普通高等教育新工科人才培养规划教材

（大数据专业）

Hadoop

大数据开发

主　编 ◇ 刘春阳　张学龙　刘丽军

副主编 ◇ 陈　勇　陈艳男　蒋中洲　王宇希

中国水利水电出版社
www.waterpub.com.cn

普通高等教育新工科人才培养规划教材（大数据专业）

Hadoop 大数据开发

主　编　刘春阳　张学龙　刘丽军

副主编　陈　勇　陈艳男　蒋中洲　王宇希

中国水利水电出版社
www.waterpub.com.cn

·北京·

内 容 提 要

本书通过原理加案例方式系统讲解了 Hadoop 大数据开发，精心安排了原理分析、环境搭建、案例开发等环节，使读者对解决大数据问题有清晰的思路。

全书共 7 章：前 6 章系统讲解大数据 Hadoop 架构，包括大数据处理平台 Hadoop、分布式文件系统 HDFS，并行计算模型 MapReduce、资源调度框架 Yarn；第 7 章是 MapReduce 应用实例，通过案例帮助读者进一步理解 Hadoop 平台。全书突出三个特点：道理简单明了、思路清晰透彻、案例新颖实用。

本书可作为普通高校大数据相关专业的教材，可供想深入了解 Hadoop 架构编程的读者参考，还可作为相关培训班的培训教材。

图书在版编目（CIP）数据

Hadoop大数据开发 / 刘春阳，张学龙，刘丽军主编
. -- 北京 : 中国水利水电出版社，2018.9
普通高等教育新工科人才培养规划教材. 大数据专业
ISBN 978-7-5170-6903-4

Ⅰ. ①H… Ⅱ. ①刘… ②张… ③刘… Ⅲ. ①数据处理软件－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第216857号

策划编辑：石永峰　　责任编辑：张玉玲　　封面设计：梁　燕

书　名	普通高等教育新工科人才培养规划教材（大数据专业） Hadoop 大数据开发 Hadoop DASHUJU KAIFA
作　者	主　编　刘春阳　张学龙　刘丽军 副主编　陈　勇　陈艳男　蒋中洲　王宇希
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座　100038） 网址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电话：（010）68367658（营销中心）、82562819（万水）
经　售	全国各地新华书店和相关出版物销售网点
排　版	北京万水电子信息有限公司
印　刷	三河市鑫金马印装有限公司
规　格	184mm×260mm　16 开本　11.5 印张　280 千字
版　次	2018 年 9 月第 1 版　2018 年 9 月第 1 次印刷
印　数	0001—4000 册
定　价	32.00 元

前　言

这是一个大数据爆发的时代，面对信息的激流、多元化数据的涌现，大数据已经为个人生活、企业经营，甚至国家与社会的发展带来了机遇和挑战，成为信息产业中极具潜力的增长点。大数据时代在众多领域掀起变革的巨浪，但我们要冷静地看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件简单地堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。

Hadoop 技术能够成功的最根本原因在于它是把传统的集中式运算转化成分布式计算的一种有效手段。Hadoop 的分布式文件系统能够以可靠快捷的方式将数据分布存储到不同计算节点中，Hadoop MapReduce 编程又能够以简单的方法为人们提供分布式编程接口，从而降低了分布式开发门槛。

本书共 7 章，不仅有详细的理论讲解，还有大量的实战操作，具体内容如下：

第 1 章深入探究大数据的概念、产生的背景和发展现状，应用案例指出了大数据面临的机遇与挑战，介绍大数据处理技术和计算模式，最后阐述大数据与云计算之间的区别和联系。

第 2 章详细介绍大数据处理平台 Hadoop 的生态系统和架构。

第 3 章讲解 Hadoop 分布式平台的搭建和验证。

第 4 章描述 HDFS 的架构、工作机制、文件读写流程和 Shell 命令。

第 5 章讲解 HDFS Windows 远程开发、HDFS Java API 接口和编程实战。

第 6 章讲解 MapReduce 编程模型、工作原理和 Yarn 资源管理。

第 7 章讲解常用的 MapReduce Java API 接口、应用实例和高级编程。

本书的编写得到北京百知教育科技有限公司的大力支持，在此表示感谢。

由于时间仓促及编者水平有限，本书难免存在不足之处，恳请读者批评指正。

编　者

2018 年 7 月

目　录

第 1 章　大数据概论

随着互联网的飞速发展，数据已经积累到了一个由量变引起质变的程度，大数据（Big Data）几乎应用到了人们发展的所有领域中，不管是云计算、物联网，还是社交网络、移动互联网等都会与大数据扯上关系。那么，什么是大数据，大数据发展的现状如何，大数据能给人们带来什么？通过本章的学习，您将会得到答案。

