 DOI: https://doi.org/10.12349/iser.v6i8.7028

Data desensitization and re-identification risk prevention mechanism under the information innovation platform

Lanbo Chen

China Electronics Technology Group (Taiji Computer Co., Ltd.), Beijing, 100000, China

Abstract

With the application of Xinchuang platform in finance, government affairs, power and other industry departments, ensuring data desensitization of important and sensitive information has become an important need. However, due to the immature development of the domestic software and hardware ecosystem, there are certain defects in the matching degree of desensitization methods, dynamic protection and audit traceback in the information and innovation environment, so that there may be the risk of data re-identification in complex business scenarios. Based on this, the following analyzes the types of data desensitization technology, as well as the problems and optimization strategies of data desensitization technology under the information innovation platform at this stage, and proposes a risk prevention mechanism under the information innovation platform, hoping to lay a methodological foundation for later technology upgrades and compliance reviews.

Keywords

information innovation platform; data desensitization; re-identify risks; guard against

信创平台下数据脱敏与重识别风险防范机制

陈兰波

中国电子科技集团（太极计算机股份有限公司），中国 · 北京 100000

摘　要

伴随着信创平台在金融、政务、电力等行业部门的应用，保障重要敏感信息的数据脱敏已成为重要需求。但由于目前国产软硬件生态发展不成熟，在信创环境下脱敏手段的匹配程度差、动态防护及审计回溯都存在一定缺陷，以至于在复杂业务场景下数据还可能会出现重识别的风险。基于此，下文针对信创环境下的数据安全需求，分析数据脱敏技术类型，以及现阶段信创平台下数据脱敏技术存在的问题与优化策略，并提出信创平台下重识别风险防范机制，希望能为后期技术升级与合规审查奠定方法论基础。

关键词

信创平台；数据脱敏；重识别风险；防范

1 引言

信创战略的持续推进，使得由国产软硬件组成，并依赖于自主可控环境作为运行底座的大批信创产品列装，成为关键信息系统的运行基础。一旦遇到大规模交互便会暴露大量敏感信息，仅靠通过静态规则进行数据脱敏的传统方式无法满足高并发、多种复杂角色访问以及跨系统的高要求。并且随着攻击手段的升级，即使已经进行过初步脱敏的数据集仍可以通过属性相关、外部分配等手段进行重新识别而暴露内容。所以必须要按照信创环境下特有的技术特点打造适应性强、实时更新，可追踪性高的脱敏及防护体系来确保整个信创环境下敏感数据安全。

【作者简介】陈兰波（1972–），男，中国北京人，硕士，从事软件工程与信息工程监理方向的研究。

2 数据脱敏技术分类

2.1 静态脱敏

静态脱敏是针对存量数据做的一种处理，通过替换、置换、泛化或扰动数据库或文件中敏感字段，是其在共享或者使用环节不具有直接敏感性。这种方法主要是用于数据分发或者构建测试环境以及在外部提供数据的时候使用，通过形成和真实数据结构相同，却不含敏感字的信息集合，让业务系统在良好的状态下运行，且不会出现重要数据的泄露[1]。在实际应用环节静态脱敏要同时考虑数据一致性以及性能的损耗，防止由于不合理设计而引发数据分布异常，降低业务逻辑验证的有效性。

2.2 动态脱敏

动态脱敏的原理是利用中间层拦截与策略引擎执行的方式来实现数据访问过程的实时脱敏。当授权用户向系统发出查询指令后，系统根据其访问身份与权限规则将输出结果

进行过滤、遮挡、替换等操作，在维持底层数据库存储结构的前提下进行差异化脱敏控制，不但能够有效避免数据泄露或者滥用，还不会过多干扰核心生产系统，因此在金融、政务、医疗等多角色、多层次的应用场景下十分适用。但是动态脱敏会显著提升系统负载能力，因此对系统架构有很高要求，需要构建稳定的高并发处理能力，并且还要精细化管控策略配置与权限的划分。

2.3 加密衍生

加密衍生主要是利用密码学原理，把包含着一些需要保护的敏感字段信息用不可逆或受控可逆的方式经过加密算法转换为其它信息的表示形式。在保证了数据安全的基础上可以通过密钥管理体系实现数据的解密和还原。因为不同的场景下对于解密的需求也是不同的，所以加密算法的选择、密钥分级以及更新，都将成为其重要的技术要点。

2.4 隐私模型

隐私模型是指基于统计学和数学建模的技术手段降低数据被重识别风险的方法，主要的方式包括 k- 匿名、l- 多样性、t- 接近性和差分隐私等技术。以上方法均通过对数据集中的敏感属性去标识化，或者向数据中添加噪声，以此来减少攻击者凭借背景知识或者多源信息关联推断出准确数据的几率。信创环境下，隐私模型常被用在数据共享开放、大数据分析或 AI 训练过程中，用于调和数据可用性与隐私保护之间的关系。隐私模型在设计过程中往往会根据业务要求以及数据分布情况选择不同的参数值，以避免数据价值损耗过大或保护强度不足 [2]。

3 信创平台下数据脱敏技术存在的问题

3.1 适配性差

在信创平台下，软硬件基础架构具备了自主可控的能力，但与之相匹配的兼容性及生态尚不够成熟，导致了目前大部分数据脱敏工具部署到信创平台上无法很好的适配。一方面由于不同数据库内核有不同的索引方式、存储方式以及的执行引擎，所以很难把某一套脱敏规则应用在不同类型的数据库上。另一方面由于一部分脱敏算法是基于脱敏对象本身的执行，在我国相关国产化操作系统的框架中，或者一些国产化中间件中运行比较慢，甚至有些存在运行卡顿的现象，消耗较多的机器资源，容易造成闲置。在不同平台间，无法做到完全一致的数据格式转换，也难以做到对应的接口协议标准一致，会造成规则失效及数据一致性降低的问题，导致脱敏技术在信创平台下较难形成高效的适配能力，安全防护也会因此而受到影响。

3.2 动态防护不足

信创平台对于数据的实时访问有着很大的需求，对于高并发的金融、电力、政务等场景，用户的访问行为较为复杂，也在不断发生变化。在这种情况下，现有的动态脱敏手段只能做到通过少量静态规则进行配置，但是无法进行自适应式的访问方式的管控，不能够做到在一个比较复杂的场景下达到一种灵活的防护的目的。比如当有多角色的用户同时访问相同的敏感数据的时候，系统无法根据实时上下文动态调整脱敏强度和范围，导致有些结果输出之后仍然存在一些比较敏感的信息 [3]。此外还有一部分脱敏的中间件本身就不具备在高并发情况下能够满足业务需求的能力，使得它的防护的效果降低。更重要的原因还在于动态脱敏与入侵检测、异常访问识别等相关联的功能融合度较差，难以针对用户访问行为进行多层次的实时防护。

3.3 审计缺失

数据脱敏实施过程中，审计机制缺陷是一个突出的问题。目前大部分已有系统的侧重点在于规则配置和执行效果上，并没有很好的考虑到在执行过程中的可回溯性和行为留痕等问题，在发现数据泄露或者重识别攻击之后很难第一时间恢复出访问路径以及脱敏处理流程，给事件应急处理带来困难。尤其是信创情况下，

监管部门的数据安全合规性严苛，如果没有一个完整的审计链条，可能导致合规风险和责任界定模糊。另外一些平台的日志采集不够完整、日志存储安全性较差、审计联动跨系统不畅等，也可能导致日志缺失、日志篡改或者因为脱敏规则的变化而无法获得实时追踪，不但不利于后期数据治理的实现，还会降低脱敏技术运行状态的长期监控能力，难以形成完善的安全闭环。

4 优化信创平台下数据脱敏技术的策略

4.1 脱敏技术与信创硬件的协同优化

信创平台下脱敏技术在软硬件兼容性、生态成熟度上依然存在一定不足，因此需要依靠技术适配、体系联动来进一步完善。一是为国产处理器和操作系统提供国产化产品适配开发接口；二是对脱敏算法进行优化，在 CPU 指令集、内存调度方式、文件系统等方面开展深入适配，降低由于调用机制的不同所引起的性能损耗；三是推动数据库厂家和脱敏工具厂商共同配合开发出符合信创平台要求的接口适配层，采用抽象化中间件适配框架统一索引调用、执行计划及事务处理逻辑，实现不同的数据库内核下脱敏规则可迁移；四是在跨平台数据交互场景中采用规范的数据接口协议和格式转换模块，保证脱敏处理的一致性和稳定性，防止由于数据结构的不同引起规则失效。

4.2 构建信息熵的风险量化模型

针对信创平台动态防护能力不足的问题，应用信息熵理论建立风险量化模型，提高实时场景下脱敏机制的灵活度和精确度。第一，对于多角色并发访问的数据，系统可根据用户行为特征、访问频次、上下文依赖关系等信息计算出信息熵值，进而判断该数据在不同访问路径下的暴露程度。当该数据的信息熵值大于阈值时，就可利用动态脱敏引擎自动增强脱敏强度，通过脱敏粒度的自适应调整实现更高效的数

据保护。第二，借助信息熵模型，可将访问方式的变化转换成量化的风险指标，并根据风险指标的不同数值为敏感信息的输出设置相应的脱敏动态阈值，从而使静态规则无法解决的问题迎刃而解。第三，与入侵检测、异常行为分析等功能安全模块结合，运用信息熵模型可基于熵值判定是否存在异常访问链路，既能够实现数据脱敏，又能起到行为监测作用，能够形成数据脱敏和行为监控双管齐下态势。

