МАГИЯ управления данными: понимание ценности и деятельности по управлению данными

Роман Лукьяненко [1]

[1] *Школа коммерции Макинтайра, Университет Вирджинии, Шарлоттсвилль, Вирджиния 22903, США, romanl@virginia.edu*

Абстрактный
В эпоху, когда доминируют информационные технологии, критическая дисциплина управления данными остается недооцененной по сравнению с инновациями, которые она обеспечивает, такими как искусственный интеллект и социальные сети. Неопределенность, окружающая то, что составляет управление данными и связанные с ним действия, усложняет усилия по объяснению его важности и обеспечению сбора, хранения и использования данных таким образом, чтобы максимизировать ценность и избегать сбоев. Целью данной статьи является устранение этих недостатков путем представления простой структуры для понимания управления данными, называемой MAGIC (Modeling, Acquisition, Governance, Infrastructuring, Consumption support). MAGIC охватывает пять ключевых видов деятельности: моделирование, приобретение, управление, инфраструктуру и задачи поддержки потребления. Разграничивая эти компоненты, структура MAGIC обеспечивает четкий и доступный подход к управлению данными, который можно использовать для обучения, исследований и практики.

Ключевые слова
Управление данными, моделирование, сбор, управление, инфраструктура, поддержка потребления, MAGIC, Modeling, Acquisition, Governance, Infrastructuring, Consumption support

Английский оригинал: Lukyanenko, R. (2024). The MAGIC of Data Management: Understanding the Value and Activities of Data Management. arXiv:2408.07607. https://arxiv.org/abs/2408.07607

1. Жизнь в эпоху магии

В научно-фантастической книге 1962 года «Профили будущего: исследование пределов возможного» Артур Кларк сформулировал свои знаменитые Три закона, из которых третий закон является самым известным: «Любая достаточно развитая технология неотличима от магии». Современный мир становится все более магическим, движимым неустанным прогрессом информационных технологий. Тем не менее, основы этого магического мира остаются плохо понятыми, а иногда даже игнорируются.

Современный мир является цифровым. Практически каждый аспект человеческого существования становится цифровым или зависит в какой-то мере от информационных технологий и информационных систем, построенных на их основе. Только за последние три десятилетия взрывное развитие информационных технологий привело к революционным изменениям в образе жизни людей. Рассмотрим несколько примеров. Интернет, который стал популярным в 1990-х годах, сделал возможной электронную коммерцию и распределенный обмен информацией. Используя Интернет, социальные сети появились в 2000-х годах, кардинально изменив способ, которым люди общаются, получают и распространяют информацию.

Совсем недавно искусственный интеллект (ИИ), приведший к таким чудесам, как ChatGPT и беспилотные автомобили, был назван «вершиной [человеческой] изобретательности» [17] . По оценкам, ИИ внесет 15 триллионов долларов в мировой ВВП к 2030 году [43] и может обеспечить мировое господство стране-лидеру в области ИИ [18] .

С резким расширением ИТ беспрецедентные требования предъявляются к вычислительным ресурсам, хранению и пропускной способности. Ответом на этот вызов являются квантовые вычисления, которые используют свойства мельчайших элементов материи – фотонов, электронов, ионов – для обработки информации и связи. Силы квантовых частиц настолько причудливы и невероятны, что их называют «Эффектом Бога» [10] . Некоторые предполагают, что квантовые вычисления «могут стать революцией для человечества, большей, чем огонь, большей, чем колесо» [31] .

В то время как новые информационные системы постоянно появляются, одно остается неизменным. Следовательно, это больше, чем совокупность искусственного интеллекта, квантовых вычислений, социальных сетей, онлайн-банкинга и Интернета. Это *цифровая информация* как таковая. Суть информационных технологий и систем, построенных с ее помощью, — это *информация* или *данные* (используемые здесь как синонимы). Без цифровой информации невозможны ни беспилотные автомобили, ни YouTube, а квантовые компьютеры — это просто куча дорогого хлама.

Чтобы цифровые данные были полезными и даже пригодными к использованию, ими необходимо *управлять* систематическим образом. В результате управление данными — подходы и методы обработки цифровой информации — является основой современных информационных технологий.

Эффективное управление данными необходимо для принятия обоснованных решений в организациях и отдельными лицами. Оно может повысить производительность, повысить конкурентоспособность, сократить расходы и отходы. И наоборот, плохое управление данными может привести к значительным сбоям и сбоям в работе. Например, утечка данных, с которой столкнулась Equifax в 2017 году, была вызвана неадекватными мерами безопасности данных, которые позволили хакерам получить доступ к конфиденциальной информации приблизительно 147 миллионов человек. Это нарушение не только нанесло ущерб репутации Equifax, но и привело к существенным финансовым штрафам и потере доверия потребителей.

Проблемы управления данными беспокоят даже лидеров информационных технологий. Таким образом, после многомиллионных инвестиций и большой шумихи онкологический центр им. М. Д. Андресона отказался от сотрудничества с системой искусственного интеллекта Watson, предоставленной IBM [46] . И это несмотря на амбициозную цель IBM использовать искусственный интеллект для лечения рака. На крах проекта MD Anderson Watson существенно повлияли проблемы качества данных и интеграции данных. Проект боролся с несоответствиями и ошибками в данных, что снижало эффективность системы Watson в предоставлении точных и безопасных рекомендаций по лечению рака [52] .

Удивительно, но в мире, где доминируют информационные технологии, ключевое знание, связанное с информационными технологиями, наука управления данными, продолжает отходить на второй план по сравнению с изобретениями, которые возможны с их помощью. Данные и, как следствие, управление данными продолжают рассматриваться как нечто второстепенное. Как показало одно исследование современных методов искусственного интеллекта: «Все хотят [создавать новый искусственный интеллект], а не [выполнять] работу с данными» [48] . Однако ИИ ничего не значит без данных. ИИ — это всего лишь творческое использование математики и статистики для извлечения дополнительных шаблонов из данных. Отсутствие цифровых данных для обучения ИИ означает отсутствие новых информационных шаблонов, отсутствие беспилотных автомобилей или роботов-доставщиков.

Относительно недостает признания управления данными по сравнению с технологиями, которые стали возможны с его помощью, такими как ИИ или квантовые вычисления. Например, с появлением инструментов ИИ, таких как ChatGPT, университеты по всему миру бросились разрабатывать новые курсы по ИИ. Напротив, в то же время было создано меньше новых курсов по управлению данными. Однако управление данными для ИИ — это не то же самое, что, например, для социальных сетей или реляционных баз данных. Хуже того, в некоторых случаях при безупречном управлении данными использование искусственного

интеллекта может быть совершенно ненужным. Можно избежать огромных инвестиций в инструменты, энергию и человеческий капитал, если данные будут храниться таким образом, чтобы напрямую позволять реализовывать цели, поставленные для искусственного интеллекта.

Существует множество причин, по которым управление данными отходит на второй план, несмотря на то, что оно является обязательным для любого типа приложения информационных технологий. Одной из них является заблуждение, что подходы и методы управления данными устоялись и могут восприниматься как должное. Простой запрос на сбои, связанные с неправильной обработкой данных, может быстро развеять этот миф. Более того, методы, приемы и инструменты управления данными постоянно развиваются. Сегодня управление данными больше не является вопросом настройки реляционной базы данных в соответствии с известными методами моделирования данных. Многие организации сталкиваются с проблемой укрощения больших объемов разнородных данных, экспериментируя с такими новыми технологиями, как озера данных и, в последнее время, хранилища данных . Однако, если все сделано неправильно, эти усилия могут дать незначительную отдачу и даже оказаться контрпродуктивными, поскольку дорогостоящая новая инфраструктура хранения может постепенно превратиться в беспорядочный, неорганизованный хаос.

Усложняет усилия по уделению должного внимания управлению данными отсутствие консенсуса относительно того, что составляет управление данными, какие действия в него вовлечены, когда начинается и заканчивается управление данными (что приводит к другим действиям). Из этого следует отсутствие доступной, простой для понимания концепции того, что такое управление данными. Эффективное управление данными требует целостного подхода, а это требует систематической и всеобъемлющей структуры.

Наша цель — предоставить простой подход к пониманию управления данными. Мы называем это MAGIC (Modeling, Acquisition, Governance, Infrastructuring, Consumption support). Он включает моделирование, приобретение, управление, инфраструктуру (новый термин, который более точно отражает связанные действия, часто называемые «хранением» или «курированием»), и действия по поддержке потребления. MAGIC охватывает ключевые действия по управлению данными, применимые к управлению данными любого типа. Если все сделано правильно, результат управления данными может быть поистине волшебным.

Ниже мы объясним MAGIC, но сначала дадим общую информацию об управлении данными.

2. Информационные системы и информационные технологии

Чтобы понять природу управления данными, полезно иметь общее представление об информационных системах и информационных технологиях, лежащих в основе необходимости управления данными.

Информационные технологии — это знания о создании и использовании информационных систем. Создание информационных систем включает в себя ряд действий от системного анализа и проектирования до внедрения и обслуживания. Этот процесс обычно включает в себя определение требований пользователей, создание системных спецификаций, разработку программных приложений и интеграцию аппаратных компонентов. С другой стороны, использование информационных систем фокусируется на применении этих систем для достижения организационных и личных целей, таких как повышение производительности, улучшение обслуживания клиентов и получение конкурентных преимуществ.

Информационные системы состоят из взаимодействующих компонентов, которые собирают, хранят, манипулируют, извлекают, обмениваются и используют данные или информацию. [2]Компоненты информационной системы включают обязательные элементы, такие как компьютерное оборудование и программное обеспечение, а также необязательные элементы, такие как пользователи и другие информационные системы и устройства.

[2]Информационная система — это тип системы. Особенностью этого является тот факт, что системы обладают эмерджентными свойствами — свойствами, которые возникают из взаимодействия их базовых компонентов [36] . В случае информационных систем эти эмерджентные свойства (например, производительность, способность извлекать информацию из данных) могут быть значительно более ценными, чем компоненты, из которых состоят эти системы.

Примерами информационных систем являются Youtube.com, Microsoft Word, ChatGPT или система приема в университет.

Компьютерное оборудование образует физическую основу информационной системы, включающую такие устройства, как серверы, настольные компьютеры, ноутбуки и сетевое оборудование, которые выполняют основные вычислительные задачи. Компьютерное программное обеспечение, еще один важный компонент, включает операционные системы, приложения и системы управления данными, которые позволяют оборудованию эффективно обрабатывать и управлять данными. Пользователи, хотя и необязательны, взаимодействуют с системой для выполнения различных задач, ввода данных и создания и использования выходных данных. Кроме того, другие информационные системы и устройства могут быть интегрированы для расширения возможностей системы, обеспечивая более полный анализ данных, улучшенную связь и оптимизированные процессы.

