ETL技术入门之ETL初认识 - xiaohai798的专栏 - CSDN博客

ETL是什么

ETL是Extract Transform Load三个英文单词的缩写 中文意思就是抽取、转换、加载。说到ETL就必须提到数据仓库。

先说下背景知识：

信息是现代企业的重要资源，是企业运用科学管理、决策分析的基础。目前，大多数企业花费大量的资金和时间来构建联机事务处理OLTP的业务系统和办公自动化系统（例如电信行业的各种运营支撑系统、购物网站系统），用来记录事务处理的各种相关数据。据统计，数据量每2～3年时间就会成倍增长，这些数据蕴含着巨大的商业价值，而企业所关注的通常只占在总数据量的2%～4%左右。因此，企业仍然没有最大化地利用已存在的数据资源，以致于浪费了更多的时间和资金，也失去制定关键商业决策的最佳契机。

在这个背景下，能够给企业所有级别的决策制定过程提供支持的所有类型数据的战略集合应运而生，他就是数据仓库。数据仓库的英文简写是Data Warehouse。数据仓库就是把OLTP系统产生的数据 整合到一起 发掘其中的商业价值和提供决策支持用。举个电信行业的例子 电信有系统每天会有客户投诉的信息、宽带群体性障碍、客户号码的停机恢复时间记录等等。这些数据都在各自的生产环境系统里面。他们每个月会把这些数据整合到一起处理加工到数据仓库里面形成报表 其中有一个功能是可以对哪些用户有离网销号的倾向做出大概的判断。这就是数据仓库的价值所在。

那么怎么把数据弄到数据仓库里去呢，其中用到的一个技术就是ETL。

下面给下ETL的详细解释定义：

ETL(Extract-Transform-Load的缩写，即数据抽取、转换、装载的过程)作为DW的核心和灵魂，能够按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。如果说数据仓库的模型设计是一座大厦的设计蓝图，数据是砖瓦的话，那么ETL就是建设大厦的过程。在整个项目中最难部分是用户需求分析和模型设计，而ETL规则设计和实施则是工作量最大的，约占整个项目的60%～80%，这是国内外从众多实践中得到的普遍共识。

ETL是数据抽取（Extract）、清洗（Cleaning）、转换（Transform）、装载（Load）的过程。是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗,最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。

于是，企业如何通过各种技术手段，并把数据转换为信息、知识，已经成了提高其核心竞争力的主要瓶颈。而ETL则是主要的一个技术手段。

做数据仓库系统，ETL是关键的一环。说大了，ETL是数据整合解决方案，说小了，就是倒数据的工具。

现在来说说ETL技术用到的工具，常用的有Informatica、Datastage、Beeload、Kettle等。目前只用过kettle，所以这里只对kettle做描述。

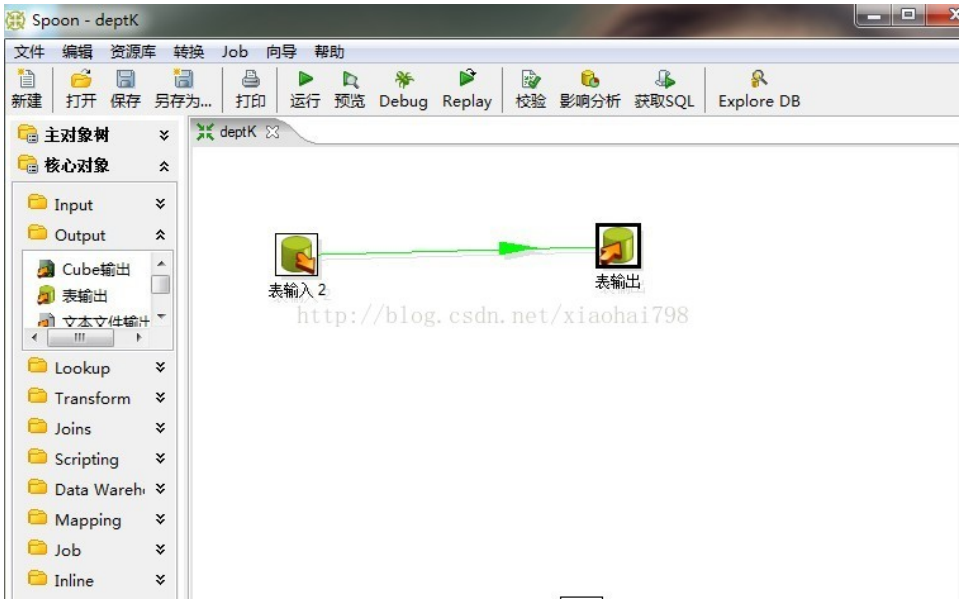
kettle是一款国外开源的ETL工具，纯java编写，可以在Window、Linux、Unix上运行，kettle 3版本需要安装 3以上都是绿色版无需安装。

