, последние использованные объекты и многое другое.

В поиске всегда участвуют только объекты хранения или визуализации данных, а интеграционные потоки можно проследить в data lineage.

Выбрав нужный объект хранения, технический специалист сначала видит его структуру, понимая реальный состав его полей и форматов, его принадлежность системе, выявленные домены данных в виде присвоенных бизнес-терминов (см. рис. 17).

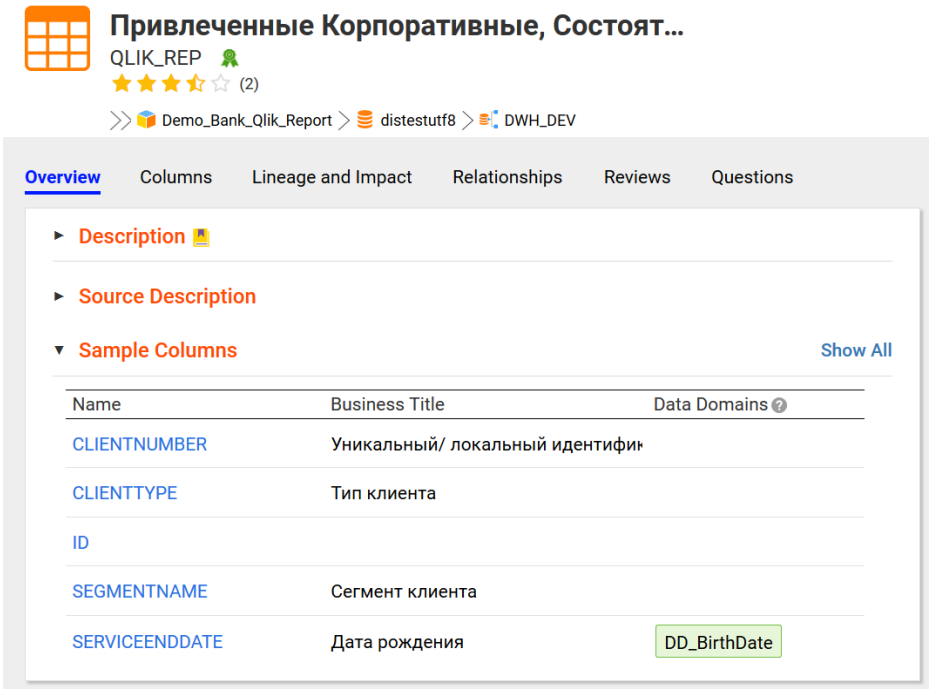

Рис. 17. Пример объекта метаданных после чтения каталогом

Детальное понимание состава полей нужно, в первую очередь, для правильного сопоставления структуры с требованиями бизнеса, в особенности, если объект является конечной структурой – витриной данных или отчетом.

6.3.5. Профилирование данных

Также важно понимать и состав данных по каждому полю. Современные каталоги позволяют выполнять профилирование данных – определение базового уровня качества, маску заполнения, частоту встречаемости значений и многое другое. Эта информация крайне нужна для понимания возможности использовать ресурс в дальнейшей работе специалистов. Довольно часто бизнес-требования выстраиваются на основе наличия поля по своей сути, но без учета реального качества данных, заполняющих его.

Приведу небольшой пример. В одном из банков требовалось выявлять дубликаты данных клиентов на основании основной банковской системы. Выбор полей был естественным – ФИО, дата рождения, место рождения. Однако в ходе профилирования выяснилось, что дата рождения заполнена только на 20%, а место – на 2%. Применение правила поиска пришлось сузить только до самых новых клиентов, для которых заполнение трех указанных полей стало обязательным. И пришлось придумывать еще много вариантов поиска дублирующихся записей на основе других признаков.

Примеры результатов профилирования данных объекта метаданных в каталоге метаданных приведены на рис. 18.

