基于元数据的数据交换系统任务模型①

李春花[1,2]，廉东本[2]

[1](中国科学院 研究生院，北京 100049)
[2](中国科学院 沈阳计算技术研究所，沈阳 110168)

摘 要：ETL 包含数据的抽取、转换、加载三个部分，是构建数据仓库的重要环节。为解决异构数据源集成问题，本文提出了基于元数据的数据交换系统，并在该基础上阐述了数据交换系统中的任务设计模型和任务调度模型。最后介绍了数据交换系统中的主要算法以及设计模式。

关键词：ETL；元数据；任务设计；任务调度；作业

Task Model in Data Exchange System Based on Meta Data

LI Chun-Hua[1,2], LIAN Dong-Ben[2]

[1](Graduate School, Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: ETL containing three parts of extracting ,transforming, loading is one core of building the Data Warehouse. For the sake of heterogeneous data source integration, this paper puts forwards a model of the data exchange system based on ETL Meta. Further, a task designing model and a task designing model are raised in in this paper. Last, the article recommends a main algorithm and the designing model in the data exchange system.

Key words: ETL; meta data; task designing; task scheduling; job

随着计算机技术、通信技术以及互联网技术的飞速发展，信息化程度已成为衡量一个国家、一个城市现代化水平和综合实力的重要标志。不同地区不同部门之间的数据交换日益频繁，数据交换需求逐步增加。目前，许多行业、单位和部门内部都逐步实现了业务信息的计算机化管理，开发了大量的软硬件平台各异的应用系统。本文提出的数据交换系统是在辽河流域开展水体污染控制及水体污染治理研究的基础上提出的。如何解决异构数据资源的集成问题，发挥数据信息应有的效能，对辽河流域水环境风险评估与预警平台课题影响重大。

数据集成工具是异构数据源集成领域研究的一个热点，数据集成工具通常也被称作 ETL 工具。这些 ETL 工具拥有各自的数据处理功能与任务制定方式。

文献[1]在传统的 ETL 框架中加入了数据质量控制模块。文献[2]提出了基于任务的数据交换平台，该数据交换平台将任务分为抽取任务、传输任务、加载任务。文献[5]提出了一种数据仓库 ETL 元模型设计方法，完成了 ETL 在概念层上的设计。文献[8]首先提出了任务单元的概念，进而阐述了任务设计模型以及基于贪婪算法的任务调度模型。

通过对 ETL 过程的学习和研究，结合数据交换过程中对任务设计的需求，本文提出了基于 ETL 元数据的数据交换模型，阐述了针对该模型的任务制定方式和任务调度方式，最后介绍了该数据交换系统的主要算法及系统性能。

1 ETL原理介绍

ETL 是 Extract-Transform-Load 的缩写，中文名称为数据抽取、转换、加载。ETL 将分布的、异构的数据源中的数据抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集中。在数据仓库

① 基金项目:国家水体污染控制与治理科技重大专项(2009ZX07528-006-05)
收稿时间:2011-07-15;收到修改稿时间:2011-08-22

实践过程中，ETL 宏观上可以看成一套数据整合解决方案，具体来讲也可以看成数据导入导出工具，是数据仓库体系结构中一个重要过程。

1.1 数据抽取

数据抽取功能是确定数据采集所涉及到的数据源并采集原始数据。数据抽取通过不用的数据接口，实现从不同的网络、操作平台、数据库及应用中抽取数据。该环节能通过对数据源的分析，抓取原始数据的元数据，为后续的数据转换等工作提供基础。

由于不同应用系统可能采用不同的数据存储技术，如关系型数据库，非关系型数据库，甚至文件系统等，因此采用何种数据访问接口来访问多种数据源是数据抽取过程需要解决的一个关键问题。制定数据抽取策略一方面要满足目标系统的需求，另一方面则必须保证不影响业务系统的性能。目前，实施数据抽取的手段主要包含增量抽取和完全抽取两种。

1.2 数据转换

数据转换是 ETL 过程中最复杂的部分，涉及较多的方法和技巧，数据转换包含数据的清洗和转换两部分功能。

数据的清洗要求对抽取来的原始数据进行有效性检查，对数据项丢失或无效的记录和相似重复记录进行处理。数据转换则根据数据抽取时获取的元数据信息和目标数据仓库中表的元数据信息来对数据项进行转换，其中包含数据的合并、汇总、过滤、转换等。

