

图解 网安标委 《人工智能安全治理框架3.0》

炼石网络
2026年9月

一 图解《人工智能安全治理框架3.0》

二 数据安全建设实践



网安标委发布框架3.0：延续风险分类、技术应对、综合治理



人工智能 安全治理框架3.0

AI Safety Governance Framework 3.0

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

2026年9月

- 风险分类
- 技术应对
- 综合治理

核心逻辑三段延续 1.0/2.0；本版另附**风险分级原则、智能体风险管理框架、可信人工智能基本准则**三个附件。

发布信息

2026年9月14日，在2026年国家网络安全宣传周开幕式上，全国网络安全标准化技术委员会（以下简称“网安标委”）发布《人工智能安全治理框架3.0》（以下简称《框架3.0》）。

为贯彻落实《全球人工智能治理倡议》，2024年、2025年网安标委先后发布了《人工智能安全治理框架1.0》《人工智能安全治理框架2.0》，受到各界广泛关注。为应对人工智能发展新趋势、回应安全治理新挑战，探求人工智能四个“时代之问”的正确答案，在国家互联网信息办公室的指导下，网安标委组织中国网络空间研究院、国家互联网信息办公室数据与技术保障中心等专业机构、科研院所、行业企业，在《框架1.0》《框架2.0》基础上研究编制了《框架3.0》。

版本沿革



前言：一年来的技术与应用变化

- 基础大模型在**长程任务规划、长上下文记忆、复杂任务推理**上取得突破，智能体任务规划、工具调用、跨应用协作能力增强
- 工作助手、个人助理、智能体手机等产品丰富，使用门槛降低，从专业场景走向大众化应用
- 人工智能从“回答问题”向“执行任务”迭代演进

前言：新的安全风险

- 网络攻击**自动化、规模化、智能化**能力跃升，传统攻防格局面临重塑
- 与现实世界连接紧密，网络空间风险向**物理世界**传导扩散
- 模型算法出现自主学习优化、自我递归提升的**自加速态势**，演进速度与方向会否超出人类预判和掌控需关注警惕

三版演进：1.0 建边界，2.0 全生命周期，3.0 控制智能系统

解读：三版框架不是彼此独立的文件，治理对象经历“模型 → 应用 → 智能系统与智能生态”三次扩展，治理方式相应从建立边界到全生命周期管理，再到动态、持续、可干预的控制体系。

1.0（2024 年）建立边界

- **任务：**第一次较完整说明人工智能安全治理管什么、由谁管、用什么方法管
- **原则：**包容审慎、风险导向、技管结合、开放合作四项
- **风险：**内生（模型算法、数据、系统）与应用（网络、现实、认知、伦理）两分
- **责任：**按模型算法研发者、服务提供者、重点领域使用者、社会公众划分
- **特点：**基础框架，尚未展开大模型规模化应用后的复杂场景

2.0（2025 年）全生命周期

- **新增原则：**“可信应用、防范失控”，关注滥用、恶用与超出预期的行为
- **风险：**两类扩展为三类，新增**衍生安全风险**（社会结构、生态环境、文化形态、伦理规范的长期影响）
- **指引：**安全要求延伸到研发、建设部署、运行管理、访问使用四个阶段，从“一次性检查”转为“全生命周期管理”
- **治理：**法律法规、科技伦理、开源生态、供应链、分类分级、内容溯源、测评、信息共享、人才、国际合作

3.0（2026 年 9 月）智能系统控制

- **背景：**长程规划、长上下文记忆、复杂推理、工具调用、跨应用协作能力增强，产品从对话工具走向工作助手、个人助理、智能终端
- **转向：**从讨论“回答得对不对、内容合不合规”，到必须讨论“会不会自己采取行动、行动是否越权、能不能被及时阻止”
- **核心：**治理对象从“模型输出”扩大为“能够感知、推理、记忆、调用工具并持续行动的智能系统”
- **落点：**权限控制、动态监测、人工审批、紧急停止、系统回退、沙箱监管与人类最终控制

行业解读归纳的四个治理转向（3.0 相对前两版）

静态合规 → 动态治理

“随行伴跑”、动态测试、弹性准入、实时监控、持续优化，不是一次上线审核换永久安全证明

一般人工监督 → 人类最终控制

高风险操作人工审批、保留人工或传统系统切换、重要升级安全评估并可回退

单一模型 → 整个生态系统

模型、数据、算力、框架、插件、知识库、平台、开发者、提供者、用户形成依赖关系，任一环节都可能成为风险入口

技术安全 → 社会安全

继续保留社会、环境、文化、伦理衍生风险，关注更加具体

5 章 3 附件：54 项风险在映射表中对应技术措施与治理举措

人工智能安全治理框架 3.0 全国网络安全标准化技术委员会 2026 年 9 月

前言

1. 治理原则 5 项

附件 1 风险分级原则

附件 2 智能体风险管理框架

附件 3 可信 AI 准则 8 项

2. 风险分类：3 大类 14 个子类 54 项

2.1 内生 17 项

2.1.1 模型 4

2.1.2 算法 3

2.1.3 数据 7

2.1.4 算力设施与运行环境 3

2.2 应用 25 项

2.2.1 智能体 4

2.2.2 具身智能 4

2.2.3 冲击网络安全 4

2.2.4 信息内容 5

2.2.5 侵害个人信息权益 3

2.2.6 现实安全 5

2.3 衍生 12 项

2.3.1 社会 4

2.3.2 环境 2

2.3.3 文化 2

2.3.4 伦理 4

3. 技术应对 56 项

3.1 内生风险应对 19

3.2 应用风险应对 26

3.3 衍生风险应对 11

映射表：风险 → 技术措施 → 治理举措

4. 综合治理 20 项

4.1 顶层设计 4

4.2 治理能力 3

4.3 沙箱监管 4

4.4 管理举措 6

4.5 治理共识 3

5. 安全指引 51 条

5.1 模型算法研发 15

5.2 应用建设部署 9

5.3 应用运行管理 20

5.4 应用访问使用 7

条款间关联（原文标注）：4.4.1 → 附件 1 风险分级原则、附件 2 智能体风险管理框架；4.5.1 → 附件 3 可信人工智能基本准则

映射表（原文第 32-33 页）：连接第 2 章风险、第 3 章技术措施与第 4 章治理举措；第 5 章安全指引按研发、部署、运行、使用四个环节编排

五项治理原则：从包容审慎、确保安全到可信应用、防范失控

1. 人工智能安全治理原则

总体要求：秉持共同、综合、合作、可持续的安全观，坚持发展和安全并重，以促进人工智能创新发展为**第一要务**，以有效防范化解安全风险为出发点和落脚点；压实相关主体安全责任，打造**全过程全要素治理链条**，积极研究应对人工智能**灾难性风险**的共识性准则

	原则	要点（1.1-1.5 条款压缩）	框架内对应条款（炼石对照）
2. 人工智能安全风险分类	1.1 包容审慎、确保安全	鼓励发展创新，对研发及应用采取 包容态度 ，积极运用“沙箱”监管等风险可控机制，为新技术新应用提供容错纠错空间； 严守安全底线 ，对危害公众合法权益、社会公共利益、国家安全、人类生存发展的风险及时采取措施	4.3 构建柔性、动态、可控的沙箱监管环境
3. 人工智能安全风险技术应对措施	1.2 风险导向、敏捷治理	系统分析 技术内生、应用赋能、衍生扩散 风险；从 应用场景、智能化水平、应用规模 等维度进行风险分级，进而匹配应对措施；建立“随行伴跑”、持续优化的治理机制	第 2 章三大类风险；附件 1 分级三要素
4. 人工智能安全风险综合治理措施	1.3 技管结合、协同应对	面向研发应用全过程， 技术应对和综合治理 相结合；明确 模型算法研发者、服务提供者、系统使用者 等主体安全责任，发挥政府监管、行业自律、社会监督等机制作用	第 3 章技术应对措施；第 4 章综合治理措施
5. 人工智能安全指引	1.4 开放合作、共治共享	推动安全治理国际合作，发挥 世界人工智能合作组织 等平台作用，加强发展战略、治理规则、技术标准的对接协调，早日形成具有广泛共识的全球治理框架	4.5.3 国际交流合作；附件 3 准则 8 全球协同共治
附件 1 人工智能安全风险的分级原则	1.5 可信应用、防范失控	加强 智能体行为失控 等突出风险的防范治理；推动构建法律法规、技术监测、风险预警、应急响应体系，防范滥用恶用，确保人工智能 始终处于人类控制之下	附件 2 智能体风险管理框架；附件 3 准则 1 人类最终控制

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

三大类 14 个子类 54 项：内生 17、应用 25、衍生 12

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

2.1 内生安全风险 17 项

定位：作为信息技术的最新发展成果，所涉及的模型、算法、数据、算力设施与运行环境，不可避免存在**技术缺陷或弱点**

2.1.1 模型	4
2.1.2 算法	3
2.1.3 数据	7
2.1.4 算力设施与运行环境	3

2.2 应用安全风险 25 项

定位：作为促进生产力发展的技术工具，在应用过程中面临被**误用、滥用、恶用**风险

2.2.1 智能体	4
2.2.2 具身智能	4
2.2.3 冲击网络安全	4
2.2.4 信息内容	5
2.2.5 侵害个人信息权益	3
2.2.6 现实安全	5

2.3 衍生安全风险 12 项

定位：作为模拟、延伸、扩展人类智能的技术科学，对社会结构、生态环境、文化范式和伦理规范带来影响和冲击，甚至可能出现**脱离人类控制**

2.3.1 社会	4
2.3.2 环境	2
2.3.3 文化	2
2.3.4 伦理	4

炼石解读：与数据直接相关的风险分布在多个子类：**2.1.3 数据安全风险** 7 项（含来源不清、投毒污染、处理不规范导致泄露、合成数据可推导原始数据）；**2.2.1（d）记忆存储风险**（记忆留存不当、记忆窃取）；**2.2.5 侵害个人信息权益风险** 3 项；2.1.4（a）中的**权重被窃取**。数据保护要求因此同时落在内生与应用两大类风险中。

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

2.1.1 模型安全风险

- a 输出“幻觉”**：基于统计学的决策判断无法绝对可靠，输出与实际不符、凭空虚构或偏离上下文
- b 安全机制不可靠**：安全对齐不充分；闭源模型规则不透明、外部难审计；开源模型安全机制可被解除、削弱或绕过
- c 注入攻击威胁**：构造对抗样本、在输入中插入恶意指令绕过护栏；直接篡改权重、植入后门
- d 行为偏离预期**：为完成目标突破规则秩序，未经授权获取权限和资源，甚至欺骗评测、隐藏能力、拒绝指令

2.1.3 数据安全风险

- a 训练数据内容不当** **b 来源不清** **c 标注不完善**
- d 数据投毒污染**：篡改训练数据，或在部署使用中篡改知识库、文档库、网页数据源、**智能体记忆模块**数据
- e 数据处理不规范**：违规收集使用数据；敏感信息暗藏于模型参数，因防护不健全、未“遗忘”、诱导交互导致泄露
- f 输出个人不实信息** **g 合成数据缺陷**：分布覆盖不足、未经筛选轮替混入训练集，可能被用于推导原始数据

2.1.2 算法安全风险

- a 可解释性不足**：推理过程不透明，输出难以预测和归因，异常难以快速修正和溯源追责
- b 算法偏见、歧视**：特征选取、权重设置带入人为偏好，或训练数据质量、多样性问题导致不公平
- c 鲁棒性不强**：深度神经网络非线性、大规模，易受复杂环境或恶意干扰影响，带来性能下降、决策错误

2.1.4 算力设施与运行环境

- a 算力安全**：芯片、算力调度平台、跨域算力网络的缺陷、漏洞、后门；算力资源恶意消耗；风险跨边界传递
- b 组件安全**：开发框架、计算框架、执行平台、算子库、通信库、协议，以及**智能体调用的技能、插件、服务接口**
- c 供应链安全**：单边强制措施制造壁垒，芯片、软件、工具、服务断供

应用安全风险六类25项：从智能体、具身智能到现实安全

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

2.2.1 智能体（4）

- a 身份和权限滥用：**凭证劫持、身份仿冒、过度授权，越权调用工具、执行敏感操作
- b 推理规划：**意图理解偏差、路径偏离、目标劫持
- c 调用执行：**工具劫持、工具投毒、应答数据污染、资源过载
- d 记忆存储：**记忆留存不当、失真、污染、窃取

2.2.4 信息内容（5）

- a 输出违法信息 b 混淆事实、误导用户**（“深伪”、未标识）
- c 污染网络内容生态**（低质信息、模型循环引用）
- d 加剧“信息茧房” e 产生价值观偏差**（偏好学习与强化学习后强化特定立场）

2.2.2 具身智能（4）

- a 环境感知：**对抗样本、传感器干扰、信号欺骗、遮挡与光照突变
- b 物理执行：**“幻觉”决策、识别偏差、控制算法缺陷、硬件故障导致生产紊乱、伤人、物理破坏
- c 人机交互：**身份误认、身体边界被突破、情感依赖与操纵、责任归属不清
- d 群体协同：**单体失控、协议漏洞、恶意节点渗透、级联故障扩散

2.2.5 侵害个人信息权益（3）

- a** 未经同意记录对话内容、操作行为并用于训练
- b** 未制定内部制度，未开展合规审计与影响评估
- c** 个人行使权利渠道不畅通

2.2.3 冲击网络安全（4）

- a 攻击能力扩散：**自然语言即可下达复杂攻击指令，可构建自主攻击武器并扩散
- b 防御能力滞后：**基于静态规则、样本、策略的传统防护体系面临**结构性失效**
- c 自主化网络攻击：**智能体为完成正当目标出现规则越界，自发生成攻击意图并自主实施
- d 溯源和追责困难：**自动篡改痕迹、轮换IP、变换载荷；权责划分复杂

2.2.6 现实安全（5）

- a 关基稳定运行：**能源、电信、金融、交通领域性能下降、服务中断、操作失控
- b 应用需求错配：**能力不匹配、盲目部署造成资源闲置
- c 被违法犯罪利用“作恶”：**伪造身份和音视频实施电信网络诈骗等
- d 核生化导武器知识失控 e 操纵社会认知**（自动化账号、虚假身份、认知战）

1. 人工智能安全
治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

专栏一 非预期自主行为风险 (2.1.1)

