# 数据映射技术在 ETL 过程中的应用①

梁吉胜[1]，李天阳[2]，王惠霞[2]，杨 锐[2]，张旭东[2]

[1](东北石油大学 计算机与信息技术学院，大庆 163318)
[2](河北汉光重工有限责任公司，邯郸 056028)

**摘 要**：为了使 ETL 系统能够高效的实现任意异构数据库之间的数据转换，需要通用的元模型支撑任意数据库之间数据转换元数据的定制。通过对数据转换中数据映射应用技术分析，重点对数据映射进行了抽象研究，定义数据映射的详细分类、基于数据映射关系的数据转换元数据描述形式，并构建支撑数据转换元数据定制的元模型，技术在中国石油数据中心大型数据迁移中应用，取得良好效果。

**关键词**：数据映射；数据转换；ETL；元模型；数据迁移

## Application of Data Mapping Technology to the ETL Process

LIANG Ji-Sheng[1], LI Tian-Yang[2], WANG Hui-Xia[2], YANG Rui[2], ZHANG Xu-Dong[2]

[1](College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Hebei Hanguang Heavy Industry Co, Ltd, Handan 056028, China)

**Abstract**: In order to enable the ETL system to effective achieve any data conversion between heterogeneous databases, we need a common meta-model for support the data conversion metadata customization. Through the analysis of the application of data mapping technical in the data conversion, research has been focused on the abstract study of data mapping, and defined the detailed classification of data mapping, and the form of data conversion metadata based on data mapping, build the meta-model which support to the custom data transformation metadata. The technology has been applied well in large-scale data migration of China's oil data centers.

**Key words**: data mapping; data conversion; ETL; meta-model; data migration

ETL(Extraction-Transformation-Loading)是信息化建设过程中不可避免的数据处理过程，数据转换是 ETL 过程中将抽取出来源数据转换为目标数据的数据处理过程[1]，是 ETL 实现的重点。传统实现 ETL 的 EAI 解决方案主要基于不同数据模型之间具体的映射编写数据交换代码实现特定数据转换，不具有通用性。目前的高效的 ETL 系统要求数据转换能够基于不同的元数据脱离具体数据库实现任意两数据库之间的数据转换，因此，需要通用的数据转换元数据描述形式，及能够支撑任意数据库间转换元数据定制的元模型[2]。本文通过对数据模型之间数据映射关系及通用数据转换需求的分析，定义基于数据映射的通用数据换转元数据描述形式，并构建支撑数定制的通用元模型。

## 1 数据映射分析定义与分类

数据映射是指同一数据领域内，存储相关数据的不同关系数据库数据模型之间的对应关系；数据映射与数据库设计模型等系统知识相关，共分为实体映射、表映射以及属性映射几个层次[3]：

①实体映射：用户可以看到的最高对等图，反映了两个数据库概念模型上的差别。

②表映射：表与实体相关，表映射是实体映射的充分反映，为数据模型数据表之间的对应关系。

③属性映射：数据映射的最底层，是对应数据表中具体属性的对应关系，是数据转换处理的最小单位。

实体映射和表映射，强调了实体完整性和引用完整性，最终实现前两种对应关系的基础是属性映射，

① 基金项目:国家自然科学基金(61170132);黑龙江省自然基金(11541008)
收稿时间:2011-11-04;收到修改稿时间:2011-12-30

因此为数据转换处理中的最小单位。

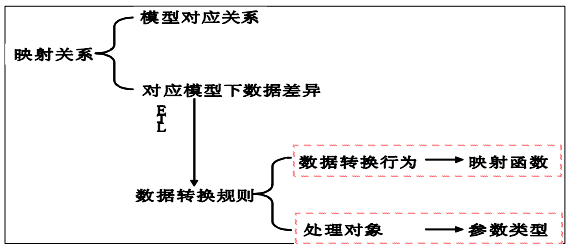


图 1　属性映射抽象图

关系模式中不同数据模型的属性映射如图 1 所示，其包含两个部分：①关系模式的模型对应关系(属性对应关系)；②模型对应下的数据差异。在 ETL 过程中数据差异表现为不同的数据转换规则。通过对数据转换规则的分析，数据转换规则可以抽象两个部分：①数据处理行为；②数据处理过程中涉及到的模型结构对象或特定参数值。通过分析不同数据模型之间的相同属性映射的数据转换规则之后，发现相同属性映射的数据处理行为相同，处理过程涉及的处理对象及对象相关属性不同；为了提高属性映射的数据转换行为的重用性，将数据处理过程涉及的对象及对象相关属性从数据处理过程中剥离出来，将数据处理行为抽象为一个数据转换函数；处理过程涉及的对象及相关属性定义为函数的参数类型(处理对象及相关属性的统一描述)，转换函数是数据转换行为的原子函数，具有通用性。不同数据模型间相同属性映射通过定义不同参数类型基于通用的转换函数实现数据转换，使数据转换脱离具体数据模型。由此可以使用转换函数信息与参数值构造 ETL 过程中的数据转换元数据，并构建支撑数据转换元数据定制的通用元模型，实现对不同数据模型的属性间数据映射的描述，基于元数据完成异构数据库之间的数据转换，使数据转换脱离具体模型。基于以上思想给出如下定义：

