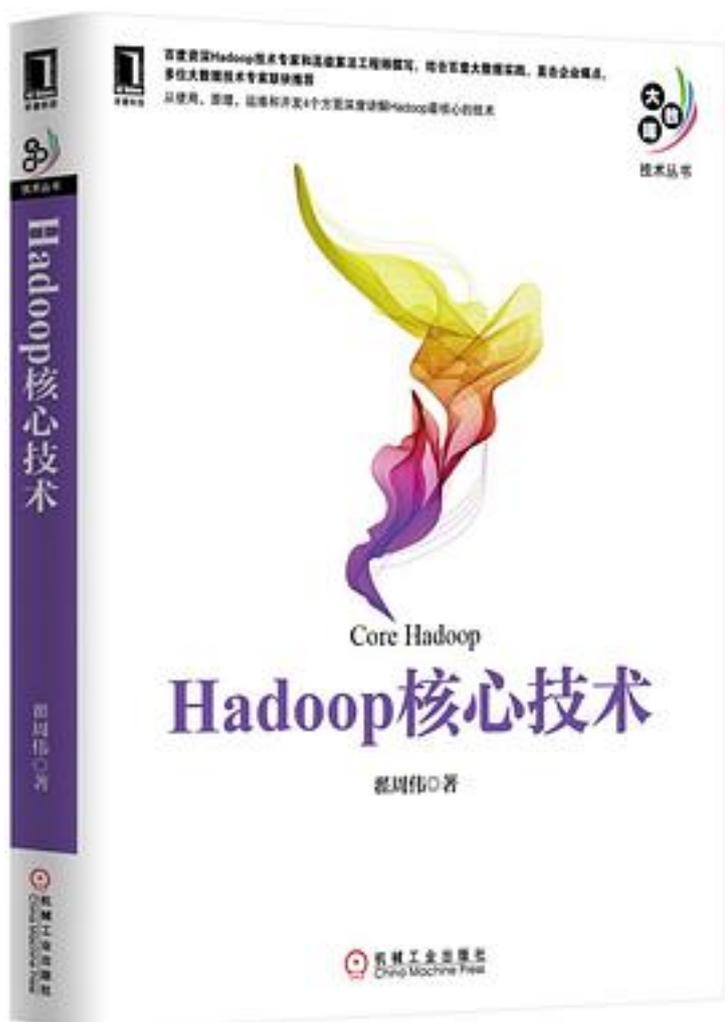


Hadoop核心技术



[Hadoop核心技术_下载链接1](#)

著者:翟周伟

出版者:机械工业出版社

出版时间:2015-4-1

装帧:平装

isbn:9787111494683

百度资深Hadoop技术专家和高级算法工程师撰写，结合百度大数据实践，直击企业痛

点，多位大数据技术专家联袂推荐！

从使用、原理、运维和开发4个方面深度讲解Hadoop最核心的技术

这是一本技术深度与企业实践并重的著作，由百度顶尖的Hadoop技术工程师撰写，是百度Hadoop技术实践经验的总结。本书使用、实现原理、运维和开发4个方面对Hadoop的核心技术进行了深入的讲解：

- (1) 使用：详细讲解了HDFS存储系统、MapReduce计算框架，以及HDFS的命令系统；
- (2) 原理：结合源代码，深度分析了MapReduce、HDFS、Streaming、Pipes、Hadoop作业调度系统等重要技术和组件的架构设计、工作机制和实现原理；
- (3) 运维：结合百度的实际生产环境，详细讲解了Hadoop集群的安装、配置、测试以及管理和运维；
- (4) 开发：详细讲解了Hadoop Streaming、Pipes的使用和开发实践，以及MapReduce的编程实践和常见问题。

与市面上已有的Hadoop相比，本书的最大不同之处是它直切企业应用和实践Hadoop技术的痛点，深入讲解了企业最需要和最头疼的技术和问题，内容上非常聚焦。

作者介绍:

