

《HBase企业应用开发实战》

图书基本信息

书名：《HBase企业应用开发实战》

13位ISBN编号：9787111478320

出版时间：2014-9

作者：马延辉,孟鑫,李立松

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

本书特色：

国内资深Hadoop技术专家实践经验结晶，完全从企业实际生产环境和需求出发，旨在帮助企业真正解决大数据的落地问题；

系统介绍HBase的功能使用、框架设计、基本原理和高级特性；详细讲解使用HBase设计大型数据应用系统的实践方法和技巧；深刻总结系统运维、监控和性能调优的最佳实践。

本书强调HBase在企业的实际应用，立足于企业的实际生产环境，旨在帮助企业切实解决大数据技术如何落地的问题。三位作者都是奋战在中国大数据技术一线的实践派专家，本书是他们实践经验的结晶。

本书内容在三个维度上具有重要特色：功能维度，从HBase的安装配置、参数设置，到数据模型、表结构设计、客户端使用、高级特性，本书做了系统且详尽的介绍；实战维度，不仅通过3个典型的应用案例详细讲解了如何使用HBase设计大型的数据应用系统，而且还结合实际生产系统讲解了HBase的集群运维、监控和性能调优；理论维度，则深入分析了HBase、框架设计、模式设计和基本原理。可谓是理论与实践完美结合，深度与广度兼备！

【名家推荐】

本书作者在Hadoop开发和运维领域工作近4年，积累了丰富的经验，同时也对Hadoop技术人员在学习过程中可能会遇到的问题有一定的了解，在此基础上写了这本书。从如何用好HBase出发，首先介绍设计原理和应用场景，让读者了解HBase适合什么场景不适合什么场景，然后再介绍应用编程、性能优化和生产环境中的运维经验，可谓由浅入深，循序渐进，值得推荐！

—— 查礼 博士

中国大数据技术大会（原Hadoop in China）主席，中国计算机学会大数据专家委员会委员，中科院计算所副研究员

近几年，大数据和开源越来越受到各行各业的关注，而作为大数据中不可替代的重中之重，Hadoop及其周边生态，也逐渐从互联网公司向传统行业过渡。本书的几位作者都是在Hadoop与大数据领域深入工作多年的践行者，既有丰富的理论知识，又有多年工作的实战经验。本书着重介绍了HBase的工作原理和设计架构，同时在实际工作的应用场景上亦着墨很重，大数据的神秘不仅仅在于具体的技术细节，更多的是由于它是个新生事物，很多人并不很清楚大数据的技术架构应如何设计，应用场景如何，而我这几位好友结合自己在实际工作中的宝贵经验，通过撰写本书为广大爱好者解答了这一难题。

本书是不可多得的理论与实践完美结合的技术书籍。

—— 向磊 phhiveadmin作者，汉云数衍创始人

大数据的概念已经逐渐深入人心，从互联网行业到传统行业，已经掀起一股“数据驱动商业价值”的热潮。大数据需要落地，需要开源技术来驱动新一轮的变革，而HBase作为大数据落地过程中的神兵利器，已经一次又一次证明了其巨大价值。本书不同于其他HBase的翻译版书籍，由来自国内互联网最前沿的实战派资深人士撰写而成，融合了自身的实战经验，更契合中国企业应用HBase技术的实情。本书由浅入深，结合理论阐述与案例剖析，如同一壶香茶，值得细细品咂。

—— 数盟社区 致力于为推崇“数据价值”的企业及个人打造最好的数据科学交流平台

作者简介

马延辉 资深Hadoop技术专家，对Hadoop生态系统相关技术有深刻的理解。曾就职于淘宝、Answers.com、暴风影音等知名互联网公司，从事Hadoop相关的技术工作，在企业级大数据系统的研发、运维和管理方面积累了丰富的实战经验。开源HBase监控工具Ella的作者。在国内Hadoop社区内非常活跃，经常在各种会议和沙龙上做技术分享，深受欢迎。现在专注于大数据技术在传统行业的落地，致力于大数据技术的普及和推广。

孟鑫 资深Hadoop技术专家，在软件行业从业近10年，对海量数据处理技术有着深刻的认识，曾负责Hadoop平台建设，在Hadoop开发和运维方面积累了大量的实战经验。于2013年获取了Cloudera的Hadoop Developer认证，多次到企业和社区去分享Hadoop、HBase等方面的技术知识和经验。对技术拥有极大的兴趣，热衷于研究各种新技术，乐于总结和分享经验及教训。目前从事管理工作，但依然热衷于产品设计和实现。

