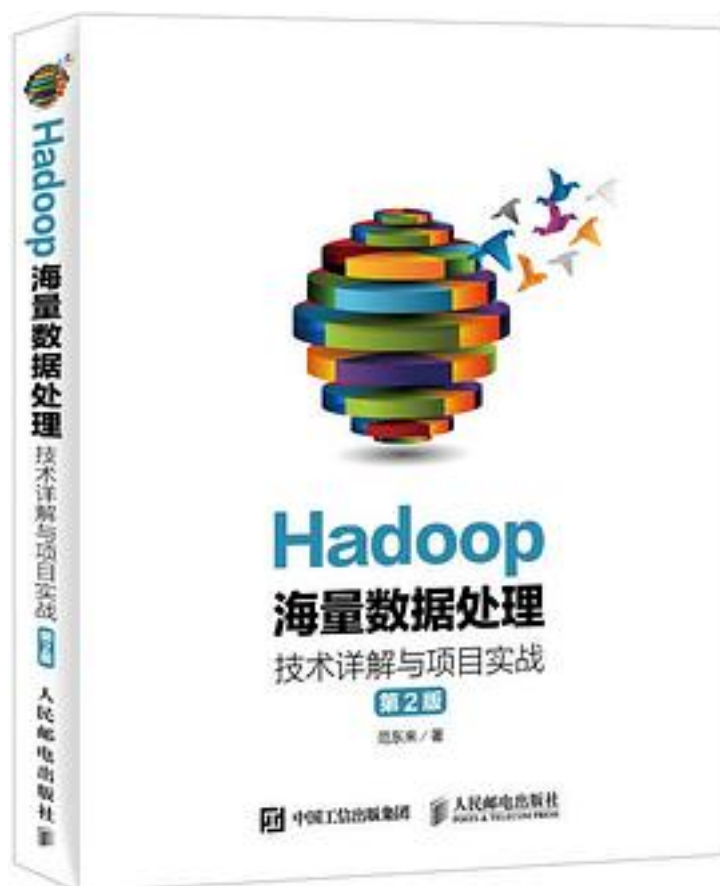


Hadoop海量数据处理（第2版）



[Hadoop海量数据处理（第2版）_下载链接1](#)

著者:范东来

出版者:人民邮电出版社

出版时间:2016-7

装帧:

isbn:9787115427465

作者介绍:

范东来，北京航空航天大学硕士，技术图书作者和译者，著有《Hadoop海量数据处理》（该书台湾繁体字版为《Hadoop：BigData技術詳解與專案實作》），译有《解读N

