ГЛАВА 1

Большие данные (Big Data)

В 2020-х годах число компаний, создающих архитектуры данных, резко возросло. Этот рост вряд ли замедлится в ближайшее время во многом потому, что данных стало доступно больше, чем когда-либо прежде: из социальных сетей, устройств интернета вещей (IoT), непрофессиональных приложений и стороннего ПО, и это лишь несколько источников. Согласно исследованию BCG (Boston Consulting Group) за 2023 год (https://oreil.ly/hpOPt), «объем генерируемых данных с 2018 по 2021 год увеличился примерно в два раза и составил около 84 зеттабайт, и ожидается, что темпы роста сохранятся». По оценкам исследователей, «в период с 2021 по 2024 год объем генерируемых данных будет увеличиваться с совокупным годовым приростом (CAGR, compound annual growth rate) в 21 % и достигнет 149 зеттабайт». Компании знают, что могут сэкономить миллионы долларов и увеличить свой доход, собирая эти данные и используя их для анализа прошлого и настоящего, а также для прогнозирования будущего, но для этого им нужен способ хранения всех этих данных.

Во всем деловом мире спешат как можно быстрее создавать архитектуры данных. Эти архитектуры должны быть готовы обрабатывать любые будущие данные — независимо от их размера, скорости обработки или типа — и поддерживать их точность. И тем из нас, кто имеет дело с архитектурами данных, необходимо четко понимать, как они работают и какие их разновидности существуют. Здесь и пригодится эта книга. Я не понаслышке знаю, к чему приводит неверное понимание сути архитектуры данных. Одна известная мне компания построила архитектуру данных стоимостью 100 миллионов долларов за два года, а потом обнаружила, что в ней использовалась неподходящая технология, она была слишком сложна и недостаточно гибка для обработки определенных типов данных. Пришлось все снести и начать все заново. Не дайте подобному случиться с вами!

Главное — донести нужную информацию до нужных людей в нужное время и в правильной форме. Для этого необходима архитектура данных, позволяющая получать, хранить, преобразовывать и моделировать данные (все это — обработка больших данных), чтобы их можно было правильно и легко использовать. Нужна архитектура, которая позволит любому конечному пользователю, не обладающему обширными техническими знаниями, анализировать данные и создавать отчеты и информационные панели (дашборды), а не рассчитывать, что специалисты в ИТ-отделе сделают это за них.

Глава 1 начинается с введения в теорию больших данных и описания некоторых ее основополагающих принципов. Затем мы обсудим, как компании используют свои данные, с акцентом на бизнес-аналитику и расширение их использования по мере развития архитектуры данных в компании.

Что такое большие данные и чем они могут быть вам полезны?

Несмотря на то что в термине *«большие данные»* (Big Data) фигурирует слово *«большие»*, он описывает не только размер данных. Он подразумевает все данные, большие или малые, внутри компании и все данные за ее пределами, которые могут быть вам полезны. Данные могут быть представлены в любом формате, а их сбор может происходить с любой частотой. Поэтому лучший способ определить понятие *«большие данные»* — считать, что это *все* данные, независимо от их размера (объема), быстроты обработки (скорости) или типа (разнообразия). В дополнение к этим критериям для описания данных можно использовать еще три: достоверность, изменчивость и ценность. Вместе они известны как «шесть *V*» больших данных (volume, velocity, variety, veracity, variability, value), как показано на рис. 1.1.

