大数据技术在数据安全治理中的应用

程伟[1,2]，马成[2]，凌捷[3]
1. 清华大学计算机科学与技术系，北京 100084；
2. 联通（广东）产业互联网有限公司，广东 广州 510320；
3. 广东工业大学计算机学院，广东 广州 510006

摘要
面对新形势下的数据安全治理挑战，顺应数据安全领域的技术发展趋势，针对大型国企在数据安全治理实际应用中突出的关键权限人员识别问题，提出了一种基于图算法的关键权限人员识别技术。该技术可以发现系统中潜在的权限影响因素，并可从多个角度衡量不同含义的权重影响力，识别结果可解释性强。针对数据安全治理中的用户与实体行为异常检测问题，提出一种基于生成对抗网络的用户与实体行为异常检测方法，实验结果表明，所提方法的精确率、召回率和F1值的平均值均优于对比基线模型方法。设计开发了数据安全平台，平台在降低数据安全风险、辅助企业合规建设、促进数据开发利用等方面起到了重要作用，已在多个数据集中管理项目中得到应用，能满足安全场景下的大数据处理需求，具有较好的应用推广价值。

关键词
数据安全治理；图算法；用户与实体行为分析；数据安全平台

中图分类号：TP315 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023074

Application of big data technology in data security governance

CHENG Wei[1 2], MA Cheng[2], LING Jie[3]
1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
2. Unicom (Guangdong) Industrial Internet Co., Ltd., Guangzhou 510320, China
3. School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China

Abstract
Facing the challenges of data security governance in the new situation and following the technological development trends in the field of data security, in response to the prominent issue of identifying key authorized personnel in the practical application of data security governance in large state-owned enterprises, this article proposes a key authorized personnel identification technology based on graph algorithm, which can discover potential authorization influencing factors in the system and measure the weight influence of different meanings from multiple perspectives, The recognition results have strong interpretability. Aiming at the problems of user and entity behavior anomaly detection in data security governance, this paper proposes a user and entity behavior anomaly detection method based on the generative adversarial network. The experimental results show that the accuracy, recall rate and average F1-

score of the proposed method are better than the comparison baseline model method. A data security platform has been designed and developed. The platform has played an important role in reducing data security risks, assisting enterprise compliance construction, promoting data development and utilization, and has been applied in multiple data centralized management projects. It can meet the needs of big data processing in security scenarios, and has good application and promotion value.

Key words
data security governance, graph algorithm, UEBA, data security platform

0 引言

数据安全事关国家安全、社会稳定、经济发展和人民福祉，守护数据安全是信息通信央企的第一责任。近年来国家陆续出台《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《关键信息基础设施安全保护条例》等多项数据安全相关监管法律法规。国家“十四五”规划也明确要求保障国家数据安全，要求做好数据资源全生命周期安全保护，建立数据分类分级管理、数据安全审查、数据安全风险评估、监测预警和应急处置等基本制度。

基于以上国家政策大环境和大型企业数据安全治理应用需求的背景，很多大型企业在积极研究设计数据安全平台，以解决在资产管理、安全防护、日志审计等应用场景下普遍遇到的数据安全治理难题。随着大数据的大规模流转、汇总存储和分析，以及各种大数据技术架构、支撑平台和大数据软件的大范围使用，企业研究数据安全治理相关技术、建设数据安全平台来统筹多维度、全流程的数据安全治理任务，优化各设备的防护策略，强化数据安全治理体系，更具紧迫性。

本文提出了基于图算法的关键权限人员识别技术，该技术可发现系统中潜在的权限影响因素，并可从多个角度衡量不同含义的权重影响力；提出一种基于生成对抗网络的异常检测方法，实验结果表明，所提方法的精确率、召回率和F1值的平均值均优于对比基线模型方法；设计开发了数据安全平台，其在降低数据安全风险、辅助企业合规建设、促进数据开发利用等方面起到了重要作用，相关技术已在广东联通的数据安全治理实际项目中得到应用，取得了良好的社会经济效果。

1 数据安全治理与大数据分析技术

1.1 安全运营大数据的特点

数据安全治理的实践路径一般为“规划–建设–运营–优化”。其中，运营阶段旨在通过不断适配业务环境和风险管理需求，持续优化安全策略措施，确保整个数据安全治理体系有效运转[1]。数据安全平台作为运营阶段的重要工具，可统一管理企业资产信息、安全运营信息、防护日志信息等安全运营数据。根据数据的应用特点，数据安全平台一般使用图数据库、事务型数据库、搜索引擎数据库存储对应信息。

企业资产信息包括主机资产信息、应用系统信息、数据源信息、人员账号信息等。由于各类信息之间存在从属或关联关系，使用图形数据库存储，可在业务中快速依据实体间的关系遍历搜索目标。

