

《大数据系统构建》

图书基本信息

书名：《大数据系统构建》

13位ISBN编号：9787111552946

出版时间：2017-1

作者：Nathan Marz,James Warren

页数：282

译者：马延辉,向磊,魏东琦

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

《大数据系统构建》

内容概要

随着社交网络、网络分析和智能型电子商务的兴起，传统的数据库系统显然已无法满足海量数据的管理需求。作为一种新的处理模式，大数据系统应运而生，它使用多台机器并行工作，能够对海量数据进行存储、处理、分析，进而帮助用户从中提取对优化流程、实现高增长率的有用信息，做更为精准有效的决策。但不可忽略的是，它也引入了大多数开发者并不熟悉的、困扰传统架构的复杂性问题。本书将教你充分利用集群硬件优势的Lambda架构，以及专门用来捕获和分析网络规模数据的新工具，来创建这些系统。它将描述一个可扩展的、易于理解大数据系统的方法——可以由小团队构建并运行。本书共18章，除了介绍基本概念，其他章节采用“理论+示例”的方式来阐释相关概念，并使用现实世界中的工具加以论证。其中，第1章介绍了数据系统的原理，给出了Lambda架构的概述，并概述了构建任何数据系统的广义方法。第2~9章集中阐述Lambda架构的批处理层。第10章和第11章集中阐述服务层，让读者了解只批量写入的特定数据库——这些数据库比传统数据库更简单，它们具有出色的性能，并具备可操作性、稳健性等特点。第12~17章集中阐述速度层，让读者更明确地了解NoSQL数据库、流处理和管理增量计算的复杂性。第18章通过综合回顾Lambda架构的相关知识，帮助读者了解增量批处理、基本Lambda架构的变种，以及如何充分利用资源。

《大数据系统构建》

作者简介

作者简介

Nathan Marz Cascalog和Storm的创始人。在2011年Twitter收购社交媒体数据分析公司BackType前，他是BackType首席工程师。在Twitter，他建立了流计算团队，提供和开发共享基础设施，为整个公司的关键实时应用提供支持。他目前是Stealth startup的创始人。

James Warren Storm8的分析架构师，精通大数据处理、机器学习和科学计算。

译者简介

马延辉，资深Hadoop技术专家，对Hadoop生态系统相关技术有着深刻的理解，在Hadoop开发和运维方面积累了丰富的经验。曾就职于阿里、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、运维和管理有着深刻的理解和丰富的实战经验。开源HBase监控工具Ella作者。现在致力于大数据技术在传统行业的落地和大数据技术的普及和推广。

向磊，前暴风影音数据平台架构师，目前在某垂直电商平台担任技术总监，惠普中国Hadoop相关课程讲师。开源项目EasyHadoop、phpHiveAdmin作者，对Hadoop及其周边生态系统的底层运维及开发、集群自动化运维、网络架构设计、集群安全、性能优化、嵌入式编程方面有较深入了解。

