PTR
PH

计算机技术
译林
精选系列

数据仓库技术指南

[美] Lou Agosta 著
潇湘工作室 译

人民邮电出版社
www.pptph.com.cn

计算机技术译林精选系列

数据仓库技术指南

[美] Lou Agosta 著

潇湘工作室 译

人民邮电出版社

计算机技术译林精选系列

数据仓库技术指南

◆ 著　　[美] Lou Agosta
译　　潇湘工作室
责任编辑　俞　彬

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号
邮编　100061　　电子函件　315@ pptph.com.cn
网址　http://www.pptph.com.cn
北京汉魂图文设计有限公司制作
北京顺义振华印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：787 × 1092　1/16
印张：21.25
字数：515 千字　　2000 年 11 月第 1 版
印数：4 001 – 7 000 册　　2001 年 2 月北京第 2 次印刷
著作权合同登记　图字：01 – 1999 – 2562 号
ISBN 7-115-08867-5/TP·1886

定价：45.00 元

内容提要

本书是直接对业务领导和数据仓库技术的新手介绍数据仓库的非常好的一本书，书中清晰而深入地说明了应用数据仓库的各种好处、风险、技术和过程。在此，读者可以了解到如何将大量信息转换为知识，来减少业务中的不确定因素。本书提出了大量数据仓库应用的个案，以帮助读者在自己的企业中衡量各种关键因素，了解如何使用数据仓库来减少供应链管理的成本，使交叉销售更为有效，具有更好的客户关系、更好的品牌发展，提高产品质量等。书中还介绍了数据仓库项目的生命周期，从计划和设计，一直到部署和优化，可能出现的问题以及如何预防这些问题。另外，本书还涉及了如下内容：创建客户和产品的统一表达；数据质量——数据仓库成功的关键因素；数据仓库设计的基础；数据仓库操作的最佳惯例；基于 Web 的数据仓库、元数据和其他新方法。

版权声明

作者序

作为作者，我当然希望读者从头至尾读这本书。但如果你需要选择先读什么内容，我希望这里提出的一些建议能对你有所帮助。“词汇表”的内容丰富而具体，这是非常有价值的资源。虽然本书尽量在第一次用到技术用语时便进行定义，但如果读者在阅读时跳过某些章节，可能会遗漏一些术语的定义，在这种情况下，读者可以参考词汇表。首先，以下几章是必读的：第1章、第2章、第5章、第6章、第12章、第13章和第18章。另外，以下章节可以对执行官员、业务领导和需要各种信息的客户等提供帮助：引言、第3章以及第19章。项目经理、数据设计师以及设计人员还应当再阅读第4章、第5章、第7章、第14章以及第15章。开发人员可详细参看第8章和第9章（顺便说一句，本书不介绍SQL或编码，读者可查阅其他文献来获得有关的帮助信息）。数据库管理员和信息管理人员会对第10章和第11章感兴趣。业务分析人员、技术专家及各个领域从事信息工作的人员可以参看第15章、第16章和第17章。

序言的真正目的是让读者了解本书的由来。作者写这本书的目的是什么呢？与书本身比起来，序言更像是一道科学的分水岭，它介于发现和证明之间。思想或发明如何成型往往与如何证明无关。许多读者不仅喜欢接受思想和观点的正确性，也希望了解思想和观点的产生背景，而且他们往往从中受益。新事物的出现可能会由于偶然闪现的灵感，或是集现有理论之大成，它对原型增添新的功能以及更多的逻辑结构，而该结构是有待于客观地进行验证并证明其是正确的。但如果您，亲爱的读者，对背景介绍不感兴趣，而是想深入了解数据仓库的具体内容，那就请马上看第1章，以后再回来看序言。作者的工作毕竟是为读者服务，您不看序言的话，作者不会感到伤了感情，您也不会因为没有连续阅读就不能连贯地理解本书。

现在，对那些仍愿意读序言的读者，我想说的是，这本书产生于三个信念。首先，就模式转换而言，数据仓库不是另一种模式转换。其次，借助于数据仓库系统的维与基本业务驱动力及业务规则之间的组合，利用数据仓库可以产生知识，而不是数据。第三，数据仓库系统既是一个信息产品，又是一种创造知识的方法。所以，它更像望远镜上的一片透镜，通过结合光学原理知识，我们便能看到并从而开始了解远方的事物，而这些是我们以前没有想过的。有时候，这要求有很好的洞察力。

