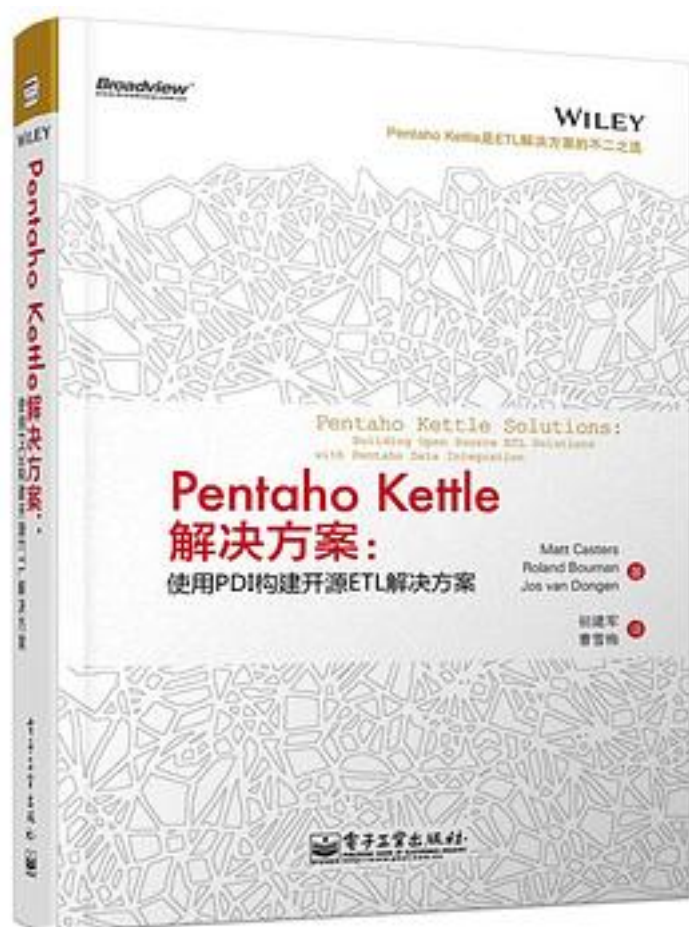


Pentaho

Kettle解决方案：使用PDI构建开源ETL解决方案



[Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案_下载链接1](#)

著者:【美】 Matt Casters (马特.卡斯特)

出版者:电子工业出版社

出版时间:2014-3

装帧:平装

isbn:9787121224454

《Pentaho

Kettle解决方案：使用PDI构建开源ETL解决方案》主要介绍如何使用开源ETL工具来完成数据整合工作。

《Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案》介绍的PDI(Kettle)是一种开源的 ETL 解决方案，书中介绍了如何使用PDI来实现数据的剖析、清洗、校验、抽取、转换、加载等各类常见的ETL类工作。

除了ODS/DW类比较大型的应用外，Kettle 实际还可以为中小企业提供灵活的数据抽取和数据处理的功能。Kettle除了支持各种关系型数据库、HBase、MongoDB这样的NoSQL数据源外，它还支持Excel、Access这类小型的数据源。并且通过插件扩展，Kettle 可以支持各类数据源。本书详细介绍了Kettle可以处理的数据源，而且详细介绍了如何使用Kettle抽取增量数据。

Kettle 的数据处理功能也很强大，除了选择、过滤、分组、连接、排序这些常用的功能外，Kettle 里的Java表达式、正则表达式、Java脚本、Java类等功能都非常灵活而强大，都非常适合于各种数据处理功能。本书也使用了一些篇幅介绍Kettle这些灵活的数据处理功能。

《Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案》后面章节介绍了如何在Kettle 上开发插件，如何使用Kettle处理实时数据流，以及如何在Amazon AWS上运行Kettle 等一些高级主题。

除了介绍PDI(Kettle)工具的使用和功能，本书还结合Kimball博士的数据仓库和ETL子系统的理论，从实践的角度介绍数据仓库的模型设计、数据仓库的构建方法，以及如何使用 PDI实现Kimball博士提出的34种ETL子系统。

作者介绍:

