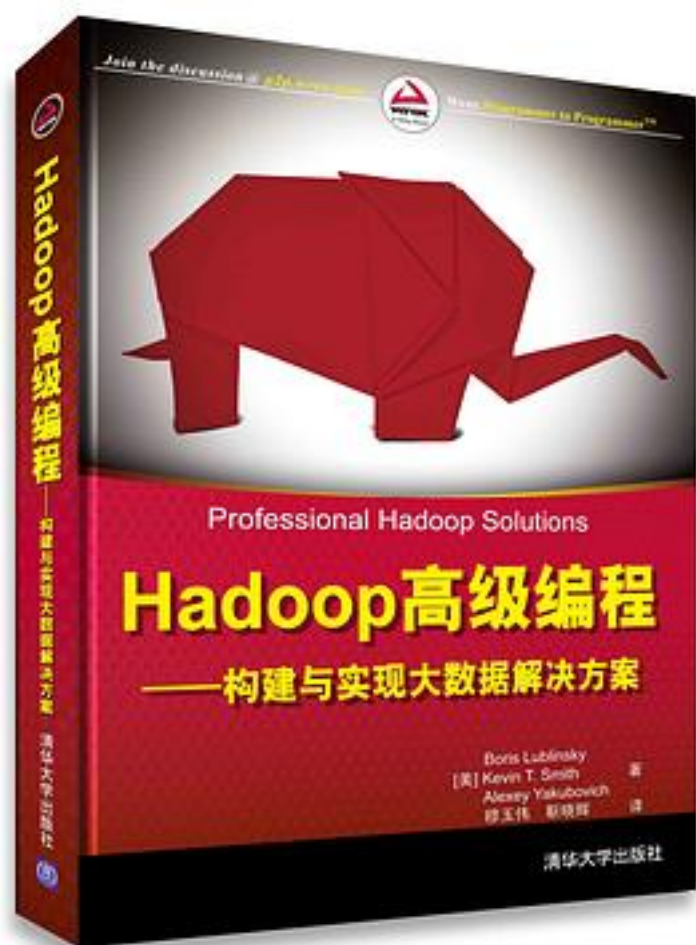


Hadoop高级编程——构建与实现大数据解决方案



[Hadoop高级编程——构建与实现大数据解决方案_下载链接1](#)

著者:(美)卢博林斯凯(Lublinsky, B.)

出版者:清华大学出版社

出版时间:2014-8-1

装帧:平装

isbn:9787302369066

如果你已经准备好要充分实施大规模可扩展性数据分析工作，那么需要知道如何利用Hadoop技术。这本《Hadoop高级编程——构建与实现大数据解决方案》可以帮助你做

到这一点！本书关注用于构建先进的、基于Hadoop的企业级应用的架构和方案，并为实现现实的解决方案提供深入的、代码级的讲解。本书还会带你领略数据设计以及数据设计如何影响实现。本书解释了MapReduce的工作原理，并展示了如何在MapReduce中重新定制特定的业务问题。在整本书中，你将会发现深入的Java代码示例，这些代码示例可以直接使用，它们均源自于已经成功地构建和部署的应用程序。

作者介绍:

Boris Lublinsky是诺基亚的首席架构师，出版了70多篇作品，包括Applied SOA: Service-Oriented Architecture and Design Strategies。

Kevin T. Smith是Novetta Solutions公司AMS部门的技术解决方案总监，他为客户构建高度安全的、面向数据的解决方案。

Alexey Yakubovich是Hortonworks的一名系统架构师，而且是对象管理组织(OMG)关于SOA治理和模型驱动架构的特别兴趣小组(SIG)的一名成员。

