数据仓库建设中的数据预处理

Data-processing in the building of data warehouse

李瑞欣 张水平（西安 空军工程大学电讯工程学院 710077）

摘要：提出了数据仓库建设中数据处理阶段遇到的若干问题，针对问题进行了分析，并提出了解决问题的方法。

关键词：数据仓库 数据集成 数据转换服务

1 引言

数据仓库的数据是从原有的分散的数据库数据中抽取出来的。操作型数据与DSS分析型数据之间差别甚大。第一，数据仓库的每一个主题所对应的源数据在原有的各分散数据库中有许多重复和不一致的地方，且来自不同联机系统的数据都和不同的应用逻辑捆绑在一起；第二，数据仓库中的综合数据不能从原有的数据库系统直接得到。因此在数据进入数据仓库之前，必须要经过统一与综合，这一步是数据仓库建设中最基础、最关键、最复杂的一步，这一点可以从数据仓库体系结构图上得知。

所要完成的工作有：

(1) 要统一源数据中所有矛盾之处，如字段的同名异义、异名同义、单位不统一、字长不一致等。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(2) 进行数据的综合和计算。数据仓库中的许多数据是从不同的源数据库或外部文件通过计算或综合生成的。

作者在数据仓库建设中担负着数据预处理的工作，在开发过程中遇到了许多的问题。这些问题有：

① 数据类型的转换
② 数据度量单位的转换
③ 同名异义字段的转换
④ 异名同义字段的转换
⑤ 数据字段间计算关系的矛盾
⑥ 数据字段空缺的解决方法
⑦ 大小写的转换
⑧ 数据字段值本身错误的校正

2 解决的方法

在数据处理的过程中，主要运用了SQL Server 2000中的DTS (Data Transformation Service) 功能。DTS是SQL Server所带的实用程序，提供输入、输出与转换数据的功能，它可以支持自然OLE DB SQL Server驱动器传输SQL Server数据之类的关系型数据。DTS还支持非关系型数据源，如文本文件、Web页面以及Foxpro或Paradox之类的ISAM数据库。DTS基于OLE DB结构，该结构使得数据仓库设计人员可以从不同的数据源复制和转换数据。例如：

(1) 直接从SQL Server和Oracle，使用本机 OLE DB 提供程序。

(2) 从ODBC源，使用 ODBC的Microsoft OLE DB 提供程序。

(3) 从 Access 2000、Excel 2000、Microsoft Visual FoxPro (r)、dBase、Paradox、HTML 和其他文件数据源。

(4) 从文本文件，使用内置 DTS 平面文件 OLE DB 提供程序。

(5) 从 Microsoft Exchange Server、Microsoft Active Directory (tm) 和其他非关系型数据源。

(6) 从第三方供应商提供的其他数据源。

2.1 DTS的体系结构及功能

DTS 转换定义被存储在Microsoft Repository、SQL Server或COM结构的存储文件中。通过OLE DB可访问相关的和无关的数据源。数据泵 (data pump) 从数据源中打开一个行集合并将每一行数据从数据源中抽取到数据泵中。数据泵运行脚本编辑功能Microsoft ActiveX (Microsoft Visual Basic Scripting Edition、Jscript) 来拷贝、确认或将数据从数据源转换到目的地。为目的单元格所赋的新值返回到泵中，并通过高速数据传输器发送到目的地。目的地可以是OLE DB、ODBC、ASCII分隔符文件和ASCII固定字段和HTML。

在DTS体系结构中，数据可以用OLE DB数据泵从数据源中抽取，并可在发送到OLE DB目的地之前选择是否转换格式。转换是一系列过程操作，在被存储在所希望的目的地之前，必须用在源行集合中。DTS数据泵提供了一个可扩展的、基于COM的体系结构，该体系结构在数据从源转移到目的地时，可以进行复杂的数据确认和转换。DTS数据泵可以在DTS 包中充分使用ActiveX脚本的功能，使得复杂过程逻辑可以用简单的、可重用的ActiveX脚本来表示。当列值通过DTS数据泵从源转移到目的地时，这些脚本可以通过所选择的脚本语言来确认、反转和转换列值。图1即为DTS的抽取、转换、录入功能表示。

