医疗数据湖建设及医疗数据治理探索*

吴 龙 严晓明 陈秀娟 麦尔丹·吐鲁甫 黎美秀 刘立宇

（广东省人民医院 广州 510080） （生命奇点（北京）科技有限公司 北京 100089）

张 帆 高云鹤 梁会营 杨小红

（广州市妇女儿童医疗中心 广州 510623） （广东省人民医院 广州 510080）

〔摘要〕 介绍医疗数据湖建设发展情况，详细阐述数据入湖流程、数据治理方式、数据湖应用路径，指出数据湖建设及数据治理能够加强医疗数据价值挖掘，提供更多维度数据分析，为医院运营管理和科研业务提供有力支撑。

〔关键词〕 医疗数据湖；数据治理；元数据；数据资产目录

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.06.002

Construction of Medical Data Lake and Exploration on Medical Data Governance

WU Long, YAN Xiaoming, CHEN Xiujuan, Guangdong Provincial People's Hospital, Guangzhou 510080, China; Maierdan. Tulufu, LI Meixiu, LIU Liyu, Gennlife (Beijing) Technology Co. Ltd., Beijing 100089, China; ZHANG Fan, GAO Yunhe, Guangzhou Women and Children's Medical Center, Guangzhou 510623, China; LIANG Huiying, YANG Xiaohong, Guangdong Provincial People's Hospital, Guangzhou 510080, China

〔**Abstract**〕 The paper introduces the construction and development of the medical data lake, expounds the process of data into the lake, data governance mode and application path of the data lake in detail, and points out that the construction of the data lake and data governance can strengthen the value mining of medical data, provide more dimensional data analysis, and provide strong support for hospital operation management and scientific research business.

〔**Keywords**〕 medical data lake; data governance; metadata; data asset catalog

1 引言

为实现医院数字化管理，各医院配套建设了数据仓库、商务智能（Business Intelligence，BI）系统等数据管理工具支撑医院运维决策。随着各临床学科发展，物联网、可穿戴设备的接入，各医疗系统间数据格式不一致、关联性不强、值域不统一、数据异构等问题越发突出。在数据管理方面现有数

〔修回日期〕 2022－06－07

〔作者简介〕 吴龙，高级工程师，发表论文 10 余篇；通信作者：杨小红，主任医师，硕士生导师。

〔基金项目〕 国家重点研发计划“医学人工智能产品检测共性关键技术及标准研究”（项目编号：2019YFB1404804）；国家重点研发计划“面向不确定需求的测试数据集配置平台研发”（项目编号：2019YFB1404803）。

据仓库模式已无法满足医院快速发展需要。随着高水平医院建设的推进，医院对运营管理风险防控、可视化监控、预测分析和精细化管理提出更高要求，数据管理需要打破不同业务系统之间的壁垒，做到数据和业务流程的融会贯通，进一步挖掘数据价值，提升医院综合决策能力[1]。医疗数据湖是可以存储医院各类原始数据的大型仓库，其数据可供存取、处理、分析及传输。数据湖从院内不同业务系统数据源获取原始数据，针对不同的入湖目的，同一份原始数据还可能有多种满足特定内部模型格式的数据副本。数据湖中被处理的数据可能是任意类型信息，包括结构化和非结构化数据。医院希望通过数据湖建设及数据治理提升医疗数据内涵质量，加强医疗数据价值挖掘，帮助临床及管理部门快速获取有用信息并通过数据分析和机器学习算法为医院运营管理和科研业务提供支撑。

2 医疗数据湖建设发展概况

2.1 数据湖发展及定义

2.1.1 发展过程 数据管理经历了数据收集、数据库、数据仓库阶段。数据库面向应用，每个应用可能仅需要一个数据库，如果一个企业有几十个应用就可能需要几十个数据库，由于这些数据库之间无法进行统一分析，因此发展出数据仓库[2]。数据仓库不面向任何应用，而是对接到应用数据库，通过提取 - 转换 - 加载（Extract - Transform - Load，ETL）进行数据抽取和汇总，并按照范式模型进行分析，得到一段时间内的数据视图。随着数据量的增加及数据类型的变化，很多非结构化数据占比越来越多。数据仓库很难继续支撑，越来越多的企业希望将原始数据以真实的初始状态保留下来，在此类需求的推动下数据湖理念逐渐形成。

