

УДК 004.6
ББК 32.973.2-018.2
К40

The Data Warehouse Toolkit: The Definitive Guide to
Dimensional Modeling, 3rd edition

Ralph Kimball and Margy Ross

Copyright © 2013 by Ralph Kimball and Margy Ross

All rights reserved. This translation published under license with the original
publisher John Wiley and Sons, Inc. via Igor Korzhenevskiy
of Alexander Korzhenevski Agency.

Wiley and the Wiley logo are trademarks or registered trademarks of John
Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries,
and may not be used without written permission.

All other trademarks are the property of their respective owners. John Wiley & Sons,
Inc. is not associated with any product or vendor mentioned in this book.

Кимбалл, Ральф.

К40 Инструментарий хранения и анализа данных : полное руко-
водство 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 по-
ведении 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

© Райтман М.А., перевод на русский язык, 2024
© Оформление. ООО «Издательство «Эксмо», 2024

Оглавление

Благодарности	23
Введение	24
Для кого эта книга	25
Структура книги	26
Глава 1. Хранение данных, анализ данных и основы размерного моделирования	27
Глава 2. Обзор методов размерного моделирования Кимбалла . . .	27
Глава 3. Розничные продажи	27
Глава 4. Склад	27
Глава 5. Закупки	28
Глава 6. Управление заказами	28
Глава 7. Бухгалтерский учет	28
Глава 8. Управление взаимоотношениями с клиентами	28
Глава 9. Управление персоналом	29
Глава 10. Финансовые услуги	29
Глава 11. Телекоммуникации	29
Глава 12. Транспортировка	29
Глава 13. Образование	29
Глава 14. Здравоохранение	30
Глава 15. Электронная коммерция	30
Глава 16. Страхование	30
Глава 17. Обзор жизненного цикла Кимбалла	30
Глава 18. Процессы и задачи размерного моделирования	30
Глава 19. Подсистемы и методы ETL	30
Глава 20. Задачи и процессы разработки и проектирования систем ETL	31
Глава 21. Аналитика больших данных	31
Веб-сайты	31
Выводы	32

1	Хранение данных, анализ данных и основы размерного моделирования	33
	Различные миры сбора и анализа данных	34
	Цели хранения и анализа данных	35
	Сравнение обязанностей менеджеров DW/BI с издательским бизнесом	37
	Введение в размерное моделирование	40
	Схема «звезда» против кубов OLAP	41
	Таблицы фактов для измерений	44
	Таблицы измерений для описательного контекста	47
	Факты и измерения, объединенные в схему «звезда»	50
	DW/BI-архитектура Кимбалла	53
	Операционные исходные системы	53
	Система извлечения, преобразования и загрузки	54
	Область представления для поддержки анализа данных	56
	Приложения по анализу данных	58
	Архитектура Кимбалла и метафора ресторана	58
	Альтернативные архитектуры DW/BI	62
	Независимая архитектура Data Mart («Витрина данных»)	62
	Веерная архитектура корпоративной информационной фабрики Инмона	64
	Гибридная веерная и кимбалловская архитектура	66
	Мифы о размерном моделировании	67
	Миф 1: размерные модели только для сводных данных	67
	Миф 2: размерные модели для отделов, а не для предприятий в целом	67
	Миф 3: размерные модели не масштабируемы	68
	Миф 4: размерные модели только для заранее определенного использования	68
	Миф 5: размерные модели не интегрируемы	69
	Еще больше причин мыслить многомерно	69
	Соглашения Agile	70
	Выводы	72
2	Обзор методов размерного моделирования Кимбалла	73
	Основные концепции	73
	Соберите бизнес-требования и реалии данных	74
	Совместные рабочие сессии по пространственному моделированию	74

