Broadview
WWW.BROADVIEW.COM.CN

数据仓库与数据挖掘
技术应用丛书

数据仓库
及其在电信领域中的应用

殷云峰 吴唯宁 李剑威 韩 洁 编著
刘 虹 审阅

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据仓库与数据挖掘技术应用丛书

数据仓库及其在电信领域中的应用

段云峰　吴唯宁　李剑威　韩　洁　编著
刘　虹　审阅

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书首先介绍了数据仓库及其一些基本概念，前 3 章详细介绍了数据仓库、数据挖掘、OLAP 分析等技术内容，并对数据仓库建设的过程及应注意的问题等进行了阐述。从第 4 章开始，概要地分析了数据仓库在电信领域中的应用情况。在第 5 章中列举了一些具体的应用案例，介绍了数据仓库技术在电信领域中的应用。第 6 章和第 7 章概要地介绍了一些数据仓库产品和进行数据仓库产品测试需要考虑的内容，也包括 TPC 测试的一些内容，最后介绍了数据仓库各项技术的发展趋势。

本书适合在电信领域从事数据仓库的读者学习参考，也可作为数据仓库爱好者的参考用书。

图书在版编目(CIP)数据

数据仓库及其在电信领域中的应用/段云峰等编著. —北京：电子工业出版社，2003.10
（数据仓库与数据挖掘技术应用丛书）
ISBN 7-5053-8881-9

Ⅰ. 数… Ⅱ. 段… Ⅲ.①数据库—基本知识 ②数据库—应用—电信—邮电企业 Ⅳ. ①TP311.13 ②F623
中国版本图书馆 CIP 数据核字（2003）第 056104 号

责任编辑： 朱沭红 齐 莉
印 刷： 北京增富印刷有限公司
出版发行： 电子工业出版社 http://www.phei.com.cn
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：21.5 字数：320 千字
版 次： 2003 年 10 月第 1 版 2003 年 10 月第 1 次印刷
印 数：5 000 册 定价：39.00 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和工具
- 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

Broadview®

点击博文视点

博文视点资讯有限公司（BROADVIEW Information Co.,Ltd.）于2003年6月18日正式成立，是信息产业部直属的中央一级科技与教育出版社——电子工业出版社（PHEI）与国内最大的IT技术网站CSDN.NET和最具专业水准的IT杂志社《程序员》合资成立的以IT图书出版为主业、开展相关信息和知识增值服务的资讯公司。

我们的理念是：创新专业出版体制；培养职业出版队伍；打造精品出版品牌；完善全面出版服务。

秉承博文视点的理念，博文视点的产品线为面向IT专业人员的出版物和相关服务。博文视点将重点做好以下工作：

（1）在技术领域开发专业作（译）者群体和高质量的原创图书

（2）在图书领域建立专业的选题策划和审读机制

（3）在市场领域开创有效的宣传手段和营销渠道

博文视点有效地综合了电子工业出版社、《程序员》杂志社和CSDN.NET的资源和人才，建立全新专业的立体出版机制，确立独特的出版特色和优势，将打造IT出版领域的著名品牌，并力争成为中国最具影响力的专业IT出版和服务提供商。

作为合资公司，博文视点的团队融合了各方面的精英力量：原电子工业出版社IT图书专业出版实力的代表部门——计算机图书事业部的团队；《程序员》杂志和CSDN网站的主创人员；著名IT专业图书策划人周筠女士及其创作群。这是一个整合专业技术人员和专业出版人员的团队；这是一个充满创新意识和创作激情的团队；这是一个不断进取，追求卓越的团队。

原PHEI计算机图书事业部以及武汉的创作团队，在合资之前就已经各自具备了很强的策划和出版实力，并创造出了优秀的出版业绩。而《程序员》杂志社则是国内最有影响力的IT专业杂志，《程序员》杂志的年发行量超过80万份。CSDN网站是国内最大的IT技术网站，到2003年5月底，CSDN网站的注册会员已超过了54万，这两者已成为中国IT技术交流与推广的最佳平台，积累了丰富的作者和读者资源。

