江苏省制造业领域面向人工智能的数据治理工作参考指引

（2026 年版）

江苏省工业和信息化厅

二〇二六年二月

前　言

习近平总书记提出，“数据基础制度建设事关国家发展和安全大局”“加快构建数据基础制度体系”。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》。

近年来，江苏认真贯彻党中央和国务院决策部署，深入开展制造业“智改数转网联”行动，率先出台《关于加快推进人工智能赋能新型工业化的行动方案（2024-2027）》，大力推动人工智能技术与制造业深度融合应用。

江苏“智改数转网联”与人工智能赋能新型工业化的发展实践深刻揭示：当模型能力日趋接近、供应日益充足，制造企业在拥抱“人工智能+”时仍感步履维艰，制约“人工智能+制造”价值释放的，不再是“有没有一个好模型”，而是“有没有能喂给模型、与业务深度耦合的高质量数据”。而当前制造业领域数据“孤岛”与“失真”、数据治理与标准化缺失、数据与应用场景脱节等痛点问题严重制约了高质量、场景化数据集的供给。

本指引基于江苏制造业数据治理现状，结合人工智能应用典型场景，提出构建面向 AI 的数据治理体系，引导企业开展数据资产的系统性治理与建设，提升将物理世界的复杂系统转化为数字世界可计算、可优化、可创新的高质量数据资产能力。

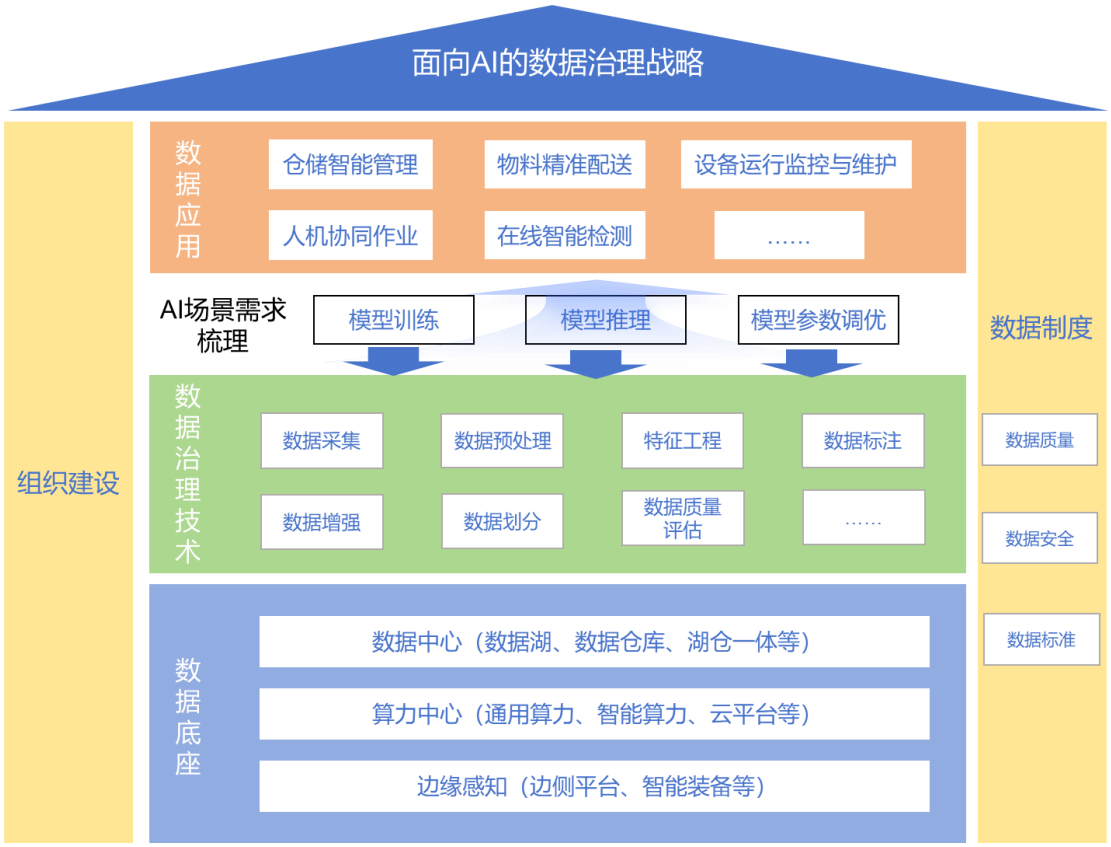

图 1　面向 AI 的数据治理体系架构

目　录

一、数据治理与人工智能应用的关系

数据治理与人工智能应用呈现“相辅相成、互为支撑”的协同关系。人工智能应用的核心是依托高质量数据完成模型训练、推理与迭代，数据治理则是保障数据质量的核心抓手，人工智能应用的深化，反向推动数据治理工作从“被动合规”向“主动价值驱动”升级。因此，企业在部署人工智能应用前，应以具体业务场景为导向，开展针对性数据治理工作，构建高质量数据集，为业务领域人工智能模型的训练与优化提供坚实支撑，切实提升模型性能与精准度。企业依托优化后模型开展业务价值研判，明确新增数据需求及现存问题，进而以内生驱动方式，推动企业持续深化数据治理优化工作，实现数据治理与人工智能应用的“协同进化”。

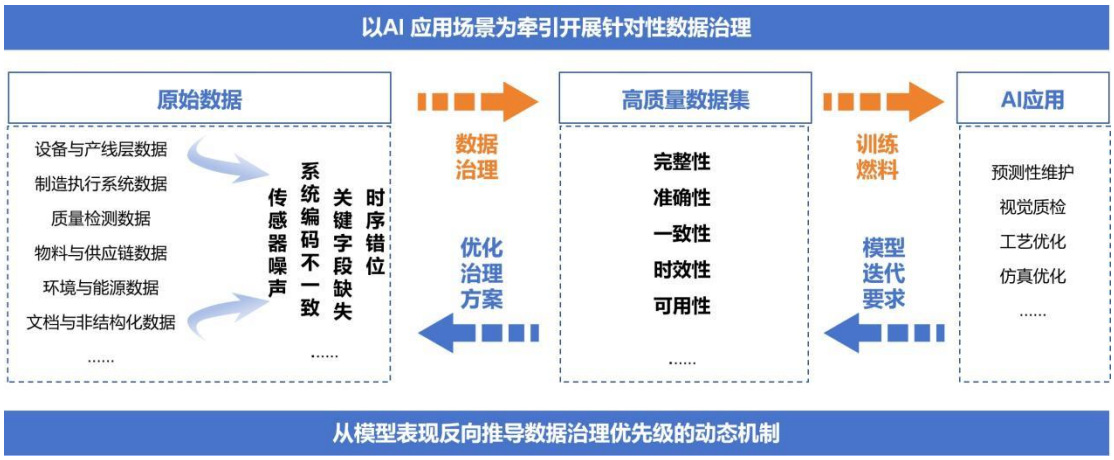

图 2　数据治理与人工智能应用的关系

二、面向 AI 的数据治理技术应用

制造业领域人工智能应用，核心依赖于多模态数据[1]、标注数据及合成数据等。这一特定数据需求，正推动数据治理技术从传统结构化数据治理向多模态数据治理方向加速迭代，同时对治理效率与质量提出更高要求。因此，数据采集、预处理、特征工程、数据标注、数据划分、数据增强六大环节成为面向人工智能场景数据治理的核心环节，对能否产出适配 AI 模型训练与应用的优质数据起到关键作用。数据存储、数据计算、数据集成、数据质量评估、数据安全保护五大环节，虽不属于面向人工智能的数据治理流程本身需执行的核心操作，更多承担着数据治理前的基础准备或全流程保障职能，但为确保数据全生命周期治理工作能高度支撑人工智能应用需求，这些环节同样是面向人工智能的数据治理工作中必须统筹考量的重要内容，需与核心治理环节协同升级、同步优化。制造业企业可结合自身技术基础、资源条件及实际业务痛点，针对性选取适配的环节落地数据治理技术，最大化挖掘数据价值，为人工智能技术在制造业的深度应用筑牢数据根基。

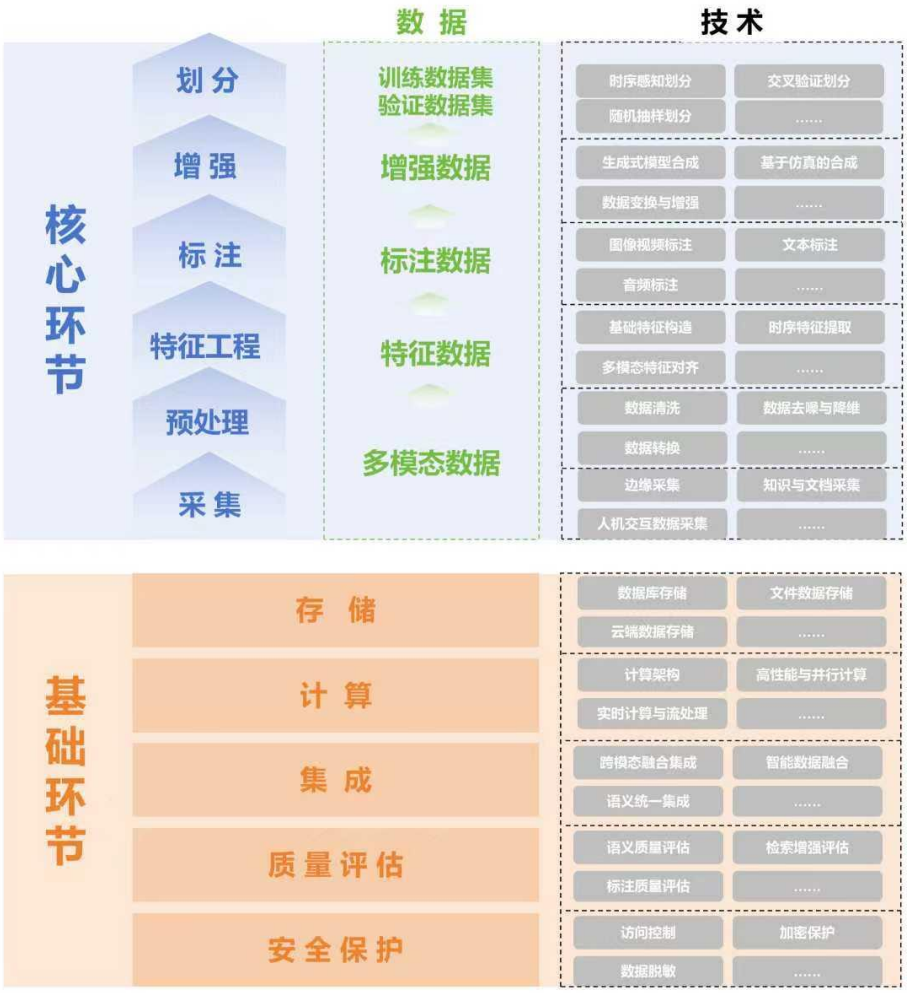

图 3　面向 AI 的数据治理技术应用框架

（一）数据采集

数据采集是为后续人工智能应用环节提供数据输入来源的环节，在确保数据时效性、完整性与真实性的前提下，为模型训练与推理构建高质量、可持续、场景化的数据供给渠道，并实现智能决策的闭环反馈。

1.解决的问题

数据“采不到”。因设备未进行或无法进行数字化改造，缺乏标准通信接口，或复杂环境下导致传感器部署难、信号

传输不稳定等问题，导致关键数据无法有效采集。**数据“采不准”**。由于传感器精度不足、系统逻辑错误、测量偏差、人为录入错误等问题，导致采集的数据与实际生产情况存在偏差，数据准确性、一致性不达标，无法满足 AI 使用需求。**数据“采不全”**。数据源虽然可以采集，但采集的数据维度缺失、粒度太粗、频率过低或关联信息缺失，导致数据无法支撑深度分析。**数据“格式乱”**。因缺乏统一的数据标准和治理规范，导致数据在命名、编码、单位、时间戳等方面不一致、不标准。**数据“分布散”。**数据分散存储在不同地域、部门、系统中，形成“数据孤岛”，无法实现全局共享与统一管理。**数据“溯源难”**。因未进行全生命周期的追踪与审计，无法明确数据来源、采集过程、流转路径与修改记录，导致数据可信度不足。

2.核心技术应用

表 1　数据采集核心技术应用

手段	技术名称	功能	解决的问题
工业通信	现场总线技术 工业以太网技术 工业光纤网络 窄带物联网 5G/5G-A	工业控制系统、设备层核心运行数据的接入及跨系统互联互通	数据“采不到” 数据“格式乱”
工业感知	工业传感器技术 工业视觉成像技术 工业声学传感技术 音频降噪技术 激光雷达扫描技术 智能仪器仪表技术	将物理世界的温度、压力、图像、形貌、声音、化学成分等信号，转化为可被采集的电信号或数字信号	数据“采不到” 数据“采不全”
信号采集与调理	信号调理技术 模数转换技术	对传感器输出的原始微弱信号进行放	数据“采不准” 数据“采不全”

手段	技术名称	功能	解决的问题
	抗干扰与屏蔽技术	大、滤波、隔离、线性化处理，并转换为高精度、高稳定性的数字信号，确保基础数据质量	
物体标识与解析	RFID 射频识别 条形码二维码 NFC 近场通信	产品、零部件全生命周期身份标识数据的采集，满足流转环节数据采集需求	数据“溯源难”
边缘采集	边缘智能网关 边缘计算框架	在边缘侧就近完成数据采集、协议解析、滤波清洗、格式标准化、本地存储与加密，并可按规则向云端同步	数据“格式乱” 数据“分布散”
移动式采集	PDA 采集 移动工控机采集 无人机巡检采集	跨设备、跨系统、跨厂区抽取数据，统一异构数据格式，剔除格式混乱引发的冗余数据	数据“采不全” 数据“分布散”
事件驱动采样	阈值触发采集技术	根据预设的规则（如数值超限、设备状态变化）或外部事件，自动启动高精度、高频率的数据采集，用于捕捉瞬态过程或异常	数据“采不准” 数据“溯源难”
知识与文档采集	OCR 与文档解析 多模态数据抽取	将操作手册、维修日志、工艺单、CAD 图纸等非结构化文档转化为机器可读的文本或向量数据	数据“格式乱” 数据“采不全”
人机交互数据采集	ASR 语音转写 专家行为捕捉 操作日志审计	采集班组长/专家的语音指令、操作习惯、维修记录文本，捕捉人的经验	数据“采不到” 数据“采不全”

3.配套工具清单

表 2　数据采集配套工具清单

工具类型	功能模块	解决问题	对应技术
工业传感器/智能仪表	传感元件 模数转换（ADC） 现场总线/以太网通信接口	数据“采不到” 数据“采不准”	工业传感器技术 智能仪器仪表技术 模数转换技术 信号调理技术 现场总线技术 工业以太网技术
数据采集（DAQ）卡/模块	高精度 ADC 信号调理电路 数字 I/O 同步时钟		模数转换 信号调理 抗干扰技术
声音/振动分析仪	高保真麦克风/加速度计 高动态范围 ADC DSP 芯片		工业声学传感技术 音频降噪技术
RTU/PLC/DCS	I/O 模块（AI/DI） 高速计数模块 专用通信处理器		现场总线 工业以太网 模数转换 信号调理 抗干扰技术
IPC（工业 PC）	插槽式 DAQ 卡/通信卡 设备驱动 数据采集服务软件		模数转换 现场总线 边缘计算框架（作为载体）
嵌入式系统	MCU 内置 ADC/外置传感电路 专用通信芯片		嵌入式系统 模数转换 信号调理
机器人/数控机床/专用智能装备	控制器 I/O 伺服/驱动单元反馈接口 专用数据端口	数据“采不到” 数据“采不准” 数据“格式乱”	工业以太网 现场总线 智能仪器仪表技术
条形码/二维码读取设备	图像传感器 解码芯片/软件 通信接口	数据“溯源难”	工业视觉成像 条形码二维码技术

工具类型	功能模块	解决问题	对应技术
RFID 读写设备	射频模块与天线 信号处理单元 通信接口		RFID 射频识别技术
工业网关	多协议驱动栈 数据映射与标准化引擎 边缘计算微服务	数据“采不到” 数据“格式乱” 数据“分布散”	边缘智能网关 5G/NB-IoT 边缘计算框架
SCADA 系统	通信驱动（IO Server）、实时数据库、HMI	数据“分布散”数据“采不全”数据“格式乱”	工业以太网 现场总线 边缘智能网关
移动数据终端（PDA/工控平板）	集成扫描头（条码/RFID）、无线通信模块（4G/5G/Wi-Fi）、定制采集 APP	数据“分布散” 数据“采不到” 数据“溯源难”	PDA 采集 5G/5G-A 条形码/RFID
无人机/巡检机器人	飞控/导航系统、任务载荷控制器（相机/传感器）、无线图传模块	数据“采不到” 数据“采不全”	无人机巡检采集 工业视觉 5G/5G-A 激光雷达
非结构化数据处理工具	OCR 识别引擎 NLP 文本解析器 语音转写工具	数据“格式乱” 数据“采不全”	OCR 与文档解析 多模态数据抽取

（二）数据预处理

数据预处理是在数据分析或建模前，对原始数据进行清洗、去噪、转换等操作，以去除数据中的噪声、异常值和重复信息，提升数据质量的环节，目的是在确保数据一致性、适配性和信息密度的前提下，将原始数据转化为适合模型训

练与推理的优化输入，以提升算法效果、加速收敛并增强模型鲁棒性。

1.解决的问题

数据“脏”。由于企业设备老化或校准失效、传感器故障、传输丢包、存储格式错误、数据冗余存储、人为操作失误等原因，可能造成数据本身存在错误、缺失、异常等质量问题，无法真实反映业务状态，更无法直接用于分析或建模。**数据“乱”**。由于企业未建立统一的数据格式标准、系统异构性强、元数据[2]管理机制缺失的原因，导致不同设备、系统及厂区采集的数据在单位、格式、时序基准、通信协议等方面标准不一致的情况，无法被分析工具或 AI 模型高效识别与解析。**数据“繁”**。因数据采集过程中缺乏针对性筛选与分类，导致原始采集的数据规模庞大、模态多元、维度交叉，直接处理大规模非结构化数据[3]会导致计算资源呈指数级消耗，且在训练或推理时，关键信息可能被稀释在大量无关文本中，导致处理速度慢、精度低、成本高。

2.核心技术应用

表 3　数据预处理核心技术应用

手段	技术名称	功能	解决的问题
数据清洗	缺失值处理技术	使用插值、填充或基于模型的预测方法修复数据中的缺失部分	数据“脏”
	噪声数据处理技术	通过滤波、平滑或小波变换等方法消除数据中的随机干扰信号	
	异常值处理技术	基于统计或机器学习算法识别并修正、剔除或	

手段	技术名称	功能	解决的问题
		标记数据中的离群点	
数据转换	数据规范化技术	将不同量纲和范围的数据缩放到统一标准区间	数据“乱”
	数据编码技术	将分类、文本等非数值数据转换为数值型特征	
	数据聚合技术	按时间窗口或业务逻辑对数据进行汇总，提升数据粒度一致性	
	数据泛化技术	使用概念分层将低层数据抽象为更高层次的类别表示	
数据去噪与降维	数据去噪技术	应用信号处理或深度学习模型分离并去除数据中的噪声成分	数据“脏”
	数据降维技术	通过主成分分析、线性判别分析等方法减少数据维度，消除冗余	数据“繁”
数据整合	数据抽取技术	从多源异构数据中提取关键字段与信息，实现初步内容对齐	数据“乱”
	数据格式统一化	将不同结构、协议的数据转换为统一格式与存储规范	
	数据冗余与相关性分析	识别并处理数据中的高度相关或重复信息字段	数据“繁”
长文档处理	文本分块技术	将操作手册按语义、段落或字符数切分成模型能消化的小片段	数据“繁”
分词处理	Tokenizer 编码技术	将人类文字转化为模型计算的 TokenID，并优化词表以覆盖行业术语	数据“乱”
格式解析	文档解析技术	解析 PDF、PPT、HTML 中的文本和表格，保持原有布局结构	
	工业图纸格式转换与解析	直接读取 CAD 文档的 DWG/DXF 图层、线条、坐标等，以及将	

手段	技术名称	功能	解决的问题
		PDF 文件转为 SVG 或矢量数据	
符号识别	图纸工程符号识别	定位工程符号（如液压阀、电机图标、粗糙度标注等）进行目标检测，以及识别特定的几何形状进行图元匹配	
预处理流程编排与管理	工作流编排引擎	可视化拖拽式流程配置，串联多源异构数据的全链路处理环节	数据“繁” 数据“乱”

3.配套工具清单

表 4　数据预处理配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
桌面数据处理工具	Microsoft Excel	数据“脏” 数据“乱”	基础清洗
	WPS 表格		
数据清洗工具	Great Expectations Cleanlab	数据“脏”	异常值处理技术 噪声数据处理技术 重复值处理技术
	Pandas Polars		缺失值处理技术 重复值处理技术
数据转换工具	Apache NiFi	数据“乱”	数据格式统一化 数据抽取
	dbt (Data Build Tool)		数据规范化 数据聚合 数据泛化
数据编码工具	category-encoders		数据编码技术
数据去噪工具	PyWavelets	数据“脏”	数据去噪技术
数据整合工具	Apache SeaTunnel	数据“乱” 数据“繁”	数据抽取 数据格式统一化 数据冗余与相关性分析

工具类型	开源框架/工具	解决的问题	对应技术
工作流编排工具	Apache Airflow Apache DolphinScheduler Kedro Prefect	数据“乱” 数据“繁”	预处理流程编排与管理
非结构化数据解析	Unstructured.io PyMuPDF	数据“乱”	文档解析技术
文本切分与编排	LangChain (Text Splitters) LlamaIndex	数据“繁”	文本分块技术
工业图纸格式转换与解析工具	ezdxf LibreDWG	数据“乱”	工业图纸格式转换与解析
图纸工程符号识别工具	YOLO Detectron2 OpenMMLab	数据“乱”	图纸工程符号识别

（三）数据特征工程

数据特征工程[4]是为后续人工智能应用提供高判别性、高可解释性输入特征的环节，目的是在确保特征稳定性、业务一致性和计算效率的前提下，通过对原始数据的深度加工与重构，最大限度地提取和构造能够表征问题本质的信息，以提升模型的准确性、泛化能力与落地可行性。

1.解决的问题

数据“价值浅”。数据的使用局限于基础统计、可视化展示等表层场景，未通过深度分析、AI 建模等方式挖掘其背后的业务关联（如工艺优化、成本降低、风险预警），导致数据资产无法转化为实际生产效益，AI 应用仅停留在“演示级”而非“生产级”。**数据“规律隐”。**制造业数据具有多维度、强耦合、非线性的特性（如设备运行数据与工艺参数、

环境因素、原材料质量深度关联），但因数据维度割裂、分析方法不足、业务理解不深等原因，隐藏在数据中的核心规律（如故障前兆、质量波动趋势、能耗优化空间）未被识别，导致 AI 模型无法学习到有效的决策逻辑。**数据“链路断”。**数据在“采集-传输-处理-存储-应用-反馈”的全生命周期中，或在“业务需求-数据采集-模型训练-业务执行-效果评估”的业务流程中，存在环节缺失、衔接不畅、反馈滞后等问题，导致数据无法形成闭环流转，无法支撑 AI 模型的持续优化和业务的动态调整。

2.核心技术应用

表 5　数据特征工程核心技术应用

手段	技术名称	功能	解决的问题
基础特征构造	统计特征构造技术	从原始数据中计算均值、方差、极值等统计量，将数值序列转化为状态描述特征	数据“价值浅”
	比率或差值特征构造技术	通过变量间的除法或减法运算，生成具有物理意义的相对特征	
	逻辑组合特征构造技术	使用与、或、非等逻辑运算符组合多个布尔条件，生成表征复合状态的分类特征	
时序特征提取	时域特征提取技术	提取信号的峰值、谷值、过零点、上升时间等，刻画信号在时间轴上的形态变化	数据“规律隐”
	频域特征提取技术	通过傅里叶变换、小波分析提取频谱能量、主频、谐波等，揭示信号的周期与振动模式	

手段	技术名称	功能	解决的问题
	序列模式特征提取技术	识别序列中的趋势、季节性、突变点、片段模式，捕获动态行为规律	
多模态特征对齐	CLIP 跨模态对齐技术	将现场图片的视觉特征映射到与文本相同的向量空间，实现“搜图”或“看图说话”	数据“链路断”
	子空间学习	假设不同模态的数据存在于一个共享的潜在子空间中通过学习一个映射函数，将各模态数据投影到此子空间，从而实现对齐	
	图神经网络对齐	将不同模态的数据表示为图结构（节点为实体，边为关系），利用 GNN 学习统一的节点嵌入，在图谱层面实现对齐	
	对抗生成式对齐	利用对抗训练机制，通过模态判别器与特征编码器的博弈，迫使编码器生成模态不变的特征表示，实现跨模态特征分布的对齐	
高阶特征构造	交互特征构造	通过四则运算组合多个原始变量，构造具有物理或业务意义的新特征（如温度×压力）	数据“价值浅”
	基于业务知识的特征构造	依据领域专家经验定义与业务目标强相关的特征	
序列与关系建模	滞后特征与窗口统计	生成历史时间点的滞后值或滑动窗口统计量，为模型提供历史上下文信息	数据“链路断”
	序列嵌入技术	使用 RNN、LSTM 或 Transformer 将变长序列编码为固定长度的向量表示，捕获长程依赖	
	图特征提取技术	从设备拓扑、物料流转等关系网络中提取节点、边及子图特征，表征系统关	

手段	技术名称	功能	解决的问题
		联	
语义特征构建	通用文本嵌入技术	将工单、日志、手册等文本转化为高维语义向量，使计算机能理解“温度高”和“过热”是相似特征	数据“规律隐” 数据“链路断”
	提示工程 结构化技术	将原始数据（如报警代码）包装成大模型能理解的自然语言提示词，包含角色设定、任务描述和上下文	
检索增强特征	RAG 分块与索引技术	将长文档切分为语义完整的片段，并提取元数据作为检索特征，为模型提供外挂知识库	
	向量索引技术	通过对高维向量特征进行编码与结构化组织，建立高效的数据结构，以实现近似最近邻搜索	
特征自动化与优化	自动特征生成	利用遗传编程、深度特征合成等方法自动搜索和生成大量候选特征	数据“价值浅”
	特征选择技术	通过过滤法、包裹法或嵌入法，从特征集中筛选出最具判别力的特征子集	数据“价值浅” 数据“规律隐”

3.配套工具清单

表 6　数据特征工程配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
基础特征构造	scikit-learn	数据“价值浅”	统计特征构造技术 交互特征构造技术
	category-encoders		分类特征编码技术 逻辑组合特征技

工具类型	开源框架/工具	解决的问题	对应技术
			术
序列建模工具	tslearn	数据“规律隐” 数据“链路断”	序列嵌入技术
	sktime		时序特征提取技术
信号处理库	PyWavelets	数据“规律隐”	频域特征提取技术
时序特征提取库	Tsfresh scipy.signal	数据“规律隐”	时域/频域特征提取技术 序列模式提取技术
	tsfel		时域/频域/统计特征提取技术
	cesium		时序特征提取与选择技术
自动化特征工程	FeatureTools	数据“价值浅”	自动特征生成技术 深度特征合成技术
	AutoFeat		自动特征构造与选择技术
多模态特征处理	OpenMLDB Feast	数据“链路断”	多模态特征对齐技术
图特征提取	PyG DGL		图特征提取技术
特征选择优化	Boruta mlxtend	数据“价值浅” 数据“规律隐”	特征选择技术
语义向量化工具	HuggingFace Transformers Sentence-Transformers OpenAI Embeddings API	数据“规律隐” 数据“链路断”	通用文本嵌入技术
提示词与编排工具	LangChain LlamaIndex DSPy		提示工程 结构化技术
向量特征存储工	Milvus		向量索引技术

工具类型	开源框架/工具	解决的问题	对应技术
具	Faiss Chroma		

（四）数据标注

数据标注是为后续人工智能应用提供高一致性的训练数据基础的环节，目的是在确保标注准确性、规范性和可扩展性的前提下，将原始数据转化为带有明确标签的训练样本，为监督学习[5]和部分半监督学习[6]算法提供“标准答案”，以驱动模型建立从输入到输出的正确映射关系。

1.解决的问题

数据“无标签”。制造业采集的原始数据（如设备运行时序数据、质检图片、工艺参数序列）未附带 AI 模型可识别的标签信息（如“设备故障类型”“产品缺陷等级”“工艺优化方向”），导致监督学习模型（如故障分类、质量分级 AI）无法建立“数据特征-业务结果”的映射关系，难以直接用于训练。**标注“不一致”。**同一批数据在不同标注者、不同时间、不同标注场景下，得到的标签结果不一致（如同一产品缺陷图片，A 专家标为“划痕”，B 专家标为“凹陷”），导致训练数据的标签可信度低，AI 模型学习到模糊的决策边界，精度下降。**知识“难固化”。**制造业在数据应用过程中产生的隐性知识（专家经验、工艺诀窍）和显性知识（数据分析结论、AI 模型规则），因缺乏标准化沉淀机制、业务融合不足等原因，无法转化为可复用、可传承的资产，导致数据价值无法持续赋能业务。

2.核心技术应用

表 7　数据标注核心技术应用

手段	技术名称	功能	解决的问题
图像、视频标注	语义分割	基于语义对场景进行分区	数据“无标签”
	拉框标注	包括直线、矩形、多边形等多种形式，取决于物体轮廓	
	关键点标注	标注关键部位及属性，如人脸识别	
	立方体标注	对 2D 图片中的物体进行 3D 标注，用于体积判断	
	3D 点云标注	对激光雷达采集的点云图标注目标对象	
	2D/3D 融合标注	对 2D 和 3D 传感器数据进行标注并建立关联	
	4D 目标追踪	按帧捕捉对象	
	光学字符识别（OCR 转写）	对图像中的文字内容进行标记和转写	
文本标注	词性标注	为词语标注词性标签	
	语义角色标注	识别句子中的谓词和论元，并标注其语义角色	
	情绪标注	用于确定文本中表达的情感极性或情绪状态	
	命名实体识别	识别并标注文本中有特定名称的实体	
	语义标注	将词汇、短语或句子与特定含义或语义信息相关联	
	关系标注	识别和描述文本中不同实体之间的关系	
	关键词抽取	从文本中提取关键信息	
音频标注	语音切割	拆分长音频为短句/片段，适配处理需求	
智能预标注	弱监督学习	利用业务规则、历史记录等弱标签自动生成初始标注	
	预训练模型微调	使用通用预训练模型对未标注数据进行预测，生成伪标	

手段	技术名称	功能	解决的问题
		签	
标注流程管理	多人标注与共识机制	通过多人独立标注、交叉验证与仲裁流程，收敛至统一、可靠的标注结果	标注“不一致”
	标注规范嵌入	制定详细、可视化的标注准则与样本库，明确各类别的判定边界与量化指标	
	自动化一致性校验	利用算法检测同一数据集中冲突或离群的标注，自动提示复审	
领域知识注入	知识图谱引导标注	利用领域知识图谱约束标注选项，确保标注的语义一致性与逻辑正确性	知识“难固化”
	交互式主动学习	系统基于不确定性采样、模型熵或决策边界多样性等量化指标，智能筛选出对模型优化最具信息增益的候选样本，引导领域专家优先标注此类高价值数据	

3.配套工具清单

表 8　数据标注配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
多模态数据标注工具	Labelbox LabelMe ImageTagger LabelImg T-Rex Label Windows 画图 记事本 VS Code Adobe Reader	数据“无标签”	多模态融合标注
专业工程软件插件	CAD 坐标标注插件 QGIS / ArcGIS 插件		
数据标注平台	CVAT Labelbox Label Studio	标注“不一致”	标注规则嵌入 多人标注与共识

工具类型	开源框架/工具	解决的问题	对应技术
	Supervisely		机制
自动化校验工具	CVAT Analytics Scale Validate		自动化一致性校验
智能预标注工具	Label Studio ML Backend CVAT AI Tools	数据“无标签”	预训练模型微调
弱监督标注工具	Snorkel		弱监督学习
预训练与清洗工具	Cleanlab ActiveClean	标注“不一致”	预训练模型微调 自动化一致性校验
知识引导标注工具	Knowlege-Aided Annotation Prodigy-like Systems	知识“难固化”	知识图谱引导标注
交互式学习框架	ModAL libact	数据“无标签” 知识“难固化”	交互式主动学习
训练推理一体化平台	Label Studio ML Backend Prodigy Hikvision AI 开放平台	知识“难固化”	弱监督学习 预训练模型微调 交互式主动学习

（五）数据增强

数据增强是为后续人工智能应用提供高仿真、强可控、规模化增强数据的环节，目的是在确保数据真实性、多样性和隐私安全的前提下，通过生成式技术填补真实数据分布的空白与不足，解决小样本、长尾分布[7]、极端场景和数据隐私等核心瓶颈，从而提升模型的泛化能力、鲁棒性与均衡性。

1.解决的问题

关键样本“极稀缺”。AI 模型训练所需的关键样本（如设备致命故障数据、产品严重缺陷数据、极端工况运行数据）

因发生概率极低、采集难度大、复现成本高，导致样本数量远无法满足模型训练需求，最终导致模型对关键场景的识别能力薄弱。**数据分布“不均衡”。**数据在类别分布、时序分布、空间分布、维度分布上存在显著差异，导致 AI 模型训练偏向多数类数据，无法公平学习不同场景的特征，最终影响模型对少数类场景的识别精度（如模型能精准识别常见故障，却漏判罕见故障）。**数据隐私与合规“门槛高”。**部分数据涉及商业秘密、核心技术、个人信息、跨境流动需求，且受行业特定监管政策约束，隐私保护与合规要求高于普通行业，导致企业需投入更高成本才能合法开展相关数据的采集与应用。

2.核心技术应用

表 9　数据增强核心技术应用

手段	技术名称	功能	解决的问题
生成式模型合成	生成对抗网络技术	通过生成器与判别器的对抗训练，学习真实数据分布并生成高保真合成样本	关键样本“极稀缺” 数据分布“不均衡”
	变分自编码器技术	学习数据潜在空间分布，通过采样与解码生成符合原始统计特性的新样本	
	扩散模型技术	通过逐步去噪过程从随机噪声生成高质量、多样化的数据样本	关键样本“极稀缺”
基于仿真的合成	物理仿真建模技术	依据物理定律与工艺原理构建仿真环境，生成带精确标签的极端工况数据	
	数字孪生模拟技术	在虚拟镜像中复现真实生产系统，模拟各类故障与异常并生成相应数据	
数据变换与增强	智能数据增强技术	对现有少数样本进行有指导的几何、光照、噪声变换，定向扩充样本多样性	数据分布“不均衡”
隐私保护合成	差分隐私合成	在合成过程中注入可控噪声，确保单个真实样本信息无法从合成数据中推断	数据隐私与合规“门槛高”
	联邦学习合成	在本地训练生成模型并仅交换模型参数，基于聚合知识生成全局代表性合成数据	
	合成数据脱敏	生成保留全局统计规律与关联结构但抹去敏感属性与可识别信息的替代数据	

3.配套工具清单

表 10　数据增强配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
生成对抗网络工具	StyleGAN2/3 CycleGAN	关键样本“极稀缺” 数据分布“不均衡”	生成对抗网络
变分自编码器工具	Pyro VAE		变分自编码器
扩散模型工具	Stable Diffusion DALL-E Mini	关键样本“极稀缺”	扩散模型
物理仿真工具	Gazebo OpenFOAM		物理仿真建模
数字孪生平台	Eclipse Ditto Apache StreamPipes		数字孪生模拟
数据增强库	Albumentations	数据分布“不均衡”	智能数据增强
差分隐私工具	TensorFlow Privacy PySyft	数据隐私与合规“门槛高”	差分隐私合成 联邦学习合成
联邦学习框架	FATE Flower		联邦学习合成
合成数据生成平台	SDV CTGAN/TVAE		合成数据脱敏

（六）数据划分

数据划分是为后续人工智能模型训练、验证及测试提供合理数据分配方案的环节，目的是在确保数据分布代表性、样本划分随机性和集合独立性的前提下，将整体数据集划分为训练集、验证集和测试集，从而支撑模型参数优化、超参数调优和泛化能力评估，保障 AI 模型训练过程科学可控、评估结果真实可信。

1.解决的问题

划分“不科学”。AI 模型训练用的数据集因划分比例失衡、随机抽样偏差、未考虑数据分布特性等原因，无法保证

训练集、验证集、测试集的样本代表性，导致模型训练出现过拟合或欠拟合问题，最终泛化能力不足。**划分“不匹配”。**数据集在划分策略、样本分层、时序划分规则等方面与 AI 模型类型、业务场景需求或模型评估标准不兼容，导致划分后的数据集无法有效支撑模型训练调优与性能验证，或评估结果无法反映真实业务效果。**划分“非动态”。**制造业生产场景具有动态变化特性，导致历史划分的数据集特征分布逐渐偏离当前生产状态，训练集与实时业务数据的分布差异持续扩大，模型基于历史划分数据集训练的效果不断衰减，最终无法支撑 AI 模型的持续迭代与业务决策优化。

2.核心技术应用

表 11　数据划分核心技术应用

手段	技术名称	功能	解决的问题
时序感知划分	时间序列划分	严格按时间顺序划分，测试集时间必须在训练集之后	划分“非动态” 划分“不科学”
	滚动时间窗划分	使用滑动时间窗口模拟模型随时间迭代更新的过程	划分“非动态”
随机抽样划分	留出法	通过随机划分提供基础的数据隔离，保证划分的简单性和可重复性	划分“不科学”
	自助法	通过有放回抽样生成多样化的训练集，评估模型在小样本下的稳定性与不确定性	划分“不科学” 划分“非动态”
交叉验证划分	K 折交叉验证	通过 K 轮次不同划分下的平均性能，提供更稳健、低偏差的模型性能评估	划分“不科学”
	分层 K 折交叉验证	在 K 折划分中强制保持各类别比例，确保评估在类别	划分“不科学”

手段	技术名称	功能	解决的问题
		不平衡的数据集上依然有效	划分“不匹配”
	留一法	每次仅用一个样本测试，最大程度利用数据，提供近乎无偏的评估	划分“不科学”
领域感知划分	设备/产线分离划分	将来自不同物理实体的数据完全分离到训练集和测试集，检验跨实体泛化能力	划分“不匹配”
	工况/批次分离划分	按不同原材料批次、工艺配方或环境条件分离数据，评估模型在不同生产条件下的稳定性	
智能诊断划分	对抗验证	训练一个分类器来区分训练集和测试集，若易区分则表明划分不合理，分布不一致，需要调整	划分“不匹配”
业务驱动划分	业务场景划分	根据实际部署计划（如“旧线训练，新线测试”“历史产品训练，新产品测试”）直接划分	划分“非动态”

3.配套工具清单

表 12　数据划分配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
基础划分工具	scikit-learn (train_test_split)	划分“不科学”	留出法 分层 K 折交叉验证
	scikit-learn (KFold, StratifiedKFold)	划分“不科学”划分“不匹配”	K 折交叉验证 分层 K 折交叉验证
高级验证工具	scikit-learn (TimeSeriesSplit)	划分“非动态”	时间序列划分
	Tslearn (TimeSeriesSplit)	划分“非动态”	时间序列划分 滚动时间窗划分
领域感知工具	自定义 Python 脚本	划分“不匹配”	设备/产线分离划分 工况/批次分离划

工具类型	开源框架/工具	解决的问题	对应技术
			分
	Great Expectations（数据验证）	划分“不匹配”	对抗验证
智能划分工具	Alibi Detect (adversarial.py)	划分“不匹配”	对抗验证
	Cleanlab (crossval.py)	划分“不科学”	分层 K 折交叉验证
时序专用工具	sktime (TemporalTrainTestSplitter)	划分“非动态”	时间序列划分 滚动时间窗划分
	Prophet（内置交叉验证）	划分“非动态”	时间序列划分
小样本工具	scikit-learn (LeaveOneOut, Bootstrap)	划分“不科学”	留一法 自助法
	imbalanced-learn (StratifiedKFold)	划分“不科学”	分层 K 折交叉验证

（七）数据存储

数据存储是为后续人工智能应用环节提供数据存取基础设施的环节，目的是在确保数据可追溯、高可用的前提下，为模型训练与推理提供稳定、高性能、低成本的数据供给与服务支撑。

1.解决的问题

结构化数据存储“不关联”。因企业生产经营中的设备参数、生产报表、质检结果、订单信息等结构化数据未建立有效的关联关系，数据之间相互孤立，无法形成完整的业务逻辑链。**特征向量数据存储“效率低”**。由于企业选取存储特征向量数据的架构选型与 AI 模型训练时数据调取请求逻辑不匹配，导致数据写入、读取、检索的效率低下，无法支

撑大规模模型训练。**非结构化数据存储“压力大”**。企业生产中产生的设备故障照片、工艺视频、质检报告文本、传感器原始波形文件等非结构化数据体积量大且增长速度快，导致数据存储容量、成本、管理压力大。**异构数据存储“难整合”**。企业生产过程中的结构化、半结构化、非结构化等类型数据多存储于不同的存储系统，由于存储协议、数据格式、元数据描述规则不一致等，导致跨类型跨系统数据难以整合分析，无法满足 AI 模型训练所需的多源异构数据需求。

2.核心技术应用

表 13　数据存储核心技术应用

手段	技术名称	功能	解决的问题
数据库存储	关系型数据库技术	通过严格的数据完整性约束和外键关联，确保核心业务数据的一致性及复杂业务关系可追溯	结构化数据存储“不关联”
	混合事务和分析处理（HTAP）数据库技术	融合 OLTP 与 OLAP 能力，在同一数据库中同时处理交易与分析，消除数据同步延迟，支撑实时决策	结构化数据存储“不关联”
	图数据库技术	以节点和边的原生存储结构，高效建模与遍历供应链、工艺路线等复杂网络关系，实现深度关联挖掘	结构化数据存储“不关联”
	向量数据库技术	针对高维向量数据提供高效的相似性搜索索引，实现 AI 场景下海量特征向	特征向量数据存储“效率低”

手段	技术名称	功能	解决的问题
		量的快速检索与匹配	
	时序数据库技术	采用时间分区、高效压缩和列式存储，专为海量时间序列数据提供高吞吐写入与低成本长期存储	非结构化数据存储“压力大”
	实时数据库技术	基于内存优先架构和确定性调度，实现生产现场毫秒级高并发数据读写，满足实时监控与控制需求	非结构化数据存储“压力大”
文件数据存储	分布式文件存储	将文件分布到多个节点，提供高吞吐访问，统一管理各类文件数据	非结构化数据存储“压力大” 异构数据存储“难整合”
	对象存储技术	通过扁平化结构存储海量非结构化数据，提供 RESTful 接口实现跨系统访问	
	软件定义存储（SDS）技术	将存储硬件与软件解耦，通过策略灵活调配异构存储资源池	
	文件数据压缩存储	对文件进行无损或有损压缩，显著减少占用的物理存储空间	非结构化数据存储“压力大”
云端数据存储技术	IaaS 云存储技术	提供虚拟化的基础设施资源，实现存储容量的弹性扩展与按需使用	非结构化数据存储“压力大” 异构数据存储“难整合”
	PaaS 云存储技术	提供数据库、中间件等存储开发平台，简化海量数据	

手段	技术名称	功能	解决的问题
		应用的构建与管理	
	SaaS 云存储技术	通过即开即用的软件服务，直接提供已整合的业务数据存储与应用	异构数据存储“难整合”
	云原生数据库	基于云环境开发，采用存算分离与微服务架构，天然支持全局数据一致与高性能访问	结构化数据存储“不关联” 特征向量数据存储“效率低”
	分布式对象存储	在对象存储基础上引入分布式架构，实现容量与性能的线性扩展及统一命名空间	非结构化数据存储“压力大” 异构数据存储“难整合”

3.配套工具清单

表 14　数据存储配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
OLTP 数据库	PostgreSQL MySQL/MariaDB	结构化数据存储“不关联”	关系型数据库技术
OLAP 数据库	Apache Doris ClickHouse StarRocks Presto/Trino		
HTAP 数据库	TiDB, Apache Doris、StarRocks		HTAP 数据库技术
图数据库	Neo4j NebulaGraph JanusGraph		图数据库技术
向量数据库	Milvus Chroma Elasticsearch (Vector)	特征向量数据存储“效率低”	向量数据库技术
时序数据库	InfluxDB （开源版） Prometheus TDengine TimescaleDB	非结构化数据存储“压力大”	时序数据库技术
实时数据库	Apache IoTDB QuestDB RocketMQ-Streams	非结构化数据存储“压力大”	实时数据库技术
分布式文件存储	CephFS GlusterFS HDFS MooseFS	非结构化数据存储“压力大” 异构数据存储“难整合”	分布式文件存储技术
对象存储	Ceph RADOS MinIO OpenStack Swift		对象存储技术
云存储服务	/		IaaS/PaaS 云存储技术

（八）数据计算

数据计算是为后续人工智能应用提供大规模数据传输和计算基础框架的环节，目的是在确保计算效率、资源弹性

和结果可靠的前提下，为模型训练与推理提供规模化、智能化、可复现的数据加工与价值转化能力。

1.解决的问题

数据计算“不协同”。企业内部不同计算节点、不同计算框架、不同业务系统的计算能力独立运行，会导致计算资源孤岛、任务调度割裂、数据流转不畅等问题，无法满足复杂的生产全流程数据分析需求。**数据计算“不按需分配”**。由于计算资源分配与业务需求不匹配，导致企业无法根据任务的优先级、计算规模、时效性要求动态调度资源。**数据计算“不实时响应”**。当企业采用的计算架构不匹配实施需求，或实时计算任务中嵌入过多非必要环节时，将导致计算结果输出滞后于业务决策窗口，无法支撑设备预测性维护、产线实时质检、工艺动态优化等低时延需求场景。**数据计算“算力不足”**。因企业可用的计算资源总量无法满足业务增长和技术升级的需求，导致面对大规模 AI 模型训练、多模态数据融合计算、复杂工艺仿真等高密度算力需求场景应用落地受阻。

2.核心技术应用

表 15　数据计算核心技术应用

手段	技术名称	功能	解决的问题
计算架构	批流融合计算技术	提供统一 API 处理历史与实时数据流，保障计算逻辑一致与结果同步	数据计算“不协同”
资源调度与编排	混合负载调度技术	在统一资源池中动态调度批处理与流式任务，实现资源隔离与共享	

手段	技术名称	功能	解决的问题
	弹性伸缩技术	根据计算负载自动扩缩容计算节点，应对业务峰值与低谷	数据计算“不按需分配”
	边缘计算编排技术	将云原生能力延伸至边缘，实现边缘算力统一管理与任务分发	
	无服务器计算	提供事件驱动的函数即服务模型，细化资源管理的粒度，实现按需分配和快速部署	
实时计算与流处理	低延迟事件处理技术	提供毫秒级事件处理与状态计算能力，实现实时响应与闭环控制	数据计算“不实时响应”
	复杂事件处理技术	提供模式匹配专用 API 或 SQL，实时识别数据流中的复杂事件序列	
	内存计算	通过非易失性内存与智能缓存技术，在保证数据持久性的前提下，将热数据与中间状态置于近计算单元的高速存储层，消除传统 I/O 瓶颈，实现兼具高可靠性与毫秒/微秒级响应的实时数据访问	
高性能与并行计算	分布式并行计算技术	将大规模计算任务分解为子任务，在集群中并行执行，以横向扩展提升总体算力	数据计算“算力不足”
	GPU 加速计算技术	利用 GPU 的众核架构进行大规模并行计算，显著加速矩阵运算、图像处理等任务	
计算卸载与优化	异构计算技术	将特定计算密集型算法（如编码、推理）固化到专用硬件，实现超高能效比和低延迟	

手段	技术名称	功能	解决的问题
	近似计算与查询优化技术	在 CPU 层面利用 SIMD 指令集进行批量数据操作，以及通过查询计划优化减少不必要的计算	

3.配套工具清单

表 16　数据计算配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
批流融合计算引擎	Apache Flink Apache Spark	数据计算“不协同”	批流融合计算技术
混合负载调度平台	Kubernetes Apache YARN Slurm		混合负载调度技术
弹性资源调度系统	Kubernetes HPA/VPA	数据计算“不按需分配”	弹性资源调度技术
边缘计算编排框架	KubeEdge OpenYurt		边缘计算编排技术
无服务器计算平台	OpenFaaS Knative		无服务器计算技术
流处理计算引擎	Apache Flink Apache Storm	数据计算“不实时响应”	低延迟流处理技术
复杂事件处理引擎	Apache Flink CEP Esper		复杂事件处理技术
内存计算系统	Apache Ignite Alluxio Redis		内存计算技术
分布式计算框架	Apache Spark Dask Ray	数据计算“算力不足”	分布式并行计算技术
GPU 加速计算平台	CUDA Toolkit ROCm oneAPI		GPU 加速计算技术
异构计算开发框架	OpenCL Vitis AI TensorRT		异构计算技术

工具类型	开源框架/工具	解决的问题	对应技术
向量化计算引擎	Apache Arrow Modin Polars		向量化与查询优化

（九）数据集成

数据集成是为后续人工智能应用提供多源原始数据治理环境的环节，目的是在确保数据一致性、关联性和可解释性的前提下，为模型训练与推理构建统一、可靠、高效的多源数据融合底座，以消除数据孤岛并支撑跨域联合分析。

1.解决的问题

数据集成“标准不统一”。企业内部的ERP、MES、SCADA、PLC 等系统往往由不同供应商在不同时期建设，其数据格式、编码、语义定义及通信协议等缺乏统一规范，导致系统间数据无法实现集中访问和分析。**数据集成“时效性差”**。由于企业缺乏实时数据采集技术或手段，导致数据从源头采集、传输到集成入库的全链路时延过长，无法满足生产实时监控、故障预警、动态工艺优化等场景的低时延需求。**数据集成“质量不可控”**。因企业在数据集成过程中未建立全流程质量管控机制，或数据在跨系统传输过程中常因网络中断、接口异常或业务规则冲突等原因，导致脏数据、错误数据、缺失数据直接流入数据库，集成后的数据直接影响生产执行与决策可靠性。**数据集成“链路不透明”**。由于企业缺少从源头到集成入库的全生命周期流转追溯和监控机制，导致集成链路中的节点状态、数据处理规则、异常情况无法可视化呈现，

影响异常数据溯源能力。**数据集成“不关联”**。企业多源异构据（图纸、文档等）缺乏有效的关联关系定义与统一语义映射，导致无法支撑跨域、深度的业务分析与智能应用。

2.核心技术应用

表 17　数据集成核心技术应用

手段	技术名称	功能	解决的问题
物理汇聚集成	数据仓库技术	对清洗后的结构化数据按主题域建模存储，保证分析数据的质量和一致性	数据集成“标准不统一” 数据集成“质量不可控”
	数据湖技术	以原始格式集中存储海量异构数据，支持灵活的数据探索与后治理	数据集成“标准不统一”
	数据湖仓一体技术	在数据湖的存储基础上实现数据仓库的 ACID 事务、强 Schema 管理与高性能分析，统一元数据治理	数据集成“标准不统一” 数据集成“质量不可控” 数据集成“链路不透明”
管道传输集成	数据管道技术	构建自动化的数据抽取、转换、加载流程，支持批流一体，实现数据从源到目的地的可靠移动	数据集成“时效性差” 数据集成“质量不可控”
逻辑虚拟集成	数据虚拟化技术	通过中间层提供统一的逻辑数据视图，实时聚合分散的源数据，避免物理移动与复制	数据集成“标准不统一” 数据集成“时效性差”
语义统一集成	元数据管理技术	集中管理技术、业务与操作元数据，形成企业数据目录，统一数据定义与语义	数据集成“标准不统一” 数据集成“链路不透明”
链路追溯集成	数据血缘分析技术	自动追踪数据从源到消费端的完整加工与流转路径，实现影响分析和根因定位	数据集成“链路不透明”

手段	技术名称	功能	解决的问题
质量管控集成	数据质量评估与清洗技术	通过规则引擎与算法对数据进行探查、清洗、监控与评分，确保数据的准确性、完整性与及时性	数据集成“质量不可控”
跨模态融合集成	向量嵌入技术 知识图谱构建技术	将非结构化数据（图纸、文档）转化为高维向量，实现语义级索引；建立结构化数据与非结构化实体之间的关联网络	数据集成“标准不统一” 数据集成“不关联”
智能数据融合	大模型编排技术	通过编排和调用一个或多个大语言模型，将 AI 的语义理解、逻辑推理和内容生成能力，嵌入到传统数据集成的各个环节，以解决异构数据源集成问题	数据集成“标准不统一” 数据集成“时效性差” 数据集成“不关联”

3.配套工具清单

表 18　数据集成配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
数据湖存储平台	Apache Hudi Delta Lake Apache Iceberg	数据集成“标准不统一”	数据湖技术
数据湖仓查询引擎	Apache Hive Presto/Trino Apache Spark SQL	数据集成“标准不统一” 数据集成“时效性差”	数据湖仓一体技术
数据仓库引擎	Apache Doris ClickHouse Greenplum	数据集成“标准不统一” 数据集成“质量不可控”	数据仓库技术
数据虚拟化平台	Dremio Presto/Trino Apache Drill	数据集成“标准不统一” 数据集成“时效性差”	数据虚拟化技术

工具类型	开源框架/工具	解决的问题	对应技术
批处理数据管道	Apache Airflow Apache NiFi Dagster	数据集成“时效性差” 数据集成“质量不可控”	数据管道技术
流处理数据管道	Apache Flink Apache Kafka Connect Apache Pulsar IO	数据集成“时效性差” 数据集成“质量不可控”	数据管道技术
元数据管理平台	Apache Atlas DataHub Amundsen	数据集成“标准不统一” 数据集成“链路不透明”	元数据管理技术
数据质量工具	Great Expectations Deequ Soda Core	数据集成“质量不可控”	数据质量评估与清洗技术
数据血缘分析工具	OpenLineage Marquez DataHub	数据集成“链路不透明”	数据血缘分析技术
大模型编排框架	LangChain LlamaIndex	数据集成“标准不统一” 数据集成“时效性差” 数据集成“不关联”	大模型编排技术

（十）数据质量评估

数据质量评估是为后续人工智能应用提供高质量数据及其评估体系的环节，目的是在确保评估维度全面性、指标科学性和流程标准化的前提下，对数据的准确性、完整性、一致性、时效性及与业务的相关性进行客观评价，从而识别数据缺陷、量化数据风险，保障输入 AI 模型的数据基础坚实可靠。

1.解决的问题

标注“不可靠”。AI 模型训练用的标签数据因标注错误、缺乏依据、验证缺失等原因，无法真实反映数据对应的业务状态（如将“设备正常数据”错误标注为“故障数据”），导致标签可信度低，模型学习到错误的“特征-标签”映射关系，最终决策失准。**数据“不匹配”。**数据在格式、维度、语义、时效、场景等方面与 AI 模型训练要求、实际生产场景需求或跨环节数据流转需求不兼容，导致数据无法直接用于建模，或建模后模型无法适配真实业务场景。**数据“动态劣化”。**制造业生产场景具有动态变化特性（如设备老化、工艺优化、原材料更换、环境波动），导致历史数据的特征分布、业务含义逐渐偏离当前实际状态，数据的时效性与有效性下降，最终无法支撑 AI 模型的持续优化与业务决策。**评估“效率低”。**传统评估方法存在高度依赖专家人工审核、难以灵活处理语义相关性、上下文一致性与领域术语的标准化、缺失闭环反馈等问题，导致评估效率较低，无法形成自动化、智能化、可闭环的持续数据质量管理流程。

2.核心技术应用

表 19　数据质量评估核心技术应用

手段	技术名称	功能	解决问题
准确性验证	基于规则的校验	通过预设的业务规则与物理阈值范围，自动识别并标记不符合预期的异常数据点	标注“不可靠” 数据“不匹配”
	统计异常检测	利用箱型图、Z-score、孤立森林等统计方法，识别偏离整体分布的离群值与噪声	

手段	技术名称	功能	解决问题
完整性评估	缺失模式分析	分析数据缺失的规律（随机缺失/系统缺失），评估缺失对分析与建模的潜在影响	数据“不匹配”
	数据采集链路监控	监控从源头到存储的各环节数据流，定位中断或延迟的采集节点	数据“动态劣化”
一致性校验	多源数据对齐验证	对比来自不同系统的同一实体数据，识别矛盾与不一致之处	数据“不匹配”
	模式一致性分析	检验数据的时间序列模式、周期性或关联关系是否符合已知的业务逻辑或历史规律	数据“动态劣化”
时效性监控	数据实时性度量	计算数据从产生到可用的时间差，监控其是否满足业务实时性要求	
标注质量评估	交叉验证与共识评估	通过多人重复标注计算一致性指标，识别标注分歧与错误	标注“不可靠”
	基于模型的噪声标签检测	利用模型预测置信度或训练损失，识别可能存在错误的标注样本	
语义质量评估	LLM-as-a-Judge 技术（大模型即裁判）	利用大模型作为裁判，对本地数据的相关性、准确性、有害性进行打分	评估“效率低”
检索增强评估	RAG 三元组评估技术	评估提问、检索到的知识、生成答案三者之间的相关度	
数据自动修复	生成式数据清洗 Agents	利用大模型的理解能力，自动纠正拼写错误、补全缺失的描述、标准化非标术语，并自动回写数据库	

