数据仓库中的数据清洗

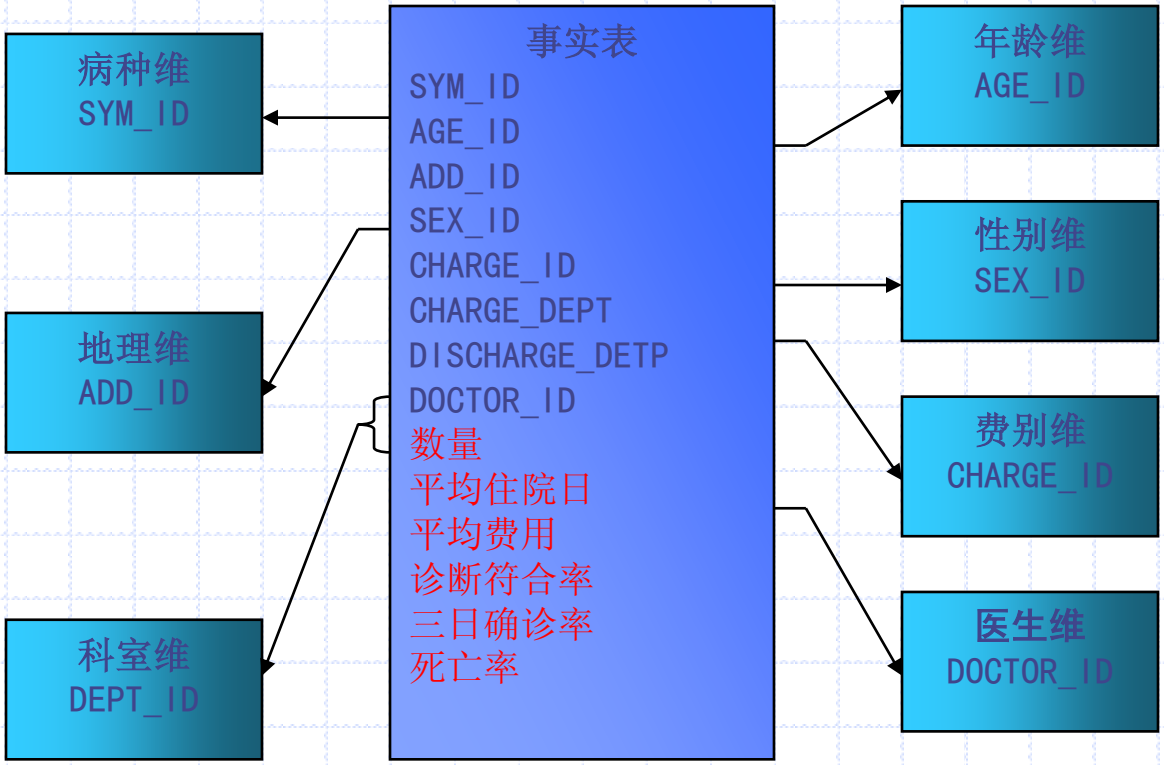
刘玉　　福州大学物理与信息工程学院
陈金雄　南京军区福州总医院

1 引言

- 随着计算机技术和计算机网络技术的发展，医院信息系统中积累了大量的业务数据。如何才能合理的利用这些数据，从中及时发现有用的知识，提高信息的利用率？
- 越来越多的医院选择建立数据仓库以提取其中有用的信息，用于分析和决策。

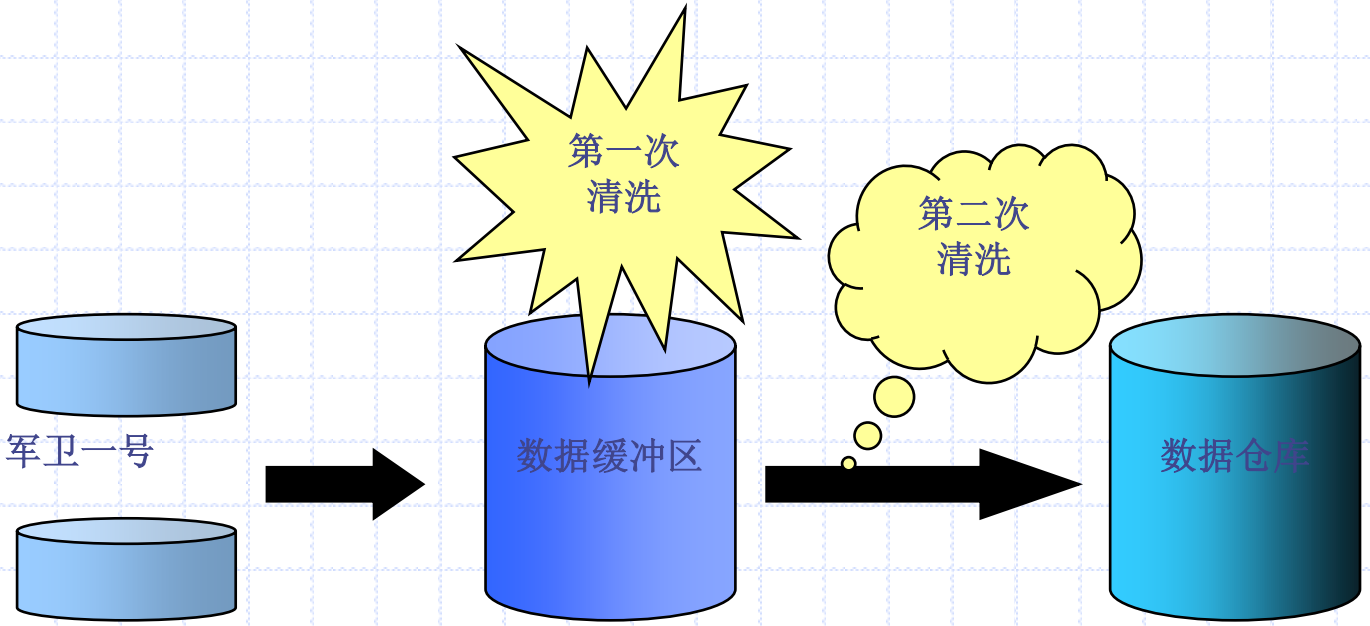
- 数据仓库包含了许多的环节：数据的清洗、转化、装载；维度的选择、度量的聚合；增量数据的提取等等。
- 数据的清洗是其中很重要的一环。

