深入BI之 Kettle 篇 第1周

2013.03.30

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

学习资料

1.http://kettle.pentaho.com/ http://wiki.pentaho.com/
2.http://infocenter.pentaho.com
3.<<Kettle Cook Book>>
4.<<Pentaho 3.2 Data Integration Beginner's Guide>>
5.<<Kettle Solution>>
6.Kettle 代码

本课程使用的 Kettle 版本：
1.Kettle 4.4 下载地址：http://kettle.pentaho.com/
2.傲飞数据整合平台 1.0.4，下载地址：
http://www.pentahochina.com

Kettle 介绍 – 课程内容

- 背景知识：ETL
- Kettle 介绍、应用情况、对比
- Kettle 基本使用

Kettle 背景知识 – ETL

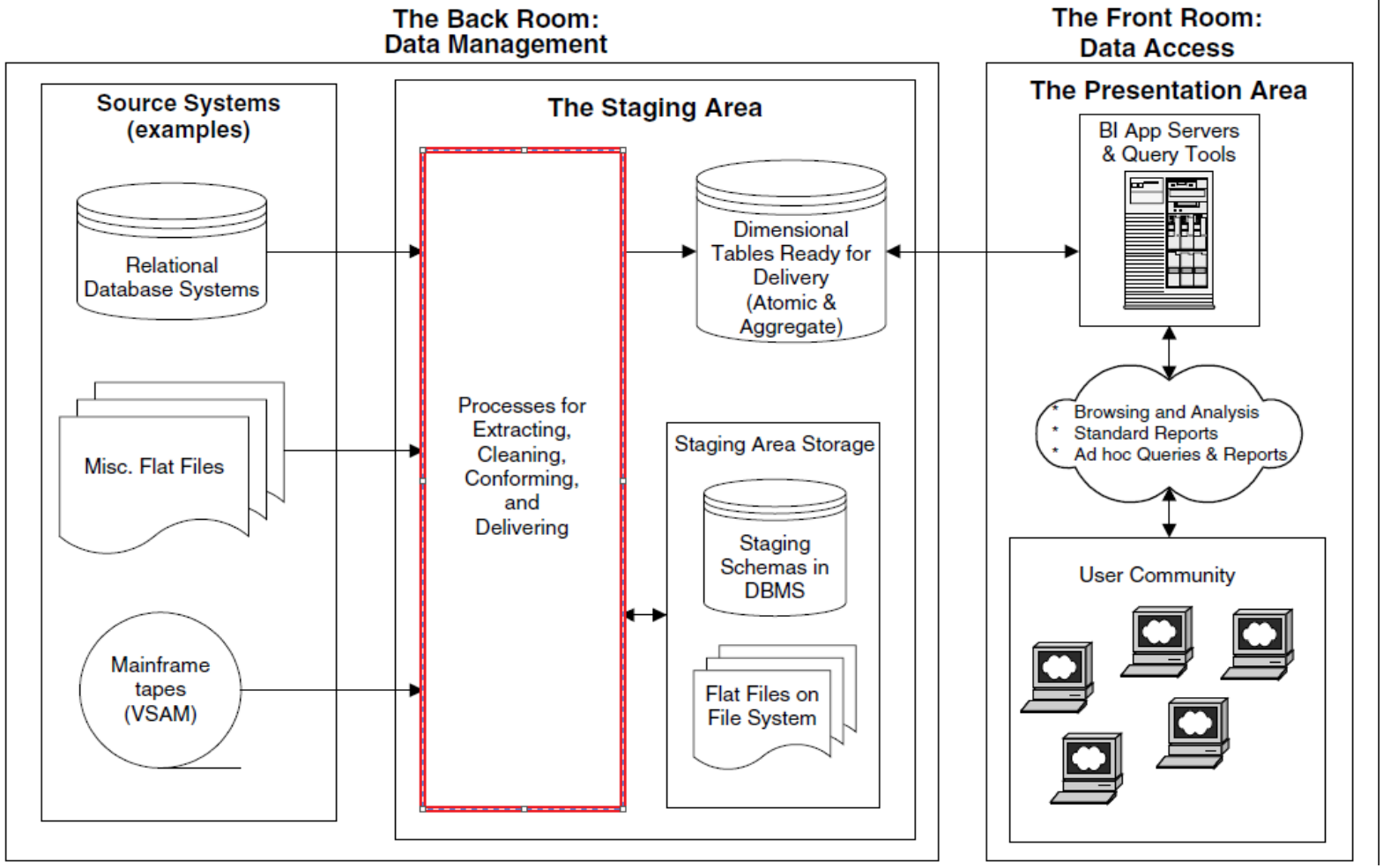
抽取(Extract)：一般抽取过程需要连接到不同的数据源，以便为随后的步骤提供数据。这一部分看上去简单而琐碎，实际上它是 ETL 解决方案的成功实施的一个主要障碍。

