

Эволюция управления данными



Андрей Ермоленко, заведующий лабораторией информационно-аналитических систем Объединенного института проблем информатики НАН Беларуси, кандидат технических наук



Андрей Говин, заместитель генерального директора по научной и инновационной работе ОИПИ НАН Беларуси, кандидат технических наук, доцент

В последние два десятилетия данные претерпели фундаментальную трансформацию своего статуса в экономической системе. С середины XX в. они рассматривались как форма представления информации в формализованном (цифровом) виде, позволяющем автоматизировать ее сбор, хранение и дальнейшую обработку средствами вычислительной техники. Такая ситуация определялась техническими ограничениями систем хранения, а также непосредственно самой инфраструктурой. Однако с начала XXI в. произошли революционные изменения материальной базы вычислительной техники и пропускной способности каналов связи, что заложило фундамент для трансформации парадигмы разработки программного обеспечения и использования данных.

Один из видных теоретиков в области создания языков программирования Никлаус Вирт в 1995 г. в форме шутки метко сформулировал ставшую заметной проблему: «программы становятся медленнее куда шустрее, чем компьютеры становятся быстрее», за этим высказыванием позже закрепился термин «закон Вирта» [1]. Этот закон часто противопоставляют знаменитому высказыванию Гордона Мура (закон Мура), который гласит, что производительность процессоров удваивается каждые 18–24 месяца из-за сочетания роста количества транзисторов и увеличения тактовых частот процессоров [2]. Следующие два десятилетия развитие информационных технологий со стороны выглядело в строгом соответствии с законом Вирта, пока не вышли на рынок первые большие языковые модели.

Появление доступных средств искусственного интеллекта в фантастически короткие сроки изменило расстановку приоритетов в информационных технологиях. Далее будет рассмотрен вопрос, как изменилось отношение к данным, которые возникают, обрабатываются и хранятся в информационных системах. Суть изменений отражена в емком высказывании Джека Ма (основателя сервиса Alibaba): «Сегодня данные – это новая нефть» [3]. Вместе с тем, чтобы получить прибыль, необходимо добыть сведения, очистить и превратить в полезный информационный продукт.

Целью настоящей статьи является системный анализ эволюции управления данными, раскрытие механизмов их трансформации в стратегический актив, а также исследование современных моделей их монетизации, обмена и вторичного использования.

Эволюция управления: от записей к стратегическому активу

Процесс совершенствования управления данными можно условно разделить на 5 этапов, каждый из которых характеризуется изменением роли этого ресурса в организационных и экономических процессах *(таблица)*.

На первых двух этапах данные выполняли преимущественно вспомогательную функцию. В архивный период они существовали в виде бумажных записей, картотек, журналов и служили для фиксации фактов хозяйственной деятельности. Их ценность определялась возможностью ретроспективного анализа и подтверждения правовых отношений.

С возникновением первых электронных вычислительных машин и реляционных баз данных эти сведения стали средством автоматизации операционных процессов: учета транзакций, управления запасами, ведения бухгалтерии. Сформировались базовые принципы управления, которые гарантировали безопасную и надежную работу: атомарность, согласованность, изолированность и долговечность – известные как ACID-свойства. Управление сводилось к корректному хранению и быстрому доступу в рамках операционных задач.

В 1990-е гг. с развитием технологий хранилищ и инструментов бизнес-аналитики (OLAP, BI-системы) данные приобрели новое качество – они стали источником знаний для принятия управленческих решений. Именно тогда зародились первые практики управления (Data Governance) – политики, стандарты, роли и процессы обеспечения качества, а также профессии, связанные с этим – Data Steward, Data Architect, Data Analyst.

