23 января 2025

Как обеспечить достоверность данных: факторы и инструменты

Ведущие эксперты ИТ-рынка — о том, какие факторы влияют на достоверность данных и какие инструменты нужны, чтобы сделать данные достоверными.

Чтобы бизнес доверял предоставляемым данным и опирался на них в принятии решений, необходимо, чтобы эти данные были, по крайней мере, достоверными. О том, что представляет собой эта характеристика качества данных, как обеспечить их достоверность в условиях имеющихся ИТ-ландшафтов и оргструктур и об инструментах, которые для этого потребуются, рассказывают опрошенные нами эксперты. Развернутый разговор на эту тему пройдет на конференции «Качество данных — 2025».

КЛЮЧЕВЫЕ ФАКТОРЫ ДОСТОВЕРНОСТИ ДАННЫХ

На достоверность данных влияют не только технологические, но и организационные факторы — об этом заявили очень многие эксперты. Более того, организационные аспекты критически важны.

«Обеспечение достоверности данных является одной из составляющих комплексного процесса управления качеством данных, — утверждает Виктор Мастеров, директор департамента НСИ и интеграции компании SOFROS. — И хотя в DAMA DMBOK2 термин “достоверность” не используется, он фигурирует в других научных дисциплинах, связанных с управлением данными. Например, согласно одному из определений, используемых в информатике, “достоверность информации определяется ее свойством отражать реально существующие объекты с необходимой точностью. Измеряется достоверность информации доверительной вероятностью необходимой точности, то есть вероятностью того, что отображаемое информацией значение параметра отличается от истинного значения этого параметра в пределах необходимой точности”. Что касается внедрения, развития и поддержания систем (в DAMA DMBOK2 используется термин “программ”) управления качеством данных, то приоритет организационно-алгоритмических факторов и решений вряд ли подлежит сомнению. При этом именно достоверность, как параметр качества данных, требует повышенного внимания к предметной области (домену данных)».

Алиса Школьникова, руководитель направления Data Governance департамента аналитических решений ГК «КОРУС Консалтинг», считает, что **компании следует в первую очередь уточнить свое понимание достоверности данных.**

Важно понять, в каких именно точках жизненного цикла данных будет измеряться их достоверность. Исходя из этого, выбираются целевая архитектура и инструменты, необходимые для обеспечения качества данных и их автоматического исправления. Главные среди нетехнологических факторов — люди и процессы: нужно определить ответственных за формирование требований к достоверности данных и тех, кто будет разбирать возникающие ошибки и минимизировать проблемы, связанные с качеством. Эти же ответственные должны корректировать меняющиеся со временем требования к качеству данных.

Алиса Школьникова,
Руководитель направления Data Governance департамента аналитических решений ГК «КОРУС Консалтинг»

«Обеспечение достоверности данных — ключевая задача при работе с информационными системами, — уверен Виталий Миронов, генеральный директор компании «Современные бизнес-аналитические решения». — Главное правило: система полезна только тогда, когда данные заслуживают доверия. Это доверие строится на прозрачных процессах, четких стандартах и оперативной реакции на проблемы». Среди технологических факторов, влияющих на достоверность данных, Миронов видит решения для сбора и обработки данных, инструменты управления качеством данных и средства

информационной безопасности. Среди организационных — квалификацию и обучение персонала, организационные процессы и корпоративную культуру, а также собственно управление качеством данных, включая контроль их качества на основе утвержденных стандартов и метрик.

Николай Скворцов, руководитель направления методологии компании «1С: Логика Данных», ссылаясь на вышедшую в 2022 году книгу выпускающего редактора DAMA DMBOK2 Лауры Себастьян-Коулман (Laura Sebastian-Coleman) «Ответы на вызовы в области управления качеством данных» («Meeting the Challenges of Data Quality Management»), обращает внимание на **пять основных аспектов управления качеством данных**.

Во-первых, организация должна хорошо знать и понимать свои данные, требования к ним и их взаимосвязи. Во-вторых, необходимо понимать, как на состояние данных, могут повлиять процессы, с помощью которых они создаются, а также способы их использования. В-третьих, нужно представлять, как выбор технологий повлияет на создание, доступность, использование и качество данных. В-четвертых, поставщики и потребители данных должны обладать знаниями, информацией и навыками, необходимыми им для доступа к данным, их понимания и интерпретации. Наконец, в-пятых, нужно формировать корпоративную культуру в части контроля данных внутри организации и подотчетности за данные в рамках их жизненного цикла. В качестве шестого, дополнительного аспекта в книге Коулман приводится наличие согласованного корпоративного словаря в области качества данных — без него может возникнуть путаница даже в простейших понятиях.

