

KETTLE基础

讲师：胡老师

用数据说话 做理性决策



大数据培训专家

www.ppvke.com

胡精英

现担任某大型电信运营商大数据团队大数据架构师和大数据应用开发工程师，专注于通信行业大数据平台架构、平台演进、大数据技术落地投入生产的深度研究。

主要负责PB 级别大数据平台架构设计、平台部署、平台运维、应用开发等大数据相关工作，以及主要负责2PB 级别实时计算平台建设和600TB 级别大数据平台升级、优化。

在国内核心期刊发表论文四篇。



课程内容

PB级数据如何提取和处理？

如何做好海量数据架构设计？

本期重点：

- 1、ETL概述
- 2、ETL部署
- 3、ETL转换设计
- 4、ETL任务设计

课程目标

ETL是大数据技术系列课程的一项前置技术，通过本次分享，使大家可以感悟ETL设计的过程价值。同时，帮助大家迈出大数据的第一步：独立设计ETL任务，可以实现操作大数据的目标。

通过这次交流，你可以有两方面的收获：

- 1、了解开源ETL工具Kettle的基本概念和使用方法；
- 2、对今后数据处理方面带来一些帮助。



PB级数据如何提取和处理？

- 1、梳理数据源
- 2、根据不同数据源采用不同方法、技术进行数据提取
 - 数据库（记录数 < 1000W，记录数 > 10000W）
 - 文本
 - 互联网数据（爬虫）
 - OGG
 - FTP
 - Flume
 - Kafka
 - JDBC
- 3、处理技术
 - 分布式文件系统
 - 分布式计算框架



如何做好海量数据架构设计？

业务场景目标：

- （1）提供海量数据存储、批处理、实时查询能力，如适应数据准备、连续批处理、企业级报表、数据挖掘、海量数据即时查询等业务场景；
- （2）提供海量数据实时计算能力，如适应在线系统、实时分析、实时推荐等业务场景；（3）提供海量数据交互式能力，如适应多维报表、下钻、数据可视化、数据探索等业务场景；
- （4）提供海量数据挖掘、图计算能力，如适应数据挖掘、大规模图计算和图挖掘等业务场景；
- （5）提供海量数据快速迁移能力，如传统数据库与大数据平台之间的数据迁移、传统数据库之间数据迁移等业务场景。

技术路线：

Hadoop生态圈技术组件



ETL概述



ETL (Extract-Transform-Load)的缩写，即数据抽取、转换、装载的过程，对于IT从业人员来说，经常会遇到大数据量的处理，转换，迁移，所以了解并掌握一款ETL工具的使用，是必不可少的。

What Is ETL?

You know of course that ETL is short for extract, transform, and load; no secrets here. But what exactly do we mean by ETL? A simple definition could be "the set of processes for getting data from OLTP systems into a data warehouse." When we look at the roots of ETL it's probably a viable definition, but for modern ETL solutions it grossly over-simplifies the term. Data is not only coming from OLTP systems but from websites, flat files, e-mail databases, spreadsheets, and personal databases such as Access as well. ETL is not only used to load a single data warehouse but can have many other use cases, like loading data marts, generating spreadsheets, scoring customers using data mining models, or even loading forecasts back into OLTP systems.

核心点：

- (1) ETL是extract、transform、load的简称
- (2) 简单定义：数据从OLTP系统汇聚到数据仓库的过程
- (3) ETL可以从异构的数据源抽取数据，同时可以将数据导入到数据集市、生成电子表格、设计数据挖掘模型、甚至将预测结果返回给OLTP系统

