云环境下
大数据分析平台
关键技术研究

戴　伟◎著

中国水利水电出版社
www.waterpub.com.cn

云环境下
大数据分析平台
关键技术研究

戴　伟◎著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

在如今的社会，大数据的应用越来越彰显它的优势，它的应用范围也越来越广，如电子商务、O2O、物流配送等，在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业也有涉及。

本书以云计算与大数据基础开篇，简单介绍了分布式文件系统 HDFS 与 NoSQL 数据库技术，重点对分布式计算框架 MapReduce、Hadoop 技术、云数据中心、大数据与数据挖掘技术进行了阐述。

本书叙述语言简洁、逻辑清楚、内容详尽，是一本值得学习研究的著作。

图书在版编目(CIP)数据

云环境下大数据分析平台关键技术研究/戴伟著.
--北京：中国水利水电出版社，2017.6
ISBN 978-7-5170-5537-2

Ⅰ.①云… Ⅱ.①戴… Ⅲ.①数据处理－研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2017)第 148924 号

书　　名	云环境下大数据分析平台关键技术研究 YUNHUANJING XIA DASHUJU FENXI PINGTAI GUANJIAN JISHU YANJIU
作　　者	戴　伟　著
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：sales@waterpub.com.cn 电话：(010)68367658(营销中心)
经　　售	北京科水图书销售中心(零售) 电话：(010)88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京亚吉飞数码科技有限公司
印　　刷	三河市佳星印装有限公司
规　　格	170mm×240mm　16 开本　14.5 印张　260 千字
版　　次	2017 年 10 月第 1 版　2017 年 10 月第 1 次印刷
印　　数	0001－2000 册
定　　价	43.50 元

凡购买我社图书，如有缺页、倒页、脱页的，本社营销中心负责调换

前　言

计算机的发展，特别是网络技术的发展催生了云计算技术的出现，云计算被认为是信息技术的一次重大变革。云计算、物联网、社交网络的发展使人类社会的数据产生方式发生了变化，社会数据的规模正在以前所未有的速度增长，出现了大量的非结构化和半结构化数据，单位也由 TB 级别跨越到了 PB、EB 级别，大量信息源产生的这些数据已远远超越目前人力所能处理的范围，人们在思索如何对这些数据进行管理及使用时，逐渐探索出一个新的领域——大数据技术。

大数据的“大”不仅指其容量，还体现在多样性、处理速度和复杂度等方面。无论人们是否关注过，海量的数据已如决堤之洪流涌入人们的生活，大数据的时代已然到来了。可以目睹的是，大数据的激流已经给个人生活、企业经营乃至国家和社会的全面发展带来了新的机遇与挑战。在如今的社会，大数据的应用越来越彰显它的优势，它占领的领域也越来越大，如电子商务、O2O、物流配送等，在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在也早已有些时日了。各种利用大数据进行发展的领域正在协助企业不断地发展新业务和创新运营模式。谷歌、Amazon、Facebook 等全球知名互联网企业作为大数据领域的先驱者，凭借自身力量进行大数据探索，甚至在必要时创造出相关工具。这些工具目前已经被视为大数据技术的基础。

随着大数据技术和市场的快速发展，驾驭大数据的呼声渐涨，蕴含在大数据中的价值使得大数据已经成为 IT 信息产业中最具潜力的蓝海，这也使得学习及掌握国际前沿的大数据处理工具和解决方案中的核心技术显得十分迫切。从全球角度来看，对大数据的认识、研究和应用还都处于初期阶段，特别是对我国来说，大数据真正落地还需要一个长期的过程。而且大数据技术有别于传统数据处理工具和技术，掌握难度较大，不仅需要 1～2 年的反复尝试，而且在实际使用中解决了大量问题之后才能正确理解它。

本书共分为 7 章，内容涵盖了云计算与大数据的基本概念，大数据的关键技术和应用。第 1 章主要对云计算与大数据基础进行阐述，内容包括云计算的概述、关键技术简介，大数据时代的机遇与挑战、大数据的技术体系、