提醒的是kettle运行 需要机器有JRE环境

Kettle这个ETL工具集，它允许你管理来自不同数据库的数据，通过提供一个图形化的用户环境来描述你想做什么。

Kettle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。

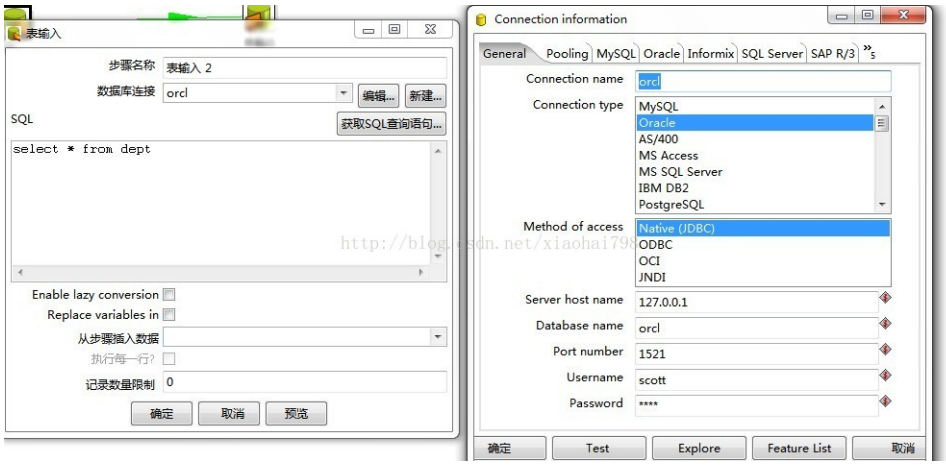
现在来看下kettle的transformation文件，一个最简单的E过程例子（windows环境）



上图文件的功能就是就是把oracle数据库一个表的数据抽取到另一表里面。

左边的图标叫表输入 右边的叫表输出 中间绿色的线代表数据流向。 表输入和表输出可由左边的菜单栏里 鼠标拖动出。

双击打开表输入是这样的：



上图左边的是打开表输入的界面

步骤名称： 即是图标下面显示的名字 可以随便填

获取sql查询语句：点击后 会树状形式展示oracle的表视图 等 选中双击后 点自己会添加到空白的sql框内。

数据库连接： 一开始没有 需要新建 有了就可以编辑了 点击编辑后会弹出上图右边的页面

connection name ： 连接起个名字 可以使数据ip地址 加实例名

connection type ： 是选择你要查询的数据库类型 mysql oracle等等

method access：是选择驱动类型 选择那个JDBC就可以了

server host name ： 是数据库的ip地址

dbname ： 是数据库实例名

Port number：是端口号

再下面就是用户名 密码了。

填好所有的信息后 可以点击test测试下能不能成功连接。成功连接即可点击确定 界面就会回到上面的左图

sql 下面的空框 是用来写你想要获取数据的sql语句(也可以由那个获取查询sql按钮自动获取) 写好后 可以点击预览（行数选少点）看下数据可正常。如果能预览数据 说明你的表输入就配好了。