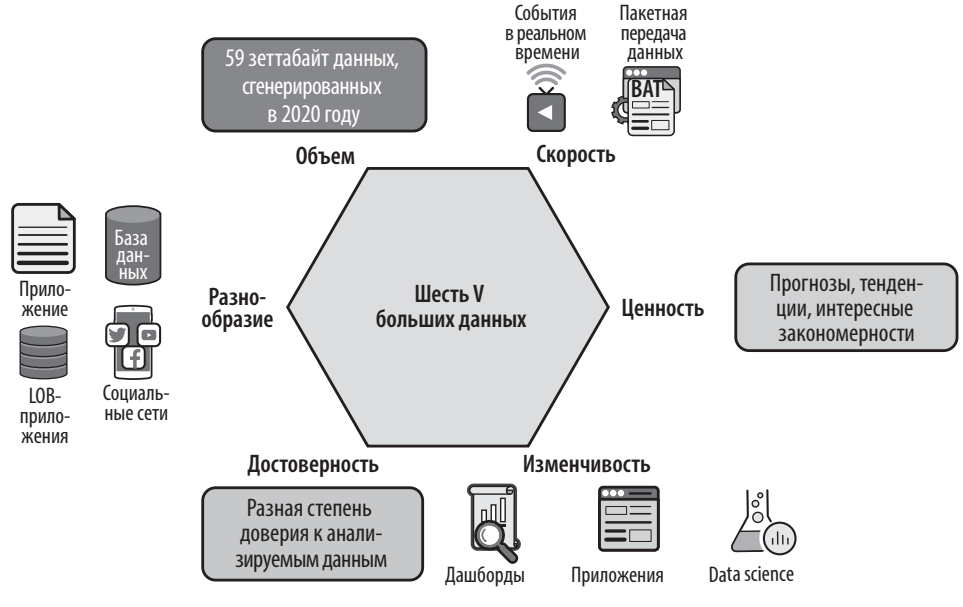

Рис. 1.1. Шесть V больших данных (источник: The Cloud Data Lake by Rukmani Gopalan [O'Reilly, 2023])

Рассмотрим каждый из этих критериев подробнее.

Объем (Volume)

Объем — это огромное количество сгенерированных и сохраненных данных. Он может составлять от терабайта до петабайта данных из самых разных источников, включая социальные сети, транзакции электронной коммерции, научные эксперименты, данные датчиков IoT-устройств и многие другие. Например, данные из системы ввода заказов могут составлять пару терабайт в день, а IoT-устройства — обрабатывать миллионы событий в минуту и генерировать сотни терабайт данных в день.

Разнообразие (Variety)

Под *разнообразием* понимается широкий спектр источников и форматов данных. Их можно разделить на *структурированные данные* (из реляционных баз данных), *полуструктурированные данные* (например, логи и документы в форматах CSV, XML и JSON), *неструктурированные данные* (например, электронные письма, документы и PDF-файлы) и *бинарные данные* (изображения, аудио, видео). Так, данные из системы ввода заказов будут структурированными, поскольку они поступают из реляционной базы данных, а данные с IoT-устройств, скорее всего, будут в формате JSON.

Скорость (Velocity)

Скорость — это скорость, с которой генерируются и обрабатываются данные. Сбор данных, производящийся редко, обычно называют *пакетной обработкой*; например, когда все заказы за день собираются и обрабатываются вечером. Однако данные могут собираться и достаточно часто или даже в режиме реального времени, особенно если они генерируются с высокой скоростью, как, например, данные из социальных сетей, с IoT-устройств и из мобильных приложений.

Достоверность (Veracity)

Достоверность — это точность и надежность данных. Большие данные поступают из самых разных источников. Использование ненадежных или неполных источников может отрицательно сказаться на качестве данных. Например, данные поступают с IoT-устройства — камеры наружного наблюдения, расположенной перед домом и направленной на подъездную дорожку, и она отправляет текстовое сообщение, когда обнаруживает человека. Не исключено, что подобное устройство сработает ложно из-за факторов окружающей среды, таких как погода, что приведет к ошибке в данных. Таким образом, при получении данных их необходимо проверять.

Изменчивость (Variability)

Под *изменчивостью* понимается согласованность (или несогласованность) данных с точки зрения их качества, содержания и смысла. Для обработки и анализа структурированных, полуструктурированных и неструктурированных форматов данных требуются разные инструменты и методики. Например, тип, частота и качество данных, поступающих с датчиков IoT-устройств, могут

сильно различаться. Датчики температуры и влажности могут генерировать данные через регулярные промежутки времени, в то время как датчики движения генерируют данные только при наличии движения.

Ценность (Value)

Важнейший *V*-фактор — *ценность* — связан с полезностью и актуальностью данных. Компании используют большие данные для получения информации и принятия решений, ведущих к созданию таких бизнес-ценностей, как повышение эффективности, снижение затрат или новые источники дохода. Например, анализируя данные о клиентах, организации могут лучше понять их поведение, предпочтения и потребности. Они могут использовать эту информацию для разработки таргетированных маркетинговых кампаний, повышения качества обслуживания клиентов и стимулирования продаж.

Собирая большие данные, компании получают информацию, которая помогает им принимать более обоснованные бизнес-решения. *Прогностический анализ* — это тип анализа данных, который предполагает использование статистических алгоритмов и машинного обучения для анализа исторических данных и составления прогнозов относительно будущих событий и тенденций. Это позволяет компаниям не просто реагировать, а действовать на опережение.

