第 8 卷 第 5 期
2019 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 8 No. 5
Sep. 2019

引文格式：
王持, 李超, 陈旭, 等. 面向医疗临床科研的大数据平台 [J]. 集成技术, 2019, 8(5): 86-96.
Wang C, Li C, Chen X, et al. Big data platform for clinical scientific research [J]. Journal of Integration Technology, 2019, 8(5): 86-96.

面向医疗临床科研的大数据平台

王 持[1,2] 李 超[1] 陈 旭[2] 洪 平[2] 郑文立[1] 沈 耀[1]
齐开悦[1] 过敏意[1]

[1](上海交通大学电子信息与电气工程学院 上海 200240)
[2](卫宁健康科技集团股份有限公司人工智能实验室 上海 200072)

摘 要 目前我国医疗信息化建设已有一定历史，各医院积累了大量电子病历临床数据，但数据结构多样。如何利用这些数据以辅助临床诊疗、科研、节约医疗资源、提升医疗效率和医疗质量，成为各医疗科研机构普遍关注的问题。该文提出了一种面向临床科研的大数据平台，构建多源数据采集方式解决信息基础设施不一致的问题：统一化存储方式应对不同数据类型、分布式计算平台提升效率与可拓展性，并对敏感数据去隐私处理；同时，构建临床科研平台辅助临床科研人员进行科研分析。根据架构搭建集群，在专病分析流程上将原本人工约 4 个月的工作简化到 15 秒左右；数据处理效率方面，由已有平台的 5 天导入 16 692 条数据提升到 10 分钟导入 15 217 026 条数据，速度与数量有了显著提升。该平台有助于完成临床数据采集，建立专病数据库、临床科研、辅助临床诊疗的闭环，为临床科研提供高效一体化的数据平台支持。

关键词 大数据；临床科研；数据安全
中图分类号 TP 392 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20190729004

Big Data Platform for Clinical Scientific Research

WANG Chi[1,2] LI Chao[1] CHEN Xu[2] HONG Ping[2] ZHENG Wenli[1] SHEN Yao[1]
QI Kaiyue[1] GUO Minyi[1]

[1](*School of Electronic Information and Electrical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)
[2](*Winning Artificial Intelligence Research*, *Winning Health Technology Group Co.*, *Ltd.*, *Shanghai* 200072, *China*)

Abstract By far, China's medical informatization construction has been a while. Each hospital has accumulated

收稿日期：2019-07-29 **修回日期**：2019-08-26
作者简介：王持，硕士研究生，研究方向为并行计算和机器学习；李超(通讯作者)，特别研究员，研究方向为计算机体系结构和系统设计，E-mail：lichao@cs.sjtu.edu.cn；陈旭，博士，研究方向为医疗影像与大数据平台；洪平，硕士，研究方向为医疗大数据平台和临床科研平台；郑文立，特别副研究员，研究方向为分布式资源管理和大数据计算系统；沈耀，副教授，研究方向为云计算分布式系统和深度学习图像处理；齐开悦，副教授，研究方向为机器学习与计算机视觉；过敏意，教授，研究方向为并行分布式计算和并行编译器。

a large amount of electronic medical clinical data, but the data structure is highly diverse. To better assist clinical diagnosis and treatment, research, save medical resources, improve medical efficiency and medical treatment quality has become a common requirement in various medical institutions. This paper proposes a big data platform for clinical research, solving the inconsistency of multi-hospital information infrastructure by constructing multi-source data collection methods, unified data storage methods to cope with different data types, and distributed data computing platforms to improve efficiency and scalability. We construct a clinical research platform to assist clinical researchers in scientific research. According to the proposed architecture, the cluster was simplified to about 15 seconds in the special disease analysis process. The data processing efficiency was compared with the existing platform. The 5 days of time for importing of 16 692 data records is reduced to 10 minutes that we can import 15 217 026 data records, significantly improving speed and quantity. This platform helps complete clinical data collection, establish a special disease database, clinical research, and assist in the closed loop of clinical diagnosis and treatment, providing an efficient and integrated data platform support for clinical research.

Keywords　big data; clinical research; data security

1　引　言

在过去十年中，国家出台了大量关于医疗信息化建设的政策。例如，2011—2012 年间出台了一系列促进医疗机构(如医院、医药厂商等)信息化的政策；2013 年开始出台区域信息化建设的政策；2015 年，《促进大数据发展行动纲要》明确了关于数据使用的总体要求。2016 年 6 月底，国务院出台《关于促进和规范健康医疗大数据应用发展的指导意见》，将医疗大数据正式纳入国家发展，对医疗大数据融合及共享开放建设，在医疗、医药、公共卫生、医保等方面的应用，以及使用安全保障等方面进行了全面规范。以上数据应用政策的释放和推进将促使医疗大数据产业加速形成，包括数据收集、融合、清洗处理到应用环节。

国家统计局提供的数据[1]显示，居民慢病患病率逐步增长，导致医疗服务的需求攀升；而医疗保险基金收入逐年下跌，人均医疗卫生费用占国内生产总值比重在世界范围内属于较低水平。目前，我国区域医疗信息化建设、数据源开放和共享化程度处于较低水平，对医疗大数据相关平台的需求很大。