1.1　大数据概述

网络和信息技术的不断发展，带动了移动设备和通信手段（如社交网站）的革新，人类生产的数据量每年都在快速增长。各行业信息化程度的提高导致业务数据正以几何级数的形式爆发，预计到 2020 年全球将总共拥有 35 亿 GB 的数据量，其收集、存储、格式、检索、分析、应用等存在诸多问题，不能再以传统的信息处理技术加以解决。

目前对大数据的准确定义尚有一些争论，这就导致大数据的定义有多种。维基百科给出的定义是：大数据是利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。美国国家科学基金会（NSF）则将大数据定义为“由科学仪器、传感设备、互联网交易、电子邮件、音视频软件、网络点击流等多种数据源生成的大规模、多元化、复杂、长期的分布式数据集”。全球知名的咨询公司麦肯锡认为：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。但它同时指出“大数据”并非总是说有数百个 TB 才算得上，根据实际使用情况，有时候数百个 GB 的数据也可以称为大数据，这主要看它的第三个维度，也就是速度或者时间维度。

我国政府、产业界和学术界也做了相应的理论研究和实践研究。2015 年 9 月，国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作。2016 年 3 月 17 日，《中华人民共和国国民经济和社会发展第十三个五年规划纲要》发布，其中第二十七章“实施国家大数据战略”提出：把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。具体包括：加快政府数据开放共享、促进大数据产业健康发展。

1.1.1　大数据产生的时代背景

随着计算机存储能力的提升和复杂算法的发展，近年来的数据量成指数型增长，这些趋势也使科学技术发展日新月异，商业模式发生了颠覆式变化。数据正在被商业化，来自网络、智能手机、传感器、嵌入式设备以及其他途径的数据形成了一项资产，产生了巨大的商业价值。苹果、亚马逊、Facebook、谷歌、阿里巴巴等利用大数据分析自己的优势，改变了竞争的基础，建立了全新的商业模式。稀缺数据的所有者利用数字化网络平台在一些市场近乎垄断，只需用独特的方式将数据整合分析，提供有价值的数据分析。2011 年全球的数据存储量就达到 1.8ZB，与 2011 年相比 2015 年大数据增长了近 4 倍，未来十年，全球数据存储量还将增长

十倍，大数据成为提升产业竞争力和创新商业模式的新途径。

大数据从产生到目前风靡全球，大致经历了以下 3 个发展阶段：

（1）20 世纪末至 21 世纪初：大数据的萌芽期。随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。

（2）21 世纪前 10 年：大数据的成熟期。Web2.0 应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的 GFD 和 MapReduce 等大数据技术受到追捧，Hadoop 平台开始大行其道。

（3）2010 年以后：大数据的大规模应用期。大数据应用渗透到各行各业，数据驱动决策，信息社会智能化程度大幅提高。

1.1.2 大数据的特征

目前，关于大数据的特征还有一定的争议，本书采用普遍被接受的 4V，即规模性（Volume）、多样性（Variety）、价值密度（Value）和高速性（Velocity）进行描述。

1. 数据量大（Volume）

非结构化数据的超大规模增长导致数据集合的规模不断扩大，数据单位已经从 GB 级到 TB 级再到 PB 级，甚至开始以 EB 和 ZB 来计数。

2. 类型繁多（Variety）

大数据的类型不仅包括网络日志、音频、视频、图片、地理位置信息等结构化数据，还包括半结构化数据甚至是非结构化数据，具有异构性和多样性的特点。

3. 价值密度低（Value）

大数据价值密度相对较低。如随着物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，存在大量不相关信息。因此需要对未来趋势与模式作可预测分析，利用机器学习、人工智能等进行深度复杂分析。而如何通过强大的机器算法更迅速地完成数据的价值提炼，是大数据时代亟待解决的难题。虽然单位数据的价值密度在不断降低，但是数据的整体价值在提高。

4. 速度快时效高（Velocity）

处理速度快，时效性要求高。需要实时分析而非批量式分析，数据的输入、处理和分析连贯性地处理，这是大数据区分于传统数据挖掘最显著的特征。

1.1.3 大数据应用案例

将大量的原始数据汇集在一起，通过数据挖掘等技术分析数据中潜在的规律，预测未来的发展趋势，有助于人们做出正确的决策，从而提高各个领域的运行效率，获得更大的收益。大数据冲击着许多行业，包括金融行业、互联网、医疗行业、社交网络、零售行业和电子商务等，大数据在彻底地改变着人们的生活。

