

# 《大数据处理之道》

## 图书基本信息

书名：《大数据处理之道》

13位ISBN编号：9787121287234

出版时间：2016-9-1

作者：何金池

页数：284

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.tushu111.com](http://www.tushu111.com)

# 《大数据处理之道》

## 内容概要

《大数据处理之道》覆盖了当前大数据处理领域的热门技术，包括Hadoop、Spark、Storm、Dremel、Drill等，详细分析了各种技术的应用场景和优缺点；同时阐述了大数据下的日志分析系统，重点讲解了ELK日志处理方案；最后分析了大数据处理技术的发展趋势。

《大数据处理之道》采用幽默的表述风格，使读者容易理解、轻松掌握；重点从各种技术的起源、设计思想、架构等方面阐述，以帮助读者从根源上悟出大数据处理之道。

《大数据处理之道》适合大数据开发、大数据测试人员，以及其他软件开发或者管理人员和计算爱好者阅读。

## 书籍目录

0 “疯狂”的大数据	1
0.1 大数据时代	1
0.2 数据就是“金库”	3
0.3 让大数据“活”起来	4
第1篇 Hadoop 军营	
1 Hadoop 一石激起千层浪	7
1.1 Hadoop 诞生——不仅仅是玩具	7
1.2 Hadoop 发展——各路英雄集结	8
1.3 Hadoop 和它的小伙伴们	10
1.4 Hadoop 应用场景	12
1.5 小结	13
2 MapReduce 奠定基石	14
2.1 MapReduce 设计思想	14
2.2 MapReduce 运行机制	19
2.2.1 MapReduce 的组成	19
2.2.2 MapReduce 作业运行流程	20
2.2.3 JobTracker 解剖	26
2.2.4 TaskTracker 解剖	34
2.2.5 失败场景分析	42
2.3 MapReduce 实例分析	43
2.3.1 运行 WordCount 程序	44
2.3.2 WordCount 源码分析	45
2.4 小结	48
3 分布式文件系统	49

3.1 群雄并起的DFS	49
3.2 HDFS文件系统	51
3.2.1 HDFS 设计与架构	52
3.2.2 HDFS 操作与API	56
3.2.3 HDFS的优点及适用场景	60
3.2.4 HDFS的缺点及改进策略	61
3.3 小结	62
4 Hadoop体系的“四剑客”	63
4.1 数据仓库工具Hive	63
4.1.1 Hive缘起何处	63
4.1.2 Hive和数据库的区别	65
4.1.3 Hive设计思想与架构	66
4.1.4 适用场景	74
4.2 大数据仓库HBase	74
4.2.1 HBase因何而生	74
4.2.2 HBase的设计思想和架构	77
4.2.3 HBase优化技巧	84
4.2.4 HBase和Hive的区别	86
4.3 Pig编程语言	87
4.3.1 Pig的缘由	87
4.3.2 Pig的基本架构	88
4.3.3 Pig与Hive的对比	90
4.3.4 Pig的执行模式	90
4.3.5 Pig Latin语言及其应用	91
4.4 协管员ZooKeeper	

96	
4.4.1	ZooKeeper是什么
96	
4.4.2	ZooKeeper的作用
97	
4.4.3	ZooKeeper的架构
98	
4.4.4	ZooKeeper的数据模型
100	
4.4.5	ZooKeeper的常用接口及操作
102	
4.4.6	ZooKeeper的应用场景分析
105	
4.5	小结
108	
5	Hadoop资源管理与调度
110	
5.1	Hadoop调度机制
110	
5.1.1	FIFO
111	
5.1.2	计算能力调度器
111	
5.1.3	公平调度器
113	
5.2	Hadoop YARN资源调度
114	
5.2.1	YARN产生的背景
114	
5.2.2	Hadoop YARN的架构
116	
5.2.3	YARN的运作流程
118	
5.3	Apache Mesos资源调度
120	
5.3.1	Apache Mesos的起因
120	
5.3.2	Apache Mesos的架构
121	
5.3.3	基于Mesos的Hadoop
123	
5.4	Mesos与YARN对比
127	
5.5	小结
128	
6	Hadoop集群管理之道
129	
6.1	Hadoop 集群管理与维护
129	

6.1.1Hadoop集群管理	129
6.1.2Hadoop集群维护	131
6.2Hadoop 集群调优	132
6.2.1Linux文件系统调优	132
6.2.2Hadoop通用参数调整	133
6.2.3HDFS相关配置	133
6.2.4MapReduce相关配置	134
6.2.5Map任务相关配置	136
6.2.6HBase搭建重要的HDFS参数	137
6.3Hadoop 集群监控	137
6.3.1Apache Ambari监控	137
6.3.2Ganglia监控Hadoop	138
6.4小结	138
第2篇Spark星火燎原	
7 Spark宝刀出鞘	141
7.1Spark的历史渊源	141
7.1.1Spark的诞生	141
7.1.2Spark的发展	142
7.2Spark和Hadoop MapReduce对比	143
7.3Spark的适用场景	145
7.4Spark的硬件配置	146
7.5Spark架构	147
7.5.1Spark生态架构	147
7.5.2Spark运行架构	149
7.6小结	151

