信息数据质量检查软件系统

金碧芳　（上海市公安局科技处 200042）

信息数据作为信息产业的基础和核心，从中发挥着重要的作用，信息数据的质量问题也就成为信息产业发展过程中急需加以重视和解决的迫切问题。因此，采取有效措施提高信息数据质量，才能够有力地推动信息产业的发展和进步，才能够赋予信息产业新的生命力。

对于信息数据中存在的质量问题，有很大一部分可以通过开发数据质量检查软件来实现对数据质量水平的控制。有些单位已经着手开发相关的数据质量检查软件，以下就是数据质量检查软件的具体的方案。

1　建设目标和任务

数据质量检查软件系统建设的目标是：根据不同业务数据的规范，对存在于不同业务系统中的数据记录根据预先设定的检查规则进行数据质量的检查，并返回相应的检查结果。

数据质量检查软件系统建设的任务是：

（1）实现数据检查方案的灵活配置和维护。

（2）定时数据检查任务调度。

（3）数据检查服务运行状况的监控。

（4）信息发布。

（5）实现系统管理功能，包括用户管理、角色管理、日志管理等。

2　技术路线

（1）Web 技术。目前 Web 技术已成为应用开发和管理的主流技术。采用基于 Web 的多层体系结构，提高了设计开发和应用部署的灵活性，相对于传统的 C/S 体系结构开发的应用系统而言，其安装、调试和管理维护的工作量将大大减少。

（2）关系型数据库。数据质量检查软件系统拟采用 ORACLE 关系型数据库建立运行资源库。

（3）负载均衡。数据质量检查的执行是比较耗时的，为了避免当数据检查任务比较多，而检查服务又比较繁忙，造成数据检查任务的排队等待，让数据质量检查任务得到及时的执行，我们考虑部署多个数据检查服务，多个数据检查服务又可以部署在多台检查服务器上，同时在任务调度时，有选择的将数据质量检查的任务分配到比较空闲的数据检查服务和检查服务器上，从而解决任务排队等待的状况，达到各个数据检查服务和数据检查服务器得到平均的任务，从而实现数据检查的负载均衡。

负载均衡功能能够动态地监听请求，并实时地将请求均衡到可用的数据检查服务上，使多个服务器可以像单个系统一样在局域网内高效地运转。

3　系统支撑环境

（1）数据库服务器操作系统采用 WINDOWS2000（或者 UNIX 或 LINUX）；

（2）应用服务器采用 WINDOWS2000（或 UNIX 或 LINUX）操作系统。

（3）客户端操作系统采用 WINDOWS 2000 或 WINDOWS 98 或 LINUX；

（4）数据库平台：配置资源库基于 ORACLE 数据库。

（5）网络平台：网络传输采用基于 HTTP 的协议，消息传输基于 TCP/IP 的 JMS。

（6）开发平台：JBUILDER X、Microsoft Visual Studio. Net 2003、JDK1. 4. 2。

（7）应用支持平台：JBOSS3. 2. 3、TOMCAT5。

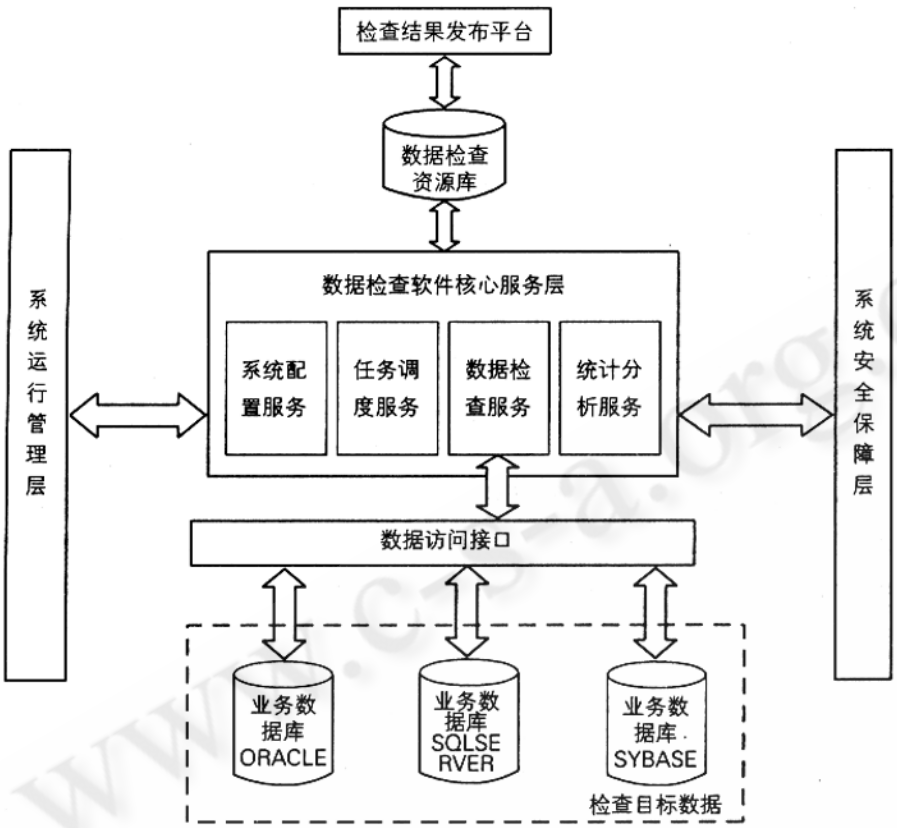
（8）硬件支撑平台：考虑硬件平台的性能、通用性以及扩展性要求，选择 IBM xSeries 365 服务器作为数据质量检查服务器。

