厦门市健康医疗大数据治理及知识

应用平台项目建设方案

2019 年 11 月

目录

第 1 章 项目概述.....9
1.1 项目名称.....9
1.2 项目建设单位.....9
1.3 编制依据.....9
1.4 项目建设目标、规模、内容、建设期.....10
1.4.1 项目建设目标.....10
1.4.2 项目建设规模.....11
1.4.3 项目建设内容.....11
1.4.4 项目建设期.....13
1.5 项目总投资及资金来源.....13
第 2 章 项目建设单位概况.....14
2.1 项目建设单位与职能.....14
2.1.1 项目建设单位概况.....14
2.1.2 项目建设单位职能.....14
2.2 背景及现状分析.....14
2.3 目前存在的问题.....15
2.3.1 数据资源缺乏整合，难以有效管理和利用.....15
2.3.2 数据质量参差不齐，标准化和规范化有待统一.....15
2.3.3 数据应用程度不深，数据价值发挥不足.....16
2.3.4 数据服务缺少平台，产业带动成效不显.....16

第 3 章 需求分析.....17
3.1 业务需求分析.....17
3.1.1 数据标准化、规范化治理及数据资源盘点需求.....17
3.1.2 健康医疗大数据助力区域临床科研建设的需求.....17
3.1.3 健康医疗大数据带动医、药、保产业协同发展的需求.....18
3.2 数据治理服务需求分析.....18
3.2.1 数据治理流程.....18
3.2.2 数据质量改进.....19
3.3 知识应用平台系统功能需求分析.....20
3.3.1 临床数据管理应用功能需求.....20
3.3.2 科研项目管理应用功能需求.....21
3.4 非功能性需求分析.....22
3.4.1 大数据存储与计算需求.....22
3.4.2 海量数据秒级查询搜索需求.....23
3.4.3 可靠性、易用性和可扩展性指标.....23
3.4.4 网络平台性能需求.....23
3.4.5 系统平台性能需求.....24
3.4.6 应用系统性能需求.....24
3.4.7 数据备份需求.....25
3.5 信息安全需求.....25
第 4 章 项目建设方案.....26

4.1 总体设计方案.....26
4.1.1 项目总体定位.....26
4.1.2 系统框架设计.....26
4.1.3 网络架构设计.....27
4.2 总体技术路线.....28
4.2.1 数据标准化技术.....28
4.2.2 多源异构数据治理技术.....28
4.2.3 自然语言处理技术.....28
4.2.4 大数据治理技术.....29
4.2.5 联机分析处理.....30
4.2.6 Kubernetes（微服务架构）.....31
4.2.7 Hadoop.....32
4.2.8 Kylin.....34
4.2.9 HBase.....36
4.2.10 MongoDB.....37
4.2.11 Kafka.....37
4.2.12 Sqoop.....39
4.2.13 安全技术.....39
4.2.14 数据可视化.....40
4.2.15 大数据挖掘与分析技术.....41
4.3 本期项目建设方案.....41

4.3.1 标准规范 41
4.3.2 信息资源规划和数据库建设方案 46
4.3.3 应用支撑平台和应用系统建设及服务方案 88
4.3.4 数据治理和存储系统建设方案 138
4.3.5 网络系统建设方案 155
4.3.6 安全系统建设方案 155
第 5 章 项目组织机构和人员培训 183
5.1 领导和管理机构 183
5.1.1 项目建设领导小组 184
5.1.2 领导小组办公室 184
5.1.3 业务子系统工作小组 185
5.1.4 项目实施机构 185
5.1.5 运行维护机构 188
5.2 技术力量和人员配置 188
5.3 人员培训方案 189
5.3.1 培训目的 189
5.3.2 培训对象 190
5.3.3 培训内容 190
第 6 章 项目实施进度 192
6.1 项目建设期 192
6.2 实施进度计划 192

第 7 章 投资概算.....193
7.1 投资概算的有关说明.....193
7.1.1 投资概算说明.....193
7.1.2 投资概算依据.....193
7.2 项目总投资概算与概算明细表.....195
7.2.1 项目总投资概算表.....195
7.2.2 硬件设备、软件及服务购置概算明细表.....196
7.2.3 应用系统及服务采购工作量概算明细表.....201
第 8 章 效益分析.....206
8.1 经济效益分析.....206
8.2 社会效益分析.....206
8.3 共享资源分析.....207
第 9 章 项目风险与风险管理.....208
9.1 风险识别和分析.....208
9.1.1 组织风险.....208
9.1.2 管理风险.....208
9.1.3 业务风险.....208
9.1.4 技术风险.....208
9.2 风险对策和管理.....208
9.2.1 组织风险防范对策.....209
9.2.2 管理风险防范对策.....209

9.2.3 业务风险防范对策 209

9.2.4 技术风险防范对策 210

第1章　项目概述

1.1 项目名称

项目名称：厦门市健康医疗大数据治理及知识应用平台项目

1.2 项目建设单位

项目建设单位：厦门市健康医疗大数据中心

1.3 编制依据

1、《“健康中国 2030”规划纲要》（中共中央、国务院）

2、《促进大数据发展行动纲要》（国发〔2015〕50 号）

3、《关于促进和规范健康医疗大数据应用发展的指导意见》（国办发〔2016〕47 号）

4、《关于印发国家健康医疗大数据标准、安全和服务管理办法（试行）的通知》（国卫规划发〔2018〕23 号）

5、《大数据产业发展规划 2016-2020》（工信部规〔2016〕412 号）

6、《促进健康产业高质量发展行动纲要（2019-2022 年）》（发改社会〔2019〕1427 号）

7、《“十三五”深化医药卫生体制改革规划》（国发〔2016〕78 号）

8、《国务院办公厅关于促进“互联网+医疗健康”发展的意见》（国办发〔2018〕26 号）

9、《厦门市促进大数据发展工作实施方案》

10、《厦门市大数据应用与产业发展规划》

1.4 项目建设目标、规模、内容、建设期

1.4.1 项目建设目标

厦门市健康医疗大数据治理及知识应用平台项目通过运用先进的医学大数据及人工智能技术，对采集和汇聚的医疗机构数据进行深度治理，将医学文本处理为归一化、结构化、标准化的医学数据，并建立统一的健康医疗大数据基础模型，提升数据质量及可用性。在此基础上建立医学知识图谱，将数据转化成医学知识，建立患者全生命周期健康数据模型，打造国内领先的集数据治理、开放、应用、运营功能于一体的健康医疗大数据平台，对医学科研领域进行赋能，为厦门城市开放式智慧式发展做出贡献。

1、构建统一的健康医疗数据资源标准规范体系

按照健康医疗领域的相关国家数据标准、省市地方标准，构建统一的健康医疗数据资源标准规范体系，建设数据治理相关标准规范，为数据治理全生命周期提供规范的数据治理标准，促进厦门市健康医疗大数据治理系统建设项目更加规范化、制度化，推动数据治理高效、有序的开展，以保证数据的统一性、科学性和可靠性。

2、提供健康医疗数据资源清洗加工服务

以目前已经完成采集的健康医疗数据为基础，推进原始数据的清洗，去除“脏”数据，加强数据比对，开展数据粗加工、精加工，推进数据的组合、关联、分析和可视化等提供程序化数据治理服务。

3、构建健康医疗数据资源共享开放体系

通过厦门市健康医疗大数据治理工作，建立完善的健康医疗相关数据共享开放体系，实现健康医疗数据跨区域、跨层级、跨部门、跨领域的数据共享和开放，为省市区各级卫健管理部门、各级政府部门、各级医疗机构、科研机构、医药企业提供标准化、高质量的数据支撑和服务保障。

4、提升临床科研数据服务能力

依托标准化、高质量数据支撑，通过本项目数据治理后的数据应用，充分释放健康医疗大数据的科研价值，通过数据深度分析和可视化展示，为医学科研、辅助诊断、趋势分析等业务提供强有力数据支撑服务，帮助医生和科研人员对文本电子病历、数值检查检验等数据进行科学分析，帮助管理者实现精准分析和科学决策，提升效率和准确性。

5、推进健康医疗产业发展

基于本项目建设的健康医疗大数据治理和知识应用平台，未来可逐步探索行业数据共享应用服务，积极发展以健康、医疗、科技、文化等核心业务为主要内容的健康医疗大数据核心业态，重点培育高端医疗设备等与核心业态紧密联系的大数据关联业态，加快形成层次分明、多方联动、协调发展、新兴繁荣的大数据产业生态体系，形成支撑全市医疗产业转型发展的源动力。

1.4.2 项目建设规模

本项目所涵盖的数据治理服务围绕目前厦门市卫健委已经完成数据采集的全市 16 家三级以上医院以及 39 家基层社区卫生服务中心的诊疗相关数据，覆盖全市 95%以上具有科研职能的公立医疗机构，数据量约 30-35T 之间。

基于知识应用平台的建设要求，数据范围主要涵盖各医疗机构的 HIS、EMR、LIS、PACS、体检、心电系统及基层社区卫生服务中心的云 HIS 及移动家庭签约系统的历史诊疗数据和系统上线后的实时新增诊疗数据。

1.4.3 项目建设内容

本项目建设内容主要包含大数据基础平台、健康医疗大数据治理服务和知识应用平台系统三部分内容。

1.4.3.1 大数据基础平台

大数据基础平台提供医学大数据的存储和计算平台服务，基于主流 Hadoop

分布式架构，具备海量数据快速处理能力，保障医学数据归集和整合，是保障数据能够快速分析、处理海量数据的手段，提供针对 TB/PB 级别数据的离线和实时处理能力。大数据计算平台向用户提供数据集成、数据开发、算法开发等可视化的一站式开发平台，图形化的部署和到运维管理平台能够更快速的解决用户海量数据计算问题，有效降低数据治理成本，并保障数据安全。

1.4.3.2 健康医疗大数据治理服务

1、建设基于医学技术的数据生产处理体系

根据不同数据应用需求，依据海量医学数据治理经验建立的数据全流程一体化生产处理系统，对数据进行自然语言识别、医学术语归一、数据结构化处理等，使其更符合医学大数据挖掘应用需求。

2、建立基于 PDCA 的数据持续治理体系

根据不同数据应用需求，建立数据质控规则，对采集、处理、应用过程中各类数据进行表、字段级的数据持续治理，以解决现有数据不规范、数据质控难、数据质量低等问题，实现数据高可用。

3、建立基于数据共享的数据开放体系

根据未来海量数据增长和业务高度发展需求，对数据采集、生产、治理等基础服务进行封装提供，面向未来各层面的业务拓展应用提供专业数据服务。

4、构建数据治理科研架构

构建集基础临床数据治理、专科数据治理为一体的医疗数据科研架构，打造疾病协同网络，提升科研效能。

1.4.3.3 健康医疗大数据知识应用平台系统

基于大数据平台经过治理后的标准数据，建设面向政府监管机构、医学院校、医疗机构、医学专家科研人员开放使用的健康医疗大数据知识应用平台系统，并在此基础上结合厦门市医学临床诊疗的专科优势及业务需求，选取 3 种专科病种进行深入数据治理，构建专病数据库，逐步强化专科优势，提高科研能力。系统

包括医学科研探索发现、区域疾病知识图谱、海量病历极速检索和分析统计、科研项目管理、科研患者数据管理、医学知识全库等功能。

1.4.4 项目建设期

项目建设周期预计 2019 年 12 月底前完成项目方案评审、立项等工作，2020 年 2 月底前完成项目招投标工作，2020 年 3 月份开工建设，建设周期 9-10 个月，在 2020 年 12 月底前完成项目验收。

1.5 项目总投资及资金来源

项目总投资：本项目涉及投资总数为 900 万元整，具体参见《项目总投资概算表》。

项目资金来源：厦门市市级财政资金。

第2章　项目建设单位概况

2.1 项目建设单位与职能

2.1.1 项目建设单位概况

厦门市健康医疗大数据中心。

2.1.2 项目建设单位职能

厦门市健康医疗大数据中心致力于卫生计生信息资源的整合，中心数据和卫生信息化应用系统的组织建设、管理、运行与维护；国家健康医疗大数据试点有关工作；药物研究开发、医学基础研究、医学信息查新咨询服务与新技术、新成果的推广应用等工作。近五年获省科学进步三等奖 2 项、厦门市科技二等奖 2 项，三等奖 1 项、省市医学科技奖 2 项。取得国家食品药品监督局保健食品批准证书、新药临床批件 3 项，承担国家自然科学基金、国家海洋公益合作项目 5 项。

2.2 背景及现状分析

截止 2018 年底，厦门市全市常住人口共计 411 万。与此对应厦门市共有医疗机构总数 1804 个（公立医疗机构 568 个），其中医院 54 个，基层医疗机构 1713 个，专业公共卫生机构 24 个（含疾病预防控制中心、专科疾病防治机构、妇幼保健机构、卫生监督所、采供血机构、急救中心等）、其他卫生机构 13 个，共同为全市人民提供医疗健康及公共卫生相关服务。

2018 年，全市医疗卫生机构总诊疗人次数达 3644.88 万人次，入院人数 605956 人呈持续增长态势。

为了更好的支持公共卫生及医疗健康服务工作的开展，厦门市卫生健康委员会及各医疗机构近年来一直持续升级和优化全市的医疗信息化建设和服务水平，

在便捷医疗、互联网医疗、智慧医疗方面取得了显著成效，同时在医疗健康大数据的采集、统计及应用方面也做了很多工作，为未来打造开放共享+共建共赢的医疗大数据新生态奠定了坚实的基础。

目前厦门市已经建立市级统一的区域医疗信息化平台，并采用 OGG 的数据采集方式实现了对全市 346 家医疗机构的医疗数据的采集工作，数据量共计 50TB，包含约 1514 万名患者约 1.6 亿次就诊记录的病历数据，目前保持平均每月新增约 250 万条就诊数据的增长量。基于已经采集到的这些数据，厦门市卫健委也开展了一些数据应用服务，包括预约挂号、院外候诊、个人健康档案查询等服务，同时在医疗监管方面也做了一些如用药信息监测等动态医疗数据监管服务，但总体来说目前的这些应用仅仅是对收集来的原始数据的简单查询和数据统计服务，由于这些海量历史数据来自各个医疗机构、不同厂商、不同标准的业务系统，数据没有进行有效的标准化、结构化和归一处理，所以在数据应用方面难以开展更深入、更精细、更高价值的应用探索。

2.3 目前存在的问题

2.3.1 数据资源缺乏整合，难以有效管理和利用

厦门市各级卫生行政管理部门、医疗机构、健康服务机构虽然在长期服务过程中积累了大量的健康医疗数据，虽然厦门完成了大部分医疗机构的数据收集工作，但收集来的数据依然存在数据形态、标准、存储格式各异，未能汇集成统一的厦门市健康医疗大数据资源，难以进行有效地管理和利用。

2.3.2 数据质量参差不齐，标准化和规范化有待统一

数据的标准化和规范化是开展健康医疗大数据挖掘分析应用的重要基础。由于全市各医疗机构管理水平及信息化建设程度存在差异，导致数据标准和规范存在有指标无数据、有数据不标准、有标准不准确、考核指标计算路径不一致、考核指标量统计不准确等问题。另一方面，由于健康医疗数据的专业性，存在大量的医学数据规则需要遵循，这也大大增加了各部门、各机构自行进行数据质量控

制的难度。

2.3.3 数据应用程度不深，数据价值发挥不足

目前厦门市健康医疗大数据的应用重点还是围绕业务流程开展，主要解决业务规范性和业务效率问题，但是从大数据发展规律来看，只有对数据、信息中蕴含的规律进行总结和提炼，形成知识，并进一步上升为智慧，为决策提供有效的依据，数据的价值才能得到更加充分的发挥。健康医疗大数据的应用同样如此，只有将分散的健康医疗大数据汇集起来，基于医学专业知识进行研究和挖掘，找出其中隐含的关于人和疾病、疾病和症状、疾病和治疗手段、疾病和生活环境等关联关系，指导疾病预防、疾病研究、疾病治疗、疾病政策制定等活动，才能有效发挥健康医疗大数据的价值。

2.3.4 数据服务缺少平台，产业带动成效不显

发展大数据的价值在于利用数据为相关工作提供决策支持和数据服务，从而带动行业水平和产业结构发展。然而，由于缺乏面向企业、产业的健康医疗大数据基础平台和开放平台，缺乏相应的数据安全保护机制和授权使用机制，实际上与健康医疗紧密相关的科研机构、保险企业、医药企业、健康服务企业、信息服务企业都无法真正利用已经积累的健康医疗大数据为企业发展和业务创新提供支持，产业带动成效不显。

第3章　需求分析

3.1 业务需求分析

3.1.1 数据标准化、规范化治理及数据资源盘点需求

目前厦门市卫健委已经完成了在市级区域医疗信息平台上对全市医疗机构历史医疗数据的批量采集和现有新产生数据的实时采集工作，但是由于数据缺乏统一有效的标准化、规范化处理，难以满足更多的数据应用需求，

通过本项目的建设，厦门市卫健委将可以对现有已经采集的来源于各医疗机构复杂的、多系统的数据进行统一化处理，通过配置数据转换规则、数据整合规则、数据映射规则等，对这些历史数据进行标准化、规范化并进行有效整合和归一处理。同时，对于历史数据中海量的自然语言文本病历信息，通过利用 NLP 处理技术，将文本形式的非结构化数据进行标准结构化处理，从而形成更便于分析、利用的数据资源。帮助各医疗机构更高效更快速地提升数据价值，建立完善的医疗信息化系统。通过数据治理过程帮助卫生主管单位更有效的对现有医疗健康数据进行统计管理和资源盘点，帮助各医疗机构数据准确地反映运营、考核等信息，更全面更理性地规划数据，更迅速响应各类管控数据，提升数据效率及价值。

3.1.2 健康医疗大数据助力区域临床科研建设的需求

基于本项目建设的数据治理及知识应用平台加强临床、科研数据整合和应用，支持提供海量病历分析检索和诊疗行为分析管理等功能，为医生及临床科研人员提供基于真实病历数据的数据挖掘分析和临床数据支持，逐步促进医疗健康相关的人工智能技术研发、临床科研成果转化等。顺应互联网创新发展趋势，提升医疗健康的数字化、智能化水平，促进大数据健康产业升级。

3.1.3 健康医疗大数据带动医、药、保产业协同发展的需求

通过本项目构建的健康医疗大数据治理及知识应用平台，促进数据综合应用，可有效促进健康医疗业务与大数据技术深度融合，构建健康医疗大数据服务产业链，推动保险、药企、健康养老产业的融合发展。通过数据治理助推健康医疗大数据运营，可通过健康数据加速药品临床研发，助推创新性保险服务，以创新健康险为依托载体，实现早筛、疾病预防、健康干预管理的院内外服务，从保“健康的人”向保“人的健康”转变。

3.2 数据治理服务需求分析

3.2.1 数据治理流程

医疗数据在医院前置机上恢复之后会经过数据生产、数据归一、数据对账、数据质量评估与反馈、问题数据更新等环节。

数据生产：数据经过采集清洗过程，会按照指定的元数据结构，生产出符合规则的数据，对数据生产的每个环节都有超时报警、重试出错等报警机制，监控平台将通过短信或电子邮件方式通知给相关人员，由技术支持人员人工干预恢复数据。

数据归一：在数据生产环节后采用相关技术将具有数据值域约定的数据项进行值域归一，并下发字典的各数据项，通过相关平台提供平台统一字典的内容以及对应医院实际数据内容，自动归一服务无法自动映射的内容，将提供异常数据查询服务，由相关人员进行人工归一。

数据对账：非实时数据采集部分在每天完成 T+1 增量数据生产之后，平台将启动自动的数据对账服务，将医院前一天产生的实时数据与当天的 T+1 增量数据进行对账。数据对账方式采取各数据域数据条目对账模式，平台目前可接受的数据指标值偏差范围在+-1%之内，即前一天的实时数据中某业务域数据条目数对账值与非实时数据采集中当天 T+1 增量数据中某业务域数据条目数对账值在+-1%之内，平台将认为对账成功，否则对账失败。

数据质量评估与反馈：数据生产之后，基于庞大的数据量，构建数据质量评价服务，将对生产后的数据进行全面的数据质量评价，评价结果将对平台数据进行各种维度的数据质量描述。

问题数据更新：针对数据质量评估报告所反馈的问题，若是由数据生产问题，则由平台数据支持人员解决，若为医院原生数据质量问题，则通知医院，由医院进行数据质量改造。

数据增量更新：平台采用固定每一天进行一次增量更新（T+1），固定 7 天进行一次全量更新的增量数据更新机制。

3.2.2 数据质量改进

基于大数据平台，数据质量治理工作与数据生产和更新同步进行，并且形成对原生数据改进闭环和对数据清洗能力改进闭环，保证数据质量的持续改进。

1、 Plan 计划环节

数据质量评估主要从数据完整性、唯一性、关联性、字典一致性、值域约束、逻辑性、规范性以及可用性等维度进行医疗数据质量评估。

2、 Do 执行环节

平台发布的数据质量评价标准将通过平台的配置工具进行配置，在原生数据通过数据采集接口进入大数据平台的时候，完成数据质控。

3、 Check 检查环节

平台提供数据质量报告功能，数据质控模块调用质控规则库对进入数据生产流程的数据做质量评估。

4、 Act 处理改进环节

数据质量处理改进环节由于涉及的医院和厦门市健康医疗大数据中心两个环节，流程参照 ITIL 实践分为三个处理流程，两个处理闭环。

三个处理流程分别是：

（1）事件处理流程：流程起点是大数据平台的数据质量报告中发现的质量事件，包括数据传输损失、数据质控违规和数据趋势异常。事件处理流程客观记录事件的发生和后续处理。局端数据运营人员从数据质量事件中筛选高频发生、

对业务影响显著、高优先级的事件做分析处理，分析结果包括对医院的数据质控任务和对平台本身的配置优化任务。

（2）任务处理：由于同一个根因可能导致多个数据质量事件，所以面向根因的任务与事件并非一一对应关系。医院段数据负责人从平台接收到质控任务之后，从任务维度对院内系统调整、数据更新、业务优化、预防改进，实现数据质量的改善。任务完成后标记任务状态。提交数据校验，校验通过后关闭任务。

（3）配置更新流程：由于部分原生数据质量问题可以通过平台后期的清洗-转换加工完成，所以数据质量事件会引发对平台清洗-转换-质控规则的持续配置更新。处理改进环节重点的质控任务分发、配置文件生成、审批和生效流程，保证了持续改进中平台配置数据的一致性和可溯源。

两个业务处理闭环分别是：

（1）数据质量改进闭环：起点是数据质量报告，通过事件分析，任务分发到数据更新，质量验证通过为止完成一个闭环。

（2）配置优化改进闭环：起点同样是数据质量报告，事件分析，任务分发后通过配置文件更新监测数据质量改进完成闭环。

数据质量需满足完整性、唯一性、关联性、字典一致性、值域以及逻辑等要求。

3.3 知识应用平台系统功能需求分析

3.3.1 临床数据管理应用功能需求

面向全市临床医生及疾病管理机构，系统需提供海量病历分析、极速检索和诊疗行为分析管理等功能，建立临床诊疗专病数据模型及数据集，为医生在临床工作和科研方向探索，全市疾病管理等方向上提供数据挖掘分析和临床数据支持。主要包括以下功能：

1、疾病探索

基于大数据挖掘及数据可视化技术，帮助临床医生从既往的真实病历数据中发现临床价值和科研价值。用户可以通过搜索本市任意疾病的关键词，查看与该

疾病主题相关的指标统计数据和数据关系图谱；搜索结果中可以查询本市相关疾病的图谱及数据统计情况。

2、病历搜索

系统需实现模糊搜索、多条件组合搜索等多种病历搜索方式，也同时支持多个复杂检索逻辑、条件树搜索、事件搜索的高级搜索方式，支持设置复杂的医学事件发生逻辑，搜索符合需求的患者集合，能够快速精确搜索符合特定要求的病历或患者，以满足临床各种查询、科研、分析场景的专业搜索需求。

3、搜索结果可视化

系统需实现高级、条件树及事件搜索；支持结果显示表格化展示，且可自定义设置显示指标数据，与科研流程打通。

4、患者全景视图

系统需实现单次就诊病历查看、患者全景视图查看、时间轴查看。

5、患者分析

基于大数据加工基础，系统需实现利用各种算法模型对本市医院已积淀下的临床数据进行初步的分析透视，包括从医生、科室以及医院维度对患者的诊断、手术、检查、检验等诊疗数据进行分析。

3.3.2 科研项目管理应用功能需求

系统需支持在线按照标准流程开展科研项目，具备多种数据统计分析算法，建立临床医学及专科病种知识全库，实现图形化方式展示科研热点趋势，展示科研文献作者图谱，提升医学研究效率，方便管理医学研究成果。主要包含以下功能：

1、研究项目管理

系统需支持在线按照标准流程开展研究项目，包括以下步骤：设置项目基本信息、纳排条件设置、观测指标设置、项目结果导出、统计分析等。支持以多种数据格式导出研究项目结果数据。

2、研究项目数据开放与共享

提供研究项目 API 数据接口申请与审批服务流程，支持接口方式获得病历数据；

3、在线统计分析

系统需支持对研究项目中设置的观察指标进行统计分析，支持研究场地下的数据关联统计。

4、重要疾病领域重点指标地图

支持基于重点专科疾病领域的医学核心观测指标的展示。

5、知识全库

系统需支持医学文献、指南共识、临床路径、药品说明书、临床试验、误诊误治、及支持基因知识库相关内容展示、浏览、搜索；支持图形化方式展示研究热点趋势；支持研究文献作者图谱展示。

6、科研架构

构建集基础临床数据治理、专科数据治理为一体的医疗数据科研架构，建立基础临床数据库、科研专病数据库，打造疾病协同网络，提升科研效能。

3.4 非功能性需求分析

3.4.1 大数据存储与计算需求

传统架构与数据技术难以支撑 TB 级数据秒级查询。本项目应基于 Spark/Hadoop 系统构建，采用大数据存储与计算技术，为各类业务应用场景和数据产品提供了丰富的数据治理能力，提供分布式文件集群系统、流式处理系统、关系型数据库集群系统、大数据查询引擎、分布式消息系统、内存数据库集群系统。形成统一、标准化、安全的数据服务接口，支持跨系统、跨应用对内对外各类数据服务场景。

3.4.2 海量数据秒级查询搜索需求

为了满足大数据时代数据激增、海量数据高速查询分析的需求，应采用Kylin、ES、MongoDB 等先进技术开发设计平台，以满足高可用、高并发、高负载的需求。

3.4.3 可靠性、易用性和可扩展性指标

系统在可靠性、易用性和可扩展性等方面，需满足以下要求：

- 可靠性要满足系统 7×24 小时不间断服务的要求。
- 系统可用率≥99.99%，即每年的不可用时间小于 9 小时。
- 采用标准接口、界面友好、使用方便。
- 支撑基础设施、软件结构、数据库等方面的设计能满足功能不断扩展，以及系统容量和用户数量不断增长的要求。

3.4.4 网络平台性能需求

要求数据传输网络畅通、快捷、可扩展。核心网络要求设备、线路均具有冗余，设备处理能力满足业务高峰期需要。整网带宽满足业务高峰期需要。
本项目建设内容较多，其业务形式多样、覆盖范围广泛，所有信息系统功能的实现都离不开顺畅的网络通信系统，因而建设功能完善、先进可靠的网络系统是实现各种系统功能的基础。项目网络需具备畅通、快捷、安全的特点，具体而言，应具备以下功能、性能特点：

1、网络可用性、可靠性保障

针对日益增长的医疗业务通信要求，需要满足通达各级医疗机构的集数据、语音、图像实时传输为一体的基础通信网络，从根本上解决网络的可用性和可靠性。

2、实现网络间互连互通，并具备良好的可扩展性

实现本级专线网节点单位之间的互连互通；实现和上、下级专线网的联接，并兼容上、下级专线网的接入设备；与电子政务网等其它网络互连互通的扩充能力。

3、高质量地支持多种应用

高质量地支持以下应用：专线电话；视频会议；数据广域传输；可扩充的其他应用。

4、实施全面的 QoS 策略，确保业务服务质量

为不同的业务提供不同的 QoS 保证，实现有效的带宽管理控制和带宽综合利用。

5、实现安全业务通道

确保不同业务传输的优先级别和 QoS 保障前提下，能实现不同业务之间的隔离和受控访问。

6、全网统一网管、分级负责

在网络建设、运营和维护管理中，实施全网统一网管、分级负责是确保更好地利用网络资源，确保网络可靠和安全运行的重要手段，网络管理应能实现网管安全联动，即全网出现异常事件时能迅速启动应急机制。

3.4.5 系统平台性能需求

要求采用通用性好、安全可靠的操作系统以及大型数据库系统，以保证系统良好的性能。

3.4.6 应用系统性能需求

1、应用系统性能应满足用户的要求，稳定、可靠、实用。

2、人机界面友好，输出、输入方便，图表生成美观，检索、查询简单快捷。

3、系统采用便于升级的模块化设计，包括采用软件升级来简化系统扩展和修改，模块组合可以根据需要来选择。

4、提供标准的网络通信应用层协议和应用基本函数及调用接口。

3.4.7 数据备份需求

平台应具有完善的备份和恢复机制，在异常情况发生时，可以快速恢复，避免数据的丢失或将其影响降到最低限度。

在数据治理过程中每一步处理流程要保留处理痕迹，确保处理后的数据可以溯源到原始数据，能够做到对每一层处理的数据及处理前的原始数据进行有效的数据备份。

3.5 信息安全需求

安全管理策略是信息安全的首要保障，一个体系化、动态化、区域化、以人为本的安全策略是必不可少的。通过安全策略，制定安全管理和安全技术的相关规范、标准及要求，才能使得信息系统的所有者、维护者及使用者的工作有章可循。

项目建设需要保障网络安全、数据安全，需要保障系统可靠运行。网络安全首先要考虑技术层面的安全性，同时考虑管理层面的安全性；数据安全主要是保证数据的原始性和完整性，包括数据不被非法修改和访问，数据的全面完整，数据的访问和修改可追踪等等，同时提供合适的数据备份策略；可靠性是指系统应具备在硬件或网络故障时的运行和修复能力，同时系统在设计时必须考虑大规模并发、不断扩展条件下的运行可靠性。

同时本项目中包含了敏感数据，这就使得信息安全保障工作显得尤为重要，需要遵照“层次化管理、多级防范”的建设思路，做好网络层、主机层、中心数据库的安全防护工作。

第4章　项目建设方案

4.1 总体设计方案

4.1.1 项目总体定位

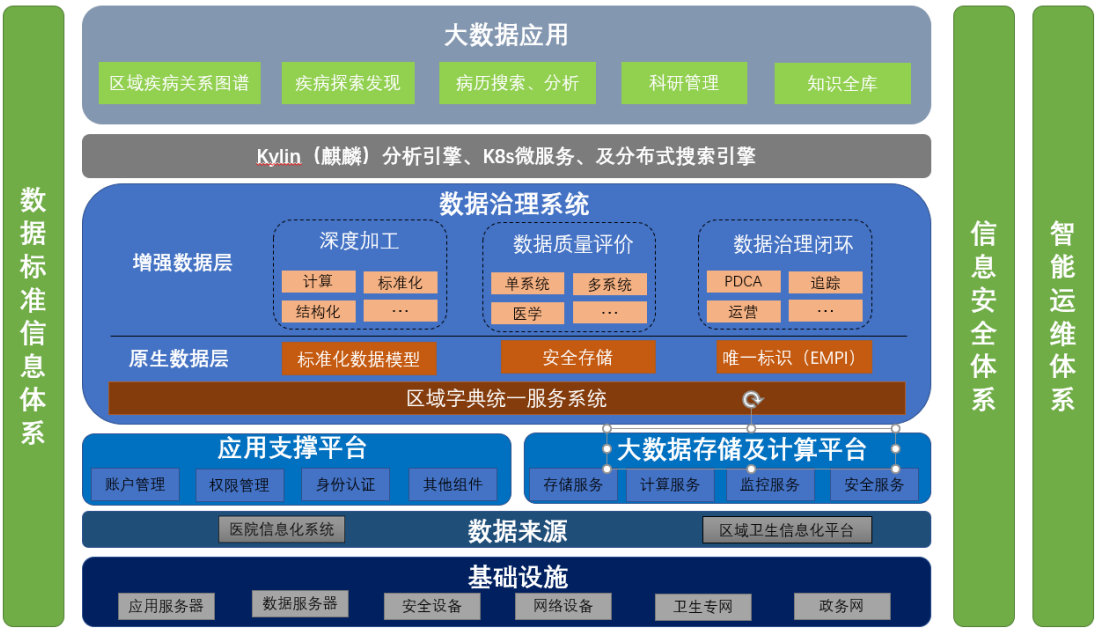
1、本项目从节约资源的角度出发，充分利旧，通过对旧的系统进行升级改造及整合建设，以达到公共卫生、医疗卫生、医学研究的信息化融合建设目标；将升级改造后的系统统一在厦门市医疗云部署，避免了一部分硬件设施的重复建设。

2、本项目将充分依托厦门市卫健委已经完成的16家三级以上医疗机构数据及基层社区卫生服务中心采集的数据来源和数据采集基础，建立基于大数据技术平台的数据治理服务和数据知识应用平台。充分预留与相关业务平台的数据接口，实现卫生计生业务数据及高质量医学挖掘数据的有效共享。

3、通过本项目的建设，对临床诊疗、公共卫生和医疗卫生资源进行整合、重组，可以实现数据资源的可理解性、可信性以及可访问性；规范统一的基本数据元素，使数据管理制度化，可以保证不同层次之间的一致性；使数据可以被“智慧健康”不同应用中的用户重复使用和灵活配置，赋能临床医学研究，方便居民日常就医诊疗，实现“一次开发，多次利用”，提高效率和灵活性，提高数据资产的价值和可复用性；为厦门市“智慧健康”建设奠定数据基础。

4.1.2 系统框架设计

本项目提供围绕全市临床数据、公卫数据、运营数据等卫生计生相关数据采集、汇聚、生产功能，对汇聚的数据进行深层次的数据治理，依托Hadoop、K8s、Elasticsearch、Kylin以及安全管理技术等应用支撑平台，建立面向全市提供服务的健康医疗大数据知识应用平台，平台整体架构设计如下图：



图：5-1 系统框架设计图

4.1.3 网络架构设计

本项目统一部署在厦门市卫健委医疗云，平台本身划分为多个集群进行部署，集群之间通过核心交换机相连，实现数据和请求的打通。

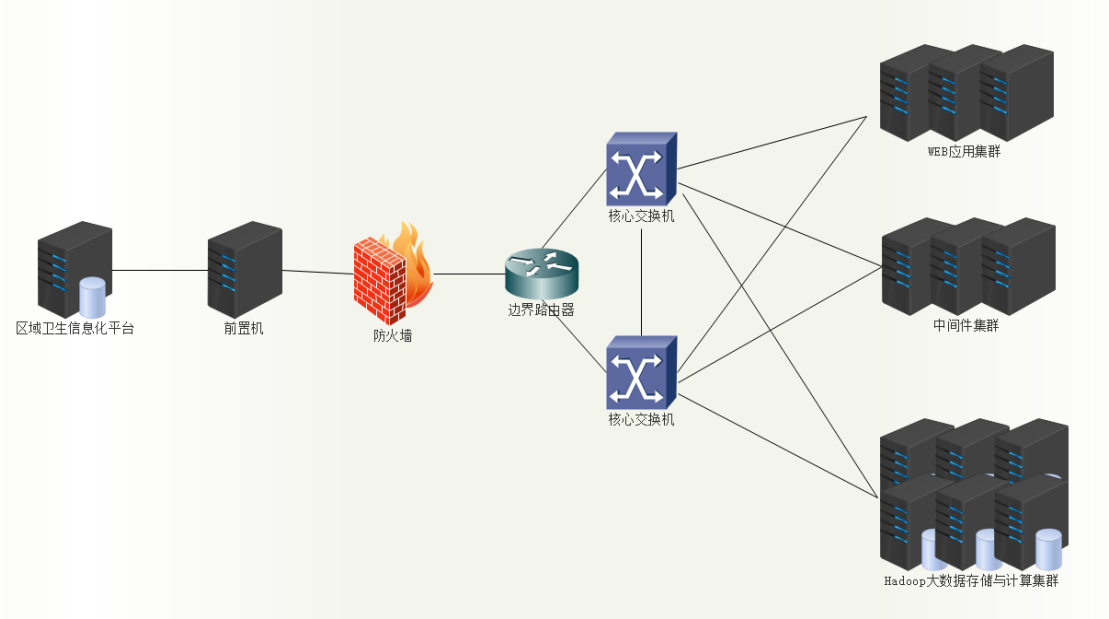


图 5-2 网络架构图

4.2 总体技术路线

4.2.1 数据标准化技术

数据标准化技术使用在数据治理阶段，在数据治理过程中将表示同一种含义的多种称呼的医疗术语，统一为标准化的名称，同时梳理不同诊断的流程关系，满足后续数据分析的需要。标准化的处理首先以标准表作为基准表，将医疗数据中实际出现医疗词往基准表上进行映射,映射过程包括机器处理加人工标注 2 个过程。当标准表不能对应实际的数据时，由专业的医生来决定是否扩充标准表。以诊断为例，标准化过程中，不同的诊断词之间的关系不同，如“白血病”和“血癌”、“艾滋病”和“获得性免疫综合症”是两组同义的诊断词，诊断词“重度脂肪肝”是“脂肪肝”的一种类型，存在包含/被包含关系。通过标准化，最终会将临床诊断与 ICD11 一一对应起来。如诊断“甲减”归一到 ICD11 疾病“甲状腺功能减退”，编码：E03.901。

