华为云MaaS安全白皮书

文档版本 1.0

发布日期 2026-09-10

华为云计算技术有限公司

指导委员会

胡玉海 华为云公有云服务部总裁

曾宇斌 华为云首席网络安全与隐私保护官

编写组组长

ZOU FENG、贾颖杰

编写组主要成员

姚晖晖、凌东东、何应钦、金培松、黄建强、尧应权、陈艳辉、倪建强、陈李、陈龙

特别鸣谢

黄焱霞、杨艳、舒彦峰、郭佳、吴军、翟明刚、陈长青、陈芳菲、陈记伟、党中兴、饶懿璇、黄一天、殷爱军、张忠永

华为云计算技术有限公司

地址: 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址: https://www.huaweicloud.com/

目 录

摘　要

随着大模型能力不断成熟，企业使用 AI 的方式正在由模型自建和单点试验，逐步转向平台化、服务化和规模化应用。MaaS（Model as a Service）将模型供给与调用、模型开发、部署与推理以及安全治理和运营能力封装为平台化服务，使企业能够根据业务需要快速选择、部署和使用模型。华为云 MaaS 以云基础设施和模型服务平台为基础，为客户提供从算力、模型到开发工具和运行管理的一体化承载。

随着 MaaS 进入生产环境，其安全保护对象也从单一模型扩展至算力底座、模型服务、数据、Agent 和运营运维等完整链路。本白皮书围绕 MaaS 应用过程中的主要安全问题，系统说明华为云 MaaS 安全治理框架、安全责任边界、平台安全保障措施，以及客户安全使用 MaaS 所需采取的安全措施和配置实践。

适用范围

本白皮书主要面向计划将大模型能力快速嵌入现有业务流程的企业和组织，为组织中 AI 及 MaaS 架构设计人员，模型与 Agent 应用研发人员，网络安全、数据安全和 AI 安全人员，平台运营运维人员，以及负责数据治理、隐私保护、风险管理和合规审查的相关人员提供参考。

本白皮书重点内容

本白皮书基于责任共担的理念，即**华为云负责云及 MaaS 托管服务本身的安全，客户负责其在 MaaS 平台上配置、上传、开发和使用内容的安全**，围绕 MaaS 从建设、使用到持续运营的安全需求，重点回答以下问题：

一是 MaaS 面临什么安全风险。分析模型服务化、多租户计算、RAG 和 Agent 等场景带来的传统风险变化及 AI 特有风险。

二是 MaaS 安全如何体系化建设。从 AI 治理原则、管理体系和全生命周期出发，构建覆盖算力底座、模型、数据、Agent 和运营运维的安全保障体系。

三是华为云与客户分别负责什么。根据双方对基础设施、平台、模型、数据和应用的实际控制范围，明确安全责任边界。

四是华为云如何保障 MaaS 平台安全。 说明华为云在算力底座、模型服务、平台数据传输与存储、持续安全运营及 Agent 平台基础安全等方面采取的安全措施。

五是客户如何安全使用 MaaS。 从模型选择与使用、数据生命周期、运营运维及 Agent 应用等方面说明客户需要落实的安全要求，并提供可直接参考的安全配置最佳实践。

客户使用 MaaS 的基本安全要求

客户在将 MaaS 用于生产业务前，建议至少完成以下安全配置：

安全连接，根据业务架构选择适当的模型服务访问方式，减少不必要的公网暴露。

身份与权限，按照最小权限原则管理账号、角色、资源权限和 API Key。

数据保护，明确数据来源、授权、敏感等级和用途，落实数据最小化、分类分级和敏感信息保护。

模型与应用安全，根据业务场景选择适当模型，对输入输出、RAG 检索和 Agent 工具调用设置必要的安全控制。

运营与审计，建立 AI 资产、日志、异常调用和安全事件的持续监测与处置机制。

通过上述控制，客户可形成“安全连接—身份授权—数据保护—模型防护—资产管理—监测审计—事件响应”的 MaaS 安全使用闭环，华为云承诺，在客户内容使用方面，**除为客户提供平台及相关服务所必需的处理外，在未获得客户明确授权的情况下，华为云不会使用客户的内容，包括用于模型训练、微调、优化等用途。**

1 AI 业务趋势与风险挑战

1.1 AI 业务发展趋势

随着大模型能力不断成熟，企业使用 AI 的方式正在由单点试验和模型自建，逐步转向平台化、服务化和规模化应用。传统自建模式通常需要企业自行承担算力建设、模型选择与训练、模型部署、安全防护及持续运营等工作，投入成本高、建设周期长，对专业人才和安全合规能力也提出较高要求。

MaaS（Model as a Service）将模型供给与调用、模型开发、模型部署与推理，以及安全治理和运营能力封装为平台化服务，使企业能够根据业务需求选择、部署和调用模型。华为云 MaaS 以云基础设施和模型服务平台为基础，为客户提供从算力、模型到开发工具和运行管理的一体化承载，帮助客户降低 AI 能力的建设和使用门槛，加快 AI 应用进入实际业务流程。

围绕 MaaS 的规模化应用，企业 AI 业务正在发生四方面变化：

- **一是模型使用方式由“自建自用”转向“按需获取”：**

企业可以通过模型调用、模型广场、专属部署和在线推理等方式获得模型能力，模型逐步成为可组合、可配置和可持续运营的云服务。

- **二是部署架构由单一环境转向云、边、端及混合架构协同：**

算力任务、模型权重和推理请求可能在不同区域、算力资源池和业务环境之间流转，企业需要同时关注服务性能、数据流向、资源隔离和跨环境安全。

- **三是应用形态由内容生成转向任务执行：**

企业开始基于模型连接知识库、业务系统、API、插件及 MCP 服务，并进一步构建 Agent。AI 应用由回答问题逐步转向规划任务、调用工具和执行业务流程，权限边界、行为边界和执行结果的可控性变得更加重要。

- **四是 AI 应用由局部试点转向核心业务融合：**

MaaS 正在支持 AI 能力进入办公、研发、客服、金融、制造、医疗和政务等场景。当模型服务被嵌入关键业务流程后，安全问题不仅影响模型输出，还可能影响业务连续性、数据合规、客户权益和企业声誉。

因此，MaaS 安全的保护对象不再局限于单一模型，而是扩展为覆盖算力底座、模型服务、Agent 应用、客户数据以及运营运维活动的完整服务链路。华为云作为 MaaS 服务提供商，需要保障责任范围内的云基础设施、MaaS 模型安全和平台安全，客户需要对其上传和使用的数据、账号权限、业务配置，以及应用安全和相关行为负责。

1.2 AI 背景下的安全风险与合规挑战

1.2.1 MaaS 安全风险

AI 的广泛应用在带来效率提升的同时，也引入了与传统 IT 安全显著不同的风险。OWASP《大语言模型应用十大风险（2025）》梳理了 LLM 应用在提示注入、模型对抗攻击、数据泄露、供应链漏洞、插件权限过载及 Agent 越权执行等方面的核心风险类别。与传统 Web 安全主要关注访问控制、注入、配置错误和组件漏洞不同，AI 安全的差异主要体现在以下方面：

- 攻击对象从系统与代码扩展到模型与语义交互过程。攻击者未必需要突破传统系统边界，也可能通过 Prompt、上下文、知识库或外部文档影响模型行为，使模型生成非预期内容、泄露敏感信息，或触发不当工具调用。
- 攻击方式从语法利用拓展并深度叠加了语义操控手段。提示注入利用模型遵循自然语言指令的特性，诱导其忽略约束或执行非预期任务，传统基于规则的特征匹配手段难以单独实现可靠拦截。
- 漏洞性质从确定性代码缺陷扩展到高度上下文依赖的模型行为偏差。模型受系统 Prompt、检索增强内容和上下文的动态耦合影响，即便输入相近，在缺乏严格约束或上下文偏移时，也可能产生偏离预期的结果，导致传统规则难以覆盖。
- 资产范围从代码和数据库扩展到模型权重、系统 Prompt、训练与微调数据、企业向量知识库、推理日志、插件配置和 Agent 执行记录。
- 风险影响从数据泄露、服务中断等技术后果，延伸到错误业务判断、越权操作和下游系统风险。

总体来看，AI 安全风险体现为攻击语义化、行为概率化、资产模型化和影响业务化。MaaS 平台除具备传统安全能力外，还需建立面向模型行为、RAG 检索、Agent 执行和安全运营的专项能力。

1.2.2 AI 合规挑战

人工智能技术正在深度进入企业业务流程和行业生产系统，其带来的合规挑战也从传统的网络安全、数据安全和个人信息保护，进一步延伸到模型治理、内容安全、算法透明、风险分级、人工监督、审计追溯和责任划分等领域。

在全球范围内，人工智能监管正在由原则性倡议和伦理指引，转向具有约束力的法律制度和可执行的合规要求。不同国家和地区虽然在监管强度、立法模式和实施节奏上存在差异，但整体方向较为一致，即通过风险分类、透明度要求、主体责任、人工监督和持续评估等机制，将 AI 风险治理嵌入产品设计、开发、部署、运营和退出全过程。

图 1-1 AI 合规趋势与挑战

中国当前主要采取“基础法律支撑、专项和场景化规则先行、逐步推进综合性立法”的治理路径。现阶段，中国 AI 监管以《网络安全法》《数据安全法》《个人信息保护法》为基础，并在算法推荐、深度合成、生成式人工智能服务等领域逐步形成专项监管规则。

《生成式人工智能服务管理暂行办法》明确了生成式 AI 服务提供者在训练数据、生成内容安全、用户权益保护、投诉举报、安全评估和算法备案等方面的责任。面向中国境内公众提供具有舆论属性或者社会动员能力的服务时，还需依法开展安全评估，并履行算法及大模型备案、变更和注销等手续。2026 年 4 月 10 日发布、2026 年 7 月 15 日起施行的《人工智能拟人化互动服务管理暂行办法》，进一步将监管延伸至具有持续情感互动特征的拟人化 AI 服务，要求建立覆盖部署、运行、升级和终止的全生命周期安全管理制度，重点防范过度依赖、情感操控等风险，并加强未成年人、老年人、敏感个人信息、交互数据、安全评估和算法备案管理。

在产业发展层面，国务院发布的《关于深入实施“人工智能+”行动的意见》提出推动人工智能与经济社会各领域深度融合，强调安全可控、分业施策和开放共享。该文件虽不直接作为行政处罚依据，但对行业合规建设、监管核查和场景治理具有重要指导作用，也反映出未来 AI 治理将更加重视应用场景、行业属性、数据安全和技术可控。

在强监管行业，企业还需满足行业专项要求。例如，国家金融监督管理总局发布的《关于银行业保险业人工智能安全开发应用的指导意见》，从治理架构、开发应用、数据治理、算力建设和风险管理等方面提出要求，强调银行保险机构建立人工智能全生命周期管理体系，并将 AI 应用风险纳入机构整体风险管理。

除通用 AI 监管外，强监管行业还会叠加更高要求。以金融行业为例，国家金融监督管理总局发布的《关于银行业保险业人工智能安全开发应用的指导意见》从治理架构、开发应用、数据治理、算力建设、风险管理、能力提升、保障与监督等方面提出要

求，强调银行保险机构应建立人工智能全生命周期管理体系，并将 AI 应用风险纳入机构风险管理体系。

欧盟采取的是“综合法案先行、配套规则持续细化”的监管路径。《欧盟人工智能法案》（AI 法案）已于 2024 年 8 月 1 日生效，并按照不同类型的义务分阶段适用。该法案采用风险分级管理思路，对禁止使用的 AI、高风险 AI、具有特定透明度风险的 AI 以及低风险或最小风险 AI 实施差异化管理，同时对通用目的 AI 模型提出透明度、版权保护、模型评估、安全保障和系统性风险治理等要求。

在 AI 法案确立基本监管框架后，欧盟正在通过实施指南、行为准则、技术标准和监管工具，进一步细化通用目的 AI、生成内容标识、高风险 AI 系统及模型安全评估等要求，推动 AI 法案由原则性制度要求转化为可执行、可检查的具体规则。与此同时，欧盟委员会于 2026 年 6 月提出《云与人工智能发展法案》（Cloud and AI Development Act，CADA）草案，拟通过促进云、算力和数据中心基础设施建设，并建立云与 AI 主权评估框架，强化欧盟 AI 产业发展的基础能力、竞争力和技术主权。

人工智能立法正由中国、欧盟等重点市场加速扩展至更多国家。秘鲁于 2023 年通过第 31814 号法律，即《关于促进人工智能应用以推动国家经济和社会发展的法律》，并于 2025 年 9 月通过第 115-2025-PCM 号最高法令批准其实施条例。韩国于 2025 年 1 月 21 日制定并公布《人工智能发展与建立信任基础框架法》，该法于 2026 年 1 月 22 日起施行，围绕人工智能产业发展、透明度、高影响人工智能和生成式人工智能安全管理等建立基础制度。越南于 2025 年 12 月 10 日公布第 134/2025/QH15 号《人工智能法》，并于 2026 年 3 月 1 日起施行，标志着其人工智能治理由政策和战略引导进一步进入专门立法阶段。

上述立法进展表明，**AI 监管正在由少数重点国家和地区向全球更多市场扩展。AI 合规已不再只是大型科技企业需要关注的特殊议题，而正逐渐成为企业开展跨境经营、部署 AI 产品与服务以及进入不同国家市场时需要普遍应对的合要求。**

总体来看，AI 合规挑战正在从“单项法规遵从”转向“跨法域、跨行业、全生命周期治理”。AI 合规既需要满足通用网络安全、数据安全和隐私保护要求，也需要支撑生成式 AI 服务监管、通用目的模型治理、行业专项监管和客户侧合规审查。由此，MaaS 安全能力不应仅体现为模型调用和平台防护能力，还应进一步沉淀为可配置的风险控制、可追溯的日志审计、可验证的安全证据，以及可面向客户交付的合规支撑材料。

2 MaaS 安全治理框架

MaaS 将云基础设施、模型服务、数据处理和业务应用连接为完整的服务链路。其安全治理不仅需要保护模型本身，还需要统筹算力底座、模型供给与运行、客户数据、Agent 及应用，以及持续运营运维等不同环节，并明确华为云、客户及相关第三方之间的责任边界。

华为云 MaaS 安全治理建立在 AI 安全治理框架之上，从治理原则、管理体系和 AI 系统全生命周期等维度提供统一的方法和要求；MaaS 安全治理则结合 MaaS 的平台化、服务化和多主体协同特点，将相关要求进一步落实到，将安全要求落实到 MaaS 平台规划、设计开发、测试验证、发布部署、运营监控和下线退出等活动中，并围绕算力底座、模型、Agent、数据及运营运维构建体系化安全能力。

因此，下图展示的 AI 安全治理框架，是华为云开展 MaaS 安全治理的基础框架和方法来源。通过将该框架与MaaS具体服务场景相结合，华为云形成覆盖治理与管理、AI系统全生命周期和 AI 安全能力的 MaaS 安全治理体系，为 MaaS 服务的安全、可信和持续运营提供统一指导。

图 2-1AI 安全治理框架

治理与管理层明确安全目标和管理机制

六大 AI 治理原则明确华为云开展 AI 开发和应用应遵循的基本要求；战略与政策、组织流程与能力、合规与风控、事件管理、保障与持续改进、意识与文化六项管理要素，则将原则转化为可执行的组织职责、制度流程和管理活动。

AI 系统全生命周期层明确安全要求的落地过程

安全要求贯穿概念、设计与开发、测试与验证、发布与部署、运营与监控及下线阶段，并通过数据流水线、模型流水线和应用流水线融入实际研发与运营流程，推动安全风险提前识别、过程控制和持续改进。

AI 安全能力层提供面向 MaaS 的技术与运营保障

算力底座、模型和 Agent 构成三项纵向能力，分别实现可信隔离与高可用、模型可控合规与可评测、Agent 可授权可审计与可管控；数据安全和运营运维安全作为两项横向能力贯穿各层。

同时，华为云针对 AI 风险在数据、模型、应用和运营环节逐层传递、相互放大的特点，构建数据安全、模型安全、内容安全、应用安全、运营运维安全五道相互衔接的安全防线。五道防线前后联动，形成从源头预防、过程控制到运行监测和闭环改进的纵深防御体系。

2.1 AI 治理原则

华为云遵循六项 AI 治理原则，并将其转化为 MaaS 服务设计、开发、运营及客户使用过程中可执行的安全要求。

图 2-2 AI 治理原则

- **AI 与人协作，服务于人：**AI 应在人的授权与监督下，与人类优势互补协同工作，服务特定目标且不损害人类自主性。
- **透明：**AI 应用应保障用户知情权和选择权，具备可理解与可解释性，使用户了解其决策原理、准确性及局限性。
- **多样性、非歧视和公平：**AI 的设计、训练与部署应避免偏见和歧视，保障数据多元与高质量，并提升模型的可解释性。

- **安全与鲁棒：**AI 系统应在失效或异常情况下将风险降至可接受范围，具备安全、可靠和韧性的防护能力。
- **数据治理与隐私保护：**AI 系统应基于高质量、具代表性的数据开发，并在全生命周期中保护用户隐私与相关权利。
- **追溯与可归责：**AI 系统应明确责任主体，具备可评估与可追溯机制，并通过分层分角色管理风险与责任。

2.2 AI 管理体系

为推动 AI 治理原则在 MaaS 服务中持续落地，华为云建立覆盖决策、执行、监督和改进的管理体系。该体系围绕六项管理要素，将治理原则转化为组织职责、管理制度、研发流程、风险控制和持续运营要求。

图 2-3 AI 管理体系

- **战略与政策：**结合 MaaS 业务目标、技术能力和风险承受水平，明确模型开发、引入、部署和应用的目标、适用范围、使用边界及风险要求，并根据技术和监管变化持续更新。
- **组织、流程与能力：**由管理层牵引，业务、产品、研发、安全、法务、隐私、合规和运营等职能协同参与，将安全要求融入需求、研发、测试、发布、运营和下线流程，并配置相应团队、工具和专家能力。
- **合规与风控：**持续识别适用于 MaaS 的网络安全、数据安全、个人信息保护、模型治理和行业监管要求，并根据数据敏感程度、模型来源、应用范围和潜在影响实施分类分级管理。
- **事件管理：**建立覆盖发现、告警、研判、处置、恢复、通知和复盘的事件管理机制，除传统网络和数据事件外，关注模型投毒、异常输出、提示注入、越权调用和内容风险等 MaaS 特有事件。
- **保障与持续改进：**通过风险评估、安全测试、红队测试、内部检查、审计和管理评审验证控制措施的有效性，并结合漏洞、事件、攻击样本、客户反馈和监管变化持续改进。

- **意识与文化：**面向管理者、产品研发人员、平台运营人员和业务使用者开展分角色培训，使相关人员理解其在数据、模型、系统配置、内容管理和事件报告方面的责任。

2.3 AI 全生命周期安全保障

华为云通过 DevSecOps 流程，将安全要求融入 MaaS 服务的数据开发、模型开发、应用开发、测试验证和运营监控环节，实现安全左移、内生安全和全链可溯，保障数据可控、模型可信和应用安全。

