基于结构图的 ETL 过程建模方法*

张忠平，赵瑞珍
(燕山大学 信息科学与工程学院 计算机应用技术系，河北 秦皇岛 066004)

摘 要：ETL 是数据仓库获得高质量数据的重要环节，在数据仓库建设过程中占有极其重要的地位。为了便于 ETL 过程的设计、维护和修改，提出一种基于结构图的 ETL 过程建模方法，并依据该方法完成了 ETL 概念模型的设计。通过图形化 ETL 过程中的元素和关联，该模型清晰直观地反映了数据的来源和流向、源数据与目标数据之间的映射和转换关系，辅助设计人员更好地进行 ETL 过程的设计和 ETL 过程的编码实现，使整个 ETL 设计过程更加方便、灵活。

关键词：数据仓库；抽取转换加载；结构图；模型
中图分类号：TP311 **文献标志码**：A **文章编号**：1001-3695(2008)11-3354-03

Architecture graph-based approach for modeling ETL process

ZHANG Zhong-ping, ZHAO Rui-zhen
(Dept. of Computer Application Technology, College of Information Science & Engineering, Yanshan University, Qinhuangdao Hebei 066004, China)

Abstract: ETL is an important part for the data warehouse to gain data with high quality, and it plays a key role in building the data warehouse system. The paper proposed a methodology for modeling ETL process based on an architecture graph, with the goal of facilitating the designment, maintenance and modification of the ETL process. On the basis of this modeling approach, the paper completed the design of ETL conceptual model. By representing the elements and relationships of the ETL process diagrammatically, it expressed the data's coming and going, as well as their mapping and transformation relationships clearly and intuitively. It also supported the ETL designer to design the ETL process and develope the code efficiently, and improved the flexibility and reliability of the ETL process designment greatly.

Key words: data warehouse; ETL; architecture graph; model

0 引言

ETL(extract-transform-load)是数据抽取转换加载过程，即从各种异构操作型数据源中抽取数据，并对抽取到的数据进行转换处理(包括数据转换、清洗和标准化)，最后加载到数据仓库中的过程。它是建立数据仓库的基石和灵魂，也是建立数据仓库的必要步骤[1,2]，在数据仓库建设过程中占有极其重要的地位。一般情况下，开发 ETL 的工作量要占整个数据仓库系统开发量的 60% ~80%[3]。

ETL 过程的设计和维护是构建一个成功数据仓库方案的关键因素，因为不正确的或误导的数据将产生错误的业务决策信息，所以在数据仓库建设的早期阶段，一个正确的 ETL 过程设计能大大提高数据质量，从而为企业提供有用的决策信息。在传统的数据库建设过程中，往往强调数据仓库本身的建模和前端数据展示，而对 ETL 的设计和建模关心不足；同时，在目前数据仓库的早期设计阶段，多数是采用 E-R 图来表达数据仓库概念模型[4]，这些模型不能很好地表达数据来源和数据间的联系，不能满足 ETL 过程设计的需要。因此，需要一种建模方法来很好地设计和维护 ETL 过程，以便让用户在统一数据模型上定义 ETL 过程，清晰地反映 ETL 过程中各个源与目标的内部结构和组成、明确数据的流动去向和转换方式、追踪数据来源与转换信息、屏蔽 ETL 过程中具体实现的差异，帮助开发人员理解 ETL 架构，辅助设计人员设计出质量较好的 ETL 过程。

数据仓库中数据的正确性和有效性需要 ETL 过程来保证，ETL 过程的正确与否关系到数据仓库的可用性，所以如何有效地为 ETL 过程设计一种模型，将是迫切需要解决的问题，并且具有广阔的应用前景。

1 基本概念

1.1 ETL 原理

ETL，即数据抽取、转换、加载的过程。主要包含以下三个方面内容[5,6]：

a) 抽取(extract)。数据抽取是捕获数据源的过程，即将数据从各种原始的业务系统中读取出来。这是所有工作的前提。

b) 转换(transform)。按照预先设计好的规则将抽取得到的数据进行转换、清洗，处理一些冗余、歧义、不完整、违反业务规则的数据，统一数据的粒度，使本来异构的数据格式统一起来。

c) 加载(load)。将转换后的数据按照计划增量或全部导

收稿日期：2007-12-20；修回日期：2008-03-12 基金项目：国家自然科学基金资助项目(60773100)；国家教育部科学技术研究重点资助项目(205014)；河北省教育厅科研计划资助项目(2006143)

作者简介：张忠平(1972-)，男，副教授，硕导，博士，博士后，主要研究方向为数据仓库、网格技术、XML 数据库(zpzhang@ysu.edu.cn)；赵瑞珍(1984-)，女，硕士研究生，主要研究方向为数据仓库、ETL 技术.

