大数据技术在网络安全风险识别与防控体系构建研究

幸天奕
广东外语外贸大学，广东 广州 510006
DOI:10.61369/SE.2026050011

摘　　要： 大数据技术在提升数据处理能力的同时，也放大了网络安全风险的复杂性与危害程度。本文从大数据环境下的网络安全风险特征出发，分析了海量多源数据采集、分布式存储与跨域流转、异构系统互联以及大规模实时处理等环节所暴露的新型风险。在此基础上，提出了基于大数据技术的风险识别关键技术，包括多源异构数据融合、机器学习异常检测、日志关联溯源及图神经网络隐蔽攻击发现。并从数据安全保护、态势感知响应、运维管理及联防联控四个维度设计了大数据驱动的网络安全防控体系。研究为大数据环境下的网络安全防护提供了系统的理论支撑与技术路径。
关 键 词： 大数据技术；网络安全；风险识别；防控体系；威胁检测

Research on the Construction of Cybersecurity Risk Identification and Prevention System Based on Big Data Technology

Xing Tianyi
Guangdong University of Foreign Studies, Guangzhou, Guangdong 510006

Abstract: While enhancing data processing capabilities, big data technology also amplifies the complexity and severity of cybersecurity risks. This paper starts from the characteristics of cybersecurity risks in the big data environment, analyzes the new types of risks exposed in the aspects of massive multi-source data collection, distributed storage and cross-domain transfer, heterogeneous system interconnection, and large-scale real-time processing. On this basis, it proposes key technologies for risk identification based on big data technology, including multi-source heterogeneous data fusion, machine learning anomaly detection, log correlation traceability, and graph neural network covert attack detection. It also designs a big data-driven cybersecurity prevention and control system from four dimensions: data security protection, situation awareness response, operation and maintenance management, and joint prevention and control. The research provides systematic theoretical support and technical paths for cybersecurity protection in the big data environment.
Keywords: big data technology; cybersecurity; risk identification; prevention and control system; threat detection

引言

网络攻击手段日趋隐蔽化、智能化和组织化，传统基于规则与签名的安全防护手段难以应对海量数据环境下的动态威胁[1]。大数据技术凭借其海量存储、快速计算与关联分析能力，为网络安全风险识别与防控提供了新的解决思路。然而，大数据平台本身的架构复杂性与数据流动的广泛性也引入了新的安全风险点，如数据采集环节的污染攻击、分布式存储的权限管理漏洞、多系统协同中的信任传递问题等。因此，有必要系统梳理大数据环境下网络安全风险的特征，研究适配的风险识别关键技术，并在此基础上构建一体化的防控体系。本文围绕上述问题展开研究，旨在为大数据技术的安全应用提供参考。

一、大数据环境下网络安全风险特征分析

（一）海量多源数据采集与汇聚带来的数据安全风险

数据采集层作为大数据系统的入口，面临来源真实性与完整性的双重挑战。攻击者可利用虚假数据源注入恶意样本，污染训练数据集，导致后续分析模型输出错误判断。不同类型传感器、日志系统、网络流量探针等采集设备的数据格式与协议差异较大，统一的汇聚节点成为攻击集散地[2]。采集频率过高时，数据缓存队列溢出可能引发数据丢失或乱序，攻击者利用这一窗口实施中间人篡改。数据汇聚阶段需要去除重复、纠正错误，这一清洗过程若存在漏洞，恶意构造的数据块可能绕过校验逻辑，进入持久化存储层。

作者简介：幸天奕，广东省兴宁市人，广东外语外贸大学，研究方向：计算机。

（二）分布式存储与跨域流转引发的安全边界模糊风险

大数据平台采用分布式文件系统与列式数据库，数据被切割为多个块副本存储在不同物理节点。传统基于网络边界的防火墙和入侵检测系统难以覆盖跨节点的数据流动路径，攻击者攻破单个数据节点即可获取部分敏感信息。跨域流转场景下，数据经过加密通道传输至不同信任域，接收域的访问控制策略若低于发送域标准，形成安全降级。数据生命周期中，冷数据与热数据的存储位置动态调整，迁移过程缺少端到端加密保护，暴露了数据在静止与传输状态之间的转换间隙。数据副本的异步一致性机制可能造成版本分歧，攻击者利用版本回滚攻击恢复已被删除的敏感记录。

（三）异构系统互联协同产生的接口与信任风险

大数据平台常与外部业务系统、第三方数据服务、云原生组件等异构系统对接。RESTful API、消息队列、数据库链接等接口暴露了较多攻击面，接口参数校验不严可导致 SQL 注入、命令执行或越权访问 [3]。各系统间的身份认证机制可能采用不同的令牌格式与有效期策略，令牌转发与重放攻击风险上升。跨系统信任关系中，一方系统被入侵后，攻击者可利用合法凭证横向移动至大数据平台。微服务架构下，服务网格的边车代理配置错误可能使内部通信流量绕过加密与鉴权。日志与审计信息分散在各异构系统中，难以形成完整的攻击链条还原。

