

<<Hadoop技术内幕>>

图书基本信息

书名：<<Hadoop技术内幕>>

13位ISBN编号：9787111445340

10位ISBN编号：7111445341

出版时间：2013-11-30

出版时间：机械工业出版社

作者：董西成

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

<<Hadoop技术内幕>>

内容概要

本书从应用角度系统讲解了YARN的基本库和组件用法、应用程序设计方法、YARN上流行的各种计算框架（MapReduce、Tez、Storm、Spark），以及多个类YARN的开源资源管理系统（Corona和Mesos）；从源代码角度深入分析YARN的设计理念与基本架构、各个组件的实现原理，以及各种计算框架的实现细节。

全书共四部分13章：第一部分（第1~2章）主要介绍了如何获取、阅读和调试Hadoop的源代码，以及YARN的设计思想、基本架构和工作流程；第二部分（第3~7章）结合源代码详细剖析和讲解了YARN的第三方开源库、底层通信库、服务库、事件库的基本使用和实现细节，详细讲解了YARN的应用程序设计方法，深入讲解和分析了ResourceManager、资源调度器、NodeManager等组件的实现细节；第三篇（第8~10章）则对离线计算框架MapReduce、DAG计算框架Tez、实时计算框架Storm和内存计算框架Spark进行了详细的讲解；第四部分（第11~13章）首先对Facebook Corona和Apache Mesos进行了深入讲解，然后对YARN的发展趋势进行了展望。

附录部分收录了YARN安装指南、YARN配置参数以及Hadoop Shell命令等非常有用的资料。

<<Hadoop技术内幕>>

书籍目录

前 言

第一部分 准备篇

第1章 环境准备 2

1.1 准备学习环境 2

1.1.1 基础软件下载 2

1.1.2 如何准备Linux环境 3

1.2 获取Hadoop源代码 5

1.3 搭建Hadoop源代码阅读环境 5

1.3.1 创建Hadoop工程 5

1.3.2 Hadoop源代码阅读技巧 8

1.4 Hadoop源代码组织结构 10

1.5 Hadoop初体验 12

1.5.1 搭建Hadoop环境 12

1.5.2 Hadoop Shell介绍 15

1.6 编译及调试Hadoop源代码 16

1.6.1 编译Hadoop源代码 17

1.6.2 调试Hadoop源代码 18

1.7 小结 20

第2章 YARN设计理念与基本架构 21

2.1 YARN产生背景 21

2.1.1 MRv1的局限性 21

2.1.2 轻量级弹性计算平台 22

2.2 Hadoop基础知识 23

2.2.1 术语解释 23

2.2.2 Hadoop版本变迁 25

2.3 YARN基本设计思想 29

2.3.1 基本框架对比 29

2.3.2 编程模型对比 30

2.4 YARN 基本架构 31

2.4.1 YARN基本组成结构 32

2.4.2 YARN通信协议 34

2.5 YARN工作流程 35

2.6 多角度理解YARN 36

2.6.1 并行编程 36

2.6.2 资源管理系统 36

2.6.3 云计算 37

2.7 本书涉及内容 38

2.8 小结 38

第二部分 YARN核心设计篇

第3章 YARN基础库 40

3.1 概述 40

3.2 第三方开源库 41

3.2.1 Protocol Buffers 41

3.2.2 Apache Avro 43

3.3 底层通信库 46

<<Hadoop技术内幕>>

- 3.3.1 RPC通信模型 46
- 3.3.2 Hadoop RPC的特点概述 48
- 3.3.3 RPC总体架构 48
- 3.3.4 Hadoop RPC使用方法 49
- 3.3.5 Hadoop RPC类详解 51
- 3.3.6 Hadoop RPC参数调优 57
- 3.3.7 YARN RPC实现 57
- 3.3.8 YARN RPC应用实例 61
- 3.4 服务库与事件库 65
 - 3.4.1 服务库 66
 - 3.4.2 事件库 66
 - 3.4.3 YARN服务库和事件库的使用方法 68
 - 3.4.4 事件驱动带来的变化 70
- 3.5 状态机库 72
 - 3.5.1 YARN状态转换方式 72
 - 3.5.2 状态机类 73
 - 3.5.3 状态机的使用方法 73
 - 3.5.4 状态机可视化 76
- 3.6 源代码阅读引导 76
- 3.7 小结 77
- 3.8 问题讨论 77
- 第4章 YARN应用程序设计方法 78
 - 4.1 概述 78
 - 4.2 客户端设计 79
 - 4.2.1 客户端编写流程 80
 - 4.2.2 客户端编程库 84
 - 4.3 ApplicationMaster设计 84
 - 4.3.1 ApplicationMaster编写流程 84
 - 4.3.2 ApplicationMaster编程库 92
 - 4.4 YARN 应用程序实例 95
 - 4.4.1 DistributedShell 95
 - 4.4.2 Unmanaged AM 99
 - 4.5 源代码阅读引导 100
 - 4.6 小结 100
 - 4.7 问题讨论 100
- 第5章 ResourceManager剖析 102
 - 5.1 概述 102
 - 5.1.1 ResourceManager基本职能 102
 - 5.1.2 ResourceManager内部架构 103
 - 5.1.3 ResourceManager事件与事件处理器 106
 - 5.2 用户交互模块 108
 - 5.2.1 ClientRMService 108
 - 5.2.2 AdminService 109
 - 5.3 ApplicationMaster管理 109
 - 5.4 NodeManager管理 112
 - 5.5 Application管理 113
 - 5.6 状态机管理 114

