EXPLORER OF FINANCE

BIG DATA MINING AND APPLICATION IN COMMERCIAL BANKS

商业银行大数据挖掘与应用

孙杨
茆家铭
陈惠民 / 著

致敬与发现

SALUTE&DISCOVERY

经济管理出版社
ECONOMY&MANAGEMENT PUBLISHING HOUSE

金融探索者文库

EXPLORER OF FINANCE

BIG DATA MINING AND APPLICATION IN COMMERCIAL BANKS

商业银行大数据挖掘与应用

孙杨 苗家铭 陈惠民 /著

经济管理出版社
ECONOMY&MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

商业银行大数据挖掘与应用/孙杨，苗家铭，陈惠民著．—北京：经济管理出版社，2018.12

ISBN 978-7-5096-5811-6

Ⅰ.①商… Ⅱ.①孙… ②苗… ③陈… Ⅲ.①数据处理—应用—商业银行—经营管理—高等学校—教材 Ⅳ.①F830.33-39

中国版本图书馆 CIP 数据核字(2018)第 294110 号

组稿编辑：宋 娜
责任编辑：张 昕 杜奕彤
责任印制：黄章平
责任校对：陈 颖

出版发行：经济管理出版社
（北京市海淀区北蜂窝 8 号中雅大厦 A 座 11 层 100038）
网 址：www.E-mp.com.cn
电 话：（010）51915602
印 刷：三河市延风印装有限公司
经 销：新华书店
开 本：720mm×1000mm/16
印 张：13.5
字 数：208 千字
版 次：2019 年 9 月第 1 版 2019 年 9 月第 1 次印刷
书 号：ISBN 978-7-5096-5811-6
定 价：98.00 元

前 言

众所周知，始于20世纪60年代的第五次信息技术革命，使大数据浪潮席卷全球。短短数年，大数据的概念和价值就得到了政府和社会的普遍认可，大数据的应用也遍及社会生活的方方面面。金融行业作为大数据挖掘和应用的行业先锋，在数据治理和应用方面都发挥着引领作用，而商业银行更是金融行业里的排头兵。笔者有幸在2013年下半年参与某商业银行总行层面以“EAST数据报送及应用”为主题的项目，亲身感知了商业银行大数据挖掘和应用的探索和成长之路。

在全球已进入信息化高速发展的时代背景下，商业银行在业务开展过程中每天都会产生海量的数据，如客户数据、业务交易数据、内部管理数据、外部数据、系统日志等。然而这些海量的数据对于商业银行而言，就像是一枚硬币的正反面，不单意味着海量的机会，也有可能成为巨大的包袱和负担。到底是机会还是负担，取决于商业银行是否具备一定的数据治理能力，是否有配套的系统支撑和人才储备等。

当前，数据资源决定未来商业银行核心竞争力的观点，已成为行业共识。但总体上看，商业银行普遍面临数据质量不高、数据治理能力不够、数据支持决策不足等问题，导致商业银行积累的存量数据如沉在海底的宝藏亟待挖掘和应用，以充分发挥其应有的价值。那么，如何开展数据治理和挖掘就成为商业银行信息化建设过程中需要破解的难题。《中国银行业信息科技“十二五”发展规划监管指导意见》指出：“商业银行要重点加强对数据治理的制度建设和流程建设，建立和完善数据治理制度体系，规范工作流程，理顺内部协作关系，提升数据质量和数据应用水平，提高数据价值创造能力。”

回顾项目历程，在银监会要求各商业银行按时向其报送符合监管

EAST 标准数据的大背景下，该商业银行投入大量人力、时间和资金，按照银监会监管 EAST 数据标准，对行内相关数据结构、数据字段、数据名称等一系列内容进行整合对接。基于这项常规报送标准化数据的工作现实，该商业银行决心以 EAST 标准化数据报送为契机，紧紧抓住行业提升数据治理水平和挖掘与应用能力的历史机遇，成立由科技、风险等部门和特邀行外专家组成的项目组来推动落实本行大数据平台系统建设。该大数据平台系统建设的目标定位为，通过接入行内标准化底层基础数据，为管理者和平台使用者提供界面友好、操作便捷的数据驾驶舱，实现对行内数据多层次、多维度、多视角、全方位的深度挖掘和应用。目前，该大数据平台建设已取得阶段性成果，使用者可轻松实现行内数据字段的任意拖拽、数据可视化、数据指标预警分析、模型代入检验等，发挥了在行内数据挖掘和应用方面的功效和作用。

