HUAWEI

企业如何构建面向智能体的混合云

编制委员会

PREPARATION COMMITTEE

主 编 单 位 华为云计算技术有限公司

编 委 顾 问 周跃峰 顾雪军 方 璞

编审组成员 涂 妍 冯吉龙 李亚为 张贝贝

主 编 人 员 郝利鹏 黄 智 李 胤 李 晓 汪 灿 柳一鑫
孙俊杰 张金东 柏 君 王 帆 王梦晴 熊洪槐
许 静 王 立 沈放之 余善英 陈 勇 冯楚灏
王艺洁 翟 叙

责 任 编 辑 王 瑞 石勇龙 秦 越 房敏娟 冯 茜 任姜静

（排名不分先后）

为智能体做好准备
企业CIO需要了解和行动的7件事

01

选择混合云架构

坚定走向多云协同架构，支持跨云MaaS调用，以最优TCO应对硬件与AI创新快速迭代

02

建立统一AI多模数据湖

AI多模数据湖是构筑企业级智能体的数据基石，为AI提供高质量企业私域数据集，并推动企业数据体系迈向支撑自主决策的新阶段

03

采用“云下稳态+云上敏态”的模型部署策略

本地化部署如气象等行业稳态大模型，针对迭代迅速的通用大模型，依托MaaS服务云端调用+机密计算技术，确保推理数据隐私与安全

04

构筑智能体安全开发与运行环境

建立一套涵盖安全开发、合规测试及动态监控的标准化运行环境，确保可记录、可追溯、可解释、可治理，为智能体规模落地提供坚实安全屏障

05

打破AI应用孤岛

建设统一的智能体平台，打破AI应用的“孤岛”与“黑箱”，最大化企业资产价值

06

用AI重塑运维能力

打造人+智能体协同的智能云管平台，构建企业级安全合规、极高可用、极优体验运维能力

07

转向精细化智能运营 option

从传统面向资源的成本运营模式，走向面向Token与智能体的效果驱动运营模式

目录 CONTENTS

01

企业面向智能体的建设趋势与混合云架构

1.1 趋势与挑战

发展趋势

智能体落地指数式爆发

企业AI进生产，正从以“模型能力”为中心，转向以“业务任务与流程”为中心的智能体驱动，并逐步由单智能体应用往多智能体协同快速演进

智能体普及加速 AI-Ready数据架构落地

随着AI-Ready数据架构成为标配，企业数据全面迈入多模态时代，数据应用从人工分析到智能体自主决策，数据管理升级为知识本体化治理

伴随智能体记忆系统的迭代完善，数据智能体正向更高阶智能形态演进

安全合规成为企业AI进生产的前置底线

Agent从个人应用向企业应用全面转型，选型已从功能导向转向安全信任导向

企业面临的问题和挑战

成本弹性缺失，全域通算智算资源统筹乏力

- 企业纯自建传统云化架构，前期固定资产投入（CAPEX）巨大，后期难以支撑向灵活运营支出（OPEX）的“自建 + 云上租用”的混合云形态平滑演进
- 缺失通算、智算与多云资源的统一调度能力，无法承接大模型、海量智能体的突发负载

数据孤岛严重，缺乏统一AI多模数据湖

- Agentic AI时代从面向传统BI报表与分析，转向面向大模型训练、增量微调、检索增强生成以及智能体调用推理协同
- 传统湖仓仅能处理结构化数据、无论数据形态、处理时效、算力架构还是扩展能力，均无法适配AI全流程的数据供给要求

缺乏灵活安全的云上云下MaaS协同能力

- 模型参数与能力日新月异，由于缺乏成熟的云上云下协同机制与机密计算能力，企业往往只能在“封闭的本地安全”与“开放的云端风险”之间艰难抉择
- 企业应考虑采用“云下稳态+云上敏态”的模型部署策略，本地化部署气象等行业稳态大模型，针对迭代迅速的通用大模型，依托MaaS服务云端部署+机密计算技术，确保推理数据隐私与安全

如何打造智能体的可信开发与运行环境

- 智能体催生攻击面放大、链式传导、追责难等新型安全挑战，安全能力已成为准入门槛
- 如何建立一套涵盖安全开发、合规测试及动态监控的标准化运行环境，为智能体的规模化落地提供坚实的安全屏障，确保可记录、可追溯、可解释、可治理

1.2 企业面向智能体的混合云落地模式与架构

在本地数据中心，混合云正在成为企业实现智能体私有化部署的最佳架构。它可以帮助政企客户在本地建立完整的智能体开发、训练与运行环境，实现数据不出域、安全可控以及核心能力自主掌握。在智能体落地过程中，我们认为企业会逐渐形成两种模式。

第一种，完全本地化模式，包括本地部署大模型，以及智能体开发态、运行态全部自建。这种方式安全性最高，但同时对企业的资金投入、研发能力和持续运营能力要求也非常高。

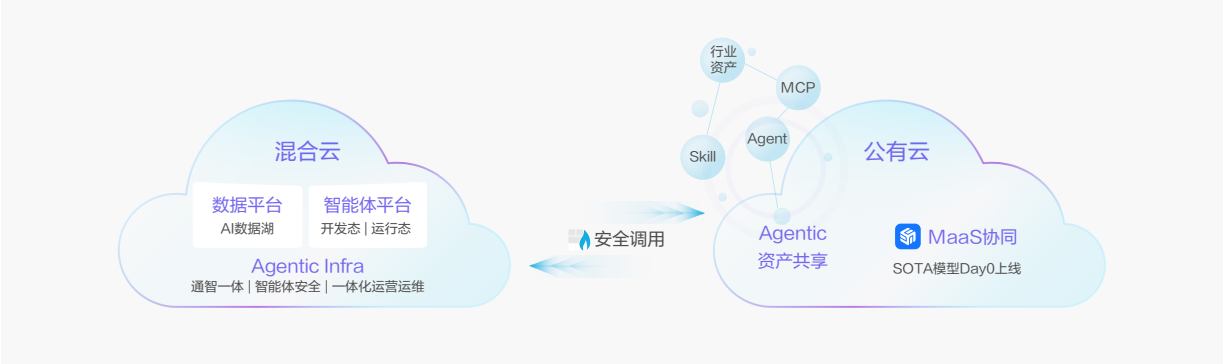
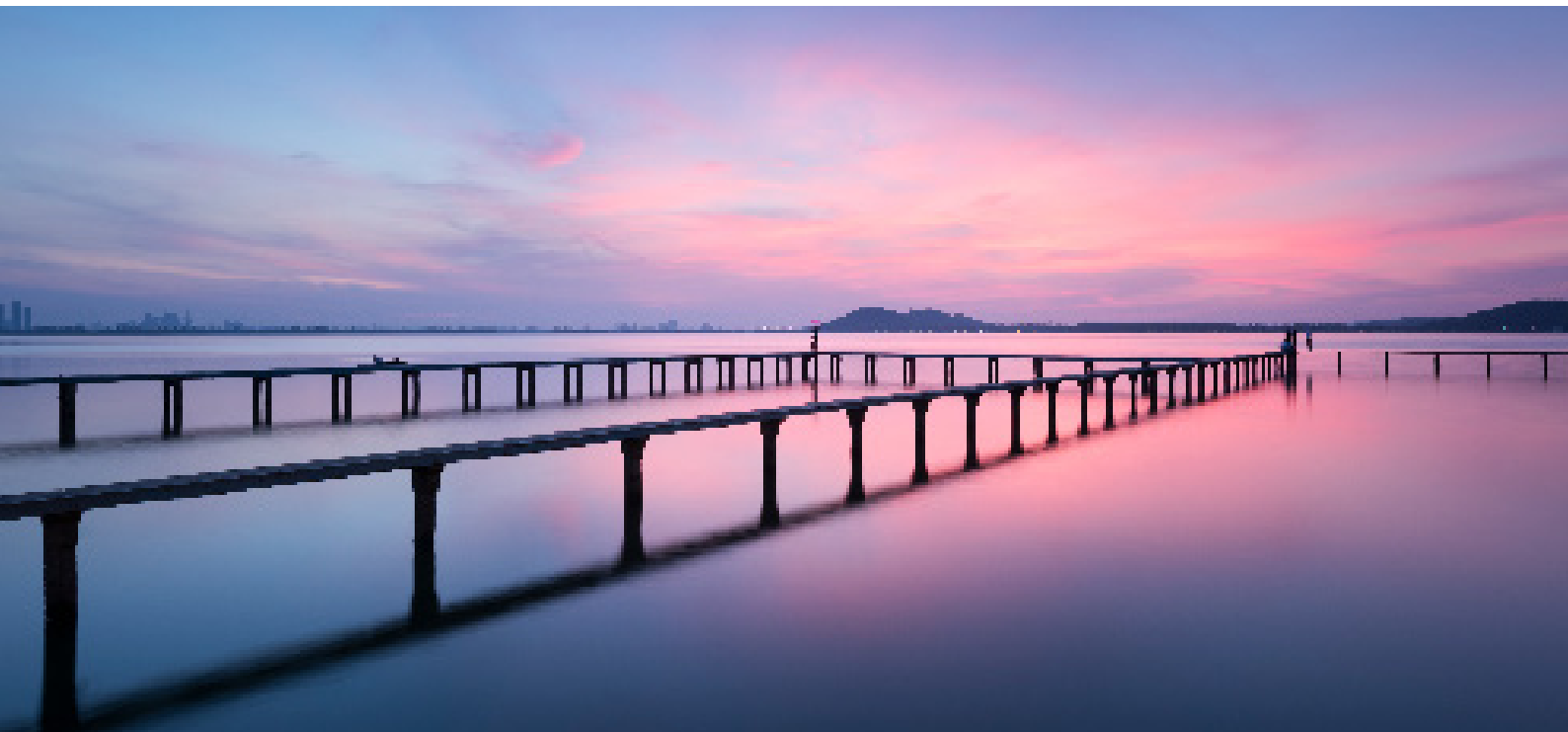


