导读：当前，数据要素市场的规模化发展正倒逼底层支撑体系的全面升级。从国家宏观布局到企业微观实践，数据基础设施建设已进入加速构建期，但概念边界模糊、制度标准缺位、核心技术能力不足等问题仍制约着整体进程。面对数据“供不出、流不动、用不好、保安全”的深层困境，仅有流通通道的搭建远远不够，更需在生产方式和数据封装方式两个维度实现根本性变革。

数据供给侧的短板正成为制约人工智能发展的突出瓶颈。高质量数据集长期依赖零散化、作坊式的生产方式，难以匹配大模型训练的规模化需求。“数据工厂”概念的提出，正是对这一矛盾的直接回应。作为一种面向人工智能应用的高质量数据集设施化、规模化、标准化生产的数据基础设施，数据工厂以储备、生产、中试的三级架构重构数据供给链路，并通过集中式、半集中式和分布式等多元形态适配不同产业场景，为高质量数据集从手工打造迈向工业化量产提供了切实可行的落地框架。

在数据流通利用的技术路径上，“数据组件”理念为跨域可信流通提供了关键解法。通过将数据要素抽象封装，将数据要素抽象为独立、轻量、可复用的最小单元，呈现出可寻址、可交换、可操作与可管控特性。辅以标准化的注册发现与互操作协议，数据组件具备了统一描述、封装发布、发现获取、动态组合与可信交互等能力，显著提升了跨域传输、访问检索与动态调度的效率，为解决数据孤岛与异构系统兼容难题提供了可验证的技术方案。

从数据组件的轻量化封装，到数据工厂的工业化生产，二者共同指向数据基础设施建设的核心命题：唯有流通侧与生产侧协同突破，方能真正打通数据赋能千行百业的“最后一公里”，支撑全社会数据要素的大规模、可信、高效流通。

——中央财经大学中国互联网经济研究院副院长、博士生导师 欧阳日辉

专栏主编：

欧阳日辉，博士，教授/研究员，博士生导师。现任中央财经大学中国互联网经济研究院副院长、中国市场学会副会长等职。兼任国家数据专家咨询委员会委员、国家数字贸易专家工作组成员、数字广东专家咨询委员会委员、海南省国民经济和社会发展“十五五”规划编制专家咨询委员会委员、清华大学新质生产力研究院专家委员会委员等职。主要研究领域为数字经济、数据要素、数字金融、数字商务，主持国家级课题 4 项，省部级课题 20 余项，发表论著 200 多篇（部）。

张向宏，教授级高级工程师，国务院特殊津贴专家，国家数据专家咨询委员会委员，全国数据标准化技术委员会数据基础设施工作组（WG6）组长，国家人工智能投资基金投资咨询委员会委员，北京、江苏、内蒙古、甘肃、无锡、东莞等 12 省市数据专家委员会委员。担任北京交通大学及北京化工大学特聘教授，主持国家级重大课题研究 50 多项，地方和企业项目研究 100 多项，参与起草国家政策文件研究制定 10 多部。

郝东林，中国移动通信联合会数据要素与区块链委员会副秘书长，多个省份“数据要素×”大赛及典型案例评审专家，清华五道口金融学院未央网专栏作家，著有《WEB3.0 时代数据资产管理》《人工智能重塑世界》等多部书籍，曾先后获评“中国金融科技企业杰出人物”、工信部数据资产运营高级管理师、国际云安全联盟 CSA STAR 云安全注册评估师等。在数据要素、人工智能、云计算、信息安全、合规隐私、风控反欺诈等领域有丰富的一线落地操盘及管理经验。

数据工厂：国家数据基础设施的新兴业态*

张茜茜[1]，殷宏宇[2]，杨　光[3]
（1. 北京物资学院　计算机与人工智能学院，北京　101126；
2. 北京联海信息系统有限公司，北京　100043；3. 中国信息安全测评中心，北京　100085）