Matt Casters是一位具有多年工作经验的独立商业智能顾问。他为许多大公司建立了无数个数据仓库和BI解决方案。在过去的8年里，Matt Casters把自己的时间都用于研发一个ETL工具——Kettle。2005年12月，Kettle成为开源软件。2006年初期，Kettle走进Pentaho。随后，Matt就职于Pentaho，成为数据集成总监。在Pentaho,他继续从事Kettle的研发工作。Matt致力于帮助建设Kettle社区，回答论坛上的提问，有时在世界会议上发表演讲。博客：<http://www.ibridge.be>。Twitter：@mattcasters。

Roland Bouman目前从事前台页面和商业智能的研发工作。他从1998年开始从事IT行业。多年来一直致力于开源软件的研发，尤其是数据库技术、商业智能以及页面开发框架。同时，Roland Bouman还是MySQL和Pentaho社区的成员。他经常参加MySQL使用者会议、OSCON、Pentaho社区等国际会议。Roland Bouman不仅是MySQL 5.1 Cluster Certification Guide和Pentaho Solutions两本书的合著者之一，也是MySQL和Pentaho相关书籍的技术评论家。技术博客：<http://rpbouman.blogspot.com>。Twitter：@rolandbouman。

Jos van Dongen是一位著名的商业智能专家、作家和演说家。他从1991年开始从事软件开发、商业智能以及数据仓库等领域的工作。Jos van Dongen曾先后就职于顶级的系统集成公司和管理咨询公司。1998，他创立了自己的咨询公司，Tholis Consulting。他为许多商业和福利组织构建了

BI和数据仓库系统。Jos为丹麦 Database Magazine撰写了新的BI研发成果，并且经常在国内和国际会议上发言。Jos van Dongen撰写了一本关于开源BI的书，并且和Roland Bouman合作编写了 Pentaho Solutions。更多信息参考：<http://www.tholis.com>。Twitter: @josvandongen。

目录: 第一部分: 开始

第1章 ETL入门.....	2
1.1 OLTP和数据仓库对比	2
1.2 ETL是什么	3
1.2.1 ETL解决方案的演化过程	4
1.2.2 ETL基本构成	5
1.3 ETL、ELT和EII	6
1.3.1 ELT	6
1.3.2 EII: 虚拟数据整合	7
1.4 数据整合面临的挑战	8
1.4.1 方法论: 敏捷BI	9
1.4.2 ETL设计	10
1.4.3 获取数据	10
1.4.4 数据质量	12
1.5 ETL工具的功能	13
1.5.1 连接	13
1.5.2 平台独立	14
1.5.3 数据规模	14
1.5.4 设计灵活性	14
1.5.5 复用性	15
1.5.6 扩展性	15
1.5.7 数据转换	15
1.5.8 测试和调试	16
1.5.9 血统和影响分析	16
1.5.10 日志和审计	16
1.6 小结	17
第2章 Kettle基本概念.....	18
2.1 设计原则	18
2.2 Kettle设计模块	19
2.2.1 转换	19
2.2.2 作业	23
2.2.3 转换或作业的元数据	28
2.2.4 数据库连接	28
2.2.5 工具	31
2.2.6 资源库	31
2.2.7 虚拟文件系统	31
2.3 参数和变量	32
2.3.1 定义变量	32
2.3.2 命名参数	33
2.3.3 使用变量	33
2.4 可视化编程	34
2.4.1 开始	34
2.4.2 创建新的步骤	35
2.4.3 放在一起	36
2.5 小结	38
第3章 安装和配置.....	39
3.1 Kettle软件概览	39
3.1.1 集成开发环境: Spoon	40

