

# 《ODPS权威指南》

## 图书基本信息

书名：《ODPS权威指南》

13位ISBN编号：9787115372411

出版时间：2014-12

作者：李妹芳

页数：360

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](http://www.tushu000.com)

# 《ODPS权威指南》

## 内容概要

ODPS (Open Data Processing Service) 是阿里巴巴自主研发的海量数据处理和分析的服务平台，主要应用于数据分析、海量数据统计、数据挖掘、机器学习和商业智能等领域。目前，ODPS不仅在阿里内部得到广泛应用，享有很好的口碑，正逐步走向第三方开放市场。

本书是学习和掌握ODPS的权威指南，作者来自阿里ODPS团队。全书共13章，主要内容包括：ODPS入门、整体架构、数据通道、MapReduce编程、SQL查询分析、安全，以及基于真实数据的各种场景分析实战。本书基于很多范例解析，通过在各种应用场景下的示例来说明如何通过ODPS完成各种需求，以期引导读者从零开始轻松掌握和使用ODPS。同时，本书不局限于示例分析，也致力于提供更多关于大数据处理的编程思想和经验分享。书中所有示例代码都可以在作者提供的网站上免费下载。本书是学习和掌握ODPS的权威指南，作者来自阿里ODPS团队。

本书包括以下重要内容：

ODPS概览及其基本知识；

如何高效地使用ODPS SQL；

MapReduce编程和进阶应用；

ODPS机器学习算法；

ODPS权限、资源和数据管理；

深入了解ODPS体系结构和高级机制。

书中所有示例代码都可以通过[https://github.com/duckrun/odps\\_book](https://github.com/duckrun/odps_book)免费下载。

本书适合想要了解和使用ODPS的读者阅读学习，对于从事大数据存储和应用以及分布式计算的专业人士来说，也是很好的参考资料。

# 《ODPS权威指南》

## 作者简介

李妹芳，阿里数据平台事业部工程师，曾译有《Linux系统编程》、《数据之美》、《数据可视化之美》等书，她喜欢儿童文学，她的微博是<http://weibo.com/duckrun>。

