Ральф Кимбалл Марджи Росс

ИНСТРУМЕНТАРИЙ ХРАНЕНИЯ И АНАЛИЗА ДАННЫХ

ПОЛНОЕ РУКОВОДСТВО ПО РАЗМЕРНОМУ МОДЕЛИРОВАНИЮ

БОМБОРА
ИЗДАТЕЛЬСТВО

УДК 004.6
ББК 32.973.2-018.2
К40

Кимбалл, Ральф.

К40 Инструментарий хранения и анализа данных : полное руководство по размерному моделированию / Ральф Кимбалл, Марджи Росс ; [перевод с английского М. А. Райтмана]. — Москва : Эксмо, 2024. — 656 с. — (Data Science. Лучшие книги о науке о данных).

ISBN 978-5-04-108040-2

Сегодня тысячи компаний собирают и сохраняют большие данные о поведении своих клиентов, ассортименте, производственном процессе и других немаловажных для бизнеса вещах. Однако, чтобы принимать обоснованные решения на основе этих данных, недостаточно их просто собрать – нужно правильно их обработать и провести грамотный анализ. Благодаря этой книге вы освоите все необходимые инструменты для хранения и анализа большого количества данных, научитесь правильно управлять ими и извлекать полезную информацию для развития бизнеса.

УДК 004.6
ББК 32.973.2-018.2

ISBN 978-5-04-108040-2

Оглавление

Введение

Отрасль хранения данных (DW, Data Warehousing) и анализа данных (BI, Business Intelligence) стала развиваться с момента публикации первого издания «Инструментарий хранилищ данных» Ральфа Кимбалла в 1996 г. Хотя ранние крупные корпоративные последователи проложили путь, концепции DW/BI с тех пор были охвачены организациями всех размеров. Отрасль создала тысячи систем DW/BI. Объем данных продолжает расти, поскольку хранилища наполнены растущим количеством атомарных данных и обновляются со все большей частотой. На протяжении нашей карьеры мы видели, что базы данных растут с мегабайтов и гигабайтов до терабайтов и петабайтов, но все же основная задача систем DW/BI остается на удивление постоянной.

Наша работа состоит в том, чтобы передать их бизнес-клиентам для помощи в принятии решений. Мы пришли к этой цели коллективно, бизнес-профессионалы всегда и везде принимают лучшие решения и окупают свои инвестиции в системы DW/BI.

С момента, когда было опубликовано первое издание книги «Инструментарий хранилищ данных», размерное моделирование было принято в качестве основной техники для представления DW/BI. Практики и эксперты единогласно признали, что представление данных должно быть основано на простоте, если вы хотите иметь какие-либо шансы на успех. Простота — фундаментальный ключ, который позволяет пользователям легко понимать базы данных и программное обеспечение для эффективного использования баз данных. Последовательно возвращаясь к перспективе, диктуемой бизнесом, и не компрометируя скорость выполнения запросов и легкость понимания структур пользователями, вы устанавливаете последовательный дизайн, удовлетворяющий аналитическим нуждам организации. Этот многомерно смоделированный фреймворк становится платформой BI. Основываясь на нашем опыте и ошеломительной обратной связи многочисленных практиков из разных компаний, таких как вы, мы верим, что размерное моделирование критически важно для успешной поддержки DW/BI.

Размерное моделирование также выступает в качестве ведущей архитектуры для создания интегрированных систем DW/BI. Когда вы используете согласованные измерения и факты набора размерных моделей, вы имеете практичный и предсказуемый фреймворк для пошагового создания комплексных систем DW/BI, которые изначально распределены.

Учитывая, как много изменений произошло в нашей отрасли, основные методы размерного моделирования, опубликованные Ральфом Кимбаллом 17 лет назад, претерпели проверку временем. Концепции, такие как согласованные измерения, медленно меняющиеся измерения, разнородные продукты, таблицы фактов без метрик, а также матрица шины предприятия продолжают обсуждаться на симпозиумах по всему земному шару. Оригинальные понятия были детализированы и расширены новыми дополнительными методами. Мы решили опубликовать это третье издание оригинальной работы Кимбалла, так как чувствовали, что будет полезно суммировать наш коллективный опыт размерного моделирования под общей обложкой. Каждый из нас концентрировался исключительно на поддержке принятия решений, хранилищах и анализе данных в течение более чем трех десятилетий. Мы хотим поделиться паттернами размерного моделирования, которые неоднократно появлялись в ходе нашей карьеры. Книга нагружена специфическими практическими рекомендациями по проектированию, основанными на реальных сценариях.

Цель этой книги состоит в том, чтобы предоставить универсальный магазин для методов размерного моделирования. Соответственно своему названию книга представляет собой набор инструментов, принципов и методов размерного проектирования. Мы стремимся удовлетворить потребности тех, кто только начинает знакомство с DW/BI, и описываем расширенные понятия для тех из вас, кто уже давно знаком с DW/BI. Мы убеждены, что эта книга выделяется глубиной освещения темы размерного моделирования. Это исчерпывающее руководство.

Для кого эта книга

Книга предназначена для разработчиков в сфере хранилищ и анализа данных, инженеров и менеджеров. Кроме того, книгу найдут полезной аналитики данных и люди, управляющие данными, активно взаимодействующие с DW/BI.