4.2.2 多源异构数据治理技术

多源异构数据治理技术使用在大数据计算集群服务器进行多源异构数据治理阶段，可以快速对前置机备份过来的医院的多源异构数据结构进行识别，并快速建立不同数据表之间的关联辨析，实现数据整合。数据整合分为患者维度、就诊维度、医嘱维度，并且通过智能的运算模型及不同的数据元标准，对数据进行清洗加工。

4.2.3 自然语言处理技术

自然语言处理技术使用在对电子病历的非结构化数据治理阶段，在电子病历医疗数据中，存在大量自然语言文本记录的数据，包括大量有价值信息，故需要利用自然语言处理技术进行数据结构化处理。本技术进行数据结构化处理将分为

三个主要层次，分别是基础数据层，自然语言挖掘算法层，以及结构化系统层。基础数据治理层主要是通过整合权威医学标准、大规模专业词库以及真实临床医学词库构建基础词库。然后挖掘算法层利用自然语言识别模型的训练进行实体识别，关键词提取，关系识别分类等工作。最后在结构化系统通过工具化人机协同工作针对临床病历数据进行疾病、手术、药品、症状、诊断、检查检验等数据项的结构化工作。结构化主要从若干个独立维度来进行,对数据依据主题字段进行划分，主要主题字段有；症状、体征、烟酒情况、病理诊断、病理表现、过敏情况、婚育状况等。根据病理或报告中不同字段的语义复杂程度和实际需求,目前结构化框架主要由抽取框架和通用框架组成。

4.2.4 大数据治理技术

大数据治理技术应用在数据采集、清洗、转换、关联、数据质控等过程，通过人机结合的方式实现高效精准的数据治理，通过大规模自动化的采集、清洗、归类、关联数据，提升数据分析利用的准确性和实用性，形成统一数据视图为后续系统提供服务。医疗机构首次接入的系统数据需要人工干预进行配置，基于机器学习的数据自动转换能力能够将大量从事数据治理的人力工作解放，投入到真正的业务实现工作中去。数据治理有以下要点：

1）迅速构建从关系表到知识图谱的目标数据模型；

2）复杂数据转换规则的设计与实现：完成机器学习规则、嵌套规则、组合规则、数据字典规则等；

3）高精度全局正确性验证：实现各种环节、各个病种的校验规则；

4）自洽性高的内部元数据管理：实现源表、目标表、映射、规则等内部存储机制；

5）适配性高的任务调度：实现满足数据增量、全量、全量式增量、回滚等任务的自动化调度。

数据治理主要参考以下规范：

GB/T 36344—2018 信息技术数据质量评价指标；

住院病案首页数据质量管理与控制指标（2016 版）；

针对医疗信息数据的数据质量评价维度，各参考规范都有明确说明，涵盖的指标各有侧重，分析对比详情如下表所示：

表 5-1 规范对照表

评价指标	GB36344	电子病历评级	ISO/IEC 25024:2015	电子健康档案数据质量评估与治理的综述研究			
				CIHI Data Quality Framework	Chan K S, Fowles J B	Weiskopf N G, Weng C	Reimer A P, Milinovich A
规范性	√	√	√				
完整性	√	√	√		√	√	√
准确性	√		√	√	√	√	√
一致性	√	√	√	√		√	√
时效性	√	√		√	√	√	√
可访问性	√		√				
合理性						√	√
可比性				√	√		
可用性				√			
可信性			√				
现实性			√				
可跟踪性			√				
可恢复性			√				

从上表看出，各参考规范选取的评价指标虽各有不同，但部分核心指标皆有覆盖，即满足规范性、完整性、准确性、一致性以及时效性。

4.2.5 联机分析处理

联机分析处理 OLAP 是一种软件技术，使用在数据应用阶段，帮助使用人员能够迅速、一致、交互地观察各类信息，以达到深入理解数据的目的。FASMI(Fast Analysis of Shared Multidimensional Information)，即共享多维信息的快速分析的特征。其中 F 是快速性(Fast)，指系统能在数秒内对用户的多数分析要求做出反应；A 是可分析性(Analysis)，指用户无需编程就可以定义新的专门计算，将其作为分析的一部分，并以用户所希望的方式给出报告；M 是多维性(Multi—dimensional)，指提供对数据分析的多维视图和分析；I 是信息性(Information)，指能及时获得信息，并且管理大容量信息。

4.2.6 Kubernetes（微服务架构）

Kubernetes 是一个全新的基于容器技术的分布式架构技术路线，使用在数据应用架构搭建阶段，Kubernetes(k8s)是 Google 开源的容器集群管理系统。在 Docker 技术的基础上，为容器化的应用提供部署运行、资源调度、服务发现和动态伸缩等一系列完整功能，提高了大规模容器集群管理的便捷性。

Kubernetes 是一个完备的分布式系统支撑平台，具有完备的集群管理能力，多扩多层次的安全防护和准入机制、多租户应用支撑能力、透明的服务注册和发现机制、内建智能负载均衡器、强大的故障发现和自我修复能力、服务滚动升级和在线扩容能力、可扩展的资源自动调度机制以及多粒度的资源配额管理能力。同时 Kubernetes 提供完善的管理工具，涵盖了包括开发、部署测试、运维监控在内的各个环节。其核心概念包括：

1、Node

Node 作为集群中的工作节点，运行真正的应用程序，在 Node 上 Kubernetes 管理的最小运行单元是 Pod。Node 上运行着 Kubernetes 的 Kubelet、kube-proxy 服务进程，这些服务进程负责 Pod 的创建、启动、监控、重启、销毁、以及实现软件模式的负载均衡。

2、Pod

Pod 是 Kubernetes 最基本的操作单元，包含一个或多个紧密相关的容器，一个 Pod 可以被一个容器化的环境看作应用层的“逻辑宿主机”；一个 Pod 中的多个容器应用通常是紧密耦合的，Pod 在 Node 上被创建、启动或者销毁；每个 Pod 里运行着一个特殊的被称之为 Pause 的容器，其他容器则为业务容器，这些业务容器共享 Pause 容器的网络栈和 Volume 挂载卷，因此他们之间通信和数据交换更为高效，在设计时我们可以充分利用这一特性将一组密切相关的服务进程放入同一个 Pod 中。

3、Service

一个 Service 可以看作一组提供相同服务的 Pod 的对外访问接口，Service 作用于哪些 Pod 是通过 Label Selector 来定义的。

- 拥有一个指定的名字（比如 my-mysql-server）；
- 拥有一个虚拟 IP（Cluster IP、Service IP 或 VIP）和端口号，销毁之前不会改变，只能内网访问；
- 能够提供某种远程服务能力；
- 被映射到了提供这种服务能力的一组容器应用上；

如果 Service 要提供外网服务，需指定公共 IP 和 NodePort，或外部负载均衡器；

4、Volume

Volume 是 Pod 中能够被多个容器访问的共享目录。

5、Label

Label 以 key/value 的形式附加到各种对象上，如 Pod、Service、RC、Node 等，以识别这些对象，管理关联关系等，如 Service 和 Pod 的关联关系。

6、RC（Replication Controller）

- 目标 Pod 的定义；
- 目标 Pod 需要运行的副本数量；
- 要监控的目标 Pod 标签（Lable）；

Kubernetes 通过 RC 中定义的 Lable 筛选出对应的 Pod 实例，并实时监控其状态和数量，如果实例数量少于定义的副本数量（Replicas），则会根据 RC 中定义的 Pod 模板来创建一个新的 Pod，然后将此 Pod 调度到合适的 Node 上启动运行，直到 Pod 实例数量达到预定目标。

4.2.7 Hadoop

Hadoop 是使用 java 编写的 Apache 开放源代码框架，在本项目中对汇总的健康医疗大数据进行分布式的集中处理。Hadoop 框架工作的应用程序可以在跨计算机群集提供分布式存储和计算的环境中工作。Hadoop 旨在从单一服务器扩展到数千台机器，每台机器都提供本地计算和存储。

Hadoop 框架包括以下四个模块：

Hadoop Common：这些是其他 Hadoop 模块所需的 Java 库和实用程序。这些

库提供文件系统和操作系统级抽象，并包含启动 Hadoop 所需的必要 Java 文件和脚本。

Hadoop YARN：这是作业调度和集群资源管理的框架。

Hadoop 分布式文件系统（HDFS）：提供对应用程序数据的高吞吐量访问的分布式文件系统。

Hadoop MapReduce：这是基于 YARN 的大型数据集并行处理系统。

自 2012 年以来，术语“Hadoop”通常不仅指向上述基本模块，而且还指向可以安装在 Hadoop 之上或之外的其他软件包，例如 Apache Pig，Apache Hive，Apache HBase，Apache Spark 等。

MapReduce

Hadoop MapReduce 是一个用于轻松编写应用程序的软件框架，它以可靠，容错的方式在大型集群（数千个节点）上处理大量数据并行处理商品硬件。

术语 MapReduce 实际上是指 Hadoop 程序执行的以下两个不同的任务：

Map Task：这是第一个任务，它接收输入数据并将其转换成一组数据，其中单个元素分解为元组（键/值对）。

Reduce Task：此任务将地图任务的输出作为输入，并将这些数据元组合并为较小的一组元组。reduce 任务总是在 map 任务之后执行。

通常输入和输出都存储在文件系统中。该框架负责调度任务，监视它们并重新执行失败的任务。

MapReduce 框架由单个主 JobTracker 和每个群集节点的一个从属 TaskTracker 组成。主管负责资源管理，跟踪资源消耗/可用性，并对从站上的作业组件任务进行调度，监控和重新执行故障任务。从站 TaskTracker 按照主机的指示执行任务，并定期向主设备提供任务状态信息。

JobTracker 是 Hadoop MapReduce 服务的单点故障，这意味着如果 JobTracker 关闭，则所有正在运行的作业都将停止。

Hadoop 分布式文件系统

Hadoop 可以直接与任何可安装的分布式文件系统（如本地 FS，HFTP FS，S3 FS 等）工作，但 Hadoop 使用的最常见的文件系统是 Hadoop 分布式文件系统（HDFS）。

Hadoop 分布式文件系统（HDFS）基于 Google 文件系统（GFS），并提供一个分布式文件系统，旨在以可靠，容错的方式在大型计算机（数千台计算机）上运行小型计算机。

HDFS 使用主/从架构，其中主机由管理文件系统元数据的单个 NameNode 和存储实际数据的一个或多个从属数据节点组成。

HDFS 命名空间中的文件被分成几个块，这些块被存储在一组 DataNodes 中。NameNode 确定块到 DataNodes 的映射。DataNodes 负责文件系统的读写操作。他们还根据 NameNode 给出的指令来处理块创建，删除和复制。

HDFS 提供了像任何其他文件系统一样的 shell，并且可以使用命令列表与文件系统进行交互。这些 shell 命令将在一个单独的章节中以及适当的示例进行介绍。

Hadoop 的优点

- Hadoop 框架允许用户快速编写和测试分布式系统。它是高效的，它自动分配数据并在机器上工作，反过来利用 CPU 核心的底层并行性。
- Hadoop 不依赖硬件提供容错和高可用性（FTHA），而是 Hadoop 库本身被设计为检测和处理应用层的故障。
- 服务器可以动态添加或从集群中删除，Hadoop 继续运行而不会中断。
- Hadoop 的另一大优点是，除了是开放源码，它是所有平台兼容的，因为它是基于 Java 的。

4.2.8 Kylin

Apache Kylin™是一个开源的分布式分析引擎，在本项目中应用在健康医疗大数据的查询分析中，提供 Hadoop/Spark 之上的 SQL 查询接口及多维分析（OLAP）能力以支持超大规模数据，最初由 eBay Inc. 开发并贡献至开源社区。它能在亚秒内查询巨大的 Hive 表。

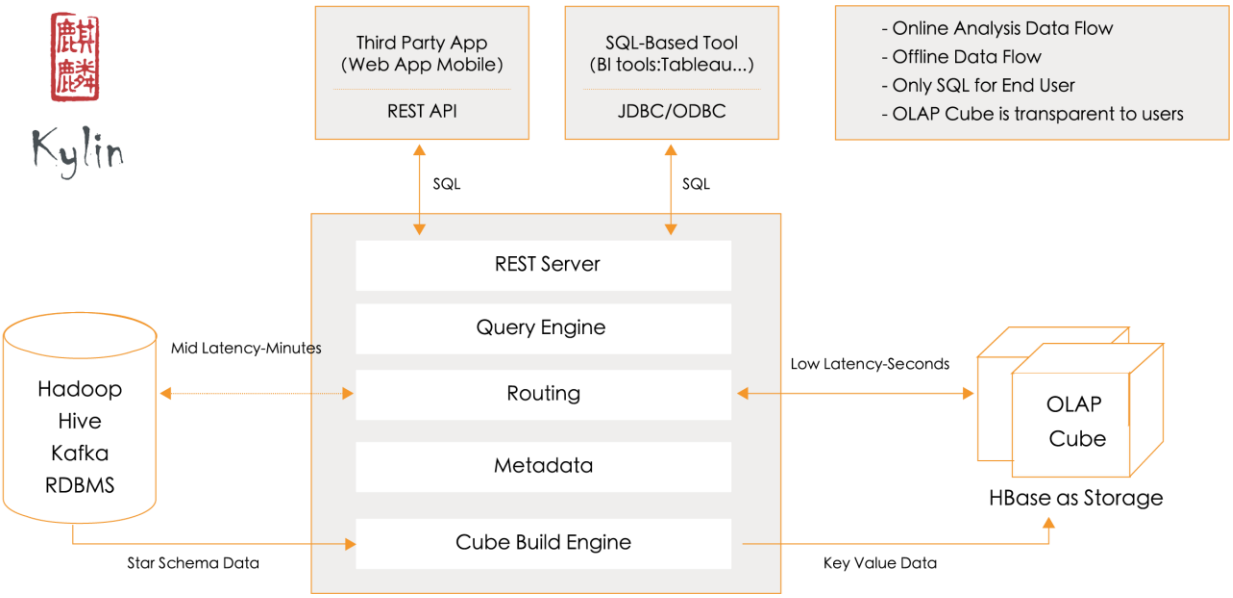


图 5-3 Kylin 技术架构图

Kylin 具有如下特性：

- 可扩展超快 OLAP 引擎：Kylin 是为减少在 Hadoop/Spark 上百亿规模数据查询延迟而设计
- Hadoop ANSI SQL 接口：Kylin 为 Hadoop 提供标准 SQL 支持大部分查询功能
- 交互式查询能力:通过 Kylin,用户可以与 Hadoop 数据进行亚秒级交互，在同样的数据集上提供比 Hive 更好的性能
- 多维立方体（MOLAP Cube）：用户能够在 Kylin 里为百亿以上数据集定义数据模型并构建立方体
- 与 BI 工具无缝整合：Kylin 提供与 BI 工具的整合能力，如 Tableau，PowerBI/Excel，MSTR，QlikSense，Hue 和 SuperSet
- 支持 LDAP、SSO

Kylin 的生态圈包括：

- Kylin 核心：Kylin OLAP 引擎基础框架，包括元数据（Metadata）引擎，查询引擎，Job 引擎及存储引擎等，同时包括 REST 服务器以响应客户端请求
- 扩展：支持额外功能和特性的插件
- 整合：与调度系统，ETL，监控等生命周期管理系统的整合
- 用户界面：在 Kylin 核心之上扩展的第三方用户界面

- 驱动：ODBC 和 JDBC 驱动以支持不同的工具和产品，比如 Tableau

4.2.9 HBase

HBase 是一个分布式的、面向列的开源数据库，在本项目中 HBase 负责存储数据工程中各个环节的数据，该技术来源于 Fay Chang 所撰写的 Google 论文“Bigtable：一个结构化数据的分布式存储系统”。就像 Bigtable 利用了 Google 文件系统（File System）所提供的分布式数据存储一样，HBase 在 Hadoop 之上提供了类似于 Bigtable 的能力。HBase 是 Apache 的 Hadoop 项目的子项目。HBase 不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是 HBase 基于列的而不是基于行的模式。HBase 利用 Hadoop HDFS 作为其文件存储系统，利用 Hadoop MapReduce 来处理 HBase 中的海量数据，利用 Zookeeper 作为其分布式协同服务。主要用来存储非结构化和半结构化的松散数据（列存 NoSQL 数据库）。

HBase 的特性

高性能

HBase 中存储了一套 HDFS 的索引，通过表名－>行健－>列族－>列限定符－>时间版本这一套索引来定位数据的位置，HBase 为每一列数据维护了一套索引规则，对于具体某一具体条数据的查询可以非常快速的通过 B＋树定位数据存储位置并将其取出。

另外，HBase 通常以集群部署，数据被分散到多个节点存储，当客户端发起查询请求的时候，集群里面多个节点并行执行查询操作，最后将不同节点的查询结果进行合并返回给客户端，提高 IO 性能。

高可用

HBase 集群中任意一个节点宕机都不会导致集群瘫痪。这取决于两方面原因：

第一方面，ZooKeeper 解决了 HBase 中心化问题；

另一方面，HBase 将数据存放在 HDFS 上面，HDFS 的数据冗余存放在不同节点，一个节点瘫痪可从其他节点取得数据，保证了 HBase 的高可用。

易扩展

Hbase 的扩展性主要体现在两个方面，一个是基于上层处理能 RegionServer 的扩展，一个是基于存储的扩展 HDFS。

无模式

使用 HBase 不需要预先定义表中有多少列，也不需要定义每一列存储的数据类型，HBase 在需要的时候可以动态增加列和指定存储数据类型。

4.2.10 MongoDB

MongoDB 是由 C++语言编写的，是一个基于分布式文件存储的开源数据库系统。在本项目中 MongoDB 应用在线索引信息的存储，在高负载的情况下，添加更多的节点，可以保证服务器性能。MongoDB 旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。MongoDB 将数据存储为一个文档，数据结构由键值(key=>value)对组成。MongoDB 文档类似于 JSON 对象。字段值可以包含其他文档，数组及文档数组。

4.2.11 Kafka

Kafka 是由 LinkedIn 开发的一个分布式的消息系统，在本项目中 Kafka 应用在将前置机的数据提取到大数据治理集群的环节。使用 Scala 编写，它以可水平扩展和高吞吐率而被广泛使用。目前越来越多的开源分布式处理系统如 Cloudera、Apache Storm、Spark 都支持与 Kafka 集成。

为何使用消息系统

解耦

在项目启动之初来预测将来项目会碰到什么需求，是极其困难的。消息系统在处理过程中间插入了一个隐含的、基于数据的接口层，两边的处理过程都要实现这一接口。这允许你独立的扩展或修改两边的处理过程，只要确保它们遵守同样的接口约束。

冗余

有些情况下，处理数据的过程会失败。除非数据被持久化，否则将造成丢失。

消息队列把数据进行持久化直到它们已经被完全处理，通过这一方式规避了数据丢失风险。许多消息队列所采用的"插入-获取-删除"范式中，在把一个消息从队列中删除之前，需要你的处理系统明确的指出该消息已经被处理完毕，从而确保你的数据被安全的保存直到你使用完毕。

扩展性

因为消息队列解耦了你的处理过程，所以增大消息入队和处理的频率是很容易的，只要另外增加处理过程即可。不需要改变代码、不需要调节参数。扩展就像调大电力按钮一样简单。

灵活性 & 峰值处理能力

在访问量剧增的情况下，应用仍然需要继续发挥作用，但是这样的突发流量并不常见；如果为以能处理这类峰值访问为标准来投入资源随时待命无疑是巨大的浪费。使用消息队列能够使关键组件顶住突发的访问压力，而不会因为突发的超负荷的请求而完全崩溃。

可恢复性

系统的一部分组件失效时，不会影响到整个系统。消息队列降低了进程间的耦合度，所以即使一个处理消息的进程挂掉，加入队列中的消息仍然可以在系统恢复后被处理。

顺序保证

在大多使用场景下，数据治理的顺序都很重要。大部分消息队列本来就是排序的，并且能保证数据会按照特定的顺序来处理。Kafka 保证一个 Partition 内的消息的有序性。

缓冲

在任何重要的系统中，都会有需要不同的处理时间的元素。例如，加载一张图片比应用过滤器花费更少的时间。消息队列通过一个缓冲层来帮助任务最高效率的执行——写入队列的处理会尽可能的快速。该缓冲有助于控制和优化数据流经过系统的速度。

异步通信

很多时候，用户不想也不需要立即处理消息。消息队列提供了异步处理机制，允许用户把一个消息放入队列，但并不立即处理它。想向队列中放入多少消息就

放多少，然后在需要的时候再去处理它们。

4.2.12 Sqoop

Apache Sqoop（SQL-to-Hadoop） 项目旨在协助 RDBMS 与 Hadoop 之间进行高效的大数据交流。在本项目中 Sqoop 用在将前置机的数据提取到大数据治理集群的环节中。用户可以在 Sqoop 的帮助下，轻松地把关系型数据库的数据导入到 Hadoop 与其相关的系统 (如 HBase 和 Hive)中；同时也可以把数据从 Hadoop 系统里抽取并导出到关系型数据库里。除了这些主要的功能外，Sqoop 也提供了一些诸如查看数据库表等实用的小工具。

理论上，Sqoop 支持任何一款支持 JDBC 规范的数据库，如 DB2、MySQL 等。Sqoop 还能够将 DB2 数据库的数据导入到 HDFS 上，并保存为多种文件类型。常见的有定界文本类型，Avro 二进制类型以及 SequenceFiles 类型。

Sqoop 中一大亮点就是可以通过 Hadoop 的 MapReduce 把数据从关系型数据库中导入数据到 HDFS。Sqoop 架构非常简单，其整合了 Hive、Hbase 和 Oozie，通过 Map-Reduce 任务来传输数据，从而提供并发特性和容错。

4.2.13 安全技术

PKI 体系

公钥基础设施（Public Key Infrastructure，简称 PKI）是目前网络安全建设的基础与核心，是电子商务安全实施的基本保障，因此，对 PKI 技术的研究和开发成为目前信息安全领域的热点。

PKI 采用证书进行公钥管理，通过第三方的可信任机构（认证中心，即 CA），把用户的公钥和用户的其他标识信息捆绑在一起，其中包括用户名和电子邮件地址等信息，以在 Internet 网上验证用户的身份。PKI 把公钥密码和对称密码结合起来，在 Internet 网上实现密钥的自动管理，保证网上数据的安全传输。

因此，从大的方面来说，所有提供公钥加密和数字签名服务的系统，都可归结为 PKI 系统的一部分，PKI 的主要目的是通过自动管理密钥和证书，为用户建立起一个安全的网络运行环境，使用户可以在多种应用环境下方便的使用加密和

数字签名技术，从而保证网上数据的机密性、完整性、有效性。数据的机密性是指数据在传输过程中，不能被非授权者偷看；数据的完整性是指数据在传输过程中不能被非法篡改；数据的有效性是指数据不能被否认。

一个有效的 PKI 系统必须是安全的和透明的，用户在获得加密和数字签名服务时，不需要详细地了解 PKI 的内部运作机制。在一个典型、完整和有效的 PKI 系统中，除证书的创建和发布，特别是证书的撤销，一个可用的 PKI 产品还必须提供相应的密钥管理服务，包括密钥的备份、恢复和更新等。没有一个好的密钥管理系统，将极大影响一个 PKI 系统的规模、可伸缩性和在协同网络中的运行成本。在一个企业中，PKI 系统必须有能力为一个用户管理多对密钥和证书；能够提供安全策略编辑和管理工具，如密钥周期和密钥用途等。

PKI 作为一组在分布式计算系统中利用公钥技术和 X.509 证书所提供的安全服务，企业或组织可利用相关产品建立安全域，并在其中发布密钥和证书。在安全域内，PKI 管理加密密钥和证书的发布，并提供诸如密钥管理（包括密钥更新，密钥恢复和密钥委托等）、证书管理（包括证书产生和撤销等）和策略管理等。PKI 产品也允许一个组织通过证书级别或直接交叉认证等方式来同其他安全域建立信任关系。这些服务和信任关系不能局限于独立的网络之内，而应建立在网络之间和 Internet 之上，为电子商务和网络通信提供安全保障，所以具有互操作性的结构化和标准化技术成为 PKI 的核心。

4.2.14 数据可视化

数据可视化在本项目中多个应用系统建设都会使用数据可视化技术，其中，这种数据的视觉表现形式被定义为，一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。数据可视化技术的基本思想，是将数据库中每一个数据项作为单个图元元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析。

数据可视化技术包含以下几个基本概念：

数据空间：是由 n 维属性和 m 个元素组成的数据集所构成的多维信息空间；

数据开发：是指利用一定的算法和工具对数据进行定量的推演和计算；

数据分析：指对多维数据进行切片、块、旋转等动作剖析数据，从而能多角度多侧面观察数据；

数据可视化：是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

数据可视化已经提出了许多方法，这些方法根据其可视化的原理不同可以划分为基于几何的技术、面向像素技术、基于图标的技术、基于层次的技术、基于图像的技术和分布式技术等等。

4.2.15 大数据挖掘与分析技术

本项目将使用大数据挖掘与分析技术，如建立疾病图谱等。将会整合和利用来自不同医疗服务机构和业务分散的信息和数据，并且随着社会的发展，围绕业务数据项的扩充也将成为必然，如何用科学的方法去整理数据，从而从不同视角为各部门的数据分析、宏观决策等提供依据，比以往更为迫切。数据仓库作为面向主题的、集成的、与时间相关的、不可修改的数据集合，将使得此项技术的应用在未来成为系统“数据整合、服务集成”的有效手段。数据仓库将能提供一个统一的数据治理平台，包括：业务信息相关数据的筛选、整理、分类、调整、汇总、普通计算、图表分析等，这些功能的应用将有效支持高级查询和决策分析需求的实现。

4.3 本期项目建设方案

4.3.1 标准规范

4.3.1.1 标准体系参考

本项目的建设过程中，会用到很多标准，在医疗行业，有很多标准已经足够成熟，可以直接拿来用，无须新建，当本项目实施过程中国家标准和行业标准涵

盖不到，却又需要有统一标准时，再由市卫健委制定相关标准，本项目暂不考虑重新制定标准。以下为参考标准：

4.3.1.1.1 基础标准

本系统应遵循和参考的基础类规范如下：

表 5-2 基础规范表

一级类目	二级类目	发布单位
基础规范	基础引用类标准	标准委等
	卫生信息数据元标准化规则	卫生部
	卫生信息数据模式描述指南	卫生部
	卫生信息数据集元数据规范	卫生部
	卫生信息数据集分类与编码	卫生部
	电子病历基本数据集编制规范(征求意见稿）	卫生部
	健康档案基本数据集编制规范（试行）	卫生部
	健康档案公用数据元标准（试行）	卫生部

4.3.1.1.2 技术类标准

本系统遵循的数据规范如下：

表 5-3 数据规范表

一级类目	二级类目	发布单位
数据规范	《卫生信息数据元目录》	卫生部
	《城乡居民健康档案基本数据集》	卫生部
	电子病历基本架构与数据标准	卫生部
	电子病历基本内容架构图	卫生部
	电子病历数据组与数据元(征求意见稿）	卫生部
	电子病历基础模板：（试行）	卫生部

一级类目	二级类目	发布单位
	国家卫生数据字典与元数据管理（试行）	卫生部

4.3.1.1.3 功能规范标准

本系统遵循和参考的功能规范如下：

表 5-4 功能规范表

一级类目	二级类目	发布单位
功能规范	妇幼保健信息系统基本功能规范（试行）	卫生部
	社区卫生信息系统功能规范（试行）	卫生部
	基本医疗卫生信息系统基本功能规范	卫生部
	……	卫生部

4.3.1.1.4 信息基础设施规范标准

信息基础设施相关的标准化主要的目的是为基础设施在选择时提供规范的约束和指导。本系统遵循和参考的基础设施规范如下：

表 5-5 基础设施规范表

一级类目	二级类目	发布单位
基础设施规范	基于健康档案的区域卫生信息平台建设技术解决方案（试行）	卫生部
	基于健康档案与区域卫生信息平台的妇幼保健信息系统技术解决方案（试行）	卫生部
	综合卫生管理信息平台建设指南（试行）	卫生部
	国家卫生信息网络直报系统设计方案	卫生部
	……	卫生部

4.3.1.1.5 安全与隐私保护类标准

本平台遵循和参考的安全规范如下：

表 5-6 安全规范表

一级类目	二级类目	发布单位
安全规范	卫生系统电子认证服务规范（试行）	卫生部
	卫生系统数字证书应用集成规范（试行）	卫生部
	卫生系统数字证书格式规范（试行）	卫生部
	卫生系统数字证书介质技术规范	卫生部
	卫生系统数字证书服务管理平台接入规范（试行）	卫生部
	……	卫生部

4.3.1.1.6 业务类标准

本系统遵循和参考的业务规范如下：

表 5-7 业务类规范表

一级类目	二级类目	发布单位
业务类规范	国家基本公共卫生服务规范	卫生部
	健康档案基本卫生服务记录表单参考用表	卫生部
	卫生部办公厅关于印发消化系统 6 个病种临床路径的通知	卫生部
	卫生部办公厅关于印发肾脏内科专业 4 个临床路径的通知	卫生部
	卫生部办公厅关于印发心血管系统 6 个病种临床路径的通知	卫生部

	慢病综合干预技术方案	卫生部
	营养与健康监测技术方案	卫生部
	肿瘤随访登记技术方案	卫生部

4.3.1.2 标准体系建议

本项目的建设过程中，会用到很多标准，在医疗行业，有很多标准已经足够成熟，可以直接拿来用，无须新建，当本项目实施过程中国家标准和省级标准涵盖不到，却又需要有统一标准时，再由市卫健委制定相关标准。

由于标准体系建设是一个庞大的工程，本项目主要遵循省级相关标准，由于医疗行业数据产生的多样性、数据类型的多样性，同时建议由市卫健委组织相关行业内专家牵头建设相关医疗数据质控标准规范即从数据产生的源头保证数据有一定的标准性、规范性，为本项目建设提供更有序、更标准、更规范的数据源，也为厦门市医疗行业标准体系为相关系统平台的数据治理、分析、应用提供夯实的数据基础。

具体涉及的质控标准规范包括但不限于下表所示：

表 5-8 建议建设质控标准规范表

序号	标准规范名称	说明
1	电子病历记录标准规范	包括病情描述、处方书写规范等
2	CT 成像标准	CT 系统的参数标定及成像标准
3	X 成像标准	X 线拍片机的成像标准
4	B 超成像标准	B 超成像参数的标准
5	核磁共振成像标准	核磁共振成像参数标准

4.3.2 信息资源规划和数据库建设方案

本项目基于厦门市卫健委已经采集的各医疗机构的历史诊疗数据进行统一的数据治理和服务，将在原有基础上建设健康医疗数据仓库和相应的主题仓库。

4.3.2.1 健康医疗数据仓库

4.3.2.1.1 数据仓库架构

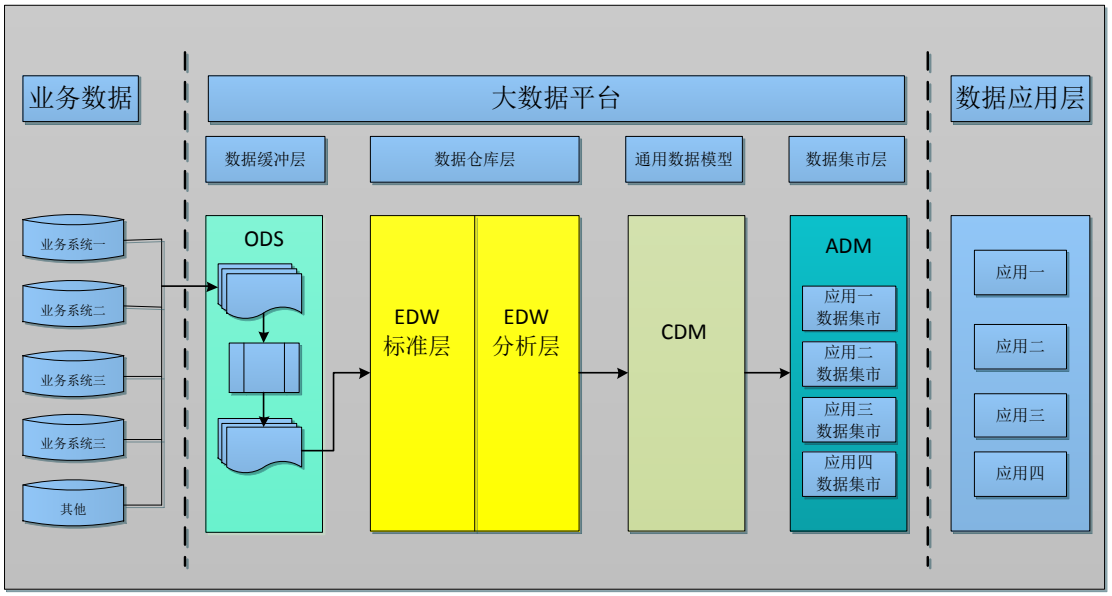


图 5-1 数据仓库架构图

层次结构说明：虚线内为数据中心分层模型

ODS 层:操作数据存储(OperationalDataStore)，保存业务系统的原始数据，基本按照原始业务系统的表结构建立。

EDW 标准层：该层以 ODS 为基础，是数据中心核心层数据模型之一，用于存放完整详细历史数据，基本按关系模型设计。其主要的目的为了统一数据格式、字段命名、数据字典，清洗数据，整合不同业务系统，提供统一的业务视图，对外提供一致性的数据服务，为整个数据中心提供标准的清洗完的整合过的最细粒度的基础数据。EDW 层以主题划分。

EDW 分析层：该层以 EDW 标准层为基础，设计目标是为 ADM/CDM 提供足够的灵活性和扩展性的基础，该层采用混合模型（关系模型和维度模型）设计。

该层数据基本为可分析的数据，粒度的提升可在该层实现。EDW 层以主题划分。

CDM：通用数据模型（也称为通用数据集市层或者通用指标层），该层以 EDW 分析层为基础，提供通用的指标模型，以维度模型为主。以业务理解为基础，不以主题划分。

ADM：数据集市层：该层的数据来自 EDW 基础层/EDW 分析层/CDM 层，采取维度建模设计，建立不同的数据子集，以满足特定部门、团队、用户或不同应用程序分析和报告需求。本层以具体的数据应用划分。

4.3.2.1.2 数据建仓工作

1、资源准备

资源准备包括开发平台的准备、在平台进行用户创建、组织管理、项目管理等资源分配准备工作，包括平台上新建用户注册 AK、增加组织成员、创建计算引擎、创建项目空间、成员角色授权、新增数据源等，注意项目空间使用开发+生产项目模板创建，数仓开发全部在开发项目中进行，资源命名规范参考数据仓库技术规范制定阶段输出的数据模型概论部分。

2、数据梳理准备

数据仓库开发的数据源头即数据接入层，因此源数据接入层数据的及时性以及准确性，决定数仓开发的数据质量；同时 DWD 层为清洗转换的中间层，对数据的易用性和准确性起到了至关重要的作用。

1） ODS 层数据表信息的梳理

数据表整体信息梳理：梳理具体的表名、表属于的主题域、表的来源系统、表的更新方式等具体详细信息，为正式开发做准备。

数据表结构梳理：数据表结构梳理是在数据表整理信息的梳理基础上，对表

的具体表结构信息进行梳理，主要包括表所属系统、表所属模块、表名、表中文名、表字段名、表字段注释、是否为主键、是否为日期字段、是否为码表字段等。

2） DWD 层数据表信息的梳理

数据表整体信息梳理：梳理 DWD 层具体表名、主题域、更新方式具体信息。

数据表结构梳理：数据表结构梳理是在 DWD 层数据表整体信息梳理后，对表的具体字段信息进行梳理，包括来源系统、子模块、主题域、ODS 层字段名、ODS 层字段类型、ODS 层字段注释、DWD 层字段名、DWD 层字段类型、DWD 层字段注释等。

3） DWS 层数据表信息的梳理

DWS 层数据表的信息梳理主要是主要操作包括基于业务整合、关联计算具体的业务表及关联表信息、贴标签的具体依据等信息。 DWS 层基础标签信息梳理：梳理 DWS 层需要打哪些基础标签。

4） ADS 层数据表信息梳理

ADS 层数据表梳理是指：结合具体业务及客户需求，收集所需的表、字段等信息。

5） DIM（维度表）数据表信息梳理

DIM（维度表）信息梳理，包括标准维度表信息、业务维度表信息、所属主题域，业务系统代码、业务系统名称、标准代码、标准名称等。

3、源数据接入层（ODS）建设

ODS 层的开发主要涉及到源数据的接入工作，由于改成对数据不做任何转换处理，所以 ODS 层的工作主要有数据抽取及数据的核查工作；在这之前我们需要结合准备阶段中 ODS 层数据表梳理的结果，来明确以下内容：

1）明确数据从哪些业务系统过来，了解各个业务系统的核心业务，明确各个业务系统中表和字段的意义。

2）了解表中各个字段的类型、取值范围、是否可以为空、格式，编码规范

等规则，并以此检查数据的存在哪些质量问题。

3）明确源系统的数据提供形式，数据库、文件、数据服务接口等，数量有多大，是否存在非结构化数据。

4）业务系统的数据是如何更新的，以及更新周期，更新记录时是否会带上时间戳。

整理 ODS 层的建表语句，统一将 ODS 层对应的表创建完成。

4、基础数据层（DWD）建设

按照 DWD 层规划，在对汇聚层数据去重、一致化、清洗转换，并进行轻度汇总、聚合数据处理，对相应普通表、分区表、拉链表、维表进行设计、开发、测试，完成基础数据层开发。