Четырехэтапный процесс размерного проектирования	75
Бизнес-процессы	75
Зернистость	76
Измерения для описательного контекста	76
Факты для измерений	77
Схема «звезда» и кубы OLAP	78
Стабильные расширения размерных моделей	78
Основные методы работы с таблицами фактов	79
Структура таблиц фактов	79
Аддитивные, полуаддитивные и неаддитивные факты	79
Пустые значения (Null) в таблицах фактов	80
Согласованные факты	80
Таблицы фактов транзакций	81
Таблицы фактов периодических моментальных снимков	81
Накопительные таблицы фактов моментальных снимков	82
Таблицы фактов без показателей	83
Агрегированные таблицы фактов, или кубы OLAP	83
Консолидированные таблицы фактов	84
Основные методы работы с таблицами измерений	84
Структура таблицы измерений	84
Суррогатные ключи измерений	85
Натуральные, стойкие и сверхнатуральные ключи	85
Детализация	86
Вырожденные измерения	86
Денормализованные плоские измерения	87
Несколько иерархий в измерениях	87
Флаги и индикаторы как текстовые атрибуты	87
Пустые атрибуты в измерениях	88
Измерение «Календарная дата»	88
Важные ролевые изменения	89
Мусорные измерения	89
Измерения в виде «снежинки»	90
Измерения с внешней опорой	90
Интеграция через согласованные измерения	90
Согласованные измерения	91
Сжатые измерения	91
Копаем вширь	92
Цепочка значений	92
Архитектура шины корпоративного хранилища данных	92

Матрица шины корпоративного хранилища данных	93
Подробная матрица шины реализации	93
Матрица возможностей/заинтересованных сторон	94
Работа с атрибутами медленно изменяющегося измерения	94
Тип 0: сохранение оригинала	94
Тип 1: перезапись	95
Тип 2: добавление новой строки	95
Тип 3: добавление нового атрибута	96
Тип 4: добавление мини-измерения	96
Тип 5: добавление мини-измерения и внешней опоры типа 1 ...	96
Тип 6: добавление атрибута типа 1 к измерению типа 2	97
Тип 7: двойные измерения типа 1 и типа 2	97
Работа с иерархиями измерений	98
Позиционные иерархии с фиксированной глубиной	98
Иерархии с пропущенными уровнями / иерархии переменной глубины	98
Неровные иерархии/иерархии переменной глубины с соединительными таблицами иерархии	99
Рваные иерархии/иерархии переменной глубины с атрибутами пути	99
Продвинутые методы работы с таблицами фактов	100
Суррогатные ключи таблицы фактов	100
Таблицы-«сороконожки» с фактами	100
Числовые значения как атрибуты или факты	101
Факты о задержке/продолжительности	101
Заголовок/строка в таблице фактов	102
Выделенные факты	102
Таблицы фактов прибылей и убытков с выделением фактов ...	102
Факты разных валют	103
Факты с множественными единицами измерения	103
Факты текущего года (Year-to-date)	104
Многопроходный SQL, чтобы избежать объединения таблиц «факт — факт»	104
Отслеживание промежутка времени в таблицах фактов	104
Факты, появляющиеся с опозданием	105
Расширенные методы работы с измерениями	105
Соединения таблиц «измерение — измерение»	105
Многозначные измерения и соединительные таблицы	106
Многозначные соединительные таблицы, меняющиеся во времени	106

Временной ряд тега поведения	107
Исследовательские группы изучения поведения	107
Агрегированные факты как атрибуты измерения	107
Динамические диапазоны значений	108
Измерение «Текстовые комментарии»	108
Несколько часовых поясов	109
Измерения типа «Показатель»	109
Измерения «Шаг»	109
Измерения с возможностью горячей замены	110
Абстрактные общие измерения	110
Измерения «Аудит»	110
Измерения, прибывающие с опозданием	111
Схемы специального назначения	112
Схемы супертипа и подтипа для гетерогенных продуктов	112
Таблицы фактов в реальном времени	112
Схемы событий ошибок	113
3 Розничные продажи	115
Четырехэтапный процесс размерного проектирования	116
Шаг 1: выбор бизнес-процесса	116
Шаг 2: объявление зернистости	117
Шаг 3: определение измерений	118
Шаг 4: определение фактов	118
Пример использования в розничной торговле	119
Шаг 1: выбор бизнес-процесса	121
Шаг 2: объявление зернистости	121
Шаг 3: определение измерений	123
Шаг 4: определение фактов	123
Подробная информация о таблице измерений	127
Измерение «Дата»	127
Измерение «Продукт»	132
Измерение «Магазин»	136
Измерение «Промоакция»	138
Прочие измерения розничных продаж	142
Вырожденные измерения для номеров транзакций	143
Розничная схема в действии	144
Расширяемость схемы розничных продаж	145
Таблицы фактов без метрик	147

