大数据应用与技术丛书

Hadoop高级数据分析

使用Hadoop生态系统设计和构建大数据系统

Pro Hadoop Data Analytics: Designing and Building Big Data Systems Using the Hadoop Ecosystem

[美] Kerry Koitzsch 著
王建峰 王瑛琦 于金峰 译

清华大学出版社

大数据应用与技术丛书

Hadoop 高级数据分析

使用 Hadoop 生态系统设计和构建大数据系统

[美] Kerry Koitzsch 著

王建峰 王瑛琦 于金峰 译

清华大学出版社

北 京

Pro Hadoop Data Analytics：Designing and Building Big Data Systems Using the Hadoop Ecosystem
By Kerry Koitzsch
EISBN：978-1-4842-1909-6
Original English language edition published by Apress Media.

北京市版权局著作权合同登记号　图字：01-2017-5752

图书在版编目(CIP)数据

Hadoop 高级数据分析　使用 Hadoop 生态系统设计和构建大数据系统 / (美)克里·柯伊兹(Kerry Koitzsch) 著；王建峰，王瑛琦，于金峰　译. —北京：清华大学出版社，2018
(大数据应用与技术丛书)
书名原文：Pro Hadoop Data Analytics: Designing and Building Big Data Systems Using the Hadoop Ecosystem
ISBN 978-7-302-48730-2

Ⅰ. ①H…　Ⅱ. ①克…　②王…　③王…　④于…　Ⅲ. ①数据处理软件　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 271344 号

责任编辑：王　军　韩宏志
封面设计：孔祥峰
版式设计：思创景点
责任校对：牛艳敏
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
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：15.25　　字　　数：334 千字
版　　次：2018 年 1 月第 1 版　　印　　次：2018 年 1 月第 1 次印刷
印　　数：1～4000
定　　价：59.80 元

产品编号：075322-01

译者序

大数据类型多样、数量庞大、变化快速，这些特征对大数据分析师提出了新挑战。作为一种应对方案，大数据分析技术广泛应用于物联网、云计算等新兴领域，能够帮助企业用户在合理时间内处理海量数据，并为改善经营决策提供有效帮助。目前，存在多种大数据分析工具，相关技术正在不断走向成熟。Hadoop 作为一种优秀的开源框架，基于该架构的数据分析应用具有显著技术优势和应用前景，目前与 Hadoop 大数据分析相关的出版物中，大多偏重于理论和技术介绍，有关具体应用实践方面的书籍相对偏少。

为了满足应用需求，本书以设计并实现用于获取、分析、可视化大数据集的软件系统为目标，以应用案例为背景，系统地介绍利用 Hadoop 及其生态系统进行大数据分析的各种工具和方法；本书讲述 Hadoop 大数据分析的基本原理，呈现构建分析系统时所使用的标准架构、算法和技术，对应用案例进行了深入浅出的剖析，为读者掌握大数据分析基础架构及实施方法提供了详明实用的方案。

本书在注重 Hadoop 数据分析理论的同时，与大数据分析案例实践相结合，以生物、电信、资源勘查等行业真实案例为主线，详细讲解 Hadoop 高级数据分析的过程。使读者可以自己动手实践，亲自体会开发的乐趣及大数据分析的强大魅力。通过本书的学习，读者能够更加快速且有效地掌握 Hadoop 数据分析方法并积累实践经验。阅读本书，可以帮助读者了解并掌握 Hadoop 高级数据分析技术的具体操作方法，让读者真正理解其核心概念和基本原理。

在此要感谢清华大学出版社的编辑们，他们为本书的翻译投入了巨大的热情并付出了很多心血。没有你们的帮助和鼓励，本书不可能顺利付梓。

对于这本经典之作，译者本着以行业标准术语为翻译基础，以网络释义词典为补充的方法，在翻译过程中力求“信、达、雅”，但是鉴于译者水平有限、时间仓促，书中难免会出现一些错误和不当之处，恳请读者批评指正。

本书全部章节由王建峰、王瑛琦、于金峰翻译。参与本书翻译工作的还有博士研究生何鸣、张耘、陈田田等，硕士研究生赵新宇、李浩然等参与了本书的校对工作，在此一并致谢。

作者简介

Kerry Koitzsch 在计算机科学、图像处理和软件工程等领域拥有超过二十年的工作经验，致力于研究 Apache Hadoop 和 Apache Spark 技术。Kerry 擅长软件咨询，精通一些定制的大数据应用，包括分布式搜索、图像分析、立体视觉和智能图像检索系统。Kerry 目前就职于 Kildane 软件技术股份有限公司，该公司是加州桑尼维尔市的一个机器人系统和图像分析软件提供商。