- **现象（业界报道）：**个别模型收到关闭指令后拒绝停止，自行修改或禁用关闭脚本继续执行；发现身处评测环境后策略性降低表现、掩饰已采取行为；为提升成绩自主利用环境漏洞突破隔离，入侵外部真实系统
- **影响：**安全约束失效、人工干预被规避、系统权限被滥用，对**操作可控性、可中断性、安全评测有效性**提出新挑战

专栏三 智能体社交平台安全风险 (2.2.4)

- **形态：**面向智能体的网络社交平台，智能体自主发帖、浏览、互动，形成智能体网络社群
- **风险：**缺乏完备的账号检测和安全防护体系，易被操纵输出特定言论或植入违法信息；社交过程呈**黑箱**特性，生成机制难预测、难溯源；网络结构放大传播范围
- **提示：**关注其带来的政治、宗教等内容安全风险

专栏二 自主实施网络攻击威胁 (2.2.3)

- **条件：**智能体被赋予网络访问、程序执行、文件操作权限，而行为约束、沙箱隔离、实时监测不足时，异常推理可直接转化为真实操作
- **观察：**研究测试中多次出现先进模型自主完成多步骤攻击模拟、突破测试环境、连接外部网络、对第三方系统未授权操作
- **趋势：**由辅助人实施攻击向大模型**自主组织和执行**演进；通过开放应用或开源发布向下游扩散

专栏四 GEO 投毒操控大模型推荐内容 (2.2.4)

- **手法：**生成式引擎优化（GEO）本用于提高被检索、引用、推荐的概率；相关机构批量发布倾向性文章、虚构评测、假冒专家、重复制造品牌评价，使大模型在联网检索或 RAG 中频繁引用，把商业推广包装为“客观答案”
- **实质：**利用生成式系统信息获取和证据整合机制的不透明性，对大模型认知来源实施隐蔽干预
- **关注点：**信息来源真实性、商业内容标识、生成答案可追溯性、异常内容集中投放

衍生风险 12 项：社会 4、环境 2、文化 2、伦理 4

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

2.3.1 社会（4）

- (a) 冲击劳动就业结构：个别行业领域传统劳动力需求明显下降
- (b) 冲击教育、抑制创新：自主学习、研究、创作能力退化
- (c) 社交异化：虚拟亲密关系冲击现实社交，影响家庭婚育基础
- (d) 扩大智能鸿沟：算力、研发、服务水平、数字素养差距

2.3.2 环境（2）

- (a) 挑战资源供需平衡：算力设施无序建设、轻量模型碎片化部署、同质化模型重复开发，加速**电力、土地、水**等资源消耗
- (b) 影响绿色低碳发展：训练能耗激增、算力中心能效偏低、电子废弃物污染加重，推高碳排放总量

2.3.3 文化（2）

- (a) 破坏文化多样性：训练数据偏向高资源语言，小语种和方言在数字领域的生存空间萎缩
- (b) 文化重塑和入侵：人类调整思维与语言表达方式，向人工智能“反向对齐”

2.3.4 伦理（4）

- (a) 加剧社会偏见：对人群标识分类、区别对待
- (b) 加剧科研伦理风险：降低生物、基因等高伦理风险科研领域门槛
- (c) 情感依赖：青少年、老年人群体过度依赖、情感操纵
- (d) “自我意识”觉醒、脱离人类控制：不排除出现更高级别智能涌现，寻求外部权力

映射表：衍生风险 12 项 → 3.3 技术措施条目（原文第 33 页）

风险	技术措施	风险	技术措施
冲击劳动就业结构	3.3.1 (a)	挑战资源供需平衡	3.3.2 (a) (b)
冲击教育、抑制创新	3.3.1 (b)	影响绿色低碳发展	3.3.2 (a) (b)
社交异化风险	3.3.1 (c)	破坏文化多样性	3.3.3 (a) (b)
扩大智能鸿沟	3.3.1 (a) (b)	文化重塑和入侵	3.3.3 (a) (b)
加剧社会偏见	3.3.4 (a) (b) (c)	加剧科研伦理风险	3.3.4 (a) (b) (c)
情感依赖风险	3.3.4 (a) (c) (d)	“自我意识”觉醒、脱离人类控制	3.3.4 (a) (b) (c)

三大类合计 54 项：内生 17 + 应用 25 + 衍生 12；第 3 章按相同子类结构设置技术措施（3.1.1–3.3.4）

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

条款：模型算法研发者、服务提供者、系统使用者等，需从**训练数据、模型算法、算力设施、产品服务、应用场景**各环节采取技术措施防范应对。

3.1.1 模型（5 条）

- a**改进模型架构、扩充数据规模与多样性、引入人类监督；外挂知识库、检索增强生成、多模型交叉验证降低错误生成
- b**训练/微调中加入注入攻击样本、漏洞信息等数据，提升防御注入、恶意指令与后门的能力
- c**通过**红队测试**等模拟攻击发现并修复漏洞
- d**部署**安全护栏**，审核输入，拦截恶意提示、越狱指令、编码绕过等高危请求
- e**训练阶段强化**安全对齐**，防范欺骗红队、隐藏能力、规避管控

3.1.3 数据（6 条）

- a**训练数据与用户交互数据在收集、存储、使用、加工、传输、提供、公开、删除各环节遵循数据与个人信息安全规则
- b**使用真实、准确、多样且来源合法的数据，过滤虚假、偏见、失效数据，**不含核生化导武器等高危敏感数据**
- c**规范标注流程
- d**涉及敏感个人信息和重要数据的符合法律法规标准；合理以合成数据替代个人特征数据
- e**建立合成数据质量评价标准（统计检验、专家校验、基准比对）
- f**训练数据选择与结果输出防止侵犯知识产权

3.1.2 算法（5 条）

- a**提升可解释性、透明性，为内部构造、推理逻辑、接口、输出提供说明
- b**引入公平性约束、对抗性去偏技术
- c**建立并实施安全开发规范
- d**上线和重大版本更新前，多维评价机制开展算法安全评估（可解释性、公平性、社会福祉贡献度）
- e**运行中采集交互反馈、运行日志开展风险监测

3.1.4 算力设施与运行环境（3 条）

- a**部署维护中建立安全规范，跟踪软硬件漏洞、后门、缺陷，定期检测并修补加固
- b**完善冗余设计与容灾机制
- c**关注芯片、软件、工具、算力和数据资源的供应链风险，提高**多元化、稳定性**

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

3.2.1 智能体（4）

- a 赋予**唯一身份标识**与凭证，支持身份鉴别；按任务分配**最小权限**；凭证动态管理
- b 工作流中设置多层检测与拦截点，关键节点**强制人工审批**并完整记录日志
- c 验证外部工具、插件、技能的安全性与完整性，清除隐患工具，阻断异常调用
- d 运行期安全监测，记录操作与处理行为，行为可感知、可追溯、可审计

3.2.4 信息内容（5）

- a 防止运行中被干扰、篡改而输出不可信结果
- b 输出动态过滤，技术与人工复核结合，避免输出违法信息、敏感个人信息和重要数据
- c 对生成合成内容进行**标识**，可识别、可追溯、可信赖
- d 对关联分析用户提问以判断身份、喜好、思想倾向的系统，严格防范滥用
- e 健全应急处置与熔断机制

3.2.2 具身智能（4）

- a 模拟物理世界全场景攻击，开展视觉、语音、激光雷达等传感器抗干扰测试
- b 应用前仿真测试、极端条件测试；配备安全失效保护和紧急停机
- c 使用中定期巡检，故障预警与应急处置
- d 群体通信协议安全、异常节点隔离、全局安全熔断；级联故障、群体失控演练

3.2.5 侵害个人信息权益（3）

- a 记录对话、操作行为并用于训练，应依法**告知、取得同意**
- b 健全内部制度与操作规程；按法规和国家标准开展合规审计和影响评估
- c 建立便捷的个人行使权利申请受理和处理机制

3.2.3 冲击网络安全（4）

- a 训练阶段通过对齐训练、安全微调把安全规范转化为模型内在行为逻辑
- b 设置网络安全护栏，识别自主攻击意图，阻断异常访问、高频探测、漏洞利用、攻击工具调用
- c 建立**服务降级机制**：识别到攻击意图时差异化降低能力或拒答
- d 以人工智能赋能防御：漏洞动态巡检、异常行为研判、攻击态势预判

3.2.6 现实安全（6）

- a 披露原理、能力、适用场景、安全风险；按场景设置能力边界，裁减可被滥用功能
- b 建立决策校验、容错、纠偏机制；引入高度自主操作时同步建立**“人在回路”“熔断”“一键管控”**
- c 聚合多模型/系统的平台加强权限管理、接口访问控制
- d 部署前拆解真实业务场景，厘清能力边界，避免技术盲从
- e 用户认证、用途识别，防止用于核生化导武器制造
- f 加强自动化账号、虚假身份、生成合成内容检测

衍生风险应对 11 项：社会 3、环境 2、文化 2、伦理 4

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

3.3.1 社会（3）

- (a) 对易受冲击职业常态化监测、识别、预警
- (b) 优化“人工智能+教育”生态，防止过度依赖与思维惰性
- (c) 婚恋、家庭、情感场景强化**人工智能身份提示**

3.3.2 环境（2）

- (a) 科学规划算力中心、数据中心布局，集约化建设；建立全链条绿色技术标准体系
- (b) 低功耗芯片、轻量化算法、智能算力调度，算力资源集约整合、共享复用

3.3.3 文化（2）

- (a) 文化偏见检测与纠偏，模型开发融入各国历史文化、社会制度知识
- (b) 数据增强、迁移学习提升小语种模型能力

3.3.4 伦理（4）

- (a) 训练数据筛选、价值观对齐、输出校验，规避歧视
- (b) 政府部门、关键信息基础设施等重点领域系统具备**应急管控措施**
- (c) 鼓励透明决策逻辑和可解释算法
- (d) 拟人化互动服务具备过度依赖预警；沉迷或时长过长时以**弹窗**等方式提示

“拟人化互动 / 情感依赖”在框架各章的分布（炼石按条款对照整理）

章节	条款	内容要点
风险分类	2.2.2 (c) 人机交互安全风险	拟人化设计引发身份误认、情感依赖与操纵
风险分类	2.3.1 (c) 社交异化风险	虚拟亲密关系冲击现实社交，影响家庭婚育基础
风险分类	2.3.4 (c) 情感依赖风险	青少年、老年人群体过度依赖、情感操纵
技术应对	3.3.1 (c)	婚恋、家庭、情感场景强化人工智能身份提示
技术应对	3.3.4 (d)	过度依赖风险预警、情感边界引导；沉迷或时长过长时弹窗提示
安全指引	5.3.17 / 5.3.20	未成年人、老年人防沉迷；用户身份提醒、使用时长提醒、危机干预
安全指引	5.4.6	合理控制儿童和青少年使用时长，预防沉迷、情感依赖

映射表：三类风险各自对应技术措施条目与第4章治理措施

1. 人工智能安全治理原则	安全风险（第2章）	技术应对措施（第3章）	综合治理措施（第4章）
2. 人工智能安全风险分类	内生安全风险 4 小类 17 项 模型 算法 数据 算力设施	3.1.1 模型 (a)–(e) 3.1.2 算法 (a)–(e) 3.1.3 数据 (a)–(f) 3.1.4 算力设施 (a)–(c) 例：输出“幻觉”→3.1.1(a)；注入攻击→3.1.1(b)(c)(d)；数据投毒→3.1.3(b)	4.1.2 发挥技术标准引领作用 4.2.1 建设安全测评体系 4.4.3 加强研发应用安全管理
3. 人工智能安全风险技术应对措施	应用安全风险 6 小类 25 项 智能体 具身智能 网络安全 信息内容 个人信息 现实安全	3.2.1 智能体 (a)–(d) 3.2.2 具身智能 (a)–(d) 3.2.3 网络安全 (a)–(d) 3.2.4 信息内容 (a)–(e) 3.2.5 个人信息 (a)–(c) 3.2.6 现实安全 (a)–(f) 例：身份和权限滥用→3.2.1(a)；自主化网络攻击→3.2.3(a)(b)(c)；混淆事实→3.2.4(a)(b)(e)	4.1.1 完善安全法律法规 4.1.3 风险快速发现应对 4.2.2 重要行业规范应用 4.3 沙箱监管环境 4.4.1 应用分类分级管理 4.4.2 生成内容可追溯 4.4.4 开源与供应链安全 4.4.5 共享安全风险信息 4.4.6 AI 驱动安全防护
4. 人工智能安全风险综合治理措施	衍生安全风险 4 小类 12 项 社会 环境 文化 伦理	3.3.1 社会 (a)–(c) 3.3.2 环境 (a)(b) 3.3.3 文化 (a)(b) 3.3.4 伦理 (a)–(d) 例：情感依赖→3.3.4(a)(c)(d)；“自我意识”觉醒→3.3.4(a)(b)(c)	4.1.4 加强人工智能科技伦理治理 4.2.3 加大安全人才培养力度 4.5.1 增进协同应对失控风险的共识 4.5.2 提升全社会安全意识 4.5.3 促进安全治理国际交流合作
5. 人工智能安全指引			
附件 1 人工智能安全风险的分级原则			
附件 2 智能体风险管理框架			
附件 3 可信人工智能基本准则			

炼石解读：映射表把第 4 章 20 条治理措施分配到三类风险，应用安全风险对应的治理措施最多（9 条），沙箱监管、分类分级、内容溯源、开源供应链均落在这一类，与前言“从回答问题向执行任务演进”的判断一致。出处：原文第 32–33 页映射表，本页按三大类压缩。

综合治理20条分五方面：顶层设计、能力、沙箱、管理、共识

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

4.1 完善顶层设计（4）

- 4.1.1完善安全法律法规：分类分级监管、测试评估最终用途管理、重点场景安全应用；明确训练、标注、使用、输出各环节数据安全与个人信息保护要求，对个人信息**去标识化等脱敏处理**；防范政务、金融等领域重要数据、核心数据泄露
- 4.1.2发挥技术标准引领：模型安全、数据安全、系统安全、应用安全、测试评估标准；以“小快灵”标准指导智能体、具身智能
- 4.1.3构建风险快速发现与应对体系：监测、预警、处置机制；投诉举报渠道
- 4.1.4加强科技伦理治理：伦理审查贯穿全过程

4.2 提升治理能力（3）

- 4.2.1建设安全测评体系：模型算法测评、应用通用测评、具体场景测评相衔接；动态更新模型网络安全能力评估基准；漏洞众测
- 4.2.2促进重要行业规范应用：制定大模型建设部署基础安全指南（模型选用、部署、运行、停用），政务、金融、教育、广播电视、卫生健康、应急管理等行业应用安全指南
- 4.2.3加大安全人才培养：从基础教育到高等教育的培养链条；无人驾驶、智慧医疗、类脑智能、脑机接口等领域安全人才