Даже если вы прямо не связаны с размерной моделью, мы считаем важным, чтобы все члены команды проекта были знакомы с понятиями размерного моделирования. Размерная модель оказывает влияние на большинство аспектов реализации DW/BI, начиная от перевода бизнес-требований через процессы извлечения, преобразования и загрузки (extract, transformation, load — ETL) и заканчивая открытием хранилища данных через приложения аналитики данных. Из-за широкого спектра применения вы должны быть сведущими в размерном моделировании независимо от того, за что именно вы ответственны — за управление проектом, анализ данных, архитектуру данных, проектирование баз данных, ETL, приложения BI или обучение и поддержку. Мы написали эту книгу таким образом, что она доступна для широкой аудитории.

Те из вас, кто читал ранние издания этой книги, найдут некоторые уже знакомые тематические исследования, однако они были значительно обновлены и обогащены содержанием, включая примеры матриц шин предприятий почти для каждого тематического исследования. Мы разработали краткие описания для новых предметных областей, включая анализ больших данных.

Содержание этой книги техническое. Впервые мы обсудим размерное моделирование в контексте реляционных баз данных с нюансами для кубов аналитической обработки в реальном времени (OLAP — online analytical processing), что отмечено в соответствующих случаях. Мы предполагаем, что вы имеете базовые понятия о реляционных базах данных, такие как таблицы, строки, ключи и объединения таблиц (join). Поскольку мы будем обсуждать размерные модели в нетрадиционном ключе, мы не станем углубляться в конкретные рекомендации по физическому проектированию и настройке любых описываемых СУБД.

Структура книги

Книга организована вокруг серий кратких описаний или примеров. Мы считаем, что разработка техник проектирования с помощью примеров — весьма эффективный способ, поскольку позволяет делиться реальным руководством и извлекать выгоду из опыта реального мира. Хотя эти примеры не предназначены для того, чтобы стать полномасштабными приложениями или решениями для отрасли, они служат основой для обсуждения паттернов, появляющихся в размерном моделировании. По нашему опыту, зачастую легче уловить главные элементы метода разработки, если отступить от слишком знакомых сложностей собственного бизнеса. Читатели предыдущих изданий благоприятно отозвались о таком подходе.

Предупреждаем вас, что мы отклоняемся от подхода учебных примеров в главе 2. Учитывая широкое промышленное принятие методов размерного моделирования, разработанных Kimball Group, мы составили официальный список наших методов наряду с краткими описаниями и указателями на более подробное освещение и иллюстрации этих методов в последующих главах. Хотя эту главу необязательно читать от начала и до конца, как другие, мы считаем, что эта центральная техническая глава может быть полезной ссылкой и может даже послужить профессиональным контрольным списком для разработчиков DW/BI.

За исключением главы 2 остальные части этой книги опираются друг на друга. Мы начинаем с базовых понятий и представляем более расширенное содержание по мере чтения. Главы следует читать по порядку каждому читателю. Например, может быть сложно осмыслить главу 16, пока вы не прочтете предыдущие главы о розничных продажах, закупках, управлении персоналом и управлении отношениями с клиентами.

Тот, кто читал предыдущее издание, может пропустить первые несколько глав. Хотя некоторые ранние факты и изучение измерений могут быть вам знакомы, мы бы не хотели, чтобы вы забегали слишком далеко вперед. Вы упустите новое в фундаментальных понятиях, если пролистаете много.

ПРИМЕЧАНИЕ Книга наполнена подсказками (такими, как это примечание), списками ключевых понятий и указателями на главы, которые будут более полезны и на которые будет легче ссылаться в будущем.

Глава 1. Хранение данных, анализ данных и основы размерного моделирования

Книга начинается с основ хранения данных, анализа данных и размерного моделирования. Мы исследуем компоненты всей архитектуры DW/BI и устанавливаем основной словарь, который будет использоваться в книге. Будут развеяны некоторые мифы и заблуждения о размерном моделировании.

Глава 2. Обзор методов размерного моделирования Кимбалла

Эта глава описывает более 75 методов размерного моделирования и паттернов. Этот официальный список методов Кимбалла включает указатели на последующие главы, где методы воплощены на примерах.

Глава 3. Розничные продажи

Розничные продажи — это классический пример, используемый для иллюстрирования размерного моделирования. Мы начинаем с классики, поскольку это то, что все понимают. Надеемся, вам не понадобится глубоко задумываться об отрасли, поскольку мы хотим, чтобы вы сфокусировались на основных понятиях размерного моделирования. Мы начнем с обсуждения четырехэтапного процесса разработки размерных моделей. Мы подробно изучим таблицы измерений, включая измерение «Дата», которое будет регулярно повторно использоваться в книге. Мы также обсудим вырожденные измерения, понятие «снежинка» и суррогатные ключи. Даже если вы не разрабатываете проект в сфере продаж, эта глава важна, поскольку наполнена основными принципами.

Глава 4. Склад

Мы останемся в отрасли розничных продаж для изучения следующего примера, но обратим ваше внимание к другому бизнес-процессу. Эта глава представляет архитектуру шины предприятия по хранению данных и матрицу шины

с согласованными измерениями. Эти понятия очень важны для любого, кто стремится к созданию интегрированной и расширяемой архитектуры DW/BI. Мы также сравним три основных типа таблиц фактов: транзакционные таблицы, таблицы фактов периодических моментальных снимков и таблицы фактов кумулятивных моментальных снимков.

Глава 5. Закупки

Эта глава укрепляет важность рассмотрения цепочки ценностей организации в процессе подготовки среды DW/BI. Мы также исследуем серии базовых и расширенных методов обработки признаков медленно меняющихся измерений и будем основываться на изменениях типа 1 (перезапись), типа 2 (добавить строку) и типа 3 (добавить столбец), когда будем представлять читателям тип 0 и типы с 4 по 7.

