Security and Safety, Vol. 3, 2024009 (2024)
https://doi.org/10.1051/sands/2024009
sands.edpsciences.org

Security and Safety in Network Simulation and Evaluation

Research Article

Open Access

Information Network

RiskTree: Decision trees for asset and process risk assessment quantification in big data platforms

Haomou Zhan[1,2], Jiawei Yang[1,2], Zhenyang Guo[1,2], Jin Cao[1,2], Dong Zhang[3,*], Xingwen Zhao[1,2], Wei You[1,2], and Hui Li[1,2]

[1] *School of Cyber Engineering, Xidian University, Xi'an 710126, China*
[2] *The State Key Laboratory of Integrated Service Network, Xi'an 710126, China*
[3] *National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100024, China*

Received: 16 May 2024 / Revised: 6 June 2024 / Accepted: 2 July 2024 / Published online: 30 July 2024

Abstract Currently, big data platforms are widely applied across various industries. These platforms are characterized by large scale, diverse forms, high update frequency, and rapid data flow, making it challenging to directly apply existing risk quantification methods to them. Additionally, the composition of big data platforms varies among enterprises due to factors such as industry, economic capability, and technical proficiency. To address this, we first developed a risk quantification assessment process tailored to different types of big data platforms, taking into account relevant laws, regulations, and standards. Subsequently, we developed RiskTree, a risk quantification system for big data platforms, which supports automated detection of configuration files, traffic, and vulnerabilities. For situations where automated detection is not feasible or permitted, we provide a customized questionnaire system to collect assets and data processing procedures. We utilize a knowledge graph (KG) to integrate and analyze the collected data. Finally, we apply a random forest algorithm to compute risk index weights, risk values, and risk levels, enabling the quantification of risks on big data platforms. To validate the proposed process, we conducted experiments on an educational big data platform. The results demonstrate that the risk index system presented in this paper objectively and comprehensively reflects the risks faced by big data platforms. Furthermore, the proposed risk assessment process not only effectively identifies and quantifies risks but also provides highly interpretable evaluation results.

Keywords Big data platform, Quantitative risk assessment, Machine learning, Big data platform, Quantitative risk assessment, Machine learning

Citation Zhan H, Yang J, Guo Z, et al. RiskTree: Decision trees for asset and process risk assessment quantification in big data platforms. Security and Safety 2024; **3**: 2024009. https://doi.org/10.1051/sands/2024009

1 Introduction

The rapid advancement of information technology in recent years has facilitated the widespread integration of big data across various sectors of society. Big data technologies have shown immense potential in diverse applications, such as commodity recommendation systems and decision analysis. However, the adoption of these technologies also brings along security concerns that have gained prominence [1]. Many big data applications have adopted open-source platforms and technologies, which were initially designed

* Corresponding author (email: zhangdong@cert.org.cn)

for use within secure and trusted internal networks. The focus during subsequent developments of big data software has predominantly been on performance, leading to inadequate considerations for overall security planning. Additionally, the proficiency in technology and management among enterprises offering big data-related services varies significantly. These factors, compounded, have resulted in a surge of security incidents related to big data in recent years.

With the open nature of Internet technology, many companies rely on different open-source projects to construct diverse types of big data platforms serving various functions. The provision of big data services built upon complex, open-distributed computing and storage architectures presents substantial challenges to conventional authentication, access control, and security auditing mechanisms [2]. Traditional data protection methodologies often prove inadequate in addressing the escalating security demands associated with ever-increasing volumes of data [3]. To address this diverse landscape, Wu *et al.* [4] utilize the Delphi method to develop a comprehensive security evaluation index system specifically designed for safeguarding enterprise big data. In a similar line of research, Zhu *et al.* [5] introduce information entropy to determine the weight of each risk index and incorporate the fuzzy comprehensive evaluation method to quantify the privacy risks associated with social networks, enabling the assessment and prediction of privacy risks. Furthermore, various nations have put forth legislative measures and regulatory frameworks to safeguard the integrity and confidentiality of user data. For instance, Europe has implemented the General Data Protection Regulation (GDPR), the United States has enacted the Sarbanes-Oxley Act, and China has implemented the Cybersecurity Law of the PRC and the Personal Information Protection Law of the PRC. This highlights the urgent need for enhancing the existing big data security standards. Therefore, designing a solution for quantitatively assessing the risks present in both assets and data processing procedures on big data platforms has become a critical issue that needs to be addressed.

This study addresses big data security challenges by exploring risk assessment theories and techniques, offering a comprehensive risk quantification approach. The main contributions are as follows.

1) Through an analysis of the general architecture and business processes of big data platforms, we designed risk index systems specifically tailored for assets and data processing procedures. These metrics provide clear and actionable guidelines for identifying and quantifying risks within big data platforms.
2) Based on the aforementioned risk index system and considering the unique characteristics of big data, we were the first to establish a comprehensive risk assessment process for big data platform assets and data processing procedures.
3) We developed a prototype system, named RiskTree, based on the foundation of RiskLens [1]. The system workflow strictly adheres to the aforementioned risk assessment process and implements the risk measurement system we designed for both assets and data processing procedures. Compared to the original RiskLens solution, RiskTree introduces optimizations that lead to significant improvements in the credibility of risk data evaluation, interpretability of the risk quantification process, and visualization of the risk quantification results.
4) To validate the proposed risk assessment method, we applied it to an educational big data platform. Experimental results indicate that, due to the use of objective and comprehensive risk data, reliable risk analysis capabilities, and efficient machine learning algorithms, the proposed method significantly reduces labor and time costs. Furthermore, the method is highly adaptable to the ever-changing big data environment, enabling continuous iteration, updates, and optimization.

The structure of this paper is organized as follows: Section II introduces the research background and prior work related to risk quantification assessment. Section III reviews the risk assessment process, experimental procedures, and results presented in RiskLens [1]. Section IV provides a detailed description of our proposed method, RiskTree, a novel risk index and assessment process tailored for big data platforms. Section V presents the experimental evaluation and results conducted on an educational big data platform at a certain university. Finally, Section VI discusses the advantages of RiskTree and concludes the paper.

Note that this paper is an extended version of the previous conference paper RiskLens [1]. We conducted further research to incorporate the latest advancements in the fields of risk assessment and big data security. Additionally, we introduced the fundamental concepts of big data, the general architecture of big data platforms, and definitions of risk assessment. Based on these latest research findings and an investigation of the educational big data platforms, we propose a new risk quantification metric system.

Besides, we redesigned the risk quantification assessment process for big data platforms, which builds upon the principles established in RiskLens. Furthermore, Based on the proposed risk quantification assessment process and the new risk index system, we developed a risk quantification system named RiskTree. The key steps of RiskTree include assessing the vulnerabilities and security measures of the big data platform through vulnerability scanning tools and questionnaires, integrating and analyzing these assessments using KGs to provide data for risk quantification, and quantifying risks using a random forest algorithm. To obtain risk data, we replaced the simulation program used in RiskLens with a locally deployed educational big data platform and designed an evaluation experiment to validate the feasibility of RiskTree's risk assessment process.

2 Background and related work

2.1 Background

The term big data first appeared in the 1980s, coined by renowned futurist Alvin Toffler in his book "The Third Wave", where he hailed big data as the most splendid symphony of the third wave. This concept was initially aimed at describing the rapid growth of massive amounts of data in various industries and the challenges and opportunities these data presented. However, as technology has evolved and been applied, the notion of big data discussed in contemporary circles has transcended Toffler's original scope. The McKinsey Global Institute, in its report "Big Data: The Next Frontier for Innovation, Competition, and Productivity", states that "big data" refers to datasets whose size exceeds the capacity of typical database software tools to capture, store, manage, and analyze [6]. In today's information era, big data refers not only to the sheer volume of data but also includes the velocity at which data is generated, the diversity of data types, and the challenge of extracting valuable information from it. These characteristics are known as the 4Vs of big data [7–10].

- ***Volume*** refers to the massive amounts of data that are measured in petabytes or even exabytes in a big data environment. In contrast, traditional data often occupies a much smaller space, typically measured in megabytes or gigabytes.
- ***Velocity*** signifies the rapid speed of data generation and processing. Many big data scenarios require real-time or near-real-time data processing and analysis, while traditional data is produced at a comparatively slower rate and is often processed periodically or in batches.
- ***Variety*** alludes to the wide range of data types, including structured data (e.g., tables in relational databases), unstructured data (e.g., text, images, audio, and video), and semi-structured data (e.g., XML and JSON). On the other hand, traditional data is mostly structured data found in relational databases.
- ***Value*** indicates that big data has a low-value density, meaning that significant data mining and analytical efforts are required to extract valuable information. In contrast, traditional data typically has a higher value density, and its contents, often pre-processed and cleaned, are more readily utilized for analysis and application.

Big data platform assets can be categorized into *data assets*, *platform assets*, and *API assets*. Their specific descriptions are as follows:

- ***Data assets*** are electronically or otherwise recorded data legally owned or controlled by organizations, including governments, enterprises, and institutions. These assets, which can be structured or unstructured, encompass various forms such as text, images, voice, video, and web pages. Data assets are valuable resources that can be measured, traded, and generate economic or social benefits, but only when effectively managed and controlled [11]. By comparison, **big data assets**, in contrast, refer to large, complex, and diverse data sets within a big data environment, comprising structured, unstructured, and semi-structured data (e.g., XML, JSON). These assets are also legally owned or controlled by organizations and require advanced management strategies to harness their value.
- ***Platform assets*** in a big data platform refer to the collection of physical and virtual devices, networks, storage systems, and computing resources that support the platform's normal operation. These assets include servers, storage devices, network equipment, operating systems, software, *etc.*, and provide the core functions of data storage, processing, and transmission. The robustness, scalability,

and availability of infrastructure assets directly impact the performance, data processing efficiency, and security of the big data platform.