在医学领域，科学研究基本是建立在基于试验设计的研究方法上，这要求医学科研人员花费大量时间进行实验、控制实验条件、记录实验数据，进而分析实验数据，得出研究结果。传统的科研数据收集方式主要是科研工作人员向医院信息科提交数据申请获取。由于临床数据同时存在结构化和非结构化的数据结果，信息人员通常只能提取局部数据，余下的数据仍需科研工作人员手工收集整理并补齐。随着科学技术的进步，各种工具及技术不断涌现，临床科研工作也开始逐步由手工处理时代向信息化时代转变。

针对以上问题，本文提出了一套用于临床科研的大数据平台。首先，以患者医疗过程相关的各类就诊信息为主，根据疾病构建专病数据中心；其次，利用深度学习技术，分析大量的临床数据特征；最后，结合大数据处理引擎，完成临床数据的快速分析处理。同时，该平台可以帮助临床医生进行数据分析：通过病历检索的方式从大量数据中调阅病例样本并分析病历分布；利用

自然语言处理技术把非结构化的文本病历进行结构化；从海量数据中快速筛选科研要求数据。另外，该平台还能进行数据分析和数据挖掘，实现更多样化的数据展示平台。

本文第 2 节简要分析医疗大数据领域国内外的研究现状；第 3 节介绍用于支撑临床科研服务的大数据平台架构与平台中的关键技术；第 4 节介绍大数据临床科研平台；第 5 节介绍本文搭建的参考平台，以及相应数据处理实验；最后，对所做的工作进行总结。

2 医疗大数据研究现状

对于医疗健康领域，信息化技术是提升医疗效率与质量的重要手段。目前，国内外大多数医院基础信息化系统的构建已经较为完善，但不同医疗机构的信息化构建水平不一，故存在信息存储方式与医疗信息库构建形式不一致的现状。对于应对逐步增长的数据复杂度和数据量问题，引入大数据技术是必经之路。

2.1 国内研究现状

在学术界，涂新莉等[2]给出了一套较为完善的普适化大数据处理流程，并明确指出其适用于医院与科研机构。不同于其他大数据应用领域，陈功等[3]指出医疗大数据的多态性、不完整性、时效性、冗余性、隐私性等特点；颜延等[4]则针对医疗大数据的数据特点进行了具体的系统分析。董诚等[5]对医疗健康大数据的典型应用场景，如公共卫生、药物副作用评估、辅助诊断与个性化治疗等进行了举例分析。黄婧等[6]针对医疗数据的数据安全问题进行分析，综合了适用于医疗信息系统的关键技术，如身份认证、数据隔离、访问控制与安全审计。王强和易应萍[7]提出一种基于观测性医疗成果协同通用数据模型的科研数据分析平台，并在子宫切除分析上进行应用。金昌晓等[8]针对大数据技术在临床科研上的应用价值进行了探讨。文献[9-10]针对专病库的建设提出了相应临床科研平台的架构。

2.2 国外研究现状

Abouelmehdi 等[11]针对大数据场景下医疗数据安全问题分析了传统隐私保护做法的缺陷：反认证、HybrEx[12]、基于身份的匿名化。Lohr 等[13]提出了一种基于可信虚拟域的方法，用以解决大数据场景下多源认证的安全问题。Khaloufi 等[14]提出了一种医疗数据的安全生命周期模型。Gandomi 和 Haider[15]提出了大数据分析在医疗辅助诊断上的应用。Singh 等[16]针对实际数据在临床科研与临床药物研发的作用进行了探讨，指出了大数据技术在精准医疗中的重要价值。文献[17-18]从临床科研人员角度指出了更好地利用大数据技术进行临床医学科研的方式。Mayo 等[19]提出利用大数据技术改善临床研究中随机对照实验的设计方法。

3 大数据平台架构

本节主要介绍大数据平台的基础架构，如图 1 所示，主要分为数据采集、数据存储、数据计算和数据安全 4 部分。接下来将详细介绍各部分使用的关键技术，图 2 展示了平台的细节。

