

一种钻井数据仓库 ETL 系统的设计

梁美红¹, 张男楠², 李 建¹, 伍 东¹, 胡永泉¹, 杨 静¹

(1. 西南石油大学 计算机科学学院, 四川 成都 610500;

2. 西南油气田信息中心, 四川 成都 610500)

摘 要:随着企业信息化的不断发展,石油单位将数据整合纳入到重点规划中。面对分散在各处的异构数据源进行数据整合并非易事,首先仅靠手工进行脏数据的清洗不但费时费力,质量也难以保证;其次,数据的定期更新也存在困难。ETL 系统为数据整合提供了令人满意的解决方案。它可以完成数据抽取、清洗、转换、装载等任务,满足了用户对异构数据源进行整合的需求,也实现了数据的后期更新。笔者对钻井数据仓库 ETL 系统的设计提出了一种基于元数据的 ETL 体系结构,并重点设计了数据准备区、ETL 管理模块、任务管理模块和元数据管理模块。该工具已在中海油田化学技术专家支持系统中得到应用。

关键词:ETL;数据仓库;元数据

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2010)03-0250-04

Design of ETL System for Drilling Data Warehouse

LIANG Mei-hong¹, ZHANG Nan-nan², LI Jian¹, WU Dong¹, HU Yong-quan¹, YANG Jing¹

(1. School of Computer Science, Southwest Petroleum University of China, Chengdu 610500, China;

2. Information Center of Southwest Oil and Gas Field Company, Chengdu 610500, China)

Abstract: With the development of information, oil unit makes a plan for data integration. It is not easy to process data integration from different data source. At first, it takes much time and money to clean dirty data by hand, and it is difficult to guarantee quality; Second, it is hard to update data on a regular time. ETL system for data integration provides a satisfying solution. It can complete many tasks, such as data extraction, cleaning, transforming and loading etc, meet the needs of heterogeneous data source integration and also realize to update data. The article designs ETL system for drilling data warehouse which is based on a metadata ETL architecture, and focus on the design of the data preparation area, ETL management module, the task management module and the metadata management module. ETL tool has been used in COSL Chemical Technology Expert Support System.

Key words: ETL; data warehouse; metadata

0 引 言

信息技术的不断推广使用,将企业带入了一个信息爆炸的时代。每日、每时、每刻都有潮水般的信息出现在管理者的面前,等待管理者去处理、去使用。为满足管理人员的决策分析需要,在数据库基础上产生了能够满足决策分析所需要的数据环境——数据仓库。而 ETL 是数据仓库中的非常重要的一环。它按照统一的规则集成数据并提高数据的价值,是负责完成数据从数据源向目标数据仓库转化的过程,是实施数据

仓库的重要步骤。在整个项目中 ETL 规则设计和实施则是工作量最大的,其工作量占整个项目的一半以上,这是国内外从众多实践中得到的普遍共识。文章根据油田实际项目,设计满足钻井业务部门级数据仓库系统应用需求的 ETL 系统。

1 ETL 简介

ETL 是数据抽取(Extract)、转换(Transform)、装载(Load)的过程。可以简单描述为:用户从数据源抽取所需的数据,经过数据的清洗、转换,最终按照预先定义好的数据格式,将数据加载到数据仓库中去。整个过程一般包括三个部分:

数据抽取:数据抽取是从数据源获取符合需要的数据的过程。即是从不同的网络、不同的操作平台、不同的数据库、不同的应用中采用不同形式的接口进行

收稿日期:2009-06-30;修回日期:2009-09-19

基金项目:国家重大专项项目(2008ZX05021-006)

作者简介:梁美红(1983-),女,四川南充人,硕士研究生,研究方向为计算机应用技术;李 建,教授,硕士生导师,研究方向为数据仓库、数据挖掘和建模仿真等。

数据的原始抽取。这个步骤是把数据放入临时的中间介等待后续步骤的清洗与转换。

数据转换:数据转换就是对数据的一种处理,即按照增强或简化它的含义的方式将源系统的数据形式转换为数据仓库的数据形式^[2]。简而言之,即是在临时中间介中按照预先设定好规则对异构数据库中抽取上来的数据进行清洗、转换、拆分、汇总。

数据装载:将经过清洗和转换的数据加载到数据仓库中指定的主题和细节库中。

2 ETL 系统的分析与设计

根据需求,本 ETL 系统是建立钻井业务部门级数据仓库,它是中海油田化学技术专家支持系统项目的一个组成部分。该钻井数据仓库主要对需要的源数据进行抽取、清洗、转换和加载处理。同时要求所涉及到的数据能通过本 ETL 系统准确、快速地加载到钻井数据仓库中。根据其专用性,它要实现如下功能:

- * 实现数据的定时自动抽取、转换和加载。
- * 能够根据用户要求清洗数据,脏数据单独处理。
- * 对部分含有特殊符号的字段数据要做字符转义处理。
- * 对有些目标数据,源表没有相应数据的需要程序自动给其赋默认值。
- * 有比较完备的日志功能,能捕获并记录转换过程中发生的异常。

另外,本 ETL 系统只负责把源数据加载到数据仓库中的 P 表,P 表位于数据准备区,它在数据仓库中是独立的,不与其它的表产生任何依赖关系。且 P 表可被认为是数据仓库的临时表,也是本 ETL 系统的目标表。当本 ETL 系统把数据加载到它里面后,会触发数据仓库中的存储过程对 P 表数据进行解析,然后根据元数据配置将相关数据装载到维表和事实表中。

经过需求分析,为了满足分布式的数据整合,本 ETL 系统需采用三层体系结构,对于数据的转换需采用基于元数据驱动的数据转换方法,这样使得数据的集成更加灵活,满足复杂的数据转换和数据清洗,同时也更容易维护。根据系统设计的思路,本 ETL 系统包括数据准备区、元数据管理模块、任务管理模块和 ETL 管理模块。系统体系结构如图 1 所示。

底层是中海油服中心数据库,它提供 ETL 系统的数据来源。中间层是数据准备区、元数据管理模块、ETL 管理模块和任务管理模块,它是数据集成的主要环节,所有的转换都在此进行。顶层钻井数据仓库是数据的目的地。数据准备区临时缓存 ETL 模块从中

中海油服中心数据库抽取出来的数据;元数据管理模块主要是对源数据库和目标数据库的元数据以及数据集成任务中的各种元数据进行获取,并将元数据存储到元数据库;任务管理模块负责任务的调度和将用户的配置信息生成任务并存储到无数据库。ETL 管理模块主要是执行元数据库中存储的任务,把各种数据转换任务解析为数据清洗、转换、加载规则,然后进行数据清洗、转换、加载。

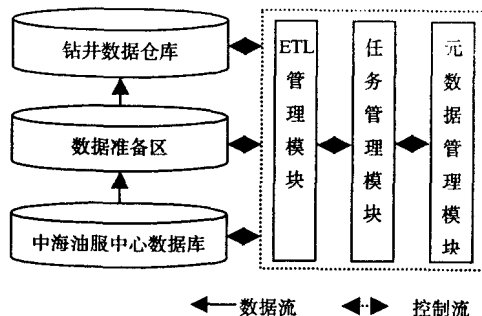


图1 钻井数据库 ETL 系统体系结构

3 ETL 系统重点功能模块设计

3.1 元数据管理模块

元数据驱动整个 ETL 过程,对元数据的管理在整个系统的数据处理过程中都是非常重要的。元数据管理模块主要实现了源、目标数据库元数据的获取、元数据的存储和元数据的查询等功能,这里主要讲述元数据的设计和元数据的获取,其中元数据的设计和内容包括确定源、目标数据库的元数据,日志元数据,脏数据和任务元数据。在设计元数据库和确定元数据的过程中,参照了国际元数据管理标准,根据油田实际的需求情况进行设计。元数据主要包括以下几方面的内容^[1-6]:

确定源、目标数据库的元数据:这类元数据包括数据库信息、数据库表信息、表字段属性。其中数据库信息包含服务器 IP 地址、服务器端口、数据库名称、用户名、密码等信息;数据库表信息包括表名、表之间的关系、表的描述以及表所包含的字段等信息;表字段属性信息包含字段名、字段取值的数据类型、字段的主外键、字段长度和精度等信息。

日志元数据:在装载任务完成后,需要日志元数据对转换任务中转换的条数、成功和失败条数,数据转换异常、数据加载异常等相关信息进行记录。日志元数据包括任务名、开始时间、结束时间、共转换条数、成功条数、失败条数、脏数据记录条数、异常详细信息等。

脏数据:用户需要分析在 ETL 过程中产生的一些不符合用户自定义的清洗规则的脏数据。用户可以判断这些脏数据是否还有用,或者对用户自定义的清洗

规则进行修改,以便使清洗规则更加贴近实际需求。

任务元数据:描述特定数据源到目标数据库的映射以及规则配置信息,它由用户通过 ETL 系统任务管理模块进行配置,它是保障 ETL 系统顺利运行的最重要的元数据信息。在进行数据集成时,系统调用相应任务元数据,并根据其提供的信息来进行数据的抽取、清洗、转换和加载。任务元数据包括源字段信息、目标字段信息、运行时间和调度时间、转换规则信息、清洗规则信息、任务类型等。

