铁路大数据治理体系研究*

王 喆
(中国铁道科学研究院集团有限公司电子计算技术研究所,北京 100081)

摘 要:随着铁路的高速发展,铁路行业积累了大量的结构化和非结构化数据并进入了大数据时代。在利用大数据技术全面提升铁路企业核心竞争力及推动铁路转型升级的愿景下,如何开展铁路大数据治理是当务之急。介绍了铁路大数据的基本现状,阐述了数据治理的基本概念和研究现状,提出了铁路大数据治理的总体框架和迭代流程。明确了大数据治理工作不仅是技术上的治理工作,更是包含了人员组织、管理制度、数据架构、应用推动四位一体的均衡的信息化架构的关键一环。该研究对于铁路大数据治理的规划设计、落地实施等具有一定的借鉴意义。

关键词:数据治理;大数据;铁路行业;数据标准
中图分类号:TP3 文献标识码:A DOI:10.19358/j.issn.2097-1788.2022.05.005
引用格式:王喆. 铁路大数据治理体系研究[J].网络安全与数据治理,2022,41(5):30-35.

Research on big data governance system of railway

Wang Zhe
(Institute of Computing Technologies, China Academic of Railway Science Corporation Limited, Beijing 100081, China)

Abstract: With the rapid development of railway, the railway industry has accumulated a large number of structured and unstructured data and entered the era of big data. Under the vision of using big data technology to comprehensively enhance the core competitiveness of railway enterprises and promote railway transformation and upgrading, how to carry out railway big data governance is a top priority. This paper introduces the basic status of railway big data, expounds the basic concepts and research status of data governance, and puts forward the overall framework and iterative process of railway big data governance. It makes clear that big data governance is not only a technical governance work, but also a key part of a balanced information architecture that includes personnel organization, management system, data architecture and application promotion. It has certain reference significance for the planning, design and implementation of railway big data governance.

Key words: data governance; big data; railway industry; data standards

0 引言

在信息技术高速发展的今天,我国智能设备、互联网、物联网技术有重要突破,数据生产和整理能力也正逐步增加,数据规模、数据类型、数据维度有显著提升,大数据的概念应运而生。

大数据是一场革命,使人们的生活方式、工作模式、思维模式发生翻天覆地的改变。大数据成为国家云计算和互联网之后对ICT产业影响最大的技术创新。通过大数据技术的使用,能使组织结构、国家治理模式、企业的决策架构、商业的业务策略以及个人的生活方式等产生深远的影响[1]。大数据最重要的应用领域之一就是预测性分析。以大数据为中心分析数据特征,以此建立合适的模型,适当在模型中增加数据,以此检验数据未来的变动趋势。经验主义将逐渐减少,基于数据的预测将成为决策的主要依据。

1 铁路大数据现状

铁路行业产生的数据主要有以下三个来源:

(1)设备日常监控数据。铁路运输的核心业务可以分为车、机、工、电、辆等几大专业,各专业都

*基金项目:中国国家铁路集团有限公司科技研究开发计划课题(K2021S001)

建设了较为完备的安全监测/监控系统，如客货车安全运行监控的5T系统，机车安全监控的6A系统，接触网安全状态监测的6C系统，监控信号设备运行状态的微机监测系统等。监控类数据以结构化数据为主，数据产生频率较高，累积数据量较大。对日常监测数据开展分析有助于评估设备的实时健康状态，预测设备未来出现故障的概率，挖掘故障原因等。

(2)铁路客货运交易平台积累的交易数据。在高速铁路高速发展的同时，信息化进度不断提速，12306网站和95306网站积累了大量的用户访问日志，订单、支付记录等结构化和半结构化信息，对这些数据的分析将有助于提高网站的运维水平、了解客户的需求、预测未来销售走势以及通过客户的订单来优化运能等。

(3)线路巡检数据。高速铁路开通之前需要进行线路联调联试，对路基、桥梁、隧道、接触网、轨道、信号系统、通信设备、噪声环境等状态进行系统评估与检测；线路运营期间，综合检测车也会对线路进行定期巡检来评估线路整体的健康状态。在联调联试和日常巡检过程中积累了门类丰富的检测数据。线路检测数据以非结构化数据(视频、图像)为主，每年数据增量可达PB级。这些数据是掌握线路整体健康状态，对线路进行全生命周期管理的重要资料，是铁路开展大数据分析的重要方向之一。

