

Star Schema完全参考手册

——数据仓库维度设计权威指南

- 维度设计原理的全面深入介绍
- 通过详细示例说明最佳实践
- 提供与体系结构无关的技术，适用于任何数据仓库

[美] Christopher Adamson 著
王红滨 王念滨 初妍 译



清华大学出版社

Star Schema 完全参考手册

——数据仓库维度设计权威指南

[美] Christopher Adamson 著

王红滨 王念滨 初妍 译

清华大学出版社

北 京

Christopher Adamson
Star Schema: The Complete Reference
EISBN: 978-0-07-174432-4
Copyright © 2011 by The McGraw-Hill Companies, Inc.

Original language published by The McGraw-Hill Companies, Inc. All Rights reserved. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

Simplified Chinese translation edition is published and distributed exclusively by Tsinghua University Press under the authorization by McGraw-Hill Education(Asia) Co., within the territory of the People's Republic of China only (excluding Hong Kong, Macao SAR and Taiwan). Unauthorized export of this edition is a violation of the Copyright Act. Violation of this Law is subject to Civil and Criminal Penalties.

本书中文简体字翻译版由美国麦格劳-希尔教育出版(亚洲)公司授权清华大学出版社在中华人民共和国境内(不包括中国香港、澳门特别行政区和中国台湾地区)独家出版发行。未经许可之出口视为违反著作权法,将受法律之制裁。未经出版者预先书面许可,不得以任何方式复制或抄袭本书的任何部分。

北京市版权局著作权合同登记号 图字: 01-2011-0932

本书封面贴有 McGraw-Hill 公司防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

Star Schema 完全参考手册——数据仓库维度设计权威指南/(美)亚当森(Adamson, C.) 著;

王红滨, 王念滨, 初妍 译. —北京: 清华大学出版社, 2012.7

书名原文: Star Schema: The Complete Reference

ISBN 978-7-302-29152-7

I. ①S… II. ①亚… ②王… ③王… ④初… III. ①数据库系统—技术手册 IV. ①TP311.13-62

中国版本图书馆 CIP 数据核字(2012)第 135063 号

责任编辑: 王 军 李维杰

封面设计: 孔祥丰

责任校对: 蔡 娟

责任印制: 王静怡

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者: 清华大学印刷厂

装 订 者: 三河市新茂装订有限公司

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 25.25 字 数: 662 千字

版 次: 2012 年 7 月第 1 版 印 次: 2012 年 7 月第 1 次印刷

印 数: 1~3000

定 价: 59.00 元

产品编号: 041894-01

作者简介

Christopher Adamson 是星型模式设计专家，并且是经验丰富的数据仓库实践者。他在世界各地演讲和授课，撰写数据仓库相关内容，并通过 Oakton 软件有限责任公司提供数据仓库咨询服务。

作为一名咨询专家，Christopher 与各个行业的客户一起工作，开发数据仓库战略，识别并确定项目优先顺序，设计并修改解决方案，以及管理实现。他服务的客户包括财富 500 强的公司、政府机关和数据仓库软件提供商。

Christopher 在世界范围内为成千上万学生讲授多维设计。他是 The Data Warehousing Institute(TDWI)的教员，为重要会议提供现场讲座，并且提供网络教学。

Christopher 撰写了许多有关数据仓库的文章和书籍。他是 *Mastering Data Warehouse Aggregates*(Wiley, 2006)一书的作者，并与人合著了 *Data Warehouse Design Solutions*(Wiley, 1998)一书。他撰写有关数据仓库的文章并为读者解答问题的博客网址为 www.StarSchemaCentral.com。

Christopher 是 Oakton 软件有限责任公司的创立者，可以通过网址 www.ChrisAdamson.net 与其联系。

致 谢

如果没有很多人的帮助，本书不可能与读者见面。首先需要感谢的是你们——阅读过我的书、参加过我的课程、通过 E-mail 向我提出过问题，以及访问过我的博客的人们。多年来，你们提出的意见和问题帮助我形成了这本有关多维设计的书籍。希望你们将发现这是一本专业的、具有实践性的并且有用的书。感谢你们的反馈，希望能够继续提出意见和建议。

