

●任瑞娟 王东方

基于分布式数据库构建分布式本体的方案设计

摘 要 分布式数据库是面向跨地区不同组织机构和企业之间进行数据交换的新的数据组织形式。采用分布式数据系统来设计实现分布式本体的方案,既能保证系统数据的分布性,又可以成功实现系统中实时更新和数据的一致性。图1。参考文献1。

关键词 分布式数据库 分布式本体 规划方案 分布式数据

分类号 G250.74

ABSTRACT Distributed database is a new form of database organization oriented to the data exchange among different institutions and enterprises. By using distributed database to realize distributed ontology, we can both ensure the distributedness of the system and the real-time updating and data consistency of the system. 1 fig. 1 ref.

KEY WORDS Distributed database Distributed ontology. Design. Distributed data

CLASS NUMBER G250.74

1 分布式数据库

分布式数据库是面向跨地区不同组织机构和企业之间进行数据交换的新的数据组织形式。它面向全部组织机构的信息,实行分而治之的集中管理,比单一的集中管理更有利,不仅能克服集中式数据库的不足,也完全适应各组织机构的结构特点。成熟的数据库技术与网络技术互相渗透,为分布式数据库的发展准备了条件。

1.1 分布式数据库的概念与特点

分布式数据库是由一组数据组成,这些数据物理上分布在计算机网络的不同节点(或称场地或站点)上,逻辑上是属于同一个系统。它有两个特点:(1)分布性。分布式数据库数据存储在网络环境中的多个物理场地上,而与集中式数据库不同;集中数据库数据存放于同一场地,确切地说是同一计算机的同一设备上。(2)逻辑整体性。分布式数据库的数据,逻辑上是集中的,是一个整体,区别于分布在网络上不同节点的集中式数据库。后者数据内部没有逻辑关系^[1]。

分布式数据库的主要优点:(1)有较高的可靠性。系统中一台机器发生故障,不会导致整个系统被破坏。(2)分布式数据库是透明的。对于每一用户来说,他所看到的都是一个统一的概念模式。主要缺点是其复杂性导致软件开发成本高、处理开销大等。在权限允许的情况下,应用程序可以同时访问和修改分布式数据库中所有的数据,而不必关心数据的存储位置。它允许用户开发的应用程序把多个数据库当

做一个完整的数据库看待,允许用户透明地操作远程数据库的数据,并可以把它们连接在一起,满足多个工作部门或地区的需求,使应用程序看起来只有一个大型数据库。一些分布式数据库如 Oracle 的分布式数据库,允许用户对某单一逻辑的数据库中数据的任何请求,并且这样的请求均可被自动分解、自动寻址、自动转换为网络请求,并在相应场地上实现相应操作。

1.2 分布式数据库的体系结构

分布式数据库有3种模式:全局概念模式、分片模式、分布模式。它与集中式数据库的结构体系对比如图1所示,这种结构保证了它的上述各种特性。

2 构建分布式数据库的规划方案

2.1 分布式数据系统功能规划

分布式本体系统的功能目标是根据网络环境下构建分布式本体的实际需要,为完成分布式本体的各模块的功能,实现系统内部包括各查询机构的资源共享及不同用户权限的控制。考虑到系统数据收集的实时性要求不高,因此采用 Browser/Server 结构来进行设计应该比较合理可行。

根据系统功能的要求,本体系统内部分为若干个局部本体库,各局部本体库又分为上层局部本体与下层局部本体两个部分。各局部本体库内部数据管理为集中式数据库,也可通过内部对象来交流信息。各局部本体库之间为分布式数据管理,但由于采用分布式数据库系统,在应用上仍能形成一个有机整体。

