



Hadoop

海量数据处理

技术详解与项目实战

第2版

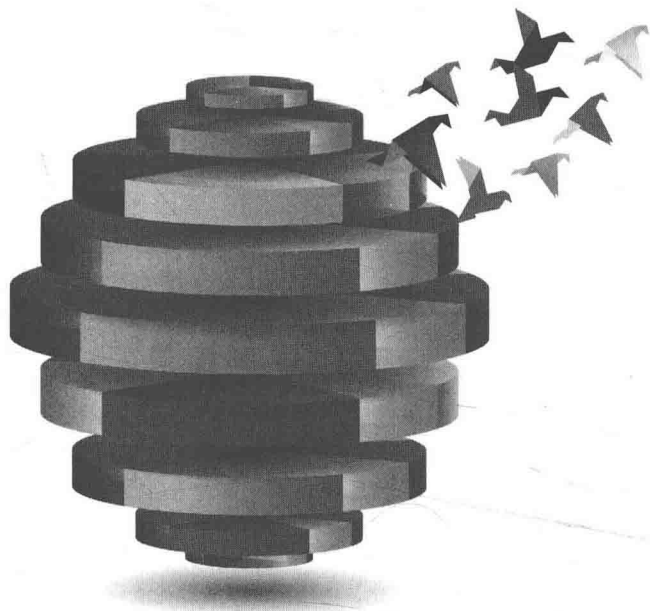
范东来 / 著



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS



Hadoop

海量数据处理

技术详解与项目实战

第2版

范东来 / 著

人民邮电出版社
北京

图书在版编目(CIP)数据

Hadoop海量数据处理：技术详解与项目实战 / 范东来著. — 2版. — 北京：人民邮电出版社，2016.8
ISBN 978-7-115-42746-5

I. ①H… II. ①范… III. ①数据处理软件 IV. ①TP274

中国版本图书馆CIP数据核字(2016)第151385号

内 容 提 要

本书介绍了 Hadoop 技术的相关知识，并将理论知识与实际项目相结合。全书共分为三个部分：基础篇、应用篇和总结篇。基础篇详细介绍了 Hadoop、YARN、MapReduce、HDFS、Hive、Sqoop 和 HBase，并深入探讨了 Hadoop 的运维和调优；应用篇则包含了一个具有代表性的完整的基于 Hadoop 的商业智能系统的设计和实现；结束篇对全书进行总结，并对技术发展做了展望。

本书结构针对学习曲线进行了优化，由浅至深，从理论基础到项目实战，适合 Hadoop 的初学者阅读，也适合作为高等院校相关课程的教学参考书。

-
- ◆ 著 范东来
 - 责任编辑 杨海玲
 - 责任印制 焦志炜
 - ◆ 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>
 - 北京鑫正大印刷有限公司印刷
 - ◆ 开本：800×1000 1/16
 - 印张：23.25
 - 字数：524 千字 2016 年 8 月第 2 版
 - 印数：3 801—6 800 2016 年 8 月北京第 1 次印刷
-

定价：59.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

第 2 版序

作为作者的老师，我很欣喜地看到作者的成长，正如这本书一样。

本书第 1 版问世后，得到了读者认可，市场反响也不错，而且其台湾繁体字版也于 2015 年问世。时间过得真快，我依然记得作者抱着第 1 版初稿到我办公室的那个下午，我们聊了很久。就一两年的时间，大数据技术已发生了巨大的变化，本书第 2 版的出版也就成了顺理成章的事。第 2 版根据最新技术做了全面修订，并新增了 YARN 和 HBase 的章节，更加全面和实用。

大数据本质上是一种思想，代表了数据的深度和广度。读者当然可以从本书学到 Hadoop 相关的技术，但这并不是最重要的，最重要的是能够用大数据的思想来思考和解决问题。

大数据天然地能够和几乎任意行业相结合，如金融、医疗、电商，但是这些都需要行业应用来支撑，需要各位读者来积极投身其中。中国的互联网土壤非常肥沃，这给了大数据非常好的发展基础。目前国内的大数据活跃程度丝毫不亚于国外，在大数据方面，中国很有机会弯道超车，成为世界一流。这是大数据最好的时代，充满了机遇与挑战。

