分布并行结构的数据仓库成绩管理系统

孟志青　游 峰　郭云飞　(湘潭大学计算机科学系　411105)

摘要：本文介绍采用了数据仓库(Data Warehouse)技术的思想和客户/服务器开放式体系结构，在现有的普通数据管理系统的基础上全面论述了开发分布并行式成绩管理系统的结构设计与实现。

关键词：分布并行　系统　模块　数据库

一、系统分析

1.问题提出

目前我国高校对成绩管理基本上都实现了计算机管理，但针对高校成绩管理的数据多样性、复杂性、准确性、可靠性，数据的大批量处理和学校各系、院的分布现况，集中式的数据库系统不能尽如人意，存在以下一些问题：

(1)数据处理时间长。由于全校所有的数据库统一由单机或服务器管理，数据量大，采用顺序查询和统计所耗时间长，尤其采用局域网互联技术，数据和文件在结点与网络之间传输量大，网络负担重，很容易成为瓶颈。

(2)数据不安全。在单机系统中，别人可以通过数据库管理软件查看或修改数据；在网络系统中，由于数据库系统与联机事物处理系统的紧藕合，也容易对数据进行非法操作。

(3)数据的一致性问题。由于存放在系统中的数据库不是只读性数据库，当有数据更新时，可能造成数据前后读取的不一致。

(4)故障恢复能力。高校的学生成绩一般需永久性存放，而在单机或网络系统里，数据只存放在单机硬盘或服务器硬盘上，一旦受到病毒侵入或其他破坏，损失不可估量，且不易全部恢复，即使有备份，重新装入也非常耗时且麻烦。

2.设计思想

为了解决上述问题，提高对数据的处理能力和安全性，为使管理系统高速可靠地运行，数据库系统对整个系统内复杂的数据进行有效地管理，以满足多功能系统对数据提出的准确、可靠、正确、快速以及使用方便的要求，我们采用了数据仓库(Data Warehouse)技术的思想和客户/服务器开放式体系结构，在现有的普通数据管理系统的基础上开发了分布并行式高校成绩管理实验系统。在广义上讲，所谓数据仓库就是一个专门的数据仓储，用来保存多个数据库或其他信息源选取的已有数据，并为上层应用提供统一的用户接口，用以完成数据的查询和分析。数据仓库系统可分为三大部分：即数据源、后端加工、前端服务。数据源提供原始数据；后端加工实施数据后处理(包括接收、储存、析取、汇总等)；前端服务面向最终用户。数据仓库技术支持并行处理的分布 DBMS，支持异构环境下的分布式数据处理的客户/服务器技术，以及实现桌面信息系统集成的方案和工具。

3.解决方案

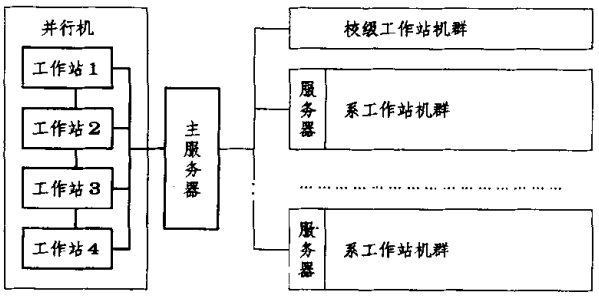
结合这两种技术来建立一个统一的、开放式的数据仓库，对全校的成绩数据，按大颗粒地存放在四台并行工作站上，例如按系存放，对全校的其他数据库，由于数据量相对小的多，则存放在网络服务器硬盘上；采用若干台工作站作分布式用户，在这些工作站上完成客户端的功能，而在网络服务器和并行工作站上做服务器的处理；将并行工作站上的成绩数据库作成只读数据库，只有在工作站上已确定的数据才传递给它们，不允许在并行工作站上修改数据。这样既降低风险，又简化项目管理的复杂性，且符合提高效率和充分利用硬件资源原则。由于将数据仓库看作是只读数据库，无需设立锁，更无需管理锁，能大大简化数据库的并发控制，改善数据的可用性和数据的安全性。且因为成绩数据库分为四部分存放，查询或统计时可同时进行，提高了速度，并且在故障恢复能力上得到质的飞越，数据的存储能力也大为改观。由于数据仓库与联机事务处理类的系统分开，网络服务器的负荷大大减少，不再成为系统的瓶颈。

二、系统总体设计

1.硬件设计

利用一台 586 微机作为专用服务器，四台 586 微机

工作站作专用并行机，若干台 586 工作站或局域网作分布式用户：



2.总系统划分

分布式并行数据库系统的目标是以最小的代价、最有效的方式、最大程度地满足用户的需求。它大致有如下四个重要标准：局部性的处理功能达到最大、有利于负载的分布、满足分布数据的可用性和可靠性要求、有足够的可用存储器容量。因此为实现上述目标，整个系统共划分为三个子系统：系级成绩管理系统、校级成绩管理系统和并行管理系统。

