

# 大数据 架构师指南

朱进云 陈 坚 王德政 编著



A GUIDE FOR  
**BIG DATA ARCHITECTS**



清华大学出版社

# 大数据 架构师指南

朱进云 陈 坚 王德政 编著

---

A GUIDE FOR  
**BIG DATA ARCHITECTS**

清华大学出版社  
北 京

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

大数据架构师指南/朱进云，陈坚，王德政 编著. —北京：清华大学出版社，2016  
ISBN 978-7-302-43516-7

I. ①大… II. ①朱… ②陈… ③王… III. ①数据处理—指南 IV. ①TP274-62

中国版本图书馆 CIP 数据核字(2016)第 079520 号

责任编辑：陈莉 高岫

封面设计：周晓亮

版式设计：方加青

责任校对：曹阳

责任印制：李红英

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](mailto:c-service@tup.tsinghua.edu.cn)

质 量 反 馈：010-62772015，[zhiliang@tup.tsinghua.edu.cn](mailto:zhiliang@tup.tsinghua.edu.cn)

印 装 者：清华大学印刷厂

经 销：全国新华书店

开 本：180mm×250mm 印 张：18.25 字 数：320 千字

版 次：2016 年 6 月第 1 版 印 次：2016 年 6 月第 1 次印刷

印 数：1~3500

定 价：45.00 元

---

产品编号：065794-01

## :::::本书编委会:::::

顾 问：赵先明

### 编委会成员：

朱进云 陈 坚 王德政 申山宏 张 强 汪绍飞  
牛家浩 刘少麟 洪 科 梁 平 薛清华 郭海生  
刘淑霞 关 涛 王利学 李 敏 周治中 管 云  
简 明 艾红芳 黄增建 郭进良 杨荣康

这是一个数据爆发的时代，宽带化、移动互联网、物联网、智能终端的普及与人工智能的兴起，促使全球数据每两年翻一番，预计2020年全球数据规模将达到44ZB，较2013年将增长10倍。有资料报告，2013年全球数据的来源基本上是消费者、企业与政府各贡献1/3。按照用户数计算，在中国，无论是互联网用户还是移动互联网用户，无论是固网宽带用户还是移动宽带用户，其规模都已经是全球第一，中国的数据拥有量的潜力为全球之冠。IDC公司曾经指出，2013年中国在全球数据占比13%，预计2020年将上升到18%。

拥有数据并不意味着坐拥金矿，数据的产生与存储还要付出成本代价，大数据只有通过数据分析与挖掘，发现知识和生成智慧才能创造价值。大数据挖掘的应用将总结事物发展规律，提升人类生产与管理活动的准确性，减少传统方式下的“试错”成本，进而提升社会的总生产效率。

大数据挖掘需要很多技术支持，反过来说也带动了海量存储、高效计算、深度学习、可视呈现等很多技术和学科的发展，它是当代信息技术的集中体现。大数据挖掘本身是产业，但其效益更多地反映在其应用到的社会管理和其他行业中，大数据之所以受重视正是因为它溢出效益明显，大数据将成为影响国家竞争力的重要因素。

美国、英国、欧盟、日本和韩国等国政府越来越重视大数据所产生的价值，鼓励使用大数据以推动社会进步，支持政府数据的公共资源化，并发布促进大数据技术发展的政策纲要。2015年中国国务院发布《促进大数据发展行动纲要》，提出了五大目标、三大任务、十大工程以及七项政策，在国家层面推动大数据的应用与落地。大数据挖掘应用正在引起各行各业的关注，成为“互联网+”行动的主要抓手，将发掘经济增长新动能。