入到数据仓库中。

从整体角度来看, ETL 的主要作用在于其屏蔽了复杂的业务逻辑, 从而为各种基于数据仓库的分析和应用提供了统一的数据接口, 这正是构建数据仓库的重要目的。

1.2　ETL 概念模型

ETL 概念模型[7]描述了整个 ETL 任务流程的逻辑结构, 这种结构独立于任何软件或者数据存储结构, 只给出一个在 ETL 过程中运行所需数据的形式表示。ETL 概念模型并不针对具体的工具制订, 与数据库系统无关、与应用程序无关、与工具无关。在这一概念级中, 不必考虑物理实现的细节, 只将注意力集中在构造源与目标实体及属性之间的映射和转换关系上。

1.3　结构图

定义 1　结构图。定义结构图为一个有向图 $G=(V, E)$。其中: V 为图的节点, E 为图的边, $V=\mathrm{E}\cup\mathrm{A}\cup\mathrm{T}\cup\mathrm{Cn}\cup\mathrm{N}\cup\mathrm{D}$, $E=\mathrm{Po}\cup\mathrm{Pr}\cup\mathrm{Cr}\cup\mathrm{Ir}$。符号 E、A、T、Cn、N、D 分别表示 ETL 模型中的实体、属性、转换、ETL 约束、注释和数据域; Po、Pr、Cr、Ir 表示元素之间的部分关系、供应关系、候选关系和实例关系。此外, 为这些元素和关联分配一个定义良好的图符, 使用这些图符来表示 ETL 模型中的要素组成部分。结构图中的节点和边的图形表示分别如表 1、2 所示。

表1　结构图的节点

名称	图符	简写	名称	图符	简写
实体	实体名	E	属性	属性名	A
转换	转换	T	ETL约束	条件	Cn
注释	注释内容	N	数据域	域名	D

表2　结构图的边

名称	图符	简写	名称	图符	简写
部分关系	——◇	Po	供应关系	——→	Pr
候选关系	{XOR} target	Cr	实例关系	·····→	Ir

2　ETL 过程建模

2.1　实例

为了便于讨论, 下面给出一个简单实例[8]。该实例描述了从源数据库 S_1 和 S_2 中抽取数据加载到数据仓库 DW 的过程。其中: S_1 是欧洲数据库; S_2 是美国数据库; DW 是欧洲数据仓库。

表 S_1. parts(Pkey, Qty, cost) 中的数据由表 PS_1(Pkey, Qty, dept) 和表 PS_2(Pkey, cost) 通过外连接得到。

表 S_2. parts(Pkey, date, Qty, cost, dept) 的数据来源可以是 AnnualParts(存储季度信息) 或 RecentParts (存储日信息) 。

表 DW. parts(Pkey, date, Qty, cost) 存储每种商品(Pkey) 成本(cost) 和数量(Qty) 的日(date) 信息。

2.2　ETL 模型中的要素描述

1) 实体　对应于源数据库中的表和文件以及数据仓库中的事实表和维表, 由表名和一组属性集描述, 如 S_1. parts(Pkey, Qty, cost) 、S_2. parts(Pkey, date, Qty, cost, dept) 和 DW. parts (Pkey, date, Qty, cost) 均表示实体。用矩形框来表示 ETL 过程中的实体, 框内注明实体的名称。

2) 属性　作用如同 ER 模型中的属性, 是实体的最小信息单位, 如表的列。实体 S_1. parts 中的属性有 Pkey、Qty、cost。

3) 转换　将数据从源数据库映射到目标数据仓库所经过的一系列操作称为转换。转换分两类: a) 过滤或数据清洗操作, 如非空值检测、主外键违反检测等; b) 转换操作, 如数据类型转换、代理键分配等。

表 3 列出了 ETL 模型中的主要转换操作, 并赋予各个操作一个图符。

表3　转换活动集

类别	名称	符号	类别	名称	符号
过滤 (Filters)	选择(select)	s	二元转换操作 (binary operation)	关联(union)	U
	非空(not null)	NN			
	主键违反 (primary key violation)	PK		连接(join)	⋈
	外键违反 (foreign key violation)	FK		差(diff)	△
元转换操作 (μnary ration)	汇总(aggregation)	g	转换操作 (transfer operation)	FTP	FTP
	投影(projection)	Π		压缩(compress)/ 解压缩(decompress)	Z/dZ
	函数应用 (function application)	f			
	代理键分配 (surrogate key assignment)	SK		加密(encrypt)/ 解密(decrypt)	Cr/dCr

4) ETL 约束　用 ETL 约束来表示某数据必须满足一定条件。如在属性或属性集上进行主键约束或非空值约束, 就表示这些属性或属性集必须满足主键约束或非空条件。

