Hadoop

开源云计算平台

刘 刚 侯 宾 翟周伟 编著

北京邮电大学出版社
www.buptpress.com

Hadoop 开源云计算平台

刘 刚　侯 宾　翟周伟　编著

北京邮电大学出版社
www. buptpress. com

内 容 简 介

本书首先介绍了云计算的基本概念以及谷歌云计算的关键技术，然后全面系统地介绍了实现云计算关键技术层的理想开源工具 Hadoop 及其应用。本书阐述了 Hadoop 中每个部分的实现机制与用法，包括 HDFS、Hadoop FS shell、Map/Reduce、Hadoop 流与管道机制、Hadoop I/O、Hadoop 命令简介、部署 Hadoop，并介绍了 Zookeeper、HBase、Pig、Hive、CloudBase、Mahout。除此之外本书还介绍了基于 Hadoop 的开发与应用。

图书在版编目(CIP)数据

Hadoop 开源云计算平台/刘刚，侯宾，翟周伟编著. --北京：北京邮电大学出版社，2011.8
ISBN 978-7-5635-2690-1

Ⅰ.①H… Ⅱ.①刘…②侯…③翟… Ⅲ.①数据处理—应用软件 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2011)第 146049 号

书　　名：Hadoop 开源云计算平台
作　　者：刘 刚　侯 宾　翟周伟
责任编辑：刘　颖
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路
发 行 部：电话：010-62282185
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：北京联兴华印刷厂
开　　本：787 mm×1 092 mm　1/16
印　　张：14.5
字　　数：355 千字
印　　数：1—3 000 册
版　　次：2011 年 8 月第 1 版　2011 年 8 月第 1 次印刷

ISBN 978-7-5635-2690-1　　定　价：26.00 元

前　言

2010年，云计算被看做是第三次IT浪潮，它成为了中国战略性新兴产业的重要组成部分。它将带来生活、生产方式和商业模式的根本性改变，是当前全社会关注的热点之一。随着谷歌、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。

2010年10月23日，中华人民共和国国家发展和改革委员会、中华人民共和国工业和信息化部《关于做好云计算服务创新发展试点示范工作的通知》中指出：为加强我国云计算创新发展顶层设计和科学布局，推进云计算中心（平台）建设，在充分考虑各地区产业发展情况的基础上，经研究中华人民共和国国家发展和改革委员会、中华人民共和国工业和信息化部拟按照自主、可控、高效原则，在北京、上海、深圳、杭州、无锡5个城市先行开展云计算创新发展试点示范工作。2010年10月18日由北京市科学技术研究院计算中心打造的云计算平台已经建成。这个“云平台”拥有每秒百万亿次的超强计算能力，是目前国内最大的工业云计算服务平台。从上述中不难看出云计算的发展受到国家、政府和科技界的高度重视，在产业化和工业化的发展道路上发展迅速。

在国家政策引导和云计算商业模式日趋完善形式下，云计算的人才缺口和市场需求形成了巨大反差。如何将云计算的相关知识介绍给读者，是我们这些科学技术工作者的责任和使命。笔者依托实验室完成的云计算的具体项目，对云计算的基本资料进行了归纳与总结，对涉及的基本问题进行了较为详细的介绍。

全书共15章，首先总体介绍了Hadoop的发展历程和整体框架，然后分步介绍Hadoop的各个子系统功能和构成机制以及与开发相关的知识，阐述了Hadoop中每个部分的实现机制与用法，包括HDFS、Hadoop FS shell、Map/Reduce、Hadoop流与管道机制、Hadoop I/O、Hadoop命令、部署Hadoop等内容。部分章节配有开发实例。

读者在阅读本书内容后，能够对云计算有一个更清晰的认识，能够较为全面地掌握云计算平台层中的海量数据处理工具Hadoop的实现机制和用法，并且可以了解到Hadoop应用领域以及如何开发基于Map/Reduce的并行应用程序。无论是对于云计算研究者、程序员还是对于系统管理员，本书都具有重要参考价值。