定义 1.属性映射：源数据库中实体 Es 的某一个属性或属性集 A（数据表的某一个或一组字段）到与含义相同的目标数据库实体 ET 属性或属性集 B 的对应关系，为 A 到 B 的属性映射，记为 M：A->B。

定义 2.映射函数：基于属性映射 M：A->B 源数据库中记录 $R_s$ 中对应于属性 A 的数据 $a_i$ 到目标记录 $R_t$ 中对应于属性 B 的数据 $b_i$ 的一系列变换、运算和统计等转化处理过程，为属性映射 M 下的映射函数 F，记为 F：$a_i$->$b_i$ 或 $b_i$=F（$a_i$）。

定义 3.参数类型：为了提高映射函数的重用性，将不同属性下相同属性映射的函数 F 处理数据 ai 过程中处理细节相同或作用相同的可变对象、属性定义与数据值抽象统一的名称，定义为参数类型。在数据转换过程中参数类型下具体值为映射函数的参数值，用 P 表示；P=<$P_1$，$P_2$，$P_3$，，，$P_i$>。引入函数参数类型概念后，映射函数变化为：$b_i$=F($a_i$)➔$b_i$=F($a_i$，P)。

基于定义 1，2，3，给出属性映射的分类与定义，定义映射函数功能，抽取参数类型，具体如下：

[关系 1]直接映射：源与目标对应属性列数据值相同。

映射函数：无需处理。

参数类型：　<V> V：源数据。

[关系 2]类型转换映射：源与目标对应映射属性列的数据值的类型不一致。

映射函数：实现支持多 DBMS 的数据类型转换。

参数类型：<源数据类型，目标数据类型>

[关系 3]单位转换映射：源与目标对应属性列的数据值的单位不一致。

映射函数：实现基于常用单位之间转换公式完成数据不同单位之间的数学运算得到目标数据值。

参数类型：<源单位，目标单位>

[关系 4]主键映射：源表主关键属性列与目标表主关键属性列相对应，但主键值不一致。

映射函数：实现将源主键值转换为目标主键值，构建中间表保存源主键值与新目标主键值之间对应关系。

参数类型：<源主键属性列名，目标主键属性列名，中间代码表名>

[关系 5]目标外键映射：源非外键属性列与目标外键属性列对应，源非外键属性列的数据值与目标外键属性列引用的参照属性列数据值相一致。

映射函数：实现基于参数类型动态 SQL 的拼接。

参数类型：<目标外键引用表名，引用属性列名，目标参照属性列名>

[关系 6]源外键映射：源外键属性列与目标非外键属性列相对应，源表外键引用中某个参照属性列数据值与目标非外键属性列的数据值相一致。

映射函数：实现基于参数类型完成动态 SQL 的拼接。

参数类型：<源外键引用表名，引用属性列名，查询参照源属性列名>

[关系 7]多值计算映射：源若干属性列数据值计算处理后与对应的目标属性列的数据值相一致。

映射函数：实现四则运算函数，实现基础四则运算，复杂运算以四则运算函数为基础，通过调用得到结果。

参数类型：<源属性列名，计算方式，优先级序号，目标属性列名> 注：计算方式为+/-/*/÷。

[关系 8]自计算映射：源属性列的数据值经过与数字运算后与对应的目标属性列的数据值相一致。

映射函数：实现四则运算函数，实现基础四则运算，复杂运算以四则运算函数为基础，通过调用得到结果。

参数类型：<源属性列名，计算方式，优先级序号，数据参数值> 注：计算方式为+/-/*/÷。

[关系 9]直接合并映射：源若干属性列数据值通过合并处理后与对应的目标属性列数据值相一致。

映射函数：实现字符数据值按序拼接。

参数类型：<源属性列名，合并顺序号，目标属性列名>

[关系 10]按符号合并映射：源若干属性列的数据值通过特定符号合并处理后与目标属性列数据值相一致。

数据转换规则：实现字符数据值按序符号拼接。

参数类型：<合并符号，合并顺序号>

[关系 11]固定截取映射：源属性列数据值通过截取处理后与对应的目标属性列数据值相一致。

映射函数：实现字符数据值的按位截取。

参数类型：<截取起始位置，截取终止位置>