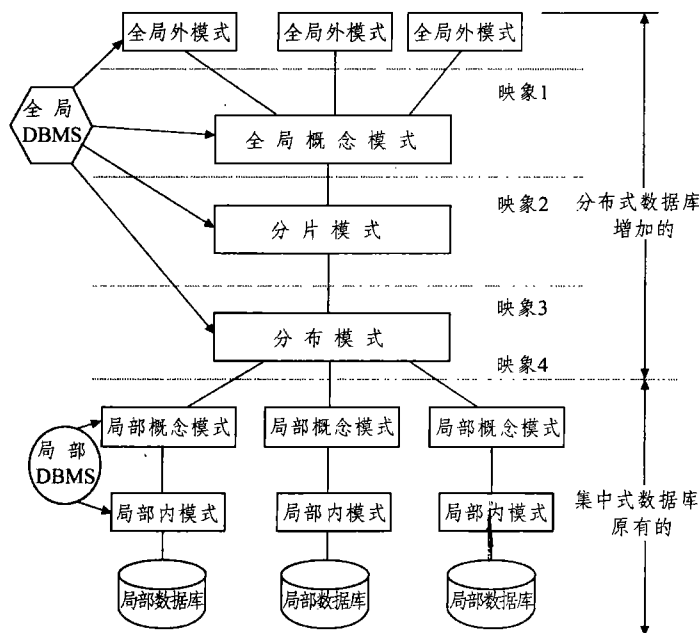


图1 分布式数据库系统的模式结构

2.2 数据分布与冗余规划

数据的分布与冗余度(复制数据副本的多少)是分布式数据库系统数据存储的分布方式。分布式数据库系统中,不同站点的数据访问引起站点间的通信。通信速度相对于存储设备的存取速度来说非常慢。通信系统存在较长的通信延时问题,处理通信和传输信息的代价是昂贵的。如何进行合适的数据分布与冗余度管理,以满足局部自治性、减少站点间通信次数和传输量、提供高效率和高可靠性,是分布式数据库设计的主要问题。

在设计系统中应采取快照方式,在各个局部本体站点上创建其他站点的副本。由于本体概念的信息适时性(更新频次)要求不高,为了减少网络传输量,刷新率设置较大。另外,对于局部本体中下层本体的内部相关的共享信息,采取复制的方式。建议在除主站点以外的其他站点上,如果硬件条件许可,推荐创建副本,这样局部本体上层及下层的站点与主站点一样,对这些共享信息有同等的操作权限,一个站点的数据所作的修改,同样会反馈到其他的站点,保持数据库的一致性。即使其他站点暂时出现故障,也不会影响本站点数据副本的正常使用,保证系统的高效和可靠。

2.3 分布式数据查询规划

分布式数据库系统应向用户提供一个统一的访

问数据接口。对用户来说,使用分布式数据库系统和使用集中式数据库系统一样,好像所有数据都存储在自己的计算机中,即实现数据分布的完全透明性。数据实际上是分布的,查询处理中需要站点之间数据的传递,多站点的数据查询使得并行化处理成为可能。对分布式查询,要充分利用处理的可并行性和对数据进行合理分布来优化,以确保查询费用最小。

在分布式本体系统中,为了能够实现数据的快速查询,减少网络流量,建议在数据库链接的基础上将一个站点的信息在其他站点创建副本,将分布式查询处理优化为对本站点的数据快速访问。

2.4 分布式数据更新规划

分布式数据库系统中,一般存在数据的多副本情形,不同站点数据的更新会引起数据的不一致。系统必须以最小的代价保持各冗余副本的一致。

本系统中,对适时性要求不高的数据如局部上层本体概念,将刷新率设置得低一些,如 1d 或 0.5d。对适时性要求较高的数据,如局部本体的下层本体库数据信息,将刷新率设置得高,如 1h 或 0.5h。对于适时性要求最高的内部交流信息,设置为立即提交方式,即使对一个副本的修改,都会被命令立即修改其他的相关数据副本。

3 分布式数据库实现方案

3.1 创建和使用数据库链接

数据链接是指两个物理数据库的连接,并允许客户机作为一个逻辑数据库访问它们。数据库链接是一个定义从一个数据库服务器到另一个数据库服务器的单向的通信路径指示器。链接指示器实际是一个数据字典入口。

