

智能化多源多维大数据融合设计在卫生应急中的应用*

李少琼 金丽珠 杜雪杰 郭 青

(中国疾病预防控制中心 北京 102206)

〔摘要〕 目的/意义 针对来源复杂、标准多样的多源多维大数据，提出不同业务应用场景的数据治理和融合技术。方法/过程 以多源多维大数据融合在应对重大突发公共卫生事件的实践为例，重点阐述多源多维大数据融合面临的挑战、应用架构设计、技术集成、实施路径、应用场景等内容，分析应用效果和存在的不足。结果/结论 通过智能化多源多维大数据融合技术的设计，实现对千亿级别高频次更新的实时监听、读取、治理和交互关联，探索有效应对突发公共应急事件的智慧化路径。

〔关键词〕 多源多维大数据融合；数据治理；卫生应急

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.06.013

Intelligent Multi-source and Multi-dimensional Big Data Fusion Design in Health Emergency Response

LI Shaoqiong, JIN Lizhu, DU Xuejie, GUO Qing
Chinese Center for Disease Control and Prevention, Beijing 102206, China

〔Abstract〕 **Purpose/Significance** To propose data governance and fusion technologies for different business application scenarios for the fusion of multi-source and multi-dimensional big data with complex sources and diverse standards. **Method/Process** Taking the practice of multi-source and multi-dimensional big data fusion in responding to major public health emergencies as an example, the paper focuses on the challenges faced by multi-source and multi-dimensional big data fusion, the design of the application architecture, technology integration, the implementation path, the application scenarios, etc., and analyzes the application effect and the deficiencies that exist. **Result/Conclusion** Through the design of intelligent multi-source and multi-dimensional big data fusion technology, real-time monitoring, reading, governance, and interactive correlation of high-frequency updates at 100 billion levels are realized, and intelligent paths are explored for effectively responding to public emergencies.

〔Keywords〕 multi-source and multi-dimensional big data integration; data governance; health emergency response

1 引言

公共卫生应急是指为了预防和减少公共卫生领域突发事件的发生，控制、减轻和消除突发公共卫生事件引起的严重社会危害而开展预防和准备、监测和预警、处置和救援、恢复和评估等工作^[1]。20

〔修回日期〕 2024-02-26
〔作者简介〕 李少琼，助理研究员，发表论文 13 篇；通信作者：郭青。
〔基金项目〕 广东省广州实验室应急攻关项目（项目编号：SRPG22-009）。

余年以来,全球陆续经历了人感染动物源性流感、基孔肯雅热、埃博拉、寨卡病毒病、中东呼吸综合征等新发、突发传染病,这些疾病迅速成为人类共同的威胁^[2-3]。在重大公共卫生应急事件应对中,相关数据资源是有效实施应急决策的前提,单来源单维度的数据难以满足科学决策需求,如何利用信息技术将多渠道来源数据融合应用,充分释放数据价值,辅助国家与科研工作者前瞻性地识别卫生应急事件的发展态势,制定合理的防控策略尤其重要^[4-5],这对多源数据智慧化融合应用提出新的挑战。

2 现状分析

多源数据有多种来源与表现形式,根据数据来源可分为内部数据与外部数据。内部数据主要包括第三方检测机构的核酸检测结果、接种单位的疫苗接种档案、法定传染病、重点慢性非传染性疾病、死亡登记等信息,外部数据包括民航航班、密切接触者、境外人员入境信息,以及现场调查等信息。根据时效性可分为积累的历史数据和信息系统运行产生的在线数据。本研究中多源多维大数据融合指在应对重大公共卫生应急事件中根据不同业务应用场景,将不同来源、不同维度和不同结构的数据进行清洗、抽取、分选等预处理,利用适当的数据融合架构和算法进行实时关联组合,提取数据特征,获得更完整的信息。数据融合被普遍采纳的定义是由美国三军组织实验室理事联合会提出的,认为数据融合是一种多层次、多角度的信息处理手段。处理过程中,为了得到精确的状态判定、身份识别以及完整及时的态势与威胁评估,将多源数据进行一系列操作,包括检测、结合、关联、估计和组合等^[6]。卫生应急事件中积累的数据具有领域宽、汇聚量大、实时性高、用户广泛、指标多样化、数据质量不同等特征。因此在融合应用中面临的主要挑