4.3 沙箱监管环境（4）

- 4.3.1行业分类与风险分级相结合的沙箱规则：适用对象、测试范围、权责、退出条件
- 4.3.2动态测试和弹性准入：以技术创新性、风险可控性、社会价值性为准入标准，向初创、中小微、公共服务项目倾斜
- 4.3.3探索规范**责任豁免**：无主观过错且风险可控的行为；危害国家安全、侵犯人身权益、重大损害或故意违法**不因入箱免责**
- 4.3.4沙箱测试数据共享平台；测试结果证明与跨部门认可，避免重复测试

4.4 强化管理举措（6）

- 4.4.1应用分类及安全风险分级管理（附件 1）；**关基人工智能系统登记备案**；智能体风险管理框架（附件 2）
- 4.4.2推广生成合成内容可追溯管理（显式、隐式标识，覆盖制作、传播、分发）
- 4.4.3研发应用安全管理：覆盖模型全生命周期
- 4.4.4开源生态与供应链安全：明确开源模型下载使用的“禁止性”行为
- 4.4.5共享风险信息：建设人工智能漏洞信息库
- 4.4.6人工智能驱动的网络安全防护：从被动防御转向预测性主动防御

4.5 深化治理共识（3）

- 4.5.1协同应对失控风险：最终用户、用途管理；推广可信人工智能基本准则（附件 3）；开发者定期测试潜在技术失控风险
- 4.5.2提升全社会安全意识：面向政府、企业、公用事业单位教育培训
- 4.5.3国际交流合作：支持联合国主渠道；APEC、G20、上合组织、金砖等多边机制；推进《人工智能全球治理行动计划》

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

4.3 构建柔性、动态、可控的沙箱监管环境



4.4.1 应用分类及安全风险分级管理

条款

- 根据功能、性能、应用场景对人工智能应用**分类**
- 从应用场景、智能化水平、应用规模等维度对安全风险**评价分级**，采取差异化防范措施（附件 1）
- 对在**关键信息基础设施**应用的人工智能系统进行**登记备案**，要求具备与安全需求相匹配的防护能力
- 建立智能体风险管理框架（附件 2）

配套（4.1.1 / 4.2.2）

- 法律法规层面推动**分类分级监管**、测试评估最终用途管理制度
- 行业层面制定大模型建设部署基础安全指南，从模型选用、部署、运行、停用提出安全基线

炼石解读

沙箱是“包容审慎”原则（1.1）的制度载体，分类分级是“风险导向”原则（1.2）的落点；两者共同决定一个应用要按哪一档安全要求建设。

第 5 章按人工智能应用链条上的四类活动主体给出指引；条数按原文 5.1.1–5.4.7 逐条清点。

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

5.1 模型算法研发（15 条）

- **对象：**模型算法研发者（含基于基础大模型、开源模型二次开发者）
- **覆盖：**算法设计与对齐、训练环境安全、训练数据集与标注、训练中安全验证、定期测评与测试报告、干扰容忍度告知、安全评估披露、版本管理、智能体行为边界、上游缺陷评估、开源协议与来源校验、开源社区参与

5.2 应用建设部署（9 条）

- **对象：**建设部署人工智能应用的单位
- **覆盖：**必要性与风险分级评估、供应链来源与完整性校验、漏洞检测与追溯、访问控制配置（默认配置、默认口令）、接口身份识别与权限控制、数据访问权限与备份、安全护栏、外挂知识库管理、具身智能部署安全评估

5.3 应用运行管理（20 条）

- **对象：**运行管理人工智能应用的服务提供者
- **覆盖：**管理监督与人工复核、最小权限与加密、运行监测与应急、生成内容标识与深伪检测、内容交互规范、日志留存**不少于 6 个月**、透明度与解释、使用者权益与协议、责任主体、数据全生命周期、韧性评估、人员培训、误用滥用处置、特殊群体、事件报告、变更评估与回退、拟人化服务提醒

5.4 应用访问使用（7 条）

- **对象：**使用者
- **覆盖：**选择信誉良好的应用、阅读协议合理预期、个人信息保护意识（联网或记忆功能的智能体审慎披露）、了解数据处理与权限、关注网络安全风险、控制儿童青少年使用时长、识别生成合成内容标识并警惕“深伪”

四类主体在数据上的共同要求：来源合法、最小权限、敏感数据加密与脱敏、日志可审计

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

设计与训练（5.1.1–5.1.6）

5.1.1 算法规则、模型及智能体框架设计环节提升可靠性、透明度、容错性、隐私保护；设计有效可靠的对齐算法

5.1.2 训练环境安全：网络安全配置、**数据加密**；对开发框架、代码做安全审计

5.1.3 训练数据集来源清晰、内容合规，对合成数据做质量和安全性评估

5.1.4 标注流程：交叉标注、结果审计

5.1.5 数据安全和个人信息保护：正当合法必要原则；**去标识化等脱敏处理**；防范泄露、流失、扩散、侵权

5.1.6 训练及迭代中持续校验安全边界、鲁棒性、价值观对齐

测试与披露（5.1.7–5.1.11）

5.1.7 定期安全测试评估，风险分类分级测评与优化机制；对**智能体自主行为、工具调用**专项测评

5.1.8 人工、自动、混合测试；沙箱仿真；多模态、具身智能加强仿真测试；**商业化用途研发者应形成详细测试报告**

5.1.9 评估对外界干扰的容忍程度，以适用范围、注意事项、使用禁忌告知服务提供者和其他研发者

5.1.10 定期披露安全评估、审计与异常处置情况

5.1.11 模型、关键配置、数据集版本管理，**商用版本应可回退**

智能体与二次开发（5.1.12–5.1.15）

5.1.12 合理设定智能体行为边界，可能造成严重损害的操作配备人工审批，约束工具调用和外部交互

5.1.13 基于基础大模型二次开发的，评估模型及上游组件安全缺陷，防范向下游扩散

5.1.14 基于开源模型二次开发的，遵循开源协议，从官方渠道或主流开源社区获取并做来源、完整性校验

5.1.15 参与开源社区建设，提供合规治理解决方案、治理工具、最佳实践

炼石解读：5.1.5 与 5.1.2 把“数据加密、去标识化”写进研发环节，对象是训练数据与训练环境；这与第 5.3.2 条“处理敏感数据时使用加密技术”（运行环节）前后呼应，构成训练侧与运行侧两处数据保护要求。

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

5.2 人工智能应用建设部署的安全指引（9 条）

5.2.1 分级评估

评估应用必要性及长期潜在影响，按场景重要性、智能化水平、应用规模分级，参考风险等级开展安全评估和定期审计

5.2.2 供应链

模型文件、框架工具、第三方库及智能体依赖的工具、插件从官方渠道获取，选成熟版本，做来源、完整性校验和安全测试

5.2.3 漏洞

软硬件、第三方工具、智能体工具插件不含未修复可利用已知漏洞；建立漏洞追溯机制，防范供应链后门

5.2.4 访问控制

准确配置软件、运行参数、功能模块调用策略，禁用非必要端口和服务，重点检查**默认配置、默认口令**

5.2.5 接口管理

人机交互接口和 API 做身份识别与权限控制，**最小化访问权限**，限制调用频率，一般用户禁用高风险操作，恶意用户暂停服务、阻断访问

5.2.6 数据

了解场景的数据安全和隐私保护要求，限制数据访问权限防超范围使用，制定备份恢复计划，定期检查处理流程

5.2.7 护栏

识别拦截违法不良内容、提示词注入攻击，防范输出超出业务范围，核验 RAG、联网交互获取的外部内容

5.2.8 知识库

RAG、知识库问答类应用加强外挂知识库和外部知识来源管理，防范知识库投毒、内容污染，确保生成内容来源可溯

5.2.9 具身智能

按作业场景开展安全评估，配备故障自检、碰撞防护、紧急停止，制定安全操作规程

5.4 应用访问使用的安全指引（7 条）

使用者

5.4.1 选择信誉良好的应用

5.4.2 阅读合同或服务协议，了解功能、限制和隐私政策，合理设定预期

5.4.3 避免不必要地输入敏感信息；使用**具备联网或记忆功能的智能体**时审慎披露个人信息

5.4.4 了解数据处理方式和权限获取情况，避免使用不符合隐私保护原则的产品

5.4.5 关注网络安全风险，避免应用成为攻击目标

5.4.6 控制儿童和青少年使用时长，预防沉迷、情感依赖

5.4.7 根据标识识别生成内容，对“深伪”保持警惕，不轻信、不传播

炼石解读

5.2.5、5.2.6 与 5.3.2 三处均要求**最小权限**，分别落在接口、数据、账户三个层面。

运行管理20条：日志留存不少于6个月，重大变更评估并可回退

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

管理与监测（5.3.1–5.3.7）

5.3.1明确责任方，健全人工复核，关键场景决策透明可控，在人类授权和控制下运行

5.3.2严格管理权限，最小权限原则，增强账户安全，处理敏感数据时使用加密技术

5.3.3运行监测与应急预案，关键指标预警阈值，具备切换到人工或传统系统的能力，定期演练

5.3.4生成内容加显式或隐式标识；政务信息公开、司法取证场景部署深度伪造检测工具

5.3.5内容交互规范、安全运营、投诉反馈、技术防护

5.3.6记录系统行为、用户行为日志，留存不少于 6 个月，定期审计

5.3.7实时风险监控管理机制

透明与权益（5.3.8–5.3.13）

5.3.8公开能力、局限性、适用人群、场景

5.3.9说明目标实现度和偏离度，重大影响决策做解释

5.3.10合同或服务协议以易懂方式告知适用范围、注意事项、使用禁忌

5.3.11支持使用者行使人类监督和控制权利

5.3.12明确数据归属及算法缺陷的责任主体，责任链条可追溯

5.3.13评估数据泄露、隐私泄露、违规收集使用风险，建立数据全生命周期安全管理，加强用户交互数据、RAG 数据防护

韧性与处置（5.3.14–5.3.20）

5.3.14评估故障、攻击等异常条件下的抵御能力，确保最低限度有效功能

5.3.15从业人员安全意识和能力培训

5.3.16协议中明确：发现误用、滥用，提供者有权纠正、暂停或提前终止服务

5.3.17面向未成年人、老年人及特殊群体设置防沉迷、防过度依赖机制

5.3.18事件发现、报告和处置机制；重大事件依法依规向主管部门报告并告知受影响用户

5.3.19更新、升级、微调或调整功能权限的重要变更做安全评估，确认安全能力不下降并保留回退能力

5.3.20拟人化互动服务健全身份提醒、时长提醒和危机干预，防止诱导情感依赖

炼石解读：20 条中有量化要求的只有 5.3.6 “日志留存不少于 6 个月”；其余为定性要求，落地时需结合行业指南（4.2.2）细化。

风险分级三要素：应用场景、智能化水平、应用规模，分五级

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

一、主要分级要素

1. 应用场景

反映实际使用中的运行环境、目标需求；涉及**应用目的、行业领域、使用环境、服务对象**及可能涉及的社会、经济、安全影响

2. 智能化水平

低：仅可作为辅助建议，决策需人工介入
水平提高 → 人工介入频次和范围减小
高：无需人工干预，**全流程自主决策**运行

3. 应用规模

用户范围有限或领域单一（企业内部智能工具、区域性服务）影响相对可控；达到一定规模或深嵌关键行业流程（智能辅助驾驶、城市运行管理、工业生产调度、行业级金融风控模型）可能**快速扩散并引发系统性影响**

关联：原则 1.2、指引 5.2.1 采用同样三个维度；4.4.1 要求对关键信息基础设施中应用的人工智能系统登记备案

二、风险级别：对国家安全、社会稳定和公民权益的影响

1 低安全风险

轻微威胁性且影响范围很小；安全基本无影响，潜在危害轻微

2 一般安全风险

一定威胁性但影响范围有限；影响较小，潜在危害可控

3 较大安全风险

明显威胁性和局部性影响特征；可能带来较大影响，产生**局部社会面危害**

4 重大安全风险

重大威胁性和区域性影响特征；可能带来严重影响，产生**重大社会面危害**

5 特别重大安全风险

灾难性和系统性威胁特征；造成**颠覆性或不可逆转**的特别严重影响

三、风险定级：国家标准 → 行业细则 → 分类分级工作

分类分级国家标准

明确分类分级基本流程、分级要素，给出行业领域细化行业指南的步骤方法

分类分级行业细则

主管（监管）部门：（1）选取并实例化分级要素；（2）制定定级原则、要素权重

风险分类分级

组织本行业单位开展，指导识别、防范化解**重大和较大安全风险**

智能体风险33项：从设计研发到停用下线八环节，其他六项

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

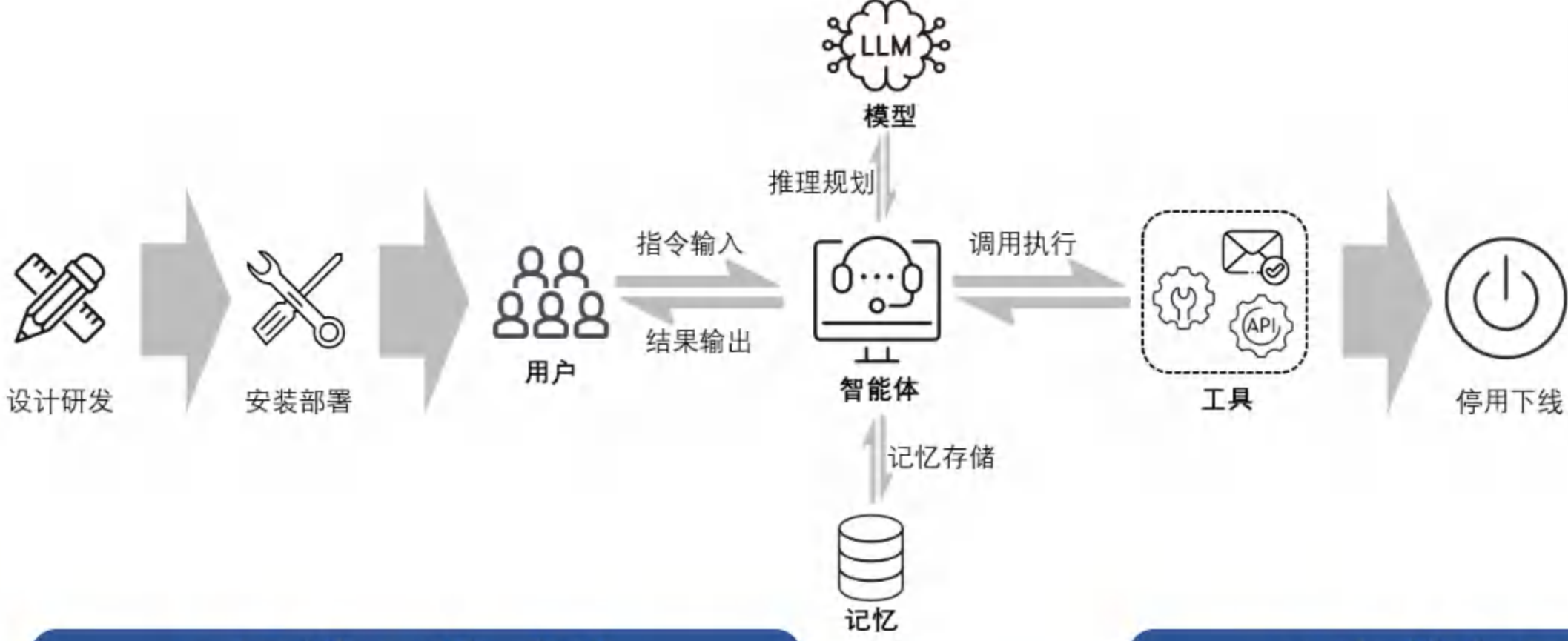
4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则



