大数据工程技术与应用

# 数据建模与DB设计

[韩] 金　范——主编

上海科学技术出版社

大数据工程技术与应用

# 数据建模与DB设计

[韩]金　范　主编

上海科学技术出版社

**图书在版编目(CIP)数据**

数据建模与DB设计 /（韩）金范主编. —上海：上海科学技术出版社，2016.10
（大数据工程技术与应用）
ISBN 978-7-5478-3242-4

Ⅰ.①数… Ⅱ.①金… Ⅲ.①关系数据库系统
Ⅳ.①TP311.132.3

中国版本图书馆CIP数据核字(2016)第208211号

**数据建模与DB设计**
[韩]金　范　主编

上海世纪出版股份有限公司
上海科学技术出版社　出版
（上海钦州南路71号　邮政编码200235）

上海世纪出版股份有限公司发行中心发行
200001　上海福建中路193号　www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092　1/16　印张 8.5
字数 170千字
2016年10月第1版　2016年10月第1次印刷
ISBN 978-7-5478-3242-4/TP·44
定价：34.00元

# 内容提要

本书重点介绍数据建模与数据库设计的理论及应用。作者从数据模型的发展历程及其必要性引入，结合作者在研究和项目实践中积累的经验，让读者理解数据建模是业务负责人与数据设计者之间沟通的工具，数据模型决定了数据处理的性能和数据管理的便利。书中首先将数据建模划分为概念建模、逻辑建模、物理建模以及最后的数据库设计四个阶段，明确了导出实体、设定实体的重要关系，并设定唯一键的数据建模流程。为了提高数据整合性和业务流程性，作者提出了范式化和反范式化过程，在构建最容易理解的数据结构的同时兼顾数据库的访问成本，寻找盈亏平衡点。

本书无论对数据分析设计领域的初学者还是实际业务的实践者，都有一定的启发和指导作用。

大数据工程技术与应用

# 编撰委员会

# 丛书序

“数支配着宇宙”——毕达哥拉斯。

大数据技术，使这句2000多年前的哲言如此形象、如此真切；大数据技术，正以前所未有的发展速度变革着人类的认知、产业和生活。

当前，我国正处于创新驱动发展、产业全面转型升级的关键阶段，大数据既是新的经济增长点，更是推进创新与发展的利器。上海产业技术研究院以服务于成果转化和产业化为使命，较早开始了大数据的应用研究和服务工作，构建了大数据应用技术平台，针对重点行业开展了一批大数据应用研究，涉及数据建模、数据分析、数据安全和数据库管理相关软件开发、测试、评价等多个方面。本丛书的出版既是前期工作探索的分享，更是进一步服务于成果转化和产业化的一个尝试。

大数据应用与产业化急需大量的工程应用技术人才。本丛书主要面向大数据工程应用的广大科技人员，在内容上汇聚了不同国家和区域、不同专业和领域的专家智慧，侧重大数据工程化知识、最佳实践和实用技巧，力求可操作性、实用性。

由于大数据技术研究和应用是一个新兴领域，发展方兴未艾，本丛书在编撰过程中因编者的知识和经验局限，必然存在许多不当之处，敬请广大读者提出宝贵意见。

2016年9月

# 前言

进入大数据时代，各行业界的数据生产急速增长，数据成为重要的资产，围绕数据的 ICT 环境也正发生着巨大的变化。

与数据处理速度的增加或存储空间的大容量化等相比，通过数据库创造新价值显得更趋重要，而随着数据库的使用形式渐渐多样化、复杂化，执行相关项目的机关或企业的难度也在逐渐加大。

人们开始更加强调数据中心系统构建方式的重要性，大规模业务系统也正式对数据建模提出要求，人们对数据的认识发生了巨大的变化。近来，即使不用强调，大家也都承认数据建模的重要性；然而，令人遗憾的是，拥有数据建模相关专业知识的人并不多，虽然在数据分析与建模、数据结构设计、数据库构建、数据处理的性能管理等各个细分领域的专家人数正逐渐增多，但对此有着综合洞察力、能够统领行业的专家却为数不多。

数据建模业务作为一个将现实业务抽象化、单纯化、明确化的过程，需要在考虑到复杂 ICT 环境、数据增加、安全及性能相关问题，以及系统修护等全方位的情况下进行操作；因此，积累这方面的知识并不容易。错误的数据设计会给整个系统带来消极影响，可能造成无法构建正常系统的致命后果。