(1)系级成绩管理系统：这是为系管理部门设计的一个功能完备的子系统，系统放置于系本地工作站或服务器，存放有在校本系学生的全部数据。系统按照多用户环境要求设计，其工作方式有两种：一种是完成本系原始数据的装入、修改、处理等工作。另一种工作方式是将装配有整套系统的某台微机(一般为指定工作站)做为整个校级分布并行系统的一个远程用户联入校级系统，每学期向校级并行工作站系统传送一次有关的成绩数据以及历史记录，并对校服务器主数据库与并行工作站数据库相关数据进行一致性与完备性检验。

(2)校级成绩管理系统：是为学校教务部门设计的多用户并行化系统，系统放在服务器上，各工作站通过服务器和并行机的数据工作，具体运行依赖于并行管理系统。主要功能是对全校的成绩、学生、课程等进行管理。

(3)并行管理系统：这是为上面两个系统实现并行化服务的一个子系统，系统程序放置在工作站上或服务器里，主要数据库是成绩库放在并行工作站内，系统启动后不断地访问服务器里的并行进程库，一旦取得指令后立即工作，工作完成后又到服务器中取另一条指令进行下一步工作。它的主要功能是具体实现成绩的接收、查询和统计等工作，将用户需要的数据传递给服务器，然后用户到服务器上读取或重新加工。

3.分布并行数据库设计

并行数据库的实现是靠数据存放在不同的存储器内实现并行化管理的，为此本系统数据库存储按局部共享和无共享混合结构，主要划分为四类数据库：系工作站数据库、服务器主数据库、并行工作数据库和并行工作站数据库。

(1)系工作站数据库：系级成绩管理系统的全部数据库，系所有的专业库、学生库、课程库、成绩库等，仅存放在本地工作站，系统对这些数据库实现增、改、删等管理，并向服务器主数据库和并行工作站数据库传递数据，其中学生库和成绩库一般仅存放在校学生的信息，根据存储器的容量确定，每学期的成绩向并行工作站数据库传送一次。

(2)服务器主数据库：校级成绩管理系统除成绩库外的全部数据库，全校的专业库、学生库、课程库等，均存放在主服务器内，这些库也是共享库。其中成绩库、学生库、班级库通过系级成绩管理系统向服务器传送，这些库的管理实际上是由系级管理系统完成，其余库由校级数据库管理系统来完成。

(3)并行工作数据库：这些库是为实现系统分布并行式管理设计的，全部都放在服务器硬盘内，有并行进程管理库、并行成绩工作库、查询工作库等，这些库内的数据为临时交换使用，一旦用过立即清除。

(4)并行工作站数据库：由于成绩库处理的记录量特别大，为了实现数据库并行管理，减少并行工作站之间的数据传递，将成绩数据库按大粒度存放，如分系存放，每个并行工作站仅存放数个成绩库并由并行管理系统管理(全部数据库格式：篇幅所限省略)。

4.并行工作设计

采用模块成对设计原理，利用并行进程管理库和其他工作库实现一个功能的运行，即一个模块分两块制作，一块为系级或校级成绩管理系统的前端(功能)模块，一块为并行管理系统的后端(功能)模块，它们之间靠数据的传递联结工作。首先介绍一下并行进程管理库 BXJCK 的结构和工作机制，它放在服务器存储器上，结构为一个二维表：

登记号	并行机号	功能名称	功能号	工作站号	标志号	状态说明
1	1000	接收成绩数据	1	12	2333	执行完毕
2	0110	查各系不合格成绩	10	4	3113	正在执行
3	1111	统计学期需补考名单	16	5	1001	执行等待

工作机制：由工作站前端模块向 BXJCK 发出请求，检查请求功能号是否存在，如无则增加一条空记录，并将得到的一个登记号，以及并行机号、功能号、工作站号和标志号 0 写入，同时将并行数据写入有关工作库，当一台并行工作站完成一个功能执行后，向 BXJCK 搜索最先登记的标志号为 0 的记录，后端模块取到登记号和功能号后，置标志号为 1，同时从工作库取数据开始执行，并将完成的结果送到服务器相应的工作库中，将 BXJCK 中该记录中相应的并行机标志号置为 2，工作站前端模块从 BXJCK 中取得标志号后，从服务器相应的工作库中取得数据后，并完成最后功能，清除服务器相应的工作库中的数据，同时将 BXJCK 中该记录删除。工作原理如下：

(1)一个前端模块一次只能向 BXJCK 提出一个调用后端模块的请求；

(2) 同一个用户工作站可同时有多个不同应用前端模块向 BXJCK 提出调用不同后端模块的请求，因此用户提出请求后，若不能得到立即响应，可转到其他应用前端模块，过一段时间再过来查看是否调用完毕；

