AI 原生技术架构发展与实践研究报告

研究编制日期：2026 年 3 月 20 日

资料整理
火龙果科技 荐

目　录

摘 要 3

第 1 章 研究背景：企业智能化的时代命题 4

第 2 章 AI 原生核心技术体系与能力图谱 9

第 3 章 企业级 AI 原生架构参考范式 15

第 4 章 AI 原生全生命周期运营管理 21

第 5 章 企业 AI 原生能力成熟度评估 24

第 6 章 传统架构向 AI 原生演进路径 27

第 7 章 AI 原生架构落地验证 36

第 8 章 未来展望与实施建议 40

参考文献 42

资料整理
火龙果科技 荐

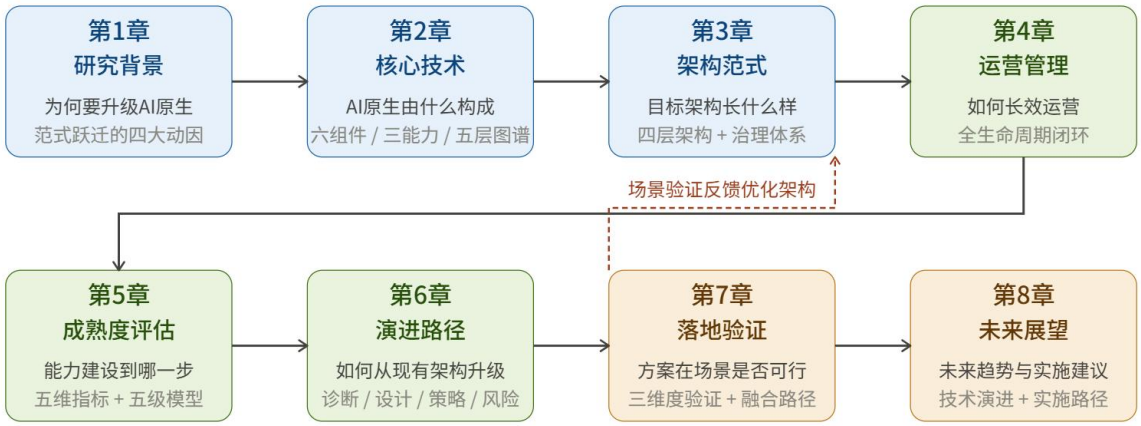
摘 要

当前，以大语言模型、多模态感知、智能体技术为核心的新一代人工智能，正推动企业信息技术体系从“数字化支撑”向“AI 原生驱动”的范式跃迁。AI 原生架构已成为企业规模化落地 AI 能力的核心底座，其建设质量直接决定了智能化转型的深度与可持续性。

本研究报告系统梳理了全球 AI 原生技术的演进脉络与产业实践现状，深入拆解了智能体、AI 网关、向量数据库、知识图谱等核心技术组件的能力边界与发展趋势；提出了企业级 AI 原生架构的通用参考范式，构建了“基础设施层-数据智能层-AI 能力开放层-应用服务层”四层能力体系；围绕全生命周期运营管理、能力成熟度评估、架构升级路径、场景化落地验证建等关键议题，形成了体系化的方法论框架与实践指引。

研究报告共分八章：第 1 章分析企业智能化转型的时代背景、核心驱动力与研究方法；第 2 章拆解 AI 原生核心技术体系与能力图谱；第 3 章提出企业级 AI 原生架构参考范式与设计方法论；第 4 章研究 AI 原生系统全生命周期运营管理范式；第 5 章构建企业 AI 原生能力成熟度评估体系；第 6 章规划传统架构向 AI 原生演进的方法路径；第 7 章阐述场景化验证方法；第 8 章展望未来趋势并提出企业实施建议。

AI原生技术架构研究报告总体研究思路



资料整理
火龙果科技 荐

第 1 章 研究背景：企业智能化的时代命题

1.1 从 AI 赋能到 AI 原生的范式跃迁

以大语言模型、多模态模型和 AI 智能体为代表的新一代 AI 技术集群，正从感知、认知、决策、执行等多个层面改变企业信息系统的构建方式与运行逻辑。这种变化不是单一技术点的突破，而是涵盖算力架构、数据组织、模型服务、应用集成等全栈技术体系的结构性重塑，要求企业在技术架构层面做出根本性回应。

当前，全球正处于新一轮科技革命与产业变革深度交织的关键时期。以人工智能为核心驱动力的新一轮科技变革，正在以前所未有的速度重塑全球产业格局与经济运行方式。能源、金融、制造等国民经济的基础性、战略性产业，面临数字化转型与智能化升级的双重历史任务。深入研判人工智能技术演进方向、产业变革态势及政策环境变化，是企业科学把握智能化战略定位和技术研究方向的基本前提。



图 1-1 企业信息技术体系的范式跃迁

推动这一范式跃迁的力量来自四个方面：一是技术侧，生成式人工智能、多模态人工智能与 AI Agent 三大技术方向相继成熟，使 AI 从“工具”变为“系统原生能力”；二是政策侧，国家“人工智能+”行动全面实施，行业监管与标准体系持续完善；三是产业侧，头部企业的 AI 原生架构实践从试点探索走向规模化部署，形成了可对标参照的行业坐标；四是需求侧，企业

资料整理
火龙果科技 荐

业务对智能化的诉求从单点功能升级为全价值链、全场景、全链条的体系化赋能。四股力量交汇，共同将“AI 原生架构”推到了企业数字化转型的核心议程。

1.2 生成式人工智能引领技术范式根本性变革

自 2022 年以来，以大语言模型为代表的生成式人工智能实现了跨越式突破，推动人工智能从判别分析时代迈入生成创造时代。大语言模型的参数规模持续突破，上下文理解能力大幅扩展，推理能力和指令遵循能力不断增强。Transformer 架构持续演进，混合专家模型、状态空间模型等新型架构不断涌现，模型训练效率和推理速度显著提升。人工智能的信息处理方式从单一文本扩展到图像、音频、视频、时序信号等多种模态，越来越接近人类的认知方式。

生成式人工智能已从实验室走向规模化行业应用。代码生成、文档撰写、数据分析、知识问答等通用场景中，人工智能的能力已接近甚至超越人类专家水平。生成式人工智能正加速向能源、金融、医疗、制造等行业渗透，行业大模型和场景大模型成为产业竞争的新焦点。模型推理成本快速下降，开源生态持续繁荣，人工智能技术的可及性和经济性显著改善，为大规模行业部署创造了有利条件。Gartner 技术成熟度曲线显示，生成式人工智能已跨过泡沫破裂期的低谷，进入稳步爬升的光明期，预计未来三至五年内将在多数行业中实现规模化部署。

然而，传统企业信息系统在设计之初并未将 AI 作为核心能力进行考量，应用架构以“人驱动系统功能完成业务”为基本假设，数据架构主要面向事务处理和统计报表，技术架构以 CPU 算力为中心。面对 AI 规模化落地的需求，这种架构形态暴露出明显的支撑短板：当前企业的技术架构仍主要面向传统业务场景设计，在算力调度弹性、模型服务能力、数据治理广度和安全防护深度等方面难以有效支撑 AI 原生应用的大规模部署和持续运营。具体而言：在算力调度层面，现有资源管理主要面向常规业务负载设计，算力分配采用相对固定的策略，对于 AI 训练和推理任务的动态突发需求缺乏弹性伸缩能力，GPU/NPU 等异构算力资源尚未实现统一池化和精细调度；在模型服务层面，模型训练环境、推理环境和业务应用环境紧密耦合，模型版本更新往往需要同步修改业务代码并重新部署应用，模型能力的复用依赖开发人员经验而非架构资产的规范指引；在数据治理层面，AI 应用大量依赖的图像、文本、音频、视频等非结构化数据尚未被纳入统一的数据资产管理体系，支撑检索增强生成的知识库构建和向量化存储缺少架构层面的数据基础；在安全防护层面，模型投毒、对抗攻击、提示词注入、有害内容生成等 AI 引入的新型安全风险尚未在现有安全架构中得到系统性覆盖。上述问题的叠加，使得核心业务领域的智能化升级缺乏清晰的技术路径指引。

这一判断同样适用于其他传统行业。凡是以“人操作系统”为基本假设构建的企业信息系统，在 AI 规模化落地时代都面临相同的结构性矛盾：AI 能力是外挂的、碎片化的、需要人工

资料整理
火龙果科技 荐

深度介入的，而业务需要的 AI 能力是内生的、体系化的、可自主运行的。解决这一矛盾的根本出路，就是以 AI 原生理念重构企业技术架构。

1.3 多模态人工智能拓展感知与认知新边界

多模态人工智能是人工智能技术演进的关键方向之一。传统人工智能模型主要处理单一模态数据，通常以文本为主，而多模态人工智能能够同时处理和理解文本、图像、音频、视频、传感器时序数据等多种数据类型，实现跨模态的感知、理解、推理和生成。多模态表示学习将不同模态的数据映射到统一的语义空间，跨模态对齐建立不同模态之间的语义对应关系，多模态融合整合多源异构信息形成综合判断。与单一模态相比，多模态的综合感知和推理能力更接近人类认知方式，判断结果也更加可靠。

多模态人工智能的技术成熟度正在快速提升。CLIP 等模型实现了图文跨模态语义匹配，Segment Anything 等模型具备零样本图像分割能力，Sora 等视频生成模型展示了多模态内容生成的巨大潜力。企业经营与生产过程中会产生海量多源异构数据，传统数据处理方式往往将这些数据割裂处理，无法充分利用不同模态之间的关联信息。多模态人工智能能够打破模态壁垒，实现不同类型数据的融合分析，从而获得更全面、更准确的态势感知和决策支持。然而，多模态人工智能对计算资源的需求远超单模态模型，对数据采集、标注、存储和管理也提出了更高要求，这些都需要企业在技术架构层面进行前瞻性规划和系统性升级。

从更普遍的意义上看，多模态能力的架构含义在于：企业数据架构必须从“面向结构化事务数据”扩展为“结构化与非结构化数据统一管理”，数据治理必须覆盖文本、图像、音频、视频等全模态数据的接入、加工、语义处理与检索利用。这是 AI 原生数据架构区别于传统数据架构的本质特征之一。

1.4 AI Agent 推动智能体从被动应答走向主动行动

AI Agent 即人工智能智能体，是当前人工智能技术发展的前沿方向，代表着 AI 从“被动应答”走向“主动行动”的范式跃迁。与传统的对话式人工智能不同，AI Agent 以大语言模型或多模态模型为核心驱动引擎，具备四大核心能力。第一大能力是自主理解与目标分解，能够将用户的模糊需求或宏观目标拆解为可执行的具体任务序列。第二大能力是路径规划与决策，能够在复杂环境中根据当前状态和约束条件规划最优执行路径。第三大能力是工具调用与环境交互，能够主动调用外部 API 接口、数据库、软件系统等工具完成具体操作。第四大能力是反思与迭代优化，能够根据执行结果的反馈信息进行自我评估和策略调整，形成持续学习的闭环。

资料整理
火龙果科技 荐

AI Agent 的理论基础可追溯至人工智能领域的经典智能体理论，但直到大模型技术的成熟才真正具备了可大规模落地的技术条件。AutoGPT、AgentGPT、MetaGPT 等多种通用智能体框架相继涌现，国内科技企业也在积极布局智能体技术与应用生态。AI Agent 的应用空间十分广阔：客户服务场景中可实现从人工为主向智能体自主服务为主的转变；运维检修场景中可实现从人工驱动向端到端自动化的转变；企业管理场景中可辅助完成经营分析报告编制、规章制度合规性审核等管理工作等。AI Agent 的引入对技术架构提出了超越传统 API 调用的新要求：应用架构需要支持智能体的定义、编排、部署和运行管理，技术架构需要支持任务规划引擎、工具调用框架和记忆管理机制的部署，安全架构需要覆盖智能体的行为审计和权限管控。这些要求推动企业技术架构从“服务导向”向“智能体导向”演进。

