

Что такое интеграция данных

В этой главе рассказывается о том, что такое интеграция данных, какую роль она играет в общем жизненном цикле данных и какое отношение имеет к стратегии данных организации. Цель главы — снабдить вас базовыми знаниями, необходимыми для эффективной реализации того или иного решения интеграции данных и приведения этого решения в соответствие с задачами, которые ставит перед собой ваша организация.

Интеграция данных и управление ими

Управление жизненным циклом данных охватывает все направления деятельности, связанные с получением и поддержанием ценности информации из данных. Эффективное управление данными должно гарантировать их точность, доступность и пригодность для дальнейшего использования и выступать главным компонентом в процессе принятия решений. Возможно, вам встречался термин *DataOps*, который представляет собой стиль управления данными, сфокусированный на взаимодействии между заинтересованными сторонами в течение всего жизненного цикла данных. Это во многом похоже на то, как в *DevOps* строится взаимодействие между командами, участвующими в разработке программного обеспечения (ПО). Подход *DataOps* акцентируется на автоматизации, поддержании качества и непрерывной доставке в процессах обработки данных, подобно тому как это делает *DevOps* в разработке ПО.

Я предпочитаю разделять управление жизненным циклом данных на три части: *интеграцию данных*, *аналитику данных* и *управление данными* (рис. 1.1). Каждое из этих направлений имеет собственные цели и состоит из процессов более низкого уровня, которые, объединяясь, формируют конвейеры

данных¹. Процессы более низкого уровня могут относиться сразу к нескольким компонентам.

Аналитика и управление не менее важны в общем жизненном цикле данных, чем интеграция. Начнем с краткого описания аналитики данных и управления ими.



Рис. 1.1. Компоненты управления жизненным циклом данных

Аналитика данных охватывает весь спектр задач, которые, как ожидается, должны выполнять специалисты по данным (дата-сайентисты) или аналитики данных — от создания визуализаций данных² до разработки моделей *машинного обучения* (machine learning, ML). Аналитику можно рассматривать как фронтенд-уровень управления данными. Именно эта составляющая управления жизненным циклом данных обычно наиболее хорошо знакома конечным пользователям и лицам, принимающим решения.



Аналитика данных (и, в частности, визуализация данных) играет примерно ту же роль для управления жизненным циклом данных, что и браузер для Интернета.

С другой стороны, управление данными относится к политикам, стандартам и практикам, гарантирующим, что данные обрабатываются правильно в течение всего своего жизненного цикла и в то же время соответствуют целям организации и стратегиям управления рисками. Сюда относятся все активности, связанные с обеспечением целостности данных, конфиденциальностью, соблюдением нормативных требований и сохранением

¹ См. обсуждение конвейеров данных в подразделе «Конвейеры» в главе 2.

² В общем случае я считаю визуализацию и бизнес-аналитику (business intelligence, BI) составными элементами аналитики данных, но иногда их объединяют с компонентами интеграции данных.

контроля над управлением данными. Например, Министерство обороны США (US Department of Defense, DOD) опубликовало семь целей управления данными, которых им необходимо достичь, чтобы стать организацией, ориентированной на данные. Согласно этим принципам, данные должны быть видимыми (visible), доступными (accessible), понятными (understandable), взаимосвязанными (linked), достоверными (trustworthy), совместимыми (interoperable) и защищенными (secure). Совокупность этих требований известна как аббревиатура VAULTIS, и, хотя они не ссылаются на конкретные законы или правила, они спроектированы так, чтобы в прямом и переносном смысле представлять Министерство обороны США.



В терминологии данных понятие *жизненный цикл* не обязательно подразумевает именно цикл.

Определение интеграции данных

Цифровые данные могут существовать в *облаке* (то есть в удаленных центрах обработки данных), в локальных вычислительных средах, на внешних жестких дисках, на флеш-накопителях и периферийных устройствах, которыми могут являться активные и пассивные датчики, смартфоны и другие системы, взаимодействующие с реальным миром. Интеграция данных — это процесс унификации и упорядочивания данных вместе с обслуживающей их инфраструктурой. Этот процесс включает активности, связанные с обнаружением, профилированием, сбором, консолидацией, объединением, очисткой, сортировкой, перемещением, миграцией, репликацией, маскированием, фильтрацией, сопоставлением, синхронизацией и автоматизацией данных¹.

