大数据应用中的数据安全治理技术

李　军
新疆民航通信网络有限责任公司　新疆　乌鲁木齐　830016

摘　要：在数字经济高速发展背景下，大数据应用的安全风险对数据全生命周期保护提出严峻挑战。本文系统分析数据采集、存储、处理及共享阶段的安全风险，从技术维度阐释数据加密、访问控制、数据审计与溯源技术及数据安全态势感知安全等关键治理技术的原理与实现路径。研究表明要构建技术、管理、法律协同的治理体系，融合对称与非对称加密优势，结合区块链溯源与AI态势感知技术，平衡数据利用与安全保护，为大数据产业健康发展提供理论与技术支撑。

关键词：大数据应用；数据安全治理；关键技术

引言：随着5G、物联网技术普及，大数据已成为驱动产业变革的核心要素。其数据量大、类型多样等特征在释放价值的同时，也因采集设备漏洞、存储介质风险等引发敏感信息泄露隐患。当前数据安全治理面临技术架构复杂性与跨域协同难题，传统单点防护技术难以应对全生命周期风险。本文立足大数据应用场景，剖析安全风险本质，探索以加密技术为基础、访问控制为核心、态势感知为支撑的治理技术体系，为构建新型数据安全防护框架提供思路。

1 大数据应用与数据安全治理概述

在数字经济蓬勃发展的当下，大数据应用凭借其独特的特点与迅猛的发展趋势，成为推动各行业变革的核心力量。大数据具有数据量大、类型多样、处理速度快和价值密度低的显著特点。从海量的用户行为数据到复杂的物联网传感数据，大数据的类型涵盖结构化、半结构化和非结构化数据；而实时分析与处理的需求，要求数据处理速度不断提升；虽然数据价值密度低，但通过深度挖掘，能为企业决策、社会治理提供关键洞察。

数据安全治理是指通过建立完善的管理体系、技术手段和操作流程，确保数据在全生命周期内的保密性、完整性和可用性，实现数据价值的安全释放。其核心目标在于平衡数据利用与安全保护的关系，既要保障数据能够被合理、高效地使用，推动业务发展，又要防范数据泄露、篡改等安全风险，维护个人隐私、企业利益和国家安全。

在大数据应用中，数据安全治理重要性主要。大数据应用涉及大量敏感信息，如个人身份信息、企业商业机密等，一旦发生安全事故，将导致严重的后果。数据安全是大数据产业健康可持续发展的基石。只有保障数据安全，才能增强用户与企业对大数据应用的信任，促进数据的开放共享与深度应用，充分释放大数据的价值[1]。强化数据安全治理，是应对大数据时代安全挑战、推动数字经济高质量发展的必然选择。

2 大数据应用中数据安全风险分析

2.1 数据采集阶段的安全风险

数据采集阶段，数据来源广泛且类型复杂，安全风险主要体现在以下数据源真实性与完整性难以保障、数据采集设备存在漏洞以及采集权限管控不严三个方面。（1）部分数据源可能被恶意篡改或伪造，如网络爬虫获取的网页数据可能被注入虚假信息，导致后续数据处理与分析的结果失真。（2）物联网设备、传感器等数据采集终端往往存在硬件或软件漏洞，易被黑客攻击控制，进而上传错误数据或窃取采集权限，破坏数据采集的正常秩序。（3）若缺乏严格的数据采集权限管理，可能导致越权采集敏感数据，侵犯用户隐私与企业权益。

2.2 数据存储阶段的安全风险

数据存储阶段的安全风险主要集中在以下存储介质安全、数据冗余备份隐患以及存储系统漏洞三个维度。（1）存储介质本身存在物理损坏、丢失或被盗取的风险，如硬盘故障、U盘遗失等，可能造成数据永久性丢失。（2）为保障数据可用性，企业通常会进行冗余备份，但过多的备份副本若未妥善管理，容易出现备份数据泄露、版本混乱等问题。（3）数据库管理系统、云存储平台等存储系统若存在未修复的安全漏洞，如SQL注入漏洞、弱密码漏洞，黑客可利用这些漏洞非法访问、篡改或删除存储的数据，严重威胁数据的保密性、完整性与可用性。

2.3 数据处理与分析阶段的安全风险

在数据处理与分析过程中，以下算法安全、计算环境安全以及数据访问控制不足是主要风险来源。（1）

部分数据处理算法可能存在逻辑缺陷或后门，在数据挖掘、机器学习模型训练等操作中，导致敏感信息泄露或模型被恶意攻击。（2）计算环境的安全性也至关重要，分布式计算框架、云计算平台等若存在安全配置错误或遭受恶意程序入侵，将影响数据处理结果的准确性与安全性。（3）若对数据处理与分析人员的权限管理松散，内部人员可能滥用权限，非法获取、篡改数据，造成数据泄露与业务损失。

2.4 数据共享与传输阶段的安全风险

数据共享与传输是大数据应用实现价值的关键环节，但也面临以下诸多安全风险。（1）数据在传输过程中，若未采用加密或加密强度不足的传输协议，数据易被窃取、篡改，如在公共Wi-Fi环境下，黑客可通过网络嗅探获取未加密传输的数据。（2）跨组织、跨平台的数据共享缺乏统一的安全标准与监管机制，接收方可能违反约定使用数据，或因自身安全防护能力不足导致数据泄露。（3）数据在共享传输时，若未对数据的使用范围、使用期限等进行严格限制，可能造成数据被过度使用或长期留存，增加数据泄露风险[2]。

3 大数据应用中的数据安全治理关键技术

3.1 数据加密技术

数据加密技术通过数学算法将原始明文数据转换为密文，只有掌握对应密钥的授权主体才能将其还原为原始数据。在大数据场景下，数据加密技术的实现依赖以下对称加密、非对称加密和哈希加密三类基础算法。（1）对称加密算法采用单一密钥完成加密和解密过程，以AES和DES为典型代表。AES作为新一代高级加密标准，支持128位、192位和256位等多种密钥长度，通过字节替换、行移位、列混合等轮函数操作，实现对数据块的高强度加密。其采用的Rijndael算法结构具备良好的并行计算特性，在GPU加速环境下可实现每秒数GB级别的数据加密处理，适用于大数据批量存储与传输场景。而DES算法虽已逐渐被AES取代，但其Feistel网络结构奠定了分组密码的设计基础，通过16轮迭代运算实现数据混淆与扩散。（2）非对称加密算法基于数学难题的计算复杂性，构建公钥与私钥的密钥对体系。RSA算法基于大整数因式分解难题，通过生成两个大素数乘积作为模值，构建公钥指数与私钥指数的数学关系，实现加密与解密操作。ECC算法则依托椭圆曲线离散对数问题，在相同安全强度下，其密钥长度仅为RSA的1/4，具备更高的计算效率与带宽利用率。（3）哈希加密算法通过单向散列函数将任意长度数据映射为固定长度哈希值，其核心特性包括雪崩效应与抗碰撞性。SHA-256算法作为SHA-2系列的典型代表，采用模块化迭代结构，通过逻辑运算、移位操作与常数相加等步骤，对512位数据块进行处理，生成256位哈希值。哈希加密在数据完整性校验中不可或缺，常与数字证书结合用于文件校验、区块链共识机制等场景。

3.2 访问控制技术

访问控制技术通过构建权限管理体系，实现对数据资源的细粒度访问控制。其核心模型包括以下自主访问控制（DAC）、强制访问控制（MAC）和基于角色的访问控制（RBAC），以及新兴的基于属性的访问控制（ABAC）。（1）DAC模型基于所有者自主授权原则，在文件系统中表现为ACL（访问控制列表）机制，每个数据对象关联一个ACL列表，记录授权用户及其权限。其实现依赖Unix/Linux系统中的chmod命令与Windows系统的安全描述符，支持读、写、执行等基本权限操作。但DAC模型存在安全缺陷，当所有者权限被非法获取时，易导致权限滥用，且缺乏统一的权限管理策略。（2）MAC模型采用强制安全标签机制，将主体与客体划分为绝密、机密、秘密等安全级别，结合范畴标签构建多维安全属性。系统依据“上读下写”原则进行权限决策，即高安全级主体只能读取低安全级客体数据，低安全级主体只能写入高安全级客体。（3）RBAC模型通过角色作为权限载体，将用户与权限解耦。其核心组件包括用户、角色、权限和会话，通过角色层次结构与角色激活机制实现权限复用与动态控制。在RBAC96模型中，定义了RBAC0（基础模型）、RBAC1（角色层次）、RBAC2（约束模型）和RBAC3（统一模型），支持互斥角色、基数约束等复杂权限管理策略。（4）ABAC模型突破传统主体-客体二元结构，将用户属性、环境属性（如时间、IP地址）和资源属性纳入权限决策要素[3]。基于XACML（可扩展访问控制标记语言）实现策略描述，通过属性-值对匹配进行权限判定，支持基于风险的动态访问控制，可根据实时安全态势调整权限策略。