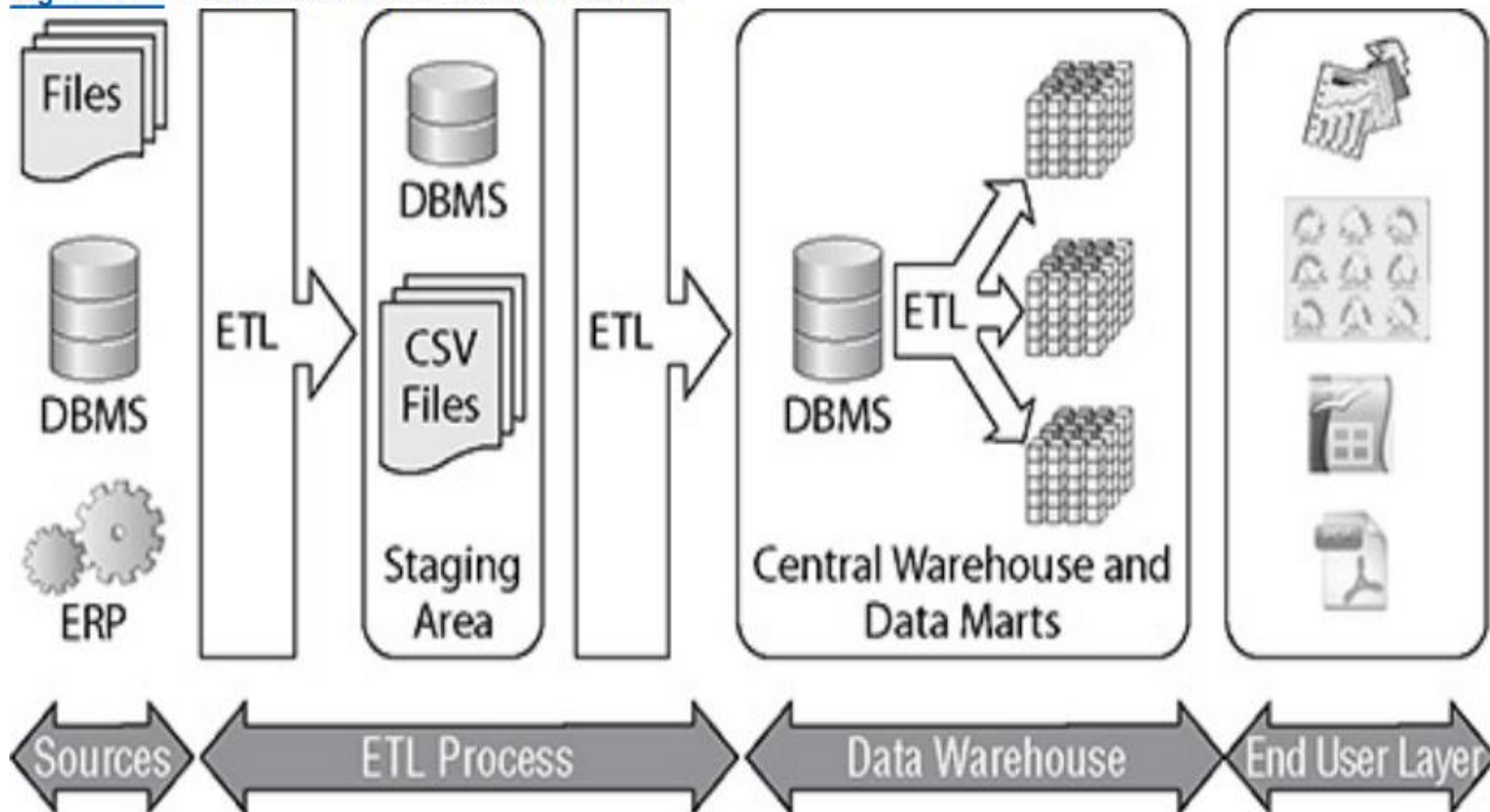


WHY WE NEED ETL ?