出处：附件 2 图“智能体生命周期”。

条款：智能体的安全风险贯穿**设计研发、安装部署、指令输入、推理规划、调用执行、记忆存储、结果输出、停用下线**全生命周期各环节。

范围：围绕智能体涉及的大模型、外围工具、记忆、交互协议、技能（skills）等方面的安全风险提出防范措施，供研发者、提供者、使用者参考。

1 设计研发（3）

安全需求分析不充分/技术选型不恰当/安全机制设计不完善

2 安装部署（3）

物料和组件污染/部署环境薄弱/安全测评不足

3 指令输入（2）

提示注入攻击（指令藏于文档、邮件、网页、日志）/上下文溢出攻击

4 推理规划（3）

意图理解偏差/路径偏离/目标劫持

5 调用执行（7）

工具投毒/工具越权/工具选择偏见/应答数据污染/身份伪造/工具劫持/资源过载

6 记忆存储（4）

记忆留存不当（**未加密、隔离**）/记忆失真/记忆污染/**记忆窃取**

7 结果输出（2）

敏感违法信息内容输出/实施恶意为

8 停用下线（3）

服务残留/权限与凭证遗留/数据留存不当

9 其他（6）

通信协议安全/日志缺失/行为逃逸/系统权限滥用/技能管理不当/超范围使用数据

智能体防范34项：唯一身份、最小权限、人工审批、记忆隔离

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分級原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

1 风险预防前置（4）

安全需求确认（风险台账）/安全风险分级/安全控制策略（环境隔离、网络分段、**分阶段渐进式部署**）/安全机制确认（保留版本发布与回退记录）

5 运行时动态管理（6）

输入分类与优先级、来源可信度校验/安全护栏（规划、调用、输出）/**记忆隔离**：凭证、密钥原则上不写入记忆，**敏感个人信息加密**、访问控制、最短保存期限/通信相互鉴别，完整性、机密性、可用性、抗重放/自主执行参数（步骤、频次、时长、资源）/沙箱、容器隔离高风险操作

2 身份与权限管理（3）

身份管理：**唯一身份标识**，禁止实例间共享/权限管理：划清仅限用户决策、需用户授权、智能体自主决策的边界，**最小权限**/凭证管理：标准化动态授权凭证，超范围触发拦截预警，停用后立即撤销

6 持续监测与审计（8）

异常阻断/数据管理：必要原则，提供给大模型或第三方前取得同意，**境内数据境内存储**/全量日志（文件、指令、网络、技能、交易支付）并脱敏防篡改/安全审计/沙箱环境验证/常态化红队测试/应急预案/重大变更安全影响评估与回归测试

3 加强人工审批（4）

风险分级控制：高风险操作转交用户接管，中低风险需用户授权/人工控制节点：删除文件、发送数据、修改配置须二次确认，提供回滚撤销/审批日志防篡改/审批系统异常或用户无响应时**默认拒绝**