翟周伟

就职于百度，资深Hadoop技术专家，专注于Hadoop&大数据，数据挖掘，自然语言处理领域。2009年便开始利用Hadoop构建商业级大数据系统，是国内该领域最早的一批人之一，负责设计过多个基于Hadoop的大数据平台和分析系统。2011年合著出版《Hadoop开源云计算平台》，并在自然语言处理领域申请过一项发明专利。

目录: 前言

基础篇

第1章 认识Hadoop 2

1.1 缘于搜索的小象 2

1.1.1 Hadoop的身世 2

1.1.2 Hadoop简介 3

1.1.3 Hadoop发展简史 6

1.2 大数据、Hadoop和云计算 7

1.2.1 大数据 7

1.2.2 大数据、Hadoop和云计算的关系 8

1.3 设计思想与架构 9

1.3.1 数据存储与切分 9

1.3.2 MapReduce模型 11

1.3.3 MPI和MapReduce 13

1.4 国外Hadoop的应用现状 13

1.5 国内Hadoop的应用现状 17

1.6 Hadoop发行版 20

1.6.1 Apache Hadoop 20

1.6.2 Cloudera Hadoop 20

1.6.3 Hortonworks Hadoop发行版 21

1.6.4 MapR Hadoop发行版	22
1.6.5 IBM Hadoop发行版	24
1.6.6 Intel Hadoop发行版	24
1.6.7 华为Hadoop发行版	25
1.7 小结	26
第2章 Hadoop使用之初体验	27
2.1 搭建测试环境	27
2.1.1 软件与准备	27
2.1.2 安装与配置	28
2.1.3 启动与停止	29
2.2 算法分析与设计	31
2.2.1 Map设计	31
2.2.2 Reduce设计	32
2.3 实现接口	32
2.3.1 Java API实现	33
2.3.2 Streaming接口实现	36
2.3.3 Pipes接口实现	38
2.4 编译	40
2.4.1 基于Java API实现的编译	40
2.4.2 基于Streaming实现的编译	40
2.4.3 基于Pipes实现的编译	41
2.5 提交作业	41
2.5.1 基于Java API实现作业提交	41
2.5.2 基于Streaming实现作业提交	42
2.5.3 基于Pipes实现作业提交	43
2.6 小结	44
第3章 Hadoop存储系统	45
3.1 基本概念	46
3.1.1 NameNode	46
3.1.2 DataNode	46
3.1.3 客户端	47
3.1.4 块	47
3.2 HDFS的特性和目标	48
3.2.1 HDFS的特性	48
3.2.2 HDFS的目标	48
3.3 HDFS架构	49
3.3.1 Master/Slave架构	49
3.3.2 NameNode和Secondary NameNode通信模型	51
3.3.3 文件存取机制	52
3.4 HDFS核心设计	54
3.4.1 Block大小	54
3.4.2 数据复制	55
3.4.3 数据副本存放策略	56
3.4.4 数据组织	57
3.4.5 空间回收	57
3.4.6 通信协议	58
3.4.7 安全模式	58
3.4.8 机架感知	59
3.4.9 健壮性	59
3.4.10 负载均衡	60
3.4.11 升级和回滚机制	62
3.5 HDFS权限管理	64
3.5.1 用户身份	64
3.5.2 系统实现	65