(3) 不同用户工作站可以同时有同一应用前端模块向 BXJCK 提出同一调用后端模块的请求；

(4)标志号均变为 2 时表示这一功能执行完毕；

(5) 在 BXJCK 的所有记录中，一个并行工作站只在一条记录中对应的标志号位为 1，因为一个并行工作站一次只能调用一个功能；

(6) 并行优先原则：若一个前端模块 BXJCK 中功能记录的标志号为两位以上，其中一位是 1 时，其余位将在下一后端模块优先调用，如标志位是 1001，则 2 和 3 号并行工作站执行完当前功能后，立即置为 1111。这样一个功能一旦调用则尽快执行完。

(7) 功能优先原则：动态估算功能执行时间，时间短的优先安排执行。

(8) 用户优先原则：对用户的权限进行处理，权限级别高的在调用同一后端模块时先响应。

三、系统结构设计

1.系级系统结构设计

本系统是一个系级学分制学生成绩管理系统。它是全校性分布并行式成绩管理系统的三大子系统之一，是一个功能完备而又具有相对独立性的小型多用户 MIS 系统。一种工作方式是既能完成本系原始数据的装填、修改，并对本系学生成绩等相关数据依据学分制有关规定进行处理，同时系统又能比较迅速地响应本地多用户不同需求的查询、统计以及产生与此有关的各种辅助性打印报表等。系统工作如图 1 所示。

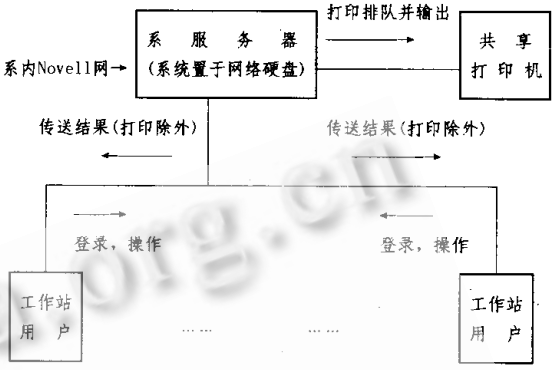


图 1

另一种工作方式即与校级另外两套系统实现无缝联接。其工作方式如图 2。将装配有整套系统(含全部有效数据)的某台微机(一般为指定工作站)做为校级分布并行系统的一个远程用户联入校级系统，一方面每学期向校级并行工作站系统主数据库(成绩库)传送一次有关的成绩数据以及历史记录；另一方面可向校级系统发送本地查询请求，有条件查询校级系统或并行工作站中的有关数据。

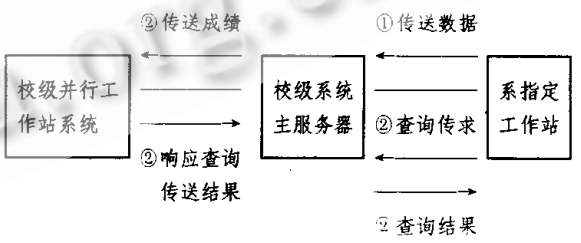


图 2

2.校级系统结构设计

校级管理系统负责全校性数据的管理，它的大致功能与系级管理系统相同，但它由于并行机制的存在，物理结构的不同，对数据处理的流程也不同。校级管理系统的结构框图如图 3。

3.并行系统结构设计

并行管理系统主要是完成对并行工作站的管理，它

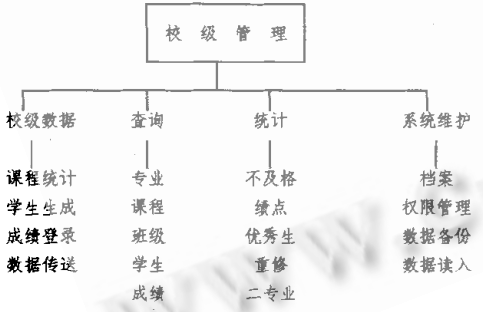
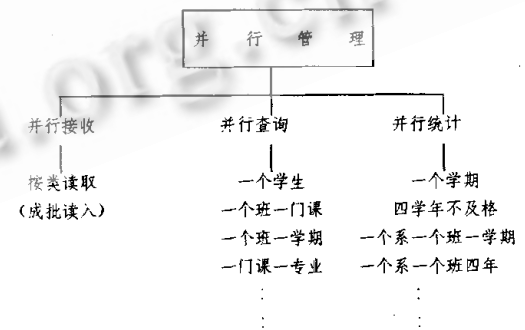


图 3



可分为接收数据、并行查询、并行统计三个功能，它相对于校级成绩管理来说是后端，并行系统结构设计如下：

参考文献

[1] 黄璇，数据库技术的发展方向，计算机工程与应用，1995,31(5),1－5

[2] 杨利，周兴铭，郑若忠，并行数据库系统的体系结构，计算机科学，1994,21(4),42－46

[3] 杨利，周兴铭，吴涛，并行查询中的进程分配与调度，计算机科学，1995,22(6),26－29

[4] 郭宜斌，数据仓库技术的基本概念和发展现况，PC世界，1996,4,26－31

（来稿时间：1997 年 7 月）