转换(Transform)：任何对数据的处理过程都是转换。这些处理过程通常包括（但不限于）下面一些操作：
移动数据
根据规则验证数据
数据内容和数据结构的修改
将多个数据源的数据集成
根据处理后的数据计算派生值和聚集值

加载(Load)：将数据加载到目标系统的所有操作。

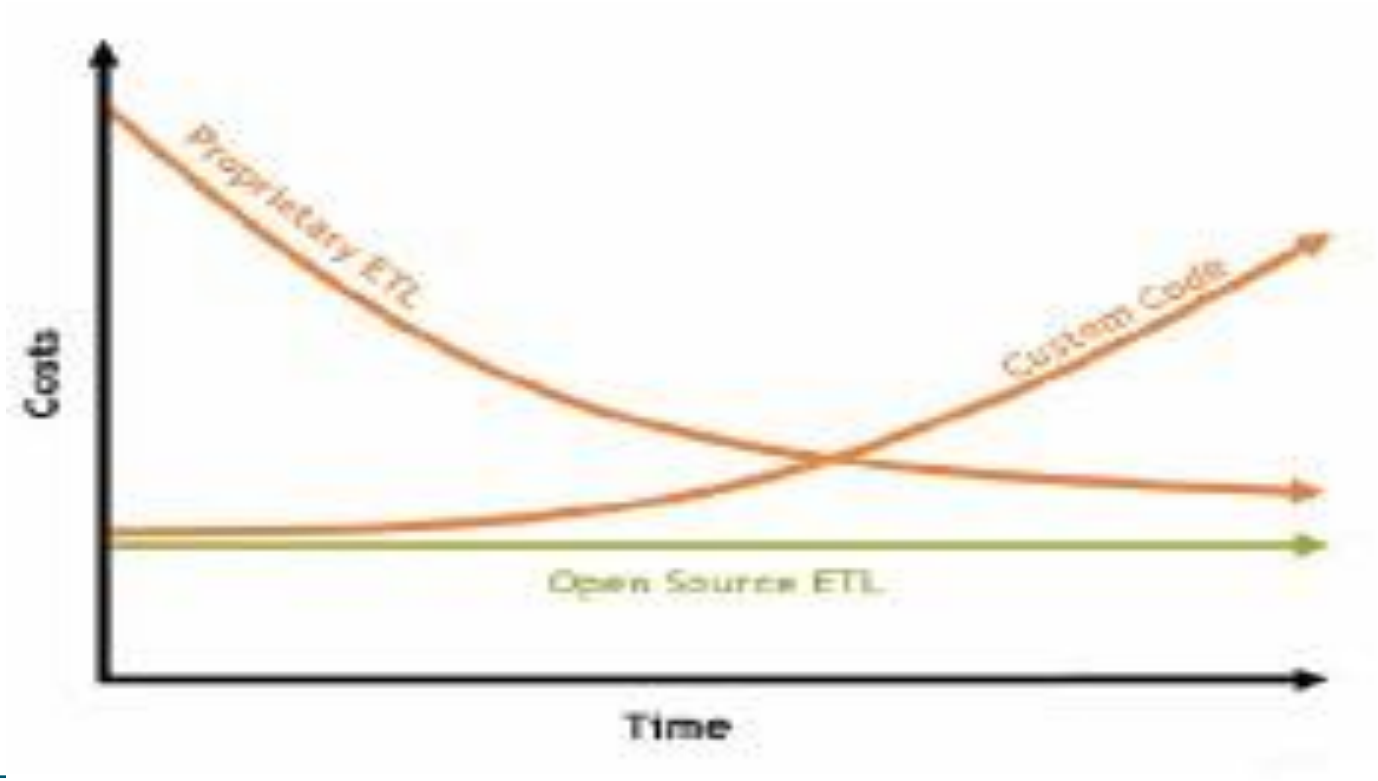
概念扩展：ELT，EII(Enterprise information integration)/Data federation