[关系 12]按符号拆分映射：源属性列数据值通过符号拆分处理与对应的目标属性列数据值相一致。

映射函数：实现字符数据值的按符号截取，并获取指定方位的分割后数据。

参数类型：<截取符号，数据方位> 注：数据方位为符号左/符号右。

[关系 13]源代码映射：源代码字段与目标非代码字段相对应，源引用的代码表中代码名称属性列的数据值与目标非代码属性列的数据值相一致。

映射函数：实现基于参数类型完成动态 SQL 的拼接。

参数类型：<源代码表名，源代码值属性列名，源代码名称属性列名>

[关系 14]目标代码映射：源非代码属性列与目标代码属性列对应，源非代码属性列的数据值与目标引用代码表中的代码名称属性列的数据值相一致。

映射函数：实现基于参数类型完成动态 SQL 的拼接。

参数类型：<目标代码表名，目标代码值属性列名，目标代码名称属性列名>

[关系 15]代码值映射：源代码属性列与目标代码属性列对应，但编码标准不一致。

映射函数：实现动态拼接 SQL。

参数类型：<源代码表名，源代码值属性列名，源代码名称属性列名，目标代码表名，目标代码值属性列名，目标代码名称属性列名>

[关系 16]常量映射：源属性列的数据值与目标属性列常量数据值相一致。

映射函数：实现常量赋值。

参数类型：<常量值>

[关系 17]组合映射：源属性列与目标属性列对应，源属性列数据值需要经过多次已存在的数据映射有序组合处理后得到目标属性列数据值。

映射函数：实现根据多次数据映射先后顺序，完成相应的数据处理，能够灵活调用任何数据映射的数据映射函数，保存中间结果，完成指定顺序的数据处理。

参数类型：<属性映射标识，顺序号>

[关系 18]依赖映射：源属性列与目标属性列对应，通过其他源属性列特定条件下获取的源属性列的数据值与目标字段数据值相一致。

映射函数：实现通过依赖属性列与依赖条件获取得到目标属性列数据值。

参数类型：<依赖字段，依赖条件>

[关系 19]记录拆分：源属性列与若干不同目标属性列对应，源表中一条记录直接或者在特定条件下拆分为若干条记录与目标表中记录相一致。

映射函数：实现根据源属性列与目标表属性列的数据映射，将源表中的一条记录转换成目标表中多条记录，任意调用其他映射函数处理数据的同时将源表一条数据记录基于主键不变进行记录拆分。

参数类型：<源属性列名，目标属性列名集合，目标属性列对应条件集合>

[关系 20]记录合并：源表属性列与若干目标属性列对应，源表中多条记录合并为目标表一条记录。

映射函数：实现根据源属性列与目标表属性列的数据映射，将源表中的多条记录转换成目标表中一条记录，任意调用其他映射函数处理数据的同时将源表多条数据记录基于主键不变进行记录组合。

参数类型：<源属性列名，目标属性列名集合，目标属性列对应条件集合>

## 2　数据映射技术应用

通过对数据映射的抽象分析，其参数类型主要为数据库结构信息与自定义的参数，不同数据库之间的数据转换实现基于不同的模型结构信息，为了使数据转换能够脱离具体的模型，在参数类型分析以后，使用数据字典元数据结合映射函数与参数类型元数据构建通用元模型，同时基于映射函数与参数定义数据转换元数据描述形式，实现基于元模型灵活定制任意数据库之间的数据转换元数据，模型结构图如图 2 所示。

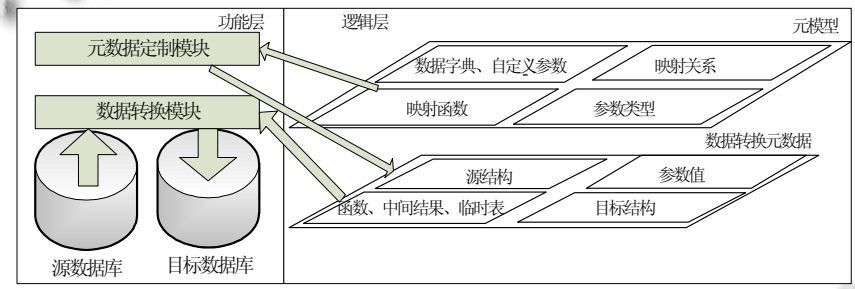


图 2　模型结构图

逻辑层分为两个部分：元模型与数据转换元数据，元模型通过功能层中的元数据定制模块[4]定制任意数据库间的数据转换元数据，功能层的数据转换模块加载数据转换元数据实现不同数据库间的数据转换。数据转换元数据用一个六元组描述：Me=<Sc，Fn，P，Tc，It，Mt>，其中 Sc 为源数据结构信息、源数据存储信息；Fn 为映射函数名称；P 为参数值集合；Tc 为目标结构、目标代码相关数据信息；It 为函数中间结果存储位置，下一函数名称等；Mt 为中间临时表名称，属性名称等信息。

