面向数据空间的数据自感知接入关键技术*

谢云龙[1]，宋雨伦[1]，徐文静[1]，孙　林[1]，高泽昕[2]

（1. 联通数字科技有限公司，北京　100031；2. 北京理工大学，北京　100081）

摘　要：数据空间技术已成为保护数据主权、释放数据要素价值的关键技术之一。针对数据空间中数据出域难、参与方之间缺乏信任、数据共享难、大规模数据交换效率低等问题，探讨了面向数据空间的数据自感知接入技术，旨在提升数据共享、流通、交换和交易的可信性、安全性、透明度和可度量性。研究通过采用 Apache Pulsar 和基于异常识别的感知技术，构建了全流程数据自感知接入平台，为搭建支持海量、多源、异构、多类型数据基础设施提供一种技术架构实现。

关键词：数据空间；智能感知；异常检测

中图分类号：TP311　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2024. 10. 006

引用格式：谢云龙，宋雨伦，徐文静，等. 面向数据空间的数据自感知接入关键技术［J］. 网络安全与数据治理，2024，43(10)：36 - 41.

Key technologies for data self-sensing access in data space

Xie Yunlong[1]，Song Yulun[1]，Xu Wenjing[1]，Sun Lin[1]，Gao Zexin[2]

（1. Unicom Digital Technology Co. , Ltd. , Beijing 100031, China; 2. Beijing Institute of Technology, Beijing 100081, China）

Abstract：Data space technology has become one of the key technologies to protect data sovereignty and release the value of data elements. Aiming at the problems such as data departure from domain, lack of trust between participants, difficulty in data sharing, and low efficiency of large-scale data exchange, this paper discusses the data self-aware access technology in data space, aiming to enhance the trustworthiness, security, transparency, and measurability of data sharing, circulation, exchange, and transaction. The research adopts Apache Pulsar and anomaly recognition-based perception technology to build a full-process data self-aware access platform, which provides a technical architecture for building a data infrastructure that supports massive, multi-source, heterogeneous, and multi-type data.

Key words：data space; intelligent perception; anomaly detection

0　引言

在新一代数字技术的推动下，全球正迅速步入数字经济时代。数字产业化与产业数字化构成了数字经济的核心内容[1-2]。随着《数据安全法》《网络安全法》和《个人信息保护法》的协同实施，以及国家层面《关于构建数据基础制度更好发挥数据要素作用的意见》等政策文件的发布，数据要素市场迎来了迅猛发展的新阶段。在这一背景下，确保数据共享、流通、交换和交易的可信性、安全性、透明度和可度量性已成为业界共识[3]。

可信数据空间（Trusted Data Matrix，TDM）作为一种新兴概念，可被视为数据资源共享的数字化基础设施[4-5]。TDM 旨在促进不同利益相关方在维护数据主权的前提下，实现数据的可信、安全、透明共享与交换。数据空间的概念最初在欧洲提出，并伴随着国际数据空间（International Data Space，IDS）参考架构的发布，为各类企业提供了产品研发的理论基础[6]。

在可信数据空间的基础上，研究者致力于激发各参与方在该空间内流通、流转数据的意愿，以实现数据价值的最大化和资源利用率的提升。然而，在现实应用场景中，由于企业自身数据的敏感性，以及参与方之间缺乏信任，数据共享面临诸多困难，导致数据流转过程中产生不必要的成本[7-8]。

为应对这些挑战，数据智能自感知技术应运而生。这种技术赋予数据主动感知自身状态、环境变化的能力，并据此自动调整和优化其行为或操作。数据智能自感知

*基金项目：国家自然科学基金（62372044）

技术的引入，使数据不再是被动存储和传输的对象，而是具备主动性和适应性，能够根据外部环境和内部变化自主进行相应的操作或决策。将数据智能自感知技术应用于可信数据空间，通过接入多源多类数据，充分利用大数据、微服务等技术实现数据统一的服务接入能力，包括批量数据接入和实时数据接入等。这不仅扩充了数据接入的方式、方法、种类及效率，而且通过封装原子服务以满足不同场景的开发需求，如基于内存数据库或数据传感器的流数据处理。采用混合编排、数据流向依赖、并发与安全控制等技术，实现中间处理逻辑的服务化，使得接入任务能够通过界面化配置、参数配置等快速完成，为数据提供方提供了统一且高效的接入标准。