- ***API assets*** in a big data platform refer to the technical components used for integrating and transferring data. These assets, including ports and APIs, facilitate communication between the platform's internal and external systems. The primary function of interface assets is to ensure that different systems, applications, and users can securely access, transmit, and exchange data on the platform. They must support security mechanisms such as data access rights management, authentication, data encryption, and transmission control to prevent unauthorized access, data leakage, and man-in-the-middle attacks. Additionally, the performance and security of interface assets are crucial for the overall stability of the platform and the security of data flow, requiring robust input/output filtering mechanisms and error-handling capabilities.

The **data processing procedures** in a big data platform encompass nine procedures: *access authentication, data collection, data transmission, data provision, data exchange, data publication, data storage, data backup and recovery, and data destruction* [12].

- ***Access authentication***: During this procedure, the platform implements multi-factor authentication and access control policies to ensure that only authorized users can access data resources.
- ***Data collection***: The platform gathers user and internet data through external devices and web crawlers, while business data is collected via internal databases, file systems, and logs. The raw data is then preprocessed through cleaning, integration, and transformation operations.
- ***Data transmission***: Collected data is transmitted to storage systems through secure protocols such as SSL, HTTPS, and FTP, ensuring safe aggregation across the internet and internal platform links.
- ***Data provision***: Preprocessed data is securely provided to authorized users or systems via APIs with strict access controls.
- ***Data exchange***: The platform facilitates data transfers between different systems, employing encryption and audit mechanisms to ensure data security and compliance.
- ***Data publication***: Data is publicly released or distributed, with de-identification and privacy protection measures in place to prevent sensitive information from being exposed.
- ***Data storage***: The platform employs distributed storage technology to shard and block large volumes of structured and unstructured data across multiple devices [13]. Protocols like ZAB, Paxos, or Raft are used to ensure data availability and consistency. Load balancing and high availability are achieved through the configuration of primary and backup master nodes [14]. Data processing involves distributed computation across the Map, Shuffle, and Reduce phases [15].
- ***Data backup and recovery***: Critical data is regularly backed up, and in case of data loss or corruption, the platform uses backup nodes to restore data, ensuring data integrity and business continuity.
- ***Data destruction***: The platform irreversibly deletes data and backups from storage media in compliance with relevant national regulations, preventing potential data breaches.

2.2 Related work

In the realm of information system risk assessment, a risk evaluation procedure consists of the identification of system vulnerabilities, potential threats, and the consequent losses stemming from these vulnerabilities and threats. Present research thrusts in the field can be bifurcated into three distinct streams: the **risk assessment index system**, the **risk assessment model**, and the **vulnerability scoring system**.

Within the ambit of the **risk assessment index system**, extant scholarly endeavors concentrate on delineating risk indices and forging a robust risk index system by dissecting the threats looming over the targeted information system. For instance, Peng *et al.* [16] proposed a data dissemination process model based on risk factors. They selected risk indices for the data dissemination process and used the Delphi method to screen and define these indicators for big data transmission operations. They then utilized the Analytic Hierarchy Process (AHP) to derive the importance of each indicator. Similarly, in reference [17], a qualitative analysis of privacy risk factors in social networks under the context of big data was conducted. The Delphi method was employed to construct an evaluation system, and the weight of the indicators was calculated using information entropy measurements. The fuzzy comprehensive evaluation

method was applied to quantitatively assess and predict the privacy risks of social networks. Likewise, Zhao *et al.* [18] adopted a fuzzy assessment approach for gauging the likelihood and ramifications of risk events, thereafter utilizing an entropy weight coefficient method to appraise the contribution of every risk determinant towards the holistic risk assessment. Besides, Lu *et al.* [19] embarked on a quantitative risk exploration for industrial control mechanisms, refining the AHP by infusing the fuzzy AHP to mitigate the challenges pertaining to judgment matrix consistency. The deployment of quantitative risk index systems furnishes an intuitive peephole into the potential hazards associated with the system in question, thereby facilitating an all-encompassing scrutiny of risk elements. Nevertheless, the majority of the prevailing methodologies are underpinned by subjective data sources, such as surveys or expert analyses. This reliance potentially seeds a degree of subjectivity into the quantification endeavor, which may, in turn, impinge upon the precision and trustworthiness of the outcomes.

As for the **risk assessment model**, current research undertakes the extraction of risk factors from information systems, constructs models, and transforms risks into model variables for analysis in order to achieve risk quantification. In reference [20], the fault tree analysis method was employed to calculate the risk values associated with information system program interchange, remote attacks, and risks in the network. Subsequently, based on the risk assessment factors and model design, a risk assessment system was developed and implemented in a network environment. For instance, Zhang *et al.* [21] presented a quantitative risk appraisal approach centered on host system security via vulnerability scanning. Their methodology involves constructing a vulnerability association graph of the host system to facilitate quantitative risk evaluation. Similarly, Xie *et al.* [22] proposed an attack tree model that scrutinizes the threat vectors of each leaf in the attack tree, enabling the computation of the threat vector of the complete attack path to derive the risk value. In another line of research, Zhang *et al.* [23] proposed a fuzzy radial basis function neural network model to numerically process network security risk factors and derive risk levels. Nan *et al.* [24] proposed a security risk analysis model that constructs a Bayesian network to calculate the likelihood and severity of security incidents. They then utilize an ant colony optimization algorithm to compute the risk propagation path. Developing further, Li *et al.* [25] proposed an Analytic Hierarchy Process-Genetic Algorithm Back Propagation model that sets the structure of the BP neural network according to the provided risk index system structure and adjusts the parameters of the neural network with a genetic algorithm. Dacier *et al.* [26] established an IT system vulnerability privilege graph model, converting it into a Markov chain to assess risks quantitatively across diverse attack scenarios. Patel *et al.* [27] proposed an enhanced fault tree method to assess the impacts of vulnerability and threats on information systems quantitatively.

While modeling the target system can yield authentic and objective risk data, the establishment and analysis of quantitative models may become proportionally complex as the system's size and complexity increase. The prevailing **vulnerability scoring systems** in operation lay down preset risk indices and evaluation scales for vulnerabilities, deploying these vulnerability scoring methodologies to quantify the threat level posed by vulnerabilities. In the current scientific and operational landscape, widely endorsed vulnerability scoring models include the Common Vulnerability Scoring System (CVSS) [28], the Threat Assessment and Remediation Analysis (TARA) [29], and the Common Weakness Scoring System (CWSS) [30]. CVSS divides vulnerability risk indices into three categories: base, temporal, and environmental. These categories are used to quantitatively assess and summarize the intrinsic characteristics of vulnerabilities, the characteristics that change over time, and the characteristics displayed in the user environment, respectively. The TARA assessment model focuses on threat assessment for selected network assets and analyzes the mitigation measures for network risks. The CWSS assessment system categorizes weak point indicators into three groups: foundational discovery, attack surface, and environment. As the information about weak points becomes more refined, the weak point scoring becomes more accurate. The vulnerability scoring model proposes a systematic evaluation index and calculation formula for system vulnerabilities. It enables comprehensive and accurate analysis of vulnerabilities and plays a crucial role in measuring and numerically expressing the severity of specific system vulnerabilities. However, the threat sources for information systems are not limited to vulnerabilities alone. The quantification of risk for the entire information system should not be confined to vulnerability scoring alone.

In recapitulation, the mainstream thrust of risk assessment research is devoted to the handpicking of risk indices and the assembly of a risk index system, grounded in the examination of threats confronting the target information system. The techniques hitherto employed for risk quantification are substantively dependent on subjective data sources, a factor that potentially introduces biases and impinges upon the

accuracy and trustworthiness of the resultant assessments. Additionally, the process of system modeling, which can be integral to risk analysis, poses significant difficulties, especially as the scope and complexity of the system amplifies. The circumstances underscore the necessity for the development of an automated and standardized quantitative risk assessment process that caters to the demand of conducting quantitative risk assessment of big data platforms. To ensure the compliance and credibility of the proposed solution, our work is conducted in accordance with China's national standard GB/T 35274-2017, Information Security Technology - Security Capability Requirements for Big Data Services. However, this does not imply that RiskTree is limited to operating under this national standard. With minor adjustments to the indicator system proposed in this paper, it can be adapted to comply with other regulatory frameworks as well.

Based on our understanding of this standard, we divided the risk assessment of big data platforms into the quantification of asset risks and data processing procedures risks. In the quantification of asset risks, we primarily consider potential risks related to data and system assets, organizational and personnel management. For the quantification of data processing procedures risks, we referred to the standard's classification of the data lifecycle and conducted threat modeling for each data processing procedure to analyze potential risks within them.

3 An overview of RiskLens

RiskLens [1] analyzes the threats faced by current big data platforms in terms of assets and data processing procedures. It proposes a primitive risk assessment process, a risk quantification metric system, and uses a random forest algorithm for risk quantification. Through experiments conducted on a simulated big data platform, RiskLens calculates the weights of the metrics and validates the feasibility of the proposed risk quantification method. RiskLens aims to provide a scientific system for quantitatively assessing big data risks. Compared to traditional risk assessment methods such as the Delphi method, the AHP, and fuzzy comprehensive evaluation, the proposed approach offers advantages in objectivity, cost efficiency, and general applicability. The following is a brief description of the main process of RiskLens.

