PART

II

Microsoft SQL Server 概述

CHAPTER

6

Microsoft SQL Server 中的商业智能

6-1 Microsoft SQL Server 入门

在安装 Microsoft SQL Server 时，第一点要注意的就是它的集成安装体验。不再需要为某些功能（如 Analysis Services）而分别执行安装程序。如果某个功能（如 Reporting Services）不可安装，则说明您的计算机不满足该功能的安装要求。可以查看帮助文档，以获得有关功能必要条件的完整介绍。在大多数配置得当的机器上，安装过程中应接受所有默认值，安装的主要功能如下：

- SQL Server 数据库引擎
- DTS
- Analysis Services
- Reporting Services
- SQL Server Management Studio（数据库管理工具集）
- Business Intelligence Development Studio（BI 应用程序开发工具集）

Reporting Services 要求在机器上安装并配置 IIS。由于 Reporting Services 是 Microsoft SQL Server Business Intelligence 功能组的一个重要组成部分，在此强烈建议执行这些配置和安装步骤。

建议用户使用 Business Intelligence Development Studio 进行开发，同时使用 SQL Server Management Studio 来操作和维护 BI 数据库对象。虽然可以在 SQL Server Management Studio 中设置 DTS 包以及 Analysis Services 多维数据集和数据挖掘模型，但 Business Intelligence Development Studio 却为设计和调试 BI 应用程序提供了更好的体验。

对于有经验的 IT 人员而言，建议从掌握新的应用程序入手，因为与升级现有 DTS 包或 Analysis Services 数据库相比，这样可以学到更多东西。如果已有一个可用的包或数据库，就会发现“重新创建”现有的包或数据会十分有用。当熟悉了这些新增工具、功能和概念之后，便可试着升级现有对象。

许多客户都借助 SQL Server 工具，使用熟悉的来自一个或多个源系统的商业智能结构来开发新的系统，使用 DTS 填充维度关联型数据仓库，然后再用数据仓库来填充 Analysis Services 数据库。但是，Microsoft SQL Server 提供了许多选项，通过消除或淡化不同的组件使其背离了这种一般化设计。

6-2 关系数据仓库

Microsoft SQL Server 关系数据库引擎包含一些对数据仓库样式应用程序设计和维护大有帮助的功能。这些功能包括：

- 对于超大型的报表而言，表分区可提高数据的加载速度，并简化维护过程。
- 轻松创建报告服务器。
- Transact-SQL 方面的改进，包括新增的数据类型和新增的分析功能。

- 联机索引操作。
- 细化备份/还原操作。
- 快速初始化文件。

6-3 SQL Server 2008 R2 概述

SQL Server 能随时随地管理数据，实现了 Microsoft 建立信息平台的愿景。它使用户可以直接在数据库中存储结构化、半结构化与非结构化的文件，如图片与音乐。SQL Server 提供了一组多样化的整合式服务，能让用户对数据进行查询、搜索、同步化、报告与分析等多种操作。用户的数据可以存储在数据中心的大型服务器上，也能存储在桌面型计算机与移动设备上，无论存储位置为何，都能让用户充分控制数据。

SQL Server 能让用户利用 Microsoft.NET 与 Visual Studio 开发定制化应用程序，以及通过 Microsoft BizTalk Server 内的面向服务架构（SOA）与业务程序来访问数据，而且信息工作者也能使用软件直接访问他们每日使用的数据，如 Microsoft Office 2007 System。SQL Server 提供高信赖度、高生产力与商业智能的信息平台，能符合用户所有的数据需求。

而对于 SQL Server 2008 R2 的发展，微软将继续原有建立信息平台的目标，通过丰富的应用，使企业能够提供或获取实时信息。SQL Server 不仅提供了一个完整的方法来管理信息平台，现在 SQL Server 2008 R2 的功能将可通过云延伸到 SQL Server 广大的平台上。在平台上，SQL Server 2008 R2 提供了一致的挖掘模型和常用工具，通过大规模的分布式数据服务，将可以提供新的商机及高可用性。

人们在信息平台上，将可使用不断扩大的信息技术和数据库，而这部分将会由专业人士不停地开发及提供服务。SQL Server 2008 R2 提供了一些独特的服务给每位使用者，而在使用了 SQL Server 2008 R2 时，SQL Azure（微软提供的云服务平台）以及微软的全球合作伙伴将会提供下列服务：