1 数据空间发展现状

数据空间是一种信息管理的新范式，它不是一种数据集成方法，相反，更像是一种数据共存方法。数据空间解决方案的目标是为所有数据源提供基础功能，无论这些数据源的集成程度如何[9-10]。例如，数字形状采样与处理器可以在其所有数据源上提供关键字搜索，类似于现有桌面搜索系统提供的搜索。当需要更复杂的操作时，如关系式查询、数据挖掘或对某些源进行监控，则可以应用额外的工作，以增量的、“即付即用”的方式更紧密地集成这些源。

数据流通的方式主要包括直连数据包/库/表、API、隐私计算等[11]。但直连数据包/库/表的方式由于数据使用方能够访问到非业务必需的其他原始数据，极大增加了数据泄露的风险。开发专用数据服务接口API，可根据具体的业务数据需求来提供数据，但这种方法不仅涉及较长的开发周期和显著的开发成本，也常因缺乏有效的数据统一监管机制而导致管理上的不足。隐私计算提供了一种在保护用户隐私情况下的“数据可用不可见”的数据共享方式，但共享双方均需要部署隐私计算平台联合建模，成本高、技术难度大。

2 数据接入服务现状

目前常见的数据接入技术选型如表1所示。数据接入作为数据分析和大数据处理的重要阶段，主要包括数据源管理、数据存储、数据计算、数仓服务和数据服务等。数据处理涉及多个方面，包括处理类型（流处理和批处理）、延迟时间（从秒到年不等）、数据服务（如业务指标、监控指标、用户画像、数据分析等）、数据仓库服务（实时数仓和离线数仓）、数据处理技术（如Apache Hive、Presto、Flink和Spark），以及数据处理的相关组件（即时编译、ETL编排、数据存储、任务调度和数据源）。这些元素共同构成了一个全面的数据接入服务框架，旨在帮助企业高效地管理和利用其数据资源，以支持决策制定和业务发展。但传统的大数据平台接入方案经常面临着数据接入分散、接入能力单一以及离线实时数据难以共同处理的难题，具体如下：

（1）接入分散

传统ETL技术中数据接入往往是分散管理的。每个数据源可能需要单独的ETL作业来进行数据抽取和转换，这导致了数据接入流程的复杂化。此外，由于每个数据源的管理和监控是独立的，维护成本增加，数据一致性和准确性也难以保证。

（2）接入能力单一

传统的ETL工具通常针对结构化数据设计，如MySQL、Oracle等关系数据库。对于半结构化和非结构化数据，如日志文件、社交媒体数据、图片等，这些工具往往难以直接处理。因此，这限制了数据接入的多样性和灵活性，使得组织无法充分利用所有可用的数据资源进行数据分析和决策支持。

（3）离线实时共存

传统ETL工具主要设计用于批处理，适合离线数据处理场景。随着业务需求向实时数据处理转变，这些工具往往无法有效支持实时数据流的处理。尽管一些工具尝试通过增加实时处理功能来解决这一问题，但通常这些解决方案在性能和复杂性上仍存在不足。

基于此，本文提出了一种面向数据空间的数据自感知技术，利用Pulsar接入实时/短周期数据，定时将当前时间点数据和前一时间点数据对比，发现差异则定期传输到库中进行数据更新。通过将数据自感知平台集成到数据源的内部环境中（其跨平台能力可屏蔽底层差异），使得应用在不同操作系统上稳定运行，保障数据传输的可靠性和安全性，通过安全认证机制确保数据在传输过程中不被窃取。对于系统的高可用性和可扩展性，数据传感器也提供了解决方案，确保系统在出现故障或需扩展时能够无缝衔接。最重要的是，引入成熟的数据传感器产品能够降低系统复杂度和建设成本，从而缩短项目建设工期，使企业更专注于业务逻辑的发展。

3 数据智能自感知技术体系与架构

3.1 功能架构

数据自感知接入平台如图1所示，以联通链平台为重要依托，基于联通数智链融合创新，结合微服务、分布式身份认证等技术，在数据资源要素流通过程中，完成数据传感器采集、数据适配器接收、数据自感知、自研判和自预警，实现全流程关键信息上链存证，达到数

