Hadoop Cloud Computing Practice

由浅入深，理论结合实践，全面、系统地介绍了Hadoop云计算的配套项目内容及其应用，具有很高的实用价值。

周　品◎主编

Hadoop 云计算实战

Hadoop Cloud Computing Practice

- 合理、完善的体系结构
- 原理与技术的完美结合
- 云计算应用的最佳手册
- 教学与科研的最新成果

清华大学出版社

Hadoop Cloud Computing Practice

周 品◎主编

Hadoop 云计算实战

Hadoop Cloud Computing Practice

清华大学出版社
北 京

内 容 简 介

本书全面介绍了云计算的基本概念、Google（谷歌）云计算的关键技术，以及 Hadoop 云计算的相关配套项目及其实战，包括 Hadoop 的 HDFS、MapReduce、HBase、Hive、Pig、Cassandra、Chukwa 及 ZooKeeper 等配套项目的实现机制、用法及应用。

本书可作为高等院校本科生和研究生的教材，也可作为广大科研人员、学者、工程技术人员的参考用书。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目（CIP）数据

Hadoop 云计算实战/周品主编. —北京：清华大学出版社，2012.10

ISBN 978-7-302-29673-7

I. ①H…　II. ①周…　III. ①数据处理-应用软件　IV. ①TP274

中国版本图书馆 CIP 数据核字（2012）第 184710 号

责任编辑：钟志芳
封面设计：刘　超
版式设计：文森时代
责任校对：柴　燕
责任印制：何　芊

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
印 刷 者：北京季蜂印刷有限公司
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：26.5　　**字　　数**：612 千字
版　　次：2012 年 10 月第 1 版　　**印　　次**：2012 年 10 月第1 次印刷
印　　数：1～4000
定　　价：46.00 元

产品编号：048237-01

前　言

2010年，云计算被看成是第三次IT浪潮，成为中国战略性新兴产业的重要组成部分，是当前关注的热点之一。随着Google（谷歌）、Amazon（亚马逊）、Salesforce的巨大成功，云计算作为IT发展的下一个方向已经基本得到了确认。云计算是如此高效，正让整个IT行业发生深刻变革。

什么是云计算？云计算是一种基于互联网的超级计算模式，在远程的数据中心，几万甚至几千万台电脑和服务器连接成一片。因此，云计算甚至可以让用户体验每秒超过100000亿次的运算能力，如此强大的运算能力几乎无所不能。用户通过电脑、笔记本、手机等方式接入数据中心，按各自的需求进行存储和运算。

云计算被视为科技业的下一次革命，它将带来工作方式和商业模式的根本性改变。首先，对中小企业和创业者来说，云计算意味着巨大的商业机遇，他们可以借助云计算在更高的层面上和大企业竞争。自1989年微软推出Office办公软件以来，人们的工作方式已经发生了极大变化，而云计算则带来了云端的办公室……

云计算在IT市场上的雏形正在逐步形成，它为供应商提供了全新的机遇并催生了传统IT产品的转变。云计算用于IT服务支出已占据五大关键市场单元的9%的份额。更重要的是，在云计算上的支出将进入加速发展阶段，2012年将实现IT支出增长总量中25%的份额。

云计算与当今同样备受关注的3G和物联网是互为支撑、交互辉映的关系。3G为“云计算”带来数以亿计的宽带移动用户。移动终端的计算能力和存储空间有限，却有很强的联网能力，如果有“云计算”平台的支撑，移动用户将获得前所未有的服务体验；物联网使用数量惊人的传感器、RFID和视频监控单元等，采集到海量的数据，通过3G和宽带互联网进行传输，如果汇聚到云计算设施进行存储和处理，则可以更加迅速、准确、智能、低成本地对物理世界进行管理和控制，大幅提高社会生产力水平和生活质量。

Hadoop是一个开源框架，遵循Google的方法实现了MapReduce算法，用以查询在互联网上分布的数据集。这个定义自然会导致一个明显的问题：什么是Map（映射），为什么它们需要被Reduce（化简）？使用传统机制分析和查询大规模数据集会非常困难，当查询自身很复杂时尤是如此。

