

面向 AI 的数据工程体系构建：基于 AI-ready DataOps 的范式转型

姜春宇¹ 艾博焕² 郭彦美¹ 崔一妍¹ 尹正¹

(1. 中国信息通信研究院人工智能研究所, 北京 100191;

2. 中国国际技术智力合作集团有限公司, 北京 100053)

摘要:人工智能(Artificial Intelligence, AI)技术的快速发展对数据工程提出了全新要求,推动数据价值形态从“数据资源”向“高质量数据集”跃迁。系统分析了从传统数据研发运营一体化(Data Operations, DataOps)向人工智能就绪(Ready for Artificial Intelligence, AI-ready)数据工程范式转型的内在机理,构建了面向 AI 的数据工程体系框架。研究首先界定了 AI-ready DataOps 内涵,指出其是通过融合 DataOps 理念,面向 AI 场景高效、高质量、安全供给数据的一种工程化方法论。在此基础上,提出了 AI-ready DataOps 能力框架,涵盖研发、交付、运维、运营四大阶段,形成步骤清晰、分工明确的数据集流水线。研究进一步阐述了“盘、建、研、管、运”五步法实践路径,为企业高质量数据集建设提供可操作的实施指南;为国家高质量数据集建设提供了理论框架与方法论指引。

关键词:数据工程;DataOps;高质量数据集;人工智能;AI-ready DataOps

中图分类号:TP311;TP18

文献标志码:A

引用格式:姜春宇,艾博焕,郭彦美,等. 面向 AI 的数据工程体系构建:基于 AI-ready DataOps 的范式转型[J]. 信息通信技术与政策, 2026,52(6):86-96.

DOI:10.12267/j.issn.2096-5931.2026.06.013

0 引言

人工智能(Artificial Intelligence, AI)技术发展催生了对高质量数据集敏捷建设的迫切需求。2024年12月,国家发展和改革委员会等部门联合印发《关于促进数据产业高质量发展的指导意见》,在政策层面提出“高质量数据集”概念,将其定位为支撑《国务院关于深入实施“人工智能+”行动的意见》的重要基础;《关于促进数据标注产业高质量发展的实施意见》进一步指出加强重点行业领域数据标注,建设行业高质量数据集。2025年2月,国家数据局召开高质量数据集建设工作启动会,标志着数据要素建设进入新阶段。这一系列政策文件的发布标志着高质量数据集建设得

到国家高度重视,为其指明产业发展方向。

当前,全球 AI 产业正处于从通用大模型向行业大模型深度演进的关键阶段。以 GPT 系列、Claude、Gemini 等为代表的大语言模型展现出强大的语言理解与生成能力,其背后依赖于海量高质量训练数据。研究表明,模型性能与训练数据的质量和规模呈现显著正相关关系,数据质量对模型效果的影响甚至超过模型架构本身的优化。这一发现从根本上改变了人们对 AI 发展驱动力的认知,即从“算法为王”转向“数据为王”。在此背景下,如何系统性地构建面向 AI 的数据工程体系框架,如何将传统数据工程方法论适配至 AI 数据生产场景,成为学术界和产业界共同关注的议题。

传统数据工程体系主要服务于商业智能(Business

Intelligence, BI) 场景, 其核心任务是实现业务数据整合、清洗与分析, 支撑经营决策和报表统计。这一体系以结构化数据为主要处理对象, 以抽取—转换—加载 (Extract-Transform-Load, ETL) 为核心加工流程, 以数据仓库为主要存储形态。然而, AI 场景对数据工程提出了截然不同的要求: 处理对象从结构化数据扩展至文本、图像、音频、视频等多模态非结构化数据; 加工流程从简单的格式转换延伸至数据清洗、精细标注、数据增强、合成生成等复杂环节; 质量要求从技术属性的一致性、完整性扩展至多样性、代表性、公平性等社会技术属性; 交付形态从静态报表转变为可直接用于模型训练的数据集。这些差异决定了传统数据工程方法论无法直接套用于 AI 数据生产场景, 亟需构建面向智能的数据工程新范式。

针对面向 AI 时代如何敏捷建设高质量数据集的挑战, 本文通过吸收 DataOps 的思想和实践, 创新性地提出 AI-ready DataOps 理念, 构建 AI-ready DataOps 能力框架, 阐明其与高质量数据集建设的内在机理, 论证从传统数据工程向 AI 数据工程的范式迁移路径, 为产业高质量数据集建设提供方法论指引。本文研究采用文献研究、理论建构与比较分析方法, 在吸收 DataOps 核心理念的基础上, 提出面向 AI 的数据工程体系框架, 并系统分析范式迁移的可行性与实现路径。

1 AI-ready DataOps 概述

DataOps 作为数据开发的新范式, 由 Lenny Liebmann 于 2014 年提出, 其核心思想是将敏捷开发、

精益制造与开发运维一体化 (Development and Operations, DevOps) 的理念融入数据工程领域^[1]。这一概念的提出标志着数据工程领域开始借鉴软件工程领域的实践, 寻求数据开发流程的标准化、自动化与持续优化。中国信息通信研究院在其发布的《DataOps 实践指南》中将 DataOps 界定为“通过对数据相关人员、工具和流程的重新组织, 打破协作壁垒, 构建集开发、治理、运营于一体的自动化数据流水线, 不断提高数据产品交付效率与质量”^[2]。如图 1 所示, 这一界定准确把握了 DataOps 的本质特征, 即通过组织变革与技术创新的双轮驱动, 实现数据工程能力的系统性提升。

