

<<大数据>>

图书基本信息

书名：<<大数据>>

13位ISBN编号：9787121202650

10位ISBN编号：7121202654

出版时间：2013-5

出版时间：电子工业出版社

作者：周宝曜，刘伟，范承工

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<大数据>>

### 内容概要

本书从实际技术解决方案出发，提出了大数据技术四层架构，即基础设施层、管理层、分析层、应用层。

在此基础上，全面剖析了当前大数据领域中的主流技术，并配以行业应用实例和一线研发人员的独到见解。

力求使读者能够通过阅读此书，全面了解当前大数据技术动态和发展趋势，并可针对自己面临的大数据问题找到可行的解决方案。

## 作者简介

周宝曜，博士，现任EMC中国研究院数据科学实验塞主任，资深主任研究员。

新加坡南洋理工大学计算机工程博士，清华大学学士和硕士。

主要从事大数据相关的蓄理集构和分析挖掘算法的研究。

曾就职于惠普中国研究院和IBM中国研究院。

已发表国际学术论文30余篇，并拥有4项美国专利和多项中国专利。

刘伟，博士，现任EMC北京研发中心总经理、EMC中国研究院院长。

此前蓄任思辩大中华区高级副总裁、中国研发中心高级总监。

惠蓄中国研究院院长等职务。

现担任CCF高级会员，中国电子学会云计算专家委员会委员，中国计算机学会大数据专家委员会委员等社会职务。

美国科罗拉多州立大掌电子工程与计算机科学博士，南开大学电子学及物理学本科及硕士。

范承工，博士，是VMware中国研发中心和EMC中国卓越研发集团的创始人，现任VMware全球高级副总裁，负责领导VMware应用存储与数据产品的全球研发团队。

加州理工学院电子工程硕士和博士，纽约古柏联舍学院电子工程学士。

国际知名分布式系统和网络存储技术专家。

曾是Rainfinity公司联合创始人之一，任首席技术官。

## 书籍目录

第一部分大数据技术概览 第1章概述2 1.1什么是大数据2 1.1.1大数据的定义及特征3 1.1.2大数据结构类型6 1.1.3大数据实例8 1.2大数据发展史10 1.3大数据技术架构12 1.4机遇与挑战14 参考文献16 第2章大数据应用18 2.1大数据驱动新应用18 2.1.1大数据生态系统18 2.1.2新的业务应用20 2.2行业应用实例22 2.2.1奥巴马的大数据22 2.2.2预测犯罪23 2.2.3数据让游戏更精彩24 2.2.4智能交通25 2.2.5大学教育27 2.2.6大数据的姻缘28 2.2.7传媒出版29 参考文献31 第3章大数据基础设施32 3.1云端大数据32 3.1.1云基础设施34 3.1.2虚拟化的三驾马车34 3.1.3云安全和云平台35 3.2计算虚拟化37 3.2.1基本概念38 3.2.2从部分虚拟化到全虚拟化39 3.2.3处理器（CPU）的虚拟化40 3.2.4内存（Memory）的虚拟化42 3.3大数据存储44 3.3.1传统存储系统时代的简单回顾44 3.3.2大数据时代的新挑战45 3.3.3分布式存储及其案例47 3.3.4云存储及其存储虚拟化48 3.3.5大数据存储的其他需求及其特点49 3.4网络虚拟化52 3.4.1网卡虚拟化52 3.4.2虚拟交换机（VirtualSwitch）54 3.4.3接入层的虚拟化54 3.4.4覆盖网络虚拟化（Network VirtualizationOverlay）56 3.4.5软件定义的网络（SDN）58 3.4.6对大数据处理的意义60 3.5基础架构的安全：云环境中面临的新的安全挑战60 3.5.1计算资源方面的安全和挑战62 3.5.2存储方面的安全和挑战64 3.5.3网络方面的安全和挑战66 3.6大数据时代的云服务69 3.6.1大数据与基础设施即服务70 3.6.2亚马逊云计算服务的解决方案71 3.6.3OpenStack解决方案72 3.6.4大数据与应用平台即服务74 参考文献77 第4章大数据管理80 4.1大数据事务处理（OLTP）80 4.1.1NoSQL83 4.1.2NewSQL94 4.2大数据分析处理（OLAP）99 4.2.1分布式大规模批量处理（MapReduce/Hadoop）100 4.2.2MPP数据库114 4.3流数据管理117 4.3.1流数据管理简介117 