BIG DATA

大数据智能分析

张华平　商建云　刘兆友◎编著

大数据、人工智能与自然语言处理三者融合贯通之作

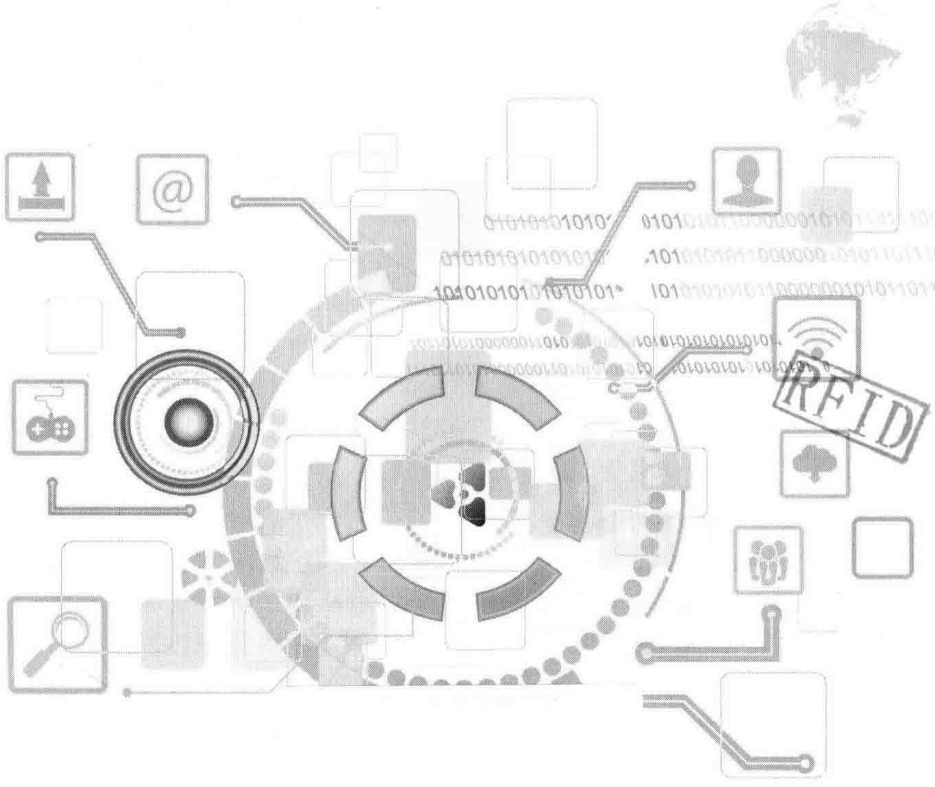
◎ 立足于作者二十年大数据智能的前沿研究与工程实践，结合北京理工大学“大数据分析与应用”研究生课程三年教学经验，体系完整，理论与实践并重。

◎ 可作为高校研究生与本科生的专业教材，也可作为大数据智能方向的科研人员和工程技术人员的实践参考书。

清華大學出版社

大数据智能分析

张华平　商建云　刘兆友◎编著



清華大學出版社
北京

内 容 简 介

大数据智能是大数据、人工智能与自然语言处理等学科交叉融合的关键技术。本书主要讲述大数据智能的框架平台、理论算法、关键技术和应用实践：在大数据与人工智能方面主要讲述了大数据智能概述、大数据技术平台与架构、传统机器学习与深度学习算法；在自然语言处理方面详细讲解了大数据精准搜索、汉语分词、新词发现、文本分类聚类、情感分析等当前热门的自然语言处理关键技术；在应用实践方面，本书进一步提供了自主研发的 NLPIR 大数据智能分析工具平台，具体介绍警情大数据、网络赌博、微博挖掘、看图说话等多个实际的大数据应用项目，也引入《红楼梦》前后作者分析、二手房房价、歌词生成等有意思的课程实践案例。

本书立足于作者近 20 年的前沿研究进展和工程实践，结合北京理工大学“大数据分析与应用”研究生课程讲授经验，体系完整，内容深入浅出，理论与实践并重，吸收了当前的技术前沿成果，同时突出原创的研究成果。本书可作为大数据、人工智能与自然语言处理方向的科研人员、高校研究生与本科生的教材，也可作为大数据智能方向的工程技术人员和爱好者的参考书。