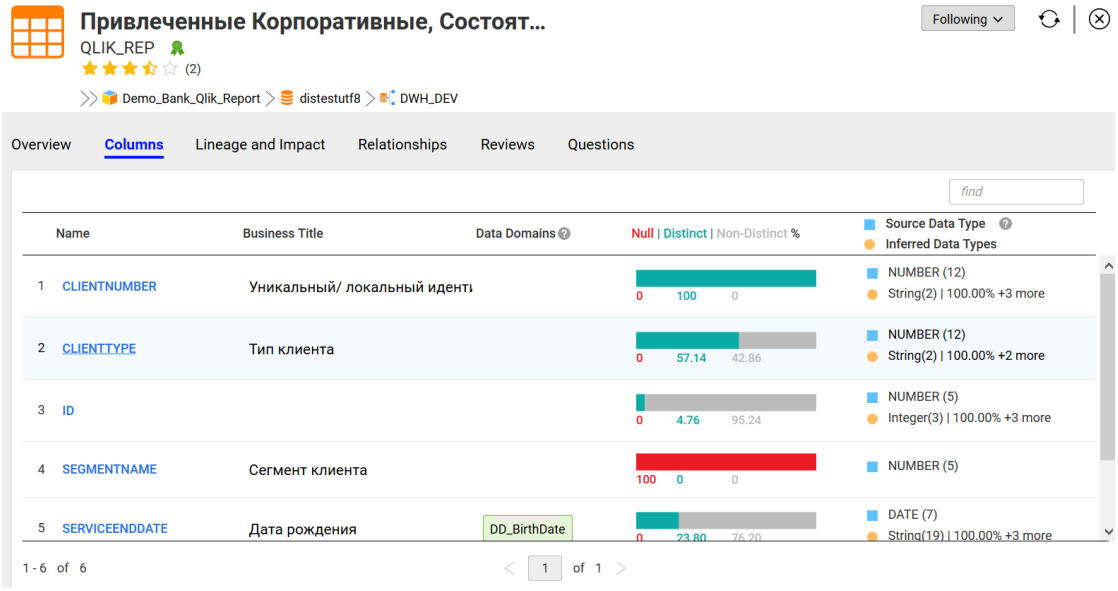

Рис. 18. Пример профиля данных

Отдельно необходимо выделить среди функций каталога возможность при подключении к нему новой системы определять, что за данные в ней находятся и нет ли таких данных в других системах. Это очень важная особенность современных каталогов данных, которая серьезно упрощает работу по описанию данных для сотрудников CDOO. Обычно подобные поисковые механизмы пользуются искусственным интеллектом, чтобы более точно предсказать, какие из данных являются близкими друг к другу по сути.

6.3.6. Data Lineage и impact analysis

Как бы то ни было, основным инструментом для работы специалистов в составе каталога остается data lineage. Это зависимость между объектами метаданных, которая выявляется по запросам, сервисам, интеграционным процессам и визуализируется каталогом. Промышленные каталоги метаданных позволяют менять размер и глубину просматриваемой области связей, так как иногда в нее попадают тысячи объектов, что заставляет ждать пользователя. Чтобы избежать задержки, обычно data lineage строится сначала в общем виде, давая возможность пользователю выбрать требуемую глубину и направление просмотра. На рис. 19 приведен пример полного детального раскрытия data lineage с учетом каждого поля изучаемого объекта метаданных.

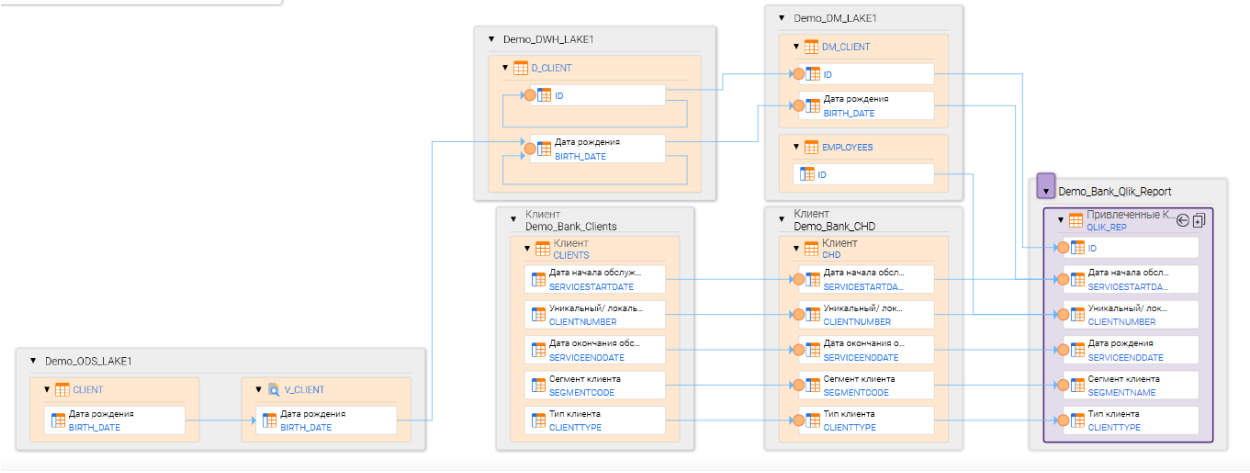

Рис. 19. Пример детального представления data lineage

Важной функцией промышленных каталогов метаданных является детализация взаимосвязей до уровня отдельных трансформаций над данными. Чаще всего, такие требования применяются при изучении потоков данных, реализованных на ETL/ELT-средствах или SQL-запросах.

Анализ влияния (или impact analysis) – это действия по изучению влияния изменений в одних системах на другие на основе построенного data lineage. Так как каталог метаданных сканирует системы обычно ежедневно, то при обновлении какой-либо из систем возможно образование разрывов в зависимостях, что однозначно указывает на появление неверных данных затем в отчетах и аналитике. Это будет сигналом к действию стюарда данных по информированию бизнес-подразделений о наличии проблем с отчетностью, а для ИТ-служб – указанием, какие нестыковки необходимо устранить. Одновременно это же служит поводом для руководителя подразделения по управлению данными к изменению организационной политики по предварительному извещению об изменениях в системах с его обязательным подтверждением.

6.3.7. Связи моделей данных

Вернемся к тому, почему каталог важен при управлении данными. Бизнес-пользователь со стороны глоссария должен иметь возможность перейти к физической реализации систем, чтобы увидеть соответствие требованиям. И с другой стороны, технический специалист должен понимать, с чем он работает. Название таблицы и поля зачастую не могут дать никакой дополнительной информации пользователю. В ряде банковских систем, особенно, западной разработки такая ситуация встречается часто.

Для таких случаев каталог может в качестве одного из ресурсов сканирования использовать сам бизнес-глоссарий, чтобы присвоить затем нужный термин нужному техническому объекту метаданных. Тогда каждый объект и сам data lineage смогут в себе содержать как технические, так и бизнес-составляющие, усиливая понимание реальной ситуации с данными (см. рис. 19).

Важно, что каталог дает возможность перейти в каждый объект для просмотра его структуры и качества, а также построить data lineage из любого выбранного объекта в нужную сторону.

