

Clickstream Data Warehousing

点击流数据仓库

[美] Mark Sweiger Mark R. Madsen 著
Jimmy Langston Howard Lombard

陆昌辉 张光剑 陈佐 张丽 等译
邓苏 审校

数据仓库与数据挖掘技术应用丛书

点击流数据仓库

Clickstream Data Warehousing

[美] Mark Sweiger Mark R.Madsen 著

Jimmy Langston Howard Lombard

陆昌辉 张光剑 陈佐 张丽 等译



邓苏 审校

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书解释了构建点击流数据仓库所需要的 Web 技术和 IT 基础设施, 并对设计、实现点击流数据仓库的整个过程提供全面的指导, 包括: 计划、人员分工以及管理整个工程; 使用创新的元模式设计模板设计点击流数据仓库; 挑选合适的数据仓库软件和存储子系统以支撑点击流数据仓库; 建立抽取、变形和装载 (即 ETL) 机制, 以及将数据传送给分析这些数据的终端用户。

本书主要面向学习或在工作中运用点击流数据仓库技术的教师、学生或工程技术人员, 特别适合对数据仓库技术有所了解, 但希望进一步提高构建点击流数据仓库能力的应用开发人员。

Copyright © 2002 by Mark Sweiger, Mark Madsen, Jimmy Langston, Howard Lombard.

All rights reserved. Authorized translation from the English language edition published by John Wiley & Sons, Inc.

本书简体中文专有翻译出版权由 John Wiley & Sons Inc. 授予电子工业出版社, 未经许可, 不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号: 图字: 01-2002-5363

图书在版编目 (CIP) 数据

点击流数据仓库 / (美) 斯韦格特 (Sweiger, M.) 等著; 陆昌辉等译. —北京: 电子工业出版社, 2004.1

(数据仓库与数据挖掘技术应用丛书)

书名原文: Clickstream Data Warehousing

ISBN 7-5053-9320-0

I. 点… II. ①斯… ②陆… III. 数据库系统 IV.TP311.13

中国版本图书馆 CIP 数据核字 (2003) 第 102122 号

责任编辑: 黄志瑜

印 刷: 北京增富印刷有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

经 销: 各地新华书店

开 本: 787×980 1/16 印张: 26.75 字数: 472.8 千字

印 次: 2004 年 1 月第 1 版

印 数: 3 000 册 定价: 58.00 元

凡购买电子工业出版社的图书, 如有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系。联系电话: (010) 68279077。质量投诉请发邮件至 zlts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

出版说明

如果没有对海量数据进行科学分析的能力，沃尔玛的老板再精明，也绝对想不到“啤酒与尿布”这两个风马牛不相及的东西之间还有着千丝万缕的联系。而将它们放在一起，竟然增加了啤酒销量，可见数据分析的巨大威力。

信息系统数年中收集了海量数据，且数据还正以指数级增长，企业迫切地需要高效、精确、科学地分析数据，以找出其背后的寓意，进而了解企业的经营状况和外部环境，做出科学的决断，在现代激烈的竞争中胜出。所以，如何将数据点石成金，更是摆在我们面前很现实也很诱人的一个问题。

现在，很多人已经意识到数据中潜在的大量商机，并踏踏实实地进行着从数据中沙里淘金的工作。特别是在信息化的大潮中，上至政府，下到企业，从银行到电信，再到网站、超市，人们都希望用数据分析这根魔杖赢得先机。与此同时，人们也在期盼着相关书籍，以便工作中学习参考。在广泛征询专家和用户的基础上，秉着选题全面、内容经典、译者严谨的原则，我们适时地推出了这套《数据仓库与数据挖掘技术应用丛书》，以飨读者。本丛书有如下几本：

- 数据仓库基础
- OLAP 解决方案：多维信息系统的构建技术
- 数据仓库工具箱：维度建模的完全指南（第二版）
- 数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法
- 数据仓库及其在电信领域中的应用
- 疑难数据仓库专家解决方案
- IBM 数据仓库和商业智能工具
- 可视化数据挖掘：数据可视化和挖掘的技术与工具
- 点击流数据仓库
- Web 数据挖掘：将客户数据转化为客户价值
- 企业信息工厂
- 机器学习与数据挖掘：方法和应用

