

《精通Hadoop》

图书基本信息

书名：《精通Hadoop》

13位ISBN编号：9787115411050

出版时间：2016-1

作者：[印] Sandeep Karanth

页数：268

译者：刘 淼,唐凯隽,陈智威

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

内容概要

本书是一本循序渐进的指导手册，重点介绍了Hadoop的高级概念和特性。内容涵盖了Hadoop 2.X版的改进，MapReduce、Pig和Hive等的优化及其高级特性，Hadoop 2.0的专属特性（如YARN和HDFS联合），以及如何使用Hadoop 2.0版本扩展Hadoop的能力。

如果你想拓展自己的Hadoop知识和技能，想应对具有挑战性的数据处理问题，想让Hadoop作业、Pig脚本和Hive查询运行得更快，或者想了解升级Hadoop的好处，那么本书便是你的不二选择。

通过阅读本书，你将能够：

- 理解从Hadoop 1.0到Hadoop 2.0的变化

- 定制和优化Hadoop 2.0中的MapReduce作业

- 探究Hadoop I/O和不同的数据格式

- 深入学习YARN和Storm，并通过YARN集成Hadoop和Storm

- 基于亚马逊Elastic MapReduce部署Hadoop

- 探究HDFS替代品，学习HDFS联合

- 掌握Hadoop安全方面的主要内容

- 使用Mahout和RHadoop进行Hadoop数据分析

作者简介

Sandeep Karanth

Scibler公司联合创始人，负责数据智能产品的架构；DataPhi Labs公司联合创始人兼首席架构师，专注于构建和实施软件系统。他拥有14年以上的软件行业从业经验，既设计过企业数据应用，也开发过新一代移动应用。他曾就职于微软总部和微软印度研究院。他的Twitter账号是@karanths，GitHub账号是<https://github.com/Karanth>。

书籍目录

第1章 Hadoop 2.X	1
1.1 Hadoop的起源	1
1.2 Hadoop的演进	2
1.3 Hadoop 2.X	6
1.3.1 Yet Another Resource Negotiator (YARN)	7
1.3.2 存储层的增强	8
1.3.3 支持增强	11
1.4 Hadoop的发行版	11
1.4.1 选哪个Hadoop发行版	12
1.4.2 可用的发行版	14
1.5 小结	16
第2章 MapReduce进阶	17
2.1 MapReduce输入	18
2.1.1 InputFormat类	18
2.1.2 InputSplit类	18
2.1.3 RecordReader类	19
2.1.4 Hadoop的“小文件”问题	20
2.1.5 输入过滤	24
2.2 Map任务	27
2.2.1 dfs.blocksize属性	28
2.2.2 中间输出结果的排序与溢出	28
2.2.3 本地reducer和Combiner	31
2.2.4 获取中间输出结果——Map侧	31
2.3 Reduce任务	32
2.3.1 获取中间输出结果——Reduce侧	32
2.3.2 中间输出结果的合并与溢出	33
2.4 MapReduce的输出	34
2.5 MapReduce作业的计数器	34
2.6 数据连接的处理	36
2.6.1 Reduce侧的连接	36
2.6.2 Map侧的连接	42
2.7 小结	45
第3章 Pig进阶	47
3.1 Pig对比SQL	48
3.2 不同的执行模式	48
3.3 Pig的复合数据类型	49
3.4 编译Pig脚本	50
3.4.1 逻辑计划	50
3.4.2 物理计划	51
3.4.3 MapReduce计划	52
3.5 开发和调试助手	52
3.5.1 DESCRIBE命令	52
3.5.2 EXPLAIN命令	53
3.5.3 ILLUSTRATE命令	53
3.6 Pig操作符的高级特性	54
3.6.1 FOREACH操作符进阶	54
3.6.2 Pig的特殊连接	58

3.7	用户定义函数	61
3.7.1	运算函数	61
3.7.2	加载函数	66
3.7.3	存储函数	68
3.8	Pig的性能优化	69
3.8.1	优化规则	69
3.8.2	Pig脚本性能的测量	71
3.8.3	Pig的Combiner	72
3.8.4	Bag数据类型的内存	72
3.8.5	Pig的reducer数量	72
3.8.6	Pig的multiquery模式	73
3.9	最佳实践	73
3.9.1	明确地使用类型	74
3.9.2	更早更频繁地使用投影	74
3.9.3	更早更频繁地使用过滤	74
3.9.4	使用LIMIT操作符	74
3.9.5	使用DISTINCT操作符	74
3.9.6	减少操作	74
3.9.7	使用Algebraic UDF	75
3.9.8	使用Accumulator UDF	75
3.9.9	剔除数据中的空记录	75
3.9.10	使用特殊连接	75
3.9.11	压缩中间结果	75
3.9.12	合并小文件	76
3.10	小结	76
第4章	Hive进阶	77
4.1	Hive架构	77
4.1.1	Hive元存储	78
4.1.2	Hive编译器	78
4.1.3	Hive执行引擎	78
4.1.4	Hive的支持组件	79
4.2	数据类型	79
4.3	文件格式	80
4.3.1	压缩文件	80
4.3.2	ORC文件	81
4.3.3	Parquet文件	81
4.4	数据模型	82
4.4.1	动态分区	84
4.4.2	Hive表索引	85
4.5	Hive查询优化器	87
4.6	DML进阶	88
4.6.1	GROUP BY操作	88
4.6.2	ORDER BY与SORT BY	88
4.6.3	JOIN类型	88
4.6.4	高级聚合	89
4.6.5	其他高级语句	90
4.7	UDF、UDAF和UDTF	90
4.8	小结	93
第5章	序列化和Hadoop I/O	95