DataOps 的核心内涵可概括为 4 个维度: 敏捷性强调快速迭代与持续交付, 通过自动化技术实现数据处理流程的快速响应, 缩短从需求提出到产品交付的周期; 精益性追求消除浪费、优化价值流, 以最小资源投入实现最大数据价值产出, 避免重复开发和无效劳动; 协作性打破业务、开发、治理、运维等传统职能壁垒, 构建跨域协同的复合型团队, 实现知识共享与能力互补; 自动化则是实现上述目标的技术基础, 通过工具链集成实现数据流水线的标准化与流程化, 降低人工干预带来的错误率和时延。这 4 个维度相互支撑、相互强化, 共同构成了 DataOps 的方法论内核, 为企业数据能力建设提供了系统性指导框架。

AI 的快速发展对数据工程提出了全新要求, 推动数据价值形态从“数据资源”向“高质量数据集”跃迁, 呈现出三方面显著特征^[3]。其一, 数据需求从业务驱动转向 AI 场景驱动。传统数据工程主要服务于经营

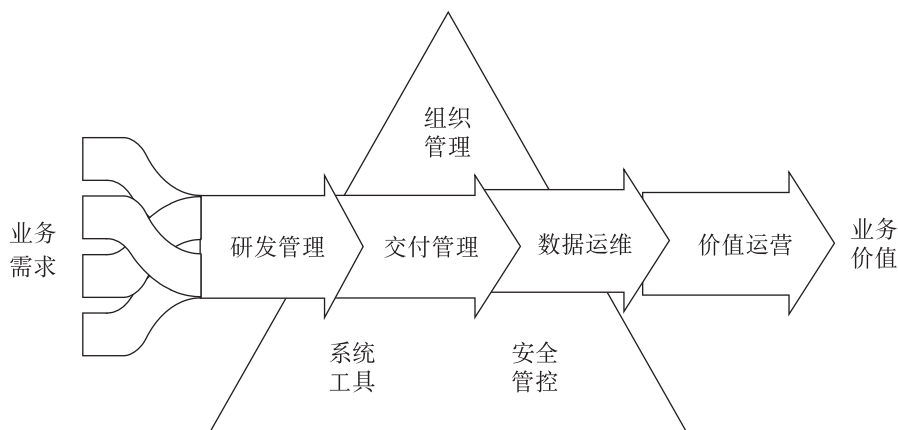


图 1 DataOps 能力框架

分析、决策支持等 BI 场景,数据需求相对明确且稳定;而 AI 场景要求数据工程直接支撑模型训练、微调与评测,数据需求随模型迭代动态变化,对数据的多样性、代表性和规模性提出更高要求。其二,数据加工环节从 ETL 向数据增强与合成延伸。传统数据工程以 ETL 为核心加工环节,主要解决数据格式统一和集成问题;AI 数据工程则需引入数据清洗、标注、增强、缩减、合成等新型加工手段,借助生成式 AI 技术补充真实世界难以获取的极端数据与隐私数据。其三,质量评估从技术属性向社会技术属性拓展。传统数据质量关注准确性、完整性、一致性等技术维度;AI 数据质量还需考虑多样性、代表性、公平性、安全性等社会技术维度,这些维度往往难以完全量化,需要引入人工判断和伦理审查。

AI-ready DataOps,即面向 AI 的数据工程体系框架,是指通过融合 DataOps 理念,面向 AI 场景高效、高质量、安全供给数据的一种工程化方法论。该理念鼓励企业加强内部组织、人员和技术协作,通过标准化、流程化、敏捷化的方式向 AI 场景供给数据,打造高质量数据集。AI-ready DataOps 并非对 DataOps 的颠覆替代,而是其在 AI 时代的范式拓展与能力升级,核心在于治理内容延伸与智能化工具引入。该概念包含 3 层要义:目标层面,以高质量数据集生产为核心产出,支撑基础语料构建、知识工程与提示词优化等 AI 数据需求;方法层面,延续 DataOps 的流程驱动与治理前置思想,将安全合规、伦理审查等要求嵌入数据全生命周

期;技术层面,构建涵盖数据采集、加工、测试、发布、运维、运营的一体化流水线,实现数据资源向 AI 可用数据资产的高效转化。

2 AI-ready DataOps 能力框架

基于产业研究和实践,本文提出 AI-ready DataOps 能力框架。框架覆盖从数据供给到多端消费的全链路,配套建设数据管理与运营平台、相关标准规范以及人才培养方案,对 AI 所需的高质量数据集建设充分赋能。框架以流程驱动为核心,将整个高质量数据集的研发过程分为研发、交付、运维和运营四大模块,针对四大模块进行分解,最终形成一条步骤清晰、分工明确的数据集流水线。如图 2 所示,这一框架设计充分考虑了 AI 场景的特殊需求,将 DataOps 的核心理念与 AI 数据生产的技术特性相结合,为企业构建面向 AI 的数据工程体系提供了可操作的实施框架。

2.1 研发阶段

研发阶段是以开发“AI 训练所需”的数据为目标,构建数据研发治理一体化流程。如图 3 所示,研发管理主要包括需求管理、设计管理及数据加工三部分。研发阶段是整个数据工程体系框架的起点,其质量直接决定了后续各环节的效率和产出质量。

需求管理需首先拉通业务、算法、开发等团队成员,对于 AI 应用场景中的数据需求依据需求采集模板进行标准化收集,而后结合业务价值、技术可行性、合规风险等维度建立评估模型,对需求进行分析评估。