从全行业实践来看，智能体导向的架构演进已成为普遍共识。客服、风控、研发、供应链等各领域的领先企业均在探索智能体与既有业务系统的集成范式，其共性要求是一致的：智能体运行框架、工具注册机制、记忆管理组件成为应用架构的新成员，智能体行为可审计、权限可管控成为安全架构的新命题。

1.5 政策驱动：数智化战略下的架构升级要求

国家层面持续深化人工智能战略布局，“十五五”规划纲要明确提出全面实施“人工智能+”行动，加强人工智能同科技创新、产业发展、民生保障、社会治理相结合，抢占人工智能产业应用制高点。算电协同、超大规模智算集群等新型基础设施纳入新基建工程，为产业数智化转型提供了底层支撑。

政策供给在近两年显著提速。2024 年，国务院国资委部署中央企业全面推进“AI+”专项行动，要求中央企业把发展人工智能放在全局工作中统筹谋划，加快布局和发展人工智能产业，在算力基础设施、高质量数据集、行业大模型等领域率先突破；2025 年，《关于深入实施“人工智能+”行动的意见》进一步明确以人工智能引领科研范式变革、赋能产业发展、提升民生福祉的实施路径。国际层面，ISO/IEC 42001 人工智能管理体系、欧盟《人工智能法案》等相继落地，人工智能治理与合规要求日趋刚性。政策与标准的双重驱动，使企业 AI 架构建设从“自发探索”进入“有章可循”的新阶段。

对于各行业企业而言，全面智能化战略不再是可选的创新方向，而是数字化转型的核心主线。技术架构需从“支撑数字化”向“原生智能化”演进，要求在系统设计之初便将 AI 能力作为核心要素，围绕数据、模型和算力进行系统性架构重构，而非在现有架构上叠加 AI 功能。这一战略转型对技术架构提出了一系列新要求：架构需要灵活接入大模型能力、支持智能体协同

资料整理
火龙果科技 荐

运行、实现数据与知识的双轮驱动、具备持续学习与自演进能力，并能够支撑全域、全场景、全链条的智能赋能。

1.6 产业实践：行业 AI 原生架构建设的现状与挑战

全球各行业头部企业已率先启动 AI 原生架构探索，互联网科技企业在模型基础设施、智能体体系、能力中台等领域形成了较为成熟的实践；能源、金融、制造等传统行业也逐步将 AI 能力深度嵌入业务全链条，从试点探索向规模化部署过渡。

全球范围内，AI 原生能力正成为重塑企业竞争格局的核心变量。装备制造领域，西门子、ABB 等企业已将人工智能深度嵌入产品研发、生产制造和运维服务全链条，实现产品全生命周期的智能化升级；能源领域，法国 EDF、美国特斯拉能源、日本东京电力等企业分别在机组运行优化、储能调度、智能巡检等场景落地了规模化 AI 应用；金融领域，全球头部银行与保险机构纷纷构建 AI 原生技术底座，支撑智能风控、智能客服、智能投研等核心场景的体系化落地；零售与消费领域，AI 原生能力已深度融入供应链调度、用户运营、商品推荐等全业务环节。

国内各行业的 AI 原生建设同样加速推进。能源行业围绕数字化战略构建人工智能中台体系，在智能巡检、智慧供应链、智慧营销等领域部署了大规模人工智能应用；金融行业持续推进 AI 技术与业务的深度融合，在风险防控、客户服务、运营管理等领域形成了体系化应用；制造行业加速布局智能工厂与智能运维，推动 AI 能力向生产现场下沉。总体来看，人工智能在各行业的应用正呈现三大共性趋势：从试点探索向规模化部署过渡，从单点应用向平台化、体系化方向发展，从技术验证向业务价值创造转变。在这一进程中，技术架构的 AI 原生能力已成为决定 AI 应用能否规模化落地的关键因素，领先企业已将架构升级作为 AI 战略的基础性工程优先推进。

AI 原生系统的建设在全球范围内仍处于快速演进之中，技术路线、架构范式、最佳实践尚未形成统一标准。各领先企业基于自身业务特征与技术积累，走出了差异化的 AI 原生架构建设路径，但由于各企业的行业属性、业务特征、技术基础存在差异，其经验的适用条件需要通过系统研究加以甄别。当前，企业在推进 AI 原生技术架构设计过程中，普遍面临三个方面的认知不足。第一，对全球 AI 原生架构的技术演进趋势和政策标准动态掌握不全面，不清楚哪些技术方向是长期趋势、哪些是短期热点。第二，对行业领先企业的具体做法、关键决策和实际效果缺乏系统性了解，不清楚同行在类似问题上做出了怎样的选择。第三，对领先实践背后的适用条件和关键成功因素缺乏深入分析，不清楚外部经验如何转化为适合企业自身的操作方案。上述问题可能导致架构设计在方向判断、技术选型和路径规划上缺乏充分的外部依据，增加决策风险。

资料整理
火龙果科技 荐

第 2 章 AI 原生核心技术体系与能力图谱

2.1 AI 原生核心技术组件解析

前序章节回答了趋势判断和经验参照的问题，回答了“往哪个方向走”，但尚未回答“依靠什么技术、按照什么标准走到那里”，这是从战略认知转向工程落地必须回应的核心命题，直接决定了架构建设的技术选型逻辑、组件协同边界与治理体系设计。AI 原生架构的本质是以大模型为核心驱动的新型技术体系，其构建涉及多类技术组件的系统性集成与协同运行。从技术层级与功能定位划分，AI 原生架构的核心技术组件主要包括：应用执行层的智能体、能力开放层的 AI 网关、数据智能层的向量数据库与知识图谱、模型引擎层的模型推理引擎，以及贯穿全栈的 AI 安全合规组件。

2.1.1 智能体：AI 原生应用的基本单元

智能体是 AI 原生应用区别于传统应用的基本运行单元，是 AI 能力的自主执行载体。它将传统应用中固化于代码的流程逻辑，转化为可编排、可组合、可进化的运行时对象，为业务流程的智能化改造提供标准化的能力载体。智能体的核心能力涵盖自主决策边界界定、多智能体协同机制、记忆与上下文管理、工具调用与外部系统集成四大维度，其技术内涵的本质是从“流程由代码定义”向“流程由目标驱动”的范式跃迁。

智能体技术的演进方向呈现三大核心趋势。第一，从单智能体演示走向多智能体协同生产化运行，多智能体的协同机制、权限治理与任务调度成为核心竞争点，企业级协同标准逐步形成。第二，从通用开发框架走向行业模板与领域工具链沉淀，场景化适配能力持续提升，垂直领域的专用智能体加速成熟，金融、能源、制造等行业的领域智能体将形成差异化竞争力。第三，从代码优先开发走向低代码化、声明式的智能体定义方式，落地门槛持续降低，逐步向业务可配置方向演进。

2.1.2 AI 网关：AI 能力开放层的统一治理枢纽

AI 网关是 AI 能力开放层的治理门户，解决大模型服务“接入分散、管控缺失”的核心问题，为多模型环境下的统一接入、安全管控、成本治理提供枢纽支撑。AI 网关的核心能力涵盖多模型路由策略、版本管理、限流熔断、输出安全过滤、成本监控与优化五大模块，是企业级多模型环境不可或缺的中台化治理组件。

资料整理
火龙果科技 荐

AI 网关的部署形态正在从独立的网关组件向“网关+治理平台”一体化方向演进，功能边界从单纯的接入转发向全链路治理延伸，与模型注册、能力目录、效果度量形成闭环，逐步成为企业 AI 能力开放层的核心治理枢纽。同时，网关智能化能力持续增强，动态路由、上下文感知、智能限流等智能化策略逐步成为标准能力。

2.1.3 向量数据库：非结构化数据的语义检索基座

向量数据库是数据智能层的新型核心组件，支撑非结构化数据的高效存储与语义检索，是 AI 原生数据架构区别于传统数据架构的标志性组件。其核心技术要素涵盖索引算法（HNSW、IVF 等近似最近邻检索算法）、相似度计算（余弦相似度、欧氏距离、内积）、混合查询（向量检索与结构化条件过滤的联合查询）、持久化与高可用机制四大维度。企业选型中重点关注混合查询能力、与模型服务的链路性能、运维成熟度、与现有数据平台的集成能力。

向量数据库技术的演进方向呈现三大核心趋势。第一，向多模态向量检索扩展，支持文本、图像、音频、视频等多模态数据的统一向量表示与跨模态检索。第二，向混合查询能力深化，强化向量检索与结构化条件、全文检索、图遍历的复合查询能力，适配更复杂的业务查询场景。第三，向数据平台一体化方向发展，与结构化数据、图数据、知识图谱实现统一治理，逐步成为企业数据智能层的标准组件。

2.1.4 知识图谱：结构化知识的语义网络基座

知识图谱是数据智能层的另一关键组件，通过实体、关系、属性的结构化建模，将领域知识转化为机器可理解、可推理的语义网络。在 AI 原生架构中，知识图谱与向量数据库形成互补关系：向量数据库负责海量非结构化数据的语义检索，知识图谱负责结构化知识的精确推理与关系挖掘，二者协同构成“语义检索+逻辑推理”的复合智能基座。知识图谱的核心技术涵盖本体建模、知识抽取、知识融合、知识推理、图存储与查询五大维度。

知识图谱技术的演进方向呈现三大核心趋势。第一，与大模型的深度融合加速，KG 增强 LLM（KG-LLM）与 LLM 增强 KG（LLM-KG）的双向增强成为研究热点，知识图谱为大模型提供事实校验与可解释性，大模型为知识图谱提供自动化构建能力。第二，神经符号结合逐步成熟，将神经网络的学习能力与符号推理的逻辑能力结合，弥补纯统计学习缺乏推理能力的短板。第三，图数据库与向量数据库的统一治理趋势明显，多模态数据（结构化属性、向量特征、图关系）的统一存储与联合查询能力成为关键竞争点。

2.1.5 模型推理引擎：AI 能力的核心执行载体

资料整理
火龙果科技 荐

模型推理引擎是模型引擎层的核心执行组件，承担大模型从离线训练产物向在线服务能力转化的关键职能。它通过模型压缩、推理加速、服务编排、异构硬件适配等技术手段，将训练完成的模型高效、稳定、可扩展地部署到生产环境，是 AI 原生架构区别于“AI 即调用 API”的标志性技术组件。

模型推理引擎呈现三大演进趋势。第一，端云协同推理成为重要方向，云端大规模推理与端侧轻量推理融合协同，模型切分与自适应推理技术逐步成熟。第二，推理优化技术持续深化，量化、稀疏化、注意力机制优化（KV Cache、Continuous Batching、PagedAttention、Speculative Decoding）等技术持续突破，推理效率与成本控制能力成为核心竞争力。第三，专用推理芯片与异构算力调度深度结合，GPU、NPU、TPU 等异构算力的统一调度与最优分配能力成为平台级关键能力。

