# 概述

ETL作业定义工具用于帮助ETL开发人员开发数据仓库基础层实体加载脚本，它的主要组成部分如下图所示：

ETL元数据资料库

ETL脚本

ETL脚本生成器

源系统元数据和数据仓库模型

Excel作业定义工作表

用户使用ETL作业定义工具进行ETL开发的主要工作流程为：

1. 将前期收集/设计的源系统元数据和数据仓库模型数据导入ETL元数据资料库，需要导入的数据的详细要求参见【[2.5 数据结构](#_数据结构)】部分。
2. 使用Excel作业定义工作表对数据仓库实体表进行字段级映射定义，并保存到ETL元数据资料库中。
3. 调用ETL脚本生成器生成数据仓库实体的加载脚本。

# ETL作业定义工作表

## 系统参数页

在“系统参数”页进行一些必要的参数设置。其中需要设置的参数包括：

* 开发人员名称：参与ETL映射开发的人员名称，需要与ETL元数据资料库中的 etl\_tasks.etl\_dvlpr\_name字段相对应。admin表示管理员用户，管理员用户可以为未分配开发人员的ETL任务指定开发人员。
* 数据源名称：在运行本ETL模板的机器上配置的ODBC数据源的名称，该数据源指向ETL元数据资料库。

## ETL作业列表页

本页显示数据仓库基础层ETL加载作业列表，仓库基础层每一个实体对应一个加载作业。用户点击“刷新”按钮显示当前的数据仓库基础层实体所对应的ETL加载作业，标题为蓝色的列允许用户修改。用户修改完毕需要点击“保存”按钮将数据保存到ETL元数据资料库中。用户可以点击“创建ETL作业表”按钮为列表中每一个ETL作业创建一个Excel Worksheet，用于编辑该作业的映射。新创建的Worksheet按照ETL作业列表中的序号字段进行命名，命名规则为“ETL\_作业序号”。如果同名的Worksheet已经存在将不会重复创建新的Worksheet。

## ETL作业模板页

本页是用于创建ETL作业Worksheet的模板，用户不应该修改本页的任何内容。

## ETL作业定义页

命名为“ETL\_作业序号”的页面用于编辑具体ETL加载作业的映射参数。本页可以分为以下几个部分：

### 目标表ETL属性

包含本作业需要加载的目标仓库实体的和ETL有关的属性，用户可以在这里指定ETL作业属性，包括加载模式、积数计算、删除模式、保留首次加载日期、备注。其中积数计算、删除模式、保留首次加载日期尚未实现。用户修改ETL作业属性之后，需要点击本节的“保存”按钮将修改后的属性保存到ETL元数据资料库。

#### 加载模式

支持的模式包括：

* 追加——适用于全量抽取。当前加载的数据将作为一个新的分区追加到目标表中，分区字段为data\_dt，表示数据加载的工作日。
* 更新——适用于增量抽取的情况。当前加载的数据将与目标表中原有数据进行主键FullOuterJoin产生当天的全量数据，关联结果中如果新加载记录不为空，就保留新加载记录，否则使用目标表原有记录。新的全量数据被加载到以当前工作日标记的新分区中（分区字段为data\_dt）。

#### 积数计算

未实现。

#### 删除模式

未实现。

#### 保留首次加载日期

未实现。

#### 分批加载

表示本数据仓库实体表的各加载批次是否需要单独执行。填1表示各批次需要单独执行，这种情况下，ETL脚本生成器将会为该实体表的每个加载批次生成一个单独的SQL脚本；填0表示批次不需要单独加载，此时ETL脚本生成器会为该实体表的所有加载批次生成一个SQL脚本。