В последнее время многие компании называют данные «новой нефтью», поскольку в современной цифровой экономике они стали невероятно ценным ресурсом, подобно нефти в индустриальную эпоху. Данные похожи на нефть во многих отношениях:

- Это сырье, которое необходимо добыть, очистить и переработать, чтобы оно стало полезным. В случае с данными речь идет об их сборе, хранении и анализе для получения информации, которая может стать основой для принятия бизнес-решений.
- Они невероятно ценны. Компании, которые собирают и анализируют большие объемы данных, могут использовать их для улучшения своих продуктов и услуг, принятия более обоснованных бизнес-решений и получения конкурентных преимуществ.
- Данные можно использовать самыми разными способами. Например, обученные на данных алгоритмы машинного обучения в дальнейшем можно применять для автоматизации задач, выявления закономерностей и составления прогнозов.
- Это мощный ресурс, способный кардинально изменить общество. Широкое распространение нефти стимулировало развитие многих отраслей промышленности и новых технологий, а появление больших данных привело к прогрессу в таких областях, как искусственный интеллект, машинное обучение и предиктивная аналитика.
- Именно вследствие всех вышеперечисленных факторов данные могут стать основой могущества и влияния.

Большие данные можно использовать для создания отчетов и дашбордов, где они подскажут, в каких точках продажи отстают, и предпринять шаги «постфактум» для повышения продаж. Машинное обучение может предсказать, в каких точках продажи упадут в будущем, что позволит предпринять упреждающие шаги, чтобы предотвратить это падение. Это называется *бизнес-аналитикой* (business intelligence, BI): процесс сбора, анализа и использования данных, помогающий бизнесу принимать более обоснованные решения.

Согласно рис. 1.2, данные можно собирать из новых источников, например IoT-устройств, блогов и социальных сетей, а также из старых источников, таких как бизнес-приложения, приложения для планирования ресурсов предприятия (ERP, enterprise resource planning) и управления отношениями с клиентами (CRM, customer relationship management). Эти данные могут быть представлены в различных форматах: CSV, JSON, Parquet и др. Они могут поступать пакетами, например раз в час, или передаваться несколько раз в секунду (это называется *потоковой передачей в реальном времени*).

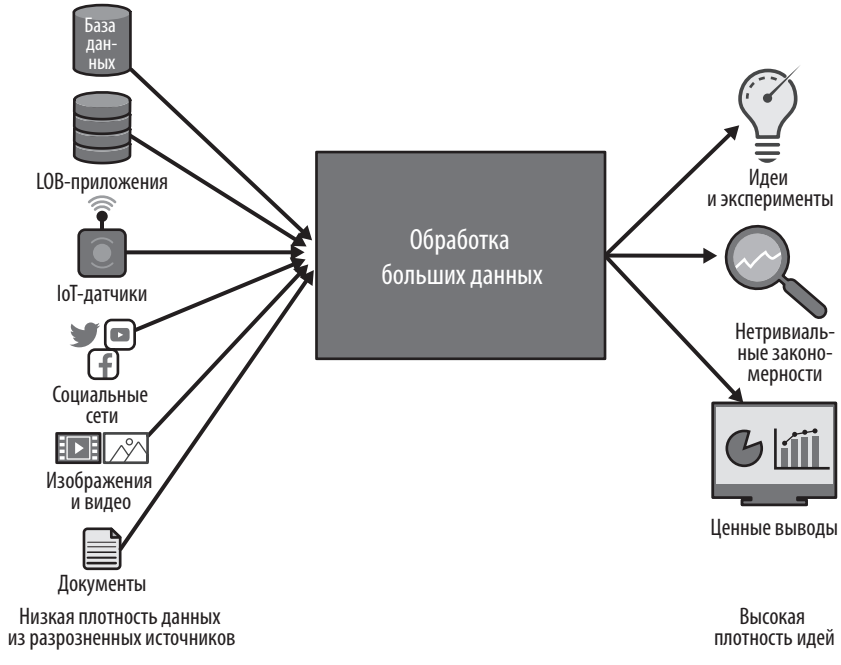

Рис. 1.2. Обработка больших данных (источник: The Cloud Data Lake by Rukmani Gopalan [O'Reilly, 2023])

Компаниям важно понимать, на каком уровне использования данных они находятся по сравнению с другими игроками. Существует концепция *зрелости данных* (data maturity), и в следующем разделе мы рассмотрим ее уровни, чтобы вы могли понять, на каком из них находится ваша компания.