- 3.5.3 超级用户 65
- 3.5.4 配置参数 65
- 3.6 HDFS配额管理 66
- 3.7 HDFS的缺点 67
- 3.8 小结 68
- 第4章 HDFS的使用 69
 - 4.1 HDFS环境准备 69
 - 4.1.1 HDFS安装配置 69
 - 4.1.2 HDFS格式化与启动 70
 - 4.1.3 HDFS运行检查 70
 - 4.2 HDFS命令的使用 71
 - 4.2.1 fs shell 71
 - 4.2.2 archive 77
 - 4.2.3 distcp 78
 - 4.2.4 fsck 81
 - 4.3 HDFS Java API的使用方法 82
 - 4.3.1 Java API简介 82
 - 4.3.2 读文件 82
 - 4.3.3 写文件 86
 - 4.3.4 删除文件或目录 90
 - 4.4 C接口libhdfs 91
 - 4.4.1 libhdfs介绍 91
 - 4.4.2 编译与部署 91
 - 4.4.3 libhdfs接口介绍 92
 - 4.4.4 libhdfs使用举例 95
 - 4.5 WebHDFS接口 97
 - 4.5.1 WebHDFS REST API简介 97
 - 4.5.2 WebHDFS配置 98
 - 4.5.3 WebHDFS使用 98
 - 4.5.4 WebHDFS错误响应和查询参数 101
 - 4.6 小结 103
- 第5章 MapReduce计算框架 104
 - 5.1 Hadoop MapReduce简介 104
 - 5.2 MapReduce模型 105
 - 5.2.1 MapReduce编程模型 105
 - 5.2.2 MapReduce实现原理 106
 - 5.3 计算流程与机制 108
 - 5.3.1 作业提交和初始化 108
 - 5.3.2 Mapper 110
 - 5.3.3 Reducer 111
 - 5.3.4 Reporter和OutputCollector 112
 - 5.4 MapReduce的输入/输出格式 113
 - 5.4.1 输入格式 113
 - 5.4.2 输出格式 118
 - 5.5 核心问题 124
 - 5.5.1 Map和Reduce数量 124
 - 5.5.2 作业配置 126
 - 5.5.3 作业执行和环境 127
 - 5.5.4 作业容错机制 129
 - 5.5.5 作业调度 131
 - 5.6 有用的MapReduce特性 132
 - 5.6.1 计数器 132
 - 5.6.2 DistributedCache 134
 - 5.6.3 Tool 135

- 5.6.4 IsolationRunner 136
- 5.6.5 Profiling 136
- 5.6.6 MapReduce调试 136
- 5.6.7 数据压缩 137
- 5.6.8 优化 138
- 5.7 小结 138
- 第6章 Hadoop命令系统 139
 - 6.1 Hadoop命令系统的组成 139
 - 6.2 用户命令 141
 - 6.3 管理员命令 144
 - 6.4 测试命令 148
 - 6.5 应用命令 156
 - 6.6 Hadoop的streaming命令 163
 - 6.6.1 streaming命令 163
 - 6.6.2 参数使用分析 164
 - 6.7 Hadoop的pipes命令 168
 - 6.7.1 pipes命令 168
 - 6.7.2 参数使用分析 169
 - 6.8 小结 170
- 高级篇
- 第7章 MapReduce深度分析 172
 - 7.1 MapReduce总结构分析 172
 - 7.1.1 数据流向分析 172
 - 7.1.2 处理流程分析 174
 - 7.2 MapTask实现分析 176
 - 7.2.1 总逻辑分析 176
 - 7.2.2 Read阶段 178
 - 7.2.3 Map阶段 178
 - 7.2.4 Collector和Partitioner阶段 180
 - 7.2.5 Spill阶段 181
 - 7.2.6 Merge阶段 185
 - 7.3 ReduceTask实现分析 186
 - 7.3.1 总逻辑分析 186
 - 7.3.2 Shuffle阶段 187
 - 7.3.3 Merge阶段 189
 - 7.3.4 Sort阶段 190
 - 7.3.5 Reduce阶段 191
 - 7.4 JobTracker分析 192
 - 7.4.1 JobTracker服务分析 192
 - 7.4.2 JobTracker启动分析 193
 - 7.4.3 JobTracker核心子线程分析 195
 - 7.5 TaskTracker分析 201
 - 7.5.1 TaskTracker启动分析 201
 - 7.5.2 TaskTracker核心子线程分析 205
 - 7.6 心跳机制实现分析 207
 - 7.6.1 心跳检测分析 207
 - 7.6.2 TaskTracker.transmitHeart-Beat() 207
 - 7.6.3 JobTracker.heartbeat() 209
 - 7.6.4 JobTracker.processHeartbeat() 212
 - 7.7 作业创建分析 213
 - 7.7.1 初始化分析 214
 - 7.7.2 作业提交分析 215
 - 7.8 作业执行分析 217
 - 7.8.1 JobTracker初始化 218