## 书籍目录

### 前言 7

### 第1章 ODPS概述 9

#### 1.1 引言 9

#### 1.2 初识ODPS 9

##### 1.2.1 背景和挑战 9

##### 1.2.2 为什么做ODPS 10

##### 1.2.3 ODPS是什么 10

##### 1.2.4 ODPS做什么 11

#### 1.3 基本概念 11

##### 1.3.1 账号 (Account) 12

##### 1.3.2 项目空间 (Project) 13

##### 1.3.3 表 (Table) 13

##### 1.3.4 分区 (Partition) 14

##### 1.3.5 任务 (Task)、作业 (Job) 和作业实例 (Instance) 14

##### 1.3.6 资源 (Resource) 14

#### 1.4 应用开发模式 15

##### 1.4.1 RESTful API 15

##### 1.4.2 ODPS SDK 18

##### 1.4.3 ODPS CLT 18

##### 1.4.4 管理控制台 18

##### 1.4.5 IDE 18

#### 1.5 一些典型场景 19

##### 1.5.1 阿里金融数据仓库 19

##### 1.5.2 CNZZ数据仓库 19

##### 1.5.3 支付宝账号影响力圈 19

##### 1.5.4 阿里金融水文衍生算法 19

##### 1.5.5 阿里妈妈广告CTR预估 20

#### 1.6 现状和前景 20

#### 1.7 小结 21

### 第2章 ODPS入门 22

#### 2.1 准备工作 22

##### 2.1.1 创建云账号 22

##### 2.1.2 开通ODPS服务 24

#### 2.2 使用管理控制台 24

#### 2.3 配置ODPS客户端 26

##### 2.3.1 下载和配置CLT 26

##### 2.3.2 准备dual表 28

##### 2.3.3 CLT运行模式 30

##### 2.3.4 下载和配置dship 30

##### 2.3.5 通过dship上传下载数据 31

#### 2.4 网站日志分析实例 32

##### 2.4.1 场景和数据说明 32

##### 2.4.2 需求分析 33

##### 2.4.3 数据准备 33

##### 2.4.4 创建表并添加分区 34

##### 2.4.5 数据解析和导入 35

##### 2.4.6 数据加工 39

2.4.7 数据分析	43
2.4.8 自动化运行	47
2.4.9 应用数据集市	49
2.4.10 结果导出	51
2.4.11 结果展现	51
2.4.12 删除数据	53
2.5 小结	53
第3章 收集海量数据	54
3.1 DSHIP工具	54
3.2 收集WEB日志	56
3.2.1 场景和需求说明	56
3.2.2 问题分析和设计	56
3.2.3 实现说明	57
3.2.4 进一步探讨	59
3.2.5 为什么这么难	61
3.3 MYSQL数据同步到ODPS	61
3.3.1 场景和需求说明	61
3.3.2 问题分析和实现	61
3.3.3 进一步探讨	63
3.4 下载结果表	63
3.5 小结	63
第4章 使用SQL处理海量数据	64
4.1 ODPS SQL是什么	64
4.2 入门示例	64
4.2.1 场景说明	64
4.2.2 简单的DDL操作	64
4.2.3 生成数据	68
4.2.4 单表查询	69
4.2.5 多表连接JOIN	71
4.2.6 高级查询	79
4.2.7 多表关联UNION ALL	87
4.2.8 多路输出 ( MULTI-INSERT )	88
4.3 网站日志分析	88
4.3.1 准备数据和表	89
4.3.2 维度表	89
4.3.3 访问路径分析	96
4.3.4 TopK查询	97
4.3.5 IP黑名单	98
4.4 天猫品牌预测	103
4.4.1 主题说明和前期准备	103
4.4.2 理解数据	104
4.4.3 两个简单的实践	106
4.4.4 问题分析和算法设计	108
4.4.5 生成特征	109
4.4.6 抽取正负样本	111
4.4.7 生成模型	114
4.4.8 验证模型	115
4.4.9 预测结果	118
4.4.10 进一步探讨	118

4.5 小结	118
第5章 SQL进阶	120
5.1 UDF是什么	120
5.2 入门示例	120
5.3 实际应用案例	122
5.3.1 URL解码	122
5.3.2 简单的LBS应用	123
5.3.3 网站访问日志UserAgent解析	125
5.4 SQL实现原理	129
5.4.1 词法分析	130
5.4.2 语法分析	130
5.4.3 逻辑分析	130
5.4.4 物理分析	136
5.5 SQL调优	137
5.5.1 数据倾斜	137
5.5.2 一些优化建议	140
5.5.3 一些注意事项	141
5.6 小结	141
第6章 通过TUNNEL迁移数据	142
6.1 ODPS TUNNEL 是什么	142
6.2 入门示例	142
6.2.1 下载和配置	142
6.2.2 准备数据	142
6.2.3 上传数据	143
6.2.4 下载数据	148
6.3 TUNNEL原理	149
6.3.1 数据如何传输	149
6.3.2 客户端和服务端如何交互	150
6.3.3 如何实现高并发	151
6.4 从HADOOP迁移到ODPS	151
6.4.1 问题分析	151
6.4.2 客户端实现和分析	152
6.4.3 Mapper实现和分析	155
6.4.4 编译和运行	157
6.4.5 进一步探讨	159
6.5 一些注意点	159
6.6 小结	160
第7章 使用MAPREDUCE处理数据	161
7.1 MAPREDUCE编程模型	161
7.2 MAPREDUCE应用场景	163
7.3 初识ODPS MAPREDUCE	164
7.4 入门示例	165
7.4.1 准备工作	165
7.4.2 问题分析	165
7.4.3 代码实现和分析	166
7.4.4 运行和输出分析	169
7.4.5 扩展：使用Combiner？	171
7.5 TOPK查询	173
7.5.1 场景和数据说明	174