战,一是信息标准多样性,各业务条线的信息系统建设标准不一致,数据元定义具有较大的不一致性,相同的数据元在不同业务对象中描述存在语义多样性,其语义结果的不同也会引起表冲突、值冲突、属性冲突等问题;二是数据来源与类型多样性,卫生应急联防联控机制成员机构包含公安、口岸、交通、民航等,其未形成卫生应急管理信息一体化协同平台,存在跨业务数据分散、数据的颗粒度和时效性不一致以及身份识别信息完整和准确性等问题,为实时的数据关联带来巨大挑战;三是海量数据实时处理能力,重大公共卫生应急事件的决策需要掌握实时动态信息,数据汇集、存储、处理、分析、展示等环节强调时效性,因此对存储空间、计算能力提出更高的要求,高效的数据处理机制需要融合应用新技术。此外,卫生应急数据包含大量个人信息、疾病信息、诊疗信息等敏感的隐私数据,与个人行为、健康危害因素等信息融合后,更加提升了数据的敏感性,数据安全问题不但关乎公民个人信息权、隐私权,更影响国家战略安全,隐私对数据安全策略也提出更高要求。

3 方法

3.1 应用架构

中国疾病预防控制中心信息系统是我国疾病预防控制和突发公共卫生应急信息管理的重要信息化支撑^[7],依托总体技术架构,基于一体化的安全保障和标准体系搭建多源多维大数据融合应用功能,将多源多维的数据集成到统一的框架与平台下,实现对高频次更新的实时监听、读取、治理、交互关联及大数据治理智能协同。以轻量级微服务分布式服务技术框架为基础,构建多源多维大数据融合技术平台,总体架构分为数据资源层、应用支撑层、应用层、接入层,见图 1。

