

# На пути к комплексному управлению метаданными

В этой главе мы кратко рассмотрим содержание всей книги. Сначала познакомимся с областями управления метаданными в компаниях и используемыми ими репозиториями метаданных — это тема части I. Затем обсудим концепцию команды по обнаружению данных, которая способна координировать эти репозитории метаданных и тем самым улучшить поиск в масштабах всей компании, — это тема части II. Наконец, кратко пройдемся по идее метасетки — децентрализованной архитектуры для метаданных, которая является темой части III.

Готовы? Поехали!

## Управление метаданными ведется повсеместно

В этой книге области, занимающиеся управлением метаданными, делятся на четыре категории<sup>1</sup>.

### *Управление ИТ*

Эта область использует репозитории метаданных для стратегического планирования архитектуры предприятия, поддержания существующей ИТ-инфраструктуры и сохранения данных о недавнем прошлом в системах резервного копирования.

### *Управление данными*

Эта область применяет набор технологий для хранения, извлечения, преобразования данных, наблюдения за ними и их приема во всем

---

<sup>1</sup> Следует подчеркнуть, что деление на области может быть и иным. Области, перечисленные в книге, не являются каноническими.

ИТ-ландшафте. В конце 2010-х и начале 2020-х годов эта часть инструментария управления данными называлась *современным стеком данных*. Однако сейчас этот термин употребляется все реже и даже был объявлен «мертвым»<sup>1</sup>. Существует несколько карт технологий управления данными, включая репозитории метаданных. Наиболее полной и известной является *MAD (Machine Learning, Artificial Intelligence, and Data) Landscape* (<https://oreil.ly/8aAmf>) Мэтта Терка и Амана Кабира.

### *Управление информацией*

Эта область имеет меньший охват, чем управление данными. Кроме того, в ней чаще используются менее сложные с технической точки зрения решения для хранения. Помимо цифровых данных, управление информацией фокусируется на нецифровых, физических объектах, таких как бумага и образцы, а также на концепциях, которые являются скорее абстрактными, чем материальными, например на бизнес-процессах и возможностях компании. Управление информацией — это не единая область, а несколько отдельных направлений. Репозитории метаданных для управления информацией в значительной степени используются для регуляторных и операционных целей, а не для инноваций (хотя ситуация меняется с появлением ИИ, поскольку неструктурированные данные очень важны для генеративного и агентного ИИ). Существует огромное количество теоретических работ, посвященных управлению информацией<sup>2</sup>. Технологии управления информацией обычно изображаются на картах RegTech (<https://oreil.ly/T5sd9>) или являются частью карт GovTech (<https://oreil.ly/EnAcA>).

### *Управление знаниями*

Это методология сбора знаний, хранящихся в сознании людей. Поэтому Международная организация по стандартизации (International Organization for Standardization, ISO) считает ее частью управления человеческими ресурсами, и для этого создана отдельная рабочая группа ISO/TC 260/WG 6 (<https://oreil.ly/R6njw>). Непосредственная задача состоит в том, чтобы собирать накапливаемые людьми знания и обеспечивать их долгосрочное хранение. С этой целью управление знаниями предполагает их хранение с помощью соответствующих технологий, использующих измерения репозитория метаданных. Эти технологии управления знаниями позволяют распространять знания среди большего числа людей

<sup>1</sup> Reis J. Everything Ends — My Journey with the Modern Data Stack. — 17 февраля 2024 г. (<https://oreil.ly/Qa7HG>).

<sup>2</sup> Обзор литературы по управлению информацией можно найти в статье Ануджа Шарма, Припендры П. Рана и Робина Нанку *Fifty Years of Information Management Research: A Conceptual Structure Analysis Using Structural Topic Modeling* // International Journal of Information Management, 58, 2021 (<https://doi.org/10.1016/j.ijinfomgt.2021.102316>).

и обучать их. Так же как управление данными и информацией, управление знаниями использует набор технологий. Наиболее полный их обзор предоставляют карты EdTech ([https://oreil.ly/k4\\_83](https://oreil.ly/k4_83)).



Все команды во всех областях пользуются репозиториями метаданных. Им может быть любое место хранения метаданных — от электронной таблицы до базы данных или сложного инструмента. Определение и обсуждение репозитория метаданных продолжается на протяжении всей части I.

