数据仓库中 ETL 技术的研究

张　宁[1]　贾自艳[2]　史忠植[2]
[1]（中国科技大学研究生院计算机学部，北京 100039）
[2]（中科院计算技术研究所智能信息处理重点实验室，北京 100080）
E-mail：lily_zn@yahoo.com.cn

摘　要　作为数据仓库的关键部件，支持数据抽取、清洗、转换和装载的工具集对任何数据仓库工程都是一个必不可少的成功因素。该文简单介绍了 ETL 技术，包括 ETL 的相关概念、ETL 在数据仓库中的功能和重要地位以及现有的研究成果，然后重点介绍了 ETL 的具体设计和实现方法。

关键词　数据仓库　ETL　数据抽取　数据转换　数据清洗　数据装载

文章编号 1002-8331-(2002)24-0213-04　**文献标识码** A　**中图分类号** TP311.13

Research on Technology of ETL in Data Warehouse

Zhang Ning[1]　Jia Ziyan[2]　Shi Zhongzhi[2]
[1](Graduate College, University of Science and Technology of China, Beijing 100039)
[2](The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract: As a key component of the data warehouse architecture, the set of tools that supports data extraction, cleansing, transformation and loading represents a critical success factor for any data warehouse project. This paper briefly introduces ETL technology, including the concepts related with ETL, ETL's function and important location in data warehouse architecture and existed research products. Then it emphasizes the design and implementation method of ETL.

Keywords: Data warehouse, ETL, Data Extract, Data Transform, Data Cleansing, Data Loading

1　引言

近年来，随着数据库技术的应用和发展，人们尝试对数据库中数据进行再加工，形成一个综合的、面向分析的环境，以更好地支持决策分析，从而形成了数据仓库（DATA WAREHOUSE，简称 DW）。业界公认的数据仓库概念创始人 W.H. INMON 在《建立数据仓库》一书中对数据仓库做了精确的定义：数据仓库是面向主题的、集成的、不可更新的（稳定性）、随时间不断变化（不同时间）的数据集合。数据仓库与传统数据库不同，它并非是一个仅仅存储数据的简单数据库，它实际上是一个以大型数据管理信息系统为基础的、附加在这个数据库系统之上的、存储了从企业所有业务数据库中获取的综合数据的，并能利用这些综合数据为用户提供经过处理后的有用信息的应用系统。数据仓库的重点与要求是能够准确、安全、可靠地从数据库中取出数据，经过加工转换成有规律信息之后，再供管理人员进行分析。

在数据仓库构筑中，传统上作业量最大、日常运行中问题最多的是从业务数据库向数据仓库抽取、变换、集成数据的作业。原因是为了从各种不同种类和形式的业务应用中抽取、变换、集成数据，并将其存储到数据仓库，要求对数据的质量进行维护和管理。ETL 工具就是在数据的抽取处理之后，进行数据的"净化提炼"处理。所谓数据的"净化提炼"就是对从多个不同业务数据库所抽取的数据，进行数据项名称的统一、位数的统一、编码的统一和形式的统一，消除重复数据。现在 ETL 工具的功能越来越高级。它具有支持数据的"净化提炼"功能、数据加工功能和自动运行功能（包括处理过程的监控、调度和外部批处理作业的启动等），支持多种数据源，能自动实现数据抽取。

2　ETL 简介

2.1　ETL 的相关概念

数据采集（ETL），即数据抽取（Extract）、转换（Transform）、清洗（Cleansing）、装载（Loading）的过程，是构建数据仓库的重要环节。用户从数据源抽取出所需的数据，经过数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。

具体来讲，数据抽取（Data extract）：是数据源接口，包括原始数据接口和外部数据接口，源数据接口从业务系统中抽取数据，为数据仓库输入数据。数据转换（Data transform）：数据转化包含对来自多个生产系统的数据源的处理，保证数据按要求装入数据仓库。数据清洗（Data cleansing）：一个确保数据集中的所有数值是一致的和被正确记录的处理过程。数据装载（Data loading）：数据装载部件负责将数据按照物理数据模型定义的表结构装入数据仓库。这些步骤包括清空数据域、填充空格、有效性检查等。

2.2　ETL 的必要性

在企业管理中，经理人员总是希望能随时随地访问到任何他们需要的信息，这就要求有一个体系结构来容纳各种格式的内部数据和外部数据，例如经营数据、历史数据、现行数据以及

基金项目：国家自然科学基金（编号：60173017，90104021）；北京自然科学基金（编号：4011003）
作者简介：张宁，在职研究生，主要研究方向为人工智能和知识发现。