Зрелость данных

Возможно, вы уже не раз слышали такой термин, как *«цифровая трансформация»*, который означает, что компания внедряет цифровые технологии, чтобы коренным образом изменить процесс извлечения ценности из данных, а также собственные операции и способы предоставления ценности клиентам. Цифровая трансформация предполагает переход от традиционных, ручных или бумажных процессов к цифровым, с использованием всех возможностей ИТ-технологий для повышения эффективности, производительности и инновационности. Важной частью этой трансформации обычно является использование данных для оптимизации бизнеса, что может означать создание профиля клиента 360[1] (https://oreil.ly/rSF6P) для повышения качества обслуживания клиентов или применение машинного обучения для повышения скорости и точности работы производственных линий.

Цифровую трансформацию можно разделить на четыре этапа, которые называются уровнями *зрелости корпоративных данных* (рис. 1.3). Хотя термин широко используется в ИТ, у меня есть свой взгляд на подобное разделение. Эти этапы описывают уровень развития и степень сложности, которых достигла организация в управлении, использовании и извлечении ценности из своих данных. Эта модель — способ оценить возможности организации по управлению данными и ее готовность к использованию расширенной аналитики, искусственного интеллекта, а также других систем, работающих с Big Data. На каждом этапе эффективность использования данных при создании бизнес-ценности и принятии решений растет. Оставшаяся часть раздела содержит описание этих этапов.

Уровень 1. Реактивный

На первом уровне данные компании находятся в разных местах, скорее всего, в виде множества электронных таблиц Excel и/или локальных баз данных, в различных файловых системах, и все это пересылается по электронной почте по всей компании. Архитекторы данных называют это *витриной таблиц* (spreadmart) (сокращение от «spreadsheet data mart» — «витрина таблиц данных»): неформализованный, децентрализованный набор данных, часто встречающийся в организациях, где используются электронные таблицы для хранения, управления и анализа данных. Отдельные сотрудники или команды обычно создают и поддерживают витрины таблиц независимо от централизованной системы управления данными или официального хранилища данных. Витрины таблиц страдают несогласованностью данных, отсутствием управления, ограниченной масштабируемостью и неэффективностью (поскольку часто требуют дублирующихся действий).

---

[1] Профиль клиента 360 — это база данных, которая включает всю историю взаимодействия клиента с брендом; данные собираются из разных источников и постоянно обновляются. — *Примеч. пер.*

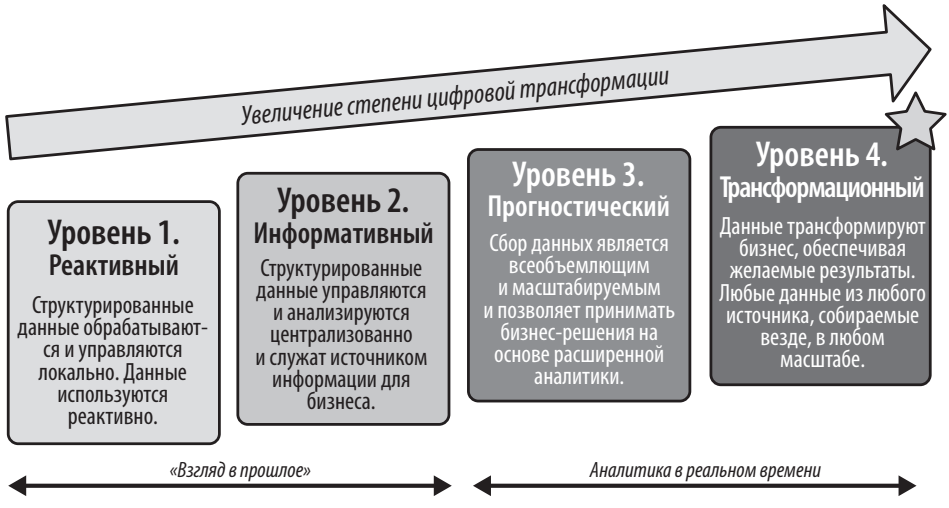

Рис. 1.3. Уровни зрелости корпоративных данных

Уровень 2. Информативный

Компании достигают второго уровня зрелости, когда начинают централизовывать свои данные, что значительно упрощает проведение анализа и создание отчетов. Уровни 1 и 2 предназначены для создания *ретроспективных (исторических) статистических отчетов* или отслеживания трендов и закономерностей в прошлом, поэтому на рис. 1.3 они обозначены как «взгляд в прошлое». На этих этапах вы только реагируете на то, что уже произошло.