Ключи таблиц измерений и фактов	148
Суррогатные ключи таблицы измерений	148
Натуральные, стойкие и сверхнатуральные ключи	151
Суррогатные ключи вырожденного измерения	152
Умные ключи измерения «Дата»	152
Суррогатные ключи таблицы фактов	153
Сопротивление стремлению к нормализации	155
Схемы «снежинки» с нормализованными измерениями	155
Внешняя опора	158
Таблицы фактов «сороконожка» со «слишком большим количеством измерений»	159
Выводы	161
 4 Склад	163
Введение в цепочку ценности	163
Модели инвентаризации	165
Периодический моментальный снимок инвентаризации	165
Полуаддитивные факты	167
Расширенные сведения о запасах	168
Операции с запасами	169
Накопительный моментальный снимок запасов	171
Типы таблиц фактов	172
Таблицы фактов транзакций	173
Таблицы фактов периодических снимков	173
Накопительные таблицы фактов моментальных снимков	174
Задержки между этапами и количество этапов	175
Накопление обновлений снимков и кубов OLAP	175
Дополнительные типы таблиц фактов	175
Интеграция цепочки ценности	176
Архитектура шины хранилища корпоративных данных	177
Понимание архитектуры шины	177
Матрица шины корпоративного хранилища данных	179
Согласованные измерения	185
Горизонтальный анализ таблицы фактов	185
Идентичные согласованные измерения	186
Сжатие согласованного измерения с подмножеством атрибутов	187
Сжатие согласованного измерения с подмножеством строк	187
Сжатые согласованные размеры на матрице шины	189

Ограниченная согласованность	190
Важность управления данными и ответственности за данные	191
Согласованные измерения и движение к Agile	193
Согласованные факты	194
Выводы	195
 5 Закупки	197
Закупки: практический пример	197
Закупочные операции и матрица шины	198
Одна или много таблиц фактов транзакций	199
Дополнительный моментальный снимок закупок	203
Основные сведения о медленно меняющихся измерениях	204
Тип 0: сохранение оригинала	205
Тип 1: перезапись	206
Тип 2: добавление новой строки	208
Тип 3: добавление нового атрибута	211
Тип 4: добавление мини-измерения	214
Гибридные методы медленно изменяющихся измерений	217
Тип 5: добавление мини-измерения и внешней опоры типа 1 . . .	218
Тип 6: добавление атрибута типа 1 к измерению типа 2	219
Тип 7: Двойные измерения типа 1 и типа 2	220
Тип 7 для незапланированных отчетов «По состоянию на»	222
Обобщение медленно меняющихся измерений	223
Выводы	224
 6 Управление заказами	225
Матрица шины управления заказами	226
Транзакции по заказам	226
Нормализация фактов	227
Ролевые измерения	228
Еще раз об измерении «Продукт»	230
Измерение «Клиент»	233
Измерение «Сделка»	237
Вырожденное измерение для номера заказа	238
Мусорные измерения	239
Паттерн заголовков/строк, которого следует избегать	241
Несколько валют	243
Факты о транзакциях с разной зернистостью	245

Еще один паттерн заголовка/строк, которого следует избегать	247
Операции по выставлению счета	248
Показатели уровня обслуживания в виде фактов, измерений или того и другого	250
Факты о прибылях и убытках	251
Измерение «Аудит»	254
Накопление снимков для конвейера выполнения заказов	256
Расчет задержек	259
Несколько единиц измерения	259
За пределами зеркала заднего вида	261
Выводы	262
 7 Бухгалтерский учет	263
Тематическое исследование по бухгалтерскому учету и матрица шин	264
Данные Главной бухгалтерской книги	265
Периодический моментальный снимок Главной бухгалтерской книги	266
План счетов	266
Заккрытие периода	267
Факты типа «С начала года и до сегодняшнего дня» (year-to-date/YTD)	269
Пересмотр нескольких валют	269
Транзакции журнала Главной бухгалтерской книги	270
Несколько календарей финансового учета	271
Детализация по многоуровневой иерархии	272
Финансовые отчеты	273
Процесс составления бюджета	274
Иерархии атрибутов измерений	278
Позиционные иерархии с фиксированной глубиной	278
Прерывающиеся иерархии переменной глубины	279
Прерывающиеся иерархии переменной глубины	280
Совместный доступ при прерывающейся иерархии	284
Неравномерная иерархия, изменяющаяся во времени	285
Изменение прерывающихся иерархий	285
Альтернативные подходы к моделированию прерывающейся иерархии	287
Преимущества подхода с соединительной таблицей для неравномерных иерархий	289