3.配套工具清单

表 20　数据质量评估配套工具清单

工具类型	开源框架/工具	解决问题	对应技术
规则校验工具	Great Expectations	标注“不可靠” 数据“不匹配”	基于规则的校验 多源数据对齐验证

工具类型	开源框架/工具	解决问题	对应技术
	Deequ		基于规则的校验 统计异常检测
统计异常检测工具	PyOD		统计异常检测
	Alibi Detect	数据“动态劣化”	数据漂移检测 模式一致性分析
数据质量监控平台	Apache Griffin	数据“动态劣化” 数据“不匹配”	数据实时性度量 多源数据对齐验证
	DataHub		数据采集链路监控 一致性校验
标注质量评估工具	Cleanlab	标注“不可靠”	基于模型的噪声标签检测
	Label Studio		交叉验证与共识评估
时序数据质量工具	TDengine	数据“动态劣化”	缺失模式分析
	Prometheus Grafana		数据实时性度量
元数据管理工具	Amundsen	数据“不匹配”	多源数据对齐验证
	OpenMetadata		数据血缘分析 一致性校验
RAG/大模型评估框架	Ragas TruLens DeepEval	评估“效率低”	LLM-as-a-Judge 技术 （大模型即裁判）
向量数据质量工具	Galileo Arize Phoenix		RAG 三元组评估技术
数据自动修复工具	LangChain LlamaIndex Ollama Transformers 库		生成式数据清洗 Agents

（十一）数据安全保护

数据安全保护为后续人工智能应用提供受控、合规、可信的数据访问与使用环境，目的是通过身份认证、访问控制、

加密脱敏、审计溯源等技术手段，防范数据泄露、篡改、滥用及非法访问风险，满足法规遵从与隐私保护要求。

1.解决的问题

数据“易泄露”。制造业核心数据（如工艺参数、设备控制代码、生产排程数据、客户定制化需求）在“采集-传输-存储-使用-销毁”全流程中，因安全防护不足、权限管理漏洞、工业场景特殊性等原因，被未授权主体非法获取、泄露或滥用，导致企业商业秘密、知识产权或客户隐私受损。**数据“被污染”。**数据在采集、传输、处理或存储过程中，因工业场景动态性、人为干预、系统异常等原因，被恶意篡改、无意干扰或错误融合，导致数据偏离真实生产状态，形成“污染数据”。这类数据会误导 AI 模型训练或导致生产决策失误。**数据“难合规”。**制造业数据因涉及核心工业数据、个人信息、跨境流动等数据特殊属性，且受工业行业特定监管政策约束，导致在采集、存储、传输、使用、销毁等环节难以满足《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》及行业标准等的要求，企业面临行政处罚、业务限制或法律诉讼的风险。

2.核心技术应用

表 21　数据安全保护核心技术应用

手段	技术名称	功能	解决的问题
访问控制	基于属性的访问控制	根据用户角色、设备状态、地理位置等多维属性动态授权数据访问权限	数据“易泄露”
	零信任网络	默认不信任任何访问请求，对每次	数据“易泄

手段	技术名称	功能	解决的问题
	架构	访问进行身份验证和权限校验	露”
加密保护	同态加密	支持在加密状态下直接进行数据计算，避免数据处理时的明文暴露风险	数据“易泄露”数据“难合规”
	传输层/存储加密	对数据在传输和存储过程中进行强加密，防止中间人攻击和物理窃取	数据“易泄露”
数据脱敏	文本去噪技术 隐私擦除	去除网页标签、乱码、重复段落自动识别并替换敏感信息，匿名化和去标识化	
	动态数据脱敏	根据用户权限实时返回不同脱敏程度的数据，平衡数据使用与安全	
	差分隐私技术	在数据中注入可控噪声，保证统计分析有效性的同时防止个体信息泄露	
	隐私实体识别	基于序列标注的自然语言处理技术，用于自动检测和分类文本中的敏感个人信息（如姓名、地址、证件号等）	
完整性保护	数字签名/区块链	为数据记录添加不可篡改的数字指纹，确保数据来源可信与内容完整	数据“被污染”
	哈希校验与防篡改	通过哈希算法验证数据在传输和存储过程中是否被修改	数据“被污染”
审计溯源	数据血缘追踪	完整记录数据的来源、流转路径和处理过程，实现全生命周期可追溯	数据“被污染”数据“难合规”
	安全信息与事件管理	集中收集和分析各类日志，实时检测异常访问和潜在威胁	数据“易泄露”
隐私计算	联邦学习	多方协同训练模型而无需共享原始数据，保障数据不出域	数据“难合规”数据“易泄露”
	安全多方计算	多个参与方在不泄露各自输入数据的前提下协同完成计算任务	数据“难合规”数据“易泄露”
大模型交互安全	AI 防火墙/围栏技术	对用户输入的提示词和模型输出的内容进行实时检测、拦截和过滤，防止恶意诱导和隐私泄露	数据“易泄露”

手段	技术名称	功能	解决的问题
模型资产保护	模型水印技术 模型环境感知	在模型权重或生成的内容中嵌入不可见的特征码，用于版权追踪和 AI 生成内容标识当模型离开特定环境，所有参数自动加密或者实现专有数据知识自我删除	
模型鲁棒性安全	对抗防御技术	识别并防御专门针对神经网络的“对抗样本”攻击，防止微小的扰动导致模型判断错误	数据“被污染”
遗忘与合规	机器遗忘技术	允许从模型中删除特定数据	数据“易泄露” 数据“被污染”

3.配套工具清单

表 22　数据安全保护配套工具清单

工具类型	开源框架/工具	解决的问题	对应技术
访问控制工具	Keycloak	数据“易泄露”	基于属性的访问控制
	OpenZiti	数据“易泄露”	零信任网络架构
加密工具	OpenSSL	数据“易泄露”	传输层/存储加密
	Microsoft SEAL	数据“易泄露” 数据“难合规”	同态加密
脱敏工具	Beautiful Soup + html2text Presidio ARX Amnesia	数据“易泄露”	文本去噪技术 隐私擦除
	Apache ShardingSphere		动态数据脱敏
	OpenDP		差分隐私技术
	Microsoft Presidio Cleanlab		隐私实体识别
完整性工具	Hyperledger Fabric	数据“被污染”	数字签名/区块链
	OpenAttestation	数据“被污染”	哈希校验与防篡改
审计溯源工具	Apache Atlas	数据“被污染” 数据“难合规”	数据血缘追踪

工具类型	开源框架/工具	解决的问题	对应技术
	Wazuh	数据“易泄露”	安全信息与事件管理
隐私计算工具	FATE	数据“难合规” 数据“易泄露”	联邦学习
	MP-SPDZ	数据“难合规” 数据“易泄露”	安全多方计算
大模型护栏工具	Lakera Guard	数据“易泄露”	AI 防火墙/围栏技术
对抗攻防工具箱	IBM Adversarial Robustness Toolbox (ART) Microsoft Counterfit	数据“被污染”	对抗防御技术
大模型审计与监控	LangSmith	数据“易泄露”	安全信息与事件管理
近似遗忘算法工具	SISA	数据“被污染” 数据“易泄露”	机器遗忘技术

三、面向 AI 数据治理的企业实践路径

围绕企业 AI 数据治理“先试点探索、再全域推广、终产业延伸”的总体思路，即从单部门单场景启动 AI 数据治理试点，总结沉淀可复制、可推广的经验做法后，逐步拓展至企业全域多部门多场景，最终延伸覆盖产业链供应链上下游，构建产业级数据治理体系。基于此，分别从单部门（单场景）试点探索、多部门（多场景）推广实施、产业链（供应链）上下游协同治理三个阶段，明确具体实践路径，制造业企业可根据自身所处的阶段及需求，开展相关工作。

（一）第一阶段：单部门（单场景）试点探索

聚焦制造业企业在单一业务场景中探索应用人工智能技术应用的实践路径，以试点场景筛选方法论为切入点，为企业结合自身实际情况，科学识别、遴选适配试点开展的高潜力场景，提供思路参考。试点场景确定后，进一步为企业面向人工智能应用的数据治理工作，提供关键实施要点，主要包括治理前期的跨职能试点团队组建、全域数据资产盘点工作，中期的数据质量标准制定、数据质量治理以及支撑场景落地的数据工程能力建设工作，后期的治理机制构建、价值挖掘及绩效评价工作。

1.AI 场景识别

（1）AI 场景识别流程

明确 AI 应用必要性，由各业务部门牵头，从影响利润、交付、合规等角度，分析企业当前在工厂建设、产品研发、

工艺设计、生产管理、生产作业、运营管理、产品服务、供应链管理等环节存在的生产经营痛点，识别高价值场景。**评估 AI 应用可行性，**组织跨部门团队会议，对候选场景利用人工智能的高质量数据供给、技术成熟度、成本投入与回收、人员协同机制等情况进行评估，按照范围小、见效快、可复制的原则，选择 1 个最小可行试点场景进行探索。

（2）场景 AI 应用原则

对标准化、规则明确且数据可采集的场景，如在线智能检测、设备运行监控与维护等场景，可重点利用 AI 技术进行人工替代。**对任务逻辑清晰但缺乏足够历史数据的场景，**如智能设计与虚拟验证闭环场景等场景，可由人工依据经验知识进行任务决策，由 AI 进行执行与迭代。**对任务复杂但高数据可用的场景，**如工艺动态优化、柔性产线快速换产等场景，可采用人机协同[8]策略，由 AI 结合多源异构数据提供解决方案，最终由人工进行决策并利用数据反向对模型调优。**对依赖经验且难以格式化的场景，**如依赖老师傅经验进行故障排查、装置精维修等环节，由 AI 技术对非结构化数据进行数字化记录。

2.试点团队建设

组建由企业领导层、业务部门、技术部门和专业服务商共同组成的项目制数据治理团队。企业领导层主要承担顶层设计职责，通常为生产副总、质量总监或工厂厂长等中高层管理者担任，核心职责在于明确 AI 应用场景与目标，并推

动企业内部组织协同与资源协调。**业务部门主要承担人工智能应用中所需理解的业务含义与规则**，通常为来自生产、质量、设备等一线部门的业务骨干，需要将工艺逻辑、质量标准、设备故障模式等业务知识，转化为 AI 可理解的结构化规则，解决“数据代表什么”的核心问题。**技术部门主要承担企业内部数据系统对接与技术保障**，通常为 IT 部门、网管部门或数字化团队骨干，配合专业服务商完成系统集成与技术调优，确保数据治理方案与企业现有 IT 架构兼容。**专业服务商主要承担数据治理工程化交付**，通常为具备成熟数据治理方法论与工具的方案提供商，需根据企业业务与技术部门的需求，提供定制化数据治理产品，负责数据治理全流程的工程化实施。

3.数据资产盘点

（1）明确数据需求范围

界定场景主数据[9]，明确 AI 场景下的核心业务实体，梳理场景内的核心主数据对象，明确所需时序传感器数据、图像数据、结构化业务数据等数据类型，如设备运行监控与维护场景应包含设备 ID 和故障事件等，需要用到机床振动/温度/电流时序数据、设备故障工单数据、维修记录数据等。**界定场景下所需的模型指标**，模型指标主要为对 AI 模型所需的数据类型、格式、精度、更新频率、标签方式等属性进行拆解，如在线智能检测需要规定工业图像的分辨率、标注精度等属性，设备运行监控与维护需要规定传感器时序数据采

样频率规则等。**界定数据时间维度和空间粒度**，空间粒度需明确数据所需采集范围，明确是设备级、产线级还是工厂级，时间粒度需明确数据更新频率为实时级、准实时、离线等类型。

（2）标准化元数据

聚焦 AI 场景内的核心数据项，对不同部门针对同一指标产生多种叫法的问题进行定义，制定统一的标识符规则，如采用“数据类型/指标缩写（图像、字符、浮点等）-业务对象/设备编号”等通用格式，实现跨部门、跨系统的数据语言统一，避免数据混淆。

（3）梳理场景数据采集现状

对场景所涉及的物理设备数据、数字化系统数据、非格式化数据、外部数据进行梳理，记录可采集数据类型、采集节点、采集频率、字段名称、通信协议、传输方式、存储位置与访问权限等特征。

（4）梳理数据权属与链路

盘点数据来源、数据流转链路、数据所有者、数据使用者等关键信息，梳理数据血缘[10]关系和权责归属，解决 AI 数据使用的合规性和责任问题。

4.质量标准制定

（1）制定数据字典

由治理团队共同**确定数据字典**[11]**的核心字段框架**，应包含数据项编号、数据项名称、数据分类、数据类型等基础属

性，业务含义、业务规则、关联业务指标等业务属性，数据来源、采集频率、数据格式、存储方式、传输方式等技术属性，以及质量要求和 AI 用途等属性。**制定数据字典的基础规则**，采取统一的数据命名、数据分类和质量标准规则，确保每项数据指标与业务对象、业务环节相对应。**建立版本更新机制**，每完成一个数据项的盘点，同步更新数据字典，避免信息遗漏。

（2）制定数据清洗质量标准

明确试点场景的业务目标及其核心数据资产要求，基于前期数据诊断，梳理试点场景核心的数据清洗问题，并定义核心数据问题的具体清洗方法，明确“识别规则-处理方法-优先级”，避免清洗操作的随意性。针对清洗后的数据，锚定 AI 模型需求，制定清洗后数据需达到的可量化质量指标阈值。

（3）制定数据整合质量标准

以试点场景业务链路为牵引，以统一主数据标识为关联基础，明确多源异构数据的关联对齐规则、异构转换规则、数据集构建规则，同时定义关联完整性、格式一致性、特征有效性的量化质量阈值，确保整合后的数据形成逻辑连贯、格式统一、可直接支撑 AI 模型训练的完整数据集。

（4）制定数据标注标准

治理团队应根据日常生产制造和质量管控经验，明确 AI 模型在标注准确率、漏检率等指标的要求，对应标注数据类

型，限定标注对象，基于已治理的主数据、元数据，明确标注对象的命名规则。

（5）制定数据安全标准

结合试点场景数据资产盘点结果，识别核心数据泄露或篡改、AI 应用时训练数据投毒、数据治理过程中数据传输导致数据泄露等风险，以业务敏感度与 AI 应用价值为核心维度，对试点场景数据进行核心敏感数据、一般敏感数据、非敏感数据等级别，界定不同保密等级的数据在治理过程中的访问与权限管控标准。

5.数据质量治理

（1）新增数据质量治理

新增数据是指数据字典制定后实时采集的增量数据，处理目标是满足 AI 模型的实时训练需求。**在数据采集环节**，明确新增数据接入方式，根据时序传感器数据、图像数据、结构化业务数据等类型确定标准化接入方式，按照数据资产情况制定接入规则，确保治理数据范围与 AI 场景应用所需训练数据一致。在接收新增数据时，针对不同数据类型配置自动化加工模板，自动校验数据格式是否符合元数据标准，自动匹配新增数据的主数据标识，确保数据归属清晰，按元数据标准自动校验数据的核心属性。**在数据清洗环节**，应消除数据中的错误、不完整、不一致和重复等问题，确保数据符合数据字典的质量标准。对缺失、错误数据，调用事前配置的自动化加工模板，自动去重、补全缺失值、过滤异常值。

对不同类型的数据，利用标识解析体系[12]等手段实现多源异构数据的源头标准化，赋予每项物理量唯一、可追溯、跨系统识别的“数字身份证”。**在数据整合环节**，通过主数据关联将多源数据整合为统一数据集。**在数据标注环节**，采用人工标注、人机协同、自动化标注[13]等多种标注模式对数据进行标注，并对重点数据进行抽样核验，确保供给 AI 应用的语料质量。针对样本量不足的场景，自动对图像数据进行旋转、裁剪、亮度调整，对时序数据进行加噪声、时间拉伸，生成新增样本。

（2）历史数据质量治理

历史数据是指数据字典制定前已采集的存量数据以及未数字化的纸质数据，处理目标是满足 AI 模型训练的样本需求。在数据清洗与标注方面，需投入大量的人力、时间用于统一格式、处理缺失值或异常值。

6.数据工程建设

（1）设计适配 AI 场景数据特性的硬件支撑架构

单一场景无需部署大型集群，优先边缘-核心两级架构，采用边缘侧处理实时、高频数据，核心侧处理批量、复杂数据，平衡实时性和算力成本。**在硬件选型上**，边缘采集层需对接工业传感器，根据所需采集数据规模情况，选取工业级边缘网关[14]或边缘计算节点[15]集群，一方面，应兼容既定工业协议与数据格式，并预留扩展接口；另一方面，可依托节点内置 NPU 算力，实现端侧 AI 模型的实时推理任务。核心

层需保障数据标注、增强的处理效率，应选取具备高速运算能力 CPU/GPU 处理器，以满足结构化与非结构化数据的计算、模型训练与推理需求。**在存储方案选择上，**实时采集数据对写入速度和查询延迟有较高要求，建议采用边缘节点存储模式，该模式需满足 1-2 周的数据保留周期，数据经处理后需及时流转至后续环节或进行清理，避免占用边缘节点存储资源。标注和训练数据应根据结构化数据与非结构化属性分离存储，结构化数据可依托数据库、数据仓库[16]进行存储，并通过唯一数据 ID 建立标注信息与原始数据的关联，非结构化或半结构化数据以对象存储为主要方式，同时需保留数据集版本信息，为模型训练效果的横向对比、训练过程回溯提供支撑。冷数据[17]需按时间周期完成清洗与分类，随后纳入归档存储体系，既实现存储资源的合理分配，也为为后续模型训练的增量数据补充提供基础支撑。**在网络配置上，**选取支持 VLAN 隔离的工业以太网交换机支持数据交换，避免与办公网络造成冲突。

（2）配置适配 AI 场景数据特性的软件环境

边缘层应实现数据实时采集与边缘预处理，支持 OPC UA/MQTT 常见协议采集，以及可配置边缘侧过滤规则，过滤超出元数据取值范围的无效数据。**处理层**开发自动化脚本，适配时序数据的清洗、特征提取，用于批量数据处理。**服务层**制定面向 AI 模型的标准化接口，为 AI 模型提供统一、高效、标准化的数据访问接口，避免模型直接对接存储。

7.治理机制构建

（1）沉淀数据治理机制及规范

沉淀已验证的核心数据资产，按已开展探索的 AI 场景分类，梳理该场景下经验证的 AI 任务所需的数据类型及精度、粒度要求，明确 AI 训练需满足的数据标注精度、质量标准，以及 AI 任务可达成的效果，最终输出对应清单，为项目纵深优化或公司新场景横向拓展提供参考。**沉淀已验证的数据治理流程，**覆盖从场景定义标准、团队构建架构、数据资产盘点、数据质量治理到 AI 应用的全环节，明确各流程的输入输出、责任人、依赖资源及时间周期等，形成标准化 SOP 流程。

（2）构建“数据-模型”闭环机制

明确数据-模型闭环运行规则，在项目数据治理团队内明确数据-模型联动负责人，负责统筹小批量验证、模型反馈、数据优化联动工作，确保闭环流程的高效执行。**数据支撑模型训练方面，**利用 EXCEL 或轻量级知识库工具管理等载体，将场景下不同模型任务与数据类型关联情况、不同数据质量与模型指标或性能情况进行关联，如在线智能检测场景中，图像标准准确率与模型检测准确率之间的关联关系，形成探索阶段的小批量验证、问题定位经验总结。**模型反馈数据优化方面，**基于模型性能和业务降本增效成果筛选有效数据，更新数据质量校验规则，提升数据质量。通过数据湖[18]回流等形式实现有效数据复用，提高数据资源利用效率。

8.治理价值挖掘

（1）明确治理价值对比基线

建立数据“治理前”和“治理后”可量化对比基准，选取 AI 应用的核心指标作为基线，对指标进行数据化标定，如设备运行监控与维护场景中故障预警准确率和预警提前时间为对比指标，分析其在“治理前”和“治理后”的量化差异，确保数据治理价值的客观评判。

（2）核算数据治理直接投入产出

在成本端，从数据治理所支出的人力成本、治理工具的采购/租赁/部署成本、新增数据采集/计算/存储等环节的设备购置/运维成本等角度核算数据治理的总投入。在收益端，将治理后业务生产效率的提高或成本的降低情况转化为经济价值，衡量 AI 模型性能优化后对核心业务降本增效的直接经济价值。

（3）探索数据治理的间接价值

利用数据打造企业信用资产，企业通过特定场景下数据治理工作，对生产运行数据、订单与交付数据、质量数据等进行治理，将分散数据转化为可追溯、不可篡改、第三方可验的高质量数据集，可作为企业核心数据资产提交银行作为授信补充依据，降低企业贷款成本。**推动数据价值变现，**通过“数据可用不可见”交易模式，向供应商、客户或交易平台提供经治理的工艺、质量、设备运行等数据，可获得间接的经济收益。

9.治理绩效评价

（1）明确数据治理绩效评价维度

在数据质量维度，应从完整性、准确性、一致性、时效性、唯一性、有效性六大质量属性进行评价。**完整性**需评估目标数据源的实际采集比例、单条数据中关键字段的缺失比例、时间维度的记录缺失等情况。**准确性**需对比传感器数据与人工测量数据的误差、业务数据与业务实际的一致性、校验数据的合理性等情况。**一致性**需检查数据格式、数据单位、数据命名等统一情况。**时效性**需计算数据从产生到采集的延迟、从采集到存储的传输延迟、数据更新频率是否满足要求。**唯一性**需记录数据重复、冗余字段、设备/产品唯一 ID 覆盖率等情况。**有效性**需检查符合业务规则的数据量、与 AI 任务相关的字段占比等情况。

在业务价值维度，应验证治理后的数据是否能支撑试点 AI 场景的业务目标，可从数据覆盖度、特征可用性、标签有效性三个角度进行评价。**数据覆盖度**应对比业务流程清单，检查是否覆盖试点场景业务全流程数据，统计试点场景中核心实体的相关数据是否完整。**特征可用性**应通过特征工程筛选出对 AI 任务有贡献的特征数量，计算特征对试点场景中目标变量的区分能力。**标签有效性**应统计有标签样本占总样本的比例，并对比由多标注人员标注的数据标签是否一致，验证标签是否与业务定义的目标一致。

在治理效率维度，应评估数据治理工作的投入产出比，可通过量化时间成本、人力成本、资源成本进行评价。**时间效率**应统计数据治理周期总时长，对比计算治理周期，计算工期偏差率。**资源效率**应统计参与治理的人力投入(人天)、治理过程中软硬件工具的采购或使用成本、数据治理过程中服务器算力、存储的占用率情况。**复用效率**应统计试点场景数据治理形成的清洗规则、标注规范可复用其他场景的比例，评估治理后的数据资产可支撑其他 AI 任务的程度。

在安全合规维度，应确保数据治理过程服务企业内部规范与外部法律法规，可从数据安全、合规性等角度进行评价。**数据安全**应重点统计敏感数据的脱敏比例、核查数据访问权限是否符合最小权限原则、记录治理过程中是否发生数据泄露。**合规性**应验证数据采集、留存等环节是否符合规范。

（2）实施数据治理绩效评价

结合数据治理绩效评价维度，依托数据治理平台的质量监控模块或自定义脚本实现指标自动化计算等方式，搭建治理评价工具，开展数据治理的多维度量化评估。针对数据质量、治理效率类量化指标，通过治理评价工具完成实时自动化计算，并不定期随机抽取一定比例数据集开展人工复核的抽样验证；针对安全合规、业务适配性中的定性指标，专项实施人工核查与打分。将治理后的数据应用于试点场景 AI 模型训练，通过对比模型实际指标与预期目标的差异进行效果

验证，若模型指标未达标，则反向追溯数据治理短板问题，同步启动二次治理工作。

（3）评价结果应用与持续优化

根据评价结果，判断是否可进入AI模型训练阶段，若不达标，需明确二次治理的优先级与方向。其中，针对治理效率低、质量差的环节，优化治理流程，并针对性地开展数据标注人员的业务知识培训、IT人员的工具使用等培训。建立数据质量监控机制，对治理后的数据进行持续跟踪，确保数据质量在AI模型全生命周期内稳定达标。

（二）第二阶段：多部门（多场景）推广实施

围绕治理范围、组织架构、标准体系、数据资产、治理能力、运行机制与价值度量等方面，推动制造业从单点AI场景向全域数据治理转型。治理范围方面，锚定“横向复制、纵向深化、生态协同”为策略，推动成熟治理模式在同类业务场景的推广与沿价值链的延伸。组织架构方面，推动从“项目型团队”向“平台型组织”转型。标准体系建设恪守“业务驱动、全域统一”原则，通过强化顶层设计、横向扩充、纵向深化与全生命周期管理，构建完整标准架构。数据资产聚焦从“数据原料”清单向“数据资产”地图升级。治理能力着力打造统一技术底座与标准化流程。运行机制迭代重点推动重心从管控向赋能、导向从合规向经营、模式从人工向智能、范围从部门向生态的四大转变。最后，构建覆盖治理过程、AI支撑与业务成效的多维度价值度量体系，并建立“评

估-分析-优化-验证”的持续改进闭环，确保治理工作能动态优化并持续释放数据价值。

1.AI 场景拓展

（1）横向复制推广

将前期试点验证成熟的数据治理流程及 AI 技术体系，全面复用至同类业务场景。以“设备运行监控与维护”试点场景为例，可将经实践检验的技术栈、数据治理流程及运维模式，从试点产线单台设备推广至全厂同类型设备，无需重构数据管道与模型框架，仅需新增数据接入节点，通过新设备数据对模型进行微调即可快速落地，有效扩大治理成效，形成规模效应。同时，依托该试点场景在 AI 数据治理中构建的时序数据处理、异常检测及预测等核心能力，将成熟治理模式复制应用于其他具备同类数据特征、需解决同类问题的业务领域，重点覆盖能源智能管控、安全一体化管控、污染在线管控等场景。

（2）纵向深化拓展

以前期试点场景为切入点，沿生产制造价值链条向上游研发、工艺环节延伸，向下游质量管控、供应链管理环节拓展，建立跨生产、质量、工艺等部门的数据协同治理机制，打通全链条数据链路，实现跨工序、跨系统的数据贯通与协同联动。以“设备运行监控与维护”试点场景为例，重点向在线智能检测、质量精准追溯、质量分析改进、工艺动态优化及先进过程控制等场景深化推广数据治理成果。

2.组织架构扩展

本阶段推动组织架构从“项目型团队”向“平台型组织”系统性转型。组织架构扩展严格遵循“业务驱动、权责清晰、梯度推进”原则，构建包含战略决策层、核心管控层、业务落地层、支撑保障层的企业级数据治理组织架构，设立数据治理委员会及数据治理办公室，其中数据治理办公室下设数据治理专职团队、跨业务治理团队、职能协同团队。

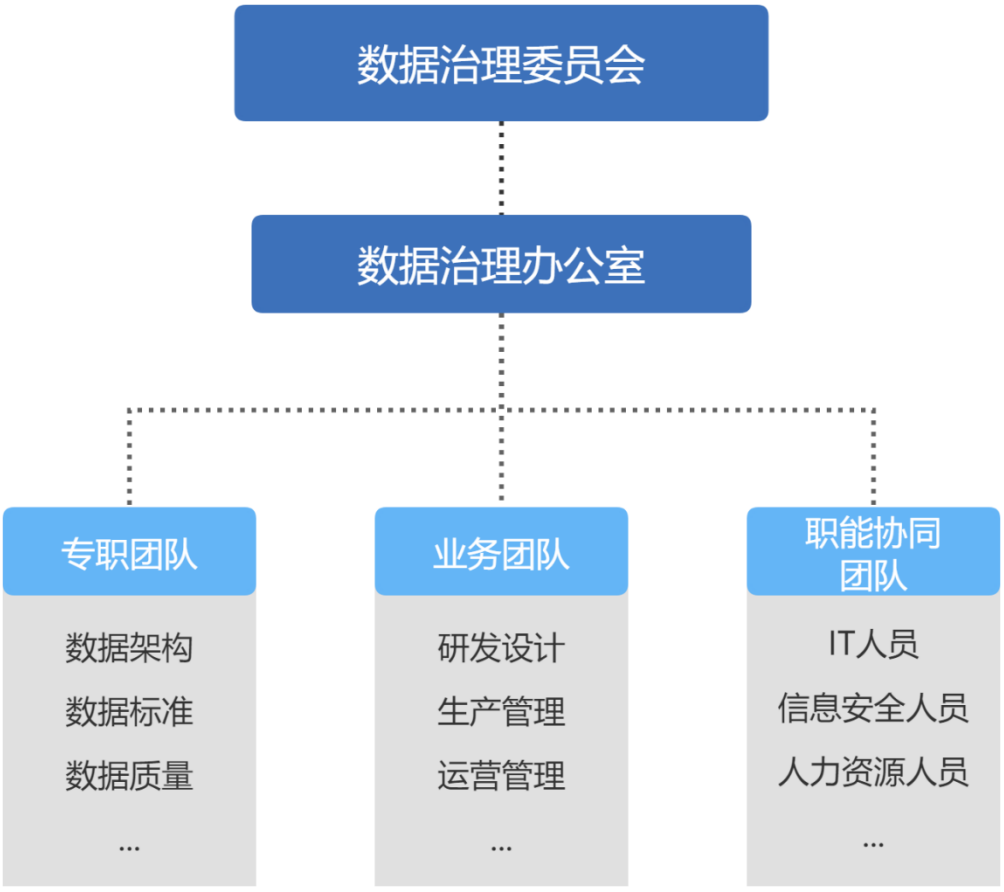

图 4　企业级数据治理组织架构图

（1）战略决策层：数据治理委员会

战略决策层由公司高层管理人员组成，包括首席执行官/首席运营官、各业务线负责人、首席技术官/首席信息官/首席数据官等。主要职责为统筹全域数据治理战略规划，审批年度目标与预算；协调跨业务线资源，解决数据壁垒等核心难题；审批、推动并仲裁数据治理相关标准，审核数据治理价值成果，推动全员形成数据治理共识。

（2）核心管控层：数据治理专职团队

核心管控层在单场景试点团队基础上扩展组建，新增数据架构师、数据标准专员、AI 模型专家等核心岗位。主要职责为提炼单场景数据治理标准（含采集规范、标注规则、模型模板等），编制形成企业级通用手册；牵头制定企业级数据标准，明确数据口径、编码规则、接口规范等，对接各业务线推动落地；搭建并持续优化企业级数据中台，支撑多场景数据整合与共享；建立全域数据质量考核指标体系，开展定期审计工作。

（3）业务落地层：数据治理业务团队

围绕拓展的 AI 业务场景，遴选研发设计、生产管理、运营管理、产品服务及供应链管理等相关业务部门人员纳入团队。主要职责为精准转化业务痛点为数据治理具体需求；按照统一标准推进本业务线数据采集、清洗、标注等工作，对接企业级数据中台；联合数据治理专职团队验证本业务线数据治理成果，及时反馈优化建议。

（4）支撑保障层：职能协同团队

团队由 IT 人员、信息安全人员、人力资源人员等组成。其中，IT 人员负责系统对接、数据链路打通及算力资源保障；信息安全人员负责制定数据安全规范，规避合规风险；人力资源人员负责组织开展全员数据治理培训，建立健全考核激励机制。

3.标准体系建设

（1）强化顶层设计

明确数据治理委员会及数据治理办公室为数据标准的审批、推动及仲裁主体。构建面向 AI 的数据标准架构，明确标准核心内容，完整框架涵盖以下类型，具体见下表。

表 23　数据标准架构内容

类型	主要内容
基础标准	主数据、业务术语、数据模型
专项标准	数据质量、数据安全、数据服务
技术标准	数据交换、数据存储、元数据
管理标准	标准本身的制定、发布、维护流程

（2）推进横向扩充

以 AI 试点场景为基础，将经验证的数据规则抽象泛化为公司级标准。深入分析试点场景“数据-业务”映射规则，剥离具体模型依赖，识别通用业务实体、状态及核心指标并固化为公司级数据标准，实现业务问题向 AI 可理解数据的

转化。以试点场景为核心，横向延伸至紧密关联业务领域，扩大标准覆盖范围。

（3）深化纵向体系

聚焦技术与管理两大维度，构建专项标准体系。在 AI 模型数据标准方面，系统总结试点场景数据标注规范、训练数据集质量要求及特征工程准则，提炼形成公司级 AI 数据准备标准，为各业务场景数据治理提供统一指引。在数据质量与安全标准方面，明确各业务域数据质量度量规则及安全分级细则，夯实数据资产目录规范化运营基础。

（4）健全全生命周期管理

制定《数据标准发布与遵从管理流程》，明确所有新建 IT 项目必须通过数据标准符合性审核，从源头保障标准落地执行。建立《数据标准异议与修订管理流程》，畅通业务部门反馈渠道，允许业务部门在标准应用过程中提出修订建议，确保标准与业务实践深度融合、动态适配。强化平台固化与迭代优化，将已发布标准嵌入数据开发平台、数据资产目录及 AI 开发平台，实现标准应用自动化、规范化。建立定期评审机制，由数据治理办公室牵头，结合业务发展变化及新技术应用情况，对数据标准进行动态修订完善。

4.数据资产盘点

（1）明确盘点目标

立足企业数据资产现状，围绕全域 AI 场景拓展需求，全面识别支撑 AI 场景的高价值数据资产，厘清数据缺口，制定数据资产盘点目标，统筹规划所需资源。

（2）明确盘点范围及内容

依据盘点目标确定盘点范围及核心内容。数据资产盘点范围涵盖组织范围、业务范围、系统范围三大维度，具体内容详见下表。

表 24　数据资产盘点范围

类型	主要内容
组织范围	相关组织及部门，包括集团本部、集团及分子公司等
业务范围	各类业务数据，涵盖生产业务、采购业务、营销业务、财务业务、人力资源业务等
系统范围	各应用系统数据，包含 ERP 系统、MES 系统、SCM 系统、CRM 系统、HR 系统等

按数据来源与属性分类梳理盘点内容，明确不同类型数据的盘点侧重，具体内容见下表。

表 25　数据资产盘点内容

类型	主要内容
基础数据	盘点数据分布的 IT 系统，区分跨系统流转共享且变化缓慢的主数据信息，以及与 IT 系统定位匹配的业务流程交易信息
衍生数据	盘点数据应用场景（如监管、统计、内部管理等），按应用场景分类衍生数据，同步梳理基础数据使用热度

外部数据	盘点外部数据需求、数据类型、数据来源、采集频率、获取成本、数据质量及数据价值评估方式等

（3）盘点前期准备

明确盘点模板。结合通用数据标准与AI专项评估标准，制定统一数据梳理模板，明确数据资产标准项，全面覆盖数据基础属性、管理属性、业务属性、AI相关属性等核心要素，保障盘点信息规范统一。

开展培训宣贯。组织开展全员培训，重点解读盘点工作目标、范围、内容、标准及模板使用方法，确保相关人员统一认知、熟练掌握盘点操作规范，为保障盘点工作质量筑牢基础。

配齐保障资源。依托数据治理委员会及数据治理办公室，明确统筹层、执行层、支撑层职责分工；配备元数据管理工具等技术支撑手段；落实专项工作经费，为盘点工作顺利推进提供全方位保障。

（4）实施盘点工作

整体摸查盘点。按照“系统-数据库表-数据字段”层级开展全维度摸查，全面掌握数据整体情况。摸查过程中同步补充完善数据基础属性、管理属性、业务属性等信息，依托元数据管理工具开展数据采集摸查，实现属性信息系统化补充完善。

元数据采集补充。全面采集盘点范围内元数据，快速识别各类数据资产并开展筛选剔除工作。组织业务人员补充完

善数据资产元数据属性，建立完整元数据档案，为后续资产编目奠定基础。

AI 价值与就绪度评估。组建由 AI 模型专家、数据治理人员及业务专家构成的联合评估小组，开展 AI 价值与就绪度评估，为每条数据资产打上关键标签，明确三方面核心信息：现有数据可支撑的现有及潜在 AI 场景；数据质量是否满足对应 AI 场景训练要求；数据形态是否适配 AI 模型应用需求。

构建资产目录。整合盘点与评估结果，构建企业级数据资产目录，目录需完整涵盖每条数据的业务信息、技术信息、AI 就绪度信息及已支撑的 AI 应用情况。建立“数据-场景”关联视图，实现数据资产与 AI 应用场景精准匹配。

（5）资产化运营与长效管理

建立常态化盘点机制。制定定期复盘制度，每季度开展局部盘点、每年开展全域复盘，确保数据资产目录与数据产生、数据质量变化、新 AI 场景上线等情况同步更新，保障数据资产信息的实时性与准确性。

强化资产动态管理。建立数据资产动态更新机制，明确数据资产新增、变更、注销的管理流程，实时跟踪数据资产状态变化。加强数据资产价值监控，持续优化数据资产配置，提升数据资产对 AI 创新应用的支撑效能。

5.治理能力升级

（1）数据集成

核心目标是实现全业务域数据“统一接入、一数一源”，破除 OT[19]/IT[20]数据壁垒。重点推进三方面工作：一是制定统一数据接入规范，明确数据采集频率、数据格式、字段命名规则等核心要求；二是构建制造数据统一总线，依托工业互联网平台或数据中台架构搭建企业级 AI 数据底座，提供数据接入、存储、清洗、特征管理、模型部署全流程统一入口，支撑 OT/IT 数据融合及多场景模型训练与推理，实现 ERP、MES、WMS、PLC/SCADA、IoT 传感器等多源异构数据统一接入；三是建立数据血缘追溯体系，完整记录数据从采集到集成的全链路信息，明确各字段来源及转换规则，为问题排查与合规审计提供支撑，最终输出全企业统一数据视图，保障跨业务域 AI 模型高效应用。

（2）数据预处理

核心目标是解决制造业数据“噪声多、缺失率高、格式乱”的痛点，输出高质量训练数据。一是推进预处理流程全企业标准化，建立预处理规则库，沉淀试点场景成熟规则并实现全业务域复用，降低重复开发成本；二是实施分级预处理，基础预处理（如格式转换、去重）由数据底座自动完成，业务规则预处理（如工艺、质量标准适配）由业务部门与 AI 团队协同制定，AI 适配预处理（如归一化、标准化）根据具体模型需求动态调整；三是强化质量管控，输出数据质量报告，量化评估数据完整性、准确性、一致性，对不达标数据回流至集成环节优化完善。

（3）特征工程

核心目标是从原始数据中提取业务强相关特征，提升模型精度与泛化能力，实现特征资产化[21]。一方面，推进全流程特征开发，系统开展特征提取、筛选与转化工作：特征提取环节，结构化数据基于制造业业务逻辑提取，非结构化数据（质检图片、运维文本、音频数据）分别提取纹理、边缘特征、故障关键词频次、频谱特征，时序数据提取滑动窗口统计特征、趋势特征；特征筛选环节，采用方差分析、互信息法、模型重要性排序等方法，去除冗余、低相关性特征；特征转化环节，通过归一化、标准化、离散化等处理，适配不同模型输入要求。另一方面，建设企业级特征库，搭建特征管理平台，统一存储特征定义、计算逻辑、关联场景，支持特征检索、复用与版本管理；按业务域分类管理特征，标注特征适用场景、模型类型、精度贡献度，实现新场景模型开发时直接调用已有特征，缩短开发周期。

（4）数据标注

核心目标是为监督学习提供精准标签，破解制造业标注成本高、歧义多的难题。一是优化全企业标注体系，搭建协同标注平台，支持图片、文本、时序数据等多类型标注，具备标注任务分配、进度跟踪、质量审核等功能；二是制定统一标注规范，明确标签体系、标注流程及歧义处理机制，杜绝不同产线标注标准不一致问题；三是提升标注效率，对相

似场景复用标注模型，对简单场景采用弱监督学习技术，减少人工标注工作量。

（5）数据增强

核心目标是通过数据扩充提升模型泛化能力，重点适配故障样本少、缺陷数据稀缺的场景。针对不同类型数据实施差异化增强策略：图像数据可模拟不同光照、油污、角度等场景；时序数据采用添加噪声、时间扭曲、缩放等算法扩充样本，重点服务小样本缺陷场景。强化增强数据质量管控，确保数据符合业务合理性；实施抽样验证，将增强数据与真实数据混合训练，对比模型精度，避免引入无效噪声。

（6）数据划分

核心目标是合理划分数据集，规避模型过拟合，保障模型在全企业场景的泛化能力。遵循三大核心原则：一是时序优先原则，考虑制造业设备、生产数据的强时间关联性，禁止随机划分，按时间顺序划分以模拟真实生产时序预测场景；二是分层抽样原则，针对故障样本占比＜5%等不平衡数据，采用分层抽样确保训练、验证、测试集中目标样本占比一致；三是跨场景迁移原则，若模型需跨产线复用，训练集纳入多产线数据，测试集包含未参与训练的新产线数据，验证模型迁移能力。明确划分标准，通用比例为训练集 70%、验证集 15%、测试集 15%，小样本场景采用 5 折/10 折交叉验证；划分后的数据集需标注数据来源、时间范围、样本分布，存入数据底座数据集管理模块，供模型训练调用。

（7）模型训练

核心目标是将单一场景训练经验升级为全企业通用训练流程，降低模型开发成本。一方面，构建企业级规模化训练体系：标准化训练策略，规范模型选型（根据场景适配选择，杜绝盲目追求复杂模型）；复用迁移学习技术，以单一场景预训练模型为基础，通过其他产线少量数据微调快速适配新场景，缩短训练时间；实施分布式训练，针对全企业级海量数据，采用 TensorFlow[22]/PyTorch[23]分布式训练框架，依托云端 GPU 集群提升训练效率。另一方面，强化训练过程管理：建立训练参数库，记录各模型超参数、训练时长、精度指标，沉淀参数优化经验；实时监控训练过程，跟踪损失函数曲线、精度变化，及时发现过拟合问题，通过早停、正则化等方式优化。

（8）模型验证

核心目标是从技术、业务双维度验证模型有效性，确保模型在不同产线、工况下稳定可用。构建多维度验证体系，具体内容详见下表。验证通过的模型输出验证报告，进入部署推理环节；验证未通过的模型返回特征工程或训练环节优化完善。

表 26　模型验证维度示例

验证维度	核心指标	验证方法
技术精度验证	分类场景的准确率、召回率、F1 值；回归场景的 MAE、RMSE	基于测试集验证，对比不同产线、工况下的精度表现

泛化能力验证	跨产线精度衰减率、新工况适应性	用未参与训练的新产线、新工况数据测试，定义合格衰减率的阈值
业务符合性验证	预测结果与工艺规则一致性	联合生产、设备部门审核模型输出
效率验证	推理时延、资源占用率	测试模型在边缘云端设备的运行速度，满足生产实时性要求

（9）模型推理

核心目标是将模型部署至生产现场支撑业务决策，建立全生命周期迭代机制。一是搭建制造业云边端协同推理架构：云端推理针对全企业级全局优化场景，模型部署于企业云端数据中心，处理海量业务数据并输出全局决策；边缘推理针对生产现场实时性场景，模型部署于边缘服务器、工控机，直接处理设备数据实现毫秒级推理，降低网络延迟；端侧推理针对小型设备，将模型轻量化后部署实现离线推理。二是建立持续监控与反馈闭环：实时监控模型推理准确率、时延、资源占用情况，精度下降至阈值时自动预警；基于生产现场新数据定期开展增量训练，更新模型版本；构建模型仓库，记录模型部署时间、适用场景、迭代历史，淘汰过时模型，保障全企业模型库有效性。

6.机制优化迭代

（1）推动治理重心从数据管控向数据赋能转变

聚焦数据价值释放，着力实现从内部“管好数据”向对外“用好数据”、驱动业务发展的转型。建立健全数据产品与服务体系，将高价值数据封装为标准化数据产品或 API[24] 服务，支撑业务及人工智能系统敏捷调用。构建业务赋能直

达通道，组建数据治理与业务部门联合创新团队，推动治理成果深度嵌入智能排产、质量预测等核心业务流程，切实发挥数据驱动效能。

（2）推动治理导向从被动合规向主动经营转变

将数据作为企业核心资产统筹推进经营运营，积极探索价值变现路径。建立数据资产化运营体系，包括数据资产盘点、估值及运营规划等工作，推动数据资产纳入企业资产负债表管理范畴。在严守数据安全与合规底线前提下，探索数据价值外溢实现路径，依托行业平台、供应链协同等模式，构建数据价值生态化交换与增值机制。

（3）推动治理模式从人工治理向智能治理转变

运用人工智能技术赋能数据治理工作，实现治理效率与质量的质变提升。引入自动化、智能化治理工具，应用机器学习技术开展数据质量探查、敏感信息识别，依托知识图谱[25]实现数据自动关联与映射。构建治理决策智能体[26]，研发基于规则引擎与历史经验的人工智能辅助系统，为数据问题提供自动化修复建议及策略推荐，提升治理决策的科学性与精准性。

（4）推动治理范围从部门级协同向生态级协同转变

打破企业内部数据壁垒，着力构建产业链协同的数据治理体系。深化跨域协同治理机制，优化完善覆盖研发、生产、供应链、营销等全链条的横纵一体化治理组织架构。探索构建产业链协同治理模式，联合核心上下游伙伴共同制定并遵

守质量数据、物流数据等关键数据的交换标准，建立数据安全与互信保障机制，实现产业链数据资源高效协同利用。

7.价值度量与持续改进

（1）构建科学完备的价值度量体系

以“数据赋能 AI、AI 驱动业务”为核心导向，建立多维度、可量化、易操作的价值度量指标体系，覆盖治理过程、AI 支撑、业务成效三大层面，实现对数据治理价值的全周期、立体化评估，指标体系可参考下表。

强化度量体系落地执行：一是建立指标基线，全面梳理现有数据治理及 AI 应用现状，采集各指标当前数据，确定度量基准值，明确各阶段提升目标；二是常态化数据采集，依托数据治理平台、AI 管理平台、业务信息系统等载体，自动采集度量指标数据，每月完成一次数据汇总核对；三是多维度分析评估，每月开展指标对比分析，每季度形成价值度量报告，精准识别成效亮点与差距短板；四是推动结果应用落地，将度量结果与部门 KPI、绩效激励挂钩，作为治理优化、资源调配、策略调整的核心依据。

表 27　价值度量指标体系示例

指标类型	指标名称	指标内容
治理过程效能指标	数据标准落地率	符合企业 AI 数据专项标准的数据占比
	数据质量合格率	涵盖数据完整性、准确性、一致性、及时性等维度，重点统计 AI 模型训练、推理所用数据的合格占比
	治理自动化率	通过智能工具完成的数据质量探查、异常修复、元数据标注等工作占比
	数据资产化率	完成盘点、估值并纳入资产台账的数据资源占比

AI 场景支撑指标	数据供给时效	从业务需求提出到数据资源到位并满足 AI 应用要求的周期
	数据复用率	不同 AI 场景共享使用同一数据资产的比例
	AI 模型性能提升度	对比治理优化前后，同一 AI 模型的预测准确率、召回率、泛化能力等核心指标变化
	模型迭代效率	依托治理优化后的数据资源，AI 模型从训练、验证到部署上线的周期缩短比例
业务价值成效指标	生产运营效率提升	如通过智能排产 AI 模型优化，生产计划达成率提升幅度；通过设备健康管理 AI 模型，设备停机时间减少比例等
	产品质量优化	通过质量预测与检测 AI 模型，产品不良率下降幅度；返工成本降低金额及比例等
	成本管控成效	数据治理相关投入与通过 AI 赋能实现的成本节约的投入产出比
	市场响应能力提升	依托 AI 驱动的供应链优化、客户需求洞察等，新产品上市周期缩短比例或订单交付及时率提升幅度

（2）建立闭环高效的持续改进机制

以价值度量结果为导向，构建“评估-分析-优化-验证”的全流程持续改进闭环，推动数据治理能力动态升级，持续适配 AI 与业务发展需求。

常态化评估诊断。定期复盘评审，每月召开数据治理成效复盘会，由数据治理委员会牵头，数据治理办公室参与，基于价值度量报告，分析治理工作存在的问题。场景化问题排查，针对 AI 应用效果不佳、业务反馈数据支撑不足的场景，开展专项诊断，排查数据质量、标准适配、供给时效等核心问题，形成问题清单。行业对标分析，每半年开展一次行业标杆调研，对比先进企业数据治理与价值转化成效，识别自身差距，明确改进方向。

精准化优化实施。排序问题优先级，结合问题影响范围、解决难度、投入成本，对问题清单进行优先级排序，优先解决高价值、易落地的问题。制定靶向性优化措施，针对数据质量问题，优化校验规则，升级数据清洗工具，补充数据采集校验环节，强化源头管控；针对标准适配问题，修订完善AI场景专项数据标准，推动标准嵌入数据采集、标注、处理全流程，强化强制落地；针对供给时效问题，优化数据资产服务化体系，扩大预制数据产品覆盖范围，提升API调用效率，完善边缘-云协同治理模式，保障实时性场景数据供给；针对价值转化不足问题，聚焦高价值业务场景，深化数据资产与 AI 模型的融合应用，推动价值落地。小步快跑试点验证，对优化措施先在1-2个典型AI场景开展试点应用，跟踪度量指标变化，验证优化成效，总结可复制的经验做法。

动态化迭代升级。体系动态更新，结合业务发展、AI技术迭代、合规要求变化，每季度修订完善价值度量指标体系、数据治理标准与流程。能力持续提升，基于优化实践经验编制《AI数据治理最佳实践手册》，推广成熟做法，加强人才培养，提升跨部门团队数据治理与AI应用融合能力。

（三）第三阶段：产业链（供应链）上下游数据治理

以“统一规则筑基、全流程管控保安全、场景化应用释价值、生态共建聚合力、闭环优化促长效”为核心框架，系统规范数据标准、治理流程、安全保障、价值度量等关键环节，为核心制造企业、上下游配套企业等相关主体提供可落

地的 AI 适配型数据治理路径，助力打破数据孤岛、提升协同效能、释放数据要素价值，夯实制造业产业链供应链智能化升级基础。

1.数据治理规则统一

（1）统一数据标准体系

核心制造企业牵头，联合上下游企业、行业协会及第三方机构，依据《中华人民共和国数据安全法》等国家法律法规和行业规范，结合产业链供应链实际业务需求及人工智能建模应用要求，建立全链条统一的数据标准体系。重点明确五大核心标准：**一是数据命名与编码标准，**对原材料、零部件、产品等核心对象制定唯一编码规则，实现全产业链供应链范围内的统一识别，适配 AI 数据检索与关联分析需求；**二是数据格式标准，**规范 Excel、CSV、JSON 等结构化或半结构化数据及图像、传感器等非结构化数据的存储、传输和展示格式，保障跨企业、跨系统数据互通兼容，满足 AI 模型数据输入规范；**三是数据元标准，**明确核心数据元的名称、类型、长度、精度、取值范围等属性，尤其细化 AI 建模关键数据元（如工艺参数、设备运行阈值、质量检测维度）的定义，确保数据定义一致；**四是数据分类分级标准，**按重要程度、敏感程度将数据划分为不同级别，明确差异化管理要求，尤其强化商业秘密、客户敏感信息等高危数据的管控标准；**五是数据标注标准，**针对 AI 训练所需的非结构化数据（如

生产场景图像、设备故障音频），制定统一的标注规范（含标注精度、标签体系、审核流程），保障训练数据质量。

（2）统一治理制度规范

制定覆盖全产业链供应链的统一数据治理制度框架，结合人工智能应用特点明确各主体权利义务、治理流程及问责机制。**核心制度包括：数据治理管理办法，**界定治理范围、组织架构及职责分工，新增 AI 数据治理专项职责条款；**数据共享管理规则，**规范数据共享的申请、审批、传输、使用及反馈流程，明确 AI 模型训练数据共享的特殊权限管理及数据使用边界，建立利益协调机制平衡各方权益；**数据安全管理制度，**明确访问控制、加密脱敏、备份恢复等安全要求，新增 AI 模型参数安全、训练数据泄露防护、算法公平性审查相关规定；**数据质量管理制度，**规定数据质量评估指标、管控流程及改进机制，补充 AI 训练数据标注质量审核、标注人员管理相关条款。同时，**建立制度动态更新机制，**结合政策变化、AI 技术发展和业务需求定期修订完善，确保制度时效性。

2.全流程数据治理实施

（1）数据梳理与资产化盘点

借鉴石油化工行业成熟实践，在全工业领域推行数字化交付[27]，打通工程建设（设计、施工、调试）与生产运维全生命周期数据链路。其核心意义在于实现数据从源头标准化，消除“信息孤岛”，确保全流程数据的一致性、可追溯性，

为 AI 模型提供完整、连续的高质量数据输入，夯实全生命周期数据治理基础。由核心企业牵头，组织上下游企业开展全链条数据梳理，结合人工智能建模需求制定详细梳理方案，明确范围、内容、责任主体及时间节点。梳理内容涵盖数据来源、格式、字段、含义、产生环节、使用主体、关联关系等核心要素，重点梳理 AI 建模所需的关键数据（如设备全生命周期运行数据、产品质量全检测数据、供应链波动历史数据）。基于梳理结果形成《产业链供应链上下游数据资产清单》，明确各类数据的归属主体、管理状态、价值等级及AI适配性。同步开展治理痛点识别，重点排查数据缺失、错误、冗余、格式不统一、共享困难、安全风险及AI适配性问题（如数据标注缺失、关键特征数据不足、数据时序完整性差），形成问题清单并明确整改优先级。

（2）全环节数据质量管控

建立覆盖数据采集、存储、传输、处理、标注、使用、共享、归档、销毁全生命周期的质量管控体系，强化AI适配性质量要求。**采集阶段**规范流程方法，明确责任主体，针对AI所需的传感器数据、图像数据等强化采集设备校准、采集频率标准化管理；自动化采集工具定期校验，人工采集强化培训考核；**标注阶段**严格执行统一标注标准，建立“标注-审核-复核”三级质控机制，确保标注准确率。**清洗阶段**针对性处理缺失、错误、冗余数据，重点补充AI训练数据的噪声过滤、异常值剔除、数据均衡化处理，确保数据标准化。**存储**

阶段选择安全可靠的存储方案，针对 AI 训练所需的海量数据采用分布式存储架构，建立定期备份与恢复测试机制；**使用与共享阶段**建立审核与二次校验机制，重点核查数据与 AI 模型的适配性，保障数据一致性与可用性。同时，构建数据质量评估指标体系，在准确率、完整率、及时性基础上，新增标注准确率、特征完整性、时序一致性等 AI 专项评估指标，利用 AI 驱动的监测工具实现实时监测与定期评估，建立问题反馈、整改跟踪及经验复盘的闭环改进机制。

（3）全链条数据安全保障

落实各企业数据安全主体责任，结合人工智能应用场景建立全链条安全防护体系。**技术层面，**实施精细化访问控制，对 AI 模型训练平台、数据标注系统采用多因素认证、最小权限原则及定期权限审计；对敏感数据及 AI 训练核心数据在传输、存储、使用全环节加密处理，采用符合国家标准的加密算法；针对 AI 模型，实施模型加密、水印嵌入等防护措施，防范模型窃取与篡改；建立数据安全监测系统，整合 AI 异常行为检测算法，实时预警数据泄露、模型滥用等风险，制定应急预案并定期开展演练；对共享数据及 AI 训练用数据实施脱敏处理，对图像、视频等非结构化数据采用模糊化、匿名化处理，去除可识别主体信息。**管理层面，**健全数据安全管理制度体系，新增 AI 数据标注人员保密协议、AI 模型开发与使用安全规范，加强安全培训提升人员 AI 数据安全意识，定期开展安全评估与审计，重点核查 AI 数据使用合

规性及模型算法公平性，及时整改安全隐患，确保数据全生命周期及 AI 应用全流程安全可控。

3.核心场景应用落地

（1）采购供应协同场景

基于治理后适配 AI 应用的高质量数据，搭建 AI 驱动的采购供应数据共享与分析模块，整合采购需求、供应能力、报价、到货验收、供应周期、供应商历史履约、市场波动等核心数据。通过 AI 算法构建供应商精准画像与动态分级模型，结合多维度数据实现供应商风险的提前预测，优化供应商选择与采购策略；基于 AI 时序预测模型，融合市场需求波动、生产计划调整等数据精准预测采购量，实现智能采购，降低库存积压与短缺风险；通过 AI 实时协同算法，共享到货进度与生产需求数据，动态优化供应调度方案，保障生产连续性。

