DASHUJU SHIDAI XIA BAOXIAN GONGSI DE CHUANGXIN ZHI LU

大数据时代下保险公司的创新之路

张　娓　著

重庆大学出版社

前言
PREFACE

我们正处于一个急速变化的时代，“互联网”“大数据”“区块链”这些层出不穷的新技术一次又一次地改变着我们认知事物的方法。伴随着科学技术的进步，日常生产、生活中产生的数据迅速膨胀，大数据不仅对传统数据处理技术提出了挑战，而且颠覆了传统的思维方式。在大数据时代下，墨守成规显然是不合时宜的，唯有创新，适应新时代带来的新变化，才能立于不败之地。

建立在大数法则基础上的保险经营与大数据技术有着天然的联系，两者的重要特征都是通过数据进行预测。大数据时代的到来对保险公司来说既有挑战也有机遇，保险公司只有在发展中不断创新，才能获取更多的机遇、创造更大的价值，在竞争中立于不败之地。

本书结合我国保险公司的现状，力求用大数据思维探索大数据时代下保险公司在产品、销售、服务、风险管理等各个经营环节的创新；本书不仅专门列出章节对走在创新前沿的公司进行案例整理及评析，而且在其他各个章节中贯穿案例举证，在保证内容通俗、有趣的前提下，尽力为读者提供一些新思路、新视角。

本书主要内容如下：

第一章从大数据在经济、生活中的运用说起，讨论什么是大数据以及运用大数据所能产生的价值，梳理大数据的类型及获取途径，明确大数据关键技术及相关技术。作为与数据密切关联的保险行业，在大数据时代下运用大数据技术进行创新，突破传统保险技术局限，无疑会给保险公司带来良好的发展机遇，但同时保险公司对大数据的运用也才刚刚起步，面临着巨大的挑战。

第二章至第六章从思维、产品、营销、服务、风险管理技术五个方面分别探讨保险公司如何利用大数据进行创新。第二章主要讲保险思维创新。大数据思维的精髓在于放弃对精确的追求，利用大量数据寻求事物的相关性。保险公司要想利用大数据创新必须将大数据思维贯穿创新过程始终。第三章主要讲利用大数据对保险产品创新。一是对保险产品定价依据的创新，二是对产品设计，如开发方式、可保风险、保险期限、缴费方式等的创新。第四章主要讲保险营销创新。探讨大数据在保险营销模式、营销渠道、营销方法三个方面的创新运用。第五章主要讲保险服务创新。利用大数据不仅可以在承保、理赔两个环节上优化客户体验，还可以在保险保全以及一些附加服务上提升服务水平。第六章主要讲风险管理技术

创新。作为专业经营风险的公司，保险公司不仅面临着一般公司的经营风险，而且面临着业务本身具备的风险。利用大数据技术，可以帮助保险公司创新风险管理技术，提高风险管理水平。

第七章紧贴第二章至第六章的内容，选取了十四个相关案例。这些案例从不同角度反映了国内外先行保险公司如何运用大数据技术突破传统保险运营局限。每个案例都配有案例评析，引导读者思考。

科技的发展日新月异，从成书到面世期间大数据技术也在不断进步，加之作者学识、眼界有限，本书存在的不足之处敬请读者批评指正。

著　者

2019 年 1 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目录
CONTENTS

第一章　改变保险公司的大数据 …… 1

第一节　大数据时代的到来 …… 1

第二节　大数据及大数据技术 …… 2

第三节　大数据时代下保险公司的机遇与挑战 …… 13

第二章　大数据时代下的保险思维创新 …… 21

第一节　大数据思维 …… 21

第二节　保险思维创新 …… 24

第三章　保险产品创新 …… 27

第一节　更加科学的产品定价 …… 27

第二节　保险产品设计创新 …… 32

第四章　保险营销创新 …… 37

第一节　保险营销模式创新 …… 37

第二节　保险营销渠道的创新 …… 40

第三节　保险营销方法的创新 …… 48

第五章　保险服务创新 …… 56

第一节　打造优质高效的承保理赔流程 …… 56

第二节　保险客户服务创新 …… 65

第六章　风险管理技术创新 …… 70
第一节　保险公司面临的主要风险 …… 70
第二节　保险公司风险管理技术创新 …… 75

第七章　保险公司创新案例 …… 86
案例一：全新的保险公司组织形式——众安保险 …… 86
案例二：平安车险依托大数据布局车主生态圈 …… 88
案例三：大数据成就退货运费险 …… 90
案例四：UBI——不仅仅是保费折扣 …… 91
案例五：可以带病投保的“糖小贝” …… 94
案例六：乐业保从大数据中发现商机 …… 96
案例七：i 云保来了，传统保险代理人路在何方 …… 97
案例八：场景营销的范例——“虚拟人生计划” …… 99
案例九：场景+新媒体成就泰康“点亮中国” …… 100
案例十：实现保险全流程互联网化的“加班险” …… 101
案例十一：车险快速理赔找“北京交警” …… 102
案例十二：DEO 用大数据提升客户体验 …… 104
案例十三：安泰保险用大数据管理健康风险 …… 105
案例十四：不用福尔摩斯也能破获骗保大案 …… 107

