基于 Hadoop 架构的医疗大数据平台应用实践和思考*

王红迁 汪 鹏 王 飞 罗 浩

（第三军医大学西南医院 重庆 400038）

〔摘要〕 介绍国内外医疗大数据战略规划、布局及 Hadoop 技术体系架构，根据西南医院实际情况，阐述构建医疗大数据平台的必要性，分析其架构设计和医疗大数据业务开展情况，指出医疗大数据业务的不足和未来前景。

〔关键词〕 医疗大数据；Hadoop；大数据平台

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.09.005

Application Practice and Thoughts of Medical Big Data Platform Based on Hadoop Architecture

WANG Hong－qian, WANG Peng, WANG Fei, LUO Hao, Southwest Hospital, Chongqing 400038, China

〔**Abstract**〕 The paper introduces the strategic planning and layout as well as Hadoop technology system architecture of medical big data at home and abroad, states the necessity of building medical big data platform and the situation of architecture design and medical big data business development according to the status of Southwest Hospital, and points out the deficiencies and prospect of the medical big data business.

〔**Keywords**〕 Medical big data; Hadoop; Big data platform

1 引言

伴随着全球医疗信息化的进步，医疗数据规模正以指数形式迅速增长。规模巨大的临床病历与健康档案、基础知识库、临床诊疗知识库、参考文献、基因数据以及个人健康数据汇聚在一起形成了医疗健康大数据，数据类型正向复杂、多样、海量、时效性的数据类型方式转变[1]。国内外的相关机构逐步认识到将医疗场景数据和大数据相关技术深度结合可对数据有新的认识，引入新的使用模式，进一步达到数据价值的最大化。在医疗信息化需求多样性、计算机信息化技术不断革新、相关资本涌入等多因素驱动下，全世界各地的医疗信息化产业正在发生翻天覆地的变化[2]。

2 国内外医疗大数据规划及布局

2.1 国外医疗大数据发展现状

目前很多国家都发布相应的法律法规，从政策角度大力推进医疗大数据技术的进步和产业的落地。如 2015 年美国奥巴马政府公布了精准医疗方

〔修回日期〕 2017－06－19

〔作者简介〕 王红迁，硕士，中级工程师，发表论文 2 篇；通讯作者：汪鹏。

〔基金项目〕 重庆市社会民生项目（项目编号：cstc2015shmszx120025）。

案，致力于使公众及时获取个性化健康信息；2014 年欧盟逐步制定关于数据价值链策略方案，同时在 2015 年开展“地平线 2020”科研规划，推动大数据在能源、制造业和医疗行业应用；2013 年英国政府实施数据能力发展战略布局[3]，2015 年又完成国民健康服务体系，完善英国国民医疗服务[4]。不仅如此，国外的一些科研机构和公司也大力布局医疗大数据产业，并且取得了一些可喜的成果。如美国公共健康协会通过使用 FluNearYou 成功预测流感疫情爆发；美国国立卫生机构建立生物医学大数据中心，核心目的就是从生物医学中获取海量数据并提炼其价值，从而增进人类对疾病的把控；IBM 公司也推出 Watson 认知系统，该系统进行个性化定制的健康指导、医疗数据预测分析和图形分析、医生最佳医疗方案；Google 公司发布了许多医疗数据应用的产品，基于大数据分析的跟踪工具 FluView、健康管理平台 Google Fit、医疗健康应用 Study Kit；微软公司同样也发布许多医疗数据成功应用的产品，健康管理平台 Microsoft Health、个人健康管理平台 HealthVault、针对医疗人员开发的信息系统 Amalga UIS。

2.2 我国医疗大数据发展现状

2.2.1 *相关政策* 目前，我国政府高度重视医疗大数据的研发与应用，2016 年 8 月 19 日习近平总书记在全国卫生与健康大会上发布关于健康的重要讲话，2016 年 8 月 26 日中共中央政治局审议并发布“健康中国 2030”全局纲要。同年国务院“十三五”布局规划也颁布“实施国家大数据战略”的目标和《国务院办公厅关于促进和规范健康医疗大数据应用发展指导意见》。政策从国务院角度发布，深刻表明了国家对医疗大数据的重视程度。