摘　要：数据要素化价值化面临“供不出、流不动、用不好”的普遍难题，其核心原因在于数据生产业态尚未成熟，高质量数据集仍以作坊式生产为主，无法满足人工智能大模型对数据的规模化需求。针对这一问题，提出“数据工厂”这一概念，将其界定为面向人工智能大模型应用，开展高质量数据集设施化、规模化、标准化生产的数据基础设施。通过梳理工业社会、信息社会和数智社会基础设施业态的演进规律，论证了数据工厂作为国家数据基础设施基本构成单元的理论逻辑。在此基础上，依据物理分布、组织方式和技术水平等特征，将数据工厂划分为集中式、半集中式和分布式三种类型，并归纳出多样化、设施化、规模化、标准化和人工智能化五大特点。研究认为，发展数据工厂能够有效突破人工智能数据供给瓶颈，推动数据产业链上下游协同，是打通数据赋能人工智能“最后一公里”的关键路径。

关键词：数据工厂；数据基础设施；高质量数据集；数据要素化
中图分类号：F49　　**文献标志码**：A　　**DOI**：10.19358/j.issn.2097－1788.2026.04.001

中文引用格式：张茜茜，殷宏宇，杨光. 数据工厂：国家数据基础设施的新兴业态[J]. 网络安全与数据治理，2026，45(4)：2－8.
英文引用格式：Zhang Qianqian，Yin Hongyu，Yang Guang. Data Factory：an emerging form of national data infrastructure[J]. Cyber Security and Data Governance，2026，45(4)：2－8.

Data Factory：an emerging form of national data infrastructure

Zhang Qianqian[1]，Yin Hongyu[2]，Yang Guang[3]
（1. School of Computer Science and Artificial Intelligence，Beijing Wuzi University，Beijing 101126，China；
2. Beijing Lianhai Information Systems Co.，Ltd.，Beijing 100043，China；
3. China Information Technology Security Evaluation Center，Beijing 100085，China）

Abstract：The valorization of data as a factor of production faces widespread challenges，including insufficient supply，restricted circulation，and ineffective utilization. The core reason lies in the immaturity of data production modes，where high-quality datasets still rely on workshop-style production that fails to meet the large-scale data demands of Artificial Intelligence（AI）large models. To address this problem，the concept of "Data Factory" is proposed and defined as a data infrastructure dedicated to the facility-based，large-scale，and standardized production of high-quality datasets for AI large model applications. By tracing the evolution of infrastructure forms across industrial society，information society，and data-intelligent society，the theoretical logic of Data Factory as a fundamental building block of national data infrastructure is established. Based on characteristics such as physical distribution，organizational structure，and technological sophistication，Data Factories are classified into three types：centralized，semi-centralized，and distributed. Five key features are identified：diversity，facility-orientation，scalability，standardization，and AI-integration. The study concludes that the development of Data Factories can effectively break through the data supply bottleneck in AI development，promote upstream and downstream collaboration in the data industry chain，and serve as a critical path to bridge the "last mile" gap between data and AI empowerment.

Key words：Data Factory；data infrastructure；high-quality dataset；data factorization

*基金项目：北京市社会科学基金（23GLC058）

0 引言

数据是数字经济时代的关键生产要素。2022 年 12 月，中共中央、国务院发布《关于构建数据基础制度更好发挥数据要素作用的意见》（“数据二十条”）[1]，首次从国家制度层面系统部署了数据产权、流通交易、收益分配和安全治理等基础制度框架，标志着我国数据要素化进入制度建设新阶段。2024 年 12 月，国家数据局发布《国家数据基础设施建设指引》[2]，明确提出要构建横向联通、纵向贯通、协调有力的国家数据基础设施体系，为数据要素的大规模流通利用提供底座支撑。与此同时，全球主要经济体也在加快数据基础设施战略布局，欧盟发布《欧洲数据战略》[3]，提出建设欧洲数据空间；美国通过“星际之门项目”大规模投资 AI 基础设施[4]。这些政策实践表明，数据基础设施已成为大国竞争的战略制高点。

然而，数据要素化价值化在实践中仍面临“供不出、流不动、用不好”的普遍难题[5]。一方面，算力、算法和数据作为人工智能的三大要素[6]，在算力和模型技术快速迭代的同时，高质量数据集的供给严重滞后，特别是 2025 年初 DeepSeek 的崛起大幅降低了大模型应用门槛，使得数据供给瓶颈更加凸显。另一方面，长期存储于政府、企业中的私域数据因安全顾虑难以流通，高质量数据集仍以作坊式、分散化方式生产，无法满足大模型对数据的规模化、标准化需求。数据产业链上下游企业难以协同，数据“采而不存、存而不治、治而不用”的现象普遍存在。