本丛书既包括商业智能（BI）的基础——数据仓库（DW），也包括数据仓库上的两类不同目的的数据增值操作——联机分析处理（OLAP）和数据挖掘（DM）；既覆盖基础理论，如数据仓库基础，又提供不同领域的解决方案，

如数据仓库在电信、银行、保险等领域的应用。

本丛书来自国外数据库领域一些著名作者的畅销书，以及国内第一线实施者的精心总结。如一直位居 AMAZON 畅销书榜的数据仓库领域的畅销书作家 Ralph Kimball 的《数据仓库工具箱：维度建模的完全指南（第二版）》、《数据仓库生命周期工具箱：设计、开发和部署数据仓库的专家方法》，数据仓库之父 William H.Inmon 的《企业信息工厂》（Corporate Information Factory）等。

丛书的译者均来自工作在该领域一线的人员，既有该领域的理论和实践经验，又具备中英文翻译的功底。且多位译者先前均已读过原著，所以，自感翻译的过程不再是枯燥，而是情趣盎然，乐在其中。

出版高品位、高品质的图书是博文视点的努力目标。希望您对我们的工作多提宝贵意见。您的意见是我们创造精品的动力源泉。

如果您希望将您的工作经验感悟等总结成书，我们将为您提供一流的服务，共创精品图书。

我们的联系方式如下：

地址：北京复兴路 47 号天行建商务大厦 604

邮编：100036

电话：010-51922832，68216158

传真：010-51922823

E-mail: jsj@phei.com.cn; zsh@phei.com.cn

博文视点资讯有限公司

2003 年 10 月

译者序

随着电子商务的普及,越来越多的从事电子商务的公司逐渐认识到用户访问公司网站所产生的点击流数据中包含大量客户信息。通过对这些信息的处理,可以帮助建立完备的客户资料与用户数据库,这引发了学术界和产业界对点击流数据仓库的广泛关注。在关于论述点击流的最有影响的两部著作中,除了 Ralph Kimball 和 Richard Merz 的《数据仓库工具箱》之外,最为详细和实用的就是 Mark Sweiger 和 Mark Madsen 著的这本《点击流数据仓库》了。

本书的特点是内容丰富、概念清楚准确、实例丰富、讲解详细。同时本书内容组织的逻辑性很强,它的每一章都是建立在前一章的基础上,引导读者在头脑中逐步形成一个点击流数据仓库的框架。然后,结合实例详细说明创建点击流数据仓库的过程及其所遇到的众多技术难题。总之,这本书系统地介绍了点击流数据仓库的所有问题,是一本值得认真阅读的好书。

本书论述了点击流数据库的基础知识及构建点击流数据仓库的全过程。读者对象是学习或者在工作中运用点击流数据仓库技术的教师、学生或工程技术人员,特别适用于对数据仓库技术有所了解,但希望进一步提高构建点击流数据仓库能力的应用开发人员。

本书由陆昌辉组织翻译,同时也负责全书的统稿和校对,参加本书翻译的有陆昌辉、张光剑、陈佐、陶勇、张丽、杨柳、郭媛妮、郭勇、易非、李蕊、凌纯清、刘瑜、蔡丽、卿立军、杭志、杨国锴、刘炼、温建邦、胡勇辉、刘泽栋、彭浩等。另外,张光剑也对全书进行了校阅。大家集思广益、畅所欲言、反复推敲,形成了对原文的基本一致和比较透彻的理解。

在翻译过程中,我们尽量融入了从事软件开发和构建数据仓库的实践体会,但由于知识局限和时间仓促,难免存在一些翻译错误和内容上的遗漏,敬请广大读者批评指正。我们会接受您的批评,并向原书作者表示歉意。