云计算作为一种前沿IT技术正处在迅猛发展阶段，限于作者的水平和对云计算认识的局限性，本书一定存在许多不足之处，恳请读者批评指正。

编　者

目　录

第1章　云计算背景与Hadoop

1.1　云计算起源与发展历程

在过去的数十年间，计算机与网络技术得到了飞速发展，极大地推动了社会的发展。计算模式也历经大型机时代的终端/主机模式(T/S)，PC时代的客户机/服务器模式(C/S)以及目前互联网时代的浏览器/服务器模式(B/S)，直到今天所谓的云计算模式。云计算似乎是“忽如一夜春风来，千国万邦云花开”。而其实早在1983年Sun Microsystems就提出了“The Network is the computer”的观点，可以认为是云计算最初的雏形。云计算最早为谷歌、亚马逊等扩建基础设施的大型互联网服务提供商所采用的一种计算服务架构。在2006年3月，亚马逊推出弹性计算云(Elastic Compute Cloud，EC2)服务。2006年8月9日，谷歌首席执行官埃里克·施密特(Eric Schmidt)在搜索引擎大会(SES San Jose 2006)首次提出“云计算”(Cloud Computing)的概念。谷歌“云计算”的概念起源于谷歌工程师克里斯托弗·比希利亚所做的“谷歌101”项目。其核心思想是一种新的分布式计算架构，具有大规模扩展、水平分布的系统特性，所拥有的资源抽象为虚拟IT服务，并可进行持续配置，同时作为公用的资源进行管理。George Gilder 2006年10月在*Wired*杂志上发表的文章(标题为“信息工厂”(The Information Factories))中对这种架构模式进行了详细介绍。Gilder所描写的服务器庄园在架构上与网格计算(Grid Computing)相似。网格计算用于松散结合的技术计算应用程序，而新的云模式则应用于互联网服务。云和网格都被设计为可非常高效地进行水平扩展。二者都能经受得起个别元素或节点的失败。二者都按使用情况收费。然而网格通常处理批作业，并且有明确的起点和终点，而云服务却可以持续运行。此外，云扩大了可用资源的类型(文件存储、数据库和Web服务)，并且将适应范围延伸至Web和企业应用程序。

对最终用户而言，云计算则意味着没有硬件购置成本，没有需要管理的软件许可证或升级，不需要雇佣新的员工或咨询人员，不需要租赁设施，没有任何种类的基建投资，并且没有隐性成本。只是一种用仪表测量出来的、根据使用情况支付的订购服务，按使用量付费。因

此，在许多情况下，各种各样的机构和个人都可以购买“计算”，而且那些已经在建超级分布式数据中心的公司则作为一种服务提供商来提供这种基础设施和服务。

云计算的重要的发展历程大致如下：

- 2007 年年末，Cloud Computing 单词在英文中出现，2008 年年初，Cloud Computing 在中文中被翻译为“云计算”。这个单词背后代表的是一种全新的商业模式趋势，是一个继 PC 时代、网络时代以后的 IT 新时代的代名词。
- 2008 年 1 月 30 日，谷歌宣布在中国台湾地区启动“云计算学术计划”，将与中国台湾“国立台大”、“国立交通大学”等学校合作，将这种先进的大规模、快速计算技术推广到校园。
- 2008 年 7 月 29 日，雅虎、惠普和英特尔宣布一项涵盖美国、德国和新加坡的联合研究计划，推出云计算研究测试床，推进云计算。该计划要与合作伙伴创建 6 个数据中心作为研究试验平台，每个数据中心配置 1 400～4 000 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宾分校、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”(Cloud Computing)商标，此举是为加强对这一未来可能重塑技术架构的术语的控制权。戴尔在申请文件中称，云计算是“在数据中心和巨型规模的计算环境中，为他人提供计算机硬件定制制造”。
- 2010 年 3 月 5 日，Novell 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”(Trusted Cloud Initiative)。
- 2010 年 7 月，美国国家航空航天局和包括 AMD、英特尔、戴尔等支持厂商共同宣布 OpenStack 计划，而微软在 2010 年 10 月表示支持 OpenStack 与 Windows Server 2008 R2 的集成。

亚马逊由于向用户提供 AWS 服务(Amazon Web Service)被视为云计算的先驱者，随着 2008 年谷歌的 App Engine 和微软的 Azure 的推出，谷歌和微软被视为云计算市场的新参与者，IBM 则在 2009 年年初推出了云计算方案，被视为企业云计算市场的开拓者。