Информационные системы могут интегрироваться с другими объектами для формирования систем более высокого порядка. Например, система управления авиакомпанией — это информационная система, которая состоит из ряда других информационных систем, таких как базы данных и специализированное программное обеспечение, но также включает в себя диспетчеров воздушного движения. Неотъемлемой частью системы управления авиакомпанией является программная система, называемая планированием ресурсов предприятия (ERP). EPR интегрирует различные бизнес-процессы, собирая и организуя данные из разных отделов, что позволяет улучшить координацию и принятие решений на основе данных.

Независимо от конфигурации, любая информационная система обрабатывает данные. Данные являются входными данными в систему, они преобразуются каким-либо образом и обычно предоставляются в качестве выходных данных для потребления человеком или компьютером. Данные являются представлениями любого объекта или события на некотором физическом носителе. Основная ценность данных заключается в их способности передавать что-либо об объекте, который они представляют, более эффективно, безопасно или даже делая возможным вообще что-либо узнать о представленном объекте. Принимая во внимание последнее, только через данные (такие как слова, нарисованные изображения) мы можем узнать содержимое чьего-либо разума. Часто информация отличается от данных как данные, которые помещены в значимый контекст таким образом, что то, что представляют данные, более или менее ясно. Мы предпочитаем не различать данные и информацию, потому что нам не удается найти случаи, когда данные лишены смысла, следуя некоторым ученым и практикам, которые рассматривают данные и информацию как синонимичные [2, 37, 54, 60, 62] . Мы также используем данные как существительное единственного числа. Данные являются формой множественного числа от латинского *datum* , но форма datum используется редко. В большинстве случаев использование единственного числа в предложении звучит более естественно.

Манипулируя данными как прокси для представляемых ими объектов, мы способны понимать и воздействовать на мир таким образом, который было бы невозможно или трудно сделать без данных. Сама ценность информационных технологий заключается в способности информационных систем эффективно обрабатывать данные в масштабе, таким образом, который был бы затруднителен для людей. Калькулятор — простой пример такой эффективности, тогда как беспилотный автомобиль — более продвинутый пример.

Чтобы понять, как данные обрабатываются информационной системой, рассмотрим пример системы онлайн-торговли. Когда клиент размещает заказ на веб-сайте электронной коммерции (информационной системе), сервер, на котором работает веб-сайт, получает различные фрагменты данных в качестве входных данных. Это может быть имя клиента, адрес, платежная информация и сведения о приобретаемых товарах. Эти данные обычно вводятся в систему через формы ввода пользователя. Затем код на сервере обрабатывает эти данные несколькими способами. Он проверяет платежную информацию, проверяет наличие товаров в инвентаре, рассчитывает общую стоимость, включая налоги и доставку, и обновляет базу данных инвентаря, чтобы отразить сокращенные запасы. Кроме того, система может анализировать данные о покупках, чтобы предоставлять рекомендации по аналогичным продуктам или обновлять профили клиентов с историей покупок. Затем преобразованные данные используются для генерации различных выходных данных. Для клиента система предоставляет

подтверждающее сообщение и квитанцию по электронной почте, содержащую сведения о покупке. Для склада система генерирует список выбора и упаковочный лист, чтобы облегчить выполнение заказа. Кроме того, в управленческих целях система может предоставлять отчеты о тенденциях продаж, уровнях запасов и покупательском поведении клиентов, которые можно использовать для дальнейшего анализа и принятия решений.

Как показывает пример системы интернет-торговли , данные являются ключевым компонентом информационной системы, и данные не материализуются из воздуха. Данные необходимо сначала собрать, а затем сохранить и использовать. Как конкретная физическая субстанция, современная цифровая информация представляет собой набор импульсов света и электрических зарядов. Эти субстанции находятся на аппаратных устройствах, таких как смартфоны или ноутбуки. При определенной обработке эти импульсы или заряды могут быть представлены в виде двоичных цифр нуля и единицы (поэтому мы называем такие данные *цифровыми данными*). При дальнейшей организации предопределенными способами эти импульсы и заряды могут образовывать более сложные узоры, которые могут интерпретироваться людьми и компьютерами как текст, изображения, видео, а также как инструкции по программированию для дальнейшей манипуляции этими узорами. Искусственный интеллект является особенно мощным методом извлечения полезных информационных узоров, которые затем могут быть преобразованы в медицинские диагнозы или инструкции для беспилотных автомобилей.

Это очень простое изложение природы цифровой информации достаточно, чтобы понять, что не так уж и просто перейти от световых импульсов к твитам об Олимпиаде, президентских выборах или прощальном туре Pink Floyd. Управление данными связано с тем, чтобы данные или информация как вход, внутренний компонент и выход информационных систем позволяли информационным системам и людям, которые их используют, реализовывать свои цели. Следовательно, управление данными гарантирует, что информационные системы будут выполнять свою магию.

3. МАГИЯ управления данными

Фреймворк MAGIC извлекает суть и в то же время упрощает понимание природы управления данными. Это фреймворк, который вырос из проекта кураторства публикаций по управлению данными в ведущем научном журнале *MIS Quarterly* [9] .

Проект кураторства управления данными MIS Quarterly концептуализировал управление данными как 5C управления данными. Он определил область управления данными как область изучения, которая исследует и разрабатывает действия и методы концептуализации, сбора, курирования, потребления и контроля данных для поддержки понимания, анализа и действий. Это определение и концептуализации, наряду с другой соответствующей литературой, обеспечили основу для MAGIC, поскольку он далее разрабатывал и совершенствовал эти концепции. Подробности проекта, наряду с другими основополагающими определениями управления данными, приведены в Приложении A.

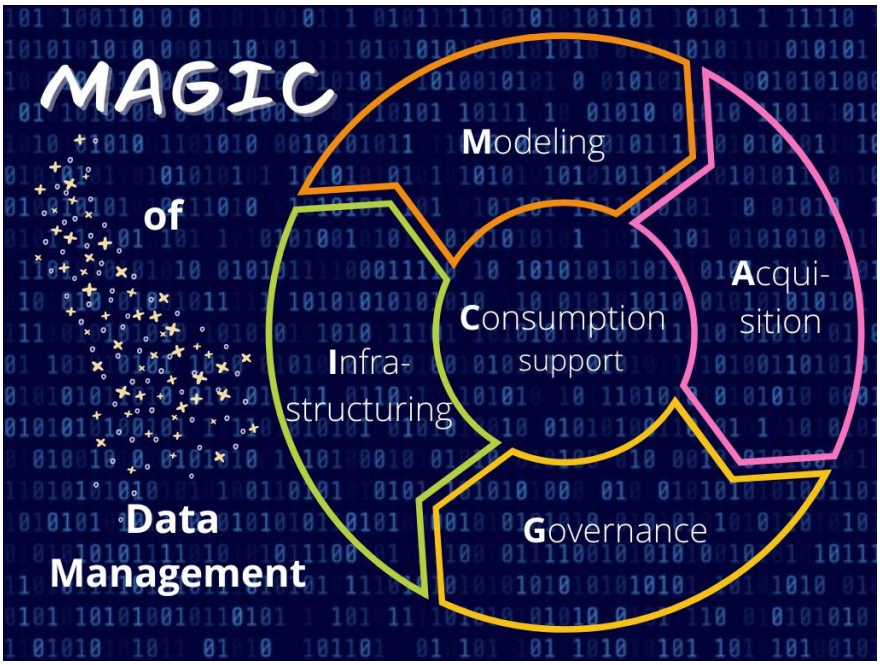

Рисунок 1. Графическая иллюстрация фреймворка MAGIC

Структура MAGIC охватывает ключевые действия по управлению данными, включая моделирование, получение, управление, инфраструктуру, поддержку потребления . Рисунок 1 графически иллюстрирует структуру MAGIC.

Действия по управлению данными могут выполняться в той же последовательности, что и буквы MAGIC, или к ним можно подходить в другом порядке, хотя моделирование неизменно является действием, которое предшествует любому другому, независимо от того, осознается ли оно формально или выполняется на подсознательном уровне. Кроме того, хотя эти действия подчеркивают отдельные проблемы моделирования и группируют связанные практики, они также частично пересекаются и подпитывают друг друга. Например, формальные методы моделирования могут использоваться для разработки протоколов управления данными, таких как использование модели компьютерной безопасности Белла-ЛаПадулы [56] . В то же время управление данными может предписывать, как формальное моделирование данных проводится в организации. Следовательно, действия MAGIC должны подходить целостно, находиться в гармонии друг с другом, и все они должны быть рассмотрены для обеспечения успешных результатов.

3.1. Моделирование

Представьте себе беспилотный автомобиль. Это сложная информационная система, которая включает в себя само транспортное средство, а также программное обеспечение на базе ИИ, которое позволяет автомобилю двигаться автономно. Чтобы программное обеспечение ИИ работало, его нужно было «обучить» на миллионах примеров человеческого вождения, которые использовались в качестве исходных входных данных. Результатом этого обучения является создание модели ИИ, которая представляет собой цифровые данные, представляющие собой сложные правила вождения. Затем, когда автомобиль начинает движение по дороге, он использует свои датчики (например, камеры, акселератор, лазеры), принимая данные, сгенерированные датчиками, в качестве входных данных, чтобы обработать эти данные в инструкции для оборудования автомобиля (ускориться, поддерживать определенную скорость, повернуть в заданный момент). Эта обработка данных управляется моделью ИИ, созданной на основе миллионов примеров предыдущего вождения.

Задолго до того, как такой автомобиль станет реальностью, он существует как идея, как воображение. Он существует как *ментальный модель* . Поскольку мы имеем дело не с обычным автомобилем, а с информационной системой (т. е. автомобилем, наполненным

технологическими компонентами, обрабатывающими данные), мы должны включить данные в качестве неотъемлемого компонента модели беспилотного автомобиля. Данные так же неизбежны, как колеса и топливо в модели беспилотного автомобиля (см. Рисунок 2 для визуализации ментальной модели).

Рисунок 2. Визуализированная ментальная модель беспилотного автомобиля.

Моделирование подразумевает создание представления объекта или события для целей понимания, коммуникации, решения проблем и проектирования. Представление — это преобразование некоторого исходного объекта во что-то другое. Например, правило «остановиться» можно представить в виде красного восьмиугольника на столбе. Представления обычно происходят в средах, отличных от их референтов (то есть язык может представлять физические объекты, физические объекты могут представлять абстрактные ментальные концепции). Этот сдвиг среды делает эти представления полезными (когда эффективнее работать в среде, отличной от исходной), но также создает проблемы, поскольку идеальный перевод с одной среды на другую невозможен. Что-то теряется, что-то еще может быть добавлено.[3]

Моделирование данных, как особый тип моделирования, представляет данные или любые связанные с ними объекты или события в целях понимания, коммуникации, решения проблем и проектирования.

Деятельность по управлению данными обычно начинается с моделирования и никогда не может избежать его. Моделирование данных может быть явным (когда создаются диаграммы, следующие строгим правилам), или неявным (когда проектировщики думают о том, как должны работать системы, а затем следуют своим идеям при создании программного кода и пользовательского интерфейса). Сам процесс размышления о беспилотном автомобиле включает в себя ментальную модель, которая, если она достаточно подробная и точная, должна учитывать незаменимую роль данных в превращении такого автомобиля в реальность.