安全运营信息包括数据生存周期各阶段部署的安全专用防护设备信息（如VPN

系统、数据库审计、API监测）、设备安全策略的执行情况以及运营工单信息。此类信息存储在事务型数据库中，以支持业务中对信息的频繁更新。

防护日志信息主要是各专用安全设备策略运行的结果日志。各设备每日监测产生大量用户和实体动作日志，这些日志大多是非结构数据，存储在搜索引擎数据库中，以应对在平台进行的中长文本检索任务。

在安全事务中，实时性很重要。数据安全平台在防护日志的解析与分析中采用流处理大数据架构，如**图1**所示。

安全专用设备将产生的日志推送至消息队列后，把编写好的解析程序和分析程序作为消费者的实时读取日志，并将分析的结果（识别的告警或统计的指标）发送至新的消息队列，再经过ETL任务处理后写入搜索引擎数据库，支撑上层应用。基于流数据的大数据架构满足安全日志处理高吞吐、低延迟的性能要求，可做到日志产生一条就分析一条，并可及时感知数据安全态势。

数据安全平台作为安全运营数据的交汇系统，掌握识别、防护、监测、响应场景下的各种类型数据，为大数据分析与挖掘提供基础条件。

1.2 关键权限人员识别技术

数据安全运营通常包括数据采集系统、数据存储系统、数据分析系统和多个业务应用系统。系统种类多，系统任务交叉，有效监控人员对各系统的操作极其困难。普遍的思路是对人员进行分级管理，重点监测关键权限人员，而如何界定和识别关键人员，目前还没有明确通用的标准和方法。

Brin S等[2]提出了一个大规模搜索引擎的原型Google，其模型中包含了一种对搜索引擎搜索结果中的网页进行排名的算法PageRank。其基本假设是：更重要的页面往往更多地被其他页面引用。算法通过对超链接集合中的元素赋权重值，实现“衡量集合范围内某一元素的相关重要性”的目的。该算法实际可以应用于任何存在元素之间相互引用的情况的集合实体。

Freeman L C[3]提出了中介中心性（betweenness centrality）的概念。顶点的中介中心性的定义为：

$$C_B(v)=\sum_{x\neq v\neq y}\frac{\sigma_{xy}(v)}{\sigma_{xy}} \qquad (1)$$

其中，v是一个节点；σ_{xy}是x和y之间最短路径的数量；$\sigma_{xy}(x)$是x和y之间通过v的最短路径的数量。

本文利用上述概念研究适合大型信息通信企业数据安全治理的关键权限人员识别技术。

1.3 安全监控预警技术

监控与审计是防范数据安全风险的重要手段，自动有效的审计方法可以及时阻

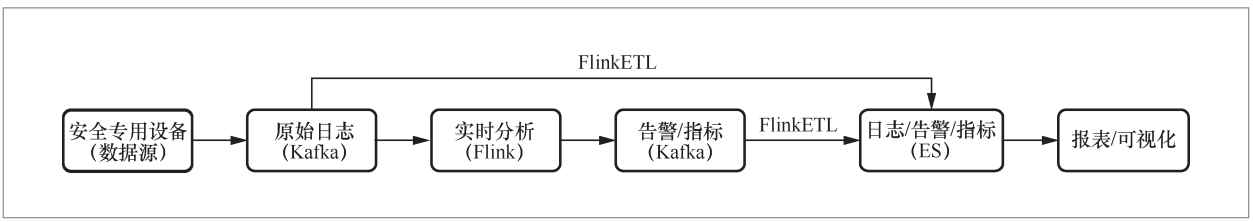

图1 日志数据处理流程

止数据安全风险转变为安全事件。2015年Gartner公司首次提出用户与实体行为分析(user and entity behavior analytics, UEBA)的概念[4],该类技术用于关联分析用户行为和设备日志,以发现潜在威胁或安全问题,当前已广泛应用于企业内部威胁分析和外部入侵检测等任务[5]。

UEBA的分析理念可分为两类:一是利用模式匹配、签名规则等方法对本次新产生日志的时间、IP、动作等内容直接分析。这种分析的特点是单独分析当前日志,即当前日志是否异常与以往的日志无关。

另一类分析方法则考虑利用回归、机器学习等技术,将当前日志与历史日志综合进行分析,参考实体以往的行为表现对当前日志做出是否出现异常情况的判断。目前许多安全产品能够从历史数据中建立(学习)行为基线,方法普遍是直接在无标签数据中做监督学习,这样的操作默认了历史数据中没有异常值,但实际生产环境中并不一定能满足该假设。

