普通高等教育
软件工程 “十二五”规划教材

12th Five-Year Plan Textbooks
of Software Engineering

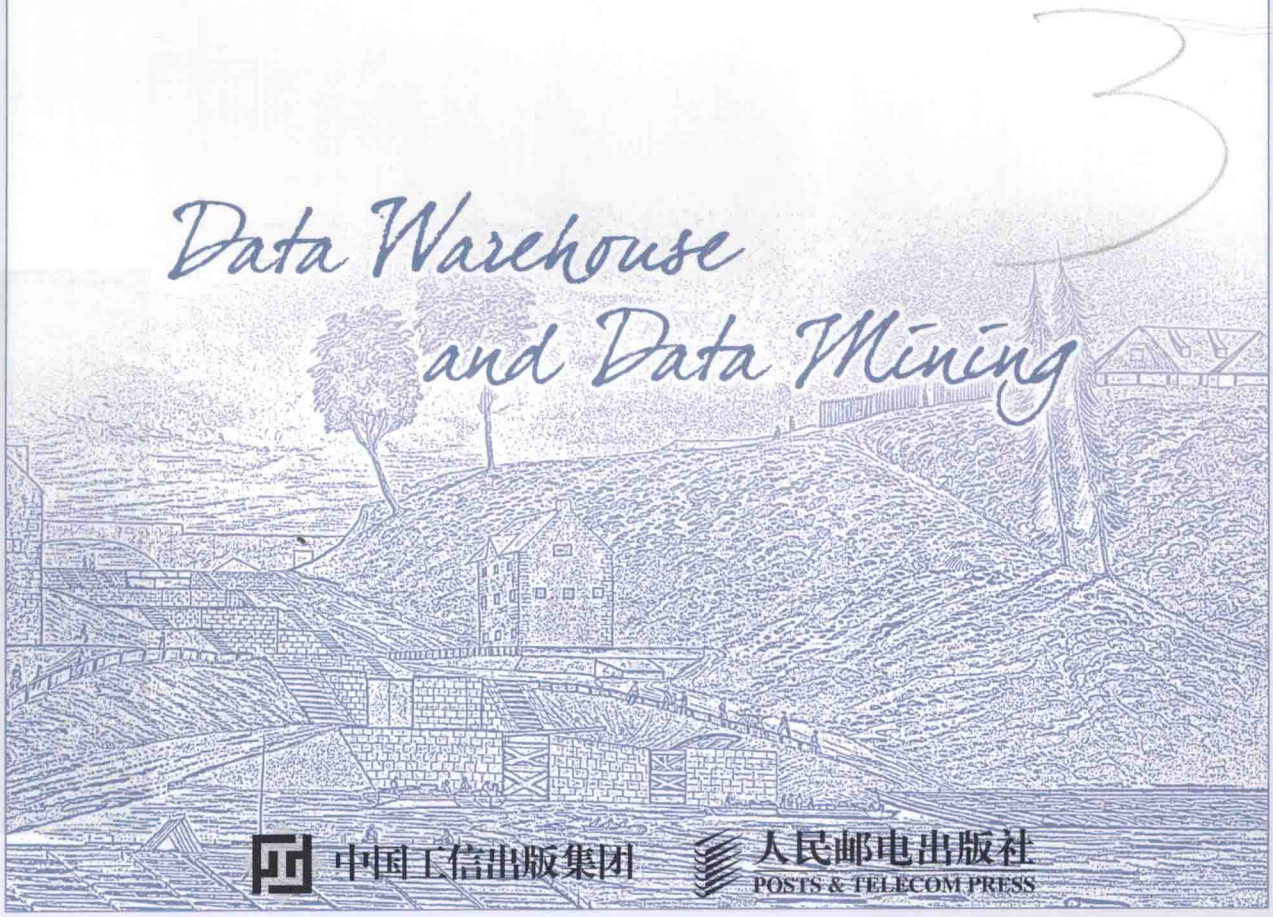
工业和信息化部
“十二五”规划教材

IBM大学合作项目

数据仓库与数据挖掘

袁汉宁 王树良 程永 金福生 宋红 ◎ 编著

Data Warehouse and Data Mining



中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

普通高等教育
软件工程“十二五”规划教材

12th Five-Year Plan Textbool
of Software Engineering

工业和信息化部
“十二五”规划教材

IBM大学合作项目

数据仓库与数据挖掘

袁汉宁 王树良 程永 金福生 宋红 ◎ 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

数据仓库与数据挖掘 / 袁汉宁等编著. -- 北京 : 人民邮电出版社, 2015.7
普通高等教育软件工程“十二五”规划教材
ISBN 978-7-115-38827-8

Ⅰ. ①数… Ⅱ. ①袁… Ⅲ. ①数据库系统－高等学校－教材②数据采集－高等学校－教材 Ⅳ. ①TP311.13 ②TP274

中国版本图书馆CIP数据核字(2015)第100204号

内 容 提 要

本书将数据视为基础资源，根据软件工程的思想，总结了数据利用的历程，讲述了数据仓库的基础知识和工具，研究了数据挖掘的任务及其挑战，给出了经典的数据挖掘算法，介绍了数据挖掘的产品，剖析了税务数据挖掘的案例，探索了大数据的管理和应用问题。

全书深入浅出，强调基础，注重应用，是软件工程及相关专业的高年级本科生、研究生的理想教材，也可作为相关领域的参考用书。

◆ 编　　著　袁汉宁　王树良　程　永　金福生　宋　红
责任编辑　邹文波
责任印制　沈　蓉　彭志环
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市潮河印业有限公司印刷
◆ 开本：787×1092　1/16
印张：13　　2015 年 7 月第 1 版
字数：338 千字　　2015 年 7 月河北第 1 次印刷

定价：39.00 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前　言

数据是信息世界的基础性资源，因为体量巨大、种类繁多、变化快速、真实质差等问题，导致难以发挥数据的价值。为此，产生了数据仓库与数据挖掘，主要研究如何管理、分析和利用数据。