<<Hadoop技术内幕>>

- 5.6.1 RMAp状态机 115
- 5.6.2 RMApAttempt状态机 119
- 5.6.3 RMContainer状态机 123
- 5.6.4 RMNode状态机 127
- 5.7 几个常见行为分析 129
 - 5.7.1 启动ApplicationMaster 129
 - 5.7.2 申请与分配Container 132
 - 5.7.3 杀死Application 134
 - 5.7.4 Container超时 135
 - 5.7.5 ApplicationMaster超时 138
 - 5.7.6 NodeManager超时 138
- 5.8 安全管理 139
 - 5.8.1 术语介绍 139
 - 5.8.2 Hadoop认证机制 139
 - 5.8.3 Hadoop授权机制 142
- 5.9 容错机制 144
 - 5.9.1 Hadoop HA基本框架 145
 - 5.9.2 YARN HA实现 148
- 5.10 源代码阅读引导 149
- 5.11 小结 151
- 5.12 问题讨论 152
- 第6章 资源调度器 153
 - 6.1 资源调度器背景 153
 - 6.2 HOD调度器 154
 - 6.2.1 Torque资源管理器 154
 - 6.2.2 HOD作业调度 155
 - 6.3 YARN资源调度器的基本架构 157
 - 6.3.1 基本架构 157
 - 6.3.2 资源表示模型 160
 - 6.3.3 资源调度模型 161
 - 6.3.4 资源抢占模型 164
 - 6.4 YARN层级队列管理机制 169
 - 6.4.1 层级队列管理机制 169
 - 6.4.2 队列命名规则 171
 - 6.5 Capacity Scheduler 172
 - 6.5.1 Capacity Scheduler的功能 172
 - 6.5.2 Capacity Scheduler实现 176
 - 6.6 Fair Scheduler 179
 - 6.6.1 Fair Scheduler功能介绍 180
 - 6.6.2 Fair Scheduler实现 182
 - 6.6.3 Fair Scheduler与Capacity Scheduler对比 183
 - 6.7 其他资源调度器介绍 184
 - 6.8 源代码阅读引导 185
 - 6.9 小结 186
 - 6.10 问题讨论 187
- 第7章 NodeManager剖析 188
 - 7.1 概述 188

<<Hadoop技术内幕>>

- 7.1.1 NodeManager基本职能 188
 - 7.1.2 NodeManager内部架构 190
 - 7.1.3 NodeManager事件与事件处理器 193
 - 7.2 节点健康状况检测 194
 - 7.2.1 自定义Shell脚本 194
 - 7.2.2 检测磁盘损坏数目 196
 - 7.3 分布式缓存机制 196
 - 7.3.1 资源可见性与分类 198
 - 7.3.2 分布式缓存实现 200
 - 7.4 目录结构管理 203
 - 7.4.1 数据目录管理 203
 - 7.4.2 日志目录管理 203
 - 7.5 状态机管理 206
 - 7.5.1 Application状态机 207
 - 7.5.2 Container状态机 210
 - 7.5.3 LocalizedResource状态机 213
 - 7.6 Container生命周期剖析 214
 - 7.6.1 Container资源本地化 214
 - 7.6.2 Container运行 218
 - 7.6.3 Container资源清理 222
 - 7.7 资源隔离 224
 - 7.7.1 Cgroups介绍 224
 - 7.7.2 内存资源隔离 228
 - 7.7.3 CPU资源隔离 230
 - 7.8 源代码阅读引导 234
 - 7.9 小结 235
 - 7.10 问题讨论 236
- 第三部分 计算框架篇
- 第8章 离线计算框架MapReduce 238
- 8.1 概述 238
 - 8.1.1 基本构成 238
 - 8.1.2 事件与事件处理器 240
 - 8.2 MapReduce客户端 241
 - 8.2.1 ApplicationClientProtocol协议 242
 - 8.2.2 MRClientProtocol协议 243
 - 8.3 MRAppMaster工作流程 243
 - 8.4 MR作业生命周期及相关状态机 246
 - 8.4.1 MR作业生命周期 246
 - 8.4.2 Job状态机 249
 - 8.4.3 Task状态机 253
 - 8.4.4 TaskAttempt状态机 255
 - 8.5 资源申请与再分配 259
 - 8.5.1 资源申请 259
 - 8.5.2 资源再分配 262
 - 8.6 Container启动与释放 263
 - 8.7 推测执行机制 264
 - 8.7.1 算法介绍 265

