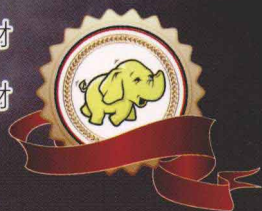


NITE 国家信息技术紧缺人才培养工程（移动云计算方向）系列教材

Uniquedu 工业和信息化部CSIP移动云计算教育培训中心官方教材



# 实战Hadoop

## ——开启通向云计算的捷径

- 云计算核心研发团队剖析Hadoop：  
**怎么装？ 怎么编程？ 怎么解决实际问题？**
- HDFS、MapReduce、HBase、Hive、Pig、Cassandra、Chukwa和ZooKeeper全覆盖

刘 鹏 主 编

黄宜华 副主编  
陈卫卫

# 实战 Hadoop

——开启通向云计算的捷径

刘 鹏 主 编

黄宜华  
陈卫卫 副主编

電子工業出版社

**Publishing House of Electronics Industry**

北京 · BEIJING

## 内 容 简 介

作为谷歌云计算基础架构的模仿实现，Hadoop 堪称业界最经典的开源云计算平台软件。本书是原著的 Hadoop 编程技术书籍，是云计算专家刘鹏教授继《云计算》教材取得成功功后，再次组织团队精心编写的又一力作，其作者均来自拥有丰富实践经验的云计算技术研发和教学团队。

该书强动手、强调实战，以风趣幽默的语言和一系列生动的实战应用案例，系统地讲授了 Hadoop 的核心技术和扩展技术，包括：HDFS、MapReduce、HBase、Hive、Pig、Cassandra、Chukwa 和 ZooKeeper 等，并给出了 3 个完整的 Hadoop 云计算综合应用实例，最后介绍了保障 Hadoop 平台可靠性的方法。

本书读者对象为各类云计算相关企业、高校和科研机构的研发人员，亦适合作为高校研究生和本科生教材。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。  
版权所有，侵权必究。

### 图书在版编目（CIP）数据

实战 Hadoop：开启通向云计算的捷径 / 刘鹏主编. —北京：电子工业出版社，2011.9  
ISBN 978-7-121-14475-2

I. ①实… II. ①刘… III. ①数据处理—应用软件—网络编程 IV. ①TP274

中国版本图书馆 CIP 数据核字（2011）第 174992 号

责任编辑：董亚峰 特约编辑：史 涛

印 刷：三河市鑫金马印装有限公司

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本：720×1 000 1/16 印张：29.5 字数：758 千字

印 次：2011 年 10 月第 2 次印刷

印 数：4 001~6 000 册 定价：59.00 元



凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 [zltts@phei.com.cn](mailto:zltts@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

服务热线：（010）88258888。

# 编 写 组

- 主 编：刘 鹏 解放军理工大学 教授
- 副主编：黄宜华 南京大学 教授
- 陈卫卫 解放军理工大学 教授
- 编 委：程 浩 华南理工大学
- 王 磊 中国矿业大学、云创存储
- 顾 荣 南京大学、云创存储
- 张 贞 华东交通大学
- 邓 鹏 湘潭大学、云创存储
- 杨晓亮 南京大学、云创存储
- 郭岩岩 云创存储
- 李 浩 云创存储
- 魏家宾 南京邮电大学、云创存储
- 王胤然 南京航空航天大学、云创存储
- 张 欣 江苏科技大学
- 王海坤 云创存储



南京云创存储科技有限公司  
Nanjing Cloud Storage Technology Co., Ltd.



## 前 言

1998 年，斯坦福大学的博士生拉里·佩奇和谢尔盖·布林在车库里创建了 Google 公司。2001 年，Google 已经索引了近 30 亿个网页。2004 年，Google 发布 Gmail，提供闻所未闻的 1GB 免费邮箱——众人还以为这是个愚人节玩笑。紧接着，Google 又发布了 Google Maps 和被称为“上帝之眼”的 Google Earth……

目前，google.com 为全世界访问量最高的站点。Google 在全球部署了约 200 多万台服务器，每天处理数以亿计的搜索请求和用户生成的约 24PB 数据，而且这些数据还在不断迅速增长。同时，Google 的 Android 智能手机操作系统已经拥有超过 40% 的美国智能手机用户，而苹果仅以 8.9% 的市场份额排名第四。社交服务 Google+ 推出不到半月，用户数量就突破 1000 万，其增长速度罕见。数辆 Google 无人驾驶汽车已经安全行驶了至少 22.5 万公里，没有发生过任何意外。Google 机器翻译服务能够实现 60 多种语言中任意两种语言间的互译……