2 基于图算法的关键权限人员识别技术

2.1 关键权限人员到关键图节点的转化

一般情况下,企业的资产信息和人员账号信息存储在关系型数据库中,基于这样的数据结构进行普通的统计分析很难发现潜在的关键人员。在数据安全治理实践中,从资产访问权限的角度,将单位内的资产信息和人员账号信息转换为图数据。由此,系统结构中关键权限人员的识别问题被转化为图中关键节点的计算问题。本节重点关注关键权限人员的识别问题。

图由有穷非空顶点集合V和顶点之间边的集合E组成,是用于描述节点之间复杂关联关系的数据结构。本文将能够发放账号的3种资产(应用系统、数据源、主机)以及账号归属的人员作为顶点集合V中的元素,并定义元素之间的有向关系,如下。

- <数据源,应用系统>:该数据源服务于该应用系统。
- <数据源,主机>:该数据源部署于该主机。
- <应用系统,主机,>:该应用系统部署于该主机。

除资产间可能建立的关系外,基于资产发放的账号和账号所属的人员,定义资产和人员之间的有向关系,如下。

- <数据源,人员>:该人员拥有此数据源的访问账号。
- <应用系统,人员>:该人员拥有此应用系统账号。
- <主机,人员>:该人员拥有此主机账号。

基于以上定义,一个可能的关系如**图2**所示。

A应用系统有2个数据源,其中数据源1和A应用系统都部署在主机R上。人员A拥有主机R的访问权限。

2.2 关键图节点的综合识别

本文提出的关键权限人员识别技术将借助相关图算法,从3个不同的维度综合衡量图中节点的重要性。

(1)基于节点影响力的衡量

数据安全治理中发现此类场景:人员A拥有主机K的账号;人员B拥有主机L的账号,同时主机L上部署了一个数据源和一个应用系统,如**图3**所示。

在这种情况下,如果只是以“人员能够访问资产的数量”为标准衡量人员在系

统中的关键程度，将会得出人员A与人员B同等重要的结论（人员A和人员B各自只拥有1个主机账号）。然而基于**图3**显然会认为人员B更重要。当资产数量丰富、资产间层级复杂时，这样有价值的信息更难以发现。

图中任意节点v的权重值称为“v的PageRank”，用符号PR(v)表示。PR值的计算式如下：

$$\mathrm{PR}(v_i)=\frac{1-d}{N}+d\sum_{v_j\in M(v_i)}\frac{\mathrm{PR}(v_j)}{L(v_j)} \quad (2)$$

其中，v_i是目标元素（节点），$M(v_i)$是链入v_i的节点集合，$L(v_j)$是节点v_j链出节点的数量，N是集合中所有节点的数量。d为阻尼系数，表示在任意时刻，该节点向下一个节点链接的概率。

令d=0.9，迭代计算**图3**中每个节点的PR值至收敛后，人员A的PR值为0.038，人员B的PR值为0.070，人员A与人员B的影响力表现出明显差异。在作者单位的具体实践中发现，PR影响力排名中，拥有生产系统访问权限的人高于拥有演示系统访问权限的人。生产系统有更多的链入节点，在结构中更重要，这导致算法认为有权限访问生产系统的人员更关键。

（2）基于信息传递路径的衡量

忽略图的方向，可将资产关系图转为无向图。在数据安全治理中常面临如下场景。

如**图4**所示，在这种情况下，人员A和人员B各自拥有2个数据源的访问权限。我们可能会简单地得出人员A与人员B同等重要的结论。但其实人员B可能更重要，他有权限访问的两个数据源来自两个不同的应用系统，可以说他连接了两个不同的系统，如果删去该节点可能导致图的一部分不连通。

图4中，应用系统A与应用系统B之间的最短路径只通过人员B而不通过人员A，两人之间的中介中心性表现出差异，明显地体现了人员B在信息流控制中更重要。在中介中心性排名中，拥有不同系统资产访问权限的共享运维人员，高于只服务于某个系统的专属运维人员，即便他们运维的