大数据挖掘不仅需要技术，更需要人才，麦肯锡公司预测，到2018年美国对大数据深度分析人才的需求与实际可供给之间相差一倍以上。我国与发达国家相比

更缺乏深度分析人才，尤其是大数据架构师。高校承担了培养人才的责任，但更需在实践中锻炼，为加速大数据架构师的成长过程，实用经验的传承十分重要。

中兴通讯对大数据的知识与工程经验进行系统性的概述，正好契合了当前大数据挖掘应用的浪潮，弥补了此类书籍的空白，为促进大数据技术的发展与应用提供了宝贵的经验。

中国工程院院士  
中国互联网协会理事长



数据并不是一个新概念，几千年来我们一直在利用数据。但数据的价值，特别是大数据的价值，最近几年才成为公众关注的焦点，是有其时代背景的。

就如同石油在几千年前就被发现了，但是其用途一直是作为日常生活或战争中的燃料，并不是特别重要的战略物资。只有内燃机被发明后，石油才成为最重要的动力能源，在最近的一百年才成为战略物资。

数据也一样。传统的数据库技术，在数据处理的能力上都有很大的局限性，超过100T这个量级，要么是处理效率急剧降低，要么是系统成本上升到难以接受的昂贵程度。所以，在大数据时代之前，数据在生产系统中的使用目的往往是单一的、即时的。大量的历史数据与过程数据，按照当时的IT技术，既无法存储，更无法处理。那些被备份到磁带机上的数据，大部分都成为死亡的数据化石。

当前大数据处理的技术，特别是云存储与云计算技术的成熟应用，为大数据的存储与处理提供了技术可能性。企业可以利用生产系统以及管理系统中产生的大量数据，对海量的数据进行存储、挖掘分析。一方面可以对生产活动进行更为准确的预测与指导，从而提高企业生产活动的准确性；另一方面还可以通过对数据价值的挖掘，产生新的业务，帮助企业充分开发数据的价值。政府也可以利用大数据来提高管理水平和效率。

2014年Gartner发布的HypeCycle曲线中，大数据技术已经越过炒作顶点。从HypeCycle曲线来看，越过炒作顶点的技术，往往是已经满足技术可行性的技术。技术进步并辅以商业模式创新，大数据在部分细分市场已经具备商业可行性，可以为企业的现在与未来带来收益。

2015年8月国务院发布了《促进大数据发展行动纲要》，将大数据的应用与落地提升到国家层面。在这种背景下，当前大数据系统建设出现一波高潮。商业级的大数据系统建设周期长，复杂度高，资金投入量大，所以需要合理的系统架构以应对未来业务需求的变化。由于业界大数据系统的建设刚起步，当前阶段急需对相关的系统架构知识以及实际项目建设经验进行共享，提升业界的整体建设水平。

纵观当前业界大数据相关的书籍，偏重于两大类型。其一是偏重于大数据理念，描绘大数据前景，说明大数据可以有哪些应用；其二是偏重于大数据基础知识，偏重于实际的编程与开发。

但在大数据项目的实际建设过程中，架构师在进行端到端方案设计时，需要对大数据庞大的知识体系进行总揽性把握，并辅以实际项目的经验，才有可能把握此类系统的关键需求与要点。而此类知识与经验，业界分享较少，只能通过各类交流活动才能获取，不仅费时费力，而且还很难将这些知识系统化。

中兴通讯作为业界知名企业，在大数据研发上投入大量资源，并具备丰富的实际工程经验。本书不仅针对大数据知识进行系统化概述，并且将实际大型项目的经验进行总结。这种无私分享的宝贵经验，正是业界所亟需的，对大数据从业者具备较好的参考价值。相信本书分享的知识与经验，对推动大数据应用与落地起到积极的促进作用。

中兴通讯股份有限公司董事长兼总裁  
赵先明



毫无疑问，这是属于大数据的时代。随着移动互联网的进步、自媒体的风行和物联网的兴起，信息传播技术和信息传播渠道得到极大发展，海量级甚至银河级的数据不断涌现，呈现出“信息爆炸”的态势。这种情况下，似乎我们获取信息变得更加容易和方便；而实际上，由于对个体有用的信息淹没在浩如烟海的无关信息中，获取“有用信息”反而变得更加困难。