参考文献 …… 109

第一章　改变保险公司的大数据

第一节　大数据时代的到来

通过处理生前在社交网络上留下的信息，逝去的人就可以复活？英国电视剧《黑镜》中就讲述了这样一个高科技故事：玛莎的爱人艾什意外去世，一家公司通过整合分析艾什生前在各种社交媒体上留下的生活信息，重新创造出“艾什”——当然它只是一个人工智能机器人。机器人根据艾什生前的数据信息预测出在特定的环境下，艾什应有的反应，从而玛莎可以像过去一样和艾什一起生活。这个故事放在过去可能看起来比较荒诞，而在将来也许并非不可实现。我们正在经历一个数据不断膨胀的时代，只要我们能找到观察问题的新角度，越来越多的东西都可以数据化，包括人。

如果说《黑镜》中的故事还被称为“科幻”，那么谷歌利用大数据预测流感却已是事实。谷歌的工程师认为，在互联网时代，人们更习惯通过网络搜索来解答各种问题。谷歌每天都会收到来自全球超过几十亿条的搜索指令，汇总这些海量的搜索记录，就可以发现它们指向的问题。2008 年谷歌推出的流感趋势系统就是通过寻找与流感相关的搜索词汇，如“发烧”“咳嗽”，成功地在 2009 年甲型 H1N1 流感暴发的前几周就预测到了流感在美国国内的传播，甚至具体预测到特定的地区。这一预测令公共卫生官员倍感震惊，由于一些条件的限制，流感病人往往会等到病情比较严重时才去医院就诊，而医院在接收病人之后，又要经过一段时间才能将信息传递给美国疾病控制与预防中心。因此，采用传统方法的疾控中心通常会在流感暴发一两周后才能得到结论。

亚历山大图书馆始建于托勒密一世（约公元前 367—前 283 年）时期，是世界上最古老的图书馆之一。据说亚历山大图书馆修建之初唯一的目的就是“收集全世界的书”，实现“世界知识总汇”的梦想，历代国王为此采取过各种正常或非常的手段使亚历山大图书馆迅速成为人类早期历史上最伟大的图书馆——据说极盛时期馆藏各类手稿逾 50 万卷，被认为收藏了人类的全部知识。而如今，如果把全世界的信息进行均分，那么每个人所拥有的信息量足以超过当年亚历山大图书馆全部藏书的 320 倍。

我们周围到底有多少数据？它们增长的速度有多快？许多人试图测量出一个确切的数字。2012 年 12 月，国际数据公司（International Data Corporation）发布了研究报告《2020 年的数字宇宙》：2005 年全球产生的数据量为 130 EB[①]，2008 年为 0.49 ZB，2009 年为 0.8 ZB，2010 年为 1.2 ZB，2011 年为 1.82 ZB，2012 年为 2.8 ZB，几乎每两年翻一番。这些数据究竟有多大？可能我们无法直观地感受到，百度公司给出了更形象的说法：百度首页导航每天要从超过 1.5 PB 的数据中进行挖掘，这些数据如果打印出来需要用到超过 5 000 亿张 A4 纸，把这些纸平铺可以铺满整个海南岛，而如果全部垒叠起来高度将超过 40 000 千米。

数据正以惊人的速度不断产生，来看看 2015 年国外一家公司展示的各大网站在 1 分钟内产生的巨大数据量：YouTube 用户每分钟上传 300 小时的新视频；Netflix 用户则每分钟观看 77 160 小时的视频；Apple 用户每分钟下载 51 000 个应用；亚马逊网站每分钟访问的用户量是 4 310 名；Uber 每分钟能获得 694 个订单；Facebook 用户每分钟点赞 4 166 667 次；Twitter 用户每分钟发布 347 222 条推文；Tinder 用户每分钟浏览 590 278 份档案；Snapchat 用户每分钟会发布 284 722 张照片。我们越来越清晰地认识到，一个新的时代已经到来，我们已经无法避免大数据对我们生活、工作的影响。

大数据一词来源于英文“Big Data”，早在 1980 年阿尔文·托夫勒（Alvin Toffler）就在他所著的《第三次浪潮》中使用过大数据一词，并将其赞颂为“第三次浪潮的华彩乐章”。2001 年高德纳（Gartner Group）公司的分析师道格拉斯·兰尼（Douglas Laney）首次从大数据特征的角度对其进行了相对明确的定义，他强调大数据必须具备 3 V 特征：即体量大（Volume）、多样化（Variety）和速度快（Velocity）。2008 年 9 月，《自然》刊登了“Big Data”的专辑，探讨如何研究、利用 PB 级容量的大数据流。大约从 2009 年开始，“大数据”才成为互联网信息技术行业的流行词汇。2012 年 1 月，达沃斯世界经济论坛发布了《大数据，大影响：国际发展的新机会》的报告，宣称数据就像货币和黄金一样，已经成为一种新的经济资产，全球进入大数据时代。

