

SQL Server 2008 R2

谢邦昌 郑宇庭 苏志雄 著

数据挖掘与商业智能 「基础及高级案例实战」

微软资深产品经理 李玉秀
数据库专家《涂抹Oracle》 李丙洋

诚挚推荐

数据挖掘领域“专家中的专家”谢邦昌倾心之作

- ◎ 智慧融化枯燥的概念
- ◎ 水煮讲解9大演算算法
- ◎ 实际案例淋漓展示如何学以致用



中国水利水电出版社
www.waterpub.com.cn

SQL Server 2008 R2 数据挖掘与 商业智能基础及高级案例实战

谢邦昌 郑宇庭 苏志雄 著



中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

本书全面介绍了数据挖掘与商业智能的基本概念与原理,包括经典理论与趋势发展,并深入叙述了各种数据挖掘的技术与典型应用。通过本书的学习,读者可以对数据挖掘与商业智能的整体结构、概念、原理、技术和发展有深入的了解和认识。

本书以 Microsoft SQL Server 2008 的数据挖掘模块进行介绍,让读者可以很快地通过书中的说明与范例,在最短的时间内就能上手。

本书分为四个部分:第一部分介绍数据仓库、数据挖掘与商业智能之间的关系。第二部分对 Microsoft SQL Server 的整体架构加以介绍,并详细阐述了直接与数据挖掘相关的两个服务:分析服务和报表服务。第三部分逐一阐述了 Microsoft SQL Server 中包含的九种数据挖掘模型。第四部分提供了四个数据挖掘的范例,通过模仿练习,读者可获得实际的数据挖掘经验,稍加修改就能在自己所处的领域中加以应用。

本书配有案例的相关素材文件,读者可以从万水书苑以及中国水利水电出版社网站下载,网址为: <http://www.wsbookshow.com> 和 <http://www.waterpub.com.cn/softdown/>。

本书为经台湾基峰资讯股份有限公司独家授权发行的中文简体版。本书中文简体字版在中国大陆之专有出版权属中国水利水电出版社所有。在没有得到本书原版出版者和本书出版者书面许可时,任何单位和个人不得擅自摘抄、复制本书的一部分或全部以任何方式包括(资料和出版物)进行传播。本书原版版权属基峰资讯股份有限公司。版权所有,侵权必究。

北京市版权局著作权合同登记号:图字 01-2011-4736 号

图书在版编目(CIP)数据

SQL Server 2008 R2数据挖掘与商业智能基础及高级
案例实战 / 谢邦昌, 郑宇庭, 苏志雄著. — 北京: 中
国水利水电出版社, 2011.9

ISBN 978-7-5084-8764-9

I. ①S… II. ①谢… ②郑… ③苏… III. ①关系数
据库—数据库管理系统, SQL Server 2008 IV.
①TP311.138

中国版本图书馆CIP数据核字(2011)第132718号

策划编辑:周春元 责任编辑:杨元泓 加工编辑:韩莹琳 封面设计:李 佳

书 名	SQL Server 2008 R2 数据挖掘与商业智能基础及高级案例实战
作 者	谢邦昌 郑宇庭 苏志雄 著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路1号D座 100038) 网址: www.waterpub.com.cn E-mail: mchannel@263.net (万水) sales@waterpub.com.cn 电话: (010) 68367658 (发行部)、82562819 (万水)
经 售	全国各地新华书店和相关出版物销售网点
排 版	北京万水电子信息有限公司
印 刷	三河市鑫金马印装有限公司
规 格	184mm×260mm 16开本 20印张 480千字
版 次	2011年10月第1版 2011年10月第1次印刷
印 数	0001—2000册
定 价	48.00元

凡购买我社图书,如有缺页、倒页、脱页的,本社发行部负责调换

版权所有·侵权必究

推荐序

过去 20 年来，企业已累积大量的商业数据，并运用数据仓库分析过去的信息，然而，对过去的了解，并不表示就拥有丰富的商业知识，数据挖掘提供预测的功能，可协助企业洞悉商机，也是如今企业提升竞争力的重要课题。

