巍巍交大　百年书香
www.jiaodapress.com.cn
bookinfo@sjtu.edu.cn

丛书策划　张荣昌
责任编辑　王　清　孟海江
封面设计

大数据、云计算、人工智能、信息安全人才培养丛书
“互联网+”新形态一体化教材

大数据
- 大数据基础
- 数据可视化
- 数据清洗与治理
- Hadoop应用与开发
- 数据挖掘基础
- SEO搜索引擎优化
- MySQL数据库
- R语言程序设计
- Go语言程序设计

云计算
- 云计算基础
- 虚拟化与容器
- 云安全运维
- 网络工程与组网技术
- 现代通信技术
- 路由交换技术
- 无线网络技术
- 现代网络SDN技术
- 数据网组建与维护
- 局域网组建与维护
- 云数据中心架构与SDN技术

人工智能
- 人工智能基础
- 物联网基础
- 深度学习
- 机器学习

信息安全
- 信息安全基础
- Linux服务器安全高级运维
- Web安全与防御
- 防火墙技术与应用
- 计算机病毒与防范
- 数据存储与恢复
- 密码学基础
- 计算机网络安全运维
- 网络设备配置与综合实战
- 无线网络安全技术
- VPN虚拟专用网安全
- 终端数据存储与恢复
- 工控安全
- 渗透测试
- 恶意代码分析
- 网络空间安全态势感知
- 数据库安全技术
- 网络安全协议
- 企业级数据安全与灾备管理
- 信息安全法律法规
- 终端数据安全及防泄密

专业基础
- Linux操作系统基础
- 计算机网络基础
- Ubuntu服务器管理
- Windows Server 2019配置与管理
- 数据结构
- Python程序设计
- Java程序设计
- C语言程序设计
- C#程序设计
- Android程序设计
- XML基础教程
- JavaScript基础教程
- Web前端开发
- OpenStack应用与开发
- Spark应用与开发
- 静态网页设计与制作
- HTML5与JavaScript程序设计
- 数据库设计与应用
- 数据库应用基础
- UML建模与设计模式
- ERP原理与应用
- 综合布线

扫描二维码
关注上海交通大学出版社
官方微信

ISBN 978-7-313-24118-4
9 787313 241184 01>
定价：55.00元

河北省“十四五”职业教育省级规划教材

大数据基础

主编◎韩鸣飞　崔　佳　李志伟

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

河北省“十四五”职业教育省级规划教材

大数据基础

DASHUJU JICHU

主编◎韩鸣飞　崔　佳　李志伟

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

本书从初学者的角度出发，分9章介绍了大数据的基础知识。具体内容包括大数据技术概述、Hadoop的安装和配置、HDFS概述、HDFS API编程、Hadoop分布式计算框架、MapReduce API编程、Hadoop数据建模、Spark概述、Spark Streaming编程。

本书可作为高等院校计算机应用、大数据技术及相关专业的教材，也可作为大数据开发入门者的自学用书，可快速提高开发技能。

图书在版编目（CIP）数据

大数据基础 / 韩鸣飞，崔佳，李志伟主编 . — 上海：上海交通大学出版社，2020（2023重印）

ISBN 978-7-313-24118-4

Ⅰ. ①大… Ⅱ. ①韩… ②崔… ③李… Ⅲ. ①数据处理—高等学校—教材 Ⅳ. ① TP274

中国版本图书馆CIP数据核字（2020）第242834号

大数据基础
DASHUJU JICHU

主　　编：韩鸣飞　崔佳　李志伟
出版发行：上海交通大学出版社
邮政编码：200030
印　　制：北京荣玉印刷有限公司
开　　本：889 mm × 1194 mm　1/16
字　　数：357千字
版　　次：2020年12月第1版
书　　号：ISBN 978-7-313-24118-4
定　　价：55.00元

地　　址：上海市番禺路951号
电　　话：6407 1208

经　　销：全国新华书店
印　　张：15

印　　次：2023年2月第3次印刷

前言

随着我国将大数据战略提升到国家战略高度，将大数据视为经济发展和转型的重要科技依据，越来越多的企业组织已将数据视为重要资产，着手开展数字化转型的系列举措，以期全面提升客户体验，推动经济增长。大数据时代，数据资产已成为一个企业的核心竞争力。但数据在手而不会运用它，就毫无价值。在当下企业数字化浪潮中，数据是企业转型的基础元素，如何将企业不同业务、类型的数据应用起来，推动企业运营，增加收入、降低成本、提高效率、控制风险等，是企业现在面临的难点。大数据技术助力企业走出数据分析的盲区，实现数据化管理与运营，经过元数据管理系统的比对、校验、转换得到一致的数据，从而进行数据整合，根据不同的需求提供差异化服务，并渗透到各个行业领域，已逐渐成为一种生产要素并发挥着重要作用。

本书以大数据工程师的实际工作为基础，从初学者的角度出发，介绍了大数据从环境搭建到使用，再到高级应用的知识。全书共 9 章，包括大数据技术概述、Hadoop 的安装和配置、HDFS 概述、HDFS API 编程、Hadoop 分布式计算框架、MapReduce API 编程、Hadoop 数据建模、Spark 概述、Spark Streaming 编程。其中，基础部分重点介绍数据组织、重要数据结构、大数据协同技术以及大数据存储技术等内容；核心部分重点介绍 MapReduce 分布式编程、数据采集与预处理、数据挖掘与分析技术、Spark 分布式内存计算框架等内容。

本书在编写上具有以下特点。

（1）通过章前的“学习目标”“知识导图”“本章导入”明确该章要学习的内容，使学生做好学习的准备；通过文中的“说明”“提示”“注意”“知识拓展”“知识链接”等模块，丰富知识内容，提高学生的学习兴趣。

（2）在精练语言的基础上，充分利用图、表尽量形象地描述知识点，帮助学生更好地理解所学内容。内容由浅及深，循序渐进，符合学习者的认知规律。

（3）计算机技术发展很快，本书着重于当前最新知识和主流技术的讲解，使学生所学知识和技术都与行业联系密切，做到学有所用。

此外，本书作者还为广大一线教师提供了服务于本书的教学资源库，有需要者可致电 13810412048 或发邮件至 2393867076@qq.com。

本书由具有多年大数据工作经验的一线工程师和具有多年大数据专业教学经验的任课教师联合编写而成，书中知识点结合具体实例和程序讲解，便于读者理解和掌握。本书可作为高等院校计算机应用、大数据技术及相关专业的教材，也可作为大数据开发入门者的自学用书，可快速提高读者的开发技能。

目录

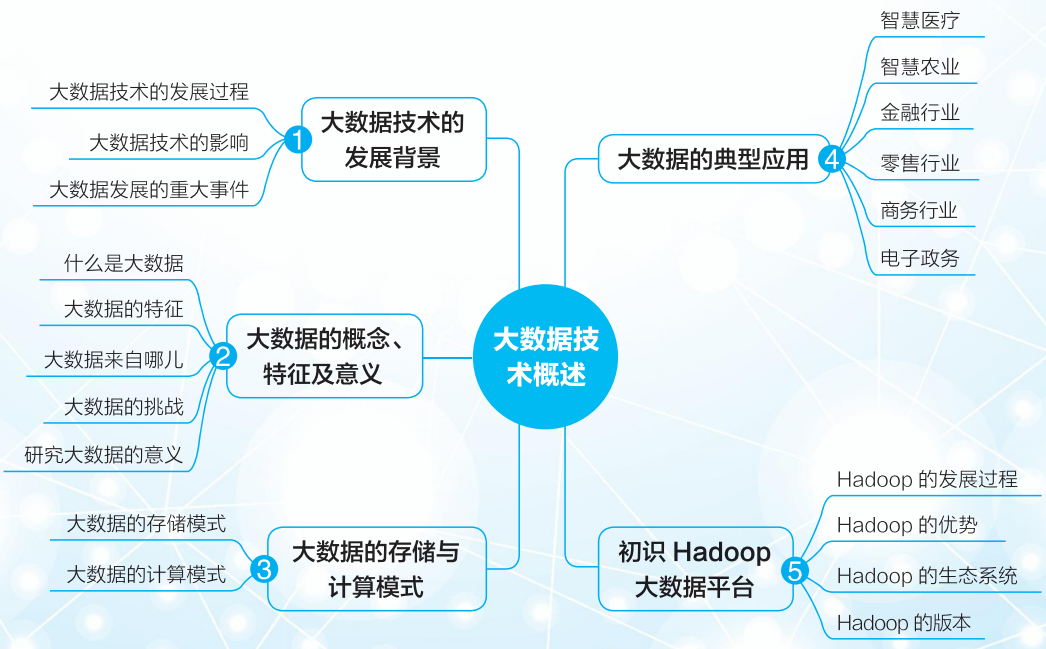
第 9 章　Spark Streaming 编程 / 211

第 1 章 大数据技术概述

学习目标

1. 了解大数据的基本概念。
2. 了解大数据的发展背景。
3. 了解大数据的存储与计算模式。
4. 掌握大数据的应用场景。

知识导图



笔记

本章导入

大数据指的是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。在商业数据、科学数据和网页数据等促使数据种类日益增多、规模呈爆炸式增长的当下，作为信息技术和计算方法迅速发展的必然产物，大数据已经开启了属于自己的时代。大数据已成为继物力和人力资源之后的又一重要资源，在社会发展过程中发挥着不可替代的作用。

1.1 大数据技术的发展背景

1.1.1 大数据技术的发展过程

大数据（Big Data）一词最早出现在 Apache Org 的开源项目 Nutch 中，当时科学家们在更新网上搜索索引的同时，将大量的数据集合描述为大数据。1980 年，当代著名思想家阿尔文·托夫勒在其《第三次浪潮》一书中将大数据描绘为信息社会的全新篇章和重要组成部分。可以看出，生产率的增长和消费的盈余激发了人们对于数据的海量挖掘和大量运用，大数据时代也随之到来了。

从古至今，处理各种不断增长的数据都是人类社会的难题。进入信息社会以来，海量增加、日益革新的数据处理问题更成为人们亟待突破的壁垒。大数据的现代发展历史可追溯到美国统计学家赫尔曼·霍尔瑞斯，他发明了一台电动机播放器来识别卡片进行数据统计，这台机器成功应用于 1890 年的人口普查工作上，耗时一年完成了预计需要 8 年的工作，成为推动全球数据处理的全新出发点。在 1943 年第二次世界大战期间，英国组织工程师发明了一种能够处理大规模数据的机器，并采用了第一台可编程的电脑进行了计算工作，旨在迅速解开纳粹设置的密码。1961 年，美国国家安全局（NSA）首先使用计算机来收集信号并自动处理信息，采用数字化的方式处理磁盘信息。20 世纪 60 年代，英国计算机科学家蒂姆·伯纳斯·李设计了超文本系统，并将其命名为万维网。自此，互联网在全世界范围内向人们提供信息共享服务。

1965 年，英特尔的创始人戈登·摩尔（Gordon Moore）通过计算机硬件研究得出摩尔定律：同面积的芯片每隔 1~2 年可容纳的晶体管数翻倍，微处理器性能也将翻倍，或成本降低一半。随着技术的发展和迭代，信息产品各种功能增强伴随着设备体积的变小，内存成本得到了极大的降低，可以使用较低的费用保存海量的数据。1988 年美国科学家马克·韦泽（Mark Weiser）提出，普适计算可以使各种微型计算设备无论何时何地都可以获得和处理数据。时至今日，智能手机、传感器、无线频率识别（RFID）标签、可穿戴设备等都可以实现数据的自动采集，为大数据的实现提供了物理基础。1997 年，美国科学家大卫·埃尔斯沃斯和迈克尔·考克斯使用“大数据”来描述超级计算机产生的大量超出主内存的信息，数据集甚至突破了远程磁盘的承载能力。

大数据时代的技术价值主要体现在数据挖掘上，即通过某种算法自动分析大量数据，从而揭示数据中隐藏的规则和趋势，并从这些数据中发现新的价值，为下一步行动提供参考和支撑。在商业领域，目前的信息技术可以记录一个产品的流向和所有消费者的情况，

笔记

然后在数据挖掘的基础上，针对顾客提供个性化的消费和服务，这样一种高度个性化消费的时代已经不陌生。数据挖掘成为越来越重要的分析预测工具——不需要撰写问卷，也不需要一一调查，费用低廉。更重要的是，这种分析没有停滞性，数据挖掘的应用使抽样技术将被降格为辅助工具。数据挖掘的优越性集中体现在“量大、多源、实时”三个方面，与数据挖掘相比，机器学习成为大数据技术的尖端和热点，其算法并不是固定的，而是带有自调适参数的，即随着计算、挖掘次数的增多，机器学习通过自动调整的算法参数，使预测结果更加准确，通过在机器上投入大量的数据，让机器变得像人一样通过学习来实现自我改进和提高，这也是这一技术命名为“机器学习”的原因。也就是说，除了数据挖掘和机器学习外，数据分析、使用技术已经十分成熟，并在数据仓库、多维联机分析处理（Multidimension OLAP）、数据可视化、基于内存分析（In-Memory Analytics）等技术的支撑下逐渐成熟化和体系化。

从 2004 年开始，Facebook、Twitter 等社交媒体相继登场，互联网开始成为实时沟通和合作的媒介。全世界网民成了数据生产者，直接促使了人类历史上最大的数据爆炸。社交媒体中产生的数据大部分都是非结构化的数据，使处理难度大大提升。乔治敦大学的教授李塔鲁（Kalev Leetaru）于 2012 年研究了 Twitter 上的数据量。研究结果表明，Twitter 上仅一天就创造了 80 亿个单词信息量，超过了过去的 50 年里《纽约时报》创造的单词量总和——30 亿个。不难发现，如今一天产生的数据总量相当于《纽约时报》100 多年产生的数据总量。回顾人类信息社会的发展历史，1966 年提出的摩尔定律成为形成大数据现象的物理基础和关键环节。正是由于晶体管体积的变小，低成本才得以实现。1989 年流行的数据挖掘技术是使大数据形成“大价值”的关键。2004 年出现的社交媒体将全世界的人都变成了潜在的数据生成工具，这是“大容量”形成的主要原因。2008 年末，“计算社区联盟”(Computing Community Consortium）在《大数据计算：在商务、科学和社会领域创建革命性突破》报告中，提醒人们不要将视野仅仅局限在技术领域，要在更广泛的领域中寻找大数据的社会价值，从而找到更多的新用途和独创性的新见解。社会领域的计算被许多学者称为“社会计算”，对社会领域数据的计算、对知识和关系的挖掘，不仅有效地加强了社会治理能力，还能够产生商业价值。

美国在 2012 年 3 月的政府报告中明确要求每个下属的联邦机构都要制定一个“大数据”发展战略，奥巴马政府首先宣布投资 2 亿美元，立刻启动大数据研究与发展项目。2013 年，英国政府投资 1.89 亿英镑发展大数据技术，2014 年又投入了 7300 万英镑。2013 年开始，日本也将开放公共数据和建设大数据作为新的 IT 国家战略，以提升数据处理能力和计算能力。中国于 2014 年开始为大数据关键技术、应用、产业和政策环境等核心要素提供支撑。2020 年，工业和信息化部发布相关通知，明确要求加快国家工业互联网大数据中心建设，鼓励各地建设工业互联网大数据分中心，利用 5G、人工智能、区块链、增强现实 / 虚拟现实等新技术支撑大数据中心发展。从整体上看，处理大规模信息的现实需求推动了大数据相关技术的快速发展。最初，国家安全是其萌生和发展的主要推进动力，而随着超级计算机的发明，大数据的存储和处理技术、大数据分析算法的研究和开发最终推动了大数据在教育、金融、医疗等各个领域的应用。

笔记

1.1.2 大数据技术的影响

“当世界开始迈向大数据时代时，社会也将经历类似的地壳运动”。大数据时代推动了以数据为中心的一系列观念、技术、应用的巨大变革，这种广泛性、根本性的变革并不是某种技术、话语和平台的简单集合，它将必然引起人类生产、交往方式以及社会管理方式、结构的变革，使社会生活发生剧烈激荡。现如今，大数据并不是遥不可及的技术，它的影响力和作用力正在迅速触及社会的每个角落，所到之处，既有对传统行业的冲击，也有基于大数据技术的产业革新，大数据技术的影响力以及作用力是每个人都能深切感受到的。

在大数据时代到来后，工业时代所形成的社会结构和政治形态都将重构。目前，智能终端、云计算、宽带网络丰富了基础设施的意义。此前，基础设施仅仅指铁路、道路、机场、港口等物理世界的建筑。土地、劳动力、资本曾经是核心生产要素，而今最有价值的资产非数据莫属。旧有的以产业链为基础的分工体系和市场体系会因为资源、制造基地和市场在时空层面发生隔离和不平衡而产生高额的费用，并受到多重限制。大数据领域每年都会有大量的新的技术上市，大规模协作和共同作业方式也在得到持续推进，大数据保存、处理、分析或可视化的有效手段在不断增加。通过将大规模数据中隐藏的信息和知识挖掘出来，大数据技术为人类社会经济活动提供保障，在提高各个领域的运行效率的同时提升了整个社会经济的集约化程度。