4　系统总体体系结构设计

按照《业务数据质量检查软件用户需求》的要求，经过一段时间的研究和技术论证，形成数据质量检查软件的体系结构如图 1 所示。

从图 1 可以看出，数据质量检查软件的框架由以

下几部分组成:系统运行管理层、系统安全保障层、数据质量检查资源库、数据质量检查核心组件(包括系统配置、检查任务代调度、数据质量检查服务、统计分析)、数据访问接口、检查结果发布平台。虚线部分是需要数据质量检查的业务数据,包括各类业务数据库数据和业务数据文件,是数据质量检查软件依赖运行的基础部分,事实上已经物理存在。



数据质量检查软件体系结构

图 1

该系统体系结构图完整、明确地定义了数据质量检查软件的系统构成和系统结构关系。通过对该图的解析,可以看到:在数据质量检查软件系统的体系结构设计中,尽可能地兼顾了目前的计算机应用系统的现状,整个软件的设计与构架实现为将来的发展提供了合理的空间。

下面对体系结构图作一简要说明:

(1) 在图1左、右两侧分别为"系统运行管理层"和"系统安全保障层"。"系统运行管理层"包括数据质量检查软件的用户管理及其授权、系统运行监控、系统日志管理等;"系统安全保障层"是指系统的安全机制、用户的身份验证和安全管理等。这两个层是为数据质量检查核心组件服务的,始终贯穿于该图中间部分数据质量检查核心的各个组件上,对各组件的进入及访问操作进行有效的控制,保证和维护各个服务正常有序地工作。

(2) 核心服务层。核心服务层由四个服务组成,分别是系统配置服务、检查任务调度服务、数据检查服务、统计分析服务,此四个服务是整个数据质量检查的核心组成部分。

(3) 数据检查资源库。数据检查资源库存放的是数据检查的配置参数、检查结果集和各类系统日志。

(4) 数据访问接口。数据访问接口为数据质量检查提供统一的数据访问接口。

(5) 检查结果发布平台。检查结果发布平台将已经通过审核的检查结果通过信息发布的方式进行对外发布。

(6) 检查目标数据。检查目标数据是各个需要检查的业务数据库的集合,支持 ORACLE/SQLSERVER/SYBASE 等主流数据库管理系统。

5 系统设计

5.1 应用功能设计

通过数据质量检查软件系统,能够提供用户对各业务系统的数据质量的进行检查,同时对检查结果进行统计、分析、发布,以及实现与之相关的各项管理功能,如用户管理、日志管理等,详细的功能结果图如图2所示。

数据质量检查软件的应用功能设计主要分为以下几个功能模块:

(1) 系统配置模块。系统配置模块包括数据质量检查规则配置、数据质量检查方案的配置、检查结果输出方案的配置、检查目标数据库配置、检查目标数据表配置。

系统配置是一个发布在局域网上的一个服务程序，用户无论在局域网上的任何可用的 PC 终端上，都可以从指定的发布位置下载到最新的配置程序，进行系统的配置操作。

（2）数据质量检查规则配置模块。数据质量检查的核心是数据检查规则的配置。

图 2

数据检查规则是一系列业务数据规范的集合，它以数据质量检查软件能够识别的形式存在于数据检查软件资源库中。

数据质量检查规则的配置就是要能够完成对于业务数据规范进行归纳和整理，并最终存放在配置资源库中。

业务数据规范千变万化，系统提供灵活和高可扩充性的配置功能，使用户能够通过配置程序完成检查规则的动态管理。

（3）数据质量检查方案配置模块。数据质量检查方案配置模块的实现是通过一个图形化的、灵活的、方便用户使用的配置操作界面，使用户根据不同的需要配置出不同的数据质量检查方案。

（4）数据质量检查模块

① 字段级检查。检查某一字段值是否符合检查标准，例如：检查某系统文件申请数据的规范，检查记录中的姓名字段不为空值、按照身份证号码的规范检查身份证号码、检查出生日期的数据是否符合标准时间规范 YYYYMMDD 格式等；

② 记录级检查。重复性检查，在中国人护照申请数据中检查是否存在相同的记录。

③ 表关联检查。检查某一表字段中的值是否在另外的表字段中存在。

（5）统计分析模块。该软件将提供统计分析服务程序，用户可以对不符合规范的业务数据记录进行定位；可以在检查结果的基础上进行分析和统计，分析不符合规范的数据记录的出现的频度和分布等，并提供以各种图形的方式显示分析的结果，如饼图、柱状图等；统计不符合规范的数据记录的数据量等。同时。通过统计分析服务，数据质量检查软件能够制作出符合用户需要的统计报表。

（6）数据访问接口模块。数据访问接口模块负责提供与各类业务系统数据库或数据文件通讯的接口，它屏蔽了数据库异构和文件访问的差异，对于数据质量检查服务程序而言通过统一数据访问接口，不管后台需要检查的目标数据是何种结构，实现数据结构的无关性。

（7）系统运行管理模块。系统运行管理包括用户管理和授权、系统运行监控、日志管理等功能。

（8）信息发布平台。系统提供根据把通过审核的数据质量检查结果，生成需要发布的数据信息，通过 WEB 的方式对外发布。

5.2 系统性能指标设计

按照本设计方案设计，结合已实施项目的经验，我们提出数据质量检查软件系统主要设计性能指标（需要根据各个单位的实际网络和数据库状况而定，这里只是在特定条件下的测试指标）：

在传输速率达百兆的局域网内，每分钟完成 0.5 千条业务数据记录的质量检查。查询响应的时间在 1 秒之内。

6 结束语

上述的数据质量检查软件方案的实现可以加强各单位对信息数据质量的控制，提高数据质量水平，推动信息产业的发展。