基于笔者的实际参与经历，本书从商业银行数据治理实践出发，综合分析当前商业银行面临的时代背景和发展现状，重点阐述商业银行在数据挖掘和应用方面的经验做法，希望可以给读者提供帮助和参考。本书主要内容如下：第一章介绍了当前大数据变革的时代背景，国内数据治理和数据人才培养现状。第二章提出了当前大数据时代面临的安全形势及需要关注的数据安全问题。第三章列举了美国、日本、欧盟在大数据方面发展的情况，以提供可学习和借鉴的经验。第四章以 3 个具有代表性的互联网金融产品为例，阐述当前国内互联网金融发展的行业现状和总体趋势。第五章阐述了互联网金融以跨界、众筹等形式与大数据融合的方式，给当前商业银行带来的机遇和挑战，以及商业银行已积极主动与互联网平台合作，抢占数据资源的现实。第六章描述了当前商业银行大数据发展前景和应用场景，并对商业银行大数据发展路径选择提出了相关建议。第七章通过案例分析了商业银行大数据平台建设的具体流程和效果。第八章从商业银行的行业监管部门角度，分析大数据平台系统的建设和应用，可为商业银行的数据治理和挖掘应用提供参考。第九章介绍了几种具体的数据挖掘技术和算法。第十章采用理论和实践应用相结合的方法，以案例分析的形式介绍了商业银行大数据挖掘技术的实际应用情况。

本书系孙杨（南京审计大学）、苗家铭（中国银保监会宿迁监管

分局）和陈惠民（南京银行总行）三人共同完成。孙杨负责全书的总纂，包括总体构思及研究框架的设计。苗家铭负责具体落实，对实际撰写工作起了至关重要的作用。陈惠民在大数据治理方面具有较高的理论水平和卓越的实践经验，系国内著名的大数据治理培训专家，主要负责对团队的研究指导及终稿的审定工作。

参与本书资料收集、数据处理等撰写工作的团队成员还有宿迁学院的戴佳俊老师、江苏银行的王嘉申先生，以及南京审计大学的王伟、颜羽、范贝蓓、高扬、王倩怡、郝茂林等同学。南京审计大学的马璇、夏戈奥及南京外国语学校仙林分校的孙琰同学对全文进行了非常细致的校对工作。同时，本书的编撰也得到了南京赛融信息技术有限公司总经理余小宁博士在数据挖掘技术等方面的指导，一并致谢。

此外，需要特别感谢经济管理出版社，由于笔者的拖延症，交稿比预订计划晚了近 2 年，顺致歉意。同时，本书出版得到了南京财经大学（昆山）花桥现代服务业研究院的资助，谨致谢意。

最后，本书在撰写过程中参阅了大量中外文献资料，笔者已尽可能地将这些文献列在了书后，以供读者延伸阅读。当然，也有少部分未能在参考文献中逐一列出。在此谨向这些文献资料的原作者表示衷心的感谢！由于笔者水平有限，再加上时间仓促，书稿难免存在缺点和不足，恳请广大读者批评指正！

孙 杨

2018 年 11 月 29 日

于南京审计大学润泽湖畔

目 录

第一章　大数据的时代变革

第一节　大数据诞生

一、数据大爆炸

19 世纪以来，每个时代都有划时代性的技术革命和标签：19 世纪的煤炭和蒸汽机，20 世纪的内燃机、石油和电力，21 世纪的信息革命和生命工程。如今，“大数据”是 21 世纪最闪耀的光环，反映的是规模大到无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。“大数据现象”表现为人类社会累积的行为数据由于规模过于庞大，甚至大到已经超乎想象，起始计量单位已不能再用 G 或 T，至少要用 PB、EB 或 ZB。美国数据公司 IDC 和 EMC 的联合调查结果表明，2011 年全球生成的数据总量已经达到 1.8ZB，而这个数值在过去的几年中又以每两年翻一番的速度增长，2015 年全球数据总量达到 8.6ZB，2017 年全球数据总量更是达到了 21.6ZB。目前，全球数据量的增长速度在每年 40% 左右。据 IBM 公司统计，全球每天生成的数据量已达 2.5EB，换算成市面上一张存储容量为 25GB 的蓝光光碟，相当于 10 亿张光碟存储的数据量。同时，在从古至今人类社会形成的既有的存量数据中，有九成左右都是近几年内生成的，这种趋势还会不断延续。预计到 2020 年，全球将生成 35ZB 的数据量，相当于 2009 年的 44 倍，2011 年的 20 倍，具体如图 1－1 所示。在 35ZB 的数