- 7.8.2 TaskTracker.startNewTask() 220
- 7.8.3 TaskTracker.localizeJob() 220
- 7.8.4 TaskRunner.run() 221
- 7.8.5 MapTask.run() 222
- 7.9 小结 223
- 第8章 Hadoop Streaming和Pipes原理与实现 224
 - 8.1 Streaming原理浅析 224
 - 8.2 Streaming实现架构 226
 - 8.3 Streaming核心实现机制 227
 - 8.3.1 主控框架实现 227
 - 8.3.2 用户进程管理 228
 - 8.3.3 框架和用户程序的交互 229
 - 8.3.4 PipeMapper和PiperReducer 230
 - 8.4 Pipes原理浅析 231
 - 8.5 Pipes实现架构 233
 - 8.6 Pipes核心实现机制 234
 - 8.6.1 主控类实现 234
 - 8.6.2 用户进程管理 235
 - 8.6.3 PipesMapRunner 235
 - 8.6.4 PipesReducer 238
 - 8.6.5 C++端HadoopPipes 238
 - 8.7 小结 239
- 第9章 Hadoop作业调度系统 240
 - 9.1 作业调度概述 241
 - 9.1.1 相关概念 241
 - 9.1.2 作业调度流程 242
 - 9.1.3 集群资源组织与管理 243
 - 9.1.4 队列控制和权限管理 244
 - 9.1.5 插件式调度框架 245
 - 9.2 FIFO调度器 246
 - 9.2.1 基本调度策略 246
 - 9.2.2 FIFO实现分析 247
 - 9.2.3 FIFO初始化与停止 248
 - 9.2.4 作业监听控制 249
 - 9.2.5 任务分配算法 250
 - 9.2.6 配置与使用 254
 - 9.3 公平调度器 254
 - 9.3.1 产生背景 254
 - 9.3.2 主要功能 255
 - 9.3.3 基本调度策略 255
 - 9.3.4 FairScheduler实现分析 257
 - 9.3.5 FairScheduler启停分析 258
 - 9.3.6 作业监听控制 260
 - 9.3.7 资源池管理 260
 - 9.3.8 作业更新策略 262
 - 9.3.9 作业权重和资源量的计算 266
 - 9.3.10 任务分配算法 267
 - 9.3.11 FairScheduler配置参数 268
 - 9.3.12 使用与管理 270
 - 9.4 容量调度器 272
 - 9.4.1 产生背景 272
 - 9.4.2 主要功能 272
 - 9.4.3 基本调度策略 274
 - 9.4.4 CapacityScheduler实现分析 274

