工程与管理科学
Engineering and Management Science
第 7 卷◆第 12 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

湖仓一体数据湖在医疗多源异构数据治理中的应用及质量优化研究

孙成龙
联通数智医疗科技有限公司　广州　510000

DOI:10.12238/ems.v7i12.16466

[摘　要] 随着“健康中国 2030”规划推进，医疗数据呈现多源异构特征，传统数据治理方案难以兼顾存储灵活性、数据质量与实时性需求。本文以医疗多源异构数据治理痛点为切入点，阐述湖仓一体数据湖的架构优势，提出“采集 - 整合 - 存储 - 服务 - 安全”全流程应用路径，并从源头控制、智能清洗、动态管控、闭环反馈四个维度构建质量优化策略。以某区域医疗大数据平台为实证对象，结果显示：实施后数据完整率从 82% 提升至 98%，跨机构数据共享耗时从 24 小时缩短至 1 小时，验证了湖仓一体架构的可行性。研究为医疗大数据治理提供技术参考，助力临床决策与科研创新。

[关键词] 湖仓一体数据湖；医疗多源异构数据；数据治理；质量优化；区域医疗大数据

一、引言

医疗数据是临床诊断、科研创新与公共卫生管理的核心资产。当前，医疗数据来源已从传统医院信息系统（HIS、LIS、PACS）扩展至物联网设备（心电监护仪、可穿戴健康设备）与公共卫生平台，形成结构化（如电子病历字段）、半结构化（如 JSON 检验报告）、非结构化（如 DICOM 影像）并存的多源异构格局（李包罗等，2010）。然而，传统数据治理方案存在明显局限：数据仓库虽能保证结构化数据一致性，但难以兼容非结构化数据；数据湖虽支持多类型数据存储，却因缺乏质量管控易形成“数据沼泽”（王健等，2022）。艾瑞咨询（2023）调研显示，85% 的三级医院存在数据孤岛，60% 的医院数据关键字段缺失率超 10%，严重制约医疗数据价值挖掘。

湖仓一体数据湖融合数据仓库的结构化管理能力与数据湖的异构存储优势，通过分层存储、流批一体计算与统一元数据管理，实现“灵活存储 + 高质量管控 + 实时响应”的协同（Apache Hudi，2023）。本文基于医疗数据治理的合规性与业务需求，系统探讨湖仓一体数据湖的应用架构，提出针对性质量优化策略，并结合区域医疗平台实证验证方案有效性，为医疗多源异构数据治理提供可落地的技术路径。

二、相关理论基础

2.1 医疗多源异构数据特征

医疗数据具有“多源化”与“异构化”双重特征：从来源看，涵盖医院核心系统（HIS、LIS、PACS、EMR）、物联网设备（监护仪、血糖仪）及公共卫生数据（疾病监测、疫苗接种）；从结构看，可分为三类（如表 1）：

