基于 Kettle 的医院多源异构数据采集应用案例分析

樊 睿[1] 王建林[1] 光 奇[1] 牛彩云[1]
FAN Rui WANG Jianlin GUANG Qi NIU Caiyun

摘 要 通过分析医院信息化建设中的多源异构数据采集上报问题。利用开源 ETL 工具 Kettle，将医院需要上报数据定时抽取、转换到 DMZ 区前置服务器，再通过 VPN 传输到上报平台。通过实际案例分析可以看出，在医院数据采集和上报工作中，Kettle 不仅可以支持关系型数据库 Oracle、SQL Server、MySQL 等，还能很好地支持后关系型数据库 Caché。Kettle 友好的用户图形界面化操作、高效稳定的数据处理功能、多源异构数据采集功能等可以优化医院数据采集流程，提高医院各种数据的上报效率。

关键词 医院信息化；多源异构数据；数据抽取；Kettle

doi：10.3969/j.issn.1672-9528.2024.02.003

0 引言

医院各功能部门众多、业务需求各异，因此建立了几十种不同的应用系统，如医院 HIS、LIS、PACS、HRP、后勤管理等，不同系统所使用的数据库也不尽相同，常用数据库有 SQL Server、Oracle、MySQL 及非关系型数据库 Caché 等，各系统又因架构及数据库等不同使得系统间相对比较独立，医院内部存在信息孤岛，使得医院在进行数据上报工作时困难重重。

医院的信息系统经过几十年的发展，历经多次更新换代，被更换掉旧系统的数据大部分也被搁置，主要原因还是在于新旧系统的数据标准不一致导致旧系统数据无法复用，一笔宝贵的数据资源被荒废，但一些特殊病种的分析又需要历史数据的支持，所以如何将历史数据有效利用起来是医院可持续发展的一项重要任务。

Kettle 作为一款开源的 ETL 工具，主要用来完成数据的抽取、清洗、转换和加载等数据处理方面的工作。Kettle 不仅提供了简单明了的图形界面，它的流程式设计也非常方便易用。Kettle 在功能上支持全面的数据访问及多平台部署，拥有优秀的插件架构扩展性，全面实现高效稳定的数据处理 [1]。Kettle 可以通过对多源异构数据采集，打破信息孤岛，实现数据在异构系统之间高效流转。它的这些特性可以解决医院信息化建设过程中多源数据到目标数据的转换、同步难题，让医院摆脱在数据上报时面临的困境，也为以后的数据集成工作提供了更多的思路。

1. 兰州大学第一医院 甘肃兰州 730000

[基金项目] 甘肃省重点研发计划 - 社会发展类，新型肺炎疫情下基于视觉控制的医疗自助系统的应用研究（项目编号：20YF8FA080）；甘肃省重点研发计划 - 工业类，基于可信区块链的数字医院电子病历共享应用研究（项目编号：23YFGA0037）

1 Kettle 概述

ETL（extract-transform-load），即数据抽取、转换、装载的过程。它是一种思想，是从不同的数据源获取数据，并通过对数据进行处理（格式、协议等转换），最后将处理后的数据提供给其他系统使用 [2-4]。

Kettle 是一款开源的 ETL 工具，又名“水壶”，通俗地理解为将不同数据源的数据放到一个壶中，经过转换以一种指定的格式流出。Kettle 支持可视化的图形用户界面（graphics user interface，GUI），以工作流的形式流转，无需安装即可在 Windows、Linux 及 Unix 系统上运行，数据抽取、转换、同步、过滤功能高效稳定 [5]。

Kettle 的功能主要由转换（transformation）和作业（job）两个核心组件完成，其中转换组件完成数据的基础转换，即数据从输入到输出的一个过程，每一个转换表示对一个或多个数据流所做的特定操作，转换是比作业粒度更小一级的容器，一个总的任务可以分解成多个作业，然后将每个作业又分解成一个或多个转换，每个转换只完成一部分工作；作业组件负责完成整个工作流的控制。转换和作业的主要区别在于转换是数据流，而作业是步骤流，作业的每一个步骤必须要等到前一个步骤完成才能执行下一个步骤，而转换会一次性启动所有控件，数据流从第一个控件开始逐个流动到最后一个控件。Kettle 概念模型如图 1 所示 [6-7]。