大数据与云计算之间的关系。在大数据时代，海量数据的增长促使人们对数据的组织和存储进行管理，由此出现海量数据存储技术——分布式文件系统 HDFS，第 2 章对此技术进行了相应的知识研究。由于传统关系型数据库存在着灵活性差、扩展性差与性能差等原因，人们开始寻求能够满足扩展性方面需求的数据库，将那些存储系统转向采用不同的解决方案、没有固定数据模式的系统称为 NoSQL，第 3 章对此内容进行重点阐述。全球每时每刻都有大量的数据生成，想要对如此多的数据进行分析处理，传统工具已明显力不从心了，为此出现了分布式计算框架 MapReduce 和 Hadoop 技术，Hadoop 是将 MapReduce 通过开源方式进行实现的框架的名称，是大数据最重要的技术，本书第 4、5 章对此部分内容进行重点阐述。第 6 章为云数据中心，云计算应用的核心技术是数据处理技术，大数据为提升云计算的应用价值提供了新的重要的技术与手段，同时，云计算为大数据提供弹性可扩展的基础设施支撑环境以及数据服务的高效模式，为此，云计算与大数据的高度融合及其深度应用已经势在必行。第 7 章为大数据与数据挖掘技术，大数据时代的重点是数据深度有效利用，也就是数据挖掘，它是各行各业都迫切需要的一项新技术和事业发展的新领域。了解并掌握数据挖掘技术并发挥它的价值，对我国大数据技术水平的增长有着重要意义。

云计算与大数据技术复杂、涉及面广，本书在撰写过程中参考并引用了大量前辈学者的研究成果和论述，在此作者向这些学者表达诚挚的敬意和谢意。云计算与大数据技术处在高速发展的技术领域之中，新技术、新方法、新架构层出不穷，加之经验水平有限，书中疏漏之处在所难免，恳请各界专家、学者、读者批评指正。

作者

2017 年 3 月

目 录

前言

第 1 章 云计算与大数据基础 …… 1

1.1 云计算概述 …… 1
1.2 云计算关键技术简介 …… 13
1.3 大数据时代的机遇与挑战 …… 20
1.4 大数据的技术体系 …… 21
1.5 大数据与云计算之间的关系 …… 25
小结 …… 27

第 2 章 分布式文件系统 HDFS …… 28

2.1 HDFS 概述 …… 28
2.2 HDFS 的体系结构 …… 33
2.3 HDFS 存取机制 …… 41
2.4 HDFS 常用命令 …… 46
2.5 HDFS 存储海量数据 …… 59
小结 …… 60

第 3 章 NoSQL 数据库技术 …… 61

3.1 NoSQL 及其与关系型数据库的比较 …… 61
3.2 列式存储和文档存储 …… 67
3.3 key-value 数据库 …… 86
3.4 图形数据库 …… 87
3.5 NewSQL 数据库 …… 88
3.6 基于 NoSQL 的 Megastore 存储系统 …… 93
小结 …… 95

第 4 章 分布式计算框架 MapReduce …… 97

4.1 MapReduce 的引入 …… 97
4.2 MapReduce 编程模型 …… 101
4.3 MapReduce 核心技术分析 …… 115
4.4 MapReduce 的应用实践 …… 117
小结 …… 120

第 5 章 Hadoop 技术 …… 121

5.1 集群上的 MapReduce 实现——Hadoop …… 121
5.2 对 Hadoop 技术的深入了解 …… 127
5.3 后 Hadoop 时代即将来临 …… 142
小结 …… 147

第 6 章 云数据中心 …… 148

6.1 云数据中心概述 …… 148
6.2 网络融合技术 …… 153
6.3 云数据中心节能技术 …… 154
6.4 虚拟化技术 …… 156
6.5 安全技术 …… 165
6.6 云数据中心的规划与建设 …… 179
6.7 大数据分析 …… 182
小结 …… 183

第 7 章 大数据与数据挖掘技术 …… 184

7.1 大数据与数据挖掘的关系 …… 184
7.2 数据挖掘的核心思想和主要功能 …… 186
7.3 数据挖掘的内容与主要方法 …… 189
7.4 复杂数据类型挖掘 …… 198
小结 …… 220