图1.1 基于混合云的智能体协同模式

第二种，我们判断也是未来更主流的模式，是基于混合云的智能体协同模式。企业核心数据、业务流程和智能体运行在本地，同时通过大模型网关、运营管理与安全治理能力，安全调用公有云最先进的MaaS服务、MCP工具以及Skill能力。这样既能够持续获得业界最新AI能力，又能保障数据安全、权限可控与资产沉淀。



因此，我们定义了企业面向智能体的混合云架构。

图1.2 企业面向智能体的混合云架构

Agile Cloud-Native基础设施层

以本地云为核心、混合多云协同的统一底座，实现通算智算资源深度融合，弹性适配业务需求，敏捷响应业务波动，多元硬件与多代次芯片架构兼容，全域云资源统一管控，保障底层架构稳定可靠。

Artificial AI-Native使能层

作为企业统一AI能力中枢，构建湖仓融合多模一体的AI数据湖底座，打通数-训-推-强全流程闭环，实现高质量私域数据+RL驱动模型的自主迭代升级，实现Token模式标准化AI能力，全域业务都能安全普惠AI能力。同时建设大模型网关，实时安全调用公有云最先进的MaaS服务、MCP工具以及Skill能力。

Agentic Agent-Native赋能层

依托Harness企业工程范式及高低码开发协同，重塑企业知识本体，承载企业全场景智能化业务流程，支撑大规模智能体并发运行，实现业务智能化落地。

Administrable智能化运营治理

针对海量智能体与Token资源做专业化治理、运营和运维，通过精细化运维和Token资源经营，有效控制成本、提升整体运营效率，让AI价值最大化。

Assured智能化安全可信

企业上云的安全底线，保障混合云网络、数据传输安全，更针对智能体新增专属安全防护，做到全场景风险可控、全程合规可信可追溯。

02

Agentic Agent-Native赋能层

2.1 智能体发展核心趋势

能力范式跃迁：从被动问答推理走向主动执行、持续自演进

传统AI应用以被动响应模式运行，依赖用户输出明确、结构化指令，仅可完成问答、文本生成、结构化数据查询等静态任务，不具备自主规划与闭环执行能力，业务价值局限于单点效率优化。

新一代企业级智能体完成核心能力升级，构建感知–记忆–规划–执行–反思全能力闭环。智能体可精准解析企业模糊业务目标，依托思维链（CoT）、思维树（ToT）完成复杂任务分层拆解，通过工具调用编排引擎自主调度API、数据库、检索增强生成知识库（RAG）及第三方服务，实现跨系统资源协同；同时基于执行反馈触发反思机制动态调整执行策略，最终达成业务闭环。

相较于传统自动化工具，智能体核心优势体现在持续自演进能力。依托增量学习、自监督微调、人类反馈强化学习（RLHF），智能体可自主优化任务规划逻辑、工具调用策略与场景适配能力，持续提升任务完成质量与运行效率，沉淀为企业可长期迭代复用的数字资产。

产业场景渗透：开源生态爆发，从个人场景全面转向企业级深度落地

2026年，OpenClaw、Hermes Agent两大开源智能体框架快速崛起、生态走向成熟。依托标准化智能体骨架、低代码编排能力、多模型兼容接口、完备工具生态，两类框架成为产业主流落地基座，打破了智能体仅用于个人测试、轻量化体验的应用边界。

结合标准化开发组件、企业级适配能力与多智能体系统（MAS）能力支撑，智能体应用场景加速延伸，从个人办公、内容创作等轻量化场景，深度融入金融、制造、政务、零售等全行业企业核心业务域，覆盖智能运营、风险管控、生产调度、客户服务、供应链协同等关键环节。目前智能体已迈入企业规模化部署、业务核心化承载、场景体系化落地新阶段，成为企业数字化转型的核心基础设施。

2.2 智能体规模化落地核心挑战

智能体规模化与复杂度激增，亟需构建高效专属工具链

随着智能体迈入规模商用阶段，其规模体量、系统复杂度及任务并发量呈指数级攀升。当前，高低码协同、多智能体互动、以及跨混合云（公有云/私有云/边缘端）的动态编排已成为主流业务形态，传统软件架构已难以适配这一新型范式。

一是高低码协同能力不足：高度依赖大模型、软件工程、业务领域复合型技术人才，缺少轻量化、可视化开发工具，业务人员难以参与共建。

二是资产标准化与复用性差：智能体逻辑、提示词、工具调用规则等核心资产无统一封装与托管能力，无法形成可复用模板，同类型场景重复开发，阻碍规模化落地。

三是异构系统集成难度大：智能体需要对接各类业务系统、数据库、第三方接口，但现有工具缺乏标准化连接器、统一网关与协议适配能力，系统对接定制成本高、兼容性差。

智能体效果对场景敏感，性能、成本、稳定性工程化难题突出

智能体业务输出具备强场景敏感性，任务完成质量高度依赖基础大模型能力、上下文窗口管理精度、外部工具调用可靠性、RAG检索效果等要素，任一要素波动都会引发整体业务效果偏差。区别于传统标准化软件，智能体动态推理、多轮迭代、自主编排的运行特性，进一步提升了工程化落地难度。

混合云部署模式下，企业规模化部署阶段主要面临三类矛盾：

一是性能瓶颈，长链路、多步骤复杂任务易出现上下文溢出、推理延迟、逻辑断点、工具调用幻觉等问题，难以匹配核心业务高时效、高闭环率的运行要求；

二是成本压力，大规模模型推理、高频工具调用、持续状态存储产生高额算力与资源开销，推理成本占智能体总拥有成本（TCO）比重超60%；

三是稳定性风险，智能体决策过程具备黑盒特征，易出现模型幻觉、逻辑推演异常、场景适配失效等问题，同时缺乏统一的运行观测、效果评估与迭代优化机制，企业难以持续量化智能体业务价值并实现能力闭环演进，无法满足企业7×24小时连续、稳定、可控的运行标准。

智能体新型安全风险凸显，安全合规成为企业落地硬性准入门槛

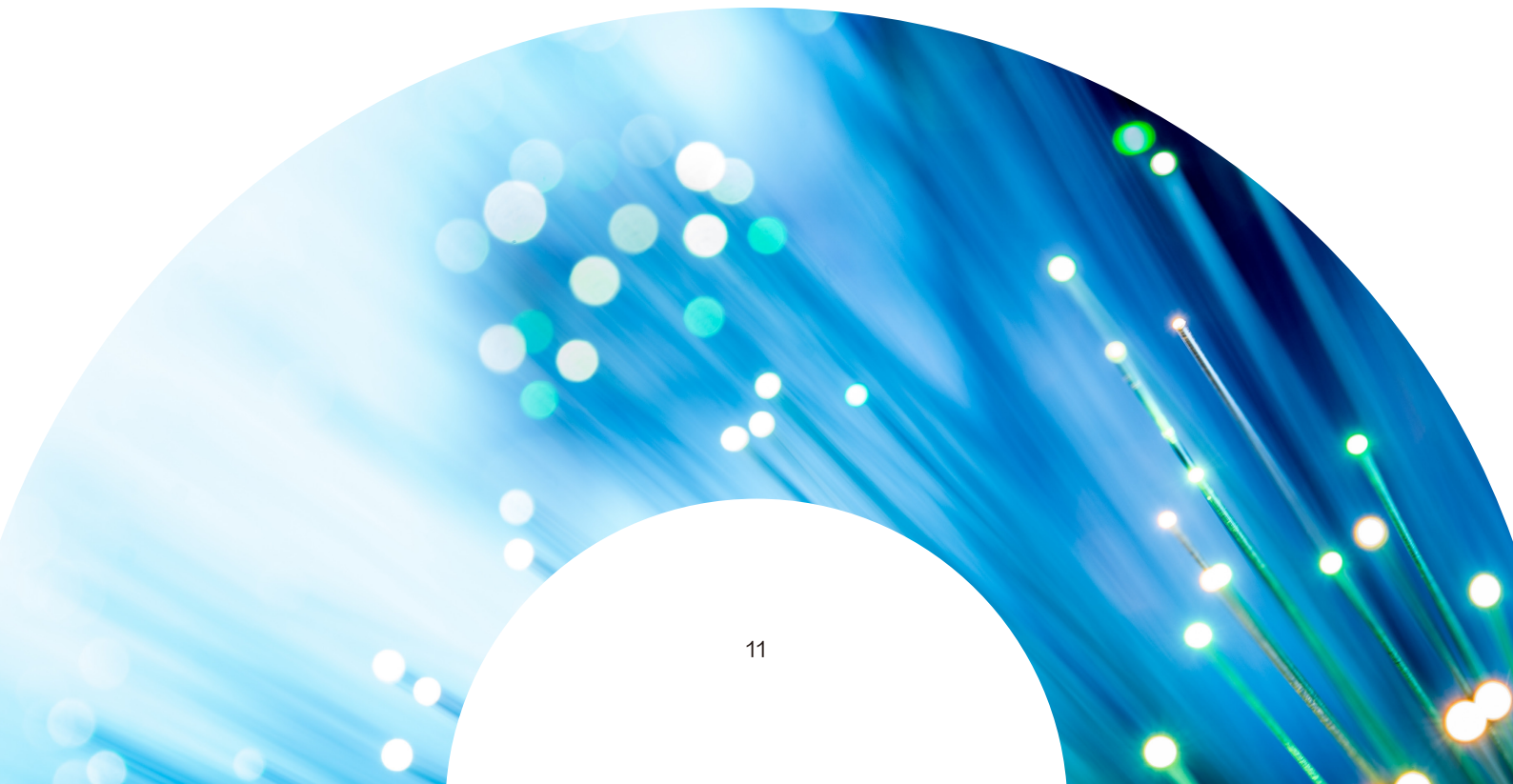
智能体自主调用、跨系统交互、自动化执行、集群化运行的运行特征，重塑传统网络安全风险形态，衍生出新型安全治理难题，主要呈现三大特征：

