## 实验6 数据的ETL

### 6.1实验目的

1.熟练掌握某一种DBMS数据装载方法（工具不限，本次介绍以SQL Server为例）；

2.体会主键、外键的作用；

3.熟练运用SQL工具清洗数据。

### 6.2内容提要

#### 6.2.1数据导入

1.**导入数据**

SQL Server 2008提供导入数据向导导入导出数据，指定数据源的具体导入数据方法如下：

选择目的数据库，如图6-1所示单击鼠标右键，在弹出菜单中选择“任务”->“导入数据”，进入SQL Server导入导出向导(图6-2所示)。

在选择数据源对话框的选择“Microsoft Excel”，设置Excel文件路径、Excel版本等信息后单击下一步，如图6-3所示。然后在图6-4所示界面选择目标后单击下一步。

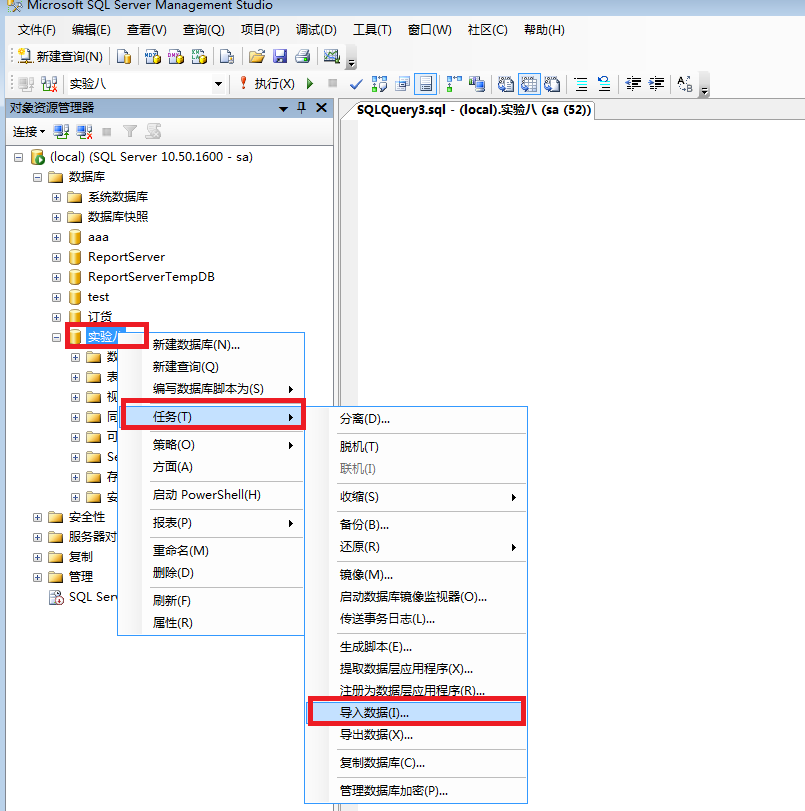