- ***Methodology***: The risk quantification assessment process of RiskLens consists of four steps: *preparation, risk identification, risk quantification analysis, and risk evaluation*. In these steps, the assessed enterprise sequentially undertakes the following tasks: defining the assessment objectives, identifying the big data platform assets at risk, modeling the data processing procedures, considering the likelihood, impact, and tolerance of potential risks, establishing a risk quantification metric system, and finally, quantifying the risks associated with the big data platform and assigning a risk rating. During the construction of the risk index system, user profile construction was selected as the focus for the risk quantification assessment. This involved identifying the assets and modeling the associated data processing procedures within the big data platform. Consequently, specific risk index systems were developed for both assets and data processing procedures. The asset risk index system encompasses metrics such as physical security, system security, big data platform component security, cloud infrastructure security, and employee security, while the data processing procedures risk index system includes metrics related to technical security and management security. Additionally, RiskLens introduces a risk calculation formula based on the weighted average of the scores for each risk index. For the experimental evaluation of RiskLens, a simulated big data cluster composed of three CentOS7 nodes was set up. Expert scoring was employed to evaluate each metric within the proposed risk index system, according to the performance and security situation of the simulated big data platform, leading to the final risk rating. The expert scoring data were analyzed using the random forest algorithm to determine the weight of each metric. The overall risk score and risk level of the big data platform were then calculated based on these weights and metric scores.
- ***Advantages***: RiskLens builds on existing risk assessment theories and integrates the characteristics of big data to propose a process for quantitative risk assessment in big data environments. The process is designed to enhance enterprises' ability to identify risks within big data platforms. It develops evaluation metric systems for big data platform assets and data processing procedures based on this process, providing a structured process for big data platform risk assessment. Experimental results suggest that the method leverages the computational power of modern computers and the learning

capabilities of machine learning algorithms, offering a potentially more objective and scalable approach with reduced labor and time costs, and the capacity to process large volumes of data continuously.

- ***Limitation***: Despite these advantages, the RiskLens approach has certain limitations. The asset risk indices were derived from analyzing the risks associated with assets involved in a certain big data platform business. However, real-world big data platforms consist of vast amounts of data, extensive software and hardware components, and numerous specific business scenarios. Therefore, the risks identified in the assets of a single operation may not adequately represent the overall asset risks of a comprehensive big data platform. Additionally, although RiskLens proposed a risk index system for data processing procedures, including procedures such as data collection, transmission, and storage, it did not thoroughly analyze the vulnerabilities and potential threats associated with each procedure. This lack of detailed analysis diminishes the logical rigor and comprehensiveness of the risk assessment. Furthermore, the data used in RiskLens to quantify risks in big data platforms were derived from simulation programs and expert evaluations. This approach may not accurately reflect the risks faced by big data platforms in real-world environments. Expert evaluations can be influenced by personal expertise, leading to subjective and potentially biased assessments, which could undermine the credibility of the final risk quantification results. Additionally, after acquiring the risk data, RiskLens directly used it to calculate the weights of risk indices and overall risk score without further processing or analyzing the data.

In response to these challenges, this paper introduces RiskTree, which builds on the strengths of RiskLens while incorporating enhancements that improve both the accuracy and comprehensiveness of risk assessments. Specifically, RiskTree categorizes big data platform assets into three groups: data, platform, and API, and conducts a thorough analysis of the risks associated with these assets under common threats and vulnerabilities. Additionally, RiskTree examines all procedures of the data processing lifecycle, identifying potential vulnerabilities and threats in each stage to provide a more accurate risk assessment.

To further enhance the objectivity and reliability of the risk quantification process, RiskTree employs KGs to correlate and store multi-source heterogeneous risk data related to assets and data processing procedures. By integrating current risk data, RiskTree offers a more comprehensive and multidimensional analysis of big data platform risks, providing a more complete and thorough dataset for risk assessment. This approach not only mitigates the subjectivity associated with expert evaluations but also better reflects the complexities and realities of real-world big data platforms.

4 The proposed scheme: RiskTree

4.1 Risk index system

To describe the risks associated with big data platform assets and data processing procedures more clearly, as well as enable more precise automated risk identification and quantification, we developed an Asset risk index system and a Data Processing Procedure risk index system. These systems were designed based on a thorough analysis of the vulnerabilities and potential threats inherent in the general architecture and business processes of big data platforms.

4.1.1 Asset risk index system

Based on the general architecture and businesses of big data platforms, assets are categorized into three types: *data assets, platform assets,* and *API assets*. We analyze the vulnerabilities of each category and the related risks.

- ***Data assets*** face significant risks due to system vulnerabilities, including unencrypted storage, inadequate data classification and grading, memory leaks, and weak encryption, which may be exploited by insider threats, malware infections, or supply chain attacks. Additionally, insufficient data cache destruction and inadequate privacy protection measures increase the risk of privacy breaches, with sensitive information potentially being tampered with or stolen. The lack of effective data backup and destruction processes may result in permanent data loss and severe privacy violations.

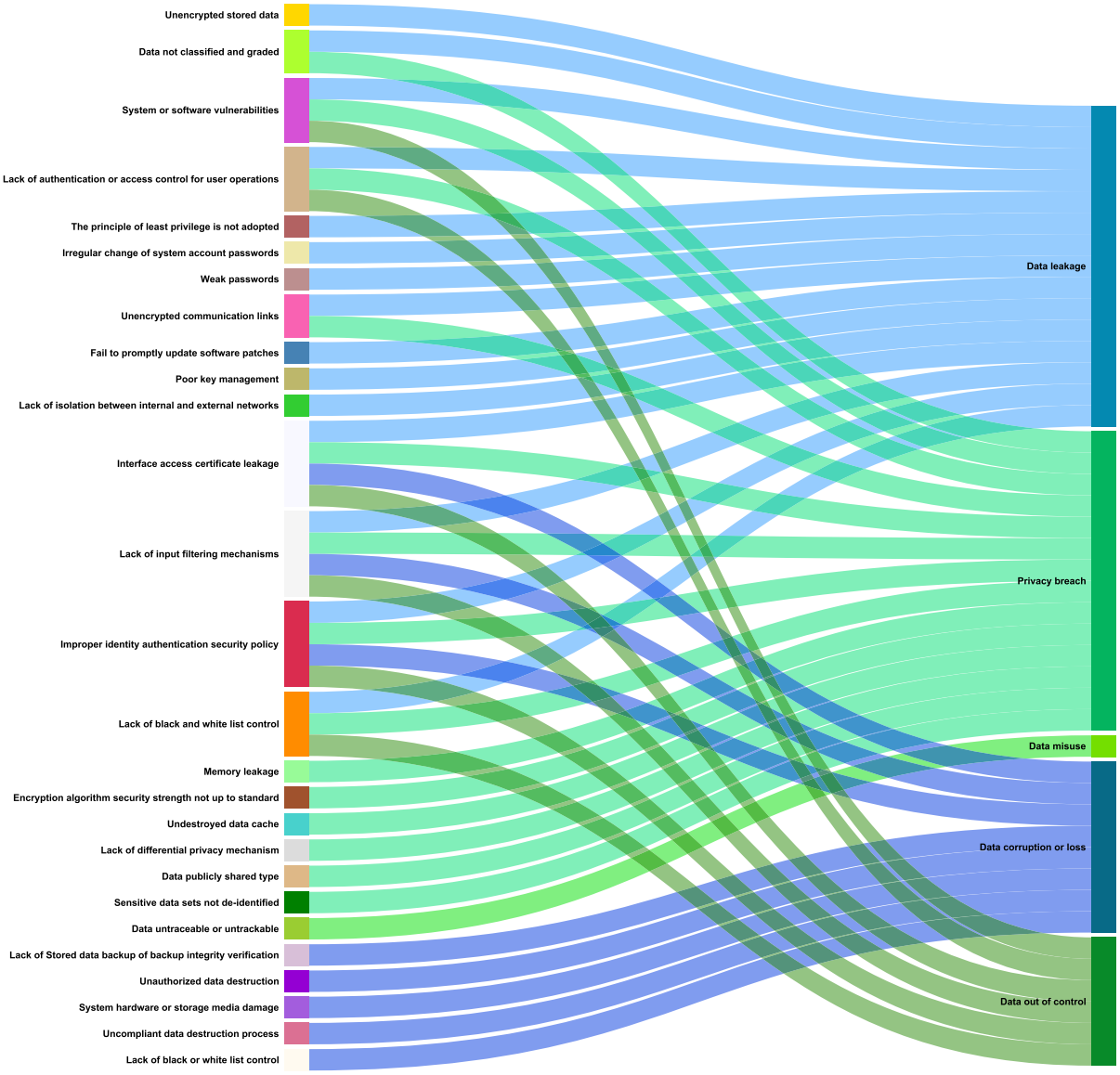


Figure 1. Risk quantitative evaluation index system of assets

- ***Platform assets*** are vulnerable to advanced persistent threats (APT), supply chain attacks, and insider negligence. Therefore, strict authentication and access control are necessary. These threats can lead to system failures or large-scale data breaches. Specifically, weak passwords, infrequent updates, and poor data management exacerbate these risks, complicating recovery efforts and potentially compromising the integrity and availability of the platform businesses.
- ***API assets*** face risks from APT and poor access management. Inadequate authentication strategies, insufficient input filtering, and the lack of proper monitoring and whitelist/blacklist controls expose APIs to risks such as unauthorized access, data breaches, privacy violations, and system overloads. These vulnerabilities increase the likelihood of data misuse and loss, making it difficult to trace and respond to malicious activities.