分布式数据库的核心是数据复制,数据复制建立在数据链接的基础上。

数据库链接允许本地用户访问远程数据。为了实现链接,分布式数据库中的任何一个数据库在网络中必须有唯一的全局数据库名。全局数据库名是区分分布式系统中不同数据库服务器的唯一标识。为了方便,暂且用局部本体 X 上层、局部本体 X 下层来表示局部本体中 2 个数据库的全局数据库名。有了唯一标识,即使不同的站点存在相同名称的数据库,也可区分。当建立全局命名以后,数据库链接对分布式数据库的用户来说本质上是透明的。即只对全局关系进行操作而不必考虑关系的分片,并且不必了解存储场地,也不必了解局部场地的数据存储模式。各产生了数据库链接建立以后,与集中式数据库相似,可以执行 SQL 语句访问远程数据对象。

3.2 创建实体化视图

在本体系统的分布式环境中有一些公用的基表,这些基表对应分类法中的一级类目。为提高系统性能,使这些基表存储在任何一个站点上(实际为冗余),并建立了远程数据库的数据链接。如果用数据链接来访问其他站点的数据库,整个系统在访问远程对象时性能就非常低,另外,还会增加网络传输量和网络费用,容易形成网络传输瓶颈。这时,如果使用视图技术(快照),将提高使用远程数据库数据应用程序的性能。视图技术为远程基表提供本地的复制,可以包含远程基表的全部和部分数据,并被自动刷新。对基于主体表的快照进行快速刷新,必须对快照建立日志。

3.3 创建数据库复制

复制技术给用户提供了一种快速访问共享数据的办法。在实际的管理应用中,随着用户使用数据库服务器规模的扩大,服务器性能会明显下降。如果网络中有多台服务器,则可以通过复制机制来自动维护相同的数据库数据。对任何一个用户来说,尽管访问的数据库服务器会不一样,得到的结果却是相同的。

分布式本体系统中,在主站点的数据库中保存共

享数据,以全局本体库服务器为主站点,将主站点上数据共享,复制到其他站点的数据库中,建立副本。

3.4 创建可串行化调度

通常情况下,数据库系统中总是有多个事务在运行。这些事务可能并发地存取相同数据,称为事务的并发操作。并发操作虽然提高了系统运行效率,但也带来了一些问题:可能导致丢失更新、错误读取、重复读取等错误,破坏数据的一致性。

如果分布式事务的并发执行能够以一定的顺序串行执行,就可使数据库处于新的一致性状态,事务的并发控制就可以正确处理并行执行的事务对数据库的冲突操作。在分布式数据系统中,可串行化调度是控制事务并发要寻求的基本方法,它让有冲突的操作串行执行,没有冲突的操作并行执行。其作用主要是协调同一时间访问同一数据库文件的多个事务之间的关系,防止这些事务之间发生冲突,产生一个可串行化调度。对分布式事务的可串行化调度(全局调度)最终会转化为子事务的可串行化调度(局部调度)。

在数据项没有副本的情况下,如果所有局部调度都是可串行化的,全局调度也称作是可串行化的。但在有多个副本的分布式数据库上,即使所有局部调度都是可串行化的,由于每个站点上的事务串行化的顺序不一样,也不能保证分布式数据库的相互一致。这里采用的方法能够解决本地事务和全局事务的冲突,同时当全局事务转化为各局部站的子事务,使子事务能够按照同一串行化执行事务,保证了数据库的一致性和完整性。

4 系统安全考虑

安全是每一个系统必须考虑的问题,分为系统安全与数据安全。用户必须要通过系统的安全认证,才能登录系统。用户以合法身份登录系统时,系统进行后台数据库的连接,根据用户类别和权限可以找开使用相应的数据表。数据库的安全是保护数据库,防止不合法使用造成数据泄露、更改或破坏。

参考文献

- 1 萨师煊,王珊.数据库系统概论(第3版).北京:高等教育出版社,2000

任瑞娟 河北大学图书馆技术部主任,副研究馆员。通信地址:河北保定。邮编 071002。

王东方 河北大学网络中心系统管理员。通信地址同上。(来稿时间:2005-12-09)