百度智能云

AI基础设施安全白皮书

AI INFRASTRUCTURE SECURITY
WHITE PAPER

专家指导委员会： 聂科峰　杜　海　王佩龙
郑　然　宋　飞　杨　阳
包沉浮　张洪海

编　　写　　组： **百度安全**
林道正　刘　杰　文　炜
韩长青　程　波　曾　远
赵飞飞　应　蕊　吴贵荣
李志伟　孙　禹　季石磊
王建奎　刘琳琳　韩　冲
吴　琼　林孟贤　曾华伟
温慧媛

百度混合云
李兆彤　史　磊　李玉双
李　阳　谢伟光　王　发
崔　凯　王晓珂　杨　正
魏　谦　孙召增　马建英
田晓利　徐　浩　陈瑛雷
赵志伟　滕阳阳　袁　穆

文心一言4.5

INTRODUCTION

引言介绍

随着人工智能技术的飞速发展，AI基础设施已成为推动产业智能化变革的核心支撑。作为承载海量数据处理、模型训练与推理任务的关键底座，其安全性直接影响人工智能应用的可靠性与可持续发展。近年来，各行业对算力的需求呈现爆发式增长，构建高效、安全、可靠的AI基础设施，已成为保障人工智能技术大规模落地的重要前提。

在这一背景下，百度积极协同客户，在全国多个地区成功部署了大规模万卡级别的AI算力集群。基于丰富的实际建设与运营经验，本白皮书立足AI基础设施在架构设计、安全防护、合规运营与持续迭代等方面的现实需求，全面梳理并阐述了经过实践验证的体系化安全解决方案与实施路径，旨在为行业提供具备参考价值的方法论与实践指南。

- 结合近两年最新安全事件与政策法规，深入解读AI基础设施的安全趋势与合规要求，为算力集群安全建设提供前瞻性指引；
- 分享在万卡级AI算力集群中的安全防护实践，涵盖模型应用安全、数据隐私保护与合规落地等方面的实战经验；
- 系统介绍覆盖规划设计、建设实施、运营管理及迭代优化的全生命周期安全框架，包括关键安全技术路线与综合管理策略。

通过本白皮书，读者将全面了解AI基础设施所面临的核心安全风险与应对措施，为构建安全、可信、合规的AI基础设施提供有力支撑。

CONTENTS

目录

01

行业背景
政策与技术发展趋势

1.1 行业背景与安全挑战

当前，全球数字经济进入以人工智能为核心驱动的新阶段。融合算力中心作为新型基础设施的核心载体，正处于从规模扩张到质量提升的关键转型时期。在国家“十四五“数字经济发展规划与“东数西算”工程战略指引下，我国算力中心建设呈现三大特征：

- 一是政策驱动显著，2025年智能算力规模预计突破1000EFLOPS，地方政府通过专项补贴、产业基金等政策加速布局；
- 二是智能化需求爆发，驱动大模型算力需求指数级增长，推动AI芯片、分布式存储等技术迭代优化，进而实现算力效率提升与成本下降；
- 三是应用场景深化，智慧电网、自动驾驶、智慧港口等垂直领域加速算力下沉，边缘计算市场规模预计2028年达到132亿美元（IDC，2023）。

然而，在快速发展的背后，安全风险日益凸显。

- 在合规层面，数据跨境流动、大模型应用备案、算法备案等监管要求持续收紧，2024年《生成式人工智能服务管理暂行办法》的实施，对算力中心的数据治理与模型合规提出更高标准；
- 在云平台层面，多云架构的复杂性导致API接口暴露、权限管理失控等问题频发，某头部运营商2024年因云平台漏洞导致的安全事件同比增长65%；
- 在大模型层面，提示词注入、模型窃取、数据投毒等新型攻击手段涌现，某开源社区2025年监测到针对大模型的恶意样本数量突破200万例。

诸如此类事件层出不穷。为此应构建AI基础设施一体化安全防护框架，涵盖合规体系构建、物理设施防护、云原生安全保障、模型应用安全防护等核心模块，为政企用户提供可落地的安全实践指南，助力行业在算力革命中筑牢安全底线。

1.2 国家政策法规解读

随着数字经济的深度发展，信息基础设施的重要性凸显，网络安全已成为国家安全的重要组成部分，相关法律法规体系也愈发完善；通过对近些年出台的网络安全法律法规，以及AI基础设施建设的各类指导意见的分析，我们梳理出当前AI基础设施建设中安全相关的政策法规主要有：

《中华人民共和国网络安全法》2017年6月施行

《网络安全法》规定了网络运营者需履行安全保护义务，强调了对关键信息基础设施的重点保护，以及监测预警和应急处置制度。压实AI基础设施安全建设的必要性，对于违反网络安全规定的行为，规定了相应的法律责任，包括警告、罚款、没收违法所得等。

《中华人民共和国密码法》2020年1月施行

《密码法》及《商用密码管理条例》规范了密码应用，AI基础设施作为关键信息基础设施时，应当使用商用密码实施保护，制定商用密码应用方案，配备必要的资金和专业人员，同步规划、同步建设、同步运行商用密码保障系统，自行或者委托商用密码检测机构开展商用密码应用安全性评估。

《中华人民共和国数据安全法》2021年9月施行

《数据安全法》强调数据全生命周期的安全保护，要求AI基础设施运营者对数据进行分类分级管理，采取加密传输、存储等措施保障数据安全。在数据采集环节，需遵循合法、正当、必要原则；在数据使用环节，要严格控制数据访问权限，防止数据滥用。

《关键信息基础设施安全保护条例》2021年9月施行

《关键信息基础设施安全保护条例》规定运营者应当落实网络安全等级保护制度要求，在网络安全等级保护的基础上，对关键信息基础设施实行重点保护，AI基础设施中的很多部分都属于关键信息基础设施范畴，需要参照。

《生成式人工智能服务管理暂行办法》2023年8月施行

《生成式人工智能服务管理暂行办法》旨在促进生成式人工智能健康发展和规范应用，对AI基础上设施中生成式人工智能服务的研发、提供、使用等环节进行全面规范，强调安全评估、算法备案、标识管理等要求，防范生成式人工智能服务安全风险。

1.3 技术前沿趋势

AI基础设施安全体系要求

增强网络安全保障能力

严格落实网络安全法律法规要求，开展通信网络安全防护工作。强化安全技术手段建设，加强对网络流量、行为日志、数据流转、共享接口等安全监测分析，推动威胁处置向风险预警和事前预防转变，建立威胁闭环处置和协同联动机制，提升威胁处置科学性、精准性和时效性。

强化数据安全保护能力

加强数据分类分级保护，根据监管要求对重要和核心数据实行精准严格管理。制定数据全生命周期安全防护要求和操作规程，配套建设数据安全风险监测技术手段，加强数据安全风险的分析、研判、预警和处置能力。

强化产业链供应链安全

加强产业链协同联动，逐步形成自主可控解决方案，鼓励算力基础设施，采用安全可信的基础软硬件进行建设，保障供应链安全。加强关键技术研发和创新，提升软硬件协同和安全保障能力。依托一体化算力应用安全保障体系，形成“云网边端”安全态势感知和网络协同防护能力。推动智能化分析和决策在未知安全风险自主捕捉和防御环节的应用，持续提升算力安全保障能力。

保障算力设施平稳运行

强化算力网络保障，对重要网络设施采用双节点、双路由配置，避免出现单点故障。加强物理设施保护，定期开展巡查巡检，制定应急预案，提高应急处置能力。对重要系统和数据，建立热备双活机制，应用仿真灰度测试、混沌工程等新技术，发掘并消除软件系统潜在隐患。

云安全

CASB（云访问安全代理）深化应用

随着企业多云、混合云战略普及，CASB作为云安全核心组件，通过实时监控云服务访问行为，强制执行数据加密、访问控制及合规策略，与AI技术结合后，可自动识别异常访问模式（如非授权数据导出），并联动威胁情报实现动态防护，成为企业上云安全的关键屏障。

CNAPP（云原生应用保护平台）全生命周期防护

针对云原生架构（容器、Serverless、微服务）的复杂性，CNAPP通过与DevOps流程无缝集成，实现“左移”安全（开发阶段嵌入安全控制）和“右移”防护（运行时实时监测），有效应对云原生环境特有的配置错误、漏洞利用及供应链攻击风险。

CSPM（云安全态势管理）持续观测

云平台安全的态势情况，采用API无代理方式持续发现与盘点云资产，结合AI与图谱分析，CSPM可识别配置漂移与潜在攻击路径，依据业务重要性与暴露面做风险分级与处置优先级排序。

AI安全

AI技术在安全防护中的应用

基于机器学习与神经网络的AI算法，能对海量安全数据进行实时分析，建立正常行为模型，快速准确识别异常行为，像通过分析网络流量数据发现潜在的DDoS攻击、恶意软件传播等威胁，且可自动学习适应变化的安全威胁，提升检测准确性与及时性，减少人工误判。同时，随着OpenAI ChatGPT与微软的Security Copilot发布，基于Transformer架构的大语言模型被引入更多安全业务场景，如恶意邮件检测、安全事件研判、恶意软件分析等，在防护效果和检测性能上有显著突破，有效解决传统机器学习和深度学习的瓶颈。

AI算法与模型自身的安全防护

针对不同AI业务场景，AI算法面临多种攻击威胁，例如图像识别业务遭遇对抗样本攻击和深度伪造攻击，AI问答业务面临提示词注入、越狱及数据泄露攻击。这要求模型生成时考虑鲁棒性，并配合体系化安全对齐机制实现算法层面的安全加固。而模型作为企业知识产权的沉淀，其自身保护至关重要，尤其在AI模型轻量化趋势下，终端模型或私有化模型在各业务场景中日益普遍，需通过可信计算、代码混淆、加密、访问控制和权限管理等安全技术，对模型分发进行全方位保护。

AI数据的安全保障

数据是AI业务的基石，也是攻击者重点关注的切入点，常见攻击包括数据投毒、数据泄露等，比如大模型训练阶段引入错误价值观内容，或数据流通过程中发生泄露。为此，需针对AI业务数据生命周期各阶段布置安全保护措施，如进行数据内容清洗、脱敏、加密存储与使用审计，形成“预防 - 检测 - 响应”闭环，确保数据在训练、推理及迭代各环节均符合数据安全与合规要求。

1.4 AI基础设施安全风险洞察

数据算力安全风险：高价值目标暴露

AI基础设施作为智能时代的算力枢纽，其稀缺的算力资源与海量敏感业务数据，成为攻击者眼中的“高价值猎物”。算力的稀缺性，使其在数字经济生态中具备战略价值，而存储的敏感数据（如政务机密、企业核心业务信息、用户隐私等），关乎组织与个人权益。AI业务场景下的算力薅羊毛，用户数据倒卖成为灰色产业的新套利点，使得攻击者持续对AI基础设施发起探测与攻击，从网络渗透、漏洞利用到暴力破解，试图突破防护边界，一旦得手，将引发数据泄露、算力劫持等重大安全事件，威胁业务稳定与数据主权。

模型技术安全风险：迭代中的安全滞后

当前AI技术迭代迅猛，新产品、新技术层出不穷，但安全建设常跟不上创新节奏。在模型开发与应用环节，安全易被忽视：一方面，快速迭代导致安全检测、验证环节压缩，模型可能存在算法漏洞、逻辑缺陷；另一方面，模型推理易输出违规内容，如暴力、色情信息等，既违反法律法规，也会对用户、社会造成不良影响。这种安全滞后，让AI基础设施面临合规风险与技术滥用威胁，需在追求创新速度的同时，筑牢模型安全防线，平衡发展与安全。

02

百度AI基础设施安全架构

2.1 AI基础设施安全架构介绍

图1.百度AI基础设施安全架构

百度AI基础设施安全架构构建了全方位、多层级的安全防护体系，以保障算力中心稳定、合规、安全运行，具体从合规要求、标准规范、核心安全域及管理与运行体系展开：

合规与标准规范

架构以法规要求为刚性约束，覆盖《网络安全法》《数据安全法》《个人信息保护法》等法规，确保业务合法合规；同时，遵循《信息安全等级保护2.0》《密码应用技术要求》《GB/T 34458 -2022云计算安全技术要求》等标准，为AI基础设施安全建设提供技术指引与评测依据。

核心安全域分层防护

模型应用安全

聚焦AI基础设施模型应用安全全流程，以内容合规把控输出边界，防范违规信息生成；以模型安全保障算法逻辑、训练过程可信，抵御投毒、篡改等攻击；依托数据安全守护训练与推理数据，实现分级加密、合规流转；通过应用安全加固业务入口，防范漏洞利用与恶意访问。

云服务安全

云服务安全提供一体化租户安全防护能力，覆盖网络、主机、应用、数据等全方位安全需求，租户可根据自身业务需求灵活选择安全服务。租户基础安全服务包括云主机安全、云原生容器安全、租户网络安全防护及租户应用合规性检查，租户合规覆盖等保合规、应用密码合规要求。同时提供租户级安全运营托管服务，满足一站式租户安全运营需求。

