大数据与人工智能技术丛书

云计算与大数据技术

◎ 吕云翔 钟巧灵 张璐 王佳玮 编著

本书特色

- 注重云计算与大数据基本概念的讲解
- 以案例的方式梳理知识脉络和要点
- 提供综合云计算实验案例

清華大學出版社

大数据与人工智能技术丛书

云计算与大数据技术

◎ 吕云翔 钟巧灵 张璐 王佳玮 编著

清华大学出版社
北京

内 容 简 介

本书在阐述云计算和大数据关系的基础上，介绍了云计算和大数据的基本概念、技术及应用。全书内容如下：第1～4章讲述云计算的概念和原理，包括云计算的概论、基础、虚拟化、应用；第5～8章讲述大数据概述及基础，包括大数据概念和发展背景、大数据系统架构概述、分布式通信与协同、大数据存储；第9～13章讲述大数据处理，包括分布式处理、Hadoop MapReduce 解析、Spark 解析、流计算、集群资源管理与调度；第14章讲述综合实践(在 OpenStack 平台上搭建 Hadoop 并进行数据分析)。

本书结合实际应用及实践过程来讲解相关概念、原理和技术，实用性较强。适合作为本科院校计算机、云计算、大数据及信息管理等相关专业的教材，也适合计算机爱好者阅读和参考。

图书在版编目(CIP)数据

云计算与大数据技术/吕云翔等编著. —北京：清华大学出版社，2018
(大数据与人工智能技术丛书)
ISBN 978-7-302-50146-6

Ⅰ. ①云… Ⅱ. ①吕… Ⅲ. ①云计算－高等学校－教材 ②数据处理－高等学校－教材
Ⅳ. ①TP393.027 ②TP274

中国版本图书馆 CIP 数据核字(2018)第 112363 号

策划编辑：魏江江
责任编辑：王冰飞
封面设计：刘　键
责任校对：时翠兰
责任印制：杨　艳

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 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：11.75　　**字　　数**：285 千字
版　　次：2018 年 10 月第 1 版　　**印　　次**：2018 年 10 月第1 次印刷
印　　数：1～1500
定　　价：35.00 元

产品编号：065820-01

FOREWORD 前言

从过去的几十年以来，计算机技术的进步和互联网的发展极大地改变了人们的工作和生活方式。计算模式也经历了从最初的把任务集中交付给大型处理机到基于网络的分布式任务处理再到目前的按需处理的云计算方式的极大改变。自 2006 年亚马逊公司推出弹性计算云(EC2)服务让中小型企业能够按照自己的需要购买亚马逊数据中心的计算能力后，云计算的时代就此正式来临，“云计算”的概念随之由 Google 公司于同年提出，其本质是给用户提供像传统的电、水、煤气一样的按需计算的网络服务，是一种新型的计算使用方式。它以用户为中心，使互联网成为每一个用户的数据中心和计算中心。