Консолидированные таблицы фактов	290
Роль OLAP и комплексных аналитических решений	292
Выводы	293
8 Управление взаимоотношениями с клиентами	295
Обзор CRM-системы	296
Операционная и аналитическая CRM	298
Атрибуты измерения «Клиент»	300
Синтаксический анализ имени и адреса	300
Некоторые соображения по поводу интернационализации имен и адресов	303
Даты, ориентированные на клиента	306
Агрегированные факты как атрибуты измерений	307
Взаимосвязь между интеллектуальным анализом данных и системой DW/BI	310
Различные счетчики в измерениях типа 2	311
Выносное внешнее измерения для атрибутов с низкой кардинальностью	312
Соображения об иерархии клиентов	313
Соединительные таблицы для многозначных измерений	314
Соединительные таблицы для разреженных атрибутов	316
Соединительная таблица для нескольких контактов с клиентами	317
Сложное поведение клиента	318
Группы по изучению поведения для когорт	318
Измерение «Шаг» для последовательного поведения	320
Таблицы фактов временного интервала	321
Пометка таблиц фактов показателями удовлетворенности	324
Пометка таблиц фактов индикаторами ненормальных сценариев	325
Подходы к интеграции клиентских данных	326
Управление основными данными при создании единого измерения «Клиент»	326
Частичная согласованность нескольких измерений «Клиент»	328
Избегание соединений таблицы фактов с таблицами фактов	329
Проверка реальности с низкими задержками	331
Выводы	332

9	Управление персоналом	333
	Отслеживание профиля сотрудников	333
	Точное время вступления в силу и истечения срока действия	336
	Отслеживание причин изменения параметров	336
	Изменения профиля как атрибуты типа 2 или события факта	337
	Периодический снимок численности персонала	338
	Матрица шин для HR-процессов	339
	Комплексные аналитические решения и модели данных	341
	Рекурсивные иерархии сотрудников	342
	Отслеживание изменений на встроенном ключе менеджера	344
	Детализация иерархий управления: вверх и вниз	344
	Многозначные атрибуты ключевых навыков сотрудников	346
	Ключевые слова для навыков сотрудников	347
	Текстовая строка для ключевого слова навыка	348
	Данные анкеты-опросника	349
	Текстовые комментарии	350
	Выводы	351
10	Финансовые услуги	353
	Тематическое исследование банковского дела и матрица шин	354
	Рассмотрение измерений для исключения «недостаточного количества измерений»	355
	Измерение «Домохозяйство»	359
	Многозначные измерения и весовые коэффициенты	360
	Пересмотр мини-измерений	362
	Добавление мини-измерений к соединительным таблицам	364
	Динамическая группировка значений в таблицах фактов	365
	Схемы супертипов и подтипов для разнородных продуктов	366
	Супертипы и подтипы банковских продуктов с общими фактами	369
	Измерения с возможностью быстрой замены	370
	Выводы	370
11	Телекоммуникации	371
	Тематическое исследование телекоммуникаций и матрица шин	371
	Общие соображения по рассмотрению и оценке проекта	374
	Сбалансируйте бизнес-требования и исходные реалии	374
	Сосредоточьтесь на бизнес-процессах	374

Зернистость	375
Единая зернистость фактов	376
Зернистость измерений и иерархии	376
Измерение «Дата»	377
Вырожденные измерения	378
Суррогатные ключи	378
Расшифровки и описания в измерениях	379
Приверженность согласованности	379
Рекомендации по рассмотрению проекта	380
Обсуждение эскизного проекта	382
Изменение существующих структур данных	385
Измерение «Географическое положение»	386
Выводы	387
 12 Транспортировка	 389
Тематическое исследование авиакомпаний	
и матрица шин для них	390
Зернистость нескольких таблиц фактов	391
Объединение сегментов в поездки	394
Таблицы связанных фактов	395
Расширения для других отраслей промышленности	396
Грузоотправитель	396
Туристические услуги	397
Объединение коррелированных измерений	398
Класс обслуживания	398
Пункты отправки и назначения	399
Дополнительные соображения о дате и времени	401
Календари для конкретных стран в качестве внешних	
выносных измерений	401
Дата и время в нескольких часовых поясах	403
Краткое описание локализации	404
Выводы	404
 13 Образование	 405
Тематическое исследование университета и матрица шин	405
Таблицы фактов накопительных моментальных снимков	406
Конвейер кандидатов	408
Конвейер предложений по исследовательским грантам	410