云平台安全

夯实AI基础设施底层基座，安全治理（漏洞、基线）实现风险主动收敛；计算、存储、网络安全保障基础设施稳定；安全合规（等保、密评）确保平台资质达标；物理安全守护机房、硬件；安全运营与数据安全，保障运维可控、数据全生命周期安全，为上层服务筑牢根基。

算力调度平台安全

算力调度平台是面向企业多角色（管理者、运维、开发）的AI算力管理平台。针对大规模算力场景需求，平台支持容器、裸金属虚机等多元基础设施，适配自建与服务托管场景，助力企业向智能化、集约化转型。针对智算调度平台安全挑战，通过构建多层级防护体系，覆盖基础设施至业务场景全链路，有效化解算力调度风险，保障算力中心稳定运行。

管理与运行体系保障

管理体系

以安全方针明确战略方向，安全组织落实岗位权责，安全规范细化操作准则，风险控制建立全周期风险闭环，从制度流程层面保障安全落地。

运行体系

通过安全服务提供专业技术支撑，安全运维保障日常稳定，安全运营（监测、响应、优化）实现动态防护，构建“规划 - 执行 - 监测 - 迭代”的持续安全能力。

整体来看，该架构以合规为纲、技术为骨、管理为翼，分层覆盖模型应用、云服务、云平台，协同管理与运行体系，形成适配 AI 基础设施特性的纵深防御体系，护航算力业务安全发展。

2.2 AI基础设施云平台安全

云平台安全			
安全治理	计算安全	安全合规	物理安全
漏洞治理	存储安全	等保合规	安全运营
基线治理	网络安全	密评合规	数据安全

图2.云平台安全

云平台作为承载算力服务的基础设施，围绕AI基础设施场景，从安全治理维度，开展云平台安全治理、漏洞治理、基线治理；覆盖计算、存储、网络安全层面；融入安全合规要求，满足等保、密评规范；联动安全运营与数据安全建设，全方位构建云平台安全防护体系。

治理

安全治理：从管理机制入手，规划云平台安全策略、流程，明确权责，让安全建设有“顶层设计”；
漏洞治理：聚焦漏洞全周期管理，涵盖扫描、验证、修复、复测，及时修复云平台（如虚拟机、容器）漏洞；
基线治理：制定云平台安全配置基线（如系统参数、账户权限），通过检测、整改，保障基础环境合规。

防护

计算安全：守护宿主机、容器等算力载体，防范算力劫持、恶意程序入侵，保障计算过程可信；
存储安全：对云存储数据加密（传输/存储）、备份，防范数据泄露、篡改，保障数据持久安全；
网络安全：构建云平台网络边界、隔离区域，检测并拦截网络攻击（如入侵、横向渗透）。

合规

安全合规：遵循国家及行业通用安全标准，规范云平台建设；
等保合规：根据网络安全等级保护2.0要求,完成云平台等保定级、备案、测评、整改,满足“一个中心、三重防护”安全要求；
密评合规：聚焦密码应用，验证云平台国密算法（SM2/SM4等）使用、密钥管理是否合规，通过密码应用安全性评估。

运营

物理安全：保障云平台物理设施（机房、服务器、网络设备）安全，防火、防盗、防雷等；
安全运营：通过态势感知、日志审计、应急响应，7x24小时监控云平台，及时处置安全事件；
数据安全：覆盖云平台数据全生命周期（采集、存储、使用、销毁），保障数据机密性、完整性、可用性，防范数据泄露、滥用。

整体来看，云平台安全以“治理-防护-合规-运营”为主线，全面覆盖建设需求，构成算力中心稳定运行的“安全底座”设计框架。

2.3 AI基础设施云服务安全

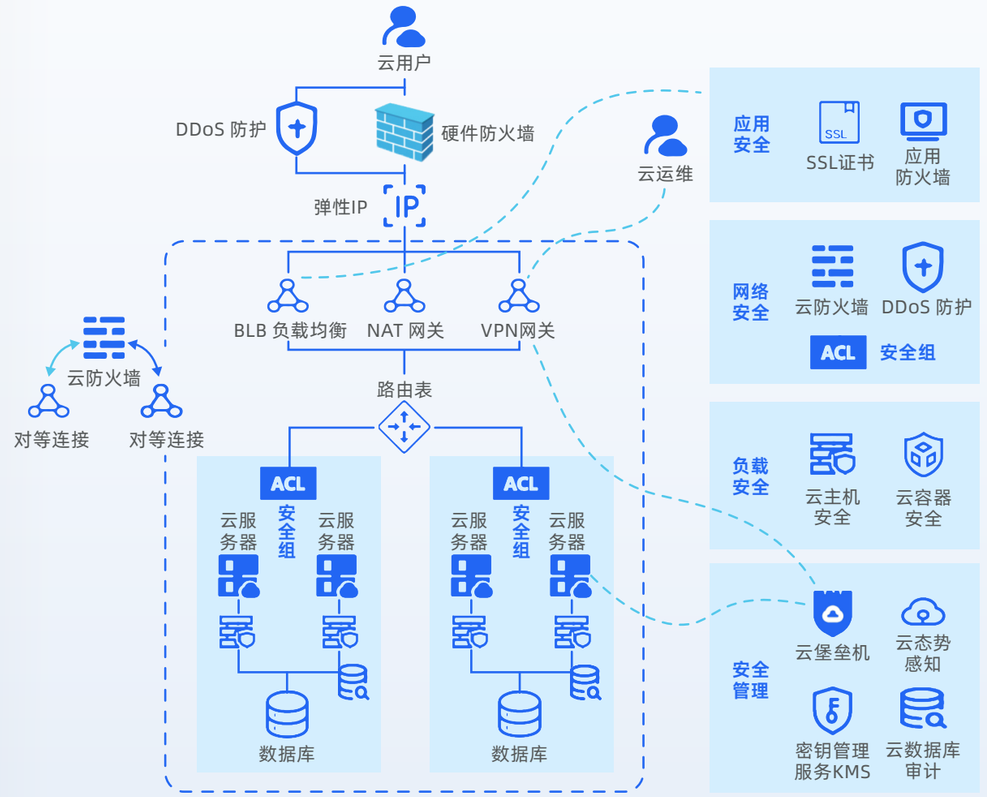

图3.云服务安全

AI基础设施云服务安全防护体系，聚焦用户访问、日常运维等场景，构建应用、网络、主机、安全管理四层防护体系，精准覆盖各层级安全防护需求。

应用层安全：守护Web业务交互

风险：云上Web服务易遭注入攻击（如SQL注入）、恶意爬虫等，威胁业务逻辑与数据安全。
防护：通过WEB应用防火墙检测过滤非法Web请求，拦截攻击；依托SSL证书加密传输链路，保障用户与业务间数据机密性、完整性，从应用入口筑牢防线。

网络层安全：隔离内外网威胁

风险：外部DDoS攻击冲击网络带宽，内部存在非法访问、横向渗透风险。
防护：以DDoS防护抵御流量攻击；云防火墙+ACL/安全组构建虚拟边界，限制跨VPC访问，精细管控业务/主机间网络交互，切断攻击传播路径。

负载层安全：加固算力载体

风险：虚拟机、容器因漏洞、弱配置，易被植入恶意程序，引发算力劫持、数据泄露。
防护：借助云主机/容器安全产品，实现登录管控、风险感知（如恶意文件检测）、入侵拦截，实时监测进程异常，守护算力载体安全运行。

安全管理层：保障安全策略落地与运营

风险：云用户自主配置易引发风险暴露，多组件告警难协同，安全管理分散。
防护：云态势感知汇聚全平台告警，关联分析识别潜在威胁；云堡垒机审计运维操作，保障管理合规；密钥管理（KMS）+数据库审计，加密敏感数据、监测数据操作，实现“监测-分析-处置”闭环，让安全策略有效落地。

四层体系层层联动，从业务交互到底层算力，从网络隔离到集中运营，构建云上全链路安全防护网。

2.4 算力调度平台安全

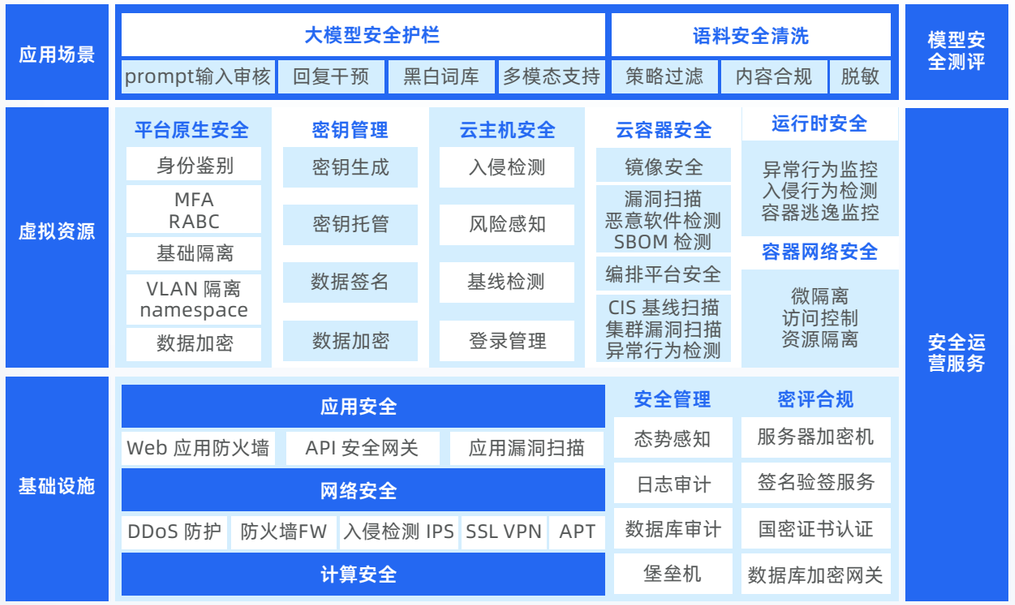

图4.算力调度平台安全

算力调度平台以高效稳定、多芯适配、轻量灵活为产品核心，是专为IT管理者、运维人员、开发人员等多角色打造的一款AI异构算力管理平台。针对大规模算力场景，算力调度平台可同时提供容器、裸金属、虚机等多种基础设施资源，满足企业自建、服务托管等多类建设场景，帮助企业快速、平稳的向新一代智能化、集约化基础设施转型。

针对大规模计算场景下异构算力平台特有的安全挑战，百度基于项目实践构建了多层级安全防护体系。该体系贯穿基础设施层、虚拟资源层及应用场景层，形成从硬件到业务场景的全链路安全保障机制，有效应对算力调度过程中的潜在风险，确保算力中心稳定运行。

基础设施安全：筑牢底层基座

计算安全
针对以容器形式接入算力调度平台的CPU和GPU服务器，底层物理服务器运行HOST OS系统，部署百度自研Hosteye Agent。借助该Agent，可实现登录管理，精准管控人员访问权限；开展漏洞扫描和修复，及时填补系统安全漏洞；进行入侵威胁检测，有效识别并拦截异常攻击，避免核心管控服务器和GPU服务器等遭受勒索、挖矿等病毒植入，为上层业务稳定运行筑牢基础防护。

网络安全——聚焦GPU算力平台防护

围绕GPU算力平台，构建网络分区分域、边界防护、流量分析、加密等多维度防护体系：

分区分域：根据GPU算力业务属性（训练、推理、数据存储等），划分独立安全域，实施“域间最小权限”访问控制，避免攻击跨域蔓延；

边界防护：在GPU算力域与外部网络、其他业务域间部署专业防火墙，结合IPS/IDS，阻断非法访问、恶意渗透，筑牢网络边界；

流量分析：通过流量可视化平台，对GPU算力平台流量进行深度解析，识别异常流量（如数据外泄、算力劫持行为）；

加密传输：采用国密算法（SM4等），对GPU算力平台的管理网络传输进行加密，保障数据传输安全。

应用安全——构建多层应用防护

- 应用防护（全平台覆盖）
  - 通过Web应用防火墙拦截SQL注入、XSS等Web攻击，API安全网关管控接口访问权限与频率，定期开展应用漏洞扫描，及时发现代码逻辑缺陷、未授权访问等隐患，从业务入口筑牢应用层安全防线，降低漏洞利用风险。

- GPU算力平台专项防护
  - 传输加密（HTTPS）：强制GPU算力平台应用访问采用HTTPS协议，基于SSL/TLS（适配国密算法）加密通信链路，防止模型训练指令、推理结果等敏感数据传输中被窃取，篡改；
  - 深度漏洞检测：针对GPU算力场景特有的应用（如模型训练框架、推理服务接口），开展定向漏洞扫描，覆盖算力调度逻辑漏洞、AI模型专属攻击面（如对抗样本注入接口漏洞），结合行业威胁情报优先修复高危漏洞。