第二节　大数据及大数据技术

一、大数据的含义

2011 年 5 月，全球知名咨询公司麦肯锡（Mckinsey）发布了《大数据：下一个创新、竞争和

① 1 EB=1 024 PB，1 ZB=1 024 EB。

生产力的前沿》的研究报告。这份长达150多页的报告系统地阐述了大数据的概念，预测了全球数据发展的趋势，详细列举了大数据的核心技术，深入分析了大数据在不同行业的应用，并且提出了政府和企业决策者应对大数据发展的策略。麦肯锡报告对大数据的定义是“大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群”。大数据被认为不仅仅是海量数据，最初由道格拉斯·兰尼所归纳的3 V特征，也被扩展至6 V加1C，即数据体量大(Volume)，类型多样化(Variety)，处理速度快(Velocity)，应用价值大(Value)，数据获取与发送的方式自由灵活(Vender)，数据分析处理后结果的真实准确性(Veracity)及处理和分析难度非常大(Complexity)。

如今大数据不仅用来描述大量数据，还涵盖了处理数据，从某种程度上说，大数据是数据分析的前沿技术。大数据可以分成大数据技术、大数据应用、大数据工程和大数据科学等领域。大数据技术是指从各种类型的大数据中快速获得有价值信息的技术，包括数据采集、存储、管理、分析挖掘、可视化等技术及其集成。大数据应用是指对特定的大数据集合，集成应用大数据技术，获得有价值信息的行为。大数据工程指大数据的规划建设运营管理的系统工程。大数据科学关注大数据网络发展、运营过程中发现和验证大数据的规律及其与自然和社会活动之间的关系。现在人们谈论得最多的是大数据技术和大数据应用。

二、大数据的价值

数据本身并不产生价值，如果数据仅仅用来储存，那么泛滥的数据将会带给我们很多麻烦。反之，如果能驾驭数据，利用大数据对实务产生帮助才能体现出数据的价值。不同领域、不同企业、不同业务的需求，数据集合和分析挖掘目标存在差异，因此它们所运用的大数据技术和大数据信息系统也可能有很大的不同，但如果坚持“对象、技术、应用”三位一体同步发展，就能够充分实现大数据的价值。目前我国规模以上公司搭建大数据平台情况的统计图如图1.1所示，预计未来我国智慧经济年均增速约为14%，截至2025年智慧经济的贡献有望达到7.8万亿。

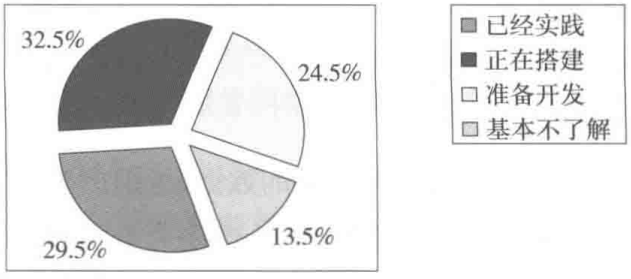

图1.1　规模以上公司搭建大数据平台情况统计图

资料来源：2016年中国大数据交易产业白皮书

最初，互联网公司可以收集到大量有价值的数据，它们通过新兴的数据处理技术，可以

解决在小规模数据基础上无法完成的事情，从而获取利益。现在，越来越多的行业都将目光转向了大数据，大数据的运用从互联网公司延伸向传统行业，营销领域、农业、通信业、医疗卫生行业、政府及公用事业行业、金融行业等都可以利用大数据创造出新的价值。生物科学第一次破译人体基因密码的时候用了10年的努力才完成30亿对碱基对的排序，而10年之后借助基因仪，这样的工作只需要15分钟就可以完成。金融行业数据类型丰富，数据质量好，信息化程度高，走在运用大数据的传统行业前列。美国股市三分之二的交易都是由建立在数学模型和算法之上的计算机程序自动完成的，这些程序会根据大数据来预测利益和降低风险。保险业和大数据也有天然的联系，保险产品定价的基础建立在大数法则之上，大数据的运用可能会对精算技术产生变革，进而影响保险产品定价。除此之外，在保险公司产品的销售、风险的控制、反欺诈等方面，大数据都可以发挥作用。图1.2是金融大数据应用市场规模的预测图，随着大数据的推广，金融行业可以找到新的增长方式。

图1.2　金融大数据应用市场规模预测

资料来源：前瞻产业研究院

三、大数据的数据类型、分布及获取途径

1. 大数据的数据类型

我们可以从不同的角度去认识数据的类型。

(1)从数据产生的主体来看

从数据产生的主体来看，可以分为个人产生的数据、组织产生的数据、机器产生的数据。个人通过发布微博、微信，发帖、点击、留言等方式会产生大量的数据。企业、事业单位、行政部门等组织会在运营过程中产生诸如销售、仓储、财务等相关数据。应用服务器日志、传感器数据、二维码数据等海量数据会由相应机器自动产生。

