浙 江 林 学 院 学 报 2006, **23**(4): 435~438
Journal of Zhejiang Forestry College

文章编号: 1000-5692(2006)04-0435-04

分布式存储的木材运输证数据集成研究

吴达胜[1], 黄美丽[1], 陈 飞[2]
(1. 浙江林学院 信息工程学院, 浙江 临安 311300; 2. 浙江省临安市锦城街道 办事处, 浙江 临安 311300)

摘要: 分布式数据存储与集成是数据安全管理的重要途径, 在森林资源监测及流通管理领域具有良好的应用前景。在林业管理中, 资源监测与电子办证涉及地域广, 研究数据挖掘技术在运输证管理中的应用具有重要意义, 分布式数据集成是对分布存储于异地的海量数据进行数据挖掘的基础。探讨了分布式存储的木材运输证数据集成模拟系统的建立过程, 包括基础实验数据的建立, 数据仓库逻辑模型的设计, 数据集成前的数据预处理, 木材运输证数据仓库的数据集成等。图 1 参 10

关键词: 森林经理学; 木材运输证; 数据集成; 数据仓库
中图分类号: S757 文献标识码: A

随着森林资源信息管理的发展[1], 基于 Web 的林证管理系统已在浙江省许多县(市), 如新昌县、临安市等积极运行, 数据管理领域的高级技术 "数据挖掘"[2~4] 是林政管理部门即将面临的需求, 在林证管理(主要是运输证、林权证、采伐证)中以木材运输证所涉及地域最广, 木材运输证的管理正逐渐步入正轨[5,6], 研究数据挖掘技术在木材运输证管理中的应用具有代表意义。如果省林业厅需要对各县(市)运输证数据进行汇总分析, 并采用数据挖掘技术寻找潜在规律, 这一需求的前提是如何有效地集成分布存储于各县(市)中的木材运输证数据, 关键问题就是要建立全局数据仓库, 并将各县(市)数据进行清洗、转换与加载。SQL Server 的 DTS (数据转换服务)具有强大的数据抽取、清洗、转换、导入/导出功能[7~10]。所述的分布式存储的木材运输证数据集成模拟系统使用 DTS 作为开发工具。

1 基础试验数据建立

为了模拟系统的分布式应用, 试验过程为: 在一个局域网中的多台数据库服务器上建立相应数据库, 考虑到工作量等原因, 试验数据仅存储于 2 台服务器上, 其中一台代表的是临安木材运输证数据库服务器, 另一台代表的是萧山木材运输证数据库服务器。数据库名称均为 "运输证", 结构一致, 数据独立, 由 4 个表组成。表结构如下: 运输证(编号, 性质, 发证地, 发货单位 ID, 收货单位 ID, 运输方式, 期限, 签发日期); 运输物(编号, 树种, 品名, 根数, 材积, 折合原木材积); 发货单位[发货单位 ID, 发货单位名称, 县, 乡(镇)名]; 收货单位(收货单位 ID, 收货单位名称, 省, 县)。

2 数据仓库逻辑模型设计

数据仓库结构的基础是分布式数据库的全局模式, 同时还要加入必要的键列属性并将源数据库表

收稿日期: 2005-07-06; 修回日期: 2006-03-27
基金项目: 浙江省教育厅资助项目(2411005023)
作者简介: 吴达胜, 副教授, 硕士, 从事森林资源信息管理与信息系统等研究。E-mail: wu62380710@263.net

结构进行适当集中或分解。在考虑维度与事实分析基础上，定义了 3 个基本表结构：发货单位维度表(发货单位序号，发货单位 ID，发货单位名称，县，乡镇名)；收货单位维度表(收货单位序号，收货单位 ID，收货单位名称，省，县)；运输证事实表(运输证序号，编号，性质，发证地，发货单位序号，收货单位序号，运输方式，期限，签发日期，树种，品名，根数，材积，折合原木材积)。其中的发货单位序号、收货单位序号和运输证序号是新定义的列，这些列的值由两部分内容组成：其一为县级林业局编号(该编号由省林业厅统一制定)；其二为单调递增的序列号，以保证该列值在任何情况下都惟一，同时又能明确地知道发货单位、收货单位和运输证来自于哪一个县(市)的林业局，确保键列的惟一性和语义的明确性，如运输证序号“0100001”代表临安市 1 号运输证，“0200001”代表萧山市 1 号运输证等。原因主要是考虑到在同一个数据库中可以保证每个表中的码(如发货单位 ID，收货单位 ID，运输证 ID)惟一，但在多个数据库数据集成后可能导致原码值不惟一的现象，如临安的发货单位 ID 与萧山发货单位 ID 是有可能重复的，为此，数据仓库设计中需要产生新的主键。

3 数据集成前的数据预处理

在各个分布式数据库中，不一定能够保证所有的数据都满足一致性要求，因此在数据被加载到数据仓库前，要对基础数据的一致性进行检查，以保证所有加载入数据仓库的数据都能满足实体完整性和参照完整性等要求，最简单的作法是应用 SQL 语句查询不满足上述要求的数据，并进行删除。