从基础设施演进的视角看，在工业社会，水厂、电厂是加工生产战略资源的基本业态；在信息社会，网络厂商、算力厂商承担了类似角色；进入数智社会，数据已成为国家战略资源，但作为数据基础设施基本业态的“数据工厂”尚未形成。现有研究对数据治理[7-8]、数据要素市场化配置[9-10]、数据流通与共享机制[11-12]以及数据确权与价值评估[13]等方面已有较多探讨，但对于如何构建面向人工智能大模型的规模化数据生产设施，尚缺乏系统的理论阐释和概念界定。

正如工业社会水有水厂、电有电厂，数据工厂正在成为数智社会的一种新兴生产业态。发展数据工厂，不仅是顺应全球数智化发展趋势的必然选择，而且对于创新国家数据基础设施新型业态，打造高质量数据集规模化供给设施，推动数据产业高质量发展，打通数据赋能人工智能“最后一公里”等方面，具有重大理论意义和实践价值。

1 数据生产业态不成熟制约数据要素化价值化

由于法律、技术、产业等三方面的障碍还未得到根本破解，当前，数据要素化价值化依旧面临“供不出、流不动、用不好”的普遍难题，突出表现在人工智能大模型面临数据供给瓶颈、高质量数据集难以实现规模化生产、数据基础设施业态千差万别、数据产业上下游协同困难等方面，亟需通过技术、组织、商业模式等多维度的联合创新，探索形成一种适应数智社会发展特点的新型数据生产业态。

1.1 人工智能发展遇到数据供给瓶颈

算力、算法和数据是人工智能的三大要素。当前，人工智能大模型技术水平快速提升，应用领域不断扩大。特别是 2025 年初 DeepSeek 异军突起，大大降低了大模型的应用门槛，拓展了大模型在各行各业的应用范围和深度。但是在初步实现“算力平权”和“模型平权”的同时，数据成为人工智能技术提升和应用拓展的关键瓶颈，特别是长期存储于政府、企业和其他机构中的私域数据。这些数据一方面具有与各种主体的业务紧密相关的高价值特性，是大模型水平提升和应用拓展必不可少的“原料”，另一方面具有与各种主体核心竞争力紧密耦合的高安全特性，是各类主体严防死守的数据安全保障“对象”。大模型若能获取高质量的私域数据，其技术水平将得以提升，应用也将更贴合业务实际需求。因此，高质量数据集的供给已成为人工智能发展的关键。

1.2 高质量数据集以作坊式生产为主

当前，高质量数据集生产还处于零散化、作坊化阶段。一方面，拥有海量数据资源的机构缺乏对大模型数据需求的认知，往往对数据放任不管或开展无明确目标的数据治理，导致数据质量低下，难以应用于大模型训练与推理；另一方面，大模型企业在训练推理过程中采取自采自用的方式，所需数据基本依赖自身采集、加工、处理，仍停留在自给自足的作坊式生产阶段，与现代社会的大生产模式极不相称。当前，大模型技术发展正处于一个关键突破关口阶段，私域数据流通已成为人工智能大模型发展的刚需，作坊式、分散化生产高质量数据集的模式已不能适应大模型技术水平提升的需求，更不能满足各行各业对大模型应用的要求。亟需创新一种新型数据生产方式，设施化、规模化、标准化生产行业高质量数据集，已成为数据企业转型升级的战略选择，并成为人工智能高质量发展的强大助推器。