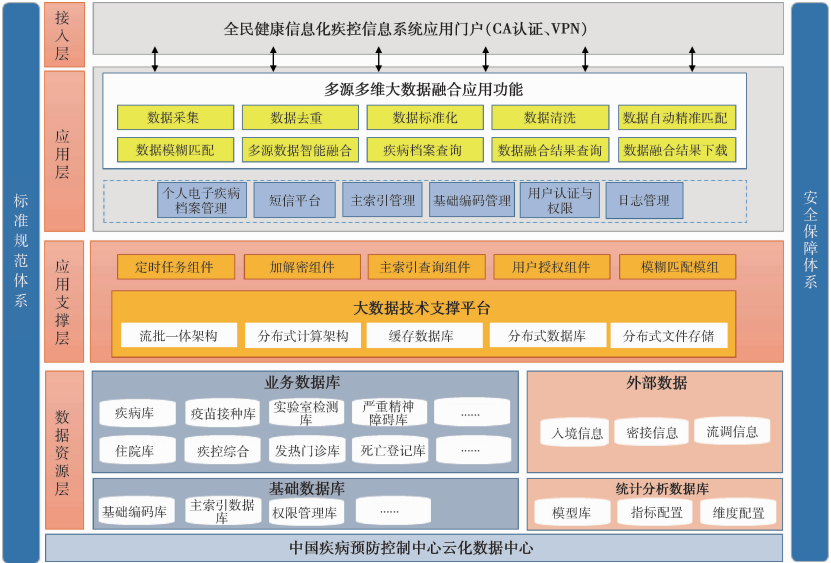


图1 总体应用架构

3.2 技术集成

在实施过程中以满足业务不断持续更新，兼顾单日超过9亿在线数据交易量与现有计算资源之间的关系，以及数据关联实时性等需求为前提，集成关键技术。

3.2.1 集成 Flink 流批一体技术架构 搭载数据处理与数据融合的模式组，实现精准的实时离线交叉数据分析和业务决策，具备容错能力和状态保持能力，确保数据计算的高效性、稳定性和一致性，可灵活扩展多业务主题处理与分析，支持大规模的集群模式，具备低延迟的流处理、高吞吐、高稳定性、数据可回溯等能力。基于 Flink 建立数据同步流程，Flink 程序启动时首先会将业务中历史数据加载到 Flink 中进行数据运算，如错误数据的清洗、数据的质量校验等，将不符合要求的数据分流到错误数据处理流中，将符合要求的数据通过身份证号码、姓名及各业务属性专有字段进行处理，保存到分布式数据库中。历史数据处理完成后，Flink 通过监听 MySQL 的 binlog 方式获取各业务中实时变化的数据内容，当数据库中的数据变更时，Flink 会将变化的数据实时读取到 Flink 处理流程中进行数据计算处理。大数据组件包括 Flink、Spark、TiDB、Doris、Hadoop、Canal、HBASE，支撑千亿级别数

据的存储、分析和查询。

3.2.2 构建多元数据的全局字典 根据不同实时业务处理与非实时分析主题实现对任意数据类型精确去重。对分业务数据及外部来源数据进行治理，提供多源海量数据去重与智能融合，具备跨业务跨部门的接口调用能力，为国家级和省级相关机构提供分业务分场景的传染病防控多源多维大数据融合结果。

3.2.3 数据融合模糊匹配 采用分布式数据库 UDF 函数方式的关联模型搭建关联匹配模型组，用于多源数据的精准匹配和模糊匹配。数据关联时通过个人身份信息执行精准的批量匹配和关联并保存结果，对个人身份信息缺失、不完整或不准确的采用字符串编辑距离 Levenshtein、正则规则、分词等算法构建模型组模糊匹配。通过计算两个字符串间的相似度，利用除个人身份信息外的其他扩展信息，如患者的现住址、人群分类等不同业务数据的特点设定匹配字段的权重，匹配结果达到一定分数时视为模糊匹配成功。

3.3 数据融合

构建患者个人信息主索引实现数据融合。患者个人信息主索引是患者基本信息的检索目录，通过唯一的患者标识信息将多个卫生医疗信息系统有效

关联起来，实现各业务间的互联互通^[8]。从中国疾病预防控制中心信息系统已汇集的海量监测业务数据中，提取不同业务个人基本信息的身份证件号建立主索引管理库，将各类业务信息中的个人基本信息与主索引建立关联，将多条业务数据的个人基本信

息合并到一条主索引上，将分散存储的全历史时期诊断信息、治疗信息、疫苗接种信息、死亡信息等数据有效关联。因此以患者个人基本信息为基点，通过全生命周期数据集合实现全景视图，达到不同业务场景数据资源有效融合。