李立松 资深Hadoop技术专家，Easyhadoop技术社区创始人之一，对HDFS、MapReduce、HBase、Hive等Hadoop生态系统中的技术有比较深入的研究，在Hadoop开发方面积累了丰富的经验。曾就职于暴风影音，负责大数据平台开发与应用，担任大数据项目负责人。现在就职于缔元信，担任Hadoop高级工程师，负责缔元信DMP平台的研发工作。

书籍目录

前 言

第一部分 基础篇

第1章 认识HBase 2

1.1 理解大数据背景 2

1.1.1 什么是大数据 3

1.1.2 为何大数据至关重要 4

1.1.3 NoSQL在大数据中扮演的角色 4

1.2 HBase是什么 6

1.2.1 HBase的发展历史 6

1.2.2 HBase的发行版本 7

1.2.3 HBase的特性 9

1.3 HBase与Hadoop的关系 10

1.4 HBase的核心功能模块 12

1.4.1 客户端Client 12

1.4.2 协调服务组件ZooKeeper 13

1.4.3 主节点HMaster 13

1.4.4 Region节点HRegionServer 13

1.5 HBase的使用场景和经典案例 14

1.5.1 搜索引擎应用 15

1.5.2 增量数据存储 15

1.5.3 用户内容服务 17

1.5.4 实时消息系统构建 18

1.6 本章小结 18

第2章 HBase安装与配置 19

2.1 先决条件 19

2.2 HBase运行模式 23

2.2.1 单机模式 23

2.2.2 分布式模式 24

2.3 HBase的Web UI 31

2.4 HBase Shell工具使用 31

2.5 停止HBase集群 33

2.6 本章小结 33

第3章 数据模型 34

3.1 两类数据模型 34

3.1.1 逻辑模型 35

3.1.2 物理模型 35

3.2 数据模型的重要概念 36

3.2.1 表 36

3.2.2 行键 37

3.2.3 列族 38

3.2.4 单元格 38

3.3 数据模型的操作 38

3.3.1 读Get 39

3.3.2 写Put 39

3.3.3 扫描Scan 39

3.3.4 删除Delete 40

3.4 数据模型的特殊属性 40

- 3.4.1 版本 40
- 3.4.2 排序 42
- 3.4.3 列的元数据 42
- 3.4.4 连接查询 43
- 3.4.5 计数器 43
- 3.4.6 原子操作 43
- 3.4.7 事务特性ACID 43
- 3.4.8 行锁 45
- 3.4.9 自动分区 45
- 3.5 CAP原理与最终一致性 46
- 3.6 本章小结 47
- 第4章 HBase表结构设计 48
 - 4.1 模式创建 48
 - 4.2 Rowkey设计 49
 - 4.3 列族定义 51
 - 4.3.1 可配置的数据块大小 51
 - 4.3.2 数据块缓存 52
 - 4.3.3 布隆过滤器 52
 - 4.3.4 数据压缩 53
 - 4.3.5 单元时间版本 53
 - 4.3.6 生存时间 54
 - 4.4 模式设计实例 54
 - 4.4.1 实例1：动物分类 54
 - 4.4.2 实例2：店铺与商品 56
 - 4.4.3 实例3：网上商城用户消费记录 57
 - 4.4.4 实例4：微博用户与粉丝 58
 - 4.5 本章小结 60
- 第5章 HBase客户端 61
 - 5.1 精通原生Java客户端 61
 - 5.1.1 客户端配置 62
 - 5.1.2 创建表 69
 - 5.1.3 删除表 70
 - 5.1.4 插入数据 70
 - 5.1.5 查询数据 72
 - 5.1.6 删除数据 76
 - 5.1.7 过滤查询 77
 - 5.2 使用HBase Shell工具操作HBase 79
 - 5.2.1 命令分类 79
 - 5.2.2 常规命令 80
 - 5.2.3 DDL命令 81
 - 5.2.4 DML命令 82
 - 5.2.5 工具命令Tools 86
 - 5.2.6 复制命令 87
 - 5.2.7 安全命令 87
 - 5.3 使用Thrift客户端访问HBase 88
 - 5.3.1 Thrift与Thrift2区别 88
 - 5.3.2 安装与部署Thrift2 89
 - 5.3.3 Python使用案例 93
 - 5.4 通过REST客户端访问HBase 95

