

面向 DataOps 的医院临床数据中心设计

俞 高^{1,2}, 刘燃峰³

- (1. 医惠科技有限公司数据中心研发部, 浙江 杭州 310051;
2. 中国科学技术大学管理学院 MBA 中心, 安徽 合肥 230000;
3. 连云港市第一人民医院信息科, 江苏 连云港 222002)

摘要:随着“三位一体”的智慧医院建设的开展, 医院临床数据中心在医院信息化建设中扮演越来越重要的角色, 当下很多临床数据中心的整体架构设计的出发点多是面向评审评级, 而忽略数据平台本身的特性及平台的可实施性。本文借鉴研发运维一体化 DataOps 理念, 采用微服务架构、面向敏捷交付构建临床数据中心, 从而有效解决临床数据中心的架构理想实施艰难的问题。

关键词: DataOps; 临床数据中心; 敏捷

中图分类号: R107

文献标识码: B

DOI: 10.3969/j.issn.1006-1959.2024.19.007

文章编号: 1006-1959(2024)19-0048-06

Hospital Clinical Data Centre Design for DataOps

YU Gao^{1,2}, LIU Ranfeng³

- (1. Data Centre R&D Department, Ewell Technology Co., Ltd., Hangzhou 310051, Zhejiang, China;
2. MBA Center, School of Management, University of Science and Technology of China, Hefei 230000, Anhui, China;
3. Information Department, the First People's Hospital of Lianyungang, Lianyungang 222002, Jiangsu, China)

Abstract: With the construction of the trinity smart hospital, the hospital clinical data center plays an increasingly important role in the construction of hospital information technology. The starting point of the overall architecture design of many clinical data centers is mostly oriented to the evaluation of ratings, while ignoring the characteristics of the data platform itself and the implementability of the platform. In this paper, we propose to draw on the concept of integrated DataOps for R&D, operations and maintenance, adopt microservice architecture, and guide the design of clinical data center products for agile delivery, so as to effectively solve the problem of difficult implementation of the architectural ideals of clinical data centers.

Key words: DataOps; Clinical data center; Agile

伴随公立医院高质量发展改革的推进, 尤其是电子病历应用水平功能测评和医院互联互通成熟度测评的普及, 医院临床数据中心已经成为了三级医院信息系统的标配。由于诊疗数据涉及业务系统多、格式与质量参差不齐, 难以有效地进行数据收集、应用, 导致临床数据的价值无法高效完成提升与转化。医院临床数据中心主要用于汇聚、加工和分析临床诊疗数据, 实现诊疗数据在系统间的互通和共享, 以支持医疗机构的管理和决策。不过, 在医院临床数据中心实施过程中, 既需要对接多个第三方厂商的若干信息系统, 又需要满足包括实时和非实时的各种数据服务需求, 因此数据中心交付难的问题一直困扰项目甲乙双方。当前建设临床数据中心面临的主要难题有: ①数据整合困难: 由于医疗机构信息建设通常缺乏长期整体规划^[1], 业务信息系统之间数据不相通、离散分布, 使得在收集围绕疾病事件/患

者就诊的全面数据时, 遭遇不少困难。即使已从多个业务系统中获得数据, 如何将医院数据进行关联与整合, 也是医院数据应用面临的难题之一。②数据清洗流程繁琐: 临床业务系统中许多数据存在录入不规范、不一致、缺失、重复、混乱、格式不规范、非结构化等情况^[2], 因此在从各个业务系统中获取海量临床专病科研所需的数据时, 对数据的清洗、转化处理较为困难。要让医疗数据生根发芽实现数据驱动, 上述问题的解决则需要依靠数据治理来解决。③数据标准化程度低: 医疗机构通常由多个系统供应商提供信息系统, 各公司系统数据库缺乏通用数据元素, 系统术语缺乏临床实用性^[3]。各类医学与医技术语使用不规范、命名不一致, 想将全部历史病历数据(非结构化和结构化数据)进行标准化、归一化处理、存储, 从而获得高质量数据, 是一件费时费力的工作。④缺乏搜索临床数据的入口: 医院存储有大量可利用的临床数据尤其是中文电子病历数据, 缺少准确地检索到所需数据的中文相似电子病历检索工具^[4], 形成“物不能尽其用”的现状。