2.1.6 AI 安全合规组件：贯穿全栈的治理基线

AI 安全合规组件是贯穿 AI 原生架构全栈的基础性保障组件，覆盖数据层、模型层、应用层、输出层全链路的安全风险与合规要求，是企业级 AI 应用规模化落地的必要前提。其核心能力涵盖数据安全与隐私保护、模型安全与鲁棒性、输出内容安全、合规审计与可追溯性四大维度，是 AI 原生架构"治理能力"的具象化体现，体现“安全左移”的治理理念。

AI 安全合规技术的演进方向呈现三大核心趋势。第一，安全合规从“事后审计”向“事前预防-事中控制-事后审计”全链路延伸，内置安全（Security by Design）成为架构设计的核心原则。第二，AI 合规标准化进程加速，行业级、国家级、国际级的 AI 合规标准逐步形成，企业级合规基线成为产品准入门槛。第三，可解释性与可追溯性成为 AI 安全的核心议题，大模型的可解释机制、决策链路的可追溯能力成为差异化竞争点，与知识图谱的结合为可解释 AI 提供新的技术路径。

企业在开展相关能力建设上，**智能体方面**，需聚焦定义规范与核心能力框架，涵盖自主决策边界、多智能体协同机制、记忆与上下文管理、工具调用与外部系统集成，分析主流框架的技术路线差异与适用场景，评估成熟度及演进方向。**AI 网关方面**，需聚焦模型服务的统一接入与治理机制，涵盖多模型路由策略、版本管理、限流熔断、输出安全过滤、成本监控与优化。**向量数据库方面**，需聚焦非结构化数据的高效存储与语义检索，涵盖索引算法、相似度计算、混合查询、持久化与高可用机制。**模型推理引擎方面**，需聚焦大模型从训练产物向在线服务转化的核心执行能力，涵盖模型压缩、推理加速、服务编排与异构硬件适配，分析开源框架、商用平台与云原生服务三条路线的适用边界，研判端云协同推理与算力调度演进方向。**知识图谱方面**，需聚焦结构化知识的语义网络建模与推理能力，涵盖本体建模、知识抽取、知识融合、图存储与查询，研究通用图谱、行业图谱与企业自建图谱的差异化路径，研判其与大模型双向

资料整理
火龙果科技 荐

增强及图-向量统一治理趋势。**AI 安全合规组件方面**，需聚焦贯穿数据、模型、应用、输出全栈的安全治理能力，涵盖数据隐私保护、模型鲁棒性、内容安全过滤与合规审计，研究安全左移、全栈治理与可解释 AI 的演进路径。

上述六大组件在架构中形成“应用执行-能力开放-模型引擎-数据智能-全栈治理”的完整能力链条，唯有通过各组件的协同设计与一体化治理，才能构建具备自主性、可靠性、可治理性的 AI 原生架构，支撑企业级智能应用的规模化落地。

2.2 AI 原生技术体系核心能力

上述组件视角回答“AI 原生技术体系由哪些技术单元构成”，而本节从能力视角回答“AI 原生技术体系应具备哪些核心能力”。AI 原生技术体系可划分为数据自驱动、模型自进化、安全可信三大核心能力维度，各维度共同构成了架构运行的底层能力支撑。

数据自驱动维度需梳理从采集、治理、特征工程到质量反馈闭环的技术链条，实现数据从业务副产品向 AI 核心生产要素的转变。在 AI 原生架构中，数据不再是流程的衍生产物，而是驱动模型训练、智能体决策、业务创新的核心生产要素，要求建立自动化、闭环化、可度量的数据供给能力。

模型自进化维度需梳理从持续训练、评估、版本管理到在线更新的闭环机制，明确 MLOps 在整个技术体系中的核心作用。模型不再是“训练-上线-固化”的单次产物，而是需要在运行中持续吸收新数据、持续评估效果、持续迭代优化的动态对象，要求建立覆盖全生命周期的模型治理能力。

安全可信维度需梳理模型安全防护、数据隐私保护、内容合规审查、可解释性保障等技术能力，为 AI 全链路运行筑牢安全底座。安全可信不是事后的修补机制，而是需要在数据、模型、应用、输出各环节内嵌的治理能力，体现"安全左移"的设计理念。

三大能力维度并非并列关系，而是相互咬合的循环体系：数据自驱动为模型自进化供给燃料，模型自进化在运行中产生新的数据需求与安全暴露面，安全可信则为前两者的运行划定边界。企业架构设计中应避免将三者割裂部署，而应通过统一的治理体系实现协同。

2.3 AI 原生技术要素图谱

在明确了三大能力维度之后，需进一步将能力维度映射到具体的架构层级，构建 AI 原生技术要素图谱，厘清基础设施层、数据层、模型层、能力开放层、应用层的层级归属、功能定位与交互逻辑，以可视化方式呈现各组件的技术依赖与数据流转路径。

资料整理
火龙果科技 荐



图 2-1 AI 原生技术体系五层要素图谱

基础设施层包括算力调度、云边端协同、弹性扩展等组件；数据层包括多源数据汇聚、数据治理、知识图谱、向量数据库等组件；模型层包括训练平台、推理服务、版本管理等组件；能力开放层包括模型注册、AI 网关、智能体编排等组件；应用服务层包括智能体运行框架、AI 原生应用框架、多端适配等组件。

基础设施层向数据层提供算力与存储支撑，数据层向模型层供给高质量数据集，模型层通过能力开放层将 AI 能力标准化开放，应用服务层消费开放能力并承载业务流程，安全可信与治理机制纵向贯穿五层。

2.4 AI 原生技术标准体系

当前，AI 原生技术标准框架的缺失是行业面临的突出问题。企业各业务域引入 AI 技术组件时缺乏统一标准，技术选型、接口规范、部署方式各自为政，可能形成新的技术孤岛。流程规范缺失影响交付效率与运行质量，评估体系缺失使建设成效缺乏统一衡量标尺。

技术标准层面，制定智能体接口规范、模型服务接入标准、向量数据库使用规范、AI 网关部署要求等核心技术标准，确保各组件选型与建设有标准可依。

流程规范层面，制定 AI 系统全生命周期的标准化作业程序，涵盖需求、设计、开发、部署、运维各阶段的关键活动与交付物。

资料整理
火龙果科技 荐

评估体系层面，建立 AI 技术组件与系统建设成效的评估指标体系，明确评估维度、方法与基准，为技术选型与成效评估提供量化依据。

安全合规层面，覆盖数据隐私保护、模型安全防护、内容合规审查、可解释性保障等安全合规要求，明确数据授权、隐私计算、内容过滤、审计追溯的标准化要求，筑牢安全底座。

此外，在国际层面，ISO/IEC JTC 1/SC 42 人工智能国际标准体系、IEEE 人工智能架构相关标准正在持续完善，成为各国技术标准的重要参考。国内也在加快智能体、大模型服务、AI 安全等领域的国家标准制定，逐步形成与国际接轨的标准体系。企业应在跟踪国际标准动态的同时，结合自身业务特征建立企业级标准体系，形成“国际标准—国家标准—行业标准—企业标准”的完整衔接。

资料整理
火龙果科技 荐

第 3 章 企业级 AI 原生架构参考范式

3.1 设计方法论：从案例剖析到方法体系

前序任务完成了趋势研判、技术研究，从外部视野构建了 AI 原生架构设计的完整认知基础，但尚未回答最核心的问题：结合企业实际，AI 原生架构具体应该如何设计。AI 原生架构设计不能仅依靠技术组件的罗列与堆砌，需要一套贯穿全生命周期的方法论来指导从需求分析到持续运维的完整过程。这套方法论的形成不能脱离对领先实践的深入剖析，那些已在 AI 原生建设上取得实质性进展的企业，其全生命周期路径蕴含着大量可借鉴的经验与可规避的教训。通过对典型案例的系统解构，可以提炼各阶段的核心动作、关键决策与成功要素，形成结构化的方法论框架。

3.1.1 典型案例的全生命周期剖析

基于领先实践研究成果，选取覆盖不同行业类型、架构模式的代表性 AI 原生系统案例，对其全生命周期进行深度剖析。系统梳理从需求发起、架构设计、开发部署到迭代优化的关键节点。

需求发起阶段，分析需求来源与定义方式，研究如何将业务价值目标转化为可执行的系统需求，识别最适合由 AI 原生系统承载的场景及关键决策权衡。

架构设计阶段，分析设计输入、原则与决策过程，研究技术方案选型判断、系统边界与模块划分、数据流与控制流设计，以及主要挑战与应对策略。

开发部署阶段，分析开发模式、工程实践与部署策略，研究开发、训练、测试、部署一体化流水线的实现，模型版本与代码版本的协同管理，以及质量保障与交付节奏的平衡。

迭代优化阶段，分析运维模式、效果监控与持续优化机制，研究运行数据驱动的持续改进、模型效果衰减预警与应对，以及业务需求变化与技术系统演进之间的动态适配。

通过上述剖析，提炼各阶段的共性规律与关键成功要素，建立系统化的案例分析库，总结形成覆盖 AI 原生系统全生命周期的闭环方法论。

案例一：微软 Microsoft 365 Copilot—SaaS 产品体系的 AI 原生重构

需求发起阶段：核心动因是解决 Office 生态用户在文档处理、数据分析、会议协作中的效率瓶颈，原有产品功能均为固化的命令式交互，无法主动提供上下文辅助与任务执行能力。项目核心目标是将 AI 能力原生嵌入办公全场景，而非作为外挂插件独立运行。关键决策权衡在

资料整理
火龙果科技 荐

于：是各产品独立开 AI 能力，还是构建统一的 AI 原生底座支撑全产品矩阵。最终选择统一底座路线，避免重复建设，保障体验一致性与安全治理统一性。

架构设计阶段：遵循“原生嵌入、统一底座、安全合规”的设计原则，构建三层架构体系：底层为大模型算力底座与向量知识层，统一管理多模型能力与企业知识；中间层为 Copilot 核心编排层，承载提示工程、工具调用、上下文管理、安全过滤等通用能力；上层为应用嵌入层，通过标准化 SDK 对接 Word、Excel、Teams 等全产品矩阵。关键决策包括采用统一 AI 网关实现多模型路由与治理，采用统一向量数据库构建企业知识语义底座，建立贯穿全链路的安全合规体系。针对多产品集成一致性的挑战，通过标准化接口规范与能力声明机制实现统一管控。

开发部署阶段：采用分阶段灰度部署策略，先面向少量企业客户开放预览功能，验证效果与稳定性后逐步扩大范围，最终实现全量用户覆盖。开发模式采用产品团队与 AI 团队协同作战模式，建立 MLOps 一体化流水线，统一管理模型版本、代码版本与配置版本，实现模型迭代与产品迭代的协同推进。质量保障采用分场景验证机制，同时覆盖功能可用性、内容准确性与合规性。交付节奏采用“小步快跑”模式，按月迭代功能优化，按季度推出核心能力升级。

迭代优化阶段：建立覆盖功能使用率、用户满意度、内容合规率、任务完成率的全链路效果监控体系。基于用户行为数据与使用反馈，持续优化提示工程策略、工具调用逻辑与上下文管理机制，针对不同行业场景推出行业化配置模板。建立模型定期增量更新机制，基于新增数据持续优化模型效果，应对数据分布漂移。最终形成“用户反馈-效果监控-模型迭代-功能升级”的完整闭环，驱动能力持续进化。

案例二：西门子工业 AI 平台—工业场景的 AI 原生架构落地

需求发起阶段：核心痛点是工业设备运维、生产流程优化高度依赖人工经验，故障预判滞后、生产效率提升遇瓶颈；原有各产线独立的自动化系统数据割裂，AI 能力单点部署、无法复用。项目目标是构建企业级工业 AI 原生底座，支撑全工厂、全价值链的智能化升级。关键决策权衡在于：是集中式云端部署，还是云边端协同部署。最终选择云边端协同路线，兼顾工业现场低时延、高可靠要求与云端算力、知识沉淀优势。