2.2 具体解决方法

在SQL Server的企业管理器中先创建

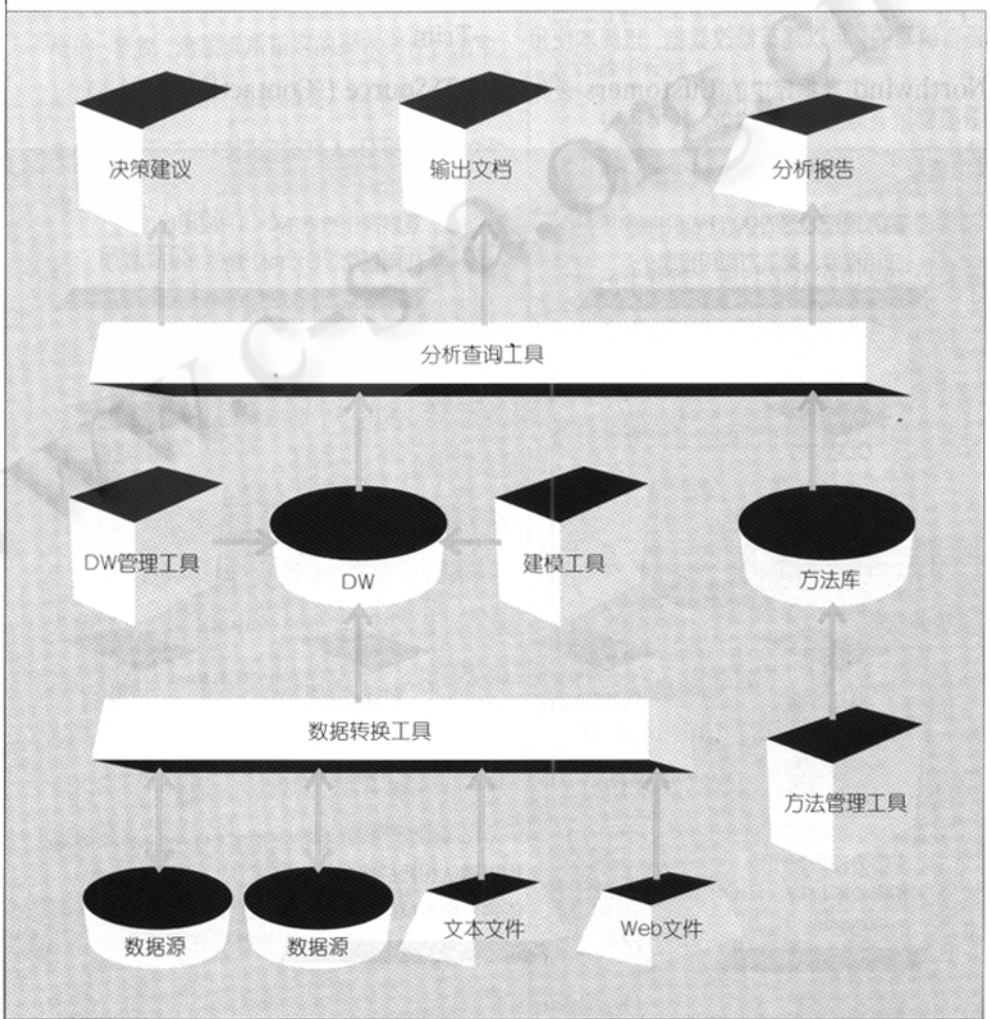


图1 将数据传送到数据仓库

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

一个数据库，选择“任务”菜单的“导入数据”栏，进入选择数据源菜单和选择导入目的菜单，设置好之后就可以进入DTS设计器进行数据的转换了。在转换中用户可以用SQL语句选择自己想要的数据字段，而舍弃那些对于数据仓库无用的字段。

在转换过程中用户可以改变字段的数据类型、字段的目的名称、大小等，使得不同来源的数据的名称、数据类型、大小等达到统一，这样就为数据的净化处理打下了坚实的基础。在使用DTS设计器进行数据转换时，系统会自动生成ActiveX脚本表达用户对数据的修改。对用户来说，也可以通过直接修改脚本的方法来达到对数据转换的目的。例如，数据的大小写的转换可以通过VB Script语言自带的大小写转换函数来实现。