7.5.2 问题分析	174
7.5.3 具体实现分析	175
7.5.4 运行和结果输出	179
7.5.5 扩展：忽略Stop Words	180
7.5.6 扩展：数据和任务统计	182
7.5.7 扩展：MR2模型	184
7.6 SQL和MAPREDUCE，用哪个？	186
7.7 小结	186
第8章 MAPREDUCE进阶	187
8.1 再谈SHUFFLE & SORT	187
8.2 好友推荐	188
8.2.1 场景和数据说明	188
8.2.2 问题定义和分析	189
8.2.3 代码实现	190
8.3 LBS应用探讨：周边定位	193
8.3.1 场景和数据说明	193
8.3.2 问题定义和分析	194
8.3.3 代码实现和分析	195
8.3.4 运行和测试	199
8.4 MAPREDUCE调试	200
8.4.1 带bug的代码	200
8.4.2 通过本地模式调试	201
8.4.3 通过Counter调试	201
8.4.4 通过log调试	202
8.5 一些注意点	203
8.6 小结	204
第9章 机器学习算法	205
9.1 初识ODPS算法	205
9.2 入门示例	205
9.2.1 通过CLT统计分析	205
9.2.2 通过XLab统计分析	207
9.3 几个经典的算法	209
9.3.1 逻辑回归	209
9.3.2 随机森林	210
9.4 天猫品牌预测	211
9.4.1 逻辑回归	211
9.4.2 随机森林	218
9.4.3 脚本实现和自动化	228
9.4.4 进一步探讨	231
9.5 小结	232
第10章 使用SDK访问ODPS服务	233
10.1 主要的PACKAGE和接口	233
10.1.1 主要的Package	233
10.1.2 核心接口	233
10.2 入门示例	233
10.3 基于ECLIPSE插件开发	235
10.4 小结	236
第11章 ODPS账号、资源和数据管理	237
11.1 权限管理	237

11.1.1 账号授权	237
11.1.2 角色 ( Role ) 授权	240
11.1.3 ACL授权特点	241
11.1.4 简单的Policy授权	242
11.1.5 Role Policy	243
11.1.6 ACL授权和Policy授权小结	245
11.2 资源管理	245
11.2.1 Project内的资源管理	246
11.2.2 跨Project的资源共享	246
11.3 数据管理	247
11.3.1 表生命周期	248
11.3.2 数据归并 ( Merge )	249
11.3.3 数据保护 ( Project Protection )	249
11.4 小结	251
第12章 深入了解ODPS	253
12.1 体系架构	253
12.1.1 客户端	254
12.1.2 接入层	254
12.1.3 逻辑层	254
12.1.4 存储/计算层	255
12.2 执行流程	256
12.2.1 提交作业	256
12.2.2 运行作业	256
12.2.3 查询作业状态	256
12.2.4 执行逻辑图	256
12.3 底层数据存储	257
12.3.1 CFILE是什么	257
12.3.2 CFILE逻辑结构	257
12.4 内聚式框架	258
12.4.1 元数据	258
12.4.2 运维管理	258
12.4.3 多控制集群和多计算集群	259
12.5 跨集群复制	260
12.5.1 数据迁移	260
12.5.2 跨集群同步	261
12.6 小结	264
第13章 探索ODPS之美	265
13.1 R语言数据探索	265
13.1.1 安装和配置	265
13.1.2 一些基本操作	265
13.1.3 分析建模	265
13.2 实时流计算	267
13.3 图计算模型	268
13.4 准实时SQL	269
13.5 机器学习平台	270
附录一 ODPS消息认证机制	271
后记	274