结构化数据：如 EMR 中的患者 ID、诊断 ICD-10 编码、检验指标数值，具有固定 Schema，易于查询但覆盖范围有限；

半结构化数据：如 LIS 生成的 JSON 检验报告、HL7 FHIR 医疗消息，Schema 灵活但需解析后才能应用；

非结构化数据：如 PACS 的 DICOM 影像、病理切片、医生病程记录文本，体积占医疗数据总量的 70% 以上，处理难度最高（张宏等，2020）。

这种异构性导致数据整合难度大，传统存储方案难以兼顾“全面接入”与“高效应用”的平衡。

2.2 医疗数据治理核心需求

医疗数据治理以“数据可用、安全合规”为目标，覆盖数据全生命周期：

数据采集：需实现多源数据实时 / 批量接入，避免遗漏关键信息（如急诊患者监护数据）；

数据质量：临床诊断与科研分析对数据准确性、完整性要求极高，例如检验值偏差可能导致误诊；

安全合规：《数据安全法》《个人信息保护法》要求医疗数据需脱敏处理，同时满足“可用不可见”的共享需求；

服务响应：临床场景需毫秒级数据查询，科研场景需批量数据支持，需兼顾实时性与批量处理能力（刘军等，2023）。

2.3 湖仓一体数据湖架构

湖仓一体数据湖基于“湖仓协同”理念，构建四层核心架构（如图 1）：

存储层：采用分层策略 —— 热数据（实时监护数据）存于 MPP 数据库（如 Greenplum），温数据（近 3 年 EMR）

存于 HDFS，冷数据（归档影像）存于对象存储（如 OSS），平衡性能与成本；

计算层：融合 Spark（批量处理）与 Flink（流处理），支持实时数据更新与批量分析；

元数据层：基于 Apache Atlas 构建统一元数据管理，记录数据血缘与字典，实现数据可追溯；

服务层：提供 REST API、JDBC 等接口，支撑临床查询、科研建模等场景（王珊等，2004）。

与传统方案相比，其优势在于：兼容多类型数据、支持实时计算、降低存储成本，契合医疗数据治理需求。

三、医疗多源异构数据治理的现状与挑战

3.1 治理现状

当前我国医疗信息化建设已覆盖大部分二级以上医院，但数据治理仍处于碎片化阶段：

数据孤岛显著：85% 的三级医院 HIS、LIS、PACS 系统独立运行，跨科室数据查询平均耗时超 15 分钟，区域医疗平台跨机构数据同步需 24 小时以上（李超等，2023）；

数据质量不达标：60% 的医院关键字段（如过敏史）缺失率超 10%，检验数据异常值未及时清理，导致科研数据可用性不足；

安全合规风险：2022 年全国医疗数据泄露事件达 32 起，主要源于访问控制不严与脱敏不彻底（张宏等，2020）；

技术与业务脱节：现有方案多聚焦技术实现，未结合临床“实时性”与科研“完整性”需求，应用效果不佳。

3.2 核心挑战

多源异构整合难：结构化数据需关系型存储，非结构化数据需对象存储，且更新频率差异大（监护数据实时更新 vs 病历每日更新），传统批量同步无法满足需求；

质量管控标准化缺失：各医院对“数据准确性”定义不一，例如部分医院允许非关键检验值缺失，且人工审核效率低（1 万条数据需 3 天，漏检率 5%）（李娜等，2021）；

安全与共享矛盾：跨机构数据共享需开放权限，但医疗数据隐私性强，现有脱敏技术易被逆推识别；

成本与性能平衡难：非结构化数据体积大，全量存储于高性能数据库成本过高，仅存于数据湖又无法满足实时查询。

四、湖仓一体数据湖的应用路径

4.1 多源数据统一采集

针对医疗数据多样性，设计分层采集方案：

结构化数据：采用 Debezium CDC 工具实时捕获 HIS、EMR 数据变更，通过 Kafka 传输至计算层，避免影响业务系统性能；

半结构化数据：用 Flume 采集 LIS 的 JSON 报告，解析“检验项目 ID、结果值、单位”等核心字段，生成结构化索引；

非结构化数据：通过 MinIO 客户端将 DICOM 影像上传至对象存储，提取元数据（患者 ID、检查日期）存入元数据层。

采集时设置实时校验规则（如血糖值需在 3.9-6.1mmol/L 范围），异常数据标记并触发告警，确保源头质量（陈亮等，2023）。

4.2 元数据驱动智能整合

以元数据为核心实现数据关联与标准化：

统一元数据模型：定义“患者 - 检验 - 诊断 - 影像”核心实体关系，统一字段命名（如“患者 ID”统一为“PATIENT_ID”）与编码（ICD-10、LOINC）；

数据血缘追踪：通过 Apache Atlas 记录数据流转路径（如“EMR 诊断数据→清洗→Greenplum→临床查询”），实现可追溯；

Schema 适配：结构化数据采用“schema-on-write”保证一致性，非结构化数据采用“schema-on-read”保留灵活性，通过 Talend ELT 工具同步至数据资产池（王健等，2022）。

4.3 分层存储与高效计算

按访问频率设计三层存储：

热数据层（近 3 个月）：存于 Greenplum，支持毫秒级查询，满足临床实时调取；

温数据层（3 个月 -3 年）：存于 HDFS，结合 Spark 批量处理，支撑科研统计；

冷数据层（3 年以上）：存于 OSS 压缩存储，降低成本。

计算层采用 Spark+Flink 流批一体架构：Flink 处理实时监护数据，Spark 处理科研批量分析，实现“实时响应 + 批量计算”协同（Apache Hudi，2023）。