安全管理

借助态势感知汇聚全平台安全告警，日志审计回溯操作行为，数据库审计监测数据访问，堡垒机管控运维权限，构建“监测-审计-管控”闭环，保障管理流程合规。

密评合规

通过服务器加密机、签名验签服务、国密证书认证、数据库加密网关，落实国密算法应用，满足密码应用安全性评估要求，筑牢密码安全防线。

虚拟资源安全：守护算力载体

平台原生安全

融合身份鉴别（MFA、RBAC）、基础隔离（VLAN、namespace）、数据加密技术，从身份、网络、数据维度，保障虚拟平台自身安全，防止资源越权与数据泄露。

密钥管理

覆盖密钥生成、托管、签名、加密全流程，为虚拟资源数据提供可靠密钥支撑，保障数据机密性与完整性。

云主机安全

通过入侵检测、风险感知、基线检测、登录管理，实时监测主机异常，拦截入侵行为，确保云主机运行环境安全。

云容器安全

围绕镜像安全（漏洞扫描、恶意软件检测等）、编排平台安全（CIS基线扫描等）、运行时安全（异常行为监控等），全生命周期防护容器平台。

应用场景安全：聚焦大模型防护

大模型安全护栏

通过prompt输入审核、回复干预、黑白词库、多模态支持，管控模型输入输出，防范违规内容生成，保障模型交互安全。

语料安全清洗

依托策略过滤、内容合规、脱敏技术，对训练语料预处理，去除敏感、违规信息，从源头保障模型训练数据安全。

该体系从底层基础设施到上层应用场景，分层防护、协同联动，为算力中心稳定运行与大模型安全应用，构建“全栈、全周期“安全屏障。即通过标准化方式将应用防护接入全平台业务，又针对 GPU 算力场景设计安全防线，精准满足算力中心核心业务需求。

03

百度AI基础设施大模型应用安全

针对大模型训练阶段、部署阶段和业务运营阶段所面临的安全挑战，提供完整的应对方案，围绕大模型网络安全、模型数据安全与隐私保护方案、模型保护方案、多模态大模型内容安全方案以及业务运营风控方案等几个核心维度提供安全防护能力；同时结合以攻促防的思路搭建了内容安全评测能力，对大模型开展例行化的安全评估。

图5.大模型应用安全架构图

3.1 大模型安全护栏建设

大模型安全护栏，分别从“语料安全、输入与模型生成安全”两个方向展开建设；首先针对大模型预训练阶段原始语料进行安全过滤，删除、屏蔽有害内容，保障大模型内生安全；其次通过对用户输入的prompt、输入行为、大模型生成的内容进行安全审核，以外围安全组件的方式构建大模型内容安全护栏。接下来分别从两个方向的主要功能、技术架构等方面进行详细介绍。

3.1.1 语料安全建设

大模型预训练语料库的质量和安全性是建设安全可靠的大模型服务最重要的一环。为了确保大模型在应用中不产生有害或不适当的内容，百度安全设计开发了一个大模型预训练语料内容安全清洗和增强模块，对大模型训练、隐私保护、合法合规具有重要作用：

- 降低重复数据导致的模型训练过拟合，提升模型泛化能力；
- 清洗风险数据（个人隐私、歧视、偏见、涉及侵权等）和不合规数据（涉黄、涉暴等），提升模型合规水平；
- 提升数据质量，清洗错误数据，提升模型质量和透明度，易于训练错误检视及模型准确性提升；

语料安全清洗过程如下图所示，首先对原始语料进行低质语料过滤，得到中间数据后进行数据去重，并通过训练好的多维清洗算子，对中间数据进行分类打标，最后将其合规化和格式化处理参与大模型预训练工作。

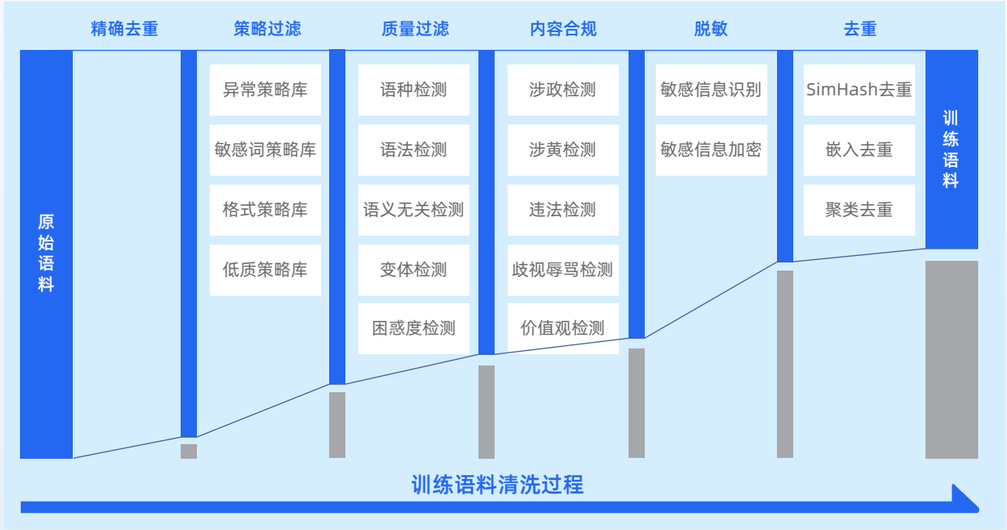

图6.大模型训练语料清洗流程

3.1.2 输入与模型生成安全

大模型输入/输出安全的技术方案主要集中在检测与过滤生成的不良内容，以及提示词注入攻击、算力消耗等恶意行为，并针对大模型特有的多模态组合风险、多轮对话中隐藏的指代风险等新型风险提供检测能力。

该解决方案突破现有的技术瓶颈，提高对新型、变种不良内容的识别能力，内置的多模态融合检测算子可识别文生图、图文生图等多模态场景下的组合风险，同时加强对大模型生成的安全范畴知识、事实错误类幻觉问题的识别准确性。此外，针对多语种、多轮对话中的上下文风险，我们也探索了更加有效的风险收敛方法。我们创新地构建了包含多语种安全、多模态组合识别、多轮对话安全、回复干预以及基于检索增强的安全代答等能力的大模型内容安全机制，该安全机制不仅使用了多道安全防线检测大模型对话全过程的风险内容，还构建了基于权威站点检索增强的安全代答模型，并通过多阶段对齐大幅度降低敏感问题的幻觉风险内容生成。

图7.大模型输入输出安全

安全分类算子

大模型输入的安全分类是指将用户输入内容进行分类，以判断其安全性和合适性。这种分类能够帮助防止不良内容的生成，保护用户免受有害、不准确或不适当的内容影响。通过有效的输入内容安全过滤，能够极大程度地减少大模型生成不安全或者负面的回复内容。

百度结合多年的业务内容安全分类实践，将输入内容划分为不同的主题类别和语义类别，由此构建出完整正交的标签体系，基于意图检测的审核模型，提供覆盖涉政、涉黄、违法等不同主题和恶意、攻击、中立、正常等不同语义的内容分类能力，能够高效检出涉政、涉黄、违法、歧视、辱骂、负面价值观等类别的不安全输入。

除此之外，针对对话内容中涉及的提示词注入攻击、违规网址、敏感信息等，模型也可以准确识别并完成拦截，避免违规信息的漏出；安全算子支持针对恶意prompt导致的模型生成过长token、接口高频访问等算力消耗攻击的检测，保障资源不被恶意消耗的同时，保障用户使用体验。

内容干预系统

大模型的内容干预是指通过人工审核、过滤技术或其他方式，干预模型输入的内容，以确保其符合特定的标准、规范和价值观。这种干预可以帮助减少有害、不准确或不恰当的内容，并提高生成内容的质量和安全性。

百度可提供完整的实时内容干预系统，内置红线必答和Query干预功能。红线必答能够很好回答常见的红线问题，确保回复内容高度安全合规，维护社会主义核心价值观；Query干预支持用户配置相应规则，通过对包含特定敏感词的快速匹配，将不安全Query引导至更加合适的处理流程中（例如标准回复模版），减少大模型在该Query输入下产生有害内容或者不正确数据。

值得注意的是，内容干预需要权衡大模型的自由创作能力与生成内容的质量和安全性之间的关系。过于严格的内容干预可能会大幅抑制大模型的创造性，而过于宽松则可能导致有害内容的生成。因此，掌握合适的内容干预尺度也对使用方提出了高要求，百度提供了相对审慎可用的预置策略，能够很好地兼顾大模型创新能力和回复内容的安全性。

3.2 大模型数据安全建设

大模型数据安全问题是关乎大模型相关企业的AI业务生存发展关键问题之一。大模型相关企业在开展新兴业务的同时，需要做好数据分类分级管控，加强敏感数据和机密数据保护，实施数据加密管控等基础保护措施，定期开展大模型业务的数据安全风险评估等，确保大模型新兴业务的持续健康发展。大模型数据安全工作在保护企业数据资产和大模型知识产权的同时，也有助于提高企业核心竞争力，成为行业数据安全标杆，树立良好的社会形象。

为积极落实《中华人民共和国数据安全法》、《工业和信息化领域数据安全管理办法(试行)》等数据安全相关要求，防范大模型全生命周期各阶段相关数据安全风险，增强大模型业务开展过程中数据安全综合能力，提升百度智能云千帆大模型平台等大模型生产服务平台的综合竞争力，百度将领先前沿数据安全与隐私保护技术与大模型生态相结合，形成Baidu AI Realm大模型安全技术框架，为百度智能云千帆大模型客户大模型业务提供端到端的数据密态管控与数据安全合规能力，覆盖大模型语料数据安全管理、大模型训练数据安全管控、大模型推理安全服务、大模型微调数据安全管理、大模型私有化数据资产保护等大模型生命周期各环节。

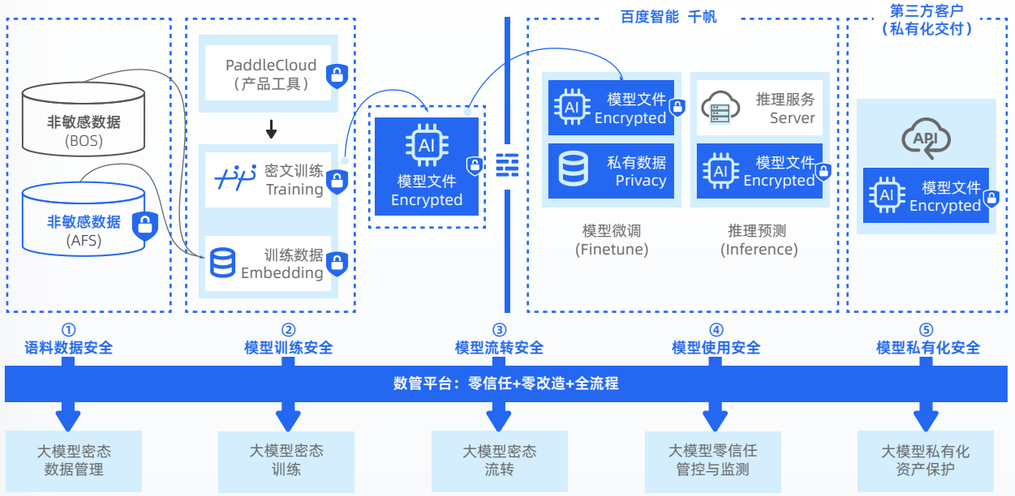

图8.大模型数据安全

3.3 大模型安全评测能力建设

大模型蓝军评测是一种主动的安全测试方法，旨在模拟攻击者的行为，评估大模型系统在真实威胁面前的安全性能与内容合规问题。蓝军安全评测的意义在于为大模型的业务运营提供全面的安全保障，增强系统的内容对抗能力，从而确保生成内容的安全性、完整性和可用性，大模型蓝军的核心能力如下:

3.3.1 建立自动化的攻击语料生成能力

我们利用开源的大型语言模型蓝军（红队）数据集及人工构建风险评测数据集作为基础，通过筛选其中具有高风险、高质量的评测题，利用改写大模型对关键词、文本段落、评测题自动化设计相似主题的评测数据，这种方法可以利用已有的数据集及互联网上收集到的数据，借助改写大模型生成能力来扩充语料库，增加测试数据的多样本和覆盖范围，提升威胁攻击预料的攻击成功率。

3.3.2 建立自动化大模型回复风险标注能力

生产了海量风险内容评测数据后，我们将评测数据输入被测大模型，获得大模型的对应回答。我们需要检测这些对应回答的风险情况，并汇总整体的回答内容风险得到被测大模型的整体风险情况。对于海量大模型输出结果做人工标注需要较大成本，因此我们探索一种可扩展的检测架构，可以自动化地对回答内容进行准确快速的风险检测:

- 模版匹配策略标注：多数大模型在检测到内容存在风险时，会使用固定的格式生成回答内容，可通过模版记录这类固定格式的回复，以快速豁免回答检测内容是否存在风险，提高风险标注效率。
- 裁判大模型：采用LoRA、P-Tuning等微调及强化训练技术，对大语言模型进行定向优化，打造面向风险标注的垂类模型，实现风险标注的自动化，全面提升标注效率与准确性。
- 黑白名单标注：基于预先配置的黑白名单词库，对大模型生成的内容进行自动化筛查和标签赋予。该方法可灵活设定敏感词（黑名单）和推荐词（白名单），实现对内容的精准管控。