## 精彩短评

- 1、阿里移动算法推荐比赛有那么一点儿用指南。。。
- 2、深入浅出，浅显易懂
- 3、差不多是ODPS的第一手资料
- 4、基本上还算一本不错的入门，虽然细节方面谈的不多，底层也不够深入，但毕竟是少有的ODPS书籍，且覆盖面很全，例子也还行
- 5、详细的实战入门；给出的案例场景都很不错。
- 6、之前一直在hadoop生态下工作，来阿里开始使用odps，也读了本书。最大的感触是，绝大多数的数据处理工作，原来都是可用sql完成的。比如mapjoin可cover大多数需要distributed cache的情况，distributed by sort by可让你轻松定义partition方式和reducer中排序方式，group by中自动应用了combiner（目测），再加上row\_number()（窗口函数），UDF等元素，工作效率大幅提升。
- 7、感觉这本书已过时，过时是因为好几年前的书，很多新的内容都没收录，但里面讲述的内容都是可用的，对于阿里内部员工而言，内部文档更详细，更新也更及时。
- 8、感觉内容组织非常混乱，看完一遍下来讲的是什么完全没印象；部分章节内容已经落伍，已经通过sql实现的功能为何还要讲如何用hadoop MR实现？每章讲一点就开始讲工具怎么安装，分析代码。应用场景太多，不是每个读者都需要，集中放在最后两章内容就够了。真正odps自己的东西讲的太少。
- 9、不是很详细，可以阅读，看看以后哪些场景，自己可以利用odps平台去做。感慨一下，odps的功能，我们仅用到他的10%都不到。对于了解odps能应用到哪些场景，这本书，还是不错的。

1、谈起ODPS，还得从阿里金融的故事说起。一直以来，阿里金融始终是ODPS的第一客户，见证了ODPS一路的成长历程。几年的坚持和信任，我们一起走了过来，而且越走越好。2010年初，集群规模只有几十台，为了完成阿里金融的信贷产品的模型计算，每天增量同步1TB左右的数据，执行几十个模型计算，运行时间在18小时左右。当时问题较多，实际上是24小时人肉运维，大家都习惯了凌晨下班，一起解决各种问题。期间的痛自不必说，但一点点的进步，都让人充满喜悦。2011年初，集群规模达到100多台，数据规模达到数百TB，模型计算任务量是原来的10倍左右，而运行时间却不到原来的1/3。集群能力完成计算任务游刃有余，大家第一次体会到一种说不清的舒畅。2012年，ODPS集群规模达到1500台，阿里金融数仓的所有数据计算都运行在上面，数据规模达到数PB，运行任务数千个。用户体验也得到不断改善。2013年，ODPS单集群规模达到5000台，阿里金融的数据仓库专家们，不再需要考虑集群方面的问题（如升级、扩容、运维等），可以专注于自己的业务，包括数据采集、ETL和数据仓库构建、BI分析和报表，通过分布式编程模型生成特征、衍生指标，通过统计和机器学习构建风险控制模型，把分析建模后的结果数据导出到线上系统服务，其中涉及数据安全性、正确性，平台稳定性和易用性等诸多方面。阿里小贷推出了“3-1-0”服务条款：3分钟申请、1秒钟获贷和0人工审批，其背后实质上是“准入资质评估、个性化授信和风险监控”，而这一切离不开海量数据计算的支撑！基于ODPS，阿里金融可以充分挖掘大数据的价值，实现数据化运营，在大促期间创下了30分钟贷款5亿的纪录！有了强大的存储和计算支持，各种创新业务不断开花结果。BI团队也逐渐把业务迁移到ODPS上，和使用SAS相比，性能上有了很大提升。阿里金融不但锤炼了ODPS，其成功也为ODPS赢得了口碑。在阿里巴巴集团内，淘宝、支付宝、阿里妈妈的业务都开始运行在ODPS集群。此外，外部的一些独立软件开发商也在使用ODPS。回首走过的路，我们充满感恩，尤其感谢阿里金融的一路陪伴。这些年的辛苦耕耘，这些年的积累和沉淀，我们也更有信心！ODPS作为一个海量数据处理平台，涉及很多前沿技术领域，包括分布式、云计算和大数据等。本书的定位是帮助ODPS用户快速了解如何使用ODPS解决其实际问题，在内容介绍上是以用户应用场景为中心，对功能和技术的介绍都是围绕并服务于这一中心。作者假设用户是带着如何使用ODPS解决自身的大数据问题来阅读本书，期望这本书能够帮助用户解决实际问题。...致谢感谢所有为本书付出努力的同事们！要感谢的人太多，在此不一一列出（见本书后记）。但我却不能不特别提到阿里巴巴研究员张东晖先生，如果没有他的指导、帮助和鼓励，就不会有这本书。最后，衷心希望这本书能带给你美好的ODPS编程之旅！

## 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](http://www.tushu000.com)