（2）生产制造优化场景

整合生产计划、工单、进度、工艺参数、设备运行、质量检测、能耗、故障历史等全维度数据，构建 AI 驱动的生产优化平台。基于机器学习算法挖掘生产数据与质量数据的深层关联，识别工艺薄弱环节并生成智能优化建议，实现工艺参数自调整；利用 AI 预测性维护模型，通过设备运行数据实时监测设备健康状态，提前预警故障风险并推送维护方案，减少停机时间；结合 AI 能耗优化模型，整合生产负荷、能耗数据及环境参数，智能调度生产任务，降低生产成本；联动

下游销售数据与 AI 需求预测模型，实现柔性生产，快速响应市场需求变化，提升生产精准度。

（3）仓储物流协同场景

打通库存、出入库、仓储位置、运输计划、轨迹、节点、成本、天气、路况等全链条数据，构建 AI 智能仓储物流协同体系。基于 AI 库存优化模型，实时共享上下游企业库存数据，结合销售预测、生产计划智能计算最优库存水平，实现库存协同调配，降低整体库存成本；利用 AI 路径规划算法，融合实时路况、运输成本、货物时效等数据动态优化物流路线，提升运输效率；通过 AI 视觉识别与物联网数据融合，实现仓储货物智能盘点、出入库自动化核验；结合生产进度与销售需求数据，利用 AI 调度算法实现物流资源精准匹配与动态调度，保障货物及时送达，减少破损与延误。

（4）产业链金融创新场景

依托治理后标准化、高可信度的交易、应收账款、应付账款、信用评级、履约历史、经营状况等数据，构建 AI 驱动的产业链金融创新平台。核心企业联合金融机构搭建数据共享平台，基于 AI 信用评估模型，融合多维度数据对中小企业进行精准信用画像，提供无抵押、高效率的信用融资服务，解决融资难问题；利用 AI 风险控制模型，实时监测企业经营数据与交易动态，实现融资风险的提前预警与精准评估，优化融资审批流程，提升融资效率；整合支付结算数据与 AI

智能清算算法，构建便捷高效的产业链支付体系，实现资金流转的自动化、精准化，优化资金流转效率。

4.产业链治理生态建设

（1）搭建协同治理组织架构

核心制造企业牵头成立产业链供应链数据治理工作小组，由高层管理人员担任组长，成员涵盖上下游主要企业负责人及技术、业务、安全、AI 算法等部门骨干，增设 AI 数据治理专项工作组。明确工作小组职责，包括制定适配 AI 应用的治理规划、协调重大问题、推动 AI 相关治理措施落地等。同时，明确各主体责任，**核心企业**承担牵头统筹责任，负责搭建协同平台、制定统一规则，主导 AI 数据治理标准的推广与落地；**上下游企业**落实主体责任，配合开展数据治理与共享，保障提供数据的 AI 适配性；**技术服务机构**提供 AI 数据标注工具、模型监控平台、数据治理自动化工具等技术支撑；**行业协会**发挥桥梁作用，推动 AI 数据治理标准推广、组织跨企业 AI 治理经验交流与培训咨询服务。

（2）构建高效数据共享协同平台

整合上下游企业信息系统资源，搭建集数据接入、存储、处理、标注、查询、分析、安全防护、AI 模型开发与部署于一体的数据共享协同平台。平台采用分布式架构与 AI 原生设计，支持多企业、多系统灵活接入，兼容结构化与非结构化数据处理，保障数据实时更新与高效传输；内置标准化数据标注模块、AI 模型训练与推理引擎，为上下游企业提供低

成本的 AI 应用工具。建立平台运营管理机制，明确数据接入规范、共享权限管理、AI 模型使用权限及运维责任，设立 AI 数据质量与模型效果专项监控模块，确保平台稳定运行与 AI 应用成效。通过平台打破数据壁垒，实现上下游企业数据互联互通，支撑 AI 驱动的全链条协同，提升协同响应能力。

（3）完善生态保障支撑体系

在组织保障方面，各企业健全内部数据治理组织架构，增设 AI 数据治理专项岗位，配备专业团队；核心企业加强跨企业沟通协调，推动形成“AI+数据治理”协同共治格局。**在技术保障方面，**加大技术投入，重点引入 AI 数据标注工具、数据治理自动化平台、AI 模型监控系统、隐私计算工具等，鼓励企业与科研机构合作开发适配产业链需求的 AI 数据治理解决方案，建立覆盖数据治理全流程与 AI 应用全生命周期的技术支撑服务体系。**在人才保障方面，**加强复合型人才培养与引进，重点培育兼具数据管理、数据分析、数据安全、制造业业务知识及 AI 算法能力的专业人才；开展“数据治理+AI 应用”多层次培训，提升相关人员的综合能力；建立人才激励机制，鼓励人才在 AI 数据治理创新中发挥积极作用。**在资金保障方面，**企业合理安排治理经费，重点保障 AI 适配性数据标准建设、共享协同平台 AI 模块开发、AI 工具采购、复合型人才培训等需求。

5.价值度量与持续改进

（1）构建多维可度量的价值度量体系

建立一套覆盖经济、效率、质量、协同与创新多个层面，且可量化、易操作的价值度量指标体系，是客观评估数据治理成效、彰显各方贡献、激发持续参与的关键。通过指标评估，推动各参与主体的绩效评估与决策优化。主要做法包括：**一是多维度设计。**指标体系需超越传统的成本节约视角，涵盖业务价值实现（如收入增长、成本降低）、运营效能提升（如效率、质量改进）、数据资产本身优化（如质量、安全水平）、协同生态效益（如链条响应速度、协同满意度）以及创新孵化能力（如新业务、新模式）五大维度，全面反映数据治理的综合价值。**二是分层级量化。**针对核心企业、上下游配套企业等不同角色，设置共性与特性相结合的指标。指标定义需清晰明确，数据来源可靠，计算方法统一，确保评估结果公平、可比。优先选用比率、百分比、时间单位等量化形式。**三是与绩效关联。**将关键数据治理与价值释放指标纳入企业及相关部门、人员的绩效考核体系。例如，将数据质量达标率、数据共享贡献度、基于数据智能应用产生的效益提升等，与激励机制挂钩，强化内部驱动力。**四是动态可扩展。**指标体系并非一成不变，需随着治理阶段深入、AI应用场景拓展而进行定期复审与动态调整，纳入能反映新发展阶段特征的新指标。

表 28　价值度量指标体系示例

维度	指标类型	指标名称	指标内容	主要关联方
业务	经济效益	采购成本节约率	（治理后周期平均采购成本-治理前基准周期平	核心企业、采购方

维度	指标类型	指标名称	指标内容	主要关联方
价值			均采购成本）/治理前基准周期平均采购成本	
		库存周转提升率	（治理后平均库存周转次数-治理前平均库存周转次数）/治理前平均库存周转次数	核心企业、上下游企业
		融资效率提升度	中小企业基于数据信用获得的平均融资审批时间缩短比例	中小企业、金融机构
	收入贡献	数据服务收入占比	通过数据产品、数据模型服务等获得的新增收入占总收入比例	核心企业、技术服务商
运营效能	效率提升	生产计划调整响应时间	从市场需求变化到生产计划完成调整的平均时间	核心企业
		物流平均延误降低率	（治理前平均延误时间-治理后平均延误时间）/治理前平均延误时间	物流企业、货主方
	质量改进	产品一次检验合格率提升	治理后产品一次检验合格率-治理前基准合格率	核心企业、生产方
		设备预测性维护准确率	AI 预测的故障中，实际发生故障的比例	设备使用方、维护方
数据资产	数据质量	关键数据项质量达标率	达到预设质量标准（完整、准确、及时、一致）的关键数据项数量占比	所有数据提供方
		AI 训练数据标注准确率	抽样检查中，标注正确的数据条目占比	数据标注方、使用方
	数据安全	数据安全事件发生率	单位时间内发生的数据泄露、滥用等安全事件次数	所有数据管理方
	数据应用	高价值数据资产调用频次	单位时间内，被 AI 模型或其他应用调用的核心数据资产次数	数据平台运营方

维度	指标类型	指标名称	指标内容	主要关联方
协同生态	协同效率	供应链订单协同满足率	上下游企业间通过数据平台协同完成订单的比例	核心企业、上下游企业
		跨企业数据共享平均耗时	从数据共享申请到数据可用（且符合安全要求）的平均时间	所有参与企业
	生态满意度	上下游伙伴协同满意度	通过调研获得的上下游企业对数据共享、协同流程的满意度评分	所有参与企业
创新孵化	创新能力	基于数据的新业务/场景数量	在一定周期内，利用治理后数据新开发的 AI 应用场景或商业模式数量	核心企业、创新单元
		数据模型复用与贡献度	某个企业开发的 AI 模型被产业链其他企业复用的次数及带来的效益	模型开发方、使用方

（2）建立闭环高效的持续改进机制

基于度量结果的持续优化是实现治理长效化的引擎，需建立一个从监测评估、到分析改进、再到反馈调整的闭环机制，并嵌入激励与收益分配，形成“度量-改进-提升-再度量”的正向循环。主要做法包括：**一是建立常态化监测与评估流程。**利用数据共享协同平台，对上述价值度量指标进行自动化或半自动化采集、计算与可视化展示。设立定期的联合评估会议，由产业链供应链数据治理工作小组牵头，各方共同审视指标达成情况，分析差距与根因。**二是设计激励相容的改进建议征集与实施机制。面向中小企业，**设立“轻量化”改进建议提交通道，鼓励其就数据标准易用性、共享流程便捷性、平台工具友好性等提出实操性建议。对于被采纳并产

生显著效益的建议，给予提供方荣誉表彰、平台服务费用减免或在收益分配中予以倾斜。**基于价值度量结果，**将评估中发现的问题转化为具体的改进课题，成立跨企业专项小组进行攻关。**三是优化数据共享与收益分配模式。**建立清晰、公平的收益分配机制，激发中小企业持续提供高质量数据、深度参与生态。在贡献度方面，不仅考虑数据提供的“量”（数据体量、种类），更强调“质”（数据质量、AI 适配性、更新频率）和“用”（数据被调用次数、支撑应用的效益大小）的贡献。利用区块链、智能合约等技术可探索贡献度的透明化记录与核算。在收益分配方面，将收益分为直接经济收益（如数据交易分成、模型使用费）、间接成本节约（如通过协同降低库存、物流成本），以及发展权益（如优先获得产业链金融支持、联合研发机会）等部分，分配方案由生态各方协商共识，并写入共享规则。**四是固化知识沉淀与标准迭代。**将改进实践中形成的有效经验和最佳实践，及时沉淀为新的操作指南、标注范例或模型优化参数，并反哺到统一数据标准与治理制度中，通过动态更新机制实现治理规则的持续进化。

四、面向 AI 典型应用场景的数据治理方案

参考江苏省工信厅发布的《江苏省制造业领域人工智能技术应用场景参考指引（2025 年版）》中入门级 9 个场景、基础级 20 个场景、进阶级的 31 个场景。同一场景不同层级的企业，由于数字化成熟度与 AI 应用的需求不同，治理的对象，数据治理的要点，以及平台（技术）工具选型原则有所差异。制造业企业可根据企业的需求与自身条件，选择适合的场景、数据治理对象、平台（技术）工具，高效开展面向 AI 应用的数据治理工作。

（一）工厂数字化规划设计

“工厂数字化规划设计”场景涉及基础级与进阶级的企业。

基础级企业针对工厂新建/改造中的空间冲突、多目标优化难和工艺匹配度低等问题，引入 AI 算法与数字孪生[28]技术，构建动态仿真模型，实现布局智能优化、冲突自动预警与多目标平衡决策，提升规划科学性与前瞻性时，数据治理的目标主要是确保数据与工厂工艺需求、空间属性精准匹配，解决“工艺匹配度低”的源头数据问题，解决多目标优化中“数据碎片化”导致的决策偏差问题，确保数据能够全面覆盖多目标分析维度，支撑动态仿真模型的精准运行，确保数据能够实时、准确反馈工厂规划过程中的动态变化，支撑 AI 算法的实时计算与冲突自动预警功能。

进阶级企业针对复杂工厂全周期效率提升、多目标深度优化的需求,引入深度 AI 自主决策与全要素数字孪生技术，构建融合跨学科知识图谱、生成式设计与实时数据驱动的规划系统时，数据治理的目标主要是实现全要素数据的统一整合与规范，支撑跨学科知识图谱构建与全要素数字孪生体的精准映射，让数据具备“支撑 AI 深度分析与自主决策”的能力，解决多目标深度优化中“数据价值挖掘不足”的问题，确保数据实时驱动与安全合规，支撑规划系统的实时响应能力与全周期风险管控，匹配复杂工厂全周期效率提升的需求。

1.治理对象

表 29　工厂数字化规划设计场景的数据治理对象

适用层级	数据类型	数据内容
基础级	厂区地理信息数据	地形、地质、周边环境等
基础级	建筑信息模型数据	建筑结构、管线布局、空间尺寸等
基础级	生产工艺数据	工艺流程、产能需求、设备参数等
基础级	物流数据	物料运输路径、搬运设备参数、流量需求等
基础级	历史项目规划数据	布局方案、成本结构、运行效果评估等
基础级	仿真模型数据	设备运行模拟、产能推演结果等
进阶级	工厂全周期数据	规划阶段的产能需求、工艺参数、空间布局数据；建设阶段的施工进度、设备安装数据；运维阶段的设备运行状态、能耗数据、故障记录等
进阶级	跨学科专业数据	机械设计参数、电气系统配置、物流路径规划、建筑结构数据、环保标准等
进阶级	同类工厂运营数据	通过联邦学习整合的产能弹性、能耗成本、空间利用率等

进阶级	数字孪生仿真数据	全要素数字孪生体的多物理场仿真结果、全周期状态模拟数据等
进阶级	生成式设计数据	多套布局方案的设备模型、路径规划、管网设计等
进阶级	跨领域知识图谱数据	行业最佳实践、规划—建设—运维关联规则等

2.平台（技术）工具

表 30　工厂数字化规划设计场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	多源异构数据的统一接入、清洗融合、质量管控、资产编目与安全共享，旨在构建全域统一、标准可信的“单一数据源”。	广联达 BIMFace、筒道云、明道云、Microsoft Power Platform、Autodesk BIM 360	华为云 DataArts Studio、阿里云 DataWorks、Informatica Intelligent Data Management Cloud、Collibra
特征工程与样本管理平台	将原始数据转化为可供机器学习算法高效识别的特征，并对训练样本进行全生命周期管理，包括标注、增强、版本控制等，是提升模型效果的关键环节。	基于 Python 开源生态，利用 Pandas,Scikit-learn 等库在 Jupyter Notebook 中进行手动的特征提取与处理、百度数据众包、京东众智、Labelbox	第四范式 FeaturePro、华为云 ModelArts、Tecton、Feast
模型开发与运维平台	为人工智能模型的开发、训练、实验、部署、监控与持续迭代提供一体化环境与工程化管理能力，即实现机器学习运维，确保模型能高效、稳定地产生业务价值。	华为云 ModelArts Lite、百度 BML、腾讯云 TI-ONE、Google Colab、Kaggle Kernels、Databricks Community Edition	华为云 ModelArts、阿里云 PAI、Amazon SageMaker、Microsoft Azure Machine Learning
数字孪生与智能应用平台	构建高保真、可交互、可仿真的虚拟工厂环境，并集成各类 AI 模型服务，以驱动智能化的设计优化、模拟推演与决策支持应用。	优锘科技 ThingJS/51VR、FlexSim、AnyLogic、	华为元图工坊、DataMesh FactVerse、英伟达 Omniverse、达索 3DEXPERIENCE

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过文件接口、数据库连接及 API 调用等方式，统一采集厂区基础地理信息数据、既有或设计的 BIM[29] 数据、核心生产工艺数据、厂内物流规划数据，并关联历史类似项目规划数据。**数据预处理。**对多源异构数据进行清洗、归一化、标准化，修正 BIM 模型中的几何错误、补全工艺参数缺失值、剔除物流数据中的异常记录，统一各类空间尺寸、坐标、物理量的单位与坐标系，与不同来源 BIM 数据的构件编码、属性格式、规范时间与版本标识，形成结构清晰、时空基准一致的基础数据集。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，构建用于智能规划分析的特征集。如“空间占用率与干涉矩阵”“工艺设备关联紧密度指数”“物流路径效率与冲突点密度”“基于历史数据的方案经济性与可行性特征向量”等。**数据标注。**结合历史项目中的设计变更记录、施工问题报告及后期运营反馈，对历史规划方案数据及其对应的运行效果进行标注，标识“布局合理”“存在管线冲突”“物流迂回”“产能瓶颈”等状态标签，为监督学习提供依据。**数据增强与划分。**针对优秀设计方案或特定类型冲突场景样本不足的问题，采用基于规则

的方案变异、参数扰动等方法进行数据增强。按项目类型或时间顺序，将数据集划分为训练集与测试集。

③第三阶段：模型训练与仿真验证

模型训练。使用标注后的数据集，训练适用于规划设计场景的 AI 模型，如用于空间冲突检测的图像识别或图神经网络模型[30]、用于物流路径优化的强化学习[31]模型、用于方案多目标评估的分类与回归模型。**仿真验证。**在测试集上评估模型性能。进一步，将模型与数字孪生仿真环境结合，输入新的规划参数，在仿真环境中模拟“设备安装与维护空间验证”“高峰物流压力测试”“生产工艺流程模拟”等场景，验证模型推荐的布局方案或优化建议在虚拟环境中的可行性与有效性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练验证通过的模型集成到工厂数字化规划设计平台中，作为智能辅助设计模块。设计人员输入规划约束条件，如用地红线、产能目标、投资限额，模型可进行在线推理，输出“推荐布局方案及其评估报告”“高概率冲突预警”“多方案对比与排序”等。**闭环进化。**建立规划-建设-运维数据反馈链路。收集设计方案在后续建设与实际运营中暴露的问题、产生的效能数据，与模型当初的预测和建议进行对比，形成反馈数据包。定期利用新的项目数据和反馈数据对模型进行增量训练与优化，提升模型在实际工程中的适用性与准确性。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建全域全周期数据湖。深度集成工厂全生命周期数据、跨学科专业数据、通过隐私计算技术安全获取的同类工厂运营基准数据、高保真数字孪生产生的多物理场仿真数据、AI 生成式设计引擎产生的海量备选方案细节数据，以及结构化、语义化的跨领域知识图谱数据。**数据预处理。**实施面向深度学习[35]的复杂数据预处理。包括对非结构化设计文档进行信息抽取与向量化，对时空序列数据进行对齐与融合，对多物理场仿真结果进行特征提取，对跨领域知识进行关联与消歧，并确保所有数据在统一的语义框架和生命周期标识下进行组织与管理。

②第二阶段：样本准备与特征工程

特征工程。应用自动特征工程与图表示学习技术。自动挖掘海量数据中深层次、高维度的关联特征，如“全生命周期成本与设计参数的隐式关系网络”“多专业约束耦合的图特征表示”“生成式设计方案的空间-功能-性能联合嵌入向量”。**数据标注。**引入专家系统、知识图谱推理和强化学习环境反馈，对复杂规划决策场景进行自动或半自动标注。例如，对“满足所有约束的帕累托最优[32]方案集”“特定运维目标下的最佳改造策略”等进行标注，用于训练高级决策模型。**数据增强与划分。**利用生成式对抗网络[33]或扩散模型[34]，生成符合物理规律与设计规范的新颖规划方案数据，极大扩

充设计空间。按照全周期阶段、工厂类型、优化目标等进行多维度的数据集划分，确保模型的广泛适用性与泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。训练深度强化学习智能体、生成式设计模型，以及基于 Transformer[36]的跨模态决策模型。使系统能够理解复杂约束，在巨大的设计空间中进行探索和优化，自主生成满足多目标、全周期最优的规划方案，并能进行方案的解释与推演。**仿真验证。**在全要素、高保真数字孪生体构成的“虚拟工厂”中进行沉浸式、全流程的仿真验证。模拟验证规划方案在“长达数十年的运行老化”“市场剧变导致的产线重构”“极端灾害条件下的韧性表现”等长周期、极端场景下的性能，实现对方案鲁棒性与可持续性的深度评估。

④第四阶段：模型部署与闭环进化

模型部署与推理。将高级 AI 模型部署为智能规划系统的核心决策引擎。支持自然语言、草图或参数化输入设计意图，系统可进行实时推理与交互式生成，输出“满足全周期目标的最优方案族及其数字孪生体”“基于实时数据的动态规划调整建议”“跨专业协同设计的自动协调方案”。**闭环进化。**构建“物理工厂-数字孪生体-AI 大脑”的实时交互与协同进化生态。物理工厂的实际运行数据持续驱动数字孪生体演化，同时反馈至 AI 模型；AI 模型不断从新的数据与交互中学习，优化决策策略，并更新知识图谱。形成数据与智能双轮驱动、规划与运营深度融合的持续自我完善体系。

（二）数字基础设施建设

“数字基础设施建设”场景涉及基础级与进阶级的企业。

基础级企业针对工厂算力分配僵化、网络负载波动大、安全防护被动等问题,引入AI驱动的智能管理与优化技术，构建融合算力动态调度、网络自适应调节、安全主动防御的数字基础设施体系,通过多维度AI分析实现资源高效利用、风险提前预警，提升工厂数字化支撑能力时，数据治理的目标主要是构建准确、稳定、标准化的基础设施运行数据体系，为AI实现资源高效利用与风险初步预警提供可信数据输入。

进阶级企业针对工厂高并发算力峰值、全域网络协同、未知威胁防御等进阶挑战，引入深度学习技术，构建一体化智能基础设施，实现算力自优化调度、网络自愈式协同、安全自主防护时，数据治理的目标主要是将静态、割裂的运维数据，转化为驱动基础设施智能体自主感知、协同决策并持续进化的“核心养料”。

1.治理对象

表 31　数字基础设施建设场景的数据治理对象

适用层级	数据类型	数据内容
基础级	服务器集群与边缘计算节点的算力数据	资源占用率、负载情况、分配记录等
基础级	网络设备的运行数据	流量大小、传输速率、负载特征等
基础级	安全防护系统的日志数据	攻击类型、威胁等级、防御记录等
基础级	业务系统的资源需求数据	算力请求、网络带宽需求等

进阶级	算力相关数据	各业务节点算力需求、GPU 集群与边缘节点资源使用状态、算力调度记录等
进阶级	网络数据	网络拓扑结构、切片配置信息、链路状态、数据传输速率与延迟等
进阶级	安全数据	安全事件记录、攻击链信息、身份认证数据、漏洞信息等
进阶级	生产关联数据	生产计划、业务类型、资源需求波动等
基础级、进阶级	数字孪生仿真数据	资源运行模拟结果、扩容升级预演数据、极端场景下的系统表现模拟、防御策略预演结果等
进阶级	跨厂区传输数据	传输内容、加密状态、权限信息等

2.平台（技术）工具

表 32　数字基础设施建设场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现多源异构数据的统一接入、清洗、融合与资产化管理，是构建高质量数据基座的“总枢纽”。	观测云、阿里云日志服务 SLS、华为云应用运维管理 AOM、Elastic Stack、Datadog	华为云 DataArts Studio、阿里云实时计算 Flink 版、Informatica Intelligent Data Management Cloud、Confluent Platform
特征工程与样本管理平台	将原始数据转化为可供机器学习算法高效识别的特征，并对训练样本进行全生命周期管理，包括标注、增强、版本控制等，是提升模型效果的关键环节。	Jupyter Notebook、Label Studio、Jupyter Notebook	第四范式 FeaturePro、百度 PaddlePaddle、Tecton、Weights & Biases
模型开发与运维平台	为人工智能模型的开发、训练、实验、部署、监控与持续迭代提供一体化环境与工程化管理能力，即实现 MLOps（机器学习运维），确保模型能高效、稳定地产生业务价值。	华为云 ModelArts Lite、百度 BML、腾讯云 TI-ONE、Google Colab、Amazon SageMaker Studio Lab	华为云 ModelArts、阿里云 PAI、Amazon SageMaker、Ray

数字孪生与智能应用平台	构建高保真、可交互、可仿真的虚拟工厂环境，并集成各类 AI 模型服务，以驱动智能化的设计优化、模拟推演与决策支持应用。	优锘科技 uThings、神州数码、FlexSim、AnyLogic、	华为云数字孪生平台、DataMesh Director、英伟达 Omniverse、微软 Azure Digital Twins

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过代理、API 等方式，统一采集服务器、网络设备、安全系统的日志与性能数据，并与业务系统的资源需求数据建立初步关联。**数据预处理。**对采集的时序数据进行清洗（处理缺失值、异常峰值）、归一化（统一单位、量纲）和标准化（统一时间戳、日志格式），确保数据的一致性与可比性。

②第二阶段：样本准备与特征工程

特征工程。基于原始数据构建关键性能指标特征，如“CPU/内存综合利用率”“网络带宽使用率趋势”“安全事件频次与等级聚合”。**数据标注。**结合历史告警与故障记录，对历史运行数据段进行标注。**数据增强与划分。**对少数类异常场景数据进行过采样等增强处理，并按时间顺序将数据集划分为训练集与测试集。

③第三阶段：模型训练与仿真验证

模型训练。使用标注好的数据集，训练时序预测模型和分类模型。**仿真验证。**在测试集上评估模型性能，并使用数

字孪生仿真数据，在模拟的“业务高峰”“常见攻击”等场景下验证模型决策的合理性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练好的模型部署为监控系统的插件或独立服务，对实时基础设施数据流进行在线推理，输出“资源扩容建议”“风险预警提示”等。**闭环进化。**收集模型预警与实际运维人员处置结果的反馈数据，定期进行模型的增量训练与调优，提升准确性。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。部署企业级数据总线与流处理平台，实时汇聚 GPU 集群、网络 SDN 控制器、全流量安全探针、生产排程系统及跨厂区数据网关的毫秒级数据流。不仅采集状态数据，更同步集成网络拓扑[37]、算力调度策略、攻击链图谱等配置与关系型数据。**数据预处理。**在流处理引擎[38]中内置业务规则与异常检测模型，对数据进行实时清洗、关联与富化。例如，将突发的算力需求波动与正在执行的生产订单自动关联；将零散的安全事件在流中实时拼接为初步的攻击链。输出的是承载丰富上下文的“实时态势数据流”，而非孤立的指标。

②第二阶段：样本准备与特征工程

特征工程。超越基础指标，构建高阶、关系型特征。例如，构建“业务服务等级协议（SLA）与资源消耗的弹性关

系模型”“网络切片间的干扰与协同特征”“用户行为与异常访问的时序关联图谱”。利用图计算技术，显式建模算力、网络、安全实体间的动态关系。**数据标注。**以高保真数字孪生系统为核心“样本工厂”。通过模拟极端业务压力、复杂网络故障、高级持续性威胁等场景，自动化、规模化生成带有精确决策标签的训练样本。引入主动学习[39]机制，优先对模型不确定的复杂边界场景进行仿真与标注，提升样本效用。

③第三阶段：模型训练与仿真验证

模型训练。采用深度强化学习与多智能体系统框架，分别训练“全局资源调度器”“网络自愈控制器”和“主动防御智能体”。训练重点不仅在于个体智能，更在于智能体间的协同策略学习，例如，学习如何在安全隔离与算力调度间做出动态权衡。**仿真验证。**在全要素数字孪生环境中，构建一个与物理基础设施同步运行的“平行系统[40]”。将训练好的 AI 决策模型在此环境中进行长期、高压的沙盘推演和攻防博弈，验证其在复杂扰动、对抗环境下的决策有效性、鲁棒性与协同稳定性，确保上线前策略的充分收敛。

④第四阶段：模型部署与闭环进化

模型部署与推理。部署集中式决策引擎，集成训练成熟的各领域智能体。引擎接收全域实时态势数据流，进行毫秒级协同推理，直接输出可执行的决策指令，并与自动化运维系统深度集成以实现指令的自动执行。闭环运营。建立系统级“感知-决策-反馈”进化闭环，采用双轨运行机制，AI 决

策引擎在实际生产环境执行决策的同时，其“影子”在数字孪生中并行推演替代方案。系统自动对比不同决策的实际效果与仿真预测，持续收集生产反馈、新型威胁数据以及环境变化。基于此，驱动特征库、样本工厂、模型参数的自动化迭代更新，形成从数据感知到决策模型再优化的完整自进化闭环，使智能基础设施具备持续的适应与学习能力。

（三）数字孪生工厂构建

“数字孪生工厂构建”场景涉及基础级与进阶级的企业。

基础级企业针对工厂数据分散关联弱、数字模型与物理实体联动性差、动态仿真能力不足等问题，引入 AI 驱动的数据融合与动态建模技术，构建具备实时映射、智能关联、基础动态仿真功能的数字孪生工厂，通过多源数据 AI 分析实现模型与物理实体的深度联动，提升工厂全要素可视性与决策支撑能力时，数据治理的目标主要是打破工厂多源数据壁垒，实现数据标准化整合与关联，为 AI 驱动的数据融合提供基础，确保数据能够精准反映物理实体状态，实现数字模型与物理实体的精准映射，解决联动性差的核心问题，解决动态仿真能力不足的数据源问题，确保数据具备实时性与可靠性，支撑数字孪生的实时映射与动态仿真功能。

进阶级企业针对工厂全生命周期协同、复杂问题自主诊断、数字孪生与业务深度耦合的需求，引入生成式 AI 与深度强化学习技术，构建自主进化系统，通过 AI 算法实现全要素数字孪生的自主建模、全流程智能决策与全生命周期持

续优化时，数据治理的目标主要是打破工厂规划、设计、施工、运营、运维各阶段的数据壁垒，保障数据在全生命周期内的连续流转与有效复用，解决全生命周期协同不足的问题，确保数据全面覆盖工厂物理实体、生产流程、业务逻辑等全要素，支撑生成式 AI 完成全要素数字孪生自主建模，实现数字孪生与业务深度耦合，解决数据质量问题导致的 AI 算法决策偏差，确保数据具备完整性、准确性、一致性，支撑深度强化学习算法实现复杂问题自主诊断与精准决策。

1.治理对象

表 33　数字孪生工厂构建场景的数据治理对象

适用层级	数据类型	数据内容
基础级	物理工厂的设备运行数据	参数变化、故障状态、能耗等
基础级	环境感知数据	温湿度、光照、粉尘浓度等
基础级	生产执行数据	产能、工序进度、物料流转等
基础级	资产信息数据	设备型号、位置、属性变更等
基础级	视觉采集数据	设备状态图像、场景影像等
基础级	各业务系统的异构数据	格式、口径不一的跨系统数据
进阶级	工厂全要素基础数据	设备 CAD 图纸、工艺参数、产线布局、建筑结构等
进阶级	全生命周期运营数据	生产计划、订单数据、设备运行状态、能耗数据、质量检测结果、维护记录等
进阶级	多模态感知数据	激光扫描点云、传感器实时监测数据、图像视频数据等
进阶级	知识与案例数据	工艺知识库、故障案例库、专家经验、行业最佳实践等
进阶级	数字孪生模型数据	动态孪生体参数、模型更新记录、虚拟仿真结果等
进阶级	极端场景数据	历史极端工况记录、应急处理方案、模拟仿真数据等

2.平台（技术）工具

表 34　数字孪生工厂构建场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现物理工厂全域异构数据的统一接入、清洗、融合、存储与资产化管理，是构建虚实同步、高质量数字孪生数据基座的“中枢系统”。	观测云、阿里云日志服务 SLS、TDengine、ThingsBoard、Apache Kafka、强思 X-ETFoundation	华为云 DataArts Studio、阿里云实时计算 Flink 版、西门子 Xcelerator 生态系统（MindSphere 等）、AVEVA System Platform
特征工程与样本管理平台	将原始工业时序、图像、文本数据转化为可供 AI 模型识别的结构化特征，并对故障、缺陷等场景的训练样本进行标注、增强、版本化管理，是提升孪生体预测与诊断能力的关键。	Jupyter Notebook、Label Studio、Apache Superset、Pandas/NumPy	第四范式 FeaturePro、华为云 ModelArts、Tecton、Prodigy、Scale AI
模型开发与运维平台	为面向孪生场景的 AI 模型（如预测性维护、视觉检测、工艺优化）提供从开发、训练、仿真验证到部署、监控、持续迭代的 MLOps 全生命周期管理能力，确保模型在虚拟与物理双空间的稳定可靠。	华为云 ModelArts Lite、百度 BML、MLflow、PyTorch Lightning	华为云 ModelArts、阿里云 PAI、Databricks Lakehouse Platform、Seldon Core、Azure Machine Learning
数字孪生与智能应用平台	构建高保真、可交互、可实时数据驱动的虚拟工厂环境，并深度集成各类 AI 模型服务，驱动智能化的实时监控、模拟推演、预测性维护与自主优化决策应用。	优锘科技 uThings、数字冰雹、Wonderware System Platform、FlexSim	华为云数字孪生平台、英伟达 Omniverse、达索 3DEXPERIENCE、PTC ThingWorx、DataMesh Director

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过部署工业物联网网关、调用系统 API 及建立数据中台等方式，统一采集物理工厂的设备实时运行数据、环境感知数据、MES 数据、EAM 数据，并接入关键区域的视觉监控图像与视频流。**数据预处理。**对采集的多源异构流式与批次数据进行清洗、对齐、归一化与格式化，处理传感器漂移[41]与通信中断导致的异常值与缺失值，统一不同数据源的时间戳至毫秒/微秒级精度，将各类工艺参数、物理量统一至标准量纲，将非结构化图像、日志文本转换为结构化或向量化表示，建立跨系统数据间的时空关联索引。强化数字化交付应用，将工程建设阶段的设计模型、设备参数、施工记录等数据通过数字化交付标准汇入数字孪生平台，实现从建设到运维的数据贯通，为数字孪生体提供精准的初始数据底座，保障虚拟与物理实体的高度映射，支撑 AI 驱动的仿真优化与全生命周期管理。

②第二阶段：样本准备与特征工程

特征工程。基于预处理后的数据，构建服务于孪生体状态表征与异常检测的特征集。如“设备健康度综合指数”“环境舒适度与安全指标”“生产节拍稳定性特征”“基于视觉的设备外观缺陷特征向量”等。**数据标注。**结合设备历史维修工单、质量异常报告、安全事故记录以及人工巡检记录，对相应的历史时序数据段、图像数据进行标注，标识“正常运行”“亚健康状态”“故障初期”“缺陷类型”等标签，为

监督学习模型提供训练样本。**数据增强与划分。**针对少数类故障或异常场景样本不足的问题，采用时序数据合成、图像几何变换与颜色扰动等方法进行数据增强。按时间顺序或设备单元，将数据集划分为训练集、验证集与测试集，确保模型评估的独立性。

③第三阶段：模型训练与仿真验证

模型训练。使用标注和特征工程后的数据集，训练适用于数字孪生场景的 AI 模型，如用于设备剩余使用寿命预测的时序预测模型、用于产品质量异常分类的图像识别模型、用于能效优化的聚类与回归模型。**仿真验证。**在测试集上评估模型的准确率、召回率等性能指标。进一步，将初步训练好的模型或规则引擎与基础的数字孪生仿真环境结合，输入模拟的“设备性能衰减”“突发性环境扰动”“生产订单混合变更”等场景参数，在虚拟环境中验证模型预测结果或决策建议的逻辑合理性与响应时效性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将验证通过的 AI 模型以微服务或组件形式部署至数字孪生平台。对实时汇聚的物理工厂数据流进行在线推理，并将推理结果实时映射至三维孪生模型中，提供可视化告警与决策辅助。**闭环进化。**建立模型输出与实际运维行动结果的比对分析机制。收集预警触发的维修结果、优化建议的实施效果等反馈数据，定期利用新的运行数据与

反馈标签对模型进行增量训练与参数优化，形成“数据采集-模型推理-行动反馈-模型优化”的持续改进闭环。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建工厂级全域数据湖。深度集成工厂全要素基础数据、全生命周期运营数据、多模态感知数据、结构化知识与案例数据，以及数字孪生体自身演化产生的海量仿真与交互数据。**数据预处理。**实施面向复杂 AI 模型的数据预处理。包括对三维点云数据进行降噪、配准与特征提取；对多模态数据进行跨模态对齐与融合编码；对长周期、高维运营数据进行自动模式发现与分段标记；利用知识图谱对非结构化文档、案例进行信息抽取与关联存储，构建统一的工厂数据语义层。强化数字化交付应用，将工程建设阶段的设计模型、设备参数、施工记录等数据通过数字化交付标准汇入数字孪生平台，实现从建设到运维的数据贯通，为数字孪生体提供精准的初始数据底座，保障虚拟与物理实体的高度映射，支撑 AI 驱动的仿真优化与全生命周期管理。

②第二阶段：样本准备与特征工程

特征工程。广泛应用自动机器学习[42]与深度学习进行端到端特征学习。利用图神经网络学习设备、工艺、物料之间的复杂关联拓扑特征；利用卷积神经网络[43]自动从多模态数据中提取高阶抽象特征；构建“全生命周期性能退化轨迹嵌入”“多目标优化决策的潜空间表示”等复合特征。**数据标**

注。采用主动学习、半监督学习与仿真环境自生成标签相结合的方式。利用强化学习智能体在仿真环境中探索产生的“状态-动作-奖励”序列作为训练数据；通过知识图谱推理自动生成部分诊断规则的标注；对极端罕见故障场景，利用生成式模型合成高保真训练数据并标注。**数据增强与划分。**利用生成对抗网络或数字孪生仿真引擎，生成涵盖广泛工况、故障模式、生产场景的合成数据，极大丰富训练样本的多样性与边界条件。按照业务场景、物理产线、产品类型等多维度进行交叉验证划分，确保模型的强泛化与鲁棒性。

③第三阶段：模型训练与仿真验证

模型训练。训练深度强化学习智能体用于自适应控制与调度优化；训练生成式模型用于工艺参数优化、备件设计或异常数据生成；训练多模态大模型用于自然语言交互问答、基于文本或语音的运维指令理解与执行。构建模型协同工作的 AI 决策中枢。**仿真验证。**在超高保真、多物理场耦合的数字孪生“元宇宙”工厂中进行大规模、加速比的仿真验证。进行“全产线重构方案验证”“未知复合故障的诊断与自愈策略推演”“市场需求突变下的供应链与生产协同弹性测试”等复杂场景的沉浸式仿真，全面评估 AI 决策系统的整体效能、安全边界与进化潜力。

④第四阶段：模型部署与闭环进化

模型部署与推理。将先进的 AI 模型集群部署为数字孪生工厂的“智能中枢”。支持基于自然语言、手势或 AR 接

口的复杂交互，系统能够进行实时、并发、协同的推理与决策，自主生成并执行如“全局能效最优的动态生产调度方案”“预测性维护与备件供应的协同策略”“新产品快速导入的产线自配置方案”。**闭环进化。**构建“物理工厂-数字孪生-AI智能体”三元协同的自主进化生态。物理工厂的实时数据驱动孪生体同步演化，并为 AI 提供在线学习样本；AI 智能体的决策在孪生体中预演优化后，再指导或自主控制物理实体行动；行动结果反馈至系统，驱动 AI 模型、孪生模型乃至底层知识图谱的持续自适应优化，实现整个系统的终身学习与智能跃迁。

（四）智能设计与虚拟验证闭环

“智能设计与虚拟验证闭环”场景涉及基础级与进阶级的企业。

基础级企业针对产品设计周期长、跨环节数据联动弱、虚拟验证覆盖不全面等问题，引入 AI 驱动的智能设计与多维度仿真技术，构建设计参数自动优化、多学科联合验证、物理原型与虚拟模型联动迭代的闭环体系，通过 AI 分析市场需求与工艺约束，提升设计方案的可制造性与验证效率时，数据治理的目标主要是打破产品设计、工艺规划、生产制造、市场调研等跨环节数据壁垒，实现数据顺畅流转与协同，确保数据与 AI 智能设计算法、工艺约束要求精准匹配，提升设计方案可制造性，补齐虚拟验证数据短板，确保数据覆盖多学科、多场景验证需求，解决虚拟验证不全面问题，确保

数据在采集、流转、使用全流程的安全与可靠，为虚实联动迭代与决策支撑提供保障。

进阶级企业针对复杂产品多性能强耦合、动态工况模拟精度不足、研发迭代效率低的问题，引入生成式 AI 与因果推理技术，构建全链路自主闭环进化系统，通过 AI 算法实现设计方案的自主生成、极端工况的精准模拟与跨领域知识的智能复用时，数据治理的目标主要是打破复杂产品研发各领域数据壁垒，实现跨领域数据的有机融合与知识沉淀，解决研发迭代中知识复用不足的问题，确保数据能够精准还原复杂产品的物理特性与动态工况，解决动态工况模拟精度不足的问题，建立数据间的因果关联链路，为因果推理技术提供可靠数据支撑，解决多性能强耦合下设计方案优化方向不明确的问题，让数据体系具备随系统进化动态更新的能力，保障全链路自主闭环进化系统的长期有效性。

1.治理对象

表 35　智能设计与虚拟验证闭环场景的数据治理对象

适用层级	数据类型	数据内容
基础级、进阶级	产品设计数据	三维模型参数、结构特征、材料属性、设计参数、CAD 模型、性能指标、多方案对比数据等
基础级、进阶级	市场需求数据	用户反馈、市场趋势、竞品分析、功能需求、隐性需求特征等
基础级、进阶级	仿真与试验数据	结构强度、热力学、流体动力学等多物理场仿真结果、极端工况模拟数据、物理试验记录、失效案例等
基础级	物理原型测试数据	性能测试、可靠性试验、参数测量等结果
基础级	制造工艺数据	设备参数、工艺要求、产能限制等

基础级	历史设计案例数据	成功方案、失效案例、改进记录等
进阶级	生产与工艺数据	制造能力参数、工艺约束、生产可行性验证结果等
进阶级	跨领域知识数据	不同行业研发经验、同类产品设计策略、技术标准规范等
进阶级	供应链数据	供应商产能、物料特性、供应稳定性等
进阶级	数字孪生数据	全生命周期虚拟验证记录、模型参数修正数据等

2.平台（技术）工具

表 36　智能设计与虚拟验证闭环场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现跨研发工具链（CAD/CAE/测试）的多源异构数据的统一接入、清洗、关联与版本化管理。	MySQL/PostgreSQL、Apache NiFi、Jira+Confluence、强思 X-ETFoundation	Siemens Teamcenter、达索 3DEXPERIENCE、PTC Windchill、Oracle Agile PLM
特征工程与样本管理平台	将原始的设计参数、仿真场数据、测试信号转化为可供 AI 模型识别的特征，并对“成功/失败”设计案例进行标注与管理。	Jupyter Notebook、ParaView/MATLAB、Label Studio	Ansys optiSLang、AWS SageMaker Data Wrangler、Tecton、Prodigy
模型开发与运维平台	为“设计代理模型”“缺陷预测模型”等 AI 应用提供开发、训练、集成验证到部署的 MLOps 能力，确保 AI 模型能稳定嵌入现有设计仿真流程并产生价值。	MATLAB、Python(Scikit-learn, PyTorch)、Google Colab、MLflow	Ansys ModelCenter、AWS SageMaker、Altair HyperWorks、神工坊
数字孪生与智能应用平台	构建参数化、可交互、多保真度的虚拟样机环境，并集成各类 AI 代理模型与优化算法，驱动智能化	达索 SIMULIA、Ansys Workbench、Altair Inspire、FlexSim	英伟达 Omniverse、微软 Azure Digital Twins、GE Digital Digital Twin、DataMesh FactVerse

	的设计方案探索、多学科权衡分析与虚拟验证。		

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过 PLM/PDM 系统接口、CAE 仿真软件数据接口、试验数据采集系统以及市场情报平台，统一汇聚市场需求数据、产品三维设计模型与参数数据、结构/热/流体等多学科仿真原始结果数据、物理原型性能测试记录，以及制造工艺能力数据库中的约束数据。**数据预处理。**对多源异构数据进行清洗、对齐、归一化与结构化处理，剔除仿真不收敛或试验设备异常导致的无效数据，补全设计参数缺失值，将不同 CAE 软件的仿真结果映射到统一的设计变量空间，统一物理试验与仿真数据的工况标识，对设计参数、性能指标进行无量纲化处理[44]，将非结构化的市场需求报告、故障案例转化为结构化标签。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，构建用于设计优化与性能预测的特征集。如“设计参数敏感性向量”“多学科性能冲突矩阵”“基于历史案例的工艺可行性评分特征”“市场需求与设计功能的匹配度特征”。**数据标注。**结合历史研发项目中的设计评审结论、物理试验失效报告、制造现场反馈问题，对过往设计方案及其对应的仿真结果、试验数据进行标注，标识“设计可行”“存在干涉”“性能不达标”“可制

造性差”等状态标签，形成监督学习样本。**数据增强与划分。**针对成功优化方案或特定失效模式样本不足的问题，采用基于设计空间采样的方法如拉丁超立方采样，生成新的“设计参数-仿真结果”配对数据，或对现有成功方案进行参数扰动以增强数据。按产品型号或项目阶段将数据集划分为训练集与验证集。

③第三阶段：模型训练与仿真验证

模型训练。使用标注和特征工程后的数据集，训练适用于研发场景的 AI 模型，如用于预测应力、温度、流量等产品性能的代理模型、用于识别设计缺陷的分类模型、用于多目标参数优化的强化学习初代模型。**仿真验证。**在独立的验证集上评估代理模型的预测精度。进一步，将 AI 模型与现有的 CAE 仿真流程集成，在虚拟环境中对模型推荐的新设计方案进行快速仿真验证，对比 AI 预测结果与高保真仿真结果，评估模型在“多约束条件优化”“跨学科性能平衡”等场景下的有效性与可靠性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将验证通过的 AI 模型部署至智能设计辅助系统。设计人员输入设计目标与约束条件后，系统可调用模型进行推理，输出“推荐的设计参数组合”“潜在的性能瓶颈预警”“基于历史案例的改进建议”等，辅助设计决策。**闭环进化。**建立“设计-仿真-试验-制造”的数据反馈链路。收集新设计方案在后续高保真仿真、物理试验及小批量

试制中产生的真实结果数据，与模型当初的预测和建议进行对比分析，形成反馈数据。定期利用新的项目数据对模型进行再训练与调优，使其持续适应新的产品类型与技术发展。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级研发数据湖。深度集成全粒度产品设计数据、高维市场与用户隐性需求数据、高保真多物理场耦合仿真及不确定性分析数据、全生命周期物理试验与在役监测数据、跨领域知识图谱、供应链能力数据，以及数字孪生体记录的全流程虚拟验证与参数修正数据。**数据预处理。**实施面向复杂 AI 模型的深度预处理。包括对非结构化知识文档进行自然语言处理与向量化嵌入；对高维仿真场数据进行降维与特征提取；对因果链明确的试验失效数据进行图结构构建；统一所有数据在“产品-配置-版本-工况”多维坐标下的标识与关联。

②第二阶段：样本准备与特征工程

特征工程。应用图神经网络、Transformer 等先进模型进行端到端特征学习。自动学习“设计参数-多物理性能-工艺约束-市场需求”之间的复杂非线性映射与高阶交互特征；构建“基于知识图谱的设计规则符合性嵌入向量”“产品全生命周期性能演化轨迹特征”。**数据标注。**采用基于强化学习的环境交互自动生成样本，智能体在仿真环境中探索设计空间获得“状态-奖励”数据；利用因果发现技术对历史成功与失

败案例进行根因分析与自动标注；构建包含物理规律的合成数据生成器，用于扩充极端、罕见工况下的训练样本。**数据增强与划分。**利用生成式对抗网络或扩散模型，学习现有优秀设计方案与仿真数据的分布，生成大量符合物理规律与设计美感的新颖概念方案数据。按照产品平台、性能维度、工况复杂性进行多层次、交叉的数据集划分，确保模型的强大泛化与创新能力。

③第三阶段：模型训练与仿真验证

模型训练。训练生成式设计模型，如基于 Transformer 或 VAE[45]的架构，实现给定设计目标下的创新方案自动生成；训练融合物理知识的神经算子[46]模型，实现超高效率的瞬态、非线性仿真预测；训练因果推断模型，用于识别设计缺陷的根本原因并提供可解释的改进方向。仿真验证。在基于高性能计算（HPC）[47]和数字孪生构建的“虚拟验证宇宙”中进行系统性验证。对生成式 AI 提出的创新方案进行“全工况、全参数、全生命周期”的加速虚拟验证，评估其在未知极端动态载荷、长期老化、多失效模式耦合等场景下的鲁棒性与优越性，实现“AI 设计-AI 验证”的内循环。

④第四阶段：模型部署与闭环进化

模型部署与推理。将高级 AI 模型集群部署为下一代智能研发平台的“核心大脑”。支持基于自然语言或高级目标的设计需求输入，系统能够进行自主推理、多轮迭代，输出“帕累托最优方案族及其详细性能预测报告”“满足新市场

需求的颠覆性概念设计”“基于供应链实时状态的韧性设计方案”。**闭环进化。**构建“物理世界-数字孪生-AI 设计引擎”三位一体的自进化生态系统。物理产品在试验与服役中产生的真实数据，持续校准数字孪生模型与 AI 代理模型的预测精度；AI 设计引擎从数字孪生提供的海量虚拟验证经验中持续学习，优化设计策略；新设计方案的制造与服役数据又反馈回系统，形成驱动产品性能持续超越的正向闭环。

（五）工艺与产品智能协同验证

“工艺与产品智能协同验证”场景仅涉及进阶级的企业。

针对产品与工艺协同研发中数据割裂、验证滞后、迭代低效的问题，引入生成式 AI 与多智能体决策技术，构建全域数字孪生驱动的协同，实现设计与工艺的同步生成、全域虚拟验证与自主协同优化时，数据治理的目标主要是打破产品设计与工艺研发两大领域的数据壁垒，实现数据全链路贯通与高效流转，确保数据实时反映设计与工艺的动态变化，支撑全域数字孪生的实时映射与虚拟验证，建立设计、工艺、资源等多维度数据的精准关联，为多智能体决策技术提供可靠的协同分析数据支撑。

1.治理对象

表 37　工艺与产品智能协同验证场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	产品设计数据	设计参数、三维模型、性能指标、变更记录等
进阶级	工艺数据	工艺方案、工序参数、制造约束规则、工艺知识图谱等

进阶级	产线数据	设备参数、产能信息、运行状态、适配性验证结果等
进阶级	市场与需求数据	市场趋势、用户需求、竞品工艺特点等
进阶级	仿真与验证数据	全工艺链虚拟验证结果、多物理场仿真数据、偏差分析记录等
进阶级	质量数据	产品质量检测结果、工艺参数与质量关联数据、失效案例等
进阶级	跨行业工艺知识数据	不同行业工艺经验、技术标准、创新方案等

2.平台（技术）工具

表 38　工艺与产品智能协同验证场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级
数据治理与集成平台	实现产品设计、工艺规划、生产制造与质量验证等跨领域、全生命周期异构数据的统一接入、清洗、关联融合与资产化管理，构建支持智能协同的单一可信数据源。	达索 3DEXPERIENCE、AVEVA System Platform、华为云 DataArts Studio、西门子 Xcelerator 生态系统
特征工程与样本管理平台	将复杂的三维模型参数、多物理场仿真结果、工艺知识图谱等数据，转化为表征“设计-工艺-质量”内在关联的深度特征，并对成功/失败协同案例进行智能化标注、合成与版本管理。	英伟达 Cosmos 系列 AI 模型、Tecton、Weights & Biases、Prodigy
模型开发与运维平台	为生成式设计-工艺协同模型、可制造性预测模型等提供从开发、训练、数字孪生验证到部署、监控的企业级 MLOps 能力，确保 AI 模型能稳定、可靠地嵌入核心研发流程。	阿里云 PAI、Databricks Lakehouse Platform、Azure Machine Learning
数字孪生与智能应用平台	构建高保真、多物理场耦合的产品与产线一体化虚拟孪生环境，并深度集成各类 AI 模型，驱动设计与工艺的同步生成、全域虚拟	达索 3DEXPERIENCE、英伟达 Omniverse、华为云数字孪生平台、PTC ThingWorx

	验证与自主协同优化等核心智能应用。	

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级工艺-产品协同数据湖。通过产品PLM、MOM、CAD/CAE、ERP等系统的标准化接口与API，统一汇聚全量结构化与非结构化数据。**数据预处理。**实施基于本体论[48]的深度数据治理。对多源异构数据进行清洗、对齐、归一化处理，处理仿真不收敛数据、设备通信异常值，将设计特征ID、工艺工序ID、检测点ID进行全局唯一映射与关联，统一各类工程参数的单位与量纲。关键步骤是构建“工艺-产品”协同知识图谱，将设计意图、工艺约束、制造能力、质量要求等概念及其复杂关系进行形式化定义与存储，为后续智能应用提供统一的语义理解基础。

②第二阶段：样本准备与特征工程

特征工程。基于知识图谱与原始数据，构建用于刻画“设计-工艺-质量”内在关联的深度特征。如“设计特征可制造性评价向量”“工艺链稳健性指数”“跨领域方案迁移相似度特征”。利用图神经网络自动学习设计元素与工艺节点之间的潜在关联模式。**数据标注。**引入生成式AI技术，构建“虚拟样本工厂”。基于历史成功案例与物理规则，通过条件生成对抗网络生成覆盖“设计变更-工艺调整”组合的合成数据。同时，利用强化学习智能体在数字孪生环境中进行探索，自

动产生大量“状态-动作-结果”的决策序列样本。对历史项目中的设计缺陷、工艺失效、质量问题进行根因分析，并借助知识图谱进行半自动的因果标注。**数据划分。**采用基于“产品族-工艺路线”的聚类方法进行数据划分，确保训练集、验证集和测试集能够覆盖不同的产品变型与工艺组合，评估模型的泛化能力与跨项目适应性。

③第三阶段：模型训练与仿真验证

模型训练。训练面向协同验证的 AI 模型集群。针对生成式设计-工艺协同智能体，基于 Transformer 或扩散模型，接收产品功能需求与约束，同步生成可行的产品概念设计与初步工艺方案；针对工艺链仿真与优化智能体，利用物理信息神经网络或深度强化学习，对生成的工艺方案进行全链路的虚拟执行与多目标优化。针对可制造性风险预测智能体，基于图注意力网络，实时预测新设计方案在现有或规划产线上可能存在的装配干涉、加工难度、质量风险。**仿真验证。**在基于全要素、高保真数字孪生构建的“虚拟制造世界”中，对 AI 协同生成的方案进行毫秒级加速的沉浸式验证。模拟验证场景需超越单点工艺，覆盖“从原材料到成品”的全工艺链动态耦合效应，以及应对设备扰动、物料偏差等不确定性因素的鲁棒性。验证结果将作为强化学习智能体的奖励信号，驱动其策略优化。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练成熟的 AI 模型集群以微服务架构部署至“工艺与产品智能协同平台”。该平台提供统一入口，设计工程师输入需求后，系统可调用多个智能体进行协同推理，实时输出“推荐的产品-工艺联合方案包”“全链路虚拟验证报告”“综合成本与风险量化评估”以及“多方案对比分析”。**闭环运营。**建立跨越虚拟与物理世界的持续学习闭环。物理世界生产执行的真实数据持续回流，用于校准数字孪生模型与 AI 预测模型。平台自动捕获每一次“人工决策否决 AI 推荐”的案例，将其作为关键样本纳入分析。同时，成功的协同方案将自动沉淀为结构化知识，丰富和更新“工艺-产品”协同知识图谱，使系统具备从历史与实时反馈中不断积累、进化的能力。

（六）生产计划优化

“生产计划优化”场景涉及基础级与进阶级的企业。

基础级企业针对生产计划人工排程效率低、需求预测不准、应对波动能力弱等问题，引入 AI 驱动的智能优化算法与实时决策模型，构建融合多源动态数据的生产计划系统，实现需求精准预测、计划自动生成与动态调整，提升生产计划的科学性与灵活性时，数据治理的目标主要是打破生产计划相关数据的分散壁垒，实现多源数据的标准化整合，为 AI 算法提供全面数据输入，解决数据滞后问题，确保数据能够实时反映生产与需求的动态变化，提升系统应对波动的能力，解决数据质量问题导致的预测偏差与计划失误，确保输入 AI

算法的数据真实可靠，规范数据全生命周期管理，保障生产计划系统稳定运行。

进阶级企业针对市场多品种小批量、订单波动剧烈的复杂场景，引入深度强化学习与全局协同决策技术，构建生产计划自主进化系统。通过 AI 算法实现需求预测的长周期精准化、生产计划的全局自优化、异常调整的实时连锁响应时，数据治理的目标主要是解决长周期预测的数据支撑不足问题，实现全周期数据的完整覆盖与有效整合，支撑深度强化学习算法挖掘需求波动规律，打破生产、销售、供应链、仓储等跨域数据壁垒，实现全域数据的深度协同关联，支撑全局协同决策技术落地，解决订单波动场景下数据滞后、失真问题，确保数据实时反映市场与生产动态，支撑异常调整的快速连锁响应，让数据体系具备随市场变化、系统进化动态调整的能力，保障生产计划自主进化系统的长期有效性。