Глава 6. Управление заказами

В этом примере мы изучим бизнес-процессы, которые, как правило, реализуются первыми в системах DW/BI, поскольку они предоставляют основные метрики исполнения бизнеса — что мы продаем каждому клиенту и по какой цене? Мы обсудим измерения, играющие множество ролей в этой схеме. Мы также исследуем наиболее частые проблемы, с которыми сталкиваются создатели моделей при работе с информацией об управлении заказами, такими как соглашения об именовании заголовков/строк, множественные значения атрибутов или единицы измерения и мусорные измерения с разными операционными индикаторами.

Глава 7. Бухгалтерский учет

Мы обсудим моделирование главной бухгалтерской книги для хранилищ данных в этой главе. Мы опишем соответствующую обработку фактов YTD (Year-to-date) и нескольких фискальных календарей, а также объединенные таблицы данных, которые объединяют данные различных бизнес-процессов. Мы также предоставим детальное руководство по иерархии признаков измерений, начиная с простых денормализованных фиксированных глубоких иерархий и заканчивая сопоставительными таблицами для навигации по более сложным рваным переменным глубоких иерархий.

Глава 8. Управление взаимоотношениями с клиентами

Многие системы DW/BI были построены на предпосылке, что вы должны лучше понимать, чтобы обслужить своих клиентов. В этой главе рассматриваются измерение «Клиент», включая стандартизацию адреса и сопоставительные таблицы

для многозначных признаков измерений. Мы также описываем паттерны моделирования сложного поведения потребителей, такие как консолидация данных клиентов из нескольких источников.

Глава 9. Управление персоналом

В этой главе рассматривается несколько уникальных аспектов моделей измерений «Управление персоналом», включая ситуацию, в которой таблица измерений начинает вести себя как таблица фактов. Мы обсуждаем комплексные аналитические решения, обработку рекурсивных иерархий управления и опросные анкеты. Сравниваются несколько методов обработки многозначных атрибутов.

Глава 10. Финансовые услуги

В банковском примере рассматривается концепция схем супертипа и подтипа для разнородных продуктов, в которых каждая сфера деятельности имеет уникальные описательные признаки и показатели эффективности. Очевидно, что необходимость обрабатывать разнородные продукты неуникальна для финансовых услуг. Мы также обсуждаем сложные отношения между счетами, клиентами и домашними хозяйствами.

Глава 11. Телекоммуникации

Эта глава структурирована несколько иначе, чтобы побудить вас критически мыслить при выполнении анализа архитектуры размерной модели. Начнем с размерного дизайна, который на первый взгляд выглядит правдоподобно. Сможете ли вы найти проблемы? Кроме того, мы исследуем особенности географических измерений местоположения.

Глава 12. Транспортировка

В этом примере мы рассмотрим связанные таблицы фактов на разных уровнях детализации, указывая на уникальные характеристики таблиц фактов, описывающих сегменты в путешествии или сети. Мы более детально рассмотрим измерения даты и времени, охватывая специфичные для страны календари и синхронизацию между несколькими часовыми поясами.

Глава 13. Образование

В этой главе мы рассмотрим несколько таблиц фактов без метрик. Кроме того, мы исследуем таблицы фактов кумулятивных моментальных снимков для обработки заявок студентов и заявок на гранты на исследования. В этой главе вы оцените разнообразие бизнес-процессов в образовательном учреждении.

Глава 14. Здравоохранение

Некоторые из самых сложных моделей, с которыми мы когда-либо работали, относятся к сфере здравоохранения. Эта глава иллюстрирует обработку таких трудных случаев, включая использование соединительных таблиц для моделирования множества диагнозов и поставщиков медицинских услуг, которые осуществляют лечение пациентов.

Глава 15. Электронная коммерция

Эта глава фокусируется на нюансах веб-данных о потоке кликов, включая их уникальные измерения. Мы также представляем измерение «Шаг», которое используется для лучшего понимания любого процесса, состоящего из последовательных шагов.

Глава 16. Страхование

Последний пример объединяет многие паттерны, которые мы обсуждали ранее в книге, в единый набор взаимосвязанных схем. Эту главу можно рассматривать как главу, которая сводит все воедино, поскольку методы моделирования накладываются друг на друга.

Глава 17. Обзор жизненного цикла Кимбалла

Теперь, когда вам удобно разрабатывать размерные модели, мы предлагаем общий обзор действий, с которыми сталкиваются в течение срока службы типичного проекта DW/BI. Эта глава представляет собой краткий обзор второго издания книги «Инструментарий хранилища данных», над которым мы работали в соавторстве с Бобом Беккером, Джой Манди и Уорреном Торнтуэйтом.

Глава 18. Процессы и задачи размерного моделирования

В этой главе изложены конкретные рекомендации по решению задач размерного моделирования в рамках жизненного цикла Кимбалла. Первые 16 глав этой книги охватывают методы размерного моделирования и паттерны проектирования, здесь же описываются обязанности, инструкции и результаты деятельности по проектированию размерного моделирования.

Глава 19. Подсистемы и методы ETL

Система извлечения, преобразования и загрузки потребляет непропорционально большую долю времени и усилий, необходимых для создания среды DW/BI. Тщательное рассмотрение передового опыта выявило 34 подсистемы, найденные в основе почти каждого хранилища данных. Эта глава начинается с требований

и ограничений, которые необходимо учитывать перед проектированием системы ETL, а затем она описывает 34 подсистемы извлечения, очистки, согласования, доставки и управления.

Глава 20. Задачи и процессы разработки и проектирования систем ETL

Эта глава посвящена конкретным тактическим действиям и антипаттернам, связанным с проектированием и разработкой ETL. Она обязательна к прочтению всем, кто отвечает за обязанности ETL.

