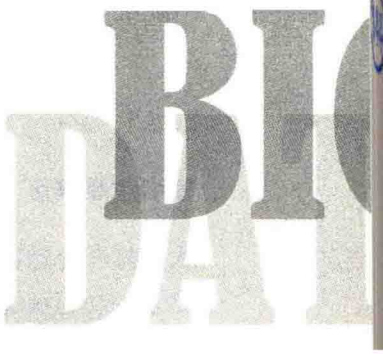
大数据在金融行业
实用案例剖析

刘世平　主编

中国财经出版传媒集团
经济科学出版社
Economic Science Press



大数据在金融行业实用案例剖析

刘世平　主编

中国财经出版传媒集团
经济科学出版社
Economic Science Press

图书在版编目（CIP）数据

大数据在金融行业实用案例剖析/刘世平主编.
—北京：经济科学出版社，2016.8
ISBN 978-7-5141-7207-2

Ⅰ.①大… Ⅱ.①刘… Ⅲ.①数据处理-应用-金融业 Ⅳ.①F83-39

中国版本图书馆CIP数据核字（2016）第206616号

责任编辑：于海汛 陈 晨
责任校对：杨 海
版式设计：齐 杰
责任印制：李 鹏

大数据在金融行业实用案例剖析
刘世平 主 编
经济科学出版社出版、发行 新华书店经销
社址：北京市海淀区阜成路甲28号 邮编：100142
总编部电话：010-88191217 发行部电话：010-88191522
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：http://jjkxcbs.tmall.com
北京季蜂印刷有限公司印装
787×1092 16开 17印张 330000字
2016年8月第1版 2016年8月第1次印刷
ISBN 978-7-5141-7207-2 定价：50.00元

大数据在金融行业实用案例剖析

主　　编　刘世平

编写人员　白　硕　刘铁斌　陈道斌　周衡昌　刘　斌　赵睿斌
陈　云　沈显克　申志华　廖祥文　田立中　姚玉辉
邱华勇　刘　捷　李华明　宋　丹　俞　立　李欣苗
刘可伋　客户关联关系研究项目组

前　言

非常高兴，这本《大数据在金融行业实用案例剖析》终于与大家见面了！非常感谢中国财经出版传媒集团副总经理、经济科学出版社社长兼总编辑吕萍同志和出版社的其他同事对本书出版的大力支持！

大数据技术发展至今，已经风靡全球。推动大数据在各个行业的应用已经成为很多国家发展的重要战略。在我国，尤其是在2015年，国务院、发改委和工信部等陆续颁布了很多重要文件，大力支持大数据产业。

20世纪80年代中期至今是大数据应用发展的重要历程。大数据应用从技术本身来看是逐渐发展、循序渐进的过程。从最早的数据仓库（Data Warehouse）、数据挖掘（Data Mining）、商业智能（Business Intelligence），一直发展到现在的大数据（Big Data），形成一个完整的体系概念，整个过程历时达30年之久。

准确来讲，美国第一次出现数据仓库的概念是在20世纪80年代中期。数据仓库当时提出的核心思想就是把企业内部分散在各个不同地方的数据进行有效整合，然后通过一系列方法提取数据中有用的信息，结合各个行业的知识，然后把这些有价值的信息应用到决策过程中。

到20世纪90年代中期，我的老东家IBM公司，系统的提出了数据挖掘的概念，同时IBM公司推出了一款叫智能挖掘（Intelligent Miner）的数据挖掘软件，这一软件的出现系统地推动了数据挖掘在各个行业的应用。

20世纪90年代末，出现了商业智能（Business Intelligence）这一新的概念。某种意义上来讲，商业智能的核心技术是数据仓库和数据

挖掘技术的整合，也就是把数据进行有效的整合，通过一系列的方法，提取数据中有用的信息，结合行业知识用于决策的过程中。商业智能的概念准确的翻译应该是“商务智能”，因为它的应用不仅包括企业单位等营利性组织，也包括政府、非政府组织（NGO）等非营利性机构。