一、可扩展平台

业务应用程序（LOB）是用于 IT 部门之间的联系和业务。LOB 应用程序可以提供安全、可靠的存储，并可集中、管理和分发数据到用户。在 SQL Server 2008 R2 中，可为企业提供一个高性能的数据库平台，此平台可靠、可扩展且易于管理。SQL Server 2008 R2 帮助 IT 部门提供更符合成本效益且具伸缩性的平台。

二、信息科技及开发效率

当信息人员面对不曾发生的需求，在现有的预算与资源下，将要如何提供最大限度的服务？为了达到这个目的，应提供信息人员所需工具以及完成作业的相关能力，以帮助他们提高效率以及简化管理，并快速发展。在 SQL Server 2008 R2 的基础下，使 DBA 和开发人员获得新的工具和能力，在最短时间内开发出可以帮助 IT 管理员的资源。SQL Server 2008 R2 提供新的工具

用于管理大型多数据库环境及改进数据库的能力，以帮助巩固价值最大化，确保简化开发和部署数据驱动的应用程序。

6-4 SQL Server 2008 R2 技术

SQL Server 2008 R2 的网站将经常性提供相关的更新以及相关的新技术，可提供用户做相关的更新。

一、分析技术

通过熟悉的工具，SQL Server 2008 R2 帮助组织建立全面的企业级分析解决方案，提供可操作的结果。

二、多数据库环境在 R2 中的应用和管理

针对多数据库的管理，SQL Server 2008 R2 将帮助企业主动有效地管理数据库环境。通过整合资源的利用，精简措施巩固和提升整个应用生命周期，使人们快速且容易使用。

三、压缩

使用 SQL Server 2008 R2 内建的数据压缩和备份压缩功能，以降低数据存储成本，并确保关键任务应用程序的最佳性能。

四、数据挖掘

SQL Server 2008 R2 授权作出明确的决定与预测分析，通过全面的数据挖掘，整合整个 Microsoft BI 平台，可扩展到任何应用程序。

五、高可用性

SQL Server 2008 R2 在技术上提供了一个全方位的方案，以最大限度地减少停机时间并保持适当水平的应用程序可用性。

六、集成服务

SQL Server 2008 R2 提供一个可扩展的企业数据整合平台，具有卓越的 ETL 和整合能力，使企业更容易地管理各式各样数据库中的数据。

七、可管理性

SQL Server 2008 R2提供了针对一个或多个SQL Server分析结果的策略性系统管理，以及工具的性能监控，故障排除和调整，使管理员能够更有效地管理它们的数据库和SQL Server 分析结果。

八、R2 中商业智能的自我管理服务

SQL Server 2008 R2 的功能，丰富的商业智能组件，可扩大到整个商业智能与企业，直观的工具和帮助最大限度地提高投资回报率和提高 IT 效率的规模。

九、R2 的主要数据服务

SQL Server 2008 R2 的主数据服务，使企业能够开始使用简单的解决方案或业务需求分析调整所需的解决方案，以逐步增加。支持多种方案利用相同的数据。

十、高性能和可伸缩性

SQL Server 2008 R2 提供全面的数据平台，其中包括技术。针对大规模的服务器和大量的数据库，可使用内置的工具来改良性能。

十一、可编程

在 SQL Server 2008 R2 中，如何使开发人员能够构建强大的下一代数据库应用程序可使用.NET 框架和 Visual Studio 团队系统。

十二、R2 中的报告服务

针对服务器的平台 SQL Server 2008 R2 提供了一个完整的报表服务，以支持多种不同的报表需求，提供整个企业需要的有关数据。

十三、安全性

SQL Server 2008 R2 提供的增强安全功能，有助于提供有效的安全管理与功能配置，强大的认证、强大的加密和密钥管理能力，并加强审计。

十四、数据空间

SQL Server 2008 R2 提供大量的数据空间，使数据库能够完整地连接、使用和扩展

数据库中的数据，并应用数据挖掘方法帮助用户做出更好的决策。

十五、R2 中的复杂事件处理能力

SQL Server 2008 R2 采用 Microsoft StreamInsight 技术，在短时间内可处理多个数据库中的大量数据。这项技术还能处理各类事件与信息查询功能。通过 StreamInsight 技术，用户能通过历史数据的信息与如今的动态数据，做出更有效的决策。