数据进入 ODS（源数据接入层）层后，会被清洗、整合、转换到 DWD（基础数据层）层，因为 DWD 层对数据的质量和一致性要求很高。按照这种情况，ODS 层作为后续流程的数据提供者，不对数据做任何的加工处理，而 DWD 层作为数据的需求者，需要根据对数据的使用目的进行相应的数据处理，然后存储到的表中，所以数据加工处理工作是在数据的需求方哪里完成的，即数据的清洗、转换和载入的任务是在这一层完成的。

ODS 层的数据经过清洗、转换、标准化等一系列操作后，需要将数据装在到 DWD 层对应的表中去,其中涉及到普通表的装载、动态分区表的装载等等。

5、基础标签层（DWS）建设

数据经过 DWD(基础数据层)层的清洗转换后，数据质量已经达到一定要求，DWS 层是以 DWD 层为主结合 ODS 层数据进行有统计汇总或算法产生的基础数据标签机按不同维度进行统计行政基础的宽表，主要操作包括基于业务整合、关联计算得到的明细数据；着力公共指标、排序聚合得到的汇总数据。

DWS 层的数据处理主要包括基于业务整合、关联计算得到的明细数据；基础标签的生成；数据的统计、排序等。

6、应用数据层（ADS）建设

ADS 层为应用数据层，主要是根据业务需求组织数据，该层支持百花齐放、尽可能都依赖 DWS，特殊情况可依赖 DWD 的数据。

ADS 层数据处理要结合具体的业务需求及客户要求，从 DWS 层表取数据（部分特需情况从 DWD 层表中取数据），编写相应的脚本。

7、公共维表（DIM）建设

DIM 表主要是为了进行码表转换作用的。其中包括标准表的建设及映射表的建设。

DIM 标准字典表数据处理：DIM 标准字典表是在原有的标准表的基础上，结合业务标准数据来不断补齐、扩充数据。具体原则是：标准里有的按标准来；标准里面完全没有的以业务标准为标准进行补充；标准里面部分有的，在原有基础上进行扩充（扩充时标准代码升位，即原来 1 位表示现在 2 位表示，原来 2 位表示 现在 3 位表示，依次类推）。

DIM 映射表数据处理：结合整理完成的“DIM 标准字典表”，一一对应相关业务表与标准表的映射关系。

8、工作流配置

工作流一般分初始化工作流和调度工作流。

初始化工作流是一次性从落地区把当天数据之前的数据全部进行处理抽取到主题区，它是一次性执行，因此属性中不需要配置。

调度工作流是在当天从落地区抽取前一天的增量数据进行抽取和处理到主题区对应的表中。由于要调度，因此需要在工作流属性中进行配置调度时间窗和依赖哪些工作流完成的前提下才能执行。

工作流属性主要分三类：

- 基本属性：在创建工作流的时候，该属性会自动创建
- 调度属性：

启动调度若是要每天自动调度执行，必须选择“启动调度”选项。

生效时间设置该工作流在哪个时间段内是有效的，可以自动调度执行。

调度周期设置工作流调度频率。可以按天、按周、按月、按小时、按分钟。

具体时间工作流在何时开始执行。

- 依赖属性：

所属项目 依赖的工作流所属的 project 名称

上游工作流 在选择完成project后，选择该project下需要依赖的工作流名称，可以选择多个依赖的工作流。

9、发布生产环境

将开发项目环境中开发好的数据表、工作流及其他资源发布到生产项目环境中，并配置调度任务，完成发布生产环境操作，完成建仓工作。

4.3.2.2 健康医疗主题数据库

本系统需根据医疗行为及综合服务监管的任务要求，按照医院实际业务需求或主题分析需求设计和建立相应的数据库，包括但不限于门诊信息库、住院信息库、医疗费用信息库、药品信息库、电子病历信息库、病案信息库等。平台底层应为标准化可扩展数据接口，后期可根据相应指标的增加逐步开通具体需求的扩展，自定义设计延伸其他信息库。健康医疗主题数据库数据范围如表 5-9 所示。

表 5-9 健康医疗主题数据库数据范围表

序号	数据范围
1	患者基本信息
2	患者院内基本信息
3	住院病案首页
4	就诊信息

5	转科历史
6	转床记录
7	科室就诊信息
8	诊断记录
9	医疗发票记录
10	医疗费用明细信息
11	中西医处方
12	药品类医嘱
13	草药处方
14	药品医嘱执行记录
15	门急诊发药
16	住院发药
17	化疗记录
18	非药品医嘱
19	检查申请单
20	检查记录
21	肺功能检查报告
22	检验记录
23	染色体核型分析报告
24	染色体 FISH 分析报告
25	基因检测报告

26	HLA 基因分型报告
27	骨髓穿刺记录
28	病理申请单
29	病理记录
30	手术申请预约
31	手术记录
32	麻醉前访视记录
33	麻醉术后访视记录
34	术后镇痛观察记录
35	手术麻醉主记录
36	手术用药记录
37	手术观察记录
38	手术事件记录
39	体外循环记录
40	输血申请单
41	输血观察记录
42	输血评价记录
43	输血不良反应上报记录
44	放疗记录
45	放疗记录 v2
46	评分记录

47	病情记录
48	出入量
49	护理观察
50	生命体征测量记录
51	护理记录
52	管路管理
53	入院评估记录
54	出院评估及指导记录
55	护理计划
56	护理不良事件报告
57	压疮护理报告
58	入院记录
59	24 小时内入出院记录
60	24 小时内入院死亡记录
61	首次病程记录
62	日常病程记录
63	上级查房记录
64	转科记录
65	抢救记录
66	会诊记录
67	有创诊疗操作记录

68	疑难危重病例讨论记录
69	交接班记录
70	阶段小结
71	术前小结
72	术前讨论
73	术后首次病程记录
74	死亡病例讨论记录
75	死亡记录
76	出院小结
77	出院记录
78	临床路径
79	检查类医嘱
80	检验类医嘱
81	处置类医嘱
82	输血类医嘱
83	手术类医嘱
84	门(急)诊病历
85	门急诊抢救记录
86	门(急)诊观察
87	院内感染登记
88	器械治疗院感监测

89	手术切口院感监测
90	随访方案
91	随访结果
92	胸外科手术申请预约
93	胸外科体外循环记录
94	胸外科手术麻醉主记录
95	胸外科手术方法详细记录
96	胸外科 ICU 记录
97	体检结果明细
98	体检检验
99	体检检查
100	体检总检结论
101	生物样本信息
102	病人维度统计信息

4.3.2.2.1 门诊信息库

门诊信息库主要包括患者门诊就诊的相关信息，包括患者就诊基本信息、挂号信息、分诊就诊信息等。

以下表格仅展示部分数据项。

表 5-10 门诊信息库部分数据项

序号	数据元名称
1	文档名称

2	医疗机构名称
3	医疗机构代码
4	患者标示
5	就诊标识
6	患者基本信息
7	医疗机构内患者标示
8	患者姓名
9	病案号
10	住院号
11	门诊号
12	性别名称
13	出生日期
14	婚姻状况名称
15	民族名称
16	户口地址
17	户口地址-省（自治区、直辖市）
18	户口地址-市（地区、州）
19	户口地址-县（区）
20	户口所在地邮编
21	职业名称
22	工作单位名称

23	工作单位的地址
24	工作单位及地址
25	工作单位地址-省（自治区、直辖市）
26	工作单位地址-市（地区、州）
27	工作单位地址-县（区）
28	工作单位邮编
29	工作单位电话
30	家庭电话
31	体重 kg
32	患者证件
33	证件类型代码
34	证件类型名称
35	证件号码
36	患者现地址信息
37	现住址
38	地址-省（自治区、直辖市）
39	地址-省（自治区、直辖市）编码
40	地址-市（地区、州）
41	地址-市（地区、州）编码
42	地址-县（区）
43	地址-县（区）编码

44	邮政编码
45	患者联系人信息
46	联系人与患者关系名称
47	联系人姓名
48	联系人电话号码
49	联系人完整地址
50	患者医疗保险信息
51	就诊卡号
52	医保卡号
53	卫生事件摘要信息
54	就诊类型
55	患者年龄
56	就诊次
57	挂号时间
58	号别
59	挂号医生 ID
60	挂号医生姓名
61	挂号科室 ID
62	挂号科室名称
63	分诊时间
64	分诊护士姓名

65	分诊护士 ID
66	初诊复诊标志
67	就诊医生 ID
68	就诊医生姓名
69	就诊状态
70	就诊时间
71	就诊科室 ID
72	就诊科室名称
73	就诊科室所属专业名称
74	医保险种类型名称

4.3.2.2.2 住院信息库

住院信息库主要包括患者住院就诊的相关信息，包括患者住院登记基本信息、病房床位信息、主管医师信息、医疗保险类型等。

以下表格仅展示部分数据项。

表 5-11 住院信息库部分数据项

序号	数据元名称	共享开放方式
1	文档名称	脱敏后开放
2	医疗机构名称	脱敏后开放
3	医疗机构代码	脱敏后开放
4	患者标示	脱敏后开放

5	就诊标识	脱敏后开放
6	患者基本信息	脱敏后开放
7	医疗机构内患者标示	脱敏后开放
8	患者姓名	脱敏后开放
9	病案号	脱敏后开放
10	住院号	脱敏后开放
11	门诊号	脱敏后开放
12	性别名称	脱敏后开放
13	出生日期	脱敏后开放
14	婚姻状况名称	脱敏后开放
15	民族名称	脱敏后开放
16	户口地址	脱敏后开放
17	户口地址-省（自治区、直辖市）	脱敏后开放
18	户口地址-市（地区、州）	脱敏后开放
19	户口地址-县（区）	脱敏后开放
20	户口所在地邮编	脱敏后开放
21	职业名称	脱敏后开放
22	工作单位名称	脱敏后开放
23	工作单位的地址	脱敏后开放
24	工作单位及地址	脱敏后开放
25	工作单位地址-省（自治区、直辖市）	脱敏后开放

26	工作单位地址-市（地区、州）	脱敏后开放
27	工作单位地址-县（区）	脱敏后开放
28	工作单位邮编	脱敏后开放
29	工作单位电话	脱敏后开放
30	家庭电话	脱敏后开放
31	体重 kg	脱敏后开放
32	患者证件	脱敏后开放
33	证件类型代码	脱敏后开放
34	证件类型名称	脱敏后开放
35	证件号码	脱敏后开放
36	患者现地址信息	脱敏后开放
37	现住址	脱敏后开放
38	地址-省（自治区、直辖市）	脱敏后开放
39	地址-省（自治区、直辖市）编码	脱敏后开放
40	地址-市（地区、州）	脱敏后开放
41	地址-市（地区、州）编码	脱敏后开放
42	地址-县（区）	脱敏后开放
43	地址-县（区）编码	脱敏后开放
44	邮政编码	脱敏后开放
45	患者联系人信息	脱敏后开放
46	联系人与患者关系名称	脱敏后开放

47	联系人姓名	脱敏后开放
48	联系人电话号码	脱敏后开放
49	联系人完整地址	脱敏后开放
50	患者医疗保险信息	脱敏后开放
51	就诊卡号	脱敏后开放
52	医保卡号	脱敏后开放
53	卫生事件摘要信息	脱敏后开放
54	就诊类型	脱敏后开放
55	患者年龄	脱敏后开放
56	入院时间	脱敏后开放
57	入科时间	脱敏后开放
58	入院科室名称	脱敏后开放
59	入院科室 ID	脱敏后开放
60	入院科室所属专业名称	脱敏后开放
61	入院病区名称	脱敏后开放
62	入院病区 ID	脱敏后开放
63	住院状态	脱敏后开放
64	主诊医师姓名	脱敏后开放
65	主诊医师 ID	脱敏后开放
66	主治医生 ID	脱敏后开放
67	主治医生姓名	脱敏后开放

68	出院时间	脱敏后开放
69	出院科室 ID	脱敏后开放
70	出院科室名称	脱敏后开放
71	出院科室所属专业	脱敏后开放
72	出院病区名称	脱敏后开放
73	出院病区 ID	脱敏后开放
74	医保险种类型名称	脱敏后开放

4.3.2.2.3 医疗费用信息库

医疗费用库主要包括患者费用结算的相关信息，包括患者门诊费用细目，住院费用细目，门诊费用结算信息，住院费用结算信息等。

以下表格仅展示部分数据项。

表 5-12 医疗费用信息库部分数据项

序号	数据元名称	共享开放方式
1	文档名称	脱敏后开放
2	文档类别	脱敏后开放
3	医疗机构代码	脱敏后开放
4	医疗机构名称	脱敏后开放
5	就诊标识	脱敏后开放
6	患者标示	脱敏后开放
7	就诊类型	脱敏后开放

8	账单项目详细	脱敏后开放
9	项目编号	脱敏后开放
10	收费项目编码	脱敏后开放
11	收费项目名称	脱敏后开放
12	药物编码	脱敏后开放
13	药物名称	脱敏后开放
14	收费状态	脱敏后开放
15	原始收费状态	脱敏后开放
16	收费单价金额	脱敏后开放
17	收费数量	脱敏后开放
18	收费时间	脱敏后开放
19	数量单位	脱敏后开放
20	收费类别名称	脱敏后开放
21	原始收费类别名称	脱敏后开放
22	原始收费类别编码	脱敏后开放
23	收费总金额	脱敏后开放
24	下收费的时间	脱敏后开放
25	下收费的人员标示	脱敏后开放
26	下收费的人员姓名	脱敏后开放
27	下收费的科室名称	脱敏后开放
28	下收费的科室 ID	脱敏后开放

29	下收费科室所属专业	脱敏后开放
30	医嘱标示	脱敏后开放
31	费用归属科室	脱敏后开放
32	费用归属科室 ID	脱敏后开放
33	医保等级	脱敏后开放
34	个人承担费用金额	脱敏后开放
35	医保范围外自费金额	脱敏后开放
36	医保范围内自负金额	脱敏后开放
37	医保支付金额	脱敏后开放
38	文档名称	脱敏后开放
39	医疗机构代码	脱敏后开放
40	医疗机构名称	脱敏后开放
41	就诊标识	脱敏后开放
42	患者标示	脱敏后开放
43	就诊类型	脱敏后开放
44	发票信息	脱敏后开放
45	发票唯一标识符	脱敏后开放
46	发票号	脱敏后开放
47	医保类型	脱敏后开放
48	医疗机构类型	脱敏后开放
49	合计金额	脱敏后开放

50	基金账户支付金额	脱敏后开放
51	个人账户支付金额	脱敏后开放
52	个人实际支付金额	脱敏后开放
53	发票状态	脱敏后开放
54	结算时间	脱敏后开放
55	费用明细分组	脱敏后开放
56	费用分组唯一标识符	脱敏后开放
57	收费分类	脱敏后开放
58	分类汇总金额	脱敏后开放
59	账单项目详细	脱敏后开放
60	项目编号	脱敏后开放
61	费用分组唯一标识符	脱敏后开放
62	收费项目名称	脱敏后开放
63	收费单价金额	脱敏后开放
64	收费数量	脱敏后开放
65	数量单位	脱敏后开放
66	收费总金额	脱敏后开放
67	医保等级	脱敏后开放

4.3.2.2.4 药品信息库

药品信息库主要包括医院各类药品的相关信息，包括药品名称、规格、剂型、

生产厂家等基本信息、药品医嘱记录等。

以下表格仅展示部分数据项。

表 5-13 药品信息库部分数据项

序号	数据元名称	共享开放方式
1	文档名称	脱敏后开放
2	医疗机构代码	脱敏后开放
3	医疗机构名称	脱敏后开放
4	就诊标识	脱敏后开放
5	患者标示	脱敏后开放
6	处方信息	脱敏后开放
7	处方编号	脱敏后开放
8	处方类别名称	脱敏后开放
9	处方费别	脱敏后开放
10	相应的医嘱详细信息	脱敏后开放
11	医嘱(组)标识	脱敏后开放
12	医嘱项目类型名称	脱敏后开放
13	医嘱数量（deprecated)	脱敏后开放
14	医嘱执行频率名称	脱敏后开放
15	医嘱执行频率中文描述	脱敏后开放
16	频率次数	脱敏后开放
17	频率周期	脱敏后开放

18	频率周期单位	脱敏后开放
19	医嘱开立者标示号	脱敏后开放
20	医嘱开立者姓名	脱敏后开放
21	医生抗菌药物处方权限	脱敏后开放
22	医嘱审核者标示号	脱敏后开放
23	医嘱审核者姓名	脱敏后开放
24	审核医生抗菌药物处方权限	脱敏后开放
25	医嘱开立日期时间	脱敏后开放
26	医嘱开立科室 ID	脱敏后开放
27	医嘱开立科室名称	脱敏后开放
28	医嘱开立科室所属专业	脱敏后开放
29	医嘱执行科室名称	脱敏后开放
30	医嘱状态	脱敏后开放
31	药品信息	脱敏后开放
32	医嘱标示	脱敏后开放
33	药物编码	脱敏后开放
34	医嘱名称	脱敏后开放
35	药物商品名	脱敏后开放
36	药物通用名称	脱敏后开放
37	药物类别名称	脱敏后开放
38	药物类型名称	脱敏后开放

39	药物使用-次剂量	脱敏后开放
40	药物使用-次剂量单位	脱敏后开放
41	药物使用-次剂型剂量	脱敏后开放
42	药物使用-次剂型剂量单位	脱敏后开放
43	药物使用-总用量	脱敏后开放
44	药物使用-总用量单位	脱敏后开放
45	用药途径名称	脱敏后开放
46	药物剂型名称	脱敏后开放
47	药物规格名称	脱敏后开放
48	处方期间	脱敏后开放
49	处方期间单位	脱敏后开放
50	药品包装数量	脱敏后开放
51	药品包装数量单位	脱敏后开放
52	抗菌药物标识	脱敏后开放
53	草药付数（deprecated)	脱敏后开放
54	生产厂商	脱敏后开放
55	制备方法（deprecated)	脱敏后开放
56	嘱托	脱敏后开放
57	药品批准文号	脱敏后开放
58	药品生产批号	脱敏后开放
59	中药配方（deprecated)	脱敏后开放

60	中药配方顺序号	脱敏后开放
61	中药药物代码	脱敏后开放
62	中药药物名称	脱敏后开放
63	中药药物通用名	脱敏后开放
64	药物剂量	脱敏后开放
65	剂量单位	脱敏后开放
66	煎药方法名称	脱敏后开放
67	生产厂商	脱敏后开放

4.3.2.2.5 电子病历信息库

电子病历信息库主要包括患者住院电子病历文书的相关信息，包括患者入院记录、患者一诉五史、入院检查结果等。

以下表格仅展示部分数据项。

表 5-14 电子病历信息库部分数据项

序号	数据元名称	共享开放方式
1	文档名称	脱敏后开放
2	文档类别	脱敏后开放
3	医嘱标示	脱敏后开放
4	医疗机构代码	脱敏后开放
5	医疗机构名称	脱敏后开放
6	就诊标识	脱敏后开放

7	患者标示	脱敏后开放
8	患者年龄	脱敏后开放
9	科室名称	脱敏后开放
10	病区	脱敏后开放
11	病房号	脱敏后开放
12	病床号	脱敏后开放
13	入院记录编号	脱敏后开放
14	记录/签发者	脱敏后开放
15	记录时间	脱敏后开放
16	记录者标示号	脱敏后开放
17	记录者姓名	脱敏后开放
18	记录者角色	脱敏后开放
19	活动记录状态	脱敏后开放
20	参与者	脱敏后开放
21	医护人员标示号	脱敏后开放
22	医护人员姓名	脱敏后开放
23	医护人员角色	脱敏后开放
24	顺位	脱敏后开放
25	参与的日期时间	脱敏后开放
26	文档标题	脱敏后开放
27	入院记录详细记录	脱敏后开放

28	文档提交时间	脱敏后开放
29	文档提交者	脱敏后开放
30	文档提交者 ID	脱敏后开放
31	入院记录	脱敏后开放
32	患者病史陈述者与患者关系	脱敏后开放
33	患者病史陈述者姓名	脱敏后开放
34	可靠程度	脱敏后开放
35	入院日期时间	脱敏后开放
36	主诉	脱敏后开放
37	体格检查-结果描述	脱敏后开放
38	现病史	脱敏后开放
39	家族史	脱敏后开放
40	既往疾病史	脱敏后开放
41	个人史	脱敏后开放
42	婚育史	脱敏后开放
43	月经史	脱敏后开放
44	生育史	脱敏后开放
45	哺乳史	脱敏后开放
46	过敏史	脱敏后开放
47	暴露史	脱敏后开放
48	母孕史	脱敏后开放

49	喂养史	脱敏后开放
50	发育史	脱敏后开放
51	中医四诊	脱敏后开放
52	发病节气	脱敏后开放
53	专科检查	脱敏后开放
54	辅助检查	脱敏后开放
55	辅助检查列表	脱敏后开放
56	检查类型名称	脱敏后开放
57	检查项目中文名称	脱敏后开放
58	检查结果	脱敏后开放
59	执行科室名称	脱敏后开放
60	检查机构	脱敏后开放
61	检查日期	脱敏后开放
62	入院诊断	脱敏后开放
63	诊断日期	脱敏后开放
64	诊断类型	脱敏后开放
65	诊断类型代码	脱敏后开放
66	诊断状态	脱敏后开放
67	诊断分类	脱敏后开放
68	诊断顺位（从属关系）	脱敏后开放
69	疾病名称	脱敏后开放

70	疾病代码	脱敏后开放
71	诊断医生	脱敏后开放
72	诊断医生标示	脱敏后开放
73	起病节气归属代码	脱敏后开放
74	望诊	脱敏后开放
75	闻诊	脱敏后开放
76	问诊	脱敏后开放
77	切诊	脱敏后开放
78	四诊摘要	脱敏后开放
79	辩证分析	脱敏后开放
80	治则治法	脱敏后开放
81	中医诊断	脱敏后开放
82	诊断标示	脱敏后开放
83	诊断日期	脱敏后开放
84	诊断类型	脱敏后开放
85	诊断状态	脱敏后开放
86	诊断顺位（从属关系）	脱敏后开放
87	病名称	脱敏后开放
88	病编码	脱敏后开放
89	证名称	脱敏后开放
90	证编码	脱敏后开放

91	是否主要诊断	脱敏后开放
92	备注	脱敏后开放
93	诊断医生	脱敏后开放
94	诊断医生标示	脱敏后开放

4.3.2.2.6 病案信息库

病案信息库主要包括患者住院病案首页的相关信息，包括患者住院基本信息、入出院信息、患者诊断及手术操作信息、离院方式、患者各类住院费用等。

以下表格仅展示部分数据项。

表 5-15 病案信息库部分数据项

序号	数据元名称	共享开放方式
1	文档名称	脱敏后开放
2	医疗机构代码	脱敏后开放
3	医疗机构名称	脱敏后开放
4	就诊标识	脱敏后开放
5	患者标示	脱敏后开放
6	活动记录状态	脱敏后开放
7	病案基本信息	脱敏后开放
8	医疗机构名称	脱敏后开放
9	医疗付费方式	脱敏后开放
10	居民健康卡号	脱敏后开放

11	住院次数	脱敏后开放
12	病案号	脱敏后开放
13	病理号	脱敏后开放
14	行政管理	脱敏后开放
15	死亡患者尸检标志	脱敏后开放
16	病案质量代码	脱敏后开放
17	质控医生	脱敏后开放
18	质控护士	脱敏后开放
19	质控日期	脱敏后开放
20	患者基本信息	脱敏后开放
21	姓名	脱敏后开放
22	性别名称	脱敏后开放
23	出生日期	脱敏后开放
24	年龄	脱敏后开放
25	新生儿出生体重(g)	脱敏后开放
26	新生儿入院体重(g)	脱敏后开放
27	身份证号	脱敏后开放
28	出生地	脱敏后开放
29	出生地-省（自治区、直辖市）	脱敏后开放
30	出生地-市（地区、州）	脱敏后开放
31	出生地-县（区）	脱敏后开放

32	籍贯	脱敏后开放
33	籍贯-省（自治区、直辖市）	脱敏后开放
34	籍贯-市（地区、州）	脱敏后开放
35	籍贯-县（区）	脱敏后开放
36	民族	脱敏后开放
37	婚姻状况	脱敏后开放
38	国籍名称	脱敏后开放
39	职业名称	脱敏后开放
40	患者电话号码	脱敏后开放
41	现住址	脱敏后开放
42	现住址-省（自治区、直辖市）	脱敏后开放
43	现住址-市（地区、州）	脱敏后开放
44	现住址-县（区）	脱敏后开放
45	现住址-省（自治区、直辖市）编码	脱敏后开放
46	现住址-市（地区、州）编码	脱敏后开放
47	现住址-县（区）编码	脱敏后开放
48	现住址邮编	脱敏后开放
49	户口地址	脱敏后开放
50	户口地址-省（自治区、直辖市）	脱敏后开放
51	户口地址-市（地区、州）	脱敏后开放

52	户口地址-县（区）	脱敏后开放
53	户口所在地邮编	脱敏后开放
54	工作单位	脱敏后开放
55	工作单位的地址	脱敏后开放
56	工作单位及地址	脱敏后开放
57	工作单位地址-省（自治区、直辖市）	脱敏后开放
58	工作单位地址-市（地区、州）	脱敏后开放
59	工作单位地址-县（区）	脱敏后开放
60	工作单位电话号码	脱敏后开放
61	工作单位邮编	脱敏后开放
62	联系人信息	脱敏后开放
63	联系人姓名	脱敏后开放
64	联系人与患者关系	脱敏后开放
65	联系人电话号码	脱敏后开放
66	联系人地址	脱敏后开放
67	入出转信息	脱敏后开放
68	入院途径	脱敏后开放
69	入院日期时间	脱敏后开放
70	出院日期时间	脱敏后开放
71	实际住院天数	脱敏后开放

72	入院科室	脱敏后开放
73	入院科室 ID	脱敏后开放
74	入院病房号	脱敏后开放
75	出院科室	脱敏后开放
76	出院科室 ID	脱敏后开放
77	出院病区名称	脱敏后开放
78	出院病区 ID	脱敏后开放
79	出院病房号	脱敏后开放
80	转科科别	脱敏后开放
81	门（急）诊诊断	脱敏后开放
82	诊断类型	脱敏后开放
83	诊断状态	脱敏后开放
84	诊断顺位（从属关系）	脱敏后开放
85	疾病名称	脱敏后开放
86	疾病代码	脱敏后开放
87	是否并发症	脱敏后开放
88	是否主要诊断	脱敏后开放
89	门（急）诊中医诊断	脱敏后开放
90	诊断类型	脱敏后开放
91	诊断状态	脱敏后开放
92	诊断顺位（从属关系）	脱敏后开放

93	病名称	脱敏后开放
94	病编码	脱敏后开放
95	证名称	脱敏后开放
96	证编码	脱敏后开放
97	是否主要诊断	脱敏后开放
98	中医诊疗	脱敏后开放
99	治疗类别	脱敏后开放
100	实施临床路径	脱敏后开放
101	使用医疗机构中药制剂	脱敏后开放
102	使用中医诊疗设备	脱敏后开放
103	使用中医诊疗技术	脱敏后开放
104	辨证施护	脱敏后开放
105	入院西医诊断	脱敏后开放
106	诊断类型	脱敏后开放
107	诊断状态	脱敏后开放
108	诊断顺位（从属关系）	脱敏后开放
109	疾病名称	脱敏后开放
110	疾病代码	脱敏后开放
111	入院中医诊断	脱敏后开放
112	诊断类型	脱敏后开放
113	诊断状态	脱敏后开放

114	诊断顺位（从属关系）	脱敏后开放
115	病名称	脱敏后开放
116	病编码	脱敏后开放
117	证名称	脱敏后开放
118	证编码	脱敏后开放
119	出院诊断	脱敏后开放
120	诊断类型	脱敏后开放
121	诊断状态	脱敏后开放
122	诊断顺位（从属关系）	脱敏后开放
123	疾病名称	脱敏后开放
124	疾病代码	脱敏后开放
125	是否并发症	脱敏后开放
126	是否主要诊断	脱敏后开放
127	入院病情	脱敏后开放
128	病情转归	脱敏后开放
129	出院中医诊断	脱敏后开放
130	诊断类型	脱敏后开放
131	诊断状态	脱敏后开放
132	诊断顺位（从属关系）	脱敏后开放
133	病名称	脱敏后开放
134	病编码	脱敏后开放

135	证名称	脱敏后开放
136	证编码	脱敏后开放
137	是否主要诊断	脱敏后开放
138	入院病情	脱敏后开放
139	损伤中毒	脱敏后开放
140	损伤中毒的外部原因	脱敏后开放
141	损伤中毒的外部原因疾病代码	脱敏后开放
142	颅脑损伤患者昏迷时间	脱敏后开放
143	入院前时长	脱敏后开放
144	入院后时长	脱敏后开放
145	病理诊断	脱敏后开放
146	诊断类型	脱敏后开放
147	诊断状态	脱敏后开放
148	辅助诊断分类	脱敏后开放
149	诊断顺位（从属关系）	脱敏后开放
150	疾病名称	脱敏后开放
151	疾病代码	脱敏后开放
152	过敏史	脱敏后开放
153	药物过敏标志	脱敏后开放
154	过敏药物	脱敏后开放
155	医疗事件参与者	脱敏后开放

156	主治医生姓名 ID	脱敏后开放
157	主治医生姓名	脱敏后开放
158	科主任姓名	脱敏后开放
159	科主任 ID	脱敏后开放
160	住院医生姓名	脱敏后开放
161	住院医生 ID	脱敏后开放
162	主任(副主任)医生姓名	脱敏后开放
163	主任(副主任)医生 ID	脱敏后开放
164	责任护士姓名	脱敏后开放
165	责任护士 ID	脱敏后开放
166	进修医生姓名	脱敏后开放
167	进修医生 ID	脱敏后开放
168	实习医生姓名	脱敏后开放
169	实习医生 ID	脱敏后开放
170	病案编码员姓名	脱敏后开放
171	病案编码员 ID	脱敏后开放
172	实验室检查	脱敏后开放
173	ABO 血型	脱敏后开放
174	Rh 血型	脱敏后开放
175	手术/麻醉	脱敏后开放
176	手术名称	脱敏后开放

177	手术操作代码	脱敏后开放
178	手术结束时间	脱敏后开放
179	手术级别	脱敏后开放
180	手术切口类别	脱敏后开放
181	手术切口愈合等级	脱敏后开放
182	手术者姓名	脱敏后开放
183	Ⅰ助姓名	脱敏后开放
184	Ⅱ助姓名	脱敏后开放
185	麻醉医师 ID	脱敏后开放
186	麻醉医师	脱敏后开放
187	麻醉方法	脱敏后开放
188	麻醉方法代码	脱敏后开放
189	离院方式	脱敏后开放
190	离院方式	脱敏后开放
191	拟接受医疗机构名称	脱敏后开放
192	治疗计划	脱敏后开放
193	出院 31 天内再住院	脱敏后开放
194	出院 31 天内再住院目的	脱敏后开放
195	费用	脱敏后开放
196	应收总费用	脱敏后开放
197	自付金额	脱敏后开放

198	一般医疗服务费	脱敏后开放
199	中医医疗服务费	脱敏后开放
200	中医辩证论治费	脱敏后开放
201	中医辨证论治会诊费	脱敏后开放
202	一般治疗操作费	脱敏后开放
203	护理费	脱敏后开放
204	其它服务费用	脱敏后开放
205	病理诊断费用	脱敏后开放
206	实验室诊断费	脱敏后开放
207	影像学诊断费	脱敏后开放
208	临床诊断费	脱敏后开放
209	非手术治疗项目费	脱敏后开放
210	临床物理治疗费	脱敏后开放
211	手术治疗费	脱敏后开放
212	麻醉费	脱敏后开放
213	手术费	脱敏后开放
214	康复费	脱敏后开放
215	中医治疗费	脱敏后开放
216	中医类（中医和民族医医疗服务）	脱敏后开放
217	中医诊断费	脱敏后开放

218	中医外治费	脱敏后开放
219	中医骨伤治疗费	脱敏后开放
220	针刺与灸法治疗费	脱敏后开放
221	中医推拿治疗费	脱敏后开放
222	中医肛肠治疗费	脱敏后开放
223	中医特殊治疗费	脱敏后开放
224	中医其他治疗费	脱敏后开放
225	中药特殊调配加工费	脱敏后开放
226	辨证施膳费	脱敏后开放
227	西药费	脱敏后开放
228	抗菌药物费用	脱敏后开放
229	中成药费	脱敏后开放
230	医疗机构中药制剂费	脱敏后开放
231	中草药费	脱敏后开放
232	血费	脱敏后开放
233	白蛋白类制品类	脱敏后开放
234	球蛋白类制品类	脱敏后开放
235	凝血因子类制品类	脱敏后开放
236	细胞因子类制品费	脱敏后开放
237	治疗一次性医用材料费	脱敏后开放
238	检查一次性医用材料费	脱敏后开放

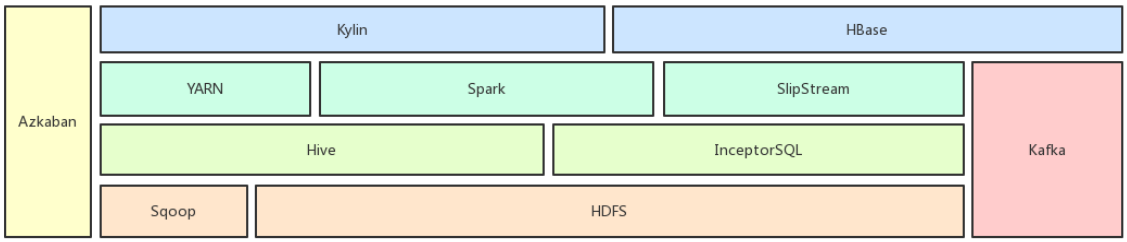
239	手术一次性医用材料费	脱敏后开放
240	其它费用	脱敏后开放
241	妇产信息	脱敏后开放
242	是否产后出血	脱敏后开放
243	术中出血量	脱敏后开放
244	术中出血量单位	脱敏后开放
245	产后出血量	脱敏后开放
246	产后出血量单位	脱敏后开放

4.3.3 应用支撑平台和应用系统建设及服务方案

4.3.3.1 应用支撑平台建设方案

4.3.3.1.1 Hadoop 大数据平台

Hadoop 是一个使用 java 编写的 Apache 开放源代码框架，它允许使用简单的编程模型跨大型计算机的大型数据集进行分布式处理。Hadoop 框架工作的应用程序可以在跨计算机群集提供分布式存储和计算的环境中工作。Hadoop 旨在从单一服务器扩展到数千台机器，每台机器都提供本地计算和存储。



如图所示：

数据储存于 HDFS 分布式文件系统上，使用 Sqoop 进行数据导入；

数据仓库使用 Hive 和 InceptorSQL；

计算引擎使用 YARN、Spark 和 SlipStream；

处理完毕的数据存放于 Kylin 和 HBase 中；

中间有 Kafka 进行数据传输，Azkaban 进行数据调度；

4.3.3.1.1.1 MapReduce

Hadoop MapReduce 是一个用于轻松编写应用程序的软件框架，它以可靠，容错的方式在大型集群（数千个节点）上处理大量数据并行处理商品硬件。

术语 MapReduce 实际上是指 Hadoop 程序执行的以下两个不同的任务：

Map Task：这是第一个任务，它接收输入数据并将其转换成一组数据，其中单个元素分解为元组（键/值对）。

Reduce Task：此任务将地图任务的输出作为输入，并将这些数据元组合并为较小的一组元组。reduce 任务总是在 map 任务之后执行。

通常输入和输出都存储在文件系统中。该框架负责调度任务，监视它们并重新执行失败的任务。

MapReduce 框架由单个主 JobTracker 和每个群集节点的一个从属 TaskTracker 组成。主管负责资源管理，跟踪资源消耗/可用性，并对从站上的作业组件任务进行调度，监控和重新执行故障任务。从站 TaskTracker 按照主机的指示执行任务，并定期向主设备提供任务状态信息。

JobTracker 是 Hadoop MapReduce 服务的单点故障，这意味着如果 JobTracker 关闭，则所有正在运行的作业都将停止。

4.3.3.1.1.2 HDFS

Hadoop 可以直接与任何可安装的分布式文件系统（如本地 FS，HFTP FS，S3 FS 等）工作，但 Hadoop 使用的最常见的文件系统是 Hadoop 分布式文件系统（HDFS）。

Hadoop 分布式文件系统（HDFS）基于 Google 文件系统（GFS），并提供一个分布式文件系统，旨在以可靠，容错的方式在大型计算机（数千台计算机）上运

行小型计算机。

HDFS 使用主/从架构，其中主机由管理文件系统元数据的单个 NameNode 和存储实际数据的一个或多个从属数据节点组成。

HDFS 命名空间中的文件被分成几个块，这些块被存储在一组 DataNodes 中。NameNode 确定块到 DataNodes 的映射。DataNodes 负责文件系统的读写操作。他们还根据 NameNode 给出的指令来处理块创建，删除和复制。

