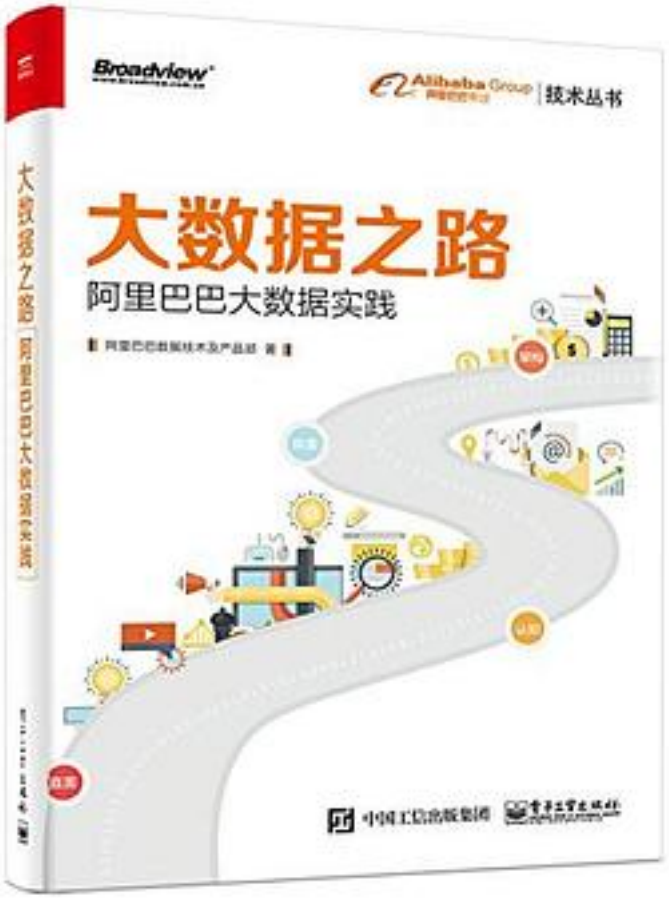
大数据之路



大数据之路_下载链接1_

著者:阿里巴巴数据技术及产品部

出版者:电子工业出版社

出版时间:2017-7-1

装帧:平装

isbn:9787121314384

在阿里巴巴集团内，数据人员面临的现实情况是：集团数据存储已经达到EB级别，部分单张表每天的数据记录数高达几千亿条；在2016年“双11购物狂欢节”的24小时中，支付金额达到了1207亿元人民币，支付峰值高达12万笔/秒，下单峰值达17.5万笔/秒

，媒体直播大屏处理的总数据量高达百亿级别且所有数据都需要做到实时、准确地对外披露……巨大的信息量给数据采集、存储和计算都带来了极大的挑战。

《大数据之路：阿里巴巴大数据实践》就是在此背景下完成的。《大数据之路：阿里巴巴大数据实践》中讲到的阿里巴巴大数据系统架构，就是为了满足不断变化的业务需求，同时实现系统的高度扩展性、灵活性以及数据展现的高性能而设计的。

《大数据之路：阿里巴巴大数据实践》由阿里巴巴数据技术及产品部组织并完成写作，是阿里巴巴分享对大数据的认知，与生态伙伴共创数据智能的重要基石。相信《大数据之路：阿里巴巴大数据实践》中的实践和思考对同行会有很大的启发和借鉴意义。

作者介绍:

目录: 第1章 总述1
第1篇 数据技术篇
第2章 日志采集 8
2.1 浏览器的页面日志采集 8
2.1.1 页面浏览日志采集流程 9
2.1.2 页面交互日志采集 14
2.1.3 页面日志的服务器端清洗和预处理 15
2.2 无线客户端的日志采集 16
2.2.1 页面事件 17
2.2.2 控件点击及其他事件 18
2.2.3 特殊场景 19
2.2.4 H5 & Native日志统一 20
2.2.5 设备标识 22
2.2.6 日志传输 23
2.3 日志采集的挑战 24
2.3.1 典型场景 24
2.3.2 大促保障 26
第3章 数据同步 29
3.1 数据同步基础 29
3.1.1 直连同步 30
3.1.2 数据文件同步 30
3.1.3 数据库日志解析同步 31
3.2 阿里数据仓库的同步方式 35
3.2.1 批量数据同步 35
3.2.2 实时数据同步 37
3.3 数据同步遇到的问题与解决方案 39
3.3.1 分库分表的处理 39
3.3.2 高效同步和批量同步 41
3.3.3 增量与全量同步的合并 42
3.3.4 同步性能的处理 43
3.3.5 数据漂移的处理 45
第4章 离线数据开发 48
4.1 数据开发平台 48
4.1.1 统一计算平台 49
4.1.2 统一开发平台 53
4.2 任务调度系统 58
4.2.1 背景 58
4.2.2 介绍 60
4.2.3 特点及应用 65