6.3.8. Выявление доменов данных

Особенным механизмом этого типа программных инструментов является выявление доменов данных и присвоение техническим объектам соответствующих терминов. Мы уже упоминали об этом в разделе о бизнес-глоссарии, который ожидает от каталога найденных объектов, подходящих под бизнес-описание, для построения связи между термином и объектом физической модели.

Есть несколько вариантов выявления такого соответствия. Рассмотрим основные из них.

1. Построение соответствия на основе названия физических объектов. Оптимальный способ, если система спроектирована так, что по названию таблиц и полей можно сделать вывод о сущности хранимых данных.
2. Построение соответствия на основе соответствия данных определенному списку. Этот метод использует обычно искусственный интеллект для выявления домена данных. Естественно, каталог нужно заранее обучать на существующих данных для более точного результата. Такой способ хорошо подходит для данных, содержащих информацию описательного характера – фамилии, имена, названия продуктов и т.п.
3. Выявление домена на основании правил и условий. Этот метод хорошо работает при возможности выявить маску заполнения данных (даты, номера телефонов, адреса электронной почты и т.д.) и наложения уточняющих условий. Хотя самое большое количество ошибок в результатах обычно дает именно этот метод. Например, как выявить дату рождения или дату транзакции? Мы быстро можем выявить, где находится дата, а вот условия, определяющие именно этот тип даты, надо серьезно продумать заранее и протестировать на используемых массивах данных.
4. Комбинированные методы. Естественно, делать ставку только на один из перечисленных выше способов выявления доменов данных часто неразумно. Обычно используется сочетание методов работы с данными и названиями метаданных. Но увеличение числа правил методов приводит к существенному увеличению времени, которое затрачивает каталог на сканирование. Для ежедневного исследования метаданных использовать такие методы нужно осторожно, чтобы успеть завершить все расчеты к началу рабочего дня специалистов (если, конечно, сканирование выполняется ночью).

Любой метод не дает точного попадания в выбранный домен. Поэтому многие каталоги определяют процент соответствия, а также имеют правила, которые дают возможность сделать присвоение физическому объекту термина при превышении определенной границы в процентах. Если процент не столь высок, чтобы сделать операцию автоматически, обычно у эксперта или архитектора данных появляется задача на ручное присвоение термина с рекомендацией от системы.

В результате выявления доменов данных каталог подтянет бизнес-описания из глоссария, проведет связи с объектами физической модели, может изменить логическую модель данных, а также оповестить заинтересованных лиц об этом.

Теперь у специалиста есть возможность не только посмотреть сам объект метаданных, но и его бизнес-описания из глоссария. Кроме того, бывает необходимо понять весь набор физических объектов, попадающих под нужный домен данных.

Вообще, самой длительной операцией при настройке каталога метаданных является именно обучение системы и построение нужного набора правил для выявления доменов и построения связей с бизнес-глоссарием.

6.3.9. Каталог каталогов

Иногда в организации уже есть один или несколько каталогов, которые умеют работать с определенными наборами метаданных. Часто крупные производители программного обеспечения выпускают возможность анализа физической модели для своих систем. Но такие решения бывает сложно расширить на изучение других систем.

Появившийся не так давно термин «Каталог каталогов» означает возможность каталога метаданных обращаться к локальному средству, работающему для узкого круга систем, для получения информации о структурах хранения и перемещения данных. Этот подход позволяет:

- не нагружать пользователей бизнес-подразделений несколькими системами для исследования метаданных;
- получить полную информацию и зависимости;
- не проводить сканирование метаданных несколько раз.

При таком подходе без дополнительной нагрузки на системы-источники метаданных каталог каталогов забирает данные из доступных «более узких» каталогов, становясь «зонтичным» решением, позволяющим сотрудникам CDOO выполнять свою работу вне зависимости от типов применяемых в организации систем (см. рис. 20).

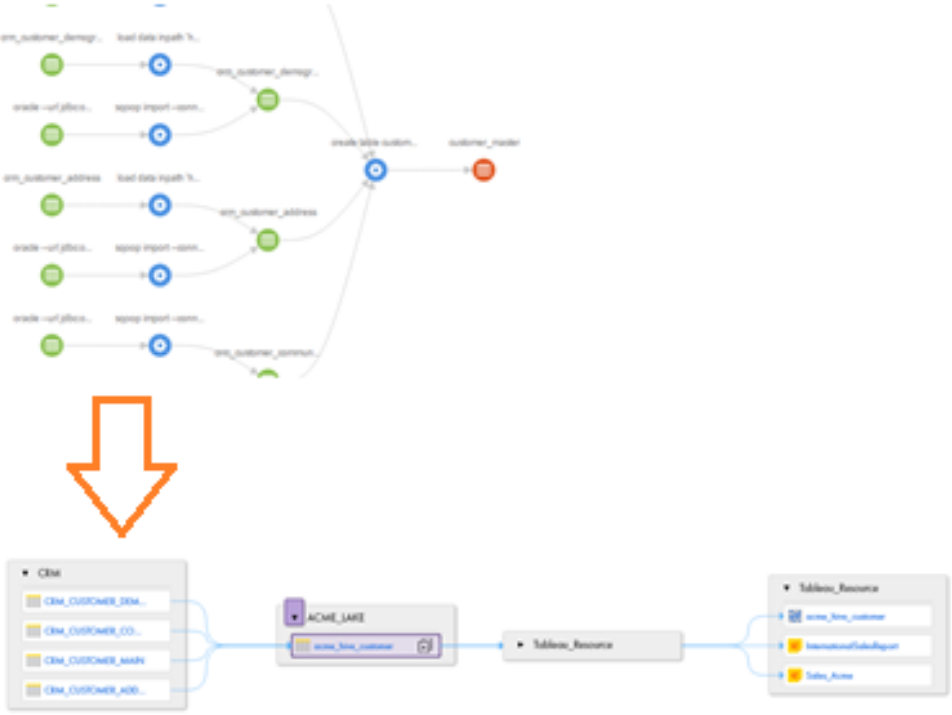

Рис. 20. Пример каталога каталогов – data lineage внутри одного решения становится частью более крупного data lineage в каталоге каталогов

6.4. Проверка качества данных

6.4.1. Функции средств обеспечения качества данных

Раньше уже затрагивался вопрос о важности понимания качества данных при рассказе о профилировании как функции каталога метаданных. Средства обеспечения качества данных не являются основными для управления данными, но вносят существенный вклад в знания о текущем их состоянии.

