PROGRAMMER TO PROGRAMMER™

Expert SQL Server 2005 Integration Services

SQL Server 2005 Integration Services 专家教程

(美) Brian Knight 著
Erik Veerman
冯 飞 译

WILEY
wrox™
清华大学出版社

Expert SQL Server 2005 Integration Services

SQL Server 2005 Integration Services 专家指南

[illegible]

[illegible]

SQL Server 2005 Integration Services 专家教程

(美) Brian Knight
Erik Veerman 著
冯 飞 译

清华大学出版社
北 京

Brian Knight, Erik Veerman
Expert SQL Server 2005 Integration Services
EISBN: 978-0-470-13411-5

图书在版编目(CIP)数据

SQL Server 2005 Integration Services 专家教程/(美)耐特(Knight, B.)，(美)弗尔曼(Veerman, E.)著；冯飞 译. —北京：清华大学出版社，2008.10

书名原文：Expert SQL Server 2005 Integration Services

ISBN 978-7-302-18554-3

I. S…　II.①耐…　②弗…　③冯…　III.关系数据库—数据库管理系统，SQL Server 2005—指南　IV.TP311.138-62

中国版本图书馆 CIP 数据核字(2008)第 140764 号

责任编辑：王　军　徐燕萍
装帧设计：孔祥丰
责任校对：成凤进
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市春园印刷有限公司
装 订 者：三河市李旗庄少明装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：24　**字　数**：584 千字
版　　次：2008 年 10 月第 1 版　　**印　　次**：2008 年 10 月第1 次印刷
印　　数：1～4000
定　　价：48.00 元

作 者 简 介

Brian Knight(SQL Server MVP、MCSE、MCDBA)，来自佛罗里达州绿湾泉，是SQLServerCentral.com 和 JumpstartTV.com 的共同创始人。他在 Jacksonville(JSSUG)开办了一个本地的 SQL Server 用户组，并且是 Professional Association for SQL Server(PASS)的董事会成员。他是 SQL Server Standard 的专栏作家，也负责数据库网站 SQLServerCentral.com 的一个专栏，并在 JumpstartTV.com 上定期发表评论。他分别与人合著和独著了 9 本 SQL Server 的书籍，包括 *Admin911: SQL Server 2000*(McGraw-Hill Companies), *Professional SQL Server 2000 DTS*(Wiley Publishing), *Professional SQL Server 2005 Administration*(Wiley Publishing)，和 *Professional SQL Server 2005 Integration Services*(Wiley Publishing)。他还在 PASS、SQL Connections 和 TechEd 等会议以及许多 Code Camps 中作演讲。他的博客地址为 www.whiteknighttechnology.com。

Erik Veerman(SQL Server MVP、MCSE、MCDBA)是 Solid Quality Learning 的顾问，专职于 SQL Server Business Intelligence(BI)平台上的培训、部署和架构解决方案工作。他是 Microsoft 的 Worldwide BI Solution of the Year 以及 SQL Server Magazine 的 Innovator Cup 获得者。他设计了大量跨广泛业务范围的 BI 解决方案，涉及通信、市场营销、零售、商业房产、金融、供应链和信息技术等。他有着大数据量、多 TB 的环境以及 SQL Server(64 位)方面的经验，使得客户可伸缩他们基于 Microsoft 的 BI 解决方案以达到最优。作为 OLAP 设计、ETL 处理和维度建模方面的专家，他是先驱者、作家和指导者。他领导了第一次用 SQL Server Integration Services(SSIS)实现 ETL 体系结构和设计，帮助在 Microsoft 的 SQL Server 2005 reference initiative(Project REAL)上推动 ETL 标准和 SSIS 的最好实践。他还是 *Profession SQL Server 2005 Integration Services*(Wiley Publishing，2006)一书的合著者。他居住在乔治亚州的亚特兰大，是本地 Atlanta SQL Server 用户组(PASS 和 INETA 用户组分支机构)的负责人。