5) 注释　注释是对实体或转换等进行描述或解释。在 ETL 模型中的作用一般包括以下几个方面: a) 设计策略或约束的解释; b) 应用函数的语义解释; c) ETL 过程中基于时间/事件的调度、监控、异常处理、错误恢复等的解释说明。

一个注释可包含多条子句, 每个子句定义为 〈类型〉:: 〈内容〉。类型有三种形式, 分别为 f: : (表示函数类型) 、e: : (表示表达式类型) 和 t: : (表示简单文本类型) 。

6) 数据域　表示模型中实体的来源, 如 S_1. parts 和 S_2. parts 分别来自源数据库 S_1 和 S_2, DW. parts 则来自数据仓库 DW。

7) 部分关系　部分关系表明实体与属性之间以及实体与数据源之间的归属关系, 即实体与其相应的属性集之间以及实体与数据源之间存在部分关系, 如源数据库 S_1 和表 S_1. parts 以及表 S_1. parts 和其属性(Pkey, Qty, cost) 之间是部分关系。

8) 供应关系　表示数据来源、流向或数据之间的关系。在实体级, 供应关系将原实体映射到目标实体, 原实体为供应者, 目标实体为消费者; 在属性级, 供应关系通过相关转换将输入属性映射到输出属性, 输入属性为供应者, 输出属性为消费者。

9) 候选关系　在数据仓库环境中, 与数据仓库中的数据有关联的源表可能是由多个候选源表或源文件提供, 通过候选关系将这些候选源标志出来, 以备 ETL 过程修改或维护之用。如源表 S_2. parts 中的数据既可以从表 AnnualParts 中得到, 又可以从表 RecentParts 中得到, 那么文件 AnnualParts 和 RecentParts 即为表 S_2. parts 的候选数据源, AnnualParts 和 RecentParts 与表 S_2. parts 之间存在候选关系。

在多个候选源中, 只选择其中一个作为这个实体的活动候选数据源。

10) 实例关系　实例之间的关系用一端带箭头的虚线表示。

2.3　ETL 过程的建模步骤

2.3.1　数据来源的确定

在数据仓库设计的最早阶段, 设计者关注两样工作: 用户需求的收集和数据源分析。依照用户需求建立目标数据模型, 根据目标数据模型进行数据源的分析和验证, 收集有关数据源的信息, 如数据源的位置、数据源所处的平台等, 并且确认哪些

是来自正式的数据源或者是非正式的数据源(正式的数据源是由业务系统提供支持;而非正式的数据源,如分析竞争对手时的市场占有调查报告等,这些是不能由现有的业务系统支持,而是来自于用户的收集与使用,这些信息往往需要一个获取信息的处理过程,将其收集到数据仓库中),从而为每个事实表与维表确定数据来源,同时确定 ETL 过程的范围。

在本文提供的实例中,为 DW. parts(Pkey, date, Qty, cost) 提供数据的数据源有 S_1. parts(Pkey, Qty, cost) 和 S_2. parts(Pkey, date, Qty, cost, dept);为 S_1. parts(Pkey, Qty, cost) 提供数据的数据源有 PS_1(Pkey, Qty, dept) 和表 PS_2(Pkey, cost);为 S_2. parts(Pkey, date, Qty, cost, dept) 提供数据的数据源可以是 AnnualParts 或 RecentParts,且表 S_1. parts 为欧洲的值和格式,S_2. parts 为美国的值和格式。

2.3.2 候选数据源的确定

确定了数据源后,必须仔细研究每个数据源的内容以及可获得性程度等。因为在某个系统中同样一个目标值的数据来源可能会有多个,这些数据必须要考虑它的其他方式的来源。这样该过程并不是一个自动化的过程,更多的是依靠经验,根据数据量、数据质量、数据内容、数据完整性等方面来确定哪个是将要使用的数据源,并选择需要的数据内容。通过对所有的数据源进行详细的分析,了解其真实的数据内容。

S_2. parts(Pkey, date, Qty, cost, dept) 的数据可以由 AnnualParts 或 RecentParts 得到,所以表 S_2. parts 有两个候选数据源,具体选哪个候选源作为将要使用的数据源还需进一步考虑。

2.3.3 源数据到目标数据的映射

数据变换过程中的数据流满足数据库、实体和属性三个层次上的映射关系,它们决定数据流的来源和去向。数据库层映射定义源数据库与目标数据库间的关系,实体层映射定义源实体与目标实体间的关系,属性层映射是数据源和目标实体层中最低层的映射关系,即表中字段间的映射关系。当数据源中的一个数据项与数据仓库的目标建立属性与属性的映射关系,就反映在实际的 ETL 过程执行中数据仓库中的数据项是从哪个特定的数据源填充,这些工作包括属性到属性的映射、属性转换两部分。在简单的情况下,源与目标之间可直接映射,但大多数情况下源与目标的映射必须经过适当的转换才能得以实现。