Глава 21. Аналитика больших данных

В заключительной главе мы сосредоточимся на популярной теме больших данных (Big Data). Мы считаем, что большие данные — естественное продолжение ваших обязанностей в области DW/BI. Мы начнем с обзора нескольких архитектурных альтернатив, в том числе MapReduce и Hadoop, и опишем, как эти альтернативы могут сосуществовать с вашей текущей архитектурой DW/BI. Затем мы рассмотрим лучшие практики управления, архитектуры, моделирования и управления большими данными.

Веб-сайты

Сайт Kimball Group наполнен дополнительными материалами и ресурсами для моделирования:

1. Зарегистрируйтесь для получения советов Кимбалла по дизайну, чтобы получить практическое руководство по размерному моделированию и темам DW/BI.
2. Получите доступ к архиву из более чем 300 советов и статей по архитектуре.
3. Узнайте о публичных и выездных занятиях в Университете Кимбалла по качественному, независимому от поставщиков обучению в соответствии с нашим опытом и публикациями.
4. Узнайте о консалтинговых услугах Kimball Group, чтобы использовать наш многолетний опыт DW/BI.
5. Задайте вопросы другим осведомленным в измерении участникам на форуме Kimball.

Выводы

Цель этой книги — рассказать о принятых в индустрии и проверенных годами методах размерного моделирования и разработки, основанных на более чем 60-летнем опыте авторов и извлеченных уроках в реальной бизнес-среде. DW/BI-системы должны основываться на потребностях бизнес-клиентов и поэтому разрабатываться и представляться с простой размерной точки зрения. Мы уверены, что вы окажетесь на один гигантский шаг ближе к успеху в области DW/BI, если купите это издание.

Теперь, когда вы знаете, куда направляетесь, пришло время погрузиться в детали. Мы начнем с учебника по DW/BI и размерному моделированию в главе 1, чтобы убедиться, что все находятся на одной волне в отношении ключевых терминов и архитектурных концепций.

1 Хранение данных, анализ данных и основы размерного моделирования

Эта глава закладывает основу для следующих. Мы начнем изучение систем хранения и анализа данных (data warehousing and business intelligence, DW/BI) с перспективы высокого уровня. Вы можете быть разочарованы, узнав, что мы не начинаем с технологий и инструментов, но прежде всего система DW/BI должна учитывать потребности бизнеса. Сформулировав четкие потребности бизнеса, мы двинемся в обратном направлении через логический, а затем физический дизайн наряду с решениями по поводу технологии и инструментов.

В этой главе мы закладываем основы понимания в отношении целей хранения и анализа данных, наблюдая при этом странное сходство между обязанностями менеджера DW/BI и издателя.

С этой высокоуровневой перспективы мы исследуем основные концепции размерного моделирования и установим основной словарь. В этой главе мы обсудим основные компоненты DW/BI-архитектуры Кимбалла, а также сравним альтернативные архитектурные подходы. К счастью, независимо от архитектуры вашей системы, в ней всегда найдется место для размерного моделирования. Наконец, мы рассмотрим общие мифы о моделировании. К концу этой главы вы поймете, что вам нужно быть наполовину администратором базы данных (DBA) и наполовину бизнес-аналитиком (MBA) при работе с проектом DW/BI.

В главе 1 обсуждаются следующие понятия:

- бизнес-ориентированные цели DW/BI;
- основные понятия и словарь размерного моделирования, включая таблицы фактов и измерений;
- компоненты и принципы DW/BI архитектуры Кимбалла;

- сравнение альтернативных архитектур DW/BI и роли размерного моделирования в каждой из них;
- заблуждения о размерном моделировании.

Различные миры сбора и анализа данных

Один из важнейших активов любой организации — ее данные. Этот актив почти всегда используется для двух целей: ведение оперативного учета и принятие аналитических решений. Проще говоря, операционные системы (OLTP) — это то место, куда вы вводите данные, а система DW/BI — это то место, откуда вы извлекаете данные.

Пользователи операционной системы крутят колеса организации. Они принимают заказы, регистрируют новых клиентов, отслеживают состояние оперативной деятельности и регистрируют жалобы. Операционные системы оптимизированы для быстрой обработки транзакций. Эти системы почти всегда имеют дело с записью одной транзакции за раз. Они предсказуемо выполняют одни и те же операционные задачи снова и снова, выполняя бизнес-процессы организации. С учетом этого факта операционные системы обычно не поддерживают историю изменения данных, а обновляют их, чтобы отразить наиболее актуальное состояние.

Пользователи DW/BI-систем, с другой стороны, наблюдают за показателями организации, чтобы оценить производительность. Они подсчитывают новые заказы, сравнивают их с заказами прошлой недели и спрашивают, почему новые клиенты зарегистрировались и на что они жаловались. Они беспокоятся, правильно ли выполняются рабочие процессы. Хотя им нужны подробные данные для поддержки их постоянно меняющихся вопросов, пользователи DW/BI почти никогда не имеют дела с одной транзакцией за раз. Эти системы оптимизированы для высокопроизводительных запросов, поскольку вопросы пользователей часто требуют поиска и сжатия сотен или сотен тысяч транзакций в набор ответов. Чтобы еще больше усложнить ситуацию, пользователи DW/BI-систем обычно требуют сохранения исторического контекста для точной оценки эффективности работы организации с течением времени.