大数据相关技术就是在这种情况下应运而生的。作为一门新兴技术，大数据技术被人熟知和掌握需要一个过程；同时，由于其始终处于一个高速发展的过程，对其认识也是不断修正提高的过程。

鉴于此，本书总结了中兴通讯大数据平台DAP团队对大数据技术的最新研究成果，结合中兴大数据平台在各行业的应用实践经验，旨在帮助读者建立系统化的大数据技术脉络，并针对业界一些似是而非的问题进行系统性的讲解与澄清。阅读完本书，读者就可以基本掌握大数据技术的系统架构和核心思想。

## 为何要写这本书

在大数据项目建设过程中，往往需要三个层次的知识。第一个层次是关于大数据是什么，能做什么等理念方面的知识；第二个层次是如果去端到端进行大数据方案设计，要厘清大数据方案所需的关注重点，并结合具体的实践案例进行说明；第三个层次是大数据相关的基础技术知识，例如，对HDFS、MR、SPARK等技术点的掌握。

第一个层次的书籍，业界已经有很多，其中以《大数据时代》为典型代表；第三个层次的书籍，业界也比较多，读者不难获得相关的学习材料。

但第二个层次的书籍，属于承上启下的层次。该层次的知识需要从实践中总结出经验与知识。由于大型项目的建设周期长，建设复杂度高，涉及面广，所以从大型项目的实践中总结出知识有较高的难度。鉴于此，市面上该层次的大数据书籍相对较少，大数据相关的从业者或建设者较难获得这方面的知识，往往只能通过各类交流活动获取这方面的知识，不仅费时费力，而且难以将这些知识系统化。

基于如上原因，我们感觉迫切需要将我们在大型项目中积累的经验总结出来，供业界同仁参考，同时，这也可以满足我们内部人员学习大数据相关知识的需求。

## 本书读者对象

如果您是IT市场营销人员，或者是企业IT主管，您可以直接阅读本书的第一部分与第三部分。通过对本书第一部分与第三部分的阅读，将帮助您建立起大数据技术概念和框架。如果您对具体的大数据技术不感兴趣，可以忽略掉第二部分纯技术的内容。

如果您是大数据技术人员，本书将会是一本较好的参考资料，有助于帮助您超越自己所从事的具体模块，将您的大数据知识体系系统化。

如果您是高校大数据相关课程的老师，由于本书较为系统，可以考虑将本书作为参考书或者教材。

如果您是大数据技术爱好者，也可以将本书作为泛读书籍，让您理解当前大数据的时代。当然，读者如果能具备一定的IT基础知识，将能够更好地汲取本书中的知识。这不仅有助于您快速理解大数据相关知识，也有助于启发您对特定专题的深入思考和独到分析。

## 本书特色

本书是首本系统化的方案实践方面书籍，系统化地阐述了大数据方案应该如何思考，以及大数据的技术基础知识，并辅以实际的案例进行说明。

以客户化的语言，描述大数据项目建设中应该重点考虑的问题。即使不是技术专家，也能很容易地理解本书第一部分的内容。

较为系统地阐述了大数据相关的体系，可以帮助读者迅速系统化大数据相关的知识。

结合实际的案例，总结在大数据建设实践中的经验与知识。

## 如何阅读本书

本书内容分为四大部分，不同的读者可以选择不同的内容进行阅读。

本书第一部分是“大数据架构师入门”，以虚构角色小明的视角，去理解大数据，理解客户的烦恼，并提出构建一个大数据系统时应该从哪些方面考虑。阅读完该部分后，读者将对大数据方案具备一定的“提问题”的能力。也就是说，如果您面前有一份大数据的建设方案，即使您以前对大数据了解甚少，也可以根据本书第3章的建议，去评判方案的完整性，评判方案的深度与广度。