实际上，MapReduce算法将查询操作和数据集都分解为组件，即映射。在查询中被映射的组件可以被同时处理（即化简），从而快速地返回结果。

Hadoop主要由以下几个子项目组成。

- Hadoop Common：是支撑Hadoop的公共部分，包括文件系统、远程过程调用（RPC）和序列化函数库等。
- HDFS：提供高吞吐量的可靠分布式文件系统，是GFS的开源实现。
- MapReduce：为大型分布式数据处理模型，是Google MapReduce的开源实现。

与Hadoop直接相关的配套开源项目还包括如下几个方面。

- HBase：支持结构化数据存储的分布式数据库，是 Bigtable 的开源实现。
- Hive：提供数据摘要和查询功能的数据仓库。
- Pig：是在 MapReduce 上构建的一种高级的数据流语言，可简化 MapReduce 任务的开发。
- Cassandra：由 Fackbook 支持的开源、高可扩展分布式数据库。是 Amazon 库层架构 Dynamo 的全分布和 Google Bigtable 的列式数据存储模型的有机结合。
- Chukwa：用来管理大型分布式系统的数据采集系统。
- ZooKeeper：用于解决分布式系统中一致性问题，是 Chubby 的开源实现。

经过近几年的发展，在所有的开源云计算系统中，Hadoop 一直稳居第一。

本书共分为 11 章，其主要内容如下。

第 1 章：介绍云计算概论，主要包括云计算的定义、产生背景、发展史、层次结构、应用研究等内容。

第 2 章：介绍 Hadoop 相关项目，主要包括 Hadoop 简介、Hadoop 系统性质、SQL 数据库与 Hadoop 的比较及 Hadoop 的配套开源项目等内容。

第 3 章：介绍 Hadoop 配置与实战，主要包括 Hadoop 的安装、运行 Hadoop、Hadoop 的 Avatar 机制及 Hadoop 实战等内容。

第 4 章：介绍 Hadoop 的分布式数据 HDFS，主要包括 HDFS 的操作、FS Shell 使用指南及 API 使用等内容。

第 5 章：介绍 Hadoop 编程模型 MapReduce，主要包括 MapReduce 基础、MapReduce 的容错性、MapReduce 实例分析及 MapReduce 作业分析等内容。

第 6 章：介绍 Hadoop 的数据库 HBase，主要包括 HBase 体系结构、HBase 数据模型、HBase 与 RDBMS 对比及 HBase 实例分析等内容。

第 7 章：介绍 Hadoop 的数据仓库 Hive，主要包括 Hive 的安装、Hive 的入口、Hive QL 详解、Hive 的服务、Hive SQL 的优化、Hive 的扩展性及 Hive 的实战等内容。

第 8 章：介绍 Hadoop 的大规模数据平台 Pig，主要包括 Pig 的安装与运行、Pig 实现、Pig Latin 语言及 Pig 实战等内容。

第 9 章：介绍 Hadoop 的非关系型数据 Cassandra，主要包括 Cassandra 的安装、Cassandra 的数据模型、Cassandra 的实例分析及 Cassandra 实战等内容。

第 10 章：介绍 Hadoop 的收集数据 Chukwa，主要包括 Chukwa 的安装与配置、Chukwa 数据流处理、Chukwa 源代码分析及 Chukwa 实战等内容。

第 11 章：介绍 Hadoop 的分布式系统 ZooKeeper，主要包括 ZooKeeper 的安装与配置、ZooKeeper 的 Leader 流程、ZooKeeper 锁服务及 ZooKeeper 的典型应用等内容。