上述三种数据在国铁集团、铁路局存储数据规模达55 PB，并且不同类型数据增量显著，众多视频和图片仅满足短期保存需求。现阶段，不管从数据资源总规模、日增数据量看，都标志着铁路已经进入大数据发展时期[2]。

2 数据治理概述

大数据分析往往涉及不同信息系统中的数据融合，除了数据量增长外，数据来源的广泛性、多样性是以往单系统数据分析所无法企及的。数据来源的广泛性带来了数据标准、含义不统一等诸多问题。为了解决上述问题，企业在开展大数据分析前必须引入数据治理体系。文献[3]对大数据治理的概念、治理要素和框架以及面临的挑战进行了探讨，提出了大数据治理的框架；文献[4]将大数据治理体系分为协同筹划、过程实施和监控评估三大主要板块，并对大数据治理的核心功能进行描述；文献[5]认为当前各行各业对大数据治理缺乏整体认识，体系建设不完善，并引入了行业通用的大数据治理体系框架；文献[6]从数据科学技术和管理实践问题两个维度对大数据治理进了讨论，提出了大数据治理的全景式框架，融合了数据生态、数据服务和数据基础；文献[7]将科学技术相关文献、学者动态、论坛热点等非结构化数据纳入科技前瞻大数据分析的数据治理范畴，并构建数据驱动的大数据治理体系，通过LDA模型实现技术趋势预测；文献[8]进而判断大数据对传统政府治理带来的影响，认为大数据治理是传统政府治理走向数字化之路的关键，并提出了大数据时代的政府公共决策体系机制的基本结构；文献[9]对大数据治理机构职能定位及配置进行了归纳，并形成了治理机构评价体系指标体系；文献[10]对大数据治理存在的安全问题进行了分析并提出了应对建议；文献[11]对网络安全中的大数据治理给出了规划建议，指导实施网络安全时如何保护数据。

上述研究多侧重于大数据治理的概念、范畴以及行业通用的治理体系建设。铁路企业是具有一定行业管理职能的传统国有企业，其大数据治理体系的建设需要依据现有信息化水平分阶段分步骤的建设。

3 铁路大数据治理体系

虽然铁路信息系统建设不断改善，但是系统之间选择各自为战，缺乏足够的数据共享量，集成程度相对较差，铁路企业需要进一步做好数据维护工作，便于进行一体化管理。目前数据管理强度落后，缺乏足够的标准化程度，出现数据不一致、数据不精准等问题，数据质量需要在后期运作中不断提升。结合上述现状进行分析，本研究将提出关于铁路大数据的治理体系，具体参考图1。

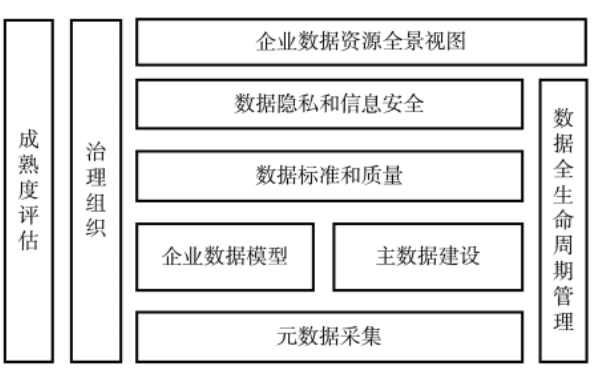

图1 铁路大数据治理体系框架

该框架涵盖了铁路企业大数据治理从认知、组织建设、工作推进及成果展示的相关环节。其中,成熟度评估是对当前企业大数据治理现状进行分析,从而有针对性地建设和调整治理组织,并开展大数据治理各项工作,实现对数据的全生命周期管理,最终通过数据资源全景视图展现治理成果;根据成果的应用反馈再修正当前的企业大数据治理成熟度,成为下一轮治理工作的基础,整个迭代流程如图 2 所示。