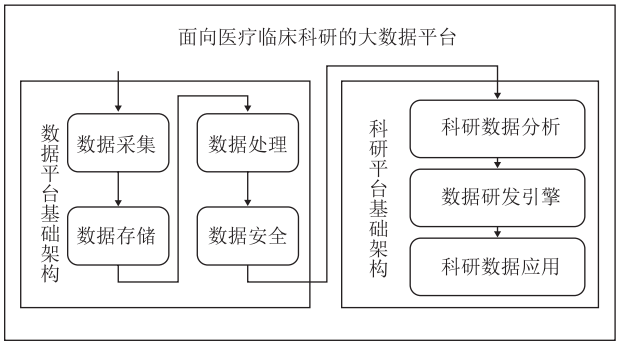

图 1 面向临床科研的大数据平台总架构

Fig. 1 Structure of big data platform for clinical scientific research

3.1 数据采集

医疗大数据平台面向的原始医疗数据不仅数

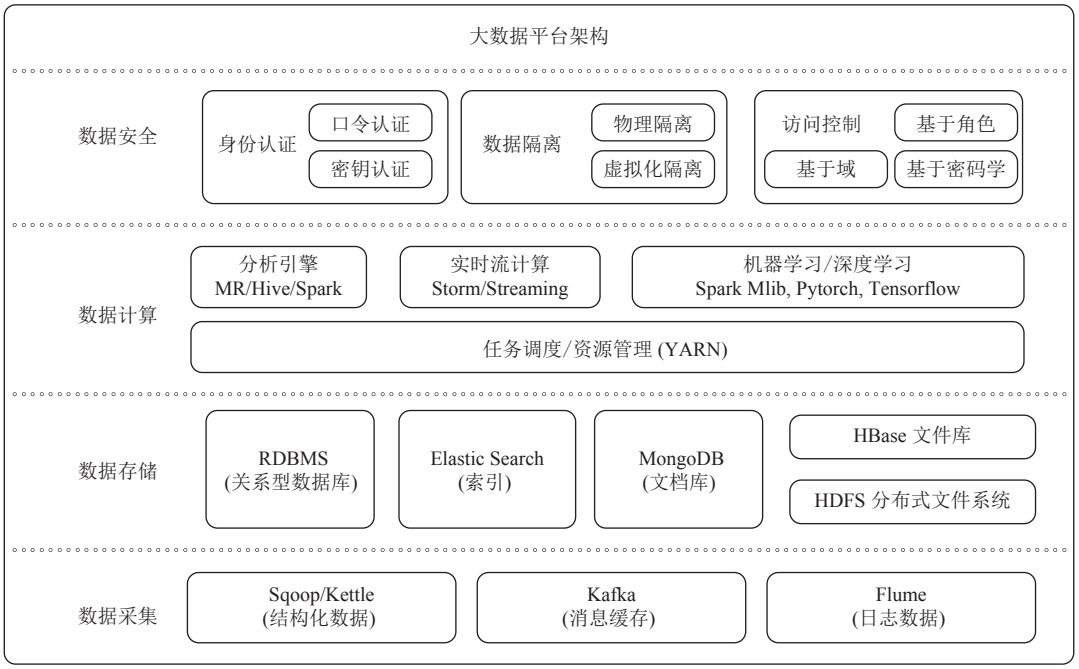

图 2　大数据平台基础架构

Fig. 2　Structure of big data platform

量大，且形式多样，不同医疗机构所使用的数据结构也不同。为正确获取医疗大数据的价值，在数据的清洗处理治理问题上需要建设相应的基础架构以应对其复杂的处理、转换与迁移工作。为应对部分可重复流程，抽取转换装载(Extract-Transform-Load，ETL)工具[20]被广泛采用。本文采用的 ETL 工具有 Sqoop、Kettle、Flume 和 Kafka。

Sqoop 是将 Hadoop 和关系数据库管理系统中的数据互相转移的工具，可以将关系型数据库中的数据导入分布式文件系统中。对于某些非关系型数据库(Not Only Structured Query Language Database，NoSQL)，Sqoop 也提供了连接器，专为大数据批量传输设计，能够分割数据集并创建 Hadoop 任务来处理每个区块。

Kettle 是由元数据驱动(Metadata-Driven)的 ETL 工具。Kettle 流程处理方式为 Transformation 和 Job 两种脚本文件。Transformation 主要用于从数据源进行数据移动与转换工作，而 Job 负责更高级别的流控制，包括执行转换、通知错误和传输文件等。

Flume 是由 Cloudera 提供的高可用、高可靠、分布式的日志采集、聚合和传输的系统，其支持在日志系统中定制数据发送方以收集数据。同时，Flume 能提供简单的数据处理功能，并支持写到各种数据接受方的能力。在本平台中，Flume 用于日志数据记录工作。

Kafka 是一款分布式流处理平台，具有以下能力：(1)以流的形式进行发布与获取数据记录，类似于企业消息系统中所使用的消息队列技术；(2)具有容错能力的流数据记录存储方式；(3)流数据记录自动处理。其特点在于支持 $O(1)$ 复杂度的磁盘数据结构，从而提供了稳定的消息存储性能以及高吞吐量。

大数据平台定义了数据采集范围和数据字段规范。首先，根据专病库采集标准，利用数据同步工具 Kettle、Sqoop 或脚本将增量病历数据同步到数据预处理区域；然后，在数据预处理区对

增量数据做数据清洗和数据质量校验；最后，通过数据分发服务，把采集到的数据分发到不同的疾病专库。同时，利用 Kafka 进行数据采集工作的消息缓存；利用 Flume 做日志记录处理操作。

经过数据清洗、字段校验和表级校验，数据交换错误及数据校验错误会被记录在数据质量管理数据库，通过数据质量组件可以进行分析和展示。最终，通过数据交换监控报告、数据校验监控报告、专题报告等数据分析视图对数据质量进行报告。

3.2 数据存储

因由数据采集部分所采集的多源数据类型多样，故医疗大数据系统中的数据存储子系统应针对数据特性提供所适配的存储方式。常见的数据存储方式主要分为结构化查询语言（Structured Query Language，SQL）结构化与 NoSQL 非结构化[21-22]。针对医疗数据特点，本文集成了 Hive、Elasticsearch、Hbase、MongoDB、SQL Server 等技术。

Hive 是构建于 Hadoop 集群之上的数据仓库应用，可以将存储于分布式文件系统上的结构化数据文件映射作为数据表，并提供完整的 SQL 查询功能，通过将 SQL 语句转换为 MapReduce 任务进行运行。该技术适用于不需要快速响应输出结果的静态历史医疗数据分析。