数据转换功能保证了数据的正确性、一致性、完整性和可靠性，为后续的工作提供了数据支持。

1.3 数据加载

数据加载负责将数据按照目标数据库元数据定义的表结构装入数据仓库，该功能对经过清洗和转换的数据进行汇总、保存、以达到数据整合的目的。

1.4 元数据

元数据是关于数据的数据，当数据在程序中不是被加工的对象，而是被用来对程序的运行起控制作用，并且可以通过值的改变而改变程序的行为时，这样的数据成为元数据。在 ETL 实际操作过程中，由元数据决定了抽取拿些数据等抽取规则，并记录了数据转换过程中的每一步详细信息。

ETL 元数据主要包含三类：业务元数据、技术元数据和过程元数据。业务元数据是为企业业务用户提供支持的元数据，而技术元数据和过程元数据是为技术人员提供支持的元数据。

2 数据交换系统的任务设计

2.1 基于元数据的系统结构

基于元数据的系统结构图如图 1 所示：

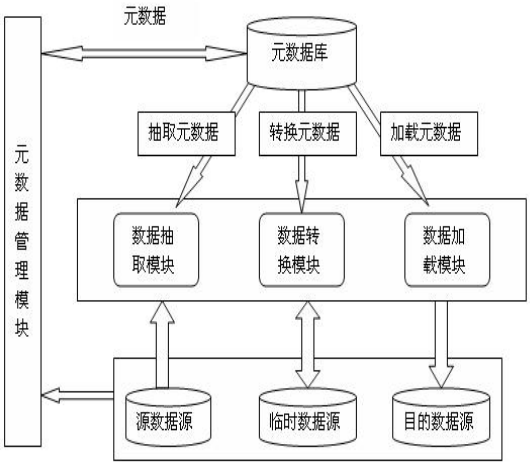


图 1 系统结构图

1) 抽取元数据，在数据抽取过程中需要有相应的映射规则以及数据源的连接信息等，我们把与数据抽取相关的元数据称为抽取元数据。

2) 转换元数据，数据转换过程完成数据格式从源数据存储格式到目标数据存储格式的转换。这个过程需要源数据存储格式和目标数据存储格式信息以及所有的数据转换规则信息。我们把以上与数据转换过程相关的元数据称为转换元数据。

3) 加载元数据，在数据加载过程中也需要映射规则等方面的元数据，我们把与数据加载有关的元数据称为加载元数据。

4) 元数据模块，元数据模块包含元数据库、元数据定义和元数据管理，元数据管理模块的主要功能是生成、保存、修改元数据。

5）任务清单，在数据交换系统中，某个任务制定完成后，该任务中的抽取元数据，转换元数据，加载元数据构成该任务的任务清单。任务调度器根据任务清单进行任务调度。

2.2 数据交换系统中的任务设计

数据转换是 ETL 过程中最复杂的部分，它包含数据格式转换、数据类型转换、数据汇总计算、数据拼

接等等。在进行某类数据转换之前必须有数据源的输入，在数据转换之后必须有数据源的输出。

在本文提出的数据交换系统中，作业（job）代表一个数据转换过程，每个作业能够完成特定的数据转换功能。多个作业的有序组合构成一个任务（task），任务也可以包含其他任务。任务完成整个工作流程的控制，作业完成针对数据的基础转换。

1）作业。

作业主要完成数据的基础转换，为实现异构数据的转换，本系统提供多类输入输出功能。一个作业可以包含多个数据输入、多个数据转换、多个数据输出，并且至少包含一个输入和一个输出。