Based on the above analysis, Figure 1 presents the risk indices for the identified assets. It outlines the risk index system for big data platform assets, with the right side depicting asset-related risks and the left side highlighting the vulnerabilities linked to each risk.

4.1.2 Data processing procedure risk index system

We divide the data processing procedures of the big data platform into nine procedures to comprehensively assess the involved risks: *access authentication, data collection, data transmission, data provision, data exchange, data publication, data storage, data backup and recovery, and data destruction.*

- ***Access Authentication***: Threats such as exposed credentials, unencrypted communication links, and weak passwords make the system vulnerable to APT, supply chain attacks, and man-in-the-middle attacks. Furthermore, without strict authentication, access control, and strong password policies, the risks of unauthorized access, data breaches, and tampering are significantly increased. Failure to enforce the principle of least privilege or monitor unauthorized activities further heightens the risk of data misuse or loss.
- ***Data collection***: Proper authentication and access control mechanisms can effectively prevent malicious or erroneous data from entering the platform. However, system vulnerabilities or the lack of data traceability can result in issues related to the authenticity and integrity of the data.
- ***Data transmission***: Transmitting data in plaintext can lead to severe security issues, particularly sensitive data leakage or tampering. Inadequate authentication configurations, the use of low-security encryption algorithms, and poor key management practices exacerbate these risks, especially when combined with threats such as malware or insider attacks.
- ***Data provision***: Once sensitive data is provided without proper desensitization, the privacy of users or the platform faces significant threats. Inadequate encryption and desensitization strategies, along with improper authorization mechanisms, further aggravate this issue.
- ***Data exchange***: Unauthorized data exchange poses serious security concerns, jeopardizing the confidentiality and integrity of the data. Existing system vulnerabilities exacerbate the likelihood of such occurrences.
- ***Data publication***: Failing to de-identify sensitive datasets before publication exposes the system to risks of privacy breaches and data misuse. The lack of traceability in the published data further exacerbates these risks.
- ***Data storage***: Data encryption and desensitization are crucial in this procedure. Using encryption and desensitization algorithms that do not align with the data's privacy level makes the system vulnerable to insider threats, malware, and ransomware attacks, which could lead to privacy breaches and misuse. The absence of a backup mechanism complicates recovery, potentially resulting in data loss or irreversible damage. This procedure is also prone to issues such as data leakage from privilege abuse and compromised data integrity.
- ***Data backup and recovery***: The primary considerations in this procedure are system authentication and whether the system's vulnerabilities have been patched, as these factors directly impact the availability and integrity of data.
- ***Data destruction***: Unauthorized data destruction and non-compliant destruction processes make the platform susceptible to APT, insider negligence, and supply chain attacks. These threats could result in the recovery or tampering of deleted data, thereby compromising data integrity.

Through the analysis of the nine data processing procedures, we identified potential security risks and vulnerabilities within each, forming the risk quantification of the data processing procedures, as shown in Figure 2. The right side of the figure presents the risks, while the left side displays the corresponding vulnerabilities.

4.2 Risk assessment process and system workflow

As shown in Figure 3, to efficiently conduct risk quantification, we designed a comprehensive risk quantification process, which serves as the foundation for the development of RiskTree. To ensure the feasibility and compliance of our designed risk quantification process, we have developed a risk assessment procedure for big data platforms based on the Chinese national standard GB/T 20984-2022, while also considering the general architecture and business operations of big data platforms. Specifically, the GB/T 20984-2022 standard outlines a four-step risk assessment process, including preparation for assessment, risk identification, risk analysis, and risk evaluation. Among these steps, risk identification requires identifying the assets, threats, existing security measures, and vulnerabilities of the evaluation target. Additionally, the

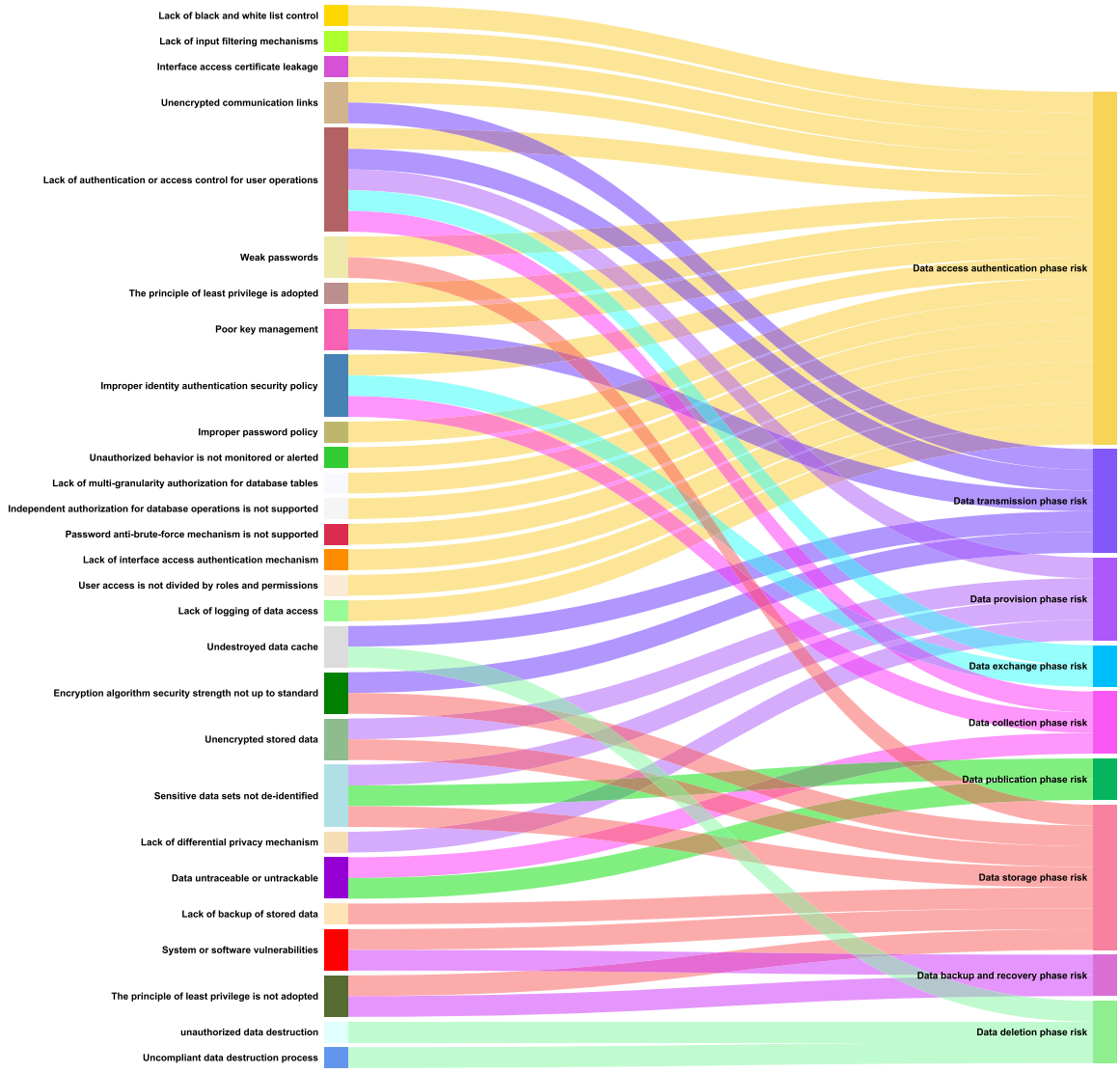


Figure 2. Risk quantitative evaluation index system of data processing procedures

standard mandates the documentation of the entire assessment process and maintaining communication and consultation throughout. In our design, we center the process around risk data from the big data platform and propose a four-step risk assessment process: assessment preparation, risk data collection, risk data analysis, and risk evaluation. Risk data collection involves sensing and preprocessing the vulnerabilities of assets and data processing procedures, while risk data analysis involves further associating and visualizing the properties and vulnerabilities of these assets and data processes. To increase the automation of the entire assessment process, we eliminated manual communication and consultation, replacing it with automatically generated and recorded intermediate assessment results at each stage. These include details of the risk quantification task from the assessment preparation stage, knowledge graphs from the risk data analysis stage, and risk quantification reports from the risk evaluation stage. The proposed risk assessment process can be summarized in the following key steps:

(1) ***Preparation***: To initiate the RiskTree assessment process, the assessor must conduct a detailed analysis of the target big data platform, clearly defining the assessment objectives and scope.
(2) ***Risk data collection***: In this step, the system utilizes vulnerability scanning tools and automated questionnaire construction. The vulnerability scanning tools are used to automatically detect vulnerabilities in the assets and data processing procedures of the big data platform, while the questionnaires

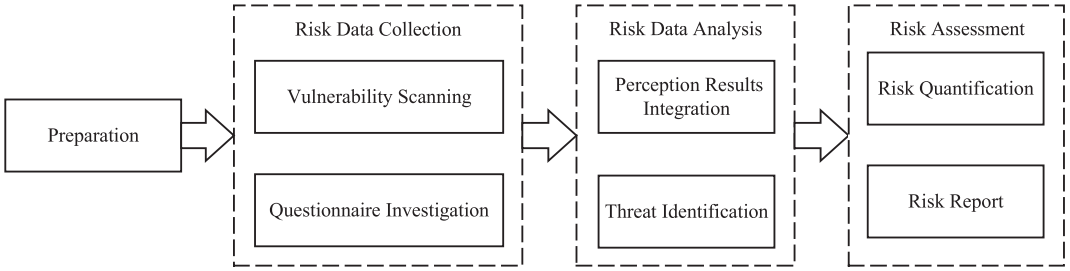


Figure 3. Risk assessment process of the big data platform

are employed to gather vulnerability data that cannot be accessed by scanning tools, along with the security policies adopted by the big data platform.