病种分析主题数据仓库的星型模型：



2 整体方案概述

- 由于病种分析涉及到的数据量很大，对这些数据的清洗将占用系统较多的资源，因此，为了不影响“军卫一号”日常的处理速度，同时保证数据尽可能的准确，在设计中采用了“二次清洗”的方法 。



3 第一次清洗

根据“脏数据”种类的不同，有四种不同的

清洗方法：

- 业务表间关联的清洗
- 业务表与公用数据字典间的关联清洗
- 表中空值的清洗
- 不合逻辑的数据的清洗

3.1 业务表间关联的清洗

- 选择主表，建立主表与辅表之间的连接。
- 利用SQL语句，将主表left outer join辅表，找出主表中不能关联至辅表的记录。
- 对不能关联的记录做具体的分析。

  对于真正的“脏数据”，可在辅表中新增一条“默认记录”，将主表中不能关联的记录全部关联到辅表中的“默认记录”。

3.2 业务表与公用数据字典间关联的清洗

以病人住院主记录中的出院科室DEPT_DISCHARGE_FROM为例说明，这种类型数据清洗的过程：

（1）提取病人住院主记录中出院科室不能与科室字典中的科室代码对应的记录，可利用sql语句的*not in*语句来找出出院科室不符合科室字典要求的记录。

(2) 制定转换规则，对步骤（1）中得到的数据进行清洗：

a. 在科室字典中新增一条记录，令科室代码DEPT_CODE＝“FF”，科室名称DEPT_NAME＝“其他科室”；

b. 病人住院主记录中不符合要求的出院科室全部更新为“FF”，将其归为“其他科室”

3.3 表中空值的清洗

对于空值的处理是数据仓库中一个常见问题，是将它作为脏数据还是作为特定一种维成员，应根据实际情况进行判断。

一般的做法是视表中空值字段有无分析价值而定：

（1）对于没有分析价值的字段，如病人住院主记录中的尸检标志、联系人姓名、联系人邮编等属性，在病种分析主题中没有分析的价值，可直接忽略，不进行清洗。

（2）对于有分析价值的数据，则必须根据实际情况对空值进行判断，转化为特定的值。如，病人住院主记录中的出入院科室等属性，主要用于构成病种分析主题中科室维度的外键，有分析价值，必须保留，若出入院科室代码为空，则将其转换为“FF”（对应于科室字典中的“其他科室”）。对于其他为空的有分析价值的属性，也可以采取类似的转换办法，给空值字段赋予特定的值。

3.4 不合逻辑数据的清洗

所谓不符合逻辑要求的数据，指的是不符合现实规律的数据，这种类型的“脏数据”主要集中在涉及到日期的字段。如病种分析中的度量之一——病人的住院天数。

4 第二次数据清洗

- 第二次的数据清洗的主要任务是从经过第一次清洗后的源表中抽取所需要的维度信息。
- 一般来说，维度都是具有层次结构的。因此，按维度层次的清洗顺序，可将维度的清洗分为三种方法：正向清洗、反向清洗及两种的综合。