Kettle背景知识– ETL在BI中的作用



Kettle背景知识– ETL 实现方式

- 手工编码，编写脚本，Java，Python
- 商业ETL 工具软件
- 开源ETL 工具软件



Kettle背景知识– ETL 商业软件

Informatica

IBM DataStage

Microsoft SSIS

Oracle ODI

Kettle背景知识– ETL 开源软件

Kettle

Talend

CloverETL

Ketl，Octopus …

Kettle 介绍 – 课程内容

- 背景知识：ETL
- Kettle 介绍、应用情况、对比
- Kettle 基本使用

Kettle 介绍 – 名称

Kettle ： Kettle is an acronym for “Kettle E.T.T.L. Environment”. This means it has been designed to help you with your ETTL needs: the Extraction, Transformation, Transportation and Loading of data.

Pentaho Data Integration（Kettle）是一款开源的ETL(Extract Transformation Load) 工具，用来完成数据的抽取，清洗、转换和加载等数据处理方面的工作。

Kettle 介绍 – 基本情况

源代码下载地址：
svn://source.pentaho.org/svnkettleroot/Kettle/trunk
官方文档：http://infocenter.pentaho.com
Bug报告地址：http://jira.pentaho.com/browse/PDI
官方论坛：
http://forums.pentaho.org/forumdisplay.php?f=135
中文论坛：http://www.pentahochina.com
当前版本：Version 4.4 (2013年)
原作者：　Matt
License:　4.3 以前 LGPL，4.3 以后改为Apache 2

Kettle 历史

2006年 Kettle 2.2， Kettle 2.3 （Kettle 开源，License 为 LGPL）
2007年 Kettle 2.4， Kettle 2.5（被Pentaho 公司收购，更名为 PDI）
2008年 Kettle 3.0 ， Kettle 3.1
2009年 Kettle 3.2 （一个使用时间较长的稳定版本）
2010年 Kettle 4.0 ， Kettle 4.1
2011年 Kettle 4.2
2012年 Kettle 4.3 ， Kettle 4.4 （License 变更为 Apache 2,支持大数据）
2013年 Kettle 5.0