6-5 SQL Server 2008 R2 新增功能

表 6-1 所示为 SQL Server 2008 R2 的新增功能。

表 6-1 SQL Server 2008 R2 新增功能

新增功能	功能介绍
SQL Server 公用程序	建立 SQL Server 公用程序控制点(UCP)：安装 SQL Server 2008 R2 Database Engine 的单一实例，然后将它升级为 UCP。UCP 是针对 SQL Server 公用程序中注册的所有实例收集的配置与性能数据的中的存储机制。UCP 是 SQL Server 公用程序的中的推理点。它可支持一些动作，例如应用中的原则，或是分析可能超出中的资源使用量原则时，所要预测的实例的资源使用量趋。
数据层应用程序	数据层应用程序会简化可支持多层或客户端－服务器应用程序的数据层对象的开发、部署与管理。DAC 会定义支持应用程序所需的所有 Database Engine 架构和实例对象，例如数据表、视图表和登录。DAC 在整个相关应用程序的开发、部署与管理周期中，都会当做单一管理单位来运作。DAC 也包含定义 DAC 部署必要条件的原则。DAC 可以部署 SQL Server 2008 R2 和 SQL Azure 的实例
与 SQL Azure 的连接	SQL Server 2008 R2 引进了从客户端公用程序连接到 SQL Azure 数据库的功能：生成和发布脚本向导，可以使用 SQL Azure 当作它所发行的脚本的来源和目标
SQL Server PowerShell 提供者	SQL Server 2008 R2 引进新的 SQLSERVER:\Utili 和 SQLSERVER:\DAC 文件夹，以支持 PowerShell 脚本中的 SQL Server 公用程序和数据层应用程序
网络连接	VIA 通信协议已被取代。未来的 Microsoft SQL Server 版本将移除这项功能。请避免在新的开发工作中使用这项功能，并规划修改目前使用这项功能的应用程序

CHAPTER

7

Microsoft SQL Server 中的数据挖掘功能

Microsoft SQL Server 平台引入了大量的数据挖掘功能，既能采用传统方式处理数据挖掘，也能采取新的方式进行数据挖掘工作。就传统方式而言，数据挖掘可以根据输入来预测未来的结果，或者尝试发现以前未识别但类似的组中的数据或集群数据间的关系。

Microsoft 数据挖掘工具与传统数据挖掘应用程序有很大的不同。首先，它们支持组织中数据的整个开发生命周期（Microsoft 将其称为集成、分析和报告）。此功能使得数据挖掘结果不再仅限于供少数专门的分析人员使用，而向整个组织开放。其次，Microsoft SQL Server 是开发智能应用程序的平台，而并非一个独立应用程序。由于可以方便地从外部访问数据挖掘模型，因而可以构建智能化的自定义应用程序。而且，该模型具有可扩展性，因此第三方可以添加自定义算法以支持特定的挖掘需求。最后，Microsoft 数据挖掘算法还可以实时运行，允许实时根据挖掘的数据集进行数据验证。

Microsoft SQL Server 中的数据挖掘功能属于商业智能技术，它可以帮助用户构建复杂的分析模型，并使其与业务操作相集成。数据挖掘可回答如下问题：

- 该客户的信用风险如何？
- 客户的特征如何？
- 人们愿意同时购买哪些产品？
- 下个月能卖出多少产品？

数据挖掘应用程序将数据挖掘模型集成到日常的业务运营之中。许多数据挖掘项目的目标是构建可供业务用户、合作伙伴和客户使用的分析应用程序，而不必理会应用程序底层的复杂计算。要实现这一目标，需要执行两个主要步骤：构建数据挖掘模型并构建应用程序。Microsoft SQL Server 使这些步骤比以往更加简单。

7-1 创建商业智能应用程序

创建智能应用程序实际上就是利用数据挖掘的各种优势，将其应用到整个数据输入、集成、分析和报告过程中。大部分数据挖掘工具都可以预测未来的结果，帮助确定不同数据元素之间的关系。这些工具中的大部分都针对数据运行，生成随后分别解释的结果。很多数据挖掘工具都是独立的应用程序，专为预测需求或识别关系而存在。

智能应用程序将获取数据挖掘的输出，将其作为输入应用到整个流程。使用数据挖掘模型应用程序的一个例子就是用于接受个人信息的数据输入窗体。应用程序的用户可以输入大量数据，如出生日期、性别、教育程度、收入水平、职业等等。属性的某些组合并不合乎逻辑；例如，七岁小孩的职业是医生且有高中文凭，这就表示有人在随便填入数据或者表明此人不具有处理数据输入窗体的能力。大部分应用程序会通过实现复杂而层层嵌套的逻辑来处理此类问题，但实际上，要确定所有此类数据组合是否有效，几乎是不可能的。