4.3.2复杂事件处理简介117 4.3.3复杂事件处理软件Esper介绍120 4.3.4大数据流处理127 4.3.5大数据摄取与处理132 参考文献133 第5章大数据分析136 5.1数据分析的演变与现状136 5.1.1数据分析的商业驱动力136 5.1.2面向分析的数据环境的演变137 5.1.3传统分析架构138 5.2大数据分析平台139 5.2.1大数据分析平台的要点140 5.2.2大数据分析平台实例：Cetas144 5.3高级分析理论与方法146 5.3.1聚类分析147 5.3.2关联规则151 5.3.3回归和分类预测160 5.4数据可视化170 5.4.1数据可视化基础171 5.4.2用数据讲故事172 5.4.3数据可视化的模式177 5.4.4数据可视化工具基础182 5.4.5大数据的可视化188 参考文献189 第6章数据科学与数据科学家193 6.1商业智能vs数据科学193 6.2数据科学家194 6.2.1大数据生态系统中的关键角色194 6.2.2数据科学家的特质195 6.3数据分析生命周期模型196 6.3.1模型概述197 6.3.2阶段1：探索发现199 6.3.3阶段2：数据准备201 6.3.4阶段3：模型规划203 6.3.5阶段4：模型建造205 6.3.6阶段5：沟通结果206 6.3.7阶段6：项目实施207 6.4使用范例：企业创新分析209 6.4.1阶段1：探索发现209 6.4.2阶段2：数据准备213 6.4.3阶段3：模型规划218 6.4.4阶段4：模型建造219 6.4.5阶段5：沟通结果223 6.4.6阶段6：项目实施223 参考文献224 第二部分大数据解决方案范例 第7章医疗大数据解决方案228 7.1医疗信息化228 7.1.1全球医疗信息化历史回顾228 7.1.2我国医疗信息化发展趋势230 7.2医疗数据综述230 7.2.1医疗数据的大数据特性231 7.2.2医疗大数据挑战和机遇233 7.3医疗大数据基础架构234 7.3.1建设原则234 7.3.2面向医疗大数据的信息基础架构方案235 7.4医疗大数据分析246 7.4.1医疗云的兴起246 7.4.2医疗云上的大数据248 7.4.3医疗大数据分析解决方案249 7.5医疗大数据的展望250 参考文献252 第8章物联网大数据解决方案253 8.1物联网253 8.1.1物联网的概念253 8.1.2物联网技术254 8.1.3物联网数据254 8.1.4物联网的机遇和挑战254 8.1.5物联网应用实例255 8.2应用行业背景256 8.2.1脱硫系统的必要性256 8.2.2脱硫系统工作原理256 8.2.3大数据时代的数据挖掘257 8.3参数分析258 8.3.1火电厂的大数据258 8.3.2脱硫相关参数259 8.4优化目标259 8.4.1脱硫参数优化259 8.4.2目标成本优化260 8.5优化方法260 8.5.1基于数据的理论与方法260 8.5.2最优化脱硫系统可调参数261 8.5.3最小化脱硫系统成本262 8.6数据相关问题262 8.6.1主要监控参数262 8.6.2业务相关假设263 8.6.3数据中存在的问题263 8.7优化目标1：脱硫运行参数最优目标值挖掘263 8.7.1数据分布直方图263 8.7.2基于历史数据的工况划分266 8.7.3FCM与模糊关联规则挖掘最优可调参数267 8.8优化目标2：最优目标成本计算269 8.8.1增压风机用电成本估计269 8.8.2石灰石成本函数270 8.9实现简介271 8.9.1基于HBase的数据模型272 8.9.2对Mahout的改进272 8.10总结272 参考文献273 第9章移动平台大数据解决方案274 9.1移动平台的大数据挑战274 9.2Instagram案例研究276 9.2.1面临的挑战277 9.2.2解决方案277 9.3MobileBack—endasaService基础279 9.4MBaaS提供商案例研究283 9.5基于PaaS的MBaaS大数据解决方案288 参考文献289 第10章社交网站大数据解决方案291 10.1大数据时代社交网站面临的挑战291 10.2Twitter解决方案294 10.2.1Twitter在线部分大数据解决方案295 10.2.2Twitter离线部分

