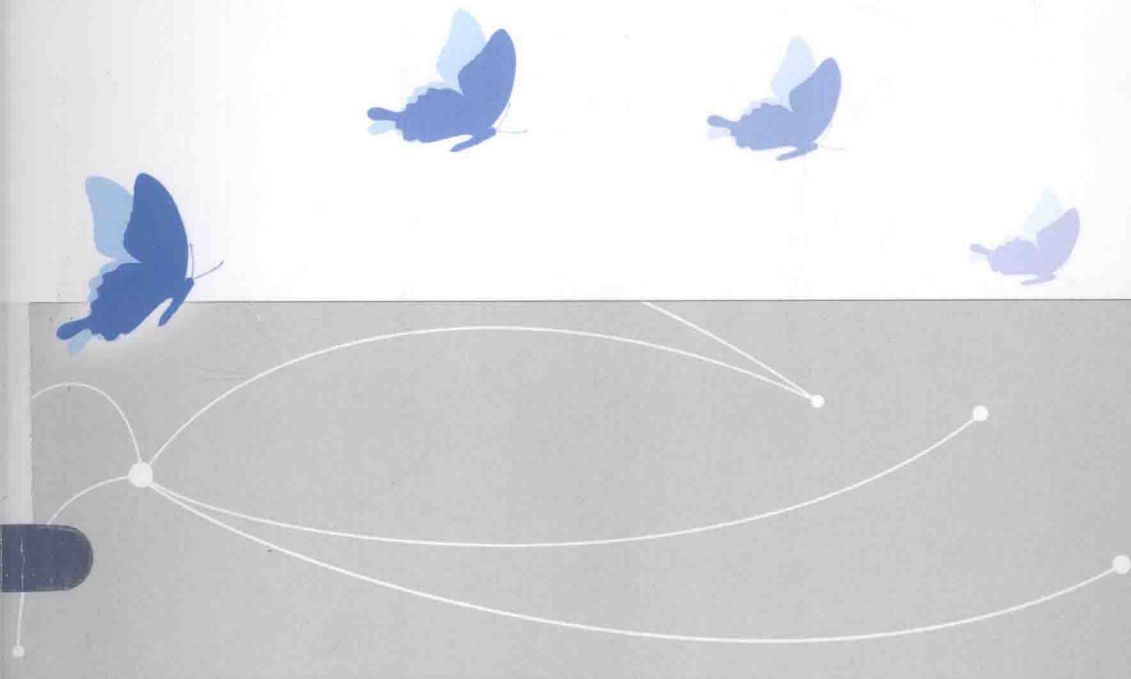
基于元数据驱动通用操作器的数据仓储构建

Constructing Data Warehouses with
Metadata-driven Generic Operators and More

[瑞士] Bin Jiang 著

郑悦林 吴西燕 余肖生 王东娟 赵美林 王缓缓 译著



WUHAN UNIVERSITY PRESS
武汉大学出版社

基于元数据驱动通用操作器的数据仓储构建

Constructing Data Warehouses with
Metadata-driven Generic Operators and More

[瑞士] Bin Jiang 著

郑悦林 吴西燕 余肖生 王东娟 赵美林 王缓缓 译著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

基于元数据驱动通用操作器的数据仓储构建/(瑞士)蒋彬著;郑悦林等译. —武汉:武汉大学出版社,2014.12

ISBN 978-7-307-14882-6

Ⅰ.基… Ⅱ.①蒋… ②郑… Ⅲ.元数据—研究 Ⅳ.G250

中国版本图书馆 CIP 数据核字(2014)第 268659 号

责任编辑:李汉保　　责任校对:汪欣怡　　版式设计:马　佳

出版发行:**武汉大学出版社**　(430072　武昌　珞珈山)

(电子邮件:cbs22@whu.edu.cn　网址:www.wdp.com.cn)

