# 1 etl入门

## 1.5 etl工具的功能

### 1.5.1 血统分析和影响分析

两个信息：转换的元数据信息，转换的数据流信息。

血统分析和影响分析基于元数据的两种信息。

血统分析是回溯机制，查看数据的来源；

影响分析是源数据字段对目标表和随后转换的影响。

# 2 kettle的基本概念

## 2.2 kettle设计模块

### 2.2.5 工具

Kitchen:运行作业的命令行工具

Pan:运行转换的命令行工具。

## 2.3 参数和变量

### 2.3.1 定义变量

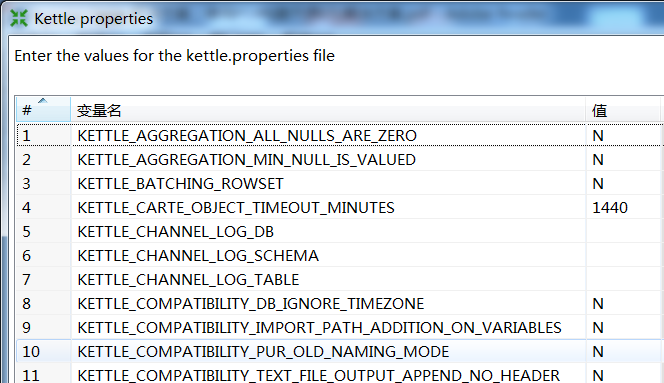
变量代表任意长度的字符串，变量可以是系统级别的（java虚拟机容器），也可以是作业内动态赋予。

初始化变量两种方式：系统设置和用户定义；

系统变量包括：jvm的变量（java.io.tmpdir）,kettle定义内部变量（Internal.Kettle.version）

用户自定义变量：多种定义方式。

{KETTLE\_HOME}/.kettle/Kettle.properties文件保存：



在作业执行的时候动态设置。

1. 作业中的set valiables步骤中定义
2. js步骤中中定义

### 2.3.2 命名参数

如果感觉一个变量，需要作为转换和作业的参数，可以定义命名参数。

命名参数的好处：参数可以通过名称显示的表示出了。可以在作业或者转换中指定位置参数的值。

a.命令行：pan或者ktchen命令行。

b.作业中的作业项。

### 2.3.3 使用变量

# 5 ETL子系统

Wiley 2002年出版，作者kimball

《the data warehouse toolkit,second edition》

Wiley 2004年出版，作者kimball

《The data warehouse etl toolkit》 ---形成了ETL里面的38个子系统

Wiley 2008年出版,作者kimball

《The data warehouse lifecycle toolkit》---ETL子系统被重构为34种子系统

## 5.1 34种子系统介绍

抽取【从数据源获取数据】：子系统1-3

清洗和更正【转换和集成数据，统一口径】：子系统4-8

发布【加载和更新数据仓库数据】：子系统9-21

管理【管理和监控ETL解决方案所有组件的处理过程】：子系统22-34

### 5.1.1 抽取

子系统1-3：数据剖析，增量数据捕获和抽取

【1】子系统1：数据剖析系统----分析不同数据源的结构和内容。

类似NULL值个数统计，行统计。

【2】子系统2：增量数据捕获系统----捕获源系统里数据的变化

Kettle中的一些步骤可以通过时间戳，快照的方式获得变化的数据。

【3】子系统3：抽取系统----从不同的数据源抽取数据，并且输入到ETL流程里。

Kimball明确区分了基于文件的和基于流的两种抽取。从kettle的角度看，这种区分方法不太恰当，无论从数据库，文件实时数据源，web services只有可以访问数据源，都可以以流的方式。抽取的主要区别是静态还是动态的问题。

### 5.1.2 清洗和更正数据

统一口径的问题：比如男女的表示方式的不同

【4】子系统4：数据清洗和质量处理系统。

可以在原系统中进行数据清洗。

可以在ETL项目中提高数据质量；

ETL项目总进行数据清洗的优势：1，在数据剖析阶段，能找出哪些错误数据，2，源系统使用的数据清洗规则，同样可以拥有etl环境中。3.业务人员可以加入到etl开发中，只有业务人员指导如何清洗数据