本书第二部分是“大数据架构师基础”，本部分将较为系统地介绍大数据相关

的基础知识。如图 I-1 所示，逐个介绍基础支撑层、计算存储层、中间件层、挖掘分析/应用层、展现层各部分内容，同时，对贯穿各层的安全和管理两大模块的相关内容做介绍，力图为读者呈现一个相对完整的大数据知识架构。

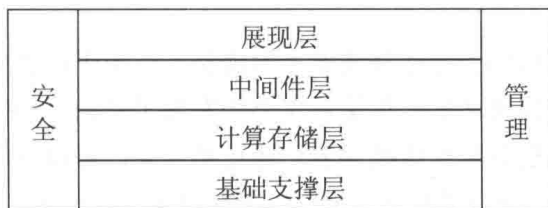


图 I-1 大数据技术框架

其中，计算存储层包括Hadoop架构、Spark架构、分析挖掘组件等内容；中间件层包括中间件的作用与意义，以及业界常用中间件及应用场景；展现层包括可视化相关的知识与内容；安全模块包括物理安全、主机安全、网络安全、数据安全等内容；管理模块包括自动部署、自动升级、自动巡检、自动维护等内容。

本书第三部分是“大数据架构师实践”，主要包括大数据开发实践中积累的一些经验，并结合案例进行阐述。这些实战中积累的知识与智慧，将帮助理论联系实际，更好地理解大数据技术。

本书第四部分是“大数据架构师拓展”，主要包括与大数据相关的其他技术。这些技术通常来说，并不属于大数据的技术范畴，但由于这些技术与大数据关系紧密，作为一名架构师，也需要系统地了解与思考这些相关的技术，才能对整个方案进行全局把握。该部分将试图对这些技术进行简单介绍，并试图说明这些技术与大数据之间的关系。

对于不需要关注具体技术的读者，则可以仅阅读第一部分“大数据架构师入门”；如果对具体的案例感兴趣，则可以阅读第三部分“大数据架构师实践”；如果是对技术感兴趣的读者，则可以阅读第二部分“大数据架构师基础”与第四部分“大数据架构师拓展”。

## 本书编写团队

大数据的知识非常广泛，不同层面的知识，以及不同技术模块的知识，很难由一个人完全掌握，所以本书是编写团队共同努力的成果。编写团队的成员都是在大数据领域担当重要工作岗位的技术骨干，大家在共同的理想与爱好下，聚集成一个团队，并为大数据架构师们完成了业界首本全面实践指导类的书籍。在此，请允许我列举参与编写的团队成员，并向他们致以诚挚的谢意。感谢他们牺牲周末与节假日的休息时间，为大家做的无私贡献。

团队成员包括：申山宏、梁平、薛清华、李敏、郭海生、杨荣康、牛家浩、刘

少麟、管云、洪科、简明、张强、艾红芳、关涛、刘淑霞、郭进良、汪绍飞、周治中、王利学、黄增建。

## 勘误与支持

尽管我们尽了各种努力来保证文章不出错误，但由于编者水平有限，加上编写时间仓促，难免会有错讹之处。如果你在书中发现了错误，例如错别字、书写错误等，请告诉我们，我们将整理成勘误表。通过勘误表，可以帮助其他读者节省阅读时间，提高阅读体验，并可以帮助我们提供更高质量的下一版。

错误反馈请发送至邮箱`zhou.zhizhong@zte.com.cn`，或者关注“中兴大数据”微信公众号(微信号ZTE\_BigData)并留言，我们将在第一时间确认反馈。勘误表可以在“中兴大数据”微信公众号上获取。