能够出版本书还要感谢我的客户，过去的和现在的。感谢你们始终不停地给我提出新的挑战。成功完成这些挑战总会成为令人满意的经验。

Mc-Graw Hill 公司的 Lisa McClain 帮助将本书付梓出版。感谢她对本书所提供的帮助。感谢所有参与本书撰写、制作、销售和发行的人员。

多年来，许多人给予我机会和帮助，没有你们本书难以在今天出版。感谢 Cameron Hendershot、Marilyn Feldman、Joe Warren、Argee Mahecha、Matt Comstock、Mike Venerable 以及 David Wells。许多同行对我的事业发展起到了影响，有些影响是在不知不觉中。感谢 Jesse Baldwin、Jim Hadley、Mike Lynott (给我讲授有关“不要煮熟青蛙”的人)、Greg Jones 和 Randall Porter。还要感谢 Chris Date 和 Laura Reeves，给了我许多有用的建议。

特别感谢 Ralph Kimball，您的许多建议、支持和鼓励对于我们来说非常宝贵，您对本领域的贡献是无法估量的。

我热爱教学，感谢 The Data Warehousing Institute(TDWI)给我提供了从事教学的机会。感谢让我获益匪浅的 TDWI 的 Wayne Eckerson、Paul Kautza、Yvonne Baho、Sandra Prescott、Brenda Woodbridge、Jennifer Noland、Nancy Hanlonand 以及 Heather Flynn。还要感谢那些在此没有提到名字的，但却以专业素质和注重细节的精神组织 TDWI 事务的人们。

感谢我的朋友和家人：你们对我的友谊和爱是最珍贵的。原谅我不能在此一一列出你们的名字，但你们应该知道我想提到的是谁。特别的感激和爱献给母亲以及 Sis、Jason、Aya、Papa、Magal、Shell、Eli 和 Sofie。也将我的爱献给我的父亲，尽管您已经离开了我，但您始终与我同行。

目 录

第 I 部分 基 础

第 1 章 分析型数据库与维度设计	3
1.1 维度设计	3
1.1.1 目的	3
1.1.2 度量和环境	5
1.1.3 事实和维度	5
1.1.4 聚合维度和事实	7
1.2 星型模式	8
1.2.1 维度表	9
1.2.2 键与历史	10
1.2.3 事实表	10
1.3 星型模式的应用	10
1.3.1 查询事实	11
1.3.2 浏览维度	12
1.4 指导原则	13
1.5 本章小结	13
1.6 补充阅读	14
第 2 章 数据仓库体系结构	15
2.1 Inmon 的企业信息化工厂	15
2.2 Kimball 的维度数据仓库	17
2.3 独立型数据集市	19
2.4 体系结构和维度设计	20
2.4.1 方法的对比	21
2.4.2 公共元素	22
2.4.3 适用于本书的术语	23
2.5 本章小结	23

2.6 补充阅读	24
----------	----

第 3 章 星型模式与多维数据集

3.1 维度表特性	25
3.1.1 代理键与自然键	25
3.1.2 丰富的维度集合	27
3.1.3 维度表中的组合维度	30
3.2 事实表特性	33
3.2.1 事实表与过程	33
3.2.2 获取事实	33
3.2.3 粒度	36
3.2.4 稀疏性	37
3.2.5 退化维	37
3.3 缓慢变化维	38
3.3.1 变化类型 1	40
3.3.2 变化类型 2	42
3.3.3 选择并实现响应类型	44
3.4 多维数据集	46
3.5 本章小结	49
3.6 补充阅读	49