是什么技术造就了这家让人惊叹的公司？是什么样的平台在支撑这些让人匪夷所思的应用？——全世界的人都很好奇。好在 Google 并不保守——从 2003 年开始，Google 连续几年发表论文，揭示其核心技术，包括 Google 文件系统 GFS、Map/Reduce 编程模式、分布式锁机制 Chubby 以及大规模分布式数据库 BigTable 等。随后，Google CEO 施密特将这类技术称之为“云计算”。所谓“云计算”，就是用网络连接大量廉价计算节点，通过分布式软件虚拟成一个可靠的高性能计算平台。之所以称作“云”，是因为我们画网络图的时候，总是将网络画成一朵云。现在，这朵云变成了我们的“计算机”，而我们的 PC、智能手机等则变成了它的终端，因此称之为“云计算”。

2004 年，正当开源搜索引擎 Nutch 和开源全文检索包 Lucene 之父 Doug Cutting 为平台的可靠性和性能深受困扰时，看到了 Google 发表的

GFS 和 MapReduce 论文，花了 2 年时间将之实现，使平台的能力得到大幅提升。2006 年，Doug Cutting 加入 Yahoo!，并将这部分工作单列形成 Hadoop 项目组。Hadoop 的名称，并不是一个正式的英文单词，而来源于 Doug Cutting 的小儿子对所玩的小象玩具牙牙学语的称呼。Hadoop 主要由以下几个子项目组成。

(1) Hadoop Common: 是支撑 Hadoop 的公共部分，包括文件系统、远程过程调用 (RPC) 和序列化函数库等。

(2) HDFS: 提供高吞吐量的可靠分布式文件系统，是 GFS 的开源实现。

(3) MapReduce: 大型分布式数据处理模型，是 Google MapReduce 的开源实现。

与 Hadoop 直接相关的配套开源项目还包括如下几个方面。

(1) HBase: 支持结构化数据存储的分布式数据库，是 Bigtable 的开源实现。

(2) Hive: 提供数据摘要和查询功能的数据仓库。

(3) Pig: 是在 MapReduce 上构建的一种高级的数据流语言，可以简化 MapReduce 任务的开发。

(4) Cassandra: 由 Facebook 支持的开源高可扩展分布式数据库。是 Amazon 底层架构 Dynamo 的全分布和 Google Bigtable 的列式数据存储模型的有机结合。

(5) Chukwa: 一个用来管理大型分布式系统的数据采集系统。

(6) ZooKeeper: 用于解决分布式系统中一致性问题，是 Chubby 的开源实现。

等等。

经过 5 年发展，在所有的开源云计算系统里，Hadoop 稳居第一。事实上，Hadoop 是如此受欢迎，全球已经安装了数以万计的 Hadoop 系统。不仅高校和小企业使用 Hadoop，连 Facebook、淘宝、360 安全卫士这样的知名企业也在大规模使用 Hadoop。2007 年，Google 开始在全球推广“Google 101”计划，即在全球知名高校给学生开设 Google 模式的云计算编程课程。Google 中国公司大学合作部在近几年也在国内的清华大学、北京大学、南京大学、上海交大、同济大学等几家著名高校中资助开设了 MapReduce 和云计算技术课程，本书的部分章节内容也正是在所开设课程内容的基础上形成。有趣的是，由于 Google 不能直接将其平台开放给学生做实验室，于是 Google 干脆用 Hadoop 来搭建实验环境——可见 Google 对 Hadoop 的认同度。

目前国内学习和使用 Hadoop 的热情高涨。在中国云计算 (<http://www.chinacloud.cn>) 网站上作的一个调查表明,网友将 Hadoop 作为云计算领域要学习的首选技术。然而,目前国内 Hadoop 书籍非常匮乏,迫切需要原著的传授 Hadoop 编程经验和解决实际问题技巧的书籍。我们的云计算技术研发团队长期战斗在存储和处理海量数据的前线,在实践中积累了一些经验。为此,我们感觉到有必要向淘宝网核心架构团队学习,将自己积累的点滴经验贡献出来与大家分享,于是萌生了创作此书的念头。

