Ralph Kimball
Richard Merz
著
张丽萍 等 译

Web数据仓库
构建指南

清华大学出版社

Web 数据仓库构建指南

Ralph Kimball
Richard Merz 著
张丽萍 等译

清华大学出版社
北京

Ralph Kimball & Richard Merz
The Data Webhouse Toolkit
EISBN: 0-471-37680-9

图书在版编目(CIP)数据

Web 数据仓库构建指南/金贝尔(Kimball, R.),莫兹(Merz, R.)著;张丽萍等译.—北京: 清华大学出版社, 2005.4

书名原文: The Data Webhouse Toolkit

ISBN 7-302-10502-2

Ⅰ.W… Ⅱ.①金… ②莫… ③张… Ⅲ.计算机网-数据库 Ⅳ.TP311.13

中国版本图书馆 CIP 数据核字 (2005) 第 011983 号

出 版 者: 清华大学出版社 **地 址**: 北京清华大学学研大厦
http: //www.tup.com.cn **邮 编**: 100084
社总机: 010-62770175 **客户服务**: 010-62776969

责任编辑: 常晓波
封面设计: 立日新
印 刷 者: 北京密云胶印厂
装 订 者: 北京鑫海金澳胶印有限公司
发 行 者: 新华书店总店北京发行所
开 本: 185 × 230 **印张**: 16.5 **字数**: 365 千字
版 次: 2005 年 4 月第 1 版 2005 年 4 月第 1 次印刷
书 号: ISBN 7-302-10502-2/TP·7128
印 数: 1 ~ 3000
定 价: 29.80 元

Teh Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse

所赢得的赞誉:

"在当今充斥着商业 Web 服务器日志分析工具的世界中,Kimball 的新作给人们带来了新鲜的活力,就如同他的处女作一样。Kimball 针对点击流数据提出的维度模型证明了标准的 RDBMS 加上超越标准的大脑才是全球电子商务站点的需要。"

——ArsDigita.com 公司 CEO Philip Greenspun

"Ralph 最新的力作"

"Ralph 的最新力作引发了 Internet 的第二次浪潮。在 Ralph 看来,公司不仅可以把 Internet 作为一种销售手段,更可将其用做一种可以增进客户关系的企业战略工具。最关键的是,此书提供了一种崭新的视角,可以帮助公司综合利用基于 Internet 的商业智能和由 Internet 带来的海量客户资料,将'吸引眼球'转变为利润。"

——IBM NUMA - Q 市场营销部全球解决方案方管

William Schmarzo

"作为一种社会现象,我们开始接受这样的现实:Internet 真的改变了一切。伟大的公司开始乐于接受这个现实,并大胆地投资于基于 Web 的信息资产。此书提出了这种设想,也给出了使之付诸实现的实践建议。"

——Oracle 全球服务部商业智能和仓库副总裁

David Fenton

感　谢

本书两位作者都为能够作为专家生活在历史中这样一个激动人心的时期而感到激动。我们在 IBM 大型机还只有不到 8kB 内存的时候就已经开始自己的职业生涯了！由 Web 引起的改变正在以惊人的速度发生，而且很明显我们还只是处于 Internet 时代的初期。因此我们首先要承认 Web 本身作为一种有趣的现象创造了如此多的机会。

本书的编写过程是一件真正的乐事，部分原因是我们从 John Wiley & Sons 的编辑 Bob Elliott 那里得到了极好支持。Bob 和他的职员是这个行业中最优秀的。我们享受着在 Wiley 办公室与 Emilie Herman、Susan McCarty 和 Brian Snapp 共同工作的乐趣。

《智能企业》（*Intelligent Enterprise*）杂志的主编 Justin Kestelyn 特别慷慨地允许我们从 Ralph Kimball 的一些文章中摘录材料以便在本书中使用。

我们还要感谢 Verio 公司的 Neal Kite 和 Kent Millington 对这项工作的鼓励和支持，此外还要特别感谢 Verio 所属 WebCom 子公司的网站管理员 James Mitchell 为第 10 章绘制的大量插图。