在社会管理方面，政府利用大数据技术对经济形式和就业信息进行全面了解，由于就业信息数据量十分庞大，为国家把控就业形式并制定相关的政策来解决相应的问题带来了一定困难，此时，大数据处理与分析技术就可以解决这一难题。在对全国就业信息进行采集后，将所有采集到的数据利用分布式文件系统存储在机器中，并利用相关的经济学原理，分析采集到的数据之间的逻辑关系，在一些优秀的大数据平台如 Hadoop、Spark 等上编写相应的数据分析程序，最后将程序部署到存储数据的机器上，数据分析开始后，将最终的结果汇总，得出相应的结论。

大数据处理和分析技术与传统抽样调查的不同之处在于，大数据技术在分析、统筹的基础上宏观挖掘所有数据的内涵和信息，不只是简单地分析某一部分数据，因此，通过大数据处理和分析得出的结论比传统的抽样调查结果更准确。更重要的是，通过对全国就业信息的分析，除了得出国家经济形式相关的结论以外，其中的数据还可以用来研判劳动力市场数据，预测哪些行业将会给就业者带来无限机遇。除此之外，每个人在日常生活中产生的各种各样的数据（如出行信息、购物信息、社交信息、身体健康状况等信息）都构成大数据的一个环节并成为处理和分析的对象。

在医疗方面，大数据处理与分析技术有着非常广泛的应用，如样本收集、临床决策系统和远程病人监护等，通过全面分析病人特征数据和疗效数据，比较多种干预措施的有效性，进而针对特定病人选择最佳治疗途径。对于患者来说，医疗提供者不同，医疗护理方法和效果也不同，成本也存在很大的差异。利用大数据技术合理分析患者的体征数据、费用数据和治疗效果数据等，并形成大型数据集，有助于医生在临床上采取最合理、最有效的治疗方法。

1.1.3 大数据发展的重大事件

作为一个 IT 行业术语，现如今的大数据已被赋予了重要的战略意义。在大数据发展

历程上，诸多重大事件影响了大数据发展的走向以及其对人们生活的改变。本小节通过对其的梳理，明确大数据发展近十年的历程和走向。

笔记

2005 年，Hadoop 项目诞生，其最初只是雅虎公司用来解决网页搜索问题的一个项目，后来因其技术的高效性，被 Apache Software Foundation 公司引入并成为开源应用。Hadoop 本身不是一个产品，而是由多个软件产品组成的一个生态系统，这些软件产品共同实现全面功能和灵活的大数据分析。从技术上看，Hadoop 由两项关键服务构成：采用 Hadoop 分布式文件系统（HDFS）的可靠数据存储服务，及利用一种叫作 MapReduce 技术的高性能并行数据处理服务。这两项服务的共同目标是，提供一个使对结构化和复杂数据的快速、可靠分析变为现实的基础。Hadoop 的独特之处在于它的编程模型简单，用户可以很快地编写和测试分布式系统。2008 年以来，Hadoop 逐渐被互联网企业广泛接受，这一开源的生态系统已成为大数据处理的主流和事实标准。

2008 年末，“大数据”这一概念得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不再局限于数据处理的机器，并提出：大数据真正重要的是新用途和新见解，而非数据本身。此组织可以说是最早提出大数据概念的机构。

2009 年，印度政府建立了用于身份识别管理的生物识别数据库，联合国全球脉冲项目已研究了如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病暴发之类的问题。同年，美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供各种各样的政府数据。该网站超过 4.45 万的数据集被用于保证一些网站和智能手机应用程序跟踪从航班到产品召回再到特定区域内失业率的信息，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。此外，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系，致力于改善在互联网上获取科学数据的简易性。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到：“世界上有着无法想象的巨量数字信息，并以极快的速度增长。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到了这种巨量信息的影响。科学家和计算机工程师已经为这个现象创造了一个新词汇：‘大数据’。库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4 TB（约 2 亿页文字量）的数据量，并在美国著名智力竞赛电视节目《危险边缘》（*Jeopardy*）上击败两名人类选手而夺冠。后来纽约时报认为这一刻为一个“大数据计算的胜利”。同年 5 月，全球知名咨询公司麦肯锡（McKinsey&Company）全球研究院（MGI）发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，大数据开始备受关注，这也是专业机构第一次全方位地介绍和展望大数据。报告指出，大数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。报告还提到，“大数据”源于数据生产和收集的能力和速度的大幅提升——由于越来越多的人、设备和传感器通过数字网络连接起来，产生、传送、分享和访问数据的能力也得到彻底变革。

2011 年 12 月，工信部发布的物联网“十二五”规划中，信息处理技术作为 4 项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，

笔记

这都是大数据的重要组成部分。

2012 年 1 月，大数据成为瑞士召开的达沃斯世界经济论坛的主题之一，会上发布的报告《大数据，大影响》宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。奥巴马政府同时宣布投资 2 亿美元于大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭，在次日的电话会议中，美国政府对数据的定义是“未来的新石油”，大数据技术领域的竞争事关国家安全和未来，并表示，国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用的能力；国家数字主权体现在对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。2012 年 4 月，美国软件公司 Splunk 在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低迷、股市持续震荡的大背景，Splunk 首日的突出交易表现尤其令人们印象深刻，首日即暴涨了一倍多。Splunk 是一家领先的提供大数据监测和分析服务的软件提供商，成立于 2003 年。Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据布局。7 月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明了在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求：例如，通过对价格关注和更好服务的渴望，个人提供数据和众包信息，并对隐私和退出权利提出需求；公共部门出于改善服务，提升效益的目的，提供了诸如统计数据、设备信息、健康指标、税务和消费信息等，并对隐私和退出权利提出需求；私人部门出于提升客户认知和预测趋势目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。

2012 年，为挖掘大数据的价值，阿里巴巴集团宣布在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台——“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。阿里巴巴集团希望通过分享和挖掘海量数据，为国家和中小企业提供价值。此举是国内企业最早将大数据提升到企业管理层高度的一次里程碑，阿里巴巴也是最早提出通过数据进行企业数据化运营的企业之一。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”为主题发布了《全球信息技术报告》(第 13 版)。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要，在接下来将对数据保密和网络管制等议题展开积极讨论。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到大数据在推动经济发展、改善公共服务，增进人民福祉，乃至保障国家安全方面的重大意义。2014 年 5 月，美国白宫发布了 2014 年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。此外，在 2014 年，“大数据”一词首次出现在我国国务院的政府工作报告中。其中指出，要设立新兴产业创业创新平台，在大数据等方面赶超先进，引领未来产业发展。“大数据”旋即成为国内热议词汇。

笔记

2015 年，国务院正式印发《促进大数据发展行动纲要》(以下简称为《纲要》),《纲要》明确指出，将推动大数据发展和应用，在未来 5~10 年打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态，标志着大数据正式上升为国家战略。

2016 年，大数据“十三五”规划出台，涉及的内容包括推动大数据在工业研发、制造、产业链全流程各环节的应用；支持服务业利用大数据建立品牌、精准营销和定制服务等。

2018 年，中国大数据技术大会在北京举办。会上指出，大数据从辅助到引领，从热点到支点，已经成为所有新旧技术、新旧模式的必备基础。2019 年，健康医疗、城镇化智慧城市、金融、互联网电子商务、制造业工业大数据在最令人瞩目的应用领域创造佳绩；在城市数据、视频数据、话音数据、互联网公开数据以及企业数据、人体数据、设备调控、图形图像数据类型方面均取得应用和技术突破。2020 年，工信部发布了《关于推动工业互联网加快发展的通知》，面对日益旺盛的工业计算需求，信息化、数字化、智能化应用市场亟待建立。同时业界人士认为，AI、5G、区块链等场景化应用可以为大数据发展打开新的成长空间。

1.2 大数据的概念、特征及意义

1.2.1 什么是大数据

大数据这个术语最早应用于 Apache Org 的开源项目 Nutch，用来表达批量处理或分析网络搜索索引产生的大量数据集。谷歌公开发布 Map Reduce 和 Google File System（GFS）之后，大数据不仅包含数据的体量，也强调数据的处理速度。在数据分析领域，大数据是前沿技术，大数据以及数据仓库、数据分析、数据安全、数据挖掘是 IT 行业时下最火爆的词汇，大数据的商业价值已经成为信息行业争相追逐的焦点。大数据包括各种互联网信息，更包括各种交通工具、生产设备、工业器材上的传感器随时随地进行测量，不间断传递着的海量的信息数据。 利用新处理模式，大数据具有更强的决策力和洞察力，能够优化流程，实现高增长率，处理海量的多样化信息资产。归根结底，大数据技术可以快速处理不同种类的数据，从中获得有价值的信息，处理速度快，只有快速才能有实际用途。随着网络、传感器和服务器等硬件设施的全面发展，大数据技术促使众多企业融合自身需求，创造出难以想象的经济效益，实现巨大的社会价值，商业价值高，各行各业利用大数据产生极大的增值和效益，表现出前所未有的社会能力，而绝不仅仅只是数据本身。所以，大数据可以定义为在合理时间内采集、处理大规模资料，帮助使用者更有效决策的社会过程。一般意义上讲，大数据指数据量巨大，通常认为数据量在 10 TB~1 PB（1 TB=1024 GB，1 PB=1024 TB）以上，达到“太字节”（2^{40}）数量级的高速、实时的数据流。在业内人士看来，大数据具有“4V+1C”特征，即数据量大（Volume）、多样（Variety）、快速（Velocity）、价值密度低（Value）以及复杂度（Complexity）。中国信息通信研究院在《大数据白皮书（2014）》中指出，作为新资源、新工具和新应用的综合体，大数据具有“资源、技术、应用”三个层次。

笔记

早在 2001 年，关于大数据这一概念的定义就已经初现端倪。META 集团的分析师道格·莱尼在研究报告中，将数据增长带来的挑战和机遇定义为三维式，即数量（Volume）、速度（Velocity）和种类（Variety）的增加。虽然这一描述最先并不是用来定义大数据的，但是诸如 IBM 和微软等企业在此后的数年间仍然使用这个“3Vs”模型来描述大数据。数量，意味着生成和收集大量的数据和日趋庞大的数据规模；速度，是指大数据的时效性，数据的采集和分析等过程必须迅速及时，从而最大化地利用大数据的商业价值；种类，表示数据的类型繁多，不仅包含传统的结构化数据，更多的还有音频、视频、网页、文本等半结构和非结构化数据。普遍来看，大数据是指无法在有限时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合。但在实际应用中，科技企业、研究学者、数据分析师和技术顾问由于各自的出发点和落脚点不同，对大数据产生了不一样的理解。通过对不同行业定义的比较了解，人们可以更好地理解大数据在技术表象背后蕴藏的深刻内涵。2010 年，Hadoop 组织将大数据定义为“普通的计算机软件无法在可接受的时间范围内捕捉、管理、处理的规模庞大的数据集”。在此定义的基础上，2011 年 5 月，全球著名咨询机构麦肯锡公司发布了《大数据：下一个创新、竞争和生产力的前沿》报告，并在其中对大数据的定义进行了扩充，认为大数据是指其大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集。该定义有两方面内涵，首先，符合大数据标准的数据集大小是变化的，会随着时间推移、技术进步而增长；其次，不同部门符合大数据标准的数据集大小会存在差别。根据麦肯锡的定义，不难发现，数据集的大小并不是判断、检验大数据的唯一标准，数据规模的扩大和传统数据库技术管理能力的难以覆盖也应被视为大数据的重要特征。

纵观人类社会的发展史，科技始终是为了满足人的需求及意愿而发展和进步的。在大数据时代，通过对数据的挖掘和分析处理，技术可以助力于人的决策过程，可以说，大数据技术的出现，开启了一次社会、经济、科技的重大转型。在 IT 时代，由于技术贯穿了数据生产、传输、储存等各个环节，技术（Technology）始终是社会关注的核心和重点；现如今，数据的价值逐步凸显，信息（Information）的重要性日益提高，数据牵引技术进步的时代将会到来。大数据不仅助推了经济发展、改变了人类生活，也为思维的创新、社会的变革贡献了独特的力量。

1.2.2 大数据的特征

当前，全球大数据正进入加速发展时期，技术产业与应用创新不断迈向新高度。大数据通过数字化丰富要素供给，通过网络化扩大组织边界，通过智能化提升产出效能，不仅是推进网络强国建设的重要领域，更是新时代加快实体经济质量变革、效率变革、动力变革的战略依托。大数据的概念并不是凭空出现的，它的前身是海量数据。但两者之间又有所区别——海量数据强调的是数据量的规模之大，并没有对其特征性进行定义；而大数据的概念包含了大数据的体积、传播速率、特征等内容。目前，业界对大数据并没有统一的定义，但却有 4 个被广泛接受的特征，如图 1-1 所示。

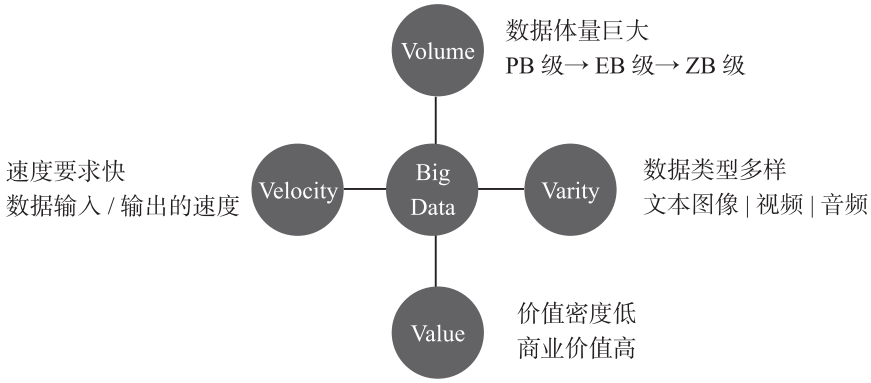


图 1-1　大数据“4V”特征

1. 数据规模大

大数据的特征首先体现为“大”，由于互联网络的广泛应用，使用网络的用户、企业、机构增多，数据获取、分享变得相对容易。用户可通过网络非常方便地获取数据，通过有意的分享和无意的点击、浏览可快速地提供大量的数据；此外，随着各种传感器数据获取能力的大幅度提高，使得人们获取的数据越来越接近原始事物本身，描述同一事物的数据激增；集成电路价格的不断降低也是大量数据得以保存下来的原因之一。随着时间的推移，存储单位从过去的 GB 级别到 TB 级别，乃至现在的 PB、EB 级别，而信息技术的高速发展，还在不断推进数据爆发性增长。社交网络（微博、推特、Facebook）、移动网络、各种智能工具、服务工具等都成为数据的来源。淘宝网逾 7 亿的会员每天产生的商品交易数据约 20 TB；Facebook 约 20 亿的月活跃用户每天产生的日志数据超过 300 TB，迫切需要智能的算法、强大的数据处理平台和新的数据处理技术，来统计、分析、预测和实时处理如此大规模的数据。

2. 数据种类繁多、复杂多变

随着传感器种类的增多及智能设备、社交网络等的流行，数据种类也变得更加复杂，其包括结构化数据、半结构化数据和非结构化数据。其中，10% 是结构化数据，存储在数据库中；90% 是非结构化数据，与人类信息密切相关。广泛的数据来源决定了大数据形式的多样性。任何形式的数据都可以产生作用，目前应用最广泛的就是推荐系统，如淘宝、网易云音乐、今日头条等平台都会通过对用户的日志数据进行分析，从而进一步推荐用户喜欢的东西。日志数据是结构化明显的数据，还有一些数据结构化不明显，如网络日志、图片、音频、视频等，这些数据因果关系弱，就需要人工对其进行标注。

3. 数据处理速度快

新时代人们从信息的被动接受者变成了主动创造者，数据从生成到消耗，时间窗口非常小，可用于生成决策的时间非常短。生活中每个人都离不开互联网，也就是说，每个人每天都在向大数据提供大量的资料。大数据的产生非常迅速，主要通过互联网传输，并且这些数据是需要及时处理的，而花费大量资本去存储作用较小的历史数据是非常不划算的，对于一个平台而言，也许保存的数据只有过去几天或者一个月之内，再远的数据就要及时清理，不然代价太大。基于这种情况，大数据对处理速度有非常严格的要求，服务器中大量的资源都用于处理和计算数据，很多平台都需要做到实时分析。数据无时无刻不在产生，速度更快的显然会在竞争中抢占先机。

