“十二五”
国家重点图书出版规划项目

学术中国 · 院士系列

未来网络创新技术研究系列

云计算大数据处理

刘鹏 于全 杨震宇 陈伟 王磊 张乃甜 编著

Cloud Computing and Big Data Processing

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

“十二五”
国家重点图书出版规划项目

学术中国 · 院士系列

未来网络创新技术研究系列

云计算大数据处理

刘鹏 于全 杨震宇 陈伟 王磊 张乃甜 编著

Cloud Computing and
Big Data Processing

人民邮电出版社
北京

图书在版编目（CIP）数据

云计算大数据处理 / 刘鹏等编著. -- 北京 : 人民邮电出版社, 2015.8
（学术中国. 院士系列. 未来网络创新技术研究系列）
ISBN 978-7-115-37810-1

Ⅰ. ①云… Ⅱ. ①刘… Ⅲ. ①计算机网络－数据处理
Ⅳ. ①TP393

中国版本图书馆CIP数据核字(2015)第067796号

内 容 提 要

本书介绍了基于云计算的大数据处理技术，重点介绍了一款高效的、实时分析处理海量数据的强有力工具——数据立方。数据立方是针对大数据处理的分布式数据库，能够可靠地对大数据进行实时处理，具有即时响应多用户并发请求的能力。

本书通过对当前主流的大数据处理系统进行深入剖析，阐述了数据立方产生的背景，介绍了数据立方的整体架构以及安装和详细开发流程，并给出了 4 个完整的数据立方综合应用实例。所有实例都经过验证并附有详细的步骤说明，无论是对于云计算的初学者还是想进一步深入学习大数据处理技术的研究和开发人员都有很好的参考价值。读者也可从本书配套网站中国云计算（http://www.chinacloud.cn）和中国大数据（http://www.thebigdata.cn）获取更多资料或求解疑难问题。

◆ 编　　著　刘　鹏　于　全　杨震宇　陈　伟　王　磊
　　张乃甜
　　责任编辑　代晓丽
　　责任印制　彭志环
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
　　邮编　100164　　电子邮件　315@ptpress.com.cn
　　网址　http://www.ptpress.com.cn
　　北京圣夫亚美印刷有限公司印刷
◆ 开本：700×1000　1/16
　　印张：13.5　　2015 年 8 月第 1 版
　　字数：265 千字　　2015 年 8 月北京第 1 次印刷

定价：78.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前 言

在互联网带来的大数据问题压力下，我们需要全新的思想，通过“积木化”的改变，重新定义计算资源的使用方式、服务的提供方式以及社会化大生产的协作过程。云计算带来了这种思想的落实机制，这种机制使我们可以组织资源以服务，组织技术以实现，组织流程以应变。而且，云计算扩大了我们对服务的定义，并带来了一种全新的计算资源管理思路，一种信息技术的系统工程理念和一次信息社会的工业化革命。

虽然云计算起步于企业界，但在发展过程中需要解决具有挑战性的技术问题。本着这种思想，我们与华为、中兴通讯、360 安全卫士、华胜天成、天威视讯和世纪鼎利等知名企业建立了紧密的联合研究关系，研究内容紧跟市场需求和技术发展，研究成果能够迅速地转化为生产力。在这本书里，我们将和大家分享其中一些研究成果。近年来，大数据技术在全世界迅猛发展，引起了全世界的广泛关注，掀起了一轮全球性的发展浪潮。大数据技术发展的主要推动力来自并行计算硬件和软件技术的发展，以及近年来行业大数据处理需求的迅猛增长。

感谢云创存储（www.cstor.cn）的研究团队为本书的撰写工作所作的贡献，特别是张晓燕、张国庆、贾文周、夏宇朗、李鹏芳、赵洪涛、钱力、刘晶晶、葛馨彤等。

欢迎大家关注我的微信公众号：刘鹏看未来（lpoutlook）。

解放军理工大学　刘鹏

2015 年 1 月 12 日

目 录

第1章 大数据挑战

由于数据规模的急剧膨胀，各行业累积的数据量越来越巨大，数据类型也越来越多、越来越复杂，已经超越了传统数据管理系统、处理模式的能力范围，于是，“大数据”概念就应运而生。

1.1 当前面临的大数据挑战

最早提出“大数据时代已经到来”的机构是全球知名咨询公司麦肯锡[1]。麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素，而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。