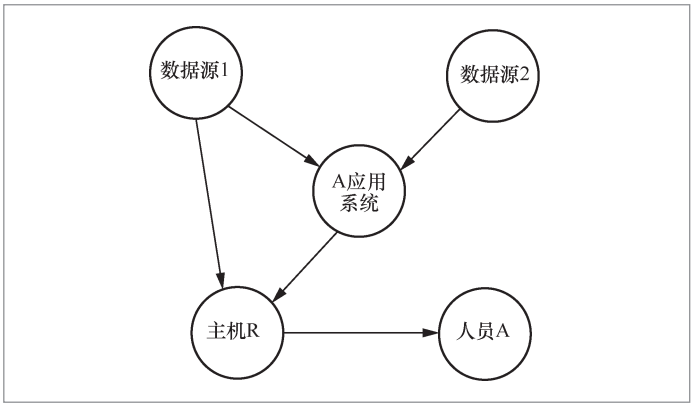

图 2 资产关系图示例

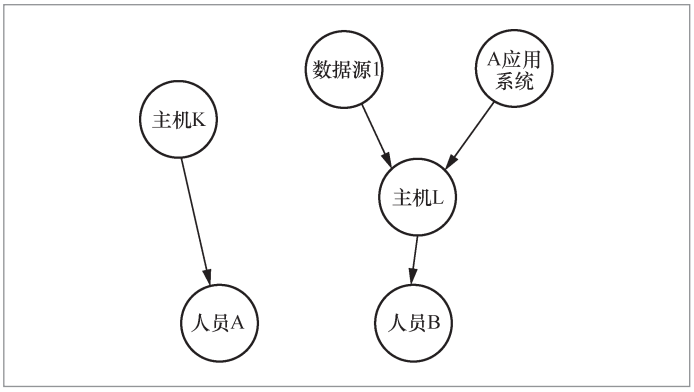

图 3 图结构示例

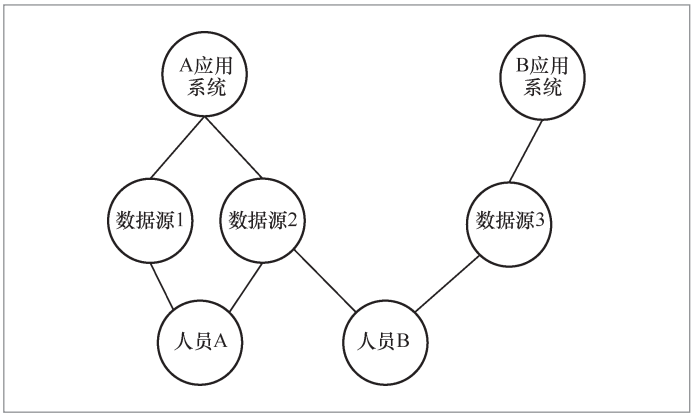

图 4 无向关系图

资产数量几乎相同，这一结果也符合安全运营常识。

（3）基于数据敏感等级的衡量

数据分级化管理已是数据安全领域的共识。工业和信息化部2022年印发的《工业和信息化领域数据安全管理办法（试行）》将数据分类分级管理作为数据安全管理的基础性要求。全国信息安全标准化技术委员2021年发布的《网络安全标准实践指南——网络数据分类分级指引》，按照数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，对个人、组织合法权益造成的危害程度，将一般数据从低到高分为1级、2级、3级、4级共4个级别。

面对关键权限人员识别的问题，本文采纳前期数据分类分级的结果，从被访问数据的敏感等级界定关键权限人员。例如："存有3级敏感数据的数据源为3级敏感数据源。有3级或以上等级敏感数据源访问权限的人，是关键权限人员"。在这种思路下，可以在构建资产关系图时为数据源节点加入"敏感等级"的描述对象（标签），再查询所有敏感数据源链出的节点，即可找到关键权限人员。

进一步，如果认为部署敏感数据源的主机也是敏感主机，则访问敏感主机的人员也是关键人员。查询这样的节点仍然可以从敏感数据源出发，遍历敏感数据源不同深度链出的节点。这样未知深度的遍历查询在关系型数据库中可能需要多层表连接操作，而在图数据库中，此类查询语句变得易于编写，且执行效率更高。

本文提出的基于图算法的关键权限人员识别技术，可以发现系统结构中潜在的权限影响因素，并且可从多个角度衡量不同含义的权重影响力，识别结果可解释性强，符合安全运营逻辑，可以更加科学高效地发现关键权限人员，进而精细化防控数据安全风险。

3 基于生成对抗网络的用户与实体行为异常检测

3.1 用户与实体行为异常检测

用户或实体的行为日志按一定的时间周期可统计为时间序列指标。例如，每个API每小时的平均访问流量、每个账号每天的登录次数。由此，用户与实体的异常行为检测问题可转化为时间序列数据的异常检测问题。

时间序列数据异常检测按训练方式可分为有监督方法和无监督方法。有监督方法需要利用有标签数据建模，这种加入了先验知识的做法本质上是一种预测方法。将模型的预测值与新样本进行比较，从而判断新样本是否存在异常。但是安全实践中多数历史日志并不带有标签[5]，直接使用带有异常值的数据做有监督训练将给下一步模型预测带来误差。无监督方法可以从无标签数据中学习一定的规律，因此无监督方法可能是更适用实际安全场景的模型训练方法。

