

**Kettle开发规范**

|  |  |
| --- | --- |
| **撰写** | **蒋晓宇、王立鑫** |
| **审核** |  |
| **版本号** | **V0.1** |

上海烟机数字技术研发中心

2021年11月24日

**目录**

[1 文档介绍 2](#_Toc88721255)

[1.1 文档目标 2](#_Toc88721256)

[1.2 术语与缩写解释 2](#_Toc88721257)

[2 ETL命名规范 2](#_Toc88721258)

[2.1 作业的命名 2](#_Toc88721259)

[2.1.1 表输出型的作业 2](#_Toc88721260)

[2.1.2 任务型的作业 3](#_Toc88721261)

[2.2 转换的命名 3](#_Toc88721262)

[2.2.1 隶属于作业脚本的转换 3](#_Toc88721263)

[2.2.2 独立于作业脚本的任务型转换 4](#_Toc88721264)

[2.3 转化的步骤的命名（日后还继续补充） 5](#_Toc88721265)

[2.4 注释 5](#_Toc88721266)

[2.5 变量的命名 6](#_Toc88721267)

[2.6 参数 7](#_Toc88721268)

[3 ETL开发原则 7](#_Toc88721269)

[3.1 版本控制原则 7](#_Toc88721270)

[3.2 异常处理原则 7](#_Toc88721271)

[3.3 日志原则 8](#_Toc88721272)

# 文档介绍

## 文档目标

本文档主要描述了KETTLE平台ETL开发过程中应遵循的原则和规定，为ETL开发人员提供规范和指导。

## 术语与缩写解释

ETL：数据的抽取（Extract），转换（Transform）和加载（Loading）

ODS：数据源层：ODS（Operational Data Store）ODS 层，是最接近数据源的一层。建立ODS层可方便在日后根据需要追溯数据。因此，对于这一层不建议做过多的数据清洗工作，而是原封不动地接入原始数据即可。名称中带有ODS前缀的库代表存放被抽取出的原始数据的专用库。

# ETL命名规范

## 作业的命名

### 表输出型的作业

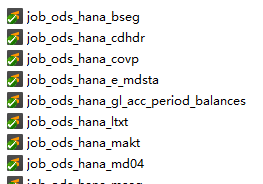
表输出型的作业应遵循的命名规则：作业\_目标库名\_目标表名。

例如：在名为job\_ods\_hana\_bseg的表输出型作业中，

job为作业标识；

ods\_hana为目标库名称；

bseg为目标表名。



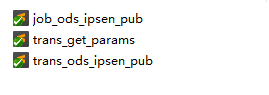
### 任务型的作业

任务型的作业应遵循的命名规则：作业\_任务描述

例如：在名为job\_ods\_ipsen\_pub的任务型作业中，

ods\_ipsen代表这是一个给ods\_ipsen库使用的脚本；

pub代表了这是一个公用抽取脚本。



## 转换的命名

### 隶属于作业脚本的转换

隶属于作业脚本的转换应遵循的命名规则：转换\_目标库名\_目标表名\_转换编号。

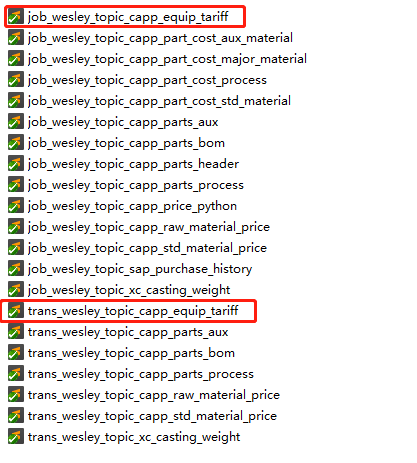
例如：在名为trans\_wesley\_topic\_capp\_equip\_tariff的转换中，

wesley\_topic为目标库名；

capp\_equip\_tariff为目标表名。

此转换名称与对应的作业文件job\_ wesley\_topic\_capp\_equip\_tariff保持一致。

**注意：**当转换为所属作业的唯一转换时，转换编号（01，02，03…）可以省略。



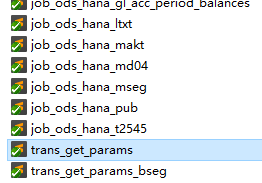
**Figure 1隶属于作业脚本的转换的命名**

### 独立于作业脚本的任务型转换

独立于作业脚本的任务型转换应遵循的命名规则：转换\_任务描述。

例如：在trans\_get\_params的任务型转换中，

get\_params代表这是一个用于获取参数的公共转换文件。



**Figure 2独立于作业脚本的任务型转换的命名**

## 转化的步骤的命名（日后待继续补充）

转换中各个步骤的名称一律使用中文、英文，或者两者的组合形式。

以下为常见的步骤的命名规范：

1. 表输入：要写清来源表的系统名、库名、表明。如有多张来源表需全部写明。
2. 表输出：需要写清输出表的系统名、库名、表明。
3. 维度查询/更新：要写清楚涉及到的更新的字段
4. 字段选择：要写清设计到的字段名称，一般不超过三个。数量特别的多的情况建议使用多个字段选择
5. 数据库查询：要写清所查询的数据库名称，查询字段以及返回字段
6. 流查询：要写清查询字段和接受字段
7. 过滤记录：要写清过滤目的、过滤字段及条件
8. 排序记录：要写清排序的字段
9. 公式、计算：要写清楚公式的作用
10. 记录集连接：要写清连接的字段名称