Провести грань между данными и информацией можно без философии, но в управлении знаниями без нее уже не обойтись. Чтобы фиксировать знания, необходимо знать, что искать, и философия дает нам ответы на этот вопрос. Мы унаследовали три типа знаний, описанных древним философом Аристотелем в «Никомаховой этике»<sup>1</sup>.

### *Эпистеме*

Научное, теоретическое знание — знание, требующее размышления.

### *Технэ*

Знание практических искусств и ремесел — знание, требующее действий и имеющее физические результаты.

### *Фронесис*

Знание этических соображений, позволяющее судить о правильности поступков.

Как вы увидите, эти типы знаний содержат универсальные истины и непосредственно прослеживаются в технологиях управления знаниями, используемых современными компаниями<sup>2</sup>.

Стефани Барнс выдвинула авангардный подход к управлению знаниями, известный как *радикальное управление знаниями* (<https://oreil.ly/VbuJY>), согласно которому знания фиксируются посредством художественных практик. Этот подход обладает большим потенциалом, поскольку знания трудно зафиксировать.

Почти во всех компаниях области управления ИТ, данными, информацией и знаниями действуют изолированно друг от друга. Такой подход влечет за собой серьезные последствия, поскольку все эти области в той или иной степени описывают ИТ-ландшафт, от которого зависит выполнение их задач.

Но есть решение, позволяющее их объединить, — новый тип команды.

<sup>1</sup> Аристотель. Никомахова этика. — Москва; Берлин: Директ-Медиа, 2020.

<sup>2</sup> Например, в системе управления знаниями Guru (<https://oreil.ly/WUgIL>).

## Команда по обнаружению данных

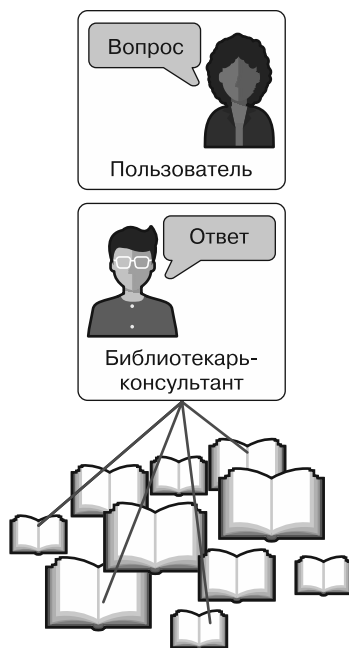
Те, кто читал мою книгу *The Enterprise Data Catalog* (O'Reilly) или следит за моей рассылкой и подкастом *Enterprise Wide Search* (<https://oreil.ly/gHvDa>), знают, что я рассматриваю технологии с новых точек зрения, опираясь на свое академическое образование в области библиотечно-информационных наук (library and information science, LIS). Это справедливо и для данной книги.



Часть II целиком будет посвящена команде по обнаружению данных. Эта команда способствует эффективной работе тщательно согласованных репозиториях метаданных. Команда по обнаружению данных также играет ключевую роль в ускорении получения знаний сотрудниками, поскольку ее ответы можно масштабировать с помощью ИИ. Поэтому такая команда становится скорее ускорителем работы, чем узким местом.

Соответственно, ключевым для этой книги является понятие *библиотекаря-консультанта*. Библиотекарь-консультант — это специалист, который помогает людям в исследованиях и обобщает информацию по определенной теме на основе множества источников. Этот специалист несет ответственность за то, чтобы предоставить наиболее глубокий и полный ответ на любой информационный запрос пользователя библиотеки. Библиотекарь-консультант оценивает множество открыто перечисленных источников, чтобы подготовить ответ даже на очень сложный информационный запрос. Эта область (<https://oreil.ly/OZ7zH>) особенно важна в образовательном и исследовательском контексте. Библиотекарь-консультант помогает исследователям и студентам решать сложные задачи, ориентируясь в непрозрачном массиве данных, который необходимо отфильтровать, чтобы получить максимально точный контекст. Этот подход отличается от «монологитного» подхода, при котором существует «единственный источник истины», способный ответить на все вопросы (рис. 1.1).

Формулирование вопросов и получение ответов на основе одного или нескольких источников информации также применимо к репозиториям метаданных. Рассмотрим пример



**Рис. 1.1.** Библиотекарь-консультант, использующий много источников для подготовки полного ответа

с *приложением*. Представим, что возник простой вопрос: что такое приложение и какие приложения есть в компании?