但是，“路漫漫其修远兮”，希望这本书能够帮到更多的人。



北京软件行业协会执行会长
北京航空航天大学软件学院教授
2016 年 6 月，北京

第1版序

这是一本大数据工程师和 Hadoop 工程师的必备书。

近年来，由于移动互联网的高速发展和智能移动设备的普及，数据累积的速率已超过以往任何时候，这个世界已经进入了大数据时代。如何高效地存储、处理这些海量、多种类、高速流动的数据已成为亟待解决的问题。

Hadoop 最早来源于全球云计算技术的领导者谷歌在 2003 年至 2006 年间发表的三篇论文。得益于学术界和工业界的大力支持，Hadoop 目前已成为最为成熟的大数据处理技术。Hadoop 利用了“分而治之”的朴素思想为大数据处理提供了一整套新的解决方案，如分布式文件系统 HDFS、分布式计算框架 MapReduce、NoSQL 数据库 HBase、数据仓库工具 Hive 等。Hadoop 打破了传统数据处理技术的瓶颈，如样本容量、样本种类，让大数据真正成为了生产力。Hadoop 目前已广泛应用于各行各业，行业巨头也纷纷推出自己的基于 Hadoop 的解决方案。今天，Hadoop 已经在电信业、能源业等有了一定的用户基础，传统数据分析架构也逐渐在向 Hadoop 进行过渡。

大数据和大数据处理技术在相互促进，大数据刺激了大数据处理技术的发展，而大数据处理技术又加速了大数据应用落地。大数据催生了一批新的产业，并产生了对 Hadoop 工程师的庞大迫切需求，而目前有关 Hadoop 的书籍和在线材料仍然太少，这更进一步加大了人才缺口。

本书章节安排合理，结构清晰，内容由浅入深，循序渐进。作者是我的学生，作为一个奋战在大数据第一线的工程师，经验非常丰富，能够更加理解并贴近开发者和读者的需求。全书涵盖了 HDFS、MapReduce、Hive、Sqoop 等内容，尤其宝贵的是包含了大量动手实例和一个完备的 Hadoop 项目实例。我相信本书对于希望学习 Hadoop 的读者来说，是一个不错的选择。



北京软件行业协会执行会长
北京航空航天大学软件学院教授、院长
2014 年 12 月，北京

为什么要写这本书

2013 年被称为“大数据元年”，标志着世界正式进入了大数据时代，而就在这一年，我加入了清华大学苏州汽车研究院大数据处理中心，从事 Hadoop 的开发、运维和数据挖掘等方面的工作。从出现之日起，Hadoop 就深刻地改变了人们处理数据的方式。作为一款开源软件，Hadoop 能让所有人享受到大数据红利，让所有人在大数据时代站在了同一起跑线上。Hadoop 很好地诠释了什么是“大道至简，衍化至繁”，Hadoop 来源于非常朴素的思想，但是却衍生出大量的组件，让初学者难以上手。

我在学习和工作的过程中，走过很多弯路也做过很多无用功，尽管这是学习新技术的必由之路，但却浪费了大量的时间。我将自己学习和工作的心得记录下来，为了帮助更多像我当年一样的 Hadoop 学习者，我决定写一本书，一本自己开始 Hadoop 职业生涯的时候也想读到的书。

本书特点有哪些

本书结构针对学习曲线进行了优化，本书由浅至深，从理论基础到项目实战。

本书最大的特点是面向实践。基础篇介绍了 Hadoop 及相关组件，包含了大量动手实例，而应用篇则包含了一个具有代表性的基于 Hadoop 的项目完整实例，该实例脱胎于生产环境的真实项目，在通过基础篇的学习后，读者将在应用篇得到巩固和升华，并对 Hadoop 有一个更加清晰和完整的认识，这也符合实践出真知的规律。

本书介绍了 Hadoop 主要组件，如 HDFS、YARN、MapReduce、Hive、Sqoop、HBase 等，还介绍了 Hadoop 生产环境下的调优和运维、机器学习算法等高级主题。

读者对象是哪些

全书内容由浅入深，既适合初学者入门，也适合有一定基础的技术人员进一步提高技术水平，本书特别适合循序渐进地学习。本书的读者对象包括：

- 准备学习 Hadoop 的开发人员；
- 准备学习 Hadoop 的数据分析师；
- 希望将 Hadoop 运用到实际项目中的开发人员和管理人员；

- 计算机相关专业的高年级本科生和研究生；
- 具有一定的 Hadoop 使用经验，并想进一步提高的使用者。

为什么要写第 2 版

本书第 1 版于 2014 年下半年完稿，次年年初出版，当时 Hadoop 正值急速发展的时期，很多组件都没有实现自己的最终形态。到目前为止，Hadoop 离我完成本书第 1 版之时已有较大变化并已基本稳定。