oSQL》。BBD（数联铭品）大数据技术部负责人，大数据平台架构师，极客学院布道师。研究方向：并行图挖掘、去中心化应用。

目录: 目录

基础篇：Hadoop基础

第1章 绪论 2

1.1 Hadoop和云计算 2

1.1.1 Hadoop的电梯演讲 2

1.1.2 Hadoop生态圈 3

1.1.3 云计算的定义 6

1.1.4 云计算的类型 7

1.1.5 Hadoop和云计算 8

1.2 Hadoop和大数据 9

1.2.1 大数据的定义 9

1.2.2 大数据的结构类型 10

1.2.3 大数据行业应用实例 12

1.2.4 Hadoop和大数据 13

1.2.5 其他大数据处理平台 14

1.3 数据挖掘和商业智能 15

1.3.1 数据挖掘的定义 15

1.3.2 数据仓库 17

1.3.3 操作数据库系统和数据仓库系统的区别 18

1.3.4 为什么需要分离的数据仓库 19

1.3.5 商业智能 19

1.3.6 大数据时代的商业智能 20

1.4 小结 21

第2章 环境准备 22

2.1 Hadoop的发行版本选择 22

2.1.1 Apache Hadoop 22

2.1.2 CDH 22

2.1.3 Hadoop的版本 23

2.1.4 如何选择Hadoop的版本 25

2.2 Hadoop架构 26

2.2.1 Hadoop HDFS架构 27

2.2.2 YARN架构 28

2.2.3 Hadoop架构 28

2.3 安装Hadoop 29

2.3.1 安装运行环境 30

2.3.2 修改主机名和用户名 36

2.3.3 配置静态IP地址 36

2.3.4 配置SSH无密码连接 37

2.3.5 安装JDK 38

2.3.6 配置Hadoop 39

2.3.7 格式化HDFS 42

2.3.8 启动Hadoop并验证安装 42

2.4 安装Hive 43

2.4.1 安装元数据库 44

2.4.2 修改Hive配置文件 44

2.4.3 验证安装 45

2.5 安装HBase 46

2.5.1 解压文件并修改Zookeeper相关配置 46

2.5.2 配置节点 46

2.5.3 配置环境变量 47

2.5.4 启动并验证	47
2.6 安装Sqoop	47
2.7 Cloudera Manager	48
2.8 小结	51
第3章 Hadoop的基石：HDFS	52
3.1 认识HDFS	52
3.1.1 HDFS的设计理念	54
3.1.2 HDFS的架构	54
3.1.3 HDFS容错	58
3.2 HDFS读取文件和写入文件	58
3.2.1 块的分布	59
3.2.2 数据读取	60
3.2.3 写入数据	61
3.2.4 数据完整性	62
3.3 如何访问HDFS	63
3.3.1 命令行接口	63
3.3.2 Java API	66
3.3.3 其他常用的接口	75
3.3.4 Web UI	75
3.4 HDFS中的新特性	76
3.4.1 NameNode HA	76
3.4.2 NameNode Federation	78
3.4.3 HDFS Snapshots	79
3.5 小结	79
第4章 YARN：统一资源管理和调平台	80
4.1 YARN是什么	80
4.2 统一资源管理和调度平台范型	81
4.2.1 集中式调度器	81
4.2.2 双层调度器	81
4.2.3 状态共享调度器	82
4.3 YARN的架构	82
4.3.1 ResourceManager	83
4.3.2 NodeManager	85
4.3.3 ApplicationMaster	87
4.3.4 YARN的资源表示模型Container	87
4.4 YARN的工作流程	88
4.5 YARN的调度器	89
4.5.1 YARN的资源管理机制	89
4.5.2 FIFO Scheduler	90
4.5.3 Capacity Scheduler	90
4.5.4 Fair Scheduler	91
4.6 YARN命令行	92
4.7 Apache Mesos	95
4.8 小结	96
第5章 分而治之的智慧：MapReduce	97
5.1 认识MapReduce	97
5.1.1 MapReduce的编程思想	98
5.1.2 MapReduce运行环境	100
5.1.3 MapReduce作业和任务	102
5.1.4 MapReduce的计算资源划分	102
5.1.5 MapReduce的局限性	103
5.2 Hello Word Count	104
5.2.1 Word Count的设计思路	104
5.2.2 编写Word Count	105

5.2.3	运行程序	107
5.2.4	还能更快吗	109
5.3	MapReduce的过程	109
5.3.1	从输入到输出	109
5.3.2	input	110
5.3.3	map及中间结果的输出	112
5.3.4	shuffle	113
5.3.5	reduce及最后结果的输出	115
5.3.6	sort	115
5.3.7	作业的进度组成	116
5.4	MapReduce的工作机制	116
5.4.1	作业提交	117
5.4.2	作业初始化	118
5.4.3	任务分配	118
5.4.4	任务执行	118
5.4.5	任务完成	118
5.4.6	推测执行	119
5.4.7	MapReduce容错	119
5.5	MapReduce编程	120
5.5.1	Writable类	120
5.5.2	编写Writable类	123
5.5.3	编写Mapper类	124
5.5.4	编写Reducer类	125
5.5.5	控制shuffle	126
5.5.6	控制sort	128
5.5.7	编写main函数	129
5.6	MapReduce编程实例：连接	130
5.6.1	设计思路	131
5.6.2	编写Mapper类	131
5.6.3	编写Reducer类	132
5.6.4	编写main函数	133
5.7	MapReduce编程实例：二次排序	134
5.7.1	设计思路	134
5.7.2	编写Mapper类	135
5.7.3	编写Partitioner类	136
5.7.4	编写SortComparator类	136
5.7.5	编写Reducer类	137
5.7.6	编写main函数	137
5.8	MapReduce编程实例：全排序	139
5.8.1	设计思路	139
5.8.2	编写代码	140
5.9	小结	141
第6章	SQL on Hadoop：Hive	142
6.1	认识Hive	142
6.1.1	从MapReduce到SQL	143
6.1.2	Hive架构	144
6.1.3	Hive与关系型数据库的区别	146
6.1.4	Hive命令的使用	147
6.2	数据类型和存储格式	149
6.2.1	基本数据类型	149
6.2.2	复杂数据类型	149
6.2.3	存储格式	150
6.2.4	数据格式	151
6.3	HQL：数据定义	152