要发挥数据挖掘最大功效，有 3 个因素不可或缺：了解算法并加以运用、具备 Domain Know-How 分析对的问题、熟悉工具的使用并与现行系统整合。此书不但针对这些关键因素有深入浅出的介绍，并搭配案例，让读者融会贯通。微软也期望通过 Microsoft SQL Server 这个容易搭建且灵活性较高的平台，让过去从未想过使用数据挖掘的企业，通过完善的架构，与 IT 信息系统深度的整合，搭配完整的分析工具及 Office 易于使用的前台界面，得以建立全新的商业智能系统以提升利润且有效降低成本，使企业可针对各种常见的商业问题自行采用数据挖掘的解决方案。

谢邦昌教授是数据挖掘领域内专家中的专家，听过他演讲的客户，对于演讲内容、实际案例分析、技术应用等，都称赞其高质量并符合企业的实际情况。而谢教授对新技术的引进及学习的积极态度，更让后生晚辈的我们敬佩及推崇。本书由数据仓库、数据挖掘与商业智能概念开始谈起，进而深入探讨如何利用 Microsoft SQL Server 创建数据仓库、数据挖掘与商业智能的平台。更重要的是，通过谢教授对九大算法的精辟介绍及实际案例说明，搭配 Microsoft SQL Server 平民化的价格与易用整合的平台及 Office，充分发挥读者的专业领域，期望能有效实现数据挖掘的三大成功关键因素，并引领读者快速进入数据挖掘神秘的殿堂。诚挚向您推荐这本不可错过的好书。

李玉秀

微软 资深产品营销经理

前 言

商业智能中有一项重要的技术，即数据挖掘的分析技术，主要是从大量数据库中寻找有意义或有价值的信息的过程。通过计算机技术或是统计分析方法论，根据整合的数据加以分析研究，发掘出隐含在数据中的特性，通过专业领域知识（Domain Know-how）整合及解释，从中找出合理且有用的信息，经过相关部门针对该模型的评估后，再提供给相关决策单位加以运用。

近年来，商业智能的运用已经逐渐受到企业的重视，将企业累积数据库通过相关信息分析，找出客户细分、消费行为、业务成本与效率等对企业极为重要的信息。通过商业智能的应用，使之更深入地了解客户，并可协助业务的开发以及增加在顾客管理上的有效性。

知识经济时代来临，企业间的竞争模式，从传统的采用压低成本与价格的杀价流血竞争，到近来倡导以创新为核心竞争力的，不论哪一种策略模式，都是不断在技术研发、生产制造、区域营销、客户服务或资源配置等运营的相关问题上，寻求问题的发生原因，并尝试找出解决方案。在整个不同运营阶段中，陆续累积的庞大数据，往往就是答案的隐身之所。因此，如何善用数据，从运营历史的记录里，挖掘出深藏其中的宝贵经验，就是“数据挖掘”（Data Mining）的目的。

相对于其他数据库系统或数据挖掘软件，微软最新推出的数据库系统 Microsoft SQL Server 借助其数据挖掘功能，具备了企业商业智能方案的性能和平实的价格，因而能在业界迅速普及。然而如何充分发挥 Microsoft SQL Server 在商业智能应用中的效果，则需要一定的专业知识和学习过程。针对业界实际需求，编写了这本教程，以期在实务应用和理论方法之间搭建一座桥梁。让读者迅速掌握现代商业智能应用的主要内容。

谢邦昌 郑宇庭 苏志雄

2011年7月

目 录

推荐序
前言

PART I 数据仓库、数据挖掘与商业智能

Chapter 1 绪论	2	3-7 数据挖掘软件介绍	22
1-1 商业智能	3	Chapter 4 数据挖掘的主要方法	24
1-2 数据挖掘	7	4-1 回归分析	25
Chapter 2 数据仓库	9	4-2 关联规则	27
2-1 数据仓库定义	10	4-3 聚类分析	27
2-2 数据仓库特性	10	4-4 判别分析	29
2-3 数据仓库架构	11	4-5 神经网络	29
2-4 创建数据仓库的目的	12	4-6 决策树	31
2-5 数据仓库的运用	14	4-7 其他分析方法	32
2-6 数据仓库的管理	14	Chapter 5 数据挖掘与相关领域的关系	34
Chapter 3 数据挖掘简介	16	5-1 数据挖掘与统计分析	35
3-1 数据挖掘的定义	17	5-2 数据挖掘与数据仓库	35
3-2 数据挖掘的重要性	17	5-3 数据挖掘与 KDD	36
3-3 数据挖掘的功能	17	5-4 数据挖掘与 OLAP	37
3-4 数据挖掘的步骤	18	5-5 数据挖掘与机器学习	37
3-5 数据挖掘建模的标准 CRISP-DM	19	5-6 数据挖掘与 Web 数据挖掘	38
3-6 数据挖掘的应用	21		