3.1.2 命令行启动: Kitchen和Pan	42
3.1.3 作业服务器: Carte	42
3.1.4 Encr.bat和encr.sh	42
3.2 安装	43
3.2.1 Java环境	43
3.2.2 安装 Kettle	43
3.3 配置	46
3.3.1 配置文件和.kettle目录	46
3.3.2 用于启动Kettle程序的shell脚本	51
3.3.3 管理 JDBC 驱动	52
3.4 小结	53
第4章 ETL示例解决方案——Sakila.....	54
4.1 Sakila	54
4.1.1 sakila示例数据库	55
4.1.2 租赁业务的星型模型	57
4.2 预备知识和一些基础的Spoon技巧	60
4.2.1 安装ETL解决方案	60
4.2.2 Spoon使用	60
4.3 ETL示例解决方案	61
4.3.1 生成静态维度	62
4.3.2 循环加载	64
4.4 小结	80
第二部分: ETL	
第5章 ETL子系统.....	82
5.1 34种子系统介绍	82
5.1.1 抽取	83
5.1.2 清洗和更正数据	84
5.1.3 数据发布	86
5.1.4 管理ETL环境	89
5.2 小结	91
第6章 数据抽取.....	92
6.1 Kettle数据抽取概览	92
6.1.1 文件抽取	93
6.1.2 数据库抽取	97
6.1.3 Web数据抽取	98
6.1.4 基于流的和实时的抽取	99
6.2 处理ERP和CRM系统	100
6.2.1 ERP 挑战	100
6.2.2 Kettle ERP插件	101
6.2.3 处理SAP数据	101
6.2.4 ERP和CDC 问题	104
6.3 数据剖析	105
6.4 CDC: 变更数据捕获	110
6.4.1 基于源数据的CDC	111
6.4.2 基于触发器的CDC	113
6.4.3 基于快照的CDC	113
6.4.4 基于日志的CDC	116
6.4.5 哪个CDC方案更适合你	117
6.5 发布数据	117
6.6 小结	118
第7章 清洗和校验.....	119
7.1 数据清洗	120
7.1.1 数据清洗步骤	121
7.1.2 使用参照表	123

7.1.3 数据校验	127
7.2 错误处理	130
7.2.1 处理过程错误	131
7.2.2 转换错误	132
7.2.3 处理数据（校验）错误	133
7.3 审计数据和过程质量	136
7.4 数据排重	137
7.4.1 去除完全重复的数据	137
7.4.2 不完全重复问题	138
7.4.3 设计排除重复记录的转换	139
7.5 脚本	142
7.5.1 公式	143
7.5.2 Java脚本	143
7.5.3 用户自定义Java表达式	144
7.5.4 正则表达式	145
7.6 小结	146
第8章 处理维度表.....	147
8.1 管理各种键	148
8.1.1 管理业务键	148
8.1.2 生成代理键	149
8.2 加载维度表	154
8.2.1 雪花维度表	154
8.2.2 星型维度表	159
8.3 缓慢变更维度	161
8.3.1 缓慢变更维类型	161
8.3.2 类型1的缓慢变更维	161
8.3.3 类型2的缓慢变更维	163
8.3.4 其他类型的缓慢变更维	167
8.4 更多维度	168
8.4.1 生成维（Generated Dimensions） ..	168
8.4.2 杂项维度（Junk Dimensions）	169
8.4.3 递归层次	170
8.5 小结	171
第9章 加载事实表.....	172
9.1 批量加载	173
9.1.1 STDIN和FIFO	173
9.1.2 Kettle批量加载	174
9.1.3 批量加载一般要考虑的问题	176
9.2 维度查询	176
9.2.1 维护参照完整性	176
9.2.2 代理键管道	177
9.2.3 迟到数据	179
9.3 处理事实表	182
9.3.1 周期快照和累积快照	182
9.3.2 面向状态的事实表	183
9.3.3 加载周期快照表	185
9.3.4 加载累积快照表	185
9.3.5 加载面向状态事实表	186
9.3.6 加载聚集表	186
9.4 小结	187
第10章 处理OLAP数据.....	188
10.1 OLAP的价值和挑战.....	189
10.1.1 OLAP 存储类型	190
10.1.2 OLAP在系统中的位置	191