电子工业出版社与《程序员》杂志和CSDN网站的合作以最有效率的方式形成了出版资源、媒体资源、网络资源的整合和互动，成为2003年IT出版界倍受瞩目的事件。

“技术凝聚实力，专业创新出版”，BROADVIEW与您携手共迎信息时代的机遇与挑战！

序

数据仓库技术起源于对大量数据进行处理的需要，是随着业务应用的需要而产生的。与传统的数据库技术相比，数据仓库为决策分析提供了更好的支持，跳出了传统联机事务处理的范畴。因此近几年来，数据仓库技术发展很快，并在各个行业都得到了很多的应用。

随着垄断格局的打破，国内电信运营商间的竞争也越来越激烈，而网络服务质量等方面的差别也在逐渐减少，单纯的价格战将对竞争的双方造成损失。因此，电信企业都在寻求改善服务质量、提高市场竞争力的方法。面对这种越来越激烈的市场竞争，电信企业迫切地需要提高企业内部的科学决策能力，增强在市场经营等方面的正确判断能力，因此，电信运营商需要数据仓库技术。

另一方面，电信运营商积累了大量的业务运营数据，这些数据都是已经电子化的数据，通过数据仓库技术，可以从这些用户数据中发现很多有价值的信息，例如用户的消费行为分析特征等。根据这些消费行为特征，市场部门就可以提供针对性更强的市场服务策略，并且节约了市场营销的成本。因此，电信企业的大量电子化数据为其建设数据仓库奠定了技术基础。

建设数据仓库系统能够极大地提高国内电信企业的业务支撑能力，丰富企业的业务应用内容，提高企业的市场竞争力，缩短与国际电信企业在运营管理能力方面的差距，为迎接进入 WTO 后更开放的、竞争更激烈的电信市场做好技术准备。

中国移动作为国际上最大的移动通信运营商，从 2001 年开始，就着手进行在数据仓库基础上的经营分析系统的建设，今年将在全国范围内开展数据仓库建设和应用工作。这种大规模的建设及应用模式，在国际上也是少见的。同时，也将为国内电信企业如何利用数据仓库技术积累重要的经验。

本书首先阐述了数据仓库、数据分析技术的一些基本概念，然后就其在电信领域中的应用进行了介绍，提出有关的 OLAP 分析和数据挖掘在电信领

域中的具体应用例子。同时，介绍了目前比较流行的几种数据仓库、OLAP和数据挖掘产品，并从使用者的角度，提出了进行数据仓库相关产品测试时要考虑的内容，并介绍了 TPC 测试。最后，本书介绍了数据仓库技术的发展历史，并阐述了将来的发展方向。

这是一本从实际出发，介绍数据仓库在电信领域中应用的书。现正逢国内电信行业要开展数据仓库建设，因此希望本书能够为有关的技术人员、业务人员提供参考。

中国移动通信集团公司副总经理

目　录

第 1 章　数据仓库概述 …… (1)

1.1　数据仓库概念 …… (1)

1.1.1　概念 …… (1)

1.1.2　与数据库的对比 …… (5)

1.2　数据仓库技术的发展 …… (7)

1.2.1　数据仓库发展历史 …… (7)

1.2.2　数据仓库的市场发展 …… (9)

1.2.3　数据集市与数据仓库的关系 …… (11)

1.2.4　数据查询技术 …… (13)

1.3　数据分析技术 …… (14)

1.3.1　OLAP 分析 …… (15)

1.3.2　数据挖掘 …… (19)

1.3.3　联机分析挖掘（OLAM） …… (24)

1.3.4　统计分析技术 …… (27)

1.3.5　数据分析技术的难点 …… (28)

1.4　数据仓库的特点 …… (30)

1.4.1　概述 …… (30)

1.4.2　面向主题等特点介绍 …… (30)

1.5　数据仓库与数据分析的关系 …… (31)

1.6　数据仓库现状 …… (32)

1.7　数据仓库应用中的几个问题 …… (34)

参考资料 …… (35)

第 2 章　构建数据仓库 …… (37)

2.1　构建数据仓库 …… (37)

2.1.1　数据仓库实施步骤 …… (37)