Kettle 特点


流程式设计
方便易用
全面优化
高效稳定
全面的数据
访问支持
插件架构
扩展性好
多种方式
应用集成
支持
多平台
商业/社区
支持

Kettle vs Talend

测试环境：

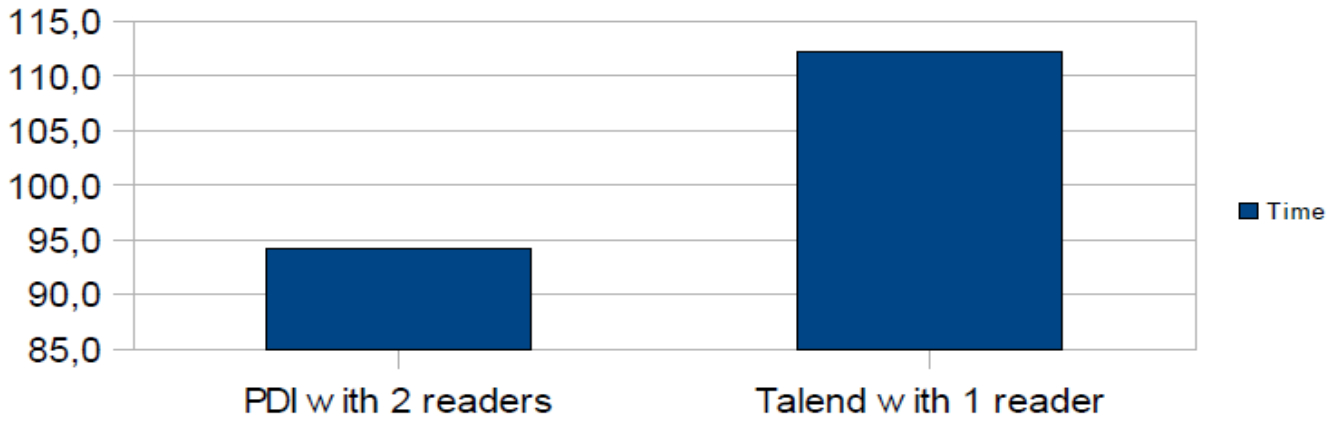
CPU Intel(R) Core(TM)2 CPU T7600 @ 2.33GHz
Disk 90GB 7200 rpm laptop disk
Memory 3.3GB, 666Mhz
OS Kubuntu 8.10 : Intrepid Ibex
Linux kernel 2.6.27-8
Filesystem used ext3
External USB disk USB 2.0, 100GB, ext3 formatted, used to write target file to
Source file: http://mattcasters.s3.amazonaws.com/customers-25M.txt
Size file 2.614.561.970 bytes
Nr of rows in file 25.000.001 with one header row

Kettle 介绍 – 对比

Kettle vs Talend

测试用例：
读取一个2.4GB 大小的 csv 文件，文件包含 25,000,000 条记录。

Description	Time
PDI with 2 readers	94,2
Talend with 1 reader	112,2
Difference : PDI is	***19% faster***



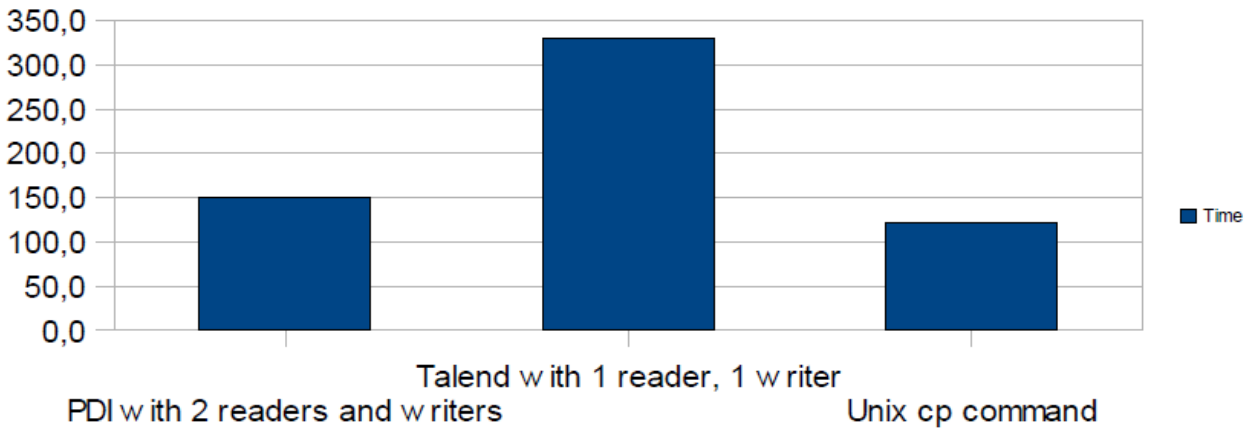
Kettle 介绍 – 对比

Kettle vs Talend

测试用例：
读取一个2.4GB 大小的 csv 文件，文件包含 25,000,000 条记录。并把这个文件写入到另一个磁盘。

Reading and writing back	**Time**
PDI with 2 readers and writers	149,3
Talend with 1 reader, 1 writer	329,4
Unix cp command	122,2

Difference : PDI is ***121% faster***



Kettle vs Talend

Talend uses a maximum of between 80-99% CPU (not multi-threaded) PDI uses a maximum of 130% CPU (1.3 CPU used) in read and write test.