Большинство средств обеспечения качества данных имеют в своем составе следующие функции:

- профилирование данных;
- детальное исследование качества данных, включая поиск связей между данными и выявление доменов данных;
- стандартизация (нормализация) данных – приведение их к требуемому виду по формату, маске, профилю;
- выявление и устранение дубликатов записей различными методами;
- отчетность по качеству.

Для целей управления данными важны только проверки качества данных, которые выявляются при профилировании или детальном исследовании. Функции обеспечения данных обычно применяются в качестве последующих действий, если выявленное качество не удовлетворяет требованиям бизнес-пользователей.

Вместе с запросом на предоставление каких-либо данных бизнес-пользователь может передать требования к качеству предоставляемых данных. Причем для одних и тех же наборов данных уровень нужный качества для разных подразделений может быть разным. Именно для стюарда данных поэтому важно видеть какие правила качества данных используются и какие результаты проверок они дают в разрезе искомой информации (см. рис. 21).

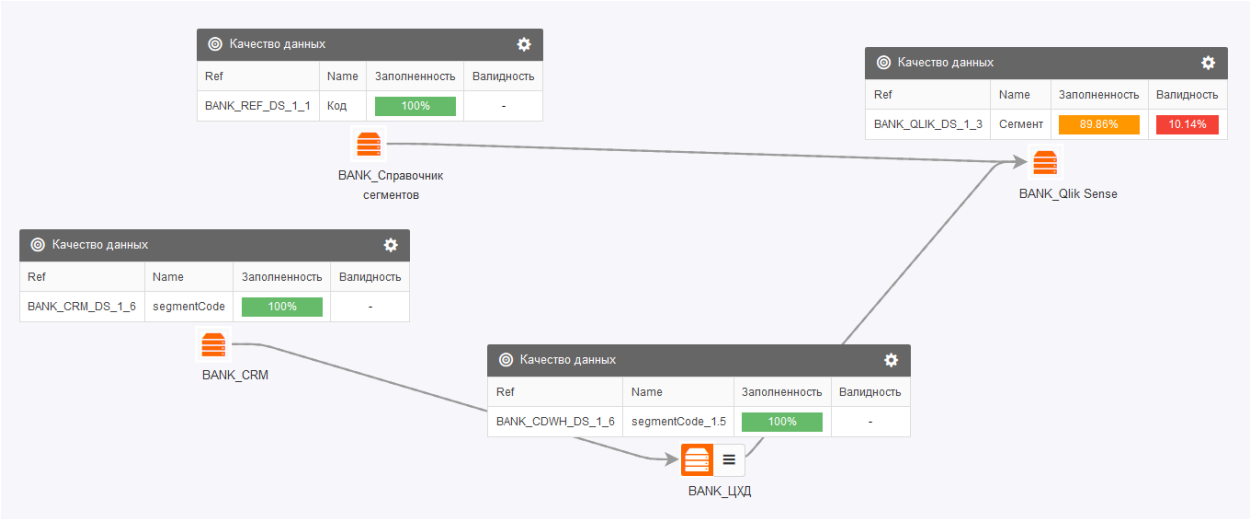

Рис. 21. Пример связей на логическом уровне с указанием реальных уровней качества данных в системах организации

6.4.2. Реестр правил качества

Как я уже указал выше, стюард данных должен понимать, как конкретно проверяются данные для понимания их качества. Для этого должен быть заведен реестр правил качества с детальным описанием сути каждого правила, а также тех объектов физической модели, на которых оно применяется. Такой реестр позволяет точно установить, правильный ли набор проверок применен к данным, а также избежать многочисленного дублирования проверок.

Во многих организациях, где функции ИТ распределены между разными подразделениями (например, свои разработчики в составе финансового блока и т.п.), часто возникает ситуация «замусоривания» интеграционных процессов проверками качества данных. Как она возникает?

При разработке хранилища данных для какого-то департамента разработчики поставили правила проверки качества данных согласно требованиям из бизнес-требований, а также собственные дополнительные процессы, которые не пропускают некачественные данные дальше. При расширении того же хранилища под нужды других подразделений никто не обратил внимания на существующие правила, и были встроены новые проверки, хотя они и являлись дубликатами уже действующих. И так далее. Отсутствие единого ответственного за проверки качества и реестра самих правил приводит к тому, что одни и те же данные многократно проверяются. Но требования по производительности подготовки данных для отчетов никто не отменял, поэтому увеличение числа правил приводит к расширению инфраструктуры – требуется всё больше мощностей, а причина кроется в незнании реальных процессов.

Даже при сканировании таких ресурсов каталогом метаданных возникает большое количество вопросов о том, что же за процессы работают с данными и что они с ними делают. Поэтому такая позиция как «офицер по качеству данных», которая обычно находится в департаменте ИТ или CDOO, должна обязательно иметь место, собирая данные обо всех правилах изучения данных и приводя их количество и функции в единый вид.