2.3.1 AI 安全流水线能力

依托数据、模型和应用三条内部研发流水线集成安全能力，将风险识别与控制前移至研发阶段，确保研发流程安全可控。

图 2-4 AI 应用全生命周期

华为云将安全要求融入数据、模型和应用三条研发流水线。在数据开发过程中，针对训练数据、微调数据、Prompt、知识库及推理输入输出，通过加密、防攻击、访问控制、合规扫描和敏感信息扫描等措施，保护数据的机密性、完整性与合规性；在模型开发过程中，围绕数据投毒、模型泄露等风险，通过威胁预判、资产保护、安全校验、运行控制和审计追溯等措施，保障模型研发过程可信可控；在应用开发过程中，针对 AI 应用由问答向 Agent 演进带来的行为失控风险，通过安全隔离、访问控制、行为检测与响应等能力，提升应用行为的可观测性和可控性。

2.3.2 持续安全保障

除将安全能力融入数据、模型和应用研发流水线外，华为云还建立覆盖供应链管理、测试验证和持续运营监控的安全保障机制，将安全控制从研发阶段延伸至第三方组件引入、产品上线及持续运行全过程，形成从风险识别、验证到监测处置的安全闭环。

供应链安全：华为云对引入的第三方模型、SDK、数据、框架和插件开展自动化评估与人工评审，对发现的安全风险及时采用安全版本、补丁或替代方案进行处置，降低第三方组件和外部依赖引入的安全风险。

测试与验证：华为云 MaaS 产品按照《华为云安全功能规范》执行上线前测试，并通过常态化红蓝演练持续识别安全风险；针对发现的共性问题开展统一排查和整改，验证相关安全控制措施的有效性。

运营与监控：产品上线后，华为云专业安全运营运维团队围绕“攻击不瘫、数据不丢、监管合规”，依托大模型内容风控态势感知平台和安全云脑开展 7×24 小时安全监测与应急响应，及时发现并处置运行过程中的安全及合规事件。

通过上述机制，华为云形成覆盖供应链引入—上线验证—运行监测—应急处置的持续安全保障闭环，使 MaaS 安全能力能够随产品运行、风险变化和攻击手法演进持续发挥作用。

3 MaaS 安全责任共担

MaaS 安全责任共担是 AI 安全责任共担原则在模型服务场景中的具体落实。华为云与客户的责任边界，以双方对云基础设施、MaaS 平台、模型、数据和应用的实际控制权、可见性及可操作性为基础，华为云负责“云及 MaaS 托管服务本身的安全”，客户负责“其在 MaaS 平台上配置、上传、开发和使用内容的安全”，华为云 AI 安全责任共担模型详见“信任中心-负责任的 AI”。

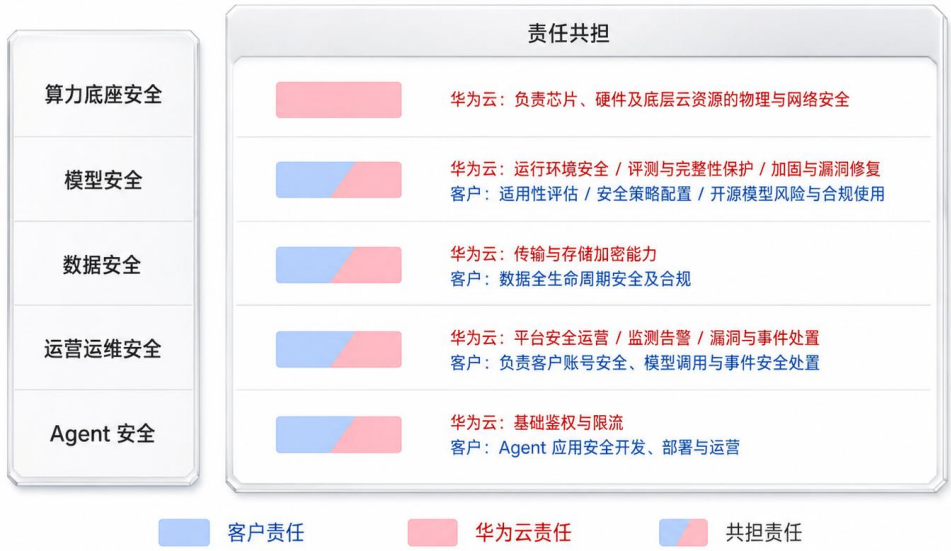

图 3-1 MaaS 安全责任共担

安全领域	华为云责任	客户责任
算力底座安全	• 负责芯片、硬件安全 • 负责底层云资源的物理与网络安全，保障计算环境可信、隔离和高可用	_
模型安全	• 保障模型、部署和推理的环境安全 • 提供模型评测、完整性保护及运行防护能力 • 对模型进行安全加固	• 评估模型适用性并正确配置安全策略，承担开源模型自身逻辑缺陷与安全风险 • 模型应用的安全及内容数据合法合规使用

安全领域	华为云责任	客户责任
		• 理解并遵守所选开源模型的许可协议 • 部署额外防护措施降低开源模型漏洞风险
数据安全	• 提供加密传输通道与基础存储加密能力，保障传输链路安全和底层存储安全	客户负责数据安全主要责任，包括： • 负责数据来源、授权和合规使用 • 负责输入/输出数据安全过滤 • 负责数据脱敏和全生命周期加密
运营运维安全	• 负责 MaaS 平台及其控制范围内组件的安全运营，包括漏洞管理、安全监控、入侵检测、日志记录、告警响应、事件处置及持续改进； • 保障平台服务连续性，并提供平台侧审计和追溯能力。	• 负责客户侧账号异常、模型调用、内容风险和业务安全事件的持续监测，根据需要配置日志、告警和安全策略； • 及时处置客户侧权限滥用、异常调用、内容风险及应用安全事件，并配合华为云开展必要的事件调查、恢复和整改。
Agent 安全	• 提供基础鉴权、限流等安全组件，防止 API 滥用	• 负责智能体应用的安全开发、部署及运营全生命周期安全

表 3-1 MaaS 安全责任共担说明

4 华为云如何保障 MaaS 安全

4.1 华为云 MaaS 面临的安全风险挑战

根据第三章 MaaS 安全责任共担边界，华为云负责云及 MaaS 托管服务本身的安全。在 MaaS 服务持续运行过程中，华为云需要重点应对其控制范围内的算力底座、模型服务、数据传输与底层存储、平台运营运维以及 Agent 基础安全组件等方面的风险。本节围绕上述责任范围梳理华为云面临的主要安全挑战，并在后续章节分别说明相应的安全保障措施。

图 4-1 MaaS 安全风险挑战

算力底座安全风险挑战

MaaS 算力底座具有高算力密度、多租户共享、软硬件栈复杂以及训练推理任务持续运行等特点，安全风险可能通过芯片、固件、驱动、操作系统、虚拟化平台、容器、AI 框架和调度系统等环节沿技术栈传递。芯片、固件及软件组件中的漏洞、后门或不可信组件可能导致计算节点被控制、权限提升或恶意代码持久化；在多租户共享环境下，如果计算、内存、存储或网络隔离机制失效，还可能造成跨租户访问、数据泄露、资源干扰或侧信道攻击。

同时，模型权重、检查点文件及推理中间结果需要在算力环境中临时加载、传输和处理，底层环境被突破可能导致相关资产泄露、篡改或被非法复制；算力劫持、恶意任务、资源抢占和拒绝服务还可能影响训练与推理性能及服务连续性。随着 GPU/NPU 集群、高速互联、分布式存储及跨区域调度规模扩大，基础设施攻击面和依赖关系进一步增加，对传输安全、可信验证、资源隔离、统一安全管理和高可用能力提出更高要求。因此，华为云需要从芯片、硬件和 AI 基础设施三个层次持续构建“可信、隔离、高可用”的 MaaS 算力底座。

模型安全风险挑战

华为云 MaaS 对第三方预置模型开展来源核验、安全评测和准入管理，并保障其在华为云 MaaS 平台托管、部署和运行环境中的安全。模型风险可能贯穿研发与引入、部署发布、推理运行、版本变更和退役全过程：在研发与引入阶段，训练数据污染、第三方模型及依赖组件安全缺陷、恶意模型文件等风险可能影响模型自身安全；在部署和运行阶段，模型权重或服务环境被篡改、模型能力边界发生异常变化以及提示注入、越狱、敏感信息泄露等新型攻击，可能影响模型服务的安全性和稳定性；模型版本升级或安全缺陷持续暴露，还可能形成新的运行风险。因此，华为云需要持续保障模型、部署和推理环境安全，并通过模型评测、完整性保护、运行防护、安全加固降低其控制范围内的模型安全风险。

数据安全风险挑战

MaaS 服务需要承载客户数据在平台接入、推理节点以及基础存储服务之间持续传输和处理，数据形态还可能包括模型上下文、缓存、临时文件及相关平台副本。对于华为云控制范围内的数据处理环境，如果传输链路保护不足、底层存储防护失效、租户隔离机制异常或平台访问控制被突破，可能导致客户数据在传输和存储过程中被窃听、篡改、非授权访问或泄露。因此，华为云需要重点保障 MaaS 平台数据传输链路和底层存储环境安全，为客户数据在平台侧处理提供基础安全保障。

运营运维安全挑战

MaaS 进入持续运行阶段后，平台侧安全风险呈现持续变化和相互关联的特点。算力资源、模型及其版本、Agent 基础服务、知识库、插件和 API 等资产会持续创建、更新和下线，资产之间复杂的调用和依赖关系增加了统一管理难度；与此同时，底层组件漏洞、模型新型攻击、异常调用以及平台配置变化等风险会不断演进，单次上线前评测难以覆盖运行期间持续出现的安全问题。

MaaS 业务调用还可能跨越模型、数据、Agent 基础组件及其他平台服务，分散的日志和复杂调用链会增加安全事件定位、影响分析和责任追溯难度。同时，业务形态、模型版本和合规要求持续变化，也要求华为云不断验证平台安全控制的有效性。因此，华为云需要持续开展 AI 资产管理、风险监测与漏洞管理、安全日志与事件管理以及安全合规与持续优化，保障 MaaS 平台及其控制范围内组件持续安全运行。

Agent 安全挑战

Agent 安全风险。 随着客户基于 MaaS 进一步构建 Agent 应用，华为云提供的 Agent 基础接口及安全组件也需要应对身份鉴权失效、异常调用和 API 滥用等平台侧风险。

4.2 算力底座安全

4.2.1 芯片安全

芯片安全是华为云 MaaS 算力底座可信的起点。华为云依托昇腾 AI 芯片的硬件安全能力，为训练和推理任务提供可信启动、执行隔离、数据保护和可信验证能力。

图 4-2 芯片安全

硬件可信根与安全启动

通过硬件可信根验证引导程序、固件及关键组件的完整性，降低底层组件被篡改或加载非授权代码的风险，并为上层服务器和计算环境建立可信基础。

可信执行与敏感信息保护

通过硬件级隔离和安全存储能力，保护训练数据、模型参数、密钥及推理中间结果，降低宿主机、其他租户或非授权进程访问敏感资产的风险。

访问控制与攻击防护

芯片内部通过安全域划分、最小权限访问及硬件隔离机制控制不同组件和计算任务之间的访问，并结合侧信道和故障攻击防护能力提升运行环境安全性。

可信度量与审计

对关键硬件和固件状态进行度量和验证，发现异常状态时触发告警或阻断，并保留必要的度量记录，支持可信验证和安全审计。

芯片供应链安全

华为在芯片设计、制造、交付和部署过程中实施安全审查、完整性验证和资产追溯，降低设计资产泄露、生产环节篡改及假冒芯片接入风险。

芯片层安全能力由华为云及相关芯片和硬件团队负责建设和维护，客户通过华为云 MaaS 使用相关能力，无需直接管理底层芯片安全配置。

4.2.2 硬件安全

在芯片安全基础上，华为云通过服务器、存储、网络设备和物理环境安全，为 MaaS 训练与推理任务提供设备可信、资源隔离和服务连续性保障。

图 4-3 硬件安全

服务器硬件安全

服务器基于可信根、可信启动和完整性度量建立从固件到操作系统的可信链，并对 BMC、固件、驱动和管理接口实施访问控制及漏洞管理。关键部件采用冗余设计，降低单点硬件故障对 MaaS 服务的影响。

存储硬件安全

对承载训练数据、模型文件、检查点以及客户主动保存至其控制范围内存储资源的数据，通过相应存储服务提供加密、完整性校验、访问隔离、备份恢复及服务质量控制能力，降低数据泄露、篡改、丢失和资源争用风险。MaaS 原生模型推理过程中不存储客户的提示词和模型推理结果，并在推理完成后立即删除。

网络硬件安全

对网络设备实施身份认证、配置基线、固件更新和管理访问控制，并通过网络隔离、加密传输、异常流量监测和 DDoS 防护保障训练、推理及平台管理流量安全。

硬件生命周期管理

对硬件采购、验收、部署、运行维护、固件和驱动更新、退役及介质清理实施全过程管理，并通过资产登记和状态监控保持硬件环境持续可信。

华为云负责其运营环境内的服务器、存储、网络设备及物理数据中心安全；客户在托管 MaaS 模式下无需直接维护底层硬件，但应结合业务连续性要求合理配置可用区、备份和恢复策略。

4.2.3 AI 基础设施安全

在芯片和硬件安全的基础上,AI 基础设施安全从系统层面构建整体安全架构,涵盖计算平台、AI 开发平台和基础设施服务。

图 4-4 AI 基础设施安全

4.2.3.1AI 计算平台安全

取华为云以 ModelArts 等 AI 开发与计算平台承载模型开发、训练和部署任务，并在计算、网络、存储和镜像等环节构建租户隔离与资产保护能力。

计算与网络隔离

通过 VPC、虚拟化、容器和文件系统隔离不同租户及不同任务的计算环境，限制跨实例、跨容器和跨网络访问，并对容器系统调用和文件访问实施运行时控制。

模型和数据资产保护

模型文件、训练数据及中间产物可存储在 OBS 等云服务中，通过加密存储、访问策略、版本管理和备份恢复能力降低非授权访问、误删除和数据损坏风险。

镜像和运行环境安全

对训练和推理镜像实施漏洞扫描、完整性校验、来源验证和仓库访问控制，防止含有漏洞或恶意代码的镜像进入 MaaS 运行环境。

资源与服务可用性

通过资源配额、任务调度、异常任务监测、弹性扩缩容及故障恢复，降低资源抢占、恶意消耗和单点故障对训练及推理服务的影响。

4.2.3.2AI 开发框架安全

AI 开发框架及其依赖组件是连接计算环境与模型开发活动的重要软件层。华为云对昇思 MindSpore 及相关框架、依赖库和组件开展安全开发、漏洞管理、版本管理和发布完整性校验，降低框架漏洞、恶意依赖和不可信组件影响 MaaS 训练与推理环境风险。

对抗训练、模型鲁棒性、隐私保护训练和模型可解释性等模型层能力，将在第四章模型安全中进一步说明；相关权威认证统一在第九章列示。

4.2.3.3基础设施服务安全

华为云通过云原生安全服务，为 MaaS 算力环境提供身份、密钥、主机、容器和网络等基础安全能力：

- 身份与权限：通过 IAM 提供统一认证、多因子认证、细粒度授权、访问密钥和身份联邦能力，控制用户和服务对 MaaS 资源的访问。
- 密钥与加密：通过 DEW、KMS 及硬件安全模块保护加密密钥，为模型文件、训练数据和云存储提供加密支撑。
- 主机与容器：通过 HSS 等能力开展资产发现、漏洞管理、基线检查、入侵检测和容器运行时防护。
- 网络与边界：通过 VPC、安全组、网络 ACL 及 Anti-DDoS 等能力实施网络隔离、访问控制、异常流量检测和拒绝服务攻击防护。
- 日志与监测：记录平台访问、资源变更和安全告警，并接入统一安全运营体系开展持续监测和事件响应。

客户应根据自身业务配置账号权限、VPC 和安全组策略、访问密钥、镜像及依赖组件，并对其计算任务和自定义运行环境的安全负责；华为云负责托管平台、底层服务及其安全机制的持续运行。

4.3 模型服务安全

华为云 MaaS 面向华为云自研模型、第三方预置模型和客户自定义模型提供模型研发、引入、部署和推理等平台能力。根据第三章 MaaS 安全责任共担边界，华为云重点保障其控制范围内的模型、部署和推理环境安全，并提供模型评测、完整性保护、运行防护以及模型安全加固和漏洞修复能力。

围绕“可控、合规、可评测”的模型安全目标，华为云将研发治理、供应链验证、模型资产保护、安全评测、发布控制和运行防护等机制贯穿模型研发与引入、部署与发布和运行三个阶段，并根据模型来源及服务方式实施差异化安全保障。

4.3.1 模型研发与引入安全

华为云 MaaS 根据模型来源实施差异化安全管理。对于华为云自研模型，安全要求贯穿数据工程、模型训练、安全对齐、模型评测和版本发布全过程；对于第三方预置模型，华为云在将其作为 MaaS 模型服务提供前开展相应的来源信息核验、供应链检查、安全评测和平台准入；对于客户自定义模型，华为云重点通过完整性检查、安全评测和平台准入机制降低模型对 MaaS 托管环境造成的安全风险。进入 MaaS 生产模型资产库的模型建立相应的模型标识、版本、安全状态及关联记录，支持后续部署、运行和问题追溯。

华为云可根据模型来源、可控范围和实际服务能力，对第三方预置模型开展相应的平台准入和风险评测，并向客户提供可用的模型或服务说明。评测结果仅反映特定版本、测试范围和测试条件下的风险情况，不构成对第三方模型及其输出安全性、准确性、真实性、完整性或业务适用性的保证。

4.3.1.1模型研发安全

华为云自研模型以盘古大模型为代表。 盘古大模型采用 L0 基础模型、L1 行业模型和 L2 场景模型的分层架构，L0 包括自然语言、视觉、多模态、预测和科学计算五类基础模型。华为云将安全要求融入自研模型的数据准备、模型训练、安全对齐、模型评测、版本发布和缺陷响应过程，并保持数据集、训练任务、模型权重、评测结果和发布版本之间的关联和追溯。

ModelArts Studio 为盘古大模型研发提供数据工程和模型开发能力。数据工程覆盖数据获取、加工、合成、标注、配比、评估和发布等环节，支持形成从原始数据到训练数据集的过程记录，为模型研发的数据质量和可追溯管理提供平台支撑。

数据治理与标注

针对华为云自研模型，研发数据治理与标注主要包括：

- 训练和评测数据应登记来源、许可、使用目的、敏感等级、保存期限和退出条件；来源不明、授权不足或用途不匹配的数据不得进入生产研发链路。
- 通过数据质量检查、重复与异常样本识别、污染和潜在投毒检测及人工复核，降低低质量或恶意数据影响模型行为的风险。
- 标注账号按项目和角色实施最小授权，敏感数据按任务需要进行去标识化，并通过多人标注、结果审核、一致性检查和异常标注复核形成质量闭环。