很多读者在阅读过程中给我来信，说书中使用的 Hadoop 版本较老(CDH3)，而且没有 YARN 和 HBase 的内容。收到读者的反馈，加之 HBase 1.0 版已经发布，CDH5 已经基本普及，我开始本书第 2 版的写作工作，于是呈现在读者面前的就是全新的《Hadoop 海量数据处理》。第 2 版主要做了以下修改和内容增补：

- 全书所有内容根据最新技术进行了修改，并修订了全书；
- 新增了 HDFS 新特性；
- 新增了关于 YARN 的一章；
- 新增了关于 HBase 的一章；
- 新增了 HBase 调优内容；
- 在项目实战中新增了有关 HBase 的内容；
- 根据最新趋势重写了技术展望的内容。

如何阅读本书

本书在章节的安排上旨在引导读者以最快的速度上手 Hadoop，而省去其他不必要的学习过程。如果你是一个有经验的 Hadoop 工程师或者是项目经理，也可以直接进入应用篇，关注项目的设计和实现；如果不是，还是建议你循序渐进地阅读本书方能获得最好的学习效果。

本书一共分为基础篇、应用篇和结束篇 3 个部分，一共 18 章。

基础篇从第 1 章至第 9 章，其中第 1 章为绪论，第 2 章为环境准备，第 3 章至第 8 章主要介绍 HDFS、YARN、MapReduce、Hive、Sqoop、HBase 的原理和使用，在此之上，第 9 章介绍 Hadoop 的性能调优和运维。读者将从基础篇获得 Hadoop 工程师的理论基础。

应用篇从第 9 章至 19 章，主要内容为一个基于 Hadoop 的在线图书销售商业智能系统的设计和实现，包含了系统需求说明、总体设计和完整的实现。应用篇会运用基础篇的知识，巩固并升华基础篇的学习效果。此外，应用篇的项目架构可以进行一些改动并推而广之，有一定的参考价值。读者将从应用篇获得 Hadoop 工程师的项目经验。

结束篇为第 20 章，将对全书进行总结，并对技术发展做了展望。

勘误和支持

写书就像是跳水，高高跳起跃入水中，但在浮出水面之前，运动员却无法知道评委的给分，而我期待读者的评价。由于作者水平有限，编写时间仓促，书中难免会出现一些错误，恳请读者批评指正。读者可以将对本书的反馈和疑问发到 ddna_1022@163.com，我将尽力为读者提供满意的回复。