PART II Microsoft SQL Server 概述

Chapter 6 Microsoft SQL Server 中的商业智能	40	数据挖掘功能	46
6-1 Microsoft SQL Server 入门	41	7-1 创建商业智能应用程序	47
6-2 关系数据仓库	41	7-2 Microsoft SQL Server 数据挖掘 功能的优势	48
6-3 SQL Server 2008 R2 概述	42	7-3 Microsoft SQL Server 数据挖掘算法	50
6-4 SQL Server 2008 R2 技术	43	7-4 Microsoft SQL Server 可扩展性	51
6-5 SQL Server 2008 R2 新增功能	45	7-5 Microsoft SQL Server 是数据挖掘与	
Chapter 7 Microsoft SQL Server 中的			

商业智能的结合.....	51	8-12 实时商业智能.....	64
7-6 使用数据挖掘可以解决的问题.....	52	Chapter 9 Microsoft SQL Server 的报表服务	
Chapter 8 Microsoft SQL Server 的分析服务		(Reporting Services)	66
(Analysis Services)	56	9-1 为何使用报表服务.....	67
8-1 创建多维数据集的结构.....	57	9-2 报表服务的功能.....	67
8-2 建立和部署多维数据集.....	58	Chapter 10 Microsoft SQL Server 的整合服务 ..	71
8-3 从模板创建自定义的数据库.....	58	10-1 SSIS 介绍.....	72
8-4 统一维度模型.....	59	10-2 操作示例.....	79
8-5 基于属性的维度.....	59	Chapter 11 Microsoft SQL Server 的	
8-6 维度类型.....	60	DMX 语言	100
8-7 量度组和数据视图.....	61	11-1 DMX 语言介绍.....	101
8-8 计算效率.....	62	11-2 DMX 函数.....	102
8-9 MDX 脚本.....	62	11-3 DMX 语法.....	107
8-10 存储过程.....	63	11-4 DMX 操作实例.....	115
8-11 关键绩效指标 (KPI)	64		

PART III Microsoft SQL Server 中的数据挖掘模型

Chapter 12 决策树模型.....	122	16-2 主要算法.....	177
12-1 基本概念.....	123	16-3 操作示例.....	179
12-2 决策树与判别函数.....	123	Chapter 17 线性回归模型	187
12-3 计算方法.....	124	17-1 基本概念.....	188
12-4 操作范例.....	126	17-2 一元回归模型.....	189
Chapter 13 贝叶斯分类器.....	134	17-3 多元回归模型.....	192
13-1 基本概念.....	135	17-4 操作范例.....	195
13-2 操作范例.....	137	Chapter 18 逻辑回归模型	203
Chapter 14 关联规则	147	18-1 基本概念.....	204
14-1 基本概念.....	148	18-2 logit 变换与 logistic 分布	204
14-2 关联规则的种类.....	149	18-3 逻辑回归模型.....	206
14-3 关联规则的算法: Apriori 算法.....	149	18-4 操作范例.....	207
14-4 操作范例.....	150	Chapter 19 神经网络模型.....	215
Chapter 15 聚类分析	160	19-1 基本概念.....	216
15-1 基本概念.....	161	19-2 神经网络模型的特点	217
15-2 层级聚类法与动态聚类法.....	161	19-3 神经网络模型的优劣比较.....	218
15-3 操作范例.....	166	19-4 操作范例.....	220
Chapter 16 时序聚类	176	Chapter 20 时序模型	228
16-1 基本概念.....	177	20-1 基本概念.....	229

20-2 时序的构成	231	20-5 参数化的时序预测模型	240
20-3 简单时序的预测	237	20-6 操作范例	243
20-4 包含趋势与季节成份的时序预测	238		