Революционные изменения произошли в 2010-е гг. с появлением технологий больших данных (big data) и облачных вычислений. Объемы информации, генерируемой цифровыми платформами (социальные сети, электронная коммерция, мобильные приложения), на несколько порядков превысили потенциал традиционных хранилищ. Такие технологии, как Apache Hadoop, Apache Spark, NoSQL, позволили обрабатывать петабайты структурированных и неструктурированных данных. На этом этапе они стали основой для создания цифровых продуктов и сервисов. Компании Google, Amazon, Facebook продемонстрировали, что сведения о пользователях, собранные в процессе предоставления бесплатных сервисов, могут быть конвертированы в значительную экономическую ценность через таргетированную рекламу и персонализацию.

С 2020-х гг. управление данными вступило в новую фазу, характеризующуюся признанием их стратегическим активом, подлежащим учету на балансе организаций, монетизации и обмену в рамках дата-экосистем. Ключевыми трендами этого этапа стали:

- ***Data as a Service (DaaS)*** *– облачная модель, с помощью которой пользователи получают доступ к нужной информации по требованию и подписке без необходимости самостоятельно собирать, хранить и настраивать инфраструктуру для нее;*
- ***Data mesh (или сетка данных)*** *– децентрализованный подход к управлению аналитическими материалами, при котором ответственность за конкретный их фрагмент делегируется владельцам (так называемое «доменное владение»);*
- ***Data Fabric (или фабрика данных)*** *– архитектура по объединению разрозненных источников информации в единую экосистему без необходимости их физической централизации;*
- ***Data Marketplace*** *– цифровые платформы для поиска, покупки, продажи и обмена сведениями между поставщиками и потребителями.*

Именно на этом этапе данные окончательно превратились из вспомогательного ресурса в самостоятельный объект экономических отношений.

Этап	Период, гг.	Роль данных	Ключевые технологии	Парадигма управления
Архивный	до 1960-х	Пассивные записи, исторические свидетельства	Бумажные носители, картотечные системы	Хранение и поиск
Операционный	1960–1980-е	Средство автоматизации транзакций	Реляционные СУБД (IMS, DB2)	Обеспечение целостности
Аналитический	1990–2000-е	Источник бизнес-аналитики	Хранилища данных (Data Warehouse), OLAP, BI	Извлечение знаний
Платформенный	2010-е	Основа цифровых продуктов	Big Data, Hadoop, облачные платформы	Создание ценности
Экосистемный	с 2020-х	Стратегический актив, товар, инфраструктура	Data mesh, Data fabric, AI/ML, API-экономика	Монетизация и обмен

Таблица. Этапы эволюции управления данными

Данные как стратегический актив

Признание данных активом требует выполнения ряда условий, сформулированных в рамках концепции «Data as an asset». Согласно определению международной некоммерческой организации The Data Management Association International (DAMA), систематизирующей знания и опыт в области управления данными, они являются активом, если [4]:

- *имеют экономическую ценность, то есть способны генерировать доход или снижать затраты;*
- *находятся под контролем организации, обладающей правами на них;*
- *идентифицируемы и измеримы, то есть могут быть учтены и оценены.*

Совет по международным стандартам финансовой отчетности International Accounting Standards Board (IASB) с 2023 г. начал проработку вопроса о включении данных в систему международных стандартов финансовой отчетности, что свидетельствует о постепенном признании их ликвидным активом на уровне глобальной финансовой системы.

Оценка их стоимости представляет собой сложную задачу, поскольку в отличие от традиционных ресурсов данные не изнашиваются физически, могут использоваться многократно и их потенциал зависит от контекста эксплуатации.