来自 Internet 服务提供商（ISP）的数据，此外还应该包含易于访问的元数据。这些源数据因为来源不同，具有大量、分散和不清洁的特点，不能为数据仓库直接使用，而对所有数据的分析、采掘活动也必须建立在一个数据清洁、结构良好的数据仓库的基础之上。这就必须由 ETL 来实现，它是数据仓库获得高质量数据的环节。总的来说，ETL 的必要性体现在以下几方面：

（1）解决数据分散问题

对于企业来说，数据主要有四个方面的来源：客户信息、客户行为、生产系统和其它相关数据。也就是说大量的数据分散在不同的 OLAP 系统中，客户的信息也分散在不同的系统中。如果只对某个系统的数据进行分析，以作为决策支持的依据，显然有信息不全面、分析不准确的缺点。ETL 可以解决这些问题。根据企业决策的需求，数据仓库将决策分析用的数据集中在一起。

（2）解决数据不清洁问题

分散的数据也带来了数据不清洁的问题。同一个客户的信息在不同系统中的数据不一致，而且有些数据可能是不真实的。另外，分散的业务系统中的数据是面向业务的，而不是面向决策的。ETL 模块解决了数据不清洁问题，并将数据转换为决策分析所需要的类型。通过对分散数据的集中、清洁和转换，数据仓库中存储着清洁、一致、全面和面向决策的数据。这些数据为了方便用户的分析与查询，设计成多维模型结构。

（3）方便企业各部门构筑数据中心

数据仓库是面向整个企业的数据应用，而针对各个部门的信息应用是构筑数据中心。数据中心的数据是按部门从数据仓库中抽取，并进行加工处理。数据中心构筑工具，就是提供从数据仓库自动进行数据的抽出、变换功能，具有 ETL 功能，可以大幅提高运行效率。

2.3 ETL 在数据仓库中的重要位置

ETL 的重要性可从其在数据仓库中所处的位置看出。那么，数据仓库都有哪些关键技术和组成部分呢？与关系数据库不同，数据仓库并没有严格的数学理论基础，它更偏向于工程。由于数据仓库的这种工程性，因而在技术上可以根据它的工作过程分为：数据的采集、存储和管理、数据的表现以及数据仓库的设计的技术咨询四个方面。图 1 即是数据仓库的框架图。

从图 1 可看出数据的采集是数据进入数据仓库的入口。由于数据仓库是一个独立的数据环境，它需要通过抽取过程将数据从联机事务处理系统、外部数据源、脱机的数据存储介质中导入到数据仓库。数据采集在技术上主要涉及互连、复制、增量、转换、调度和监控等几个方面。数据仓库的数据并不要求与联机事务处理系统保持实时的同步，因此数据抽取可以定时进行，但多个采集操作执行的时间、相互的顺序、成败对数据仓库中信息的有效性则至关重要。

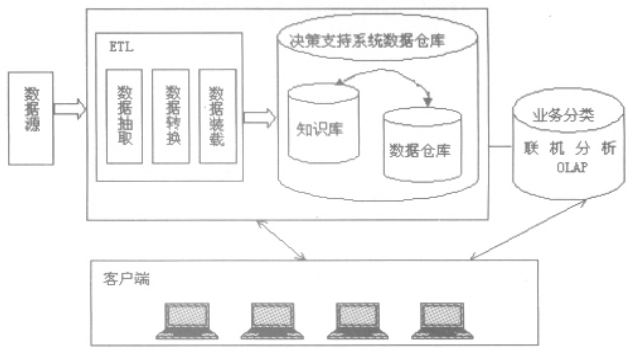

图 1　数据仓库整体框架图

总之，ETL 是数据仓库的一个主要的方面，它从运作的资源中抽取、转换和装载数据到数据仓库或者数据集市中，用于以后的分析。

2.4 ETL 的发展现状

至今为止，数据仓库的实用化已走过了近十年的历程，应用领域遍及通信、证券、银行、税务、保险等行业。而各大数据库厂商纷纷宣布产品支持数据仓库并提出一整套用以建立和使用数据仓库的产品，比如 INFORMIXGONGSIDE 公司的数据仓库解决方案，ORACLE 公司的数据仓库解决方案，Microsoft 公司的数据仓库解决方案等等。各厂商的产品具有不同的特点，它们的 ETL 工具也各有其优势和不足。表 1 是对几个主要数据库厂商的 ETL 工具的简单比较。