本书立意新颖，结构合理；强调基础，注重应用，重视实例；坚持理论联系实践，从学生中来，为学生服务；和市场结合，为市场服务；既重视理论知识的讲解，又强调应用技能的培养，重点介绍面向专业领域的典型案例。全书面向数据的特点和未来趋势，利用软件工程的思想，根据数据仓库与数据挖掘的内在联系，在一个统一框架内，深入浅出地讲述了数据仓库和数据挖掘的核心内容。全书共分 9 章，依次介绍了数据利用的发展过程，总结了数据仓库和数据 ETL 的基础知识，详述了数据仓库和数据 ETL 的工具，研究了数据挖掘的任务及其挑战性问题，给出了经典的数据挖掘算法，介绍了数据挖掘的工具与产品，剖析了税务数据挖掘的案例，最后结合实际工具，就大数据的管理和应用给出了编者的独特见解。

在本书正式出版前，我们在教学实践中一直采用其前身——《数据仓库与数据挖掘讲义》。该讲义自 2010 年以来，先后在武汉大学、北京理工大学、武汉理工大学使用。使用期间，学生通过对具体实例的学习和实践，掌握了数据仓库和数据挖掘中必要的知识点，达到了学以致用的目的。同时，典型案例不断被完善，在案例分析翔实度、软件工程思想体现方面也逐步改进和补充。本书由多位教师集体编著，实现了高校教师和 IBM 软件高级工程师的深度合作。北京理工大学的袁汉宁副教授、王树良教授、金福生副教授、宋红教授有着多年的教学实践经验，熟知教学规律，了解学生需求；IBM 软件集团中国区合作伙伴技术支持（BPTS）高级信息工程师程永有丰富的软件研制和咨询经验，精通软件技能，了解社会经济的实际需求。更为重要的是，在本书的编写过程中，我们自始至终邀请学生全程参与，包括课堂讨论，作业讲评，内容规划，资料搜集，书稿一读再读等。感谢博士研究生李延、王大魁、李草原、李东伟等的积极参与，辛勤劳动，在资料采集中融会贯通，在讨论中作出建设性发言，在阅读中及时发现问题并给出有益的修改建议。讲义经过多年的教学使用，效果相当不错，深受教师和学生的欢迎和好评，甚至被视为专业必备宝典，不离左右。

本书备受同行推崇，入选工业和信息化部“十二五”规划教材，获得 IBM 大学合作项目书籍出版资助。同时，本书还获得国家自然科学基金（61173061, 61472039, 71201120）、高等学校博士学科点专项科研基金（20121101110036）的支持，在此一并致谢。

本书可作为高等学校软件工程及相关专业高年级本科生、研究生的数据仓库和数据挖掘实用教程，也可供相关领域的广大科研、工程人员和高校师生参考。

由于写作时间仓促，编者水平有限，书中难免有疏漏之处，敬请大家指正。如果有关于本书改进的建议和意见，请直接发送邮件至 yhn6@bit.edu.cn 进行交流。感谢您的支持。

在本书的编著过程中，编者袁汉宁、王树良的妈妈彭明珍女士驾鹤西去，请允许以此书追忆，厚恩犹在，音容长存，笑貌铭记，皆宛如昨日的曾经。

编　者

2015 年 3 月

目 录

第 1 章 数据仓库和数据挖掘概述

最近几年，越来越多的企业专注于优化自己的业务，以求在市场竞争中获取更大更持久的竞争优势，而这些都离不开对企业现有数据的分析和利用。为了创造更多的价值，规避更多的风险，以及更好地优化业务，数据仓库的构建和数据挖掘应用的构建就被提上了日程。数据仓库和数据挖掘的企业级应用大体经历了 3 个阶段：传统数据仓库时代、动态数据仓库时代和数据中心时代，而数据中心又可以分为关系型数据中心、非关系型数据中心（基于 Hadoop 或企业内容管理）和混合型数据中心（大数据平台）。

1.1 概述

当前企业面临着业务不确定性和竞争日益提高，同时客户忠诚度却不断下降的问题，企业难以获得长期的优势，除非全面地利用分析能力。零散地利用分析技术难以帮助企业获得全面的洞察，并降低企业内不同业务部门间有效协作。越来越多的企业专注于全面利用分析技术，以便在市场竞争中获取更大更持久的竞争优势。例如，银行希望知道如何有效地识别信贷风险，发现欺诈和洗钱等不合法行为，更高效地进行交叉销售和提升销售；电信公司希望知道如何对市场业务发展和竞争环境进行精准分析，从而为市场决策提供深入的分析支撑，提升营销活动的精确性，提高客户满意度，培育新的商务模式等；保险公司希望知道哪些理赔客户骗保的可能性更高，以及哪些客户是高价值低风险的客户群等。以上这些问题的解决都离不开企业数据的支撑和对现有数据的分析和利用（李德仁，王树良，李德毅，2013；程永，2013；Inmon，2005；Han，Kamber，Pei，2013；Executive Office of the President，2014）。

