生物医学信息实验室数据安全计算环境设计与应用

徐志鹏[12]，李静[2]，戴启明[1]，裴浩宇[12]，徐永武[1]

(1.复旦大学人类表型组研究院张江复旦国际创新中心，上海 201203；2.上海国际人类表型组研究院，上海 200433)

摘要：由于生物医学信息实验室数据的敏感性，根据国家法律法规要求和现行的科研伦理要求，需要在支持科研活动的过程中，确保数据的安全可靠共享使用。通过分析现有生物医学数据安全共享技术不足，以及 DSMM 数据安全过程域设计要求，研究安全防护技术，设计生物医学信息实验室数据应用、数据防泄密、安全审计等方面诉求的安全计算环境。在标准沙箱技术基础上，结合生物医学信息要求以及实验室管理要求，设计包含动态资源调度、安全审计、流程审批等技术工具套件，支撑实验室数据安全共享，解决数据加密和高效计算这一两难问题，使实验室的数据使用符合安全能力成熟度模型和管理要求，提供具有高实用价值、高性价比的实验室数据基础设施。

关键词：安全沙箱；数据共享；信息安全；成熟度模型

中图分类号：TP311 **文献标识码：A**

Design and Application of a Secure Computing Environment in the Biomedical Informatics Laboratory

Xu Zhipeng[12], Li Jing[2], Dai Qiming[1], Pei Haoyu[2], Xu Yongwu[1]

(1. State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 201203; 2.International Human Phenome Institutes (Shanghai), Shanghai 200433)

Abstract: Owing to the sensitive nature of data within biomedical informatics labs, there's a critical need to ensure secure and compliant data sharing and utilization, abiding by national legal mandates and prevailing research ethics. An analysis of the existing data-sharing methodologies in these labs reveals significant security gaps. Following the DSMM (Data Security Maturity Model) design specifications, we investigated contemporary security defense mechanisms and crafted a secure computing environment that addresses specific requirements like data utilization, leakage prevention, and security auditing within the biomedical informatics setting. On the basis of

standard sandbox technology, combined with biomedical information requirements and laboratory management requirements, including dynamic resource scheduling, security audit, process approval and other technical tools suite is designed, all tailored to enhance secure data sharing within the lab. This integration effectively tackles the dual challenges of data encryption and optimized computation. With this fortified data computing environment in place, the lab's data operations are aligned with the leakage prevention criteria of the DSMM, ensuring a robust and cost-effective data infrastructure.

Keywords: security sandbox; data sharing; information security; maturity model

1 引言

目前在各行各领域中，数字化转型及大数据应用的建设，已经成为信息化向数字化转型的关键路径，从数据中挖掘价值，为场景提供数字化解决方案，通过互通互联方式整合多方资源，以提高资产价值，已经成为发展趋势。生物医学信息科学研究也进入数据驱动的新范式时代，但生物医学领域大量牵涉个人信息、人类遗传资源信息等重要或敏感的数据信息，具有更高的合规要求。

数字化转型过程中如何保障数据安全是重中之重，约束条件包括隐私计算、数据安全等多方面；数据管理包括数据来源可确认、使用范围可界定、流通过程可追溯等；安全防护框架，态势感知及主动防御能力都需要可信计算核心技术作为支撑。所以，在保护数据安全的同时，又能提供高效的计算能力，需要创新数据服务模式来形成可行性路径。

2 系统设计

2020 年实施的 GB/T 37988-2019 《信息安全技术数据安全能力成熟度模型》，提出了数据安全保护体系，本文依据体系的成熟度模型提出了利用安全计算沙箱、隐私计算、数据汇交等技术，实现数据可用不可带出，解决数据信任和隐私保护、溯源等利用难题，让数据安全地流动起来。

本文在提出数据可用不可带出模式的同时，基于安全沙箱技术，自主研发安全管理套件，构建了完整的安全计算环境，实现数据从采集、访问、传输和流转的全流程闭环管控，解决生物医学数据分析和应用上遇到的风险问题。

2.1 研究思路

本方案从数据共享及控制限制方面的不足，只能实现部分数据安全为出发点，利用沙箱技术为基础能力，优化包括存储管理方式、网络访问控制和数据流转控制等关键流程，解决数据安全问题[1]。