В первом издании «Инструментарий хранилища данных» Ральф Кимбалл посвятил целую главу описанию дихотомии между мирами оперативной обработки и хранения данных. В настоящее время широко признано, что система DW/BI имеет совершенно другие потребности, клиентов, структуры и ритмы, нежели операционные системы записи информации. К сожалению, мы все еще сталкиваемся с так называемыми системами DW/BI, которые скорее просто копии операционных систем записи, хранящихся на отдельной аппаратной платформе. Хотя эти среды могут учитывать необходимость изолировать

операционную и аналитическую среды по соображениям производительности, они ничего не делают для устранения других внутренних различий между двумя типами систем. Бизнес-клиенты не в восторге от удобства использования и производительности, предоставляемых этими псевдохранилищами данных. Такие самозванцы оказывают плохую услугу DW/BI, потому что они не признают, что у их клиентов потребности совершенно не такие, как у пользователей операционной системы.

Цели хранения и анализа данных

Прежде чем мы углубимся в детали размерного моделирования, полезно сосредоточиться на основных целях хранения и анализа данных. Цели легко можно сформулировать, прогулявшись по холлу любой организации и прислушавшись к менеджерам. Эти повторяющиеся темы существуют уже более трех десятилетий:

- «Мы собираем тонны данных, но не можем получить к ним доступ»;
- «Нам нужно сгруппировать и отфильтровать данные различными способами»;
- «Люди, принимающие важные для бизнеса решения (бизнес-клиенты), должны иметь доступ к необходимой им информации»;
- «Просто покажи мне, что важно»;
- «Мы проводим целые собрания, где спорим, у кого верные цифры, а не принимаем решения»;
- «Мы хотим, чтобы люди использовали информацию для поддержки принятия решений на основе фактов».

Исходя из нашего опыта скажем: эти проблемы все еще настолько универсальны, что определяют основные требования к системе DW/BI. Теперь превратим эти предложения по управлению бизнесом в требования.

- **Система DW/BI должна сделать информацию доступной.**
  Содержание системы DW/BI должно быть понятным. Данные должны быть интуитивно понятными и очевидными для бизнес-клиента, а не только для разработчика. Структуры и метки данных должны имитировать мыслительные процессы и словарь бизнес-клиентов. Бизнес-клиенты хотят разделять и объединять аналитические данные в бесконечных комбинациях. Инструменты и приложения аналитики данных, обеспечивающие доступ к данным, должны быть простыми и удобными в использовании. Они также должны возвращать результаты запроса пользователю с минимальным

временем ожидания. Мы можем обобщить это требование, сказав «просто и быстро».

- **Система DW/BI должна представлять информацию последовательно**. Данные в системе DW/BI должны быть достоверными, тщательно собранными из различных источников, очищенными, проверенными на качество и выпущенными только тогда, когда они пригодны для использования пользователем. Согласованность также подразумевает использование общих меток и определений для содержимого системы DW/BI, которые используются в разных источниках данных. Если два показателя производительности имеют одинаковое имя, они должны означать одно и то же. И наоборот, если две меры не означают одно и то же, они должны быть помечены по-разному.

- **Система DW/BI должна адаптироваться к изменениям.** Потребности пользователей, условия бизнеса, данные и технологии могут меняться. Система DW/BI должна быть спроектирована так, чтобы корректно обрабатывать это неизбежное изменение и чтобы не превращать в недействительные существующие данные или приложения. Существующие данные и приложения не должны изменяться или разрушаться, когда бизнес-сообщество задает новые вопросы или когда новые данные добавляются в хранилище. Наконец, если необходимо изменить описательные данные в системе DW/BI, вы должны соответствующим образом учесть эти изменения и сделать их прозрачными для пользователей.

- **Система DW/BI должна предоставлять информацию своевременно.** Поскольку система DW/BI более интенсивно используется для принятия оперативных решений, сырые данные, возможно, придется преобразовать в полезную информацию за часы, минуты или даже секунды. Команда DW/BI и бизнес-клиенты должны иметь реалистичные ожидания в отношении того, что означает предоставление данных, когда у них мало времени для их очистки или проверки.

- **Система DW/BI должна быть бастионом, защищающим информационные активы.** Информационные «жемчужины» организации помещаются в хранилище данных. Как минимум это хранилище содержит информацию о том, что и кому вы продаете, а также по какой цене — это потенциально опасные детали в руках не тех людей. Система DW/BI должна эффективно контролировать доступ к конфиденциальной информации организации.

- **Система DW/BI должна служить авторитетной и заслуживающей доверия основой для более эффективного принятия решений.**
  Хранилище данных должно иметь правильные данные для поддержки принятия решений. Наиболее важные результаты работы системы DW/BI — решения, которые принимаются на основе представленных аналитических данных. Эти решения обеспечивают влияние на бизнес и ценность, придаваемую системе DW/BI. Первоначальное название, предшествующее названию DW/BI, — по-прежнему лучшее описание того, что вы разрабатываете: система поддержки принятия решений.

- **Деловое сообщество должно принять систему DW/BI, чтобы считать ее успешной.**
  Неважно, что вы создали элегантное решение, используя лучшие в своем роде продукты и платформы. Если бизнес-сообщество не принимает эту DW/BI-среду и не использует ее активно, вы не прошли вступительный тест. В отличие от реализации операционной системы, где у бизнес-клиентов нет выбора, кроме как использовать новую систему, использование DW/BI иногда необязательно. Бизнес-клиенты будут использовать систему DW/BI, если это простой и быстрый источник полезной информации.

