Data Mining & Business Intelligence

SQL Server

数据挖掘与商业智能基础及案例实战

适用于SQL Server 2012/2014

谢邦昌／著

- 基础全面+理论通俗+操作详细
- 涵盖数据挖掘的主要方法、SQL Server的各相关服务、数据挖掘各种模型及其评估
- 配合众多知识点案例＆8大完整应用案例
- 适用于SQL Server 2012/2014

微软产品营销经理
周慕义专业推荐
著名大数据、数据挖掘、统计学专家**谢邦昌**
教授倾情分享

中国水利水电出版社
www.waterpub.com.cn

SQL Server 数据挖掘与商业智能基础及案例实战

谢邦昌　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。

本书共四部分：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 的整体架构进行介绍，并详细阐述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 Microsoft SQL Server 中包含的九种数据挖掘模型；第四部分提供四个数据挖掘的案例以及数据挖掘模型的评估，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/。

北京市版权局著作权合同登记号：图字 01-2015-4778 号

图书在版编目（CIP）数据

SQL Server数据挖掘与商业智能基础及案例实战 / 谢邦昌著. -- 北京：中国水利水电出版社，2015.8
ISBN 978-7-5170-3541-1

Ⅰ. ①S… Ⅱ. ①谢… Ⅲ. ①关系数据库系统 Ⅳ. ①TP311.138

中国版本图书馆CIP数据核字(2015)第190956号

策划编辑：周春元　　责任编辑：杨元泓　　封面设计：李　佳

书　　名	SQL Server 数据挖掘与商业智能基础及案例实战
作　　者	谢邦昌　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	三河市铭浩彩色印装有限公司
规　　格	184mm×240mm　16 开本　22.75 印张　515 千字
版　　次	2015 年 8 月第 1 版　2015 年 8 月第 1 次印刷
印　　数	0001—3000 册
定　　价	58.00 元

推荐序

Microsoft 商业智能解决方案为整个组织提供突破性的洞察能力，也为端对端商业智能解决方案树立了一套新标准。通过遍及整个组织的数据探索功能，提供新的洞察能力。过去 20 年企业已累积了大量的商业数据，并运用数据仓库来分析过去的信息，然而，过去了解，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，也是现今提升竞争力的重要课题。

要发挥数据挖掘最大功效，有 3 项要素不可或缺：了解算法并加以运用、具备 Domain Know-How（译者注：领域专业知识整合及解释）的能力、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要素有深入浅出的介绍，还搭配了实践案例帮助读者融会贯通。Microsoft SQL Server 2014 在继承旧版本的关键任务功能的基础上，为您的关键任务应用程序提供了突破性的效能、可用性和管理性。SQL Server 2014 针对在线事务处理（OLTP）和数据仓库提供了把核心数据库内置于内存中的（In-Memory）新功能，完善了我们现有的内存中数据仓库和 BI 功能，成为市场上最全方位的内存数据库解决方案。

谢邦昌教授一直是业界推广数据挖掘技术的先行者，不仅拥有长期的教学经验，丰富的实践及项目顾问经验，其在商业智能与海量数据处理方面的专业知识更是有口皆碑，也是我个人崇拜的良师。要想一窥最新数据挖掘技术与算法的神奇与奥妙，本书绝对是您最佳的选择，让我们一起来探索崭新信息平台与数据探索的绝妙境界吧！诚挚推荐您阅读这本不可错过的好书。

周慕义 Jack Chou
微软　产品营销经理

序

Microsoft 商业智能中一项重要的技术为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。透过机器学习技术或是统计分析方法论，根据整合的资料加以分析探索，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，数据量的增加速度越来越快，加上商业智能的运用早已受到企业的重视。将企业累积的数据库，透过大量的信息与相关信息的分析，更能找出顾客区分、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。

随着知识经济时代来临，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。不论哪一种策略模式，都是不断在技术研发、制造生产、营销销售、客户服务或资源分配等相关问题上，寻求问题的发生原因并尝试找出解决方案。在不同运营阶段，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，从运营的历史记录中，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。

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 2014 可为您的关键任务应用程序提供突破性的性能、可用性和可管理性。SQL Server 2014 还针对在线事务处理（OLTP）和数据仓库提供了把核心数据库内置于内存中（In-Memory）的新功能，完善了现有数据仓库和商业智能的功能。借助这些功能，极大提升了企业在商业智能处理方面的性能与效率。然而如何充分发挥 Microsoft SQL Server 在商业智能应用中的效力，则需要一定的专业知识和学习过程。针对业界实务上的需求，我们编写了这本教程，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

目　录

PART II Microsoft SQL Server 概述

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

PART IV Microsoft SQL Server 数据挖掘应用实例

PART

I

数据仓库、数据挖掘与商业智能

CHAPTER

1

绪论

1-1 商业智能

1-1-1 什么是商业智能

根据 2014 年 IDC 报告，2013 年的全球数据量有 4.4ZB，预计 2020 年时，全球数据量将增至 44ZB。在如此庞大的数据当中，究竟如何才能挖掘出对决策者真正有用的信息，是现在大家所关注的问题。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。通过商业智能的应用，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，客户关系管理）的运用更具成效，也有助于未来 KM（Knowledge Management，知识管理）的落实（潘启铭，2002）。商业智能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，以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利用在线实时分析（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让企业生存下去，因此企业主们必须随时随地根据所

掌握的信息做出实时的决定，但是事后回过头来审视这些当时的决策，会发现其中既包含了有效解决问题的决策也包含了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，透过数据仓库的汇总，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营运策略制定的依据。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，然而这些决策所依据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，商业智能的操作层工具可以说是商业智能中最重要的核心，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，若以商业智能在客户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，必须清楚地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，达成企业的需求，而终将面临失败。

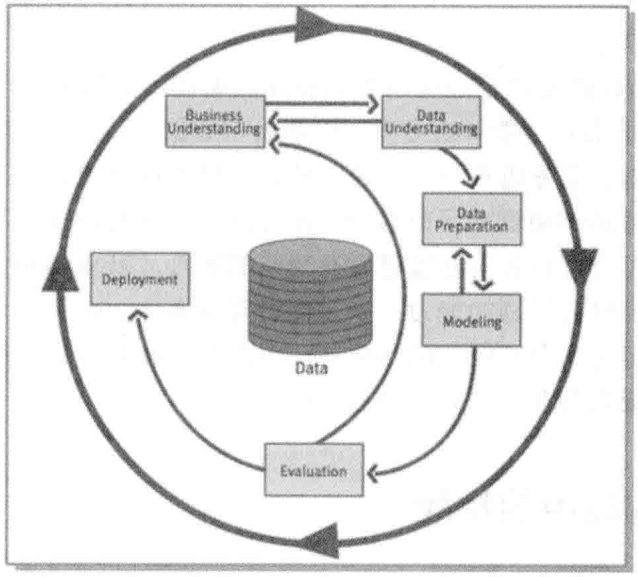

图 1-1　商业智能流程（数据来源：www.crisp-dm.org）

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增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，导致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的，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则因资金不足、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）约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）缺乏对元数据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，很多企业感觉到它们被数据淹没了；它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仓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。通过使用数据仓库，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，数据仓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？