本书根据作者在操作现场的各种经历编写而成，无论是对于刚刚开始接触建模的初学者，还是已有多次建模经验的专业人士，都有一定的参考价值。作者希望读者们能够以本书的数据建模知识为基础，让大数据时代中最核心的数据建模在信息化系统建构中充分发挥其重要作用。

本书由金范编写，曹艳珺、张青对本书进行了认真校对，周兆明、王一帆、邱雯等参与了资料的收集、整理、录入等工作。此外，本书的编写得到了上海产业技术研究院大数据专家委员会等相关单位的大力支持和指导，上海产业技术研究院的组织协调也使本书得以顺利出版，在此一并表示衷心感谢。

金　范

2016 年 6 月

# 目 录


第1章 数据建模概述 1

·1.1 数据模型的概念 2
·1.2 数据模型的必要性 3
·1.3 业务理解的重要性 4
·1.4 建模的基本规则 5
1.4.1 没有 100 分的数据模型 5
1.4.2 端正数据建模员的作用 6
1.4.3 改善数据处理性能的建模 6

·1.5 数据结构与业务流程的验证 7
·1.6 建模的不同阶段 8
1.6.1 概念数据建模(conceptual data modeling) 9
1.6.2 逻辑数据建模(logical data modeling) 9
1.6.3 物理数据建模(physical data modeling) 10
1.6.4 数据库设计(database design and architecture) 11

·1.7 自上而下式 (top-down) 建模与自下而上式 (bottom-up) 建模 11
·1.8 结构化查询语言 (SQL) 11
·1.9 数据库构建阶段 12

**第2章 数据建模流程** 13

·2.1 导出实体 (entity) 14
2.1.1 实体的定义 14
2.1.2 选定实体候选 15
2.1.3 实体分析方法 18

·2.2 设定关系 (relationship) 19
2.2.1 关系图式的解释 20
2.2.2 设定关系时的原则 21
2.2.3 个体创建与关系的关联性 24
2.2.4 关系 M 的范围 25
2.2.5 关系与数据的完整性 25

·2.3 选定唯一键 (unique identifier) 26
2.3.1 业务中有意义的复合属性 27
2.3.2 人造键(artificial key)的使用 27
2.3.3 唯一键属性的顺序 30
2.3.4 唯一键的继承 32

·2.4 导出属性 (attribute) 32
2.4.1 属性的种类 33
2.4.2 属性的验证 35

**第3章 范式化 (normalization) 和反范式化 (de-normalization)** 37

·3.1 范式化 38
3.1.1 第 1 范式 38
3.1.2 第 2 范式 39
3.1.3 第 3 范式 40

·3.2 反范式化 40

3.2.1 属性的重复 40
3.2.2 实体的合并与分离 41
3.2.3 分布式环境的实体重复 43

**第4章 实体种类与特性** 45

·4.1 主实体 (primary entity) 46
·4.2 关联实体 (associative entity) 48
·4.3 历史记录实体 (historical entity) 51
·4.4 父型/子型 (super/sub type) 实体 52
·4.5 排他 (arc) 关系实体 54
·4.6 递归 (recursive) 关系实体 55
·4.7 1：1关系实体 57
·4.8 M：M关系实体 58
·4.9 各种角色 (role) 60
·4.10 列优先和行优先 62

**第5章 数据建模实例** 63

·5.1 上位实体的设定 64
·5.2 客户实体 67
·5.3 商品实体 71
·5.4 关系实体的唯一代码设定 77
·5.5 派生属性的生成 80
·5.6 属性的分离和重复 82
·5.7 历史记录实体和关系实体的选择 83
·5.8 递归 (recursive) 结构的实体 86
·5.9 会员管理 90
·5.10 历史记录变更管理 94
·5.11 1：M中M的范围 (1) 100
·5.12 1：M中M的范围 (2) 104
·5.13 为业务流程改变数据结构 106

·5.14 维持阶层结构的一贯性 109
·5.15 分布式环境中的注意事项 113

第6章 数据建模习题 117

·6.1 习题 1 118
·6.2 习题 2 119
·6.3 习题 3 119

参考文献 121

# 第1章

# 数据建模概述

## 1.1 数据模型的概念

请大家想象一下自己家中房间、客厅、仓库等处散落着的各种物品。为了方便地使用书籍、衣服、鞋子、被子、电子产品、玩具、学习用品等，必须适时适当地整理，为此需要收纳空间，必须在每个收纳空间内系统化地整理物品。我们整理物品不仅是为了知道物品平时所在位置并在需要时可以轻松找到它，而且为了能够有效地做多件事。儿童房内放置孩子使用的玩具、儿童书籍、学习用品等，厨房内摆放厨房用品、食材等，相互关联的物品放置在同一个空间内会更为有效。如果是家里所有人共同使用的物品，放置在约定好的位置固然不错，但也可以根据需要在必要的位置重复放置多个。但如果独自生活家当又不多的话，比起将物品放在多个位置分开收纳，全都笼统地放在一处反而更加方便。