与 Ralph Kimball 的前两本工具书类似，本书在将原稿提交给 Wiley 之前经过了详尽地审查。尽管这样的审查步骤会给作者带来更多的工作，但是这些努力使得本书精益求精。此次我们有许多非常出色的审查人员，这里也要感谢他们，他们的名字（按照字母顺序）为：Caroline Ahdab、Maurice Frank、Mike Heathman、Joy Mundy、Margy Ross、Susan Roston、Dick Smallwood、Mark Sweiger、Jerry Tattershall 和 Warren Thornthwaite。

我们还要感谢 Julie Kimball、Ralph Kimball 的生意伙伴和妻子，她在本书的整个写作过程中起着不可或缺的作用。Julie 谈好了商业合同，安排了审查人员，并激励我们前进。

最后，Ralph 和 Richard 还想要感谢我们的妻子 Julie Kimball 和 Toni Merz 在本书写作过程中给予的坚定支持和热情，感谢我们已长大成人的孩子 Sara Hayden Smith、Brian Kimball、Alexey Merz 和 Chris Merz 能够容忍他们父亲的脾气。

目　录

第 2 部分 把数据仓库引入 Web

引　言

就在刚刚过去的一年里，由 Web 产生的影响把 IT 的任务从支持应用继承转变成了发送内容、信息和事务处理能力，而这一切都是通过浏览器界面实现的。或者看起来是这个样子。IT 组织在此之前从来没有受到这么大的压力，要接受多种新的思维方式并重组如此多的界面。

在应对更大的挑战——响应 Web 需求的过程中，千年虫问题只是一个短暂的停顿。Web 不仅仅是用来连接分布式处理设备的一项技术，它还是一种更新更廉价的通信形式。

为变革构建基础

在我们的历史中，有过多次使通信费用骤然降低的变革。在所有这些事件中，当花费明显下降时，通信量就会突然增加，人们通过新的媒介手段变得更具文化性，社会也由此不断改变。通过下面的列表以及图 I.1 展示的内容我们可以知道：

- 16 世纪的印刷术把阅读和出版引入了社会的方方面面
- 19 世纪 40 年代出现的便士邮递（Penny Post）使人们仅花 1 便士就能把一封信寄到英格兰境内的任何地方
- 19 世纪 40、50 年代出现在欧洲和美国的电报是第一种即时长距离通信的形式
- 20 世纪早期出现在美国的乡村免费邮政派送（Rural free postal delivery）把新闻、商务和个人通信带给了遍布在整个国家的许多人
- 20 世纪 20 年代出现的电话使得任何市民都可以与其他市民进行实时交谈
- 20 世纪 30 年代，无线电广播可以在瞬间跨国度传播新闻、文化、语言和言论
- 20 世纪 50 年代，电视延续了由无线电广播所产生的影响，区别在于使用了更引人注目的媒体
- 20 世纪 90 年代的 Web 则是对邮局、电话、无线电广播和电视的一种动力强劲通往高带宽方向的继承发展

20 世纪 90 年代的 Web 革命与早期革命的最大不同可能在于此前者发展的惊人速度。在六年里，这个世界的重要组成部分已经改变了它的通信方式、交易方式，以及使用信息的方式。

这一次，轮到作为 IT 专业人员的我们来为变革构建基础了。我们要使旧应用程序适用于 Web，这样每个人都可以在任何地方使用 Web 进行输入和输出。我们要能够为我们的客

户、生意伙伴以及内部雇员提供适用于 Web 的界面。我们还要能从档案中得到 Web 要求的各种格式的多媒体数据(包括数据库、电子数据表、未组织的文本、图像、地图,甚至可能是音频和视频等等)。与此同时,我们还应该处理由于保护和适当发布所有这些信息而引起的安全机密性问题。

构建这个新的 Web 基础结构既令人兴奋同时也让人毫无头绪。我们应该从哪儿开始?我们如何才能把问题缩小到恰当的范围,使之易于处理并且使得我们可以针对这一挑战应用已有的系统构建技巧?