图书在版编目(CIP)数据

大数据智能分析/张华平等编著. —北京：清华大学出版社，2019
ISBN 978-7-302-53117-3

Ⅰ. ①大…　Ⅱ. ①张…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 167582 号

责任编辑：白立军
封面设计：杨玉兰
责任校对：焦丽丽
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**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
课件下载：http://www.tup.com.cn,010-62795954
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：20.5　　**字　　数**：499 千字
版　　次：2019 年 10 月第 1 版　　**印　　次**：2019 年 10 月第1次印刷
定　　价：59.80 元

产品编号：081294-01

前　言

大数据智能是指从客观存在的全量超大规模、多源异构、实时变化的微观数据中，利用自然语言处理、信息检索、机器学习等技术抽取知识，转化而来的决策智慧的方法与过程。大数据智能涉及大数据、人工智能、自然语言处理三个相互支撑的关键技术。其中，大数据为大数据智能提供了关键的数据基础与大数据计算平台，是大数据智能的驱动力；人工智能为大数据提供了算法基础，是大数据智能的核心；自然语言处理直接面对数据中的语义内容，是人工智能“皇冠上的明珠”，直接决定大数据智能的广度与深度。

大数据、人工智能与自然语言处理是当前的研究热点，并被金融投资与商业开发广为追捧。习近平总书记在党的十九大报告中明确指出：“推动互联网、大数据、人工智能和实体经济深度融合。”2017 年 7 月 20 日，国务院印发《新一代人工智能发展规划》，明确了我国发展人工智能的战略目标，到 2030 年，人工智能核心产业规模超过 1 万亿元，带动相关产业规模超过 10 万亿元。自然语言处理(Natural Language Processing，NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP 的目标是让机器在理解语言上像人类一样智能，最终目标是弥补人类交流(自然语言)和计算机理解(机器语言)之间的差距。人工智能已经成为现代科学“皇冠上的明珠”，而自然语言处理号称是“人工智能皇冠上的明珠”。微软全球副总裁、著名人工智能专家沈向洋在中国计算机大会上明确表示：“懂语言者得天下。”自然语言理解已经成为人工智能研究与投资的重点，孕育着改变世界未来的产业机会。Gartner 发布的 2017 年商业智能和分析平台魔力象限报告中明确提出：到 2020 年，自然语言生成和人工智能将占现代商业智能平台标准特性的 90%，同时 50%的分析排队查询来自于搜索、自然语言处理或话音发起，或自动生成。2017 年始语音及自然语言处理是创业最火热的领域之一，占据人工智能领域的 11%。

大数据、人工智能或者自然语言处理的单一课程已经相对成熟，但很难适应当前科学技术的发展趋势，无法解决当前社交网络为代表的应用需求，将三者融会贯通的教材相对匮乏。笔者从 2016 年开始在北京理工大学为研究生一年级开设了必修课“大数据分析与应用”，尝试将大数据智能的教学体系进行完善，希望能出一本大数据智能的教材，这个想法得到北京理工大学“十三五”教材的立项支持。

本书共计14章。第1章对大数据智能进行了概述，并阐述关于大数据、人工智能与自然语言处理的主要体系架构，分享了作者关于大数据与人工智能的观点；第2章是大数据技术平台与架构，介绍了Hadoop、Spark等常用的云计算平台，介绍了机器学习、大数据可视化与大数据挖掘等常用工具；第3章介绍经典的机器学习与数据挖掘方法，包括管理规则挖掘、分类与聚类，最后为读者准备了常用的数据挖掘工具；第4章介绍经典的深度学习算法与平台，包括神经网络、循环神经网络(Recurrent Neural Network，RNN)、卷积神经网络(Convolutional Neural Network，CNN)、基于长短记忆的神经网络(Long Short-Term Memory，LSTM)、序列到序列模型(Sequence-to-Sequence)、注意力模型(Attention Model)以及生产对抗网络(Generative Adversarial Network，GAN)等，具体介绍了TensorFlow与PyTorch两个常用的深度学习平台；第5章介绍了作者研制的JZSearch大数据精准搜索引擎；第6章为汉语分词，主要介绍基于层次隐马尔可夫模型的浅层词法分析，以及基于深度学习的汉语分词算法；第7章介绍基于角色标注的命名实体识别；第8章介绍新词发现，主要基于上下文邻接熵与互信息的社交网络新词发现；第9章是文本分类与聚类；第10章重点推荐了基于Top N关键词的热点话题发现算法；第11章是当前关注比较多的情感分析算法，介绍情感词的自动发现与权重自动计算方法，详细介绍基于树模型的无监督情感分析算法与基于深度神经网络的短文本情感倾向性分析方法；第12章介绍文本摘要的关键技术；第13章与第14章都是大数据智能应用部分，其中第13章具体介绍了作者在警情大数据、网络赌博、微博挖掘、看图说话等多个实际的大数据应用项目，第14章是《红楼梦》前后作者分析、二手房房价、歌词生成等有意思的课程实践案例。