因为目前运行的应用系统是用用户花费了很大精力和财力构建的、不可替代的系统，尤其系统中的数据是非常之宝贵。但由于不同原始数据库中的数据的来源、格式不一样，导致了系统实施、数据整合出现问题。ETL就是用来解决这一问题的。



Figure 1-1: Classic data warehouse architecture



定义：ETL分别是“Extract”、“Transform”、“Load”三个单词的首字母缩写。也就是“抽取”、“转换”、“装载”，但我们日常往往简称其为数据抽取。

ETL包含了三方面：

“抽取”：将数据从各种原始的业务系统中读取出来，这是所有工作的前提。

“转换”：按照预先设计好的规则将抽取得数据进行转换，使本来异构的数据格式能统一起来。

“装载”：将转换完的数据按计划增量或全部导入到目标数据库中。



主流商业ETL工具介绍

Data Stage: 为开放式、可延伸的结构，其简单易见的设计工具让开发人员可以增加资料来源、目标，无须重新建立应用程序，因而减低了成本时间及资源。

Informatica: 为国外知名资料整合厂商，其最大的优点在于 re-usable，因为可重复使用设计好的 transform 不需每次重新规定，大幅缩短开发时程及人力，管理元数据的功能较同类产品比较强，为目前市面上极佳的资料转换工具。

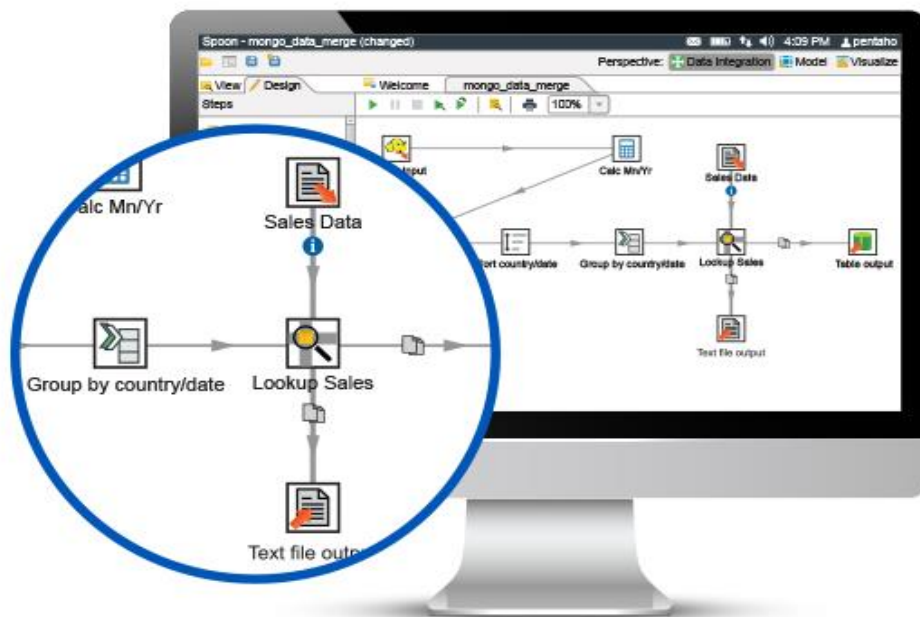
厂商	市场占有率	引擎驱动	易用性	功能
Data Stage	排名紧跟后者	为编码方式	较低	辅助功能略强
Informatica	领先几个百分点	为元数据方式	较高	主要功能一流



Pentaho Data Integration

Power to access, prepare and blend all data

Pentaho data integration prepares and blends data to create a complete picture of your business that drives actionable insights. The complete data integration platform delivers accurate, "analytics ready" data to end users from any source. With visual tools to eliminate coding and complexity, Pentaho puts big data and all data sources at the fingertips of business and IT users alike.



开源



ETL部署



Pentaho Data Integration(Kettle)介绍及部署



PDI是一款国外开源的ETL工具，通常我们习惯称它为Kettle，纯java编写，绿色无需安装，数据抽取高效稳定。



PDI中有两种脚本文件，transformation和job，transformation完成针对数据的基础转换，job则完成整个工作流的控制。



Pentaho Data Integration(Kettle)介绍及部署



下载地址

🌐 PDI下载可以到：<http://community.pentaho.com/projects/data-integration/>
取得最新版本 (**pdi-ce-7.0.0.0-25.zip**)

Data Integration - Kettle

Data Integration (or **Kettle**) delivers powerful Extraction, Transformation, and Loading (ETL) capabilities, using a groundbreaking, metadata-driven approach.