第 II 部分 多种星型模式

第 4 章 过程处理中的事实表

4.1 事实表与业务过程	53
4.2 具有不同时间的事实	54
4.2.1 单一事实表带来的问题	55
4.2.2 以不同事实表的方式建模	57

4.3	具有不同粒度的事实	58
4.3.1	单一事实表带来的问题	58
4.3.2	以不同事实表的方式 建模	61
4.4	从多个事实表中分析事实	61
4.4.1	连接事实表的危险	62
4.4.2	横向钻取	63
4.4.3	横向钻取的实现	67
4.5	本章小结	70
4.6	补充阅读	71
第 5 章	维度一致性问题	73
5.1	星型模式之间的协同	73
5.2	维度与横向钻取	75
5.2.1	造成横向钻取失败的 原因	75
5.2.2	维度表不必相同	78
5.3	一致性维度	79
5.3.1	维度一致性的类型	80
5.3.2	规划一致性	85
5.4	结构与一致性	86
5.4.1	维度数据仓库	87
5.4.2	企业信息化工厂	89
5.4.3	独立型数据集市	90
5.5	本章小结	92
5.6	补充阅读	92

第 III 部分 维度设计

第 6 章	深入学习维度表	97
6.1	将维度分组到表中	98
6.1.1	关联维度属性的两种 方法	98
6.1.2	对维度分组时的考虑	100
6.2	分解大型维度	102
6.2.1	任意分隔维度表	103
6.2.2	分割维度方法的替代 方法	104

6.2.3	微型维度缓解 ETL 瓶颈 和过度增长	105
6.3	角色维度和别名使用	110
6.3.1	单一表，多重关系	110
6.3.2	利用别名来存取角色	111
6.4	避免空值	113
6.4.1	空值带来的问题	113
6.4.2	避免空外键值	117
6.4.3	特殊情况行的用途	118
6.5	行为维度	121
6.5.1	在查询期间将事实转 换为维度	121
6.5.2	设计并使用行为维度	122
6.5.3	行为维度设计要素	123
6.6	本章小结	123
6.7	补充阅读	124

第 7 章	层次结构与雪花模式	127
7.1	钻取	127
7.1.1	钻取的定义	127
7.1.2	钻取的实质	128
7.2	属性层次和钻取	129
7.2.1	属性层次	129
7.2.2	在属性层次中钻取	130
7.2.3	钻取的其他方法	130
7.2.4	文档化属性层次结构	132
7.3	雪花模式	135
7.3.1	避免采用雪花模式	136
7.3.2	采用雪花架构	138
7.4	支架表	140
7.4.1	重复组	140
7.4.2	通过支架表消除 重复组	142
7.4.3	支架表和缓慢变化 处理	143
7.5	本章小结	145
7.6	补充阅读	145

第 8 章 深入学习缓慢变化技术 147**8.1 时间戳维度 147**

8.1.1 维度的时间点状态 148

8.1.2 时间戳解决方案 150

8.2 类型 3 变化 1548.2.1 研究具有新旧维度值的
所有事实 154

8.2.2 类型 3 解决方案 156

8.3 混合缓慢变化 159

8.3.1 冲突的需求 159

8.3.2 混合响应 160

8.3.3 评估和扩展混合方法 163

8.4 本章小结 164**8.5 补充阅读 165****第 9 章 多值维度与桥接表 167****9.1 标准的一对多关系 167****9.2 多值维度 169**

9.2.1 简化关系 169

9.2.2 为多值维度构建
桥接表 171**9.3 多值属性 177**

9.3.1 简化多值属性 178

9.3.2 使用属性桥接表 179

9.4 本章小结 185**9.5 补充阅读 186****第 10 章 递归层次与桥接表 187****10.1 递归层次 187**

10.1.1 涉及其他行的行 188

10.1.2 报表的难度 189

10.2 递归层次的扁平化 190

10.2.1 扁平化层次 190

10.2.2 扁平化带来的问题 191

10.2.3 适合扁平化的情况 193

10.3 层次桥接表 193

10.3.1 层次桥接表的设计 194

10.3.2 使用桥接表 197

10.3.3 双重计算 201

10.3.4 解决多对多关系 204

10.3.5 可能存在的误用 207

10.4 变化与层次桥接表 20710.4.1 维度或桥接表的
类型 1 变化 207

10.4.2 维度的类型 2 变化 208

10.4.3 层次的类型 2 变化 212

**10.5 关于层次桥接表的相关
讨论 214**

10.5.1 装饰桥接表 214

10.5.2 多父节点 215

10.5.3 多层次 215

10.6 本章小结 216**10.7 补充阅读 217****第 IV 部分 事实表设计****第 11 章 事务、快照和累积快照 221****11.1 事务事实表 221**