«Достоверность данных определяется как степень соответствия данных требованиям пользователя, — продолжает Олег Гиацинтов, технический директор DIS Group. — Чтобы ее достичь, необходимо понимать требования каждого пользователя к достоверности, определить общие подходы к оценке и обеспечению качества данных, закрепить в регламентах требования к проверкам, назначить ответственных за качество и применять технические решения, позволяющие оценивать и улучшать качество данных согласно требованиям пользователей».

Как отмечает Полина Сорокина, консультант практики «Стратегия данных и управление данными» компании Axenix, достоверность, по сути, означает степень соответствия данных происходящим в реальности событиям и позволяет судить, насколько собранные данные заслуживают доверия. Впрочем, добавляет она, достоверность данных весьма относительна: «Об одной и той же записи в таблице два разных специалиста могут вынести противоположные суждения. Поэтому в крупных организациях, чтобы избежать разночтений, назначают владельцев данных — компетентных специалистов, способных корректно сформулировать критерии качества данных, в том числе их достоверности. Проблемы с качеством могут возникать на всех этапах их жизненного цикла данных, следовательно, нужны проверки качества на всем пути “следования” данных — от их ввода до загрузки в конечные аналитические продукты. Важно выстроить процесс управления качеством данных».

Со своей коллегой соглашается Наталья Кудрявцева, функциональный архитектор компании Navicon: «Обеспечение достоверности данных — задача, решаемая на протяжении всего их жизненного цикла. Ключевым моментом является степень доверия к источнику данных. Кроме того, существенную роль играют конкретные способы получения информации: например, данные, записанные на слух, более подвержены ошибкам, чем полученные путем сканирования документов. На качество данных влияет также квалификация и мотивация специалистов, занимающихся их вводом, удобство применяемого ими интерфейса и наличие контроля вводимых данных. Важно, чтобы пользователь мог быстро и легко находить информацию и использовать ее повторно вместо того, чтобы вводить заново».

Андрей Бойко, коммерческий директор B2B-Center, уточняет: «Чтобы данные были достоверными, они должны быть востребованными, точными, согласованными, своевременными, доступными и интерпретируемыми. Среди технологических факторов, влияющих на эти характеристики, важны правильно спроектированные архитектура и СУБД, обеспечивающие очистку и автоматическое тестирование целостности данных еще на этапе их создания. Также существенное значение имеют прозрачность источников и пригодность данных для реализации конкретных задач, стандартизация моделей данных и низкое количество противоречий в структуре и семантике. Среди важных нетехнологических факторов — стандартизация данных, их валидация, определение метрик качества и его непрерывный мониторинг».

Ирина Мягкова, заместитель директора по развитию компании «РЕЛЭКС», также рекомендует обратить внимание на **СУБД**, поскольку они **играют**

важную роль в обеспечении безопасности данных: «Как правило, для этих целей применяются СУБД с закрытым кодом, имеющие сертификаты ФСТЭК или Минобороны, такие, например, как отечественная СУБД ЛИНТЕР БАСТИОН».

Михаил Рощин, заместитель директора отделения департамента НСИ и интеграционных сервисов IBS, выделяет **три основных аспекта обеспечения достоверности данных**: «Во-первых, необходимо определить наборы правил для контроля качества данных и их обновления, а также метрики для отслеживания и регламентирования процессов. Во-вторых, нужны технологии — программные средства для осуществления проверок. Наконец, необходима команда специалистов для регулярного отслеживания инцидентов с качеством данных и их устранения».

Арсен Кудзиев, руководитель отдела аналитического департамента компании PARMA TG, первым из значимых факторов считает **наличие специалистов, ответственных за внедрение политик управления данными, и предоставление им необходимых ресурсов и полномочий**: «Компании, где существует позиция CDO или аналогичная должность, уже сделали главный шаг на пути обеспечения достоверности данных». Следующий по значимости фактор — последовательное внедрение практик работы с данными: определение ключевых доменов данных, разработка процессов и регламентов работы с данными, управление сопутствующими рисками, определение ролей и полномочий сотрудников, отвечающих за отдельные домены и их сегменты. Технологии являются обеспечивающим инструментом для повышения достоверности данных».

