**中山大学附属第八医院**

**数据湖仓一体化系统需求**

2025年7月

**目录**

1. 项目背景与目标 1

1.1. ‌项目背景‌ 1

1.2. ‌项目目标‌ 1

1.3. 项目名称 2

2. 功能需求 2

2.1. 系统需求 2

2.2. 硬件需求 8

2.3. 对接需求 9

2.4. 服务需求 9

2.5. 运维服务需求 10

# 项目背景与目标

## ‌项目背景‌

随着医院信息系统的建设和使用，医院积累了大量的医疗数据，包括结构化数据（如电子病历、检验检查结果）、半结构化数据（如HL7、FHIR格式的临床文档）以及非结构化数据（如医学影像、基因测序数据、手术视频等）。这些数据蕴含着巨大的临床和科研价值，但传统的数据仓库架构在处理这些多模态、高复杂度的医疗数据时存在局限性，导致数据孤岛现象严重，数据价值无法被充分挖掘和利用。

为了打破这一困境，提高我院医疗数据的管理效率和利用价值，拟通过湖仓一体的技术方案，构建一个统一、高效、可扩展的多模态数据处理平台，实现医疗数据的全面整合、高效管理和智能应用，为临床诊疗、科研创新及医院管理提供全量数据支持。

## ‌项目目标‌

**‌构建统一数据平台‌：**实现结构化、半结构化、非结构化数据的统一存储与管理，消除数据孤岛，提高数据可用性。

**‌提升数据处理能力‌：**支持批处理与流处理，实现数据的实时入湖与高效分析，满足临床诊疗与科研创新的实时数据需求。

**‌优化数据治理体系‌：**建立元数据管理、数据治理与数据质量管控体系，确保数据的准确性、完整性和安全性。

**‌推动数据智能应用‌：**支持AI、BI等智能应用的开发与部署，为临床诊疗、科研创新及医院管理提供智能化决策支持。

**‌降低运营成本‌：**通过冷热数据分层与成本优化策略，降低数据存储与处理成本，提高资源利用效率。

**提供全面科研数据支持‌：**整合多模态医疗数据，为科研项目提供全面、准确的数据资源，加速科研成果的产出。

## 项目名称

数据湖仓一体化系统‌

# 功能需求

## 系统需求

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 产品名称 | 模块 | 功能 | 数量 | 功能描述 |
| 1 | 湖仓一体数据平台 | 数据管理 | 数据湖管理 | 1 | 支持对数据湖统计的展示，展示存储总数、文件总数、库总数、表总数等存储信息。各存储信息的变化趋势，存储格式分布等汇总信息，同时支持新增和查看数据湖的路径；支持影像（CT/MRI/DICOM）、基因组学（FASTQ/VCF）、音视频（内窥镜录像、手术直播）、电子病历（半结构化文本）等结构化、半结构化、非结构化数据的原生存储；通过Hudi等数据表格式管理软件，支持 ACID 事务，基于Hudi的时间旅行(Time Travel)功能，对历史病历、影像分析的回溯；支持统一元数据与权限管理。 |
| 2 | 数据源管理 | 1 | 支持接入各类数据源类型，包括常见的关系型数据库、非关系型数据库、大数据存储、半结构化存储等 |
| 3 | 元数据管理 | 1 | 对元数据进行检索、自动采集和管理等操作支持元数据自动发现与版本溯源，支持元数据与数据生命周期联动支持查看用户有权限的元数据列表，对元数据进行主题、库、表级别的管理，支持库的新增和删除、表的新增和删除、及库表的权限管理具备完整元数据登记，支持技术、业务、安全等多维元数据关联。 |
| 4 | 资产管理 | 1 | 对数据资产进行全面管理和展示，包括资产首页、资产地图等，资产包括主要包括数据接口实时运行情况、今日资产、累计资产、任务调度以及冷热表运行情况；同时能够多维度展示数据资产运行情况。 |
| 5 | 数据加工 | 数据集成 | 1 | 支持离线和实时的数据同步与采集：1）支持采用DataX用于实现离线数据、异构数据间的同步，采用Spark用于增量处理数据，采用Sqoop,有效解决hadoop与关系型数据库之间传输问题；2）通过集成 Apache Flink 和Flink CDC技术，实现高效、低延迟的实时数据同步。实时捕获源数据库的变更事件，并将这些变更以毫秒级延迟同步到目标系统3）系统既能处理大批量的数据，也能处理持续流动的数据。通过“实时入湖”将数据存到数据湖里，通过Flink CDC做增量处理，最后通过Trino实现秒级查询。从以前ETL到现在的秒级临床反馈，全场景需求都能满足。 |
| 6 | 数据开发 | 1 | 离线开发和实时开发数据，包括工作流创建、执行等，通过可视化拖拽自定义编写脚本创建输入工作流，将数据导入到资产库中；系统可通过任务配置文件（yml）与维表策略配置，实现任务动态加载、变更策略更新、实时调度等能力，增强整体系统的灵活性和可运维性。 |
| 7 | 数据清洗 | 1 | 对数据进行清洗、转换和加载等操作，支持清洗残缺数据、错误数据和重复数据，并支持数据清洗时进行数据校验，确保数据过滤的正确性和准确性 |
| 8 | 数据转换 | 1 | 对数据进行格式转换和标化操作，对业务字符类型、日期、数字等类型数据进行统一的格式转换，保障数据格式的一致性存储。对各业务使用的值域，通过主数据系统的映射操作进行数据标化转换，保障数据含义的一致性存储 |
| 9 | 数据采集监控 | 1 | 对工作流和任务实例进行监控和管理。可快速发现问题排查问题定位到失败工作流，排查解决，手动重跑工作流 |
| 10 | 数据治理 | 数据地图 | 1 | 数据地图基于元数据采集提供各类数据检索、数据血缘、数据质量、版本信息查看等功能 |
| 11 | 数据血缘 | 1 | 监控元数据和相关任务变化状态，进行血缘解析，提供数据在生产流转过程中的血缘关系 |
| 12 | 数据质量 | 1 | 对数据质量进行规范管理、质控项管理、主题管理等，生成质控报告，进一步保障数据湖的数据质量 |
| 13 | 数据安全 | 数据安全 | 1 | 对数据进行脱敏、水印管理，支持建立多层级数据访问权限控制体系，进一步保障数据的安全性和可靠性 |
| 14 | 数据服务 | 数据服务 | 1 | 对数据服务接口进行统一管理、接口管理、角色管理等 |
| 15 | 临床数据中心 | 建模与部署 | 1 | 涵盖对患者基本信息、就诊信息、门诊处方、临床诊断、住院病历（支持非结构化与半结构化数据）、住院医嘱、输血及配发血记录、手麻记录、检验与检查申请单、门诊病历、体征记录、护理病历（支持非结构化数据）、检验报告、PACS报告及病理报告等各类医疗信息的模型设计与运营库数据集成，以实现这些信息的连续性存储与管理。 |
| 16 | 临床数据中心 | 1 | 设计CDR数据仓库，依据国家卫健委标准与医院数据质量，涵盖数据对象等要素。通过CDR平台建设，确立医院数据标准与规范，采用面向服务架构（SOA）实现高度可扩展性。同时，保障数据资产持续保护，确保数据利用不受业务系统更换干扰，使医院在系统选择上拥有主动权 |