据总量中，我国生成的数据量约占总量的40%。

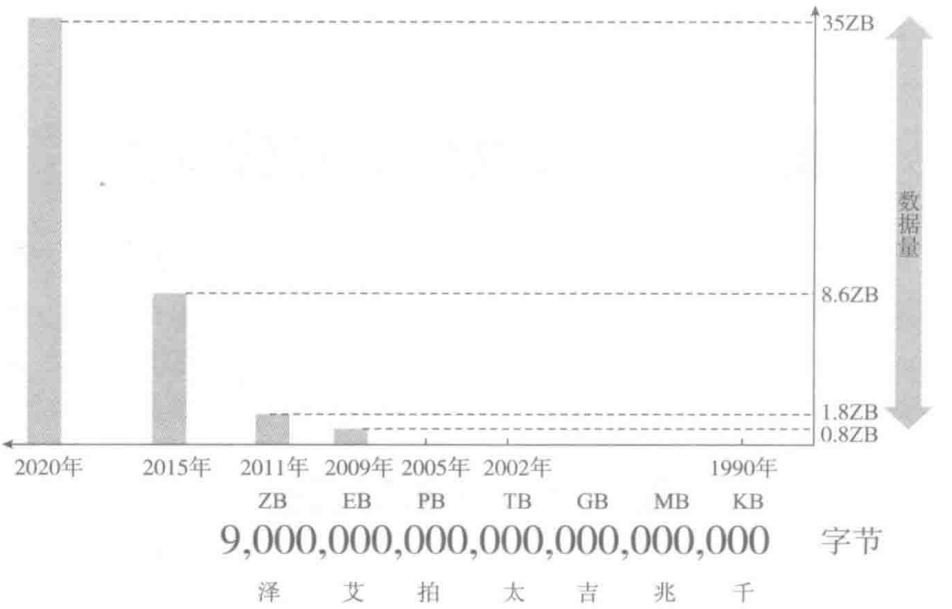


图1-1　全球生成数据总量及预测

“35ZB”究竟是多么庞大的数据量？1ZB相当于10的21次方，35ZB的数据，需要40000亿盘高清录像带才能收录得下，一个人如果日夜24小时、一年365天不间断地观看，需要94000万年才能看完；如果用64G内存的U盘来保存的话，需要5750亿个，如果把它们排列起来，其长度可以环绕地球904圈；还相当于中国全国人民每分钟发1条微信状态，连续发539520年的数据总量。如果将上述提及的单位换算成日常生活中较为常用的GB，那么1TB相当于2^{10}GB（千），1PB相当于2^{20}GB（百万），1EB相当于2^{30}GB（十亿）。除此之外，与大数据爆炸有关的预测还有很多，而且每隔几个月这些预测又会再更新调整一次。

大数据变革在各行各业掀起了轩然大波，恰如瓦特的蒸汽机在工业上引起的革命，大数据也成为一场革命，创造了一个行业，也迅速改造了众多行业。20世纪90年代互联网在中国兴起，经过十几年的发展探索，在当今大数据潮流的冲击下，也逐渐成为一个颇受关注的大数据领域。如微信现已成为国内最大的社交软件，使用智能手机的用户基本上都会注册使用微信。2017微信数据报告显示，2017年底我国拥有7.53亿的手机网民规模，微信已实现对国内移动互联网用户的

大面积覆盖。2017 年微信登录人数达 9.02 亿，较 2016 年增长 17%，日均发送微信次数为 380 亿，这说明微信现已成为国内最大的移动流量平台。目前，微信已完全融入了国内网民生活，培养出用户高度的依赖性，占据网民上网 23.8% 的时间，排在第二位的腾讯视频仅占据 4.9% 的时间。《创新生态共同体助力经济新动能——2017 年微信经济社会影响力研究》显示，2017 年由微信驱动的信息消费达 2097 亿元，拉动流量消费达 1191 亿元，占行业流量收入的 34%。仅是微信 2017 年统计得出的数据量，就已大到让人难以想象和计量，所以据此可以认为，数据大爆炸催生的大数据时代已经来临，任何行业和个人都不能置身事外，如果不能适应 21 世纪这个数据时代，那就只能被淹没在数据的海洋里。