Elasticsearch 是一个建立在全文搜索引擎 Lucene 基础上的实时分布式搜索和分析引擎。它可快速处理大规模数据，不仅包含全文搜索功能，还可实现分布式实时文件存储。通过 Elasticsearch 进行索引电子病历数据，在医疗系统中可提供基于相似病症的病历检索功能，同时还能支持实时数据统计。

Hbase 可被用来存储海量图片、文档等小文件，具有全局名字空间的优势。用一个单独的列簇存储文件内容，用其他列簇存储文件的类型、大小、创建时间、修改时间等标准属性及应用相关的属性信息，大表能支持文件多属性的实时查询。

MongoDB 是一个基于分布式文件存储，介于关系数据库和非关系数据库之间的数据库产品，适用于存储复杂结构的数据类型。病历文档和后结构化的指标数据由于数据结构不固定，非常适合用其存储，同时能提供属性索引和强大的查询功能。

3.3 数据计算

在数据计算方面，本文主要使用 Hive、Spark 作为分布式数据的分析引擎；Storm 作为实时流计算工具；Spark Mlib 作为分布式机器学习、深度学习框架；Yarn 进行任务调度和管理。

Hive 是基于 Hadoop 文件系统上的数据仓库架构，能为数据仓库管理提供大量功能，如数据存储管理、大型数据查询和分析能力。另外，还可将结构化数据的数据文件转化为一张数据表，并提供简单的查询功能。

Spark 是一个通用内存并行计算框架，主要用来构建大型、低延迟的数据分析应用程序。其扩展了广泛使用的 MapReduce 计算模型，高效支撑更多计算模式，包括交互式查询和流处理，主要特点是能够在内存中进行计算。针对实时性要求高的分析行为，需要 Spark 参与计算。

Storm 是一个分布式实时可靠的、容错的数据流处理系统计算系统。利用 Storm 可以很容易做到可靠地处理无限的数据流，像 Hadoop 批量处理大数据一样，Storm 可以实时处理数据。另外，Storm 会把工作任务委托给不同类型的组件，每个组件负责处理一项简单特定的任务。

Spark Mllib 是 Spark 的机器学习（Machine Learning）库，旨在简化机器学习的工程实践工作，并方便扩展到更大规模。Mllib 由一些通用的学习算法和工具组成，包括分类、回归、聚类、协同过滤和降维等，同时还包括底层的优化原语和高层的管道。

3.4　数据安全

医疗大数据平台所存储的数据内容具有特殊性，且存在各种形式的泄露危险[23]，因而医疗大数据平台的数据安全部分需要着重设计。本文在平台中使用到的数据安全关键技术有身份认证、数据隔离、访问控制、数据去隐私。

3.4.1 身份认证

用于大数据平台的身份认证技术主要包括口令认证和密钥认证[24-26]。当存在数据共享需求时，需要进行相应的身份认证技术保证合法性。

口令认证技术为传统基础认证技术，实现形式简单，但安全性不高，泄露后存在较高的安全风险。密钥认证技术主要分为对称加密与非对称加密方式两类。对称加密技术基本机制为加密、解密的密钥相同，其中较具代表性的算法有数据加密标准、高级加密标准，优点在于速度快，缺点在于其仅适合较小范围的网络。非对称加密技术主要使用数字签名以及公钥基础设施技术，如 RSA 算法(Ron Rivest、Adi Shamir、Leonard Adleman Algorithm)、安全散列算法系列算法。该技术主要优点为其一对多机制，签名者使用签名密钥产生签名，任何人都可使用公开密钥进行验证，同时保证数据源认证和数据完整性、不可否认性，但缺点在于运算代价较大。

3.4.2 数据隔离

本小节主要讨论以下数据隔离技术：物理存储隔离和虚拟化隔离[27-29]。

基于物理存储的隔离方式主要通过物理存储分离的方式进行数据隔离。该方式从物理意义上杜绝了不同用户之间数据存储泄露的危险。但物理隔离方式耗费的实际使用量较多，存在数据利用率不充分的问题。

基于虚拟化的数据隔离技术主要通过虚拟机或容器形式[30]实现。具体实现如下：(1)进程隔离，通过进程标识符使不同用户之间难以获取到对方的进程信息，从而使进程间数据隔离；(2)文件系统隔离，通过挂载不同的文件系统结构以隔离数据；(3)网络隔离，通过将不同用户的存储数据划分到不同的网段进行隔离。

3.4.3 访问控制

本小节主要讨论的技术如下：基于角色的访问控制、基于域的访问控制、基于密码学的访问控制[31-33]。

基于角色的访问控制通过预先设定不同用户相对应的角色，给予相应的访问权限，访问时系统进行权限确认。目前，在权限分配方式上存在基于角色挖掘的访问控制方法[30]，但面向医疗健康大数据仍应采用传统的管理员管理方式以保证安全性。

基于域的访问控制模型主要面向分布式系统间不同域管理者进行协调工作的模型，其基础设计源自基于角色的访问控制[34]。针对医疗健康领域的多机构问题，基于域的访问控制应用较为广泛。

目前，基于密码学的访问控制[35-37]中广泛使用的是基于属性的加密(Attribute-Based Encryption，ABE)算法，其统一使用属性来描述访问控制中的基本概念，如请求者、资源、条件。ABE 算法具体又可分为基于密钥策略的属性加密 KP-ABE[38]以及基于密文策略的属性加密 CP-ABE。针对医疗数据中的特殊共享资源，该方式提供了匿名访问的途径。