印刷:武汉中远印务有限公司

开本:720×1000　1/16　　印张:21.5　字数:307 千字

版次:2014 年 12 月第 1 版　　2014 年 12 月第 1 次印刷

ISBN 978-7-307-14882-6　　定价:55.00 元

中文版序

信息技术的飞速发展和深入应用不断推进社会的进化。我们说我们处在信息社会，一个重要的标志性特征就是技术融合(或称做IT融合)。也就是说，当前所处的信息社会与以往的工业社会、农业社会相比较，处于一个前所未有的技术大融合的阶段。这体现在技术的两个演化维度上面：技术透明性(transparency)和技术渗透性(pervasiveness)。技术透明性是指由于技术水平的提高和广泛的业务应用，在越来越多的产品和服务中，顾客甚至都感觉不到(也无需过多了解)内在的技术细节。例如，手机用户并不需要了解无线通信基站的工作方式和信号发送指标；Internet用户也无需了解网络通信的多层协议结构。对于最终用户而言，他们只需了解技术所呈现出的效用即可，因此技术对于用户来说具有透明性。技术渗透性是指技术对人类社会和生活的方方面面的影响深度。对于企业来讲，许多传统的运作管理逐渐变成了面向数据的管理，许多传统的业务决策逐渐变成基于数据分析的决策。例如，通过深度商务分析(business analytics)更好地了解客户、业务和竞争对手，以开展精准营销、优化运营管理、保持和提升战略优势。

近年来，全球数据量正呈现出前所未有的爆发式增长态势。国际数据公司(IDC)的研究报告预计2020年全球被创建和被复制的数据总量将达到35ZB。与此同时，数据复杂性也急剧增长，其多样性(多源、异构、富媒体等)、低价值密度(大量不相关信息、知识"提纯"难度高)、实时性(流数据，需实时生成、存储和分析)等复杂特征日益显著。"大数据"(即以超规模、多样性、低价值密度、实时性为显著特征的数据)成为重要话题，并在业界和学界引起广泛关注。

数据仓储(Data Warehouse，也称数据仓库)作为一类重要数据

平台，在过去20余年间得到了理论和实践上的长足发展。在大数据背景下，随着对于数据的获取、组织、分析和利用等应用需求的快速增长，数据仓储的重要性正在进一步凸显。数据仓储面向分析处理(analytical processing)，强调数据整合、切分回溯以及多维视图，重点支持“为什么发生”这类分析型管理问题的求解。

《基于元数据驱动通用操作器的数据仓储构建》一书根据作者在数据仓储领域丰富的专业知识及相关经验，对数据仓储的概念、框架和构建进行了较为全面的阐述和讨论。该书恰当地刻画了数据仓储与操作型数据、分析型数据、数据分析器、业务改进器的关系，并围绕预备域、处理域和存储域对于数据仓储的构建进行了详细探讨。特别值得一提的是，该书介绍了一种基于元数据驱动通用操作器(MGO)的数据仓储构建方法，旨在通过面向元数据本身(而不是面向具体数据(值)内容)获得数据和相关程序/功能的独立性。这是一个很有价值的构建思路，一方面可以增强系统适应性，通过对元数据的操作而减少对于诸多相关数据和具体功能的操作；另一方面可以提升系统完整性，通过对元数据的操作而避免诸多相关数据和功能在内容和操作中的不一致性。

本书不仅提供了基于元数据驱动通用操作器的数据仓储构建的概念和思路，而且对于一系列相关操作器也给出了较细致的介绍和方法描述。从整体撰写内容和风格看，本书可以作为计算机科学和工程、信息系统应用相关专业的课程教材，也对于从事IT咨询和实施、数据仓库构建和大数据分析等应用的企业IT管理者和专业人员具有参考价值。

相信广大读者可以从本书中获得许多启迪。

陈国青 *

2014年4月于清华园

---

* 陈国青，现为清华大学经济管理学院EMC讲席教授，2005年度受聘国家教育部长江学者特聘教授。担任国家教育部高等学校管理科学与工程类专业教学指导委员会主任委员、国家信息化专家咨询委员会成员、国际信息系统学会中国分会(CNAIS)创始主席(2005—2013)。

译 者 序

2011 年，蒋彬博士出版了数据仓储领域具有里程碑意义（Bill Inmon）的著作——*Constructing Data Warehouses with Metadata-driven Generic Operators and More*。该书不仅对数据仓储的定义、类型、特征及要求进行了完整而准确的论述，对目前两种主流的数据仓储开发方法——Top-Down 以及 Bottom-Up 进行了分析，根据自己近 20 年数据仓储的从业实践提出了一种行之有效的方法——Top-On，并从体系结构、组件算法与技术、数据仓储的构建范式、数据仓储的生态环境四个方面对 Top-On 方法进行了详细的介绍，集理论性与实践性于一体，“应为每个认真、严肃的数据仓储实践者案头必备之物”（Bill Inmon）。

2011 年底，蒋彬博士受聘为三峡大学特聘教授。在与蒋彬博士深入交流并拜读原著后，我们被其精妙的构思、严谨的体系及模型的可实现性所折服，遂决定翻译该书，为国内数据仓储建设介绍一种新的思路及方法。由于语言表述方式不同，我们与原作者蒋彬博士沟通达成共识：翻译时不采用逐句翻译，而是以相对完整的语句群为单位，在不改变原意的情况下重新表述，以符合汉语的语言习惯。