参考文献 …… 221

第1章　云计算与大数据基础

过去几年里，云计算已成为新兴技术产业中最热门的领域之一，也是继个人计算机、互联网变革后的第三次信息技术浪潮，它将使人类生活、生产方式和商业模式等产生根本性的变革。云计算技术的发展使得人们汇聚、存储和处理数据的能力超过以往，从数据中提取价值的能力也在显著提高。云计算的蓬勃发展开启了大数据时代的大门。随着互联网、移动互联网、物联网、数码设备等的快速发展，更多的智能终端、传感设备等接入到网络，由此产生的数据及增长速度将超过历史上的任何时期，社会化信息正步入大数据(Big Data)时代，“大数据”的概念逐渐成为发展的趋势，这种趋势为理解这个世界和作出决策开启了一扇大门。

1.1　云计算概述

1.1.1　什么是云计算

也许大家都知道瞎子摸象的故事。话说一位商人牵来一头大象，几个瞎子饶有兴趣地对大象抚摸起来。摸到象腿的说，大象像大木桩；摸到耳朵的说，大象像大葵扇；摸到象牙的说，大象像大萝卜；摸到尾巴的说，大象像绳子……云计算诞生初期，人们对它的认识，真有点像瞎子摸象，各有各的说法。

有人说，虚拟化就是云计算；有人说，分布式计算就是云计算；也有人说，把一切资源都放在网上，一切服务都从网上取得就是云计算；更有人说，云计算是一个简单的甚至没有关键技术的东西，它只是一种思维方式的转变，等等。

先来看看为什么用“云”来命名这个新的计算模式，以及云计算中的“云”是什么。

一种比较流行的说法是当工程师画网络拓扑图时，通常是用一朵云来抽象表示不需表述细节的局域网或互联网，而云计算的基础正是互联网，所

以就用了“云计算”这个词来命名这个新技术。另外一个原因就是，云计算的始祖——亚马逊将它的第一个云计算服务命名为“弹性计算云”。

其实，云计算中的“云”不仅是互联网这么简单，它还包括了服务器、存储设备等硬件资源和应用软件、集成开发环境、操作系统等软件资源。这些资源数量巨大，可以通过互联网为用户所用。云计算负责管理这些资源，并以很方便的方式提供给用户。用户无须了解资源具体的细节，只需要连接上互联网，就可以使用了。例如，人们使用网络硬盘，只需连接上服务提供商的网站，就可以使用了，不需要知道存放文件的机器型号、存放位置、容量等。存储空间不够？再申请就可以了。

1.1.2 云计算的特征

云计算如今被热炒，很多商家不管是与不是，都把自己的产品贴上云标签，使得云产品满天飞，甚至以假乱真！那么，什么样的产品及其应用才算是云计算呢？答案是具备云计算特征。

云计算的主要特征有：

其一，以网络为依托，通过网络提供服务。云计算所依托的网络主要是互联网，根据需要，也可以是广域网、局域网、企业网及专用网等。

其二，以虚拟技术为基础，用虚拟技术整合软硬件资源和计算能力。

其三，服务透明化。用户使用服务时，无须知道资源的结构、实现方式和所在的位置。

其四，按需自动服务。用户通过云计算可自动获得满足用户需求的计算资源、计算机能力和相关服务。

上述 4 条是云计算的主要特征，也是云计算的核心。此外，诸如高可靠性、高扩展性、低应用成本等，是对云计算的要求，或云计算应该达到的目标，而非云计算的核心特征。

1.1.3 云计算的结构和层次

1. 云计算参考架构

云计算参考架构中包含了五类重要的用户角色：云用户、云提供商、云载体、云审计和云代理，其中每个角色都是一个实体，既可以是个人也可以是机构，参与云计算的事务处理或任务执行。不同的用户在云计算中扮演不同的角色，它们是云计算的主体和推动力量。

