



大数据技术与应用专业规划教材



大数据云服务技术 架构与实践

© 李天目 编著

清华大学出版社





大数据技术与应用专业规划教材



大数据云服务技术 架构与实践

◎ 李天目 编著

清华大学出版社
北京

内 容 简 介

本书是从大数据和云计算相结合的视角,系统地介绍大数据云架构技术与实践的专业图书,全书分为五篇 19 章,分别介绍大数据云计算的概论、关键技术、体系架构、云架构实践与编程和安全。本书层次清晰,结构合理,主要内容包括大数据云计算关系、大数据应用价值、分布式计算、NoSQL 数据库、机器学习、虚拟化、Docker 容器、Web 2.0、绿色数据中心、基础设计即服务(IaaS)、平台即服务(PaaS)、软件即服务(SaaS)、容器即服务(CaaS)、大数据云架构搭建、Spark 大数据编程、大数据和云计算面临的安全威胁、保障大数据安全、应用大数据保障安全等。

本书可作为高年级本科生和研究生教材,也可作为广大科学技术人员和计算机爱好者的参考书。

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

大数据云服务技术架构与实践/李天目编著. —北京:清华大学出版社,2016

(大数据技术与应用专业规划教材)

ISBN 978-7-302-45460-1

I. ①大… II. ①李… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 274722 号

责任编辑:魏江江 赵晓宁

封面设计:刘 键

责任校对:焦丽丽

责任印制:沈 露

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

课 件 下 载: <http://www.tup.com.cn>, 010-62795954

印 装 者:北京嘉实印刷有限公司

经 销:全国新华书店

开 本:185mm×260mm

印 张:21.75

字 数:516 千字

版 次:2016 年 11 月第 1 版

印 次:2016 年 11 月第 1 次印刷

印 数:1~2000

定 价:49.50 元

移动互联网、电子商务及社交媒体的快速发展使得人类需要面临的数据量呈指数增长。根据 IDC《数字宇宙》(Digital Universe) 研究报告显示,2020 年全球新建和复制的信息量将超过 40ZB,是当前的 10 倍;而中国的数据量则会在 2020 年超过 8ZB,比当前增长 20 倍。数据量的飞速增长带来了大数据技术的发展和市场的繁荣,同时在学术界,关于大数据的科研工作如火如荼,越来越多的学者投入到大数据云计算研究之中,相关文献呈指数增长。

不断积累的大数据包含着很多在小数据量时不具备的深度知识和价值,带来巨大的技术创新与商业机遇。大数据分析挖掘将为行业/企业带来巨大的商业价值,实现各种高附加值的增值服务,进一步提升行业/企业的经济效益和社会效益。谈到大数据,不可避免地要提及云计算,云计算结合大数据,这是时代发展的必然趋势。有人把云计算和大数据比作是一个硬币的两面。云计算是大数据的 IT 基础和平台,而大数据是云计算范畴内最重要、最关键的应用。大数据必然架构在云上才能高效运作并对外服务,两者之间缺一不可,相辅相成,相互促进。

大数据云计算正在快速发展,相关技术热点也呈现百花齐放的局面,业界各大厂商纷纷制定相应的战略,新的概念、观点和产品不断涌现。大数据和云计算作为新一代 IT 技术变革的核心,必将成为广大学生、科技工作者构建自身 IT 核心竞争能力的战略机遇。因而作为高层次 IT 人才,学习大数据和云计算知识,掌握相关技术迫在眉睫。然而,大数据和云计算密切相关,当前国内外相关的资料还相当少,缺乏比较系统完整的论述。目前在我国急需一本教材能够全面而又系统地讲解大数据和云计算,普及大数据和云计算知识,推广云计算中大数据的应用,解决大数据架构的实际问题,进而培养高层次大数据人才。

在这样的背景下,作者着眼于大数据和云计算有机结合的视角,从理论探索和应用实践两个方面来撰写本书,适合对大数据和云计算具有初步认识并希望对大数据云架构进行深入、全面了解,并进行实践的计算机信息相关专业高年级本科生使用,同时本书也将成为广大专业工程技术人员不可缺少的参考资料。本书分为 5 篇 19 章,第 1~第 3 章为大数据云计算概论篇,第 4~第 10 章为大数据云计算关键技术篇,第 11~第 14 章为云计算架构篇,第 15 和第 16 章为大数据云架构实践与编程篇,第 17~第 19 章为大数据安全篇。