Web 数据仓库

过去的 10 年中,在 Web 革命形成气候之前,IT 组织已经学习了如何向内部分析人员和管理人员发布组织的数据资源。发布就是数据仓库的中心任务。

数据仓库是 IT 的核心功能之一。通过多种方式,数据仓库在基于 OLTP(联机事务处理)系统"采集进数据"后,实现"读出数据"。根据在数据仓库方面十多年的经验,我们已经对什么是数据仓库及 IT 如何引入技术以有效发布所有这些数据有了相对成熟的理解。

Web 革命并没有取代人们对数据仓库的需求。事实上,Web 革命把每个人对通过 Web 浏览器界面无缝发布各种信息的期望大大提高了。数据仓库数据的拥护者从内部管理人员扩展到了周边客户、合作伙伴以及更多的内部员工。Web 关注"顾客经验"(customer experience),这使得许多组织进一步意识到要了解顾客并给予顾客有用信息。

由于在许多情况下控制或分析 Web 经验的引擎必定是数据仓库,因此 Web 革命把数据仓库推到了主要舞台上。为了实现这一强化功能,必须对数据仓库进行调整。数据仓库的本质要求和它过去 10 年中的样子有所不同。图 I.2 为网站顾客和网库之间的关系。

我们将把数据仓库的重生称为 Web 数据仓库。

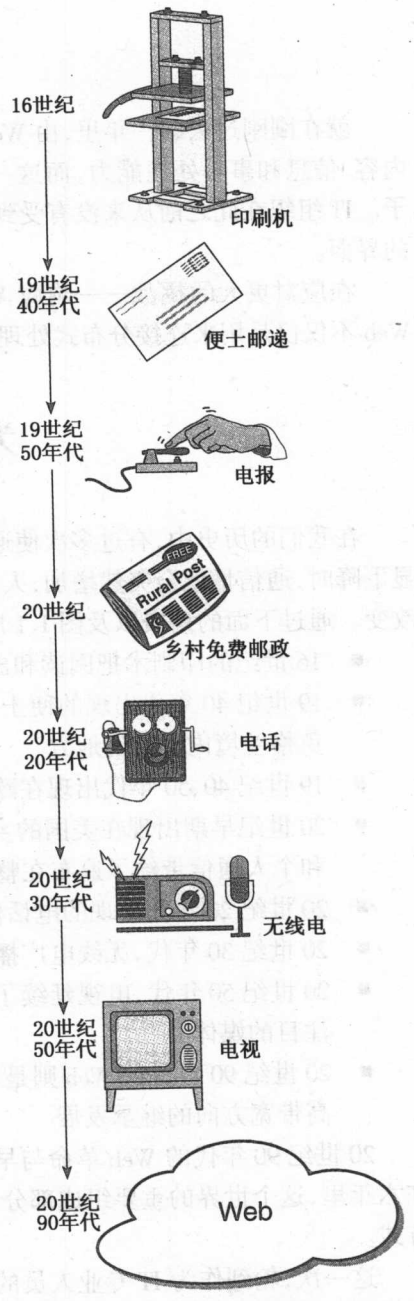

图 I.1 对社会产生深刻影响的媒体

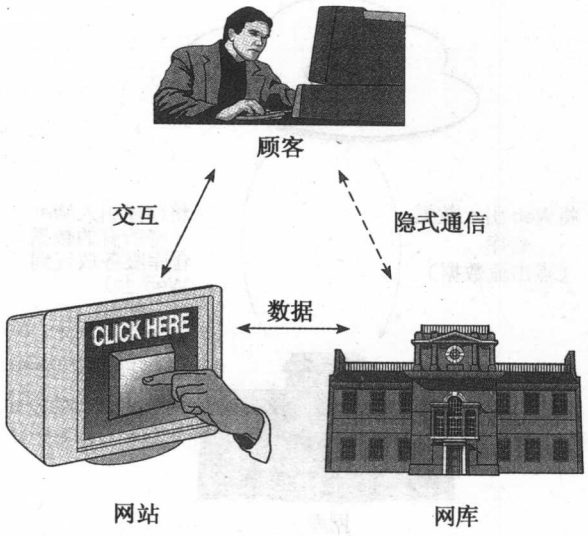

图 I.2 顾客、网站和网库

本书内容

本书是关于 Web 数据仓库的。我们将定义 Web 数据仓库并说明它与传统数据仓库的区别。最重要的是,我们会详细说明如何建立 Web 数据仓库,或者是从零开始,或者更可能的是改写并扩展现有的数据仓库使之成为完全意义上的 Web 数据仓库。