架构设计阶段：构建云边端协同的分层架构，边缘侧部署现场数据采集节点与轻量推理引擎，满足低时延控制类场景需求；云端部署模型训练平台、知识管理体系与智能体编排层，承载复杂分析、全局优化类能力。核心组件包括工业 AI 网关、时序向量数据库、工业场景智能体框架。关键决策是制定统一的数据模型与接口标准，打通不同设备、不同产线的数据壁垒。针对工业场景高可靠、高安全的约束，采用“核心能力边缘下沉、非实时能力云端处理”的分级部署策略，保障生产运行稳定性。

资料整理
火龙果科技 荐

开发部署阶段：采用“试点产线验证—多产线复制—全工厂推广”的渐进式部署路径。开发模式采用 OT 运营团队、IT 技术团队与 AI 算法团队跨领域协作，建立适配工业场景的 MLOps 体系，兼顾模型准确率与现场运行可靠性。部署节奏遵循“先非核心、后核心”的原则，先在设备巡检、质量检测等非核心场景验证，再逐步渗透到生产控制、故障预警等核心生产场景。

迭代优化阶段：建立设备级、产线级、工厂级三级效果监控体系，核心监控指标包括故障预测准确率、生产效率提升率、质量不良率下降幅度。基于现场持续产生的运行数据开展增量训练，持续适配不同设备工况、产线工艺变化。沉淀设备运维、质量管控等通用场景的智能体模板与知识组件，在集团内各工厂、各业务线复用推广。最终形成“现场数据采集-模型迭代优化-能力沉淀复用-业务价值提升”的闭环。

通过对上述案例剖析，可以提炼出 AI 原生系统建设的若干共性规律：

规律一，架构形态均呈现“分层解耦”特征。 基础设施、数据、智能能力、应用服务各层级职责清晰、协同运作，通过标准化接口契约实现上下层解耦，保障各层可独立演进。

规律二，核心能力均通过“中台化”方式沉淀复用。 通用能力下沉、共性能力沉淀，避免各业务域重复建设，提升整体研发效率与治理一致性。

规律三，安全可信均贯穿架构全链路。将合规要求内嵌于架构设计而非事后补救，覆盖数据、模型、接口、应用全层级的风险防控。

规律四，迭代优化均形成闭环。 案例中的四个阶段并非线性串行，而是以迭代优化阶段的反馈为起点形成闭环：运行数据揭示的效果衰减驱动新一轮需求发起，业务变化带来的能力缺口触发架构再设计，流水线的持续交付能力决定迭代节奏的上限。

3.1.2 与企业架构方法的有机融合

跨案例共性规律为提炼 AI 原生架构设计原则提供了实践基础，但方法论的落地还需要与企业既有的企业架构方法论进行有机融合。目前多数大型企业已建立一套基于 TOGAF 的企业架构方法体系，在应用架构、数据架构、技术架构各层面形成了成熟的设计规范。AI 原生架构的引入不是推翻这套体系，而是在现有方法基础上进行扩展与升级。如何在应用架构层面实现业务流程的智能化重构与 AI 能力的微服务化封装，如何在数据架构层面实现知识图谱构建与非结构化数据有效治理，如何在技术架构层面实现云边端协同的技术栈选型与部署架构设计，这些都需要在充分结合企业架构方法论的前提下给出明确的设计指引。

具体而言，可从应用、数据和技术三个方向完善相关设计方法。

资料整理
火龙果科技 荐

在应用架构方面，重点分析 AI 介入业务流程的适宜位置和价值切入点，结合智能化改造需求重新梳理业务活动、流程节点及触发机制，并建立面向 AI 能力服务化的封装思路，规范推理服务的调用接口、运行质量、版本维护和生命周期管理方式，进一步明确智能服务与既有业务服务之间的组合、协同及调用关系。

在数据架构方面，重点探索知识资源的体系化组织方式，研究领域概念、实体及其关联关系的建模规则，完善知识资产与现有业务数据之间的映射和联动机制。同时针对文本、图片、语音、视频等多类型信息，建立覆盖接入、加工、语义处理、向量转换、存储及检索利用的数据治理方法，并明确向量存储等新型数据能力在整体架构中的作用边界。

在技术架构方面，结合总部、云侧、边缘节点等不同运行环境，研究计算资源部署、模型服务布置和云边协同训练推理的技术组织方式，并充分考虑关键业务对低时延、高稳定性和连续运行的要求。

在此基础上，对智能体运行框架、统一 AI 访问入口、向量数据服务、模型开发训练环境等关键能力进行整体编排，明确各组件职责、连接方式及协作机制，最终沉淀形成一套可用于指导后续 AI 原生系统设计和建设的架构方法体系。

3.2 AI 原生架构设计核心原则

企业级 AI 原生架构设计需遵循分层解耦、能力复用、弹性扩展、安全可信的核心原则，在保障系统高可靠、高安全运行的基础上，最大化释放 AI 技术的价值。

分层解耦原则要求各层级之间通过标准化接口交互，上层依赖下层能力但不耦合下层实现，确保各层可独立演进、灵活替换。

能力复用原则要求通用能力下沉、共性能力沉淀，避免重复建设，提升整体研发效率。

弹性扩展原则要求算力、数据、模型能力均可根据业务需求动态伸缩，适配 AI 负载的波动性特征。

安全可信原则要求安全治理贯穿架构全层级，覆盖数据、模型、接口、应用全链路。

3.3 分层能力架构体系设计

技术架构总体框架在整个技术体系中处于承上启下的核心地位。向上承接 AI 原生架构设计方法的指引，将设计原则与规范转化为具体的架构形态与组件定义；向下指导各业务域、各系统的 AI 原生架构设计与建设实施，为技术选型、系统开发、能力部署提供统一的架构基准与

资料整理
火龙果科技 荐

规范约束。缺少这一总体框架的统一规定，各业务域在 AI 原生建设中将缺乏共同遵循的蓝图，可能出现架构形态各异、组件选型分散、层级划分混乱、交互方式不统一等问题，最终形成新的技术孤岛。

企业级 AI 原生总体架构涵盖基础设施层、数据智能层、AI 能力开放层、应用服务层四大核心层级及贯穿各层的统一治理体系。各层级划分需遵循清晰的职责边界，避免功能重叠与职责模糊。层级之间的依赖关系与调用链路需明确界定，上层对下层的调用遵循规范的接口契约，下层对上层的支撑具备明确的服务边界。



图 3-1 企业级 AI 原生架构总体框架

3.3.1 基础设施层

基础设施层定义支撑 AI 原生系统运行的算力基础设施架构，涵盖通用与异构算力的统一管理与调度、云边端协同部署、算力弹性扩展与成本优化策略，明确向上提供的算力服务接口与服务质量要求，为上层应用提供稳定、弹性、高效的算力底座。

3.3.2 数据智能层

数据智能层定义面向 AI 的数据架构与知识体系，涵盖多源异构数据的采集、存储、治理与服务全链条，结构化与非结构化数据的统一管理，向量数据库、图数据库等新型组件的定位

资料整理
火龙果科技 荐

与使用规范，以及知识图谱构建、数据语义层建设、高质量数据集管理与供给的架构设计，明确向上层提供的数据与知识服务接口。

3.3.3 AI 能力开放层

AI 能力开放层定义 AI 能力的标准化封装与开放治理架构，涵盖模型训练平台、推理服务、版本管理与持续迭代机制，AI 能力以标准化 API 进行注册、发布、调度与治理的规范，智能体的定义、编排、运行与监控框架，AI 网关在统一接入、路由、治理与安全防护中的定位与功能，以及能力效果度量与持续优化机制。

3.3.4 应用服务层

应用服务层定义 AI 原生应用的架构范式与开发运行框架，涵盖应用的模块化组织、智能体与传统应用的协同集成、AI 能力嵌入业务流程的架构模式、多端适配与体验一致性保障，以及应用层的可观测性、可配置性与可扩展性设计。

3.3.5 一体化治理体系

统一治理体系贯穿架构各层级，是保障架构规范、稳定、安全运行的制度支撑，主要包含安全保障、运维运营、标准规范三大体系。

安全保障体系覆盖数据安全、模型安全、内容安全、访问控制四大领域，建立全链路的风险防控机制。运维运营体系涵盖全链路可观测性、智能运维、成本管理与资源优化，实现从被动运维向主动运营的转变。标准规范体系包括各层级的接口标准、数据标准、安全标准与运维标准，确保各组件、各系统的互联互通与规范运行。

三大体系与四层架构的关系是“纵向贯穿、横向协同”：安全保障体系在每层都有具体控制点（如基础设施层的算力隔离、数据智能层的隐私计算、能力开放层的输出过滤、应用服务层的权限管控）；运维运营体系将各层的可观测数据汇聚为统一运营视图；标准规范体系则将层间接口契约固化为可执行的规范条款。

基于上述总体框架的整体视图构建与规范输出，将各层级设计整合为统一的总体框架视图，以可视化方式呈现各层级、各组件的结构关系与交互链路，使架构设计人员、系统建设人员与运维管理人员能够据此理解架构的整体形态与运行逻辑。同时以规范的架构文档形式输出，明确各层级、各组件的定义、职责、接口与约束，为后续详细设计与建设实施提供可直接引用的架构规范。

资料整理
火龙果科技 荐

第 4 章 AI 原生全生命周期运营管理

4.1 从 IT 运维到 AI 运营的范式转变

AI 原生系统的全生命周期管理与传统信息系统有本质区别，是从架构设计走向长效运营的关键延伸功能，是确保 AI 原生架构不仅建得起、更能用得好且持续产生价值的重要保障。传统系统上线后以日常运维为主，变更频率低。而 AI 原生系统的核心是数据持续驱动、模型持续演进、效果持续优化，模型性能会随数据分布变化而衰减，业务场景演化要求能力持续扩展。这些特征决定了管理与运营必须超越传统 IT 运维，建立涵盖敏捷需求管理、研发运营一体化、模型持续学习、智能监控自愈等全新维度的管理机制。



图 4-1 AI 原生系统全生命周期运营闭环

这一运营闭环由需求管理、研发运营、持续学习、监控自愈、能力治理、价值度量等环节构成，各环节以数据和效果为纽带相互驱动：需求管理将业务价值转化为可验证的能力假设，研发运营一体化将假设快速交付为运行中的服务，持续学习机制保障服务效果不随时间衰减，监控自愈体系守护运行质量，能力治理统筹共享复用，价值度量则将运行效果反馈给需求管理，形成螺旋上升的运营循环。缺少任何一环，运营闭环都会退化为传统的“上线即固化”模式。

4.2 敏捷需求管理与价值验证

资料整理
火龙果科技 荐

明确 AI 原生系统需求从业务价值识别、可行性评估、优先级排序到迭代交付全过程的管理方法。包括如何将模糊的业务问题转化为可验证的 AI 能力需求，如何建立需求假设与验证之间的闭环机制，如何通过快速原型验证需求的真实价值，以及如何建立贯穿需求定义到效果评估的价值度量指标体系。