При этом они обладают разной ценностью, определяемой следующими характеристиками:

- ***уникальность*** *(отсутствие таких материалов у конкурентов);*
- ***актуальность*** *(релевантность во времени, соответствие текущим бизнес-задачам);*
- ***качество*** *(точность, полнота и непротиворечивость);*
- ***масштабируемость*** *(возможность включения в различные сценарии);*
- ***взаимосвязанность*** *(интеграция с другими наборами).*

При этом особую значимость представляют так называемые темные данные (dark data) – информация, которая собрана организацией в ходе своей деятельности, но не задействована для аналитики и принятия решений. По оценкам IBM, до 80–90% корпоративных данных относятся к категории «темных», что создает значительный нереализованный потенциал [5]. По какой же причине они потенциально очень ценны? Это связано с тем, что применяемые математические подходы обработки данных позволяют существенно улучшить ее результаты при дополнительном задействовании сведений, связанных с исходными. С одной стороны, известно, что существует проблема сложности информационного объекта, которая была сформулирована более 50 лет назад А.Н. Колмогоровым в алгоритмической теории информации (где было доказано, что задача вычисления алгоритмической сложности объекта алгоритмически неразрешима). Это обстоятельство, так называемая колмогоровская сложность, определяет, что модель объекта никогда не будет в полной мере описывать все его свойства. Но с другой стороны, большинство инструментов, связанных с big data, тесно интегрированы с ИИ и алгоритмами машинного обучения, которые в свою очередь опираются на теорию вероятности, и прежде всего на теорему Томаса Байеса. В самом общем виде ее смысл сводится к утверждению, что изначальное (априорное) суждение (событие А), дополненное новой информацией (событие В), даст более приближенную к реальности картину [6]. Другими словами, логика Байеса предопределяет уточнение вероятности того или иного явления, учитывая и ранее известные факты, и новые наблюдения. В качестве дополнительной информации как раз и выступают темные данные: промежуточные отчеты, логи и системные журналы, сведения с датчиков, записи с камер видеонаблюдения и др.

Монетизация данных

Процесс, отвечающий за преобразование накопленной информации в настоящий источник дополнительной прибыли и конкурентных преимуществ перед другими компаниями, – не что иное, как монетизация данных. Она дает бизнесу возможность эффективно использовать большие data-массивы, превращая их в полноценный актив на рынке [7]. При этом прямая монетизация – это получение дохода от продажи или лицензирования данных третьим сторонам, а косвенная – создание ценности за счет работы с ними для улучшения собственных бизнес-процессов, продуктов и услуг.

В 2025 г. расходы на глобальном рынке инструментов big data составили 394,7 млрд долл. В 2026 г. объем рассматриваемого рынка, как ожидается, достигнет 447,68 млрд долл. По мнению аналитиков Fortune Business Insights, в дальнейшем среднегодовой темп роста в сложных процентах (CAGR) прогнозируется на уровне 12,8% [8].

Прямая монетизация данных может реализовываться через следующие бизнес-модели:

- ***Data Marketplace*** *– создание рыночных площадок для купли-продажи информации (AWS Data Exchange – облачный сервис от Amazon Web Services, позволяющий легко находить, подписываться на сторонние наборы сведений и использовать их в облаке, или Bloomberg Data License – корпоративная подписка от компании Bloomberg на массивы финансовых материалов для автоматической загрузки во внутренние системы и приложения организации);*
- ***Data as a Service (DaaS)*** *– облачная модель управления и дистрибуции данных, предоставление их по подписке через API (Bloomberg Terminal, финансовые материалы), Nielsen Holding (показатели маркетинговой компании о потребительском поведении);*
- ***продажа наборов данных*** *– разовая передача прав на них для исследовательских, маркетинговых и отраслевых целей;*
- ***лицензирование данных*** *– предоставление прав на них на определенных условиях (по территории, сроку, объему).*

При этом модели прямой монетизации не были самоцелью и появились для получения дополнительной прибыли.

Однако гораздо большую экономическую ценность формирует косвенная монетизация. Например, персонализация продуктов и услуг использует отчеты о поведении пользователей для адаптации предло-

жений. Компания Netflix, по собственным оценкам, экономит более 1 млрд долл. ежегодно за счет рекомендательной системы, основанной на информации о предпочтениях потребителей. Монетизация данных через продажу рекламных мест на основе этого лежит в основе таргетированной рекламы, на чем построены бизнес-модели Google, YouTube, Яндекс.