传统沙箱技术通过在计算机系统中创建一个隔离的环境，让应用程序或进程在其中运行，但其访问系统和敏感资源被严格限制。这种隔离性使得应用程序内的潜在威胁无法波及服务器的其他部分，从而确保服务器数据和系统的安全，实现了“防外不防内”的要求。但此安全机制中，却不能解决数据利用中常见数据防泄密的问题。研究人员之间的合作研究必须克服数据泄露造成隐私潜在风险，以及将敏感数据转移到外国组织的潜在风险。

2.2 解决方案

本方案通过如下纬度保障数据安全：

（1）匿名化和脱敏：在数据沙箱中，数据会经过匿名化和脱敏处理，以避免直接暴露个人敏感信息。这样可以确保数据在共享过程中不会泄露隐私[2]。

（2）访问控制：数据沙箱可以实施严格的访问控制策略，只允许经过授权的用户或实体访问特定数据，只有合法用户才能获取数据[3]。

（3）数据分割：数据沙箱可以将数据分割成不同的部分，使得不同用户只能访问特定的数据片段。这种分割可以减少数据泄露的风险[4]。

2.3 设计原则

1）设计原则：不能见(分级分类，授权可见)，不能取(数据可控，不出沙箱)，不能毁(储存加密，备份留存)，不能赖(审计查询，操作留痕)。

2）设计流程

（1）审批流程贯通整个数据使用过程，并根据实际需要进行灵活设置。包括：数据使用的权限审批；安全个性化分析软件的审批；分析结果导出安全环境的审批；同时提供操作日志用于审计和查询。

（2）面向生物医学信息提供计算工具，如 Python、R、Matlab 等，可以满足用户不同的计算需求，提供可视化操作界面。

（3）根据实际需要申请计算资源，提供独立用户在超级计算集群上的配置工作信息。

3 平台操作

3.1 用户操作流程

1）权限获取：用户向管理员申请从而开通个人沙箱账号，用户从外部网络通过堡垒机的交叉授权获得沙箱环境内的各个运算主机访问权限。

2）环境搭建：在堡垒机系统内，用户可在超算集群上选择自己使用的沙箱系统，沙箱计算环境中包含多个配置的沙箱系统，包括 GUI sandbox、Computing sandbox、HPC compatible environment sandbox、GPU node sandbox。其中用户可在 HPC compatible environment sandbox、GPU node sandbox 中进行软件环境的搭建工作，例如自定义运算工具的安装，用于自己特殊的科研需求。用户可以将自己的实验数据导入沙箱环境与沙箱环境中的原始数据一起运算，整个环境搭建流程是处于联网状态。

3）数据计算：实验数据储存在沙箱环境中，用户搭建好环境，确定好所需的运算资源后，通知管理员，管理员将沙箱环境断网，调配好运算资源，用户开始运算。管理员可根据用户对算力的要求通过堡垒机系统外的 PBS 运算调动系统增减虚拟机数目灵活配置沙箱系统的算力资源。利用沙箱的用户在使用同一沙箱系统时相互隔离并且运行独立，每个用户根据需求由管理员调度不同的计算资源。

4）结果导出：管理员可以通过堡垒机对所有数据用户操作进行管理，包括对登录用户的身份进行审核、监控用户在沙箱系统中的操作、确认用户在沙箱系统中导出的数据结果是否违反规定。

5）操作审计：平台设置有独立的操作审计员，审计员具有审计账号，但没有平台操作权限或管理员权限，即只看不能操作。审计员可以看到平台上用户的操作，并进行审计，避免违规的数据使用和操作。