1.1.1 数据仓库和数据挖掘的目标

通过总结多个行业不同客户的需求，发现构建数据仓库和应用数据挖掘具有一些共同的目标，具体包括以下几个方面。

（1）通过跨系统实现数据共享，解决信息孤岛问题，提升数据质量；

（2）构建企业信息单一视图，实现结构化、半结构化和非结构化数据的统一管理和洞察；

（3）提供完善的业务模型挖掘、定义和管理，并在此基础上提供实时决策支持；

（4）提供准确有效的客户特征管理机制，为客户细分、销售提升、交叉销售、市场营销和客户维护挽留等提供深入洞察；

（5）构建企业级数据仓库、主数据管理、企业内容管理和大数据管理等，为企业提供统一的

数据服务；

（6）构建完整统一的元数据管理体系，制定完善的元数据管理策略，为企业提供统一高效的元数据管理服务；

（7）构建数据治理体系，保证数据的一致性，解决信息的冗余、冲突和缺失等问题；

（8）提供高效、实时和准确的多维数据分析、报表统计、即时查询、多媒体分析、流分析和内容分析等功能，为企业运营分析提供全面支持；

（9）提供简洁易用的数据挖掘和预测分析支撑，为企业分析提供全面支持；

（10）提供协同工作、规则引擎和事件处理功能，为基于全面分析能力的各种应用间有效协作提供支撑；

（11）提供完善的 IT 安全管理、综合监控和企业资产管理等。

除了以上目标，在构建数据仓库和实施数据挖掘过程中，同时也面临着各种挑战，例如如何构建企业面向数据文化，如何打破组织壁垒，如何控制整合项目的实施周期和风险，如何克服整合在技术上的复杂度等。

1.1.2 数据仓库与数据挖掘的发展历程

在“应用议程”时代，企业构建了众多的业务系统，为了满足市场竞争、企业管理和监管的需要，企业开始构建众多的报表查询系统。随着时间的推移，这些报表查询系统越来越不能满足企业的需求。例如，查询访问性能比较慢，报表统计相对固定，难以满足企业灵活的业务需求，无法进行多维分析等。于是一些企业开始尝试使用传统数据仓库技术进行商务智能（Business Intelligence，BI）系统的构建，也就是使用 ETL（Extract，Transform，Load）或 ETCL（Extract，Transform，Clean，Load）工具实现数据的导出、转换、清洗和装入工作，使用操作型数据存储（Operational Data Store，ODS）存储明细数据，使用数据集市和数据仓库技术实现面向主题的历史数据存储，使用多维分析工具进行前端展现，以及使用数据仓库工具提供的挖掘引擎或基于单独的数据挖掘工具进行预测分析等。相比之前的报表查询系统，传统数据仓库技术具有以下优点：

（1）通过完善的数据清洗转换保证了操作型数据的准确性和一致性；

（2）通过数据仓库技术提升了 BI 系统的性能；

（3）通过多维分析展现工具，给客户提供了全面的多维分析、报表统计和即席查询等功能；

（4）通过数据挖掘技术，帮助客户灵活地进行预测分析。

后来，传统数据仓库技术在越来越多的行业得到了大规模的应用，对提升企业运营效率，提高企业竞争力以及降低风险等起到了非常大的作用。同时，随着业务的发展，使用传统数据仓库技术的企业又开始面临新的问题，例如：

（1）随着竞争的进一步加剧，企业需要对市场变化及时进行响应，对数据仓库时效性的要求越来越高，而传统数据仓库中的数据都是批量定期更新的，难以满足时效性的要求；

（2）越来越多的一线用户需要使用数据仓库，而传统数据仓库用户通常只针对高端管理层或少数管理人员，许多一线用户无法访问数据仓库，例如银行，就有成千上万的客户经理和客户代表期望访问数据仓库；

（3）业务系统越来越需要传统数据仓库主动提供相应的分析能力，而传统数据仓库通常不会主动推送分析能力。

于是企业开始使用动态数据仓库技术（见图 1-1）解决上述问题，相比传统数据仓库技术，

动态数据仓库具有以下优点。

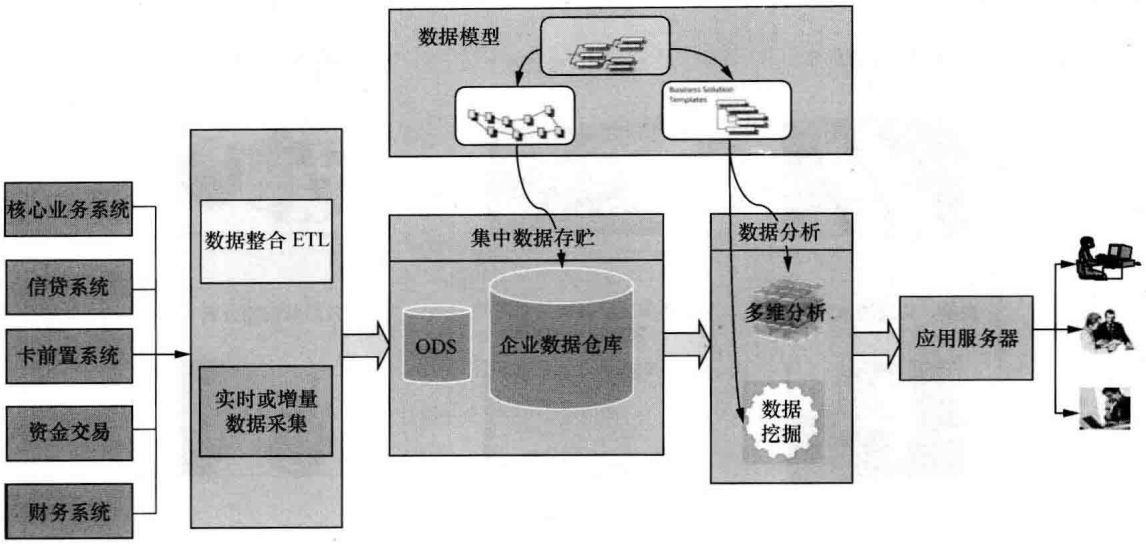


图 1-1　动态数据仓库参考架构

（1）一线用户可以动态（或者说实时）地访问数据仓库，以便获取其所需的信息；

（2）使用动态数据加载方式。相比传统数据仓库采用批量形式加载数据，动态数据仓库通常以准实时的方式连续加载数据（以增量数据加载为主），最低可以到秒级的时间间隔，从而在根本上保证数据仓库数据的实时性；