3.3.3 建立大模型安全评测框架

为助力大模型内容风控系统升级，促进大模型生态健康有序发展，我们构建了大模型安全评测框架：通过设计全面且科学的评测标准与可量化的评估指标，输出包含评测方法、测试数据集、核心指标的详细报告，并基于实际检测风险点提出针对性改进建议，为风控升级提供技术依据；同时，在评测框架中融入自动化评测内容生成与自动化大模型回答评估能力，大幅提升风险识别效率与量化精准度。通过定期开展评测，实时追踪大模型内容安全能力的动态变化，及时定位大模型潜在的内容安全漏洞，全面保障大模型的安全内容输出能力。

04

百度AI基础设施安全合规

4.1 AI基础设施合规需求分析

当前AI基础设施建设在《算力基础设施高质量发展行动计划》等法规、制度促进和激励下，如雨后春笋般涌现；同时经过多年发展迭代，与建设初期相比，当前AI的基础设施建设呈现规模化、重运营、强监管的趋势特点，而这些趋势特点也从不同维度凸显了AI基础设施安全合规体系建设的重要性。

规模化	重运营	强监管
通过对当前AI基础设施建设项目披露信息的梳理，AI基础设施建设明确从分散的、小型的建设模式向规模化、万卡级别建设模式转变。 随着资产规模、应用数量的增加，需要同步考虑及规划的合规要求更多（涵盖云上、云下、跨域、多应用等），对建设方的经验也提出更高要求。	此前AI基础设施通常为先建设后运营，当前整体向“建运一体”模式转变，在建设初期就规划好消纳途径和规模，常需面向行业/公众提供算力租赁、订阅等服务，以提高算力利用效率；多样化的模型适配、应用场景、网络配置及安全管理需要，对AI基础设施的持续性合规、持续性安全提出了更高挑战。	AI基础设施建设大量使用GPU服务器、高性能网络设备等高价值资产，并且对电力供应等基础设施也有较高要求，因此在建设中，通常需要属地政府的大力支持，甚至由政府部门牵头指导；政务场景、关键行业相关的AI基础设施，必然要求满足相关行业、场景的合规要求，要求安全体系建设在合规及业务效率之间取得平衡。

基于百度参与的多个AI基础设施建设项目，以及与AI基础设施承建方、需求方的深度沟通，我们梳理了当前AI基础设施建设的合规要求，绘制形成**AI基础设施安全合规矩阵**，矩阵中列举了AI基础设施建设需要满足的核心常见合规要求，主要分为与基础设施同步规划、同步建设、同步使用的等保合规建设、密评合规改造的基础合规要求，以及大模型内容合规、数据安全合规、关键信息基础设施安全保护等核心的应用场景合规要求；前者通常会作为AI基础设施建设的必备要求和验收条件，后者随着AI基础设施的使用场景、算力消纳路径的确定，逐步进行安全能力的考量和建设。

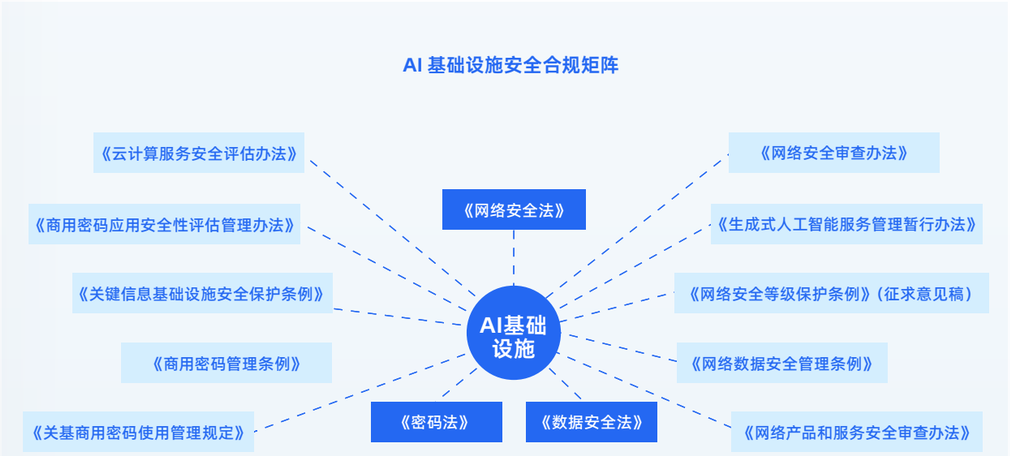

图9.AI基础设施安全合规

4.2 AI基础设施大模型安全合规

4.2.1 大模型安全合规背景

新的技术带来新的风险，当前大模型技术在各行业广泛应用的同时，也将新的技术、伦理等风险引入，而这些风险已超出企业自律和技术自治的能力范围，为此,国家网信办发布一系列管理办法及要求，为模型安全划定底线、统一要求、提供指导，保障国家安全、社会稳定和公众利益。

类型	施行日期	名称	简介及意义
国家标准	2025年11月	《GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求》	**百度深入参与**的“全国信息安全标准化技术委员会”牵头制定的安全基本要点，针对大模型的语料安全、生成内容安全、模型安全、安全措施做了详细要求，是大模型合规备案、安全评测最重要依据文件。
国家网信办	2023年8月	《生成式人工智能服务管理暂行办法》	提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估，并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。
国家网信办	2020年3月	《网络信息内容生态治理规定》	本规定所称网络信息内容生态治理，是指政府、企业、社会、网民等主体，以培育和践行社会主义核心价值观为根本，以网络信息内容为主要治理对象，以建立健全网络综合治理体系、营造清朗的网络空间、建设良好的网络生态为目标，开展的弘扬正能量、处置违法和不良信息等相关活动。

4.2.2 大模型安全合规建议

根据《办法》的相关要求，结合对当前主要应用场景的调研和了解，我们将目前在中国境内使用的大模型，**按开放范围**初步区分为两类合规场景并给出合规建议：

- 面向公众开放（含API接口、网页、App等形式）：必须进行大模型/算法备案并符合安全评估合规（符合GB/T 45654 GB/T 45438等标准的相关要求）方可对公众提供服务。
- 企业内部使用：可豁免备案，但建议参照GB/T 45654、GB/T 45438等标准进行合规建设，避免机密信息泄露、恶意信息窃取等模型攻击行为。

面向公众开放的大模型服务需在上线前完成大模型备案的相关工作，典型的对公开放的大模型如百度自研的千帆大模型等均已完成大模型及算法备案，**百度更是首批通过大模型备案的模型厂商，对大模型合规备案具有丰富的实践经验。**（注：本章节重点讨论大模型作为业务系统独特的合规要求，除此之外作为信息系统也应遵循和符合《网络安全法》《数据安全法》等常见的网络安全合规要求。）

类型	大模型备案
全称	生成式人工智能（大语言模型）上线备案
主要法规要求	《生成式人工智能服务管理暂行办法》
提交途径	线下提交至属地网信部门，如：所在地为北京市的向北京市网信办提交申报材料
审核部门	由企业所在属地网信办初步审核，最后中央网信办终审
备案对象	微调+ToC 涉及“舆论属性或者社会动员能力”AIGC产品的；面向境内公众提供生成文本、图片、音频、视频等内容的服务的
提交资料	《生成式人工智能上线备案申请表》、《附件 1：安全评估报告》、《附件 2：模型服务协议》、《附件 3：语料标注规则》、《附件 4：关键词拦截列表》、《附件 5：评估测试题集》
审核方式	材料审核+技术测试
备案结果	国家网信办官网公告或当地网信办下发备案告知书 https://www.cac.gov.cn/2024-04/02/c_1713729983803145.htm

大模型备案从流程上主要分为备案准入评估、属地报备指导对接、自评估及材料准备、省级初审、国家复审、备案及公示六个阶段，其中主要的技术难点工作包括大模型合规能力建设、安全自评估以及配合网信部门的测试评估。

因此，为配合AI基础设施中大模型的合规应用，百度基于自身丰富的备案经验，提供大模型备案的全流程咨询服务，服务具体可包含安全方案设计、评估方案撰写、大模型内容安全评测等工作，旨在协助大模型快速、高效的完成备案相关工作。

- 安全方案设计：百度大模型安全领域的专家团队，结合百度多年的大模型安全实践经验，发现模型安全不合规项并提出改进建议，协助完成大模型内容安全方案设计（方案建议可参考3. 百度大模型应用安全）；
- 大模型内容安全评测：根据《办法》的要求并结合GB/T 45654中的风险项，从多角度对大模型服务展开内容安全评测，涵盖政治敏感、违法犯罪、歧视偏见、侵犯版权、侵犯隐私、伦理道德、不当内容以及恶意利用等维度，利用数十万的评测数据集以及高级提示词攻击指令，全方位对大模型服务做安全评测，并产出详细内容安全评测报告；
- 评估方案撰写：协助进行模型安全的自评估工作，结合评测报告、合规要求及方案设计形成大模型安全自评估报告。

4.3 AI基础设施等保密评合规实践

4.3.1 等保合规

网络安全等级保护（等保）是我国网络安全战略核心制度，通过法律化、制度化手段保障网络安全。《网络安全等级保护条例》要求AI基础设施运营者按等保要求（多数云平台为三级）合规建设及测评，否则将面临罚款、业务暂停等处罚。

而针对在AI基础设施推广应用的金融、医疗、政务等关键行业领域，行业监管机构也对AI基础设施等关键信息基础设施的合规性出台了明确指引/标准（如《金融行业网络安全等级保护实施指引》），未通过等保测评的AI基础设施将不能承载相关行业的业务，因此可以说，AI基础设施的等保合规是其正常运行和对外运营的前提。

AI基础设施的整体等保合规工作，按照环境隔离、责任分工等因素，区分为**云平台自身等保合规，以及为云平台上业务应用系统提供符合等保合规要求的安全服务能力。**

4.3.1.1 百度云平台等保合规体系

百度云平台等保合规体系，主要围绕以云管理平台Console、云运维平台NoahEE为核心的云平台软件系统及其管理使用的基础设施进行安全能力建设，除基础的安全管理要求外，技术方面主要根据相关标准分为安全物理环境、安全区域边界、安全通信网络、安全计算环境和安全管理中心五个方面。

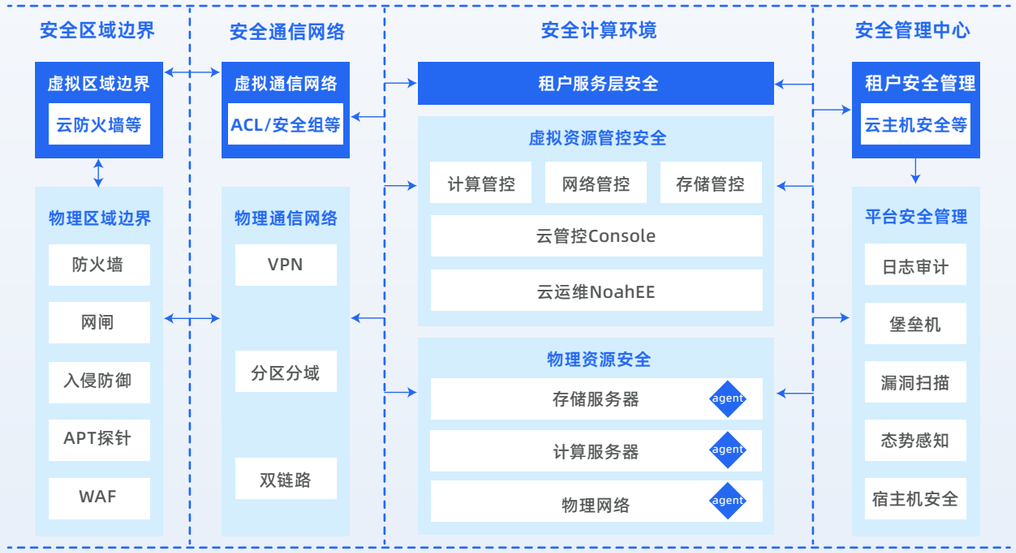

图10.云平台等保合规安全架构

安全物理环境

安全物理环境与被保护系统所部署的数据中心相关，在云平台建设过程中，需考虑机柜数量、通信距离、电力价格等因素，选择国内具备资质的运营方提供的符合等保要求的标准机房。