1. 大数据在互联网企业的应用

互联网是最早利用大数据进行精准营销的行业，通过大数据不仅可以为企业进行精准营销，还可以快速友好地对用户实施个性化解决方案。IBM 大数据提供的服务包括数据分析、

文本分析、蓝色云杉（混搭供电合作的网络平台）、业务事件处理和商业化服务。基于对大数据价值的沉淀，依据信用体系等，马云将集团下的阿里金融与支付宝两项核心业务合并成立阿里小微金融；另外，为了便于在内部解决数据的交换、安全和匹配等问题，阿里集团还搭建了一个数据交换平台，在这个平台上，各个事业群可以实现数据的内部流转，实现价值最大化。

由于互联网的数据较为集中，数据量足够大，数据种类较多，因此未来互联网数据应用将会有更多的想象空间，包括预测流行趋势、消费趋势、地域消费特点、客户消费习惯、各种消费行为的相关度、消费热点、影响消费的重要因素等。

2. 大数据在医疗行业的应用

医疗行业拥有大量的病例、病理报告、治愈方案、药物报告等。如果这些数据可以被整理和应用将会极大地帮助医生和病人。人们面对的数目及种类众多的病菌、病毒，以及肿瘤细胞，都处于不断进化的过程中。在发现诊断疾病时，疾病的确诊和治疗方案的确定是最困难的。

借助于大数据平台可以收集不同病例和治疗方案，以及病人的基本特征，建立针对疾病特点的数据库。在医生诊断病人时可以参考病人的疾病特征、化验报告和检测报告，参考疾病数据库来快速帮助病人确诊，明确定位疾病。在制定治疗方案时，医生可以依据病人的基因特点，调取相似基因、年龄、人种、身体情况的有效治疗方案，制定出适合病人的治疗方案，帮助更多人及时进行治疗，同时这些数据也有利于医药行业开发出更加有效的药物和医疗器械。

3. 大数据在金融行业的应用

金融行业的数据具有交易量大、安全级别高等特点。银行在做信贷风险分析的时候，需要大量数据进行相关性分析，但是很多数据来源于政府各个职能部门，包括工商税务、质量监督、检察院法院等，这些数据短期仍然是无法拿到的。

摩根大通通过使用大数据技术以满足日益增多的需求，如诈骗检验、IT 风险管理和自助服务；存储大量非结构化数据，允许公司收集存储 Web 日志、交易数据和社交媒体数据，以方便以客户为中心的数据挖掘和数据分析工具的使用。光大银行将在线营销方案、微博营销、客户行为分析（包括电话语音、网络的监控录像等）和风险控制与管理（结构化非结构化数据整合，分析系统存在 IT 风险或者钓鱼网站防欺诈）等。建设银行充分跟进大数据时代的脚步，建立善融商务企业商城，在该平台上，每一笔交易，银行都有记录并且能鉴别真伪，可作为客户授信评级的重要依据。中信银行采用大数据方案，可以结合实时、历史数据进行全局分析，风险管理部门现在可以每天评估客户的行为，并决定对客户的信用额度在同一天进行调整，原有内部系统、模型整体性能显著提高。

4. 大数据在零售行业的应用

零售行业大数据应用有两个层面：一个层面是零售行业可以了解客户的消费喜好和趋势，进行商品的精准营销，降低营销成本；另一个层面是依据客户购买的产品，为客户提供可能购买的其他产品，扩大销售额，也属于精准营销范畴。另外，零售行业还可以通过大数据掌握未来消费趋势，有利于热销商品的进货管理和过季商品的处理。零售行业的数据对于产品生产厂家是非常宝贵的，零售商的数据信息将会有助于资源的有效利用，降低产能过剩，厂商依据零售商的信息按实际需求进行生产，减少不必要的生产浪费。

未来考验零售企业的不再只是供应关系的好坏，而是要看挖掘消费者需求，以及高效整合供应链满足其需求的能力，因此信息科技水平的高低成为获得竞争优势的关键要素。不论是国际零售巨头，还是本土零售品牌，要想顶住日渐微薄的利润率带来的压力，在这片领域立于不败之地，就必须思考如何利用大数据为顾客带来更好的消费体验。