1.3 数据基础设施基本业态还未成型

人类社会进入工业社会以后，以专业分工为基础

的社会化大生产成为基本生产方式，基础设施作为基础性、全面性支撑底座，在维持社会化大生产过程中发挥着越来越重要的作用，而加工生产各种要素资源的工厂，成为各不同发展阶段基础设施的组成单元和基本业态。在工业社会，水、电、油、气、交通等是国家战略资源，联结这些战略资源的水网、电网、油气管道网、交通运输网等，是工业社会的基础设施，加工生产这些资源的各种水厂、电厂、炼油厂、车辆厂等工厂，是各种工业基础设施的基本生产业态；在信息社会，网络、算力等成为国家战略资源，网络基础设施和算力基础设施是信息社会的基础设施，加工生产网络和算力资源的网络厂商、通信厂商、芯片厂商、服务器厂商、软件厂商等工厂，是网络基础设施和算力基础设施的基本生产业态；在当前的数智社会，数据已成为国家战略资源，数据基础设施成为数智社会的基础设施，专门从事数据采集汇聚、计算存储、生产加工、管理运营等活动，将原始数据资源加工成高质量数据集的数据工厂，正在成为数据基础设施的基本生产业态。当前，作为数据基础设施基本业态的数据工厂还未成型。表 1 展示了不同社会阶段基础设施业态的演进对比。

表 1　不同社会阶段基础设施业态演进对比

对比维度	工业社会	信息社会	数智社会
战略资源	水、电、油、气	网络、算力	数据
基础设施	水网、电网、交通网	网络基础设施、算力基础设施	数据基础设施
基本业态	水厂、电厂、炼油厂	网络厂商、芯片厂商、软件厂商	数据工厂
生产对象	工业产品	网络与算力资源	高质量数据集
生产方式	流水线大规模生产	模块化标准化生产	设施化、智能化生产

1.4　数据产业链上下游企业难以协同

当前，数据要素化还处于初级发展阶段，大多数企业还没有实现业务的数据驱动，数据业务还没有成为企业的主业，企业内部的信息中心、IT 业务流程部门、数据中心等数据相关部门大多数都不是企业的核心部门，多数企业在数据方面的投入还不多，企业数据治理水平参差不齐。同时，数据产业链条包含采集汇聚、计算存储、生产加工、流通服务、管理运营、安全治理等多个环节，这导致企业很难完成数据全流程治理，数据采而不存、存而不治、治而不用的现象普遍存在。亟需打造一种集数据“采存算管用”全流程各环节功能于一体的数据产业新型业态，在自身边界框架内，实现对数据全链条各环节功能的全面覆盖，形成“原始数据进、高质量数据出”的数据生产与创新模式，推动数据产业链上、中、下游在企业内部实现协同融合。

2　数据工厂是数智化发展阶段的基本生产业态

当前，数据要素市场化配置改革还处于发展初级阶段，数据在数字经济中的关键要素作用还未得到充分发挥，其根本原因是还未探索出与数智社会发展相适应的生产方式和商业模式。数据工厂作为集“采存算管用”各种功能于一体的最基本生产单元，将在促进数据要素化价值化方面发挥关键作用。

2.1　数据工厂是国家数据基础设施基本构成单元

国家数据基础设施是一个“供得出、流得动、用得好、保安全”的数据安全流通环境，其根本目标就是在这个安全流通环境中，将高价值的数据流通出来，以赋能人工智能应用。而数据工厂就是通过搭建一个安全环境，将原始数据加工成高质量数据集，并应用于人工智能训练和推理等各种场景。因此，数据工厂本质上就是数据基础设施的最小组成单元，每个数据工厂都是一个“微型”的数据安全高效流通环境，当众多生产高质量数据集的数据工厂建成，并通过相关标准规范实现互联互通后，便形成了横向联通、纵向贯通、协调有力的国家数据基础设施。

2.2　数据工厂是突破人工智能发展瓶颈有效手段

当前，高质量数据集的供给规模远远跟不上算力和模型的技术迭代速度，成为制约人工智能技术水平提升和应用领域拓展的关键瓶颈。一个重要原因是，一方面这些数据是业务发展的“副产品”，数据“采存算管用”等数据治理活动是企业的“副业”，而大多数企业数据加工处理能力弱，高质量数据集生产能力不足而无力提供；另一方面高价值数据往往与企业核心业务紧密耦合，有加工能力的机构很难获得这些高价值数据进行处理并对外提供。建设独立业态的数据工厂，通过数据脱敏、匿名化处理等手段，实现数据与业务解耦，并通过专业化、规模化的社会大生产方式，能实现高质量数据集的规模化生产和供给，从根本上突破人工智能技术水平提升和应用范围拓展的“最初一公里”瓶颈。