1.1.1 大数据急剧膨胀

在麦肯锡的报告发布后，大数据迅速成为计算机行业争相传诵的热门概念，也引起了其他行业的高度关注。随着互联网技术的不断发展，“数据本身是资产”这一观点在业界已经达成共识。事实上，全球互联网巨头都已意识到大数据时代数据的重要意义，包括 EMC、惠普、IBM 和微软在内的全球 IT 巨头，纷纷通过收购大数据相关厂商来实现技术整合，可见其对大数据的重视。

IBM 公司称，全球每天生成的数据量达 2.5 EB，而且其增长速度在不断加快。大家或许对 EB 这一单位不太熟悉，如果换算成家庭使用的普通蓝光光盘，2.5 EB 相当于 10 亿张容量为 25 GB 的蓝光光盘。而且令人吃惊的是，人类迄今为止生成的数据中，有 90%是在近两年内产生的。除了电子邮件、社交网络中输入的数

据、用手机随手拍摄的图片以及向娱媒投稿的视频外，还有各种传感器收集的数据以及网上购物的日志数据等。现在，每秒就会产生 1 万条微博等信息，这些数据被称为非结构化数据，在生成的数据中约占 80%。非结构化数据今后还将会爆发性增加，在今后 5 年内将增加 800%。

另外，根据 IDC（国际数据公司）的监测统计，2011 年全球数据总量已经达 1.8 ZB，而这个数值还在以每两年翻一番的速度增长，预计到 2020 年全球将总共拥有 35 ZB 的数据量，增长近 20 倍。

IDC 研究表明，数字领域存在着 1.8 万亿 GB 的数据。企业数据正在以 55%的速度逐年增长。如今，只需两天就能创造出自文明诞生以来到 2003 年所产生的数据总量。现代企业正在经历规模化、多样化、高速化的数据挑战，大数据已成为重要的时代特征。IBM 全球副总裁兼大中华区软件集团总经理胡世忠表示："大数据正带来一场信息社会的变革[2]。"大量的结构化及非结构化数据以及流数据的广泛应用，致使企业需要重新思考已有的 IT 模式；与此同时，大数据又将推动企业进行又一次基于信息革命的业务转型，使企业能够借助大数据带来的大洞察，获取更多的商业价值和发展机会，在激烈的市场竞争中脱颖而出。2005 年，人们创造的信息量达到了 150 EB，而到 2011 年，这一数字达到了 1.8 ZB，预计到 2020 年，这一数字将突破 35 ZB。这便是大数据时代的来临。这样的数据量是巨大的，是 IT 刚开始时所无法想象的，不过企业必须找到最好的、有竞争力的方式来解决这些数据。《麻省理工学院斯隆管理评论》和 IBM 商业价值研究院联合举行的 2011 年新智能企业全球高管调查和研究项目指出，2011 年，58%的企业已经将分析技术用于在市场或行业内创造竞争优势，实现业务价值，这一数据比 2010 年增加了 21%。毫无疑问，大数据时代的机遇和挑战同时摆在了企业面前。

国家工业和信息化部软件服务业司司长陈伟为大数据概括出 4 个方面特征：第一是体量大，是一个数据全集的概念；第二是类型多，包括结构化数据、半结构化数据、非结构化数据等多种类型，其中视频数据在目前占到了 90%以上的总额；第三要求速度快，需要以秒级为目标进行实时动态处理；第四是价值密度，由于大量有用和可能没用的数据并存，所以大数据的目的是从庞大的数据集合中寻找有价值的数据和知识。"以交互数据为例，目前一些自媒体平台，比如新浪微博，每天都有超过 2 500 万条的微博信息在发布，里面有很多有价值的信息尚未得到发掘"，中国电子信息产业发展研究院副总工程师李峻认为，"在这样庞大的非结构化数据背后，如何利用大数据技术，从海量堆积的交互数据当中发现带有趋势性、前瞻性的信息，就能够发现并产生巨大的社会价值和商业价值。"

1.1.2 大数据智能分析

随着互联网科技日益成熟，各种类型的数据增长将会超越历史上任何一个时

期；用户想要从这庞大的数据库中提取对自己有用的信息，就离不开大数据分析技术和工具。越来越多企业开始使用 Hadoop 平台处理大量数据。在 2014 年 Hadoop 峰会上发布的数据显示，2009 年 Hadoop 服务提供商总共只有 9 家，而 2014 年已经超过了 120 家。