图6-1



图6-2

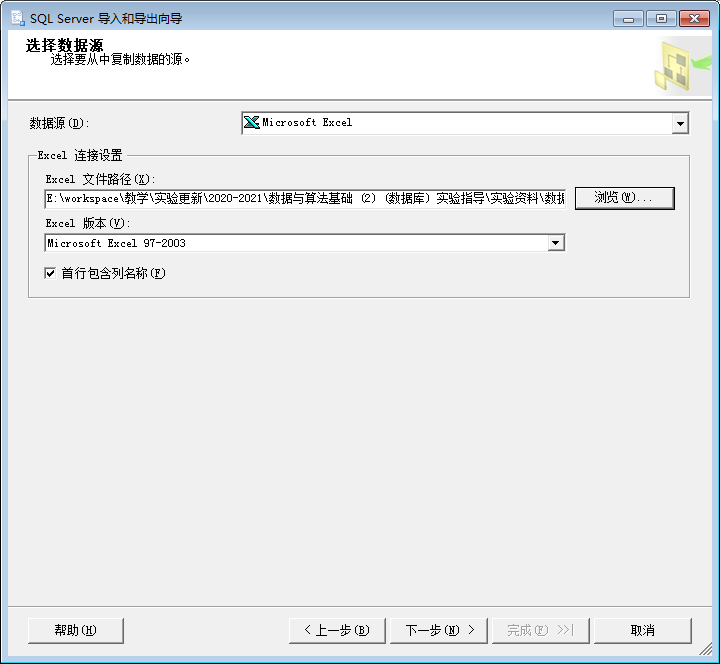


图6-3

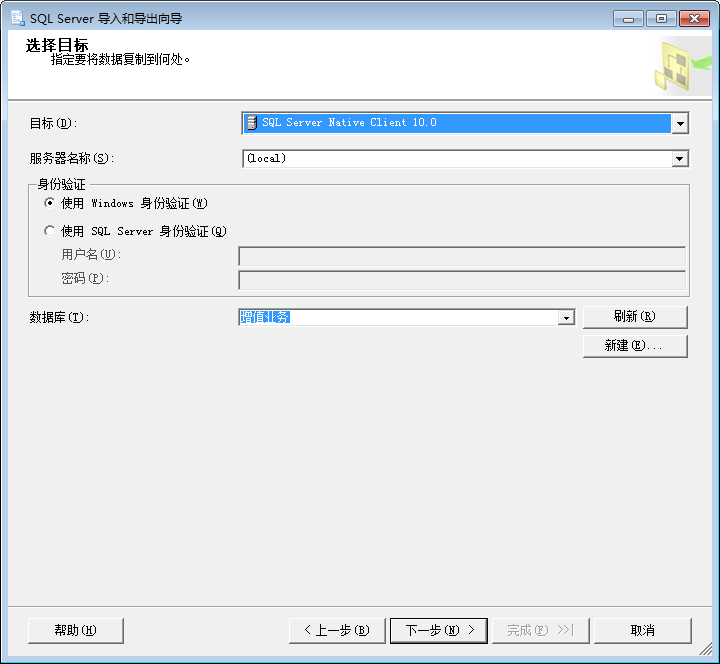


图6-4



图6-5

在图6-5所示界面单击下一步，进入图6-6所示“选择源表和源视图”界面，选中原始数据，在目标列中可选择已有表或新建表。

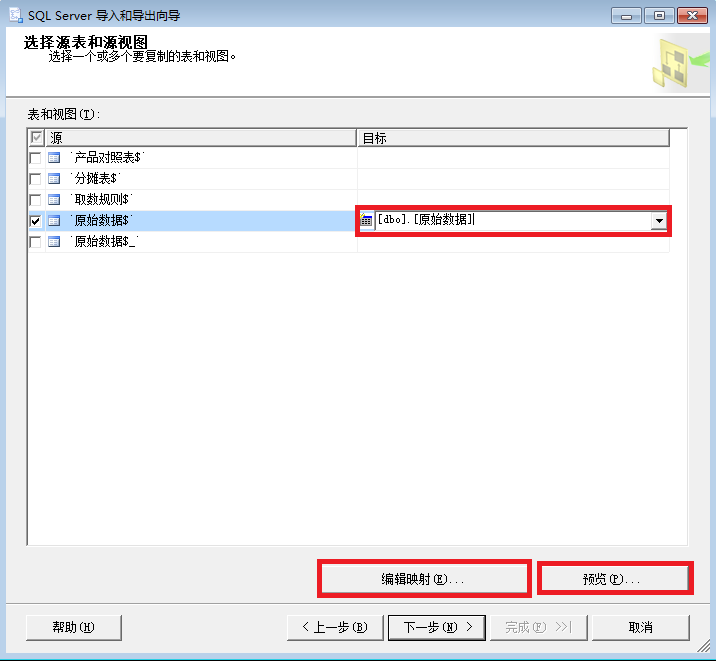


图6-6

表名设置好后单击图6-6所示“编辑映射”按钮，弹出如图6-7所示“列映射对话框”，此窗口可修改**目标列名称、数据类型**等，设置完后点确定返回图6-6，然后可以在图6-6中点击“预览”按钮在预览数据对话框中查看相应数据，如图6-8所示，点击确定返回图6-6所示界面，点击下一步，按照图6-9~6-11所示完成数据导入操作。