Хотя каждое требование в этом списке значимо, последние два наиболее важны и, к сожалению, часто игнорируются. Успешное хранение и анализ данных требуют большего, чем быть звездным архитектором, техником, разработчиком моделей или администратором базы данных. Благодаря инициативе DW/BI вы одной ногой стоите в своей зоне комфорта информационных технологий (ИТ), в то время как ваша другая нога находится в незнакомой среде деловых пользователей. Вы должны управлять обеими, модифицируя некоторые проверенные навыки, чтобы приспособиться к уникальным требованиям DW/BI. Очевидно, что вам нужно иметь недюжинный талант, чтобы вести себя так, как будто вы гибрид DBA и MBA.

Сравнение обязанностей менеджеров DW/BI с издательским бизнесом

Исходя из целей DW/BI давайте сравним обязанности менеджеров DW/BI с обязанностями главного редактора. Как редактор высококачественного журнала, вы будете иметь широкие возможности по управлению содержанием, стилем и доставкой журнала. Любой, кто занимает эту должность, занимается следующими видами деятельности:

- Понимать читателей:
  - определить их демографические характеристики;
  - найти, что читатели хотят видеть в этом журнале;
  - определить «лучших» читателей, которые обновят свои подписки и купят товары у рекламодателей журнала;
  - найти потенциальных новых читателей и ознакомить их с журналом.
- Убедиться, что журнал обращается к читателям:
  - выбирать интересный и востребованный контент;
  - принимать решения о макете и визуализации, чтобы максимизировать удовольствие читателей;
  - поддерживать высококачественные стандарты письма и редактирования, придерживаясь единого стиля представления;
  - постоянно контролировать точность статей и претензий рекламодателей;
  - адаптироваться к изменяющимся профилям читателей и обеспечить доступность новых материалов от писателей и участников.
- Поддерживать публикации:
  - привлекать рекламодателей и вести журнал с прибылью;
  - публиковаться на регулярной основе;
  - поддерживать доверие читателей;
  - удовлетворять требования владельцев бизнеса.

Вы также можете определить пункты, которые должны быть нецелевыми для главного редактора. Это может быть создание журнала на основе конкретной технологии печати или использование энергии руководства исключительно для повышения эффективности работы. Сюда отнесем навязывание технического стиля письма, который читатели не могут легко понять, или создание запутанного, перегруженного макета, который трудно читать.

Если издательский бизнес построен на основе эффективного обслуживания читателей, то журнал, вероятно, будет успешным. И наоборот: просмотрите список и представьте, что произойдет, если вы пропустите какой-либо элемент — в конце концов у журнала начнутся серьезные проблемы.

Можно провести яркие параллели между обычным издателем и менеджером DW/BI. Руководствуясь потребностями бизнеса, менеджеры DW/BI должны публиковать данные, которые были собраны из различных источников и отредактированы для обеспечения качества и согласованности. Основная обязанность заключается в обслуживании читателей, иначе известных как бизнес-клиенты. Образ публикации подчеркивает необходимость сосредоточиться на ваших клиентах, а не просто фокусироваться на продуктах и процессах. Хотя вы используете технологию для доставки системы DW/BI, эта технология — в лучшем случае

средство для достижения цели. Таким образом, технологии и методы, использованные для разработки системы, не должны прямо отражаться в ваших главных должностных обязанностях.

Теперь измените обязанности издателя журнала на обязанности менеджера DW/BI:

- Понимать бизнес-клиентов:
  - понимать свои должностные обязанности, цели и задачи;
  - определять решения, которые бизнес-клиенты хотят принимать с помощью системы DW/BI;
  - определять «лучших» пользователей, которые принимают эффективные решения;
  - найти потенциальных новых пользователей и ознакомить их с возможностями системы DW/BI.
- Предоставлять бизнес-клиентам высококачественную, актуальную и доступную информацию и аналитику:
  - выбирать наиболее надежные и действенные данные для представления в системе DW/BI, тщательно отобранные из множества возможных источников данных в вашей организации;
  - сделать пользовательские интерфейсы и приложения простыми и управляемыми паттернами, явно согласованными с профилями когнитивной обработки пользователей;
  - убедиться, что данные точны и им можно доверять, последовательно маркируя их по всему предприятию;
  - постоянно контролировать точность данных и анализа;
  - адаптироваться к изменяющимся профилям пользователей, требованиям и бизнес-приоритетам, а также обеспечивать доступность новых источников данных.
- Поддерживать среду DW/BI:
  - использовать часть кредита для бизнес-решений, принятых с использованием системы DW/BI, и использовать эти успехи, чтобы оправдать расходы на персонал и текущие расходы;
  - регулярно обновлять систему DW/BI;
  - поддерживать доверие бизнес-клиентов;
  - удовлетворять требования бизнес-клиентов, исполнительных спонсоров и ИТ-менеджмента.

Если вы хорошо справитесь со всеми этими обязанностями, вы станете отличным менеджером DW/BI! И наоборот: просмотрите список и представьте, что

произойдет, если вы пропустите какой-либо пункт — в итоге у среды начнутся серьезные проблемы. Теперь сопоставьте этот взгляд на работу менеджера DW/BI с собственным описанием работы. Скорее всего, предыдущий список больше ориентирован на пользователей и бизнес-проблемы и может даже не походить на работу в сфере ИТ. По нашему мнению, именно это делает хранение и анализ данных интересными.

Введение в размерное моделирование

Теперь, когда вы понимаете цели системы DW/BI, давайте рассмотрим основы размерного моделирования. Размерное моделирование широко признано в качестве предпочтительного метода представления аналитических данных, поскольку оно одновременно учитывает два требования:

- предоставлять данные, понятные бизнес-клиентам;
- обеспечивать высокую производительность выполнения запросов.