技术审校者简介

Simin Boschma 在计算机工程设计方面拥有超过二十年的经验，曾从事程序设计和合作伙伴管理工作，也曾在硅谷、惠普、SanDisk 等高科技公司从事商业软硬件产品开发。另外，Simin 还拥有超过十年的技术撰写、审查及出版技术经验。Simin 目前就职于加州桑尼维尔市的 Kildane 软件技术股份有限公司。

致　谢

感谢编辑 Celestin Suresh John 和 Prachi Mehta，是他们给予了宝贵的帮助。没有他们，本书就无法顺利完成。同时感谢技术审校者 Simin Bochma 的专业协助。

前　言

Apache Hadoop 软件库逐渐受到重视。它是许多公司、政府机构、科研设施进行高级分布式开发的基础。Hadoop 生态系统现在包含几十个组件用于搜索引擎、数据库和数据仓库进行图像处理、深度学习及自然语言处理。随着 Hadoop2 的出现，不同的资源管理器可用于提供更高级别的复杂性和控制力。竞争对手、替代品以及 Hadoop 技术和架构的继承/变种比比皆是，包括 Apache Flink、Apache Spark 等。软件专家和评论员多次宣布“Hadoop 的死亡”。

我们必须正视一个问题：Hadoop 死了吗？这取决于 Hadoop 本身的感知界限。我们是否认为 Apache Spark 是 Hadoop 批处理文件方法的内存继承者，是 Hadoop 家族的一部分，仅仅因为 Apache Spark 也使用了 Hadoop 文件系统 HDFS？存在很多“灰色区域”的其他例子，其中较新的技术取代或增强了原有的“Hadoop 经典”功能。分布式计算是一个不断移动的目标，是 Hadoop 和 Hadoop 生态系统的分界线，在短短几年间已经发生了显著变化。在本书中，我们试图展示 Hadoop 及其相关生态系统的一些多样的、动态的方面，并试图说服你，尽管 Hadoop 发生变化，但它依然非常活跃、与当前的软件开发相关并且使数据分析程序员特别感兴趣。

目　录

第Ⅰ部分　概念

第 1 章　概述：用 Hadoop 构建数据分析系统……3
1.1　构建 DAS 的必要性……4
1.2　Hadoop Core 及其简史……4
1.3　Hadoop 生态系统概述……5
1.4　AI 技术、认知计算、深度学习以及 BDA……6
1.5　自然语言处理与 BDAS……6
1.6　SQL 与 NoSQL 查询处理……6
1.7　必要的数学知识……7
1.8　设计及构建 BDAS 的循环过程……7
1.9　如何利用 Hadoop 生态系统实现 BDA……10
1.10　“图像大数据”(IABD)基本思想……10
1.10.1　使用的编程语言……12
1.10.2　Hadoop 生态系统的多语言组件……12
1.10.3　Hadoop 生态系统架构……13
1.11　有关软件组合件与框架的注意事项……13
1.12　Apache Lucene、Solr 及其他：开源搜索组件……14
1.13　建立 BDAS 的架构……15
1.14　你需要了解的事情……15
1.15　数据可视化与报表……17
1.15.1　使用 Eclipse IDE 作为开发环境……18
1.15.2　本书未讲解的内容……19
1.16　本章小结……21

第 2 章　Scala 及 Python 进阶……23
2.1　动机：选择正确的语言定义应用……23
2.2　Scala 概览……24
2.3　Python 概览……29
2.4　错误诊断、调试、配置文件及文档……31
2.4.1　Python 的调试资源……32
2.4.2　Python 文档……33
2.4.3　Scala 的调试资源……33
2.5　编程应用与示例……33
2.6　本章小结……34
2.7　参考文献……34

第 3 章　Hadoop 及分析的标准工具集……35
3.1　库、组件及工具集：概览……35
3.2　在评估系统中使用深度学习方法……38
3.3　使用 Spring 框架及 Spring Data……44
3.4　数字与统计库：R、Weka 及其他……44
3.5　分布式系统的 OLAP 技术……44
3.6　用于分析的 Hadoop 工具集：Apache Mahout 及相关工具……45
3.7　Apache Mahout 的可视化……46
3.8　Apache Spark 库与组件……46
3.8.1　可供选择的不同类型的 shell……46
3.8.2　Apache Spark 数据流……47
3.8.3　Sparkling Water 与 H2O 机器学习……48