Существует множество проблем управления, связанных с моделированием данных. Эти проблемы обычно направлены на то, чтобы гарантировать, что модели предоставляют адекватные представления, чтобы системы могли надлежащим образом собирать, хранить и использовать данные. Эти проблемы предполагают, что:

- Модели являются точными, полными и актуальными представлениями, поэтому на их основе могут быть построены соответствующие компоненты системы [4, 41, 59] ;
- Модели и моделирование полезны и эффективны, обеспечивая высокую отдачу от усилий по моделированию [21, 44, 57] ;
- Модели доступны, инклюзивны и просты в использовании, поэтому все заинтересованные стороны могут участвовать и формировать моделирование и последующие действия [6, 24, 34, 35] .

[3] Вспомним, что данные — это также представление. Фактически, любая физическая модель — это также данные. Пока мы можем получить доступ к ментальным моделям в мозге человека, ментальные модели также могут быть данными. Действительно, модели как данные часто становятся входными данными в информационных системах, например, в инженерии на основе моделей или автоматизации роботизированных процессов.

Для решения этих проблем были разработаны и постоянно оцениваются формальные подходы, включающие явное моделирование [5] . Некоторые примеры популярных методов моделирования включают диаграммы сущностей-связей (ERD), унифицированный язык моделирования (UML) и модель и нотацию бизнес-процессов (BPMN). Кроме того, постоянно разрабатываются новые методы и языки моделирования. Важно следить за этими разработками, поскольку они могут обеспечить лучшие решения проблем моделирования.

При построении сложных или высокорисковых систем явные модели данных практически неизбежны. Эти модели предоставляют обзор существующих систем, фиксируют требования к новой системе, описывают тип данных, которые будут входными данными, объясняют преобразования данных и указывают типы выходных данных, которые система будет генерировать. Затем эти модели можно показать другим для проверки, коммуникации , в качестве дорожной карты для создания программного кода и пользовательского интерфейса, а также они документируют функции системы формальным образом.

Особенно часто проводится явное моделирование при построении баз данных. Базы данных хранят данные систематическим образом, что влияет на степень доступности и безопасности данных. Для обеспечения этих свойств важно следовать проверенным методам моделирования баз данных, которые существуют с 1970-х годов [13, 29, 30, 51] .

Диаграммы «сущность-связь» (ERD) и унифицированный язык моделирования (UML) являются мощными инструментами, используемыми при проектировании и моделировании информационных систем, хотя они служат немного разным целям и используются в разных контекстах. ERD в основном используются для моделирования структуры данных системы путем представления сущностей (например, таблиц в базе данных), их атрибутов и отношений между ними [8] . Например, ERD для системы управления клиентами может включать такие сущности, как «Клиент», «Заказ» и «Продукт», с отношениями, показывающими, как клиенты размещают заказы, а заказы включают продукты. ERD помогают визуализировать схему базы данных и имеют решающее значение при проектировании базы данных, гарантируя, что все необходимые отношения данных учитываются перед реализацией.

С другой стороны, UML является более всеобъемлющим языком моделирования, который охватывает различные типы диаграмм для представления различных аспектов программных систем, включая структуру, поведение и взаимодействия [27] . UML включает диаграммы классов, которые похожи на ERD и используются для моделирования статической структуры системы путем отображения классов, их атрибутов, методов и отношений. Кроме того, UML включает диаграммы вариантов использования для фиксации функциональных требований, диаграммы последовательности для моделирования взаимодействий с течением времени и диаграммы состояний для представления состояний и переходов объектов в системе. Эта универсальность делает UML стандартным инструментом для разработки программного обеспечения, предоставляя унифицированный способ моделирования сложных систем с разных точек зрения, облегчая коммуникацию между заинтересованными сторонами и поддерживая жизненный цикл разработки от анализа и проектирования до внедрения и обслуживания.

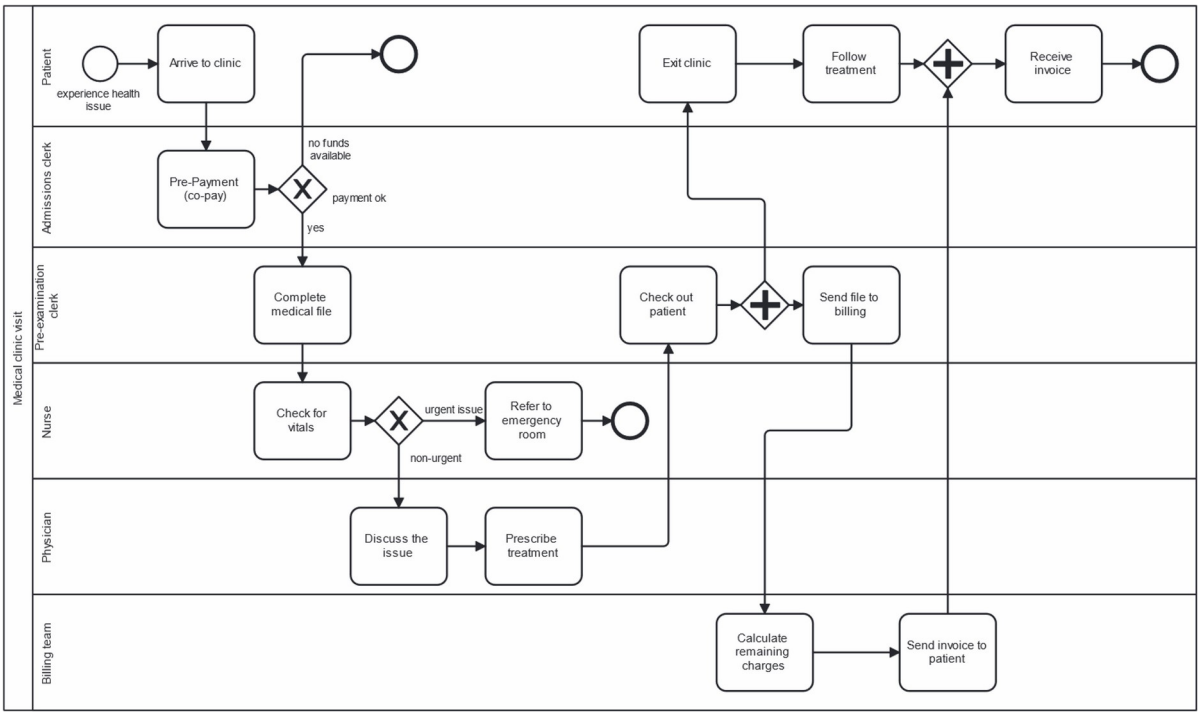

Рисунок 3. Диаграмма BPMN посещения частной клиники с точки зрения пациента. Простой процесс выявляет удивительные повороты, которые могут быть не очевидны без конкретной модели.

Модель и нотация бизнес-процессов (BPMN) — популярный подход к моделированию для представления бизнес-процессов, которые поддерживают или реализуют информационные системы [12] . Он предоставляет набор символов, которые позволяют заинтересованным сторонам, включая бизнес-аналитиков и технических разработчиков, наглядно и эффективно визуализировать и сообщать о процессах. Диаграммы BPMN обычно включают такие элементы, как события, действия, шлюзы и потоки, а также объекты данных, такие как базы данных. Эти элементы позволяют организациям явно документировать, анализировать и оптимизировать свои бизнес-процессы. Используя BPMN, компании могут улучшить сотрудничество, улучшить понимание процессов и повысить эффективность операций. На рисунке 3 показан пример диаграммы BPMN, представляющей типичный процесс посещения частной клиники .

Моделирование предшествует любой другой деятельности по управлению данными, поскольку перед тем, как приступить к любой другой деятельности по управлению данными, ее необходимо сначала обдумать, то есть смоделировать. Этот акт мышления может создать явную физическую модель (например, диаграмму BPMN, описывающую элементы управления данными) или остаться неявным, создавая ментальные модели (например, ментальный план для типов данных, которые система будет собирать). МАГИЯ управления данными начинается с моделирования.

3.2. Приобретение

Фаза моделирования обычно завершается внедрением моделей данных в конкретные компоненты информационных систем. Это очень похоже на строительство настоящего дома после создания и проверки строительных чертежей. Это порождает другие виды деятельности по управлению данными: приобретение, управление и инфраструктура.

Начнем с приобретения данных, как точки, когда данные рождаются или получают новую жизнь, но важно подчеркнуть, что само приобретение должно подлежать соответствующему контролю. Кроме того, для успешного приобретения данных требуется инфраструктура для их

хранения. Однако приобретение — это отдельная деятельность, которая, хотя и осуществляется с учетом других видов деятельности, также формирует их.

Приобретение решает вопросы, связанные со сбором и получением данных в качестве входных данных для информационных систем. Это включает рассмотрение данных, необходимых для достижения желаемых целей информационной системы, как и где их получить, и как организовать процесс сбора данных.

Осуществимость сбора данных является важным соображением. Вполне возможно, что существует несоответствие между тем, что желательно, и тем, что возможно с цифровыми данными, даже в цифровом мире. Например, может быть желательно получить секретные документы о событии, но это невозможно сделать. Поэтому проект по управлению данными может быть прекращен на этапе получения, если становится невозможным получить желаемые данные экономически, своевременно, безопасно, этично или юридически.

Если предположить, что сбор данных осуществим, то следующим вопросом является то, следует ли приобретать новые или существующие данные. Это основано на анализе затрат и выгод сбора данных с нуля или получения существующих вариантов данных или их комбинации. Информационная система может собирать новые данные, запрашивая данные у своих пользователей или из окружающей среды. Например, беспилотный автомобиль может получать свои навигационные инструкции со смартфона пользователя, который хочет добраться до определенного пункта назначения.

Данные также могут быть получены из другой информационной системы, которая создана для взаимодействия и обмена уже существующими данными. Следовательно, банки, предоставляющие кредиты, могут получать кредитные отчеты клиентов из кредитных бюро, таких как Equifax или TransUnion. Больницы и клиники часто взаимодействуют со страховыми компаниями для проверки покрытия пациента и получения уже существующих медицинских записей. Например, когда пациент посещает нового поставщика медицинских услуг, поставщик может получить историю болезни пациента, результаты лабораторных анализов и предыдущие методы лечения из других больниц или клиник, используя электронные системы медицинских записей. Платформы электронной коммерции часто взаимодействуют с платежными шлюзами и службами доставки для оптимизации операций. Когда клиент совершает покупку, сайт электронной коммерции может использовать платежные процессоры, такие как PayPal (paypal.com) или AliPay (alipay.com), чтобы проверить и обработать транзакцию. Туристические сайты, такие как Expedia.com или Kayak.com, взаимодействуют с системами бронирования авиабилетов (такими как Amadeus или Sabre), чтобы предоставлять клиентам варианты рейсов, цены и наличие мест. Гостиничные сети используют централизованные системы бронирования, которые взаимодействуют с онлайн-турагентствами, такими как Expedia.com, Booking.com или Airbnb.com, для обновления наличия номеров и управления бронированиями. В цифровом мире существует высокая вероятность того, что данные не будут собираться с нуля, а могут быть получены из существующих источников.