致谢

感谢电子科技大学的赵勇教授、北京航空航天大学的孙伟教授和邵兵副教授，从您们身上我学到了严谨的学术精神和做人的道理。

感谢清华大学苏州汽车研究院大数据处理中心的林辉主任，您的锐意进取精神一直深留我心。

感谢周俊琨、肖宇、赵虎、李为、黄普、朱游强、熊荣、江彦平，没有你们的帮助和努力，本书不可能完成。

感谢我的父母和外婆这些年来在生活上对我无微不至的关怀和无时无刻的支持，你们辛苦了；感谢吴静宜和她的家人对我的支持；感谢范若云哥哥，是你改变了我。

感谢人民邮电出版社的杨海玲编辑在本书出版过程中给予我的指导和一如既往的信任，感谢庞燕博士为审阅本书第1版付出的辛勤劳动。

感谢所有在我求学路上帮助过我的人。

范东来

2016年6月于成都

目 录

基础篇：Hadoop 基础

第 1 章 绪论	2	2.1.3 Hadoop 的版本	23
1.1 Hadoop 和云计算	2	2.1.4 如何选择 Hadoop 的版本	25
1.1.1 Hadoop 的电梯演讲	2	2.2 Hadoop 架构	26
1.1.2 Hadoop 生态圈	3	2.2.1 Hadoop HDFS 架构	27
1.1.3 云计算的定义	6	2.2.2 YARN 架构	28
1.1.4 云计算的类型	7	2.2.3 Hadoop 架构	28
1.1.5 Hadoop 和云计算	8	2.3 安装 Hadoop	29
1.2 Hadoop 和大数据	9	2.3.1 安装运行环境	30
1.2.1 大数据的定义	9	2.3.2 修改主机名和用户名	36
1.2.2 大数据的结构类型	10	2.3.3 配置静态 IP 地址	36
1.2.3 大数据行业应用实例	12	2.3.4 配置 SSH 无密码连接	37
1.2.4 Hadoop 和大数据	13	2.3.5 安装 JDK	38
1.2.5 其他大数据处理平台	14	2.3.6 配置 Hadoop	39
1.3 数据挖掘和商业智能	15	2.3.7 格式化 HDFS	42
1.3.1 数据挖掘的定义	15	2.3.8 启动 Hadoop 并验证安装	42
1.3.2 数据仓库	17	2.4 安装 Hive	43
1.3.3 操作数据库系统和数据仓库 系统的区别	18	2.4.1 安装元数据库	44
1.3.4 为什么需要分离的数据仓库	19	2.4.2 修改 Hive 配置文件	44
1.3.5 商业智能	19	2.4.3 验证安装	45
1.3.6 大数据时代的商业智能	20	2.5 安装 HBase	46
1.4 小结	21	2.5.1 解压文件并修改 Zookeeper 相关配置	46
第 2 章 环境准备	22	2.5.2 配置节点	46
2.1 Hadoop 的发行版本选择	22	2.5.3 配置环境变量	47
2.1.1 Apache Hadoop	22	2.5.4 启动并验证	47
2.1.2 CDH	22	2.6 安装 Sqoop	47
		2.7 Cloudera Manager	48
		2.8 小结	51

第3章 Hadoop 的基石: HDFS.....	52	4.5 YARN 的调度器	89
3.1 认识 HDFS.....	52	4.5.1 YARN 的资源管理机制.....	89
3.1.1 HDFS 的设计理念.....	54	4.5.2 FIFO Scheduler	90
3.1.2 HDFS 的架构	54	4.5.3 Capacity Scheduler	90
3.1.3 HDFS 容错	58	4.5.4 Fair Scheduler	91
3.2 HDFS 读取文件和写入文件.....	58	4.6 YARN 命令行	92
3.2.1 块的分布	59	4.7 Apache Mesos	95
3.2.2 数据读取	60	4.8 小结	96
3.2.3 写入数据	61	第5章 分而治之的智慧: MapReduce... 97	
3.2.4 数据完整性.....	62	5.1 认识 MapReduce	97
3.3 如何访问 HDFS	63	5.1.1 MapReduce 的编程思想	98
3.3.1 命令行接口.....	63	5.1.2 MapReduce 运行环境	100
3.3.2 Java API	66	5.1.3 MapReduce 作业和任务	102
3.3.3 其他常用的接口	75	5.1.4 MapReduce 的计算资源划分 ..	102
3.3.4 Web UI	75	5.1.5 MapReduce 的局限性	103
3.4 HDFS 中的新特性	76	5.2 Hello Word Count	104
3.4.1 NameNode HA	76	5.2.1 Word Count 的设计思路	104
3.4.2 NameNode Federation	78	5.2.2 编写 Word Count	105
3.4.3 HDFS Snapshots	79	5.2.3 运行程序.....	107
3.5 小结	79	5.2.4 还能更快吗.....	109
第4章 YARN: 统一资源管理和调度平台	80	5.3 MapReduce 的过程	109
4.1 YARN 是什么	80	5.3.1 从输入到输出	109
4.2 统一资源管理和调度平台范型	81	5.3.2 input	110
4.2.1 集中式调度器.....	81	5.3.3 map 及中间结果的输出	112
4.2.2 双层调度器.....	81	5.3.4 shuffle	113
4.2.3 状态共享调度器	82	5.3.5 reduce 及最后结果的输出	115
4.3 YARN 的架构	82	5.3.6 sort	115
4.3.1 ResourceManager	83	5.3.7 作业的进度组成	116
4.3.2 NodeManager	85	5.4 MapReduce 的工作机制	116
4.3.3 ApplicationMaster	87	5.4.1 作业提交	117
4.3.4 YARN 的资源表示模型		5.4.2 作业初始化	118
Container	87	5.4.3 任务分配	118
4.4 YARN 的工作流程	88	5.4.4 任务执行	118
		5.4.5 任务完成	118
		5.4.6 推测执行	119