作业的结构如下图 2 所示：

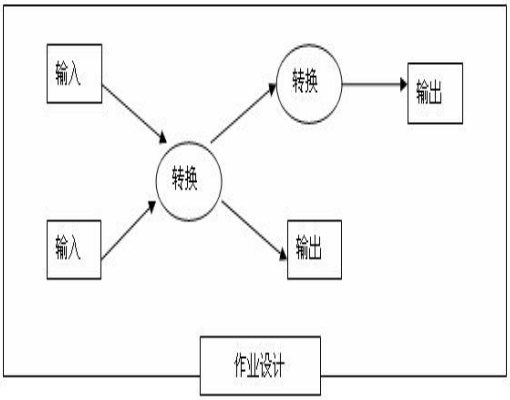
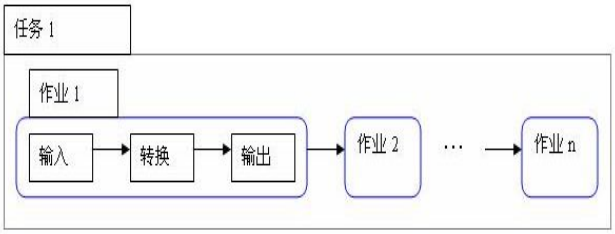


图 2　作业模型图

2）任务。

任务主要完成对工作流程的控制，本系统中任务的制定规则为：任务包含作业，任务也可以包含其他任务。任务中的作业完成数据的基础转换，这些作业可以单独执行，而任务是将不同作业进行有序连接，规定了这些作业的执行顺序。任务制定完成后，与之相关的所有抽取元数据、转换元数据、加载元数据构成了该任务的任务清单。

任务的结构如下图 3 所示：



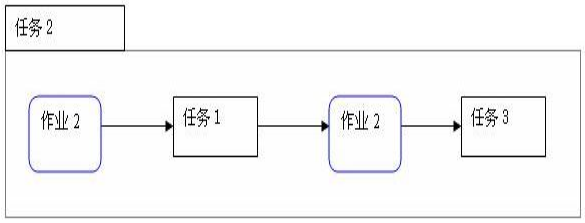


图 3　任务模型图

2.3 数据交换系统中的任务调度

1）数据交换系统中的任务调度模型如图 4 所示：

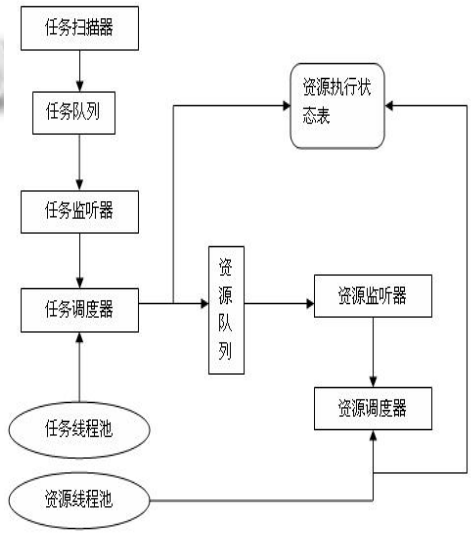


图 4　任务调度模型图

3）任务调度流程

任务扫描器将需要执行的任务添加到任务队列中，任务监听器时刻监听任务队列，监听器把监听到的任务按照特定的规则放入任务调度器中。

任务调度器为每个任务分配任务线程以执行该任务。任务线程执行的主要结果是形成资源队列和资源状态表。

资源监听器、资源调度器以及资源线程池在资源执行状态表的配合下的实现对资源的调度和执行，一个任务中的所有资源全部执行完标志着该任务执行结束。

2.4 主要算法及技术特点

1）主要算法

根据任务的制定规则，任务结构是一个树形结构。作业的执行过程本质上是各个作业所对应的资源的执行。在具体算法实现中，根据树的先序遍历算法来确定任务中的作业顺序、任务清单的整体结构以及资源