一是攻击面全域放大，智能体需对接各类业务系统、API接口与外部工具，多维度交互链路拓宽攻击入口，易诱发提示注入、越权调用、恶意指令执行、核心数据泄露等风险；

二是风险链式传导，单点智能体出现安全漏洞后，可凭借跨系统协同、自主调度能力实现风险横向扩散，诱发系统性业务风险，威胁企业整体数据与业务安全；

三是责任界定困难，智能体动态决策、自动迭代、黑盒运行的特点，导致执行链路溯源难度提升，安全事件发生后难以界定责任主体，无法匹配企业合规管控要求。

现阶段，安全合规能力已从智能体附加功能，升级为企业级智能体规模化落地的刚性准入门槛。缺失全维度安全防护体系的智能体部署，将为企业带来严重的业务风险与合规隐患。



2.3 企业级智能体体系化建设思路

面向智能体时代，企业应用架构正从单一系统向多智能体协同演进，同时受数据主权、业务连续性、成本优化及创新敏捷性等因素驱动，混合云逐步成为智能体落地的主流基础设施形态。因此，企业需构建覆盖开发、运行、运营、安全及协同治理的一体化智能体平台，实现跨公有云、私有云及边缘环境的统一管理与协同运行。

图2.1 智能体平台整体架构

第一阶段：智能体工程化开发能力—构建一站式、高低码协同、原生Harness的智能体开发平台

针对混合云环境下智能体开发效率偏低、定制成本高、标准化程度不足等痛点，需搭建一站式智能体开发平台，覆盖需求梳理、可视化编排、开发调试、测试发布、组件复用全生命周期流程。

平台采用高低码协同架构，兼顾落地效率与功能灵活性：低码模式支持业务人员通过可视化拖拽、模板复用快速搭建标准化场景智能体，降低业务落地门槛；高码模式开放全量API与SDK，支撑技术团队开发复杂业务逻辑，适配高复杂度、高定制化核心场景。

平台原生集成Harness工程能力，内置标准化智能体骨架、记忆组件库、工具编排引擎、RAG增强模块，统一开发规范、简化集成流程，从根源解决智能体开发碎片化、兼容性弱、迭代效率低等问题，全面提升企业批量开发与组件复用能力。

第二阶段：智能体弹性运行能力—打造高性能、极致弹性的智能体运行时环境

智能体运行时呈现动态化、高波动、高并发的典型特征，对资源弹性调度能力提出极高要求。平台全面兼容公有云、私有化及本地边缘部署等多种架构，适配不同企业的IT基础设施与安全合规要求；同时内置智能模型路由与混合推理引擎，可按需调度不同能力、不同部署形态的大模型协同工作，有效平衡性能、成本与安全性，全面覆盖各类业务场景的智能体落地需求。

运行平台可采用microVM+PicoOS双重隔离架构，构建多层安全防护体系。一方面实现会话级隔离，为每一个智能体会话分配独立沙箱，保障会话数据相互隔离、互不干扰；另一方面依托虚拟机级隔离打造硬件级安全边界，有效抵御恶意代码对内网的渗透攻击。同时整体采用羽量级设计，运行资源开销极低，兼顾安全与轻量化特性。

运行平台可依托物理集群预热、软硬件协同的全链路深度优化，大幅提升系统运行效率。架构实现沙箱秒级冷启动，极致优化后启动耗时可达百毫秒级；结合网络Serverless化改造，系统具备强大的高并发承载能力，可支持每秒十万级别的沙箱创建与链路连通。

运行平台可内置多种记忆策略，并支持自定义配置，可适配各类复杂业务场景，有效提升智能体运行效率与响应准确度。其中短期记忆可灵活配置1~365天的有效期，长期记忆则采用持久化存储方案，保障跨会话对话的连贯性。

第三阶段：智能体持续运营能力—搭建AgentOps驱动的智能体快速迭代运维平台

针对智能体黑盒运行、状态可视性弱、效果无法量化、迭代缺乏依据等运维痛点，需打造面向智能体全生命周期的AgentOps体系，实现全链路精细化管控。

平台具备全维度可观测能力，实时采集底层资源状态、智能体任务链路、模型推理过程、工具调用记录、决策日志及异常信息，可视化还原全流程执行过程，实现故障快速定位与根因分析。在面向混合云场景时，平台更需实现本地环境全域统一观测、统一评测与统一运营，避免智能体能力分散管理形成新的运维孤岛。

平台也需要搭建量化评估体系，从技术稳定性、任务完成率、资源利用率、人效提升等维度设置指标，实现智能体运行效果、业务价值的量化、对比与溯源。同时具备持续可演进能力，基于运维数据与用户反馈构建迭代闭环，通过增量微调、提示工程优化、知识库迭代、工具链调整及模型升级等方式，持续打磨智能体能力；配套自动化测试与发布流水线，压缩迭代周期，快速响应业务需求变化，形成“观测—评估—优化—发布”的持续演进闭环。

在持续运营与迭代过程中，平台还需帮助企业将智能体开发经验、业务知识与最佳实践沉淀为可复用资产，构建统一资产库，实现能力复用与规模化复制。平台需帮助企业沉淀资产库，资产库采用分层分类的管理架构，全面覆盖智能体开发全生命周期的各类核心资产。在基础层，资产库提供通用提示词模板、工具连接器、API接口、数据处理组件等基础能力资产，支持开发者快速搭建智能体的基础框架；在业务层，资产库沉淀了行业专属知识库、业务流程模板、智能体角色定义、多智能体协同策略等行业化资产，可直接适配金融、政务、制造、零售等不同行业的业务场景；在应用层，资产库支持完整智能体实例的打包与发布，开发者可基于已有的成熟智能体进行快速定制与二次开发。

同时，资产库内置完善的版本管理、权限控制、搜索检索与评价体系，支持资产的迭代更新、权限分级管理与精准查找，确保资产的安全性、准确性与可用性，并推动优秀能力在不同智能体、不同业务场景及不同环境间快速复用。

第四阶段：智能体安全治理能力—构建端到端、多层次的全链路安全防护体系

围绕混合云环境下智能体全生命周期，打造开发、部署、运行、审计一体化多层次安全防护方案，系统性化解新型安全风险。开发阶段践行安全左移理念，集成代码扫描、依赖审计、风险检测工具，从源头规避配置漏洞、权限冗余、数据泄露等隐患。

部署阶段严格落地最小权限原则，通过细粒度权限管控、角色隔离、接口白名单等机制，最大限度收缩攻击面。运行阶段实现行为实时监测、异常智能识别、风险自动阻断，内置安全护栏、提示注入防护、数据脱敏能力，防范越权操作、恶意调用与风险链式扩散。配套全链路安全审计体系，完整留存智能体决策、调用、操作日志，做到安全事件可溯源、责任可界定，全面满足企业合规要求。针对跨云数据流转与跨环境智能体协同场景，建立统一身份认证、访问控制与数据安全治理体系。

第五阶段：智能体协同自治能力—构建云上云下一体化协同创新体系

在智能体规模化建设过程中，企业普遍面临“技术创新速度”与“数据安全边界”的平衡挑战。尤其是在金融、政务等对数据安全与合规要求极高的行业，这一矛盾表现得更为突出。一方面，大模型技术迭代日新月异，行业智能体需要持续接入最新的模型能力、通用资产与工具生态，才能保持技术领先性与业务竞争力；另一方面，企业核心业务数据、知识资产及生产系统往往具有较高敏感性，尤其是金融、政务等行业，更需严格遵循数据不出域、可审计、可追溯的合规要求。因此，单一的纯本地部署或纯云端部署模式均无法同时满足这两大核心诉求。

云上云下协同架构则为解决这一行业难题提供了有效路径，它既充分发挥了本地部署的数据价值与安全优势，又共享了云端生态的规模效应与技术红利，实现“云上创新，本地运行”的协同模式，有效解决了传统部署模式下“安全与效率不可兼得”的痛点。

具体而言，该架构通过三大核心机制实现协同赋能：

一是云上开发，本地部署。企业可基于云端统一的智能体开发平台与资产库，获取通用提示词模板、工具连接器、行业知识库、预训练模型等标准化资源，快速完成智能体开发与验证，并将成熟智能体部署至本地运行，确保核心业务数据全程不出域。

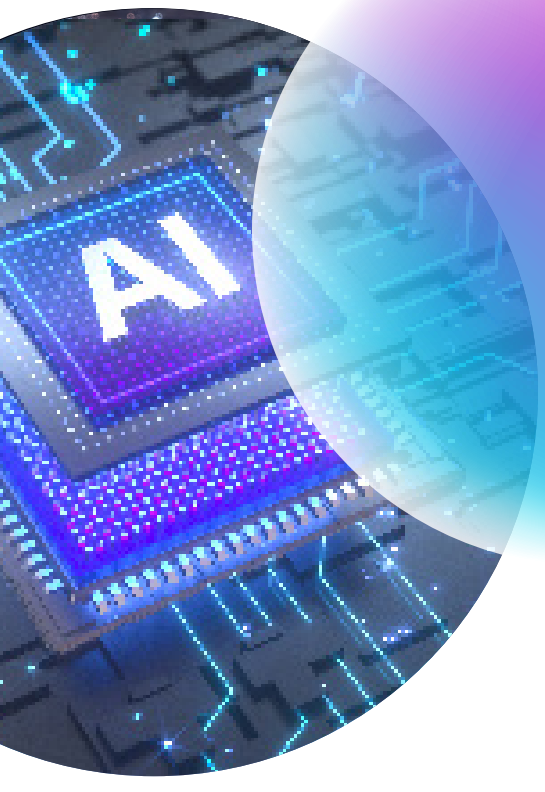
二是云上创新，本地升级。云端持续汇聚最新大模型能力、工具生态与算法优化成果，通过安全同步机制向本地环境持续输送创新能力，使企业智能体能够在不暴露核心数据的前提下快速获得技术升级。

三是云上弹性，本地运营。对于模型训练、批量推理、仿真测试等阶段性高负载场景，企业可按需调用云端弹性资源，在保障关键业务本地运行的同时，充分利用云端规模化算力优势，实现成本与效率的动态平衡。

通过云上创新与本地运行相结合的协同模式，企业能够在保障数据安全与合规要求的基础上，持续获得云端技术创新红利，实现智能体能力快速构建、持续演进与规模化推广，最终达成安全、效率与成本的综合最优。

03

Artificial AI-Native使能层



3.1 大模型网关

3.1.1 趋势与需求

随着企业级智能体（Agent）深度融入生产活动，大语言模型已成为企业提升运营效率、加速业务创新的核心驱动力。在这一趋势下，企业日均Token消耗量正经历从百万级向亿级乃至百亿级的指数级跨越。

在此背景下，如何安全、高效、可控地使用大模型API，已成为当前企业数字化转型的关键课题，如果直接允许员工或Agent调用模型提供商的API，将面临严重的安全隐患、成本失控风险以及合规问题，因此，企业亟需建立一套与自身组织架构和经营策略相匹配的Token分配体系，确保Token在企业内部实现最高效、合规的分配，从而达到Token效益的最大化 。

企业建设大模型网关的核心驱动力主要体现在以下三个方面：

业务形态演进：从“单模型调用”到“多智能体协同”

随着企业内Agent数量爆发式增长，多智能体协作、多大模型联合调度已成为业务常态。这对底层API的管理、路由、熔断、观测等能力提出了系统性要求，驱动网关成为智能体基础设施的必然组件。

IT资源使用主体变化：从“面向IT部门”到“面向全体员工”

传统IT资源主要由技术部门管控，而在Agent时代，Token作为一种新型生产资料，直接面向业务和全体员工供给。企业对Token的配额、审计、成本归属、访问控制等分配治理需求日益迫切，网关成为实现“Token即资源”精细化管理的关键载体。

投资回报考量：从“整体感知”到“量化度量”

企业对AI投入的关注点，正由模糊的总体ROI判断，转向可拆解、可追踪、可优化的量化指标体系。大模型网关能够提供按模型、按场景、按团队、按Agent的成本与效果数据，使ROI评估从“感觉值”进化为“经营指标”。

3.1.2 关键能力

企业大模型网关需要具备以下三大核心能力，以支撑企业级AI应用的规模化落地。

统一接入与路由

通过将多个底层账号能力抽象为“虚拟Key”，网关能够对上层应用保持一致的接入方式，从而彻底消除企业内部的“影子AI”现象。

- 模型抽象层：屏蔽不同大模型厂商API的底层差异，对外提供统一的标准接口（如兼容OpenAI接口），降低应用开发与迁移成本。
- 智能路由：基于任务的具体类型、成本预算、限制、延迟要求以及模型自身能力，动态、自动地选择当前最优模型。
- Fallback（降级）机制：当主模型发生故障或响应超时，系统自动切换至备用模型，有效保障核心业务的连续性。
- 多模型编排：支持串联、并联、条件分支等多种复杂的模型调用模式，满足多样化的业务场景需求。

多租户与身份认证

- 多级租户模型：构建“平台级→企业级→部门级→用户级”的多级租户架构，精准映射企业复杂的组织架构。

- 虚拟Key机制：为每个租户或Agent颁发独立的虚拟凭证，实现严格的权限隔离与精确的用量归因。

- 工作负载身份：支持Kubernetes Service Account、云实例角色等非人员身份认证机制，保障IT应用、Agent机机调用的安全。

- MFA与设备可信：针对高风险操作或高密级数据访问，强制要求二次认证（MFA）或可信设备校验。

成本与配额管控

建立多维度的标签体系（涵盖用户、团队、项目、应用、场景等），支撑从宏观总账到单次Token消耗的全链路下钻分析与管控。

- 多维标签体系：支持系统自动打标与强制打标相结合，并允许人工手动补充，确保计费维度的完整性。

- Token Plan模式：将Token交付标准化、服务化，支持订阅制，并兼容预留配额、预付费、后付费及混合计费模式。

- 精细化配额：支持按模型类型、用户身份、团队归属、应用维度及时间窗口等设置多层次的Token使用上限。

- 动态预算控制：实现预算的梯级预警（如80%、95%、100%三级告警），并在超预算时自动触发熔断机制，同时支持灵活的预算转移。

- 成本分摊与结算：按标签聚合成本数据，支持企业内部的IT计价与财务结算（Showback/Chargeback）

3.1.3 建设思路

第一阶段：企业统一管理Token，面向企业组织和IT统一供给Token

大模型供应商统一管理

企业应建立标准化的供应商生命周期管理体系，确保底层模型能力的安全、稳定与高性价比。

供应商准入：建立全面的模型评测体系，涵盖能力评估、安全评估、成本评估及SLA评估，形成企业内部的合格供应商清单。

凭据管理：统一、集中存储供应商API Key，支持自动轮转与加密存储（如集成KMS/Vault），并实现多账号的负载分担。

健康度监控：实时监控各供应商服务的运行状态（包括延迟、错误率、限频触发情况等），建立动态的供应商健康评分体系。

供应商路由：支持供应商级别的故障切换（主备模式）、用于灰度测试新模型的加权流量分配，以及成本优先的智能调度策略。

生命周期管理：涵盖模型版本管理（升级通知与兼容性测试）、合同与计费系统对接，以及供应商退出时的数据彻底清理。

企业员工和应用Token分级分配

企业需综合考量角色职责、应用场景、数据敏感度及成本预算四个维度，建立差异化的Token供给策略。

个人办公Agent主要用于文档处理、日程管理、信息检索与简单数据分析。按每用户每日8次会话计算， 每用户每日消耗约 1.2M Tokens。

- 交互特征：中等复杂度（调用1-2个工具），需结合RAG（检索增强生成）获取企业内部知识。

- Token消耗拆解：
  单次会话输入包含系统提示词（约1,500Tokens）、RAG上下文（约2,000Tokens）、多轮历史累积及工具调用结果。
  考虑中文字符（约1.5倍Token膨胀）及Agent多步迭代（约2.5倍乘数），单次完整会话消耗约150K Tokens。

编码Agent属于重度Token消耗场景，需要处理庞大的代码库上下文，并进行多次自主迭代和自我修正。按每开发者每日4次深度编码任务计算， 每用户每日消耗约 6.8M Tokens。

- 交互特征：高复杂度（多步骤、多工具），需要深度推理和大规模代码生成。

- Token消耗拆解：
  单次任务涉及长上下文（代码库背景约8,000 Tokens）、多次工具迭代（如编译测试、报错分析，单次任务约5次迭代）及复杂的思维链（CoT）推理。
  结合代码Token特性（约1.8倍膨胀）及高复杂度乘数（约3.0倍），单次编码任务消耗高达1.7M Tokens。

基于Agent使用场景和Token消耗模型建立企业员工Token分配策略：

等级	适用对象	模型能力	配额上限	数据权限	审批要求
L1（基础）	普通员工、内部基础问答	基础轻量级模型	低（如 100K Tokens/天）	仅限脱敏数据	无需审批
L2（标准）	业务骨干、日常文档处理	主流商业模型	中（如 1M Tokens/天）	内部非敏感数据	部门主管审批
L3（增强）	研发团队、代码生成辅助	高级模型+代码专项能力	高（如 6M Tokens/天）	内部数据+代码库	项目经理审批
L4（专属）	核心业务、深度数据分析	旗舰级模型+超长上下文	定制化配额	核心敏感业务数据	安全委员会+业务双审