前　言

曾有一些书介绍过 SQL Server Integration Services(SSIS)——事实上，本书的前篇 Professional SQL Server 2005 Integration Services(Indianapolis:Wiley，2006)就对此工具作了详细介绍。不过，尽管许多相关技术书籍都写得很好，也是初次尝试时极好的参考指导，但将这一技术应用于解决方案并不是那么简单的事情。

例如，您在当地五金店购买的电动工具都会带有用户手册。但仅看一眼目录索引即可知道，手册的目的只是介绍工具的旋钮和按钮。例如电动锯的手册就可能会描述如何抬升和降低锯条，如何弯曲手臂来切 45° 角。但往往不会介绍如何学习制造柜子或床。显然，您不能指望手册来教会您这一切，特别是如果该电动锯有成百种不同的用途，那就更不可能了。

正如您所料，SSIS 的情况与之类似。目前的在线文档和 SSIS 书籍很好地讲述了如何使用 FTP 任务，如从远程服务器中拖文件，以及如何连接文件将数据提取到表中。但它们并没有介绍如何整合这些内容，将 SSIS 应用于特定目的。当然，SSIS 的许多用户很高兴地发现可根据在线文档实现一些标准的一次性任务。但在使用该工具构建解决方案时，需要更多的信息。

本书有关于应用——即应用 SSIS 的功能帮助预想、开发和实现数据处理需求。

本书读者对象

在本书中，我们将把 SSIS 功能应用到一些常见的行业领域，包括数据仓库提取、转换和加载(ETL)，数据集成 ETL，高级 ETL 开发和管理。因此，本书针对 3 种主要的 SSIS 用户：

- ETL 数据架构师
- ETL 开发员
- 负责 SSIS 的数据库管理员(DBA)

阅读本书的大部分人都有其他行业 ETL 工具的背景，如 Ab Initio、Informatica 和 Ascential。而其他人可能最为熟悉 DTS(Data Transformation Service)这个 ETL 工具。也可能有一些人是 SQL(如 T-SQL 或 PL-SQL)脚本专家，能运用关系引擎中的脚本构建处理算法。如果有此背景，那一定熟悉一般的 ETL 概念，而本书将向您介绍 SSIS 中的数据仓库 ETL。

如果你是 ETL 的新手，但由于您所在公司做新的改革或是出于自身需要来学习这一技

能，那将从本书中获得 ETL 的基本知识，以及如何将 SSIS 用于这一目的。

本书内容

不同于那些只注重功能介绍的书，本书的目的不是向您介绍所有属性、所有功能，而只是教会您如何将 SSIS 组件应用到 ETL 任务。

注意：

本书解答了下列问题：如何使用 SSIS 构建企业 ETL 解决方案，使得该解决方案可伸缩和可执行、很好地处理错误，以及给管理人员提供正确的信息来管理和监控处理数据处理。

本书结构

为了介绍如何将 SSIS 应用于不同任务，我们有意安排了本书的章节，以方便读者理解。本书首先提供了有关脚本和数据提取的所有背景知识和基础信息。然后详细介绍了数据仓储 ETL、错误处理、管理和数据集成的内容。由于 DTS 的流行性，还有一章讲到了从 DTS 到 SSIS 的迁移，另有一章提到伸缩 SSIS。下面简单介绍一下本书的结构：