第5章 实时技术 68
5.1 简介 69
5.2 流式技术架构 71
5.2.1 数据采集 72
5.2.2 数据处理 74
5.2.3 数据存储 78
5.2.4 数据服务 80
5.3 流式数据模型 80
5.3.1 数据分层 80
5.3.2 多流关联 83
5.3.3 维表使用 84
5.4 大促挑战&保障 86
5.4.1 大促特征 86
5.4.2 大促保障 88
第6章 数据服务 91
6.1 服务架构演进 91
6.1.1 DWSOA 92
6.1.2 OpenAPI 93
6.1.3 SmartDQ 94
6.1.4 统一的数据服务层 96
6.2 技术架构 97
6.2.1 SmartDQ 97
6.2.2 iPush 100
6.2.3 Lego 101
6.2.4 uTiming 102
6.3 最佳实践 103
6.3.1 性能 103
6.3.2 稳定性 111
第7章 数据挖掘 116
7.1 数据挖掘概述 116
7.2 数据挖掘算法平台 117
7.3 数据挖掘中台体系 119
7.3.1 挖掘数据中台 120
7.3.2 挖掘算法中台 122
7.4 数据挖掘案例 123
7.4.1 用户画像 123
7.4.2 互联网反作弊 125
第2篇 数据模型篇
第8章 大数据领域建模综述 130
8.1 为什么需要数据建模 130
8.2 关系数据库系统和数据仓库 131
8.3 从OLTP和OLAP系统的区别看模型方法论的选择 132
8.4 典型的数据仓库建模方法论 132
8.4.1 ER模型 132
8.4.2 维度模型 133
8.4.3 Data Vault模型 134
8.4.4 Anchor模型 135
8.5 阿里巴巴数据模型实践综述 136
第9章 阿里巴巴数据整合及管理体系 138
9.1 概述 138
9.1.1 定位及价值 139
9.1.2 体系架构 139
9.2 规范定义 140
9.2.1 名词术语 141

9.2.2 指标体系 141
9.3 模型设计 148
9.3.1 指导理论 148
9.3.2 模型层次 148
9.3.3 基本原则 150
9.4 模型实施 152
9.4.1 业界常用的模型实施过程 152
9.4.2 OneData实施过程 154
第10章 维度设计 159
10.1 维度设计基础 159
10.1.1 维度的基本概念 159
10.1.2 维度的基本设计方法 160
10.1.3 维度的层次结构 162
10.1.4 规范化和反规范化 163
10.1.5 一致性维度和交叉探查 165
10.2 维度设计高级主题 166
10.2.1 维度整合 166
10.2.2 水平拆分 169
10.2.3 垂直拆分 170
10.2.4 历史归档 171
10.3 维度变化 172
10.3.1 缓慢变化维 172
10.3.2 快照维表 174
10.3.3 极限存储 175
10.3.4 微型维度 178
10.4 特殊维度 180
10.4.1 递归层次 180
10.4.2 行为维度 184
10.4.3 多值维度 185
10.4.4 多值属性 187
10.4.5 杂项维度 188
第11章 事实表设计 190
11.1 事实表基础 190
11.1.1 事实表特性 190
11.1.2 事实表设计原则 191
11.1.3 事实表设计方法 193
11.2 事务事实表 196
11.2.1 设计过程 196
11.2.2 单事务事实表 200
11.2.3 多事务事实表 202
11.2.4 两种事实表对比 206
11.2.5 父子事实的处理方式 208
11.2.6 事实的设计准则 209
11.3 周期快照事实表 210
11.3.1 特性 211
11.3.2 实例 212
11.3.3 注意事项 217
11.4 累积快照事实表 218
11.4.1 设计过程 218
11.4.2 特点 221
11.4.3 特殊处理 223
11.4.4 物理实现 225
11.5 三种事实表的比较 227
11.6 无事实的事实表 228

11.7 聚集型事实表 228
11.7.1 聚集的基本原则 229
11.7.2 聚集的基本步骤 229
11.7.3 阿里公共汇总层 230
11.7.4 聚集补充说明 234
第3篇 数据管理篇
第12章 元数据 236
12.1 元数据概述 236
12.1.1 元数据定义 236
12.1.2 元数据价值 237
12.1.3 统一元数据体系建设 238
12.2 元数据应用 239
12.2.1 Data Profile 239
12.2.2 元数据门户 241
12.2.3 应用链路分析 241
12.2.4 数据建模 242
12.2.5 驱动ETL开发 243
第13章 计算管理 245
13.1 系统优化 245
13.1.1 HBO 246
13.1.2 CBO 249
13.2 任务优化 256
13.2.1 Map倾斜 257
13.2.2 Join倾斜 261
13.2.3 Reduce倾斜 269
第14章 存储和成本管理 275
14.1 数据压缩 275
14.2 数据重分布 276
14.3 存储治理项优化 277
14.4 生命周期管理 278
14.4.1 生命周期管理策略 278
14.4.2 通用的生命周期管理矩阵 280
14.5 数据成本计量 283
14.6 数据使用计费 284
第15章 数据质量 285
15.1 数据质量保障原则 285
15.2 数据质量方法概述 287
15.2.1 消费场景知晓 289
15.2.2 数据加工过程卡点校验 292
15.2.3 风险点监控 295
15.2.4 质量衡量 299
第4篇 数据应用篇
第16章 数据应用 304
16.1 生意参谋 305
16.1.1 背景概述 305
16.1.2 功能架构与技术能力 307
16.1.3 商家应用实践 310
16.2 对内数据产品平台 313
16.2.1 定位 313
16.2.2 产品建设历程 314
16.2.3 整体架构介绍 317
附录A 本书插图索引 320
・・・・・・(收起)