(2)从数据存储形式来看

从数据存储的形式来看，大数据可以分为结构化数据、半结构化数据和非结构化数据。

结构化数据以固定字段驻留在一个记录或文件内，是可以用二维表结构来逻辑表达实现的数据，存储在关系数据库里。半结构化数据介于结构化和非结构化数据之间。不同于结构化数据的先有结构再有数据，半结构化数据是先有数据再有结构，它的格式较为规范，一般都是纯文本数据，很多 XML、JSON 等格式的文件就属于这一类。非结构化数据没有一个事先定义的数据模型或不是以事先预定好的方式进行组织。它存储在非结构数据库中，突破了关系数据库结构定义不易改变和数据定长的限制。它没有标准格式，包括所有格式的办公文档、文本、图片、音频、视频等。

2. 大数据的行业分布

海量的数据主要分布如下：

(1)以 BAT 为代表的互联网公司

百度占有 70% 以上的搜索市场份额，拥有庞大的搜索数据；阿里巴巴拥有 90% 以上的电商数据；腾讯拥有大量通过社交、游戏等领域积累起来的文本、音频、视频和关系类数据。

(2)电信、金融、电力、石化系统

仅从银行卡来看，2013 年全国“银联”银行卡发行量就已经接近 40 亿张，每天有近 600 亿元通过银联的银行卡交易。不仅如此，开户信息数据、在线交易数据、金融系统自身运营的数据等，使国内银行系统每年产生的数据能达到数十 PB，保险系统生成的数据也接近 PB 级别。

(3)公共安全、医疗、交通领域

如今道路监控摄像头星罗棋布，每天会产生大量的视频数据，临床信息、健康档案、疾病监控也是大数据的来源。

(4)气象与地理、教育、政务等

以气象卫星和多普勒天气雷达为代表的遥感遥测业务每天会产生 TB 级的观测数据，绝大多数中央部委和省级政府部门的核心业务都有数据库支撑，这些数据库涵盖全国的人口数据、企事业社会团体的相关信息等，拥有巨大的数据存储。

(5)商业销售、制造业、农业、物流等其他行业

目前这些行业处于数据积累期，随着网络的普及，这些行业的数据会越来越多。

3. 获取大数据的途径

想要利用大数据创造价值，首先要获得基础的数据，获得数据的途径通常有以下几种。

(1)内部途径

通过一些数据采集工具或者软件，对自身生产经营过程中所产生的内部数据进行收集。如经授权后在自己的官网上收集用户的 Cookie 数据，通过 Cookie 跟踪统计用户访问该网站的习惯，如习惯访问网站的时间，访问了哪些页面，在每个网页停留的时间等，即使在用户没

此为试读，需要完整PDF请访问：www.ertongbook.com

有登录的情况下，也能识别用户身份，获取相应的信息。APP 也是获取用户移动端数据的一种有效手段，如果将自身 SDK 内置在 APP 中，用户访问时甚至不用访问 APP 内容都能将信息汇总给指定服务器，获知用户终端的相关信息，如用户安装了什么样的应用，以及有多少个应用等。汇总这些数据，进行分析处理后能得到有用的信息。

已经有保险公司主动创建自己的数据库。MetLife 保险公司已经投资了 3 亿美金建立一个新式系统，其中的第一款产品是一个基于 MongoDB 的应用程序，它将所有客户信息放在同一个地方。MongoDB 汇聚了来自 70 多个遗留系统的数据，并将它合并成一个单一的记录。它运行在两个数据中心的 6 个服务器上，目前存储了 24 TB 的数据，囊括了 MetLife 的全部美国客户，它的更新几乎是实时的。MetLife 还计划将它的国际客户数据也纳入其中。

2014 年初，中国人寿上海数据中心正式竣工投产。该数据中心总用地 80 亩[①]，总建筑面积 13.1 万平方米，由 3 栋 8 层主楼及 2 栋单层能源动力区组成。机房可用面积约 1 万平方米，可靠性达 99.995%。这一以技术、服务、安全三大体系建设为手段，布局大数据、云计算、虚拟化、移动互联等先进技术领域的数据中心被认为是中国人寿信息化建设的一个重要里程碑，标志着中国人寿信息科技水平跨入同业领先行列。

泰康人寿基于云数据中心构建大数据平台，为包括寿险类、年金类、资产类以及养老社区等全业务领域提供全方位随动的技术支撑。泰康在武汉光谷、北京长安街和北京中关村建有三个数据中心，其中，中关村数据中心分为地上和地下两层，建筑面积 8 000 平方米，机房地板面积约 2 300 平方米，是整个数据系统的“枢纽”和“心脏”。在云中心启用的同时，泰康人寿移动互联部和大数据部也宣告正式成立。这两个部门在数据信息中心原有的数据服务、支持服务、信息技术基础设施等职能的基础上，更侧重于与大数据及移动互联网相关的应用规划、系统开发、上线维护、技术研究和市场推广规划等工作。