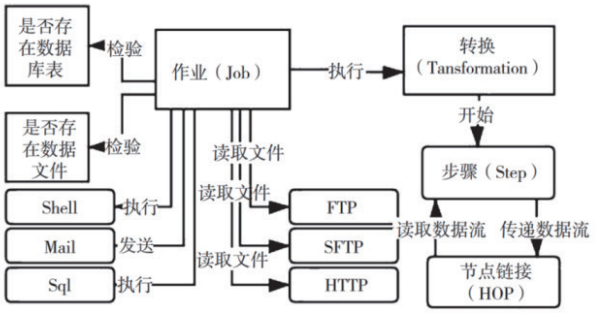

图 1 Kettle 概念模型图

2 国家抗肿瘤药物临床应用监测网数据上报案例分析

为掌握我国抗肿瘤药物临床应用情况，进一步加强肿瘤规范化诊疗管理，国家卫生健康委于 2018 年 12 月 12 日印发了《关于开展全国抗肿瘤药物临床应用监测工作的通知》（国卫办医函〔2018〕1108 号），并委托国家癌症中心开发了全国抗肿瘤药物临床应用监测网。医院需要上报抗肿瘤药物采购记录、抗肿瘤药物使用记录、患者就诊基本信息表等共计 19 类 27 张表，涵盖门急诊住院病人全流程所有医嘱、费用、电子病历、检查及检验数据。

医院数据现状分析数据上报存在以下三大问题。

（1）补报 2013—2021 年历史数据时，HIS 系统存在新旧两套系统的问题，旧 HIS 系统使用 Oracle 数据库存储数据，而新 HIS 系统使用后关系型 Caché 数据库开发设计，如何将两套异构系统的数据同时抽取上报给肿瘤药物检测网比较困难。

（2）运行中的 PACS 系统、病理系统、心电系统等都是独立的应用系统，分别使用了 Oracle 数据库、SQL Server 数据库及 My SQL 数据库，多源异构的系统对数据上报工作增加了难度。

（3）2022 年开始每个月的增量数据上报工作量大，人工完成上报工作效率低，需要分析研究出一种能自动上报且高效率高质量的数据上报形式。

经过调研分析发现开源 ETL 工具 Kettle 的特性能够解决上述问题，Kettle 能支持各种数据源的连接，能在同一个作业中连接多个不同的数据库，将多源数据同时进行抽取转换到目标库中，这一特性可以解决医院问题一和问题二中多个数据源系统的数据采集问题。在操作系统中配置 Kettle 任务定时运行计划，抽取采集数据到前置机数据库中，能够解决问题三。具体方案为：Kettle 工具将医院各数据库中的数据采集、抽取、转换并发送到医院内网前置机的 MySQL 数据库中，再通过定时任务将数据发送到 DMZ 区的监测网外网前置机数据库，最后通过 VPN 加密通道将数据上报到监测网。网络架构图及数据采集流程图如图 2 和图 3 所示。