图6-7

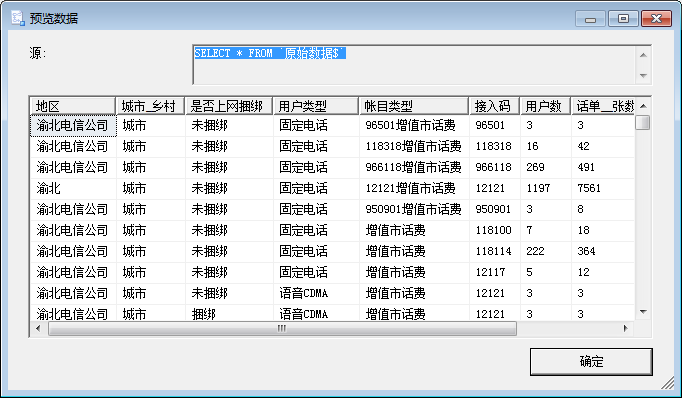


图6-8

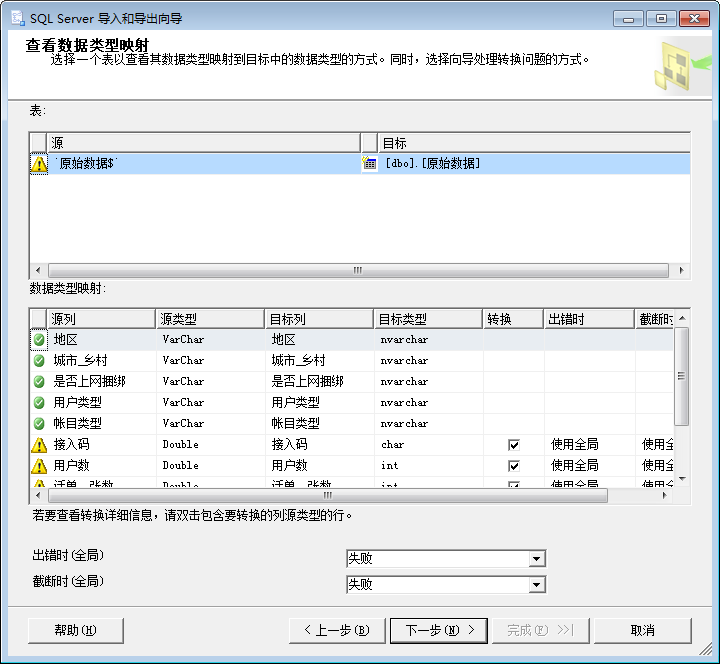


图6-9



图6-10

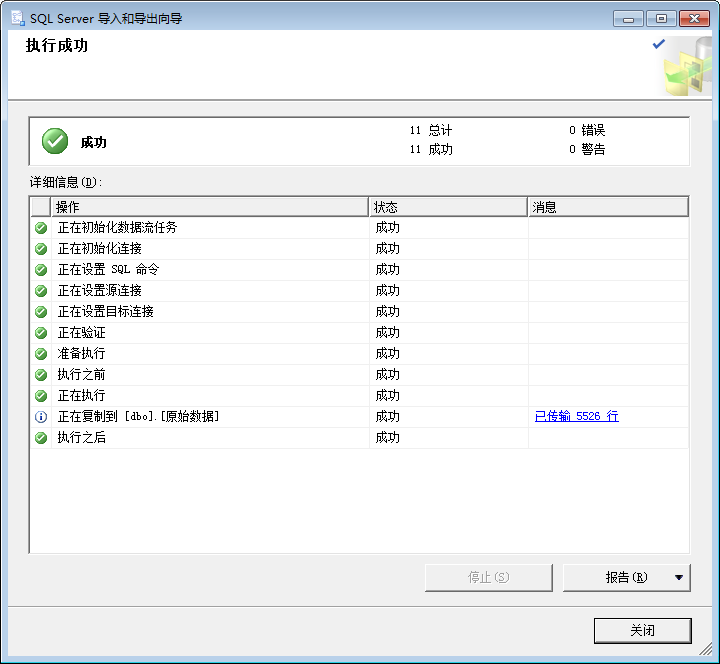


图6-11

#### 6.2.2数据转换

要得到正确的统计报表，ETL (数据的抽取、转换和装载)是一个关键的步骤，该步骤是系统能否正常运行的关键。在对原始数据进行ETL处理时常会遇到以下问题：

1) 原始数据来源于多个系统，生成的数据格式不统一。如该例中BSN系统中地区“巴南电信公司”，在PHS预付系统中却为“巴南”，实际上他们都是指同一家公司。

2) 系统中部分维表数据书写不规范。出现部分括号为全角书写、部分数据半角书写，这样都会导致计算机识别有误。

3) 不可见字符，包括空格等。对原始数据进行处理时，不可见字符是不容易察觉也是及其容易被忽略的，它会导致统计结果不按正常情况显示。

本实验可通过以下方法对数据进行清洗处理：

1．地区维转换。可建立转换表替换数据中地区信息表述不一的情况。

2．全角半角问题可用字符串处理函数REPLACE替换。详细说明见SQL SERVER帮助。

3．不可见字符处理，如空格要用字符串处理函数LTRIM、RTRIM函数综合实现。

4．可采用Insert into … Select 批量写入Select 语句指定数据到相应表中。

### 6.3实验任务及步骤

#### 6.3.1数据装载

**1.源数据**

运行SQL Server 2008，通过导入数据向导，将数据源文件夹“A602电话QQ统计报表”、“A629增值市话和统计报表”、“个人信息处理考查”中原始数据分别导入到数据库相应表中。