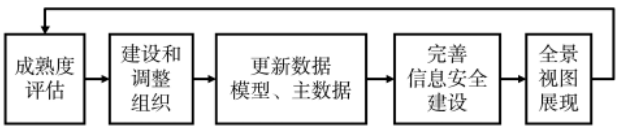

图 2 铁路大数据治理迭代流程

3.1 治理成熟度评估

企业大数据治理成熟度分析是企业大数据治理现状的基础,也是判断企业与最终发展目标距离的依据,可以将大数据治理模型划分成组织、策略、能力等架构。分析铁路企业大数据治理成熟度分为初始阶段、起步阶段、发展阶段、成熟阶段和创新阶段,铁路企业大数据治理成熟度阶段示意图如图3 所示。

铁路从行政架构可分为国铁集团本级、铁路局两级,各级内部又有不同的专业划分,不同的机构、专业之间在人员能力、组织机构、对数据建设的重视程度与现状都是不同的。需要对不同的机构及下属的不同专业部门进行成熟度评估,依据部门现状以及铁路大数据发展的总体规划,制定本部门未来 1~3 年数据治理的目标,并且本着急用先行的策略,找到能力和目标之间的差距,按部就班地实施整体规划。

3.2 治理组织建设

组织机构建设对数据治理过程有重要意义,这也是所有企业共识,也是数据治理的核心。因此,在铁路企业开展数据治理需要在决策层组建由国铁集团高层管理人员组成的数据治理委员会;在领导层,分别由国铁集团信息管理部门以及各业务部门领导、业务专家等人员组建铁路局数据工作小组;在实施层,由各业务部门工作人员和信息系统研发维护人员组成数据治理项目实施组,具体负责数据治理工作同业务系统的对接和实施工作。实施组根据当前企业数据治理成熟度,可以包括主数据工作组、数据全景视图发布工作组、数据质量标准工作组等。整体组织机构如图 4 所示。

3.3 元数据采集

元数据可结合具体用途进行划分:业务元数据、技术元数据。技术元数据主要用于保持系统技术细节,可进行大数据平台和仓库的开发。业务元数据则站在业务的角度分析系统数据,能为使用者、实际系统建立语义层。

目前,铁路主数据中心和各铁路局应用中心运行的信息系统大约 2 500 个。对上述信息系统元数据的采集是构建铁路企业级数据模型,梳理数据标准的基础。通过建设铁路数据服务平台,通过抓取数据库结构或者由信息系统定时推送两种方式实