陆昌辉

2003 年 3 月于长沙

引 言

这本书向你介绍了如何设计和建立一个数据仓库来对网站上的用户行为进行分析。网站上的每个用户的每一次点击都会被记录在网页服务器的日志文件中，通过把这些点击流历史信息加载到一个设计良好的点击流数据仓库中，便可以转化为商业智能的珍藏品。不管进行纯粹的电子商务，还是从事传统的商业活动，只要是把网站作为众多商务渠道之一，那么这本书中所介绍的技术就可以使用。最终的点击流数据仓库既可以作为一个单独的实体，又可以成为大型企业数据仓库系统的一部分。

点击流数据仓库不同于早期的数据仓库，如财务数据仓库、供销数据仓库或者支持客户关系管理的数据仓库。比如说，点击流数据仓库包含了一些新的基于网页的技术和底层结构，这些技术和结构是很多使用传统数据仓库的组织所不熟悉的。还有，比起那些典型的传统的数据仓库环境，点击流数据仓库有更为丰富的源数据，增加了数据提取和模式设计的复杂性。最后，点击流数据仓库是巨大的，它通常包含 10 亿行左右的事实表，它把底层的硬件技术和结构推向了极限。

如果你有一个网站，那么点击流数据仓库同样是你商业上的必需品。网站上发生的任何事情对企业来说都是完全透明的，除非你下载并分析最终的点击流。没有点击流数据仓库，就像乘坐的飞机没有窗户或驾驶舱内没有仪器工具一样，你会毫无方向地乱飞，最终是会坠机的。对于那些没有点击流数据仓库的企业活动，我们会感到不可思议，当然这只是时间的问题。这些企业中的某些最终会倒闭，我们不会感到吃惊。因为它们没有深入研究商业的走向，冒昧地以国内的工业界为例。点击流数据仓库还没有成为主流，这是因为没有它，你可能仍然在以最快的速度前进。

读者范围

这本书适合任何将要参与数据仓库工程项目（包含基于 Web 的数据）的人员，尤其适用结构设计师、数据库管理员、数据建模人员和工程管理人员，因为它介绍了丰富的基于 Web 数据主题的详细知识，这些知识在其他地方是找不到的。同时这本书的内容也给工程设计人员提供一个通用的框架，该框

架是任何点击流数据仓库工程的必备基础。那些准备参与实现点击流数据仓库工程的技术顾问同样要读这本书，因为他们是这个工程的关键性人物。

我们还要建议那些网站开发人员和网站管理员读读这本书，以便于更好地理解商业用户的需求和有关的数据问题。这些信息会说明网站是如何实现的，这样的话，网站开发和管理人员就会提高点击流分析的积极性而不再排斥它。通常网站的开发人员是负责电子商务信息技术创新的，如果网站不被设计成能分析用户行为的方式，那么进行适当的网页分析就会变得很困难，甚至是不可能的。

那些 Web 应用服务器产品和个人专有化工具的设计者应该真正地读懂这本书。这些产品中的相当一部分并不能很好地支持对点击流数据的捕获。因为这些产品要动态地生成网页，所以它们要提供记录网页生成的点击流信息的能力，这样才能跟踪并分析用户行为的历史信息。如果没有这种能力的话，那么使用这些工具的商业活动在管理电子商务的时候就会面临巨大的挑战。

必备知识

这本书假定读者已经熟悉了多维数据建模和星型模式设计的基本知识。有关这些知识的介绍推荐一本由 Ralph Kimball (Wiley, 1996) 写的《数据仓库工具箱》。本书还要求读者了解数据仓库开发实践和技术的基本知识，推荐参考书为 Ralph Kimball, Laura Reeves, Margy Ross 和 Warren Thornthwaite 写的《数据仓库生命周期工具箱》。

在写这本书的时候，市面上只有一本介绍点击流数据仓库的书，那就是 Ralph Kimball (Wiley, 2000) 写的《数据仓库工具箱》。我们认为此书对创业者来说是一本很好的指导书，它完全可以与《数据仓库工具箱》相媲美，它对底层的 Web 技术和实现点击流数据仓库所需的设计技术进行了详细的介绍。

