

# 基于异构数据库的税务信息查询系统

## Design and implement of taxation information system based on heterogeneous database

高 剑 (浙江大学 310027)

**摘要:** 税务信息查询系统是一个提供纳税资料基础信息、并辅助领导进行决策的综合查询平台。本文详细介绍该系统的设计与实现,并且着重分析了系统采用的基于异构数据库的数据抽取和数据更新技术。

**关键词:** 异构数据库 数据抽取

目前,税务系统各个应用系统大多采用不同机器、不同操作系统,不同的数据库平台,它们存在着重复劳动、应用程序水平层差不齐、需要手工干预、自动化程度不高、无法实现数据共享等问题。

随着税务信息化建设的推进,各应用系统之间进行业务整合和数据集中的要求日益迫切,为此,充分

监控,为领导层提供查询分析并作出决策服务。

### 1 系统平台

通过数据分析,以及对现有资源的调查了解,决定采用如下系统平台:

(1) 采用基于浏览器的三(N)层体系结构,应用

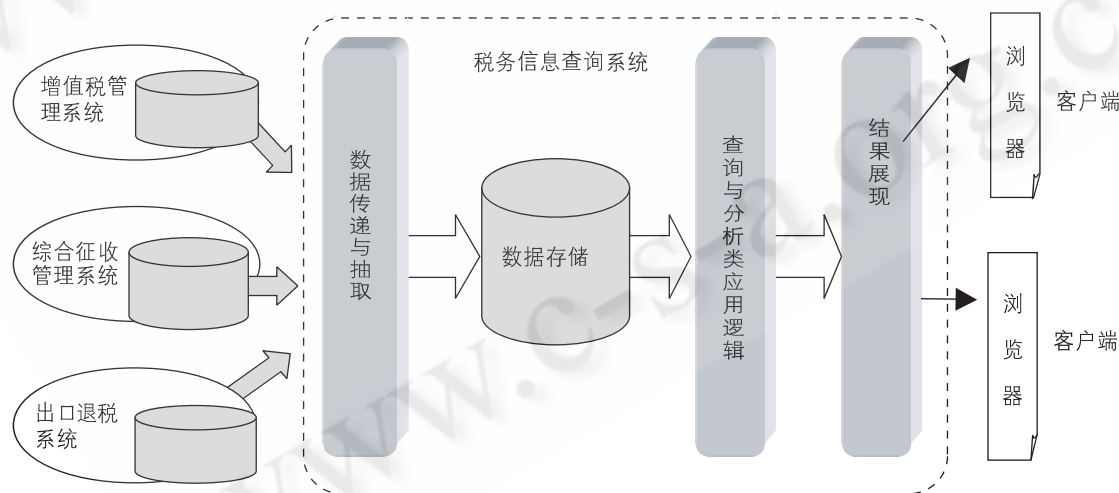


图 1

利用现有资源进行系统整合是当前信息化建设的一项主要任务。通过税务信息查询系统,一是满足基层税务机关人员执法的需要,使基层税务机关人员能够全面、快速和准确地查询出纳税人的全面涉税信息,为基层执法服务。二是满足各级税务管理机关管理方面的需要,使管理层能够对纳税人的涉税活动进行全面的

服务器采用 WEBLOGIC。

(2) 支持多种数据源,实现跨数据库、跨平台的数据抽取。

(3) 数据库采用 ORACLE 8i。

(4) 运行环境采用小型机 (IBM S80) 作为数据库服务器,PC SERVER (Dell PowerEdge) 作为中间层应用

服务器,PC(联想 开天)作为 Oracle Transparent Gateway for SQL Server 服务器。

## 2 系统设计

### 2.1 系统总体架构

税务信息查询平台从逻辑功能上可分为三部分:

数据抽取、数据存储、数据展现。如图 1 所示。

### 2.2 数据分析

现有综合征收管理系统、增值税管理系统、出口退税系统是基于不同操作系统、不同数据库平台,对纳税人涉税行为信息的查询功能是分布在不同的应用系统中。具体见表 1。

表 1

系统	数据库	操作系统	数据分布方式	数据增量周期
综合征收管理系统	SYBASE	Solaris、AIX	为地市级集中	实时
增值税管理系统	ORACLE	UNIX	为地市级集中	实时
出口退税管理系统	SQL SERVER	WINDOWS 2000	为地市级集中	实时

### 2.3 数据抽取

由于综合征管系统、增值税管理信息系统、出口退税系统等税收业务系统采用的数据库、开发平台差异较大,而且各系统的数据集中程度也不同,数据抽取就成为税务综合查询系统实现中的关键工作。

数据抽取的方式一般有复制、使用抽取工具、中间文件或中间库的导出/传输/导入、数据操作记录的重写等。其中数据操作记录的重写又有几种方式,包括:redo log 转为 SQL、用触发器记录变化等。各种方式初步分析如下:

(1) 数据库复制。此方法就是对同样的数据库平台之间使用数据库产品本身自带的复制功能,实现源到目标的数据抽取。对于跨数据库平台的数据复制,可以采用数据库厂商的复制网关或代理产品来实现。

利:可实现增量复制;自动程度较高,实时性强,效率较高。

弊:不同数据库平台间的复制比较复杂,需要专门的网关或代理产品;源与目标数据的映射关系简单,需要较好的网络条件。

(2) 使用数据抽取工具。此方法是使用现成的数据抽取工具产品。一般此类工具都能够定义不同的数据源和目标、定义源数据到目标数据的复杂的映射关系(包括计算关系),并实现定时的自动抽取。

利:可实现复杂的源与目标数据之间的映射关系;自动程度较高;适用于数据仓库的应用。

弊:自身不能实现增量复制,增量复制有赖于源系

统的结构;需要专门购买产品。

(3) 中间文件或中间库的导出/传输/导入。此方法是不同系统间数据传递的最普遍的方法。一般需要根据目标数据定义一个标准的中间文件或中间库的格式(此格式一般包含格式规定和数据项的规定),并以中间文件作为数据传递的载体实现不同或远程的数据转入。

利:不受平台和系统的限制,较为通用;可以不用专门的产品支持;可实现较为复杂的映射关系;对不同平台间的数据类型转换的实现比较灵活。

弊:实时性较弱;数据转出、数据转入需要专门处理,抽取质量有赖于编程质量;增量复制有赖于源系统的结构。

(4) 日志转为 SQL。此方法即在源系统中将日志转为 SQL 文件,然后将此 SQL 文件传输到目标系统中,再执行相应的 SQL 命令,从而实现源与目标数据的一致。实际上数据库复制就是采取此方法。此方法平时并不常用。

利:可实现增量复制。

弊:实时性弱;不是每个数据库平台都支持,需要工具;需要操作人员有较强的能力。

(5) 触发器记录变化,在目标端重现。此方法其实是对使用中间文件和中间库转数据的一个改进。它可以实现增量数据转入,即在源系统中加入触发器,将数据库中的变化记录在中间库中。根据中间库的记录在目标数据库重现数据修改。

利:可实现增量复制。

弊:需要修改源系统,并对源系统的性能有影响。

影响抽取效率的主要因素有两个:一是抽取的方式,尤其是面对综合征管系统这样数据量较大的系统,每次抽取采用全面抽取,数据量巨大,一次抽取的时间必然很长,对于间隔较短的抽取,应采取增量抽取的方式。增量抽取最直接的技术就是复制,对于跨平台的抽取来说,最适宜的方式则是增量数据的导出。这就需要考虑影响抽取效率因素之二即源系统

本身的结构是否支持增量抽取。增量数据导出需要源系统中的标有明确操作时间字段并且无物理数据删除操作,或者在源系统中增加触发器记录增量数据或操作。

对各数据抽取方式的综合比较,就通用性、效率、灵活性、实时性而言,数据库复制方式实现了增量数据抽取的工作自动化,而且效率较高,保证了系统运行的实时性,应该是首选。具体针对各不同数据库应用系统决定采用抽取方式如下表 2:

表 2

系统	数据库	抽取工具与方式	抽取周期
综合征管系统	SYBASE	1. 通过 SYBASE 复制网关服务器 DirectConnet 实现数据库复制 2. 通过 ORACLE 网关服务器 GateWay 调用存储过程进行数据抽取	实时或定时
增值税管理系统	ORACLE	ORACLE 数据库复制(快照)	实时
出口退税系统	SQL SERVER	通过 ORACLE 网关服务器 GateWay 调用存储过程进行数据抽取	实时或定时

## 2.4 数据存储

由于税务信息查询系统是从各种数据采集系统中抽取数据,所抽取和存储的数据都是不断变化着的,在逻辑存储上不采用数据仓库的存储方式,而采用一般的数据库存储方式;同时,与数据采集系统不同,平台所实现的功能仅限于对数据的查询,没有对数据的更新与修改,并且要对抽取到的明细数据进行汇总加工,因此数据又具有一定的历史性。

数据的结构以综合征收管理系统的数据库表结构和设计原则作为主要参照标准,并结合其他需要抽取的数据,确定规范的主题数据表。按照税务信息查询系统的需要确定统一标准。

## 2.5 数据展现

数据展现实现税务信息查询系统所要求的查询和分析的功能。其功能是在从不同的数据采集系统中抽取数据并整理存储后,实现查询分析的应用逻辑,并将结果在浏览器端展现出来。技术上采用符合 J2EE 标准的三层体系结构,通过 JSP + JavaBean 实现。无需安装客户端软件,可直接实现上级对下级的监控。

