一种保持数据完整性的数据集成机制的探讨

杨永强 马世龙 靳 文

(北京航空航天大学 计算机学院, 北京 100191)

摘 要: 目前,企业数据往往分散于各个独立的业务系统中,因此不能很好的实现数据的统一管理和集成,基于对分布异构数据资源特点的分析,提出了一种保持数据完整性的数据集成机制. 该机制通过三级模式 (本地模式、局部模式和全局模式)的映射建立了层次化的数据模型,屏蔽了数据的异构性,并提出了一个数据完整性检查算法,即在数据集成中保证了基于全局模式的数据完整性,从而达到了既能实现数据资源的共享,又能将分布异构的数据资源形成有机整体的目的,最后以地震行业前兆数据集成平台的某一应用场景为例,说明了该数据集成机制具有保持数据完整性的能力.

关 键 词: 分布式数据; 数据集成; 数据完整性

中图分类号: TP 311. 133. 1

文献标识码: A **文 章 编 号**: 1001-5965(2008)09-1045-04

Discussion of a data integration mechanism with data integrality

Yang Yongqiang Ma Shilong Jin Wen

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: At present, the enterprise data is often scattered in various independent business systems, therefore the unified data management and integration is not well realized. Based on the analysis of the distributed and heterogeneous data characteristics, a data integration system with data integrality was proposed. The system introduced the mapping mechanism of three data mode (inner mode, local mode and global mode), established the architecture of hierarchical data model. Because of the architecture, the system could shield of heterogeneous data in the distributed and heterogeneous data environment. The system also proposed a data integrity check algorithm on the global mode. This algorithm would ensure the data integrity in the distributed and heterogeneous data environment. So the system could not only share the data resources, but also integrate the distribution of data resources. Finally, an application example in the earthquake data integration platform was described and analyzed, which showes this mechanism has the data integrity capability.

Key words: distributed data; data integration; data integrality

地震行业的信息化建设要求对全国数据资源进行充分共享、互联互通与协同工作,然而由于各地信息化建设的参差不齐,导致了数据资源具有分布性、异构性以及自治性等特点,即在数据资源集成过程中产生如下问题:首先由于数据资源的分布性和异构性导致了信息孤岛[1]的产生,即无法实现数据共享;其次由于部分集成数据之间存在着语义的关联 (即数据的参照完整性,参照完整性主要是指表间的联系,通过参照完整性约束可以维护相关数据之间语义的关联),因此在对集成数据进行操作过程中,需要保证数据之间的参照完整性,然而分布式数据资源的自治性妨碍了数据参照完整性约束的建立,并且由于集成数据的动态性 (数据源的动态加入和退出),使得数据源之间的约束关系发生变化 (约束关系的取消和约束关系的建立),从而使得应用逻辑对约束关系的维护变得更加复杂,因而如何保证数据完整性[2]已成为当前地震行业中数据集成的难点.

收稿日期: 2007-09-07

作者简介: 杨永强 (1982 -),男,江苏泰兴人,硕士生, younger0616@sina.com.

集成数据的完整性是指,通过在集成数据之间建立全局约束关系,保证了集成数据的参照完整性,实体完整性以及其它自定义完整性,从而在数据集成中将语义相关的数据进行关联,实现了数据的整合.

数据集成主要是将相互关联的异构数据进行集成,并以透明的方式对数据进行访问和处理.目前主要有2种数据集成方案:①将分散异构的数据移植到一个核心数据库中.该方法必须对异构数据进行数据模式和数据类型的统一转换[3],并且将转换后的数据导入到核心数据库中,即此方案会占用很大的物理存储空间,并且无法保证底层数据与集成数据之间数据的一致性.②利用中间件技术[4]来实现异构数据集成,该方法并不需要改变原始数据的存储和管理方式,其向下协调各数据库系统,向上提供统一数据访问接口,从而实现了异构数据的集成检索[5],但是在集成数据之间无法将语义相关的异构数据进行关联,即无法保证集成数据的数据完整性,从而无法实现数据的一体化管理.

本文通过分析上述两种集成方案提出了一种保持数据完整性的数据集成机制,即通过三级模式(本地模式、局部模式和全局模式)的映射机制,屏蔽了数据的异构性.通过在全局模式中建立约束关系,实现了集成数据的关联,并通过任务引擎的解析执行,保证了集成数据的完整性.