二、大数据特征

IBM 和 IDC 两家公司对大数据基本特征的概括——“4V”是业界最具代表性的，即海量化（Volume）、时效高（Velocity）、多样化（Variety）、价值化（Value）。海量化指数据量巨大，超出常规，需要用至少 TB 或 PB 单位来衡量；时效高强调的是数据增长速度快，在数据获取、传输、处理、利用等层面的高速高效；多样化指数据的来源丰富，不仅包括以往大量的便于存储的结构化数据，也包括像文字、图片、音频、视频、地理位置和点击流等非结构化的数据；价值化强调大数据价值密度低，从海量数据中提取和挖掘有用的微量信息如浪里淘沙，数据产生的速度越快，单位容量里的数据所含的价值就越低，但正是这些看起来价值密度很低的数据却能挖掘出意想不到的价值。

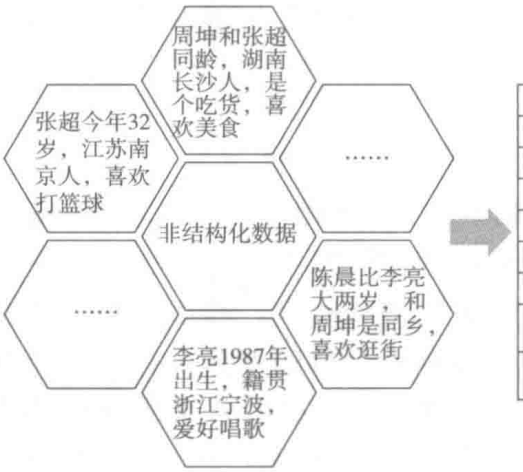
关于大数据特征的概括，有几点需要特别说明：

（1）大数据的“大”是相对的。“大”作为形容词具有相对性，如刘翔跑步跨栏的速度非常快，与他相比一般运动员的速度就是慢的。几十年后再谈大数据，会有更高的标准。在大数据概念普及之前，经济、金融、军事、气象、医疗等领域，早就使用了类似大数据的研究手段。

（2）大数据目前只是开端。大数据概念由被人们认知到熟识仅短短几年时间，有专家预测，未来 5 年内，全球每秒生成的数据量将会激增至现在的 8 倍，每个人平均拥有 7 件可穿戴设备连上互联网，那时候，大数据发展竞赛才进入正式阶段，目前是要建立大数据思维，

做好充分的技术准备。

（3）八成数据都不适合分析。在每天生成的数据中有80%左右都属于非结构化数据。“非结构化数据”是指图片、音频、视频等信息，包括微信上发的状态和留言评论，爱奇艺视频网站上的声音、影像数据，淘宝网上的购物日志数据等。这些“非结构化数据”不适合分析，而与之相对的“结构化数据”则更适合用来分析研究，如图1－2所示。未来，计算机擅长处理的结构化数据量的增加幅度要远远低于非结构化数据量的增加幅度。



结构化数据

序号	姓名	年龄	籍贯	爱好	……
1	陈晨	34	湖南长沙	逛街	
2	李亮	31	浙江宁波	唱歌	
3	张超	32	江苏南京	打篮球	
4	周坤	32	湖南长沙	吃美食	
5					
6					
⋮					
⋮					

图1－2　非结构化数据整合结构化数据

（4）核心数据永远稀缺。大数据时代每天会生成海量化的数据，其中大部分价值不高的数据的长期积累不仅会造成数据泛滥，增加存储成本，还会耗费大量分析所需的时间成本。那么，如何巧妙地收集非结构化的数据、整合多种类型的数据，使海量数据背后的隐性知识显性化成为大数据时代的关键。其实，大数据的核心就是，收集到以往无法收集到的数据，看见以往看不见的事物。大数据从显性知识中提取隐性知识的过程如图1－3所示。