从学习原理上，时间序列数据异常检测又可分为一般统计学习方法和深度学习方法。不同于进行统计机器学习时需要人工构造训练特征，深度学习包含的自编码器（auto encoder）可以进行自动特征选择，这样的非线性权重模型在特征提取方面具有明显优势。在安全场景中，企业内部环境复杂、使用的设备多种多样，多源异构数据难以融合，在仅有单一统计指标的情况下构造特征困难，深度学习方法在特征提取上有明显的优势。

本文采用的无监督的深度学习方法，较适合生产环境下的安全分析场景，有效地实现了对用户与实体行为的异常检测。

3.2 基于生成对抗网络的方法

2020年Geiger A等人[6]提出了一种利用生成对抗网络(GAN)重建信号并进行异常检测的方法,为时间序列异常检测提供了新的思路。

不同于有监督模型,生成式模型旨在寻找一个数据分布到另一个数据分布的映射[7]。在此处是希望用一个模型捕捉时间序列的低维表示,再用另一个模型从低维空间中重建时间序列,而异常信息则会在数据映射到低维空间时丢失。

图5为Geiger A等人提出的TadGAN训练过程的简化图示。其中信号数据用深色矩形表示。训练时需要学习两个映射函数:编码器E和解码器(生成器)G。编码器E用于将原始信号X映射到潜在特征空间,并输出至解码器G。解码器G同时将编码信号和白噪声信号Z作为输入,生成重建信号G_Z。鉴别器C_X迫使重建信号与原始信号模式相同。鉴别器C_Z迫使编码信号与白噪声Z处于同类潜在空间。这样的生成式训练策略保证了生成器G可以重建原始信号X的分布,而不过拟合X中的异常值。综合考虑鉴别器C_X的得分与重构误差(重建信号C_Z与原始信号间的误差),则可发现原始信号中的异常[6]。

TadGAN方法采用双向长短期记忆模型(LSTM)作为编码器E和解码器G的架构,本文尝试使用基于注意力机制(attention mechanism)[8]的Transformer中的Encoder和Decoder作为编码器E和解码器G的架构,在具体实践中发现并验证了该方法具有更好的效果。为了满足该方法Encoder的输入,本文重新设计了针对安全场景的时间序列矩阵采样方法,将24小时作为滑动窗口大小和步长大小,将采样时间序列将采样的子序列作为单个向量。不同时间间隔的数据集按此规则动态采样。

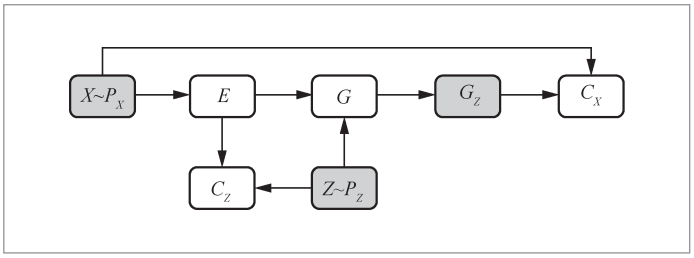

图5 TadGAN训练策略

本文构建了3个数据集进行对比实验。其中,数据集A_1为某API每1小时返回流量的平均值;数据集A_2为某内部安全域每30 min访问流量的总和;数据集A_3为某VPN账号每12小时内登录次数的总和,并由业务安全专家人工标注数据中的异常点。数据集的信息见**表1**。

对于A_1数据集,单个向量的维度为24,A_3数据集单个向量的维度为2。每次依顺序采样7个向量形成矩阵,作为Encoder的输入。下一次则将时间窗口后移24小时再次采样7个向量作为输入。这样设计考虑了UEBA分析中行为动作是以24小时和7天为周期发生的现实情况。

实验的比对对象选择TadGAN(LSTM)、Arima模型(autoregressive integrated moving average model)和某商业UEBA软件DC(匿名)作为基线方法进行实验。对于各模型预测的异常时间窗口,如果预测窗口包含任何已知的异常点,则记录一个TP;如果预测窗口不包含任何已知的异常点,则记录一个FP;如果已知异常点不包含于任何预测窗口,则记录一个FN。

表1 数据集信息

项	A1	A2	A3
数据点数量/个	1 976	3 791	172
异常点数量/个	145	267	14

实验数据来自广东联通实际安全运营生产环境，根据实验记录计算出各模型的精确率（precision）、召回率（recall）和F1-Score，见**表2**。

从实验结果可见，本文方法的召回率、F1值在3个数据集上都有提高，在数据量较少的A3数据集上出现了Arima的精确率略高于本文方法的情况。本文方法的精确率、召回率和F1值3个评估指标的平均值均优于其他3个基线模型的值。