元模型用五元组描述[5]：M=<D，S，R，F，A>，其中 D 为数据字典信息，包括数据库位置及属性相关信息、数据库中数据表、表字段、字段属性信息等；S 为自定义的符号标示集合，包括运算符号；R 为数据映射信息，包括数据映射标识、数据映射名称；F 为映射函数信息，函数所属数据映射、函数标识、函数物理位置、函数调用方式；A 为参数类型信息，参数所属函数、参数序号、参数属性、参数存储位置。

元模型中包含数据字典信息存储结构，能够存储任意数据库的结构信息，结合属性映射，映射函数及参数类型描述信息，通过元数据定制模块能够定制任意数据库之间的转换元数据，基于数据转换元数据驱动数据转换模块实现任意数据库之间的数据转换，元模型具有通用性，同时，当出现新的数据映射，构建相应的映射函数，抽取参数类型，存储到元模型中，通过定制即可进行数据转换，元模型具有可扩展性。

采用该技术开发的 ETL 系统与传统 ETL 系统相比较具有如下优势：(1)实用性高：使用数据字典元数

据，通过不同数据字典定制数据转换元数据能够实现不同数据库的数据迁移，具有通用性。(2)扩展性好：使用基于数据映射的数据转换元数据，不同数据模型之间的数据迁移过程中，数据映射模式繁多，当出现一种新的映射模式只需要编写独立的映射函数，在元模型中添加函数及参数类型信息，定制之后确定使用函数及参数值即可实现新映射的数据转换。(3)易维护，使用元数据驱动的方式实现 ETL 功能，维护工作大部分为修改元数据，使维护简化。

ETL 系统在“大庆油田公司数据中心建设项目”中得到了较好应用，数据中心数据模型构建完成后，需从 A2 数据库、勘探开发数据库、压裂曲线数据库、计划统计数据库、经营管理数据库中迁移业务数据，以保证数据中心的正常启动运行，该项目使用该系统完成了 A2 数据库 62 张数据表 1223 万条生产数据、勘探开发数据库 37 数据表 413 万条基础数据、压裂曲线数据库 16 张数据表 369 万条压裂数据、计划统计数据库 22 张数据表 160 万条统计数据、经营管理数据库 18 张数据表 35 万经营数据的数据迁移任务，应用证明系统具有较好的实用性和较高的应用价值，目前系统正准备在中石油推广。

## 3 结语

本文重点研究了数据映射，对数据映射进行抽象，定义了映射函数与参数类型，并通过数据映射的分析，给出了基于数据映射的数据转换元数据的描述形式，并构建了通用的元数据定制元模型，技术与系统在实际应用中证明具有良好的通用性、适应性、可移植性、可扩展性，能够对数据迁移需求做出快速反应，减轻用户的负担。

## 参考文献

1 宋杰,王大玲,鲍玉斌.一种元数据驱动的 ETL 方法的研究.小型微型计算机系统,2007,28(12):2167−2173.
2 孙伟,张忠能.ETL 架构研究.微型电脑应用,2005,21(3):13−15.
3 胡晓鹏,李晓航,李岗.一种基于 xml 映射规则的数据迁移方法设计和实现.计算机应用,2005,25(8):1849−1852.
4 熊辉,刘彦峰,郭大庆.分布式异构数据库迁移系统的设计与实现.计算机工程,2008,34(4):57−59.
5 Zhao XF, Huang ZQ. A formal framework for reasoning of Meta-data based on CWM.The 25th International Conference on Conceptual Modeling. 2006.371−384.

(上接第 138 页)

科学版),2006,12(4):3.
2 陈尚松,杜旭英,俞欢军.基于 Struts+Hibernate+Spring 框架的毕业设计管理系统.计算机工程与设计,2008,29(15):54−57.
3 陈刚.基于 SSH 的 J2EE 开发平台研究与应用[硕士学位论文].成都:四川师范大学,2007.
4 陆荣幸,郁洲,阮永良,王志强.J2EE 平台上 MVC 设计模式的研究与实现.计算机应用研究,2003,20(3):145.
5 周杨.AJAX 应用的典型设计模式.计算机系统应用,2011,20(1):128−132.
6 田珂,谢世波,方马.J2EE 数据持久层的解决方案.计算机工程,2003,29(22):93−95.
7 肖竟华,何洁.PKI 技术在网上报税系统中的应用.电脑与信息技术,2006,14(4):40−42.
8 李一鸣,张剑,李哲,黄鑫.WebSphere 性能问题的发现及其处理对策.电脑知识与技术,2009,5(6):93−94.