大模型调用安全合规

构建覆盖网络层、数据层及应用层的全方位大模型安全防护体系。

网络边界防护：针对极高安全要求的业务场景，建议通过云厂商提供的专线或私有网络（VPC）接入大模型API，彻底切断数据在公共互联网上的暴露风险。

敏感信息（PII）脱敏：部署大模型专属防火墙服务，确保所有请求在离开企业内网、进入公有云环境之前，自动识别并剥离任何可识别个人身份的敏感信息（PII）。

访问控制与最小权限原则：基于RBAC模型，按部门、员工、Agent角色精确配置其允许调用的模型类型及可用功能，严格落实最小权限原则。

第二阶段：精细化Token成本经营，实现Token ROI精准评估

建立企业级的Token成本核算体系，确保每一笔消耗都可追踪、可归因、可计价，实现降本增效。

建立成本分摊模型：支持按实际消耗量、按固定比例分摊、按业务优先级等多种计费规则，满足财务级精度的内部结算要求。

多维标签聚合：对API Key附加用户、部门、项目、Agent、场景等多维度标签，支持Token计量数据

的任意维度下钻与多维聚合分析。

实时成本计量：对每笔请求进行实时成本计算（公式：输入Token × 输入单价 + 输出Token × 输出单价），为实时预警提供数据支撑。

成本基准库：沉淀并记录各模型在不同典型场景下的标准成本基线（例如DeepSeek V4 Pro在代码生成场景的平均成本），用于异常调用的快速检测与拦截。

3.2 训推一体使能平台

3.2.1 趋势与需求

面对Token消耗的指数级跨越，模型的使用方式主要存在两种典型模式：一是调用MaaS (Model as a Service) API，二是在本地部署并进行持续微调。两者并非相互取代，而是可以根据业务场景协同使用。在当前政企实践中，后者更为常见（尽管从长期趋势看，未来可能会逐步向MaaS迁移）。在这种本地部署与微调的模式下，企业应用AI由“模型嵌入范式”向“智能体驱动范式”的跃迁，不是纯粹的技术取代，而是有机的能力分工与协同。智能体的核心是规划、记忆和执行动作，而模型的核心是理解、推理和指令生成。企业业务的发展和变化，不仅需要智能体的流程革新，更需要模型通过持续微调学习其行业知识和专业内核，防止技能断层。两者的协同进化需要形成一个“数据加工 → 模型微调 → 智能体升级 → 业务效果评估”的闭环飞轮，需要建立一个统一平台覆盖数据加工、模型微调、部署监控、评估反馈的全流程，将模型开发标准化、工程化，支撑业务快速演进。

3.2.2 关键能力

综上所述，企业建设训推一体能使能平台应包含以下关键能力：

企业智能体大模型高性能推理能力

低时延和高稳定的大模型是支撑智能体流畅运行的关键，为此平台需要具备以下能力：

高性能模型推理能力：基于模型量化、算子融合、投机推理等模型推理技术，结合分布式集群及异构多代次算力资源调度技术，构建低时延、高吞吐、高稳定的模型推理能力。

资源弹性伸缩能力：基于智能体模型资源配额及实时资源利用率进行最优资源配比计算，实现资源池的弹性扩缩容，保障模型承接流量不溢出。

推理服务流控能力：基于智能体对于不同模型的调用需求，通过流量控制能力匹配不同模型流量峰谷特征，确保智能体对于不同模型的稳定调度。

模型多版本管理能力：通过建立模型版本生命周期管理能力，清晰管理模型迭代路径，历史版本可回溯。

企业大模型高效后训练开发能力

企业大模型后训练的目标是将通用基座模型转化为深度理解自身业务、具备领域知识、能执行复杂策略的“专属智能引擎”，更好支撑业务智能体的落地。为此平台需要具备以下能力：

监督微调训练能力：基于企业提供的指令数据（如客服应答、合同审核、代码生成）及行业通识数据，结合平台提供的数据配比能力，进行微调训练，使模型输出对齐企业期望的格式和质量标准，形成企业自有知识模型。

强化学习训练能力：基于微调训练增强后的企业模型及治理后的复杂推理数据，通过强化学习训练使模型具备更强场景理解和分析能力，形成企业自有专家模型。

高质量数据集生产能力

企业大模型后训练及智能体的应用都依赖高质量的语料数据，为此平台需要具备以下能力：

数据基础治理能力：依托企业构建的AI多模数据湖能力，实现表格等结构化以及文本、图片、视频、音频等非结构化数据的统一基础治理。同时通过训推一体使能平台提供的智能化算子与工作流编排能力实现数据特征的自动提取、存储与复用，支撑模型高质量训练数据供给。

数据合成与增强能力：针对企业业务原始数据量的场景，通过AI模型采样、数据合成及多重数据质检策略生成高质量后训练/SFT数据，有效补充企业整体数据集。

3.2.3 建设思路

企业应用大模型和智能体赋能业务转型是一个分段推进，逐步收敛的过程。总体上分为平台建设、场景赋能和专家评估三个关键环节：平台建设阶段基于统一平台拉通数据供给、模型开发、模型应用和业务反馈等环节，实现高效开发，协同演进；场景赋能阶段通过“试点沉淀→规模扩散→治理闭环”逐步落地；专家评估环节，由跨领域专家团队负责标准制定、模型审核与平台维护，并建立业务与开发的双向反馈机制，驱动模型持续优化。具体来说，对企业实施建议包括：

第一阶段：建设统一平台，实现技术共享

平台建设统一国产化算力适配的业界主流训推开发框架，企业内各组织开发智能体基于统一模型技术栈标准进行，促进技术共享，避免内部技术壁垒。

平台提供统一业务数据与模型语料数据的转换标准，同类智能体开发场景使用数据实现归一化、模版化治理，形成高质量可复用数据集。

平台提供统一模型效果评价标准体系，智能体开发团队与模型开发团队统一价值观，避免模型产生负反馈。

第二阶段：标杆场景试点，总结统一架构

选择流程边界清晰、决策依赖明确、反馈闭环短的场景进行模型和智能体应用的先期试点，总结每个场景的最优模型与智能体组合方案，形成“场景—模型—工具”适配架构，支持长期持续演进。

构建企业各智能体开发团队统一的工具调用模版、提示词模板、记忆接口等标准，规范模型演进路线。

第三阶段：场景规模扩散，应用与开发闭环

构建由业务架构师、数据专家、模型专家、智能体专家、合规专家组成的企业级AI评估团队，共同负责标准制定、平台维护、模型和智能体审核，以集中化的方式，管理和推动AI的规模化、规范化落地。

构建“业务-开发”双向反馈机制，业务的结果反馈持续反哺模型，使模型的思维越发贴近业务。

3.3 AI多模数据湖

3.3.1 趋势与需求

在大模型与智能体Agent加速规模化落地的背景下，数据平台正在经历一次根本性定位重构。过去，数据平台的核心价值主要体现在支撑BI分析、经营决策与流程数字化；而在AI时代，数据平台开始从传统数据分析系统，演进为赋能AI的智能数据基础设施。这一变化的本质，在于数据消费对象发生了根本转移——从面向BI报表与分析，转向面向大模型训练、增量微调、检索增强生成以及智能体调用推理协同。传统湖仓平台仅能处理结构化数据、依赖单一CPU算力，无论数据形态、处理时效、算力架构还是扩展能力，均无法适配AI全流程的数据供给要求。

在此背景下，企业对新一代智能数据平台提出明确需求：

1）具备全域多模态数据工程能力，可加工文本、图片、音视频以及原有的结构化数据，构建标准化多模态训练数据集，支持跨模态检索；

2）实现异构算力统一调度，高效承载数据预处理、特征提取、模型推理等AI负载；

3）依托存算分离与统一元数据，实现海量数据低成本存储与全域数据统一管理；

4）配套智能体专属存储组件，留存交互状态与记忆数据，反哺模型迭代；

5）支持混合云分级调度，将高弹性、低安全等级的数据集预处理、离线训练任务分流至公有云，兼顾计算效率与成本优化，同时将核心敏感数据留在本地，确保安全可控。

3.3.2 关键能力

围绕高质量AI数据集生产、模型训练数据供给、智能体调用模型推理三大核心目标，新一代平台融合大数据与AI技术，构建五大核心能力，形成从数据接入、加工治理、数据集构建，到模型训练、智能体推理的全链路支撑体系，推动企业数据体系从“支撑业务分析”迈向“驱动智能生成与自主决策”的新阶段。

图3.1 AI多模数据湖架构

多模态数据处理与跨模态检索能力

平台集成多模态数据处理引擎，支持库、表结构化数据以及文本、图片、音视频、文档等非结构化数据统一加工，完成数据清洗、格式转换、特征提取、标注等全流程，批量产出标准化高质量多模态训练数据集，基于统一向量湖实现多模态数据向量化存储与跨模态检索，方便算法人员筛选样本、智能体调用模型时快速匹配关联数据。同时，兼容SQL、Python、DataFrame多种交互范式，适配算法与数据工程师工作流程，不同语言环境无缝打通，深度集成主流AI开源生态，减少环境适配成本，提升数据集加工、模型调试与智能体开发效率。

基于分布式调度框架的异构算力统一调度能力

核心采用Ray或具备同等能力的AI异构调度引擎，构建企业级异构算力调度平台，实现CPU、GPU、NPU资源的统一管理与弹性伸缩，突破单机算力瓶颈，将复杂Python任务、多模态处理任务、模型推理任务无缝分布到集群所有节点。平台根据任务类型自动分配最优算力，将通算任务调度至CPU资源池，AI负载调度至GPU/NPU资源池，同时提供算力配额、任务优先级、故障自动恢复等能力，提升资源利用率，保障核心任务稳定运行。

云原生存算分离与统一元数据管理能力

多模数据处理引擎需基于云原生存算分离架构部署，以支撑海量非结构化数据的低成本存储与弹性计算；企业原有存量湖仓可以选择进行改造，若需实现存量结构化数据与新增非结构化数据的统一管理与统一存储，可考虑对存量湖仓进行存算分离演进，核心数据存储层基于对象存储承载海量训练数据集、原始多模态数据，具备PB至EB级扩展能力，大幅降低海量数据的存储成本。计算引擎可依托Serverless架构实现秒级弹性扩缩，灵活应对数据集批量生产、模型集中训练带来的算力峰值。统一元数据中心对数据集、数据表、向量索引进行统一管理，实现多计算引擎数据共享，保障模型、智能体跨数据源关联查询与数据调用的准确性。