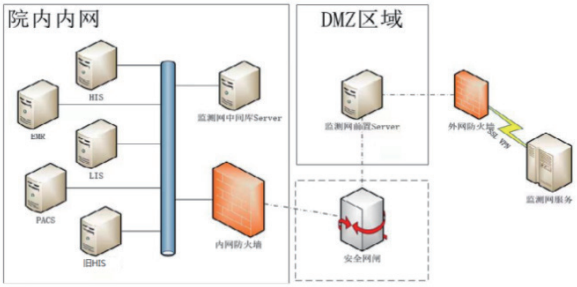

图 2 网络架构图

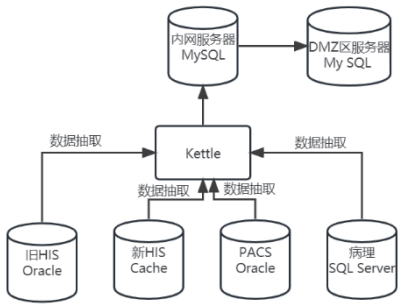

图 3 数据采集流程图

2.1 数据库连接

Kettle 提供了 OCI、ODBC、JDBC、JNDI 等四种不同的数据库连接方式，其中 OCI 只适用于 Oracle。ODBC（open database connectivity）开放数据库互联，是微软公司开发和定义的一套数据库访问标准，在使用 ODBC 链接数据库时需要先在 ODBC 数据源管理器中进行相应数据库的创建配置。JDBC 和 JNDI 是比较普遍使用的两种链接方式，JDBC（Java database connectivity），是 Java 语言中用来规范客户端程序如何来访问数据库的应用程序接口，提供了诸如查询和更新数据库中数据的方法。使用 JDBC 方式连接时首先要将链接数据库的驱动文件放到“\data-integration\lib”目录下，在 Kettle 中新建数据库连接时要在界面上填写主机名、数据库名、端口号和用户名密码等字段。JNDI（Java naming and directory interface）的逻辑是，通过配置文件维护数据库连接信息，并赋予数据库连接信息一个名称，程序通过该名称引用数据库连接信息从而访问后台数据库，具体配置文件为 \data-integration\simple-jndi 目录下的 jdbc.properties 文件，在该文件中填写配置 JNDI 连接名称的 type、driver、url、user、password 的值，在 Kettle 新建连接时直接填写文件中配好的连接名即可 [8]。

Kettle 可以为一个作业同时提供多个数据源的连接，将医院新旧 HIS 系统、PACS、病理等不同系统数据库连接到一个作业中，解决了医院数据多源异构的抽取问题。医院使用的 Oracle、SQL Server、MySQL 等数据库使用 JNDI 方式进行连接，因为使用 JNDI 方式连接时，脚本里仅有连接名

称，文件转移时基本没有风险，会比 JDBC 更安全。医院现用 Caché 数据库则通过 JDBC 的方式进行配置连接。

医院在建立数据库连接前先分析各种现用数据库的数据抽取方式，关系型数据库 Oracle、SQL Server、MySQL 等可以通过在数据库中建立视图或者编写 SELECT 语句的方式抽取数据，所以只用在原始库为 Kettle 建立数据库中表或者视图的只读用户，而后关系型 Caché 数据库用视图或者 SQL 提取数据的效率太低，则采用 M 语言编写的接口程序方法来抽取数据。为了避免抽取数据影响生产库正常业务，Kettle 将数据连接建到 HIS 系统的 Mirror 库上，因为镜像数据库毫秒级同步了业务数据库中的数据，所以 Kettle 从 Mirror 库上提取的数据与业务库中完全一致。数据库连接建立完成后，是只对当前转换或作业有效的，如果需要后面新建作业和转换都可以使用该连接，则需要把该连接设置为“共享”，数据库连接属性如图 4 所示。

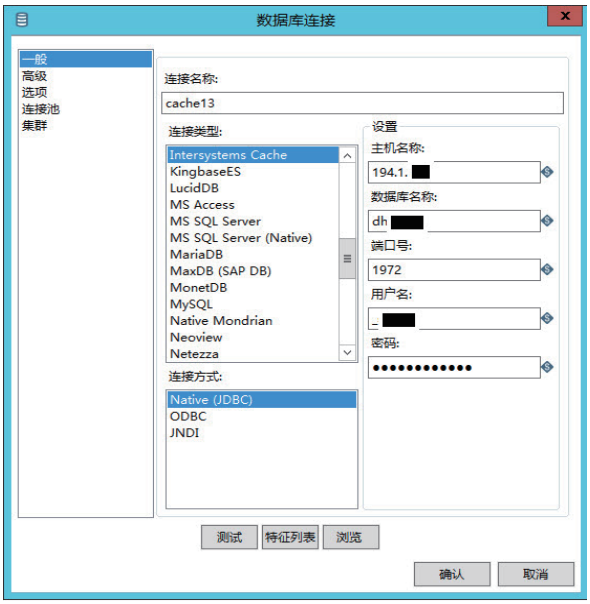

图 4 数据库连接属性

2.2 建立转换

转换是 ETL 解决方案中最主要的部分，处理抽取、转换、加载各阶段各种数据的操作。转换包含一个或多个步骤，其中读取文件或数据，过滤数据，将数据加载到数据库等操作都是步骤。转换里的步骤通过“Hop- 跳”来连接，“Hop- 跳”定义了一个单向通道，允许数据从一个步骤向另一个步骤流动，转换流程具体负责实施对原数据与目标数据的映射关系操作 [9]。