Обмен данными в дата-экосистемах

Дата-экосистема – это систематизированная среда взаимодействия множества участников (организаций, государственных структур, исследовательских институтов, граждан), в рамках которой осуществляется обмен информацией для создания совместной ценности. В отличие от традиционных рыночных отношений, дата-экосистемы характеризуются:

- *сетевой структурой (множественностью связей между участниками);*
- *синергетическим эффектом от объединения данных;*
- *стандартизацией (использованием единых протоколов и форматов обмена);*
- *доверием (наличием механизмов безопасности и конфиденциальности).*

Примеры дата-экосистем: европейская облачная инициатива Gaia-X по образованию единой децентрализованной облачной системы в целях снижения технологической зависимости от американских облачных провайдеров; российское облачное платформенное решение для федеральных и региональных органов власти «ГосТех»; отраслевые экосистемы в здравоохранении, построенные на базе международного стандарта HL7 FHIR для обмена медицинской информацией.

Эффективный обмен данными требует их унификации с помощью ключевых стандартов и инициатив:

- ***FAIR-принципов (Findability, Accessibility, Interoperability, Reusability)*** *– наборов правил обеспечения доступности и пригодности сведений для их повторного использования, разработанных научным сообществом в 2014 г., одобренных на саммите G20 в Ханчжоу в 2016 г.;*
- ***HL7 FHIR (Fast Healthcare Interoperability Resources)*** *– протокола обмена медицинскими данными, основанного на REST API и форматах JSON/XML, уже стандарта «де-факто» в здравоохранении;*
- ***OpenAPI Specification (OAS)*** *– стандарта для описания RESTful API в виде единого машино- и человекочитаемого файла в формате JSON, обусловливающего интероперабельность между информационными системами;*
- ***ISO 8000*** *– серии международных стандартов, определяющих требования к качеству данных.*

Вторичное использование данных (secondary data use)

Эта концепция предполагает применение информации, собранной для одних целей, для достижения иных, первоначально не запланированных задач и приобретает особую значимость в условиях, когда основная экономическая ценность сведений раскрывается именно при их нетривиальной реализации в следующих областях:

- ***научные исследования*** *– клинические данные, полученные в ходе диагностики заболеваний и лечения пациентов, могут быть полезны для ученых, разрабатывающих новые протоколы терапии, для предиктивной аналитики в прогнозировании эпидемий или при создании лекарственных препаратов;*
- ***государственное управление*** *– органы власти генерируют огромные объемы информации в процессе оказания услуг, администрирования, статистического учета, а вторичное их использование позволяет повышать качество обслуживания, оптимизировать бюджетные расходы.*

Для этого данные должны быть доступны – открыты благодаря стандартизированным протоколам, а их поиск должен обеспечиваться через метаданные и идентификаторы, а также интероперабельны – совместимы между собой.

FAIR-принципы позволяют организациям раскрыть потенциал темных данных и сформировать основу для их рециклинга.

Одновременно развивается новый тренд – генерация с помощью ИИ синтетической информации. Для этих целей искусственные наборы данных создаются на базе реальных сведений организации. При этом они математически идентичны, сохраняют все необходимые статистические закономерности, но не содержат персональной информации. Такой подход помогает решить проблему нехватки качественных материалов для искусственного интеллекта с одновременным выполнением требований регуляторов в отношении конфиденциальности. Такие данные активно используются не только для обучения собственных моделей ИИ, но и продаются на маркетплейсах без риска нарушения приватности.

Вызовы и перспективы

Совершенствование процесса управления сопровождается формированием сложной законодательной системы, отражающей баланс между коммерциализацией данных и защитой прав их субъектов. Современное регулирование стремится к равновесию между тремя группами интересов:

- ***субъектов данных*** *– право на конфиденциальность, контроль над своими сведениями, информированное согласие;*
- ***бизнеса*** *– возможность монетизации данных, инноваций, создания новых продуктов;*
- ***общественных*** *– развитие науки, здравоохранения, государственного управления.*

Вместе с тем нормативно-правовая база еще только начала создаваться, что влечет многочисленные дисбалансы интересов.