Размерное моделирование — это давно разработанный метод, упрощающий работу с базами данных. В каждом случае в течение более пяти десятилетий ИТ-организации, консультанты и бизнес-клиенты естественным образом стремились к простой размерной структуре, чтобы соответствовать фундаментальной потребности человека в простоте. Простота имеет решающее значение, поскольку она гарантирует, что пользователи смогут легко понять данные, а также позволяет программному обеспечению быстро и эффективно осуществлять навигацию и возвращать результаты.

Представьте себе руководителя, который описывает свой бизнес следующим образом: «Мы продаем продукты на различных рынках и измеряем наши результаты с течением времени». Разработчики моделей внимательно прислушиваются к акценту на продукт, рынок и время. Большинство людей интуитивно представляют такой бизнес как куб данных, грани которого обозначены как «Продукт», «Рынок» и «Время». Представьте себе нарезку вдоль каждого из этих измерений. Точки внутри куба — это места, где хранятся измерения, такие как объем продаж или прибыль для этой комбинации продукта, рынка и времени. Способность визуализировать что-то настолько абстрактное, как набор данных, конкретным и осязаемым способом — секрет понятности. Если эта перспектива кажется слишком простой, хорошо! Модель данных, которая начинается с простого, имеет шанс остаться простой в конце проекта. Модель, которая начинается сложнее, несомненно, будет слишком сложной в конце, что приведет к снижению производительности выполнения запросов и сокращению количества бизнес-клиентов. Альберт Эйнштейн сформировал основную философию,

управляющую размерным дизайном, когда сказал: «Все следует упрощать до тех пор, пока это возможно, но не более того».

Хотя размерные модели часто создаются в системах управления реляционными базами данных, они довольно сильно отличаются от моделей третьей нормальной формы (3NF), которые стремятся устранить избыточность данных. Нормализованные структуры 3NF делят данные на множество отдельных объектов, каждый из которых становится реляционной таблицей. База данных заказов на продажу может начинаться с записи для каждой строки заказа, но превращаться в сложную паутину в виде модели 3NF, возможно, состоящей из сотен нормализованных таблиц.

Промышленность иногда называет 3NF-модели моделями отношения сущностей (ER). Диаграммы отношений сущностей (ER-диаграммы или ERD) — это чертежи, которые сообщают о взаимосвязях между таблицами. И 3NF-модели, и размерные модели могут быть представлены в виде ER-диаграмм, поскольку состоят из объединенных реляционных таблиц. Ключевое различие между 3NF-моделями и размерными моделями заключается в степени нормализации. Поскольку оба типа моделей могут быть представлены как ERD, мы не будем называть 3NF-модели ER-моделями. Вместо этого мы называем их нормализованными моделями, чтобы минимизировать путаницу.

Нормализованные структуры 3NF чрезвычайно полезны для оперативной обработки, поскольку транзакция обновления или вставки затрагивает базу данных только в одном месте. Нормализованные модели, однако, слишком сложны для запросов анализа данных. Пользователи не могут понимать, перемещаться или помнить нормализованные модели, которые напоминают карту системы автострад Лос-Анджелеса. Аналогичным образом большинство реляционных СУБД не могут эффективно запрашивать нормализованную модель. Сложность непредсказуемых запросов пользователей перегружает оптимизаторы базы данных, что приводит к катастрофическим результатам.

ПРИМЕЧАНИЕ Размерная модель содержит ту же информацию, что и нормализованная модель, но упаковывает данные в формате, который обеспечивает понятность для пользователя, производительность запросов и устойчивость к изменениям.

Схема «звезда» против кубов OLAP

Размерные модели, реализованные в реляционных СУБД, называются схемами типа «звезда» из-за их сходства со звездообразной структурой. Размерные модели, реализованные в размерных средах баз данных, называются кубами

оперативной аналитической обработки (OLAP — online analytical processing), как показано на рис. 1-1.

Если ваша среда DW/BI включает в себя схемы типа «звезда» или кубы OLAP, она управляет концепциями измерений. И звезды, и кубы имеют общий логический дизайн с узнаваемыми измерениями, однако физическая реализация отличается.

Когда данные загружаются в куб OLAP, они сохраняются и индексируются с использованием форматов и методов, разработанных для размерных данных. Агрегирование или предварительно рассчитанные сводные таблицы часто создаются и управляются механизмом куба OLAP. Следовательно, кубы обеспечивают превосходную производительность запросов благодаря предварительным расчетам, стратегиям индексации и другим способам оптимизации. Бизнес-клиенты могут производить свертку и развертку, добавляя или удаляя атрибуты из своих анализов с превосходной производительностью, не выполняя новых запросов. Кубы OLAP также предоставляют более аналитически устойчивые функции, которые превосходят функции, доступные в SQL. Недостатком в том, что вы платите производительностью за эти возможности, особенно с большими наборами данных.

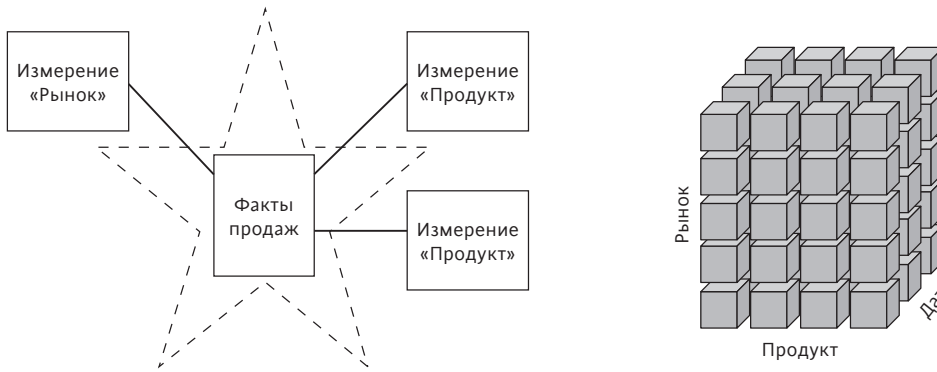

Рис. 1-1. Схема «звезда» в сравнении с кубом OLAP

К счастью, большинство рекомендаций в этой книге не зависят от платформы размерных реляционных баз данных. Хотя возможности технологии OLAP постоянно совершенствуются, мы обычно рекомендуем загружать подробную атомарную информацию в схему типа «звезда», дополнительные кубы OLAP затем заполняются из схемы типа «звезда». По этой причине большинство методов размерного моделирования в этой книге сформулированы в терминах реляционной схемы «звезда».