Если вы спросите команду корпоративных архитекторов, она даст четкий ответ на вопрос, что считать приложением. Помимо определения приложения, у них есть полный перечень всех приложений в инструменте управления архитектурой предприятия (enterprise architecture management, EAM), который является для нее репозиторием метаданных и единственным источником истины об ИТ-ландшафте (возможно, команда знает о других источниках, но не использует их в повседневной работе).

Однако соответствуют ли сведения в инструменте EAM данным из других репозиториях метаданных? Что произойдет, если мы спросим другую команду о том, что такое приложение и какие приложения у нас есть, а она воспользуется для ответа своим репозиторием метаданных? Стоит ли ожидать, что сведения в этих репозиториях метаданных будут совпадать? Или что приложения вообще будут в них зарегистрированы? К сожалению, нет, так как репозитории метаданных создаются и поддерживаются изолированно друг от друга командами по всей компании, которые почти или совсем не взаимодействуют.

Чтобы решить эту проблему, необходимо действовать как библиотекарь-консультант: координировать репозитории метаданных, чтобы давать надежные ответы (рис. 1.2).



**Рис. 1.2.** Предоставление обоснованного ответа на вопрос о приложении с использованием данных из нескольких репозиториях метаданных

Представьте, что репозитории — это элементы сетки, все они связаны друг с другом и образуют коллективную истину вместо изолированных, противоречащих друг другу истин. Именно такую ситуацию, если говорить кратко, книга призвана создать в ИТ-ландшафте вашей компании. Мы обсудим это в главах 3–6.

Необходимо понять, что репозиториями метаданных обычно управляют как монолитами. Давайте разберемся, что это значит, и впервые познакомимся с новой децентрализованной архитектурой управления метаданными.

## Метасетка — третья волна децентрализации данных

Если кратко, *метасетку* можно определить как архитектуру метаданных, предназначенную для их децентрализации и совместного использования в нескольких репозиториях метаданных<sup>1</sup>.

Таким образом, метасетка устанавливает основной набор элементов метаданных, который извлекается из одного репозитория метаданных и становится доступным для совместного использования в нескольких других репозиториях. А теперь вернемся к традиционному управлению данными и посмотрим, чем метасетка отличается от привычного подхода.

Свод знаний по управлению данными DAMA (DAMA-DMBOK) ([https://oreil.ly/Q5\\_t8](https://oreil.ly/Q5_t8)) представляет две архитектурные схемы для управления метаданными (рис. 1.3). Обе схемы включают в себя портал метаданных, который обеспечивает поиск метаданных в типичном наборе репозиториях метаданных, рассматриваемых в области управления данными. Это можно сделать через промежуточный слой, известный как корпоративный репозиторий метаданных, либо непосредственно в репозиториях метаданных.

---

<sup>1</sup> Я должен подчеркнуть, что такую децентрализацию нужно рассматривать в масштабе всей компании, где источники данных крайне разнородны. До этих волн (трех, считая метасетку) аналогичные идеи были сформулированы для открытого Интернета, например в статье *The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities* (<https://oreil.ly/fWQVH>) Тима Бернерса-Ли и Джеймса Хендлера и в самой концепции связанных данных (<https://oreil.ly/CTw6S>). Однако эти идеи уходят корнями в гораздо более ранние времена, например к концепции «Мунданеума» (<https://oreil.ly/mc524>) бельгийского документалиста Поля Отле, созданной в 1920-х годах.