（3）采用事件驱动和主动推送的方式为业务系统提供分析能力，例如银行的信贷风险管理员，当审批某人的贷款请求时，关于该申请人的相关风险评级等信息就会被主动推送过来。

动态数据仓库技术给企业带来了实时分析能力，对提升企业竞争力作出了非常大的贡献，为了全面地提供分析能力，并让这些分析能力自动得到应用，就需要将其嵌入多的业务流程中。于是，很多企业开始在传统数据仓库和动态数据仓库基础上构建数据中心。通过数据中心的构建，企业从传统的交易系统（记录系统）和各种差分系统（Different System）逐渐转向构建创新系统，企业将分析技术慢慢融入其核心战略制定和日常运营管理中（见图 1-2）。数据和分析技术不再仅仅用来管理财务预测、年度预算分配、供应链优化和流线型运营，而是更多地用来增强客户体验，制定更智慧的营销活动，评估员工的绩效和制定组织的战略目标等（程永，2013）。

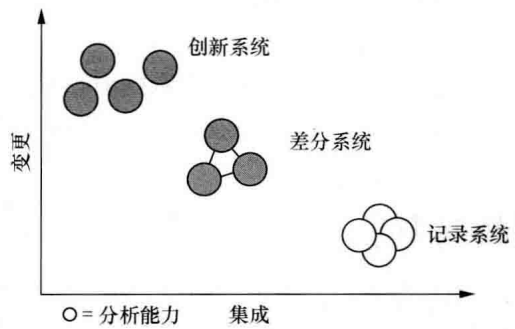


图 1-2　构建新系统示意

通过构建新一代数据中心，可以在各行各业实现智慧的分析洞察（见图 1-3），例如在交通行

业进行实时交通流优化、公交线路优化、基于交通流量预测进行出行线路推荐等；在银行业进行反欺诈、反洗钱和风险管理整合等；以及在零售行业预测客户购买意向等（程永，2013）。

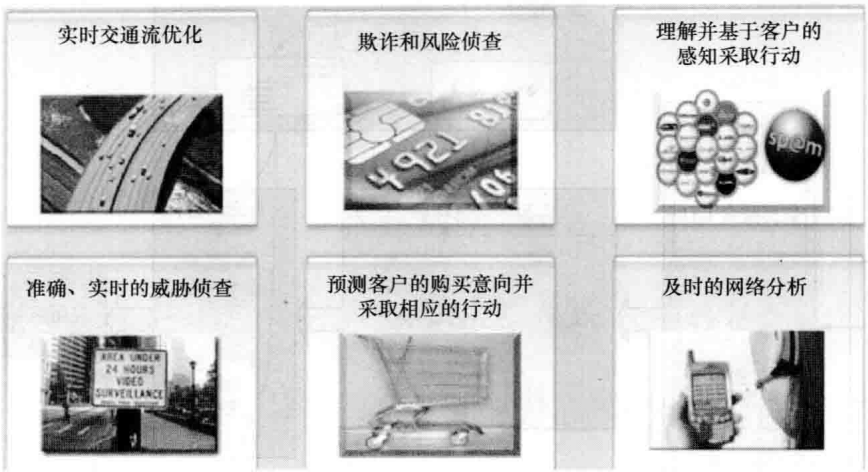


图 1-3　各行业分析示例

1.2　数据中心

根据处理数据类型的差异，数据中心可以分为 3 种类型：关系型数据中心、非关系型数据中心（基于 Hadoop 或企业内容管理）和混合型数据中心（大数据平台）。通过构建数据中心，特别是混合型数据中心（大数据平台），企业将实现贯穿整个信息供应链的所有信息的单一视图。同时在各种运营分析和预测分析的帮助下，企业管理层可以在全局层面拥有对业务的可见性，获得深入的洞察力（李德仁，王树良，李德毅，2013；程永，2013；　Meyer-Schoenberger，Cukier，2013）。

1.2.1　关系型数据中心

如图 1-4 所示，关系型数据中心通常以数据仓库或关系型数据库为基础构建数据存储层，数据以关系型数据为主（结合少量半结构化和非结构化数据），实现企业全部或部分结构化数据的物理或逻辑集中，通过完善的数据治理和统一元数据管理构建企业信息单一视图（关系型），并提供全面的分析能力，为企业快速决策、风险管理以及个性化服务提供支持，帮助企业优化业务流程，构建创新型应用。

如图 1-5 所示，与动态数据仓库技术相比，关系型数据中心不再局限于在现有应用程序的基础上提供有限的分析能力，而是基于企业信息单一视图提供全面的分析能力，并在分析能力之上全面构建创新型应用。

1.2.2　非关系型数据中心

在数据仓库以及后来的关系型数据中心发展的同时，企业也从未放弃对大量非结构化数据（全世界 80%的信息资产是非结构化的）的管理。开始的时候多采用企业内容管理的方式将大量的音频、视频、图像、文本和电子扫描件等非结构化数据进行管理，并通过内容分析获取对非结构化数据的洞察。