各章内容如下:

第 1 章大数据概述,介绍大数据产生背景,大数据发展历程,大数据概念和特点,大数据应用场景,大数据研究展望等内容。

第 2 章大数据和云计算,介绍大数据和云计算的关系,云计算的概念,云计算的由来,云

计算类型,云计算的商业模式。

第3章大数据应用价值,介绍大数据的应用价值,数据的生成、分析、存储、分享、检索、消费都在大数据的生态系统中进行,应用大数据技术,让数据参与决策,发掘找到大数据真正有效的价值,进而改变人们的未来,革新生活模式,产生社会变革。

第4章分布式计算框架,介绍构成大数据云计算的主要关键技术——分布式计算技术,以及 Hadoop、Spark、Flink 等分布式大数据计算框架。

第5章 NoSQL 数据库,介绍 NoSQL(NoSQL = Not Only SQL),这是一项全新的数据库技术,然后引出分布式数据库的重要理论 CAP,最后介绍 HBase。

第6章机器学习,介绍机器学习的概念、分类和发展历程,简要介绍多种机器学习算法。

第7章虚拟化,介绍构成云计算的主要关键技术——虚拟技术,它整合多种计算资源,实现架构动态化,并达到集中管理和动态使用物理资源及虚拟资源,以提高系统结构的弹性和灵活性,降低成本、改进服务、减少管理风险等目标。

第8章 Docker 容器,介绍 Docker 容器相关的概念、优势、由来和实现原理。

第9章 Web 2.0,介绍构成云计算主要的关键技术 Web 2.0,是因特网的一次理念和思想体系的升级换代,由来自上而下的由少数资源控制者集中控制主导的因特网体系转变为自下而上的由广大用户集体智慧和力量主导的因特网体系。

第10章绿色数据中心,介绍构成云计算的主要关键技术——绿色数据中心,是指数据机房中的 IT 系统、机械、照明和电气等能取得最大化的能源效率和最小化的环境影响。

第11章基础设施即服务,介绍云计算环境中的 IaaS(Infrastructure as a Service),分析 Amazon 公司的 IaaS 案例。

第12章平台即服务,介绍云计算环境中的 PaaS(Platform as a Service),分析 Google App Engine 和 Windows Azure Platform 的 PaaS 案例。

第13章软件即服务,介绍云计算环境中的 SaaS(Software as a Service),分析 Salesforce 的 SaaS 案例。

第14章容器即服务,介绍云计算环境中的 CaaS(Container as a Service),阐述 Kubernetes 和 Mesos 容器调度框架,分析互联网公司 SAE 容器云和互联网公司“去哪网”容器云。

第15章大数据云架构搭建,介绍分布式的 Hadoop 与 Spark 集群搭建和基于 Docker 容器的 Spark 大数据云架构。

第16章 Spark 大数据编程,介绍使用 IntelliJ IDEA 构建 Spark 开发环境,并列举应用 Spark 计算框架的 WordCount 和基于 Spark Streaming 股票趋势预测案例。

第17章大数据云计算面临的安全威胁,介绍大数据云计算面临的各种安全威胁,阐述不同行业大数据安全的需求,指出大数据安全应该包括保障大数据安全和大数据用于安全两个层面的含义。

第18章保障大数据安全,介绍保障大数据安全的相关技术和相关实践。

第19章应用大数据保障安全,介绍应用大数据保障安全,包括大数据安全检测及应用,安全大数据,基于大数据的网络态势感知和视频监控数据的安全应用等方面内容。

在本书最后给出了相关的参考文献,有兴趣的读者可以进一步阅读。此外,关于虚拟化技术,作者认为普通虚拟化和容器虚拟化是完全不同的两种技术,大数据云平台多采用容器

架构,所以 Docker 容器技术作为本书独立一章,并且在第 4 篇详细介绍基于 Docker 容器的大数据云架构实践。在方兴未艾的大数据云计算时代,统一的标准和解决方案还未成形,不同人在不同背景下的需求和观点是不一样的,我们花费一年多的时间努力编著本书,希望能提供比较深入的见解,每一个对大数据和云计算感兴趣的读者都能学有所得。