魏东琦，博士，长期从事软件研发工作，现就职于中国地质调查局西安地质调查中心，参加、承担过多项科研项目。现致力于地质行业与大数据技术融合的相关研究工作。

书籍目录

译者序

前言

关于本书

致谢

第1章 大数据的新范式1

1.1 本书是如何组织的2

1.2 扩展传统数据库3

1.2.1 用队列扩展3

1.2.2 通过数据库分片进行扩展4

1.2.3 开始处理容错问题4

1.2.4 损坏问题5

1.2.5 到底是哪里出错了5

1.2.6 大数据技术是如何起到帮助作用的5

1.3 NoSQL不是万能的6

1.4 基本原理6

1.5 大数据系统应有的属性7

1.5.1 鲁棒性和容错性7

1.5.2 低延迟读取和更新8

1.5.3 可扩展性8

1.5.4 通用性8

1.5.5 延展性8

1.5.6 即席查询9

1.5.7 最少维护9

1.5.8 可调试性9

1.6 全增量架构的问题10

1.6.1 操作复杂性10

1.6.2 实现最终一致性的极端复杂性11

1.6.3 缺乏容忍人为错误12

1.6.4 全增量架构解决方案与 Lambda架构解决方案13

1.7 Lambda架构14

1.7.1 批处理层15

1.7.2 服务层16

1.7.3 批处理层和服务层满足几乎所有属性16

1.7.4 速度层17

1.8 技术上的最新趋势19

1.8.1 CPU并不是越来越快20

1.8.2 弹性云20

1.8.3 大数据充满活力的开源生态系统20

1.9 示例应用：SuperWebAnalytics.com21

1.10 总结22

第一部分 批处理层

第2章 大数据的数据模型24

2.1 数据的属性25

2.1.1 数据是原始的28

2.1.2 数据是不可变的30

2.1.3 数据是永远真实的33

2.2 基于事实的数据表示模型34

- 2.2.1 事实的示例及属性34
- 2.2.2 基于事实的模型的优势36
- 2.3 图模式39
 - 2.3.1 图模式的元素39
 - 2.3.2 可实施模式的必要性40
- 2.4 SuperWebAnalytics.com的完整数据模型41
- 2.5 总结42
- 第3章 大数据的数据模型：示例44
 - 3.1 为什么使用序列化框架44
 - 3.2 Apache Thrift45
 - 3.2.1 节点46
 - 3.2.2 边46
 - 3.2.3 属性47
 - 3.2.4 把一切组合成数据对象47
 - 3.2.5 模式演变48
 - 3.3 序列化框架的局限性49
 - 3.4 总结50
- 第4章 批处理层的数据存储51
 - 4.1 主数据集的存储需求52
 - 4.2 为批处理层选择存储方案53
 - 4.2.1 使用键/值存储主数据集53
 - 4.2.2 分布式文件系统54
 - 4.3 分布式文件系统是如何工作的54
 - 4.4 使用分布式文件系统存储主数据集56
 - 4.5 垂直分区58
 - 4.6 分布式文件系统的底层性质58
 - 4.7 在分布式文件系统中存储SuperWebAnalytics.com的主数据集60
 - 4.8 总结61
- 第5章 批处理层的数据存储：示例62
 - 5.1 使用HDFS62
 - 5.1.1 小文件问题64
 - 5.1.2 转向更高层次的抽象64
 - 5.2 使用Pail在批处理层存储数据65
 - 5.2.1 Pail基本操作66
 - 5.2.2 序列化对象到Pail中67
 - 5.2.3 使用Pail进行批处理操作69
 - 5.2.4 使用Pail进行垂直分区69
 - 5.2.5 Pail文件格式与压缩71
 - 5.2.6 Pail优点的总结71
 - 5.3 存储SuperWebAnalytics.com的主数据集72
 - 5.3.1 Thrift对象的结构化Pail73
 - 5.3.2 SuperWebAnalytics.com的基础Pail74
 - 5.3.3 用于垂直分区数据集的分片Pail75
 - 5.4 总结78
- 第6章 批处理层79
 - 6.1 启发性示例80
 - 6.1.1 给定时间范围内的页面浏览量80
 - 6.1.2 性别推理80
 - 6.1.3 影响力分数81

- 6.2 批处理层上的计算82
- 6.3 重新计算算法与增量算法84
 - 6.3.1 性能85
 - 6.3.2 容忍人为错误86
 - 6.3.3 算法的通用性86
 - 6.3.4 选择算法的风格87
- 6.4 批处理层中的可扩展性87
- 6.5 MapReduce：一种大数据计算的范式88
 - 6.5.1 可扩展性89
 - 6.5.2 容错性91
 - 6.5.3 MapReduce的通用性92
- 6.6 MapReduce的底层特性94
 - 6.6.1 多步计算很怪异94
 - 6.6.2 手动实现连接非常复杂94
 - 6.6.3 逻辑和物理执行紧密耦合96
- 6.7 管道图——一种关于批处理计算的高级思维方式97
 - 6.7.1 管道图的概念97
 - 6.7.2 通过MapReduce执行管道图101
 - 6.7.3 合并聚合器101
 - 6.7.4 管道图示例102
- 6.8 总结103
- 第7章 批处理层：示例104
 - 7.1 一个例证105
 - 7.2 数据处理工具的常见陷阱106
 - 7.2.1 自定义语言107
 - 7.2.2 不良的可组合抽象107
 - 7.3 JCasalog介绍108
 - 7.3.1 JCasalog的数据模型109
 - 7.3.2 JCasalog查询的结构110
 - 7.3.3 查询多个数据集111
 - 7.3.4 分组和聚合器113
 - 7.3.5 对一个查询示例进行单步调试114
 - 7.3.6 自定义谓词操作117
 - 7.4 组合121
 - 7.4.1 合并子查询122
 - 7.4.2 动态创建子查询123
 - 7.4.3 谓词宏125
 - 7.4.4 动态创建谓词宏128
 - 7.5 总结130
- 第8章 批处理层示例：架构和算法131
 - 8.1 SuperWebAnalytics.com批处理层的设计132
 - 8.1.1 所支持的查询132
 - 8.1.2 批处理视图132
 - 8.2 workflow概述135
 - 8.3 获取新数据137
 - 8.4 URL规范化137
 - 8.5 用户标识符规范化138
 - 8.6 页面浏览去重142
 - 8.7 计算批处理视图142