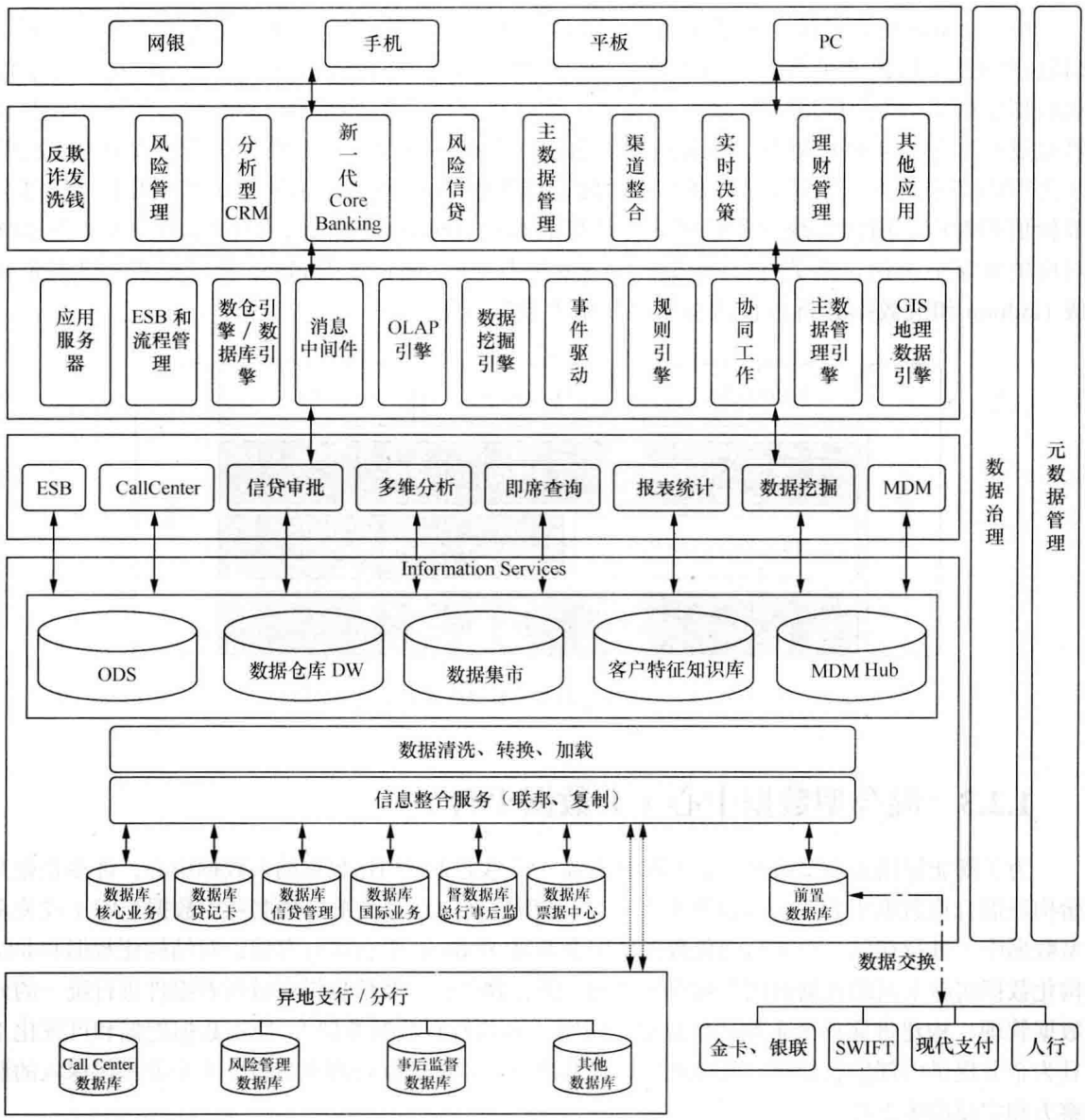


图 1-4　关系型数据中心体系结构实例

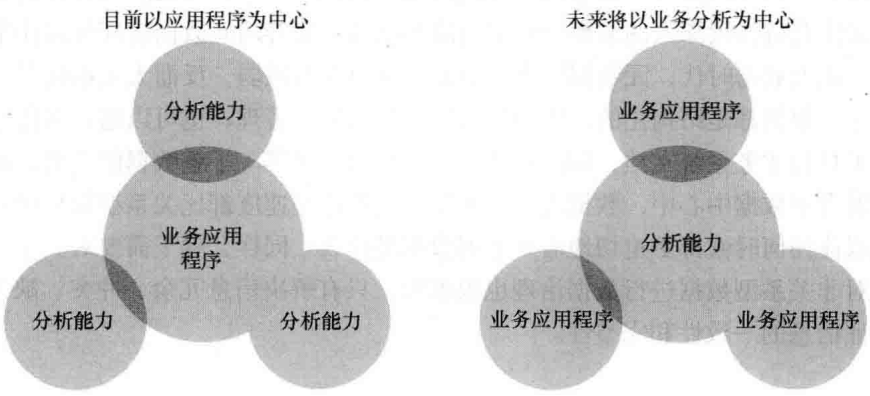


图 1-5　数据中心场景

随着 Hadoop 的兴起，越来越多的企业倾向于将海量的“每字节已知价值（指每字节数据包含的已知价值或可被量化的价值，是一种相对描述）”较低的原始、半结构化和非结构化数据使用企业级 Hadoop 平台进行存储、管理和分析。企业级 Hadoop 可以跨廉价机器和磁盘进行大规模扩展以处理大数据问题，通过内置在环境中的冗余性，数据冗余地存储在整个集群内多个地方（默认存储三份）。因此，企业级 Hadoop 平台的“每计算成本（与每字节价值相对应，通常传统业务系统的每计算成本比 Hadoop 系统高，但同样的传统业务系统对应的每字节价值也高）”同样较低。图 1-6 是典型的 Apache Hadoop 开源框架，更多企业级 Hadoop 和流数据分析内容请参见第 9 章大数据管理。