智能体专属状态与记忆存储能力

部署智能体状态存储引擎，可专门存储智能体调用模型过程中产生的上下文信息、交互记忆、任务状态及调用日志的，支持毫秒级读写、增量更新和语义检索。它不仅保障智能体连续执行任务，还能将沉淀数据整理为专项数据集，反哺模型再训练，形成“智能体运行—数据沉淀—模型优化”的闭环，为持续提升智能体能力提供核心支撑。

混合云任务调度与分级运行能力

平台支持按数据安全等级、任务类型进行跨云调度。将多模态数据集批量预处理、大规模离线模型训练等高弹性、低安全要求的任务调度至公有云，按需取用弹性算力，任务结束后即时释放资源；涉密数据集处理、核心智能体实时模型推理等高安全任务严格运行在本地集群。平台实现跨云任务监控、状态同步与异常告警，全域管控数据集生产与模型调用任务，平衡算力成本、运行弹性与数据安全。

3.3.3 建设思路

平台整体遵循存量复用、平滑演进、AI导向、分步迭代的原则，在企业现有传统大数据湖仓基础上进行架构改造与能力升级，分三个阶段逐步将平台从传统业务分析底座，转型为面向模型训练、智能体推理的AI数据专属供给平台，最大限度保护企业现有IT投资，不影响已有的核心业务稳定运行。

第一阶段：核心能力建设，打造AI多模数据湖

本阶段为平台建设核心，引入多模态数据处理引擎与AI异构调度引擎，依托多模引擎拓展数据处理边界，全面接入非结构化数据，唤醒企业80%的沉睡数据，支持结构化与非结构化数据关联融合分析，搭建自动化多模态数据加工流水线，实现全品类AI训练数据集标准化制作，同时构建统一向量湖支撑

跨模态检索。借助AI异构调度引擎搭建CPU、GPU、NPU混合算力池，高效承载多模态解析、特征向量化、大规模数据集并行加工、模型推理等各类AI负载。至此，平台可规模化生产高质量多模态数据集，全面支撑大模型训练与调优。

第二阶段：基础架构优化，可选完成存算分离与统一元数据管理

建设AI数据湖能力的同时，也必须以稳定现有生产业务为前提，根据企业实际需求决定是否将存量湖仓结构化数据迁移至对象存储完成存算分离演进，若完成改造可通过统一元数据实现结构化与非结构化数据的统一管理与跨源关联查询，打通原有湖仓系统与新增AI数据链路的元数据壁垒，实现多引擎数据共享与基础数据治理，保障数据集标准统一。

第三阶段：闭环赋能，全面支撑智能体规模化落地

完善平台全量能力，构建“数据生产—数据集供给—模型训练—智能体调用”的闭环体系：通过部署智能体状态存储引擎，智能体的运行状态、交互记忆及任务日志得以高效存储，形成可直接用于二次训练的高质量数据沉淀；引入混合云分级调度，将高弹性、低安全敏感任务分流至公有云，优化算力成本，同时通过全链路数据安全、动态脱敏和操作审计保障核心数据资产安全。

平台从数据接入、加工治理，到模型训练与智能体推理，全生命周期全链路适配，持续稳定输出高质量数据集，支撑大模型迭代优化与智能体自主决策，实现企业数据平台从传统分析型架构向AI多模数据湖的平滑升级，打造面向智能化时代的核心竞争力。

04

Agile Cloud-Native基础设施层

4.1 趋势与需求

随着大模型技术持续突破，企业数智化建设正从“应用驱动”向“智能驱动”演进，从“模型调用”迈向“智能体协作”。传统云基础设施主要面向Web应用、微服务等请求-响应式负载，擅长服务编排和资源管理，无法适配智能体长周期运行、强状态依赖、多轮迭代、链式工具调用的核心特征，存在状态治理缺失、上下文无法持续保障、跨系统协同调度能力不足等短板，难以支撑智能体稳定、可治理的生产级运行。

为此，企业需构建敏捷云原生（Agile Cloud-Native）基础设施，原生支持长任务编排、状态持久化、上下文管理、工具治理与全链路追踪；同时适配智能体全新算力架构，实现通算承载业务调度、智算支撑AI推理决策的深度融合，推动IT架构从单一资源池升级为通智一体化算力底座。

智能体规模化落地驱动企业基础设施形成四大核心发展趋势：

通算智算融合成为常态，双算力互补共生、全程联动支撑多智能体全流程作业；

面向智能体的精细化弹性伸缩成为核心能力，解决智能体负载波动大、阶段资源差异大的问题，实现通算智算资源按需调度、动态复用；

多元算力统一纳管成为刚需，适配硬件快速迭代、多代次算力长期共存的格局，破除资源孤岛、提升整体利用率；

多云协同成为行业标配，依托私有云承载核心合规业务、公有云提供前沿模型与算力，实现智能体跨云协同运行。整体而言，传统静态固化的云架构已无法适配AI生产场景，企业必须升级打造具备通智融合、弹性伸缩、异构调度、多云协同能力的敏捷云原生底座，支撑全域智能体稳定、高效、规模化自主运行。

4.2 关键能力

智能体时代对云基础设施提出了新的要求：资源更多元，管理更统一，调度更智能，供给更弹性。通过通算智算融合管理、通算智算弹性伸缩、异构硬件统一调度、多云算力统一纳管，四种能力充分协同构成智能体原生的基础设施，向上屏蔽异构硬件复杂性，向下榨取每一分算力的极致效率。智能体不再受限于单一云、单一硬件或固定资源规模，而是根据任务需求灵活调用最合适的算力资源，实现“算力随需而动、资源全局最优、智能体高效运行”。

通算与智算融合管理，支撑智能体全链路运行

智能体的运行并不只依赖单一的AI加速资源，而是需要通算与智算协同配合。模型推理、向量检索、图像识别、语音理解等任务需要NPU、GPU等智算资源支撑；而任务调度、业务逻辑执行、数据清洗、接口调用、应用服务运行等环节则依赖CPU、内存、存储和网络等通算资源。

因此，需要构建通算与智算融合的资源管理体系，将CPU、NPU、GPU、存储、网络等资源统一纳入调度框架。平台可根据智能体任务的类型、优先级、时延要求和资源消耗特征，自动匹配合适的算力组合，实现通算与智算之间的协同供给。通过这种方式，智能体可以在同一资源底座上完成从感知、理解、推理到执行的全流程运行。

通算与智算弹性伸缩，让智能体按需使用算力

智能体业务具有明显的动态性和不确定性。不同时间段、不同任务类型、不同用户规模下，对算力的需求可能出现快速波动。例如，大规模并发问答、复杂工具调用、多智能体协同、批量内容生成等场景，会在短时间内拉高智算和通算需求；而在低峰时段，资源需求又会明显下降。

因此，算力底座需要具备面向通算与智算的统一弹性伸缩能力。平台可根据任务队列长度、模型调用频率、NPU利用率、CPU负载、响应时延等指标，动态扩缩容相关资源，实现计算资源随业务负载自动调整。对智能体而言，底层资源不再是固定、刚性的供给，而是可以按需获取、即时扩展、灵活释放的弹性能力，从而支撑更高并发、更复杂任务和更低成本运行。

多代次硬件统一管理，提升异构算力利用效率

伴随AI算力技术的高速演进，新一代超节点架构硬件持续规模化落地，依托底层架构重构与互联体系升级，实现了算力网络能力的跨越式创新。相较于传统分散式算力硬件，新一代超节点通过全新的集群架构设计与高速互联体系，大幅提升集群通信带宽、降低跨节点传输时延，具备更强的算力聚合能力与分布式协同性能，彻底革新了传统算力集群的组网与运行模式。在企业实际环境中，算力基础设施往往呈现多代次并存的状态。既有历史建设的CPU服务器、上一代NPU/GPU集群，也有新引入的高性能AI加速卡、国产化超节点算力等。不同代次、不同架构、不同厂商的硬件在性能、驱动、框架适配和调度方式上存在差异，容易形成资源割裂的情况。

面向这一问题，需要通过统一资源抽象和异构硬件适配能力，将不同代次的通算与智算硬件纳入同一资源池进行管理。平台可对硬件能力进行标准化标识，包括计算性能、显存容量、指令集、加速框架、能耗水平等，并根据任务需求进行智能匹配。这样既能充分发挥新一代硬件的高性能优势，也能让存量硬件继续承担适配性任务，提升整体资源利用率和投资回报。

多云算力统一纳管，实现“哪里有算力，哪里就可用”

面向多云并存的现实环境，企业需要打破不同云平台之间的资源壁垒，将公有云、私有云、专有云等多云算力资源统一纳入管理体系。通过统一资源视图、统一调度策略和统一权限治理，屏蔽底层云平台差异，实现跨云资源的统一发现、统一分配与统一运维。

这种模式能够帮助企业充分利用已有云资源，避免算力孤岛和资源闲置。当某一云环境资源紧张时，智能体任务可自动调度至其他可用资源池；当业务对安全、成本或性能有不同要求时，也可根据策略选择最合适的云资源承载。最终实现多云算力“统一可见、统一可管、统一可用”。

4.3 建设思路

Agile Cloud-Native基础设施层的建设应遵循“整合—架构—调度—优化”的渐进式路径，在兼容企业现有基础设施的前提下，逐步向智能体原生算力范式演进, 为全域智能体应用筑牢底层支撑。