2.2 数据仓库项目的需求分析 …… (39)
2.3 数据仓库体系结构与实施框架 …… (40)
2.3.1 体系结构与实施框架概述 …… (40)
2.3.2 统一的数据仓库 …… (41)
2.3.3 数据集市 …… (42)
2.4 数据源的分析 …… (44)
2.4.1 来自业务系统的实时数据 …… (45)
2.4.2 汇总数据 …… (45)
2.5 数据仓库模型设计 …… (46)
2.5.1 数据仓库的建模技术 …… (46)
2.5.2 实体关系建模 …… (47)
2.5.3 维度建模 …… (50)
2.5.4 实体关系建模与维建模的关系 …… (52)
2.5.5 数据仓库模型设计工具 …… (52)
2.6 数据抽取/转换/加载（ETL）过程 …… (66)
2.6.1 数据抽取（Extraction） …… (66)
2.6.2 数据转换（Transformation） …… (67)
2.6.3 数据加载（Load） …… (68)
2.7 构建数据仓库的几个关键问题 …… (69)
2.7.1 数据仓库粒度问题 …… (69)
2.7.2 ETL 的处理策略 …… (71)
参考资料 …… (72)
第 3 章 数据分析技术（OLAP/数据挖掘介绍） …… (73)
3.1 简介 …… (73)
3.2 联机分析处理（OLAP）技术 …… (74)
3.2.1 维的介绍 …… (75)
3.2.2 多维分析（OLAP）方法和工具 …… (76)
3.2.3 OLAP 数据的处理方式 …… (82)
3.3 数据挖掘技术 …… (83)
3.3.1 数据挖掘技术简介 …… (83)

3.3.2　数据挖掘的步骤 …… （85）
3.3.3　数据挖掘的体系结构 …… （89）
3.3.4　数据挖掘的分析模型 …… （90）
3.3.5　数据挖掘的具体算法和常用技术 …… （94）
3.3.6　应用举例 …… （97）
3.4　OLAP 与数据挖掘的对比 …… （104）
3.5　数据挖掘与其他技术的对比 …… （106）
3.5.1　与专家系统的对比 …… （106）
3.5.2　与统计分析的对比 …… （107）
3.5.3　与人工智能的对比 …… （109）
3.6　联机数据挖掘 …… （110）
3.6.1　OLAM 的简介 …… （110）
3.6.2　OLAM 的体系结构 …… （111）
3.6.3　OLAM 的应用 …… （112）
3.7　统计分析方法 …… （114）
3.8　数据分析要与业务结合 …… （117）
参考资料 …… （118）
第 4 章　数据仓库在电信领域中的应用 …… （121）
4.1　概述 …… （121）
4.1.1　数据仓库在电信领域的应用 …… （121）
4.1.2　中国移动的经营分析系统 …… （124）
4.1.3　电信行业数据仓库的应用情况 …… （131）
4.1.4　电信行业选择数据仓库的必然 …… （133）
4.2　电信行业数据仓库及数据分析的特点 …… （134）
4.3　电信行业数据仓库的应用内容 …… （136）
4.3.1　OLAP 在电信行业中的应用 …… （137）
4.3.2　数据挖掘在电信行业中的应用 …… （140）
4.3.3　电信应用中数据挖掘与 OLAP 的对比 …… （145）
4.4　项目实施组织 …… （145）
4.4.1　数据仓库项目的阶段划分 …… （146）

4.4.2 项目管理 …… （150）
4.4.3 应该注意的一些问题 …… （154）
4.5 CRM 与数据仓库 …… （155）
参考资料 …… （160）
第 5 章 数据仓库实施举例 …… （161）
5.1 数据仓库模型实施案例 …… （161）
5.1.1 大客户资料分析主题的数据仓库构建 …… （162）
5.1.2 客户流失分析主题的数据仓库构建 …… （165）
5.1.3 网络状况分析主题的数据仓库构建 …… （166）
5.2 ETL 模块的结构 …… （168）
5.2.1 数据采集清洗子系统 …… （170）
5.2.2 源数据变换、重整、汇总子系统 …… （171）
5.3 OLAP 分析实例 …… （172）
5.3.1 确定分析主题 …… （173）
5.3.2 确定分析方法 …… （173）
5.3.3 定义维度 …… （174）
5.3.4 根据具体的分析主题构造分析立方体（MOLAP） …… （176）
或星型结构（ROLAP） …… （176）
5.3.5 解释分析结果 …… （178）
5.4 数据挖掘方法论 …… （181）
5.4.1 SAS 的 SEMMA 方法论 …… （181）
5.4.2 SPSS 的 CRISP-DM 方法论 …… （183）
5.4.3 IBM 的通用数据挖掘方法论 …… （188）
5.5 电信业务数据挖掘实例 …… （194）
5.5.1 应用决策树算法（Decision Tree）进行客户流失分析 …… （194）
5.5.2 应用聚类分析对电信客户进行细分 …… （201）
5.5.3 应用数据挖掘预防电信欺诈 …… （209）
参考资料 …… （215）
第 6 章 数据仓库产品及解决方案举例 …… （217）
6.1 IBM 公司的相关系列产品简介 …… （218）

