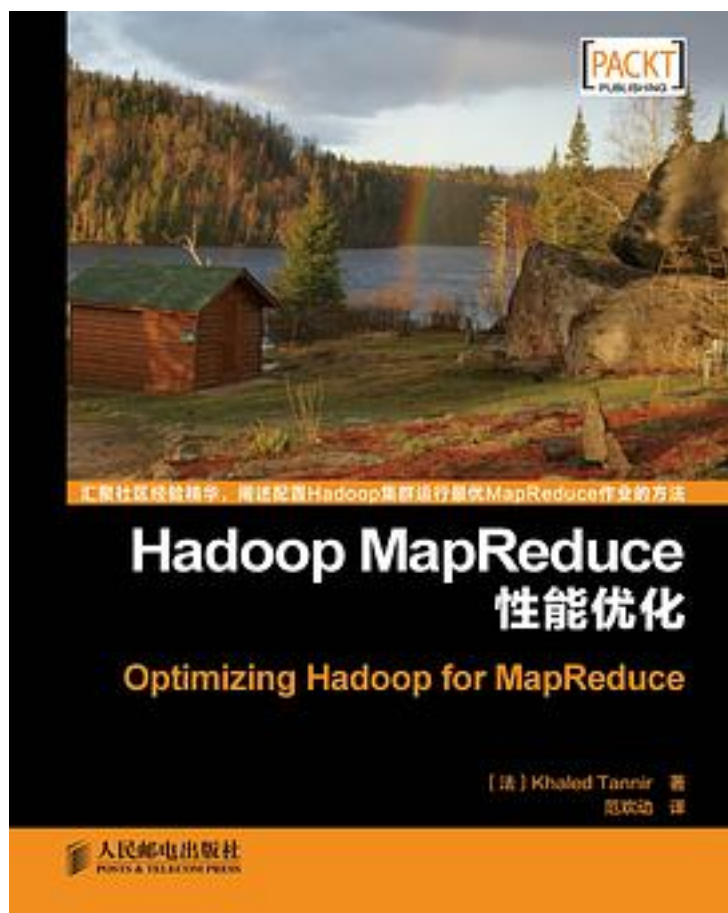


Hadoop MapReduce性能优化



[Hadoop MapReduce性能优化_下载链接1](#)

著者:[法]卡勒德·坦尼尔 (Khaled Tannir)

出版者:人民邮电出版社

出版时间:2015-3

装帧:

isbn:9787115381279

大数据时代，MapReduce的重要性不言而喻。Hadoop作为MapReduce框架的一个实现，受到业界广泛的认同，并被广泛部署和应用。尽管Hadoop为数据开发工程师入门和编程提供了极大便利，但构造一个真正满足性能要求的MapReduce程序并不简单。数据量巨大是大数据工作的现实问题，而对低响应时间的要求则时常困扰着数据开发工程师。

师。

本书采用原理与实践相结合的方式，通过原理讲解影响MapReduce性能的因素，透过实例一步步地教读者如何发现性能瓶颈并消除瓶颈，如何识别系统薄弱环节并改善薄弱环节，讲解过程中融合了作者在优化实践过程中积累的丰富经验，具有很强的针对性。读完本书，能让读者对Hadoop具有更强的驾驭能力，从而构造出性能最优的MapReduce程序。

Hadoop性能问题既是程序层面的问题，也是系统层面的问题。本书既覆盖了系统层面的优化又覆盖了程序层面的优化，非常适合Hadoop管理员和有经验的数据开发工程师阅读。对于初学者，本书第1章也作了必要的技术铺垫，避免对后面章节的理解产生梯度。

作者介绍:

作者介绍

Khaled

Tannirhas从1980年开始从事计算机相关工作。他是微软认证的开发人员（MCSD），他在领导软件解决方案的开发和实施以及技术演说方面，拥有20多年技术经验。如今，他是一名独立IT咨询师，并在法国、加拿大的许多大公司担任基础设施工程师、高级研发工程师、企业/解决方案架构师等职务。

他在Microsoft .NET、Microsoft服务器系统、Oracle Java技术等领域拥有丰富的经验，并且熟练驾驭在线和离线应用系统设计、系统转换以及多语言的互联网/桌面应用程序开发。

Khaled

Tnnirhas总是热衷于探索和学习新的技术，并基于这些技术在法国、北美、中东等地区寻求商机。他现在拥有一个IT电子实验室，实验室中配备了很多服务器、监控器、开源电子板（如Arduino、Netduino、RaspBerry Pi和.Net Gadteer），还有一些装有Windows Phone、Android和iOS操作系统的智能设备。