1.治理对象

表 39　生产计划优化场景的数据治理对象

适用层级	数据类型	数据内容
基础级、进阶级	订单数据	订单类型、数量、交付日期、订单信息、交期要求、优先级、产品类型等
基础级、进阶级	生产资源数据	设备参数、负荷状态、产能信息、维护记录等
基础级、进阶级	物料数据	库存数量、库存水平、采购周期、供应商信息、供应链状态、物料需求计划等
基础级	设备数据	产能、运行状态、维护计划等
基础级	生产工艺数据	工序流程、工艺约束、生产周期等
基础级	历史生产计划数据	计划方案、执行情况、调整记录等

基础级	实时动态数据	设备故障、订单变更、物料延迟等
进阶级	市场需求及预测数据	市场趋势、宏观经济指标、行业周期、区域政策、季节因素历史销售数据等
进阶级	异常事件数据	设备故障记录、物料延迟情况、质量问题、异常调整方案等
进阶级	多厂区协同数据	各厂区产能、计划分配、资源共享情况等
进阶级	数字孪生仿真数据	生产流程模拟结果、计划执行预演数据、异常场景推演记录等

2.平台（技术）工具

表 40　生产计划优化场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现订单、产能、库存、工艺等多源异构数据的统一接入、清洗、关联与资产化管理。	观测云、TDengine、网易数帆 DataOps 平台	华为云 DataArts Studio、阿里云实时计算 Flink 版、达索 3DEXPERIENCE
特征工程与样本管理平台	将生产时序数据（如订单流、设备状态）与业务规则转化为供优化算法与 AI 模型使用的特征，并对“计划成功/失败”样本进行标注、增强与管理。	Jupyter Notebook、Label Studio、Pandas/NumPy	Tecton、Weights & Biases、Prodigy
模型开发与运维平台	为需求预测、排产优化、异常模拟等模型提供从开发、训练、仿真验证到部署、监控的 MLOps 全生命周期管理能力，确保模型在生产环境中的稳定与持续优化。	华为云 ModelArts Lite、MLflow、PyTorch Lightning	华为云 ModelArts、阿里云 PAI、Databricks Lakehouse Platform
数字孪生与智能应用平台	构建可交互、可模拟、数据驱动的虚拟产线/工厂环境，深度集成计划优化模型，驱动计划方案仿真推演、瓶颈预	FlexSim、AnyLogic、Wonderware System Platform	英伟达 Omniverse、达索 3DEXPERIENCE、华为云数字孪生平台

	演、动态调整与可视化决策等核心智能应用。		

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过 ERP、MES、WMS、SCM 等系统的数据库接口或中间件，统一抽取订单数据、历史销售数据、物料库存与采购数据、设备基础产能与状态数据、标准工艺路线数据，以及近期的计划执行与调整记录。**数据预处理。**对多源业务数据进行清洗、对齐、归一化与时间序列化，修正手工录入错误、填补因系统未对接导致的关键字段缺失，将不同系统中的物料编码、设备编码、订单号进行映射统一，将数量、时间、金额等指标统一量纲，将所有数据按计划周期，如日、周，进行规整对齐，形成结构清晰、可用于分析的基础数据表。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，构建用于预测和优化的基础特征集。例如，构建“订单紧急程度指数”“物料齐套率预测”“设备综合利用率趋势”“基于历史相似度的需求波动特征”。**数据标注。**结合历史计划评审记录、计划延误报告及生产调度日志，对历史计划方案及其对应的执行结果进行关联标注，标识“计划可行”“计划过于激进导致延误”“计划保守造成闲置”等状态标签。数据增强与划分。针对市场突变或重大设备故障等少数异常场景样本不足的问题，

采用基于时间序列分解与重组的方法进行数据增强。按时间顺序，将数据集划分为训练集（用于学习规律）与测试集（用于验证模型效果）。

③第三阶段：模型训练与仿真验证

模型训练。使用标注和特征工程后的数据集，训练适用于计划场景的 AI 模型，如用于短期需求预测的时序模型、用于初步排产优化的启发式算法或整数规划模型、用于识别潜在延误风险的分类模型。**仿真验证。**在测试集上评估需求预测模型的准确率。利用离散事件仿真软件或简易规则引擎，构建产线基础仿真模型，输入 AI 生成的计划方案，模拟“订单临时插入”“关键设备突发停机”等场景，验证计划的鲁棒性与可执行性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将验证通过的模型集成至现有计划系统或作为独立模块部署。计划员输入初步约束后，系统可调用模型进行辅助推理，输出“未来周期需求预测值”“推荐的主生产计划草案”“高风险的订单或资源瓶颈预警”。**闭环进化。**建立计划执行反馈机制。收集实际生产进度、订单完成情况与计划预测值的偏差数据，定期利用新的数据对预测模型进行重训练与校准，对优化模型的参数进行调优，形成初步的“计划-执行-反馈-优化”数据闭环。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级生产运营数据湖。实时接入全域数据：高细粒度订单流数据[49]、物联网平台采集的设备实时状态与性能参数、供应链协同平台的物料在途与供应商绩效数据、外部整合的市场大数据与宏观经济指标、多工厂协同数据，以及数字孪生平台产生的海量仿真推演数据。**数据预处理。**实施面向实时决策的流批一体数据处理[50]。对实时数据流进行在线清洗、窗口聚合与复杂事件检测；对多源异构数据利用知识图谱技术进行语义关联，例如建立“订单-产品-物料-设备-工序-工厂”的全景关联网络；对数据进行不确定性量化，为后续的鲁棒优化提供输入。

②第二阶段：样本准备与特征工程

特征工程。应用图神经网络、注意力机制等先进技术进行深度特征提取。自动学习“市场信号-订单模式-生产扰动”之间的动态关联特征；构建“全局资源负载均衡度向量”“供应链韧性评估指标”“多目标（交期、成本、能耗）权衡的帕累托前沿特征”。**数据标注与合成。**采用深度强化学习环境自生成样本。智能体在与高保真数字孪生环境的交互中，产生海量的“状态（当前计划与资源）-动作（调整决策）-奖励（综合效益）”序列，作为训练数据。同时，利用生成对抗网络模拟极端市场波动、复合型供应链中断等罕见但高影响的“黑天鹅”场景数据，用于增强模型的抗风险能力。**数据划分。**采用基于场景的划分方法，确保训练集能覆盖平稳

期、促销期、危机期等多种运营模式，测试集包含未曾见过的、复杂的组合式扰动场景，以检验系统的泛化与应急能力。

③第三阶段：模型训练与仿真验证

模型训练。训练深度强化学习智能体作为核心计划决策引擎。该智能体以全局生产状态为输入，以调度指令为输出，以综合经济效益、客户满意度、资源平稳度等为优化目标，通过与环境交互自主学习最优决策策略。同时，训练融合外部信息的深度需求感知网络。**仿真验证。**在基于全要素数字孪生构建的“虚拟工厂集群”中进行百万次级的加速仿真与对抗性测试。在仿真中注入各种历史及虚拟的极端扰动，验证智能体在“全球供应链局部断裂”“市场需求一夜翻倍”“多厂区产能协同博弈”等超复杂场景下的决策质量、收敛速度与稳定性，确保其策略优于任何固定规则或传统优化算法。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练成熟的深度强化学习智能体及配套模型，以“决策即服务”的形式部署于云边协同架构中。系统能够实时感知内外部状态变化，进行毫秒级推理，动态生成并发布全局最优的滚动生产计划，同时向具体执行单元下达精准指令。**闭环运营。**构建“物理-虚拟”双向实时映射的增强学习闭环。真实生产系统的执行结果不断反馈，用于微调数字孪生模型和在线校准智能体策略。系统自动分析每一次人工干预背后的隐性知识，并将其转化为规则或约束，

注入到训练环境中。所有成功的优化策略均被沉淀至企业计划策略库，形成可复用的组织智慧，实现系统的永续自主进化。

（七）生产执行智能联动优化

“生产执行智能联动优化”场景仅涉及进阶级的企业。

进阶级企业针对生产全要素联动不足、异常响应滞后、资源调度低效等问题，引入多智能体协同决策与全域数字孪生技术，构建全链路自主运行的智能中枢，实现生产要素的迅速联动、扰动自适应调整与全流程智能优化时，数据治理的目标主要是构建标准化、高时效性、高可信度的全域生产数据底座，保障多智能体间的数据交互效率与数字孪生模型的实时映射精度，为生产要素联动决策提供高质量的数据支撑。

1.治理对象

表 41　生产执行智能联动优化场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	生产全要素数据	设备运行参数、人员状态、物料位置与数量、环境温湿度等
进阶级	工序与进度数据	各工序加工状态、工件流转记录、全工序数字线程数据等
进阶级	资源调度数据	AGV 路径、机器人任务、设备负荷分配等
进阶级	异常事件数据	设备故障、质量异常、紧急插单等事件记录及影响分析
进阶级	决策与调整数据	调度方案、参数调整记录、优化结果等
进阶级	数字孪生数据	生产状态实时映射数据、场景模拟结果等
进阶级	跨车间知识数据	历史调度经验、异常处理案例、优化策略等

2.平台（技术）工具

表 42　生产执行智能联动优化场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级企业
数据治理与集成平台	负责生产全要素、工序、资源调度等七类数据的实时采集、清洗、融合与质量管理，构建统一、可信的数据湖。	华为云物联网平台、阿里云 DataWorks、TDengine
特征工程与样本管理平台	提供从原始数据到模型训练样本的全流程加工能力，包括特征构造、数据标注、增强与版本管理。	Apache Flink、Label Studio、Feast
模型开发与运维平台	支撑多智能体协同决策等 AI 模型的开发、训练、部署与全生命周期管理，实现高效的 MLOps。	百度 PaddlePaddle、华为云 ModelArts、MLflow
数字孪生与智能应用平台	构建高保真虚拟生产环境，用于策略仿真验证、多智能体协同训练，并作为连接数据、模型与物理世界的“智能决策与指挥中枢”。	达索 3DEXPERIENCE、英伟达 Omniverse

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过工业物联网平台与企业服务总线，实时汇聚生产全要素、工序进度、资源调度等七类数据至统一数据湖，确保数据全量接入。**数据预处理。**执行数据清洗（处理缺失、异常值）、格式标准化、关键字段（如设备 ID、订单号）一致性对齐，并对时序数据进行戳同步，为后续分析提供干净、一致的数据集。

②第二阶段：样本准备与特征工程

特征工程。基于业务知识，从原始数据中构造高级特征。例如，将设备振动时序数据转换为频域特征；将 AGV[51]路径

与任务列表聚合为“动态负载率”；构建反映“人–机–环”耦合关系的综合指标。**数据标注。**对历史异常事件、调度决策等结果数据进行分类与归因标注，形成“工况—结果”监督学习标签。**数据增强与划分。**针对小样本异常场景，采用合成少数类过采样[52]等技术进行数据增强。按时间或比例将处理后的数据集划分为训练集、验证集与测试集。

③第三阶段：模型训练与仿真验证

模型训练。使用划分后的训练集，驱动多智能体强化学习等算法进行模型训练。将跨车间知识数据（历史案例、策略）作为规则约束或先验知识注入训练过程，加速收敛。**仿真验证。**在独立的验证集与测试集上评估模型性能（如预测准确率、调度效率提升率）。同时，在数字孪生环境中，将模型置于“紧急插单”“设备突发故障”等仿真场景进行压力测试与鲁棒性验证，完成虚拟空间的“推理”预演。

④第四阶段：模型部署与闭环进化

模型推理与部署。将通过验证的模型部署为在线服务，接入实时数据流，对生产扰动进行实时推理，并输出调度建议或控制参数至执行系统。**闭环运营。**持续采集模型推理结果产生的决策与调整数据及其对应的实际优化结果，将这些新的“反馈数据”作为增量数据，回流至数据湖。以此驱动特征集的迭代、标注体系的完善以及模型的周期性再训练，形成数据与智能共进的飞轮。

（八）仓储智能管理

“仓储智能管理”场景涉及入门级、基础级与进阶级的企业。

入门级企业针对人工搬运效率低、库存记录易出错、库位利用率不高等问题，部署自动化设备和基础感知技术，结合 WMS 系统实现物料自动识别、流程标准化与库存数据自动采集，提升基础作业效率和库存准确性时，数据治理的目标主要是保障自动化设备与 WMS 系统的数据交互精准、流程衔接顺畅，为提升基础作业效率和库存准确性筑牢数据根基。

基础级企业在自动化基础上，针对复杂订单拣选路径冗余、库存呆滞料积压、存储策略缺乏灵活性等问题，引入深度 AI 优化算法与预测模型，结合 WMS 系统实现拣选路径动态规划、库位智能分配、多形态物料混存优化及自动化盘点，提升库存周转效率与空间利用率时，数据治理的目标主要是构建适配 AI 算法与预测模型运行的标准化、高关联性、时序化数据体系，消除算法输入数据的噪声与偏差，保障模型训练与推理的精准性，为仓储场景的智能决策提供稳定可靠的数据输入。

进阶级企业针对复杂动态仓储环境中预测精度不足、协同能力有限、自适应调整滞后等问题，引入多智能体协同决策与全域数字孪生技术，构建全链路自主运行的智能仓储系统，通过 AI 算法实现需求精准预判、资源全域协同、流程自

适应调整时，数据治理的目标主要是打造全域协同、实时映射、高可信度的仓储数据闭环体系，打通物理仓储与数字孪生空间的数据链路，保障多智能体间的数据交互效率与决策指令的精准下达，为智能仓储系统的自主运行提供全维度的数据支撑。

1.治理对象

表 43　仓储智能管理场景的数据治理对象

适用层级	数据类型	数据内容
入门级、基础级、进阶级	仓储核心数据	库存信息、物料属性、存储位置、出入库记录等
入门级、基础级、进阶级	作业执行数据	机器人运行状态、AGV 路径、分拣记录、设备负荷等
入门级、基础级、进阶级	需求与订单数据	订单信息、生产计划、市场趋势、短时需求预测结果等
入门级、基础级、进阶级	供应链数据	各节点库存特征、供应关系、调拨记录、合约约束等
入门级、基础级、进阶级	环境与设备数据	仓储环境参数、智能货架状态、感知设备监测数据等
进阶级	数字孪生数据	仓储全要素模拟数据、极端场景推演结果、策略预演记录等
基础级、进阶级	历史知识数据	订单规律、库存波动趋势、设备调度记录、历史作业案例、优化策略、异常处理经验等

2.平台（技术）工具

表 44　仓储智能管理场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	负责多源数据的采集、清洗、整合与质量管控，构	阿里云 DataWorks、华为云 DataArts Lite、主流 WMS/ERP 厂商	阿里云 DataWorks、华为云 DataArts Studio、	Apache Doris, StarRocks、西门子 MindSphere

	建统一可信的数据基座。	提供的原生数据集成模块、Apache NiFi, Kettle）	TDengine, InfluxDB	
特征工程与样本管理平台	提供特征提取、构造、标注与版本管理能力，将数据转化为模型燃料。	FineBI, Tableau Prep、WMS/SCM 系统内置的报表与数据分析模块、Excel 高级功能/Python 脚本	华为 ModelArts、百度 BML、Amazon SageMaker Data Wrangler、Label Studio	Feast、Hopsworks、DataRobot, H2O.ai
模型开发与运维平台	支撑 AI 模型的开发、训练、部署、监控与迭代全生命周期管理。	WMS/自动化设备供应商提供的优化算法包、云端托管的 AutoML 服务	阿里云 PAI，腾讯云 TI-ONE、MLflow、Kubeflow、scikit-learn, XGBoost	英伟达 Isaac Lab，Unity ML-Agents、Databricks，Azure Machine Learning、Ray RLlib
数字孪生与智能应用平台	构建仓储虚拟映射，进行仿真推演、协同训练，并驱动智能应用。	WMS/SCADA 系统的图形化监控界面、部分国产 SCADA 工具、基础版三维可视化工具	基于 Three.js,Unity 的定制化看板、ThingJS,智轻云	达索 3DEXPERIENCE，西门子 Process Simulate、DataMesh FactVerse、英伟达 Omniverse

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。定义并实施 RFID、条形码扫描设备、AGV 与 WMS 之间的标准化数据接口协议，确保物料 ID、库位编码、数量等核心信息的准确、无丢失传输。**数据预处理。**建立数据质量检核规则，对采集的库存数据进行重复、缺失、

逻辑冲突的实时清洗与告警，确保 WMS 中库存记录的唯一性与真实性。

②第二阶段：样本准备与特征工程

特征工程。基于标准化的出入库、库位数据，加工计算“库位利用率”“分拣作业单平均耗时”“库存数据准确率”等关键绩效指标特征。**数据标注、增强与划分。**此阶段以流程规则验证为主，标注工作主要用于标记“异常操作事件”作为案例样本。

③第三阶段：模型训练与仿真验证

模型训练与验证。重点在于业务流程验证。利用历史数据，验证 WMS 内置的存储上架、拣货路径等基础规则在不同数据输入下的执行正确性与效率，确保流程标准化。

④第四阶段：模型部署与闭环进化

模型推理与闭环进化。核心是稳定执行。确保基于实时采集数据触发的自动化作业指令（如根据入库单分配库位）被可靠执行，并将执行结果（确认上架）数据闭环反馈至 WMS，更新库存账目，形成“数据驱动作业、作业产生数据”的基础业务闭环。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。除设备数据外，集成 ERP 订单数据、历史出入库明细、物料主数据（尺寸、品类）等，形成多维度数据关联。**数据预处理。**进行深入的时序对齐（将订单、库存、

作业事件按时间序列对齐）、异常值检测与平滑（如处理传感器瞬时波动），并处理数据不平衡问题（如呆滞料与热销料样本不均），为算法提供高质量输入。

②第二阶段：样本准备与特征工程

特征工程。构建复杂特征，如“物料热度时序趋势”“基于订单关联的共拣概率”“库区间的搬运成本矩阵”“多维（品类、体积、重量）相似度”等，为路径规划与库位分配提供深度输入。**数据标注。**对历史订单进行“最优拣选路径”标注，对物料进行“周转等级（快/中/慢）”和“存储策略类别”标注。**数据增强与划分。**采用合成少数类过采样技术增强“呆滞料转活跃”等少数场景样本；按时间窗口划分数据集，确保训练集与测试集的时序独立性。

③第三阶段：模型训练与仿真验证

模型训练。使用标注好的特征集，训练如协同过滤推荐算法（用于关联拣选）、时序预测模型（用于需求预测）、组合优化算法（用于路径规划）。**模型验证。**采用交叉验证评估模型泛化能力，并使用离线历史数据回测，对比模型推荐策略与实际执行结果的效能提升。

④第四阶段：模型部署与闭环进化

模型推理。将训练好的模型部署为微服务，接收实时订单和库存状态，在线推理出动态拣选路径、推荐上架库位等决策建议。**闭环进化。**建立模型效果监控看板，采集决策执

行后的实际效能数据（如实际行走距离、出库效率），定期将新数据注入训练集，启动模型的迭代再训练与优化。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。实现所有仓储元素的全域、全量、实时数据接入，并统一至时空基准坐标系下。数据预处理。实施毫秒级时间戳同步与空间位置标定，对多源数据进行实时融合（如将视觉识别结果与 RFID 位置信息融合），保障数字孪生体映射的高保真度与一致性。

②第二阶段：样本准备与特征工程

特征工程。构造用于多智能体协同的高阶特征，如“全域动态交通密度热力图”“任务冲突概率预估”“基于仿真的未来状态预测特征”。**数据标注。**对复杂异常事件（如多车死锁、高峰期系统响应迟滞）进行根本原因与解决策略的序列标注。**数据增强与划分。**利用数字孪生引擎，生成大量极端、罕见场景的仿真数据（如突发设备故障），扩充训练样本的多样性与复杂性。

③第三阶段：模型训练与仿真验证

模型训练。在数字孪生环境中，采用多智能体强化学习等算法，让 AGV 调度、货架搬运、订单分配等智能体在仿真环境中进行海量次博弈与协同训练。**模型验证。**在孪生环境中设计极限压力测试场景（如瞬时订单洪峰、多节点故障），验证智能体集群决策的鲁棒性、协同效率与自恢复能力。

④第四阶段：模型部署与闭环进化

模型推理。将训练成熟的“多智能体决策大脑”部署上线，实现对复杂动态环境的实时自主决策与调度。**闭环进化。**建立双向数据驱动闭环：物理世界执行数据实时驱动孪生体更新；同时，在孪生体中持续进行策略探索与仿真优化，将已验证的更优策略动态加载至物理系统，实现整个仓储系统的自主、持续进化。

（九）物料精准配送

“物料精准配送”场景涉及入门级、基础级与进阶级的企业。

入门级企业针对人工配送效率低、错送率高、路径规划不合理等问题，部署 AGV/AMR[53]等自主移动设备，依托 AI 定位与环境感知技术，结合调度系统实现物料点到点自动配送，通过 AI 算法处理任务指令，提升配送准时性与准确性时，数据治理的目标主要是搭建适配 AGV/AMR 运行与调度系统的标准化基础数据体系，保障设备感知数据、任务指令数据与站点物料数据的精准匹配，打通设备端与系统端的数据交互链路，为 AI 定位导航与任务调度算法提供高质量的基础数据输入。

基础级企业针对厂区环境动态变化、配送任务复杂度提升导致的路径低效、集群协同不足等问题，引入 AI 感知与智能决策技术，构建融合多模态环境数据的配送系统，通过 AI 算法实现动态路径规划、集群协同调度与数字孪生仿真优

化，提升物料配送的精准度与动态适应能力时，数据治理的目标主要是构建多模态、高关联、时序化的厂区配送数据体系，消除环境感知与任务调度的数据割裂，保障 AI 算法对动态场景的精准研判能力，为数字孪生仿真提供高保真的数据输入，支撑配送系统的智能决策与动态优化。

进阶级企业针对大规模复杂生产物流中预测滞后、调度低效、适应能力弱的问题，引入多智能体深度协同与全域数字孪生技术，构建全链路自主运行的智能配送系统，通过前沿 AI 算法实现需求精准预判、全局协同调度、动态自适应调整时，数据治理的目标主要是打造覆盖“需求-资源-路径-执行”全链路的高可信、强关联、实时化数据闭环体系，打通多智能体间的数据交互壁垒，保障全域数字孪生模型与物理配送系统的精准映射与实时推演，为智能配送系统的自主决策与全局优化提供全维度的数据支撑。

1.治理对象

表 45　物料精准配送场景的数据治理对象

适配层级	数据类型	数据内容
入门级、基础级	物料数据	类型、数量、目的地、优先级等
入门级、基础级	设备运行数据	AGV/AMR 的位置、状态、负载、电量等
入门级、基础级	配送任务数据	任务内容、时间要求、优先级变化等
基础级	路径数据	历史轨迹、拥堵情况、标线标识等
基础级、进阶级	生产与需求数据	生产工单、设备状态、工序进度、物料需求计划等
进阶级	仓储关联数据	库存信息、物料位置、出入库记录等
进阶级	配送执行数据	AMR 运行状态、任务分配、路径信息、交通流状态等
入门级、基础级、进阶级	环境感知数据	激光雷达点云、视觉图像、超声波信息、动态障碍物轨迹、车间场景图像、障碍物信息、定位数据等
进阶级	跨系统数据	MES 生产进度、WMS 库存状态、ERP 订单需求等
进阶级	数字孪生数据	配送系统模拟数据、扰动场景推演结果、调度策略预演记录等
基础级、进阶级	历史知识数据	任务完成效率、异常处理记录、历史配送案例、优化策略、异常处理经验等

2.平台（技术）工具

表 46　物料精准配送场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	负责多源数据的采集、清洗、整合与质量管控，构建统一可信的数据基座。	ThingsBoard、腾讯云 IoT Hub、Apache NiFi	华为云 ROMA Connect、TDengine、阿里云 DataWorks	阿里云 DataWorks、Hologres、华为云 DataArts Studio 、Apache Doris
特征工程与样本管理平台	提供特征提取、构造、标注与版本管理能力，	Python、Jupyter Notebook、LabelImg	Amazon SageMaker Data Wrangler、华为云 ModelArts、Label Studio	Feast、阿里云 PAI、Snorkel

	将数据转化为模型燃料。			
模型开发与运维平台	支撑AI模型的开发、训练、部署、监控与迭代全生命周期管理。	ROS 内置导航栈、腾讯云 TI-ONE、scikit-learn	百度 PaddlePaddle、阿里云 PAI、MLflow	Ray RLlib、英伟达 Omniverse Replicator、Kubeflow
数字孪生与智能应用平台	构建仓储/厂区虚拟映射，进行仿真推演、协同训练，并驱动智能应用。	自家 WMS/调度系统的可视化监控大屏、轻流/简道云 ECharts	51WORLD 数字孪生平台、优锘科技 ThingJS、Unity 3D	英伟达 Omniverse、达索 3DEXPERIENCE、DataMesh Director

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过标准协议集成 AGV/AMR 的实时状态数据（位置、电量、速度）、任务指令数据（起点、终点、物料编码）及 WMS 的站点物料数据。**数据预处理。**清洗设备上报数据中的噪声与跳点；统一物料编码、站点编码与系统主数据的一致性；对任务指令进行完整性校验。

②第二阶段：样本准备与特征工程

特征工程。构建“站点间物理距离”“历史任务平均耗时”“设备实时负载率”等基础路径规划与调度特征。**数据**

标注与划分。对历史任务执行结果进行“成功/超时/失败”标注，作为模型评估样本。按时间顺序划分数据集。

③第三阶段：模型训练与仿真验证

模型训练与验证。使用经典算法或轻量机器学习模型，在历史数据上进行静态路径规划与任务分配模型的训练与离线验证，确保其在典型场景下的有效性。

④第四阶段：模型部署与闭环进化

模型推理与闭环。将模型嵌入调度系统，对新的配送任务进行实时推理并生成指令。系统自动采集任务实际完成时间与状态，与预测结果比对，形成基础绩效闭环，用于监控与人工优化规则。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。除设备与任务数据外，同步集成激光雷达、视觉相机等多模态传感器数据，以及厂区电子地图、工位状态等环境数据。**数据预处理。**对多源感知数据进行时间同步、空间标定与坐标系统一；进行动态障碍物检测与跟踪，并将其转化为结构化的事件流数据。

②第二阶段：样本准备与特征工程

特征工程。构造复杂特征，如“动态交通密度热力图”“预测性通行耗时（基于历史拥堵模式）”“多车协同冲突概率”。**数据标注与增强。**对历史数据中因动态障碍导致的

路径重规划、交通拥堵等场景进行标注。使用仿真或数据合成技术增强各类罕见交通状况的样本。

③第三阶段：模型训练与仿真验证

模型训练。在数字孪生构建的虚拟厂区中，使用融合特征集，训练基于深度学习的动态路径规划模型和多车协同避撞算法。**模型验证。**在数字孪生环境中模拟高峰人流、临时占道等复杂场景，对模型策略进行大规模压力测试与鲁棒性验证，确保其安全性与效率。

④第四阶段：模型部署与闭环进化

模型推理与进化。将验证后的模型部署上线，实现实时环境感知下的动态调度与路径重规划。建立模型效果监控体系，将实际运行中遇到的新障碍模式作为反馈数据，驱动模型定期迭代更新。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。实现全链路数据毫秒级接入，包括生产计划与物料需求、全厂库存状态、所有移动单元（AGV/AMR/叉车）的细粒度状态、全局环境感知网络数据及跨区域传输信息。**数据预处理。**实施跨域数据实体关联（如将需求关联至具体物料、库位和可用运力），并进行超高维数据的实时降维与特征初筛，满足多智能体模型输入需求。

②第二阶段：样本准备与特征工程

特征工程。构建支撑全局博弈与优化的特征，如“基于需求预测的全局负载均衡度”“多智能体通信与协商有效性指标”“扰动传播影响链预测向量”。**数据标注与增强。**对历史大规模扰动事件（如全线急单、关键设备集群故障）进行根因与处置策略的序列标注。利用全域数字孪生，生成海量极端、随机和对抗性场景的仿真数据，构成深度强化学习的训练环境。

③第三阶段：模型训练与仿真验证

模型训练。在全局数字孪生中，采用多智能体深度强化学习等框架，让代表不同车间、工序、物流资源的智能体进行自主博弈、协商与协同训练，以优化全局履约率和资源利用率。**模型验证。**在孪生体中设计系统性极限测试，如模拟供应链断裂、订单组合剧变等，验证智能体集群的整体涌现智能、抗毁性与自恢复能力。

④第四阶段：模型部署与闭环进化

模型推理与决策。将训练成熟的“多智能体决策大脑”作为调度核心，实现面对动态需求与扰动的全自动、全局优化调度。**闭环进化。**建立双向数据-仿真驱动闭环：物理世界数据实时优化孪生模型；同时在孪生体中持续进行策略探索与离线进化，将更优策略无缝部署至实际系统，实现从数据到智能的永续进化。

（十）危险作业自动化

“危险作业自动化”场景涉及入门级、基础级与进阶级的企业。

入门级企业针对高危区域人工操作安全风险高、异常处置滞后等问题，部署搭载AI感知与决策技术的工业机器人，结合监控系统实现危险作业自动化执行、环境异常 AI 预警及远程干预，降低人员暴露风险并提升作业安全性时，数据治理的目标主要是搭建适配工业机器人与监控系统运行的标准化、高可靠基础数据体系，保障 AI 感知数据的精准性与预警决策数据的有效性，打通设备端与监控端的数据交互链路，为高危场景的自动化作业与安全预警筑牢数据屏障。

基础级企业针对复杂且环境存在变化的风险场景，引入AI驱动的多模态感知与远程精准操控技术，融合智能作业单元、AR/VR交互与数字孪生决策的自动化系统，通过AI分析实现环境动态识别、远程精细化操作与安全智能决策，提升危险作业的自动化水平与安全性时，数据治理的目标主要是构建多模态、高关联、可追溯的风险作业数据体系，消除感知、操控、决策环节的数据割裂，保障AI分析的精准性与数字孪生模型的可靠性，为动态环境下的安全作业与智能决策提供高质量数据支撑。

进阶级企业针对高度复杂、动态、不可预测危险场景中机器人自主性不足、协同能力弱、风险预判滞后的问题，引入多模态深度感知与群体智能决策技术，构建全链路自主运

行的智能作业系统,通过AI算法实现危险环境全自主认知、多机器人协同作业、预测性风险规避时，数据治理的目标主要是打造覆盖“感知-认知-决策-执行-反馈”全链路的高可信、强关联、自进化数据闭环体系，打通多模态感知数据与群体智能决策模型的深度耦合链路，保障多机器人协同的一致性与风险预判的前瞻性，为智能作业系统的全自主运行提供全维度、高质量的数据支撑。

1.治理对象

表 47　危险作业自动化场景的数据治理对象

适配层级	数据类型	数据内容
入门级、基础级、进阶级	环境感知数据	可见光/红外/热成像视频、3D 激光点云、毫米波雷达数据、声纹/超声波数据、有毒有害/易燃易爆气体浓度、温湿度、气压、辐射强度等。
入门级、基础级、进阶级	作业对象与过程数据	作业对象三维模型、作业规程、任务分解步骤、历史作业过程录像与日志、工艺参数等。
入门级、基础级、进阶级	机器人本体与协同数据	单机实时位姿、关节状态、电机电流/扭矩、电池电量、故障码；多机间通信消息、任务分配记录、相对位置、协作轨迹等。
基础级、进阶级	外部系统与上下文数据	来自 MES/ERP 的工单与资产信息；来自气象/地质部门的预警数据；厂区地图与基础设施 BIM 模型等。
基础级、进阶级	仿真推演与知识数据	高风险作业场景的仿真模型、历史事故案例报告、专家处置经验规则、风险预案库等。

2.平台（技术）工具

表 48　危险作业自动化场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	负责多源、高可靠性数据的采集、清洗、整合与质量管控，构建危险作业场景下统一可信的数据基座。	ThingsBoard、华为云 IoTDA、MQTT Broker	华为云 ROMA Connect、TDengine、阿里云 DataWorks	阿里云实时计算 Flink 版、华为云 DataArts Studio、Apache Pulsar
特征工程与样	提供针对危险场景的特征提取、构	Python、Jupyter Notebook、VoTT	CVAT、Scale AI、PyTorch/TensorFlow	英伟达 DALI、Snorkel Feast

本管理平台	造、安全标注与版本管理能力，将数据转化为模型燃料。			
模型开发与运维平台	支撑安全关键型 AI 模型的开发、训练、高可靠部署与严格生命周期管理。	ROS、腾讯云 TI-ONE、scikit-learn	PyTorch/TensorFlow 阿里云 PAI、MLflow	Ray RLlib、英伟达 Isaac Sim、Kubeflow
数字孪生与智能应用平台	构建高保真危险环境虚拟映射，用于安全策略仿真验证、协同训练及智能监控指挥。	组态王、ECharts、基础版三维场景查看器	Unity 3D 或 Unreal Engine、51WORLD 数字孪生平台、微软 Azure Digital Twins	英伟达 Omniverse、ANSYS Twin Builder、达索 3DEXPERIENCE

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过工业协议集成机器人的本体传感器数据（位姿、电流、关节扭矩）、固定点监控传感器数据（特定气体、火焰、红外）及安全联锁系统的状态信号。**数据预处理。**对传感器数据进行冗余校验与滤波，抑制干扰；统一报警事件编码与设备唯一标识；对关键安全信号实施毫秒级延迟监测与补偿。

②第二阶段：样本准备与特征工程

特征工程。构建“传感器读数距安全阈值的动态余量”“机器人运动状态突变指标”“历史同期风险事件发生概率”等基础安全特征。**数据标注与划分。**对历史监控录像与日志中的典型异常事件（如泄漏初期、人员误入）进行精确起止时间标注。按场景类型划分数据集。

③第三阶段：模型训练与仿真验证

模型训练与验证。使用标注样本，训练基于阈值的动态预警模型或简单的时序异常检测模型。在仿真或历史回放中严格验证其检测率、误报率与响应延迟，确保安全底线。

④第四阶段：模型部署与闭环进化

模型推理与闭环。将模型嵌入监控系统，对实时数据流进行在线推理，触发声光报警或自动急停。所有预警事件及其处置结果（包括误报）必须完整记录，形成安全事件案例库，用于定期复盘与规则优化。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。同步集成多源视觉(可见光、红外、热成像)、声纹/超声、激光雷达点云等感知数据，以及机器人的高精度位姿与力控数据、AR/VR 远程操控指令流。**预处理。**对多模态数据进行时空配准与融合（如图像与点云对齐）；对操控指令数据进行意图解析与标准化；清洗感知数据中的环境噪声（如光照变化、无关反射）。

②第二阶段：样本准备与特征工程

特征工程。构造复杂特征，如“多视角下的目标物三维姿态估计”“基于声纹的设备内部异常诊断特征”“操作者的意图与机器人状态匹配度”。**数据标注与增强。**对多模态数据中的风险源（如裂缝、腐蚀、高温点）进行像素级或点云级精细标注。使用生成对抗网络等技术合成不同光照、遮挡条件下的风险场景数据。

③第三阶段：模型训练与仿真验证

模型训练。利用融合特征集，训练多模态风险识别模型、机器人视觉伺服操控模型，并在数字孪生中训练人机协同安全决策算法。**模型验证。**在数字孪生环境中，模拟环境突变、操作延迟等场景，对整套系统的作业精度、安全冗余与鲁棒性进行系统化验证。

④第四阶段：模型部署与闭环进化

模型推理与进化。将模型部署为远程作业系统的辅助智能模块，提供实时环境识别结果与安全操作建议。建立专家评价系统，收集远程操作员的反馈评分与修正动作，驱动模型向更符合人类专家判断的方向迭代。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。实现所有作业机器人、无人机、固定感知节点的全量、原生多模态数据（高清视频流、原始点云、频谱数据）的实时接入，同步集成全局环境物理化学参数及历史

灾害案例数据。**数据预处理。**实施跨模态数据的特征级早期融合；对海量高维数据进行在线降维与关键信息提取，以满足群体智能模型的实时处理需求；构建全局一致的世界动态模型。

②第二阶段：样本准备与特征工程

特征工程。构建支撑群体博弈与预测的特征，如“多机器人感知置信度与信息熵融合特征”“风险传播路径与影响范围预测向量”“群体任务分配与冲突消解效用函数”。**数据标注与增强。**对极少发生的重大连锁故障或灾难性场景进行全链路因果标注。利用高保真物理仿真[54]引擎，生成海量包含非线性相互作用、极端条件演变的灾难推演数据，作为核心训练资源。

③第三阶段：模型训练与仿真验证

模型训练。在超实时数字孪生中，采用多智能体深度强化学习、模仿学习[55]与因果推理结合的框架，使机器人群体学会在未知风险下协同探索、主动避险、互助救援等高级策略。**模型验证。**在孪生体中设计“最坏情况”对抗性测试，如模拟感知全面退化、通信中断、突发连锁反应等，验证群体智能的极端环境生存能力、自主重构能力与安全优先的决策逻辑。

④第四阶段：模型部署与闭环进化

模型推理与决策。将训练成熟的“群体智能大脑”作为系统核心，实现面对不可预测风险的全自主作业调度与安全

控制。**闭环进化。**建立“平行系统”级进化机制：真实作业数据用于校准孪生模型；同时，一个始终在孪生体中并行运行的“影子”智能体集群，持续进行超越当前安全规程的策略探索与压力测试，其产生的经过充分验证的“超视距”安全策略可审慎导入实际系统，实现安全能力的超前进化。

（十一）安全一体化管控

“安全一体化管控”场景涉及基础级与进阶级的企业。

基础级企业针对生产现场安全监控碎片化、风险识别滞后、应急响应低效等问题，引入 AI 驱动的多模态数据融合与智能决策技术，构建覆盖全流程的安全一体化管控系统，通过深度 AI 分析实现风险动态感知、智能预测与协同处置，提升安全管控的整体性与前瞻性时，数据治理的目标主要是构建全流程贯通、多模态融合、高可信度的安全管控数据体系，保障 AI 算法对风险的动态识别精度与预测能力，为安全一体化管控系统的智能决策与协同处置提供全维度数据支撑。

进阶级企业针对复杂场景下隐性风险识别难、应急处置效率低的问题，引入多模态 AI 感知与全域协同决策技术，构建融合深度学习、知识图谱与数字孪生的安全一体化智能管控平台，通过 AI 分析实现风险全域感知、隐患精准预警、应急自主处置与全流程智能闭环，提升安全管控的前瞻性与精准性时，数据治理的目标主要是打造“全域覆盖、深度关联、高保真映射、自进化迭代”的安全数据生态体系，打通

多模态感知数据与智能决策模型的深度耦合链路，保障隐性风险特征的精准提取与数字孪生场景的实时推演，为安全管控平台的自主闭环运行提供全维度、高价值的数据支撑。

1.治理对象

表 49　安全一体化管控场景的数据治理对象

适配层级	数据类型	数据内容
基础级、进阶级	多源感知数据	工业相机捕捉的现场图像视频、传感器采集的环境参数与设备状态数据、智能穿戴设备记录的人员状态信息等
基础级、进阶级	风险与事件数据	风险识别记录、隐患信息、安全事件详情、处置过程记录等
基础级、进阶级	知识与方案数据	安全知识图谱数据、历史风险案例、应急处置方案、设备安全参数等
进阶级	时间序列数据	设备状态变化趋势、环境参数演变数据等
进阶级	数字孪生数据	风险演化模拟结果、处置方案预演数据等
进阶级	跨系统关联数据	设备控制系统信息、人员管理数据等

2.平台（技术）工具

表 50　安全一体化管控场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
数据治理与集成平台	负责多源、多模态安全数据的采集、融合、治理与可信存储，构建统一、高质量的安全数据基座。	华为云 DataArts Studio、阿里云 DataWorks、Apache NiFi	阿里云实时数仓 Hologres、Cloudera Data Platform、Confluent Platform
特征工程与样本管理平台	提供从原始安全数据到模型特征与训练样本的加工、标注与管理能力，提炼风险信号。	Label Studio Amazon SageMaker Ground Truth、Python Pandas/Scikit-learn	Feast、Tecton、Snorkel
模型开发与运维平台	支撑安全 AI 模型的开发、训练、评	百度 PaddlePaddle、	英伟达 NGC、Ray Domino Data Lab

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
	估、部署与全生命周期管理，实现安全可靠的MLOps。	MLflow、TensorFlow Extended	
数字孪生与智能应用平台	构建高保真安全态势虚拟映射，用于风险推演、预案仿真、智能决策与协同指挥。	51WORLD 数字孪生平台 优锘科技 ThingJS Unity Industrial	达索 3DEXPERIENCE 英伟达 Omniverse ANSYS Twin Builder

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过企业服务总线与物联网平台，整合来自视频监控、烟感/温感等消防传感器、有毒有害气体探测器、门禁系统、人员定位卡、电子巡检记录及 MES 工单的多维数据。**数据预处理。**统一各系统时间戳；对视频流进行抽帧与关键事件切片；对传感器数据进行阈值滤波与缺失值插补；将非结构化的巡检日志文本进行关键信息（地点、问题、状态）结构化提取。

②第二阶段：样本准备与特征工程

特征工程。构造直观有效的特征，如“区域人员密度与动态聚集趋势”“特定作业与危险气体浓度的时空关联性”“设备高温点与周边易燃物距离”。**数据标注。**组织安全专家对历史视频片段、传感器报警记录进行复核，标注“真实火情”“违规闯入”“安全防护设备未正确使用”等正负样

本。**数据增强与划分。**对少数类别风险事件样本，通过加噪、仿射变换（对图像）等方法进行增强。按时间或厂区区域划分训练集与测试集，防止数据泄漏。

③第三阶段：模型训练与仿真验证

模型训练。使用标注样本，分别训练视频异常行为识别模型、多传感器融合的早期火灾预警模型、基于时序数据的区域风险等级预测模型。**模型验证。**在独立的测试集上验证各模型性能（如精确率、召回率）。通过数字孪生进行初步场景复现，验证预警信息是否能准确触发预设的应急联动规则（如报警、通知责任人）。

④第四阶段：模型部署与闭环进化

模型推理。将训练好的模型部署为微服务集群，对实时数据流进行并行推理，并将风险预警事件推送至统一安全管控平台。**闭环进化。**建立处置反馈流程，安全员在平台中确认预警、记录处置结果。这些“预警-处置-结果”配对数据定期回流，用于模型的增量学习与优化，降低误报，提升针对新型风险的识别能力。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。在基础级之上，深度集成设备内部传感器高频数据、工业控制系统状态数据、全员生物识别与行为日志、供应链与维保记录，并全部接入数字孪生平台。

数据预处理。对多模态数据进行特征级融合[56]与时空对齐（如将人员精确位置叠加到设备三维模型上）；利用知识图谱对零散的安全隐患描述进行实体与关系抽取，构建结构化知识。

②第二阶段：样本准备与特征工程

特征工程。构造深度关联与因果推断特征，如“基于知识图谱的‘设备老化-工艺参数偏离-泄漏风险’传导链特征”“人员疲劳指数与操作失误率的关联特征”“多系统日志联合分析下的异常操作序列模式”。**数据标注。**对难以直接观测的“隐性风险”进行间接标注，例如，将“事后确认的重大未遂事故”所对应事前一段时间的所有多模态数据进行序列标注。**数据增强。**利用数字孪生与物理仿真引擎，主动生成大量极端、罕见、复杂的“黑天鹅”事件场景数据（如多米诺骨牌效应、隐蔽空间泄漏扩散），作为核心训练资源。

③第三阶段：模型训练与仿真验证

模型训练。采用图神经网络学习安全知识图谱中的复杂关系；采用时空序列预测模型学习风险演化规律；在孪生环境中，采用深度强化学习训练“风险感知-预案生成-资源调度”的端到端应急决策模型。**模型验证。**在数字孪生中构建“平行安全系统”，将训练好的模型置于其中，与基于传统规则的模型进行全天候、高并发的对抗性推演与压力测试，验证其在复杂不确定环境下的决策优越性、鲁棒性与安全性。

④第四阶段：模型部署与闭环进化

模型推理与决策。将经过严苛验证的自主决策模型作为平台“智能内核”，实现从风险预警到自动启动应急预案、调度应急资源、引导人员疏散的闭环。**闭环进化。**建立“四驱”自进化引擎：真实数据持续优化感知模型；处置效果数据优化决策模型；新案例与法规反哺知识图谱；平行系统中的持续博弈探索产生超越现有经验的前瞻性安全策略，经审批后导入实际系统，实现安全管控能力的自主迭代与超越。

（十二）能源智能管控

“能源智能管控”场景涉及基础级与进阶级的企业。

基础级企业针对工厂能耗波动大、多能源协同低效、绿电利用率低等问题，引入 AI 驱动的智能分析与优化技术，构建覆盖全流程的能源智能系统，通过深度 AI 分析实现能耗精准预测、多能源协同调度与动态优化，提升能源利用效率时，数据治理的目标主要是构建标准化、高关联、时序化的全流程能源数据体系，保障 AI 分析模型的输入质量与决策有效性，为能耗预测、协同调度提供可靠的数据支撑。

进阶级企业针对工厂多能源协同难、供需失衡、能效低的问题，引入多模态预测与深度强化学习技术，构建全链路自主运行的智能能源综合管控平台，通过 AI 算法实现多能源协同优化、供需动态平衡与全周期能效提升时，数据治理的目标主要是打造覆盖“能源生产-传输-存储-消耗-再生”全链路的高可信、强关联、自进化能源数据生态体系，打通多

能源类型、多系统链路的数据深度耦合链路，保障多模态预测模型与深度强化学习算法的输入质量，支撑平台对能源供需的精准研判与全域协同优化决策，为全周期能效提升筑牢数据根基。

1.治理对象

表 51　能源智能管控场景的数据治理对象

适配层级	数据类型	数据内容
基础级、进阶级	多能源基础数据	电力、燃气、水等各能源介质的计量数据、管网参数、设备能耗参数等
基础级、进阶级	生产关联数据	生产计划、工序能耗需求、设备运行状态等
进阶级	环境与市场数据	气象数据、可再生能源发电曲线、能源市场价格波动等
进阶级	调度与优化数据	能源调度策略、供需平衡记录、梯级利用路径、优化结果等
基础级、进阶级	异常与历史数据	异常能耗记录、处置方案、历史能耗数据、改进经验等
基础级、进阶级	数字孪生数据	全能源系统模拟数据、极端场景推演结果、调度策略预演记录等
进阶级	跨系统数据	能源管理系统信息、设备控制系统数据等

2.平台（技术）工具

表 52　能源智能管控场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
数据治理与集成平台	负责多源异构能源数据（电、气、热、生产）的采集、清洗、关联与可信存储，构建高质量能源数据基座。	阿里云 DataWorks、华为云 DataArts Studio、TDengine	Apache Doris、Cloudera Data Platform、Timecho DB
特征工程与样本管理平台	提供从原始数据到模型特征与训练样本的加工、标注与	Amazon SageMaker Data Wrangler、Jupyter Notebook+Pandas/	Feast、Tecton、Snorkel

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
	管理能力，提炼能源优化信号。	Scikit-learn、Label Studio	
模型开发与运维平台	支撑能源预测与优化模型的开发、训练、评估、部署与全生命周期管理，实现可靠 MLOps。	百度 PaddlePaddle、阿里云 PAI、MLflow	Ray RLlib、英伟达 AI Enterprise、Kubeflow
数字孪生与智能应用平台	构建高保真能源系统虚拟映射，用于策略推演、仿真优化、智能调度与全景监控。	51WORLD 数字孪生平台、优锘科技 ThingJS、图扑软件 HT	达索 3DEXPERIENCE、ANSYS Twin Builder、DataMesh Director

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过能源管理系统或物联网平台，整合智能电表、燃气表、水表等计量数据、主要用能设备（空压机、HVAC）的运行状态与功率数据，以及基础生产计划数据。**数据预处理。**统一各计量设备数据采集频率与时间戳；对数据进行清洗与插补，处理通信中断导致的缺失值；剔除明显异常值；将不同能源单位统一折算为标准煤或等效电耗。

②第二阶段：样本准备与特征工程

特征工程。构造直观有效的业务特征，如“单位产品能耗”“分时电价下的用电成本曲线”“设备负载率与能效关联指标”“基于天气预报的制冷/热负荷预测特征”。**数据标注。**结合生产日志与运维记录，标注历史数据中的“非生产性能耗尖峰”“设备低效运行时段”等。**数据增强与划分。**

对不同季节、不同生产模式下的能耗曲线进行数据增强。严格按时序划分数据集，防止未来信息泄露，确保模型泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。使用标注后的时序特征，训练多元时序预测模型（如用于预测未来24小时全厂用电负荷），以及基于规则的或简单优化的调度策略模型（如在高电价时段建议启用备用储能）。**模型验证。**在测试集上评估预测模型的误差率。通过模拟仿真，验证调度策略在历史数据上的理论经济性与可行性，并与实际历史操作对比。

④第四阶段：模型部署与闭环进化

模型推理。将预测模型部署，每日自动生成能耗预测报告；将优化模型以“专家建议”形式嵌入能源管理系统，供调度人员参考执行。**闭环进化。**系统自动对比预测值与实际值、建议策略与实际执行结果的偏差，将偏差数据作为新的训练样本，定期启动模型的滚动训练与更新，使模型持续适应生产变化。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。在基础级之上，深度集成分布式光伏/风电的发电预测与实时出力数据、储能系统（SOC、充放电功率）的毫秒级状态数据、实时电力市场交易价格数据、全厂生产计划与柔性负荷可调数据，并全部接入能源数字孪生平台。

据预处理。实现多源异构数据的“五维”统一（时间、空间、量纲、品质、语义）。利用因果分析初步探索生产扰动与能源波动的关联关系，为特征工程提供方向。

②第二阶段：样本准备与特征工程

特征工程。构造全局优化与深度博弈特征，如“多能源耦合转换效率矩阵特征”“基于强化学习的‘状态–动作–价值’特征”“市场风险收益与物理安全约束的权衡特征”“碳排放在线核算与影子价格特征”。**数据标注。**对历史成功/失败的复杂调度案例进行全程决策序列与最终效果标注，作为模仿学习的样本。**数据增强。**利用高保真数字孪生与物理仿真，主动生成海量极端场景数据（如极端天气导致新能源出力剧变、主力设备突发故障、市场价格剧烈波动），为深度强化学习提供充足的“探索”环境。

③第三阶段：模型训练与仿真验证

模型训练。采用多模态融合的时序预测模型（融合气象、市场、生产数据）精确预测供需。在孪生环境中，采用多智能体深度强化学习，训练以全局经济性、碳排、安全为综合目标的协同调度主模型[57]。**模型验证。**在能源数字孪生中构建“平行仿真系统”，将训练好的策略模型与多种基准策略（如传统规则、人工经验）进行长期、多场景的对抗性推演与压力测试，从统计学上验证其优越性、鲁棒性与风险抵御能力。

④第四阶段：模型部署与闭环进化

模型推理与决策。将通过严苛验证的模型作为平台核心决策引擎，实现从分钟级到日前级的自动优化决策、实时控制指令下发与市场自动报价。**闭环进化。**建立“双环”进化机制：内环基于实时反馈数据对预测与决策模型进行在线微调；外环在平行仿真系统中持续进行策略探索与前沿优化，寻找超越当前策略的更优解，经安全评估后注入生产系统，实现能源管控系统的永续自主进化。

（十三）碳资产全生命周期管理

“碳资产全生命周期管理”场景仅涉及进阶级的企业。

进阶级企业针对碳资产全生命周期管理场景，面向碳排放数据采集、碳足迹追踪和碳资产核算等业务活动，围绕碳排放计量难、碳足迹追踪效率低等问题，建立 AI 数字化碳管理系统，应用碳排放精细化检测、碳排放指标自动核算、碳捕获利用与封存等技术，实现碳的追踪、分析、核算和交易，挖掘碳资产利用价值，降低单位产值碳排放量时，数据治理的目标主要是构建覆盖“碳产生-碳计量-碳追踪-碳核算-碳交易”全生命周期的标准化、高可信、可溯源碳数据体系，打通碳数据与能源数据、生产数据、供应链数据的深度关联链路，保障 AI 数字化碳管理系统对碳信息的精准解析与智能决策，为碳资产精细化运营、合规核算及价值挖掘提供全维度高质量数据支撑。

1.治理对象

表 53　碳资产全生命周期管理场景的数据治理对象

适配层级	数据类型	数据内容
进阶级	碳监测数据	多类型碳监测设备采集的碳排放实时数据、激光光谱与红外传感融合数据等
进阶级	全生命周期碳足迹数据	原材料开采、生产加工、物流运输、废弃物处理等各环节碳排放数据
进阶级	生产与能源数据	生产计划、设备运行参数、能源消耗数据、碳捕集设备状态等
进阶级	碳交易与资产数据	碳价波动记录、碳资产持有量、交易记录、政策合规数据等
进阶级	数字孪生数据	碳流动模拟数据、减排策略预演结果、碳价波动影响推演记录等
进阶级	历史与优化数据	历史碳数据、减排方案及效果、自进化模型迭代记录等

2.平台（技术）工具

表 54　碳资产全生命周期管理场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级企业
数据治理与集成平台	负责全域数据接入与融合、数据质量管控与数据资产目录管理。	Apache NiFi、Talend、阿里云 DataWorks Informatica MDM、Collibra、华为云 DataArts Studio、InfluxDB、TDengine
特征工程与样本管理平台	负责专业特征构建、数据标注与管理、数据增强、数据集划分与版本控制。	Feathr、Amazon SageMaker Feature Store Label Studio
模型开发与运维平台	负责多模态模型开发、强化学习训练、仿真验证与合规测试、模型全生命周期管理。	Databricks、Azure Machine Learning、百度 PaddlePaddle、Ray RLlib, OpenAI Gym MLflow, MLflow Models

数字孪生与智能应用平台	负责碳流程数字孪生、智能应用部署、闭环反馈与进化、决策可视化与协同。	ANSYS Twin Builder、微软 Azure Digital Twins、树根互联根云 Mendix、OutSystems、Tableau、Power BI、阿里云 DataV、Apache Kafka

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过物联网平台与系统接口，全域集成三类核心数据：一是直接监测数据，包括在线监测系统的实时排放浓度、流量，碳捕集设施运行参数。二是间接核算数据，包括能源管理系统的各类能源消耗明细；物料管理系统的原料投入与产品产出数据。三是关联数据，包括供应链环节的碳足迹数据、生产计划与工艺数据。**数据预处理。**对监测数据进行异常值识别、剔除和基于标准物质的校准追溯；对核算数据执行逻辑一致性校验（如能量/物料平衡）。统一各类数据的时间戳、计量单位及排放因子引用来源。依据国家核算标准，对活动水平数据进行分类编码。

②第二阶段：样本准备与特征工程

特征工程。构建核算驱动特征，如“单位产品全流程碳强度”“基于实时工况的碳排放流率”“能源-碳排关联弹性系数”；构建交易与资产特征，如“碳配额履约风险敞口”“碳资产预期收益曲线”“碳成本内部转移定价”。**数据标注。**组织碳管理专家，对历史数据进行多维标注，标记监测

数据异常、核算数据偏差等；标注历史成功减排案例对应的工况与操作策略；标记核算报告中易受核查关注的关键数据点及支撑证据链。**数据增强与划分。**针对碳交易、碳捕集等小样本场景，利用数字孪生模拟不同市场政策、技术条件下的数据，进行合理增强。按时间序列划分数据集，保证时序模型的训练有效性。

③第三阶段：模型训练与仿真验证

模型训练。训练预测模型，使用特征数据，训练多模态碳排放预测模型（融合生产计划、能源消耗、实时监测）；训练优化模型，利用强化学习，在数字孪生中训练“碳成本最小化”或“碳资产价值最大化”的全局优化决策模型，动态调整生产与能源调度。**仿真验证。**在孪生环境中，模拟不同核算边界与方法学下的碳排放结果，验证 AI 核算引擎的合规性与准确性。模拟碳价剧烈波动、核查政策变更等极端场景，测试优化模型的鲁棒性、风险抵御能力与经济性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将验证后的模型部署为碳管理系统的核心引擎，实现碳排放的实时监测、自动核算、履约预警。模型对生产与能源系统提供低碳优化调度建议，并与碳交易平台接口联动，支持自动化交易决策辅助。**闭环运营与进化。**持续采集模型决策产生的实际效果数据，如减排措施的实际碳减排量、碳交易的实际收益/成本、第三方核查的反馈意见。将反馈数据作为新的黄金样本，回流至数据湖，驱动特征集

更新、模型迭代优化，并反哺数字孪生模型，使其更贴近现实，形成“监测-核算-决策-交易-反馈-优化”的碳数据价值闭环。

（十四）污染在线管控

“污染在线管控”场景涉及基础级与进阶级的企业。

基础级企业针对生产过程中污染排放溯源难、处理设施协同低效、超标预警滞后等问题，引入 AI 驱动的智能监测与优化技术，构建覆盖全流程的污染在线管控系统，通过 AI 分析实现污染物精准识别、处理过程动态优化与排放趋势预测，提升污染管控的精准性与前瞻性时，数据治理的目标主要是构建标准化、高可信、全流程贯通的污染管控数据体系，保障 AI 分析模型对污染物的识别精度与趋势预测能力，为污染在线管控系统的精准管控与动态优化提供可靠数据支撑。