Сергей Полехин, владелец продукта PIX BI, также уверен, что **на достоверность данных влияют, прежде всего, люди, которые за них отвечают**. «Поэтому за каждым набором данных должен быть закреплен ответственный, который будет знать, почему данные именно такие по составу и форме, откуда они взяты, как часто обновляются, и сможет, исходя из четко поставленной функциональной задачи, правильно оценить уровень их достоверности, — дополняет свой ответ Полехин. — Среди технологических факторов на достоверность данных влияют методы их сбора, хранения и обновления. Для подтверждения достоверности важно проверять не только конкретный набор данных, но и возможность его корректного сопоставления с данными из других источников».

«Обеспечить достоверность данных можно только комплексно: важно разработать четкий регламент актуализации данных, регулярно проверять источники, внедрять эффективные инструменты управления данными, которые легко встраиваются в бизнес-процессы и способствуют работе, а не мешают ей, — подчеркивает Григорий Бокштейн, ведущий эксперт по управлению данными компании TData. — Именно такой подход, заложенный в нашу платформу управления данными, позволил “Ростелекому”, одному из наших ключевых клиентов, повысить уровень достоверности данных, попадающих в хранилище и конечные отчеты».

Андрей Андриченко, директор по развитию компании «ЭсДиАй Солюшен», особо отмечает **важность разработки единых стандартов описания форматов представления данных и обмена ими**: «Стандартизировать

термины и определения, необходимые для создания библиотеки шаблонов, содержащих наборы характеристик для различных групп однородных информационных объектов, позволяет описанная в стандарте ГОСТ ИСО 22745 методология открытых технических словарей (Open Technical Dictionary, OTD), реализованная в среде MDM: сначала должна быть унифицирована терминология предметной области, после чего на основе единого глоссария терминов можно создавать шаблоны, принадлежащие различным уровням классификационной иерархии каталогов и нормативно-технических справочников. В атрибуты шаблонов необходимо включить информацию о допустимых значениях. Библиотека стандартизованных шаблонов обеспечивает возможность обмениваться качественной информацией независимо от специфики программной реализации приложений».

По мнению Александра Учаева, менеджера по продукту «1С:MDM» компании «1С», **нетехнологические факторы все активнее переходят в разряд технологических**: «Так, автоматизация изначально была направлена на устранение человеческих ошибок. Технологические же факторы разнятся, исходя из типа самих данных: если для основных и справочных наиболее важным является проверка соответствия классификаторам и каталогам из доверенных источников, то для транзакционных данных на первый план выходит комплексная взаимная верификация с использованием сервисов крупных вендоров».

Иван Вахмянин, управляющий партнер Visiology, обращает внимание на то, что д**анные всегда неточны — они отражают реальность, и искажения в отражении неизбежны:** «Следовательно, абстрактная достоверность данных

в отрыве от конкретной бизнес-задачи не имеет смысла. Поэтому самое важное в работе с данными — четко формулировать бизнес-задачи, определять требования к данным и при этом ясно понимать, кто их будет использовать и для принятия каких именно решений. Опыт успешных проектов показывает, что на постановку задачи, работу с бизнес-заказчиком и проработку требований должны быть направлены не менее 20-30% трудозатрат аналитиков».

ИНСТРУМЕНТЫ ДЛЯ ОБЕСПЕЧЕНИЯ ДОСТОВЕРНОСТИ ДАННЫХ

Переходя к обсуждению инструментов, эксперты отметили важность не только технологических средств, но также процессов и оргструктур.

Инструменты не исчерпываются ИТ-продуктами. Это еще и набор правил, которые должны быть сформулированы и регламентированы, и люди, ответственные за работы по обеспечению достоверности данных. Среди технологических инструментов самый, пожалуй, важный — источники данных. И чем больше предъявляется требований к источникам, тем проще поддерживать чистоту данных. Если по каким-то причинам обеспечивать достоверность данных в источниках невозможно, необходимо позаботиться о достоверности данных в КХД и BI-системах.