笔记

4. 数据价值密度相对较低

数据价值密度相对较低是大数据的核心特征之一。在社会日常运转所产生的数据中，有价值的数据所占比例很小。也就是说，数据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相应比例增长。恰恰相反，从海量数据中挖掘稀疏珍贵的信息，挖掘大数据的价值类似沙里淘金。例如，商场的监控视频，连续数小时的监控过程中有可能有用的数据仅仅只有几秒钟。然而，相比于传统的小数据，大数据最大的价值在于通过从大量不相关的各种类型的数据中，挖掘出对未来趋势与模式预测分析有价值的数据，并通过机器学习方法、人工智能方法或数据挖掘方法深度分析，发现新规律和新知识，并运用于农业、金融、医疗等各个领域，从而最终达到改善社会治理、提高生产效率、推进科学研究的目的。

1.2.3 大数据来自哪儿

某种程度上来说，现代社会为人们创造的是一个数据驱动的环境，不论技术怎么更新换代，数据都有着不可替代的地位，而大数据的核心就是数据本身，没有数据做支撑的大数据平台就是一个空壳。具体来看，无论是公司内部的数据还是外部的数据都可以构成大数据平台的来源数据，如数据库、日志、前端埋点、爬虫等。

在大数据技术风靡起来前，关系型数据库系统（Relational Database Management System，RDMS）是主要的数据分析与处理途径，发展到今天数据库技术已经相当完善。当大数据出现的时候，行业就在考虑能否将数据库处理数据的方法应用到大数据中，于是 Hive、Spark SQL 等大数据 SQL 产品就这样诞生。虽然出现了 Hive 大数据产品，但是在生产过程中业务数据依旧使用 RDMS 进行存储，这是因为产品需要实时响应用户的操作，在毫秒级完成读写操作，而大数据产品不是应对这种情况出现的。一般来说，业务数据使用实时和离线采集数据来将数据抽取到数据仓库中，然后再进行后续数据处理和分析，一些常见的数据库导入工具有 Sqoop、Datax 和 Canal 等。

日志系统将系统运行的每一个状况信息都使用文字或者日志的方式记录下来，这些信息可以被理解为业务或是设备在虚拟世界的行为的痕迹，通过日志对业务关键指标以及设备运行状态等信息进行分析。以大数据日志收集常用的工具 Apache Flume 为例，Agent 主要有三个组件：Source、Channel、Sink。Source 主要负责收集数据，封装数据为事件（Event）后发送到 Channel，数据来源可以是企业服务器、文件系统、云、数据存储库等。通常，读取速度比写入速度快，因此，人们需要一些缓冲区来匹配读写速度差异。基本上，Channel 提供一个消息队列的功能，用于存储 Source 发送的事件，对事件进行消息排序，发送到 Sink。Sink 从 Channel 收集数据，将数据输送至大数据存储设备，如 HDFS、Hive、Hbase 等。Sink 也可以作为新的 Source 输入源，两个 Agent 进行级联，根据需求开发各种处理结构。

现在的互联网公司越来越关注转化、新增、留存，而不是简单地统计 PV、UV。当下，分析数据来源通过埋点获取，前端埋点分为三种：手工埋点、可视化埋点、自动化埋点。手工埋点指的是前端需要返回数据的位置调用写好的埋点 SDK 的函数，按照规范传入参数通过 HTTP 方式传入后代服务器中。这种方式可以下钻并精准采集数据，但工程量巨大。自动化埋点也称为无埋点，即是无须埋点，在全部位置都设置埋点，对用户所有操

作进行采集，这种方式将通过统一的 SDK 返回数据，再选择需要的数据进行分析，这种方式将加大服务器的压力，采集许多不需要的数据，浪费资源。在实践中，可以采用对部分用户或者部分简单操作页面进行全埋点采集。可视化埋点是介于手工埋点和自动化埋点之间的方式，通过可视化交互设置埋点，可以理解为人为干预的自动化埋点形式。埋点问题不应该通过单一的技术方案来解决，在不同场景下我们需要选择不同的埋点方案。

笔记

另外，爬虫也是数据的重要来源之一。爬虫的数据成为公司重要战略资源，通过获取同行的数据跟自己的数据进行支撑对比，管理者可以更好地做出决策。而且越难爬虫获取的竞争对手的数据，对于公司来说越有价值。数据采集本身不是目的，只有采集到的数据可用、能用，且能服务于最终应用分析的数据采集才是根本。数据采集的科学性决定了这个数据分析报告是不是有使用价值，只有当数据采集具有科学性，客观、严密的逻辑性时，建立在这样的数据分析基础之上出来的结论才具有现实的价值和意义。

1.2.4 大数据的挑战

在大数据时代已经到来的当下，人们不能仅仅看到大数据的价值和广阔前景，也应意识到在其管理和维护的过程中潜在的问题和面临的挑战。面对数据量大、数据形式多样化的海量非结构化数据，如何存储这些快速增长的、海量的数据，如何对大数据进行分析处理、挖掘出价值是个巨大的挑战。

在存储方面，大数据通常可达到 PB 级的数据规模，因此，海量数据存储系统也一定要有相应等级的扩展能力。与此同时，存储系统的扩展一定要简便，可以通过增加模块或磁盘柜来增加容量，甚至不需要停机。当前互联网中的数据向着异质异构、无结构趋势发展，图像、视频、音频、文本等异构数据每天都在以惊人的速度增长，不断膨胀的信息数据使系统资源消耗量日益增大，运行效率显著降低。海量异构数据资源规模巨大，新数据类型不断涌现，用户需求呈现出多样性。目前，海量异构数据一般采用分布式存储技术。当前的存储架构仍不能解决数据的爆炸性增长带来的存储问题，静态的存储方案满足不了数据的动态演化所带来的挑战，因而在海量分布式存储和查询方面仍然需要做进一步研究。

数据安全是互联网中大数据管理的重要组成部分。然而随着网络规模不断扩大，数据和应用呈现出指数级增长趋势，给动态数据安全监控和隐私保护带来了极大的挑战。大数据分析往往需要多类数据相互参考，而在过去并不会有数据混合访问的情况，因此大数据应用也催生出一些全新的、亟待解决的安全问题。大数据的发展需要加大信息的开放程度，设计出新的信息收集设备，并为海量数据的存取和分析提供支持。由于数据存储和应用方式出现的新的变化，可能带来的副作用是 IT 基础架构将变得越来越一体化和外向型，对数据安全和知识产权构成更大风险。若企业不了解大数据内涵，就更增加了其风险成本。因此，企业需要关注完整的数据生命周期，包括数据质量、数据保留度、数据整合、数据安全性和信息隐私等内容。数据隐私作为安全的重要环节也备受关注，隐私问题包括两个方面：一方面是个人隐私的保护，随着数据采集技术的发展，在用户无法察觉时，个人的兴趣、习惯、身体特征等隐私信息可以被更容易地获取；另一方面，即使得到用户的许可，个人隐私数据在存放、传输和使用的过程中，也有被泄露的风险。大数据的分析能力导致看似简单的信息可能会被挖掘出其中的隐私，因此大数据时代下的隐私保护将成为

笔记

新的命题。

外部业务需求的数据转换也是挑战之一。由移动智能终端、物联网、云计算引发的大数据趋势，不仅改变了人们的生活方式，也要求企业重新设计运作模式，以数据驱动满足新的外部业务需求。但是，通常业务管理人员和后台技术人员使用的语言是不同的，业务管理人员会加入自己领域的术语和解释，技术人员会从系统实现的角度解释需求，两者的转换变得较为困难。因此，需要了解面向业务级的数据应用，针对不同业务部门的具体需求，统一业务语义模型和数据逻辑建模，根据需求合并、汇总业务数据，满足业务分析、挖掘和查询需求的变化。

在技术运用方面，大数据的发展也面临着一些困难。在实际生产中，有些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，难以用传统的方法描述与度量。而现有的数据处理方法仅适用于结构化数据，无法将大量的非结构化数据与结构化数据进行统一、整合。如何对跨业务平台的数据进行关联，并全面实时地给出分析结果，也是大数据技术面临的一个挑战。

数据质量影响着大数据的利用，低质量的数据不仅浪费了传输和存储资源，甚至也无法利用。制约数据质量的因素有很多，生成、采集、传输和存储的过程都可能影响数据质量。数据质量具体表现在数据的准确性、完整性、冗余性、一致性。虽然有很多提升数据质量的措施，但是数据质量的问题是不可能完全根除的。因此，需要一种可以对数据质量进行自动检测，并可以自行修复部分质量问题数据的方法。在大数据技术运用过程中，必须对大数据进行清理、准备、验证，检查其是否合规并不断维护，以防错误数据形成不准确的信息为业务决策带来危险。

1.2.5 研究大数据的意义

大数据的研究和分析应用具有十分重大的意义和价值。被誉为“大数据时代预言家”的维克托·迈尔·舍恩伯格在其《大数据时代》一书中列举了大量翔实的大数据应用案例，并分析预测了大数据的发展现状和未来趋势，提出了很多重要的观点和发展思路。他表示：“大数据开启了一次重大的时代转型”，并强调大数据将带来巨大的变革，改变人们的生活、工作和思维方式，改变商业模式，影响经济、政治、科技和社会等各个层面。由于大数据行业应用需求日益增长，未来越来越多的研究和应用领域将需要使用大数据并行计算技术，大数据技术将渗透到每个涉及大规模数据和复杂计算的应用领域。不仅如此，以大数据处理为中心的计算技术还将对传统计算技术产生革命性的影响，广泛影响计算机体系结构、操作系统、数据库、编译技术、程序设计技术和方法、软件工程技术、多媒体信息处理技术、人工智能以及其他计算机应用技术，并与传统计算技术相互结合产生很多新的研究热点和课题。

从国家战略角度来看，大数据是与自然资源、人力资源一样重要的战略资源，是一个国家数字主权的体现。大数据时代，国家层面的竞争力将部分体现为一国拥有大数据的规模、活性以及对数据的解释、运用的能力。一个国家在网络空间的数据主权将是继海、陆、空、天之后另一个大国博弈的空间。大数据的研究对捍卫国家网络空间的数字主权、维护社会稳定、推动社会与经济可持续发展有着独特的作用。在大数据领域的落后，意味

着失守产业战略制高点，意味着国家安全将在网络空间出现漏洞。大数据介入了人类探索现实自然界的过程，并用计算机处理人类的探索和发现。在这一过程中，随着巨量数据的产生，人类在不知不觉中创造了一个更复杂的数据自然界，同时，人、社会和宇宙的历史将变为数据的历史。

在科学研究方面，目前所有的科学研究领域都能形成相应的数据科学，不论是在自然科学领域还是社会科学领域，数据科学都能够在极大程度上支撑研究工作。大数据引起了学术界对科学研究方法论的重新审视，正在引发一场科学研究思维与方法的革命。另外，大数据的出现催生了一种新的科研模式，即面对大数据，科研人员只需从数据中直接查找、分析或挖掘所需要的信息、知识和智慧，甚至无须直接接触需研究的对象，大数据必将会为科学的研究和发展带来范式的革新和理念的进步。未来，针对大数据科学的研究将会在大数据网络发展和运营过程中发现和验证大数据的规律并揭示更多自然和社会活动的奥秘。

笔记

1.3 大数据的存储与计算模式

1.3.1 大数据的存储模式

面对大数据结构化、半结构化和非结构化海量数据的存储和管理，传统的数据中心无论是在性能、效率方面，还是在投资收益、安全方面，都已经远远不能满足新兴应用的需求，数据中心业务急需新型大数据处理中心来支撑。除了传统的高可靠、高冗余、绿色节能之外，新型的大数据中心还需具备虚拟化、模块化、弹性扩展、自动化等一系列特征，才能满足具备大数据特征的应用需求。这些史无前例的需求让存储系统的架构和功能都发生了前所未有的变化。由于轻型数据库无法满足对其存储以及复杂的数据挖掘和分析操作的要求，通常使用分布式文件系统、No SQL 数据库、云数据库等。

分布式系统包含多个自主的处理单元，通过计算机网络互连来协作完成分配的任务，其分而治之的策略能够更好地处理大规模数据分析问题。主要包含以下两类。

（1）分布式文件系统：存储管理需要多种技术的协同工作，其中文件系统为其提供最底层存储能力的支持。分布式文件系统 HDFS 是一个高度容错性系统，被设计成适用于批量处理，能够提供高吞吐量的数据访问。

（2）分布式键值系统：用于存储关系简单的半结构化数据。典型的分布式键值系统有 Amazon Dynamo，获得广泛应用和关注的对象存储技术（Object Storage）也可以视为键值系统，其存储和管理的是对象而不是数据块。

关系数据库由于无法满足海量数据的管理需求、无法满足数据高并发的需求、高可扩展性和高可用性的功能特点等弱点，已经无法覆盖当前需要。No SQL 数据库可以支持超大规模数据存储，灵活的数据模型可以很好地支持应用，具有强大的横向扩展能力，典型的 No SQL 数据库如图 1-2 所示。

笔记

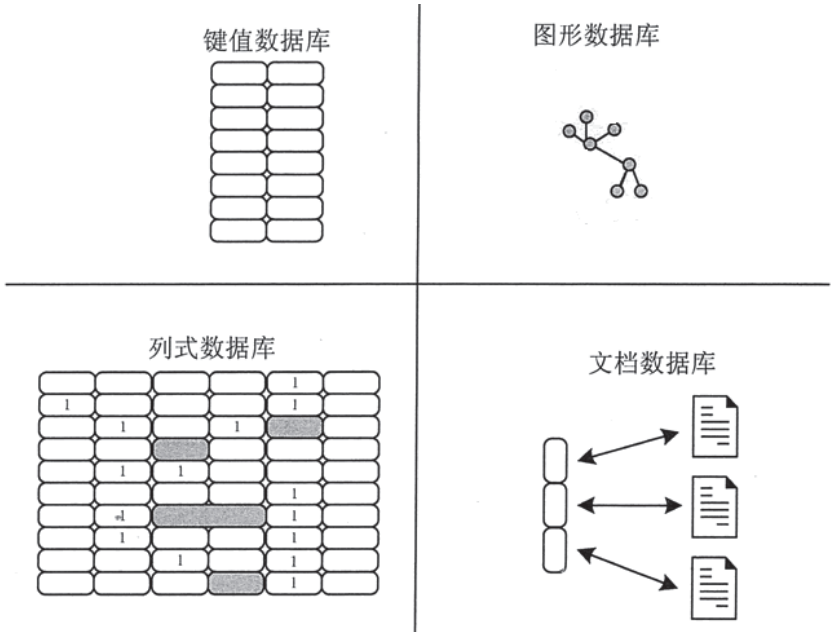


图 1-2　No SQL 数据库

云数据库是基于云计算技术发展的一种共享基础架构的方法，是部署和虚拟化在云计算环境中的数据库。与传统存储设备不同，云存储不是一个硬件，而是一个由网络设备、存储设备、服务器、软件、接入网络、用户访问接口以及客户端程序等多个部分构成的复杂系统。该系统以存储设备为核心，通过应用层软件对外提供数据存储和业务服务。随着视频监控的高清化和网络化，当前存储和管理的视频数据量已有海量之势，云存储技术是突破 IP 高清监控存储瓶颈的重要手段。云存储作为一种服务，在未来大数据应用领域有着广阔的应用前景（见图 1-3 ）。

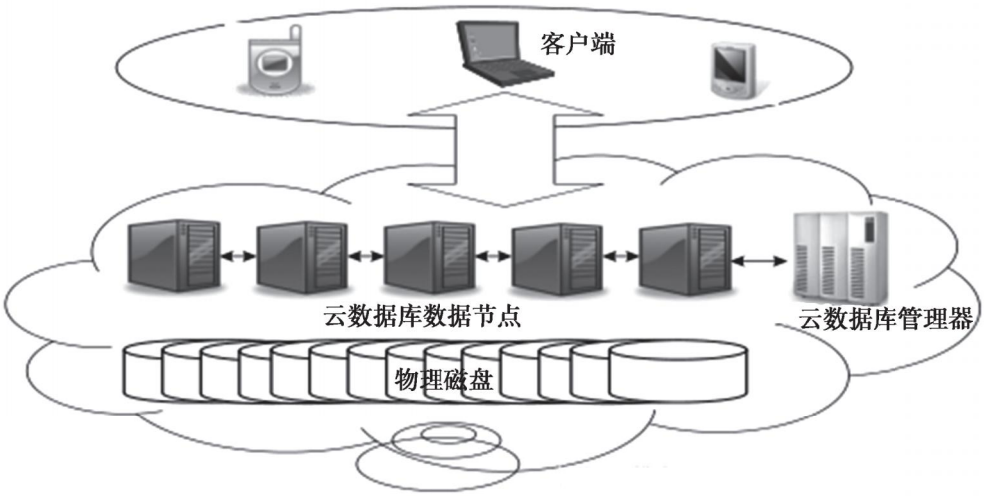


图 1-3　云数据库存储原理

云数据库具有动态可扩展性、高可用性、采用多租形式和支持资源有效分发等特点，还具有较低的使用代价、易用性、高性能、免维护、安全等特征。从数据模型的角度来说，云数据库并非一种全新的数据库技术，而只是以服务的方式提供数据库功能。云数据库所采用的数据模型可以是关系数据库所使用的关系模型（微软的 SQLAzure 云数据库都采用了关系模型）。同一个公司也可能提供采用不同数据模型的多种云数据库服务。