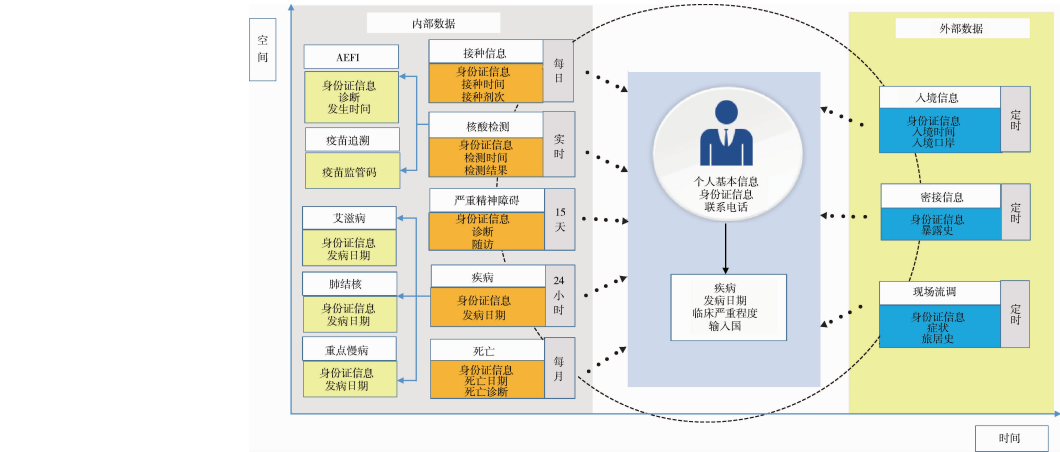


图2 数据融合

3.4 服务层的融合

服务层结合不同业务场景需求为用户提供数据汇聚融合后的查询界面、接口、服务与应用。在服务层建立触发器机制，设计感知器准则，当新增信息或信息发生变化时触发数据关联机制，按精准匹配、模糊匹配等规则动态智能更新，能够基于时间、空间关联的关系融合，保持数据动态性和时效性。

4 结果

4.1 流程机制

卫生应急精准决策和科学研究支撑数据具备来源广泛、类型复杂等特征，信息来源于医院、基层医疗机构、第三方检测机构、疾控机构、疫苗接种单位、民航、工信等多部门，数据类型包括疾病诊断信息、实验室检测结果、暴露史、免疫接种史、临床症状、病情、死亡原因等多类型。各类数据的监测需求不同，精度、地域空间和时间频次具有多维特点。实际应用中以业务需求为导向确立并定义数据源作为输入流，存储至各类业务数据库中，对归集的数据源预处理。一是数据选取，在原始数据

库中选取与任务相关的数据项。二是属性一致化处理，判断数据的标准化，利用平台的统一元数据描述框架，将不同类型数据的多样化元数据模式重新标识转换为该元数据模型。三是数据标准化，数据清理、填补空缺值、处理错误数据、消除噪声数据、去除孤立点等有效提高数据质量。四是数据归约，在保留数据完整性的前提下缩小处理规模，对数据去重提高效率，预处理完成的数据根据不同的业务场景存储到分布式数据库中，数据流入模型组融合关联，通过触发器自动感知数据更新情况，实现定时、实时的数据处理和结果呈现。本研究构建了多源多维大数据智能融合流程机制，为卫生应急提供及时准确的数据支撑。

4.2 用户授权

根据不同的数据特性、业务场景和应用需求及数据的业务类型、实时性、安全等级要求采用不同的数据结果获取和查询方式，对于省统筹区域平台、外部机构通过开放接口的方式，实现数据的调用和查询。用户授权管理采用基于角色的访问控制（role – based access control，RBAC）模型，结合业务域、功能点、控制点三维立体用户管理模型^[9]，

联接业务主题分类和机构职责,根据机构内部职责分配与指定权限,使多机构的多人员在不同时间范围内有多种可能的权限,保障不同业务用户获取不同业务管理的功能操作权限,通过控制点严格控制用户查询、个人隐私信息下载和敏感数据的权限,保障隐私数据的安全性,防止数据泄漏,构建科学合理的用户授权管理机制。

4.3 数据安全

卫生应急数据包含个人隐私信息,在多源数据关联分析过程应用患者个人基本信息属性标识,融合后数据提炼不同业务主题特征,对数据安全提出更高的要求。遵循国家标准《GB/T 22239—2019 信息安全技术网络安全等级保护基本要求》,安全保障重点体现在数据保密性方面,数据基于 HTTPS 协议的 SSL 传输,重要信息以加密形式存储,敏感数据存储到数据库时,基于国密算法,针对每个敏感数据项加密,加密后的数据存储到数据库中。此外,设有严格的信息访问权限控制机制,未经授权查看个人隐私的用户查询信息时强制去隐私化,已授权用户仅能执行查询操作,需要下载时信息摆渡至云桌面,分析的结果数据审核通过后,才可以下载至本地计算机。数据每日自动备份,Web、应用服务器和数据库服务器采用热冗余方式部署,构建了安全、可靠的数据安全保护机制。