安全通信网络

在百度云平台的物理网络设计中，遵循以下措施以保障符合安全通信网络的要求，包括：

- 设备、性能与带宽冗余，关键网络链路及设备保持冗余，采取双链路模式，使用双主/主备的高可用策略，避免单点故障；针对网络流量变动较大的链路，如互联网出口、核心业务区保持常态带宽占用低于50%；针对关键的串联网络设备，选型保持性能冗余，业务峰值占用低于设备整体性能的80%，日常均值占用低于70%；
- 分区分域和边界隔离，百度云平台的物理网络设计，根据区域功能划分为互联网接入区、管理区、核心业务区等不同的网络分区，并通过配置网络策略、分配不同网段等方式实现区域基础隔离；而针对不可信区域，如互联网区、专线接入区与核心业务区之间的网络，通过部署「防火墙」、「网闸」、「光闸」的方式进行更高强度的网络隔离和更精细的边界访问控制；
- 加密传输链路，通过部署「VPN」设备搭建远程访问云平台的加密传输链路，从用户接入身份的安全性（数字证书、账号口令等多因素身份验证）、终端设备的合法性（用户端部署）、访问业务系统的权限合法性（服务端鉴权）、业务数据传输的安全性（对称加密算法+非对称加密算法，保障传输的完整性和保密性）等多个层面保障用户跨互联网远程接入的安全。

此外为满足云计算安全扩展的要求，百度云平台的虚拟网络设计支持：

- 面向云上用户开放 VPC 配置能力，不同用户之间网络隔离，针对VPC内部可通过ACL、安全组等方式进行业务访问路径和安全策略的精细配置。

近年来随着云计算技术和应用的愈发成熟，等保测评对于云计算安全扩展要求的测评趋向于更加严格，根据2025年最新的高风险项指引，云计算安全扩展要求中的多项要求被标记为高风险。

安全区域边界

在百度云平台网络中，建立安全区域边界的主要方式包括：

- **边界防护与访问控制**，在互联网出口、核心业务区边界、专线接入区等位置部署「防火墙」，实现跨区的受控访问，默认拒绝非授权的跨区访问请求，并且能够基于五元组信息、应用协议信息、会话信息等进行灵活的访问策略配置；针对更高隔离要求、较低业务交互要求的场景，如政务云访问场景，也可部署「网闸/光闸」进行更强力的边界防护和访问控制；
- **网络入侵、恶意代码传播等攻击防护**，在关键网络节点部署「入侵防御系统」，基于规则或高级检测引擎发现网络中的异常流程和潜在攻击行为，进行安全告警并支持通过策略修改或联动防火墙的方式，阻断恶意攻击流量；
- **高级威胁攻击防护**，针对具有互联网可访问业务的高风险场景，在互联网出口或物理服务器中部署「APT探针」，针对新型、可持续高级威胁攻击，能够基于攻击特征和智能检测引擎进行更深度的安全检测，并通过流量镜像的方式采集和存储原始流量（包含五元组信息），便于后续的研判分析和溯源；
- **安全审计**，为实现针对重要用户、安全事件的安全审计，关键的网络设备、安全设备应将日志统一转发至「日志审计系统」中，进行日志的统一存储和分析，必要时可通过日志审计与存储服务如百度BOS的打通，保障关键日志存储时长满足六个月的合规要求，并具有多重备份。

此外为满足云计算安全扩展的要求，百度云平台的虚拟网络区域边界支持以下功能：

- 百度云平台为云上用户提供「云防火墙」，结合基础的ACL、安全组功能，实现虚拟网络边界的访问控制能力；
- 针对入侵攻击行为，百度云平台提供「云流量审计」，针对流量中包含的攻击特征、恶意文件进行检测和告警；
- 针对云上敏感操作，百度云平台为云上用户提供「云堡垒机」，用户可以通过堡垒机进行虚拟机的配置修改，保障敏感操作均被记录。

安全计算环境

在百度云平台的等保合规建设中，保护对象包含云平台系统及其纳管的物理硬件资源，以及主要的管控和公共服务，具体可采用以下措施，打造百度云平台的安全计算环境：

- 云平台系统（在等保测评中，通常指云管理平台Console、云运维平台NoahEE）。
  - 具备完善的身份鉴别配置能力，需开启多因素身份验证、口令复杂度要求等策略增强身份鉴别能力，并且具有验证码、会话超时限制等功能，避免暴力破解和信息窃取；
  - 在访问控制方面，云平台系统支持基于用户账号分配对应的权限，云运维人员需要基于权限进行账号能力的控制，避免非授权访问、越权访问行为；
  - 此外用户在云平台系统的操作行为均会被记录和保存在云平台中，并且支持联动存储进行持久化和备份，存储系统支持基于「密钥管理系统」中的数据密钥创建加密存储桶，进行数据的加密存储；
  - 在数据传输的保密性和完整性上，云平台系统支持配置「SSL证书」，与用户终端之间建立基于HTTPS的加密传输链路；
  - 而针对潜在的外部入侵攻击和恶意代码侵入，云平台的Web访问请求经过「WAF」的安全检测，能够识别并阻断常见的OWASP top10的攻击行为，保障云平台的应用安全；
  - 而针对云平台的脆弱性（弱口令、Web漏洞、组件漏洞等），通过部署的「漏洞扫描系统」进行定期的检测，并基于检测报告进行漏洞的闭环修复，保障云平台的安全稳定运行。
- 其他关键设备和系统（关键网络设备、安全设备、服务器和数据库系统等）
  - 关键设备和系统的运维操作，均需通过「堡垒机」进行统一授权和管理，完善身份鉴别和访问控制能力，而在远程运维场景下需先登录「VPN」完成初次身份鉴别后再登录堡垒机获取对应资产的运维权限；
  - 关键设备和系统的日志应统一集中到「日志审计」进行统一的收集、存储和备份，满足关键日志保留180天的要求，并支持日志的检索和分析；
  - 关键服务器如计算宿主机、存储服务器、云管控服务器等，均部署「宿主机安全」，检测入侵攻击行为、防范恶意代码扩散，且宿主机安全支持登录管理能力，能针对服务器中的异常登录行为告警。

此外为满足云计算安全扩展的要求，保障安全计算环境，除上述能力外云平台还应支持以下能力：

- 云平台与云上租户系统之间应保持明确的隔离与边界划分，云平台管理者无法在未授权情况下，修改云上用户的数据，并且需要保障在云上用户删除数据时，数据将被彻底清除；
- 针对云上虚机和镜像安全，除云平台本身应具备镜像完整性校验、镜像加固、操作审计等功能外，应提供「云主机安全」能力，进行虚机上入侵攻击行为和恶意代码传输的检测；
- 此外针对密钥托管和云数据加密需求，云平台应提供「云密钥管理服务」，从而保障云服务或云上用户应用可调用密钥进行数据加密存储、数据加密传输。

安全管理中心

云管控和云运维组成的云平台系统是整个云的管理中心，能够基于不同用户账号的角色分配对应的权限，服务于系统管理、审计管理、安全管理等不同场景需要，而针对各种各样的安全设备、服务可以通过部署「态势感知」作为云平台安全管理中心的补充。

- 基于态势感知及「日志审计」「堡垒机」「数据库审计」等审计设备，能集中收集和分析来自不同设备、系统的日志信息，实现对网络中各类操作行为的全面审计记录，便于追踪溯源。
- 此外基于「态势感知」对安全设备、网络设备的管控能力，可以实现安全策略的统一下发和配置修改，保障策略落地的准确性和一致性，快速响应安全事件。

此外为满足云计算安全扩展的要求，建立安全的管理中心，除上述能力外百度云平台还提供以下能力：

- 云平台系统能进行统一的资源管理，与云用户之间可以基于自身责任的不同，进行各自资产的配置管理和操作审计。

4.3.1.2 面向等保合规的百度云服务

运行在云平台上的各类业务应用，需满足等保合规要求并通过测评。业务应用合规的责任主体为该应用的使用和运营方，具体而言，公有云、行业云场景下为独立第三方，私有云场景下通常为对应的业务部门或行业用户。这些责任主体需依托云平台提供的安全产品/服务或采购第三方安全产品来通过等保测评。不过，通常情况下，云租户业务应用的等保合规建设能够基于云平台的合规基础，并且等保定级等级要小于等于云平台的定级等级。

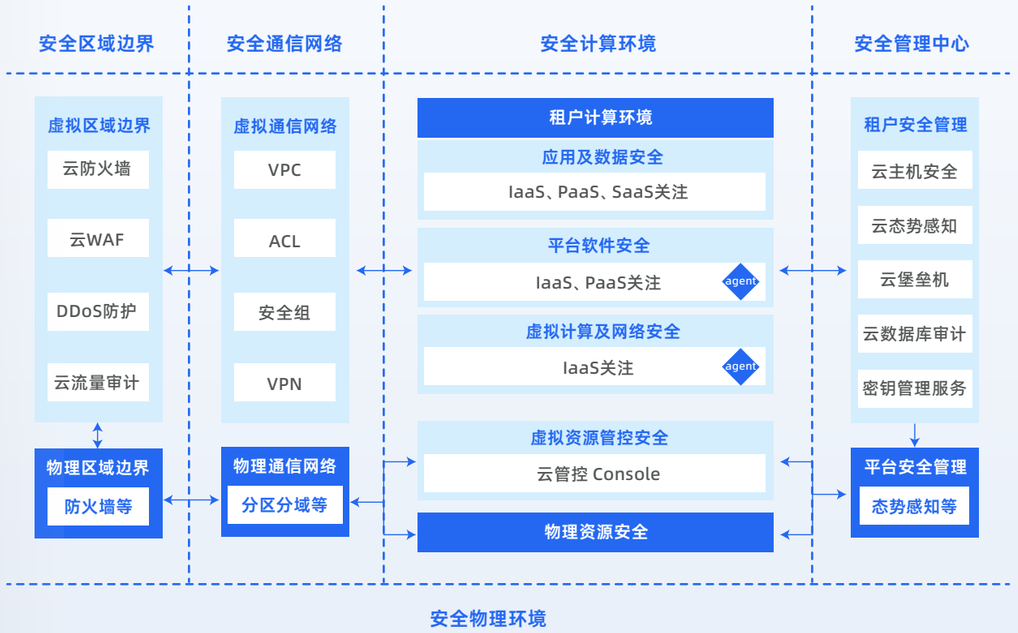

图11.云服务等保合规安全架构

因此云服务的等保合规重点关注用云模式和“一个中心、三重防护”安全能力的建设情况，分别为：

虚拟区域边界

在已有物理网络边界的基础上划分虚拟区域边界满足不同云租户的边界访问控制需要，主要使用包括「云防火墙」（提供VPC边界的流量管控，进行网络入侵行为检测等）、「DDoS基础防护」（提供DDoS攻击流量检测和清洗功能）、「云WAF」（提供Web侧的安全防护边界，阻止恶意Web攻击流量）、「云流量审计」（提供流量镜像采集和深度分析等功能）等云上安全产品，构建安全的区域边界并实现合规。

虚拟通信网络

在物理网络基础上将云上流量通过VxLAN等方式进行封装转发，实现隔离的虚拟通信网络，而虚拟网络内部基于私有网络VPC进行基础的用户隔离，而针对VPC网络内部可以适用ACL进行网段划分隔离不同的业务环境，并通过安全组进一步限制虚机的访问控制；此外针对传输机密需求，采用SSL/TLS协议对数据在通信网络中的传输进行加密，防止数据在传输过程中被窃取、篡改或监听，并基于云VPN 构建远程访问的安全通信网络。

租户计算环境

借助「云主机安全」对云主机实施系统防护与深度加固，有效抵御恶意代码入侵主机系统，利用其漏洞扫描功能，及时发现并修复系统漏洞，防止黑客利用漏洞发动攻击。「云WAF」可针对Web应用进行深度防护，精准识别并阻断SQL注入、跨站脚本攻击（XSS）、命令注入、文件包含等OWASP TOP 10常见Web攻击，防止攻击者利用Web应用漏洞获取敏感数据、篡改页面内容或控制服务器。「密钥管理KMS」生成的数据密钥对重要数据进行加密存储，保障存储的完整性和保密性；并可使用如SSL证书，保障数据传输过程中的保密性与完整性，防止数据被窃取或篡改。

租户安全管理

主要包括安全设备的集中管控及平台整体的安全审计，其中集中管控可使用云平台上配置的云监控BCM，监控计算、网络及原生云安全服务的运行状态，确保实例的稳定正常运行，此外针对云上各类安全告警可通过「云态势感知」进行统一的收集和呈现，便于关联分析和处置；而在安全审计方面，可以使用云平台的云审计BCT、「云堡垒机」和「云数据库审计」针对云平台的敏感操作、虚机的运维操作和数据库的调用行为等重要操作进行记录和审计。

4.3.2 密评合规

商用密码应用安全性评估（以下简称“密评”）是国家为了保障网络空间安全，特别是关键信息基础设施安全，通过法律法规强制要求对重要信息系统中的密码应用进行合规性、正确性和有效性的专业评估，是落实《密码法》的重要手段，也是构建国家网络安全综合防护体系的关键环节。因此在当前AI基础设施建设中，需要遵循《关键信息基础设施安全保护条例》《商用密码管理条例》等合规要求，在AI基础设施建设、规划和运营阶段同步考虑密评相关工作，保障所使用的密码算法符合商用密码应用要求。