(3) ***Risk data analysis***: In this step, KGs are used to integrate and correlate various forms of vulnerability data. Then this process further analyzes potential threats to the big data platform.

(4) ***Risk assessment***: Design and implement a risk quantification model based on the risk index system. Query the vulnerability data stored in the KG to calculate the risk scores for each asset and data processing procedures within the big data platform. Then the assessment result, including the potential threats risk scores and risk levels for each asset and data processing procedures will be integrated into a risk report.

4.3 System design

When applying the aforementioned metric system to their platform, enterprises should first clearly define their assessment objectives and scope. Subsequently, the RiskTree process will proceed with the following steps.

4.3.1 Risk data collection

At this procedure, RiskTree employs vulnerability scanning tools and questionnaires to collect operational data and identify risks associated with data, platform, business assets, and various data processing procedures on the big data platform. The non-intrusive scans ensure that normal system operations and businesses are not disrupted, though they may have limitations in scanning depth and could result in false positives or missed vulnerabilities. The assessment primarily relies on existing data, logs, configuration files, and passive network traffic monitoring, minimizing the system's impact and aligning with continuous risk assessment requirements.

To address the limitations of non-intrusive scanning in detection capabilities, additional questionnaires are provided for further data collection. This aims to identify assets, data processing procedures, vulnerabilities, or security policies that scanning tools might miss. Based on the risk index system, 33 basic questions are designed to cover vulnerabilities related to data, platform, API, and data processing procedures, focusing on key areas such as access control, authentication, encryption, privacy protection, and log management. For example, questions ask whether the platform API enforces access authentication, if data is encrypted during storage, and whether access activities are logged. These queries help uncover critical security practices and potential vulnerabilities, offering a more complete view of the platform's risk landscape.

Additionally, the questionnaire is customized based on specific assets or data processing procedures being evaluated. For example, a general question like *"Is network isolation implemented?"* would be tailored to *"Is network isolation implemented for the Hive4 database at IP address 10.10.22.4?"*. By gathering detailed information, this method compensates for the limitations of non-intrusive scans, ensuring a more comprehensive and accurate risk assessment.

We will invite developers, testers, and operations personnel from the big data platform to participate in consultations and questionnaire responses during the risk data collection phase. Developers are responsible for the design and implementation of the platform components, as well as developing core

modules such as data collection, storage, processing, and application. They are familiar with the platform architecture, data flow paths, and processing procedures, and can provide detailed information on assets and data processing, as well as potential code defects and security vulnerabilities in data interfaces during development. Testers are primarily responsible for verifying the functionality and performance of the big data platform, ensuring the system's accuracy and stability when handling large-scale data. They can identify security issues discovered during testing and provide feedback on the effectiveness of various security measures. Operations personnel are responsible for the daily maintenance and management of the platform, ensuring its continuous operation. They are able to describe potential operational risks in processes such as data transmission and storage, and can point out risks related to operational errors or misconfigurations, such as data loss or inadequate permission management.

Through the questionnaire survey of the aforementioned personnel, RiskTree can comprehensively collect information on the attributes and vulnerabilities of assets and data processing procedures that cannot be detected by automated vulnerability scanning tools. Although the responses from developers, testers, and operations personnel may carry a degree of subjectivity, their answers remain the most reliable and valuable information available at this stage. These individuals possess rich professional expertise and practical experience, and are the ones most familiar with the system architecture, data processing procedures, and potential risks. While their responses may reflect personal perspectives, their judgments are based on deep system knowledge, providing critical insights for the risk assessment process. Moreover, we have meticulously designed the questionnaire to cover various aspects of risk in the big data platform, guiding respondents to focus on key risk-related issues, thereby reducing the potential bias introduced by subjectivity to a certain extent.

4.3.2 Risk data analysis

In RiskTree, we use KGs to integrate risk data from various sources and formats, transforming it into structured data. The nodes in the graph represent detailed information about assets, data processing procedures, and associated vulnerabilities, while the edges define relationships between these elements, such as the connections between assets, vulnerabilities, and data processing procedures.

We developed two types of risk KGs: the Asset Risk KG and the Data Processing Procedures Risk KG. These graphs visually depict the primary assets, data processing procedures, vulnerabilities, and their interconnections within the big data platform, facilitating the automated construction of risk KGs during data integration.

As shown in Figure 4, the Asset Risk KG is built using 11 ontologies, including data source, data table, data field, server, hardware, operating system, software, business system, port, API, and vulnerability. Similarly, as illustrated in Figure 5, the Data Processing Procedures Risk KG is composed of 10 ontologies, such as data collection, data storage, data backup and recovery, data destruction, data access authentication, data transmission, data exchange, data provisioning, data publication, and vulnerability.

When constructing the risk KGs from multi-source heterogeneous data based on the risk KG templates, we faced the following challenges:

1) ***Classification and integration of sensed data***: Sensed data comes from vulnerability scanning tools and questionnaires, including detailed information about ten types of big data platform assets, nine data processing procedures, and related vulnerabilities. The first challenge is to automatically classify and integrate this complex, multi-source, heterogeneous data and convert it into the nodes and edges of the KG.
2) ***Automated construction of KGs***: Once the sensed data is converted into nodes and edges for the KG, the second challenge is to automatically translate this data into Cypher query parameters to construct a complete risk KG in Neo4j.
3) ***Querying the KG***: After constructing the KG, we need to query it to provide the necessary risk data for assessment, such as associated assets and vulnerabilities for a specific asset, or the preceding and subsequent procedures and vulnerabilities for a specific data processing procedure. The third challenge is to automatically generate Cypher queries based on the risk assessment requirements and format the query results appropriately.

To address the challenges mentioned, we developed a KG module using Java, which automates the process of obtaining sensed data, converting it into nodes and edges within the KG, and executing

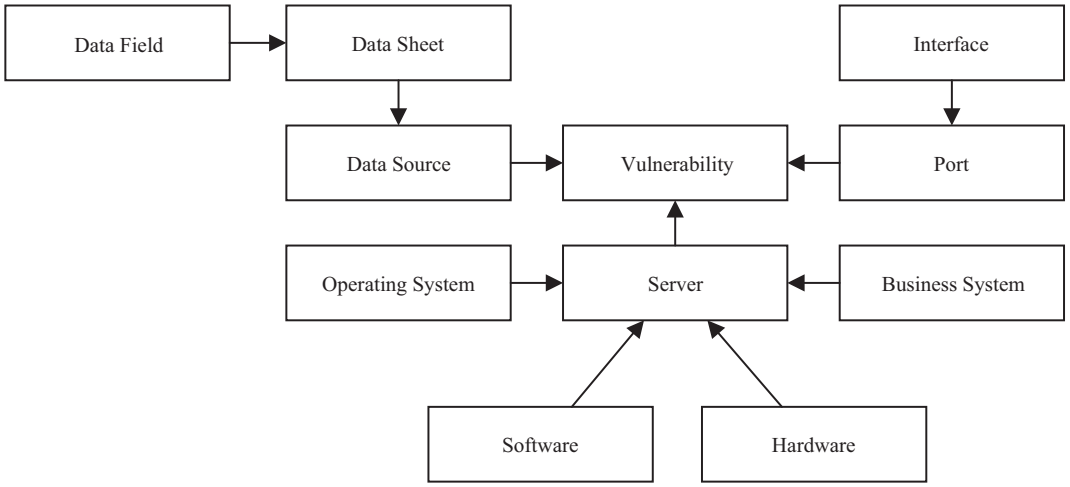


Figure 4. Assets risk KG template

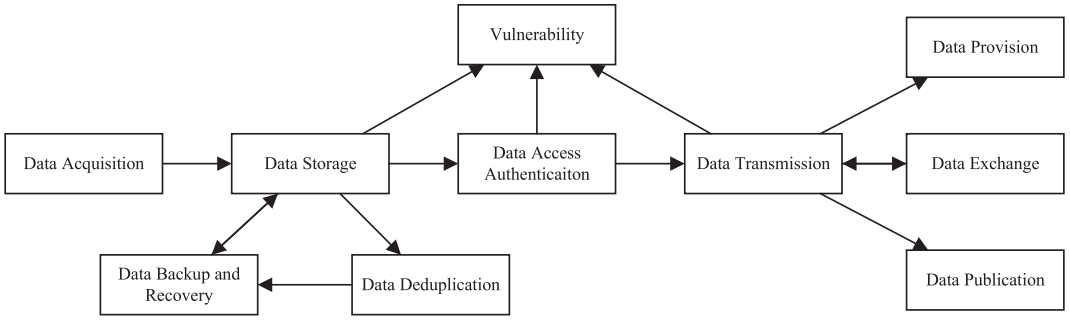


Figure 5. Data processing procedures risk KG template

automated queries on the graph. This module consists of four submodules: Sensed Data Reading, KG Construction, and KG Query.

Sensed Data Reading Submodule: This submodule defines data structures for each type of asset and data processing procedure. It dynamically executes SQL queries on the MySQL database to retrieve risk data. The data is then categorized and stored in the appropriate data structures according to predefined parsing rules. Finally, the query results are encapsulated in JSON format and sent via an API to the KG Construction submodule.