仅靠 Hadoop 服务无法解决企业的大数据问题，很多传统的数据库管理系统开始整合 Hadoop 服务，以便更好地为企业服务。例如惠普、戴尔、甲骨文、IBM 等知名公司分别都有针对自家需求的 Hadoop 服务。

在相关大数据分析处理技术出现前，IT 公司经理们通常要对公司数据进行删选以便用户查询和分析。现在，各种大数据分析工具既方便用户查询数据，又能避免泄露公司机密；同时，所有原始数据都将完好保存。

阻碍大数据分析技术或是使用 Hadoop 的原因之一就是缺乏相应的技术、环境/数据安全以及可行性。幸好，许多开源和专利软件社区都已经着手解决这些问题。

目前，有一半以上的企业还在利用磁盘进行数据存档、备份和恢复。但随着大数据分析技术日渐成熟，磁盘终将被淘汰。

Hadoop 平台对业务的针对性较强，现粗略地从几个角度将大数据分析的业务需求分类，针对不同的具体需求，应采用不同的数据分析架构。

按照数据分析的实时性，分为实时数据分析和离线数据分析两种。

实时数据分析一般用于金融、移动和互联网 B2C 等产品，往往要求在数秒内返回上亿行数据的分析，从而达到不影响用户体验的目的。要满足这样的需求，可以采用精心设计的传统关系型数据库组成并行处理集群，也可以采用一些内存计算平台或 HDD 的架构，这些无疑都需要比较高的软硬件成本。目前比较新的海量数据实时分析工具有 EMC 的 Greenplum、SAP 的 HANA 等。

对于大多数反馈时间要求不是那么严苛的应用，比如离线统计分析、机器学习、搜索引擎的反向索引计算、推荐引擎的计算等，应采用离线分析的方式，通过数据采集工具将日志数据导入专用的分析平台。但面对海量数据，传统的 ETL 工具往往彻底失效，主要原因是数据格式转换的开销太大，在性能上无法满足海量数据的采集需求。互联网企业的海量数据采集工具，有 Facebook 开源的 Scribe、LinkedIn 开源的 Kafka、淘宝开源的 TimeTunnel、Hadoop 的 Chukwa 等，均可以满足每秒数百 MB 的日志数据采集和传输需求，并将这些数据上载到 Hadoop 中央系统上。

海量数据级别指的是对于数据库和 BI 产品已经完全失效或者成本过高的数据量。海量数据级别的优秀企业级产品也有很多，但基于软硬件成本的原因，目前大多数互联网企业采用 HDFS（Hadoop 分布式文件系统）来存储数据，并使用 MapReduce 进行分析。本文稍后将主要介绍 Hadoop 上基于 MapReduce 的一个多

维数据分析平台。

根据不同的业务需求，数据分析的算法也差异巨大，而数据分析的算法复杂度与架构是紧密关联的。举个例子，Redis 是一个高性能的 Key-Value 数据库，它支持 List、Set、SortedSet 等简单集合，如果数据分析需求通过排序、链表就可以解决，同时总的数据量不大于内存（准确地说是内存加上虚拟内存再除以 2），那么使用 Redis 会达到非常惊人的分析性能。还有很多易并行（Embarrassingly Parallel）问题，计算可以分解成完全独立的部分，或者很简单地就能改造出分布式算法，比如大规模脸部识别、图形渲染等，这样的问题自然是使用并行处理集群比较适合。而大多数统计分析、机器学习问题都可以用 MapReduce 算法改写。MapReduce 目前最擅长的计算领域有流量统计、推荐引擎、趋势分析、用户行为分析、数据挖掘分类器、分布式索引等。

1.1.3 大数据深度挖掘

智库百科是这样描述数据挖掘（Data Mining）的：数据挖掘又称数据库中的知识发现，是目前人工智能和数据库领域研究的热点问题，所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知并有潜在价值的信息的非平凡过程。数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，做出正确的决策。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。这个定义包括 4 层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

与数据挖掘相近的同义词有数据融合、人工智能、商务智能、模式识别、机器学习、知识发现、数据分析和决策支持等。

何为知识？从广义上理解，数据、信息也是知识的表现形式，但是人们更多的是把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，例如关系数据库中的数据；也可以是半结构化的，例如文本、图形和图像数据；甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决