1 系统结构

本文提出的系统架构从下到上主要分为5个层次,如图1所示.①数据资源层:各种分布异构数据资源的集合,如关系型数据库,非关系型数据库以及数据文件等.②资源接入层:将数据资源层中各个数据源以服务的形式进行接入,并且公共模式适配器[6]可以屏蔽底层数据的模式异构以及各个数据源执行语句的异构.③数据汇聚层:通过数据转换屏蔽掉了数据的语义异构,并将转换后的数据以统一的格式进行访问.④任务引擎层:基于已注册的约束关系对任务进行解析执行,即在保证数据完整性的前提下完成一次任务的执行.⑤应用层:提交任务的用户或上层应用.

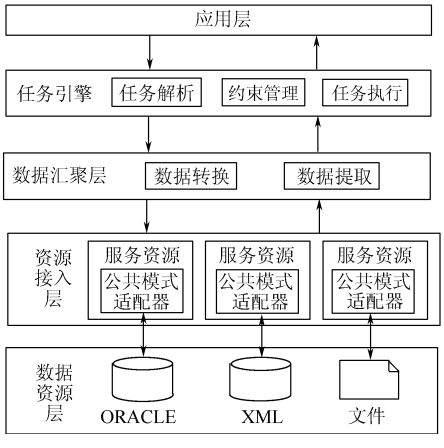


图1 系统架构图

本系统通过资源接入层以及数据汇聚层实现了三级模式的映射,并且通过任务引擎层对任务的解析,保证了任务执行过程中数据的完整性,从而实现了数据资源的有效集成.

2 系统关键技术

2.1 三级模式及其映射

本文提出了本地模式、局部模式和全局模式,如图2所示.①本地模式:实际存在的数据资源模式,包括关系型数据库(如Oracle,Mysql)以及非关系型数据库(如XML数据库)或以文件方式进行存储的数据资源.②局部模式:将本地数据资源进行数据模式的转换,统一转换成为关系型模式,并且将SQL语句解析成为本地模式所能接受的执行语句(如XML数据库中的Xquery语句).③全局模式:屏蔽局部模式之间数据语义的异构,并将局部模式进行整合,对上层,通过统一的全局模式实现了集成数据的透明访问.通过三级模式的映射,建立了三层数据模型,从而降低了异构数据集成的复杂性,提高了数据集成的灵活性.

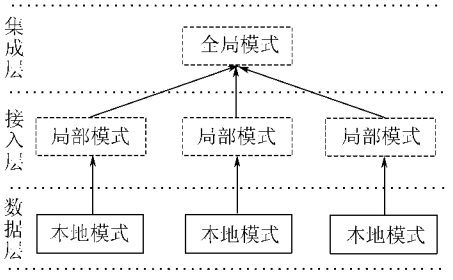


图2 三级模式映射

2.2 数据完整性

集成数据的完整性分为两个层次:局部数据完整性和全局数据完整性.局部数据完整性是指各个数据源内部的数据完整性,全局数据完整性是指不同数据源之间的数据完整性.通过各个数据源的自治性可以实现局部数据完整性,但是全局数据完整性的实现则需要一种新的机制对其进行保障,因此本文主要是对全局数据完整性进行研究.

本文通过在全局模式中的虚拟表(下文通称为数据对象)之上,建立全局的约束关系,并基于

约束关系对提交的任务进行解析执行,从而保证了全局数据的完整性.

由于实体完整性和非级联的参照完整性是级联参照完整性的一个子集,因此在本文中只对级联参照完整性关系进行讨论.

级联参照完整性约束关系具有传递性,因此通过约束树构造子算法(算法 1)可以将其构造成一棵约束关系树.

算法 1中相关的概念:

任务主表(M):提交的任务所操作的数据对象.

正向约束(+):外码表到主码表的方向,如 A到 B是正向约束,则表示为 $A \rightarrow B$.

反向约束(-):主码表到外码表的方向. 如 A到 B是反向约束,则表示为 $A \leftarrow B$.

约束集合(C):约束关系的集合.

算法 1

1) $C \leftarrow \varnothing$, $i \leftarrow 1$. 以 M为根节点构造约束树 TC,将 M视为待扩展节点,并将 M视为 TC的第 1层. 转第 2)步.

2) 若 i大于 TC的深度,则 TC构造完毕,否则转第 3)步.