作者简介: 俞高(1986.12-), 男, 安徽定远县人, 硕士, 高级工程师, 主要从事医疗大数据研究

DataOps 方法在各个行业的应用研究也取得了一定的进展。在制造业和公共事业行业,公共事业组织考虑采用 DataOps 来建立数据驱动的文化并在市场上获得竞争优势^[5]。在医疗健康领域,Bahaa S 等^[6]为数据科学和分析项目提出了一种新的 DataOps 生命周期,并应用于使用加州大学欧文分校的心脏病数据集的医疗保健案例研究。提倡协作、质量控制和更快地交付的 DataOps,在汲取了软件开发中的 DevOps 方法并与敏捷和精益生产等相结合之后,可以有效的促进临床数据中心的建设。因此,本文试图利用 DataOps 理念构建一套科学、标准、规范的面向医院的临床数据中心,解决临床数据中心建设过程中遇到的数据整合难、清洗流程繁琐等问题,从而提升医院临床数据中心项目的交付质量和效率。

1 DataOps 简述

1.1 DataOps 定义 DataOps 由 Lenny Liebmann^[7]在2014

年首次提出,Ereth J^[8]、Atwal H^[9]、Vargas-Rueda L^[10]、Mainali K^[11] 等学者在其研究中都提到了自己对 DataOps 概念的理解,表 1 为各个学者对 DataOps 的理解。总体来说,DataOps 是一种面向流程的方法,它由人驱动,而不是技术驱动。对于专注于数据的数据科学家和开发人员来说,它允许其在不放弃数据治理要求的情况下开展敏捷的数据处理流程。DataOps 为数据编排、自动化和协作提供了最佳实践,旨在提高工作效率。作为新兴的概念,DataOps 正成为热点。

1.2 DataOps 的实施准则 综合上述专家的论述,可以看出 DataOps 与敏捷软件工程、DevOps 实践、精益生产等概念相近。如图 1 所示,DataOps 将敏捷、DevOps、精益生成等技术用于数据科学或数据分析项目中,来提升数据处理质量获取更大收益。相比传统的数据处理方式,DataOps 更强调数据过程的自动化和协同化。

表 1 DataOps 定义

研究者	文献名称	DataOps 定义
Julian Ereth	DataOps-Towards a Definition	DataOps 是一组实践、流程和技术,将数据集成和面向流程的数据视角与敏捷软件工程的自动化和方法相结合,以提高质量、速度和协作,并促进持续改进的文化。
Harvinder Atwal	Practical DataOps-Delivering Agile Data Science at Scale	当我们将原始数据转化为有用的数据产品视为一个需要高度协作、自动化和持续改进的端到端的流水线过程,DataOps 是一套面向这样流水线过程的解决方案。
Gartner Inc	Definition of DataOps-Gartner Information Technology Glossary	DataOps 是改进跨组织的数据管理者和消费者之间的数据通信、集成和自动化的一种数据协作管理实践。
G. Alley	What is a Data Pipeline	DataOps 是一种通过自动化、测试、协调、协作开发、容器化和持续监控使得数据分析过程统一化的解决方案。
The DataOps Manifesto	The DataOps Manifesto	DataOps 是一种从数据收集到数据处理后的信息传递的分析过程,是以更好的方式开发和交付数据分析项目的方法。
G. Anadiotis	DataOps: Changing the world one organization at a time	DataOps 是一种提升低效的团队、系统和数据之间的沟通和整合的数据管理方法。
Kiran Mainali	DataOps: Towards Understanding and Defining Data Analytics Approach	DataOps 是一个合作团队使用适当的工具和技术以有效的方式从数据中获取价值的过程。