运行环境

- 🌐 OS : Windows/Linux 32/64BIT
- 🌐 安装java 环境**1.6或以上版本**



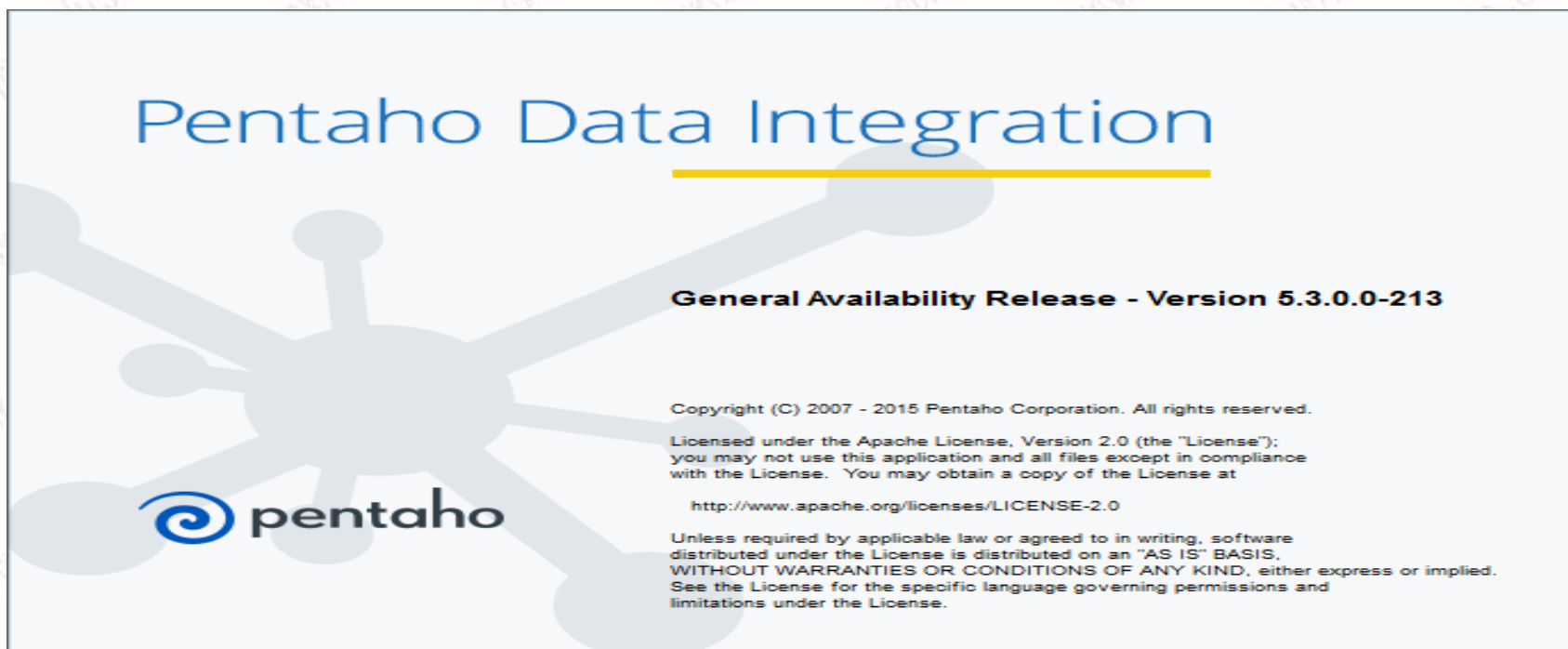
Pentaho Data Integration(Kettle)介绍及部署



运行PDI

下面是不同平台上运行PDI 所支持的脚本：

- Windows 平台运行PDI执行Spoon.bat
- Linux 平台运行PDI执行spoon.sh



Pentaho Data Integration(Kettle)介绍及部署



PDI资源库

 RDBMS数据库存储

ORACLE


MySQL

Microsoft®
SQL Server® 2008

IBM
DB2

PostgreSQL

其他。。

 XML数据文件存储



Pentaho Data Integration(Kettle)介绍及部署



PDI日志

可以设置**日志级别**，如下：

- 没有日志：不显示任务输出
- 错误日志：仅仅显示错误输出
- 最小日志：使用最小的日志
- 基本日志：缺省的日志级别**
- 详细日志：给出日志输出的细节**
- 调试日志：调试日志的输出
- 行级日志：最细粒度的日志

日志级别

- 基本日志
- 没有日志
- 错误日志
- 最小日志
- 基本日志
- 详细日志
- 调试
- 行级日志(非常详细)



Pentaho Data Integration(Kettle)介绍及部署



PDI运行方式



单机运行方式



集群运行方式



ETL转换设计





文本文件输入

Text file input : 文本文件输入, 可以支持多文件合并, 有不少参数。



表输入

Table input : 数据表输入, 实际上是视图方式输入, 因为输入的是sql语句。当然, 需要指定数据源(数据源的定制方式在后面讲一下)