【华为云服务】**ModelArts Studio** 数据工程支持数据获取、加工、合成、标注、配比、评估和发布。数据标注支持标注任务、审核任务和任务管理，并可根据数据类型使用多人标注、标注审核、任务移交及文本 AI 预标注能力。

环境、供应链与权重安全

- **研发环境隔离：**华为云对自研模型研发以及 MaaS 控制范围内的模型处理环境实施研发、测试和生产环境隔离，通过项目、网络、计算、存储、身份和密钥等边界降低环境间非授权访问及凭据滥用风险。
- **模型供应链安全：**对华为云自研模型研发所引入的第三方库、框架、镜像和组件，以及拟进入 MaaS 托管环境的第三方模型和客户自定义模型，根据对象类型开展来源信息、完整性、恶意文件、漏洞、自定义算子及可执行代码等必要安全检查，降低不可信组件影响 MaaS 模型研发和运行环境的风险。
- **模型资产保护：**对华为云控制和托管范围内的模型权重、检查点和适配器实施访问控制、传输与存储保护、完整性校验和关键操作审计，降低模型资产在研发、部署和运行过程中被非授权访问、篡改或泄露的风险。

【华为云服务】**统一身份认证服务 IAM** 支撑用户、用户组、委托、项目隔离与最小权限控制。

【华为云服务】**数据加密服务 DEW** 可用于密钥托管、加密和轮换，降低权重、数据集及相关凭据被越权读取的风险。

【华为云服务】**云审计服务 CTS** 记录关键云资源管理操作，支撑模型资产和权限变更追溯。

4.3.1.2模型筛选与引入流程

对拟作为第三方预置模型提供，或由客户上传至 MaaS 托管环境的模型，华为云根据模型来源类型和平台控制范围实施相应的安全准入，包括候选模型登记、来源信息核验、完整性与供应链检查、安全评测以及风险分级和准入审批。对于可能影响 MaaS 平台或托管服务安全的风险，根据风险采取拒绝准入、要求整改或限制服务范围等措施。

其中，华为云自研模型按照内部研发、评测和发布流程完成安全准入；第三方预置模型由华为云实施相应的平台引入和准入管理；客户自定义模型则由华为云重点开展平台安全准入，其模型来源、许可及业务适用性仍由客户按照第三章责任边界负责。

以下流程主要适用于第三方预置模型和客户自定义模型的引入；华为云自研模型按照内部研发、评测和发布流程完成安全准入。

图 4-5 模型筛选与引入流程

下表描述华为云 MaaS 平台侧模型引入和安全准入的主要控制环节，具体责任仍以第三章责任边界为准。

流程环节	主要检查	输出
候选模型登记	模型名称、版本、来源、用途、责任人、目标用户和数据范围	候选模型记录
来源与许可审查	提供方身份、许可证、商业使用与再分发限制、地域及行业限制	许可审查结论
完整性与供应链检查	签名/哈希、模型文件格式、镜像和依赖漏洞、恶意代码、自定义算子	扫描报告和物料清单
模型安全评测	内容安全、隐私复现、鲁棒性、事实性、偏见公平性、越狱与模型抽取	评测报告和缺陷清单
风险分级与审批	业务影响、用户范围、数据敏感度、能力边界、补偿控制和残余风险	准入、整改、限制或拒绝决定
模型资产入库	绑定模型、权重、镜像、评测、模型卡、责任人及版本标识	可追溯模型资产

表 4-1 模型筛选与引入流程说明

4.3.2 模型部署与发布安全

对由华为云提供或托管的 MaaS 模型服务，华为云在模型上线前建立安全准入、风险评测、发布审批、版本管理、灰度验证、回滚和下线机制。发布对象与模型权重、推理环境、安全策略、评测结果和版本标识建立关联，确保模型部署和发布过程可识别、可验证、可追溯，并在发现重大安全问题时具备回退或下线条件。

4.3.2.1安全准入与评测

- 华为云根据模型类型、服务范围和风险，对拟提供或托管的 MaaS 模型服务开展相应的安全评测，并设置必要的发布条件、缺陷整改和残余风险管理要求。评测可覆盖内容安全、隐私、鲁棒性、事实性、偏见公平性和模型能力边界等维度。

- 红队测试覆盖越狱、直接及间接提示词注入、系统提示泄露、敏感信息复现、有害输出、模型抽取和多模态绕过；记录攻击库版本、覆盖率、缺陷等级和复测退出条件。
- 金融、医疗、政务、教育和未成年人场景实施专项测试，对面向高风险或强监管场景提供的模型服务，可结合实际服务范围开展相应专项评测和风险验证。
- 对华为云对外提供的模型服务形成相应的模型说明或服务说明，根据实际服务情况提供模型版本、主要能力、适用范围、已知限制及必要风险提示，为客户了解和选择模型提供基础信息。

4.3.2.2华为云业务实践

华为云 MaaS 按照模型来源实施差异化管理。自研模型、第三方预置模型和客户自定义模型均纳入统一的模型资产、评测和发布管理，但责任边界、检查深度和可开放能力根据来源及风险确定。

【华为云服务】MaaS 承载模型资产、在线推理和相关平台管理能力，为模型部署、发布和运行提供相应的平台支撑；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】云审计服务 CTS 记录 MaaS 及相关云资源的关键管理操作，支持发布与变更追溯。

4.3.2.3发布、灰度与回滚

- 模型发布经审批后，先采用小流量灰度或影子评估，对安全命中、误拦截、事实性、响应时延和故障率进行验证，达到退出条件后再扩大流量。
- 模型权重、推理镜像、系统提示或内容安全策略发生重大变更时，应重新开展安全评测；低风险展示性变更可采用简化流程。
- 保留已验证版本、配置和回滚路径，并定期验证流量切换、旧版本封禁和紧急下线能力。

4.3.3 模型运行与应用安全

模型进入在线推理和持续服务阶段后，华为云重点保障其控制范围内的模型、推理环境和 MaaS 模型服务安全。华为云通过模型原生安全能力、推理链路安全机制、平台身份与访问控制、租户和资源隔离以及持续的模型安全缺陷处置，降低模型运行过程中提示词攻击、异常输出、敏感信息泄露、非授权访问和模型安全缺陷等风险。

在整体 MaaS 模型应用链路中，华为云同时提供可供客户选择和配置的安全护栏及相关云安全能力；客户侧业务权限、安全策略和最终使用要求将在 5.2 节进一步说明。

下图展示模型从输入、推理到输出与应用执行的整体分层防护关系，其中华为云负责其控制范围内的平台及模型运行安全能力；图中的客户应用侧控制由客户按照第三章责任边界落实，相关内容将在 5.2 展开。

图 4-6 模型运行与应用安全分层防护

4.3.3.1内容安全

华为云 MaaS 在模型推理链路中提供华为云大模型防火墙及相应的输入输出安全能力，对进入模型的 Prompt、上下文和图片，以及模型生成的文本、图片等内容进行安全检测，降低恶意输入、违规内容、敏感信息泄露及异常模型输出对 MaaS 服务和客户业务造成的影响。相关能力可根据模型类型、应用场景、部署区域及实际服务能力进行配置。

图 4-7 大模型防火墙

在文本内容检测方面，华为云大模型防火墙可对输入和输出中的色情、暴恐、违禁、恶意及其他不符合内容安全要求的信息进行识别，并关注个人身份信息、设备信息、账号凭据及企业敏感信息等内容的暴露风险。针对多语言应用场景，可结合实际支持范围，对不同语种内容实施相应检测和处置，降低攻击者利用小语种或混合语言绕过安全策略的风险。

针对大模型特有的语义攻击，相关安全能力可识别直接提示词注入、越狱攻击、系统提示词探测、反向诱导、恶意指令及代码等风险。在 RAG、Memory 和 Agent 场景中，还需关注外部文档、知识库内容及长期记忆中隐藏的间接提示词注入，防止不可信内容影响模型行为，进一步诱导模型泄露敏感信息、绕过既定规则或触发非预期的工具调用。对于连接工具和业务系统的 Agent，可结合 Agent 安全能力识别恶意诱导、工具滥用及 Agent 劫持等风险，相关控制见 4.6 节。

在图片及多模态内容检测方面，华为云大模型防火墙可根据实际服务能力，对图片语义和图片中的 OCR 文字进行检测，识别图片中的违规内容、敏感信息及隐藏指令，并关注文本与图片组合形成的跨模态绕过风险。通过对不同模态内容进行综合检测，降低攻击者利用图片文字、视觉提示或模态差异规避单一文本检测的风险。

华为云 MaaS 可根据实际服务能力支持对模型输入和输出分别实施检测，并对流式响应内容进行持续检测，减少仅在完整响应生成后开展检测所带来的风险暴露。相关能力可通过 API 等方式接入模型调用链路，并结合 WAF、API 网关等入口安全能力，对传统 Web 攻击、恶意请求和模型语义风险实施分层防护。

华为云支持根据业务场景配置内容安全策略，对违规内容、提示词注入和敏感信息等检测类别设置相应处置动作，包括拦截、记录、脱敏或安全代答等。针对不同应用、模型和业务风险，可关联差异化策略集，并通过豁免规则及策略优化机制降低误报对正常业务的影响。客户应结合自身业务场景、用户群体和适用要求，合理配置检测类别、处置方式及业务侧人工复核机制。

内容安全检测结果可形成相应的安全日志和风险记录，记录命中的风险类别、检测方向和处置结果，并支持对不同攻击及内容风险进行分类统计。华为云可结合平台实际可观测范围，对检测数量、拦截数量和风险变化进行监测，为安全告警、事件调查和策略持续优化提供依据。

【华为云服务】华为云大模型防火墙可为文本、图片等内容提供安全检测能力，帮助识别违规内容及相关内容风险；具体支持的检测类别、语种、输入输出方式、处置动作和区域范围，以目标区域、控制台及正式产品文档为准。

4.3.3.2运行隔离与访问控制

华为云在 MaaS 平台侧对不同租户的模型资产、数据和推理资源实施必要隔离，并通过平台身份认证、模型服务访问控制和资源隔离机制，降低跨租户访问、非授权调用和资源相互影响等风险。对模型托管、在线推理及相关管理接口实施相应的服务访问保护，保障华为云控制范围内的模型运行环境和服务接口安全。

对客户账号、角色、API Key、网络访问策略、流量及配额等可配置项的安全管理，由客户根据自身业务需要配置，华为云可提供相应平台和云服务能力，具体将在 5.2 说明。

4.3.3.3模型运行安全边界

- 对华为云提供或托管的模型服务，当发现模型能力边界发生重大变化、出现严重安全缺陷或存在可能影响 MaaS 安全运行的重大风险时，华为云根据风险采取安全加固、调整防护策略、限制相关能力、回滚至已验证版本或下线相关模型服务等措施。模型相关安全问题纳入运营运维安全流程持续跟踪和处置。
- 对华为云提供和控制范围内的模型服务，退役时关闭相应服务入口，并按照平台资产和数据管理要求处理模型权重、适配器、运行镜像、缓存及其他托管资产，降低已下线模型继续暴露或被非授权使用的风险。

4.4 平台数据传输与存储安全

数据安全是华为云 AI 安全框架中的横向安全能力。从数据三态看，MaaS 数据安全覆盖传输态、存储态和使用态。根据第三章 MaaS 安全责任共担边界，华为云重点保障其控制范围内的数据传输链路和底层存储安全，通过安全传输、网络隔离、存储加密及密钥保护等能力保护传输态和存储态数据；对于进入计算和推理环境的使用态数据，华为云依托算力底座的租户、任务和运行环境隔离等能力保障其处理环境安全。客户对自身数据来源、授权、合规使用及具体数据保护策略承担相应责任，相关内容见 5.3 节。

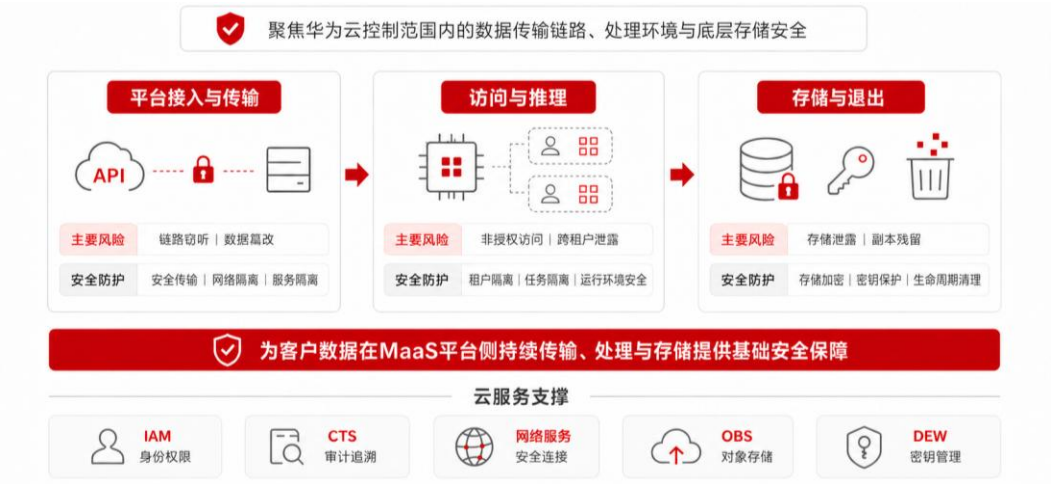

图 4-8 平台数据传输与存储安全

4.4.1 数据传输安全

客户数据通过 API、对象存储等方式进入华为云 MaaS，并在 MaaS 平台、存储服务以及相关训练和推理环境之间流转。华为云对其控制范围内的数据传输链路实施安全保护，通过安全通信协议、网络隔离和服务隔离等机制，保护数据在接入和平台内部流转过程中的机密性与完整性。

对不同部署场景，华为云提供相应的安全数据连接能力，支持客户根据实际架构通过公网加密链路或受控网络连接 MaaS 服务，并降低数据在跨网络、跨服务传输过程中的暴露风险。客户侧数据传输范围、数据内容以及具体使用要求由客户根据自身业务进行管理，具体见 5.3 节。

数据进入 MaaS 计算和推理环境后，由传输态进入使用态。华为云依托算力底座及 MaaS 平台提供的租户隔离、任务隔离和运行环境安全机制，保护数据在内存、显存、缓存及相关处理环境中的安全，降低不同租户和任务之间发生非授权数据访问的风险。相关算力底座安全能力见 4.2 节，本节不再重复展开。

【华为云服务】**统一身份认证服务 IAM** 提供身份认证和细粒度权限管理能力，可为 MaaS 相关用户和资源访问提供身份与权限控制支撑；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**云审计服务 CTS** 可记录相关云资源的关键管理操作，为数据接入、权限及资源变更提供操作追溯能力；具体可审计事件范围以正式产品文档为准。

【华为云服务】华为云网络服务可通过 HTTPS、**VPC 终端节点、VPN 或云专线**等方式为不同场景提供安全的数据连接能力，具体采用方式由实际部署架构确定。

4.4.2 数据存储与销毁安全

MaaS 在训练、评测、模型托管及客户显式配置的存储场景中，可能产生相应的任务数据和托管资产。华为云按照适用的数据留存规则和资源生命周期要求，对上述数据和资源实施清理及退出管理。对于 MaaS 原生模型推理，客户提示词和模型推理结果仅在推理过程中被处理，不形成平台持久化副本，并在推理完成后立即删除。MaaS 业务运行过程中可能产生缓存、临时文件、平台副本等数据。华为云按照 MaaS 数据留存规则及平台资源生命周期要求，对其控制范围内的数据和相关资源实施相应的留存、清理和退出管理，使平台侧数据按照既定规则完成生命周期处理。

华为云负责保障其控制范围内托管存储及底层存储环境的安全，并根据相应平台和存储服务能力对数据实施基础存储加密。客户对数据敏感等级、自身可配置的加密策略、保存期限和业务副本等负责，相关要求见 5.3 节。

【华为云服务】**对象存储服务 OBS** 可为 MaaS 相关数据提供对象存储及服务端加密等基础存储能力；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**数据加密服务 DEW** 提供密钥管理和加密相关能力，可为 MaaS 涉及的静态数据保护提供密钥安全支撑；具体支持范围以正式产品文档为准。

4.5 平台持续安全运营

MaaS 进入持续运行阶段后，华为云控制范围内的算力资源、MaaS 平台组件、模型服务及 Agent 相关基础服务持续发生资产、版本、配置和运行状态变化，安全风险也随平台演进和攻击手法持续变化。根据第三章 MaaS 安全责任共担边界，华为云围绕自身控制范围内的 MaaS 平台及托管组件开展持续安全运营，保障相关安全机制、监测能力和平台服务持续有效运行。

华为云从 AI 资产管理、风险监测与漏洞管理、安全日志与事件管理、安全合规与持续优化四个方面构建运营运维安全闭环，实现“资产可见—风险发现—事件处置—持续改进”，持续保障 MaaS 平台及其控制范围内组件的安全运行。

图 4-9 MaaS 持续安全运营

4.5.1 AI 资产管理

AI 资产管理是华为云开展 MaaS 持续安全运营的基础。随着 MaaS 平台持续迭代，华为云自身建设、运营及托管范围内的算力资源、平台组件、模型及版本、模型服务、API 接口和 Agent 基础服务等 AI 资产会持续新增、变更和下线，资产状态及依赖关系动态变化。华为云通过统一资产管理机制持续掌握相关资产的类型、版本、运行状态、责任主体、安全状态和关联关系，降低 AI 资产快速变化带来的管理盲区。

华为云对自身建设、运营和托管范围内的 AI 资产建立统一资产台账，重点覆盖算力及运行环境、MaaS 平台组件、华为云提供或托管的模型及版本、推理服务、平台 API、Agent 基础服务以及相关安全组件等。资产台账记录资产类型、责任主体、版本、部署状态、安全状态、风险等级及上下游依赖关系，为漏洞管理、风险监测、事件分析和安全策略实施提供统一资产基础。

华为云结合平台资产发现和资产识别能力持续更新 MaaS 平台资产信息。当相关资产发生新增、版本升级、配置变更、部署调整或下线时，同步更新资产状态，并建立算力资源、平台组件、模型服务、API 和 Agent 基础服务之间的依赖及调用关系，为风险评估和安全运营提供持续更新的资产视图。

对模型版本、平台组件、API 及相关 MaaS 服务建立版本和变更记录；对于重大版本升级、关键组件变更、高权限服务或其他可能影响安全边界的变化，根据风险触发相应的安全复核。对于发现的未登记组件、异常资产或未纳管服务，及时纳入资产管理范围并根据风险采取复核、限制或处置措施。

