基于 ETL 的数据集成优化研究与实现

王世水，王元元，高应波
（贵州大学 计算机科学与信息学院，贵阳 550025）

摘 要：通过分析数据源的数据量和异构数据库系统环境等情况，提出基于 ETL 技术的异构数据集成优化方案. 对实验验证数据与现有集成方法进行对比和效能评估得出结果：该解决方案对异构数据源的集成效率较高，尤其是针对海量数据效果尤其明显.

关键词：异构数据；ETL；数据集成；xml
中图分类号：TP311 **文献标志码**：A **文章编号**：1671-024X(2013)03-0078-04

Research and realizations of data integration ETL-based optimization

WANG Shi-shui, WANG Yuan-yuan, GAO Ying-bo
(College of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract: By analyzing the amount of data sources to be integrated, the heterogeneous database system optimization solutions based on the integration of heterogeneous data ETL technology are presented. Through the experimental verification of data compared with existing integration methods and performance assessment, it is concluded that the solution for efficient integration of heterogeneous data sources, especially for massive data effect is particularly evident.

Key words: heterogeneous data; ETL; data integration; xml

数据库系统是目前存储、检索信息最为方便高效的手段之一，大多数企业都有自己的数据库系统，并将信息存储在数据库中. 然而在数据库技术迅猛发展的今天，数据库系统的趋势是由小型化向大型化、集中式向分布式方向发展，数据量由少量向海量方向转变. 如用户要获取相应的数据信息，就不得不花费大量的时间和精力从各个系统的海量异构数据中查询提取需要的数据，并对于这些相互孤立的数据做进一步处理，这不仅给终端用户带来极大不便，甚至由于信息的滞后导致经济上的损失. 要解决这些问题，就必须对分散存储在不同网络中的异构数据系统中的数据进行高质高效的数据集成. 多年来，许多研究人员致力于解决此问题. 目前，用于实现异构数据集成的方法除了传统的集成方法[1]外，还出现了一些基于 ETL 数据集成的优化方法[2-4]. 这些方法针对海量数据的集成效率不高，且采用一次性 ETL 处理过程不能消除全部脏数据；在数据的异构性上，只考虑某一方面的系统数据集成过程的具体集成技术，在处理多种异构数据集成方面没有一个综合的、完整的解决方案，导致在系统实现起来存在一定的局限性；另外，集成技术和集成方法上主要集中在数据库管理系统本身所带的工具的和其他的 ETL 工具软件. 这在数据集成效率和数据质量方面很难满足要求. 为此，在传统和现有的基于 ETL 的数据集成方法上进行优化显得非常必要. 本方法重点论述了基于 ETL 集成优化的处理过程，并对海量数据的处理流程做了改进；均衡网络中各数据库服务器的负载并提高系统整体效率；最后通过纯 SQL 语言编写 ETL 处理程序进行验证实现.

1 数据库系统异构性表现形式

在对异构数据进行集成时，首先需要掌握这些异构数据的具体形式，以及针对不同的异构数据采取相应的数据集成策略和相应的技术手段，异构数据库系

收稿日期：2012-12-31 **基金项目**：贵州省科学技术基金项目(黔科合 J 字[2012]2136 号)
第一作者：王世水(1978—)，男，硕士，讲师.
通信作者：高应波(1972—)，男，硕士，副教授. E-mail：csc.ybgao@gzu.edu.cn

统的异构性主要表现在以下几个方面.

(1) 计算机软硬件运行环境的异构. 计算机软硬件环境的异构主要表现在其参与运行的硬件性能、操作系统、网络类型以及数据库管理系统等的异构性.

(2) 数据文件的异构. 在数据系统中,形成数据的文件格式可能各不相同,如 xls、txt 等各类型数据文件. 要实现数据的集成可以通过 2 种途径实现:一是实现数据库和各种文件类型的转换;二是实现数据的透明访问. 通过 DBMS 所提供的数据库转换工具和 API 接口可实现这两点.

(3) 元数据的异构. 元数据[5-6]的异构在数据集成中是比较常见的一种异构类型. 其特点主要表现在语法上的异构和语义上的异构. 语法上的异构通常指数据源和最终形成目的数据之间的命名规则和数据类型上的差异;语义上的异构比较常见的就是字段拆分与合并,字段数据格式以及数据源在按不同的维度构建实体模型时引起的差异.