3.4.4 数据去隐私

电子病历中包含大量医疗知识，其本身是重要的研究资源，但同样包含了很多隐私信息，如患者个人基本信息、家庭住址、联系方式、经济状况、健康状况和病史等。其中，重点及难点在于对隐私信息的识别。常用方法主要有基于规则、基于机器学习以及规则与机器学习相结合的方法[39]。

4　临床科研平台架构

本文通过之前大数据平台所提供的服务，

提取到患者诊疗相关数据，对其进行转换、加工、清洗，并对其中非结构化数据处理，从而依据不同的疾病类型构建出专病数据中心。同时，通过基于专病的病例搜索，依据不同指标检索，导出表单供研究人员分析。另外，还能基于深度学习、自然语言处理的相似病例服务及药物识别服务等，与医生对接，辅助医生参考历史治疗方案，分析疗效，减轻医师工作量。

根据临床科研人员的需求，本文临床科研平台(图 3)提供了科研数据分析模块，并整合了用于科研人员的数据研发引擎，以便实现临床科研数据的应用。

4.1 科研数据分析

对于从数据平台获取到的清洗后数据，需要针对具体科研应用进行统计学数据分析与可视化分析，以便科研人员能够快速直观地获取数据信息。该流程主要包括连续型/离散型变量数值统计、异常值分析、数据可视化分析。

连续型/离散型变量数值统计是利用统计学方法进行描述所使用的变量，具体为计算数据的分布类型，参数分布情况(均值、方差、四分位数、梯度、散度、旋度等)，期望离群值数量(置信度区间等)。

异常值分析在科研数据分析中具有挑选可利用变量的价值。目前有多种异常值分析方法，包含基础分析缺失程度、基于距离的异常值分析方法[40]、基于可视化的异常值分析方法(如GEPHI)、基于分类的异常值探测方法、基于聚类的异常值探测方法[41]。

数据可视化分析主要包含以下形式：文本可视化、网络可视化、时空数据可视化、多维数据可视化[42]。其在医疗科研分析中的意义在于帮助科研人员直观地找到数据之间的关联性，具有指导价值。

4.2 数据研发引擎

针对具体的科研应用要求，需要结合经过科研数据分析流程后的特定数据、问题模型、特定领域的算法，将其转化为实际应用[43]。这一流程需要数据研发引擎的支撑。本小节将讨论具体实验与应用所需要的引擎。

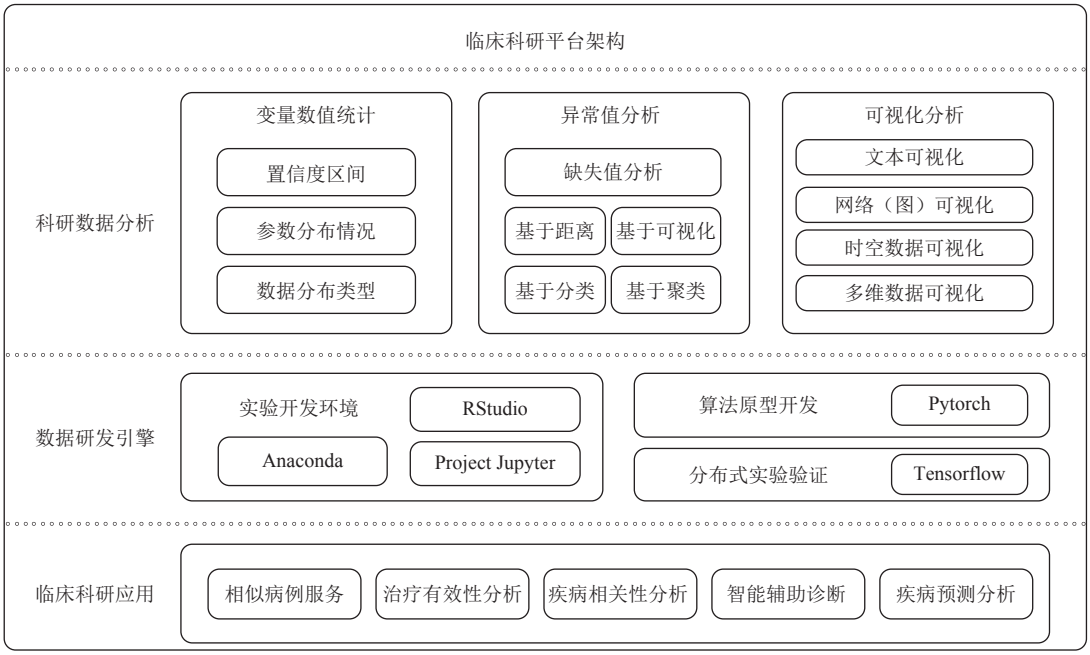

图 3 科研平台基础架构

Fig. 3 Structure of research platform

科研人员需要能够进行快速迭代科研实验的开发环境。Anaconda 是 Python/R 集成开发环境，科研人员能够直接从其平台开始进行算法研发、项目协作。Rstudio 主要面向统计学背景的科研人员，特点在于其提供了较为友好的数据可视化方案。Project Jupyter 提供了一种科研开发标准，提供交互式阶段可视化开发，适用于大规模开发环境部署，用于为科研人员提供快速实验的基础。