5.1 Hadoop数据序列化	95
5.1.1 Writable与WritableComparable	96
5.1.2 Hadoop与Java序列化的区别	98
5.2 Avro序列化	100
5.2.1 Avro与MapReduce	102
5.2.2 Avro与Pig	105
5.2.3 Avro与Hive	106
5.2.4 比较Avro与Protocol Buffers/Thrift	107
5.3 文件格式	108
5.3.1 Sequence文件格式	108
5.3.2 MapFile格式	111
5.3.3 其他数据结构	113
5.4 压缩	113
5.4.1 分片与压缩	114
5.4.2 压缩范围	115
5.5 小结	115
第6章 YARN——其他应用模式进入Hadoop的引路人	116
6.1 YARN的架构	117
6.1.1 资源管理器	117
6.1.2 Application Master	118
6.1.3 节点管理器	119
6.1.4 YARN客户端	120
6.2 开发YARN的应用程序	120
6.2.1 实现YARN客户端	120
6.2.2 实现AM实例	125
6.3 YARN的监控	129
6.4 YARN中的作业调度	134
6.4.1 容量调度器	134
6.4.2 公平调度器	137
6.5 YARN命令行	139
6.5.1 用户命令	140
6.5.2 管理员命令	140
6.6 小结	141
第7章 基于YARN的Storm——Hadoop中的低延时处理	142
7.1 批处理对比流式处理	142
7.2 Apache Storm	144
7.2.1 Apache Storm的集群架构	144
7.2.2 Apache Storm的计算和数据模型	145
7.2.3 Apache Storm用例	146
7.2.4 Apache Storm的开发	147
7.2.5 Apache Storm 0.9.1	153
7.3 基于YARN的Storm	154
7.3.1 在YARN上安装Apache Storm	154
7.3.2 安装过程	154
7.4 小结	161
第8章 云上的Hadoop	162
8.1 云计算的特点	162
8.2 云上的Hadoop	163
8.3 亚马逊Elastic MapReduce	164

8.4	小结	175
第9章	HDFS替代品	176
9.1	HDFS的优缺点	176
9.2	亚马逊AWS S3	177
9.3	在Hadoop中实现文件系统	179
9.4	在Hadoop中实现S3原生文件系统	179
9.5	小结	189
第10章	HDFS联合	190
10.1	旧版HDFS架构的限制	190
10.2	HDFS联合的架构	192
10.2.1	HDFS联合的好处	193
10.2.2	部署联合NameNode	193
10.3	HDFS高可用性	195
10.3.1	从NameNode、检查节点和备份节点	195
10.3.2	高可用性——共享edits	196
10.3.3	HDFS实用工具	197
10.3.4	三层与四层网络拓扑	197
10.4	HDFS块放置策略	198
10.5	小结	200
第11章	Hadoop安全	201
11.1	安全的核心	201
11.2	Hadoop中的认证	202
11.2.1	Kerberos认证	202
11.2.2	Kerberos的架构和工作流	203
11.2.3	Kerberos认证和Hadoop	204
11.2.4	HTTP接口的认证	204
11.3	Hadoop中的授权	205
11.3.1	HDFS的授权	205
11.3.2	限制HDFS的使用量	208
11.3.3	Hadoop中的服务级授权	209
11.4	Hadoop中的数据保密性	211
11.5	Hadoop中的日志审计	216
11.6	小结	217
第12章	使用Hadoop进行数据分析	218
12.1	数据分析工作流	218
12.2	机器学习	220
12.3	Apache Mahout	222
12.4	使用Hadoop和Mahout进行文档分析	223
12.4.1	词频	223
12.4.2	文频	224
12.4.3	词频 - 逆向文频	224
12.4.4	Pig中的Tf-idf	225
12.4.5	余弦相似度距离度量	228
12.4.6	使用k-means的聚类	228
12.4.7	使用Apache Mahout进行k-means聚类	229
12.5	RHadoop	233
12.6	小结	233
附录	微软Windows中的Hadoop	235

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