10.1.3 Kettle OLAP选项	191
10.2 Mondrian	192
10.3 XML/A服务	194
10.4 Palo	197
10.4.1 建立Palo 连接	198
10.4.2 Palo 架构	199
10.4.3 读Palo数据	200
10.4.4 写Palo数据	202
10.5 小结	204
第三部分：管理和部署	
第11章 ETL开发生命期.....	206
11.1 解决方案设计	206
11.1.1 好习惯和坏习惯	206
11.1.2 ETL流设计	209
11.1.3 可重用性和可维护性	209
11.2 敏捷开发	210
11.3 测试和调试	214
11.3.1 测试活动	214
11.3.2 ETL测试	215
11.3.3 调试	218
11.4 解决方案文档化	220
11.4.1 为什么实际情况文档很少	220
11.4.2 Kettle的文档功能.....	221
11.4.3 生成文档	222
11.5 小结	223
第12章 调度和监控.....	224
12.1 调度	224
12.1.1 操作系统级调度	225
12.1.2 使用Pentaho 内置的调度程序	228
12.2 监控	232
12.2.1 日志	232
12.2.2 邮件通知	234
12.3 小结	237
第13章 版本和移植.....	238
13.1 版本控制系统	238
13.1.1 基于文件的版本控制系统	239
13.1.2 内容管理系统	240
13.2 Kettle 元数据	240
13.2.1 Kettle XML 元数据	241
13.2.2 Kettle 资源库元数据	242
13.3 管理资源库	244
13.3.1 导出和导入资源库	244
13.3.2 资源库升级	245
13.4 版本移植系统	245
13.4.1 管理XML文件	245
13.4.2 管理资源库	246
13.4.3 解决方案参数化	246
13.5 小结	248
第14章 血统和审计.....	249
14.1 批量血统抽取	250
14.2 血统	251
14.2.1 血统信息	251
14.2.2 影响分析信息	252
14.3 日志和操作元数据	254

14.3.1 日志基础	254
14.3.2 日志架构	255
14.3.3 日志表	257
14.4 小结	262
第四部分：性能和扩展性	
第15章 性能调优.....	264
15.1 转换性能：找到最弱连接	264
15.1.1 通过简化找到性能瓶颈	265
15.1.2 通过度量值找到性能瓶颈	266
15.1.3 复制数据行	267
15.2 提高转换性能	269
15.2.1 提高读文本文件的性能	269
15.2.2 写文本文件时使用延迟转换	271
15.2.3 提高数据库性能	272
15.2.4 数据排序	275
15.2.5 减少CPU消耗	276
15.3 提高作业性能	280
15.3.1 作业里的循环	280
15.3.2 数据库连接池	281
15.4 小结	281
第16章 并行、集群和分区.....	283
16.1 多线程	283
16.1.1 数据行分发	284
16.1.2 记录行合并	285
16.1.3 记录行再分发	285
16.1.4 数据流水线	286
16.1.5 多线程的问题	287
16.1.6 作业中的并行执行	289
16.2 使用Carte子服务器	289
16.2.1 配置文件	289
16.2.2 定义子服务器	290
16.2.3 远程执行	291
16.2.4 监视子服务器	291
16.2.5 Carte安全	291
16.2.6 服务	292
16.3 集群转换	293
16.3.1 定义一个集群模式	293
16.3.2 设计集群转换	294
16.3.3 执行和监控	295
16.3.4 元数据转换	296
16.4 分区	298
16.4.1 定义分区模式	299
16.4.2 分区的目标	300
16.4.3 实现分区	300
16.4.4 内部变量	301
16.4.5 数据库分区	301
16.4.6 集群转换中的分区	302
16.5 小结	302
第17章 云计算中的动态集群.....	303
17.1 动态集群	303
17.1.1 建立动态集群	304
17.1.2 使用动态集群	306
17.2 云计算	306
17.3 EC2	307

17.3.1 如何使用EC2	307
17.3.2 成本	307
17.3.3 自定义AMI	307
17.3.4 打包新AMI	310
17.3.5 中止AMI	310
17.3.6 运行主节点	310
17.3.7 运行子节点	311
17.3.8 使用EC2 集群	312
17.3.9 监控	313
17.3.10 轻量原则和持久性	314
17.4 小结	314
第18章 实时数据整合	315
18.1 实时ETL介绍	315
18.1.1 实时处理面临的挑战	316
18.1.2 需求	316
18.2 基于流的转换	317
18.2.1 一个基于流的转换实例	318
18.2.2 调试	321
18.2.3 第三方软件和实时整合	321
18.2.4 Java 消息服务	322
18.3 小结	324
第五部分：高级主题	
第19章 Data Vault管理	326
19.1 Data Vault 模型介绍	327
19.2 你是否需要Data Vault	327
19.3 Data Vault的组成部分	328
19.3.1 中心表	328
19.3.2 链接表	329
19.3.3 附属表	329
19.3.4 Data Vault 特点	331
19.3.5 构建 Data Vault 模型	331
19.4 将Sakila的例子转换成Data Vault 模型	331
19.4.1 Sakila中心表	331
19.4.2 Sakila 链接表	332
19.4.3 Sakila 附属表	333
19.5 加载Data Vault 模型：简单的ETL解决方案	334
19.5.1 安装Sakila Data Vault	335
19.5.2 安装ETL方案	335
19.5.3 创建一个数据库账户	335
19.5.4 ETL解决方案的例子	335
19.5.5 加载 Data Vault 表	341
19.6 从Data Vault 模型更新数据集市	341
19.6.1 ETL解决方案例子	342
19.6.2 dim_actor 转换	342
19.6.3 dim_customer 转换	343
19.6.4 dim_film 转换	346
19.6.5 dim_film_actor_bridge 转换	347
19.6.6 fact_rental 转换	347
19.6.7 加载星型模型里的所有表	349
19.7 小结	349
第20章 处理复杂数据格式	350
20.1 非关系型和非表格型的数据格式	350
20.2 非结构化的表格型数据	351
20.2.1 处理多值字段	351