KG Construction Submodule: Based on the risk graph templates, this submodule has predefined rules for generating Cypher queries to create nodes and edges for each asset, data processing procedure, vulnerability, and their interactions. Upon receiving the preprocessed JSON data, this submodule extracts the key parameters required to construct nodes (e.g., asset type, data type, vulnerability ID) and generates the corresponding Cypher queries. These queries are sent to Neo4j through a session created using the 'Neo4jLink' class to build the KG nodes. The submodule then continues to parse the JSON data, identifying relationships between assets, data processing procedures, and vulnerabilities based on predefined fields (e.g., asset relationships, data flow sequences, vulnerability associations). It generates Cypher queries to establish these relationships within the KG and sends them to Neo4j.

KG Query Submodule: This submodule defines several Cypher query templates based on the risk data needed for assessment. After executing a query, the results are encapsulated in JSON format and returned.

4.3.3 Risk assessment

After evaluating various risk quantification methods, we chose to continue using the Random Forest algorithm adopted in RiskLens for calculating the risks associated with big data platform assets and data

processing procedures. Random Forest is an ensemble learning method, that improving classification or regression accuracy by building multiple decision trees and integrating their results. Each tree is trained using random sampling of the data and random selection of features at each node, with the final prediction derived from the aggregated results of all trees, typically through voting or averaging. Based on the risk index weights calculated using the random forest algorithm, we proposed new risk quantification formulas for the big data platform. Specifically, for asset risk calculation, we introduced Equation (1), which further accounts for the impact of asset type and the presence of sensitive and non-sensitive data on the overall asset risk. For calculating the influence of asset type on risk, we introduced Equation (2), which considers the vulnerabilities and threats associated with different types of assets. Similarly, for data processing procedure risk calculation, we proposed Equation (3), which incorporates the effects of vulnerabilities and threats on risk.

$$Risk = \sum_{i=1}^{n} Type_i * \left\lfloor \left(log_2 \left(pro_{sen} * 2^{val_{sen}} + pro_{n\_sen} * 2^{val_{n\_sen}} \right) \right) \right\rceil \tag{1}$$

$$Type_i = \sum_{j=1}^{m} \left(Weak_j * Weight_j * \sum_{k=1}^{p} Threat_k \right) \tag{2}$$

$$Val_i = \sum_{j=1}^{m} \left(Weak_j * Weight_j * \sum_{k=1}^{p} Threat_k \right) \tag{3}$$

In Equation (1), the variables are defined as follows: Type represents the risk associated with the asset type. pro_{sen} is the amount of sensitive data, val_{sen} is the sensitivity level of the data, pro_{n_sen} is the amount of non-sensitive data, and val_{n_sen} is the sensitivity level of non-sensitive data. Equation (2) provides the calculation for the risk associated with the asset type, where Weak is the asset's vulnerability, Weight represents the vulnerability weight calculated by the random forest algorithm and Threat is the threats faced by the asset. Each vulnerability corresponds to multiple threats. In Equation (3), Weak refers to the vulnerabilities in the data processing procedure, Weight is the vulnerability weight and Threat represents the threats faced during that procedure.

5 Evaluation

5.1 Setup

To demonstrate the feasibility and effectiveness of RiskTree, we set up an educational big data platform. This platform consists of web servers, database servers, network devices, and storage devices, with Apache Hadoop and Spark serving as the primary tools for big data processing and analysis. Data storage is managed through HDFS, access authentication is provided by LDAP, and Apache NiFi is used for data collection management. The platform also integrates core business systems such as a Learning Management System and a Student Information System, utilizing big data technologies to enable personalized learning recommendations, learning outcome assessments, and educational data visualization. Additionally, we reviewed the overall system configuration and implemented security policies to ensure the safety of data storage, the effectiveness of access authentication, and the real-time, efficient operation of data collection processes. We conducted scans of all assets and data processing procedures within the 10.10.22.0/29 subnet and the asset with the IP address 10.10.22.71. The tools used for scanning included nmap, hydra, and our own custom-written MySQL database scanning script. The types of assets scanned included data sources, data tables, data fields, server hosts, software, operating systems, business systems, hardware, and ports. The data processing procedures scanned encompassed data storage, access authentication, collection, transmission, backup and recovery, and deletion.

5.2 Risk data collection

In our experiment, RiskTree utilized Nmap[1], Hydra[2], and a custom MySQL scanning script to assess vulnerabilities across the big data platform's software, hardware, ports, and data processing procedures,

[1] https://nmap.org/
[2] https://github.com/vanhauser-thc/thc-hydra

Id	Asset or Processing Procedure's Name	Update Time	Detail Information
gzAE-1111-1111-t201	/api.example.com/v1/users	2024-08-08 21:48:43	port id: gzAD-1111-1111-t201
gzAD-1111-1111-t201	7777	2024-08-08 21:45:49	server id: gzAA-1111-1111-201 asset type: port
gzAA-1111-1111-201	192.168.1.1	2024-08-08 21:44:30	mac address: none asset type: server
gzAA-1111-1111-202	192.168.1.2	2024-08-08 21:44:30	mac address: none asset type: server
gzAE-1111-1111-112	/api.example.com/v1/users	2024-08-08 21:42:25	port id: gzAD-1111-1111-152
gzAE-1111-1111-113	/api.example.com/v1/users/123/orders	2024-08-08 21:42:25	port id: gzAD-1111-1111-159
gzAE-1111-1111-114	/api.example.com/v1/users	2024-08-08 21:42:25	port id: gzAD-1111-1111-125
gzAE-1111-1111-115	/api.example.com/v1/devices/12345	2024-08-08 21:42:25	port id: gzAD-1111-1111-162
HDFS_1000	HDFS_1000	2024-07-31 00:30:24	software version: 2.7.3 port id: null

Figure 6. Sanning result of big data platform assets and data processing procedures

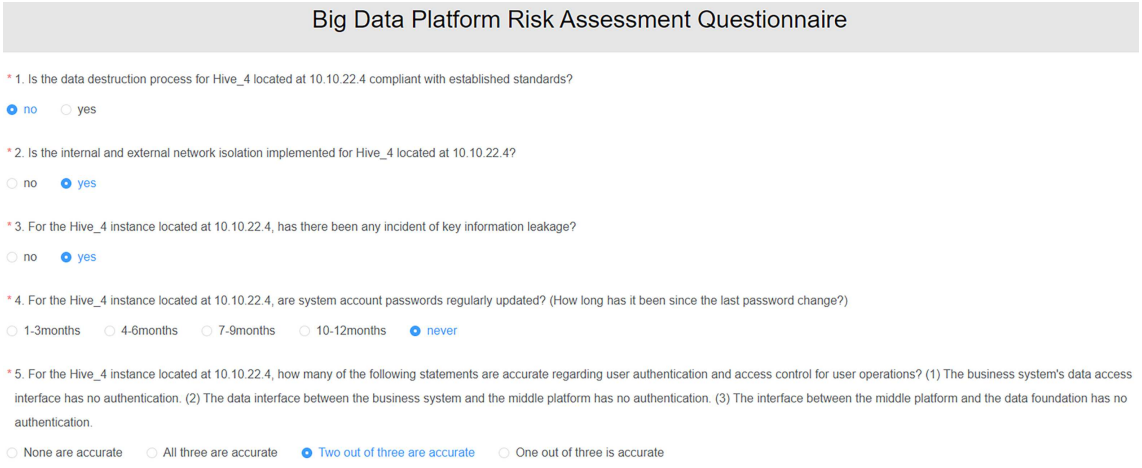


Figure 7. Questionnaire on big data platform

including storage, access authentication, and data collection. The scanning results were categorized and stored in a database for further analysis. As shown in Figure 6, each result is assigned a unique ID, name, last update time, and detailed information. The ID acts as a unique identifier for the scan result in the KG construction, while the name provides a summary of the asset or data processing procedure. The detailed information includes the basic characteristics of the asset or procedure, along with their relationships.

Following the risk quantification assessment tasks, we identified the specific assets and data processing procedures that required monitoring. From the 33 predefined questions, we selected relevant ones and tailored them to the specific asset or data processing procedure by using it as the subject in the generated questionnaire. This approach ensured that the questions were clear and easy for security personnel to understand. The automatically generated questionnaire used in the experiment is illustrated in Figure 7.

In this experiment, a total of 9 types of big data platform assets were detected, including 13 servers, 4 data sources, 12 ports, 1 hardware component, 4 operating systems, 10 data tables, 10 data fields, 2 software applications, and 2 business systems. These assets had a total of 75 identified vulnerabilities. Additionally, 6 types of data processing procedures were detected, including 1 data storage process, 2 data access authentication processes, 5 data collection processes, 2 data transmission processes, 1 data backup and recovery process, and 1 data destruction process. These procedures had a total of 12 identified vulnerabilities.

5.3 KG construction

The asset and data processing procedure risk knowledge graphs (KGs) constructed during the experiment are illustrated as follows. As shown in Figure 8a, in the asset risk KG, nodes represent assets and vulnerabilities, while edges denote relationships between assets or associations between assets and vulnerabilities. A total of 58 assets across 9 categories were identified: red nodes represent 75 vulnerabilities; orange nodes represent 13 servers; light blue nodes represent 4 data sources; light brown nodes represent 12 ports; light green represents 1 hardware asset; light pink nodes represent 4 operating systems; blue nodes represent 10 data tables; yellow nodes represent 10 data fields; light purple nodes represent 2 software instances; and green nodes represent 2 business systems. The data processing procedures risk KG constructed during the experiment is shown in Figure 8b. In this graph, nodes represent data processing procedures and their associated vulnerabilities, with edges indicating the relationships between them. It covers 12 vulnerabilities across six types of data processing procedures: red nodes represent 12 vulnerabilities; green represents the data storage process with 1 node; orange represents access authentication with 2 nodes; yellow represents data collection with 5 nodes; light green represents data transmission with 2 nodes; blue represents data backup and recovery with 1 node; and light purple represents data destruction with 1 node.