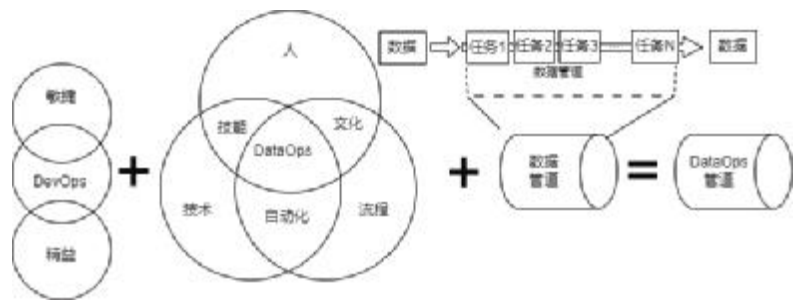


图 1 DataOps 示意图

数据类项目具有区别其他项目的特异性,数据分析的需求往往变化很快。以前处理的一些数据管道对于新的业务很容易过时而不适用,因此,往往需要从头开始一个新的数据任务来处理新需求。为了应对不断变化的需求和数据,敏捷和 DevOps 等被引入数据分析项目中,形成 DataOps 方法。其落地过程还需要结合人、技巧、技术、流程、文化等。总体来说,DataOps 的实施准则主要涉及以下几点:

- ① DataOps 文化:首先,通过确定组织中的人员和文化来启动 DataOps;其次,针对 DataOps 建立与组织内部的流程和工具一致的管理机制。
- ② 自动化:通过各种工具和技术将数据处理分析过程进行集成,从而使数据分析流程自动化运行。
- ③ 版本控制:版本控制对于数据、文档和代码的跟踪是必不可少的。通过版本控制,不同的团队成员可以方便沟通数据分析工作,避免由于版本原因导致的问题。
- ④ 容器化和复用:有复用的可能性,就不必要浪费时间重做同样的事情。通过将程序容器化提升数据分析过程的复用有助于减少因外部环境而导致的失败风险。
- ⑤ 设置多个环境:为生产环境和开发环境设置一个独立的环境,频繁的需求更新在开发环境进行,这不会对正在生产环境产生影响。在开发环境中,每个团队成员有自己的工作环境,这样每个人都可以独立工作而不影响其他环节。
- ⑥ 测试和试验:在将开发环境发布到生产环境之前,需要做一个充分的测试。没有测试,就无法保证数据处理过程的质量。
- ⑦ 持续集成和部署:在测试环境就将各数据流程的工作集成起来,在通过测试后,使用持续部署,将其发布到生产环境中。
- ⑧ 持续监控:定期监控开发环

境和生产环境,以跟踪追溯各个数据处理流程的性能、输入和输出质量。通过统计分析持续监测的数据,进一步改进各个数据处理流程。

⑨ 沟通与合作:不断地与客户、团队成员及利益相关者沟通,使信息能够更快地传递。在工具、作业任务和人之间建立协作工作平台,保障得到更好的结果。

总之,DataOps 提倡协作、质量控制和更快地交付项目。在汲取了软开发中的 DevOps 方法,并与敏捷和精益生产等相结合之后,再加上落地实施过程中需要遵守的原则和指南,DataOps 的推广落地已经是有章可循。

2 基于 DataOps 的临床数据中心设计

2.1 建设思路 面向数据全生命周期的 DataOps 是一种以价值最大化为目标的最佳实践,该方法聚焦于协同从数据需求输入到交付物输出的全链路过程,包含了敏捷化、标准化、自动化、智能化的特征。基于 DataOps 的系统开发遵循迭代和增量的方法,允许管理者和利益相关者在软件的开发期间优先考虑用户价值的特性,并跟踪用户的反馈。根据客户的持续反馈进行的小迭代,确保项目保持令人满意的进展。

DataOps 倡导协同式、敏捷式的数据资产管理^[12](图 2)。它需要通过建立数据管道,明确数据流转过程及环节,并且该过程能够自动化进行,从而可以缩短数据项目的周期,并持续改进数据质量,降低管理成本,加速数据价值释放。具体到医疗机构,业务部门对应了医务科、质控科及临床业务科室等提出数据需求的业务部门;IT 部门则对应信息部门或者医院 IT 供应商;如今越来越多的医疗机构开始成立专门的数据管理部门,部分医疗机构成立“医学大数据或人工智能中心”“运营部”等承担对应的数据管理工作。