На уровне 2 решение, созданное для сбора данных, обычно не очень хорошо масштабируется. Как правило, объем и типы данных, которые оно может обрабатывать, ограниченны, и оно может принимать данные относительно редко (к примеру, каждую ночь). Большинство компаний находятся на уровне 2, особенно если их инфраструктура все еще остается *локальной*[1].

[1] Под «локальной» понимается такая ИТ-инфраструктура организации — серверы, системы хранения данных и сетевое оборудование, — размещение и управление которой производятся в собственных физических помещениях организации, обычно называемых дата-центрами. В этом их отличие от облачных сервисов, где ресурсы размещаются и управляются сторонними провайдерами, такими как Azure, Amazon Web Services (AWS) или Google Cloud Platform (GCP), в удаленных дата-центрах. О преимуществах перехода от локальной инфраструктуры к облачной я расскажу в главе 16, а пока примите к сведению, что переход с локальных серверов на облачные является важной частью цифровой трансформации большинства предприятий.

Уровень 3. Прогностический

На третьем уровне находятся компании, которые уже перешли на облачные технологии и создали систему, способную обрабатывать большие объемы данных, разные типы данных и данные, которые поступают часто (каждый час или в потоковом режиме). Они также улучшили процесс принятия решений, внедрив машинное обучение (расширенную аналитику), чтобы принимать решения в режиме реального времени. Например, когда пользователь заходит в книжный интернет-магазин, система может рекомендовать ему дополнительные книги на странице оформления заказа, основываясь на его предыдущих покупках.

Уровень 4. Трансформационный

Наконец, на четвертом уровне компания создает решение, способное обрабатывать любые данные, независимо от их объема, скорости поступления и типа. Новые данные можно легко и быстро использовать в работе, поскольку архитектура способна их обрабатывать и имеет инфраструктуру для их поддержки. Это решение позволяет конечным пользователям, не обладающим техническими знаниями, легко создавать отчеты и дашборды с помощью выбранных ими инструментов.

В этой книге мы сосредоточимся на уровнях 3 и 4. В частности, процесс, когда конечные пользователи сами создают отчеты, называется *бизнес-аналитика самообслуживания*, о ней речь пойдет в следующем разделе.

Бизнес-аналитика самообслуживания

В течение многих лет, если конечному пользователю в организации требовался отчет или дашборд, ему приходилось собирать все свои материалы (необходимые исходные данные, а также описание того, как должен выглядеть отчет или дашборд), заполнять форму запроса в ИТ-отдел и ждать. ИТ-отдел создавал отчет. Процесс включал извлечение данных, загрузку их в хранилище, построение модели данных и, наконец, формирование отчета или дашборда. Конечный пользователь просматривал его и либо утверждал, либо запрашивал изменения. Это часто приводило к длинной очереди запросов, так что в итоге ИТ-отдел становился серьезным «узким местом». Конечным пользователям требовались дни, недели или даже месяцы, чтобы начать извлекать пользу из данных. Сейчас этот процесс называют «традиционной бизнес-аналитикой», поскольку в последние годы появилось нечто лучшее — бизнес-аналитика самообслуживания (self-service BI).

Цель любого решения архитектуры данных, которое вы создаете, должна заключаться в том, чтобы каждый конечный пользователь, независимо от своих технических навыков, мог быстро и легко запрашивать данные и создавать отчеты и дашборды. Для выполнения этой работы не нужно привлекать ИТ-специалистов — пользователи должны справляться с этим самостоятельно.

Чтобы этого добиться, придется проделать большую предварительную работу: пообщаться со всеми конечными пользователями, чтобы выяснить, какие данные им нужны, а затем построить архитектуру данных с учетом их потребностей. Но результат будет стоить того, поскольку позволит сэкономить время на создание отчетов. Такой подход избавляет от очередей и переписки с ИТ-отделом, сотрудники которого, как правило, мало что понимают в конкретных данных. Вместо этого конечный пользователь, который лучше всех разбирается в данных, получает прямой доступ к ним, подготавливает их, создает модель данных, формирует отчеты и проверяет их правильность. Такой рабочий процесс гораздо более продуктивен.