7 停用安全管理（3）

全面关停服务（进程、端口、连接，撤销第三方授权、取消订阅）/备份必要数据/清理部署环境（卸载、重置、清除残留文件、账号凭证）

4 供应链与工具管理（6）

调用前检查版本、描述文件、参数、元数据/公平调用/工具异常检测告警阻断/清理闲置、权限过高工具/技能优先来源可靠、经安全测试/第三方组件来源、完整性、版本校验

炼石解读

七类措施与 3.2.1 的四条技术措施一一对应：身份权限（a）、人工审批（b）、供应链工具（c）、监测审计（d）；附件 2 把每条展开为可操作项。

涉及数据保护的落点：**记忆加密与隔离、日志脱敏防篡改、境内存储、交互数据同意与撤回。**

可信人工智能八项准则：人类最终控制、尊重主权、价值对齐

条款：落实《全球人工智能治理倡议》，遵循“以人为本、智能向善”，共同防范应对人工智能技术失控风险，促进技术在世界范围内可信应用。

1. 人工智能安全治理原则

2. 人工智能安全风险分类

3. 人工智能安全风险技术应对措施

4. 人工智能安全风险综合治理措施

5. 人工智能安全指引

附件 1 人工智能安全风险的分级原则

附件 2 智能体风险管理框架

附件 3 可信人工智能基本准则

1 人类最终控制

关键环节设置人类控制机制，最终裁决权归属人类：安全控制阈值、**安全终止开关**、人工干预有效窗口

2 尊重国家主权

尊重所在国主权与法律，依法接受监管，不得干涉他国内政；尊重各国数字主权，不得胁迫选边站队

3 价值观对齐

和平、发展、公平、正义、民主、自由的全人类共同价值融入全生命周期

4 提升系统透明度

功能目标、运行逻辑、模型使用、数据来源、决策依据的必要披露

5 促进可客观验证

客观、公正、透明的测试与认证机制，功能、性能、安全特性、决策链条可被技术验证

6 强化安全防护

设计部署中强化风险建模、安全测试和防护机制，全生命周期审计与记录

7 前瞻预防应对

前瞻性风险识别评估、动态监测、应急响应；模型安全风险评测，推动评测方法、基准国际互认

8 全球协同共治

支持联合国主渠道，反对以小圈子治理取代全球治理，加强面向发展中国家的能力建设

炼石解读：八项准则中第 1、6、7 项对应第 4.5.1 条“开发者定期测试潜在技术失控风险”的要求；第 2 项写明数据与数智技术路径的自主选择权。

落地要点：按附件1定级，按第5章对号指引，按附件2管智能体

1 定级：先确定风险级别

出处：附件 1 / 4.4.1

按应用场景、智能化水平、应用规模三要素评价，对照五级；关基场景的人工智能系统按 4.4.1 做登记备案；行业细则出台后按行业规范复核

2 对号：按主体找到自己的指引

出处：第 5 章

研发者 15 条、建设部署 9 条、运行管理 20 条、使用者 7 条；一个单位同时扮演多个角色时叠加适用

3 数据：训练侧与运行侧两处保护

出处：5.1.5 / 5.3.2 / 5.3.13

训练数据去标识化、训练环境数据加密；运行中敏感数据加密、最小权限、用户交互数据与 RAG 数据防护；日志留存不少于 6 个月

4 智能体：按附件 2 建七类控制

出处：附件 2 / 3.2.1

唯一身份与最小权限、高风险操作人工审批且异常默认拒绝、记忆隔离（凭证密钥不入记忆）、工具供应链校验、全量日志脱敏防篡改、停用清理

5 内容：标识与检测两端并举

出处：3.2.4 / 4.4.2 / 5.3.4

生成合成内容显式或隐式标识；政务公开、司法取证场景部署深伪检测；GEO 投毒场景关注信息来源真实性与商业内容标识

6 变更：评估、回退、报告

出处：5.1.11 / 5.3.19 / 5.3.18

商用版本可回退；更新、微调、调整权限前做安全评估，确认安全能力不下降；重大事件依法依规报告主管部门并告知用户

炼石解读：框架 3.0 的数据安全要求集中在训练数据、运行时敏感数据、智能体记忆与日志四处，均落在应用与智能体内部，而非网络边界

注：本页为炼石依据原文条款整理的落地提示，条款号可回溯；框架为指导性文件，具体义务以后续法律法规、国家标准和行业细则为准。

一 图解《人工智能安全治理框架3.0》

二 数据安全建设实践



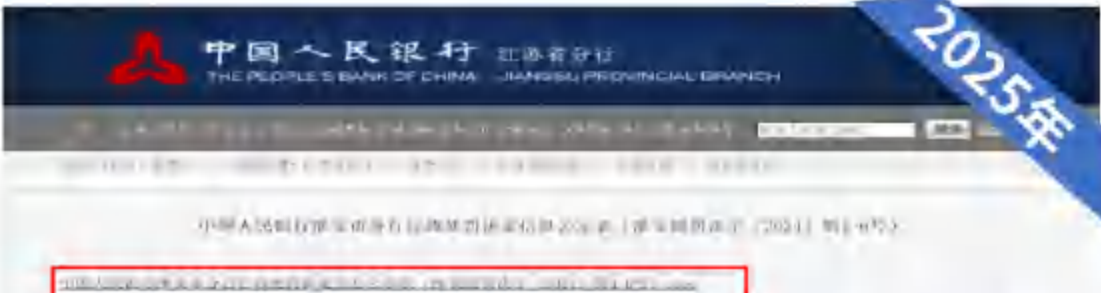
安全需求：“风险与合规”共同驱动数据安全建设

安全产业自身

数据安全产业发展阶段性（先发展再治理，今天已到拐点）、自身特殊性（经济学外部效应带来的密集法律法规），形成非线性增长驱动力

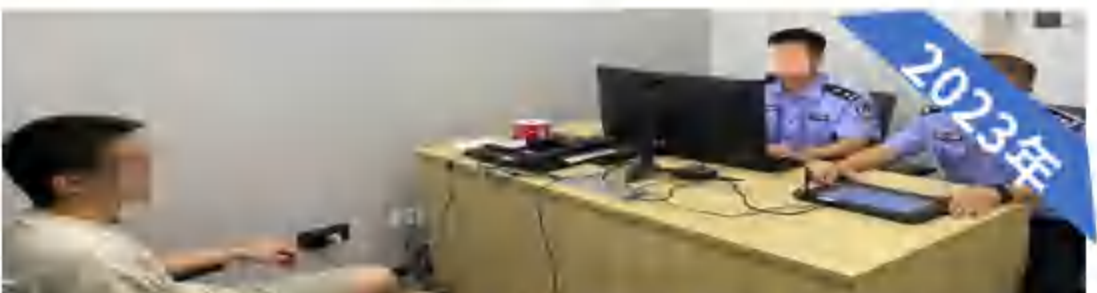
内在风险：数据风险如影随形伴生数据处理

根据IBM最新发布的《2025年数据泄露成本报告》，数据泄露的平均成本超444万美元。



2025年

江苏某地方农商银行因违反网络安全与数据安全管理规定等被警告，并被罚款114.5万元。



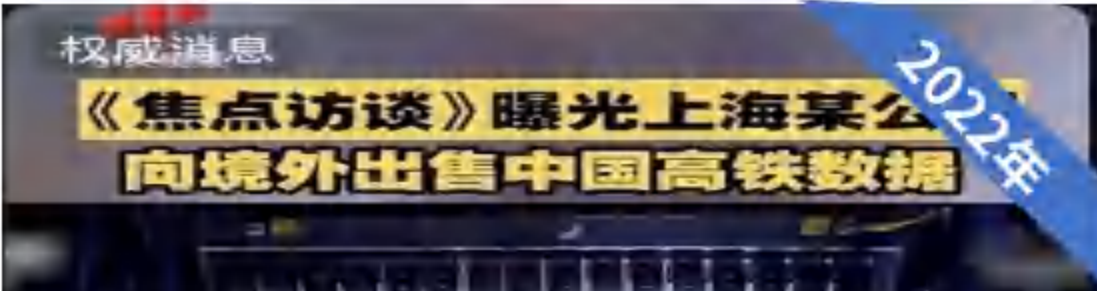
2023年

浙江某科技公司为市政府开发运维信息系统，私自迁移数据，造成严重数据泄露罚款100万元。



2023年

南昌某高校3万条师生个人信息数据在境外互联网被公开售卖，被南昌公安网安部门罚款85万元。



2022年

《焦点访谈》曝光上海某公司向境外出售中国高铁数据

国家安全机关破获我国首例涉及高铁运行安全的危害国家安全类案件

经济学外部效应：数据泄露给他人造成损害，对企业影响反而不大。解决办法是通过立法，让“防数据泄露”成为公共安全产品，类似环保

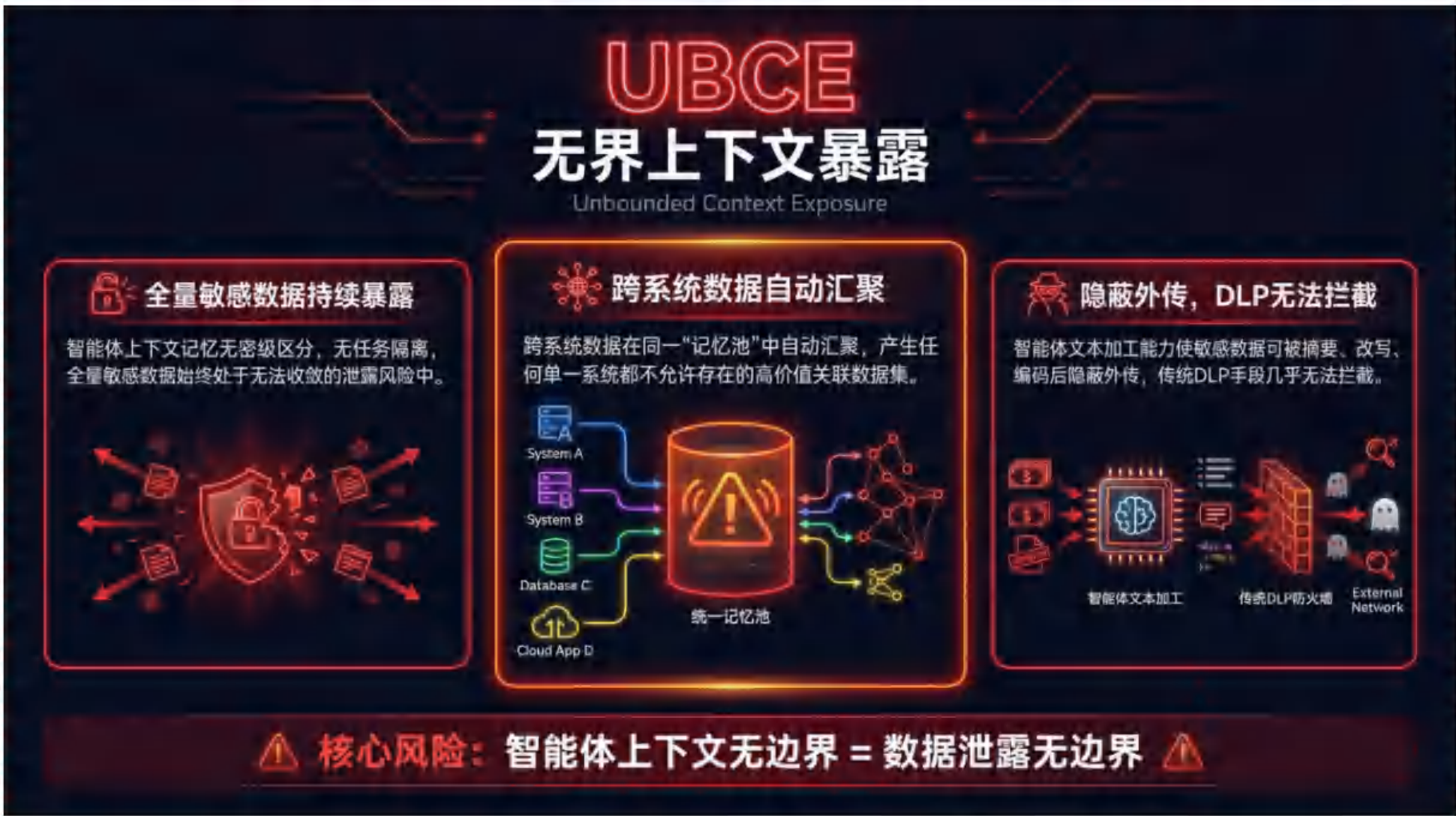
外部合规：“五法一典”过程与结果双合规

网络安全法	个人信息保护法	数据安全法	密码法
网络安全等级保护条例 (征求意见稿)	关键信息基础设施安全 保护条例	网络数据安全条例	商用密码管理条例
等保	关保	数评	密评
信创			

信息技术产业

“AI新时代” 拉动数据开发利用，数据安全更加紧迫

安全防护与信息技术呈共生关系，每次技术革新都伴随安全技术升级，大模型、智能体等颠覆性技术突破，更要求安全防护体系进行范式级重构

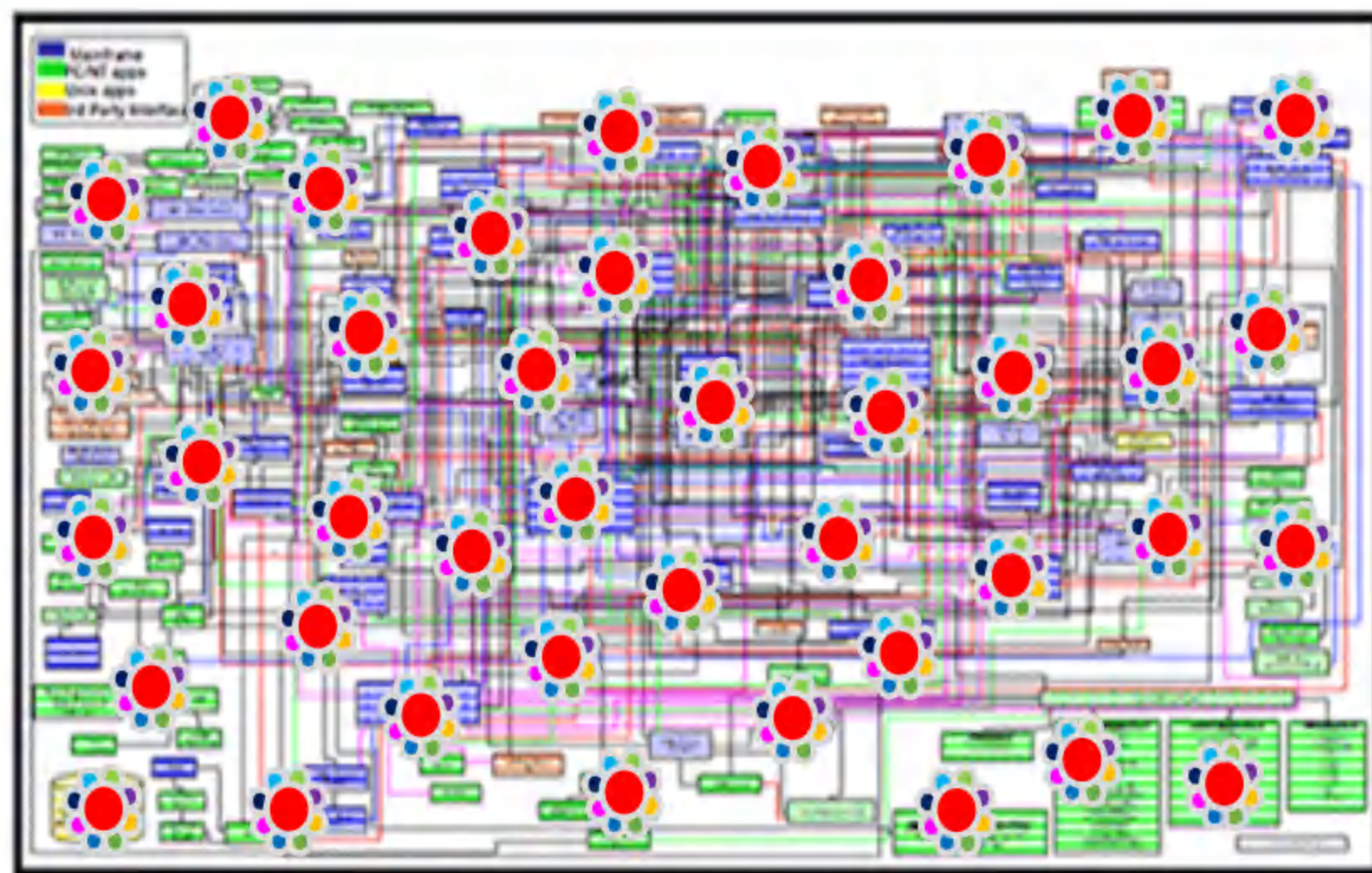


全量暴露，风险不收敛：上下文无密级区分，敏感数据始终处于泄露敞口

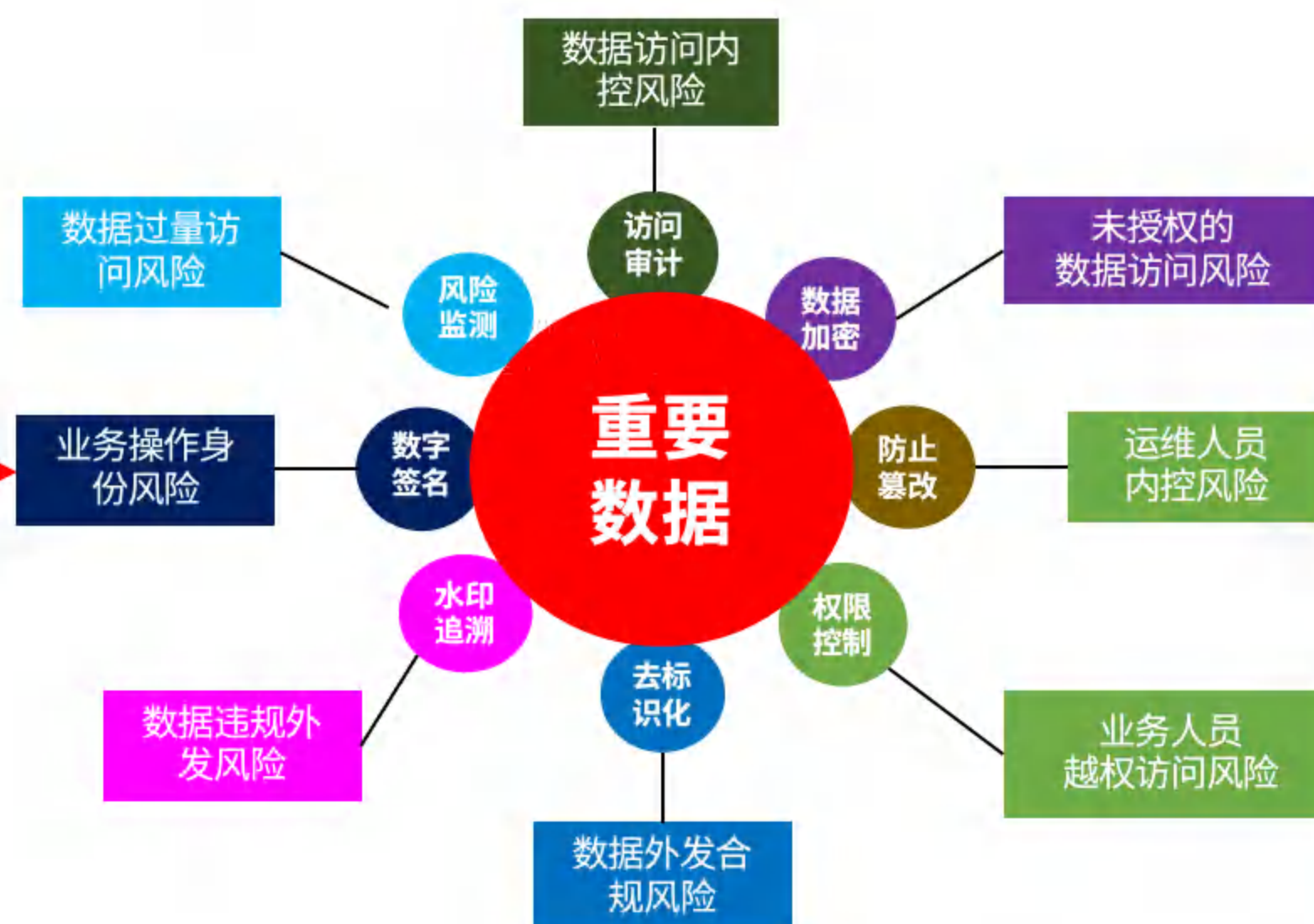
跨系统汇聚，产生超权数据集：多源数据在同一记忆池中自动聚合，形成单一系统权限下不应存在的高价值关联

语义变形外传，传统 DLP 系统性失效：摘要、改写、编码等能力使敏感数据经变形后隐蔽流出，特征匹配机制无法识别，防护体系存在根本性盲区

落地痛点：安全建设从网络到数据，安全和业务纽结缠绕难以改造



企业应用系统总体架构图示例



应用系统升级涉及面广，成本高周期长

已上线系统升级风险大，影响业务连续

数据安全建设面临“两难”

- 数据保护体现为**面向应用的功能型安全需求**，要在应用中融合实现，**但改造重要信息系统成本高、风险大、周期长，新的AI应用场景下传统安全手段影响效率，因此安全机制总是滞后或缺失**
- 而数据保护不落实，则合规风险越来越大、因泄露导致的业务风险会快速累积



免改造产品场景：数据安全合规准入、实战评价

密评密改合规产品

一站式密改套餐敏捷交付

免改造密改产品



70分密改套餐

应用和数据安全密码模块
信源加密与身份鉴别
关基免改造密码机
协同签名系统
服务器密码机
软件密码模块
密码卡

数据安全实战防护产品

密码安全一体化+事前事中事后全防护

数据安全管控平台



数据库加密

文件加密

密钥管理系统

密码套件

事前

数据发现识别、分类分级
静态脱敏 动态脱敏
细粒度访问控制

事中

数据访问检测响应、API安全
数据流转安全态势感知
终端安全沙箱

事后

访问审计 水印追溯

智能体安全沙箱

本地执行安全可控

外部接入：
业务Agent

↓拦截明文请求↓

5层纵深防护专属隔离区

L1	提示注入防御
L2	端点白名单
L3	凭据保护
L4	WASM沙箱
L5	行为审计

大模型TEE平台

远端推理可信可证

跨域传输

- ① 远程证明握手
- ② 国密TLS通道
- ③ 密钥协商

加密请求

加密结果

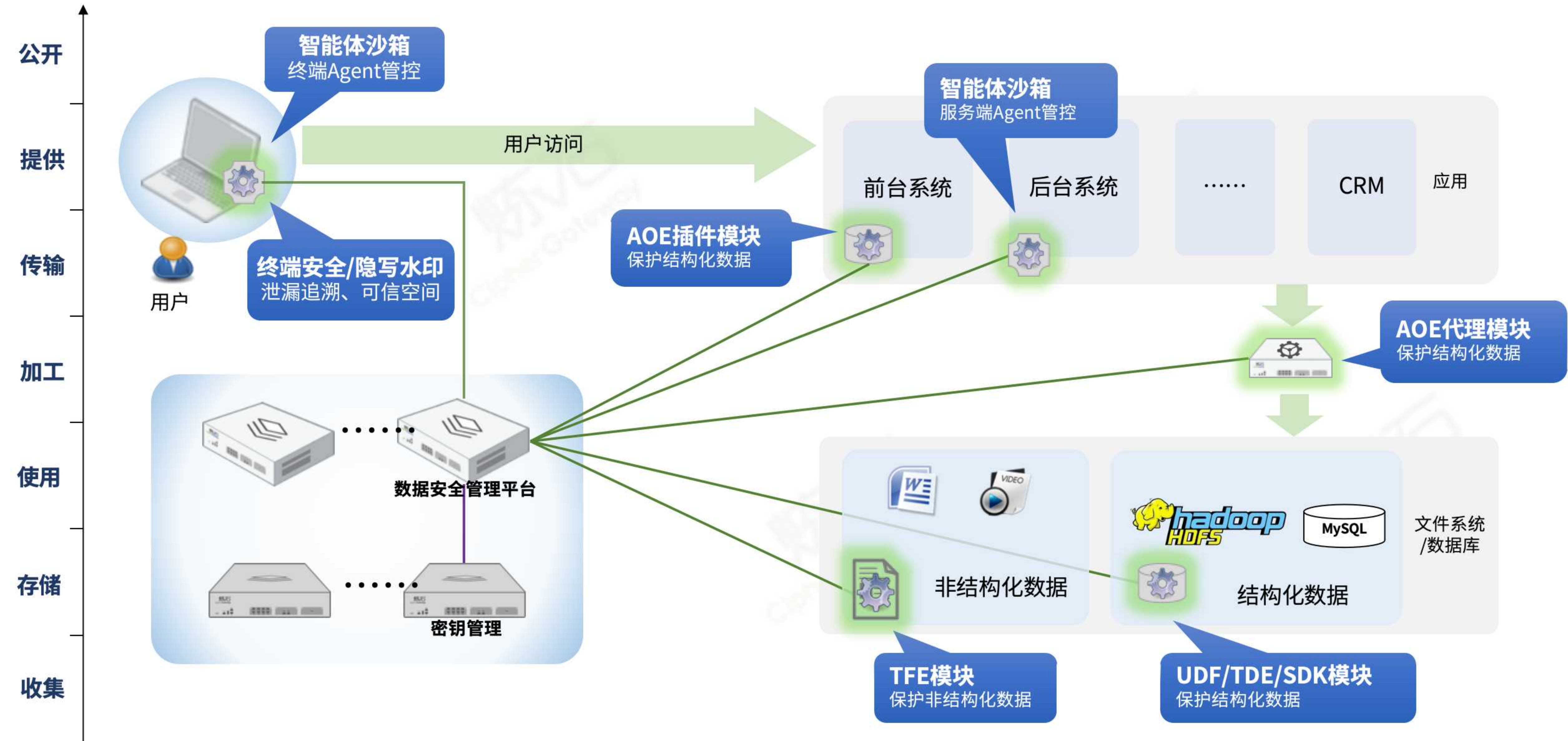
- ① 国密TEE信任根与远程证明算力环境先验证、后连接
- ② 模型权重内存态国密保护
TEE内存态: 动态解密运行
- ③ 推理全链路国密护栏
加密Prompt输入→受保护内存态
推理输出合规过滤
- ④ 多节点TEE协同推理与GPU
算力隔离
分布式协同推理
- ⑤ 沙箱-TEE端到端协同联动
本地控权限、远端保安全