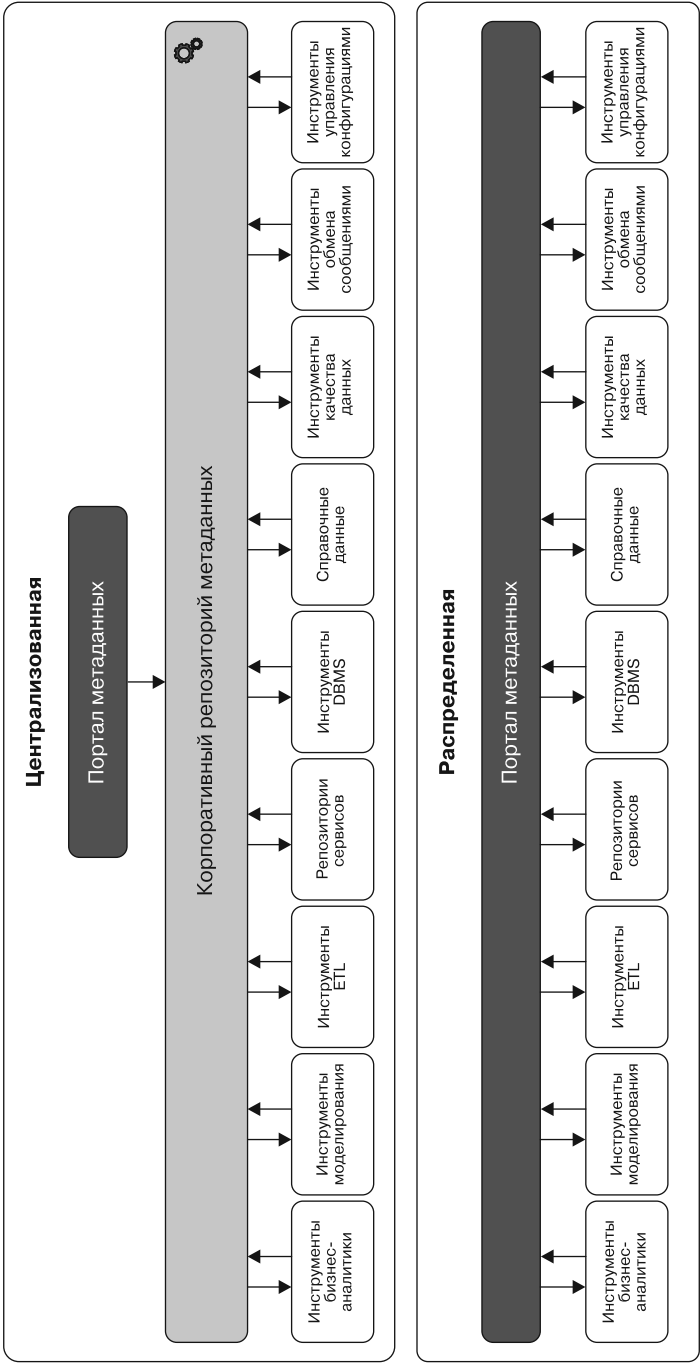


Рис. 1.3. Два стиля управления метаданными, описанные в справочнике DAMA-DMBOK Международной ассоциации DAMA<sup>1</sup>

<sup>1</sup> DAMA-DMBOK: Свод знаний по управлению данными. 2-е изд. — М., 2021.

Однако обе архитектуры несовершенны. Централизованная архитектура метаданных предполагает сбор всех метаданных в общем репозитории, в то время как в распределенной архитектуре поиск ведется напрямую по разным источникам. Но ни одна из них не решает проблему потенциального пересечения метаданных в различных репозиториях, например при описании конечных точек в инструментах извлечения, преобразования и загрузки данных (extract, transform, load, ETL) и бизнес-аналитики (business intelligence, BI). В результате возникают частично избыточные и нескоординированные репозитории. Обе архитектуры носят технический характер и не учитывают важнейший аспект управления метаданными: они исходят из того, что каждый источник содержит уникальные метаданные, не пересекающиеся с метаданными из других источников.

Таким образом, каждый репозиторий метаданных можно считать монолитом (подробнее об этом говорится в части II), изолированным от других, а в случае централизованного репозитория метаданных над ними появляется еще один монолит. Это означает, что метаданные, присутствующие в нескольких репозиториях, не рассматриваются как независимая сущность — они рассматриваются только в контексте того репозитория, в котором представлены. В итоге одни и те же метаданные снова и снова регистрируются в разных репозиториях, почти всегда с ошибками и расхождениями.

В связи с этим в книге предлагается децентрализовать метаданные с помощью метасетки. В отличие от предыдущих методов метасетка позволяет работать с метаданными как с набором небольших, легко управляемых фрагментов. Чтобы лучше понять концепцию метасетки, кратко рассмотрим две более ранние волны децентрализации, которым успешно удалось разбить огромные монолиты данных в компаниях. Этими двумя волнами были *микросервисы* и *сетка данных*.

## Микросервисы

Микросервисы разбивают монолиты операционных данных. *Операционные данные* — это данные, необходимые для работы компании. Они обеспечивают бесперебойное функционирование всей цепочки создания ценности. Традиционно они находятся в крупных технологических компонентах, таких как системы планирования ресурсов организации (enterprise resource planning, ERP), системы управления взаимоотношениями с клиентами (customer relationship management, CRM), системы управления информацией о продуктах (product information management, PIM) и системы управления контентом (content management systems, CMS) (рис. 1.4).