下面的ActiveX脚本是用VBScript语言编写的，用以逐行修改数据。该脚本使用Northwind数据库中的Customers表作为数据源，并将数据移入Northwind数据库中的新目的表。该脚本验证源数据中的几列，并在将行插入目的表之前转换某些列数据。这种转换将Company Name改为大写字符，剪裁姓和名中的起始空格和尾随空格，并用字符串“unknown”填充Region字段(如果该字段为空)。

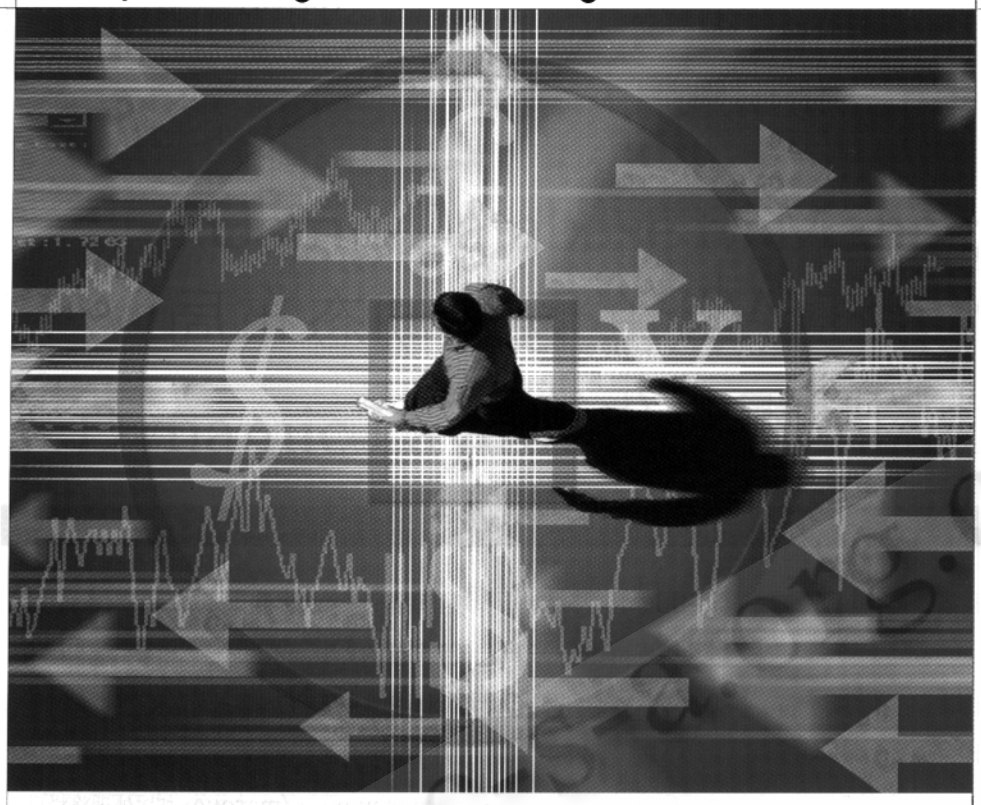
```
If DTSSource ("CompanyName") <> "" Then
DTSDestination ("CustomerID") = DTSSource ("CustomerID")
' 把小写变换成大写
DTSDestination ("CompanyName")= Ucase(DTSSource ("CompanyName"))
'
'整理ContactName前后的空格
DTSDestination ("ContactName") =Trim (DTSSource ("ContactName"))
DTSSource("ContactTitle") =DTSSource ("ContactTiltle")
DTSDestination ("Address") = DTSSource("Address")
DTSDestination ("City")= DTSSource ("City")
检查源Region域是否为空，如果为空，则以unknown填充。'
If IsNull (DTSSource("Region").value then
DTSDestination ("Region") = "unknown" Else
DTSDestination ("Region") = DTSSource("Region")
End if
DTSDestination("PostalCode") =DTSSource("PostalCode")
DTSDestination("Country") = DTSSource("Country")
DTSDestination("Phone") = DTSSource ("Phone")
DTSDestination("Fax") = DTSSource ("Fax")
Main = DTSTransformStat-OK
Else
Main = DTSTransformStat-SkipRow
End If
```