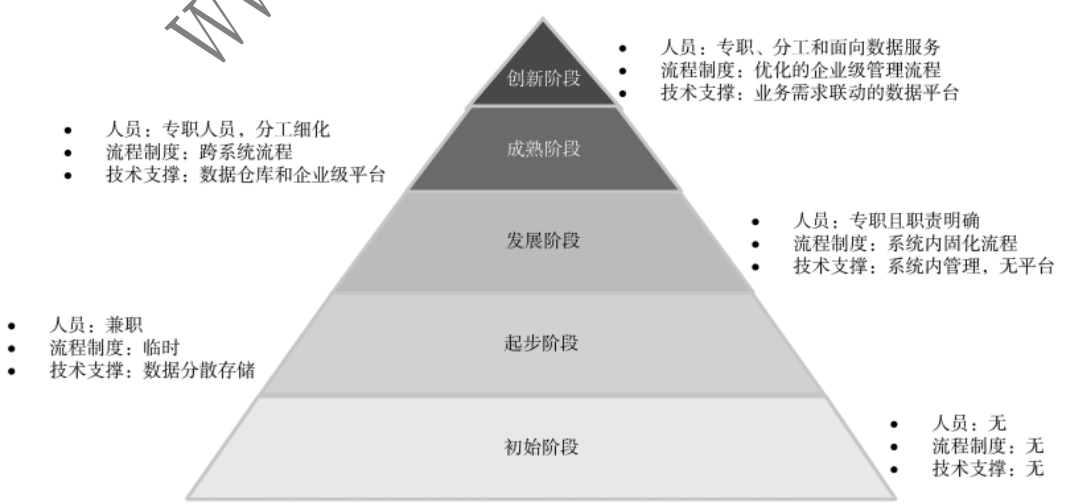

图 3 铁路企业大数据治理成熟度阶段示意图

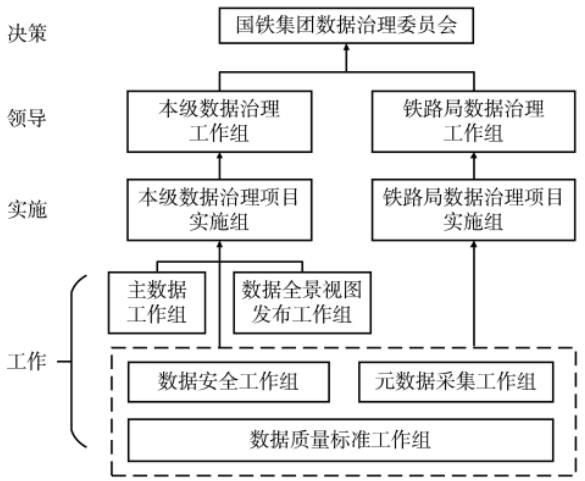

图 4　铁路大数据治理组织架构图

现元数据的采集和更新，在统一平台内部进行汇集整理。

3.4 企业数据模型

铁路行业的信息系统建设面临着业务和信息系统存在差异，出现适应变革灵活性不足的问题，如站段的工务安全生产管理系统中存储了铁轨、道岔等基础设施的日常监测信息，联调联试对这些设备产生的检测数据则存储在另外的信息系统中，这种条块化的 IT 架构造成了信息共享困难、运营和投资成本升高等问题。建设企业级数据模型用于企业的重要业务元素以及这些元素之间的关系，能够清楚地了解企业的数据结构和业务规则，能为 IT 人员和业务人员建立互动平台，是实现业务智能的重要基础。

在建设铁路公司数据模型时，需要划分多个层次：主题域模型、概念数据模型、逻辑数据模型、物理数据模型。主题域模型主要用于判断业务抽象多个实体的相互关系；顶级实体细分成更多子实体后形成概念数据模型；设计出每个实体的属性定义之后形成了逻辑数据模型，通常是满足第三范式的；逻辑数据模型同具体大数据平台的结合形成了物理数据模型。图 5 是本研究提出的铁路企业主题域模型示例。

3.5 主数据建设

近年来，随着铁路信息化建设的逐步深入，信息系统已覆盖客货营销、运输组织、经营管理等各个领域，基础设施及设备检测方面，铁路的工务、电务、供电、车辆和机务等部门积累了铁路线路、通信信号、机车车辆等各种设施设备的海量数据。这些系统之间存在着大量的共用信息，如车型、车号、物资编码、车站名称等。铁路开展主数据管理首先需要判断上述数据要素，并创建数据目录信息；然后，判断主数据管理模式，根据铁路组织机构的特点，核心系统主数据采用集中型管理，次要型系统采用协同性管理的方式更容易实施；之后，还需要确定数据所有者，创建完善的数据管理组织，做好主数