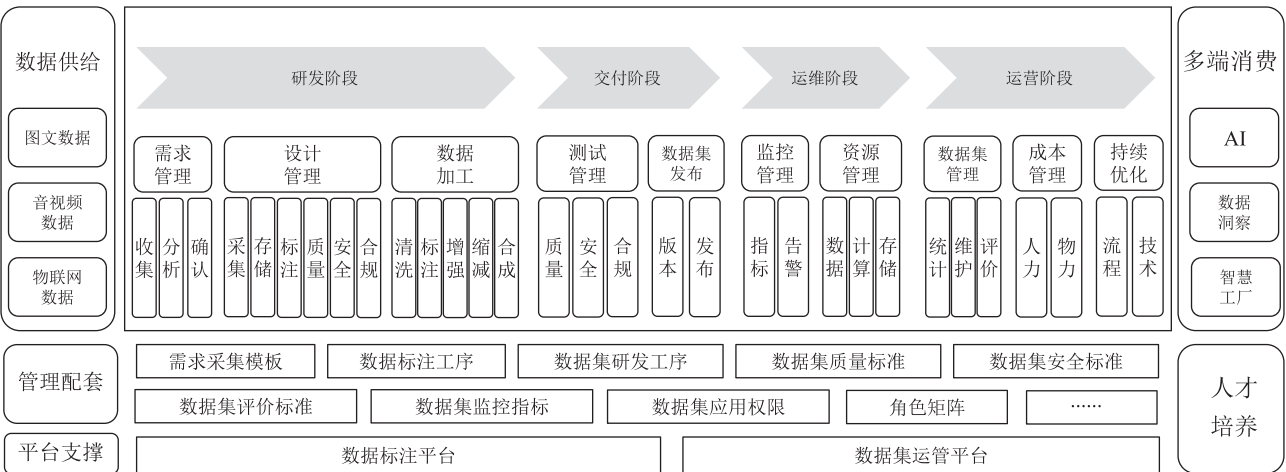


图 2 AI-ready DataOps 能力框架

最后明确需求优先级并与业务团队确认交付标准。需求管理的核心在于建立标准化的需求收集和评估机制,确保数据研发工作与业务目标和 AI 应用场景紧密对齐。具体而言,数据需求采集模板应涵盖数据集的目标应用场景、预期规模、质量要求、交付时间、使用权限等关键信息,为后续的设计和加工环节提供明确的输入。需求评估则需要综合考虑业务价值、技术可行性、资源投入、合规风险等多个维度,建立科学的评估模型,确保资源向高价值需求倾斜。

设计管理阶段,DataOps 的治理前置思想需要组织在数据研发阶段就对后续数据采集、存储、标注等环节进行合理设计。除了传统的数据质量、安全外,AI-ready DataOps 还要求企业融入数据合规管控。在该阶段,企业需要制定如数据采集策略、存储策略、标注规范、数据集质量标准等文件,确保后续工作有序开展。设计管理的价值在于将治理成本从后期修正转移到前期设计,虽然增加了前期投入,但可以显著降低后期修正的成本和风险。数据采集策略需要明确数据来源、采集方式、采集频率、数据格式等关键要素;存储策略需要考虑数据的访问模式、保留期限、备份策略、成本优化等因素;标注规范需要定义标注的对象、方法、质量标准、验收流程等内容;数据集质量标准则需要明确质量评估的维度、指标、阈值和检测方法。

数据加工是供给 AI 所需数据的核心环节,涵盖数

据清洗、标注、增强、缩减及合成。该过程需依据前期制定的标准操作规范,利用智能化工具对原始数据进行深度处理。数据清洗环节采用基于规则与机器学习相结合的方法去除噪声内容,包括去除超文本标记语言(Hyper Text Markup Language, HTML)标签、广告文本、个人隐私信息等,并通过文本质量评分模型(如基于困惑度、语言模型打分)过滤低质量语料。数据标注环节引入预标注模型与主动学习机制,利用大语言模型生成初始标注结果并经人工审核修正,形成机器预标与人工精标协同的混合增强智能模式,降低人工标注成本的同时提升标注一致性与准确性^[4]。数据增强环节综合运用回译、同义替换等传统方法以及基于生成对抗网络和扩散模型的合成技术,扩充数据规模与多样性,有效补充真实世界难以采集的长尾样本与极端场景数据。数据合成技术可补充真实世界难以获取的极端数据与隐私数据,增强模型输出结果的可靠性,提升模型泛化能力。

2.2 交付阶段

交付阶段的目标是将研发好的数据安全、合规、高质量地供给下游应用。如图 4 所示,交付阶段主要包含测试管理及数据集发布两部分。交付阶段是连接数据研发与数据消费的桥梁,其质量直接影响下游应用的效果和数据价值的实现。