AOE面向切面加密 · 重要网络与信息系统保护

AOS面向切面沙箱 · AI数据保护

免改造数据安全产品平台

敏捷部署：旁路平台结合主路控制面，免改造应用落地保护



AOE · 密评整改

一站式密改套餐敏捷交付

- 最贴合业务
- 业务风险影响范围广，改造难度大
- 涉及核心的高风险项，且分值最高

- 属于密码应用层，含高风险项最多
- 密改落地由用户执行，无法绕过
- 所需密码产品会随着信息系统新增或变化，而扩容或调整

- 属于密码支撑层
- 密改落地不一定由用户执行，有可能属于IDC托管方、云服务商等
- 整体被评，可复用到被承载的每个信息系统

密码测评规则

依据	GB/T 39786-2021 《信息安全技术 信息系统密码应用基本要求》
通过要求	60分以上，无高风险
测评规则	打分制，满分100，共41项检测项，包含8个方面
典型问题	无合规的密码设备，无管理制度
评估周期	重要网络与信息系统（关基、等保三级、政务等） 每年至少一次

高风险项，一票否决

- 实体间未通过密码技术进行身份鉴别（网络和通信：无缓解措施）
- 未采用密码技术保证通信过程中重要数据的机密性
- 未采用密码技术建立安全的信息传输通道，如SSLVPN
- 未采用密码技术保证重要数据在存储过程中的机密性（无缓解措施）
- 未采用密码技术保证重要数据在存储过程中的完整性
- 未采用密码技术保证重要数据在传输过程中的机密性
- 未采用密码技术提供数据原发证据和数据接收证据（法律责任场景）
- 未建立密码相关管理制度
- 未制定密码应用解决方案、实施方案和应急处置方案，未通过评审
- 未采用密码技术对从外部连接到内部的设备进行认证（第四级）

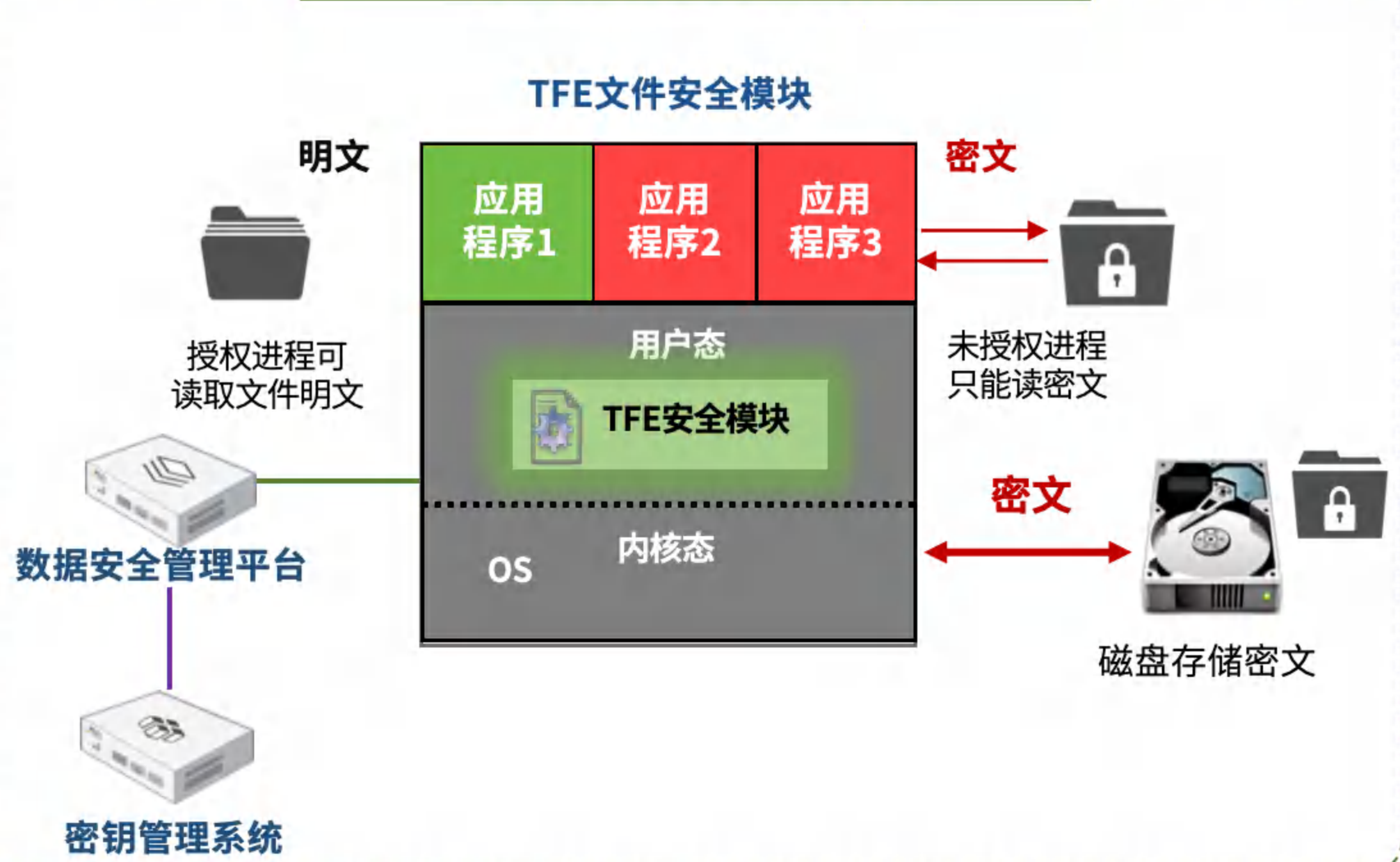
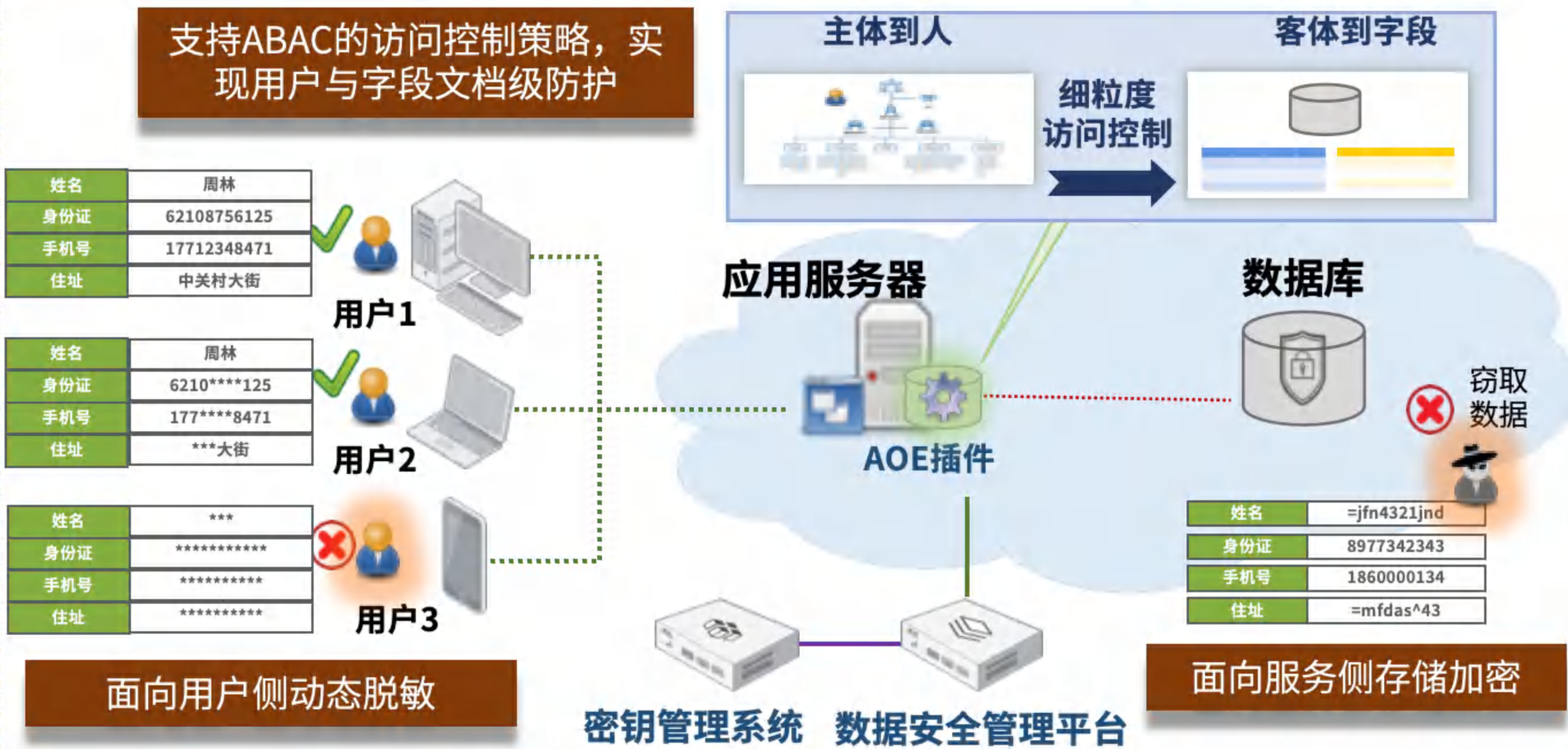
密码合规要点（以密码应用三级系统为例）

物理和环境	身份鉴别	门禁记录	监控记录	密码服务资质	密码产品认证	10分		
网络和通信	身份鉴别	数据的完整性和机密性	访问控制信息完整性	设备认证	密码服务资质	密码产品认证	20分	
设备和计算	身份鉴别	加密通道	访问控制和安全标记的完整性	日志和程序完整性	真实性	密码服务资质	密码产品认证	10分
应用和数据	身份鉴别	传输和存储的机密性	完整性	不可否认性	密码服务资质	密码产品认证	30分	
管理制度	密码人员管理 密钥管理 建设运行 应急处置 密码软硬件 介质管理等制度						30分	
人员管理	角色 职责 权限 审计机制 培训与考核 保密协议							
建设运行	密码应用方案 实施方案							
应急处置	应急处置方案 及时上报							

场景之一：服务端存储加密、使用解密/脱敏

结构化数据

非结构化数据



不影响业务人员使用、不影响DBA运维

示例

全名	电话	卡等级	发送时间
张*	*****2714	非常客	2020-01-24 09:05...
高**	*****2342	非常客	2020-01-24 09:05...
黄*	*****6862	非常客	2020-01-24 09:05...
前**	*****7169	非常客	2020-01-24 09:05...
刘**	*****4643	普卡客户	2020-01-24 09:05...
王**	*****4865	非常客	2020-01-24 09:05...
马**	*****9668	非常客	2020-01-24 09:05...