2011 年，麦肯锡提出了大数据（Big Data）的概念，主要强调了大数据对未来企业的竞争、创新、生产效率等的提升将会起到重要作用。我对大数据的理解是更加系统全面的电子化和全社会范围的互联互通。电子化指的是思维模式，这个模式让我们在决策过程中更加依赖于客观数据做出更有效的决策。另外很重要的一点是，大数据在社会范围内的互联互通，其中互联网和移动设备在互联互通方面起到了非常重要的作用。当然，随着技术的进步和发展，物联网在互联互通方面也将会发挥更大的作用。

大数据发展到今天，与技术的发展密切相关，最关键的就是云计算、互联网、存储技术和数据处理与分析能力的提升。一方面，存储设备本身的价格降低、存储设备容量的扩展、数据分析和处理能力的提升，对大数据的发展有着巨大的推动作用。另一方面，移动技术的发展，移动互联网、云计算的发展，对数据的收集和整理起了很大的推动作用。

大数据目前已广泛应用到各个行业，包括金融、电信、制造业、医疗、能源、零售业、餐饮业以及政府的税务管理、海关管理以及精准扶贫等。大数据在企业的应用主要有几个不同的阶段，首先是从 ERP 到 CRM，再到互联网以及今天的大数据应用，这同时也是企业利用数据做出决策的逐步深化的过程。从技术角度来讲，大数据最主要的应用是利用现代的数据挖掘和数据分析技术，从海量的信息中获取有用信息用于决策过程中。获取信息的方法可以归纳为四类：查询；统计和报表分析；多维分析；数据挖掘。多维分析和数据挖掘是大数据技术今天得到广泛关注和提升的非常关键的部分。

本书讨论的重点不是大数据技术本身（包括数据库、数据的存储、数据架构、数据仓库、数据挖掘技术等）。本书主要讨论的是如何利用大数据技术去解决我们在实际工作中遇到的问题，尤其是金融行业中

遇到的各种问题，比如风险管理、市场营销、精准决策、精准营销等。本书的初衷就是让读者通过这些经典案例，去体会大数据是如何解决企业所面临的实际问题，给读者一个直观而又务实的体验，同时可以模拟这些方法解决遇到的具体问题，起到“依葫芦画瓢”的示范作用。

本书主要是基于第三届“全国金融大数据战略与应用研讨会”主要嘉宾发言基础上的提炼与提升，本书案例中所进行的大数据分析的相关数据均由各章编写人员收集整理而来，数据来源真实可靠。在此，我要感谢太原市委和市政府对前三届“全国金融业大数据战略与应用研讨会”给予的支持与帮助，也要感谢我在中科院大学的领导——原中国科学院大学党委书记邓勇博士和副校长王颖博士对会议的大力支持！还要感谢我们吉贝克的同仁们，他们在会议的筹划和组织方面付出了很多辛勤的劳动，谢谢大家！最后我要感谢我的家人，谢谢他们对我的工作的无限理解和支持。

本书共分15章，主要是大数据在金融行业的经典应用案例。作为本书的主编，在此我谨向本书的编写人员致以诚挚的谢意，他们分别是来自国家开发银行、中国工商银行、平安银行、平安人寿、上海农商银行、上海财经大学、福州大学、兴业证券、上海证券交易所、温州市金融局、国家信息中心、中国智慧城市发展研究中心、国家千人计划等的专家、金融企业高管、政府部门负责人等。编写人员包括李华明、客户关联关系研究项目组、申志华、周衡昌、田立中、刘铁斌、陈道斌、宋丹、陈云、俞立、李欣苗、刘可伋、刘斌、邱华勇、廖祥文、白硕、沈显克、刘捷、姚玉辉和赵睿斌。感谢他们的努力和付出！

刘世平

2016年7月于上海

目　录

第 1 章

大数据在金融行业的应用

——客户综合管理

刘世平　李华明*

一、什么是大数据?