5. 大数据在农业的应用

大数据在农业的应用主要是指依据未来的商业需求预测来进行农牧产品的生产，降低菜贱伤农的概率。同时大数据的分析将会更加精确预测未来的天气气候，帮助农牧民做好自然灾害的预防工作。大数据同时也会帮助农民依据消费者的消费习惯来决定增加哪些品种的种植，减少哪些品种的生产，提高单位种植面积的产值，同时有助于快速销售农产品，完成资金回流。

由于农产品不容易保存，因此合理种植和养殖就显得十分重要。如果没有规划好，容易出现鸡蛋过剩、苹果过剩、大蒜过剩、莲藕过剩等伤农事件。借助于大数据提供的消费趋势报告和消费习惯报告，政府将为农牧业生产提供合理引导，建议依据需求进行生产，避免产能过剩造成的不必要资源和社会财富浪费。农业关乎到国计民生，科学的规划将有助于社会整体效率的提升，大数据技术可以帮助政府实现农业的精细化管理，实现科学决策。

6. 大数据在交通行业的应用

近年来，我国的智能交通已实现了快速发展，许多技术手段都达到了国际领先水平。但是，问题和困境也非常突出，从各个城市的发展状况来看，智能交通的潜在价值还没有得到有效挖掘。对交通信息的感知和收集有限，对存在于各个管理系统中的海量数据无法共享运用、有效分析，对交通态势的研判预测乏力，对公众的交通信息服务很难满足需求。这其中很重要的问题是对于海量数据尤其是半结构、非结构数据无能为力。

目前，交通的大数据应用主要在两个方面：一方面可以利用大数据传感器数据来了解车辆通行密度，合理进行道路规划包括单行线路规划；另一方面可以利用大数据来实现即时信号灯调度，提高已有线路通行能力。

7. 大数据在教育行业的应用

教育行业中的考试数据、学籍数据、教师数据、事业数据、经费数据、人口数据、研究数据等都分散在不同的机构和政府部门，很难形成大数据，这是需要统筹考虑解决的问题。

教育中有两个特定的领域会用到大数据：教育数据挖掘和学习分析。教育数据挖掘应用统计学、机器学习和数据挖掘的技术和开发方法，对教学和学习过程中收集的数据进行分析，检验学习理论并引导教育实践。学习分析应用信息科学、社会学、心理学、统计学、机器学习和数据挖掘的技术来分析从教育管理和服务过程中收集的数据，学习分析创建的应用程序直接影响教育实践。

8. 大数据在政府机构的应用

政府利用大数据技术可以了解各地区的经济发展情况、各产业发展情况、消费支出和产品销售情况，依据数据分析结果科学地制定宏观政策，平衡各产业发展，避免产能过剩，有效利用自然资源和社会资源，提高社会生产效率。在以下几个方面，可以进一步协助发挥政府机构的职能作用：

（1）重视应用大数据技术，盘活各地云计算中心资产，把原来大规模投资产业园、物联网产业园的政绩工程改造成智慧工程。

（2）在安防领域，应用大数据技术，提高应急处置能力和安全防范能力。

（3）在民生领域，应用大数据技术，提升服务能力和运作效率，以及个性化的服务，比如医疗、卫生、教育等部门。

（4）解决在金融、电信等领域数据分析的问题，提高国家的金融、电信安全水平，预防电信诈骗。

1.1.4 大数据的机遇与挑战

随着近年来大数据的不断升温，人们也逐渐意识到大数据中提到的数据是全部数据，而不是随机采样；预测是大体方向，而不是精确制导。随着对大数据研究的不断深入，人们越来越意识到对数据的利用可以为其生产生活带来巨大的便利，同时也带来了不小的挑战。

1. 大数据的安全与隐私问题

在互联网上浏览网页，就会留下一连串的浏览痕迹；注册登录网站需要输入个人的重要信息，例如用户名、登录密码、手机号，甚至是身份证号、住址、银行卡等信息。通过相关的数据分析，就可以轻易挖掘出人们的行为习惯和个人重要信息。如果这些信息运用得当，可以帮助相关领域的企业随时了解客户的需求和习惯，便于企业调整相应的产品生产计划，取得更大的经济效益。但若是这些重要的信息被不良分子窃取，随之而来的就是个人信息泄露、财产丢失等安全性问题。