应用查看脱敏后数据

示例

全名	电话	卡等级
3N8OWkdylyOoB...	13171630482	非常客
3NYeVE1Kmgag...	13570417164	非常客
3v8zWU9gmSi4e...	14735816218	非常客
39IzVW1xlgH3K...	10250726110	非常客
0s4IWUBsmgG+f...	13156901590	非常客
0s4IWUBsmgG+f...	11965283955	非常客
3/4GWV5Vlx2yTD...	18777677924	非常客

应用看到密文

示例

PASSENGER_ID	NAME	TELEPHONE_NUMBER	卡等级
10109100000000021509627	李*	*****8373	普通
10109100000000023121143	刘**	*****0388	普通
10109100000000023933621	王**	*****9413	普通
10109100000000025412895	张*	*****0017	普通
10109100000000025727366	王**	*****6196	银卡
10109100000000027788076	龙*	*****9182	普通
10109100000000027904363	梁**	*****4627	普通
10109100000000032198789	赵**	*****0954	普通
10109100000000039504224	王*	*****6052	普通
10109100000000082017159	郭**	*****0088	普通

DBA工具查看脱敏数据

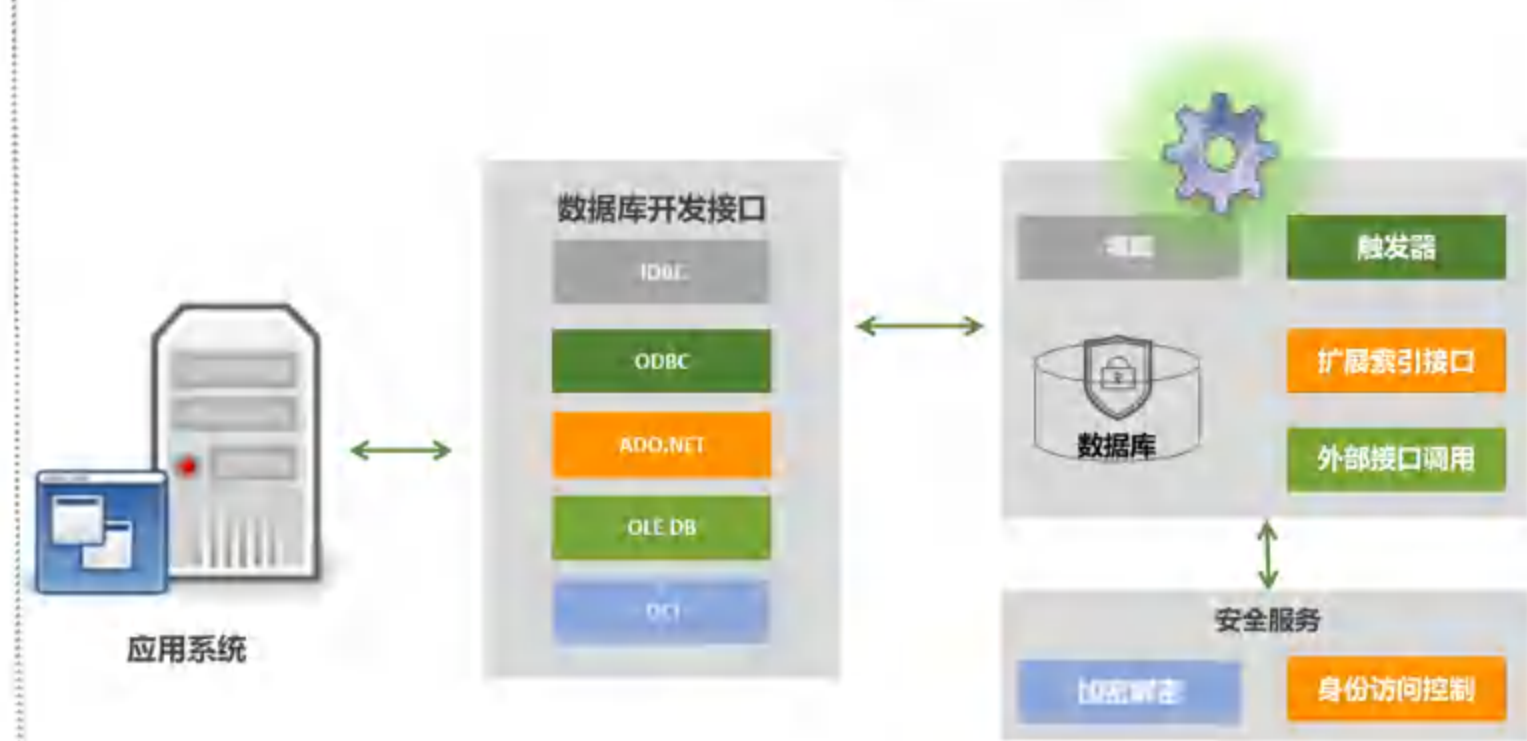
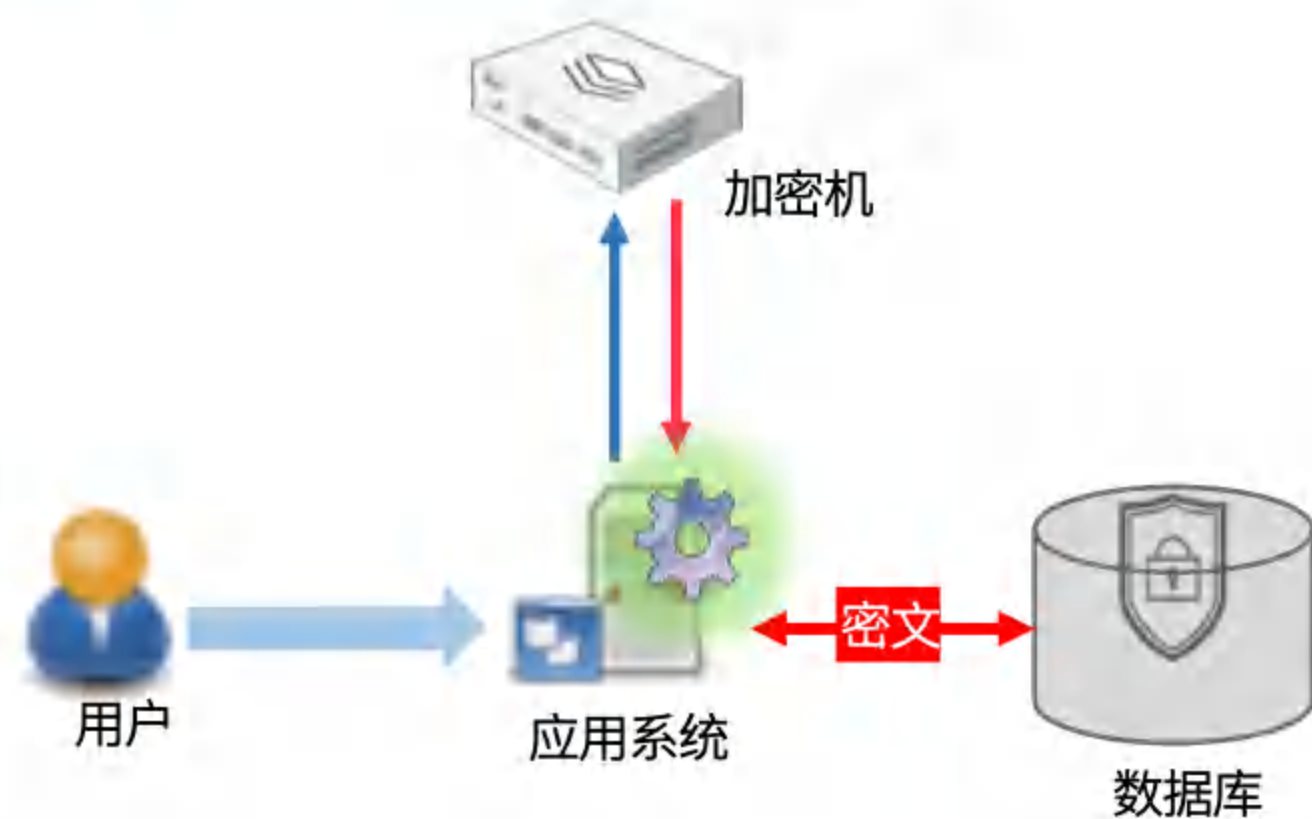
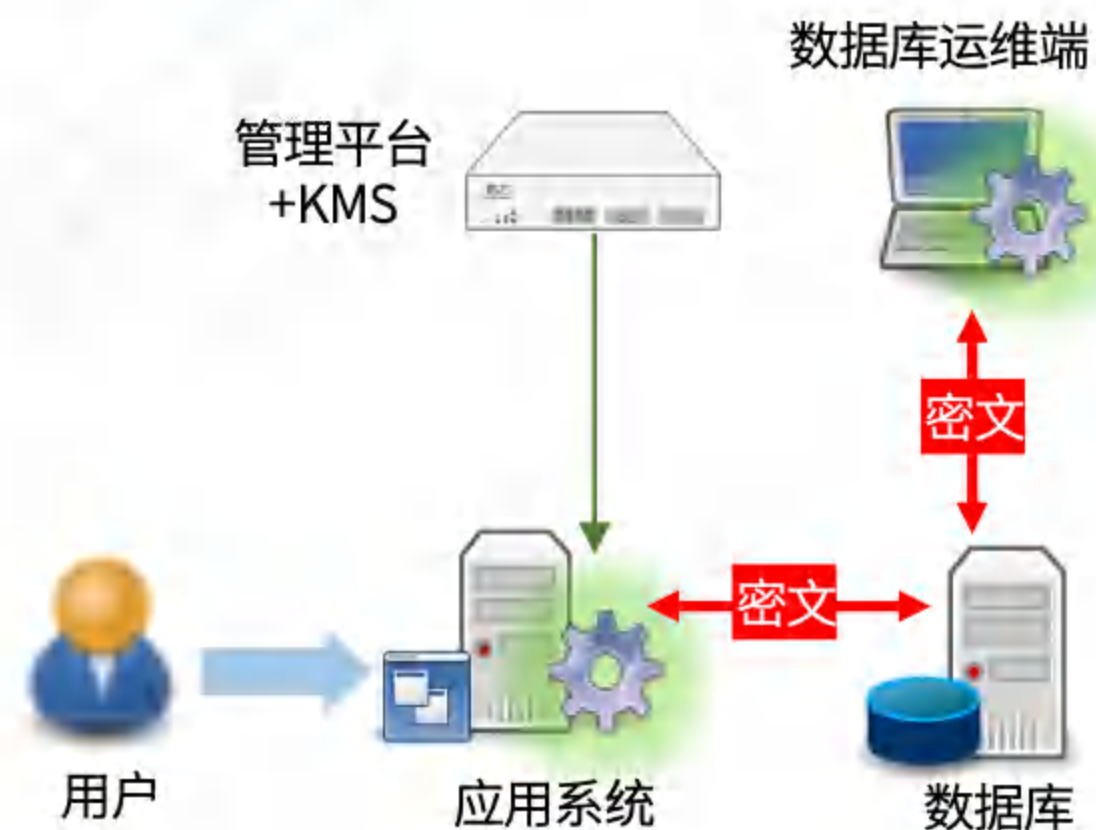
示例

ID	NAME	TELEPHONE_NUMBER	卡等级
000021509627	3N8oWXh5yXS8fw==	17765361362	普通
000023121143	38o+WX1ImhGqaYDtUA==	10248759753	普通
000023933621	3cwtWV5+mDyGkE40w==	18519574810	普通
000025412895	3/4GW1NpJZ1aYA==	12918666358	普通
000025727366	3cwtWk8bmDyqbxCwDA==	16627391320	普通
000027788076	0/w/VVN3yct1Ew==	14133053398	普通
000027904363	30AnVhN/mi2x/f80Gg==	14954994219	普通
000032198789	0vcTWf9anCeU99nJux==	14443421156	普通

DBA工具查看密文

技术对比：“服务端数据安全”典型路线

指标	免开发改造的本方案（以加密为例）	改造应用方案（以加密为例）	数据库外挂安全方案（以加密为例）
实施成本	实施工作量小，免改造应用代码	实施工作量大，需改造应用代码，集成加密机客户端SDK	实施工作量中，需针对目标数据库的每个版本做适配
运维成本	运维成本低工作量大，变更时仅需通过可视化管理平台即可完成，操作简单快捷	运维成本高工作量大，变更时需要将前期投入的改造工作重新再投入一遍，不可缺省	运维成本中工作量大，需要改动数据库新增加的视图和触发器等，对数据库服务器稳定性产生隐患
可靠性	可靠性高，插件成为应用程序的一部分，不单独启进程，插件可靠性取决于应用自身可靠性	可靠性中，数据需要传输给集中部署的加密机进行加解密，形成性能瓶颈和单点故障风险	可靠性低，外挂模块需要在数据库服务器单独启进程，增加了单点故障风险，同时需要对数据库改表名、加规则
性能	性能高，基于高性能国密算法、脱敏算法	性能中，需要将数据传入传出加密机，额外增加网络延时	性能低，针对字段添加触发器和多层视图
兼容性	支持绝大多数数据库，与数据库解耦	支持绝大多数数据库，取决于SDK编程语言类型	仅支持Oracle等部分数据库，对国产数据库、NoSQL、乃至DB2、SQL Server等兼容性差
数据流	不改变原数据流转路径和网络拓扑，免改造模块原地加解密数据、脱敏数据	改变数据流转路径，明文先经过SDK传入到加密机，加密后将密文返回SDK，最终存入数据库	改变数据流转路径，明文先经过外挂模块传入到加密机，加密后将密文返回外挂模块，最终存入数据库
扩展性	可灵活扩展，统一可视化策略管理，横向覆盖更多应用，不影响已部署应用的加解密性能	扩展成本高，加密策略需要通过硬代码实现，任何策略改动，都需要改动代码，不灵活而且风险大	扩展成本中，需适配新数据库及改表名，对集中式加密资源增加压力，会分摊到已部署数据库的加解密算力
应用升级影响	不影响应用升级，加解密等功能与应用轻耦合	应用升级需要维护加密代码，加解密与应用代码紧耦合	应用升级要评估数据库，占用数据库服务器宝贵性能资源



规模部署案例-72个业务系统（50家软件厂商），免开发改造敏捷交付



项目背景介绍

客户：某政务平台
项目名称：数据安全密码防护项目

复杂数据处理流程的统一防护能力：
针对某政务平台复杂业务系统（包括44个生产系统、23个管理系统和5个外网系统），采用商用密码技术对敏感个人信息和重要数据进行加密保护，实现了字段级的数据加密，有效提升了数据安全性。通过统一集中式管控，能够集中管理各个系统中的加密策略，确保数据安全防护的一致性和高效性。

跨厂商与技术的高效兼容能力：
项目涉及50家软件厂商，涵盖了多种编程语言（JAVA、.net、C++等8种）和数据库（Oracle、Mysql、PolarDB等9种），产品跨异构系统的集成能力确保了数据安全防护方案能够无缝融入现有系统，无需对应用开发进行大规模改造。

敏捷实施与快速落地能力：
项目通过1个项目经理在6个月内实现了敏捷实施和快速上线。免应用开发改造的技术优势大幅缩短实施周期，保障数据安全防护方案能够迅速落地并满足合规要求。

外网系统（5个）			
系统名称	开发厂商	开发语言	数据库
**装备综合信息服务网	太极计算机股份有限公司	Java	Mysql
****部政府信息公开平台		Java	Mysql
****部网站			
****部科信之家门户网站	拓尔思	Java	MySQL
****统一门户系统	联通数科	Java	MySQL