从2011年6月，麦肯锡咨询公司发布研究报告《大数据：下一个竞争、创新和生产力的前沿领域》并首次系统阐述大数据的概念以来，“大数据（Big Data）”一词已经无处不在。然而，时至今日大数据的概念仍然存在混淆，业界对大数据的争论似乎仍未停止，大数据被用于承载了各种名目众多的概念，典型的包括海量的数据、社交媒体分析、下一代数据管理能力、物联网数据、实时数据等。关于大数据讨论最多的声音是“到底多大才算大数据?”实际上对数据的量上并没有一个绝对的界限，数据量的大小，是相对于当时数据的存储能力、计算能力和分析能力而言的，20世纪90年代，10G存储量的数据在当时的存储和计算能力下可以称得上“大数据”，而在今天一部智能手机的存储量也远超过了10G。那么到底什么是大数据？笔者引用目前业界普遍接受的，来自于高德纳公司（Gartner）的定义，即大数据的特征具体涵盖了“4V”的内容。

（1）数据量庞大（Volume）：存储单位从PB扩展到ZB；IT系统、互联网、移动互联网、物联网等每天都在产生大量新生数据，过去的两年间产生的数据占到了所有数据的90%。

（2）产生速度快（Velocity）：数据变化与处理的频度由天加速到秒/毫秒；订单、支付、欺诈、微博、监控视频、传感器、信令每时每刻都在不停地产生数据。

---

* 刘世平，博士，中国科学院大学金融科技中心主任，福州大学特聘教授，淮南师范学院吉贝克大数据学院名誉院长，吉贝克信息技术（北京）有限公司董事长。研究方向：大数据、商业智能（数据仓库、数据挖掘）和风险管理。李华明，吉贝克信息技术有限公司证券行业及金融大数据分析服务线领导人。研究方向：数据治理、DW/BI及大数据分析的技术与应用。

（3）数据多样化（Variety）：数据种类繁多包括数据库表、格式文本、自然语言文本、电子表格、声音、图片、视频等。

（4）数据价值（Value）：通过对大数据的处理和分析，可以发掘出巨大的价值，包括商业价值和社会价值。

在上述定义中，前3个“V”是对数据本身特征的描述，那么是否数据量足够大，数据产生的速度足够快，数据类型足够多就是大数据了呢？笔者认为，大数据核心的本质价值在于第四个“V”，即数据价值的实现，也就是说当海量的、实时的各种不同类型的数据产生后，只有通过大数据的处理和分析，发掘出数据背后隐藏的本质规律并将其应用于商业决策过程，并进一步推动商业模式的变革，大数据才真正得以诞生，数据价值是成就大数据的关键所在。

二、在金融行业的应用

金融行业由于自身行业特性，在大数据的应用方面具备得天独厚的优势，在长期的业务开展过程中金融企业积累了海量的高价值数据。以数据强度居于众多行业之首的银行业为例，据统计银行每创造100万美元的收入，平均就会产生820G存储量的数据。面对如此种类繁多、数据巨大、不断产生的数据，金融企业应该如何分类、整合、存储、分析和应用，是摆在每一个金融企业面前的新问题。

1. 如何认清金融行业的海量数据，如何对这些数据进行归类，每类数据有什么特点

我们从两个维度来看，第一个维度是数据的结构化程度，根据数据结构化程度的不同，我们把企业的数据划分为结构化数据、半结构化数据和非结构化数据；第二个维度，我们从对数据的运用角度，把数据划分为静态数据和动态数据。我们把两个维度交叉，可以把企业的数据划分为三种不同类型。