- 6.3.1 Hive中的数据库 152
- 6.3.2 Hive中的表 154
- 6.3.3 创建表 154
- 6.3.4 管理表 156
- 6.3.5 外部表 156
- 6.3.6 分区表 156
- 6.3.7 删除表 158
- 6.3.8 修改表 158
- 6.4 HQL：数据操作 159
 - 6.4.1 装载数据 159
 - 6.4.2 通过查询语句向表中插入数据 160
 - 6.4.3 利用动态分区向表中插入数据 160
 - 6.4.4 通过CTAS加载数据 161
 - 6.4.5 导出数据 161
- 6.5 HQL：数据查询 162
 - 6.5.1 SELECT...FROM语句 162
 - 6.5.2 WHERE语句 163
 - 6.5.3 GROUP BY和HAVING语句 164
 - 6.5.4 JOIN语句 164
 - 6.5.5 ORDER BY和SORT BY语句 166
 - 6.5.6 DISTRIBUTE BY和SORT BY语句 167
 - 6.5.7 CLUSTER BY 167
 - 6.5.8 分桶和抽样 168
 - 6.5.9 UNION ALL 168
- 6.6 Hive函数 168
 - 6.6.1 标准函数 168
 - 6.6.2 聚合函数 168
 - 6.6.3 表生成函数 169
- 6.7 Hive用户自定义函数 169
 - 6.7.1 UDF 169
 - 6.7.2 UDAF 170
 - 6.7.3 UDTF 171
 - 6.7.4 运行 173
- 6.8 小结 173
- 第7章 SQL to Hadoop: Sqoop 174
 - 7.1 一个Sqoop示例 174
 - 7.2 导入过程 176
 - 7.3 导出过程 178
 - 7.4 Sqoop的使用 179
 - 7.4.1 codegen 180
 - 7.4.2 create-hive-table 180
 - 7.4.3 eval 181
 - 7.4.4 export 181
 - 7.4.5 help 182
 - 7.4.6 import 182
 - 7.4.7 import-all-tables 183
 - 7.4.8 job 184
 - 7.4.9 list-databases 184
 - 7.4.10 list-tables 184
 - 7.4.11 merge 184
 - 7.4.12 metastore 185
 - 7.4.13 version 186
 - 7.5 小结 186
- 第8章 HBase:HadoopDatabase 187