Решения для интеграции данных, как правило, создаются с использованием набора программных инструментов, включая, например, средства управления качеством данных и механизмы управления данными. Наиболее эффективные инструменты интеграции также могут автоматизировать некоторые необходимые процессы, а также помогать идентифицировать и подключать источники данных, сопоставлять элементы данных между исходной и целевой системами и реализовывать правила трансформации, чтобы обеспечить согласованность и точность данных. Мы рассмотрим некоторые такие инструменты в части II.

Обычный процесс интеграции данных состоит из нескольких ключевых этапов, которые отвечают за то, чтобы данные из различных источников

¹ Это не исчерпывающий, равно как и не взаимоисключающий список.

были тщательно и эффективно скомбинированы для последующего анализа и формирования отчетов. Мы подробно разберем многие из этих этапов в следующих главах, а пока достаточно знать, что все начинается с определения источников данных, таких как базы данных (БД), электронные таблицы, облачные сервисы и программные интерфейсы приложений (application programming interface, API). Потом выполняется извлечение данных, цель которого — забрать данные из идентифицированных источников, часто с использованием разнообразных инструментов и методов. Затем происходит сопоставление данных, при котором создаются *схемы*, обеспечивающие соответствие элементов данных из различных систем и согласованность названий полей и структур.

Для выявления ошибок и несоответствий проводятся валидация данных и контроль качества. На этапе преобразования извлеченные данные приводятся к единому формату с помощью процессов кондиционирования, обогащения и нормализации. После преобразования выполняется этап загрузки, где конвертированные данные передаются в систему хранения (data store) либо используются сразу для дальнейшей обработки или анализа как в пакетном, так и в потоковом режиме.

Для поддержания интегрированного потока данных применяется синхронизация данных, которая позволяет периодически или почти в реальном времени¹ обновлять набор данных по мере поступления новых данных. При этом особую важность приобретает соблюдение нормативных требований, а также обеспечение целостности и безопасности данных, особенно при работе с чувствительной информацией.

Легкость обнаружения и удобство использования интегрированных данных повышаются благодаря метаданным, которые предоставляют контекстную информацию об их происхождении и значении. Интегрированные наборы данных становятся доступными и пригодными для анализа с помощью инструментов бизнес-аналитики (business intelligence, BI) и аналитических платформ, которые позволяют получать полезную информацию и извлекать идеи, служащие основой для принятия решений и формирования бизнес-стратегии организации.

Таким образом, интеграция данных сочетает технические процессы, инструменты и стратегии для обеспечения согласованности, точности и доступности данных из потенциально разнородных источников для дальнейшего анализа и принятия решений. Мы рассмотрим подробнее эти процессы в следующих главах.

¹ Правила обсуждаются в главах 3 и 5.

Почему интеграция данных важна

Основными факторами, которые приводят к необходимости поиска решений для интеграции данных, являются:

- значительный рост объемов данных;
- растущее разнообразие типов данных;
- рост ожиданий обработки в почти реальном времени¹;
- все более широкое использование облачных сервисов;
- изменения в нормативно-правовой среде;
- непрерывное развитие технологий.



Согласно оценкам Precedence Research, «мировой рынок интеграции данных в 2023 году оценивался в 13,6 млрд долларов США и, как ожидается, достигнет 43,38 млрд долларов США к 2033 году, со среднегодовым темпом роста 12,32 %». Я не знаю, какой процент от этих сумм относится к государственным закупкам, однако, судя по тому, что я наблюдаю со своей текущей позиции (в качестве руководителя направления анализа и инженерии данных), количество запросов от государственных агентств (запросы информации (RFI), предложений (RFP) и т. д.) по тематике интеграции данных продолжает расти относительно общего числа поступающих запросов. Учитывая текущую сумасшедшую популярность генеративного искусственного интеллекта (GenAI), который опирается на огромные обучающие наборы данных, можно предположить, что необходимость в эффективных решениях по интеграции данных в ближайшем будущем не исчезнет.