HDFS 提供了像任何其他文件系统一样的 shell，并且可以使用命令列表与文件系统进行交互。这些 shell 命令将在一个单独的章节中以及适当的示例进行介绍。

Hadoop 的优点：

Hadoop 框架允许用户快速编写和测试分布式系统。它是高效的，它自动分配数据并在机器上工作，反过来利用 CPU 核心的底层并行性。

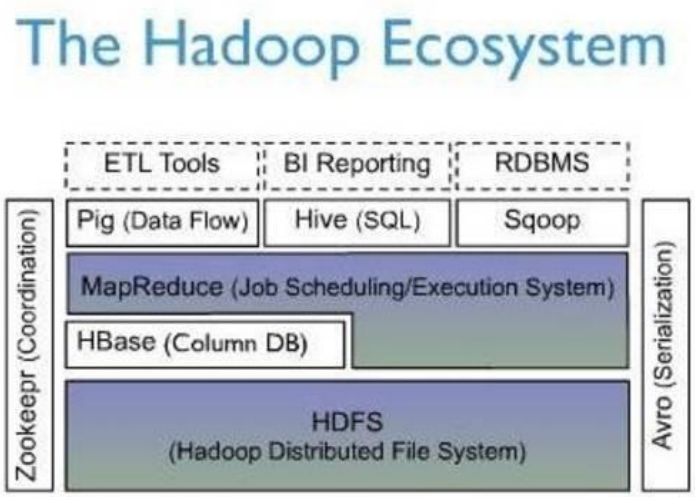
Hadoop 不依赖硬件提供容错和高可用性（FTHA），而是 Hadoop 库本身被设计为检测和处理应用层的故障。

服务器可以动态添加或从集群中删除，Hadoop 继续运行而不会中断。

Hadoop 的另一大优点是，除了是开放源码，它是所有平台兼容的，因为它是基于 Java 的。

4.3.3.1.1.3 HBase

HBase 是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的 Google 论文“Bigtable：一个结构化数据的分布式存储系统”。就像 Bigtable 利用了 Google 文件系统（File System）所提供的分布式数据存储一样，HBase 在 Hadoop 之上提供了类似于 Bigtable 的能力。HBase 是 Apache 的 Hadoop 项目的子项目。HBase 不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是 HBase 基于列的而不是基于行的模式。HBase 利用 Hadoop HDFS 作为其文件存储系统，利用 Hadoop MapReduce 来处理 HBase 中的海量数据，利用 Zookeeper 作为其分布式协同服务。主要用来存储非结构化和半结构化的松散数据（列存 NoSQL 数据库）。



Hadoop 生态系统

上图描述了 Hadoop 生态系统中的各层系统，其中 HBase 位于结构化存储层，Hadoop HDFS 为 HBase 提供了高可靠性的底层存储支持，Hadoop MapReduce 为 HBase 提供了高性能的计算能力，Zookeeper 为 HBase 提供了稳定服务和故障转移机制。

此外，Pig 和 Hive 还为 HBase 提供了高层语言支持，使得在 HBase 上进行数据统计处理变的非常简单。 Sqoop 则为 HBase 提供了方便的 RDBMS 数据导入功能，使得传统数据库数据向 HBase 中迁移变的非常方便。

HBase 的特性：

- 高性能

HBase 中存储了一套 HDFS 的索引，通过表名－>行健－>列族－>列限定符－>时间版本这一套索引来定位数据的位置，HBase 为每一列数据维护了一套索引规则，对于具体某一具体条数据的查询可以非常快速的通过 B＋树定位数据存储位置并将其取出。

另外，HBase 通常以集群部署，数据被分散到多个节点存储，当客户端发起查询请求的时候，集群里面多个节点并行执行查询操作，最后将不同节点的查询结果进行合并返回给客户端，提高 IO 性能。

- 高可用

HBase 集群中任意一个节点宕机都不会导致集群瘫痪。这取决于两方面原因：

第一方面，ZooKeeper 解决了 HBase 中心化问题；

另一方面，HBase 将数据存放在 HDFS 上面，HDFS 的数据冗余存放在不同节点，一个节点瘫痪可从其他节点取得数据，保证了 HBase 的高可用。

- 易扩展

Hbase 的扩展性主要体现在两个方面，一个是基于上层处理能 RegionServer 的扩展，一个是基于存储的扩展 HDFS。

- 无模式

使用 HBase 不需要预先定义表中有多少列，也不需要定义每一列存储的数据类型，HBase 在需要的时候可以动态增加列和指定存储数据类型。

4.3.3.1.1.4 Kafka

Kafka 是由 LinkedIn 开发的一个分布式的消息系统，使用 Scala 编写，它以可水平扩展和高吞吐率而被广泛使用。目前越来越多的开源分布式处理系统如 Cloudera、Apache Storm、Spark 都支持与 Kafka 集成。

Kafka 主要设计目标如下：

1、以时间复杂度为 O(1)的方式提供消息持久化能力，即使对 TB 级以上数据也能保证常数时间复杂度的访问性能

2、高吞吐率。即使在非常廉价的商用机器上也能做到单机支持每秒 100K 条以上消息的传输

3、支持 Kafka Server 间的消息分区，及分布式消费，同时保证每个 Partition 内的消息顺序传输

4、支持离线数据治理和实时数据治理

5、支持在线水平扩展

Kafka 是 Apache 下的一个子项目，是一个高性能跨语言分布式发布/订阅消息队列系统，而 Jafka 是在 Kafka 之上孵化而来的，即 Kafka 的一个升级版。具有以下特性：快速持久化，可以在 O(1)的系统开销下进行消息持久化；高吞吐，在一台普通的服务器上既可以达到 10W/s 的吞吐速率；完全的分布式系统，Broker、Producer、Consumer 都原生自动支持分布式，自动实现负载均衡；支持 Hadoop 数据并行加载，对于像 Hadoop 的一样的日志数据和离线分析系统，但又要求实时处理的限制，这是一个可行的解决方案。Kafka 通过 Hadoop 的并行

加载机制统一了在线和离线的消息处理。Apache Kafka 相对于 ActiveMQ 是一个非常轻量级的消息系统，除了性能非常好之外，还是一个工作良好的分布式系统。

4.3.3.1.1.5 Sqoop

Apache Sqoop（SQL-to-Hadoop） 项目旨在协助 RDBMS 与 Hadoop 之间进行高效的大数据交流。用户可以在 Sqoop 的帮助下，轻松地把关系型数据库的数据导入到 Hadoop 与其相关的系统 (如 HBase 和 Hive)中；同时也可以把数据从 Hadoop 系统里抽取并导出到关系型数据库里。除了这些主要的功能外，Sqoop 也提供了一些诸如查看数据库表等实用的小工具。

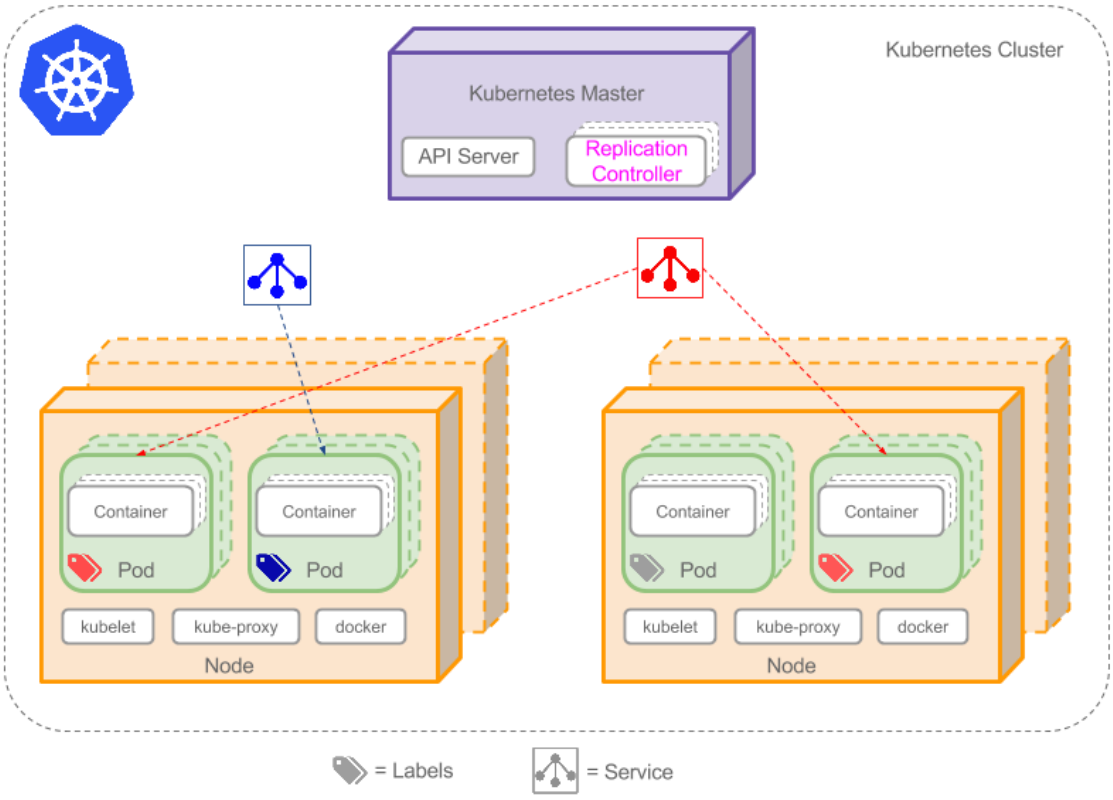
理论上，Sqoop 支持任何一款支持 JDBC 规范的数据库，如 DB2、MySQL 等。Sqoop 还能够将 DB2 数据库的数据导入到 HDFS 上，并保存为多种文件类型。常见的有定界文本类型，Avro 二进制类型以及 SequenceFiles 类型。

Sqoop 中一大亮点就是可以通过 Hadoop 的 MapReduce 把数据从关系型数据库中导入数据到 HDFS。Sqoop 架构非常简单，其整合了 Hive、Hbase 和 Oozie，通过 Map-Reduce 任务来传输数据，从而提供并发特性和容错。

4.3.3.1.2 k8s 微服务架构

Kubernetes 是一个全新的基于容器技术的分布式架构领先方案。Kubernetes(k8s)是Google开源的容器集群管理系统。在Docker技术的基础上，为容器化的应用提供部署运行、资源调度、服务发现和动态伸缩等一系列完整功能，提高了大规模容器集群管理的便捷性。

k8s 是一个完备的分布式系统支撑平台，具有完备的集群管理能力，多扩多层次的安全防护和准入机制、多租户应用支撑能力、透明的服务注册和发现机制、内建智能负载均衡器、强大的故障发现和自我修复能力、服务滚动升级和在线扩容能力、可扩展的资源自动调度机制以及多粒度的资源配额管理能力。同时 k8s 提供完善的管理工具，涵盖了包括开发、部署测试、运维监控在内的各个环节。其核心架构如下图所示：



1、Node

Node 作为集群中的工作节点，运行真正的应用程序，在 Node 上 k8s 管理的最小运行单元是 Pod。Node 上运行着 k8s 的 Kubelet、kube-proxy 服务进程，这些服务进程负责 Pod 的创建、启动、监控、重启、销毁、以及实现软件模式的负载均衡。

Node 包含的信息：

Node 地址：主机的 IP 地址，或 Node ID。

Node 的运行状态：Pending、Running、Terminated 三种状态。

Node Condition：…

Node 系统容量：描述 Node 可用的系统资源，包括 CPU、内存、最大可调度 Pod 数量等。

其他：内核版本号、k8s 版本等。

2、Pod

Pod 是 k8s 最基本的操作单元，包含一个或多个紧密相关的容器，一个 Pod

可以被一个容器化的环境看作应用层的“逻辑宿主机”；一个 Pod 中的多个容器应用通常是紧密耦合的，Pod 在 Node 上被创建、启动或者销毁；每个 Pod 里运行着一个特殊的被称之为 Pause 的容器，其他容器则为业务容器，这些业务容器共享 Pause 容器的网络栈和 Volume 挂载卷，因此他们之间通信和数据交换更为高效，在设计时我们可以充分利用这一特性将一组密切相关的服务进程放入同一个 Pod 中。

一个 Pod 里的容器之间仅需通过 localhost 就能互相通信。

一个 Pod 中的应用容器共享同一组资源：

PID 命名空间：Pod 中的不同应用程序可以看到其他应用程序的进程 ID；

网络命名空间：Pod 中的多个容器能够访问同一个 IP 和端口范围；

IPC 命名空间：Pod 中的多个容器能够使用 SystemV IPC 或 POSIX 消息队列进行通信；

UTS 命名空间：Pod 中的多个容器共享一个主机名；

Volumes（共享存储卷）：Pod 中的各个容器可以访问在 Pod 级别定义的 Volumes；

Pod 的生命周期通过 Replication Controller 来管理；通过模板进行定义，然后分配到一个 Node 上运行，在 Pod 所包含容器运行结束后，Pod 结束。

k8s 为 Pod 设计了一套独特的网络配置，包括：为每个 Pod 分配一个 IP 地址，使用 Pod 名作为容器间通信的主机名等。

3、Service

在 k8s 的世界里，虽然每个 Pod 都会被分配一个单独的 IP 地址，但这个 IP 地址会随着 Pod 的销毁而消失，这就引出一个问题：如果有一组 Pod 组成一个集群来提供服务，那么如何来访问它呢？Service！

一个 Service 可以看作一组提供相同服务的 Pod 的对外访问接口，Service 作用于哪些 Pod 是通过 Label Selector 来定义的。

拥有一个指定的名字（比如 my-mysql-server）；

拥有一个虚拟 IP（Cluster IP、Service IP 或 VIP）和端口号，销毁之前不会改变，只能内网访问；

能够提供某种远程服务能力；

被映射到了提供这种服务能力的一组容器应用上；

如果 Service 要提供外网服务，需指定公共 IP 和 NodePort，或外部负载均衡器；

4、Volume

Volume 是 Pod 中能够被多个容器访问的共享目录。

5、Label

Label 以 key/value 的形式附加到各种对象上，如 Pod、Service、RC、Node 等，以识别这些对象，管理关联关系等，如 Service 和 Pod 的关联关系。

6、RC（Replication Controller）

目标 Pod 的定义；

目标 Pod 需要运行的副本数量；

要监控的目标 Pod 标签（Lable）；

k8s 通过 RC 中定义的 Lable 筛选出对应的 Pod 实例，并实时监控其状态和数量，如果实例数量少于定义的副本数量（Replicas），则会根据 RC 中定义的 Pod 模板来创建一个新的 Pod，然后将此 Pod 调度到合适的 Node 上启动运行，直到 Pod 实例数量达到预定目标。

4.3.3.1.3 Elasticsearch 分布式搜索引擎

ElasticSearch 是一个分布式、高扩展、高实时的搜索与数据分析引擎。它能很方便的使大量数据具有搜索、分析和探索的能力。充分利用 ElasticSearch 的水平伸缩性，能使数据在生产环境变得更有价值。ElasticSearch 的实现原理主要分为以下几个步骤，首先用户将数据提交到 Elastic Search 数据库中，再通过分词控制器去将对应的语句分词，将其权重和分词结果一并存入数据，当用户搜索数据时候，再根据权重将结果排名，打分，再将返回结果呈现给用户。

Elasticsearch 可以用于搜索各种文档。它提供可扩展的搜索，具有接近实时的搜索，并支持多租户。”Elasticsearch 是分布式的，这意味着索引可以被分成分片，每个分片可以有 0 个或多个副本。每个节点托管一个或多个分片，并充当协

调器将操作委托给正确的分片。再平衡和路由是自动完成的。“相关数据通常存储在同一个索引中，该索引由一个或多个主分片和零个或多个复制分片组成。一旦创建了索引，就不能更改主分片的数量。

Elasticsearch 使用 Lucene，并试图通过 JSON 和 Java API 提供其所有特性。它支持 facetting 和 percolating，如果新文档与注册查询匹配，这对于通知非常有用。另一个特性称为“网关”，处理索引的长期持久性；例如，在服务器崩溃的情况下，可以从网关恢复索引。Elasticsearch 支持实时 GET 请求，适合作为 NoSQL 数据存储，但缺少分布式事务。

4.3.3.1.4 Kylin 分布式分析引擎

Apache Kylin™是一个开源的分布式分析引擎，提供 Hadoop/Spark 之上的 SQL 查询接口及多维分析（OLAP）能力以支持超大规模数据，最初由 eBay Inc. 开发并贡献至开源社区。它能在亚秒内查询巨大的 Hive 表。

Kylin 具有如下特性：

- 可扩展超快 OLAP 引擎：Kylin 是为减少在 Hadoop/Spark 上百亿规模数据查询延迟而设计
- Hadoop ANSI SQL 接口：Kylin 为 Hadoop 提供标准 SQL 支持大部分查询功能
- 交互式查询能力：通过 Kylin，用户可以与 Hadoop 数据进行亚秒级交互，在同样的数据集上提供比 Hive 更好的性能
- 多维立方体（MOLAP Cube）：用户能够在 Kylin 里为百亿以上数据集定义数据模型并构建立方体
- 与 BI 工具无缝整合：Kylin 提供与 BI 工具的整合能力，如 Tableau，PowerBI/Excel，MSTR，QlikSense，Hue 和 SuperSet
- Job 管理与监控
- 压缩与编码
- 增量更新
- 利用 HBase Coprocessor

- 基于 HyperLogLog 的 Dinstinc Count 近似算法
- 友好的 web 界面以管理，监控和使用立方体
- 项目及表级别的访问控制安全
- 支持 LDAP、SSO

Kylin 的生态圈包括：

Kylin 核心：Kylin OLAP 引擎基础框架，包括元数据（Metadata）引擎，查询引擎，Job 引擎及存储引擎等，同时包括 REST 服务器以响应客户端请求

扩展：支持额外功能和特性的插件

整合：与调度系统，ETL，监控等生命周期管理系统的整合

用户界面：在 Kylin 核心之上扩展的第三方用户界面

驱动：ODBC 和 JDBC 驱动以支持不同的工具和产品，比如 Tableau

4.3.3.1.5 MongoDB 分布式文件存储

MongoDB 是由 C++语言编写的，是一个基于分布式文件存储的开源数据库系统。在高负载的情况下，添加更多的节点，可以保证服务器性能。MongoDB 旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。MongoDB 将数据存储为一个文档，数据结构由键值(key=>value)对组成。MongoDB 文档类似于 JSON 对象。字段值可以包含其他文档，数组及文档数组。

MongoDB 主要特点：

MongoDB 是一个面向文档存储的数据库，操作起来比较简单和容易。

可以在 MongoDB 记录中设置任何属性的索引。

可以通过本地或者网络创建数据镜像，这使得 MongoDB 有更强的扩展性。

如果负载的增加（需要更多的存储空间和更强的处理能力），它可以分布在计算机网络中的其他节点上这就是所谓的分片。

MongoDB 支持丰富的查询表达式。查询指令使用 JSON 形式的标记，可轻易查询文档中内嵌的对象及数组。

MongoDB 使用 update()命令可以实现替换完成的文档（数据）或者一些指定的数据字段 。

MongoDB 中的 Map/reduce 主要是用来对数据进行批量处理和聚合操作。

Map 和 Reduce。Map 函数调用 emit(key，value)遍历集合中所有的记录，将 key 与 value 传给 Reduce 函数进行处理。

Map 函数和 Reduce 函数是使用 JavaScript 编写的，并可以通过 db.runCommand 或 mapreduce 命令来执行 MapReduce 操作。

GridFS 是 MongoDB 中的一个内置功能，可以用于存放大量小文件。

MongoDB 允许在服务端执行脚本，可以用 Javascript 编写某个函数，直接在服务端执行，也可以把函数的定义存储在服务端，下次直接调用即可。

MongoDB 支持各种编程语言:RUBY，PYTHON，JAVA，C++，PHP，C#等多种语言。

MongoDB 安装简单。

4.3.3.1.6 安全技术平台

4.3.3.1.6.1 PKI 体系

公钥基础设施（Public Key Infrastructure，简称 PKI）是目前网络安全建设的基础与核心，是电子商务安全实施的基本保障，因此，对 PKI 技术的研究和开发成为目前信息安全领域的热点。

PKI 采用证书进行公钥管理，通过第三方的可信任机构（认证中心，即 CA），把用户的公钥和用户的其他标识信息捆绑在一起，其中包括用户名和电子邮件地址等信息，以在 Internet 网上验证用户的身份。PKI 把公钥密码和对称密码结合起来，在 Internet 网上实现密钥的自动管理，保证网上数据的安全传输。

因此，从大的方面来说，所有提供公钥加密和数字签名服务的系统，都可归结为 PKI 系统的一部分，PKI 的主要目的是通过自动管理密钥和证书，为用户建立起一个安全的网络运行环境，使用户可以在多种应用环境下方便的使用加密和数字签名技术，从而保证网上数据的机密性、完整性、有效性。数据的机密性是指数据在传输过程中，不能被非授权者偷看；数据的完整性是指数据在传输过程中不能被非法篡改；数据的有效性是指数据不能被否认。

一个有效的 PKI 系统必须是安全的和透明的，用户在获得加密和数字签名服

务时，不需要详细地了解 PKI 的内部运作机制。在一个典型、完整和有效的 PKI 系统中，除证书的创建和发布，特别是证书的撤销，一个可用的 PKI 产品还必须提供相应的密钥管理服务，包括密钥的备份、恢复和更新等。没有一个好的密钥管理系统，将极大影响一个 PKI 系统的规模、可伸缩性和在协同网络中的运行成本。在一个企业中，PKI 系统必须有能力为一个用户管理多对密钥和证书；能够提供安全策略编辑和管理工具，如密钥周期和密钥用途等。

PKI 发展的一个重要方面就是标准化问题，它也是建立互操作性的基础。目前，PKI 标准化主要有两个方面：一是 RSA 公司的公钥加密标准 PKCS（Public Key Cryptography Standards），它定义了许多基本 PKI 部件，包括数字签名和证书请求格式等；二是由 Internet 工程任务组 IETF（Internet Engineering Task Force）和 PKI 工作组 PKIX（Public Key Infrastructure Working Group）所定义的一组具有互操作性的公钥基础设施协议。在今后很长的一段时间内，PKCS 和 PKIX 将会并存，大部分的 PKI 产品为保持兼容性，也将会对这两种标准进行支持。

PKI 的发展非常快，已经从几年前的理论阶段过渡到目前的产品阶段，并且出现了大量成熟技术、产品和解决方案，正逐步走向成熟。PKI 的发展受应用驱动的影响，比如，早期的 Internet 商务和 Web 安全要求主要依赖于 SSL，并要求应用首先对证书进行处理，所以，在很多公司的消息和群组产品中都提供了公钥和证书系统，如 Exchange 和 Notes 等。另外，基于标准的基础设施和应用也同样促进了 PKI 的发展，它能够保证基于 Internet 的安全消息传送的可交互性，如 S/MIME 等。

PKI 作为一组在分布式计算系统中利用公钥技术和 X.509 证书所提供的安全服务，企业或组织可利用相关产品建立安全域，并在其中发布密钥和证书。在安全域内，PKI 管理加密密钥和证书的发布，并提供诸如密钥管理（包括密钥更新，密钥恢复和密钥委托等）、证书管理（包括证书产生和撤销等）和策略管理等。PKI 产品也允许一个组织通过证书级别或直接交叉认证等方式来同其他安全域建立信任关系。这些服务和信任关系不能局限于独立的网络之内，而应建立在网络之间和 Internet 之上，为电子商务和网络通信提供安全保障，所以具有互操

作性的结构化和标准化技术成为PKI的核心。

PKI 的应用非常广泛，包括在 web 服务器和浏览器之间的通讯、电子邮件、电子数据交换（EDI）、在 Internet 上的信用卡交易和虚拟私有网（VPN）等。

一个简单的 PKI 系统包括证书机构 CA、注册机构 RA 和相应的 PKI 存储库。CA 用于签发并管理证书；RA 可作为 CA 的一部分，也可以独立，其功能包括个人身份审核、CRL 管理、密钥产生和密钥对备份等；PKI 存储库包括 LDAP 目录服务器和普通数据库，用于对用户申请、证书、密钥、CRL 和日志等信息进行存储和管理，并提供一定的查询功能。

4.3.3.2 健康医疗大数据治理服务方案

4.3.3.2.1 数据治理服务范围

厦门市健康医疗大数据治理服务主要面向市卫健委区域卫生信息平台已经完成数据采集的16家三级以上医疗机构HIS、EMR、LIS、PACS、体检、心电系统的诊疗数据及39家基层社区卫生服务中心的HIS及移动家庭签约系统的现有诊疗数据，以及这些医疗机构通过平台不断采集到的新增数据。

4.3.3.2.2 数据调研服务

4.3.3.2.2.1 调研规划

（1）调研内容

对需要治理的业务数据涉及到的数据范围、以及这些数据的来源系统、数据产生的业务部门等进行梳理，对这部分数据资源进行调研，内容包括：业务系统、表名、表中文名、更新方式、更新周期、是否共享、共享条件、表结构、字段编码对应表（字段取值说明）等。

（2）调研方法

根据以上确定的调研内容，可以通过如下调查方法进行调查工作：

从部门的业务梳理入手，即梳理各个业务处室的具体业务流程，以及流程涉及的数据需求，了解数据提供方、提供方式、采集方式等内容；

通过调研内容把业务部门承担的每一件业务事项、每一项工作任务都概括为各种要素，如业务名称、发生日期、参与人员、输入输出信息、涉及信息系统、相关类别等等，再利用这些要素整理成对应的资源信息目录。

（3）调查手段

对业务和信息资源的调查与梳理，采用统一的调研表，进行填报。使用该方法，业务人员在操作上相对方便。针对各个重点部门的需求调研表如附件。

4.3.3.2.2.2 调研计划

调研涉及范围广、调研工作繁重，针对这种情况，需要统一安排，分步进行，为此我们采用了摸排调研和详细调研、调研报告编制三个阶段完成对各单位的信息化系统建设状况及需求的调研。

（1）摸排调研阶段：

根据有关的标准，我们设计了一系列的基本需求调查表格，对各医疗机构信息资源及应用系统现状进行全面摸排，初步掌握全市医疗信息资源分布基本情况等。

（2）详细调研阶段：

根据摸排工作情况了解，对各医疗机构进行实地详细调研，本阶段重点调研内容：

调研并详细分析医疗机构拥有的信息资源种类、属性、用途、采集方式、提供者、共享需求等等。

调研并详细分析结构化信息资源的库、表、字段等信息及相关应用系统建设情况。

整理各种非结构化信息资源种类、数量，包括各种文字、图片、音视频。

（3）调研报告编制阶段:

汇总各医疗机构信息资源调研结果，绘制全市健康医疗信息资源总体分布图、深入分析共享资源的使用者、提供者及应用需求。

4.3.3.2.2.3 业务梳理

部门内数据资源的产生和使用来源于业务办理，为了全面掌握信息资源情况，有必要从业务入手进行信息资源调查。从业务入手，切入点是本单位业务职责，通过先理清业务职责，再理清信息资源，达到完成信息资源调查的目的。

从业务入手进行梳理和调查，要业务进行细分，根据业务类型调整调查表指标项的选择，最后完成业务信息调查表和资源信息调查表的填写。

梳理业务需要对业务进行细分，业务细分之后才能获得填写调查表的内容。业务调查表最难填的指标项是业务事项，而合适的业务事项名称是通过业务细分得到的，常用的业务细分方法有：按业务内容划分、按业务环节划分。

4.3.3.2.2.4 调研材料准备

包括发文文件、调研表格模板等。

（1）发文文件

通过发文，通知各家单位需要配合的工作内容，事先提供要准备的模板内容以及相应工作配合要求。

（2）调研表设计

根据调研内容、单位业务和信息特点着手研究设计调研表。

调研表可依据健康医疗数据资源的用途进行设计，调研表的内容要覆盖信息资源的采集机制、共享汇聚方式、更新周期等。

调研表可以分为业务调研表和信息资源调研表两种类型，可以两种表格配套使用。业务调研表要描述部门的业务基本情况，与业务相关的信息资源情况，以及它们之间的内在关系，以便从根本上把握信息资源。通过调研表可了解产生信息资源的业务事项、信息内容、信息责任单位、信息共享需求和信息使用等方面信息。信息资源调研表是对资源的具体描述和细化，信息资源调研表主要回答信息资源是什么，由什么组成、由谁管理、供谁使用、使用期限等问题。

4.3.3.2.2.5 基础及详细资源调研

调研包含基础调研和详细资源调研。

基础调研：由市卫健委牵头带领实施方与各个业务处室进行初步沟通，了解拥有的资源情况，建立联系机制，明确相应地接口人，沟通交流如何填写初步调研表等内容。初步调研主要是为了了解部门所拥有的资源的情况进行一个摸底工作，了解业务现状，为后面的详细调研做铺垫。

详细调研：是针对前面基本调研的结果，针对具体每个业务系统、每项资源情况进行资源的梳理，针对资源的字段、共享开放的情况进行整理确定。

调研表格参考如下，具体包括但不限于如下列举内容：

表 5-16 应用系统调查表

应用系统调查表						
填表人：			日期：			
序号	系统名称（必填）	使用范围（必填）	所属单位（必填）	网络环境（必填）	开发商（必填）	备注（选填）
示例 1						
1						

表 5-17 数据调研反馈表模板

xxx 数据调研反馈表							
序号	数据资源名称（必填项）填写数据资源名称	数据资源提供方		数据资源格式		更新周期	
		所属处室（选填项）	所属业务系统（必填项）	数据资源格式分类（必填项）下拉选择格式	数据资源格式类型（必填项）下拉选择数据仓库类型	更新周期频率（必填项）下拉选择更新	其他更新周期频率描述（选填项）

						周期	
示例 1	项目表	某处室	项目库系统	数据仓库	SQLServer	每月	
1							

表 5-18 数据公开申请表

序号	数据资源名称（必填项）填写数据资源名称	表名（数据仓库方式提供时必填）填写对应的数据仓库表名	数据项信息					共享属性		开放属性	
			数据项名称（必填项）对应数据仓库表中字段中文名称	代码（必填项）对应数据仓库表中字段代码	数据类型（长度）（必填项）对应数据仓库表中字段数据类型和长度	枚举值（必填项）	是否允许为空（必填项）	共享类型（必填项）	共享条件（必填项）	是否向社会开放（必填项）	开放条件（选填项）
示例 1	XXX 公开申请表	info_apply	申请 ID	apply_id	varchar(36)	无	否	无条件共享	无	是	无
			组织	organ	varcha	无	是	无条	无	是	无

			名称	_name	r(255)			件共享			
			组织机构代码	organ_code	varchar(31)	无	是	无条件共享	无	是	无
			营业执照信息	organ_licence	varchar(1023)	无	是	无条件共享	无	是	无
			法人代表	artificial_person	varchar(63)	无	是	无条件共享	无	是	无
			联系人姓名	contact_name	varchar(63)	无	是	无条件共享	无	是	无
			联系人电话	tel_	varchar(127)	无	是	无条件共享	无	是	无

4.3.3.2.2.6 调研结果梳理与编制

对各处室的调研结果进行梳理，对结果进行编制汇总。为下一步数据采集进行信息实时数据采集做准备。

4.3.3.2.3 数据统一、标准化、规范化服务

结合数据资源管理现状，基于数据处理系统，制定主数据、元数据、数据模型、数据标准等，完成健康医疗信息化大平台数据模型设计。通过数据处理，解决传统实现模式中由于数据标准与主数据、编码、数据质量相互隔离而导致的标准无法贯彻应用以及标准重复固化等问题；提供多种数据采集和数据接入方式，实现主数据管理和编码管理的完全服务化。

数据处理流程是一个闭环、不断优化的流程。整个数据处理的开发建立包括逻辑模型设计、物理模型设计、数据映射规则设计等过程，具体包括数据清洗、数据关联、数据映射、数据转换、数据归一、数据整合、数据索引、病历结构化数据处理、数据质控规则的制定。

需要用数据标准来指导模型设计，需要以元数据管理为核心，来准确把握数据加工的过程质量状态，以数据质量管理为保障来提供数据质量问题监控和改进的手段，同时，还可以通过主数据管理来提供核心的共享数据。

主数据是健康医疗大数据治理系统的核心数据，大多数业务环节和功能模块都在使用，其特点是持续增长但单条记录变动较少。

目前这些基础数据中，每一类数据大多都同时分布在多个系统当中，数据由各个信息系统独自维护和控制，这样就造成了系统之间基础数据的不完整、不一致，甚至不准确的情况。而且系统之间基础数据的同步机制也不完善，同步不及时、同步的过程无法监控，出现问题难以发现。

在主数据管理中，要从主数据组织、管理、共享几个方面着手。必须要建立一个集中存储的、统一管理的主数据管理系统，对分布在各个系统中的这些基础数据进行集中存储，同时建立系统之间这些基础数据的同步机制，保证各个系统之间的数据变化能被及时的跟踪和记录，保证这些基础数据在生成、传递、变更、存储、利用过程中的唯一性、完整性、准确性、及时性，形成一个统一的主数据管理平台，实现数据统一入口、统一校验、统一存储、统一分发的主数据管理模式。同时，主数据的准确性也可以极大地提高的与此相关的业务数据分析的准确性、可信性和一致性。

4.3.3.2.3.1 元数据管理

4.3.3.2.3.1.1 数据字典统一

根据国家、省医疗行业标准，制定数据字典，由于各医院系统厂商不一，业务系统繁多，其制定的数据规则也不相同，当采集各医院原始数据后，需对其进行处理，转换为本系统数据标准，为此，需建立数据字典。

目前，医疗行业一般涉及几十上百类文书，几千个字段，每个字段对应一个

字典。文书类型一般有：患者基本信息、患者院内基本信息、住院病案首页、就诊信息、转科历史、转床记录、科室就诊信息、诊断记录、医疗发票记录、医疗费用明细信息、中西医处方、药品类医嘱、草药处方、药品医嘱执行记录、门急诊发药、住院发药、化疗记录、非药品医嘱、检查申请单、检查记录等等。

每个文书里对应不同个数的字段，根据文书的复杂程度，如患者基本信息字段个数就会比较少，检查记录字段个数就会较多，如疾病名称（1 个字典）包括：霍乱，由于霍乱生物型所致、古典生物型霍乱、霍乱，由于埃尔托生物型所致、埃尔托生物型霍乱、霍乱轻型、霍乱中型、霍乱重型、霍乱暴发型、伤寒、伤寒杆菌败血症、伤寒并发脑膜炎、伤寒复发、伤寒并发肺炎、伤寒迁延型、伤寒逍遥型、伤寒并发腹膜炎、伤寒并发肠穿孔、伤寒并发肠出血、伤寒并发中毒性肝炎、伤寒并发支气管炎以及伤寒并发胆囊炎等几万个疾病名称。这仅仅是几千个字段中的一个，就包含了几万个名称，可想而知，数据字典的工作量是非常大的，但也是必要的。

4.3.3.2.3.1.2 数据模型设计

数据模型设计要遵循以下原则：继承性、稳定性、前瞻性和动态性。继承性指遵循已有的概念，并且在已有的主体基础上进一步细化，而不要贸然提出新的模型概念。稳定性指数据模型要对业务功能进行抽象概括，确保核心要求稳定。前瞻性是指数据模型要有一定的前瞻性，在设计上适当超前，给予业务发展一定的可扩容空间。动态性指物理模型和概念模型的对应关系，要及时更新及维护逻辑关系，保持数据模型的动态调整能力。

数据模型一般分为物理模型、概念模型和主题域模型 3 类。

物理模型（Extract-Transform-Load-Data-Record,ETLDR）是在计算机信息系统中依托特定实现工具的数据结构转换，是从关系型数据库中将数据转化为费关系型数据形式的一个模型，通过清洗、映射、脱敏、加密、后结构化等技术手段实现。

概念模型也叫通用数据模型，以数据实体及其之间的关系为基本构成单元的模型，是参考行业标准设计的通用数据模型。构建一个通用数据模型的目的是为

了规范数据的内容，方便数据的使用。通用数据模型会定义每个字段的数据格式、字段长度、值域、内容约束。在生成通用数据模型时，需要对原始数据进行清洗，将不符合要求的数据进行格式转换、计算、字典映射等。

主题域模型是节后应用场景来建立的个性化模型。在数据治理中，主题域数据模型包含有如下的类型：1、患者模型（patient profile，PP）对药品、检验、检查、手术、诊断等数据进行归一和结构化处理，结合患者所有信息按照患者维度方式进行聚合形成的模型；2、疾病模型（disease profile，DP）结合疾病树的构建、疾病数据标准化等进行聚合形成的模型；3、主体模型（disease tree，DT）面相统计分析运营等应用场景，包括指标计算、运营效率或疾病分布等，聚合相关数据形成的模型。

为保证数据标准的普适性，数据模型制定需参照 IHE、HL7、CDA、电子病历基本架构与数据标准等一系列国内外通用标准，通过 HL7 的 RIM 模型进行顶层设计，并落地细化医疗各类数据字段，确保医疗数据规划的标准可以覆盖医疗相关的信息。

同时，数据标准的设计应该涵盖医院主流的业务系统，包括但不限于：

1、居民健康档案数据

健康档案的基本内容主要由个人基本信息和主要卫生服务记录两部分组成。

1）个人基本信息

包括人口学和社会经济学等基础信息以及基本健康信息。其中一些基本信息反映了个人固有特征，贯穿整个生命过程，内容相对稳定、客观性强。主要有：

人口学信息：如姓名、性别、出生日期、出生地、国籍、民族、身份证件、文化程度、婚姻状况等。

社会经济学信息：如户籍性质、联系地址、联系方式、职业类别、工作单位等。

亲属信息：如子女数、父母亲姓名等。

社会保障信息：如医疗保险类别、医疗保险号码、残疾证号码等。

基本健康信息：如血型、过敏史、预防接种史、既往疾病史、家族遗传病史、

健康危险因素、残疾情况、亲属健康情况等。

建档信息：如建档日期、档案管理机构等。

2）主要卫生服务记录

健康档案与卫生服务活动的记录内容密切关联。主要卫生服务记录是从居民个人一生中所发生的重要卫生事件的详细记录中动态抽取的重要信息。按照业务领域划分，与健康档案相关的主要卫生服务记录有：