2.3　数据工厂是高质量数据集的规模化生产设施

数据工厂建设和运营的最终目标就是能规模化生产高质量数据集。与当前高质量数据集生产方式相比，数据工厂具有三个显著特点：一是设施化。数据工厂

作为国家数据基础设施的基本组成单元，将当前各行业、各企业使用的各种技术、工具进一步规范化、平台化、设施化，形成智能化、自动化的采集汇聚、清洗标注、加工处理、检验测试等工具和设施，将原始数据资源生产成高质量数据集。二是规模化。作为一种独立的数据生产业态，数据工厂不再是传统企业的副业，也不再是人工智能企业不得不兼顾的非核心环节，数据是数据工厂的核心业务，其生产原料是海量的多元异构的原始数据，最终产品是大规模、业态多样、种类繁多的高质量数据集。三是标准化。数据工厂是数智社会的一种通用生产业态和商业模式，在生产工具、生产流程、生产过程等方面的标准化程度较高，其生产的高质量数据集不仅仅要满足个别人工智能企业的需求，更多的是面向各行业各领域各企业的需要，提供全社会各类主体都能使用的高质量数据集。

2.4 数据工厂是数据全生命周期的综合产业形态

当前，数据产业包括数据资源企业、数据技术企业、数据应用企业、数据服务企业、数据安全企业和数据基础设施企业等六种业态，还未形成一个独立的产业形态，对数据要素化价值化的支撑作用还未充分发挥出来。数据工厂通过数据编织、密态计算、隐私保护计算等数据安全流通技术，构建起数据安全流通环境，并从数据采集汇聚的源头做起，通过研发突破数据清洗、标注、增强、检测等核心数据技术，创新形成各种以高质量数据集为主的数据产品和服务。数据工厂的生产流程覆盖了数据产业全部六种业态，是数据产业的综合产业形态。发展数据工厂对加快打造独立数据产业形态，为数据要素化价值化形成全面产业支撑，具有极其重要的意义。

3 数据工厂的涵义、类型和特点

3.1 数据工厂的涵义

数据工厂是人类社会进入数智化发展新阶段，面向人工智能大模型应用，开展高质量数据集设施化、规模化、标准化生产的数据基础设施。数据工厂是国家数据基础设施的基本构成单元，是突破人工智能发展瓶颈的有效手段，是高质量数据集的规模化生产设施，是数据全生命周期的综合产业形态。

3.2 数据工厂的类型

根据物理分布、组织方式和技术水平等特征，数据工厂可以分为集中式数据工厂、半集中式数据工厂和分布式数据工厂三种不同类型，如表 2 所示。数据工厂的三种形式特点各异、各有优势，并将在今后很长一段时间内长期共存。

表 2　三种数据工厂类型对比分析

对比维度	集中式数据工厂	半集中式数据工厂	分布式数据工厂
核心特征	物理集中	技术平台统一、加工场所分散	逻辑统一、物理解耦
数据流动	数据汇聚到中心	平台部署到数据所在地	数据不出域，结果返回
适用场景	大规模通用数据集生产	多区域、多行业场景	国防、金融、医疗等高安全行业
典型代表	美国星际之门、帕西尼超级数据工厂	Scale AI	Palantir

3.2.1 集中式数据工厂

集中式数据工厂是指将“供应—生产—交付”等数据加工生产全过程，都集中在一个固定物理区域内的数据生产业态。集中式数据工厂最大的特点是“物理集中”，即数据资源集中供给、数据生产集中加工、数据产品集中交付。集中式数据工厂是未来五年之内数据工厂的主流方式之一。欧洲数据中心以德国、英国、荷兰、爱尔兰等国为核心，集中部署了数百个超大规模数据中心，并在 GDPR 数据主权法规驱动下持续扩张，2025 年欧盟进一步提出 300 亿美元兆瓦级 AI 数据中心建设计划[3]。帕西尼具身智能超级数据工厂于 2025 年 6 月在天津投用，占地近 12 000 m^2，是全球最大的具身智能数据采集与模型训练基地，通过 150 个标准化采集单元年产近 2 亿条多模态训练数据，将数据采集、加工与交付集中于同一物理空间[14]。库帕思语料超级工厂是上海市组建的国内首家 AI 语料战略性平台企业，以集中化方式将原始数据加工为高质量大模型训练语料，日生成语料接近 1 TB，覆盖具身智能、金融、医疗等多个行业领域[15]。贵州主枢纽存力中心数据要素保障基地依托国家“东数西算”工程，以贵安新区为核心，采用“前店后厂”模式，集中汇聚 47 个重点数据中心，可调度算力达 4.5 EFLOPS、存力 980 PB，是面向全国的算力保障枢纽[16]。这些项目和工厂都是集中式数据工厂的典型代表。