Реестр правил качества обычно ведется в двух инструментах:

- в бизнес-глоссарии – описание правил проверки (см. рис. 22);
- в инструменте обеспечения качества данных – бизнес-представление проверок в средствах обеспечения качества данных и их техническая реализация (см. рис. 23).

Ссылка	Имя	Описание	Имя атрибута	Измерено в	Тип	Критичность	Результат
BANK_DQ_1	Корректность заполнения номера лицевого счета	Номер в числовом формате, 20-ти значный	numberAccount	BANK_IBSOCore	Валидность	1	100.00%
BANK_DQ_2	Корректность заполнения номера лицевого счета	Номер в числовом формате, 20-ти значный	numberAccount	BANK_IBSODeposit	Валидность	1	33.33%
BANK_DQ_3	Проверка валидности заполнения поля ClientNumber	Проверка соответствия номера клиента формату, заданному в справочнике клиентски... | Развернуть |	ClientNumber	BANK_IBSODeposit	Валидность	2	0.00%
BANK_DQ_4	Проверка валидности заполнения поля ClientType	Проверка того, что тип клиента указан верно и соответствует справочнику типов к... | Развернуть |	ClientType	BANK_CRM	Валидность	2	100.00%
BANK_DQ_5	Проверка валидности заполнения поля ClientNumber (CRM)	Проверка соответствия номера клиента формату, заданному в справочнике клиентски... | Развернуть |	ClientNumber	BANK_CRM	Валидность	2	100.00%
BANK_DQ_6	Проверка валидности заполнения поля ServiceStartDate	Проверка корректности указания даты начала обслуживания: дата начала обслуживан... | Развернуть |	ServiceStartDate	BANK_CRM	Валидность	2	75.61%

Рис. 22. Реестр правил качества данных в бизнес-глоссарии

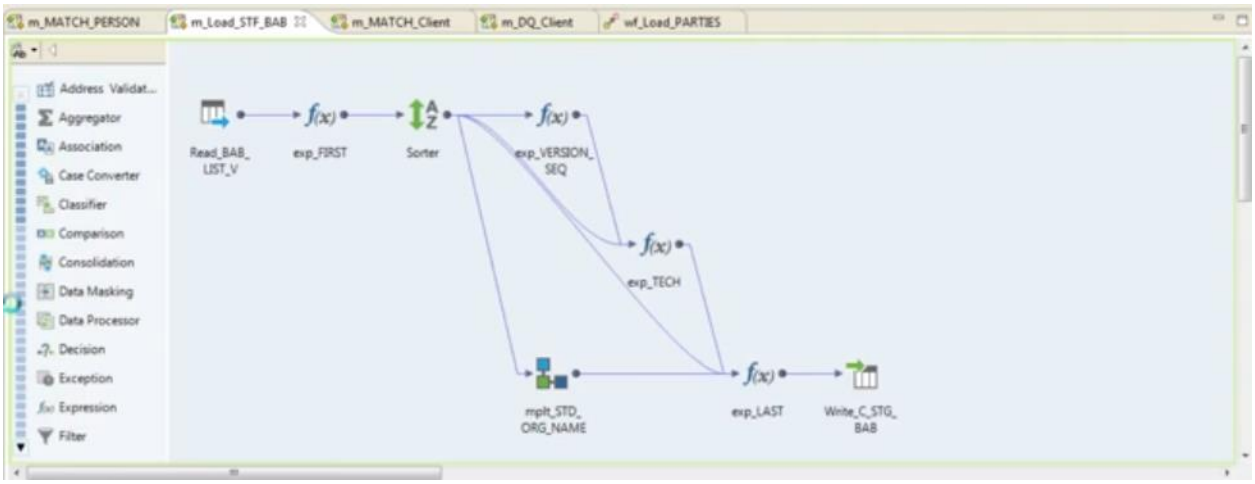

Рис. 23. Техническое представление правила проверки качества в инструменте управления качеством данных

Для каждого правила качества в реестре обязательно также указываются заинтересованные лица (стейкхолдеры) и условия проверки.

Инструмент обеспечения качества данных должен давать возможность бизнес-пользователю возможность создания собственных правил без привлечения ИТ-разработчика, а также использовать более ранние разработки.

Такая функция может быть реализована и в бизнес-глоссарии, давая возможность пользователю создать новое правило проверки с автоматической генерацией его в средстве обеспечения качества по ранее выбранному шаблону.

Оба этих случая очень важны для ускорения работы с данными без проволочек между подразделениями. Например, стюард при просмотре в глоссарии зависимостей между терминами обращает внимание, что нужные ему данные проверялись только на системе-источнике и непосредственно перед построением отчета, и результаты проверок существенно различаются. Значит, он однозначно будет разбираться, в какой из промежуточных систем или интеграционных процессов качество резко меняется и по какой причине. И ему потребуется создать дополнительный набор, скорее всего, идентичных правил проверки, чтобы найти место возникновения проблемы.

С другой стороны, неконтролируемое создание подобных правил может привести к «замусориванию», поэтому регламент обеспечения качества должен быть введен в компании с обязательным мониторингом изменений, которые происходят с правилами качества данных.

6.5. Критичные данные и их защита

6.5.1. Выявление критичных данных и информирование об уровне их защиты

Еще одним поддерживающим типом технологий для Data Governance являются средства выявления критичных данных для организации и инструменты их защиты.

Часто функция информационной безопасности не выходит за рамки одноименного подразделения, выходя наружу только в виде согласований требований от бизнеса, технических заданий и самой программной реализации.