(2)外部免费平台

政府部门可以提供一些官方数据。过去十多年来政府开展了大量电子政务及信息化工作，积累了大量数据。这些数据和公众的生产生活息息相关，政府的开放使用会大大降低大数据的获取成本。2012 年以来，北京、上海、武汉、无锡、贵阳等城市先后发布地方政府开放数据平台。在政府数据开放初期，主要面向一些大企业。2013 年 2 月 25 日，国家食品药品监督管理局的三大药品数据库，总计 20 余万个权威药品信息全面入驻百度，与百度合作实施“安全用药，搜索护航”战略。2014 年 5 月 27 日，中国气象局公共气象服务中心与阿里云达成战略合作，共同搭建“中国气象专业服务云”，为有气象数据需求的企业提供专业化的云计算服务。2014 年 10 月 15 日，贵州省政府与阿里牵头的企业合建云计算基础设施“云上贵州”，将大数据应用在交通等领域。2015 年 1 月 13 日，阿里健康宣布将药品监管网的基础设

① 1 亩≈667 平方米。

施从甲骨文数据库迁移到阿里云平台，阿里将利用大数据技术帮助解决假药问题。2015 年 9 月，根据国务院《促进大数据发展行动纲要》的要求，各级政府开始向众多领域开放数据。此外，一些行业协会、俱乐部等也会提供半官方数据，一些民间平台也会提供一些免费数据。这些免费渠道是获取数据的来源之一，如淘宝网、京东、唯品会等平台会免费开放一些数据。

（3）外部收费平台

我国大数据市场供给的主力还是互联网企业、传统 IT 厂商和大数据企业三方，如图 1.3 所示，大数据市场初步形成三角形供给结构。需要的大数据可以通过购买的方式来获取。目前我国的数据交易平台有三种类型：

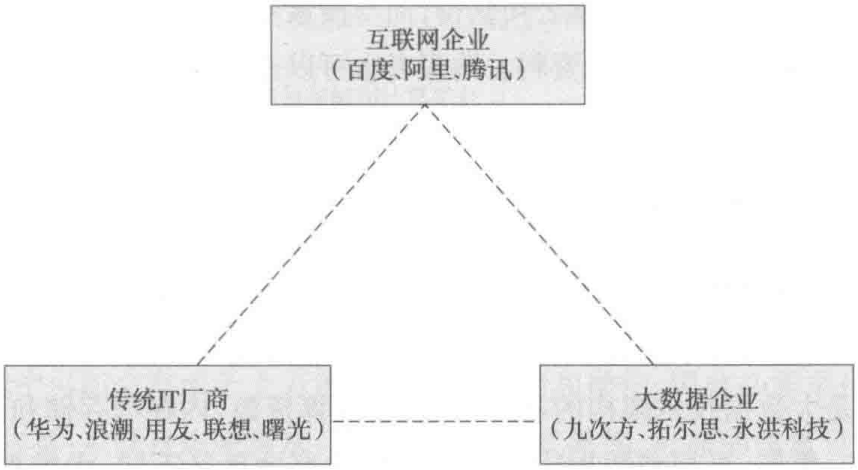

图 1.3　我国大数据市场的供给结构

资料来源：前瞻产业研究院

一种是交易所平台，以贵阳大数据交易所为代表，包括湖北长江大数据交易所、陕西西咸新区大数据交易所等。贵阳大数据交易所是全球第一个大数据交易所，采用市场化运作模式，为金融、医疗、电商、交通等 30 多个领域提供数据交易服务。交易所实行会员制，具有会员资格的企业才能通过交易所进行大数据交易。目前贵阳大数据交易所已有腾讯、京东、华为、中国人寿、中国联通等 300 余家会员单位。

一种是产业联盟性质的交易平台，以中关村数海大数据交易平台为代表。中关村数海大数据交易平台是由中关村大数据产业联盟于 2014 年承建。中关村大数据产业联盟成立于 2012 年 12 月，一直致力于推动大数据产业的发展。其核心价值定位是打造“智库、传播、资本”三位一体的新兴科技服务业模式，为政府、学术界和产业界搭建桥梁。中关村数海大数据交易平台的模式是通过开放的应用程序接口（API）进行数据录入、检索、调用，为政府机构、科研单位、企业乃至个人提供数据交易和使用。在确保数据不涉及个人隐私、不危害国家安全，同时获得数据所有方授权的情况下，为数据所有者提供大数据变现的渠道；为数据开发者提供统一的数据检索、开发平台；为数据使用者提供丰富的数据来源和数据应用。

还有一种是专注于互联网综合数据交易和服务的平台，以数据堂为代表。数据堂成立于2011年，是国内首家专注于互联网综合数据交易和服务的公司，总部位于北京，目前在南京、镇江、天津、保定等地设有多个专业数据处理中心，并在北美硅谷设有分公司。数据堂致力于融合和盘活各类大数据资源，实现数据价值最大化，推动相关技术、应用和产业的创新。数据堂旗下有三大核心业务：数据定制、数据商城、移动应用数据服务。它的数据采集范围遍及全球30多个国家，合作伙伴遍布世界10多个国家，已成功为包括百度、腾讯、阿里巴巴、Microsoft、Canon、Intel等国内外多家企业提供数据定制服务。