Talend is faster in the single threaded reading of a file.

The file size (2.4GB) is large enough to NOT fit into the cache.

Kettle 介绍 – 对比

Kettle vs Informatica

相似点：

•Pentaho 和 Informatica 都提供了大量的转换步骤、脚本功能，都可以处理复杂的ETL 转换。

•通常情况下 Informatica 比 Kettle 更快。Informatica 有下推优化，缓存查询等提高性能的手段。但是如果你对 Kettle 和数据库有足够的了解，做一些调整，你可以提高 Kettle 的速度，在一些情况下可以达到甚至超过 Informatica 的速度。

Kettle 的优点：

•Kettle 的易用性比 Informatica 好，需要的培训要少很多。

•Kettle 不需要像 Informatica 那样大的前期投入。

•Kettle 的插件架构支持快速定制开发

Informatica 的优点：

•Informatica 的错误报告功能比 Kettle 更友好，更容易定位错误。Kettle 通常只把异常抛出，需要实施人员有更丰富的经验。

•Informatica 比 Kettle 有更好的监控工具和负载均衡等企业级应用功能，更适合大规模的ETL 应用。。

Kettle 介绍 – 国内应用

公安部

Kettle 介绍 – 课程内容

- 背景知识：ETL
- Kettle 介绍、特点、应用情况
- Kettle 基本使用

Kettle 介绍 – Kettle 基本使用

Kettle 的几个子程序的功能和启动方式

Spoon.bat: 图形界面方式启动作业和转换设计器。
Pan.bat: 命令行方式执行转换。
Kitchen.bat: 命令行方式执行作业。
Carte.bat: 启动web服务，用于 Kettle 的远程运行或集群运行。
Encr.bat: 密码加密

转换和作业

Kettle 的 Spoon 设计器用来设计转换（Transformation）和 作业（Job）。

- 转换主要是针对数据的各种处理，一个转换里可以包含多个步骤（Step）。
- 作业是比转换更高一级的处理流程，一个作业里包括多个作业项（Job Entry），一个作业项代表了一项工作，转换也是一个作业项。

保存作业

用户通过 Spoon 创建的转换、作业、数据库连接等可以保存在资源库和 XML 文件中。

- 转换文件以 ktr 为扩展名，作业文件以 kjb 为扩展名
- 资源库可以是各种常见的数据库。可以在 Spoon 中自动创建资源库，资源库默认用户名和密码是 admin/admin。

Kettle 介绍 – 输入步骤简介

输入类步骤用来从外部获取数据，可以获取数据的数据源包括，文本文件（txt，csv，xml，json）数据库、 Excel 文件等桌面文件，自定义的数据等。对特殊数据源和应用需求可以自定义输入插件。

例子：生成随机数步骤

Kettle 介绍 – 转换步骤简介

转换类步骤是对数据进行各种形式转换所用到的步骤。

例子：
字段选择
计算器
增加常量

Kettle 介绍 – 流程步骤简介

流程步骤是用来控制数据流的步骤。一般不对数据进行操作，只是控制数据流。

例子：
过滤步骤

Kettle 介绍 – 连接步骤简介

连接步骤用来将不同数据集连接到一起。

例子：
笛卡尔乘积

Kettle 介绍 – 输出步骤简介

输出步骤是输出数据的步骤，常见的输出包括文本文件输出、表输出等，可以根据应用的需求开发插件以其他形式输出。

例子：
表输出

Kettle 介绍 – 书面作业

生成 100 个随机数，随机数取值于[0，100）之间，计算小于等于 50 的随机数个数和 大于50 的随机数个数。

并把这两个统计数字放在数据库表的一行的两列中，即输出的结果有一行，一行包括两列，每列是一个统计值。

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间