Но, согласитесь, что бизнес-пользователю (особенно, стейкхолдеру) тоже требуется знание, являются ли используемые им данные критичными для бизнеса компании в целом, а также, насколько хорошо они защищены.

Средства выявления критичности данных являются довольно новыми на рынке. В основном, они в себе содержат следующие функции:

- определение доменов данных (функция каталога метаданных);
- ведение реестра критичных данных с привязкой к доменам данных (см. рис. 24);
- определение зависимостей критичных данных с другими (функция каталога метаданных);
- выявление уровня защищенности данных с учетом связей и примененных средств защиты;
- интеграция с системами мониторинга обращений к данным для выявления аномалий по обращению к критичным данным;
- информирование о существующем уровне защищенности и возникающих аномалиях.

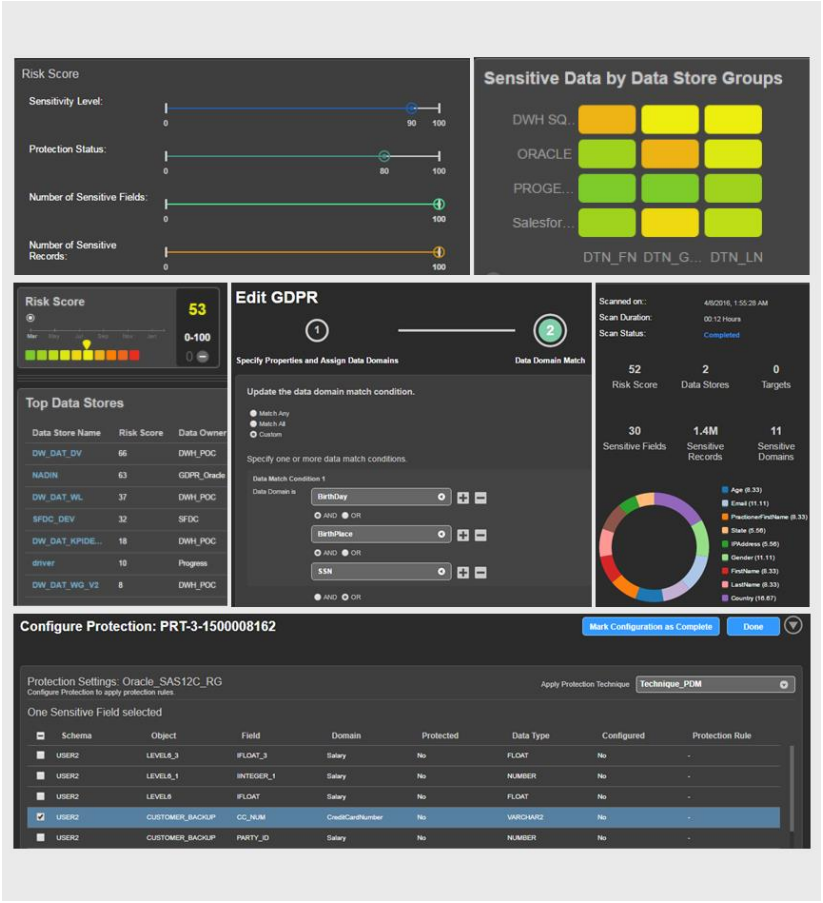

Рис. 24. Реестр критичных данных и оценка их защищенности

Для управления данными такие средства несут дополнительную информацию, насколько термин, набор данных, атрибут являются критичными для бизнеса данными. Обычно эти данные доступны через бизнес-глоссарий в качестве дополнительных параметров, полученных путем интеграции.

При необходимости пользователь должен иметь возможность информировать владельцев системы о необходимости изменения уровня критичности, если, например, ранее данные рассматривались доступными для всех.

6.5.2. Средства защиты данных

Решения, ограничивающие доступ к данным, крайне разнообразны. Но, в рамках деятельности CDOO, мы обратим внимание на две группы средств защиты данных, которые не решают вопросы разграничения доступа, а основываются на их обезличивании или маскировании.

Таких решений обычно выделяют два типа:

- инструменты статического обезличивания данных;
- инструменты динамического маскирования данных.

Эти решения применяются по запросу со стороны CDO в случае понимания, что часть данных, критичных для бизнес-подразделения согласно его запросу, может быть доступна другим пользователям или внешним разработчикам.

6.5.3. Средства статического обезличивания

Средства статического обезличивания представляют собой программные решения по физическому искажению данных. Обычно основной сферой их применения являются среды для разработки и тестирования внутренними и внешними командами. Особенность таких решений состоит в том, что они должны создать данные, крайне похожие, валидные для тестирования, но без возможности получить по ним доступ к реальной информации.

Валидность данных определяется на основе их структуры, применяемой маски значений, условий и других требований. Например, если поле содержит фамилии клиентов, то оно и должно после обезличивания содержать фамилии, но другие, и лучше с привязкой к полу клиента. Или, если даты транзакций находятся в определенном диапазоне, то и новые искаженные даты должны попадать в этот диапазон.

Важной функцией таких систем обязательно должна являться возможность создания единой модели обезличивания данных в разных системах-источниках. Под этим я понимаю однотипность применяемых принципов создания новых данных для одних и тех же доменов данных.

Вернемся к примеру с фамилиями. В трех системах, не связанных друг с другом интеграционными процессами, есть имена клиентов. Новая разработка основывается на этих данных как на логических связях. Если при создании тестовых данных не учесть эту особенность, то связи могут потеряться, и тестирование будет неполным. То есть, проблемы проявятся уже в процессе промышленной эксплуатации.