大数据购买者可以通过广告联盟的方式获取搜索用户的数据资料，当用户搜索一些关键词时，该用户的数据资料便由搜索公司获得，而与搜索内容相关联的数据需求方因事前购买了广告位而获得这些用户的数据资料。购买者也可以选取拥有稳定、完整、连续的数据资源的公司进行长期的战略合作。

四、大数据技术的类型

大数据技术可以包括大规模并行处理(MPP)数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网、可扩展的存储系统等。图1.4是一个典型的大数据技术栈，底层是大数据技术架构的基础层，涵盖计算资源、内存与存储和网络互联，具体表现为计算节点、集群、机柜和数据中心。与以往的存储孤岛不同，大数据基础设施必须在容量、性能和吞吐量方面都可以线性扩展，成为具有共享能力的高容量储存池。第二层是管理层，主要对结构化和非结构化数据进行管理，实现数据的实时传送、查询和计算。管理层既包括数据的存储和管理，也涉及数据的计算。数据存储和管理包括文件系统、数据库和类似YARN的资源管理系统。数据的计算处理如Hadoop、MapReduce、Spark等，以及在此之上的各种不同计算范式，如批处理、流处理和图计算等，包括如BSP、GAS等衍生的编程模型的计算模型。第三层是分析层，能够通过统计学的数据挖掘和机器学习算法对数据集进行分析和解释，从而获得对数据价值深入的理解，包括数据分析和可视化。数据分析包括简单的查询分析、流分析以及更复杂的分析(如机器学习、图计算等)。一般的可视化是对分析结果的展示，交互式可视化还可以形成迭代的分析和可视化，使分析获得新的线索。第四层是应用层，为终端用户提供决策和服务，帮助用户实现竞争优势，也是大数据价值的体现。

1. 大数据关键技术

麦肯锡报告详细列举了“大数据的关键技术”，包括A/B测试、数据挖掘、遗传算法、神经网络、时间序列预测模型、BigTable、Hadoop、标签云、Clustergram、历史流、空间信息流等技术和应用。可以从大数据采集和传输、大数据预处理、大数据储存、大数据分析挖掘、大数据应用这五个方面来认识大数据关键技术。

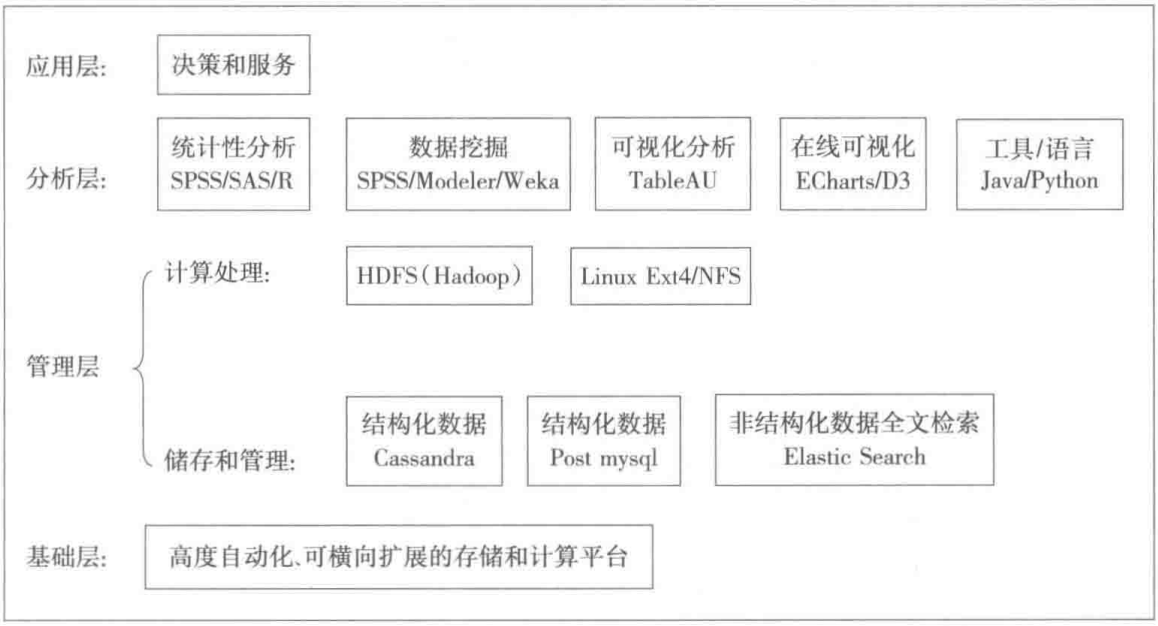

图 1.4　大数据技术栈

(1)大数据采集和传输技术

大数据采集是通过多个数据库获得结构化、半结构化及非结构化的海量数据的过程。由于在采集过程中可能会有成千上万的用户并发访问和操作,因此必须采用专门的方法采集大数据,主要包括:系统日志采集法,通过一些分布式架构、可靠的海量日志聚合系统,支持在系统中定制各类数据发送方,在收集数据的同时还可以对数据进行简单处理,能满足每秒数百兆日志数据的采集和传输。很多互联网企业都有此类数据采集工具,如 Hadoop 的 Chukwa, Cloudera 的 Flume 等;网络数据采集法,借助网络爬虫或网站公开 API 等方式,从网站上获取数据信息,能够将非结构化数据、半结构化数据从网页中提取出来,并以结构化的方式将其存储为统一的本地数据文件;其他数据采集法,针对保密性要求较高的数据,通过和相关机构合作,采用特定系统接口等方式采集。