为了解决此问题，企业可以使用数据挖掘来查询现有的数据，据此构建有效数据组合的规则。每个组合都给予一个可靠程度计分。组织然后就可以构建数据输入程序，使用数据挖掘模型进行实时数据输入验证。该模型将根据现有总体数据给输入计分，并返回输入的可靠程度。接着应用程序可以根据预先确定的可靠程度阈值来决定是否接受输入。

此例说明了使用可以实时运行的数据挖掘引擎的好处：可以编写能利用数据挖掘的

强大功能的应用程序。数据挖掘并非最终结果，它成为整个过程的一部分，在集成、分析和报告的每个阶段都起到一定的作用。

数据挖掘可以用在数据集成过程的前端，以验证输入，也可以在分析阶段使用数据挖掘。数据挖掘提供了分组或聚类的功能，例如，可以根据关键词将类似的消费者或文档归入同一个组中。然后可以将这些聚类送回到数据仓库，从而可以使用这些分组执行分析。一旦知道了分组情况并将其送回到分析循环中，分析人员就可以使用它们来采用以前不可能的方式查看数据。

智能应用程序的一个主要目标就是使得每个人都可以使用数据挖掘模型，而不再是分析人员的专利。过去，数据挖掘一直是具有统计学或操作研究背景的专家的领域。为支持此类用户而构建了很多数据挖掘工具，但这些工具并不能方便地与其他应用程序集成。因此，在数据挖掘产品本身之外使用数据挖掘信息的能力非常受限制。不过，通过使用跨越整个过程且将模型和结果对其他应用程序开放的工具，企业可以创建能在任何阶段使用数据挖掘模型的智能应用程序。

平台采用集中的服务器存储数据挖掘模型和结果，这是平台有利于创建智能应用程序的另一方面。这些模型通常具有高度的专用性，且保密性较高。将其存储在服务器上，可以防止其分散到组织外部。所带来的额外的好处就是，通过为模型提供共享位置，公司可以为每个模型保持单一版本，而不会在每个分析人员的桌面上存在多个版本。具有“事实的单一版本”是数据仓库的目标之一，此概念也可以扩展到数据挖掘，因此创建的模型也具有单一版本，并针对特定业务进行了改良。

Microsoft SQL Server 中数据挖掘功能的目标是构建具备以下特征的工具：

- 简单易用
- 可提供一整套的功能
- 可轻松嵌入到产品应用程序中
- 紧密集成其他的 SQL Server BI 技术
- 能够扩展数据挖掘应用程序的市场

可以肯定，本书的每位读者几乎都曾“使用”过数据挖掘应用程序。例如在线购买音乐，并收到了“购买此产品的其他客户”的建议；或者食品店在收据上打印个性化优惠券。所有这些，都是从使用数据挖掘应用程序中得到的好处。时至今日，这种应用程序的开发已集中于解决大型公司所面临的最大问题，这些公司能够承受分析能力的匮乏以及巨额的开发费用，而这些都是过去用传统方法构建数据挖掘应用程序所需面对的。正如 Microsoft 的 OLAP 技术已推动了 OLAP 市场增长一样，我们期望能够将数据挖掘技术推广开来，使那些在过去不能开发这种应用程序的企业和部门也能够加入到其开发行列中来。

使用 Microsoft SQL Server 中的数据挖掘工具开发一套数据模型，然后在这些模型的基础上随意执行预测。这是所有数据挖掘的模式：开发、模型发现和模型预测。

7-2 Microsoft SQL Server 数据挖掘功能的优势

Microsoft SQL Server 数据挖掘功能具有优于传统数据挖掘应用程序的众多优势。正

如前面所讨论的，Microsoft SQL Server 数据挖掘功能与所有 SQL Server 产品实现了集成，包括 SQL Server、SQL Server Integration Services 和 Analysis Services。SQL Server 数据挖掘工具不是公司运行以输出（稍后将独立于分析过程的其他部分对其进行分析）的单个应用程序。数据挖掘功能嵌入到整个过程中，可以实时运行，且结果可以发送回集成、分析或报告过程。不过，如果这些功能难于使用，则没有什么实际意义。幸运的是，Microsoft 特别关注工具的易用性。

