

# 《Storm企业级应用：实战、运维骸

## 图书基本信息

书名：《Storm企业级应用：实战、运维和调优》

13位ISBN编号：9787111503384

出版时间：2015-6

作者：马延辉,陈书美,雷葆华

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)

## 内容概要

### 【编辑推荐】

国内资深大数据专家根据Storm最新技术撰写，基于实际生产环境，从实战、运维和调优3个维度对Storm进行了详细的讲解

全面介绍Storm的架构、原理、核心概念、操作和数据流模型；6个不同领域的经典案例完整呈现大型数据应用系统的设计；系统总结了Storm常见运维故障的处理以及常用的技巧和最佳实践

### 【内容简介】

这是一本真正能让读者通过真实的企业应用实践掌握如何应用Storm处理各种大数据业务的著作。作者是国内顶尖的大数据专家，深谙初学者的难处和企业的实际需求，在此基础上巧妙地安排和组织了本书的内容，旨在让读者能达到事半功倍的效果。

本书没有过于繁琐和高深的理论，先系统介绍了Storm的设计思想、核心组件和适用场景，生产环境的搭建和部署，以及核心概念和数据模型；然后通过实时语音处理、网络流量分析、实时路况分析、数据质量实时监控、交通路况监控、广告实时流量统计等6典型的企业级案例讲解了如何应用Storm处理各种类型的大数据业务，每个案例都包括背景介绍、系统架构、模块设计和逻辑实现等组成部分，讲解十分详尽；最后从运维角度总结了Storm在安装部署、启动、运行等环节可能会出现各种故障并给出了解决方案，以及Storm的各种技巧和最佳实践。

## 作者简介

马延辉 资深Hadoop技术专家，曾就职于淘宝、Answers.com、暴风等互联网公司，从事Hadoop相关的研发和运维工作，对大数据技术的企业级落地、研发、运维和管理方面有着深刻理解和丰富的实战经验。开源HBase监控工具Ella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。现在致力于大数据技术在传统行业的落地，以及大数据技术的普及和推广。著有畅销书《HBase企业应用开发实战》，HBase领域公认最有价值的著作之一。

陈书美 高级数据分析工程师，对Storm实时数据计算、Kafka消息系统析有深入的研究和丰富的实践经验，并对HDFS、MapReduce、Hive、HBase等Hadoop生态系统内的技术有系统且深刻的了解。曾就职于暴风影音，负责大数据应用等开发工作；目前就职于百度，从事数据分析工作。

雷葆华 武汉绿色网络信息服务有限公司副总经理，负责公司大数据、SDN/NFV等新产品的研发工作。是业界知名的云计算专家，中国电子学会云计算专委会委员。创业之前曾任中国电信北京研究院云计算产品线总监，是中国电信云计算工作主要发起者和推动者之一，在CDN、P2P、IDC等方面都有深入研究。提交专利28项，已授权专利8项，发表多篇有影响力的论文和文章，多次获得部级科技进步奖励。