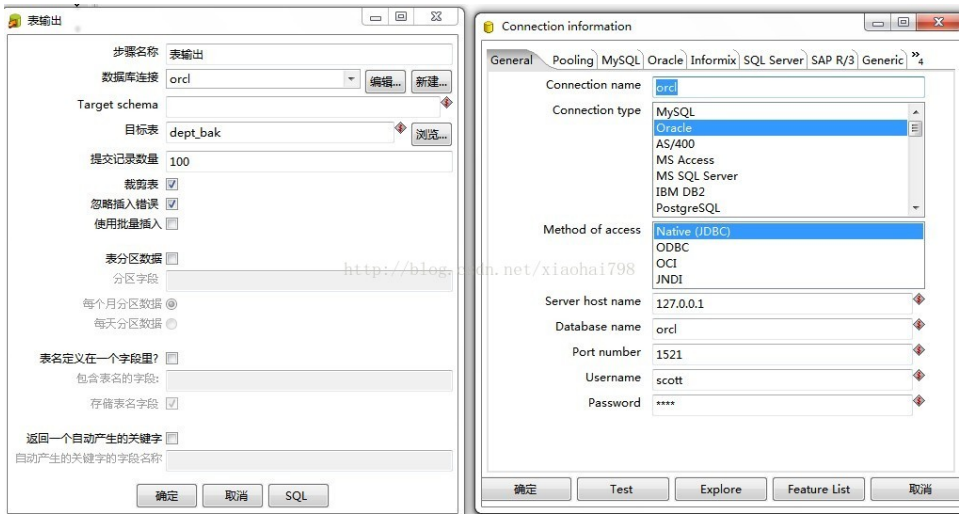
上图左边

ennable lazy conversion 勾上后就表示延迟转换：这样在从数据库读取的数据就能保持原有字符集，不会默认强制使用utf8。

replace variables in ：表示如果sql框里的sql有变量的时候 会使用环境变量替代它 像table_201407 这样带日期的固定格式表名 可以使用这个实现自动化

记录数量限制：默认为0 若设为大于0的任何值 则无论sql怎么写的 输入表只有设置的行数那么多。

现在来看表输出：



上图左边为双击打开表输出的 界面。

步骤名称：表输出图标下面显示的名字 可用数据库ip 用户名 表名

数据库连接 ：没有需要新建 可以新建几个 新建好的可以编辑

target schema ：目标表或者视图的用户

目标表：可以自己输入，也可以从浏览里面选择

提交记录数量：批量一次提交的数据量或者非批量插入数据量的限制值

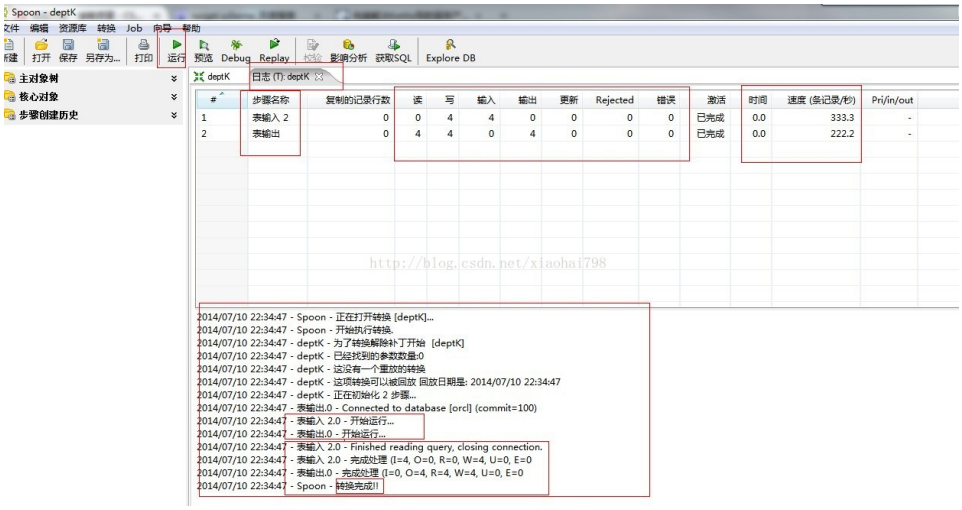
裁剪表：插入之前 有 truncate 操作。

忽略插入错误 ：这是非批量插入的功能，非批量插入时 若有一天数据插入报错 后面的数据还可以正常插入。

使用批量插入：点上即是批量 否则为非批量。

点击数据库连接的编辑后 会弹出上图右图 与表输入的一样 填写数据库的tns信息 及用户名密码。点击test可以连接后 点击确定

回到上图左边界面 再点击确定 即配好了一个转换 点击运行后：



可以从日志看 该转换有没有finished、 每个步骤的耗时、速度 、平均每秒多少行、 总共插入了多少记录数等。