5.4.7 MapReduce 容错	119	6.2.4 数据格式	151
5.5 MapReduce 编程	120	6.3 HQL: 数据定义	152
5.5.1 Writable 类	120	6.3.1 Hive 中的数据库	152
5.5.2 编写 Writable 类	123	6.3.2 Hive 中的表	154
5.5.3 编写 Mapper 类	124	6.3.3 创建表	154
5.5.4 编写 Reducer 类	125	6.3.4 管理表	156
5.5.5 控制 shuffle	126	6.3.5 外部表	156
5.5.6 控制 sort	128	6.3.6 分区表	156
5.5.7 编写 main 函数	129	6.3.7 删除表	158
5.6 MapReduce 编程实例: 连接	130	6.3.8 修改表	158
5.6.1 设计思路	131	6.4 HQL: 数据操作	159
5.6.2 编写 Mapper 类	131	6.4.1 装载数据	159
5.6.3 编写 Reducer 类	132	6.4.2 通过查询语句向表中插入数据	160
5.6.4 编写 main 函数	133	6.4.3 利用动态分区向表中插入数据	160
5.7 MapReduce 编程实例: 二次排序	134	6.4.4 通过 CTAS 加载数据	161
5.7.1 设计思路	134	6.4.5 导出数据	161
5.7.2 编写 Mapper 类	135	6.5 HQL: 数据查询	162
5.7.3 编写 Partitioner 类	136	6.5.1 SELECT...FROM 语句	162
5.7.4 编写 SortComparator 类	136	6.5.2 WHERE 语句	163
5.7.5 编写 Reducer 类	137	6.5.3 GROUP BY 和 HAVING 语句	164
5.7.6 编写 main 函数	137	6.5.4 JOIN 语句	164
5.8 MapReduce 编程实例: 全排序	139	6.5.5 ORDER BY 和 SORT BY 语句	166
5.8.1 设计思路	139	6.5.6 DISTRIBUTE BY 和 SORT BY 语句	167
5.8.2 编写代码	140	6.5.7 CLUSTER BY	167
5.9 小结	141	6.5.8 分桶和抽样	168
第 6 章 SQL on Hadoop: Hive	142	6.5.9 UNION ALL	168
6.1 认识 Hive	142	6.6 Hive 函数	168
6.1.1 从 MapReduce 到 SQL	143	6.6.1 标准函数	168
6.1.2 Hive 架构	144	6.6.2 聚合函数	168
6.1.3 Hive 与关系型数据库的区别	146	6.6.3 表生成函数	169
6.1.4 Hive 命令的使用	147	6.7 Hive 用户自定义函数	169
6.2 数据类型和存储格式	149		
6.2.1 基本数据类型	149		
6.2.2 复杂数据类型	149		
6.2.3 存储格式	150		

6.7.1	UDF	169	8.4	HBase 的架构模式	193
6.7.2	UDAF	170	8.4.1	行键、列族、列和单元格	193
6.7.3	UDTF	171	8.4.2	HMaster	194
6.7.4	运行	173	8.4.3	Region 和 RegionServer	195
6.8	小结	173	8.4.4	WAL	195
第 7 章	SQL to Hadoop : Sqoop	174	8.4.5	HFile	195
7.1	一个 Sqoop 示例	174	8.4.6	Zookeeper	197
7.2	导入过程	176	8.4.7	HBase 架构	197
7.3	导出过程	178	8.5	HBase 写入和读取数据	198
7.4	Sqoop 的使用	179	8.5.1	Region 定位	198
7.4.1	codegen	180	8.5.2	HBase 写入数据	199
7.4.2	create-hive-table	180	8.5.3	HBase 读取数据	199
7.4.3	eval	181	8.6	HBase 基础 API	200
7.4.4	export	181	8.6.1	创建表	201
7.4.5	help	182	8.6.2	插入	202
7.4.6	import	182	8.6.3	读取	203
7.4.7	import-all-tables	183	8.6.4	扫描	204
7.4.8	job	184	8.6.5	删除单元格	206
7.4.9	list-databases	184	8.6.6	删除表	207
7.4.10	list-tables	184	8.7	HBase 高级 API	207
7.4.11	merge	184	8.7.1	过滤器	208
7.4.12	metastore	185	8.7.2	计数器	208
7.4.13	version	186	8.7.3	协处理器	209
7.5	小结	186	8.8	小结	214
第 8 章	HBase:HadoopDatabase	187	第 9 章	Hadoop 性能调优和运维	215
8.1	酸和碱：两种数据库事务方法论	187	9.1	Hadoop 客户端	215
8.1.1	ACID	188	9.2	Hadoop 性能调优	216
8.1.2	BASE	188	9.2.1	选择合适的硬件	216
8.2	CAP 定理	188	9.2.2	操作系统调优	218
8.3	NoSQL 的架构模式	189	9.2.3	JVM 调优	219
8.3.1	键值存储	189	9.2.4	Hadoop 参数调优	219
8.3.2	图存储	190	9.3	Hive 性能调优	225
8.3.3	列族存储	191	9.3.1	JOIN 优化	226
8.3.4	文档存储	192	9.3.2	Reducer 的数量	226
			9.3.3	列裁剪	226