### 源表列表

用户在源表列表中填写当前数据仓库实体表的源表信息，包括以下内容：

#### 加载批次

如果数据仓库实体表的数据来自多个相互独立的源主表，则需要为这些源主表指定不同的加载批次。加载批次从1开始递增。加载过程中需要用到的代码表也需要列入源表中，其加载批次与同其关联的主表批次相同。如果代码表同时被多个加载批次使用，可以将代码表的加载批次设置为0，表示被多个加载批次共同使用的公共代码表。

#### 源表名

源表的名称，此名称应该与ETL元数据资料库中src\_table\_analysis和src\_column\_analysis中的table\_name字段保持一致。

#### 源数据库名

源表所在数据库（或schema）的名称。此名称应该与ETL元数据资料库中src\_table\_analysis和src\_column\_analysis中的schema\_name字段保持一致。

#### 源系统

源表所来自的源系统的名称。此名称应该与ETL元数据资料库中src\_table\_analysis和src\_column\_analysis中的sys\_name字段保持一致。

#### 源表描述

源表的说明。来自ETL源数据资料库中src\_table\_analysis的description属性。用户不需要填写这个字段。

#### 源表别名

源表别名可以被用在关联条件、过滤条件或字段映射表达式中，作为当前源表的引用，主要用于区分同一批次不同源表中的同名字段。

#### 接口名

接口名是源表的物理名称，在生成的加载脚本中，使用接口名访问源表。接口名应定义为“数据库名.表名”的形式。

#### 连接次序和连接类型

如果对于一个加载批次，需要其中各源表进行关联得到最终加载结果，那么需要指定这些源表的关联次序和关联条件。其中排第一位的源表不用指定关联次序和关联条件，之后的每一个源表都需要指定关联次序（从1开始递增），以及与前一个表的关联类型，可以指定的关联类型包括：

Inner Join

Left Outer Join

Right Outer Join

Full Outer Join

#### 连接条件

当前源表与其他源表关联使用的关联条件。关联条件中可以使用为源表指定的别名来区分不同源表的同名字段。

#### 筛选条件

表示用于对当前源表进行筛选的条件。源表的筛选发生在源表的关联之前。

#### 增量抽取

表示当前源表是否采用增量抽取。1表示增量，0表示全量。一般来说，如果源表采用增量方式抽取，数据仓库实体应采用更新方式加载，否则应采用追加方式加载。

用户填写源表属性之后，应该点击“保存”按钮将修改后的数据存入ETL元数据资料库。

### 目标表ETL规则

用户在填写本部分内容之前，应确保源表列表已经填写并保存到ETL元数据资料库中。

填写目标表ETL规则就是为当前仓库实体指定加载各批次（及组别）时各字段应被赋予的值或表达式。在开始填写之前，应该先点击本节的“刷新”按钮，将原有的ETL规则数据从ETL元数据资料库中取回，如果库中没有当前实体的ETL规则数据，程序会根据当前提供的源表列表中的批次，自动为每个批次生成一套空白的ETL映射规则。

以下说明ETL规则中各字段的含义：

#### 字段编号

字段编号是数据仓库实体属性的唯一编号。字段编号不需要用户填写，如果用户需要自己添加新的规则条目（如为新的加载组别提供规则），则需要确保新规则记录中的字段编号和字段名称的对应关系与其他规则记录保持一致。

#### 字段名称

字段名称是数据仓库实体属性的字段名称。这个字段不需要用户填写。如果用户需要自己添加新的规则条目（如为新的加载组别提供规则），则需要确保新规则记录中的字段编号和字段名称的对应关系与其他规则记录保持一致。

#### 字段中文名

字段名称是数据仓库实体的属性名称，一般为中文。这个字段不需要用户填写。如果用户需要自己添加新的规则条目（如为新的加载组别提供规则），则需要确保新规则记录中的字段编号和字段名称的对应关系与其他规则记录保持一致。