11.1.1 描述事件 221

11.1.2 事务事实表的特性 222

11.2 事实表快照 223

11.2.1 挑战：研究状态 223

11.2.2 快照模型 225

11.2.3 快照注意事项 229

11.3 累积快照事实表 23311.3.1 挑战：研究事件之间的
时间间隔 233

11.3.2 累积快照 236

11.3.3 累积快照的注意事项 240

11.4 本章小结 244**11.5 补充阅读 244****第 12 章 无事实的事实表 247****12.1 无事实的事件 247**

12.1.1 什么也不度量 247

12.1.2 无事实的事实表 248

12.1.3 使用无事实的
事实表 249

12.1.4 添加事实 250

12.2	条件、范围或资格	252
12.2.1	对条件建模的原因	252
12.2.2	用于条件的无事实的 事实表	254
12.2.3	比较行为和条件	255
12.2.4	缓慢变化维度和 条件	258
12.3	本章小结	258
12.4	补充阅读	259
第 13 章	特殊类型的星型模式	261
13.1	特殊类型属性	261
13.1.1	操作型系统	261
13.1.2	分析型系统	262
13.2	核心和自定义星型模式	263
13.2.1	核心和自定义维度表	263
13.2.2	核心和自定义事实表	267
13.2.3	其他注意事项	270
13.3	使用通用属性	271
13.3.1	通用属性	271
13.3.2	使用通用设计	272
13.4	本章小结	273
13.5	补充阅读	274

第 V 部分 性能

第 14 章	导出模式	277
14.1	重构维度数据	277
14.1.1	应用导出模式	278
14.1.2	已经接触过的导出 模式	279
14.1.3	导出模式的开销	280
14.2	融合事实表	281
14.2.1	预先计算横向钻取 结果	281
14.2.2	简化过程比较	282
14.2.3	改进性能	283
14.2.4	不能实现横向钻取的 支持工具	283
14.2.5	单一过程分析	283

14.2.6	包含一个不可共享的 维度	284
14.3	旋转事实表	285
14.3.1	对旋转数据的需求	285
14.3.2	旋转数据带来的好处	286
14.3.3	旋转数据带来的问题	287
14.4	切片事实表	287
14.4.1	为星型模式建立切片	287
14.4.2	使用切片事实表	288
14.4.3	开始切片	289
14.5	集合操作事实表	289
14.5.1	比较数据的两个集合	289
14.5.2	几种可能的比较	290
14.5.3	选择预计算集合操作	291
14.6	本章小结	291
14.7	补充阅读	292

第 15 章 聚集

15.1	聚集的基础知识	293
15.1.1	汇总基本数据	294
15.1.2	使用聚集	297
15.1.3	加载聚集	300
15.1.4	作为聚集的多维 数据集	302
15.2	使聚集不可见	303
15.2.1	聚集导航	303
15.2.2	建立聚集	306
15.3	可选的汇总设计方法	307
15.3.1	可转换汇总也是一种 可用的方法	307
15.3.2	避免单一表设计	308
15.4	本章小结	310
15.5	补充阅读	310