云计算参考架构中的五类角色之间的相互关系如图 1-1 所示。

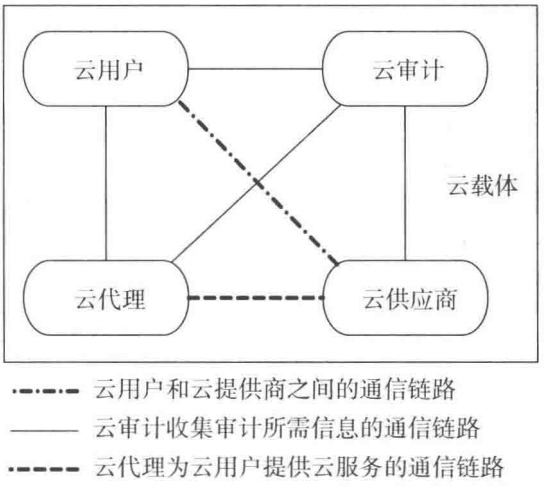

图 1-1　云计算中各类角色之间的交互关系及通信链路

(1)云用户

云用户为云服务的使用者,它们与云提供商保持业务联系,使用云提供商提供的各种云服务,可以是个人也可以是机构,如政府、教育机构或企业客户等,它们租用而不是购买云服务提供商提供的各种服务,并为之付费。

云用户是云服务的最终消费者,也是云服务的主要受益者。云服务为云用户提供的服务主要包括:浏览云提供商的服务目录;请求适当的服务;云提供商建立服务合同;使用服务。

在云计算中,云用户和云服务提供商按照约定的服务等级协议进行通信。这里,服务等级协议(Service Level Agreement,SLA)指在一定开销下为保障服务的性能和可靠性,服务提供商与用户间定义的一种双方认可的协议。云用户使用 SLA 来描述自己所需的云服务的各种技术性能需求,如服务质量、安全、性能失效的补救措施等,云提供商使用 SLA 来提出一些云用户必须遵守的限制或义务等。

云用户可以根据价格及提供的服务自由地选择云提供商。服务需求不同,云用户的活动和使用场景就不同。

由于云计算环境提供三大类服务,即软件即服务(Software as a Service,SaaS)、平台即服务(Platform as a Service,PaaS)和基础设施即服务(Instruction as a Service,IaaS)。相应地,根据用户使用的服务类型,可以将云用户分为三类,即 SaaS 用户、PaaS 用户和 IaaS 用户。

1)软件即服务 SaaS

SaaS 用户通过网络使用云提供商提供的 SaaS 应用,它们可以是直接

使用软件的终端用户，可以是向其内部成员提供软件应用访问的机构，也可以是软件的管理者，为终端用户配置应用。SaaS 提供商按一定的标准进行计费，且计费方式多样，如可以按照终端用户的个数计费，可以按用户使用软件的时间计费，可以按用户实际消耗的网络带宽计费，也可以按用户存储的数据量或者存储数据的时间计费。