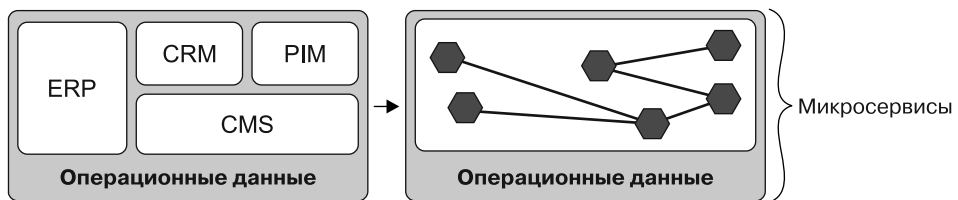
Архитектура, изображенная на рис. 1.4, оказалась под давлением со стороны новых подходов к разработке программного обеспечения (ПО), в частности принципов, представленных еще в 2001 году в Манифесте гибкой разработки

программного обеспечения (Manifesto for Agile Software Development, <https://agilemanifesto.org>). Эти подходы в итоге эволюционировали в то, что мы сейчас называем микросервисами<sup>1</sup>. Данный сдвиг был вызван осознанием того, что крупным технологическим компонентам не хватает скорости и гибкости, чтобы компании могли эффективно развиваться, масштабироваться и адаптироваться к изменениям. Это приводит к эффекту технологической зависимости, когда компании становятся настолько зависимыми от технологий, которые обеспечивают их цепочку создания ценности, что теряют способность изменять эту цепочку. В результате, когда меняется внешняя среда, компания постепенно утрачивает конкурентоспособность.

Микросервисы были предложены как успешная альтернатива, позволяющая дробить эти крупные компоненты (монолиты) на как можно более мелкие *сервисы*, под которыми понимается технология, способная выполнять определенное действие. Эта архитектура называется гексагональной, потому что каждый сервис упакован как продукт и визуализируется в виде шестиугольника (рис. 1.5).



**Рис. 1.4.** Крупные технологические компоненты для операционных данных



**Рис. 1.5.** Микросервисы для операционных данных

Архитектура микросервисов обладает высокой адаптивностью: благодаря своей гибкости она обеспечивает высокую скорость, масштабируемость и возможность быстрой перенастройки. Компании Amazon, Netflix и Uber используют архитектуру микросервисов (на самом деле компания Amazon первой применила микросервисы с подходом «сначала API»). Это позволяет им адаптироваться к поведению и потребностям клиентов, одновременно добавляя новые сервисы на свои платформы. В конечном счете архитектура микросервисов послужила основой для нового подхода — сетки данных.

<sup>1</sup> Основные источники, посвященные микросервисам: *Fowler M. Microservices — A Definition of This New Architectural Term* (<https://oreil.ly/5YJOQ>); *Ньюмен С. Создание микросервисов*. 2-е изд. — СПб.: Питер, 2023; *Bodet G. Scrum en action* (Pearson).

## Сетка данных

*Сетка данных* (data mesh) разбивает монолиты аналитических данных. Уроки, вынесенные из использования микросервисов, легли в основу сетки данных, идея которой возникла в конце 2010-х годов<sup>1</sup>. В концепции сетки данных была предложена архитектура на основе небольших автономных единиц аналитических данных. В отличие от операционных данных аналитические данные не обеспечивают работу цепочки создания ценности в компании. Они используются для анализа деятельности компании и способствуют инновациям благодаря аналитическим сценариям использования на основе машинного обучения (machine learning, ML) и искусственного интеллекта (ИИ). Традиционно аналитические данные хранились, потреблялись и предоставлялись с помощью крупных, довольно сложных технологических платформ для работы с данными, таких как:

- хранилища данных (data warehouses) для BI;
- озера данных (data lakes) для BI, ML и ИИ;
- озеро-хранилища (data lakehouse) — гибриды хранилищ и озер данных.