8 Spark核心RDD	153
8.1RDD简介	153
8.1.1什么是RDD	153
8.1.2为什么需要RDD	154
8.1.3RDD本体的设计	154
8.1.4RDD与分布式共享内存	155
8.2RDD的存储级别	155
8.3RDD依赖与容错	157
8.3.1RDD依赖关系	157
8.3.2RDD容错机制	160
8.4RDD操作与接口	161
8.4.1RDD Transformation操作与接口	162
8.4.2RDD Action操作与接口	164
8.5RDD编程示例	165
8.6小结	166
9 Spark运行模式和流程	167
9.1Spark运行模式	167
9.1.1Spark的运行模式列表	167
9.1.2Local模式	168
9.1.3Standalone模式	169
9.1.4Spark on Mesos模式	171
9.1.5Spark on YARN 模式	173
9.1.6Spark on EGO 模式	175
9.2Spark作业流程	177
9.2.1YARN-Client模式的作业流程	

178	
9.2.2	YARN-Cluster模式的作业流程
179	
9.3	小结
181	
10	Shark和Spark SQL
183	
10.1	从Shark到Spark SQL
183	
10.1.1	Shark的撤退是进攻
183	
10.1.2	Spark SQL接力
185	
10.1.3	Spark SQL与普通SQL的区别
186	
10.2	Spark SQL应用架构
187	
10.3	Spark SQL之DataFrame
188	
10.3.1	什么是DataFrame
188	
10.3.2	DataFrame的创建
188	
10.3.3	DataFrame的使用
190	
10.4	Spark SQL运行过程分析
190	
10.5	小结
192	
11	Spark Streaming流数据处理新贵
193	
11.1	Spark Streaming是什么
193	
11.2	Spark Streaming的架构
194	
11.3	Spark Streaming的操作
195	
11.3.1	Spark Streaming的Transformation操作
196	
11.3.2	Spark Streaming的Window操作
197	
11.3.3	Spark Streaming的Output操作
198	
11.4	Spark Streaming性能调优
198	
11.5	小结
200	
12	Spark GraphX图计算系统
201	



12.1图计算系统	201
12.1.1图存储模式	202
12.1.2图计算模式	203
12.2Spark GraphX的框架	206
12.3Spark GraphX的存储模式	207
12.4Spark GraphX的图运算符	208
12.5小结	211
13 Spark Cluster管理	212
13.1Spark Cluster部署	212
13.2Spark Cluster管理与监控	213
13.2.1内存优化机制	213
13.2.2Spark日志系统	213
13.3Spark 高可用性	215
13.4小结	216
第3篇其他大数据处理技术	
14 专为流数据而生的Storm	218
14.1Storm起因	218
14.2Storm的架构与组件	220
14.3Storm的设计思想	222
14.4Storm与Spark的区别	224
14.5Storm的适用场景	225
14.6Storm的应用	226
14.7小结	227
15 Dremel和Drill	228
15.1Dremel和Drill的历史背景	228

15.2Dremel的原理与应用	230
15.3Drill的架构与流程	232
15.4Dremel和Drill的适用场景与应用	234
15.5小结	234
第4篇大数据下的日志分析系统	
16 日志分析解决方案	236
16.1百花齐放的日志处理技术	236
16.2日志处理方案ELK	238
16.2.1ELK的三大金刚	238
16.2.2ELK的架构	240
16.2.3ELK的组网形式	242
16.3Logstash日志收集解析	245
16.3.1Input Plugins及应用示例	246
16.3.2Filter Plugins及应用示例	248
16.3.3Output Plugins及应用示例	249
16.4ElasticSearch存储与搜索	250
16.4.1ElasticSearch的主要概念	251
16.4.2ElasticSearch Rest API	252
16.5Kibana展示	253
16.6小结	255
17 ELK集群部署与应用	256
17.1ELK集群部署与优化	256
17.1.1ELK HA集群部署	256
17.1.2ElasticSearch优化	257
17.2如何开发自己的插件	259

17.3ELK在大数据运维系统中的应用	261
17.4ELK实战应用	262
17.4.1ELK监控Spark集群	262
17.4.2ELK监控系统资源状态	263
17.4.3ELK辅助日志管理和故障排查	263
17.5小结	264
第5篇数据分析技术前景展望	
18 大数据处理的思考与展望	266
18.1大数据时代的思考	266
18.2大数据处理技术的发展趋势	267
18.3小结	270

# 《大数据处理之道》

## 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.tushu111.com](http://www.tushu111.com)