密评相关改造当前主要参考《GB/T 39786-2021信息安全技术信息系统密码应用基本要求》及配套的各类标准进行，在该标准中，将密码应用工作主要切分为应用和数据、设备和计算、网络和通信、物理和环境四个技术部分和管理制度、人员管理、建设运行、应急处置四个管理部分，结合百度在AI基础设施建设中的实践经验，可将密评技术改造工作分为云平台密评改造，及云上应用（如模型平台、模型应用等）密评改造两部分。

4.3.2.1百度云平台密评合规改造

基于GB/T 39786-2021标准体系，结合百度专有云ABC Stack云平台实际情况，我们建设了密码基础资源池，并对云平台核心系统（如云管理平台Console、云运维平台NoahEE）实施密码改造，最终构建起覆盖核心系统真实性、完整性、机密性及不可否认性的商用密码服务体系。值得说明的是，百度云平台已于2022年首批通过新版密评测评，具备丰富的合规改造经验。

- 密码基础资源池建设，增加具有“商用密码产品认证证书”的专用密码产品，为云平台及云服务提供可调用的密码服务接口，支持商密证书生成、签名验签服务、数据加解密、数据特征值生成等能力，组成云平台国密基础服务体系；
- 云平台密码应用改造，基于各密码产品构建的密码基础服务平台为上层云平台密码应用提供密码技术支撑，以便针对身份鉴别、敏感数据传输和存储加密、关键信息传输和存储完整性验证等密码应用场景进行商密的改造、增强。

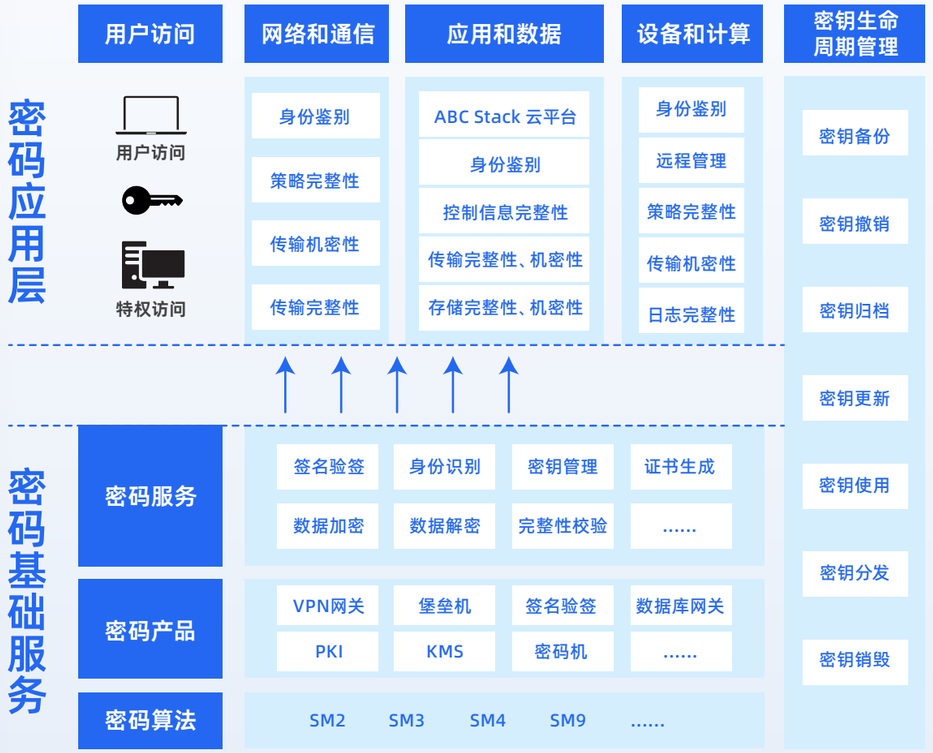

图12.云平台密评合规安全架构

物理和环境安全

基于密评要求需要替换或改造成符合国密要求的国密电子门禁系统，采用消息鉴别码和数字签名机制进行机房出入人员的身份鉴别，以及对电子门禁系统进出记录等数据进行完整性保护。此外通过部署国密视频监控系统或对已有的视频监控系统进行国密化改造，能够基于国密算法对视频监控数据存储和使用过程进行完整性验证，符合密评和要求，避免被有意篡改或删除。

网络与通信安全

网络与通信的密码应用安全改造，需要在网络接入的边界（虚拟边界）部署国密的VPN综合安全网关，配合部署在用户PC中的远程终端，对访问网络、云平台系统（云Console）的用户进行身份鉴别和权限控制，对传输的数据进行机密性和完整性保护。

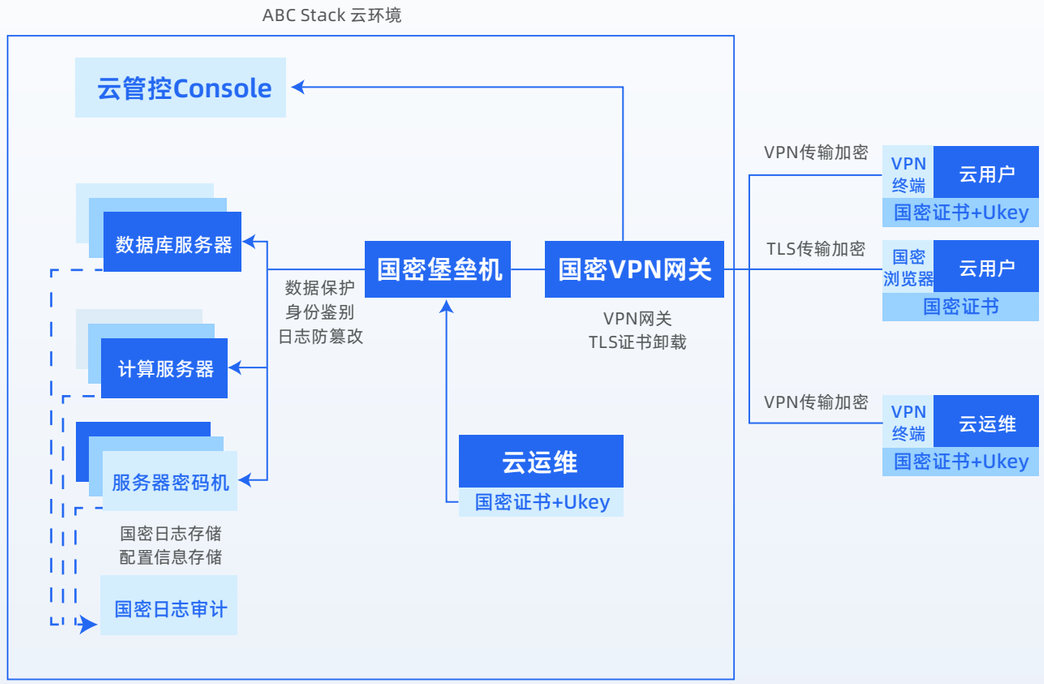

图13.云平台密评网络通信安全

云用户与百度云ABC Stack云平台的通信安全

- 身份鉴别：在本地机房部署VPN综合安全网关，并向单位用户配发已导入关联国密证书的智能密码钥匙（以下简称国密Ukey），用户终端使用国密Ukey+VPN客户端登录VPN网关，以实现对云Console的访问，基于VPN客户端与VPN网关的双向身份鉴别实现对用户身份真实性的保障。
- 通信数据完整性和通信过程中重要数据的机密性：所有针对云平台的访问行为均经过VPN综合安全网关的代理，避免对公网的直接暴露，而基于VPN网关与客户端之间建立的基于国密算法的非对称加密传输信道，能够保障身份鉴别等重要数据传输的完整性和机密性。
- 网络边界访问控制信息的完整性：VPN综合安全网关作为经过认证的商用密码产品，通过自带国密加密卡能够在访问控制信息写入、存储时生成消息鉴别码，通过对特征值的校验保障完整性不被破坏，完成对访问控制信息进行完整性保护。
- 安全接入认证：基于VPN客户端及对应的国密证书，不仅能够针对用户身份进行真实性校验，也可以针对登录设备进行安全接入的认证检测，拒绝未授权设备的访问。（非必须项可根据实际需要决定是否纳入测评范围）

运维人员与云平台底层硬件基础设施的通信安全

- 身份鉴别：在本地机房部署VPN综合安全网关，并为运维人员配发含有运维权限的特殊证书，并导入国密Ukey，实现运维人员登录的安全身份鉴别，保障真实性并赋予对应权限。
- 通信数据完整性和通信过程中重要数据的机密性：在机房部署VPN综合安全网关，实现运维人员与云平台底层硬件基础设施之间通信数据的机密性和完整性保护。
- 网络边界访问控制信息的完整性：在机房部署VPN综合安全网关，利用VPN综合安全网关本身的机制完成对访问控制信息进行完整性保护。
- 安全接入认证：运维人员通过VPN综合安全网关连接到密码基础设施机房，采用VPN综合安全网关实现安全接入认证。（非必须项，可根据实际需要决定是否纳入测评范围）

设备与计算安全

设备和计算安全主要指云管理/云运维人员在对云平台进行运维时，需要针对运维人员进行身份真实性鉴别，并保障网络环境中主机系统、存储资源以及其上所承载的应用程序等重要数据的机密性和完整性；在百度云平台环境中设备主要包括通用设备（计算服务器、存储服务器、数据库服务器等）、网络设备（路由器、交换机等）及安全设备（垒机、VPN等）。

- 身份鉴别及访问控制信息：部署国密堡垒机，运维人员基于账户密码+UKey进行国密堡垒机的登录，并在登录具体设备时，校验签名证书，保障身份鉴别的真实性，而针对国密堡垒机上访问控制信息可以基于自带的国密加密卡实现完整性的校验避免被恶意篡改；
- 远程管理：针对云平台服务器、网络设备、安全设备的远程管理和运维操作，需要先登录VPN综合安全网关，通过端到端加密链路保护，建立信息传输的安全通道；
- 日志记录：部署国密日志审计系统，系统资源、平台设备和密码设备的日志统一发送和存储到日志审计中，调用密码机的密码服务接口，基于国密算法保障日志记录的完整性；
- 重要可执行程序：云平台中关键的可执行程序，一般包括实例镜像、快照等，针对由云平“台调用签名验签服务，基于签名验签保障重要可执行程序（如操作系统镜像）的完整性和真实性。

应用与数据安全

百度云平台由两个主要的子系统构成，包括云管控Console平台和云运维NoahEE平台，因此应用和数据的密评改造主要针对云Console、云NoahEE两个主要子系统和配套基础设施进行密码应用安全性的改造。

图14.云平台密评应用与数据

改造工作包括密码设备上架及云平台改造两部分，首先需要在云平台区域部署服务器密码机、签名验签服务器和数据库加密网关，并且在云上的公共管理VPC环境中，将数字证书认证系统和国密密钥管理系统部署在虚机计算实例中，为云平台应用搭建专属的密码资源。随后基于资源池的SDK或API接口，进行平台身份鉴别流程改造、平台访问控制信息完整性校验改造、重要数据（在云平台中，主要是身份鉴别信息、配置信息）传输机密性和完整性改造、重要数据存储机密性和完整性改造等工作，以保障云平台对商用密码应用的合规有效。

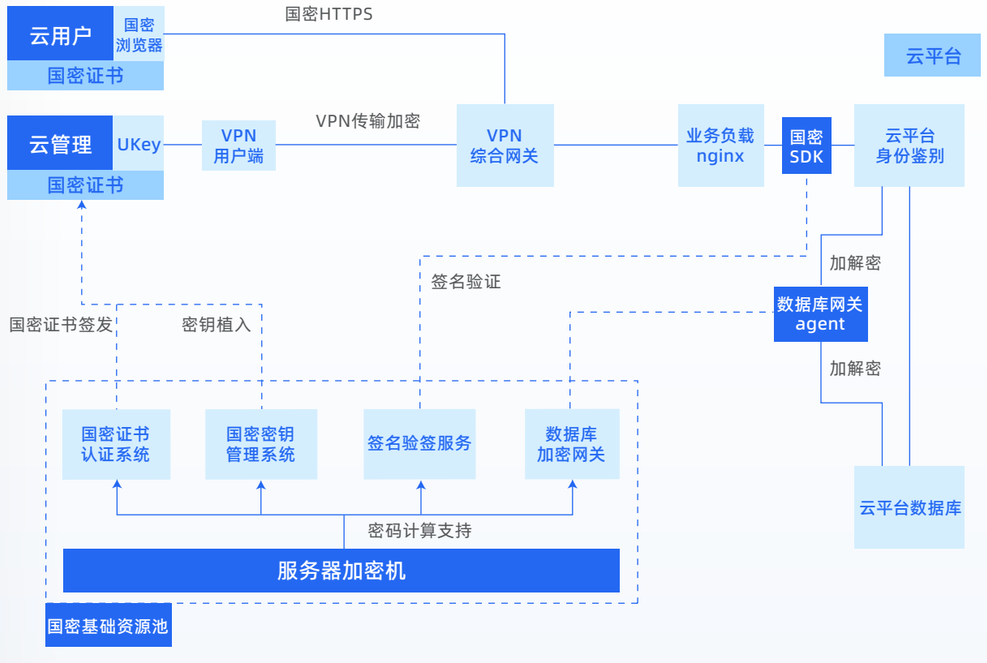

图15.云平台密评应用与数据安全流程

- 身份鉴别：云平台管理和操作人员，使用带有证书（由国密证书认证系统签发）的UKey登录Console和NoahEE，在密评改造后云平台系统支持通过调用签名验签服务接口的方式，验证证书及Ukey的特征口令，保障云平台登录用户身份的真实性和不可否认性；
- 数据存储完整性和机密性：部署数据库加密网关，通过代理云平台访问基础数据库的链路，进行写入、读取数据过程的加解密，使用HMAC-SM3对敏感数据如用户访问权限控制列表进行存储完整性保护；使用SM4针对重要数据（如身份信息、操作审计数据等）存储进行加密，平台系统调用需要敏感数据需要经过代理的加解密和完整性校验。
- 数据传输完整性和机密性：在应用层通过改造使用基于签名证书的SM2+HMAC-SM3算法进行传输完整性的保障；而考虑到应用侧传输加密的性能和可用性要求，利用网络侧的VPN端到端加密链路（SM2、SM3、SM4）+应用侧RSA2048 算法，保障数据传输机密性符合商用密码应用安全性的要求。