8.1 酸和碱：两种数据库事务方法论	187
8.1.1 ACID	188
8.1.2 BASE	188
8.2 CAP定理	188
8.3 NoSQL的架构模式	189
8.3.1 键值存储	189
8.3.2 图存储	190
8.3.3 列族存储	191
8.3.4 文档存储	192
8.4 HBase的架构模式	193
8.4.1 行键、列族、列和单元格	193
8.4.2 HMaster	194
8.4.3 Region和RegionServer	195
8.4.4 WAL	195
8.4.5 HFile	195
8.4.6 Zookeeper	197
8.4.7 HBase架构	197
8.5 HBase写入和读取数据	198
8.5.1 Region定位	198
8.5.2 HBase写入数据	199
8.5.3 HBase读取数据	199
8.6 HBase基础API	200
8.6.1 创建表	201
8.6.2 插入	202
8.6.3 读取	203
8.6.4 扫描	204
8.6.5 删除单元格	206
8.6.6 删除表	207
8.7 HBase高级API	207
8.7.1 过滤器	208
8.7.2 计数器	208
8.7.3 协处理器	209
8.8 小结	214
第9章 Hadoop性能调优和运维	215
9.1 Hadoop客户端	215
9.2 Hadoop性能调优	216
9.2.1 选择合适的硬件	216
9.2.2 操作系统调优	218
9.2.3 JVM调优	219
9.2.4 Hadoop参数调优	219
9.3 Hive性能调优	225
9.3.1 JOIN优化	226
9.3.2 Reducer的数量	226
9.3.3 列裁剪	226
9.3.4 分区裁剪	226
9.3.5 GROUP BY优化	226
9.3.6 合并小文件	227
9.3.7 MULTI-GROUP BY和MULTI-INSERT	228
9.3.8 利用UNION ALL 特性	228
9.3.9 并行执行	228
9.3.10 全排序	228
9.3.11 Top N	229
9.4 HBase调优	229
9.4.1 通用调优	229

9.4.2 客户端调优	230
9.4.3 写调优	231
9.4.4 读调优	231
9.4.5 表设计调优	232
9.5 Hadoop运维	232
9.5.1 集群节点动态扩容和卸载	233
9.5.2 利用SecondaryNameNode恢复NameNode	234
9.5.3 常见的运维技巧	234
9.5.4 常见的异常处理	235
9.6 小结	236
应用篇：商业智能系统项目实战	
第10章 在线图书销售商业智能系统	238
10.1 项目背景	238
10.2 功能需求	239
10.3 非功能需求	240
10.4 小结	240
第11章 系统结构设计	241
11.1 系统架构	241
11.2 功能设计	242
11.3 数据仓库结构	243
11.4 系统网络拓扑与硬件选型	246
11.4.1 系统网络拓扑	246
11.4.2 系统硬件选型	248
11.5 技术选型	249
11.5.1 平台选型	249
11.5.2 系统开发语言选型	249
11.6 小结	249
第12章 在开发之前	250
12.1 新建一个工程	250
12.1.1 安装Python	250
12.1.2 安装PyDev插件	251
12.1.3 新建PyDev项目	252
12.2 代码目录结构	253
12.3 项目的环境变量	253
12.4 如何调试	254
12.5 小结	254
第13章 实现数据导入导出模块	255
13.1 处理流程	255
13.2 导入方式	256
13.2.1 全量导入	256
13.2.2 增量导入	256
13.3 读取配置文件	257
13.4 SqoopUtil	261
13.5 整合	262
13.6 导入说明	262
13.7 导出模块	263
13.8 小结	265
第14章 实现数据分析工具模块	266
14.1 处理流程	266
14.2 读取配置文件	266
14.3 HiveUtil	268
14.4 整合	268
14.5 数据分析和报表	269
14.5.1 OLAP和Hive	269