对于异名同义字段，可以先统一合适的字段名称，然后针对各个情况具体的进行转换。比如某个源表中用‘年龄’字段表示顾客的年龄，而另一个源表则用‘出生日期’字段表示顾客的年龄，如何统一两种表达方法呢？针对这种情况，认为用‘年龄’字段表示较为合适，下面是具体的转换脚本：

```
Function Main()
DTSDestination("姓名")=DTSSource ("姓名")
```

图2 将数据传送到数据仓库

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn



DTSDestination("人员编号") = DTSSource("人员编号")
DTSDestination("性别")= DTSSource("性别")
DTSDestination("年龄") = DateDiff(yy, DTSSource("出生日期").value,getdate()
DTSDestination("单位")=DTSSource("单位")
Main = DTSTransformStat-OK
End Function

对于字段空缺或者是字段间的计算关系矛盾，那么只能检查出来后人工校对或通过调查研究之后查缺补漏，使之能达到数据仓库建设的要求。

2.3 DTS数据转换功能的注意事项及功能局限性

当使用 DTS 导入导出向导和 DTS 设计器创建数据转换包时，应考虑以下事项：

如果 text、ntext 和 image 类型的字段长度超过 8388602 字节，则 DTS 的复制 SQL Server 对象任务将截断超出的部分。DTS 设计器或 DTS 导入/导出向导不显示任何错误信息，而是显示任务已成功完成。唯一的失败指示是一条写入日志文件的日志消息，此文件的名称为 <server>.<database>.log，位于"复制 SQL Server 对象任务属性"对话框的"复制"选项卡所指定的脚本文件目录中。此日志消息详细说明表和列，但不指出发生截断的行。无任何错误记录写入 DTS 错误文件或 SQL Server 日志。

尽管DTS 功能比较的强大，而且提供了图形界面，使得数据的转换变得较为的容易，但是由于数据来源的多样性和复杂性，DTS 并不能完成所有的转换和校验功能。DTS 的局限性有：

(1) 当数据字段之间存在计算关系的矛盾时，在用DTS转换的过程中，可能会出现矛盾不能够被识别出来从而造成最后录入数据仓库中的数据依然存在错误的情况。在这种情况下就要求数据仓库设计人员在设计数据仓库的过程中能预先考虑到种种情况，可以编制数据转换脚本来弥补DTS 功能的不太灵活的缺点。

(2) 如果数据字段本身存在错误，而且又与其他字段没有计算关系时，DTS也是无能为力的。这种情况可以通过一些经验或者规则、规定或专家知识人工来达到识别纠正的目的。这样经过处理之后录入到数据仓库中的数据才能客观的反映现实情况。

3 总结

数据仓库的建设中的数据处理是一项繁琐而复杂的工程，一时的成功并不代表应用过程中不出现问题，往往是边建设边应用，在应用中发现问题，添加功能。所以数据仓库的建设是一项长时间的劳动。

但是这并不表示数据仓库数据的数据处理问题只能长期用手工来完成，可以通过SQL Server 的代理服务使得数据的转换集成达到自动化的程度。在DTS设计器中，可以把要做的转换集成工作步骤按一定的顺序连在一起打成一个DTS包，告诉系统一个步骤一个步骤所要完成的工作，数据的更新操作可以定时完成，比如放在周末或夜晚完成，这些都可以在DTS设计器中设定。

使用SQL Server 代理服务，管理任务可以由设定定时执行哪些任务而实现自动化，并且管理任务可以由定义工作和警告集而实现程序化管理。通过SQL Server 代理服务自动完成的动作，管理员能为它们日复一日的操作任务构造一个健壮的、自管理的环境。这样管理员就能有时间去管理那些不能自动管理的复杂任务。

[参 考 文 献]

1 王珊等，数据仓库技术与联机分析处理[M]，北京科学出版社，1999。

2 W.H.Inmon Building the Data Warehouse[M]，北京机械工业出版社，2000。

3 Dianne Siebold Visual Basic 开发指南——SQL Server篇[M]，北京电子工业出版社，2000。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn