基于共享库的数据集成方案改进

毛 赟,徐宏炳
(东南大学 计算机科学与工程学院,江苏 南京 210096)

摘 要:基于共享数据库的方案是现在高校进行数据集成普遍采用的模式。但随着全局业务的增多,这种方案不可避免地会遇到共享数据库性能瓶颈的问题。从有效利用业务数据库、降低共享数据库业务压力出发,在研究了已有的改进方案之后,对基于共享数据库方案的框架进行改进,提出了一个同时使用共享数据库模式和中介模式,能够进行负载均衡的新的数据集成框架模型。对模型的优劣性进行了分析。

关键词:数据集成;共享数据库;业务数据库;负载平衡;中介模式
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2008)07-0170-03

Improving Solution of Data Integration Based on Shared Database

MAO Yun, XU Hong-bing
(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Method based on shared database is wildly used in data integration of campus. This method will encounter capability bottleneck inevitably while the globe business increasing. Aiming pressure release of shared database by using business database effectively, a new data integration framework model is introduced after studying of current improving methods. The model is developing upon shared database model. It uses shared database model and mediator model simultaneously by balancing of the load. After that, analysis of the new model is made.

Key words: data integration; shared database; business database; load balancing; mediator model

0 引 言

基于共享数据库的数据集成技术[1]是以共享数据库存储需要共享的数据实现数据集成,并通过元数据提供一个统一的数据模式。共享库和数据仓库不同点在于,前者提供当前的业务数据,不维护历史数据;而后者主要面向高层决策,存储统计信息。共享库更像是一个传统的全局数据库,负责把需要共享的数据从业务库抽取上来。当前该技术已经在复旦大学、浙江大学、同济大学等高校得到广泛应用。

但随着全局应用的增多,共享库业务压力增大,需要共享的数据也不断积累,明显成为整个系统的瓶颈。与此同时,业务库却可能完全闲置,甚至修改和添加的数据仅仅需要同步到共享库就不再使用。如何缓解共享库压力,有效利用业务库资源是需要切实考虑的问题。

收稿日期:2007-10-18
作者简介:毛 赟(1982-),男,江苏无锡人,硕士研究生,研究方向为数据集成和应用集成;徐宏炳,教授,研究方向为数据库设计与应用、数据挖掘技术、企业数据集成。

1 改进方式

数据集成改进方案大体上有以下两种:

(1)改用中介模式/全局模式:作为现在最典型的数据集成方式,它通过提供一个统一的数据逻辑视图来隐藏底层的数据细节,具体的局部数据源利用包装器为中介模式提供规范的数据。根据跨库视图的构造方式,中介模式可分为 GAV(Globe-as-View)[2], LAV(Local-as-View), GLAV(Globe-Local-as-View), BAV(Both-as-View)[3,4]等实现方式。中介模式基本上不存储任何转化后的数据,固化视图虽然也有所研究,但还是处于研究阶段。中介模式的缺点是处理查询时,效率比较低下,存在较大的延时。改用中介模式将使原来共享库方式下的资源无效。

(2)使用负载平衡:负载平衡是为提高性能和克服现有设备中的缺陷而将某些负载分配到多个链路、服务器、处理器或其他设备的过程。负载平衡能够有效地解决共享库的性能限制。但共享库数据量庞大,简单的同步多个共享库备份,分派请求的做法会浪费大量的资源。

上面两种方案虽然能解决或暂时解决共享库的瓶

颈问题,但改用中介模式的方法否定了原有的集成方案和成果;负载平衡的方法不能有效利用闲置的业务库资源,而且同步的粒度太大,资源浪费明显。文中的改进方案采用负载平衡的想法,在数据集成共享库繁忙时把全局请求分配到空闲的业务库上。在全局请求通过业务库进行响应时可以看作是一套完整的中介模式。可以说这种改进方案结合了中介模式和负载平衡,互补之间的缺点。

2　整体框架

文中框架采用局部申请、集中分配的方式设计,在确保实现目标的同时尽量保证系统的简单性。这种方式来源于动态负载平衡中的接收者启动(RID)算法[5]。但不同于一般 RID 算法,这种方式采用集中分配从而防止掠夺现象的发生。负载平衡的负载参数很多,但国际上多使用运行队列任务数作为负载指标[6]。本框架也采用它作为负载强度的判断依据。集成框架见图 1。