4.3 零信任审计的透明日志注册表设计

针对信创平台下的脱敏过程无审计链路的问题，可引入零信任理念，搭建基于透明日志注册表的审计链路，实现实时、完整的审计并满足法律法规要求。第一，零信任思维强调“永不默认信任”，在审计过程中可通过多源身份核验以及对脱敏操作的签名控制保证每次脱敏都有责任人，确保访问行为可控。第二，

透明日志注册表以链的形式存储脱敏策略变更，规则调用、数据访问路径等信息，只要生成日志条目，就无法被篡改，便于之后溯源分析时查证。第三，为解决跨系统跨数据库环境下复杂审计的问题，可以统一制定日志标准接口，不同模块的日志数据自动汇集并进行多种方式的交叉比对验证。对于彼此之间相对孤立的记录可能会出现的追踪断点的问题，通过在日志中加入时间戳、哈希值、加 / 解密校验等方法，可以很好的弥补这一问题带来的安全隐患，在保障存储的同时实时查看和监控。透明日志注册表作为审计日志在监管端的体现，除了实现数据的真实留痕，也将为日后有关行业管理部门的监督检查提供可靠的监管抓手。

5 信创平台下重识别风险防范机制

5.1 基于信息熵的脆弱性建模

在信创平台敏感数据保护框架下，可以利用信息熵来进行重识别风险量化建模的实践探索。信息熵是刻画不确定性的特征量之一，可以通过分析数据分布的均匀性得到数据在不同维度上的暴露情况。对单一属性字段而言，可以通过该字段取值概率分布推导得到其熵值大小，进而判断该属性字段的信息强度，并据此来确定在脱敏环节中的扰动或泛化处理强度大小。若针对多属性联合场景分析，则需要考虑条件熵、联合熵的变化趋势，找到组合状态下可能出现的一些隐匿关联链路，从而判断数据集整体面临的安全威胁状态。除此之外，建模的过程中需要将数据集进行熵值变换后的风险矩阵划分成不同的敏感程度等级，设置阈值驱动的风险监控机制，在属性分布发生变化时实时更新风险程度值。

5.2 实施重识别风险动态评估

在建设信创平台的数据安全体系过程中，要求综合考量多维度参数来动态量化重识别风险的动态评估机制。第一，从数据属性分布、访问行为特征及外部辅助信息中选取关键性指标并联合作用构建联合评估模型，然后通过观测数据应用频率以及访问路径等属性变量是否存在偏差，判断是否存在潜在重识别风险。第二，采用时序分析技术，把用户的交互过程中行为的变化纳入到风险评估当中，通过观察并判断出趋势的变化幅度以及是否存在异动的现象，来动态调整脱敏的等级。第三，风险评估模型还需要与其他的平台审计日志、会话监控等功能模块进行对接，并以多方数据为依据来判定风险水平，以防出现因为单一因素而得出结论的情况发生。

5.3 绘制信创场景风险热力图

重识别风险的空间化表达是指利用风险热力图直观表达出不同的业务场景下的薄弱区域（比如交易数字最为集中的位置），高危节点（比如高并发多业务场景汇聚点、跨库联合查询或多接口调用入口）。在实际应用过程中，需根据信息剩余的数据量以及关联性构建风险指标体系，并以权重映射到业务流程每一个环节。通过分析访问日志、交互流量和角色权限的内容构建出风险数值矩阵，再将数值矩阵的内容转化为可视化的热力图。工作人员能够根据风险热力图的颜色强度代表风险发生的频率高低，来判断诸如高并发交易、跨库联合查询或外部接口调用等不同应用对于风险发生程度大小的不同影响。

6 结语

信创平台下的数据脱敏与重识别防范机制是技术层面的创新任务，同时也是数据安全治理体系的一个重要环节。通过协同优化脱敏算法与国产软硬件、建立信息熵驱动的脆弱性模型、引入动态风险评估与透明审计框架，并辅以风险热力图的可视化表达，可以逐步形成覆盖全链路的立体化防控体系。以此能够弥合现有脱敏工具与信创平台间存在的差异问题，为今后的敏感数据安全增加一道防线，也为有关行业未来的制度化与规范化经营，以及与时俱进的技术迭代进化提供正确的发展思路。

参考文献

[1] 刘圣龙,黄秀丽,江伊雯,等.面向多方数据融合分析的隐私计算技术综述[J].网络与信息安全学报, 2024, 10(6):24-36.

[2] 龚安.信创平台数据备份与恢复系统[C]//2021年国家网络安全宣传周网络安全产业发展论坛.中国网络安全产业联盟中国电子技术标准化研究院, 2021.

[3] 付威.信创领域的数据库应用及创新[J].软件和信息服务（原：软件世界）, 2021, 000(8):2.