- 第 1 章介绍了 SSIS 带给 ETL 和集成的价值主张，还回顾了 SSIS 基础知识。
- 第 2 章重点介绍了高级脚本知识。本章提前介绍是因为理解何时使用脚本以及如何实现脚本对于体系结构决策来说很重要。另外还介绍了一些场景，展示了脚本对于更高级需求的有效性，而这些无法通过其他现成组件轻松设计实现。
- 第 3 章重点介绍了数据提取和谱系(ETL 的核心概念)，即 ETL 中的 E。该章主要介绍设计提取，包括增量提取、从源到目标跟踪数据。
- 第 4~6 章介绍了数据仓储 ETL，主要是因为大部分新数据仓库都将 SSIS 用于 ETL 过程，许多已有的 ETL 解决方案用 SSIS 重写了。第 4 和第 5 章介绍了关系数据库转换及维度表和事实表的加载方法。第 6 章介绍了 SSIS 中的 SQL Server 2005 Analysis Services 支持。
- 第 7 章介绍了错误和事件处理以及可重启性。它提供了设计解决方案的能力，这样的解决方案可优雅应付错误处理和通过可重启性使执行变得轻松。
- 第 8~9 章介绍了支持和生产环境之间移动的最好实践。还介绍了管理这一过程的方法，以及程序包配置和执行。
- 第 10 章介绍了异构集成。这样就可以从非 SQL Server 系统或文件(如 Oracle、DB2、Sybase、Teradata 和非 ANSI 代码页面文件等)中拖出数据或向其中推入数据。本章介绍了在集成这些系统时会涉及的内容。
- 第 11 章介绍了如何利用 SSIS 功能使得迁移的程序包变得更好。可能有些人会将基于 DTS 的 ETL 迁移至 SSIS。由于产品的体系结构不同，迁移可能会需要一些处理。第 11 章将讨论包含在 SSIS 中使用程序包在内的诸多内容。

- 第 12 章讨论了如何最好地利用内存，何时使用关系型引擎，在何处执行 SSIS 程序包，以及有什么好的加载技术可用于目标。

使用本书的要求

由于本书讲解的是 SQL Server 2005 Integration Services，所以如果您有 SQL Server 2005 的开发人员版，那就会得到更多的内容，包括示例应用程序和工具。本书针对 SQL Server 2005 SP2 而作(2007 年 2 月发布)。如果您还没有 SQL Server 2005 许可版，那你不好先从 Microsoft 下载站点下载试用期为 120 天的版本先行试用(http://download.microsoft.com)。

源代码

在完成本书的示例时，可以选择手动输入代码或者直接使用本书附带的源代码文件。本书中用到的所有源代码可以从 www.wrox.com(或 www.tupwk.com.cn/downpage)下载。进入该站点后，只需找到本书的名称(使用 Search 框或者书名列表)，单击本书详细页面上的 Download Code 链接，就可以得到本书所有的源代码。

注意：
由于很多书有相似的名称，所以用 ISBN 搜索更为容易。本书的 ISBN 是 978-0-470-13411-5。

下载了代码后，用您喜欢的压缩工具把它解压缩。此外，也可以去 Wrox 的主下载页面 www.worx.com/dynamic/books/download.aspx 找到本书或其他 Wrox 出版的书的代码。

勘误表

尽管我们已竭尽所能来确保在正文和代码中没有错误，但人无完人，错误难免会发生。如果您在 Wrox 出版的书中发现了错误(例如拼写错误或者代码错误)，我们将非常感谢您的反馈。发送勘误表将节省其他读者的时间，同时也会帮助我们提供更高质量的信息。

到 www.wrox.com 站点上，用 Search 框或者标题列表找到本书的名称，在详细页面上点击 Book Errata 链接就能找到本书的勘误表。在这个页面中可以看到所有被提交的本书的勘误表，它们是由 Wrox 的编辑发布的。在 www.worx.com/misc-pages/booklist.shtml 中有完整的书的列表，其中包括每本书的勘误表。

如果您在书的勘误表页面上没有看到您发现的错误，可以到 www.wrox.com/contact/techsupport.shtml 上填写表单，把您发现的错误发给我们。我们会检查这些信息，如果属实就把它添加到本书的勘误表页面上，并在本书随后的版本中更正错误。

p2p.wrox.com

如果想和作者或者其他人讨论，请加入在 http://p2p.worx.com 的 P2P 论坛。该论坛是基于 Web 的系统，您可以发布关于 Wrox 出版的书和相关技术的消息，与其他读者或技术人员交流。该论坛有预定功能，在您选择的感兴趣的主题有新帖子时，会邮件通知。Wrox 的作者、编辑、其他业界专家和像您一样的读者都会出现在这些论坛中。