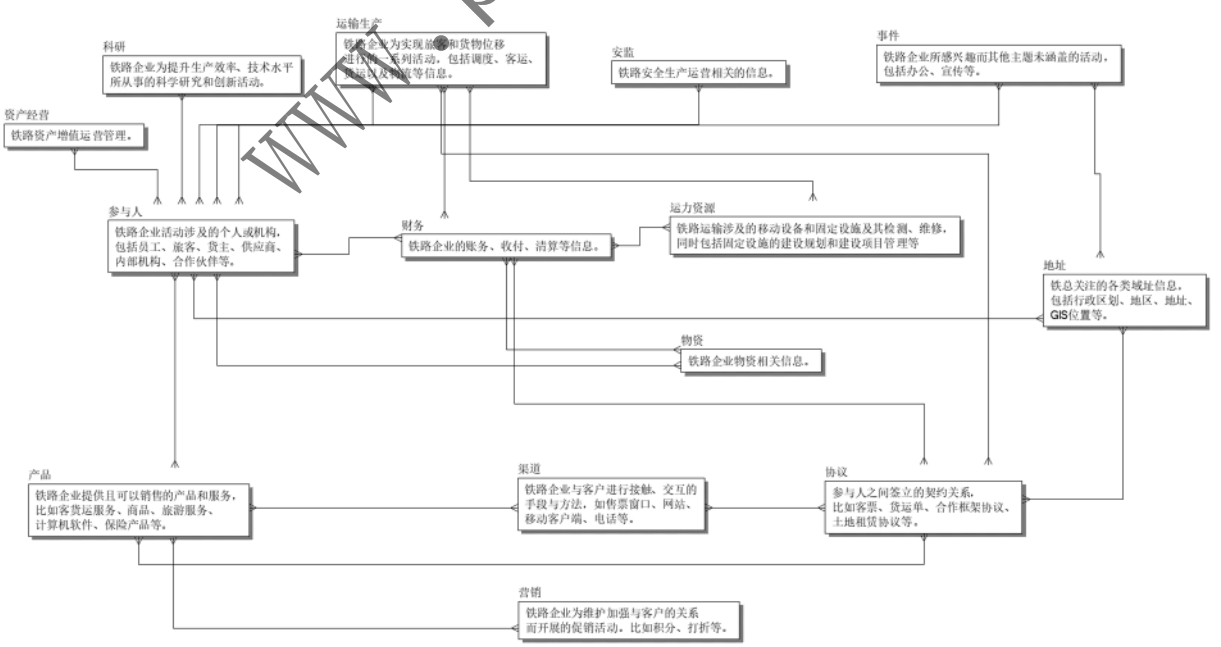

图 5　铁路企业主题域模型

据流程的规范工作;最后,基于以上标准和原则建设主数据管理系统,实现铁路全行业的主数据管理。

3.6 数据标准和质量

建设铁路数据质量管理体系,就是要创建企业数据管理工具,提升数据管理质量,将铁路相关的指标作为切入口,客观分析数据的成熟性,并对数据进行集中抽取,以此满足标准化管理的需求,组织数据稽查工作,提升优化方法的质量,做好数据清洗、数据清除等工作,降低数据多头管理矛盾和问题,进而建立数据资产,通过创建企业数据质量管理制度、管理规范等方式,促进价值数据属性的提升,使业务运营和经营分析质量得到提升。本文基于铁路企业现状提出了数据质量管理的全流程,如图 6 所示。

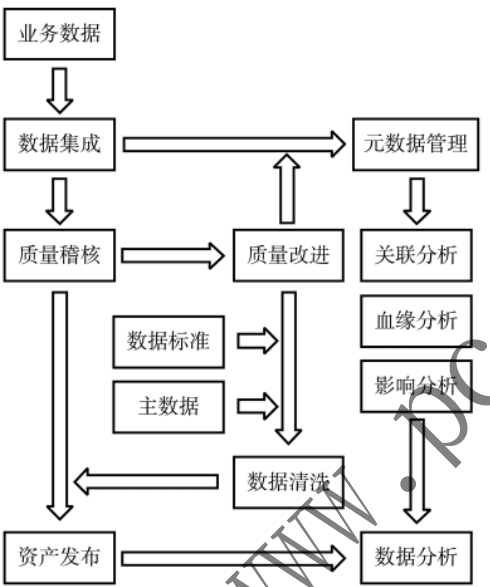

图 6 铁路企业数据质量管理流程设计

铁路大数据场景下,来自各业务系统的数据会统一汇集至铁路大数据服务平台并开展数据质量稽核。数据质量稽核根据定义的数据稽核规则对平台上的数据合规性进行校验,应具备离线分析和内存准实时分析的能力从而处理 TB 级别数据量,并给出数据质量稽核报告。

3.7 大数据隐私与信息安全

铁路业务信息系统中存在着大量的个人隐私数据,包括:乘车人信息、企业职工社保信息、员工履历信息、医疗信息等;还存在着跟铁路企业建设运营相关的涉密数据,如高铁桥梁隧道建设期参数、安全事故信息、设备故障详情等。由于开展大数据分析业务所需的数据集中汇聚,给数据安全带来的更大的安全风险。

对上述敏感数据的保护需要首先建设数据的安全分级体系,包括划定敏感数据范围,指定隐私数据及信息安全管理委员会作为相关责任主体,制定网络安全管理制度、密码安全管理制度、数据备份安全管理制度,划分平台使用人员权限等;其次从技术角度,做好数据访问权限控制,对结构化数据应支持粒度为单元格级的访问控制,不同涉密等级的人员只能访问对应密级的数据,对数据的任何操作和访问都需要被系统记录并存档;另外,还要制定针对特权用户(例如数据库管理员、平台运维人员等)的数据安全管理策略,以监控特权用户对敏感数据的访问,用户对数据的访问记录应以日志的形式存储在大数据服务平台中作为审计依据。