<<Hadoop技术内幕>>

- 8.7.2 推测执行相关类 266
- 8.8 作业恢复 267
- 8.9 数据处理引擎 269
- 8.10 历史作业管理器 271
- 8.11 MRv1与MRv2对比 273
 - 8.11.1 MRv1 On YARN 273
 - 8.11.2 MRv1与MRv2架构比较 274
 - 8.11.3 MRv1与MRv2编程接口兼容性 274
- 8.12 源代码阅读引导 275
- 8.13 小结 277
- 8.14 问题讨论 277
- 第9章 DAG计算框架Tez 278
 - 9.1 背景 278
 - 9.2 Tez数据处理引擎 281
 - 9.2.1 Tez编程模型 281
 - 9.2.2 Tez数据处理引擎 282
 - 9.3 DAG Master实现 284
 - 9.3.1 DAG编程模型 284
 - 9.3.2 MR到DAG转换 286
 - 9.3.3 DAGAppMaster 288
 - 9.4 优化机制 291
 - 9.4.1 当前YARN框架存在的问题 291
 - 9.4.2 Tez引入的优化技术 292
 - 9.5 Tez应用场景 292
 - 9.6 与其他系统比较 294
 - 9.7 小结 295
- 第10章 实时/内存计算框架Storm/Spark 296
 - 10.1 Hadoop MapReduce的短板 296
 - 10.2 实时计算框架Storm 296
 - 10.2.1 Storm编程模型 297
 - 10.2.2 Storm基本架构 302
 - 10.2.3 Storm On YARN 304
 - 10.3 内存计算框架Spark 307
 - 10.3.1 Spark编程模型 308
 - 10.3.2 Spark基本架构 312
 - 10.3.3 Spark On YARN 316
 - 10.3.4 Spark/Storm On YARN比较 317
 - 10.4 小结 317
- 第四部分 高级篇
- 第11章 Facebook Corona剖析 320
 - 11.1 概述 320
 - 11.1.1 Corona的基本架构 320
 - 11.1.2 Corona的RPC协议与序列化框架 322
 - 11.2 Corona设计特点 323
 - 11.2.1 推式网络通信模型 323
 - 11.2.2 基于Hadoop 0.20版本 324
 - 11.2.3 使用Thrift 324

<<Hadoop技术内幕>>

- 11.2.4 深度集成Fair Scheduler 324
- 11.3 工作流程介绍 324
 - 11.3.1 作业提交 325
 - 11.3.2 资源申请与任务启动 326
- 11.4 主要模块介绍 327
 - 11.4.1 ClusterManager 327
 - 11.4.2 CoronaJobTracker 330
 - 11.4.3 CoronaTaskTracker 333
- 11.5 小结 335
- 第12章 Apache Mesos剖析 336
 - 12.1 概述 336
 - 12.2 底层网络通信库 337
 - 12.2.1 libprocess基本架构 338
 - 12.2.2 一个简单示例 338
 - 12.3 Mesos服务 340
 - 12.3.1 SchedulerProcess 341
 - 12.3.2 Mesos Master 342
 - 12.3.3 Mesos Slave 343
 - 12.3.4 ExecutorProcess 343
 - 12.4 Mesos工作流程 344
 - 12.4.1 框架注册过程 344
 - 12.4.2 Framework Executor注册过程 345
 - 12.4.3 资源分配到任务运行过程 345
 - 12.4.4 任务启动过程 347
 - 12.4.5 任务状态更新过程 347
 - 12.5 Mesos资源分配策略 348
 - 12.5.1 Mesos资源分配框架 349
 - 12.5.2 Mesos资源分配算法 349
 - 12.6 Mesos容错机制 350
 - 12.6.1 Mesos Master容错 350
 - 12.6.2 Mesos Slave容错 351
 - 12.7 Mesos应用实例 352
 - 12.7.1 Hadoop On Mesos 352
 - 12.7.2 Storm On Mesos 353
 - 12.8 Mesos与YARN对比 354
 - 12.9 小结 355
- 第13章 YARN总结与发展趋势 356
 - 13.1 资源管理系统设计动机 356
 - 13.2 资源管理系统架构演化 357
 - 13.2.1 集中式架构 357
 - 13.2.2 双层调度架构 358
 - 13.2.3 共享状态架构 358
 - 13.3 YARN发展趋势 359
 - 13.3.1 YARN自身的完善 359
 - 13.3.2 以YARN为核心的生态系统 361
 - 13.3.3 YARN周边工具的完善 363
 - 13.4 小结 363

<<Hadoop技术内幕>>

- 附录A YARN安装指南 364
- 附录B YARN配置参数介绍 367
- 附录C Hadoop Shell命令介绍 371
- 附录D 参考资料 374

<<Hadoop技术内幕>>

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>