**注意**：导入过程中，可能产生导入失败，这事可对数据进行简单处理，如使用在excel或者word中使用“选择性粘贴->文本”去掉数据中的格式设置。

如还出现问题，具体问题要具体解决。该过程也是ETL中较关键的一步。

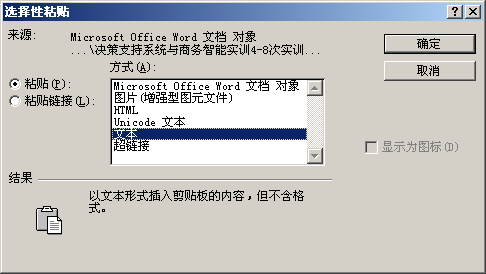


图6-12

**2.合同、财务科目及取数规则**

1) 如表6-1所示，录入合同、财务科目及取数据规则细则。

2) 如表6-2所示，录入财务科目编码表。

表6-1合同、财务科目及取数据规则细则

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 科目代码 | 统计归并帐目科目编码 | 号码 | 合作单位名称（SP） | 电信分成（%） | SP分成（%） | 所属业务 | 列收方式 | 取数报表名称 | 接入号 | 帐目类型 |
| 60010104增值业务 | 1576，1577，1578，1580，1581 | 118318 | 重庆音浮科技有限公司 | 50 | 50 | 电话QQ | 先列收后分成 | A602电话QQ统计报表 | 118318 | 固定电话，后付费非包月小灵通，卡式电话 ，预付费非包月小灵通 |
| 60010104增值业务 | 1576，1577，1578，1580，1581 | 966118 | 贵阳朗玛信息技术有限公司 | 50 | 50 | 电话QQ | 先列收后分成 | A602电话QQ统计报表 | 966118 | 固定电话，后付费非包月小灵通，卡式电话 ，预付费非包月小灵通 |
| 60010106综合信息应用服务 | 1893，1906，1919，1933，1946 | 12121 | 重庆唯胜气象信息有限公司 | 70 | 30 | 12121业务 | 先列收后分成 | A629增值市话和统计报表 | 12121 |  |

表6-2财务科目编码表

|  |  |  |
| --- | --- | --- |
| 科目代码 | 统计归并帐目科目编码 | 统计归并帐目科目名称 |
| 60010104增值业务 | 1576 | 城市 |
| 60010104增值业务 | 1577 | 乡村 |
| 60010104增值业务 | 1578 | 小灵通 |
| 60010104增值业务 | 1580 | 未捆绑 |
| 60010104增值业务 | 1581 | 捆绑 |
| 60010106综合信息应用服务 | 1891 | 城市 |
| 60010106综合信息应用服务 | 1904 | 乡村 |
| 60010106综合信息应用服务 | 1917 | 小灵通 |
| 60010106综合信息应用服务 | 2308 | 未捆绑 |
| 60010106综合信息应用服务 | 2309 | 捆绑 |

#### 6.3.2数据处理

处理原始中不规范数据、不可见字符等。

**1.处理地区列数据**

由于数据来源于不同系统，对同一个地区存在不同的表述，如数据中“巴南”，“巴南电信公司”，“巴南电信”都表示同一个地区，应将其统一为一个名称，如统一为“巴南”。

可以为地区建立如表6-2所示“地区替换表”，再与原始数据表连接查询自动规范地区列数据。

表6-2地区替换表

|  |  |  |
| --- | --- | --- |
| 序号 | 原地区 | 规范后地区 |
| 1 | 巴南 | 巴南电信公司 |
| 2 | 巴南电信公司 | 巴南电信公司 |
| 3 | 巴南电信 | 巴南电信公司 |
| … | 渝北 | 渝北电信公司 |
| … | 渝北电信公司 | 渝北电信公司 |
| … | 渝北电信 | 渝北电信公司 |
| **…** | **……** |  |

**2.处理用户类型列数据**

如表6-2所示建立用户类型与产品类型对照表，再与“原始数据”表连接查询自动规范用户类型数据。

表6-2 用户类型对照表

|  |  |  |
| --- | --- | --- |
| 序号 | 用户类型 | 产品类型 |
| 1 | 语音CDMA | 移 动 |
| 2 | 后付费小灵通 | 小灵通 |
| 3 | 预付费小灵通 | 小灵通 |
| … | PHS | 小灵通 |
| … | 固定电话 | 固 话 |
| … | 卡式电话 | 固 话 |
| **…** | **……** |  |

**3.数据全角半角问题**

对数据进行分析，主要是括号存在全角、与半角共存在数据中，可用字符串处理函数REPLACE替换。

**4.处理数据中不可见字符**

本实验不可见字符有：空格，可使用SQL Server 提供的去空格函数LTrim()、RTrim()。

#### 6.3.3 实验总结

1.记录实验全过程SQL语句，并写出实验报告。

2.详细记录实验过程中遇到的问题，以及问题的解决方法。

### 6.4实验拓展

体会主键、外键的作用，并思考其对数据录入时产生何种影响？