译者全部来自三峡大学，分工如下：郑悦林翻译第 4 章、第 7 章、第 8 章，吴西燕翻译英文序、前言及第 1 章、第 10 章、第 11 章，余肖生翻译第 2 章、第 6 章、第 9 章，王东娟翻译第 3 章、第 5 章、第 12 章及全书插图，王缓缓翻译第 11 章，第 12 章，赵美林翻译第 13 章、第 14 章。全书由郑悦林与吴西燕统稿。

经过长达两年的集体学习讨论、分工翻译、交换修改、分块审查，译稿才最终完成。翻译小组每个成员都对全稿进行了多轮阅读并提出修改意见。可以说，整个书稿凝聚了每一位译者的心血。

在翻译过程中，覃兵文、姜艳静等老师参与了部分初稿的翻译；李碧涛老师多次参与讨论，为翻译工作和翻译内容提出了宝贵的意见，在此对这些老师表示衷心的感谢！

原著作者蒋彬博士自始至终指导着整个翻译工作。他踏实的工作作风、严谨的治学态度给我们留下了深刻的印象，也深深地影响着我们每一个人。

最后感谢陈国青教授百忙之中阅读书稿并为本书撰写中文序！

译　者

2014 年 5 月

序

在数据仓储发展初期，绝大部分数据处理都通过主机系统和事务操作来完成，那时的数据库理论家们对数据仓储不屑一顾。联机事务处理(OLTP)为当时的主流，任何不以 OLTP 为中心的数据处理都被打入另册，不予考虑。

随后，数据仓储开始被商界重视。随着营销系统、库存管理系统、客户应用系统的出现，商界人士将技术人员引入了数据仓储领域。

很快，随之而来的是数据集市、维度建模、企业信息工厂及数据仓储 2.0(DW2.0)。今天，数据仓储已作为通常智慧为人们所接受。业界在商务智能和数据仓储上的投资早已超过在 OLTP 上的总投资。

数据仓储业以惊人的速度成熟。蒋彬博士的著作——《基于元数据驱动通用操作器的数据仓储构建》是数据仓储业发展成熟的极好实例。

在阅读蒋彬博士著作的过程中，以下三点给我留下了深刻的印象，即书的：

——完整性；

——实用性；

——创新性。

蒋彬博士的著作代表了数据仓储发展进程中最新和最完整的一步。本书应为每个认真、严肃的数据仓储实践者案头必备之物。我热切地欢迎这一始于数十年前的数据仓储发展之路的

最新里程碑。

Bill Inmon①

Colorado(科罗拉多)

2011 年 8 月 22 日

① Bill Inmon 被公认为“数据仓储之父”。关于他在这个领域贡献的更多内容，参见 http://en.wikipedia.org/wiki/Bill_Inmon。

前　言

即使经过了近30年的实践，数据仓储的构建、扩展和维护对许多企业来说依然是个挑战。无论是否采用工具或其他辅助手段，这件事仍是昂贵、费时和有风险的。造成这些问题的主要原因之一是重复。

如果对数据仓储构建、扩展或维护过程中所进行的活动进行分析，读者或许会发现一些事情是值得注意的。对我来说，这就是重复。不仅在开发阶段，而且在设计说明阶段、测试阶段、文档阶段，“复制—粘贴—搜索—替换—调整—验证”操作链在成千上万次地重复。事实上，尤其在当今瞬息万变的业务世界里，正是这种重复使得数据仓储的构建、扩展和维护具有不可思议的挑战性。

基于这简单而根本的观察，本书介绍一个全新的数据仓储构建方法。采用该方法，数据仓储的整个加载和更新机制由12个小型元数据驱动通用操作器组成。即使面对非常复杂和大规模的企业型数据仓储，这一方法也是有效的。使用这些操作器，上文所提及的重复可以有效地给予消除。由此，数据仓储的构建、扩展和维护将实质上变得更便宜、更快捷、更安全，20倍的效能提升也不再是不可思议的了。新方法在实践中的运用已清晰地表明，从某种程度上说，新方法使数据仓储构建在整个数据仓储项目中不再像从前那样是一个突出的问题了。