<<大数据>>

大数据解决方案297 10.3LinkedIn解决方案297 10.4Facebook解决方案300 10.5国内社交网络解决方案302  
10.5.1腾讯大数据解决方案302 10.5.2新浪微博大数据解决方案303 参考文献304 第11章大数据未来展  
望306 11.1大数据发展趋势306 11.2新的机遇与挑战307 参考文献310

## 章节摘录

版权页：插图：3.4.6对大数据处理的意义 相对于普通的应用，大数据的分析与处理对网络有着更高的要求。

要求涉及从带宽到延时，从吞吐率到负载均衡，以及可靠性、服务质量控制等方方面面。

同时随着越来越多的大数据应用部署到云计算平台中，对虚拟网络的管理需求就越来越高。

首先，网络接入设备虚拟化的发展，在保证多租户服务模式的前提下，还能同时兼顾高性能与低延时、低CPU占用率。

其次，接入层的虚拟化保证了虚拟机在整个网络中的可见性，使得基于虚拟机粒度（或大数据应用粒度）的服务质量控制成为可能。

覆盖网络的虚拟化，一方面使得大数据应用能够得到有效的网络隔离，更好地保证了数据通信的安全；另一方面也使得应用的动态迁移更加便捷，保证了应用的性能和可靠性。

软件定义的网络更是从全局的视角来重新管理和规划网络资源，使得整体的网络资源利用率得到优化利用。

总而言之，网络虚拟化技术通过对性能、可靠性和资源优化利用的贡献，间接提高了大数据系统的可靠性和运行效率。

3.5基础架构的安全：云环境中面临的新的安全挑战 云计算在安全方面引入了怎样的新问题呢？

那么首先得回答云计算的本质是什么？

所谓云计算，宏观来讲就是按需利用“远端”的资源（诸如CPU，内存，网络，存储等）进行“计算”，以达到和“本地”计算等价甚至更好的目的，诸如省钱，管理便捷，可扩展性强，绿色环保等。

然而天下没有免费的午餐，云计算在数据安全方面引入了很多新问题。

譬如在云计算基础架构服务层（Infrastructure as a Service, IaaS），主要有以下两类问题：新的安全问题，诸如信任问题（特指租客和云服务商之间），多租客之间的资源隔离问题；对已有的安全攻击，IaaS是否更容易被攻击？

或者存在新的技术方法去避免这些攻击。

安全问题中的信任和隔离问题，源于云计算的新模型。

由于资源使用者和管理者角色的分离，衍生IaaS的使用者和IaaS提供者之间的信任问题。

我们把云资源的使用者，称为云租户。

比如一个小型公司租赁了Amazon上的EC2服（主要指虚拟机），并在EC2上搭建了一个网站，那么这个公司就是AmazonEC2的租户，而使用网站的用户只是这个小公司的客户。

由于资源不由租客完全控制，那么租客就有疑问：怎么确定租赁的资源仅仅为我所用，而不被其他租客或者云管理员非法使用，导致数据的丢失或者泄露。



#### 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>