- 8.7.1 给定时间范围内的页面浏览量143
- 8.7.2 给定时间范围内的独立访客143
- 8.7.3 跳出率分析144
- 8.8 总结145
- 第9章 批处理层示例：实现147
 - 9.1 出发点147
 - 9.2 准备工作流148
 - 9.3 获取新数据149
 - 9.4 URL规范化152
 - 9.5 用户标识符规范化153
 - 9.6 页面浏览去重159
 - 9.7 计算批处理视图159
 - 9.7.1 给定时间范围内的页面浏览量159
 - 9.7.2 给定时间范围内的独立访客161
 - 9.7.3 跳出率分析163
 - 9.8 总结165
- 第二部分 服务层
- 第10章 服务层概述168
 - 10.1 服务层的性能指标169
 - 10.2 规范化/非规范化问题的服务层解决方案172
 - 10.3 服务层数据库的需求173
 - 10.4 设计SuperWebAnalytics.com的服务层174
 - 10.4.1 给定时间范围内的页面浏览量175
 - 10.4.2 给定时间范围内的独立访客175
 - 10.4.3 跳出率分析176
 - 10.5 对比全增量的解决方案177
 - 10.5.1 给定时间范围内的独立访客的全增量方案177
 - 10.5.2 与Lambda架构解决方案的比较182
 - 10.6 总结183
- 第11章 服务层：示例184
 - 11.1 ElephantDB的基本概念184
 - 11.1.1 ElephantDB中的视图创建185
 - 11.1.2 ElephantDB中的视图服务185
 - 11.1.3 使用ElephantDB186
 - 11.2 创建SuperWebAnalytics.com的服务层188
 - 11.2.1 给定时间范围内的页面浏览量188
 - 11.2.2 给定时间范围内的独立访客数量191
 - 11.2.3 跳出率分析191
 - 11.3 总结192
- 第三部分 速度层
- 第12章 实时视图194
 - 12.1 计算实时视图195
 - 12.2 存储实时视图197
 - 12.2.1 最终一致性198
 - 12.2.2 速度层中存储的状态总量198
 - 12.3 增量计算的挑战199
 - 12.3.1 CAP原理的有效性199
 - 12.3.2 CAP原理和增量算法之间复杂的相互作用201
 - 12.4 异步更新与同步更新202

- 12.5 过期实时视图203
- 12.6 总结205
- 第13章 实时视图：示例206
 - 13.1 Cassandra的数据模型206
 - 13.2 使用Cassandra208
 - 13.3 总结210
- 第14章 队列和流处理211
 - 14.1 队列211
 - 14.1.1 单消费者队列212
 - 14.1.2 多消费者队列214
 - 14.2 流处理214
 - 14.2.1 队列和工作节点215
 - 14.2.2 队列和工作节点的缺陷216
 - 14.3 更高层次的一次一个的流处理217
 - 14.3.1 Storm模型217
 - 14.3.2 保证消息处理221
 - 14.4 SuperWebAnalytics.com速度层223
 - 14.5 总结226
- 第15章 队列和流处理：示例227
 - 15.1 使用Apache Storm定义拓扑结构227
 - 15.2 Apache Storm集群及其部署230
 - 15.3 保证消息处理232
 - 15.4 实现SuperWebAnalytics.com给定时间范围内的独立访客的速度层233
 - 15.5 总结237
- 第16章 微批量流处理239
 - 16.1 实现有且仅有一次语义240
 - 16.1.1 强有序处理240
 - 16.1.2 微批量流处理241
 - 16.1.3 微批量流处理的拓扑结构242
 - 16.2 微批量流处理的核心概念244
 - 16.3 微批量流处理的扩展管道图245
 - 16.4 完成SuperWebAnalytics.com的速度层246
 - 16.4.1 给定时间范围内的页面浏览量246
 - 16.4.2 跳出率分析247
 - 16.5 另一个跳出率分析示例251
 - 16.6 总结252
- 第17章 微批量流处理：示例253
 - 17.1 使用Trident253
 - 17.2 完成SuperWebAnalytics.com的速度层257
 - 17.2.1 给定时间范围内的页面浏览量257
 - 17.2.2 跳出率分析259
 - 17.3 完全容错、基于内存及微批量处理265
 - 17.4 总结266
- 第18章 深入Lambda架构268
 - 18.1 定义数据系统268
 - 18.2 批处理层和服务层270
 - 18.2.1 增量的批处理270
 - 18.2.2 测量和优化批处理层的资源使用276
 - 18.3 速度层280

18.4 查询层281

18.5 总结282

《大数据系统构建》

精彩短评

- 1、高屋建瓴，又苦口婆心地介绍了一遍Lambda，是我看过这个领域最好的书或者文章。只是又想理论，又想实战这件事情是行不通的，扣一星就是因为实战部分太鸡肋了。
- 2、大数据技术集大成者

《大数据系统构建》

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