在数据仓储领域最有影响的两个人物Bill Inmon和Ralph Kimball之间一直有一个经久不息，且当前仍持续进行的争论①。

---

① 关于这场争论更多的内容，参见http://www.b-eye-network.mobi/view/14115。

争论的焦点就是哪种数据仓储的构建方法更好：是 Bill Inmon 主张的自上而下的方法，还是 Ralph Kimball 主张的自下而上的方法。根据一般工程学的教条，前者更加理性。然而，这种方法在过去的实践中却经常失败，因为与之相应的项目费用太高，耗时过长且不能如期交付第一个实质性的成果。至少从短期的观点来看，后一种方法则并非如此。如果数据仓储的构建不再如上文所述是一个难题，那么继续争论下去就没有意义了：即自上而下的方法应该胜出。

过去，我一再注意到，不同企业的数据仓储环境中用到了大量的体系结构选择，如系统组件、算法、技术等都一次次被重复地再次发明，而实质上它们并没有多少新意，这是另一种重复。尽管这种努力付出带来了乐趣，但这种乐趣对我们的客户和投资人来说实在不便宜。为了减少这种重复发明的必要性，本书收集、描述并在复杂数据仓储体系结构的环境下分析了 20 多个常用的通用算法。尽管这些算法奠定了上文提及的通用操作器的内容基础，但当使用传统方法构建数据仓储时，包括构建非常简单的数据仓储，这些算法也能单独采用。此外，本书包括 30 多个构建练习。如果读者在阅读的过程中完成了这些练习，那么，合上本书之前，即使面对一个错综复杂的数据仓储读者也已具备坚实的基础。由此可知，新方法并非不可驾驭。

我们在大学里学习了大量的计算机科学知识，如实体—联系分析、关系理论、范式、数据建模、算法和数据结构、复杂性分析、图论、编程语言、编译原理、事务管理、系统体系结构、软件工程，等等。有多少知识我们已自觉地运用到数据仓储的构建之中？据我的观察，答案是："几乎没有"。事实上，对以上知识而言，数据仓储的构建是一个极具综合性的领域。因此，它非常适合于学生运用刚刚学到的知识进行项目练习，由此获得对这些知识更好的理解。本书考虑了这一点并提供了这一可能。

数据仓储这个术语已经有 20 年的历史，最先由 Bill Inmon 在 1991 年提出。然而，如果读者问问身边的 5 位资深人士——其同事、其老板、其投资人或者其客户——什么是数据仓储？读者很可

能会得到六种不同的答案，这可能很有趣。但是，如果我们意识到数据仓储是一项严肃的工程事务，而且在大多数情况下对企业有着重要的战略意义，这种状况就不再让人觉得轻松自在了。这也许就是过去虽然有成千上万篇论文、报告、博客以及类似的文章，但在这一领域却没有取得什么实质性进展的原因之一。一般而言，我们一再地重复着相同的内容，其实是穿着不同的衣服参加不同的聚会而已。

作为数据仓储顾问，我们一再告诉我们的客户，通过数据仓储，我们可以从不同的数据源中找到唯一的真理。尽管我们已经产生了成千上万的讨论源，但是正如上文所提到的，我们自己却还不能确定一个数据仓储的唯一定义。对我们来说，这难道不是一个讽刺吗？因此，本书的另一个目的就是努力澄清几个重要，但却混乱的术语，如数据仓储①、商务智能②和时间性，我清楚地知道这是一项充满挑战和风险的工作。

也许，我触发了一个数据仓储构建的范式转换③，由此也可能导致数据仓储机。为了此次尝试，我期望得到读者的理解，更重要的是得到读者的支持！

不过，到底何为数据仓储，至少是根据……

蒋　彬

Niederglatt ZH（瑞士）

2011 年 7 月 7 日

---

① 数据仓储术语的通俗定义或解释的不完全的列表，参见 http://www.google.ch/search?q=define:DataWarehouse&hl=de&defl=zh-TW&sa=X&ei=JyEjTfSBNtqN4gbp7KTFAQ&ved=oCAYQpQMoAA&defl=zhCN&defl=en。

② 商务智能术语的通俗定义或解释的不完全的列表，参见 http://www.google.ch/search?q=define:Business-Intelligence&hl=de&defl=zh-TW&sa=X&ei=CSojTbiWKJW44AbEtL2GAg&ved=oCAcQpQMoAA&defl=zh-CN&defl=en。

③ 关于 Thomas Samuel Kuhn 提出的范式转换的更多信息，参见 http://en.wikipedia.org/wiki/Thomas_Kuhn。

目　录

第一篇　设计问题、概念和体系结构

第二篇 组件、算法与技术