

《hadoop技术详解》

图书基本信息

书名：《hadoop技术详解》

13位ISBN编号：9787115333322

10位ISBN编号：7115333327

出版时间：2013-11-27

出版社：人民邮电出版社

作者：Eric Sammer

页数：250

译者：刘敏,麦耀锋,李冀蕾等

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《hadoop技术详解》

内容概要

本书将向读者详细介绍Hadoop的各项操作，从最初的设计，到安装、设置，以帮助读者提供稳定持续的系统表现。而对于那些希望通过学习Hadoop工作原理以提高MapReduce工作效率的开发者来说，也将会从本书收益。

作者简介

Eric Sammer目前是Cloudera公司的首席方案架构师，协助客户规划、配置、开发和使用Hadoop以及相关的大型项目。他在开发和运营分布式的、高并发的数据摄取和处理系统方面很有经验。在过去十年里，他参加了开源社区并且为许多项目做出了贡献。

书籍目录

第1章 简介

1

第2章 HDFS

6

2.1 目标和动机

6

2.2 设计

7

2.3 守护进程

8

2.4 读写数据

10

2.4.1 数据读取流程

10

2.4.2 数据写操作流程

11

2.5 管理文件系统元数据

13

2.6 NameNode的高可用性

14

2.7 NameNode联盟

16

2.8 访问与集成

17

2.8.1 命令行工具

18

2.8.2 用户空间文件系统（FUSE）

21

2.8.3 表示状态传输（REST）的支持

21

第3章 MapReduce

23

3.1 MapReduce的若干阶段

24

3.2 Hadoop MapReduce简介

30

3.2.1 后台程序

31

3.2.2 出错处理

33

3.3 YARN

35

第4章 规划一个Hadoop集群

37

4.1 挑选Hadoop的发行版本

37

4.1.1 Apache Hadoop

37

4.1.2 Cloudera的Apache Hadoop发行版本

38

4.1.3 版本和功能

38

4.1.4 我应该使用哪个版本

40

4.2 硬件选型

41

4.2.1 主节点硬件的选择

42

4.2.2 工作节点的硬件选择

43

4.2.3 集群的大小

45

4.2.4 刀片服务器、存储区域网络 (SAN) 和虚拟化

47

4.3 操作系统的选择和准备

49

4.3.1 部署规划

49

4.3.2 软件

50

4.3.3 主机名、DNS和标识

51

4.3.4 用户、组和特权

54

4.4 内核调整

56

4.4.1 vm.swappiness

56

4.4.2 vm.overcommit_memory

57

4.5 磁盘配置

58

4.5.1 选择文件系统

58

4.5.2 挂载选项

60

4.6 网络设计

60

4.6.1 Hadoop中的网络使用：回顾

60

4.6.2 1 Gb与10 Gb网络

62

4.6.3 典型的网络拓扑

63

第5章 安装和配置

67

5.1	安装Hadoop	67
5.1.1	Apache Hadoop	68
5.1.2	CDH	72
5.2	配置概述	76
5.3	环境变量和Shell脚本	80
5.4	日志配置	82
5.5	HDFS	84
5.5.1	识别和定位	84
5.5.2	优化与调整	86
5.5.3	格式化NameNode	89
5.5.4	创建/tmp目录	91
5.6	NameNode的高可靠性	92
5.6.1	隔离 (Fencing) 选项	93
5.6.2	基本配置	95
5.6.3	自动失效备援配置	96
5.6.4	格式化和引导NameNode启动	99
5.7	NameNode联盟 (Federation)	105
5.8	MapReduce	113
5.8.1	识别和定位	113
5.8.2	优化和调整	115
5.9	机架拓扑	122
5.10	安全	125
第6章	用户标识、身份验证和授权	126
6.1	用户标识	127
6.2	Kerberos和Hadoop	

128	
6.2.1	Kerberos
128	
6.2.2	Hadoop上的Kerberos支持
130	
6.3	授权
143	
6.3.1	HDFS
144	
6.3.2	MapReduce
146	
6.3.3	其他工具和系统
149	
6.4	集成试试
153	
第7章	资源管理
156	
7.1	何谓资源管理
156	
7.2	HDFS配额
156	
7.3	MapReduce 调度器
159	
7.3.1	先进先出 (FIFO) 调度器
160	
7.3.2	公平调度器
162	
7.3.3	计算能力调度器 (Capacity Scheduler)
174	
7.3.4	未来发展
181	
第8章	集群维护
183	
8.1	Hadoop流程管理
183	
8.1.1	用初始化脚本管理进程
183	
8.1.2	手动管理进程
184	
8.2	HDFS维护任务
184	
8.2.1	添加一个DataNode
184	
8.2.2	卸载DataNode
185	
8.2.3	用fsck来检查文件系统的一致性
185	
8.2.4	HDFS块数据均衡
190	

8.2.5	处理坏磁盘	192
8.3	MapReduce维护任务	193
8.3.1	添加tasktracker	193
8.3.2	卸载tasktracker	193
8.3.3	终结MapReduce 作业	194
8.3.4	终结MapReduce任务	194
8.3.5	处理列入黑名单的tasktracker	195
第9章	故障分析与排查	196
9.1	鉴别诊断 (Differential Diagnosis)	196
9.2	故障和问题	197
9.2.1	人类 (自己)	198
9.2.2	配置错误	198
9.2.3	硬件故障	199
9.2.4	资源枯竭	200
9.2.5	主机标识和命名	200
9.2.6	网络分区	200
9.3	“ 计算机插好了么？ ”	201
9.4	治疗和护理	203
9.5	实战案例	206
9.5.1	神秘的瓶颈	206
9.5.2	127.0.0.1这个地址不存在	209
第10章	监控	213
10.1	概览	213
10.2	Hadoop度量(Metrics)	214
10.2.1	Apache Hadoop 0.20.0和CDH3 (metrics1)	

214	
10.2.2	Apache Hadoop 0.20.203及之后的版本、CDH4(metrics2)
221	
10.2.3	SNMP
222	
10.3	健康监控
222	
10.3.1	主机级别的检查
223	
10.3.2	所有Hadoop进程
225	
10.3.3	HDFS检查
226	
10.3.4	MapReduce检查
229	
第11章	备份与恢复
232	
11.1	数据备份
232	
11.1.1	分布式拷贝 (distcp)
233	
11.1.2	并行提取数据
235	
11.2	NameNode元数据
237	
附录	弃用的配置属性
239	

精彩短评

- 1、《Hadoop配置参数详解》
- 2、翻了下，不能算详解，毕竟薄薄的一本

精彩书评

1、适合初学者，在运维的角度讲解了集群的规划和部署，说句真心话讲得比较浅，不过对于没有运维经验的人还是有一定思路的开阔（主要是看第四章就可以了，这章是本书的特色）。本书看完的收获是：如何规划一个Hadoop集群：1.通过数据增长量和作业完成时间来规划集群规模2.硬件选型NOTE：SAN,NAS这种集中存储不适合hadoop.3.网络拓扑的规划4.操作系统目录权限的管理防止，更加安全的控制后台进程内存参数的调优：noatime,降低swap5.hadoop版本的选型hadoop集群的部署：这方面很多书都讲到了，主要说一下在配置优化方面的东西：1.mapred.local.dir和datanode.local.dir共享磁盘要设置du.reserved防止数据盘被task的临时文件写满，要为mr预留20%-30%的空间2.master后台rpcserver线程的增加3.数据传输带宽的限制4.以及mapreduce相关的调优，其中一点我还真不知道mapre.child.ulimit可以限制进程内存5.机架感知

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:www.tushu000.com