7-2-1 易于使用

通过 Microsoft SQL Server，Microsoft 努力将数据挖掘从博士们的实验室中搬出来，让负责设置和运行数据模型的开发人员和数据库管理员（DBA）、任何分析人员、决策者或使用模型输出的其他用户也可以使用数据挖掘，而不需要具有任何专门知识。

例如，一家使用 Microsoft SQL Server 早期版本的公司希望创建一个交叉销售应用程序。交叉销售会根据人们的购买模式和当前购买的产品向其推荐产品。例如，某个消费者购买了特定女影星主演的三部电影，则该顾客可能对同类电影中该女影星主演的电影更感兴趣。另一方面，对科幻小说和恐怖电影感兴趣的消费者可能不会对爱情影片的交叉促销感兴趣。

为了实现交叉销售程序，该公司求助于 DBA，而不是分析人员。DBA 使用 Microsoft SQL Server 新数据挖掘功能设置了一个预测模型，该模型将根据各种因素（包括历史销售资料和消费者的个人信息）进行建议。这个已就绪的模型每秒钟可就此特定的消费者产生一百万个预测。最终结果：实现新模型后，推荐产品的销售额翻了一番。

7-2-2 简单而丰富的 API

Microsoft SQL Server 中的数据挖掘功能具有一个非常强大却甚为简单的 API，这使得创建智能应用程序非常简单。利用该 API，无需了解每个模型的内部细节及其工作原理，就可从客户端应用程序调用预测模型。这使得开发人员可以根据分析的数据调用引擎并选择能提供最佳结果的模型。返回的数据已被标记，即数字值在一系列属性中返回。这使得开发人员可以使用简单资料，而不用面对新的数据格式。

访问数据挖掘结果非常简单，通过使用一种与 SQL 相似的语言即可（称为 Data Mining Extensions to SQL 或 DMX）。其语法设计非常适合已经了解 SQL 的人员使用。例如，DMX 查询可以与如下所示类似。

```
SELECT TOP 25 t.CustomerID
FROM CustomerChurnModel
NATURAL PREDICTION JOIN
OPENQUERY（'CustomerDataSource', 'SELECT * FROM Customers'）
ORDER BY PredictProbability（[Churned],True）DESC
```

7-2-3 可伸缩性

Microsoft SQL Server 中最重要的数据挖掘功能就是其处理大型数据集的能力。在众多数据挖掘工具中，分析人员必须创建有效的随机数据样本，并对该随机样本运行数据挖掘应用程序。尽管生成随机样本听起来非常容易，但统计学家可以提出大量的理由，说明为什么生成有效且真正具有随机性的样本是非常困难且充满风险的。

Microsoft SQL Server 允许模型对整个数据集运行，从而消除了采样方面的挑战。这意味着分析人员不必创建样本集，算法将在所有数据上有效，从而能提供最为准确的结果。

7-2-4 数据挖掘算法

所有数据挖掘工具（包括 Microsoft SQL Server Analysis Services）都采用了多种算法。当然，Analysis Services 是可扩展的；第三方 ISV 可以开发算法，并将所开发的算法无缝地融入到 Analysis Services 数据挖掘框架之中。根据数据和目标的不同，应该采用不同的算法，而且每种算法都可用于解决多个问题。

数据挖掘工具擅长解决多种类型的问题。表 7-1 概括了分析问题的大致分类。

表 7-1 数据挖掘在解决方法上的分类

分析问题	示例	Microsoft 算法
分类：为事例分配预定义的级别（如“好”与“差”）	➤ 信用风险分析 ➤ 客户流失分析 ➤ 客户挽留	➤ 决策树 ➤ 贝叶斯算法 ➤ 神经网络
拆分：开发一种按相似事例分组的分类方法	➤ 客户数据分析 ➤ 邮件推销活动	➤ 聚类 ➤ 顺序群集
关联：相关性高级计算	➤ 购物车分析 ➤ 高级数据研究	➤ 决策树 ➤ 关联规则
时间序列预测：预测未来	➤ 预测销售 ➤ 预测股票价格	➤ 时序
预测：根据相似事例（如现有客户）的值预测新方案（如新客户）的值	➤ 提供保险率 ➤ 预测客户收入 ➤ 预测温度	➤ 全部
偏差分析：发现事例或群体与其他事例和群体之间的差别	➤ 信用卡欺骗检测 ➤ 网络入侵分析	➤ 全部