1.3.2 大数据的计算模式

笔记

大数据计算是发现信息、挖掘知识、满足应用的必要途径，也是大数据从收集、传输、存储、计算到应用等整个生命周期中最关键、最核心的环节，只有有效的大数据计算，才能满足大数据的上层应用需要，才能挖掘出大数据的内在价值，才能使大数据具有意义。大数据计算主要包括批量计算和实时计算，大数据计算模式是指根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象（Abstraction）和模型（Model）。大数据计算模式主要有批量计算、交互式计算、流式计算等，常用的大数据计算模式主要有批处理计算（MapReduce、Hadoop、Spark）、查询分析计算（HBase、Hive、Dremel、Cassandra、Impala Shark、HANA）、图计算（Pregel、Giraph、Trinity、GraphX PowerGraph）、流计算（Scribe、Flume、Storm、S4、Spark Streaming）、迭代计算（Twister、Spark）、内存计算（Spark、HANA、Dremel）等。针对大数据处理多样性的需求，可以采用不同的计算模式，并应用与大数据计算模式相对应的大数据计算系统和工具。

批处理计算是数据分析工作中比较常见的一类数据处理，其关键技术主要有 Pig、ZooKeeper、Hive、HDFS、Mahout 和 YARN，主要针对的是大规模数据的批量处理。MapReduce 大数据批处理技术是最具有典型的代表技术，用来执行大规模数据的处理任务，可用于大规模数据集的并行计算。MapReduce 计算模式作为一个主流的计算模式，可以进行大数据线下批处理，是一种支持分布式计算环境的并行处理，采用的是分治策略，即将一个大数据集分割为多个小尺度子集，然后让计算机程序靠近每个子集，同时并行完成计算处理。MapReduce 编程界面简单易用，能够在普通商业计算机集群上有效地处理超大规模数据，是当前大数据计算的一个主流计算模式，有力地推动了大数据技术以及应用向前发展。

大数据查询分析计算的关键技术主要是 HBase 和 Hive，是大数据计算模式当中常用的计算模式之一，其典型的代表产品如 Facebook 公司开发的 Cassandra、Cloudera 公司的实时查询引擎 Impala 以及 Google 公司的 Dremel，能够提供面向大数据存储管理、查询分析的新的技术方法和系统，可在数据体量极大时提供实时或准实时的数据查询分析能力，满足企业日常的管理需求。

图计算主要针对大规模图结构数据的处理，其关键技术主要有数据融合和图分割等，是一种可以表达为有向图的大数据计算，如社交网络数据等。一般对于大型图的计算，需要利用图计算模式，图计算相关产品较多，其中比较典型的代表产品有 Pregel、Trinity、GraphX、GraphLab 等。Pregel 是一种基于 BSP（Bulk Synchronous Parallel）模型实现的并行图处理系统，主要用于图遍历、最短路径和 PageRank 计算等。对于大型图的计算，通常使用的图处理软件主要有两种，第一种是基于遍历算法和实时的图数据库，第二种图处理软件主要是基于 BSP 模型实现的并行图处理系统。DT 时代，很多大数据都是以大规模的图或网络形式呈现，此外，还有非图结构的大数据，通常也会被转换为图模型后进行处理和分析。对于图计算，国内外的一些大型企业公司及研究机构等都进行了大量的研究，如 Google 推出的图并行计算框架 Pregel、Apach 开源社区在 2012 年启动的图并行计算的项目 Hama 等。

笔记

流计算（Stream Computing）是一种处理实时动态数据的计算模型。流计算主要是针对流数据的实时计算，能够对来自不同数据源的和连续到达的数据流进行实时处理，通过实时分析处理，给出有价值的分析结果。流计算的关键技术主要有高可用技术和数据传输以及系统资源调度等，存储体系为 HDFS 和 GFS，采用的计算模型是流计算模型，流计算的代表性产品主要有 Storm、S4、IBM InfoSphere Streams 和 IBM StreamBase 等，此外，还有百度开发的通用实时流数据计算系统 Dstream 以及淘宝开发的通用流数据实时计算系统。流计算系统一般要满足高性能、海量式、实时性、分布式、易用性和可靠性等多种需求，高性能方面主要是处理大数据的基本要求方面，海量式主要是支持的数据规模可以达到 PB 级或 PB 级以上，实时性主要是延迟时间方面，可以达到秒的级别，甚至是毫秒级，分布式主要是支持大数据的基本架构，必须能够平滑扩展，易用性主要是指能够快速进行开发和部署，可靠性主要是指能够可靠地处理流数据。

交互式计算的关键技术为 Hash 表和列存储结构，存储体系主要有 GFS、HDFS、NoSQL 等，其计算模型为 MapReduce + 算法，计算平台为 Hadoop、Google 等，典型的代表产品有 Dremel、Drill 以及 Power Drill 等。迭代计算的关键技术为 Spark，存储体系为 Twister，计算模型为基于内存的 RDD 数据集模型，计算平台为 Spark，其典型的代表产品有 Spark 等。内存计算其关键技术为内存数据库和列存储格式以及读写分离，存储体系为集中式存储，计算模型为大内存计算，计算平台有 Spark 和 HANA 等。

迭代计算的关键技术为 Spark，存储体系为 Twister，计算模型为基于内存的 RDD 数据集模型，其典型代表产品为 Spark，是大数据计算模式中常用的计算模式之一。当前服务器可配置的内存容量不断提高，其内存的价格也不断下降，采用内存计算完成快速的海量数据的处理已经成为大数据计算的一个主流趋势。

1.4 大数据的典型应用

1.4.1 智慧医疗

面对着人口老龄化、城镇化、慢性病增加、人口结构不合理等巨大挑战和城市与乡镇医疗水平的不均衡，医院的运营压力和管理压力日益增加。提前进入老龄化社会对我国的医疗卫生服务也提出了非常大的挑战，增加了对本来就供不应求的医疗服务的需求。面对这些挑战，医院亟须转变运营模式，迫切需要大数据、互联网、人工智能等技术辅助医疗行业，实现患者与医务人员、医疗机构、医疗设备之间的互动，逐步实现医疗管理信息化、智能化。

智慧医疗的启用可以实现医疗数据的共享共用。传统医疗系统是典型的垂直业务信息系统，每个医院和卫生服务机构都有自己的信息系统。大部分的资讯系统来自不同的厂商，而且数据格局不一致，数据规范不一致，数据形容方式也不一致，从而造成了“信息孤岛”。基于大数据的智慧医疗服务可以让所有医疗卫生服务机构共用一个信息系统，共享共用医疗数据。此外，共享医疗数据可为医院提供优质的医疗资源管理，可实现患者足不出户就把病治好的愿望，可为政府建立居民健康档案的数据来源。智慧医疗还可以优化医疗资源配置，我国的医疗资源分布极不平衡，在大城市的大型医院有优质的医疗资源，小医院或者基层卫生服务机构的医疗资源却匮乏。基于大数据的智慧医疗服务应用大数据

手段为患者提供跨机构网上问诊、在线预约等业务服务，将各级医疗卫生机构的优质医疗资源进行整合并实现共享，在让医疗资源得到了优化的配置的同时，提高了医疗办事效率。在患者医疗数据方面，医疗数据的来源是庞大数量的个体患者信息，隐藏着一定的商业利益，如果个人医疗数据遭到泄露，将会使患者隐私权受到侵害。基于大数据的智慧医疗服务是对不同层级的个人数据使用不同的保护形式，设置一定的访问权限，对患者个人医疗数据实行分层级整理，在一定程度上也可以有效地防止患者个人数据被泄露。

笔记

1.4.2 智慧农业

随着互联网技术的发展和社会的进步，大数据在农业领域也得到了广泛应用，并对我国农业实现网络化、智慧化、精准化生产产生了促进作用。大数据提升了我国农业科技水平，并通过挖掘农业深层价值，实现了农业快速发展。互联网大数据的应用，让我国农业资源得到了充分利用和整合，打破了传统农业信息化服务的限制，同时加强了各方之间的联系，如政府和农业企业等。早在 20 世纪 80 年代，传统农业就已经与互联网信息技术相融合，形成了计算机农业、数字农业、精准农业和智慧农业，实现了现代互联网信息技术在传统农业领域的应用。其中，计算机农业实质是利用软件技术搭建农业专家系统，使计算机具有农业专家水平，并以软件复制的形式，让普通农民随时使用；数字农业是精准农业和智慧农业的基础，这个阶段农业领域的发展已初步使用了大数据技术，具体包括农业土地数据、农业用水数据、农业生产气候数据、农业肥料数据等；精准农业则根据数字农业所提供的基础数据，对农作物土壤、土地等静态指标进行分析，通过计算机得出最优化投入，并施以动态控制，以达到农产品产出最大化；智慧农业则是农业发展的最高级形态，也是现代化农业发展的最终目标。

农业的分类方式不同对大数据的获取、采集、分析方法不尽相同，因此本小节从种植业、林业和畜牧业 3 个方面对农业大数据应用问题进行分析。

（1）种植业。种植业作为农业最重要的组成部分之一，在我国农业部门中所占的比重最大，相应的数据收集工作也更加繁重，尤其是种植业中的粮油等农作物发展状况，对我国国民经济发展水平、国家粮食安全以及民众生活质量具有重要影响。目前，通过智慧农业的应用，建立起规范的种植业数据信息化管理规章制度。因此，数据的合理管理在农业大数据中非常重要，有效的数据管理让大数据工作效率更高，这不仅需要高端的数据分析设备，人才在其中发挥的管理作用也不可或缺。只有发挥出人的主观能动性和积极性，才能最好地利用种植业大数据。

（2）林业。林业是一个复杂的生态系统，林业的发展受到很多环境因素的影响，造成林业内部系统不断更新，数据也随之发生动态的变化，而林业大数据的建设采取统一的数据尺度，在制定一个系统的标准的基础上，对不同环境的林业数据进行采集，大大提高了林业大数据的准确性和科学性。

（3）畜牧业。由于我国目前的畜牧业产业没有采用统一的标准进行管理，它的市场也比较复杂，没有很好地受到监督管理，大数据在畜牧业方面的应用可以使搜集到的加工、分享与实际利用等环节相对分散的相关农业数据集中起来，形成完整的数据分析系统。

总体来看，智慧农业即是对农业生产过程进行智慧化管理，根据大数据提供的相关信息对其进行分析与整合处理，将其利用在农业生产的具体过程中，胜于传统经验对农业

笔记

生产的指导与帮助，能直接提升农业管理与决策的准确性，提高农业产业的整体收入。首先，大数据可以对各项农业资源进行智慧化管理。土地资源、水资源、各类生物资源、相关生产资料等都是农业生产中的重要资源，在社会经济发展不断加速的影响下，使得农业发展的资源日益减少，所以，利用大数据对现有的农业资源进行全面的整合与处理尤为重要，能直接提升农业产业的优质高效性。其次，智慧农业也包含对农业生态环境进行智慧化的管理。土地情况、水资源情况、大气情况、病虫害情况以及自然灾害、自然环境等方面的因素都对农业生产的实际效益有着重要的影响，而在大数据的作用下，能对这些影响因素进行系统全面的精细化管理，保障农业生产的有序进行。再次，对农产品的安全性进行智慧化的管理也尤为重要。在农业生产的各个环节都需要安全管理，大数据能实现对农产品的整个生产链进行严格的质量监管与控制，在安全风险预警机制的作用下确保各类农产品安全问题均能得到有效的管理与应对。最后，智慧农业可以实现对农业生产设施设备进行智慧化的管控，在物联网技术与大数据技术的支持下，能对农业生产的相关设备与设施的具体运行状态实施监控与管理，不仅能及时发现设施设备运行过程中的问题，还能提高农业设备调度与管理的科学合理性，增强农业生产与经营决策的正确性。

1.4.3 金融行业

在金融领域，利用大数据平台对数据进行采集、加工和分析，能够实现金融机构数据采集的标准化、数据统计的自动化、数据应用分析的共享化，并在以下几个方面发挥重要作用。

（1）加大对决策制订的支持力度。通过快速、有效的数据共享和数据获取方式，大幅缩短决策支持周期，显著提升金融业应急响应水平；通过统一的基础代码、严格的校验规则等设置，保证数据的有效性和时效性。因此，能够为切实防范和化解区域性金融运行风险、维护金融稳定提供有力的支撑。提供多种数据采集手段提升数据采集效率和质量，如总行集中系统导出数据快速入库、金融机构报表数据上报等；通过预设数据模板和校验规则、后台监控等功能，加强数据校验；提供入库数据自动统计汇总功能，实现数据的基本加工。

（2）提高数据分析效率。采用数据可视化技术设计专题分析系统代替以往人工统计的方式，避免了部门间数据索取、人工校验和分析统计的过程，以地图、排名图、趋势图等更为直观、丰富的形式展现业务指标数据，并结合数据钻取手段从多个维度进行指标数据的展开分析，显著提升了数据分析的效率。金融大数据的应用还可以满足各类用户获取数据的需求，通过规定不同的角色，满足不同人员对数据的个性化需求。例如，面向管理人员提供专题分析系统、面向业务专家提供多维查询工具、面向业务人员提供报表定制功能。通过灵活的配置和简便的操作，提升了各类用户获取数据的效率。

（3）充分发挥现有数据价值。可以通过建立省级、市级数据中心，形成金融数据共享平台，实现跨地区、跨业务、跨部门的数据共享利用，并采用专用工具对数据进行综合分析与挖掘，充分发挥现有数据的价值。对于客户信息数据，可以将这些信息集中在大数据管理平台，对客户进行分类，依据其他的交易数据，进行产品开发和决策支持。例如，可以依据客户年龄、职业、收入、资产等，针对部分群体推出信用消费、抵押贷款、教育储蓄、投资产品、养老产品等，为客户提供针对人生不同阶段的金融服务。

1.4.4 零售行业

笔记

在零售行业，要实现大数据的精准助力，企业需能认识到收集客户相关数据的价值，懂得如何从这些数据中获得洞察力，继而创造智慧的、主动的、与客户相关的交互通路。公司明确如何有效地使用客户数据做决策，将洞察力转变为平台销售业绩的增长。对其而言，“大”并不是大数据中最重要的，最重要的应该是企业如何来驾驭这些大数据，企业对大数据进行分析，及随之采取的业务改进措施才是最重要的。拥有大量的数据本身并不会增加任何价值，只有全面地利用数据、统计和定量分析以及基于事实的管理，才能在当前复杂的环境中做出更明智的决策。零售企业在洞察大数据的结构和多元性的基础上，对其进行搜集、分析和执行，建立起以驱动客户参与为目标的市场推广模式。这种推广模式为公司内部职责建立了可量化标准，协助营销人员整合市场化流程、更高效地制订计划、执行任务和证明其业务价值，从而实现市场推广营销成果量化。

具体来说，大数据可以从以下 5 个方面驱动零售行业的发展和变革。

（1）客户细分：基于大数据海量数据和分析技术的支持，零售企业可以对客户进行群体细分，针对不同的客户群体制定相应的营销计划，以达到事半功倍的效果。

（2）客户挖掘：通过客户日常行为数据分析挖掘新需求。智能手机的普及让企业获得的客户行为数据呈爆发式增长，各大社交平台的迅猛发展更为之提供了一大助力。在云计算与大数据分析技术的支持下，企业可以将客户日常浏览行为和交易行为数据全部整合在一起，对客户进行深度分析，挖掘客户新的需求。

（3）数据储存：随着互联网时代的发展和 4G 网络的普及，企业对于储存空间的需求日益增长，不论是企业经营数据还是用户消费数据，都需要长久储存才能发挥其价值。

（4）客户关系管理：在大数据时代飞速发展的背景下，越来越多的企业意识到客户管理的重要性，企业通过不同的角度去深度分析客户，通过分析数据来挖掘新客户、提升客户黏性、降低客户流失率等。

（5）市场分析：根据已有的销售数据，对市场进行分析，从“人”“货”“场”三个角度来分析销售数据产生原因。例如，是否应该推出促销活动，产品的陈列、包装是否需要优化，以及客户购买动机等。