测试管理过程中需依据前期设定的标准规范,对

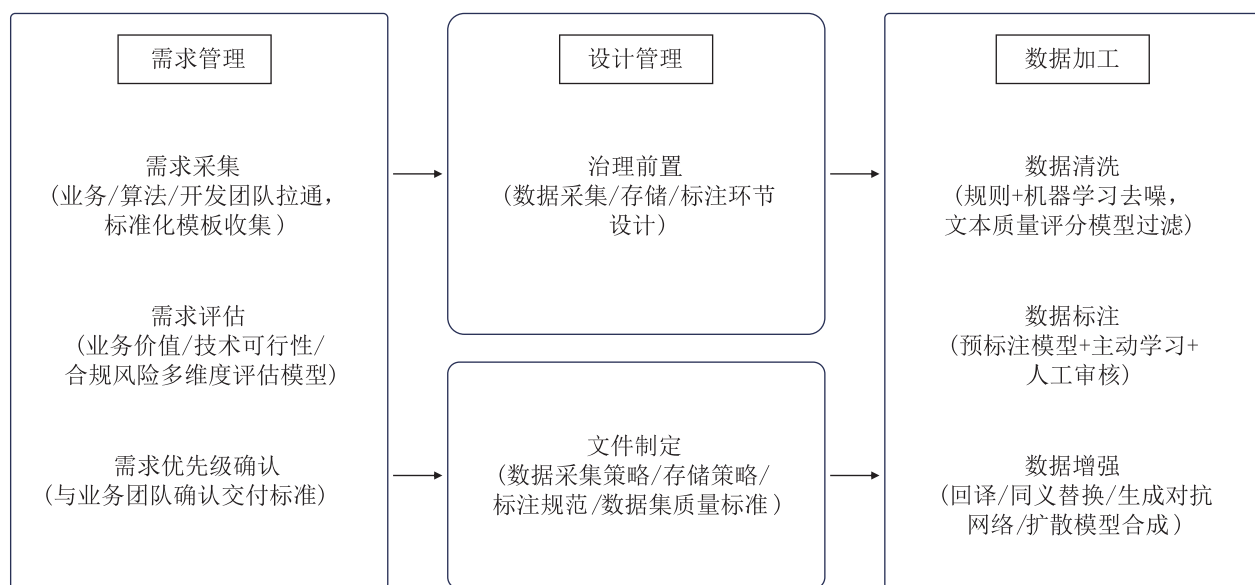


图 3 AI-ready DataOps 研发阶段能力说明

数据的质量、安全、合规方面进行测试。质量测试包括检测数据是否准确、完整,评估数据分布的多样性和代表性,验证标注结果的一致性和准确性^[5]。安全测试评估数据访问控制是否有效实施,验证敏感数据是否经过适当的脱敏处理,检查数据传输和存储的加密机制。合规测试检验数据内容是否符合伦理要求,验证数据来源的合法性,确保数据集符合相关法律法规和行业标准。测试管理需要建立自动化测试框架,将质量检测、安全扫描、合规审查等环节纳入自动化流水线,实现快速、全面的测试覆盖。同时,还需要建立测试用例管理体系,确保测试的全面性和可追溯性。

数据集发布阶段,数据通过测试后,企业需要将数据组合成 AI 训练所需的数据集,并按照既定流程将数据集发布至指定平台或存储位置,供后续多场景的应用。在此过程中,企业需对数据集进行版本管理,明确版本号、更新内容、更新时间等信息,为后续追溯与管理提供支持,并且还需要建立版本命名规范、版本变更记录、版本兼容性管理等机制,确保数据集的可追溯性和可复现性。数据集发布还应包含完整的元数据描

述,包括数据集的来源、规模、分布、质量指标、使用协议等信息,便于下游用户理解和使用数据集。发布流程需要定义发布审批、发布验证、发布通知等环节,确保发布的规范性和可控性。

2.3 运维阶段

数据运维阶段是以保障数据高可用、流程高效稳定为核心目标,构建持续监控与动态管理的运维体系。如图 5 所示,运维管理主要包括监控管理和资源管理两部分。运维阶段是确保数据工程体系框架持续稳定运行的关键,其核心在于建立全面的监控体系和高效的资源管理机制。

监控管理需要设置数据集监控指标,借助自动化监控工具对数据进行全生命周期的实时监控。指标覆盖数据集质量、安全、性能、成本等多维度,包括错误率、访问频率、存储成本、数据新鲜度等关键指标。一旦数据指标偏离正常范围立即触发告警机制,通过系统消息通知、短信通知等方式及时告知相关人员以便迅速处置。监控管理还应建立数据质量趋势分析能力,通过历史数据发现潜在问题,实现从被动响应到主动预防的转变。具体而言,监控指标体系应包括数据

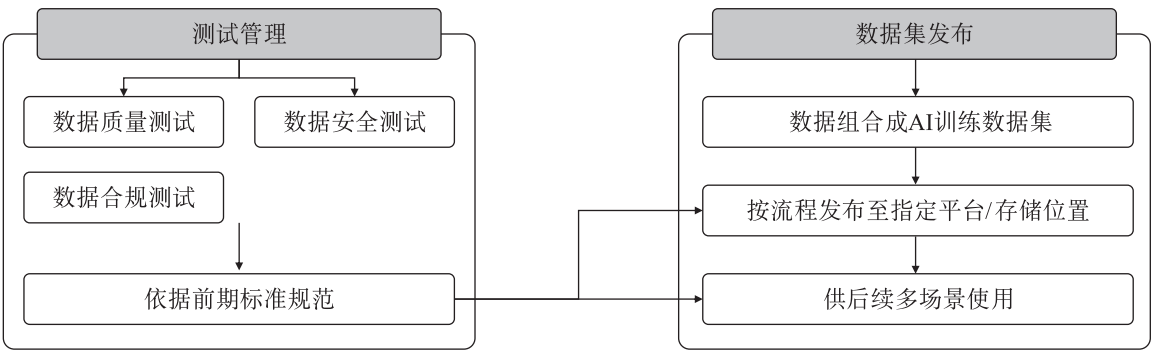


图 4 AI-ready DataOps 交付阶段能力说明

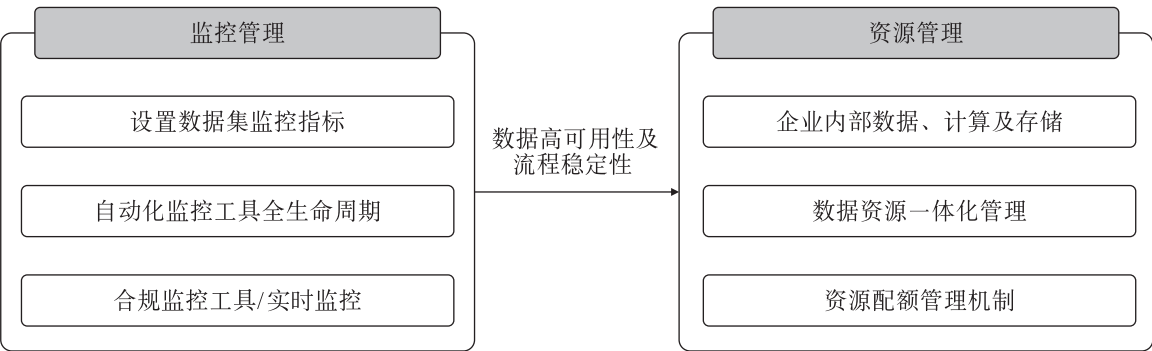


图 5 AI-ready DataOps 运维阶段能力说明

质量指标(如完整性、准确性、一致性)、安全指标(如访问异常、权限变更)、性能指标(如响应时间、吞吐量)、成本指标(如存储成本、计算成本)等多个维度。

资源管理对于企业内部的数据、计算及存储资源进行统一化管理。数据资源方面,维护数据的及时性及有效性,定期清理过期数据,确保数据资源的可用性。计算资源方面需设置弹性调度规则,根据数据流量、任务负载、任务优先级等情况实现自动扩容缩容的动态调整,优化计算资源利用率。存储管理方面,结合数据的使用频率、重要性等合理规划存储资源,通过分层存储策略将热数据、温数据、冷数据分别存储在不同成本的存储介质上,优化存储成本。资源管理还需要建立资源配额管理机制,为不同项目、团队分配合理的资源配额,避免资源争抢和浪费,同时支持资源的动态调整和回收。