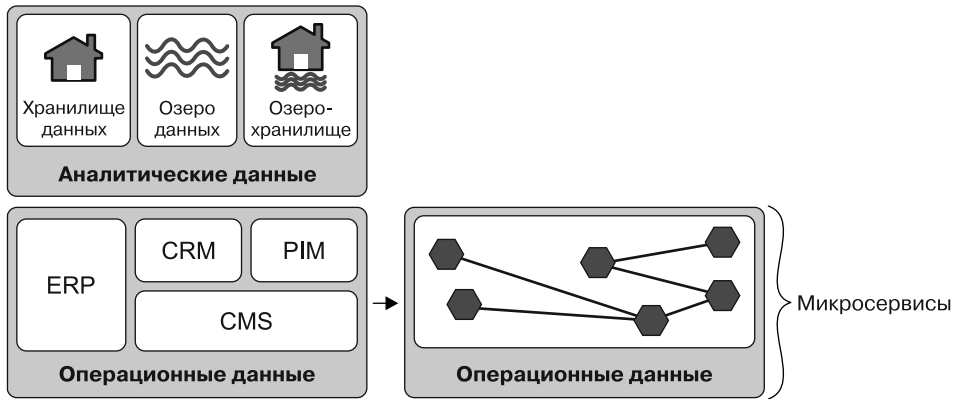
По мере роста потребности в аналитических данных в течение последних 20 лет все три типа платформ для работы с данными стали узкими местами из-за низкой скорости работы, нехватки у сотрудников знаний в предметной области и низкого качества данных.

У архитектуры, представленной на рис. 1.6, множество недостатков. Во-первых, по своей структуре она повторяет архитектуру операционных данных, показанную на рис. 1.4: аналитические данные тщательно размещаются на платформе, которая работает очень медленно и не способна ни масштабироваться, ни перестраиваться. Централизованная платформа данных не успевает за растущим спросом на аналитические данные — поставщики не могут достаточно быстро загружать на платформу данные требуемого качества, а потребители не могут эффективно получать к ним доступ и использовать их.

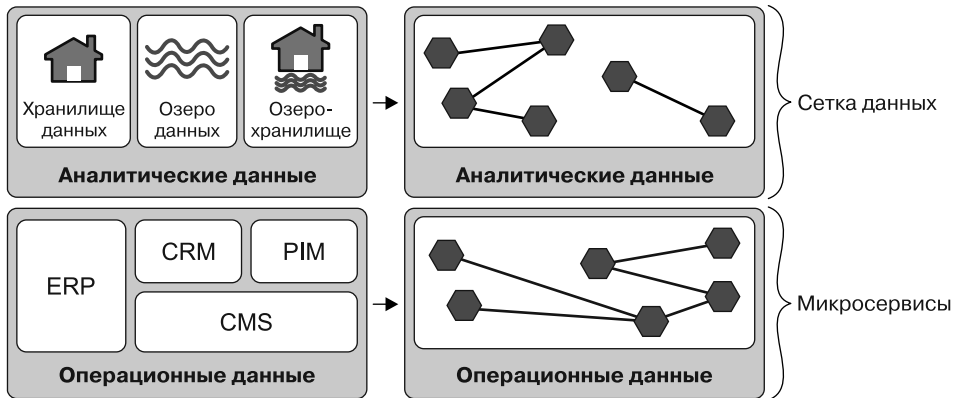
Сетка данных была предложена как альтернатива, предусматривающая для аналитических данных архитектуру, подобную архитектуре микросервисов для операционных данных<sup>2</sup>. Централизованная платформа данных разбивается на бизнес-области с продуктами данных (рис. 1.7).

<sup>1</sup> Dehghani Z. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. — 20 мая 2019 г. ([https://oreil.ly/V\\_xLY](https://oreil.ly/V_xLY)); Dehghani Z. Data Mesh Principles and Logical Architecture. — 3 декабря 2020 г. (<https://oreil.ly/3ulT4>).

<sup>2</sup> На эту тему были одновременно опубликованы две книги: Дехгани Ж. Data Mesh: новая парадигма работы с данными. — Астана, 2023; Стренгхольм П. Масштабируемые данные: высоконагруженные архитектуры, сетки и фабрики данных. 2-е изд. — Астана: «Спринт Бук», 2025.



**Рис. 1.6.** Централизованные платформы данных для аналитических данных превращаются в узкие места



**Рис. 1.7.** Сетка данных для аналитических данных

Архитектура сетки данных масштабируется быстрее централизованной архитектуры и позволяет четче представить каждую область благодаря отказу от единой модели, характерной для централизованного решения. В сетке данных используется множество моделей, каждая из которых относится к своей предметной области, вместо единой централизованной канонической модели данных.