第一阶段：打造通算智算融合统一算力池

全面推进企业全域算力资源池化整合，统筹纳管企业存量通用算力与AI专用算力，深度打通CPU、NPU等各类异构算力资源，建成通算、智算深度融合的一体化算力资源池。依托标准化服务封装与底层资源抽象能力，屏蔽底层硬件差异，为上层智能体应用提供统一接入、可调度、可度量的稳定算力供给体系。

第二阶段：基于Agent需求实现通算智算双向弹性伸缩

围绕智能体任务并发规模、运行周期、模型调用频次、业务等级等核心维度，制定智能化调度策略与规则体系，以智能体实际使用需求为核心导向，推动通用算力、AI智算算力同步实现按需分配、动态扩缩、自动回收与循环复用。借助通算、智算双向弹性伸缩能力，强化业务峰值时段的资源保障能力，降低业务低峰期的算力闲置，实现算力资源价值最大化。

第三阶段：搭建训推一体调度架构，实现多代次异构算力统一纳管

面向智能体全生命周期场景所需的Agentic Model微调、评测验证、多智能体高并发在线推理等多类型任务，构建全域训推一体化算力调度架构。在多代次通算、智算资源基础上，构建统一资源池，结合任务优先级、时延要求、资源负载特征及业务SLA要求进行精细化调度，实现训练、推理类算力资源统一编排、动态分配与协同复用。配套完善资源隔离、优先级管控、智能调度等机制，在保障在线推理任务高实时性、高稳定性的同时，最大化模型训练、增量微调等离线任务的资源利用率，有效规避算力割裂、资源抢占、重复建设等问题。

第四阶段：推进多云协同资源整合，深化全链路性能优化，完成场景化算力适配持续升级

面向企业主流智能体业务场景，立足多云协同架构，深化实现全域算力资源池化整合，同时对推理链路、通信传输、缓存机制、算力负载开展深度优化，全面提升高频交互、多轮推理、多智能体协同等复杂场景的响应效率。结合各类业务负载特点落地场景化算力定制适配，形成分类分级的算力供给方案，持续强化多云架构下智能体底座的运行性能、稳定性与成本效益。

05

Assured智能化安全可信

5.1 趋势与需求

AI建设模式正从企业重资产自建、封闭独享的私有算力集群，转向公有云弹性算力与私有云敏感数据、核心模型相结合的混合云架构。这种模式在大幅降低企业AI建设成本、提升算力利用效率的同时，也彻底打破了传统的物理安全边界。企业的训练数据、模型参数和推理请求需要在私有云、公有云、边缘节点之间频繁跨域流转，数据在传输、存储、计算的每一个环节都面临泄露和被窃取的风险。

传统的网络隔离和访问控制手段在混合云环境下效果大打折扣，不同云厂商的安全标准不统一、接口不兼容，导致安全策略难以统一部署和执行。特别是对于金融、医疗、政务等对数据合规性要求极高的行业，如何在利用公有云算力的同时，确保敏感数据不出域、不泄露，成为企业面临的首要难题。

当前企业普遍面临以下核心顾虑：

- 云厂商是否能够看到企业训练数据和自研模型；
- 数据在训练和推理过程中是否存在失控风险；
- 模型权重文件是否可能被窃取或非法复制；
- 云上AI平台与私有云、IDC之间的数据流动是否安全可控；
- 高敏业务是否能够实现独立隔离运行；
- 智能体是否会绕过安全限制，执行高危操作，导致生产系统和数据被破坏；
- 智能体出现异常行为执行时，不清楚行为指令从何而来。

5.2 关键能力

全密态数据流转，客户自持密钥

结合政企 AI 基础设施向混合云演进的整体趋势，针对前文提出的数据跨域管控、推理隐私保护、全栈可信等核心安全挑战，围绕混合云（公有云 + 专属云 + 边缘云）典型建设场景，落地全密态数据流转 + 客户自持密钥核心安全架构，依托端到端加密、HYOK 客户自持密钥、数据出域自销毁、数据流转可观测、机密推理五大关键技术能力，构建覆盖数据全生命周期、全流转链路的安全防护体系，兼顾业务弹性、算力效能与合规要求。

智能体安全可信，模型输入输出保护

构建安全可信的 Agent 开发运行基础设施，构建坚实底座，以基础设施保障硬件和系统层的安全保护策略强制执行，杜绝 Agent 应用和数据被非授权访问和篡改。同时基于 Agent 的实际运行环境和业务侧的相关约束，对 Agent 进行角色和权限定义，对异常操作进行检测和阻断，同时进行全局安全审计，出现问题时及时回溯。

构建模型输入与上下文层面的安全，通过用户输入护栏与上下文护栏协同机制，对 Prompt 注入、越狱攻击、多轮渐进式诱导及角色边界绕过进行实时检测与拦截。同时对系统 Prompt、Tool 调用返回结果以及多轮对话历史等上下文要素进行一致性与可信性校验，防止外部注入信息与上下文污染进入模型推理链路，影响其决策稳定性与行为可靠性。

AI基础设施安全，软硬芯协同，全链路信任可证明

以一个中心、七层防线构建混合云安全基础，打造纵深防御体系。一个中心指安全运营中心，提供安全数据湖、安全智能分析平台、高级威胁检测、资产风险管理、安全态势感知、安全编排与自动化响应等能力，实现安全威胁快速自动化响应能力。七层防线指物理安全、身份认证、网络安全、应用安全、主机安全、数据安全和运维安全，为系统安全运行

提供有力的保障，结合一个中心形成联防联控。

伴随 AI 算力资源从私有化部署向混合多云架构的战略转型，AI 基础设施的信任模型已从封闭可控的内部系统扩展为开放分布的多边算力体系。当算力资源不再由单一实体独占，供应链中的任意环节——从芯片硬件底层固件、网络传输链路、虚拟化管理层到 AI 应用容器——都可能成为攻击的突破口。传统的信任链模型（将信任根植于操作系统或云平台管理层）已难以满足 AI 基础设施全域全栈的安全需求。业界正加速形成以“软硬芯协同、硬件根信任、全栈可证明”为核心的新一代信任链体系。

5.3 建设思路

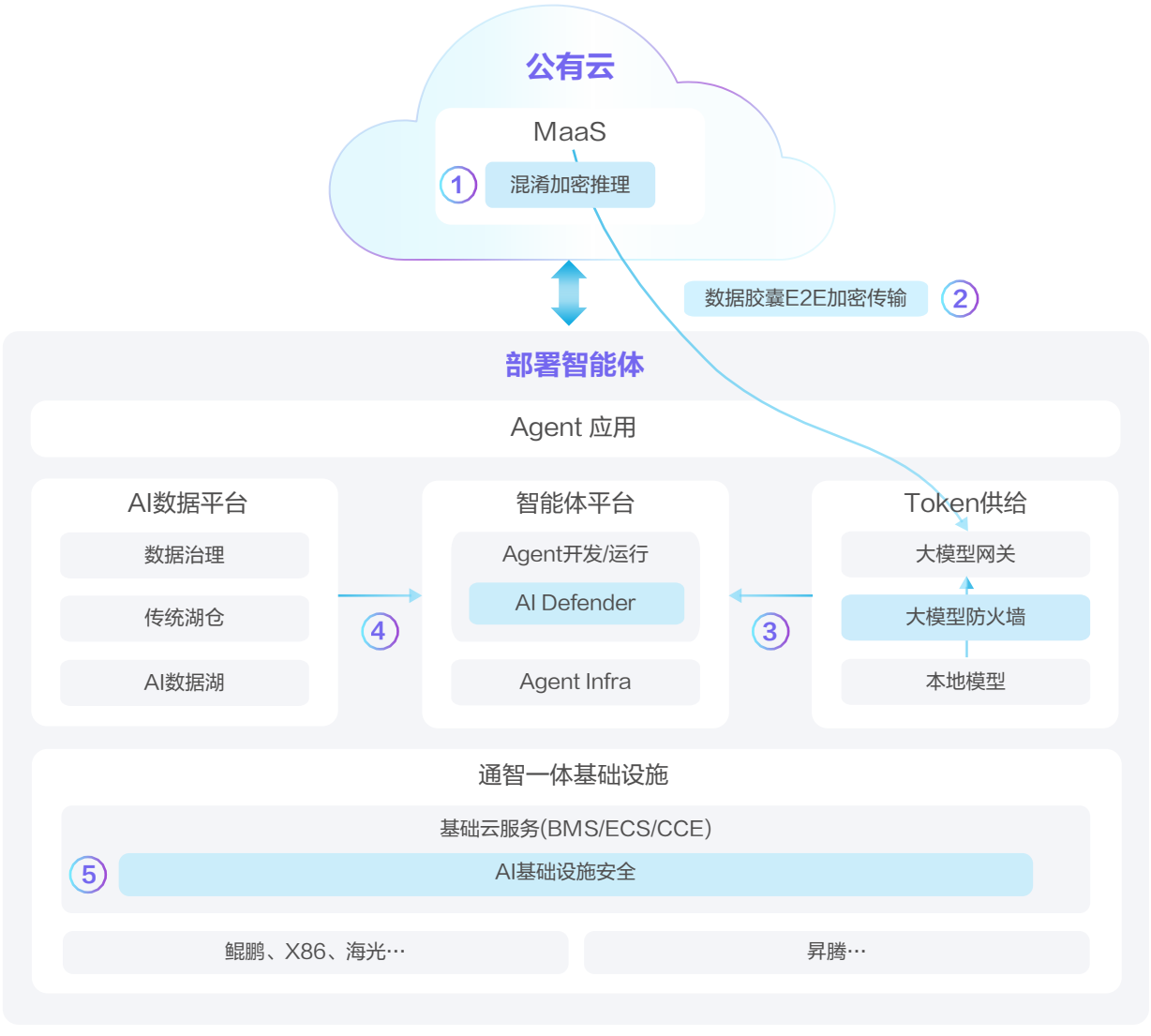


图5.1 混合云智能体安全架构

第一阶段：通智一体基础设施安全

混合云通智一体基础设施安全建设体系，以网络安全等级保护基本要求为指导进行建设，按照一个中心 + 七层防线进行方案设计，构建以安全运营中心的全网安全态势感知能力，协同七层防线能力持续强化纵深防御体系。从 AI 资产保护沙箱安全环境与综合威胁防护三方面夯实底层安全。统一 AI 资产图谱持续纳管模型、数据集与 API 端点，识别后门与 CVE 漏洞，联动流转监控自动隔离异常外发；沙箱环境采用内置 TEE 与安全启动链的硬件级可信，配合毫秒级轻量隔离沙箱，支撑智能体低延迟运行；基础设施安全覆盖虚拟化 / 容器漏洞防护、OWASP LLM 应用层攻击防御，并引入安全智能体驱动的智能化威胁检测，形成闭环防护。

