基于数据仓库的医院管理查询系统的构思

穆荔　刘良斌

摘要：目的　设计数据仓库和动态信息相结合的多维和平面一体的医院管理信息查询系统。方法　以已在应用的 HIS 院长查询系统为引子，确定数据仓库主题和设计事实表和维度表。重点分析 ETL 数据转换规则和数据源质量的管理措施以提升信息的可信度。　结果　产生了本院管理查询系统需求分析说明书和 BI 方案概要设计说明书。

关键词：HIS，数据仓库，数据质量

0.引言

按照卫生部《医院信息系统基本功能规范》建立的院长综合查询分系统，随着 HIS 应用的深入而逐渐成为管理层监控业务运作的有效工具，面向管理的数据仓库项目由此进入我院信息化建设的实施期。在项目的需求调研和方案设计阶段，我们讨论了在不增加业务系统负荷的前提下满足管理层的信息需求、根据信息输出目标确定数据源和梳理业务系统流程等问题。确定了以数据仓库为基础，集成动态信息的直接查询，形成多维和平面为一体的查询系统的目标。

1.数据仓库的设计思路

1.1　数据仓库对历史数据和信息利用的意义

HIS应用的深入和扩展积累了大量的数据，而不断增加的在数据库上进行查询统计的联机分析处理业务(OLAP)，对以实时性要求很高的HIS也产生着越来越大的影响。目前对HIS的历史数据处理方法通常有：1）备份后删除，服务器处理事务的压力得到减负但转移的数据难以被利用。2）删除历史数据的同时建立历史服务器，提供对历史数据的查询。3）将历史数据导入数据仓库中[1]，利用OLAP技术提供的多维分析和动态报表功能，，从多层面对历史数据进行分析，使历史数据不再只用于检索。显然，从实用角度考虑，建立数据仓库是在不增加业务系统负荷的前提下满足管理层信息需求的一种较适宜的解决方案。

1.2　数据仓库体系架构规划

数据仓库的业务需求来自管理人员，而管理人员通常不会考虑技术实现的可能性和复杂程度，所以数据仓库规划的制定者必须综合考虑业务需求和技术实现的难度，优先实现紧迫程度高、技术风险低的需求。自顶向下的设计开发方式要求先计划和实施一个完整的数据仓库结构，这种

方式建设周期长，开销大，对技术开发要求较高。自底向上的方式可先从部门级应用开始创建，然后逐步发展，最终完成全院级数据仓库。这种方式的优点是既能快速见效，又能保持数据仓库的完整架构。我们是以在应用的HIS院长查询系统为引子，展开需求调研和进行主题的规划。这种方法的好处是，管理人员对院长查询系统已有感性认识和应用体会，容易理解数据仓库规划的制定者展现的主题定义和相关事实，也容易提出新的有价值信息需求，以便设计人员充实数据仓库规划。

1.3　数据仓库星形模型设计

“星形模型”是一种常用的数据建模方式[2]，按“数据立方”结构组织和提供数据。星形方案的图形模型中央是一个事实表（Fact Table），呈星形辐射状分布的维表(Dimension Table)从属于这个事实表。事实表包含一些有价值的事实，一般都是数值或其他可以进行计算的数据。维表则具有一些静态的信息，大都是文字、时间等类型的数据。例如住院单病种费用立方的星形模型，图 1。目前我们设计了 38 个立方体。