7-3 Microsoft SQL Server 数据挖掘算法

Microsoft SQL Server 中可以使用很多算法（见表 7-2）。

表 7-2　Microsoft SQL Server 的算法

模型	描述
决策树	决策树算法将基于训练集中的值计算输出的概率。例如，20～30 岁年龄组中每年收入超过 60,000 美元，且有自己的房子的人比没有自己房子的 15～19 岁年龄组的人更可能需要别人提供整理草坪的服务。以年龄、收入和是否有房子等信息为基础，决策树算法可以根据历史数据计算某个人需要整理草坪的服务的概率
关联规则	关联规则算法将帮助识别各种元素之间的关系。例如，在交叉销售解决方案中就使用了该算法，因为它会记录各个项之间的关系，可以用于预测购买某个产品的人也会有兴趣购买何种产品。关联规则算法可以处理异常大的目录，经过了包含超过五十万种商品的目录的测试。
Naïve Bayes	Naïve Bayes 算法用于清楚地显示针对不同数据元素特定变量中的差异。例如，数据库中每个消费者的 Household Income（家庭收入）变量都会不同，可以作为预测未来购买活动的参数使用。此模型在显示特定组间的差异方面尤为出色，如那些流失的消费者和那些未流失的消费者
时序集群	时序集群算法用于根据以前时间的顺序分组或聚类数据。例如，Web 应用程序的用户经常按照各种路径浏览网站。此算法可以根据浏览站点的页面顺序对用户进行分组，以帮助分析消费者并确定是否某个路径比其他路径具有更高的收益。此算法还可以用于进行预测，例如预测用户可能访问的下一个页面。请注意，时序集群算法的预测能力是许多其他数据挖掘供货商所无法提供的功能
时序	时序算法用于分析和预测基于时间的数据。销售额是最常见的使用时序算法进行分析和预测的数据。此算法将发现多个数据序列所反映出来的模式，以便企业确定不同的元素对所分析序列的影响
神经网络	神经网络是人工智能的核心。它们旨在发现数据中其他算法没有发现的关系。神经网络算法一般比其他算法更慢，但它可以发现各种并不直观的关系
文字探勘	文字探勘算法出现在 SQL Server Integration Services 中，用于分析非结构化的文本数据。利用此算法，各个公司可以对非结构化数据进行分析，如消费者满意度调查中的“comments”（注释）节

7-4　Microsoft SQL Server 可扩展性

Microsoft SQL Server 包括了大量可以立即使用的算法，而 Microsoft SQL Server 所使用的模型允许任何供货商向数据挖掘引擎添加新模型。这些模型将与 Microsoft SQL Server 提供的模型处于同等位置。第三方的算法还可以享有其他功能所带来的优势：可以使用 DMX 对其进行调用，且易于整合到分析和报告过程的任何部分中。

7-5　Microsoft SQL Server 是数据挖掘与商业智能的结合

集成阶段包括从不同的源获得数据、传输数据和将其加载到一个或多个源中。传统数据挖掘工具在集成阶段几乎没有任何作用，因为正是在这个阶段取得数据，将其准备好用于挖掘。尽管这个听起来像先有鸡还是先有蛋的问题，Microsoft 对于此阶段的处理方法相当直接：取得数据、合并数据、数据挖掘，然后将数据挖掘的结果应用到目前和所有将来的数据。而且，数据挖掘算法可以帮助各个公司发现已经存在于数据中的多

余数据，或者在传统的提取、转换和加载（ETL）过程中生成的多余数据。

在集成阶段，如果可以接受插补值，则也可以接受模型所提供的缺失值。这些值可能来自前一段时间，也可以预测未来的活动。Microsoft 数据挖掘工具可以从集成阶段动态生成数字，而不是仅在集成完成后才能提供，这一点颇具优势。

数据挖掘工具与 SQL Server Integration Services 实现了整合。这意味着在数据传输和转换阶段，可以根据数据挖掘模型的预测输出分析和修改资料。例如，可以动态地分析文件或数据字段，并根据文件内的关键词放入恰当的数据库中。

7-5-1 数据分析

典型的数据挖掘工具将在构建了数据仓库后产生结果，而这些结果独立于在数据仓库上完成的其他分析。还将产生预测或标识关系，但数据挖掘模型的结果通常独立于数据仓库中使用的数据。