业务中使用的数据模型也与此类似，即：了解何种数据由何人使用、如何使用，将数据分类整理、保管，对相互关联的数据设定关系，从而使人能够轻松联系起来查找并获取数据。

使用简单业务流程的系统中，用户不多且事务也不多时，纵使不系统化地整理数据，数据处理也不存在什么问题。但从现在起我们要处理的数据范围并非家中的物品，至少应该比喻成不停地运输、装载大量物品的大型物流仓库。

为了开展业务，存在一些无论用什么方法都必须要记录下来的数据。业务初期所需的数据可视为业务开展所需的最小的重要信息集合。业务逐渐扩大后重要信息也增加，与之相关的信息集合也逐渐增多，管理数据的方式也日渐复杂。如果从业务流程角度管理数据，业务初期的效率会有所提高，但随着数据逐渐增多，各处的数据会发生重复且很难维持数据准确性，必然会导致难以有效地管理数据。因为这种现实情况，看待数据的角度必须是数据本身，导出业务所需的集合(全体集合)后，再划分出具体的集合，进而导出集合间的业务关联性，这被称为以数据为中心的业务分析。这种以数据为中心的处理方法正成为系统构建方法的标准。

为了保持以单纯的角度看待数据，需要将业务中产生的数据全体集合进行具体划分。尽可能单纯地看待业务数据时，则全部业务都有能够交易的商品、都有生产交易该商品的主体，主体间的交易中一定存在几种形态的合同，以此合同为依据发生实际交易，而合同则由合同主体、交易场所等数据构成。

如图 1-1 所示，如果用大集合对数据分类，则业务领域的数据集合非常简单。但从全体数据集合中分离出各个具体集合并确定其数据结构的过程却并不简单。数据结构决定

数据处理性能与数据管理便利性。因此，生成数据结构与数据集合需要综合考虑各种要素。为了使这一过程有效进行，需要理解数据集合，区分数据与流程，并具备分析数据的能力。准确理解数据集合的概念，就拥有一双在任何复杂情形中都能区分数据与流程的慧眼，获得读取核心数据、统领整个数据结构的能力。

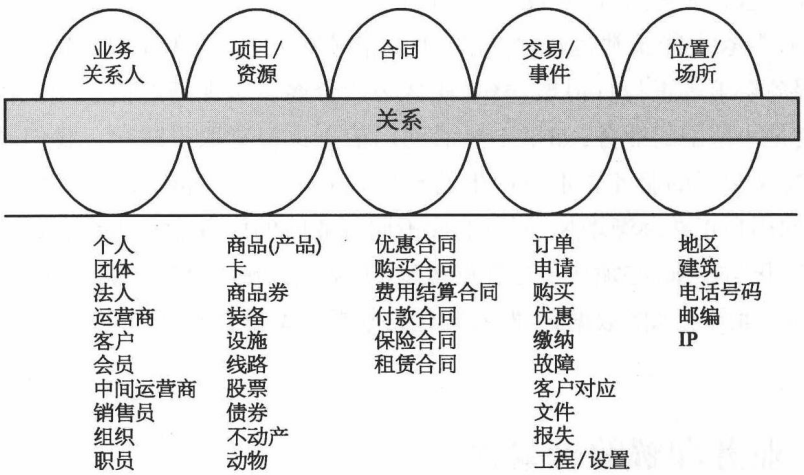


**图 1-1　数据集合分类**

## 1.2　数据模型的必要性

直到 20 世纪 80 年代后期，数据模型、关系数据库(relational database)都还是相当陌生的术语。当时大规模数据中心的数据管理还以 ISAM、VSAM、HDB 等为主，关系数据库作为试验阶段的数据管理方法，也只用于主要业务以外的周边业务，其市场反应非常冷淡，在性能方面的表现尤为不佳。

此后大型系统集成企业开始引进系统构建方法论，其中包括数据分析。随着数据分析方法论的发展，人们提出了数据模型概念。随着应用系统构建方法论的计算机辅助软件工程(computer aided software engineering，CASE)工具开始流行，数据模型的必要性也开始体现出来。

20 世纪 90 年代起，各种关系数据库陆续进入市场，以数据为中心的系统构建方式的重要性进一步显现出来。随着数据建模经验的积累，开始出现这一领域的专家，数据建模开始正式应用于大型业务系统中，人们对数据的认识也开始发生巨大改变。