只有高速的传输技术才能保证数据及时载入分析平台、提供实时有效的数据供用户使用,保持数据分析系统的运算峰值和运作效率。随着数据量的不断攀升,需要更加快速的传输技术与之相适应。在探索如何提高传输速度上,人们取得了一次又一次的进步。现在,利用光纤传输,每秒能轻松传输几十 TB 的数据。在无线数据传输方面,2017 年 Facebook 使用毫米波技术,在 13 km 距离上实现了破纪录的 36 GB 点对点数据速率。

(2)大数据预处理技术

大数据预处理是对采集到的原始数据进行清理,将杂乱无章的数据转化为相对单一且便于处理的数据,提高数据质量,为之后的数据分析奠定基础。大数据预处理技术主要包括数据清理、数据集成、数据转换以及数据规约四种类型。

数据清理主要采用 ETL(Extraction/Transformation/Loading)和 Potter's Wheel 等清洗工具

对遗漏数据进行填充处理、对错误数据进行降噪处理、对不一致数据进行更正处理。

来自多个数据集合的数据会因为命名的差异导致对应的实体名称不同、数据属性命名不同导致数据冗余、不同来源的同一实体具有有冲突的数据值，数据集成可以解决这些问题，它将不同来源的数据合并存放到一个一致的数据存储库中。

数据转换是对数据中存在不一致的情况进行处理，主要包括统一数据名称及格式、进行字段的组合、分割或计算等。依据业务规则对异常数据进行清洗后能保证后续分析结果的准确性。

数据归约能最大限度地精简数据量，使数据集变小，但同时仍能基本保持原数据的完整性。具体方法主要包括数据方聚集、维规约、数据压缩、数值规约和概念分层等。

(3)大数据储存技术

采集到的数据需要存储起来，建立相应的数据库，方便进行管理和调用。传统的数据存储和管理以结构化数据为主，通常关系数据库系统就能够满足需要。而大数据以半结构化和非结构化数据为主，结构化数据为辅，在应用上也需要对不同类型的数据综合分析，因此，传统的数据库已经远不能满足需要。要解决大数据储存的关键问题，需要重点解决复杂结构化、半结构化和非结构化大数据管理与处理技术。

根据数据类型的不同，大数据存储技术大致可以分为三类：第一类主要针对大规模的结构化数据。目前最佳选择是MPP(Massive Parallel Processing)数据库，它可以有效支撑起PB量级的结构化数据的存储和分析。通过高效的分布式计算，MPP可以在低成本下实现对分析类应用的支持，具有高性能和高扩展性特点。第二类主要针对半结构化和非结构化数据。更擅长对此类数据进行分析的是Hadoop，它能处理传统关系数据库较难处理的数据和场景。利用Hadoop开源的优势，通过扩展和封装Hadoop来实现对互联网大数据存储、分析的支撑。第三类主要针对结构化和非结构化混合的数据，需要集合MPP并行数据库集群与Hadoop集群，发挥各自应对结构化或非结构化数据的优势，实现对EB量级数据的存储和管理。

另外，大数据安全技术也必不可少。只有大数据安全技术得以突破，数据的真伪鉴别、数据销毁、隐私保护、数据的复制与转移等问题才能得到有效的解决。

(4)大数据分析挖掘技术

数据挖掘就是从大量的数据中提取出隐含其中的、有价值的信息的过程。这些原有数据可能并不完全甚至是有噪声的、模糊的随机数据，通过数据挖掘能从中获得人们事先并无预想的信息。

数据挖掘的技术有很多种，按照不同的标准可以对其进行分类。根据挖掘任务可以分为分类或预测模型发现、关联规则发现、异常和趋势发现等；根据挖掘对象可以分为关系数据库、空间数据库、时态数据库、多媒体数据库、异质数据库等；根据挖掘方法可以分为机器

学习法、统计方法、神经网络方法和数据库方法等。

大数据分析挖掘技术是对原有数据挖掘和机器学习技术进行改进，开发出新型数据挖掘技术，如数据网络挖掘、特异群组挖掘、图挖掘等。大数据分析挖掘技术将侧重在可视化分析、数据挖掘算法、预测性分析、语义引擎等方面取得突破。