После определения источника данных, еще одной проблемой управления данными является обеспечение того, чтобы правильный объем данных достаточного качества был предоставлен в качестве входных данных для системы, чтобы она могла выполнять соответствующие преобразования этих данных для реализации поставленных целей для систем. Поэтому качество данных является первостепенной проблемой сбора данных.

Качество данных относится к таким параметрам данных, как точность, полнота, надежность, релевантность и своевременность [61] . Низкое качество данных может привести к неточному анализу, ошибочным стратегиям, испорченной репутации и финансовым потерям.

Одним из ключевых аспектов качества данных является точность, что означает, что данные должны представлять реальные сущности или события, которые они должны отображать. Неточные данные могут возникать из-за человеческой ошибки, сбоев системы или устаревшей информации. Полнота является еще одним критическим фактором; она гарантирует наличие всех необходимых точек данных и отсутствие пропущенных значений, которые могли бы повлиять на анализ. Кроме того, релевантность гарантирует, что собранные данные соответствуют конкретным целям и процессам принятия решений.

Соображения относительно качества данных напрямую формируют процессы получения данных и инфраструктуру получения данных. Некоторые указания относительно того, как формировать эти процессы, должны исходить из предшествующей деятельности по моделированию. В частности, например, диаграммы UML могут определять поля данных как атрибуты, а также могут включать типы данных и желаемые преобразования. В то же время, это руководство является неполным, поскольку необходимо сделать много дополнительных проектных решений, чтобы перевести это высокоуровневое руководство в конкретные функции, такие как теги HTML, элементы сценариев CSS, алгоритмы JavaScript и т. д. [38] .

Плохо спроектированные пользовательские интерфейсы могут существенно способствовать низкому качеству данных, внося опечатки и ошибки из-за неясных инструкций и меток. Когда поля ввода неоднозначно помечены или не имеют надлежащих указаний, пользователи могут неправильно истолковать, какие данные требуются. Например, поле, просто помеченное как «Имя» без указания «Имени» или «Фамилии», может привести к путанице и неправильному вводу данных. Аналогичным образом, недостаточные инструкции об ожидаемом формате (например, формат даты, формат номера телефона) могут привести к неправильному вводу данных пользователями, что приведет к ошибкам, которых можно было бы легко избежать с помощью лучшего дизайна пользовательского интерфейса.

Кроме того, макет и дизайн интерфейса могут привести к ошибкам или помешать сбору полных данных. Загроможденный интерфейс со слишком большим количеством элементов может перегрузить пользователей, затрудняя им поиск правильных полей ввода или кнопок, тем самым увеличивая вероятность ошибок. Непоследовательное размещение элементов, таких как кнопки и поля, также может сбивать пользователей с толку, заставляя их нажимать не ту кнопку или вводить данные не в то поле. На мобильных устройствах маленькие сенсорные цели и сверхчувствительные поля могут затруднить для пользователей точное нажатие, что приводит к опечаткам.

Рекомендуется следовать лучшим практикам в области дизайна пользовательского интерфейса и взаимодействия с пользователем [например, 40, 50] не только для того, чтобы гарантировать, что информационные системы просты и приятны в использовании, но и для минимизации проблем с качеством данных, возникающих из-за проблем с интерфейсом и дизайном процесса. Существует множество неочевидных решений, которые могут быть эффективными в предотвращении определенных типов проблем с качеством данных. Одним из них является геймифицированный дизайн, то есть проектирование информационных систем с соблюдением принципов игрового дизайна, даже если эти системы не являются играми. Такие проекты могут минимизировать проблемы с качеством данных за счет внедрения интуитивно понятных элементов управления (например, красочных и очевидных вариантов навигации) и четких механизмов обратной связи (например, звучание фанфар при успешном выполнении действия).

Хотя качество данных играет важную роль во время получения, проблемы качества данных пронизывают весь цикл управления данными и должны учитываться при моделировании, управлении, инфраструктурировании и подготовке данных для потребления. Распространенной, но часто упускаемой из виду причиной низкого качества данных является недостаточное или предвзятое моделирование данных, которое может игнорировать определенные требования пользователей или лишать права голоса определенные категории пользователей, что затрудняет для них предоставление своего пользовательского ввода добросовестным образом [33] . Обеспечение качества данных включает внедрение процессов и стандартов, которые регулируют сбор данных, а также хранение и использование, что мы рассмотрим позже, как часть других видов деятельности MAGIC.

3.3. Управление

Чтобы гарантировать, что данные надлежащим образом получены, хранятся и используются эффективно и ответственно, необходимо внедрить соответствующие элементы управления на разных этапах жизненного цикла данных, что приводит к управлению данными — основной деятельности по управлению данными. Управление данными, как правило, является

организационной деятельностью, но его принципы также могут применяться к проектам по управлению персональными данными, поскольку они касаются законности, этичности и прозрачности использования данных.

Управление данными включает в себя создание и внедрение политик, стандартов и процедур, которые обеспечивают доступность данных, удобство использования, целостность и безопасность. Внедряя надежное управление данными, организации и отдельные лица могут повысить качество данных, соблюдать правила и принимать более обоснованные решения на основе надежных данных. В организациях эффективный контроль данных требует сотрудничества между различными отделами, включая ИТ, соответствие и бизнес-подразделения, для согласования методов управления данными с целями организации.

Одним из важнейших аспектов управления данными является определение права собственности на данные и подотчетности. Назначение управляющих данными или владельцев гарантирует, что есть назначенные лица, ответственные за определенные наборы данных, что помогает поддерживать качество и согласованность данных. Например, в организации здравоохранения управляющий данными может отвечать за записи пациентов, обеспечивая их точность, актуальность и доступность при соблюдении правил конфиденциальности, таких как HIPAA. [4]Это четкое разграничение ролей помогает упростить процессы управления данными и способствует формированию культуры подотчетности в организации.

Другим важным компонентом управления данными является установление политик и стандартов данных. Эти политики определяют, как данные должны собираться, храниться и использоваться, гарантируя, что все заинтересованные стороны будут придерживаться передовых практик. Например, финансовое учреждение может иметь политики для управления шифрованием данных, контролем доступа и сроками хранения данных для соответствия таким нормам, как GDPR [5]или PCI DSS. [6]Определяя четкие стандарты, организации могут снизить риск утечки данных и повысить общую безопасность своих информационных активов.

Безопасность инфраструктуры данных является критически важным аспектом управления данными и включает в себя множество проблем. Все конфиденциальные данные должны быть зашифрованы. Шифрование данных — это метод изменения формы данных таким образом, чтобы к ним могли получить доступ только уполномоченные стороны с правильной процедурой расшифровки. Новая проблема — защита данных от будущего использования квантовых компьютеров. Известная как постквантовая криптография, это область исследований и практики, которая разрабатывает криптографические методы, которые будут защищены от потенциальных угроз, создаваемых квантовыми компьютерами. Квантовые компьютеры могут решать определенные математические задачи, которые классические компьютеры, возможно, никогда не смогут решить. Сюда входят такие задачи, как факторизация больших целых чисел, основа для шифрования RSA или вычисление дискретных логарифмов, которое используется в криптографии эллиптических кривых. Постквантовая криптография направлена на разработку новых криптографических алгоритмов, которые остаются безопасными даже при наличии возможностей квантовых вычислений. Эти алгоритмы (например, криптография на основе решетки, криптография на основе хэша, криптография на основе кода) основаны на математических задачах, которые, как считается, находятся за пределами возможностей квантовых компьютеров. Заблаговременно подготовившись и перейдя на эти квантово-устойчивые алгоритмы, организации и отдельные лица смогут защитить свои данные и коммуникации от будущих квантовых угроз, обеспечив долгосрочную безопасность в постквантовом мире.

Безопасность также включает определение доступа и разрешений на доступ к данным. Внедрение надежных средств контроля доступа, таких как многофакторная аутентификация и доступ на основе ролей, помогает ограничить круг лиц, которые могут просматривать или изменять данные. Регулярные аудиты безопасности и оценки уязвимостей выявляют и устраняют потенциальные слабые места в системе хранения. Кроме того, избыточность данных

[4]Закон США о переносимости и подотчетности медицинского страхования 1996 года (HIPAA) — это федеральный закон, который требует защиты конфиденциальной медицинской информации от раскрытия без согласия или ведома пациента.

[5]Общий регламент Европейского союза о защите данных (GDPR) регулирует порядок обработки и передачи персональных данных физических лиц в ЕС.

[6]Стандарт безопасности данных в индустрии платежных карт (PCI DSS) представляет собой набор рекомендаций, призванных помочь организациям, обрабатывающим информацию о кредитных картах, обеспечить ее безопасность и сохранность.

и регулярное резервное копирование имеют решающее значение для восстановления в случае потери или повреждения данных. Использование передовых средств обнаружения и реагирования на угрозы также может помочь отслеживать и устранять подозрительные действия в режиме реального времени.

Безопасность также подразумевает защиту помещений, сетей и местоположений центров обработки данных от несанкционированного доступа и обеспечение их устойчивости к психологическим атакам, таким как поджог или кража. Нередко эти проблемы игнорируются, особенно учитывая, что команды, работающие над управлением данными, иногда не обладают экспертными знаниями в вопросах «физической» безопасности. Целостное управление данными подразумевает рассмотрение «физических» и «цифровых» аспектов безопасности и обеспечение их тщательной организации и интеграции.

Управление качеством данных также является ключевым направлением в управлении данными. Организациям необходимо внедрять процессы мониторинга, оценки и улучшения качества данных с течением времени. Это может включать методы профилирования, очистки и проверки данных для выявления и исправления ошибок в наборах данных. Например, розничная компания может регулярно оценивать данные своих клиентов, чтобы гарантировать точность и актуальность контактной информации. Отдавая приоритет качеству данных, организации могут повысить надежность своей аналитики и отчетности, что приведет к более эффективному принятию решений и стратегическому планированию.

Соответствие требованиям и управление рисками также являются неотъемлемыми частями управления данными. Организации и отдельные лица должны гарантировать, что их методы работы с данными соответствуют соответствующим правилам и отраслевым стандартам. Это может включать проведение регулярных аудитов и оценок для выявления потенциальных рисков соответствия и реализацию корректирующих действий по мере необходимости. Например, фармацевтическая компания должна придерживаться строгих нормативных требований при управлении данными клинических испытаний. Создав структуру управления данными, которая ставит соответствие приоритетом, организации могут снизить риски и избежать дорогостоящих штрафов.