策支持。在这种需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。

从商业角度的定义，数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

简而言之，数据挖掘是一类深层次的数据分析方法。数据分析本身已经有很多年的历史，而过去数据收集和分析的目的是用于科学研究。另外，由于当时计算能力的限制，对大数据量进行分析的复杂数据分析方法受到了很大限制。现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据不再是为了分析而收集，而是由纯机会的（Opportunistic）商业运作产生。分析这些数据也不再是单纯出于研究的需要，更主要是为商业决策提供真正有价值的信息，进而获得利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少。从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。

因此，数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的、先进的、有效的方法。

1.1.4　业务与技术脱节

据 Gartner 调查统计，80%的企业业务正运行在 IT 系统上，IT 影响力已经遍布于企业的各个角落。离开 IT 技术，业务也将无法正常运行。因此，随着 IT 在业务中的地位越来越重要，企业对于 IT 的要求也越来越高。这种转变说明 IT 再也不是停留在基础设施层面上的单纯技术，而是需要为企业的业务成效贡献力量的核心因素之一。让众多 CIO 感到苦恼的是，如此巨大的 IT 投入，仍无法动态满足业务的发展需求，IT 与业务的脱节让全球经济危机的严冬更加寒冷。

1.2　大数据催生云计算

现在已经从过去资本经济的时代进入数字经济的时代。特别是我们所看到的虚拟世界、物理世界与人类社会相关联时，已经创造出了更多与以前不一样的数据。所以有学者说，18 个月翻一番的数据量导致存储和处理能力的提高开始落后

于现有数据增长的幅度，导致现在的知识社会中面临着最大的瓶颈。在这个瓶颈下，过去的数据（以商业数据为主）是确定了的数据。数据管理能力、数据处理能力、高可靠安全服务能力这 3 个能力的局限性和发展空间，为现在的数据处理带来了新的机会，也就是数据和经济社会密切相连。因需求而催生的技术往往最具生命力和广泛、残酷的市场竞争性，行业洗牌，甚至国家竞争力都附着于此。也正因此，云计算被称为是继大型计算机、个人计算机、互联网之后的第四次 IT 产业革命，它不仅改变了网络应用的模式，也将成为带动 IT、物联网、电子商务等诸多产业强劲增长、推动信息产业整体升级的基础。

1.2.1 云计算不是偶然

2006 年 Google 推出了“Google 101 计划”，并正式提出“云”的概念和理论。随后亚马逊、微软、惠普、雅虎、英特尔、IBM 等公司都宣布了自己的云计划，“云安全”、“云存储”、“内部云”、“外部云”、“公共云”、“私有云”等一堆让人眼花缭乱的概念在不断冲击人们的神经。那么到底什么是云计算技术？云计算技术的产生、概念、原理、应用和前景又在哪里？

1．云计算思想的产生

在传统模式下，企业建立一套 IT 系统不仅需要购买硬件等基础设施，还需要购买软件的许可证，需要专门的人员维护。当企业的规模扩大时，还要继续升级各种软硬件设施以满足需要。对于企业来说，计算机等硬件和软件本身并非是他们真正需要的，它们仅是完成工作、提供效率的工具而已。对个人来说，我们想正常使用电脑需要安装许多软件，而许多软件是收费的，对不经常使用该软件的用户来说购买是非常不划算的。能否有这样的服务，能够提供我们需要的所有软件供我们租用？这样我们只需要在使用时付少量租金即可租用这些软件服务，为我们节省许多购买软硬件的资金。

我们每天都要用电，但不是每家自备发电机，它由电厂集中提供；我们每天都要用自来水，但不是每家都有井，它由自来水厂集中提供。这种模式极大地节约了资源，方便了我们的生活。面对计算机带来的困扰，是否可以像使用水和电一样使用计算机资源？这些想法最终导致了云计算的产生。

云计算的最终目标是将计算、服务和应用作为一种公共设施提供给公众，使人们能够像使用水、电、煤气和电话那样使用计算机资源。

云计算模式即为电厂集中供电模式。在云计算模式下，用户的计算机会变得十分简单，不大的内存、不需要硬盘和各种应用软件，就可以满足我们的需求，用户的计算机除了通过浏览器给“云”发送指令和接收数据外，基本上什么都不用做，便可以使用云服务提供商的计算资源、存储空间和各种应用软件。这就像