PART IV Microsoft SQL Server 数据挖掘应用实例

Chapter 21 决策树模型实例	253	Chapter 24 时序模型实例	292
Chapter 22 逻辑回归模型实例	260	24-1 时序模型实例一	293
22-1 回归模型实例一	261	24-2 时序模型实例二	297
22-2 回归模型实例二	266	Chapter 25 如何评估数据挖掘模型	303
22-3 回归模型实例三	270	25-1 评估图节点介绍 Evaluation	
Chapter 23 神经网络模型实例	275	Chart Node	304
23-1 神经网络模型实例一	276	25-2 在 SQL Server 中如何评估模型	307
23-2 神经网络模型实例二	281	25-3 规则度量：支持度与可信度	311

I

数据仓库、数据挖掘 与商业智能

1

CHAPTER

绪论

1-1 商业智能

1-1-1 什么是商业智能

根据 IBM 的统计,在电子商务时代,数据量正以每年 1.3 倍的速度迅速激增,然而在大量数据中,真正被利用来分析与运用的部分却不到一成,因此如何将这些庞大的数据快速地转换成决策者所需的信息,以作为提升企业运营所需的企业智能,已经成为经营管理的一大挑战(栾斌,2002),而商业智能的应用也随着信息量的增加而逐渐受到企业界的重视。通过商业智能的运用,企业可将原始的客户数据做更深入的分析,进而建立有效的预测模式及客户市场区分,使 CRM(Customer Relationship Management)的运用更具成效,也有助于未来 KM(Knowledge Management)的落实(潘启铭,2002)。商业智能的概念是指利用组织化及系统化的流程来取得、分析、发布对其商业活动有重大影响的信息;利用商业智能的协助来预测客户或竞争者的行动,及市场活动或趋势的变化情形(Hannula & Pirttimaki, 2003)。

所谓商业智能是指企业利用信息科技以企业内部及外部既有的数据库数据为基础,根据所需解决的问题进行数据的汇整,整合成数据仓库后,利用适当的工具进行数据处理及包含在线实时分析(OLAP)及数据挖掘(Data Mining)等技术分析数据,将所发现的潜在特性或是建立的预测模型传递给决策者,以协助其进行决策的制定,并达到企业目标的一个程序。

远擎管理顾问公司(2002)认为,商业智能是一种利用信息科技,将分散于企业内部、外部的结构化数据加以汇整,并依据某些特定需求进行分析与运算,再以最优的方法将结果呈现给决策者、管理者或是知识工作者的一种分析机制。换言之,企业将可通过商业智能的使用,使得企业中的决策者得以获得适当的信息,以协助其作出最正确的决策。

而栾斌(2002)则认为将企业内各种数据转换为有意义的信息,提供企业了解现状或是预测未来,更能让企业快速掌握关键商机,将不同平台的异质性数据,通过智能型的转换分析,产出结构化知识的整合交互式分析工具,以利企业内部决策、判断、分析的依据基础,使企业改善决策制订的方法与过程就是商业智能。

1-1-2 商业智能作用及意义

商业智能之所以重要,探究其原因,不外乎是由于企业同业间的彼此激烈竞争,企业经营者为求生存不得不竭尽所能协助企业持续生存下去,因此企业主们必须随时随地地根据所掌握的信息做出实时的决定,但是事后回过头来审视这些实时的决策,会发现其中包含了有效解决问题以及无法解决问题的决策,除了决策者自身的个性会影响决策外,影响决策有效性最重要的因素就是做决策时所掌握信息的充分性及正确性,而商业

智能的含义就是指通过企业所拥有的数据，透过数据仓库的汇整，结合在线实时分析及数据挖掘分析技术挖掘出潜藏在数据库中的有用信息，并将其提供给决策者或部门主管作为平时营运策略的制定，而当企业面临危机或亟需立即做出重大决策时，更能依据数据仓库所提供的正确数据及时做出正确的决策，协助企业顺利解决问题，化危机为转机，因此更可以看出商业智能的重要性，王茁在《商业智能》一书中更提到“商业智能所争取的就是充分利用企业在日常经营过程中搜集的大量资料，并将它们转化为信息和知识来免除企业的瞎猜行为和无知状态。”

对于一般企业来说，商业智能主要可以应用在：

(1) 了解营运状况：商业智能可以帮助企业了解自身营运状况及其推动力量，协助使用者清楚了解产品未来趋势、运营上出现哪些异常情况和哪些行为正对业务产生影响。

(2) 衡量绩效：商业智能可以用来确立对员工的期望，帮助他们跟踪并管理其绩效。

(3) 改善关系：商业智能亦可透过客户关系管理的整合运用，有效地为客户、员工、供应商、股东和大众提供关于企业及其业务状况的有用信息，从而提高企业的知名度，强化整体信息的一致性。利用商业智能，企业可以在问题变成危机之前，很快地检测出问题所在并提出相关建议方案加以解决。商业智能也有助于加强顾客忠诚度，一个参与其中并充分掌握信息的顾客更加有可能会购买您的产品或提供的服务。