4.4 应用效果

通过多源多维大数据融合的技术架构,实现对千亿级别高频次更新的实时监听、读取、治理和交互关联,提供了实时的定制化分析需求支撑,数据融合应用体现出海量数据汇聚关联带来的数据全面性、可靠性和潜在知识关联的协同作用。从指标关联、时间关联、空间关联等维度有效统筹各类数据资源,利用海量多源数据强融合的优势,在重大传染病防控和突发公共卫生事件应急工作中发挥信息服务的重要价值。基于指标关联的融合,能够体现不同业务领域数据之间的相关关系^[10]。基于空间关联的关系融合,在一定程度上发现聚类性和潜在的风险,为各地统筹医疗物资储备、保障民生物资供

应提供参考,取得良好的社会效益。实践工作实现了多源多维大数据融合应用,为突发公共卫生事件应急工作提供重要数据支撑服务。

5 结语

探索大数据架构智慧化融合技术的集成和应用,突破传统数据管理模式,“以人为核心”的信息管理模型挖掘疾病与其他因素之间的关系,有效解决了数据分散、各类数据标准化和准确性不足而难以共享应用的限制。在应对突发公共卫生应急事件中充分利用已经积累的业务数据和外部共享数据资源,以真实的业务需求驱动,集成应用流批一体计算架构、图数据库、大数据建模、自然语言处理及机器学习等技术构建数据实时交互、协同共享的平台。通过平台支撑基于海量实时多源数据的汇集、预处理、融合关联、分析展示,对辅助决策发挥大数据的重要赋能作用,对多源数据融合的基础理论、技术方法以及实践应用具有重要意义,在数据处理能力方面基于现有资源实现多源大数据处理高吞吐量与高稳定性,可支撑超千亿级别不同来源业务数据的实时智能化融合,成为数据深度挖掘的重要基础。

随着多源头数据融合和实践的应用,驱动数据融合从理论层面进一步深入。本研究现阶段的应用和验证主要集中在应对突发公共卫生应急事件中获得的多源多维卫生应急数据,以结构化和半结构化数据为主;下一步将总结现有应用基础,加强模型训练和迭代优化,利用人工智能、多模态数据融合等新技术进一步提升数据关联的精确性。此外,集成加密机和 SSL 通信加密协议等安全措施保证数据传输阶段的机密性和完整性,为有效应对重大公共卫生应急事件探索智慧化数据融合路径。

作者贡献: 李少琼负责论文撰写及修改;金丽珠负责资料收集;杜雪杰负责论文审核;郭青负责论文指导。

利益声明: 所有作者均声明不存在利益冲突。

(下转第 84 页)

工、个性化分析和创新服务方面仍有较大提升空间。另一方面，气象敏感性疾病错综复杂，要实现全面的气象敏感性疾病的预警和风险评估需深入探究气候变暖对疾病发病规律及其病理的影响机理，构建气象部门与卫生、环保等多部门间的协同合作体系，共同推进相关综合研究。

作者贡献：郑环负责论文初稿撰写和修改；陈强负责论文润色；张睿、孟郁洁负责文献调研；赵嘉、金丽珠负责文献整理；杜雪杰负责技术支持；王松旺负责论文修改与审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

- 1 胡利娟.《中国气候变化蓝皮书（2022）》发布中国升温速率高于全球平均水平 [J]. 中国科技财富, 2022 (8): 48.
- 2 王嘉鑫, 石彦军, 卢山, 等. 我国华东与西南县域主要气象敏感性疾病变化特征及其医疗费用研究 [J]. 沙漠与绿洲气象, 2019, 13 (6): 133 - 140.
- 3 肖冰霜, 马玉霞, 郑晓东, 等. 南京市最低气温对不同人群循环系统疾病死亡人数的影响 [J]. 兰州大学学报 (自然科学版), 2017, 53 (3): 401 - 406.
- 4 董继元, 陈永聪, 张本忠, 等. 兰州市气温对脑卒中发