儿童保健：出生医学证明信息、新生儿疾病筛查信息、儿童健康体检信息、体弱儿童管理信息等。

妇女保健：婚前保健服务信息、妇女病普查信息、孕产期保健服务与高危管理信息、产前筛查与诊断信息、出生缺陷监测信息等。

疾病预防：预防接种信息、传染病报告信息、结核病防治信息、艾滋病防治信息、寄生虫病信息、职业病信息、伤害中毒信息、行为危险因素监测信息、死亡医学证明信息等。

疾病管理：高血压、糖尿病、肿瘤、重症精神疾病等病例管理信息，老年人健康管理信息等。

医疗服务：门诊诊疗信息、住院诊疗信息、住院病案首页信息、成人健康体检信息等。

2、临床数据信息

数据来源包括但不限于HIS、EMR（电子病历）、LIS、RIS、移动护理、手术麻醉、重症监护、病案系统、体检等。

HIS：患者（含门诊、住院）的基本信息、就诊情况、病历、诊断、医嘱、用药、耗材、手术、输血、检查、检验等信息；

EMR：门诊患者的门诊病历，住院患者的入院病历、病程、术前讨论、术后情况、出院小结、会诊记录等全部文书。按照卫健委的电子病历标准，共计17个大类共62个活动记录；

LIS：检查患者基本信息、身份信息、检查项目、检查细项、细项结果及正常值范围；

RIS：包括但不限于病理信息管理系统、超声信息管理系统、电生理信息管理系统（心电）、内镜信息管理系统、骨穿信息管理系统、肺功能信息管理系统、

核医学信息管理系统等系统的患者基本信息、身份信息、检查项目、检查方式、检查所见、检查结论等信息；

移动护理系统：包括但不限于体征采集、医嘱执行、护理记录、出入量记录、护理评估、健康宣教、护理计划、不良事件上报、压疮评估,交班报告、标本采集、输血记录等信息；

手术麻醉信息管理系统：包括但不限于手术发生前后的系统记录的术前访视、手术麻醉单、术后恢复，手术医嘱，术中护理，术中麻醉，手术申请，手术安排，等信息；

重症监护信息系统：包括但不限于病人信息的采集、存储、展现、包括重症监护记录单，管道管理，护理操作记录，评分记录，药品执行记录，给药记录，出入量记录，出入液体总量记录，转出转入记录等信息。

病案管理系统：包括但不限于编目后的病案首页信息等；

体检：患者基本信息、单位基本信息、体检项目清单、各项检查结果及正常值范围、各科室检查结论、终检结论等；

其他系统：单独科室进行的检查（如 DSA、24 小时动态心电）：检查患者基本信息、身份信息、检查报告及原始文件。

其他基于数据融合所需要的其它数据和信息等，比如社会保障保险信息、商业保险信息等。

3、健康医疗大数据中心基本数据集

基本数据集是指构成某个卫生事件(或活动)记录所必需的基本数据元集合。与健康医疗相关的每一个卫生服务活动（或干预措施）均对应一个基本数据集。基本数据集标准规定了数据集中所有数据元的唯一标识符、名称、定义、数据类型、取值范围、值域代码表等数据元标准，以及数据集名称、唯一标识符、发布方等元数据标准。

针对健康医疗的主要信息来源，目前已有健康医疗相关的卫生服务基本数据集标准共 32 个。按照业务领域（主题）分为 3 个一级类目：基本信息、公共卫生、医疗服务。其中“公共卫生”包含 4 个二级类目：儿童保健、妇女保健、疾病控制、疾病管理。

下表列出了健康医疗相关卫生服务基本数据集标准目录。如：《出生医学证

明基本数据集》的数据集标识符为“HRB01.01_V1.0”，表示该数据集标准属于“健康档案领域(HR)”中的一级类目“公共卫生(B)”下的二级类目“儿童保健(01)”，数据集顺序号为“01”，数据集版本号为“1.0”。

表 5-19 健康医疗相关卫生服务基本数据集标准目录

序号	一级类目	二级类目	数据集标准名称	数据集标识符
1	A 基本信息		个人信息基本数据集	HRA00.01_V1.0
2	B 公共卫生	01 儿童保健	出生医学证明基本数据集	HRB01.01_V1.0
3			新生儿疾病筛查基本数据集	HRB01.02_V1.0
4			儿童健康体检基本数据集	HRB01.03_V1.0
5			体弱儿童管理基本数据集	HRB01.04_V1.0
6		02 妇女保健	婚前保健服务基本数据集	HRB02.01_V1.0
7			妇女病普查基本数据集	HRB02.02_V1.0
8			计划生育技术服务基本数据集	HRB02.03_V1.0
9			孕产期保健服务与高危管理基本数据集	HRB02.04_V1.0
10			产前筛查与诊断基本数据集	HRB02.05_V1.0
11			出生缺陷监测基本数据集	HRB02.06_V1.0
12		03 疾病控制	预防接种基本数据集	HRB03.01_V1.0
13			传染病报告基本数据集	HRB03.02_V1.0

序号	一级类目	二级类目	数据集标准名称	数据集标识符
14			结核病防治基本数据集	HRB03.03_V1.0
15			艾滋病防治基本数据集	HRB03.04_V1.0
16			血吸虫病病人管理基本数据集	HRB03.05_V1.0
17			慢性丝虫病病人管理基本数据集	HRB03.06_V1.0
18			职业病报告基本数据集	HRB03.07_V1.0
19			职业性健康监护基本数据集	HRB03.08_V1.0
20			伤害监测报告基本数据集	HRB03.09_V1.0
21			中毒报告基本数据集	HRB03.10_V1.0
22			行为危险因素监测基本数据集	HRB03.11_V1.0
23			死亡医学证明基本数据集	HRB03.12_V1.0
24		04 疾病管理	高血压病例管理基本数据集	HRB04.01_V1.0
25			糖尿病病例管理基本数据集	HRB04.02_V1.0
26			肿瘤病例管理基本数据集	HRB04.03_V1.0
27			精神分裂症病例管理基本数据集	HRB04.04_V1.0
28			老年人健康管理基本数据集	HRB04.05_V1.0
29	C 医疗服		门诊诊疗基本数据集	HRC00.01_V1.0

序号	一级类目	二级类目	数据集标准名称	数据集标识符
30	务		住院诊疗基本数据集	HRC00.02_V1.0
31			住院病案首页基本数据集	HRC00.03_V1.0
32			成人健康体检基本数据集	HRC00.04_V1.0

4.3.3.2.3.1.3 代码字段标准化

1、标准化逻辑

根据实际数据，标记出代码字段，统一做代码转换处理。新增字段名为“${原字段名}dm”，填充原字段的代码值；原字段填充转换后的中文值。

2、处理要点

“证件类型、性别、民族、籍贯、行政区划、派出所”属于基础类字段，务必确保完成代码转换。

没有码表的字段，也需要新增“${原字段名}dm”字段，填“@@_代码值”。

所有行政区划的代码转换，一般很难统一成标准表，会记录行政区划的来源。

多值列字段，要借助 UDF 完成多值列的代码转换。

有性别和出生日期字段的，根据身份证号，提取性别和出生日期字段；没有身份证号的，根据码表转性别。

性别只根据码表转换，不根据身份证号提取，字段为空，保留不变。

标准代码表名：dm_bzdmys；尽量用小表广播 mapjoin 的方式进行代码标准化的 SQL 处理。

4.3.3.2.3.2 主数据管理

在主数据管理中，要从主数据组织、管理、共享几个方面着手。必须要建立一个集中存储的、统一管理的主数据管理系统，对分布在各个系统中的这些基础

数据进行集中存储，同时建立系统之间这些基础数据的同步机制，保证各个系统之间的数据变化能被及时的跟踪和记录，保证这些基础数据在生成、传递、变更、存储、利用过程中的唯一性、完整性、准确性、及时性，形成一个统一的主数据管理平台，实现数据统一入口、统一校验、统一存储、统一分发的主数据管理模式。同时，主数据的准确性也可以极大地提高的与此相关的业务数据分析的准确性、可信性和一致性。

4.3.3.2.4 数据汇聚服务

4.3.3.2.4.1 服务概述

本数据汇聚方案及标准依照国家卫健委已发布的各类标准规范，结合各业务系统制定的标准规范，并且基于对厦门市卫生信息化建设现状及未来建设方向中卫生信息及相关业务数据的理解，包括对已有数据产生方式的调研，详细阐述大数据治理服务的数据采集与集成的过程，就如何从现有区域卫生信息化平台进行数据采集服务，提出详细的解决方案。

标准化数据采集与集成系统区域卫生信息化平台已经采集到的现有数据按照本建设项目统一规定的数据接口标准，经过数据抽取、校验、清洗转换后传输到监管部门的数据存储中心；同时对流程管理、业务调度、异常数据处理、数据质控乃至数据提供机构/系统的系统改造和接口开发、采集过程中涉及的关键问题均需提出详细合理的解决方案。

本方案在设计同时需充分考虑后续监管业务应用的可得性和时效性，提供有针对性的不同的数据采集方式。例如针对在某些统计在当前实际卫生信息化建设中尚未建立对应的业务系统，需提供手工上报的方式，由相关机构部门按时按照标准上报，丰富系统数据的展现内容及数据统计分析需求；同时提供自动化数据采集接口，针对系统不同应用的及时性需求，分别提供实时和定时的数据化采集方案，例如：卫生监督业务实时性要求较高，相关数据要求实时采集；医院管理，绩效考核等业务实时性要求相对不高，理论上不超过 24 小时，相关数据可以要求定时采集方式进行，同时在实施阶段根据不同的业务数据采用不同的数据采集方式，可以降低因大量数据采集工作并发导致的网络和系统负载压力。

4.3.3.2.4.2 数据获取方式

常用的数据采集获取方式包括：

（1）数据仓库备份方式

部署前置数据仓库，数据采集通过前置库与数据源的备份库进行连接，数据仓库按照规定的时间策略，如按时、按天、按周等频率进行数据全量备份、增量备份，实现数据采集和更新。

（2）日志解析方式

部署前置数据仓库，基于数据仓库日志解析的数据备份复制方式，通过解析源数据仓库的在线日志或归档日志获得数据的变化，再将这些变化应用到前置机的备份库，从而实现医院业务数据仓库和前置机数据仓库之间的同步。

（3）数据 API 服务方式

通过卫健委给出的 API 接口标准实现在线数据采集和传输，接口采用统一的开发标准，独立于平台、独立于软件供应商的标准，实现跨部门、跨业务和跨平台的数据采集。

4.3.3.2.4.3 数据备份方式

基于厦门市各级各类医疗机构、区县卫生信息系统及其他业务系统中不同的数据仓库，本方案除需适配医疗行业常用的 Oracle、SQL Server、MySQL、InterSystems Cache、DB2、Sybase 等主流数据仓库以外，还应针对项目特点制定针对性的备份策略和管理规划，以保证数据备份的稳定运行，尤其是对于特殊的数据仓库，数据集成引擎同时支持 OGG 等商业化软件数据仓库同步的方式。

备份策略的选择，要统筹考虑需备份的总数据量、线路带宽、数据吞吐量、时间窗口以及对恢复时间的要求等因素。目前主流的备份策略主要有全量备份、增量备份 2 种。各个备份策略的特点及差异如下表所示：

	全量备份	增量备份

备份方法	备份所有文件	备份自从上一次备份后的全部改动和新文件
备份速度	最慢	最快
恢复速度	最快	最慢
空间要求	最多	最少

4.3.3.2.4.3.1 数据漂移

（1）数据仓库表中用来表示数据记录更新时间的时间戳字段（假设这类字段叫 modified_time）；

（2）数据仓库日志中用来表示数据记录更新时间的时间戳字段（假设这类字段叫 log_time）；

（3）数据仓库表中用于记录具体业务过程发生时间的时间戳字段（假设这类字段叫 proc_time）；

（4）标识数据记录被抽取到时间的时间戳字段（假设这类字段叫 extract_time）当然这种还是比较少用的；

理论上这几个时间应该是一致的，但是在实际生产中，这几个时间往往会出现差异，可能的原因有以下几点：

（1）由于数据抽取是需要时间的，extract_time 往往会万余前三个时间。当然这个应用的必要场景我所料及的不多；

（2）前台业务系统手工订正数据时候未更新 modified_time，在我们的系统中这种现象仍然没有杜绝；

（3）由于网络或者系统压力问题，log_time 或者 modified_time 会晚于 proc_time；

4.3.3.2.4.3.2 全量备份（Full Backup）

是对整个系统包括系统文件和应用数据进行的完全备份，这种备份方式的优点是数据恢复所需的时间短，缺点是备份数据总有大量的内容是重复的，这些重复的数据浪费了大量的文件空间，无形中增加了数据备份的成本；再者，由于需要备份的数据量相当大，因此备份所需的时间相对较长。

4.3.3.2.4.3.3 增量备份（Incremental Backup）

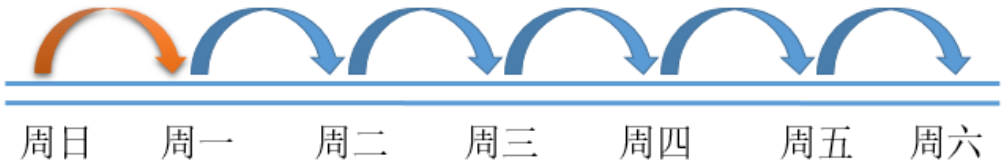
指每次备份的数据只是相当于上一次备份（全量、增量、差异）后增加的和修改过的数据。这种备份的优点很明显：没有重复的备份数据，节省文件空间，又缩短了备份时间。但是它的优点在于恢复数据比较麻烦，需进行多次数据恢复才能恢复至最新的数据状态。

4.3.3.2.4.4 数据备份策略

面对本项目庞大的数据量，做全量备份所需时间相当长，需按照项目现状制定多个基于海量数据的备份策略，项目实施技术人员和数据上报部门技术人员可结合现场实际情况进行选择，可选择的备份策略方法如下：

4.3.3.2.4.4.1 备份策略一（全量备份+增量备份/差异备份）

备份策略方法一（全量备份+增量备份）：如图所示



说明：

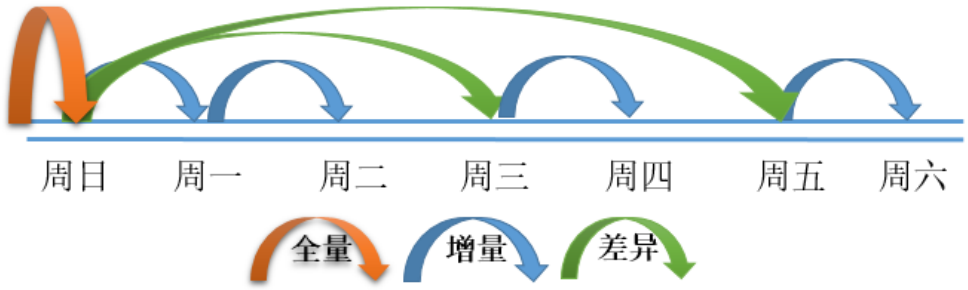
每周在访问量比较小（如：周六、周日）的时候做一次全量备份

每天对业务数据做一次差异备份或增量备份

每次业务数据做大幅度调整后应立即做一次全量备份

每天对备份服务器的 catalog 日志备份

4.3.3.2.4.4.2 备份策略方法二（全量备份+增量备份+差异备份）：



备份策略二

说明：

每周日进行全量备份

周一增量备份：周日到周一期间变更或者新增的数据

周二增量备份：周一到周二期间变更或者新增的数据

周三差异备份：周日到周三期间变更或者新增的数据（含周一和周二的增量备份）

周四增量备份：周三到周四期间变更或者新增的数据

周五差异备份：周日到周五期间变更或者新增的数据（含周三的差异备份和周四的增量备份）

周六增量备份：周五到周六期间变更或者新增的数据

4.3.3.2.4.5 数据抽取

为提高中心端数据恢复效率和减低恢复成本，本方案应综合考虑项目现状，在前置机端即对各级各类医疗机构、区县卫生信息系统及其他业务系统的传统数据上的数据向 hadoop 数据仓库进行迁移，这样在数据汇聚的时候，本系统只需要处理同一种数据仓库备份类型即可。

按照项目建设方案，在前置机上需安装与所在系统/系统相同的数据仓库管理软件，在完成定时数据采集过程中，系统处理流程如下：

系统将按照事先约定的备份策略完成数据备份，并将备份数据传送到前置机；

前置机数据集成引擎发现新增备份文件后，启动数据迁移流程；

前置机数据集成引擎将备份数据还原至备份文件相同的数据仓库管理系统中；

数据迁移模块将会把备份文件数据迁移至前置机的 Hadoop 数据中。

4.3.3.2.4.6 数据脱敏

根据项目建设需求，为了保护个人隐私，保护数据代表的对象利益不受侵害，本数据治理系统在前置机即需对数据进行脱敏处理。

数据脱敏模块可以根据不同的应用场景对数据进行脱敏处理，提供动态脱敏和静态脱敏两种技术手段：

动态脱敏：是指在数据调用时使用时进行脱敏；

静态脱敏：是指在数据使用前进行脱敏处理，可以在原始数据层、通用数据仓库层、主题库层等数据资源库上进行脱敏。

按照项目建设需求，本系统内所有的脱敏均不可逆，数据脱敏后不能进行反脱敏。通常需要脱敏的信息包括：患者信息、医务人员信息、医疗机构信息，本系统还可以根据项目建设过程中的新需求进行调整，常规脱敏内容如下：

患者：需要按 HIPPA 要求，对患者隐私信息进行脱敏处理，包括患者姓名、出生地、工作单位、工作单位及地址、工作电话、身份证号、家庭电话、现住址、户口地址、联系人姓名、联系人地址、联系人电话；

医务人员：医务人员姓名、医务人员 ID；

医疗机构：医疗机构名称、医疗机构代码。

本系统需按照建设需求需要对所有出现以上信息的地方进行脱敏，包括结构化、非结构化数据。

4.3.3.2.5 基于人工智能及自然语言处理技术的智能化数据治理服务

4.3.3.2.5.1 数据治理规范

厦门市健康医疗大数据治理服务应梳理建立数据治理相关标准规范，为数据治理全生命周期提供规范的数据处理标准,促进厦门市健康医疗大数据治理系统服务项目更加规范化、制度化，推动数据治理高效、有序的开展，确保项目运用取得实效。

（1）《数据采集管理流程规范》

定义明确的数据采集流程及数据采集方式，包括采集数据源方式、数据同步方式、数据同构传输方式、数据异构传输方式、采集数据脱敏方式、数据转换管理方式、采集作业管理方式、采集监控管理方式、采集系统配置、采集用户管理等内容。

（2）《数据治理流程规范》

定义明确的数据治理流程及数据治理方式，包括数据采集管理、数据目录管理、数据清洗、数据关联、数据映射、数据转换、数据归一、数据整合、数据索引、数据增强、服务协议转换、数据服务订阅等。

（3）《数据质量控制规范》

定义明确的数据质量控制标准及方式，数据信息采集管控、规则配置管理、流程调度控制、数据质量考核、数据质量评估、数据质量控制规则。

4.3.3.2.5.2 数据清洗服务

数据清洗是对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

数据清洗是将数据仓库精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。数据清理标准模型是将数据输入到数据清理处理器，通过一系列步骤“ 清理”数据，然后以期望的格式输出清理过的数据。数据清理从数据的准确性、完整性、一致性、唯一性、实时性、有效性几个方面来处理数据的丢失值、越界值、不一致代码、重复数据等问题。在进行数据初次清洗时需要人工针对现有采集到的 16 家三级以上医疗机构的数据制定数据清洗规则。

4.3.3.2.5.3 数据关联服务

数据关联是将恢复并清洗后的数据表中的数据，进行表与表之间的关联，还原原本数据表之间的关系。在进行数据初次关联时需要人工针对现有 16 家三级以上医疗机构已经采集的诊疗数据制定数据关联规则。

数据整合及数据融合的最后一步，需要人工针对现有 16 家三级以上医疗机构采集来的标准化后的数据进行数据融合，建立融合规则。在各级医疗机构的数据，通过数据转换成统一的数据模型及统一的数据结构之后，可以将不同医疗结构的数据整合在一起，形成最终融合汇聚的健康医疗大数据。融合后的数据，将作为以后大数据应用的数据基础。不同的应用，只需要在通用数据模型中，选取应用所需的字段，形成应用数据层，即可支持相对应的应用。

4.3.3.2.5.4 数据映射服务

数据映射（Data Mapping）即给定两个数据模型，在模型之间建立起数据元素的对应关系。数据映射是很多数据集成及数据融合任务的第一步。在进行数据映射时，需要人工针对现有 16 家三级以上医疗机构采集来的数据标准之间建立数据映射规则。

在完成数据映射之后，将数据从原有的数据模型加载到全新的数据模型的过程。数据转换分为两部分，一部分是数据结构的转换，另外一类是数据存储格式的转换。在进行数据转换时，需要人工针对现有 16 家三级以上医疗机构采集来的数据建立数据结构对应关系和数据储存格式。

4.3.3.2.5.5 数据归一服务

数据归一的过程即人工针对现有采集来的 16 家三级以上医疗机构的业务数据统一数据标准，在数据治理过程中建立对应关系。数据归一及对医疗数据实现主数据管理，也就是需要对数据进行标准化处理，主要参考国家标准和国家卫健委数据元标准，对目前医院自定的数据元，根据现有国家标准时，能够方便地进行修正。其他应用系统所使用的字典必须以数据中心标准字典为准，各个应用系统可以根据自身的需要选择性接收系统发出的元数据同步消息。

4.3.3.2.5.6 数据索引服务

数据主索引服务。包括数据主索引算法配置、唯一标识的产生、匹配和交叉引用管理、标识及基本信息的更新通知等。如：患者主索引基于身份证（社保卡）、姓名、性别、联系方式等进行算法匹配。建立患者主索引后，可以将患者将市内所有的信息进行关联。数据主索引在数据的使用过程中的不断优化，需要人工进行干预，人工干预主要指手动的拆分或者合并同一患者，或者患者主动提供不同医疗机构的病历 ID 信息申请关联，审核通过后，更新数据主索引。

4.3.3.2.5.7 自然语言处理处理服务

自然语言处理处理需要人工将全市医疗机构所有电子病历数据进行结构化规则编制，在规则经人工制定完毕之后，系统就可以按照规则进行自动化的病历结构化生产和处理。

结构化主要从若干个独立维度来进行，对数据依据主题字段进行划分，主要主题字段有：症状、体征、烟酒情况、病理诊断、病理表现、过敏情况、婚育状况等。根据病理或报告中不同字段的语义复杂程度和实际需求，目前结构化框架主要由正则抽取和通用框架组成。

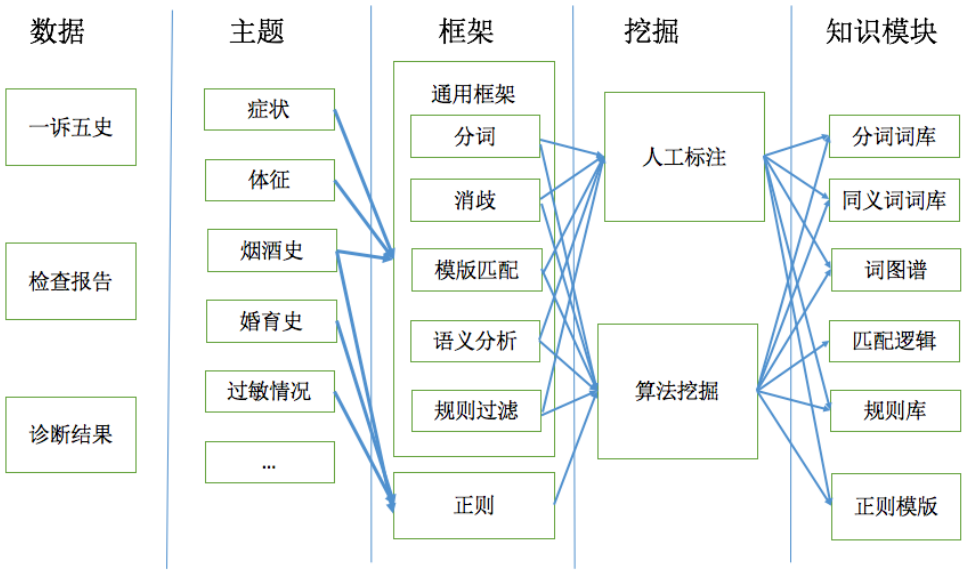


图 5-5 结构化整体流程示意图

正则抽取主要针对语义比较单一或规则性较强的字段，可以采用正则表达式

直接对文档中的关键信息进行抽取。例如,对于句子“月经情况:初潮 15 岁,5/26 天,末次月经 2016-2-14”,可通过正则表达式对日期的模式进行提取末次月经的时间,结果为{‘末次月经’:‘2016-2-14’}。

对于比较复杂的文本，例如患者症状，通用框架基于一些基本语法规则、医疗常识基础上,通过自然语言处理的技术方法来分析文本的隐含语义和上下文结构关系,主要包括:分词、消歧、模版匹配、语义分析、规则过滤等技术方法。

结构化主要针对病历中的大段文本内容，主要包括但不限于一诉五史、检查报告等，目前已经处理的主要包括:症状、体征、烟酒情况、病理诊断、病理表现、过敏情况、婚育状况等。相关字段如下：症状、吸烟史、吸烟量、吸烟现状、烟龄、饮酒史、饮酒类型、饮酒现状、日饮酒量、酒龄、生育状态、孕、产、流、育、是否绝经、初潮时间、经期、是否规律、是否痛经、是否有血凝块、是否颜色异常、是否绝经、是否停经、过敏史、过敏原名称等。

其他的关键字段，比如超声心动关注的三尖瓣反流压差及病理中的 MSH2、MSH6、高低卫星等都可以根据医疗机构的需求快速的进行结构化处理。

健康医疗数据中，存在大量的自然语言文本记录的数据，其中也包括大量有价值的信息，需要利用自然语言处理技术进行数据的增强。

自然语言处理子系统分为三个主要层次。分别是基础数据层，自然语言挖掘算法层，以及结构化系统层。在基础数据处理层主要通过整合权威医学标准，大规模专业词库以及真实临床医学词库构建基础词库。之后在挖掘算法层利用自然语言识别模型的训练进行实体识别，关键词提取，关系识别分类等工作。之后在结构化系统通过工具化人机协同工作针对临床病历数据进行疾病，手术，药品，症状，诊断，检查检验等数据项的结构化工作。这里涉及的所有自然语言处理子模块的算法均可以几种不同方式呈现。包括可调用 API 接口以及可人工参与的可视化工具。此外，针对此项目还需要文本分类、文本聚类、信息检索、信息推荐过滤等方面的基础技术积累，可供在项目需要时进行实验与技术支撑。

4.3.3.2.6 数据质控服务

数据质量评估是非实时性的定期或不定期对源或系统进行的数据质量评价，由数据质量管理人员根据需要发起，根据评估指标和评估方法，对数据的数据质量进行评价，得到评估结果以作为数据质量改进的参考和依据。

（1）评估指标

根据各指标对数据质量影响程度的不同，设置指标权重及评分规则。根据基础数据的特点，评估的指标主要从数据属性、模型关系和数据业务特征等方面进行设置，从而实现对数据的数据质量状况评价。评估指标包括但不限于以下方面的内容：

表 5-20 评估指标表

评估指标	计算方法	计算对象	维度归属	注释
空值率	空值数/值总数	单个字段	完整性	如：NULL 值等，用于检查具有业务含义字段的空值情况。
空字段率	空字段数/字段总数	全表	有效性	如：无值的字段，用于检查数据仓库表字段设计的利用率
数值重复率	重复数值数/数值总数	单个字段	重复性	采用全匹配方式统计数据重复比率，用于检查字段值的重复情况，如业务上不允许重复的字段
值域合法率	合法值数/有值总数	单个字段	合法性	如某字段的取值范围为枚举值（a,b,c,e,f）或值域范围(1-100)或例外值(3、8 除外)，则可统计合法数值比率
业务主键	业务主键	单个字段	合法性	业务上要求唯一的字段数

唯一性	唯一的数值数/数值总数			值比率评估报告

4.3.3.2.7 数据治理科研架构构建

在数据治理服务的基础上，构建集基础临床数据治理、专科数据治理为一体的医疗数据科研架构，建立基础临床数据库、专科科研数据库，打造疾病协同网络，提升科研效能。

（1） 基础临床数据库

建设包含门诊信息库、住院信息库、医疗费用信息库、药品信息库、电子病历信息库、病案信息库等全生命周期基础临床数据库。

（2） 专科科研数据库

研究基本常见疾病标准数据集，结合厦门市医学临床诊疗的专科优势及业务需求，构建 3 种专科疾病数据库。

4.3.3.3 健康医疗大数据知识应用系统建设方案

4.3.3.3.1 疾病探索

基于大数据挖掘及数据可视化技术，帮助临床医生从既往的真实病历数据中发现临床价值和研究价值。用户可以通过搜索本院的任意疾病的关键词，查看与该疾病主题相关的指标统计数据和数据关系图谱；搜索结果中可以查询本院相关的图谱及数据统计情况。

4.3.3.3.1.1 疾病图谱

疾病图谱可展现与查看的疾病主题关键词强相关的诊疗关键词，以及各个诊疗关键词互相关联的多层级关系网络。每个节点的大小与统计类型的权重及分布概率相关。本模块共统计五种类型的关键词，包括伴随诊断、用药、手术、检查、检验、症状。其中伴随诊断取自病历主诊断为当前主题词下的其他伴随诊断的 ICD10 标准名称。症状/体征取自病历主诉内容中结构化提取的症状/体征名。

任意连线连上的两个节点之间代表有数据关联，在节点上可以查看具体病例例数统计数据。

每个节点允许再次点击下钻查看该节点相关的数据关系。

4.3.3.3.1.2 疾病特征分布

疾病特征分布主要对选定的疾病主题词的某些指标特征以及诊疗业务相关数据进行分布统计，包括性别、年龄、伴随诊断、症状/体征、用药、手术、检验、检查。

4.3.3.3.1.3 疾病指标趋势

疾病指标趋势主要对选定的疾病主题词的某些指标特征以及诊疗业务相关数据进行趋势分析。支持选择按半年、一年及三年进行相关趋势的展现。主要包括门诊人次、住院人次、平均住院日、手术例数、用药 TOP5、手术 TOP5、检查 TOP5、检验 TOP5。

4.3.3.3.2 病历搜索

支持模糊搜索、多条件组合搜索等多种搜索方式，也同时支持多个复杂检索逻辑、条件树搜索、事件搜索的高级搜索方式，支持设置复杂的医学事件发生逻辑，搜索符合需求的患者集合，能够快速精确搜索符合特定要求的病历或患者，以满足临床各种查询、研究、分析场景的专业搜索需求。

4.3.3.3.2.1 关键词搜索

关键词搜索提供便捷的快速关键词搜索入口，通过医疗专业字典分析、切词等技术处理，检索出符合请求条件的病历结果。搜索关键字可以按照用药信息、诊疗信息、基本信息、检查、检验信息、出院记录等维度。

用户完成搜索操作后，系统可在后台迅速检索相关数据，将搜索结果展示到页面。系统需提供病历、患者 2 种视图模式。

为方便用户浏览，搜索结果展示页还在搜索结果的左侧提供了灵活的筛选

功能，可按照就诊类型（住院、门诊、急诊）、入院时间、入院科室、出院科室、就诊年龄、性别等维度进行结果筛选。

4.3.3.3.2.2 高级搜索

提供高级搜索功能，用于描述多个复杂的检索逻辑和条件，精确召回想查找的病历或患者；

高级搜索包括了逻辑关系/搜索主题/搜索条件/值域范围四个建立高级搜索条件的变量：

逻辑关系：包括 AND/NOT/OR 三种逻辑（默认值为 AND），另外所有的逻辑条件均是从下往上全部包含关系。

搜索主题：包括诊断/用药/检验/检查等诸多主题。

搜索条件：

【AND】 且，要求同时两个条件同时满足；

【OR】 或，要求被连接的多个条件满足任何一个或者同时满足均可

【NOT】 （【NOT】可以视为排除标准）

值域范围：限定的疾病或病症范围

患者维度：是指单病人的所有病历共同满足搜索条件，则该病人被召回。

就诊维度：指病人单份病历满足所有的搜索条件则该病人被召回，且符合所有搜索条件的病历可以被查看。

4.3.3.3.2.3 条件树搜索

用于描述具有复杂节点关系的检索逻辑和条件，精确召回想查找的病历或患者；

支持条件节点间进行多种逻辑关系组合；逻辑关系包括“且”、“或”；

支持纳入、排除组合搜索患者的方式；

支持 2 种筛选粒度：同病人、同病历。

4.3.3.3.2.4 事件搜索

支持条件树纳排的范围内进行二次事件搜索，通过加入相对时间或绝对时间、发生次数等条件，综合组成事件高级搜索方式。通过定义 T0 事件与事件前后的条件，精确搜索患者。

须支持多个事件关联搜索；可限定多个事件间的前后发生顺序，支持起始事件的发生次数的限定：首次、末次、任意次、第 2-9 次；

须支持 6 大类起始事件的事件类型的设置：诊断、手术、用药医嘱、检验、检查、就诊；

须支持入组条件的发生次数计算方式设置：累计、连续、任意。

4.3.3.3.3 搜索结果可视化

支持高级、条件树及事件搜索；支持结果显示表格化展示，且可自定义设置显示指标数据，与研究流程打通。

1、支持按照就诊类型、出入院科室、出入院时间等条件对搜索结果进行筛选。

2、能够为搜索结果病历展现病历摘要。

3、实现与“卫生健康科教大数据管理中心中心”的联动，搜索结果可以直接导入研究项目。

4、支持对搜索结果进行多维度统计分析，并用统计图形进行展现，支持明细展示。

5、支持对搜索历史信息的记录与搜索结果的收藏。

4.3.3.3.4 患者全景视图

基于临床疾病数据管理中心，系统需建立临床患者 360 视图，可展示个人

基本信息、入院记录、出院记录、病程记录、生命体征、病理、检验、检查、医嘱等。针对专科单一病种，系统需支持一键查看患者科研全景数据-患者详情；查看针对专病可定制的患者诊疗时间轴，直观了解患者基于此专病的特征性诊疗过程和结局以及患者的原始病历。

视图展示方式多样，可展示个人全景视图，展示视图维度包括：时间轴、就诊信息、检验、检查、病理、手术、文书记录、医嘱等，并可以筛选科室和就诊类型。

4.3.3.3.4.1 时间轴

时间轴是该患者在该院所有就诊信息的图形化总览，包括了所有就诊信息，诊断信息，用药信息，手术信息，检查信息，检验等信息，在该界面可以通过时间滚动条来调整数据的范围。

对较复杂的病情，较长治疗周期的患者数据，时间轴是很好的展示整个诊疗行为、病情变化的模式，同时将针对此患者、此疾病和诊疗的关键特征通过图形化可关联对比的模式给予呈现，方便医生快速获取有效信息和特征，无论对于病例分析学习、科研发现、查房看诊、治疗方案讨论都有非常大的价值。

以时间轴的方式展示患者的全治疗周期，记录患者在每一个时间节点的诊断、用药、体征数据、检查、检验、治疗、手术等数据，例如可以通过大量患者时间轴的堆叠，得出医院常规的诊疗路径，以及特定患者的个性化方案，结合着患者治疗效果的对比，可以作为知识库为患者提供更加科学的治疗方案。

4.3.3.3.4.2 就诊信息

就诊信息需按时间顺序排列该患者所有的门诊/住院/急诊的数据。

4.3.3.3.4.3 检验

检验需按照该患者的信息默认以检验单的样式显示患者所有就诊的检验信息，可以通过就诊的类型，检验的标本及检验的结果进行灵活的筛选，还可以通过检验指标名称搜索进行快速的定位。

4.3.3.3.4.4 检查

按照检查的类型，检查的部位需将所有的检查进行汇总显示，可以灵活方便的看到同类型检查，同部位在不同时间点的比较，还能将检查报告进行关键词的搜索。

4.3.3.3.4.5 病理

病理版块主要针对患者的病理报告做分类展示，内容主要需涵盖基本信息，肉眼所见，病理所见以及病理诊断等信息，可通过病理信息搜索进行快速查找。

4.3.3.3.4.6 手术

手术版块需全面记录患者的手术记录，包括手术基本信息、手术诊断、手术麻醉、术中用药以及手术过程描述等信息，还可以通过手术信息搜索进行快速的定位。

4.3.3.3.4.7 文书记录

文书记录需详细记录患者的入/出院的文书记录信息，包括但不限于病案首页、入院记录、出院记录、死亡记录、首次病程、查房记录以及会诊记录单等，可通过文书记录信息搜索进行快速定位。

4.3.3.3.4.8 医嘱

医嘱版块需详细记录患者住院期间的医嘱详细信息，包括医嘱时效、用药剂量、用药频率以及下医嘱科室等信息，支持按时效、科室以及频率快速筛选，也可通过输入医嘱信息进行关键词筛选。

4.3.3.3.5 患者分析

基于大数据加工基础，利用各种算法模型对本区域已积淀下的临床数据进行初步的分析透视。

1、支持从医生维度、科室维度及医院维度对患者诊疗数据进行分析。

2、支持医生以治疗过程中的角色（如：术者、主治医生）为视角分析数据。

3、支持从诊断、手术、检查、检验等业务领域为视角进行分析。

4、支持多种统计图形来展示分析结果统计。

4.3.3.3.6 研究项目管理

支持在线按照标准流程开展研究项目，包括以下步骤：设置项目基本信息、纳排条件设置、观测指标设置、项目结果导出等。支持以多种数据格式导出研究项目结果数据。常规的项目研究流程如下图所示：