4 数据安全平台设计

数据安全平台是以数据为中心，面向数据全生存周期构建的安全管理与防护体系。其核心是在合规监管和具体业务的驱动下，以数据发现和数据分类分级为基础，以降低数据安全风险为目标，融合多种数据安全技术实现数据安全治理的平台化数据安全防护。

4.1 平台功能架构设计

本文基于广东联通数据安全治理的实际应用背景，设计了数据安全平台。通过数据安全平台，可实现对数据安全能力的集中管理、闭环运营，形成覆盖数据全生命周期安全的纵深防御管理体系。数据安全平台主要由5个中心组成，包括资产管理中心、能力管控中心、分析监管中心、安全运营中心、态势感知中心，平台的功能架构如**图6**所示。

各中心主要包括以下功能。

- 资产管理中心：识别或登记各类资产信息，并基于行业规则实现数据分类分级，形成资产目录、资产关系图。
- 能力管控中心：标准化对接异构安全设备，基于资产分级结果远程下发安全策略，实现多设备信息联动。包括设备统一纳管、状态监控、安全日志数据处理等功能模块，以及数据库加密、数据脱敏、API风险监测、数据分类分级、数据库访问控制等安全工具的纳管对接。
- 分析监管中心：支持低代码创建多种流处理模型，实时分析安全日志，发现威胁与异常行为。包括原始日志、告警管理、风险管理、事件管理、模型管理等模块。支持开箱即用的安全检测分析规则，提供开放式的规则管理模型，贴合安全人员的实际使用需求。
- 安全运营中心：提供国家标准的安全运营流程及量化指标，在线合规审查、生成安全报告、事件闭环管理，综合提升运营效率与准确性。内置流程引擎，实现工单的流程自定义，可一键生成，全程跟进。从而实现从风险、预警、运营事项安排到具体的人员的跟踪处理，实现责任到人，有效跟进风险处置。

表2 各种方法的对比实验

基准	A1			A2			A3			平均		
	精确率	召回率	F1-Score	精确率	召回率	F1-Score	精确率	召回率	F1-Score	精确率	召回率	F1-Score
本文方法	0.80	0.70	0.75	0.75	0.69	0.72	0.68	0.65	0.66	0.74	0.68	0.71
TadGAN(LSTM)	0.80	0.71	0.75	0.72	0.67	0.69	0.66	0.62	0.64	0.73	0.67	0.69
Arima	0.75	0.67	0.71	0.62	0.66	0.64	0.69	0.60	0.64	0.69	0.64	0.66
DC	0.71	0.68	0.69	0.65	0.60	0.62	0.64	0.58	0.61	0.67	0.62	0.64