2. 对现有技术的挑战

（1）对现有数据库管理技术的挑战。

传统的数据库部署不能处理 TB 级别的数据，也不能很好地支持高级别的数据分析。急速膨胀的数据体量即将超越传统数据库的管理能力。如何构建全球级的分布式数据库，可以扩展到数百万的机器，数以百计的数据中心，上万亿行的数据，是今后大数据处理需要解决的问题。

（2）对经典数据库技术的挑战。

经典数据库并没有考虑数据的多类别，SQL（结构化数据查询语言）在设计的一开始是没有考虑非结构化数据的。

（3）实时性的技术挑战。

传统的数据仓库系统和各类 BI（Business Intelligence）应用对处理时间的要求并不高，因此这类应用往往运行一两天获得结果依然是可行的。但实时处理的要求是区别大数据应用和传统数据仓库技术、BI 技术的关键差别之一。

（4）对网络架构、数据中心、运维的挑战。

人们每天创建的数据量正呈爆炸式增长，但就数据保存来说，我们的技术改进不大，而数据丢失的可能性却不断增加。如此庞大的数据量首先在存储上就会是一个非常严重的问题，硬件的更新速度将是大数据发展的基石。

1.2 大数据处理技术

面对大数据的全新特征，既有的技术架构和路线已经无法高效地处理如此海量的数据，而对于相关组织来说，如果投入巨大采集的数据无法及时处理反馈有效信息，那将是得不偿

失的。可以说，大数据时代对人类的数据驾驭能力提出了新的挑战，也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。

大数据技术的不同技术层面功能和计算模式如表 1-1 和表 1-2 所示。

表 1-1　大数据技术层面功能

技术层面	功能
数据采集	利用 ETL 工具将分布的异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL 数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

表 1-2　大数据计算模式

计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark 等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb 等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala 等

1.3　大数据与云计算

云计算（Cloud Computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法，过去往往用云来表示网络，后来也用来表示互联网和底层基础设施的抽象。狭义云计算指 IT 基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需资源；广义云计算指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需服务。这种服务可以是 IT 和软件、互联网相关，也可以是其他服务。它意味着计算能力也可以作为一种商品通过互联网进行流通。

大数据，或称海量数据，指的是所涉及的资料量规模巨大到无法通过目前的主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助政府、企业经营决策的依据。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构。它的特色在于对海量数据的挖掘，

但它必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。云计算和大数据的关系如图 1-1 所示。

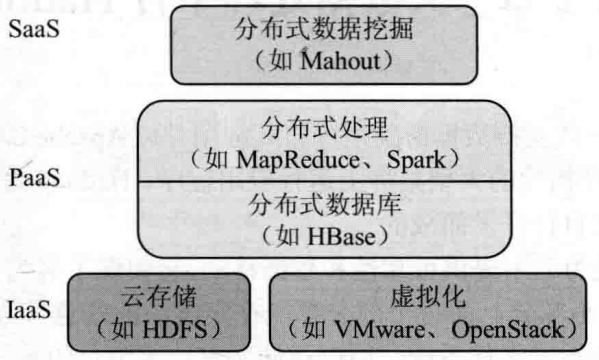

图 1-1 云计算和大数据的关系图

简单来说，云计算是硬件资源的虚拟化，而大数据是海量数据的高效处理。虽然从这个解释来看也不是完全贴切，但是却可以帮助对这两个名字不太明白的人很快理解其区别。当然，如果解释更形象一点的话，云计算相当于我们的计算机和操作系统将大量的硬件资源虚拟化后再进行分配使用。

可以说，大数据相当于海量数据的“数据库”，通观大数据领域的发展我们也可以看出，当前的大数据发展一直在向着近似于传统数据库体验的方向发展，一句话就是，传统数据库给大数据的发展提供了足够大的空间。

大数据的总体架构包括三层：数据存储、数据处理和数据分析。数据先要通过存储层存储下来，然后根据数据需求和目标来建立相应的数据模型和数据分析指标体系对数据进行分析产生价值，而中间的时效性又通过中间数据处理层提供的强大的并行计算和分布式计算能力来完成。三者相互配合，这让大数据产生最终价值。

云计算未来的趋势是：云计算作为计算资源的底层，支撑着上层的大数据处理；大数据的发展趋势是，实时交互式的查询效率和分析能力。

1.4 本章小结

大数据包含庞杂的知识体系，在具体学习相关技术之前，有必要对其有清晰直观的认识。大数据具有规模性、多样性、价值密度和高速性的特征。它虽然在金融行业、互联网、医疗行业、社交网络等方面改变着人们的生活，但是也对人们的信息安全和现有技术提出了挑战。大数据技术主要包含数据采集、数据存储和管理、数据处理与分析、数据隐私和安全等层面，而常用的计算模式有批处理计算、流计算和图计算等。本章最后阐述了大数据与云计算之间的区别和联系，使读者对两者有个清楚的了解。