- 5.4.1 启动服务 95
- 5.4.2 使用REST访问example表 96
- 5.5 使用MapReduce批量操作HBase 97
 - 5.5.1 三种访问模式 98
 - 5.5.2 实现MapReduce API 98
 - 5.5.3 HBase作为输入源示例 99
 - 5.5.4 HBase作为输出源示例 101
 - 5.5.5 HBase作为共享源示例 103
- 5.6 通过Web UI工具查看HBase状态 106
 - 5.6.1 Master状态界面 106
 - 5.6.2 RegionServer状态界面 107
 - 5.6.3 ZooKeeper统计信息页面 109
- 5.7 其他客户端 110
- 5.8 本章小结 110
- 第二部分 实战篇
- 第6章 整合SQL引擎层 114
 - 6.1 NoSQL背景知识 114
 - 6.1.1 什么是NoSQL 114
 - 6.1.2 将SQL整合到HBase的原因 115
 - 6.1.3 基于HBase的SQL引擎实现 116
 - 6.2 Hive整合HBase的实现 119
 - 6.2.1 认识Hive 119
 - 6.2.2 Hive整合HBase的环境准备 122
 - 6.2.3 Linux环境下重新编译Hive 123
 - 6.2.4 Hive参数配置 125
 - 6.2.5 启动Hive 127
 - 6.2.6 Hive与HBase整合后的框架如何使用 127
 - 6.2.7 HBase到Hive的字段映射 133
 - 6.2.8 多列与Hive Map类型 134
 - 6.3 查询引擎Phoenix 137
 - 6.3.1 认识Phoenix 138
 - 6.3.2 Phoenix安装环境准备 141
 - 6.3.3 Phoenix安装部署 142
 - 6.3.4 Phoenix源码编译 143
 - 6.3.5 Phoenix中SQLLine的快速使用 149
 - 6.3.6 使用JDBC访问Phoenix 153
 - 6.4 对象映射框架Kundera 155
 - 6.4.1 认识Kundera 155
 - 6.4.2 Kundera的客户端API快速使用 158
 - 6.4.3 Kundera模块介绍 161
 - 6.4.4 Kundera的REST访问方式 162
 - 6.5 分布式SQL引擎Lealone 165
 - 6.5.1 认识Lealone 165
 - 6.5.2 Lealone的安装部署 166
 - 6.5.3 通过JDBC访问Lealone 168
 - 6.5.4 通过Python访问Lealone 169
 - 6.5.5 Lealone特有的建表语法 170
 - 6.6 本章小结 171
- 第7章 构建音乐站用户属性库 173

- 7.1 案例背景 173
 - 7.1.1 音乐站 173
 - 7.1.2 需求概述 175
 - 7.1.3 需求范围和系统边界 175
 - 7.1.4 需求详述 176
 - 7.1.5 名词解释 180
- 7.2 概要设计 181
 - 7.2.1 设计目标 181
 - 7.2.2 数据规模假设 181
 - 7.2.3 功能指标 182
 - 7.2.4 系统流程 182
- 7.3 表结构设计 183
 - 7.3.1 功能抽象 183
 - 7.3.2 逻辑结构 184
 - 7.3.3 Rowkey设计 188
 - 7.3.4 列族设计 188
 - 7.3.5 版本定义 188
 - 7.3.6 优化属性定义 188
- 7.4 数据加载 189
 - 7.4.1 加载流程 189
 - 7.4.2 Mapper类 190
 - 7.4.3 Main类 192
 - 7.4.4 运行 193
- 7.5 数据检索 193
 - 7.5.1 HBaseTable 193
 - 7.5.2 HBaseAdmin 193
 - 7.5.3 几种检索类型 195
- 7.6 后台查询 198
 - 7.6.1 二级索引实现 198
 - 7.6.2 后台查询系统 205
- 7.7 本章小结 206
- 第8章 构建广告实时计算系统 208
 - 8.1 理解广告数据和流处理框架 208
 - 8.1.1 网络广告的几大特性 209
 - 8.1.2 网络广告的数据类型 210
 - 8.1.3 流处理框架 211
 - 8.1.4 背景与需求描述 217
 - 8.2 概要设计 218
 - 8.2.1 设计目标 219
 - 8.2.2 主要功能 219
 - 8.2.3 系统架构 219
 - 8.3 详细设计 221
 - 8.3.1 表结构设计 221
 - 8.3.2 功能模块设计 222
 - 8.4 核心功能实现 223
 - 8.4.1 规划集群环境部署 223
 - 8.4.2 安装ZooKeeper集群 225
 - 8.4.3 安装Kafka分布式集群 228
 - 8.4.4 实现Kafka生产者 231