（1）第一类是海量静态的结构化数据，像企业里面的企业资源计划（ERP）系统、客户关系（CRM）系统、人力资源管理（HR）系统、财务系统里面所存储的数据。

（2）第二类是海量静态的非结构化和半结构化数据，如企业里面的文本、报告、音频、视频、社交网络、邮件等。

（3）第三类是动态的海量数据，如企业里的数据网络点击率、日志文件，实时传感信息，实时路况信息，实时行情信息等。

金融机构所面临的三种类型海量数据，如图 1 - 1 所示。

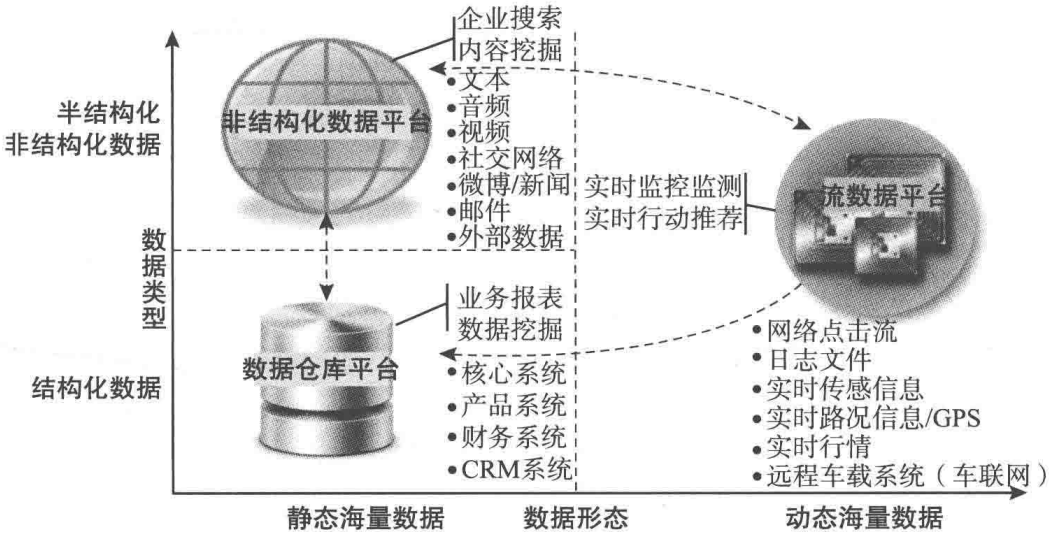


图 1 - 1　金融机构所面临的三种类型海量数据

2. 金融企业如何应用这三种不同类型的海量数据；三种不同类型的海量数据对分析手段和分析能力的要求有哪些不同

（1）对静态结构化数据，技术最为成熟，应用也比较广泛。通过建立企业级数据仓库（EDW）将企业范围内的结构化数据按主题进行整合集中存储，在上面可以依据业务单元、业务领域的不同建立数据集市（Data Mart），有了数据仓库和数据集市，就可以构建各种 BI 应用，包括标准化的报表满足业务和监管要求；通过即席查询（Ad - Hoc）和多维分析（OLAP）进行数据的钻取和多维度分析；通过数据挖掘，建立企业的预测能力等。

（2）对静态的结构化和非结构化数据，企业可以建立非结构化的数据平台，把文本、影像、微博、社交网络数据进行集中存储，依据企业内部维度，建立企业内部的 Google，称之为企业搜索。此外，可以对企业内部的非结构数据进行内容计算，内容计算涉及的技术主要包括自然语言处理，分词、句法的分析，关键实体识别、归类与索引、机器学习等。内容计算的典型应用包括舆情分析与基于内容的人际关系分析等。如吉贝克公司研发的“精准推荐产品”即可以对企业领导的讲话、新闻报道、内部工作手稿等资料进行文本挖掘分析，通过分词、句法分析，关键实体识别等文本分析技术提炼出能够反映企业决策层重点关注的“关键术语”；然后根据这些“关键术语”，利用爬虫技术，从国内外专业财经网站爬取与关键术语密切相关的新闻和文章，及时推送给企业领导，以帮助其有效地观察社会状态，并辅助决策，及时发现预警信息。