Алиса Школьникова,
Руководитель направления Data Governance департамента аналитических решений ГК «КОРУС Консалтинг»

Сорокина видит **ключ к обеспечению достоверности данных в сочетании выстроенных процессов управления качеством данных и рационального применения программных средств для их проверки**: «Для сбора проверок, мониторинга их исполнения и устранения инцидентов в едином информационном пространстве существуют специальные инструменты. Наиболее передовые из них позволяют быстро оценивать качество данных найденных в каталоге таблиц или отчетов, не требуя дополнительных переходов и переключения между приложениями».

По мнению Бокштейна, **самым простым ответом на вопрос об инструментах для обеспечения достоверности данных было бы классическое перечисление продуктов, однако не все так просто**: «Рассмотрим пример мобильного оператора. Data Governance позволяет продемонстрировать бизнесу, какие данные у него есть (например, абонентская база, количество и расположение вышек и т.д.), и связать их между собой. Master Data Management позволяет унифицировать НСИ. Средства Data Quality — проверить качество данных. Однако одного лишь внедрения этих инструментов недостаточно, очень важно, чтобы у бизнеса были налажены внутренние бизнес-процессы. Здесь важна работающая связка “инструмент–методология”. Поэтому более корректно будет утверждать, что необходим приведенный выше набор инструментов, но с условием, что эти инструменты должны бесшовно встраиваться в текущие бизнес-процессы и поддерживать их развитие, чтобы эти процессы не пришлось менять ради выстраивания работы с данными. Именно такой подход позволил легко внедрить наши продукты в крупнейших компаниях различных отраслей и реализовать проекты, которые были удостоены престижных наград».

Скворцов делает **особый акцент на создании единого информационного пространства в области качества данных**: «Оно представляет собой не столько инфраструктурную, сколько социотехническую систему: в рамках пространства очень важно обеспечить эффективное взаимодействие разнородных рабочих групп, участвующих в создании, поставке, обработке и использовании данных, и здесь помогут подходы, заимствованные из социологии. В качестве важного типа инструментов для создания единого

информационного пространства следует особо выделить так называемые пограничные объекты (boundary objects), обеспечивающие коммуникацию между группами. Они предоставляют информацию о том или ином аспекте работы с данными, которая может быть немного по-разному интерпретирована внутри каждой группы в зависимости от направления ее деятельности, но общее понимание ключевых моментов остается неизменным. К пограничным объектам можно отнести любые артефакты, способствующие устранению разницы в представлениях различных рабочих групп о корпоративных данных, требованиях к ним и их взаимосвязи: глоссарии, модели данных, диаграммы SIPOC (Suppliers, Inputs, Processes, Outputs, Consumers — “поставщики, входы, процессы, результаты, потребители”) и т.п.»

По наблюдениям Кудзиева, **на выбор инструментов влияет множество факторов: конкретная бизнес-задача, ИТ-инфраструктура, количество преобразований данных, особенности их агрегации** и пр.: «Перед внедрением технологических инструментов нужно провести предпроектный аудит, по итогам которого сформировать список необходимых компонентов. Базовый инструментарий обычно включает в себя онтологии и глоссарии данных, MDM-средства, решения для отслеживания происхождения данных и инструменты управления качеством данных. В условиях импортозамещения крупные организации достаточно успешно применяют решения на базе открытого ПО: DataHub, OpenMetaData, DBT + Great Expectations и др.».

Мастеров считает, что **говорить о каком-то универсальном и фиксированном наборе инструментов было бы методологически**

необоснованно: «Согласно базовому принципу, сформулированному в DAMA DMBOK2, “инструменты следует выбирать с учетом системной архитектуры и планируемых настроек еще на фазе планирования программы качества данных предприятия”. Применительно к достоверности наиболее важное значение имеют инструменты формирования запросов к данным, шаблоны правил качества данных и репозитории метаданных».

Согласно замечанию Гиацинтова, **инструменты для обеспечения достоверности данных следует выбирать в зависимости от метрик качества**: «В первую очередь понадобятся инструменты класса Data Quality, имеющие в своем составе внушительный арсенал возможностей для оценки и анализа качества данных по основным бизнес- и техническим характеристикам и приведения данных в порядок. Для сложных метрик качества, таких, как контролируемость и согласованность данных, применяются решения Data Governance, позволяющие выявлять суть и логику трансформации данных. Для простейших проверок обычно используются стандартные средства интеграции данных — ETL/ELT и им подобные».