- 8.4.5 安装Storm分布式集群 233
- 8.4.6 查看集群节点部署情况 240
- 8.4.7 基于Storm-kafka中间件实现计算逻辑 240
- 8.4.8 如何使用HBase中统计数据 251
- 8.5 本章小结 252
- 第三部分 高级篇
- 第9章 核心概念 254
- 9.1 核心结构 254
- 9.1.1 B+树 255
- 9.1.2 LSM树 255
- 9.1.3 两种结构本质区别 257
- 9.2 底层持久化 258
- 9.2.1 存储基本架构 258
- 9.2.2 HDFS文件 259
- 9.2.3 Region切分 264
- 9.2.4 合并 265
- 9.2.5 HFile格式 266
- 9.2.6 KeyValue格式 269
- 9.3 预写日志 270
- 9.3.1 概要流程 270
- 9.3.2 相关Java类 271
- 9.3.3 日志回放 274
- 9.3.4 日志一致性 275
- 9.4 写入流程 276
- 9.4.1 客户端 276
- 9.4.2 服务器端 281
- 9.5 查询流程 286
- 9.5.1 两种查询操作 286
- 9.5.2 客户端 286
- 9.5.3 服务器端 287
- 9.6 数据备份 291
- 9.6.1 备份机制架构 292
- 9.6.2 故障恢复 292
- 9.7 数据压缩 294
- 9.7.1 支持的压缩算法 295
- 9.7.2 使用配置 295
- 9.8 本章小结 296
- 第10章 HBase高级特性 297
- 10.1 过滤器 297
- 10.1.1 过滤器的两类参数 297
- 10.1.2 比较器 298
- 10.1.3 列值过滤器 300
- 10.1.4 键值元数据过滤器 300
- 10.1.5 行键过滤器 303
- 10.1.6 功能过滤器 303
- 10.1.7 Thrift使用过滤器 304
- 10.1.8 过滤器总结 309
- 10.2 计数器 310
- 10.2.1 使用Shell操作计数器 310

- 10.2.2 基于单列的计数器 312
- 10.2.3 多列计数器 313
- 10.3 协处理器 314
 - 10.3.1 认识协处理器 315
 - 10.3.2 观察者Observer 316
 - 10.3.3 终端EndPoint 318
 - 10.3.4 协处理器部署 320
- 10.4 Schema设计要点 323
 - 10.4.1 行键设计 323
 - 10.4.2 列族设计 325
- 10.5 二级索引 325
 - 10.5.1 Client-managed方式 326
 - 10.5.2 ITHBase实现 326
 - 10.5.3 IHBase实现 329
 - 10.5.4 Coprocessor方式 329
 - 10.5.5 MapReduce两种方式 330
- 10.6 布隆过滤器 330
 - 10.6.1 基本概念 331
 - 10.6.2 配置布隆过滤器 332
 - 10.6.3 使用布隆过滤器 333
- 10.7 负载均衡 333
 - 10.7.1 全局计划 334
 - 10.7.2 随机分配计划 337
 - 10.7.3 批量启动分配计划 337
 - 10.7.4 通过Shell控制负载均衡 338
- 10.8 批量加载 338
 - 10.8.1 准备数据：importtsv 338
 - 10.8.2 加载数据：completebulkload 340
- 10.9 本章小结 340
- 第11章 集群运维管理 341
 - 11.1 HBase常用工具 341
 - 11.1.1 文件检测修复工具hbck 342
 - 11.1.2 文件查看工具hfile 346
 - 11.1.3 WAL日志查看工具hlog 348
 - 11.1.4 压缩测试工具CompressionTest 349
 - 11.1.5 数据迁移工具CopyTable 350
 - 11.1.6 导出工具export 351
 - 11.1.7 导入工具Import 351
 - 11.1.8 日志回放工具WALPlayer 351
 - 11.1.9 行数统计工具RowCounter 352
 - 11.2 Region和RegionServer管理 353
 - 11.2.1 大合并工具major_compact 353
 - 11.2.2 Region合并工具Merge 354
 - 11.2.3 下线节点 354
 - 11.2.4 滚动重启 355
 - 11.3 性能指标Metrics 356
 - 11.3.1 Master Metrics 357
 - 11.3.2 RegionServer Metrics 357
 - 11.3.3 RPC Metrics 358