第 2 章　大数据处理平台 Hadoop

Apache Hadoop 是一款支持数据密集型分布式应用并以 Apache 2.0 许可协议发布的开源软件框架，支持在商品硬件构建的大型集群上运行应用程序。Hadoop 是根据 Google 公司发表的 MapReduce 和 GFS 论文自行开发而成的。

Hadoop 框架透明地为应用提供可靠性和数据移动，它实现了名为 MapReduce 的编程范式：应用程序被切分成许多小部分，而每个部分都能在集群中的任意节点上执行或重新执行。此外，Hadoop 还提供了分布式文件系统，用以存储所有计算节点的数据，这为整个集群带来了非常高的带宽。MapReduce 和分布式文件系统的设计，使得整个框架能够自动处理节点故障。通过本章学习，您将会掌握 Hadoop 生态系统和版本变迁等知识。

2.1　Hadoop 生态系统

Hadoop 是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 2.0 版本引入了 HA（High Avalability，高可用性）和 Yarn（资源调度），这是与 Hadoop 1.0 的最大区别。Hadoop 1.0 生态系统如图 2-1 所示。

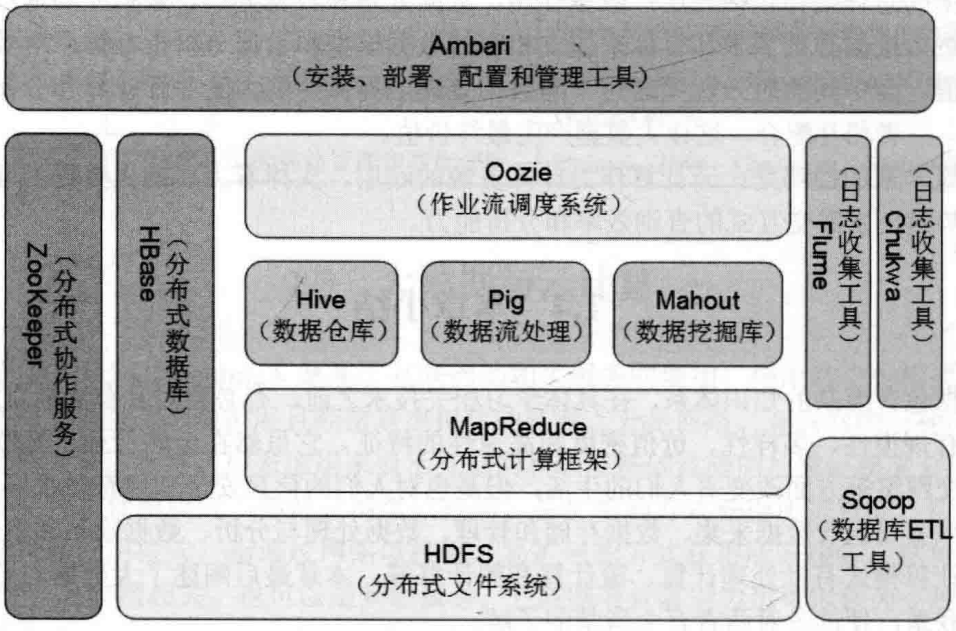

图 2-1　Hadoop 1.0 生态系统

Hadoop 2.0 主要由三部分组成：HDFS 分布式文件系统、MapReduce 编程模型和 Yarn 资源管理。Hadoop 2.0 生态系统如图 2-2 所示。