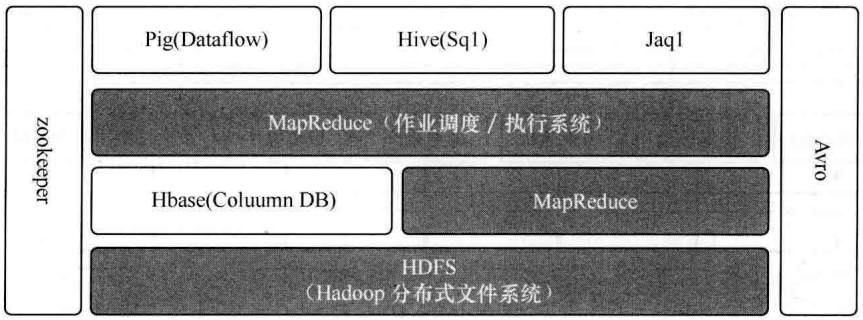


图 1-6　基于 Hadoop 的非关系型数据中心简单示例

1.2.3　混合型数据中心（大数据平台）

为了既能使用关系型数据（仓）库的高效，又想把非结构化数据纳入数据中心，许多企业开始构建混合型数据中心（即大数据平台）。混合型数据中心对关系型数据使用数据仓库（或关系型数据库）进行存储，对非结构化数据使用企业级 Hadoop 平台进行存储，对结构化数据和非结构化数据实施全面的数据治理，对业务规则、流程和逻辑以及信息供应链所有组件进行统一的元数据管理，构建贯穿整个企业的信息单一视图（涵盖所有数据类型），使用数据挖掘和可视化工具为企业提供内部信息导航，通过流计算工具进行实时地数据处理和分析，为企业提供深入的洞察力和实时决策支持。

在构建混合型数据中心的过程中，大数据的管理不应该仅仅侧重存储、分析和可视化，更应该注重元数据管理和数据治理。如果没有合理的元数据管理，企业将无法从大数据分析中获取有效信息，更无法获得持续深入的洞察力和实时决策支持。完善的元数据管理可以让企业数据更加完整和准确，在大数据时代，元数据管理的重要性非但没有减弱，反而大大增强了。在构建关系型数据中心时，数据都是结构化的，即便没有完整的元数据管理，也可以通过多种方法（使用数据探索分析工具和业务系统文档，询问业务人员和技术人员等）了解数据的含义，而在非关系型数据中心和混合型数据中心中，数据无论是容量、类型还是速度都比关系型数据中心大（快）得多，企业比以往任何时候都更迫切地需要了解数据是什么。同样，除了需要对关系型数据进行数据治理外，对非关系型数据进行数据治理也很重要，只有解决信息冗余、冲突、缺失和错误等问题，才能保证信息的一致性和完整性。

1.3　混合型数据中心参考架构

如图 1-7 所示，以构建银行新一代数据中心为例，假设银行上下级之间存在三级结构，分别是总行、异地分行/支行和支行/网点。其中核心账务系统、贷记卡系统、信贷审批管理系统、国际业务和票据中心等数据存放在总行数据中心，异地支行/分行也会有部分数据，例如呼叫中心、风险管理和事后监督等。总行和金卡、银联、SWIFT、国家现代化支付系统和人行等之间存在数据交互，数据中心可以为银行提供一个关于业务的全面准确的视图，帮助银行更加有效地进行反欺诈和反洗钱管理、风险管理、信贷管理、理财管理和实时决策等（李德仁，王树良，李德毅，2013；程永，2013；Meyer-Schoenberger，Cukier，2013；IBM 中国官方网站，2014；IBM 开发者园地，2014；DB2 V10.1 信息中心，2014）。