3.2.2 半集中式数据工厂

半集中式数据工厂是指采用相对一致的技术、工具、平台和方法，在不同区域开展不同行业领域、不同应用场景的数据规模化生产方式。半集中式数据工厂最大的特点是“技术平台统一、加工场所分散”，即，将相对统一的技术平台部署在不同区域，生产不同场景应用需求的高质量数据集。半集中式数据工厂是我国数据工厂发展的重点方向，预计在未来五年实

现爆发式增长。大量专业数据标注公司和人工智能企业都将开发数据工厂的统一技术平台，并在全国多区域布局建设。当前的国家数据标注基地也将向半集中式数据工厂转型升级。美国 Scale AI 就是典型的半集中式数据工厂，它通过旗下的 Remotasks 和 Outlier 等子平台，将数据标注业务分别部署到东南亚、非洲等多个区域，依托统一的数据标注引擎、模型评测工具和 AI 应用接口，面向 Google、Meta、OpenAI 等客户以及美国国防部、卡塔尔政府等机构的具体应用场景，提供数据标注、模型评估与 AI 应用开发等全流程数据加工服务[17]。

3.2.3 分布式数据工厂

分布式数据工厂是指数据的“供应—生产—交付”等加工生产全过程分布在一个广泛、分散的网络空间内，原始数据不需要汇聚在一个集中场所，数据加工也不需要集中在一个物理空间，而是通过数据虚拟化和连接器框架等数据编织技术，将分散在不同区域的数据库、数据仓、数据湖中的数据连接在一起，并将查询转换为源系统能够理解的语言，在数据原地进行处理，只将结果返回给平台，形成“逻辑统一，物理解耦”的分布式数据加工生产业态。分布式数据工厂不需要将原始数据移出其原先的数据库仓，可以完全实现“原始数据不出域，数据可用不可见”的数据资源开发利用，特别适用于国防、金融、医疗、交通等安全要求极高行业的高质量数据集生产加工。分布式数据工厂是数据工厂发展的必然发展方向，将在未来五年的技术突破中实现快速增长。众多技术创新创业型企业将持续突破数据编织、数据虚拟化、数据连接框架等数据前沿技术，与此同时，越来越多的政府机关、国有企事业单位及大型龙头行业企业将采用分布式数据工厂模式，既能实现数据的高价值加工流通，又能确保对原始数据的控制权。Palantir 就是典型的分布式数据工厂，其 Foundry 平台通过 Data Connection 框架支持连接 200 多种数据源，在客户原有系统上就地完成批处理、流式处理及虚拟表查询，并借助数据虚拟化实现双向近实时数据流转，无需集中搬迁数据即可完成跨系统整合与分析[18]。

3.3 数据工厂的特点

数据工厂作为数智社会的一种新兴生产业态，目前还处于刚刚起步的萌芽发展阶段，不同企业根据自身基础和条件，开展了形式多样的数据工厂探索实践。总体来看，数据工厂具有多样化、设施化、规模化、标准化、人工智能化五方面显著特点。

3.3.1 多样化

当前，数据工厂的形态表现出多样化特点，包括集中式、半集中式和分布式三种类型。集中式数据工厂是指将数据集中汇聚、集中生产、集中交付或应用的数据生产方式。美国星际之门、欧洲数据中心、帕西尼具身智能超级数据工厂、库帕思语料超级工厂、贵州主枢纽存力中心数据要素保障基地等都是典型的集中式数据工厂。

半集中式数据工厂是指将数据引擎、生产工具、测评平台和 AI 应用平台等组成的一套集成式技术工具，部署在不同区域、不同行业和不同企业，面向具体应用场景对数据进行加工处理。美国 Scale AI 就是典型的半集中式数据工厂。

分布式数据工厂指依托数据虚拟化和数据编织技术，通过支持多种数据源连接协议的强大连接器框架，将数据处理平台直接连接到客户现有数据系统中，在数据原地进行处理，只将结果返回给平台。美国的 Palantir 公司就是典型的分布式数据工厂。