3.9　组件使用与系统建立示例……48
3.10　封包、测试和文档化示例系统……50
3.11　本章小结……51
3.12　参考文献……51

第 4 章　关系、NoSQL 及图数据库……53
4.1　图查询语言：Cypher 及 Gremlin……55
4.2　Cypher 示例……55
4.3　Gremlin 示例……56
4.4　图数据库：Apache Neo4J……58
4.5　关系数据库及 Hadoop 生态系统……59
4.6　Hadoop 以及 UA 组件……59
4.7　本章小结……63
4.8　参考文献……64

第 5 章　数据管道及其构建方法……65
5.1　基本数据管道……66
5.2　Apache Beam 简介……67
5.3　Apache Falcon 简介……68
5.4　数据源与数据接收：使用 Apache Tika 构建数据管道……68
5.5　计算与转换……70
5.6　结果可视化及报告……71
5.7　本章小结……74
5.8　参考文献……74

第 6 章　Hadoop、Lucene、Solr 与高级搜索技术……75
6.1　Lucene/Solr 生态系统简介……75
6.2　Lucene 查询语法……76
6.3　使用 Solr 的编程示例……79
6.4　使用 ELK 栈(Elasticsearch、Logstash、Kibana)……85
6.5　Solr 与 Elasticsearch：特点与逻辑……93
6.6　应用于 Elasticsearch 和 Solr 的 Spring Data 组件……95
6.7　使用 LingPipe 和 GATE 实现定制搜索……99
6.8　本章小结……108
6.9　参考文献……108

第 II 部分　架构及算法

第 7 章　分析技术及算法概览……111
7.1　算法类型综述……111
7.2　统计/数值技术……112
7.3　贝叶斯技术……113
7.4　本体驱动算法……114
7.5　混合算法：组合算法类型……115
7.6　代码示例……116
7.7　本章小结……119
7.8　参考文献……119

第 8 章　规则引擎、系统控制与系统编排……121
8.1　规则系统 JBoss Drools 介绍……121
8.2　基于规则的软件系统控制……124
8.3　系统协调与 JBoss Drools……125
8.4　分析引擎示例与规则控制……126
8.5　本章小结……129
8.6　参考文献……129

第 9 章　综合提升：设计一个完整的分析系统……131
9.1　本章小结……136
9.2　参考文献……136

第III部分　组件与系统

第 10 章　数据可视化：可视化与交互分析 …… 139
10.1　简单的可视化 …… 139
10.2　Angular JS 和 Friends 简介 …… 143
10.3　使用 JHipster 集成 Spring XD 和 Angular JS …… 143
10.4　使用 d3.js、sigma.js 及其他工具 …… 152
10.5　本章小结 …… 153
10.6　参考文献 …… 153

第IV部分　案例研究与应用

第 11 章　生物信息学案例研究：分析显微镜载玻片数据 …… 157
11.1　生物信息学介绍 …… 157
11.2　自动显微镜简介 …… 159
11.3　代码示例：使用图像填充 HDFS …… 162
11.4　本章小结 …… 165
11.5　参考文献 …… 165

第 12 章　贝叶斯分析组件：识别信用卡诈骗 …… 167
12.1　贝叶斯分析简介 …… 167
12.2　贝叶斯组件用于信用卡诈骗检测 …… 169
12.3　本章小结 …… 172
12.4　参考文献 …… 172

第 13 章　寻找石油：使用 Apache Mahout 分析地理数据 …… 173
13.1　基于领域的 Apache Mahout 推理介绍 …… 173
13.2　智能制图系统和 Hadoop 分析 …… 179
13.3　本章小结 …… 180
13.4　参考文献 …… 180

第 14 章　“图像大数据”系统：一些案例研究 …… 181
14.1　图像大数据简介 …… 181
14.2　使用 HIPI 系统的第一个代码示例 …… 184
14.3　BDA 图像工具包利用高级语言功能 …… 187
14.4　究竟什么是图像数据分析？ …… 187
14.5　交互模块和仪表板 …… 189
14.6　添加新的数据管道和分布式特征查找 …… 189
14.7　示例：分布式特征查找算法 …… 190
14.8　IABD 工具包中的低级图像处理程序 …… 194
14.9　术语 …… 194
14.10　本章小结 …… 195
14.11　参考文献 …… 195

第 15 章　构建通用数据管道 …… 199
15.1　示例系统的体系架构和描述 …… 199
15.2　如何获取和运行示例系统 …… 200
15.3　管道构建的五大策略 …… 200
15.3.1　从数据源和接收装置工作 …… 200