本书的结构

这本书分为两部分。第 1 部分（即第一篇）叫做“点击流数据仓库的构建基础”，该部分详细地介绍了 Web 技术及其底层结构，因为它与点击流数据仓库有关。作者发现在很多情况下，数据仓库工程的设计人员不熟悉新的 Web 技术和它的底层结构。这些知识的缺乏会导致工程进度的不连贯，用户使用的热情不高，甚至导致工程的失败。没有这些知识，工程设计人员就会发现很难与网站的开发方交流，给问题的解决带来更多不确定的因素。

第 1 部分有 4 章：

第1章“典型电子商务体系结构”，它描述了电子商务信息系统体系结构中的组成元件，并说明它们如何与点击流数据仓库建立联系。内容包括终端用户系统、因特网服务提供商（ISPs）、Web 服务器、应用服务器、内容缓冲服务器、广告引擎、搜索引擎、商业事务服务器、公用因特网、通用载体、私有内部网，当然还有点击流数据仓库。这种规范的体系结构将贯穿于本书后面的所有章节中。

第2章“Web 应用环境”，它描述了不常用的无状态 Web 应用环境，并介绍了一些基本概念，像超文本传输协议（HTTP）、HTTP 头部域、查询字符串、通用网关接口（CGI）、Cookies、Web 服务器日志记录、脚本语言、Web 服务器和应用服务器。这本书剩下的部分假设读者已经熟悉了这些概念，对体系结构基础来说，这一章是关键性的一章。

第3章“点击流数据源和 Web 服务器日志文件”，它在深层次上分析点击流数据仓库的基于 Web 的数据源。它包含的内容有 Web 服务器日志文件的基础知识和标准日志文件的格式，并对在面对基于 Web 的数据时会遇到的很多问题进行了详细的分析，比如说在处理其他数据源（像缓冲服务器、Web 应用服务器和媒体服务器）之前所碰到的困难。日志文件数据是主要的点击流数据源，因此我们认为这一章会成为有关格式和日志文件数据信息的基准源。

第4章“用 Cookies 和其他机制跟踪用户身份”，它重点研究了如何在网页上和数据库里建立并跟踪用户身份的关键问题。这一章描述了用于管理用户对话，识别用户并跟踪他们的访问，以及有关数据管理问题的机制。同时也讨论了有关用户身份和用户隐私的商业和伦理的问题。所有这些问题在任何一个点击流数据仓库的实现过程中都需要认真考虑。

如果读者已经对基于 Web 的基础结构和第1章中所涉及的数据有了深刻的理解，那么他就应该从第2部分开始学习“逐步建立点击流数据仓库”。这部分涵盖了这本书中剩下的所有内容，实际可以成为一本有关如何设计并实现点击流数据仓库的手册。它包含了所有的问题，从工程的人员分配和管理到模式设计，到抽取转换和装载，再到终端用户的分析。正像以数据仓库为中心的设计人员会在第1章中学到很重要的有关 Web 结构的问题一样，以网站为中心的设计人员，像网站的设计者和网页的管理者，他们会在第2章中学到有关为了支持一个好的点击流数据仓库，网站需要有哪些设计特色的问题。我们真诚地希望这本书能成为网站设计人员和数据仓库设计人员之间交

流的桥梁，能为点击流数据仓库的工程带来更好的成果。也就是说，第 2 部分的读者群理所应当是点击流数据仓库的设计人员。

第 2 部分（即第二篇）有 5 章。

第 5 章是“点击流数据仓库工程的计划、管理及其人员配置”，它包括一个点击流数据仓库工程所有必经阶段的概括性描述。在“学习难点”部分，我们深入讨论了在一个典型的工程中如何避免那些典型的问题和缺陷。该章的结尾部分探讨了以下问题：工程中的角色，人员需求，以及如何把一个工程小组组织起来，其中包括一些内在的组织图。

第 6 章“点击流数据仓库的元模式”，是本书中最重要的章节之一。这一章描述了点击流数据仓库的元模式，它是一种通用的模板，用来指导任何点击流数据仓库模式的逻辑设计。这个模式包括用户行为事实表和 10 个可能维的表，包括用户维、内容维、行为维、用户时间维、财政时间维、物理位置维、Web 地理位置维、网站地理位置维、内部促销维和外部促销维。该模板像一座桥，架起了商业用户和模式设计者之间沟通的渠道；它确保一个点击流数据仓库中所有重要的事实、维度和属性都会在逻辑模式的设计中被考虑到。