根据资产重要程度、服务权限、安全状态和业务影响实施差异化管理，对承载关键 MaaS 能力、高权限接口或重大业务影响的平台资产实施重点监测和安全管理。

图 4-10 AI 资产动态管理与风险监测

4.5.2 风险监测与漏洞管理

风险监测与漏洞管理用于持续发现和处置华为云控制范围内 MaaS 平台及托管组件运行过程中出现的安全风险。华为云结合平台资产信息、安全日志、漏洞信息、安全告

警和运行状态开展关联监测，持续识别平台异常访问、组件漏洞、新型 AI 攻击、内容合规风险和服务异常，并通过风险评估、告警、处置、修复和复核形成闭环。

华为云围绕自身控制范围内的算力底座、MaaS 平台、模型服务及 Agent 基础服务建立统一风险监测视图，关联相关资产、漏洞、访问行为、安全告警和运行状态。在算力底座和平台层持续关注计算节点、容器、镜像、网络及平台组件的漏洞、异常访问和资源异常；在模型及 Agent 相关平台能力层，根据平台实际可观测范围关注针对 MaaS 服务的异常请求、攻击行为、接口滥用，以及异常输出、有害或违规内容、敏感信息泄露等内容安全与合规风险，并形成相应安全告警，为风险定位和影响分析提供依据。

为应对 MaaS 调用关系复杂、风险信号分散的问题，华为云可在自身可观测范围内关联身份认证、平台访问、模型服务调用、组件运行、安全检测和漏洞等信息，将不同组件产生的安全信号关联至对应资产和服务，辅助识别异常活动的来源、影响范围及相关风险。对需要进一步调查的安全事件，由安全日志与事件管理机制开展调用链追溯和事件处置。

在内容合规方面，客户是内容合规的第一责任方，为支持客户对内容数据合法合规使用，华为云作为 Maas 平台提供方，根据适用法律法规、监管要求及《MaaS 模型即服务服务声明》，通过人工或技术方式对客户内容进行必要的安全合规审核和处置。平台可能对模型推理的输入输出内容，按涉政、涉黄、涉暴、涉诈等维度进行实时内容安全检测。当安全检测命中风险时，平台可在适用服务声明和产品能力范围内记录必要的非内容类事件元数据，例如风险类别、检测方向、处置结果、模型或服务版本、事件时间及命中次数，不记录客户输入输出正文及其内容摘要。若出现重大合规风险，平台将通过客户经理向租户推送内容合规风险通知，告知违规类型、触发次数和涉及模型，并附处置建议；同时平台侧自动对涉事模型实施内容安全策略增强，必要时临时限制高风险生成能力。

华为云对自身控制及托管范围内的计算节点、容器、镜像、AI 框架及相关 MaaS 平台组件持续开展漏洞发现和风险评估，并结合漏洞严重程度、可利用性、资产重要程度和服务影响进行分级处置。漏洞修复后开展必要验证；对于可能影响 MaaS 平台安全或服务连续性的高风险漏洞，及时采取缓解、修复、隔离或其他必要措施。

华为云结合新型攻击样本、漏洞信息、红队测试、安全事件和平台变化持续优化平台侧检测规则和安全防护策略，并在必要时开展策略验证，降低安全策略变更对正常 MaaS 服务的影响。

4.5.3 安全日志与事件管理

MaaS 业务调用可能跨越身份认证、平台接口、模型服务、数据处理和 Agent 基础组件，分散的日志和复杂调用关系会增加安全事件定位和影响分析难度。华为云在自身控制和可观测范围内建立平台侧日志记录、关联分析和安全事件响应机制，通过关联关键操作、安全告警和调用信息，提高 MaaS 平台安全事件的可观测、可追溯和可处置能力。

华为云根据平台实际可观测范围记录必要的身份认证、资源访问、模型及服务版本、配置变更、API 调用、安全检测、管理操作和安全告警等信息，并对日志中的敏感内容实施必要的最小化、脱敏、访问控制和安全保护。

当发生平台侧安全异常时，可结合租户、用户、模型、服务、时间和事件等维度关联相关日志，对关键调用路径进行分析，辅助定位事件来源、相关资产和影响范围，为事件调查和责任界定提供平台侧依据。

华为云建立覆盖告警、研判、分级、处置、恢复和复盘的安全事件响应机制。对发生在华为云控制范围内的漏洞利用、异常访问、平台攻击、服务异常等安全事件，根据风险采取限制访问、策略调整、组件修复、服务隔离或恢复等必要措施。对于涉及华为云与客户控制边界的事件，根据第三章责任共担原则，在双方实际控制和可见范围内开展必要的信息共享与协同处置。

图 4-11 MaaS 安全事件管理

4.5.4 安全合规与持续优化

安全合规与持续优化是华为云 MaaS 运营运维安全的闭环。华为云持续跟踪适用于自身 MaaS 平台和服务的网络安全、数据安全、个人信息保护、AI 治理及相关监管要求，并结合平台资产、漏洞、安全事件、攻击样本和服务变化持续评估自身安全控制的有效性。

华为云根据平台风险、重大版本变化、关键组件变更和适用监管要求，对 MaaS 平台及相关算力底座、模型服务和 Agent 基础服务开展必要的安全测试、风险评估、红队测试、合规检查或其他验证活动，并对发现问题开展整改和复核。

华为云结合漏洞信息、攻击样本、安全事件、红队测试和平台变化持续优化检测规则、安全策略和运营流程，使安全能力能够随 MaaS 平台和风险环境持续演进。

根据服务能力和适用范围，华为云可提供相应的安全评测、认证、审计等证明材料，为客户开展自身安全评估和合规审查提供支持。相关认证和证明材料在第七章统一说明。

4.6 Agent 平台基础安全保障

客户基于华为云 MaaS 进一步构建 Agent 应用后，Agent 可能通过平台接口调用模型及相关服务。根据第三章 MaaS 安全责任共担边界，华为云在 Agent 安全方面主要提供平台侧的基础身份鉴权、访问控制、配额和限流等安全组件，降低非授权访问、异常调用和 API 滥用对 MaaS 平台及托管服务造成的安全影响；Agent 应用本身的身份权限设计、知识库访问、Skill/MCP 及工具调用、业务执行和持续运营安全由客户负责，相关内容见 5.5 节。

华为云通过平台身份认证和访问控制机制，对 Agent 访问 MaaS 相关平台服务实施必要的身份校验，并结合平台接口的配额和限流机制约束异常或过量调用，降低凭据滥用、异常请求以及 API 资源消耗对平台安全性和服务稳定性的影响。对于华为云控制范围内发现的异常接口调用，可结合平台安全运营机制开展必要的监测和处置。

【华为云服务】**统一身份认证服务 IAM** 可提供身份认证和权限管理能力，为 Agent 访问相关华为云资源提供基础身份与访问控制支撑；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**AgentArts** 可提供 Agent 相关的平台接入、身份校验、访问控制、配额和限流等基础能力，为 Agent 调用 MaaS 及相关平台服务提供安全支撑；具体能力和支持范围以目标区域、控制台和正式产品文档为准。

5 华为云如何助力客户安全的使用 MaaS

5.1 客户使用 MaaS 可能面临的安全风险挑战

根据第三章 MaaS 安全责任共担边界，客户负责其在 MaaS 平台上配置、上传、开发和使用内容的安全。客户使用 MaaS 时，主要需要关注模型选择与应用、数据处理、客户侧持续运营以及基于 MaaS 构建 Agent 等方面的安全风险，并结合自身业务场景正确配置和使用华为云提供的安全能力。本节围绕客户责任范围梳理主要安全挑战，后续章节将分别说明华为云可提供的安全能力和支持。

图 5-1 客户安全风险挑战

模型安全风险挑战

客户使用华为云 MaaS 时，可根据业务需要选择第三方预置模型或部署客户自定义模型。不同模型在能力边界、适用场景、版本来源和安全风险方面存在差异，客户需要结合自身业务影响和使用方式评估模型适用性，并正确配置相应安全策略。

在模型选择与引入阶段，客户可能面临模型能力与业务需求不匹配、开源或自定义模型存在已知安全缺陷、模型及依赖组件来源不可信，以及开源模型许可要求与实际使用方式不匹配等风险；在模型配置和应用阶段，安全策略配置不足、访问权限过大、提示词注入、越狱、敏感信息泄露、有害或错误输出等问题可能进一步影响客户业务。随着模型版本更新和业务持续运行，模型能力边界、风险状态及适用性还可能发生变化，需要客户持续评估模型及其应用的安全影响。

阶段	主要风险	控制重点
研发与引入	模型能力与业务不匹配、开源/自定义模型安全缺陷、来源或依赖风险、许可协议不匹配	模型适用性评估、来源与版本确认、License 核查、必要的安全测试
部署与发布	安全策略配置不足、权限过大、能力边界不清、版本变更引入新风险	正确配置安全策略、权限控制、上线前验证、必要的额外防护
运行与应用	提示词注入、越狱、敏感信息泄露、有害或错误输出以及业务误用	输入输出安全控制、内容防护、业务规则、人工复核及使用边界
变更与退役	模型能力变化、旧版本持续使用、客户模型或凭据残留	变更后重新验证、旧版本停用、凭据撤销及客户模型资产清理

数据安全风险挑战

MaaS 显著扩大了客户数据的类型、规模和使用范围。客户提供或使用的数据既可能作为训练、微调和评测数据进入模型处理流程，也可能作为 Prompt、对话上下文和知识库数据参与在线推理，并进一步被 RAG、Agent 及外部工具调用。随着数据在不同模型、应用和外部系统之间持续流转，数据来源与授权不足、敏感信息泄露、越权使用、处理范围扩大以及删除不完整等传统风险被进一步放大，同时还可能出现数据投毒、模型记忆、越权检索及 Agent 数据外传等 MaaS 场景下更加突出的风险。

客户首先面临数据来源、授权和使用目的风险。评测数据、知识库、用户上传文件和外部数据可能存在来源不明、授权不足、敏感信息混入、质量不足或恶意投毒等问题。知识库文档在解析、切分和向量化过程中还可能丢失原有权限属性，使原本受限的数据在后续检索中被非授权访问。因此，客户需要重点关注数据的合法来源、授权范围、业务用途、敏感程度和输入数据安全。

数据会在模型调用、知识检索和应用处理过程中流转。 客户侧权限、存储和传输配置不当，可能导致敏感数据被非授权访问、错误共享或泄露；对于 RAG 和 Agent 场景，知识库权限设置不当、Memory 长期保留以及 MCP 或外部工具调用还可能进一步扩大数据访问和外传范围。客户需要结合数据敏感程度落实输入输出过滤、脱敏、访问权限以及自身控制范围内的数据加密和流转管理。

数据使用完成后，还可能存在关联副本和派生数据残留风险。除原始数据外，客户还需要关注数据集版本、知识库文档、向量索引、会话上下文、Memory 以及客户控制范围内的其他副本；已经参与模型训练的数据还可能持续影响模型行为。若未结合实际数据位置、保存期限和业务要求完成关联清理，已不再需要的数据可能继续被访问或使用。因此，客户需要根据自身数据生命周期要求管理相关数据的留存、删除和退出。

运营运维安全挑战

客户将 MaaS 持续用于生产业务后，需要持续关注自身控制范围内的账号、模型调用、内容输出和业务应用安全状态。账号权限配置不当、访问凭据泄露、模型调用量或来源异常、安全策略未及时更新等问题，可能导致权限滥用、非预期模型调用；模型输出和客户应用中的内容风险如果缺乏持续监测，也可能在业务运行过程中持续积累和扩大。

同时，客户侧模型、应用和 Agent 的版本、配置以及业务场景会持续变化，仅依赖上线前检查难以覆盖后续风险。若未根据自身业务需要配置必要的日志、告警和安全策略，发生异常调用、内容风险或应用安全事件后，可能难以及时识别、调查和处置。因此，客户需要持续监测自身账号、调用和业务安全事件，及时处理权限滥用、异常调用和内容风险，并在涉及 MaaS 平台的事件中配合华为云开展必要的调查、恢复和整改。

Agent 安全风险挑战

客户基于华为云 MaaS 进一步连接企业知识库、Memory、Skill、MCP 服务和业务系统构建 Agent 后，安全边界将由模型输入输出扩展至身份授权、知识访问、工具调用和业务执行。Prompt 注入、越狱和模型能力偏差可能影响 Agent 的任务理解与规划，并进一步转化为非预期执行行为；Agent 连接 MCP、Skill 和外部 API 后，恶意工具、过度权限或危险参数还可能造成越权访问、错误操作或资源滥用。

同时，RAG、Memory 和上下文中的敏感信息可能因权限控制不足、恶意内容或外部调用发生越权检索和数据外泄；多 Agent 协同及第三方组件还会增加身份、权限和供应链关系的复杂度，形成权限传递、调用链失控和责任难以追溯等风险。因此，客户基于 MaaS 构建 Agent 时，需要重点关注 Agent 身份权限、知识库访问、Skill/MCP 及工具调用以及实际执行行为的安全，并在后续开发、部署和运营过程中持续管理相关风险。

针对上述风险，华为云分别从模型、数据、运营运维和 Agent 等方面提供相应的 MaaS 安全能力，帮助客户在自身责任范围内加强安全管理，具体将在 5.2～5.5 节进一步说明。

5.2 模型选择与使用安全

客户使用华为云 MaaS 时，需要根据自身业务目的、风险等级和使用方式选择合适的模型，并对自身可控制的模型配置、模型应用及最终业务使用承担安全责任。对于第三方或开源模型，客户还需要关注模型自身安全缺陷、许可条件和业务适用范围；对于客户自定义模型，则需要对模型来源、版本及相关组件进行管理。

华为云通过模型说明、安全评测、模型托管、访问控制、大模型安全护栏及相关云安全能力，帮助客户识别模型风险、配置必要的安全控制并持续安全使用 MaaS 模型服务。华为云提供的平台侧模型安全保障见 4.3 节，本节重点说明客户如何利用相关能力履行自身模型安全责任。

5.2.1 模型选择与适用性管理

客户在选择华为云 MaaS 提供的模型时，应结合业务目的、用户范围、数据敏感程度和潜在业务影响，综合评估模型的能力边界、适用场景和已知限制，避免仅依据模型通用能力或性能指标直接应用于高风险业务。华为云可通过模型说明、服务说明及安

全评测等信息帮助客户了解不同模型的能力和风险，客户仍需要结合自身业务完成最终适用性判断。

对于 MaaS 提供的第三方或开源模型，客户还应关注模型版本、已知安全缺陷及相关使用条件；如模型涉及开源许可要求，应确认其商业使用、地域、行业或其他使用限制与实际业务场景相符。对于模型自身已知但暂无法消除的安全风险，客户应结合业务影响采取必要的使用限制或补充防护措施。

客户可结合华为云提供的模型说明、安全评测结果、适用范围和已知限制等信息开展模型选型。对于金融、医疗等高风险业务，还应结合自身场景开展必要的业务适用性验证，不应仅以模型通过平台通用安全评测作为业务上线的唯一依据。华为云针对 MaaS 所提供模型实施的平台侧安全管理和保障机制见 4.3 节。

5.2.2 模型配置与安全防护

客户在确定使用模型后，需要根据业务风险正确配置模型访问权限和安全策略。华为云提供相应的身份认证、访问控制、模型安全护栏及相关云安全能力，客户可结合人员、应用和模型服务的实际使用方式实施最小权限管理，并合理限制模型调用、安全策略配置及其他高权限操作的授权范围。

图 5-2 模型配置与安全防护

对模型相关权限，客户应区分模型调用、API Key 管理、安全策略配置及其他高权限操作，避免单一账号拥有不必要的权限；对于 API Key 等访问凭据，应进行安全保存、权限限制和生命周期管理，避免将凭据明文放置于前端代码、Prompt 或非受控日志中。客户还可根据实际业务容量配置合理的调用频率、并发、Token 或其他可用的资源限制，降低异常调用和凭据滥用造成的业务及资源风险。

对于所选开源模型自身存在且客户无法直接消除的安全风险，客户应结合业务场景评估其影响，并根据需要配置额外防护措施。华为云可提供大模型安全护栏、访问控制及其他适用安全能力，帮助客户降低提示词攻击、异常调用和敏感信息泄露等风险；相关安全策略仍需由客户结合自身业务需求进行配置和验证。

【华为云服务】**统一身份认证服务 IAM** 可帮助客户对 MaaS 相关用户、角色和服务身份实施认证与最小权限管理。

【华为云服务】**MaaS** 可在推理链路中提供可配置的大模型安全护栏，客户可根据实际业务场景和目标区域支持能力配置相应安全策略。具体支持范围以控制台和正式产品文档为准。

5.2.3 模型应用与持续使用安全

模型进入实际业务应用后，客户需要结合应用场景、服务对象、数据敏感程度和潜在业务影响，对模型输入输出、安全策略和最终使用方式进行持续管理。华为云 MaaS 提供相应的模型输入输出安全、身份与访问控制等能力，帮助客户降低提示词攻击、有害内容、敏感信息泄露和异常调用等风险；客户负责根据自身业务需求正确配置相关安全策略，并对模型输出的最终业务使用负责。

图 5-3 模型应用与使用安全

5.2.3.1输入输出与应用安全

客户应根据自身应用场景和风险水平配置适当的模型输入输出安全策略，重点关注提示词注入、越狱、有害或违规内容、敏感信息泄露以及多模态输入中的隐藏指令、OCR 内容和跨模态绕过等风险。华为云提供的 MaaS 大模型安全护栏等能力可帮助客户对相关风险进行识别和处置，客户应结合业务要求验证拒答、阻断、脱敏或安全改写等策略是否符合实际使用需要。

对模型输出进一步进入客户业务系统、自动化流程或其他应用组件的场景，客户还应结合输出类型和业务影响实施必要的结构校验、权限控制和安全验证，避免未经验证的模型输出被直接作为高风险业务操作或可信指令执行。

【华为云服务】大模型防火墙可帮助客户对模型输入输出中的提示词攻击、有害或违规内容、敏感信息等风险进行检测和处置，并可以支撑西班牙语、葡萄牙语等多种小语种，具体能力范围以正式产品文档为准。

5.2.3.2业务使用与人工监督

华为云提供的模型安全能力能够帮助降低模型输入、推理和输出过程中的安全风险，但不能替代客户自身的业务规则、事实核验和必要的人工决策。客户应根据模型能力边界和具体业务风险确定模型输出的使用方式，并对最终业务结果负责。

对涉及人身、财产、公共利益或其他重大权益的业务，客户应根据实际风险设置必要的人工复核、权限限制、安全降级或其他控制措施，避免模型输出未经验证直接触发重要业务决策或高风险操作。对于模型生成内容的合法合规使用，客户还应结合自身业务场景和适用要求进行判断和管理。

5.2.3.3模型变更与退出管理

客户应及时关注华为云 MaaS 发布的模型上线、版本更新、能力调整、下线及相关安全公告，了解当前所使用模型的版本和服务状态。当模型发生重大版本变化、能力边界调整或安全状态变化时，客户应结合自身业务影响重新验证模型适用性、输入输出安全策略及相关业务控制是否仍然有效，并根据验证结果及时完成模型升级或调整使用方案。

当客户需要在华为云 MaaS 内更换所使用的模型时，可通过应用或服务配置更新所选择的模型，并对新模型的能力、输出表现和既有安全策略进行必要验证，确保模型切换后仍满足业务和安全要求。