近来，即使不强调，大家也都承认数据建模的重要性。但遗憾的是拥有数据建模相关专业知识的人并不多，虽然在数据分析与建模、数据结构设计、数据库构建、数据处理的性

能管理等各个细分领域的专家正逐渐增多，但对此有着综合洞察力、能够统领行业的专家并不多。如果了解有关数据模型的功能性知识但并不理解数据结构如何影响实际业务处理，则实施有效的建模时必然会心有余而力不足。我们必须认识到，为培养这种专家，传授专门知识固然是必需的，但同时还必须不断为他们创造能够持续将相关知识应用于实务的机会，而实际上这并非易事，需要投入大量时间和关注。可能有人会强调，现在使用的业务系统即使没有专家也能正常运转，性能管理也很顺畅。但是，在我们所生活的今天，数据、事务、各种服务要求等正与日俱增，整个社会必然会逐步实现高度信息化。在这个信息化社会中，纵使是小范围的业务，如果不灵活利用信息也很难取得成功。我们看到为有效应对逐渐变得复杂多样的服务要求所产生的新形式业务流程正持续增加，由此可知，使用这些业务流程的用户正在逐渐增加，同行业或不同行业间的服务竞争也会日趋激烈。为了应对这种环境变化与要求，必须从长远观点构建 IT 基础设施。从这个角度来看，可以说为了实现超越现实、面向未来的数据管理，系统化的数据模型是必不可少的。

## 1.3 业务理解的重要性

数据模型指的是“表现业务中使用的数据结构的设计图”。因此，为了确立数据结构，必须有最准确了解相应业务知识的核心业务负责人的积极参与。虽然业务负责人了解需要何种数据，但他们无法了解何种数据应以何种形态构成、如何构成。换句话说，他们在理解数据方面存在明显局限。

虽然对业务流程的理解程度并非数据建模的绝对要素，但它严重地影响确定数据结构的过程。作为实施数据建模的人(数据建模员，data modeler)，最理想的人选无疑是兼备渊博的业务知识与专业数据建模技术的人，但这在现实中大多是不可能的。所以大部分的建模进行过程必须通过业务负责人与系统设计者间的适当沟通方能顺利进行，业务范围较广泛时可能会有数十人参与。进行数据建模时，为了明确业务领域中使用的数据并通过适当沟通导出要求，充分审核现有数据与新要求数据从而做出明确的设计，数据建模员必须理解业务流程。特别是在建模初期阶段，业务负责人必须详细陈述意见，帮助建模员完成业务分析。此过程中，实体关系图(entity relationship diagram, ERD)是确保顺畅沟通、表现标准化图式的优秀工具。业务负责人向数据建模员说明业务知识，数据建模员告知负责人如何反映要求事项以及后期如何实现，必要时告知负责人业务流程中必须要反映的事项等，以此提高业务负责人对数据的理解程度以及数据建模员对业务的理解程度，从而逐步提高数据模型的质量。

过去要求建模的客户中，偶尔也会有人质疑说“我花钱请你们工作，你们应该自己看着办，为什么还要我帮忙呢”。当然，客户中也存在核心业务负责人抗议说“我目前的工作已

经很辛苦，为什么连数据建模都要我帮忙呢”。但是，请大家静心想一想。未能充分理解业务知识的数据建模员在完成建模后离开，之后才发现关键问题该怎么办呢？数据结构的问题会再次泛滥成混乱的数据。而数据建模的质量问题最终必然会留给留下来的业务负责人与系统操作者，让他们承担后果。这不仅仅是交接层面的问题，更需要我们转变思维、转变认识，即“这是我的工作，建模的质量由我决定”“建模员是帮助我的人”。

## 1.4　建模的基本规则

笔者首次接触数据建模是刚参加工作时利用 CASE 工具做的一个项目，为了使用工具，必须先理解系统构建方法论与数据模型。当时最先接触的方法论是詹姆斯·马丁(James Martin)的信息工程学(information engineering)，当时系统构建方法论和数据建模都还非常陌生，查找相关书籍并理解其中的内容都很困难，有疑问时几乎也无处可问。自那时起经历了数十载的光阴，现在笔者的感受是，从最开始到随着时间流逝做新项目、讲新课或读新书，每年看待数据的角度也都在发生变化。遗憾的是，此前的建模专家应该也为这些烦恼所困扰，但苦于没有学习这些经验的途径而走了相同的弯路，浪费了很多时间。以这期间的经验思考时，因为数据建模员经验不同、对建模的态度与认识也各不相同，因而会有各自的体系，但自然而然也会存在一些顽固的错误坚持。为了帮助大家不再陷入同样的不合理逻辑陷阱，下面将为大家整理几点作为数据建模员应该遵守的基本规则。