图 1-2 所示为 SaaS 基于构件库的架构设计。图 1-3 所示为 SaaS 平台逻辑架构。

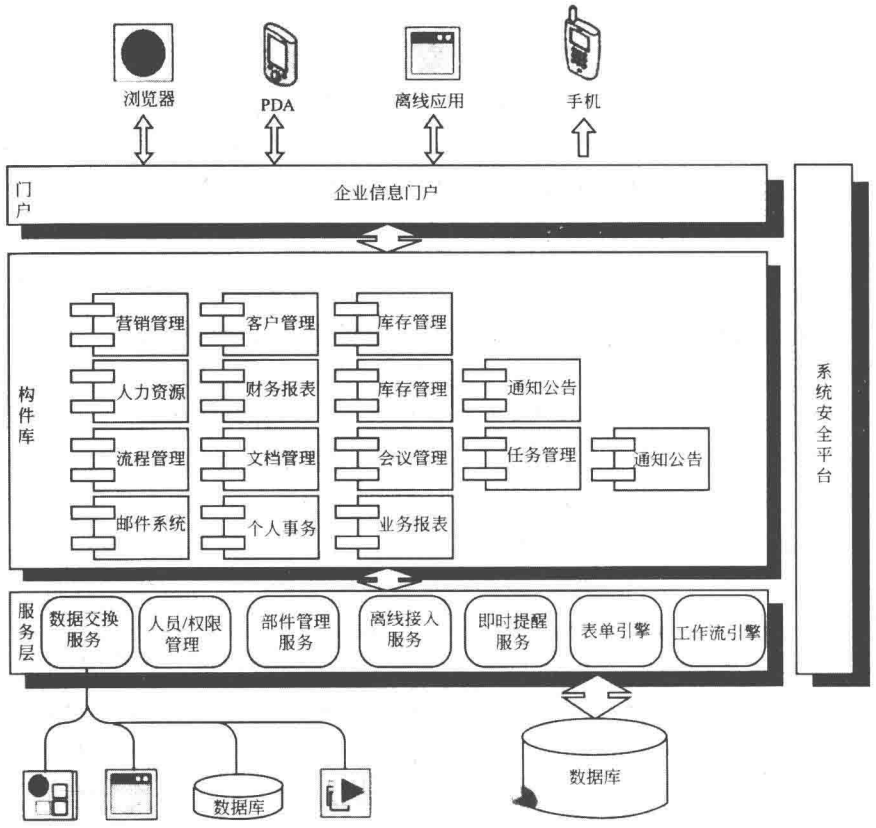

图 1-2 基于构件库的架构设计

2）平台即服务 PaaS

PaaS 用户可以使用云服务提供商提供的工具和可执行资源部署、测试、开发和管理托管在云环境中的应用。PaaS 用户可以是设计和开发各种软件的应用开发者，可以是运行和测试基于云环境的应用测试者，可以是在云环境中发布应用的部署者，也可以是在云平台中配置、监控应用性能的管理者。PaaS 提供商按照不同的形式进行计费，如根据 PaaS 应用的计算量、

数据存储所占用的空间、网络资源消耗大小及平台的使用时间来计费等。

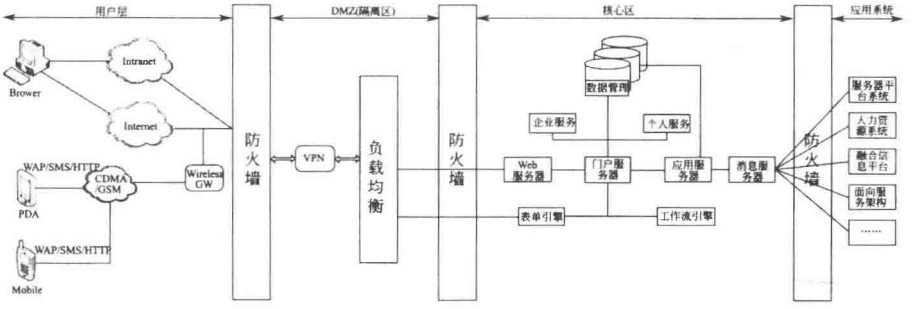

图 1-3　SaaS 平台逻辑架构

3)基础设施即服务 IaaS

IaaS 用户可以直接访问虚拟计算机，通过网络访问存储资源、网络基础设施及其他底层计算资源，并在这些资源上部署和运行任意软件。IaaS 用户可以是系统开发者，系统管理员，以及负责创建、安装、管理和监控 IT 基础设施运营的 IT 管理人员。IaaS 用户具有访问这些计算资源的能力，IaaS 提供商根据其使用的各种计算资源的数量及时间来进行计费，如虚拟计算机的 CPU 小时数、存储空间的大小、消耗的网络带宽、使用的 IP 地址个数等。

(2)云提供商

云服务的提供者，负责提供其他机构或个人感兴趣的服务，可以是个人、机构或者其他实体。云提供商获取和管理提供云服务需要的各种基础设施，运行提供云服务需要的云软件，并为云用户交付云服务。云提供者的主要活动包括以下 5 个方面：服务的部署、服务的组织、云服务的管理、安全和隐私，如图 1-4 所示。