与传统需求的“确定式交付”不同，AI 需求天然具有假设属性：上线前无法精确预知模型效果，需求验证贯穿交付全过程。因此敏捷需求管理的核心动作包括：需求阶段即定义成功标准（如准确率阈值、采纳率目标、响应时长承诺），以可证伪的方式描述需求；优先级排序引入“价值×可行性×数据就绪度”三维评估，避免高价值但数据基础不具备的需求占用资源；以两到四周的迭代节奏交付可用增量，每轮迭代结束时对照成功标准做出继续、调整或终止的决策；建立需求假设台账，记录每个已验证与已证伪的假设，沉淀为组织的学习资产。

4.3 研发运营一体化（MLOps）体系

研发运营一体化是 AI 原生系统运营的核心基础，重点关注 MLOps 与 CVOps 的企业落地方法。MLOps 方面包括模型开发、训练、测试、部署、监控、治理全流程的标准化与自动化管理，明确数据科学家与运维人员的协作模式与职责边界，建立从实验环境到生产环境的模型发布流水线，实现模型版本与代码版本的协同管理。CVOps 方面包括计算机视觉模型特有的数据标注管理、模型压缩加速、边缘端部署等管理方法，最终形成适用于实际实际条件的研发运营一体化操作指南。

4.4 模型持续学习与版本治理

针对 AI 模型效果持续衰减的特征，需建立模型持续学习与版本治理的机制。持续学习方面需聚焦模型持续训练与更新的触发条件与频率策略，明确数据反馈机制、增量学习策略与效果评估方法，确保模型持续适应业务场景变化与数据分布漂移。版本治理方面需聚焦模型版本命名规范、分级管理制度、上线审批流程与回滚策略，明确版本变更的影响分析、风险评估与发布管理要求，建立多版本并行运行与灰度切换的管理机制，确保更新过程中服务不间断与效果平稳过渡。

4.5 全链路智能监控与自愈运维

监控体系需覆盖从基础设施层到应用服务层的全链路观测能力，包括算力资源利用率、模型推理性能、数据质量变化、业务效果指标等多维度实时监控。

资料整理
火龙果科技 荐

监控层级	核心监控指标	典型异常场景
资源层	算力利用率、显存占用、节点健康度、任务排队时长	算力争抢、节点故障、队列拥塞
服务层	推理延迟、吞吐量、调用成功率、token 消耗与成本	性能劣化、限流触发、输出异常
效果层	模型准确率漂移、数据分布偏移、业务采纳率、用户反馈	效果衰减、概念漂移、体验退化

自愈能力需建立异常自动诊断、问题自动修复或辅助修复的闭环流程，针对算力波动、模型漂移、接口异常等常见问题，实现自动化定位与处置，最大限度减少人工干预，提升系统运行稳定性。自愈流程的设计应遵循分级响应原则：一级异常（服务不可用）触发自动切换与流量迁移；二级异常（性能降级）触发弹性扩容与路由调整；三级异常（效果漂移）触发持续学习流程并推送人工复核。所有自愈动作均需记录审计日志，确保自动化行为可追溯、可解释。

4.6 AI 安全合规与伦理风险运营

AI 原生系统引入了传统信息安全之外的新型风险，需建立常态化的安全合规运营机制。合规管理方面，跟踪落实算法备案、生成式人工智能服务管理等监管要求，建立上线前合规审查与上线后持续监督流程。内容安全方面，建立输入输出双向的内容安全策略，覆盖提示注入与越狱攻击防护、违规内容拦截、幻觉检测与标注。数据安全与隐私方面，严格训练数据与用户数据的来源合法性审查，落实个人信息保护要求，对敏感数据实施不出域的技术管控。伦理与责任方面，明确 AI 决策的适用边界与人工复核机制，建立责任界定规则与伦理审查流程，对自动化决策全程留痕。

4.7 跨域能力共享与生态共建

生态共建方面明确 AI 能力在各业务域之间的共享复用机制，建立统一的能力注册、发现、调用与治理框架，避免重复建设、碎片化发展，实现从单点建设向生态化协同的转变。并通过能力中台的模式沉淀通用 AI 能力，实现能力的一次建设、多处复用，提升整体建设效率与资源利用率。同时建立外部合作伙伴的接入规范与协作流程，整合产业生态能力，补充能力短板。

资料整理
火龙果科技 荐

第 5 章 企业 AI 原生能力成熟度评估

5.1 多维度评估指标体系

科学的评估体系是衡量 AI 原生建设成效、指引优化方向的核心工具。系统梳理企业运营智能化的核心能力维度，可构建涵盖技术支撑力、数据驱动力、业务融合度、治理管控力、价值实现效果五大维度的评估指标体系。

技术支撑力维度评估算力资源供给调度、多源异构数据处理、AI 模型部署运维、系统高可用与弹性扩展等能力，明确各能力项的关键评估要点与衡量标准。

数据驱动力维度评估数据的完整性、准确性、及时性，数据的跨域流通与共享程度，数据转化为知识资产的成熟度，以及数据驱动决策的覆盖范围与深度。

业务融合度维度评估 AI 对业务流程的渗透率、AI 替代人工完成业务活动的比例、业务决策中 AI 的参与程度，并针对不同业务分别明确差异化指标。

治理管控力维度评估模型全生命周期治理、AI 安全合规管控、智能体权限与行为治理、AI 技术标准与制度体系建设等能力，为 AI 原生能力的安全合规运行划定边界与规则。

价值实现效果维度涵盖效率提升、质量改善、成本节约、风险控制、客户满意度等可量化价值指标，以及战略价值、创新价值等定性维度，形成定性与定量相结合的价值评估框架。

基于上述维度，形成涵盖各能力域的运营智能化核心指标库，明确每项指标的计算公式、数据来源、采集频率与评价标准。

评估维度	核心评估内容	代表性指标
技术支撑力	算力供给调度、数据处理、模型部署运维、高可用与弹性扩展	算力利用率、资源供给时效、服务可用率
数据驱动力	数据质量、跨域流通、知识资产化、数据驱动决策	数据完整性比率、跨域共享覆盖度、知识图谱覆盖率
业务融合度	流程渗透、人机替代、决策参与	AI 流程渗透率、AI 替代人工比例、决策辅助采纳率
治理管控力	模型全生命周期治理、AI 安全合规、智能体行为管控、标准制度体系	模型版本合规率、输出内容合规率、Prompt 注入防护成功率、智能体权限管控覆盖率
价值实现效果	效率、质量、成本、风险、满意度与战略创新价值	处理时长变化、差错率变化、单位成本变化

资料整理
火龙果科技 荐

5.2 成熟度分级模型

在标准指标体系基础上，可将运营智能化水平划分为逐级递进的成熟度等级，体现从初始级、发展到优化级的递进逻辑。低等级侧重基础能力建立与单点验证，中等级侧重系统化建设与跨场景推广，高等级侧重全面融合、持续优化与价值创造。各等级需具备明确的整体特征描述，使企业能够快速判断自身所处的大致阶段。在各项核心指标上设定量化阈值或定性特征要求。



图 5-1 企业 AI 原生能力成熟度五级模型

成熟度等级	整体特征	能力建设重点
一级·初始级	单点探索阶段，AI 应用以试点为主	基础能力建立与单点验证，积累数据与经验
二级·成长级	系统建设阶段，AI 能力开始嵌入核心业务流程	标准化建设，打通数据与模型基础链路
三级·系统级	体系化阶段，AI 能力覆盖多个业务域并平台化沉淀	跨场景推广，能力复用与中台化
四级·融合级	深度融合阶段，AI 融入业务流程与决策闭环	业务流程重构，人机协同运行
五级·优化级	价值创造阶段，AI 能力持续自演进并驱动模式创新	持续优化、自学习闭环、生态协同

量化阈值适用于 AI 渗透率、数据处理效率、模型推理准确率等可精确测量的指标，需设定各等级的具体数值区间；定性特征适用于组织协同程度、创新能力水平等难以精确量化的指标，需通过行为锚定或特征描述的方式明确各等级的典型表现。

资料整理
火龙果科技 荐

此外，模型还需明确等级提升的路径与条件，即各等级跃迁需具备的关键能力与前置条件，以及成熟度评估的实施方法与操作流程，包括数据采集方式、结果计算方法、报告生成规范等，确保评估具备可操作性与可重复性。

5.3 试点评估实施方法

模型设计阶段的合理性需在真实业务场景中验证，指标体系的适用性需在具体评估实践中检验。脱离真实场景的桌面推演难以发现指标定义、数据采集、等级判定等方面的潜在问题，这些缺陷只有在实际评估中才会暴露。

开展试点评估具有多重不可替代的价值。模型验证方面，通过选取典型场景开展实际评估，可检验成熟度模型的适用性与可操作性，发现指标定义不清晰、数据采集不可得、等级判定不合理之处，为模型优化提供实践反馈。差距识别方面，通过深入评估可精准定位各场景的实际瓶颈与技术短板，明确当前所处等级，为架构设计优化提供精准输入。标杆示范方面，试点评估形成的报告与建议可为后续规模化推广提供方法参照，为各业务领域智能化建设提供标杆对比。

首先，选择典型业务场景。基于成熟度模型，选取典型业务场景开展评估试点。场景选取需要遵循代表性原则，所选场景应能反映智能化建设的主要特征与关键挑战；差异性原则，所选场景的业务特征、数据基础、技术条件差异显著，可全面检验模型在不同类型业务环境中的适用性；实施条件原则，所选场景应具备必要的评估数据基础与配合条件。确定试点场景后，依据成熟度评估模型的操作流程，系统开展评估数据的采集、分析与判断工作，确保评估过程遵循模型规范、评估结论具备客观基础。

其次，开展能力对标与差距的系统分析。在完成各场景成熟度等级评定基础上，对各场景在各项核心指标上的表现进行深入分析。通过将指标表现与成熟度模型各等级的标准要求逐项对标，明确各场景在技术支撑力、数据驱动力、业务融合度、治理管控力、价值实现效果维度的具体差距。差距分析需形成清晰的能力短板清单与根因判断，明确每个差距的性质、严重程度与制约影响。

最后，形成评估报告输出与改进建议提出。基于差距分析结论，输出场景分级评估报告，并提出针对性的能力提升建议与实施路径。提升建议需针对具体差距明确解决方向与关键举措，区分短期可快速见效的改进项与中长期需系统性建设的能力提升项。实施路径需明确优先级顺序、阶段划分与里程碑节点。评估报告的结构化输出需确保评估结论能够为后续架构优化与能力建设提供精准输入。

资料整理
火龙果科技 荐

第 6 章 传统架构向 AI 原生演进路径

6.1 总体演进框架

AI 原生架构升级不同于单点技术替换或局部系统改造，其本质是企业技术体系从“传统数字化”向“AI 原生”的能力跃迁，涉及架构形态、运行模式、技术栈、组织能力的系统性重构。这一跃迁不能脱离企业现有基础盲目推进，需建立在精准的现状诊断之上，并遵循"诊断—设计—实施—风险管控"的科学路径。



图 6-1 传统架构向 AI 原生演进路径

6.2 现状诊断与差距识别方法

AI 原生架构升级不能脱离企业现有基础盲目推进，需建立在精准的现状诊断之上。通过对现有系统的全面评估，识别能力短板与技术债务，才能明确升级的起点与方向。现状诊断包含业务侧调研与技术侧评估两个视角：现状调研从业务侧发现问题，能力评估则需要从技术侧诊断根因，两者结合才能形成从现象到本质的完整认知链条。

6.2.1 业务侧 AI 应用现状调研方法

从各行业企业的普遍实践来看，AI 应用正逐步渗透到生产运营、经营管理、客户服务等各类业务场景，其中客服与财务是当前 AI 落地实践较为集中、应用探索相对活跃的两大业务领域。一方面，客服作为直面客户的前端核心触点，服务诉求量大、重复度高，降本提质的业务诉求迫切，叠加全链路交互数据积累充分，成为多数企业最早试水 AI 的场景；另一方面，财务