6.1.1 数据仓库产品 …… （218）
6.1.2 OLAP 分析产品 …… （220）
6.1.3 数据挖掘产品 …… （224）
6.2 Oracle 相关系列产品简介 …… （227）
6.2.1 数据仓库产品 …… （227）
6.2.2 OLAP 产品 …… （229）
6.2.3 数据挖掘产品 …… （232）
6.3 Sybase 相关系列产品简介 …… （234）
6.3.1 设计组件 Warehouse Architect …… （235）
6.3.2 元数据管理软件 Warehouse Control Center …… （236）
6.3.3 数据仓库引擎 Adaptive ServerIQ …… （237）
6.3.4 Warehouse Studio 的其他特性 …… （239）
6.3.5 Warehouse Studio 的商业应用 …… （240）
6.3.6 Sybase 电信业数据仓库应用案例介绍 …… （241）
6.4 SAS 的数据仓库产品简介 …… （245）
6.4.1 定义数据仓库及其主题 …… （246）
6.4.2 传送和汇总整理数据 …… （246）
6.4.3 更新汇总数据 …… （247）
6.4.4 建立、管理和取用查看 Metadata …… （247）
6.4.5 设置数据市场 …… （248）
6.4.6 SAS 电信业数据仓库应用案例介绍 …… （248）
6.5 Informix 的数据仓库产品简介 …… （250）
6.5.1 数据抽取工具 Datastage …… （250）
6.5.2 数据仓库引擎 Redbrick …… （251）
6.5.3 分析工具 Metacube …… （251）
6.5.4 Informix 电信业数据仓库成功案例 …… （251）
6.6 NCR Teradata 的整体解决方案简介 …… （253）
6.6.1 NCR 可扩展数据仓库 …… （254）
6.6.2 NCR Teradata 数据仓库电信业解决方案介绍 …… （254）
6.6.3 NCR Teradata 电信业数据仓库成功案例 …… （259）

6.7 CA 数据仓库产品介绍 …… (261)
6.8 数据仓库相关专业工具的提供厂商介绍 …… (263)
6.8.1 MicroStrategy …… (263)
6.8.2 Brio …… (264)
6.8.3 Business Object …… (275)
6.8.4 Cognos …… (276)
6.8.5 Informatica …… (279)
6.8.6 Hyperion …… (282)
6.8.7 SPSS …… (285)
参考资料 …… (289)
第 7 章 数据仓库的测试与选型 …… (291)
7.1 数据仓库测试概述 …… (291)
7.2 数据库、数据仓库测试标准 …… (291)
7.2.1 权威的数据仓库测试机构——TPC 测试介绍 …… (291)
7.2.2 TPC-D …… (295)
7.2.3 TPC-H …… (296)
7.2.4 其他数据仓库测试标准 …… (300)
7.3 数据仓库选型所要考虑的因素 …… (301)
7.4 数据仓库测试过程以及测试方案 …… (302)
7.4.1 测试过程 …… (302)
7.4.2 测试指标以及测试流程 …… (302)
7.5 自行组织测试要注意的问题 …… (307)
参考资料 …… (308)
第 8 章 数据仓库的发展与展望 …… (309)
8.1 数据仓库未来发展方向 …… (309)
8.2 OLAP 发展的历史 …… (310)
8.3 OLAP 发展方向 …… (312)
8.4 数据挖掘研究历史及现状 …… (314)
8.4.1 数据挖掘（知识发现）的历史 …… (314)
8.4.2 从专家系统的演进 …… (315)