Excel输入

Excel input : Excel表数据输入。



获取系统信息

Get system info : 取系统信息, 就是取一些固定的系统环境值, 如本月最后一天的时间, 本机的IP地址之类。



生成记录

Generate Rows : 生成多行。这个需要匹配使用, 主要用于生成多行的数据输入, 比如配合Add sequence可以生成一个指定序号的数据列。



转换设计——表输入

选择表输入，点击鼠标右键，选择编辑步骤。

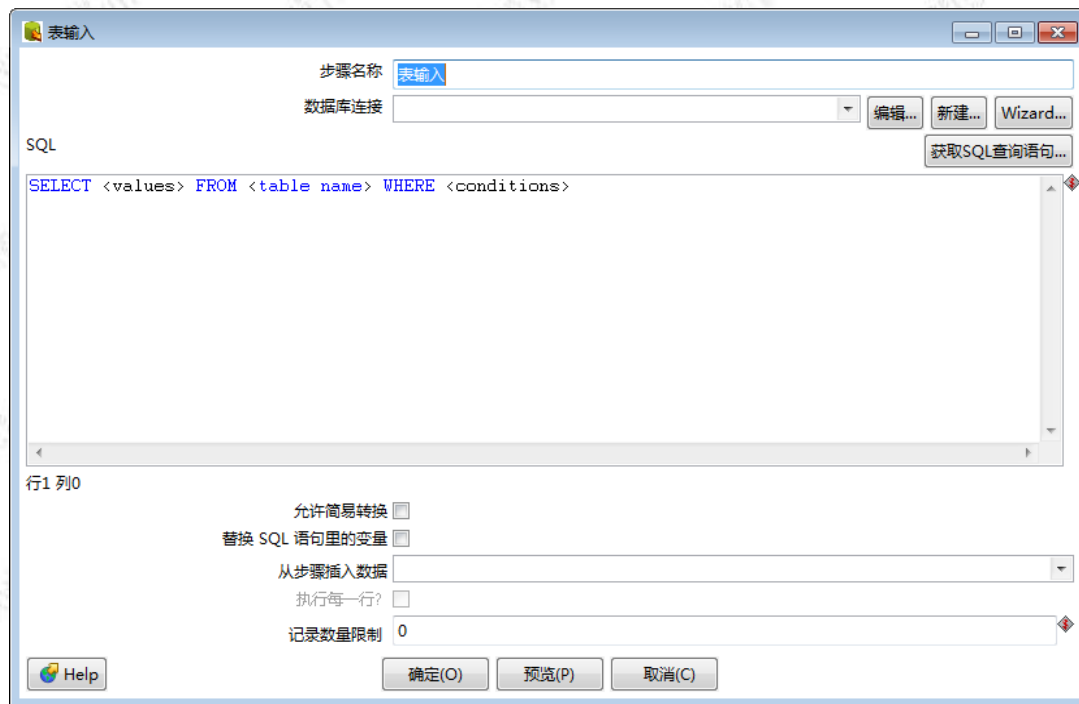
步骤名称可以更改，一般更改为和输入表相关的名称。

数据库连接：选择一个已建好的数据库连接，也可以新建一个。

点击“获取SQL查询语句”，可弹出数据库浏览器，选择自己需要的表或视图。

选择好表或视图后，SQL 区域会显示相应的SQL，如选择在SQL里包含字段名，你所选择的表的所有字段均会显示。

在SQL区域用户可手动修改SQL语句。





文本文件输出

Text file output : 文本文件输出。



表输出

Table output : 输出到目的表。



插入 / 更新

Insert/Update : 目的表和输入数据行进行比较，然后有选择的执行增加，更新操作。



XML输出

XML Output : xml文件输出。



Excel输出

Excel Output : Excel 文件输出。



转换设计——表输出

表输出

步骤名称: 表输出

数据库连接: [] [编辑...] [新建...] [Wizard...]

目标模式: [] [浏览(B)...]

目标表: [] [浏览(B)...]

提交记录数量: 1000

裁剪表: ☐

忽略插入错误: ☐

指定数据库字段: ☐

主选项 **数据库字段**

表分区数据: ☐

分区字段: []

每个月分区数据: ☒

每天分区数据: ☐

使用批量插入: ☒

表名定义在一个字段里?: ☐

包含表名的字段: []

存储表名字段: ☒

返回一个自动产生的关键字: ☐

自动产生的关键字的字段名称: []

[Help] [确定(O)] [取消(C)] [SQL]



转换设计——转换组件（1/2）



