

kettle入门(五) 之kettle抽取gz格式文本详细案例 - xiaohai798的专栏 - CSDN 博客

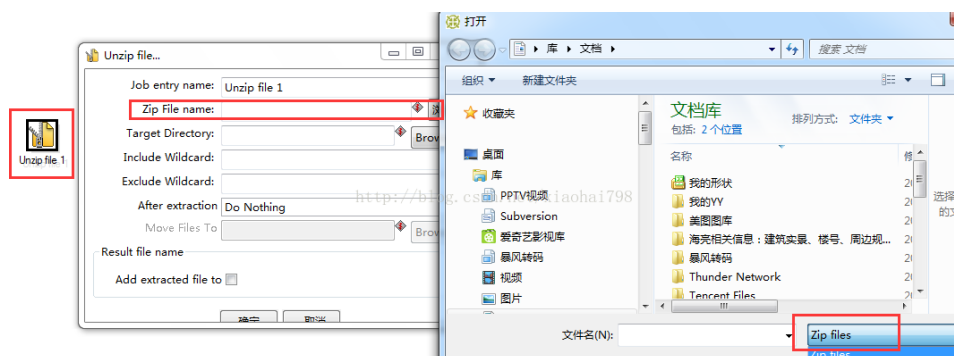
背景:

ods平台的一个很简单的数据共享需求:

运营商的某个部门每天定时送gz格式的HLR文本数据到FTP服务器的固定目录下。然后ods每天定时去取然后录入到RDBMS的表中, 开放给其他系统查询调用, 这种称作数据库表接口。

需求很简单, 但是因为以前只用过文本输入做txt 或者csv、excel , 所以一时就想怎么先把gz格式解压出来, 再用文本文件输入, 首先想到了用 kettle3自带的unzip 功能

如下图:



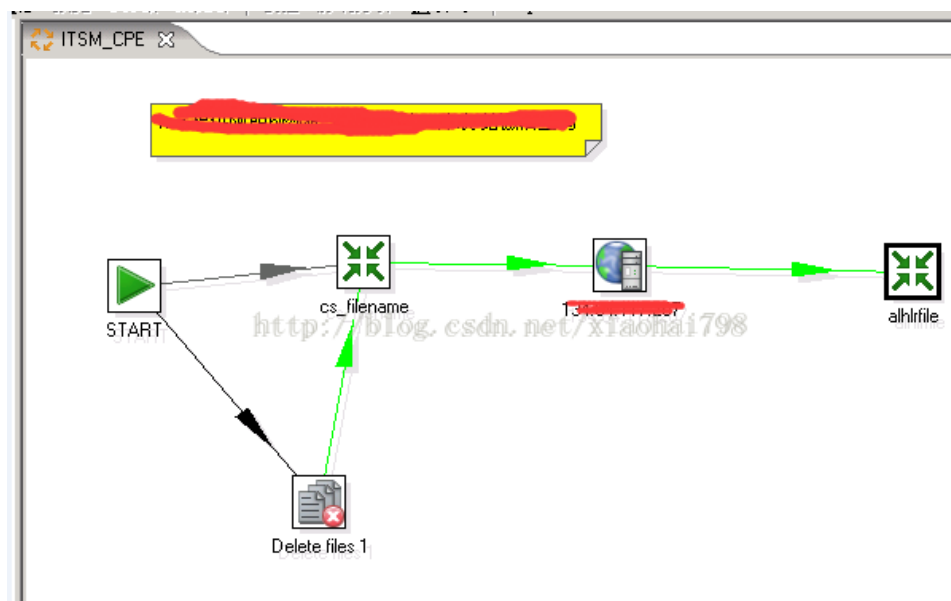
结果发现 解压不了gz格式的, 后来又想到使用shell命令, 但是在后台putty可以 但是用当前系统etl平台封装执行shell命令后就有问题。搞了较长时间问题没搞定。

后来偶然发现 文本文件输入 本身就可以直接读gz格式文本, 感觉前面都浪费了时间

解决:

下图所示的kjb 就是实现该功能的job , 步骤如下:

- 1 START
- 2 删除本地服务器的历史文件 (Delete files) 防止历史文件占用机器存储
- 3 设置时间变量(cs_filename)
- 4 使用时间变量正则匹配下载FTP文本到本地(带有地球的那个图)
- 5 从本地匹配需要的文本录入RDBMS (alhlrfile)

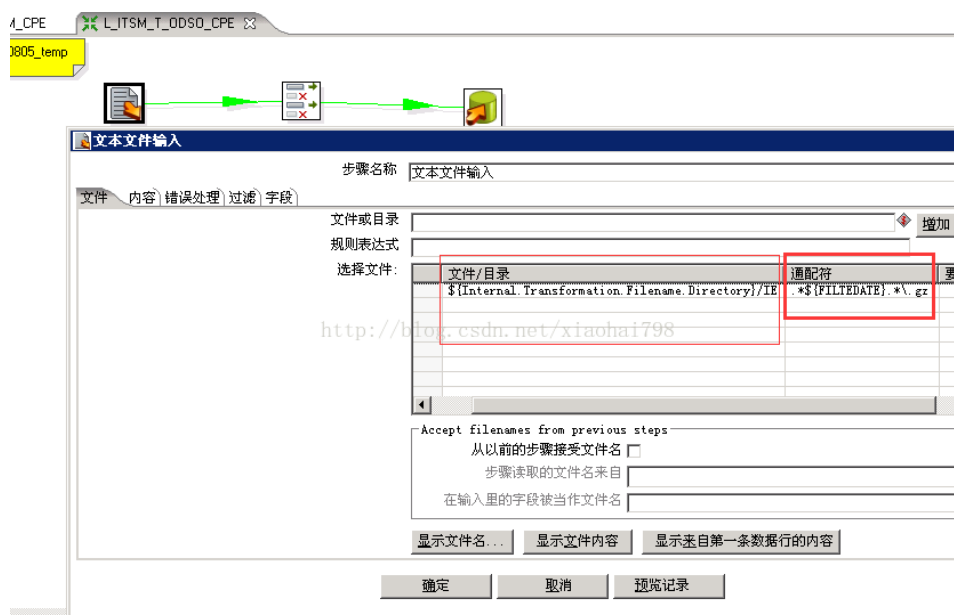


下图实现的功能是 从本地匹配需要的文本录入RDBMS，画红圈的目录部分指当前脚本文件ktr所在的目录下的IE目录，通配符用来指定读取匹配的文本。

ex:每天读取前一天的文本 则 时间变量设置为 系统时间的前一天 2015-05-16 就是2015-05-15，文本就是2015-05-15.txt.gz

关于时间变量的设置 参考：

[kettle入门\(四\) 之kettle中设置任意时间变量的详细案例](http://blog.csdn.net/xiaohai798)



下图文本文件输入 就是 直接把gz格式的文本录到RDBMS的图元。

如下图：

文件类型：选择上csv即可（与上有系统 协商是文本）

分隔符、封闭字符：按照实际的文本内容来

头部数量：指文本头部 非正式数据的内容（如 字段属性名等）的行数

compression: 默认none 应选择GZIP，若有中文 格式选择UNIX，编码方式选择 UTF-8，若没有则默认

文本文件输入

步骤名称 文本文件输入

文件 内容 错误处理 过滤 字段

文件类型 CSV

分隔符 , 插入TAB

封闭字符

在被封闭的字段里运行? ☐

逃逸字符

头部 ☒ 头部行数 1

尾部 ☐ 尾部行数 1

包装行? ☐ 以时间包装的行数 1

分页布局 (printout)? ☐ 每页记录行数 80

文档头部行 0

Compression GZip

没有空行 None

在输出包括字段名? Zip

输出包含行数? ☐ 行数字段名称

Rownum by file? ☐

格式 Unix

编码方式 UTF-8

记录数量限制 0

解析日期时候是否严格要求? ☒

本地日期格式 zh_cn

确定 取消 预览记录

选好之后，可以点击预览记录 看看设置是否正确，数据是否正常。

下图是文本数据全量录入到oracle数据库的示例：