更进一步,大数据和云计算是新一代 IT 技术变革的核心,是中国建立自己 IT 体系的战略机遇,阅读本书,期待读者既能从宏观角度更全面地认识大数据云架构,同时也能从微观技术实践角度接触大数据和云计算,更深入地学习和掌握大数据和云计算知识。

本书适合于从头到尾阅读,也可以按照喜好和关注点挑选独立的章节阅读。希望本书的介绍能加深读者对云计算的理解。

由于编者水平有限,书中不妥之处在所难免,恳请读者批评指正。

编 者

2016 年 5 月

第 1 篇 大数据云计算概论

第 1 章 大数据概述	3
1.1 大数据产生与发展	3
1.1.1 大数据产生背景	3
1.1.2 大数据发展历程	3
1.1.3 当前大数据	5
1.2 大数据概念与特征	6
1.2.1 大数据概念	6
1.2.2 大数据的特点	7
1.3 大数据应用	7
1.3.1 企业内部大数据应用	8
1.3.2 物联网大数据应用	8
1.3.3 面向在线社交网络大数据的应用	9
1.3.4 医疗健康大数据应用	9
1.3.5 群智感知	10
1.3.6 智能电网	10
1.4 大数据的研究与展望	11
1.5 本章小结	12
第 2 章 大数据和云计算	14
2.1 大数据和云计算的关系	14
2.1.1 大数据和云计算关系概述	14
2.1.2 云计算是大数据处理的基础	15
2.1.3 大数据是云计算的延伸	16
2.2 云计算概念	16
2.2.1 云的兴起	16
2.2.2 云计算的定义及其特点	17
2.2.3 云计算名称的来历	19

2.3	云计算类型	19
2.3.1	基础设施类	19
2.3.2	平台类	21
2.3.3	应用类	22
2.3.4	以所有权划分云计算系统类型	23
2.4	云计算商业模式	24
2.4.1	商业模式是云计算的基石	24
2.4.2	云计算的市场规模	24
2.4.3	云计算商业模式分析	24
2.5	本章小结	25
第3章	大数据应用价值	26
3.1	大数据在电子商务中的应用	27
3.1.1	大数据是电子商务发展要素	27
3.1.2	电子商务大数据的实用措施	27
3.1.3	电子商务大数据的转型路径	28
3.2	大数据在金融的应用	29
3.2.1	大数据金融的提出	29
3.2.2	大数据金融的功能	34
3.2.3	大数据金融的挑战	35
3.2.4	大数据金融创新	37
3.3	大数据在媒体的应用	39
3.3.1	传统媒体的不足	40
3.3.2	大数据驱动传统媒体的升级	40
3.3.3	大数据引领新媒体发展	41
3.4	大数据在医疗上的应用	43
3.4.1	大数据改进临床决策支持系统	44
3.4.2	大数据助推医疗产品研发	44
3.4.3	大数据催生新医疗服务模式	45
3.5	大数据在教育上的应用	47
3.5.1	大数据教育与传统教育的优势	47
3.5.2	大数据教学模式的不断改善	47
3.5.3	教育大数据市场的广阔前景	48
3.5.4	大数据变革教育应用的实践措施	49
3.6	本章小结	49