字段选择

Select values：对输入的行记录数据的字段进行更改（更改数据类型，更改字段名或删除），数据类型变更时，数据的转换有固定规则，可简单定制参数。可用来进行数据表的改装。



过滤记录

Filter rows：对输入的行记录进行指定复杂条件的过滤。用途可扩充sql语句现有的过滤功能。但现有提供逻辑功能超出标准sql的不多。



排序记录

Sort rows：对指定的列以升序或降序排序，当排序的行数超过5000时需要临时表。



增加序列

Add sequence：为数据流增加一个序列，这个配合其它Step(Generate rows, rows join)，可以生成序列表，如日期维度表(年、月、日)。



空操作（什么也不做）

Dummy：不做任何处理，主要用来作为分支节点。



Join Rows (cartesian product)

Join Rows：对所有输入流做笛卡儿乘积。



转换设计——转换组件（2/2）



分组

Group by：分组，用途可扩充sql语句现有的分组，聚合函数。



JavaScript代码

Java Script value：使用mozilla的rhino作为脚本语言，并提供了很多函数，用户可以在脚本中使用这些函数。



增加常量

Add constants：增加常量值。



去除重复记录

Unique rows：去掉输入流中的重复行，在使用该节点前要先排序，否则只能删除连续的重复行。



计算器

Calculator：提供了一组函数对列值进行运算，使用该方式比用户自定义JAVA SCRIPT脚本速度更快。



合并记录

Merge Rows：用于比较两组输入数据，一般用于更新后的数据重新导入到数据仓库中。



转换设计——过滤记录

这个步骤根据条件和比较符来过滤记录。

发送true数据给步骤：指定条件返回true的数据将发送到此步骤

发送false数据给步骤：指定条件返回false 的数据将发送到此步骤。

True 和false 步骤必须指定。



过滤记录

步骤名称: 过滤记录

发送true数据给步骤:

发送false数据给步骤:

条件:

<field> = <field>

<value>

Help 确定(O) 取消(C)