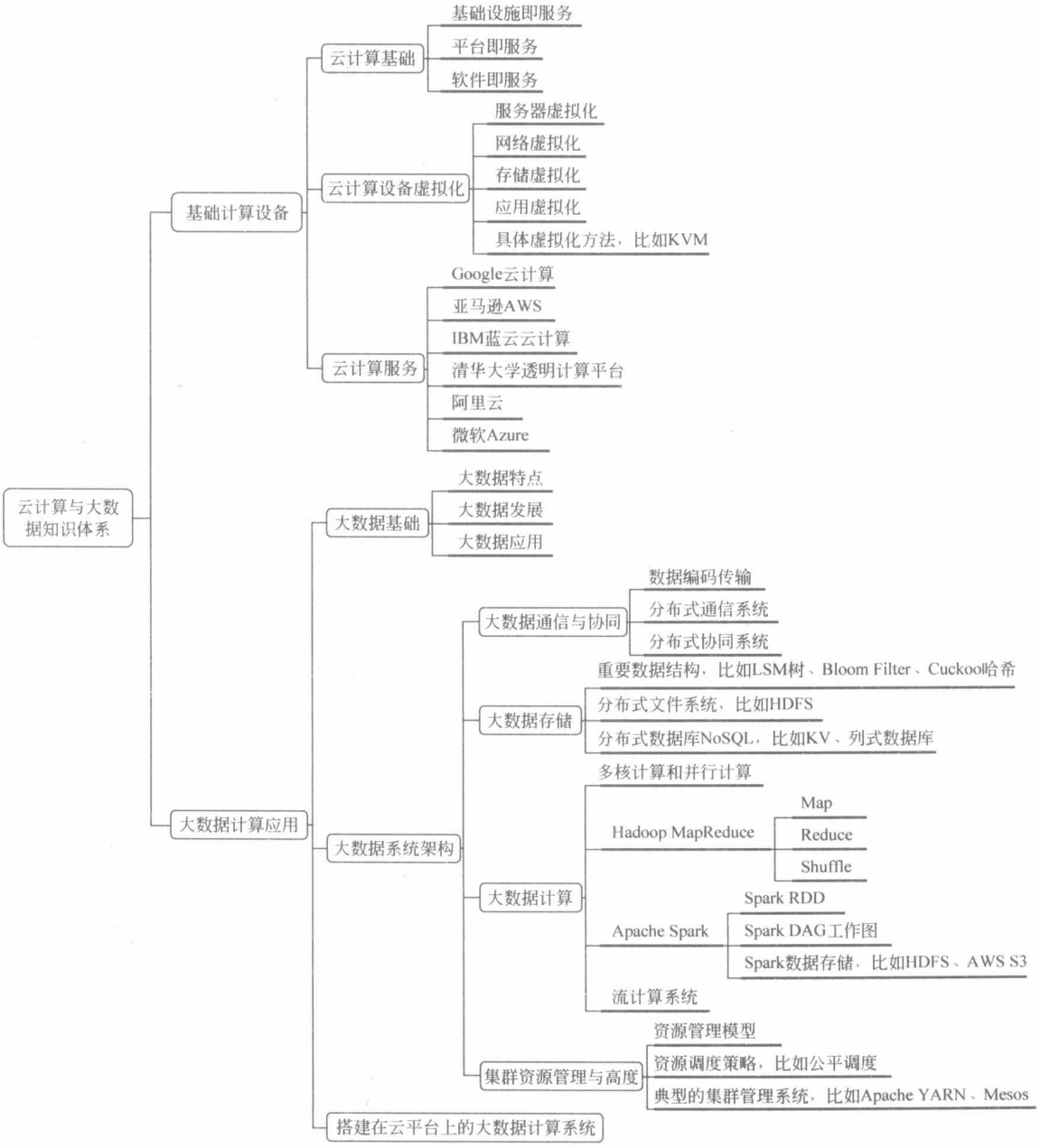
互联网技术不断发展，各种技术不断涌现，其中大数据技术已成为一颗闪耀的新星。我们已经处于数据世界，互联网每天产生大量的数据，利用好这些数据可以给我们的生活带来巨大的变化以及提供极大的便利。目前大数据技术受到越来越多的机构重视，因为大数据技术可以创造出巨大的利润，其中典型代表是个性化推荐以及大数据精准营销。

本书的各章内容如下：第 1～4 章讲述云计算的概念和原理，包括云计算的概论、基础、虚拟化、应用；第 5～8 章讲述大数据概述及基础，包括大数据概念和发展背景、大数据系统架构概述、分布式通信与协同、大数据存储；第 9～13 章讲述大数据处理，包括分布式处理、Hadoop MapReduce 解析、Spark 解析、流计算、集群资源管理与调度；第 14 章讲述综合实践(在 OpenStack 平台上搭建 Hadoop 并进行数据分析)。

本书对云计算和大数据的概念和基础讲解详细，力求通过实例进行描述，并可通过综合实践篇章将理论联系实际，适合计算机相关专业的读者，以及计算机爱好者阅读和参考。本书的作者为吕云翔、钟巧灵、张璐、王佳玮，另外，曾洪立、吕彼佳、姜彦华进行了素材整理及配套资源制作等。

在本书的编写过程中，我们尽量做到仔细认真，但由于我们的水平有限，书中还是可能会出现一些疏漏与不妥之处，在此非常欢迎广大读者进行批评指正。同时也希望广大读者可以将自己读书学习的心得体会反馈给我们(yunxianglu@hotmail.com)。