1.2 云计算定义与体系

由于云计算性能的扩张性，现在还无法给云计算一个统一的定义，维基百科中的定义为：云计算将 IT 相关的能力以服务的方式提供给用户，允许用户在不了解提供服务的技术、没有相关知识以及设备操作能力的情况下，通过 Internet 获取所需要的服务。Vaquerolm等人在 *A break in the clouds: Towards a cloud definition* 给出了一个定义，认为云是一个包含大量可用虚拟资源(例如硬件、开发平台以及 I/O 服务)的资源池。这些虚拟资源可以根据不同的负载动态地重新配置，以达到更优化的资源利用率。这种资源池通常由基础设施提供商按照服务等级协议(Service Level Agreement，SLA)采用按时付费(Pay-Per-Use，PPU)的模式开发管理。

关于云计算的体系，Lamia Youseff 等人在 *Toward a Unified Ontology of Cloud*

Computing 文献中提出了一种五层体系结构，如图 1-1 所示。

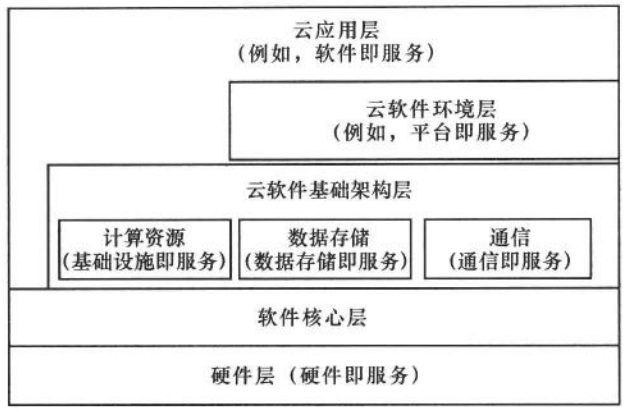

图 1-1　云计算五层体系结构

云计算体系可以分成 5 个层次：云应用层、云软件环境层、云软件基础构架层、软件核心层、硬件层。最上层是云应用层，是云提供给用户服务的接口，这种模型简化了云计算服务提供商的应用部署，一个云应用是部署在服务提供商的计算中心而不是用户的机器之中，可以称这种模型为软件即服务（SaaS），很有名的两个 SaaS 的例子就是 Salesforce 的 CRM（Customer Relationships Management system）和谷歌 Apps；第二个层次就是云软件环境层，也可以称之为软件平台层，这个层的用户是云应用的研发者，软件平台提供商提供研发者一种研发环境以及很好的 API，提供这个层次的服务可以称之为平台即服务（PaaS），在这个层中最有名的就是 Google's App Engine，另一个范例就是 Salesforce Apex Language。

第三个层次就是云软件基础设施层，这个层提供基础资源给上一层。这一层提供的云服务有 3 种：计算资源、数据存储和通信。虚拟机是最常见的计算资源提供形式，可以称这种服务为基础设施即服务（IaaS），因此虚拟化技术是云计算的关键理论技术之一，目前开源的虚拟化技术以 XEN 和 KVM 为代表，同时 VMWare 虚拟化在企业虚拟化应用中也很广泛，在基础设施即服务中最出名的就是亚马逊弹性云计算平台（Amazon's Elastic Compute Cloud，EC2）和 Enomalism 弹性云计算平台（Enomalism Elastic Computing Infrastructure）。数据储存是第二个基础设施资源，提供存储服务，也就是很有名的数据存储即服务（DaaS），例如谷歌的分布式文件系统 GFS。最有名的 DaaS 提供商就是亚马逊的 S3 和 EMC 存储管理服务。

在云计算中用户对于服务质量（QoS）有很高的要求，而云服务是实时通过网络通信提供的，因此通信对于云平台至关重要，所以提出了通信即服务（CaaS）的概念。很多研究者研究了提供 QoS 通信服务的构架设计、协议以及解决方案。比较有名的 CaaS 范例就是微软的连接服务框架（CSF），还有就是基于 CaaS 的可能应用，包括 VoIP、Audio & Videoconferencing、IM 等。

第四个层次是软件核心层，这一层提供组成云实际物理服务的基本软件的管理。这一层的实现可以是操作系统核心、系统管理程序（Hypervisor）、虚拟机监视器或集群中间件。云计算的最底层便是硬件层，这一层的用户通常是有巨大 IT 需求的大企业，可以称提供的这种云服务为硬件即服务（HaaS）。最大的 Haas 范例便是 Morgan Stanley 和 IBM 建立的

Sublease Contract，本层相关技术有 Remote Scriptable Boot-loaders（PXE，Uboot，IBM Kitty Hawk）等。