我们将看到 Web 数据仓库有两种特性,这些将体现在本书的两个部分中,请参见图I.3。前半部分为将 Web 引入 Web 数据仓库。当个体通过浏览器与远程网站进行交互时,Web 本身就是一种极好的行为数据源。尽管这种点击流(dickstream)数据在许多情况下是原始且未经修饰的,但是它具有提供使用 Web 媒体的任何人所做每个动作详细细节的潜能。像亚马逊那样的点击流数据形成的原始数据源已经成为现有已知的、最大的文本与数字数据库,这使得从电信到保险这样著名的大型数据库都显得相形见绌。当我们把 Web 引入 Web 数据仓库时,我们把这个巨大的、未经提炼的数据源导入 Web 数据仓库使之自我分析,除此之外,也可以使之与现有的较为传统的数据源保持一致并互相结合。

本书的整个前半部分是关于了解和掌握点击流资源,以便把这些资源存储在 Web 数据仓库中并予以有效地利用。

Web 数据仓库的第二个特性将在本书的第二部分中进行介绍,该特性把现有的数据仓库引入 Web。描述这个过程的最简洁的方式是我们已经不再处于客户端/服务器环境下,我们是处于适用于 Web 的环境中,并且体系结构中有比过去更多的层。

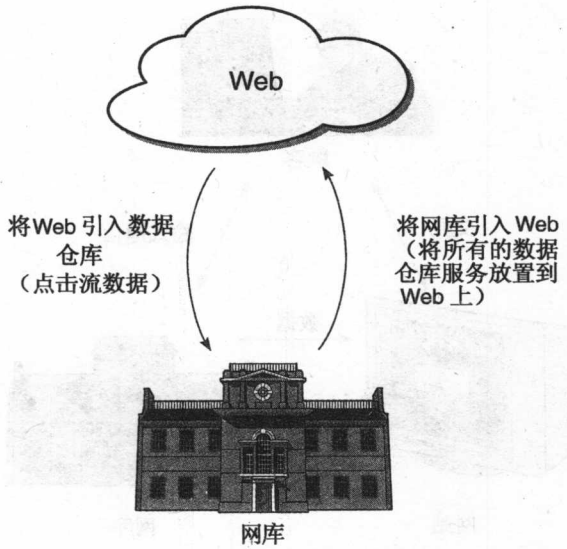

图 I.3 网库的两种特性

把数据仓库引入 Web 意味着要使所有的数据仓库接口通过 Web 浏览器就能使用。如一来，整个过程可分成数据记录、简单特殊查询、复杂报告、正式应用程序开发，以及最后的数据库和系统管理。

把数据仓库引入 Web 也意味着要一次并彻底解决完全分布式环境的问题。Web 数据仓库是比完全集中式数据仓库方法更有意义的替代产品。Web 数据仓库与 Internet 本身一样，也难以实现集中化。但是不管怎样，我们必须设法接受这样一种设计理念：要允许分布在 Web 之上的分散岛式数据仓库以一种有效的方式彼此可见并通信。本书的目的之一就是要说明这种理念的含义。我们将学习如何建立并控制分布式 Web 数据仓库。

本书不对 intranet 网库和 Internet 网库进行更多的区分，这是因为我们认为二者在设计问题、技术和开放程度方面实质上是一样的。我们确信在实际操作中我们无法限制一个组织只使用纯粹的 intranet 方法。几乎从一开始，人们就想通过 Web 连接骨架远程使用他们的浏览器。而这时，人们就会遇到 Web 配置的所有问题。所以我们从一开始就可能需要处理这类较大的问题。

本书适用对象

本书适用于那些像之前介绍过的在工厂组织中担当各种 Web 和数据仓库职责、打算建立部分或整个 Web 数据仓库的设计人员和项目经理。如果你正是这样的专业人员，那么你或你的员工可能需要：