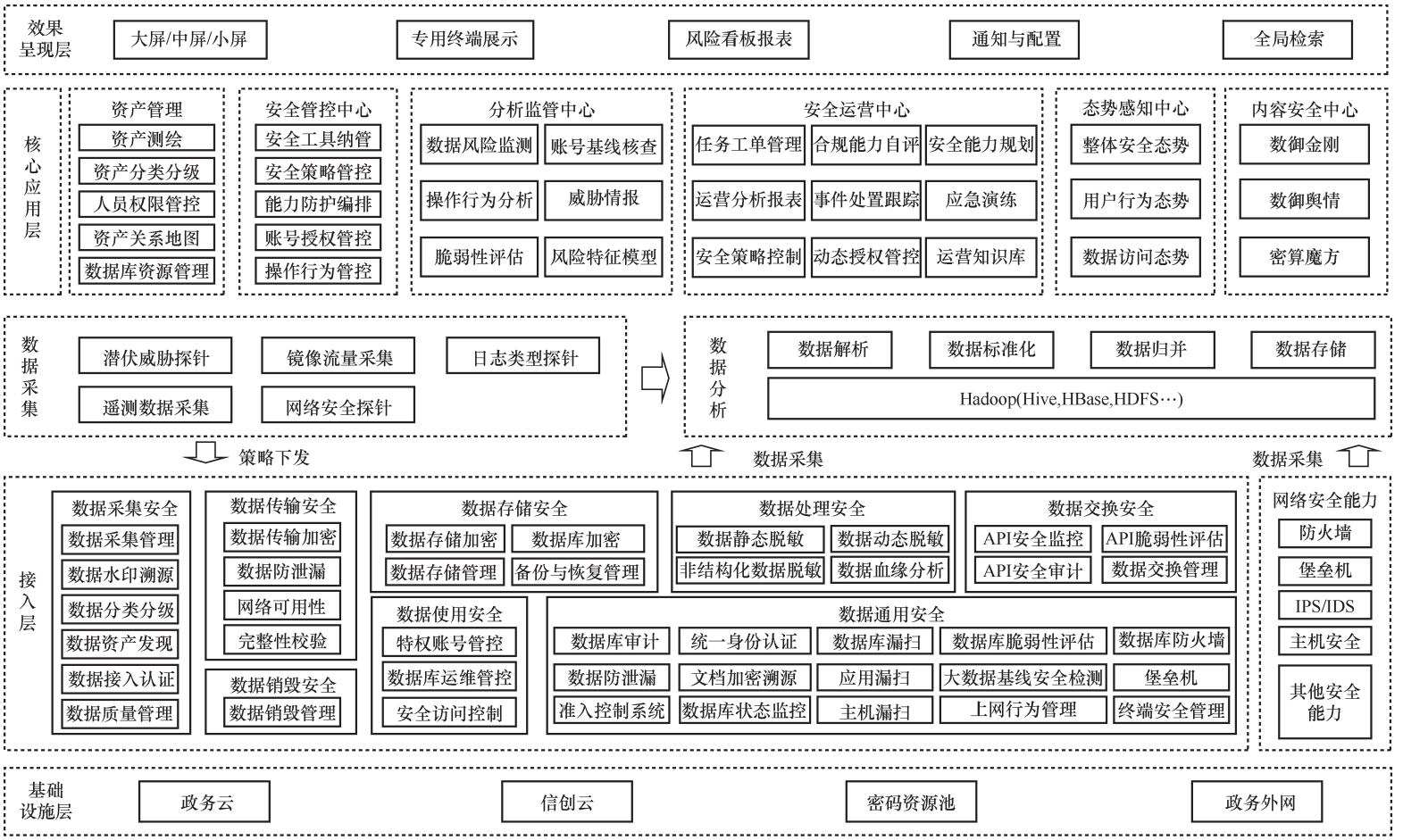
效果呈现层
大屏/中屏/小屏
专用终端展示
风险看板报表
通知与配置
全局检索
核心应用层
资产管理
资产测绘
资产分类分级
人员权限管控
资产关系地图
数据库资源管理
安全管控中心
安全工具纳管
安全策略管控
能力防护编排
账号授权管控
操作行为管控
分析监管中心
数据风险监测
账号基线核查
操作行为分析
威胁情报
脆弱性评估
风险特征模型
安全运营中心
任务工单管理
合规能力自评
安全能力规划
运营分析报表
事件处置跟踪
应急演练
安全策略控制
动态授权管控
运营知识库
态势感知中心
整体安全态势
用户行为态势
数据访问态势
内容安全中心
数御金刚
数御舆情
密算魔方
数据采集
潜伏威胁探针
镜像流量采集
日志类型探针
遥测数据采集
网络安全探针
数据分析
数据解析
数据标准化
数据归并
数据存储
Hadoop(Hive,HBase,HDFS…)
策略下发
数据采集
数据采集
接入层
数据采集安全
数据采集管理
数据水印溯源
数据分类分级
数据资产发现
数据接入认证
数据质量管理
数据传输安全
数据传输加密
数据防泄漏
网络可用性
完整性校验
数据销毁安全
数据销毁管理
数据存储安全
数据存储加密
数据库加密
数据存储管理
备份与恢复管理
数据使用安全
特权账号管控
数据库运维管控
安全访问控制
数据处理安全
数据静态脱敏
数据动态脱敏
非结构化数据脱敏
数据血缘分析
数据交换安全
API安全监控
API脆弱性评估
API安全审计
数据交换管理
数据通用安全
数据库审计
统一身份认证
数据库漏扫
数据库脆弱性评估
数据库防火墙
数据防泄漏
文档加密溯源
应用漏扫
大数据基线安全检测
堡垒机
准入控制系统
数据库状态监控
主机漏扫
上网行为管理
终端安全管理
网络安全能力
防火墙
堡垒机
IPS/IDS
主机安全
其他安全能力
基础设施层
政务云
信创云
密码资源池
政务外网

图 6　平台功能架构

● 态势感知中心：对资产分布、数据风险、用户行为等多类指标进行可视化呈现。

以资产管理中心的数据资产发现功能为例，内置数据资产发现、中间件资产发现能力，且可通过与工具对接，获取安全工具整理的API资产，实现API资产发现。基于平台内置的流程引擎，提供资产认领、工单下发功能，快速流程化地对资产完成责任认领，最终形成数据资产清单，为平台资产梳理提供支持。**图7**是数据资产发现的界面截图。