Microsoft 工具与整个过程实现了整合。正如可在 SQL Server Integration Services 中使用数据挖掘一样，在 Analysis Services 和 SQL Server 中也可以看到数据挖掘带来的好处。不管公司选择使用关系数据还是 OLAP 数据，数据挖掘在分析阶段带来的优势都十分明显。通过归一化数据模型（UDM），才能以透明的方式对关系数据和 OLAP 数据进行分析，而数据挖掘则对此分析起到了促进作用。

当分析特定数据元素时，如产品之间的关系如何以及如何根据购买模式和网站浏览模式对消费者进行分组，各种数据挖掘模型可以确定如何将这些客户或产品划分为对分析有意义的组。当把这些组发送回分析过程时，数据挖掘引擎允许分析人员和用户根据这些集群进行划分和细化。

7-5-2 报告

一旦建模完成，创建了正确的模型，数据挖掘的重点就从分析转到了结果上，而且更重要的是通过将结果在正确的时间送到正确的人手中，来将这些结果应用到工作中。Microsoft SQL Server 中实现了数据挖掘和报告的集成，可以通过简单灵活且可伸缩的方式向组织中的任何人提供预测结果。

通过充分利用 Microsoft SQL Server Reporting Services，预测模型的结果通过将报告嵌入 Microsoft SharePoint Services，可以轻松地部署到打印报告、Microsoft Office 文件或内网中。例如，一个部门可以方便地看到产品销售的智能预测，或将最可能购买某个产品的消费者列表分发到呼叫中心。他们甚至可以看到显示消费者购买或不购买产品的十大原因，从而合理地分配销售人力。Microsoft 通过以易于理解的方式向用户报告、提供有意义的数据，可以轻松地使用数据挖掘的智能和强大功能。

7-6 使用数据挖掘可以解决的问题

谈到数据挖掘可以解决的业务问题时，很多人都会想到购物车分析或发现数据间的

关系，这些在以前都已经广为人知了。实际上，很多问题都可以通过数据挖掘得到解决，但要处理这些问题，重要的是要认识到数据挖掘可以适用于集成、分析和报告过程的任何阶段。

7-6-1 构建挖掘模型

模型的构建、培训和测试过程是创建应用程序过程中最为困难的一部分。正如下面要讨论的，实际开发应用程序是一个简单的编程过程。在开始构建数据挖掘模型之前，应当已经收集和整理了数据，这些数据极有可能位于数据仓库中。Microsoft SQL Server 可以从关系数据库或 Analysis Services 多维数据中查看数据。

开发数据挖掘模型的最佳人选是同时具备业务和技术技巧的人员。模型的开发人员将会从其统计背景中获益、了解企业面临的关键业务问题、对数据和关系产生极大的好奇心，同时还能够利用 Microsoft SQL Server 工具处理和存储数据。现有数据仓库小组中的成员最有可能遇到这些标准。

作为数据挖掘的初学者，应在构建原型模型的同时，计划花费数周时间来研究数据、工具以及可供选择的算法。使用一台具备数据库管理权限的开发服务器。构建模型的最初阶段是探索阶段：用户可能会希望以不同的方法来重新构建数据和实验。当然，用户肯定希望从少量数据子集开始，并在开发愈加清晰的模型设计时扩展数据集。在原型阶段，不要为如何构建一个“可供生产使用”的应用程序而担心。使用 DTS 或执行任何所需数据处理最为舒适的任何工具。保存一份记录有必要转换的高级日志，但不要期望所做的一切都能成为永久应用程序的一部分。

用户应当准备两套数据：一套用于开发模型，另一套用于测试模型的精确度，从中选择适合业务问题的最佳模型。在考虑如何划分数据子集时，要确保没有引入任何偏差。例如，从十个客户中选择一个客户，或根据姓氏的第一个字符区分，或根据其他任意属性区分。

开发数据挖掘模型的过程涉及选择以下内容：

- 输入数据集
- 输入字段
- 数据挖掘算法
- 算法在计算过程中所用到的参数

如果不知道哪种类型的算法适合处理业务问题，请先从“决策树”或“贝叶斯”下手研究资料。如果不知道要包括哪些属性，就选择所有属性。使用依赖关系网络视图，从中获得可帮助用户简化复杂模型的视图。

在原型开发阶段，用户可能希望构建相关模型，以便评估最佳算法和模型。使用挖掘准确度图表评估在预测中效果最佳的模型。用户可能还希望构建相关模型，对相同的数据执行不同类型的分析。这些模型在作为相关模型时的处理速度要比作为独立定义模型时的处理速度快。