### 1.4.1　没有 100 分的数据模型

“没有 100 分的数据模型，所以只需按照数据建模员的喜好自行构建即可”“无论如何业务都会运转”“如果这样对的话那那样也是对的，请不要干涉”，以上是我们经常能听到的一些话。这些话是有一定建模经验、觉得自己建模工作做得还不错的人说的。看起来也像那么回事，但这就是陷阱。虽然没有 100 分的模型，但也不存在 0 分的模型。没有 100 分的模型是因为不可能在了解相应业务的全部规则、拥有数据建模专业知识后再去建模，因为大部分一定规模以上的数据模型都无法由一人独自完成。不存在 0 分的模型是说即使数据建模做得不对，也不会出现无法编写应用程序、业务流程瘫痪的情况。但有一点很明确，那就是我们能够区分做得好的数据建模和不及格的数据建模。做得好的数据建模能妥善应对数据积累以及业务流程变更与增加引起的各种状况，而不恰当的数据建模只能运用复杂、性能低下的应用逻辑(application logic)或结构化查询语言(structured query language，SQL)，引起临时数据或重复数据泛滥，最终导致系统无论如何努力也无法改善。如此一来，可能会导致整个系统性能低下、各处数据不一致、业务负责人无法信任数据准确性等一系列最坏情况。因为这些原因，在完成建模时至少要通过建模验证解决问题后方能进入下

一阶段。即使忽略该过程直接进行，如果能在下一阶段发现问题并立即更正也实属万幸，但不幸的是模型中的问题大多在项目最后的测试阶段或系统运行阶段才被发现，此时已经到了无法挽回的境地。

因错误数据模型经常导致的几种问题分析如下。

- 多个组织中报告给经营者的资料分类系统不同，数据不一致。
- 为了业务流程，根据需要随时制作临时表使用。
- 如想准确地选出报告，必须由了解多年数据历史的人操作。
- 管理层要求综合统计信息时需要部门间的协商与数据确认。
- 作为业务组织单位，业务系统要求分离或要求新设辅助系统。

### 1.4.2 端正数据建模员的作用

业务负责人被之前经验所束缚造成偏见，以一己之力统领数据建模，程序员或程序分析员为达到个人目的影响数据结构或因所谓强权业务部门的利害（损益）影响数据结构，则数据模型必将面目全非。之所以产生这种现象，是因为数据建模员自己不是领导也不做决定，而是如同象棋中的卒子般依赖他人对数据相关事宜做出的决定，仅起到在 ERD 中反映相关事宜的作用。这种情况下，数据建模员只是单纯的技术人员而非专家。数据建模必须以明智的专家为中心。那么建模员是象棋中的将帅吗？他可以总是对有关数据的全部事宜做出最终决定并在相互间存在分歧时做出判断吗？这样的数据模型无法信任。数据建模员必须在各种专家主张自己的见解时起到协调领导作用。建模员协调拥有过去经验的业务系统负责人、程序分析负责人及业务部门负责人间的意见，是制定经所有人同意的数据模型的执行者。为忠实地履行这一职责，建模员必须根据多种形态的数据结构预测可能会对流程与数据管理产生何种影响，并对此进行充分说明，从而使所有人都能朝着同一方向做出决定。

“不要自己决定，让大家去决定”，这是数据建模员的正确姿态。必须做出所有人都能达成一致意见的决定，即使不能如此，也要拥有让大家对建模员的决定可协商一致的领导能力。只有如此，方能称为真正意义上的数据专家。为了能够践行这种职责，必须以信任为基础。事实上，在项目初期获取参与者信任的最好方法就是传授经验。

### 1.4.3 改善数据处理性能的建模

如图 1-2 所示，多个要素都会影响系统性能，但实际上给数据库系统性能造成最大影响的部分是分析与设计阶段投入多少时间、整个数据与流程结构的构成是否良好。

虽然数据结构给整体业务处理性能造成巨大影响，但有时也会因为极其微小的性能原因造成不合理地确定数据结构的情况。这是过去验证某客户公司的建模时发生的事例。对公司组织建模时将分店定义为分店实体、店铺定义为店铺实体、代理店定义为代理店实体、共同部分定义为组织实体，各实体（entity）的关系以独立关系连接。因感觉模型有些出