本书由张华平、商建云、郭涛与刘兆友合著。全书成果主要涉及张华平、商建云所在的大数据搜索与挖掘实验室，有些章节内容来自实验室近20年发表的学术论文与研究生毕业论文。其中，商建云完成了大数据应用等内容以及部分相关工作综述等的撰写工作；郭涛审阅了全书，并完成最后的统稿工作；刘兆友作为本书的助手，认真完成全部初稿的整理工作。在本书的写作与相关科研课题的研究工作中，得到多方面的支持与帮助，并采用了作者指导的研究生李清敏、张瑞琦、陈晓阳、李然、高莘、杨耀飞、李蕾、孙梦姝、王琦、季欣怡、张红瑾、Asif Khan等学生的毕业论文及发表的文章；同时采用了北京理工大学“大数据分析与应用”部分同学的课程作业，均在相应的内容部分进行了标识。中国科学院信息工程研究所的张卫博、周艳对深度学习等部分内容进行了修订，在此谨向这些文献的作者以及为本书提供帮助的老师、同仁和课题组成员致以诚挚的谢意和崇高的敬意。本书亦得到2018年国家自然科学基金(编号：61772075)、北京理工大学“十三五”教材的资助。

最后，感谢我的太太与家人的支持，家庭的无私支持成就了这部书的写作与出版，本书同时献给望望与杨杨。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

张华平

2019年3月

目　录

第1章

大数据智能概述

本章将从数据的智能演化过程、大数据、人工智能与自然语言处理4部分对大数据智能进行综合介绍。

1.1 数据的智能演化过程

我们先从数据的4个层次来洞察数据的智能演化过程，如图1-1所示。

数据：用于表示客观事物的未经加工的原始素材，在计算机系统中，数据以二进制信息单元0、1的形式表示。换句话说，只要占据硬盘空间的，都可以认定为数据，但是，数据有可能是杂乱无章的，无法表达实际含义。例如，笔者传给读者2MB的数据，存储到计算机后，如果读者完全不了解这个数据的格式，就无从了解其实际含义。

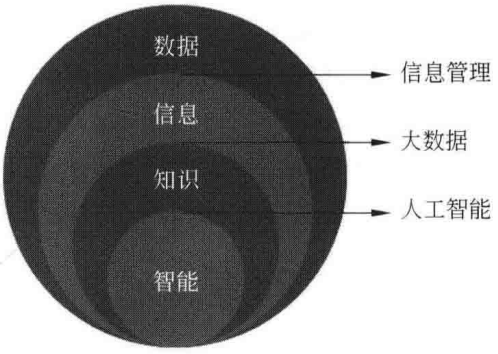


图1-1 数据的智能演化过程

信息：1948年，数学家香农在题为《通信的数学理论》的论文中指出："信息是用来消除随机不定性的东西"。继续上面的例子，如果这2MB数据是一个人的JPEG格式的图片，采用相应软件打开后，就可以看到其中包含11张图片。在这个过程中，这个数据就代表了有价值的信息，读者消除了一个不确定性的东西：数据的格式是什么？图片中的人究竟长什么样？这个过程一般可以称为信息管理。信息管理就是人对信息资源和信息活动的管理。信息管理的过程包括信息收集、信息传输、信息加工、信息存储与信息检索。

知识：人类在实践中认识客观世界的成果，知识是人类从各个途径中获得的经过提升总结与凝练的系统认识。知识本身是信息从量变到质变的产物，也是大数据的本质所在，即从多个表层信息中，利用大数据技术获取表层信息背后进一步的深度知识。依然以刚才的例子来说，11张普通的"表哥"照片单个来看都是低价值密度的表层信息，但当11张普通信息以一定的方式组织起来，并聚焦到其所戴手表时，很容易挖掘出背后的知识："表哥"拥有多款名贵手表，存在贪腐嫌疑。这个知识的价值远远超出了原始信息，这也是大数据的增值过程。