2.4 运营阶段

价值运营阶段的目标是充分发挥数据价值、高效利用组织资源,并持续优化数据服务。如图 6 所示,价值运营主要包括数据集管理、成本管理及持续优化三部分。运营阶段是数据工程体系框架的价值实现环节,其核心在于将数据资源转化为业务价值,并通过持续优化提升运营效率。

数据集管理需要对企业内部的多个数据集进行统筹管理。对于数据集的类型、分布等情况进行统计,掌握企业数据集资源。同时,需要依据数据集权限管理规范、数据集安全处置规范、数据集存储策略等制度对数据集进行定期维护。此外,还需通过用户反馈、量化分析等方式对数据集的效用进行评价,保障企业数据集持续发挥价值。数据集管理还应建立数据资产目录,实现数据集的可发现性和可共享性,促进数据在企业内部的流通和价值释放^[6]。数据资产目录需要包含数据集的基本信息、质量评分、使用统计、用户评价等

内容,支持多维度的检索和筛选。

成本管理针对数据开发、运维、管理全流程,从人力、物力两方面进行成本梳理。首先梳理数据处理流程,评估流程设计合理性,消除冗余环节及重复操作。其次,在人力方面,评估人员分配合理性,避免人力冗余或不足;在物力方面,核算硬件设施、软件工具等物力成本,优化资源采购及使用策略。通过全方位的成本管理有效降低成本,实现资源的最优配置。成本管理还需要建立成本分摊机制,将数据工程的成本合理分摊到各个业务单元,促进成本意识的提升和资源的合理使用。同时,建立成本效益分析机制,评估各项投入的产出效果,指导资源的优化配置。

持续优化需要延续 DataOps 理念,结合组织战略方向、业务发展需求及技术发展趋势等,持续优化数据治理体系。内部建立用户反馈机制,持续收集用户建议并分析应用。从敏捷、精益的角度不断优化流程链路,并通过引入新的技术工具提升智能化数据治理水平。持续优化还应关注行业最佳实践和新兴技术发展,及时将先进的理念和技术引入数据工程体系框架,保持体系的先进性和竞争力。持续优化需要建立优化需求收集、评估、实施、验证的闭环机制,确保优化工作的系统性和有效性。同时,还需要建立关键绩效指标体系,量化评估优化效果,为持续改进提供数据支撑。

3 AI-ready DataOps 实践路径

基于 AI-ready DataOps 的目标,本文提出高质量数据集建设的五步法实践路径。企业从总体规划层面和具体数据集建设方面可按照“盘、建、研、管、运”(即数据管理应用现状盘点、规范体系与技术体系建设、数据集研发、数据集管理、全链路运营)五步进行实施。五步法路径具有很强的实践指导价值,企业可以根据自身情况选择合适的切入点,逐步推进数据工程体系

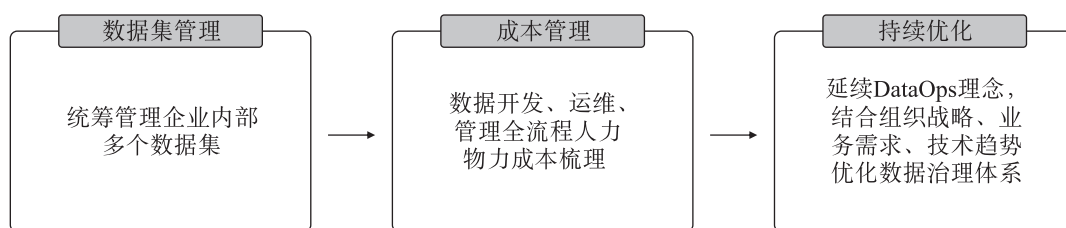


图 6 AI-ready DataOps 运营阶段能力说明

的建设和升级。

3.1 数据管理与应用现状盘点

数据管理与应用现状盘点的目标是理清内外部数据资源的来源、规模、质量、类别情况,明确应用与采集的格式、质量、内容要求,开展全域数据(结构化、非结构化)资源、管理应用需求的盘点工作。如图 7 所示,盘点是数据工程建设的起点,其全面性和准确性直接影响后续工作的效果。

盘点内部数据资源需要针对结构化数据、非结构化数据、半结构化数据、时序数据等类型的数据开展内部数据资源的盘点工作,全面掌握数据的分布、格式、标注、质量等情况。盘点外部数据资源需要针对外部采购、订阅、采集等类型的数据开展外部数据资源的盘点工作,全面掌握外部数据的来源、质量等信息。盘点业务智能化场景需要针对已经建设或需要建设的业务智能化场景进行盘点,掌握当前及未来业务数智化转型场景全景图,明确智能化手段赋能业务的价值链条及建设优先级。盘点数据采集优化需求需要基于内外部数据资源情况以及业务智能化升级场景,梳理数据采集优化需求,明确其在获取阶段的数据可用性、质量水平、安全保障及标准合规等信息。

该阶段的关键输出包括数据资源地图和业务场景地图。数据资源地图基于内外部数据资源盘点底稿进行分析,形成全域数据资源地图,掌握内外部结构化、非结构化和半结构化数据的分布、归属、权限和描述等关键信息,便于数据的使用、交换、管理和运营。业务