5.4 Risk quantitative assessment

During the risk quantification process, we queried the risk KGs to obtain detailed information on various assets and data processing procedures, as well as their corresponding vulnerability information. Using the Random Forest algorithm, we calculated the respective risk values. The risk quantification results for a specific port-type asset are shown in Figure 9. These results include the asset type, IP address, identified vulnerabilities and their risk values, associated risks and their levels, and an overall risk score. Similarly, the risk quantification results for data processing procedures are illustrated in Figure 10, where each procedure's risk value is provided, allowing for the calculation of corresponding risk levels.

In the asset risk quantification results, a higher risk score indicates greater risk. In this experiment, the asset with the highest risk is a port under the IP address 10.10.22.69, which has several identified vulnerabilities. Based on the risk scores associated with these vulnerabilities and the asset risk quantification formulas 1 and 2 presented in Section 4.3.3, it is determined that this port asset presents a high risk of data leakage, privacy leakage, and data corruption or loss, with a total risk score of 283. Similarly, in the data processing procedure risk quantification results, a higher risk score also indicates greater risk. The data transmission process within the big data platform is identified as having the highest risk in this experiment. The vulnerabilities present in this process result in a total risk score of 86, as analyzed using the risk scores and data processing procedure risk quantification formula 3 in Section 4.3.3.

5.5 Discussion

Through experimental evaluations of various components of RiskTree, we have demonstrated its effectiveness in assessing the risks associated with big data platform assets and data processing workflows. Additionally, we have highlighted the advantages of RiskTree when compared to other risk assessment methods, including RiskLens, as discussed in detail below.

Compared with the previous risk quantification system, RiskLens, RiskTree not only delves into multiple subnets, significantly expanding the scope of exploration, but it also quantifies risks for specific attributes and vulnerabilities of an asset or data processing procedure. By calculating the risk score for each individual vulnerability and providing an overall risk score, RiskTree offers a more granular quantification compared to RiskLens, which only considers assets or data processing procedures as a whole and computes a total risk score. This finer granularity enables security analysts to respond to risks more quickly and accurately. Furthermore, RiskTree employs knowledge graphs to visualize the vulnerabilities of assets and data processing procedures, providing security analysts with a more intuitive way to identify the sources of risk within the big data platform.

As shown in Table 7 of RiskLens [1], compared to traditional manual methods of risk data preprocessing and analysis, the risk data collection techniques and risk knowledge graph technology used in this

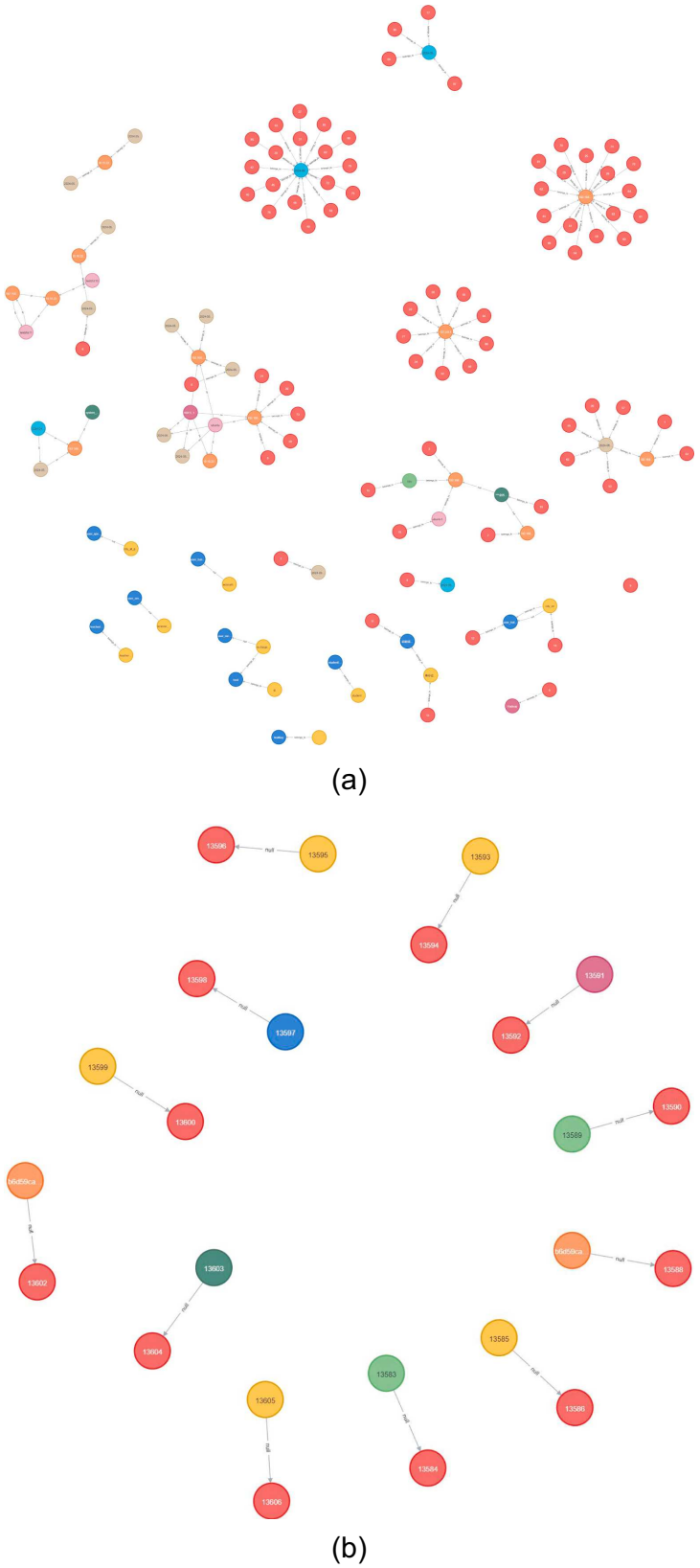


Figure 8. The KG constructed in the experiment: (a) KG of assets risks; (b) KG of data processing procedures risks

ID	Type	IP	Server	Vulnerabilities	Risks	RiskScore
	port	192.168.1.1		Lack of differential privacy mechanisms: 10 Unencrypted storage of data: 10 System hardware or storage media failure: 10 Data untraceable or unauditable: 10 Sensitive data sets not de-identified: 10 Lack of input filtering mechanisms: 8 Lack of user authentication and access control: 10 Role and permission segregation for user access: 10 The principle of least privilege: 4 Support for password brute-force attack prevention mechanisms: 10 Lack of whitelist/blacklist control: 6 Authorization for data destruction: 10 Monitoring and alerting of unauthorized actions: 10 Failure to destroy data caches: 10	Data breach: High Privacy breach: High Data corruption or loss: Medium Data misuse: Low Data control loss: Low	263

Figure 9. Risk assessment result of big data platform asset

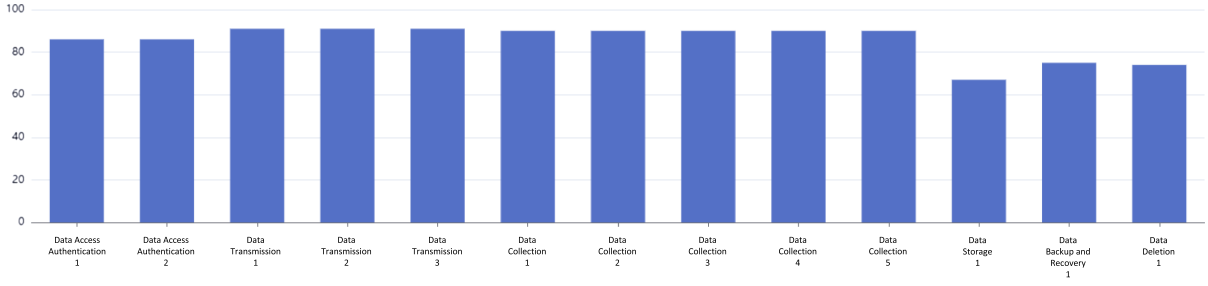


Figure 10. Risk assessment result of data processing procedures

study offer several significant advantages in the field of big data platform risk quantification. By using specialized vulnerability scanning tools to collect risk data from the big data platform, this approach ensures objectivity and significantly reduces the bias that may arise from human subjective interpretations, such as expert evaluations. The automatic processing and analysis of risk data using risk knowledge graphs eliminates the need for extensive manual analysis by security personnel. As a result, our proposed approach significantly reduces operational costs and conserves human resources. Additionally, one of the key advantages of machine learning methods is their suitability for handling large-scale, high-dimensional data, offering strong adaptability and robustness to meet the demands of risk quantification in various big data scenarios. As risk data evolves and updates, the model can be continuously trained and optimized. The Random Forest algorithm dynamically assigns weights to risk indicators, adapting to changes in risk data while maintaining high accuracy and flexibility, enabling effective and timely responses to new risks and the dynamic environment of big data platforms. In contrast, traditional methods such as the Hierarchical Approach, Delphi Method, and Fuzzy Judgment Method rely more heavily on subjective expert judgments, have limited scalability when dealing with large-scale data platforms, and often require time-consuming re-analysis and recalculations when risk data changes. Moreover, their weight allocation is overly dependent on manual settings, which can introduce bias and reduce the accuracy and robustness of the results.