本书由刘鹏教授负责总体设计、组织全书写作和对内容把关调整,黄宜华教授负责第 5 章和整体润色,陈卫卫教授负责全书逐字审校,程浩完成第 3、4 章,王磊完成第 14 章,顾荣完成第 11 章,张贞完成第 1、2 章,邓鹏完成第 6 章并与王胤然完成第 10 章,杨晓亮完成第 12 章,郭岩岩与王海坤完成第 13 章,李浩完成第 8 章,魏家宾完成第 7 章,张欣完成第 9 章。南京云创存储科技有限公司对本书的编写给予了大力支持,在紧张从事云计算系统研发的同时,仍然派出几位研发人员参与本书写作。

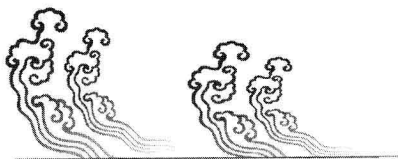
衷心感谢中国电子学会云计算专家委员会、中国通信学会云计算与 SaaS 专家委员会、江苏省计算机学会、江苏省云计算论坛专家委员会和电子工业出版社对本书出版的大力支持。感谢多位知名院士、专家在百忙之中阅读书稿并给予推荐。在此拜谢我的硕士生导师谢希仁教授和博士生导师李三立院士对我的辛勤培养。

本书的读者服务网页为: <http://www.chinacloud.cn/hadoop.html>, 该网页提供书中源码下载、相关信息和问题交流。如读者需要较为全面地学习云计算技术,欢迎阅读本书的姊妹篇《云计算(第二版)》,刘鹏主编,电子工业出版社 2011 年 5 月出版。

由于作者水平有限,加之时间较紧,书中难免会存在写作不到位甚至错误的之处,敬请读者批评指正。意见和建议请发邮件到: [cloudforum@163.com](mailto:cloudforum@163.com)。新浪微博互动交流请关注: <http://weibo.com/cloudgrid>。

解放军理工大学 刘 鹏

2011 年 9 月



# 目 录

## 第 1 章 神奇的大象——Hadoop

1.1 初识神象	2
1.2 Hadoop 初体验	4
1.2.1 了解 Hadoop 的构架	4
1.2.2 查看 Hadoop 活动	7
1.3 Hadoop 族群	10
1.4 Hadoop 安装	11
1.4.1 在 Linux 系统中安装 Hadoop	11
1.4.2 在 Windows 系统中安装 Hadoop	21
1.4.3 站在象背上说“hello”	29
1.4.4 Eclipse 下的 Hadoop 应用开发	30
参考文献	34

## 第 2 章 HDFS——不怕故障的海量存储

2.1 开源的 GFS——HDFS	36
2.1.1 设计前提与目标	36
2.1.2 HDFS 体系结构	37
2.1.3 保障 HDFS 可靠性措施	39
2.2 HDFS 常用操作	42
2.2.1 HDFS 下的文件操作	42
2.2.2 管理与更新	45
2.3 HDFS API 之旅	48



2.4 实战：用 HDFS 存储海量视频数据	55
2.4.1 应用场景	55
2.4.2 设计实现	55
参考文献	58

### 第 3 章 分久必合——MapReduce

3.1 MapReduce 基础	60
3.1.1 MapReduce 编程模型	60
3.1.2 MapReduce 的集群行为	62
3.2 样例分析：单词计数	64
3.2.1 WordCount 源码分析	64
3.2.2 WordCount 处理过程	67
3.3 MapReduce，你够了解吗	69
3.3.1 没有 map、reduce 的 MapReduce	69
3.3.2 多少个 Reducers 最佳	72
3.4 实战：倒排索引	74
3.4.1 倒排索引简介	74
3.4.2 分析与设计	76
3.4.3 倒排索引完整源码	79
参考文献	83

### 第 4 章 一张无限大的表——HBase

4.1 HBase 简介	85
4.1.1 逻辑模型	85
4.1.2 物理模型	86
4.1.3 Region 服务器	87
4.1.4 主服务器	89
4.1.5 元数据表	89
4.2 HBase 入门	91
4.2.1 HBase 的安装配置	91
4.2.2 HBase 用户界面	97

4.3	HBase 操作演练	100
4.3.1	基本 shell 操作	100
4.3.2	基本 API 使用	103
4.4	实战：使用 MapReduce 构建 HBase 索引	105
4.4.1	索引表蓝图	105
4.4.2	HBase 和 MapReduce	107
4.4.3	实现索引	108
	参考文献	112