进阶级企业针对复杂生产场景下污染物成分复杂、溯源困难、处理低效的问题，引入多模态 AI 感知与深度强化学习技术，构建全链路智能管控平台，通过 AI 算法实现污染物精准识别、污染源快速定位、处理过程自适应优化与风险提前预警时，数据治理的目标主要是打造覆盖“污染产生-多模态监测-精准溯源-智能处理-风险预警”全链路的高保真、强关联、自进化污染管控数据体系，打通多模态感知数据与深度强化学习模型的深度耦合链路，保障复杂场景下污染物

识别的精准度与污染源溯源的高效性，为全链路智能管控平台的自主决策与自适应优化提供全维度高质量数据支撑。

1.治理对象

表 55　污染在线管控场景的数据治理对象

适配层级	数据类型	数据内容
基础级、进阶级	污染物监测数据	光谱、色谱等多模态分析数据、传感器采集的污染物浓度与成分数据、实时监测记录等
基础级、进阶级	生产关联数据	生产工艺参数、设备运行状态、原材料特性、工序操作记录等
基础级、进阶级	污染源与溯源数据	污染物排放特征、污染源定位信息、影响因素关联数据等
基础级、进阶级	处理设施数据	处理设备运行参数、处理效果记录、优化调整方案等
基础级、进阶级	环境与扩散数据	气象条件、地形环境、污染物扩散模拟数据等
基础级、进阶级	应急与预警数据	风险预警记录、应急减排方案、处置效果评估等
基础级、进阶级	历史与优化数据	历史污染数据、处理工艺改进案例、自进化模型迭代记录等

2.平台（技术）工具

表 56　污染在线管控场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
数据治理与集成平台	负责多源异构污染数据的采集、融合、治理与可信存储，构建统一、高质量的污染管控数据基座。	Apache NiFi、阿里云 DataWorks、InfluxDB	Cloudera Data Platform、Confluent Platform、华为云 DataArts Studio
特征工程与样本管理平台	提供从原始数据到模型特征与训练样本的加工、标注、增强与管理能	Label Studio（开源多功能数据标注平台）Python Pandas/Scikit-learn	Feast、Tecton、Snorkel

支撑平台/工具	核心功能	工具示例	
		基础级企业	进阶级企业
	力,提炼污染风险与优化信号。		
模型开发与运维平台	提供模型开发、训练、验证、部署与运维的全生命周期管理环境，是AI 能力的核心生产线。	MLflow、Azure Machine Learning、H2O.ai	Ray & RLlib、Kubeflow Domino Data Lab
数字孪生与智能应用平台	构建污染管控流程的虚拟映射,支撑仿真推演、智能决策应用开发与闭环进化,是智能落地与价值呈现的载体。	Grafana、ThingsBoard、组态王/力控	ANSYS Twin Builder、微软 Azure Digital Twins、西门子 Process Simulate

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。以企业排污许可证信息为统一索引，全面集成三类核心数据源：一是末端监测数据，包括废气、废水排口的在线监测系统实时浓度与流量数据，确保数据采集与传输符合《污染物自动监测监控系统数据传输技术要求》（HJ 212—2025）等国家标准规范；二是过程工况数据，采集关键生产设备与污染治理设施（如风机、水泵、加药系统）的运行状态、启停信号与关键工艺参数（如 pH 值、压力）；三是基础管理数据，整合生产台账、原辅材料消耗记录及简单的能源消耗数据。**数据预处理。**对在线监测数据进行标准化清洗，依据仪器量程与历史规律剔除明显异常值，对通信中

断造成的缺失值进行标记与插补。统一所有数据源的时间戳至秒级，并强制关联至对应的污染治理设施与生产批次。通过逻辑规则校验（如治理设施运行状态与排放浓度的关联性）初步识别可疑数据。

②第二阶段：样本准备与特征工程

特征工程。基于业务逻辑构建可直接反映管控效果的特征指标，如“单位产品污染物产生强度”“治理设施同步运行率”“关键工艺参数与排放浓度的关联曲线”。引入外部关联特征，如利用治污设施用电监控数据，构建“设施能耗—处理效率”关联特征，为非现场监管提供辅助判断。**数据标注。**组织环保工艺工程师与运维人员，结合历史巡检记录、手工监测报告与运维工单，对历史数据集进行关键标注：标注“排放浓度异常波动时段”“治理设施非计划停运或低效运行事件”“因原辅料或工况变化导致的特征污染因子变化”等。**数据增强与划分。**针对季节性生产或不同产品批次导致的排放模式差异，采用周期叠加、小幅扰动等方法进行数据增强，丰富模型见过的场景。严格按时间先后顺序划分训练集、验证集与测试集，严防未来信息泄露，确保模型对时序规律的泛化学习能力。

③第三阶段：模型训练与仿真验证

模型训练。使用标注后的时序特征数据，训练以预测为核心的基础模型。包括：训练多元时序回归模型，用于预测未来数小时至一天的污染物排放浓度趋势；训练基于规则的

分类模型，用于识别排放异常模式（如连续超标、浓度骤升）并触发预警。**模型验证。**在独立测试集上评估预测模型的平均绝对误差、均方根误差等指标。通过历史事件回放，验证异常识别模型的准确率与召回率，确保其能有效发现已知类型的异常工况，避免误报对生产造成干扰。

④第四阶段：模型部署与闭环进化

模型推理。将验证合格的预测与异常识别模型部署至污染在线管控系统，实现排放浓度的实时展示、短时趋势预测与超标预警信息自动推送。优化建议以“操作提示”形式推送至中控室，辅助人工决策。**闭环进化。**系统自动记录模型预警与实际人工确认或处置结果，定期将产生偏差的案例作为新样本，回流至训练流程，对模型参数进行微调更新，使其持续适应生产工艺的缓慢变迁。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。在基础级数据底座上，进行深度与广度拓展。深度集成多模态监测数据：包括污染物组分谱图数据、监测传感器网络数据、定期巡检获得的空间分布数据等。广度融合关联数据：集成全流程物料与元素平衡数据、高清视频监控智能流、实时气象微站数据等。所有数据统一接入企业级环境数字孪生平台。**数据预处理。**对光谱、图像等非结构化数据，进行降噪、校准与特征初提取。构建以“排放源-管网

-处理设施-排口”为核心拓扑的时空关联图谱，将离散数据点映射为图谱节点与关系，为后续的智能溯源奠定基础。

②第二阶段：样本准备与特征工程

特征工程。构建面向深度分析与自主决策的高阶特征。包括：多模态融合特征，如“光谱特征与浓度序列的关联向量”“视频烟雾纹理与 VOCs[58]排放的映射特征”；溯源推理特征，如“基于风场、扩散模型的反向贡献度特征”“管网节点水质突变传导关系特征”；优化决策特征，如“多污染因子协同去除的帕累托前沿特征”“基于强化学习的‘环境状态-调控动作-长期成本/收益’价值特征”。**数据标注。**对历史重大污染事件或成功优化案例进行全链路、细颗粒度等进行标注，标注内容包括异常事件的初始触发点位与扩散路径、专家处置过程中的决策逻辑与权衡考量等。**数据增强。**充分利用高保真数字孪生平台，基于物理化学机理模型[59]，主动生成海量在现实中难以获得或成本高昂的样本数据。特别是模拟极端工况、设备突发故障、多种污染物异常耦合等小概率高风险场景，为强化学习模型提供充足的“探索-利用”训练环境。

③第三阶段：模型训练与仿真验证

模型训练。采用多模态融合深度学习模型（如图神经网络与时序卷积网络结合），实现对复杂污染物来源与构成的精准识别与定量解析。在数字孪生环境中，采用多智能体深度强化学习，训练以“综合合规成本最小化”（涵盖能耗、

药耗、碳排、风险）为目标的全局协同优化主模型，该模型能自主生成从生产源头调优到末端治理的协同调控策略。**模型验证。**在数字孪生环境中构建“平行仿真与对抗测试系统”。将训练好的决策模型与多种基准策略（如经典优化算法、专家规则库、历史操作模式）进行长期、多回合的对抗性推演。在虚拟环境中模拟碳市场政策突变、新增特别排放限值、原材料剧变等压力场景，从统计学上综合评估优化模型的鲁棒性、经济性优越性及风险抵御能力，确保其决策优于现有最佳实践。

④第四阶段：模型部署与闭环进化

模型推理与决策。将通过严苛验证的智能模型作为全链路智能管控平台的“核心决策引擎”。该引擎能够实现：分钟级响应的污染源快速定位与溯源报告自动生成；小时级或班次级的处理工艺参数自适应优化与设定值下发；对市场风险与政策变化做出前瞻性响应的战略级减排路径规划。**闭环进化。**基于实时反馈的监测结果与成本数据，对预测与优化模型进行在线微调与校准。在平行仿真系统中，持续引入新的工艺知识、政策约束与优化目标，让 AI 进行超越当前策略的前沿探索，寻找潜在更优解。经安全评估与合规审查后，将稳定的新策略注入生产系统，推动污染管控系统实现永续自主进化。

（十五）柔性产线快速换产

“柔性产线快速换产”场景仅涉及进阶级的企业。

进阶级企业针对多品种混线生产中个性化需求响应慢、换产协同低效的问题,引入多智能体自主决策与生成式 AI 工艺重构技术，构建全链路 AI 驱动的柔性换产系统，实现产线不停机自主换产、工艺参数自优化与全域协同时，数据治理的目标主要是构建一个“时-空-质”三元统一、全要素实时映射、知识自演化驱动的“数据-决策”一体化基座，通过海量多模态数据在“订单-物料-设备-工艺-人员”复杂网络间无损、无感、无缝流转与深度耦合，为多智能体的分布式协同决策提供一致、可信的全局态势感知，并为生成式 AI 的工艺创新提供高质量、强关联、可回溯的领域知识燃料，最终支撑换产系统实现从“感知-响应”到“预测-重构”的根本性跨越。

1.治理对象

表 57　柔性产线快速换产场景的数据治理对象

适配层级	数据类型	数据内容
进阶级	产线与设备数据	智能装备运行状态、模块化加工中心参数、柔性输送线调度信息、设备通信网络拓扑数据等
进阶级	工件与产品数据	工件三维特征、装配公差、新产品参数、个性化需求指标等
进阶级	换产流程数据	换产任务分配、进度记录、工艺路径、参数调整记录等
进阶级	感知与检测数据	工业相机图像、力传感器数据、激光扫描结果、质量检测数据等
进阶级	数字孪生数据	换产全流程预演记录、时序仿真结果、潜在问题修正方案等
进阶级	知识与历史数据	工艺知识图谱、历史换产案例、自进化模型迭代记录、换产优化经验等

2.平台（技术）工具

表 58　柔性产线快速换产场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级企业
数据治理与集成平台	负责“订单-物料-设备-工艺-人员”复杂网络中海量多模态数据的实时、无损采集、融合治理与可信存储，构建全域实时映射的“数字线程”。	Apache Kafka、Apache Flink、Cloudera Data Platform
特征工程与样本管理平台	将原始数据加工为驱动多智能体协同决策与生成式 AI 工艺重构的高维特征与知识样本，是提炼“数据燃料”与“知识图谱”的核心。	Feast、Neo4j、Snorkel
模型开发与运维平台	提供多智能体强化学习与生成式 AI 模型的开发、训练、验证与全生命周期管理环境。	Ray & RLlib、Kubeflow、PyTorch Lightning / Hugging Face
数字孪生与智能应用平台	构建高保真虚拟产线，实现“预测-重构”级仿真推演，并承载智能应用完成自主决策。	英伟达 Omniverse、微软 Azure Digital Twins、Grafana

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。建立全域数据接入中枢，集成五维核心数据流。一是“需求-订单”数据流，实时接入 APS、MES 及个性化订单平台的动态需求、产品谱系与交期约束。二是“物料-物流”数据流，通过物联网平台与 RFID 技术，采集原材料、在制品、工装夹具的实时身份、位置、状态与流转路径。三是“设备-控制”数据流，深度连接 PLC、CNC、机器人控制器，毫秒级采集设备运行参数、伺服状态、换产动作序列与

故障代码。四是“工艺-质量”数据流，集成 CAPP 系统、工艺知识库、在线检测设备的工序参数、质量标准与实时检测结果。五是“人员-协同”数据流，记录操作员数字终端交互日志、AR 辅助作业数据及跨部门协作通讯关键信息。**数据预处理。**实施“时-空-质”三元统一工程。时间统一：以高精度网络时间协议为基准，对齐所有数据源的时间戳至毫秒级，并建立严格的事件时序逻辑。空间统一：构建车间级统一坐标系，将设备位置、物料点位、AGV 轨迹等空间信息进行映射与关联。质量统一：对多源数据进行即时清洗，利用基于工艺规则与统计模型的方法，识别并处理传感器漂移、通信中断导致的异常值与缺失值；对所有工艺参数、计量单位进行标准化转换与语义对齐，确保数据在不同系统间的一致性与可比性。

②第二阶段：样本准备与特征工程

特征工程。构建多层次、高维度的特征体系。提炼如“产线整体换产就绪度指数”“多订单混排下的资源冲突热力图”“基于实时物流效率的瓶颈预警指标”等全局态势特征；为调度、工艺、设备等智能体，构造如“设备 Agent 的本地效能与全局贡献度关联特征”“物料 Agent 的履约路径优化空间特征”“基于博弈论的资源竞合关系特征”等多智能体协同特征；为生成式 AI 构建深度特征，如“产品设计特征-历史工艺参数-加工质量”的隐式关联向量“工艺约束与优化目标（如节拍、能耗）的权衡图谱”等。**数据标注与增强。**专

家组织工艺专家、设备专家与生产调度专家，对历史成功/失败的换产案例进行多维度标注；利用高保真生产系统数字孪生，主动生成海量在现实生产中难以获取或成本极高的样本，特别是模拟极端订单组合、紧急插单等小概率高挑战场景下的换产过程数据，丰富训练样本的多样性与复杂性。**数据集划分。**采用“时空分层抽样”策略划分训练集、验证集与测试集。严格保证时间序列的连续性不被破坏，并确保各类换产模式、产品族系、季节周期等因素在数据集中分布均衡，防止模型过拟合，保障其在未知场景下的泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。实施双引擎驱动训练。采用多智能体深度强化学习或基于博弈论的协同算法，在数字孪生环境中训练全局调度优化主模型。各智能体（调度、设备、工艺等）通过与环境及其他智能体持续交互，学习以实现“总换产时间最小化”“综合成本最优”为目标的协同策略；训练基于Transformer 或扩散模型等架构的生成式 AI 模型。利用“产品特征-工艺参数-加工效果”的关联样本，使其学习工艺设计的深层规律与约束，能够针对新产品或新需求，自动生成或优化推荐可行的、高质量的工艺参数方案。**模型验证。**在数字孪生环境中构建“平行仿真验证体系”。将训练好的协同决策模型与生成式工艺模型进行联合测试。通过模拟不同生产场景、市场扰动和内部异常，从统计学上全面评估决策有效性、系统鲁棒性、工艺创新可靠性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将通过验证的 AI 模型以“云-边-端”协同架构进行部署。协同决策模型作为“中枢大脑”，集成至柔性换产指挥系统，实时下发调度指令；生成式工艺模型作为“工艺参谋”，嵌入工艺设计系统，提供实时优化建议。系统具备分钟级甚至秒级的实时推理与响应能力，直接或通过确认后驱动产线执行自主换产。**闭环运营与进化。**系统自动采集每次实际换产的全链路执行数据、操作员反馈及与模型决策的差异。利用这些在线反馈数据，对模型进行快速微调与校准，适应产线的慢时变[60]特性。在平行数字孪生系统中，持续利用积累的新数据与先进的算法，进行新策略、新工艺的探索性训练与仿真测试，寻找超越当前最优解的方案。经安全与性能评估后，将成熟的策略注入生产系统，从而实现整个柔性换产系统在持续实践中永续自主进化。

（十六）工艺动态优化

“工艺动态优化”场景仅涉及进阶级的企业。

进阶级企业针对复杂生产中工艺参数耦合性强、多环节协同难、优化滞后的问题，引入多模态 AI 建模与深度强化学习技术，构建全链路自主运行的工艺在线优化系统，实现工艺参数的实时自优化、多环节协同寻优与全流程智能进化时，数据治理的目标主要是构建能够深度刻画复杂工艺内在耦合关系、支撑实时协同寻优与自主进化的“高保真关联数据基座”。该基座需超越传统的数据采集与记录，实现从静

态参数到动态耦合关系、从单点数据到全链路状态、从历史记录到未来推演的根本性转变。

1.治理对象

表 59　工艺动态优化场景的数据治理对象

适配层级	数据类型	数据内容
进阶级	多环节工艺数据	各工序温度、压力、速度等工艺参数、工序衔接状态、参数调整记录等
进阶级	感知与实时数据	传感器采集的设备运行数据、工业相机图像、环境参数、高频工艺状态数据等
进阶级	质量与效率数据	关键质量指标、生产效率、成本核算数据等
进阶级	原材料与下游数据	原材料特性、下游需求指标、供应链状态等
进阶级	数字孪生数据	不同参数组合的工艺效果模拟、优化方案预演结果等
进阶级	知识与历史数据	工艺机理知识、工艺知识图谱、历史优化案例、自进化模型迭代记录等

2.平台（技术）工具

表 60　工艺动态优化场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例 进阶级企业
数据治理与集成平台	负责实现覆盖“原料-设备-过程-质量”全价值链多模态工艺数据的实时、高精度采集、融合与统一治理。	Apache Kafka、Apache Flink、西门子 MindSphere
特征工程与样本管理平台	专注于将原始工艺数据转化为能够深度刻画参数耦合关系与优化潜力的高阶智能特征。	Feast、Neo4j、Snorkel
模型开发与运维平台	提供面向工艺动态优化的专用 AI 模型研发、训练与全生命周期管理环境，确保模型从实验、验证到部署上线的效率、可追溯性及对工艺安全约束的合规性。	Ray RLlib、Kubeflow、PyTorch Lightning
数字孪生与智能应用平台	负责构建基于机理的高保真工艺数字孪生，并提供平行仿真环境用于对 AI 优化策略进行长周期、多场景的对抗验证与压力测试。同时，该平台承担将 AI 能力封装为实时优化应用、构建闭环反馈数据流、并驱动整个优化系统持续自主进化的关键职能。	Aspen Plus/Dynamics、微软 Azure Digital Twins、AVEVA System Platform

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。建立工艺数据融合中枢，系统集成四类关键数据源。一是“原料-配方”数据，集成实验室信息管理系统的物料成分、物性数据与生产订单的配方信息。二是“设备-工况”数据，通过物联网平台深度采集产线装备的实时传感器读数（如温度、压力、转速、电流）、控制器设定值及健康状态（如振动、温度）。三是“过程-时序”数据，高频采集各工艺单元（反应器、分离塔、轧机等）的关键过程变量

时序曲线，确保捕捉动态瞬态过程。四是“质量-性能”数据，集成在线检测仪表（如近红外、机器视觉）的实时质量数据与最终产品的实验室全项检验报告，形成质量闭环。**数据预处理。**实施“时空语义”三重对齐与耦合关系初探。以统一的高精度时钟源同步所有时序数据，并将数据映射到具体的工艺单元与物料批次，建立“批次-设备-时间”三维索引。对来自不同系统的同类参数进行单位统一、量纲归一化及语义标定。在清洗异常值与插补合理缺失值的基础上，运用格兰杰因果检验、互信息分析等方法，对关键工艺参数之间、过程参数与质量指标之间的关联性与滞后性进行初步量化分析，为后续深度特征工程提供方向性指引。

②第二阶段：样本准备与特征工程

特征工程。构建多粒度、强关联的特征体系。基于预处理阶段的关联分析，构建以工艺参数为节点、以耦合强度与滞后时间为边的“工艺影响网络”特征，并提取网络拓扑指标。将时序过程数据、设备状态图谱与质量光谱数据等进行融合编码，生成统一的状态表征向量。构造面向强化学习的状态与动作特征，如“当前工况与最优操作面的距离特征”“多目标（收率、能耗、质量）权衡的帕累托前沿逼近度特征”。**数据标注与增强。**组织工艺专家与操作专家，对历史生产数据中的“黄金批次”（高产、优质、低耗）和“异常工况”进行精细化标注，标注内容包含导致结果的关键操作序列、参数调整拐点及专家决策逻辑。利用基于第一性原理

或经验公式构建的高保真工艺机理模型，在数字孪生环境中进行海量仿真，生成涵盖不同原料边界、设备效率衰减、环境干扰下的工艺过程数据，极大地扩充优化样本的多样性与边界完备性。**数据集划分。**采用“基于工艺模式的分层时序划分”策略。首先根据产品类型、原料批次、生产阶段等划分不同的工艺模式，然后在每种模式下严格按照时间顺序划分训练集、验证集和测试集，确保模型既能学习到不同模式下的特异性，又能有效评估其面对未知时间序列的泛化与预测能力。

③第三阶段：模型训练与仿真验证

模型训练。实施“预测-决策”双模型协同训练。训练能够融合多源数据、精准预测未来短期内工艺状态（如关键质量指标）的深度学习模型（如时空图卷积网络），为优化决策提供前瞻性感知。将各个工艺单元或控制回路建模为智能体，在数字孪生环境中，采用多智能体深度强化学习算法，训练其协同寻找全局最优策略。其目标是最大化长期综合收益（如总经济效益、稳定性），同时满足多重工艺约束。**模型验证。**在工艺数字孪生中构建“平行仿真与压力测试验证体系”。该体系将训练好的优化模型置于海量虚拟场景中进行测试。通过对比优化模型策略与基准策略在长期运行下的统计性能，从优化效果、约束满足率、抗扰鲁棒性三个维度进行全面评估，确保其具备上线运行的资格。

④第四阶段：模型部署与闭环进化

模型推理与部署。采用“边云协同、人机协同”的部署架构。将轻量化的实时预测模型与快速响应的局部优化智能体部署在边缘侧，实现毫秒至秒级的实时闭环微调。将全局协同优化模型部署在云端，负责分钟至小时级的全局策略计算与下发。系统初期以“建议”模式运行，将优化参数推荐给操作员确认后执行，随着置信度提升，逐步过渡到在安全边界内的自动执行模式。**闭环运营与进化。**系统持续监测优化指令的实际执行效果，将预测值与实际值、预期收益与实际收益的偏差数据实时回流，用于对预测模型和优化策略进行在线校准与微调，适应装置的慢时变特性。在数字孪生平行系统中，利用积累的新数据与更先进的算法，持续进行新策略的探索与训练。

（十七）先进过程控制

“先进过程控制”场景仅涉及进阶级的企业。

进阶级企业针对复杂工艺中多变量强耦合、动态扰动频发、控制精度不足的问题，引入深度强化学习与数字孪生实时控制技术，构建全链路自主响应的先进过程控制系统，实现多变量协同的精确控制与全流程智能进化时，数据治理的目标主要是构建能够实时刻画多变量动态耦合关系、支撑毫秒级控制决策与系统自演进的“高保真控制数据基座”。该基座需超越传统的监控数据记录，实现从静态模型到动态演

化、从滞后分析到超前预判、从单点优化到全局协同的根本性转变。

1.治理对象

表 61　先进过程控制场景的数据治理对象

适配层级	数据类型	数据内容
进阶级	多变量控制数据	各工艺环节温度、流量、压力等控制变量数据、变量耦合关系记录、控制策略与参数调整记录等
进阶级	实时传感数据	高精度传感阵列采集的工艺状态数据、设备运行参数、环境影响因素数据等
进阶级	扰动与补偿数据	突发扰动事件记录、扰动根源分析、补偿策略及执行效果等
进阶级	数字孪生数据	不同控制策略下的系统动态模拟结果、控制指令预演记录、极端工况仿真数据等
进阶级	工艺与机理数据	工艺机理知识、控制知识图谱、设备特性参数等
进阶级	历史与进化数据	历史控制数据、控制模型迭代记录、自进化系统优化经验等

2.平台（技术）工具

表 62　工厂数字化规划设计场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级企业
数据治理与集成平台	负责实现多变量控制数据、实时传感数据、扰动数据等六大类异构数据的毫秒级实时接入、高精度时序同步与统一治理，构建支撑控制决策的单一可信数据源。	Apache Kafka、OSIsoft PI System、华为云数据湖
特征工程与样本管理平台	专注于从海量实时数据流中提取表征过程动态耦合、扰动传播的高维特征，并管理由专家标注和数字孪生生成的训练样本，为强化学习提供高质量“状态–动作”对。	Feast、Label Studio、阿里云 PAI
模型开发与运维平台	提供深度强化学习等先进算法的开发、训练与全生命周期管理环境，支持在数字孪生中进行大规模仿真训练，并确保模型版本、性能的持续监控与可控迭代。	Ray RLlib、Kubeflow、百度 PaddlePaddle

数字孪生与智能应用平台	构建高保真工艺过程数字孪生，用于控制策略的仿真验证与压力测试；同时作为智能控制应用的载体，实现从仿真推演到在线闭环控制的平滑过渡与持续进化。	ANSYS Twin Builder、微软 Azure Digital Twins、西门子 Process Simulate

3.治理方案

（1）进阶级企业数据治理方案

①第一阶段：数据集成与标准化预处理

数据集成。通过工业物联网平台与实时数据总线，同步汇聚六大类数据至统一的高性能数据湖。核心是毫秒级接入多变量控制数据与实时传感数据，确保控制指令流与过程状态流的精准对齐；同步集成扰动事件、数字孪生仿真结果、工艺机理知识图谱及历史演进数据，构建覆盖“物理-虚拟”“当前-历史”的完整数据集合。**数据预处理。**执行面向实时控制的数据清洗与强化。重点对高频传感与控制数据进行时序同步对齐，消除毫秒级抖动；基于工艺安全边界与统计规则识别并处理异常值；将多源异构数据映射至统一的“设备-回路-变量”语义模型，并进行量纲归一化，为刻画动态耦合关系提供标准、干净的数据基底。

②第二阶段：样本准备与特征工程

特征工程。从原始时序数据中构造能表征过程动态的核心特征。例如，将多变量控制数据转换为反映回路间动态耦合强度的矩阵特征；将实时传感数据与扰动记录聚合为“扰动-响应”关联特征；基于数字孪生数据构建“策略-预期效果”的虚拟验证特征。**数据标注。**组织领域专家，对历史数据中

的“成功控制案例”“扰动抑制过程”以及“参数优化时段”进行归因标注，标记关键操作动作与边界条件，形成高质量的监督学习与模仿学习样本。**数据增强与划分。**利用高保真数字孪生，主动模拟海量极端工况与扰动组合，生成在现实中难以获取的对抗性样本，以增强模型的鲁棒性。按生产周期或运行阶段，将整体数据集严格按时序划分为训练集、验证集和测试集，防止数据泄漏。

③第三阶段：模型训练与仿真验证

模型训练。使用标注与增强后的数据集，驱动深度强化学习等算法进行模型训练。将工艺机理知识（如物质能量平衡约束）与历史优化经验作为先验知识或奖励函数的一部分注入训练过程，引导智能体学习安全、高效的控制策略。**仿真验证。**在独立的测试集上评估模型的预测与控制精度。同时，在数字孪生环境中构建“平行系统”，将训练好的控制模型置于模拟的传感器故障、剧烈扰动、模型失配等极端场景下进行长周期、高强度的压力测试与鲁棒性验证，全面评估其安全性、稳定性与经济性表现，完成上线前的虚拟“压力考核”。

④第四阶段：模型部署与闭环进化

模型推理与部署。将通过严苛验证的控制模型部署为实时推理服务。采用“云-边”协同架构，云端模型负责设定值优化，边缘端模型负责毫秒级回路控制，以“建议-确认-执行”的渐进模式接入现有 DCS 系统，实现智能决策对物理过程

的闭环干预。**闭环运营与进化。**系统持续采集模型控制指令的实际执行效果数据，包括过程响应曲线、产品质量波动与能耗变化，形成“决策-反馈”闭环。将这些反馈数据与对应的工况数据作为新的黄金样本，回流至数据湖。定期利用增量数据驱动特征优化与模型的再训练与微调，使控制系统能够适应工艺漂移并持续探索更优策略，形成自主进化的智能飞轮。

（十八）人机协同作业

“人机协同作业”场景涉及入门级、基础级以及进阶级的企业。

入门级企业针对单一工序人工重复劳动大、精度不稳等问题，引入 AI 协作机器人，利用机器视觉与力反馈技术实现人机协同，机器人执行固定轨迹作业，工人负责复杂判断与灵活操作，AI 安全监测保障作业安全，提升效率与稳定性时，数据治理的目标主要是构建支撑机器人基础作业与简单 AI 感知的标准化、流程化、可靠化的数据环境。该环境需确保机器人运行所需的轨迹数据、视觉定位数据、简单安全规则数据能够被准确采集、稳定传输与清晰定义，为机器人完成固定任务和实现基本的安全监测提供清晰、无误的数据指令与环境反馈。

基础级企业针对协同灵活性不足、任务衔接低效、认知负荷高等问题,引入 AI 驱动的动态感知与自适应决策技术，构建融合多模态人机交互、实时安全调控、数字孪生仿真的

协同系统，通过深度 AI 分析实现任务动态分配、动作预判与智能辅助，提升人机协同的流畅度与效率时，数据治理的目标主要是构建支持多模态感知融合、动态任务建模与实时安全闭环的协同数据基座。该基座需实现对人、机、环多源异构数据的同步采集、关联对齐与统一表征，形成可被 AI 系统理解的“协同场景数字镜像”，以支撑动态任务分配、人机动作预测与实时风险研判。

进阶级企业针对复杂生产场景下人机分工模糊、协作效率低、安全保障不足的问题，引入具身智能与多模态深度理解技术，构建全链路自主协同的人机伙伴系统，实现机器人的深度环境认知、任务自主拆解、人机动态适配与安全智能防护时，数据治理的目标主要是构建能够刻画复杂协作意图、支持具身智能认知与决策、驱动系统自主进化的“人机共生数据生态”。该生态需超越单向的数据采集，支撑机器人形成接近人类的场景理解、任务协作与安全共情能力。

1.治理对象

表 63　人机协同作业场景的数据治理对象

适配层级	数据类型	数据内容
入门级、基础级、进阶级	人机状态数据	工人的工作状态、技能水平、生理指标、情绪状态等；智能协作机器人的运行参数、任务执行进度、故障信息等
入门级、基础级、进阶级	任务与环境数据	生产任务详情、子任务拆解记录、作业环境参数、设备布局等
基础级、进阶级	交互与指令数据	自然语言指令、手势动作、表情信息、交互响应记录等
入门级、基础级、进阶级	安全防护数据	人机相对位置信息、安全域边界数据、预警记录、碰撞风险分析等

适配层级	数据类型	数据内容
基础级、进阶级	数字孪生数据	协同过程预演记录、行为预测结果、分工方案优化模拟等
基础级、进阶级	历史与进化数据	历史协同案例、任务完成效率、自进化模型迭代记录、协作优化经验等

2.平台（技术）工具

表 64　人机协同作业场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	负责实现从机器人本体数据、多模态传感器数据到人工交互指令等全链路人机协同数据的实时接入、同步对齐、统一存储与质量治理，为上层应用提供高可信、低延迟的标准化数据服务。	OPC UA 服务器/客户端、Mosquitto、InfluxDB	Apache Kafka ROS 2、主流云厂商 IoT 核心套件	Cloudera Data Platform Confluent Platform（企业级 Kafka）数据湖仓一体解决方案
特征工程与样本管理平台	专注于将原始的多模态时序数据转化为能够刻画人机状态、意图与协作关系的高阶特征，并对专家标注的协作样本进行版本化管理和合成增强。	Python、Jupyter Notebook	Feast / Tecton Label Studio、Albumentations	英伟达 TAO Toolkit、Snorkel AI

模型开发与运维平台	提供从模仿学习、强化学习到多模态大模型等算法的人机协同模型开发、训练、调优与全生命周期管理环境，支持在仿真和真实场景中进行高效的模型迭代与A/B测试。	Scikit-learn、TensorFlow/PyTorch、MLflow	Ray RLlib PyTorch Lightning、Kubeflow/Azure ML	Unity ML-Agents、Meta Habitat/AI2-THOR、内部MLOps平台
数字孪生与智能应用平台	构建高保真的人机协同虚拟仿真环境，用于复杂协作策略的训练、验证与压力测试；同时作为智能协同应用的运行载体，实现从虚拟训练到实体部署的平滑过渡与持续优化。	CoppeliaSim (V-REP）、Unity/UE简易版、轻量级组态软件	英伟达Isaac Sim、微软Azure Digital Twins、达闼Cloud Brain	英伟达Omniverse 西门子Process Simulate、定制化全息仿真环境

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过机器人控制器接口与基础物联网网关，接入三类核心数据：机器人状态数据（关节位置、速度、力矩）、简单视觉定位数据(用于工件拾取或放置的2D坐标)，

以及预设的安全规则数据(如电子围栏边界、最大速度限制)。同时记录工人通过示教器或简单界面触发的交互指令数据。**数据预处理。**对机器人状态数据进行滤波以平滑噪声；将视觉坐标数据转换到统一的机器人基坐标系下；对安全规则数据进行结构化存储与版本管理。统一所有数据的时间戳至毫秒级，确保动作与感知在时间上的基本同步。

②第二阶段：样本准备与特征工程

特征工程。构造基础作业特征，如“机器人末端执行器的实际运动轨迹与预设轨迹的偏差”“完成一次拾放循环的周期时间”“关节力矩的实时变化曲线”。**数据标注。**由产线班组长或工程师，对历史运行日志中的“作业成功/失败”事件（如拾取失败、放置精度超差）进行简单标注，并关联当时的状态数据。**数据增强与划分。**对有限的异常样本，通过轻微扰动预设轨迹参数或模拟视觉遮挡等方式进行少量数据增强。按生产批次或工作日划分数据集。

③第三阶段：模型训练与仿真验证

模型训练。主要训练基础的安全监测模型（如基于规则或简单分类算法，判断机器人是否即将进入安全禁区）和异常检测模型（如通过时序模型识别关节力矩异常，预示潜在卡阻）。**模型验证。**在离线数据集上验证安全模型的预警准确率和异常检测模型的召回率。在机器人仿真软件中，模拟典型异常场景，测试安全模型的响应是否及时、准确。

④第四阶段：模型部署与闭环进化

模型推理与部署。将验证后的模型固化为机器人控制系统的内置功能模块，如实时安全监控线程或周期性的设备健康自检报告。**闭环运营。**系统记录每次安全预警或异常报警是否属实，以及人工处理结果，定期（如每月）汇总分析，用于评估模型效果并作为未来优化规则或阈值的依据，形成初步的数据反馈闭环。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。在入门级基础上，扩展集成多模态人机状态数据（如工人的动作骨骼关键点、语音指令、注意力朝向估计）、高维环境数据（3D 点云、多视角视频流）、实时力觉/触觉数据，以及来自数字孪生的仿真任务流数据。所有数据通过统一的数据总线接入。**数据预处理。**实现多源数据的“时空-语义”统一。采用统一时钟源进行毫秒级同步；将视觉、点云、机器人坐标系统一至世界坐标系；对语音指令进行识别与结构化；对动作骨骼数据进行降噪与标准化。

②第二阶段：样本准备与特征工程

特征工程。构建协同场景特征，如“人机相对位置与速度向量”“工人手势意图编码”“当前任务阶段的多模态融合表征”“基于历史交互预测的工人下一步动作概率分布”。**数据标注。**组织人机工程专家与熟练工人，对大量人机协同视频片段进行标注，标注内容包括“任务阶段”“工人意图”

“最优/次优机器人辅助动作”，形成高质量的模仿学习样本。**数据增强与划分。**利用数字孪生，在虚拟环境中模拟不同体型工人、不同任务节奏、不同环境光照下的海量协同场景，生成合成数据。按复杂任务类型划分数据集，确保各类场景在训练和测试集中均有体现。

③第三阶段：模型训练与仿真验证

模型训练。训练多模态融合感知模型，以理解协同场景；训练任务动态分配模型（如基于强化学习），决定何时由谁执行何动作；训练人机动作预判模型，使机器人能提前准备。**模型验证。**在独立测试集上评估各模型的精度。在高保真数字孪生环境中进行系统性验证，模拟紧急中断、工人误操作、新型任务等复杂场景，全面评估协同系统的流畅性、效率提升以及安全性，确保其优于固定模式的协同。

④第四阶段：模型部署与闭环进化

模型推理与部署。将模型群部署为“协同大脑”微服务，通过边缘计算设备实时处理多模态数据流，并向机器人和工人 AR 设备发送辅助指令。**闭环运营。**系统持续采集实际协同过程中的多模态数据、AI 决策、工人反馈（显式或隐式）及最终任务效能指标。这些数据自动回流，用于对预判模型、分配模型进行每周或每月的增量学习与优化，使系统不断适应工人的个性化习惯。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。在基础级之上，实现更深度的融合。集成神经生理数据（如工人肌电信号、眼动轨迹，用于更精准的意图识别）、机器人多模态感知的原始高维流数据、复杂任务的知识图谱数据，以及记录长期协作演化的历史与进化数据。构建企业级的“人机协同数据湖”。**数据预处理。**侧重于高维数据的表征学习与对齐。利用自监督学习等方法，从原始感知数据中提取具有语义意义的紧凑表征。建立跨模态的共享语义空间，实现语言指令、视觉场景、动作序列在语义层面的对齐与互译。

②第二阶段：样本准备与特征工程

特征工程。构建共生级特征，如“人机联合任务的心智模型向量”“对不确定环境和模糊指令的鲁棒性场景理解特征”“长期协作中形成的人机默契度指标”。**数据标注。**标注工作转向对协作策略与长期效果的评估。专家需对长期的协作会话进行宏观标注，评价协作策略的优劣、安全共情的水平，为强化学习提供高阶奖励信号。**数据增强。**核心是利用元学习与因果学习框架，在数字孪生中自动构建海量反映复杂因果关系的训练场景，驱动模型学习可泛化的协作常识与物理常识。

③第三阶段：模型训练与仿真验证

模型训练。训练具身智能体，使其不仅能感知，还能在仿真环境中通过“动手”尝试来理解任务物理约束；训练分层强化学习模型，底层负责精细动作，高层负责协作策略与沟通；训练人机双向意图理解与协商模型。**模型验证。**验证重点从“性能”转向“能力”。在数字孪生中设计开放式、涌现式的测试任务，评估系统面对前所未见任务时的分解与解决能力、在人机意图冲突时的协商能力，以及在动态风险中的主动共情防护能力。

④第四阶段：模型部署与闭环进化

模型推理与部署。将训练成熟的“人机伙伴”作为自主智能体部署。它具备高度的自主权，能与工人进行自然、持续的对话与协作，共同应对复杂任务。闭环进化。建立“社会学习”式进化机制。系统不仅从自身交互数据中学习，还能从观察其他工人-机器人团队、吸收外部知识库中学习。进化目标是使协作模式逼近甚至超越最佳的人类团队，实现真正的共生进化，形成具备持续创新能力的人机协作生态。

（十九）在线智能检测

“在线智能检测”场景涉及入门级、基础级以及进阶级的企业。

入门级企业针对规则明确的产品外观或尺寸缺陷，部署工业相机和光源，通过卷积神经网络等视觉算法提取图像特征并进行自动化检测，结果自动记录并联动分拣，提升效率

与准确性时，数据治理的目标主要是构建标准化、流程化、可靠化的数据环境，确保用于视觉算法训练的图像样本数据、用于模型推理的实时图像流以及检测结果数据，能够被清晰定义、准确采集、稳定传输与统一存储，为缺陷检测提供高质量、无歧义的数据输入与结果输出保障，支撑固定规则下的自动化判定。

基础级企业针对复杂多样化外观缺陷、部分可量化物性分析不足的问题，引入 AI 学习与多模态数据融合技术，构建融合深度学习模型、多传感器协同与数字孪生仿真的智能检测系统，通过 AI 分析实现复杂缺陷精准识别、物性量化分析与动态检测策略优化，提升检测覆盖范围与精度时，数据治理的目标主要是构建支持多模态数据融合、复杂缺陷特征提取与检测策略动态优化的协同数据基座。该基座需实现多源异构数据的同步采集、时空对齐与统一表征，形成可被深度学习模型高效理解的跨模态缺陷数字镜像，支撑复杂缺陷的精准识别与量化分析。

进阶级企业针对全流程质量管控中缺陷根因难定位、质量干预滞后、小样本场景适应差的问题，引入多模态融合学习与数字孪生质量推演技术，构建全链路自主运行的在线智能检测系统，实现质量缺陷的实时溯源、趋势预判、主动干预与全流程智能进化时，数据治理的目标主要是构建能够刻画质量缺陷全生命周期、支持质量态势感知与推演、驱动检测系统自主进化的“质量数据智能体”。该体系需超越单点

检测，实现从原材料、生产过程、设备状态到最终缺陷的全链路数据因果关联；需构建数字孪生驱动的质量仿真与根因追溯环境；并需建立从质量干预结果中持续学习的反馈闭环，最终实现质量管控从“事后检验”到“事前预防”与“实时调控”的根本性转变。

1.治理对象

表 65　在线智能检测场景的数据治理对象

适配层级	数据类型	数据内容
入门级、基础级、进阶级	多模态检测数据	机器视觉图像、光谱分析数据、声学检测记录、三维扫描结果等缺陷特征数据
入门级、基础级、进阶级	生产与工艺数据	各环节工艺参数、设备运行状态、原材料特性、工序操作记录等
入门级、基础级、进阶级	质量与缺陷数据	缺陷类型、位置、严重程度、根因分析结果等
进阶级	预测与干预数据	质量趋势预测结果、风险预警记录、工艺参数调整方案及执行效果等
进阶级	数字孪生数据	不同工艺参数下的质量状态模拟、检测策略预演结果、干预方案仿真记录等
基础级、进阶级	小样本与历史数据	新产品/新缺陷的少量样本数据、历史检测案例、自进化模型迭代记录、质量改进经验等

2.平台（技术）工具

表 66　在线智能检测场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	实现多源数据的采集、清洗、集成与统一管理，构建高质量、可信的数据底座。	通用工业相机厂商的配套 SDK 与软件、开源的 Apache NiFi	阿里云 DataWorks/瓴羊 Dataphin、蚂蚁集团 DataFab	星环科技 TDS、网易数帆 EasyData

特征工程与样本管理平台	提供从原始数据到模型特征的转换、标注、增强与版本管理能力，为模型训练准备“燃料”。	LabelImg、CVAT 等开源标注工具	BeagleData Kaleido、度小满 Etron 自动化建模平台	星环科技 Sophon Base 的可视化建模模块
模型开发与运维平台	提供算法开发、模型训练、评估、部署及全生命周期管理的环境，支撑 AI 模型高效迭代。	本地化的 PyTorch、TensorFlow、Jupyter Notebook	星环科技 Sophon Base、MLflow	星环科技 Sophon LLMOps、嘉为蓝鲸 AIOps 平台
数字孪生与智能应用平台	构建高保真虚拟环境进行仿真验证与策略推演，并承载智能检测应用，实现闭环决策。	Halcon、OpenCV 的机器视觉仿真模块	博维数孪 Create/PlayTwins、飞渡科技 DTS	树根互联根云平台、华为河图 Cyberverse、阿里云数字孪生引擎

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过工业相机与 PLC 接口，主要接入两类核心数据：标准产品图像数据和在线检测图像流数据。同时，记录检测系统输出的判定结果数据及简单的设备触发信号。**数据预处理。**对采集的图像执行标准化处理，包括统一分辨率、色彩空间转换、光照归一化以及去除噪声。为每张图像关联唯一的产品序列号、时间戳及相机位置信息，确保数据可追溯。

②第二阶段：样本准备与特征工程

特征工程。基于预训练的卷积神经网络模型，提取图像的高层语义特征。同时，可构造简单的统计特征，如图像特定区域的像素值均值、方差或直方图分布，作为辅助。**数据标注。**由质检人员使用标注工具，对训练图像中的缺陷区域进行精确框选或分割，并分类标注缺陷类型。标注过程需制定明确的标注规范，确保一致性。**数据增强与划分。**针对正样本（缺陷样本）不足的问题，采用基础的数据增强技术，如随机旋转、翻转、裁剪、亮度对比度调整等，扩充训练集。按产品型号或生产批次划分数据集，确保训练集和测试集分布一致。

③第三阶段：模型训练与仿真验证

模型训练。使用标注和增强后的数据集，训练一个目标检测或图像分类模型。训练目标是在测试集上达到预设的准确率与召回率指标。**模型验证。**在独立的验证集上评估模型性能，重点考察其对各类已知缺陷的识别准确率与泛化能力。可在图像层面模拟轻微的视角变化、光照波动，测试模型的鲁棒性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将训练好的模型转换为优化格式，部署在边缘计算设备或工控机上，实现毫秒级实时推理。检测结果通过接口自动触发分拣装置或记录至 MES 系统。**闭环运营。**系统定期（如每日）统计误检、漏检案例，由人工复

核后，将修正的标注数据作为新增样本，定期（如每周）对模型进行增量训练或微调，以持续适应生产环境的微小变化。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。扩展接入多模态传感数据，包括高分辨率 2D 图像、3D 点云/结构光数据、X 光图像、近红外光谱数据等。同时，集成产线工艺参数数据和数字孪生仿真数据。**数据预处理。**实现多源数据的时空严格对齐。确保同一产品在通过不同传感器时，其数据在时间和空间坐标系上能够精确匹配。对 3D 点云进行去噪、配准和特征提取；对光谱数据进行降维和标准化。

②第二阶段：样本准备与特征工程

特征工程。构建跨模态融合特征。例如，将 2D 图像的纹理特征与 3D 点云的高度/曲率特征进行融合；将外观图像与 X 光内部结构图像的特征进行关联；构建反映缺陷与工艺参数关联性的复合特征。**数据标注。**标注工作升级为多模态联合标注。质检专家需在同一产品的多模态数据视图上，协同标注同一缺陷，并补充标注缺陷的量化属性。**数据增强与划分。**利用数字孪生，在虚拟环境中模拟不同材质、不同缺陷形态、不同成像条件下的海量多模态数据样本。数据集按缺陷的复杂程度和出现场景进行分层划分。

③第三阶段：模型训练与仿真验证

模型训练。训练多模态融合深度学习模型，如基于注意力机制的网络，以综合判断缺陷。训练缺陷量化回归模型，预测缺陷的物理尺寸或严重程度。探索元学习或小样本学习技术，以快速适应新产品或新缺陷。**模型验证。**在数字孪生环境中构建虚拟检测产线，将训练好的模型置于模拟的复杂工况下进行系统性验证，评估其识别精度、量化准确性与系统鲁棒性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将多模态模型部署为微服务集群，通过高速数据总线接收和处理各传感器数据流，输出包含缺陷类型、位置、量化信息的综合检测报告。**闭环运营。**系统持续采集实际检测数据、模型判断结果，以及后续的人工复检或实验室测量结果。利用这些反馈数据，不仅可以优化现有模型，更能用于训练一个检测策略优化模型，动态调整不同传感器的启用、检测算法的参数或检测流程的优先级，实现系统效能的持续提升。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。实现“质量数据全域融合”。在基础级之上，进一步集成全流程数据，构建企业级“质量数据湖”。**数据预处理。**侧重于跨域数据的因果关联与知识提取。利用时序分析、因果发现等方法，初步挖掘生产参数波动与最终缺陷

出现之间的潜在关联关系。对非结构化的运维日志、工艺文档进行自然语言处理，提取关键实体与事件。

②第二阶段：样本准备与特征工程

特征工程。构建“质量根因溯源特征”与“质量风险预测特征”。例如，构建反映“特定设备参数序列异常模式”的特征；构建“多工序质量指标传播链”的图网络特征；基于历史数据构建“缺陷发生前兆”的早期预警特征。**数据标注。**标注重点转向对质量事件链的归因分析[61]。专家需对历史上发生的重大质量事故或频发的缺陷，进行全链路回溯分析，标注出关键的根本原因节点，形成高质量的因果学习样本。**数据增强。**核心是利用数字孪生进行质量推演与根因实验。在孪生体中主动“注入”各种假设的设备故障、工艺偏差或原料问题，模拟其对最终质量的影响，生成海量“原因-结果”配对数据,用于训练强大的质量根因诊断与预测模型。

③第三阶段：模型训练与仿真验证

模型训练。训练质量缺陷实时溯源模型，能够在新缺陷出现时快速锁定可疑工序与设备。训练质量趋势预判模型，基于实时生产数据预测未来一段时间内的质量风险。训练主动干预策略模型，通过强化学习在数字孪生中学习如何调整工艺参数以避免缺陷发生。**模型验证。**验证在全链路、长周期、小概率场景下进行。在数字孪生中模拟完整生产周期的运行，评估系统对未知新型缺陷的溯源能力、对缓慢漂移导致的质变风险的预警能力，以及干预策略的经济性与安全性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将上述模型群集成为“质量大脑”，作为核心服务部署。它不再仅是检测终端，而是能够实时监控全流程、主动预警、并给出或直接执行优化建议的智能决策中心。**闭环进化。**建立“质量数据飞轮”。每一次质量事件的真实根因确认、每一次干预措施的实际效果，都作为强化“质量大脑”认知的反馈数据，自动回流至数据湖和数字孪生。系统利用这些数据持续迭代所有模型，并能在孪生体中主动探索更优的质量控制策略，从而实现质检系统从“感知-判断”到“认知-决策-进化”的质变，最终驱动全流程质量的自主、持续优化。

（二十）质量精准追溯

“质量精准追溯”场景涉及入门级、基础级以及进阶级的企业。

入门级企业针对纸质记录追溯效率低、易丢失等问题，部署基础 QMS 系统融合 AI 技术，为关键物料/半成品/成品赋予唯一标识，通过 AI 机器视觉自动扫描识别，实现电子化数据采集与追溯链，替代纸质记录并提升效率时，数据治理的目标主要是构建一个以“一物一码”为核心、支撑数据自动采集与单向关联的标准化、可检索数据环境。该环境需确保赋码规则统一、扫码数据准确、基础信息字段完整，为形成电子化、可查询的单向追溯链条提供清晰、无误的数据

基础，实现从“纸质分散记录”到“电子集中归档”的初级转变。

基础级企业针对跨系统数据割裂、追溯效率低、根因定位难等问题，引入 AI 驱动的全流程数据融合与智能分析技术，构建融合多源异构数据、智能决策与数字孪生仿真的质量追溯系统，通过 AI 分析实现全链条数据关联、质量波动预警与根因自动定位，提升追溯精准度与效率时，数据治理的目标主要是构建一个能够打通“人、机、料、法、环、测”多系统数据壁垒、支撑多维度关联分析与根因初步定位的融合化、关联化数据基座。该基座需实现跨系统数据的自动抽取、清洗、对齐与关联，形成以产品批次为主线的“全要素数据关联图谱”，为 AI 模型进行质量波动关联分析、异常模式识别与常见根因定位提供高质量、强关联的数据输入。

进阶级企业针对复杂质量波动中根因诊断难、风险预警滞后、追溯链条断裂的问题，引入因果推理与全域数字孪生追溯技术，构建全链路自主运行的质量精准追溯系统，实现质量波动的秒级根因定位、潜在风险预判与全链条智能追溯时，数据治理的目标主要是构建一个能够深度刻画质量波动产生、传导与演化全过程，支撑实时因果推断与预测性决策的“质量认知数据智能体”。该体系需超越数据关联，致力于建立覆盖供应链、生产、检测全链路的高保真因果数据模型；需构建基于数字孪生的实时质量推演与根因实验环境；并需形成从干预结果中持续学习因果知识的反馈闭环，最终

实现质量追溯从“事后关联查询”向“事中实时诊断”与“事前风险预见”的跃升，驱动质量管理体系的自适应与自优化。

1.治理对象

表 67　质量精准追溯场景的数据治理对象

适配层级	数据类型	数据内容
入门级、基础级、进阶级	全要素生产数据	“人机料法环测”各环节数据，如人员操作记录、设备运行参数、原材料信息、工艺标准、环境参数、检测数据等
入门级、基础级、进阶级	质量波动与缺陷数据	质量异常记录、缺陷特征、波动趋势、根因诊断结果等
入门级、基础级、进阶级	产品标识数据	产品数字身份证信息、RFID 标签数据、二维码记录、批次与序列号关联数据等
基础级、进阶级	追溯链条数据	各环节流转记录、工序交接信息、全生命周期轨迹数据等
进阶级	预测与风险数据	潜在质量风险预警、受影响产品追溯结果、干预方案及效果等
基础级、进阶级	数字孪生数据	质量风险场景模拟、追溯路径推演结果、干预方案预演记录等
基础级、进阶级	历史检测数据	检测案例、缺陷与工艺关联记录、优化方案等
进阶级	进化数据	历史追溯案例、根因分析经验、知识图谱迭代记录、自进化模型优化数据等

2.平台（技术）工具

表 68　质量精准追溯场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	实现从物料编码、过程参数到检验结果等全链路质量数据的统一采集、清洗、关联与可信存储，为追溯体系构建	主流 QMS/ERP 系统内置模块、基础 ETL 工具	网易数帆 EasyData、腾讯云 WeData、瓴羊 Dataphin	星环科技 TDS、企业级数据湖仓平台

	标准化的数据基座。			
特征工程与样本管理平台	将原始数据转化为表征质量因果关系的特征，管理缺陷根因标注、多因子关联样本，为分析模型提供“高质量数据燃料”。	人工标注结合 Excel/简单脚本处理	BeagleData Kaleido、度小满 Etron 自动化建模平台	星环科技 Sophon Base、自动化因果特征发现平台
模型开发与运维平台	提供从根因分析模型、预测模型训练、仿真验证到持续部署的全生命周期管理，支撑质量诊断算法的迭代与可靠运行。	Python、Jupyter Notebook	星环科技 Sophon Base、MLflow	星环科技 Sophon LLMOps、嘉为蓝鲸 AIOps
数字孪生与智能应用平台	构建高保真生产过程虚拟镜像，用于质量波动推演、根因追溯仿真；并承载智能追溯看板、决策建议等应用。	基础 2D/3D 可视化工具	飞渡科技 DTS、华为河图 Cyberverse、百度智能云开物	树根互联根云平台、阿里云数字孪生引擎

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过基础 QMS 系统接口与工业相机 SDK，集成三类核心数据：为关键物料/半成品/成品生成的唯一标识数据、AI 视觉系统扫描识别的实时扫码记录数据，以及人工在关键工序点通过终端录入的基础批次与工序数据。**数据预处理。**对 AI 视觉扫码数据进行准确性与完整性校验，自

动修正或标记识别模糊、错误的记录；统一并标准化所有物料编码、批次号、工序代码的命名规则与格式；将所有数据按时间顺序与批次号进行初步关联对齐，形成结构化的电子记录。

②第二阶段：样本准备与特征工程

特征工程。从集成后的结构化数据中，构造用于追溯查询与分析的基础特征，如“物料/产品全生命周期流转路径”“各关键节点时间戳序列”“工序与操作员关联记录”等。**数据标注。**由质量管理人员对历史数据中的“异常流转事件”进行识别与标注，为后续简单的异常检测模型提供监督标签。**数据增强与划分。**对有限的异常事件样本，采用重采样或基于规则的轻微扰动（如模拟部分字符识别错误）进行少量数据增强。按生产时间顺序或产品批次，将数据集划分为训练集与测试集，确保时间序列的连续性。

③第三阶段：模型训练与仿真验证

模型训练。主要训练基于规则的数据校验引擎或简单的分类模型，用于自动识别标识重复、批次信息逻辑冲突、必填字段缺失等基础数据质量问题。**模型验证。**在独立的测试数据集上，验证模型的准确率与召回率，确保其能有效替代人工进行基础数据核对，支撑追溯链条的完整性与准确性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将验证后的规则或模型以微服务或内置模块形式，集成至 QMS 系统的数据录入与校验环节，实

现实时数据质量监控与自动预警。**闭环运营与进化。**系统持续记录模型的预警信息与人工复核确认结果，定期将确认为真实异常的案例作为新的训练样本，对模型进行增量训练或规则优化，形成数据质量持续提升的初级闭环。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过企业服务总线或数据中台，全面集成来自 ERP（物料主数据、采购批次）、MES（工艺参数、设备状态、工单信息）、WMS（仓储出入库）、LIMS（检验报告、光谱数据）以及环境监控系统的多源异构数据流。同时，接入更丰富的产线传感器数据。**数据预处理。**实施跨系统的“时空-业务”三重对齐。以生产工单和产品批次为主键，清洗并关联各系统数据；统一不同来源数据的时间戳至毫秒级；解析并标准化非结构化的检验报告、维修日志文本信息；解决不同系统间物料编码、设备 ID 的映射与一致性问题。