本系统涉及的源和目标数据库都是以 ORACLE 为平台的,所以在表信息元数据获取方面采取直接在程序里用 SQL 语句访问数据库数据字典的方式。

3.2 ETL 管理模块

ETL 管理模块是 ETL 过程的核心,它解析元数据配置信息,并根据要求完成数据从源数据库到目标数据库的抽取、转换和加载。ETL 管理模块的程序流程如图 2 所示。

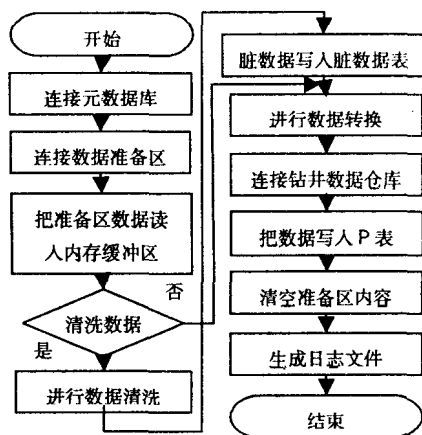


图 2 ETL 管理模块程序流程图

ETL 管理模块主要执行元数据库中存储的数据转换任务,从元数据库中取出数据转换任务并把它解析为数据清洗、转换、加载规则,然后进行数据的清洗、转换、加载^[7]。首先根据任务信息的源字段和目标字段的映射关系动态地生成 SELECT 语句,并把数据准备区中数据字段内容读取到内存缓冲区里,等待进行下一步的处理;然后根据清洗规则对相关数据进行清洗,再根据转换规则进行数据转换;最后根据加载规则生成动态的 INSERT 语句把数据加载到钻井数据仓库目标 P 表里,当本次数据转换加载任务全部执行完后,程序就清空数据准备区的内容。在整个处理过程中,程序将把不符合清洗规则的数据插入到元数据库中的脏数据表中,而日志文件也将记录转换、加载失败的数据信息。

3.3 数据准备区

数据准备区的设计对 ETL 系统的实现起着至关

重要的作用^[2],在数据转换过程中,首先将抽取的数据存储在数据准备区当中,数据转换就不需要再访问源数据库而直接利用数据准备区的数据进行转换操作。然后再把转换后的数据临时写入另一个数据准备区中,等待 ETL 管理模块把数据加载到目标数据仓库中。数据准备区是按照数据采集需求建立的小型关系数据库。选择小型关系数据库的原因有二点:一是程序直接连接钻井数据仓库分析设计库系统,它的运行效率较高;二是中海油服中心库和钻井数据仓库都是关系数据库,方便管理。

关于这个小型关系数据库中表的表结构形式有二种:一种是针对表数据来源于源数据库的多张表,设计它们的表字段和对应的数据仓库快照表相同,并且两个数据准备区表的字段属性是和相应源表字段属性完全一样的;另一种是针对数据来源只涉及一个表,设计这些表的表结构、字段属性和相应源数据库表的表结构、字段属性相同。这样做的目的是要尽可能地减少在从源数据库抽取数据时要进行复杂的数据清洗、转换而对它的不良影响。另外,数据准备区表中存放的是日常的增量数据,当每一次它里面的数据被成功地转换完之后,这些表内容就会被清空。

3.4 任务管理模块

任务管理模块记录各种配置信息,它贯穿于整个 ETL 过程的各个模块^[8]。它实现了两个功能:

一是任务的配置,任务的配置在定义 ETL 任务之前,程序先扫描源和目标数据库,获得相关元数据并存储在元数据库中,之后用户根据这些元数据定义 ETL 任务,并根据优先级把任务存储到元数据库中。

二是任务的调度及管理,配置好的任务存储在元数据库中,以便将来实现定时运行,以此来追加日常增量的数据。任务调度的策略是由时间机制触发的,根据油田需求,可有立即和设定默认的调度时间两种策略。当运行任务的时间到时,触发机制首先触发任务管理模块调度 ETL 处理模块运行数据追加任务,把上一次数据抽取时间到现在抽取时间之间的新增变化数据,抽取到数据准备区。另外,可以根据需要对已有的配置信息进行删除,也可更新已有的配置,并保存在元数据库中,以便调用。