本书可作为广大在校本科生和研究生的学习用书，也可作为广大科研人员、学者、工程技术人员的参考用书。

本书由周品主编，参与编写的还有丁伟雄、雷晓平、李娅、杨文茵、何正风、赵新芬、赵书梅、栾颖、刘志为、周灵、余智豪、赵书兰和崔如春。

由于作者水平有限，加之时间紧迫，书中难免会存在不足之处，敬请广大读者批评指正。

编　者

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第 1 章　云计算概论

很少有一种技术能够像“云计算”（Cloud Computing）这样，在短短的两三年间就产生巨大的影响力。Google、Amazon、IBM 和微软等 IT 巨头们以前所未有的速度和规模推动云计算技术和产品的普及，一些学术活动迅速提上议事日程。

1.1　云计算概述

“云计算”被称为继个人计算机、互联网之后的第三次信息化革命，通过与相关技术创新要素、商业模式创新要素形成新革命，通过与相关技术创建要素、商业模式创建要素形成有机互动，“云计算”将成为推动电信业乃至广义 ICT 产业下一轮突破发展的重要驱动力。

1.1.1　云计算的定义

云计算是在 2007 年第三季度才诞生的新名词，但仅仅过了半年多，其受到关注的程度就超过了网格计算（Grid Computing），如图 1-1 所示。

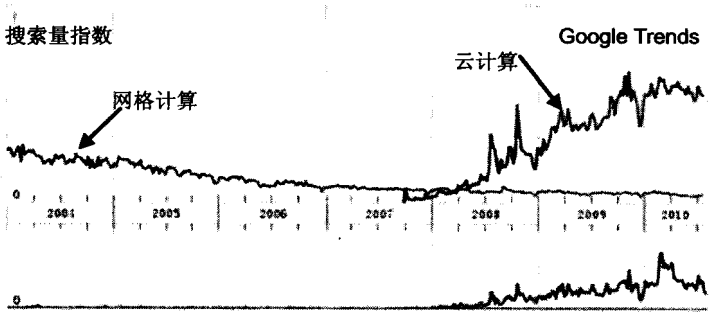

图 1-1　云计算和网格计算在 Google 中的搜索趋势图

云计算迄今为止还没有统一的定义，不同的组织从不同的角度给出了不同的定义，根据不完全的统计至少有 25 种以上。例如，Gartner 认为，云计算是一种使用网络技术并由 IT 使能而具有可扩展性和弹性能力作为服务提供给多个外部用户的计算方式；美国国家标准与技术实验室对云计算的定义为：“云计算是一个提供便捷的通过互联网访问一个可定制的 IT 资源共享池能力的按使用量付费模式（IT 资源包括网络、服务器、存储、应用、服务），这些资源能够快速部署，并只需要很少的管理工作或很少的与服务供应商的交互。”随着应用场景的变化和使能技术的发展，关于云计算的定义还在不断产生新的观点。

云计算将网络上分布的计算、存储、服务构件、网络软件等资源集中起来，基于资源虚拟化的方式，为用户提供方便快捷的服务，其可以实现计算与存储的分布式与并行处理。如果把“云”视为一个虚拟化的存储与计算资源池，那么云计算则是这个资源池基于网络平台为用户提供的数据存储和网络计算服务。互联网是最大的一片“云”，其上的各种计算机资源共同组成了若干个庞大的数据中心及计算中心。

但是，云计算并不是一个简单的技术名词，并不仅仅意味着一项技术或一系列技术的组合。其所指向的是 IT 基础设施的交付和使用模式，即通过网络以按需、易扩展的方式获得所需的资源（硬件、平台、软件）。提供资源的网络被称为“云”。从更广泛的意义上来看，云计算是指服务的交付和使用模式，即通过网络以按需、易扩展的方式获得所需的服务，这种服务可以是 IT 基础设施（硬件、平台、软件），也可以是任意其他的服务。无论是狭义还是广义，云计算所秉承的核心理念是“按需服务”，就像人们使用水、电、天然气等资源的方式一样。这也是云计算对于 ICT 领域乃至于人类社会发展最重要的意义所在。