(4) 创造获利机会：掌握各种商务信息的企业可以出售这些信息获取利润。但是，企业需要发现信息的买主并找到合适的传递方式。

近年来，企业发展的节奏越来越快，商业复杂性越来越高。虽然许多因特网的企业都消失了，但是因特网的速度不仅没有减慢反而更加突显出其意义。不论企业规模的大小，都需要面对瞬息万变的市场趋势，并根据既有信息做出决策，然而这些决策所依循的是正确无误的信息，由此可见，企业经营管理中信息的重要性仅次于人才的重要性。

1-1-3 商业智能架构

有许多人会将商业智能误认为企业中技术层的电子化解决方案，然而商业智能却是整合了“管理”、“决策”及“信息科技”等三项要素的有效分析机制（远勤管理顾问公司，2002），因此企业必须从策略层的观点来看商业智能，才能了解其重要性。就应用层来看，因为现今信息科技与因特网的兴起，商业智能的应用范畴日益增加，不论是企业界中众人熟知的客户关系管理、供应链管理、企业资源规划，或者是知识管理，都是商业智能实际的运用。为了使企业中的决策人员实时地取得正确及所需的数据，商业智能的作业层工具可以说是商业智能中最重要核心，这些工具包含了数据仓库（Data Warehouse）、在线实时分析、数据挖掘等。

在实际应用中，若以商业智能在顾客关系管理上的应用为例，企业常通过数据仓库的技术汇整来自于不同数据库的信息，进而利用数据挖掘的技术进行各项分析，并以此针对客户过去购买记录、个人基本数据等，分析客户的产品贡献度、细分市场，以便于营销方案的制定，或是针对不同特性客户进行交叉销售（Cross Selling）与向上销售（Up Selling）以提升顾客的产值。

商业智能在于企业中的实施流程（商业智能流程）如图 1-1 所示。由图中可以了解，

企业引入商业智能应用方案过程前，必须清楚的了解企业本身对于引入商业智能的需求是什么，也就是必须理清企业引入商业智能解决方案的原因、整合的组织层级、各部门支持的程度和企业本身对此的重视程度等，若企业管理层不重视，各部门不提供协助，或是主管的层级不够，即使商业智能解决方案再完整，也无法解决企业的问题，无法达成企业需求，而终将面临失败。

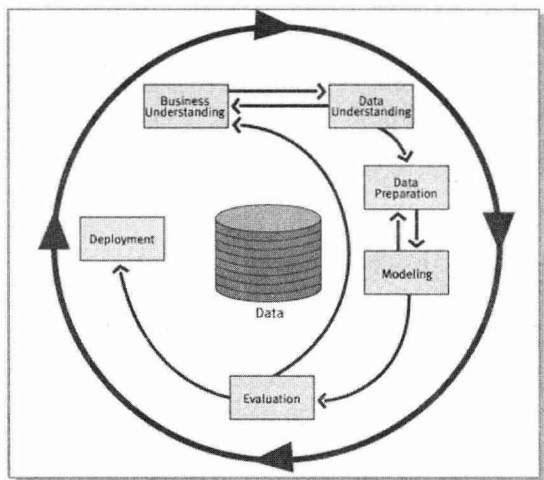


图 1-1 商业智能流程（数据来源：www.crisp-dm.org）

企业本身既然已经了解引入商业智能的目的及其需求，下一步则是要了解企业本身所拥有的数据，商业智能的解决方案不外乎是针对企业既有的数据透过加值分析探索出潜藏的特性，因此对企业本身数据的掌握就更显重要。