事物之间具有错综复杂的关系，需要透过表面才能看到事物的本质。例如，在文艺复兴时期，科学家伽俐略将他的尖形天文望远镜对准月球并看到山（如同地球上的上一样），而不是天堂般的美景时，由于他具有深刻的洞察力，从而理解了包括地球、太阳和月球在内的统一天体系统。但是，当时有学问的宗教学者透过这一奇怪的装置观察时，由于受眼光所限，他们看到的不是山，对他们来说，这些设计是如此陌生，而认为是魔鬼的作品。不要以为这是小事，因为这件事，可怜的伽俐略被宗教裁判所逮捕了。尽管数据仓库设计者和数据采集者不会有这么悲惨的命运，但他们在涉及数据质量的问题上仍应当非常小心。

这里的关键是，本书（不同于它的证明）的出发点可在 3 个动态术语中找到："模式（paradigm）"、"组合（alignment）"和"知识（knowledge）"。这些变化无常的词语经常被人们所滥用。在这里，我们将仔细推敲这些术语，严格定义并精心使用它们。

"模式转换"非常有趣，1962 年，在一本介绍科学知识是如何发展的书中第一次提出这个术语，这本书是很有争议的，它就是托马斯•库恩所著的 *The Structure of Scientific Revolution*。在无数科学知识发展的例子中，库恩选择了从地心说到日心说的转移。从地心说到日心说的转移是一个基本的转移。这是模式转换的绝好例子。实际上，日心说是过去一千年中最伟大的发现之一（1643 年，哥白尼的 *On the Revolution of the Heavenly Bodies（天体演变）*）。许多旧体系中的事实包含在新体系中。以前许多有待澄清的新的事实，也被包含了进来。而且，框架的建立（构筑）还结合了其他许多事实。这一例子强调了，发现是如何产生与发现是如何被证实的关系不大（如果不是无关的话）。哥白尼的模式转换（他的新理论）实际上是一个复杂的逻辑推测，几乎像它要取代的体系一样混乱而且违反直觉。例如，哥白尼继续推论说，行星的轨迹是一个圆而不是椭圆。但如果不利用椭圆运动，新理论所做的许多推论就无法成立。物理和数学再经过两个世纪的逐步发展，才令人满意地证实了我们现在按照严格的科学标准认为是正确的东西。的确，证明中的一个重要部分就是构造数据仓库，来存放 Tycho Branche 和他的学生 Johannes Kepler 对行星的连贯一致的观察数据（出版于 1625 年的 *Tabulae Rudolfinae（Rudof's Tables）*，在 Prince Rudolf 之后）。但随着之后一系列持续、逐步的改进（包括伽俐略、开普勒和牛顿著作中的现代科学史），这个新的模式终于开始看起来像是一个突破了。

从而我们得到了编写本书的本质信念：数据仓库不是模式转换。在企业数据和图形用户界面（GUI）之间——在太阳和地球之间——诱人而肤浅的类比自有它的魅力。但这太有限了。强调的重点并没有从大型机转移到表示层，或从以数据为中心转移到以 GUI 为中心。在完整的计算系统体系结构中，开始就有这两者，而且，它们仍是一个完整的计算系统体系结构中不可缺少的部分。知道这点，也许会使一些认为"模式转换"开始代表着突破或解决方案的读者失望。另一方面，那些被商业出版物中的模式转换搞得疲惫不堪的读者，在得知仍然有可能有很大的进步时，会感到放心。无论哪种情况，我们都不是指像客户机/服务器、关系数据库、网络计算机或基于 Web 的 Internet 解决方案。确切地说，数据仓库是一个系统体系结构，它将设计和产品构造成各种形式。但在每种形式中，我们都要处理决策中需理解并应用的将数据（客户、产品、市场、信息）作为公司资产进行管理的技术规则。数据就是业务这一概念是有着深厚基础的。无论是当作者于 1980 年第一次遇到它，还是在 60 年代后期当统计方法（和其他方法）第一次运用到决策中，这个概念始终是正确的。

促使作者编写本书的第二个思想是“组合”。通过数据仓库产生知识而不仅是产生更多数据的主要方法，是定义仓库的数据维与基本的业务规则及驱动力之间的组合。这一思想指引我们用语言来描述世界。在每一天，语言是所有符号系统（这其中包括计算系统）中最普通的一种，用来表达并约束世界中的各种情形。