3.3.2 设施化

设施化是数据工厂的一个显著特征，主要表现在三方面：其一，数据工厂本身就是一个设施。数据工厂是国家数据基础设施的最基本组成单元，本质上是一个设施集合。其二，数据工厂的主要构成要素都是设施。构成数据工厂的数据储备、加工、中试等各种车间，其核心组成部分也都是各种工具、平台等设施。其三，数据工厂服务常常以平台设施方式提供。如 Scale AI 的通用数据引擎、Palantir 的 Foundry 平台、库帕斯的语料运营平台等。正如 Palantir 公司 CEO Alex Karp 所说：“我们不是在提供一套工具，而是在构建一个数字时代的基础设施。”

3.3.3 规模化

区别于当前自给自足的高质量数据集作坊化生产方式，数据工厂是一种建立在专业分工基础上的社会化大生产方式，更加适应数智时代对高质量数据集的规模化需求。数据工厂的规模化特征主要表现在三方面：一是数据规模化采集汇聚。数据工厂特别是集中式数据工厂，需要海量的、集中的数据源和供给渠道，确保数据上游供给通畅。二是数据规模化自动加工。数据工厂的数据清洗、标注、增强、合成、管理等活动，绝大部分工作都是由机器自动化、智能化、批量化完成，只有很小一部分最后的判断、决策工作由人工实现。三是数据规模化场景应用。数据工厂生产出来的高质量数据集，无论是应用于人工智能大模型训

练或智能体迭代，还是直接应用于企业生产管理，应用场景越大，用户数量越多，释放价值越大。

3.3.4 标准化

当前，高质量数据集多由人工智能企业或其他用户自行加工生产，不同用户加工生产出来的数据集各具特点，表现出显著的个性化与非标准化特征。而数据工厂生产出来的高质量数据集，要面向社会广大用户服务，不同用户在数据工厂中购买或使用的数据集都是相同的，表现出显著的通用化、标准化特征。数据工厂的标准化特征主要表现在三方面：一是数据资源标准化。数据工厂的上游数据大多数都是多源异构的，都需要按照统一的分类分级、统一的目录化体系、统一的加工处理规范，形成标准化的数据资源体系。二是数据加工处理标准化。数据工厂的数据清洗、标注、增强、合成、管理等活动，都是按照统一的数据加工处理标准实施的。三是数据产品和服务标准化。数据工厂使用统一标准的数据资源、按照统一方法加工处理后，形成的高质量数据集是一种标准化的数据产品和服务，能为不同用户提供服务。

3.3.5 人工智能化

人工智能化是数据工厂的重要特征。数据工厂是数智化时代的新兴业态，人工智能是数据工厂的最常用工具，广泛应用于数据工厂中数据“采存算管用”各环节，如数据采集智能化、数据加工处理智能化、数据质量检测智能化、数据应用智能化等。例如，Scale AI 打造了企业生成式 AI 评估平台、多诺万人工智能平台和生成式 AI 平台，Palantir 推出了 AIP 人工智能平台，这些平台通过接入 OpenAI 等大语言模型，利用 AI 技术实现智能代理与流程自动化。

4 结论

本文针对数据要素化价值化面临的“供不出、流不动、用不好”难题，提出“数据工厂”概念，将其界定为面向人工智能大模型应用，开展高质量数据集设施化、规模化、标准化生产的数据基础设施。研究表明：第一，数据生产业态不成熟是制约数据要素化的核心问题，高质量数据集的作坊式生产模式已无法满足大模型的规模化需求；第二，数据工厂是数智社会基础设施的基本构成单元，集“采存算管用”功能于一体，是突破人工智能数据供给瓶颈、实现数据全生命周期管理的综合产业形态；第三，数据工厂可分为集中式、半集中式和分布式三种类型，具有多样化、设施化、规模化、标准化和人工智能化五大特点，三种类型将长期共存，并以中外典型案例为上述理论提供了实证支撑。建议将数据工厂纳入国家数据基础设施建设规划，推动多种建设模式协同发展，加快数据编织、隐私保护计算等关键技术突破，培育数据要素市场生态。

参考文献

[1] 中共中央，国务院. 关于构建数据基础制度更好发挥数据要素作用的意见［EB/OL］.（2022－12－02）［2026－03－10］. https：//www. gov. cn/zhengce/2022－12/19/content_5732695. htm.