（四）大规模实时处理引入的性能安全与资源竞争风险

流处理框架如 Spark Streaming、Flink 需要在高吞吐与低延迟之间平衡，资源调度器分配计算资源时，攻击者可提交大量短作业耗尽槽位，造成合法作业饥饿。内存计算模式下，算子执行过程中的中间结果保留在堆外内存，若未及时释放，内存碎片累积触发频繁垃圾回收，系统响应超时。同一物理机上运行多个租户的计算任务时，缓存侧信道攻击可窃取其他租户的密钥材料。反压机制处理上游数据积压时，背压传播可能导致整个拓扑停滞，攻击者通过突发流量触发这一状态，实现拒绝服务 [4]。实时模型的增量更新过程中，新旧模型版本切换瞬间存在短暂的不一致窗口，攻击者利用该窗口提交绕过旧模型但被新模型拦截的恶意流量。

二、基于大数据的网络安全风险识别关键技术

（一）多源异构安全数据的融合处理与特征提取技术

网络环境中产生防火墙日志、终端检测告警、DNS 解析记录、流量元数据等不同类型的安全数据，时间粒度与字段语义差异较大。融合处理的第一步是构建统一的数据模型，将各源数据映射至标准化的事件结构，保留原始证据的同时添加标注字段。时间对齐采用滑动窗口机制，补偿各数据源的时间戳漂移。特征提取阶段，对数值型字段计算统计特征，包括均值、方差、峰度及突变点数量。对类别型字段进行独热编码或嵌入向量表示，保留高基数特征的语义关系。协议头部字段中的标志位组合可转化为二进制特征向量，用于识别畸形包探测行为。流记录中的五元组信息计算熵值变化，反映扫描或 DDoS 攻击前的地址空间随机化趋势。

（二）基于机器学习的网络异常行为与攻击检测技术

有监督学习方法依赖标记好的正常与攻击样本，训练分类器如随机森林、梯度提升树，用于检测已知攻击类型。特征选择阶段使用互信息或递归特征消除，剔除冗余特征，降低过拟合风险。无监督学习适用于标记缺失场景，孤立森林算法在特征空间中随机切割，异常点因较少划分而路径更短。自编码器神经网络通过重构误差判断异常，正常流量重构误差低，攻击流量因偏离训练分布而产生高误差。半监督学习结合少量标记数据与大量未标记数据，使用生成对抗网络生成逼近真实分布的负样本，提升检测器泛化能力。时序卷积网络处理流量序列中的长程依赖，检测慢速扫描或低速率拒绝服务攻击。

（三）面向海量日志的智能威胁关联分析与溯源技术

安全运营中心每天接收百万至亿级日志条目，直接分析效率低下。日志解析阶段使用字典树或正则引擎将原始文本转换为结构化键值对，提取时间戳、源 IP、目的 IP、操作类型等关键字段。事件聚合按照时间窗口和共同属性对相似日志归并，减少数据量 [5]。关联分析采用因果图模型，将每个系统调用或网络连接视为节点，依据时序与依赖关系构建有向边。攻击路径追溯时，从可疑事件节点反向遍历，标记所有前驱节点形成攻击链。基于规则引擎的关联配置灵活，例如检测到某主机外发大量数据后，立即关联该主机之前十分钟内的认证失败记录。图数据库存储关联后的威胁情报，支持多跳查询，快速定位受影响资产范围。

（四）基于图神经网络的隐蔽攻击链路发现与风险预测技术

网络实体间存在通信关系、信任关系、数据流关系，构成异构图结构。图注意力网络为不同类型的边分配可学习的权重，模型聚焦于最具威胁传播可能性的连接。节点特征包括设备类型、开放端口、历史告警次数，边特征包含通信频率、传输数据量。消息传递过程中，每一层聚合邻居节点的信息，更新当前节点的表征向量。多层堆叠后，节点表征蕴含了多跳邻域的结构信息。针对高级持续性威胁，攻击者通过低慢小行为逐步渗透，传统检测无法串联离散事件。图神经网络将长时间跨度的登录操作、配置文件修改、数据访问事件映射到同一张图上，通过子图匹配识别与已知攻击模式相似的路径。模型输出每一节点被入侵的概率以及预测的下一步攻击目标，辅助安全团队提前阻断。