当客户不再使用华为云 MaaS 并切换至其他 MaaS 服务商时，应根据新的服务接口调整应用调用配置，包括更新 API 名称、接口路径及相关调用参数，并及时停用不再使用的华为云 MaaS 访问凭据和相关配置，避免旧接口或凭据持续暴露。新的 MaaS 服务上线前，客户仍应按照自身业务和安全要求重新完成模型适用性和相关安全配置验证。同时，客户还应根据业务需要完成相关数据和资产的导出、备份与迁移，并在确认迁移完成后清理不再使用的资源和访问凭据；相关数据导出、删除及生命周期管理要求见 5.3 节。

5.3 客户数据生命周期安全

数据安全贯穿 MaaS 的数据采集、传输与存储、处理与加工及销毁全过程。根据第三章 MaaS 安全责任共担边界，客户承担自身数据安全的主要责任，需要对进入 MaaS 的数据来源、授权和合规使用负责，并结合业务场景落实输入输出数据过滤、敏感数据保护、访问控制、加密及生命周期管理等措施。

华为云在保障 MaaS 平台传输链路和底层存储安全的基础上，同时提供身份权限、数据识别与脱敏、加密存储、密钥管理、审计追溯等安全能力，帮助客户围绕算力底座、模型和 Agent 中的数据使用场景落实自身数据安全责任。华为云平台侧的数据安全保障见 4.4 节，本节重点说明客户如何安全管理和使用其数据。

图 5-4 客户数据生命周期安全

5.3.1 数据采集安全

客户将数据接入或用于 MaaS 前，应明确数据来源、授权范围和使用目的，确保数据获取及后续使用符合适用的法律法规、合同约定及自身数据管理要求。对于来源不明、授权不足、超出既定用途或包含非必要敏感信息的数据，应在进入 MaaS 前进行必要处理。在业务允许的情况下，应避免直接向模型提交可识别个人身份的信息；确因业务需要处理个人信息时，应根据处理目的、数据敏感程度和适用要求，在数据进入 MaaS 前采取必要的脱敏、去标识化或其他保护措施，并避免将与业务目的无关的个人信息提供给模型。

客户应根据数据类型和敏感程度开展必要的数据识别、分类分级和敏感信息检查，并按照最小必要原则控制进入 MaaS 的数据范围。对于训练、评测、知识库及其他业务数据，还应关注数据质量、完整性和可信性，避免低质量、错误或恶意数据影响后续模型和应用使用。

在将数据接入 MaaS 前，客户还应按照最小必要原则控制实际上传和使用的数据范围，并根据业务需要对不同数据集、知识库和使用场景实施相应的访问授权。华为云提供的数据治理和访问控制能力可帮助客户完成上述管理，但数据来源合规、授权管控、数据用途及最终使用责任仍由客户承担。

【华为云服务】**数据安全中心 DSC** 可用于数据资产发现、分类分级、敏感数据识别和脱敏，并可帮助客户识别拟用于模型训练、评测或知识库的数据中的敏感信息，降低敏感数据被不当使用的风险；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**大模型防火墙**可帮助客户识别模型输入输出中的个人数据、敏感信息及相关内容风险，客户应根据业务要求配置并验证相应安全策略；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**统一身份认证服务 IAM** 提供身份认证和权限管理能力，可帮助客户控制 MaaS 相关数据、知识库及资源的访问范围；具体支持范围以正式产品文档为准。

5.3.2 数据传输与存储安全

客户数据在使用 MaaS 过程中，可能在客户业务系统、华为云 MaaS 服务、存储服务及相关应用之间持续流转，并以业务数据、评测数据、知识库数据、模型上下文、向量数据、会话记录和日志等不同形态进行传输和保存。由华为云控制的 MaaS 数据传输链路和托管存储环境，其基础传输和存储安全由华为云负责，具体见 4.4 节；对于客户自行创建、配置和控制的网络连接以及其他存储资源，客户负责根据数据敏感程度合理配置访问权限、加密策略、密钥、数据副本和保存期限。

在数据传输方面，客户应根据自身部署架构选择适当的安全连接方式，并控制实际传输的数据范围；对于跨网络或连接外部系统的数据，应避免传递与业务无关的敏感信息。在客户控制的存储环境中，应根据数据等级配置访问权限和加密措施，并合理管理数据版本、副本及生命周期。

对模型推理使用过程中形成的缓存、向量数据、会话上下文、Memory、日志以及其他派生数据，客户还应识别其实际存储位置和访问范围，并根据业务需要设置相应的保存期限和保护要求，避免原始数据受到保护但派生数据长期暴露。

【华为云服务】**对象存储服务 OBS** 可用于存放客户自行管理的数据集、评测集及其他 MaaS 相关数据。对于客户自行创建和控制的 OBS，客户负责配置桶策略、访问权限、

加密、版本控制和生命周期规则；由华为云 MaaS 平台控制的托管存储，其平台侧存储安全责任见 4.4 节。具体支持范围以正式产品文档为准。

【华为云服务】**数据加密服务 DEW** 提供密钥管理和加密相关能力。对于客户控制的数据和存储资源，客户可结合数据敏感程度和业务要求配置相应加密措施，并管理客户侧密钥的访问和生命周期；具体支持范围以正式产品文档为准。

【华为云服务】在网络层面可通过 HTTPS 协议，或 **VPC 终端节点、VPN、云专线**等服务为不同业务场景提供安全连接能力，客户可结合实际部署架构选择适当的数据传输方式；具体支持范围以目标区域和正式产品文档为准。

5.3.3 数据处理与加工安全

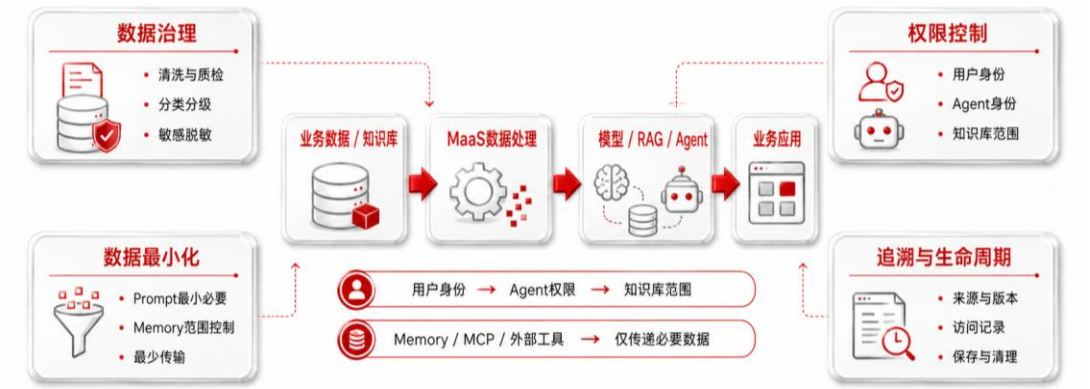

图 5-5 数据处理与加工安全

数据进入 MaaS 后，客户应结合数据质量和合规要求开展必要的数据清洗和治理，包括识别并处理错误、重复、异常或不适当数据，对缺失或不完整信息进行必要修正，并对敏感信息、个人数据及不合规内容实施过滤、脱敏或其他保护措施，降低不当数据进入后续模型处理和应用环节的风险。

需要开展数据标注的场景，客户应对标注范围、标签内容和标注结果进行必要管理，并保留相应的版本和过程记录。数据用于模型评测、推理或其他 MaaS 应用前，客户还应结合实际场景开展必要的数据安全和质量评估，重点关注数据完整性和准确性、数据偏差、内容合规性以及标注结果准确性；对于不满足要求的数据，应进行必要的清洗、修正或停止使用。

数据进入实际使用阶段后，客户应按照最小必要原则控制数据访问和使用范围，并根据数据敏感程度采取权限限制、脱敏以及必要的输入输出数据安全过滤措施。在 RAG 和 Agent 场景中，应根据当前用户和 Agent 身份限制知识库检索范围，避免不同部门、用户或密级数据发生越权访问；对于 Memory、MCP 及外部工具调用，应限制实际发送的数据范围，避免完整上下文或非必要敏感信息被传递至第三方。

客户还应保留必要的数据来源、版本、处理及访问记录，支持数据使用过程追溯，并结合权限控制、脱敏等机制减少原始敏感数据的直接暴露。数据保存期限、到期清理和相关审计要求在 5.3.4 进一步说明。

【华为云服务】**数据安全中心 DSC** 可帮助客户识别 MaaS 处理数据中的敏感信息，并结合实际场景实施分类分级、敏感数据识别和脱敏等措施；具体支持范围以正式产品文档为准。

【华为云服务】**大模型防火墙**可帮助客户识别模型输入输出中的敏感信息及相关内容风险，客户应根据业务要求配置并验证相应安全策略；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**AgentArts** 可提供知识库、Agent 及相关资源管理能力，帮助客户管理知识数据访问和 Agent 应用中的数据使用范围；具体支持范围以正式产品文档为准。

【华为云服务】**数据治理中心 DataArts Studio** 可提供数据质量管理等能力，帮助客户开展数据清洗、质量检查及相关数据治理；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**ModelArts** 提提供数据集管理、标注及相关数据处理能力，为客户开展数据准备、标注和质量管理提供平台支撑；具体支持范围以目标区域、控制台和正式产品文档为准。

5.3.4 数据销毁安全

MaaS 数据在使用过程中可能形成原始数据、数据集版本、知识库、向量索引、会话上下文、Memory 及其他派生数据。客户在变更 MaaS 使用方式、迁移业务或退订服务前，应根据业务连续性、数据保存和合规要求，识别需要保留的数据及相关资产，并结合实际产品能力完成必要的数据导出、备份或迁移，同时验证导出数据的完整性和可用性。

对于客户可控制的数据、模型、知识库及相关资源，华为云可通过 MaaS 及相关云服务提供相应的数据访问、下载、导出或迁移能力，帮助客户完成服务退出前的数据准备。客户应在退订前完成必要的数据保留和迁移，并及时停用不再使用的 API Key、访问权限及相关配置。

在完成数据迁移或业务退出后，客户应根据数据使用目的、合规要求和敏感程度，对其控制范围内的数据及副本进行清理。华为云按照适用的服务规则和数据生命周期机制，对其控制范围内不再需要保留的客户数据及相关平台侧数据实施清理或删除；对于客户自行创建或控制的 OBS、数据库、知识库、日志及其他关联资源，仍需由客户结合实际数据位置完成相应清理。具体可导出的数据类型、格式、保留期限和删除机制，以客户实际使用的产品能力、服务条款及正式产品文档为准。

【华为云服务】**对象存储服务 OBS** 支持对象生命周期和版本管理等能力，客户可根据自身数据保存要求配置生命周期规则，对客户控制的存储数据实施到期处理；具体支持范围以正式产品文档为准。

【华为云服务】**ModelArts、AgentArts** 提供相关数据集、知识库及资源管理能力，可帮助客户结合实际数据位置开展相应生命周期管理；具体支持范围以正式产品文档为准。

【华为云服务】**云审计服务 CTS** 可记录相关云资源的关键管理操作，为数据生命周期管理的审计与追溯提供支撑；具体可审计范围以正式产品文档为准。

5.4 运营运维安全

客户将 MaaS 持续用于生产业务后，需要对自身控制范围内的账号、模型调用、安全配置、内容风险和业务应用安全状态进行持续运营。客户负责持续监测客户侧账号异

常、模型调用、内容风险和业务安全事件，并根据自身业务需要配置日志、告警和安全策略，及时处置权限滥用、异常调用、内容风险及应用安全事件。

华为云在保障 MaaS 平台持续安全运营的基础上，通过资产及调用信息、身份权限、审计日志、监控告警和内容安全等能力，帮助客户开展自身责任范围内的 AI 资产管理、风险监测、安全事件处置和持续合规管理。华为云平台侧运营运维安全保障见 4.5 节，本节重点说明客户如何利用相关能力持续安全使用 MaaS。

在此基础上，客户可结合华为云提供的 AI 安全态势管理和智能化安全运营能力，将 MaaS 相关资产、风险、日志、告警和安全事件进行关联管理，并借助安全智能体提升资产识别、风险发现、威胁研判和处置策略生成效率，形成“资产可见、风险可查、事件可溯、处置可闭环”的持续运营机制。

图 5-6 客户侧运营运维安全

5.4.1 AI 资产管理

客户使用 MaaS 后，应对自身选择、配置、创建和使用的 AI 相关资产建立持续管理机制，明确资产责任人、业务用途、版本、权限、运行状态及关联关系。随着客户持续增加模型服务、接入点、数据集、知识库和业务应用，相关资产和配置也会动态变化，客户需要及时掌握资产新增、变更和退出情况，避免出现无人负责、长期闲置或超出预期使用范围的未纳管资产。

客户应建立自身 MaaS 资产台账，至少覆盖所使用的模型服务及版本、服务接入点、API Key 等访问凭据，并根据实际使用场景进一步纳入知识库、Agent、Skill/MCP 和相关外部连接。客户可记录资产名称、业务用途、版本或状态、责任主体、使用范围及关联关系，为后续风险检查和变更管理提供依据。

为提升 MaaS 资产发现和关联管理效率，客户可结合华为云提供的安全指纹智能体能力，辅助识别模型服务、Agent、知识库、向量数据库、Skill/MCP、API 及相关连接等 AI 资产，并结合资产类型、版本、运行状态、权限配置和调用特征形成资产指纹。通过资产之间的调用和依赖关系，可进一步辅助识别未纳管、闲置或异常资产，并及时

纳入客户自身 MaaS 资产台账和后续安全管理范围。参考方案中的安全指纹智能体即用于 AI 资产自动发现、属性采集和资产指纹台账建立。

客户应定期根据 MaaS 平台中的实际资源、配置和调用情况与自身资产台账开展核对。当模型版本发生变化、增加或停用服务接入点、创建或删除 API Key、调整知识库或业务应用配置时，应同步更新资产信息。对已不再使用的模型服务、接入点、API Key 及其他客户控制资产，应及时停用或清理，避免闲置资产和历史凭据持续暴露。

对关键 MaaS 资产，客户还应结合业务容量建立正常使用基线并配置合理的调用配额。对于高权限 API Key、涉及敏感数据的知识库、高影响业务应用以及其他高风险资产，可根据业务重要程度实施更严格的权限、调用和变更管理，并在发生异常使用时优先开展检查和处置。

【华为云服务】**云审计服务 CTS** 可根据实际支持范围记录 IAM 授权变更、模型服务开通或退订、接入点变更、API Key 创建或删除等关键管理操作，为客户开展 MaaS 资产变更追溯和台账核对提供支持；具体可审计事件范围以目标区域、控制台和正式产品文档为准。

5.4.2 风险监测与漏洞管理

客户将 MaaS 持续用于生产业务后，应对自身控制范围内的账号、模型调用、内容输出和业务应用安全状态开展持续监测，避免仅依赖上线前检查判断业务是否安全。客户应结合自身业务正常使用模式建立必要的监测指标和风险基线，对偏离正常状态的行为及时识别和调查。

在账号和凭据方面，客户应关注异常登录、权限变化、API Key 异常使用以及其他可能表明账号或凭据被滥用的风险；在模型调用方面，可结合调用量、调用频率、调用来源、调用时段、模型及接入点等维度识别异常模式，并关注调用量突增、非预期来源或异常模型调用造成的安全和费用风险。

在内容和模型使用方面，客户应根据自身业务场景持续关注提示词注入、越狱、有害或违规内容、敏感信息泄露等风险，并结合华为云提供的模型输入输出安全能力和客户应用侧安全策略开展监测。在客户自行构建 Agent、连接第三方组件或将模型能力集成至业务应用的场景，还应关注 Agent 行为偏移、异常工具调用、知识库越权访问、第三方组件异常以及相关业务安全事件。

在持续风险监测基础上，客户可结合华为云提供的智能化安全运营能力提升风险发现和分析效率。漏洞挖掘智能体可辅助针对客户控制范围内的 AI 应用、Agent、插件及相关组件开展风险发现和安全检查；安全观测智能体可持续关联模型调用、Agent 行为、工具调用、数据访问和其他安全信号，辅助发现偏离正常使用基线的异常活动；威胁研判智能体可对多源告警进行聚合、去重和关联分析，辅助还原风险链路、判断影响范围和风险等级；策略生成智能体可根据研判结果形成相应的安全策略、权限调整、风险整改或应急处置建议，帮助客户提高风险处置效率。客户仍应结合自身业务场景和实际控制范围，对相关风险和处置措施进行最终判断和落实。上述四类能力与参考方案中“风险发现—持续观测—威胁研判—策略生成”的智能化运营链路一致。

对于客户控制范围内的应用代码、Agent、Skill/MCP、第三方插件及相关组件，客户应持续关注其漏洞及安全更新，并结合漏洞严重程度、实际暴露情况和业务影响及时开展升级、修复、限制访问或其他必要处置。对于华为云 MaaS 平台及托管组件本身的漏洞发现和修复，由华为云按照第四章相应责任进行管理。

当发现账号异常、调用异常、内容风险或业务安全事件时，客户应结合资产、身份、调用和业务信息开展初步分析，并根据风险及时采取限制访问、权限收敛、API Key 轮换、调用限制、安全策略调整等措施，防止风险进一步扩大。

【华为云服务】**安全云脑**：云原生统一安全运营，基于风险管理全生命周期闭环，提供漏洞挖掘、安全观测、威胁研判、策略生成四大智能体，帮助客户主动排查风险，持续检测模型调用及业务访问；具体支持范围以目标区域、控制台和正式产品文档为准。

5.4.3 安全日志与事件管理

客户应根据自身业务风险、审计要求和事件响应需要，对客户控制范围内的 MaaS 相关操作、模型调用、安全检测和业务应用日志进行必要记录和管理，并合理设置日志保存期限、访问权限和保护措施。日志中可能包含账号信息、Prompt、模型调用内容或其他敏感信息，客户还应结合实际情况采取相应保护措施。

在日常运营中，客户应关注与账号登录、权限和 API Key 变更、模型服务调用、知识库访问、Agent 及业务应用运行相关的日志和安全告警。当出现异常登录、调用量突增、异常来源、内容风险或业务安全事件时，应结合相关日志、告警和资产信息开展调查，确认事件来源、影响范围和处置要求。

对需要进一步分析的安全事件，客户可结合统一安全运营平台及智能化分析能力，将相关日志、告警、资产和风险信息进行关联，辅助还原事件链路、判断影响范围，并根据实际情况开展处置和复盘。对于涉及华为云平台、客户应用或第三方组件的跨边界事件，双方按照第三章确定的责任边界和实际控制范围开展必要的信息共享和协同处置。

客户应保留必要的日志、告警、调查和处置记录，用于后续事件复盘、安全评估和合规审计，并对日志访问实施最小权限和必要保护，避免敏感日志本身成为安全风险。

【华为云服务】**云审计服务 CTS** 可根据实际支持范围记录 IAM 授权变更、模型服务开通或退订、接入点变更、API Key 创建或删除等 MaaS 管理面关键操作，帮助客户开展操作审计和事件追溯；具体可审计事件范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**对象存储服务 OBS** 可用于集中保存客户转储的审计日志，客户可根据自身安全和合规要求配置访问权限、保存期限及相应保护措施；具体支持范围以正式产品文档为准。

【华为云服务】**云监控服务 CES** 可帮助客户针对相关服务指标配置监控和告警，并将关键异常及时通知指定责任人；具体支持指标、告警方式和区域范围以正式产品文档为准。