目录: 目录

第1章 大数据和Hadoop生态系统 1

1.1 当大数据遇见Hadoop 2

1.1.1 Hadoop：直面大数据的挑战 3

1.1.2 商业世界中的数据科学 4

1.2 Hadoop生态系统 6

1.3 Hadoop核心组件 7

1.4 Hadoop发行版 9

1.5 使用Hadoop开发企业级应用 10

1.6 小结 14

第2章 Hadoop数据存储 15

2.1 HDFS 15

2.1.1 HDFS架构 15

2.1.2 使用HDFS文件 19

2.1.3 Hadoop特定的文件类型 21

2.1.4 HDFS联盟和高可用性 26

2.2 HBase 28

2.2.1 HBase架构 28

2.2.2 HBase结构设计 34

2.2.3 HBase编程 35

2.2.4 HBase新特性 42

2.3 将HDFS和HBase的组合用于高效数据存储 45

2.4 使用Apache Avro 45

2.5 利用HCatalog管理元数据 49

2.6 为应用程序选择合适的Hadoop数据组织形式 51

2.7 小结 53

第3章 使用MapReduce处理数据 55

3.1 了解MapReduce 55

3.1.1 MapReduce执行管道 56

3.1.2 MapReduce中的运行时协调和任务管理 59

3.2 第一个MapReduce应用程序 61

3.3 设计MapReduce实现 69

3.3.1 将MapReduce用作并行处理框架 70

- 3.3.2 使用MapReduce进行简单的数据处理 71
- 3.3.3 使用MapReduce构建连接 72
- 3.3.4 构建迭代式MapReduce应用程序 77
- 3.3.5 是否使用MapReduce 82
- 3.3.6 常见的MapReduce设计陷阱 83
- 3.4 小结 84
- 第4章 自定义MapReduce执行 85
 - 4.1 使用InputFormat控制MapReduce执行 85
 - 4.1.1 为计算密集型应用程序实现InputFormat 87
 - 4.1.2 实现InputFormat以控制Map的数量 93
 - 4.1.3 实现用于多个HBase表的InputFormat 99
 - 4.2 使用自定义RecordReader以自己的方式读取数据 102
 - 4.2.1 实现基于队列的RecordReader 102
 - 4.2.2 为XML数据实现RecordReader 105
 - 4.3 使用自定义输出格式组织输出数据 109
 - 4.4 使用自定义记录写入器以自己的方式写入数据 119
 - 4.5 使用组合器优化MapReduce执行 121
 - 4.6 使用分区器控制Reducer执行 124
 - 4.7 在Hadoop中使用非Java代码 128
 - 4.7.1 Pipes 128
 - 4.7.2 Hadoop Streaming 128
 - 4.7.3 使用JNI 129
 - 4.8 小结 131
- 第5章 构建可靠的MapReduce应用程序 133
 - 5.1 单元测试MapReduce应用程序 133
 - 5.1.1 测试Mapper 136
 - 5.1.2 测试Reducer 137
 - 5.1.3 集成测试 138
 - 5.2 使用Eclipse进行本地应用程序测试 139
 - 5.3 将日志用于Hadoop测试 141
 - 5.4 使用作业计数器报告指标 146
 - 5.5 MapReduce中的防御性编程 149
 - 5.6 小结 151
- 第6章 使用Oozie自动化数据处理 153
 - 6.1 认识Oozie 154
 - 6.2 Oozie Workflow 155
 - 6.2.1 在Oozie Workflow中执行异步操作 159
 - 6.2.2 Oozie的恢复能力 164
 - 6.2.3 Oozie Workflow作业的生命周期 164
 - 6.3 Oozie Coordinator 165
 - 6.4 Oozie Bundle 170
 - 6.5 用表达式语言对Oozie进行参数化 174
 - 6.5.1 Workflow函数 175
 - 6.5.2 Coordinator函数 175
 - 6.5.3 Bundle函数 175
 - 6.5.4 其他EL函数 175
 - 6.6 Oozie作业执行模型 176
 - 6.7 访问Oozie 179
 - 6.8 Oozie SLA 180
 - 6.9 小结 185
- 第7章 使用Oozie 187
 - 7.1 使用探测包验证位置相关信息正确性 187
 - 7.2 设计基于探测包的地点正确性验证 188
 - 7.3 设计Oozie Workflow 190

- 7.4 实现Oozie Workflow应用程序 193
 - 7.4.1 实现数据准备Workflow 193
 - 7.4.2 实现考勤指数和聚类探测包串Workflow 201
- 7.5 实现 Workflow行为 203
 - 7.5.1 发布来自java动作的执行上下文 204
 - 7.5.2 在Oozie Workflow中使用MapReduce作业 204
- 7.6 实现Oozie Coordinator应用程序 207
- 7.7 实现Oozie Bundle应用程序 212
- 7.8 部署、测试和执行Oozie应用程序 213
 - 7.8.1 部署Oozie应用程序 213
 - 7.8.2 使用Oozie CLI执行Oozie应用程序 215
 - 7.8.3 向Oozie作业传递参数 218
- 7.9 使用Oozie控制台获取Oozie应用程序信息 221
 - 7.9.1 了解Oozie控制台界面 221
 - 7.9.2 获取 Coordinator作业信息 225
- 7.10 小结 227
- 第8章 高级Oozie特性 229
 - 8.1 构建自定义Oozie Workflow动作 230
 - 8.1.1 实现自定义Oozie Workflow动作 230
 - 8.1.2 部署Oozie自定义Workflow动作 235
 - 8.2 向Oozie Workflow添加动态执行 237
 - 8.2.1 总体实现方法 237
 - 8.2.2 一个机器学习模型、参数和算法 240
 - 8.2.3 为迭代过程定义Workflow 241
 - 8.2.4 动态Workflow生成 244
 - 8.3 使用Oozie Java API 247
 - 8.4 在Oozie应用中使用uber jar包 251
 - 8.5 数据吸收传送器 256
 - 8.6 小结 263
- 第9章 实时Hadoop 265
 - 9.1 现实世界中的实时应用 266
 - 9.2 使用HBase来实现实时应用 266
 - 9.2.1 将HBase用作图片管理系统 268
 - 9.2.2 将HBase用作Lucene后端 275
 - 9.3 使用专门的实时Hadoop查询系统 295
 - 9.3.1 Apache Drill 296
 - 9.3.2 Impala 298
 - 9.3.3 实时查询和MapReduce的对比 299
 - 9.4 使用基于Hadoop的事件处理系统 300
 - 9.4.1 HFlame 301
 - 9.4.2 Storm 302
 - 9.4.3 事件处理和MapReduce的对比 305
 - 9.5 小结 305
- 第10章 Hadoop安全 307
 - 10.1 简要的历史：理解Hadoop安全的挑战 308
 - 10.2 认证 309
 - 10.2.1 Kerberos认证 310
 - 10.2.2 委派安全凭据 318
 - 10.3 授权 323
 - 10.3.1 HDFS文件访问权限 323
 - 10.3.2 服务级授权 327
 - 10.3.3 作业授权 329
 - 10.4 Oozie认证和授权 329
 - 10.5 网络加密 331