资料整理
火龙果科技 荐

作为企业核心管控环节，业务规则标准化程度高、数据结构化基础好，AI 应用的效果易衡量、风险可控，同样成为 AI 实践落地的重点领域。目前客服领域围绕客户服务智能化、精准营销、客户需求预测、智能问答等场景开展了 AI 应用探索，但各场景落地深度与技术成熟度参差不齐，整体呈现碎片化试点特征。财务领域在智能报销审核、财务风险预警、智能报表分析等方面逐步引入 AI 能力，但受限于业务标准化程度和数据治理基础，整体系统性与规模化水平仍有较大提升空间。这两大业务领域的应用形态与现存问题，是当前企业 AI 落地整体阶段的典型缩影，其所反映的碎片化部署、基础支撑不足、能力不成体系等问题，同样普遍存在于其他业务场景中。如不经系统的全业务域调研评估，后续架构设计将在目标设定、差距识别、需求对接、资源投放等方面缺乏依据。

围绕核心业务开展 AI 应用现状调研，系统摸排已落地场景、应用成效、用户反馈与运维保障情况。调研维度包括落地覆盖范围、应用成效、使用反馈、运维保障四个方面，全面掌握业务侧的真实诉求与现存痛点。

落地覆盖范围方面，梳理各业务域已开展的 AI 应用场景清单，明确每项 AI 技术的类型、建设时间、部署方式与覆盖范围，识别已覆盖与尚未覆盖的业务环节及各场景间的衔接关系，形成 AI 应用场景全景图。

应用成效方面，收集各场景的效率提升、质量改善、成本节约等可量化指标，结合业务人员的主观评价与满意度反馈，综合判断各场景的实际价值，识别效果达预期与未达预期的场景及原因。

使用反馈方面，了解一线人员对 AI 工具的易用性、响应速度、结果准确性、流程适配程度及异常处理能力的体验，识别功能缺陷与改进期望，确保架构设计能回应使用者层面的真实诉求。

运维保障方面，调研各 AI 应用的运维模式、资源投入、问题响应机制、模型更新频率、效果监控手段与数据回流机制，评估运维体系能否支撑 AI 应用的持续稳定运行与效果优化。

在调研信息收集的基础上，研究需结合业务发展需求，系统梳理当前 AI 应用落地过程中的业务痛点、需求堵点与能力缺口。

在业务痛点方面，需识别业务人员在现有工作流程中遇到的效率瓶颈、质量难题与决策困境，分析哪些痛点有望通过 AI 技术得到有效缓解或解决，明确各痛点的优先级与解决紧迫程度。

资料整理
火龙果科技 荐

在需求堵点方面，需识别业务部门已提出但尚未得到有效响应的 AI 需求，分析需求未能落地的具体原因，包括技术条件不成熟、数据基础不具备、投入产出不匹配、组织流程不支撑等，明确不同类型堵点的分布特征与解决路径。

在能力缺口方面，需系统分析现有 AI 应用能力与业务发展需求之间的差距，明确哪些能力需要从无到有建设、哪些能力需要从有到优提升、哪些能力需要从单点向规模化扩展，形成结构清晰的能力缺口分析结论。

最终，将分析结论形成按业务领域、问题类型、严重程度、紧急程度分类标注的结构化需求清单，确保调研结论能够被后续评估与设计工作有效承接使用，避免调研与后续环节之间的信息断层。

6.2.2 技术侧架构能力评估

上述现状业务侧调研揭示了业务层面“发生了什么”，但 AI 原生架构设计需要的是技术层面的系统评估，明确现有系统在应用、数据、技术架构层面对 AI 原生的支撑能力处于什么水平、存在哪些技术短板、与标准要求之间存在多大差距。业务调研发现问题，能力评估需从技术侧诊断根因，两者结合才能形成从现象到本质的完整认知。

业务调研中发现的 AI 应用成效不佳、需求响应缓慢、系统协同困难等问题，其根源往往不在业务层面，而在于底层架构对 AI 原生的支撑能力不足。某 AI 场景推进困难，业务归因可能是数据获取不便，但技术根因可能是数据架构存在孤岛或应用架构耦合度过高。某模型上线后效果衰减快，业务感知是工具不好用，但技术根因可能是缺乏模型持续迭代的架构闭环。如仅停留在业务问题描述而不深入架构根因诊断，后续优化方向将缺乏准确的技术判断依据。

围绕架构根因诊断，聚焦 AI 原生架构能力评估分析，具体包括以下几个方面。第一，系统耦合度评估，分析现有系统间的依赖关系与集成方式，识别制约 AI 灵活接入的瓶颈。第二，数据孤岛评估，分析各系统间数据共享流通现状，识别制约 AI 训练数据供给的障碍。第三，算力调度评估，分析现有技术架构对异构算力的管理调度能力。第四，模型开发运维评估，分析是否具备模型持续集成、持续交付与持续监控能力。第五，安全机制评估，分析现有安全架构对 AI 新型安全风险的覆盖程度。

在此基础上，需针对核心系统建立体系化的技术侧架构能力评估框架，形成专项评估维度和指标体系，将上述五大方向落地为可量化、可比对的具体评判标准，针对核心系统的建立技术侧架构能力评估框架，形成专项评估维度和指标体系。技术侧架构能力评估框架应该覆盖应用架构、数据架构、技术架构三个层面，每个层面结合 AI 原生特性要求细化具体评估维度与评判标准。

资料整理
火龙果科技 荐

应用架构层面包括模块化与解耦程度、API 接口标准化水平、业务流程弹性编排能力、智能化功能集成方式与可扩展性。

数据架构层面包括多源数据汇聚整合能力、非结构化数据处理能力、数据质量保障水平、数据服务对 AI 训练推理的支撑能力。

技术架构层面包括异构算力资源调度能力、模型服务部署管理水平、平台组件标准化与可扩展性、高可用与弹性扩展能力。

安全架构层面覆盖 AI 原生安全特性，包括数据安全、模型安全保障及 API 安全管控等。评估维度设计需确保系统覆盖标准框架各项要求，使评估结论具备充分的覆盖面与可量化性。

6.2.3 全维度评估实施

在评估框架确定的基础上，对相关核心系统开展逐项专项评估。评估采用文档审查、系统演示、数据采集、技术访谈等多种方式，确保每项评估结论具备充分的事实依据。

系统耦合度方面，评估现有系统间的依赖关系与集成方式，识别因耦合度过高制约 AI 灵活接入的瓶颈，分析松耦合改造的可行性与复杂度。

数据孤岛方面，评估各系统间数据共享流通现状，识别因数据分散且缺乏有效整合而制约 AI 训练数据供给的障碍，分析数据贯通所需的调整方向。

算力调度方面，评估现有技术架构对异构算力的管理调度能力，识别因算力供给不协同导致推理延迟或训练效率低下的问题。

模型开发运维方面，评估是否具备模型持续集成、持续交付与持续监控能力，识别训练环境与部署环境的割裂问题。

安全机制方面，评估现有安全架构对 AI 新型安全风险的覆盖程度，识别数据隐私保护、模型安全保障、内容合规审查等方面的薄弱环节。评估过程中还需发现其他可能存在的短板，确保问题识别的全面性。

6.2.4 差距分析与提升方向

基于上述评估结果，将各维度现状水平与 AI 原生技术标准框架要求逐项对照，形成清晰的差距分析结论，明确每个维度上现有系统所处能力等级、与标准要求的差距大小及对 AI 能力建设的制约程度。在此基础上提出重点提升方向的优先级排序，综合考虑差距严重程度、对 AI 落地的制约程度及实施可行性与投入，严重制约且相对易解决的短板优先提升，差距存在但不影响核心场景推进的纳入中长期规划。

资料整理
火龙果科技 荐

6.3 升级方案设计

6.3.1 升级方案设计的前置基础

技术架构升级是复杂的系统工程，涉及技术路线调整、平台能力重构、系统集成改造、组织能力转型等多个维度。若无系统的升级方案，可能面临技术风险、资源风险、业务风险和组织风险等多重挑战。升级方案设计首先需要解决的，是如何将分散于各专项研究的成果系统整合为统一、完整、可直接使用的输入基础，为升级方案设计提供结构化的认知框架。

在前序内容中，AI 原生核心技术及能力图谱明确了关键技术的能力特征与标准要求，企业级 AI 原生架构参考范式构建了目标架构的形态蓝图，AI 原生系统全生命周期运营管理形成长效运营的管理机制，能力成熟度评估建立了架构成效的评估基准。上述研究从不同维度刻画了 AI 原生技术架构的目标状态与建设要求，共同构成了指导技术架构升级的完整认知体系，升级方案的设计需要从结合上述研究成果，进行成果整合，最终形成整体的升级方案。

成果整合可按三个环节推进。第一环，前序成果关键要素的系统提取与结构化归集，将分散在不同成果中的关键信息进行结构化整理：关键技术层面，提取智能体、AI 网关、向量数据库、知识图谱等关键技术的核心能力特征、成熟度判断、演进趋势及标准规范要求，形成技术要素清单；架构设计方法层面，提取设计方法论中的核心原则、关键设计活动、决策要点及各层面设计规范，形成方法要素清单；目标架构框架层面，提取总体框架中各层级功能定位、关键组件定义、交互关系与接口规范，形成架构要素清单；运营智能化评估体系层面，提取核心评估维度、关键指标定义、成熟度等级划分标准与评估方法，形成评估要素清单。

第二环，要素关联逻辑的梳理与 AI 原生核心能力矩阵的构建：以目标架构框架的层级结构为骨架，将各层级所需核心能力进行逐一定义与整合，基础设施层整合算力调度、云边端协同、资源弹性扩展等能力要求，数据智能层整合多源数据治理、知识图谱构建、非结构化数据处理、高质量数据集管理等能力要求，AI 能力开放层整合模型服务管理、智能体运行框架、AI 网关治理等能力要求，应用服务层整合 AI 原生应用开发框架、智能体与传统应用集成等能力要求，统一治理体系整合安全保障、运维运营、标准规范等能力要求；能力矩阵需为每项能力明确其技术支撑组件、设计方法依据及对应评估指标，使各前序成果在统一框架下形成关联呼应。

第三环，架构升级基线要求、目标标准与核心约束的明确：基线要求明确从当前架构状态出发启动升级的前提条件，包括现有技术基础、资源条件与组织准备。目标标准基于目标架构框架和能力矩阵，明确各项核心能力在升级完成后应达到的标准水平，形成可检验的目标参照。核心约束识别升级过程中需遵循的技术约束、安全约束、业务约束与资源约束，包括与现有系统的兼容性要求、安全合规红线、业务连续性保障及资源投入合理边界。

资料整理
火龙果科技 荐

6.3.2 升级方案的核心构成

升级方案应包含技术路线调整、平台能力重构、系统集成改造、组织能力转型四个核心组成部分，每个部分明确设计要点，确保方案的系统性与可实施性。

技术路线调整方案方面，针对现状评估识别的技术短板与升级目标要求，制定由现状技术路线向 AI 原生技术路线的迁移路径。设计要点包括：技术选型决策依据（如智能体框架、AI 网关、向量数据库等关键技术选型准则）、新旧技术路线过渡策略（如灰度切换、双轨运行）、技术债务清理计划（如废弃组件梳理与下架节奏）、技术规范更新要求（如接口规范、数据规范、安全规范的同步升级）。