[2] 国家发展改革委，国家数据局，工业和信息化部. 国家数据基础设施建设指引［EB/OL］.（2024－12－31）［2026－03－10］. https：//www. gov. cn/zhengce/zhengceku/202501/content_6996487. htm.

[3] European Commission. A European strategy for data［R/OL］.（2020－02－19）［2026－03－10］. https：//www. europarl. europa. eu/RegData/etudes/ATAG/2021/690527/EPRS_ATA（2021）690527_EN. pdf.

[4] OpenAI. Announcing the Stargate Project［EB/OL］.（2025－01－21）［2026－03－10］. https：//openai. com/index/announcing-the-stargate-project/.

[5] 国家数据局等十七部门. “数据要素 x”三年行动计划（2024—2026 年）［EB/OL］.（2024－01－05）［2026－03－10］. https：//www. cac. gov. cn/2024－01/05/c_1706119078060945. htm.

[6] 中国信息通信研究院. 中国人工智能发展报告 2020［R］. 北京：中国信息通信研究院，2020.

[7] 尚宏利，强国令. 基于可信数据空间的领域数据资产治理研究［J/OL］. 图书与情报，1－12［2026－03－10］. https：//link-cnki-net-s. webvpn. bwu. edu. cn/urlid/62. 1026. G2. 20260224. 1311. 002.

[8] 陆志鹏. 基于数据元件的领域数据治理工程化路径研究[J]. 网络安全与数据治理，2026，45（1）：42－47.

[9] 谷城，张树山，张佩雯，等. 数据赋能：数据要素市场化配置与企业韧性塑造[J]. 中国软科学，2025（12）：166－176.

[10] 李健，薄俊鹏，董雪璠. 新质生产力视阈下数据要素的协同机制与治理创新：综述及展望［J/OL］. 管理评论，1－13［2026－03－10］. https：//doi. org/10. 14120/j. cnki. cn11－5057/f. 20260204. 001.

[11] 朱前涛. 三权分置下数据要素流通的现实困境、体系建构与推进路径[J]. 中国流通经济，2025，39（12）：43－57.

[12] 林浩鼎. 数据要素流通交易中的权责界定与安全责任分配研究[J]. 网络安全和信息化，2025（12）：15－17.

[13] 王琳琳，付丽霞. 数据知识产权登记的实践争议与优化路径——基于 20 个省市数据知识产权登记管理办法的分析[J]. 数字经济与法治，2025（2）：184－207.

[14] 中国新闻网. 全球最大具身智能数据工厂落地天津［EB/

OL]. (2025 - 06 - 23) [2026 - 03 - 10]. https://www.chinanews.com.cn/cj/2025/06 - 23/10436767.shtml.

[15] 新浪财经. 库帕思：专注 AI 语料，以“数据炼金术”赋能大模型时代 [EB/OL]. (2025 - 03 - 05) [2026 - 03 - 10]. https://finance.sina.com.cn/roll/2025 - 03 - 05/doc - inenqtkr1902664.shtml.

[16] 贵州省人民政府. 数据要素加速转化为现实生产力 [EB/OL]. (2024 - 07 - 08) [2026 - 03 - 10]. https://www.guizhou.gov.cn/home/gzyw/202407/t20240708_85073192.html.

[17] Sacra Research. Scale AI: revenue, valuation & funding [EB/OL]. [2026 - 03 - 10]. https://sacra.com/c/scale - ai/.

[18] Palantir Technologies. Connecting to data: Foundry documentation [EB/OL]. [2026 - 03 - 10]. https://www.palantir.com/docs/foundry/data - integration/connecting-to-data.

（收稿日期：2025 - 03 - 11）

作者简介：

张茜茜（1990 - ），通信作者，女，博士，副教授，主要研究方向：数据要素、数据基础设施。E-mail：zhangqianqian@bwu.edu.cn。

殷宏宇（1986 - ），男，本科，工程师，主要研究方向：数据要素、数据管理。

杨光（1980 - ），女，博士，副研究员，主要研究方向：数据要素、数据安全。

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