本 ETL 系统结合实际情况,在保持专用的简单易操作性的同时也兼顾了日后需求发生变化后进行处理的灵活性。

任务管理模块的程序流程如图 3 所示。

4 结束语

文章结合中海油田化学专家支持系统的实际情况

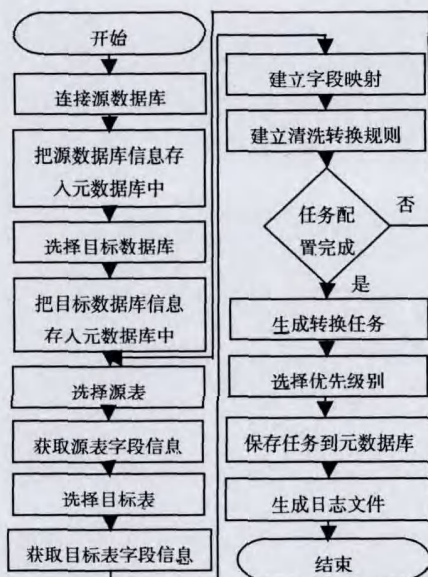


图3 任务管理模块程序流程图

设计了基于元数据的三层体系结构钻井数据仓库 ETL 系统,实现了数据源和目标数据库的分离,使得数据集更加灵活,可靠性更高。

增加的数据准备区可以大大地减小 ETL 过程对数据源系统的影响,使用基于元数据的数据转换方法实现了把钻井数据源数据导入到钻井数据仓库,并且所有的元数据都存储在元数据库中,提高了程序执行

(上接第 127 页)

图 3a 中,参数 σ 为 1.1,阈值为 0.8。在图 3b 和图 3c 中,阈值分别为 0.6 和 0.5。从图中可以看出,使用 3×3 的窗口得到的图像显示出了太多的细节,其中包括了一些虚假边缘,因此将不利于图像的分割。然而由 5×5 的窗口得到的图像的效果与 Canny 算子相当,并且保持了较高的细节保存。

5 结束语

提出了一种新的边缘检测算子,即用支持向量机对图像的像素进行分类来判别边缘。这种算子减少了支持向量的数目从而降低了执行的时间。同时,可以通过改变窗口的尺寸来改善性能,用较大的窗口得到的结果与标准的 Canny 算子相当。

由于支持向量机在图像边缘检测的过程中具有更大的灵活性,因此可以在训练样本中增加一些信息。比如,可以在图像中加入高斯噪声来改善噪声图像的边缘检测。这是以后要做的工作。

参考文献:

- [1] 徐彤阳,姚跃华,朱志勇.一种基于支持向量机的图像边缘检测方法[J].微机发展(现更名:计算机技术与发展),

的效率,有利于元数据的管理。该系统在实现其专用性目的的同时也具有一定的灵活性,有利于程序的维护和软件的二次开发。

参考文献:

- [1] 陈京民.数据仓库与数据挖掘技术[M].北京:电子工业出版社,2007.
- [2] 伍东,李建.钻井数据仓库 ETL 工具的研究与实现[DB/OL].2007-07.CNKI 中国优秀硕士学位论文全文数据库.
- [3] 郑洪源,周良.基于 CWM 的标准 ETL 的设计与实现.吉林大学学报:自然科学版,2006,24(1):50-54.
- [4] 王立刚,刘文煌.构造数据仓库系统的元数据[J].计算机工程与应用,2001,37(16):94-96.
- [5] 张宁,贾自艳,史忠植.数据仓库中 ETL 的研究[J].计算机工程与应用,2002(24):213-216.
- [6] Dyche J. Warehouse D, Metadata and Middleware[J]. EAI Journal,2000(9):71-76.
- [7] Vassiliadis P, Simitsis A, Skiadopoulos S. On the Logical Modeling of ETL Processes[C]//Pidduck A B. CAISE 2002. [s.l.]:[s.n.],2002:782-786.
- [8] Vassiliadis P, Vagena Z, Skiadopoulos S, et al. Arktos: Towards modeling, design, control and execution of ETL Processes[J]. Information System,2001,26(8):537-561.

2005,15(1):87-90.

- [2] 梅跃松,杨树兴,莫波.基于 Canny 算子的改进的图像边缘检测方法[J].激光与红外,2006,36(6):501-503.
- [3] 周德龙.图像模糊边缘检测的改进方法[J].中国图像图形学报,2001,23(4):34-48.
- [4] 王玉震,李雷.基于 SVR 的图像增强方法[J].计算机技术与发展,2009,19(1):60-62.
- [5] Konishi S, Yuille A L, Coughlan J M. A statistical approach to multi-scale edge detection[J]. Image Vision Comput, 2003,21(1):37-48.
- [6] Rishi R R, Chaudhuri P. Thresholding in edge detection: a statistical approach[J]. IEEE Tran on Image Processing, 2004,13(7):927-936.
- [7] Eli P. Feature detection algorithm based on a visual system model[J]. Proceedings of the IEEE, 2002,90(1):78-94.
- [8] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2004.
- [9] Gomez - Moreno H, Maldonado - Bascon S, Lopez - Ferreras. Edge detection in noise images using the support vector machine[M]//IWANN, Lecture Notes on Computer Science. Heidelberg:Springer - Verlag, 2001:685-692.
- [10] Chang C C, Lin C J. LIBSVM: Introduction and benchmarks [EB/OL].2001. <http://www.csie.ntu.edu.tw/~cjlin/pap>.