Как отмечает Мартин Клеппман в книге *Designing Data-Intensive Applications*, необходимость интеграции данных «часто становится очевидной, если посмотреть на потоки данных шире — в рамках всей организации»². Подобные процессы наблюдаются и в федеральных органах власти США, так как все

¹ Я предпочитаю не использовать термин *реальное время*, потому что в действительности ничто не происходит мгновенно (за исключением, наверное, квантовой запутанности). Однако некоторые специалисты приравнивают *почти реальное время* к минутам, а *реальное время* — к секундам (или даже к более коротким интервалам). Эти понятия относительные и во многом субъективные, а также в некоторой степени зависят от сценария использования. Тем не менее я предпочитаю выражение *почти реальное время*.

² Клеппман М. Высоконагруженные приложения. Программирование, масштабирование, поддержка.

больше старших должностных лиц начинают придерживаться культуры, ориентированной на данные. И с учреждением Федерального совета главных специалистов по данным (CDO) и Главного управления цифрового и искусственного интеллекта (CDAO) Министерства обороны правительство США предпринимает согласованные усилия по внедрению ориентированного на данные подхода.

Вот некоторые из основных преимуществ эффективной интеграции данных:

- более полное использование данных;
- повышение качества данных и обеспечение их целостности;
- улучшение согласованности, точности и доверия;
- снижение количества изолированных хранилищ данных (data silos) и распространение практики совместного принятия решений;
- более масштабные и передовые идеи;
- повышение эффективности управления;
- упрощенное масштабирование инфраструктуры данных.

Эволюция интеграции данных

Основная задача интеграции данных заключается в том, чтобы обеспечить успешную консолидацию и упорядочивание данных из различных источников для поддержания процесса принятия управленческих решений и достижения операционных целей организации. Однако технологический прогресс и меняющийся ландшафт данных привели к заметным инновациям в этой области, а развитие методов интеграции данных отражает все более широкие технологические, организационные и методологические сдвиги в информационных технологиях и управлении данными.

Изначально задачи интеграции данных выполнялись исключительно с помощью программного кода, написание которого, как правило, отнимало много времени. Код мог содержать ошибки и плохо поддаваться масштабированию. Интеграции также были в основном ограничены однородной средой, где источники и приемники данных часто принадлежали одному поставщику или строились на единой технологической платформе.

В конце 1990-х стали появляться коммерческие инструменты, которые автоматизировали множество ручных процессов в интеграции данных, значительно сократив сроки реализации и снизив сложность проектов, связанных с интеграцией данных. В тот же период широкое распространение получили

хранилища данных, создание которых требовало надежных решений по интеграции данных для агрегирования и согласования данных из разных операционных систем для целей аналитики.

Стремительное развитие веб-технологий и облачных сервисов в 2000-х привело к появлению множества новых видов и источников данных, что требовало более гибких и масштабируемых решений в области интеграции данных. Программные интерфейсы приложений и веб-сервисы сыграли важную роль в обеспечении более динамичной интеграции данных между разрозненными системами и платформами почти в реальном времени.

Появление в 2010-х технологий больших данных, таких как базы данных Hadoop и NoSQL, требовало, чтобы решения интеграции данных могли масштабироваться горизонтально и обрабатывать огромные объемы данных. А спрос на аналитику с минимальной задержкой привел к появлению технологий потоковой интеграции данных, способных обрабатывать данные в движении и предоставлять аналитические результаты практически без промедления.

В последние годы все большую популярность приобретают такие концепции, как ткань данных (data fabric) и сетка данных (data mesh), делающие акцент на более децентрализованный подход к интеграции данных и управлению данными, в котором логика интеграции становится неотъемлемой частью самой информационной среды. Мы обсудим это подробнее в главе 4, а пока просто запомните, что децентрализация обычно приводит к развитию более гибкой и масштабируемой архитектуры.