Создание столь простого в применении решения по работе с данными приводит к появлению бизнес-аналитики самообслуживания. Сформировать отчет должно быть так же просто, как перетащить поле в документе. Конечные пользователи не должны разбираться, как объединять данные из разных таблиц, или беспокоиться о том, что отчет обрабатывается слишком медленно. Создавая систему обработки данных, всегда спрашивайте себя: *легко ли будет другим людям формировать собственные отчеты?*

Итоги

Прочитав эту главу, вы узнали, что такое Big Data и как они могут помочь вам и вашей организации принимать более эффективные бизнес-решения, особенно в сочетании с машинным обучением. Вы научились описывать большие данные с помощью шести V, а также узнали, что такое зрелость данных и как определять ее уровни. Наконец, вы поняли разницу между традиционной бизнес-аналитикой (BI) и бизнес-аналитикой самообслуживания: цель последней состоит в том, чтобы каждый мог использовать данные для быстрого и простого создания отчетов и получения аналитических выводов.

Теперь позвольте вкратце описать, что вас ждет в следующих главах. В главе 2 я расскажу о том, что такое архитектура данных, и представлю общий обзор изменений типов архитектур данных со временем. В главе 3 покажу, как провести дизайн-сессию по разработке архитектуры, помогающую определить, какая архитектура данных лучше всего подходит для ваших целей.

В части II «Общие понятия архитектуры данных» более подробно рассказывается о различных архитектурах. В главе 4 речь пойдет о том, что такое хранилище данных и чем оно не является, а также для чего его нужно использовать. Я остановлюсь на подходе «сверху вниз», отвечу на вопрос, пришел ли конец реляционным хранилищам данных, и расскажу о способах заполнения хранилища. В главе 5 описывается, что такое озеро данных и почему нужно его использовать. В ней также обсуждается подход «снизу вверх», а затем более подробно рассматривается проектирование озера данных и случаи, когда стоит использовать несколько озер данных.

Глава 6 посвящена общим концепциям архитектуры данных, связанным с хранением данных, включая витрины данных (data mart), хранилища оперативных

данных, управление мастер-данными и виртуализацию данных. В главе 7 рассматриваются общие понятия архитектуры данных, связанные с проектированием, в том числе оперативная обработка транзакций (OLTP, Online Transaction Processing) в сравнении с оперативной аналитической обработкой (OLAP, Online Analytical Processing), оперативные данные в сравнении с аналитическими, симметричная многопроцессорная архитектура (SMP, symmetric multiprocessing) в сравнении с массово-параллельной (MPP, massive parallel processing), а также лямбда-архитектуры, каппа-архитектуры и полиглотное хранение (polyglot persistence). В главе 8 речь идет о моделировании данных, в том числе реляционном и размерном моделировании, дискуссии между Кимбаллом (Kimball) и Инмоном (Inmon), общей модели данных и моделям типа Data Vault (свод данных). А в главе 9 вы прочтете о приеме данных, в том числе об операциях ETL (extract, transform, load — извлечение, преобразование, загрузка) в сравнении с ELT (extract, load, transform — извлечение, загрузка, преобразование), обратном ELT, пакетной обработке в сравнении с обработкой в реальном времени, а также об управлении данными.

Часть III посвящена конкретным архитектурам данных. В главе 10 описано современное хранилище данных и пять этапов его создания. В главе 11 рассматриваются архитектурный паттерн «ткань данных» (Data Fabric) и варианты его использования. В главе 12 обсуждается архитектура озера-хранилища данных (data lakehouse), а таже плюсы и минусы отказа от использования реляционных хранилищ данных.

Главы 13 и 14 посвящены архитектурам сеток данных (data mesh) — здесь есть о чем поговорить! Глава 13 фокусируется на децентрализованном подходе и четырех принципах построения сетки данных, а также описывает, что такое домены данных и продукты данных. Глава 14 посвящена проблемам и вызовам, связанным с построением сеток данных, и развенчивает некоторые распространенные мифы о них. Она поможет проверить, готовы ли вы к внедрению сетки данных. В заключение мы поговорим о том, каким может быть будущее сетки данных.

В главе 15 рассматриваются причины успеха и неудач различных проектов, а также описывается, как правильно организовать команду для создания архитектуры данных. Наконец, глава 16 посвящена обсуждению открытого исходного кода, преимуществ облачных технологий, основных облачных провайдеров, мультиоблачности и программных фреймворков.

Пришла пора в корне изменить ваши представления о данных. Готовы?