第 VI 部分 工具及文档

第 16 章 设计与商业智能

16.1	商业智能与 SQL 生成	315
16.1.1	SQL 生成器	315
16.1.2	SQL 生成器的限制	318

16.2	适用于语义层的指导原则	319	17.4.1	需要清理什么	357
16.2.1	需要避免的特性	320	17.4.2	清理维度数据	358
16.2.2	需要使用的特性	322	17.4.3	含有无效细节信息的事实	358
16.3	使用商业智能工具的 SQL 生成器	323	17.5	内部处理列	360
16.3.1	多星型模式	323	17.5.1	维度表中的内部处理列	360
16.3.2	半可加事实	328	17.5.2	内部处理与事实表	361
16.3.3	浏览查询	329	17.6	本章小结	362
16.3.4	桥接表	331	17.7	补充阅读	363
16.4	使用基于多维数据集的商业智能	336			
16.4.1	以多维数据集为中心的商业智能工具	336	第 18 章	如何设计并文档化维度模型	365
16.4.2	自动生成多维数据集	338	18.1	维度设计与数据仓库生命周期	365
16.5	本章小结	340	18.1.1	维度设计的战略位置	365
16.6	补充阅读	341	18.1.2	何时做维度设计	367
第 17 章	设计与 ETL	343	18.2	设计活动	367
17.1	ETL 过程	343	18.2.1	规划设计工作	368
17.1.1	ETL 是一种复杂的任务	343	18.2.2	开展访谈指导	369
17.1.2	ETL 过程使用的工具	344	18.2.3	设计维度模型	372
17.1.3	结构与 ETL 过程	344	18.2.4	计划的优先级	377
17.2	加载星型模式	344	18.2.5	文档化结果	378
17.2.1	高层依赖	344	18.3	文档化维度模型	379
17.2.2	加载维度表	345	18.3.1	需求文档	379
17.2.3	加载事实表	350	18.3.2	顶层设计文档	381
17.3	优化加载	354	18.3.3	详细设计文档	386
17.3.1	变化数据的识别	354	18.3.4	逻辑设计与物理设计	389
17.3.2	简化处理	355	18.4	本章小结	389
17.4	清理数据	357	18.5	补充阅读	390

第 I 部分

基 础

第 1 章：
分析型数据库与维度设计

第 2 章：
数据仓库体系结构

第 3 章：
星型模式与多维数据集

分析型数据库与维度设计

本书描述了适用于分析型数据库设计的一系列强大和有效的技术。采用这些技术能够以一种非常简单的方式实现复杂的业务分析工作：为业务过程度量建模。

业务过程的维度模型由两个组件构成：度量以及度量的环境。这些组件也称为事实和维度。在数据库设计时，利用这些组件将方便各种类型的分析型应用的实现。维度模型应用于关系数据库环境时被称为星型模式。应用于多维数据库环境时，维度模型被称为多维数据集。只要数据仓库的任何一部分包含了一个星型模式或多维数据集，就说明数据仓库采用了维度设计。

本章介绍维度设计的基本概念，这些基本概念可用于响应分析型系统所具有的独特需求。度量的概念定义为术语“事实”和“维度”，并应用于数据库设计或星型模式中。本章还展示了如何通过星型模式的基本交互来支持范围广泛的度量环境，并定义了用于指导开发维度解决方案的两个简单原则。

1.1 维度设计

从业务过程的角度考虑，信息系统可以划分为两个主要的类别：一类用于支持业务过程的执行；另一类用于支持业务过程的分析。维度设计的理念直接来源于分析型系统特有的需求。每个维度模型的核心都是业务度量的集合，业务度量反映了对业务过程的评价，并描述各种度量的环境。

1.1.1 目的

分析型系统和操作型系统具有完全不同的目的。操作型系统支持业务过程的执行，而分析型系统支持对业务过程的评价。不同的目的反映了分析型系统和操作型系统应用环境的差异，因此指导两类系统设计的原则也存在差别。

1. 操作型系统

操作型系统直接支持业务过程的执行。它通过获取有意义的事件和事务的细节来构建业务活动的记录。例如，销售系统需要获取订单、发货、利润等信息；人力资源系统需要获得雇员雇用