智能：指知识进一步归纳总结后的更普世的规律，可演化为更多的知识，用来指导客观实践。例如，我们可以从"表哥"的知识中，举一反三，利用已有的各类报道，自动挖掘更多人

的奢侈消费，从而获取电子侦察智慧，随时发现更多问题，防患于未然。这个过程，人们常称为人工智能过程。

1.2 大数据

1.2.1 大数据的概念

关于大数据如何定义，不同机构有不同的定义。其中，研究机构 Gartner 的定义：大数据是指需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。麦肯锡的定义：大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合。舍恩伯格·维克托的《大数据时代》中的定义：大数据是指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的全域分析方法。

无论哪种定义，都可以看出，大数据并不是一种新的产品，也不是一种新的技术，就如同21世纪初提出的“海量数据”的概念一样，大数据只是数字化时代出现的一种现象。那么海量数据与大数据的差别何在？从翻译的角度来看，“大数据”和“海量数据”均来自英文，Big Data 翻译为“大数据”，Large-scale Data 翻译为“大规模数据”，Very Large Data 翻译为“超大规模数据”，Massive Data 则翻译为“海量数据”。从组成的角度来看，海量数据包括结构化和半结构化的数据，大数据除此以外还包括非结构化数据和交互数据。大数据由海量交易数据、海量交互数据和海量数据处理三大主要技术趋势汇聚而成，其规模和复杂程度超出了常用技术，按照合理的成本和时限捕捉、管理及处理这些数据集的能力。

1.2.2 大数据的特征

大数据的特征包含 4 个层面。第一，数据体量巨大。从 TB 级别跃升到 PB 级别。第二，数据类型繁多。例如，网络日志、视频、图片、地理位置信息等。第三，价值密度低。以视频为例，在连续不间断的监控过程中，可能有用的数据仅仅有一两秒。第四，流动速度快。1秒定律，最后这一点和传统的数据挖掘技术有本质的不同。业界将大数据的特征归纳为4V，即 Volume、Variety、Value、Velocity。

1. 数据体量巨大（Volume）

大数据通常指 10TB（1TB=1024GB）规模以上的数据量。之所以产生如此巨大的数据量，一是由于各种仪器的使用，使人们能够感知到更多的事物，这些事物的部分甚至全部数据可以被存储；二是由于通信工具的使用，使人们能够全时段联系，机器—机器（M2M）方式的出现，使得交流的数据量成倍增长；三是由于集成电路价格降低，使很多东西都向智能化发展。

2. 数据类型繁多（Variety）

随着传感器种类的增多及智能设备、社交网络等的流行，数据类型也变得更加复杂，不仅包括传统的关系数据类型，也包括以网页、视频、音频、E-mail、文档等形式存在的半结构化的和非结构化的数据。

3. 价值密度低(Value)

数据量呈指数级增长的同时，隐藏在海量数据中的有用信息却没有以相应比例增长，反而使人们获取有用信息的难度加大。

4. 流动速度快(Velocity)

通常理解的数据流动速度是指数据获取、存储及挖掘有效信息的速度。由于现在处理的数据是 PB 级代替了 TB 级，“超大规模数据”和“海量数据”也有规模大的特点，数据是快速动态变化的。因此，形成流式数据是大数据的重要特征，数据流动的速度快到难以用传统的系统去处理。

大数据的 4V 特征表明其不仅仅是数据海量，对于大数据的分析将更加复杂，更追求速度，更注重实效。

大数据独立发展形成特有的市场化与规模化，也充分带动了其他行业与大数据的广泛、充分融合，从而推进大数据的全面落地。大数据从产业到行业的成熟将推动更多传统企业向科技智能化转型，将推进政府政务大数据发展，也将鞭策大数据行业在中国平稳落地。

大数据之大，还在于数据结构的有容乃大——它不再需要传统的数据库表格来整齐排列，几乎可以无所不包地记录、存储和计算各种规则的结构化数据和不规则的非结构化数据，于是便有了逐步演变为一个数字化世界的可能。