1.4.5 商务行业

在大数据的时代背景下，电子商务的经营模式发生了很大的变化，由传统的管理化的运营模式变为以信息为主体的数据化运营模式，电子商务的管理与各类经济环节都变得数据化，并且贯穿在整个电子商务环节中，小到基础材料的采购，大到资产运行及订单的完成。电子商务通过对大数据专业分析技术的运用，能够对消费者的消费习惯及消费心理进行归纳分析与预测，从而对电商产品的市场调度、供需程度进行一系列的建议指导，降低电商生产成本，提高效益。另一方面，在电商的经营中，大数据时代的到来可以使整个电商行业的信息资源共享变得方便快捷。电子商务的各环节有效地利用大数据的整合处理技术，在整个产品生产供应环节中实现各种数据信息的及时共享，从而更好地吸引消费者，促进产品销售，实现电子商务企业的产业结构转型优化与完善。以往被认为毫无利用价值的数据资料将是炙手可热的资源，电子商务模式下产生的数据资源不仅可以为自己所用，

笔记

也可以为电子商务企业创造相应的商业利益。各电子商务企业利用数据信息，开发数据分析业务、提供数据可视化服务以及数据资源共享等，扩展电子商务经营渠道，为企业增加效益。电子商务行业的大数据应用有以下几个方面：精准营销、个性化服务、商品个性化推荐。

精准营销指的是互联网企业使用大数据技术采集有关客户的各类数据，并通过大数据分析建立“用户画像”来抽象地描述一个用户的信息全貌，从而可以对用户进行个性化推荐、精准营销和广告投放等。当用户登录网站的瞬间，系统就能预测出该用户今天为何而来，然后从商品库中将合适的商品找出来，并推荐给他。大数据支持下的营销核心在于，让企业的业务在合适的时间，通过合适的载体，以合适的方式，推送给最需要此业务的用户。首先，大数据营销具有很强的时效性。在互联网时代，用户的消费行为极易在短时间内发生变化，大数据营销可以在用户需求最旺盛时及时进行营销策略实施。其次，可以实施个性化、差异化营销。大数据营销可以根据用户的兴趣爱好、在某一时间点的需求，做到对细分用户的一对一的营销，让业务的营销做到有的放矢，并可以根据实时性的效果反馈，及时调整营销策略。最后，大数据营销可以对目标用户的信息进行关联性分析。大数据可以对用户的各种信息进行多维度的关联分析，从大量数据中发现数据项集之间有趣的关联和相关联系。

另外，电子商务具有提供个性化服务的先天优势，可以通过技术支持实时获得用户的在线记录，并及时为他们提供定制化服务。许多电商都已经尝试依靠数据分析，在首页为用户提供全面的个性化的商品推荐。海尔和天猫提供了让用户在网上定制电视的功能，顾客可以在电视机生产以前选择尺寸、边框、清晰度、能耗、颜色、接口等属性，再由厂商组织生产并送货到顾客家中。这样的个性化服务受到了广泛欢迎。和服务个性化并行的是商品个性化。随着电子商务规模的不断扩大，商品数量和种类快速增长，顾客需要花费大量的时间才能找到自己想买的商品。个性化推荐系统通过分析用户的行为，包括反馈意见、购买记录和社交数据等，分析和挖掘顾客与商品之间的相关性，从而发现用户的个性化需求、兴趣等，然后将用户感兴趣的信息、产品推荐给用户。个性化推荐系统针对用户特点及兴趣爱好进行商品推荐，能有效地提高电子商务系统的服务能力，从而保留客户。大数据在电子商务中的应用如此广泛，因此企业掌握了大数据，就掌握了在这个瞬息万变时代中处于不败之地的秘籍。

1.4.6 电子政务

在电子政务早期发展阶段，社会数据和政务数据相互隔离。随着智慧电子政务的不断升级、大数据技术的发展以及各个部门对于数据资源使用需求的上涨，实现政务数据与社会数据的深度融合也成为一种必然趋势。政务大数据技术的运用打破了政务数据和社会数据中间的壁垒，政府可以利用第三方提供的数据，如互联网公司数据、金融机构数据等，更加全面和动态地了解社会舆情、民生信息；而社会机构也可以利用政府公开的一些公共服务数据、行政许可数据，实现政府和社会机构双方的共赢。

随着政府部门事务性工作的不断增加，仅依靠人工对相关数据进行收集、分类、整合、处理等工作不仅效率低、速度慢，还容易出现人为性差错，数据结果的人为性因素较大。此时，依托大数据技术在多元数据收集、处理方面的优势，及计算机网络技术下的信

息共享平台建设，能够帮助政府通过网络获取社会各领域的相关数据，并对数据资源进行有效整合，形成庞大的数据库资源。然而，对于数据库来说，只有得到利用才能体现其价值，在这样一种情况下，政府部门就充分利用了大数据交换共享平台的优势，建立以政府事务为中心的社会基础数据库，为政府相关工作的开展提供横向、纵向信息的全方位共享。在区域间政府工作交流方面，大数据共享交换平台能够突破传统政务工作的空间限制，进而促进跨地区政府部门信息资源整合与交流下的业务开展。

笔记

部分地方政府开展了大数据技术支持下政府网站的大数据分析工作。为准确掌握网站的浏览情况，大多数网站都会对用户的日常浏览情况进行数据分析，相关分析要素包括用户访问的路径、不同网页的停留时间、浏览网页的具体时间等，通过对以上要素的研究，能够对用户需求、习惯进行准确分析，并能够对后期网站缺陷的具体调整提供指导性意见。通过用户浏览网站后留下的大量信息，网站一方可以将用户信息存入数据库中，并利用大数据技术对相关信息进行分类，以实现网站信息对用户的精准推送。而且，经过大数据处理后的数据信息逐渐成为政府行政决策的重要依据，并能够在一定程度上保证行政决策的有效性和科学性。为更好地发挥电子政务的优势，在大数据交换共享平台的建设方面，需要对这一平台的信息资源目录体系进行完善，制定政府间统一的大数据交换共享平台使用标准，规范政府在使用大数据交换共享平台的各种行为，以实现对数据资源的合理、高效利用。因此，大数据交换共享平台的使用，不仅便于政府工作的开展，也促进了社会管理工作有条不紊的展开，社会环境的稳定得以实现。

1.5 初识 Hadoop 大数据平台

1.5.1 Hadoop 的发展过程

Hadoop 起源于 Apache Nutch 项目，始于 2002 年，是 Apache Lucene 的子项目之一。Google 在“操作系统设计与实现”（Operating System Design and Implementation，OSDI）会议上公开发表了题为“MapReduce：简化大规模集群上的数据处理”的论文之后，受到启发的 Doug Cutting 等人开始尝试实现 MapReduce 计算框架，并将它与 NDFS（Nutch Distributed File System）结合，用以支持 Nutch 引擎的主要算法。由于 NDFS 和 MapReduce 在 Nutch 引擎中有着良好的应用，它们于 2006 年 2 月分离出来，成为一套完整而独立的软件，并被命名为 Hadoop。到了 2008 年年初，Hadoop 已成为 Apache 的顶级项目，包含众多子项目，被应用到包括 Yahoo 在内的很多互联网公司。

1.5.2 Hadoop 的优势

Hadoop 是一个能够对大量数据进行分布式处理的软件框架。Hadoop 以一种可靠、高效、可伸缩的方式进行数据处理。Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理，按位存储和处理数据的能力值得人们信赖。同时，Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。Hadoop 还是可伸缩的，能够处理 PB 级数据。

此外，Hadoop 依赖于社区服务，因此它的成本比较低，任何人都可以使用。

笔记

Hadoop 是一个能够让用户轻松架构和使用的分布式计算平台，用户可以轻松地在 Hadoop 上开发和运行处理海量数据。Hadoop 还具有以下几个优点。

1. 高速性

Hadoop 拥有独特的存储方式，用于数据处理的工具通常在与数据相同的服务器上，从而导致能够更快地处理数据。如果用户正在处理大量的非结构化数据，Hadoop 能够有效地在几分钟内处理 TB 级的数据，而不是像以前处理 PB 级数据都要以小时为单位。

2. 容错能力佳

使用 Hadoop 的一个关键优势就是它的容错能力。Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。当数据被发送到一个单独的节点时，该数据也被复制到集群的其他节点上，这意味着在故障情况下，存在另一个副本可供使用，非单点故障。

3. 灵活性更好

Hadoop 能够使企业轻松访问到新的数据源，并可以分析不同类型的数据，从这些数据中产生价值，这意味着企业可以利用 Hadoop 的灵活性从社交媒体、电子邮件或点击流量等数据源中获得宝贵的商业价值。此外，Hadoop 的用途非常广，如对数处理、推荐系统、数据仓库、市场活动分析以及欺诈检测等。

4. 高可扩展性

Hadoop 是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。作为一个高度可扩展的存储平台，它可以存储和分发横跨数百个并行操作的廉价的服务器数据集群。不同于传统的关系型数据库系统不能扩展到处理大量的数据，Hadoop 是能给企业提供涉及成百上千 TB 的数据节点上运行的应用程序。

5. 成本效益

Hadoop 还为企业用户提供了极具成本效益的存储解决方案。传统的关系型数据库管理系统的问题是并不符合海量数据的处理器，不能够符合企业的成本效益。许多公司过去不得不假设哪些数据最有价值，然后根据这些有价值的数据设定分类，如果保存所有的数据，那么成本就会过高。虽然这种方法可以短期内实现工作，但是随着数据量的增大，这种方式并不能很好地解决问题。与一体机、商用数据仓库以及 QlikView、Yonghong Z-Suite 等数据集市相比，Hadoop 是开源的，项目的软件成本因此会大大降低。Hadoop 带有用 Java 语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的应用程序也可以使用其他语言编写，如 C++。

1.5.3 Hadoop 的生态系统

Hadoop 作为一个能够对大量数据进行分布式处理的软件框架，具有可靠、高效、可伸缩的特点。Hadoop 的核心是 HDFS 和 MapReduce，Hadoop 2.0 还包括 YARN。Hadoop 项目包括很多组件：Hadoop Common、Hadoop Distributed File System、Hadoop YARN 和 Hadoop MapReduce。这些组件系统共同提供给用户并支持附加的 Hadoop 工程的工具，让用户有能力实时处理大数据集，在这里 Hadoop 自动调度任务和管理集群资源。当下，Hadoop 已经成长为一个庞大的生态体系，只要是与海量数据相关的领域，都有 Hadoop 的身影。Google 在 2003 和 2004 年分别发表了两篇论文，其影响之一就是产生了 Apache

Hadoop 这个开源项目。同时，其他众多的开源项目，如 HBase、Hive、Hadoop++、Pig、Zookeeper 和 Sqoop 等，自然形成了围绕 Hadoop 的生态系统，为大数据提供了一个完整的、多种选择的解决方案。

笔记

MapReduce：MapReduce 是使用集群的并行、分布式算法处理大数据集的可编程模型。Apache MapReduce 是从 Google MapReduce 派生而来的，在大型集群中简化数据处理。当前的 Apache MapReduce 版本基于 Apache YARN 框架构建。YARN ＝“Yet-Another-Resource-Negotiator”，可以运行非 MapReduce 模型的应用。YARN 是 Apache Hadoop 想要超越 MapReduce 数据处理能力的一种尝试。

HDFS：The Hadoop Distributed File System 是提供跨多个机器存储大型文件的一种解决方案。Hadoop 和 HDFS 都是从 Google File System（GFS）中派生的。Hadoop 2.0 之前，NameNode 是 HDFS 集群的一个单点故障（SPOF）。利用 Zookeeper，HDFS 高可用特性解决了这个问题，提供选项来运行两个重复的 NameNodes，在同一个集群中，使用同一个 Active/Passive 配置。

ZooKeeper：ZooKeeper 是 Hadoop 的正式子项目，是一个针对大型分布式系统的可靠协调系统，提供的功能包括配置维护、名字服务、分布式同步、组服务等。ZooKeeper 的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户。Zookeeper 是 Google 的 Chubby 一个开源的实现，是高效和可靠的协同工作系统。Zookeeper 能够用来 Leader 选举、配置信息维护等。在一个分布式的环境中，人们需要一个 Master 实例或存储一些配置信息，以确保文件写入的一致性等。

HBase、Sqoop 和 Flume：Hadoop 核心还是一套批处理系统，数据加载进 HDFS、处理，然后检索。对于计算这或多或少有些倒退，但通常互动和随机存取数据是有必要的。HBase 作为面向列的数据库运行在 HDFS 之上。HBase 以 Google BigTable 为蓝本，项目的目标就是快速在主机内数十亿行数据中定位所需的数据并访问它。HBase 利用 MapReduce 来处理内部的海量数据。同时 Hive 和 Pig 都可以与 HBase 组合使用，Hive 和 Pig 还为 HBase 提供了高层语言支持，使得在 HBase 上进行数据统计处理变得非常简单。但为了授权随机存储数据，HBase 也做出了一些限制：Hive 与 HBase 的性能比原生在 HDFS 之上的 Hive 要慢 4~5 倍；同时，HBase 大约可存储 PB 级的数据，与之相比 HDFS 的容量限制达到 30 PB；HBase 不适合用于 ad-hoc 分析，HBase 更适合整合大数据作为大型应用的一部分，包括日志、计算以及时间序列数据。Sqoop 和 Flume 可改进数据的互操作性和其余部分。Sqoop 的功能主要是从关系数据库导入数据到 Hadoop，并可直接导入到 HFDS 或 Hive，而 Flume 设计旨在直接将流数据或日志数据导入 HDFS。

1.5.4 Hadoop 的版本

Hadoop 的发行版除了有 Apache Hadoop 外，Cloudera、Hortonworks、MapR、华为、DKhadoop 等都提供了自己的商业版本。商业发行版主要是提供了更为专业的技术支持，这对于大型企业更为重要，不同发行版都有自己的一些特点，本小节就各发行版作简单对比介绍。

DKHadoop：有效地集成了整个 Hadoop 生态系统的全部组件，并深度优化，重新编译为一个完整的性能更高的大数据通用计算平台，实现了各部件的有机协调。因此，DKH

笔记

相比开源的大数据平台，在计算性能上有了高达 5 倍（最大）的性能提升。DKHadoop 将复杂的大数据集群配置简化至 3 种节点（主节点、管理节点、计算节点），极大地简化了集群的管理运维，增强了集群的可用性、可维护性、稳定性。

Hortonworks：Hortonworks 的主打产品是 Hortonworks Data Platform（HDP），其同样是 100% 开源的产品，其版本特点在于，HDP 包括稳定版本的 Apache Hadoop 的所有关键组件；安装方便，HDP 包括一个现代化的、直观的用户界面的安装和配置工具。HDP 除了常见的项目外还包含了 Ambari，一款开源的安装和管理系统；HCatalog，一个元数据管理系统，HCatalog 现已集成到 Facebook 开源的 Hive 中。Hortonworks 的 Stinger 开创性极大地优化了 Hive 项目。Hortonworks 为入门提供了一个非常好的、易于使用的沙盒。Hortonworks 开发了很多增强特性并提交至核心主干，使得 Apache Hadoop 能够在包括 Windows Server 和 Windows Azure 在内的 Microsoft Windows 平台上本地运行。

Cloudera：CDH 是 Cloudera 的 Hadoop 发行版，完全开源，比 Apache Hadoop 在兼容性、安全性、稳定性上有增强。Cloudera 版本层次更加清晰，且它提供了适用于各种操作系统的 Hadoop 安装包，可直接使用 apt-get 或者 yum 命令进行安装，更加方便。

第 2 章 Hadoop 的安装和配置

学习目标

1. 掌握 Hadoop 的基本安装步骤。
2. 掌握 Hadoop 的单机模式。
3. 掌握 Hadoop 的分布式模式。

知识导图

笔记

本章导入

Hadoop 是一个开源的、可运行于大规模集群上的分布式计算平台，它主要包含分布式并行编程模型 MapReduce 和分布式文件系统 HDFS 等功能，已经在业内得到广泛的应用。借助于 Hadoop，程序员可以轻松地编写分布式并行程序，将其运行于计算机集群上，完成海量数据的存储与处理分析。

2.1 Hadoop 简介

Hadoop 是 Apache 软件基金会旗下的一个开源分布式计算平台，为用户提供了系统底层细节透明的分布式基础架构。Hadoop 是基于 Java 语言开发的，具有很好的跨平台特性，并且可以部署在廉价的计算机集群中。Hadoop 的核心是分布式文件系统和 MapReduce。

Hadoop 是行业公认的大数据标准开源软件，在分布式环境下提供了海量数据的处理能力。几乎所有主流厂商都围绕 Hadoop 提供开发工具、开源软件、商业化工具和技术服务，如谷歌、微软、思科、淘宝等，都支持 Hadoop。