6 Conclusion

This paper aims to build upon RiskLens by further exploring and improving risk quantification assessment for big data platforms. To achieve this, we conducted an in-depth analysis of the vulnerabilities and potential threats inherent in big data platforms and proposed a more scientific and comprehensive risk index system for assets and data processing procedures. We designed and implemented a risk-sensing method that combines vulnerability scanning with questionnaires and developed a risk data integration and visualization approach based on KGs. To validate the feasibility and accuracy of the proposed approach, we set up an educational big data platform. The experimental results on this platform demonstrate the advantages of our approach. First, the risk data is sourced from a locally deployed big data platform, providing a more accurate reflection of real-world conditions compared to the simulated expert scoring data used in the RiskLens. Second, our approach employs more rigorous risk data collection methods,

including vulnerability scanning tools and questionnaires, as well as more advanced risk data analysis and integration techniques, such as risk KGs. Unlike RiskLens, which directly used expert scoring data to train a random forest algorithm, our approach offers a more precise and comprehensive quantification of the risks faced by big data platform assets and data processing procedures. Finally, we introduced risk visualization methods, including risk KGs and risk reports, which allow security personnel to more intuitively and efficiently understand the risk landscape of big data platforms, facilitating the rapid identification of security vulnerabilities and enabling targeted security measures.

Conflicts of interest

The authors declare that they have no conflict of interest.

Data Availability

No data are associated with this article.

Authors' Contributions

Haomou Zhan and Jiawei Yang collaboratively conceptualized and designed the research schemes, and contributed to the drafting of the paper. Zhenyang Guo was in charge of the implementation, which encompassed coding, carrying out experiments, and data analysis. Jin Cao and Wei You provided advisory support throughout the study, offering their extensive expertise in both technical and theoretical dimensions. Dong Zhang made significant contributions to the implementation of the big data platform simulation program and the collection of risk data. Hui Li and Xingwen Zhao reviewed the proposed schemes, providing insightful critiques that enhanced the approaches and the presentation of the findings.

Acknowledgements

Thanks to the anonymous reviewers for their helpful comments.

Funding

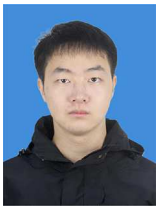
This work is supported by the National Key R&D Program of China (No. 2022YFB3103401), and the National Natural Science Foundation of China (No. 62172317, U23B2024).

References

[1] Zhan HM, Yang JW and Guo ZY et al. RiskLens: A novel way to quantify the risk for big data platform enhanced by machine learning. In: Proceedings of International Conference on Network Simulation and Evaluation, 2023, 228–242.
[2] Special Working Group-Big Data Standard (SWG-BDS). White paper on big data security standardization. Netinfo Secur 2018.
[3] Hu K, Liu D and Liu MH. Research on security connotation and response strategies for big data. Telecommun Sci 2014; **30**: 112–117.
[4] Wu JY, Zhang YF and Xiong S. Construction of enterprise big data security evaluation index system based on SHEL model. Sci Technol Manag Res 2021; **41**: 144–151.
[5] Zhu G, Feng M and Chen Y et al. Research on fuzzy evaluation of privacy risk for social network in big data environment. Inf Sci 2016; **34**: 94–98.
[6] Key Laboratory of Big Data Strategy. Concept and development of big data. China Terminol 2017; **19**: 43–50.
[7] Big data. Commun Technol 2015; **48**: 361.
[8] Liu ZH and Zhang QL. Research overview of big data technology. J Zhejiang Univ (Eng Sci) 2014; **48**: 957–972.
[9] Chen RM. Challenges, values and coping strategies in the era of big data. Proc China Int Inf Commun Exhib 2012; **17**: 14–15.
[10] Zhang K. Four changes of modern universities from the perspective of "4V" of big data. Can Soc Sci 2015; **11**: 292–297.
[11] Big Data Technology and Standard Committee. White Paper on Data Asset Management practices 2023.
[12] Li ZH. Big Data Technology Architecture: Core Principles and Application Practices 2021.
[13] Ghemawat S, Gobioff H and Leung ST. The Google file system. ACM Sigops Oper Syst Rev 2003; **37**: 29–43.
[14] Shvachko K, Kuang H and Radia S et al. The hadoop distributed file system. In: IEEE Symposium on Mass Storage Systems and Technologies, 2010, 1–10.
[15] Dean J and Ghemawat S. MapReduce: Simplified data processing on large clusters. Commun ACM 2008; **51**: 107–113.

[16] Peng YT, Song SY and Fang DY. Research on risk identification and evaluation indicators for big data dissemination process. Sci Technol Manag Res 2018; **38**: 78–82.
[17] Zhu G, Feng MN and Chen Y et al. Research on fuzzy evaluation of privacy risk for social network in big data environment. Inf Sci 2016; **34**: 94–98.
[18] Zhao DM, Zhang YQ and Ma JF. Fuzzy risk assessment of entorpy-weight coefficient method applied in network security. Comput Eng 2004; **30**: 21–23.
[19] Lu HK, Chen DQ and Peng Y et al. Quantitative research on risk assessment for information security of industrial control system. Autom Instrum 2014; **35**: 21–25.
[20] Cheng XY, Wang YM and Liu ZL. A quantitative risk assessment model for information security. J Air Force Eng Univ (Nat Sci Ed) 2005; **6**: 56–59.
[21] Zhang Y, Fang B and Yun X. A quantitative risk assessment approach for host system security. Comput Eng 2005; **31**: 147–149.
[22] Xie F, Lu T and Guo X et al. Security analysis on cyber-physical system using attack tree. In: Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2013, 429–432.
[23] Zhang JL. Network security risk dynamic evaluation method research. Comput Simul 2016; **33**: 356–360.
[24] Feng N, Wang HJ and Li M. A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. Inf Sci 2014; **256**: 57–73.
[25] Li Z, Xu W and Shi H et al. Security and privacy risk assessment of energy big data in cloud environment. Comput Intell Neurosci 2021; **2021**: 1–11.
[26] Dacier M, Deswarte Y and Kaâniche M. Models and tools for quantitative assessment of operational security, Springer, US 1996.
[27] Patel SC, Graham JH and Ralston PAS. Quantitatively assessing the vulnerability of critical information systems: A new method for evaluating security enhancements. Int J Inf Manag 2008; **28**: 483–491.
[28] National Institude of Standards and Technology, Forum of Incident Response and Security Teams. Common Vulnerability Scoring System Version 3.1, 2019, https://www.first.org/cvss/v3-1/
[29] Wynn J, Whitmore J and Upton G et al. Threat assessment and remediation analysis (TARA) methodology description version 1.0, Bedford, MA, 2011.
[30] MITRE Corporation. Common Weakness Scoring System (CWSS) v1.0.1, 2023, https://cwe.mitre.org/cwss/cwss_v1.0.1.html
[31] Eke CI, Norman AA and Shuib L et al. A survey of user profiling: State-of-the-art, challenges and solutions. IEEE Access 2019; **7**: 144907–144924.
[32] Kanoje S, Girase S and Mukhopadhyay D. User profiling trends, techniques and applications. arXiv preprint arXiv:1503.07474, 2015.

Haomou Zhan received the B.S. degree from Xidian University, Xi'an, China, in 2023. He is currently working toward the M.D. in the School of Cyber Engineering, Xidian University, China. His research interests include Knowledge Graph and privacy protection.



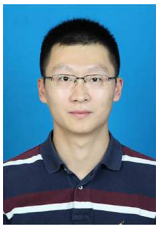
Jiawei Yang received the B.S. degree from Xidian University, Xi'an, China, in 2023. He is currently working toward the M.D. in the School of Cyber Engineering, Xidian University, China. His research interests include machine learning and user authentication.

Zhenyang Guo received the B.S. degree from Xidian University, Xi'an, China, in 2018. He is currently working toward the Ph.D. degree in the School of Cyber Engineering, Xidian University, China. His research interests include machine learning and user authentication.

Jin Cao received the B.S. degree and Ph.D. degrees from Xidian University in 2008 and 2015, respectively. Since July 2015, he has been a professor in the school of Cyber Engineering, Xidian University, Xi'an Shaanxi, China. His interests are in wireless network security and application cryptography.

Dong Zhang is an Engineer at the National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), focusing on directions such as artificial intelligence and cybersecurity.

Xingwen Zhao received the B.S. and M.S. degrees from the School of Telecommunications Engineering, Xidian University, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, in 2011. He is currently an Associate Professor with the School of Cyber Engineering, Xidian University. His research interests include machine learning (or artificial intelligent) based network security, multi-party data sharing, anonymous authentication, broadcast encryption, traitor tracing, key agreement.



Wei You received the Diplôd'Ingéur and Ph.D. degrees from Télécom Bretagne, Brest, France, in 2010 and 2014, respectively. He is currently working as a Lecture with the School of Cyber Engineering, Xidian University since May 2014. His research interests include space information security and security risk assessment.

Hui Li received the B.Sc. degree from Fudan University, China, in 1990, M.A.Sc. and Ph.D. degrees from Xidian University in 1993 and 1998. Since June 2005, he has been the professor in the school of Cyber Engineering, Xidian University, Xi'an Shaanxi, China. His research interests are in the areas of cryptography, wireless network security, information theory and network coding.