Эффективное управление данными обеспечивает соблюдение нормативных требований, что становится все более важным в условиях строгих законов о защите конфиденциальности, таких как GDPR, CCPA [7]или PIPL. [8]Компании, которые внедряют комплексные стратегии управления данными, могут обеспечить соблюдение этих требований, избегая крупных штрафов и юридических последствий.

Появляется все большее соображение относительно прозрачности данных, которая подчеркивает открытость в отношении практики сбора, обработки и использования данных. Организации и заинтересованные лица должны четко сообщать о своей политике в отношении данных, включая то, как данные собираются, хранятся, передаются и используются. Предоставление пользователям легкого доступа к своим данным и возможности их исправления или удаления способствует формированию чувства контроля и доверия. Кроме того, минимизация данных является ключевым принципом, который выступает за сбор только тех данных, которые необходимы для определенной цели, тем самым снижая риск неправомерного использования и обеспечивая соблюдение законов о конфиденциальности.

Наконец, эффективная коммуникация и обучение являются важнейшими компонентами управления данными. Организации должны гарантировать, что все сотрудники понимают важность управления данными и обладают знаниями и навыками для соблюдения политик и процедур в отношении данных. Это может включать проведение учебных занятий, семинаров и кампаний по повышению осведомленности для продвижения культуры, основанной на данных. Например, технологическая компания может внедрить программу обучения по контролю данных для своих сотрудников, чтобы гарантировать, что они понимают передовые методы обработки данных и важность конфиденциальности данных. Развивая культуру

[7]Закон Калифорнии о защите конфиденциальности потребителей 2018 года (CCPA) предоставляет потребителям больше контроля над личной информацией, которую компании собирают о них, и содержит рекомендации для организаций по реализации этого закона.

[8]Закон Китая о защите персональных данных 2021 года (PIPL) устанавливает правила обработки персональных и конфиденциальных данных, включая правовые основания и требования к раскрытию информации.

управления данными, организации могут дать своим командам возможность принимать обоснованные решения и вносить вклад в общий успех организации.

3.4. Инфраструктура

Инфраструктурирование — это новый термин в дискурсе управления данными. Он точно отражает важную деятельность по управлению данными таким образом, как традиционные концепции не могут. Когда данные собираются или приобретаются из внешних источников, им нужно место для сохранения в течение желаемого периода времени. В результате курирование [9] или хранение данных были предложены в качестве деятельности по управлению данными. Однако курирование подразумевает тщательный отбор и организацию, которые некоторые современные решения для хранения (например, озера данных или память с произвольным доступом, обсуждаемые ниже) не требуют. Кроме того, соображения об инфраструктуре сбора и получения данных, местоположении центров обработки данных, их компоновке, вопросах связи и сетей выходят за рамки простых соображений хранения или курирования данных. Все это отражено в деятельности по инфраструктурированию — разработке и внедрению физических и организационных структур и объектов, необходимых для сбора и хранения данных.

Для получения данных необходимо наличие существующей инфраструктуры. При обсуждении получения мы затронули вопрос интерфейсов и процессов сбора данных. После того, как их форма и вид определены, их необходимо внедрить в определенный программный код и хранить и управлять ими на компьютерном оборудовании (например, на локальном сервере или в облаке). Эти интерфейсы доставляют данные с клиентских компьютеров (ноутбуков, смартфонов, датчиков Интернета вещей) в хост-системы (обычно отличные от клиента), которые временно или постоянно хранят данные.

Физические аспекты хранения данных относятся к компонентам инфраструктуры, используемой для хранения цифровой информации. Сюда входят различные типы оборудования, такие как оперативная память, жесткие диски (HDD), твердотельные накопители (SSD), ленточные накопители и решения для облачного хранения, каждое из которых имеет свои преимущества и ограничения.

Даже если данные удаляются без постоянного использования, их нужно где-то хранить, чтобы использовать в момент сбора. Обычно такое хранение включает в себя энергозависимые механизмы, такие как память с произвольным доступом. Эта память обычно очищается, как только данные больше не нужны, или новые данные заменяют старые. Тем не менее, в большинстве информационных систем данные хранятся более постоянно. Жесткие диски обычно используются из-за их большой емкости и экономической эффективности, что делает их подходящими для хранения больших объемов данных. Напротив, твердотельные накопители обеспечивают более высокую скорость чтения и записи, что делает их идеальными для приложений, требующих быстрого доступа к данным. Ленточные накопители, хотя и медленнее, часто используются для долгосрочного архивирования из-за их большой емкости и долговечности. По мере роста потребностей в хранении данных многие организации все чаще обращаются к облачному хранилищу, которое предоставляет масштабируемые решения без необходимости значительных инвестиций в физическую инфраструктуру.

Для систематической организации данных, которые должны храниться для эффективного доступа и безопасного контроля, была разработана экосистема различных технологий. Для хранения транзакционных и операционных данных многие организации полагаются на реляционные базы данных, традиционную технологию хранения. В то же время альтернативное семейство баз данных, NoSQL, неуклонно завоевывает долю рынка благодаря своим преимуществам, связанным с масштабируемостью и гибкостью. Для хранения данных для нужд аналитической отчетности традиционное решение реляционного хранилища данных продолжает оставаться популярным. В то же время для хранения огромных объемов гетерогенных данных и улучшения поддержки быстрорастущей области искусственного интеллекта озера данных становятся ценным подходом.

Реляционные базы данных являются традиционной и по-прежнему широко используемой формой технологии хранения данных, которая организует данные в таблицы или отношения, которые могут быть связаны на основе предопределенных отношений. Реляционная модель хранения баз данных была изобретена в 1970-х годах [11] и с тех пор является неотъемлемой частью мировой инфраструктуры данных, доказав свою безопасность и надежность. Эти базы данных используют язык структурированных запросов (SQL) для определения и обработки данных, что упрощает выполнение запросов (для чтения, записи и изменения данных). Известными примерами поставщиков реляционных баз данных являются Microsoft SQL Server, MySQL, PostgreSQL и Oracle Database. Реляционные базы данных идеально подходят для приложений, требующих структурированных данных и свойств ACID (атомарность, согласованность, изоляция, долговечность) для обеспечения надежной обработки транзакций и высокой целостности данных. Их основанная на схеме конструкция обеспечивает согласованность, но может ограничивать гибкость в обработке изменяющихся структур данных и неструктурированных или полуструктурированных данных.

В ответ на недостатки реляционного подхода при обработке больших объемов неструктурированных или полуструктурированных данных в качестве основной альтернативы появились базы данных NoSQL. Эти базы данных предлагают гибкие схемы и разработаны для масштабируемости и высокой производительности [3, 22, 23, 47] . Базы данных NoSQL подразделяются на четыре основных типа на основе их моделей данных, включая хранилища документов (например, MongoDB), хранилища ключей и значений (например, Redis), хранилища семейств столбцов (например, Apache Cassandra) и графовые базы данных (например, Neo4j). В отличие от реляционных баз данных, базы данных NoSQL не требуют фиксированной схемы, что делает их подходящими для приложений с динамическими требованиями к данным, таких как платформы социальных сетей, аналитика в реальном времени и искусственный интеллект. На рисунке 4 представлен пример структуры графовых данных.

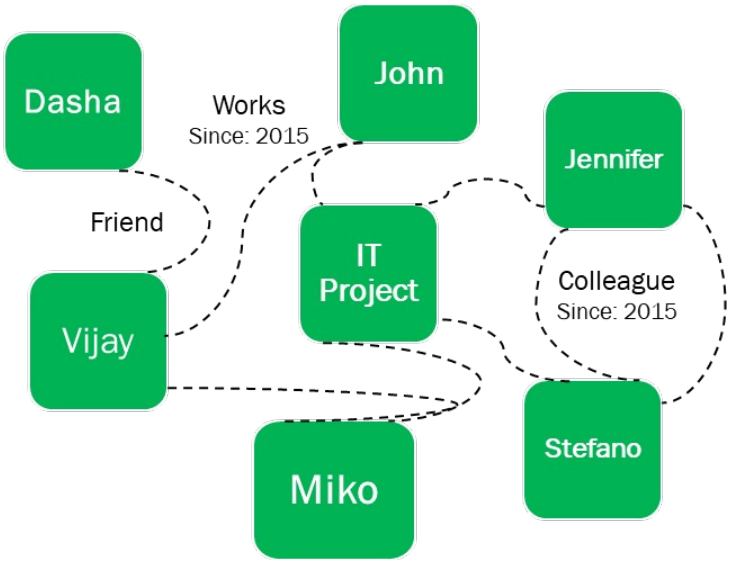

Рисунок 4. Пример графовой модели, показывающей узлы (люди, проект) и ребра (отношения). В некоторых графовых базах данных ребра могут иметь свойства.

Хранилища данных — это специализированные системы хранения, предназначенные для аналитической обработки и отчетности. Они консолидируют и хранят большие объемы структурированных данных из различных источников, позволяя выполнять сложные запросы и бизнес-аналитику [26] . Хранилища данных используют процессы извлечения, преобразования, загрузки (ETL), чтобы гарантировать чистоту, согласованность и оптимизацию данных для анализа. Они поддерживают онлайн-аналитическую обработку (OLAP) для многомерного анализа и оптимизированы для операций с большим объемом чтения. Примерами решений для хранилищ данных являются Amazon Redshift, Google BigQuery и Snowflake. Предоставляя

централизованный репозиторий исторических данных, хранилища данных позволяют организациям принимать решения на основе данных и получать информацию о тенденциях и производительности бизнеса.

Хранилища данных, особенно традиционных типов, работают на основе фиксированной схемы и испытывают трудности с поддержкой неоднородных данных, таких как текст, изображения и видео. Эти ограничения привели к появлению озер данных и связанных с ними технологий. Озера данных предназначены для хранения огромных объемов необработанных, неструктурированных, полуструктурированных и структурированных данных в их собственном формате [19, 20, 53] . Озера данных поддерживают подход «схема при чтении», позволяющий интерпретировать и структурировать данные во время анализа, а не во время приема. Эта гибкость делает озера данных идеальными для аналитики больших данных, машинного обучения и проектов по науке о данных. Такие технологии, как Apache Hadoop и Amazon S3, обычно используются для создания озер данных. Популярные платформы озер данных включают Amazon Lake Formation, Microsoft Azure Data Lake Storage, Google BigLake , Cloudera Data Platform, Databricks Delta Lake и Dremio . Предоставляя единое хранилище данных, озера данных позволяют организациям выполнять расширенную аналитику и извлекать ценную информацию из разнообразных наборов данных, но для обеспечения качества и удобства использования данных требуются надежные методы управления данными.

Серьезной проблемой является потенциальная возможность превращения озер данных в болота данных. Учитывая, что озера данных могут хранить огромные объемы иногда слабо связанных и плохо организованных несвязанных данных, если ими не управлять тщательно, они могут быстро стать загроможденными и беспорядочными, что затруднит поиск соответствующих данных и их связывание с другими данными. Следовательно, новая практика заключается в объединении озера данных с хранилищем данных, что может наложить структуру на объекты данных озера данных. Этот подход называется data lakehouse [25] . Databricks — известный поставщик хранилищ данных . Data Lakehouse — это последний эпизод (еще один важный эпизод — реляционное и NoSQL) в многолетней борьбе между структурой и отсутствием структуры в хранении данных.