## 第 5 章 更上一层楼——MapReduce 进阶

5.1	简介	114
5.2	复合键值对的使用	115
5.2.1	把小的键值对合并成大的键值对	115
5.2.2	巧用复合键让系统完成排序	117
5.3	用户定制数据类型	123
5.3.1	Hadoop 内置的数据类型	123
5.3.2	用户自定义数据类型的实现	124
5.4	用户定制输入/输出格式	126
5.4.1	Hadoop 内置的数据输入格式和 RecordReader	126
5.4.2	用户定制数据输入格式与 RecordReader	127
5.4.3	Hadoop 内置的数据输出格式与 RecordWriter	133
5.4.4	用户定制数据输出格式与 RecordWriter	134
5.4.5	通过定制数据输出格式实现多集合文件输出	134
5.5	用户定制 Partitioner 和 Combiner	137
5.5.1	用户定制 Partitioner	137
5.5.2	用户定制 Combiner	139
5.6	组合式 MapReduce 计算作业	141
5.6.1	迭代 MapReduce 计算任务	141
5.6.2	顺序组合式 MapReduce 作业的执行	142
5.6.3	具有复杂依赖关系的组合式 MapReduce 作业的执行	144

5.6.4	MapReduce 前处理和后处理步骤的链式执行	145
5.7	多数据源的连接	148
5.7.1	基本问题数据示例	149
5.7.2	用 DataJoin 类实现 Reduce 端连接	150
5.7.3	用全局文件复制方法实现 Map 端连接	158
5.7.4	带 Map 端过滤的 Reduce 端连接	162
5.7.5	多数据源连接解决方法的限制	162
5.8	全局参数/数据文件的传递与使用	163
5.8.1	全局作业参数的传递	163
5.8.2	查询全局 MapReduce 作业属性	166
5.8.3	全局数据文件的传递	167
5.9	关系数据库的连接与访问	169
5.9.1	从数据库中输入数据	169
5.9.2	向数据库中输出计算结果	170
	参考文献	172

## 第 6 章 Hive——飞进数据仓库的小蜜蜂

6.1	Hive 的组成	174
6.2	搭建蜂房——Hive 安装	176
6.3	Hive 的服务	182
6.3.1	Hive shell	182
6.3.2	JDBC/ODBC 支持	183
6.3.3	Thrift 服务	184
6.3.4	Web 接口	185
6.3.5	元数据服务	186
6.4	HiveQL 的使用	187
6.4.1	HiveQL 的数据类型	187
6.4.2	HiveQL 常用操作	188
6.5	Hive 示例	196
6.5.1	UDF 编程示例	196

6.5.2	UDAF 编程示例	198
6.6	实战：基于 Hive 的 Hadoop 日志分析	200
	参考文献	209
<b>第 7 章 Pig——一头什么都能吃的猪</b>		
7.1	Pig 的基本框架	211
7.2	Pig 的安装	212
7.2.1	开始安装 Pig	212
7.2.2	验证安装	213
7.3	Pig 的使用	214
7.3.1	Pig 的 MapReduce 模式	214
7.3.2	使用 Pig	216
7.3.3	Pig 的调试	219
7.4	Pig Latin 编程语言	224
7.4.1	数据模型	224
7.4.2	数据类型	225
7.4.3	运算符	226
7.4.4	常用操作	228
7.4.5	用户自定义函数	231
7.5	实战：基于 Pig 的通话记录查询	231
7.5.1	应用场景	231
7.5.2	设计实现	232
	参考文献	238
<b>第 8 章 Facebook 的女神——Cassandra</b>		
8.1	洞察 Cassandra 的全貌	240
8.1.1	目标及特点	240
8.1.2	体系结构	241
8.1.3	存储机制	243
8.1.4	数据操作过程	244

8.2 让 Cassandra 飞	247
8.2.1 Windows 7 下单机安装	247
8.2.2 Linux 下分布式安装	249
8.3 Cassandra 操作示例	253
8.3.1 客户端命令代码跟踪	253
8.3.2 增删 Cassandra 节点	262
8.3.3 Jconsole 监控 Cassandra	263
8.4 Cassandra 与 MapReduce 结合	266
8.4.1 需求分析	266
8.4.2 编码流程分析	267
8.4.3 MapReduce 的核心代码	268
参考文献	269