三、大数据驱动的网络安全防控体系构建

（一）以大数据平台为核心的数据安全与隐私保护机制

数据分类分级作为保护基础，依据业务属性与合规要求划分公开、内部、敏感、机密四级。不同级别数据在采集时施加不同强度的完整性校验，机密数据增加数字签名或区块链存证。存储加密采用列级加密与文件系统加密两层结构，密钥由统一密钥管理服务保管，密钥使用与数据访问日志关联审计。数据脱敏在输出环节动态执行，根据请求者角色决定掩码规则，身份证号显示后四位、手机号隐藏中间四位。差分隐私技术在统计查询输出中添加拉普拉斯噪声，保证单条记录的增减不影响查询结果分布，

抵御差分攻击。数据水印以不可见方式嵌入数值型字段的最低有效位，用于泄露溯源时提取水印定位责任主体。

（二）基于态势感知与智能预警的动态协同响应体系

态势感知平台聚合终端、网络、应用、身份四类传感器的遥测数据，计算全局安全评分。评分输入包括攻击尝试频率、异常流量占比、认证失败率、漏洞暴露面变化。时间序列预测模型（长短时记忆网络）依据历史评分曲线预测未来十五分钟的风险趋势，预警等级分为蓝、黄、橙、红四档。动态协同响应触发后，编排引擎调用预定义的处置剧本。低风险预警触发增强日志采样率与缩短检测间隔，中风险预警触发可疑 IP 临时封禁与多因素认证强制启用，高风险预警触发隔离受影响工作负载与启动备用取证链路。响应动作的状态反馈至态势感知平台，形成闭环评估处置有效性。跨组织协同场景下，加密共享告警摘要与处置建议，联盟内成员同步更新黑名单。

（三）大数据环境下的安全运维管理与溯源审计机制

运维管理建立变更控制流程，任何对大数据平台组件配置、访问控制列表、数据管道的修改需经过双人审批，修改前后生成配置基线快照。自动化运维工具执行批量补丁安装前，先在隔离环境中验证补丁兼容性与回滚方案。会话管理要求运维人员使用堡垒机登录，所有键盘输入与屏幕输出录屏留存，防止绕过审计。溯源审计采用不可篡改的审计日志链，每条日志包含前一条日志的哈希值，形成区块链式结构。审计字段覆盖操作主体、操作客体、时间戳、源 IP、操作类型、返回结果。针对数据泄露事件，审计系统支持模糊查询，根据泄露数据片段推断可能的查询语句与执行时间范围。长期存储的审计日志每季度抽样验证完整性，防止日志被篡改而不被发现。

（四）数据驱动与规则驱动相结合的联防联控防御架构

规则驱动部分依托专家经验提炼的检测规则，覆盖已知攻击模式，具有低误报、高确定性的特点。规则以 Sigma 或 YARA 格式编写，部署在流量入口与主机终端。数据驱动部分使用前述机器学习与图神经网络模型，覆盖未知或变种攻击。两类检测引擎并行运行，检测结果送入仲裁模块。仲裁模块采用置信度加权投票，规则引擎命中高危规则时权重较高，数据驱动引擎给出高异常分数但无规则匹配时需二次确认。联防联控机制中，数据驱动引擎发现的异常行为自动提取特征，经安全分析师确认后转化为新规则，补充规则库。反之，规则引擎频繁误报的阈值在数据驱动引擎的统计分布指导下调整。整体架构支持横向扩展，新增数据源或业务系统后，规则库与模型同步增量更新。

四、结论

本文围绕大数据技术在网络安全风险识别与防控体系构建中的核心问题展开研究。首先分析了大数据环境下四类典型风险特征，明确了传统防护手段的失效边界。其次提出了四项基于大数据的风险识别关键技术，涵盖了数据融合、机器学习检测、日志关联与图神经网络预测，提升了对隐蔽攻击的发现能力。然后建立了针对大数据平台自身安全的风险评估与研判方法，从组件脆弱性、数据生命周期评估到知识图谱协同研判，形成了完整的安全评估闭环。最后构建了大数据驱动的防控体系，整合了数据安全保护、态势感知响应、运维审计与联防联控机制。研究成果为大数据平台的安全规划、日常运营与应急响应提供了系统性的技术参考，推动网络安全防护从被动应对向主动感知与智能决策转变。

参考文献

[1] 张文聪 . 基于大数据的异构融合通信网络可信接入安全风险自动识别方法 [J]. 自动化与仪器仪表 ,2024,(07):164-167.

[2] 贾晓旭 . 基于可解释人工智能的数据安全风险识别研究 [J]. 信息系统工程 ,2024,(01):50-54.

[3] 于川 . 大数据技术在网络安全风险挖掘中的应用 [J]. 大众标准化 ,2023,(15):164-166.

[4] 陈佳雯 , 严利民 . 基于风险反馈的网络安全密钥可信度识别仿真 [J]. 计算机仿真 ,2023,40(08):393-397.

[5] 余战秋 . 基于攻防微分博弈的网络安全风险溯源识别 [J]. 河北北方学院学报 (自然科学版),2022,38(01):8-12.