9.4.5 CapacityScheduler启停分析 275

9.4.6 作业监听控制 277

9.4.7 作业初始化分析 277

9.4.8 任务分配算法 278

9.4.9 内存匹配机制 279

9.4.10 配置与使用 280

9.5 调度器对比分析 283

9.5.1 调度策略对比 283

9.5.2 队列和优先级 283

9.5.3 资源分配保证 283

9.5.4 作业限制 284

9.5.5 配置管理 284

9.5.6 扩展性支持 284

9.5.7 资源抢占和延迟调度 284

9.5.8 优缺点分析 285

9.6 其他调度器 285

9.6.1 HOD调度器 285

9.6.2 LATE调度器 286

9.7 小结 288

实战篇

第10章 Hadoop集群搭建 290

10.1 Hadoop版本的选择 290

10.2 集群基础硬件需求 291

10.2.1 内存 291

10.2.2 CPU 292

10.2.3 磁盘 292

10.2.4 网卡 293

10.2.5 网络拓扑 293

10.3 集群基础软件需求 294

10.3.1 操作系统 294

10.3.2 JVM和SSH 295

10.4 虚拟化需求 295

10.5 事前准备 296

10.5.1 创建安装用户 296

10.5.2 安装Java 297

10.5.3 安装SSH并设置 297

10.5.4 防火墙端口设置 298

10.6 安装Hadoop 298

10.6.1 安装HDFS 299

10.6.2 安装MapReduce 299

10.7 集群配置 300

10.7.1 配置管理 300

10.7.2 环境变量配置 301

10.7.3 核心参数配置 302

10.7.4 HDFS参数配置 303

10.7.5 MapReduce参数配置 306

10.7.6 masters和slaves配置 313

10.7.7 客户端配置 313

10.8 启动和停止 314

10.8.1 启动/停止HDFS 314

10.8.2 启动/停止MapReduce 315

10.8.3 启动验证 315

10.9 集群基准测试 316

10.9.1 HDFS基准测试 316

- 10.9.2 MapReduce基准测试 317
- 10.9.3 综合性能测试 318
- 10.10 集群搭建实例 319
 - 10.10.1 部署策略 319
 - 10.10.2 软件和硬件环境 320
 - 10.10.3 Hadoop安装 321
 - 10.10.4 配置core-site.xml 321
 - 10.10.5 配置hdfs-site.xml 322
 - 10.10.6 配置mapred-site.xml 322
 - 10.10.7 SecondaryNameNode和Slave 324
 - 10.10.8 配置作业队列 324
 - 10.10.9 配置第三方调度器 325
 - 10.10.10 启动与验证 327
- 10.11 小结 327
- 第11章 Hadoop Streaming和Pipes编程实战 328
 - 11.1 Streaming基础编程 328
 - 11.1.1 Streaming编程入门 328
 - 11.1.2 Map和Reduce数目 331
 - 11.1.3 队列、优先级及权限 332
 - 11.1.4 分发文件和压缩包 333
 - 11.1.5 压缩参数的使用 336
 - 11.1.6 本地作业的调试 338
 - 11.2 Streaming高级应用 338
 - 11.2.1 参数与环境变量传递 339
 - 11.2.2 自定义分隔符 340
 - 11.2.3 自定义Partitioner 343
 - 11.2.4 自定义计数器 347
 - 11.2.5 处理二进制数据 347
 - 11.2.6 使用聚合函数 351
 - 11.3 Pipes编程接口 352
 - 11.3.1 TaskContext 352
 - 11.3.2 Mapper 353
 - 11.3.3 Reducer 354
 - 11.3.4 Partitioner 354
 - 11.3.5 RecordReader 355
 - 11.3.6 RecordWriter 356
 - 11.4 Pipes编程应用 357
 - 11.5 小结 359
- 第12章 Hadoop MapReduce应用开发 360
 - 12.1 开发环境准备 360
 - 12.2 Eclipse集成环境开发 361
 - 12.2.1 构建MapReduce Eclipse IDE 361
 - 12.2.2 开发示例 363
 - 12.3 MapReduce Java API编程 368
 - 12.3.1 Mapper编程接口 369
 - 12.3.2 Reducer编程接口 370
 - 12.3.3 驱动类编写 372
 - 12.3.4 编译运行 373
 - 12.4 压缩功能使用 374
 - 12.4.1 Hadoop数据压缩 374
 - 12.4.2 压缩特征与性能 374
 - 12.4.3 本地压缩库 375
 - 12.4.4 使用压缩 376
 - 12.5 排序应用 378

- 12.5.1 Hadoop排序问题 378
- 12.5.2 二次排序 378
- 12.5.3 比较器和组合排序 380
- 12.5.4 全局排序 381
- 12.6 多路输出 382
- 12.7 常见问题与处理方法 384
- 12.7.1 常见的开发问题 384
- 12.7.2 运行时错误问题 386
- 12.8 小结 387
- • • • • (收起)

[Hadoop核心技术_下载链接1](#)

标签

Hadoop

大数据

计算机

技术

数据

评论

国内作者写的最好的Hadoop图书之一。

文字表达很好，讲解的很清楚

[Hadoop核心技术_下载链接1](#)

书评

[Hadoop核心技术_下载链接1](#)