针对快速实验的算法迭代要求，应进行原型分析。Pytorch 是由 Facebook 开放的机器学习计算框架，其模型定义方式较为灵活，具有动态语言特性，适合进行原型开发。当算法原型提出后，应使用分布式计算实验的算法框架进行实验验证。Tensorflow 是由 Google 提出的采用数据流图(data flow graphs)用于数值计算的分布式计算框架，支持多种异构平台，主要用于机器学习与深度神经网络方面的研究与部署。

4.3　临床科研应用

本文针对临床科研中的特点，分析了临床科研大数据平台所能应用的实际场景。

4.3.1 相似病例服务

临床医疗中各医院系统遇到的病症信息较为繁杂，但具体到某些特定症状又存在大量相似现象。当病人之间状况接近或相同时，通常治疗方案比较相似。因此，使用相似病例选择算法在历史数据中找到语义相似度最高的病例，可为医生提供参考，并帮助医生提高诊疗质量[44]。

4.3.2 治疗有效性分析

历史医疗数据中存在大量用药后患者治疗症状信息，通过历史性数据使用软件(如 Revman)进行 Meta 分析，可以得出具体药物应用在特定病例后患者的身体状况、急激素水平变化情况，从而判断是否具有疗效，这有助于指导医生为患者选取药物[45]。

4.3.3 疾病相关性分析

不同疾病之间可能存在并发症或互斥性症状，以往的相关性分析方式效率较低[46]。通过大数据技术可以提供大量可参考病例，帮助科研人员分析不同疾病之间的相关性信息，有助于医生在遇到相似病例时为患者提供防范意见。

5　实　验

根据上文所提出的架构搭建了相应的大数据与临床科研平台，并在生产环境中进行相应实验。针对平台运行状况，给出了大数据平台集群软硬件运作环境、针对专病数据场景所做的具体数据处理实验，以及相应讨论。

5.1　集群环境

为提供搭建该大数据平台的参考案例，表 1 列出了目前已部署的大数据平台集群硬件配置；表 2 提供了该集群运行所需的软件版本号。

表 1　集群硬件配置表

Table 1　Cluster hardware table

类别	内容
服务器数	8 台
服务器内存	64 GB
操作系统	Ubuntu16.04
硬盘空间	4 TB
CPU	E5-2680V2
GPU	GTX 1080TI

注：具体参数为单台服务器配置参数

Ubuntu 是一款以桌面应用为主的开源 GNU/Linux 操作系统，其底层基于 Debian GNU/Linux，支持众多处理器架构，适用于异构的分布式集群搭建，是目前被广泛应用的服务器操作系统。因其便捷的包管理机制，Ubuntu 非常适合用于自动化配置来搭建大规模集群。此硬件配置表仅供参考。

表 2 中关于大数据平台的软件部分兼容性较好，最基本需要 JAVA 1.8 版本。后续 Hadoop 生态的软件版本若未按照表 2 给出版本，需保证第一位大版本号不变，方可确保大数据集群环境

运行正常。关于科研平台部分的软件，因涉及到不同 GPU 硬件的计算架构不同，故需确保是 Nvidia 显卡，且硬件计算架构高于 SM30，才能保证后续 CUDA 计算库的正常安装与使用。

表 2 集群软件版本表

Table 2 Cluster software version table

类别	内容
JAVA	1.8
Hadoop	2.9.2
Zookeeper	3.4.0
Hive	2.3.4
Spark	2.3.0
ElasticSearch	7.0
MongoDB	3.6.0
MySQL	5.6
Kafka	3.1.1
Flume	1.7.0
Kettle	8.2.0.0
CUDA	10.0
Python	3.6.4
Tensorflow	1.10
Pytorch	1.0.0

5.2 实验案例

本文针对传统方法与本平台在脓毒症专病上的临床科研应用流程进行对比。在重症加强护理病房内存在部分脓毒症患者，临床科研人员可以根据过往患者检查指标进行分析，预测病情加剧程度及死亡概率，对危险患者进行加强护理，降低死亡概率。下面对比传统人工方法与经过大数据平台优化后进行科研数据分析的流程与时间差别。

传统数据分析流程为：首先，科研人员从各医院收集脓毒症数据，各院信息科导出数据并传输(耗时未知)，并手动设定统一筛选指标，去除不合要求的数据记录(过程繁琐)；然后，查询专病相关文献，根据相关文献与诊疗经验排除不符合要求的数据记录(参考较少)；接着，查询缺失值处理相关文献，手动设定缺失值过滤指标(参考较少)；最后，根据筛选所得数据及经验建立机器学习模型，预测各指标所影响的死亡概率。

利用本平台可优化为：首先，科研人员从大数据平台直接调取 2 380 条脓毒症记录(耗时 10 秒)，自动化筛选得到 172 条检测指标(如基础信息、护理体征数据、临床症状等)；然后，根据已有专病库存储的大量过往医师专病排除标准进行排除，入选 2 204 条数据(耗时 1 秒)；接着，利用知识库内存在的多种缺失值过滤方式排除指标缺失过多的数据，剩余 368 条(耗时 1 秒)；最后，科研平台提供内置分析方法(如逻辑斯蒂回归、决策树等)辅助科研人员快速分析得到结论。

