<<Hadoop实战手册>>

图书基本信息

书名：<<Hadoop实战手册>>

13位ISBN编号：9787115337955

10位ISBN编号：7115337950

出版时间：2014-3

出版时间：人民邮电出版社

作者：(美)Jonathan R. Owens　Jon Lentz　Brian Femiano

译者：傅杰 赵磊 卢学裕

<<Hadoop实战手册>>

内容概要

这是一本hadoop实用手册，主要针对实际问题给出相应的解决方案。
《hadoop实战手册》特色是以实践结合理论分析，手把手教读者如何操作，并且对每个操作都做详细的解释，对一些重要的知识点也做了必要的拓展。
全书共包括3个部分，第一部分为基础篇，主要介绍hadoop数据导入导出、hdfs的概述、pig与hive的使用、etl和简单的数据处理，还介绍了mapreduce的调试方式；第二部分为数据分析高级篇，主要介绍高级聚合、大数据分析等技巧；第三部分为系统管理篇，主要介绍hadoop的部署的各种模式、添加新节点、退役节点、快速恢复、mapreduce调优等。

《hadoop实战手册》适合各个层次的hadoop技术人员阅读。
通过阅读《hadoop实战手册》，hadoop初学者可以使用hadoop来进行数据处理，hadoop工程师或者数据挖掘工程师可以解决复杂的业务分析，hadoop系统管理员可以更好地进行日常运维。
《hadoop实战手册》也可作为一本hadoop技术手册，针对要解决的相关问题，在工作中随时查阅。

作者简介

jonathan r. owens：软件工程师，拥有java和c++技术背景，最近主要从事hadoop及相关分布式处理技术工作。
目前就职于comscore公司，为核心数据处理团队成员。
comscore是一家知名的从事数字测量与分析的公司，公司使用hadoop及其他定制的分布式系统对数据进行聚合、分析和管理，每天处理超过400亿单的交易。

jon lentz：comscore核心数据处理团队软件工程师。
他更倾向于使用pig脚本来解决问题。
在加入comscore之前，他主要开发优化供应链和分配固定收益证券的软件。

brian femiano：本科毕业于计算机科学专业，并且从事相关专业软件开发工作6年，最近两年主要利用hadoop构建高级分析与大数据存储。
他拥有商业领域的相关经验，以及丰富的政府合作经验。
他目前就职于potomac fusion公司，这家公司主要从事可扩展算法的开发，并致力于学习并改进政府领域中最先进和最复杂的数据集。
他通过教授课程和会议培训在公司内部普及hadoop和云计算相关的技术。

傅杰，硕士，毕业于清华大学高性能所，现就职于优酷土豆集团，任数据平台架构师，负责集团大数据基础平台建设，支撑其他团队的存储与计算需求，包含hadoop基础平台、日志采集系统、实时计算平台、消息系统、天机镜系统等。
个人专注于大数据基础平台架构及安全研究，积累了丰富的平台运营经验，擅长hadoop平台性能调优、jvm调优及诊断各种mapreduce作业，还担任china hadoop submit 2013大会专家委员、优酷土豆大数据系列课程策划&讲师、easyhadoop社区讲师。

赵磊，硕士，毕业于中国科学技术大学，现就职于优酷土豆集团，任数据挖掘算法工程师，负责集团个性化推荐和无线消息推送系统的搭建和相关算法的研究。
个人专注于基于大数据的推荐算法的研究与应用，积累了丰富的大数据分析与数据挖掘的实践经验，对分布式计算和海量数据处理有深刻的认识。

卢学裕，硕士，毕业于武汉大学，曾供职腾讯公司即通部门，现就职于优酷土豆集团，担任大数据技术负责人，负责优酷土豆集团大数据系统平台、大数据分析、数据挖掘和推荐系统。
有丰富的hadoop平台使用及优化经验，尤其擅长mapreduce的性能优化。
基于hadoop生态系统构建了优酷土豆的推荐系统，bi分析平台。