Apache Hadoop 版本分为两代：第一代 Hadoop 称为 Hadoop 1.0；第二代 Hadoop 称为 Hadoop 2.0。第一代 Hadoop 包含 0.20.x、0.21.x 和 0.22.x 三大版本，其中，0.20.x 版本最后演化成 1.0.x 版本，变成了稳定版本，0.21.x 和 0.22.x 版本则增加了 HDFSHA 等重要的新特性。第二代 Hadoop 包含 0.23.x 和 2.x 两大版本，它们完全不同于 Hadoop 1.0，是一套全新的架构，均包含 HDFSFederation 和 YARN 两个组件。本书采用 Hadoop 2.7.1 版本。

除了免费开源的 Apache Hadoop 以外，还有一些商业公司推出的 Hadoop 发行版。2008 年，Cloudera 成为第一个 Hadoop 商业化公司，并在 2009 年推出第一个 Hadoop 发行版。此后，很多大公司也加入了 Hadoop 产品化的行列，如 MapR、Hortonworks、星环等。一般而言，商业化公司推出的 Hadoop 发行版也以 Apache Hadoop 为基础，但是，前者比后者具有更好的易用性、更多的功能和更高的性能。

2.2 安装 Hadoop 前的准备

安装 Hadoop 之前的一些准备工作包括创建 hadoop 用户、更新 APT、安装 SSH 和安装 Java 环境等。

2.2.1 创建 hadoop 用户

一定要按照该以下方法创建 hadoop 用户，并且使用 hadoop 用户登录 Linux 系统，然后再开始学习后面的内容。

1. 新建用户

使用 adduser 命令，具体代码如下。

```
sudo adduser hadoop
passwd hadoop
```

输入密码之后，依次输入“y”确定即可。

笔记

2. 添加用户组

在创建 hadoop 用户的同时也创建了 hadoop 用户组，下面将 hadoop 用户添加到 hadoop 用户组。具体代码如下。

```
sudo usermod -a -G hadoop hadoop
```

前面一个 hadoop 是组名，后面一个 hadoop 是用户名。完成后查询，代码如下。

```
cat /etc/group
```

3. 赋予 root 权限

先切换到 root 用户，代码如下。

```
sudo nano /etc/sudoers
```

然后修改文件。在 root ALL=(ALL) ALL 下方添加如下代码。

```
Hadoop ALL=(ALL) ALL
```

保存退出，hadoop 用户就拥有了 root 权限了。

2.2.2 更新 APT

为了确保 Hadoop 安装过程顺利进行，用 hadoop 用户登录 Linux 系统后打开一个终端，执行下方命令更新 APT 软件。

```
$sudo apt-get update
```

2.2.3 安装 SSH

SSH 是 SecureShell 的缩写，是建立在应用层和传输层基础上的安全协议。SSH 是目前较可靠、专为远程登录会话和其他网络服务提供安全性的协议。利用 SSH 协议可以有效防止远程管理过程中的信息泄露问题。SSH 最初是 UNIX 系统上的一个程序，后来迅速扩展到其他操作平台。SSH 由客户端和服务端的软件组成，服务端是一个守护进程，它在后台运行并响应来自客户端的连接请求，客户端包含 SSH 程序以及 SCP（远程复制）、Slogin（远程登录）、SFTP（安全文件传输）等其他的应用程序。

Hadoop 的名称节点（NameNode）需要启动集群中所有机器的 Hadoop 守护进程，这个过程需要通过 SSH 登录来实现，而 Hadoop 并没有提供 SSH 输入密码登录的形式，因此，为了能够顺利登录集群中的每台机器，需要将所有机器配置为“名称节点可以无密码登录”。

Ubuntu 默认已安装了 SSH 客户端，因此，这里还需要安装 SSH 服务端，在 Linux 的终端中执行以下命令。

```
$sudo apt-get installopenssh-server
```

笔记

安装后，可以使用如下命令登录本机。

```
$ssh localhost
```

执行该命令后会出现如图 2-1 所示的提示信息（SSH 首次登录提示），输入“yes”，然后按提示输入密码“hadoop”，就登录到本机了。

```
hadoop@DBLab-XMU:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is a9:28:e0:4e:89:40:a4:cd:75:8f:0b:8b:57:79:67:86.
Are you sure you want to continue connecting (yes/no)? yes
```

图 2-1　SSH 登录提示信息

此处在理解上会有一点费解，也就是说，原本人们登录进入 Linux 系统以后，就是在本机上，这时，在终端中输入的每条命令都是直接提交给本机去执行，然后，又在本机上使用 SSH 方式登录到本机，这时，在终端中输入的命令是通过 SSH 方式提交给本机处理。如果换成包含两台独立计算机的场景，SSH 登录会更容易理解。例如，有两台计算机 A 和 B 都安装了 Linux 系统，计算机 B 上安装了 SSH 服务端，计算机 A 上安装了 SSH 客户端，计算机 B 的 IP 地址是 59.77.16.33，在计算机 A 上执行命令“ssh 59.77.16.33”，就实现了通过 SSH 方式登录计算机 B 上面的 Linux 系统，在计算机 A 的 Linux 终端中输入的命令都会提交给计算机 B 上的 Linux 系统执行，也就是说，在计算机 A 上操作计算机 B 中的 Linux 系统。现在，只有一台计算机，就相当于计算机 A 和 B 都在同一台机器上，所以，理解起来就会有点费解。

由于这样登录需要每次都输入密码，配置成 SSH 无密码登录会比较方便。在 Hadoop 集群中，名称节点要登录某台机器（数据节点）时，也不可能人工输入密码，所以，也需要设置成 SSH 无密码登录。

首先，执行“exit”命令退出刚才的 SSH，就回到了原先的终端窗口；然后，可以执行命令“ssh-key gen”生成密钥，并将密钥加到授权中。代码如下。

```
$cd~/.ssh/   # 若没有该目录，请先执行一次 ssh localhost
$ssh-key gen-trsa   # 会有提示，按 Enter 键即可
$cat./id_rsa.pub>>./authorized_keys    # 加入授权
```

此时，再执行“ssh localhost”命令，无须输入密码就可以直接登录了，如图 2-2 所示。

```
hadoop@DBLab-XMU:~/.ssh$ ssh localhost
Welcome to Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-49-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

40 packages can be updated.
40 updates are security updates.

Last login: Thu Apr 30 21:20:50 2015 from localhost
hadoop@DBLab-XMU:~$
```

图 2-2　SSH 登录后的提示信息

2.2.4 安装 Java 环境

笔记

Hadoop 是基于 Java 语言开发的，本书的 Hadoop 应用程序也是采用 Java 语言编写的，因此，需要安装 Java 环境。Java 环境可选择安装 Oracle 的 JDK，或者安装 OpenJDK。下面介绍两种不同的安装方式，建议优先选择第一种方式，如果第一种方式失败，再选择第二种方式。

1. 第一种安装方式

直接通过如下命令安装 OpenJDK7。

```
$sudo apt-get install openjdk-7-jreo penjdk-7-jdk
```

上述安装过程需要访问网络下载相关文件，请保持联网状态。安装好 OpenJDK 后，需要找到相应的安装路径，这个路径是用于配置 JAVA_HOME 环境变量的。需要执行如下命令。

```
$dpkg -L openjdk-7-jdk|grep'/bin/javac'
```

该命令会输出一个路径，除去路径尾的“/bin/javac”，剩下的就是正确的路径了。例如，上面的命令执行后输出路径为 /usr/lib/jvm/java-7-openjdk-amd64/bin/javac，则人们需要的路径为 /usr/lib/jvm/java-7-openjdk-amd64#。

接下来，需要配置 JAVA_HOME 环境变量，为方便起见，这里直接在 ~/.bashrc 这个文件中进行设置，采用这种配置方式时，只对当前登录的单个用户生效，当该用户登录以及每次打开新的 Shell 时，它的环境变量文件 .bashrc 会被读取。在 Linux 终端中输入下面命令打开当前登录用户的环境变量配置文件 .bashrc。

```
$vim~/.bashrc
```

在文件最前面添加如下单独一行（注意，= 前后不能有空格），然后保存退出。

```
$exportJAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

接下来要让环境变量立即生效，执行如下命令。

```
$source?/.bashrc      # 使变量设置生效
```

执行上述命令后，可以检查一下是否设置正确，命令如下。

```
$echo$JAVA_HOME     # 检验变量值
$java-version
$$JAVA_HOME/bin/java-version          # 与直接执行 java-version 一样
```

如果设置正确，执行“$JAVA_HOME/bin/java-version”命令后，会输出 Java 的版本信息，并且与 java-version 的输出结果一样。

至此，Hadoop 所需的 Java 运行环境就顺利安装完成。

2. 第二种安装方式

根据大量计算机安装 Java 环境的情况发现，部分计算机按照上述第一种安装方式会

笔记

出现安装失败的情况，这时，可以采用这里介绍的另外一种安装方式，命令如下。

```
$sudo apt-get install default-jre default-jdk
```

上述安装过程需要访问网络下载相关文件，请保持联网状态。安装结束以后，需要配置 JAVA_HOME 环境变量，在 Linux 终端中输入如下命令打开当前登录用户的环境变量配置文件 .bashrc。

```
$vim~/.bashrc
```

在文件最前面添加如下单独一行（注意，= 前后不能有空格），然后保存退出。

```
$exportJAVA_HOME=/usr/lib/jvm/default-java
```

接下来要让环境变量立即生效，执行如下命令。

```
$source?/.bashrc      # 使变量设置生效
```

执行上述命令后，可以检查一下是否设置正确，命令如下。

```
$echo$JAVA_HOME     # 检验变量值
$java-version
$$JAVA_HOME/bin/java-version # 与直接执行 java-version 一样
```

至此，就成功安装了 Java 环境，下面就可以进入 Hadoop 的安装。

2.3 安装 Hadoop

Hadoop 安装包括 3 种模式。

（1）单机模式。只在一台机器上运行，存储采用本地文件系统，没有采用分布式文件系统 HDFS。

（2）伪分布式模式。存储采用分布式文件系统 HDFS，但是，HDFS 的名称节点和数据节点都在同一台机器上。

（3）分布式模式。存储采用分布式文件系统 HDFS，而且，HDFS 的名称节点和数据节点不同。

2.3.1 下载安装文件

本书采用的 Hadoop 版本是 2.7.1，可以到 Hadoop 官网下载安装文件，hadoop-2.7.1.tar.gz 文件需要保存到“%/home/hadoop/ 下载 /”目录下。需要注意的是，如果是在 Windows 系统下面下载安装文件 hadoop-2.7.1.tar.gz，则需要通过 FTP 软件上传到 Linux 系统的“%/home/hadoop/ 下载 /”目录下，这个目录是本书所有安装文件的中转站。

下载完安装文件以后，需要对文件进行解压。按照 Linux 系统使用的默认规范，用户安装的软件一般都是存放在“/usr/local/”目录下。使用 hadoop 用户登录 Linux 系统，打开一个终端，执行如下命令。

```
$sudotar-zxf~/下载/hadoop-2.7.1.tar.gz-C/usr/local        #解压到U/usr/local目录中的$cd/usr/local/
$sudomv./hadoop-2.7.1/./hadoop            #将文件夹名改为hadoop
$sudochown-Rhadoop./hadoop    #修改文件权限
```

笔记

Hadoop 解压后即可使用，可以输入如下命令来检查 Hadoop 是否可用，成功则会显示 Hadoop 版本信息。

```
$cd/usr/local/hadoop
$./bin/hadoopversion
```

2.3.2 单机模式配置

Hadoop 的默认模式为非分布式模式（本地模式），无须进行其他配置即可运行。

Hadoop 附带了丰富的例子，运行如下命令可以查看所有例子。

```
$cd/usr/local/hadoop
$./bin/hadoopjar./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar
```

上述命令执行后，会显示所有例子的简介信息，包括 grep、join、wordcount 等。此处选择运行 grep 例子，可以先在“/usr/local/hadoop”目录下创建一个文件夹 input，并复制一些文件到该文件夹下；然后，运行 grep 程序，将 input 文件夹中的所有文件作为 grep 的输入，让 grep 程序从所有文件中筛选出符合正则表达式“dfs[a\.+”的单词，并统计单词出现的次数；最后，把统计结果输出到“/usr/local/hadoop/output”文件夹中。完成上述操作的具体命令如下。

```
$cd/usr/local/hadoop
$mkdir input
$cp./etc/hadoop/".xml./input      #将配置文件复制到input目录下
$./bin/hadoopjar./share/hadoop/mapreduce/hadoop-mapreduce-examples-".jargrep./%#('#(, $9"[&-
8.]+,
$c&(./output/"                    #查看运行结果
```

执行成功后，结果如图 2-3 所示，输出了作业的相关信息，输出的结果是符合正则表达式的，单词“dfsadmin”出现了 1 次。

图 2-3　grep 程序运行结果

需要注意的是，Hadoop 默认不会覆盖结果文件，因此，再次运行上面实例会提示出错。如果要再次运行，需要先使用如下命令把 output 文件夹删除。

笔记

```
$rm -r ./output
```

2.3.3 伪分布式模式配置

Hadoop 可以在单个节点（一台机器）上以伪分布式的方式运行，同一个节点既作为名称节点（NameNode），又作为数据节点（DataNode），读取的是分布式文件系统 HDFS 中的文件。

1. 修改配置文件

需要配置相关文件，才能够让 Hadoop 在伪分布式模式下顺利运行。Hadoop 的配置文件位于 /usr/local/hadoop/etc/hadoop/ 中，进行伪分布式模式配置时，需要修改 2 个配置文件，即 core-site.xml 和 hdf-site.xml。

可以使用 VIM 编辑器打开 core-site.xml 文件，它的初始内容如下。

```
〈 configuration 〉
〈 /configuration 〉
```

修改以后，core-site.xml 文件的内容如下。

```
<configuration 〉
〈 property 〉
nameJhadoop.tmp.dir</nameJ
<value 〉 file:/usr/local/hadoop/tmp</value>
descriptionJAbaseforothertemporarydirectories.</description></property>
<property>
Wname 〉 fs.defaultFSW/name 〉
<valueJhdfs://localhost:9000</value 〉
</property 〉
</configuration 〉
```

在上面的配置文件中，“ hadoop.tmp.dir”用于保存临时文件，若没有配置“ hadoop.tmp.dir”这个参数，则默认使用的临时目录为 /tmp/hadoo-hadoop，这个目录在 Hadoop 重启时有可能被系统清理掉，导致一些意想不到的问题，因此，必须配置这个参数。“ fs.defaultFS”这个参数，用于指定 HDFS 的访问地址，其中，9000 是端口。

同样，需要修改配置文件 hdfs-site.xml，修改后的内容如下。

```
<configuration>
<property>
name>dfs.replication</name>
<value 〉: </value 〉
/property>
property>
name>dfs.namenode.name.dir</name>
value>file:/usr/local/hadoop/tmp/dfs/name</value>
</property>
property>
name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop/tmp/dfs/data</value>
</property>
</configuration>
```

笔记

在 hdfs-site.xml 文件中，“ dfs.replication”这个参数用于指定副本的数量。因为在分布式文件系统 HDFS 中，数据会被冗余存储多份以保证可靠性和可用性，但是，由于这里采用伪分布式模式，只有一个节点，只可能有 1 个副本，所以，设置“ dfs.replication”的帧为 1。“ dfs.namenode.name.dir”用于设定名称节点的元数据的保存目录，“ dfs.datanode.data.dir”用于设定数据节点的数据保存目录，这两个参数必须设定，否则后面会出错。配置文件 core-site.xml 和 hdfs-site.xml 的内容，也可以直接到本书官网的“下载专区”的“代码”目录下的“第 3 章”子目录下的“伪分布式模式配置”子目录中下载。

需要指出的是，Hadoop 的运行方式（运行在单机模式下还是运行在伪分布式模式下）是由配置文件决定的，启动 Hadoop 时会读取配置文件，然后根据配置文件来决定运行在什么模式下。因此，如果需要从伪分布式模式切换回单机模式，只需要删除 core-site.xml 中的配置项即可。

2. 执行名称节点格式化

修改配置文件以后，要执行名称节点的格式化，命令如下。