9.3.4 分区裁剪	226	9.4.2 客户端调优	230
9.3.5 GROUP BY 优化	226	9.4.3 写调优	231
9.3.6 合并小文件	227	9.4.4 读调优	231
9.3.7 MULTI-GROUP BY 和 MULTI-INSERT	228	9.4.5 表设计调优	232
9.3.8 利用 UNION ALL 特性	228	9.5 Hadoop 运维	232
9.3.9 并行执行	228	9.5.1 集群节点动态扩容和卸载	233
9.3.10 全排序	228	9.5.2 利用 SecondaryNameNode 恢复 NameNode	234
9.3.11 Top N	229	9.5.3 常见的运维技巧	234
9.4 HBase 调优	229	9.5.4 常见的异常处理	235
9.4.1 通用调优	229	9.6 小结	236

应用篇：商业智能系统项目实战

第 10 章 在线图书销售商业智能系统

10.1 项目背景	238
10.2 功能需求	239
10.3 非功能需求	240
10.4 小结	240

第 11 章 系统结构设计

11.1 系统架构	241
11.2 功能设计	242
11.3 数据仓库结构	243
11.4 系统网络拓扑与硬件选型	246
11.4.1 系统网络拓扑	246
11.4.2 系统硬件选型	248
11.5 技术选型	249
11.5.1 平台选型	249
11.5.2 系统开发语言选型	249
11.6 小结	249

第 12 章 在开发之前

12.1 新建一个工程	250
12.1.1 安装 Python	250
12.1.2 安装 PyDev 插件	251

12.1.3 新建 PyDev 项目

12.2 代码目录结构	253
12.3 项目的环境变量	253
12.4 如何调试	254
12.5 小结	254

第 13 章 实现数据导入导出模块

13.1 处理流程	255
13.2 导入方式	256
13.2.1 全量导入	256
13.2.2 增量导入	256
13.3 读取配置文件	257
13.4 SqoopUtil	261
13.5 整合	262
13.6 导入说明	262
13.7 导出模块	263
13.8 小结	265

第 14 章 实现数据分析工具模块

14.1 处理流程	266
14.2 读取配置文件	266
14.3 HiveUtil	268
14.4 整合	268

14.5 数据分析和报表	269	16.5.1 网站分析的指标	300
14.5.1 OLAP 和 Hive	269	16.5.2 网站分析的决策支持	301
14.5.2 OLAP 和多维模型	270	16.6 小结	301
14.5.3 选 MySQL 还是选 HBase	272	第 17 章 实现购书转化率分析模块	302
14.6 小结	273	17.1 漏斗模型	302
第 15 章 实现业务数据的数据清洗 模块	274	17.2 处理流程	303
15.1 ETL	274	17.3 读取配置文件	303
15.1.1 数据抽取	274	17.4 提取所需数据	304
15.1.2 数据转换	274	17.5 编写转化率分析 MapReduce 作业	305
15.1.3 数据清洗工具	275	17.5.1 编写 Mapper 类	306
15.2 处理流程	275	17.5.2 编写 Partitioner 类	308
15.3 数据去重	276	17.5.3 编写 SortComparator 类	309
15.3.1 产生原因	276	17.5.4 编写 Reducer 类	310
15.3.2 去重方法	277	17.5.5 编写 Driver 类	312
15.3.3 一个很有用的 UDF: RowNum	277	17.5.6 通过 Python 模块调用 jar 文件	314
15.3.4 第二种去重方法	279	17.6 对中间结果进行汇总得到最终 结果	314
15.3.5 进行去重	279	17.7 整合	316
15.4 小结	282	17.8 小结	316
第 16 章 实现点击流日志的数据清洗 模块	283	第 18 章 实现购书用户聚类模块	317
16.1 数据仓库和 Web	283	18.1 物以类聚	317
16.2 处理流程	285	18.2 聚类算法	318
16.3 字段的获取	285	18.2.1 k -means 算法	318
16.4 编写 MapReduce 作业	288	18.2.2 Canopy 算法	319
16.4.1 编写 IP 地址解析器	288	18.2.3 数据向量化	320
16.4.2 编写 Mapper 类	291	18.2.4 数据归一化	321
16.4.3 编写 Partitioner 类	295	18.2.5 相似性度量	322
16.4.4 编写 SortComparator 类	295	18.3 用 MapReduce 实现聚类算法	323
16.4.5 编写 Reducer 类	297	18.3.1 Canopy 算法与 MapReduce	323
16.4.6 编写 main 函数	298	18.3.2 k -means 算法与 MapReduce	323
16.4.7 通过 Python 调用 jar 文件	299	18.3.3 Apache Mahout	324
16.5 还能做什么	300	18.4 处理流程	324