作　者



CONTENTS 目录

第1章

云计算概论

本章介绍云计算的定义，旨在让读者对云计算有一个宏观的概念，然后介绍云计算的产生背景，接着介绍云计算的发展历史。通过本章的学习，读者将对云计算有一个初步的认识。

1.1 什么是云计算

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电信网，后来也用来表示互联网和底层基础设施的抽象。因此，云计算甚至可以让人们体验每秒10万亿次的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。用户可通过计算机、笔记本、手机等方式接入数据中心，按自己的需求进行运算。

对云计算的定义有多种说法。对于到底什么是云计算，至少可以找到100种解释。现阶段广为接受的是美国国家标准与技术研究院(NIST)的定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件、服务)，这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。

1.2 云计算的产生背景

云计算是继20世纪80年代大型计算机到客户/服务器的大转变之后的又一种巨变。

云计算是分布式计算(Distributed Computing)、并行计算(Parallel Computing)、效用

计算（Utility Computing）、网络存储（Network Storage Technologies）、虚拟化（Virtualization）、负载均衡（Load Balance）、热备份冗余（High Available）等传统计算机和网络技术发展融合的产物。

1.3 云计算的发展历史

1983年，太阳微系统公司（Sun Microsystems）提出“网络是计算机”的概念，2006年3月，亚马逊公司（Amazon）推出弹性计算云（Elastic Compute Cloud，EC2）服务。

2006年8月9日，Google公司首席执行官埃里克·施密特（Eric Schmidt）在搜索引擎大会（SES San Jose 2006）首次提出云计算的概念。Google“云端计算”源于Google工程师克里斯托弗·比希利亚所做的Google 101项目。

2007年10月，Google与IBM公司开始在美国大学校园，包括卡内基·梅隆大学、麻省理工学院、斯坦福大学、加州大学伯克利分校及马里兰大学等，推广云计算的计划，这项计划希望能降低分布式计算技术在学术研究方面的成本，并为这些大学提供相关的软硬件设备及技术支持（包括数百台个人计算机及BladeCenter与System x服务器，这些计算平台将提供1600个处理器，支持包括Linux、Xen、Hadoop等开放源代码平台）。而学生则可以通过网络开发各项以大规模计算为基础的研究计划。

2008年1月30日，Google公司宣布在中国台湾启动“云计算学术计划”，与台湾台大、交大等学校合作，将云计算技术推广到校园的学术研究中。

2008年2月1日，IBM公司宣布将在中国无锡太湖新城科教产业园为中国的软件公司建立全球第一个云计算中心（Cloud Computing Center）。

2008年7月29日，雅虎、惠普和英特尔公司宣布一项涵盖美国、德国和新加坡的联合研究计划，推进云计算的研究进程。该计划要与合作伙伴创建6个数据中心作为研究实验平台，每个数据中心配置1400～4000个处理器。这些合作伙伴包括新加坡资讯通信发展管理局、德国卡尔斯鲁厄大学Steinbuch计算中心、美国伊利诺伊大学香槟分校、英特尔研究院、惠普实验室和雅虎。

2008年8月3日，美国专利商标局网站信息显示，戴尔正在申请云计算商标，此举旨在加强对这一未来可能重塑技术架构的术语的控制权。

2010年3月5日，Novell公司与云安全联盟（CSA）共同宣布一项供应商中立计划，名为“可信任云计算计划”。

2010年7月，美国国家航空航天局和包括Rackspace、AMD、Intel、戴尔等支持厂商共同宣布OpenStack开放源代码计划，微软公司在2010年10月表示支持OpenStack与Windows Server 2008 R2的集成；而Ubuntu已把OpenStack加至其11.04版本中。

2011年2月，思科公司正式加入OpenStack，重点研制OpenStack的网络服务。

2013年，我国的IaaS（基础设施即服务）市场规模约为10.5亿元，增速达到了105%，显示出旺盛的生机。IaaS相关企业不仅在规模、数量上有了大幅提升，而且吸引了资本市场的关注，UCloud、青云等IaaS初创企业分别获得了千万美元级别的融资。