2.2.2 *相关研究进展* 目前，我国一些科研机构和医疗机构都逐步开展医疗大数据的研究，目的是顺应当前医疗信息化的发展形势，将大量沉淀的医疗数据进行有效充分的利用，从而实现健康中国的长远目标。如由复旦大学及相关的各个附属医院组建上海精准医疗大数据中心，已深入展开对生物数据的研究和实验，而且根据研究成果建立食管癌疾病的临床信息辅佐决策系统，研发精准医学临床辅助服务决策系统；由复旦大学附属中山医院和华大基因联合成立中华精准医学中心，已通过基因大数据，建立在多领域具有国际领先水平的基因组学科研和应用转化中心；贵州省建立了省级大数据精准医学实验室，针对医疗影像数据，开发研制影像学人工智能医疗服务体系；清华大学数据科学研究院下属专门成立医疗健康大数据研究中心，开展整合海量医疗健康大数据，为居民、医生、政府管理提供辅助支持，采取对临床数据和基因数据整合分析，从而完成对高危疾病精准预防、诊断与治疗[4]。

3 Hadoop 技术构成

3.1 概述

目前，医疗行业大部分机构都选择基于 Hadoop 技术扩展和封装的医疗大数据平台，一个特别重要的原因是 Hadoop 开源[5]，可解决数据多源异构的问题，具备高可靠性、高扩展性、高效性、高容错性的口碑，能够降低成本的同时确保技术的可靠与发展的延续，同时 Hadoop 体系下的技术成熟且全面，可很好地满足相应业务的技术要求。

3.2 数据收集

Hadoop 架构下常用的数据收集技术有 Flume、Sqoop、ETL 等。其中 Flume 是一种可定制化的收集数据，具有高可靠、高可用并且支持分布式的数据采集传输系统；Sqoop 可以很方便地将关系型数据库中数据导出到 Hadoop 分布式文件系统（Hodoop Distribute Flie System，HDFS），反之也是很方便；Kettle 是一种开源的 ETL 工具，支持多源数据库并且具有图形化的界面，可高效的提供实时数据流支持。

3.3 数据存储

Hadoop 架构下常常运用的数据存储技术有 HDFS、Hbase 与 MongoDb 等。其中，HDFS 是分布式文件存储系统，支持多类型数据，具有高可用性；Hbase 是列式分布式数据库，底层也是用 HDFS

存储数据；Mongodb 面向集合的文件存储数据库，支持的数据结构松散，具有面向对象的查询特点。

3.4 数据分析

Hadoop 架构下，经常使用处理数据的技术有 MapReduce、Spark、Storm、Hive、Pig 等。其中，MapReduce 为一种分布式计算框架，主要提供实时性要求不太高的数据处理场景；Storm 是一种流式处理系统，目前逐渐被 Spark 取代；Spark 是对 MapReduce 的优化，启动内存分布式数据集，可提供实时性和离线性数据处理；Hive 和 Pig 可通过类 SQL 语句让大数据查询变得更易用。

3.5 机器学习

Hadoop 架构下常用的机器学习技术有 Mahout 等。其中 Mahout 已整合多种机器学习的算法，可在分布式集群上部署使用，简化机器学习的入门难度，提升机器学习的性能。

3.6 资源调度

Hadoop 架构下常用的资源调度技术有 Mesos、Yarn、Docker 等。其中 Yarn 与 Mesos 是新一代资源管理框架，高效支持 Spark 和 Docker 等服务；Docker 容器化为服务高效稳定运行提供保证，同时为服务快速扩展提供保障。

3.7 数据检索

通用的检索技术有 SolrCloud、Elasticsearch。都可提供高并发大容量分布式检索及过滤，其中 Elasticsearch 相对更轻量级、实时性更好。

3.8 数据缓存

常用的数据缓存，可以用缓存类数据库，如 Redis，也可以安全可靠的消息队列，如 Kafka、RMQ 等。

3.9 集群监控

常用的集群监控技术有 Ganglia、Nagios、ClouderManger 等，其中 Ganglia 是一个跨平台可扩展的，高性能计算系统下的分布式监控系统，但是没有报警机制，出现问题不能够及时报警；Nagios 最大的特点是其强大的管理中心，所有的监控、报警功能都是由相关插件完成的；ClouderManger 是非开源的监控工具，安装操作简单。

4 构建医疗大数据平台的必要性

西南医院作为全国知名的三甲医院，经过 20 多年的数字化医院建设，信息系统已经积累了大量的医疗数据，目前已有 400TB 的临床数据[6]，其中包括 3 723 多万条医疗就诊记录、249 多万份电子病历文档资料、58 多万份标准化存储的生物样本、110 多万份与沙区共享的居民电子健康档案等。目前医疗数据量呈现井喷的趋势，但还缺乏对这些海量医疗数据有效深度的利用。实际上目前我国很多医院现有的信息化技术和手段已经难以满足医疗领域中的智能高效知识需求，无法实现统一的全方位的医疗辅助知识的获取，为此大力研发基于医疗大数据的实践应用是十分必要的，并且大力发展医疗大数据应用是迈向精准医疗的关键技术。总而言之，实现基于医疗大数据的应用，是实现医疗活动从经验科学向数据科学的发展，深化医疗改革，实现健康中国的重要一环。结合开展医疗大数据平台建设过程中的实践经验，从架构、进展以及未来的改进之处等方面进行逐步论述。