表 1　数据接入技术选型对比

能力类型	对比项目	Kafka	RabbitMQ	RocketMQ	Pulsar
基础能力	架构	单体	单体	单体	计算存储分离
	消费模式	拉	推 + 拉	推 + 拉	推 + 拉
	重试队列	不支持	不支持	支持	支持
	死信队列	不支持	支持	支持	支持
	海量 Topic	不支持	不支持	支持	支持
	延迟消息	不支持	支持	支持	支持
	堆积能力	海量	一般	海量	海量
	消息回溯	支持	不支持	支持	支持
	多协议支持	私有协议	AMQP、MQTT 等协议	私有协议	私有协议、AMQP、MQTT 等
	安全机制	支持	支持	支持	支持
	事务性消息	支持	支持	支持	支持
运营能力	吞吐量	非常高	一般	高	非常高
	低延迟	一般	非常好	好	好
	可靠性	多副本异步刷盘	主备模式	多副本同/异步刷盘	多副本同/异步刷盘
	一致性	SR 算法	主从模式	主从模式	Quorum 算法
	可用性	较高	一般	较高	高
服务能力	多租户	不支持	支持	不支持	支持
	故障恢复	需平衡数据	横向扩容	需手动同步配置	友好（即时扩容）
	动态扩容	较友好	不友好	不友好	友好
	数据清理	Topic 级别	Topic 级别	集群级别	Topic 级别
	安全机制	身份认证 + 权限	身份认证 + 权限	不支持	身份认证 + 权限
使用场景	典型场景	高吞吐大数据场景	传统业务场景	高性能电商场景	金融场景和大数据场景

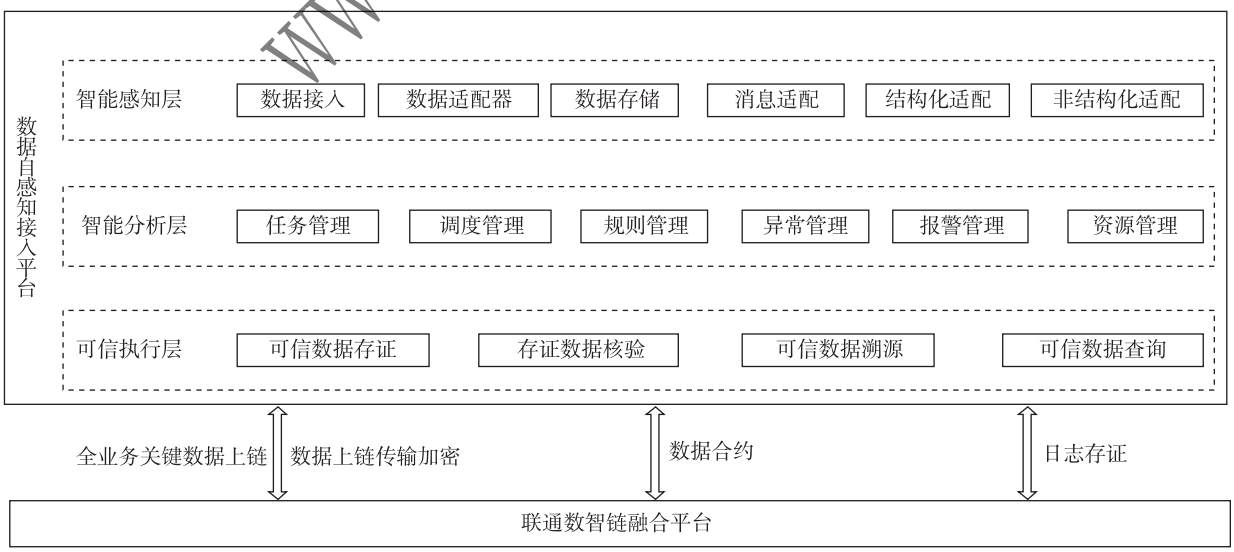

图 1　数据智能自感知功能架构

据自感知接入全流程可存证、可溯源。

智能感知层：利用先进的数据采集技术和感知算法，实时监控和识别数据环境的变化，为智能分析提供实时数据流。主要包括数据接入、数据适配器、数据存储、消息适配、结构化适配、非结构化适配等功能。

智能分析层：采用机器学习和深度学习模型，对收集的数据进行分析和学习，生成数据驱动的洞察和预测。主要包括任务管理、调度管理、规则管理、异常管理、报警管理等功能。

可信执行层：结合可信计算和区块链技术，确保数据处理和交易的安全性和透明性。主要包括可信数据存证、存证数据核验、可信数据溯源、可信数据查询等功能模块。

3.2 数据智能自感知关键技术

在数据智能感知领域，定义数据变化的规则和方法至关重要。本文通过对周期性采集的实时或近实时数据进行系统分类与处理，实现动态状态的实时分析。然后利用异常检测技术及时识别并响应异常情况，保障系统的持续性、稳定性和可靠性。随着人工智能技术的迅猛发展，机器学习和深度学习模型在异常检测领域的应用日益广泛，为系统数据异常的精准判定提供了新的视角与方法论[12-16]。具体而言：