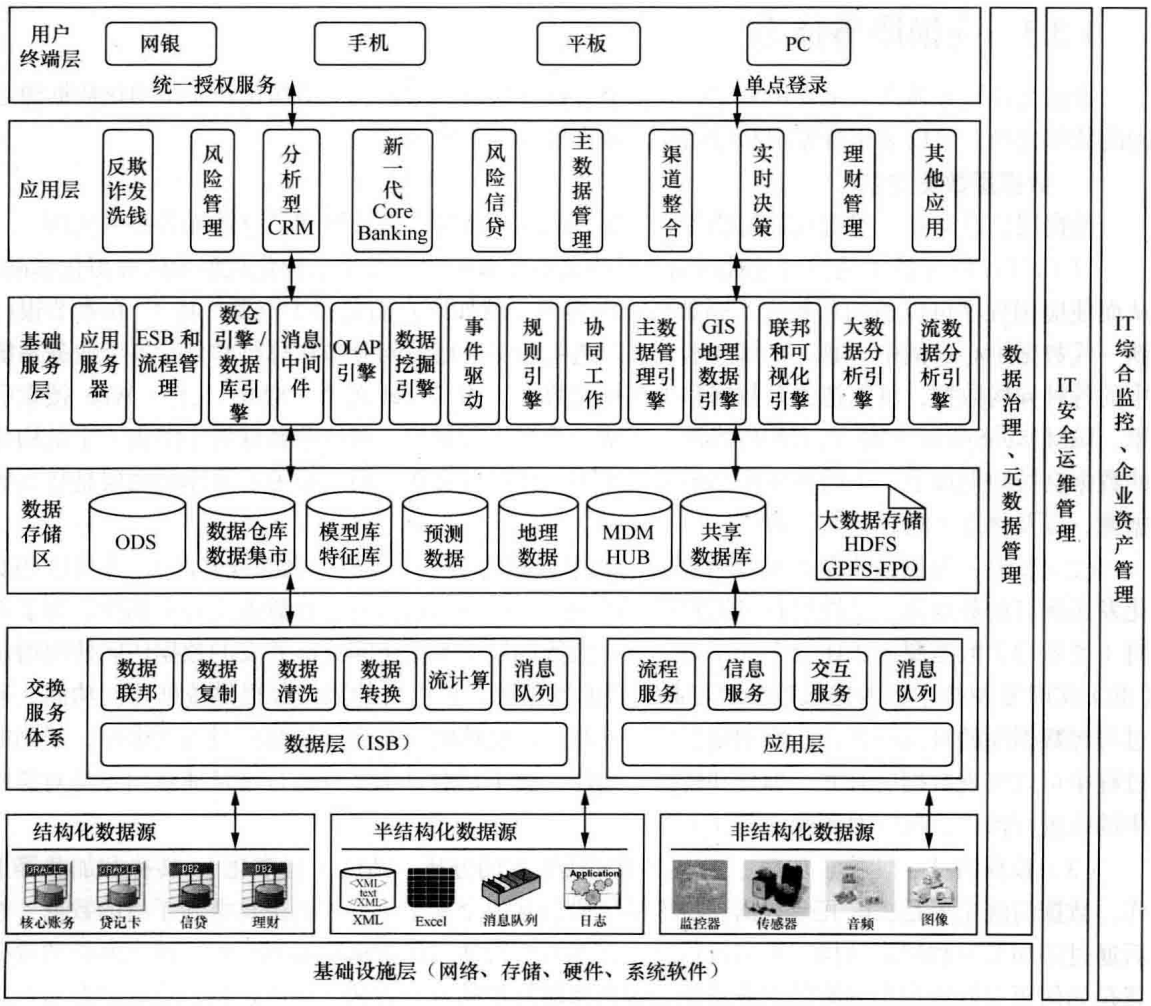


图 1-7　数据中心整体体系结构

银行新一代数据中心整体分为十层，分别为基础设施层、数据源层、交换服务体系、数据存储区、基础服务层、应用层、用户终端层、数据治理、元数据管理层、IT 安全运维管理和 IT 综合监控、企业资产管理等。

1.3.1 基础设施层

基础设施层主要包括整个企业所涉及的硬件、系统软件、网络设备和各种存储等，实现的方式可以基于企业私有云的方式，也可以基于公有云的方式，从而实现自动化、虚拟化和标准化管理等。

1.3.2 数据源层

数据源层主要包括结构化、半结构化和非结构化数据源。

（1）结构化数据源主要指各种关系型数据库，例如 DB2、Oracle 和 MS Sql Server 等。

（2）半结构化数据源主要指各种包含半结构化数据（如 XML、Excel、文本和日志等）的数据源。

（3）非结构化数据源主要指包含如图像、音频、视频和扫描件等非结构化数据的数据源。

1.3.3 交换服务体系

数据交换服务体系层主要用来完成数据中心存储层与结构化、半结构化和非结构化数据源之间的数据交换，可以采用数据层和应用层两种实现方式来实现。

1. 数据层数据交换

数据层信息交互主要通过数据联邦、复制、清洗、转换、流计算和消息传输等技术实现。

（1）联邦：联邦技术是指通过对同构或异构关系型数据源以及半结构化数据源的虚拟化基础，从而使应用程序可以访问和集成不同数据和内容源（就如同它们是单个资源一样）。在本节银行新一代数据中心示例中，通过联邦技术，可以透明和实时地访问分布在总行和分行各个业务系统中的各种异构数据，可以把关系数据和半结构化数据（如 Excel 文件、XML 文件、Web 搜索引擎、IBM WebSphere MQ 查询和内容源）组成一个逻辑数据库，对这些数据源中的表（半结构化的数据会被映射成表）可以像操作本地数据库表一样进行操作，而不必关心操作的底层是什么数据源，以及物理上处在什么位置等。

（2）复制：复制技术是指捕获数据源端表中数据的变化（增加、删除或修改），并将这些变化发送到目的数据源，最终将其应用到相应的表中。在本节银行新一代数据中心示例中，为了不对（关系型）数据源造成比较大的压力，针对大数据量的数据访问或高并发的数据访问使用增量（准）实时复制技术将数据从数据源复制到目的数据源，再由目的数据源提供数据访问功能。通过对源数据库的日志捕获，获取增量数据，并基于消息传输机制将其复制到目的数据库，复制的过程中可以实现数据的合并、拆分和转换等操作。由于是对日志文件进行增量捕获而不是对源库中的表进行操作，所以对源库业务压力不大。

（3）数据清洗、转换、加载：主要用来完成数据的分析、清洗（标准化）、转换和加载等工作。数据清洗主要是去除冗余数据，将零散字段合并成全局记录，并解决重叠和矛盾的数据，然后通过添加关系和层次结构完善丰富信息。在本节银行新一代数据中心示例中，首先面临的挑战就是如何更有效地识别现有的业务系统，包括系统使用的分类方法、层次结构、数据分布和数据字典等。如果数据字典不完整或缺失，就要通过工具（例如 InfoSphere Discovery）找出其数据的存储结构以及各个表之间的主外键关联、各表之间的转换关系等。数据的分布情况同样可以使用工具来完成（例如 InfoSphere Information Analyzer）。在对现有数据足够了解的基础上（完成了数据分析），制定数据的清洗规则以及转换规则，其中，清洗规则又分为两种情况：①清洗规

则是明确的；②清洗规则是模糊的。清晰的数据清洗规则比较好处理，直接制定数据转换规则并借助数据转化工具（例如 InfoSphere　DataStage）进行转换即可。针对模糊的清洗规则就需要使用数据概率算法匹配重复的数据记录，并自动地将数据转换为经过检验的标准格式（可以借助 InfoSphere　QualityStage 来实现），消除数据源中的重复内容，确保数据的一致性。例如在银行不同业务系统中都存有地址信息，某业务系统中某条客户记录地址信息为“北京市朝阳区北四环中路盘古大观 21 层”，另一业务系统中某条客户记录了地址信息为“北京市北四环（中）路盘古大厦 21 层”。通过手工方式，可以判断这两个地址实际上是同一个地址，但计算机会当成两个地址来处理，这时候就需要用到模糊匹配功能，具体处理请参见本书第 4 章 InfoSphere QualityStage 介绍部分。

（4）流计算：通过使用流计算工具（例如 IBM InfoSphere　Steams），可以实时处理、过滤和分析流数据，实现业务实时预警和事件处理等，同时可以将获取的高价值结构化分析结果存储到数据仓库中，以便进一步分析，还可以将大量有价值数据（去除噪声后的原始、半结构化和非结构化数据）存储到企业级 Hadoop 中，以便后续分析处理。