4 木材运输证数据仓库的数据集成

由于 SQL Server 2000 的数据转换服务(DTS)易用，功能强，实验就在这一环境中展开。

4.1 建立连接

连接指向数据源和数据目的，由于一个数据源只能是一个表或一系列的 SQL 语句(执行结果也是一个表)，原系统中的发货单位、收货单位数据需要导入到数据仓库中的 2 个维度表(发货单位维度表和收货单位维度表)，其余两个表(运输物、运输证)数据要导入到数据仓库中的事实表(运输证事实表)，因此系统设计中对每个数据库数据导入至少需要建立 3 个源连接和一个目的连接(图 1)。

4.2 执行数据转换任务

对于每个从源到目的的数据转换都要建立一个数据转换任务(图 1)，以实现单个源到目的的数据转换任务。在这一步骤中，维度表的转换是简单的，只要选择“源”和“目的”表，建立源到目的列的相应转换映射就可以了。事实表的数据转换任务是本设计中的一个技术难点，由于涉及到源数据库中的“运输物”“运输证”的事实数据及目的数据仓库中的“发货单位维度表”及“收货单位维度表”的键列数据，因此要实现数据库与数据仓库中的 4 个表的连接。SQL 语句如下：

```
select 运输证. *, 运输物. 树种, 品名, 根数, 材积, 折合原木材积, 运输证数据仓库. DBO. 发货
单位. 发货单位序号, 运输证数据仓库. DBO. 收货单位. 收货单位序号
from 运输证
inner join 运输物
on 运输证. 编号= 运输物. 编号
inner join 发货单位
on 发货单位. 发货单位 ID= 运输证. 发货单位 ID
inner join 运输证数据仓库. DBO. 发货单位
on 运输证数据仓库. DBO. 发货单位. 发货单位 ID= 发货单位. 发货单位 ID
and 运输证数据仓库. DBO. 发货单位. 县= 发货单位. 县
inner join 收货单位
on 收货单位. 收货单位 ID= 运输证. 收货单位 ID
inner join 运输证数据仓库. DBO. 收货单位
on 运输证数据仓库. DBO. 收货单位。收货单位 ID= 收货单位. 收货单位 ID
and 运输证数据仓库. DBO. 收货单位. 县= 收货单位. 县
```

以上的转换是将位于本地的数据库导入到数据仓库中，其转换相对容易。

另一个技术难点在于如何将不同数据库服务器上的数据集成到本地数据仓库中。在这个过程中，维度表的转换也相对容易实现，只需在定义源连接时，将数据库服务器连接到萧山运输证数据库服务器即可。在连接属性中，笔者将数据库服务器设置为代表萧山运输证数据库服务器的 HML，其他的设计过程与本地数据库的数据导入是一样的。在解决异地数据库事实表数据导入过程中，笔者遇到了一个难题，即在使用 SQL 语句产生数据源时需要使用以下格式进行访问(服务器名称．数据库名．数据库拥有者．表名)，同时在访问列时，还要在末尾加上“．列名”，这样的操作在 SQL Server 中是不允许的，因为它规定这样的引用不可以超过 4 级。为此，笔者将 HML 服务器上的运输物表和运输证表分别取了别名。这个转换与临安运输证数据库事实数据转换略有不同。代码如下：

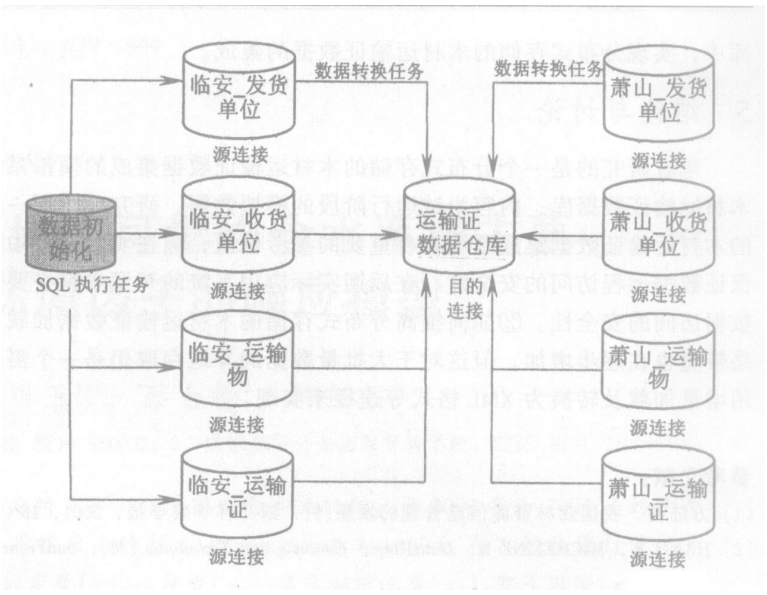

图 1　木材运输证数据仓库

Figure 1　Timber transportation licence data warehouse

```
select 运输证 1.　*，运输物 1. 树种，品名，根数，材积，折合原木材积，
发货单位．发货单位序号，收货单位序号
from HML．运输证．DBO．运输证 as 运输证 1
inner join HML．运输证．DBO．运输物 as 运输物 1
on 运输证 temp．编号= 运输物 temp．编号
inner join 发货单位
on 发货单位．发货单位 ID= 运输证 temp．发货单位 ID
inner join 运输证．DBO．发货单位
on 运输证．DBO．发货单位．发货单位 ID= 发货单位．单位 ID
and 运输证．DBO．发货单位．县= 发货单位．县
inner join 收货单位
on 收货单位．收货单位 ID= 运输证 temp．收货单位 ID
inner join 运输证．DBO．收货单位
on 运输证．DBO．收货单位．收货单位 ID= 收货单位．收货单位 ID
and 运输证．DBO．收货单位．县= 收货单位．县
```