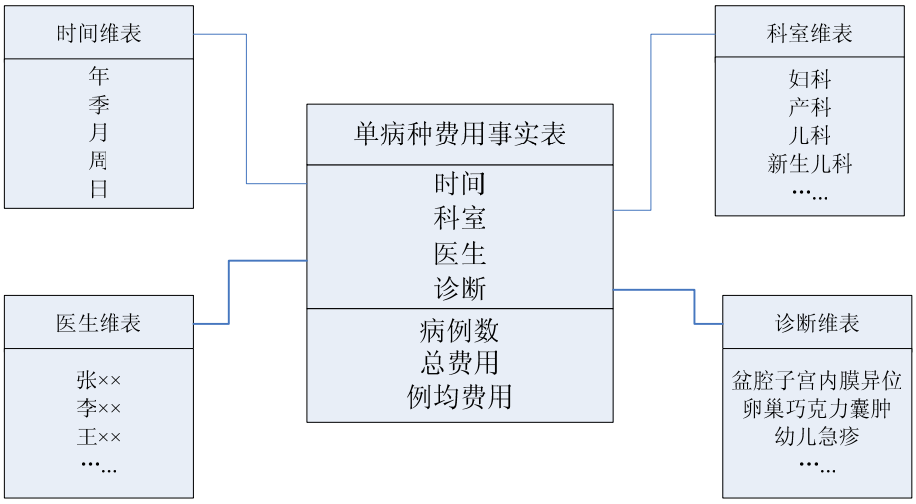

图1　住院单病种费用立方模型

2.与信息相关的数据管理

通常构建数据仓库需要分析数据源、数据准备、星形模型、OLAP服务与应用表达等内容。我们认为把这个过程分为数据层、应用层和表达层三个层面较容易理解，每一层再展开为具体的环节（图2）。数据层是系统的基础，数据的质量和取舍关系到最终展现给管理人员的信息价值。所以，我们详细讨论了对管理信息相关数据的质量管理问题。

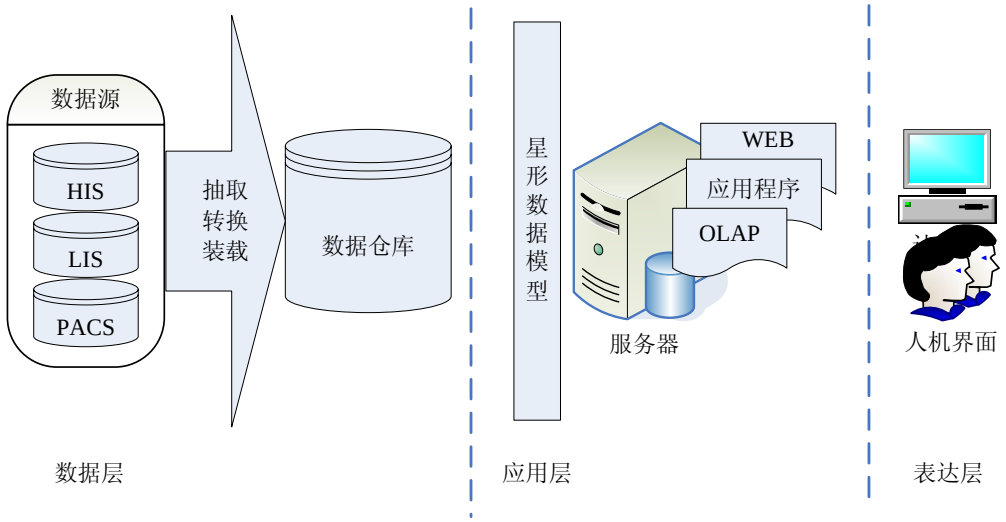

图2　系统总体结构

2.1 数据源的确定及业务流程的梳理

数据源是首先被关注的。我们根据对管理人员信息需求的调研分析设定了查询主题，与主题相关的数据源确定则需要对业务系统进行清点，缺乏数据源的需要通过业务流程的梳理加以补充。例如门诊等待时间，在我院的信息系统中，门诊挂号、医生接诊、开出处方、患者交费、取药、采血、治疗等环节的时间数据源在 HIS 里；检验标本登记和报告发出的时间数据源在 LIS 里；患者到检查科室报到、开始受检和检查报告发出的时间数据源在 PACS 里。我们分别从 HIS、LIS、PACS 中抽取这些时点数据，组成门诊等待时间事实表。

对于住院医疗考核的一些时效性指标和安全指标，以及运转病历监控，正在运行的 CIS 存储的数据不能完全满足。我们同时对与电子病历的生成、质控点和流转有关的业务流程进行梳理，充实相关数据的采集点。这样的数据仓库主题建立与 HIS 流程改造的结合互动，使信息系统对生产的和管理的支持都得到了提升。