(5)大数据应用技术

大数据技术通过挖掘隐藏在海量数据中的信息，最终为人们的社会经济活动提供支持，提高各行业的运作效率，从而大大提高整个社会经济的集约化程度。在我国，大数据将重点应用于以下三大领域：商业智能、政府决策、公共服务。商业智能可以有效提高企业经营活动的效率。如对消费者行为及趋势的分析与预测、提供个性化的购物体验以提高客户忠诚度、制订适合的广告策略等。大数据运用在政府决策上可以提高政府决策的科学性和时效性。借助民意调查、听证会等形式，大数据可以帮助建立政府与公民之间的双向信息流动机制，产生出共同的政务信息，作出的决策更符合民意。大数据在公共服务领域的应用可以涵盖教育、医疗、社会保障、环境保护等各个方面。通过信息和数据的共建共享，能够避免资源的闲置与重复供给，降低了成本，提高了公共服务供给的效率。

2. 大数据相关技术

数据的不断膨胀和技术的飞速发展已经对国家治理、经济运行和人们的生活各方面都产生了巨大的影响，大数据时代下，互联网、移动互联网、物联网、云计算、人工智能、区块链等技术都和大数据技术紧密相关、互相影响，推动着国家管理、企业生产和人们生活发生本质的变化。

(1)互联网和移动互联网

互联网是将计算机网络相互连接在一起，并在此基础上发展出覆盖全世界的相互连接在一起的全球网络结构。互联网能够不受空间限制来进行信息的交换，并且更新速度快、使用成本低。大量的互联网使用者也催生了海量的数据。这些数据可以以视频、图片、文字等半结构化或非结构化的形式存在，这也促进了大数据分析技术的发展。另一方面，互联网也是大数据分析应用最广泛的领域之一，比如依托大数据分析发展起来的搜索引擎、面向互联网用户的精准营销等。

移动互联网是移动通信和互联网的结合。移动互联网包含 3 个层面：终端、软件和应用，通过智能移动终端，采用移动无线通信方式来获取业务和服务。随着宽带无线接入技术和移动终端技术的发展，移动互联网用户数量不断攀升，据统计，截至 2018 年 12 月，中国手机网民规模已经达到 8.17 亿人。人们可以利用各种智能移动终端，如智能手机、平板电脑、电子书等，随时随地在网上交流信息，用户规模的增长带来了移动互联网市场的繁荣，也推动了数据的大爆炸。

互联网和移动互联网的数据都具有大量化、多样化和快速化的特点，是目前大数据信息

采集的主要来源，采集信息的范围、速度、数量、类型也直接影响大数据应用功能最终效果的发挥。

(2)物联网

物联网是通过射频识别装置(Radio Frequency Identification，RFID)、传感器、红外线感应器、全球定位系统、激光扫描器等信息传感设备，按约定的协议，把任何物品与互联网相连接，以进行信息交换和通信，从而实现智慧化识别、定位、跟踪、监控和管理的一种网络体系。[①]物联网把客户端延伸到了任何物品与物品之间的信息交换，可以看成互联网的延伸和扩展，因此其核心和基础仍是互联网。

物联网数据囊括了所有连接于网络上的物品，比起主要以人和服务器为数据产生来源的互联网有更大的数据量，以及更多样的数据类型。物联网所带来的大数据也正在引起社会的广泛关注。物联网的核心商业价值也是将物与物之间产生的大数据进行智能化的处理、分析，最后运用到各种商业模式中，如智慧城市、智慧交通、智慧家居、智慧医疗、智慧物流等。将大数据和物联网结合起来，可以以物联网促进大数据技术的发展，以大数据的应用带动物联网更快地向前发展。

(3)云计算

云计算是一种基于互联网的计算方式，凡是共享的软硬件资源和信息都可以通过这种计算方式根据需求提供给计算器或者其他设备。云计算相当于把起到“主机”作用的计算、服务和应用由云服务提供商的服务器集群提供，而使用者只需要“显示器”就好，并且这种“主机”是可以由多人共享的。根据开放对象的不同，可以分为向公众开放的公有云和仅供企业或组织内部使用的私有云。

云计算的产生和大数据不无关系，正是因为传统的计算架构难以处理日益膨胀的数据，才促成了云计算的产生。以亚马逊为例，亚马逊需要对大量的网上用户的数据资料进行整理、挖掘和提炼，而仅靠传统的计算方法是无法完成的，因此催生出亚马逊的海量数据处理平台，进而又在2006年推出亚马逊云计算服务(Amazon Web Services)，以Web服务的形式向企业提供IT基础设施服务。

可见，云计算的基础是大数据，没有大量的数据，云计算的计算能力也不能得以发挥，而大数据需要利用云计算强大的数据存储技术、数据管理技术、数据计算能力来处理数据、挖掘信息，以便提供更加周到、及时的服务。

(4)人工智能

人工智能是集计算机科学、控制论、信息论、仿生学、心理学、语言学等多个学科于一体的综合性学科，是用计算机来模拟、延伸、扩展人的智能。人工智能不仅有科学研究价值，还

① 参见：娄岩. 大数据技术概论[M]. 北京：清华大学出版社，2017年版，第21页。

此为试读，需要完整PDF请访问：www.ertongbook.com