2.1.2 定义 数据湖（Data Lake）一词最早由美国互联网企业于 2011 年提出[2]，其最早定义为以原始格式存储数据的存储库或系统，是企业级数据解决方案。随着大数据技术的融合发展，数据湖不断演变，汇集了各种技术，包括数据仓库、实时和高速数据流、数据挖掘、深度学习、分布式存储等技术[3]，逐渐发展成为可以存储所有结构化和非结构化任意规模数据并可以运行不同类型数据的大数据工具，是可以对大数据进行处理、实时分析和机器学习等操作的统一数据管理平台[3]。

2.2 数据湖与数据仓库的区别

数据仓库通常从业务系统中提取，在将数据加载到数据仓库之前会对数据进行清理与转换[4]。在数据抓取中数据湖会获取半结构化和非结构化数据[2]，而数据仓库则是获取结构化数据并将其按模型进行组织的[4]。数据湖适合深入分析非结构化数据，而数据仓库因为具有高度结构化的特点而较适用于生成数据指标、报表、报告等。数据湖与数据仓库理念不同，相对于数据仓库注重数据管控，数据湖更倾向于数据服务。

2.3 医疗数据湖建设面临问题与挑战

2.3.1 数据情况错综复杂 医院业务系统因为升级换代、更换厂商等原因，造成不同时期的数据在不同系统中，或者系统升级换代的过渡时期，两套系统同时使用，难以区分业务数据重叠还是分散在不同系统中。数据在抽取、汇聚、分析过程中出现找不到、读不懂、获取难、不敢信等情况。

2.3.2 标准不统一 医院同一业务在不同时期、不同系统中术语不统一，进行数据分析时处理数据、统一术语标准成为最耗时费力的工作。例如诊断、手术操作、药品、检验项目等在不同时间段都存在不同标准术语集，使用这些数据就需要先统一标准集，每次处理业务数据都需要考虑同一业务在不同时期标准字典，还需进行数据格式统一和数据汇总。

2.3.3 数据使用不方便 临床数据分布在不同系统中，各系统数据之间的关联、条件查询缺乏系统支撑。不同系统中的数据缺少外键关联或者外键关联规则不统一，导致各系统关联规则不一致、规则复杂等。例如要查询临床科研数据往往要访问多个业务系统，且各业务系统数据库之间的外键规则不统一，需要关联中间表，查询繁琐、执行效率低。

3 数据入湖

3.1 概述

医疗数据湖是对医疗原始数据的汇聚，数据入湖过程中不对数据做转换、清洗和加工，保留数据原始特征，为后期数据的加工和消费提供丰富可能。数据入湖是数据消费的基础，必须遵从一定入湖标准。

3.2 数据入湖前准备

3.2.1 发布数据标准 入湖数据要有对应的业务数据标准。业务数据标准包括数据资产目录（数据资产目录是元数据的集合，相当于可用数据清单）、数据定义及规则（物理表结构、字段、长度及业务属性描述等）、责任主体，这些标准是医院对数据的共同理解，一旦明确发布需要被共同遵守[5]，如对时间域设置固定的数据长度，值域设置固定的格式“YYYY－MM－DD”即年－月－日，对性别设置标准代码库，业务系统中的“男”“男性”“male”“man”“1”等，都对应标准代码库中“男”。

3.2.2 定义数据密级 医疗数据入湖的必要条件。根据数据资产的重要程度定义不同密级，不同密级数据对应不同数据消费要求。数据密级决定了数据可以共享的级别及用户。

3.2.3 元数据注册 将需要进入医疗数据湖的业务元数据和数据湖的技术元数据进行关联，包括逻辑实体和物理表的对应关系，如超声系统数据库网络地址与数据湖资产目录注册关联，业务表的结构、业务属性和表字段的对应关系与数据湖技术元数据关联。