场景地图基于应用场景需求盘点底稿进行分析,形成业务数智场景地图,掌握当前及未来业务数智升级的价值、优先级等信息。此外,该阶段需形成统一的数据需求模板,供后续业务需求提出时复用,能够清晰地提出数据采集交换、质量等要求,便于数据加工供给方高效、高质量地交付数据。

3.2 规范体系与技术体系建设

规范体系与技术体系建设阶段遵从“先设计、再开发”的理念,提前制定数据集的质量、安全、运营等要求,标准化工作流程,定制化工具平台。开展制度规范、工艺流程、功能平台的建设与优化。这一阶段是数据工程体系框架的制度基础,其完善程度直接影响数据工程体系框架的规范性和可持续性。

该阶段的主要活动包括明确标注目标、建设数据集评价指标、对标技术功能满足度、构建工艺流程。明确标注目标需要针对数据集应用场景需求,在开展数据标注工作前明确数据标注的对象、规模、应用场景等信息^[7]。建设数据集评价指标需要基于标准指引、行业经验以及业务需要,明确数据集质量、安全、标准等评价指标。上述指标涵盖数据质量(准确性、完整性、一致性、时效性)、数据安全(隐私合规、脱敏程度、访问权限)及标准规范(规范性、元数据完整性、标注一致性)等维度,在借鉴标准与外部最佳实践的基础上,充分考虑业务场景需求,保证评价的可执行性。对标技术功能满足度需要针对数据集研发、运营和管理等工作所需要的平台与功能进行分析,明确当前技术平

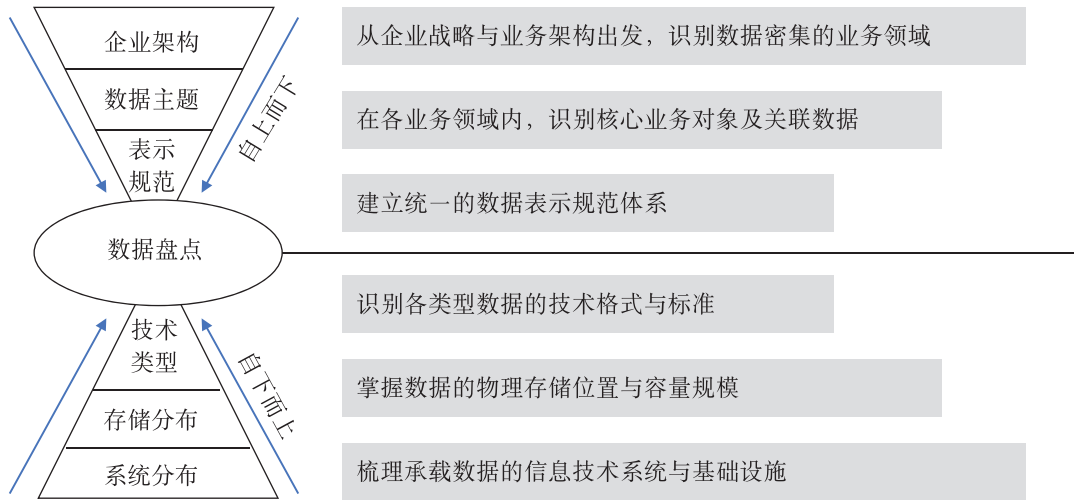


图 7 数据盘点工作思路

台与预期功能的差异。构建工艺流程需要针对数据集研发、运营、管理等工作涉及的团队组织、制度规范、实施环节进行流程构建,保障各环节的紧密配合、高效协同^[8]。

该阶段的关键输出包括数据集评价体系、数据集研发管理运营工序、数据标注及数据集管理运营平台。数据集评价体系通过量化方式评价数据集建设成果,客观评价数据集建设的质量、安全等^[9],还应形成完整、可行、标准化的高质量数据集研发、管理、运营工序,保障研发与治理要求的协调性和融合性,保障每个环节的标准化、高质量与可信^[10]。数据标注及数据集管理运营平台提升高质量数据集研发、管理、运营的线上化与一体化水平。此外,该阶段还需要形成数据集质量、安全、合规、存储需求采集表,数据标注质量需求采集表,以及平台功能需求采集表等输入文档,为后续工作提供指导。

3.3 数据集研发

数据集研发阶段依托前两步工作的关键输出,如数据资源地图、需求收集模板、数据集评价体系等开展标准化的数据集研发工作,产出高质量数据集。数据集研发是数据工程体系框架的核心环节,其产出质量直接决定了数据价值的实现程度。

数据预处理是基于高质量要求与数据需求,依托技术平台开展数据预处理工作,包括数据的清洗、标注等工作。清洗与标注后的数据满足该项工艺的质量要求,满足数据需求方的标注规模等要求。数据集的增强、缩减、合成是基于数据需求和场景建设需求,按需开展数据集的增强、缩减与合成工作,以满足业务在场景智能化升级阶段的数据需求。数据增强技术可以扩充数据规模与多样性,提高模型泛化能力;数据缩减技术可以去除冗余数据,提升训练效率;数据合成技术可以补充真实世界难以获取的极端数据、隐私数据等,增强模型输出结果的可靠性。

数据集评价是在完成数据预处理、加工等阶段后,基于数据集的评价体系开展数据集质量、安全、合规等方面的评价,基于数据集在大模型上的应用表现开展效果评价,确保数据集能够满足数据集需求方对数据集的高质量要求^[11]。该阶段的关键输出是高质量数

据集,基于数据需求,按照研发工序对原始数据资源进行清洗、标注、增强、缩减、合成,过程经过质量、安全、合规等要求的检查,最后通过数据集的评价体系检查后交付输出高质量数据集。数据集评价需要建立多维度的评估体系,包括技术质量维度(准确性、完整性、一致性)、AI 特化维度(多样性、代表性、公平性)、安全合规维度(隐私保护、伦理审查)等^[12]。