4.3.2.2面向密评改造的百度云密码服务

在AI基础设施建设过程中，为保障和支持后续的模型应用场景，通常需要部署模型训练推理平台、行业模型应用等多个业务系统，因此百度云也为云上应用提供了成套的云密码服务，能够满足云上应用密评改造需求，并基于标准化的密码服务，降低整体改造工期。

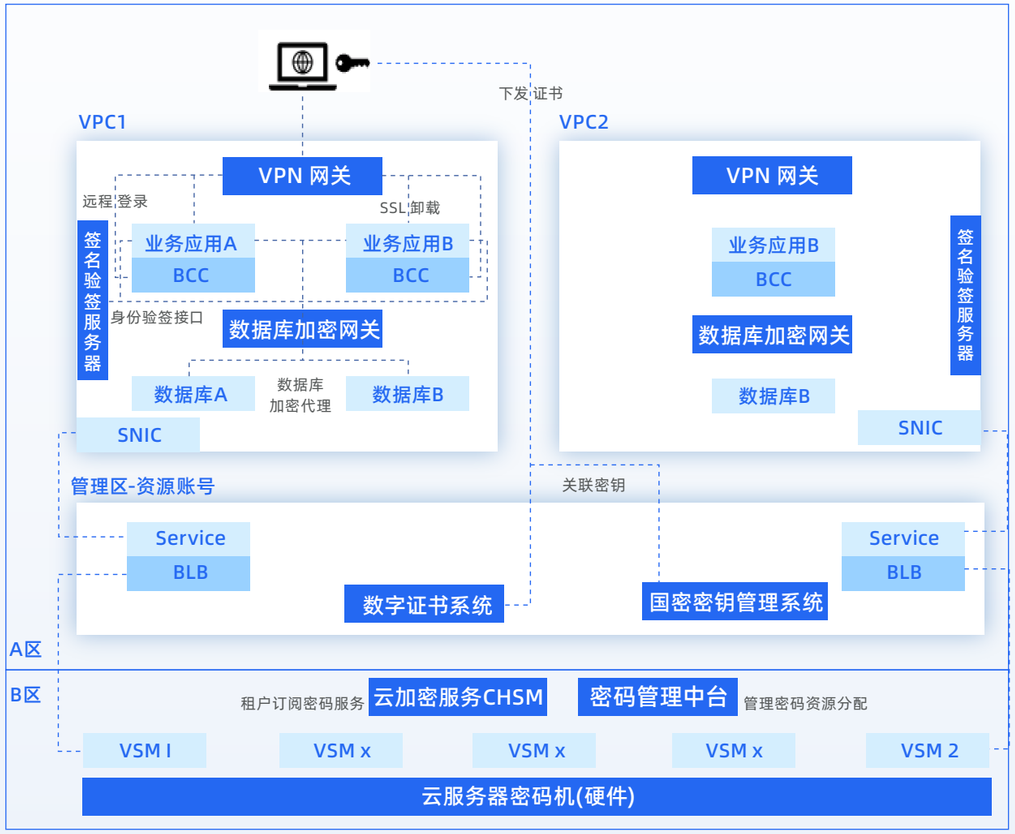

图16.云服务密评业务逻辑

云上网络与通信安全

通过在虚拟私有云（VPC）内部署国密VPN安全网关，构建符合国密标准的加密通信通道，系统性解决云用户与云资源间的传输安全需求，建立端到端的国密加密访问链路，依托国密算法实现三重核心防护：

- 保障访问人员真实性，通过国密VPN安全网关建立与运营、管理等人员终端的传输加密链路，相关人员登录VPN网关需在账号密码的基础上，使用智能密码钥匙（U-key）或国密证书进行身份的再次验证，而VPN安全网关也将基于此前分发的U-Key中的国密证书的信息，来确认登录人员的身份信息，能够有效保障访问人员的真实性。
- 保障数据传输完整性，国密VPN安全网关在通信过程中执行SM3-HMAC运算，为每个传输数据包生成唯一特征值（消息认证码）。当数据抵达目标端时，网关立即进行特征值比对验证，确保从用户终端到云资源的全路径数据完整无损。
- 实现数据全程机密性防护，网关创建国密VPN加密隧道，采用双算法协同机制：通过SM2算法进行密钥协商与数字证书认证，建立安全会话通道；利用SM4对称算法对传输载荷进行高强度加密。用户终端发出的数据在进入互联网前即被转化为密文，仅在抵达VPC边界的VPN网关时才进行解密卸载。

国密VPN网关需要部署在VPC中的虚机实例中，并通过挂载EIP的方式暴露访问方式，便于运营人员、运维人员等通过VPN登录到云内VPC环境，相应的访问控制信息将保存在国密VPN安全网关中；此外国密VPN网关需要调用云密码机提供的接口，进行国密算法运算，进行访问控制信息的完整性保障，并支撑进行身份鉴别、传输加解密等，而且通过原生的云密码机能力，租户可独占虚拟密码机实例，保障密码资源和敏感信息的租户隔离。

云上应用与数据安全

为构建符合国密标准的云上应用安全体系，需系统性实施以下改造措施：

- 在身份认证层面，所有通过国密加密通道访问应用的用户（包括普通使用者、系统管理员及安全审计员），须采用国密U-Key+数字证书或国密安全浏览器+数字证书的双因子验证机制。该机制将为每位有国密需求的用户颁发国密数字证书，当用户连接国密VPN通道时，系统通过SM2等国密算法实时核验证书真伪，从根本上杜绝身份冒用风险，确保登录者身份的真实性。
- 在防篡改层面，云上应用需通过调用VPC内部署的签名验签服务，对自身存储的访问控制策略执行HMAC-SM3运算，相当于给关键配置生成独特的“数据指纹”。每次调用策略时自动校验指纹完整性，可即时发现非法篡改行为。同时，对于身份信息、个人敏感数据等重要信息资源，同样采用HMAC-SM3算法生成校验值并进行校验。
- 在数据传输层面：通过在VPC边界部署国密传输加解密服务（国密SSL卸载网关），自动对进出应用的重要业务数据实施SM4实时加密，使传输内容始终处于密文状态，仅在抵达应用内部时才进行安全解密，便于应用之间调用，能够在满足密码应用要求的前提下，尽量降低对应用的改造。
- 在数据存储层面：实施双重保险机制，一方面通过数据库加密网关建立proxy代理层，当应用访问数据库时，网关自动对身份证号、交易金额等关键字段进行SM4加密/解密，实现“数据使用即解密、存储即加密”的透明防护；另一方面启用云平台存储产品的国密加密功能，基于SM4算法对云硬盘、对象存储中的静态数据进行全盘加密。覆盖身份凭证、核心业务记录、审计日志、个人信息等敏感数据类型，其保护范围可根据业务风险动态调整，满足不同应用对数据存储过程中完整性、机密性的要求。

05

百度AI基础设施安全管理与运营

随着数字化、智能化进程的不断加快，在AI基础设施、平台服务到上层应用的各个层面都面临着不断升级的安全挑战。例如，多租户共享的环境增加了隔离防护难度，AI模型训练过程中的海量数据和模型资产需要严密保护，大规模集群的内部网络和对外边界防护变得愈发复杂，同时还需满足严格的法律法规和合规性要求。在此背景下，如何对算力中心实施有效的安全运营管理，成为运营单位亟待解决的重要课题。

5.1 AI基础设施安全运营管理

AI基础设施安全运营管理强调对威胁态势的实时感知和对攻击事件的全链路处置。它通过安全态势感知平台与部署在平台各处的安全防护产品（如主机防护、网络防火墙等）的协同，以及专业安全团队的7x24小时值守运营，来实现对安全威胁的高效发现和综合处置。安全运营管理贯穿安全事件发生的整个生命周期，可分为事前、事中和事后三个阶段，每个阶段各有侧重又相互衔接，形成闭环的防护体系：

百度AI基础设施安全运营的目标主要是围绕四个方面进行：

- 更高的风险可视性：通过资产梳理、漏洞扫描、威胁检测等方式，全局掌控算力中心风险分布和威胁态势。
- 更快的威胁响应速度：依托安全监测与告警能力，配合专业运营团队或驻场人员，保证在最短时间内完成事件处置。
- 更强的防御韧性：通过持续的安全运维和产品加固，保障算力中心算力的运行稳定性与合规性，抵御高强度持续攻击。
- 更低的安全使用门槛：在最大限度提升算力中心防护安全水位的同时，为算力租户提供易用且灵活的安全配置、安全服务与合规支持，降低整体安全成本。

图17.AI基础设施安全运营体系

5.2 安全运营成功的关键点

要想在算力中心复杂的环境中真正落地高效的安全运营管理，还需要从人员、流程、技术多个角度具备相应的能力和资源。总结业界实践经验，百度算力中心依托以下几个关键要素达成安全运营的成功：

丰富的日志与数据的采集

- 汇总防火墙、IPS/NDR、WAF、HIDS、系统日志等海量数据；
- 与外部威胁情报平台联动，引入域名/IP/URL/Hash等恶意情报；
- 应用UEBA等技术进行用户与实体行为分析，识别滥用算力的异常行为。

科学高效的告警研判与分析流程

- 威胁检测：基于规则匹配、流量审计、异常行为基线、机器学习等多种检测手段进行综合分析；
- 告警分类：按照告警严重程度和类型进行自动或人工分级；
- 事件研判：由运营人员对可疑告警进行深度分析，判断是否为误报并追踪攻击链。

分级分层的安全运营策略

- 普通级别告警：自动处置或在服务时间内由驻场人员进行检查；
- 高级别告警：在30分钟内启动安全应急流程，由专人牵头进行封堵、溯源、修复等操作，必要时联系相关部门进行快速决策；
- 紧急重大事件：7×24小时值守，必要时进行现场支援、停机隔离、数据备份及系统恢复等。

5.3 持续运营改进与业务保障

百度算力中心提供的安全运营管理服务日常通过周报/月报的形式与算力运营单位共同进行运营效果的优化提升，通过聚焦关键风险、主要告警趋势、处置情况、改进建议等，包括但不限于不断完善检测规则、缩短误报与漏报之间的差距；优化事件处置流程与自动化响应脚本，提升协同效率；引入更多安全能力（如SOC、SOAR），实现安全运营效果的升级和质变。客户可以通过百度安全运营管理的成果，系统的梳理算力中心安全态势演变、安全投资ROI、重大事件回顾等，帮助和支持进行未来的安全规划。

06

百度AI基础设施安全实践案例

6.1 某地方万卡集群算力中心安全建设案例

安全痛点与建设目标

某省级政府为推进人工智能产业发展，主导建设超大规模算力中心该中心部署万卡级GPU集群，承担算力售卖、多租户运营等核心业务，覆盖智慧城市、科研教育等新场景，涉及算力基础平台、支持平台及运维运营平台的协同运转。核心安全挑战集中在三方面：

- 互联网攻击风险：互联网环境中存在海量恶意攻击，如网络渗透、算力盗用等风险，对算力中心的业务连续性和数据安全构成威胁；
- 合规压力突出：需满足等保2.0三级、密码应用安全性评估（密评）等刚性要求，覆盖硬件、网络、数据全环节；
- 运营复杂度高：万卡集群的大规模运维、新业务场景的动态安全适配，需建立高效的安全运营机制。

为此，中心以“构建可管可控的互联网攻击防御机制、实现合规达标基础上的风险可控、建立高效及时的运营响应体系”为目标，着力打造全栈式安全防护能力。

安全建设方案：体系化防护与动态管控

图18.某算力平台安全能力架构

互联网攻击防护能力

对于高敏感业务采用专线网闸实现物理隔离与逻辑访问控制，结合抗DDoS系统抵御流量型攻击。通过部署互联网防火墙、入侵防御系统（IPS）及APT流量检测设备，形成从流量过滤到异常行为分析的纵深防御机制，有效阻断外部威胁渗透。