过去几年里，腾讯、百度等互联网巨头纷纷推出了各自的开放平台战略。新浪SAE等PaaS（平台即服务）的先行者也在业务拓展上取得了显著的成效，在众多互联网巨头的介入

和推动下，我国 PaaS 市场得到了迅速发展，2013 年市场规模增长近 20%。但由于目前国内 PaaS 仍处于吸引开发者和产业生态培育的阶段，大部分 PaaS 都采用免费或低收费的策略，因此整体市场规模并不大，估计约为 2.2 亿元人民币，但这并不妨碍人们对 PaaS 的发展前景抱有充足的信心。

无论是国内还是国外，SaaS（软件即服务）一直是云计算领域最为成熟的细分市场，用户对于 SaaS 的接受程度也比较高。2015 年，SaaS 市场增长率达到 117.5%，市场规模增长至 8.1 亿元人民币。

2015 年以来，云计算方面的相关政策不断。2015 年年初，国务院发布了《国务院关于促进云计算创新发展培育信息产业新业态的意见》，明确了我国云计算产业的发展目标、主要任务和保障措施。2015 年 7 月，国务院又发布了《关于积极推进“互联网＋”行动的指导意见》，提出到 2025 年，“互联网＋”成为经济社会创新发展的重要驱动力量。2015 年 11 月，工业和信息化部印发《云计算综合标准化体系建设指南》。

1.4 如何学好云计算

云计算是一种基于互联网的计算方式，要实现云计算则需要一整套的技术架构，包括网络、服务器、存储、虚拟化等。云计算目前分为公有云和私有云。两者的区别只是提供服务的对象不同，一个是企业内部使用，一个则是面向公众。目前企业中的私有云都是通过虚拟化来实现的，建议可以了解一下虚拟化行业的前景和发展。

虚拟化目前分为服务器虚拟化（以 VMware 为代表）、桌面虚拟化（思杰比 VMware 的优势大）、应用虚拟化（以思杰为代表）。学习虚拟化需要的基础如下。

（1）操作系统。了解 Windows 操作系统（如 Windows Server 2008、Windows Server 2003、Windows 7、Windows 8、Windows 10 等）的安装和基本操作、AD 域角色的安装和管理、组策略的配置和管理。

（2）数据库的安装和使用（如 SQL Server）。

（3）存储的基础知识（如磁盘性能、RAID、IOPS、文件系统、FC SAN、iSCSI、NAS 等）、光纤交换机的使用、使用 Open E 管理存储。

（4）网络的基础知识（如 IP 地址规划、VLAN、Trunk、STP、Etherchannel）。

习题

1. 美国国家标准与技术研究院（NIST）是如何定义云计算的？
2. 云计算的发展历史经历了哪些过程？
3. 虚拟化指的是什么？

第2章

云计算基础

本章主要介绍关于云计算的各种基础知识，包括分布式计算、云计算的基本概念、实现云计算的几种关键技术以及云交付和部署模式，同时介绍云计算有哪些优势、面临的挑战以及几种典型的云应用。通过本章的学习，读者应能够对云计算有一个基本的认识。

2.1 分布式计算

分布式计算是一种计算方法，和集中式计算是相对的。随着计算技术的发展，一些应用需要巨大的计算能力才能完成，如果采用集中式计算，则需要耗费很长的时间才能完成。而分布式计算将应用分解成许多更小的部分，分配到多台计算机进行处理，这样可以节省整体计算时间，大大提高计算效率。云计算是分布式计算技术的一种，也是分布式计算这种科学概念的商业实现。

分布式计算的优点就是发挥“集体的力量”，将大任务分解成小任务，分配给多个计算节点同时去计算。分布式计算将计算扩展到多台计算机，甚至是多个网络，在网络上有序地执行一个共同的任务，当然离不开 Web 技术，但在分布式计算发展起来之前的网络协议并不能满足分布式计算的要求，于是产生了 Web Service 技术。