2　基于 ETL 技术集成优化的解决方案

通过对目前 ETL 数据集成的研究情况和异构数据环境的分析,我们可以根据不同的异构环境采取相应的集成策略. 如果待集成的是海量数据, 那么可以通过与 DBMS 高度兼容的 SQL 语句将其数据导出与目标数据接近的数据文件,通过 SQL 语句将其直接加载到临时数据库或目标数据库进行进一步处理,这样能够大大提高 ETL 的处理速度与效率. 其数据集成模型如图 1 所示.

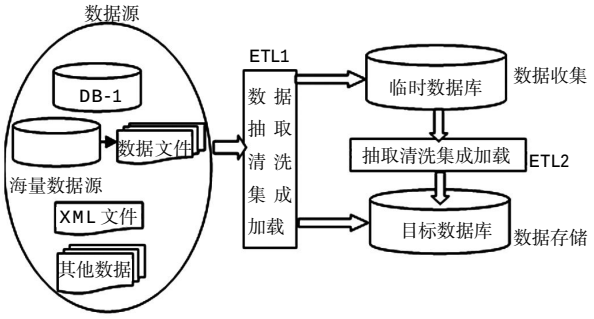

图 1　基于 ETL 集成模型

Fig.1　ETL-based integration model

2.1　ETL 数据集成过程分析

ETL 数据集成即是把数据源数据提取、清洗和转换并最终加载到目标数据库中, 成为联机分析处理、数据挖掘的基础. 目标数据库中的数据必须消除异构数据中的不一致性,它是数据集成的最终目标,并为下游系统服务.

(1) 数据源:在进行系统集成时, 异构的数据源多种多样. 如果数据源数据量庞大,使用 ETL1 对其进行直接数据抽取与加载等操作,其网络负荷和效率将极其低下. 为了保证数据质量和效率, 把源数据库系统中数据尽可能生成与目标数据相应的结构化数据文件,在根据其相似度考虑加载到临时数据库或是目标数据库中. Xml 文件和其他数据文件也可以采取相应的策略进行数据处理并加载到临时数据库或目标数据库中.

(2) ETL1 是对数据源进行的第一次 ETL 处理. 如数据源能够通过一次 ETL 处理就能生成与目标数据一致,则可直接加载到目标数据库;反之则加载到临时数据库进行二次 ETL 处理后再加载到目标数据库.

(3) 临时数据库即是内存数据库, 作为临时存放数据的地方,在后续的数据处理过程中不需要再做读写 I/O 的操作. 该功能是根据每个数据源的配置信息动态的调用客户自定义的一些基于业务规则的转换需求,如类型转换、数据计算等以实现多变的业务需求. 并完成类型转换和数值计算等操作, 最后根据需要来完成数据完整性的检查.

(4) ETL2 是对临时数据库中的数据做进一步清洗、转换集成,最终加载到目标数据库中,使之形成最终统一的正确的数据. 通过这一步的实施, 即可保证目标数据库中的数据不存在脏数据.

(5) 目标数据库是最终数据的集合. 它是存放经过一次 ETL 和二次 ETL 处理后的数据,其数据可以直接来自于数据源或临时数据库,这些数据是已经全部完成集成后的数据集合,提供给下游系统使用.

2.2　数据集成中需要处理的关键问题

(1) 在数据集成时, 必须掌握数据源与目标数据的映射关系,搞清楚数据源与目标数据之间的关系,处理好命名冲突、格式冲突、数据类型冲突、结构冲突以及语法语义等冲突,如表 1 所示的异构冲突现象.

表 1　异构数据基本映射表

Tab.1　Heterogeneous data mapping table

数据源	映射关系	目标数据	备注
Area_nam	字段名映射	Area_id	字段名称异构
varchar	数据类型映射	int	数据类型异构
北京	数据值的映射	010	字段值表示异构
单价:100	数据合并映射	金额:400	语义异构:字段
合并或拆分	数量:4		

(2) XML 文件和数据库的映射结构是嵌套的树形结构,而关系数据库则是简单、平面的二维表结构.

