Sybase IQ

卢东明 主编
罗永强 等 编著
王 珊 审

商务智能数据库

Sybase IQ

技术与应用

清华大学出版社

Sybase IQ

商务智能数据库 Sybase IQ 技术与应用

卢东明　主编
罗永强　等 编著
王　珊　审

清华大学出版社
北京

内 容 简 介

本书以深入浅出的方式介绍分析型数据库引擎 Sybase IQ 和商务智能的理论基础与应用。全书共14章，介绍 Sybase IQ 的历史、体系结构、安装、开发、视图、表、索引、数据加载和卸载，Java 程序开发，Sybase IQ 的性能优化、集群解决方案及配置原理，基于 Sybase IQ 进行非结构化的数据处理，使用中常见的问题及处理办法等。

本书可作为计算机及信息管理专业本科生及研究生的参考教材，也可作为信息化管理者、IT 咨询顾问、IT 技术人员的实用参考手册。

图书在版编目（CIP）数据

商务智能数据库 Sybase IQ 技术与应用 / 卢东明主编；罗永强等编著．—北京：清华大学出版社，2009.11
ISBN 978-7-302-21180-8

Ⅰ. 商… Ⅱ. ①卢… ②罗… Ⅲ. 关系数据库—数据库管理系统，Sybase IQ Ⅳ. TP311.138

中国版本图书馆 CIP 数据核字(2009)第 178115 号

责任编辑：张　民　赵晓宁
责任校对：焦丽丽
责任印制：何　芊

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 装 者：北京市清华园胶印厂
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：17.75　　**字　　数**：410 千字
版　　次：2009 年 11 月第 1 版　　**印　　次**：2009 年 11 月第 1 次印刷
印　　数：1～3000
定　　价：29.00 元

大学计算机基础教育规划教材

近　期　书　目

大学计算机基础(第3版)("国家精品课程"、"高等教育国家级教学成果奖"配套教材)
大学计算机基础实验指导书("国家精品课程"、"高等教育国家级教学成果奖"配套教材)
大学计算机应用基础("国家精品课程"、"高等教育国家级教学成果奖"配套教材)
大学计算机应用基础实验指导("国家精品课程"、"高等教育国家级教学成果奖"配套教材)
C程序设计教程
Visual C++程序设计教程
Visual Basic程序设计
Visual Basic.NET程序设计(普通高等教育"十一五"国家级规划教材)
计算机程序设计基础——精讲多练C/C++语言(普通高等教育"十一五"国家级规划教材)
微机原理及接口技术(第2版)
单片机及嵌入式系统(第2版)
数据库技术及应用——Access
SQL Server数据库应用教程
Visual FoxPro 8.0程序设计
Visual FoxPro 8.0习题解析与编程实例
多媒体技术及应用(普通高等教育"十一五"国家级规划教材)
计算机网络技术及应用(第2版)
计算机网络基本原理与Internet实践
Java语言程序设计基础(第2版)(普通高等教育"十一五"国家级规划教材)
Java语言应用开发基础(普通高等教育"十一五"国家级规划教材)

商务智能数据库 Sybase IQ 技术与应用 序

数据仓库、联机分析技术、数据挖掘是支撑商务智能的三大技术支柱。数据仓库系统则是它们的基础设施。数据仓库服务器用来存储和管理海量数据，包括企业的历史数据和当前数据，细节数据和综合数据，以支持众多用户的联机查询、联机分析处理和数据挖掘。

人们发现从数据仓库大量集成的一致业务数据中通过统计、分析、挖掘，可以找到隐藏在其中的有用信息和商业（或业务）规律，帮助企业进行业务分析和正确决策。因此数据仓库应用日益广泛深入，数据仓库存储和处理的业务数据呈指数级增长，并发查询和分析处理的用户也越来越多。与此同时，数据仓库的性能问题越来越成为系统的瓶颈。

数据仓库服务器一般是关系数据库管理系统或扩充的关系数据库管理系统。传统的关系数据库系统主要针对联机事务处理，关注的是更新数据库，保证事务的原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）和持续性（Durability），即 ACID 特性。而数据仓库系统则应该针对读取海量数据，以及基于海量数据上的统计和分析。

从近几年的 Gartner 报告看，数据仓库引擎出现了列式数据库产品的新方向。Sybase IQ 是 Sybase 公司推出的支持数据仓库的关系型列式数据库，由于其大量的专利技术和 15 年以上产品成熟度，确定了 Sybase IQ 在列式数据库中的领先地位。针对数据仓库应用，Sybase IQ 的架构与大多数关系型数据库不同，其设计与执行优先考虑查询性能，其次是完成批量数据更新的速度。它不是按行（或记录）存储数据而是按列存储数据。这种存储方法适应了数据仓库查询的特点，为提高性能和系统的扩展性提供了优势。

由于同一列数据具有相同的数据类型，所以数据更容易压缩，压缩比更高，不但节省存储空间而且提高了查询性能。

为支持数据仓库中大量并发用户的即席查询，Sybase IQ 采用了多种索引技术，例如位图（Bitmap）索引技术、BitWise 技术、FP（Fast Projection）索引、LF（Low Fast）索引、HNG（High Non Group）索引、HG（High Group）索引、WD（Word）索引、CMP（Compare）索引等。