3.3 结构化数据入湖流程

3.3.1 概述 结构化数据是指以二维表结构表达和实现的数据，其遵循严格的数据格式和长度规范，通常在关系型数据库中存储和管理，见图 1。

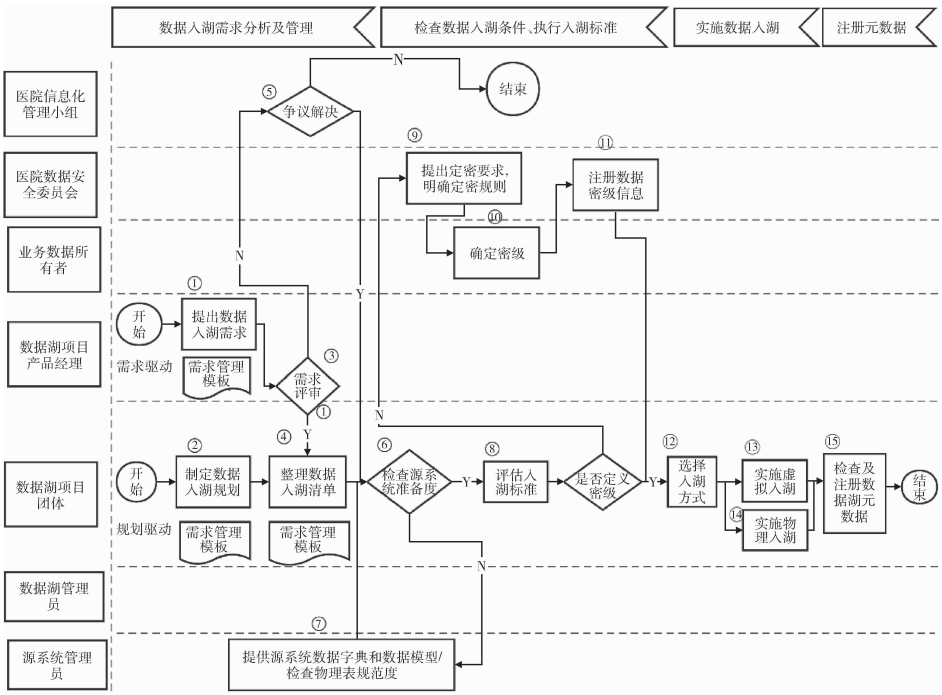

图 1 结构化数据入湖流程

3.3.2 医疗数据入湖需求分析 医疗数据入湖需求分为由数据管理部门发起的主动规划类需求和由数据消费方发起的被动类需求，每个入湖申请都要以使用目的为导向，有针对性地提出今后使用方

向，如临床科研、医院运维管理。发起人需要提供规划清单，并由信息系统工程师提供信息系统分组、业务对象、逻辑实体、源系统物理表和物理字段、业务属性对应的界面截图等信息，经过业务系统部门负责人和数据湖项目建设负责人联合评审通过。

3.3.3 数据入湖条件和标准评估 检查数据源是数据入湖的前提条件，检查需要源系统的工程师提供数据字典和数据模型，并检查源系统的物理表规范度，评估源系统的数据质量[5]。评估标准包括明确数据所有者、发布数据标准、认证数据源、定义数据密级、评估入湖数据质量，不满足上述任一入湖标准则需要源系统完成整改，满足要求后方可实施数据入湖。

3.3.4 实施数据入湖 数据湖管理员根据数据消费场景选择入湖方式，原则上不要求历史数据，数据量小且实时性要求高的场景可优先考虑虚拟入湖；要求历史数据的且数据量大、实时性要求不高的场景，优先考虑物理入湖。数据入湖由数据湖承建商实施，并负责设计集成方案和数据质量检测方案，同医院信息部门一起完成测试和上线验证。

3.4 非结构化数据入湖流程

3.4.1 概述 医疗非结构化数据包括医学影像、音频、视频、生命体征检测波形数据、可穿戴设备数据、物联网设备数据及信息系统数据库日志等异构的格式文件。相较于结构化数据，非结构化数据更难通过标准化理解。因此医疗非结构化数据管理不仅包含文件本身还包含对文件的描述属性，即非结构化的元数据信息。例如文件标题、格式、所有者、设备信息等基本特征，非结构化数据入湖包括基本特征入湖、文件解析内容入湖、文件关系入湖、原始文件入湖，见图 2。

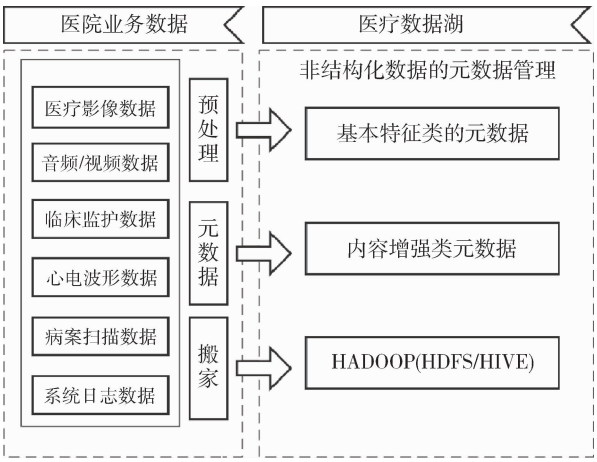

图 2 非结构化数据入湖流程

3.4.2 基本特征数据入湖 基本特征数据入湖过程中，数据内容仍存储在源系统，数据湖中仅存储非结构化数据的基本特征及元属性。非结构化数据的基本特征元属性包括文件唯一标识、文件类型（图片、音频、视频）、创建者、文件内容描述、创建或发布时间、版本、标识、来院、关联、密级等。