## Метасетка

Метасетка (meta grid) представляет собой иной тип децентрализации, чем микросервисы и сетка данных. Я предлагаю с помощью метасетки совместно использовать метаданные из различных репозиториях метаданных в рамках

сетки, что позволит сделать управление метаданными более надежным и в конечном счете добиться успеха в управлении метаданными в целом.

*Репозиторий метаданных* — это зеркало ИТ-ландшафта вашей компании. Подробнее мы поговорим об этом в главе 2. Репозитории метаданных представляют собой монолиты в том смысле, что каждый из них претендует на роль единственного и абсолютного источника истины об ИТ-ландшафте. Однако ни один из них таковым не является. Все они хранят определенное представление об ИТ-ландшафте — собственный взгляд, который выдается за единственно верный. Монолитность репозитория метаданных на самом деле оказывается иллюзией полного представления об ИТ-ландшафте, которую каждый репозиторий стремится создать. С помощью этой книги я собираюсь развеять эту иллюзию и предложить вместо нее более методический и целостный подход к метаданным.

На рис. 1.8 показаны репозитории метаданных, изображенные одновременно как централизованные инструменты и как изолированные, нескоординированные репозитории.

Микросервисы и сетка данных опираются на схожие подходы и архитектурные шаблоны, но решают разные задачи: первые обеспечивают выполнение цепочки создания ценности, вторая — создание аналитических сценариев использования. Метасетка придерживается схожих подходов и архитектурных принципов, но решает собственную задачу. Метасетка предназначена для получения более надежного представления об ИТ-ландшафте с помощью согласованных репозитория метаданных. Создание надежного, скоординированного представления об ИТ-ландшафте во всех репозиториях метаданных способствует успешному управлению ИТ, данными, информацией и знаниями. Метасетка, рассматриваемая в этом контексте, показана на рис. 1.9.

Обратите внимание на то, что все репозитории метаданных будут относиться к одной из областей управления: ИТ, данными, информацией или знаниями. Эти четыре области управления — единственные области, представленные в метасетке, и каждой из них соответствуют отдельные репозитории метаданных. Архитектура метасетки подробно описывается в части III книги.



Рисунок 1.9 иллюстрирует основную концепцию и суть книги. Метасетка представляет собой нечто вроде сетки данных и микросервисов, но для метаданных. Она проще, компактнее и предъявляет меньше требований. Изображения метасетки содержатся в главах 10 и 11.

Важно отметить, что эта книга не охватывает абсолютно все существующие репозитории метаданных. Лучше представить метасетку как нечто, способное постоянно расширяться и включать все новые репозитории.