Физическая компоновка центров обработки данных играет решающую роль в хранении данных. Центры обработки данных должны быть спроектированы для оптимизации производительности, надежности и безопасности. Это включает в себя такие соображения, как контроль температуры, электропитание и избыточность для предотвращения потери данных в случае отказа оборудования. Кроме того, физические меры безопасности, такие как контроль доступа, наблюдение и системы пожаротушения, жизненно важны для защиты конфиденциальных данных от несанкционированного доступа или повреждения. Поскольку объем данных продолжает расти, организации должны тщательно оценивать свои решения для физического хранения, чтобы убедиться, что они могут эффективно хранить, управлять и защищать свои данные, приспосабливаясь к будущему росту.

Сетевое взаимодействие и подключение являются частью управления данными, поскольку они формируют и ограничивают возможность обмена информацией. Важной инфраструктурой подключения являются интерфейсы прикладного программирования (API). API — это протоколы, которые позволяют информационным системам взаимодействовать друг с другом для обмена данными. С диверсификацией ландшафта хранения информации все больше доступа к данным и манипулирования ими осуществляются через API (в отличие от традиционного SQL). Следовательно, разработка API может стать либо посредником, либо узким местом в обмене данными [53] .

Ключевым фактором, который необходимо учитывать при хранении данных, является воздействие на окружающую среду. Зеленое хранение данных относится к экологически чистым методам и технологиям, используемым для управления хранением данных таким образом, чтобы минимизировать потребление энергии, сократить выбросы углерода и способствовать устойчивости [45] . Поскольку спрос на хранение данных продолжает расти экспоненциально, традиционные центры обработки данных подвергаются пристальному вниманию из-за их значительного потребления энергии и воздействия на окружающую среду. Зеленое хранение данных направлено на решение этих проблем путем принятия стратегий,

которые не только повышают эффективность, но и уменьшают экологический след операций по хранению данных.

Одним из основных компонентов зеленого хранения данных является использование энергоэффективного оборудования и инфраструктуры. Это включает использование SSD вместо традиционных жестких дисков, поскольку SSD обычно потребляют меньше энергии и выделяют меньше тепла. Кроме того, центры обработки данных все чаще используют энергоэффективные системы охлаждения и возобновляемые источники энергии, такие как солнечная или ветровая энергия, чтобы еще больше сократить потребление энергии. Решения виртуализации и облачного хранения также играют важную роль в зеленом хранении данных, оптимизируя использование ресурсов, позволяя нескольким виртуальным серверам работать на одном физическом сервере, тем самым снижая потребность в дополнительном оборудовании.

Инфраструктурирование включает в себя реализацию планов управления данными и процедур, связанных с безопасностью. Это включает в себя установку физической инфраструктуры и написание необходимого программного кода для обеспечения безопасности данных, доступа и разрешений, а также установку и обслуживание физических элементов безопасности, таких как здания, замки, внедрение физических элементов управления безопасностью.

Инфраструктуризация — это непрерывный процесс, поскольку он подразумевает частые изменения в объектах данных для того, чтобы соответствовать растущим требованиям использования данных и меняющимся требованиям. Поэтому не редкость порождать озеро данных в хранилище данных [25] , когда становится очевидным, что изначальное озеро данных превращается в болото данных (проблемы, которые мы обсудим позже). Изменения в инфраструктуре также распространены в попытке оптимизировать производительность поиска информации или минимизировать воздействие на окружающую среду. Таким образом, хотя это тесно связано с приобретением, это деятельность, которая также существует сама по себе.

3.5. Мероприятия по поддержке потребления

То, что данные хранятся и организованы, не означает, что к ним можно получить доступ и использовать их. Конечная деятельность по управлению данными направлена на то, чтобы сделать данные пригодными для использования, поддерживая их конечное потребление.

Существует множество различных подходов к улучшению данных для принятия решений, которые могут быть реализованы как часть усилий по управлению данными. Их отличие от действий по манипулированию данными, проводимых как часть потребления данных, заключается в их необычности для города по отношению к конкретной задаче. Это усилия по обеспечению того, чтобы данные могли поддерживать организационное или личное принятие решений в целом, в различных случаях и для различных целей. Метафорически эти действия можно рассматривать как океанский прилив, который одновременно поднимает все лодки. При выполнении этих действий будущие манипуляции данными и их потребление становятся более эффективными и действенными. Здесь мы выделяем некоторые из этих действий.

Оптимизация запросов. Традиционный подход к извлечению хранимой информации из баз данных заключается в использовании запросов, написанных на языках доступа к данным. Наиболее распространенным языком доступа и манипулирования данными является SQL (Structured Query Language), изначально созданный для поддержки реляционных баз данных. С ростом баз данных NoSQL появились новые языки доступа к данным, такие как Dynamo Query Language (DQL) для DynamoDB, MongoDB Query Language (MQL) и Cypher для Neo4J. Однако наиболее развитые подходы к оптимизации по-прежнему остаются для SQL.

Оптимизация запросов — это процесс повышения производительности запроса к базе данных путем минимизации использования ресурсов и времени выполнения при максимальной эффективности. Это включает анализ и преобразование операторов SQL, которые выполняются в отношении базы данных, чтобы гарантировать, что они извлекают желаемые результаты наиболее эффективным образом. Оптимизация запросов может включать такие методы, как переписывание запросов, создание или изменение индексов и корректировка

конфигураций базы данных для улучшения плана выполнения, которому следует ядро базы данных при обработке запроса.

Важным аспектом оптимизации запросов является использование планов выполнения, которые описывают шаги, которые база данных будет выполнять для выполнения запроса. Система управления базами данных (СУБД) генерирует эти планы на основе различных факторов, включая структуру данных, существующие индексы и сложность запроса. Оптимизаторы оценивают различные стратегии, чтобы определить наиболее эффективный способ извлечения запрошенных данных, часто учитывая стоимость операций, таких как объединения, сканирования и фильтры. Используя эффективные методы оптимизации запросов, организации могут значительно сократить время ответа для сложных запросов, повысить общую производительность системы и улучшить пользовательский опыт, в конечном итоге обеспечивая лучшее принятие решений на основе данных.

Оценка и улучшение качества данных. Оценка и оптимизация данных для будущего использования подразумевает постоянную оценку качества хранимых данных, включая их точность, полноту, согласованность, релевантность и своевременность. Распространенным методом является профилирование данных, которое включает анализ структуры, типов и содержимого набора данных для выявления потенциальных проблем. Это может включать проверку на наличие пропущенных значений, дублирующих записей и выбросов, которые могут указывать на ошибки ввода данных или аномалии. Такие инструменты и методы, как описательная статистика и визуализация данных, могут использоваться для получения информации о наборе данных и выделения областей, которые могут потребовать внимания.

В идеале организации должны установить конкретные критерии и ориентиры для оценки качества данных на основе предполагаемого использования набора данных [32] . Это включает сравнение набора данных с авторитетными источниками или стандартами для обеспечения точности и согласованности. Например, если набор данных используется для демографического анализа, он должен соответствовать признанным демографическим классификациям. Кроме того, релевантность должна оцениваться путем оценки того, соответствуют ли данные текущим потребностям организации или проекта. Наконец, необходимо учитывать своевременность данных, гарантируя, что они актуальны и отражают самую последнюю доступную информацию. Систематически оценивая эти измерения, организации могут выявлять проблемы качества данных, расставлять приоритеты в усилиях по исправлению ситуации и в конечном итоге повышать надежность и удобство использования своих наборов данных для принятия решений.

Частью обзора и оценки качества данных является определение того, следует ли архивировать или удалять данные, если они больше не нужны. Архивированные данные могут храниться на более медленных носителях и дополнительно сжиматься для дальнейшей оптимизации эффективности хранения и снижения общего воздействия на хранение и окружающую среду.

Интеграция данных и взаимодействие. Интеграция данных и взаимодействие возникают как существенная проблема управления данными, поскольку ни данные, ни системы, которые обрабатывают данные, не существуют изолированно. Чтобы быть пригодными для использования, смоделированные, полученные и размещенные данные должны быть доступны для своих пользователей. Это приводит к интеграции данных – интеграции данных с пользователями данных и другими системами – как к общей деятельности по управлению данными.

Несмотря на свою важность и повсеместность, интеграция данных часто упускается из виду как деятельность по управлению данными. Действительно, ни одно из определений управления данными, приведенных в Приложении А, не упоминает ее явно. Тем не менее, недостаточное внимание к вопросам интеграции данных часто приводит к катастрофическим сбоям. Среди проблем, которые обрекли проект MD Andreson-Watson (упомянутый ранее), была неспособность справиться со сложностью интеграции огромных объемов медицинских записей из различных источников. В том же духе генеративные инструменты ИИ, такие как ChatGPT, галлюцинируют, то есть создают бессмысленную информацию [28] . Галлюцинации ИИ часто возникают из-за неспособности должным образом интегрировать различные источники данных, используемые для обучения этих технологий.

Чтобы сделать данные полезными, интеграция охватывает контроль версий, управление метаданными и отслеживание происхождения данных. Контроль версий гарантирует, что изменения в моделях данных документируются и могут быть отменены при необходимости, в то время как управление метаданными предоставляет контекст и информацию о данных, такую как их источник, структура и цель. Отслеживание происхождения данных помогает понять поток данных на протяжении всего жизненного цикла, позволяя отслеживать происхождение данных и обеспечивать соответствие политикам управления данными.

Управление метаданными включает в себя систематическую организацию и обслуживание метаданных, которые являются данными, описывающими и предоставляющими контекст для других данных. Этот процесс имеет решающее значение для улучшения управления данными, качества и удобства использования в организации. Эффективное управление метаданными помогает пользователям понимать происхождение, значение и взаимосвязи данных, позволяя им принимать обоснованные решения и использовать данные более эффективно. Например, в организации здравоохранения метаданные могут включать информацию о записях пациентов, такую как источник данных, структура полей данных и частота обновления данных. Эта контекстная информация гарантирует, что специалисты в области здравоохранения могут доверять данным и использовать их для точного ухода за пациентами.

Ключевым компонентом управления метаданными является создание централизованного репозитория или каталога метаданных. Этот репозиторий действует как единый источник истины для всех метаданных в организации, что упрощает для заинтересованных сторон поиск и доступ к соответствующим данным. Например, компания, предоставляющая финансовые услуги, может внедрить систему управления метаданными, которая каталогизирует различные наборы данных, включая записи транзакций, профили клиентов и данные о соответствии нормативным требованиям. Эта система будет предоставлять подробные описания, происхождение данных, классификацию данных и информацию о владельце данных, позволяя аналитикам и сотрудникам по обеспечению соответствия быстро находить необходимые данные и понимать их контекст и ограничения. Централизуя метаданные, организации могут улучшить сотрудничество между командами, сократить избыточность и улучшить обнаружение данных.