第2篇 大数据云计算关键技术

第4章	分布式计算框架	53
4.1	分布式计算基本概念	53

4.1.1	分布式计算与并行计算	53
4.1.2	分布式计算和并行计算的比较	54
4.2	Hadoop 系统介绍	55
4.2.1	Hadoop 发展历程	55
4.2.2	Hadoop 使用场景和特点	56
4.2.3	Hadoop 项目组成	57
4.3	分布式文件系统	57
4.3.1	分布式文件系统概述	57
4.3.2	HDFS 架构	58
4.3.3	HDFS 设计特点	59
4.4	MapReduce 计算模型	60
4.4.1	MapReduce 概述	60
4.4.2	MapReduce 应用实例	61
4.4.3	MapReduce 实现和架构	62
4.5	分布式协同控制	63
4.5.1	常见分布式并发控制方法	63
4.5.2	Google Chubby 并发锁	64
4.6	Spark 计算框架	66
4.6.1	Spark 简介	66
4.6.2	Spark 生态系统	67
4.7	Flink 计算框架	72
4.7.1	Flink 简介	72
4.7.2	Flink 中的调度简述	73
4.7.3	Flink 的生态圈	74
4.8	本章小结	74
第 5 章	NoSQL 数据库	76
5.1	NoSQL 数据库概述	76
5.1.1	NoSQL 数据库的 4 大分类	76
5.1.2	数据库系统 CAP 理论和 BASE 理论	78
5.1.3	NoSQL 的共同特征	79
5.2	Hbase 数据库	80
5.2.1	HBase 简介	80
5.2.2	HBase 访问接口	80
5.2.3	HBase 数据模型	81
5.2.4	MapReduce on HBase	82
5.2.5	HBase 系统架构	83
5.3	本章小结	87

第 6 章 机器学习 88

6.1 机器学习概述..... 88

6.1.1 机器学习分类 88

6.1.2 机器学习发展历程 92

6.2 机器学习常用的算法..... 93

6.2.1 回归算法 93

6.2.2 基于实例的算法 93

6.2.3 正则化方法 94

6.2.4 决策树算法 94

6.2.5 贝叶斯方法 94

6.2.6 基于核的算法 95

6.2.7 聚类算法 95

6.2.8 关联规则学习 96

6.2.9 遗传算法 96

6.2.10 人工神经网络..... 97

6.2.11 深度学习..... 97

6.2.12 降低维度算法..... 97

6.2.13 集成算法..... 98

6.3 本章小结..... 98

第 7 章 虚拟化 99

7.1 虚拟化概述..... 99

7.1.1 虚拟化发展历史 99

7.1.2 虚拟化技术的发展热点和趋势..... 100

7.1.3 虚拟化技术的概念..... 101

7.2 虚拟化的分类 102

7.2.1 从实现的层次划分..... 102

7.2.2 从应用的领域划分..... 105

7.3 应用虚拟化 109

7.3.1 应用虚拟化的使用特点..... 109

7.3.2 应用虚拟化的优势..... 110

7.3.3 应用虚拟化要考虑的问题..... 111

7.4 桌面虚拟化 111

7.4.1 桌面虚拟化优势..... 111

7.4.2 桌面虚拟化使用条件..... 112

7.5 服务器虚拟化 112

7.5.1 服务器虚拟化架构..... 112

7.5.2 CPU 虚拟化 113

7.5.3	内存虚拟化	115
7.5.4	I/O 虚拟化	117
7.6	网络虚拟化	118
7.6.1	传统网络虚拟化技术	118
7.6.2	主机网络虚拟化	119
7.6.3	网络设备虚拟化	121
7.7	存储虚拟化	125
7.7.1	存储虚拟化概述	125
7.7.2	按照不同层次划分存储虚拟化	126
7.7.3	按照实现方式不同划分存储虚拟化	128
7.8	本章小结	130
第 8 章	Docker 容器	131
8.1	Docker 容器概述	131
8.1.1	Docker 容器的由来	131
8.1.2	Docker 定义	133
8.1.3	Docker 的优势	134
8.2	Docker 的原理	134
8.2.1	Linux Namespace(ns)	135
8.2.2	Control Groups(cgroups)	136
8.2.3	Linux 容器(LXC)	137
8.2.4	AUFS	137
8.2.5	Grsec	140
8.3	Docker 技术发展与应用	140
8.3.1	Docker 解决的问题	140
8.3.2	Docker 的未来发展	141
8.3.3	Docker 技术的局限	141
8.4	本章小结	141
第 9 章	Web 2.0	142
9.1	Web 2.0 产生背景和定义	142
9.1.1	Web 2.0 产生背景	142
9.1.2	Web 2.0 的概念	143
9.1.3	Web 2.0 和 Web 1.0 比较	143
9.1.4	Web 2.0 特征	145
9.2	Web 2.0 应用产品	145
9.2.1	Web 2.0 主要应用产品	146
9.2.2	主要产品的区别	147
9.3	Web 2.0 相关技术	148