支持在线按照标准流程开展研究项目，包括以下功能：

4.3.3.3.6.1 设置项目基本信息

如项目名称，项目起止时间等项目基本信息；

4.3.3.3.6.2 患者入组设置

支持系统纳排：入组方式以一组或者多组对研究对象设置纳入标准或排除标准；

同时支持人工导入功能：可直接通过输入姓名、病案号、住院号、门诊号、就诊号、病历号等信息人工导入患者；

4.3.3.3.6.3 观测指标设置

支持上千量级标准化指标选取、高级选取（按相对医疗事件基线时间制定化观测指标）选取以及通过结构化编辑器（根据标注信息机器学习来一键提取数值型、是否型等）等方式灵活提取所需的观测指标等功能；

4.3.3.3.6.4 结构化编辑器

支持通过文本提取技术与文本结构化技术，为医院从非结构化医疗文本数据中提取结构化数据，以提高用户提取医学结构化数据效率和能力，帮助用户快速获取结构化信息。

1、字段设置

字段类型分为三种：

是否型——提取是否包含某些内容，例如是否包含吸烟、是否包含头痛等

数值型——提取数值，例如提取肺结节大小，肠管多长。

特征型——提取部分文本，例如提取肿瘤发生的部位，如肺左叶，右叶等

2、信息提取-支持手工标注：

核心词输入框：输入精确描述提取结果的关键词，对提取条件做限定。

提取结果输入框：“数值型”和“特征型”字段均限制输入一个结果进行标注。是否型标识“是或否”

3、信息提取-智能提取：

标注完成后保存回到信息提取页，支持一键【提取当前页】和【提取所有页】进行提取；

4、提取完成页：提取完成后支持分来源统计，以及输出统计图表

4.3.3.3.6.5 研究对象详情

提供并支持两个入口视图查看患者信息：传统视图与多维视图

1、传统视图：

传统视图作为默认视图，可以看到患者入组途径、病案号、门诊号、性别以及相关病历数（通过搜索纳排被搜索命中的病历数）、所有病历数（患者所有就诊病历数）；同时还有该患者对应的诊断数、手术数、检验数、检查数、病历数、用药数

2、多维视图：

在多维视图中支持“基础信息表”+“观测指标表-就诊、用药、手术、检验、检查”等信息表展示，可将患者对应相关病历的入组途径、病案号、性别、就诊年龄、就诊科室、就诊类型、主诊断名称、就诊时间；另外在基础信息表中支持透出“通过纳排搜索添加的指标内容”，便于查询；

4.3.3.3.6.6 研究项目数据导出

当完成“观测指标设置”后，支持“导出预览”功能，在数据预览页面即可查阅到患者以及选择指标数据与展示格式。

支持以多种数据展示格式导出研究项目结果数据；将患者基本信息、诊断、用药、检验、检查、手术、病理等信息分为不同的 Sheet 展示；例如将一次就诊的所有数据平铺至一行方式；以患者+就诊+日期维度，一个结构化指标字段为一列对应展示数据；以及将相对医疗事件指标以二维表的格式等导出格式；

4.3.3.3.7 研究项目数据开放与共享

1、支持通过“我的研究项目”功能进行患者入组；

2、提交 API 申请：选择数据内容（根据所选择观测指标和全部数据层指标字段）、申请有效期；

3、支持后台管理员审批功能

4、支持生产以 Webservice 格式的 API 数据接口，可通过该接口获取病历数

据；

4.3.3.3.8 在线统计分析

支持对研究项目中设置的观察指标进行描述性统计分析；可将观测指标变量进行选择并添加统计分析栏中，点击“开始分析”即可自动化出分析结果，如分类型变量指标，其结果的形式展示为--表格（分类名称、总数、占比）和统计图表（饼图、条形图）。

系统需支持将该研究组中患者及就诊信息以及选择的观测指标数据，全部导入到统计分析模块中，可看到每个指标变量的描述性统计结果，并以表格与统计图的形式予以展示。

另外，指标变量为连续性数据，支持通过频数直方图或箱线图展示并将均值、标准差、中位数的统计结果在表格中展示，同时将正态分布检验结果予以展示。

4.3.3.3.9 重要疾病领域重点指标地图

支持重要疾病领域的医学核心观测指标的展示：基于医学分类领域，例如癌症领域、心血管领域、呼吸领域、内分泌领域中的 TOP 疾病，将该领域中研究场景下医学核心观测指标及对应的数据进行离线计算，并予以展示统计结果；如在“肿瘤领域”中，将展示：统计入院时间统计、诊断名称、患者性别、T 分期、M 分期、N 分期、肿瘤转移部位等统计图表；

4.3.3.3.10 知识全库

知识全库需主要包括的知识内容为中英文文献、指南共识、临床路径、药品说明书、临床试验、误诊误治等内容，上述信息支持检索与查看，部分知识可查看下载原文。此外根据用户特征信息进行智能推荐，并能够学习用户对于推荐内容的喜好程度进行深度学习，将医生更加需要的知识推荐给用户，包括由大数据智能技术对文献分析产生的文献研究热点趋势图和对应的文献作者图谱等。

4.3.3.3.10.1 知识推荐

知识库需设置有不同类别知识内容的推荐等整合功能，包括文献、指南共识、

临床路径、药品说明书、临床试验、误诊误治等。推荐功能：根据该医生/科室主要的诊疗疾病及其领域内对应的热点趋势、文献推荐，其推荐的内容源需涵盖万方和 PubMed 多个数据源，并且会对于推荐的内容打上推荐理由，可查看推荐的文献详情。

4.3.3.3.10.2 研究热点趋势图

研究热点趋势图需根据 PUBMED 文库中近十年相关疾病研究中各个不同重要 TOPIC 的河流图，并且每个 TOPIC 下的 keywords 能够进行点击后查看搜索后的具体文献情况。

热点趋势图的研究疾病，优先选择用户本身的主要诊疗疾病，其次选择用户所在科室的主要诊疗疾病，若用户本身和科室都没有治疗疾病，则选择医院的主要诊疗疾病作为显示内容，科研热点趋势图最多显示用户的前三个疾病对应的近十年文献研究热点，为用户提供热点走势，了解专业领域发展情况，为科研提供思路启发。

4.3.3.3.10.3 文献

文献搜索支持根据用户画像信息推荐文献，文献搜索的功能需包括中英文切换（默认为英文）、按照相关性或者时间排序、能够统计该搜索词的文献热度情况，部分搜索词还能够看到其相关搜索的词，并且能够通过点击相关搜索词进行二次搜索。

对于不同的搜索条件，提供相关患者与病历的功能，能够在医院中查询到符合条件的患者数量和病历数量。

此外，可以通过文献高级搜索进行精确查找。支持按主题、关键词、摘要、作者、作者单位、期刊、发表时间等多维条件进行检索。另外在按主题、关键词、摘要、作者单位搜索时，支持中英文互译来进行搜索。搜索结果同文献简单搜索功能。

4.3.3.3.10.4 指南共识

指南共识的推荐，和文献推荐一样，需根据医生/科室的主要诊疗疾病进行

指南的推荐，推荐的指南主要包括指南、共识、解读三部分内容，且有英文、中文、译文不同的类型。

搜索结果能够按照指南、共识、解读不同的类型进行分类切换，能够按照中文、英文、译文以及发布的年份进行指南的筛选。也可以按照相关性或者时间顺序进行排序的变换。

4.3.3.3.10.5 临床路径

临床路径页主要是临床路径搜索框和针对用户画像提供的临床路径推荐内容，并可下载对应的临床路径文档。

4.3.3.3.10.6 药品说明书

药物说明书默认展示其主要用药信息，用户可以通过其药品名称直接进入到药品说明书详情页。对于存在多种结果的用品名称，则显示与其相关的用药列表，可以查看具体的药品说明书。

4.3.3.3.10.7 临床试验

在知识全库的临床试验中，可查询到全球权威机构发布临床试验内容与信息（美国 ClinicalTrials 与中国临床试验注册中心 ChiCTR），供临床和科研景下使用；在进入临床试验 Tab 页中，可根据医生/科室的主要诊疗疾病与临床试验中研究疾病进行匹配推荐，可以在搜索框中对研究标题进行检索；同时在检索后可以进行筛选功能：可按来源 ClinicalTrials、ChiCTR 筛选；也可按试验类型：观察性试验、干预性试验等试验类型筛选；还可按招募状态：尚未招募、招募中、已完成等状态进行筛选；此外还可以按相关性排序或按发布时间排序；

4.3.3.3.10.8 误诊误治

误诊误治模块的数据来源要求官方提供，累计误诊病例不少于 2500 种，支持误诊误治疾病分类与推荐内容展示，此外同文献推荐功能根据医生/科室的主要诊疗疾病与误诊误治疾病进行匹配推荐，可查询该疾病对应的误诊误治详情。

4.3.4 数据治理和存储系统建设方案

本项目不包括计算机存储资源投入，需要另外的大数据计算平台提供医疗健康大数据的存储和计算服务，具备海量数据的处理能力，保障医疗数据归集和整合的落盘，保障数据能够快速，为医疗用户提供便捷的分析、处理海量数据的手段，从而达到简单高效地分析海量数据的目的，提供针对 TB/PB 级别数据的离线和实时处理能力。

大数据计算平台采用分布式架构，存储能力和计算能力可横向扩展。数据存储采用分布式文件系统，支持多种存储格式，具有高可靠性和高性能，支持列式存储，具有高压缩比。支持与常见数据源如 Database、Hadoop、HDFS 和文件进行数据交换，以及流式数据导入。基于统一的内存迭代计算架构，提供 SQL、Graph 计算、MapReduce、机器学习等多种数据治理接口和框架，具有高吞吐高性能的数据治理能力。提供 SQL 查询界面内置丰富的函数库，让数据运营人员可以关注数据治理业务逻辑本身。

大数据计算平台向用户提供数据集成、数据开发、算法开发等可视化的一站式开发平台，图形化的部署和到运维管理平台能够更快速的解决用户海量数据计算问题，有效降低医疗成本，并保障数据安全。

4.3.4.1 平台架构

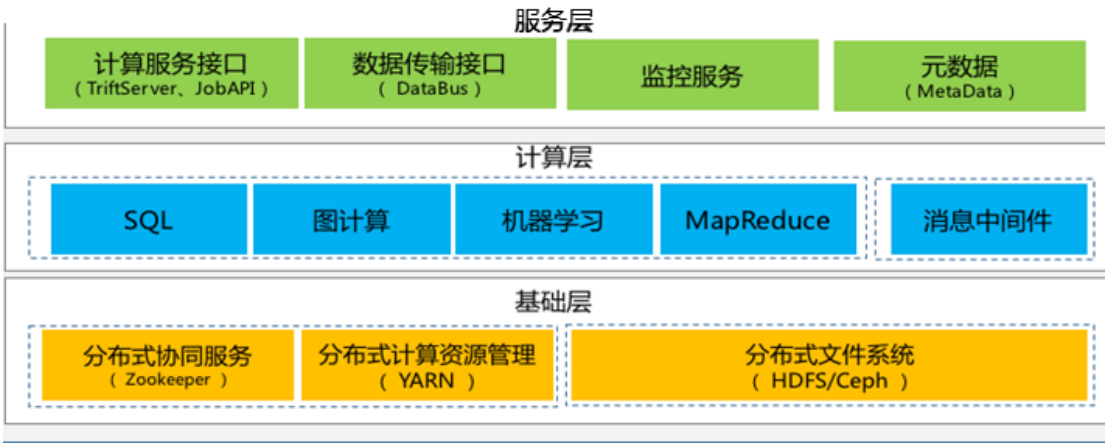


图 5-6 大数据计算平台架构图

基础层提供了计算资源、数据持久化服务。包括：

协同服务 Zookeeper，为集群中各种分布式应用如 HDFS、YARN 和 Kafka 等提供一致性服务。

计算资源管理 YARN，管理整个集群的计算资源，针对集群执行的各类计算作业进行管理和调度。

分布式文件系统，为平台提供了数据持久化存储，目前支持 HDFS 和 Ceph。其中 HDFS 是 Hadoop 生态主流的文件系统，适合大文件的顺序读写，Ceph 更擅长对小文件的访问，更适合有随机读写的场景。

计算层，提供了统一的基于内存迭代计算架构，支持 SparkSQL（包含 DataFrame、DataSet 编程接口，自定义函数功能和 HiveSQL 兼容），图计算 GraphX，机器学习 ML 和 MapReduce 等各种常用处理框架。

服务层：通过服务层向用户或者上层应用提供服务和 API，主要包括：

计算服务接口：ThriftServer 是基于 thrift 实现的通用 RPC 框架，对外提供服务，允许第三方应用通过 jdbc 等方式访问集群。Job API 可以让用户提交自定义计算作业。

监控服务：提供系统运维监控基础服务。数据传输接口 DataBus：DataBus

作为数据通道，允许用户通过编程接口导入流式数据。

元数据：MetaStore 作为元数据组件，提供元数据存取服务。

运维管理平台：运维监控服务，提供运维功能。

4.3.4.2 平台功能

4.3.4.2.1 数据存储

平台支持数据导入、数据传输和交换，以及流式数据实时导入，提供编程接口，支持流式数据实时导入，编程简单，高可靠，高性能。

分布式文件系统(HDFS)是适合运行在通用硬件上的分布式文件系统，是一个高度容错性的系统，适合部署在廉价的机器上，但能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。

HDFS 是一个主从结构，一个 HDFS 集群是由一个名字节点，它是一个管理文件命名空间和调节客户端访问文件的主服务器，当然还有一些数据节点，通常是一个节点一个机器，它来管理对应节点的存储。HDFS 对外开放文件命名空间并允许用户数据以文件形式存储。它具有以下特点：

主要高吞吐量访问：HDFS 的每个 block 分布在不同的 rack 上，在用户访问时，HDFS 会计算使用最近和访问量最小的服务器给用户提供。由于 block 在不同的 rack 上都有备份，所以不再是单数据访问，所以速度和效率是非常快的。

高容错性：HDFS 通过多方面保证数据的可靠性，多分复制并且分布到物理位置的不同服务器上，数据校验功能、后台的连续自检数据一致性功能，都为高容错提供了可能。

容量扩充：因为 HDFS 的 block 信息存放到 namenode 上，文件的 block 分布到 datanode 上，当扩充的时候，仅仅添加 datanode 数量，系统可以在不停止服务的情况下做扩充，不需要人工干预。

平台支持存储包括所有的关系型数据库、Hadoop 数据源及各种类型的文件，

数据传输安全可靠易用，支持数据回滚、断点续传、实时增量，性能卓越，并能通过水平扩展大幅提升数据存储和传输的性能。

4.3.4.2.2 数据计算

基于内存迭代计算架构，弹性分布式数据集是框架中的核心概念。可以将弹性分布式数据集视作数据库中的一张表，可以保存任何类型的数据。平台将数据存储在不同分区上的 RDD 之中。RDD 可以帮助重新安排计算并优化数据治理过程。它还具有容错性，因为 RDD 知道如何重新创建和重新计算数据集。RDD 是不可变的。你可以用变换（Transformation）修改 RDD，但是这个变换所返回的是一个全新的 RDD，而原有的 RDD 仍然保持不变。提供了 SQL（包含 DataFrame、DataSet 编程接口，自定义函数功能和 Hive SQL 兼容）、图计算 Graph、机器学习 ML、MapReduce 等各种常用处理框架。

4.3.4.2.2.1 SQL 计算

SQL 通过 JDBC API 将数据集对外提供出去，可以在数据上执行类似 SQL 的查询，同时还可以用 SQL 对不同格式的数据（如 JSON，Parquet 以及数据库等）执行 ETL，将其转化，然后提供给特定的查询。

用户可通过 SQL 语句的形式操作表，与传统 SQL 编程类似，语法高度兼容 HIVESQL，同时，提供 SQL 编程界面，让数据开发人员可以只需关注数据治理本身的业务逻辑。

平台支持 SQL 计算服务，提供了在线 IDE 供开发者编辑、调试 SQL 代码，而且还提供了强大的智能联想和高亮显示功能，方便开发者进行代码写作和阅读。

基于 DAG（有向无环图）执行模式和内存迭代计算架构，减少落盘环节；

支持数据治理流程的开发和调度；

算法开发，提供运营人员通过图形化的拖拽实现机器学习算法的开发；

基于统一的内存迭代计算架构，支持 SQL、MapReduce、Graph、机器学习等

并行计算框架，快速实现分布式并行计算；

提供简单易用内置丰富函数的 SQL 数据库，提供 SQL 编程界面和可拖拽的图形化算法开发界面，让数据运营人员可以更加关注数据的处理逻辑本身。

采用列式存储，数据压缩比高，更节省 IO；平台兼容 Hive，提供丰富的内置函数；

可适用 JDBC 接口访问，适用方便；支持 DataFrame 和 DataSet 编程接口，支持 Java、Python 和 Scala 等多种编程语言。

4.3.4.2.2.2 MapReduce 计算

MapReduce 是当前比较流行的分布式计算软件框架，常用于大规模数据的处理，具有线性可伸缩的特点。MapReduce 先把大数据集进行分治，即把大量的输入数据划分成较小的分片(通常称为 split)，并对每个分片输入分配一个 Map 任务进行处理，充分利用了计算机集群的并行计算性能；对所有 Map 任务过程处理得到的计算结果经过 shuffle(对各个 Map 任务的输出结果做排序、合并等操作并分发到各个 Reduce 任务的过程)交给相应的 Reduce 任务进行归约并得到最终的计算结果。

MapReduce 框架为用户屏蔽了数据的物理位置和分片实现等底层细节，用户只需要关注 Map 和 Reduce 的计算过程本身，编写自己的 Map 和 Reduce 程序，在作业提交页面配置作业参数并执行作业。

主要功能如下：

支持 DataFrame 和 DataSet 编程接口；

支持 Java、Python 和 Scala 等多种编程语言，提供 Java 版本的 MapReduce 编程接口供用户编写 Map 和 Reduce 程序；

用户可以在开发平台相应页面提交编写好的 MapReduce 程序 JAR 包进行执行并查看结果。

大数据计算平台提供了 Java 版本的 MapReduce 编程接口供用户编写 Map 和 Reduce 程序。用户可以在系统开发平台相应页面提交编写好的 MapReduce 程序 JAR 包进行执行并查看结果。

4.3.4.2.2.3 Graph 计算

Graph 是用于图计算和并行图计算，通过引入弹性分布式属性图（Resilient Distributed Property Graph），一种顶点和边都带有属性的有向多重图，扩展了 Spark RDD，提供了基础操作符集合和经过优化的 Pregel API 变体。此外，Graph 还包括一个持续增长的用于简化图分析任务的图算法和构建器集合。

图计算的特点是数据吞吐量本身相对不大，更看重迭代的效率。Graph 计算提供类似 Pregel 的 API，基于 RDD 数据模型进行迭代。

采用内存网络替代磁盘 IO，换来更快的性能

面向图数据而设计，适合图算法开发。

4.3.4.2.2.4 ML 机器学习计算

大数据计算平台提供常用的机器学习算法库，比 Hadoop 体系的 Mahout 速度更快（数量级的性能提升）。

提供 DataFrameAPI，更好的数据抽象，更丰富的数据类型，加快开发节奏；

支持多种语言如 Scala，Python，Scala；

支持 Pipeline，可以构建出复杂的机器学习工作流；

内存迭代计算，效率更高。

4.3.4.2.2.5 数据类型

支持整型和浮点型数据类型，支持布尔类型，也支持字符串和日期类型。

表 5-21 数据类型表

数据类型	长度
TINYINT	1 字节有符号整数
SMALLINT	2 字节有符号整数
INT	4 字节有符号整数
BIGINT	8 字节有符号整数
FLOAT	4 字节单精度浮点数
DOUBLE	8 字节双精度浮点数
DECIMAL	38 位精度浮点数
STRING	字符串
VARCHAR	变长字符串
TIMESTAMP	日期和时间，精度为微秒
DATE	日期类型，精度为天
BOOLEAN	布尔类型，true 或者 false
BINARY	字节序列

1）显示类型转换

显式类型转换是用 cast 将一种数据类型的值转换为另一种类型的值的行为。

表 5-22 显式类型转换表

From/To	整型	浮点型	STRING	VARCHAR	TIMESTAMP	DATE	BOOLEAN	BINARY
整型	-	Y	Y	Y	Y	Y	Y	N
浮点型	Y	Y	Y	Y	Y	Y	Y	N
STRING	Y	Y	-	Y	Y	Y	Y	Y
VARCHAR	Y	Y	Y	-	Y	Y	Y	Y
TIMESTAMP	Y	Y	Y	Y	-	Y	Y	N
DATE	N	N	Y	Y	Y	-	N	N
BOOLEAN	Y	Y	Y	Y	Y	N	-	N
BINARY	N	N	Y	Y	N	N	N	-

其中，'Y'表示可以转换，'N'表示不可以转换，'-'表示不需要转换。

2）STRING、TIMESTAMP 和 DATE 类型之间的转换

支持 STRING 类型、TIMESTAMP 类型和 DATE 类型之间的相互转换。

4.3.4.2.2.6 运算符

1、关系运算符

平台支持数据库关系运算符操作，满足常用 SQL 操作的需求，如下表所示：

表 5-23 关系运算符说明表

操作符	说明
A=B	如果 A 或 B 为 NULL，返回 NULL；如果 A 等于 B，返回 TRUE，否则返回 FALSE
A==B	同 A=B
A<=>B	如果 A、B 都不为 NULL，返回结果与 A=B 相同；如果 A、B 都为 NULL，返回 TRUE；如果 A、B 中一个为 NULL 一个不为 NULL，返回 FALSE
A<>B	如果 A 或 B 为 NULL，返回 NULL；如果 A 不等于 B，返回 TRUE，否则返回 FALSE
A!=B	同 A<>B
A<B	如果 A 或 B 为 NULL，返回 NULL；如果 A 小于 B，返回 TRUE，否则返回 FALSE
A<=B	如果 A 或 B 为 NULL，返回 NULL；如果 A 小于等于 B，返回 TRUE，否则返回 FALSE
A>B	如果 A 或 B 为 NULL，返回 NULL；如果 A 大于 B，返回 TRUE，否则返回 FALSE
A>=B	如果 A 或 B 为 NULL，返回 NULL；如果 A 大于等于 B，返回 TRUE，否则返回 FALSE
A [NOT] BETWEEN B	如果 A、B、C 三者之一为 NULL，返回 NULL；如果 A 大于等于 B 且小于等于 C，返回 TRUE，否则返回 FALSE。使用关键字 NOT 翻

AND C	转此语义。
A IS NULL	如果 A 为 NULL，返回 TRUE，否则返回 FALSE
A IS NOT NULL	如果 A 不为 NULL，返回 TRUE，否则返回 FALSE
A LIKE B	如果 A 或 B 为 NULL，返回 NULL，A 为字符串，B 为要匹配的模式，如果匹配，返回 TRUE，否则返回 FALSE。B 的表达式说明如下：'%'匹配任意多个字符，'_'匹配单个字符。要匹配'%'或'_'需要用转义符表示'\%'，'_'。例如： 'aaa' like 'a__' = TRUE 'aaa' like 'a%' = TRUE'aaa' like 'aab' = FALSE 'a%b' like 'a\%b' = TRUE 'axb' like 'a\%b' = FALSE
A RLIKE B	A 是字符串，B 是字符串常量正则表达式； 如果匹配成功，返回 TRUE，否则返回 FALSE； 如果 B 为空串会报错退出；如果 A 或 B 为 NULL，返回 NULL；
A IN B	B 是一个集合，如果 A 为 NULL，返回 NULL，如 A 在 B 中则返回 TRUE，否则返回 FALSE 若 B 仅有一个元素 NULL，即 A IN (NULL)，则返回 NULL。若 B 含有 NULL 元素，将 NULL 视为 B 集合中其他元素的类型。 B 必须是常数并且至少有一项，所有类型要一致

2、算法运算符

平台支持 SQL 操作的常用算法运算符，包括加减乘除等操作。如下表所示：

表 5-24 算法运算符说明表

操作符	说明

A + B	如果 A 或 B 为 NULL，返回 NULL；否则返回 A + B 的结果。
A - B	如果 A 或 B 为 NULL，返回 NULL；否则返回 A - B 的结果。
A * B	如果 A 或 B 为 NULL，返回 NULL；否则返回 A * B 的结果。
A / B	如果 A 或 B 为 NULL，返回 NULL；否则返回 A / B 的结果。
A % B	如果 A 或 B 为 NULL，返回 NULL；否则返回 A 模 B 的结果。
+A	仍然返回 A。
-A	如果 A 为 NULL，返回 NULL，否则返回-A。

3、逻辑操作符

平台支持业务逻辑操作符的计算，如下表所示：

表 5-25 逻辑操作符表

操作符	**说明**
A and B	TRUE and TRUE=TRUE TRUE and FALSE=FALSE FALSE and TRUE=FALSE FALSE and NULL=FALSE NULL and FALSE=FALSE TRUE and NULL=NULL NULL and TRUE=NULL

操作符	说明
	NULL and NULL=NULL
A or B	TRUE or TRUE=TRUE TRUE or FALSE=TRUE FALSE or TRUE=TRUE FALSE or NULL=NULL NULL or FALSE=NULL TRUE or NULL=TRUE NULL or TRUE=TRUE NULL or NULL=NULL
NOT A	如果 A 是 NULL，返回 NULL 如果 A 是 TRUE，返回 FALSE 如果 A 是 FALSE，返回 TRUE

4.3.4.2.2.7 视图操作

创建视图时，必须有对视图所引用表的读权限。

视图只能包含一个有效的 SELECT 语句。

视图可以引用其它视图，但不能引用自己，也不能循环引用。

不可以向视图写入数据，例如使用 INSERT INTO 或者 INSERT OVERWRITE 操作视图。

当视图建好以后，如果视图的引用表发生了变更，有可能导致视图无法访问，例如：删除被引用表。用户需要自己维护引用表及视图之间的对应关系。

视图可以包含 ORDERBY 和 LIMIT 从句，如果查询该视图的语句也包含这些从句，查询层级的从句将会在视图从句之后计算。例如视图指定了 LIMIT 5，相关查询语句为 select from v LIMIT 10，则最多 5 行记录被返回。

4.3.4.2.3 计算平台监控管理

大数据计算平台提供图形化监控管理功能模块，以对大数据计算平台的配置维护与运行状态监控预警为核心，主要实现自身软硬件环境运行状态的监控与预警，以及自身资源、作业、安全、项目等方面的运维管理功能。

包括监控大盘、集群管理、服务中心、资源管理 、作业管理、日志中心、告警中心、版本信息八大功能模块。

4.3.4.2.3.1 计算监控总览

运维大盘显示运维平台重要信息，包括集群服务器数量、集群 整体磁盘使用率、集群整体 CPU 使用率、主机概览、服务概览、自定义监控项。

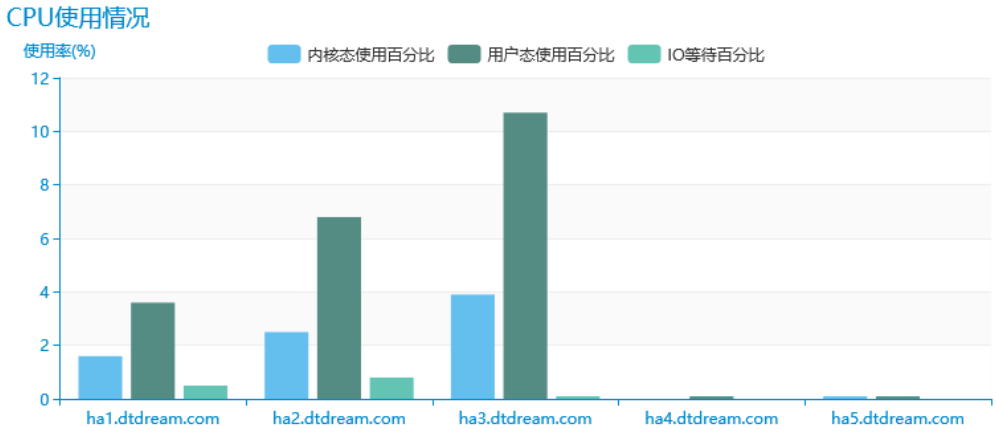
4.3.4.2.3.2 计算集群管理

集群管理主要包括主机监控、主机管理和集群时钟。

主机监控

主机监控显示各个主机的 CPU 使用率、磁盘使用率、内存使用率、负载情况，包括实时图和趋势图。

例如：



主机管理

主机管理显示各个主机的主机名、IP 地址、操作系统、CPU 核数、磁盘总量、已使用磁盘量、内存总量、平均负载、版本、组件等信息，实现了开启、关闭、重启组件以及主机删除功能。

集群时钟

大数据平台集群时钟采用 NTP 同步方式同步时钟，默认为本地时钟源，也支持添加外部时钟源，集群时钟板块主要分上下两个板块，上面的可以添加外部时钟源，下面的主要展示各个主机的时钟同步情况。需要注意的是，添加的外部时钟源时钟精度需要高于本集群，本集群才会向该时钟源同步。

4.3.4.2.3.3 计算服务监控

服务中心包括服务监控、服务管理、业务自检、冒烟测试等功能。

1、服务监控

服务监控实现对 HDFS、YARN 等核心服务的监控，可以查看到这些服务一些关键指标以了解服务的状况，比如 HDFS 空间使用率、损坏的数据块、YARN 的失败 app 数量、DATABUS 的流量情况、SPARK thrift 的内存空间使用情况等。

图 5-7 服务指标情况

2、服务管理

服务管理实现了用户对服务的管理，包括服务信息概览、修改服务配置、服务日志查看、服务的启动和关闭。

总体包括服务的概览、服务配置、日志查看和自恢复配置。服务概览就是显示该服务的组件状况，可以连接到相应主机；服务配置支持修改服务的配置文件，这里以版本的形式记录文件的配置修改情况，没修改一次会新生成一个版本，不支持历史版本的删除；日志查看支持查看近期的日志，若是需要更详细的日志，需要下载相应的打包日志；自恢复配置用来配置服务的自恢复设置，系统会根据服务的自恢复配置定期检测该服务，需要注意的是，并不是每个服务都支持服务的自恢复。

3、业务自检

业务自检主要检测健康状况，包括主机的 CPU、磁盘、内存等各项指标、服务的组件是否正常运行、服务组件进程的内存占用情况等。

4、冒烟测试

冒烟测试主要检测集群的一些功能是否正常。一般集群安装完成需要运行一次冒烟测试，若冒烟测试通过，则集群的基本功能就没有问题。

4.3.4.2.3.4 计算资源管理

采用分布式的资源管理系统，用以提高分布式的集群环境下的资源利用率，这些资源包括内存、IO、网络、磁盘等。支持多种计算框架，基于一个全局的资源管理器，所有接入的框架要先向该全局资源管理器申请资源。 提供资源隔离与动态调整。不同的框架中的不同任务往往需要的资源（内存，CPU，网络 IO 等）不同，它们运行在同一个集群中，会相互干扰，资源统一管控与调度平台提供一种资源隔离机制避免任务之间由资源争用导致效率下降。

资源管理主要包括分片管理和资源池管理。

1、分片管理

分片管理用于定义长服务运行的最小资源单位。

2、资源池管理

资源池管理实现了用户对多租户的管理，包含各个租户的资源配比以及长服务启动的配置。

4.3.4.2.3.5 计算作业管理

作业管理主要列出正在运行的作业，以及作业详情。

4.3.4.2.3.6 计算日志中心

日志中心主要包括服务日志、数据审计日志、操作审计日志。

服务日志模块提供大数据计算平台内部服务日志的打包和下载。

数据审计日志模块提供 HDFS 和 SparkSQL 的访问记录查看。

操作审计日志模块提供运维管理平台上所有操作的记录查看。

4.3.4.2.3.7 计算告警中心

1、主机告警

主机告警提供主机告警的阈值设置，主要包括 CPU、内存、磁盘、KAFKA 磁盘的阈值设置

2、服务告警

服务告警主要列出各个服务的告警设置。这里主要列出了 HDFS、YARN、ZOOKEEPER、DTSpark、DATABUS 这几个重要服务的告警设置。

3、告警通知设置

本系统支持邮件告警通知。邮件告警通知需要设置邮件服务器，

4、告警详情

可列出所有的告警内容，并按照严重、警告、未知、正常的告警级别从前往后排序。告警详情支持告警搜索。另外，上面的红色框框会显示有几个严重告警。

4.3.4.3 平台性能

平台采用分布式集群架构，集群规模可以根据需要灵活扩展，扩展过程业务不中断，最大单集群规模可扩展至 200 个节点。

所有数据三副本存储，保障数据高可靠性，可靠性高达 99.9999%。

无单点故障，保障服务高可用。

支持与已知的各种 Database 之间进行数据交换，支持流式数据导入，数据吞吐量可达 120MB/S。

采用分布式文件系统，支持多种存储格式，支持列式存储。

基于内存的迭代计算架构，支持 SQL、MapReduce、Graph、机器学习等并行计算框架。

内置丰富的SQL函数库，提供SQL编程界面和可拖拽的图形化算法开发界面，让数据运营人员可以只需关注数据治理本身的业务逻辑。

4.3.5 网络系统建设方案

本项目统一部署在卫健委医疗云，网络部分按照卫健委信息中心的统一规划进行实施，不再单独规划和建设项目外部网络系统。

4.3.6 安全系统建设方案

为有效推进大数据融合分析平台项目的信息安全工作，全面落实信息安全等级保护政策要求，成立健康医疗部门信息安全工作组，作为信息安全工作的执行单位，负责信息安全相关工作的规划、决策与监督，以及信息安全工作任务的全面落实。

4.3.6.1 安全服务

4.3.6.1.1 主要安全设计

在该建设项目中，安全依据《信息系统等级保护安全设计技术要求》（GB/T 25070-2010）三级防护要求，以及《国务院关于印发政务信息资源共享管理暂行办法的通知》等要求进行规划与设计。

以信息安全等级保护相关文件及 ISO27001/GBT22080 为指导，结合系统安全现状及未来发展趋势，建立一套完善的安全防护体系。通过体系化、标准化的信息安全风险评估，积极采取各种安全管理和安全技术防护措施，落实信息安全等级保护相关要求。

从技术与管理上提高网络与信息管理系统安全防护水平，防止信息网络瘫痪，防止应用系统破坏，防止业务数据丢失，防止信息泄密，防止终端病毒感染，防止有害信息传播，防止恶意渗透攻击，确保信息管理系统安全稳定运行，确保业务数据安全。

本次项目安全系统首先依托云平台提供的安全机制，实现平台安全。在此基础上本项目主要实现对应用软件的安全管理，各类安全提供主体分类如下：

表 5-26 安全提供主体分类表

序号	安全类型	提供主体
1	物理和环境安全	数据应用部分云平台负责，其他由系统承建方负责
2	网络和通信安全	云平台负责
3	设备和计算安全	租用部分云平台负责，项目采购硬件部分系统承建方负责
4	应用安全	系统承建方负责
5	数据安全	项目建设单位负责

4.3.6.1.2 物理和环境安全

厦门市健康医疗大数据治理及知识应用平台项目主数据采集、主数据存储、主数据备份等由大数据前置机、大数据一体机实现。本项目应用服务部分数据部署在云平台上，由平台统一提供安全等保三级服务，并提供云盾等相应的安全应用服务。本系统物理和环境安全设计及需求如下：

1、物理位置选择

机房场地应选择在具有防震、防风和防雨等能力的建筑内；机房场地应避免设在建筑物的顶层或地下室，否则应加强防水和防潮措施。

2、物理访问控制

机房出入口应配置电子门禁系统，控制、鉴别和记录进入的人员。机房出入口应安排专人值守，需进入机房的来访人员应经过申请和审批流程，并限制和监控其活动范围。

3、防盗窃和防破坏

将设备或主要部件进行固定，并设置明显的不易除去的标记；将通信线缆铺设在隐蔽处，可铺设在地下或管道中；设置机房防盗报警系统或设置有专人值守的视频监控系统。

4、防雷击

将各类机柜、设施和设备等通过接地系统安全接地；采取措施防止感应雷，例如设置防雷保安器或过压保护装置等。

5、防火

机房应设置火灾自动消防系统，能够自动检测火情、自动报警，并自动灭火；机房及相关的工作房间和辅助房应采用具有耐火等级的建筑材料；应对机房划分区域进行管理，区域和区域之间设置隔离防火措施。

6、防水和防潮

采取措施防止雨水通过机房窗户、屋顶和墙壁渗透；采取措施防止机房内水蒸气结露和地下积水的转移与渗透；安装对水敏感的检测仪表或元件，对机房进行防水检测和报警。

7、防静电

安装防静电地板并采用必要的接地防静电措施；采取措施防止静电的产生，例如采用静电消除器、佩戴防静电手环等。

8、温湿度控制

设置温、湿度自动调节设施，使机房温、湿度的变化在设备运行所允许的范围之内。

9、电力供应

在机房供电线路上配置稳压器和过电压防护设备；提供短期的备用电力供应，至少满足设备在断电情况下的正常运行要求；设置冗余或并行的电力电缆线

路为计算机系统供电。

10、电磁防护

电源线和通信线缆应隔离铺设，避免互相干扰；对关键设备实施电磁屏蔽。

4.3.6.1.3 网络和通信安全

本系统网络和通信安全设计及需求如下：

1、网络架构

保证网络设备的业务处理能力满足业务高峰期需要；保证网络各个部分的带宽满足业务高峰期需要；划分不同的网络区域，并按照方便管理和控制的原则为各网络区域分配地址；避免将重要网络区域部署在网络边界处且没有边界防护措施；提供通信线路、关键网络设备的硬件冗余，保证系统的可用性。