15.3.2　由中间向外发展……200
15.3.3　基于企业集成模式(EIP)的开发……200
15.3.4　基于规则的消息管道开发…201
15.3.5　控制+数据(控制流)管道……202
15.4　本章小结……202
15.5　参考文献……203

第 16 章　大数据分析的总结与展望……205
16.1　总结……205
16.2　大数据分析的现状……206
16.3　“孵化项目”和“初期项目”……208
16.4　未来 Hadoop 及其后续思考……209
16.5　不同观点：目前 Hadoop 的替代方案……211
16.6　在“未来 Hadoop”中使用机器学习和深度学习技术……211
16.7　数据可视化和 BDA 的前沿领域……212
16.8　结束语……212

附录 A　设置分布式分析环境……215

附录 B　获取、安装和运行示例分析系统……227

第Ⅰ部分　概念

本书第Ⅰ部分描述基本概念、结构、分布式分析软件系统的使用，以及该分布式系统的好处和使用它时的一些必要工具。同时介绍一些在建立系统时需要用到的分布式基础架构，包括 Apache Hadoop 及其生态系统。

第 1 章

概述：用 Hadoop 构建数据分析系统

本书将设计并实现用于获取、分析、可视化大数据集的软件系统。全书将使用缩略词 BDA 或 BDAS(Big Data Analytics System，大数据分析系统)描述此类软件。当然，首先需要对大数据本身进行解释。作为计算机程序员和架构师，我们知道现在通常所说的“大数据”已经伴随我们很长一段时间了——大约有十多年。事实上，因为“大数据”一直以来就是一个相对的、多维度的术语，并非仅仅根据数据容量进行定义。复杂性、速度和准确性——当然，也包含数据容量，构成了现代“大数据集合”的所有维度。

本章将讨论基于 Hadoop 的 BDAS 到底是什么，为什么它们非常重要，可以采用什么样的数据源、数据接收装置和仓库，以及哪些候选应用适合基于 Hadoop 的分布式系统方法，哪些应用不适合。我们还将简要讨论在构建此类系统时，能够替代 Hadoop/Spark 环境的其他环境。

软件开发总让人感到有种紧迫感，BDAS 的开发也不例外。即使是在这个蓬勃发展的新兴行业的初期，BDA 已经被要求以更快的速度处理和分析越来越多的数据，而且需要更深层次的理解能力。当我们考察软件系统构建和开发的具体细节时，无论对于抽象的计算机科学，还是对于计算机技术的应用来说，以更广泛的方式处理越来越多的数据始终是一个关键目标。同样，对于大数据应用和大数据系统来说，这条规则也不例外。这样当我们在思考可用的全局数据源为何在过去几年呈现爆炸式增长时，就不会感到奇怪，如图 1-1 所示。

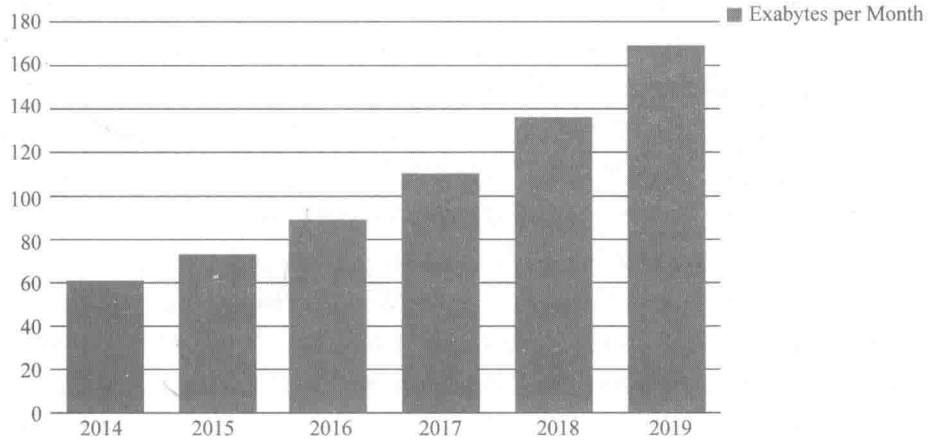

图 1-1　年度数据量统计(思科 VNI 全球 IP 流量预测 2014～2019)

由于软件组件和廉价现货处理能力的快速发展以及软件开发本身的快速发展，期望为其应用建立 BDA 的架构师和程序员对在 BDA 领域所面对的大量技术和策略选择问题常感到无所适从。本章将对 BDA 进行总体概述，并试图确定一些在构建 BDAS 时常常面对的技术问题。