②第二阶段：样本准备与特征工程

特征工程。构建深度反映质量关联的复合特征。例如：“跨工序参数影响链特征”，量化上游工序的工艺参数波动对下游关键质量指标的潜在影响；“多维度异常协同特征”，将特定时间窗口内的设备振动异常、温度偏移与最终产品缺陷进行关联建模。**数据标注。**组织工艺、质量和设备专家，对历史发生的典型质量事故或周期性波动进行根因回溯分析标注。标注内容不仅包括最终缺陷，更需标记出从原材料

到成品的全链条中，导致问题的关键异常数据节点、序列及组合模式。**数据增强与划分。**利用数字孪生技术，基于历史数据和工艺机理模型，仿真生成不同原材料批次、设备性能衰减、工艺边界条件下的生产过程与质量数据，扩充训练样本的多样性与覆盖范围。按产品家族或故障模式对数据集进行分层划分，确保各类场景在训练与测试中均有体现。

③第三阶段：模型训练与仿真验证

模型训练。训练时序图神经网络或多变量关联分析模型，用于挖掘跨工序、跨参数的质量异常传播路径。训练集成学习分类模型，用于综合多维度特征，对质量波动进行根因分类。**模型验证。**在独立的测试集上评估模型的根因定位准确率。更重要的是，在高保真数字孪生环境中，构建虚拟质量事故场景，验证模型能否在模拟的全链路数据中，准确关联并定位到预设的根因节点，并评估预警的时效性与误报率。

④第四阶段：模型部署与闭环进化

模型推理与部署。将训练好的模型群部署为质量追溯系统的智能分析引擎，实时关联多源数据流，动态输出质量波动预警及根因假设图谱，并与 MES、Andon 系统联动。**闭环运营与进化。**建立人机协同反馈环。系统将 AI 推荐的根因假设推送给工程师确认，并将确认结果连同对应的完整数据上下文，作为高质量的反馈数据回流至数据湖。定期使用这些增量数据对模型进行再训练与微调，使其持续适应产线变化，提升分析精度。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级“质量数据宇宙”，全量集成供应链协同数据（供应商物料谱系、过程数据）、制造全链路毫秒级传感器时序数据、高清视频流、声纹数据，以及市场端的客户体验数据、社交媒体舆情等。实现内外部数据、结构与非结构化数据的全域融合。**数据预处理。**在融合基础上，运用因果发现算法对海量高维数据进行初步分析，自动探索和识别变量间潜在的因果结构关系，为后续构建可解释的因果模型提供先验图模型。

②第二阶段：样本准备与特征工程

特征工程。构建面向因果推断的高阶特征。例如，构建“潜在因果变量对”的特征表示，以及用于估计“干预效应”的反事实特征。从全域数据中提取能够表征复杂系统状态的深度表征特征。**数据标注。**标注重点转向对 “干预-结果”因果对的确认。专家需对历史成功或失败的质量改进措施进行深度评估，明确标注所采取的具体干预动作（如调整参数X）与产生的量化质量结果（如指标Y的变化量），形成因果效应评估的黄金标准。**数据增强。**核心是利用全域数字孪生进行“因果实验场”模拟。在虚拟环境中，主动、安全地实施各种在现实中难以进行的干预（如极端工艺参数组合、模拟新型原材料缺陷），精确记录其引发的全链路连锁反应，

生成海量、标注清晰、无混杂偏差的“原因-结果”仿真数据，为训练鲁棒的因果模型提供核心燃料。

③第三阶段：模型训练与仿真验证

模型训练。训练基于结构因果模型或双机器学习的因果推断模型，使其能基于观测数据，定量估算不同生产因素（如原料属性 A、设备参数 B）对最终质量指标（如缺陷率 C）的因果效应大小。训练深度强化学习模型，在数字孪生中学习在复杂约束下的最优质量干预策略。**模型验证。**验证在极端复杂与反事实场景下进行。在数字孪生中模拟出现实中未见过的、多因素非线性交织的质量问题，测试因果模型能否准确推断出主导根因及其贡献度。同时，构建“平行追溯与决策系统”，让 AI 模型与人类专家在相同虚拟场景下进行根因定位与决策竞赛，从准确性、速度、经济性等多维度进行系统评估。

④第四阶段：模型部署与闭环进化

模型推理与部署。将因果模型与策略模型部署为实时质量管控系统的“认知决策核心”。系统不仅能秒级定位当前质量波动的根因，更能预测不同干预措施的潜在结果，并推荐综合最优的决策方案，甚至可在授权范围内自动执行微调指令。**闭环运营与进化。**建立“社会学习”式进化机制。每一次真实世界的干预决策及其带来的长期质量、成本、效率结果，都被自动采集、评估并转化为因果知识，持续反馈至数据宇宙与数字孪生。这使得系统不仅能从自身经验中学习，

还能通过吸收外部知识、模拟推演，不断发现更深层次的质量规律与更优的控制策略，实现质量管理体系的自主、持续、前瞻性进化，最终形成具有韧性的质量智能体。

（二十一）质量分析与改进

“质量分析与改进”场景涉及入门级、基础级以及进阶级的企业。

入门级企业针对生产过程中显性质量问题数据分散、分析依赖人工的问题，利用 AI 技术构建质量数据分析系统，通过标准化模板整合分散数据，实现质量问题自动分类、结构化记录与分布分析，提升质量改进效率时，数据治理的目标主要是构建标准化、结构化和集中化的数据环境，确保分散在各处的、显性的质量问题数据能够被统一范式采集、清晰定义和集中管理，为 AI 系统提供进行自动分类与分布分析所需的一致、可计算的数据基础，实现从“人工经验驱动”到“数据基础驱动”的初步转变。

基础级企业针对多因素影响的复杂质量波动，引入 AI 驱动的深度挖掘与智能决策技术，构建融合全流程数据关联、机器学习根因定位与数字孪生仿真的质量改进系统，通过 AI 分析实现复杂质量波动的多维度解析、智能方案推荐与改进闭环管理，提升质量改进的精准性与效率时，数据治理的目标主要是构建支持多源数据融合、复杂因子关联与改进策略评估的质量改进数据基座，为 AI 模型进行深度根因定位与

方案智能推荐提供高质量、强因果关联的分析输入，支撑数据驱动的持续改进闭环。

进阶级企业针对潜在质量风险识别滞后、根因定位模糊、改进策略僵化的问题，引入深度因果学习与自进化知识图谱技术，构建全链路自主运行的质量分析与改进系统，实现质量趋势的精准预判、风险根因的智能定位、改进策略的自适应生成与全流程智能进化时，数据治理的目标主要是构建能够感知质量态势、洞察深层因果、驱动策略自适应优化的“质量智能体”。该体系需超越显性问题，实现对潜在质量风险与隐性模式的预见；需建立可解释、可演化的因果知识图谱；并形成从策略执行结果中持续学习并自我更新的反馈闭环，最终实现质量改进从“被动响应”到“主动预测”与“自主优化”的根本性跨越。

1.治理对象

表 69　质量分析与改进场景的数据治理对象

适配层级	数据类型	数据内容
入门级、进阶级	多模态质量数据	工业相机图像、设备传感器数据、生产工艺参数、检测结果等
入门级、进阶级	质量风险与缺陷数据	潜在风险记录、缺陷特征、质量异常趋势、根因定位结果等
入门级、进阶级	生产全要素数据	人员操作信息、设备运行状态、原材料特性、环境参数、工序流转记录等
入门级、基础级、进阶级	文档数据	质量报告、改进案例、技术文档
进阶级	知识数据	质量报告、改进案例、技术文档、自进化质量知识图谱数据等

适配层级	数据类型	数据内容
基础级、进阶级	改进策略与效果数据	动态改进方案、实施记录、效果评估结果等
基础级、进阶级	数字孪生数据	质量模型模拟数据、改进策略预演结果、方案优化仿真记录等
基础级、进阶级	历史质量数据	问题案例、根因分析、解决方案、改进效果记录等
进阶级	迭代数据	改进经验、模型参数迭代记录等

2.平台（技术）工具

表 70　质量分析与改进场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级企业	基础级企业	进阶级企业
数据治理与集成平台	负责从分散记录、多源业务系统到全域物联数据的高效集成、清洗、关联与统一管理，构建高质量、强关联、可追溯的质量分析数据基座。	主流 QMS/MES 系统内置数据模块、腾讯文档、简道云	阿里云 DataWorks、网易数帆 EasyData、华为云 ROMA	星环科技 TDS、Cloudera Data Platform
特征工程与样本管理平台	提供从原始数据到深度分析特征的转换能力，支持质量案例的标注、因果样本的管理与仿真数据的合成，为模型训练准备高质量输入。	人工标注结合 Excel/Python Pandas	BeagleData Kaleido、Label Studio	星环科技 Sophon Base 的可视化因果探索模块、自动化特征工程平台
模型开发与运维平台	提供从统计分析、机器学习到深度因果学习等模型的开发、	Python、R 语言、Jupyter Notebook	星环科技 Sophon Base、MLflow	星环科技 Sophon LLMOps、华

	训练、评估、部署与全生命周期管理环境,支撑质量智能算法的持续迭代。			为云 ModelArts
数字孪生与智能应用平台	构建生产过程的高保真虚拟镜像,用于质量波动推演、根因实验与改进策略仿真;并承载智能分析报告、决策建议等应用。	Tableau Public、Power BI 等基础数据可视化工具	达索 3DEXPERIENCE、百度智能云开物	西门子 Process Simulate、ANSYS Twin Builder、微软 Azure Digital Twins

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过标准化数据接口或模板，系统集成分散在纸质记录、电子表格及简单数据库中的显性质量问题记录数据，包括缺陷描述、发生位置、产品批次等。同时，整合基础的生产过程记录数据，如关键工序的检验结果。**数据预处理。**执行数据的结构化转换与清洗。将非结构化的缺陷描述文本，通过自然语言处理技术[62]进行关键词提取与标准化分类;统一各来源数据的编码体系(如缺陷代码、产品型号);处理缺失值与明显异常记录，建立统一的质量问题主数据。

②第二阶段：样本准备与特征工程

特征工程。构造用于分类与分布分析的基础特征,如“缺陷类型分布频率”“按生产线/班次/产品的缺陷发生热力图”“关键工序合格率时序变化”等。**数据标注。**由质量工程师

对历史缺陷描述进行标准化分类标签的复核与确认，形成高质量的监督学习样本集，用于训练自动分类模型。**数据增强与划分。**对少数类缺陷样本进行重采样，以平衡分类模型训练。按时间顺序将数据集划分为训练集与测试集，确保模型评估的时间有效性。

③第三阶段：模型训练与仿真验证

模型训练。训练文本分类模型（如基于 BERT 的微调模型），用于自动将新上报的缺陷描述归类到标准分类体系。训练统计过程控制模型，自动识别关键质量指标的异常波动。**模型验证。**在独立测试集上验证分类模型的准确率与召回率，验证 SPC 模型[63]对已知历史异常点的识别能力，确保分析工具的基础可靠性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将分类模型与 SPC 模型集成至质量数据分析系统，实现新问题的自动归类与异常波动的自动报警。生成标准化的质量分析报告模板。**闭环运营与进化。**系统收集用户对自动分类结果的修正反馈，以及对分析报告的实际应用评价，定期用于模型的迭代优化，提升系统的实用性与准确性。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过数据中台或企业服务总线，深度集成来自 MES（工艺参数、设备状态）、ERP（物料批次）、QMS

（检验数据、客诉记录）、设备物联网传感器以及数字孪生仿真数据的全流程多源数据。**数据预处理。**实现跨系统数据的“时空-批次”关联对齐。以生产工单和产品序列号为轴心，清洗、关联并同步制造全过程数据；对传感器高频时序数据进行降采样与特征初提取；统一多源数据的语义定义，构建企业级质量数据模型。

②第二阶段：样本准备与特征工程

特征工程。构建深度关联特征，如“多工序参数波动传播链特征”，量化上游工艺变化对下游质量指标的累积影响；构建“人-机-料-法-环综合影响因子特征矩阵”。**数据标注。**组织跨部门专家（工艺、设备、质量）对历史上已解决的复杂质量波动案例进行根因回溯标注，不仅标注最终确定的根本原因，还需标注调查过程中排除的疑似因素及决策依据，形成用于训练根因推理模型的复杂样本。**数据增强与划分。**利用数字孪生仿真环境，主动模拟在不同原材料特性、设备性能状态、环境扰动组合下的生产过程，生成对应的质量结果数据，极大扩充用于关联与因果分析的样本库。按问题复杂度和工艺类型对数据集进行分层划分。

③第三阶段：模型训练与仿真验证

模型训练。训练集成学习或图神经网络模型，用于从高维特征中识别导致质量波动的关键因子组合。训练推荐系统模型，基于历史成功改进案例的特征，为相似质量问题推荐潜在的改进措施方案。**模型验证。**在独立的测试集上评估根

因定位模型[64]的 Top-K 准确率[65]。在数字孪生环境中构建虚拟改进项目，验证推荐系统提出的改进方案在仿真执行后的预期效果，评估其可行性与有效性。

④第四阶段：模型部署与闭环进化

模型推理与部署。将模型部署为质量改进系统的“分析大脑”，当系统识别到复杂波动时，自动触发多维度解析，输出根因假设图谱及改进措施推荐列表，并与项目管理工具集成，跟踪改进闭环。**闭环运营与进化。**建立改进效果反馈链路。系统自动关联改进措施的实施记录与实施后的质量绩效数据，将“措施-效果”配对数据作为新的强化学习样本，持续回流，用于优化根因定位与方案推荐模型，形成从“分析”到“行动”再到“学习”的增强闭环。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级“质量宇宙数据湖”，全域集成供应链端物料谱系与过程数据、制造端全要素毫秒级传感与视频数据、使用端产品运行状态与用户反馈数据，以及外部行业标准、学术文献中的质量知识。**数据预处理。**在实现全域融合的基础上，运用因果发现算法对海量时序与关系数据进行自动扫描，初步构建变量间的潜在因果结构网络，为后续深度因果学习提供先验约束与方向。

②第二阶段：样本准备与特征工程

特征工程。构建面向因果推断与策略生成的元特征。例如，构建用于反事实推理的“干预表示特征”；从数据中学习并构建能够表征系统健康状态的“深度表征特征”；动态维护和更新“质量因果知识图谱”中的实体与关系特征。**数据标注。**专家工作聚焦于对因果假设与策略有效性的终极判定。对系统自动发现的潜在因果链进行确认或修正；对自主生成的改进策略进行前瞻性评估与伦理、风险标注，形成驱动系统高阶进化的“监督信号”。**数据增强。**核心是利用“因果数字孪生”进行大规模反事实实验与策略探索。在虚拟环境中，主动实施各种现实中难以尝试的干预组合，观察系统长期演变，生成用于训练鲁棒因果模型和策略优化模型的“合成经验”数据。

③第三阶段：模型训练与仿真验证

模型训练。训练结构因果模型与深度因果推断模型，使其不仅能识别相关性，更能定量评估不同因素对质量结果的因果效应。训练基于强化学习与进化算法的策略自适应生成模型，使其能在满足多重约束下，自主探索并优化改进策略。**模型验证。**验证在极端复杂、长周期、反事实场景下进行。在数字孪生中模拟未知的新型材料失效模式、设备交互性故障等，测试系统能否提前预警、准确定位并生成有效策略。进行“虚拟历史重演”，评估若采用 AI 策略，历史重大质量事故是否可以避免或减轻。

④第四阶段：模型部署与闭环进化

模型推理与部署。将因果模型与策略模型作为“自主质量智能体”的核心，直接嵌入实时质量管控流。智能体具备持续监测、实时诊断、预测风险、生成并动态调整优化策略的能力，在授权范围内可实现闭环控制。**闭环运营与进化。**建立“集体智慧”进化范式。智能体不仅从自身交互中学习，还能吸收人类专家经验、外部知识库，并在数字孪生中进行永无止境的“思想实验”。每一次决策及其长远影响都转化为因果知识，持续丰富和修正知识图谱，使系统成为一个能够预见未知、创造新知、永续进化的“质量伙伴”，最终实现质量管理的完全自主化与智能化。

（二十二）设备运行监控与维护

“设备运行监控与维护”场景涉及入门级、基础级与进阶级的企业。

入门级企业针对生产线上关键设备核心运行参数监控不及时、数据分散等问题，部署传感采集模块，通过人工智能技术识别关键参数监测异常，依托集中平台实现数据自动采集、实时报警，提升设备监控及时性与准确性时，数据治理的目标主要是打破关键设备运行数据分散存储、难以统筹的壁垒，实现数据向集中平台的有效汇聚，解决数据监控不及时的问题，确保数据能够实时反映设备运行状态，为实时报警提供时效支撑，解决数据失真问题，确保输入 AI 算法的数据真实准确，提升异常识别的精准度。

基础级企业针对复杂工况下设备隐性异常难识别、故障预警滞后的问题，引入深度 AI 感知与预测技术，构建融合多维度数据、边缘智能与数字孪生仿真的监控系统，通过 AI 分析实现隐性异常早期识别、健康度动态评估与远程智能调控，提升设备监控的精准性与前瞻性时，数据治理的目标主要是打破单一参数监控的局限，实现多维度数据的有机融合，为深度 AI 感知隐性异常提供全面数据输入，解决复杂工况下数据失真问题，确保数据能够精准反映设备实际运行状态，支撑数字孪生模型精准仿真与 AI 提前预警，解决故障预警滞后问题，确保数据实时流转与联动，支撑动态评估与及时调控。

进阶级企业针对设备全生命周期中数据割裂、预测滞后、维护被动、效能未达最优的问题，引入多模态深度融合与数字孪生全工况仿真技术，构建全链路自主运行的设备智能监控系统，通过 AI 算法实现设备状态的全域感知、故障的精准预判、维护的自主优化与效能的持续提升时，数据治理的目标主要是打破设备设计、采购、安装、运行、维护、报废全生命周期各阶段的数据壁垒，实现数据全链路贯通与有效复用，解决单一数据类型支撑不足的问题，通过多模态数据融合挖掘隐性故障特征，提升全域感知与预判精准度，解决预测滞后问题，确保数据精准反映设备实际状态并实时流转，支撑数字孪生全工况仿真与维护自主优化，让数据体系随设

备状态变化、系统运行需求动态进化，保障智能监控系统的长期自主有效运行。

1.治理对象

表 71　设备运行监控与维护场景的数据治理对象

适用层级	数据类型	数据内容
入门级、基础级、进阶级	设备运行数据	设备运行参数、振动、温度、压力、转速、声音、电流等多维度状态数据、运行时长与负载记录等、工业相机采集的设备状态图像等
入门级	设备状态数据	运行、停机、故障等
基础级	设备基础信息	型号、参数等
基础级、进阶级	维护相关数据	维护记录、维修方案、备件更换信息、维护资源调度情况、故障根因分析、维护效果等
进阶级	环境与生产数据	环境温湿度、粉尘浓度等参数、生产计划、设备负载分配等
入门级、基础级、进阶级	故障与预警数据	故障类型、故障模式、发生时间、部位、根因分析、预警记录等
基础级、进阶级	数字孪生数据	设备全工况仿真数据、异常演化路径模拟、故障演化路径预演结果、维护方案模拟效果等
进阶级	知识图谱数据	设备故障—维护方案—效能影响关联关系、历史解决方案等
进阶级	全生命周期数据	设备采购信息、安装调试记录、历次维护与故障数据、效能变化趋势、模型迭代记录等

2.平台（技术）工具

表 72　设备运行监控与维护场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级	基础级	进阶级
数据治理与集成平台	从单一设备到全域产线多源异构数据的统一接入、清洗、融合与资产化管理。	ThingsBoard、Databuff、TDengine	浪潮海岳物联网平台 inIoT、Wonderware、阿里云日志服务 SLS	华为云 DataArts Studio、西门子 Xcelerator 生态系统、AVEVA System Platform
特征工程与样本管理平台	将设备运行的时序、振动频谱、图像等原始数据，转化为表征健康、异常、劣化趋势的结构化特征，并对故障样本进行标注、增强与管理。	Jupyter Notebook、Pandas/NumPy、Label Studio	Apache Superset、MLflow、Prodigy	英伟达 Cosmos AI 模型、Tecton、华为云 ModelArts Data+
模型开发与运维平台	为预测性维护、视觉检测、故障诊断等 AI 模型提供从开发、训练、仿真验证到部署、监控的 MLOps 全生命周期管理能力，确保模型持续可靠。	Python+Scikit-learn、PyTorch Lightning、Google Colab	华为云 ModelArts Lite、百度 BML、MLflow	华为云 ModelArts、阿里云 PAI、Databricks Lakehouse Platform、Azure Machine Learning
数字孪生与智能应用平台	构建从单体设备到整条产线的高保真、可交互、数据驱动的虚拟孪生环境，深度集成 AI 模型，驱动智能监控、仿真推演、预测性维护与自主决策应用。	Wonderware InTouch HMI、FlexSim	优锘科技 uThings、数字冰雹、Wonderware System Platform	达索 3DEXPERIENCE、英伟达 Omniverse、华为云数字孪生平台、PTC ThingWorx

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过部署振动、温度、压力等基础传感器与数据采集模块，统一采集关键设备的运行参数数据。通过OPC UA、Modbus 等协议，与现有的 SCADA 系统、MES 系统对接，获取设备启停状态、报警代码及生产批次信息。**数据预处理。**对采集的时序数据进行基础清洗、归一化与时间对齐，处理因信号干扰产生的瞬时尖峰、补全因通讯中断造成的短时数据缺失，将不同传感器的物理量统一至标准化范围，将所有数据同步至统一时间戳，确保数据可被集中平台稳定接收与展示。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，计算基础统计特征，如“关键参数最近 N 分钟均值/方差”“与标准设定值的持续偏差”。构建简单的复合指标，如“设备综合运行指数”。**数据标注。**结合 SCADA 系统的历史报警记录与维修工单，对发生报警时段前后的设备运行数据段进行手动标注，标识“正常”“超限报警”“未知波动”等基础标签。**数据划分。**按时间顺序，将标注后的数据集划分为训练集与测试集，用于后续简单的异常识别模型训练。

③第三阶段：模型训练与仿真验证

模型训练。使用统计过程控制方法[66]或简单的机器学习模型，如孤立森林，基于历史正常数据学习各参数的正常波动范围，或训练二分类模型区分“正常”与“已知报警”状态。**仿真验证。**在测试集上评估模型的误报率与漏报率。可

在监控平台中设置模拟数据注入功能，验证报警规则触发的及时性与准确性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练好的阈值或规则模型以配置项形式部署至集中监控平台。平台对实时数据流进行在线计算与比对，触发“参数超限”“趋势异常”等实时报警，并通过看板、短信等方式推送。**闭环进化。**建立报警处理反馈日志。记录每次报警的确认情况、处理措施与实际结果，定期人工复审这些日志，优化报警阈值或调整规则，初步积累故障-现象关联数据。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。扩展传感器维度，增配振动频谱分析、声学、高清工业相机等，采集设备的多模态运行数据。通过边缘网关，集成设备基础信息库、详细的计算机化维护管理系统历史维护记录，并与 MES 的生产工单、工况信息深度关联。**数据预处理。**在边缘侧进行更复杂的数据处理。对振动信号进行时频域转换，提取频谱特征；对声学与图像数据进行降噪与增强；对不同频率、不同单位的传感数据进行特征级融合与对齐。建立初步的设备数字孪生模型，用于生成特定工况下的仿真基准数据。

②第二阶段：样本准备与特征工程

特征工程。应用信号处理与深度学习自动提取深度特征。例如，从振动频谱中提取“共振频率偏移量”“谐波能量分布变化”；基于卷积神经网络从设备外观图像中提取“螺栓松动表征”“油液渗漏区域”；构建“多传感器信息融合的健康度退化趋势向量”。**数据标注与增强。**结合专家经验与故障根因分析报告，对历史数据进行更精细的标注，如“早期磨损”“不对中初期”“润滑不良”。采用合成少数类过采样技术或基于生成对抗网络生成稀缺的故障样本，解决样本不均衡问题。**数据划分。**采用分层抽样，确保训练集和测试集能覆盖设备不同健康状态以及主要工况，验证模型的泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。训练更先进的 AI 模型，如用于剩余使用寿命预测的 LSTM 或 Transformer 模型，用于多故障模式分类的深度学习模型，以及用于异常检测的自编码器。仿真验证。在测试集上评估模型性能。将训练好的预测模型与设备数字孪生模型结合，在仿真环境中模拟“载荷阶跃变化”“部件性能渐变退化”等复杂场景，验证模型预警的提前量、准确性以及在不同虚拟工况下的稳健性。

④第四阶段：模型部署与闭环进化

模型部署与推理。采用云边协同架构。将轻量级模型部署至边缘网关进行实时异常检测，将重型预测模型部署在云端。系统可输出“设备健康评分”“潜在故障模式及概率”

“预估剩余可用时间”及“维护建议优先级”。**闭环进化。**建立模型预测结果与后续维护行动、实际故障发生情况的自动关联分析机制。利用新的维护记录与运行数据，对预测模型进行定期的增量学习与在线校准，不断缩小预测误差，优化健康评估模型。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级设备运维数据湖。全量接入设备全生命周期数据、实时多模态运行数据、环境与精细化生产数据、供应链备件数据，以及高保真数字孪生持续产生的仿真与推演数据。利用知识图谱技术，结构化治理故障模式、影响及诊断分析、维修规程等非结构化知识。**数据预处理。**实施流批一体、时空对齐的深度数据治理。利用知识图谱建立“设备部件-传感器-故障模式-维护动作-效能指标”的语义关联网络，实现数据在语义层面的深度融合与智能检索。对数据进行不确定性量化与置信度标注。

②第二阶段：样本准备与特征工程

特征工程。应用图神经网络、注意力机制等认知计算技术。自动学习跨设备、跨系统的群体性退化规律与传播模式；构建“维护策略-效能回报关联特征”；从知识图谱中推理生成“针对新型复合故障的诊断路径特征”。**数据标注与合成。**系统具备自进化标注能力。通过强化学习智能体在数字孪生环境中的自主探索，产生海量“状态-干预-结果”的决策序列

数据。利用物理信息生成式模型，合成任何历史未发生的、极端或边界条件下的故障演化数据，用于训练系统的未知故障应对能力。**数据划分。**采用基于故障模式、设备类型、工况组合的多维度交叉验证划分策略，确保系统具备应对全新场景的“零样本”或“少样本”快速适应能力。

③第三阶段：模型训练与仿真验证

模型训练。训练具备自主决策能力的多智能体系统。针对全域态势感知智能体，融合多源数据，生成统一设备健康视图；针对故障预测与根因推理智能体，结合数据驱动模型与知识图谱推理；针对维护策略优化智能体，基于深度强化学习，权衡成本、安全、库存与生产损失，生成最优维护计划。**仿真验证。**在数字孪生环境中进行加速应力测试。模拟整条产线、整个工厂在数十年运行中可能遭遇的所有工况、退化、突发事件组合，验证自主决策系统在长期运行中的全局最优性、稳定性以及对“黑天鹅”事件的韧性。

④第四阶段：模型部署与闭环进化

模型部署与推理。将上述智能体集群部署为“设备智能运维中心”。中心不仅能发布预警和报告，更能直接输出“自主生成的预防性维护工单”“动态调整的备件采购建议”“基于能效最优的设备运行参数优化设定值”，并可向下游CMMS、ERP、控制系统下达可执行指令。**闭环进化。**构建“物理设备-数字孪生-运维知识-决策智能”四体协同的永续进化生态。真实世界的每一次交互都驱动数字孪生模型与 AI

模型的迭代；运维知识图谱在每次决策后自动吸收新的案例与规则；系统能够从跨工厂，甚至跨行业的设备群数据中学习普适性规律，实现从企业级优化向生态级协同的跃迁。

（二十三）智能经营决策

“智能经营决策”场景仅涉及进阶级的企业。

针对工厂经营中资源配置低效、决策依赖经验、市场响应滞后等问题，引入多模态融合决策与全域数字孪生推演技术，构建全链路自主运行的智能经营决策系统，通过 AI 算法实现资源全域协同调度、风险收益动态平衡、决策全流程智能进化时，数据治理的目标主要是打破工厂经营各环节的数据壁垒，实现全链路数据贯通，为资源全域协同调度提供全面数据支撑，解决单一数据类型支撑不足的问题，通过多源多模态数据融合挖掘经营风险与收益特征，提升决策科学性，解决市场响应滞后问题，确保数据精准反映经营动态并实时流转，支撑全域数字孪生推演与快速决策，让数据体系随市场环境变化、经营需求升级动态进化，保障智能经营决策系统的长期自主有效运行。

1.治理对象

表 73　智能经营决策场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	多领域经营数据	生产数据、财务报表、销售数据、供应链信息、人力资源数据等
进阶级	市场与外部数据	市场趋势、竞争对手信息、政策法规、宏观经济指标等
进阶级	决策与执行数据	资源配置方案、决策指令、执行进度、效果评估结果等
进阶级	风险与收益数据	风险识别记录、风险概率、影响范围、收益预测、实际收益等
进阶级	数字孪生数据	不同决策方案的仿真结果、场景推演记录、风险预演数据等
进阶级	知识图谱数据	经营要素关联关系、历史决策案例、业务规则等
进阶级	历史与迭代数据	历史经营数据、决策模型迭代记录、自进化系统优化经验等

2.平台（技术）工具

表 74　智能经营决策场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级
数据治理与集成平台	实现跨财务、生产、供应链、市场等多域异构数据的统一接入、清洗、融合与资产化管理。	华为云 DataArts Studio、瓴羊 Dataphin、微软 Azure Purview、袋鼠云数栈
特征工程与样本管理平台	将复杂的经营时序数据、市场文本、业务关系网络转化为可供 AI 模型识别的深度特征，并对“决策情景-结果”样本进行智能化标注、增强与管理。	Kaleido、Beagledata
模型开发与运维平台	为需求预测、资源优化、风险模拟等多智能体决策模型提供从协同开发、仿真训练、数字孪生验证到生产部署、监控、持续迭代的企业级 MLOps 能力，确保决策模型稳定可靠。	阿里云 PAI、百度 BML、EasyModel、蓝卓 supOS
数字孪生与智能应用平台	构建高保真、可模拟、数据驱动的企业经营全景数字孪生环境，深度集成各类决策模型，驱动战略推演、风险预演、资源调度优化等核心智能应用，实现“决策沙盘”式管理。	DataMesh FactVerse、易知微 EasyV、英伟达 Omniverse、达索 3DEXPERIENCE

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级经营数据湖。通过建立统一数据中台，利用 API、ETL、流处理等手段，统一汇聚来自 ERP、MES、SCM、CRM、HRM 等内部业务系统的结构化数据，同时整合外部市场情报、行业数据、宏观经济指标、政策法规等非结构化与半结构化数据。建立与全域数字孪生平台的实时数据通道，确保虚拟推演与物理世界的信息同步。**智能**

化预处理。对汇聚的海量多源异构数据进行深度治理。实施智能化清洗、语义对齐、时空归一化，利用异常检测算法识别并处理财务数据中的离群值、补全供应链数据中的缺失链路，将不同系统中的“客户”“订单”“成本”等核心概念通过本体论进行统一映射与关联，统一所有数据的时间颗粒度与地理/组织维度。核心是初步构建经营知识图谱，将“产品-订单-物料-设备-客户-供应商-人员”等实体及其复杂业务关系进行结构化存储，为后续智能分析奠定语义基础。

②第二阶段：样本准备与特征工程

特征工程。基于知识图谱与原始数据，利用图神经网络、自然语言处理等技术，提取用于高阶决策的深度特征。如构建“供应链韧性指数”“市场机会-产能匹配度向量”“跨部门资源协同效率特征”“基于历史案例的决策风格与效果嵌入表示”。从非结构化市场报告和政策文件中提取情绪倾向、风险关键词等语义特征。**数据标注。**基于历史重大经营决策会议纪要、项目复盘报告、审计结果以及决策产生的实际财务业务成果，对历史决策情景与决策动作进行关联标注，标识决策的“成功”“部分成功”“失败”等结果标签，并标注关键决策要素与约束条件。利用生成式 AI 技术，结合经营规律与蒙特卡洛模拟[67]，合成大量在历史中未曾出现但符合商业逻辑的“极端市场情景”“‘黑天鹅’事件”或“创新型业务模式”样本数据，用于增强决策模型的鲁棒性与前瞻性。**数据划分。**采用基于时间窗口和业务场景的混合划分

策略。确保训练集覆盖完整的商业周期，测试集包含近期的、未见过的复杂组合情景，以检验模型面对新挑战的泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。训练面向复杂经营决策的 AI 模型集群。针对全域态势感知与预测智能体，融合多模态数据，实时生成企业综合健康度仪表盘，并对关键指标进行多周期、多情景预测。针对策略生成与优化智能体，基于深度强化学习，在给定的经营目标和约束下，自动生成资源配置、产品定价、投资组合等策略方案。针对风险仿真与评估智能体，利用因果推断与概率图模型，识别潜在经营风险链，并量化不同决策方案下的风险暴露与潜在损失。**仿真验证。**在构建的“企业级经营数字孪生体”中进行大规模、加速比的仿真推演。该孪生体需模拟从原材料采购、生产制造到产品销售、资金回笼的全价值链动态。将 AI 生成的策略方案输入孪生体，在虚拟环境中推演未来多个季度甚至数年的经营结果，评估其在“原材料价格暴涨”“竞争对手突袭”“政策法规剧变”等上百种扰动情景下的稳健性、鲁棒性与长期价值，实现“决策前验”。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练验证成熟的 AI 模型集群部署为“智能经营决策中枢”的核心引擎。该中枢提供自然语言交互界面，决策者可以输入决策目标、约束条件或直接提问。中枢调用多智能体协同推理，输出“推荐策略方案包”“多

方案对比推演报告”“风险收益热力图”及“关键行动建议”。闭环运营。建立“决策-执行-反馈-学习”的强闭环。物理世界决策的执行结果实时回流，用于校准预测模型和优化策略生成模型。系统自动捕获每一次“人工否决 AI 建议”的上下文与最终理由，将其作为高质量反馈样本进行学习。所有成功的决策逻辑与推演过程自动沉淀，持续丰富和优化经营知识图谱与案例库，使决策中枢不仅从数据中学习，更能从组织的历史经验与专家智慧中学习，实现持续自主进化。

（二十四）数智精益管理

“数智精益管理”场景仅涉及进阶级的企业。

针对经营过程中全要素协同不足、浪费识别滞后、改善方案复制难的问题，引入全要素数字孪生与深度强化学习技术，构建全链路数智精益管理系统，通过前沿 AI 算法实现浪费的实时识别、资源的全域优化、改善的自主生成与全流程智能进化时，数据治理的目标主要是打破经营全链路人、机、料、法、环、测各要素的数据壁垒，实现全要素数据贯通与协同，解决浪费识别滞后问题，确保数据实时反映经营过程状态并精准可靠,支撑前沿 AI 算法快速识别各类浪费，解决改善方案复制难问题，通过数据追溯与复用为改善经验沉淀提供支撑，让数据体系随经营场景变化、精益管理需求升级动态进化，保障数智精益管理系统长期有效运行。

1.治理对象

表 75　数智精益管理场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	全要素生产经营数据	人员定位与作业数据、设备状态参数与利用率数据、物料流转记录与库存信息、工艺标准与执行数据、环境参数等
进阶级	浪费与改善数据	浪费识别记录、浪费根源分析、改善方案、实施进度与效果评估等
进阶级	多工厂与多产线协同数据	各工厂/产线精益指标、资源共享情况、经验复制记录等
进阶级	数字孪生数据	全域精益数字孪生体映射数据、改善方案仿真结果、收益量化验证记录等
进阶级	知识图谱数据	精益管理规则、历史改善案例、浪费与原因关联关系等
进阶级	历史与进化数据	历史精益数据、模型参数迭代记录、系统适应能力提升轨迹等

2.平台（技术）工具

表 76　数智精益管理场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级
数据治理与集成平台	实现覆盖人员、设备、物料、工艺、环境的全要素、多模态异构数据的统一接入、实时清洗、深度关联与资产化管理。	西门子 Xcelerator 生态系统、AVEVA System Platform、华为云 DataArts Studio
特征工程与样本管理平台	将人员动线、设备微停机、物料滞留等原始数据，转化为表征七大类浪费（如等待、搬运、过度加工）的深度时空特征，并对“浪费场景-改善方案”样本进行智能化标注、合成与版本管理。	Tecton、Scale AI、Prodigy、华为云 ModelArts Data+
模型开发与运维平台	为深度强化学习智能体（自主改善模型）及多模态浪费识别模型提供从协同开发、数字孪生环境训练、仿真验证到生产部署、监控的企业级 MLOps 能力，确保模型能稳定驱动自主优化。	英伟达 AI 企业级套件、华为云 ModelArts、阿里云 PAI、微软 Project Bonsai
数字孪生与智能应用平台	构建高保真、可模拟、包含人员与物料动态的全域精益数字孪生环境，深度集成 AI 模型，驱动浪费	英伟达 Omniverse、达索 3DEXPERIENCE、DataMesh Director、PTC ThingWorx

	实时可视化、改善方案仿真推演、自主调度优化等核心智能应用。	

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级精益运营数据湖。部署 UWB 定位、振动传感器、RFID、视觉相机等 IoT 传感器，打通 MES、WMS、EMS、HR 系统以及设备管理系统，实现人员实时位置与作业节拍、设备全维度状态与全局设备效率、物料实时流转与库存水位、工艺参数执行曲线、环境能耗等全要素数据的统一汇聚。同步接入多工厂、多产线的协同运营数据。**数据预处理。**对海量多模态流数据进行深度情境融合与对齐。实施基于业务规则的清洗，并利用时空对齐引擎，将所有数据在统一的“时间-工位-产品-人员”四维坐标下进行精准关联。核心是构建初代全域数字孪生体，将物理实体与数据流实时映射，并基于此孪生体，对原始数据进行情境化打标，为浪费识别提供可直接理解的语义上下文。

②第二阶段：样本准备与特征工程

特征工程。基于情境化数据，利用计算机视觉（CV）、信号处理、图计算等技术，自动提取表征七大类浪费（过度生产、等待、搬运、加工、库存、动作、缺陷）的深度特征。例如，从人员动线数据中提取“无效移动距离与频率”；从设备状态序列中提取“微型停机（<1 分钟）模式”；从物料流图中计算“在制品库存周转率与瓶颈点”；构建“人-机-料

协同效率矩阵”。**改善知识图谱构建与样本标注。**系统化构建精益改善知识图谱。将“浪费现象-根本原因-改善工具-标准作业-改善效果”等概念、规则、历史案例进行结构化关联存储。利用此图谱，对历史数据进行半自动标注，系统自动识别潜在的浪费模式，由精益专家确认并关联根因与改善措施，形成高质量的“情境-浪费-方案”标注样本库。同时，基于物理规则与历史数据分布，利用生成式 AI 合成罕见或复合型浪费场景数据，用于增强模型的识别广度。**数据划分。**采用基于“价值流-产品族”的划分策略，确保训练集能覆盖主要产品类型的完整价值流，测试集包含新产品导入或产线重组等新情境，以验证模型的泛化与迁移能力。

③第三阶段：模型训练与仿真验证

模型训练。训练自主改善智能体。智能体以全域数字孪生体的实时状态为观察空间，以可执行的改善动作（如调整人员站位、微调设备参数、重排序生产队列、触发物料补给）为动作空间，以综合效率提升（如人均产出、OEE、库存周转率）和浪费减少为奖励函数。智能体通过与高保真数字孪生仿真环境的交互试错，自主学习在复杂动态环境中实现多目标优化的最优策略。**仿真验证。**在构建的高保真、多智能体协同仿真的精益数字孪生环境中，对智能体生成的改善方案进行毫秒级加速的“先验验证”。不仅验证单一改善点，更模拟改善方案实施后，在全价值流中可能引发的连锁反应，评估其对上下游工序、物料流、人员负荷的全局影响，确保

方案的稳健性与全局最优性。同时，可进行多方案并行对比推演，量化不同方案的潜在收益与风险。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练成熟的自主改善智能体及配套的浪费识别模型，以“精益优化服务”形式部署。系统 7x24 小时运行，实时监控运营状态，动态输出三类指令：实时预警类，如“XX 工位等待时间超过阈值，原因为上游物料未齐套”；辅助决策类，如“为应对今日订单混合，推荐采用 B 套排产序列，预计可减少换型时间 15%”；自主执行类，如在安全策略允许下，自动下发指令调整 AGV 路径或向协作机器人发送新作业指令。**闭环运营。**构建“物理现场-数字孪生-改善智能体-知识图谱”四轮驱动的永续进化生态。现场实施改善方案的真实效果数据实时回流，用于校准孪生模型和优化智能体策略。每一次成功的改善案例，其完整上下文、实施过程与量化结果，由系统自动提炼、结构化后沉淀至精益改善知识图谱，形成可全局复用的组织资产。系统能够主动识别不同工厂、产线间的相似情境，并推荐经过验证的改善方案，实现精益最佳实践的自动传播与自适应复制。

（二十五）规模化定制

“规模化定制”场景仅涉及进阶级的企业。

针对多品种小批量生产中需求转化低效、设计与生产脱节、成本居高不下的问题，引入生成式 AI 与全域协同优化技术，构建全链路智能定制系统，通过 AI 算法实现需求的

深度解析、设计的自主生成、生产的柔性适配与全流程成本优化时，数据治理的目标主要是打破需求、设计、生产各环节的数据壁垒，实现全链路数据顺畅流转，提升需求转化效率，实现设计数据与生产数据的精准匹配，保障设计方案可落地、生产可柔性适配，解决成本居高不下问题，通过成本数据与全链路数据的关联，为 AI 算法实现全流程成本优化提供支撑，让数据体系随客户需求变化、产品品种拓展、生产能力升级动态进化，保障智能定制系统长期有效运行。

1.治理对象

表 77　规模化定制场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	客户需求数据	文字描述、图片参考、语音记录等多模态需求信息、核心诉求提炼结果、标准化设计参数等
进阶级	设计与方案数据	定制方案、模块化设计参数、可制造性验证结果、模块复用记录等
进阶级	生产与调度数据	定制订单特性、设备状态、物料库存、生产调度策略、工艺参数、换产记录等
进阶级	成本与效益数据	采购成本、调试成本、生产制造成本、成本优化方案、效益评估结果等
进阶级	数字孪生数据	“需求—设计—生产”全流程模拟数据、定制方案预演结果、优化记录等
进阶级	知识图谱与反馈数据	客户需求与方案关联关系、历史定制案例、交付反馈、自进化知识图谱迭代记录等
进阶级	跨系统数据	CRM 系统客户信息、ERP 系统资源数据、规模化定制管理平台数据等

2.平台（技术）工具

表 78　规模化定制场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级
数据治理与集成平台	实现客户多模态数据（文本、图像、语音）、模块化设计数据与柔性生产资源数据的统一接入、语义解析、关联融合与资产化管理。	华为云 DataArts Studio、阿里云 DataWorks、西门子 Teamcenter、达索 3DEXPERIENCE
特征工程与样本管理平台	将非结构化客户描述、设计草图、工艺知识转化为机器可理解的深度语义特征与向量，并对“需求-成功方案”配对数据进行智能化标注、合成生成与版本管理，为生成式 AI 提供高质量“养料”。	Tecton、Scale AI、Prodigy、Labelbox
模型开发与运维平台	为生成式设计模型、需求-生产匹配模型、柔性调度优化模型提供从多模态预训练、数字孪生环境强化学习到生产环境部署、A/B 测试的全栈式 MLOps 能力。	英伟达 AI 平台、阿里云 PAI、华为云 ModelArts、Hugging Face Enterprise
数字孪生与智能应用平台	构建贯穿虚拟客户体验、产品设计仿真、柔性产线运作的全链路数字孪生环境，深度集成生成式 AI，驱动个性化方案实时渲染、可制造性即时验证、定制订单全流程模拟等智能应用。	英伟达 Omniverse、达索 3DEXPERIENCE

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级定制数据湖。通过数据中台统一接入来自 CRM、在线定制平台、社交媒体等多渠道的客户非结构化数据；集成 PLM/PDM 系统中的模块化设计库、历史方案与工艺知识；实时汇聚 MES、APS、WMS 中的生产资源状态；同步接入 ERP 中的成本数据与供应链信息。**数据预处理。**实施面向多模态理解的深度数据治理。对非结构化客

户需求，利用自然语言处理与计算机视觉技术进行意图提取与结构化解析。对设计数据，进行模块化解构与特征编码，将每个设计模块转化为带有一系列可制造性、成本、工艺约束的特征向量。对所有数据实体（客户、需求特征、设计模块、生产资源）进行唯一标识与关联，初步构建“需求-设计-资源”关联图谱。

②第二阶段：样本准备与特征工程

特征工程。基于语义化数据，构建支撑智能匹配与生成的深度特征。需求侧，构建“客户偏好向量”“需求可制造性复杂度预估特征”等；设计侧，构建“设计模块兼容性矩阵”“模块–工艺–成本关联特征向量”“美学风格嵌入向量”等。生产侧，构建“生产线柔性度指数”“动态产能与负载特征”等。**知识增强样本构建与标注。**系统化构建规模化定制知识图谱，将“客户语义-设计模块-制造工艺-成本结构-交付反馈”进行关联。利用此图谱与历史成功案例库，对海量“需求-最终采纳方案”配对进行自动或半自动标注，形成高质量训练样本。针对小众或创新型需求样本不足的问题，采用条件生成式对抗网络[68]，在知识图谱的约束下，合成符合商业逻辑与物理规律的新“需求-方案”配对数据，极大扩充设计空间的探索范围。**数据划分。**采用基于客户群体、产品大类的分层抽样划分策略，确保训练集能覆盖主流需求模式，测试集包含新兴的、跨品类的长尾定制需求，以检验系统的创新与泛化能力。

③第三阶段：模型训练与仿真验证

模型训练。训练面向规模化定制的 AI 模型集群。针对需求-设计生成智能体，基于多模态 Transformer 或扩散模型，接收客户原始输入，直接生成符合设计规范与制造约束的多个候选方案示意图及参数化模型，实现“需求到初步设计”的端到端生成。针对设计-生产匹配与优化智能体，利用图神经网络与强化学习，对生成的设计方案进行可制造性深度分析、成本模拟，并从模块库中推荐最优的模块组合与工艺路线，在满足个性化需求的同时最大化生产共性与效率。针对全局资源调度智能体，基于实时生产资源状态与订单池，对定制订单进行动态分组、排序与资源分配，实现柔性生产线的负载均衡与交付期优化。仿真验证。在构建的“从需求到交付”的全链路数字孪生体中进行方案可行性、经济性与交付可靠性的加速仿真验证。虚拟孪生体需模拟从方案详细设计、物料准备、柔性线生产到物流发运的全过程。在此环境中并行推演多个 AI 生成方案，量化评估其生产周期、综合成本、质量风险，并与传统设计流程进行对比，验证智能系统在效率、成本与创新性上的综合优势。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练成熟的 AI 模型集群部署为“智能定制引擎”，嵌入在线定制平台与内部设计生产系统。前端面向客户提供交互式、引导式的智能共创新体验；后端面向设计、生产部门自动输出“推荐方案包”“精准成本报价”

“最优生产指令”及“物料采购建议”。系统支持“人在环路”的协同模式，专家可对 AI 方案进行微调与确认。闭环运营。构建“市场-设计-生产-服务”数据驱动飞轮。每一笔定制订单的最终交付数据、客户满意度反馈、生产实际成本与耗时，均实时回流至数据湖。系统自动分析“设计预期”与“实际结果”的偏差，用于持续校准成本预测模型、优化设计规则。成功的定制案例自动沉淀至知识图谱，转化为新的设计灵感或模块化选择。系统能够感知市场趋势变化，主动生成新的模块或方案建议，驱动产品模块库的迭代进化，从而实现从响应定制到引领个性化需求的跨越。

（二十六）产品精准营销

“产品精准营销”场景涉及基础级、进阶级的企业。

基础级企业针对客户信息碎片化、需求洞察浅、营销策略经验依赖的问题，引入 AI 驱动的多维度分析与智能决策技术，构建融合客户知识图谱、自然语言深度解析与动态营销优化的精准营销系统，通过分析实现客户隐性需求挖掘、个性化策略生成与营销效果动态优化，提升营销精准度与响应效率时，数据治理的目标主要是打破客户信息分散存储、多源异构的壁垒，实现客户数据的集中整合，为客户知识图谱构建提供基础支撑，解决需求洞察浅的问题，通过客户数据的精准关联挖掘潜在需求特征，为 AI 多维度分析提供支撑，解决营销策略经验依赖问题，确保数据真实可靠，为 AI 生成个性化营销策略提供高质量数据输入，让数据体系随客

户需求变化、营销场景升级动态调整，保障精准营销系统持续有效运行。

进阶级企业针对客户需求隐性化、市场预测滞后、营销转化低效的问题，引入多模态用户洞察与深度强化学习技术，构建全链路智能营销系统，通过 AI 算法实现客户需求的深度挖掘、营销策略的自主生成、营销执行的动态优化与全流程效能提升时，数据治理的目标主要是打破单一数据类型局限，通过多模态数据融合挖掘隐性需求特征，为多模态用户洞察技术提供全面支撑，解决市场预测滞后问题，确保数据实时反映客户动态与市场变化，为深度强化学习算法生成精准策略提供高质量数据输入，解决营销转化低效问题，让数据体系随营销过程动态进化，支撑深度强化学习算法持续优化营销执行策略，防范数据安全风险，确保营销数据全生命周期合规可控，保障智能营销系统长期稳定运行。

1.治理对象

表 79　产品精准营销场景的数据治理对象

适用层级	数据类型	数据内容
基础级、进阶级	客户行为数据	交易记录、社交媒体互动信息、浏览行为轨迹、客户服务沟通记录、浏览轨迹、购买历史、产品偏好等等
基础级、进阶级	客户属性数据	基础信息、消费能力、行业特征等
基础级、进阶级	需求数据	显性需求描述、隐性需求标签、场景化需求特征等
基础级、进阶级	市场与竞品数据	市场趋势分析、竞品营销动态、价格变动、宏观经济指标等
基础级、进阶级	营销执行数据	营销策略方案、个性化内容、报价记录、渠道投放数据、客户响应反馈、转化效果等
基础级	历史营销案例数据	成功经验、失败教训、优化方案等
进阶级	产品与成本数据	产品特性、成本结构、库存状态、价值评估参数等
进阶级	数字孪生数据	营销场景模拟结果、策略预演效果、投入产出比仿真记录等
进阶级	知识图谱数据	客户需求与产品关联关系、营销案例、渠道特性知识、自进化迭代记录等
进阶级	跨系统数据	CRM 系统客户信息、ERP 系统成本数据、营销平台执行数据等

2.平台（技术）工具

表 80　产品精准营销场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现线上、线下、社媒、服务等全域异构客户数据与市场数据的统一接入、清洗、融合、存储与资产化管理，是构建360°客户视图的“总枢纽”。	GrowingIO/CDP、神策数据、火山引擎增长分析	华为云 DataArts Studio、阿里云 DataWorks、网易数帆 DataOps
特征工程与样本管理平台	将客户行为序列、文本评论、图像视频等多模态数据，转化为表征兴趣偏好、购买意向、价值潜力的结构化特征，并对“用户-营销互动”样本进行标注、增强与版本管理。	Jupyter Notebook、Pandas/NumPy、Label Studio、Apache Superset	第四范式 FeaturePro、华为云 ModelArts Data+、Tecton、Prodigy
模型开发与运维平台	为客户细分、响应预测、个性化推荐等营销 AI 模型，提供从开发、训练、仿真验证到部署、监控、持续迭代的 MLOps 全生命周期管理能力。	华为云 ModelArts Lite、腾讯云 TI-ONE、MLflow、Google Colab	华为云 ModelArts、阿里云 PAI、Databricks Lakehouse Platform、Azure Machine Learning
数字孪生与智能应用平台	构建可交互、可模拟、数据驱动的虚拟市场与客户旅程环境，深度集成 AI 模型，驱动营销策略模拟、预算分配优化、个性化内容生成与智能触达等核心应用。	Tableau、Power BI、Adobe Analytics、Google Analytics 360	百度营销“观星盘”、英伟达 Omniverse

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过部署客户数据平台，利用 API、SDK 等方式，统一接入来自电商平台、线下门店 POS、CRM 系统、社交媒体、官网等渠道的客户交易、属性及行为数据。同步

整合内部 ERP 中的产品与成本数据，以及经清洗的外部行业报告数据。**数据预处理。**对多源异构数据进行清洗、去重、归一化与标签化处理，处理 ID 不一致、信息缺失、异常值，识别并合并同一客户在不同渠道的身份标识，统一日期、金额等格式，基于规则对客户进行初步分群，如“高价值客户”“潜在流失客户”。核心是构建统一的用户唯一标识体系的客户主数据，为后续分析提供一致的基础。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，利用统计分析与非监督学习，构建用于客户洞察与预测的特征集。如构建“客户生命周期价值预估”“产品偏好向量”“内容兴趣标签”“购买周期与活跃度指标”。利用自然语言处理技术，从客服记录、社交媒体评论中提取“情感倾向”“核心诉求关键词”等文本特征。**数据标注。**结合历史营销活动的转化效果数据，如点击率、转化率、投入产出比，对过往的营销触达记录（客户特征、营销内容、渠道）进行关联标注，标识“高响应”“低响应”“负面反馈”等结果标签，形成监督学习样本。**数据增强与划分。**针对高价值转化或特定小众客群样本不足的问题，采用基于相似度的样本合成进行数据增强。按时间顺序将数据集划分为训练集与测试集，确保模型评估的时间有效性。

③第三阶段：模型训练与仿真验证

模型训练。使用标注后的数据集，训练基础的 AI 模型，如用于预测客户响应概率的机器学习分类模型、用于客户精细化分群的聚类模型、用于推荐产品的协同过滤模型。**策略模拟与验证。**在测试集上评估模型的预测准确率与召回率。进一步，可构建简单的营销效果模拟器，基于历史转化率，对不同客群、不同渠道、不同内容的组合策略进行投入产出比的模拟计算与对比，辅助营销人员制定更具数据依据的初步策略。

④第四阶段：模型部署与闭环进化

模型部署与推理。将训练验证通过的模型集成至营销自动化平台或 CDP 中。营销人员通过系统界面，选择目标客群与营销目标，系统调用模型进行辅助推理，输出“目标客户推荐列表”“个性化内容建议”“最佳触达渠道预测”及“预计转化率区间”。**闭环进化。**建立营销活动执行数据（曝光、点击、转化）的自动回流机制。定期利用新产生的营销效果数据对模型进行增量训练与参数调优，使模型能够适应市场与客户偏好的变化，形成初步的“数据-策略-执行-反馈-优化”运营闭环。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级营销数据智能平台。深度集成全渠道实时行为流数据（网站/APP 点击流、广告曝光与点击流）、多模态交互数据（语音客服录音转文本、直播与短视频互动

数据、AR/VR 体验数据）、外部生态数据（第三方 DMP 数据、全网舆情监测、竞品数字足迹）。实现与供应链数字孪生、产品数字孪生的数据联动。**实时预处理与情境构建。**实施流批一体的实时数据处理。对实时行为流进行会话分割与意图识别；对多模态数据（图像、视频、语音）进行深度特征提取与跨模态对齐，如将用户评论的文本情感与对应产品的视觉特征关联）。核心是构建动态的“客户情境图谱”，实时融合客户“当前所处场景”“历史偏好”“实时情绪”“社交影响力”等多维信息，形成深度洞察。

②第二阶段：样本准备与特征工程

特征工程。应用深度学习与图神经网络技术，提取用于深度决策的认知特征。如构建“客户决策心理路径模拟特征”“跨渠道旅程连贯性与摩擦点分析”“微观市场细分与趋势预测特征”“内容创意元素与客群响应关联的嵌入向量”。**生成式样本构建与自动化标注。**系统化构建营销知识图谱，将“客户画像-产品卖点-创意内容-渠道特性-转化归因”进行深度关联。利用此图谱与强化学习环境，自动化生成海量的“虚拟客户-虚拟营销动作-虚拟回报”交互序列作为训练数据。同时，利用生成式 AI，基于成功案例自动生成符合品牌调性、适配不同客群的营销文案与创意变体，极大丰富营销策略的探索空间。**数据划分。**采用基于客户旅程阶段与市场细分的多维交叉验证划分，确保模型不仅能预测短期转化，更能理解并优化客户的长期价值与全生命周期体验。

③第三阶段：模型训练与仿真验证

模型训练。训练自主营销决策智能体。该智能体以实时“客户情境图谱”和“营销资源状态”（预算、库存、渠道容量）为观察空间，以个性化内容生成、渠道选择、出价策略、优惠力度等为动作空间，以长期客户价值最大化和营销投资回报率为奖励函数。智能体通过在与高保真“营销数字孪生”环境的交互中持续学习最优策略。**营销数字孪生推演验证。**构建“客户-市场-竞品”交互的营销数字孪生体。该孪生体模拟真实市场环境中海量虚拟客户的复杂决策行为。在此环境中对智能体生成的海量营销策略进行并行、加速的 A/B/n 测试与蒙特卡洛仿真，预演策略在“竞品突然降价”“热点事件爆发”“宏观经济转向”等复杂动态场景下的表现，量化评估其鲁棒性、适应性与长期收益，实现“策略未上线，效果已预知”。

