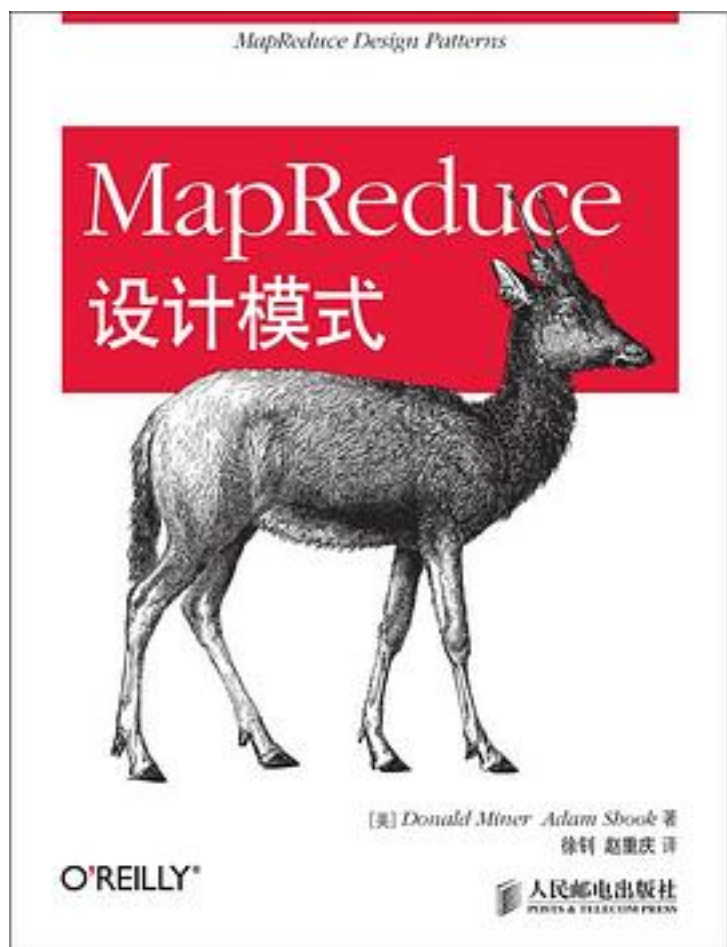


MapReduce设计模式



[MapReduce设计模式_下载链接1](#)

著者:[美]Donald Miner

出版者:人民邮电出版社

出版时间:2014-9-1

装帧:平装

isbn:9787115360946

MapReduce作为一种分布式海量数据处理的编程框架，已经得到业界的广泛关注。随着Hadoop的普及，MapReduce目前已经成为海量数据处理的最基础但也是最重要的方法之一。

《MapReduce设计模式》是一本关于设计模式的书，为读者提供解决问题的模板或通用指南。书中主要介绍编程模式，即如何利用MapReduce框架解决一类问题，重在提供解决问题的方法和思路。作者花大量篇幅介绍各种模式的原理及实现机制，并给出相应的应用实例，让读者对每种模式能有更直观的理解。

由于本书不会过多涉及底层框架及MapReduce API，所以希望读者阅读《MapReduce设计模式》之前，能够对Hadoop系统有所了解，知道如何编写MapReduce程序，并了解MapReduce程序框架的工作原理。《MapReduce设计模式》面向中高级MapReduce开发者，涵盖了绝大部分MapReduce编程可能面对的场景，相信初学者和专家同样可以在本书中得到一些启示。

作者介绍:

Donald Miner目前是EMC Greenplum的解决方案架构师，为实现与使用基于Greenplum的大数据系统的用户提供咨询和帮助。在加入Greenplum之前，Miner博士作为承包商为美国政府部署和构建了多个规模巨大且涉及关键任务的Hadoop集群。他还参与了教学，在马里兰大学巴尔的摩分校（UMBC）讲授Hadoop方面的业界前沿相关的课程以及各种人工智能课程。Miner博士在UMBC获得了计算机科学的博士学位，读博士期间他的研究主题为机器学习，博士论文的主题是多智能体系统。

Adam Shook是ClearEdge IT Solutions公司的软件工程师，从事一些大数据技术工作，如Hadoop、Accumulo、Pig以及ZooKeeper。Shook在马里兰大学巴尔的摩分校（UMBC）获得了计算机科学的学士学位，并得到了一份为游戏工作室构建一个全新的高性能图像引擎的工作。为了寻求新的挑战，他在UMBC就读研究生，主要专注的研究方向是分布式计算的相关技术。他很快找到了一份开发工作，作为美国政府承包商，从事大规模的Hadoop部署。Shook参与了Hadoop和Pig的开发及培训课程的指导。在繁忙工作的间隙他喜欢参与相关项目，玩视频游戏。