平台能力重构方案方面，针对 AI 原生核心能力矩阵中各项能力的建设要求，设计平台层能力的建设蓝图。设计要点包括：核心能力建设优先级排序（依据差距分析结论）、平台组件化与标准化设计（确保能力可复用、可组合）、平台运营机制设计（如模型管理、效果监控、版本管理流程）、平台与上层应用及底层基础设施的接口设计。

系统集成改造方案方面，针对存量系统与新建平台的集成需求，设计渐进式集成策略。设计要点包括：存量系统盘点与改造分级（保留、改造、替换三类清单）、集成接口设计（统一 API 网关、AI 网关的双层接入架构）、数据迁移与一致性保障方案、集成过程业务连续性保障措施（如灰度发布、回滚机制）。

组织能力转型方案方面，针对 AI 原生架构对组织能力的新要求，设计配套的组织转型路径。设计要点包括：人员能力提升计划（如 AI 架构师、智能体工程师等新岗位能力模型与培养路径）、跨部门协同机制设计（如平台团队与业务团队的协作流程）、变革管理策略（如沟通机制、培训机制、激励机制）、运维运营能力升级（如 AI 场景专属运维团队建设）。

6.3.3 AI 原生技术全景图构建

技术方案解决的是各项技术的具体实现问题，尚未从全局视角回答 AI 原生技术架构的整体形态应呈现为何种面貌。技术全景图是将分散的技术选型、融合路径、集成规范等成果整合为一张完整架构蓝图的顶层设计，其价值体现在：提供统一架构视图，将分散于各报告中的技术认知整合为完整蓝图，使各角色在同一视图下形成准确理解；揭示端到端协同逻辑，系统梳理数据采集到评估优化的全链路流程，使各环节依赖关系与数据流转清晰可见；呈现跨域联动机制，展现技术底座、方法论、评估体系在不同业务域的交叉联动关系；指导规模化推广，全景图是向实施团队传递技术方案的最有效载体，可大幅降低知识传递成本。

（1）全景图的层级结构

资料整理
火龙果科技 荐

全景图以四层架构为基本骨架。基础设施层呈现异构算力资源的构成与统一调度管理机制、云边端协同部署架构及与统一技术底座的对接关系。数据层呈现多源异构数据的采集、存储、治理与服务全链条，结构化与非结构化数据的统一管理机制，以及知识图谱、向量数据库等新型组件的架构定位。模型层呈现企业级大模型体系的接入方式、AI 能力中台的模型服务治理框架、模型训练推理与持续迭代的能力支撑体系，以及 AI 网关在多模型统一接入与路由中的定位。应用层呈现智能体在不同业务场景中的运行框架、AI 原生应用的架构范式及 AI 能力与业务应用的集成规范。四层结构需确保层级边界清晰、上下层调用关系明确、核心组件完整无遗漏。

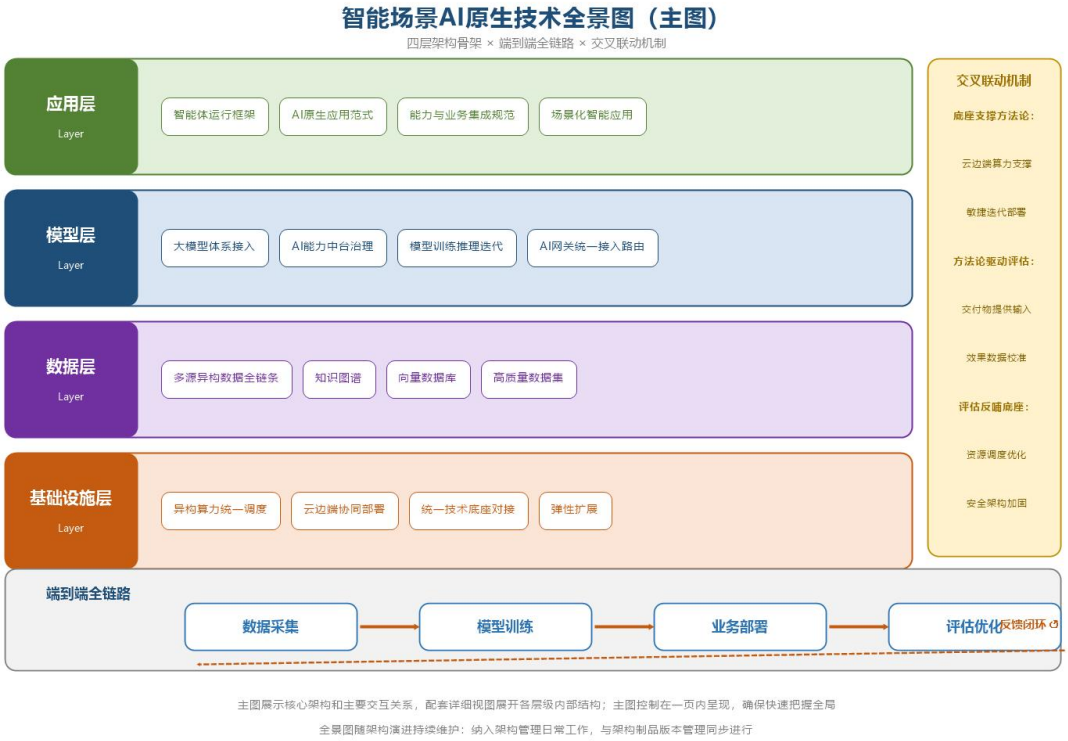


图 6-1 智能场景 AI 原生技术全景图

（2）端到端全链路流程

开展端到端全链路流程与技术支撑点的系统梳理，以业务价值流动为主线，系统梳理数据采集、模型训练、业务部署到评估优化的端到端全链路流程。数据采集环节标注多源异构数据的统一采集与标准化处理、数据标注、质量校验等技术支撑点。模型训练环节标注基础设施层的异构算力支撑、数据层的高质量数据集供给、模型层训练与生产环境的统一管理。业务部署环节标注 AI 能力通过中台和网关向业务应用标准化部署与开放，应用层实现 AI 能力与业务流程的深度融合。评估优化环节标注运营智能化评估体系对 AI 应用的效果度量，及反馈闭环如何反哺数据标注优化、模型版本迭代与应用体验改进。全链路需确保各环节的数据流、控制流与反馈流清晰呈现。

资料整理
火龙果科技 荐

（3）交叉联动机制

全景图需呈现技术底座、建设运营方法论与评估体系三者间的动态联动关系。技术底座与建设运营方法论的联动，呈现云边端协同算力如何支撑敏捷迭代部署、AI 能力中台如何支撑能力复用等。建设运营方法论与评估体系的联动，呈现各阶段交付物如何为评估提供输入数据，及运维运营阶段效果数据如何驱动评估校准。评估体系与技术底座的联动，呈现评估结果如何驱动资源调度策略优化、安全架构加固升级等。交叉联动机制的呈现需使阅读者能够理解架构各要素之间的动态协同关系，而非仅看到静态的组件堆叠。

6.4 渐进式升级策略与路径规划



图 6-2 传统架构向 AI 原生演进的三阶段路线

升级方案明确“做什么”，升级策略与路径规划解决的核心问题不是“目标架构是什么”，而是“从当前架构到目标架构应该按照什么路径、采取什么策略、依据什么节奏达成，以及如何应对升级过程中的风险”。

首先，与现有技术路线的系统性对比分析。针对现有数字化技术路线，涵盖企业架构、数据模型、基础设施等核心领域，与 AI 原生目标架构要求进行系统性对比。企业架构层面，分析现有应用、数据、技术架构的设计逻辑、技术选型与运行模式，与目标架构在模块化程度、接口标准化水平、弹性扩展能力、智能化支撑能力等方面的差距。数据模型层面，分析现有数据模型对结构化数据以外多模态数据的覆盖能力、数据治理体系对 AI 训练与推理场景的支撑程度、数据服务对实时性与高并发的满足能力等方面的不足。基础设施层面，分析现有算力资源

资料整理
火龙果科技 荐

的类型与规模、算力调度的灵活性与效率、基础设施对云边端协同部署的支持能力等方面的差距。

然后，基于现状与目标的差距分析，综合评估技术可行性、经济合理性与潜在风险，制定分阶段、分领域、分组件的渐进式升级策略。

综合评估技术可行性、经济合理性与潜在风险，制定分阶段、分领域、分组件的渐进式升级策略。技术可行性方面评估各项升级工作在现有技术条件下的可实现程度，区分可立即启动、需等待条件成熟、需分阶段推进等不同类型。经济合理性方面评估升级投入与预期收益之间的关系，确保升级方案的投入产出比合理。

基于上述评估，分阶段策略明确各阶段时间跨度、核心任务、里程碑与交付物；分领域策略针对不同业务域特征制定差异化方案，优先在痛点突出、价值明显的领域推进；分组件策略明确各组件的替换、改造或重建方式。在此基础上规划清晰的演进路线图，确保升级过程路径清晰、节奏可控。

6.5 升级风险识别与应对框架

架构升级过程中面临多维度风险，需建立系统性的风险应对机制，确保风险管控贯穿升级全过程。

技术风险关注新旧系统切换的数据一致性、新技术栈性能不确定性、接口兼容性等；业务风险关注核心业务流程连续性、用户习惯变化、高峰期窗口限制等；组织风险关注人员技能差距、跨部门协同难度、变革阻力等。

针对各类风险制定系统性应对策略，涵盖预防措施、应急响应与恢复措施三个层面，确保风险管控贯穿升级全过程。

风险类别	典型风险	应对策略要点
技术风险	新旧系统切换数据不一致、新技术栈性能不确定、接口不兼容	迁移校验机制、灰度切换、回退预案
业务风险	核心流程中断、用户习惯冲击、停机窗口受限	双轨并行、分批切换、业务影响评估
组织风险	人员技能差距、跨部门协同难、变革阻力	技能培训、专职团队、变革沟通机制

资料整理
火龙果科技 荐

第 7 章 AI 原生架构落地验证

架构设计的理论完备性不等于实际业务场景中的有效性，必须通过典型场景中开展技术底座、建设运营方法论与评估体系的适配性验证，识别架构方案在真实业务环境中的适配短板并制定补充策略，确保方案经得起实践检验。

场景验证的价值体现在三个方面：发现设计缺陷，识别架构设计阶段基于通用假设做出的判断在具体场景约束下的适用性不足；积累实践经验，为后续规模化推广提供可借鉴的实践范本；建立落地信心，经过验证的架构方案相比纯理论方案更具说服力。

7.1 场景验证的价值与选取方法

开展验证场景的选取与方案论证框架的建立，选取业务中处于不同成熟度的智能场景作为验证对象。场景选取遵循成熟度差异原则，兼顾已具备较好基础的成熟场景和尚处探索阶段的成长场景，全面检验方案对不同起点条件的适配能力；业务代表性原则，覆盖核心业务域的核心业务特征与关键智能化需求；实施可行性原则，具备必要的实施条件与配合意愿。确定场景后，结合业务特性建立涵盖技术底座、建设运营方法论及评估体系三个维度的验证框架，明确各维度验证要点、判断标准与评估方法。

制定明确的场景选择标准。宜选取两个以上场景，确保场景之间在业务特征、技术复杂度、成熟度上形成互补。以能源企业为例，系统调度场景侧重高并发、实时交互类 AI 应用，财务场景侧重高合规、数据处理类 AI 应用，通过差异化场景的交叉验证检验方案的通用性。验证结论需明确标注适用条件和范围边界，说明哪些结论可推广至其他业务域、哪些结论仅适用于特定场景类型。