2、通信传输

采用校验码技术或密码技术保证通信过程中数据的完整性；采用密码技术保证通信过程中敏感信息字段或整个报文的保密性。

3、边界防护

保证跨越边界的访问和数据流通过边界防护设备提供的受控接口进行通信；能够对非授权设备私自联到内部网络的行为进行限制或检查；能够对内部用户非授权联到外部网络的行为进行限制或检查；限制无线网络的使用，确保无线网络通过受控的边界防护设备接入内部网络。

4、访问控制

在网络边界或区域之间根据访问控制策略设置访问控制规则，默认情况下除允许通信外受控接口拒绝所有通信；删除多余或无效的访问控制规则，优化访问控制列表，并保证访问控制规则数量最小化；对源地址、目的地址、源端口、目的端口和协议等进行检查，以允许/拒绝数据包进出；能根据会话状态信息为进出数据流提供明确的允许/拒绝访问的能力，控制粒度为端口级；在关键网络

节点处对进出网络的信息内容进行过滤，实现对内容的访问控制。

5、入侵防范

在关键网络节点处检测、防止或限制从外部发起的网络攻击行为；在关键网络节点处检测、防止或限制从内部发起的网络攻击行为；采取技术措施对网络行为进行分析，实现对网络攻击特别是新型网络攻击行为的分析；检测到攻击行为时，记录攻击源 IP、攻击类型、攻击目的、攻击时间，在发生严重入侵事件时应提供报警。

6、恶意代码防范

在关键网络节点处对恶意代码进行检测和清除，并维护恶意代码防护机制的升级和更新；在关键网络节点处对垃圾邮件进行检测和防护，并维护垃圾邮件防护机制的升级和更新。

7、安全审计

在网络边界、重要网络节点进行安全审计，审计覆盖到每个用户，对重要的用户行为和重要安全事件进行审计；审计记录应包括事件的日期和时间、用户、事件类型、事件是否成功及其他与审计相关的信息；对审计记录进行保护，定期备份，避免受到未预期的删除、修改或覆盖等；确保审计记录的留存时间符合法律法规要求；能对远程访问的用户行为、访问互联网的用户行为等单独行为进行审计和数据分析。

8、集中管控

划分出特定的管理区域，对分布在网络中的安全设备或安全组件进行管控；能够建立一条安全的信息传输路径，对网络中的安全设备或安全组件进行管理；对网络链路、安全设备、网络设备和服务器等的运行状况进行集中监测；对分散在各个设备上的审计数据进行收集汇总和集中分析；对安全策略、恶意代码、补丁升级等安全相关事项进行集中管理；能对网络中发生的各类安全事件进行识别、报警和分析。

通过采购云平台的安全服务，提供的网络和通信安全保障如下：

- DDoS 攻击检测：检测 DDoS 攻击行为，包括网络层 DDoS 及应用层 DDoS，针对每次 DDoS 攻击事件进行流量成分分析及流量图保留。

- web 攻击监测：通过分析应用层流量，实现对常见 web 攻击行为的识别，如 SQL 注入、XSS、上传漏洞等。

- 紧急事件分析：紧急事件包括：肉鸡行为、暴力破解成功、后门、DDoS 攻击事件和异常网络连接等，提供必要的智能分析，协助用户处理紧急事件。

4.3.6.1.4 设备和计算安全

本系统设备和计算安全设计及需求如下：

1、身份鉴别

对登录的用户进行身份标识和鉴别，身份标识具有唯一性，身份鉴别信息具有复杂度要求并定期更换；具有登录失败处理功能，应配置并启用结束会话、限制非法登录次数和当登录连接超时自动退出等相关措施；当进行远程管理时，应采取必要措施，防止鉴别信息在网络传输过程中被窃听；采用两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用动态口令、密码技术或生物技术来实现。

2、访问控制

对登录的用户分配账户和权限；重命名或删除默认账户，修改默认账户的默认口令；及时删除或停用多余的、过期的账户，避免共享账户的存在；进行角色划分，并授予管理用户所需的最小权限，实现管理用户的权限分离；由授权主体配置访问控制策略，访问控制策略规定主体对客体的访问规则；访问控制的粒度应达到主体为用户级或进程级，客体为文件、数据库表级；对敏感信息资源设置安全标记，并控制主体对有安全标记信息资源的访问。

3、安全审计

启用安全审计功能，审计覆盖到每个用户，对重要的用户行为和重要安全事件进行审计；审计记录应包括事件的日期和时间、用户、事件类型、事件是否成功及其他与审计相关的信息；对审计记录进行保护，定期备份，避免受到未预期的删除、修改或覆盖等；确保审计记录的留存时间符合法律法规要求；对审计进程进行保护，防止未经授权的中断。

4、入侵防范

遵循最小安装的原则，仅安装需要的组件和应用程序。关闭不需要的系统服务、默认共享和高危端口；通过设定终端接入方式或网络地址范围对通过网络进行管理的管理终端进行限制；能发现可能存在的漏洞，并在经过充分测试评估后，及时修补漏洞；能够检测到对重要节点进行入侵的行为，并在发生严重入侵事件时提供报警。

5、恶意代码防范

采用免受恶意代码攻击的技术措施或可信验证机制对系统程序、应用程序和重要配置文件/参数进行可信执行验证，并在检测到其完整性受到破坏时采取恢复措施。

6、资源控制

限制单个用户或进程对系统资源的最大使用限度；提供重要节点设备的硬件冗余，保证系统的可用性；对重要节点进行监视，包括监视 CPU、硬盘、内存等资源的使用情况；能够对重要节点的服务水平降低到预先规定的最小值进行检测和报警。

通过采购云平台的安全服务，为部署在云平台的有关设备提供的安全保障如下：

服务器安全：

- 基线检查：实现 windows、linux 操作系统的安全基线检查，包含弱口令、账户安全、可疑自启动项/服务项、可疑进程等关键点的检测。

- 网站后门查杀：通过规则匹配和语义动态解析方式对服务器中存在的后门木马进行精准查杀，提供隔离及忽略处置功能。

- 异地登陆告警：异地登录告警功能通过分析和记录用户常用登录位置，识别常用的登录区域（精确到地市级）。其中对于连续登录 6 次的登录地址将自动识别为常用登陆地不再告警，用户也可自行进行常用登陆地设置。

- 主机安全运维：可批量对防护主机进行运维操作，所有的运维命令以及执行记录均将全量审计记录，不可被删除。

- 暴力破解检测：对黑客进行暴力破解的行为进行实时检测，支持在 Windows 和 Linux 环境下对 SSH、RDP、Ftp、MySQL、MS SQL Server 等等常见服务的暴力破解行为进行监控。

补丁管理：

- 漏洞信息展示：基于主机扫描，发现主机漏洞并给出漏洞修复方式。

- 通过采购云平台的云盾和和云灾备服务，为部署在云端的的应用服务器提供安全保障。

- 威胁分析：

- 普通攻击类型包括：SQL 注入、XSS 攻击、代码 / 命令执行、本地文件包含、远程文件包含、脚本木马、上传漏洞、路径遍历、拒绝服务、越权访问、CSRF 和其他等。

- 攻击威胁分析包括：定点 Web 攻击、针对性主机密码爆破、撞库攻击，CMS 异常登陆等分析模型，能够分析出针对性攻击次数以及针对性攻击者人数，提供 TOP5 受威胁资产。

弱点分析：

- 应用漏洞分析：针对 web 应用层漏洞进行扫描，并提供扫描结果立即验证手段与对应的修补建议，周期性结合 nat 资产、主机资产自动化进行扫描，并对已经发现的应用漏洞进行统一验证，验证后进行应用漏洞状态的刷新。

- 主机漏洞分析：针对主机层漏洞进行扫描发现，并提供扫描结果与对应的修补建议。

- 弱口令分析：针对 web、ssh、FTP 等通用系统的登陆进行弱口令扫描，支持自定义弱口令添加，每晚结合 nat 资产、主机资产自动化进行扫描，并对已经发现的弱口令进行统一验证，验证后进行弱口令发现时间的刷新。

- 配置项检测：依据对外业务页面访问情况进行扫描发现，对 web 页面配置项的泄露进行告警，并在每晚进行已经发现的配置项泄露情况进行验证，对发现时间进行刷新。

情报收集：

- 重要漏洞情报：为用户提供近期最新软件的重要漏洞信息与详细内容链接，包含漏洞信息、信息来源、影响范围、修复方法等信息，并通过与弱点扫描模块的关联，实现自动化发现受影响主机资产信息。

- 行业安全新闻：为用户提供安全业内重大的安全事件通报，新闻推送。

安全运营：

- B-waf 阻断：可设置对 web 攻击行为是观察模式还是阻断模式。

- 暴力破解功能阻断：可设置对暴力破解攻击行为是观察模式还是阻断模式。

数据灾备：

- 实时数据传输及完整设备支持：采用远程复制技术，实现数据实时复制，网络具备自动或集中切换能力，业务处理系统就绪或运行中。

4.3.6.1.5 应用安全

应用和数据安全中要求的身份鉴别、访问控制、安全审计、软件容错和资源控制、数据保密性等多个方面跟应用程序自身设计的安全性有极大的关系。

涉及到大量不同用户所使用的应用程序自身也必然具备上述各项功能。

应用程序在开发过程中具备安全编码意识是非常重要的。开发人员应普遍接受安全开发培训，在应用系统开发的过程中，让源代码自身尽可能地安全，减少因此而产生的攻击面。同时，通过应用程序自带的功能实现访问控制、安全审计、软件容错、资源控制和数据加密等功能。

要求软件开发商定时对代码进行安全审计，进行阶段测试，即时排除安全漏洞，并采用软件开发标准体系完成代码的走查和评审。

对于开发文档须做到配置完整，在移交工作开发文档时注意不能泄露文档内容，并按照系统集成安全规范对开发人员进行管理。

软件开发避免直接使用开源产品，保护程序的安全性，并在测试过程中监控程序对后台资源的调用情况，发现异常情况做到及时清理和安全风险的预防。

所以，必须做好以下几个方面的应用安全措施：

1、授权管理

功能权限和数据权限分离。功能权限基于角色和权限点控制。整个平台所以权限点，提供了非常细致的功能控制粒度。同时提供自定义角色能力，使用户可以进行团队分工和协作。

基于数据包进行数据授权管理。支持组织间数据交换，项目间数据交换和项目内对成员的数据授权控制。支持数据授权有效期控制，支持授权生效过程中动态增删数据内容。

生产环境和开发环境数据授权隔离，通过授权可控制个人用户对生产数据的可见范围，生产数据对个人只读不可写，并且混合数据可用不可看。

2、日志审计

日志审计模块记录所有用户通过系统对功能模块、数据的操作日志，包括用户的帐号、权限和认证的管理日志以及系统自身服务调用等的系统日志，内容包含操作产品、操作人、操作目标和操作行为等信息。遇到特殊安全事件和系统故障，日志审计可以帮助管理员进行故障快速定位，并提供客观依据进行追查和

恢复。

3、代码安全

数据开发平台可提供代码加密功能，加密后的代码需要密钥解密才能查看到源代码，确保核心代码的安全。

代码安全另外还包括工作流的备份和恢复，数据开发平台可以将整个工作流以及工作流相关的节点、资源文件、调度依赖等进行离线备份和恢复，确保代码不丢失。

4、身份鉴别

对登录的用户进行身份标识和鉴别，身份标识具有唯一性，鉴别信息具有复杂度要求并定期更换；提供并启用登录失败处理功能，多次登录失败后应采取必要的保护措施；强制用户首次登录时修改初始口令；用户身份鉴别信息丢失或失效时，应采用技术措施确保鉴别信息重置过程的安全；采用两种或两种以上组合的鉴别技术对用户进行身份鉴别，且其中一种鉴别技术至少应使用动态口令、密码技术或生物技术来实现。

5、访问控制

提供访问控制功能，对登录的用户分配账户和权限；重命名或删除默认账户，修改默认账户的默认口令；及时删除或停用多余的、过期的账户，避免共享账户的存在；授予不同账户为完成各自承担任务所需的最小权限，并在它们之间形成相互制约的关系；由授权主体配置访问控制策略，访问控制策略规定主体对客体的访问规则；访问控制的粒度应达到主体为用户级，客体为文件、数据库表级、记录或字段级；对敏感信息资源设置安全标记，并控制主体对有安全标记信息资源的访问。

6、安全审计

提供安全审计功能，审计覆盖到每个用户，对重要的用户行为和重要安全事件进行审计；审计记录应包括事件的日期和时间、用户、事件类型、事件是否

成功及其他与审计相关的信息；对审计记录进行保护，定期备份，避免受到未预期的删除、修改或覆盖等；确保审计记录的留存时间符合法律法规要求；对审计进程进行保护，防止未经授权的中断。

7、软件容错

提供数据有效性检验功能，保证通过人机接口输入或通过通信接口输入的内容符合系统设定要求；在故障发生时，应能够继续提供一部分功能，确保能够实施必要的措施；在故障发生时，应自动保存易失性数据和所有状态，保证系统能够进行恢复。

8、资源控制

当通信双方中的一方在一段时间内未作任何响应，另一方应能够自动结束会话；能够对系统的最大并发会话连接数进行限制；能够对单个账户的多重并发会话进行限制。

4.3.6.1.6 数据安全

1、数据安全体系设计

图 5-8 数据安全体系表

安全管理体系由平台自身的安全实现层、平台提供的安全服务层、租户可

选的安全产品层和数据共享开放的安全策略层来构成。

（1）平台安全实现层

该层主要由底层平台本身提供相关能力，由云平台提供安全保障。保障平台在代码实现和部署配置时的产品自身安全性。

（2）平台安全服务层

该层与平台安全实现层类似，一般由底层平台本身提供相关的能力。

为租户和其用户提供平台基础性的安全服务能力，如：租户资源隔离、数据加密服务、用户访问控制和行为安全审计等。

此部分由云平台提供技术和安全保障

（3）安全产品层

该层主要由卫健委信息中心机房的安全产品提供相应的能力，与平台安全实现层和平台安全服务层是对接关系。

为用户提供可选的、集成的安全产品或工具，帮助用户根据其自行定义的安全策略对其拥有的系统、数据进行安全防护和运维管理。提供数据内容的脱敏工具，确保数据加工的过程能被评审、测试和批准，确保不对原有数据产生破坏，同时防止木马程序泄露数据。

（4）数据共享开放的安全策略层

该层主要由数据服务开放平台提供相关的安全能力。在本次项目中可以由已部署实施的数据交换子系统和数据服务子系统提供。

数据交换共享安全策略，保障各个租户的数据资源在分类、上架、交换、共享、开放等各个环节上的安全可控。数据分类分级策略描述数据的分类情况，访问和使用这些数据的要求；数据分类的原则是基于数据的价值、保密性和敏感性给数据合适的安全级别的基础。数据分类情况记录于数据目录，与数据描述联系在一起。数据上架管理策略定义了数据上架的安全规范和元数据信息要求，对

于不同密级的数据资产给与不同的安全上架策略、查找和访问策略的控制。

2、数据采集安全

数据采集系统仅完成数据同步/传输过程，整体数据传输过程完全控制于数据采集系统同步集群模型下，从而实现同步的通道以及同步数据流对用户完全隔离，保障数据同步过程中安全。如下图所示：

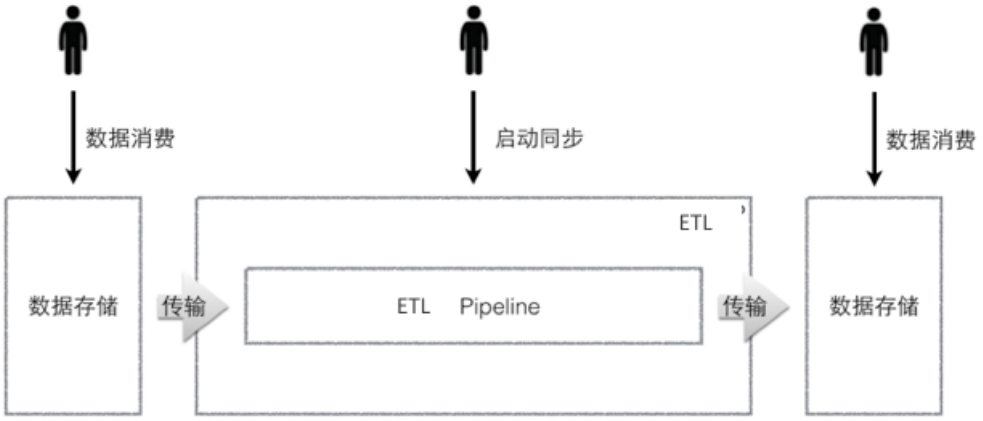


图 5-9 数据同步过程图

- 角色和权限

数据采集系统将用户角色仅分为两类: Pipeline Owner、数据采集系统管理员。数据采集系统避免了复杂的权限模型体系，只提供最基本的权限模型，保证对接应用可以按照自己业务逻辑设计权限系统。

- 多租户隔离

数据采集系统支持系统多租户隔离。数据采集系统的用户相互之间是无法管控对方的 Pipeline 及下属所有对象信息，包括 Job 配置信息、Job 传输数据流信息。同时，不同用户不同 Pipeline 下的 Job 运行环境互相隔离，从而保障数据采集系统租户与租户之间、Job 与 Job 之间的数据安全。

- 数据源签权

数据采集系统仅能负责用户对数据采集系统请求 API 鉴权，但无法负责对用户请求数据同步的源端和宿端的权限进行鉴别。同时，为了避免用户鉴权信息(例如 AccessId、AccessKey 等敏感字段)泄露，数据采集系统本身提供了一套安

全非对称加密方式，保证用户敏感信息不会存在泄密风险。

3、数据治理安全

基于大数据平台的数据治理系统提供数据治理安全，围绕数据治理过程中的活动、数据的访问和处理相关环节提供自动化保护。

- 连接到数据治理系统用户的身份集中认证

用户登录时会记录登录系统的用户和操作信息，包括登录的时间、位置等相关信息，用于日后审计查询。

- 数据治理系统的活动日志

用户在数据治理系统中的每次的操作信息，都会记录到日志，包括运行调度工作流、创建表、浏览表数据、数据编目等相关操作，这些日志信息可以作为日后审计和分析的依据。基于这些全面的活动日志信息，进行数据挖掘，可以实现可疑行为分析。

- 对项目空间、数据库访问的基于角色的权限控制

用户在数据治理系统中的角色和权限都是由管理员预先分配和定义，用户登录后在项目空间的活动权限，包括查看空间的对象、创建和删除表、进行工作流调度等活动都由不同的权限控制。

- 数据敏感度级别分类

数据的责任人可以定义和标识敏感的数据，基于数据表或者数据列字段设置敏感度等级。系统要求支持设置 9 个级别的敏感度，结合用户的敏感度级别即可以控制用户对数据访问的权限。

- 支持使用沙箱机制实现对数据的访问

数据的实时查询分析，以及跨部门或者团队的共享，一般设定不直接对原始数据的访问。当专业用户如数据科学家、业务分析师、运营人员、其他团队需要使用数据进行测试、探索、分析、挖掘时，通常可以给这些用户分配专用的沙箱并且将需要的数据放入沙箱，必要时进行数据的脱敏，以供其进行数据相关的活动，同时将这些活动计入行为日志。沙箱机制可以实现用户活动之间完全隔离，

互不影响，同时根据治理策略，以生命周期来管理这些沙箱。

- 数据实体访问权限控制

用户对自己拥有的项目空间、数据表等具有完全的管理权限和责任。对于数据表，不论该数据表位于哪个项目空间，只要是用户具有相关的权限，则可以设置其安全等级，配置 ACL 授权或者策略授权，以限制其他人员对该表的访问。表的权限包括 Describe，Select，Alter，Update，Drop 等。对于项目管理员角色的用户，还可以设置是否允许对象创建者进行授权、是否限制项目空间的数据外流、ACL 授权是否开启、Policy 授权是否开启等灵活的安全功能设置。

4、数据存储安全

（1）自动存储容错机制，保障数据高可靠性

基于数据多副本备份、自动存储容错、系统错误监控、故障自动迁移等技术，确保数据的安全性，数据可用性达到 99.99%。

在大数据计算平台中，数据采用 3 副本，并且保障各副本数据保存在不同的计算节点中。

（2）租户隔离和用户权限控制，保障数据计算和访问安全

在租户层面，大数据计算平台支持租户资源隔离，包括 CPU、内存等，确保不同租户间数据计算的安全；大数据计算平台支持基于 ACL 的用户权限管理，可以配置灵活的数据访问控制策略，防止数据越权访问。

5、数据开发安全

（1）授权管理

功能权限和数据权限分离。功能权限基于角色和权限点控制。整个平台上百个个权限点，提供了非常细致的功能控制粒度。同时提供自定义角色能力，使用户可以进行团队分工和协作。

基于数据包进行数据授权管理。支持组织间数据交换，项目间数据交换和项目内对成员的数据授权控制。支持数据授权有效期控制，支持授权生效过程中

动态增删数据内容。

生产环境和开发环境数据授权隔离，通过授权可控制个人用户对生产数据的可见范围，生产数据对个人只读不可写，并且混合数据可用不可看。

（2）日志审计

日志审计模块记录所有用户通过系统对功能模块、数据的操作日志，包括用户的帐号、权限和认证的管理日志以及系统自身服务调用等的系统日志，内容包含操作产品、操作人、操作目标和操作行为等信息。遇到特殊安全事件和系统故障，日志审计可以帮助管理员进行故障快速定位，并提供客观依据进行追查和恢复。

（3）数据脱敏

数据开发子系统可提供数据脱敏安全服务，内置多种常用的数据脱敏方案，帮助用户对某些敏感信息通过脱敏规则进行数据的变形，在保存数据原始特征（比如数据类型、长度和属性等）的同时改变它的数值，确保应用程序可在使用脱敏数据的开发与测试过程中正常运行，从而防止敏感数据被滥用、被泄露的风险，保护敏感数据免于未经授权的访问。帮助安防系统提高安全性和保密等级、满足安全性的规范要求，以及由管理/审计机关所要求的隐私标准，以防止敏感数据被滥用、被泄露的风险。

（4）代码安全

数据开发子系统可提供代码加密功能，加密后的代码需要密钥解密才能查看到源代码，确保核心代码的安全。

代码安全另外还包括工作流的备份和恢复，数据开发子系统可以将整个工作流以及工作流相关的节点、资源文件、调度依赖等进行离线备份和恢复，确保代码不丢失。

4.3.6.1.7 安全策略和管理制度

1、安全策略

应制定信息安全工作的总体方针和安全策略，说明机构安全工作的总体目标、范围、原则和安全框架等。

2、管理制度

对安全管理活动中的各类管理内容建立安全管理制度；对要求管理人员或操作人员执行的日常管理操作建立操作规程；形成由安全策略、管理制度、操作规程、记录表单等构成的全面的信息安全管理制度体系。

3、制定和发布

指定或授权专门的部门或人员负责安全管理制度的制定；安全管理制度应通过正式、有效的方式发布，并进行版本控制。

4、评审和修订

应定期对安全管理制度的合理性和适用性进行论证和审定，对存在不足或需要改进的安全管理制度进行修订。

4.3.6.1.8 安全管理机构和人员

1、岗位设置

成立指导和管理信息安全工作的委员会或领导小组，其最高领导由单位主管领导委任或授权；设立信息安全管理工作的职能部门，设立安全主管、安全管理各个方面的负责人岗位，并定义各负责人的职责；设立系统管理员、网络管理员、安全管理员等岗位，并定义部门及各个工作岗位的职责。

2、人员配备

配备一定数量的系统管理员、网络管理员、安全管理员等；配备专职安全管理员，不可兼任。

3、授权和审批

根据各个部门和岗位的职责明确授权审批事项、审批部门和批准人等；针对系统变更、重要操作、物理访问和系统接入等事项建立审批程序，按照审批程

序执行审批过程，对重要活动建立逐级审批制度；定期审查审批事项，及时更新需授权和审批的项目、审批部门和审批人等信息。

4、沟通和合作

加强各类管理人员之间、组织内部机构之间以及信息安全职能部门内部的合作与沟通，定期召开协调会议，共同协作处理信息安全问题；加强与公安机关、各类供应商、业界专家及安全组织的合作与沟通。

5、审核和检查

定期进行常规安全检查，检查内容包括系统日常运行、系统漏洞和数据备份等情况；定期进行全面安全检查，检查内容包括现有安全技术措施的有效性、安全配置与安全策略的一致性、安全管理制度的执行情况等；制定安全检查表格实施安全检查，汇总安全检查数据，形成安全检查报告，并对安全检查结果进行通报。

6、人员录用

指定或授权专门的部门或人员负责人员录用；被录用人员的身份、背景、专业资格和资质等进行审查，对其所具有的技术技能进行考核；与被录用人员签署保密协议，与关键岗位人员签署岗位责任协议。

7、人员离岗

及时终止离岗员工的所有访问权限，取回各种身份证件、钥匙、徽章等以及机构提供的软硬件设备；办理严格的调离手续，并承诺调离后的保密义务后方可离开。

8、安全意识健康医疗和培训

对各类人员进行安全意识健康医疗和岗位技能培训，并告知相关的安全责任和惩戒措施；针对不同岗位制定不同的培训计划，对信息安全基础知识、岗位操作规程等进行培训；定期对不同岗位的人员进行技能考核。

9、外部人员访问管理

确保在外部人员物理访问受控区域前先提出书面申请，批准后由专人全程陪同，并登记备案；确保在外部人员接入受控网络访问系统前先提出书面申请，批准后由专人开设账户、分配权限，并登记备案；外部人员离场后应及时清除其所有的访问权限；获得系统访问授权的外部人员应签署保密协议，不得进行非授权操作，不得复制和泄露任何敏感信息。

4.3.6.1.9 安全运维管理

1、环境管理

指定专门的部门或人员负责机房安全，对机房出入进行管理，定期对机房供配电、空调、温湿度控制、消防等设施进行维护管理；

建立机房安全管理制度，对有关机房物理访问，物品带进、带出机房和机房环境安全等方面的管理作出规定；

不在重要区域接待来访人员和放置包含敏感信息的纸档文件、移动介质等。

2、资产管理

编制并保存与保护对象相关的资产清单，包括资产责任部门、重要程度和所处位置等内容。其中，租用的硬件设备、网络设备等资产属于云平台所有。智慧健康医疗大数据融合分析平台内部标准化的主题数据资产属于市卫生健康局所有。有关软件系统资产属于市卫生健康局所有。

根据资产的重要程度对资产进行标识管理，根据资产的价值选择相应的管理措施；

对信息分类与标识方法作出规定，并对信息的使用、传输和存储等进行规范化管理。

3、设备维护管理

对各种设备（包括备份和冗余设备）、线路等指定专门的部门或人员定期进行维护管理；

建立配套设施、软硬件维护方面的管理制度，对其维护进行有效的管理，包括明确维护人员的责任、维修和服务的审批、维修过程的监督控制等；

确保信息处理设备必须经过审批才能带离机房或办公地点，含有存储介质的设备带出工作环境时其中重要数据必须加密；

有存储介质的设备在报废或重用前，应进行完全清除或被安全覆盖，确保该设备上的敏感数据和授权软件无法被恢复重用。

4、漏洞和风险管理

采取必要的措施识别安全漏洞和隐患，对发现的安全漏洞和隐患及时进行修补或评估可能的影响后进行修补。

5、网络和系统安全管理

划分不同的管理员角色进行网络和系统的运维管理，明确各个角色的责任和权限；

指定专门的部门或人员进行账户管理，对申请账户、建立账户、删除账户等进行控制；

建立网络和系统安全管理制度，对安全策略、账户管理、配置管理、日志管理、日常操作、升级与打补丁、口令更新周期等方面作出规定；

制定重要设备的配置和操作手册，依据手册对设备进行安全配置和优化配置等；

详细记录运维操作日志，包括日常巡检工作、运行维护记录、参数的设置和修改等内容；

指定专门的部门或人员对日志、监测和报警数据等进行分析、统计，及时发现可疑行为；

严格控制变更性运维，经过审批后才可改变连接、安装系统组件或调整配置参数，操作过程中应保留不可更改的审计日志，操作结束后应同步更新配置信

息库；

严格控制运维工具的使用，经过审批后才可接入进行操作，操作过程中应保留不可更改的审计日志，操作结束后应删除工具中的敏感数据；

严格控制远程运维的开通，经过审批后才可开通远程运维接口或通道，操作过程中应保留不可更改的审计日志，操作结束后立即关闭接口或通道；

保证所有与外部的连接均得到授权和批准， 应定期检查违反规定无线上网及其他违反网络安全策略的行为。

6、恶意代码防范管理

提高所有用户的防恶意代码意识，告知对外来计算机或存储设备接入系统前进行恶意代码检查等；

对恶意代码防范要求做出规定，包括防恶意代码软件的授权使用、恶意代码库升级、恶意代码的定期查杀等；

定期验证防范恶意代码攻击的技术措施的有效性。

7、配置管理

记录和保存基本配置信息，包括网络拓扑结构、各个设备安装的软件组件、软件组件的版本和补丁信息、各个设备或软件组件的配置参数等；将基本配置信息改变纳入变更范畴，实施对配置信息改变的控制，并及时更新基本配置信息库。

8、备份与恢复管理

识别需要定期备份的重要业务信息、系统数据及软件系统等；规定备份信息的备份方式、备份频度、存储介质、保存期等；根据数据的重要性和数据对系统运行的影响，制定数据的备份策略和恢复策略、备份程序和恢复程序等。

9、安全事件处置

及时向安全管理部门报告所发现的安全弱点和可疑事件；制定安全事件报告和处置管理制度，明确不同安全事件的报告、处置和响应流程，规定安全事件

的现场处理、事件报告和后期恢复的管理职责等；在安全事件报告和响应处理过程中，分析和鉴定事件产生的原因，收集证据，记录处理过程，总结经验教训；对造成系统中断和造成信息泄漏的重大安全事件应采用不同的处理程序和报告程序。

10、应急预案管理

规定统一的应急预案框架，具体包括启动预案的条件、应急组织构成、应急资源保障、事后健康医疗和培训等内容；制定重要事件的应急预案，包括应急处理流程、系统恢复流程等内容；定期对系统相关的人员进行应急预案培训，并进行应急预案的演练；定期对原有的应急预案重新评估，修订完善。

4.3.6.2 安全测评

平台在上线前要由第三方安全测评单位进行完备的性能和安全性测试，并出具安全测评报告，安全测评报告应包含物理、完了、设备、应用和数据等安全性测试相关内容，以便采取措施应对发现的安全问题。

对平台安全性能定期进行等级测评，发现不符合相应等级保护标准要求的及时整改，在发生重大变更或级别发生变化时进行等级测评，确保测评机构的选择符合国家有关规定。

本次项目建设的系统主要部署于云平台，故物理安全、网络安全、主机安全等基础安全部分由云平台负责保障。

本次项目主要考虑范围为应用安全、数据安全和系统建设管理部分，由云平台配合项目承建方共同完成安全测评工作。

4.3.6.2.1 应用安全

1、身份鉴别（S3）

（1）应提供专用的登录控制模块对登录用户进行身份标识和鉴别；

（2）应对同一用户采用两种或两种以上组合的鉴别技术实现用户身份鉴别；

（3）应提供用户身份标识唯一和鉴别信息复杂度检查功能，保证应用系统中不存在重复用户身份标识，身份鉴别信息不易被冒用；

（4）应提供登录失败处理功能，可采取结束会话、限制非法登录次数和自动退出等措施；

（5）应启用身份鉴别、用户身份标识唯一性检查、用户身份鉴别信息复杂度检查以及登录失败处理功能，并根据安全策略配置相关参数。

2、访问控制（防火墙）

（1）应提供访问控制功能，依据安全策略控制用户对文件、数据库表等客体的访问；

（2）访问控制的覆盖范围应包括与资源访问相关的主体、客体及它们之间的操作；

（3）应由授权主体配置访问控制策略，并严格限制默认帐户的访问权限；

（4）应授予不同帐户为完成各自承担任务所需的最小权限，并在它们之间形成相互制约的关系。

（5）应具有对重要信息资源设置敏感标记的功能；

（6）应依据安全策略严格控制用户对有敏感标记重要信息资源的操作；

3、安全审计

（1）应提供覆盖到每个用户的安全审计功能，对应用系统重要安全事件进行审计；

（2）应保证无法单独中断审计进程，无法删除、修改或覆盖审计记录；

（3）审计记录的内容至少应包括事件的日期、时间、发起者信息、类型、描述和结果等；

（4）应提供对审计记录数据进行统计、查询、分析及生成审计报表的功能。

4、剩余信息保护

（1）应保证用户鉴别信息所在的存储空间被释放或再分配给其他用户前得到完全清除，无论这些信息是存放在硬盘上还是在内存中；

（2）应保证系统内的文件、目录和数据库记录等资源所在的存储空间被释放或重新分配给其他用户前得到完全清除。

5、通信完整性

应采用密码技术保证通信过程中数据的完整性。

6、通信保密性

（1）在通信双方建立连接之前，应用系统应利用密码技术进行会话初始化验证；

（2）应对通信过程中的整个报文或会话过程进行加密。

7、抗抵赖

（1）应具有在请求的情况下为数据原发者或接收者提供数据原发证据的功能；

（2）应具有在请求的情况下为数据原发者或接收者提供数据接收证据的功能。

8、软件容错

（1）应提供数据有效性检验功能，保证通过人机接口输入或通过通信接口输入的数据格式或长度符合系统设定要求；

（2）应提供自动保护功能，当故障发生时自动保护当前所有状态，保证系统能够进行恢复。

9、资源控制

（1）当应用系统的通信双方中的一方在一段时间内未作任何响应，另一方

应能够自动结束会话；

（2）应能够对系统的最大并发会话连接数进行限制；

（3）应能够对单个帐户的多重并发会话进行限制；

（4）应能够对一个时间段内可能的并发会话连接数进行限制；

（5）应能够对一个访问帐户或一个请求进程占用的资源分配最大限额和最小限额；

（6）应能够对系统服务水平降低到预先规定的最小值进行检测和报警；

（7）应提供服务优先级设定功能，并在安装后根据安全策略设定访问帐户或请求进程的优先级，根据优先级分配系统资源。

4.3.6.2.2 数据安全

1、数据完整性

（1）应能够检测到系统管理数据、鉴别信息和重要业务数据在传输过程中完整性受到破坏，并在检测到完整性错误时采取必要的恢复措施；

（2）应能够检测到系统管理数据、鉴别信息和重要业务数据在存储过程中完整性受到破坏，并在检测到完整性错误时采取必要的恢复措施。

2、数据保密性

（1）应采用加密或其他有效措施实现系统管理数据、鉴别信息和重要业务数据传输保密性；

（2）应采用加密或其他保护措施实现系统管理数据、鉴别信息和重要业务数据存储保密性。

4.3.6.2.3 系统建设管理

1、软件开发

（1）应确保开发环境与实际运行环境物理分开，开发人员和测试人员分离，测试数据和测试结果受到控制；

（2）应制定软件开发管理制度，明确说明开发过程的控制方法和人员行为准则；

（3）应制定代码编写安全规范，要求开发人员参照规范编写代码；

（4）应确保提供软件设计的相关文档和使用指南，并由专人负责保管；

（5）应确保对程序资源库的修改、更新、发布进行授权和批准。

2、产品采购和使用

（1）应确保安全产品采购和使用符合国家的有关规定；

（2）应确保密码产品采购和使用符合国家密码主管部门的要求；

（3）应指定或授权专门的部门负责产品的采购；

（4）应预先对产品进行选型测试，确定产品的候选范围，并定期审定和更新候选产品名单。

3、系统安全管理

（1）应根据业务需求和系统安全分析确定系统的访问控制策略；

（2）应定期进行漏洞扫描，对发现的系统安全漏洞及时进行修补；

（3）应安装系统的最新补丁程序，在安装系统补丁前，首先在测试环境中测试通过，并对重要文件进行备份后，方可实施系统补丁程序的安装；

（4）应建立系统安全管理制度，对系统安全策略、安全配置、日志管理和日常操作流程等方面做出具体规定；

（5）应指定专人对系统进行管理，划分系统管理员角色，明确各个角色的权限、责任和风险，权限设定应当遵循最小授权原则；

（6）应依据操作手册对系统进行维护，详细记录操作日志，包括重要的日

常操作、运行维护记录、参数的设置和修改等内容，严禁进行未经授权的操作；

（7）应定期对运行日志和审计数据进行分析，以便及时发现异常行为。

第5章　项目组织机构和人员培训

5.1 领导和管理机构

厦门市健康医疗大数据治理及知识应用平台建设是一项大规模的系统工程，它不仅涉及到技术实现的方法和手段，还涉及到实施期间各项资源的管理与调配。为了保证信息化建设的实施成功，需要专门成立项目建设领导小组,负责项目建设的重大决策,并负责审批项目建设中的重要文件,指导项目建设。