（1）异常识别方法：构建正常数据的行为模型，并设定合理的阈值。通过将当前感知数据与模型的历史状态进行比较，若数据点超出预设阈值，则判定为异常。例如，贝叶斯网络方法通过分析数据的概率分布来进行异常判断，利用贝叶斯定理来计算数据点的概率值，若该值低于预定阈值，则识别为异常。

（2）异常预测方法：基于历史数据的演变规律来预测未来可能出现的异常。通过对数据的时间序列特性、趋势等进行深入分析，构建预测模型，可以在异常发生前进行预警，从而实现系统的及时调整和优化。

本文综合运用上述两种方法，对数值型数据和非数值型数据（如二元数据，取值为 0 或 1）实施差异化的检测策略。对于数值型数据，本研究采用了基于时间序列的元组建模技术，通过比较当前时刻与前一采样点的数值变化，来识别数据的异常波动。通过这种融合策略，提升数据智能感知层的监测能力，实现对系统状态的全面、准确和实时的评估，为数据驱动的决策提供坚实的技术支撑，具体如式（1）所示：

$$\mathrm{Al}_t=\sqrt{(\max\ (\mathrm{oc1}_{vt}-o1_{vn}))^2+(\max(\mathrm{oc2}_{vt}-o2_{vn}))^2+\cdots+(\max\ (\mathrm{oc}N_{vt}-oN_{vn}))^2} \tag{1}$$

其中，Al_t 为 t 时刻数值型数据的状态异常度；$d_{vt}=(\mathrm{oc1}_{vt},\ \mathrm{oc2}_{vt},\ \cdots,\ \mathrm{oc}N_{vt})$ 表示 t 时刻的系统状态感知数据，$\mathrm{oc}i_{vt}$ 表示第 i 个感知对象 t 时刻的数据；$d_{vn}=(o1_{vn},\ o2_{vn},\ \cdots,\ oN_{vn})$ 为上一阶段数据感知状态，oi_{vn} 表示第 i 个感知对象上一时刻的数据；N 为数据的属性个数。

对于非数值型数据，定义第 k 个感知对象状态数据的异常度为 A_{dk}，具体计算如式（2）所示：

$$A_{dk}=\begin{cases}0, & \mathrm{dK}_{vt}=\mathrm{dK}_{vn}\\ 1, & \mathrm{dK}_{vt}\neq\mathrm{dK}_{vn}\end{cases} \tag{2}$$

其中 dK_{vt}、dK_{vn} 分别为当前以及之前的数值。

4 数据智能自感知性能分析应用成效

为了进一步验证本文提出的数据智能自感知技术，搭建了实验平台 A 和 B 做相关测试，具体参数详见表 2。其中平台 A 作为部署路由管理的中心，平台 B 用于部署消息处理节点。考虑到其核心数充足以及为了消除网络宽带速率的影响，实验中的生产者、消费者客户端均运行在实验平台 B 上。

实验使用多个生产者、消费者客户端，模拟重负载情况下面向可信数据空间的数据自感知技术系统消息读写性能，并在相同配置与测试条件下，与 RocketMQ 性能进行对比，说明基于该系统处理机制的有效性。其中，无流量控制版本的 RocketMQ 生产者客户端连续调用消息发送接口向服务端发送消息。面向可信数据空间的数据自感知技术系统使用具有流量控制机制的生产者客户端发送消息。消费者客户端使用消息拉取接口从服务端连续拉取消息。

表 2 实验平台性能测试服务器参数

平台	规格	机器信息
实验平台 A	CPU 型号及频率	Intel（R）Core（TM）CPUi7 - 11700 2.50 GHz
	内核数量	16
	DRAM 大小	16 GB DDR4 内存
	硬盘	WDC WDS500G2B0C - 00PXH0 512 GB
实验平台 B	CPU 型号及频率	Intel（R）Xeon（R）CPU E5 - 2683 v3 2.00 GHz
	内核数量	56
	DRAM 大小	128 GB DDR4 内存
	硬盘	Samsung SSD 970 PRO 512 GB