Результаты этого с одной стороны – рост числа кибератак и утечек информации, что чревато рисками для организаций и граждан (по сведениям IBM, убытки от этого в 2023 г. составили 4,45 млн долл.), а с другой стороны – пустую нишу заполняют данные низкого качества (по оценкам Gartner, они обходятся организациям в среднем в 12,9 млн долл. ежегодно).

Проблема усугубляется ростом объемов информации и усложнением источников. По оценкам Gartner, к 2027 г. более 70% всего цифрового контента в сети Интернет будет синтетическим, что делает доказательство истинности столь же важным, как и выявление ложности. Подрывают доверие к цифровым технологиям и подвергают организации новым рискам гиперреалистичные диппфейки. Их обнаружение формирует новую дисциплину кибербезопасности, способную распознавать, анализировать и проводить криминалистическую экспертизу сфабрикованных, манипулированных, созданных или обработанных с помощью ИИ мультимедийных материалов, включая изображения, аудио, видео и интерактивные видео в реальном времени, когда основной целью выступает обман или дезинформация. Эта растущая угроза стимулирует спрос на технологии, которые смогут проверить подлинность и снизить риски, связанные с синтетическими и ложными сведениями.

В ближайшие годы можно прогнозировать перспективы управления данными, которые будут определяться следующими трендами:

- *развитие AI/ML-технологий увеличит спрос на качественные сведения и одновременно будут появляться новые возможности для их обработки;*
- *генеративный ИИ потребует огромных объемов материалов для обучения, что актуализирует вопросы лицензирования и этики;*
- *совершенствование технологий обучения ML-моделей на распределенных данных без их централизации решит проблемы конфиденциальности и безопасности;*
- *сформируются национальные и региональные подходы к управлению, отражающие культурные и правовые особенности.*

***

Данные прошли фундаментальную эволюцию от пассивных записей до стратегического актива, определяющего конкурентоспособность организаций в цифровой экономике. Признание этого требует новых методов оценки, учета и управления big data. Традиционные бухгалтерские подходы оказываются недостаточными, что стимулирует другие концепции и стандарты на международном уровне.

Капитализация осуществляется через разнообразные модели: от прямой продажи данных до косвенной монетизации через персонализацию и оптимизацию бизнес-процессов. При этом вторая, как правило, создает большую ценность, чем первая.

Вторичное использование сведений представляет собой значительный нереализованный потенциал, особенно в таких областях, как здравоохранение, наука и государственное управление. Это потребует применения FAIR-принципов и развития законодательной базы.

Нормативно-правовое регулирование активно совершенствуется, ключевая задача при этом – достичь баланса интересов между коммерциализацией данных, защитой прав граждан и общества.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Цитаты известных личностей // https://ru.citaty.net/avtory/niklaus-virt.
2. Закон Мура: что это и когда перестанет работать //https://trends.rbc.ru/trends/futurology/652569579a794753ae9c5d60?from=copy.
3. Данные: новая нефть или обыкновенный хайп? // https://skillbox.ru/media/code/dannye_eto_novaya_neft_ili_obyknovennyy_khayp.
4. DAMA-DMBOK: свод знаний по управлению данными // https://dokumen.pub/dama-dmbok-2nbsped-9785969304048-g-3192039.html.
5. Что такое «темные данные»? // https://www.ibm.com/think/topics/dark-data.
6. «Формула всего»: универсальная теорема Байеса для объяснения реальности // https://econs.online/articles/bookshelf/formula-vsego-universalnaya-teorema-bayesa-dlya-obyasneniya-realnosti.
7. Монетизация данных // https://www.decosystems.ru/monetizacziya-dannykh.
8. Большие данные (Big Data) мировой рынок // https://www.tadviser.ru/index.php.