С точки зрения Вахмянина, **необходимым условием для обеспечения достоверности данных является формирование «единой точки правды»**: «Для этого можно создать общее КХД или использовать встроенное в BI-платформу хранилище, такое как СУБД ClickHouse под управлением механизма “ДанКо”. Также требуются механизмы ETL, обеспечивающие не только загрузку, но и проверку и сопоставление данных. В задачах проверки достоверности данных большую пользу приносит искусственный интеллект

— эти механизмы будут развиваться в ближайших релизах Visiology».

Андриченко также высоко оценивает перспективы ИИ в управлении данными: «Его применение в может значительно повысить достоверность данных. По мнению клиентов, подсистема машинного обучения, созданная специалистами нашей компании и интегрированная в программный комплекс Semantic MDM, позволяет примерно в три раза сократить время обработки заявок на включение новых позиций в MDM-систему и при этом существенно снизить количество ошибок».

Миронов помещает в свой портфель **программных и методических инструментов средства для верификации и очистки данных и автоматизации проверки качества данных**, в том числе платформы для оценки надежности источников информации (например, FactCheck.org) и системы контроля происхождения и целостности изображений (такие как Exif Viewer, FotoForensics, TinEye, Google Reverse Image Search). Кроме того, он рекомендует выстроить проведение регулярных аудитов и проверок качества данных, обучение персонала методам работы с данными и их верификации, а также внедрение стандартов, процедур и инструкций, обеспечивающих последовательность и надежность процессов обработки и проверки данных.

Полехин перечисляет **три вида инструментов, необходимых для обеспечения достоверности данных**: средства проверки данных на корректность, решения для устранения в них шумов и информации, не нужной для решения конкретных бизнес-задач, а также инструменты для сопоставления данных, позволяющие понять, насколько они соотносятся

между собой и пригодны ли для совместного использования.

Бойко рекомендует **сконцентрировать внимание на системах мониторинга, очистки и валидации данных, а также платформах MDM**: «Крайне важен последовательный и комплексный подход, включающий профилирование данных, определение правил и метрик, документирование стандартов в области качества данных и их внедрение».

Игорь Моисеев, директор по развитию DataCatalog (входит в группу Arenadata), также выступает за **комплексный подход, охватывающий различные аспекты управления информацией**: «В первую очередь необходимы средства управления качеством данных, инструменты профилирования данных и каталоги метаданных. Эти системы востребованы в инфраструктуре каждой организации, хранящей и обрабатывающей информацию».

Кудрявцева предлагает использовать инструментальный набор, включающий механизмы поиска (в том числе с частичным совпадением вводимых данных и определением ограничений, предотвращающих дублирование), средства ввода данных и их автоматизированного контроля, а также механизмы интеграции для проверки вводимых данных на соответствие достоверным источникам.

Рощин особо выделяет **«коробочные» решения для проверки и контроля качества данных** — такие, как «Плюс7 ФормИТ DQ», рекомендуемое для импортозамещения продукта Informatica Data Quality, и инструменты для

реализации правил контроля качества данных на базе существующих интеграционных средств и создания отчетов с метриками качества данных.

Очень конкретно определяет набор средств обеспечения достоверности данных Учаев. Помимо специализированных продуктов класса MDM, таких как «1С:MDM Управление мастер-данными КОРП» и «1С:MDM Управление нормативно-справочной информацией», он включает в инструментальный портфель **отраслевые продукты, содержащие готовые интерфейсы для обмена со специализированными финансовыми сервисами, кредитными бюро, государственными реестрами и скоринговыми системами, а также готовые специализированные сервисы верификации и стандартные библиотеки «1С»**, позволяющие разработчикам быстро добавлять в создаваемые продукты функции обеспечения достоверности данных.

Мягкова, в свою очередь, предлагает учитывать важность СУБД как средства обеспечения безопасности хранения и передачи информации, а также аудита вносимых изменений. В частности, СУБД ЛИНТЕР БАСТИОН уже более 20 лет применяется в системах, предъявляющих высокие требования к уровню защиты информации, поскольку в них обрабатывается информация, содержащая сведения высокой степени секретности.

Как видим, для обеспечения достоверности данных требуется комплексный подход, охватывающий не только технологические, но также методологические и организационные инструменты.