Именно поэтому в системах статического маскирования данных присутствует функция единой модели обезличивания, дающая затем в процессе разработки учесть и логические связи, не представленные на физическом уровне.

6.5.4. Средства динамического маскирования

Инструменты динамического маскирования применяются для ограничения доступа к реальным данным. Они используются для защиты часто изменяемых данных и структур. Для департамента управления данными основным поставщиком метаданных часто являются хранилища и озера данных. А их структура (это особенно важно для озер данных) постоянно меняется, из-за чего классические средства ограничения доступа могут запаздывать или не срабатывать.

Почему это важно? Озера часто используются бизнес-подразделениями для своих аналитических исследований. В озере данных работают сайентисты (аналитики, математики), которые в рамках своих операций постоянно создают новые структуры хранения, наполняя их данными для проверки гипотез. Их работа динамична, а ограничения со стороны подразделений безопасности с необходимостью согласования каждого шага сведут эффективность их работы «на нет».

Для таких случаев хорошо применять решения по динамическому маскированию, которые не обращаются к данным, не меняют данные или права доступа к ним. Эти

средства анализируют запросы к данным и заменяют их в случае несоответствия требованиям (например, в части критичности) пользователя, времени обращения, системы, состава полей и многих других.

Вне зависимости от того, из какой среды поступило обращение, система динамического маскирования выявляет наличие в запросе защищаемых данных, остальные параметры, как описано ранее, и создает при необходимости другой запрос к данным или возвращает отказ.

Примечательно, что такое решение оказывает минимальное влияние на производительность по предоставлению данных, так как с самими данными оно не работает, затрачивая свое время на обработку запросов. Но и тут нужно быть осторожным, так как в высоконагруженных системах, где количество запросов к данным может исчисляться тысячами в секунду, скорость анализа будет существенно ниже требований к производительности.

7. Примеры внедрений

Рассмотрим несколько примеров удачных и неудачных, на взгляд автора, внедрений.

Одна из крупных нефтяных компаний поставила перед собой цель в увеличении объемов продаж на автозаправочных станциях. Цель, в общем, довольно стандартная, если не считать того, как она достигалась до внедрения проекта по управлению данных.

Большинство аналитиков, гипотезы которых должны были изменить подход к продажам, занимались изучением, где данные взять, как они связаны между собой и можно ли их вообще использовать с учетом их качества.

Для себя компания определила ряд особенностей, усложняющих решение задачи:

- наличие кросс-функциональных моделей данных с большим числом источников (более ста);
- высокая доля внешних источников данных;
- многократное дублирование данных;
- высокая чувствительность моделей data science к качеству данных;
- отсутствие «единой версии правды» данных;
- несвязность бизнес-архитектуры с ролевой моделью;
- высокая скорость изменений;
- большой объем работы в «песочницах», который не передается затем в промышленную эксплуатацию.

Первым шагом началось решение задач, связанных с качеством данных:

- создание единой точки приемки обращений по нарушениям качества данных в контуре BI-системы;
- постоянный мониторинг и контроль качества данных, использование тематических панелей и закладок в BI-системе;

- формирование базы знаний о нарушениях в данных, выработка правил выявления нарушений и контрольных соотношений по проверке данных;
- контроль автоматизации правил проверки корректности данных в контуре смежных систем;
- разработка методологии для нормативно-справочной информации, развитие и сопровождение системы управления НСИ.

Подход «умного озера данных» для выхода из сложившейся ситуации состоял в следующем:

- все работы аналитиков перенесены в единую «песочницу» - озеро данных;
- создание единого управляемого каталога бизнес-правил по качеству данных;
- внедрение методологии и инструментов для связывания правил по качеству и наборов данных;
- мониторинг качества на цепочках происхождения данных от источника до пользовательских витрин и BI-приложений;
- построение связей бизнес-терминов и правил по качеству данных в бизнес-глоссарии;
- создание пользовательского портала для просмотра глоссария, происхождения данных и правил по качеству данных, а также статистики по каждому правилу;
- вовлечение бизнес-сотрудников в развитие системы.

В результате удалось за несколько лет работы достичь следующего:

- снижение затрат и сроков для вывода новых продуктов на рынок до 40%;
- более 50% активных аналитических проектов используют единую платформу;
- большая часть всех таблиц и витрин хранилища охвачены правилами и метриками по качеству данных;
- созданы инструменты онлайн-мониторинга и статистики качества данных для витрин;
- объекты глоссария (бизнес-термины) связаны с правилами по качеству данных;
- портал управления данными доступен для всех сотрудников;
- ведется постоянная работа по закреплению владельцев данных.

Важно отметить, что при реализации проекта не выделялось отдельно направление управления данными или ведения озера данных. Эти проекты поддерживают основных пользователей – аналитиков – в их работе, ускоряя ее, давая им прозрачные и качественные данные, а также полное понимание возможности их использования.

Еще один пример – из телекоммуникационной компании.

За десять лет использования хранилища данных в нем накопилось довольно много разнообразных данных. Отдельной службы по управлению данными хранилища не велось, поэтому разрастание объемов вело к постоянной необходимости наращивания серверных мощностей и мест для его установки. При этом было необходимо сохранять нужную производительность, чтобы не задерживать основной бизнес компании.

Катализатором для начала решения проблемы послужило изменение требований регулятора по резкому увеличению сроков хранения данных о звонках и других видах коммуникаций. Среди действий по решению поставленных задач одно касалось выявления реальной ситуации по используемости данных хранилища.