④第四阶段：模型部署与闭环进化

自主系统部署与实时决策。将训练成熟的智能体部署为“实时智能营销引擎”。该系统能够 7x24 小时自动运行，实时捕捉市场机会与客户意图变化，自动执行诸如“动态调整程序化广告出价”“向高意向客户推送个性化产品视频”“在客户犹豫时自动发放定制化优惠券”等优化动作，实现营销的“自动驾驶”。**生态化进化与价值网络协同。**构建“真实市场-数字孪生-决策智能体-知识图谱”的四轮驱动进化生态。真实世界的每一次营销互动与商业结果都实时反馈，用于校

准孪生模型和优化智能体。系统能够自动从跨业务线、跨品牌的成功实践中提炼模式，沉淀至营销知识图谱。更进一步，可与合作伙伴在隐私计算框架下进行联邦学习，在数据不出域的前提下共同优化模型，实现从企业智能到生态协同智能的跃迁。

（二十七）远程运维服务

“远程运维服务”场景涉及基础级、进阶级的企业。

基础级企业针对产品运维依赖现场、故障处理低效、资源浪费等问题，引入 AI 驱动的智能诊断与协同决策技术，构建融合多模态数据感知、智能故障预测与数字孪生仿真的远程运维系统，通过 AI 分析实现故障早期预警、根因自动定位与远程精准指导，大幅降低现场服务频次，提升运维效率时，数据治理的目标主要是打破单一数据类型局限，通过多模态数据融合为远程智能诊断提供全面支撑，减少对现场数据采集的依赖，解决故障处理低效问题，确保数据实时反映产品状态并真实可靠，支撑远程精准决策，解决资源浪费问题，通过全链路数据贯通明确故障根因与资源配置逻辑，防范数据安全风险，确保运维数据全流程可追溯，保障远程运维系统长期稳定运行。

进阶级企业针对复杂设备隐性故障难识别、突发故障处理滞后、维护资源错配的问题，引入多模态融合诊断与数字孪生远程协作技术，构建“监测-诊断-预测-维护-进化”全链路智能远程运维平台，通过 AI 算法实现故障的早期预警、

精准定位、远程协同处理与维护策略自优化时，数据治理的目标主要是打破设备运维各环节数据壁垒，实现全链路数据深度联动，为多模态融合诊断挖掘隐性故障特征提供全面支撑，解决隐性故障难识别问题，通过多模态数据精准融合提升数据价值密度，为 AI 诊断算法提供高质量数据输入，解决突发故障处理滞后问题，确保数据实时流转且真实可靠，为远程协同处理提供高效数据支撑，解决维护资源错配问题，推动数据体系随运维需求动态进化，支撑维护策略持续优化。

1.治理对象

表 81　远程运维服务场景的数据治理对象

适用层级	数据类型	数据内容
基础级、进阶级	设备状态监测数据	振动、温度、电流等传感器实时数据、工业相机图像、设备日志、声学信号、隐性故障特征记录等
基础级、进阶级	故障与诊断数据	故障类型、发生时间、部位、异常特征、故障现象描述、根因分析结果、诊断记录等
基础级、进阶级	维护与协作数据	维护计划、维修步骤、操作记录、维修效果评估、远程协作记录、工程师与现场人员的沟通文本、AR 标注信息、现场操作反馈、维护效果评估等
基础级、进阶级	设备基础信息数据	型号参数、设计图纸等
进阶级	全生命周期数据	设备出厂信息、安装调试记录、历次维修数据、故障历史、健康度变化轨迹等
进阶级	资源与计划数据	维护资源调度、生产计划、备件库存信息、需求预测等
基础级、进阶级	数字孪生数据	设备三维模型数据、故障模拟结果、维修方案效果模拟、远程协作场景仿真记录等
基础级、进阶级	知识图谱数据	故障模式与解决措施关联、维修经验、设备手册解析内容、自进化迭代记录等

2.平台（技术）工具

表 82　远程运维服务场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现设备多模态传感器、运维系统、知识文档等全域异构数据的统一接入、清洗、关联融合与资产化管理。	ThingsBoard、TDengine、阿里云物联网平台	华为云 DataArts Studio、西门子 Xcelerator 生态系统、AVEVA System Platform
特征工程与样本管理平台	将设备振动、图像、日志、文本等原始数据，转化为表征健康状态、故障模式、退化趋势的结构化特征，并对“故障-维修”场景样本进行标注、增强与版本管理。	Jupyter Notebook、Label Studio、Pandas/NumPy	华为云 ModelArts Data+、Tecton、Prodigy、Scale AI
模型开发与运维平台	为故障诊断、预测性维护、远程指导等 AI 模型提供从开发、训练、仿真验证到部署、监控、持续迭代的 MLOps 全生命周期管理能力。	华为云 ModelArts Lite、百度 BML、MLflow	华为云 ModelArts、阿里云 PAI、Azure Machine Learning、Seldon Core
数字孪生与智能应用平台	构建高保真、可交互、数据驱动的设备数字孪生体与远程协作环境，深度集成 AI 模型，驱动三维可视化监控、故障模拟推演、AR 远程指导与维护策略优化等核心智能应用。	优锘科技 uThings、Wonderware System Platform、PTC Vuforia	英伟达 Omniverse、达索 3DEXPERIENCE、PTC ThingWorx、微软 Azure Digital Twins

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过在关键设备上加装振动、温度、电流等传感器及工业相机，利用边缘网关统一采集设备多模态运行数据。通过标准协议（如 OPC UA、MQTT）对接设备 PLC、SCADA 系统获取运行日志与报警代码。同步接入 CMMS 中的维修工单、备件库存数据，以及产品基础信息库。**结构化预处理。**对采集的流式与批量数据进行清洗、对齐，滤除传感器噪声、处理通讯丢包，统一多源数据时间戳至毫秒级，并进行特征初提取，对振动信号进行 FFT 转换[69]获取频谱，对图像进行边缘检测与缺陷区域初判。建立“设备-传感器-报警-工单”的初步数据关联，为后续分析提供结构化输入。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，利用信号处理与图像处理技术，构建用于故障识别的特征集。如从振动频谱中提取“特征频率幅值变化”“边带能量”；从电流波形中提取“谐波畸变率”；从热成像图中提取“温度梯度异常区域”；构建“多传感器健康度综合指标”。**数据标注。**结合历史维修报告、故障代码及专家经验，对历史数据段进行半自动标注，标识“正常”“轴承磨损早期”“润滑不良”“叶片结垢”等故障模式标签，形成监督学习样本。**数据增强与划分。**针对某些特定故障模式样本稀少的问题，采用时序数据变换（如加噪、伸缩）和图像数据增强（如旋转、裁剪）等方法扩充样本。按设备序列号或时间顺序，将数据集划分为训练集与测试集。

③第三阶段：模型训练与仿真验证

模型训练。使用标注后的数据集，训练用于故障分类的机器学习模型（如随机森林、支持向量机）及用于异常检测的自编码器模型。训练基于规则的简单根因推理模型。**仿真验证。**在测试集上评估模型分类准确率与召回率。进一步利用初步构建的设备三维数字孪生模型，将模型诊断出的故障部位、类型进行三维可视化呈现，并关联预置的维修知识库，模拟生成初步的维修指导步骤（如拆装顺序、注意事项），供专家在线审核与修正，验证诊断结果的可解释性与指导性。

④第四阶段：模型部署与闭环进化

模型部署与远程服务。将验证通过的诊断模型与知识库集成至远程运维平台。当系统检测到设备异常或触发预警时，自动生成包含“疑似故障类型”“可能根因”“故障部位三维高亮图示”“初步维修建议”的电子工单，并推送至现场工程师的移动终端（如 AR 眼镜、平板），实现“数据+模型”驱动的远程精准指导。**反馈优化闭环。**建立现场维修结果反馈机制。工程师在完成维修后，通过移动终端反馈实际故障原因、维修操作及效果。平台将这些反馈数据与之前的预警、诊断记录进行自动比对分析，定期利用新的验证数据对诊断模型进行增量训练与校准，优化预警阈值，提升诊断准确性。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级设备健康管理数据湖。扩展传感器类型（如声学、超声波、油液监测），并深度融合 MES 的生产任务与工况数据、ERP 的备件供应链数据、气象与环境数据。实现与高保真设备数字孪生体的实时数据同步与指令交互。**实时情境构建。**实施流式复杂事件处理。对多模态实时数据流进行在线特征提取、融合与情境判断，如将“特定频率振动能量上升”“油温轻微偏高”“当前处于高负载生产阶段”关联为一个“亚健康退化”情境事件。核心是构建动态的“设备健康情境图谱”，实时融合物理状态、运行负载、维护历史、环境因素，形成对设备健康状况的深度、上下文感知。

②第二阶段：样本准备与特征工程

认知级特征工程。应用深度学习与图神经网络进行端到端特征学习与关联挖掘。如利用 CNN-LSTM 网络[70]从原始振动信号中直接学习退化趋势特征；利用 GNN 分析“部件-故障-症状-维修措施”知识图谱，学习故障传播路径与维修方案的有效性特征；构建“设备全生命周期健康轨迹嵌入向量”。**自生成样本工厂与自动化标注。**系统化构建并持续进化运维知识图谱。利用强化学习智能体在与高保真数字孪生环境的交互中，自主探索设备在各种工况、负载、磨损状态下的行为，自动生成海量“状态-动作-结果”序列数据。同时，利用生成式 AI 合成极端、罕见故障模式下的多模态数据（如模拟特定裂纹扩展的声发射信号），极大丰富训练数据的边界

与多样性。**数据划分。**采用基于设备类型、故障模式、工况组合的交叉验证划分，确保模型具备强大的泛化能力，能够识别从未在历史数据中出现过的新型复合故障。

③第三阶段：模型训练与仿真验证

深度预测与决策模型训练。训练下一代 AI 模型集群。针对剩余使用寿命预测智能体，基于 Transformer 或深度生存分析模型，实现高精度的个体化设备 RUL 概率预测。针对多智能体协同维护优化智能体，每个关键设备或产线作为一个智能体，通过多智能体强化学习，在综合考虑自身健康状态、生产计划、备件库存、专家资源的情况下，协同学习出全局最优的预防性维护计划与资源调度策略。针对远程协同诊断智能体，基于多模态融合和知识图谱推理，实现故障的自动定位与诊断，并能自动召集、调度异地专家资源，在数字孪生环境中开展虚拟会诊。**孪生协同仿真与决策推演。**在超高保真、多物理场耦合的设备数字孪生体中进行大规模仿真。不仅模拟故障演化，更模拟完整的远程协作维修过程：数字孪生体同步真实设备状态，异地专家通过 AR/VR 接口接入同一孪生环境，在其中进行虚拟拆卸、测量、方案讨论与操作演练，验证维修方案可行性后再指导现场执行，实现“先虚后实，虚实协同”。

④第四阶段：模型部署与闭环进化

自主运维平台部署。将成熟的 AI 模型集群部署为“自主运维大脑”。该平台能够自动执行动态健康评估与预警、预测性维护工单自动生成与派发、备件需求自动预测与申领、远程专家资源智能匹配与协同会话召集。平台甚至可以在安全策略允许下，对设备进行远程参数调整或启停控制。**生态化进化与集体智慧沉淀。**构建“物理设备群-数字孪生网络-运维智能体-工业互联网平台”的四层进化生态。海量设备在运行中产生的群体数据，通过隐私计算技术在保障数据主权的前提下，用于训练更强大的全局预测模型。每一次成功的远程诊断与维修案例，其完整过程被自动抽象、沉淀至运维知识图谱，并可通过工业互联网平台向同类设备用户安全共享最佳实践，实现从企业级智能运维到行业级运维知识共创的跃迁。

（二十八）客户主动服务

“客户主动服务”场景涉及入门级、基础级与进阶级的企业。

入门级企业针对客户咨询渠道分散、人工效率低、响应慢的问题，搭建 AI 客服平台，整合渠道，通过自然语言处理技术和计算机视觉基础处理文本与图像，实现简单问题自动回复、复杂问题智能派单与进度追踪，提升服务效率与响应速度时，数据治理的目标主要是打破各咨询渠道数据壁垒，实现客户咨询数据集中整合，为 AI 客服全域响应提供基础

支撑，解决人工效率低问题，通过规范文本与图像数据格式，提升 AI 识别与处理精度，解决响应慢问题，确保数据实时流转且真实可靠，提升派单效率与追踪精准度。

基础级企业针对客户需求多样化、服务个性化不足、问题响应低效等问题，引入 AI 驱动的深度交互与主动服务技术，构建融合多模态数据感知、知识图谱深化与智能决策的客户主动服务系统，通过 AI 分析实现精准需求理解、个性化服务推送与产品改进闭环，提升客户满意度与服务效率时，数据治理的目标主要是打破客户数据分散壁垒，实现全域数据整合，为精准理解多样化需求提供全面数据支撑，解决服务个性化不足问题，通过多模态数据规范融合提升数据价值密度，支撑 AI 精准匹配服务需求，解决响应低效问题，确保数据实时反映客户动态且真实可靠，支撑主动服务与智能决策，推动服务与产品持续优化，让数据体系随客户需求变化动态进化，支撑产品改进闭环形成。

进阶级企业针对客户隐性需求未被深度挖掘、产品服务迭代与客户期望脱节、共创参与不足的问题，引入多模态融合理解与数字孪生共创技术，构建全链路智能服务系统，通过 AI 算法实现客户需求的深度解析、服务方案的自主生成、产品迭代的协同共创与全流程体验优化时，数据治理的目标主要是打破单一数据类型局限，通过全域多模态数据深度融合挖掘隐性需求特征，为多模态融合理解技术提供全面支撑，解决产品服务迭代与客户期望脱节问题，通过数据深度关联

与可信校验，为 AI 自主生成服务方案、精准匹配迭代方向提供高质量数据支撑，解决共创参与不足问题，确保数据实时反映客户共创需求与体验反馈，支撑全流程体验动态优化，推动数据体系与产品服务迭代协同进化，形成“数据驱动-共创优化-反馈迭代”的良性循环。

1.治理对象

表 83　客户主动服务场景的数据治理对象

适用层级	数据类型	数据内容
入门级	客户基本信息数据	姓名、联系方式、所购产品信息等
入门级	服务工单数据	工单编号、问题类型、处理状态、责任人等
基础级、进阶级	多模态客户需求数据	语音通话记录、文字反馈内容、行为轨迹数据、图像视频素材、隐性需求提炼结果、标准化需求参数等
入门级、基础级、进阶级	客户服务数据	文本咨询、图像信息、通话记录、邮件内容、社交平台留言、在线聊天记录等服务记录、问题处理过程、解决方案、满意度评价等
基础级	客户画像数据	基本信息、消费偏好、产品使用习惯等
基础级、进阶级	产品相关数据	产品型号、功能特性、常见问题、解决方案、产品手册、维修案例、迭代历史、改进方案等
基础级、进阶级	共创与交互数据	客户参与方案调整记录、AR/VR 交互信息、共创反馈等
基础级、进阶级	市场与迭代数据	市场反应、产品迭代效果、客户反馈的痛点、改进建议、客户反馈转化记录等
基础级、进阶级	数字孪生数据	“客户-产品-服务”模拟数据、改进方案仿真结果等
基础级、进阶级	知识图谱数据	客户问题与根因关联、服务经验、需求与方案匹配关系、自进化迭代记录等

2.平台（技术）工具

表 84　客户主动服务场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例		
		入门级	基础级	进阶级
数据治理与集成平台	实现客户交互渠道、业务系统等多源异构数据的统一接入、清洗、融合（One-ID 客户视图）、存储与资产化管理。	腾讯云智聆、网易七鱼、环信	神策数据、火山引擎增长分析、GrowingIO/CDP	华为云 DataArts Studio、阿里云 DataWorks、Databricks
特征工程与样本管理平台	将客服文本、语音、图像、行为序列等多模态数据，转化为表征客户意图、情绪、偏好、价值的结构化特征，并对“服务-结果”样本进行标注、增强与管理。	Jupyter Notebook、Pandas/NumPy、SnowNLP	Label Studio、Prodigy、火山引擎 VeDI	Tecton、Scale AI、Snorkel
模型开发与运维平台	为意图识别、智能推荐、服务机器人等 AI 模型提供从开发、训练、验证到部署、监控、持续迭代的 MLOps 全生命周期管理能力。	腾讯云 TI 平台、百度 EasyDL、MLflow	阿里云 PAI、华为云 ModelArts Lite、Hugging Face Spaces	阿里云 PAI、华为云 ModelArts、Databricks Lakehouse Platform
数字孪生与智能应用平台	构建可交互、可模拟的虚拟客户、产品与服务环境，深度集成 AI 模型，驱动服务策略模拟、体验预演、AR 远程协作与智能营销等应用。	企业微信/钉钉、腾讯云慧眼	DataV、网易有数、Tableau	英伟达 Omniverse、DataMesh Director

3.治理方案

（1）入门级企业

①第一阶段：数据集成与标准化预处理

数据集成。部署统一客服平台或中间件，通过 API、网页插件等方式，整合官网、微信公众号、小程序、电商平台等文本咨询入口。对接现有工单管理系统，获取客户基本信息、产品购买记录与历史工单。结构化预处理。对咨询文本进行基础清洗、标准化和关键词提取，去除无关字符、纠正错别字，统一时间格式、工单状态代码，基于规则匹配常见问题关键词。将非结构化咨询初步分类，如“产品咨询”“故障报修”“投诉建议”，并与工单系统进行自动关联，形成结构化的“咨询-工单”初始记录。

②第二阶段：样本准备与特征工程

特征工程。基于预处理数据，构建用于问题分类的简单特征集。如利用 TF-IDF[71]提取咨询文本的关键词向量；基于产品型号、问题关键词组合构建分类规则特征。**数据标注与划分。**由人工对历史咨询记录进行分类标注。对标注好的数据集按时间顺序或随机划分为训练集与测试集，用于训练基础分类模型。

③第三阶段：模型训练与仿真验证

模型训练。使用标注数据集，训练文本分类模型。训练意图识别模型，用于区分“需自动回复”“需转人工”“需创建工单”等意图。**流程验证。**在测试集上评估分类准确率

与意图识别准确率。在客服平台模拟常见咨询场景，验证自动回复话术的恰当性及工单自动创建、流转的流程通畅性。

④第四阶段：模型部署与闭环进化

模型部署与服务上线。将训练好的模型集成至客服平台，实现 7x24 小时在线自动应答。对于复杂问题，系统自动生成标准化工单并分配给相应坐席或工程师，同时向客户发送工单进度通知。**闭环优化。**收集自动回复的客户满意度评分、工单解决时长等数据。定期分析模型误判案例，优化分类规则和关键词库，对模型进行增量训练，提升自动化处理的准确率。

（2）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。建设企业级客户数据平台，统一接入呼叫中心语音转文本、企业邮箱、社交媒体官方账号等全渠道交互数据。同时，独立对接产品主数据管理系统，获取结构化产品知识库。深度特征提取与画像构建。对语音文本进行情感极性分析；对客户历史所有交互记录进行序列分析，提取服务偏好特征。基于此，独立构建动态的客户细分画像，包含“价值等级”“产品偏好”“服务敏感度”“潜在流失风险”等标签，形成服务决策的基础。

②第二阶段：样本准备与特征工程

知识驱动的特征工程。构建“产品-问题-解决方案”初级知识图谱，将产品功能模块、常见故障代码、标准操作指南

进行关联。利用此图谱，对客户咨询进行语义解析，提取“可能涉及的故障模块”“历史相似案例”等深度特征。**高质量服务样本库建设。**选取过去一年内被标记为“圆满解决”且客户满意度高的服务案例，由专家团队进行深度复盘，标注出“核心需求”“关键解决步骤”“服务亮点”等，形成高质量“成功服务”样本库，用于训练服务推荐模型。

③第三阶段：模型训练与仿真验证

模型训练。训练个性化服务推荐模型，基于客户画像与实时问题，从知识图谱和成功案例库中推荐最优解决方案或服务产品；训练主动服务触发模型，基于产品使用数据（如联网设备运行日志）与客户画像，预测客户潜在服务需求（如备件更换、功能升级），并生成外呼或推送提醒。测试验证。将训练好的模型在小范围真实客户群体中进行 A/B 测试，对照组采用原有服务模式，实验组采用模型推荐的服务策略，严格对比客户满意度、问题解决时长等核心指标，验证模型有效性。

④第四阶段：模型部署与闭环进化

智能服务中台独立部署。将模型集群部署为“智能服务推荐中台”，以 API 形式赋能全体客服坐席。坐席工作台实时显示客户画像与推荐方案，并自动提示可能的主动服务机会。**产品改进价值闭环。**建立独立的分析流程，定期从智能中台提取高频问题、共性痛点，形成结构化产品改进需求报

告，直接向产品经理团队推送，驱动产品迭代，形成的“服务反馈-产品优化”价值流。

（3）进阶级企业

①第一阶段：数据集成与标准化预处理

全要素、全链路数据融合。建设体验数据平台，全面接入客户语音、视频、AR/VR 交互、产品物联网传感器数据、使用行为序列等全模态数据。同时，直接对接市场情报系统与竞品数据库。**实时情境感知与动态图谱构建。**运用多模态大模型技术，实时融合分析客户交互中的语音情绪、文本意图、图像内容及产品实时状态。构建动态的“客户-产品-场景”情境图谱，深度刻画客户在特定使用场景下的体验、阻碍与潜在期望。

②第二阶段：样本准备与特征工程

生成式特征与方案探索。基于情境图谱，利用生成式 AI 自动生成对客户隐性需求的多维度解释假说，并进一步生成多种可能的创新服务概念或产品改进原型描述。构建虚拟共创沙盒。利用数字孪生技术，为明星产品或复杂系统构建高保真虚拟模型。在此基础上，通过模拟海量不同用户角色、不同使用环境下的交互，生成包含异常、创新使用方式在内的合成行为与反馈数据，极大扩展训练数据的边界和多样性。

③第三阶段：模型训练与仿真验证

智能体训练。针对需求深度解析与概念生成智能体，基于多模态输入，深度推理客户未言明的根本需求，并生成初

步的服务创新或产品改进概念方案。针对数字孪生协同仿真智能体，在虚拟产品孪生体中，模拟不同改进方案对产品性能、用户体验及服务流程的影响，进行量化评估与迭代优化。**沉浸式共创环境验证。**在基于 VR/AR 技术的“虚拟共创实验室”中，邀请种子客户与内部专家，以虚拟化身形式对智能体生成的概念方案进行沉浸式体验、评估与再设计，验证共创流程的可行性与价值。

④第四阶段：模型部署与闭环进化

部署开放式客户共创平台。将上述能力整合为面向公众或核心客户社区的开放式平台。客户可提交创意、参与虚拟产品测试、对改进方案投票。AI 作为“共创协理”，提供灵感激发、方案细化与模拟验证支持。**驱动价值网络自进化。**形成“客户创意-智能深化-虚拟验证-共识形成-落地转化”的快速循环。成功的共创成果直接转化为研发输入或新服务产品，并反馈激励参与客户。平台持续从海量共创互动中学习，形成不断进化的群体智慧知识网络，最终驱动企业从产品制造商向客户共创型生态组织演进。

（二十九）供应商数字化管理

“供应商数字化管理”场景涉及基础级、进阶级的企业。

基础级企业针对供应商信息分散、评价标准不统一、风险识别滞后的问题，引入 AI 驱动的智能分析与决策技术，构建融合供应商知识图谱、多源数据融合与动态评价模型的数字化管理平台，通过 AI 分析实现供应商智能分级、风险

动态预警与全流程协同，提升供应链管理的精准性与前瞻性时，数据治理的目标主要是打破供应商信息分散存储、多源异构的壁垒，实现全域数据集中整合，为供应商知识图谱构建提供基础支撑，解决评价标准不统一问题，通过数据标准化处理提升数据质量，为 AI 动态评价模型提供高质量数据输入，解决风险识别滞后问题，确保数据实时反映供应商状态并真实可靠，为风险动态预警提供及时支撑，让数据体系随供应链管理需求动态进化，保障数字化管理平台持续有效运行。

进阶级企业针对复杂供应商网络中风险隐蔽性强、评价维度固化、寻源协同低效的问题，引入多模态风险感知与深度强化学习技术，构建全链路智能供应商管理系统，通过 AI 算法实现供应商风险的实时预警、评价的动态多维、寻源的精准智能与全流程效能提升时，数据治理目标主要是打破单一数据类型局限，通过全域多模态数据深度融合挖掘隐蔽风险特征，为多模态风险感知技术提供全面支撑，解决评价维度固化问题，通过数据深度关联与可信校验，为深度强化学习算法实现动态多维评价提供高质量数据支撑，解决寻源协同低效问题，确保数据实时反映供应商状态与寻源需求变化，支撑精准寻源与全流程协同，推动数据体系与供应商管理全流程协同进化，形成“数据驱动-智能决策-执行反馈-迭代优化”的良性循环。

1.治理对象

表 85　供应商数字化管理场景的数据治理对象

适用层级	数据类型	数据内容
基础级、进阶级	供应商基础与动态数据	供应商资质信息、企业信用、产能规模、合作历史记录、财务报表、生产能力数据、实时履约状态等
基础级、进阶级	多系统业务数据	订单信息、合同文本、沟通记录、物流跟踪数据、质检结果、交付周期、采购金额、付款记录等
基础级、进阶级	外部环境数据	行业动态、市场趋势、舆情信息、政策法规变化、同类供应商表现、原材料价格波动等
基础级、进阶级	风险与评价数据	风险事件记录、审计报告、行业风险趋势、风险评估结果、潜在风险点预警、评价指标、评分结果、历史评价记录、评价维度权重调整记录等
进阶级	寻源与供应链数据	采购需求、供应商组合方案、供应链成本、效率及稳定性指标、韧性模拟结果等
进阶级	数字孪生数据	“供应商-物料-供应链”仿真数据、不同选择方案的预演结果、风险应对策略模拟效果等
进阶级	知识图谱数据	供应商关联关系、风险传导路径、评价标准与案例、自进化迭代记录等

2.平台（技术）工具

表 86　供应商数字化管理场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例	
		基础级	进阶级
数据治理与集成平台	实现客户多渠道、多模态交互数据的统一接入、清洗、融合（One-ID 客户视图）、存储与资产化管理，是构建企业级客户数据基座、支持个性化服务的“中枢系统”。	火山引擎 VeCDP、中关村科金 CDP	华为云 DataArts Studio、阿里云实时计算 Flink 版
特征工程与样本管理平台	将客户对话文本、交互行为、产品使用等原始数据，转化为表征客户意图、偏好、价值、情绪的结构化特征，并对“服务-结果”样本进行高效标注、合成生成与版本化管理。	Jupyter Notebook、Label Studio、Apache Superset	Tecton、Prodigy、Scale AI
模型开发与运维平台	为意图识别、智能推荐、服务策略生成等 AI 模型提供从开发、训练、仿真验证到部署、监控、持续迭代的企业级 MLOps 全生命周期管理能力。	华为云 ModelArts Lite、百度 BML、MLflow	华为云 ModelArts、阿里云 PAI、Databricks Lakehouse Platform
数字孪生与智能应用平台	构建可交互、可模拟、可共创的“客户-产品-服务”全链路数字孪生环境，深度集成 AI 模型，驱动服务策略推演、虚拟产品体验、AR/VR 远程协同与智能营销等核心应用。	企业微信/钉钉、腾讯云慧眼	英伟达 Omniverse、达索 3DEXPERIENCE、DataMesh Director

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。通过数据中台或 ETL 工具，统一接入来自 ERP 系统的订单、交货、质量数据；SRM 系统的基础信息与资质文件；财务系统的付款与成本数据。整合外部公开的企业信用信息、行业风险报告等结构化数据。**标准化预处理。**对多源数据进行清洗、对齐、归一化，修正人工录入错误、统一供应商编码与名称，将不同系统的采购物料编码、订单号进行关联映射，将交付周期、合格率等绩效指标统一量纲与统计口径）。核心是构建统一的供应商主数据，并初步建立“供应商-物料-交易”的基础关联关系。

②第二阶段：样本准备与特征工程

量化特征工程。基于预处理数据，构建用于评价与风险识别的量化特征集。如计算“近 12 个月平均订单履约率”“质量批次合格率波动系数”“交付周期稳定性指数”；从舆情报告中提取“负面新闻提及频次”；基于财务数据计算“速动比率变化趋势”。**风险样本标注与划分。**结合历史风险事件记录，如重大质量问题、交付严重延迟、供应商破产及人工评估结果，对相应时间段内的供应商运行数据（绩效特征、舆情特征）进行标注，标识“高风险”“中风险”“低风险”等标签。对标注后的数据集，按时间顺序划分为训练集与测试集。

③第三阶段：模型训练与仿真验证

模型训练。使用标注数据集，训练适用于供应商场景的AI模型。如训练供应商综合绩效评分模型，基于多个量化特征的加权聚合或机器学习回归；供应商风险分类模型，如逻辑回归、随机森林用于区分高风险与低风险供应商。**策略模拟验证。**在测试集上评估模型的准确率与召回率。可在平台中构建简单的规则引擎，模拟基于模型评分结果的供应商分级策略，并回溯分析该策略与历史实际管理决策的吻合度与潜在改进空间。

④第四阶段：模型部署与闭环进化

模型部署与智能应用。将验证通过的模型集成至新建或现有的供应商数字化管理平台。平台可自动输出“供应商动态绩效看板”“风险预警清单”“推荐供应商分级结果”，并为采购决策提供数据支持。**闭环进化。**建立管理决策反馈机制。收集基于模型预警所采取的行动及其后续实际效果，定期利用新的交易数据与反馈标签对风险模型进行重训练与校准，对绩效模型的权重进行优化调整。

（2）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级供应链网络数据湖，深度集成物联网数据（如关键物料在途GPS轨迹、仓储温湿度）、供应商端实时生产状态数据（通过协同平台）、高维外部数据（全球海运/空运数据、地缘政治风险指数、社交媒体舆情流、实

时大宗商品价格）。实现与供应链数字孪生体的数据同步。**实时情境构建与关联。**实施流批一体的复杂事件处理。实时关联如“某地区极端天气预警”“某港口拥堵指数上升”“供应商 A 主要产线停机维修公告”“在途物料 B 预计延迟到港”等多源事件，构建动态的“供应商-物料-物流-环境”风险情境图谱。对非结构化合同、审计报告进行自然语言处理，提取关键条款与风险条款。

②第二阶段：样本准备与特征工程

网络化深度特征工程。应用图神经网络、时序预测等先进技术。提取“供应商在网络中的中心性与依赖度特征”“风险沿供应链拓扑结构的传导概率特征”；构建“多源替代供应商组合的韧性指数”；从海量文本中提取“潜在合规与 ESG 风险语义特征”。**生成式风险样本工厂。**系统化构建并持续进化供应链风险知识图谱，整合历史案例、行业报告、专家经验。利用生成式对抗网络与基于代理的建模，在数字孪生环境中模拟生成各种极端、复合型供应链中断场景，如“疫情封锁+原材料暴涨+汇率剧烈波动”，及其对供应商网络的影响数据，构建超大规模、多样化的风险样本库，用于训练模型的预测与抗压能力。

③第三阶段：模型训练与仿真验证

模型训练。训练面向供应链韧性的 AI 决策模型。针对多模态风险感知与预测智能体，融合结构化与非结构化数据，实时评估并预测单个供应商及供应商网络的整体风险敞口。

针对动态寻源与库存协同优化智能体，基于深度强化学习，以总拥有成本、供应链韧性、可持续性等多目标为优化方向，在动态约束下（需求波动、产能限制、库存水平）自主生成最优的供应商选择、订单分配及安全库存策略。**供应链数字孪生推演验证。**在构建的“全球供应商-制造-物流”高保真数字孪生体中进行亿级规模的蒙特卡洛仿真。将 AI 生成的寻源与库存策略注入孪生体，在虚拟世界中模拟未来数年可能遭遇的各种随机与极端事件，量化评估策略在成本、服务水平和韧性等多维目标下的长期表现与稳健性。

④第四阶段：模型部署与闭环进化

智能系统部署与自主决策。将成熟的模型集群部署为“智能供应商管理中枢”。该系统能够自动执行：实时风险预警与预案推送、动态供应商绩效看板与自动分级、基于多目标优化的采购建议生成、在规则允许下的自动订单分发与协同指令下达。**生态协同进化。**构建“物理供应链网络-数字孪生网络-管理智能体-工业互联网平台”四层协同进化生态。通过与核心战略供应商建立安全数据通道，在保护商业机密的前提下，共享预测性需求与产能信息，实现协同计划。系统持续从内外部网络的实际运营中学习，沉淀最佳实践至知识图谱，并能将成功的风险管理与协同模式通过工业互联网平台向生态内合作伙伴赋能，实现从企业级智能管理到供应链网络级协同智能的跃迁。

（三十）采购计划协同优化

“采购计划协同优化”场景仅涉及进阶级的企业。

针对市场波动大、需求预测滞后、计划与供应链各环节协同不足的问题，引入多模态融合预测与数字孪生跨域协同技术，构建全链路智能采购计划系统，通过 AI 算法实现需求的精准预判、计划的动态优化、上下游的实时协同与全流程效能提升时，数据治理的目标主要是打破市场、需求、采购、生产、仓储、供应商等各环节的数据壁垒，实现全链路数据顺畅流转，提升供应链跨域协同能力，解决需求预测滞后问题，通过多源多模态数据融合挖掘市场波动规律，提升需求预判精准度，应对市场波动大的挑战，确保数据实时反映市场与供应链动态，支撑采购计划动态调整与上下游实时协同，让数据体系随市场环境变化、供应链升级需求动态进化，保障智能采购计划系统长期有效运行。

1.治理对象

表 87　采购计划协同优化场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	需求与市场数据	客户订单信息、生产计划、市场趋势分析、宏观经济指标、物料需求预测结果等
进阶级	库存与生产数据	实时库存状态、库存预警信息、生产约束条件、产能数据等
进阶级	采购与成本数据	采购计划、采购成本明细、批量折扣信息、运输成本、应急采购记录等
进阶级	供应商数据	供应商报价、交付能力、履约记录、状态变化信息等
进阶级	跨域协同数据	生产异常报告、各系统协同指令、跨部门协作记录等
进阶级	数字孪生数据	不同采购计划的仿真结果、协同策略预演效果、供应链指标模拟数据等
进阶级	知识图谱与迭代数据	采购知识图谱关联关系、历史采购案例、市场波动应对经验、自进化模型迭代记录等

2.平台（技术）工具

表 88　采购计划协同优化场景的数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例
		进阶级
数据治理与集成平台	实现跨域（需求、生产、采购、供应商、市场）多源异构数据的统一接入、清洗、语义关联与资产化管理，是构建一体化、可追溯的供应链协同数据基座的核心。	达索 3DEXPERIENCE、华为云 DataArts Studio、西门子 Xcelerator 生态系统
特征工程与样本管理平台	将市场波动时序、供应商履约文本、物流异常等多模态数据，转化为表征供应风险、协同效率、成本结构的深度特征，并对“风险-应对”场景样本进行智能化标注、合成生成与版本管理。	Tecton、华为云 ModelArts Data+、Prodigy
模型开发与运维平台	为多模态需求预测、动态采购优化、供应链仿真推演等 AI 模型，提供从开发、数字孪生环境训练与验证到生产部署、监控的全链路 MLOps 能力，确保决策模型在高动态环境下的稳定可靠。	华为云 ModelArts、阿里云 PAI、Azure Machine Learning
数字孪生与智能应用平台	构建涵盖供应商网络、生产节点、物流路由的全局供应链数字孪生环境，深度集成预测与优化模型，驱动需求-供给模拟、风险预演、动态寻源与多级库存优化等智能决策应用。	达索 3DEXPERIENCE、英伟达 Omniverse、华为云数字孪生平台

3.治理方案

（1）进阶级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级供应链数据湖。通过数据中台统一接入来自 ERP 的销售与生产计划、CRM 的客户订单与预测、WMS 的实时库存与在途数据、SRM 的供应商绩效与报价信息；整合外部市场大数据平台、宏观经济数据库、行业指数以及航运/物流平台的实时运价与时效数据。建立与制造

数字孪生、物流数字孪生的实时数据通道。**数据预处理。**实施面向不确定性管理的深度数据治理。对时序数据进行多尺度分解与异常模式识别；对非结构化数据进行自然语言处理与关键事件抽取；对所有数据进行时空对齐与实体解析。核心是初步构建供应链知识图谱，将“物料-需求-库存-产能-供应商-物流”等实体及其动态关系进行结构化存储，为协同优化提供语义网络基础。

②第二阶段：样本准备与特征工程

特征工程。基于知识图谱与预处理数据，构建支撑全局优化的深度特征。需求侧，构建“需求波动性指数”“跨产品需求关联特征”“外部市场信号与内部需求的领先滞后关系向量”等。供给侧，构建“供应商综合韧性评分”“物流通道可靠性指数”“多源采购组合风险对冲特征”。协同侧，构建“计划-库存-采购协同度指标”“采购提前期与生产节拍匹配度特征”“成本-服务水平多目标权衡曲面特征”等。**风险增强样本构建与标注。**系统化构建采购风险与协同知识图谱，整合历史供应中断案例、价格暴涨事件、紧急调拨记录及其应对措施与结果。利用此图谱对历史采购决策数据进行标注，标识决策的“成本最优”“风险规避成功/失败”“协同效率高/低”等标签。为应对“黑天鹅”事件样本稀缺，采用生成式对抗网络与蒙特卡洛模拟相结合，在知识图谱约束下，合成各类极端供应风险场景及其关联的供应链数据，用于增强模型的抗风险与应急决策能力。**数据划分。**采用基于

时间周期与风险场景的混合划分策略。训练集需包含完整的经济周期波动，测试集则聚焦于近期的、复杂的多重风险交织场景，以检验系统在高不确定性环境下的决策稳健性。

③第三阶段：模型训练与仿真验证

模型训练。训练面向供应链协同的 AI 模型集群。针对多模态需求感知与预测智能体，融合内部订单与外部的宏观、舆情数据，进行概率性、多情景的需求预测，并输出预测不确定性区间。针对动态采购计划优化智能体，基于深度强化学习，以总拥有成本最小化、服务水平达标、供应链韧性最大化为综合目标，在动态约束下（库存、产能、供应商状态）生成滚动采购计划与订单分配策略。针对供应商协同与风险预警智能体，利用图神经网络实时分析供应商网络状态，预测潜在履约风险，并自动生成协同建议。**仿真验证。**在构建的“供应商-工厂-仓库-客户”全链路供应链数字孪生体中进行压力测试与方案寻优。该孪生体需模拟从供应商生产、跨国运输、工厂生产到成品分销的动态过程。将 AI 生成的采购计划与协同策略注入孪生体，在虚拟环境中加速推演未来数月乃至数年的运营情况，承受数百种随机扰动（需求突变、港口拥堵、汇率波动等），全面评估计划在成本、韧性、可持续性等多维度的表现。

④第四阶段：模型部署与闭环进化

模型部署与协同决策服务。将训练验证成熟的 AI 模型集群部署为“智能采购协同中枢”。该中枢与 ERP、SRM、

APS 等系统深度集成，能够实时响应变化，输出三类核心指令：前瞻性预警，如“预计 X 物料在未来 3 个月供应紧张，建议启动备选供应商认证”；动态计划建议，如“根据当前库存与最新需求预测，建议将 Y 物料的采购量上调 15%，并提前 2 周下单”；自动协同指令，如在规则允许下，自动向选定供应商发布订单、触发物流预约。**闭环运营。**构建“物理供应链-数字孪生-决策智能体”三位一体的学习闭环。真实世界的每一次供应事件的执行结果都回流用于校准孪生模型和优化智能体策略。系统自动分析每一次人为干预决策背后的深层逻辑，将其转化为结构化规则或新的训练样本。所有成功的风险应对与协同案例，自动沉淀至供应链知识图谱，使系统能够跨越单一工厂，从整个供应商网络的历史经验中学习，实现从企业智能到网络协同智能的进化。

（三十一）供应链智能调度与物流协同

“供应链智能调度与物流协同”场景仅涉及进阶级的企业。

针对供应链全域透明度低、风险响应滞后、物流资源协同低效的问题，引入多模态风险感知与数字孪生全域协同技术，构建全链路智能供应链系统，通过 AI 算法实现供应链风险的实时感知、物流资源的动态优化、跨环节的自主协同与全流程韧性提升时，数据治理的目标主要是打破供应链采购、生产、仓储、物流、分销、终端、供应商等全环节的数据壁垒，实现数据全链路顺畅流转，提升供应链全域透明度，

解决风险响应滞后问题，通过多源多模态数据融合挖掘风险特征，提升风险感知的及时性与精准度，解决物流资源协同低效问题，确保数据实时反映供应链动态并真实可靠，支撑物流资源动态优化与跨环节协同，让数据体系随供应链环境变化、风险形态升级、协同需求优化动态进化，保障智能供应链系统长期有效运行与韧性持续提升。

1.治理对象

表 89　供应链智能调度与物流协同场景的数据治理对象

适用层级	数据类型	数据内容
进阶级	供应链全域数据	供应商产能数据、仓储库存信息、物流运输状态、订单需求变化、外部环境数据等
进阶级	风险与异常数据	潜在风险信号、风险类型、蔓延路径预测、异常处理记录等
进阶级	物流与调度数据	物流资源信息、物流路线规划、仓储调度方案、订单优先级、成本约束参数等
进阶级	数字孪生数据	不同调度方案的模拟结果、风险应对策略预演效果、供应链指标仿真数据等
进阶级	知识图谱与迭代数据	供应链各环节关联关系、历史调度案例、风险应对经验、自进化模型迭代记录等

2.平台（技术）工具

表 90　供应链智能调度与物流协同场景的

数据治理平台（技术）工具

支撑平台/工具	核心功能	工具示例 进阶级
数据治理与集成平台	实现覆盖供应商、工厂、仓库、物流全节点的多源异构数据（运营数据、IoT 传感数据、外部环境数据）的统一接入、实时清洗、深度关联与资产化管理。	华为云 DataArts Studio、阿里云 DataWorks、Cloudera Data Platform
特征工程与样本管理平台	将物流时序、风险文本、网络拓扑等原始数据，转化为表征运输可靠性、网络脆弱性、协同效率的深度时空与图谱特征，并对“风险-应对”场景样本进行智能化标注、合成生成与版本管理。	Tecton、Scale AI、Prodigy
模型开发与运维平台	为多智能体协同调度、风险预测、路径动态优化等 AI 模型提供从协同开发、数字孪生环境训练与验证到生产部署、监控的全链路 MLOps 能力，确保模型在复杂动态环境中的决策稳定。	华为云 ModelArts、阿里云 PAI、Azure Machine Learning、Ray
数字孪生与智能应用平台	构建涵盖物理流、信息流、资金流的高保真供应链网络数字孪生环境，深度集成 AI 模型，驱动全局可视化监控、风险模拟推演、调度方案动态优化与自主协同决策等核心智能应用。	英伟达 Omniverse、达索 3DEXPERIENCE、AnyLogic、FlexSim

3.治理方案

（1）基础级企业

①第一阶段：数据集成与标准化预处理

数据集成。构建企业级供应链网络数据湖。通过部署物联网关与建立供应链数据中台，统一接入内部 ERP、WMS、TMS、MES 系统的结构化数据，并深度整合外部多源数据：包括供应商协同平台的实时产能与履约状态、第三方物流的 GPS 轨迹与运输状态、港口/海关的公开作业数据、气象与交

通实时信息、市场舆情及地缘政治风险情报。建立与各节点（工厂、仓库）数字孪生体的实时数据通道。**数据预处理。**实施面向复杂事件处理的流式数据预处理。对多源异构流数据进行实时清洗、对齐，过滤传感器漂移、纠正人工录入偏差，将所有数据统一映射到“订单-物料-载具-地理位置”四维时空坐标，并与复杂事件关联，如将“某港口拥堵新闻”“某航线船舶延迟到港 AIS 信号”“仓库特定物料库存下降速率加快”关联为一条潜在的“供应中断风险链”。核心是初步构建供应链网络知识图谱，将“供应商-物料-工厂-仓库-运输路线-客户”等实体及其动态业务关系进行结构化存储，为智能分析提供语义网络基础。

②第二阶段：样本准备与特征工程

特征工程。基于知识图谱与情境化数据，利用图计算、自然语言处理等技术，提取支撑风险预警与协同优化的深度特征。包括网络拓扑特征，构建“节点关键度（介数中心性）”“路径冗余度”“网络脆弱性指数”；风险感知特征，构建“多模态风险信号融合向量（融合舆情情感分析、物流延迟模式、天气异常指数）”“风险传导概率与影响范围预估特征”。协同优化特征，构建“多目标（成本、时效、碳足迹）权衡曲面特征”“动态需求与柔性运力匹配度”“跨设施负载均衡与缓冲库存协同特征”。**风险样本工厂构建与标注。**系统化构建供应链风险与应对知识图谱，整合历史中断事件、近失事件及其应对措施的完整案例。利用此图谱，结合历史

物流调度日志与异常报告，对过去的海量运营决策点进行半自动化回溯标注，标识决策的“规避风险”“放大风险”“实现最优”等结果标签。为应对极端罕见风险样本稀缺，采用“生成式对抗网络+基于代理的建模”方法，在数字孪生环境中模拟生成成千上万种包含复合型风险的虚拟供应链场景及其完整数据序列，构建一个超大规模的“风险-决策-结果”样本工厂。**数据划分。**采用基于“风险场景-网络子图”的划分策略。训练集需覆盖主要产品流经的供应链子网络及常见风险模式；测试集则设计为包含全新的、跨网络的级联风险冲击，以检验模型在面对未知结构扰动时的泛化与自适应能力。

③第三阶段：模型训练与仿真验证

模型训练。训练面向供应链网络协同调度的 AI 模型集群。针对全域风险感知与预测智能体，基于图神经网络与时空注意力机制，实时融合多模态风险信号，预测潜在中断的发生概率、波及路径与业务影响。针对动态调度与资源优化智能体，采用多智能体深度强化学习，每个智能体（如仓储智能体、运输智能体）在观察局部状态的同时，通过通信与协作，共同学习实现全局成本、服务水平和韧性最优的实时调度策略（如库存动态调配、运输路径实时重规划、订单优先级调整）。针对自主协同谈判智能体，在规则约束下，模拟与外部合作伙伴（如供应商）进行简单的资源交换或承诺谈判，以优化全局网络效能。**数字孪生压力测试与推演。**在

构建的“供应商-物流-生产-分销”高保真供应链数字孪生体中进行亿级规模的加速蒙特卡洛仿真与对抗性测试。该孪生体需模拟实体流动、信息流、资金流的相互作用。将 AI 生成的调度策略注入孪生体，在虚拟世界中承受持续随机的“压力测试”，如模拟持续一年的随机需求波动、设施随机故障、燃油价格波动，并主动注入“对抗性攻击”，如模拟关键枢纽同时失效）全面、量化地评估调度策略的稳健性、适应性及长期韧性。

④第四阶段：模型部署与闭环进化

模型部署与自主协同服务。将训练成熟的 AI 模型集群以“云-边”协同架构部署为“智能供应链调度中枢”。云端中枢进行全局优化计算与风险监控，边缘节点（区域配送中心、工厂）执行实时响应。系统输出三类指令：一是风险预警与预案，如“预计华南区域强降雨将影响 X 线路运输，建议提前启动 Y 备用线路并调整 Z 仓库出库计划”；二是实时调度指令，如“根据当前网络状态，建议立即将 A 仓库的冗余库存调拨至 B 仓库，并指派车辆 C 执行”；三是协同建议，如“建议与供应商 D 协商将下周订单提前发货，以对冲潜在船期延误风险”。**闭环进化与知识沉淀。**构建“物理供应链网络-数字孪生网络-多智能体系统”三元耦合的永续学习生态。真实世界每一次扰动与应对的实际效果数据，持续用于校准孪生体模型和优化智能体策略。系统自动捕获并分析不同节点、不同合作伙伴的本地化优秀实践与应急智慧，将其

抽象、提炼后沉淀至供应链网络知识图谱。这使得系统不仅能从自身历史中学习，更能从整个生态网络的集体经验中学习，实现从企业级智能到生态级协同智能的进化，使供应链网络具备“免疫记忆”般的韧性提升能力。

五、面向 AI 的数据基础设施建设方案

企业的数据基础设施由硬件、软件、模型算法、标准规范、机制设计等构成，从数据要素价值释放的角度出发，面向企业提供数据采集、汇聚、传输、加工、流通、利用、运营、安全等服务。网络设施、算力设施与数据基础设施紧密相关，并通过迭代升级，不断支撑数据的流通和利用。

面向 AI 的数据基础设施是制造业实现人工智能应用的重要支撑，企业通过泛在的感知网络、弹性的算力调度和系统的数据治理，将碎片化的工业数据转化为标准化、高质量、易获取的数据资产，为 AI 模型提供持续可靠的“数据燃料”与“计算动力”，使企业能够从单点智能试点走向全域智能协同，助力实现数据驱动的生产优化、质量提升和业务创新。

（一）边缘感知

边缘感知是 AI 数据基础设施的数据源，是连接物理制造世界与数据智能世界的第一道桥梁，负责采集生产环境中人、机、料、法、环、测等全要素的实时状态信息，将物理世界的连续信号转化为离散数据，为上层的 AI 应用提供原始“数据燃料”。

感知层关键在于形成“感知-采集-传输”的完整链条，由传感器与仪器、边缘设备与智能终端、标识与定位系统、工业通信系统构成。根据企业数字化基础与转型目标，可采取不同策略，建立适用于企业实际的感知系统。

1.入门级

企业数字化基础薄弱、自动化水平参差不齐，设备老旧居多、信息化系统未实现全覆盖、以手工记录数据为主，AI应用仅限于个别技术验证性项目，如基于规则的单点图像识别尝试。该阶段的企业 AI 应用尚处探索期，需建立基础数据采集与处理能力。

建议以“单点切入，由点到面”的思路进行感知层建设，以建设小范围的数据网络为下一阶段目标，重点解决数据“有无”和“联通”问题，从低成本、轻量化物联网改造入手，避免大规模硬件投入。

（1）核心组件

传感器与仪器：基础物理量传感器、简易工业相机、能耗计量仪表。

边缘设备与智能终端：数据采集和协议转换设备，工业平板电脑，手持扫码终端。

标识与定位系统：条码识别设备、RFID 基础设备、区域定位触发器。

工业通信系统：工业以太网交换机，工业 Wi-Fi 接入点，现场总线通信模块。

（2）基础条件

企业关键生产设备需具备传感器安装的物理接口和供电条件，车间网络需实现基础覆盖（百兆以太网或 Wi-Fi），关键工序需有明确的数据采集需求定义，运维团队需具备基础的设备接线和网络配置能力。

（3）部署方法

优先选择故障率高、影响大的关键设备，采用非侵入式安装方式快速部署标准化传感器套件，通过 4G/工业 Wi−Fi 等无线网络传输数据，重点解决核心参数的可视化监控问题，在 1−2 周内完成首个试点并验证数据价值。

2.基础级

企业数字化基础良好，拥有自动化产线和控制系统（PLC/SCADA），关键工艺参数和设备状态已在线，已部署若干单点 AI 应用，但数据维度限制模型性能提升，多模态融合应用缺乏基础。该阶段的企业 AI 应用进入规模化推广期，需建立体系化数据支撑能力。

建议以“打通整合，场景深化”的思路进行感知层建设，以主要产线的感知升级为下一阶段目标，重点解决 OT 与 IT 的数据壁垒问题，构建产线级数据采集网络，实现数据在边缘侧的预处理和实时分析。

（1）核心组件

传感器与仪器：多参数传感器、高分辨率工业相机和视觉处理器、温湿度等多环境参数监测仪。

边缘设备与智能终端：智能边缘网关、工业触控一体机、手持式移动分析仪。

标识与定位系统：高频 RFID 读写系统、UWB 或蓝牙定位基站、基于深度学习的视觉识别装置。

工业通信系统：工业环网交换机、专用频段工业无线系统、支持实时数据交换的工业以太网设备。

（2）基础条件

企业产线需全面支持标准工业通信协议（如 OPCUA、ModbusTCP），车间需完成工业环网建设，设备维护团队需熟悉 PLC/SCADA 系统数据接口，拥有明确的设备数据资产清单和采集频次要求。

（3）部署方法

对整条产线进行感知网络规划，通过工业协议网关统一采集 PLC、SCADA 系统数据，补充视觉、声学等新型传感器填补数据空白，采用工业以太网+无线 AP 的混合网络架构，建立标准化的设备接入规范和安装工艺，实现 OT 数据的全面数字化。

3.进阶级

企业实现高度自动化与数字化，关键数据已实现全面采集，已广泛应用 AI 于质量控制、设备预测等领域，但现有感知体系的实时性、同步精度与数据保真度无法满足新一代自主智能系统与实时仿真优化对数据质量的要求。

建议以“体系创新，智能驱动”的思路进行感知层建设，以建设工厂级实时优化与跨链协同的“感知-控制”一体化智能网络为目标，重点解决全域全要素数据采集实时性、精度与反馈问题，支撑实时数字孪生与闭环自主决策。

（1）核心组件

传感器与仪器：多智能传感器阵列、多光谱成像检测设备、微机电系统和纳米级检测仪器。

边缘设备与智能终端：高性能边缘服务器、增强现实工业终端、巡检机器人和无人机系统。

标识与定位系统：全息标识管理系统、厘米级精度实时定位基站、支持运动物体识别和追踪的智能定位系统。

工业通信系统：支持时间敏感网络的确定性工业交换机、5G/6G 工业专网基站设备、支持多协议融合的统一通信管理平台。

（2）基础条件

企业需已完成生产环境的 5G/TSN 网络规划与覆盖，关键设备需具备数字原生接口或改造条件，拥有专业的传感技术团队，建立了完善的传感器生命周期管理和校准体系，与设备供应商建立感知能力联合开发机制。

（3）部署方法

以工厂级顶层架构设计为基础，制定涵盖设备、产线到车间的三级感知标准与网络拓扑图。通过与设备供应商开展“数字原生”联合研发，将 μs 级同步传感器、标准数据接口及初始数字孪生模型预集成为新一代智能装备的出厂标准配置。配套建设以 5GTSN 为骨干的确定性融合网络，部署支持多模态数据实时融合与轻量 AI 推理的智能边缘节点。

最终通过统一管理平台实现所有感知资产的数字孪生化全生命周期管控。

（二）算力中心

算力中心是制造业 AI 数据基础设施的智能计算与处理核心，是支撑人工智能模型训练、推理和数据处理任务的关键动力系统，为 AI 算法提供必需的运算能力和存储资源，将原始数据转化为智能洞察和执行指令，是实现制造智能化从数据到价值转化的计算基础。

算力中心关键在于搭建多层次计算架构，由基础计算单元、存储系统、传输网络、虚拟化与容器化[72]工具、资源调度系统构成。根据企业数字化基础与转型目标，可采取不同策略，建立适用于企业实际的算力架构。

1.入门级

企业数据分散在多个孤立系统中，缺乏统一的数据处理平台，AI 项目往往需要从各个系统手动抽取数据，进行复杂的 ETL 处理后才能用于模型训练。数据处理流程非标准化，重复工作量大,无法支撑 AI 模型的快速迭代和规模化应用。关键在于构建以云端数据湖为核心的基础数据处理架构，建立统一数据存储层，实现关键数据源的自动化接入，建立基础数据质量检查机制。

（1）核心组件

基础计算单元：通用计算服务器、基本配置的 GPU 计算卡、工业边缘计算设备。

存储系统：机械硬盘存储阵列、企业级固态硬盘、基础备份存储设备。

传输网络：千兆工业以太网交换机、基础防火墙设备、无线接入点。

虚拟化与容器化工具：服务器虚拟化软件、容器运行时环境、镜像仓库。

资源调度系统：基础作业调度器、监控告警系统、资源配置管理工具。

（2）基础条件

企业需具备稳定可靠的互联网接入（专线带宽≥20Mbps），关键业务系统（如ERP）具备数据导出接口，网络带宽满足数据传输需求，业务部门能够提供明确的数据质量要求，至少配备1名具备SQL和脚本处理能力的数据处理人员。

（3）部署方法

选择单一公有云平台，部署对象存储作为数据湖基座。针对每个数据源开发独立的数据接入管道，每日定时执行数据抽取和加载任务。建立基础的数据验证规则，对入库数据进行完整性检查。数据处理流程全部通过配置化工具实现，降低技术门槛。