在 http://p2p.wrox.com，您会找到很多不同的论坛，它们不但有助于您阅读本书，还有助于您开发自己的应用程序。加入论坛的步骤为：

(1) 到 http://p2p.wrox.com 上单击 Register 链接。

(2) 阅读使用说明，单击 Agree 按钮。

(3) 填写加入必需的信息和其他您愿意提供的信息，单击 Submit 按钮。

(4) 您将收到一封 E-mail，描述如何验证您的账户并完成加入过程。

注意：

不加入 P2P 也可以阅读论坛里的消息。但是如果要发布自己的消息，就必须加入。

加入之后，就可以发布新的消息和回复其他用户发布的消息。可以随时在 Web 上阅读论坛里的消息。如果想让某个论坛的新消息以 E-mail 的方式发给您，可以单击论坛列表里论坛名字旁边的 Subscribe to this Forum 图标。

要了解如何使用 Wrox P2P 的更多信息，请阅读 P2P FAQs，其中回答了论坛软件如何使用的问题，以及许多与 P2P 和 Wrox 出版的书相关的问题。要阅读 FAQ，单击任何 P2P 页面里的 FAQ 连接即可。

目　录

第 1 章

绪　言

本书主要是以讲解实际应用为主，特别讲解了应用 SQL Server 2005 Integration Services(SSIS)的功能，帮助设计、开发和实现数据处理需求。本书的讨论主要集中于 SSIS 如何帮助实现数据集成和处理需求。

SSIS 最擅长的数据处理核心技术就是 ETL，即提取(Extraction)、转换(Transformation)和加载(Loading)。该 ETL 技术在多年以来被赋予了多种不同的意义。有从将数据从一个地方移至另一地方的一般性观点，也有说是数据仓储 ETL 的特定应用。事实上，ETL 主要就是商业智能(BI)和数据仓库处理。

本章为 DBA 所需的广义 ETL 提供了重要的背景信息，以及基本的数据仓储 ETL 概念。另外，还包括了对 SSIS 功能的应用回顾，为本书研究和应用 SSIS 功能打下了基础，可帮助实现您个人的数据集成和处理需求目标。

1.1　选择合适的工具

如果您想进行家庭装修，那就有机会在当地的家居装潢店的工具区转悠。上千种工具被用来实现各种功能，有时，使用方式较复杂。

任何做杂役的新手都可以体会到这样一句名言：工欲善其事，必行利其器。同样的概念适用于数据处理。毫无疑问，对于任一情况，都会有一个特定的工具来进行处理。考虑一下您的组织机构内不同类型的数据处理需求：

- 系统间的数据同步
- 从 ERP 系统提取数据

- 即席报告
- (同构和异构)复制
- PDA 数据同步
- 遗留数据集成
- 供应商和合作方数据文件集成
- 业务数据线
- 客户和雇员目录同步
- 数据仓库 ETL 处理

正如您可能所知道的那样，在进行数据处理时，有许多工具供选择。有些用于特定的目的(如文件夹同步工具)，而其他一些工具用于执行不同情形下的多种功能。因此，就有了一个很常见的问题——哪种工具最符合任务所需的业务和逻辑需求。

看一下 Microsoft 工具组中包含的工具。可以使用 Transact SQL(TSQL)手动编码加载，用 Host Integration Server 与不同的数据源进行通信，用 Biztalk 以事务的方式协调消息，或用 SSIS 以批处理方式加载数据。这些工具都在数据世界中起着作用。

尽管可能有功能重叠，但每种工具都有明确的针对性和目标功能。当您习惯于某一技术时，如果另外有一工具更适合，您会倾向于超出其最佳应用范围来应用这一技术。您肯定听说过这样一句话：如果您是一个铁锤，您看什么都是钉子。例如，C#开发人员可能想构建应用程序来做 SSIS 可能要花费一小时完成的事。每个人要解决面对的困难都需要时间和能力。没有人是个全能型的专家。因此，开发人员和管理人员都应该努力研究工具和技术，使得它们在不同的情形下互为补充。