5 医疗大数据平台架构设计

5.1 硬件

医疗大数据平台的建设不仅考虑目前局部的需求和功能，而且从扩展性、统一性、标准性、安全性等角度进行统一设计和实施，实现该平台与医院现有各业务数据库无缝衔接，医院所有相关的医疗科研活动以及后续需求都可以使用该平台，从而达到更低成本、更高灵活性。图 1 所示是医疗大数据平台的硬件整体架构设计，硬件架构包括在线数据服务器集群、离线数据服务器集群、业务服务器集群，此外还包括流量控制服务器、防火墙服务器、

VPN 服务器、堡垒机等，这些服务器集群架构达到数据可相对安全控制的应用。服务器区是数据处理的核心，完成具体医疗大数据业务；架构中配置多个防火墙，将互联网区与医院内网区隔离，同时医院内网区内部也部署多个防火墙，将不同功能区隔离，进一步保证各个业务区的相对独立安全；架构中部署网闸，用来保证军网数据安全以及医院内网中数据不外泄；部署堡垒机，保证所以进入医院内网的人的任何行为都有记录，对远端及本地操作者进行审计；部署流量控制器，也是为了保证数据的安全不泄露；部署漏洞扫描系统，主要帮助管理员发现漏洞威胁、及时了解全网安全状态；部署日志审计系统，具备范式化功能，对异构日志格式的统一化，见图 1。

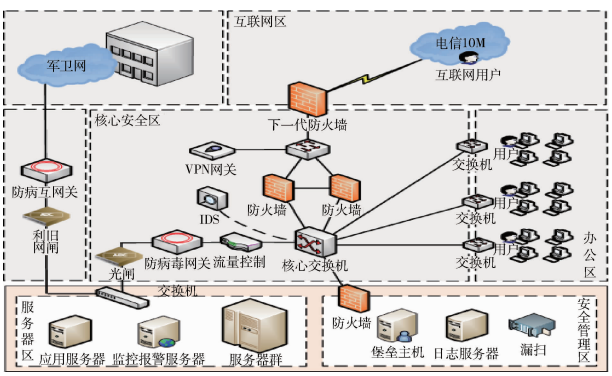

图 1 硬件整体架构

5.2 软件

医疗大数据平台的软件整体架构，基于主流的云计算及大数据技术，采用 Hadoop2.0 技术框架及 Spark 并行计算框架，以及业界领先 Docker 容器化封装技术。平台具体采用 Kettle 和 Sqoop 用于数据收集、用 HDFS 分布式文件系统用于存储原始数据、用 MapReduce 和 Hive 和 Spark 用来分析数据、用 Yarn 和 Docker 用来资源调度和资源封装、用内存数据库 Redis 用来数据缓存、Mongodb 用来存储分析后需快速检索的数据、Elasticsearch 用来构建分布式检索系统、Mahout 用来机器学习实现聚类分类、采用 Ganglia 和 Nagios 监控集群本身性能。平台将解决数据收集、数据存储、数据清洗、数据展示等方面的工作，医院原有的医疗信息子系统的数据通过 ETL 实时导入到大数据集群，这样就为后续的应用提供了原始数据支撑，目前整个系统已经完成医院信息系统和电子病历系统数据的处理，下一步是将所有数据和相关医疗文献等数据导入大数据集群。目前医疗大数据平台中，原始数据在 HDFS 中容灾保存，对采用 MapReduce 和 Hive 和 Spark 等框架这些原始数据进行各种处理和规则的建立，预处理后的数据根据使用的目的，分别存储在 HDFS 和 Mongodb，一方面可保证各种实时性能高的业务需求，另一方面也保证其他的各种性能要求一般的业务需求，为以后系统接入各种业务提供了扩容性。同时整个系统引入 Docker 虚拟化技术，提高机器资源的利用率，降低用户使用成本。除此之外，采用私有云的架构设计，进一步保证内部数据的安全性能。同时整个架构引入许多额外的保障系统，如集群内部安全认证体系、系统管理部署体系等，这样也进一步保证系统数据的安全，同时方便整个系统的运维。该平台基本功能包括：ETL、医疗数据结构化处理、医疗数据关联规则建立、医疗数据清洗归一、医疗数据脱密、医疗数据挖掘分析、索引构建、机器学习、智能推荐等。软件整体架构，见图 2。