通常，数据拥有者业务人员，源系统开发人员，etl开发人员，需要共同的完成数据的清洗。

【5】子系统5：错误事件处理

目的：记录下ETL过程中的每一个错误；

Kettle中很多功能能处理错误事件日志，不需要独立的错误事件模块来处理。

【6】子系统6：审计维度

审计维度表是一类特殊的维度表，审计维度表包括了对事实表变更的元数据；

如数据加载日期，数据质量指标等；可以带来很大好处。

如多维数据仓库使用data vault带来的好处一样；

【7】子系统7：排除重复记录系统

排除重复数据，或者把不同系统里面互相冲突的数据统一；

客户数据是最容易产生重复数据的地方；

大多数情况下，可以采用模糊查询、正则表达式匹配，soundex函数以及各种数据挖掘技术解决这类问题；

DQ Guru是开源的数据去重的工具；

【8】子系统8：数据一致性

经过数据质量提高的步骤之后，交给数据一致性系统

目的是：使得来源于多个业务系统的事实数据遵照相同的维度；

处理方法：解决两个系统的事实表指向同一个维度表的常用方法是，维度表中保留从不同系统带来的自然键。

### 5.1.3 数据发布

更新维度表有很多方式，需要生成查询正确的维度键，取保维度表在事实表加载之前，加载完毕。加载事实数据，可能需要更新事实数据。

【9】缓慢变更维度处理

缓慢变更维度处理scds是多维数据仓库或者总线架构的基础；

维度表保存了对事实表进行分析或分组的信息；

例如，如果业务系统修改了客户所在的城市，缓慢变更维度的过程也会根据不同的规则了变更数据仓库中的数据维度

a. 覆盖：直接用新值代替旧值

b. 增加新行：把当前行标记为old并且设置一个结束的时间戳

同事创建一个新行，并且设置一个开始的时间戳

1. 增加新列：给表增加一个新列来存放新值，同时保留原来的值不变
2. 增加一个小维度表：把经常变更的表熟悉从主维度表里面分离出来，保存在自己的表里。
3. 分离历史表：把每次变化保存到一个历史表中，同时保存变化的类型和变化的时间。

这样历史表就可以回答变化情况

1. 混合型：a,b,c三种方式混合使用

【10】子系统10：代理键生成系统

Kettle里的增加序列的步骤，可以自动生成或者会议数据库里的序列来生成代理键

这种步骤适用于创建或者第一次加载维度表。而不适用于以后修改维度表。

可以使用维度查询/更新步骤和联合查询步骤来解决更新维度的问题。

这两个步骤有三种方式来生成代理键。

1. 使用表里现在代理键的最大值+1
2. 使用数据库序列
3. 使用一个自增的字段

【11】子系统11：层次维度构建

构建固定的，可变深度的，可有级别缺失的层次维度系统。

层次可以让用户分析查看维度不同级别上的数据；

最简单的层次概念就是时间维度的层次。

在时间维度里，所有级别的深度都是固定一样的。

组织结构的维度更复杂一些，通常是不平衡的或者称为可变深度的或级别缺失的。

比尔省-市-县-区 有些地区是没有区的，造成缺失级别。

【12】子系统12：特殊维度生成系统

* 杂项维度也称为垃圾维度：

一些零散的属性，分析需要，但是不适合放在其他维度表里。

例如状态标志，yes/no 都可以放在杂项维度表里。

* 小维度：

从大维度表里分离出经常发生变化的一些属性，单独放在一个小维度表中

这些维度称为SCD4

* 收缩的或者上卷的维度：

普通维度表的子集，是根据普通维度表创建和更新的。

这种维度适用于聚集的数据

* 静态维度：

通常是小的字典表，这类表在源系统中没有对应的数据，如状态码的描述或者性别。

* 用户自定义维度：

源系统中没有的，而报表需要自定义的描述，分组和层次。

可以是自动的维度。唯一区别就是，这些维度是用户来维护的，不是通过ETL过程来维护。

【13】子系统13：事实表加载

全名：事务粒度、周期快照粒度、累积快照粒度事实表的加载系统

加载事实表之前，需要把数据准备好

* 事务粒度事实表：以一个事务或者事件作为单位，例如电话呼叫记录
* 周期快照事实表：事实表并不保存全部数据，只保存固定时间间隔的数据例如每天或者每月的库存水平或者每个月的账号余额；
* 累积快照事实表：当有新的数据时候，更新事实表中的数据，数据仓库总是保存最新的数据。例如订单过程，付款日期，收获日期，实际发货日期，会随着进程不断更新事实表中的记录；