- 病的滞后效应研究 [J]. 气候变化研究进展, 2017, 13 (4): 366 - 374.
- 5 张莹, 王式功, 贾旭伟, 等. 气温与 PM2.5 协同作用对疾病急诊就诊人数的影响 [J]. 中国环境科学, 2017, 37 (8): 3175 - 3182.
- 6 杨丝絮, 彭丽, 叶晓芳, 等. 气温与 PM2.5 协同作用对心脑血管疾病死亡的影响研究 [J]. 上海预防医学, 2023, 35 (7): 660 - 666.
- 7 ZHANG R, MENG Y, SONG H, et al. The modification effect of temperature on the relationship between air pollutants and daily incidence of influenza in Ningbo, China [J]. Respiratory research, 2021, 22 (1): 153.
- 8 李传玺, 刘起勇, 马伟. 广州市极端降水事件对不同特征人群登革热发病的影响 [J]. 山东大学学报: 医学版, 2021, 59 (12): 151 - 157.
- 9 马盼, 王式功, 尚可政, 等. 气象舒适条件对呼吸系统疾病的影响 [J]. 中国环境科学, 2018, 38 (1): 374 - 382.
- 10 徐飞龙, 吕佳, 吴佳荣, 等. 中医药科学数据管理平台设计与实现 [J]. 医学信息学杂志, 2023, 44 (12): 78 - 82.
- 11 张俊. 基于 Redis 实现关系型数据库内存化研究 [D]. 成都: 四川师范大学, 2021.
- 12 孟郁洁, 戚晓鹏, 马睿, 等. 流行病学动态数据采集平台在公共卫生调查中的应用 [J]. 医学信息学杂志, 2013, 34 (6): 18 - 22.

(上接第 78 页)

参考文献

- 1 徐婷, 鲍勇, 王韬. 中国公共卫生应急管理体系的变迁与效果分析 [J]. 中国公共卫生, 2020, 36 (12): 1704 - 1706.
- 2 于佳会, 刘佳静, 郑建明. 多源多维数据融合研究态势: 理论、方法与应用 [J]. 情报杂志, 2022, 41 (5): 133 - 138, 207.
- 3 邓建高, 吴灵铭, 齐佳音, 等. 突发公共卫生网络舆情信息传播博弈分析 [J]. 现代情报, 2021, 41 (5): 139 - 148.
- 4 刘樑, 金瑞丰, 魏华. 突发公共卫生事件研究的趋势和特征 —— 基于 CiteSpace 的知识图谱分析 [J]. 电子科技大学学报 (社会科学版), 2021, 23 (4): 46 - 52.
- 5 STEINBERG A N, BOWMAN C L, WHITE F E. Revisions to the JDL data fusion model [C]. Orlando: Sensor Fusion: Ar-

- chitectures, Algorithms, and Applications III, 1999.
- 6 苏雪梅, 赵自雄, 赵嘉. 疾病预防控制信息化建设、应用与规划 [J]. 医学信息学杂志, 2023, 44 (5): 1 - 6.
- 7 宋磊. EMPI 在医院信息系统中的研究与构建 [J]. 医学信息学杂志, 2016, 37 (4): 63 - 65, 82.
- 8 郭青, 赵自雄, 苏雪梅, 等. 基于电子疾病档案的疾病动态监测全周期管理模型研究 [J]. 中国卫生信息管理, 2020, 17 (4): 411 - 415.
- 9 国家卫生健康委办公厅关于进一步加强疫情防控期间医务人员防护工作的通知 [EB/OL]. [2023 - 09 - 26]. <http://www.nhc.gov.cn/zycgj/s7659/202002/75c6e88ecbeb42a9a26acb538383e2fc.shtml>.
- 10 WANG F, AN Z, RODEWALD L, et al. Guangdong's study of the effectiveness of China's inactivated vaccines against the SARS - CoV - 2 B.1.617.2 (Delta) variant [J]. China CDC weekly, 2021, 3 (34): 728 - 730.