表 3 对脓毒症分析时两种分析流程的时间消耗进行了对比。由表 3 可知，大数据平台统一了数据收集方式，解决了各院数据系统不通的问题，利用数据处理引擎加快了数据处理效率；同时，辅助科研人员筛选排除无效数据，提高了临床科研分析速度，规避了人工处理可能出现的数据处理不当的问题。另外，该大数据平台高效一体化地整合了临床科研行为，快速得出临床指导结论，协助科研人员更好地开展后续临床诊疗工作。

5.3 处理性能测试

除脓毒症外，本文还建立了多个专病库，用以针对临床科研大数据平台的数据处理能力测试。表 4 提供了目前本平台在专病数据上的数据处理结果汇总。通过从多源数据中心进行数据采

表 3 专病分析时间对照表

Table 3 Special disease analysis time comparison table

方法类型	时间消耗			
	数据抽取	指标筛选	指标标准过滤	缺失值过滤
传统方法	约 10 天	约 120 天	约 3 天	约 1 天
大数据平台优化	10 秒	15 秒	1 秒	1 秒

集，利用关系数据库管理系统和 ElasticSearch 进行存储，根据关键词索引计算初步抽取数据记录。将初步筛选后的数据记录进行自然语言分析处理后，本平台成功从 15 217 026 条专病数据记录中正确将疾病类别分出，并针对敏感信息重标定标识进行数据去隐私化处理，达到了便于后续临床科研人员针对专病数据记录进行专病分析的目的。

表 4　专病数据处理汇总表

Table 4　Special disease data processing summary table

疾病名称	病例数量
肺炎	45 169
肝炎	1 233
脓毒症	2 380
炎症性肠病	271
心肌受损	3 431
心内科	11 262
疑难重症	166
急性呼吸窘迫综合征	2 479
腹泻	567

注：处理数据来自数据平台中的 15 217 026 条专病数据记录，已针对敏感数据隐去了详细描述

对比罗辉等[10]提到的平台，其一次性导入 16 692 例数据共花费 5 天时间，而本平台导入并处理 15 217 026 例数据仅耗时 10 分钟，吞吐量与效率相比较为优异。相较甘伟等[9]提到的平台，本文所搭建的平台运用了 Docker 技术进行集群内部多机数据隔离，数据安全性更能得到保障。

6　总　结

根据目前医疗领域存在的问题，本文利用大数据技术为临床科研提供相应的帮助。为解决医疗数据多源、数据结构混乱的特点，提出了数据清洗流程以及相应大数据分布式数据存储方式；为解决科研需求与数据计算问题，提出了相应的数据计算解决框架。针对医疗领域的数据隐私敏感问题，给出了相应的数据安全加密解决方案。特定于临床科研人员，本文为其提供了临床科研平台，方便科研人员更好地进行数据探索与应用分析。

当今国内外对医疗健康领域的大数据具体应用都十分重视，下一步工作是希望能够更有效地帮助临床科研人员在科研应用上的探索。

参 考 文 献

[1] 国家统计局. 中国统计年鉴 [EB/OL]. 2018[2019-07-29]. http://www.stats.gov.cn/tjsj/ndsj/2018/indexch.htm.

[2] 涂新莉, 刘波, 林伟伟. 大数据研究综述 [J]. 计算机应用研究, 2014, 31(6): 1612-1616, 1623.

[3] 陈功, 范晓薇, 蒋萌, 等. 数据挖掘与医学数据资源开发利用 [J]. 北京生物医学工程, 2010, 29(3): 323-328.

[4] 颜延, 秦兴彬, 樊建平, 等. 医疗健康大数据研究综述 [J]. 科研信息化技术与应用, 2014, 5(6): 3-16.

[5] 董诚, 林立, 金海, 等. 医疗健康大数据: 应用实例与系统分析 [J]. 大数据, 2015, 1(2): 78-89.

[6] 黄婧, 王云光, 皮冰斌. 健康医疗大数据的安全保障技术研究 [J]. 计算机时代, 2018, 317(11): 49-52.

[7] 王强, 易应萍. 临床医疗大数据治理和应用 [J]. 医学信息学杂志, 2018, 39(8): 2-6.

[8] 金昌晓, 计虹, 席韩旭, 等. 大数据科研分析平台在临床医学研究中的应用探讨 [J]. 中国数字医学, 2019, 14(2): 37-39.

[9] 甘伟, 徐明明, 陈联忠, 等. 大数据临床科研平台的设计与实现 [J]. 中国数字医学, 2019, 14(2): 40-43.

[10] 罗辉, 薛万国, 乔岫. 大数据环境下科研专病数据库建设 [J]. 解放军医学院学报, 2019, 5: 1-6.

[11] Abouelmehdi K, Beni-hssane A, Khaloufi H, et al. Big data security and privacy in healthcare: a review [C] // Proceedings of the 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, 2017: 73-80.

[12] Ko SY, Jeon K, Morales R. The hybrex model for confidentiality and privacy in cloud computing [C] // Proceedings of the 3rd USENIX Conference on Hot Topics in Cloud Computing, 2011: 8.

[13] Lohr H, Sadeghi AR, Winandy M. Securing the E-health cloud [C] // ACM International Health Informatics Symposium, 2010: 220-229.