#### 数据类型

数据仓库实体属性的数据类型。不需要用户填写，程序自动根据dw\_table.data\_type字段的值填写。

#### 主键

标识当前属性是否是实体的主键字段。1表示主键，0表示非主键。这个字段不需要用户填写，系统自动根据dw\_table.is\_pk字段的值填写。

#### 分区键

标识当前属性是否是实体的分区键字段。0表示非分区键，>0表示分区键的次序（data\_dt为固定第一分区键，1表示第二分区键，2表示第三分区键，依次类推）。这个字段不需要用户填写，系统自动根据dw\_table.is\_partition\_key字段的值填写。

注意在目前的设计中，要求所有数据仓库基础层实体都具备第一分区键data\_dt，表示数据加载日期。实体的其他分区键字段（如果有）都依次列在data\_dt之后，构成完整的分区键。data\_dt字段不需要列入数据仓库模型的实体属性列表。如果实体表需要分别执行各批次加载，则实体表中必须定义代表加载批次（源系统号）的分区字段，且此分区字段的分区次序必须为1，即作为data\_dt之后的第二分区字段。

#### 积数周期

目前未实现，用户不需要填写。

#### 参与拉链比较

标识当前属性是否属于拉链比较字段。1表示参与拉链比较，0表示不参与。这个字段不需要用户填写，系统自动根据dw\_column.chain\_compare字段的值填写此字段。

#### 加载批次及组别

如果数据仓库实体的数据来自多个源主表，这些源主表需要通过多个批次分别加载，每个批次加载一个源主表。如果在一个批次内，源主表的一条记录加载到数据仓库实体后，形成多条记录，就需要在这个批次内定义多个加载组别。加载批次和组别表示为“批次号..组别号”的形式，组别号默认为0。

假设需要加载人口信息，主键为人口ID，源表记录中有人员ID和配偶ID两个字段（假设都不为空，且ID都不相同），人员ID和配偶ID所对应的人都需要被加载到人口信息，因此源表中每条记录在目标表中产生两条人口记录，ID分别为源数据中的人员ID和配偶ID，这时可以通过定义两个组别来分别将人员ID和配偶ID映射到目标表人口ID，来实现一条源表记录产生两条目标表记录。当有多个组别时，不必为所有字段定义所有组别的映射，如果某个字段组别0和组别1的映射相同，则组别1的映射可以省略，系统在找不到该字段组别1的映射时，会自动使用这个字段组别0的映射来代替。

#### 源表名

表示当前映射对应的源表表名，需要以“源系统名.源数据库名.源表名”的形式填写，如果对应多个源表的多个字段，源表名称之间用逗号分隔。源系统名、源数据库名和源表名应取自src\_table\_analysis表的sys\_name, schema\_name和table\_name字段。

如果当前映射的取值为常量，则源表名可以不填。

#### 源表字段

表示当前映射对应的源表字段名，如果对应多个源表的字段，源表字段名称之间用逗号分隔。源表字段名称应取自src\_column\_analysis.column\_name字段的值。如果当前映射的取值为常量，则源表字段可以不填。

#### 源表字段类型

用户不需要填写，系统根据src\_column\_analysis.data\_type字段的值自动填写。

#### 计算表达式

如果数据仓库实体属性的值直接取自某一个源表字段的值，则只需要在映射中正确填写源表名和源表字段即可，否则，目标字段的值由计算表达式确定。用户填写的计算表达式将被用在加载语句的select子句中，用来计算数据仓库实体属性的取值。

由于Excel的特殊规定，单元格内容的开头需要用两个单引号表示一个单引号。

#### 表达式备注

用于填写对计算表达式的解释性内容，用户可根据需要填写，也可不填。

填写完目标表的ETL规则后，用户需要点击本节的“保存”按钮将新增和修改过的ETL映射规则保存到ETL元数据资料库中。下次需要编辑这些规则时，需要先点击本节的“刷新”按钮从ETL元数据资料库中取回上次保存的内容。