4.1 反向清洗

以病种维为例，说明反向清洗的过程。病种维采用层次结构，按病种大类、病种中类及疾病名称构成父子层次维度，且一种疾病只对应唯一的病种大类和病种中类。

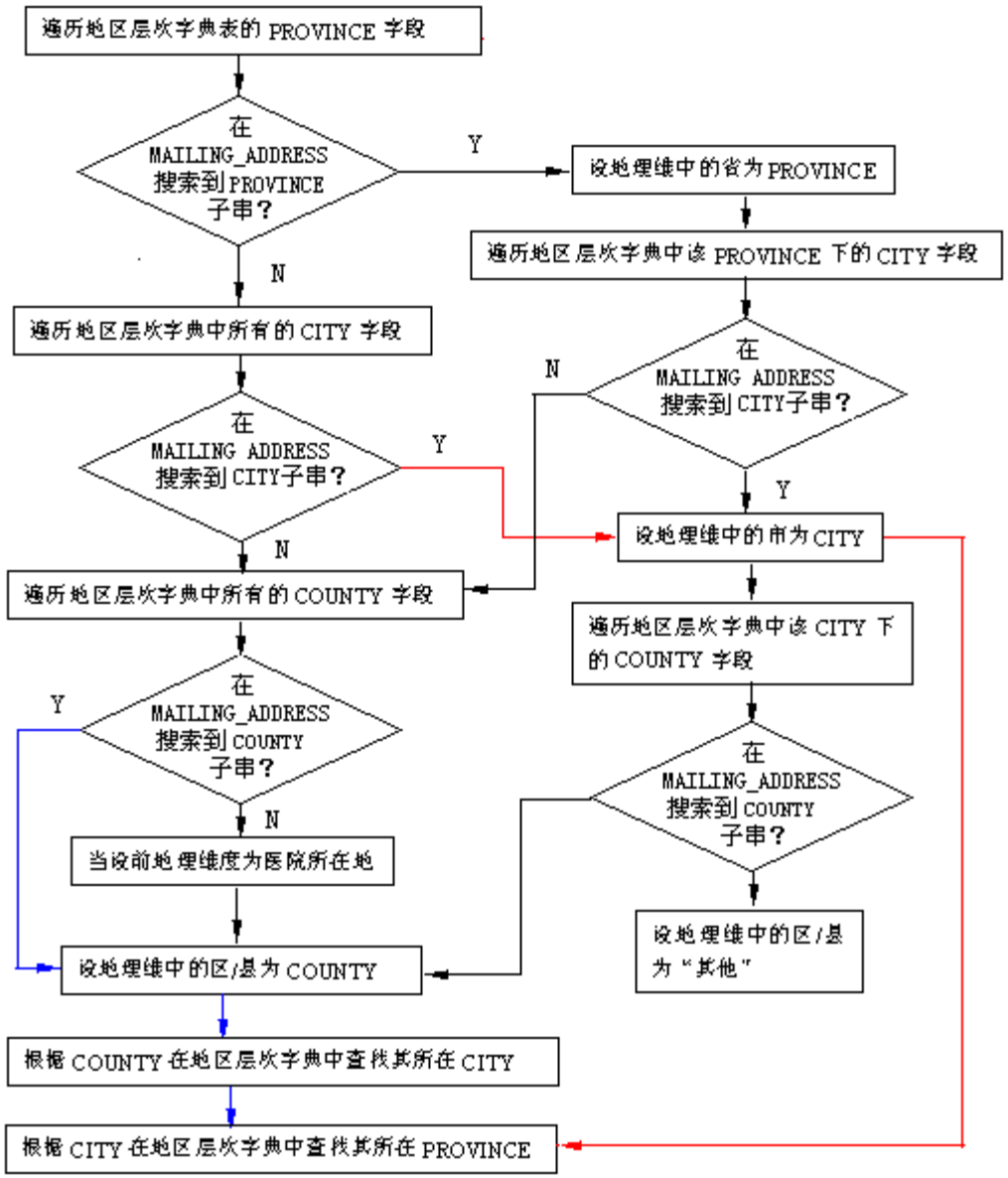
字段名称	类型	长度	说明
ID	NUMBER	4	主键，自动增量填写
SYM_BIG_TYPE	STRING	20	病种大类，按ICD9编码
SYM_MID_TYPE	STRING	20	病种中类，按ICD9编码
DIGNOSIS_CODE	NUMBER	10	与诊断分类记录中的DIAGNOSIS_CODE相关联

4.2 综合清洗

- 以地理维的清洗为例，说明维度综合清洗的过程。地理维用于存储病人所在地的信息。
- 根据需求，地理维应分为省、市、区/县三个层次。

- 源表中没有属性直接指明病人的地理信息，但可以通过对病人住院主记录中的通信地址MAILING_ADDRESS的清洗而得到。
- 为了从通信地址中提取地理维中三个层次的信息，需要在数据缓冲区增加一张地区层次字典LEVEL_AREA_DICT 。

字段名称	类型	长度	说明
ID	NUMBER	4	主键，自动增量填写
PROVINCE	STRING	20	记录各省的信息
CITY	STRING	20	记录各省中市的信息，一个市只能对应一个省
COUNTY	NUMBER	10	记录中各市中区/县的信息，一个区/县只对应一个市



5 结语

- 本文绍了病种分析中的数据清洗的具体方法，对数据仓库中其他主题数据的清洗，也可以采用类似的办法进行。但在数据清洗方面，仍然有许多值得思考和优化的地方，以进一步提高清洗的效率。如，病种分析每月的数据量多达几百万条记录，大表之间的关联必然存在性能和稳定性的问题，必须优化它们的关联；另外，对于大数据的处理无疑会占用太多的系统资源，出错的几率非常大，如何做到有效错误恢复也是个问题。