说明：转换说明

- 1、从1个数据库表中读取数据
- 2、把数据写到文本文件中



Output Stream : 一个Output Stream 是离开一个步骤时的行的堆栈。

Input Stream : 一个Input Stream 是进入一个步骤时的行的堆栈。

Hop: 一个Hop 代表两个步骤之间的一个或者多个数据流。一个Hop总是代表着一个步骤的输出流和一个步骤的输入流。

Note: 一个Note 是一个转换附加的文本注释信息。



ETL任务设计





START

START：任务流的开始。



成功

成功：任务正常执行的标志。



创建文件

创建文件：创建一个新文件。



检查表是否存在

检查表是否存在：检查数据库中表是否存在。



SQL

SQL：执行SQL语句。



Shell

Shell：执行Shell脚本。



FTP 上传

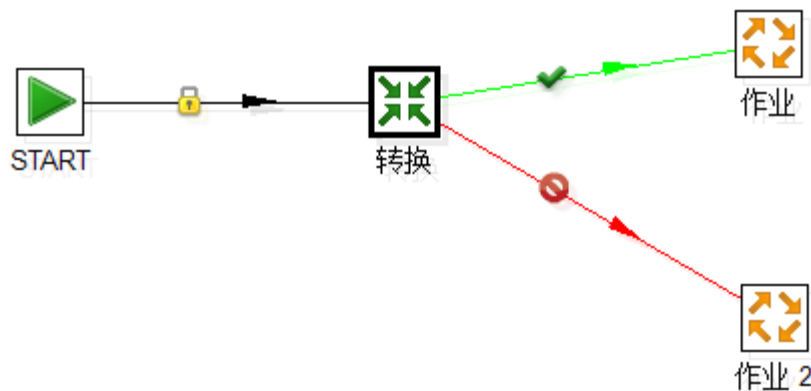
FTP上传：FTP上传数据文件。

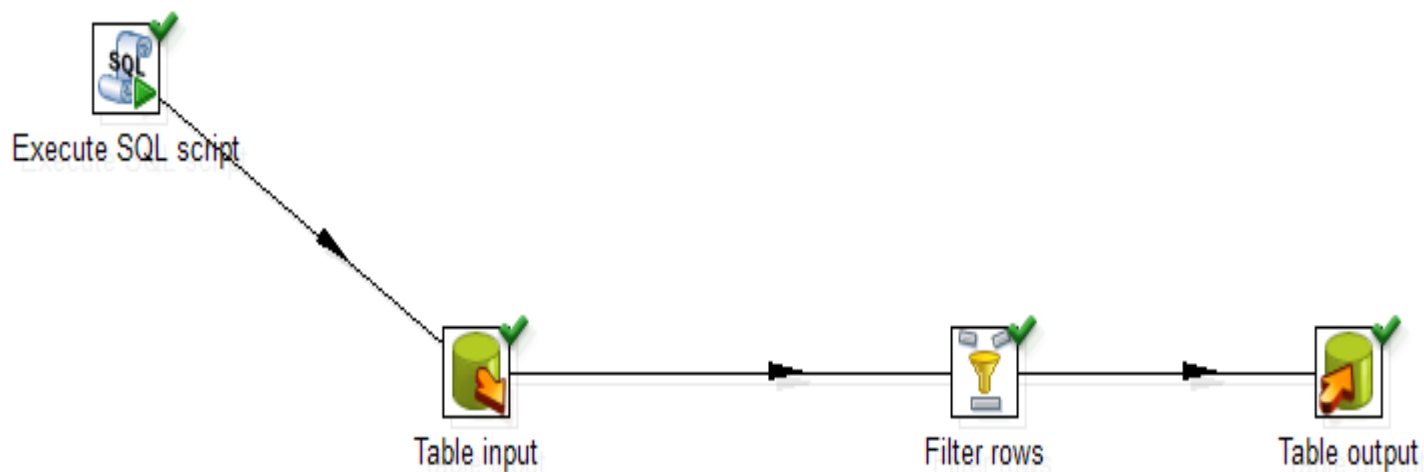


Job Entry：一个Job Entry 是一个任务的一部分，它执行某些内容。

Hop: 一个Hop 代表两个步骤之间的一个或者多个数据流。一个Hop总是代表着两个Job Entry 之间的连接，并且能够被原始的Job Entry设置，无条件的执行下一个JobEntry,直到执行成功或者失败。