## 致谢

感谢中兴大数据平台DAP团队的所有成员，你们多年的潜心研究和积累是本书的基石。

感谢所有评审本书，并对本书提出过建议的朋友，你们的帮助对我们非常重要。

感谢关心本书的各界朋友，你们的关心与期望是我们的动力，更是对我们全心全意写好这本书的鞭策。

## 第一部分 大数据架构师入门

### 第1章 大数据概述 3

- 1.1 什么是大数据 4
- 1.2 大数据的本质 6
- 1.3 大数据技术当前状态 8
- 1.4 大数据的技术发展趋势 11

### 第2章 大数据项目常见场景 13

- 2.1 实验型部署场景 14
- 2.2 中小型部署场景 16
- 2.3 大型部署场景 19

### 第3章 大数据方案关键因素 23

- 3.1 数据存储规模与数据类型 24
- 3.2 数据来源与数据质量 25
- 3.3 业务特征 26
- 3.4 经济可行性 27
- 3.5 运维管理要求 28
- 3.6 安全性要求 29
- 3.7 部署要求 31

3.8 系统边界 32

3.9 约束条件 34

3.10 要点回顾 34

## 第二部分 大数据架构师基础

### 第4章 Hadoop基础组件 39

- 4.1 Hadoop简介 40
- 4.2 Hadoop版本演进 41
- 4.3 Hadoop2.0生态系统简介 42
- 4.4 Hadoop分布式文件系统  
HDFS 43
- 4.5 Hadoop统一资源管理框架  
YARN 48
- 4.6 Hadoop分布式计算框架  
MapReduce 52
- 4.7 Hadoop分布式集群管理系统  
ZooKeeper 57

### 第5章 Hadoop其他常用组件 61

- 5.1 Hadoop数据仓库工具Hive 62

5.2 Hadoop分布式数据库HBase 65

5.3 Hadoop实时流处理引擎  
Storm 70

5.4 Hadoop交互式查询引擎  
Impala 74

5.5 其他常用组件 78

## 第6章 Spark内存计算框架 83

6.1 内存计算与Spark 84

6.2 Spark的主要概念 86

6.3 Spark核心组件介绍 96

6.4 Spark与Hadoop之间的  
关系 100

6.5 要点回顾 104

## 第7章 大数据分析 105

7.1 数据时代 107

7.2 先进分析 109

7.3 架构与平台 112

7.4 数据分析流程 116

7.5 要点回顾 119

## 第8章 大数据中间件层 121

8.1 中间件层简介 122

8.2 中间件层产品介绍 123

8.3 中间件层的应用 137

8.4 中间件层的发展 140

8.5 要点回顾 144

## 第9章 可视化技术 145

9.1 可视化技术引言 146

9.2 什么是数据可视化 147

9.3 数据可视化设计 151

9.4 数据可视化的发展趋势 160

9.5 要点回顾 161

## 第10章 大数据安全 163

10.1 安全体系 164

10.2 大数据系统安全 168

10.3 要点回顾 180

## 第11章 大数据管理 181

11.1 数据管理的范围和定义 182

11.2 开源软件的管理能力 183

11.3 ZTE中兴大数据管理框架 187

11.4 大数据管理展望 192

11.5 要点回顾 192

## **第三部分 大数据架构师实践**

### 第12章 大数据项目实践 195

12.1 大数据项目架构关键  
步骤 197

12.2 架构师实践思考 209

### 第13章 大数据部署实践 213

13.1 中兴通讯DAP大数据平台  
功能和架构 214

13.2 DAP平台特点 215

13.3 某银行成功案例 216

## 第四部分 大数据架构师拓展

### 第14章 分布式系统与大数据的

#### 关系 225

- 14.1 分布式系统概述 226
- 14.2 分布式系统关键协议和算法  
概述 233
- 14.3 分布式系统和大数据 237

### 第15章 数据库系统与大数据的

#### 关系 241

- 15.1 数据库系统的历史 242
- 15.2 各类系统求同存异 254
- 15.3 数据库的发展展望 255

### 第16章 云计算与大数据的关系 257

- 16.1 虚拟化概述 258
- 16.2 OpenStack云管理架构  
实现 263
- 16.3 大数据基于云计算IAAS(包括  
Docker)部署的探讨 270

### 后记 273





第一部分

# 大数据架构师入门

---

A GUIDE FOR

**BIG DATA ARCHITECTS**

---