- 10.6 使用Rhino项目增强安全性 332
 - 10.6.1 HDFS磁盘级加密 333
 - 10.6.2 基于令牌的认证和统一的授权框架 333
 - 10.6.3 HBase单元格级安全 334
- 10.7 将所有内容整合起来——保证Hadoop安全的最佳实践 334
 - 10.7.1 认证 335
 - 10.7.2 授权 335
 - 10.7.3 网络加密 336
 - 10.7.4 敬请关注Hadoop的增强功能 336
- 10.8 小结 336
- 第11章 在AWS上运行Hadoop应用 337
 - 11.1 初识AWS 338
 - 11.2 在AWS上运行Hadoop的可选项 339
 - 11.2.1 使用EC2实例的自定义安装 339
 - 11.2.2 弹性MapReduce 339
 - 11.2.3 做出选择前的额外考虑 339
 - 11.3 理解EMR-Hadoop的关系 340
 - 11.3.1 EMR架构 341
 - 11.3.2 使用S3存储 343
 - 11.3.3 最大化EMR的使用 343
 - 11.3.4 利用CloudWatch和其他AWS组件 345
 - 11.3.5 访问和使用EMR 346
 - 11.4 使用AWS S3 351
 - 11.4.1 理解桶的使用 352
 - 11.4.2 使用控制台浏览内容 354
 - 11.4.3 在S3中编程访问文件 355
 - 11.4.4 使用MapReduce上传多个文件到S3 365
 - 11.5 自动化EMR作业流创建和作业执行 367
 - 11.6 管理EMR中的作业执行 372
 - 11.6.1 在EMR集群上使用Oozie 372
 - 11.6.2 AWS 简单工作流 374
 - 11.6.3 AWS数据管道 375
 - 11.7 小结 376
- 第12章 为Hadoop实现构建企业级安全解决方案 377
 - 12.1 企业级应用的安全顾虑 378
 - 12.1.1 认证 380
 - 12.1.2 授权 380
 - 12.1.3 保密性 380
 - 12.1.4 完整性 381
 - 12.1.5 审计 381
 - 12.2 Hadoop安全没有为企业级应用原生地提供哪些机制 381
 - 12.2.1 面向数据的访问控制 382
 - 12.2.2 差分隐私 382
 - 12.2.3 加密静止的数据 383
 - 12.2.4 企业级安全集成 384
 - 12.3 保证使用Hadoop的企业级应用安全的方法 384
 - 12.3.1 使用Accumulo进行访问控制保护 385
 - 12.3.2 加密静止数据 394
 - 12.3.3 网络隔离和分隔方案 395
 - 12.4 小结 397
- 第13章 Hadoop的未来 399
 - 13.1 使用DSL简化MapReduce编程 400
 - 13.1.1 什么是DSL 400
 - 13.1.2 Hadoop的DSL 401

- 13.2 更快、更可扩展的数据处理 412
 - 13.2.1 Apache YARN 412
 - 13.2.2 Tez 414
- 13.3 安全性的改进 415
- 13.4 正在出现的趋势 415
- 13.5 小结 416
- 附录 有用的阅读 417
 - • • • • [\(收起\)](#)

[Hadoop高级编程——构建与实现大数据解决方案_下载链接1](#)

标签

Hadoop

大数据

计算机

云计算

hadoop

计算科学

计算机技术

纸书

评论

很一般

hadoop实战必读

[Hadoop高级编程——构建与实现大数据解决方案_下载链接1_](#)

书评

[Hadoop高级编程——构建与实现大数据解决方案_下载链接1_](#)