3) 对 TC中的第 i层的待扩展节点从左到右依次进行反向约束查找. 查找相应的约束关系 C_j以及对应的外码表 V_j,若 $C_j \in C$,则不在此方向上进行拓展,否则,将 V_j作为当前查找结点的最右子节点加入到 TC中,并将 V_j视为待扩展节点,且 $C \leftarrow C \cup \{C_j\}$,转第 4)步.

4) 对 TC中的第 i层的待扩展节点从左到右依次进行正向约束查找. 查找相应的约束关系 C_j以及相应的主码表 V_j,若 $C_j \in C$,则不在此方向上进行拓展,否则,将 V_j作为当前查找结点的最右子节点加入到 TC中,并将 V_j视为完成节点,且 $C \leftarrow C \cup \{C_j\}$,转第 5)步.

5) 将 $i \leftarrow i+1$,转第 2)步.

图 3展示了一棵约束树的构造过程.

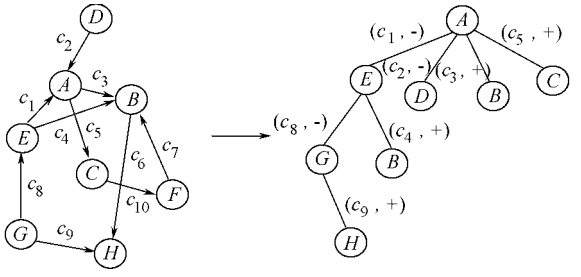


图 3　约束关系树的转换

2.3　任务引擎

任务引擎主要是对提交的任务进行解析和执行,通过解析操作可以得到对应的约束关系树,并且通过任务树构造子算法(算法 2)可以对约束关系树进行裁剪,得到相应的任务树.

算法 2中相关的概念:

剪枝:将 A树中的 B节点进行"剪枝",是指在 A树中将以 B节点为根节点的子树进行删除.

正向拓展(+):节点通过正向约束与其子节点相联.

反向拓展(-):节点通过反向约束与其子节点相联.

算法 2

1) 查找约束关系树 TC的根节点 R,若任务类型是插入或约束关系注册检查,则转第 2)步. 若任务类型是删除则转第 3)步,若任务类型为更新则转第 4)步.

2) 以 R开始对 TC进行层次遍历,将所有反向拓展的子节点进行"剪枝"操作. 任务树构造结束.

3) 以 R开始对 TC进行层次遍历,将所有正向拓展的子节点进行"剪枝"操作. 任务树构造结束.

4) $U \leftarrow \varnothing$,以 R开始对 TC进行层次遍历,将所有反向拓展出来的节点加入 U中,若 U中重复出现两个相同的节点,则表明约束条件存在二义性,直接返回任务执行失败提示. 否则,任务树合法性验证通过,任务树构造结束.

通过上述算法,可以将约束关系树解析成为任务树. 在任务树中,每个节点都是一个子任务所要执行的数据对象,并且通过"剪枝"操作将无用的子任务从任务树中删除,从而得到了一棵较优的任务树,提高了任务的执行效率. 图 4展示了当任务类型为删除时任务树的构造过程.

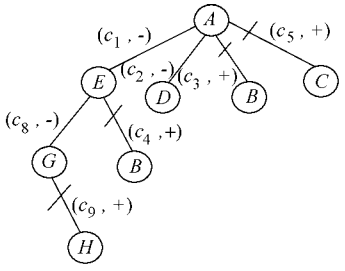


图 4　删除操作的剪枝

任务执行引擎的调度策略是对任务树从下到上、从左到右进行层次遍历,即首先执行最底层的子任务集,只有当本层的子任务全部执行成功之后,才能开始执行上一层的子任务集,以此类推,实现了任务树的层次遍历与执行. 当向上遍历到树的根节点并执行成功之后,将任务提交,否则将

任务回滚,从而完成了一次任务的执行.

2.4 数据完整性的保证机制

本系统通过约束关系的注册检查以及数据操作的机制保证了数据的完整性.

约束关系注册检查是在约束关系注册时对约束关系的合法性进行验证,若在当前数据中存在不符合约束关系的数据时,则将约束关系视为非法,否则将接受此约束关系的建立.

本系统中对数据操作的方式主要分为两种,即全局操作与局部操作.