三、大数据问题

2008年谷歌公司通过分析500万条被美国人频繁搜索的词语，发

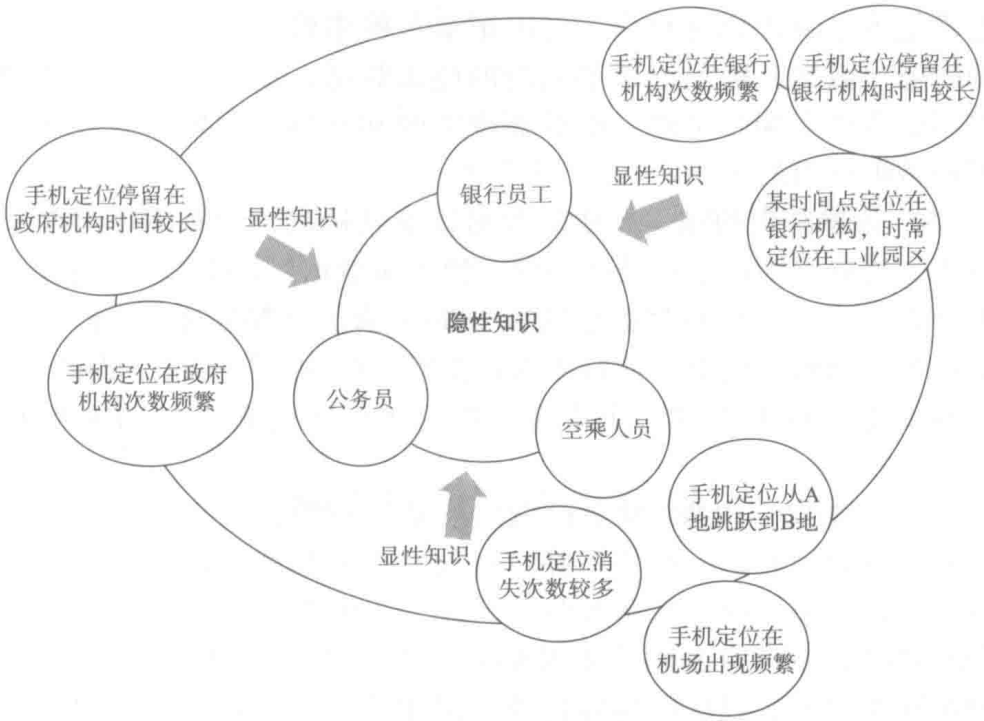


图 1－3　大数据从显性知识中提取隐性知识

现其与流感疾病关联度很高，谷歌公司后经对比 2003～2008 年美国疾病中心掌握的季节性流感时期的数据并建立数据模型，成功地预测了 2009 年冬季流感的传播趋势，具体到美国各州县。根据疾病防疫中心的事后评估，谷歌公司本次预测的准确度超过 95%，这项研究成果在 2009 年 2 月被发表在了《自然》杂志上。但这个成功的应用场景在 2013 年的流感预测中却出现严重错误。

大数据在信息获取、传递、处理、利用等功能应用上日益多样化，为人类生产方式、生活方式、思维方式的变革发挥了正向推进作用。如今，围绕大数据研究的课题堆积如山，人们在享用大数据便利的同时，也应注意到在收集、分析、利用大数据上会存在的各种问题，既要充分利用，又不陷入大数据思维定式。

（1）数据共享问题。日常生成的数据不仅量大，而且结构也很复杂，它们分布在不同的地理区域、行政部门或企业平台，整合起来较为困难。比如腾讯储存了人们在 QQ 和微信上通信、生活的数据，阿

里巴巴记录了消费购物数据，百度记录了搜索数据，移动运营商记录了即时通信数据，医院记录了人们的健康数据，这些不同类型的数据被不同的主体存储和掌握，很难实现关联和共享，这种数据孤岛现象会阻碍数据价值的实现。

（2）数据割裂问题。生成的数据缺乏结构化，物理实体与虚拟实体及虚拟实体之间缺乏有效的映射，给多源数据整合带来了挑战。IDC报告中提出，2012年全球数字信息中80%的信息都是视频、声音和图像文件等非结构化信息，这对数据转化和分析的手段提出了很高要求，只有不断更新技术手段，才能保证数据被处理过程中的真实性和完整性。

（3）数据安全问题。“数据就是财富”这种认识目前已成为政府部门和各行业的共识。在共享经济时代，企业和商家都积极主动地通过各种方式采集消费者数据并整合运用，人们在享受这些大数据带来便利的同时，也将自身的信息和隐私有形或无形地透露出去。另外，政府在社会管理宏观层面需要记录、采集公民的隐私信息，这本无可厚非，但如果这些涉及隐私的数据被不法分子获取，将给个人带来很大的安全问题。