其结构的差异性使得在存储 XML 数据时需要按一定的映射规则进行转换，并使之能够恢复到原 XML 文件. Xml 文档到数据库的转换有两大类映射方法，模型映射和结构映射，其中基于结构的映射方法又可以分为基于 DTD 的结构映射方法和基于 XML Schema 的结构映射方法.

(3) 任务调度[2,7]就是规定对数据源进行处理的先后顺序. 在对数据集成时，涉及的数据很多，有些数据是孤立的，有些数据之间是相互关联的，但由于涉及到一些数据依赖或参照完整性问题，就必须对这些异构数据的相互关系给予理顺，并按一定的先后执行顺序进行 ETL 处理，从而保证数据的正确性. 对起始任务可以采用时间驱动的形式来启动，之后的任务采用事件驱动的形式进行自动执行来完成 ETL的全过程.

3　数据集成与转换规则

3.1　针对临时数据库中的数据转换

就目前大多数数据库系统来说，都自带有相应的工具软件，然而通过其自带工具则效率相对较低. 可以针对不同的数据库系统，采用与各种数据库系统高度兼容的 SQL 语句进行相应的 ETL 操作，这将大大提高系统的数据处理能力. 如临时数据库中存在表 2 的数据，而在目标数据库中存在表 3 的数据，在 ETL2 过程中可以通过 SQL 编写的存储过程来进行实现.

表 2　源数据表

Tab.2　Source data table

地区名称	数量	单价
成都	20	36
北京	10	40
贵阳	12	20

表 3　目标数据表

Tab.3　Target date table

Area_id	Amout
028	720.0
010	400.0
0851	240.0

通过 SQL 编写存储过程实现表 2 到表 3 的目标清洗转换，其实现过程的处理代码如下：

```
Create procedure OldDateToTable
 As
 Inser into area_amout_tab(area_id,amout)
 Select case when 名称='北京' then '010'
      when 名称='成都' then '028'
      else '0851'
      end,
      cast(单价 as numeric(10,2))* 数量
 From 源数据表 [where 条件]
 Go
```

3.2　XML 文件集成转换的实现

XML 文档中的源文件如下.

```
<?xml version="1.0" encoding="UTF-8"?>
 <students>
  <student ID="120901">
    <NAME>wss</NAME>
    <TEL>(0851)3564134</TEL>
  <EMAIL>wss@e163.com</EMAIL>
  </<student>
  <student ID="120902">
    <NAME>zhangsan</NAME>
    <TEL>1364458646</TEL>
  <EMAIL>2324@gzu.edu.cn</EMAIL>
  </<student>
 </students>
```

其数据组成结构如同一棵倒立的树，如图 2 所示.

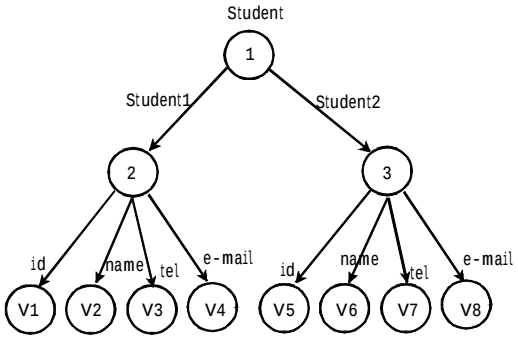

图 2　XML 文档数据结构图

Fig.2　XML data structure diagram

根据其特点，其数据结构模式采用边模型[8-9]的结构定义：Edge(source,ordinal,label,flag,target). 其中：source 域和 target 域表示引出结点和引入结点；ordinal 域表示该边在兄弟边中的位置序号；label 与用来存储边标记（即改边所指向结点的标记名）；flag 属性用来反映边所指向的结点类型.

4　效率性能分析

4.1　实验环境

采用分布式的实验环境，以 5 台机器为例，其中 4 台作为数据源系统. 目标服务器运行的是 sybaseIQ15.1，其他 4 台数据源机器上运行的是数据库管

理系统分别是 sybaseiq15.1、oracle11g、sqlserver2005 及 MySQL,其他软硬件环境一致.