Таблицы фактов без фактов	410
События приема	410
Регистрация на курсы	411
Использование объекта	415
Посещаемость студентов	416
Увеличение образовательных аналитических возможностей	417
Выводы	418
14 Здоровоохранение	419
Тематическое исследование здравоохранения и матрица шин	419
Выставление счетов и платежей по претензиям	423
Ролевое измерение «Дата»	426
Многозначные диагнозы	426
Супертипы и подтипы для запросов оплаты	429
Электронные медицинские записи	430
Измерение «Тип измерения» для разреженных фактов	431
Текстовые комментарии произвольной формы	432
Изображения	433
Использование инвентаря помещений, оборудования	433
Работа с ретроактивными изменениями	434
Выводы	434
15 Электронная коммерция	437
Источники данных для потока кликов	437
Проблемы с данными из потоков кликов	438
Размерные модели потока кликов	442
Измерение «Событие»	444
Измерение «Сеанс»	444
Измерение «Направление»	445
Таблица фактов сеанса потока кликов	446
Таблица фактов события страницы потока кликов	449
Измерение «Шаг»	452
Агрегированные таблицы фактов потока кликов	452
Google Analytics	453
Интеграция потока кликов в матрицу шин интернет-магазина	454
Прибыльность по всем каналам, включая веб	456
Выводы	460

16 Страхование	461
Изучение предметной области «Страхование»	462
Цепочка создания стоимости страхования	463
Проект матрицы шины	465
Транзакции по страховому полису	465
Ролевые измерения	466
Медленно меняющиеся измерения	467
Мини-измерения для больших или быстро меняющихся атрибутов	468
Многозначные атрибуты измерений	469
Числовые атрибуты как факты или измерения	469
Вырожденные измерения	470
Таблицы измерений с низкой кардинальностью	470
Измерение «Аудит»	470
Таблица фактов транзакций по полису	470
Гетерогенные продукты: супертипы и подтипы	471
Дополнительная стратегия, накапливающая моментальный снимок	472
Премияльный периодический страховых взносов	473
Согласованные измерения	473
Согласованные факты	474
Факты о предоплате	474
Пересмотр гетерогенных супертипов и подтипов	475
Пересмотр многозначных измерений	476
Более подробная информация об изучении страховых случаев	476
Обновленная матрица страховой шины	477
Подробная матрица шины реализации	478
Операции с претензиями	478
Транзакция в сравнении с мусорным измерением «Профиль претензии»	481
Накопительный моментальный снимок для претензий	481
Накопление моментальных снимков для сложных рабочих процессов	482
Накопительный моментальный снимок во времени	483
Моментальный снимок вместо периодического	484
Консолидированный периодический снимок полисов/претензий	484
События без фактов, связанные с несчастными случаями	485
Типичные ошибки размерного моделирования, которых следует избегать	486

Ошибка 10: размещение текстовых атрибутов в таблице фактов	487
Ошибка 9: ограничивать подробные дескрипторы для экономии места	487
Ошибка 8: разделение иерархий на несколько измерений	487
Ошибка 7: игнорировать необходимость отслеживать изменения измерений	488
Ошибка 6: решение всех проблем с производительностью с помощью большого количества оборудования	488
Ошибка 5: использование натуральных ключей для соединения измерений и фактов	489
Ошибка 4: пренебрежение декларированием и соблюдением зернистости таблиц фактов.	489
Ошибка 3: использовать отчет для разработки размерной модели	489
Ошибка 2: ожидать, что пользователи будут запрашивать нормализованные атомарные данные	490
Ошибка 1: терпеть неудачу при попытке построить согласованные измерения и факты	490
Выводы	491

17 Обзор жизненного цикла хранилища данных

по Кимбаллу	493
Дорожная карта жизненного цикла	494
Дорожная карта и верстовые столбы	495
Мероприятия по запуску жизненного цикла	496
Планирование и управление программами/проектами	496
Определение бизнес-требований	501
Отслеживание технологии жизненного цикла	508
Технический архитектурный дизайн	508
Выбор и установка продукта	511
Отслеживание данных жизненного цикла	513
Размерное моделирование	513
Физический дизайн	513
Проектирование и разработка ETL	515
Отслеживание приложений BI жизненного цикла	516
Спецификация приложения BI	516
Разработка приложений BI	517
Мероприятия по завершении жизненного цикла	517
Развертывание	518