以往的企业中，各部门常会根据自身的需求，通过促销活动等方式搜集顾客的信息，但往往是根据部门自身需求而制定的，缺乏整体性考虑，无法将信息与企业整体的客户数据库进行整合，形成许多重要的客户信息都是不完整的，所分析出的信息容易造成偏差，无法真正为企业解决问题。商业智能的核心工作在于根据企业数据库整合成可以作为分析使用的数据仓库，再进一步通过分析技术来探索数据，而对于数据仓库的建立，《Building the Data Warehouse》的作者 William Inmon 认为数据仓库必须具备“面向主题 (Subject-oriented)”、“集成性 (Integrated)”、“时变性 (Time-variant)”及“稳定性 (Non-volatile)”四个特性，事实上数据仓库是有别于传统的数据库系统，且是企业必须特别注意的。

企业在构建商业智能基础的过程中，实时数据查询分析功能扮演着非常重要的角色（远勤管理顾问公司，2002）。简单来说，在线实时分析就是能让用户依据本身决策需求来浏览数据、动态且实时地产生其所需的报表，以提高分析效率的技术。事实上，它除了能提供在线实时数据分析模块外，更重要的是能展示多维度 (Multi-Dimensional) 的数据。

然而商业智能的另一项重要技术是数据挖掘的分析技术，主要是在大量数据库中寻找有意义或有价值的信息的过程。通过机器学习技术或是统计分析方法论，根据整合的数据加以分析探索，发掘出隐含在数据中的特性，通过专业领域知识 (Domain

Know-how) 整合及解释, 从中找出合理且有用的信息, 经过相关部门针对该模型的评估后, 再提供给相关决策单位加以运用。

近年来, 商业智能的运用已经逐渐受到企业的重视, 例如 ING 安泰人寿自 1998 年起, 逐步导入 IBM 的商业智能解决方案, 逐渐累积数据库, 透过相关信息的分析, 找出顾客群体、消费行为、业务成本与效率等对其公司极为重要的信息。通过商业智能的应用, 使 ING 安泰人寿能够更深入了解客户, 并可协助业务的开发以及增加在顾客管理上的有效性。另外, 可口可乐公司亦透过商业智能的导入, 以 mySAP.com 作为基础平台, 统整财务信息, 提升财务规划的能力, 以强化管理市值达 200 亿美元的企业管理能力。上述例子都是企业运用商业智能的成功典范, 因此在产业竞争越来越激烈的环境下, 如何运用商业智能将成为企业强化竞争力的关键之一。

1-1-4 商业智能中的挑战

商业智能活动在美国和欧洲发展的程度较其他地区发达, 商业智能已经变成企业 e 化的主要项目之一。欧美企业希望能够通过商业智能充分利用企业以往对信息技术的投入、改善决策、提高利润、提高营运效率和增强信息透明度。然而针对欧美企业应用商业智能的目的而言, Gartner 在 2002 年进行的商业智能调查中发现, 美国企业与欧洲企业对于商业智能工具的使用略有不同, 美国企业主要是利用商业智能做在线实时分析, 而欧洲企业则是透过商业智能进行高级分析。

纵观欧美企业对商业智能的应用层面, 较可惜的是, 商业智能的运用并未被广泛地提升到策略层面, 致使企业即使使用商业智能, 也不一定能成功地运用商业智能。有些企业的商业智能或数据仓库项目实现了预期的效益, 有些企业这方面的项目因资金不足、人员不足, 或因采取了未能跟策略性的营运目标整合一致的方法而终遭失败。由于缺乏有效、适当的规划, 很多项目变得僵化、孤立, 无法适应不断变化的市场环境, 最后不得不停止。

美国著名商业智能专家 Shaku Atre 于 2003 年所提出的商业智能白皮书中明确指出企业的商业智能项目之所以失败, 主要有下面的 10 个原因:

(1) 未能认识到商业智能项目是跨部门的商务整合计划, 未能理解商业智能不同于那些孤立的解决方案。

(2) 缺乏积极参与的支持者或支持者在企业中没被充分授权。

(3) 缺乏来自业务部门的代表或参与者不够积极主动。

(4) 缺乏有技术、有执行能力的人员或者未充分利用人力。

(5) 缺乏有反馈机制的软件开发方法。

(6) 缺乏分工、缺乏方法论。

(7) 缺乏业务分析或活动标准。

(8) 缺乏对“劣质数据影响一切”的认知和对策。

(9) 缺乏对 Metadata 的必要性认知和使用方式。

(10) 过分依赖分散的方法和工具。

从上述原因中, 不难看出其主要原因是企业没有把商业智能看成是影响企业兴衰和

存亡的大事。如果企业把商业智能和数据仓库看成是策略性问题，而不是一般性或不重要的问题，就会提升实行商业智能项目成功的可能性。

微软公司（Microsoft）的 Microsoft SQL Server 是一个完整的商业智能（Business Intelligence, BI）平台，为用户提供了可用于构建典型和创新的分析应用程序所需的各种特性、工具和功能。其中引入了大量新的数据挖掘功能，允许企业给出这些问题和其他问题的答案。本书将讨论数据挖掘可以解决的各种问题，并介绍 Microsoft SQL Server 处理这些问题的模式。