全局操作是通过全局模式来完成对数据的修改和删除等操作,任务引擎实现了基于全局模式的任务解析与执行,从而在全局操作中保证了集成数据的完整性.

局部操作是通过局部模式来完成对数据的修改和删除等操作,通过局部模式与全局模式的映射关系,可以将操作的局部对象映射成为全局对象,并以全局操作的模式实现了对操作任务的解析执行,从而在局部操作中,保证了集成数据的完整性.

3 用例与意义

本系统应用于全国地震前兆数据集成平台中,实现了对分布异构数据的集成管理. 本系统具有很好的可扩展性,即当加入一个数据资源时,只需要建立全局视图以及相关的约束关系,就可以在不用改变上层应用程序的数据处理逻辑的情况下,将其透明的接入到系统中来,从而保证了本系统具有很好的可扩展性.

图5表明了在系统中注册了2种全局数据统

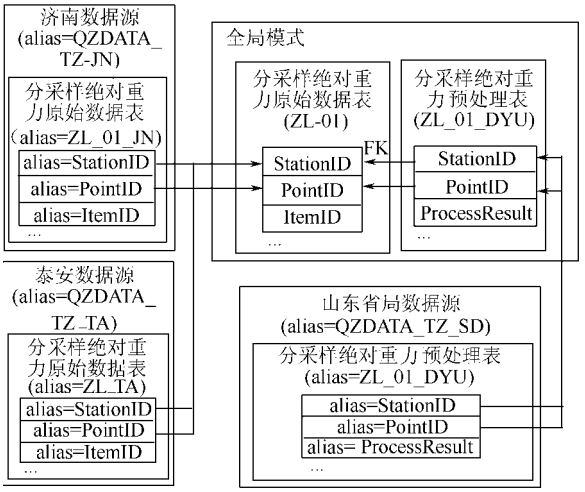


图 5 映射关系

一视图,即重力数据统一视图(ZL_01)和预处理数据统一视图(ZL_01_DYU),并且在这两种数据资源之间注册了相关的约束关系,当上层应用提交了一个对原始重力数据进行修改的任务,如将重力数据中 PointID 为“1”的重力数据更新成“a”,通过任务引擎的解析会同时也将相关预处理数据的 PointID 由“1”更新成“a”,从而实现了集成数据的一致性,保证了数据的完整性,因此避免了无意义的(不存在相关原始数据)预处理数据的产生,满足了数据集成的需求.

4 结 论

本文提出的保持数据完整性的数据集成机制,将数据进行层次化的抽象,并在全局模式中建立全局的约束关系,通过完整性检查算法实现了任务的解析与执行,即在异构数据集成中保证了数据的完整性,并通过全局操作与局部操作的机制,实现了对集成数据的一体化管理.

参考文献 (References)

[1] 何涛,刘君强,张学斌. 异构数据源数据集成的研究[J]. 计算机工程与科学, 2006, 28(9): 132-135
He Tao, Liu Junqiang, Zhang Xuebin. Research of the data integration of heterogeneous data sources [J]. Computer Engineering and Science, 2006, 28(9): 132 - 135 (in Chinese)

[2] Ramakrishnan R, Gehrke J. 数据库管理系统[M]. 第二版. 北京:清华大学出版社, 2002: 39 - 64
Ramakrishnan R, Gehrke J. Database management systems [M]. 2nd ed. Beijing: Tsinghua University Press, 2002: 39 - 64 (in Chinese)

[3] Chercenak A, Foster I, Kesselman C, et al. The data grid: towards an architecture for the distributed management and analysis of large scientific datasets [J]. Journal of Network and Computer Applications, 2001, 23 (3): 187 - 200

[4] 朱韵箴,李双庆. 基于 XML 的分布式数据交换中间件研究[J]. 计算机应用, 2002, 22(12): 16 - 18
Zhu Yunchi, Li Shuangqing. Research of distributed data exchange middleware based on XML [J]. Computer Applications, 2002, 22(12): 16 - 18 (in Chinese)

[5] Collins S R, Navathe S, Mark L. XML schema mappings for heterogeneous database access [J]. Information and Software Technology, 2002, 44(4): 251 - 257

[6] Antonioletti M, Atkinson M, Baxter R, et al. The design and implementation of grid database services in OGSA-DAI [J]. Concurrency and Computation: Practice and Experience, 2005, 17(2): 357 - 376

(责任编辑:张 嵘)