8.5　数据挖掘发展方向 …… （316）
参考资料 …… （318）

第 1 章　数据仓库概述

1.1　数据仓库概念

1.1.1　概念

1．概述

传统的数据库技术是单一的数据资源[4]，即以数据库为中心，进行从事务处理、批处理到决策分析等各种类型的数据处理工作。近年来，计算机技术正在向着两个不同的方向拓展：一是广度计算，二是深度计算。广度计算的含义是把计算机的应用范围尽量扩大，同时实现广泛的数据交流，互联网就是广度计算的特征。另一方面，人们对以往计算机的简单数据操作提出了更高的要求，希望计算机能够更多地参与数据分析与决策制定等领域。特别是数据库处理可以大致地划分为两大类：操作型处理和分析型处理（或信息型处理）。这种分离划清了数据处理的分析型环境与操作型环境之间的界限，从而由原来的以单一数据库为中心的数据环境发展为一种新环境：体系化环境。这种分离的结果，导致了数据仓库技术的出现和迅速发展。

20 世纪 80 年代中期，“数据仓库之父”William H.Inmon 先生在其《建立数据仓库》一书中定义了数据仓库的概念，随后又给出了更为精确的定义：数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合。与其他数据库应用不同的是，数据仓库更像一种过程，即对分布在企业内部各处的业务数据的整合、加工和分析的过程，而不是一种可以购买的产品。

数据仓库结构如图 1-1 所示。

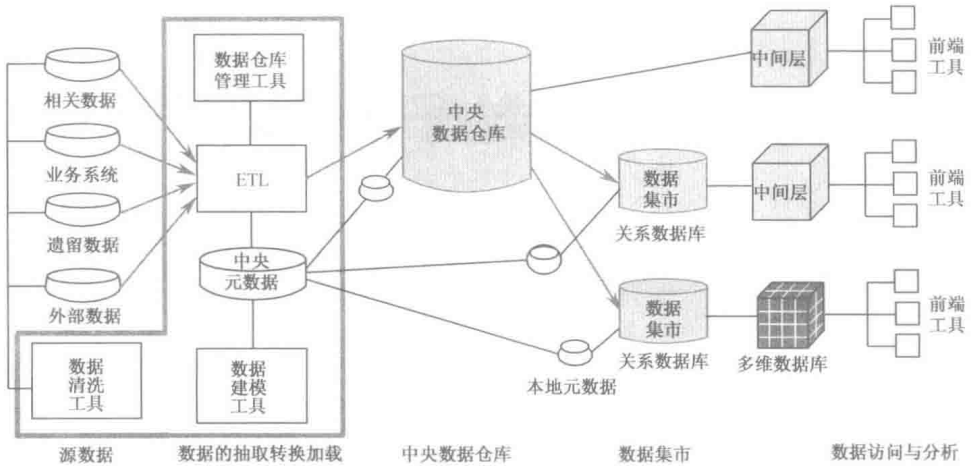

图 1-1　数据仓库结构图

从图 1-1 中可以看出，数据仓库系统包括如下几个部分：

- 源数据部分，数据抽取、转换和装载（ETL）部分，以及中心数据仓库部分。经过这些环节，可以完成将数据从源数据装载到数据仓库中的过程。
- 数据集市，根据部门的需要，可以从数据仓库中形成数据集市，以满足部门级数据分析的需要。
- 数据访问和分析部分：在数据访问和分析过程中，可以采用 OLAP 分析及数据挖掘技术进行分析，得出有关的分析结果。

数据仓库是从数据库系统发展而来的，传统的数据库系统[3]，体现了事务处理过程的优势，人们选择关系数据库是为了方便地获得信息。通过 C.J.Date 博士的经典之作《An Introduction to Database Systems》可以发现，今天数据仓库所要提供的正是当年关系数据库所要倡导的。然而，由于关系数据库系统在联机事务处理应用中获得的巨大成功，使得人们已不知不觉将它划归为事务处理的范畴。过多地关注于事务处理能力的提高使得关系数据库在面对联机分析应用时又遇到了新的问题，即今天的数据仓库对关系数据库的联机分析能力提出了更高的要求，采用普通关系型数据库作为数据仓库在功能和性能上都是不够的，它们必须有专门的改进。因此，数据仓库与数据库的区别不仅仅表现在应用的方法和目的方面，同时也涉及到产品和配置上的不同。