【14】子系统14：代理键管道

负责抽取正确的代理键，用于加载事实表。

Kettle的流查询步骤和数据库查询步骤代替了查询代理键。

【15】多值维度桥接生成系统

处理不同深度的层次时候需要桥接表。

比如一个公司，有子公司和子子公司，每一级都可能去购买商品，如果想看这个公司一共购买了多少商品，就需要使用桥接表来实现。

当多个事实表和其他维度表关联的时候，需要使用桥接表。

例如汇总一个演员有多少电影票房搜人，就需要电影和电影演员维度之间建立一个桥接表。

【16】迟到数据处理

维度查询和更新，默认有valid\_from,valid\_to字段；

【17】子系统17：维度管理系统

中心控制系统，用来准备和向数据仓库发布正确的维度；

负责组织管理所有和维度相关的任务

【18】子系统18：事实表管理系统

负责创建、组织、管理和事实表相关的任务。

子系统17和18字啊一起结伴工作。事实表的管理需要获取维度管理的维度。

【19】子系统19：聚集创建

聚集表是为提高数据库的性能。聚集表的管理，oracle等有自动聚集导航功能。

【20】子系统20：OLAP Cube创建系统

OLAP有特殊的存储结构，当加载的时候，可以预先聚集数据。

一些OLAP数据库只能写，不能更新。

其他OLAP数据库可以更新事实表，但是有自己的更新机制

【21】子系统21：数据整合系统

这个子系统拥有从数据仓库获取数据，并把数据发送到其他环境中，通常用于离线数据分析或者其他特殊目的。如给特定用户发送报表。

### 5.1.4 管理ETL环境

14个子系统来完成管理功能

【22】子系统22：作业调度：

社区版本的kettle不提供自己的调度功能，而是依赖pentaho bi的调度功能或者linux crong功能

【23】子系统23：备份系统

备份ETL过程中产生的中间数据也应该是ETL方案的一部分

【24】子系统24：恢复和重新启动子系统

ETL失败到时候，可以重新启动，避免数据丢失，而且需要避免数据重复

【25,26】子系统25：版本控制系统

子系统26：开发到测试，到生产的版本移植系统

ETL系统的各个部分应该有个主版本号，ETL系统作为一个整体，也需要有个主版本号。

【27】子系统27：工作流监控

Kettle中日志监控就是工作流监控

【28】子系统28：排序系统

对于排序，分组合并操作，都要事先对数据进行排序，kettle中有排序步骤，但是对于大文件，需要单独一个排序工具

【29】子系统29：血统和依赖分析，或者叫血缘和影响性分析。

血统分析：从处理后的数据，往前追溯，查看数据的起源，中间经过哪些处理步骤。

影响性分析和血统分析相反。

在数据质量验证方面：血统和依赖分析非常重要

【30】子系统30：问题报告系统

【31】子系统31：并行管道系统。

并行方式处理大量的转换和数据。

【32】子系统32：安全系统

数据仓库保存所有数据，是最容易产生数据风险的地方。

ETL过程访问多个数据源，ETL方案也是易被攻击的地方。

【33】子系统33：合规报告系统。

确保ETL过程符合规章制度，意味对数据进行详细的审计。

审计包括：血统分析，基于时间戳的备份，审计表（数据值的修改），日志（谁访问了数据），data Vault是提供审计的数据模型。

【34】子系统34：元数据资源库管理系统。

目标是：捕获到ETL过程相关的业务，过程和技术元数据。

# 6 数据抽取