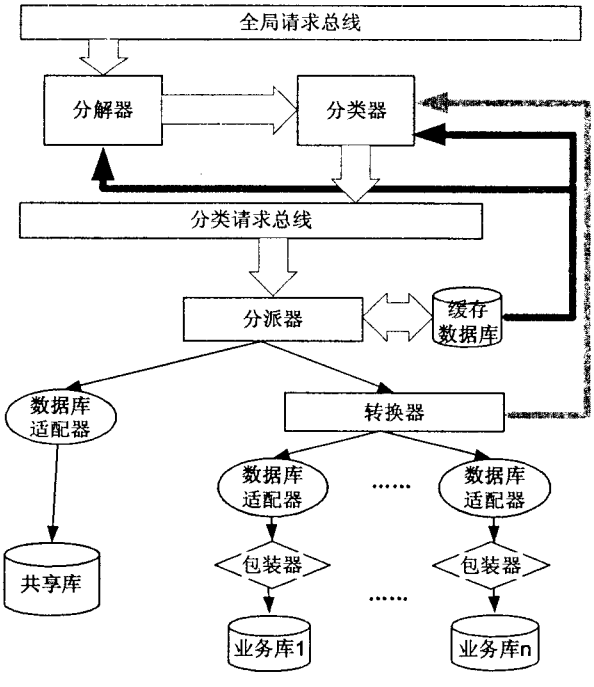

图 1　集成框架

全局请求总线用于接受全局应用对底层全局数据的复杂访问,包括嵌套查询等。这些数据将由分解器做进一步处理。分解器主要用于把所有嵌套访问解套,分解为多个非嵌套的子访问请求,并标明这些子访问之间的事务关系。同时它还负责在繁忙状态下把需要同时访问业务库和共享库的请求分解为仅访问业务库或共享库的多个请求。分类器对分解器输出的子访问请求进行分类。分类按照需要使用的业务数据库进行。分类器的分类依据有两个源,第一个源在初始化时对转化器的转化规则进行分析,得出全局各属性的权威数据库。第二个源是缓存数据库。缓存数据库中存放着数据库异步时数据更新、删除对应的是共享数据库还是业务数据库。缓存数据库的来源是分派器,在整个框架中,分派器起到了集中分配的作用。分派器接受来自各数据库的令牌,并保存起来。当有访问请求时,分派器根据请求的分类和相应数据库的令牌数决定请求的分派。如果业务库和共享库都有令牌则按概率随机分配,令牌多的概率相应较大。任务分派后相应数据库令牌减一。如果是更新或删除操作还要把分派到的数据库记录进缓存库中。转换器的工作比较复杂,主要任务涉及到查询解析、重写[7]与优化、并行处理以及结果汇总。数据库适配器对相应数据库的请求任务进行管理和监测。数据库适配器管理一个定长的任务队列,当任务队列周期性向分派器发送队列空余位置数的令牌。当队列任务数接近上限时还将触发适配器向分派器发送满载信息,获取满载信息的分派器将把相应数据库令牌清零。数据源包装器是中介模式的组成部分,它主要负责对数据进行转换,使之符合中介模式。

系统运行时有空闲、繁忙和同步三种状态。处于空闲状态的系统,业务库和共享库数据同步,所有更新和删除请求将同时传送给业务库和共享库,查询请求按概率随机传给业务库或共享库处理。当系统由空闲状态转为繁忙状态,分派器将把更新和删除请求按概率分派到数据库。但前提是需要更新或删除的数据在业务库和共享库中同步。如果数据不同步,分派器将根据由分类器从缓存库获得的分类结果指定操作的数据库。同步状态可能在两种情况下出现。一种情况是繁忙状态下系统同步周期,这种情况下系统会同步部分数据。这种情况是为了保持系统的效率问题,异步数据过多会导致分派器按新数据的存放地点选择数据库,从而忽略了系统的效率。第二种情况是繁忙状态下分派器获得了较多的令牌,也就是实际数据库系统已经有较多的空闲。为了使系统回复到空闲状态,需要使异步的数据库同步。三种状态的动作差异主要表现在分派器上,判断由分派器上各数据库的令牌数决定。图 2 表示了三种状态所有可能的迁移。