Применение каталога метаданных показало реальную картину по существующим таблицам и отчетам, процессам их наполнения, базовому уровню их качества. Одновременно было проведено исследование по количеству и частоте обращения к данным через системы отчетности.

Результат не заставил себя ждать, показав, что около 70% всех структур и отчетов не используются вообще или частично. Основной причиной для такого результата явилось постепенное изменение источников и требований к данным от бизнеса, но в течение всего срока жизни хранилища. Однако при выполнении очередного изменения старые структуры оставляли «на всякий случай», а потом про них забывали.

Такая «чистка» хранилища данных затратила около полугода времени и затраты одного человека на указанный срок, существенно скорректировав затраты на инфраструктуру. Кроме того, отслеживание изменений в хранилище данных стало отдельным процессом в ИТ.

Однако есть примеры не самые удачные, которые важно знать, чтобы избежать ряда ошибок. Я приведу один из них, который наблюдал не раз и в банках, и в производственных компаниях.

В одном из блоков компании создается служба по управлению данными. Причем ее основной задачей становится не решение насущных проблем, а создание методологии и описание текущего состояния бизнеса. Обычно на написание всего набора документов уходит порядка года, после чего служба выходит с предложением внедриться в текущие проекты, чтобы провести их по новой методологии.

Сразу становится ясно, что при отсутствии выстроенных процессов хотя бы на небольшом прототипе эта инициатива начинает подвергаться жесткой критике со стороны как внедряющих подразделений, так и самих бизнес-пользователей. Они все понимают, что сроки проекта будут увеличены, а затраты возрастут. В результате инициатива не приживается, а подразделение распускается. Созданная концептуальная модель данных устаревает довольно быстро, как и методология, которая не была привязана к конкретной ситуации в компании.

Аналогичные проблемы могут возникнуть, когда одно из подразделений проводит более практические исследования, но также не привязанные к реальной задаче в компании. Результаты таких действий обычно кладут «на полку».

8. Небольшие рекомендации

В завершение приводится ряд рекомендаций, которые, с точки зрения автора, помогут выстроить целостную модель Data Governance в организации.

1. Не ставить управление данными во главу угла. Это важная, но поддерживающая функция в организации.
2. Создать совместные проектные команды из представителей бизнес-подразделений и ИТ для создания общих видения и стратегии управления данными.
3. Определить ключевые задачи для управления данными, инициативы по их решению и включить их в корпоративную стратегию по управлению данными.
4. Определить политику управления изменениями в области данных.
5. Выбрать первые проекты, исходя их существенной значимости, наличия доступных данных, интереса от бизнес-пользователей и коротких сроков решения (3-5 месяцев на первый проект).
6. Не начинать управление данных с общеописательных инициатив. Они обычно не приживаются в компаниях.
7. Вовлечь влиятельных бизнес-спонсоров со стороны руководства организации.
8. Разработать комплексную программу проектов, которые принесут дополнительную ценность организации.
9. Внедрить политики управления данными и назначения владельцев данных на высшем уровне.
10. Не затевать сразу резкое изменение бизнес-процессов. В каждой организации они будут своими, часто не похожими на идентичные по сути бизнес-процессы в других компаниях. Изменение процессов должно происходить плавно без негатива со стороны бизнес-подразделений.
11. Двигаться от проекта к проекту, постепенно описывая данные подразделений. Не делать попыток сразу работать со всеми.
12. Для каждого проекта описывать данные на всех уровнях моделей представления, а также их качество, исходя из нужности для бизнеса. Не описывать все, что имеется.
13. Вовлекать бизнес-пользователей как экспертов для уточнения концептуальной и логической моделей данных.
14. Обучать сотрудников бизнес-подразделений по мере появления результатов для них.
15. Не пытаться внедрить сразу все виды ПО для управления данными. Это только задержит старт инициатив по управлению данными. Первичное изучение данных и требований к ним даст понимание, что конкретно нужно сейчас и последовательность внедрения других продуктов.
16. Оценить наборы правил качества – существующие и требуемые. Отсечь ненужные наборы правил.
17. Создать реестр правил качества и максимально переиспользовать существующие правила для идентичных проверок.
18. Не делать отдельно проекты по очистке данных, если до этого не изучены реальные параметры и проблемы качества, а также не изучено, где еще в компании есть такие данные.

В каждой компании управление данными идет по-своему. Я не думаю, что есть единый рецепт для быстрого эффекта. Но как раз постепенное увеличение доли проектов с использованием Data Governance сделает вашу работу быстрее и прозрачнее, что и приведет всю компанию к успеху. Ведь каждый из нас, работая более оптимально, и дает такой результат для компании в целом.

И пусть у вас все получится!

9. Ссылки.

1. Chief Data Officer. Synthèse des ateliers. Informatica, 2017.
2. Informatica Velocity. Data Governance Maturity Assessment. Susan Wilson, 2012.
3. Ток-шоу Game Changers. Выпуск №3. Светлана Бова, Александр Тарасов. DIS Group, 2019.
4. Designing a data transformation that delivers value right from the start. Chiara Brocchi, Davide Grande, Kayvaun Rowshankish, Tamim Saleh, Allen Weinberg. McKinsey & Company, 2018.
5. Методология моделирования Data Governance. Вебинар №8. Александр Тарасов, Станислав Уштей. DIS Group, 2018.
6. Практика управления качеством данных. Иван Черницын. Газпром нефть, 2020.
7. The Chief Data Officer Handbook. Sunil Soares. MC Press Online, 2014.
8. Data Management Book of Knowledge. 2nd Edition. DAMA International. Technics Publications, 2017.