1.3 云计算关键技术

云计算的主要特点是数据密集型的计算方式，同时还具有移动计算的特点，即移动计算到数据，而不是移动数据到计算，因为将 CPU 计算移动到数据的代价更小。因此，一般来说，云计算的关键支撑技术包括虚拟化技术、并行计算、分布式存储、分布式数据管理等，下面进行简要介绍。

1.3.1 虚拟化技术

抽象地说，虚拟化技术是通过新增的虚拟中间层截获上层软件对底层接口的调用，并对该调用重新作出解释和处理，以实现异构环境中资源的可共享、可管理和可协同，并支持应用大规模部署、迁移和运行维护。通过虚拟机，可以在原有的硬件资源和操作系统上仿真一台虚拟计算机，使软件可以不经修改直接运行在虚拟机中，这样就可以以虚拟机的形式出租 CPU 的计算能力，如亚马逊的弹性云计算。

在虚拟化技术中最重要的技术就是 VMM 技术，即虚拟机监控器，依靠 VMM 可以实现虚拟计算机的所有功能。按照虚拟机的宿主环境可将虚拟机划分为两种：一种就是依赖于宿主操作系统的，虚拟机作为操作系统的应用程序存在，如 VMWare 和 Virtual PC；另一种就是直接依赖于宿主硬件平台本身，例如 IBM 的 VM/360 系统、剑桥大学的 Xen 以及 VMWare 面向企业用户推出的 VMWare ESX Server 等，这类虚拟机可以直接访问硬件资源，所以性能与没有虚拟化的机器相当。

通过虚拟化技术可以实现硬件物理资源的逻辑抽象，将云中计算机物理资源形成一个统一的可以调度分配的云资源池，通过虚拟化技术可以提高整个云中机器或集群的资源利用率，能根据用户的计算需求快速动态地分配系统资源。

1.3.2 分布式计算和并行计算

并行处理是高性能计算机的一个前沿研究领域，是云计算的关键理论技术支撑。云计算的一个重要的特点就是计算的并行和分布式。用户的应用在云中是以分布式的并行计算模式进行的，这样可以在很大程度上提高运行效率的同时也利于计算资源的负载均衡。分布式计算将大的问题分解为许多小的子问题，然后将这些小的子问题分配给多台计算机（通常为一个集群）进行处理，最终将计算得到的结果进行收集综合而得到大问题的结果。并行计算和分布式计算思想基本类似，分布式计算强调的是空间，而并行计算强调的是时间的同步，很多实际的问题都是可以分解为并行计算的多个子任务。

目前并行模式可粗略地归纳为 3 类，即共享存储模式、分布存储模式、共享存储与分布存储混合模式。相应的程序设计也可以归为 3 类，即共享程序设计（如 OpenMP）、基于消息传递程序设计（如 PVM 和 MPI）和混合编程模式（如 MPI＋OpenMP）。其中 MPI 模式在

学术研究领域应用较多，而在商业领域，云计算大多采用的是谷歌云计算系统中的 Map/Reduce 并行编程模型，云计算强调的就是简单的编程模型，而 Map/Reduce 就是一种高效的、简单的并行编程模式，也是一种高效的任务调度器，Map/Reduce 这种编程模型并不仅适用于云计算，在多核和多处理器、Cell Processor 以及异构机群上同样有良好的性能。利用 Map/Reduce，程序员得能够轻松地编写紧耦合的程序，运行时能高效地调度和执行任务，在实现时，在 Map 函数中指定对各分块数据的处理过程，在 Reduce 函数中指定如何对分块数据处理的中间结果进行归约。用户只需要指定 Map 和 Reduce 函数来编写分布式的并行程序，不需要关心如何将输入的数据分块、分配和调度，同时系统还将处理集群内节点失败以及节点间通信的管理等。

Hadoop 开源云计算平台中就实现了这种 Map/Reduce 的计算模型，因此利用 Hadoop 可以实现并行计算。

1.3.3　分布式存储

分布式存储在云计算中也叫云存储，云计算平台必须保证高可用、高可靠和经济性。面对海量数据采用分布式存储的方式来存储数据，采用冗余存储的方式来保证存储数据的可靠性，即为同一份数据存储多个副本。同时，云计算需要同时满足大量用户的访问需求，并行地提供服务。因此，云计算的数据存储技术必须具有高吞吐率和高传输率的特点。