图 2 DataOps 敏捷协同的一体化管理

2.2 系统架构 为了实现图 2 的 DataOps 敏捷协同一体化管理机制,需要特定的临床数据中心平台来支撑,它需要包括工作流业务流程工具、测试和监控工具、部署自动化工具,也会涉及代码和数据版本控制工具、数据分析和可视化工具、数据治理工具等。

为了满足 DataOps 产品架构方面采用基于微小化的整体建设架构理念,以支撑自动化运维管理、灵活、动态、安全的基础架构,将数据中心管理平台和数据应用分解成为模块化应用,见图 3。通过这种模式的建设,能够随着整体系统运营时间的增长,从而使得

医院有效分析和利用医疗大数据的能力不断提升^[13]。

平台采用 Spring Cloud 微服务框架,在该框架基础之上开发了数据服务配置功能。Spring Cloud 提供了微服务的配置管理、服务发现、断路器、智能路由、微代理、控制总线、全局锁、分布式会话和集群状态管理等,基于 Spring Cloud 微服务框架的数据平台满足了各类院内、互联网、医联体等服务的高并发、稳定性等需求。微服务基础底座依托 Kubernetes 资源配额管理、容器编排、应用部署、集群管理能力,采用了国内外先进、多源丰富的云原生开源社区与企业服务技术栈兼容融合,构建微服务基础底座,实现对整体系统的彻底块间解耦,提供灵活的软件复用和服务重构能力。

2.3 核心模块 能够支撑 DataOps 落地的临床数据中心平台需要具备敏捷开发、快速迭代的要求。为了满足上述要求,本临床数据中心设计了数据集成引擎、数据可视化引擎、自然语言处理引擎和数据治

理引擎 4 个模块,它们以低代码、易操作的优势被公司实施人员和医疗机构技术人员所接纳。

2.3.1 集成引擎模块 面向 DataOps 的数据研发环节,需要通过串联数据模型设计、数据标准设计、数据质量设计、数据集成、数据存储、数据加工等流程^[14],建成数据研发一体化能力,为此需要一个集成上述功能的数据集成引擎工具。如图 4 所示,数据集成引擎模块是结合数据仓库建模、采集和实施的方法论,满足 Oracle、SqlServer 等各种主流数据源数据的整合要求,提供了更加便捷、更智能的服务。数据集成的核心任务是要将互相关联的分布式异构数据源集成到一起,使用户能够以透明的方式访问这些数据源。有别于其他行业的数据集成工具,该工具通过内置医疗行业临床、运营和科研仓库数据采集模板,降低了数据采集的运维门槛,使得复杂的工作简单化,杂乱的流程统一化。