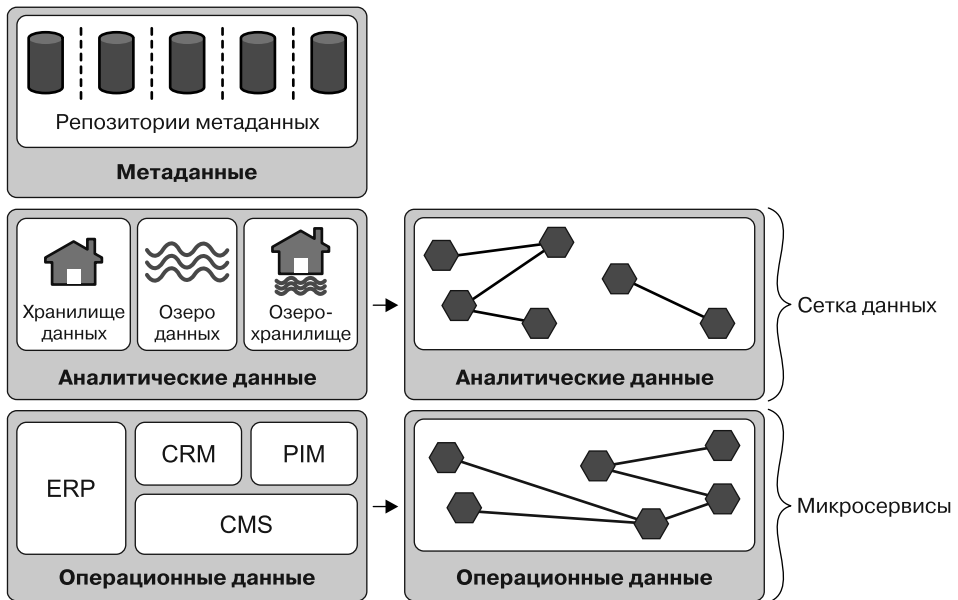


Рис. 1.8. Репозитории метаданных для работы с метаданными

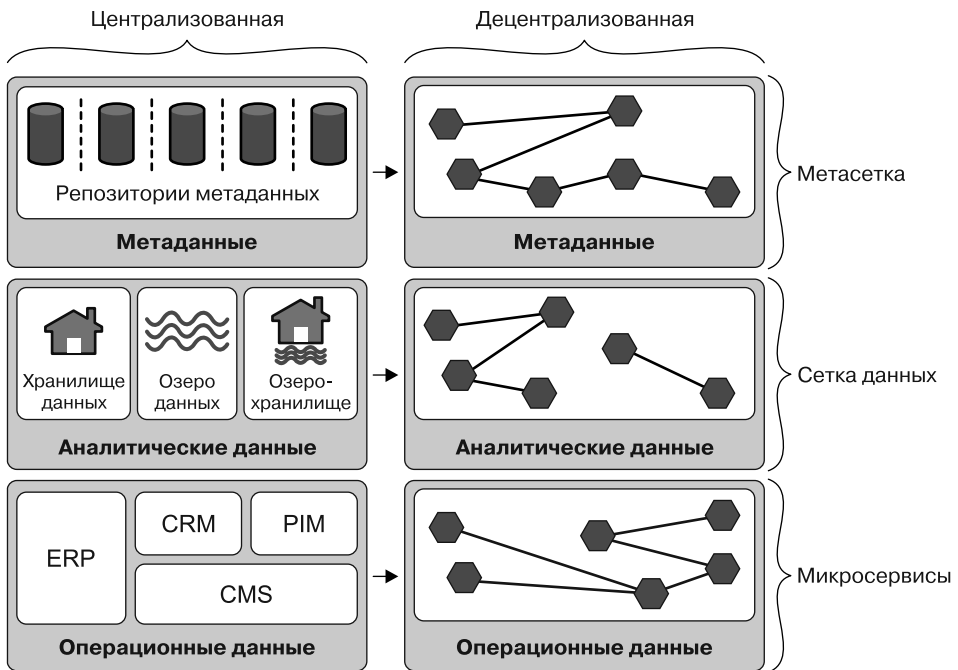


Рис. 1.9. Третья волна децентрализации данных — метасетка

Настройка архитектур микросервисов и сетки данных может быть сложной, требовать серьезной технической проработки и глубокого теоретического осмысления. Для сравнения: архитектура метасетки проста и ее легко совершенствовать, однако связанные с ней организационные изменения могут быть значительными.

## Резюме

В этой главе были представлены краткое введение и общий обзор того, что вы найдете в книге. Самое важное, что следует уяснить: управление метаданными нужно понимать как способ координации множества существующих репозиторий метаданных в компаниях. Затем мы обсудили управление метаданными и репозитории метаданных. Я объяснил, что существуют репозитории метаданных для управления ИТ, данными, информацией и знаниями, и рассказал о команде по обнаружению данных, которая будет играть ключевую роль в вашей компании — помогать создавать полезное представление об ИТ-ландшафте, объединяя усилия уже существующих в компании команд. Наконец, я выдвинул идею метасетки, которая призвана решить реальную проблему управления метаданными и помочь получить более целостное представление об ИТ-ландшафте.

Основные выводы этой главы.

- Управление метаданными необходимо переосмыслить, чтобы выйти за рамки традиционных подходов к управлению данными, поскольку они не отражают реального положения дел в компаниях.
- В сфере управления ИТ существует множество репозиторий метаданных, каждый из которых по-своему описывает ИТ-ландшафт.
- В областях управления информацией и знаниями также используются репозитории метаданных, содержащие различные представления об ИТ-ландшафте.
- Прообразом команды по обнаружению данных послужили библиотекари-консультанты, которые для предоставления исчерпывающих ответов на сложные запросы используют не один, а несколько источников.
- Подобным образом команда по обнаружению данных учится давать ответы об ИТ-ландшафте на основе множества источников вместо одного, чтобы эти ответы были максимально полными.
- Метасетка — это третья волна децентрализации, идущая по стопам микросервисов и сетки данных.
- Метасетка меньше и проще, чем микросервисы и сетка данных.