顺序。所以任务的树形结构在逻辑上可以做如下图所示的扩展：

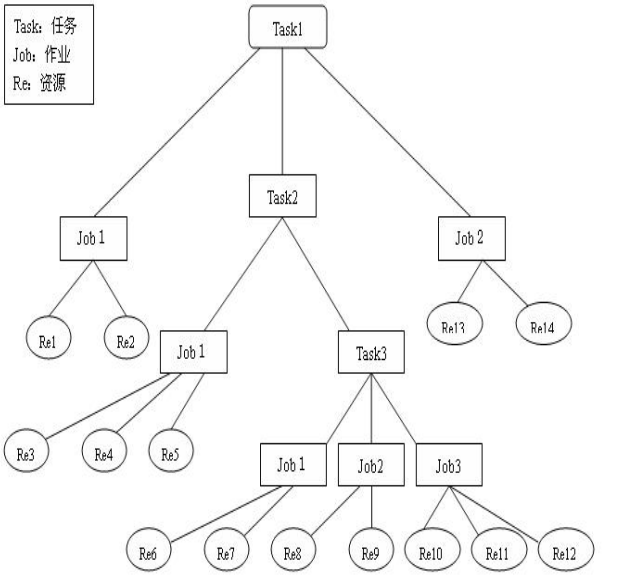


图 5 任务的树形结构图

2） 技术特点

本系统将数据抽取、数据转换、数据加载三部分细化为相互独立的数据操作对象。作业是对象的组合，任务是作业的组合。这种方式能够保证任务制定的多样性，消除了作业之间的依赖性，易于功能扩展和系统维护。

任务调度模块通过任务调度器和资源调度器的合理设计保证了数据操作对象的有序执行。任务线程池和资源线程池提高了调度的高效性。

任务设计模块采用基于 FLEX 的 MVC 设计模式，该设计模式把应用程序的输入、处理、和输出分开，用户可以在 VIEW 部分轻松的制定、查看、修改任务。采用 Flex 技术使得任务设计模块具有更好的网络交互能力，MXML 和 ActionScript 的结合完成了绚丽高效的任务设计页面。

3 结语

本文首先介绍了 ETL 中的数据抽取、数据转换、数据加载原理，并详细介绍了数据交换过程中的元数据概念。在 ETL 原理以及元数据的基础上提出了基于元数据的数据交换系统架构，进一步详细介绍了该数据交换系统的任务设计模型和任务调度模型。最后对任务执行的核心算法以及系统的设计模式进行了介绍。

元数据是 ETL 处理数据的主要依据，所以如何定义和管理元数据是数据交换系统的重点。在本文提出的基于元数据的数据交换系统中，任务设计模型能够将抽取元数据、转换元数据、加载元数据组合到一起形成任务说明书，任务调度模型进而根据任务说明书执行各项任务。 综上，基于元数据的任务设计模型能够简化任务制定流程，消除数据转换前后步骤间的依赖性，易于后期功能的扩展和修改。

本文提出的基于元数据的任务设计模型还有以下扩展：1）在异构数据源的基础上扩展更多的数据转换功能。2）提高制定任务的简易性和便捷性。3）能够在任务制定过程中提供给用户一些关键元数据，例如流程信息，sql 语句，字段说明等等。

参考文献

1 李庆阳,彭宏.面向数据质量的 ETL 框架的设计与实现.计算机工程与设计,2010,9:2057–2060.

2 贾樊星,陈梦东,刘连忠.基于任务的数据交换平台.计算机工程,2008,34(19):61–66.

3 王亚玲,刘迪,曹占峰,等.基于优先级的任务调度模型在数据交换平台中的应用.中国电力,2008,41(4):84–87.

4 杨丽,廉东本.基于 SOA 的数据交换平台的设计.计算机系统应用,2011,20(5):30–33.

5 宋旭东,闫晓岚,刘晓冰,杨莉国.数据仓库 ETL 元模型设计.计算机仿真,2010,9:106–108.

6 王预.数据仓库技术及其设计与开发流程.兰台世界:下半月,2010,9:24–25.

7 刘娜,段智敏.基于 SOA 异构数据交换平台的设计与实现.企业技术开发,2009, 28:30–34.

8 王珊,陈琨.ETL 中基于贪婪算法的任务调度方法研究.微电子学与计算机,2009,26(7):130–134.

9 Ferigato C, Masera M. Design of a Platform for Information Exchange on Protection of Critical Infrastructures. Critical Information Infrastructures Security, 2008,51–41:337–338.

10 Alessandro Lapadula, Rosario Pugliese and Francesco Tiezzi. Regulating Data Exchange in Service Oriented Applications. International Symposium on Fundamentals of Software Engineering, 2007,4767:223–239.