自然，语言的使用不限于在实际世界里进行简单的表达（这里为避免累赘，“语言”是代表所有类型计算系统应用的符号）。而且，在与业务规则的组合中，语言既可以作为一种工具，又具有实际作用。它是在各方之间产生关系的一种手段，在业务过程中，它嵌入到系统中，起到协调和交流的作用。从而，“组合”成了计算系统（特别是数据仓库）表现客户、产品和市场之间业务交互的方法的“代理人”。数据仓库表现而且（就像设计者喜欢说的那样）“揭示”目标市场、产品和消费者之间的关系。这些关系反过来作为一种结构，帮助构造者和数据仓库操作人员找到并定义系统的真实世界参照物。例如，需求规划者的预测几乎完全是系统的人工产物。根据它运作的世界就像是一个奇迹。然而，它并不完全是相对的——业务情形的准确特征确实表达出来了，而且系统客观地参照着环境中的事情。如果这像循环，那么它就是确实像。系统设计者（和构造者）是这个循环中的重要部分，这个循环是良性的而不是恶性的。这就暗示了为什么这个过程的本质是反复循环的——数据仓库系统与其使用环境之间的组合不仅仅是被人们发现的，它是被构造出来的（当然，给定的业务实践活动无论在任何情况下总是适用的。在由自动化系统实现或支持的时候，这些商业实践活动是经过抽象、捕获，有时经过了转换，从而超出了人们的认识）。最终得到的系统不仅表示了业务情形，还使得人们能访问它。设计者的艺术就是知道何时停止迭代，因为环境的最小特征已经在系统体系结构中捕获，剩余特征与业务规则相去甚远。从而，我们有了三种事物（系统、世界和系统设计者（和建造者））的双向组合。结果就是知识。

这将我们带到第三个意思多变的词，“知识”。如今，知识意味着取得专利的知识产权到咨询公司内部网上使用 PowerPoint 表示的内容等。直观地说，说某物为知识则意味着该物具有“尊严”，知识表示“优秀”，表示“高度的声望”。在第 6 章题为“信息产品”的一节中，揭示了这种直观性。本书采用的方法暗示着数据仓库系统就是信息产品。而且，数据仓库的确是一个知识来源。它是一个“启用者（enabler）”。它是产品可能成为知识的条件。但是，它本身不是知识。相反，知识产生于问题的提出与答案提供者的交互中。知识不在那些用 SQL 或其他用户交互方法来形成查询的职员头脑中。知识在问题产生者和答案提供者的相互作用中。知识在问题与答案之间的关系中。总之，知识是在信息与行为的协调中产生，这些信息与行为反映在公司承诺回答的问题中，其中至少有一些问题还未产生。因而，数据仓库系统是业务过程中不可或缺的部分，这一业务过程的结果和成果就是知识。

从技术或是业务的角度来看，信息和知识间的关系有所不同。从技术的角度来看，知识与信息是连续的。当信息的质量提高时，它越来越接近于知识。知识就是水平线上的一点，信息始终朝着这个点前进。信息就是一直处于确定的改善过程中的数据。如果这一过程一直延续下去，得到的结果就是知识。从业务的观点来看，知识与信息有着本质的不同。无论信息的质量有多高，都存在一个深渊，将信息与知识分离开。如果不加入一些特别的东西来加强信息，即使是最好的信息也绝不会产生知识。为了产生知识，需要在信息中加入某种东西。这种东西便是承诺。当信息成为业务决策的基础时，其中就暗含着并流动着承诺。数据仓库

系统通过提供机械而实际的知识，提出并解答支持决策问题。本书特别把重点放在了解客户、了解产品品牌在市场中的行为方面。在信息供应链中，无用的信息得到清除，然后产生知识，这并不是奇迹。但有些时候确实像一个奇迹，因为那些特殊公司的承诺不在新闻标题或是表面上体现出来。结果会产生应用于最终目标的有用的知识。知识产生有助于商业利益的结果。这进一步需要实际知识，在这里知识成为了承诺的基础，数据仓库协调了业务过程，处理了基本的业务规则，回答了基本的支持决策问题。第二种通过承诺定义知识的方式要求我们在业务环境中理解知识。数据仓库对知识的这一承诺在决策支持中起到了重要的作用。

目 录

第二部分 设计和建造

第四部分 应用与推测

数据仓库引言：在不确定性与知识之间

本引言主要内容:

- 学习与不确定性共存;
- 模式转换的困扰;
- 通过知识减小不确定性;
- 数据仓库将知识作为特殊的表示;
- 知识的种类;
- 基本的业务规则;
- 三个规则;
- 数据仓库表达业务;
- 复杂的现象，简单的原理;
- 业务与数据仓库的组合;
- 业务的数据仓库映射;
- 不能压缩业务知识;
- 事实的单一版本;
- 知识清单：将“决策”放回“决策支持”。

客户不想要更多的数据……他们想要的是更多的知识。

Andy Kaufman

I.1 学习与不确定性共存

每个年代都必须从全新的角度来重新学习同一课程——每个年代，每一代。现在，我们将学习与不确定性共存。这在我们的生活和业务世界里是确实的。我们在时间的水平轴里引进了产品和服务、忠实用户的奇思怪想以及客户方面的投机行为——所有这些加起来便成为了业务领导者、管理人员以及员工所认为的事务的不确定性。