例如，许多组织将 Biztalk 用于很多目的，不仅仅是 B2B 通信的处理和工作流过程自动化。他们也会困惑，为什么没有将 Biztalk 扩展来满足组织的千兆字节数据仓库 ETL 的需要。简单地说，批量 BI 处理的合适工具是 ETL 工具，如 SSIS。事实上，在图 1-1 中，SSIS 为利用其高性能的数据管道(pipeline)提供了很好的平台。

图 1-1 列出的过程很简单，实际上 SSIS 所做的就是提供使过程有效和可伸缩的技术，以及处理数据错误的功能。

本章较详细地讲述了 ETL 概念，首先是 SSIS 示例。在深入到后继章节作的专家级的介绍之前，要提醒的是，ETL 概念和 SSIS 功能将帮助巩固了解 SSIS 应用程序前所需的背景知识。

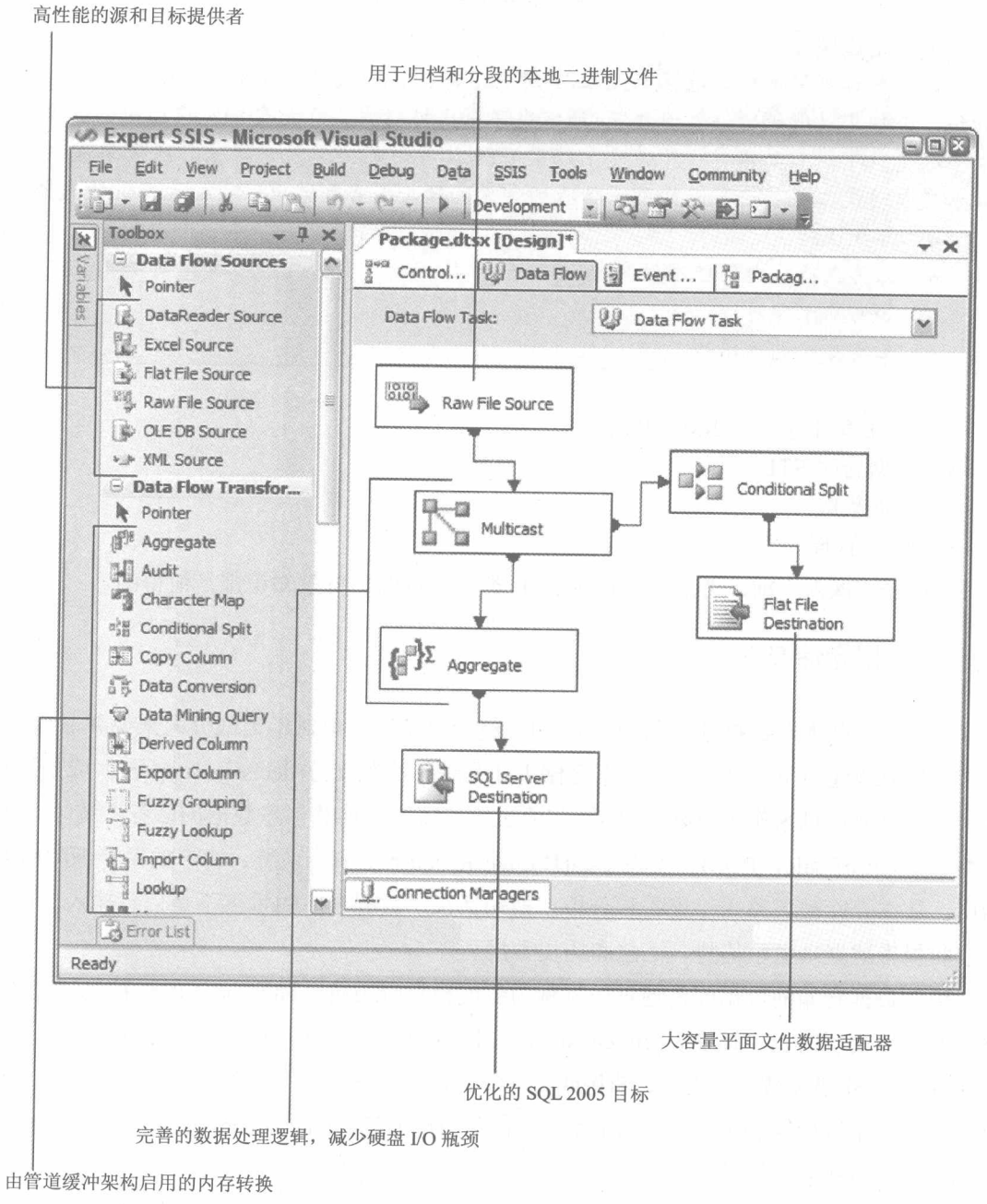

图 1-1　SSIS 高性能数据管道

注意工具选择

在有些客户环境中，选择 ETL 工具时并不考虑行业技术的可用性、支持以及学习曲线。即使工具能够创造“奇迹”，通常也不是小型魔术，只是会掏空您公司的腰包。大多数情况下，需要花费数千美元来购买 ETL 工具，而且需要花很长时间来掌握、实现和支持这类工具。除了一些关于功能的常见问题，还需要考虑下列关于工具的问题：

- 自身技能
- 工具在行业内的使用倾向性
- 如何易于学习
- 工具支持的简单性

本书主要集中于三类 SSIS 用法：

- 数据仓库 ETL
- 数据集成