和升迁的信息；财务系统需要获得关于企业的金融资产和债务的信息。由这些系统记录的活动通常称为事务，而这类系统本身通常称为联机事务处理(OLTP)系统，或简称为事务系统。

为方便业务过程的执行，操作型系统必须能够实现包括插入、更新和删除等在内的数据库交互操作类型。这些交互操作的焦点几乎总是原子的：特定的订单、运输、退款等。这些交互操作具有高度的可预测特性。例如，订单输入系统需要提供对产品、客户、销售人员列表的管理；输入订单；打印订单汇总、发票、装箱单等；并且跟踪订单状态。

操作型系统关注过程执行，因此在事情发生改变时可能需要更新相关数据，并在数据操作有效期结束后，清除或归档数据。例如，当一个客户地址发生变动时，他或她原先的地址不能再作为运送产品或发送票据的地址，地址数据被简单地重写了。

在关系数据库设计领域，广泛使用并被大家认可的最好的操作型系统的模式设计方法是第三范式。这种设计方法采用实体-关系模型进行描述，实体-关系模型通常简称为 E-R 模型。与其他合适的数据库技术相结合，这种设计方法可以高性能地、以保持一致性和可预测的方式完成对原子事务的插入、更新和删除操作。开发人员通常将事务处理的这一特性称为 ACID 属性——原子性、一致性、隔离性和持久性。

2. 分析型系统

与操作型系统关注业务过程执行不同，分析型系统主要支持对业务过程的评价。本月订单的趋势与上个月相比有何不同？与本季度的目标相比，这种趋势说明什么问题？某一市场营销策略对销售有何影响？谁是我们最佳的客户？这些问题涉及对整个订单流程的度量，无法从单个的订单中获得答案。

与分析型系统的交互只需要利用对业务过程中包含的大量数据进行的查询操作就可以实现。在分析型系统中，不需要创建或修改信息。与操作型系统查询通常涉及单个的事务不同，分析型系统查询通常涉及大量的事务。分析型系统对特定问题的查询结果几乎是不可预测的，并可能随着时间的变化而发生改变。在操作型系统中不再使用的历史数据对于分析型系统来说仍然很重要。表 1-1 对操作型系统和分析型系统的主要差别进行了总结。

表 1-1 操作型系统与分析型系统的主要差别

	操作型系统	分析型系统
目的	执行业务过程	度量业务过程
主要的交互类型	插入、更新、查询、删除	查询
交互范围	单个事务	聚合事务
查询模式	可预测，稳定的	不可预测，变化的
时间关注	当前的	当前的和历史的
设计优化	更新并发性	高性能查询
设计原则	基于第 3 范式(3NF)的实体-关系(ER)设计	维度设计(星型模式或多维数据集)
其他称谓	事务系统 联机事务处理(OLTP)系统 源系统	数据仓库系统 数据集市(Data Mart)

维度建模的原则专门用于解决分析型系统特有的需求。维度设计用于优化对大量事务而不是单个事务的查询。支持并发和高性能更新不是维度设计的主要目标。维度设计支持对历史数据的维护,即使这些数据被操作型系统修改或删除过。

1.1.2 度量和环境

维度设计的基本原则非常简单。维度设计支持对业务过程的分析,这是通过对业务过程度量进行建模来实现的。

无论是通过倾听谈话还是阅读报表或图表,都很容易识别度量这一概念。考虑下列业务问题:

- 按照产品分类计算一月份的毛利是多少?
- 不同教育程度的平均账户结余是多少?
- 营销人员去年总的病假天数是多少?
- 销售商未付的欠款是多少?
- 供应商的收益率如何?