1.1.2 云计算产生的背景

有人说云计算是技术革命的产物，有人说云计算只不过是已有技术的重新包装，是设备厂商或软件厂商“换汤不换药”的一种商业策略。笔者认为，云计算的发展是需求推动、技术进步及商业模式转换共同作用下的结果。

（1）需求是云计算的动力。

IT 设施要成为社会基础设施，现在面临高成本的瓶颈，这些成本至少包括人力成本、资金成本、时间成本、应用成本、环境成本。云计算带来的益处是显而易见的：用户不需要专门的 IT 团队，不需要购买、维护、安放有形的 IT 产品，可以低成本、高效率、随时、快捷地按需使用 IT 服务；云计算服务提供商可以极大提高资源（硬件、软件、空间、人力、物力、资源等）的利用率和业务响应速度，有效聚合产业链。

（2）技术是云计算发展的基础。

云计算自身核心技术的发展，如硬件技术、虚拟化技术（计算虚拟化、网络虚拟化、存储虚拟化、桌面虚拟化、应用虚拟化）、海量存储技术、分布式并行计算、多用户构架、自动管理与部署；云计算赖以存在的移动互联网技术的发展，如高速和大容量的网络、无处不在的接入、灵活多样的终端、集约化的数据中心 Web 技术。

（3）商业模式是云计算的内在要求，是用户需求的外在体现，并且云计算技术为这种特定商业模式提供了实现可能性。

从商业模式的角度看，云计算的主要特征是以网络为中心、以服务为产品形态、按需使用与付费，这些特征分别对应于传统的用户自建基础设施、购买有形产品或介质（含liscence）、一次性买断。

纯粹从技术角度看，云计算是很多技术自然发展、精心优化与组合的产物，是这些技术的集大成者；另一方面，如果同时考虑到商业模式，那么可断言，云计算将给整个社会的信息化带来革命性的改变。所以，在此绝不能离开技术谈云计算，否则有“忽悠”之嫌；也不能离开商业模式谈云计算，否则云计算就是“无源之水，无根之木”。

1.1.3 云时代谁是主角

纵观整个 ICT 产业发展历程，每一次计算模式的变革都会引发一场产业变革，同时也会造就一批“名星”厂商。主机时代，IBM 风光无限，称霸一时；互联网时代，微软、Intel 即为名副其实的行业主导者。然而，云计算时代，谁才是主角?

1. Google

Google 是最早提倡和实践云计算技术的企业之一，其互联网搜索服务便建立在云计算基础架构之上。经过多年的发展，Google 云计算技术逐渐成熟，针对自身特点建立了一套极其有效的商业模式与产品、服务组合。2011 年 8 月份，Google 以 6820 万美元收购企业级 IP 通信解决方案提供商 GIPS；2011 年 5 月 10 日在旧金山召开“Google I/O 开发者大会”，发布了以企业级 Google App 为核心的云计算产品。种种事实表明，Google 云计算目标并不只在于个人用户，它的野心在于覆盖从个人用户至企业用户的广大空间。Google 还积极与其他云计算企业合作。作为全球最具影响力的高科技企业之一的 Google，正以一种先行者的姿态拥抱云计算时代的到来。

Google 标志如图 1-2 所示。

2. IBM

IBM 是云计算领域中名副其实的巨头，2007 年高调启动“蓝云”计划，推出一系列云计算产品。2008 年，IBM 在云计算领域的累积投入超过了 10 亿美元，将其云计算产品和服务扩展到亚洲、欧洲、非洲、美洲市场。为了进一步抢占全球云计算市场，从 2009 年开始，IBM 加大了在云计算领域上的投入。有消息透露 IBM 将投资 200 亿美元进行并购、开发云计算终端、推出网络软件……摆出了一副势在必得的架势。IBM 在 IaaS、PaaS、SaaS 3 个层面都有方案推出，公有云、私有云、混合云一个不落。近两年来，IBM 的“智慧”战略如火如荼，智慧地球、智慧城市、智慧通信、智慧医疗……一切都是智慧的。IBM 智慧的云计算也是其智慧战略中的重要组成部分，其云智慧正不断向云计算领域延深。