4.4 场景化数据服务

基于整合数据提供三类服务：

临床决策服务：向医生工作站推送患者多源数据（监护 + 病历 + 影像），急诊诊断耗时从 40 分钟缩短至 20 分钟；

科研数据服务：构建疾病数据集（如糖尿病、冠心病），提供脱敏数据下载，模型训练周期从 2 个月缩至 2 周；

区域共享服务：实现跨机构数据同步（如转诊患者数据），

达成“患者不动数据动”（陈亮等，2023）。

4.5 全流程安全管控

在架构各环节嵌入安全机制：

加密防护：传输 TLS 1.3 加密，存储 AES-256 加密；

访问控制：RBAC 权限分配（医生仅访问管辖患者数据），双人授权操作；

隐私保护：去标识化 + 假名化处理，结合联邦学习实现“数据可用不可见”；

审计日志：记录所有操作，保留 6 个月以上，便于追溯（张宏等，2020）。

五、数据质量优化策略

5.1 采集阶段源头控制

制定《医疗数据接入标准》，明确三类数据要求：

结构化数据关键字段非空，格式符合 ISO 标准（日期 YYYY-MM-DD）；

半结构化数据需包含“检验项目 ID、结果值”等核心字段；

DICOM 影像需含患者 ID、检查日期等元数据。

接入时用 Flink SQL 实时校验，异常数据反馈源头整改，仪表盘实时监控接入率（低于 95% 触发告警）（李娜等，2021）。

5.2 清洗阶段智能化优化

引入机器学习提升清洗精度：

缺失值填充:KNN 算法基于同人群均值填充数值型数据，决策树预测填充分类数据（准确率 92%）；

异常值识别：孤立森林检测异常检验值，结合临床知识判断是否保留；

冗余删除：哈希值比对重复数据（如重复检验报告），保留最新版本；

文本标准化：NLP 解析病程记录，提取症状并编码（ICD-11）（李娜等，2021）。

5.3 元数据驱动动态管控

建立动态质量规则库：

定义规则（如“患者 ID 唯一”“检验单位匹配项目”）；

每小时扫描数据资产池，生成完整率、准确率报告；

异常自动触发处理（如 ID 重复通知管理员，处理周期缩至 24 小时）；

随业务更新规则（如新增“疫苗接种日期”规则）（王珊等，2004）。

5.4 持续评估与反馈闭环

构建质量评估体系（如表 2），每月评估并反馈临床、科研部门，优化流程。某医院依反馈新增“随访数据规则”，完整率从 95% 升至 98%（陈亮等，2023）。

一级指标	二级指标	目标值
完整性	关键字段缺失率	≤2%
准确性	检验值异常率	≤3%
一致性	跨系统一致率	≥98%
时效性	更新延迟	≤30 分钟
合规性	隐私符合率	100%

六、实证分析

以某省区域医疗平台（覆盖 30 家医院）为对象，2022 年应用湖仓一体方案，对比实施前后指标（如表 3）：

指标	实施前	实施后	提升幅度
关键字段完整率	82%	98%	16%
检验值准确率	85%	96%	11%
跨机构共享耗时	24 小时	1 小时	95.8%
临床查询耗时	15 分钟	2 分钟	86.7%
安全事件发生率	1.2 次 / 月	0 次 / 月	100%

实施后，急诊救治时间缩短 30%，3 个省级科研项目完成数据收集，流感预警响应缩至 2 天，验证方案有效性（陈亮等，2023）。

七、结论与展望

本文提出的湖仓一体数据湖应用路径与质量优化策略，有效解决医疗多源异构数据“整合难、质量低、共享慢”问题：分层存储平衡性能与成本，元数据驱动实现可追溯，智能化清洗提升数据质量，实证显示核心指标显著改善。

未来可进一步探索:结合 AI 大模型自动生成质量规则，引入边缘计算预处理物联网数据，融合联邦学习解决跨机构隐私保护，推动医疗大数据价值深度挖掘（刘军等，2023）。

[参考文献]

[1]王珊，王会举，覃雄派，等。数据仓库和商务智能的研究与进展 [J].计算机学报，2004，27（10）：1343-1351.

[2]李包罗，薛万国。医院信息系统建设与应用 [M].北京：人民军医出版社，2010.

[3]王健，刘伟，张宏。湖仓一体架构在区域医疗数据治理中的应用 [J].中国数字医学，2022，17（8）：45-49.

[4]李娜，赵刚，孙明。机器学习在医疗数据清洗中的应用研究 [J].计算机工程与应用，2021，57（12）：234-240.