2.基础级

企业已建立基本的数据仓库和处理流程，但无法满足AI应用对实时性、高并发和复杂数据处理的需求。数据响应业

务变化缓慢，难以支撑 AI 应用场景。关键在于构建流批一体的数据处理架构,建成支撑实时 AI 应用的数据处理能力。

（1）核心组件

基础计算单元：高性能计算服务器集群、专业 GPU 计算系统、AI 推理加速设备、可编程逻辑器件。

存储系统：全闪存存储阵列、分布式存储系统、数据分层管理软件、备份恢复系统。

传输网络：高性能数据中心交换机、低延迟网络设备、工业协议转换网关、网络管理系统。

虚拟化与容器化工具：企业级虚拟化平台、容器编排平台、服务发现系统、配置管理工具。

资源调度系统：高性能计算作业调度系统、容器资源调度器、资源监控分析平台、计费管理系统。

（2）基础条件

企业需拥有标准化数据中心基础设施已建立数据治理组织和工作机制，IT 团队需具备虚拟化、容器化和 GPU 服务器管理能力，业务部门能够明确实时数据处理的性能指标。

（3）部署方法

在现有数据仓库基础上，新增流处理集群。建立统一的数据接入层，支持批量和实时两种接入模式。部署特征平台，将常用的特征计算逻辑标准化、服务化。建立端到端的数据血缘视图，实现数据处理全链路可观测。部署实时数据质量监控，设置关键指标的异常预警规则。

3.进阶级

企业面临海量多模态数据（图像、视频、点云、时序数据）的处理挑战，需要支撑数字孪生、工业大模型等复杂 AI 应用，传统数据处理架构在性能、成本和扩展性方面遇到瓶颈，关键在于构建智能数据计算平台，实现多模态数据处理和智能编排，支撑千亿参数工业大模型的训练数据预处理体系。

（1）核心组件

基础计算单元：超大规模 AI 训练集群、存算一体设备。

存储系统：大规模分布式存储系统、计算型存储设备、持久内存系统、智能数据管理平台。

传输网络：确定性时间敏感网络设备、光交换系统、高速互联设备、软件定义网络控制器。

虚拟化与容器化工具：多云容器管理平台、服务网格系统、无服务器计算框架、GPU 虚拟化系统。

资源调度系统：智能资源编排引擎、联邦调度系统、预测性调度算法、能效优化管理系统。

（2）基础条件

企业需拥有专业的 AI 算力架构师和高性能计算团队，与芯片厂商、云服务商建立深度技术合作，制定 3-5 年的长期算力发展规划，年度算力投资预算充足。

（3）部署方法

构建支持多种计算范式的异构集群，针对图像、文本、时序等不同类型数据优化处理流水线。部署数据编织架构，通过知识图谱和语义层实现数据的智能发现和自动集成。建立隐私计算平台，支持在数据不出域的前提下进行联合建模。部署智能编排引擎，基于数据血缘和资源状态动态优化数据处理流程。建立数据产品运营机制，将数据处理能力包装成可度量的数据服务。

（三）数据中心

数据中心是制造业 AI 数据基础设施的数据价值转化中心，是企业级的数据能力复用平台，通过系统化地汇聚、治理、建模和服务化各类数据资产，将原始数据转化为可复用、可共享、可运营的数据智能，为前端 AI 应用和业务决策提供标准化、高质量的数据服务支撑，是实现数据驱动智能制造的核心枢纽。

数据中心关键在于构建数据价值链，由数据集成模块、数据治理模块、数据运营模块、数据开发模块构成。根据企业数字化基础与转型目标，可采取不同策略，建立适用于企业实际的数据中心。

1.入门级

企业数据来源分散、缺乏标准，数据视图和管理流程尚未统一。关键在于实现最基础的云端数据汇聚，优先打通 1-

2 个最核心业务系统的数据，快速响应管理层的数据分析需求。

（1）核心组件

数据集成模块：自动化数据采集工具、标准化数据接口、基础数据同步机制。

数据治理模块：关键业务实体统一编码系统、基础数据字典、数据质量校验规则引擎、基础访问权限控制系统。

数据运营模块：可视化业务指标看板工具、移动端数据查看应用、基础数据使用监控系统。

数据开发模块：数据清洗转换工具、SQL 查询分析环境、定时任务调度器。

（2）基础条件

企业需具备核心业务系统（如 ERP、MES）的基本数据导出接口，IT 团队至少具备基础的数据库管理和脚本开发能力，管理层需明确至少 1-2 个优先解决的数据痛点场景，并具备年度数据平台建设预算。

（3）部署方法

采用云端 SaaS 化数据中台服务快速部署。对接 ERP 等核心业务系统数据，建立基础数据仓库和可视化报表；采用低代码配置方式实施数据集成和治理规则，重点围绕生产、质量等关键场景构建数据服务，完成从数据接入到业务应用验证的完整闭环。

2.基础级

企业各业务系统相对独立，虽已建立部分数据仓库和报表体系，但数据口径不一，跨部门数据共享困难。关键在于建设企业级统一数据资产平台，实现数据的一次加工、多处复用，并提供统一的、高质量的数据服务。

（1）核心组件

数据集成模块：实时工业数据采集平台、多源异构数据融合引擎、可视化数据管道编排系统。

数据治理模块：全链路元数据管理系统、自动化数据质量监控平台、数据分类分级与动态脱敏系统。

数据运营模块：企业级数据资产目录、数据服务门户系统、数据价值评估体系工具。

数据开发模块：标准化特征工程平台、自助式数据分析沙箱、机器学习算法集成环境。

（2）基础条件

企业需已完成主要业务系统的数据标准化工作，拥有专业数据团队（3-5 人）负责平台建设和运营，生产网络与办公网络已实现安全隔离，确保具备用于数据中台建设和持续优化的预算。

（3）部署方法

采用混合云架构部署企业级数据中台。构建统一数据湖和实时数据集成管道；分阶段实施各业务域数据资产建设，同步建立数据治理体系和运营机制；通过 API 服务化方式逐步开放数据能力，建成支撑企业级数据应用的服务平台。

3.进阶级

企业已建成成熟的数据平台，数据治理体系完善，数据驱动决策的文化初步形成，而数据驱动创新相对不足。关键在于将数据中台演进为“智能中台”或“企业大脑”，深度融合知识图谱、实时决策引擎、仿真优化等能力，形成可支持自主决策和业务创新的智能中枢。

（1）核心组件

数据集成模块：智能数据发现与识别平台、联邦数据集成系统、毫秒级实时数据流处理引擎。

数据治理模块：知识图谱驱动的数据关联系统、隐私计算平台、自动化合规审计与风险预警系统。

数据运营模块：数据产品全生命周期管理平台、数据服务交易市场、AI 驱动的数据智能运营分析系统。

数据开发模块：自动机器学习平台、低代码 AI 开发环境、多模态数据融合分析系统、智能标注模块。

（2）基础条件

企业需建立完善的数据治理组织和制度，拥有专业的数据科学家和 AI 算法团队，具备数据中心基础设施，与产业链合作伙伴建立数据共享机制，具有用于数据智能创新的年度预算。

（3）部署方法

建设智能数据中台作为企业数字化转型的核心引擎。采用数据编织架构实现多源数据的智能发现和集成；构建数据

产品工厂机制，将数据能力产品化、市场化；部署隐私计算平台支持生态数据协作，建成行业领先的数据智能创新平台。

附件一、数据治理典型案例

（一）单场景

1.案例名称：AI+AR 赋能智慧巡检与预测性维护的数据治理实践案例

案例企业：灯塔工厂

该企业是家用电器制造业企业，主营业务包括研发、生产、销售电风扇、空调、冰箱、洗衣机、厨电和小家电等全品类家电产品以及全屋智能解决方案。

背景描述：

设备稳定运行是保障生产线高效运转的核心前提。当前，大型制造企业设备故障处置工作多依赖技术人员的专业经验，而受设备结构复杂、精度要求严苛等因素制约，传统检修模式的局限性日益凸显，不仅需投入大量人力、物力与时间成本，还易因人为操作偏差引发二次问题。

治理方案：

（1）数据治理场景识别

聚焦设备运行监控与维护场景，推动设备运维模式从“事后补救”向“事前防范”转型，实现停机风险提前识别、前置处置。系统沉淀设备健康模型、告警规则及运维经验，形成可复用标准化业务规则，深度融入智能巡检、运维管理及预测性维护全流程，涵盖产线与设备仿真建模、资产健康评估、告警与工单自动联动、运维资源智能调度、预测性维护实施、基于故障知识图谱的根因定位、

AR 与 AI 融合运维辅助等方向，最终达成从单台设备风险预警到产线级运维协同的规模化升级。

（2）解决的基本路径

①强化组织建设。建立数据治理委员会，组建由业务、IT、数据治理、设备工程、生产、质量等部门构成的跨部门矩阵组织，统筹推进数据治理工作。明确数据所有者（数据 Owner），为设备数据、维修数据等关键数据域，由业务部门专家担任数据 Owner，负责数据标准定义、数据质量审核及业务语义分歧化解与规范统一。

②完善制度体系。以《企业数据治理总纲》为顶层设计，明确数据权责归口，衍生《设备主数据管理规范》《传感器数据质量标准》等配套细则，形成与业务流程深度嵌套的动态操作指南。通过明确数据 Owner 职责、规范数据采集上传、界定数据质量问题上报整改流程、建立数据标准变更评审机制，将治理要求转化为岗位具体行动。构建“制定规范-执行落地-监督考核-优化改进”闭环，内置度量反馈机制，定期评审主数据成熟度、数据质量分数等治理成效，驱动制度迭代优化，并将可显性化、规则化治理内容封装至平台，实现数据治理自动化、智能化。

③数据资产盘点。全面开展数据资产普查，精准识别预测性维护相关数据源、数据资产及对应责任人。构建企业数据资产目录，集成元数据管理、数据血缘分析及数据

价值评估功能，便利业务人员与数据科学家快速发现、理解和运用数据。

④推进治理活动落地。一是设备主数据与元数据管理，设备主数据与元数据管理。建立唯一可信设备资产库，为压缩机电机、主轴等关键设备赋予唯一编码，统一型号、规格、供应商、安装位置等核心属性，构建规范设备物模型[73]。**二是多源异构数据集成与多模态融合与对齐，**深度关联、对齐时序、事务、文档等多模态数据并完成语义融合，构建面向“设备健康管理”主题的、上下文丰富的宽表或数据图，打造设备全生命周期“数字孪生体”及运行物理规则“规约机理库”，形成 AI 模型可直接调用的高精度、高质量数据构件。**三是数据质量管理，**建立数据质量规则库、多维度数据成熟度评估模型及数据集检测评估平台，持续监控度量数据完整性、准确性、一致性、时效性。针对场景需求，基于数据质量规则库，对数据成熟度问题实施自动化清洗、插补及异常检测，保障 AI 模型输入数据洁净可靠；对治理后数据进行数据集封装及检测评估认证，支撑 AI 模型在跨基地、跨产线、跨设备间冷启动与快速适配推广，提升模型泛化能力与应用迁移效率。**四是数据标准与规范制定，**统一数据字典、结构化设备物模型及语义网络，配套嵌入工作流的治理工具、专业成熟度评估模型及评测工具，建立覆盖设备资产、故障代码、维修操作、数据集认证全生命周期的标准体系，构建

数据生态“通用语言”。**五是非结构化知识数据治理，**采集、识别、分类、打标设备手册、图纸、维修视频、专家经验等非结构化数据，与结构化主数据、设备孪生模型关联映射及时空对齐，构建企业级设备故障知识图谱与统一语义层。依托云边端架构，赋能 AI 模型加持具身终端，实现感知与智能适配、任务生成、日程提醒、预测性维护、根因分析及知识自动化沉淀，形成任务与数据闭环。**六是数据安全与权限管控，**依据核心工艺参数、设备实时运行数据等数据敏感度及业务影响程度，划分公开、内部、秘密、核心等级别密级，实施差异化加密、脱敏、存储及流转策略。升级权限管控机制，从角色访问控制演进为动态上下文感知的属性基访问控制，结合用户角色、设备位置、访问时间等多维属性实现细粒度授权，配套数据水印与全链路操作日志审计，确保数据访问可追溯、可预警。在坚守“最小必要原则”前提下，打破部门壁垒，实现数据安全可控共享，筑牢 AI 应用信任数据基座。

⑤健全治理机制。建立每周数据治理例会制度，常态化跟踪任务进展、协调解决跨部门协作问题。构建“定义标准-质量监控-发现问题-分析根因-整改优化-度量效果”闭环管理机制，打造自学习、自完善的自治系统，推动治理成果固化为组织制度基因。

⑥强化价值度量。兼顾数据治理技术指标（数据质量分数、数据资产数量等）与业务价值度量，将治理成效与

预测性维护项目 KPI 强绑定，重点考核设备综合效率提升、维修成本降低、非计划停机时间减少等核心业务指标。

实施成效：

通过面向 AI 的数据治理实践，实现数据对 AI 应用的适配就绪，为 AI+AR 智慧巡检与预测性维护落地筑牢坚实数据根基。系统上线运行后，在经济效益、管理效能等方面成效显著。

运维成本显著降低。一是非计划停机时间减少 25%。依托 AI 模型提前数天至数周预判设备潜在故障，支持维护团队利用生产间隙开展计划性维修，有效规避突发故障导致的生产线停摆，经测算，每年减少非计划停机损失超 800 万元。二是维修成本降低 30%。推动运维模式从“定期更换、可能过度维护”的预防性维护，转向“按需维护”的预测性维护，减少不必要的备件消耗与人工投入，同时通过精准预判规避小故障扩大化，延长设备使用寿命。三是备件库存成本优化 20%。基于故障模式精准预测，科学规划备件采购与库存水平，有效降低资金占用成本。

生产效率与产量稳步提升。设备综合效率提升 5 个百分点，得益于停机时间大幅压缩、设备稳定性持续增强，生产线可用率及性能开动率显著提高，直接拉动产能增长，在市场需求旺盛阶段，工厂月均产量提升约 3.5%。

安全与质量管控效能提升。一是设备相关质量事故下降 60%。通过预测性维护有效规避温控系统失灵等设备故障，杜绝由此引发的产品批次性质量缺陷，提升产品一次合格率。二是维修安全事故实现清零。借助 AR 智能眼镜为维修人员提供直观步骤化安全操作指引及远程专家支持，彻底杜绝误操作引发的二次故障及各类维修安全事故。

组织能力与员工素养全面赋能。一是平均维修时间缩短 40%。维修人员通过 AR 眼镜快速定位故障点、查看三维拆装动画，并与远程专家实时协同，大幅提升维修效率，有效弥补新手工程师经验短板。二是实现知识沉淀与传承。将专家维修经验数字化并融入 AR 指导系统，推动隐性知识显性化、可持续传承，整体提升维修团队技术水平。

（二）多场景

1.案例名称：医药包装企业数智化转型的数据治理实践案例

案例企业：江苏省外制造业企业，暂未参评智能工厂

该企业是药品制造产业链核心配套企业，专注于药品折叠纸盒、药用铝管等包装产品研发生产，业务覆盖全球医药包装市场。

背景描述：

该企业在推进数字化转型与智能化升级过程中，面临一系列数据治理层面的突出瓶颈，严重制约 AI 技术与生产经营的深度融合：生产、仓储、质量、物流等核心业务系统各

自独立，设备、工艺、订单等关键数据分散存储形成孤岛，跨环节数据流通壁垒显著；不同生产线、异构设备的数据格式缺乏统一规范，印刷、覆膜、赋码等工序的工艺参数统计口径存在差异，进一步抬高了全流程数据整合的门槛；受设备运行状态、现场环境波动等因素影响，部分关键工艺数据存在缺失、异常等问题，直接降低了 AI 质量预测、智能生产调度等模型的精度与可靠性；与此同时，海量生产数据未经过系统化治理与深度挖掘，无法有效沉淀为可复用的工艺知识图谱与智能决策模型，数据资产的价值潜力未能释放，生产管控仍以人工经验驱动为主，智能化转型进程受阻。

治理方案：

（1）数据治理场景识别

围绕数据质量不足制约 AI 模型落地、生产管控依赖人工经验的痛点，企业以标准化治理后的高质量数据为核心支撑，精准对接生产优化、质量管控、智能物流等核心 AI 应用需求，提升 AI 质量预测、智能生产调度、物流路径优化等模型的精度与可靠性。通过数据与 AI 技术的深度融合，推动生产模式从传统“经验驱动”向科学“数据驱动”转型，实现生产效率与管控水平的双重提升。

（2）解决的基本路径

①制定数据治理总体架构。该企业构建了“1+2+4”人工智能数据治理总体架构，即 1 个数据中台、2 大治理维度、4 大核心环节，实现数据从采集到价值转化的全生命周期闭

环管理。**“1 个数据中台”。**依托数智工业平台，该企业搭建了专属数据中台，整合生产设备、业务系统、仓储物流等多源数据，提供数据存储、计算、调度等一体化支撑，兼容16 条全自动生产线的设备协议与业务系统接口，实现数据实时汇聚与统一管控。**“2 大治理维度”。**在业务层治理方面，聚焦印刷、覆膜、赋码、裁切等核心生产环节，以及原料入库、成品出库、物流调度等仓储环节，构建场景化数据治理体系；在数据层治理方面，针对设备数据、工艺数据、质量数据、订单数据等多类数据，建立分类分级治理机制，明确数据权责与流转规则。**“4 大核心环节”**覆盖数据采集接入、清洗转换、标注加工、安全应用全流程，每个环节嵌入智能工具与标准规范，确保数据治理高效落地。

②数据采集。在数据接入方面，通过边缘计算节点接入16 条智慧化全自动生产线设备、AGV 物流机器人、质量检测设备等各类数据源，支持 SMT 设备、印刷机等 50+种工业设备协议解析。**在数据采集传输方面，**采用 5G+工业互联网技术，实现生产工艺参数、设备运行状态等数据的实时采集与传输，采集延迟控制在毫秒级，保障数据时效性。**在业务系统贯通方面，**打通 ERP、MES、WMS 等核心业务系统数据接口，实现订单信息、生产计划、库存数据、质量检测结果等数据的自动化同步，打破系统间数据壁垒。

③数据清洗与转换。一是建立质量校验规则。针对医药包装生产特性，制定工艺参数阈值校验、生产逻辑校验、数

据完整性校验等 8 类质量规则，自动识别异常数据与缺失值。**二是应用智能清洗工具。**部署智能清洗模块，通过多模型算法自动完成数据去重、补全、格式标准化转换，数据清洗效率提升 40%以上。三是统一数据标准。制定医药包装行业数据元标准，将不同设备、不同系统的异构数据转换为统一规范格式，实现跨环节数据互联互通。

④数据标注与加工。一是建立标注规则。针对药品包装质量检测、生产工艺优化、物流路径规划等 AI 场景，建立专属标注规则，形成缺陷特征、工艺参数阈值、物流节点等标注数据集。**二是低代码标注工具应用。**提供工业级低代码标注平台，支持批量标注与人工修正结合，针对印刷瑕疵、尺寸偏差等质量问题，快速构建高质量训练数据集，标注效率提升 50%。**三是知识模型沉淀。**整合行业专家经验与生产实践数据，构建工艺知识图谱与质量预测模型，将隐性经验转化为结构化数据资产。

⑤数据安全与应用。一是分级分类管控。按数据敏感程度将订单信息、工艺配方、质量数据等划分为核心数据与一般数据，对核心数据实施加密存储与严格的访问权限管控。**二是全链路安全防护。**建立数据传输加密、存储加密、访问审计的全链路安全机制，实现数据操作全程追溯，保障医药包装数据合规性。**三是 AI 场景深度应用。**基于治理后的数据，支撑多项 AI 应用落地，包括质量缺陷智能检测、生产调度优化、物流效率提升等，实现数据价值最大化。

实施成效：

数据经过治理后，质量显著提升。工厂数据完整性从 72%提升至 98.5%，数据准确性达到 99.2%，数据标准统一率实现 100%，为 AI 应用提供了高质量数据支撑。

在质量管控方面，基于治理后的数据训练质量预测 AI 模型，药品包装缺陷检出率提升至 99%，不良率降低 30%，有效保障医药包装产品合规性。

在生产效率方面，通过生产调度 AI 模型优化工艺参数与生产流程，订单交付周期缩短 25%，人均产出提升 30%，实现规模化高效生产。

在物流成本方面，AGV 物流机器人基于治理后的物流数据优化路径规划，仓储物流效率提升 20%，库存管理成本降低 15%。

2.案例名称：智能制造中面向工业模型应用的数据治理实践案例

案例企业：2025 年卓越级智能工厂

该企业是钢铁生产企业，集采矿、炼铁、炼钢、轧钢及深加工于一体，主营产品广泛应用于基建（桥梁、高铁）、汽车、机械制造、能源装备等领域。

背景描述：

企业在已完成 ERP、MES 等核心系统部署，建成企业级数据库基础之后，通过打造“仓湖一体”数据中台，整合 100 余套系统数据，绝大程度上解决了企业“数据孤岛”问题，

形成覆盖铁钢轧全流程的高质量数据集。为推动企业数智化发展，企业进一步探索“数据驱动决策、算法优化生产”的智能化转型，构建数据工作整体框架。

治理方案：

（1）数据治理场景识别

在钢铁制造流程中，高炉是将铁矿石转化为铁水的关键设备，为后续炼钢提供主要原料，但高炉内部反应复杂、不可见，操作高度依赖老师傅经验，“黑箱化”问题严重，对燃料、产量和设备均容易造成不良影响。另外，传统质检存在漏检、误判风险，检测主观性强、效率低，影响产品交付质量。该企业紧扣钢铁制造核心流程中的经验依赖强、过程不可控、质量波动大的难题，探索基于智能感知网络的智能制造解决方案。

（2）解决的基本路径

①通过“1 套数据体系+3 项工作建设+3 个能力打造”，构建数据工作整体框架。“1 套数据体系”即全流程统一数据体系，编制了《数据治理沟通机制》《元数据管理办法》《主数据管理办法》《指标数据管理办法》等 30 余项数据管理制度。“3 项工作建设”包括搭建可视化分析平台支持数据驱动决策、建设数据中心打通数据传输链路、部署专用传感器与工业相机等设备采集生产数据。“3 个能力打造”包括通过数据感知实现全流程数据实时捕获、通过三大平台监

控数据完整性与准确性、通过公司级零信任架构保障数据安全。

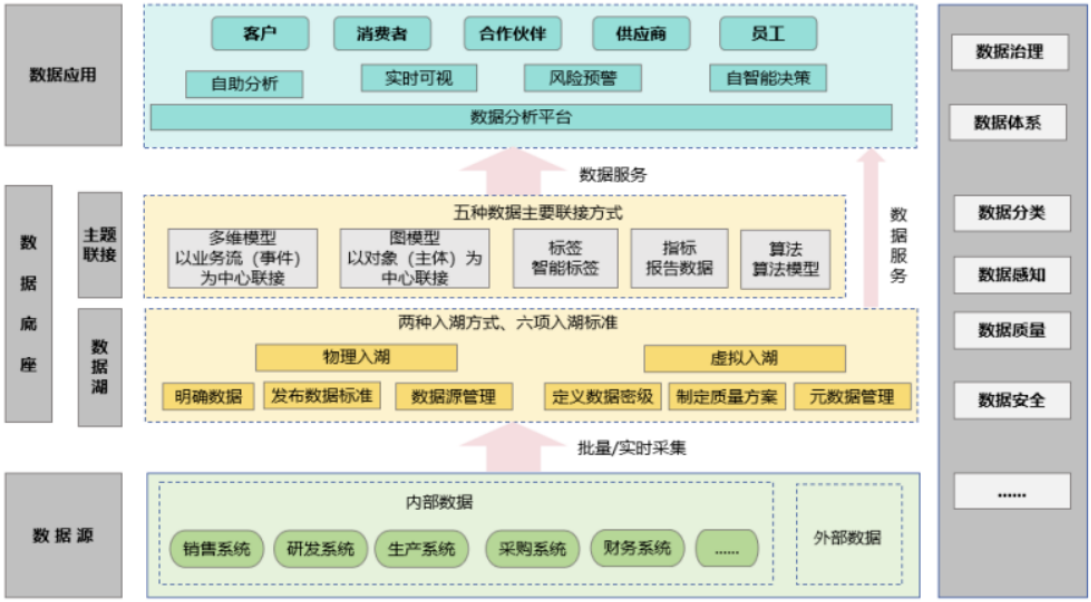

图 5　数据工作整体框架

②构建“组织-制度-技术-文化”四位一体的保障体系，确保治理工作“有人管、有章循、有技撑、有共识”。在组织架构方面，成立由集团董事长领导的数字化委员会进行顶层设计，设立配备 CDO 和 8 名专职工程师的数据治理办公室负责统筹协调，引入 DCMM 评估体系推动管理转向“标准驱动”；在各分厂设立“数据专员”，形成“总部统筹+分厂执行”的二级管理体系；通过“数据治理联席会”机制推动跨部门协作，解决数据壁垒问题。**在制度体系方面**，发布的 30 余项标准体系和管理制度覆盖数据全生命周期，确保数据统一管理。如《主数据管理办法》确保源头“一数一源、动态更新”；《数据质量考核细则》将数据准确率、完整率纳入部门 KPI 考核体系，挂钩绩效；《数据分类分级指南》

和《工业数据安全防护规范》则对敏感数据（如工艺参数、客户信息）实施严格管控，要求“双人审批”。**在技术平台方面**，部署主数据管理平台实现核心数据的全生命周期管理，年处理异常数据超100万条；利用数据质量管理平台自动监控千余项关键指标并实时预警，响应时间缩短至1小时；搭建可视化分析平台支持数据驱动决策，让数据直接服务于降本增效。**在数据文化方面**，实施业务技术复合型骨干培养计划，举办数据大赛，通过各类渠道广泛传播数据驱动价值，使“用数据说话、用数据决策”成为全员共识。

③建设支撑大模型应用的数据工程。在数据采集方面，拓宽采集范围并保障质量，部署专用传感器、工业相机、边缘计算网关等设备，部署几十万余个实时采集点，实现生产数据毫秒级采集。同时建立数据校验机制，确保源头数据精准可靠。在算力支撑方面，搭建“私有云+公有云”混合架构，支撑数万级模型并行计算。

实施成效：

场景落地方面，开发及集成工业模型500余个，覆盖近20个业务领域、100余个场景，形成“数据采集-模型计算-决策执行”闭环，全面优化生产效率、产品质量和能源消耗等关键指标。

在解决高炉“黑箱化”问题方面，企业通过在新建高炉本体部署的高精度传感器，实时采集温度、压力、煤气流、料面分布等数百项参数，并利用AI模型对炉内状态进行精

准感知与预测，通过“工业大脑”指挥中枢展示可视化看板和优化建议，实现炉况透明化、调控智能化。

在解决质量检测主观性问题方面，企业通过在车间的 AI 质检区域部署的工业相机天网系统，叠加由 500 万张缺陷图片训练的专用质检模型，实现每秒 300 帧的速度扫描钢材表面，精准识别各类 0.3 毫米的划痕与针尖大小的气泡，拦截缺陷品近 4 万吨，实现海外客户零理赔的优异表现。

3.案例名称：特钢企业多场景的数据治理实践案例

案例企业：2025 年卓越级智能工厂

该企业是特钢生产企业，主营高端轴承钢、高档汽车钢等产品，广泛应用于交通、石化、航天等领域。

背景描述：

随着企业规模扩大、产品结构向高端化升级，叠加钢铁行业生产流程长、工艺复杂、供应链冗长的固有特性，传统生产运营模式已难以适配成本控制、品质提升、高效协同的发展需求，企业亟需推动 AI 与研发、生产、质量等环节的深度融合，挖掘数据要素价值，增强企业竞争优势。

治理方案：

（1）数据治理场景识别

该企业基于特钢生产全流程，形成覆盖生产过程、能耗分析、质量分析、智能排产、物流管理、设备监测、安全管控、环保监测等多方面智能管控场景图谱。重点围绕制造工艺定制设计、工艺智能控制、高精度智能轧制、质量实时监

测等多个高价值 AI 应用场景进行数据治理，实现特钢生产核心环节的 AI 深度赋能覆盖。

（2）解决的基本路径

该企业制定了工业数据治理实践指南，以工业互联网平台和大数据平台为基础，构建统一数据架构数据治理体系，打通企业研发、生产、能源、物流、环保、安全、设备、质量、运营全流程数据，以高质量数据支撑企业生产运营和战略决策。该企业数据治理建设分为四个阶段，第一阶段完成大数据平台的基础构建工作，将企业数据平台化，打造数据引擎，构建数据底座；第二阶段逐步完善平台能力，构建治理和管理体系；第三阶段全面展开，深化应用，实现 AI 赋能；第四阶段实现深层次挖掘和利用，洞察市场，支撑企业战略发展。具体建设路径如下：

①制定数智化战略。确定了三大战略目标、3 大战略方针、5 大战略举措，通过四线并行、以点带面、先易后难的实施路线支撑企业构建 6 大核心竞争力。

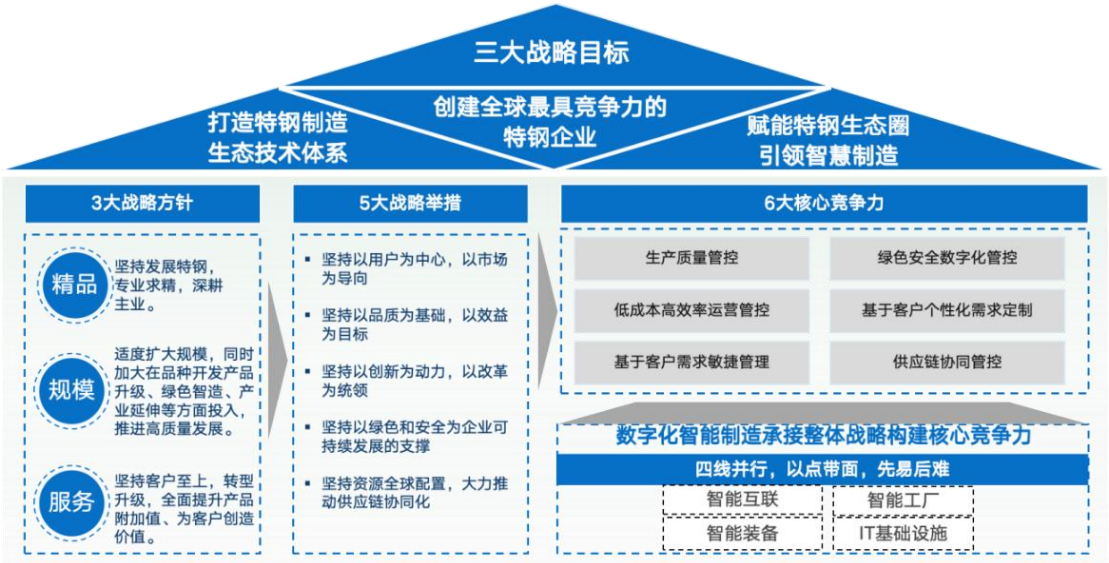

图 6　数智化战略

②设计工业互联网数据架构。建立统一数据架构，依次进行生产制造基础数据采集、边缘存储和计算，再构建统一的数据平台和应用模型，支撑业务应用分析，打通企业业务数据流程，形成产销协同、质量研发、制造执行、能碳安环、企业管理和数据互联等领域应用，最终支撑企业战略决策。

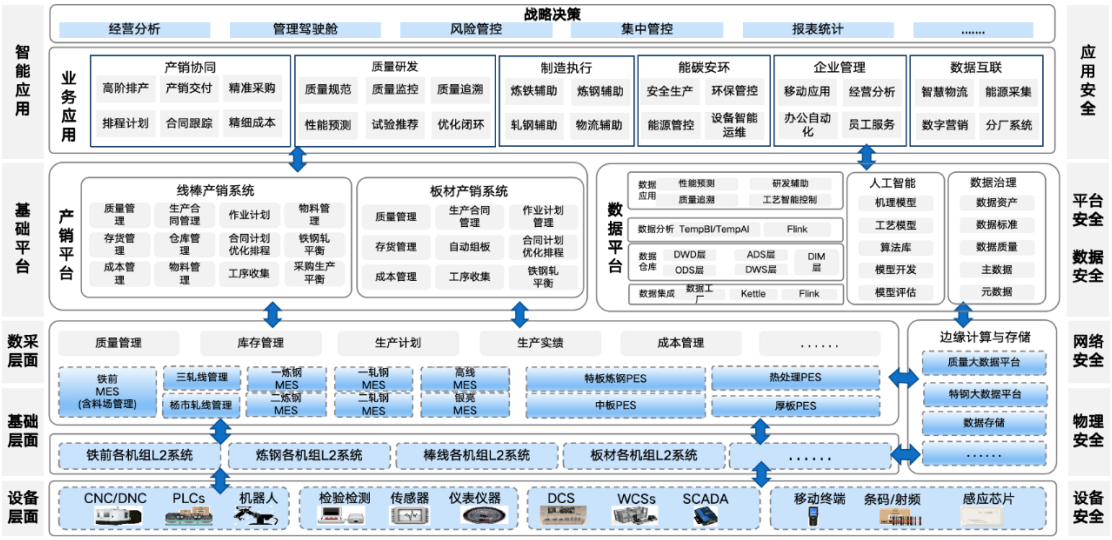

图 7　工业数据架构图

③打造数据治理体系。构建数据治理体系支撑企业的数据标准化、数字化，打通研发、制造、运营和维护数据断点，为企业数据应用夯实基础。该企业采取多种关键举措构建工业数据治理体系，**一是建立了企业级核心大脑—全流程数智管控中心，**包括四大中心：铁前操控中心、炼轧操控中心、设备诊断中心、大数据 IDC 中心。实现了集生产过程、能耗分析、质量分析、智能排产、物流管理、设备监测、安全管控、环保监测等多方面的系统融合，是公司生产运营的核心大脑；**二是全面盘点企业数据资产，**设计覆盖企业全领域的

数据资产五级分类目录，其中 L1 级资产 19 项，L2 级资产 106 项；经过自下而上地梳理企业系统应用情况、关联关系、业务流程及数据字典等，形成数据资产目录；**三是建立企业大数据 IDC 中心和私有云平台，**支持信息化系统内存储、数据中心存储和私有云存储，为企业信息化搭建高效、安全、可扩展的存储体系；**四是重塑组织架构，建立高效数字化管理组织。**成立由总经理部领导的数字化转型领导团队，设立智能制造中心，制定实施路线图，并建立跨部门的协调机制。设立敏捷工作组，实施数字化技术和业务流程优化。在数据体系建设过程中，任命企业数据 OWNER 和数据管家，通过“传帮带”模式，培养数据人才。以工业互联网平台为基础，整合内部数字化能力，与清华、华为、麦肯锡、京诚数科等知名高校企业合作，形成生态合作圈。**五是建立一套完善的数字化制度保障体系，**包括数据管理和信息化管理文件，确保数字化转型科学、规范、有序推进。通过制定规章制度、数据标准和开发规范支撑数据建设，有效指导企业数据建设落地实施，整合数据资源，支持跨部门的业务协同。

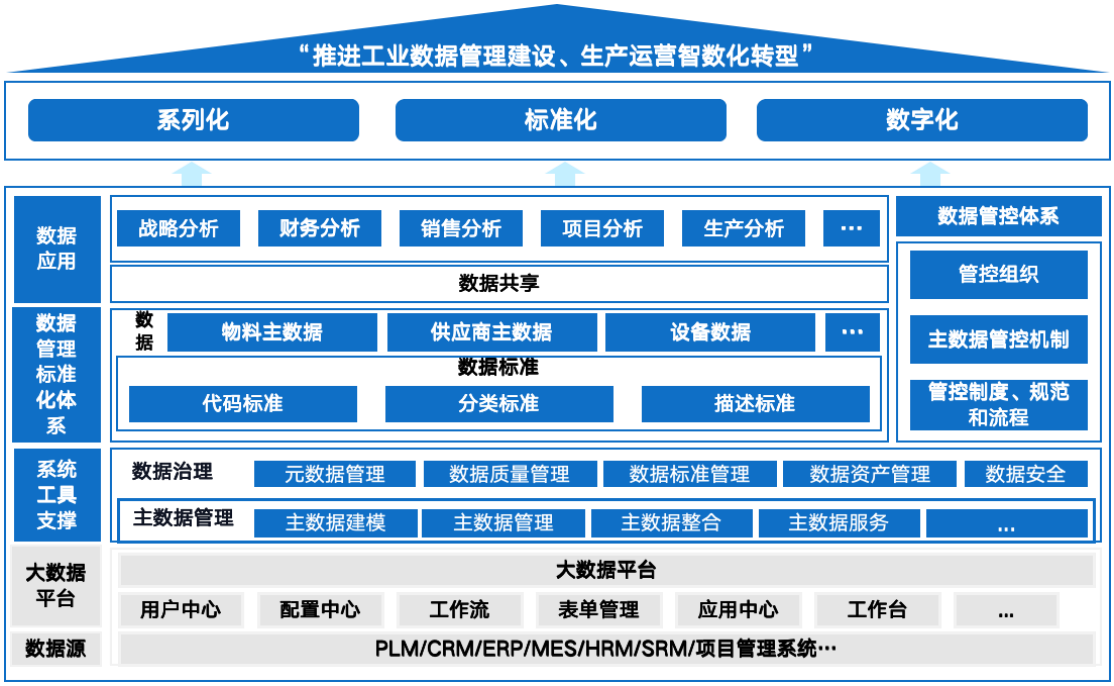

图 8　数据治理体系图

实施成效：

企业以数据要素和人工智能为突破口，在纵向工序维度，贯通铁、钢、轧生产全流程，打造智能化示范产线；在横向业务维度，覆盖研发、质量、能源等全领域，布局数字化系统应用。通过工业数据治理体系的构建，帮助企业实现效率和效益的双重提升，最终实现产量提升 14.4%，定制化订单数量提升 35.3%，交货周期缩短 20.3%，质量损失节约 46.7%，吨钢能耗降低 10.5%。

（三）企业全域

1.案例名称：钢铁行业企业生产及管理全域数据治理的实践案例

案例企业：2025 年领航级智能工厂

该企业所属钢铁行业，主营业务聚焦于高强度、高韧性、高疲劳、耐候、耐磨、耐腐蚀等高附加值优质特殊钢研发生产，已形成中厚板、棒材、高速线材、带钢、异型钢、复合板等产品系列。

背景描述：

该企业基于钢铁生产全域、全流程，围绕研发设计、生产制造、辅助决策、智能运维、运营管理等关键环节，逐步形成包含铁前、炼铁、炼钢、轧钢等全流程生产场景以及安全、园区、市场采购、经营调度等管理场景的全景图，实现钢铁流程高价值场景全覆盖。

治理方案：

（1）数据治理场景识别

该企业基于“突出主业、体现价值、知识密集、语料丰富、技术可行”五个原则，系统分析战略决策、业务运营各领域以大模型为核心的人工智能应用机会，制定“价值、成本、风险、数据、技术”5个维度、16个指标的成熟度评估框架，量化评估场景应用成熟度，逐步形成上述场景全景图。

（2）解决的基本路径

①统一数据管理组织。由董事长亲自挂帅，建立数字化管理委员会，并在委员会之下设立数字应用研究院与人工智能研究院，负责数字化转型顶层设计、数字人才变革与选拔、数据治理与数据资产化、人工智能等数字技术应用场景孵化

落地工作，建设以业务主管为主导的数据 Owner 体系，承载公司数据战略落地。

②统一数据管理规范。发布企业级数据管理制度，构建数据管理质量体系，形成 1 个总纲、3 个政策，8 大类 28 个管理规范、35 个数据管理流程。

③统一数据平台底座。公司依托“云-边-端”架构建立统一的数据平台底座。在端侧，基于先进装备集成 26 条产线上 PLC、SCADA 系统、传感器、智能仪表等，实时采集温度、压力、能耗、设备状态等数据。在边缘，统一规划工业互联网平台，实现工业现场海量异构数据的高效采集、安全可靠的分布式存储以及实时与离线计算能力。在云端，构建实时湖仓，实现结构化数据统一汇集和存储，为企业在数据清洗、数据开发、数据服务等提供一站式技术支撑。该数据平台底座高效支撑了智慧运营中心、铁区一体化中心、钢轧一体化中心等公司级应用建设。

实施成效：

该企业通过在数字化转型中开展的数据治理与人工智能应用工作，沉淀了 10000+数据资源表、2500 余业务指标、百万 OT 点数据资产，构建了安全可控的数据汇聚、治理、多模态融合能力与 AI 高质量数据集，高效支撑自助数据分析与 AI 场景构建。与开发的 3000 余模型配合形成数据与智能飞轮效应，以数据与智能驱动重塑钢铁生产运营模式。

在研发环节，采用机器学习对影响钢材性能的过程工艺数据进行智能分析，采用 AI 模型对近 70 个钢种的力学性能进行在线预测，准确率达 95%以上，通过提前预测产品性能，及时进行工艺优化，提升产品质量。在制造环节，针对传统配矿“业务不闭环、约束不全面、求解效率低、降本不极致”等行业难题，基于“工艺机理-性能预测-反馈优化”架构开发一体化智能配矿模型，研发“跨时域非线性梯度自适应寻优”算法，考虑工艺约束 30+项，10s 内求解，与传统模式相比，吨钢降本 4~6 元。行业首创智能金相检测系统，具备缺陷样本数据生成与图像数据自动标注能力，结合对模型精度和泛化能力的提升，实现非金属夹杂物、晶粒度、脱碳层等识别评级，准确率≥90%。在营销环节，精准分析客户需求，快速生成专业营销方案，高效响应客户，创建智能化仓储结算，缩短订单签订时间 40%。

同时，公司以“数据+模型”为载体高质量推进数据资产入表，2024 年成为全国首批数据资产入表上市公司之一，截止 2025 年底累计入表超 3300 万元，实现数据产品和数据资产登记，并探索工业数据资产交易实现路径。

附件二、数据治理服务商清单

序号	平台型服务商	核心覆盖的数据治理环节
1	树根互联股份有限公司	数据采集、数据集成、数据存储、数据计算、数据预处理、数据集质量评估、数据安全保护等
2	宝信软件股份有限公司	数据集成、数据集（高质量语料库）、数据计算（智算平台）等
3	徐工汉云技术股份有限公司	数据采集、数据集成、数据计算等
4	东方国信科技股份有限公司	数据采集、数据集成、数据预处理、数据安全等
5	用友（上海）工业互联网科技发展有限公司	数据集成、数据预处理、数据安全等
6	中兴通讯股份有限公司	数据采集、数据集成、数据存储、数据计算等
7	星环信息科技（上海）股份有限公司	数据集成、数据存储、数据预处理等
8	朗坤智慧科技股份有限公司	数据集成、数据预处理
9	苏州慧工云信息科技有限公司	数据采集、数据集成、数据预处理、数据集质量评估等
10	南京凯奥思数据技术有限公司	数据采集、数据集成、数据预处理、特征工程、数据集质量评估等
11	苏州真趣信息科技有限公司	数据采集、数据集成、数据预处理、数据集质量评估等
12	苏州协同创新智能制造装备有限公司	数据采集、数据集成、数据预处理等
13	南京钢铁股份有限公司	数据采集、数据预处理、数据集质量评估、数据安全保护等
14	南京科远智慧科技集团股份有限公司	数据采集、数据集成、数据预处理、数据集质量评估等
15	无锡华光环保能源集团股份有限公司	数据采集、数据预处理、数据集质量评估等
16	江苏沙钢高科信息技术有限公司	数据采集、数据集成、数据预处理、数据集质量评估、数据安全保护等
17	苏州汇川技术有限公司	数据采集、数据集成、数据预处理、数据集质量评估、数据安全保护等

18	南京南瑞信息通信科技有限公司	数据采集、数据存储、数据计算、数据集成、数据预处理、数据集质量评估、数据安全保护等
19	江苏亨通数字智能科技有限公司	数据采集、数据集成、数据预处理、数据安全保护等
序号	**AI 算法服务商**	**核心覆盖的数据治理环节**
20	北京阿丘科技有限公司	数据增强、数据标注
21	常州微亿智造科技有限公司	数据预处理、特征工程、数据标注/增强/划分等
22	思必驰科技股份有限公司	数据预处理、特征工程、数据标注/增强/划分等
23	云知声智能科技股份有限公司	数据预处理、特征工程、数据标注/增强/划分等
24	云从科技集团股份有限公司	数据预处理、特征工程、数据标注/增强/划分等
25	北京旷视科技有限公司	数据预处理、特征工程、数据标注/增强/划分等
26	北京地平线信息技术有限公司	数据预处理、特征工程、数据标注/增强/划分等
27	苏州天准科技股份有限公司	数据预处理、特征工程、数据标注/增强/划分等
28	达智汇（苏州）科技有限公司	数据预处理、特征工程、数据标注/增强/划分等
序号	**基础设施服务商**	**核心覆盖的数据治理环节**
29	中兴通讯股份有限公司	数据采集、数据存储、数据计算、数据集成、数据预处理、数据安全保护等
30	中移（苏州）软件技术有限公司	数据采集、数据存储、数据计算、数据集成、数据预处理、数据安全保护等
31	江苏敏捷科技股份有限公司	数据安全保护
32	启明星辰信息技术集团股份有限公司	数据安全保护
33	南京安秉信息安全有限公司	数据安全保护
34	南京众智维信息科技有限公司	数据安全保护
35	江苏通付盾科技有限公司	数据安全保护
36	华云数据控股集团有限公司	数据存储、数据计算、数据安全保护等

37	江苏恒云太信息科技有限公司	数据存储、数据计算、数据安全保护等
38	新华三技术有限公司	数据采集、数据存储、数据计算、数据集成、数据安全保护等
39	浪潮卓数大数据产业发展有限公司	数据采集、数据存储、数据计算、数据集成、数据预处理、数据集质量评估等
40	苏州中飞遥感技术服务有限公司	数据采集、数据预处理、数据集质量评估等

附件三、专业名词介绍

序号	概念	缩写	核心含义
1	多模态数据	Multimodal Data	来自不同来源或形式的数据，如图像、文本、时序信号、音频等。
2	非结构化数据	UD(Unstructured Data)	难以用常规的行列形式表示，如文本类、多媒体类、网页内容、日志文件等类型数据。
3	元数据	Metadata	少数类别样本极多、多数类别样本极少的现象。
4	特征工程	FE(Feature Engineering)	从原始数据中提取、优化特征的过程，旨在提升机器学习模型的性能和可解释性。
5	监督学习	SL(Supervised Learning)	
6	半监督学习	SSL(Semi-Supervised Learning)	利用少量标注数据与大量未标注数据联合训练模型的学习范式。
7	长尾分布	Long-tailed Distribution	主要利用一组已知类别的样本来训练模型，使模型能够预测新样本的输出。
8	人机协同标注	/	在数据标注过程中，人工与算法相互协作共同完成高质量标注任务的工作范式。
9	主数据	Master Data	描述、标识、组织、管理及检索其他数据资源的数据，其本质是“关于数据的数据”。
10	数据血缘关系	DL(Data Lineage)	满足跨部门业务协同需要的、反映核心业务实体状态属性的组织机构的基础信息。
11	数据字典	DD(Data Dictionary)	对数据的数据项、数据结构、数据流、数据存储、处理逻辑等进行定义和描述的说明。
12	标识解析体系	IRS(Identifier Resolution System)	给物理/数字对象赋全球唯一“身份证”（标识编码），并通过解析系统实现跨域信息定位与共享，支撑全链路协同、溯源等应用。

序号	概念	缩写	核心含义
13	自动化标注	/	利用算法、模型或规则系统，无需或仅需极少人工干预，对原始数据进行标注的工作范式。
14	边缘网关	EG(Edge Gateway)	部署于网络边缘侧的通信设备，属于物联网系统的中间层设备，主要负责终端设备与云端服务器的连接。
15	边缘节点	EN(Edge Node)	边缘计算的核心逻辑实体，通过抽象整合网关、控制器、服务器等设备的基础功能，形成具有实时数据处理、本地存储与智能决策能力的通用技术载体。
16	数据仓库	DW(Data Warehouse)	企业级信息系统中的核心基础设施，其根本目的是集成、存储和管理来自多个异构数据源的历史性、结构化数据，以支持高效地分析、报表与决策支持。
17	冷数据	Cold Data	对于离线类不经常访问的数据。
18	数据湖	Data Lake	集中存储原始数据的架构，支持任意规模的结构化、半结构化和非结构化数据。
19	OT 数据	Operational Technology Data	由运营技术系统（如 PLC、DCS 等）产生、采集和使用的数据。
20	IT 数据	Information Technology Data	由信息技术系统（如 ERP、CRM、办公系统等）在支撑企业业务运营、管理决策与用户交互过程中产生、处理和存储的结构化或半结构化数据。
21	特征资产化	/	将特征作为可管理、可复用、可计量的企业资产进行运营，超越传统“模型附属物”定位。
22	TensorFlow	/	Google 开发的机器学习框架，具有灵活性、可移植性等特点。
23	PyTorch	/	Facebook 开发的开源的深度学习框架，支持动态计算图和自动微分功能。

序号	概念	缩写	核心含义
24	API（应用程序编程接口）	API(Application ProgrammingInterface)	为操作系统或者框架提供的接口。
25	智能体	Agent	能感知环境、自主决策、执行任务的 AI 软件实体。
26	知识图谱	KnowledgeGraph	一种结构化的语义知识库，以图模型（节点和边）描述现实世界中的实体（人、地点、概念等）、属性及其关系，形成网状知识结构。其核心目标是为机器提供可理解、可推理的知识表。
27	数字化交付	/	在工程建设项目中，同步交付物理设施与结构化、标准化的全生命周期数字资产。
28	数字孪生	DT(Digital Twin)	通过实时数据驱动，在虚拟空间中构建与物理实体或系统全生命周期同步、高保真、可交互、可仿真的动态数字镜像。
29	BIM（建筑信息模型）	BIM(Building Information Modeling)	用于在建筑、工程与施工全生命周期中创建、管理、共享和应用富含语义信息的三维数字模型。
30	图神经网络	GNN(Graph Neural Network)	一类专门用于处理图结构数据（节点与边）的深度学习模型，通过消息传递机制聚合邻居信息，学习节点、边或整图的嵌入表示。
31	强化学习	RL(Reinforcement Learning)	一种机器学习范式，智能体通过与环境交互试错，根据获得的奖励信号学习最优策略，以最大化长期累积回报。
32	帕累托最优	Pareto Optimal	指在多目标优化中，无法在不损害某一目标的前提下改进另一目标的状态。
33	生成式对抗网络	GAN(Generative AdversarialNetwork)	生成器与判别器对抗训练生成数据。
34	扩散模型	DM(Diffusion Model)	基于概率生成机制的深度学习模型。

序号	概念	缩写	核心含义
35	深度学习	DL(Deep Learning)	一类基于多层非线性神经网络的机器学习方法，通过端到端方式从原始数据中自动学习层次化特征表示。
36	Transformer 架构	/	一种基于自注意力机制（Self-Attention）的深度神经网络架构。
37	网络拓扑	Network Topology	描述计算机网络中节点（如设备、服务器）与通信链路之间连接方式和逻辑/物理布局的结构形式，常见类型包括星型、环型、总线型、网状等。
38	流处理引擎	SPE(Stream Processing Engine)	用于实时接收、处理和分析连续高速数据流（如日志、传感器数据、交易记录）的计算系统，支持低延迟响应与窗口聚合，典型代表包括 Apache Flink、Apache Kafka Streams、Spark Streaming。
39	主动学习	AL(Active Learning)	一种机器学习范式，模型在训练过程中主动选择最具信息量或不确定性的未标注样本，交由人工标注后加入训练集，以最小标注成本实现性能最大化。
40	平行系统	Parallel System	物理系统与高保真虚拟系统同步运行，用于策略预演、压力测试与持续进化。
41	传感器漂移	Sensor Drift	感器输出随时间缓慢偏离真实值的现象。
42	自动机器学习	AutoML(Automated Machine Learning)	利用自动化方法完成机器学习流程中的特征工程、模型选择、超参数调优、神经网络架构搜索等环节，降低 AI 应用门槛。
43	卷积神经网络	CNN(Convolutional Neural Network)	专为处理网格结构数据（如图像、视频）设计的深度神经网络，通过卷积层提取局部空间特征、池化层降低维度，具有参数共享和平移不变性优势。

序号	概念	缩写	核心含义
44	无量纲化处理	Dimensionless Processing	对原始数据进行尺度变换，消除量纲和数量级差异，使不同特征具有可比性。
45	VAE 架构	VAE(Variational Autoencoder)	一种生成式深度学习模型，通过引入概率隐变量和变分推断，将输入数据编码为潜在空间的概率分布，并从中采样解码重建数据，用于生成、去噪和表示学习。
46	神经算子	Neural Operator	用于学习函数到函数映射的深度学习模型，可高效求解复杂物理仿真。
47	高性能计算	HPC(High-Performance Computing)	利用超级计算机或计算集群，通过并行处理大规模数值计算任务，以解决科学、工程和人工智能等领域中对算力要求极高的复杂问题。
48	本体论	Ontology	在信息科学中指对领域概念及其关系的形式化定义，是构建知识图谱的基础。
49	流数据	Streaming Data	持续、高速、无界生成的数据序列，通常按事件驱动方式实时处理，强调低延迟与状态管理。
50	流批一体数据处理	Unified Stream-Batch Processing	一种数据处理架构，使用同一套引擎和 API 同时支持低延迟的流式数据处理与高吞吐的批量数据处理，实现逻辑一致、运维简化。
51	AGV（自动引导运输车）	AGV(Automated Guided Vehicle)	一种无需人工驾驶、通过电磁、激光、视觉或 SLAM 等导航技术在工厂、仓库等环境中自主移动并完成物料搬运任务的智能移动机器人。
52	合成少数类过采样技术	SMOTE(Synthetic Minority Over-sampling Technique)	一种针对类别不平衡问题的数据增强方法。

序号	概念	缩写	核心含义
53	ARM	Advanced RISC Machine	一种基于精简指令集（RISC）的低功耗处理器架构，广泛应用于移动设备、嵌入式系统及边缘计算节点。
54	物理仿真	Physical Simulation	基于物理定律构建数学模型，在虚拟环境中模拟真实世界物体或系统的动态行为。
55	模仿学习	IL(Imitation Learning)	一种从专家示范中学习策略的机器学习方法，通过行为克隆或逆强化学习，使智能体复现专家行为。
56	特征级融合	Feature-Level Fusion	在对图像进行初步处理后，从中提取出边缘、形状、轮廓等关键特征，并对这些特性进行融合。
57	主模型	Primary Model	在模型集成或多任务学习中起主导作用的核心模型，或指大语言模型（LLM）中未经微调的原始基础版本，用于后续适配特定下游任务。
58	VOCs	VOSC(Volatile Organic Compounds)	挥发性有机化合物，指在常温下易蒸发的一类有机化学物质。
59	机理模型	Mechanistic Model	基于物理、化学或生物过程的第一性原理建立的数学模型，具有强可解释性和外推能力，区别于纯数据驱动模型。
60	慢时变	Slowly Time-Varying	描述系统参数或数据分布随时间缓慢变化的特性。
61	归因分析	Attribution Analysis	一种用于识别和量化不同因素对最终结果贡献度的分析方法。
62	自然语言处理技术	NLP(Natural Language Processing)	计算机理解、分析、生成人类语言的技术体系。
63	SPC 模型	SPC Model(Statistical Process Control Model)	基于统计过程控制理论构建的监控与诊断模型，用于实时检测生产或业务流程中的异常波动，确保过程处于受控状态。

序号	概念	缩写	核心含义
64	根因定位模型	Root Cause Localization Model	通过分析系统日志、指标或拓扑关系，自动识别导致故障或性能下降的根本原因的智能模型。
65	Top-K 准确率	/	一种更宽松、更实用的分类模型评估指标。它不要求模型预测的“第一名”必须正确，只要正确的答案出现在模型预测的“可能性最高”的前 K 个结果中，就算预测正确。
66	统计过程控制方法	SPC(Statistical Process Control)	利用控制图、过程能力指数等统计工具监控和控制生产过程稳定性的质量管理方法，旨在及时发现特殊原因变异并防止缺陷产生。
67	蒙特卡洛仿真	MCS(Monte Carlo Simulation)	通过大量随机抽样和统计实验模拟复杂系统行为或估计数学期望的方法。
68	条件生成式对抗网络	cGAN(Conditional Generative Adversarial Network)	GAN 的扩展形式，在生成器和判别器中引入额外条件信息，使生成过程可控。
69	FFT 转换（快速傅里叶变换）	Fast Fourier Transform	一种高效计算离散傅里叶变换（DFT）及其逆变换的算法，将时域信号转换为频域表示。
70	CNN-LSTM 网络	CNN-LSTM Network	一种混合深度学习架构，先用卷积神经网络提取空间局部特征，再通过长短期记忆网络建模时间序列依赖。
71	TF-IDF	Term Frequency-Inverse Document Frequency	一种用于衡量词语在文档中重要程度的统计方法，结合词频（TF）与逆文档频率（IDF），常用于信息检索、文本挖掘和关键词提取。
72	容器化	Containerization	一种轻量级虚拟化技术，将应用程序及其依赖打包为标准化单元（容器），实现跨环境一致运行。

序号	概念	缩写	核心含义
73	物模型	Thing Model	对物理世界中一个设备、资产或对象的数字化抽象，用于描述其属性、服务、事件等能力与状态。