18.5 提取数据并做归一化.....	325	18.9.1 一份不适合聚类的数据.....	337
18.6 维度相关性.....	327	18.9.2 簇间距离和簇内距离.....	337
18.6.1 维度的选取.....	327	18.9.3 计算平均簇间距离.....	338
18.6.2 相关系数与相关系数矩阵.....	328	18.10 小结.....	339
18.6.3 计算相关系数矩阵.....	328		
18.7 使用 Mahout 完成聚类.....	329	第 19 章 实现调度模块.....	340
18.7.1 使用 Mahout.....	329	19.1 工作流.....	340
18.7.2 解析 Mahout 的输出.....	332	19.2 编写代码.....	341
18.7.3 得到聚类结果.....	334	19.3 crontab.....	342
18.8 得到最终结果.....	335	19.4 让数据说话.....	343
18.9 评估聚类结果.....	337	19.5 小结.....	344

结束篇：总结和展望

第 20 章 总结和展望.....	346	20.4 Pregel 系技术.....	349
20.1 总结.....	346	20.5 Docker 和 Kubernetes.....	350
20.2 BDAS.....	347	20.6 数据集成工具 NiFi.....	350
20.3 Dremel 系技术.....	348	20.7 小结.....	351
参考文献.....	352		

基础篇：Hadoop 基础

本书的第一部分相当于工具的使用手册，将会介绍 Hadoop 的核心组件：HDFS、YARN、MapReduce、Hive、Sqoop 和 HBase，并在此基础上，进一步学习 Hadoop 性能调优和运维。通过这部分的学习，读者将获得 Hadoop 工程师的理论基础。

第 1 章

绪论

这是最好的时代，这是最坏的时代；这是智慧的时代，这是愚蠢的时代；这是信仰的时期，这是怀疑的时期；这是光明的季节，这是黑暗的季节；这是希望之春，这是失望之冬……

——狄更斯《双城记》

本章作为绪论，目的是在学习 Hadoop 之前，让读者理清相关概念以及这些概念之间的联系。

1.1 Hadoop 和云计算

Hadoop 从问世之日起，就和云计算有着千丝万缕的联系。本节将在介绍 Hadoop 的同时，介绍 Hadoop 和云计算之间的关系，为后面的学习打下基础。

1.1.1 Hadoop 的电梯演讲

如果你是一名创业者或者是一名项目经理，那么最好准备一份“电梯演讲”。所谓电梯演讲，是对自己产品的简单介绍，通常都是 1~2 分钟（电梯从 1 层~30 层的时间），以便如果你恰巧和投资人挤上同一部电梯的时候，能够说服他投资你的项目或者产品。

在做 Hadoop 的电梯演讲之前，先来恶补一下 Hadoop 的有关知识。来看看 Hadoop 的发布者 Apache 软件基金会（ASF）对 Hadoop 的定义：Hadoop 软件库是一个框架，允许在集群中使用简单的编程模型对大规模数据集进行分布式计算。它被设计为可以从单一服务器扩展到数以千计的本地计算和存储的节点，并且 Hadoop 会在应用层面监测和处理错误，而不依靠硬件的高可用性，所以 Hadoop 能够在一个每个节点都有可能出错的集群之上提供一个高可用服务。

从上面的定义可以看出 Hadoop 的如下几个特点。