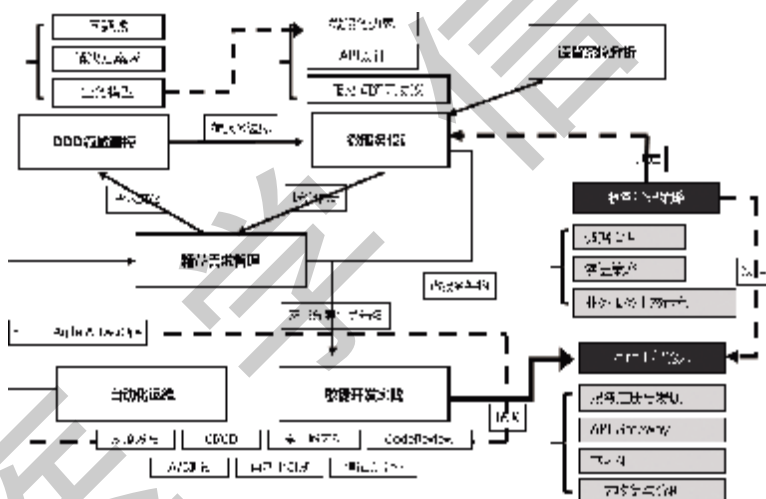


图 3 临床数据中心微服务架构示意图

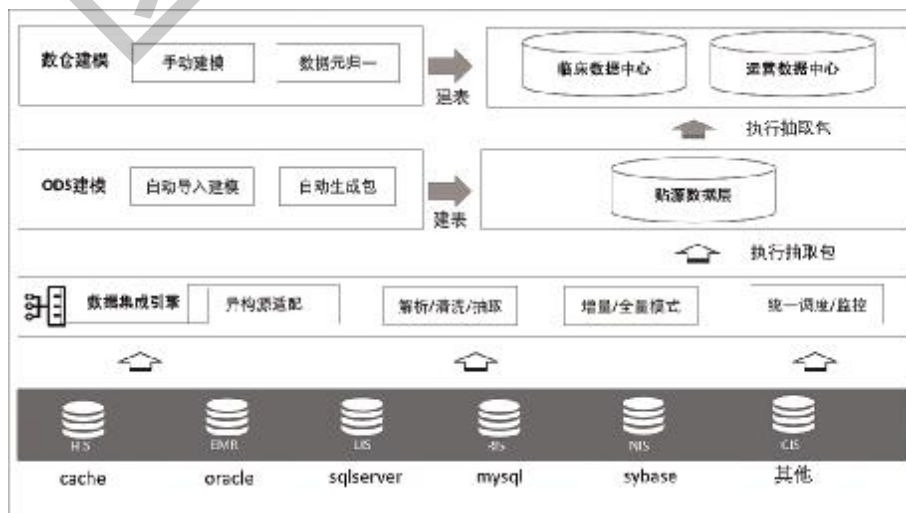


图 4 集成引擎架构示意图

2.3.2 数据可视化引擎 商业智能可视化技术在对医院抗菌药物应用、公立医院绩效考核管理等医院数据分析领域有广泛的应用^[15,16],但指标的统计分析往往会根据需求的变化而变化,DataOps 倡导的持续集成和部署正是解决该问题的有效方式,因此便于用户自助分析的可视化引擎必不可少。

可视化集成引擎采用 asp.net、bootstrap 等技术,汲取 Echart、FineRepor 等第三方可视化库开发而

成。其核心功能主要包括对柱形图、折线图等图例的封装,根据不同需求自定义展示模块,然后用不同模块自由设计数据分析报表。此外,引擎基于运营数据仓库实现可视化报表的自助设计,可以快速生成、方便操作,可根据医院、个人喜好生成各种仪表盘,见图 5。基于可视化数据分析平台,可以开发出适配电脑端和移动端的各种应用产品,这样就可以大大节约数据产品应用开发工作。