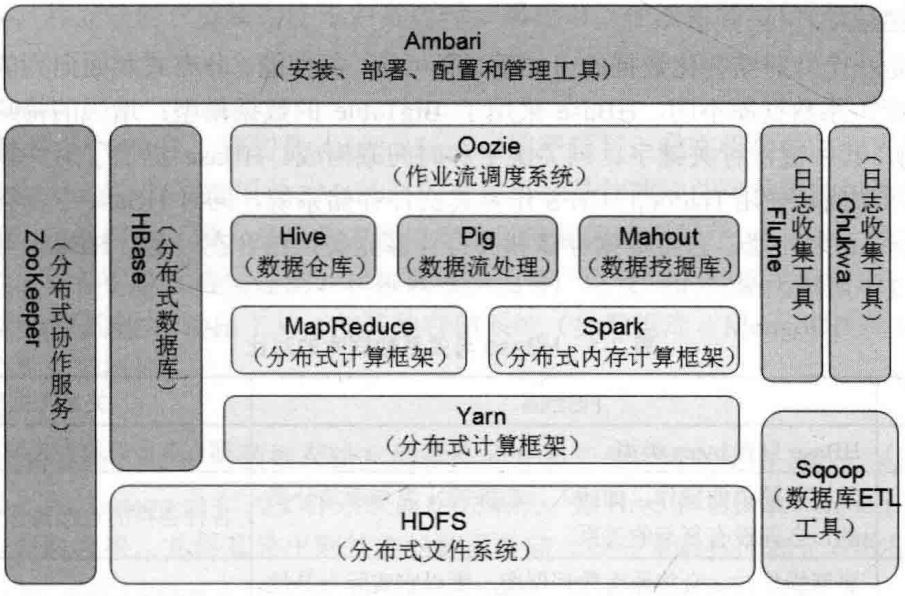

图 2-2 Hadoop 2.0 生态系统

由图 2-1 和图 2-2 可以看出，整个 Hadoop 家族由以下几个子项目组成：

（1）HDFS。

对于分布式计算，每个服务器必须具备对数据的访问能力，这就是 HDFS（Hadoop Distributed File System）所起到的作用。在处理大数据的过程中，当 Hadoop 集群中的服务器出现错误时，整个计算过程并不会终止，同时 HDFS 可以保障在整个集群中发生故障错误时的数据冗余。当计算完成时将结果写入 HDFS 的一个节点之中，HDFS 对存储的数据格式并无苛刻的要求，数据可以是非结构化或其他类别的，而关系数据库在存储数据之前需要将数据结构化并定义 Schema。

（2）MapReduce。

MapReduce 是一个计算模型，用于大规模数据集的并行运算。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。MapReduce 的重要创新是当处理一个大数据集查询时会将其任务分解并在运行的多个节点中处理。当数据量很大时就无法在一台服务器上解决问题，此时分布式计算的优势就体现出来了，将这种技术与 Linux 服务器结合可以获得性价比极高的替代大规模计算阵列的方法。Hadoop MapReduce 级的编程利用 Java APIs，并可以手动加载数据文件到 HDFS 中。

（3）ZooKeeper。

ZooKeeper 是一个分布式应用程序协调服务，是 Hadoop 和 HBase 的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括配置维护、域名服务、分布式同步、组服务等。

ZooKeeper 集群提供了 HA，可以保证在其中一些机器死机的情况下仍可以提供服务，而且数据不会丢失；所有 ZooKeeper 服务的数据都存储在内存中，且数据都是副本。ZooKeeper 集群中包括领导者（leader）和跟随者（follower）两种角色，当客户端进行读取时，跟随者的服务器负责给客户端响应；客户端的所有更新操作都必须通过领导者来处理。当更新被大多数 ZooKeeper 服务成员持久化后，领导者会给客户端响应。

（4）HBase。

HBase 是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库。与传统关系数据库不同，HBase 采用了 BigTable 的数据模型：增强的稀疏排序映射表（key/value），其中键由行关键字、列关键字和时间戳构成。HBase 提供了对大规模数据的随机、实时读写访问，使用 Hadoop HDFS 作为其文件存储系统，同时 HBase 中保存的数据可以使用 MapReduce 来处理，它将数据存储和并行计算完美地结合在一起。HBase 与关系数据库的对比如表 2-1 所示。

表 2-1　HBase 与关系数据库的对比

对比项	HBase	关系数据库
数据类型	HBase 只有 bytes 类型	拥有丰富的数据类型和存储方式
数据操作	只有很简单的操作，如插入、删除等，表与表是分离的，之间没有复杂的关系	各种各样的连接操作和函数
数据维护	更新操作时，会将原有数据保留，所以它实际上是插入了新数据	直接修改原数据
存储方式	基于列存储的，每个列族都有自己的文件，不同的列族是分开的	基于表结构和行来存储
可扩展性	支持随意的扩展，而不需要改变表内原有的数据	修改表结构需要复杂的操作
事务	没有复杂的事务支持，只有简单的行级事务	ACID 保证
索引	没有二级索引	拥有丰富的索引支持