## 数据结构

本模板的运行依赖于ETL元数据资料库中下列表中的数据：

|  |  |  |  |
| --- | --- | --- | --- |
| 表名 | 中文名称 | 是否导入 | 说明 |
| src\_table\_analysis | 源系统表级分析结果表 | 是 | 存放源系统各表的表级属性。其中必填项包括：  schema\_name  table\_name  sys\_name  cn\_name  is\_inc\_ext  stbl\_name  各字段的说明参见模型文件。 |
| src\_column\_analysis | 源系统字段级分析结果表 | 是 | 存放源系统各表的字段级属性。其中必填项包括：  column\_name  sys\_name  schema\_name  table\_name |
| etl\_tasks | ETL任务表 | 否 | 存放ETL映射任务属性，一般每一张需要加载的仓库表对应一个ETL映射任务。其中必填项包括：  task\_name  schema\_name  table\_name  sys\_name  etl\_dvlpr\_name  serial\_no  利用admin用户执行刷新和保存ETL任务列表为这个表填入数据。 |
| dw\_tables | 数据仓库表 | 是 | 定义数据仓库中的实体。必填项包括：  schema\_name  table\_name  sys\_name  phy\_name  load\_mode  subject\_name  is\_single\_source  各字段的说明参见模型文件。 |
| dw\_table\_mapping | 数据仓库表级映射 | 否 | 定义源系统表和数据仓库表之间的映射关系。必填项包括：  load\_batch  schema\_name  src\_schema  table\_name  src\_table\_name  sys\_name（固定为’DW’）  src\_sys\_name  join\_order  table\_alias  join\_type  各字段的说明参见模型文件。 |
| dw\_column\_mapping | 数据仓库字段级映射 | 否 | 定义数据仓库基础层实体的字段级映射。这个表的数据来自用户在映射模板中填写的映射内容。 |
| dw\_subjects | 数据仓库主题表 | 是 | 定义数据仓库的主题。必填项包括  subject\_name |
| dw\_columns | 数据仓库实体属性表 | 是 | 定义数据仓库实体的属性。必填项包括：  column\_name  schema\_name  table\_name  sys\_name  column\_id  data\_type  phy\_name  is\_pk  chain\_compare  is\_partition\_key |
|  |  |  |  |

# ETL元数据资料库

ETL元数据资料库用于存储和数据仓库ETL过程有关的元数据，

## ETL元数据资料库的构成

目前的版本中的元数据可以分为源系统调研资料、数据仓库模型和ETL映射定义几部分：

### 源系统调研资料

### 数据仓库模型资料

### ETL映射定义

## 安装设置

1. 创建MySQL数据库（建议库名为etl\_metadata），使用utf8字符集。并创建用于访问该库的用户，并赋予足够的访问权限和远程访问权限。
2. 使用ETL开发工具包中提供的etl\_metadata.sql创建ETL元数据资料库中用到的表。
3. 根据需要导入已整理好的源系统资料和数据仓库模型资料。

# ETL脚本生成器

ETL脚本生成器能够根据ETL元数据资料库中提供的信息，为指定的数据仓库实体表生成加载脚本。目前的版本只支持为Hive平台生成脚本。在生成的加载脚本中，每个加载批次对应一条完整的HQL语句，如果加载批次包含多个组别，每个组别对应HQL中的一条SELECT子句。

## 用法说明

1. 将etlman-1.0-SNAPSHOT-bin.zip解压。
2. 在展开的目录下，编辑job-script.sh，配置ETL元数据资料库访问参数。
3. 在展开的目录下执行job-script脚本生成任务脚本，用法为：

       ./job-script.sh <实体名称> <输出脚本文件目录>

       例如：

              ./job-script.sh 人口 ./load\_scripts

程序执行完毕后，生成的脚本位于当前路径的load\_scripts子目录下。如果实体表的“分批加载”属性为0，则脚本文件名的格式为“<ETL任务名>.sql”，否则脚本文件被命名为“<ETL任务名>\_<批次号>.sql”。