20.2.2 处理重复的字段组	352
20.3 半结构化和非结构化数据	353
20.4 键/值对	358
20.5 小结	362
第21章 Web Services.....	363
21.1 Web 页面和Web Services	363
21.2 数据格式	365
21.2.1 XML	365
21.2.2 HTML	366
21.2.3 JavaScript Object Notation	367
21.3 XML例子	369
21.3.1 XML例子文件	369
21.3.2 从XML中抽取数据	371
21.3.3 生成XML文档	378
21.4 SOAP例子	384
21.4.1 使用“Web服务查询”步骤	385
21.4.2 直接访问 SOAP服务	386
21.5 JSON例子	389
21.5.1 Freebase项目	389
21.5.2 使用Kettle 抽取Freebase数据	392
21.6 RSS	396
21.6.1 RSS结构	396
21.6.2 Kettle对RSS的支持	398
21.7 小结	403
第22章 Kettle集成.....	404
22.1 Kettle API	404
22.1.1 LGPL协议	404
22.1.2 Kettle Java API	405
22.2 执行存在的转换和作业	406
22.2.1 执行一个转换	406
22.2.2 执行一个作业	407
22.3 应用程序中嵌入Kettle	408
22.3.1 Pentaho 报表	408
22.3.2 把数据放到转换里	410
22.3.3 动态转换	413
22.3.4 动态模板	416
22.3.5 动态作业	416
22.3.6 在Kettle里执行动态ETL	419
22.3.7 Result	419
22.3.8 替换元数据	420
22.4 OEM版本和二次发布版本	421
22.4.1 创建PDI的OEM版本	421
22.4.2 Kettle的二次发布（Forking）	422
22.5 小结	423
第23章 扩展Kettle.....	424
23.1 插件架构	424
23.1.1 插件类型	425
23.1.2 架构	425
23.1.3 前提	425
23.2 转换步骤插件	428
23.2.1 StepMetaInterface	428
23.2.2 StepDataInterface	434
23.2.3 StepDialogInterface	434
23.2.4 StepInterface	440

23.3 用户自定义 Java 类步骤 444

23.3.1 传递元数据 444

23.3.2 访问输入和字段 445

23.3.3 代码片段 445

23.3.4 例子 445

23.4 作业项插件 446

23.4.1 JobEntryInterface 446

23.4.2 JobEntryDialogInterface 448

23.5 分区插件 448

23.6 资源库插件 450

23.7 数据库类型插件 450

23.8 小结 451

附录A Kettle生态群..... 452

附录B Kettle 企业版特性..... 456

附录C 内置的变量和属性参考 457

• • • • • [\(收起\)](#)

[Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案 下载链接1](#)

标签

KETTLE

ETL

数据分析

数据仓库

BI

数据挖掘

数据

计算机

评论

相关资料好少啊。都打算自己写一本了

聚焦在整体解决方案上，对ETL本身，尤其是缺乏必要文档的Kettle工具本身介绍有限

好久没有看工具书，最近在做ETL方面的工作，通读了一遍，基本上都是大方面的解决方案，关于Kettle的细节点没有说得很清楚，感觉排版也不是特别好

当前工作时派上用场的大宝典。

kettle的资料不多，还好这本写得比较全

[Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案_下载链接1](#)

书评

[Pentaho Kettle解决方案：使用PDI构建开源ETL解决方案_下载链接1](#)