（5）Hive。

Hive 是基于 Hadoop 的一个数据仓库工具，由 Facebook 开源，最初用于海量结构化日志数据统计，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，可以将 SQL 语句转换为 MapReduce 任务运行。通常用于进行离线数据处理（采用 MapReduce），可以认为是一个从 HQL（Hive QL）到 MapReduce 的语言翻译器。

Hive 的特点如下：

- 可扩展。Hive 可以自由地扩展集群的规模，一般情况下不需要重启服务。
- 支持 UDF。Hive 支持用户自定义函数，用户可以根据自己的需要来实现。
- 容错。良好的容错性，节点失效时 SQL 依然可以正确执行到结束。
- 自由的定义输入格式。默认 Hive 支持 txt、rc、sequence 等，用户可以自由地定制自己想要的输入格式。
- 可以根据字段创建分区表，如根据日志数据中的日期。

（6）Pig。

Pig 是一个高级过程语言，它简化了 Hadoop 常见的工作任务，适合于使用 Hadoop 和 MapReduce 平台来查询大型半结构化数据集。通过允许对分布式数据集进行类似 SQL 的查询，Pig 可以简化 Hadoop 的使用。Pig 可以加载数据、表达转换数据和存储最终结果。Pig 内置的操作使得半结构化数据变得有意义（如日志文件），同时 Pig 可以扩展使用 Java 中添加的自定义数据类型并支持数据转换。

可以避免用户书写 MapReduce 程序，由 Pig 自动转成。任务编码的方式允许系统自动去

优化执行过程，从而使用户能够专注于业务逻辑，用户可以轻松地编写自己的函数来进行特殊用途的处理。

（7）Mahout。

Mahout 起源于 2008 年，最初是 Apache Lucent 的子项目，它在极短的时间内取得了长足的发展，现在是 Apache 的顶级项目。Mahout 的主要目标是创建一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout 现在已经包含了聚类、分类、推荐引擎（协同过滤）和频繁集挖掘等广泛使用的数据挖掘方法。除了算法，Mahout 还包含数据的输入/输出工具、与其他存储系统（如数据库、MongoDB 或 Cassandra）集成等数据挖掘支持架构。

（8）Sqoop。

Sqoop 是 Hadoop 与结构化数据存储互相转换的开源工具。可以使用 Sqoop 从外部的数据存储将数据导入到 Hadoop 分布式文件系统或相关系统，如 Hive 和 HBase。Sqoop 也可以用于从 Hadoop 中提取数据，并将其导出到外部的数据存储（如关系数据库和企业数据仓库），如 MySQL、Oracle、SQL Server，还可以通过脚本快速地实现数据的导入/导出。

（9）Flume。

Flume 是 Cloudera 提供的一个高可用、高可靠、分布式的海量日志采集、聚合和传输的系统。Flume 支持在日志系统中定制各类数据发送方，用于收集数据。同时，Flume 提供对数据进行简单处理，并写到各种数据接收方（可定制）的能力。它将数据从产生、传输、处理并最终写入目标路径的过程抽象为数据流，在具体的数据流中，数据源支持在 Flume 中定制数据发送方，从而支持收集各种不同协议数据。同时，Flume 数据流提供对日志数据进行简单处理的能力，如过滤、格式转换等。总的来说，Flume 是一个可扩展、适合复杂环境的海量日志收集系统。

（10）Chukwa。

Chukwa 是一个开源的用于监控大型分布式系统的数据收集系统，构建在 Hadoop 的 HDFS 和 MapReduce 框架之上。Chukwa 还包含了一个强大和灵活的工具集，可以用于展示、监控和分析已收集的数据。

（11）Oozie。

Oozie 是一个工作流引擎服务器，用于运行 Hadoop MapReduce 任务工作流（包括 MapReduce、Pig、Hive、Sqoop 等）。Oozie 工作流通过 HPDL（Hadoop Process Defintion Languge）来构建，工作流定义通过 HTTP 提交，可以根据目录中是否有数据来运行任务，任务之间的依赖关系通过工作流来配置，任务可以定时调度。JobConf 类要配置的内容，通过在工作流（XML 文件）中定义，实现了配置与代码的分离。

（12）Ambari。

Ambari 是一种基于 Web 的工具，用于创建、管理、监视 Hadoop 的集群，支持 Hadoop HDFS、Hadoop MapReduce、Hive、Hcatalog、HBase、ZooKeeper、Oozie、Pig 和 Sqoop 等的集中管理。

2.2 Hadoop 架构

Hadoop 是一个存储和处理大规模数据的开源软件框架，实现在大量计算机组成的集群中