【华为云服务】**安全云脑**可帮助客户结合实际接入范围，将 MaaS 相关安全信号纳入统一安全运营视图，辅助开展多源告警关联、风险研判、事件调查和协同处置；具体支持范围以目标区域、控制台和正式产品文档为准。

5.4.4 安全合规与持续优化

客户应结合所在国家或地区、所属行业、数据类型、用户群体和 MaaS 具体应用场景，持续识别适用于自身业务的网络安全、数据安全、个人信息保护、生成式人工智能及行业监管要求，并将相关要求落实到模型使用、数据处理、账号权限、日志管理、内容安全和业务应用等客户控制范围内的安全措施中。

客户应根据自身风险和合规要求定期开展 MaaS 使用安全评估，并在模型发生重大变化、业务场景发生重大调整、关键权限或数据处理方式变化，以及发生重大安全事件后，根据实际风险重新开展必要评估。评估可重点关注模型适用性、数据合规使用、访问权限、安全策略、日志告警、事件响应以及客户应用和 Agent 相关控制措施的有效性，并对发现问题进行整改和复核。

客户应根据业务和审计要求留存必要的安全评估、配置变更、事件记录、关键安全配置和证明材料，包括适用的安全评估记录、关键安全配置记录、审计日志、告警和事件处置记录，以及华为云提供的相关产品证明和服务材料。

客户还可将风险监测、事件调查和处置复盘过程中形成的告警、风险信息和整改经验，用于持续调整自身监测基线、安全策略和运营流程，使安全控制能够随 MaaS 业务和风险变化持续优化，形成“监测—研判—处置—优化”的持续改进闭环。

【华为云服务】**云审计服务 CTS** 可为客户提供支持范围内的云资源关键管理操作记录，帮助客户形成操作审计和安全评估证据；具体可审计事件范围以正式产品文档为准。

5.5 Agent 安全

客户基于 MaaS 构建 Agent 后，模型能力将进一步连接知识库、Memory、Skill、MCP、API 及业务系统，AI 应用也由内容生成进一步扩展至工具调用和业务执行。根据第三章 MaaS 安全责任共担边界，客户负责 Agent 应用安全开发、部署和运营全生命周期安全，需要重点管理 Agent 身份权限、知识访问、工具与外部系统连接、业务执行行为以及运行审计，防止 Prompt 注入、权限过大、越权检索、恶意工具调用和非预期业务操作等风险。

华为云提供基础鉴权、访问控制、配额、限流以及相关数据、模型和安全运营能力，帮助客户落实 Agent 安全控制；相关平台基础安全能力见 4.6 节。本节重点说明客户基于华为云 MaaS 构建和使用 Agent 时应落实的安全要求。

5.5.1 Agent 身份与访问控制

客户构建 Agent 时，应区分自然人、业务应用、服务账号和 Agent 等不同身份，并根据实际业务为 Agent 配置独立、可识别、可追溯的身份，避免多个 Agent 或业务场景长期共用高权限账号或访问凭据。Agent 访问模型、知识库、API、MCP 及其他工具时，应按照最小权限原则配置所需权限，使其仅能够访问完成当前业务任务所必要的资源和操作。

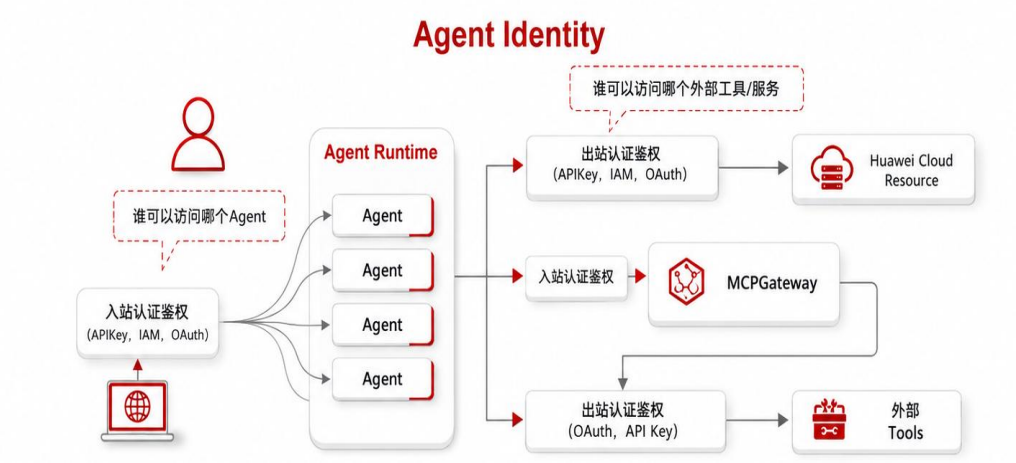

图 5-7 Agent 身份与访问控制

客户应结合 Agent 业务角色、数据敏感程度和操作影响建立权限边界，对模型调用、知识库检索、工具使用和外部系统访问分别实施必要的授权。高权限 Agent 以及能够执行数据修改、交易、审批、消息发送、系统配置等高风险操作的 Agent，应进一步限制其授权范围，并根据业务风险设置用户确认、人工审批或其他权限提升机制。

Agent 身份、权限、访问凭据及关联资源应纳入生命周期管理。当 Agent 用途、责任人、连接系统或权限发生变化时，应及时重新评估并调整权限；Agent 下线或不再使用相关工具和外部系统时，应及时撤销对应权限和凭据，降低历史授权持续有效的风险。

【华为云服务】**统一身份认证服务 IAM** 可帮助客户对相关华为云资源实施身份认证和权限控制，并按照最小权限原则配置 Agent 可访问的资源范围；具体支持范围以正式产品文档为准。

【华为云服务】**AgentArts** 可提供 Agent 身份及相关资源权限管理能力，帮助客户管理 Agent 与平台资源之间的访问关系；具体支持范围以目标区域、控制台和正式产品文档为准。

5.5.2 RAG 与企业知识库安全

客户将企业文档、业务数据或外部内容接入 Agent 知识库时，应明确知识来源、适用范围和访问权限，并尽可能保留原有用户、部门、项目、租户和密级等权限属性。在 RAG 检索过程中，应同时结合当前用户身份和 Agent 身份实施访问控制，确保 Agent 仅检索当前请求主体有权访问的知识内容，降低跨用户、跨部门或跨密级越权检索风险。

对于网页、用户上传文件或其他外部内容，客户还应关注提示词注入、恶意指令和不可信数据对 Agent 行为的影响，避免外部内容被错误解释为系统指令并进一步触发知识泄露或非预期操作。知识库内容、权限或授权状态发生变化时，应及时同步更新相关检索权限和索引，避免已经失效的授权继续生效。

Agent 使用知识库、RAG 结果、会话上下文或 Memory 向模型及外部服务传递数据时，客户应按照最小必要原则限制数据范围，避免将完整上下文或非必要敏感信息传递给无关模型、工具或第三方服务。更完整的数据来源、分类分级、脱敏和生命周期要求见 5.3 节。

【华为云服务】**AgentArts** 可提供知识库接入、文档处理、知识检索及相关资源管理能力，帮助客户构建 Agent 知识访问流程；具体权限控制和支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**数据安全中心 DSC** 可帮助客户对知识库及相关数据开展敏感数据识别、分类分级和脱敏；具体支持范围以正式产品文档为准。

5.5.3 Skill、MCP 与插件安全

客户为 Agent 连接 Skill、MCP 服务、插件、API 或其他外部工具时，应结合业务目的选择可信组件，并在使用前了解相关组件的来源、功能范围、访问权限和外部依赖。对于客户自行开发、部署或接入的 Skill、MCP 服务及第三方工具，客户负责其安全使用以及相关组件和依赖的安全管理。

客户应按照最小权限原则限制 Agent 通过工具可以访问的数据、接口和业务系统，并避免向 Agent 提供超出实际任务需要的高权限凭据。对于能够修改业务数据、调用高权限接口、执行代码、发起交易或对外发送信息的工具，应进一步限制可调用的方法、参数和资源范围，并根据业务风险增加必要的用户确认或人工审批。

客户还应关注工具描述、MCP 返回内容及第三方服务响应中可能包含的恶意指令或不可信内容，避免 Agent 将外部内容直接作为高可信指令执行。对于来源变化、版本更新、权限扩大或发现安全缺陷的 Skill、MCP 及外部工具，应重新评估其风险并及时调整、限制或停止使用。

【华为云服务】**AgentArts** 可提供 Agent 与相关工具、知识及平台资源连接和管理能力，帮助客户管理 Agent 应用中的资源调用关系；具体支持的 Skill、MCP、网关及权限能力以目标区域、控制台和正式产品文档为准。

5.5.4 运行隔离与审计

Agent 与普通模型调用的重要区别在于，模型输出可能进一步被用于任务规划、工具调用和实际业务操作。客户应将模型输出视为需要进一步验证的非确定性结果，不应在缺少必要校验和权限控制的情况下直接将其作为高风险业务操作或可信执行指令。

在 Agent 执行业务操作前，客户应结合操作类型和潜在影响实施必要的参数校验、权限检查和安全验证。对于涉及人身、财产、重要数据修改、对外发送、交易、审批、系统配置或其他重大权益的高风险操作，应根据实际风险增加用户确认、人工审批、权限提升或终止机制，使 Agent 不能仅凭模型一次输出直接完成重大或不可逆操作。

客户还应限制 Agent 可执行的代码、工具、网络访问和资源消耗范围，对动态代码、不可信任务或高权限工具采用必要的受限运行方式，并结合业务容量配置调用次数、并发、超时和其他资源限制，防止异常任务、递归调用或工具滥用造成资源异常消耗或业务影响。

在运行阶段，客户应根据业务风险记录必要的 Agent 身份、模型调用、知识检索、工具调用、关键参数、用户确认及执行结果等信息，为异常行为发现、调用链分析和业务事件追溯提供依据。相关日志应按照数据敏感程度设置访问权限和保存期限，并避免记录不必要的敏感信息。

当发现 Agent 权限滥用、异常工具调用、知识越权访问、非预期业务执行或其他应用安全事件时，客户应及时采取限制调用、撤销凭据、收窄权限、暂停相关工具或 Agent 服务等措施，并结合自身事件响应流程开展调查和整改；涉及华为云 MaaS 平台的事件，应按照第三章责任共担边界配合华为云开展必要的调查和恢复。

Agent 及其知识库、工具连接、权限和访问凭据应纳入应用生命周期管理。当 Agent 业务用途、模型、知识库、Skill/MCP 或外部系统发生重要变化时，客户应重新检查相关权限和安全控制；Agent 退出使用时，应及时撤销身份、访问凭据和相关权限，并按照 5.3 的数据生命周期要求处理知识库、Memory 和其他客户控制范围内的数据资产。

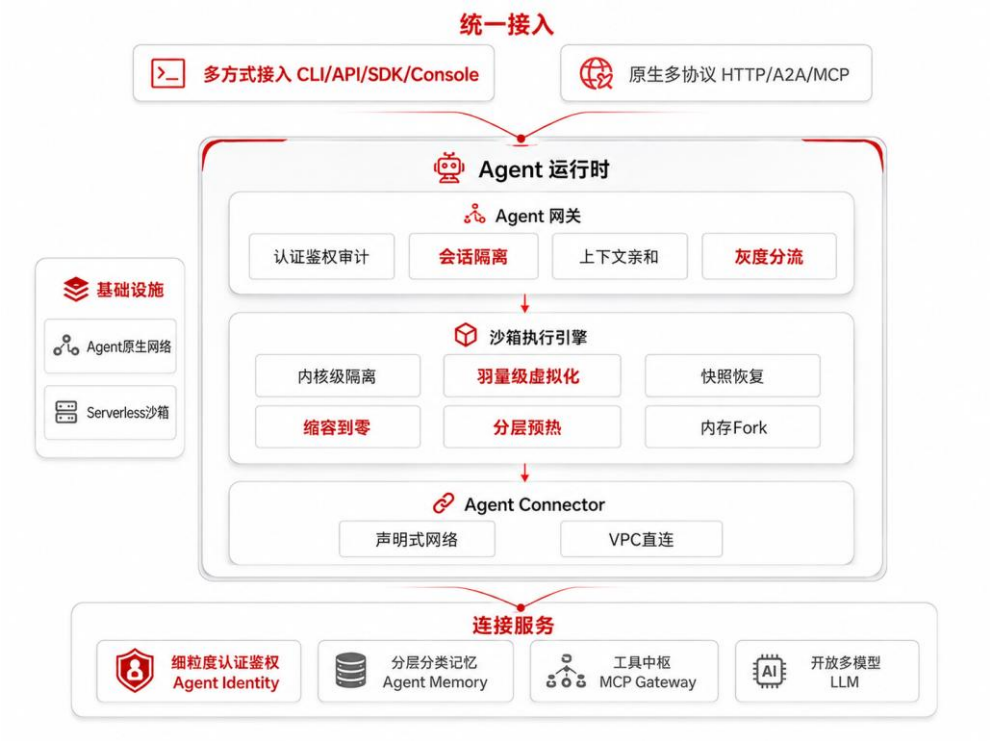

图 5-8 Agent 运行隔离

【华为云服务】**AgentArts** 可提供 Agent 运行环境及相关管理能力，客户可结合实际支持范围配置身份校验、访问控制、资源限制及相关运行安全能力；具体支持范围以目标区域、控制台和正式产品文档为准。

【华为云服务】**云审计服务 CTS** 及相关日志、监控能力可在支持范围内为客户开展相关资源操作追溯和安全运营提供支撑；具体日志、事件和监控范围以正式产品文档为准。

5.5.5 意图与执行安全

客户基于 MaaS 构建和部署 Agent 时，除了身份认证、权限控制和资源隔离，客户还应将 Agent 视为具备自主操作能力的数字员工，通过华为云智能体卫士，结合岗位职责建立行为管控基线，对运行过程中的关键行为和执行意图实施动态安全防护，防止自主规划、语义误解或任务执行偏离导致角色越界和高风险操作。

客户应结合企业组织架构和岗位职责，限定智能体的数字岗位角色和行为基线，明确其可访问的数据、可调用的工具和可执行的业务操作。例如，财务类 Agent 不应访问核心代码库，编码类 Agent 不应执行资金操作，防止 Agent 在自主规划过程中突破岗位职责边界产生逻辑越权。

客户还应为 Agent 建立明确的数据与指令边界，将系统指令、用户指令和外部数据进行明确区分。外部文档、网页内容、工具返回结果及其他第三方数据应作为不可信或低可信数据处理，不得通过内容注入获得系统级指令的语义地位，也不得覆盖既有任

务目标、权限约束或安全策略。Agent 在多轮任务执行过程中，应持续校验当前子目标与用户授权任务的一致性，避免外部数据或异常上下文诱导 Agent 偏离原始任务。

客户可通过华为云智能体卫士对智能体的运行行为和执行意图进行检测，结合任务上下文、用户身份、知识检索、工具调用及目标资源等信息，识别数据擦除、权限提升、敏感信息外发等高风险行为，并在实际执行前实施实时拦截或限制。对于涉及数据删除、资金交易、权限变更、审批或对外发送等高风险操作，应结合企业业务审批流程设置人工确认、人工审批或动态熔断机制。

当 Agent 偏离岗位行为基线或触发高风险意图时，应及时暂停相关任务、限制工具调用或将操作转交人工控制点，形成机主防、人主控的动态安全闭环。

6 MaaS 安全实践

前述章节分别从算力底座、模型、数据和运营运维等维度介绍了华为云 MaaS 安全能力。本章进一步从客户实际使用场景出发，说明这些安全能力如何组合应用：一是客户基于 MaaS 模型能力构建 Agent 时，如何将安全边界延伸至知识库、工具和业务系统；二是金融机构在强监管环境下使用 MaaS 时，如何结合数据、模型、平台及业务风险落实安全控制。

6.1 行业案例：某金融行业 MaaS 安全实践

6.1.1 客户需求

随着大模型逐步应用于智能理赔、员工办公助手、代码开发等核心业务场景，金融机构在使用 MaaS 服务过程中，对数据传输安全、网络边界隔离以及数据使用范围提出了更高要求。传统通过互联网访问 MaaS 服务的方式虽然可以采用 HTTPS 等技术保障传输安全，但对于金融、国资等安全敏感型客户而言，业务数据经过公网链路仍可能难以满足其内部安全管理和风险控制要求。

基于此，客户希望在使用云上大模型能力的同时，进一步降低业务数据暴露于公网的风险，实现客户数据在企业数据中心与 MaaS 服务之间的内网传输，并结合网络隔离、访问控制和传输加密等安全措施，构建更加安全、可控的大模型访问环境。

6.1.2 MaaS 安全实践

针对客户安全需求，华为云基于 VPC 和云连接 CC 构建 MaaS 内网访问方案。客户可将智能理赔、办公助手、代码开发等业务系统部署在自身 IDC 或云上 VPC，通过“客户 IDC—云连接 CC—客户 VPC—MaaS VPC”的访问链路，将客户业务网络与 MaaS 服务网络安全打通，使模型调用及相关业务数据能够通过云内网络完成传输，减少对公网访问链路的依赖。

在网络安全方面，通过 VPC 实现客户业务环境与其他网络环境之间的逻辑隔离，并结合安全组、访问控制等能力限制 MaaS 服务访问范围。在数据传输过程中，可进一步通过专属网络链路及加密机制保护数据传输安全，降低数据被窃听、篡改或非授权访问的风险，从而形成“网络隔离+内网访问+链路加密”的多层防护体系。

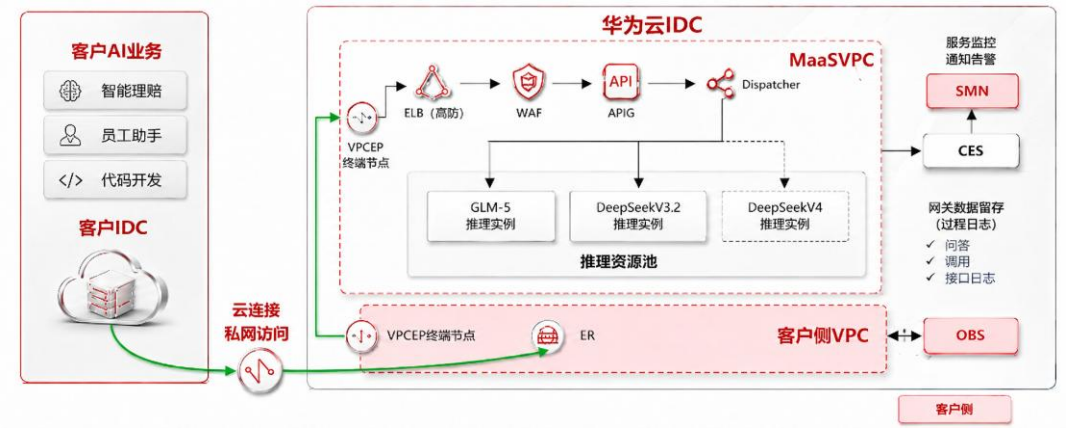

图 6-1 MaaS 内网传输方案示例

6.1.3 客户价值

通过上述方案，客户能够在获得大模型能力的同时，将安全控制进一步延伸至模型服务访问链路。一方面，通过 MaaS 内网专属访问实现业务数据不经公网传输，降低公网暴露面；另一方面，结合端到端传输加密、网络隔离、访问控制、内容审核和提示词安全防护等能力，对模型调用全过程实施安全保护。