## 第 9 章 Chukwa——收集数据的大乌龟

9.1 初识 Chukwa	271
9.1.1 为什么需要 Chukwa	271
9.1.2 什么是 Chukwa	272
9.2 Chukwa 架构与设计	274
9.2.1 代理与适配器	276
9.2.2 元数据	277
9.2.3 收集器	278
9.2.4 MapReduce 作业	279
9.2.5 HICC	280
9.2.6 数据接口与支持	280
9.3 Chukwa 安装与配置	281
9.3.1 Chukwa 安装	281
9.3.2 源节点代理配置	284
9.3.3 收集器	288
9.3.4 Demux 作业与 HICC 配置	289
9.4 Chukwa 小试	291

9.4.1 数据生成	291
9.4.2 数据收集	292
9.4.3 数据处理	292
9.4.4 数据析取	293
9.4.5 数据稀释	294
9.4.6 数据显示	294
参考文献	295

## 第 10 章 一统天下——ZooKeeper

10.1 Zookeeper 是个谜	297
10.1.1 ZooKeeper 工作原理	298
10.1.2 ZooKeeper 的特性	301
10.2 ZooKeeper 安装和编程	303
10.2.1 ZooKeeper 的安装和配置	303
10.2.2 ZooKeeper 的编程实现	306
10.3 ZooKeeper 演练：进程调度系统	308
10.3.1 设计方案	308
10.3.2 设计实现	309
10.4 实战演练：ZooKeeper 实现 NameNode 自动切换	318
10.4.1 设计思想	319
10.4.2 详细设计	319
10.4.3 编码	321
10.4.4 实战总结	329
参考文献	329

## 第 11 章 综合实战 1——打造一个搜索引擎

11.1 系统工作原理	331
11.2 网页搜集与信息提取	333
11.2.1 网页搜集	334
11.2.2 网页信息的提取与存储	337

11.3 基于 MapReduce 的预处理	338
11.3.1 元数据过滤	339
11.3.2 生成倒排文件	341
11.3.3 建立二级索引	353
11.3.4 小节	357
11.4 建立 Web 信息查询服务	358
11.4.1 建立前台查询接口	358
11.4.2 后台信息查询与合并	359
11.4.3 返回显示结果	360
11.5 系统优化	361
11.5.1 存储方面的优化	361
11.5.2 计算方面的优化	362
11.6 本章总结	363
参考文献	364
<b>第 12 章 综合实战 2——生物信息学应用</b>	
12.1 背景	366
12.2 总体框架	368
12.3 系统实现	370
12.3.1 序列数据库的切分和存储	370
12.3.2 构造单词列表和扫描器	375
12.3.3 Map: 扫描和扩展	376
12.3.4 主控程序	378
12.4 扩展性能测试	381
12.5 本章总结	382
参考文献	383
<b>第 13 章 综合实战 3——移动通信信令监测与查询</b>	
13.1 分析与设计	385
13.1.1 CDR 数据文件的检测与索引创建任务调度	388
13.1.2 从 HDFS 读取数据并创建索引	389

13.1.3 查询 CDR 信息	390
13.2 实现代码	391
13.2.1 CDR 文件检测和索引创建任务调度程序	392
13.2.2 读取 CDR 数据和索引创建处理	397
13.2.3 CDR 查询	402
13.3 本章总结	407
参考文献	407

## 第 14 章 高枕无忧——Hadoop 容错

14.1 Hadoop 的可靠性	409
14.1.1 HDFS 中 NameNode 单点问题	409
14.1.2 HDFS 数据块副本机制	410
14.1.3 HDFS 心跳机制	411
14.1.4 HDFS 负载均衡	412
14.1.5 MapReduce 容错	413
14.2 Hadoop 的 SecondaryNameNode 机制	414
14.2.1 磁盘镜像与日志文件	414
14.2.2 SecondaryNameNode 更新镜像的流程	414
14.3 Avatar 机制	418
14.3.1 系统架构	419
14.3.2 Avatar 元数据同步机制	420
14.3.3 故障切换过程	423
14.3.4 Avatar 运行流程	426
14.3.5 Avatar 故障切换流程	431
14.4 Avatar 实战	436
14.4.1 实验环境	436
14.4.2 编译 Avatar	437
14.4.3 Avatar 安装和配置	440
14.4.4 Avatar 启动运行与宕机切换	452
参考文献	456



神奇的大象——Hadoop



- 1.1 初识神象
- 1.2 Hadoop 初体验
- 1.3 Hadoop 族群
- 1.4 Hadoop 安装