在众多的产品中，普遍认为 DTS 是系统最易用、扩展性最好、编程效率最高的数据抽取工具。而在我国，对 ETL 的研究开发甚少，还没有一个成型的完善的 ETL 工具应用于数据仓库的系统中。

3 ETL 的具体设计

实际上，无论何时，在数据仓库中为了存储、检索或者表达的目的，数据需要从一种形式转换到另一种形式时，转换就需要被考虑，而且数据的抽取、转换和装载都可以刻画成转换操

表 1　ETL 工具的比较

数据库厂商	ETL 工具	优点	缺点
IBM	Visual Warehousing	数据源广泛，在大数据量的抽取中具有速度优势；提供编程接口和调用外部程序的功能，按计划自动执行数据抽取；提供对 Cube 处理的功能；提供 Agent 把数据抽取分布到工作站、小型机、大型机等各种平台	界面不够友好，在处理复杂的数据源时面临较多的工作量
Oracle	Oracle Warehouse Builder	提供功能包括：模型构造和设计；数据提取、移动和装载；元数据管理；分析工具的整合；数据仓库管理；具有开放可延伸的框架	不能把数据抽取扩充到 Unix 工作站、小型机、大型机，流程繁琐，不易使用
Microsoft	DTS	从广泛的数据源抽取数据，提供市场上最有效的编程方式，以及工作流的任务处理方式；提供调用外部程序的功能和强大、丰富的被外部程序调用的对象库；按计划自动执行数据抽取	
Informax	ArdentDataStage	提供工作流的方式，可以实现内部编程	数据抽取功能的处理方式过于简单，程序的高效性和准确性方面的保证措施太少
CA Platinum	Inforbump	抽取速度较快，非工作流的工作方式	用户面临过大工作量，程序高效性和准确性方面的保证措施太少

作。因此,对于数据仓库来说,转换是一个核心环节。

3.1 ETL 模块的组织结构

由于对数据的各种操作均可以刻画为转换操作,所以在 ETL 的设计过程中,笔者将对数据的转换操作定义为一系列的转换活动,ETL 模块的组织结构可用图 2 表示。

在数据的转换过程中,一个转换将源对象利用一种转换规则转换成一组目标对象。源对象和目标对象都是数据对象集中的元素。数据对象集中的元素能够是任何类型的数据元素,但是典型的是表、列或表示在内存中暂存对象的模型元素。数据对象集可以是不同转换的源和目标。具体来说,在同一个逻辑单元中,一个给定的数据对象集可以是一个转换的目标和一个或多个转换的源。经常,转换也可以产生一系列的临时数据。那些必须一起执行的转换被归类到相应的转换任务中。在执行时,转换步骤(transformation steps)被用来协调转换任务之间执行的控制流。每个转换步骤执行单一的转换任务。转换步骤可以进一步被归类成转换活动(transformation activities)。在每个转换活动中,它的转换步骤的执行序列可以通过使用步骤优先依赖(step precedence dependency)或者优先约束(precedence constraint)被定义为确定性的,或者通过使用数据依赖(data dependency)被定义为不确定的。

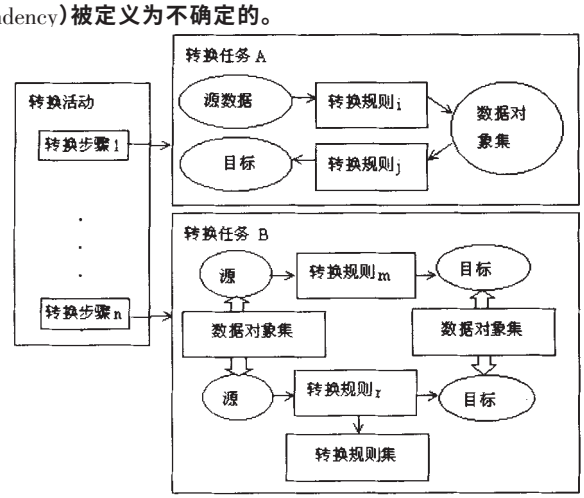

图 2 ETL 组织结构

3.2 ETL 相关元数据

元数据的设计合理与否直接影响 ETL 系统的性能,根据 ETL 组织结构,其元数据可定义为图 3。

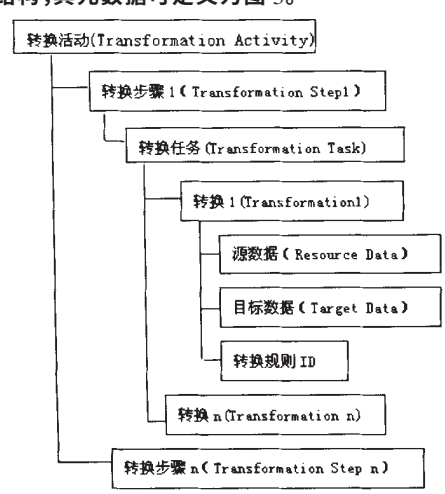

图 3 ETL 元数据

3.3 设计 ETL 工具模块需要考虑的问题

通过分析和比较各个数据库厂商的 ETL 工具,并根据商务的需求和条件,在 ETL 工具的设计和开发时通常需要考虑:系统数据的知识基础是否充分;数据抽取/装载操作是一次完成,还是不断/反复操作;关注的焦点是数据内涵的质量,还是补偿式的传递数据;数据质量问题是特殊性的(例如是针对客户或销售额),还是普遍性的;使用现成集成套件工具,还是自己有针对性的开发;数据抽取/转换是集中管理,还是分布管理;数据抽取/转换是通过参数控制,还是编程控制。

3.4 ETL 相关的操作类型

ETL 工具功能的强大在很大程度上取决于转换规则集的健全,通过分析项目数据和项目的需求,发现数据源和数据仓库的数据都是采用关系数据库来存放的,因此现在制定的转换规则都是针对关系数据库来定义的。同时,这里的规则集设计采用开放的方式进行管理,也就是说用户可以根据自己的需求添加转换算法。现在定义的规则集如下,用户如果需要复杂的数据转换,可以通过一系列的转换组合来达到目的,其实这个转换组合就是前面讲到的一个转换活动。总结 ETL 常见的操作有以下几种:

(1)数据的有效性检查

为避免数据冗余,要认识到数据装入数据仓库之前,应该对数据进行有效性检查,这是很重要的。如果没有进行数据的有效性检查,就有可能破坏依赖于数据仓库的商务分析的完整性,帮助检查数据的有效性的最好方法是源系统专家。源系统专家包括具有技术专业知识和非技术知识的人士。检查数据仓库中数据的有效性是一个非常耗时但必不可少的过程。建议该过程应高度自动化。SQL Server7 中有许多内置功能,可自动进行数据有效性检查。

(2)清除数据

有效性检查是决定是否符合给定标准的过程。标准是依赖于安装的,为某个站点开发和执行的标准可能在其他地方毫无意义。如果数据不在给定的界限之内,它就成为这里称作 scrubbing(清除)过程的对象。清除数据包括对那些在给定范围之外的数据采取纠正措施。

(3)数据格式化

数据仓库中的数据来自于多种业务数据源,这些数据源可能是在不同的硬件平台上,使用不同的操作系统,因而数据以不同的格式存在不同的数据库中。如何向数据仓库中加载这些数量大、种类多的数据,已成为建立数据仓库所面临的一个关键问题。所以,在数据迁移的过程中,通常需要将操作数据转换成另一种格式以更加适用于数据仓库设计,这也就是数据的格式化的过程。提取处理是数据仓库成功的关键。在提取过程中,数据会被格式化,并分发给需要从操作环境中共享数据的资源。

(4)数据转换

在大多数情况下,转换是将数据汇总,以使它更有意义。在转换结构中,确保能找出一种最好的方法保证数据从传统的数据存储器到数据仓库的同步。同步结构应当把重点放在转换语言的标准化、数据移动平台、通信策略和支持策略方面。数据仓库与操作数据存储器之间的同步过程能够采取不同的结构。除寻找自动化转换操作的工具之外,还应估计数据转换的复杂性。大多数传统的数据存储方法缺乏标准,常常有些不规则的东西让开发员摸不着头脑。工具正在不断改进以有助于转换过程的自动化,包括复杂问题,如掩匿的数据、传统标准的缺乏及

不统一的关键数据。元数据存储的工作是定义和解释数据资源和数据标准。因此,在操作数据上执行的转换过程应该用元数据存储中定义的标准数据格式放置数据。

该文从转换的难易程度总结定义了数据变换的几个基本类型,每一类都有自己的特点和表现形式:

()简单变换

简单变换是所有数据变换的基本构成单元。顾名思义,它是数据变换中最简单的形式,这些变换一次改变一个数据属性而不考虑该属性的背景或与它相关的其他信息。它包括如下的转换类型:数据类型转换,日期/时间格式的转换,字段解码。