3.3 数据脱敏技术

数据脱敏技术实现涵盖以下静态脱敏与动态脱敏两种模式，并结合多种脱敏算法与新兴隐私保护技术，构建全场景数据保护体系。（1）静态脱敏技术在数据使用前进行一次性处理。数据屏蔽技术通过正则表达式匹配敏感字段，如将身份证号后四位替换为“****”，支持通配符与分组捕获功能；数据泛化采用聚类算法对精确数据进行抽象，如将经纬度坐标映射至地理区域；数据替换通过建立替换字典，将真实数据替换为伪造但符合分布特征的数据，可采用K-Means聚类算法生成相似数

据集。（2）动态脱敏技术在数据访问时实时执行脱敏策略。基于查询重写技术，在SQL语句执行前解析语义，对敏感字段添加脱敏函数，如对SELECT语句中的手机号字段应用掩码函数。动态脱敏支持多维度策略配置，可根据用户角色、访问时间、IP地址等条件触发不同脱敏规则。差分隐私技术通过向数据添加拉普拉斯或高斯噪声，在数学层面保证个体数据不可区分性，其ε-差分隐私模型通过控制噪声强度实现隐私预算管理，确保数据可用性与隐私保护的平衡。同态加密技术则在密文状态下直接进行数据处理，无需解密即可完成统计分析，为数据脱敏提供了全新的技术路径。

3.4 数据审计与溯源技术

数据审计与溯源技术通过构建操作追踪与数据流转记录体系，实现安全事件的事后追责与风险预警。（1）数据审计系统采用旁路镜像或探针采集技术，实时捕获数据库操作语句、文件访问日志等数据。操作记录模块通过解析SQL语法树，提取操作类型、表名、字段名等信息，采用二进制日志解析技术实现全量操作记录。异常检测模块基于机器学习构建基线模型，采用孤立森林算法识别离群操作，通过时序分析检测异常操作序列。合规性检查模块依据GDPR、等保2.0等法规要求，对数据操作进行合规性验证，支持策略模板库与规则引擎动态配置。（2）数据溯源技术在大数据环境下面临多源异构数据融合与跨域追踪难题。区块链技术通过分布式账本记录数据操作，每个区块包含前一区块哈希值、操作记录与时间戳，形成不可篡改的操作链。智能合约实现溯源规则自动化执行，支持数据操作权限验证与来源追溯。数字水印技术通过离散余弦变换（DCT）或奇异值分解（SVD）在数据中嵌入鲁棒性水印，水印信息包含数据来源标识与操作日志索引。

3.5 数据安全态势感知技术

数据安全态势感知技术通过多源数据融合分析，实现安全状态的实时监测与风险预测。其技术架构包含以下数据采集、智能分析与可视化展示三个核心模块，并融入AI技术提升预测能力。（1）数据采集模块采用分布式日志采集框架（如Flume、FileBeat）实现异构数据源接入，支持Syslog、JSON、XML等多种格式。通过流量镜像技术捕获网络数据，利用NetFlow协议提取会话特征。（2）分析处理模块构建多层级分析架构：数据预处理层采用正则表达式、词法分析进行数据清洗；特征提取层通过TF-IDF、词向量模型提取安全特征；关联分析层采用贝叶斯网络、D-S证据理论融合多源告警信息；异常检测层基于深度学习构建LSTM、GAN模型，识别未知攻击模式。（3）态势展示模块采用D3.js、ECharts等可视化框架，构建三维拓扑图、热力图等展示界面。风险评估模型结合CVSS评分与攻击路径分析，量化安全风险等级[4]。预测分析模块基于时间序列预测算法（如ARIMA、Prophet）预测安全事件发生概率，利用因果推理技术分析风险传播路径。

结束语：大数据安全治理需技术创新与管理优化双轮驱动。研究证实，数据加密与脱敏技术可实现敏感信息防护，区块链溯源与AI态势感知能提升风险预警能力。未来应聚焦跨域数据共享安全标准制定，推动联邦学习、隐私计算等技术融合，建立动态自适应治理模型。强化法律合规与人才培养，形成技术、管理、法律三位一体的治理生态，为数字经济高质量发展筑牢安全基石。

参考文献

[1]高磊,赵章界,宋劲松,等.大数据应用中的数据安全治理技术与实践[J].信息安全研究,2022,8(4):326-332.

[2]张万里.大数据应用中数据安全治理技术研究[J].信息系统工程,2023(11):125-128.

[3]贾若飞.大数据应用中数据安全治理技术研究[J].中国设备工程,2023(2):26-28.

[4]方子诚.大数据技术在数据安全治理中的应用[J].中国宽带,2023,19(10):131-133.