- SSIS 管理

在进一步深入之前，考虑一下每种 ETL 类型的目的和背景知识都是值得的。

1.1.1 数据仓储 ETL

可能有一些人对数据仓库和相关的 ETL 概念很精通，但对于大部分人来说并非如此。所以这里将概述一下数据仓库。数据仓储主要是进行决策支持(decision support)，或是通过有组织地获取信息来作出更好的决策。与通过快速事务来捕获信息数据的事务系统(如销售点终端(point of sale，POS)、人力资源(Human Resource，HR)系统、客户关系管理(CRM))相比，数据仓库被调整用于报告和分析。换句话说，数据仓库的重点不是信息输入，而是提取和报告数据来显示趋势、汇总和历史数据。

用做数据仓储的数据库是通过一种称为维度模型(dimensional model)的结构来创建的，它涉及两种类型的表。维度表(dimensional table)存储描述实体的信息数据或属性。事实表(fact table)捕获描述数量、级别、销售或其他统计数字的度量或数值数据。一个数据仓库可能包括许多个维度表和事实表。图 1-2 以一种称为星型模式(star schema)结构显示了几个维度表和一个事实表之间的关系，

本书的重点并不在于维度表和事实表的设计，而是将其他仓库的数据装入这些结构中。数据仓储的 ETL 处理包括从源系统或文件提取数据，在数据上执行事务逻辑使数据关联、清除或合并，以及加载数据仓库环境进行报告和分析(见图 1-3)。