3.8 数据全生命周期管理

实现铁路数据全生命周期管理必须建立数据生命周期管理体系,应采用数据湖的形式存储和管理 PB 级别的数据。数据湖的特点是不对汇聚的数据进行加工,保留原始数据格式,在使用之前根据业务需要开展加工和处理。在大数据量场景下为了节省存储成本,应根据数据的产生时间和使用频度将数据分为冷、温、热数据。冷数据可以采用低成本存储方式,热数据采用高速存储,确保数据的高可用性。另外,还需要制定全路统一的数据生命周期管理,建立一体化的管理标准,针对目前的数据进行更细化的管理,明确管理标准、管理方案、管理制度,使数据管理工作保持科学性、系统性、统一性等。规范中还要定义数据清理原则、数据清理周期以及监督规范执行的人员,从而保证数据全生命周期管理工作的正常运作。

3.9 企业数据资源全景视图

建设铁路数据资源目录系统,不仅可以作为数据治理成果展示的平台,还是企业数据资源共享交换的门户。该门户连接大数据平台,将纳入大数据治理的数据资源以目录的形式对外发布。通过企业级的数据资源全景视图,可以使得企业所掌握的数据资源情况一目了然,是数据交换与共享的基础,也为铁路盘活数据资源提供了有效保证。

4 结论

综上所述,大数据治理工作是一项系统工程,

不可能一蹴而就，从企业评估自身能力开始，到组织机构变革创新、政策制定、流程重建等，都是较为详细的工作项目。从大数据分析的角度看，大数据治理缺少激动人心的业务创新，更多的是枯燥无味、苦练内功的持续投入。大数据治理工作的特点决定了企业大数据业务不可能迅速见效，领导层的决心和企业上下的协调一致是实现数据真正治理以及挖掘大数据价值的不二法门。

参考文献

[1] 大数据标准化白皮书[Z].全国信息技术标准化技术委员会大数据标准工作组，2020.
[2] 邹丹，马小宁，王喆.铁路大数据平台架构研究[J].铁路计算机应用，2019，28(8)：1-4.
[3] 郑大庆，范颖捷，潘蓉，等.大数据治理的概念与要素探析[J].科技管理研究，2017，37(15)：200-205.
[4] 甘似禹，车品觉，杨天顺，等.大数据治理体系[J].计算机应用与软件，2018，35(6)：1-8，69.
[5] 代红，张群，尹卓.大数据治理标准体系研究[J].大数据，2019，5(3)：47-54.
[6] 印鉴，朱怀杰，余建兴，等.大数据治理的全景式框架[J].大数据，2020，6(2)：19-26.
[7] 王俊，王修来，庞威，等.面向科技前瞻预测的大数据治理研究[J].计算机科学，2021，48(9)：36-42.
[8] 廖振民.大数据治理：传统政府治理的变革之道[J].桂海论丛，2018，34(2)：114-119.
[9] 赵豫生，林少敏，郑少翀.大数据治理机构职能及其评价指标体系构建研究[J].中国行政管理，2020(7)：70-77.
[10] 李冬，万磊，费建章.大数据治理中的安全问题研究[J].信息与电脑，2017(6)：192-193.
[11] 杨隆志，李洁，诺伊莉莎，等.网络安全中的大数据治理[J].信息安全与通信保密，2021(7)：56-66.

(收稿日期：2022-10-01)

作者简介：

王喆(1981-)，男，博士，副研究员，主要研究方向为云原生技术、大数据等。

(上接第9页)

less Communications & Mobile Computing Conference (IWCMC)，2018：542-547.
[36] FAN X，GOU G，KANG C，et al.Identify OS from encrypted traffic with TCP/IP stack fingerprinting[C]// 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)，2019：1-7.
[37] SONG J，CHO C H，WON Y.Analysis of operating system identification via fingerprinting and machine learning[J].Computers & Electrical Engineering，2019，78：1-10.
[38] KUMAR A，SONI I，ANAND KUMAR M.Operating system fingerprinting using machine learning[C]//Proceedings of International Conference on Intelligent Cyber-Physical Systems，2022：157-167.

(收稿日期：2022-09-12)

作者简介：

邵磊(1994-)，男，硕士，主要研究方向：网络安全。

余晓(1973-)，通信作者，女，硕士，讲师，主要研究方向：网络管理、云计算、网络安全。E-mail：pp_xyu@seu.edu.cn。

吴剑章(1972-)，男，硕士，讲师，主要研究方向：网络管理、物联网、云计算、虚拟化技术、机器学习，网络安全。

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