（4）数据表象问题。过于依赖已获得的、能得到的数据容易引起对事件本质的错误判断。设想，如果人类数据的分析能力发展到可以预测个体犯罪何时发生，那时会产生一个可怕的伦理问题，即个人需要对即将发生的犯罪负责而不是对已发生的违法行为负责，这听起来非常荒唐，违背逻辑和法律精神。再如现在社会上的投诉问题，A地由于经济发展水平落后，市民生活安逸简单，政府部门接到各方面信访投诉的案件较少；B地经济发展迅速，市民生活节奏快、吸引外来人口多、各行业发展程度高，政府部门经常接到信访投诉。如果仅根据B地市民投诉的案件数量明显多于A地就得出“B地政府施政能力不如A地，群众对政府满意度不如A地”的结论是很不严谨的。

（5）数据导向问题。数据是冰冷的，执迷于数据表面数值含义而忽略数据的内涵和实质这种情况一直都存在，而非专属于大数据时代。最常见的例子是地方政府官员的政绩考核，通过GDP的增长来评定一个官员的执政能力和水平，这种数据导向问题容易造成地方官员过分

注重 GDP 数字本身，而忽略执政为民这一根本指引，无法做到真正为人民服务。

类似对大数据问题的思考还有很多，大数据虽是一个很重要的概念，代表了一个重要的趋势，但也绝对不是放之四海而皆准的万能钥匙。努力在适当的领域应用它、拓展它；不适当的领域就停下来，这应是所有人面对这个新概念、新领域和新思潮应有的负责任的态度。当下，大数据似乎成了“万金油”，但人们也应避免“当你手里有了斧头，看什么都像木头”这种倾向，时刻保持清醒头脑。

第二节　大数据变革

大数据时代的来临并不仅仅是因为数据大爆炸所表现的数据量的增加，更重要的是当下有关数据存储、传输、处理和分析的技术已经发展到了一个新的高度。正如 Intel 公司创始人戈登·摩尔的“摩尔定律”所描述的那样，“集成电路上能容纳的晶体管数量约每隔 18 个月翻一番”。直到现在，这种情况还一直在持续。之所以说现在是大数据时代，是因为各项技术的进步推动着大数据的发展，支撑着大数据的时代变革。

一、数据处理技术的发展

（一）处理器性能的提升

CPU（Central Processing Unit）中央处理器作为信息产品系统中必不可少的核心元件，在计算机运行过程中负责对指令的执行和数据的处理。计算机所有的操作都受 CPU 控制，对数据的算术运算、逻辑运算和其他信息的处理完全依靠 CPU 进行，CPU 性能的高低直接决定着数据处理能力的大小。CPU 的性能体现了计算机发展的程度，是信息社会发展的重要标志。可以毫不夸张地说，CPU 的发展支撑着整个信息产业的发展，CPU 技术的不断改进给信息产业创造了更多发展的机会。CPU 性能的发展过程如图 1 -4 所示。