4.2　实验方法与实验数据

分别在各台数据源服务器上随机生成用于测试用户数据 10 万条记录和 1 个 xml 文档用于测试用数据,然后分别对各个数据源服务器上采用传统方法和现有的 ETL 优化方法以及本方案中的数据集成方法测试,通过多次执行集成测试的执行时间,然后取得它们的各自平均集成时间值进行效能分析. 其取得的测试数据如表 4 所示.

表 4　数据集成时间对比表

Tab.4　Data integration time comparison

数据库	传统方法	现有方法	本方法
Sqlserver2005	739.6	132.5	45.3
Sybaseiq15.1	235.4	86.4	25.4
MYSQL	559.6	139.9	60.54
Oracle11g	293.4	65.9	26.5

4.3　验证结果

4.3.1　数据集成时间比较

使用 sybaseIQ15.1 作为目标数据库管理系统,分别对应不同版本的异构源数据库管理系统进行测试,对多次获取其到目标 DBMS 数据库中的数据读取时间、转换时间、清洗时间和装载时间之和的平均值进行比较. 现有优化的 ETL 的数据集成方法和本方法的集成效率差距非常明显,其实验效率结果如图 3 所示.

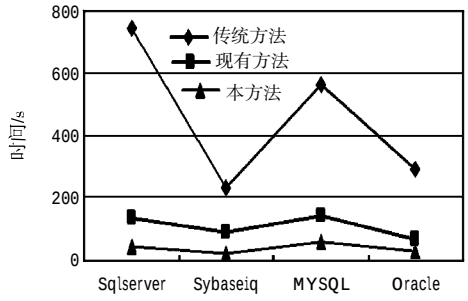

图 3　数据迁移时间图

Fig.3　Data miqration time diagram

4.3.2　验证结果分析

上述实验结果表明,本方法与现有方法和传统方法的数据集成方法相比,本方法技术简单、高效;其效率与现有方法相比, 前者是后者的 3 倍. 此外本方法还能够消除脏数据, 其优越性非常明显. 在实现时的其他方面比较情况如表 5 所示.

表 5　ETL 集成方法对比情况表

Tab.5　ETL integration method contrast table

集成方法	实现难易	效率	脏数据
本方法	易	高	无
现有方法	难	较低	很少
传统方法	简单	很低	多

5　结束语

采用本 ETL 优化技术的集成处理方法,能够方便地将异构数据源集成到统一的目标数据库中. 其利用数据库管理系统与 SQL 语句高度兼容的特点,使用纯 SQL 语句编写的程序对其异构数据进行相应的 ETL 处理,其不但能够方便快捷地提高本身的数据处理效率,同时也能够消除脏数据,保证数据源与目标数据的一致性,在对海量数据源的集成时,其效果非常明显. 相对传统和现有的数据集成方法来说, 本方法有很高的优越性和良好的应用价值.

参考文献:

[1]　徐俊刚,裴莹. 数据 ETL 研究综述[J]. 计算机科学,2011,38(4):12-20.

[2]　苌程,李善平. 基于 ETL 的金融数据集成过程模型[J]. 计算机工程与设计,2010,31(9):2070-2072.

[3]　张靖,雷航,唐雪飞,等. ETL 应用优化设计与实现研究[J]. 微电子学与计算机,2012,29(4):134-137.

[4]　彭璐.基于数据仓库的 ETL 过程优化[J].计算机与数字工程,2010,38(5):166-169.

[5]　杨宏英, 林长松. 异构数据集成系统的应用模式与技术实现[J]. 微电子学与计算机,2006,23(8):70-72.

[6]　郑丹青. 基于元数据的空间数据仓库 ETL 系统设计与研究[J]. 吉林师范大学学报:自然科学版,2010(2):43-45.

[7]　JI Shuiwang, YE Japing. Kernel uncorrelated and regularized discriminant analysis: A theoretical and computational study[J]. IEEE Transactions on Knowledge and Data Engineering,2008,20(10):1131-1142.

[8]　丁月华,杨敏,文贵华,等. 基于 XML 的异构数据源集成与交换的实现[J]. 计算机应用与软件,2006,23(10):134-143.

[9]　DAN Asit, JOHNSON Robert, ARSANIANI Ali. Information as a service:Modeling and realization[C]//IEEE International Workshop on Systems Development in SOA Environments (SDSOA'07),2007:2-6.