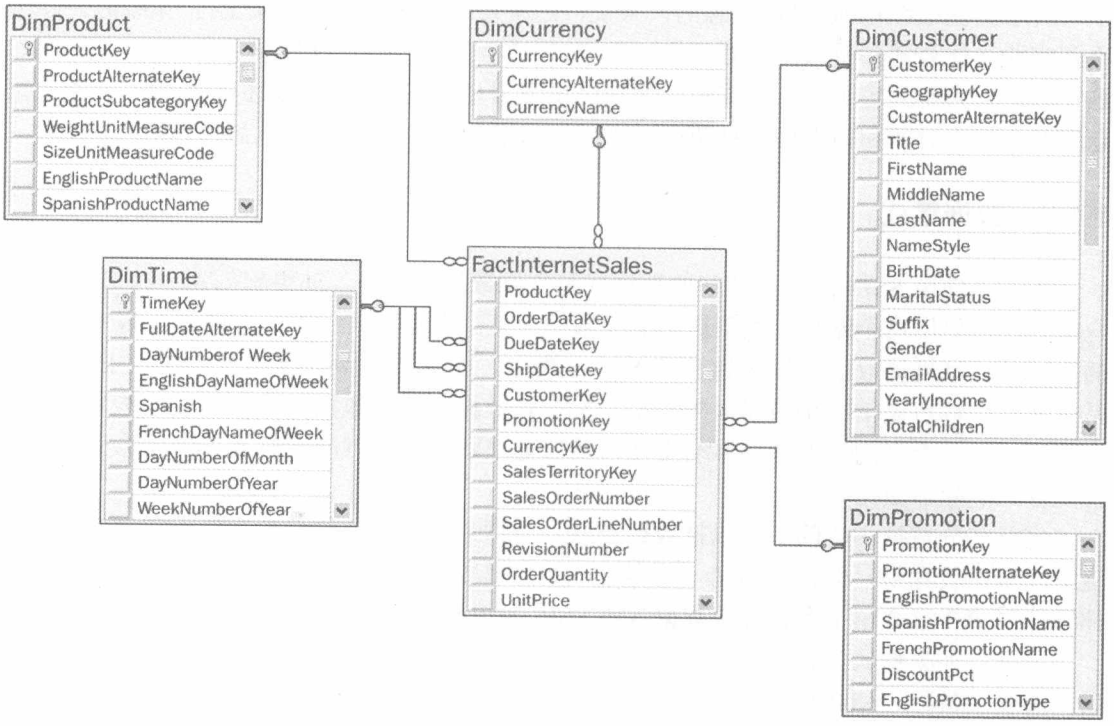

图 1-2　星型模式

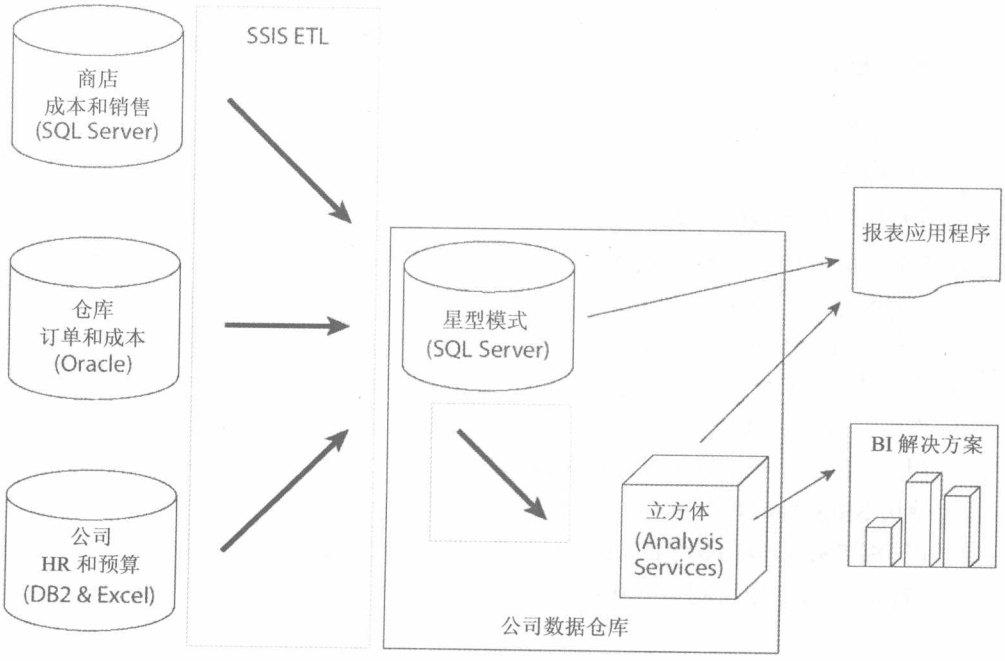

图 1-3　ETL 系统