- 14.5.2 OLAP和多维模型 270
- 14.5.3 选MySQL还是选HBase 272
- 14.6 小结 273
- 第15章 实现业务数据的数据清洗模块 274
 - 15.1 ETL 274
 - 15.1.1 数据抽取 274
 - 15.1.2 数据转换 274
 - 15.1.3 数据清洗工具 275
 - 15.2 处理流程 275
 - 15.3 数据去重 276
 - 15.3.1 产生原因 276
 - 15.3.2 去重方法 277
 - 15.3.3 一个很有用的UDF: RowNum 277
 - 15.3.4 第二种去重方法 279
 - 15.3.5 进行去重 279
 - 15.4 小结 282
- 第16章 实现点击流日志的数据清洗模块 283
 - 16.1 数据仓库和Web 283
 - 16.2 处理流程 285
 - 16.3 字段的获取 285
 - 16.4 编写MapReduce作业 288
 - 16.4.1 编写IP地址解析器 288
 - 16.4.2 编写Mapper类 291
 - 16.4.3 编写Partitioner类 295
 - 16.4.4 编写SortComparator类 295
 - 16.4.5 编写Reducer类 297
 - 16.4.6 编写main函数 298
 - 16.4.7 通过Python调用jar文件 299
 - 16.5 还能做什么 300
 - 16.5.1 网站分析的指标 300
 - 16.5.2 网站分析的决策支持 301
 - 16.6 小结 301
- 第17章 实现购书转化率分析模块 302
 - 17.1 漏斗模型 302
 - 17.2 处理流程 303
 - 17.3 读取配置文件 303
 - 17.4 提取所需数据 304
 - 17.5 编写转化率分析MapReduce作业 305
 - 17.5.1 编写Mapper类 306
 - 17.5.2 编写Partitioner类 308
 - 17.5.3 编写SortComparator类 309
 - 17.5.4 编写Reducer类 310
 - 17.5.5 编写Driver类 312
 - 17.5.6 通过Python模块调用jar文件 314
 - 17.6 对中间结果进行汇总得到最终结果 314
 - 17.7 整合 316
 - 17.8 小结 316
- 第18章 实现购书用户聚类模块 317
 - 18.1 物以类聚 317
 - 18.2 聚类算法 318
 - 18.2.1 k-means算法 318
 - 18.2.2 Canopy算法 319
 - 18.2.3 数据向量化 320
 - 18.2.4 数据归一化 321

- 18.2.5 相似性度量 322
- 18.3 用MapReduce实现聚类算法 323
 - 18.3.1 Canopy算法与MapReduce 323
 - 18.3.2 k-means算法与MapReduce 323
 - 18.3.3 Apache Mahout 324
- 18.4 处理流程 324
- 18.5 提取数据并做归一化 325
- 18.6 维度相关性 327
 - 18.6.1 维度的选取 327
 - 18.6.2 相关系数与相关系数矩阵 328
 - 18.6.3 计算相关系数矩阵 328
- 18.7 使用Mahout完成聚类 329
 - 18.7.1 使用Mahout 329
 - 18.7.2 解析Mahout的输出 332
 - 18.7.3 得到聚类结果 334
- 18.8 得到最终结果 335
- 18.9 评估聚类结果 337
 - 18.9.1 一份不适合聚类的数据 337
 - 18.9.2 簇间距离和簇内距离 337
 - 18.9.3 计算平均簇间距离 338
- 18.10 小结 339
- 第19章 实现调度模块 340
 - 19.1 工作流 340
 - 19.2 编写代码 341
 - 19.3 crontab 342
 - 19.4 让数据说话 343
 - 19.5 小结 344
- 结束篇：总结和展望
- 第20章 总结和展望 346
 - 20.1 总结 346
 - 20.2 BDAS 347
 - 20.3 Dremel系技术 348
 - 20.4 Pregel系技术 349
 - 20.5 Docker和Kubernetes 350
 - 20.6 数据集成工具NiFi 350
 - 20.7 小结 351
- 参考文献 352
 - • • • • (收起)

[Hadoop海量数据处理（第2版）_下载链接1](#)

标签

大数据

Hadoop

hadoop

评论

写的挺不错，有原理有实践

[Hadoop海量数据处理（第2版）_下载链接1_](#)

书评

新来现在公司的时候怕程序员笑我嘛都不懂，就去公司旁的中关村图书大厦随便买了一本看目录还挺丰富的书，跑数无聊的时候翻翻，到今天五个月过去终于翻完了。作者喜欢模仿人家在每篇开头放个名言或者歌词什么的，老实说真的很牵强。而且，作为一个技术人员，行文中成语多有误用...

[Hadoop海量数据处理（第2版）_下载链接1_](#)