（3）动态的海量数据，通常也称为“流数据”。对流数据的分析处理方式主要是流计算。流计算从传感器，网络日志，网上点击率实时采集数据，结合企业的业务规则，实时提取满足业务规则的数据，并依据这些数据进行实时地分析监控和业务干预。例如在银行的实时监控场景中，大数据模式可以应用于风险管理，包括信用卡诈骗、保险诈骗、证券交易诈骗、程序交易等，能够实时跟踪发现。

3. 从业务的角度，大数据在金融行业的应用主要体现在哪些方面

从业务的角度，大数据在金融行业的应用可以分为客户管理、风险管理和运营优化3个方面，如图1－2所示。大数据在金融行业的典型应用场景举例，如表1－1所示。

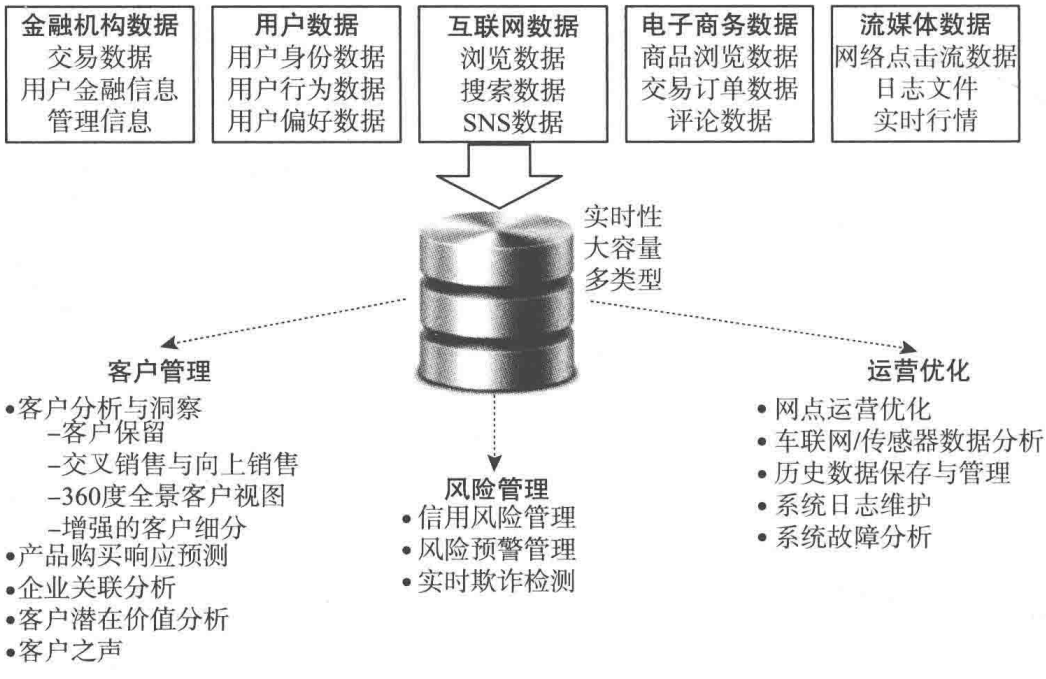


图1－2　大数据在金融行业的主要应用

表1－1　　大数据在金融行业的典型应用场景举例

	应用场景		描述
客户分析	客户分析与洞察	客户流失预测与客户保持	应用预测算法，建立客户流失预测模型，提前识别流失可能性高的客户，并预测造成客户可能流失的主要原因。金融企业依据客户流失预测结果，结合客户的发展潜力和贡献度，及早采取干预手段，进行客户保持