3.4 数据集管理

数据集管理阶段依托前序工作成果,上架数据集资产,洞察分析数据集规模和数据集分布等构成,维护数据集版本、权限,依托数据集管理平台开展数据集的统一管理。数据集管理是实现数据资产化和价值化的关键环节,其核心在于建立统一的数据资产管理能力和完善的数据治理机制。

洞察数据集构成需要基于公司已建设的数据集,结合日常应用情况,深入分析数据集的构成元素、数据类型结构以及业务域分布情况。通过量化分析,识别数据集中各类数据的占比、关联性和重要性,以及在不同业务域中的分布特点。针对分析结果,提出优化数据集构成、调整数据分布、共享不同业务间数据集等策略,提升数据集的全面性,在内部形成数据集资源流通共享^[13]。数据集洞察需要建立多维度的分析视角,包括数据类型分布、业务域覆盖、质量评分分布、使用频率统计等,为数据资产的优化配置提供决策支持。

数据集持续测试需要基于数据集评价体系,针对数据集的质量、安全、合规等要求进行持续的检验检测,随着数据集评价体系的不断完善,保障研发后数据集的保鲜。维护数据集版本需要随着数据集的更新对其进行维护,掌握数据集的历史版本,版本差异,当前版本数据集的描述、使用、归属、权限等信息。上架数据集资产需要将业务场景应用效果好、应用场景广泛、具有高价值的数据集,作为资产进行上架,便于实现充分的内部开放共享以及在数据要素市场、可信数据空间中的交易流通。数据集资产管理还需要建立资产价值评估机制,从使用频次、应用效果、成本投入等维度评估数据资产的价值,为数据资产的交易和流通提供定价参考。

3.5 全链路持续运营

全链路持续运营阶段依托前序工作成果,持续不

断跟踪、维护及优化数据集的评价、使用效果、建设与维护成本和高质量数据集价值。全链路持续运营是数据工程体系框架的持续改进环节,其核心在于建立持续优化的机制,确保数据工程体系框架的持续演进和价值最大化。

维护各评价指标需要基于高质量数据集的不断学习、应用、治理,维护数据集评价指标体系、数据集管理评价体系等评价指标,确保高质量数据集的建设、应用、维护等工作不断优化。跟踪数据、存储、计算资源需要随着高质量数据集的大规模建设和利用,不断跟踪、维护数据资源的建设情况,按需丰富、整合相关数据资源,控制数据资源成本^[14]。统筹数据集存储资源,减少冗余存储、超期存储,优化数据存储成本。统筹数据集计算资源,按需配置计算资源,保障计算资源的最优化配置。资源跟踪需要建立资源使用监控体系,实时掌握各类资源的使用情况,及时发现资源“瓶颈”和浪费,为资源优化提供数据支撑。

关注应用成本与价值需要精益化管理数据集建设、维护等环节的人力与设备成本,建立“数据集-智能化应用-业务价值提升”的价值链路评价体系,持续关注、优化高质量数据集的投入产出比。持续优化需要对研发工艺、数据资源配置、工具平台建设、业务场景应用、评价体系等进行优化,实现高质量数据集赋能业务、服务产业的价值最大化^[15]。该阶段的关键输出包括数据集运营模块、数据集运营大屏、数据集价值度量模块,形成成熟的数据集运营能力。数据集运营大屏需要展示高质量数据集规模、分布、评价等核心信息,促进高质量数据集的应用、评价与优化。数据集价值度量模块需要建立涵盖使用频次、应用效果、成本投入等多维度的价值评估模型,量化数据集的业务贡献与资产价值。

4 结束语

本文通过系统分析 AI 时代数据工程体系框架的新特征与新要求,构建了面向 AI 的数据工程体系框架,阐明了其与高质量数据集建设的内在机理。研究首先厘清了 AI-ready DataOps 的内涵,指出其是通过融合 DataOps 理念,面向 AI 场景高效、高质量、安全供

给数据的一种工程化方法论。这一定位突破了传统数据工程将数据视为静态资源的管理范式,将数据工程的价值创造从支撑业务决策转向驱动智能计算,为数据工程领域的发展指明了方向。

在此基础上,本文提出了 AI-ready DataOps 能力框架,涵盖研发、交付、运维、运营四大阶段,形成步骤清晰、分工明确的数据集流水线。研发阶段构建数据研发治理一体化流程,涵盖需求管理、设计管理和数据加工 3 个核心环节,确保数据研发工作的规范性和高质量。交付阶段实现数据的安全合规发布,通过测试管理和数据集发布确保数据集的可用性和可追溯性。运维阶段保障数据的高可用性,通过监控管理和资源管理确保数据工程体系框架的稳定运行。运营阶段最大化发挥数据价值,通过数据集管理、成本管理和持续优化实现数据价值的持续释放。这一框架设计充分考虑了 AI 场景的特殊需求,将 DataOps 的核心理念与 AI 数据生产的技术特性相结合,为企业构建面向智能的数据工程体系提供了可操作的实施框架。

研究进一步阐述了“盘、建、研、管、运”五步法实践路径,为企业高质量数据集建设提供可操作的实施指南。数据管理与应用现状盘点理清数据资源现状,形成数据资源地图和业务场景地图,为后续工作奠定基础。规范体系与技术体系建设奠定制度基础,形成数据集评价体系、研发管理运营工序和技术平台。数据集研发产出高质量数据集,通过数据预处理、增强合成和评价验证确保数据集质量。数据集管理实现数据资产的统一管理,通过洞察分析、持续测试和版本管理保障数据资产价值。全链路持续运营保障数据价值的持续释放,通过指标维护、资源跟踪和持续优化实现价值最大化。五步法路径具有很强的实践指导价值,企业可以根据自身情况选择合适的切入点,逐步推进数据工程体系的建设和升级。