著有《云计算解码》、《CDN技术详解》、《SDN核心技术剖析和实战指南》等多部畅销著作并获得多个奖项。

## 书籍目录

前 言

基 础 篇

第1章 认识Storm 2

1.1 什么是实时流计算 2

1.1.1 实时流计算背景 3

1.1.2 实时计算应用场景 3

1.1.3 实时计算处理流程 4

1.1.4 实时计算框架 5

1.2 Storm是什么 11

1.2.1 Storm出现的背景 12

1.2.2 Storm简介 12

1.2.3 Storm的设计思想 13

1.2.4 Storm与Hadoop的角色和组件比较 14

1.3 Storm核心组件 15

1.3.1 主节点Nimbus 15

1.3.2 工作节点Supervisor 15

1.3.3 协调服务组件ZooKeeper 16

1.3.4 其他核心组件 16

1.4 Storm的特性 16

1.5 Storm的功能 18

1.6 本章小结 19

第2章 开始使用Storm 20

2.1 环境准备 20

2.1.1 系统配置 20

2.1.2 安装ZooKeeper集群 22

2.2 启动模式 26

2.2.1 本地模式 26

2.2.2 分布式模式 26

2.3 安装部署Storm集群 26

2.3.1 安装Storm依赖库 27

2.3.2 安装Storm集群 28

2.3.3 启动Storm集群 31

2.3.4 停止Storm集群 33

2.4 创建Topology并向集群提交任务 33

2.4.1 创建Topology 34

2.4.2 向集群提交任务 36

2.5 本章小结 36

第3章 核心概念和数据流模型 37

3.1 Tuple元组 37

3.1.1 Tuple描述 37

3.1.2 Tuple的生命周期 38

3.2 Spout数据源 39

3.2.1 Spout介绍 39

3.2.2 Spout实例 40

3.3 Bolt消息处理器 42

3.3.1 Bolt介绍 42

3.3.2 Bolt实例 45

- 3.4 Topology拓扑 47
    - 3.4.1 Topology实例 48
    - 3.4.2 Topology运行 51
  - 3.5 Stream消息流和Stream Grouping消息流组 55
    - 3.5.1 Stream消息流 55
    - 3.5.2 Stream Grouping消息流组 55
  - 3.6 Task任务 56
  - 3.7 Worker工作者进程 56
  - 3.8 Worker、Task、Executor三者之间的关系 57
  - 3.9 事务 57
  - 3.10 数据流模型 58
    - 3.10.1 数据流模型简介 58
    - 3.10.2 Storm数据流模型 60
  - 3.11 本章小结 61
- 实 战 篇
- 第4章 实例1：移动互联——语音“实时墙” 64
    - 4.1 业务背景 64
      - 4.1.1 案例背景 64
      - 4.1.2 设计目标 65
      - 4.1.3 数据格式 66
      - 4.1.4 硬件配置 68
    - 4.2 系统架构与模块设计 68
      - 4.2.1 整体架构 69
      - 4.2.2 数据采集 70
      - 4.2.3 数据实时处理 70
      - 4.2.4 存储设计 70
      - 4.2.5 Web实时展示 71
      - 4.2.6 硬件部署图 72
    - 4.3 核心模块实现 73
      - 4.3.1 实时处理业务逻辑实现 73
      - 4.3.2 Web展示实现 80
      - 4.3.3 最终效果呈现 88
    - 4.4 本章小结 89
  - 第5章 实例2：运营商——网络流量流向实时分析 90
    - 5.1 业务背景 90
      - 5.1.1 案例背景 91
      - 5.1.2 设计目标 91
      - 5.1.3 数据规模预估 92
      - 5.1.4 数据格式 92
      - 5.1.5 统计分析需求 93
    - 5.2 系统架构与模块设计 94
      - 5.2.1 整体架构 94
      - 5.2.2 数据源 95
      - 5.2.3 日志采集 96
      - 5.2.4 数据存储 96
      - 5.2.5 数据处理 97
      - 5.2.6 目标存储和扩展服务 97
      - 5.2.7 结果Web展示 97
    - 5.3 核心模块实现 98

- 5.3.1 模拟数据实现 98
- 5.3.2 日志采集和存储实现 102
- 5.3.3 数据处理实现 105
- 5.3.4 Web展示实现 111
- 5.4 本章小结 114
- 第6章 实例3：交通——基于GPS的实时路况分析 115
  - 6.1 业务背景 115
    - 6.1.1 案例背景 115
    - 6.1.2 设计目标 116
    - 6.1.3 数据格式 118
    - 6.1.4 实时路况分析方法 118
  - 6.2 系统架构和模块设计 118
  - 6.3 核心模块的实现 121
    - 6.3.1 安装Kafka集群 121
    - 6.3.2 Flume整合Kafka 124
    - 6.3.3 实时处理数据 125
    - 6.3.4 Web页面展示 127
  - 6.4 本章小结 129
- 第7章 实例4：互联网——数据质量实时监控 130
  - 7.1 业务背景 130
    - 7.1.1 案例背景 130
    - 7.1.2 设计目标 132
    - 7.1.3 数据格式 132
  - 7.2 系统架构与模块设计 133
    - 7.2.1 整体架构 133
    - 7.2.2 结果Web展示 135
  - 7.3 核心模块实现 135
    - 7.3.1 模拟数据 135
    - 7.3.2 实时处理业务逻辑的实现 141
    - 7.3.3 Web界面实现 147
    - 7.3.4 最终效果图 150
  - 7.4 本章小结 152
- 第8章 实例5：交通——超速频发路段监控 153
  - 8.1 业务背景 153
    - 8.1.1 案例背景 153
    - 8.1.2 数据类型 155
  - 8.2 系统架构和模块设计 157
  - 8.3 核心模块实现 158
    - 8.3.1 实现入口类Main 158
    - 8.3.2 数据源SocketSpout的实现 159
    - 8.3.3 实时处理MapSearchBolt和SpeedProcessBolt的实现 161
    - 8.3.4 目标存储DataBaseLoadBolt的实现 169
  - 8.4 本章小结 171
- 第9章 实例6：互联网——广告实时流量统计 172
  - 9.1 广告实时流量统计系统架构 172
    - 9.1.1 广告数据 172
    - 9.1.2 详细需求描述 174
    - 9.1.3 系统架构 175
  - 9.2 表结构与模块设计 177