```
$cd/usr/local/hadoop
$./bin/hdfsnamenode-format
```

如果格式化成功，会看到“ has been successfully formatted”和“ Exiting with status 0”的提示信息，如图 2-4 所示。若为“Exiting with status1”，则表示出现错误。

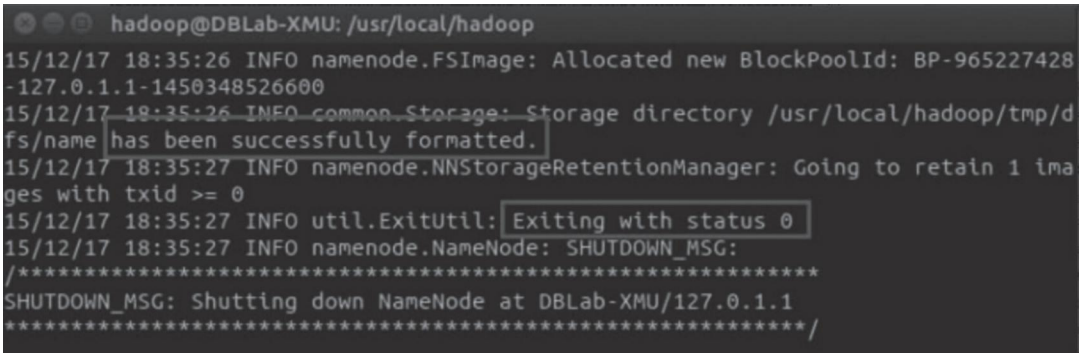


图 2-4　执行名称节点格式化后的提示信息

如果在执行这一步时提示错误信息“ Error ： JAVAHOME is not set and could not be found”，则说明之前设置 JAVA_HOME 环境变量时没有设置成功，要按前面的内容介绍先设置好 JAVA_HOME 变量，否则，后面的过程都无法顺利进行。

3. 启动 Hadoop

执行以下命令启动 Hadoop。

```
$cd/usr/local/hadoop
$./sbin/start-dfs.sh  #start-dfs.sh 是一个完整的可执行文件，中间没有空格
```

如果出现如图 2-5 所示的 SSH 提示，输入“yes”即可。

笔记

```
hadoop@DBLab-XMU:/usr/local/hadoop$sbin/start-dfs.sh
Startingnanenodeson[localhost]
localhost:startingnamenode,loggingto/usr/local/hadoop/logs/hadoop-hadoop-nalocalhost:startin
gdatanode,loggingto/usr/local/hadoop/logs/hadoop-hadoop-daStartingsecondarynanenodes[0.0.0.0]
Theauthenticityofhost10.0.0.0(9.9.0.9)1can'tbeestablished.
ECDSAkey._finqerprintisa9:28:e9:4e:89:40:a4:cd:75:8f:0b:8b:57:79:67:86.
Areyousureyouwanttocontinueconnecting(yes/no)?!yes
```

图 2-5　启动 Hadoop 后的提示信息

提示

这个警告提示信息可以忽略，并不会影响 Hadoop 的正常使用。

启动时可能会出现如下警告信息。

```
WARNutil.NativeCodeLoader:Unabletoloadnative-hadooplibraryforyourplatform***usingbuiltin-
javaclasseswhereapplicableWARN
```

如果启动 Hadoop 时遇到输出非常多“ssh：Could not resolve hostname xxx”的异常情况，如图 2-6 所示，这并不是 SSH 的问题，可以通过设置 Hadoop 环境变量来解决。

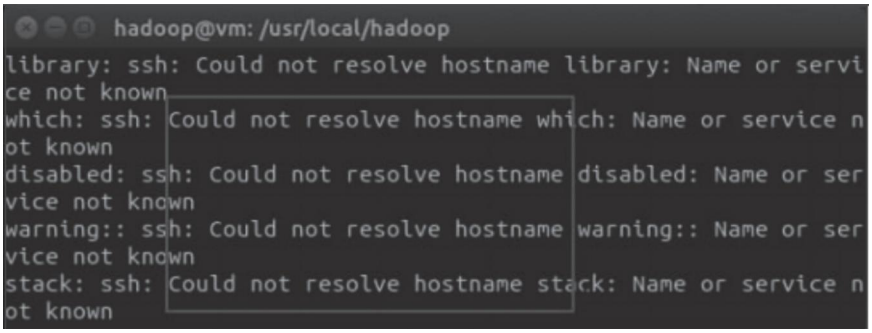


图 2-6　Hadoop 启动后的错误提示信息

首先，按“Ctrl+C”快捷键中断启动过程；然后，使用 VIM 编辑器打开文件~/.bashrc，在文件最上边的开始位置增加如下两行内容（设置过程与 JAVA_HOME 变量设置一样，其中，HADOOP_HOME 为 Hadoop 的安装目录）。

```
exportHADOOP_HOME=/usr/local/hadoop
exportHADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
```

保存该文件以后，务必要执行命令“source~.bashrc”使变量设置生效；然后，再次执行命令“./sbin/start-dfs.sh”启动 Hadoop。

Hadoop 启动完成后，可以通过命令“jps”来判断是否成功启动，命令如下。

```
$jps
```

若成功启动，则会列出如下进程：NameNode、DataNode 和 SecondaryNameNode。如果看不到 SecondaryNameNode 进程，执行命令“./sbin/stop-dfs.sh”关闭 Hadoop 相关进程；然后再次尝试启动。如果看不到 NameNode 或 DataNode 进程，则表示配置不成功，仔细检查之前的步骤，或通过查看启动日志排查原因。

通过 start-dfs.sh 命令启动 Hadoop 以后，就可以运行 MapReduce 程序处理数据，此时是对 HDFS 进行数据读写，而不是对本地文件进行读写。

4. Hadoop 无法正常启动的解决方法

一般可以通过查看启动日志来排查原因。启动时屏幕上会显示类似如下信息。

```
DBL&b-XMU:startingnarnenode,loggingto/usr/local/hadoop/logs/hadoop-hadoop-namenode-DBLab-XMU.out
```

其中，“DBL&b-XMU”对应的是机器名（用户的机器名可能不是这个名称），不过，实际上启动日志信息记录在下方文件中。

```
/usr/local/hadoop/logs/hadoop-hadoop-namenode-DBLab-XMU.log
```

所以，应该查看这个后缀为“.log”的文件，而不是“.out”文件。此外，每一次的启动日志都追加在日志文件之后，因此需要拉到日志文件的最后面查看，根据日志记录的时间信息，就可以找到某次启动的日志信息。

当找到属于本次启动的一段日志信息以后，出错的提示信息一般会出现在最后面，通常是写着 Fatal、Eiror、Warnin，或者 JavaException 的地方。可以在网上搜索一下出错信息，寻找一些相关的解决方法。

如果执行“jps”命令后找不到 DataNode 进程，则表示数据节点启动失败，可尝试如下的方法。

```
$./sbin/stop-dfs.sh   # 关闭
$rm-r./tmp  # 删除 tmp 文件，注意：这会删除 HDFS 中原有的所有数据
$./bin/hdfsnamenode-format   # 重新格式化名称节点
$./sbin/start-dfs.sh  # 重启
```

5. 使用 Web 界面查看 HDFS 信息

Hadoop 成功启动后，可以在 Linux 系统中（不是 Windows 系统）打开一个浏览器，在地址栏输入地址“http://localhost:50070”（见图 2-7）就可以查看名称节点和数据节点信息，还可以在线查看 HDFS 中的文件。

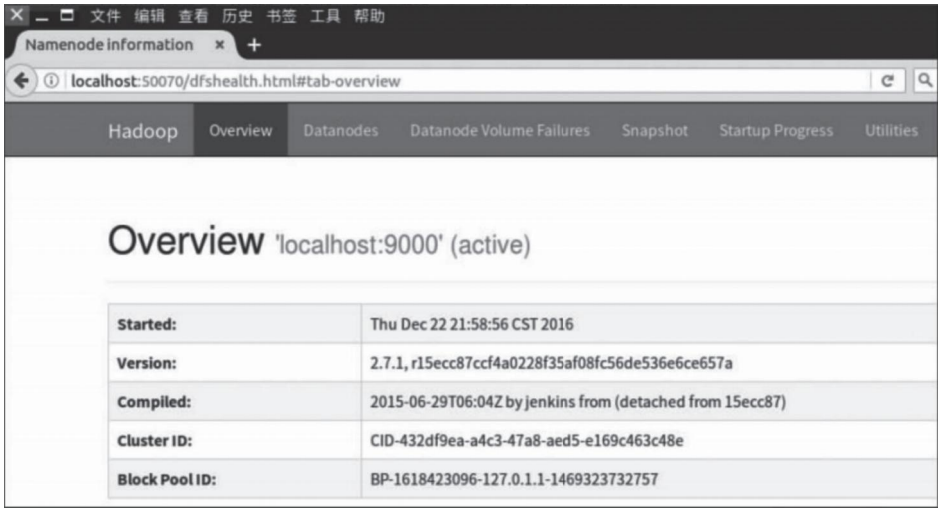


图 2-7　HDFS 的 Web 管理界面

6. 运行 Hadoop 伪分布式实例

在上面的单机模式中，grep 例子读取的是本地数据，在伪分布式模式下，读取的则是

笔记

注意

这会删除 HDFS 中原有的所有数据，如果原有的数据很重要，请不要这样做。不过对于初学者而言，通常这个时候不会有重要数据。

笔记

分布式文件系统 HDFS 上的数据。要使用 HDFS 首先需要在 HDFS 中创建用户目录（本书全部统一采用 hadoop 用户名登录 Linux 系统），命令如下。

```
$cd/usr/local/hadoop
$./bin/hdfsdfs-mkdir-'/user/hadoop
```

上面的命令是分布式文件系统 HDFS 的操作命令，会在后面作详细介绍，目前只需要按照命令操作即可。

接着需要将本地文件系统的“/usr/local/hadoop/etc/hadoop”目录中的所有 .xml 文件作为输入文件，复制到分布式文件系统 HDFS 中的“/user/hadoop/input”目录中，命令如下。

```
$cd/usr/local/hadoop
$./bin/hdfsdfs-mkdirinput      # 在 HDFS 中创建 hadoop 用户对应的 input 目录
$./bin/hdfsdfs-put./etc/hadoop/*.xmlinput        # 把本地文件复制到 HDFS 中
```

复制完成后，可以通过如下命令查看 HDFS 中的文件列表。

```
$./bin/hdfsdfs -Is input
```

执行上述命令以后，可以看到 input 目录下的文件信息。现在就可以运行 Hadoop 自带的 grep 程序，命令如图 2-8 所示。

```
hadoop@DBLab-XMU:/usr/local/hadoop$ bin/hdfs dfs -cat output/*
1       dfsadmin
1       dfs.replication
1       dfs.namenode.name.dir
1       dfs.datanode.data.dir
```

图 2-8　在 Hadoop 伪分布式模式下运行 grep 程序的结果

需要强调的是，Hadoop 运行程序时，输出目录不能存在，否则会提示如下错误信息。

```
org.apache.hadoop.mapred.FileAlreadyExistsException:Outputdirectoryhdfs://localhost:9000/user/hadoop/outputalreadyexists
```

因此，若要再次执行 grep 程序，需要执行如下命令删除 HDFS 中的 output 文件夹。

```
$./bin/hdfs dfs -rm -r output        # 删除 output 文件夹
```

7. 关闭 Hadoop

如果要关闭 Hadoop，可以执行以下的命令。

```
$cd/usr/local/hadoop
$./sbin/stop-dfs.sh
```

下次启动 Hadoop 时，无须进行名称节点的初始化（否则会出错），也就是说，不要再次执行 hdfsnamenode-format 命令，每次启动 Hadoop 只需要直接执行 start-dfs.sh 命令即可。

8. 启动 YARN

需要说明的是，本部分关于 YARN 的内容可以仅作为了解，不需要实际操作，本书在后面章节的所有操作环节都不会启动 YARN。如果读者按照下面介绍的 YARN 配置方

法进行了实际操作，那么，在操作结束后，请删除相关的 YARN 配置文件，恢复原来配置，确保后面其他步骤不会出现问题。

笔记

假设刚才已经关闭了 Hadoop，现在按照上面介绍的方法，使用如下命令再次启动 Hadoop。

```
$cd/usr/local/hadoop
$./sbin/start-dfs.sh
```

启动成功后，使用 Jps 命令查看进程，会发现只有以下几个进程。

```
SecondaryNameNode
Jps
DataNode
NameNode
```

可以看出，在上面运行 Hadoop 的过程中并没有启动 YARN。YARN 是 Hadoop 2.0 以后新增的专门负责资源管理与任务调度的组件，MapReduce 作业可以运行在 YARN 之上，由 YARN 负责为 MapReduce 作业进行资源管理和任务调度。但是，YARN 对于分布式模式（真正由多台机器构成的集群环境）才有意义，伪分布式模式下，YARN 无法真正发挥作用，因此，伪分布式模式下不需要借助于 YARN 为 MapReduce 作业进行资源管理和任务调度，而可以直接借助于 Hadoop 自身内置的 mapred.LocalJobRunner 来为 MapReduce 作业进行资源管理和任务调度。也就是说，不启动 YARN，MapReduce 程序也能够正常运行。不过，此处仍然要介绍一下伪分布模式下如何配置和启动 YARN，让 YARN 来负责资源管理与任务调度。

首先，修改配置文件“/usr/local/hadoop/etc/hadoop/mapred-site.xml”，此文件可以从“mapred-site.xml.template”重命名后得到，命令如下。

```
!$cd/usr/local/hadoop
$mv./etc/hadoop/mapred-site.xml.template./etc/hadoop/mapred-site.xml
```

然后，使用 VIM 编辑器打开 mapred-site.xml 文件，把配置修改为如下内容。

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

接着，修改配置文件“/usr/local/hadoop/etc/hadoop/yarn-site.xml”，把配置修改为如下内容。

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
</configuration>
```

提示

开启历史服务器，才能在 Web 中查看任务运行情况。

现在就可以启动 Hadoop 和 YARN 了，命令如下。

```
$./sbin/mr-jobhistory-daemon.shstarthistoryserver
```

开启后，可以通过 Jps 命令查看当前运行的进程，可以看到多了“NodeManager”和“ResourceManager”两个后台进程，如图 2-9 所示。

图 2-9　后台进程

启动 YARN 之后，运行实例的方法与伪分布式模式是一样的，仅仅是资源管理、任务调度方式不同。观察日志信息可以发现，不启用 YARN 时，是 mapred.LocalJobRunner 在跑任务；启用 YARN 之后，是 mapred.YARNRunner 在跑任务。启动 YARN 还有个好处，那就是可以通过 Web 界面查看任务的运行情况，如图 2-10 所示。

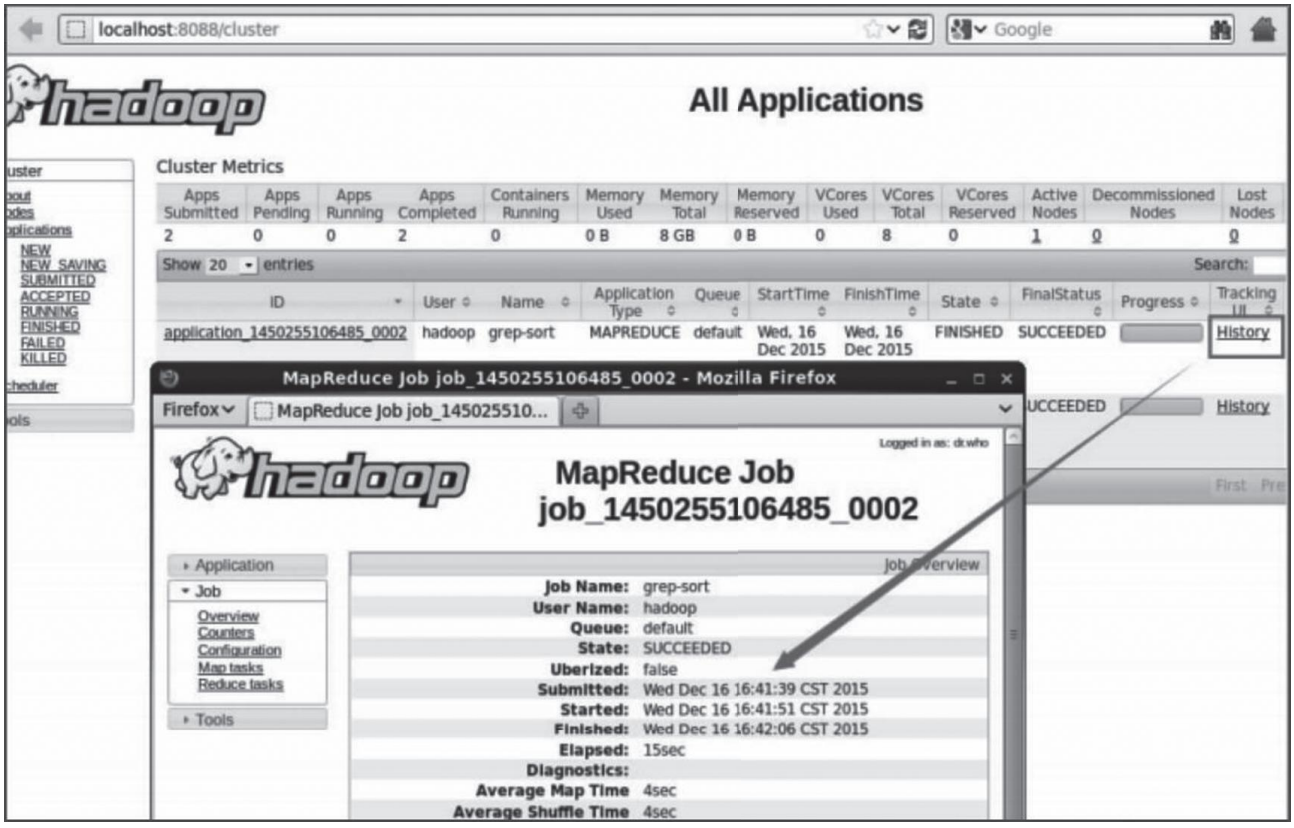


图 2-10　通过 Web 界面查看任务的运行情况

关闭 YARN 和 Hadoop 的命令如下。

笔记

```
$./sbin/stop-yarn.sh
$./sbin/mr-jobhistory-daemon.shstophistoryserver
$./sbin/stop-dfs.sh
```