<<Hadoop实战手册>>

书籍目录

《hadoop实战手册》
第1章　hadoop分布式文件系统——导入和导出数据　1
1.1　介绍　1
1.2　使用hadoop shell命令导入和导出数据到hdfs　2
1.3　使用distcp实现集群间数据复制　7
1.4　使用sqoop从mysql数据库导入数据到hdfs　9
1.5　使用sqoop从hdfs导出数据到mysql　12
1.6　配置sqoop以支持sql server　15
1.7　从hdfs导出数据到mongodb　17
1.8　从mongodb导入数据到hdfs　20
1.9　使用pig从hdfs导出数据到mongodb　23
1.10　在greenplum外部表中使用hdfs　24
1.11　利用flume加载数据到hdfs中　26
第2章　hdfs　28
2.1　介绍　28
2.2　读写hdfs数据　29
2.3　使用lzo压缩数据　31
2.4　读写序列化文件数据　34
2.5　使用avro序列化数据　37
2.6　使用thrift序列化数据　41
.2.7　使用protocol buffers序列化数据　44
2.8　设置hdfs备份因子　48
2.9　设置hdfs块大小　49
第3章　抽取和转换数据　51
3.1　介绍　51
3.2　使用mapreduce将apache日志转换为tsv格式　52
3.3　使用apache pig过滤网络服务器日志中的爬虫访问量　54
3.4　使用apache pig根据时间戳对网络服务器日志数据排序　57
3.5　使用apache pig对网络服务器日志进行会话分析　59
3.6　通过python扩展apache pig的功能　61
3.7　使用mapreduce及二次排序计算页面访问量　62
3.8　使用hive和python清洗、转换地理事件数据　67
3.9　使用python和hadoop streaming执行时间序列分析　71
3.10　在mapreduce中利用multipleoutputs输出多个文件　75
3.11　创建用户自定义的hadoop writable及inputformat读取地理事件数据　78
第4章　使用hive、pig和mapreduce处理常见的任务　85
4.1　介绍　85
4.2　使用hive将hdfs中的网络日志数据映射为外部表　86
4.3　使用hive动态地为网络日志查询结果创建hive表　87
4.4　利用hive字符串udf拼接网络日志数据的各个字段　89
4.5　使用hive截取网络日志的ip字段并确定其对应的国家　92
4.6　使用mapreduce对新闻档案数据生成n-gram　94
4.7　通过mapreduce使用分布式缓存查找新闻档案数据中包含关键词的行　98
4.8　使用pig加载一个表并执行包含group by的select操作　102
第5章　高级连接操作　104

5.1 介绍 104
5.2 使用mapreduce对数据进行连接 104
5.3 使用apache pig对数据进行复制连接 108
5.4 使用apache pig对有序数据进行归并连接 110
5.5 使用apache pig对倾斜数据进行倾斜连接 111
5.6 在apache hive中通过map端连接对地理事件进行分析 113
5.7 在apache hive通过优化的全外连接分析地理事件数据 115
5.8 使用外部键值存储(redis)连接数据 118
第6章 大数据分析 123
6.1 介绍 123
6.2 使用mapreduce和combiner统计网络日志数据集中的独立ip数 124
6.3 运用hive日期udf对地理事件数据集中的时间日期进行转换与排序 129
6.4 使用hive创建基于地理事件数据的每月死亡报告 131
6.5 实现hive用户自定义udf用于确认地理事件数据的来源可靠性 133
6.6 使用hive的map/reduce操作以及python标记最长的无暴力发生的时间区间 136
6.7 使用pig计算audioscrobbler数据集中艺术家之间的余弦相似度 141
6.8 使用pig以及datafu剔除audioscrobbler数据集中的离群值 145
第7章 高级大数据分析 147
7.1 介绍 147
7.2 使用apache giraph计算pagerank 147
7.3 使用apache giraph计算单源最短路径 150
7.4 使用apache giraph执行分布式宽度优先搜索 158
7.5 使用apache mahout计算协同过滤 165
7.6 使用apache mahout进行聚类 168
7.7 使用apache mahout进行情感分类 171
第8章 调试 174
8.1 介绍 174
8.2 在mapreduce中使用counters监测异常记录 174
8.3 使用mrunit开发和测试mapreduce 177
8.4 本地模式下开发和测试mapreduce 179
8.5 运行mapreduce作业跳过异常记录 182
8.6 在流计算作业中使用counters 184
8.7 更改任务状态显示调试信息 185
8.8 使用illustrate调试pig作业 187
第9章 系统管理 189
9.1 介绍 189
9.2 在伪分布模式下启动hadoop 189
9.3 在分布式模式下启动hadoop 192
9.4 添加一个新节点 195
9.5 节点安全退役 197
9.6 namenode故障恢复 198
9.7 使用ganglia监控集群 199
9.8 mapreduce作业参数调优 201
第10章 使用apache accumulo进行持久化 204
10.1 介绍 204
10.2 在accumulo中设计行键存储地理事件 205
10.3 使用mapreduce批量导入地理事件数据到accumulo 213

10.4 设置自定义字段约束accumulo中的地理事件数据 220
10.5 使用正则过滤器限制查询结果 225
10.6 使用sumcombiner计算同一个键的不同版本的死亡数总和 228
10.7 使用accumulo实行单元级安全的扫描 232
10.8 使用mapreduce聚集accumulo中的消息源 237

<<Hadoop实战手册>>

版权说明