3.4.3 文件解析内容入湖 文件解析内容入湖是对元数据的文件内容进行文本解析、拆分后入湖。入湖过程中原始文件仍存储在源系统，数据湖中仅存储解析后的内容增强元数据的描述[5]。如医院早期手写病历，经过扫描归档后，归档目录中仅包含患者住院号、住院时间、住院科室信息。这些数据入湖时经过对扫描文档的文字识别及人工鉴别后（因工作量较大，仅对有科研价值的病历进行属性补充），增加了患者主要诊断、主诉、检验检查等信息，为后续科研检索提供服务。

3.4.4 文件关系入湖 文件关系入湖过程中原始文件仍存储在源系统，数据湖中仅存储文件的关系等增强元数据。如重症监护系统在建设时医院尚未

建设临床数据仓库（Clinical Data Repository，CDR）系统，导致早期积累的 PDF 特护单不能在 CDR 中关联调用。这些历史数据入湖时，通过重新建立元数据关联实现特护单在 CDR 中调取。

3.4.5 原始文件入湖 原始文件入湖是从源端将原始文件搬入数据湖，在数据湖中存储原始文件并进行全生命周期管理。

4 数据治理

4.1 元数据管控

传统的数据仓库将数据存储在关系表中，而数据湖则使用平面结构。每个数据元素分配唯一标识符，并用一组元数据标签进行标记[6]。如一条医嘱数据在医院信息系统（Hospital Information System，HIS）中存放在医嘱表中，并设有对应的主外键关联其他表；在进入数据湖后，需要对医嘱数据进行数据湖唯一标识分配并增加数据标签为“医嘱数据”，同时更新医嘱表的主外键关系。经过元数据管理，之前互不相通的业务系统数据可以实现关联检索。

4.2 数据资源目录管理

数据资源目录包含业务术语表关联、标签管理、数据分类、数据来源和全文检索[7]。每个进入数据湖的系统都需要提供系统数据库配置信息，表结构、表描述及表之间的关联关系等，经过自动化和人工操作更新数据湖资源目录[8]。自动化的工作会设计相应模型，利用机器学习实现数据自动分类和打标签。

4.3 数据清洗

通过属性错误检测进行筛选，筛选出属性错误的数据，根据已发布的数据标准进行清洗[9-10]，如时间格式错误、性别描述错误、身份证号格式错误等。除属性错误清洗外，数据清洗还包括不完整数据清洗，相似重复记录清洗，都需要对数据进行不完整或相似性重复检测并根据规则进行清洗。

5 数据湖应用路径

5.1 建立高效的数据同步工具

建立数据中心服务器集群，通过基于 Hadoop 技术扩展和封装的医疗大数据平台解决数据多源异构问题。在数据湖生产平台中可以看到各项作业的代码、配置、运行状态、运行日志等，并在实时采集系统资源状态的同时进行智能动态分配。在该体系下集群资源得到充分利用的同时，系统稳定性也得到保证，数据安全与平台运行都处于可知、可控状态。增量数据更新和高效的资源利用充分保证了数据的实时性。

5.2 通用数据模型实现数据集成

使用通用数据模型（Common Data Model，CDM）作为大数据平台数据存储的模型，覆盖了医院绝大部分业务与系统，将多源异构数据转换为统一数据模型，利用数据中台存储数据模型转模规则，并通过统一调度平台执行作业的方式实现高效的数据转换和存储，且过程可知、可控。

5.3 建立数据标准化服务

利用自然语言处理（Natural Language Processing，NLP）技术实现术语的字典映射，将不同时期、不同系统中非标准术语进行标准化处理，提高数据质量和可用性。数据标准定义参照国家卫生健康委员会以及国际标准如国际疾病分类（International Classification of Diseases，ICD）第 9 次、第 10 次修订本等，建立代码、数据元的分类标准，依数据规范要求制定详细的代码标准和数据元分类标准，为数据存储、访问、整合提供一致性保障，见图 3。