图示

中度可信度描述已自动生成

**Figure 3转化的步骤的命名**

## 注释

需要为每个转换写该转换对应的mapping文件和开发的版本历史时，使用【注释】功能。

文本

描述已自动生成

**Figure 4转换的注释**

## 变量的命名

变量的命名规则使用ETL\_RC，如下所示：

**注意：**变量不能通过外部命令传递，如果要外部传递进变量，需要使用参数。

图形用户界面, 表格

描述已自动生成

**Figure 5变量的命名**

## 参数

参数分为命名参数和位置参数，进行参数传递时一般建议使用命名参数。设置完参数，ETL作业（转换）就可以使用该参数，并对其动态赋值。参数设置如下：

图形用户界面, 应用程序, 表格, Excel

描述已自动生成

# ETL开发原则

## 版本控制原则

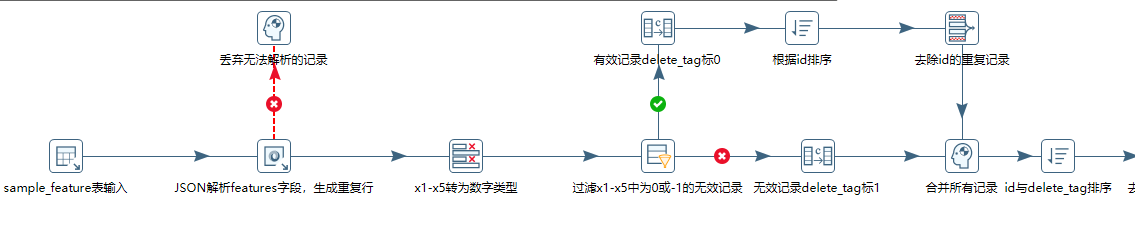
ETL开发的版本应与mapping的版本保持一致。此外，上线的版本和MAPPING的版本同时在SVN里保留备份。

测试，正式环境使用资源库

为了对元数据进行管理，必须使用资源库。确定测试版本和上线版本，然后保存在资源库里进行测试和上线。

## 异常处理原则

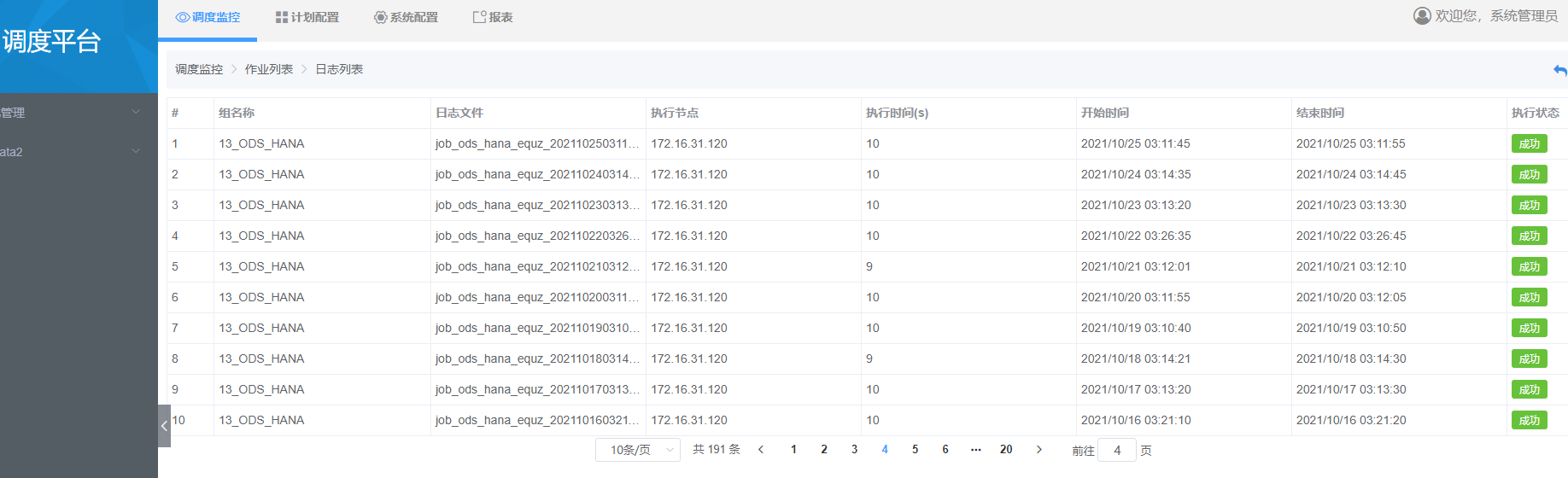
异常处理可以有效捕捉脚本中存在的逻辑错误，解决数据质量问题，并有助于处理异常和脚本的迭代开发。因此，应对需要的转换组件进行错误处理，避免数据丢失及脚本报错阻塞。



## **日志原则**

获取日志的常用方式有两种：

1. 将运行的输出作为日志。这样做的优点时直观地展示了转换的运行情况，可在调度平台中实现。呈现信息如下：



1. 可以将每个转换的运行日志配置到数据库。但是不建议这么做，因为不够直观。

图形用户界面, 应用程序

描述已自动生成