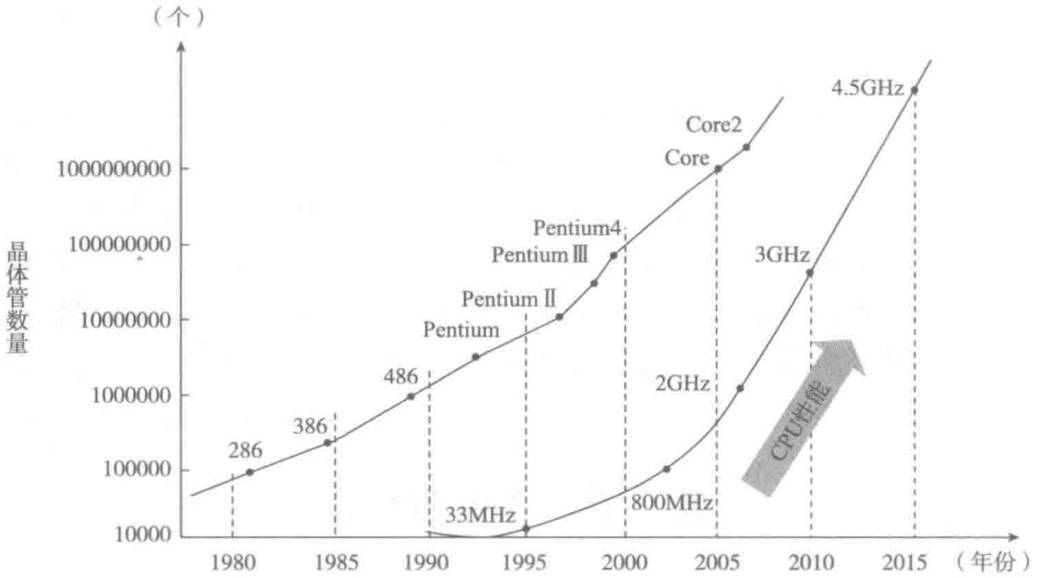


图 1-4　CPU 性能飞跃式发展

CPU 内部由数以亿计的微型晶体管共同组成控制单元、逻辑单元和存储单元，晶体管的数目及密度极大地影响着 CPU 的数据处理能力。如“摩尔定律”所描述的，“集成电路上可容纳的晶体管数量约每隔 18 个月翻一番”，性能也将提升一倍。虽然在当时这只是一个推测的理论，但根据 40 多年来 CPU 发展的历史看，理论基本和现实情况相一致。1971 年 Intel 推出世界上第一款微处理器 4004，包含 2300 个晶体管。1978 年 Intel 推出 8088 芯片，是第一块成功用于个人电脑的 CPU，内含 29000 个晶体管。1989 年 Intel 推出 80486 芯片，它的特殊意义在于这块芯片首次突破了 100 万个晶体管的界限，集成了 120 万个晶体管。1999 年 Intel 发布奔腾（Pentium）Ⅲ处理器，含有 950 万个晶体管。2005 年 Intel 第一个主流双核处理器奔腾 D 诞生，含有 2.3 亿个晶体管。2013 年酷睿 i7 - 4960X 问世，晶体管数量达到 18.6 亿个。2018 年酷睿 i9 - 7980XE 问世，晶体管数量直达百亿级。

（二）存储设备性能的提升

不仅是 CPU，数据存储设备的发展也是日新月异，表现在数据存储效率不断提升、容量不断增加、价格成本反而不断降低。目前，运

用闪存技术的 SSD（Solid State Drive）固态硬盘逐渐普及，打开了高速处理海量数据的大好局面。SSD 作为一种存储器，是运用闪存技术的驱动装置，比起旋转轴式的 HDD（Hard Disk Drive）机械硬盘，其在读写速度、耗电量和稳定性等方面的优势更加明显，如图 1－5 所示。传统 HDD 读取速度的极限是 200M/s，一般写入速度很难突破 100M/s。SSD 不用磁头，几乎没有寻道时间，在传输速度上最高可达 500M/s，读取速度在 400～600M/s，写入速度同样可以高达 500M/s，下载一部容量为几个 G 的高清电影只需几秒。SSD 的快绝不仅仅体现在持续读写上，其随机读写速度的快最直接地体现在了绝大部分的日常操作中。同时，还有极低的存取时间，最常见的 7200 转 HDD 的寻道时间一般为 12～14ms，而 SSD 可以轻易达到 0.1ms 甚至更低。另外，在重要的数据安全方面，SSD 没有盘片，即使在高速移动甚至翻转倾斜的情况下也能正常使用，而且在发生碰撞和震荡时数据丢失的可能性较小，数据安全性高。传统的 HDD 是通过磁头读取盘片来完成数据读写的，在高速旋转过程中盘片和磁头碰撞容易造成数据受损，运输过程中颠簸也容易使盘片受损造成数据丢失。

数据安全	访问速度	其他方面
固态硬盘：没有盘片 使用条件相对宽松 机械硬盘：磁头读取盘片 不能运输颠簸、碰撞	固态硬盘：读取速度400~600M/s 写入速度500M/s 机械硬盘：读取速度极限200M/s 写入速度不超过100M/s	固态硬盘：体积小、质量轻 噪声小、功耗低 机械硬盘：体积大、会震动 有噪声、功耗高

图 1－5　固态硬盘对比传统机械硬盘

二、云计算技术的发展

云计算技术炙手可热，被认为是互联网行业发展的下一个风口，其发展和普及也为大数据变革刮来东风。简单地说，云计算可以实现把本地电脑上的工作任务，安排到远程电脑上去计算处理。它最早兴起于美国，微软、谷歌、亚马逊等公司都为云计算做过前期技术铺垫。近几年在阿里巴巴、百度等互联网公司的倡导下，云计算的概念慢慢在国内兴起。阿里巴巴作为最早引入和使用云计算的公司之一，在建