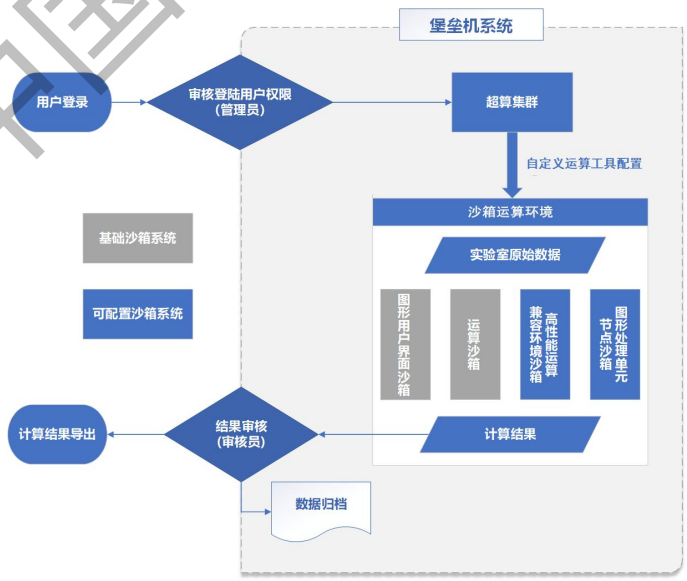

图 1　平台操作流程图

3.2 用户操作界面

1）登录沙箱

整个操作流程基于浏览器操作，并以集群账号进行操作。

图 2　沙箱登录操作展示图

2）沙箱信息界面

打开主文件夹后即可找到 sandbox_link 的快捷连接（进而找到存储在沙箱中的数据）。

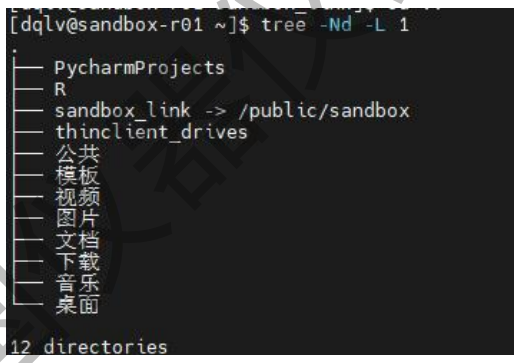

图 3　沙箱信息展示图

3）沙箱软件操作

沙箱中初始安装了数据分析常用的软件，文本编辑常用软件等，方便用户操作。

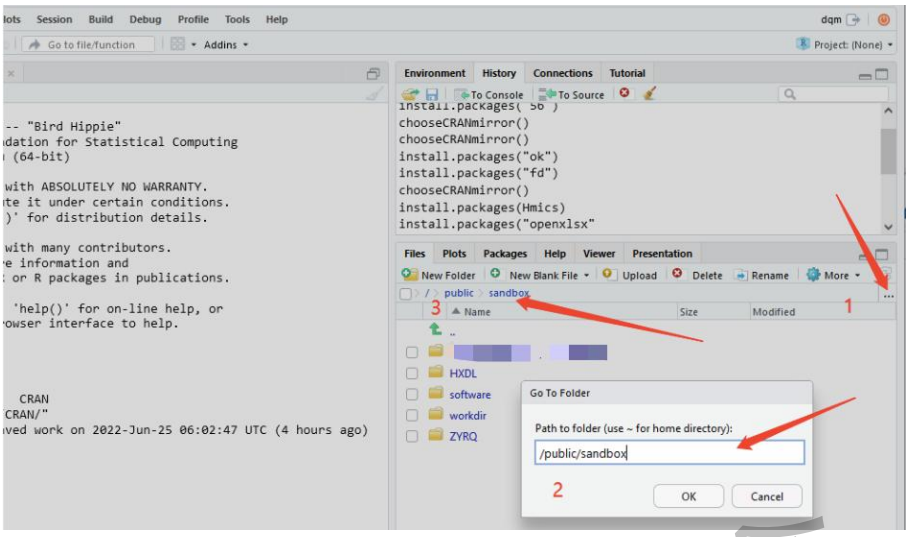

图 4　沙箱预装公共软件展示图

4 小结

本方案为科研人员利用生物医学信息数据提供了指导，并论证了安全计算环境的机制，在控制数据泄漏的风险和保护研究参与者的隐私方面起着重要的作用。

截至 2022 年 6 月，利用本方案搭建的安全计算环境，已有 116 位注册用户，完成 150 次运算结果导出，导出结果数据容量达到 85G，能够顺利同时支持多项科研工作的开展。结果表明，本方案设计的生物医学信息实验室数据安全计算环境在保证安全性的同时，具有较高的可用性.

参考文献：

[1]王忠春,陈庆荣,刘婷. 大数据下新型安全沙箱技术运用分析与研究[J].网络空间安全, 2020,013(006):89-97.

[2]黄莹. 人工智能时代下的隐私权保护问题研究[J]. 文化创新比较研究, 2020(012):86-87.

[3]万兆泽,刘尚焱. 数据网络的安全 [J]. 计算机科学 , 2010(003):10-13.

[4]吴超."互联网+医疗"信息安全问题及对策[J].集成电路应用, 2019(3):17-20.