目前的分布式存储系统有网络存储 NAS、FAS、Prospero 与 P2P 的海量存储系统。NAS 是一种特殊的集成了操作系统和存储设备的专用数据存储服务器；NAS 是让客户机和服务器的任意集合共享一个公用的文件系统，严格来说不算是分布式文件系统，但却是分布式文件系统的鼻祖；FAS 通过一个统一的名字空间（目录树）将所有文件服务器组织在一个两层结构中。用户只要和 FAS 中任何一台文件服务器相连，就可以接受整体服务；Prospero 利用一个具有多名字空间、文件名过滤等功能的名字服务，将众多的文件服务器组织成一个逻辑整体的分布式文件系统；基于 P2P 的海量存储系统，最具影响力的是谷歌的 GFS(Google File System)。GFS 是一个管理大型分布式数据密集型计算的可扩展的分布式文件系统。它使用廉价的商用硬件搭建系统并向大量用户提供容错的高性能的服务。GFS 系统由一个 Master 和大量块服务器构成。Master 存放文件系统的所有元数据，包括名字空间、访问控制、文件分块信息、文件块的位置信息等。GFS 中的文件切分默认为 64 MB的块进行存储。在 GFS 文件系统中，采用冗余存储的方式来保证数据的可靠性。每份数据在系统中保存 3 个以上的备份。为了保证数据的一致性，对于数据的所有修改需要在所有的备份上进行，并用版本号的方式来确保所有备份处于一致的状态。

Hadoop 中的 HDFS(Hadoop Distributed File System)是谷歌的 GFS 的 Java 开源实现。大部分 IT 厂商，包括雅虎、英特尔的云计划采用的都是 HDFS 的数据存储技术。

1.3.4　分布式海量数据管理

云计算的显著特点是对海量的数据存储、读取后可进行大数据量的分析，数据的读操作频率远大于数据的更新频率。虽然分布式存储可以提供用户的并行访问，但是直接建在分布式文件系统上的存储往往具有延迟性，很难保证用户的低延迟需要，因此就需要对云中的海量数据进行优化管理。

目前比较有名的行优化管理技术为谷歌云计算系统中使用的 BigTable 技术，BigTable 是一个基于分布式文件系统的非关系数据库，采用基于列的方式来管理数据，谷歌对 BigTable 的定义为：BigTable 是一种为了管理结构化数据而设计的分布式存储系统，系统中数据可以扩展到非常大的规模，例如在数千台商用服务器上可达到 PB(Peta Bytes)规模，因此利用 BigTable 就可以高效地管理云中的海量数据。

关于 BigTable 的进一步了解可以参考谷歌的相关文献。

1.4 Hadoop 与云计算

从上述的云计算关键技术中可以看出分布式并行计算、分布式存储以及分布式数据管理是实现云技术的关键技术，而 Hadoop 就是一个实现了谷歌云计算系统的开源系统，包括并行计算模型 Map/Reduce、分布式文件系统 HDFS 以及分布式数据库 HBase，同时 Hadoop 的相关项目也很丰富，包括 ZooKeeper、Pig、Chukwa、Hive、HBase、Mahout 等，这些项目都使得 Hadoop 成为一个很大的家族。目前使用 Hadoop 技术实现云计算平台的有 IBM 的蓝云，雅虎、英特尔的云计划，还有中国移动的 BigCloud，百度云计算以及阿里巴巴云计算平台。

如果从云计算体系构架上看 Hadoop 的云计算的环境，Hadoop 技术属于云体系中的一种 PaaS 技术，Hadoop 可以给用户提供一种分布式计算和分布式存储的编程环境。

1.5 谁在使用 Hadoop

Hadoop 是一个开源的高效的云计算实现平台，其不仅在云计算领域用途广泛，同时在海量数据处理、数据挖掘、机器学习、科学计算等领域也越来越受到青睐，以下将列出著名企业使用 Hadoop 的情况，这些数据大多数统计的时间是 2008 年，目前 Hadoop 的使用远远超出这些。

1.5.1 外国 Hadoop 应用

1. 雅虎

雅虎是 Hadoop 的最大支持者，大约有两万台计算机，超过 10 万个 CPU 在运行 Hadoop。最大的一个机群有 2 000 个节点(每个节点 2×4 CPU boxes w,4 TB 磁盘)用于支持广告系统和 Web 搜索的研究，用于可扩展性测试，以便支持更大机群上的 Hadoop 开发。

2. Facebook

Facebook 使用 Hadoop 存储内部日志与多维数据，并以此作为报告、分析和机器学习