本书全面地介绍了 Sybase IQ 的技术特点和体系结构，Sybase IQ 的操作使用方法，基于 Sybase IQ 的数据仓库设计以及常见故障处理方法等。

本书作者多年从事分析型系统研究和开发工作，担任过数据仓库开发的高级程序员、系统分析员、需求分析师、项目经理和系统架构师等职务，参与了金融、电信和政府

等多个行业的数据仓库建设，对国内的数据仓库建设现状及未来有深刻的了解。作者结合其数据仓库行业经验，就如何进行基于 Sybase IQ 的数据仓库项目建设提出了自己的见解。

本书的出版，对 Sybase IQ 所采用的列式数据库新技术的普及和 Sybase IQ 的使用将不无裨益。

于中国人民大学

2009 年 9 月

王珊，中国人民大学信息学院教授、博士生导师，国家级教学名师，历任中国计算机学会副理事长、中国计算机学会数据库专委会主任、教育部第五届科学技术委员会委员和中国科学技术协会第六届全国委员会委员等。

前言

商务智能数据库 Sybase IQ 技术与应用

前几天看到一份杂志，一个国际知名的自行车厂商在讲到如何让自行车的性能达到国际竞争水平时说："4 年的辛勤耕耘就是为了研究如何能在风洞中减少百分之一的流动阻力，这真的很疯狂。但最后，它可能让赛车手快上四分之三秒。"

其实，当前 IT 的发展又何尝不是在这样一种境遇呢？为了一项技术的发展，往往许多公司的工程师们绞尽脑汁，夜以继日地工作，不过为了让现有的技术水平提高百分之一，有时只是万分之一。科技的每一次进步、每一次飞跃都带给我们振奋，因为它带来生产力的提升，可能原 10 个人的工作量变为 5 个人完成，也可能过去一周的工作变为 3 天做完，也可能预算 6000 万元的项目变为 5000 万元竣工。然而，相比这个自行车厂商来说，在 IT 界更令人兴奋的是，有时 IT 的一个飞跃就像"量子跳跃"一样，给人们带来全新的能量和全新的思考。

关系型数据库的发展至今，经历了 30 多年，各个厂商在其发展的不同阶段都做出了不同的贡献，但是很多数据库的核心技术和理念变化并不多，如 30 年后的今天，我们仍然使用 SQL 语言，虽然此 SQL 非彼 SQL，现在的 ANSI SQL/92 比 SQL 语言的初期有了很多功能性的扩充和完善，但是突破性的变化并未发生。影响数据库性能最大的存储层面技术在 30 年来一直以行式存储垄断各家公司的产品，OLTP 系统对数据库的要求在最短的时间里处理最多的并发事务（Transaction），使得行式数据库的"短小精悍，快进快出"的优点发挥得淋漓尽致。随着近些年来分析系统的日益重要，商务智能 BI 技术的广泛使用，历史数据的大量积累，行式数据库的局限性也反映得越来越明显。

以 Sybase IQ 为代表的列式数据库正是因这种业务发展的需求而产生的一个创新的技术，是数据库 30 多年的历史上第一次在存储层面进行的革命性的飞跃，数据不再是传统地按行存储和访问，而是按列存储，按列访问的。这一底层根本的变革带来了许多使用上的变化，例如分析时常用的在几个列上的查询不再浪费大量的系统 I/O；相同的数据类型以及类似的数据特征使得在列上的数据压缩变得不仅可能，而且高效；这些特点在 Sybase IQ 的实用案例中得到了充分的体现，列式数据库 Sybase IQ 的价值更是通过分析查询时间、硬件系统资源直至机房能耗的大量节省来体现。

这些特点使得列式数据库 Sybase IQ 在国内外的大型数据库和数据仓库系统中有了用武之地，很多过去不敢想象的数据量在 Sybase IQ 的魔力下运行自如。我们希望本书能够给列式数据库在中国的普及和正确使用发挥效力，使这一国际 IT 界的先进技术尽快在中国发扬光大，为各企业的信息化增添武器，提升企业的核心竞争力。

本书共 14 章，从 Sybase IQ 的基本原理开始，系统地介绍 Sybase IQ 的使用中必须了解的功能和特性，其中结合了大量 Sybase IQ 在国内外部署及使用过程中遇到的经验，是一本理论结合实践的书。本书由卢东明主编，第 1～5、7、10、11、14 章由罗永强撰写，第 6、12、13 章由刘广军撰写，第 8、9 章由夏小涛撰写。

因水平有限，书中难免存在缺点和不足，望读者不吝赐教。

编　者

2009 年 9 月

商务智能数据库 Sybase IQ 技术与应用 目录

第1章 商务智能基础

随着20世纪90年代后期Internet的兴起与飞速发展，我们进入了一个新的时代，商业的有效性比以往更受关注。将信息放到网上并不能自动地产生销售或利润。大量的业务仍然需要通过其他已建立的渠道面对面的进行。最需要关注的就是商业绩效和市场竞争——不论是在网上还是在大街上，必须有效地针对客户，以合适的价格提供合适的产品与服务以及更具人性化的交流。这需要商务智能(Business Intelligence，BI)。