1-2 数据挖掘

在信息科技发展日进千里的今天，数据处理与存储管理的问题，在软件技术与速度不断的改良，以及硬件设备的购置成本大幅降低之下，都变得简单了，也因此间接带动了企业在与营运相关的数据库的部署与投资。

而所谓“知识经济”时代的来临，企业间的竞争模式，从传统的“红海策略”（采用压低成本与价格的杀价流血竞争），到近来倡导以“创新”为核心竞争力的“蓝海策略”，不论哪一种策略模式，都是不断在技术研发、制造生产、营销、客户服务，或资源配置等营运的相关问题上，寻求问题的发生原因，并尝试找出解决方案。而在整个运营阶段中，陆续累积的庞大数据，往往就是答案的隐身之所。因此，如何善用数据资料，从营运历史的记录里，挖掘出深藏其中的宝贵经验（金矿），就是“数据挖掘”（Data Mining）的目的。

企业在尝试分析其数据时都面临若干问题。一般而言，并不缺乏数据。事实上，很多企业感觉到他们被数据淹没了；他们没有办法完全利用所有的数据，将其变成有用的信息，尤其是当数据从不同的操作系统涌入时，如何得到一致性的信息，是一直困扰企业运营的问题。为了处理这方面的问题，开发了数据仓库技术，让企业将源于各不同操作系统间的数据，加以利用并将其变成有用的信息。

一个适当运作的数据库是具有惊人强大功能的解决方案。公司可以对信息进行分析，并加以利用，以进行明智的决策。通过使用数据库，可以为您提供以下问题的答案：

- ▶ 哪些产品最受 15~20 岁的女性欢迎？
- ▶ 特定消费者的订单前置时间和按时交付的百分比与所有消费者的平均值相比如何？
- ▶ 病房花在每个患者身上的成本和时间是多少？
- ▶ 在签约阶段停滞时间超过十天的项目所占的百分比为多少？
- ▶ 如果某个特定的实验室在某类特定的药品上投入了较多的资金，临床试验结果是否显示病人健康状况好于其他实验室？

除了这些通常可通过使用分析应用程序得出答案的问题之外，数据库还支持各种数据交换格式。分析应用程序设计供分析人员使用，分析人员会对数据进行分类，研究有助于管理与决策的分析结果；报表应用程序会生成书面报表或在线报表，这些报表供

功能要求略低的用户使用，提供静态内容，或提供有限的深入挖掘功能；另对于业务决策者而言，计分卡是非常强大的功能，可以提供公司关键绩效指标（Key Performance Indicator, KPI）的概况，使决策者知道其身处何处。

尽管数据仓库功能强大而实用，但其自身有一个局限：它实质上反映的是过去的历史。由于数据仓库经常在特定周期或时间进行数据加载和处理，因此它只是表示一个时间上的快照（Snapshot）。即使是建构了实时（Real-Time）或近似实时（Near Real-Time）的数据仓库，其数据仍然只表示当前和历史的数据，无法达到“预测”的需要，因此为了发现数据的因果关系，数据仓库需要利用其他科学方法，进行定量的分析。

与传统的统计分析方法不同，“数据挖掘”不是让人提出假设，然后据此去找相关数据，而是让数据仓库确定数据关联性，并允许采用以往不同的模式对数据进行分析。透过数据挖掘，可以得出诸如以下这样的问题的答案：

- ▶ 客户将购买什么产品？哪些产品将一起销售？
- ▶ 公司如何预测哪些消费者可能会流失？
- ▶ 市场状况如何，将会如何发展？
- ▶ 企业如何对其网站使用模式进行最佳的分析？
- ▶ 组织如何确定营销活动是否成功？
- ▶ 什么是分析非结构化数据（如无格式文件）的最好技术？

2

数据仓库