2012年，他协助组织并出席了法国波尔多大学的复杂数据挖掘国际论坛——EGC 2012。

他还是《RavenDB 2.x Beginner's Guide》一书的作者。

译者简介

范欢动

信息技术领域的一名老兵。1994年获得电力系统及其自动化专业学士学位。1997年获得信号与信息处理专业硕士学位。兴趣广泛，涉及技术领域和非技术领域，但大多浅尝辄止。参与过航天、通信、电子出版和金融领域的信息处理、软件架构与开发，以第一作者或合作者身份拥有三项实用新型专利，并获得第十八届北京市优秀青年工程师称号。曾在佳讯飞鸿电气股份有限公司、英国雅讯（xarios）北京代表处、博云科技等公司担任产品技术总监。目前，终于把全部的兴趣聚焦到了数据分析，尤其是大数据分析上，主要目标是利用数据分析手段发现模式，从而降低决策风险、改善产业价值链。冀望在大数据时代与读者一起学习并分享大数据相关技术，共同发掘大数据带来的价值。

目录: 目录

第1章 了解Hadoop MapReduce 1

1.1 MapReduce模型 1

1.2 Hadoop MapReduce概述	3
1.3 Hadoop MapReduce的工作原理	4
1.4 影响MapReduce性能的因素	5
1.5 小结	8
第2章 Hadoop参数概述	9
2.1 研究Hadoop参数	9
2.1.1 配置文件mapred-site.xml	10
2.1.2 配置文件hdfs-site.xml	15
2.1.3 配置文件core-site.xml	18
2.2 Hadoop MapReduce性能指标	19
2.3 性能监测工具	20
2.3.1 用Chukwa监测Hadoop	21
2.3.2 使用Ganglia监测Hadoop	21
2.3.3 使用Nagios监测Hadoop	21
2.4 用Apache Ambari监测Hadoop	22
2.5 小结	23
第3章 检测系统瓶颈	25
3.1 性能调优	25
3.2 创建性能基线	27
3.3 识别资源瓶颈	30
3.3.1 识别内存瓶颈	30
3.3.2 识别CPU瓶颈	31
3.3.3 识别存储瓶颈	32
3.3.4 识别网络带宽瓶颈	33
3.4 小结	34
第4章 识别资源薄弱环节	35
4.1 识别集群薄弱环节	35
4.1.1 检查Hadoop集群节点的健康状况	36
4.1.2 检查输入数据大小	37
4.1.3 检查海量I/O和网络阻塞	38
4.1.4 检查并发任务不足	39
4.1.5 检查CPU过饱和	40
4.2 量化Hadoop集群	41
4.3 正确配置集群	44
4.4 小结	47
第5章 强化map和reduce任务	49
5.1 强化map任务	49
5.1.1 输入数据和块大小的影响	51
5.1.2 处置小文件和不可拆分文件	51
5.1.3 在Map阶段压缩溢写记录	53
5.1.4 计算map任务的吞吐量	55
5.2 强化reduce任务	57
5.2.1 计算reduce任务的吞吐量	58
5.2.2 改善Reduce执行阶段	59
5.3 调优map和reduce参数	60
5.4 小结	64
第6章 优化MapReduce任务	65
6.1 使用Combiner	65
6.2 使用压缩技术	68
6.3 使用正确Writable类型	72
6.4 明智地复用类型	74
6.5 优化mapper和reducer的代码	76
6.6 小结	78
第7章 最佳实践与建议	81

- 7.1 硬件调优与操作系统推荐 81
 - 7.1.1 Hadoop集群检查表 81
 - 7.1.2 Bios调优检查表 82
 - 7.1.3 OS配置建议 82
- 7.2 Hadoop最佳实践与建议 83
 - 7.2.1 部署Hadoop 83
 - 7.2.2 Hadoop调优建议 84
 - 7.2.3 使用MapReduce模板类代码 86
- 7.3 小结 90
 - • • • • (收起)

[Hadoop MapReduce性能优化_下载链接1](#)

标签

Hadoop

编程

程序设计

性能优化

性能

大数据

MapReduce

评论

虽然是讲的Hadoop1，不过仍旧给我一些启发。从mapreduce的每个阶段去分析和优化是一个很好的方法

一本只有90页小册子，能获得一些启发。

错误： P67第二行 应该是 Map output records P53 图5-3 percent拼错 P54 表5-1 tasktracker拼错 p62 表5-3 reducer数这列26.25≈26

[Hadoop MapReduce性能优化_下载链接1](#)

书评

[Hadoop MapReduce性能优化_下载链接1](#)