Note: 一个Note 是一个任务附加的文本注释信息。





执行SQL语句

步骤名称: Execute SQL script

数据库连接: Sampledata [编辑...] [新建...] [Wizard...]

SQL script to execute. (statements separated by ;) Question marks will be replaced by arguments.

```
truncate table N;  
truncate table F;  
truncate table E;  
truncate table D;
```

行3 列17

☐ 执行每一行?
☐ Execute as a single statement
☐ 变量替换
☐ Bind parameters?
☐ Quote Strings?

参数:

#	作为参数的字段
1	

包含插入状态的字段:
包含更新状态的字段:
包含删除状态的字段:
包含读状态的字段:

[Help] [确定(O)] [取消(C)] [获取字段]



表输入

步骤名称

数据库连接

SQL

```
SELECT
CUSTOMERNUMBER
/ CUSTOMERNAME
/ CONTACTLASTNAME
/ CONTACTFIRSTNAME
/ PHONE
/ ADDRESSLINE1
/ ADDRESSLINE2
/ CITY
/ STATE
/ POSTALCODE
/ COUNTRY
/ SALESREPEMPOYEEENUNBER
/ CREDITLIMIT
FROM CUSTOMERS_TABLE_OUT
```

行1 列0

允许简易转换 ☐


替换 SQL 语句里的变量 ☐

从步骤插入数据

执行每一行? ☐

记录数量限制



 过滤记录


步骤名称: Filter rows


发送true数据给步骤:

发送false数据给步骤:

条件:

CUSTOMERNAME IS NOT NULL



 Help

确定(O) 取消(C)



表输出

步骤名称: Table output

数据库连接: Sampledata [编辑...] [新建...] [Wizard...]

目标模式: [浏览(B)...]

目标表: CUSTOMERS_TABLE_OUT [浏览(B)...]

提交记录数量: 100

裁剪表: ☐

忽略插入错误: ☐

指定数据库字段: ☐

主选项 | **数据库字段**

表分区数据: ☐

分区字段: []

每个月分区数据: ☒

每天分区数据: ☐

使用批量插入: ☒

表名定义在一个字段里?: ☒

包含表名的字段: CUSTOMERNAME

存储表名字段: ☐

返回一个自动产生的关键字: ☐

自动产生的关键字的字段名称: []

[Help] [确定(O)] [取消(C)] [SQL]



转换属性

转换 命名参数 日志 日期 依赖 杂项 监控

转换名称: test new TableOutput with table name in field

转换文件: E:\迅雷下载\pdi-ce-5.1.0.0\data-integration\samples\transformations\Table Output - Tablename in field.ktr

描述:

扩展描述:

状态:

版本:

目录: /

创建者:

创建日期: Wed Feb 26 16:25:18 CST 2014

最近修改的用户:

最近修改日期: Wed Feb 26 16:25:18 CST 2014

确定(O) SQL 取消(C)





ETL开发需求描述：

将website.txt中**百度**的网址放到baidu.txt，**新浪**的网址放到sina.txt

website.txt内容如下（字段间以“|”分隔）

NAME	URL
百度	http://image.baidu.com/
新浪	http://www.sina.com.cn/
新浪	http://sports.sina.com.cn/nba/
QQ	http://www.qq.com/
百度	https://www.baidu.com/





大数据培训专家

讲师：胡老师

咨询电话：4000-707-620

学习QQ群：87353699



用数据说话 做理性决策

The End!