（1）消息写性能

为了模拟真实的消息处理系统，本文选择在异步写 - 异步发送的场景下进行数据自感知系统性能的测试。

为评估异步写－异步发送模式下，基于 Pulsar 与 RocketMQ 数据自感知系统消息写性能差异，测试消息长度分布在 8 KB、16 KB、32 KB 时写吞吐量与生产者数量的关系，结果如图 2～图 4 所示。

整体而言，发送者数量不变的情况下，消息长度越长，写吞吐量越大；消息长度一定的情况下，在发送者超过一定数量后，写吞吐量先升后降。当发送者数量为 32 个时，写吞吐量会下降的原因在于，随着异步生产者数量增加，其发送的写请求迅速堆积在消息写请求队列中。在 RocketMQ 默认配置下，写请求到达消息处理节点后超过 5 s 消息未落盘则认为处理失败，因此大量消息写请求处理失败。而 Pulsar 优秀的请求流量控制策略使其可以避免写请求过多，以及随之而来的写请求超时导致的系统过载问题，能够更充分地利用非易失性存储器 NVMe 固态盘的写带宽。

（2）消息读性能

在不同消息长度下，对基于 Pulsar 的自感知系统读吞吐量进行测试，消费者数量为 16，测试结果如图 5 所示。基于 Pulsar 自感知系统的读吞吐量部分情况下略低于基于 RocketMQ 的数据自感知系统，但同样显示出基于 Pulsar 自感知系统具有优秀的消息读性能。