1.2.3　大数据带来的决策方式的革命

近半个世纪以来，人们经历了计算机时代计算方式的革命、互联网时代信息传播方式的革命、大数据时代决策方式的革命，如表 1-1 所示。

表 1-1　半个世纪技术的悄然革命

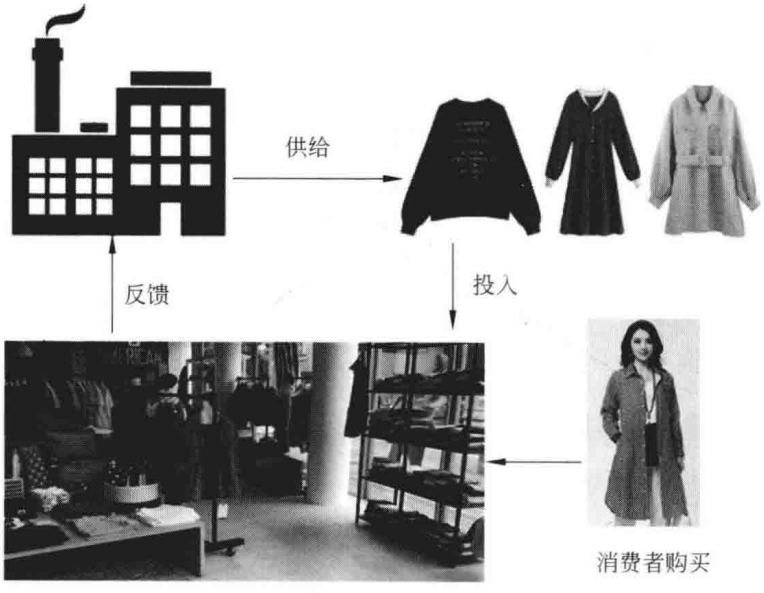
发 生 时 间	时　代	带来的技术革命
20 世纪 70 年代	计算机时代	计算方式的革命
20 世纪 90 年代	互联网时代	信息传播方式的革命
21 世纪初	大数据时代	决策方式的革命

计算方式大概经历了手工计算、机械计算、电子计算**等不同的发展阶段**。在运用计算机进行计算之前，所有的计算载体(如算盘等)只有计算功能，没有存储功能。计算机出现后，实现了既可以计算又可以记忆的功能，成为了人类的“外脑”。20 世纪 60 年代，数据一般存储在文件中，由应用程序直接管理；20 世纪 70 年代人们构建了关系数据模型，数据库技术为数据存储提供了新的手段；20 世纪 80 年代中期，数据仓库由于具有面向主题、集成、时变和非易失等特点，成为数据分析和联机分析的重要平台。

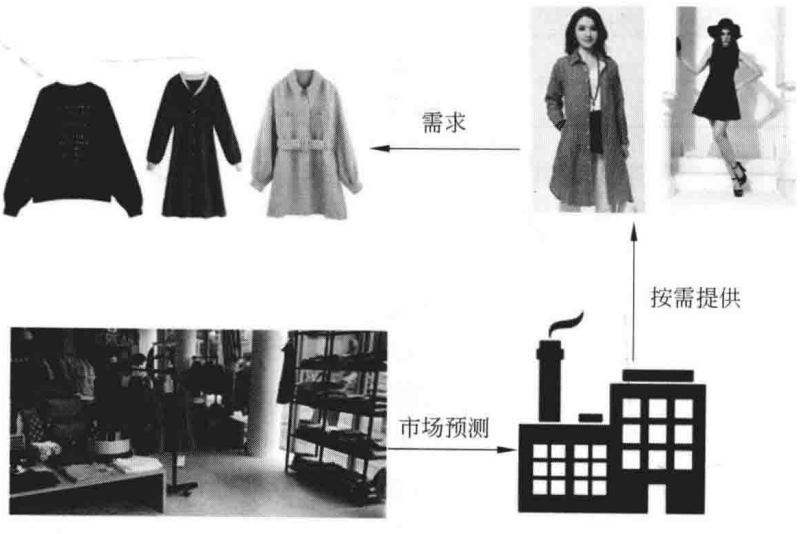
信息传播方式大概经历了口头传播、口头加书面(如报刊)传播、音频/视频和文字传播、新媒体传播**等不同发展阶段**。在互联网出现之前，信息传播都有一定的滞后性，如书信要邮递，报刊、电视节目都是先编辑或录制，均要经过一定的延时才会到达信息接收者手中。到了 20 世纪 90 年代，互联网出现，书信慢慢减少，电子邮件即刻到达。报刊、电视经历了多媒体的融合；拍客、博客的出现也使得人们从被动的信息接收者变成了信息的发布者和接收