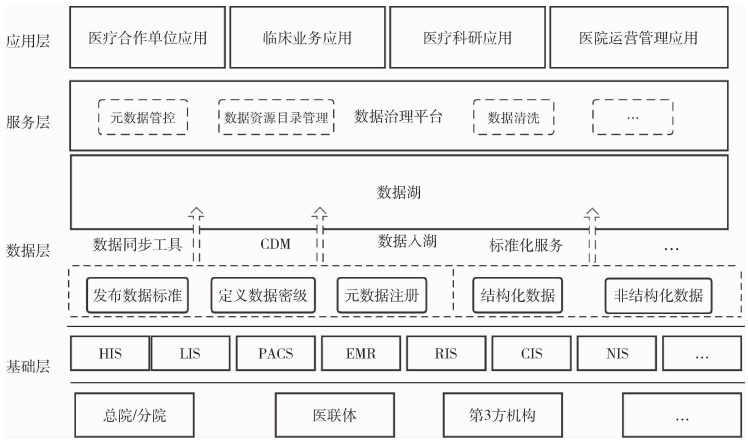

图 3 数据湖建设框架

6 结语

目前借助数据湖所要达成的目标涉及不止一种数据技术，汇集了包括数据仓库、实时和高速数据流技术、数据挖掘、深度学习、分布式存储等技术在内的多种技术，已经从一种“大数据存算方案”进阶到“大数据存算 + 处理分析 + 资产治理 + 安全隐私 + 数据变现”一揽子方案。在数字经济时代，从数据仓库到数据湖不仅是数据存储架构的变革，更是大数据思维方式的升级。数据湖能为医院赋能，帮助医院优化运营模型，为医院科研提供更多维度数据分析，有助于医院提升运营管理和科研能力。

参考文献

1 王韶锋，赵善斌，杨静. 医院数据治理与数据质量提升研究［J］. 现代医院，2021，21（11）：1761 – 1763.

2 尹西明，林镇阳，陈劲，等. 数字基础设施赋能区域创新发展的过程机制研究——基于城市数据湖的案例研究［EB/OL］.［2022 – 06 – 07］. http://kns.cnki.net/kcms/detail/12.1117.g3.20220509.1258.002.html.

3 李硕，卢华明. 基于数据湖的环境大数据存储模型［J］. 北京信息科技大学学报（自然科学版），2021，36（6）：81 – 86.

4 任仲晟. 基于数据仓库的数据挖掘技术［J］. 数字技术与应用，2021，39（9）：59 – 61.

5 华为数据管理部. 华为数据之道［M］. 北京：机械工业出版社，2020.

6 王志勇，吴骋，王立鹏，等. 医疗大数据背景下的数据治理与质量监管［J］. 中国数字医学，2021，16（4）：92 – 96.

7 叶琳，罗铁清. 医疗数据治理综述［J］. 计算机时代，2021（5）：10 – 12.

8 王志勇，吴骋，王立鹏，等. 医疗大数据背景下的数据治理与质量监管［J］. 中国数字医学，2021，16（4）：92 – 96.

9 吴信东，董丙冰，堵新政，等. 数据治理技术［J］. 软件学报，2019，30（9）：2830 – 2856.

10 邓军增. 医院健康医疗数据治理探讨［J］. 医学信息学杂志，2021，42（8）：14 – 17.

（上接第 7 页）

9 龙思哲. 基于数据中台的医院信息系统数据治理方案的探讨［J］. 当代医学，2021，27（29）：193 – 194.

10 叶琳，罗铁清. 医疗数据治理综述［J］. 计算机时代，2021（5）：10 – 12.

11 董方杰，李岳峰，杨龙频，等. 我国卫生健康信息标准工作进展与展望［J］. 中国卫生信息管理杂志，2019，16（4）：400 – 405.

12 王韶锋，赵善斌，杨静. 医院数据治理与数据质量提升研究［J］. 现代医院，2021，21（11）：1761 – 1763.

13 Wang Q，Reps J M，Kostka K F，et al. Development and Validation of a Prognostic Model Predicting Symptomatic Hemorrhagic Transformation in Acute Ischemic Stroke at Scale in the OHDSI Network［J］. PLoS ONE，2020，15（1）：e0226718.

14 洪娜，刘飞，张梦阳，等. OHDSI 通用数据模型在肿瘤大数据中的应用探索［J］. 中国数字医学，2021，16（11）：24 – 28.