4.2 数据安全平台的功能特色

本文提出的基于图算法的关键权限人员识别技术和基于生成对抗网络的用户与实体行为异常检测方法，已融合到数据安全平台，平台可登记应用资产、数据源、主机资产、人员账号。其中，新增应用系统时需填写其部署的主机IP地址；新增数据源时，需填写所属应用系统，以及其所在的主机IP地址。登记的实体信息为节点，从属关系为边，直接写入原生多模型数据库。平台后端默认在资产关系发生变动的5小时后自动执行上述3种图算法，每种算法执行完毕后，筛选出人员节点，按计算结果降序排序。

在平台“人员管控”模块中可以选择任一衡量标准（算法结果）排序，清单化管理业务系统中的人力资源。关键权限人员在列表中排名靠前，提醒安全审计员分级管控。另外，信息详情页能够可视化展示任意一名人员在资产关系图中的位置，帮助安全审计员直观了解人员及资产的关联关系，充分发挥图数据结构的优势。

数据安全平台统一收集各设备的监控日志后，可以按一定的时间窗口将某个用户与实体的行为动作统计为时间序列指标。在平台的“模型管理”模块可以对任意已创建的指标应用该算法创建模型，算法的实现采用麻省理工大学Data to AI实验室的开源项目[9]。模型上线后按指定的时间间隔自动进行异常检测。

图 7 平台数据资产发现界面

5 结束语

基于本文方法设计研发的联通“数御”数据安全平台，获评2022年中国信息通信研究院的“数据安全-数据安全治理优秀案例”、2022年度DAMA中国数据治理奖“数据治理优秀产品奖”等奖项。项目研发的技术和平台不仅在广东联通体系内部得到应用，还在地市政务服务数据管理局等政企客户项目中落地应用。对于统筹、集中管理安全资源的客户，通过联通“数御”数据安全平台的资产梳理能力、跨部门层级协作能力以及内嵌的大数据分析与挖掘技术，可有效满足大型企业集中数据安全管控的需求。

在广东联通的数据安全治理应用实践中发现，在安全监管中一次异常可能包含多个设备的上下文信息。生产中若能将同一用户或实体在不同设备中的行为日志关联起来，丰富日志证据链，将有效提升异常检测的准确性。目前，这方面的研究优化工作还在持续推进。

近年来，数据活动日益丰富，数据安全治理愈发复杂，大数据技术在安全领域的作用越来越重要，人工智能技术在自动机器学习（AutoML）、生成式预训练模型（GPT）等方向取得新进展。未来，将进一步探索AI技术进步为数据安全治理带来的新价值，推动数据安全治理向智能化方向发展。

参考文献:

[1] 中国信息通信研究院. 数据安全治理实践指南（2.0）[R]. 2023.
China Academy of Information and Communications Technology. Practical guidelines for data security governance (2.0)[R]. 2023.

[2] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1): 107-117.

[3] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35.

[4] Gartner. Market guide for user and entity behavior analytics[R]. 2018.

[5] 崔景洋, 陈振国, 田立勤, 等. 基于机器学习的用户与实体行为分析技术综述[J]. 计算机工程, 2022, 48(2): 10-24.
CUI J Y, CHEN Z G, TIAN L Q, et al. Overview of user and entity behavior analytics technology based on machine learning[J]. Computer Engineering, 2022, 48(2): 10-24.

[6] GEIGER A, LIU D Y, ALNEGHEIMISH S, et al. TadGAN: time series anomaly detection using generative adversarial networks[C]// Proceedings of 2020 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2021: 33-43.

[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. [S.l.:s.n.], 2017: 5998-6008.

[9] SMITH M J, SALA C, KANTER J M, et al. The machine learning bazaar: harnessing the ML ecosystem for effective system development[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2020: 785-800.

作者简介

程伟（1976- ），男，清华大学计算机科学与技术系博士生，联通（广东）产业互联网有限公司副总经理、高级工程师，主要研究方向为云计算、边缘计算及网络安全。

马成（1996- ），男，联通（广东）产业互联网有限公司软件开发工程师，主要研究方向为数据安全。

凌捷（1964- ），男，博士，广东工业大学计算机学院教授（二级）、博士生导师，兼任广东省大数据安全与服务工程技术研究中心主任、广东省电子政务信创企业重点实验室学术委员会主任等职。主要研究方向为网络信息安全、大数据安全、人工智能安全等，出版相关学术论著4部，在国内外重要期刊和国际会议上发表学术论文100多篇，获授权发明专利超过60件，获广东省科学技术奖一等奖1次、广东省科学技术奖二等奖2次，获南粤教书育人优秀教师等称号。

收稿日期： 2023-05-09
基金项目： 广州市重点领域研发计划项目（No.202007010004）
Foundation Item: Guangzhou Key Field Research and Development Project (No.202007010004)