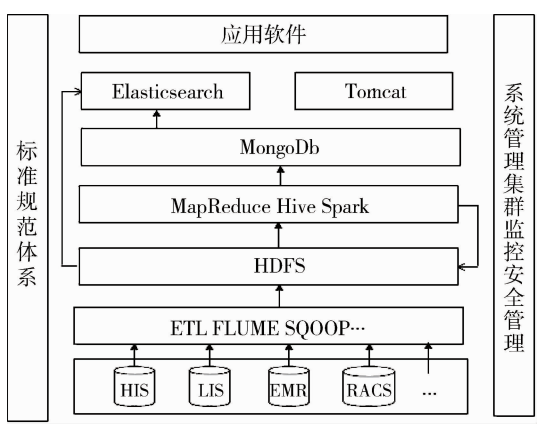

图 2 软件整体架构

6 医疗大数据业务系统进展

6.1 概述

立足于该医疗大数据平台已经开展很多工作，取得部分成果。目前已基于医疗大数据开发出为医生服务的 3 大知识服务系统，使医生既能单独使用

这些系统，又能实现与日常工作信息系统的无缝融合，使之在诊疗过程中获得强大的智能知识外挂，为诊疗活动提供参考借鉴。

6.2 多维度、细粒度的全景式医疗大数据搜索系统

该架构是运用 Elasticsearch 技术框架对关键词匹配与搜索，设计实现可跨数据域和异构数据等高效检索，整个架构也具备多条件组合关联等搜索，通过进行数据领域、范围的权限分配，使得系统管理者可控制用户的权限，保障数据的安全。系统提供通过对搜索的结构进行各种维度的统计，生成图表。同时，给出管理历史和常用的搜索条件表达式，提高用户的搜索效率。

6.3 临床科研信息系统

该系统是迄今为止的所有临床活动数据为核心，通过该系统可以满足临床医生对临床数据的分类、汇总，提供简单的分析。该系统目前以患者全生命周期数据为主线，建立了 4 个管理系统，包含数据的收集、元数据的维护、对象的统一管理、采集数据形成病种库的维护[7]。该应用软件为临床医生和科研人员进行科研活动提供数据支持。

6.4 基于智能化临床循证知识推送的诊疗决策支持信息系统

该软件系统在充分分析现有的医疗数据的基础上，计算为临床医生进行医疗诊断的时候进行知识发现与推送，从而给出全方位多角度的提醒，帮助医生的临床诊断，减少误诊率。该应用系统已成为医生诊疗过程中必不可少的辅佐平台，使其成为资深专家经验的传承平台，年轻医生学习提高的平台[7]。

7 结语

目前医疗大数据平台的建设和应用刚步入正轨，有很多工作还需要完善和开展新的应用系统的研发。（1）医疗大数据平台接入的医疗数据还不全面，目前正在积极将基础知识库、临床诊疗知识库、参考文献、从历史病例挖掘形成的知识、基因数据、患者穿戴数据等全面接入该系统。（2）目前基于该平台的应用开发还是比较少，接下来进一步从医生辅助支持、患者健康管理、医院管理等角度做深度的研究及应用。（3）目前基于该平台实现的专病管理还比较少，下一步可以从一些常见的慢性病入手，逐步建立基于大数据平台的各种专病管理体系。（4）目前正在积极与华大基因等机构合作，目的是基于该平台实现相关基因数据的应用开发和管理。（5）正在完善医疗大数据系统中安全系统的建立，为区域医疗大数据资源共享做准备。（6）积极参与重庆市的医疗大数据产业的重点工程的建设，包括健康医疗大数据应用基础平台、生物医学大数据中心、医疗个性化服务基础平台、医疗卫生管理与服务应用、健康医疗大数据保障机制等。

参考文献

1 王震寰. 计算医学——应对大数据的挑战向临床转化［J］. 蚌埠医学院学报, 2014, 39（1）: 1 - 2.

2 Murdoch T B, Detsky A S. The Inevitable Application of Big Data to Health Care［J］. JAMA, 2013, 309（13）: 1351 - 1352.

3 汪鹏, 吴昊, 罗阳, 等. 医疗大数据应用需求分析与平台建设构想［J］. 中国医院管理, 2015, 35（6）: 40 - 42.

4 陈敏, 刘宁. 医疗健康大数据发展现状研究［J］. 中国医院管理, 2017, 37（2）: 46 - 48.

5 王玉龙, 曾梦岐. 面向 Hadoop 架构的大数据安全研究［J］. 信息安全与通信保密, 2014,（7）: 83 - 86.

6 汪鹏, 王飞, 王毅琳, 等. 医疗大数据临床应用的探索与实践［J］. 中国数字医学, 2016, 11（9）: 8 - 10.

7 王建强, 仲晓伟, 杨飞. 数据挖掘在医疗临床路径中的应用［J］. 现代医院, 2011, 11（3）: 1 - 3.