需要注意的是，在经过上述测试以后，如果以后在伪分布式模式下启动 Hadoop 的时候不想同时启动 YARN，就务必把配置文件 mapred-site.xml 重命名，改成 mapred-site.xml.template，需要用时改回来即可。否则，如果该配置文件存在，使用 start-dfs.sh 命令启动 Hadoop 以后却未开启 YARN，则运行程序时会提示“ Retrying connect to server：0.0.0.0/0.0.0.0%032”错误信息。

9. 配置 PATH 变量

前面在启动 Hadoop 时，都要加上命令的路径，例如，/sbin/start-dfs.sh 这个命令中就带上了路径，实际上，通过设置 PATH 变量，可以在执行命令时，不用带上命令本身所在的路径。例如，打开一个 Linux 终端，在任何一个目录下执行 ls 命令时，都没有带上 ls 命令的路径。实际上，执行 ls 命令时，是执行 /bin/ls 这个程序，之所以不需要带上路径，是因为 Linux 系统已经把 ls 命令的路径加到 PATH 变量中，当执行 Is 命令时，系统是根据 PATH 这个环境变量中包含的目录位置，逐一进行查找，直至在这些目录位置下找到匹配的 ls 程序（若没有匹配的程序，则系统会提示该命令不存在）。

知道了这个原理以后，同样可以把 start-dfs.sh、stop-dfs.sh 等命令所在的目录 /usr/local/hadoop/sbin 加到环境变量 PATH 中，这样，以后在任何目录下都可以直接使用命令 start-dfs.sh 启动 Hadoop，不用带上命令路径。具体操作方法：先使用 VIM 编辑器打开 ~/.bashrc 这个文件，然后在此文件的最前方位置处加入如下单独一行命令。

```
Export PATH=$PATH:/usr/local/hadoop/sbin
```

在后面的学习过程中，如果要继续把其他命令的路径也加入 PATH 变量中，也需要继续修改 ~/.bashrc 文件。当后面要继续加入新的路径时，只要用英文冒号隔开，把新的路径加到后方即可，例如，如果要继续把“ /usr/local/hadoop/bin”路径增加到 PATH 中，只要继续追加到后方，如下所示。

```
PATQt$PATQ:/usr/local/hadoop/sbin:/usr/local/hadoop/bin
```

添加后，执行命令 source~/.bashrc 使设置生效。设置生效后，在任何目录下启动 Hadoop，都只需直接输入 start-dfs.sh 命令即可。同理，停止 Hadoop，也只需要在任何目录下输入 stop-dfs.sh 命令即可。

2.3.4 分布式模式配置

当 Hadoop 采用分布式模式部署和运行时，存储采用分布式文件系统 HDFS，而且，HDFS 的名称节点和数据节点位于不同机器上。这时，数据就可以分布到多个节点上，不同数据节点上的数据计算可以并行执行，这时 MapReduce 的分布式计算能力才能真正发挥作用。

为了降低分布式模式部署的难度，本书简单使用两个节点（两台物理机器）来搭建

笔记

集群环境：一台机器作为 Master 节点，局域网 IP 地址为 192.168.1.121 ；另一台机器作为 Slave 节点，局域网 IP 地址为 192.168.1.122。由 3 个以上节点构成的集群，也可以采用类似的方法完成安装部署。

Hadoop 集群的安装配置大致包括以下几个步骤。

（1）选定一台机器作为 Master。

（2）在 Master 节点上创建 hadoop 用户、安装 SSH 服务端、安装 Java 环境。

（3）在 Master 节点上安装 Hadoop，并完成配置。

（4）在其他 Slave 节点上创建 hadoop 用户、安装 SSH 服务端、安装 Java 环境。

（5）将 Master 节点上的 /usr/local/hadoop 目录复制到其他 Slave 节点上。

（6）在 Master 上启动 Hadoop。

上述步骤中，关于如何创建 hadoop 用户、安装 SSH 服务端、安装 Java 环境、安装 Hadoop 等过程，已经在前面介绍伪分布式安装的时候作了详细介绍，按照之前介绍的方法完成步骤（1）~步骤（4），这里不再赘述。在完成步骤（1）~步骤（4）的操作以后，才可以继续进行下面的操作。

1. 网络配置

假设集群所用的两个节点（机器）都位于同一个局域网内。如果两个节点使用的是虚拟机安装的 Linux 系统，那么两者就都需要更改网络连接方式为“桥接网卡”模式，才能实现多个节点互连，如图 2-11 所示。此外，一定要确保各个节点的 MAC 地址不能相同，否则会出现 IP 冲突。在第 2 章曾介绍过采用导入虚拟机镜像文件的方式安装 Linux 系统，如果是采用这种方式安装 Linux 系统，就有可能出现两台机器的 MAC 地址是相同的，因为一台机器复制了另一台机器的配置，因此，需要改变机器的 MAC 地址（见图 2-11），可以单击界面右边的“刷新”按钮随机生成 MAC 地址，这样就可以让两台机器的 MAC 地址不同了。

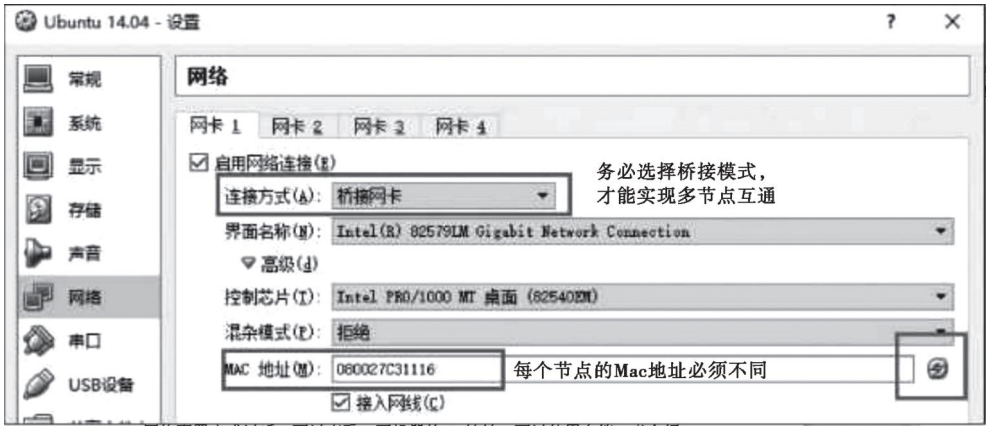


图 2-11　网络连接方式设置

网络配置完成以后，应查看一下机器的 IP 地址，可以使用 ipconfig 命令查看。本书在同一个局域网内部的两台机器的 IP 地址分别是 192.168.1.121 和 192.168.1.122。

由于集群中有两台机器需要设置，在接下来的操作中，一定要注意区分 Master 节点和 Slave 节点。为了便于区分 Master 节点和 Slave 节点，可以修改各个节点的主机名，这样，在 Linux 系统中打开一个终端以后，在终端窗口的标题和命令行中都可以看到主机名，就比较容易区分当前是对哪台机器进行操作。在 Ubuntu 中，在 Master 节点上执行如

下命令修改主机名。

```
$sudovim/etc/hostname
```

执行上面命令后，就打开了“/etc/hostname”这个文件，这个文件里面记录了主机名，例如，本书在第 2 章中安装 Ubuntu 系统时，设置的主机名是 dblab-VirtualBox，因此，打开这个文件以后，里面就只有 dblab-VirtualBox 这一行内容，可以直接删除，并修改为 Master（注意，此处是区分大小写的）；然后保存并退出 VIM 编辑器，这样就完成了主机名的修改。需要重启 Linux 系统才能看到主机名的变化。

要注意观察主机名修改前后的变化。在修改主机名之前，如果用 hadoop 用户登录 Linux 系统，打开终端，进入 Shell 命令提示符状态，会显示如下内容。

```
hadoop@dblab-VirtualBox:?$
```

修改主机名并且重启系统之后，用 hadoop 用户登录 Linux 系统，打开终端，进入 Shell 命令提示符状态，则会显示如下内容。

```
hadoop@Master:~!
```

这时就很容易辨认出当前是于 Master 节点上进行操作，不会和 Slave 节点产生混淆。

然后执行如下命令打开并修改 Master 节点中的“/etc/hosts”文件。

```
$sudovim/etc/hosts
```

可以在 hosts 文件中增加如下两条 IP 和主机名映射关系。

```
192.168.1.121 Master
192.168.1.122 Slave1
```

修改后的效果如图 2-12 所示：

图 2-12　修改 IP 和主机名映射关系后的效果

完成了 Master 节点的配置后，接下来要继续完成对其他 Slave 节点的配置修改。本书只有一个 Slave 节点，主机名为 Slave1。参照上面的方法，把 Slave 节点上的“/etc/hostname”文件中的主机名修改为 Slave1，同时，修改“/etc/hosts”的内容，在 hosts 文件中增加如下两条 IP 和主机名映射关系。

```
192.168.1.121 Master
192.168.1.122 Slave1
```

修改完成以后，重新启动 Slave 节点的 Linux 系统。

笔记

注意

一般 hosts 文件中只能有一个 127.0.0.1 映射，其对应主机名为“localhost”，如果有多余 127.0.0.1 映射，应删除，特别是不能存在“127.0.0.1Master”这样的映射记录。修改后需要重启 Linux 系统。

笔记

这样就完成了 Master 节点和 Slave 节点的配置，然后，需要在各个节点上都执行如下命令，测试是否相互 ping 得通，如果 ping 不通，后面就无法顺利配置成功。

```
$pingMaster-c3      # 只 ping3 次就会停止，否则要按 Ctrl+C 键中断 ping 命令
$pingSlave1-c3
```

例如，在 Master 节点上 ping Slave1，如果 ping 通的话，会显示如图 2-13 所示的结果。

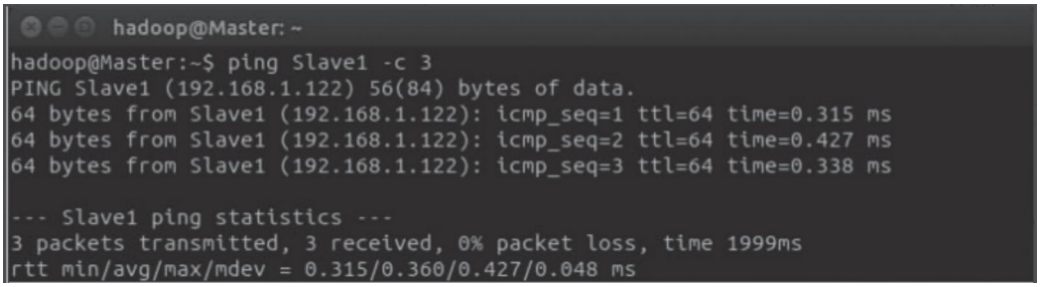


图 2-13　ping 命令效果

2. SSH 无密码登录节点

图 2-13 为使用 ping 命令的效果节点的公匙，如果之前已经生成过公钥，必须要删除原来生成的公钥，重新生成一次，因为前面对主机名进行了修改。具体命令如下。

```
$cd~/.ssh           # 如果没有该目录，先执行一次 sshlocalhost
$rm./id_rsa*        # 删除之前生成的公匙（如果已经存在）
$ssh-keygen-trsa    # 执行该命令后，遇到提示信息，一直按 Etner 键就可以
```

为了让 Master 节点能够无密码 SSH 登录本机，需要在 Master 节点上执行如下命令。

```
$cat./id_rsa.pub>>./authorized_keys
```

完成后可以执行命令 sshMaster 来验证一下，可能会遇到提示信息，只要输入“ yes”命令即可，测试成功后，执行“exit”命令返回原来的终端。

接下来在 Master 节点将上述公匙传输到 Slavel 节点，命令如下。

```
scp~/.ssh/id_rsa.pubhadoop@Slave1:/home/hadoop/
```

上面的命令中，“ scp”是 securecopy 的简写，用于在 Linux 下远程复制文件，类似于 cp 命令，不过，cp 只能在本机中复制。执行 scp 时会要求输入 Slave 1 上 hadoop 用户的密码，输入完成后会提示传输完毕，如图 2-14 所示。

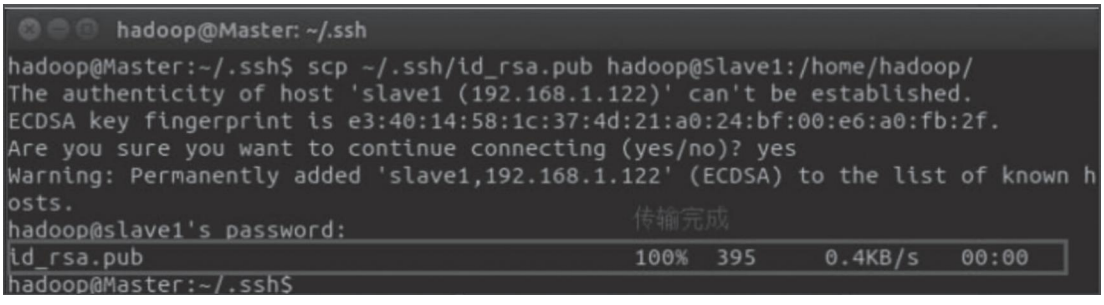


图 2-14　执行 scp 命令后的效果

接着在 Slave 1 节点上将 SSH 公钥加入授权，命令如下。

```
$mkdir~/.ssh          # 如果不存在该文件夹需先创建，若已存在，则忽略本命令
$cat~/id_rsa.pub >> ~/.ssh/authorized_keys
$rm~/id_rsa.pub       # 用完以后就可以删掉
```

如果有其他 Slave 节点，也要执行将 Master 公钥传输到 Slave 节点以及在 Slave 节点上加入授权这两步操作。

这样，在 Master 节点上就可以无密码 SSH 登录到各个 Slave 节点了，可在 Master 节点上执行如下命令进行检验。

```
$ssh Slave1
```

执行该命令后的效果如图 2-15 所示。

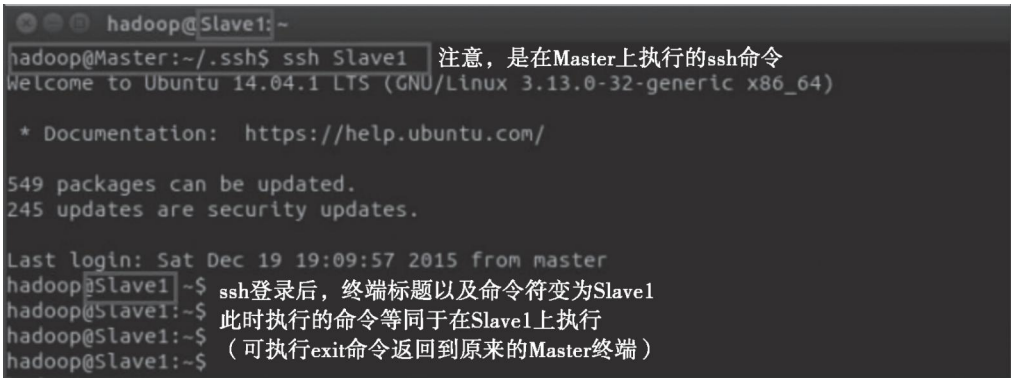


图 2-15　ssh 命令执行后的效果

3. 配置 PATH 变量

在前面的伪分布式安装内容中，已经介绍过 PATH 变量的配置方法，可以按照同样的方法进行配置，这样就可以在任意目录中直接使用 hadoop、hdfs 等命令了。如果还没有配置 PATH 变量，就需要在 Master 节点上进行配置。首先，执行命令 vim~/.bashrc，也就是使用 VIM 编辑器打开 ~/.bashrc 文件；然后，在该文件最上方的位置处加入如下命令。

```
Export PATH=$PATH:/usr/local/hadoop/bin:/usr/local/hadoop/sbin
```

保存后执行命令 source~/.bashrc，使配置生效。

笔记