4.3　建立 SQL 执行任务

为了保证每次数据加载都与源数据库中数据一致，需要建立 1 个 SQL 执行任务。作用是实现数据初始化(图 1)，即将数据仓库中表内容清空。

4.4　定义工作流属性

以上步骤仅仅建立了单个执行任务，其间的操作流属性并未定义，系统不能明白哪步工作先做哪步工作后做，工作流属性就要完成定义执行步骤的顺序。我们把“数据初始化”定义为第 1 次执行任务，本地数据库的维度表数据转换定义为第 2 次执行任务，本地事实数据导入为第 3 次执行任务，异地(指萧山)数据转换任务必须在以上任务执行的基础上再执行。

4.5　运行 DTS

在已经定义了以上的 DTS 后，只要运行 DTS 就能将 2 个分布于异地的数据库内容导入到数据仓

库中，实现分布式存储的木材运输证数据的集成。

5 结论与讨论

笔者研究的是一个分布式存储的木材运输证数据集成的模拟系统，数据库的结构来源于已运行的木材运输证数据库，内容为试运行阶段的模拟数据，研究对于下一步开发具有实用价值的分布式存储的木材运输证数据集成系统有着重要的参考价值，但在实际运行中需要进一步解决以下问题：①如何保证数据远程访问的安全性。在后期实际应用系统的开发中有必要采用对数据审计、加密等手段确保数据访问的安全性。②如何提高分布式存储的木材运输证数据加载的性能。随着硬件技术的发展，网络带宽也在逐步增加，但这对于大批量数据的异地存取仍是一个需要重点考虑的问题，可以将数据采用增量加载及转换为 XML 格式等途径来实现。

参考文献：

[1] 方陆明. 我国森林资源信息管理的发展[J]. 浙江林学院学报，2001，**18**(3)：322－328.
[2] HAN J W，MICHELINE K. *DataMing*：*Concepts* and *Techniques*[M]. SanFrancisco：Morgan Kaufman Pubishers，2001：10－120.
[3] 吴达胜，方陆明，唐丽华，等. XML 技术在森林资源信息管理系统异构数据集成中的应用[J]. 浙江林学院学报，2003，**20**(4)：403－407.
[4] 吴达胜. 分布式数据挖掘在森林资源信息管理中的应用[J]. 福建林学院学报，2004，**24**(4)：340－343.
[5] 许岚. 木材运输证是运输木材的惟一合法凭证——评析一起纤维板运输行政纠纷案[J]. 林业资源管理，2001(3)：15－18.
[6] 易超. 木材检查站发放木材运输证在森林资源管理中的作用[J]. 江西林业科技，2004(6)：53－54.
[7] 杨飞. 基于 DTS 对象模型的数据转移的实现[J]. 电脑与信息技术，2004(5)：27－30.
[8] 王胜德，杨学强. 利用 DTXS 实现异构数据库的数据交换[J]. 计算机应用，2003，**23**(7)：132－134.
[9] 武彦峰，朱仲英. 基于 DTS 组件的数据仓库的数据抽取工具的设计与实现[J]. 微型电脑应用，2004，**20**(3)：1－5.
[10] 程晓华，胡凯. 数据转换服务(DTS)在话费网络查询系统中的应用[J]. 南昌大学学报：工科版，2004，**26**(2)：97－100.

Integration of distributed data of timber transportation licence

WU Da-sheng[1]，HUANG Mei-li[1]，CHEN Fei[2]

(1. School of Information Engineering，Zhejang Forestry College，Lin'an 311300，Zhejiang，China；2. Office of Jincheng Town，Lin'an City，Lin'an 311300，Zhejiang，China)

Abstract：The storage and integration of distributed data，an important way of safety management，has an optimistic application prospect in the area of monitoring and flowing management of forest resources which involves management of a great deal of distributed data. Meanwhile，Data Mining，the advanced technology in data management，is the requirement with which the departments of the forest administration are about to be faced. Therefore，research on the application of data mining technology in the management of transportation licence is representative. Furthermore，the integration of distributed data is the bases of data mining of distributed mass data. In this paper，the process of building the simulate system of the timber transportation licenceis discussed，which in cludes setting-up of the basic experimental data. Logical model design of the data warehouse，data preprocessing before integration，the integration of the data warehouse of timber transportation licence. [Ch，1 fig. 10 ref.]

Key words：forest management；timber transportation licence；data integration；data warehouse