第 7 章“实现合适的点击流数据仓库的技术基础”，它包括那些你在任何其它的物理数据库设计中，以及任何点击流数据仓库的技术框架中都找不到的信息。我们发现在大多数有关数据仓库的书中，作者掩盖了对物理数据库设计和技术框架问题的解释，他们都把这个非常复杂的主题留做读者的习题。该章讨论了以下内容：

- 有效的大容量批处理加载技术；

- 表分割，包括范围、散列和复合分割；

- 索引，包括 B 树索引、位图索引、功能索引和索引分割；

- 连接，包括星型连接、Oracle 的星型转换和分割的连接；

- 维度聚合的管理，包括创建候选的聚合、数据库优化器的聚合意识、聚合导航和物化视图；

- 数据库的并行性，包括如何用块、关键字和散列分割来并行化数据库的操作以及并行数据库处理；

- 对 SQL 进行新扩展以支持点击流数据仓库，包括 SQL 语句中的前 N 级，用于创建聚合的 ROLLUP 操作，用于交叉表的 CUBE 操作；

- 磁盘驱动和逻辑卷标管理，包括串联、剥离、镜像、RAID 阵列等，及如

何把表、索引、表空间、日志卷标和临时表空间映射到卷标和磁盘驱动上去；

如何从各种各样的销售商那里选取产品，包括数据库软件销售商、逻辑卷标管理软件销售商和磁盘子系统销售商。

第 8 章“建立点击流的抽取、转换和装载机制”，描述了对点击流数据仓库进行抽取、转换和装载的基本知识。这一章详细讨论了建立一个点击流的抽取、转换和装载机制（ETL）所必需的 8 个步骤。同时它也在元模式的基础上通过一个详细的例子说明在 ETL 期间如何把源数据处理成点击流数据仓库中的数据。

第 9 章“点击流数据仓库中的数据分析”，为如何在大的点击流数据仓库中进行查询的问题提供实际可行的解决办法，包括对关系型联机分析处理（OLAP）、多维 OLAP 和混合 OLAP 查询环境的讨论。这一章也说明了所选的 OLAP 查询环境如何利用有效的查询技术（像分割消除、物化视图及在服务器端进行汇总计算）来满足终端用户的性能期望。

本书网站

这本书有一个网站，即 www.wiley.com/compbooks/sweiger。该网站是用来扩展本书的信息，并为点击流数据仓库社区提供一个寻找更多启发和交流的场所。虽然这个网站还需要一定时间进一步加以完善，但是它已经以电子版的形式提供了最好的额外信息，比如说微软的一个工程项目——点击流数据仓库的工程计划，有用的脚本及由本书作者和其他人所写的文章的链接。我们同样提供那些很新的、可能会经常变动的信息，这些信息对这本书而言可能不太合适。我们还提供一些更一般的有关数据仓库的信息。

致谢

我要感谢 Sun 微系统的 Elizabeth Schaedler，他最初的鼓励是我写这本书的动力。我还要感谢 Mark Madsen，因为是他帮助我完成这项工作，没有他就没有这本书。

我还要感谢 Stuart, Vinita Nelson, John 和 Linda Gensler，感谢他们在我艰难的写作过程中给予我的支持。

最后，我要感谢 John Wiley&Sons 的 Bob Elliott，感谢他在帮助我们出版这本书时所做出的努力。

Mark Sweiger

2002 年 1 月

首先，我要感谢我的妻子 Kim，是她在我工作过程中一直支持着我，并几个月来忍受着黑夜里敲击键盘的声音。接着，我要感谢的人是 Bob Kondoff 和 Ralph Kimball。没有 Bob 的鼓励和督促，我不会进入这一领域并最终取得成功。至于 Ralph Kimball，虽然他本人可能并没有意识到，但是他却在我第一次进行数据仓库工程的模式设计时，教会了我几乎所有的东西。另外，我还应感谢《DecisionWorks》的作者 Bob Becker，从他那里学到了他书中没有讲到的一些知识。