Соглашение об именовании при развертке OLAP

При развертывании данных в кубах OLAP необходимо учитывать некоторые моменты:

1. Схема типа «звезда», размещенная в реляционной базе данных, — это хорошая физическая основа для создания куба OLAP и обычно рассматривается как более стабильная основа для резервного копирования и восстановления.
2. Кубы OLAP традиционно отличались исключительными преимуществами в производительности по сравнению с реляционными СУБД, но это различие стало менее важным благодаря достижениям в области компьютерного оборудования, такого как устройства и базы данных в памяти, а также в области программного обеспечения реляционных СУБД, такого как столбчатые базы данных.
3. Структуры данных куба OLAP сильнее отличаются у разных поставщиков, чем реляционные СУБД, поэтому подробные сведения о развертывании часто зависят от того, какой поставщик OLAP выбран. Как правило, переносить приложения BI между различными инструментами OLAP сложнее, чем переносить приложения BI на разные реляционные базы данных.
4. Кубы OLAP обычно предлагают более сложные параметры безопасности, чем СУБД, например ограничивают доступ к подробным данным, но предоставляют более открытый доступ к сводным данным.
5. Кубы OLAP предлагают значительно более широкие возможности анализа, чем реляционные СУБД, которые обременены ограничениями SQL. Это может быть основным оправданием использования продукта OLAP.
6. Кубы OLAP исправно поддерживают медленно меняющиеся измерения типа изменения 2 (которые обсуждаются в главе 5), но кубы часто необходимо обрабатывать частично или полностью с использованием альтернативных техник медленно меняющихся измерений, когда данные перезаписываются.
7. Кубы OLAP стабильно поддерживают транзакционные таблицы фактов и таблицы фактов периодических моментальных снимков, но не обрабатывают таблицы фактов кумулятивных моментальных снимков из-за ограничений на перезапись данных, описанных в предыдущем пункте.
8. Кубы OLAP обычно поддерживают сложные рваные иерархии неопределенной глубины, такие как организационные диаграммы или ведомости материалов, с использованием собственного синтаксиса запросов, превосходящего по функциональным возможностям подходы, применяемые в РСУБД.

9. Кубы OLAP могут накладывать детальные ограничения на структуру ключей измерений, которые реализуют более детальные иерархии по сравнению с реляционными базами данных.
10. Некоторые продукты OLAP не поддерживают размерные роли или псевдонимы, поэтому необходимо определять отдельные физические измерения.

Мы вернемся в мир размерного моделирования на реляционной платформе, рассматривая два ключевых компонента схемы типа «звезда».

Таблицы фактов для измерений

Таблица фактов в размерной модели хранит измерения производительности, полученные в результате событий бизнес-процессов организации. Вы должны стараться хранить данные самых детальных измерений, возникающие в результате бизнес-процесса, в одной размерной модели. Поскольку эти данные в подавляющем большинстве — самые большие наборы данных, их не следует реплицировать в нескольких местах для нескольких организационных функций по всему предприятию. Предоставление бизнес-клиентам из нескольких организаций доступа к единому централизованному репозиторию для каждого набора данных измерений обеспечивает использование согласованных данных по всему предприятию.

Термин «факт» представляет собой бизнес-измерение. Представьте себе, что вы стоите на рынке, наблюдаете, как продаются продукты, и записываете количество единиц товара и объем продаж в долларах для каждого продукта в каждой транзакции. Эти измерения регистрируются при сканировании продуктов в регистре, как показано на рис. 1-2.

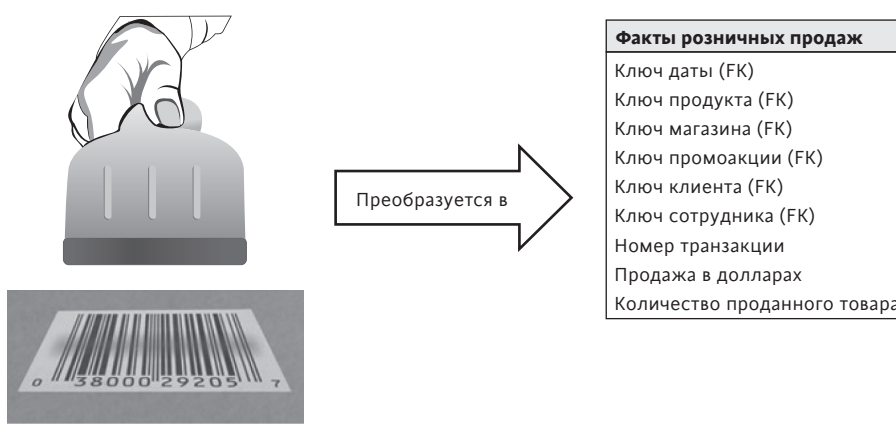

Рис. 1-2. События измерения бизнес-процессов переводятся в таблицы фактов