- 9.3.1 Web 2.0 的设计模式 148
- 9.3.2 Web 标准 149
- 9.3.3 向 Web 标准过渡 151
- 9.4 本章小结 155

第 10 章 绿色数据中心 156

- 10.1 绿色数据中心概述 156
 - 10.1.1 云数据中心发展阶段 156
 - 10.1.2 绿色数据中心架构 157
 - 10.1.3 云数据中心需要整合的资源 158
- 10.2 数据中心管理和维护 158
 - 10.2.1 实现端到端、大容量、可视化的基础设施整合 159
 - 10.2.2 实现虚拟化、自动化的管理 159
 - 10.2.3 实现面向业务的应用管理和流量分析 160
- 10.3 本章小结 161

第 3 篇 云计算架构

第 11 章 基础设施即服务 165

- 11.1 IaaS 概述 165
 - 11.1.1 IaaS 的定义 165
 - 11.1.2 IaaS 提供服务的方法 166
 - 11.1.3 IaaS 云的特征 166
 - 11.1.4 IaaS 和虚拟化的关系 167
- 11.2 IaaS 技术架构 167
 - 11.2.1 资源层 167
 - 11.2.2 虚拟化层 168
 - 11.2.3 管理层 169
 - 11.2.4 服务层 170
- 11.3 IaaS 云计算管理 170
 - 11.3.1 自动化部署 170
 - 11.3.2 弹性能力提供技术 171
 - 11.3.3 资源监控 172
 - 11.3.4 资源调度 173
 - 11.3.5 业务管理和计费度量 174
- 11.4 Amazon 云计算案例 175
 - 11.4.1 概述 175
 - 11.4.2 Amazon S3 176
 - 11.4.3 Amazon Simple DB 177

11.4.4	Amazon RDS	178
11.4.5	Amazon SQS	178
11.4.6	Amazon EC2	179
11.5	本章小结	180
第 12 章	平台即服务	182
12.1	PaaS 概述	182
12.1.1	PaaS 的由来	182
12.1.2	PaaS 的概念	183
12.1.3	PaaS 模式的开发	183
12.1.4	PaaS 推进 SaaS 时代	185
12.2	PaaS 架构	186
12.2.1	PaaS 的功能	187
12.2.2	多租户弹性是 PaaS 的核心特性	187
12.2.3	PaaS 架构的核心意义	188
12.2.4	PaaS 改变未来软件开发和维护模式	190
12.3	Google 的云计算平台	191
12.3.1	设计理念	192
12.3.2	构成部分	192
12.3.3	App Engine 服务	193
12.4	Windows Azure 平台	194
12.4.1	Windows Azure 操作系统	194
12.4.2	SQL Azure	195
12.4.3	.NET 服务	196
12.4.4	Live 服务	196
12.4.5	Windows Azure Platform 的用途	197
12.5	本章小结	197
第 13 章	软件即服务	199
13.1	SaaS 概述	199
13.1.1	SaaS 的由来	199
13.1.2	SaaS 的概念	200
13.1.3	SaaS 与传统软件的区别	201
13.1.4	SaaS 模式应用于信息化优势	202
13.1.5	SaaS 成熟度模型	203
13.2	模式及实现	204
13.2.1	SaaS 商业模式	204
13.2.2	SaaS 平台架构	206
13.2.3	SaaS 服务平台的主要功能	207