同样我要感谢 Tom Dopirak，几年来他一直在默默地支持我。JP Gordon 和 Alice Liu 给我提供了产品信息，避免我犯一些低级错误。我还要感谢 Alex Lubarov, Dave Stauffer 和 Larry Murdock，他们和我一起参与了 Web 开发工程，并为我提供很多有关这本书的技术问题。

最后要感谢的是 Mark Sweiger，他多年来一直和我工作在一起，并协助我完成这本书的写作任务。

Mark Madsen

2002 年 1 月

对《点击流数据仓库》的高度评价

“对于研究点击流数据的严谨的数据仓库设计者来说，《点击流数据仓库》是一本好书。作者以清晰的风格介绍了这个重要的客户行为数据源中的复杂问题，并结合工程实践中关于多维数据仓库设计技术的知识写出了这么一本很有用的书。”

Ralph Kimball

President, Ralph Kimball Associates

“Mark Sweiger 和他的伙伴创建了一个能从电子商务产生的所有点击流数据中赢利的基础结构。点击流数据的激增要求在数据仓库中能够捕捉这种数据来确保满足你客户的需求。本书的作者先从分析一个典型的电子商务结构的特点开始，接着解释如何通过创建一个点击流数据仓库来增强它的性能。对于那些在电子商务的基础结构上有很大投资并打算增加其 ROI 的公司，这本书是必读的。”

Ron Powell

Publisher and Editorial Director, DM Review

“点击流数据仓库属于现代竞争性商业中最重要的资产之一。但是很不幸，由于缺乏对于这个相对奇怪资产（也就是说在使它的商业价值达到最大化的过程中没有任何挑战）的计划、开发和管理的相关可用信息，可能会引发一些不确定因素，从而威胁到那些经验最丰富的数据仓库应用开发团体。在这本新书中，Mark Sweiger, Mark Madsen, Jimmy Langston 和 Howard Lombard 补充了 Ralph Kimball 和 Richard Merz 具有突破性成果的著作《Web 数据库工具箱》的部分内容，对点击流数据仓库这个统一体的各个方面给出了详细实用的建议，范围涉及体系结构、数据源、用户身份追踪、工程管理和基础结构。《点击流数据仓库》将会成为这个课题的经典手册。”

Justin Kestelyn

Editor-in-Chief, Intelligent Enterprise

“这本书还包括了用以确保公司从他们网站的点击流数据获得尽可能大的利益的所有过程。通过正确捕捉和分析点击流数据所得到的结论能很容易地为那些实现了点击流数据仓库的公司带来竞争上的优势。这本书除了介绍所有用于点击流数据仓库的结构性单元外，还给出了对于设计网站的建议，从而可以获得最有用的点击流数据用于分析。不管是作为初学者的帮助性材料，还是作为那些已经创建了点击流数据仓库的人员的参考材料，这本书都将成为必需的非软件性工具。”

Jean Schauer

Editor-in-Chief, DM Review

“和电子商务数据打交道的每个人都应该读读《点击流数据仓库》这本书。从某种意义上来说，它是一本很受欢迎的参考材料。它不仅指明了从 Web 数据建立数据仓库的潜在缺陷，还提供了一些很好的解决方法。这本书的技术性很强，对项目的参与者很有帮助，同时它的结构清晰简明，对项目的管理者也很有用。我就很喜欢这本书。”

Herb Edelstein

President, Two Crows Corporation

“如果你正在思考如何避免重复很多商业网站中的错误，那么你就必须读这本书。成功的商业管理需要好的商业决策。而创建点击流数据仓库则是在电子商务中获取实用的商业决策的惟一途径。”

Elizabeth Schaedler

Enterprise Systems Business Development Manager, Sun Microsystems

“如果你是一个正在考虑或正参与开发点击流数据仓库工程的人员，那么你所需要的知识都包含在这本书中。书的内容按逻辑顺序来进行组织，每一章都是建立在前面一章的基础之上，使读者在头脑中形成有关点击流数据仓库的一个框架，包括它的体系结构、各个单元、有用的工程练习及成功的因素。关于以直接可理解的方式处理 Web 结构、Web 事务及内部用什么来捕捉，组织和使用有用的点击流数据的知识，这对任何新手或经验丰富的数据仓库管理人员来说都是十分有用的。并且书中还有大量的例子和建议来详细说明怎样解决那些在实现点击流数据仓库过程中所遇到的很多技术难题。

John Gensler

Senior Consultant, KPMG Consulting, Inc.