以上的每一个问题关注的是一个业务过程:销售、账户管理、出勤率、应付款、收益。这些以过程为中心的问题并不关注个别的活动或事务。要回答上述问题,需要关注一组事务。

更为重要的是,上述问题揭示了如何对不同的业务过程进行度量。对销售的研究包含对毛利的度量。金融机构度量账户结余,人力资源部门度量缺勤天数,财务部门度量应付款情况,采购经理关心收益情况。

在缺乏应用环境的情况下,度量没有任何意义。当谈及“销售是10000美元”时,这一信息对你来说并无实际的意义。因为你并不知道此处所指的销售是指单件产品还是多件产品,是代表一次的交易额,还是到目前为止公司整个的销售情况。没有应用的环境,度量就没有实际意义。

与度量类似,环境通常也反映在业务问题或报表中。例如,前面提到的问题中,毛利处于产品分类和时间(一月份)环境中;病假天数存在于部门(市场)和时间(去年)环境中;付款情况处于付款状态(未付款)和销售商环境中。

度量和环境这两个简单概念构成了维度设计的基础。所有的维度设计方案对业务过程的描述都是通过获取度量内容和度量所在的环境来构建的。

1.1.3 事实和维度

在维度设计中,将度量称为“事实”,将环境描述称为“维度”。在开展维度设计时,需要将信息需求分类到事实和维度中。具体的分类方法是从需求陈述或问题,以及报表中获取分类依据。在理解了如何发现分类的依据后,将信息需求分类到事实和维度的工作将非常简单。

1. 通过对语言描述的分析来发现事实和维度

对于常见的语言描述,在词汇“通过”、“根据”、“按照”后通常能够发现维度。考虑问题“按照产品分类,在一月份订单额度是多少?”。“按照产品分类”这一表述,很清楚地说明问题提出者需要对所包含的每个产品类别的订单额度进行统计。因此,产品类别是一个维度。

类似的,表示“在……时间”的词汇通常也表明存在维度。在此情况下,紧接的词语通常是维度的实例值。例如问题“按照产品分类,在一月份订单额度是多少?”。此例中,“在”后面跟

着“一月份”。一月份就是维度的值或实例。尽管维度名称没有显示地表现出来，但我们可以推断出维度名称为月份。

事实通常表现为数值，通常人们期望考查不同细节层次的事实。可以通过发现问题中期望开展“上卷”、“分类”等事件来识别事实。例如考查前述的问题“按照产品分类，在一月份订单额度是多少？”。在此，订单额度可以推断出是数值型，可以按照分类汇总。

并非所有的数值都是事实。有时，数值数据元素也可能是维度，关键是要看应用的场景。如果数值数据元素可用于描述细节的不同层次，那么为事实。如果数值数据元素提供了环境，那么为维度。“通过订单数量获得利润额度”这样的问题中包含两个数量数据元素。利润额度可以包含对不同层次的细节的描述，问题提出者要求按照订单数量对利润额度进行汇总。显然，“利润额度”是事实。如何理解订单数量呢？它也是一个数值，但是问题提出者并未要求将订单数量累加到一起。相反，订单数量主要用于描述利润额度表述的环境。因此，订单数量是维度。

其他一些常见的作为维度的数值数据元素还包括尺度、年龄、电话号码、文档编号以及单位数量(例如单位开销、单位价格等)等。判断这些元素是维度的线索在于它们的应用情况。这些数值本身不能上卷或下钻，但它们描述了其他元素上卷或下钻的环境。

2. 在报表中发现事实和维度

事实和维度如果存储在数据库中并以报表的形式展现，那么也可以根据它们的使用情况加以识别。维度充当“过滤器”或“查询谓词”。例如图 1-1 所示的报表，按照西部区域和 2009 年 1 月过滤细节数据。这些过滤器隐含了一些维度：地区、月份和年度。