第二阶段：模型、数据与智能体安全

针对模型、智能体安全进行安全保护。模型的提示词攻击、恶意意图和权限越界，部署输入过滤、行为分析与角色限定三层防御。提示词防护通过复杂度监测、规则与机器学习对抗直接 / 间接注入，本地模型加密混淆上云减少成本消耗；意图威胁检测构建全任务链行为序列分析，识别恶意步骤组合（如读取→查询→转账），关键节点设沙箱门禁自动挂起；企业 Agent 角色限定遵循最小权限，办公智能体仅限日程管理，编程智能体禁止推送生产环境，行为识别模型实时拦截“文员写代码、工程师转账”等偏离操作并动态降权。

第三阶段：云上云下协同安全

构建全密态流转、密钥自持、数据离域自销毁与前置隔离的一体化防线。端到端加密确保全链路密文，客户通过 HYOK 专钥专用，本地模型混淆加密后上云推理；数据胶囊绑定统一策略，仅在私有云 + 可信专区内可用，离域即失效并在任务完成后自动销毁；安全前置隔离区对所有模型、Skill 和 Agent 代码先扫描后运行，安全智能体持续巡检云上云下 API 调用，实现资产准入可控与异常行为主动阻断。

06

Administrable智能运营运维

6.1 趋势与需求

随着智能体迈向集群化、规模化部署，企业传统运维体系已无法适配新型业务特征。传统运维聚焦服务器、网络、容器等底层资源状态监控，而智能体的核心风险更多源于自主执行过程的行为异常，包括推理幻觉、工具误用、权限越界、上下文污染、任务死循环与异常决策等非资源类故障，此类问题隐蔽性强、无明确硬件报错，导致人工排查难度大、定位周期长、业务风险高。在此背景下，智算运维正从“ 资源运维 ”向“ 业务运维 ”和“ 智能体全生命周期运维 ”升级。运维能力不再局限于资源可用性监控，而是进一步覆盖智能体开发、发布、运行、评估、优化和下线全过程，强调对智能体执行轨迹、行为状态、任务结果和业务影响的持续观测与智能诊断，以此降低人工排障成本，提升智能体运行的稳定性与可控性。

智能体多轮推理、动态工具调用的运行特性，也带来了资源消耗不确定、成本突发波动等全新运营治理难题，传统粗放式资源统计模式，无法满足按智能体、业务场景、组织维度的精细化计量与成本归因需求。

为此，企业运营体系正向成本精细化、场景化运营、安全合规运营全面转型，通过精准计费、预算熔断、成本归因、行为审计等能力，实现智能体算力消耗可控、行为可管、合规可溯。整体而言，自动化、智能化、全流程可控，已成为智能体规模化落地场景下运维运营能力建设的核心趋势。

6.2 关键能力

为满足面向智能体的运维运营需求，企业需构建覆盖五项关键能力的 Administrable 智能运营运维体系。其中，运维领域聚焦于全链路可观测、故障自愈与全生命周期管理，运营领域聚焦于精细化计量与安全合规保障。五项能力协同构成从“ 看见问题 ”到“ 自动修复 ”再到“ 持续优化 ”的完整治理闭环。

全链路可视化监控能力

构建覆盖算力资源、智能体进程、模型推理、任务流转、工具调用、网络链路等多维度的统一监控体系，实现智能体运行状态的全景可视与实时感知。通过对 CPU、NPU/GPU、内存、存储、网络、Token 消耗、推理时延、调用成功率、任务执行状态等指标进行持续监测，及时发现资源瓶颈、性能异常、行为异常和服务故障。结合可视化看板、链路追踪和实时告警机制，支撑运维人员快速掌握智能体集群运行态势，提升异常发现与响应效率。

智能化故障自愈能力

面向智能体运行故障、模型服务异常、算力资源故障和网络链路异常，建立自动识别、快速定位、智能处置和故障回溯能力。通过异常检测、根因分析、告警聚合和策略编排，自动识别智能体循环执行、推理失败、工具误用、服务超时、资源不可用等问题，并支持任务重试、服务切换、资源迁移、实例重启、限流降级和策略回滚等自愈动作。通过自动化闭环处置，缩短故障恢复时间，降低人工排障成本，提升智能体服务连续性与稳定性。

全生命周期运维管理能力

建立覆盖智能体部署、发布、运行、监控、迭代、升级、回滚和下线的全生命周期运维管理机制，支撑多智能体、多模型、多版本、多环境的统一管理。通过标准化部署流程、版本管理、配置管理、灰度发布、变更审计和运行评估，保障智能体从开发测试到生产运行的平稳交付。面向多智能体集群场景，支持统一编排、集中监控、批量运维和策略管理，降低大规模智能体运维复杂度，提升运维标准化和自动化水平。

精细化智算运营能力

构建面向智能体业务场景的精细化运营体系，支持对算力资源、模型调用、Token 消耗、工具使用 、任务执行和用户访问等运营数据进行统一采集、统计和分析。通过用量计量、成本分摊、利用率分析、负载评估和成本归因，识别高成本任务、低效资源占用和异常消耗链路。结合预算管理、配额控制、成本预警和场景化运营报表，实现资源使用可度量、成本消耗可控制、运营效果可评估，推动智算资源价值最大化。

安全合规治理能力

构建覆盖算力资源访问、智能体操作行为、模型调用链路和数据流转过程的安全合规治理能力。通过身份认证、权限分级、访问控制、操作审计、日志留存和数据追溯，实现对智能体运行过程的全过程管控。针对企业私有化部署场景，支持敏感数据保护、权限最小化、异常访问告警和合规审计，确保智能体在调用模型、访问知识库、使用工具和处理业务数据时满足企业安全策略与监管要求，实现安全可控、责任可追溯、风险可处置。

6.3 建设思路

围绕精细化运营与一体化运维治理两条主线，智能运营运维体系建设应按照“ 统一监控、自动运维、量化运营、安全合规 ”的路径逐步推进，形成覆盖资源、模型、智能体的全流程管理闭环。

第一阶段：搭建统一智算运维监控平台

整合算力资源监控、模型服务监控、业务链路监控和智能体运行监控能力，构建统一的全景监控视图。平台应覆盖 CPU、NPU/NPU、内存、存储、网络、模型推理时延、Token 消耗、任务执行状态、工具调用情况等关键指标，实现从底层资源到上层业务的全链路可观测。在此基础上，建立分级预警机制，对资源瓶颈、服务异常、行为异常和成本异常进行实时告警，为故障定位和运营决策提供数据支撑。

第二阶段：构建自动化一体运维体系

面向智能体部署、运行、升级和故障处置过程，建设自动化运维能力，落地故障自动检测、根因分析、自动恢复、版本灰度更新、策略回滚和资源自动调度等功能。通过标准化运维流程和自动化处置策略，减少人工干预，缩短故障恢复时间，提升多智能体集群运行的稳定性和连续性。同时，结合任务负载与资源状态，动态调整算力分配，保障关键业务优先运行。

第三阶段：搭建精细化智算运营体系

建立面向智能体的精细化计量与运营分析机制，对算力使用量、Token 消耗、模型调用次数、工具调用频次、资源利用率和业务产出效果进行统一统计。通过算力成本核算、成本分摊、资源利用率分析、负载优化和场景运营评估，识别高成本环节、低效资源和异常消耗，实现运营数据可度量、成本可归因、预算可控制、资源可优化，支撑企业开展量化运营与持续优化。

第四阶段：完善安全合规治理规范

面向企业私有化部署要求，建立统一的权限管控、身份认证、日志审计、数据流转追溯、风险预警和应急处置体系，保障智能体在模型调用、工具使用、知识访问和业务执行过程中的安全可控。同步沉淀标准化运营运维流程，明确监控、告警、变更、发布、回滚、审计和应急响应规范，形成制度化、流程化、可追溯的管理闭环，支撑智能体长期稳定、安全合规运行。

扫码了解更多

扫码获取电子版

商标声明

、HUAWEI、华为、 是华为技术有限公司商标或者注册商标。

在本手册中以及本手册描述的产品中，出现的其它商标，产品名称，服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺，华为不对您在本文档基础上做出的任何行为承担责任。华为可能不经通知修改上述信息，恕不另行通知。