## 3 具体实现

下面以出口退税系统为例具体阐述数据抽取及更新的实现。

由于源数据库与目标数据库异构,因此需采用数据库网关或自行编程实现。安装数据库网关:使用 Oracle Transparent Gateway for SQL Server 在源和目标数据库间建立 DBLINK,可以实现数据库级的直连。其优点是对应用透明,编程量小;缺点是安装及配置较复杂。

### (1) 透明网关参数配置

在 C:\OraGateway\tg4msql\ADMIN\下复制原有的 inittg4msql.ora 文件到新文件 initckts.ora。修改 listener.ora 文件增加监听,文件路径是 C:\OraGateway\network\ADMIN\listener.ora,增加下面黑体的部分。

```
# LISTENER.ORA Network Configuration File: c:\OraGateway\network\admin\listener.ora
```

```
# Generated by Oracle configuration tools.
```

```
LISTENER =
```

```
(DESCRIPTION_LIST =
```

```
(DESCRIPTION =
```

```
(ADDRESS_LIST =
```

```
(ADDRESS = (PROTOCOL = TCP) (HOST = IP_ADDRESS) (PORT = 1521))
```

```
)
```

```
(ADDRESS_LIST =
```

```
(ADDRESS = (PROTOCOL = IPC) (KEY = EXTPROC1))
```

```

)
)
)
SID_LIST_LISTENER =
(SID_LIST =
(SID_DESC =
(SID_NAME = PLSExtProc)
(ORACLE_HOME = c:\OraGateway)
(PROGRAM = extproc)
)
(SID_DESC =
(SID_NAME = ckts)
(ORACLE_HOME = c:\OraGateway)
(PROGRAM = tg4msql)
)
)
)

```

IP\_ADDRESS 为税务信息系统数据库所在主机的 IP 地址。

修改完毕后在 Windows 服务列表中重新启动监听服务: OracleOraHomeGatewayTNSListener

(2) 目标数据库的 tnsnames.ora 配置。在目标数据库增加访问出口退税审核系统数据库的 tnsname, 按以下步骤操作(以 UNIX 平台为准, Windows 平台参照进行):

① 以 oracle 用户登录数据库服务器;

② 修改 \$ORACLE\_HOME/network/admin/tnsnames.ora 文件, 在最后增加以下代码, 加入出口退税审核系统的 TNS:

```

ckts =
(DESCRIPTION =
(ADDRESS_LIST =
(ADDRESS = (PROTOCOL = TCP)(HOST = IP_ADDRESS_OF_CKTS)(PORT = 1521))
)
(CONNECT_DATA =
(SERVICE_NAME = SID)
)
)
)

```

其中黑体字为 tnsnames, IP\_ADDRESS\_OF\_CKTS 为出口退税数据库的 IP 地址、1521 为监听端口号、SID 为

数据库网关所配置的 SID\_NAME。

③ 使用 `tnsping ckts` 命令测试 tnsnames 配置是否正确, 如果返回信息中包括类似以下的内容, 表示配置成功:

```

Attempting to contact ( ADDRESS = ( PROTOCOL =
TCP)( HOST = 1. 1. 1. 1)( PORT = 1521))
OK ( 20 msec)

```

(3) 在目标数据库上建立 DBLINK

建立指向“出口退税审核系统”的 DBLINK, 名称为“CKTS”:

```

CREATE DATABASE LINK CKTS
CONNECT TO USER_NAME
IDENTIFIED BY PASSWORD
USING 'ckts'

```

其中的 USER\_NAME 和 PASSWORD 是出口退税审核系统为税务信息系统数据抽取建立的用户。

在出口退税审核系统数据库中给税务信息系统建立的用户赋权后, 测试通过 DBLINK 是否能够正确访问数据, 执行命令 `select count(*) from TABLE_NAME@ckts`, 如果能够返回查询结果表明配置成功。

## 4 总结

在技术发展上, 数据抽取所涉及的单个技术环节都已相对成熟, 其中有一些是躲不开编程的, 但整体的集成度还很不够。目前市场上所提供的大多是数据抽取工具。这些工具通过用户选定源数据和目标数据的对应关系, 会自动生成数据抽取的代码。但数据抽取工具支持的数据种类是有限的; 同时数据抽取过程涉及数据的转换, 它是一个与实际应用密切相关的部分, 其复杂性使得不可嵌入用户编程的抽取工具往往不能满足要求。因此, 实际的实施过程中往往不一定使用抽取工具。整个抽取过程能否因工具的使用而纳入有效的管理、调度和维护则更为重要。

在我们这个税务信息系统中较好的解决了源数据库和目标数据库之间的数据抽取问题, 利用成熟的数据库网关产品实现了异构数据之间的数据共享, 为以后更深层次的数据利用、数据挖掘从技术上开辟了一条可行的道路。

## 参考文献

1 Oracle8i Database Online Document Release