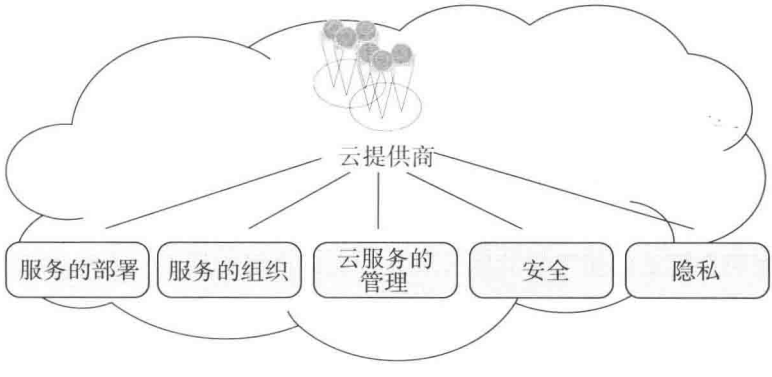

图 1-4　云提供者的主要活动

1)SaaS 环境云提供商

在云基础设施上部署、配置、维护和更新各种软件应用，确保能按照约定的服务级别为云用户提供云服务。SaaS 提供商承担维护、控制应用和基础设施的大部分责任，SaaS 用户不需要安装任何软件，它们对软件拥有有限的管理控制权限。

2)PaaS 环境云服务提供商

负责管理平台的基础设施，运行平台的云软件，如运行软件执行堆栈、数据库及其他的中间件组件等。PaaS 提供商通常也为 PaaS 用户提供集成开发环境(IDE)，软件开发工具包(SDK)，管理工具的开发、部署和管理等。PaaS 用户具有控制应用程序的权限，也可能具有对托管环境进行各种设置的权限，但无权或者受限访问平台之下的底层基础设施，如网络、服务、操作系统和存储等。

3)IaaS 环境提供商

IaaS 提供商需要位于服务之下的各种物理计算资源，包括服务器、网络、存储和托管基础设施等。IaaS 提供商通过运行云软件使 IaaS 用户能通过服务接口、计算资源抽象如虚拟机、虚拟网络接口等访问 IaaS 服务。反过来，IaaS 用户使用这些计算资源如虚拟计算机来满足自己的基础计算需求。和 SaaS、PaaS 用户相比，IaaS 用户能够从更底层上访问更多的计算资源，因此对应用堆栈中的软件组件具有更多的控制权，包括操作系统和网络。另一方面，IaaS 提供商具有对物理硬件和云软件的控制权，使其能配置这些基础服务，如物理服务器、网络设备、存储设备、主机操作系统和虚拟机管理程序等。

云服务的要求不同，云用户的活动和使用场景就不同。例如，云用户直接向云提供商发送服务请求，云服务提供商接收到云用户请求后，进行相应的处理，并将云服务直接交付给云用户，不经过任何中间机构或个人。云载体负责将云服务从云提供商传输给云用户。这里的云提供商需要两种不同的服务等级协议(图 1-5)：①用于和云用户之间的通信(使用 SLA1)；②用于和云载体之间的通信(使用 SLA2)。云提供商为了保证能够按照 SLA1 为云用户提供高质量的服务，通常需要和云载体建立一定的服务约定，因此它们采用 SLA2 来向云载体提出其在能力、灵活性、功能方面的要求，如云提供商利用 SLA2 要求云载体为其提供专用的、加密的连接以保证云用户能够按照合同正确使用和消费云服务。云载体将按照 SLA2 来为云提供商提供高质量的通信服务。

(3)云载体

云载体作为中介机构负责提供云用户和云提供商之间云服务的连接和

传输，负责将云提供商的云服务连接和传输到云用户。云载体为云用户提供通过网络、电信和其他设备访问云服务的能力，如云用户可以通过网络设备如计算机、笔记本、移动电话、移动网络设备等访问云服务。