大数据之路_下载链接1_

标签

大数据

阿里巴巴

数据

架构

数据分析

数据仓库

计算机

技术

评论

遇到有价值的内容，就说限于篇幅……然后就没有了，写书还有限于篇幅这种说法，那你写来做甚？

阿里巴巴把这本书当成产品文档来写了，使用的术语和一般技术书籍在内延上会有细微区别，可能是行业限制。整体来说，对阿里巴巴现有的技术架构介绍得比较清楚，各技术环节使用什么产品，达到什么效果，都描述得较清晰，只要有一定售前经验的从业人员都能理解。然而技术演进思路不甚了了，建议可以结合《淘宝技术这十年》结合起来一起看。至于大数据推进为目的的阅读者，基本上洗洗睡吧，不要对这么书抱太大的期望，只能用来做扫盲，而且阅读者本身要有一定开发经验才容易有点心得。

泥水味好浓

教科书级别

真心读不下去，不值8分

以前团队出的书，写得不错。

系统地从规范，模型，平台，应用等多层次对阿里大数据产品实践做了一遍梳理，值得一读。

数据产品体系的第一部分主要讲技术构建经验，从浏览器和app的数据采集开始，到数据的同步处理，离线数据的处理，实时数据的处理，到数据服务的架构演进和实践，以及数据挖掘的平台和算法建设。第二部分关注数据仓库的建模分析技术，维度设计和事实表设计部分经验值得关注，第三部分的数据管理关注的数据元数据，计算，存储和数据质量，最后是数据应用的案例，作为阿里经验的分享不乏真知灼见，值得阅读。

平台这边有不少值得借鉴,需要再整理一下能实际落地应用的部分.
维度设计和事实表设计那里太枯燥了读不进去

看个热闹。

把阿里巴巴内部大数据建设的每一个点都拿出来讲了，全面到令人发指，从技术选型，到管理规范，从设计原理，到字段命名。看这本书相当于借阿里的经验去了解大数据全景。看了太多“in action”，“guide book”，"principle”的书，偶尔看看这种“建设经验”，角度清奇，2017年的书可惜印刷量太少，已经绝版了。

数据服务这一章带给了我很多新知识。阿里是国内少有的，不是瞎搞概念，而是真的在大数据上做了很多实事的公司了。极有可能是国内做得最好的了。

---
kaisi借我的一本书。。

年度技术书推荐. 横向来说涵盖了数据业务的各个方面,
纵向来说有各个技术方案的背景,原因,演化路径. 限制. 非常难得.
最近正在做流式计算部分, 有一些需求想不清楚能不能接, 看到阿里都都不支持,
我们也果断不支持了....

本书可以作为进入大数据的一本入门书（不涉及技术问题，仅涉及使用场景），可以比较快速的了解大数据在阿里（其他公司同理）的使用方式和场景。其中技术内容不多，场景和想法比较多。

适合有数据技术基础从业人员观看。

冰冻三尺，非一日之寒。阿里数据体系全图。业务不涉及，很多章节只能脑补。

参考架构，道略少

公司出的书，写到这个细致的程度，仁至义尽了

读到第一章收集数据部分，我们现在数据收集人肉工作量太大，纯业务埋点，无无痕埋点。闭环没闭上。书中能看到自己当时经历的那一部分只是阿里大数据大版图中很小很小的一块。庆幸现在的平台。全书读完，发现主要集中在ODPS内容。后半部分的数据仓库方法论读起来太干，快翻。

大数据之路_下载链接1_

书评

数据是公司的资产已经成了事实上的信仰，从数据洞察商业规律，为决策提供支撑，创造价值，为商业赋能，一直是IT的愿景使命和不懈追求之一。在小数据时代，各大企业、机构的探索和努力方向主要体现在BI和数据仓库等应用上，对于当时数量相对有限、结构严谨有序的数据，这些工具...

这本书作为我的2018开年第二本阅读的技术书籍，读完之后感觉受益良多
第一，对于整个大数据的体系有了更多且清晰的认知
第二，对于不同系统的逻辑处理方式给予了引导
第三，毕竟是阿里多年技术的累计产出，而且都是阿里技术大牛写的，干货相当多
最后，如果对于大数据方向想有...

HW产品配置系统部部长推荐语：
几年前，有人提出“人类正从IT时代走向DT时代”，社会上也不少人著书立说，纷繁解读，大家一时躁动不止。今天，以物联网、云计算、大数据、人工智能等为代表的新技术革命正在渗透着各行各业，并在悄悄的深深的影响、改变着我们的生活。出门...

大数据之路_下载链接1_