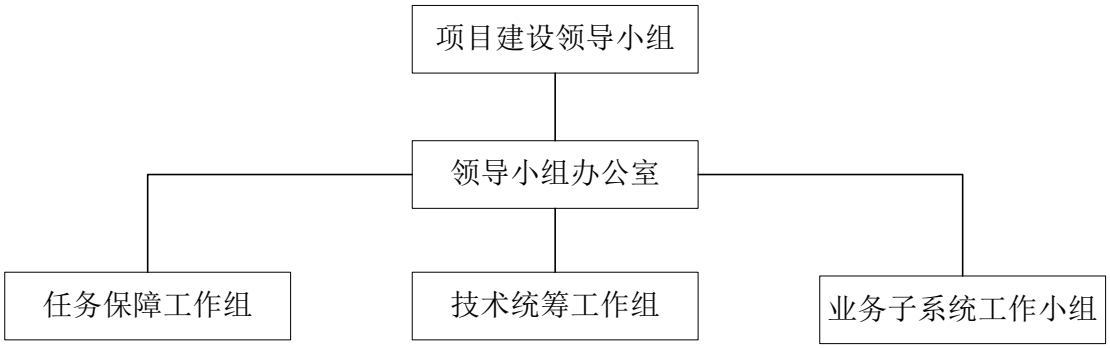


图 6-1 项目组织机构图

在选定供应商后，项目建设领导小组增加供应商的专家，为领导的项目建设决策提供咨询服务。

项目建设管理机构为厦门市健康医疗大数据中心，主要有以下职责：

1、负责从宏观上把握本项目建设的方向；

2、协调系统建设过程中资源调用；

3、系统建设过程中各种行政事务和重大技术问题的决策；

4、项目建设中的方案、变更以及资金拨付的审查与批复等。

5.1.1 项目建设领导小组

成立“厦门市健康医疗大数据治理及知识应用平台项目”建设工作领导小组。

主要职责：

- 研究、决定和部署“厦门市健康医疗大数据治理及知识应用平台项目”重大事项，研究、分析和决定民政业务信息系统的需求；
- 审定“厦门市健康医疗大数据治理及知识应用平台项目”技术标准规范、可行性研究报告、设计技术方案和概算报告、实施方案；
- 指导检查各业务处室各单位本项目建设相关工作；
- 研究部署“厦门市健康医疗大数据治理及知识应用平台项目”建设工作的其它重大事项。

领导小组每 2 周至少一次例会，领导小组组长应经常关注、参与和指导实施工作，及时处理各种问题。

5.1.2 领导小组办公室

领导小组下设办公室，负责领导小组日常工作。

主要职责：

- 负责编制“厦门市健康医疗大数据治理及知识应用平台项目”项目可行性研究报告、项目初步设计方案和概算报告、项目实施方案和技术标准规范；
- 负责“厦门市健康医疗大数据治理及知识应用平台项目”项目的组织管理工作；
- 负责检查与考评各业务处室各单位本项目建设相关工作；
- 负责与承建单位对接与组织实施。

办公室内设任务保障、技术统筹两个 3 人工作组

1、任务保障工作组

工作任务：主要承担项目建设中属行政管理范畴的工作，负责项目建设中公司内部、外部的工作协调，组织项目招标、项目验收、项目建设督察等。

2、技术统筹工作组（集中办公）

工作任务：主要承担项目建设总体设计组织和技术方案审核。负责与承建单位的技术对接，组织建立数据标准、交换与共享、应用支撑系统、系统集成、系统上云等；负责业务需求收集与审核，负责可研报告、技术方案的编制等。

5.1.3 业务子系统工作小组

成立业务子系统工作小组。

业务子系统工作小组成员由各业务处室各单位自行组织，小组组长由相关处室单位主要负责同志担任，并确定一名联络员，名单报项目领导小组办公室。

工作任务：

结合本业务处室单位工作，提出明确具体的业务需求；

- 负责制订本处室单位业务条线的业务规范和工作流程，确定数据规范；
- 协调各业务处室各单位，推进厦门市健康医疗大数据治理系统建设工作；
- 配合任务保障工作组、技术统筹工作组共同做好本业务子系统的建设和组织管理工作；
- 负责对本业务子系统的质检、验收、应用推广和考评工作。

5.1.4 项目实施机构

由厦门市卫生健康医疗大数据中心具体组织本项目的有关建设工作。在组织专家进行方案论证的基础上，形成厦门市健康医疗大数据治理及知识应用平台项目规划方案。根据规划方案，形成建设实施方案。按照建设实施方案进行招标，确定系统承建单位，通过规范的操作程序，保证一流的公司和技术人员参与。在建设实施方案确定的基础上，由中标方负责项目建设任务实施。建设机构要严格按照总体方案的规定产品规格，进行系统建设实施工作。整个项目的建设过程要严格按照项目管理技术和规范进行管理，项目建设单位要密切与建设项目承担单位的合作，协调各方面力量，确保工程的进度和质量达到预期要求。

厦门市健康医疗大数据治理及知识应用平台建设项目实施机构主要为厦门市健康医疗大数据中心通过招标方式中标的系统开发商和网络安全设备提供商等。项目实施机构的设立将体现统一领导、分工明确、职责清楚、层次分明、同时又能协调配合的原则。根据本项目各部分工作内容性质，建议项目实施组织机构图如下：

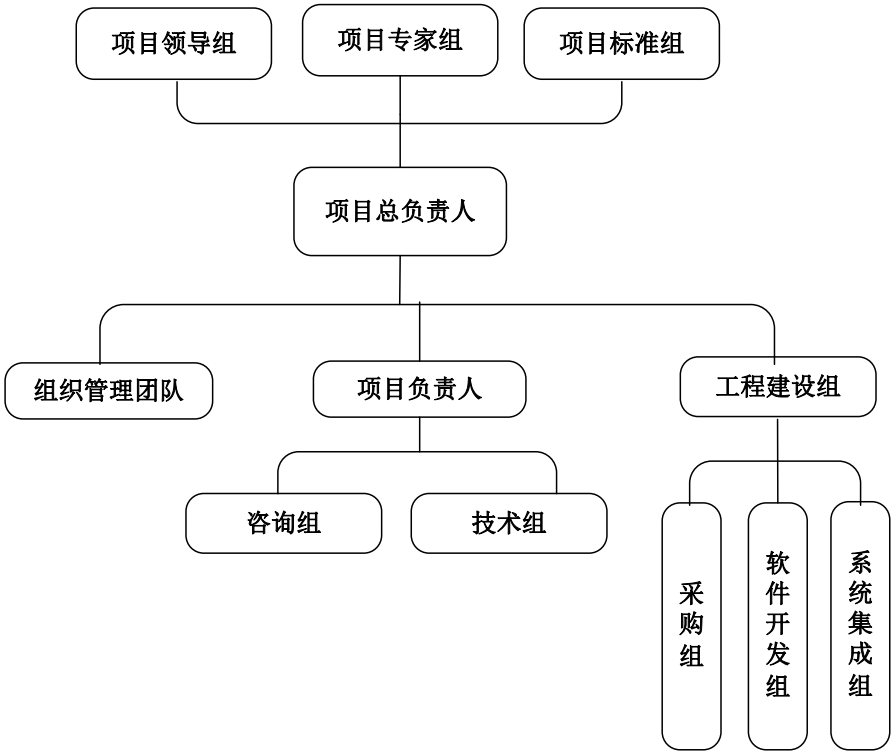


图 6-2 项目实施组织架构图

各岗位职责说明如下：

1、项目领导组

整个项目的工作进度、项目质量等的审核批准和监督；对项目实施中的重大决策做出决定；负责合理调配人力及物力资源，以确保本项目高质量、高效率、顺利的进行实施。

2、项目专家组

对整个项目提供技术咨询和指导，把握重要技术关键点并进行评估与确认，

对项目发展方向规划进行明确，对阶段性的成果进行审核。专家指导组将对项目组的工作进行协调和指导。

3、项目标准化组

负责试点建设项目的信息系统标准、服务规范和监管流程标准的研究与编制工作。

4、项目总负责人

负责各合作单位的协调以及与领导组、专家组合标准组的沟通，控制项目总体进度和质量。

5、项目负责人

组织各专业或专业负责人会议，进行内容策划和协调；负责编制项目计划书和设计大纲；控制项目总体进度；负责相关质量记录的收集和管理。

6、组织管理团队

负责项目组与厂家以及政府负责相关部门的沟通、协调。

7、项目咨询组

负责试点建设项目的产业化运营的商业模式策划，运营方案筹划，产业带动策划等。

8、项目技术组

负责项目顶层设计规划和试点项目的建设方案，方案获批后，与工程建设组共同完成项目实施方案。

9、项目工程建设组

在项目建设方案和项目实施方案基础上，完成试点项目所需的软硬件采购及集成，包括硬件、软件、集成小组。

创新思路，建立完善的项目实施管理制度。统筹考虑厦门健康医疗大数据治

理系统的发展规划、市场环境及项目实施全过程的管理需求因素，制定合理可行的项目招投标、融资、监督管理、评估考核、运行维护等制度，理顺项目全流程管理体制机制，确保项目的顺利实施与良好运行。

5.1.5 运行维护机构

本项目建成后，整体归属厦门市健康医疗大数据中心管理，负责网络、系统的正常运行和维护。

系统的运行维护必须建立健全的组织机构和切实可行的运行管理机制，才能保证系统的正常稳定运行。

运行维护机构是系统投入运行后保证系统在任何情况下都能正常、稳定、可靠运作的保障，必须由常设的部门或组织负责系统的维护和管理，并且需要足够的资金予以支持。

运行维护机构的工作内容有：

- 建立健全系统运行维护规章制度，保证规章制度的贯彻、实施；
- 对系统运行实行全时、全程监控，建立完整的运行维护记录；
- 确保系统安全、顺畅运行，快速处理和解决系统故障；
- 为使用者提供培训和技术支持；
- 不断优化结构、更新设备和升级应用，改善系统的运行效率。

5.2 技术力量和人员配置

厦门市健康医疗大数据治理及知识应用平台项目涉及的专业技术包括标准规范体系建设、网络管理、应用系统开发，其他还包括业务梳理、需求分析等。因此必须配备相应的专业技术人员做到对口维护。

为有效支撑项目的建设和维护，厦门市健康医疗大数据中心会需成立一支经验丰富、技术过硬、管理规范的信息化人才队伍，专业技术人员，拟配置具有专业技能、专业职称的网络技术人员、安全技术人员、计算存储技术人员、网络维

护人员等等对项目运行进行全力支撑，专业技术人员拟配置如下，人员由厦门市健康医疗大数据中心抽调专人与项目实施单位共同组成。人员配置表如下：

表 6-1 人员配置表

人员配置		人数	技术职称分布
项目管理人员		1	高级工程师 1
技术管理人员	网络技术人员	2	工程师 1
	安全技术人员	2	高级工程师 1
	厦门健康医疗大数据治理系统技术人员	3	高级工程师 1
	应用运维人员	2	高级工程师 2
	数据库管理人员	2	高级工程师 1
	数据录入与采集人员	4	工程师 4
	资料整理人员	2	工程师 2

5.3 人员培训方案

5.3.1 培训目的

用户培训是保证整个项目实施运行的重要环节。为使厦门市健康医疗大数据治理及知识应用平台项目达到预期目标，并充分发挥其效能，必须对参加本项目建设、使用、运行维护的业务人员、技术人员、管理人员提供相应的培训，确保

在项目建设、系统部署、系统管理、运行维护等各个阶段都有具备相应能力的人员参与保障，实现项目预期目标。针对技术人员，经过各类计算机专业技术培训，可以有效地提高对计算机系统的管理水平、操作水平和维护能力，从而确保计算机系统的安全运行。针对管理人员和业务人员，经过各类应用系统的操作使用培训，可以有效地提高对管理信息系统的管理水平、操作水平和维护能力，从而确保日常业务系统的正常运作。用户培训的宗旨是实用性与针对性的结合，规范化与系统化的结合，切实有效的提高用户的专业与技术水平。

5.3.2 培训对象

为确保本项目建设任务顺利实施以及建成后达到预期的目标，在项目建设和试运行阶段考虑对参与项目建设和系统运行维护的各类管理人员、专业技术人员和操作系统使用人员进行系统培训。

1、管理人员：管理的培训应以掌握基本建设程序、相关政策法规为主，了解招投标理论，将项目管理方法与实践相结合，更好地进行项目的管理和控制。

2、业务人员：业务人员的培训在建设前期以掌握基本建设程序、相关政策法规为主，了解招投标理论，将项目管理方法与实践相结合，更好地进行项目的管理和控制；在建设后期以掌握软件信息交换、信息发布等基本业务流程等为主。

3、技术人员：技术人员经过培训能够掌握软件系统进行维护、修改和升级，掌握设备安装与调试的方法，熟悉数据备份的方法、掌握系统故障诊断与排除的方法、熟悉系统性能监测与分析的方法等专业技能，熟练掌握新系统的业务处理流程，熟练使用应用软件系统的计算机操作。

5.3.3 培训内容

1、计算机应用基础培训

了解计算机的组成和网络基本知识；掌握计算机操作系统的基本知识和使用技巧；掌握浏览器使用和在网上搜索信息的方法；掌握常用软件的安装和卸载；

掌握计算机病毒的基本知识和常用的检测杀毒方法；掌握网络安全的基本概念；学习信息系统的管理规定。

计算机应用基础培训以本地培训为原则，由现有信息技术力量自行负责解决，不单独估算投资。

2、系统管理培训

系统项目中涉及到的软件和设备较多、对技术人员的技术能力要求较高，为了项目的顺利实施和今后日常维护工作的正常进行，相关技术人员需具备较全面的知识。为此，需制定一个循序渐进、由浅入深、全面的培训计划，为项目培养雄厚的技术力量。系统管理培训主要内容包括：

（1）网络管理：使系统管理员了解掌握网络的硬件及软件的构成；局域网、广域网介绍；TCP/IP 协议详解；子网划分；路由器的使用及常用命令；交换机的系统介绍；网络设计及维护。

（2）系统安全：了解掌握计算机系统安全策略；安全体系结构与安全模型；公钥基础设施；身份认证技术；访问控制与防火墙；网络与通信安全；操作系统安全；应用系统安全；漏洞扫描与入侵检测。

（3）数据库管理：了解数据库的运行方法及原理，能够实现简单的备份和恢复；能够用 SQL 语句实现数据库的检索和维护。

（4）应用系统管理和维护：从管理员的角度对各类应用系统和专网门户系统进行系统管理和维护监控。

第6章　项目实施进度

6.1 项目建设期

厦门市健康医疗大数据治理及知识应用平台项目是一项系统工程，涉及面广、实施难度较高，综合考虑实际业务需要、项目管理能力和项目风险控制等各种因素，本项目建设期预计需要 1 年，从 2020 年 1 月-2020 年 12 月平台正式完成上线使用。

6.2 实施进度计划

1、 项目审批立项阶段：

第一个月至第二个月：开展厦门市健康医疗大数据治理及知识应用平台项目前期准备工作；完成项目相关方案编制；完成前期各部门的审批程序。

2、 项目实施阶段：

第三个月至第四个月：完成项目招投标、谈判工作；

第五月至十个月：中标供应商进行项目需求分析、系统软件开发，系统软件进行部署及全面测试，各应用系统上线试运行等；其中在数据治理阶段预计花费四到五个月时间。

3、 项目试运行和验收阶段：

第十一个月至第十四个月：项目技术培训和工程验收工作。

第7章　投资概算

7.1 投资概算的有关说明

7.1.1 投资概算说明

1、本项目投资依据国家建设项目投资的有关规定编制，投资遵循“符合规范、结合实际、经济合理、不重不漏、计算正确”的指导原则。

2、费率按国家颁布的有关规范、规定及文件执行。只对本期建设方案进行。

3、本项目建设投资主要包括软件及服务购置费用，项目相关费用均按国家标准计取，根据当地市场标准，结合业主询价情况，部分费用做了下浮说明。

4、方案设备价格（如有）参照厂商报价和有关资料，设备的运杂费包含在设备费中，不另单列。系统设备及配套设施价格按目前主要厂商询价测算。

5、应用系统软件研购置费用：参照相关应用系统软件需要的工作量和人工费用。

6、数据治理服务服务费用：参照实际治理服务的相关工作量和人员工资进行概算。

7.1.2 投资概算依据

- 中华人民共和国信息产业部《电子建设工程概（预）算编制办法及计价依据》，2015 年；
- 中华人民共和国信息产业部《电子建设工程定额》，2015 年；
- 中华人民共和国国家发展改革委员会《国家电子政务工程建设项目管理暂行办法》，2007 年；
- 中华人民共和国国家发展改革委员会、建设部《工程勘察设计收费标准（2002 年修订本）》，2002 年；
- 中华人民共和国国家发展改革委员会、建设部《建设工程监理与相关服务收费标准》，2007 年；

- 《基本建设项目建设成本管理规定》财建〔2016〕504 号);
- 《建设项目前期工作咨询收费暂行规定》计价格〔1999〕1283 号
- 软件成本模型（COCOMOII+FPA)，2000 年。

软件开发费用在参照类似工程费用基础上，根据需求分析和业务功能工作人月，标准方面参考同行业系统开发的市场报价基础上，按 2 万元/人月计算。

7.2 项目总投资概算与概算明细表

7.2.1 项目总投资概算表

单位：万元

项目名称：厦门市健康医疗大数据治理及知识应用平台			
项目投资概算表			
序号	费用构成及名称	金额	备注
一	**工程建设费用**		
1	应用系统工程		
2	系统软件		
3	计算机网络系统工程费		
4	主机和存储系统租赁费		
5	信息安全系统工程费		
6	综合布线系统工程费		
7	机房工程费		
8	系统集成费		
9	硬件购置费		
10	其他工程费		
二	**服务费用**	**900**	
1	项目管理费		
2	咨询费		
3	工程监理费	20	
4	第三方测评费		
7	标准规范		
8	数据治理服务费用	720	
9	知识应用平台建设服务费用	160	
三	合 计(一+二)	**900**	

说明：表单内容非每项必填，项目建设单位可根据实际情况选择填写

7.2.2 硬件设备、软件及服务购置概算明细表

单位：万元

项目名称： 厦门市健康医疗大数据治理及知识应用平台项目							
硬件设备、软件及服务购置概算明细表（按系统划分）							
序号	设备及软件名称	主要性能指标	参考品牌及型号	单价	数量	总价	说明
	总计					880	
一	网络系统						
（一）	网络设备						
1							
（二）	网络系统软件						
1							
	小计						
二	数据处理和存储系统						
（一）	服务器设备						
1							
（二）	数据处理软件						
1							
（三）	存储设备						
1							
（四）	存储软件						
1							
	小计						
三	应用支撑系统（指通用工具、产品，通用应用系统）						
1							
	小计						
四	应用系统						
（一）	疾病探索						
1	疾病图谱			10	1	10	
2	疾病特征分布			5	1	5	
3	疾病指标趋势			5	1	5	

（二）	疾病搜索						
1	关键词搜索			5	1	5	
2	高级搜索			5	1	5	
3	条件树搜索			7	1	7	
4	事件搜索			7	1	7	
（三）	搜索结果可视化						
1	搜索结果可视化			10	1	10	
（四）	患者全景视图						
1	时间轴			3	1	3	
2	就诊信息			3	1	3	
3	检验			3	1	3	
4	检查			3	1	3	
5	病理			3	1	3	
6	手术			3	1	3	
7	文书记录			3	1	3	
8	医嘱			3	1	3	
（五）	患者分析						
1	患者分析			8	1	8	
（六）	研究项目管理						
1	设置项目基本信息			4	1	4	
2	患者入组设置			4	1	4	
3	观测指标设置			4	1	4	
4	结构化编辑器			4	1	4	
5	研究对象详情			4	1	4	
6	研究项目数据导出			4	1	4	
（七）	研究项目数据开放与共享						
1	研究项目数据开放与共享			7	1	7	
（八）	在线统计分析						
1	在线统计分析			7	1	7	
（九）	重要疾病领域指标						

	地图						
1	重要疾病领域指标地图			12	1	12	
（十）	知识全库						
1	知识推荐			6	1	6	
2	研究热点趋势			6	1	6	
3	文献			2	1	2	
4	指南共识			2	1	2	
5	临床路径			2	1	2	
6	药品说明书			2	1	2	
7	临床试验			2	1	2	
8	误诊误治			2	1	2	
	小计			**160**		**160**	
五	**信息资源建设（指数据采集、目录、质量及标准的管理工具）**						
1							
	小计						
六	**数据库**						
（一）	数据库服务器						
1							
（二）	数据库软件						
1							
	小计						
七	**终端系统**						
（一）	终端设备						
1							
（二）	终端软件						
1							
	小计						
八	**安全系统**						
（一）	安全设备						
1							
（二）	安全软件						
1							
	小计						

九	**备份系统**						
（一）	备份设备						
1							
（二）	备份软件						
1							
	小计						
十	**标准规范**						
1							
	小计						
十一	**数据治理服务**						
（一）	数据调研服务						
1	调研计划			1	1	1	
2	调研材料准备			1	1	1	
3	详细调研			10	1	10	
4	调研结果编制			3	1	3	
（二）	数据采集服务						
1	数据采集对接			8	1	8	
2	数据抽取分析			18	1	18	
3	数据脱敏			12	1	12	
（三）	智能数据治理服务						
1	数据清洗						
1.1	16 家医院 HIS 系统			16	1	16	
1.2	16 家医院 EMR 系统			12	1	12	
1.3	16 家医院 LIS 系统			12	1	12	
1.4	16 家医院 PACS 系统			12	1	12	
1.5	16 家医院体检系统			12	1	12	
1.6	一套云 HIS 系统			16	1	16	
2	数据关联						
2.1	16 家医院 HIS 系统			16	1	16	
2.2	16 家医院 EMR 系统			12	1	12	
2.3	16 家医院 LIS 系统			12	1	12	
2.4	16 家医院 PACS 系统			12	1	12	
2.5	16 家医院体检系统			12	1	12	
2.6	一套云 HIS 系统			16	1	16	
3	数据映射						
3.1	16 家医院 HIS 系统			14.4	1	14.4	

3.2	16 家医院 EMR 系统			10.8	1	10.8	
3.3	16 家医院 LIS 系统			10.8	1	10.8	
3.4	16 家医院 PACS 系统			10.8	1	10.8	
3.5	16 家医院体检系统			10.8	1	10.8	
3.6	一套云 HIS 系统			14.4	1	14.4	
4	数据归一						
4.1	16 家医院 HIS 系统			18	1	18	
4.2	16 家医院 EMR 系统			13.5	1	13.5	
4.3	16 家医院 LIS 系统			13.5	1	13.5	
4.4	16 家医院 PACS 系统			13.5	1	13.5	
4.5	16 家医院体检系统			13.5	1	13.5	
4.6	一套云 HIS 系统			18	1	18	
5	数据索引						
5.1	16 家医院 HIS 系统			13	1	13	
5.2	16 家医院 EMR 系统			9.75	1	9.75	
5.3	16 家医院 LIS 系统			9.75	1	9.75	
5.4	16 家医院 PACS 系统			9.75	1	9.75	
5.5	16 家医院体检系统			9.75	1	9.75	
5.6	一套云 HIS 系统			13	1	13	
6	数据结构化处理						
6.1	16 家医院 HIS 系统			38	1	38	
6.2	16 家医院 EMR 系统			28.5	1	28.5	
6.3	16 家医院 LIS 系统			28.5	1	28.5	
6.4	16 家医院 PACS 系统			28.5	1	28.5	
6.5	16 家医院体检系统			28.5	1	28.5	
6.6	一套云 HIS 系统			38	1	38	
（四）	数据质控服务						
1	数据质控服务						
1.1	16 家医院 HIS 系统			20	1	20	
1.2	16 家医院 EMR 系统			30	1	30	
1.3	16 家医院 LIS 系统			10	1	10	
1.4	16 家医院 PACS 系统			10	1	10	
1.5	16 家医院体检系统			10	1	10	

1.6	一套云 HIS 系统			10	1	10	
	小计					**780**	

7.2.3 应用系统及服务采购工作量概算明细表

应用系统及服务采购工作量概算明细表								
序号	应用系统或服务名称	工作量核算（人月数）					单价	总价
		需求分析	系统设计	软件开发	系统测试	部署实施		
一	数据治理服务						2	0
（一）	数据调研服务						2	0
1	调研规划	0.2				0.3	2	1
2	调研材料准备	0.2				0.3	2	1
3	详细调研	1				4	2	10
4	调研结果编制	0.5				1	2	3
（二）	数据采集服务						2	0
1	数据采集对接	0.5	0.5	1	0.5	1.5	2	8
2	数据抽取分析	2	1	2	0.5	3.5	2	18
3	数据脱敏	2	1	1	0.5	1.5	2	12
（三）	智能数据治理服务							0
1	数据清洗服务	2	1	5	1	31		0
1.1	16 家医院 HIS 系统	0.4	0.2	1	0.2	6.2	2	16
1.2	16 家医院 EMR 系统	0.3	0.15	0.75	0.15	4.65	2	12
1.3	16 家医院 LIS 系统	0.3	0.15	0.75	0.15	4.65	2	12
1.4	16 家医院 PACS 系统	0.3	0.15	0.75	0.15	4.65	2	12
1.5	16 家医院体检系统	0.3	0.15	0.75	0.15	4.65	2	12
1.6	1 套云 HIS 系统	0.4	0.2	1	0.2	6.2	2	16

2	数据关联服务	2	1	5	1	31		0
2.1	16 家医院 HIS 系统	0.4	0.2	1	0.2	6.2	2	16
2.2	16 家医院 EMR 系统	0.3	0.15	0.75	0.15	4.65	2	12
2.3	16 家医院 LIS 系统	0.3	0.15	0.75	0.15	4.65	2	12
2.4	16 家医院 PACS 系统	0.3	0.15	0.75	0.15	4.65	2	12
2.5	16 家医院体检系统	0.3	0.15	0.75	0.15	4.65	2	12
2.6	1 套云 HIS 系统	0.4	0.2	1	0.2	6.2	2	16
3	数据映射服务	2	1	4	1	28		0
3.1	16 家医院 HIS 系统	0.4	0.2	0.8	0.2	5.6	2	14.4
3.2	16 家医院 EMR 系统	0.3	0.15	0.6	0.15	4.2	2	10.8
3.3	16 家医院 LIS 系统	0.3	0.15	0.6	0.15	4.2	2	10.8
3.4	16 家医院 PACS 系统	0.3	0.15	0.6	0.15	4.2	2	10.8
3.5	16 家医院体检系统	0.3	0.15	0.6	0.15	4.2	2	10.8
3.6	1 套云 HIS 系统	0.4	0.2	0.8	0.2	5.6	2	14.4
4	数据归一服务	3	2	5	2	33		0
4.1	16 家医院 HIS 系统	0.6	0.4	1	0.4	6.6	2	18
4.2	16 家医院 EMR 系统	0.45	0.3	0.75	0.3	4.95	2	13.5
4.3	16 家医院 LIS 系统	0.45	0.3	0.75	0.3	4.95	2	13.5
4.4	16 家医院 PACS 系统	0.45	0.3	0.75	0.3	4.95	2	13.5
4.5	16 家医院体检系统	0.45	0.3	0.75	0.3	4.95	2	13.5
4.6	1 套云 HIS 系统	0.6	0.4	1	0.4	6.6	2	18
5	数据索引服务	2	1	4	1.5	24		0
5.1	16 家医院 HIS 系统	0.4	0.2	0.8	0.3	4.8	2	13
5.2	16 家医院 EMR 系统	0.3	0.15	0.6	0.225	3.6	2	9.75

5.3	16 家医院 LIS 系统	0.3	0.15	0.6	0.225	3.6	2	9.75
5.4	16 家医院 PACS 系统	0.3	0.15	0.6	0.225	3.6	2	9.75
5.5	16 家医院体检系统	0.3	0.15	0.6	0.225	3.6	2	9.75
5.6	1 套云 HIS 系统	0.4	0.2	0.8	0.3	4.8	2	13
6	非结构化数据处理服务	7	4	12	3	69		0
6.1	16 家医院 HIS 系统	1.4	0.8	2.4	0.6	13.8	2	38
6.2	16 家医院 EMR 系统	1.05	0.6	1.8	0.45	10.35	2	28.5
6.3	16 家医院 LIS 系统	1.05	0.6	1.8	0.45	10.35	2	28.5
6.4	16 家医院 PACS 系统	1.05	0.6	1.8	0.45	10.35	2	28.5
6.5	16 家医院体检系统	1.05	0.6	1.8	0.45	10.35	2	28.5
6.6	1 套云 HIS 系统	1.4	0.8	2.4	0.6	13.8	2	38
（三）	数据质控服务						2	0
1	数据质控服务	4	2	8	2	29	2	90
	小计						2	720
二	健康医疗大数据知识应用系统						2	0
（一）	疾病探索						2	0
1	疾病图谱	0.3	0.7	3.5	0.4	0.1	2	10
2	疾病特征分布	0.3	0.5	1.2	0.4	0.1	2	5
3	疾病指标趋势	0.3	0.5	1.2	0.4	0.1	2	5
（二）	疾病搜索						2	0
1	关键词搜索	0.3	0.5	1.2	0.4	0.1	2	5
2	高级搜索	0.3	0.5	1.2	0.4	0.1	2	5
3	条件树搜索	0.3	0.7	2	0.4	0.1	2	7
4	事件搜索	0.3	0.7	2	0.4	0.1	2	7
（三）	搜索结果可视化						2	0

1	搜索结果可视化	0.5	1	2.5	0.5	0.5	2	10
（四）	患者全景视图						2	0
1	时间轴	0.2	0.2	0.8	0.2	0.1	2	3
2	就诊信息	0.2	0.2	0.8	0.2	0.1	2	3
3	检验	0.2	0.2	0.8	0.2	0.1	2	3
4	检查	0.2	0.2	0.8	0.2	0.1	2	3
5	病理	0.2	0.2	0.8	0.2	0.1	2	3
6	手术	0.2	0.2	0.8	0.2	0.1	2	3
7	文书记录	0.2	0.2	0.8	0.2	0.1	2	3
8	医嘱	0.2	0.2	0.8	0.2	0.1	2	3
（五）	患者分析						2	0
1	患者分析	0.5	1	2	0.4	0.1	2	8
（六）	研究项目管理						2	0
1	设置项目基本信息	0.2	0.5	1	0.2	0.1	2	4
2	患者入组设置	0.2	0.5	1	0.2	0.1	2	4
3	观测指标设置	0.2	0.5	1	0.2	0.1	2	4
4	结构化编辑器	0.2	0.5	1	0.2	0.1	2	4
5	研究对象详情	0.2	0.5	1	0.2	0.1	2	4
6	研究项目数据导出	0.2	0.5	1	0.2	0.1	2	4
（七）	研究项目数据开放与共享						2	0
1	研究项目数据开放与共享	0.5	0.5	2	0.4	0.1	2	7
（八）	在线统计分析						2	0
1	在线统计分析	0.5	0.5	2	0.4	0.1	2	7
（九）	重要疾病领域指标地图						2	0
1	重要疾病领域指标地图	1	0.5	4	0.4	0.1	2	12
（十）	知识全库						2	0
1	知识推荐	0.3	0.2	2	0.4	0.1	2	6
2	研究热点趋势	0.3	0.2	2	0.4	0.1	2	6

3	文献	0.1	0.1	0.6	0.1	0.1	2	2
4	指南共识	0.1	0.1	0.6	0.1	0.1	2	2
5	临床路径	0.1	0.1	0.6	0.1	0.1	2	2
6	药品说明书	0.1	0.1	0.6	0.1	0.1	2	2
7	临床试验	0.1	0.1	0.6	0.1	0.1	2	2
8	误诊误治	0.1	0.1	0.6	0.1	0.1	2	2
	小计						2	160
	总计						2	880

第8章　效益分析

8.1 经济效益分析

1、 避免重复建设，减少数据冗余和资源浪费

本项目基于之前已经统一完成数据采集工作的区域医疗信息化平台的数据进行数据治理，而无需在建设大数据平台时重新从各医疗机构进行数据采集，实现卫生信息化的统一建设、统一管理、全方位服务，加强信息资源有效利用，避免重复建设和资金、资源浪费，提高政府投资效益。本项目从整体上对公共卫生信息化工作进行统筹规划，分步实施，减少了各医疗卫生单位自行搭建信息化平台带来的重复建设。

2、 助力临床医学研究，促进科研成果转化

利用健康医疗大数据治理及知识应用平台的数据资源的业务应用体系，改善医学研究工作流程，提高临床试验效率，加速成果转化。以平台的建设为切入点，从数据治理及统计分析等方面，探讨其在临床研究尤其是真实世界研究中的应用效果。利用自然语言处理、机器学习等人工智能技术，深度挖掘临床研究中的数据内在价值，多层次、多角度满足不同医学研究需求，有着广阔的应用前景。

3、 探索健康医疗大数据共享服务，带动区域医疗健康大产业协同发展

通过本项目的建设，在健康医疗大数据科研应用的基础上，在后期逐步探索健康医疗大数据面向科研之外的更多应用场景，如新药研发、保险设计、健康消费品研发、健康管理服务、健康养老服务等，促进区域医疗健康大产业的数据共享和产业协同发展，产生价值巨大的间接经济效益。

8.2 社会效益分析

项目建成后，将进一步完善厦门市医疗健康服务体系，加强疾病数据管理能力，提升临床医学研究水平，切实提高临床诊疗服务和临床科研服务能力，促进数据资源的合理分配，最终为打造“健康中国厦门样本”提供优质高效的卫生健

康服务。

1、 有利于推动区域健康医疗大数据中心的建设和发展

本项目在区域医疗信息化平台数据采集的基础上利用人工智能和大数据技术对数据进行智能化、标准化处理，为未来面向更多业务场景的健康医疗挖掘与分析提供数据基础，形成区域内高可用的健康医疗数据，有力推动区域健康医疗大数据中心的建设和发展。

2、 有利于政府公共卫生决策和惠民利民

依托本项目建设，进行健康医疗大数据清洗、挖掘和分析，除数据用于临床科研应用之外，数据本身以及临床科研成果数据都可以为政府提供公共卫生技术支撑和决策支持，同时能够为公众提供更加高效、科学、个性化的医疗健康服务，提升居民医疗健康服务获得感。

3、 有利于区域健康医疗大数据建设和应用水平保持国家领先水平

通过区域健康医疗大数据平台建设，以及对数据的广泛治理和高效利用，能够促进各医疗卫生机构信息化建设的进度，统一和规范各医疗卫生机构信息化建设的标准。全面提升本区域医疗卫生信息标准化能力，促进提高整个区域医疗卫生信息化水平提升，是确保本区域医疗信息化建设水平处于全国领先行列的有效手段。

8.3 共享资源分析

通过本项目的建设，区域海量真实的患者诊疗数据得到有效整合和标准化治理，从而极大提升了这些数据的整体利用价值，除了数据可供个人、临床医生查看诊疗数据资料外，还可以将经过脱敏处理后的数据在有效管理的范围内开放给科研机构、医药企业、医疗保险公司、健康服务公司等医疗健康产业相关参与者，促进区域健康医疗大数据的互通、共享。

第9章　项目风险与风险管理

9.1 风险识别和分析

本项目的风险包括组织风险、管理风险、业务风险和技术风险等。

9.1.1 组织风险

组织风险主要包括由于组织内部成员对目标未达成一致，管理高层对项目不重视，工程参与人员知识与技能欠缺、团队合作精神不足、人员激励机制不当等因素导致建设队伍不稳定，建设资金不足，与其它项目存在资源冲突等。

9.1.2 管理风险

管理风险主要包括项目管理的基本原则使用不当，计划草率、质量差，进度和资源配置不合理等。

9.1.3 业务风险

业务变化可能产生的风险主要包括业务流程的改变、职能部门的调整等。

9.1.4 技术风险

技术风险主要包括技术目标过高，技术标准发生变化，复杂、高新技术或非常规方法应用的潜在问题等。

9.2 风险对策和管理

为确保工程成功，将在本项目建设中采取有效的风险管理，消除各类风险的不良影响，确保实现工程建设目标。

本项目的风险防范主要侧重于组织风险防范、管理风险防范、业务风险防范

和技术风险防范四个方面。

9.2.1 组织风险防范对策

厦门市健康医疗大数据中心现有行政组织架构能够支撑本项目的项目管理，并为本项目建设成立了领导小组为核心的决策机制，能够有效地保障本项目高效的项目管理。现有的业务、技术力量能够承担本工程建设任务，同时还将依靠由信息技术专家组成的专家组，为本项目的建设的提供技术与管理问题的咨询和指导。

9.2.2 管理风险防范对策

为确保工程管理的高效率，将对项目进行有效策划，制定并落实严格的项目实施具体计划，应用先进管理工具和方法提高进度计划管理、跟踪水平。同时将借鉴行业项目管理实践的经验，合理估算项目工作量，明确项目间依赖关系和先后顺序，突出关键项目，进一步分解项目工作任务，使每个里程碑阶段均应有工作量估算、时间进度、以及可操作、可管理和可检查的交付物。

在工程建设过程中，根据工作需要为工程决策层和管理层人员举办项目管理培训，分级确定专职工程负责人并充分授权，同时聘请部分经验丰富的外部人力资源协助工程管理。切实加强对各业务处室技术人员的知识和能力培训，稳定骨干人员，增强团队的凝聚力。

9.2.3 业务风险防范对策

在本项目应用系统部署之前，将对厦门市健康医疗大数据治理系统的业务流程进行优化和规范；在业务流程设计的过程中，将进行业务差异分析，充分了解各类用户的业务需要，设计某些可灵活配置的业务流程，满足各类用户的特殊需求。

9.2.4 技术风险防范对策

全面落实信息系统安全体系，尽快全面组织落实本项目安全体系方案。对本项目各应用系统，强制要求按照安全设计方案的要求，采用身份认证，对重要数据进行加密、签名，加强安全记录和审计。