分布式计算的另一种应用是 Web Service，Web Service 是一个平台独立的、低耦合的、自包含的、基于可编程的 Web 的应用程序，可使用开放的 XML（标准通用标记语言下的一个子集）标准来描述、发布、发现、协调和配置这些应用程序，用于开发分布式的、互操作的应用程序。

如图 2-1 所示，Web Service 的体系结构是基于 Web 服务提供者、Web 服务请求者、Web 服务注册

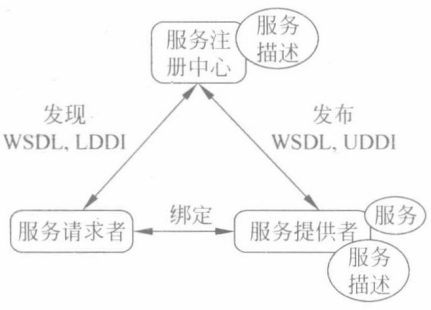


图 2-1 Web Service 的体系结构

中心三个角色和发布、发现、绑定三个动作构建的。简单地说，Web 服务提供者就是 Web 服务的拥有者，等待为其他服务和用户提供自己已有的功能；Web 服务请求者就是 Web 服务功能的使用者，利用 SOAP 消息向 Web 服务提供者发送请求以获得服务；Web 服务注册中心的作用是把一个 Web 服务请求者与合适的 Web 服务提供者联系在一起，它充当管理者的角色，一般是 UDDI(Universal Description Discovery and Integration)。这三个角色是根据逻辑关系划分的，在实际应用中，角色之间很可能有交叉：一个 Web 服务既可以是 Web 服务提供者，也可以是 Web 服务请求者，或者二者兼而有之，显示了 Web 服务角色之间的关系，其中，“发布”是为了让用户或其他服务知道某个 Web 服务的存在和相关信息；“发现”是为了找到合适的 Web 服务；“绑定”则是在提供者与请求者之间建立某种联系。

简单地说，这种技术的功能和中间件的功能有相似之处：Web Service 技术是屏蔽掉不同开发平台开发的功能模块的相互调用的障碍，从而可以利用 HTTP 和 SOAP 使商业数据在 Web 上传输，可以调用这些开发平台不同的功能模块来完成计算任务。这样看来，要在互联网上实施大规模的分布式计算，就需要 Web Service 作支撑。

2.2　云计算的基本概念

云计算已经成为一个大众化的词语，似乎每个人对于云计算的理解各不相同，第 1 章已经对云计算有一个宏观的概念和通俗的理解，如图 2-2 所示，云计算的“云”就是存在于互联网上的服务器集群上的资源，它包括硬件资源(服务器、存储器、CPU 等)和软件资源(应用软件、集成开发环境等)，本地计算机只需要通过互联网发送一个需求信息，远端就有成千上万的计算机为用户提供需要的资源并将结果返回给本地计算机。这样，本地计算机几乎不需要做什么，所有的处理都在云计算提供商所提供的计算机群来完成。简而言之，云计算是一种商业计算模型，它将计算任务分布在大量计算机构成的资源池上，使用户能够按需获取计算力、存储空间和信息服务。

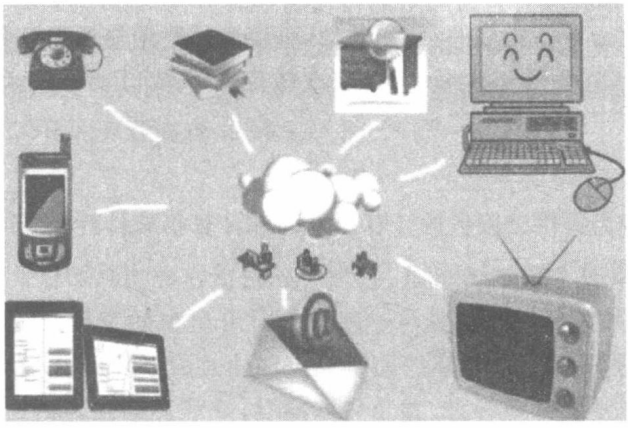


图 2-2　云计算

最简单的云计算技术在网络服务中已经随处可见，例如搜索引擎、网络信箱等，使用者只需要输入简单的指令即能得到大量信息。