- 9.2.1 表结构设计 177
- 9.2.2 功能模块设计 178
- 9.3 核心模块实现 179
  - 9.3.1 部署物理集群环境 179
  - 9.3.2 Kafka生产者逻辑的实现 181
  - 9.3.3 使用Storm-kafka实现业务逻辑 182
  - 9.3.4 使用HBase存储并实现统计 193
- 9.4 本章小结 194
- 技巧篇
- 第10章 Storm常见故障及解决方法 196
  - 10.1 安装部署故障 196
    - 10.1.1 “ no jzmq in java.library.path ” 异常 196
    - 10.1.2 “ No rule to make target ” 异常 198
    - 10.1.3 “ cannot access org.zeromq.ZMQ ” 异常 198
    - 10.1.4 缺少pkg-conf?ig异常 198
    - 10.1.5 “ java.lang.Unsatisf?iedLinkError ” 异常 199
    - 10.1.6 “ java.lang.NoClassDefFoundError : clojure.core.protocols\$ ” 异常 199
    - 10.1.7 “ Error : cannot link with -luuid , install uuid-dev ” 异常 199
    - 10.1.8 “ bad interpreter : No such f?ile or directory ” 异常 200
    - 10.1.9 “ org.zeromq.ZMQException : Invalid argument ” 异常 200
  - 10.2 启动故障 201
    - 10.2.1 “ java.io.FileNotFoundException ” 异常 201
    - 10.2.2 “ java.io.EOFException ” 异常 202
  - 10.3 运行时故障 202
    - 10.3.1 “ Nimbus host is not set ” 异常 203
    - 10.3.2 “ AlreadyAliveException ( msg : xxx is alreadyactive ) ” 异常 203
    - 10.3.3 无法序列化log4j.Logger异常 203
    - 10.3.4 “ Failing message ” 异常 203
    - 10.3.5 “ java.io.NotSerializableException ” 异常 204
    - 10.3.6 “ java.lang.NoClassDefFoundError ” 异常 205
    - 10.3.7 “ java.net.NoRouteToHostException ” 异常 206
    - 10.3.8 “ java.net.UnknownHostException ” 异常 206
    - 10.3.9 重复defaults.yaml资源文件异常 207
    - 10.3.10 “ KeeperException\$NoNodeException ” 异常 208
    - 10.3.11 “ A fatal error has been detected by the Java Runtime Environment ” 错误 209
    - 10.3.12 “ java.lang.ArrayIndexOutOfBoundsException ” 异常 212
    - 10.3.13 DRPC空指针异常 212
    - 10.3.14 Storm Thrift读取数据报错 212
  - 10.4 本章小结 214
- 第11章 Storm使用技巧和最佳实践 215
  - 11.1 核心组件使用要点 215
    - 11.1.1 Spout和Bolt 215
    - 11.1.2 ZooKeeper集群尽量独立 219
    - 11.1.3 Thrift服务的应用场景 220
    - 11.1.4 序列化机制的使用场景 220
  - 11.2 集群配置技巧 220
    - 11.2.1 默认参数配置 220
    - 11.2.2 日志信息 223
    - 11.2.3 合理配置JVM参数 223

- 11.3 集群运维技巧 224
  - 11.3.1 Storm计算结果的存储位置 224
  - 11.3.2 Storm集群动态增删节点 224
  - 11.3.3 关闭Storm相关进程 224
  - 11.3.4 Storm UI显示内容的问题 224
- 11.4 项目开发技巧 225
  - 11.4.1 使用assembly插件打包 225
  - 11.4.2 依赖JAR冲突 228
- 11.5 保证消息的可靠处理 228
  - 11.5.1 消息失败后的处理 228
  - 11.5.2 主动干预可靠性 229
  - 11.5.3 处理重复的Tuple 229
- 11.6 理解DRPC原语 230
  - 11.6.1 DRPC workflow 230
  - 11.6.2 LinearDRPCTopologyBuilder实现类 231
  - 11.6.3 DRPC的两种模式 231
- 11.7 快速理解一致性事务 232
  - 11.7.1 Trident框架的使用 233
  - 11.7.2 Trident框架的细节 234
  - 11.7.3 事务性Spout 236
  - 11.7.4 状态State 238
- 11.8 本章小结 241

## 精彩短评

- 1、扫了一下
- 2、周末去图书馆看了storm相关的几本书，觉得这本有些凑字基础的话不如《从零开始学Storm》  
使用场景及模式，不如《Storm分布式实时计算模式》  
再看看徐明明先生的storm博客，  
和 并发变成网的storm入门翻译<http://ifeve.com/getting-started-with-stom-index/>  
基本上能做到storm入门了

### 精彩书评

1、这本书主要举出了6个实例，来说明storm的应用。我只做了其中的三个实例，做不下去了。。。书中的代码有明显的错误，而且有不少地方表述不清。此书浪费的太多的空间来讲程序的流程，尽管只是一个很简单的程序，还很煞有介事的画出了不少图。再说说它的代码吧，可以直接从github上下载。可是呢？代码中的数组千疮百孔，如果自己不做任何修改的画，根本跑不起来。最令人可恨的是，写了web页面来展示数据，之前数据也存储进入了hbase，但是居然页面并没有读取hbase，而是直接从.csv文件中获取很少的数据。另外，代码的书写规范很烂的。很多代码都考虑不周。这本书的第一作者原本是做hadoop开发的，估计对storm也不是特别了解清楚吧。代码估计也是随意弄弄。本想写邮件给第一作者，居然没有留邮箱，相关的论坛也关闭了。故此在豆瓣上留下一些评论，提醒大家。如果大家想了解一些storm的一些例子，可以看看。但是本书的内容以及代码确实糟糕的很呐！！！期待有好的storm开发书出现。

## 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)