()清洁和刷洗

目的是为了保证前后一致地格式化和使用某一字段或相关的字段群。清洁和刷洗是两个可互换的术语,指的是比简单变换更复杂的一种数据变换。在这种变换中,要检查的是字段或字段组的实际内容而不仅是存储格式。一种清洁是检查数据字段中的有效值。这可以通过范围检验、枚举清单和相关检验来完成。数据刷洗的另一主要类型是重新格式化某些类型的数据,这种方法适用于可以用许多不同方式存储在不同数据来源中的信息,必须在数据仓库中把这类信息转换成一种统一的表示方式。

()集成

集成是将业务数据从一个或几个来源中取出,并逐字段地将数据映射到数据仓库的新数据结构上。要把从全然不同的数据源中得到的业务数据结合在一起,真正的困难在于将它们集成为一个紧密结合的数据模型。这是因为数据必须从多个数据源中提取出来,并结合成为一个新的实体。这些数据来源往往遵守不同的业务规则,在生成新数据时,必须考虑到这一差异。数据的集成可大致分为两类:字段水平的简单映射和复杂集成。字段水平的简单映射在必须执行的数据变换总量中占去了大部分。这种映射的定义是指数据中的一个字段被转移到目标数据字段中的过程。在这过程中,这个字段可以利用前面讨论过的任何一种简单变换进行变换,它可以被刷洗或重新格式化。在一般的数据仓库中,数据转移和集成中的 10%~20%要比从源字段到目标字段的简单移动复杂一些。为了将源数据变换为目标数据,这些复杂集成必须做更多的分析,包括通用标识符问题、目标元素的多个来源、数据丢失问题、衍生数据/计算数据等等。

()聚集和概括

大多数数据仓库都要用到数据的某种聚集和概括。这通常有助于将某一实体的实例数目减少到易于驾驭的水平,也有助于预先计算出广泛应用的概括数字,以使每个查询不必计算它们。概括是指按照一个或几个业务维将相近的数值加在一起。聚集指将不同业务元素加在一起或为一个公共总数。在数据仓库中它们是以相同的方式进行的。聚集还可以去除数据仓库中的过时细节。在许多情况下,数据在一定时期内要以很具体的方式存放着,一旦数据到了某一时限,对所有这些细节的需求就大大减弱了。此时,这些非常具体的数据应该传送到离线存储器或近线存储器中,而数据的概括形式则可以存放在数据仓库中。目前可以得到的数据刷洗工具中,许多都已内置了概括功能,尤其是在时间维上进行聚集的功能。当然,不管如何做到这一点,重要的是用户能够轻松地访问元数据。

4 结束语

数据仓库必须以大量的、日积月累的数据为基础,必须以运行的、不断更新的数据库为主要源泉。因此,通过 ETL 系统建好、维护好基础数据库是创建数据仓库的前提。一个为企业决策者提供快捷、准确、全面的有价值信息的数据仓库系统必然以高效的 ETL 作为基础。(收稿日期:2002 年 7 月)

参考文献

1. 美)Harjinder S GILL 著.王仲谋,刘书舟译.数据仓库-客户/服务器计算指南[M].北京:清华大学出版社
2.王珊等编著.数据仓库技术与联机分析处理[M].科学出版社,1999
3.张澜,康增培.数据仓库白皮书概念篇.http://www.ccidnet.com/tech/paper/2001/03/02/58_1770.html
4.张伟.各家数据仓库产品的评估.2001
5.Michael F Jennings.Strategies for Custom Data Warehouse ETL Processing.2000
6.Common Warehouse Metamodel (CMW)Specification.http://www.cwm-forum.org/spec.htm,2001

(上接 201 页)

图 2 视频帧检索结果

6 结束语

以直方图为工具实现对压缩视频基于内容检索的方法原理简捷、实现简便。但如果能与其它方法结合起来使用效果将更佳。在这种方法里还有很多问题需进一步研究:直方图的优化存储、相关镜头的聚类、镜头可变数目代表帧的提取和用户查询接口的设计等,同时如何减少计算量、缩短处理时间也需要重点考虑。(收稿日期:2001 年 10 月)

参考文献

1.rman F,Hsu A,Chiu M Y.Image processing on compressed video data for large video databases[J].ACM Multimedia,1993:267~272
2.aitao Jiang,AbdelSalam Helal.Scene change detection techniques for video databases systems[J].Multimedia Systems,1998;6):186~195
3.Stephen W Smoliar,HongJiang Zhang.Content_based Video Indexing and Retreval[J].Multimedia,1994
4.曹莉华,胡晓峰,李国辉.基于内容检索中的视频处理技术研究[J].计算机工程与应用,1998,34(6):31~41