2. 应用层数据交换

应用层数据交换主要基于程序接口、适配器、ESB 总线和 Web Service 等多种技术实现。同时，数据层很多数据交换作业也可以发布为 Web Service，从而允许用户在应用层对其进行调用。在本节银行新一代数据中心示例中，可以使用 Websphere MB 和 MQ 构建支持各种协议和数据格式的企业服务总线，各系统可以通过服务使用企业服务总线进行交互。系统间的信息格式、传输协议和采用技术的差异，以及物理位置的不同等问题都将被企业服务总线屏蔽，还可以将服务按照业务流程的需要重新进行编排，以便满足业务的需要。

1.3.4　数据存储区

数据存储区是数据中心所有数据的集中（物理或逻辑）存放地，主要用来存放各种历史数据（结构化和非结构化数据）、预测数据和汇总数据等，客户特征库和模型库也在数据存储区，其他数据还有主数据管理集线器（MDM　Hub）相关的主数据，地理信息系统相关的地理数据以及需要共享的数据等。需要注意的是，数据存储区的结构是一种逻辑描述，客户实际部署时需要根据具体情况进行部署。

（1）操作型数据存储（Operational Data Store，ODS）：ODS 存放了数据中心需要用到的业务明细数据，通常是一个可选项，用来在业务系统和数据仓库之间形成一个隔离层。ODS 中的明细数据将被进一步加工、集成和汇总，并加载到数据仓库或数据集市中。ODS 可以用来提供部分业务系统查询的功能（转移业务系统压力），并转接一些数据仓库中不能完成的功能（例如偶尔对明细数据的访问需求，DW 层通常都是存储汇总过的数据，偶尔对细节数据的查询可以转移到 ODS 来完成）。ODS 中存放的明细数据通常和原业务系统中的数据略有差别，例如多个下属分支机构相同的数据会汇总到一张表中，某些代码会被转换成新的值，某些缺失的数据会被补齐，某些不一致的数据会被标准化等。

（2）数据仓库（Data Warehouse，DW）：数据仓库是用来存储面向主题的、集成的、相对稳定的和反映历史变化的数据，用于支持决策管理。而数据集市（DataMarts）则是为了特定的应用目的或应用范围，而从数据仓库中独立出来的一部分数据，也可称为部门数据或主题数据（subject area）。数据仓库建模一般使用星型模型或雪花模型，例如简单的星型模型包含一个事实表，并以事实表为中心关联多个维表；雪花模型是星型模型的扩展，就是一个或多个维表没有直接连接到

事实表上，而是通过其他维表连接到事实表，像雪花一样连接在一起。

（3）特征库和模型库：特征库主要用来存储经过客户分析生成的每个客户分群的群组特征，除了传统的客户属性如年龄、单位、工作年限和地理位置等，群组特征还包括客户选择商家的心理特征、购买商品的可能性以及客户对公司的累积价值等。基于客户特征库，企业可以将每个客户视为个体提供个性化的服务。模型库主要用来存放数据挖掘建模生成的业务模型，这些业务模型经过评估合格后，将被用来支持企业的各种业务流程和决策。

（4）预测数据：主要用来存放依据业务模型预测的各种数据，特别是无法明确描述出规则的业务模型，其预测的数据会直接存放在数据存储区。

（5）地理数据：通常会单独存储在一个空间数据库，别的应用程序可以通过接口或 Web Service 方式调用 GIS 地理数据引擎获得想要的地理数据。

（6）共享数据：存放需要进行共享的数据，可以和数据仓库或 ODS 等存储在一起，也可以分开存储。通常一些组织存在数据共享的需求，但又不想将数据仓库中的数据直接共享出来，而会单独部署一个数据库存储共享数据。

（7）MDM Hub：用来存储主数据，为主数据管理引擎提供数据支撑。

（8）大数据：大数据主要指各种原始、半结构化和非结构化数据，为大数据分析提供数据支持。大数据的存储常采用 HDFS 或 GPFS File Place Optimizer （GPFS-FPO）文件系统。

1.3.5 基础服务层

基础服务层主要包括构建应用所需的各种核心引擎，除了传统的应用服务器、关系型数据库（数据仓库）引擎、ESB 和流程整合引擎、消息中间件以外，还包括 OLAP 引擎、数据挖掘引擎、规则引擎、协同引擎、事件驱动、主数据管理引擎和 GIS 引擎等，另外还有针对原始、半结构化和非结构化数据的大数据分析引擎，针对所有的结构化、半结构化和非结构化数据的联邦和可视化工具以及针对流数据的流分析引擎。

（1）应用服务器：作为连接前端和后端应用程序的中间件层，即业务系统的应用服务器，可以帮助企业对诸如安全性、可扩展性、可管理性、成本、简易性等众多因素进行综合考虑，尤其适用于复杂的混合环境。

（2）企业服务总线：通过在企业架构中引入企业服务总线，减少了点到点的互连数目。使用与平台无关的企业服务总线集成多个应用程序、网络和设备类型，从而让用户安全可靠地处理业务。为客户应对快速变化的业务准备了灵活的 IT 技术框架，提高了用户业务的灵活度。

（3）数据仓库/数据库引擎：主要用来为存储在数据仓库中的海量历史数据提供多维分析以及数据检索、聚合、压缩、建模和更新等。

（4）业务流程管理（Business Process Management，BPM）引擎：BPM 提供了一个平台，支持嵌入在应用程序和系统中的业务功能，在高于应用程序间集成以及数据集成的级别进行交互和集成。

（5）消息中间件：提供准确、安全、可靠的消息传递，保证客户的数据不会遗失或重复，支持异步处理和并行消息传递，可以通过集群来完成横向或纵向的扩展，提供一致的编程接口，保证应用独立于网络或系统故障，支持文件传输。

（6）联机分析处理（On-Line Analytical Processing，OLAP）引擎：以多维度方式分析数据，弹性地提供上钻（Roll-up）、下钻（Drill-down）、切片和切块（Slice & Dice）、旋转（Pivot）等操作，呈现整合性决策信息（运营分析）。OLAP 主要用于大规模数据分析和统计计算，为决策提