建立源与目标之间的映射关系具体分为以下七个步骤:

a) 根据结构图的定义画出 ETL 过程中涉及的数据实体及其相关属性;

b) 确定实体间的连线,画出实体间的连线与箭头,标明实体的关系;

c) 根据原始数据和目标数据确定中间结果的数据结构;

d) 画出源与目标之间的数据来源与流向的连线和箭头,标明数据间的关系;

e) 根据需要确定各属性间的转换,标明转换过程与转换类别;

f) 确定输入的各个参数值;

g) 精练数据结构图。

2.4 ETL 概念模型

根据本文提供的实例,按照以上设计步骤,运用结构图中的图符和形式化规则,画出基于结构图的 ETL 概念模型如图 1 所示。

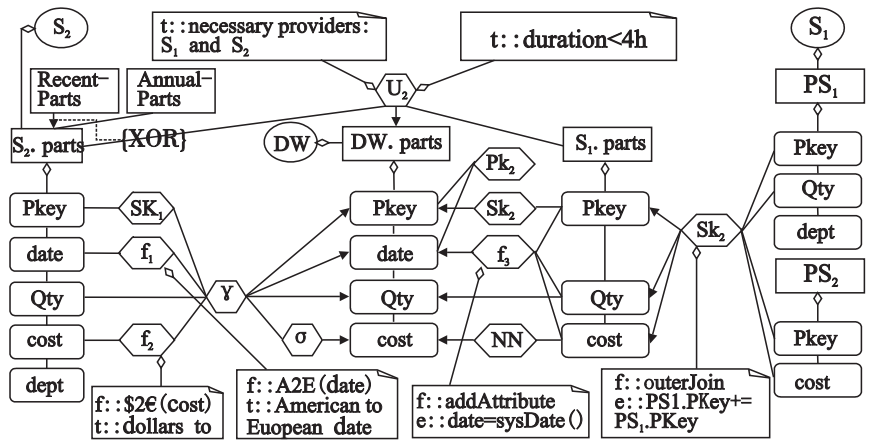

图1 ETL概念模型

利用结构图中定义良好的图符来表示 ETL 过程中的元素和关联,从而图形化整个 ETL 过程,并按照一定规则完成 ETL 概念模型的设计。从图 1 可以形象直观地看出,整个 ETL 过程涉及到了哪些数据源、源数据,通过怎样的转换操作映射到目标数据、转换操作的语义是什么以及属性或属性集上必须满足什么样的约束条件等设计 ETL 过程所必需的信息。

3 结束语

在数据仓库的早期设计阶段,提出一种基于结构图的 ETL 过程建模方法来建立图形化的 ETL 过程框架。本文首先给出了结构图的定义,图形化表示 ETL 模型中的元素;其次,对 ETL 模型中的元素进行详细描述;最后,给出相应的设计 ETL 模型的主要步骤,并基于具体实例完成了 ETL 概念模型的设计。通过该设计方法使整个 ETL 设计过程规范化、重用化,清晰地指导开发人员在 ETL 过程的代码开发,提高工作效率。

参考文献:

[1] HANG Xu-feng, SUN Wei-wei, WANG Wei, *et al*. Generating incremental ETL processes automatically[C]//Proc of the 1st International Multi-Symposiums on Computer and Computational Sciences. Washington DC: IEEE Computer Society, 2006: 516-521.

[2] 毛彧. 银行数据仓库系统中 ETL 的总体设计与实现[J]. 信息与电子工程, 2007, 5(4): 292-295.

[3] 尤玉林, 张宪民. 一种可靠的数据仓库中 ETL 策略和架构设计[J]. 计算机工程与应用, 2005, 41(10): 172-175, 229.

[4] LEOPOLDE Z, MATILDE C. A model driven approach for data warehouse conceptual design[C]//Proc of the 7th International Baltic Conference on Databases and Information Systems. Piscataway: Institute of Electrical and Electronics Engineers Computer Society, 2006: 114-121.

[5] JUAN T, SERGIO L M. An UML based approch for modeling ETL processes in data warehouses[C]//Proc of Conceptual Modeling. Chicago: Notes in Computer Science, 2003: 307-320.

[6] 郭晓红. ETL 实施过程研究[J]. 网络与信息, 2007, 21(7): 67-67.

[7] LI Ze-hai, SUN Ji-gui, YU Hai-hong, *et al*. Common cube-based conceptual modeling of ETL processes[C]//Proc of the 5th International Conference on Control and Automation. New York: Institute of Electrical and Electronics Engineers Inc, 2005: 131-136.

[8] SIMITSIS A, VASSILIADIS P, SELLIS T. Logical optimization of ETL Workflows[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 17(10): 150-161.