目录: 第1章 设计模式与MapReduce 1

- 1. 1 设计模式 2
- 1. 2 MapReduce简史 3
- 1. 3 MapReduce和Hadoop简介 4
- 1. 4 Hadoop示例：单词计数 6
- 1. 5 Pig和Hive 10
- 第2章 概要模式 12
 - 2. 1 数值概要 12
 - 2. 1. 1 模式描述 12
 - 2. 1. 2 数值概要示例 16
 - 2. 2 倒排索引概要 30
 - 2. 2. 1 模式描述 30
 - 2. 2. 2 倒排索引示例 32
 - 2. 3 计数器计数 34
 - 2. 3. 1 模式描述 34
 - 2. 3. 2 计数器计数示例 36
- 第3章 过滤模式 39
 - 3. 1 过滤 40
 - 3. 1. 1 模式描述 40
 - 3. 1. 2 过滤示例 43
 - 3. 2 布隆过滤 45
 - 3. 2. 1 模式描述 45

3. 2. 2 布隆过滤器示例	48
3. 3 Top 10	53
3. 3. 1 模式描述	53
3. 3. 2 Top 10示例	58
3. 4 去重	60
3. 4. 1 模式描述	60
3. 4. 2 去重示例	63
第4章 数据组织模式	65
4. 1 分层结构	65
4. 1. 1 模式描述	65
4. 1. 2 分层结构示例	69
4. 2 分区	76
4. 2. 1 模式描述	76
4. 2. 2 分区示例	79
4. 3 分箱	81
4. 3. 1 模式描述	81
4. 3. 2 分箱示例	83
4. 4 全排序	85
4. 4. 1 模式描述	85
4. 4. 2 全排序示例	88
4. 5 混排	92
4. 5. 1 模式描述	92
4. 5. 2 混排示例	93
第5章 连接模式	96
5. 1 连接简介	97
5. 2 reduce端连接	102
5. 2. 1 模式描述	102
5. 2. 2 reduce端连接示例	104
5. 2. 3 使用布隆过滤器的reduce端连接	110
5. 3 复制连接	112
5. 3. 1 模式描述	112
5. 3. 2 复制连接示例	114
5. 4 组合连接	116
5. 4. 1 模式描述	116
5. 4. 2 组合连接示例	119
5. 5 笛卡儿积	121
5. 5. 1 模式描述	121
5. 5. 2 笛卡儿积示例	124
第6章 元模式	131
6. 1 作业链	131
6. 1. 1 关于驱动程序	132
6. 1. 2 作业链示例	133
6. 1. 3 关于shell脚本	142
6. 1. 4 关于JobControl	145
6. 2 链折叠	149
6. 2. 1 ChainMapper方法和ChainReducer方法	153
6. 2. 2 链折叠示例	153
6. 3 作业归并	158
作业归并示例	160
第7章 输入和输出模式	166
7. 1 在Hadoop中自定义输入和输出	166
7. 1. 1 InputFormat	167
7. 1. 2 RecordReader	168
7. 1. 3 OutputFormat	169

- 7. 1. 4 RecordWriter 170
- 7. 2 生成数据 170
 - 7. 2. 1 模式描述 170
 - 7. 2. 2 生成数据示例 172
- 7. 3 外部源输出 177
 - 7. 3. 1 模式描述 177
 - 7. 3. 2 外部源输出示例 179
- 7. 4 外部源输入 183
 - 7. 4. 1 模型描述 183
 - 7. 4. 2 外部源输入示例 185
- 7. 5 分区裁剪 190
 - 7. 5. 1 模式描述 190
 - 7. 5. 2 分区裁剪示例 192
- 第8章 最后的思考与设计模式的未来 203
 - 8. 1 数据的本质趋势 203
 - 8. 1. 1 图像、音频和视频 203
 - 8. 1. 2 流式数据 204
 - 8. 2 YARN的影响 204
 - 8. 3 作为库或者组件的模式 205
 - 8. 4 读者可以帮到什么 205
- 附录 布隆过滤器 207
 - • • • • ([收起](#))

[MapReduce设计模式 下载链接1](#)

标签

大数据

hadoop

mapreduce

MapReduce

设计模式

计算机

Hadoop

互联网

评论

大部分内容已经知道了，读着没什么新意，入门书而已

非常适合学习MR的入门书籍，常见的用法都讲到了。示例代码大部分都是基于2.0+的API，现在也能够使用。不过MR本身并不复杂，而且相比Hive和Pig这类高级工具开发效率低性能也不见得好很多，实际上用到的并不多

印象最深的还是用 MapReduce 实现全排序、各种 Join、笛卡尔积

大数据里的设计模式。

醍醐灌顶，常用到的大数据计算思维都有。

作为mapreduce入门，这本书真的很不错，之前书刚上来就买了，然后发现其实在国外2012年时候就有了，有种差距好大的感觉

[MapReduce设计模式_下载链接1](#)

书评