- 11.3.4 JVM Metrics 359
- 11.3.5 集群属性Metrics 360
- 11.4 监控系统Ganglia 360
 - 11.4.1 HBase监控指标 360
 - 11.4.2 安装、部署和使用Ganglia 361
- 11.5 HBase管理扩展JMX 366
 - 11.5.1 如何使用JMX 366
 - 11.5.2 基于JMX的监控工具Ella 368
- 11.6 报警工具Nagios 371
- 11.7 故障处理 376
 - 11.7.1 问题咨询渠道 377
 - 11.7.2 常用日志信息 377
 - 11.7.3 常用故障调试工具 379
 - 11.7.4 客户端故障排查 384
 - 11.7.5 MapReduce故障排查 386
 - 11.7.6 网络故障排查 387
 - 11.7.7 RegionServer相关问题解决 387
 - 11.7.8 Master相关问题解决 391
 - 11.7.9 ZooKeeper相关问题解决 392
- 11.8 集群备份 392
 - 11.8.1 冷备份 393
 - 11.8.2 热备份之Replication 393
 - 11.8.3 热备份之CopyTable 393
 - 11.8.4 热备份之Export 393
- 11.9 本章小结 393
- 第12章 性能调优 395
 - 12.1 硬件和操作系统调优 395
 - 12.1.1 配置内存 395
 - 12.1.2 配置CPU 396
 - 12.1.3 操作系统 396
 - 12.2 网络通信调优 399
 - 12.2.1 配置交换机 399
 - 12.2.2 添加机架感知 401
 - 12.3 JVM优化 402
 - 12.3.1 Java垃圾回收算法 402
 - 12.3.2 Java垃圾收集器 403
 - 12.3.3 垃圾回收器的选择 405
 - 12.3.4 JVM参数设置 406
 - 12.4 HBase查询优化 408
 - 12.4.1 设置Scan缓存 408
 - 12.4.2 显式地指定列 409
 - 12.4.3 关闭ResultScanner 410
 - 12.4.4 禁用块缓存 410
 - 12.4.5 优化行键查询 410
 - 12.4.6 通过HTableTool访问 410
 - 12.4.7 使用批量读 411
 - 12.4.8 使用Filter降低客户端压力 412
 - 12.4.9 使用Coprocessor统计行数 412
 - 12.4.10 缓存查询结果 413

- 12.5 HBase写入优化 413
 - 12.5.1 关闭写WAL日志 413
 - 12.5.2 设置AutoFlush 414
 - 12.5.3 预创建Region 415
 - 12.5.4 延迟日志flush 419
 - 12.5.5 通过HTableTool访问 419
 - 12.5.6 使用批量写 420
- 12.6 HBase基本核心服务优化 421
 - 12.6.1 优化分裂操作 421
 - 12.6.2 优化合并操作 423
- 12.7 HBase配置参数优化 423
 - 12.7.1 设置RegionServer Handler数量 423
 - 12.7.2 调整BlockCache大小 425
 - 12.7.3 设置MemStore的上下限 426
 - 12.7.4 调整影响合并的文件数 427
 - 12.7.5 调整MemStore的flush因子 427
 - 12.7.6 调整单个文件大小 427
 - 12.7.7 调整ZooKeeper Session的有效时长 428
- 12.8 分布式协调系统ZooKeeper优化 428
 - 12.8.1 配置ZooKeeper节点数 428
 - 12.8.2 独立ZooKeeper集群 429
- 12.9 表设计优化 430
 - 12.9.1 开启布隆过滤器 430
 - 12.9.2 调整列族块大小 430
 - 12.9.3 设置In Memory属性 432
 - 12.9.4 调整列族最大版本数 434
 - 12.9.5 设置TTL属性 435
- 12.10 其他优化 436
 - 12.10.1 关闭MapReduce的预测执行功能 436
 - 12.10.2 修改负载均衡执行周期 438
- 12.11 性能测试 438
- 12.12 本章小结 441
- 附录A HBase配置参数介绍 442
- 附录B Phoenix SQL语法详解 451
- 附录C YCSB编译安装 468

《HBase企业应用开发实战》

精彩短评

- 1、看完了，里面很多东西在HBase权威指南都有。。
- 2、本书结合实际案例做了hbase的介绍。比较通俗易懂。最大的优点就是可以结合本书做自己相关项目的开发。
- 3、HBase一本参考书，主要看中理解表模型和一些案例看到，客户端编程部分，HBase发展太快，已经略显跟不上
- 4、这本书适合入门，实战部分有点启发，但是后面的深层解读方面就做得差一些，总的来说，用来了解一下 HBase 还是很有用的

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