在构建和测试原型后，便可以构建和测试实际数据挖掘模型。在将数据输入数据挖掘引擎前，如果需要转换数据，那么为了要实现这些操作，应当开发可供生产用的操作

流程。在某些情况下，可能要选择从 DTS 管道直接植入挖掘模型。如果在少量数据的基础上开发原型，将需要在整套培训数据的基础上重新评估备选模型。

7-6-2 构建数据挖掘应用程序

在 Business Intelligence Development Studio 中开发和研究数据挖掘模型可使企业获得巨大的价值。用户可以浏览模型，了解数据与业务之间的关系，并使用该信息促进决策的制定。但是，其最大的价值还是来自可以影响公司日常操作的数据挖掘应用程序：例如，向客户推荐产品、记录客户信用风险，或根据预测的库存不足下订单的数据挖掘应用程序。要开发可操作的数据挖掘应用程序，需要跳出 Business Intelligence Development Studio 的圈子，并用 Microsoft Visual Studio 或选择的其他开发环境编写代码。

大部分企业客户都将面向客户的数据挖掘应用程序实施为基于 Web 的 Win32 应用程序，如 ASP 网页。数据挖掘模型业已构建完毕，而且应用程序也可以根据客户的选择或在 Web 商务应用程序中输入的内容，为客户执行预测。这可能是十分简单的应用程序；唯一不寻常的部分是发布预测查询。

数据挖掘应用程序开发人员不一定就是开发数据挖掘模型的人员。应用程序开发人员应具备一流的开发技能，而对业务或统计知识的需求则相对较低。

Microsoft 的数据挖掘技术大大地简化了构建自动化数据挖掘应用程序的过程。其中共有两个步骤：

- 开发数据挖掘预测查询，其 DMX 语法在“数据挖掘”规范的 OLE DB 中定义。不需要手工编写 DMX，用户只需单击 Business Intelligence Development Studio 编辑器左栏上的“挖掘模型预测”图标即可。“挖掘模型查看器”图形化工具会有助于开发预测查询。
- 在数据挖掘应用程序中使用预测查询。如果应用程序只使用 DMX 便可完成预测，则项目应包括 ADO、ADO.Net 或 ADOMD.Net 等类引用（建议在 Beta 1 之后的开发中使用 ADOMD.Net）。如果用户正在构建一个更为复杂的应用程序（例如要显示用户挖掘模型查看器，如“决策树查看器”），将需要包括 Microsoft.AnalysisServices 和 Microsoft.Analysis- Services.Viewers 类。

有些客户（主要是独立软件供应商）希望创建可生成数据挖掘模型的应用程序，这种应用程序可能会替代在 Business Intelligence Development Studio 中开发挖掘模型，但可能只适用于特定的领域，如 Web 分析。在这种情况下，开发项目就需要包括 Microsoft.DataWarehouse.Interfaces，以便获得对 AMO（Analysis Management Objects，分析管理对象）的访问权限。

7-6-3 DMX 范例

数据挖掘过程包括三个步骤，分别为：创建数据挖掘模型、培训模型和根据模型预测行为，这三个步骤都可通过简单、类似 SQL 编程语言的 DMX 来实现。范例语法如下所示；DMX 的完整使用方法可从联机帮助中获得。

- 创建数据挖掘模型

```
CREATE MINING MODEL CreditRisk
(CustID LONG KEY,
Gender TEXT DISCRETE,
Income LONG CONTINUOUS,
Profession TEXT DISCRETE,
Risk TEXT DISCRETE PREDICT)
USING Microsoft_Decision_Trees
```

- 培训数据模型

```
INSERT INTO CreditRisk
(CustId, Gender, Income, Profession, Risk)
SELECT CustomerID, Gender, Income, Profession, Risk
From Customers
```

- 根据数据挖掘模型预测行为

```
SELECT NEWCUSTOMERS.CUSTOMERID, CREDITRISK.RISK,
PREDICTPROBABILITY(CREDITRISK)
FROM CREDITRISK PREDICTION JOIN NEWCUSTOMERS
ON CREDITRISK.GENDER=NEWCUSTOMER.GENDER
AND CREDITRISK.INCOME=NEWCUSTOMER.INCOME
AND CREDITRisk.Profession=NewCustomer.Profession
```