В настоящее время возможности ИИ и машинного обучения внедряются в инструменты интеграции данных с поразительной скоростью. В первую очередь они используются для автоматизации комплексных задач интеграции, таких как маркировка данных, сопоставление и преобразование данных, что значительно улучшает эффективность и точность интеграции. Дополнительная гибкость, масштабируемость и экономическая эффективность стали возможны благодаря переходу к облачно-ориентированным приложениям и бессерверным вычислениям.

Кроме того, поскольку данные все чаще генерируются и обрабатываются Интернетом вещей (IoT)¹ на периферии, значительно возрастает и важность периферийной интеграции. Вместо передачи больших объемов неопределенных данных в центральные репозитории обработка и интеграция

¹ Интернет вещей включает датчики, микропроцессоры и другие устройства, такие как смартфоны и смарт-телевизоры. Он обычно ассоциируется с устройствами «на периферии».

происходят (или по крайней мере начинаются) ближе к источнику данных, что способствует уменьшению задержки на следующих этапах и смягчению ряда проблем передачи данных.

В последнее время в сфере данных и технологий стали доминировать большие языковые модели (large language model, LLM) и агентный (или автономный) ИИ (agentic AI). Эти ИИ-инструменты уже помогли ускорить, а в некоторых случаях даже полностью автоматизировать интеграцию данных. Например, Amazon Q дает своим пользователям возможность создавать процессы извлечения, преобразования, загрузки (extract, transform, load, ETL) данных с использованием обработки естественного языка (natural language processing, NLP). Современные системы ИИ уже способны достаточно точно распознавать различные форматы данных и объединять их без участия человека. Однако от перехода к полному доверию ИИ нас пока отделяют несколько существенных недостатков. Мы обсудим их в главе 3.

Эволюция интеграции данных — свидетельство непрерывно меняющегося ландшафта управления данными и постоянного стремления организаций адаптироваться к новым условиям для извлечения максимальной пользы из своих информационных ресурсов. Интеграция данных остается одной из наиболее востребованных областей и продолжает быстро эволюционировать от ручных процессов к автоматизации на основе ИИ.

Сценарии использования и углубленные исследования интеграции данных

Интеграция данных — это в определенной мере неотъемлемая часть управления данными, она применяется в различных отраслях для решения многих бизнес-задач и проблем. Рассмотрим наиболее распространенные сценарии ее использования.

Хранилища данных

Интеграция данных играет критически важную роль при построении хранилищ данных, которые служат централизованными репозиториями данных для аналитики и отчетности.

Озера данных

В средах, работающих с большими данными, методы интеграции используются для передачи структурированных, неструктурированных и полуструктурированных данных от разрозненных локальных платформ к гибкому централизованному хранилищу.

Бизнес-аналитика (business intelligence) и отчетность

Интеграция данных критически важна при подготовке комплексных ВІ-отчетов и дашбордов, которые предоставляют информацию о различных показателях эффективности бизнеса, в частности в области продаж, маркетинга, финансов и операционной деятельности.

Обработка данных Интернета вещей

Интеграция данных от IoT-устройств позволяет организациям мониторить и управлять подключенным оборудованием, анализировать данные от датчиков и зачастую автоматизировать процессы на основе информации, получаемой почти в реальном времени.

Чтобы наглядно продемонстрировать преобразующую силу интеграции данных за пределами административных барьеров, перед тем как завершить главу, рассмотрим три кратких исследования, в которые я сам был непосредственно вовлечен или о которых знаю из надежных источников¹.

Здравоохранение

В сфере здравоохранения интеграция данных из электронных медицинских карт, лабораторных систем и систем медицинской визуализации позволяет медицинским работникам формировать надежные, консолидированные отчеты об истории болезни пациента для постановки более точных диагнозов и более эффективного планирования лечения. Интегрированные системы данных могут предупреждать врачей о потенциально опасных лекарственных взаимодействиях и выявлять значимые тенденции в состоянии пациентов, что в целом улучшает качество медицинской помощи.