Интеграция данных может быть облегчена использованием онтологий доменов. Онтологии являются формальными представлениями знаний в пределах определенного домена, определяющими концепции, отношения и правила, которые управляют данными в этом домене [21, 39] . Они предоставляют общий словарь и структурированную структуру для понимания и интеграции разнообразных наборов данных, облегчая взаимодействие между различными системами и приложениями. Предоставляя общее понимание данных, онтологии помогают преодолеть разрывы между разнородными системами, облегчая эффективное объединение и анализ данных.

Одним из преимуществ использования онтологий для интеграции данных является их способность моделировать сложные отношения. Например, в секторе здравоохранения онтология может определять отношения между пациентами, методами лечения, лекарствами и результатами. Сопоставляя данные из различных источников, таких как электронные медицинские карты, базы данных клинических испытаний и страховые претензии, с этой онтологией, поставщики медицинских услуг могут достичь единого представления информации о пациентах. Такая интеграция позволяет проводить более комплексный анализ, улучшать процесс принятия решений и улучшать результаты для пациентов.

Онтологии также поддерживают взаимодействие данных, обеспечивая семантическое обоснование, что позволяет системам выводить новые знания на основе существующих отношений, определенных в онтологии. Например, если онтология определяет, что «кардиолог — это тип врача» и «врач может выписывать лекарства», система, использующая эту онтологию, может автоматически вывести, что «кардиолог может выписывать лекарства». Эта возможность повышает способность различных систем работать вместе и обмениваться идеями.

Обогащение данных. Последняя деятельность, которая становится все более популярной, — это обогащение данных, преобразования и дополнения к исходному набору данных, которые в целом повышают его способность поддерживать принятие решений, анализ и

действия. Например, организация, которая широко использует машинное обучение ИИ, может заниматься общими преобразованиями функций и инженерными работами.

Инженерия признаков включает в себя создание и преобразование входных данных в значимые признаки, которые могут улучшить производительность моделей машинного обучения. Например, в наборе данных, содержащем временные метки транзакций, можно извлечь такие признаки, как час дня, день недели или даже праздники, чтобы зафиксировать временные закономерности. В наборе данных с текстовыми данными такие методы, как извлечение количества слов, n-грамм или оценок настроений, могут превратить неструктурированный текст в структурированные признаки [16] . Если в наборе данных отсутствуют значения, эти значения могут быть вменены [7] или могут быть предприняты усилия для получения отсутствующих значений.

Чтобы сделать данные дружественными к ИИ для различных задач, проектирование признаков может проводиться до того, как данные будут использованы учеными по данным и инженерами по проектированию ИИ. Обогащая исходный набор данных соответствующими признаками, проектирование признаков может значительно повысить точность и интерпретируемость моделей машинного обучения, что приводит к лучшему принятию решений и получению более глубокого понимания. Естественно, эти усилия необходимо рассматривать в контексте других приоритетов, таких как стоимость и влияние энергии на хранение дополнительных данных. В то же время проектирование общих признаков может также привести к общему сокращению ресурсов хранения и вычислений, поскольку это единовременное усилие, которое заменяет множественные действия по преобразованию признаков различными командами машинного обучения.

3.6. Помимо управления данными

Результатом управления данными является предоставление данных, которые могут быть использованы в качестве входных данных для принятия решений, анализа и действий. Эффективное, действенное и ответственное управление данными позволяет организациям и людям раскрывать возможности информационных технологий и использовать их в полной мере. Это может вывести организации на значительную известность и даже изменить характер отраслей и рынков. Например, новаторский подход Netflix к развлечениям породил новую отрасль, вытеснив традиционных гигантов проката CD и DVD, таких как Blockbuster. Этот подвиг стал возможным благодаря инновационным и надежным методам управления данными, которые позволили Netflix делиться контентом напрямую и предлагать персонализированные рекомендации миллионам зрителей. В аналогичном ключе компании по всему миру конкурируют на основе данных [15] , получая конкурентное преимущество за счет лучшего моделирования данных, приобретения, управления, инфраструктуры и готовности и поддержки потребления.

Управление данными может значительно повысить эффективность работы. Правильно управляемые данные облегчают интеграцию и доступ между различными отделами, способствуя сотрудничеству и инновациям. Использование Amazon управления данными для оптимизации операций своей цепочки поставок является ярким примером. Интегрируя данные из различных источников и применяя расширенную аналитику, Amazon может оптимизировать управление запасами, сократить время доставки и повысить удовлетворенность клиентов.

Управление данными — это не только бизнес или организационная деятельность. Поскольку цифровая информация формирует повседневную жизнь практически каждого человека на планете, хорошие навыки управления данными могут значительно улучшить личные финансовые и социальные результаты. Например, хорошо организованные папки на персональном ноутбуке для личных документов, таких как счета, квитанции и гарантии, позволяют быстро находить и ссылаться на важную информацию и избегать дорогостоящих счетов или ошибок в налоговых декларациях. Эффективное управление данными также может уменьшить информационную перегрузку путем фильтрации и приоритизации соответствующих данных, позволяя людям сосредоточиться на высокоприоритетных задачах и быстро принимать обоснованные решения. Аналогичным образом, просто организуя

фотографии и видео на смартфоне осмысленным и интуитивно понятным способом, становится проще делиться особыми моментами с другими. Эти улучшения приводят к лучшему управлению временем, улучшенному принятию решений и повышению общей производительности в повседневной деятельности.

Управление данными не заканчивается, когда данные готовы к последующему использованию. Мониторинг и оценка эффективности данных и, как следствие, управления данными способствуют обучению и совершенствованию практик. Следовательно, специалисты по управлению данными должны активно искать обратную связь о том, в какой степени предоставленные ими данные помогли создать личную и организационную ценность, и вносить необходимые коррективы в практику управления данными.

4. Выводы и будущее МАГИИ

В эпоху, все больше формируемую информационными технологиями, важнейшая дисциплина управления данными часто борется за то признание, которого она заслуживает, особенно по сравнению с новаторскими инновациями, которые она способствует, такими как искусственный интеллект, социальные сети и аналитика данных. Несмотря на свою важную роль в обеспечении этих достижений, управление данными часто воспринимается как второстепенная задача, что приводит к отсутствию инвестиций и внимания в этой жизненно важной области. Усугубляет эту проблему двусмысленность, окружающая определение управления данными и различных видов деятельности, которые оно охватывает. Эта путаница затрудняет усилия по донесению значимости эффективных методов управления данными до заинтересованных сторон в организациях и в обществе.

Для решения этих проблем в данной статье представлена всеобъемлющая и простая структура для понимания управления данными под названием MAGIC. Эта структура включает пять ключевых видов деятельности: моделирование, получение, управление, инфраструктуру и поддержку потребления. Каждый из этих компонентов играет важную роль в общем управлении данными, обеспечивая структурированный подход, который способствует ясности и согласованности в практиках, связанных с данными. Разграничивая эти виды деятельности, структура MAGIC не только улучшает понимание управления данными, но и служит ценным инструментом для обучения, исследований и практического применения в различных организационных контекстах.

Используя фреймворк MAGIC, организации могут способствовать более глубокому пониманию управления данными и его ключевой роли в использовании всего потенциала информационных технологий. Кроме того, этот фреймворк снабжает специалистов по данным, исследователей и преподавателей общим языком и методологией для обсуждения и внедрения практик управления данными. При этом мы стремимся преодолеть разрыв между управлением данными и технологиями, которые оно поддерживает, в конечном итоге продвигая более целостный подход к данным как к жизненно важному активу. С помощью этой инициативы мы надеемся поднять разговор об управлении данными на новый уровень и поощрить возобновление внимания к его важности в мире, все больше управляемом данными.

Ссылки

1. Abraham, R. et al.: Data governance: A conceptual framework, structured review, and research agenda. International journal of information management. 49, 424–438 (2019).
2. Ballou, D.P., Pazer, H.L.: Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. Information Systems Research. 6, 1, 51 (1995).
3. Bugiotti, F. et al.: Database design for NoSQL systems. Presented at the International Conference on Conceptual Modeling (2014).
4. Burton-Jones, A. et al.: Assessing representation theory with a framework for pursuing success and failure. MIS Quarterly. 41, 4, 1307–1333 (2017).
5. Burton-Jones, A. et al.: Guidelines for Empirical Evaluations of Conceptual Modeling Grammars. Journal of the Association for Information Systems. 10, 6, 495–532 (2009).

6. Castellanos, A. et al.: Basic Classes in Conceptual Modeling: Theory and Practical Guidelines. Journal of the Association for Information Systems. 21, 4, 1001–1044 (2020).
7. Castellanos, A. et al.: Improving machine learning performance using conceptual modeling. In: AAAI Symposium on Combining Machine Learning and Knowledge Engineering in Practice. pp. 1–4 , Stanford, CA (2021).
8. Chen, P.: The entity-relationship model - toward a unified view of data. ACM Transactions on Database Systems. 1, 1, 9–36 (1976).
9. Chua, C.E.H. et al.: Data Management. MISQ Quarterly Online. 1–10 (2022).
10. Clegg, B.: The God effect: Quantum entanglement, science's strangest phenomenon. Macmillan, London, UK (2006).
11. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM. 13, 6, 377–387 (1970).
12. Compagnucci, I. et al.: Trends on the Usage of BPMN 2.0 from Publicly Available Repositories. Presented at the International Conference on Business Informatics Research (2021).
13. Couger, J.D., Knapp, R.W. eds: System analysis techniques. John Wiley & Sons, New York NY (1974).
14. DAMA et al.: DAMA-DMBOK: Data Management Body of Knowledge. Technics Publications, Sedona, AZ (2017).
15. Davenport, T.H.: Competing on analytics. harvard business review. 84, 1, 98–108 (2006).
16. Duboue, P.: The Art of Feature Engineering: Essentials for Machine Learning. Cambridge University Press, Cambridge, UK (2020).
17. Filippouli, E.: AI: The Pinnacle of our Ingenuity, https://www.globalthinkersforum.org/news-and-resources/news/ai-the-pinnacle-of-our-ingenuity, last accessed 2022/09/28.
18. Gill: Whoever leads in artificial intelligence in 2030 will rule the world until 2100, https://www.brookings.edu/blog/future-development/2020/01/17/whoever-leads-in-artificial-intelligence-in-2030-will-rule-the-world-until-2100/, last accessed 2021/09/25.
19. Gopalan, R.: The Cloud Data Lake. OReilly Media, Inc, Sebastopol, CA (2022).
20. Gorelik, A.: The enterprise big data lake: Delivering the promise of big data and data science. O'Reilly Media (2019).
21. Guizzardi, G., Proper, H.A.: On Understanding the Value of Domain Modeling. In: Proceedings of 15th International Workshop on Value Modelling and Business Ontologies (VMBO 2021). (2021).
22. Harrison, G.: Next Generation Databases: NoSQL, NewSQL, and Big Data. Apress, New York, NY, USA (2015).
23. Hewasinghage, M. et al.: Modeling strategies for storing data in distributed heterogeneous NoSQL databases. Presented at the International Conference on Conceptual Modeling (2018).
24. Hvalshagen, M. et al.: Empowering Users with Narratives: Examining The Efficacy Of Narratives For Understanding Data-Oriented Conceptual Models. Information Systems Research. 34, 3, 890–909 (2023).
25. Inmon, B., Srivastava, R.: Rise of the Data Lakehouse. Technics Publications, New York NY (2023).
26. Inmon, W.H. et al.: DW 2.0: The architecture for the next generation of data warehousing. Elsevier, New York NY (2010).
27. Jacobson, I. et al.: The unified software development process. Addison-Wesley, Reading MA (1999).
28. Ji, Z. et al.: Survey of hallucination in natural language generation. ACM Computing Surveys. 55, 12, 1–38 (2023).
29. Kent, W.: Data and reality: basic assumptions in data processing reconsidered. North-Holland Pub. Co., Amsterdam, Netherlands (1978).
30. Klimbie, J.W., Koffeman, K.L.: Data Base Management: Proceedings of the IFIP Working Conference on Data Base Management. North-Holland, London (1974).