续表

	应用场景		描述
客户分析	客户分析与洞察	交叉销售与向上销售	基于金融企业现有的客户资源，通过对客户的行为（交易、偏好）、购买的产品、交易的时间以及购买产品间的相关性进行分析，借助产品关联分析成果向客户销售金融机构的其他产品
		360 度客户全景视图	除了企业本身的交易数据以外，利用外部的社交网络数据，通过客户的社交网络分析，对客户在社交网络上面做客户画像，通过社交网络、客户网站、微博、评论等等社交媒体信息，把客户在网络上的信息归类，包括客户评价讨论、客户的倾向性信息，还有客户情绪的信息，同时考虑行为的数据，客户的属性，对所有的数据进行整合，形成客户 360 度全景视图
		增强的客户细分	根据客户属性划分的客户集合，即将客户分为具有不同需求和特征的若干人群，并对不同的人群给予不同的产品组合和市场供给，以达到区别营销目标的过程；客户细分的工作步骤为首先对能反映客户特征的内部、外部数据进行收集和分析，在此基础上，根据不同业务应用目的选取易获取、质量高、可维护的数据作为细分维度，设计客户细分模型，根据模型对客户进行细分，通过细分结果验证和修订细分变量，优化分群模型，形成最终的客户细分结果
	产品购买响应预测		客户购买响应预测通过分析客户在金融机构的相关历史纪录、客户的基本信息和其他相关的信息，建立预测产品营销客户可能响应状况的模型，依据响应的可能性为每个客户进行评分，预测客户接受银行主动营销某种产品的概率，找出最有可能响应特定产品的客户群组。通过客户购买响应模型，金融机构可以选择反应概率超过一定程度的客户为营销对象，从而提高营销成功率，降低营销成本
	企业关联分析		从工商局，交易所，以及证监会、银监会、安全部门、公安部门发布的监管文件，新闻、出版物、社交媒体数据抓取企业关系、交易对手风险暴露以及风险事件信息，全面刻画企业的社交网络图并应用于企业的风险管理、营销等不同业务领域
	客户价值潜力分析		选取客户关键社会属性进行客户分群以及客户圈子分析，结合客户历史交易数据，预测客户潜在价值，计算客户价值提升指数
	客户之声		将客户在微博上的言论，在呼叫中心（Call Center）的通话记录，在网上商城对产品的评论等信息进行抓取和集成，通过语义分析，提取关键词，建立分析模型，识别出关键问题。应用客户之声，及早发现客户所抱怨的问题，及早介入进行舆论导向，发现新的产品诉求和新的销售机会，提升客户忠诚度

续表

	应用场景	描述
风险管理	信用风险管理	集成金融企业客户社交网络信息，移动互联网，基于位置的服务，O2O 等动态、海量、实时数据，结合传统上的宏观经济信息、市场信息、行业信息、客户信息、财务信息、历史交易信息等，从金融资产到社交资产，综合识别企业的风险因子
	风险预警管理	除金融企业内部信息以外，从外部非结构化数据（法院、税务、小贷公司黑名单）中提取有效信息，并根据信息组合将同一来源的外部数据细分到不同类别，从而进行预警、评分
	实时欺诈检测	以行为建模理论为基础，结合经验数据，明确反渗漏和欺诈的规则并确立模型
运营优化	网点运营优化	基于业务量的预测，综合考虑金融机构的业务量分布，内部工作人员资源现状，物理设施的限制等因素，最优化金融机构网点设置，窗口资源和人力资源排班，以达到降低运营成本，提升服务水平和提升员工满意度的目的
	车联网/传感器数据分析	通过对驾驶者总行驶里程、日行驶时间等数据，以及急刹车次数、急加速次数等驾驶行为在云端的分析，有效帮助保险公司全面了解驾驶者的驾驶习惯和驾驶行为，有利于保险公司发展优质客户，提供不同类型的保险产品
	历史数据保存与管理	应用大数据分布式数据存储，实现低成本存储金融机构海量历史数据，高效率的查询与应用历史数据
	系统日志维护	从金融机构各种源系统各种日志源上收集日志，存储到中央存储系统（可以是 NFS、分布式文件系统等）上，以便于进行集中统计分析处理
	系统故障分析	基于设备监控大数据分析，实现智能化故障原因分析、性能容量动态阈值分析、实时交易路由分析、业务交易实时跟踪、面向业务服务的全方位监控、量化的业务影响性分析、实时业务全景分析

三、应用案例

案例背景：某证券股份有限公司拥有遍布全国的 220 家营业部，业务经营涉