2.2 数据的转换规则

数据仓库 60%-80%的工作在ETL，即数据的抽取（Extract）、转换（Transform）和装载（Load）。ETL的过程就是数据从不同异构数据源流向统一的目标数据的过程。其间，数据的抽取、清洗、转换和装载形成串行或并行的过程。ETL的核心又在于转换这个过程，而抽取和装载一般可以作为转换的输入和输出。

数据转换的规则被提出来重点讨论。数据转换的规则是依赖目标数据的，目标数据有多少字段，就有多少条规则。以下规则在我们的开发中都会被用到：

1）直接映射，原来是什么就是什么，原封不动照搬过来，例如病人姓名。

2）字段运算，数据源的一个或多个字段进行数学运算得到目标字段，这种规则一般应用于

数值型字段。例如西药收入、中成药收入、中草药收入相加得到药品收入。

3）参照转换，在转换中通常要用数据源的一个或多个字段作为 Key，去一个关联数组中去搜索特定值，而且应该只能得到唯一值。例如通过员工号关联员工编码表，取得员工姓名并做对应转换。

4）字符串处理，从数据源某个字符串字段中经常可以获取特定信息，例如身份证号。对字符串的操作通常有类型转换、字符串截取等。由于字符类型字段的随意性也造成了脏数据的隐患，所以在处理这种规则的时候，一般要加上异常处理。

5）空值处理，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？还要看应用的具体情况。但是无论如何，对于可能有 NULL 值的字段，不要采用“直接映射”的规则进行处理，一定要对空值进行判断，建议是将它转换成特定的值。例如操作员字段的 NULL 值转换为“XXXX”。

6）日期转换，在数据仓库中日期值有时会有特定的，不同于日期类型值的表示方法，例如使用 8 位整型 20050101 表示日期。而在数据源中，这种字段基本都是日期类型的，这样的转换一般通过一些共通函数来处理。

7）日期、时间运算，基于日期、时间，我们通常会计算日差、月差、时长等。一般通过数据库系统提供的日期运算函数或开放人员自定义的日期运算函数集来处理。例如前面提到的门诊等待时间，就是通过门诊诊疗流程中定义的多个日期时间采集点的数据，通过运算得到的时长。

8）聚集运算，事实表中的度量字段，有时是通过数据源一个或多个字段运用聚集函数得来的，这些聚集函数一般为 SQL 标准函数，包括 sum,count,avg,min,max。

9）既定取值，这种规则的特殊在于它不依赖于数据源字段，对目标字段取一个固定的或是依赖系统的值。例如：数据抽取时间。

构建基于数据仓库的医院管理查询系统对数据准确性有较高要求。除了合理、准确运用上述数据转换的规则提高数据质量外，事实上，有些数据质量问题首先在数据源那里，已经很难保证了。比较常见的有下面几类：

1）数据格式错误，例如缺失数据、数据值超出范围或是数据格式非法等。对于一些处理大数据量的数据源系统，通常会舍弃一些数据库自身的检查机制，例如字段约束等。他们希望尽可能将数据检查在入库前保证，但是这一点是很难确保的。

2）数据一致性，同样，数据源系统为了性能的考虑，会在一定程度上舍弃外键约束，这通常会导致数据不一致。例如有些编码在编码表中找不到等。

上述两类情况可以通过不断发现问题，完善数据源系统的设计，加强业务运作管理来提高数

据质量。

3）业务逻辑的合理性，这一点很难说对与错。有些可以通过优化业务逻辑来解决，有些是业务逻辑本身无法解决的，要通过数据分析者的经验，结合统计理论来对质量不好的数据进行取舍。仍然以分析门诊等待时间的问题为例，在门诊看病流程中，经常会出现病人从一个环节到另一个环节的过程中，因为各种个人原因而没有及时衔接，甚至中断的情况发生。例如，病人挂号后，嫌排队人多或有急事，放弃就诊或办完其它事情后才来就诊。这样必然造成一个时间间隔的特例，从而严重影响数据分析目标（各环节平均等待时间）的准确性。对于这样情况，我们根据经验舍去了各环节时间间隔最短的20%的数据和时间间隔最长的20%的数据，取中间较为合理的60%的数据参与分析，取得比较合理的分析结果数据，从而对门诊流程的优化有较大的参考价值及指导意义。