31. Lango, L.: The Revolutionary Tech Supercharging Gains In the Age of AI, https://investorplace.com/hypergrowthinvesting/2024/01/putting-ai-on-the-fast-track-to-sure-fire-success/, last accessed 2024/07/27.
32. Lee, Y.W. et al.: AIMQ: A methodology for information quality assessment. Information & Management. 40, 2, 133–146 (2002).
33. Lukyanenko, R. et al.: Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-generated Content. MISQ. 43, 2, 634–647 (2019).
34. Lukyanenko, R. et al.: Inclusive Conceptual Modeling: Diversity, Equity, Involvement, and Belonging in Conceptual Modeling. In: ER Forum 2023. pp. 1–4 Springer, Lisbon, Portugal (2023).
35. Lukyanenko, R. et al.: Principles of universal conceptual modeling. In: EMMSAD 2023. pp. 1–15 Springer, Saragosa, Spain (2023).
36. Lukyanenko, R. et al.: System: A Core Conceptual Modeling Construct for Capturing Complexity. Data & Knowledge Engineering. 141, 1–29 (2022).
37. Lukyanenko, R. et al.: The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-generated Content. Information Systems Research. 25, 4, 669–689 (2014).
38. Lukyanenko, R., Parsons, J.: Design Theory Indeterminacy: What is it, how can it be reduced, and why did the polar bear drown? Journal of the Association for Information Systems. 21, 5, 1–30 (2020).
39. McDaniel, M., Storey, V.C.: Evaluating Domain Ontologies: Clarification, Classification, and Challenges. ACM Computing Surveys. 53, 1, 1–40 (2019).
40. Norman, D.A.: The design of everyday things. Bsic Books, New York, NY (2002).
41. Olivé, A.: Conceptual modeling of information systems. Springer Science & Business Media, Berlin, Germany (2007).
42. Oracle Inc: What is Data Management?, https://www.oracle.com/database/what-is-data-management/, last accessed 2024/07/28.
43. Rao, A.S., Verweij, G.: Sizing the prize: What's the real value of AI for your business and how can you capitalise, (2017).
44. Recker, J. et al.: From Representation to Mediation: A New Agenda for Conceptual Modeling Research in A Digital World. MIS Quarterly. 45, 1, 269–300 (2021).
45. Recker, J.: Toward a design theory for green information systems. Presented at the System Sciences (HICSS), 2016 49th Hawaii International Conference on (2016).
46. Rodriguez, J.: Some AI Lessons from Watson's Failure at MD Anderson, https://jrodthoughts.medium.com/some-ai-lessons-from-watsons-failure-at-md-anderson-9b895cf70840, last accessed 2024/07/28.
47. Sadalage, P.J., Fowler, M.: NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Pearson Education, New York NY (2013).
48. Sambasivan, N. et al.: "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Presented at the proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021).
49. SAP: What is data management? | Definition, importance, & processes, https://www.sap.com/products/technology-platform/what-is-data-management.html, last accessed 2024/07/28.
50. Shneiderman, B.: Designing the user interface. Pearson Education, Boston, MA (2003).
51. Storey, V.C. et al.: Conceptual Modeling: Topics, Themes, and Technology Trends. ACM Computing Surveys. 55, 14s, 1–38 (2023).
52. Storey, V.C. et al.: Explainable AI: Opening the Black Box or Pandora's Box? Communications of the ACM. 1–6 (2022).
53. Strengholt, P.: Data Management at scale. O'Reilly Media, Inc, New York (2020).
54. Strong, D.M. et al.: Data quality in context. Communications of the ACM. 40, 5, 103–110 (1997).
55. Tableau: What Is Data Management? Importance & Challenges | Tableau, https://www.tableau.com/learn/articles/what-is-data-management, last accessed 2024/07/28.

56. Tang, Z. et al.: A self-adaptive Bell–LaPadula model based on model training with historical access logs. IEEE Transactions on Information Forensics and Security. 13, 8, 2047–2061 (2018).
57. Thalheim, B.: Modelology—The New Science, Life and Practice Discipline. In: Information Modelling and Knowledge Bases XXXV. pp. 1–19 IOS Press, Netherlands (2024).
58. Timonera, K.: What is Data Management? A Guide to Systems, Processes, and Tools, https://www.datamation.com/big-data/what-is-data-management/, last accessed 2024/07/28.
59. Wand, Y., Weber, R.: On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal. 3, 4, 217–237 (1993).
60. Wang, R.Y. et al.: A framework for analysis of data quality research. Knowledge and Data Engineering, IEEE Transactions on. 7, 4, 623–640 (1995).
61. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems. 12, 4, 5–33 (1996).
62. Zarraga-Rodriguez, M., Alvarez, M.J.: Experience: information dimensions affecting employees' perceptions towards being well informed. Journal of Data and Information Quality (JDIQ). 6, 2–3, 1–14 (2015).

Приложение A: Основы фреймворка MAGIC

Фреймворк MAGIC (Modeling, Acquisition, Governance, Infrastructuring, Consumption support) вырос из проекта по курированию публикаций по управлению данными в научном журнале MIS Quarterly [9] , который был выполнен командой ученых по управлению данными со всего мира. Проект был запрошен MIS Quarterly, ведущим рецензируемым академическим журналом, который публикует исследования по управлению информационными системами и широко рассматривается как ведущий журнал в дисциплине информационных систем.

Проект кураторства управления данными MIS Quarterly концептуализировал управление данными как 5C управления данными. Он определил область управления данными как область изучения, которая исследует и разрабатывает действия и методы концептуализации, сбора, курирования, потребления и контроля данных для поддержки понимания, анализа и действий. Это определение и концептуализации легли в основу MAGIC, поскольку он далее разрабатывал и совершенствовал эти концепции. Во-первых, структура MAGIC расширяет концептуализацию до более широкой деятельности моделирования. Во-вторых, она расширяет деятельность по сбору данных до более широкой деятельности по получению данных, которая также включает получение существующих данных из других систем. Несколько узконаправленная деятельность по кураторству данных заменяется новой концепцией инфраструктуры данных.

Структура MAGIC не включает деятельность по потреблению данных. Мы считаем, что это выходит за рамки управления данными. Управление данными, как и любая область практики, должно иметь четко определенные границы. Потребление данных происходит в результате и как следствие управления данными, но само по себе является отдельной деятельностью, которая включает использование данных для принятия решений, понимания и действий. Однако бесспорной является необходимость поддержки потребления различными методами управления данными, такими как управление метаданными. Следовательно, MAGIC усовершенствовал деятельность по потреблению, назвав ее поддержкой потребления.

Наконец, деятельность контроля сохраняется, но ее сфера существенно расширяется, чтобы лучше учитывать развивающиеся практики управления данными. Кроме того, в MAGIC используется более традиционная метка управления, в соответствии с отраслевыми и исследовательскими практиками [1] .

В дополнение к фреймворку 5Cs of Data Management, MAGIC рассмотрел ведущие определения управления данными. Некоторые из них приведены в Таблице A1. Мы рассмотрели эти определения, чтобы гарантировать, что основные аспекты управления данными охвачены MAGIC, как их понимают ведущие мыслители в области управления данными. Сравнение определений в Таблице A1 с MAGIC показывает, что фреймворк Magic

охватывает соображения ключевых определений управления данными, тогда как любое данное определение в таблице не является таким всеобъемлющим, как фреймворк MAGIC.

Таблица А 1Популярные определения управления данными

Определение управления данными	**Источник**	**Наш анализ**
Управление данными — это разработка, выполнение и контроль планов, политик, программ и практик, которые предоставляют, контролируют, защищают и повышают ценность данных и информационных активов на протяжении всего их жизненного цикла.	ДАМА Интернешнл [14]	• Сосредоточьтесь на ценности данных, а не на самих данных • Неясно, какие действия, связанные с данными, являются частью управления данными. • Моделирование, как существенная деятельность, не упоминается явно (подразумевается, если понятие «планирование» трактовать как моделирование)
Управление данными — это практика безопасного, эффективного и экономически выгодного сбора, хранения и использования данных.	Оракул [42]	• Использование данных, как правило, выходит за рамки управления данными. • Не упоминается ценность или цели данных или причина управления данными. • Моделирование, важная деятельность, не упоминается явно
Управление данными — это ИТ-дисциплина, ориентированная на прием, подготовку, организацию, обработку, хранение, поддержку и защиту данных на всем предприятии.	Datamation [58]	• Фокус внимания на организации, но управление данными также является проблемой для отдельных лиц • Не упоминается ценность или цели данных или причина управления данными. • Прием данных — один из многих типов сбора данных (наряду со сбором с нуля). • Моделирование, важная деятельность, не упоминается явно • Организация предлагается после сбора, тогда как на самом деле организация (как часть моделирования данных) формирует сбор
Управление данными — это практика сбора, организации, защиты и хранения данных организации с целью их анализа для принятия бизнес-решений.	Табло от Salesforce [55]	• Фокус внимания на организации и бизнесе, но управление данными также является проблемой для отдельных лиц • Цель управления данными – принятие решений – ясна • Моделирование, важная деятельность, не упоминается явно • Организация предлагается после сбора, тогда как на самом деле организация (как часть моделирования данных) формирует сбор
Управление данными — это практика сбора, организации, управления и доступа к данным для поддержки	САП [49]	• Организация предлагается после сбора, тогда как на самом деле организация (как часть моделирования данных) формирует сбор

производительности, эффективности и принятия решений.		• Управление как деятельность является циклическим по отношению к основной концепции • Сосредоточьтесь на результатах: производительности, эффективности и принятии решений • Доступ к данным — это действие, выходящее за рамки управления данными; однако подготовка данных для эффективного доступа — это