理论层面,本文拓展了数据工程的研究边界,将 DataOps 的适用范围从传统 BI 场景延伸至 AI 训练场景,提出了涵盖基础语料、知识数据与指令数据 3 类形态的数据集工程化生产机制。这一理论贡献填补了数据工程领域在 AI 数据生产方面的研究空白,为后续研究提供了理论基础和分析框架。实践层面,本文为国

家高质量数据集建设提供了方法论指引,提出的能力参考框架和实践路径可支撑企业评估现状、规划路径并实施落地。特别是关于治理前置、人机协同标注、主客观融合评价等策略,为破解当前数据集建设中的质量管控与规模效率矛盾提供了可行方案。

本文侧重于理论框架构建与宏观路径设计,对特定行业(如医疗、金融)的特殊化需求分析有待深化;提出的能力体系尚需更多实证检验以验证其普适性与有效性。未来研究可进一步关注 3 个方向:一是探索结构化与非结构化数据、AI 与 BI 场景深度融合的全域数据工程体系框架,打破“数据孤岛”,实现数据价值的最大化释放;二是深化 DataSecOps 研究,实现数据安全性与运营的深度协同,在保障安全的前提下提升运营效率;三是追踪 AI-Native DataOps 演进,研究 AI 驱动的自治化数据管理模式,探索数据工程的未来发展方向^[16]。

参考文献

- [1] LIEBMANN L. Three reasons why DataOps is essential for big data success[R], 2014.
- [2] 大数据技术标准推进委员会. DataOps 实践指南(3.0)[R], 2025.
- [3] 中国信息通信研究院,清华大学计算社会科学与国家治理实验室,中国人工智能产业发展联盟. 人工智能高质量数据集建设指南[R], 2025.
- [4] 蔡莉,王淑婷,刘俊晖,等. 数据标注研究综述[J]. 软件学报, 2020,31(2):302-320.
- [5] 韩京宇,徐立瑞,董逸生. 数据质量研究综述[J]. 计算机科学, 2008,35(2):1-5.
- [6] 江涛. 数据资产的交易效应、服务收益与交易生态[J]. 学术月刊, 2024(1):70-80.
- [7] 燕江依,李荪,樊威,等. 新一代数据标注产业对“人工智能+”范式创新的作用机理与实践路径研究[J]. 信息通信技术与政策, 2025, 51(8): 26-34.
- [8] 王泽宇,闫树,王丹阳. 数据资产化基础概念研究[J]. 信息通信技术与政策, 2025,51(2):53-59.
- [9] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810.04805, 2019.

- [10] HUMBLE J, FARLEY D. Continuous delivery: reliable software releases through build, test, and deployment automation[M]. Boston: Addison-Wesley Professional, 2010.
- [11] 蔡莉,朱扬勇. 从数据质量到数据产品质量[J]. 大数据, 2022,8(3):26-39.
- [12] BASS L, WEBER I, ZHU L. DevOps: a software architect's perspective[M]. Boston: Addison-Wesley Professional, 2015.
- [13] ARMBRUST M, DAS T, SUN L, et al. Delta lake: high-performance ACID table storage over cloud object stores[J]. Proceedings of the VLDB Endowment, 2020, 13(12):3411-3424.
- [14] 中国信息通信研究院,中电信人工智能科技(北京)有限公司. 数据标注产业发展研究报告[R], 2025.
- [15] BAROCAS S, HARDT M, NARAYANAN A. Fairness and machine learning: limitations and opportunities[M]. Cambridge: MIT Press, 2023.
- [16] AMERSHI S, BEGEL A, BIRD C, et al. Software engineering for machine learning: a case study[J]. IEEE Software, 2019, 36(4): 81-89.

作者简介:

- 姜春宇** 中国信息通信研究院人工智能研究所高级工程师,主要从事大数据技术、数据库、数据要素流通、数据治理等方面的研究工作
- 艾博焕** 中国国际技术智力合作集团有限公司工程师,主要从事数据运营、多模态数据管理等方面的研究工作
- 郭彦美** 中国信息通信研究院人工智能研究所工程师,主要从事 DataOps、数据要素、非结构化数据治理等方面的研究工作
- 崔一妍** 中国信息通信研究院人工智能研究所工程师,主要从事 DataOps、数据编织、主动元数据管理等方面的研究工作
- 尹正** 通信作者。中国信息通信研究院人工智能研究所工程师,主要从事 DataOps、数据治理、数据编织、多模态数据管理等方面的研究工作

Construction of data engineering architecture for AI: a paradigm shift based on AI-ready DataOps

JIANG Chunyu¹, AI Bohuan², GUO Yanmei¹, CUI Yiyan¹, YIN Zheng¹

(1. Artificial Intelligence Research Institute, China Academy of Information and Communications Technology,
Beijing 100191, China;

2. China International Intellectech Group Co.,Ltd., Beijing 100053, China)

Abstract: The rapid advancement of Artificial Intelligence (AI) technology has introduced novel requirements for data engineering, catalyzing a paradigm shift in data value morphology from data resources to high-quality datasets. This paper systematically analyzes the underlying mechanisms driving the transformation from traditional DataOps to AI-ready data engineering paradigms, and constructs an intelligence-oriented data engineering architecture. The study first delineates the connotation of AI-ready DataOps, conceptualizing it as an engineering methodology that integrates DataOps principles to efficiently, securely, and reliably supply data for AI scenarios. Building upon this foundation, an AI-ready DataOps capability reference framework is proposed, encompassing four critical phases: research and development, delivery, technical operations, and data governance, thereby establishing a clearly defined, step-by-step dataset pipeline. Furthermore, the study elaborates a five-step practical implementation methodology—inventory, construction, processing, governance, and operations—offering actionable guidelines for enterprise-level high-quality dataset construction. This study provides theoretical frameworks and methodological guidance for national high-quality dataset infrastructure development.

Keywords: data engineering; DataOps; high-quality dataset; AI; AI-ready DataOps

(收稿日期:2026-03-31)