Order Report Western Region January 2009 (cont'd)					
Category	Product	SKU	Quantity Sold	Cost	Order Dollars
Packaging	Box - Large	011-4822	700	\$ 950.53	\$ 1,100.00
	Box - Medium	011-4899	1,250	\$ 1,001.84	\$ 1,380.00
	Box - Small	011-5744	1,200	\$ 1,200.72	\$ 1,330.00
	Clasp Letter	011-1729	400	\$ 352.82	\$ 356.00
	Envelope #10	021-0011	2,000	\$ 2,017.46	\$ 2,080.00
	Envelope Bubble	021-0012	1,200	\$ 866.51	\$ 1,212.00
	<i>All Packaging</i>			\$ 6,389.88	\$ 7,458.00
Pens	Gel Pen Black	017-1999	5000	\$ 116.39	\$ 120.00
	Gel Pen Blue	017-2444	2990	\$ 600.88	\$ 624.91
	Silver Pen	017-3001	50	\$ 128.46	\$ 130.00
	<i>All Pens</i>			\$ 845.73	\$ 874.91
Grand Total				\$207,229.42	\$214,896.91
Page 10 of 10					

图 1-1 某订单报表的示例页

在报表中，维度还用于定义分组或“分类层次”，或者用于区分分类汇总的层次。从订单报表中可以看出，“Category”用于将报表划分为不同的区域，单个的分类，例如“Packaging”或“Pens”，

则被用于标记包含分类汇总的行。这些特征表明“Category”是一个维度。

维度也可以通过它们在控制度量聚集时的使用情况得以发现。在订单报表中,为每一个产品名称和SKU都定义了相关的度量(SKU是一个零售业术语,用于区分不同的产品,又称为“库存保持单位”)。因此,Product和SKU是维度。需要注意的是,维度值也可以用于数据排序。

事实也可以根据它们在查询和报表中的使用情况来加以识别。被聚集、汇总或分类汇总的元素都是事实。图1-1所示的报表清楚表明,“Cost”和“Order Dollars”都是事实。这些数值被定义于包括产品、分类和总计的不同的层次中。考查这个报表,还会发现另一个事实——Quantity Sold,尽管并未被分类汇总,但可以推断这个数值是通过聚集获得的。在本例中,这个数值看起来像是从单个订单中聚集的订购产品的数量。

图1-1所示的报表类型有时被称为扁平表格。还有一些报表可能以矩阵形式表示交叉参考的事实和维度,甚至以图形化的方式展现事实。图中标绘的值是事实,轴线上标明的值、条形图的标号、归属等是维度。

在有些场景中,判断一个数据元素是事实还是维度比较困难。有时一个数据元素既可以当作事实,也可以当作维度。在第6章“深入维度表”中讨论的行为属性,就是从事实中获取维度。之后在第11章“事务、快照和累积快照”中描述了从维度值中构建事实的技术。实际上可以发现,一些过程中并没有明确的度量。第12章“无事实的事实表”描述如何用无事实的事实表处理这类问题。您可能发现:一些主要的业务项不太容易汇总。第3章“星型模式与多维数据集”考虑了不可加事实,而第11章讨论了各种半可加事实。

1.1.4 聚合维度和事实

维度设计负责组织事实和维度并将它们存储到数据库中。对于维度集合中的元素来说,它们相互之间往往存在着独立于事实的关联关系。因此,通常需要将它们聚集到同一个关联表中以反映它们之间自然的聚类关系。同样,具有相同细节级别的事实往往也可以聚集在一起。

图1-2所示为对订单过程分析后得到的事实和维度列表。在这个示例中,产品通过分销商卖到零售商店,因此客户主要以公司形式存在。罗列在图中左侧的事实可以根据右侧所列的维度值开展上卷操作。

注意:维度值之间存在一些自然的

Facts	Dimensions
Order Dollars	Product
Cost Dollars	Product Description
Quantity Ordered	SKU
	Brand Code
	Brand
	Brand Manager
	Category Code
	Category
	Order Date
	Month of Order
	Quarter of Order
	Fiscal Period of Order
	Year of Order
	Salesperson
	Salesperson ID
	Territory
	Territory Code
	Territory Manager
	Region
	Region Code
	Region VP
	Customer
	Customer ID
	Headquarters State of Customer
	Billing Address
	Billing City
	Billing State
	Billing Zip
	SIC Code of Customer
	Industry Name of Customer

图1-2 订单过程中包含的事实和维度