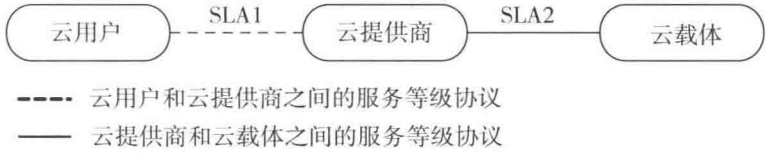

图 1-5　云提供商需要的两种不同服务等级协议

云服务一般是通过网络、电信或者传输代理来提供的，这里传输代理指的是提供高容量硬盘等物理传输介质的商业组织。为了确保能够按照与用户协商的服务等级协议(SLA)为用户提供高质量的云服务，云提供商将和云载体建立相应的服务等级协议，如在必要的时候要求云载体为云提供商和云用户之间建立专用的、安全的连接服务。

(4)云审计

云环境中的审计是指通过审查客观证据验证服务是否符合标准。云审计者是可独立评估云服务，信息系统操作、性能和安全的机构，能够从安全控制、隐私及性能等多个方面对云服务提供商提供的云服务进行评估。

例如，云审计负责对云服务提供商提供的云服务的实现和安全进行独立的评估，因此云审计需要同时与云提供者和云消费者进行交互。如图 1-6 所示，这里的云用户是直接向云提供商请求服务，而不是通过云代理或者其他机构使用云服务，因此云审计在收集审计所需要的信息时，仅需要与云用户和云提供商进行通信，但是在存在云代理或其他中间机构时，为了准确完成审计工作，云审计可能需要收集更多的审计信息，包括从云代理或其他中介机构那里获取信息。

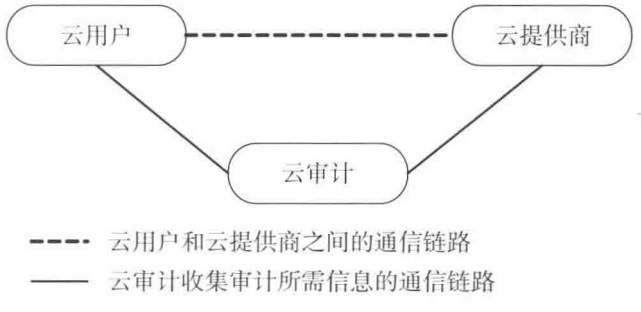

图 1-6　云审计关系流程示例

(5)云代理

云环境中的代理机构,负责管理云服务的使用、性能和分发的实体,也负责在云提供者和云用户之间进行协商。此时,云用户不再需要直接向云提供商请求服务,而可以向云代理请求服务。

例如,如图 1-7 所示,云代理获取云提供商 1 和云提供商 2 的服务,并通过提升现有的服务或者组合不同的服务来产生新的服务,提供给云用户,并进行计费。对于云用户而言,云提供商是透明的,它们直接和云代理进行交互,使用云代理提供的云服务。

云代理提供的云服务包括服务中介、集成、增值三类。

图 1-7 云用户通过云代理使用云服务示例

2. 云计算技术体系

由于云计算的服务分为 IaaS、PaaS 和 SaaS 三种类型,不同的厂家又提供了不同的解决方案,因此目前还没有一个统一的技术体系架构。综合不同厂家的方案,给出一个供商榷的云计算技术体系架构,如图 1-8 所示。该技术架构概括了不同解决方案的主要特征,每一种方案或许只实现了其中部分功能,或许还有部分相对次要的功能尚未概括进来。

如图 1-8 所示,云计算技术体系架构分为四层,由下而上分别为物理资源层、资源池层、管理中间件层和 SOA(Service-Oriented Architecture, SOA)构建层。

云计算通常提供 IaaS、PaaS 和 SaaS 三个层次的服务,不同的服务所涉及的核心技术存在较大差异,图 1-9 给出了三个层次的服务所涉及的技术和典型应用。

3. 云服务部署

云计算有三种不同的部署模式,分别为公有云、私有云和混合云。在介绍云服务部署模式之前,先对安全边界进行阐述。如图 1-10 所示,安全边界能够对访问进行限制:安全边界内部的实体能够自由地访问安全边界内的资源,而安全边界外的实体只有在边界控制设备允许的情况下才能访问安全边界内的资源。典型的边界控制设备包括防火墙、安全卫士和虚拟专