[14] Khaloufi H, Abouelmehdi K, Beni-hssane A,

et al. Security model for big healthcare data lifecycle [C] // The 9th International Conference on Emerging Ubiqutous Systems and Pervasive Networks, 2018: 294-301.
[15] Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics [J]. International Journal of Information Management, 2015, 35(2): 137-144.
[16] Singh G, Schulthess D, Hughes N, et al. Real world big data for clinical research and drug development [J]. Drug Discovery Today, 2018, 23(3): 652-660.
[17] Zhang Z. Big data and clinical research: perspective from a clinician [J]. Journal of Thoracic Disease, 2014, 6(12): 1659-1664.
[18] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential [J]. Health Information Science and Systems, 2014, 2: 3.
[19] Mayo CS, Matuszak MM, Schipper MJ, et al. Big data in designing clinical trials: opportunities and challenges [J]. Frontiers Oncology, 2017, 7: 187-193.
[20] 徐俊刚, 裴莹. 数据 ETL 研究综述 [J]. 计算机科学, 2011, 38(4): 15-20.
[21] 林子雨, 赖永炫, 林琛, 等. 云数据库研究 [J]. 软件学报, 2012, 23(5): 1148-1166.
[22] 李学龙, 龚海刚. 大数据系统综述 [J]. 中国科学: 信息科学, 2015, 45(1): 1-44.
[23] 陈鹤群. 大数据环境下医疗数据隐私保护面临的挑战及相关技术梳理 [J]. 电子技术与软件工程, 2014, 16: 51-53.
[24] 杨宇. 基于 PKI 身份认证系统的研究和实现 [D]. 成都: 电子科技大学, 2009.
[25] 周棟淞, 杨洁, 谭平嶂, 等. 身份认证技术及其发展趋势 [J]. 通信技术, 2009, 42(10): 183-185.
[26] 房晶, 吴昊, 白松林. 云计算安全研究综述 [J]. 电信科学, 2011, 27(4): 37-42.
[27] 杨勇, 王强. 云服务数据隔离技术 [J]. 信息安全与通信保密, 2012(2): 57-59, 66.
[28] 张遥, 王森林. Docker 安全性研究 [J]. 网络安全技术与应用, 2017(8): 32-33.
[29] 常天天, 陈兴蜀, 罗永刚, 等. 面向 Hive 的基于安全域的数据隔离保护框架 [J]. 山东大学学报(理学版), 2019, 54(3): 1-9.
[30] 李昊, 张敏, 冯登国, 等. 大数据访问控制研究 [J]. 计算机学报, 2017, 40(1): 72-91.
[31] Kallahalla M, Riedel E, Swaminathan R, et al. Plutus: scalable secure file sharing on untrusted storage [C] // Proceedings of the 1st USENIX Conference on File and Storage Technologies, 2003.
[32] 余波, 台宪青, 马治杰, 等. 云计算环境下基于属性和信任的 RBAC 模型研究 [J]. 计算机工程与应用, 2019, 17(2): 1-13.
[33] Bacon J, Moody K, Yao W. A model of OASIS role-based access control and its support for active security [J]. ACM Transactions on Information and System Security, 2002, 5(4): 492-540.
[34] Hur J, Noh DK. Attribute-based access control with efficient revocation in data outsourcing systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2011, 22(7): 1214-1221.
[35] 马康, 陈松政. 基于密码的访问控制研究 [J]. 计算机应用研究, 2012, 29(1): 305-307, 315.
[36] 李晓峰, 冯登国, 陈朝武, 等. 基于属性的访问控制模型 [J]. 通信学报, 2008, 29(4): 90-98.
[37] 袁春, 文振焜, 张基宏, 等. 基于密码学的访问控制和加密安全数据库 [J]. 电子学报, 2006, 34(11): 2043-2046.
[38] Yu S, Wang C, Ren K, et al. Achieving secure, scalable, and fine-grained data access control in cloud computing [C] // Proceedings of the 29th Conference on Information Communications, 2010.
[39] 程健一, 关毅, 何彬. 基于 SVM 和 CRF 双层分类器的英文电子病历去隐私化 [J]. 智能计算机与应用, 2016, 6(6): 17-19, 24.
[40] Knorr ME, Ng TR. Algorithms for mining distance-based outliers in large datasets [D]. Canada: University of British Columbia, 1998.
[41] 陶盈春, 张红丽, 徐健. 异常值探测在大数据分析中的应用研究 [J]. 情报科学, 2018, 36(3): 75-80.
[42] 任磊, 杜一, 马帅, 等. 大数据可视分析综述 [J]. 软件学报, 2014, 25(9): 1909-1936.
[43] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述 [J]. 模式识别与人工智能, 2014, 27(4): 327-336.
[44] 胡敬远. 基于关联数据的医疗决策支持系统的研究 [D]. 上海: 上海交通大学, 2015.
[45] 李存存, 王晶晶, 陈潮, 等. 坤泰胶囊与激素替代疗法治疗更年期综合征有效性和安全性比较的 Meta 分析 [J]. 中国中西医结合杂志, 2013, 33(9): 1183-1190.
[46] 金晓燕, 雷虹, 胡美华. 结直肠癌患者术后化疗期间癌因性疲乏与疾病不确定感的相关性分析 [J]. 护理管理杂志, 2011, 11(1): 3-4, 23.