同时，针对金融客户对数据保护的重点关注，可通过数据不用于模型训练、服务侧数据处理控制以及相关安全合规认证等措施，进一步明确数据使用边界。整体形成覆盖“访问、传输、数据、内容和模型”的多层安全防护体系，帮助金融机构更加安全、可控地将 MaaS 能力应用于智能理赔、内部办公及研发等业务场景。

6.2 MaaS 安全配置最佳实践

本章适用于通过华为云 MaaS 服务构建和运行 AI 应用的客户，覆盖网络安全、身份管理、数据保护、资产管理以及日志记录与威胁检测。各项配置应结合实际业务、区域可用能力和组织合规要求实施。

6.2.1 网络安全

6.2.1.1调用通道选型

1-1 | 优先通过 VPCEP 建立内网链路

业务系统部署在华为云 VPC 内时，应优先通过 VPC 终端节点（VPCEP）接入 MaaS，使调用流量通过内网传输，减少公网暴露。VPC 外或跨云场景可结合云专线（DC）或 VPN 扩展内网可达范围。

可参考：**https://support.huaweicloud.com/bestpractice-vpcep/vpcep_05_1001.html**

1-2 | 将公网 HTTPS 作为备选通道并减少暴露

无法通过内网接入时，可使用公网 HTTPS 端点。传输协议和服务端证书校验要求详见 1-3 与 1-4；后端代理和前端凭据保护要求详见 1-5。通道优先顺序为：VPCEP 内网、专线或 VPN 延伸的内网、公网 HTTPS（完成出口收敛后）。不得由前端通过公网 HTTPS 直接调用模型端点。

6.2.1.2传输加密

1-3 | 强制使用 TLS 1.2 及以上版本

所有调用应使用 TLS 1.2 及以上版本，条件允许时优先启用 TLS 1.3；禁用明文 HTTP，并避免协议或密码套件降级。

1-4 | 严格校验服务端证书

客户端 SDK 或 HTTP 库应启用证书和主机名校验，不得配置 verify=False 或忽略 TLS 错误。出向流量经过代理或网关时，应确保中间环节不存在未经授权或不受控的 TLS 终止与重新加密。

6.2.1.3调用链路收敛

1-5 | 通过后端代理调用，前端不直连模型端点

前端不得直接持有凭据并请求 MaaS 端点。请求应由后端、BFF 或网关代理，前端仅与自有后端通信。有关凭据安全托管和定期轮换的具体要求， 分别参见 2-5 与 2-6。

1-6 | 统一收敛出向出口

所有出向调用应收敛至少量固定出口，例如统一 NAT、出向代理或专用子网，以便集中实施安全控制和监测。固定出口 IP 可作为配置 API Key 来源 IP 白名单的依据；有关白名单的具体要求参见 1-8。

1-7 | 出向访问控制（最小化目的地与端口）

出口侧仅放行访问 MaaS 端点所需的目的地址和端口，例如 HTTPS 端口 443 或 VPCEP 内网地址，并限制与业务无关的外联。

6.2.1.4访问来源控制

1-8｜为 API Key 配置调用来源 IP 白名单

创建 API Key 时，应将允许的调用来源 IP 绑定至白名单。有关统一收敛出向出口的要求参见 1-6。建议将白名单设置为 后端统一出向出口的固定公网 IP，使 API Key 与出口对应。即使 API Key 泄露，白名单之外的来源也无法直接使用该凭据发起调用，因此建议将此项作为基线要求。

可参考：**https://support.huaweicloud.com/intl/zh-cn/model-call-maas/model-call-049.html**

1-9｜自有入口侧来源/地域管控作为补充

面向终端用户的来源 IP 和地域限制，应在自建 WAF 或 API 网关侧实施。该控制适用于“用户到自有应用”的访问链路；1-8 规定的来源 IP 白名单适用于“自有后端到 MaaS”的调用链路。

6.2.1.5自有 AI 应用入口防护

1-10｜WAF 入口防护

应根据业务暴露面在自有应用入口部署 WAF，防护注入、恶意爬取和恶意请求载荷，并针对大报文和超长上下文配置合理限制。

可参考：**https://support.huaweicloud.com/waf/index.html**

1-11｜Anti-DDoS/DDoS 高防

公网入口应启用基础 Anti-DDoS 防护；重要业务应结合风险评估和业务连续性要求配置增强型 DDoS 防护。

可参考：**https://support.huaweicloud.com/productdesc-aad/aad_01_0001.html**

1-12｜APIG 统一入口

自有应用可通过 API 网关统一对外提供服务，集中实施 TLS、限流、并发控制、访问控制和可观测性管理。

可参考：**https://support.huaweicloud.com/apig/index.html**

编号	检查项	控制责任	达标判据
1-1	VPCEP 内网接入	需客户开通或配置	VPC 内业务经 VPCEP 内网调用，流量不经公网
1-2	公网 HTTPS 备选	平台默认提供	无内网条件时使用 HTTPS，且不在前端暴露端点和凭据
1-3	TLS 1.2 及以上版本	平台默认提供	客户端启用 TLS 1.2 及以上版本，且不允许降级
1-4	证书校验	客户自行落实	开启证书+主机名校验，无 verify=False

编号	检查项	控制责任	达标判据
1-5	前端不直连	客户自行落实	前端仅连自有后端，凭据不落前端
1-6	出口收敛	客户自行落实	出向调用集中到少数固定出口
1-7	出向访问控制	客户自行落实	出口仅放行 MaaS 端点必需目的地/端口
1-8	API Key 来源 IP 白名单	需客户开通或配置	创建 API Key 时绑定固定出口 IP，并将其纳入安全基线
1-9	自有入口来源/地域管控	客户自行落实	终端用户来源限制在 WAF/APIG 落地
1-10	自有入口 WAF	需客户开通或配置	自有 AI 入口已接入 WAF
1-11	DDoS 防护	平台默认提供+需客户开通或配置	公网入口已启用基础防护，重要业务按风险配置增强防护
1-12	APIG 统一入口	需客户开通或配置	对外经 APIG，具备限流与统一治理

表 6-1 网络安全核查清单

6.2.2 身份管理

6.2.2.1最小权限与职责分离

2-1｜使用细粒度权限隔离 API Key 签发和删除操作

利用 MaaS 的 API Key 细粒度权限实施职责分离：仅向少数 API Key 管理员授予 modelarts:apikey:create（创建或编辑）与 modelarts:apikey:delete（删除）；普通使用者不应获得上述权限，只能使用已签发的 API Key，无权自行签发或吊销。

2-2｜按应用/环境隔离凭据边界

不同应用和不同环境（开发、测试、生产）应使用相互独立的 API Key，并配置独立的权限范围与来源约束。有关 API Key 权限范围及来源 IP 白名单的具体要求，分别参见 2-3 与 2-4。通过上述隔离措施， 可将凭据泄露的影响范围限制在单一应用和单一环境内。

6.2.2.2API Key 全生命周期管理

2-3｜签发时设置自定义访问范围

创建 API Key 时，权限范围应优先选择“自定义访问”而非“全部访问”。应仅允许 API Key 访问对应业务所需的模型服务或接入点，降低凭据泄露后访问其他模型服务的风险。

2-4｜签发时绑定来源 IP 白名单

创建 API Key 时，应同步绑定来源 IP 白名单，并将后端统一出向出口的固定 IP 配置为允许来源。有关来源 IP 白名单的配置要求参见 1-8。

2-5｜API Key 创建后立即纳入安全托管

若 API Key 明文仅在创建时展示，创建后应立即将其导入凭据管理服务，例如云凭据管理服务（CSMS）或数据加密服务（DEW），并验证应用能够安全读取。不得通过截图、即时通信工具、普通文本文件或其他非受控渠道保存和传递明文凭据。

2-6｜定期轮换，灰度替换不中断

应根据风险等级制定固定轮换周期，并在疑似泄露、人员变动或权限调整时立即轮换。可采用“创建新 Key、灰度切流、确认旧 Key 无调用流量、删除旧 Key”的顺序完成替换，以降低业务中断风险。若平台未提供自动过期能力，应由客户通过删除并重新创建的方式主动实施轮换。

2-7｜泄露应急预案

应预先制定书面流程。发现 API Key 疑似泄露时（例如收到依据 5-3 规定的检测规则产生的异常调用告警），应依次吊销泄露的 API Key、签发并适当收窄权限范围和来源 IP 的新 API Key、切换业务流量、排查泄露路径并开展复盘；同时应定期演练该流程。

编号	检查项	控制责任	达标判据
2-1	API Key 签发和删除权限隔离	需客户开通或配置	创建和删除权限仅授予指定的 API Key 管理员
2-2	应用和环境隔离	客户自行落实	各应用、各环境使用独立 API Key
2-3	自定义访问范围	需客户开通或配置	生产环境 API Key 仅能访问业务所需的模型服务或接入点
2-4	来源 IP 白名单	需客户开通或配置	签发时绑定固定出口 IP；配置要求参见 1-8
2-5	明文安全留存	平台默认提供+客户自行落实	仅创建时可见；即时导入托管，不明文散播
2-6	定期轮换	客户自行落实	有周期；灰度替换不中断
2-7	泄露应急	客户自行落实	有书面流程并演练

表 6-2 身份管理核查清单

6.2.3 数据保护

6.2.3.1数据分类与最小化

3-1｜对输入提示词的数据进行分类分级

应在应用侧建立数据分类分级和准入规则。公开数据可按业务需要调用；内部数据和敏感数据，例如个人信息、证件或银行卡信息、密钥口令、未公开源代码及核心商业数据，原则上不应直接输入模型。确因业务需要处理时，应按照 3-2 规定的数据保护要求实施脱敏或去标识化，并完成必要的授权与合规评估。

3-2｜敏感字段调用前脱敏/去标识化

在请求离开客户后端前，应对可识别个人身份的信息或高敏感字段进行脱敏、掩码、假名化或占位符替换。例如，可将真实姓名或证件号替换为令牌，并在获得响应后于客户侧还原。脱敏应在客户后端完成，不应依赖平台处理。

6.2.3.2传输中数据保护

3-3｜传输全程加密，禁用明文与弱协议

MaaS API 接入点使用 TLS 加密。有关传输通道选型、TLS 版本和服务端证书校验的具体要求，参见 1-1 至 1-4。数据在传输过程中应始终处于加密状态，禁用明文协议以及 TLS 1.0、TLS 1.1 等弱协议。

编号	检查项	控制责任	达标判据
3-1	已建立“可入提示词”数据分级规则并在应用侧执行	客户自行落实	规则文档存在；敏感级默认拦截
3-2	敏感字段在客户后端脱敏后才送入，不依赖平台	客户自行落实	抽样请求无明文 PII/凭据
3-3	全部调用走 HTTPS 并校验证书，无弱协议	需客户开通或配置	抓包/配置核查为 TLS1.2+且校验开启

表 6-3 数据保护核查清单

6.2.4 资产管理

6.2.4.1资产清点与台账

4-1｜基于平台入口建立 MaaS 资产台账

应利用控制台中的模型服务或订购列表、接入点配置、API Key 列表以及调用监控与统计等入口完成资产清点，并建立客户侧统一台账。台账字段建议包括资产类型、唯一标识、所属团队或应用、开通或创建时间、状态和负责人。可记录 API Key 的权限范围、来源 IP 白名单、标签和创建时间等非敏感属性，但不得记录明文凭据。

4-2｜定期核对台账与平台实际资产

应定期（例如每月）比对平台资产列表与客户侧台账，识别未纳入台账的模型服务或 API Key。此类资产可能源于人员变动后的遗留资源或未清理的临时测试资源。发现后，应将其纳入台账管理，或按照资产处置流程下线模型服务、删除 API Key。

6.2.4.2配额、用量与成本可见性

4-3｜用平台调用监控与配额视图掌握用量基线

应利用平台提供的调用监控、统计和配额视图，按模型服务、接入点或 API Key 等维度建立正常用量基线，包括调用量、频率和配额占用。用量基线应定期更新，并作为异常检测阈值的依据。

4-4｜为资产配置合理配额上限，防失控消耗

应结合业务容量和峰值需求，为模型服务或调用配置合理的配额上限，降低单一应用异常或 API Key 泄露导致调用失控和费用异常增长的风险。

编号	检查项	控制责任	达标判据
4-1	已建立覆盖模型服务/接入点/API Key 的资产台账	需客户开通或配置+客户自行落实	台账存在、字段完整、与平台入口对应
4-2	定期与平台对账，识别并处置未纳管资产	需客户开通或配置+客户自行落实	有对账记录；无游离于台账外的资产
4-3	已掌握各资产用量基线（借平台监控/配额视图）	需客户开通或配置	有按维度的用量基线记录
4-4	关键资产配置了合理配额上限	需客户开通或配置	配额已设且与业务量匹配

表 6-4 资产管理核查清单

6.2.5 日志记录与威胁检测

6.2.5.1日志采集与留存

5-1｜开通 CTS 云审计并覆盖 MaaS 管理面关键事件

应开通华为云云审计服务（CTS），并根据实际可审计事件范围，覆盖 IAM 授权变更、模型服务开通或退订、接入点变更、API Key 创建或删除以及委托操作等管理面关键事件，为身份和资产管理提供审计依据。

可参考：**https://support.huaweicloud.com/cts/index.html**

5-2｜日志转储独立留存、设保留期、防篡改

应将 CTS 事件转储至独立的对象存储服务（OBS）桶集中留存，设置符合组织合规要求的保留期限和防篡改控制，并通过职责分离避免操作人员删除或修改与自身操作相关的审计日志。

6.2.5.2威胁检测

5-3｜基于用量基线检测凭据滥用信号

应以 4-3 规定的用量基线为参照，针对以下偏离配置检测规则（使用平台调用监控或云监控 CES 指标；规则和阈值由客户确定）：一是异常调用量或频率，例如调用量突增、超过配额或异常高频；二是异常来源，例如来自 1-8 规定的 IP 白名单之外的调用尝试；三是异常时段或模式，例如非业务时段的规律性调用、异常的模型与接入点组合。

5-4｜启用内容安全审核，检测输入/输出层滥用

针对违规内容生成、提示词注入和诱导输出敏感信息等风险，如目标区域的 MaaS 平台提供内容安全审核能力，应按需启用并接入其风险信号；对于平台未覆盖的场景，客户应在应用侧对输入和输出实施补充校验。

5-5｜纳入平台侧防护日志（WAF/Anti-DDoS）作为威胁视图补充

按照 1-10 和 1-11 部署的 WAF 与 Anti-DDoS 防护可产生攻击拦截日志。应将相关日志纳入整体威胁视图，并与调用面异常开展关联分析，例如识别同时触发 WAF 拦截和异常调用的来源，以提高研判准确度。

6.2.5.3告警、响应与联动

5-6｜配置云监控告警，关键威胁实时触达责任人

应使用华为云云监控服务（CES），针对 5-3 规定的关键指标和事件配置阈值、突变及白名单外调用等告警，并通过短信、邮件或 Webhook 实时通知资产台账中登记的责任人。告警信息应明确关联具体资产及其负责人。

可参考：**https://support.huaweicloud.com/ces/index.html**

5-7｜建立威胁响应预案并联动处置措施

应制定并演练威胁响应预案，将告警与既有处置措施联动。发现异常后，应按照 2-7 规定的凭据泄露应急流程吊销或收窄疑似泄露的 API Key，按照 4-4 规定的配额管理要求临时调整配额，并依据 1-8 与 1-9 的访问来源控制要求收紧白名单和入口控制；随后排查泄露路径、复盘原因并更新相关策略。

5-8｜定期复盘日志与告警并持续优化控制策略

应定期复盘审计日志和告警记录，评估检测规则的有效性，并据此优化 1 至 4 所规定的配置策略，形成持续改进闭环。

编号	检查项	控制责任	达标判据
5-1	CTS 已开通并覆盖 MaaS 管理面关键事件	需客户开通或配置	CTS 事件列表含 IAM/服务/Key 变更
5-2	日志独立转储、设保留期、防篡改、操作者不可自删	需客户开通或配置+客户自行落实	转储桶独立；保留期与权限隔离到位
5-3	已针对异常调用量、来源和时段建立凭据滥用检测规则	需客户开通或配置+客户自行落实	检测规则已建立，并以按照 4-3 建立的用量基线为依据
5-4	已启用平台内容审核（如提供）+应用侧内容校验	平台默认提供+需客户开通或配置+客户自行落实	内容风险信号被采集与处置
5-5	平台侧 WAF/Anti-DDoS 日志纳入威胁视图	平台默认提供	防护日志被关联分析
5-6	关键威胁已配置告警并实时触达具体负责人	需客户开通或配置	CES 告警规则存在、通知链路验证有效
5-7	有威胁响应预案且联动吊销/配额/白名单等手段	客户自行落实	预案文档存在、经过演练
5-8	定期复盘日志和告警并持续优化安全配置	客户自行落实	有复盘记录和安全配置持续改进记录

表 6-5 日志记录与威胁检测核查清单

7 权威认证

7.1 大模型获取的权威安全认证

华为云 MaaS 依托全功能、高性能、高可靠、优体验等核心优势，通过了多项权威机构的认证与评估，验证了华为云 MaaS 在安全、可信、性能、服务能力等维度的行业领先性，为各方选型提供参考。

认证名称	认证说明
大规模预训练模型可信要求检验证书	中国信息通信研究院依据《大规模预训练模型技术和应用评估方法》开展的“可信 AI”系列评估。 华为云大模型产品凭借丰富的行业应用及卓越的安全可信能力，成为首个获得卓越级（5 级）评分的大模型产品，在人工智能技术能力和产品能力标准符合性验证中，获得优秀级（4+级）评分。
大模型服务能力成熟度等级证书	《人工智能 大模型 第 3 部分：服务能力成熟度评估》由中国电子技术标准化研究院联合 50 余家企事业单位共同编制，是大模型 ToB 服务能力领域的首个国家级标准。华为云成为首批通过中国电子技术标准化研究院大模型服务等级能力成熟度评估 3 项（全部）测试的企业，成熟度达到协同优化级。
可信 AI 评估证书（高质量 Token 服务系列-服务质量）	中国信息通信研究院依据《人工智能词元(Token)服务质量能力要求》开展“可信 AI-Token 服务质量评估”。华为云 MaaS 模型即服务经评估达到企业级通用场景 Token 服务质量要求，获颁首批“可信 AI-高质量 Token 服务”评估证书。
人工智能安全国家标准符合性自评估证书	中国网络安全标准化技术委员会发布《网络安全技术 生成式人工智能服务安全基本要求》，华为云参与生成式人工智能服务安全自评估试点，经评估验证，模型安全能力符合标准要求，获得中国电子技术标准化研究院颁发的证书。