## 硬件需求

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 名称 | 配置描述 | 数量 | 单位 |
| 服务器和存储 |
| 1 | HDFS存储计算节点 | 2颗24核以上处理器，1024内存，2块480G SSD硬盘，8块3.84T 硬盘，双口10Gbps网卡，双口25Gbps网卡，冗余电源，原厂3年保修。 | 1 | 台 |
| 2 | ds数据服务节点 | 2颗24核以上处理器，1024内存，2块480G SSD硬盘，8块3.84T 硬盘，双口10Gbps网卡，双口25Gbps网卡，冗余电源，原厂3年保修。 | 1 | 台 |
| 3 | HADOOP管理和应用主机 | 2颗24核以上处理器，1024内存，2块480G SSD硬盘，8块3.84T 硬盘，双口10Gbps网卡，双口25Gbps网卡，冗余电源，原厂3年保修。 | 1 | 台 |
| 网络 |
| 4 | 业务网交换机 | 三层交换机，24个10Gbps光口，配置24个万兆SFP+模块，24对OM4 LC-LC光纤跳线，冗余电源，原厂3年保修。 | 1 | 台 |
| 5 | 存储网交换机 | 三层交换机，24个25Gbps光口，配置24个万兆SFP28模块，24对OM4 LC-LC光纤跳线，100Gbps直连线，冗余电源，原厂3年保修。 | 1 | 台 |

## 对接需求

系统提供跟院内系统厂商的数据对接服务，包括接口及数据库对接联调，新增院内厂商不再收费。

## 服务需求

系统提供不低于壹年免费质保，售后运维服务期间，提供7×16小时热线电话、远程网络等服务方式。远程无法解决的问题，应在接到院方服务请求后12小时内到达现场解决。

在服务期内，非功能性的程序优化、提升均应提供免费升级。

## 运维服务需求

1. 实时监控：运维监控工程师7\*16小时对系统运行大屏进行监控，保证系统异常时第一时间发现问题并及时处理。
2. 日常运维：对服务器上日志文件、应用、操作系统等状态做检查，及时发现可疑问题做分析排查。
3. 故障管理：阿系统故障时准确预警，自动发送手机短信邮件等通知。根据故障情况给出准确故障级别，自动分析及提示运维人员排查。
4. 应急预案：明确的故障时流程规范，有效处置突发事件，反应迅速、处置有力，将正在发生或已发生事故的损害程度减轻到最低。
5. 运维制度规范：严格要求运维各岗位人员，明确其职责及异常处理行为规范。
6. 系统资源优化：不断优化系统各资源配置，监控性能参数，持续性系统调优。