验证维度	验证要点	判断标准
技术底座	云边端协同算力支撑能力、多模态数据实时处理能力、弹性扩展与延迟表现	峰值负载下满足性能与服务质量要求
建设运营方法论	跨域业务协同有效性、高频迭代优化适用性	能指导数据共享、能力复用与流程改造落地
评估体系	安全合规覆盖度、效率效能指标有效性	关键要求具备可操作、可衡量的指标

7.2 三维度适配性验证实施

资料整理
火龙果科技 荐

确定验证场景后，结合业务特性建立涵盖技术底座、建设运营方法论及评估体系三个维度的验证框架，明确各维度验证要点、判断标准与评估方法。

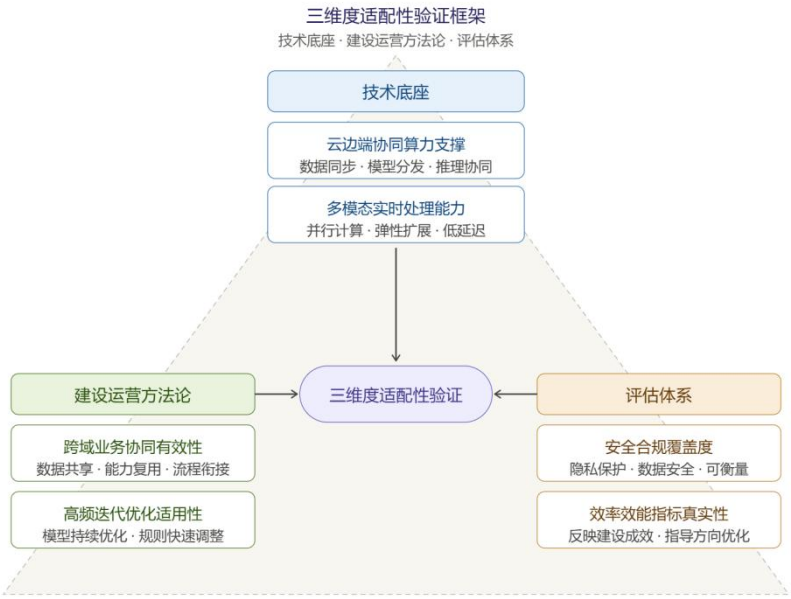


图 7-1 AI 原生架构落地三维度适配性验证框架

技术底座验证维度，评估云边端协同算力对网点分散、终端多样部署需求、集中化与分布式协同调度需求的支撑能力，验证云边端数据同步、模型分发与推理协同机制的有效性；评估多模态数据实时处理对语音交互、图像识别、票据影像、结构化报表等并行处理与实时计算的能力，验证数据链路在流量突增时的弹性扩展与延迟表现。

建设运营方法论验证维度，评估方法论在跨域业务协同方面的有效性，验证其能否指导数据共享、能力复用、流程衔接的智能化改造；评估在高频迭代优化方面的适用性，验证其能否适应场景模型效果持续优化及规则快速调整的响应需求。

评估体系验证维度，检验评估体系对安全合规的覆盖度，验证用户隐私保护、数据安全合规等关键要求是否具备可操作、可衡量的指标；检验效率效能维度指标能否真实反映建设成效，评估结果能否有效指导建设方向优化。

7.3 适配短板识别与优化策略制定

基于验证结果，系统识别方案在具体场景中暴露的适配短板，涵盖技术底座支撑不足、方法论应对复杂性局限、评估体系价值衡量缺失等方面。分析每项短板的根本原因，判断是对场景特殊需求考虑不足、技术选型偏差、方法适用边界不清还是评估维度遗漏。据此制定针对性

资料整理
火龙果科技 荐

补充策略，包括技术底座配置调整、方法论适用边界明确与补充指引、评估体系维度扩展与指标优化，确保方案经场景验证与优化调整后能有效适配实际需求。

7.4 关键技术与现有体系融合路径

适配性验证回答了“是否适用”及“哪些方面需要调整”，但尚未回答在技术实现层面如何将关键技术组件与公司现有技术体系深度融合。AI 原生架构落地是在公司现有技术体系基础上的演进式升级，在适配性验证确认了架构方案可行性的前提下，需要考虑各项关键技术组件如何与公司现有技术体系进行深度融合，目前较多大型企业均已建成大模型体系、统一技术底座、企业级 AI 能力中台等技术基础设施，这些现有体系构成了架构升级不可忽视的现实约束。在验证确认方案可行性的前提下，需进一步回答：智能体、AI 网关、向量数据库等组件如何与企业级模型体系有效对接，如何与统一技术底座在算力调度、数据管理、服务治理等层面深度融合，如何与企业级 AI 能力中台在能力注册、服务编排、运营治理等环节形成分工协作，需从以下三个方面考虑与当前技术架构的融合：

一是关键技术组件技术选型确认与细化。智能体方面，明确适用于业务场景的智能体框架选型建议，涵盖定义规范、编排方式、运行框架、协同机制及与 AI 能力中台的集成模式。AI 网关方面，明确功能定位、部署方式、性能要求与选型标准。向量数据库方面，明确技术选型建议，涵盖索引算法选择、部署架构设计、性能容量规划。选型需综合考虑验证中发现的技术适配性问题、性能瓶颈与优化方向，确保结论具备充分验证依据。

二是关键技术组件之间的深度融合设计。设计各技术组件间的深度融合方案。智能体与 AI 网关融合需明确模型服务调用标准化流程、多模型动态路由机制、输出结果统一治理规范及网关在智能体通信协调中的定位。AI 网关与向量数据库融合需明确查询请求通过向量数据库完成知识检索与上下文增强的流程，检索结果返回的标准化格式及数据流性能优化策略。智能体与向量数据库融合需明确智能体利用向量数据库进行长期记忆存储检索、通过向量检索增强知识覆盖与推理准确性的机制。各组件交互需在公司统一技术底座服务治理框架下规范运行。

三是与公司现有技术体系的融合路径与集成规范设计。与已有模型体系融合，明确 AI 网关与模型体系的对接方式，包括模型服务接入标准、调用协议、接口规范与鉴权机制。与统一技术底座融合，明确对算力调度、存储服务、网络通信、安全管控的集成要求，特别是异构算力资源向底座注册与调度的标准方式。与企业级 AI 能力中台融合，明确各类 AI 能力向中台注册、发布、共享与治理的规范，以及中台对多域 AI 能力调用的统一路由与编排机制。与业务应用集成，明确同步与异步调用的适用场景与规范、业务上下文传递标准、返回结果解析与展示规范。

资料整理
火龙果科技 荐

任何企业推进 AI 原生架构升级，都需先盘点既有技术资产，再设计新组件与既有体系的融合方式与集成规范，避免“新体系孤岛化、旧体系边缘化”的双重陷阱。

资料整理
火龙果科技 荐

第 8 章 未来展望与实施建议

8.1 未来技术演进趋势展望

未来三到五年，AI 原生技术将持续快速演进，智能体技术将从单智能体向多智能体协同生态发展，模型推理成本持续下降，多模态融合能力持续增强。架构层面将进一步向解耦化、服务化、弹性化方向演进，标准化程度持续提升。

从技术侧看，四个趋势值得关注。一是多智能体系统走向生产化运行，智能体间协作协议、记忆共享机制与行为治理规范将成为企业架构的新组件；二是模型服务的成本治理走向精细化，以 token 为计量单位的算力经济将倒逼 AI 网关的计费、配额与调度能力持续增强；三是端侧与边缘智能加速落地，大模型蒸馏、量化与云边协同推理使 AI 能力延伸至业务末梢；四是 RAG 与知识工程的深度融合，企业知识资产化将成为数据架构建设的主线，知识图谱与向量检索双轮驱动的知识服务体系逐步成型。

从组织侧看，企业 IT 部门的角色将从“系统建设者”转向“能力运营者”，AI 能力的规划、治理、度量与复用成为新的核心职能；业务部门与 IT 部门的协作界面将围绕“智能体编排”重构，业务人员通过低代码方式定义智能体将成为常态；同时，AI 治理组织（如模型风险委员会、数据与 AI 治理办公室）将成为企业治理架构的标配。

对于企业而言，AI 原生架构将从可选的创新探索变为必备的基础能力，成为企业数字化基础设施的核心组成部分。架构的建设重心也将从技术可行性验证，转向规模化落地、价值化运营、生态化协同。

8.2 企业推进实施建议

对于计划推进 AI 原生架构建设的企业，建议遵循“规划先行、底座优先、试点验证、逐步推广”的实施路径。

顶层规划先行：先明确整体架构框架与标准体系，避免各自为政、重复建设，形成统一的技术路线与建设规范。

公共底座优先：优先建设算力、数据、能力中台等公共基础能力，形成共享能力底座，支撑各业务域快速应用。

典型场景试点：选取代表性场景开展验证，检验架构方案的可行性，沉淀最佳实践后再规模化推广。

资料整理
火龙果科技 荐

持续迭代优化：建立持续运营、持续评估、持续优化的闭环机制，实现架构能力的持续演进与价值提升。

在实施节奏上，建议把握三个要点。一是价值锚定：每一阶段的建设都以可度量的业务价值为锚点，避免“为架构而架构”的技术自嗨；二是风险预算：将安全合规、业务连续性的评估前置到方案设计阶段，为新技术引入预留验证周期；三是能力沉淀：将场景验证中形成的方法、模板、组件及时纳入能力中台与标准体系，确保先行经验可复制、可推广。

8.3 结语

AI 原生架构不是一次性的技术改造，而是企业面向智能时代的系统性能力重构。它要求企业在认知上完成从“AI 作为功能”到“AI 作为架构第一性原理”的转变，在方法上打通从场景识别、架构设计、持续运营到价值度量的完整闭环，在组织上建立与之匹配的协作机制与治理体系。

本报告提出的四层架构范式、全生命周期运营机制、成熟度评估体系、渐进式演进路径与场景验证方法，构成了一个相互衔接、可分步实施的完整方法论体系。企业可以据此校准自身所处的成熟度阶段，选择适配的演进策略，在可控风险下稳步走向 AI 原生。展望未来，人工智能技术的演进仍将持续加速，但“以架构确定性应对技术不确定性”的基本逻辑不会改变——这正是 AI 原生架构研究的价值所在。

资料整理
火龙果科技 荐

参考文献

[1] Gartner. AI-Native Architecture: The Foundation for Enterprise AI Scaling [R]. 2025.

[2] 华为、中国信息通信研究院.云原生 AI 技术架构白皮书[R]. 2024.

[3] 阿里云.AI 原生应用架构白皮书

[4] 中国信息通信研究院.人工智能研发运营体系（MLOps）实践指南 [R]. 2023.

[5] 中科算网算泥社区、中国工业互联网研究院.智能体技术发展报告 [R]. 2026.

[6] ISO/IEC JTC 1/SC 42. Artificial intelligence systems overview [R]. 2024.

[7] ISO/IEC 42001. Artificial intelligence management system [S]. 2023.

[8] NIST. AI Risk Management Framework (AI RMF 1.0) [R]. 2023.

[9] 国务院.关于深入实施 “人工智能 +” 行动的意见 [Z]. 2025.

[10] 中国信息通信研究院.中小企业数字化转型升级路径指南 [R]. 2025.

[11] 全球计算联盟.2025 向量数据库白皮书 [R]. 2025.

[12] 中国中文信息学会.知识图谱发展报告（2022）[R]. 2022.

资料整理
火龙果科技 荐