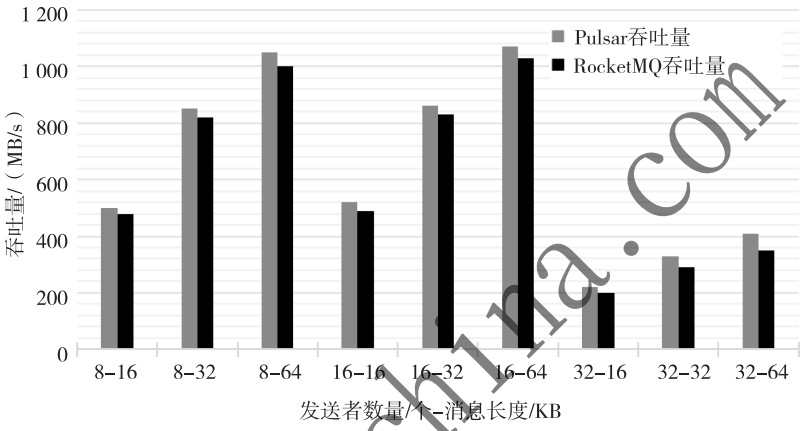

图 2　写吞吐量与生产者数量及消息长度关系

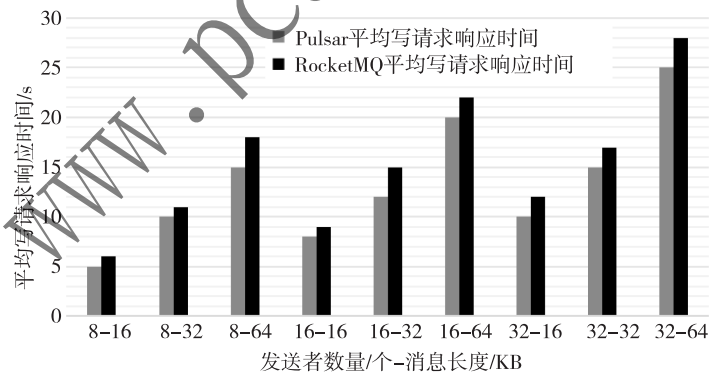

图 3　平均写请求响应时间与生产者数量关系及消息长度对比

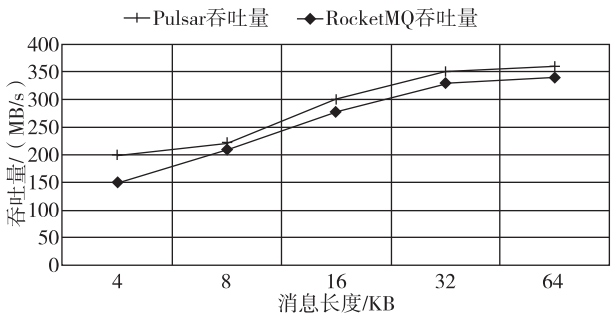

图 4　写吞吐量与消息长度的关系

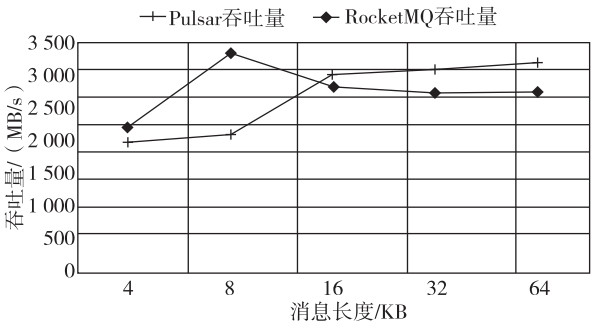

图 5　读吞吐量与消息长度的关系

实践中，在X省股交的“区块链+隐私计算+股权交易”创新应用试点项目中，利用数据自感知技术，实现金融资产交易系统、四板挂牌展示系统和股权登记托管交易系统业务数据轻量化灵活接入。同时利用区块链的防篡改、可追溯、多方共识的特性，以及隐私计算对数据隐私保护的能力，让数据共享交换的各方都参与区块链数据的授权、存储和维护，实现数据变化的实时探知、数据访问的全程留痕、数据共享的有序关联，形成一整套基于区块链的数据智能自感知新秩序。

5 结束语

本文重点介绍面向可信数据空间的数据智能自感知技术，通过利用Pulsar技术高效进行数据采集，通过异常函数智能感知数据变化，不仅能够提升数据处理的安全性和效率，还可以在多个行业中推动智能化转型。未来的研究将集中在优化数据智能自感知算法、提升系统的可扩展性和可靠性，以及探索更广泛的应用场景。

参考文献

[1] 苟兴朝. 数字经济、结构转型与共同富裕——基于30省（区，市）面板数据的经验分析［J］. 西南交通大学学报（社会科学版），2024，25（4）：14－35.

[2] 凌巧. 数字经济、创新活跃度与共同富裕［J］. 统计与决策，2024，40（13）：11－15.

[3] 王澳然，周尚万. 数据要素市场的高质量发展：问题及实现路径——基于马克思流通理论的视角［J］. 当代经济，2024，41（7）：46－52.

[4] 杨云龙，张亮，杨旭蕾. 可信数据空间助力数据要素高效流通［J］. 邮电设计技术，2024（2）：57－61.

[5] 程建润. 打造可信数据空间推动数据价值释放［J］. 软件和集成电路，2024（1）：16－17.

[6] 吕指臣，卢延纯，马凤娇. 数据空间建设：理论逻辑、发展现状与实践路径［J/OL］. 北京工业大学学报（社会科学版），1－14［2024－09－11］. http://kns.cnki.net/kcms/detail/11.4558.G.20240903.1219.004.html.

[7] 王衍之，黄静思，王剑晓，等. 数据要素流通与收益分配机制研究：以风电场景融合气象数据为例［J］. 管理评论，2024，36（6）：30－41.

[8] 尹绮，王成. 数据要素化背景下医院医疗健康数据流通研究［J］. 中国卫生信息管理杂志，2024，21（3）：342－348.

[9] FRANKLIN M，ALON H，DAVID M. From databases todataspaces：a new abstraction for information management［J］. ACM Sigmod Record，2005，34（4）：27－33.

[10] 范淑焕，侯孟书. 数据空间：一种新的数据组织和管理模式［J］. 计算机科学，2023，50（5）：115－127.

[11] 中国信通院. 数据要素白皮书（2023）［Z］. 2023.

[12] SHARMA R，ATYAB M. Introduction to Apache Pulsar［J］. 2022. DOI：10.1007/978－1－4842－7839－0_1.

[13] 刘勇. 基于人工智能的交通流量异常检测与管理［J］. 汽车画刊，2024（3）：194－196.

[14] 赵勇. 基于机器学习的风力发电现场异常检测的应用研究［J］. 价值工程，2024，43（23）：120－123.

[15] 周密，陈烨，焦良葆，等. 基于人工智能的边缘计算设备智能监控和维护系统［J］. 信息化研究，2024，50（4）：66－72.

[16] 程文. 数据驱动的自适应感知方法的研究与实现［D］. 西安：西安电子科技大学，2020.

（收稿日期：2024－08－28）

作者简介：

谢云龙（1976－），男，本科，高级工程师，主要研究方向：大数据、区块链。

宋雨伦（1981－），男，博士，高级工程师，主要研究方向：区块链、数据科学及人工智能。

徐文静（1990－），男，博士，工程师，主要研究方向：区块链、隐私计算以及数据要素流通。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部