者。每个人、每部手机都成为信息源。加上互联网上的人人连接、人物连接及物物连接相互产生了传播放大效应，带来了信息传播方式的革命。以前不会引起重大反响的事件，今天通过互联网的迅速传播，影响巨大。

决策方式经历了从直觉和经验到更加依赖数据和分析的转变。21 世纪初，借助计算机软件和硬件的不断发展，信息产生方式发生了变化，企事业单位和互联网上都积累并实时产生大量的数据，从而出现了具有 4V 特征的大数据。用数据分析的方法进行辅助决策是这个时代的特点。图 1-2(a)显示了以产品或服务提供者为中心的 BSFD(Business，Supply，Feedback，Duration)决策流程；图 1-2(b)显示了 CDPR(Consumer，Demand，Prediction，Realtime)以消费者为中心的决策流程。



(a) 以产品或服务提供者为中心的BSFD决策流程



(b) 以消费者为中心的决策

图 1-2　大数据分析进行辅助决策的方式

决策方式可以从决策主体、决策依据、决策机制和决策效率 4 个方面来描述。表 1-2 显示了大数据带来的决策方式变革。

表 1-2　大数据带来的决策方式变革

名　称	描　述
决策主体	以商业供给侧为中心→以消费需求侧为中心(Business→Consumer)
决策依据	以商业供给能力为依据→以用户需求为依据(Supply→Demand)
决策机制	反馈机制→预测机制(Feedback→Prediction)
决策效率	期间决策→实时决策(Duration→Realtime)

在大数据的背景下，决策主体从关注商务转移到关注消费者，决策依据不再是生产者的生产能力，而主要是消费者的需求，得益于大数据的支撑，决策机制从反馈机制转变为预测机制，决策效率也从半年、一年决策变为实时决策。

笔者的亲身经历充分证实了这一点。1994 年，笔者在日本看到路边有人用不同的砝码记录一天的人流情况，打听后得知他们将此数据带回去分析，用于预测今后的人流状况，进而指导人们的出行或决定道路的改建计划。2013 年，笔者在美国，开车在路上行驶，导航仪随时提供道路上的信息。如今，在北京大家可以看到实时的路况信息，实时指导人们的出行道路选择及为道路的改建计划提供参考。

可以看出，决策方式的变化来源于计算方式和信息传播方式的改变。

1.2.4　大数据面临的挑战及其对应的技术概览

鉴于大数据的特征和决策方式的革命要求，在处理大数据时会面临新的挑战，新的数据搜集、存储、传输和处理模式会随之产生，下面结合新的挑战论述大数据技术。

1. 数据量的指数级增长挑战数据存储能力

在大数据的决策中，90%基于数据，10%基于直觉，因此要收集大量的数据，包括实时数据。收集数据量巨大、数据种类繁多、数据潜在商业价值高，这些均要求使用专门的数据库技术和专用的数据存储设备。传统的数据库追求高度的数据一致性和容错性，缺乏较强的扩展性和较好的系统可用性，不能有效存储视频、音频等非结构化和半结构化的数据。目前，数据存储能力的增长远远赶不上数据的增长，设计最合理的分层存储架构成为信息系统的关键。

在相关技术中，比较具有代表性的是 Apache 软件基金会开发的 Hadoop。以 MapReduce 和 Hadoop 为代表的非关系数据分析技术，凭借其适合非结构处理、大规模并行处理和简单易用等优势，在互联网搜索和其他大数据分析技术领域取得重大进展，成为主流技术。

MapReduce 是 2004 年谷歌公司提出的用来进行并行处理和生成大数据的模型，是一种线性的、可伸缩的编程模型。其可扩展性得益于 Shared-nothing 结构、各节点间的松耦合性和较强的软件级容错能力。MapReduce 被设计在处理时间内解释数据，所以对非结构化、半结构化的数据处理非常有效。针对 MapReduce 并行编程模型的易用性，产生了多种