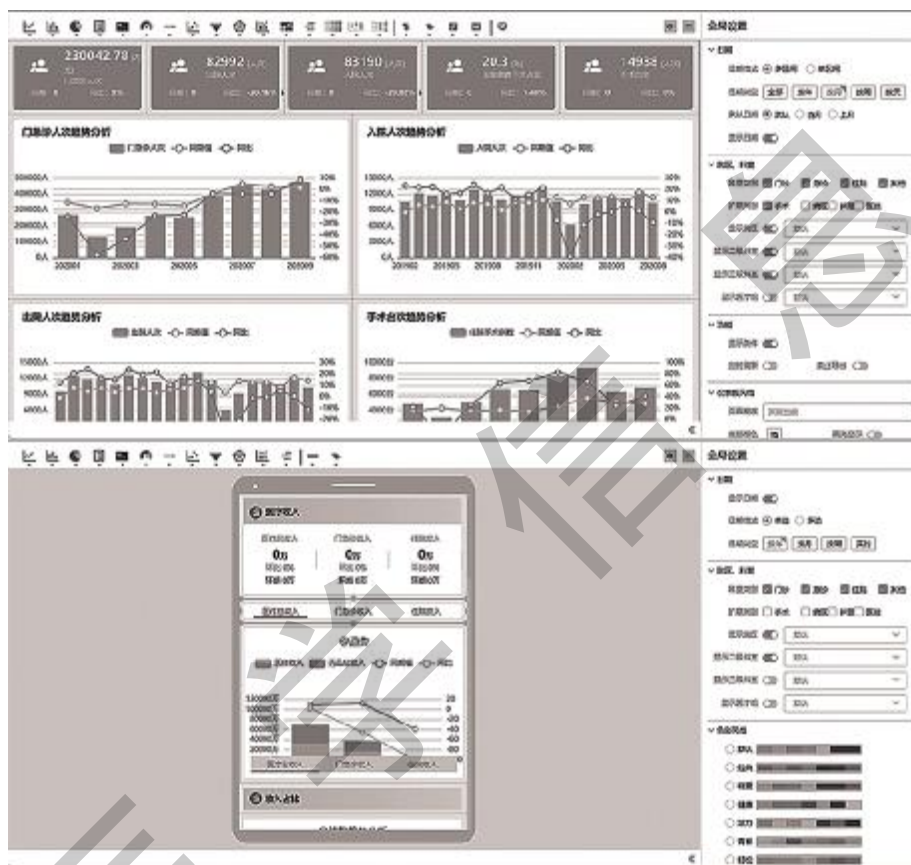


图 5 基于可视化引擎的仪表盘开发示意图

2.3.3 自然语言处理引擎 精准医学的核心是广泛收集患者个体相关的数据,其中电子病历数据是最重要的数据之一^[17]。以电子病历为主体的医疗大数据被广泛应用于医药研发、临床决策支持、个性化医疗、流行病监测和预警等^[18],上述数据应用的前提都需要经历非机构化的病历数据进行结构化的过程,因此自然语言处理不可或缺。自然语言处理引擎由算法标注平台、机器学习平台、自然语言处理平台、结构化抽取平台以及可视化展示 5 大板块构成,实现对电子病历非结构化数据的文档分类、段落分类、医学术语和实体属性关系识别,将临床数据中的文本段落信息转化成可以被医院数据分析项目有效利用的结构化标准数据。

2.3.4 数据治理引擎 数据治理可以有效地提高数据质量、共享信息和保护敏感数据,而且数据治理是贯穿数据整个生命周期。在 DataOps 概念中,数据治理可以作为开发、部署、操作和监控工作流的一部分,它是 DataOps 最重要的组成部分之一^[19]。数据治理引擎模块通过数据标准管理、元数据管理、数据质量管理、数据资产管理等功能实现数据标准规范化、数据关系脉络化、数据质量量化的目标,最终实现医院数据资产规范化管理,使得医院数据资产更加稳定,更可靠,可管理,更易懂,提高医院的信息化管理水平。

通过上述 4 个模块实现代码线上流转,构建了 CI/CT/CD 能力,支持自动化集成、部署和测试,支持

对数据全链路的监测和预警,建立全链路数据安全管控,实现面向 DataOps 的全周期的数据管理。

3 总结

当前,DataOps 实施是研究者重点关注的领域。Vargas-Rueda L 等^[3]分析了 DataOps 概念及其在数据管道中的实现,为了解和实施 DataOps 提供了指导。Demchenko Y^[20]采用敏捷服务和应用程序开发模式来快速响应市场需求和技术变化的情况,并提供了用于 DevOps 和 DataOps 的基于云服务的平台和工具的示例,介绍了通过云服务开发及基于云原生微服务的自动化配置加速 DevOps 软件开发实践。可以说,DataOps 的实践和推广已经没有了技术壁垒。

医疗行业内构建基于 DataOps 的临床数据中心平台,在产品功能、交付流程等方面也可以取得一定效果。产品功能方面,更加贴近用户需求,产品设计更加科学合理。如可视化的数据集成功能,提供了一体化的数据流水线作业平台,降低了交付门槛;数据可视化分析工具为客户自助分析提供了有效工具,增强了产品的可扩展性,也有利于拓展更多的数据应用。交付流程方面,构建数据开发运维一体化(DataOps)的数据开发交付运维范式,将敏捷、精益等理念融入数据开发交付过程,通过对数据相关人员、工具和流程的重塑,打破协作壁垒,构建新的数据开发、治理、运营、交付一体化的流水线,提高数据产品交付效率和质量。