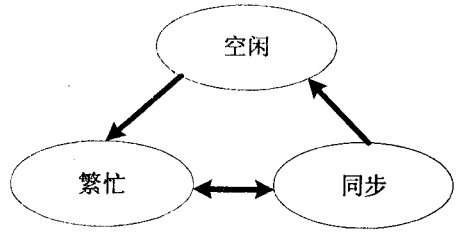

图 2　系统状态迁移

3 方案优劣性分析

框架优点如下：

(1)框架在一定程度上利用了业务数据库资源，缓解了共享数据库压力。

(2)框架有效利用原有共享数据库模式相关资源，保留了共享数据库模式查询速度快的优势。

(3)框架在负载平衡方面可以做扩展，把负载平衡的粒度降低到业务库甚至业务库共享数据的级别，从而使负载平衡的性价比更高。

(4)框架模块化后，各模块可以复用。

本框架缺点如下：

1) 框架在异步状态下需要定期同步将来查询操作经常使用的项。周期性的同步操作在系统繁忙时导致框架的性能优化下降。

2) 框架比较复杂，仅适用于已使用共享库方案进行数据集成的项目。对于从头开始数据集成的项目应当考虑更优的解决方案。

3) 非技术缺点。在行政上，业务库可能存在访问或修改限制，需要工程实施人员针对框架对业务库的依赖这一特性与用户进行沟通。

4 结束语

现今高校数字化校园建设不仅需要在业务系统之间实现共享需求，还要建立统一一致的全校数据模式，理顺业务系统之间的数据关系，对共享数据进行规范化和标准化工作，保证学校的统计数据来源的一致性[8]。使用共享数据库方式做数据集成是高校最优选择。文中的框架为高校的共享数据库方式提供了升级方案，在今后高校的数据集成中将会有较大的应用前景。本框架的后续工作将重新考虑系统异步状态下的工作方式，进一步有效利用业务数据库资源。

参考文献：

[1] 王天亮，陈 刚，徐宏炳.基于共享数据库的数据共享数据技术[J].计算机工程与设计，2007，28(8)：1923－1926.

[2] Lenzerini M. Data integration: A theoretical perspective[C]// In: Pops L, editor. Proceedings of 21st ACM SIGACT－SIGMOD－SIGART Symposium on Principlrs of Database System. New York: ACM Press, 2002: 233－246.

[3] McBrien P, Poulovassilis A. Data integration by bidirectional schema transformation rules[C]//In: Dayal U, Ramamritham K, Viyayaraman TM, eds. Proceedings of 19th International Conferences on Data Engineering. New York, NY, USA: IEEE Computer Society, 2003: 227－238.

[4] Jasper E, Tong N, McBrien P. Generating and optimizing views from both as view data integration rules[C]//In: Barzding J, Caplinskas A, eds. Proceedings of 6th International Baltic Conference on Databases and Information Systems. Amsterdam: IOS Press, 2005: 3－19.

[5] Willebeek－LeMair M H, Anthony P. Reeves Strategies for Dynamic Load Balancing on Highly Parallel Computers[J]. IEEE Transactions on Parallel and Distributed System, 1993, 4(9): 979－993.

[6] 李冬梅，施海虎.负载平衡调度问题的一般模型研究[J].计算机工程与应用，2007，43(8)：121－125.

[7] Arens Y, Knoblock C A, Shen W M. Query reformulation for dynamic information integration[J]. Journal of Intelligent Information Systems, 1996, 6(2－3): 99－130.

[8] 周长春，徐宏炳，张小伟.基于共享数据库的数据集成方案的改进[J].计算机工程与设计，2007，28(8)：1917－1919.

(上接第 169 页)

采用 C/S 的程序结构。同时在硬件配置和系统设计中还充分考虑系统的发展和升级，使系统具有较强的扩展能力。

(5)经济性：即在实用的基础上做到最经济，以最小的投入获得最大的效益。包括在硬件和软件配置、系统开发和数据库建立上都充分考虑投入和经济效益。

5 结束语

煤矿采掘衔接生产计划管理系统是按照煤矿现场要求的矿井采掘衔接和生产报表管理的辅助决策与管理基础上开发的，在满足回采要求的前提下，模拟生成掘进衔接方案，经交互式调整，才确定最终采掘衔接方案。此种处理方法提高了系统的灵活性和适用性，有利于对煤矿生产布局进行科学分析与评判，保证煤矿数据的安全，提高矿区生产管理的工作效率。

参考文献：

[1] 臧立岩，邢存恩. 煤矿采掘计划计算机自动编制系统的研究[J]. 山西煤炭，2005，25(2)：11－14.

[2] 蒋国安，王新华，李兴华. 矿井采掘计划的编制与检验[M]. 北京：煤炭工业出版社，1992.

[3] 李仲学，廖荣怀. 地下煤矿采掘衔接计划系统的一种实现[J]. 中国矿业，1996，5(6)：54－57.

[4] 陈鸿章. 矿山系统工程的基本方法与信息论的应用[M]. 北京：煤炭工业出版社，2001.

[5] 陈建宏，古德生，罗周全，等. 采矿 CAD 中图元属性表述方法的研究[J]. 金属矿山，2001(8)：9－11.