Общество информационных и управленческих систем в здравоохранении (Healthcare Information and Management Systems Society, HIMSS) выделяет три типа архитектуры сетей обмена медицинской информацией².

Централизованная

Данные о пациенте собираются и курируются в централизованном репозитории, таком как хранилище или озеро данных.

¹ Никакая информация, содержащаяся в этой книге, не является конфиденциальной, контролируемой или засекреченной.

² Общество информационных и управленческих систем в здравоохранении (Healthcare Information and Management Systems Society, HIMSS). Interoperability in Healthcare. Healthcare Information and Management Systems Society, Inc. (HIMSS). <https://www.himss.org/resources/interoperability-healthcare>.

Федеративная

Обмен данными осуществляется между взаимосвязанными, но независимыми базами данных.

Гибридная

Включает в себя варианты обоих типов (централизованной и федеративной), что позволяет пользоваться преимуществами обоих типов архитектур.

Налоговое администрирование

Налоговая служба США (The US Internal Revenue Service, IRS) интегрирует данные из множества источников для контроля соблюдения налогового законодательства. Комбинируя данные из налоговых деклараций, финансовых учреждений и других федеральных ведомств, IRS может выявлять несоответствия и аномалии, что помогает определять сокрытые доходы, факты мошенничества и другие нарушения, связанные с соблюдением правил и норм.

Кроме того, IRS сравнительно недавно перенесла свои данные из локальных хранилищ на базе мейнфреймов в Amazon Web Services (AWS) GovCloud¹, где данные организованы как *витрина данных (data mart)* в удобных для пользователей форматах, что упрощает выполнение запросов и формирование отчетности.

Иммиграционный и пограничный контроль

Министерство внутренней безопасности США (The US Department of Homeland Security, DHS) использует интеграцию данных для улучшения работы иммиграционного и пограничного контроля. Объединяя сведения из различных источников, например из заявлений на визу, пассажирских списков и биометрических систем, DHS может более эффективно выявлять потенциальные угрозы безопасности. Такой комплексный подход помогает предотвращать нелегальную иммиграцию, укрепляет национальную безопасность и одновременно способствует развитию законных поездок и торговли. Хорошо организованный конвейер интеграции данных, обеспечивающий обработку заявлений на неиммиграционные визы в Бюро консульских дел Государственного департамента США, взаимодействует с Национальным центром проверки (NVC) Таможенного и пограничного контроля (СВР). Благодаря этому, как считается, значительно сократилось число заявителей на визу, ожидающих административной проверки.

¹ Amazon Web Services (AWS) GovCloud — это набор регионов AWS, которые находятся как в физической, так и в логической сетевой изоляции от всех регионов AWS. Мы поговорим об AWS более подробно в части II.

Резюме

В этой главе вы познакомились с областью интеграции данных и рассмотрели ее значение в широком контексте управления данными и корпоративной стратегии работы с ними. Мы провели различие между интеграцией, аналитикой и управлением данными, а также разобрались, как они вместе создают единую экосистему. Интеграция данных, как мы выяснили, необходима для того, чтобы сводить воедино разные источники, поддерживать точность и непротиворечивость данных, а также обеспечивать оперативную обработку и принятие решений.

Сфера интеграции данных продолжает активно развиваться под влиянием технологического прогресса и меняющихся потребностей в данных. Эволюция интеграции данных от индивидуальных программных решений до современных тенденций в автоматизации на базе ИИ и облачно-ориентированных приложений отражает постоянно растущую сложность и масштаб проблем управления данными. Поскольку объемы генерируемой информации продолжают увеличиваться, потребность в надежных, масштабируемых и гибких решениях для интеграции данных почти наверняка будет расти.

В последующих главах мы подробнее разберем конкретные примеры, концепции и методологии, используемые в реализации эффективных решений интеграции данных.