2.3 数据质量监控措施的制度化

即便是一个设计和规划良好的数据仓库，如果其中存在着大量的噪声数据，展现的信息对管理层来说就可能是一堆垃圾。我们在 ETL 过程的数据清洗环节虽然有诸多的规则可循，但如果原始数据就存在许多质量问题，纵使有再多的清除噪声数据手段也还是使产出信息的可信度大打折扣。所以，建立相关的数据质量监控措施以减少源头的脏数据出现是很有必要的。

有文章对“军卫一号”的数据质量问题进行过研究[3]，罗列了一些不良数据的起因，并提出了数据把关的办法。广东也有同行对医院信息系统建立了数据完整性约束机制。我们也在系统应用不断深入的同时，适时推出一些数据管理措施。在影响数据质量的诸多原因中，“数据歧视”是不易改观的。但如果岗位录入的数据与绩效考评挂勾，数据录入的操作就会被重视。例如病患的基本资料，住址和联系电话多不完整和不真实，这个数据不仅影响数据仓库中的客户源主题，对门诊流程中的门诊日志、疾病报告、处方等也有直接的影响。这个数据的录入环节在挂号处，后续要使用这个数据的部门有门诊医生、药房、院感科等。门诊日志和疾病报告卡的完整性还要接受CDC的检查。这样，医务科、财务科、门诊部等职能部门就联手出台了相应的管理规定，信息科提供挂号病人和相关资料对照表，财务科直接管到录入者。信息科还提供资料完整性的改进情况对照表，使管理部门对规定实施前后有对比数据说话，从而使制度的执行有落实，数据质量也就上来了。

3.基于数据仓库的管理查询系统

管理人员对信息查询的认识往往是从报表开始的，但数据仓库并不是为业务报表而设计的。需要指出的是，数据仓库的分析工具在固定格式的报表再现上有时不如专门定制的程序，所以习惯使用 HIS 的院长查询系统改为使用数据仓库，可能有一个接受过程。数据仓库的强项在于提供

联机的业务分析手段，虽然因为数据仓库的使用，会使管理人员逐步摆脱对固定报表的依赖，代之以丰富和动态的联机查询和分析，但也可能因为使用者知识面的局限而引起对数据仓库应用的困惑。

另外，由于数据要从业务数据源通过 ETL 过程到数据仓库，这必然有一个延迟，少则一天，甚至更长时间。也就是说，数据仓库不能实现随机的统计分析，如果完全将管理查询依靠数据仓库，就存在即时数据无法掌握的问题，势必影响对突发问题的掌控和决策的及时性。

因此，我们在设计中仍然保留 HIS 院长查询系统的模块，与数据仓库结合构成管理查询系统。使系统同时具备了多维立体查询和平面实时查询的功能。由于我们原 HIS 院长查询系统为 C/S 结构，新设计的基于数据仓库的管理查询系统为 B/S 结构。为了避免用户在使用系统的过程中要分别登录到不同系统，并且要在不同系统之间来回切换的麻烦，我们将把 HIS 院长查询系统中管理人员关心和需要的模块制作成动态网页，这部分数据源仍为业务数据库。同时，在以数据仓库为数据来源的医院管理查询系统的人机界面上通过链接访问这些动态网页，从而在用户不知觉的情况下完成涉及两个数据源的两个系统的融合。

4.结语

经过两年的需求分析和方案设计，产生了本院管理查询系统需求分析说明书和 BI 方案概要设计说明书。目前项目处于开发阶段。我们相信，通过充分酝酿的前期构思会使开发和应用的推动都相对顺利一些。

参考文献：

1.刘晓辉,李小华. 基于数据仓库技术维护过期数据的研究. 医疗卫生装备,2006,27(10):41-42

2. 周俊辉. 利用数据仓库整合银行信息资源. 中国金融电脑,2006,2:23-25

3. 王志勇,吴骋,余志明. “军卫一号”数据仓库建设中数据质量问题的研究. 医学信息,2006年19（9）:1503-1505

作者简介：

穆荔，广东省妇幼保健院，科长/副主任医师，020-61118674，muli01@21cn.com

刘良斌，广东省计算中心，项目经理/工程师，020-83548647-807，liulb@gdcc.com.cn