从企业的内部信息化建设来看，大多数公司已经完成了日常业务的信息化，长期的业务运行也积累了大量的历史数据，如何有效利用企业的历史数据，形成分析结论，指导未来的生产经营是企业信息化建设的一个新的方向。从企业的外部环境看，不断变化的电子商务，对电子商务态度的转变，新技术的出现如无线和移动网络，对竞争与基本业务性能的进一步关注，个性化的潮流，所有这些因素使企业对商务智能更加渴望，对商务智能的有效性及其部署成本愈加关注。商务智能正步入一个新的时代，我们需要重新考虑一个适应时代需求的商务智能解决方案。

1.1 商务智能

商务智能作为正式的概念，是由美国Gartner公司的分析师Howard Dresner于1989年首次创造的这个名词，而早在1985年，美国宝洁公司就已经设计并开始利用类似商务智能的信息体系了。然而从概念到大规模的推广，还是有很长一段路要走。1996年，Gartner公司提出了信息民主(Information Democracy)的概念，极大地推动了商务智能的应用。它认为商务智能是“使企业在竞争市场中保持领先地位的关键所在。正确的商务决策是以准确和及时的信息为基础的，而不是靠直觉。数据分析、报告及查询工具可帮助企业用户成功穿越数据海洋，并从中得到有价值的综合信息”。

自从关系数据库出现以后，联机事务处理(Online Transaction Processing)系统迅速发展，然而当人们初步设想从数据中提炼知识，却发现难以实现，而原因往往是找不到完整的数据，因为业务数据存储在分散的、异构的数据库中。所以人们开始思考把数据格式统一并集中到一起的方法，这就是数据仓库(Data Warehouse)的由来。随着DW的海量数据的积累，就出现了数据挖掘(Data Mining)的概念。然而DM仅仅是产生了一堆统计报告，管理人员往往发现这些报告没有用，或不适用，无法对企业的经营决策带来帮助，于

是数据库营销出现了。BI 开始与 CRM 相结合，对客户的购买行为进行有针对性的统计，并发现其中的规律，从而为企业的营销管理决策提供支持。为了更好地与企业的业务流程结合，报告展示层面的技术不断完善，并达到了可以定制和设计数据分析模式的高度，直接为企业提供上至战略规划，下至业务操作的支持。

综上所述，商务智能的产生和发展，正是数据分析与企业运营、业务流程紧密结合的过程，为企业创造持续的竞争优势。2004 年的商务智能峰会确定了商务智能即是数据仓库之上的查询、报表和多维数据分析。正如峰会发言人所概括的“BI 是一堆技术工具的集合”。2007 年的商务智能峰会对商务智能的重定义：“将商务智能视为一个伞状的概念，它包括了分析应用、基础架构和平台及良好的实践。”可见，数据仓库、数据标准等平台已经涵盖在商务智能范畴里，商务智能已不再是前端展现工具；商务智能不再仅属于技术的范畴了。

1.2 商务智能的应用层次

随着市场竞争的加剧和计算机应用的普及，企业对计算机应用的需求已日益扩展，他们已不再满足于计算机能帮助他们迅速地处理具体业务，而要求计算机能帮助他们从积累的大量业务数据中探索业务活动的规律性、市场的运作趋势，并为他们参与市场竞争做出重要的决策。概括地说，即要求计算机系统为他们实现商务智能。完备的商务智能应包含 5 个层次，由低到高依次如下：

(1) 查询。对企业范围数据的各种形式的访问。

(2) 报表。查询结果的格式化展现。

(3) 在线分析(Online Analytical Processing)。多视角、多层次的数据分析。

(4) 知识挖掘。探索、揭示企业数据之间的内在联系和规律。

(5) 预测。对企业未来发展趋势的推测、模拟和仿真。

这 5 个层次互相关联，每一层次都满足企业的部分需求。查询报表可使企业了解自己的过去和现在；OLAP 分析本质上是一种复杂的查询，它所提供的多维视图使企业对自己的了解更全面、更深入；查询结果和 OLAP 视图都可以报表的形式展现和保存下来，供以后或上级部门使用；知识挖掘运用各种高级分析方法和模型分析数据之间的内在关联，从而使企业准确把握自己的运作规律；而预测则在知识挖掘的基础上推测企业的未来状况，使企业可以根据预测提前规划未来，保证自己在竞争中的优势地位。

1.3 企业级数据仓库

企业级数据仓库(Enterprise Data Warehouse，EDW)是一个企业信息化水平到一定阶段的必然产物。企业的信息化一般是从办公自动化、管理信息系统、企业资源计划等一路发展而来。企业信息化的早期，一般是期望通过计算机替代人工繁琐计算，帮助企业运行业务。多年的信息系统运行为企业积累了大量的数据，这些数据中包含了企业内部管理和外部竞争的有用信息，但是如果这类数据缺乏有效管理，对企业来说反而是一个负