IBM 标志如图 1-3 所示。

图 1-2 Google 标志

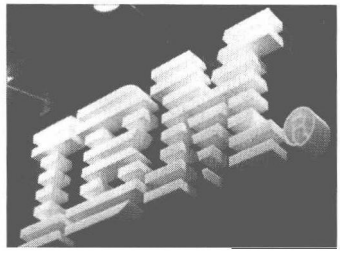

图 1-3 IBM 标志

3. 微软

云+端、软件+服务是对微软云计算的最佳诠释，其云计算平台 Windows Azure 被认为是 Windows NT 之后，16 年来最重要的产品，关乎微软的未来。微软 CEO 史蒂夫·鲍尔默

多次公开表示，云计算是微软的又一次机遇。与其他云计算厂商相比，微软在用户上有着明显的优势。微软的操作系统和操作习惯，不管是在个人用户，还是企业用户中都有很广泛的影响力。微软可以借助这些优势迅速推广其云计算产品和服务，微软基于云计算的解决方案正在得到越来越广泛的应用。近日有消息称微软云计算用户已经超过 Google App 用户。这表明，微软云计算产品的势力范围正在逐步扩大，将对产业链上下游产生深远影响。

微软标志如图 1-4 所示。

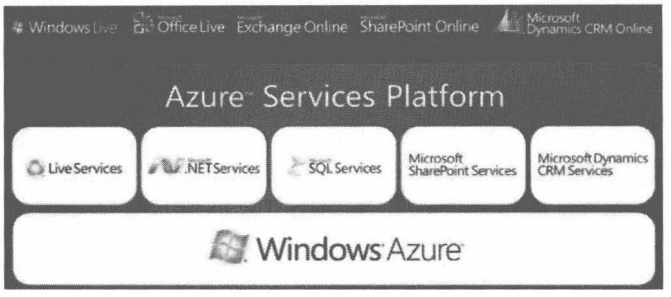

图 1-4 微软标志

1.1.4 云计算的特征

之所以称为“云”，是因为它在某些方面具有现实中云的特征，例如：

- 云一般都较大。
- 云的规模可以动态伸缩，它的边界是模糊的。
- 云在空中飘忽不定，无法也无须确定它的具体位置，但它确实存在于某处。同时还因为云计算的鼻祖之一亚马逊（Amazon）公司将大家曾经称为网格计算的东西，取了一个新名“弹性计算云”（Elastic Computing Cloud），并取得了商业上的成功。

有人将这种模式比作从单台发电机供电模式转向了电厂集中供电的模式。其意味着计算能力也可以作为一种商品进行流通，就像天然气、水及电一样，使用方便，费用低廉。但最大的不同在于，它是通过互联网进行传递的。

云计算是并行计算（Parallel Computing）、分布式计算（Distributed Computing）及网格计算（Grid Computing）的发展，或者说是这些计算科学概念的商业实现。云计算是虚拟化（Virtualization）、效用计算（Utility Computing）、将基础设施作为服务 IaaS（Infrastructure as a Service）、将平台作为服务 PaaS（Platform as a Service）和将软件作为服务 SaaS（Software as a Service）等概念混合演进并跃升的结果。从研究现状上看，云计算具有以下特点：

（1）超大规模。“云”具有相当的规模，Google 云计算已经拥有 100 多万台服务器，亚马逊、IBM、微软和 YAHOO!（雅虎）等公司的“云”均拥有几十万台服务器等。“云”能赋予用户前所未有的计算能力。

（2）虚拟化。云计算支持用户随时、随地、使用各种终端获取服务。所请求的资源来自“云”，而不是固定的有形的实体。应用在“云”中某处运行，但实际上用户无须了解应用运行的具体位置，只需要一台笔记本或一个 PDA，就可以通过网络服务来获取各种能力超强的服务。

此为试读，需要完整PDF请访问：www.ertongbook.com