用网。通过对重要资源设置安全边界，机构既能够实现对这些资源的访问控制，又能够实现对这些资源使用情况的监控。更进一步，通过更改配置，机构可以根据需求改变设备的安全边界，如根据业务情况的变化阻止或允许不同的协议或数据格式。

图 1-8　云计算技术体系架构

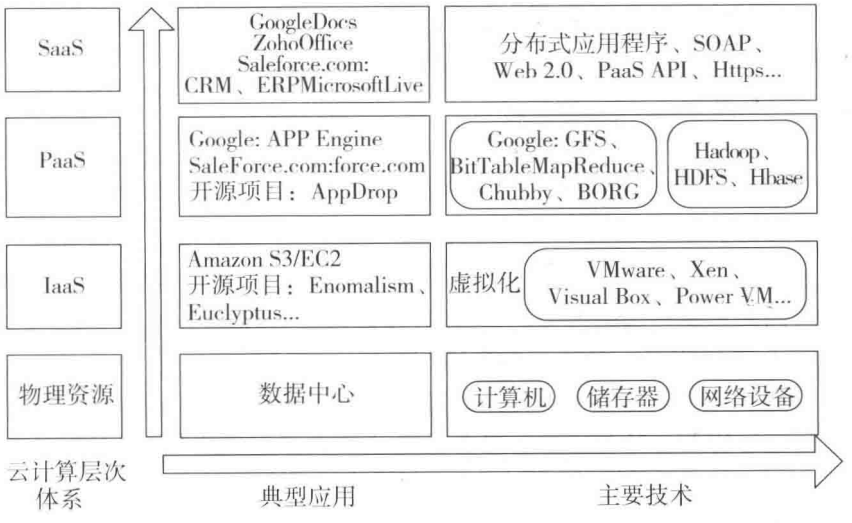

图 1-9　三个层次云计算技术及典型应用

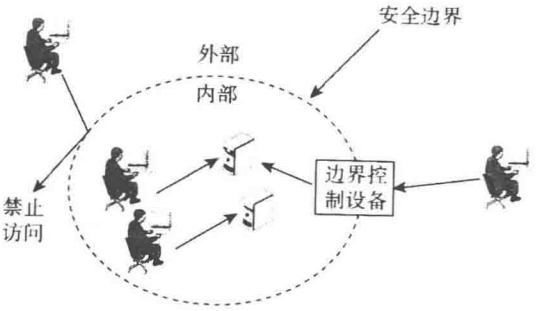

图 1-10 安全边界

不同的云部署模式具有不同的安全控制边界，因此云用户对云资源也具有不同的执行权限。

(1)公有云

在公有云中，云提供商负责公有云服务产品的安全管理及日常操作管理等，用户对云计算的物理安全、逻辑安全的掌控及监管程度较低。图 1-11 给出了一个公有云应用实例图。使用公有云服务的用户既可以是个体用户，也可以是机构用户。个体用户仅需一个能上网的终端设备，如笔记本电脑、手机或 iPad 等通过互联网即可访问云服务；机构用户通过本单位的边界控制设备访问云服务。

①边界控制设备能限制和管理内部用户对公有云的访问。

②边界控制设备也能保护内部设备免受外部攻击。

目前，典型的公有云有微软的 Windows Azure Platform、亚马逊的 AWS、Salesforce. con，以及国内的阿里巴巴、用友伟库等。

(2)私有云

私有云有下列两种部署方式。

①将私有云部署在企业数据中心的防火墙内，由云用户自己管理，称为自建私有云。

②将私有云部署在一个安全的主机托管场所，如外包给托管公司，由托管公司负责云基础设施的维护和管理，称为托管私有云。

图 1-12 给出了一个简单的自建私有云。

对图 1-12 进行分析可知，安全边界既覆盖了云用户的内部资源，也覆盖了私有云资源。私有云可以集中在单个云用户站点内部，也可以分布在多个私有云用户的站点之间。安全边界的存在使得云用户有机会对站点内的私有云资源进行控制。

图 1-13 描述了一个托管私有云。