7.2 华为云获取的权威认证

华为云计算技术有限公司已通过多项国内外管理体系认证，旨在提供坚实的合规性保障体系，涵盖 AI 治理、数据安全、隐私合规等多个关键领域，彰显华为云的 AI 合规能力。

认证名称	认证说明
ISO/IEC 42001	为应对人工智能的兴起及其带来的挑战，国际标准化组织(ISO)和国际电工委员会(IEC)发布了全球首个 ISO/IEC 42001 人工智能管理体系标准，华为云已经通过 ISO/IEC 42001 认证，证书具备 ANAB 国际权威认可标识。 该认证覆盖华为云人工智能产品，包括人工智能模型、人工智能云平台、人工智能应用助手等业务的研发、测试、运维及服务支持领域。 获得 ISO/IEC 42001 认证，是企业 AI 治理、合规与可持续发展达到国际先进水平的重要证明，充分展示了华为 AI 治理已全面对标国际标准，树立了负责任 AI 治理的标杆。
数据安全管理认证	数据安全管理认证（DSM 认证）是由中国市场监督管理总局和国家互联网信息办公室基于国家标准《信息安全技术 网络数据处理安全要求》（GB/T 41479）及相关标准规范开展的认证。 该认证表明华为云在数据收集、存储、使用、加工、传输、提供、公开等处理活动中符合严苛的标准要求。
数据管理能力成熟度等级认证	DCMM（Data management Capability Maturity Model）是《数据管理能力成熟度评估模型》GB/T 36073-2018 国家标准，是中国数据管理领域首个正式发布的国家标准。 华为云获得数据管理能力成熟度达到最高级别 5 级，证明了华为云在数据平台和数据管理服务的领先地位。

此外，华为云还通过了 ISO/IEC 27001、ISO/IEC 27701、ISO/IEC 27018、ISO/IEC 29151、ISO/IEC 22301 等权威认证/审计，通过华为云官网的【合规中心】-【华为云合规认证全景图】可以查阅华为云已获取的更多认证。除了这些第三方认证，【合规中心】还提供了【基于国家/地区的合规遵从性指导】和【基于行业的合规遵从指导】，以帮助解答您在合规过程中可能遇到的问题。

华为云致力于构建安全可信的服务，通过独立第三方权威机构的测评和认证，持续为客户提供可放心使用、最佳体验的智能产品和服务，共同推动 AI 向善发展。

8 附录

8.1 华为云 MaaS 安全能力清单

能力域	能力	MaaS内置能力 （默认安全能力）	客户可选能力 （高阶安全能力）	安全产品服务
资产与访问安全	AI资产识别与统一管理	√	√	AgentArts
	控制台及API身份认证	√		
	细粒度权限控制	√	√	IAM
	委托授权与临时凭证	√	√	IAM ModelArts委托授权
数据保护	敏感个人信息静态加密		√	DEW
	数据传输加密	√		
	敏感数据脱敏	√	√	DSC
	模型及资产完整性校验	√		
	租户数据隔离	√		
平台与运行防护	Web应用攻击防护	√		
	主机与计算环境防护	√		
	漏洞扫描与修复	√		
	大流量攻击防护	√		
	数据库安全防护	√		
	模型输入输出内容安全	√	√	大模型防火墙
	资源隔离与接口访问限制	√		
安全运营	关键操作审计与追溯	√	√	CTS
	在线服务和模型监控告警	√	√	CES
	威胁感知与事件响应	√	√	安全云脑
	故障检测与自动恢复	√		
	服务安全更新	√		

8.2 常见问题

- **问题 1：使用华为云 MaaS 时，我主要需要关注哪些安全风险？**

  答：使用 MaaS 时，您重点需要关注自身控制范围内的模型选用与应用、数据处理、账号与访问凭据、模型输入输出、业务应用及 Agent 运行等安全风险。例如，模型能力与业务场景不匹配，提示词注入、越狱、有害或错误输出，敏感信息泄露，API Key 泄露或异常调用，以及 RAG、Agent 场景中的越权检索、工具滥用和非预期业务执行等。对于芯片、服务器、底层云资源、MaaS 平台及华为云托管组件等华为云控制范围内的安全风险，由华为云按照 MaaS 安全责任共担边界实施相应安全保障。

- **问题 2：使用 MaaS 时，华为云和我分别承担哪些安全责任？**

  答：MaaS 安全责任主要依据双方对相关资源、数据和业务的实际控制权、可见性及可操作性进行划分。华为云负责云及 MaaS 托管服务本身的安全，包括算力底座、MaaS 平台和托管模型运行环境、平台侧数据传输与底层存储、平台安全运营以及 Agent 基础安全组件等。您负责自身在 MaaS 上的配置和使用活动，包括数据来源、授权和合规使用，模型选择和业务适用性判断，账号权限和 API Key 管理，安全策略配置，模型输入输出，以及自建应用和 Agent 的安全。

- **问题 3：华为云 MaaS 平台本身提供哪些安全保障？**

  答：华为云围绕 MaaS 服务链路建立覆盖算力、模型、数据和持续运营的安全保障机制，包括可信计算与资源隔离、租户隔离、安全传输与存储保护、身份与访问控制、模型安全评测与运行防护、模型输入输出安全、日志审计、风险监测、漏洞管理和安全事件响应等。同时，您可以根据业务需要配置或使用 IAM、大模型防火墙、DSC、DEW、CTS 等相关安全能力。具体支持能力以目标区域、实际使用的模型、控制台和正式产品文档为准。

- **问题 4：华为云如何帮助我降低模型使用过程中的安全风险？**

  答：华为云对其提供或托管的 MaaS 模型服务实施相应的模型安全管理，包括模型来源及供应链检查、安全评测、完整性保护、部署与运行防护、安全缺陷处置和版本管理等。在客户使用侧，您可以结合华为云提供的模型说明、安全评测和已知限制等信息选择适合业务场景的模型，并通过身份与访问控制、大模型安全护栏、调用限制等能力降低提示词注入、越狱、有害内容、敏感信息泄露和异常调用等风险。对于金融、医疗等高风险业务，您还应结合自身业务规则开展必要的适用性验证、事实核验和人工监督，不应仅以平台通用安全评测替代业务侧安全判断。更多 AI 安全解决方案请查看**《AI 安全解决方案白皮书》**。

- **问题 5：使用模型广场中的第三方模型或第三方模型 API 时，需要注意什么？**

  答：对于模型广场中的第三方模型，您应了解并遵守相应模型的许可协议、使用限制和其他第三方要求，并结合自身业务评估模型的适用性及相关风险。对于通过华为云云商店提供的第三方模型 API 服务，相应服务由第三方商家提供，相关服务权利义务由您与第三方商家根据对应协议确定。对于第三方模型及其输出，您仍应在实际业务使用前进行必要的安全性、准确性、合规性和业务适用性判断，不应未经验证直接用于重要业务决策或高风险操作。

- **问题 6：调用 MaaS 如何进行认证？API Key 如何安全管理？**

答：使用 API Key 鉴权，在华为云统一身份体系下生成管理。伙伴转售可按客户分配 API Key 并做资源隔离与计费对账。凭据请妥善保管、避免公开渠道泄露、建议开启监控与轮换；异常（如 401）联系技术支持恢复。

- **问题 7：华为云如何保障我在 MaaS 中使用内容数据及个人数据的安全？**

  答：在华为云控制范围内，华为云对 MaaS 的数据传输链路、托管及底层存储环境提供相应安全保障，通过安全传输、网络与租户隔离、存储加密、密钥保护、运行环境隔离和审计等机制降低数据被非授权访问、篡改或泄露的风险。您仍需对自身数据安全承担相应责任，包括确保数据来源合法、获得必要授权、明确使用目的，并根据数据敏感程度落实数据最小化、敏感信息识别与脱敏、访问权限、加密和生命周期管理。对于您自身数据中涉及的个人数据或其他敏感信息，应尽量减少非必要的进入模型，确需处理时，可在进入 MaaS 前采取脱敏等保护措施。华为云还可提供 IAM、数据安全中心 DSC、数据加密服务 DEW、云审计服务 CTS 等能力，帮助您落实自身责任范围内的数据保护措施。

- **问题 8：华为云会使用我的个人数据、提示词、上传数据或模型输出进行模型训练或优化吗？**

  答：根据华为云《MaaS 模型即服务服务声明》，“您的内容”包括您在平台处理的上传数据、提示词以及模型输出等。除为您提供平台相关服务外，在未获得您明确授权的情况下，华为云不会使用您的内容，包括用于模型训练、优化等；如“您的内容”中包含个人数据，同样按照上述规则处理，涉及法律法规、监管要求等情形的处理，则按照服务声明及相关协议执行。具体的数据处理规则、双方权利义务及最新条款，请以华为云**《MaaS 模型即服务服务声明》**为准。

- **问题 9：使用共享资源池时，我的数据会被其他租户访问吗？**

  答：华为云 MaaS 通过租户隔离、资源隔离以及运行环境安全机制对不同租户的数据、模型资产和计算资源实施隔离，降低跨租户非授权访问风险。共享底层资源并不意味着不同客户之间可以共享或访问彼此的数据。与此同时，您仍应妥善管理自身账号、API Key、数据访问权限和安全配置，避免因客户侧权限配置不当或凭据泄露导致数据被非授权访问。华为云对客户内容的处理范围，则按照双方协议、MaaS 服务声明以及客户授权和适用法律法规执行。

- **问题 10：正在使用的 MaaS 模型发生版本升级、能力调整或下线时怎么办？**

  答：模型可能由于版本迭代、第三方停止维护、安全风险、监管要求或其他原因发生升级、能力调整或下线。根据当前 MaaS 服务声明，对于模型下线日前三个月内存在调用记录的用户，在没有法律法规或监管特殊要求的情况下，华为云将按照声明约定提前通知模型下线信息。您应及时关注华为云官网、控制台及相关通知，并在规定期限内完成模型升级或业务迁移。更换模型后，还应重新验证模型能力、输入输出安全策略和业务控制是否仍然有效；不再使用原有服务时，应同步停用不再需要的访问凭据和相关配置。

- **问题 11：海外使用 MaaS 时，如何考虑合规准入、数据跨境和区域部署？**

  答：出海使用 MaaS 时，客户应首先确认目标市场可使用的 MaaS 区域、具体模型、服务条款及实际数据处理位置，并结合业务场景、数据类型和当地适用法律判断合规要求。

  使用本平台进行模型推理时，华为云将调配中华人民共和国境内的算力资源，并在中华人民共和国境内处理您的内容，包括但不限于您的输入的“提示词”、模型推理结

果。推理过程中，本平台不会存储您的内容，我们将在推理完成后立即删除您的内容。您需要对输入的“提示词”以及与此相关数据跨境合规性负责，并同意委托华为云作为数据处理者处理数据，包括处理输入的“提示词”、模型推理结果以及数据跨境事宜。具体要求，请以华为云**《MaaS 模型即服务服务声明》**为准。

8.3 华为云安全隐私服务

客户负责客户数据安全，为保障客户内容安全，客户可以根据客户内容适合的安全级别，采取额外的安全措施，额外的安全措施可以来源华为云，也可来源第三方。

华为云理解客户的安全与隐私保护需求，并结合自身丰富安全与隐私保护实践及技术能力，提供了相关的安全与隐私保护相关云服务供客户选择。云服务涵盖网络、数据库、安全、管理与部署工具等产品，相关产品的数据保护、数据删除、网络隔离、权限管理、容灾备份、安全审计等功能可帮助客户加强安全保障。

- **安全合规**

产品名称	产品介绍	核心功能
Web 应用防火墙 Web Application Firewall (WAF)	WAF可对网站业务流量进行多维度检测和防护，结合深度机器学习智能识别恶意请求特征和防御未知威胁，阻挡诸如SQL注入或跨站脚本等常见攻击。 客户可使用WAF保护其网站或服务器免受外部攻击，避免这些攻击影响Web应用程序的可用性、安全性或过度消耗资源，降低数据被篡改、失窃的风险。	安全防护
大模型防火墙 Large Model Firewall	全面守护AI应用安全，防范提示词攻击，确保内容合规与数据隐私。针对智能助手、智能体等AI业务，实现输入输出流量安全管控，支持包括提示词注入、敏感信息泄露、模型越狱、算力DoS、智能体劫持、RAG投毒等威胁检测能力，保障大模型及智能体业务平稳运行与数据合规。	AI安全防护 内容安全防护 大模型安全防护
云防火墙 Cloud Firewall (CFW)	CFW是新一代的云原生防火墙，提供云上互联网边界和VPC边界的防护，包括实时入侵检测与防御、全局统一访问控制、全流量分析可视化、日志审计与溯源分析等，同时支持按需弹性扩容、AI提升智能防御能力、灵活扩展满足云上业务的变化和扩张需求，让用户快速灵活应对威胁。 云防火墙服务是为用户业务上云提供网络安全防护的基础服务。	资产保护 访问控制 在线防御
企业主机安全 Host Security Service (HSS)	HSS 能提供资产管理、漏洞管理、基线检查、入侵检测等功能，能够帮助企业更方便地管理主机安全风险，实时发现并阻止黑客入侵行为。 客户可通过HSS更方便地管理主机、容器的安全风险，实时发现勒索、挖矿、渗透、逃逸等入侵行为，以满足等保合规的要求。	资产管理 漏洞管理 入侵检测
数据库安全服务 Database Security Service (DBSS)	DBSS 是一款智能的数据库安全服务，基于机器学习机制和大数据分析技术，提供数据库审计，SQL 注入攻击检测，风险操作识别等功能。 客户可通过DBSS检测潜在风险，保障云上数据库的安全。	安全审计
密码安全中心 Data Encryption	DEW 是一款综合的云上数据加密服务，提供专属加密、密钥管理、密钥对管理等功能。其密钥由硬件安	数据加密

产品名称	产品介绍	核心功能
Workshop (DEW)	全模块保护，并与华为云其他服务集成。客户也可以借此服务开发自己的加密应用。 客户可采用DEW进行密钥全生命周期集中管理，保障数据存储过程中的完整性。	
DDoS防护 Anti-DDoS Service (AAD)	AAD 是一款保护互联网服务器免受大流量 DDoS 攻而击导致不可用的增值服务。 客户可以通过AAD产品配置高防IP，将攻击流量引流到高防IP清洗，确保源站业务稳定可靠。	安全防护
数据安全中心 Data Security Center (DSC)	DSC 是新一代的云原生数据安全平台，提供数据分类分级，数据安全风险识别，数据水印溯源，数据脱敏等基础数据安全能力。 客户可通过DSC整合数据安全生命周期各阶段状态，构建云服务全景图，保护数据采集、存储/传输、使用、交换/销毁的安全。	数据分级分类 数据脱敏 数据水印
云堡垒机 Cloud Bastion Host (CBH)	CBH 是华为云的一款 4A 统一安全管控平台，为企业提供集单点登录、统一资产管理、多终端访问协议、文件传输、会话协同等功能于一体的运维管理服务。 客户可通过CBH对云主机进行远程运维，提高客户的访问控制安全能力，保护资源运维和系统管理的安全性，降低系统和运维资源被非法入侵的风险。	权限管理
云证书与管理服务 Cloud Certificate Manager (CCM)	CMM 是一个为云上海量证书颁发和全生命周期管理的服务，提供 SSL 证书管理和私有证书管理服务。 客户可通过CCM提高对SSL证书和私有证书的保密性和安全性，提升访问和传输通道的安全，降低数据在传输和访问过程中被非法入	证书管理
安全云脑 SecMaster	安全云脑基于云原生安全，提供全面的日志采集、安全治理、智能分析、态势感知、编排响应等快速闭环的安全信息和事件管理能力，实现自动化安全运营，助客户守护云上安全。	态势感知 安全运营

- **存储**

产品名称	产品介绍	核心功能
云备份 Cloud Backup and Recovery (CBR)	CBR为云上的弹性云服务器、裸金属服务器、云硬盘、云下VMware虚拟化环境和本地文件目录，提供简单易用的备份服务，当发生病毒入侵、人为误删除、软硬件故障等事件时，可将数据恢复到任意备份点。	数据备份
云硬盘备份 Volume Backup Service (VBS)	VBS 为云硬盘创建在线永久增量备份，并对加密盘发备份数据自动加密，并可将数据恢复到任意备份点，增强数据可用性。 VBS可降低病毒入侵、人为误删除、软硬件故障等事件的发生的可能性，保护数据安全可靠，降低数据被非法篡改的风险。	数据备份
云服务器备份 Cloud Server Backup Service (CSBS)	CSBS 可同时为云服务器下多个云硬盘创建一致性在线备份。 CSBS可降低病毒入侵、人为误删除、软硬件故障等事件的发生的可能性，保护数据安全可靠，降低数据被非法篡改的风险。	数据备份

- **管理与监督**

产品名称	产品介绍	核心功能
统一身份认证服务 Identity and Access Management (IAM)	提供身份认证和权限管理功能，可以管理用户（比如员工、系统或应用程序）账号，并且可以控制这些用户对其名下资源的操作权限。 客户可通过IAM采取适合的用户管理、身份认证和细粒度的云上资源访问控制等措施，防止对内容数据进行的未授权修改。	权限管理
云审计服务 Cloud Trace Service (CTS)	为客户提供云账户下资源的操作记录，实现安全分析、合规审计、问题定位等场景。 客户可以通过配置CTS对象存储服务，将操作记录实时同步保存至CTS，以便保存更长时间的操作记录，保障数据主体的知情权、实现快速查找。	安全审计
云监控服务 Cloud Eye Service(CES)	为客户提供一个针对弹性云服务器、带宽等资源的立体化监控平台。 客户可通过CES全面了解华为云上的资源使用情况、业务的运行状况，并及时收到异常报警做出反应，保证业务顺畅运行。	安全审计
云日志服务 Log Tank Service (LTS)	提供日志收集、实时查询、存储等功能，无需开发即可利用日志做实时决策分析，提升日志处理效率，帮助用户轻松应对日志实时采集、查询分析等日常运营、运维场景。 客户可通过LTS保留对个人信息的操作记录，保障数据主体的知情权。	安全审计
配置审计 Config	配置审计（Config）服务提供全局资源配置的检索，配置历史追溯，以及基于资源配置的持续的审计评估能力。 客户可通过Config查看资源详情、资源之间的关系、资源历史，并可以通过配置合规规则来对资源进行合规性检查，确保云上资源配置变更符合预期。	安全审计

- **网络**

产品名称	产品介绍	核心功能
虚拟专用网络 Virtual Private Network (VPN)	VPN 用于搭建客户本地数据中心与华为云 VPC 之间便捷、灵活，即开即用的 IPsec 加密连接通道。 客户可通过VPN实现灵活一体，可伸缩的混合云计算环境，并且由于VPN的加密特性，提高了客户的安全	安全传输
虚拟私有云 Virtual Private Cloud (VPC)	VPC 是客户在华为云上的隔离的、私密的虚拟网络环境。客户可以自由配置 VPC 内的 IP 地址段、子网、安全组等子服务，也可以申请弹性带宽和弹性 IP 搭建业务系统。 VPC是客户的云上私有网络，各客户之间100%隔离，增强云上数据的安全性。	网络隔离