此外,伴随医疗云平台的发展,医疗行业内的 DataOps 发展更将提速。受限于医疗机构的内外网分离的影响,当前临床数据中心平台在持续迭代、持续集成和持续发布交付的流程仍有很大提升空。基于医疗云平台的 DataOps 可以完全落实 DevOps 软件开发实践。

参考文献:

- [1]陆斌杰,孔宪明,翁子寒,等.上海仁济医院数据中心建设探索[J].中国数字医学,2010,5(8):71-73.
- [2]王强,易应萍.临床医疗大数据治理和应用[J].医学信息杂志,2018,39(8):1-6.
- [3]赖俊恺,廖茜雯,姚晨,等.临床研究中真实世界数据标准化的障碍以及建议:一项定性研究[J].英国医学杂志中文版,2022,25(11):640-646.
- [4]于家畦,康晓东,白程程,等.一种新的中文电子病历文本检索模型[J].计算机科学,2022,49(z1):32-38.
- [5]Sahoo PR,Premchand A.Dataops in manufacturing and utilities industries [J].International Journal of Applied Information

Systems,2019,12(6):1-6.

- [6]Bahaa S,Ghalwash AZ,Harb H.DataOps Lifecycle with a Case Study in Healthcare [J].International Journal of Advanced Computer Science and Applications,2023,14(1):136-144.
- [7]Liebmann L.3 reasons why DataOps is essential for big data success [J].IBM Big Data & Analytics Hub,Retrieved October,2014,28:2020.
- [8]Ereth J.DataOps - Towards a Definition[J].LWDA,2018,2191:104-112.
- [9]Atwal H.Practical DataOps:Delivering agile data science at scale[M].Berkeley:Apress,2019.
- [10]Vargas - Rueda L,Pongutá - Díaz S,González - Sanabria JS. DataOps,una alternativa para la gestión de Data Pipelines[J].Revista Ibérica de Sistemas e Tecnologías de Informação,2020 (E38):259-269.
- [11]Mainali K.DataOps:Towards understanding and defining data analytics approach[D].Stockholm:KTH Royal Institute of Technology,2020.
- [12]大数据技术标准推进委员会.数据资产管理实践白皮书(6.0版)[M].北京:现代出版社,2023.
- [13]马猷,陈丽,郝冀皖,等.利用 kubernetes 集群搭建基于容器技术的分布式架构数据中心研究[J].中国数字医学,2021,16(12):43-48.
- [14]Tamburri DA,Heuvel WJVD,Garriga M.DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching [C]//2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI).2020:391-394.
- [15]赵雨,李哲青,张波,等.基于 Power BI 的数据分析在三级公立医院绩效考核管理中的研究与应用[J].现代医院,2023,23(5):741-746.
- [16]钱进,张劲松.医院门诊抗菌药物应用管理中大数据可视化分析方法的应用和探索[J].中国医院药学杂志,2017,37(18):1850-1856.
- [17]韦玉芳,施维,尚于娟,等.基于电子病历数据的临床表型提取及其应用进展[J].医学信息学杂志,2017,38(8):6-10.
- [18]孟琳,马金刚,刘静,等.医疗大数据的应用与挑战[J].医疗卫生装备,2018,39(10):71-74,88.
- [19]Torre - Bastida AI,Gil G,Miñón R,et al.Technological Perspective of Data Governance in Data Space Ecosystems[M]//Data Spaces:Design,Deployment and Future Directions.Cham:Springer International Publishing,2022:65-87.
- [20]Demchenko Y.From DevOps to DataOps: Cloud based Software Development and Deployment [C]//The International Conference on High Performance Computing and Simulation (HPCS 2020).2020:10-14.

收稿日期:2023-09-20;修回日期:2023-10-09

编辑/成森