医院新旧 HIS 系统的数据产生于不同架构基于不同数据库的业务系统，多数据源异构主要表现为字段定义和类型的不一致，要将新旧系统的数据提取转换成肿瘤药物监测网需要的数据编码格式并上传到监测网的数据库中，不仅耗时而且相对较难度，而 Kettle 所具有的转换功能正好包含字段选择对照、字符串替换、去除重复记录等操作，在数据抽取时对数据进行清洗转换，能将新旧系统的字段转换成检测网需要的标准编码格式，很好地解决多源数据到目标数据的转换与同步的难题。以病人基本信息表为例，旧 HIS 系统使用的 Oracle 数据库因为只存放历史数据，且对 SQL 语句的执行效率高，Kettle 在转换的表输入模块中直接编写提取数据的 SQL 就能高效地完成数据的抽取任务，而新 HIS 系统使用 Caché 数据库因为是树状存储结构，在 SQL 解析方面效率不高，如果直接用 SQL 抽取数据会影响抽取效率，但是 Caché 数据库自带的 M 语言可以编提取数据的方法，在 Kettle 中使用“CALL ”的方式高效抽取数据，需要抽取的字段全部封装到 patientinfo('startdate', 'enddate') 方法中，在表输入的 SQL 内容中直接编写“CALL Instance. patientinfo('startdate', 'enddate')” 语句就能抽取数据。具体转换如图 5 所示。

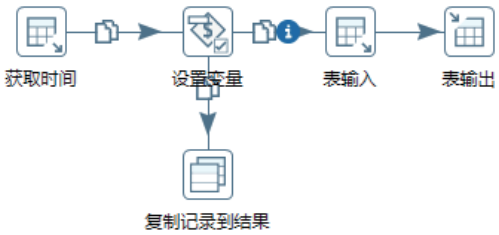

图 5 转换流程建立

2.3 创建作业

作业基于工作流模型，协调数据源、执行过程和相关依赖性的 ETL 活动，完成整个工作流的控制。一个作业可由一系列的转换或者子作业组成，其作用是将每一个转换或子作业按照各自固有的顺序执行，维持整个工作流的秩序。

本例在设计时充分利用了 Kettle 处理多源异构数据的优势，首先在子作业中创建了多个转换处理新旧 HIS 系统中不同数据源的数据、处理 PACS 系统数据的子作业、处理病理系统数据的子作业、处理 LIS 系统数据的子作业等，如图 6 所示。

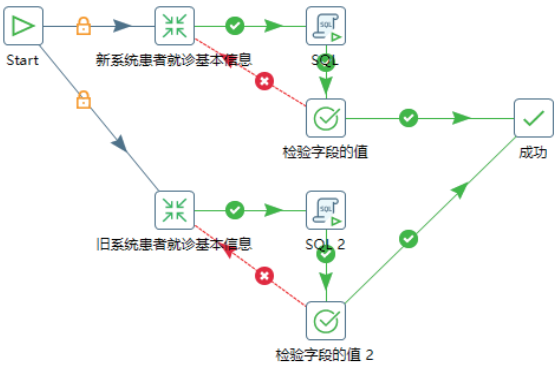

图 6 患者就诊基本信息子作业流程图

然后建立一个整体的作业将各个子作业整合到一个作业中，通过该作业将这些子作业按照工作流的模式关联起来，整体运行将数据抽取到监测网的内网服务器上。最后通过定时任务推送到监测网外网服务器上。父作业创建如图 7 所示。

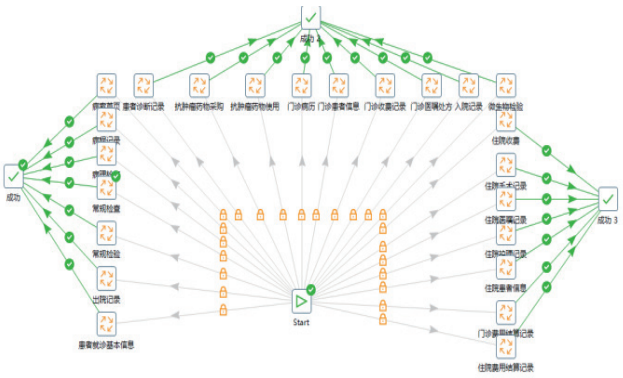

图 7 整体作业流程图

2.4 自动任务配置

操作系统中建立每天的自动执行任务，定时执行 Kettle 任务，抽取上报数据到内网前置机的 My SQL 数据库中，任务配置的具体内容如下：

D:

cd D:\data-integration

kitchen.bat /file D:/ 肿瘤监测数据上报 .kjb /level=Basic /logfile D: \log\log.log

文件内容包含 Kettle 的安装路径、执行作业路径及报错日志的存放路径。每天通过自动执行任务抽取的数据存储到内网前置机的数据库中，通过定时任务再将数据传输到外网服务器的数据库，肿瘤药物监测网通过配置 VPN 工具采集外网服务器数据库中的数据到检测网服务器。

3 Kettle 在处理医院多源异构数据中的意义

通过分析医院的数据现状发现，大多数医院都会存在以下两个方面的问题。

（1）医院因为特殊的行业需求，会使用到很多的信息系统，系统之间虽然有一定的集成，但是信息系统使用的数据库却相对独立，数据存在着信息孤岛，这将对医院的数据集成和数据提取使用分析造成一定的困难。

（2）经过几十年的发展，大部分医院的信息系统都有过数次的更新迭代，被替换掉的信息系统中存在着大量的数据资源，这些数据库相对比较独立，对旧系统数据的利用相对困难。

本文通过肿瘤药物检测网数据上报案例分析可以看出，Kettle 的按照工作流程拖放和设计图形符号操作功能、直观地数据处理和集成过程展示、多数据源的数据集成功能等，能为上述医院面临的数据问题提供解决方案，并且能为医院数据提取汇集工作及数据集成工作提供思路。

4 Kettle 在医院应用的前景探讨

我院正在建设的数据湖应用中将继续探索 Kettle 的应用场景，将 Kettle 与 KAFKA 消息队列的功能互相结合应用到大数据中心建设中，首先通过 DataXone 技术分析数据库日志文件，将获得的实时数据流提供给 KAFKA 消息队列，KAFKA 通过不同的 Topic 将数据提供给消费者 Kettle，Kettle 又将数据清洗转换后提供给临床数据中心 CDR。整个过程利用了 Kettle 工具的跨平台特性，基于不同的数据源特性构建出合适的转换脚本，对数据进行相应的清洗、标准化处理后将实时数据落库到医院临床数据中心，这种方式可以为数据中心提供实时的数据，为医院搭建能为临床科室提供运营指标的目标库及一个可为医院专病数据库提供各种医疗数据的数据湖。

参考文献：

[1] 黑马程序员 . 数据清洗 [M]. 北京 : 清华大学出版社 ,2020.

[2] 刘充 . 基于 Kettle 的高校多源异构数据集成研究及实践 [J]. 电子设计工程 , 2015,23(10):24-26.

[3] 季亚婷 , 刘乐群 . 基于 Kettle 的高校多源异构数据整合实践 [J]. 合肥师范学院学报 , 2019(6):59-61.

[4] 张孟春 . 面向数据集成的分布式 ETL 研究与设计 [J]. 软件导刊 , 2017,16(11): 197-199.

[5] 陈健 , 左秀然 , 杨国良 . 基于 Kettle 的医院多源异构数据集成研究及分析 [J]. 中国数字医学 , 2018,13(3):35-37.

[6] 韦亚军 , 张文文 , 李冬青 . 基于 Kettle 的数据转换同步方法研究 [J]. 软件导刊 , 2022,21(8):126-131.

[7] 赵亚伟 . 一种基于 Kettle 的无损增量数据同步方法研究 [J]. 软件导刊 , 2019, 18(10):55-58.

[8] 赵建勋 . 基于 Kettle 的数据整合研究与实践 [J]. 西安文理学院学报 (自然科学版), 2020, 23(3):28-31.

[9] 李莉娇 . 基于 Kettle 的专项项目库数据同步方法研究 [J]. 信息系统工程 , 2019(12):34-35.

【作者简介】

樊睿（1983—），女，甘肃通渭人，硕士研究生，工程师，研究方向：信息安全。

（收稿日期：2023-11-30）