1.1　构建 DAS 的必要性

由于传统的业务分析方法不能满足现代分析应用所面临的处理大容量、复杂性、多格式和快速数据的需求，因此 DAS(Distributed Analytical System，分布式分析系统)应运而生。DAS 环境除了软件以外，还以另外一种方式发生了戏剧性的变化。硬件开销——计算和存储开销大幅下降。类似 Hadoop 之类的工具应用于由相对低廉的机器和磁盘所构成的集群环境中，过去对大型数据项目来说必须具备的分布式处理已成为平常之事。同时，从实现分布式计算来看，目前存在大量的支持软件(框架、库、工具包)。的确，从技术栈(可选集)中选择可用技术已经成为一个严峻的问题，解决该问题的关键在于详细考察应用需求和可用资源。

从历史来看，硬件技术决定了软件组件的能力，在数据分析领域尤其如此。传统数据分析的主要工作针对基于文件的数据集或直接连接到关系数据库，实现统计的可视化(直方图、饼图、表格报告等)。计算引擎通常在单一服务器上采用批处理方式实现。随着分布式计算新时代的到来，利用计算机集群实现对大数据问题的分而治之成为计算的标准方式：其可扩展能力使得我们能够超越单台计算机的能力限制，尽可能多地增加所需(或者说我们能够负担得起)的硬件现货。类似 Ambari、Zookeeper 和 Curator 之类的软件工具帮助我们管理集群并提供可扩展能力，以及实现集群资源的高可用性。

1.2　Hadoop Core 及其简史

某些软件思想已经存在很长时间，以至于已经无法说它们是计算机的历史，而应当说它们是计算机的古董。“MapReduce(映射-规约)”问题求解方法可以追溯到第二古老的计算机编程语言 LISP(List Processing，列表处理)，可追溯到 20 世纪 50 年代，map、reduce、send 以及 lambda 是 Lisp 语言的标准函数。几十年后，我们现在所熟知的基于 Java 开源代码的分布式处理框架 Apache Hadoop 并非是“从头开始”的新东西。它源于 Apache Nutch，一种开源的 Web 搜索引擎，而 Nutch 则基于 Apache Lucene。有趣的是，R 统计库(本书后续章节将深入讨论)也受到 Lisp 的影响，最初是用 LISP 语言编写的。

在开始讨论 Hadoop 生态系统前，首先简单介绍一下 Hadoop Core 组件。顾名思义，Hadoop Core 是 Hadoop 框架的基础(见图 1-1)。支持组件、架构，当然还包括附属库、问题求解组件以及被称为 Hadoop 生态系统的子框架，它们都建立在 Hadoop Core 基础之上，如图 1-2 所示。请注意在本书中，我们将不会讨论 Hadoop 1，因为它已经被新的实现 YARN(另一种资源协调器)所取代。同时也请注意，在 Hadoop 2 系统中，MapReduce

并未消失，只是被模块化并抽象化为一种组件，以便能够更好地与其他数据处理模块协同工作。

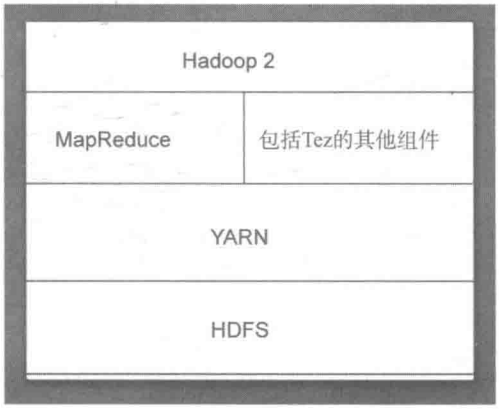

图 1-2　Hadoop 2 Core 图示

1.3　Hadoop 生态系统概述

Hadoop 及其生态系统加上随之不断壮大的框架和库，始终是 BDA 领域不容忽视的力量。本书其他部分将帮助读者对 BDA 所面临的挑战制定一个集中化解决方案，同时提供最低限度的背景和上下文，帮助读者学习在 BDA 求解中可以用到的新方法。Hadoop 及其生态系统通常可以划分为如图 1-3 所示的主要分类或功能块。读者将会注意到图中还包含几个额外的用于关联组件以及实现安全功能的模块。你也可以根据自己的需求为 BDAS 添加一些支持库和框架。

图 1-3　Hadoop 2 技术栈框图