安全强合规：兼顾平台与租户业务安全合规

云平台天然满足等保、密评等合规要求，为安全运营提供有力安全能力支撑，云上租户安全能力丰富，如：云防火墙、DDOS基础防护、云主机安全、云WAF等可为租户不同业务提供不同等级的安全防护需求。

实战化安全运营：一体化管控与达标保障

通过搭建安全运营中心，实现多平台日志的汇聚与分析，借助智能算法识别各类安全风险，提升响应效率。聚焦算力安全运营场景，提供丰富安全运营报告。

实施效果：安全与业务协同增长

上线1年内，该算力中心实现：

- 安全层面：有效拦截各类针对算力集群的恶意攻击，成功防范算力劫持与数据越界访问等风险；
- 业务层面：服务租户数量稳步增长，算力售卖规模突破预期，安全合规优势成为业务拓展的重要支撑；
- 行业影响：形成可推广的政府算力中心安全运营模式，输出多项地方安全规范，树立区域安全标杆。

6.2 某广电AIGC平台安全建设案例

项目背景与安全需求

AIGC技术商用化加速，广电行业凭借海量版权音视频数据和庞大用户群占据优势，但算力和模型存短板。某广电客户为借AIGC突破发展，规划新建云智一体专有云平台，基于全国产化软硬件实现自主可控，利旧老设备并迁移全部应用，分四阶段推进智能化战略，其中30多个IPTV核心应用需3个月内零改造且不中断迁移，难度极大。安全需求支撑AIGC应用，应对算力密集型场景安全挑战；保障IPTV应用零改造、不中断迁移及后续运营安全；满足自主可控要求，应对多场景安全威胁；配备成熟云安全管理平台及系列产品，覆盖多场景安全；符合关键信息基础设施安全保护要求，满足等保、密评强合规标准。

安全建设方案

图19.融媒体生产云安全体系框架

云安全资源池：多元能力聚合与服务化输出

基于 “一个中心、三重防护” 理念，构建集防护、检测、响应于一体的安全资源池，深度整合防火墙、态势感知、日志审计、漏洞扫描等能力，实现：

- 能力耦合：与云平台原生架构深度融合，支持虚拟机、物理资源、网络资源的安全策略联动；
- 生态兼容：开放接口对接第三方安全厂商，快速引入新场景防护能力；
- 服务化交付：为租户提供“等保合规套餐”“多租户隔离防护”等自服务化能力，适配多业务等场景需求。

分层防护体系：覆盖全环节安全需求

- **计算环境安全：主机与数据防护**
  - 主机安全：通过资源池部署主机安全防护软件，覆盖Windows/Linux系统，实现病毒查杀、漏洞管理、入侵检测（如监测系统层 / 应用层攻击、漏洞信息）；
  - 数据安全：依托数据库审计、国密加密能力，对用户数据、大模型训练数据实施分级加密存储、全流程审计（覆盖云内/ 云外数据库访问行为，实时监控风险）。

- **区域边界安全：南北 / 东西向精细管控**
  - 南北向防火墙：部署于互联网出口，融合直播、信源边界防护，支持应用识别、入侵防御、病毒拦截，实现租户接入链路“加密认证 + 行为管控”；
  - 东西向防火墙：划分 VPC 安全域，对集群内部流量实施微隔离，防范多租户数据越界、算力盗用，保障万卡集群内资源“安全共享”。
- **安全管理中心：主动运营闭环构建**

  搭建云安全管理平台 + 态势感知平台，实现“风险预测 - 防御控制 - 检测分析 - 响应处置”闭环：
  - 云安全管理平台：统一管控安全资源池策略，汇聚多租户日志、漏洞、资产数据，通过关联分析、威胁情报联动，智能识别集群级风险（如大规模漏洞利用、租户越权访问）；
  - 态势感知平台：解析主机/网络流量，检测威胁、攻击等异常行为，与资源池防护能力实时联动（如发现攻击行为，自动触发防火墙策略拦截、主机隔离）。

价值体现：安全与业务协同升级

- 快速合规：通过云安全资源池“套餐化交付”，平台快速完成等保密评核心能力落地，适配政府主导项目的合规刚需；
- 场景适配：云平台兼容多方安全能力,快速构建异构化的安全能力提供多租户隔离、大模型内容合规等能力，支撑智慧城市、科研教育场景安全复用算力资源，加速业务创新；
- 运营提效：安全管理中心实现“风险自动识别 - 响应”闭环，保障新业务场景“动态安全适配”。

6.3 某头部移动设备厂商大模型内容安全建设案例

项目背景与安全需求

大模型能力快速发展，终端算力加速渗透，大模型技术推动智能终端迈向新高度。某头部移动设备厂商计划在国内推出首款AIPC和pad，通过大模型技术辅助办公场景的文件查找、翻译、总结等能力，主打纯离线运行、高隐私保护等产品主张，端侧的大模型相比云端大模型场景面临更严重的内容合规风险，也将面临更严苛的备案审核要求：

- 终端设备算力资源有限，仅能为安全模块提供少量GPU甚至纯CPU的审核资源。
- 对于安全模块的初始化时间、审核时间、内存占用等性能要求较高。
- 端侧大模型安全审查更为严苛，且要求端侧的用户违规数据必须存储与上报。

安全建设方案

大模型安全护栏lite版本是一套专门面向低算力的终端大模型安全解决方案。为各类终端场景加载离线大模型服务时，提供算力消耗低、安全能力无损的终端大模型内容安全方案，解决离线场景下的内容安全审核与干预问题，同时支持违规用户封禁和违规数据存储，为各类终端大模型保驾护航。

图20.大模型内容安全体系框架

方案优势

超低算力需求：可运行在纯CPU环境，最低180Mhz主频CPU，不超过1024MB内存支持。

全离线运行：生成内容安全检查时无需联网，超低响应时间满足终端设备的极速体验需求。

多平台支持：全线支持X86，ARM架构，原生支持Linux，Android平台，国产化适配。

可回复干预：构建超灵活的回复干预手段及干预平台，结合OTA通道实现轻量级快捷处理风险。

违规处置：支持上报违规日志，且可以灵活配置违规用户封禁策略。

实施效果

- 助力客户成为国内率先备案的端侧大模型内容安全产品，抢占客户市场，凸显企业技术实力。
- 客户云端、PC、pad产品均使用百度大模型内容安全能力，风控标准一致，便于风险数据统一管理。
- 端侧预置了违规用户封禁能力，强化风险管控，满足监管要求。
- 提供线上干预平台，支持灵活调整策略阈值、封禁阈值、知识库与词表数据，并生成端侧打包文件，快速干预处置策略。

07

总结与未来展望

7.1 总结

AI基础设施平台作为数字经济时代的核心设施，其安全防护体系的构建需直面“算力集中化、业务云化、模型智能化”带来的复合型挑战。本文档围绕百度AI基础设施的安全解决方案，系统阐述了大规模算力中心安全建设的完整框架，核心可概括为以下三点：

全维度安全体系已形成闭环

从底层基础设施到上层模型应用，百度算力中心构建了“边界-平台-租户-密码-模型-运营”六层联动的安全防护体系：通过边界安全筑牢网络入口防线，平台基础安全夯实底层合规底座，租户安全适配多场景需求，密码合规体系实现“一次适配、全栈合规”，模型安全覆盖大模型全生命周期，安全运营服务则形成“监测-分析-处置-优化”的持续保障闭环。这一体系既满足《网络安全法》《数据安全法》《等保2.0》《密评》等刚性合规要求，又精准应对数据泄露、算力劫持、模型投毒等新型威胁。

技术与管理深度融合是核心支撑

在技术层面，依托云原生安全、AI威胁检测、国密算法等前沿技术，实现了从物理环境到虚拟资源、从数据传输到模型推理的全链路防护；在管理层面，通过安全组织架构、制度流程、人员培训等体系化设计，确保技术能力有效落地。尤其在GPU算力平台、大模型应用等核心场景，通过“通用防护+专项强化”的差异化策略，平衡了安全与业务效率。

安全运营驱动持续进化

以安全运营中心（SOC）为枢纽，通过资产管理、告警值守、漏洞管理等模块，实现了安全风险的“可视、可管、可控”。结合实战化演练与威胁情报联动，保障了安全体系对动态威胁的快速响应，为算力中心稳定运行提供了“合规+防护+运营”三位一体的坚实支撑。

7.2 未来展望

随着AI技术的深度渗透、量子计算的广泛应用以及产业数字化的加速，算力中心安全将迎来新的演进方向：

智能化防御成为核心能力

大语言模型将深度赋能安全运营，通过自然语言处理解析威胁情报、生成自动化处置方案，结合UEBA技术提升未知威胁识别能力。AI Security Agent智能安全运营助手将实现自主感知环境、分析风险、决策响应（如隔离威胁、更新策略），并通过学习持续优化，替代人工处理重复任务（如邮件钓鱼分类、告警优先级排序），释放安全人员精力聚焦高阶威胁狩猎与战略规划。同时，AI模型自身安全防护将更精细化，从对抗样本检测到伦理审查，形成全生命周期管控。

量子抗性与密码技术升级

面对量子计算对传统加密的冲击，需提前布局抗量子密码技术，推动国密算法与抗量子算法融合，构建“量子时代”密码合规体系。密钥管理向分布式、零信任架构演进，保障多云与边缘场景下的密钥安全。

生态协同与合规深度融合

安全将从“单点防护”转向“生态共治”，通过威胁情报共享与跨厂商联动构建安全生态。合规要求将深度嵌入业务流程，从“被动达标”转向“主动合规”，结合隐私计算、联邦学习等技术实现“数据可用不可见”。

百度云安全的持续突破

百度将依托AI、云原生、密码技术积累，深化安全产品与智算场景适配，推出轻量化、弹性化安全服务；通过开源社区与行业标准共建，推动安全技术普惠化，助力客户在合规基础上释放智能算力价值。

未来，算力中心安全将不仅是“风险防御的盾牌”，更成为“业务创新的基石”。百度智能云将持续以技术创新响应时代需求，为算力中心的安全、合规、高效运行保驾护航，推动数字经济在安全可控的环境中蓬勃发展。

缩略词	说明
AIGC	AIGC 是“Artificial Intelligence Generated Content”的缩写，意为“人工智能生成内容”，是一种利用人工智能技术来生成内容的方式，涉及多个技术领域，如自然语言处理、机器学习、深度学习等。
API	**Application Programming Interface，应用程序编程接口：**是不同软件应用之间实现交互的一套规则、协议和工具，规定了数据请求、响应的格式和方式，让不同系统能高效协同工作。
CASB	部署在用户与云服务之间的安全工具，通过管控访问权限、检测数据泄露、确保合规性等，保护企业在使用云服务时的数据安全。
CSPM	通过持续监控云环境的配置、资源使用和合规状态，识别安全漏洞、不合规行为等风险，帮助企业维护云安全态势。
CNAPP	**Cloud-Native Application Protection Platform，云原生应用保护平台**整合多种云安全工具，覆盖云原生应用从开发、部署到运行的全生命周期，提供漏洞检测、runtime 防护等一体化安全保障。
WAF	专门针对 Web 应用的安全防护工具，可识别并拦截 SQL注入、XSS 等常见的Web 攻击，保护Web 应用和服务器安全。
IDS	通过监测网络流量或系统日志，分析识别未授权的访问、攻击等异常行为，并发出告警，帮助管理员及时响应安全威胁。
CIS	一个非营利组织，发布全球公认的安全基准和最佳实践，指导企业和组织配置系统、网络及应用的安全设置。
DDOS	攻击者控制大量分布式设备向目标发送海量请求，耗尽目标的带宽、服务器资源等，使其无法正常响应合法用户的请求。
SQL	一种用于管理关系型数据库的标准化语言，可实现数据的查询、插入、更新、删除等操作，是数据库交互的核心工具。
XSS	攻击者在网页中注入恶意脚本，当用户访问该网页时，脚本被执行，可能导致用户数据泄露、会话劫持等安全问题。
SSL	一种用于在网络通信中建立加密连接的安全协议，通过对传输数据进行加密，确保数据在客户端和服务器之间传输的机密性和完整性（现已逐步被 TLS 替代，但常被统称为 SSL/TLS）。
KMS	负责加密密钥全生命周期管理的系统，包括密钥的生成、存储、分发、轮换、销毁等，确保加密密钥的安全性和可用性。
APT	由有组织的攻击者发起的长期、定向攻击，具有隐蔽性强、目标明确、持续时间长等特点，通常针对特定机构或企业窃取敏感信息。
MFA	一种身份验证方法，要求用户提供两种或两种以上的验证因素（如密码、手机验证码、指纹等），以增强身份认证的安全性，降低账号被盗风险。
VXLAN	一种网络虚拟化技术，通过在现有 IP 网络上封装二层数据帧，构建跨物理网络的大二层虚拟网络，支持多租户隔离和大规模虚拟机迁移。


百度智能云


百度安全