“对于如何写实现指南，该书是一个好的模板。在 Ralph Kimball 博士等人著的《Web 数据库工具箱》这本书的基础上，Mark Sweiger 和他的同事们深入研究了现实世界中复杂的 Web 数据和电子商务。这本书之所以特别有用，是因为它从深度和广度两方面来综合考虑问题。如果你正负责实现一个点击流数据仓库，这本书将是你确保成功的最好选择。”

Laura L. Reeves

Principal, Starsoft Solutions, Inc.

“任何需要有效处理有关如何传递有用点击流数据仓库的专业人员都会发现这本书很好地解决了网站点击流的技术难题。作者并没有介绍他们自己的数据仓库，而是概括了到目前为止大家熟知的有关多维数据模型和数据库开发实践的基础知识，并把它作为发布一个有效点击流解决方案的基础。”

Bob Becker

Vice President, Decision Works Consulting, Inc.

“《点击流数据仓库》建立在前几年引入的数据仓库概念的基础上，同时增加了关键的数据元素——点击流，它是目前很多市场商人急需的东西。我们中的某些人可能会甜蜜地回忆起以前的那些日子——我们每周要从事务系统中提取数据，提供商业运行所需的所有决策支持。不管你是否喜欢，这些日子将一去不复返。商业管理者正急切地要求能查看到在 Web 日志上积累的数据。这本书的作者做了一件令人震惊的工作，他们引进了新的概念，再次讨论了那些经过检验的正确理论，并在更深的层次上寻找问题的答案。因此受到了高度的评价。”

Dave Stauffer

Director of Site Development, Walmart.com

“这本书从新的角度研究了传统的数据仓库问题。如果你的业务中包括 Web 介绍并且你很在意它的成功，那么这本书将是必读的。它是为每个参与数据仓库开发工程的人员设计的，它描述了成功实现该工程所能遇到的所有难题，并提供了其解决方法。我期待着能把自己所学到的新知识运用到以后的工程中。”

Curtis McClendon

Oracle Team Manager, Qwest Cyber. Solutions Inc.

“《点击流数据仓库》在传统的数据仓库中增加了一个 Web 维，除了增加销售额外，还能分析跟踪并预测用户的行为。”

Gary Hallmark
Architect, Oracle Corporation

目 录

第一篇 点击流数据仓库的构建基础

第 1 章 典型电子商务体系结构	(3)
▶ 1.1 电子商务体系结构简述	(3)
▶ 1.2 Internet 服务提供商——ISP	(5)
▶ 1.3 多种 Internet 连接服务	(6)
▶ 1.4 多种物理 Web 服务器	(7)
▶ 1.5 不同类型的备份 Web 服务器	(7)
▶ 1.6 集群商业交易系统	(9)
▶ 1.7 点击流数据仓库	(9)
▶ 1.8 规范的电子商务体系结构	(12)
▶ 1.9 小结	(12)
第 2 章 Web 应用环境	(14)
▶ 2.1 状态无关的 HTTP 事务模型	(14)
▶ 2.2 HTTP 事务间的信息交互	(16)
▶ 2.3 查询字符串	(18)
▶ 2.4 Cookies, 用户标识以及 Web 服务器日志记录	(18)
▶ 2.5 站点点击、页面访问和用户会话	(19)
▶ 2.6 使用 CGI 调用其他可执行程序	(20)
▶ 2.7 使用脚本语言记录用户行为	(22)
▶ 2.8 网页服务器、应用服务器及动态生成 Web 页	(24)
▶ 2.9 动态生成 Web 页和搜索引擎	(26)
▶ 2.10 小结	(27)
第 3 章 点击流数据源和 Web 服务器日志文件	(28)
▶ 3.1 Web 服务器日志	(28)