- 13.2.4 SaaS 服务平台关键技术 210
- 13.3 Salesforce 云计算案例 213
 - 13.3.1 Salesforce 云计算产品组成 213
 - 13.3.2 Salesforce 云计算特点 214
- 13.4 本章小结 215
- 第 14 章 容器即服务 216**
 - 14.1 容器云服务 216
 - 14.1.1 云平台架构层次 216
 - 14.1.2 容器云 217
 - 14.1.3 容器云的特点 217
 - 14.2 Kubernetes 应用部署 220
 - 14.2.1 Kubernetes 架构 220
 - 14.2.2 Kubernetes 模型 221
 - 14.2.3 内部使用者的服务发现 222
 - 14.2.4 外部访问 Service 224
 - 14.3 Mesos 应用 225
 - 14.3.1 Mesos 体系结构和工作流 225
 - 14.3.2 Mesos 流程 226
 - 14.3.3 Mesos 资源分配 226
 - 14.3.4 Mesos 优势 227
 - 14.4 基于 Kubernetes 打造 SAE 容器云 228
 - 14.4.1 Kubernetes 的好处 228
 - 14.4.2 容器云网络 228
 - 14.4.3 容器云存储 230
 - 14.5 基于 Mesos 去哪儿网容器云 230
 - 14.5.1 背景 230
 - 14.5.2 应用 Mesos 构建容器云 231
 - 14.5.3 云环境构建 231
 - 14.6 本章小结 234

第 4 篇 大数据云架构实践与编程

- 第 15 章 大数据云架构搭建 237**
 - 15.1 分布式 Hadoop 与 Spark 集群搭建 237
 - 15.1.1 Hadoop 集群构建 237
 - 15.1.2 Spark 集群构建 250
 - 15.2 基于 Docker 大数据云架构 256
 - 15.2.1 简介 256

15.2.2	Docker 和 Weave 搭建	257
15.2.3	Hadoop 集群镜像搭建	257
15.2.4	集群部署与启动	261
15.2.5	基于 Ambari 管理平台的镜像搭建	263
15.2.6	桌面系统 XFCE 搭建	265
15.3	本章小结	266
第 16 章	Spark 大数据编程	267
16.1	Spark 应用开发环境配置	267
16.1.1	使用 IntelliJ 开发 Spark 程序	267
16.1.2	使用 Spark Shell 开发运行 Spark 程序	272
16.2	Spark 大数据编程	272
16.2.1	WordCount	272
16.2.2	股票趋势预测	274
16.3	本章小结	280
第 5 篇 大数据安全		
第 17 章	大数据云计算面临的安全威胁	283
17.1	大数据云计算的安全问题	283
17.1.1	大数据基础设施安全威胁	283
17.1.2	大数据存储安全威胁	284
17.1.3	大数据云架构网络安全威胁	286
17.1.4	大数据带来隐私问题	286
17.1.5	针对大数据的高级持续性攻击	287
17.1.6	其他安全威胁	288
17.2	不同领域大数据的安全需求	288
17.2.1	因特网行业	289
17.2.2	电信行业	289
17.2.3	金融行业	290
17.2.4	医疗行业	290
17.2.5	政府组织	291
17.3	大数据安全内涵	291
17.3.1	保障大数据安全	292
17.3.2	大数据用于安全领域	293
17.4	大数据安全研究方向	294
17.4.1	大数据安全保障技术	294
17.4.2	大数据安全应用技术	295
17.5	本章小结	296

第 18 章 保障大数据安全	297
18.1 大数据安全的关键技术	297
18.1.1 非关系数据库安全策略	297
18.1.2 防范 APT 攻击	299
18.2 大数据安全保障实践	304
18.2.1 大数据采集与存储的安全防护	305
18.2.2 大数据挖掘与应用的安全防护	306
18.2.3 大数据安全审计	307
18.2.4 大数据安全评估与安全管理	308
18.2.5 数据中心的安全保障	308
18.3 本章小结	310
第 19 章 应用大数据保障安全	311
19.1 大数据安全检测及应用	311
19.1.1 安全检测与大数据的融合	311
19.1.2 用户上网流量数据的挖掘与分析	312
19.2 安全大数据	313
19.2.1 数据挖掘方法	314
19.2.2 挖掘目标及评估	315
19.3 基于大数据的网络态势感知	315
19.3.1 态势感知定义	315
19.3.2 网络态势感知	316
19.3.3 基于流量数据的网络安全感知	316
19.3.4 基于大数据分析的网络优化	318
19.3.5 网络安全感知应用实践	319
19.4 视频监控数据的安全应用	320
19.4.1 视频监控数据的处理需求	320
19.4.2 视频监控数据挖掘技术	320
19.4.3 海量视频监控数据的分析与处理	322
19.5 本章小结	322
参考文献	324

第 1 篇 大数据云计算概论