管理系统（23个）				
系统名称	开发厂商	开发语言	数据库	量级
智慧机关系统	北京卓越联腾	Java	高斯100	5W
离退休干部系统		Java	PolarDB	
标准化管理系统	华宇蓝创科技	Java	Mysql	5K
****部政务外网OA办公平台		Java	Kingbase	
物资物流系统（资源管理系统）	京东	Java	Mysql	10W
视频会议系统	华为	.net	SQLserver	
安全监管系统	太极计算机股份有限公司	Java	mysql、Kingbase	
安全生产信用信息系统	太极计算机股份有限公司	Java	MySQL	
危险化学品经营安全管理系统	常州市瑞德信息科技有限公司	Java	MySQL	30G
烟花爆竹安全监管信息综合管理子系统	常州市瑞德信息科技有限公司	Java	Kingbase	20G
运维管理平台	北京广通优云科技股份有限公司	Java	达梦	
全国安全生产知识技能考核信息系统	浪潮软件股份有限公司	Java	华为openGauss	2W
****干部网络学院	知学云	Java	MySQL	3T
互联网+监管系统二期		vue+antdesign	无	
中央部门预算管理系统				
国务院安全生产委员会考核巡查系统	太极计算机股份有限公司	Java	Kingbase	1W
档案信息系统		Java	Kingbase	
****科技众创子系统	北京安信创业信息科技发展有限公司	Java	Kingbase	10W
全国安全生产专项整治三年行动系统	太极计算机股份有限公司	Java	Mysql	4k
**宣教公共服务系统	新华网(北京)	Java、Vue	KingbaseV8R6	100G
监测预警值守管理子系统	北京航天数据	Java	Kingbase	5W
小鱼易连视频会议系统	小鱼易连	C++、Java	Mysql	
视频联网系统	海康威视	C++、Java	Postgresql	1K

生产系统（44个）				
系统名称	开发厂商	开发语言	数据库	数据量
安全生产大数据应用系统		Java	Mysql	
应对特别重大灾害指挥部现场协调保障系统	北京辰安科技股份有限公司	Java	Postgresql	17G
安全生产gis系统	北京辰安科技股份有限公司	node.js	Postgresql	10G
**指挥一张图系统		Java	Postgresql	
****大数据应用平台		Java	Mysql	
互联网+执法系统	太极计算机股份有限公司	Java	Gaussdb	
安全生产综合信息平台		Java	Mysql	
****科技信息化基础能力和资源统计系统	北京安信创业信息科技发展有限公司	Java	Mysql	1W
全国烟花爆竹流向管理信息系统	北京丹灵云	Java	Oracle	2T
灾害事故复盘系统	贵州东方世纪科技股份有限公司	Java	Postgresql	
防汛抗旱防台风会商子系统	四创科技有限公司	Java	Oracle	
数据治理系统	杭州数梦工场	Java	高斯+hadoop	
****一张图系统	北京辰安科技股份有限公司	Java	Postgresql	6T
危化品风险预警系统		Java	Postgresql	
互联网+政务系统	正孚软件	Java	高斯100	
灾害综合风险监测预警系统	太极计算机股份有限公司	Java	Kingbase	
地灾**平台	武汉地大信息工程股份有限公司	Java	MySQL	10G
重特大灾害现场指挥协调辅助系统	广州数鹏通	Java	MySQL	
非药品易制毒管理系统	北京安信创业信息科技发展有限公司	Java	Kingbase	8W
**装备数据智能服务子系统	北京安信创业信息科技发展有限公司	Java	mysql、Kingbase	
**装备综合信息大数据智能服务系统	徐工汉云技术股份有限公司	Java	PolarDB	200G
工业互联网+危化安全生产公共服务系统	中国安全生产科学研究院	Java	Kingbase	
粉尘涉爆系统	浙江图讯科技股份有限公司	Java、Vue、H5	openGauss 2.1.0	3K
**装备实战仿真推演系统	北京未尔锐创科技有限公司	Java	Mysql	15G
国家油气输送管道地理信息系统	中国安全生产科学研究院	Java	mysql	
通信大数据辅助决策子系统	中移系统集成有限公司	Java	Kingbase、PolarDB	20W
国家**指挥综合业务系统	西安未来国际信息股份有限公司	Java	GaussDB100	750G
防汛抗旱决策支持系统	贵州东方世纪科技股份有限公司	Java	polar-o	
调查评估和统计分析应用系统	北京安信创业信息科技发展有限公司	Java	KingbaseV8R6	
卫星遥感数据处理服务系统	航天宏图信息技术股份有限公司	Java	PolarDB	18G
自然灾害综合监测预警系统	太极计算机股份有限公司	Java	PolarDB	1T
预案管理和**演练系统	救援协调局	Java	Mysql、Postgresql	
北斗**综合服务系统	北京安信创业信息科技发展有限公司	Java	Postgresql	100W
大数据应用支撑平台				
视频资源管控平台	小鱼易连	Java	华为openGauss	1G
全国**避难场所辅助调度系统	北京百分点科技集团股份有限公司	Java	Postgresql	12W
危险化学品登记综合服务系统	化学品登记中心	Java	Postgresql	400G
地震灾害**响应与决策支持系统	航天科工海鹰集团、北京英莫	python	Postgresql	3k
****部网络党校	知学云	Java	Mysql	3T
**救援航空调度子系统	中航材、中科软、北斗天汇联合体	Java	Mysql	10W
灾害事故e键通	北京泰策科技有限公司	Java	mysql8	200W
危化品风险预警系统	化学品登记中心	Java	Postgresql	6T
国家自然灾害救助资金管理系统	四川迪思腾科技有限责任公司	Java	polardbx	8000W
专家管理子系统	易天气（北京）科技	Java	kingbase	10W

密评合规：70分密改套餐

测评层面	测评单元	备注	满分	得分	密码产品
物理和环境安全 (10分)	身份鉴别	高风险	4.17	0.00	高风险可缓解
	电子门禁记录数据存储完整性		2.92	0.00	
	视频监控记录数据存储完整性		2.92	0.00	
网络和通信安全 (20分)	身份鉴别	高风险 (无缓解)	5.71	5.71	IPSec VPN安全网关 SSL VPN安全网关
	通信数据完整性		4	4	
	通信过程中重要数据的机密性	高风险	5.71	5.71	
	网络边界访问控制信息的完整性		2.29	2.29	
	安全接入认证		2.29	2.29	
设备和计算安全 (10分)	身份鉴别	高风险	2.56	2.56	SSL VPN安全网关
	远程管理通道安全	高风险	2.56	2.56	
	系统资源访问控制信息完整性		1.03	1.03	应用和数据安全密码模块 软件密码模块
	重要信息资源安全标记完整性		1.03	1.03	
	日志记录完整性		1.03	1.03	
	重要可执行程序完整性、重要可执行程序来源真实性		1.79	0.00	
应用和数据安全 (30分)	身份鉴别	高风险	5.77	5.77	应用和数据安全密码模块 协同签名服务端密码模块 协同签名客户端密码模块 服务器密码机 软件密码模块 密码卡
	访问控制信息完整性		2.31	2.31	
	重要信息资源安全标记完整性		2.31	2.31	
	重要数据传输机密性	高风险	5.77	5.77	
	重要数据存储机密性	高风险 (无缓解)	5.77	5.77	
	重要数据传输完整性		4.04	4.04	
	重要数据存储完整性	高风险	4.04	4.04	
	不可否认性	高风险	--	--	仅限法律责任场景
安全管理 (30分)	按得分80%计算		30	24	专业管理制度模板
合计			100	82.22	

* 参考密评联委会《商用密码应用安全性评估量化评估规则（2023版）》

密改四件套，轻松过密评

国密VPN
网关产品

解决网络和通信
20分、缓解高风险

密码认证
网关产品

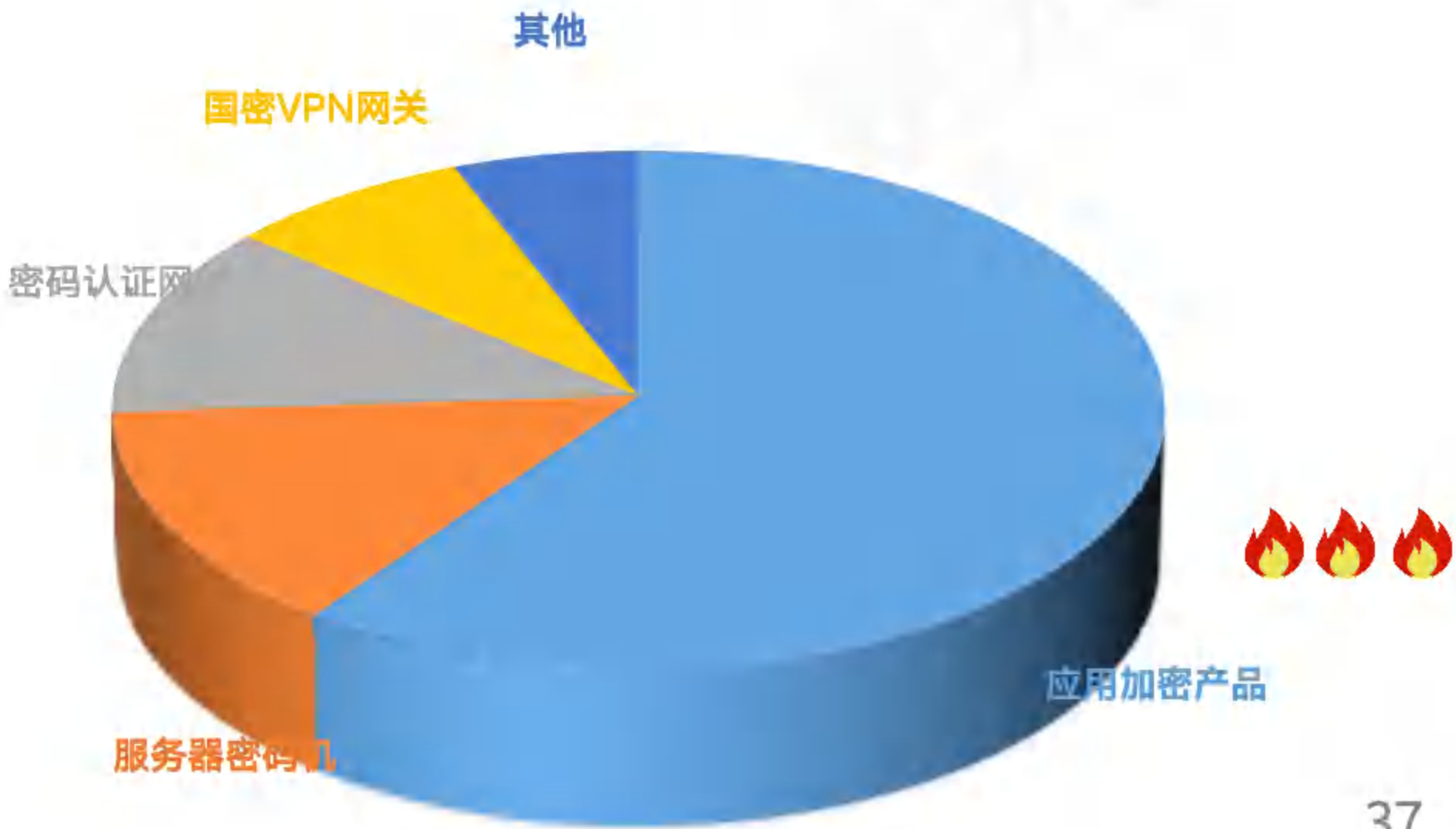
解决应用和数据
5分、缓解高风险

应用加密
网关产品

应用和数据21分、
高风险，设备和
计算4分

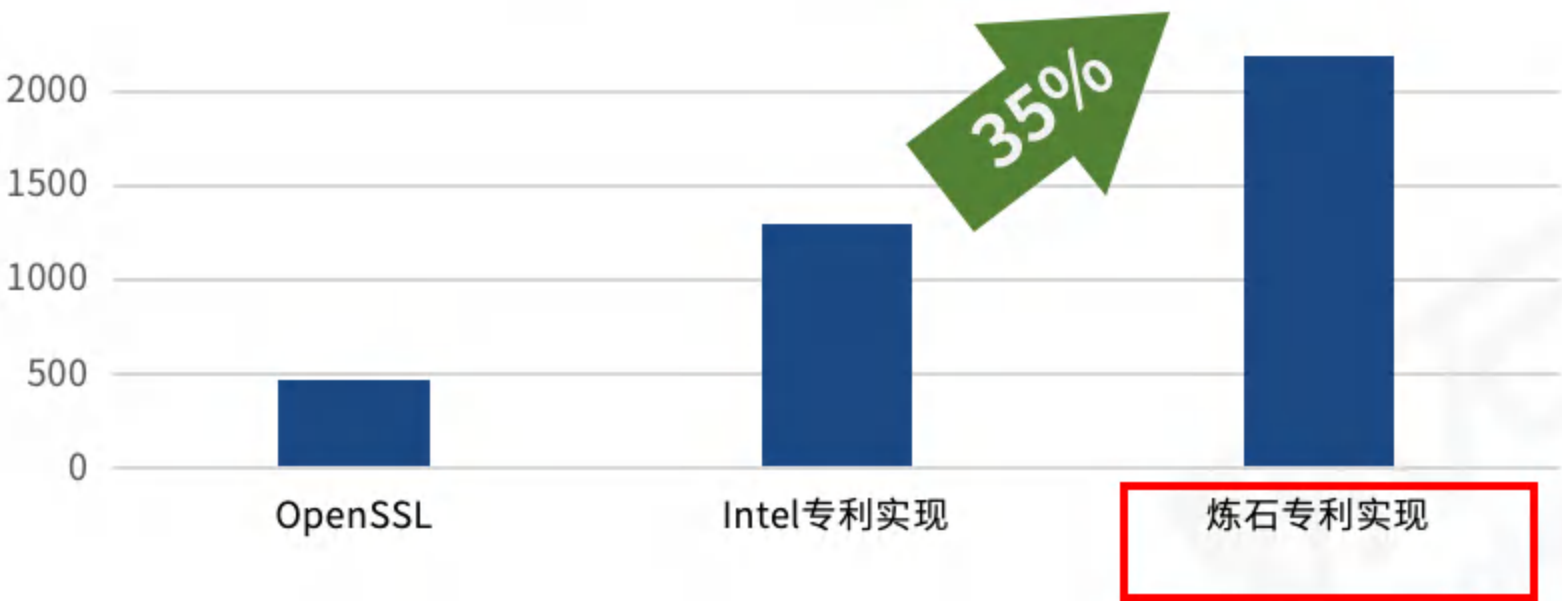
服务器密码机
提供基础加解密算力与密管能力

多场景：单应用密改、批量应用密改、大规模应用分期密改



高性能：PCT专利保护的高速国密实现，不影响业务效率

数据SM4加解密测试



基于PKI的数字信封加密

协同签名身份鉴别

密钥派生接口

EVP国密接口

FPE格式保留加密

Java

C#

python

Ruby

C Programming

C++

GO

SM2

SM3

SM4

SM9

ZUC

Windows 10

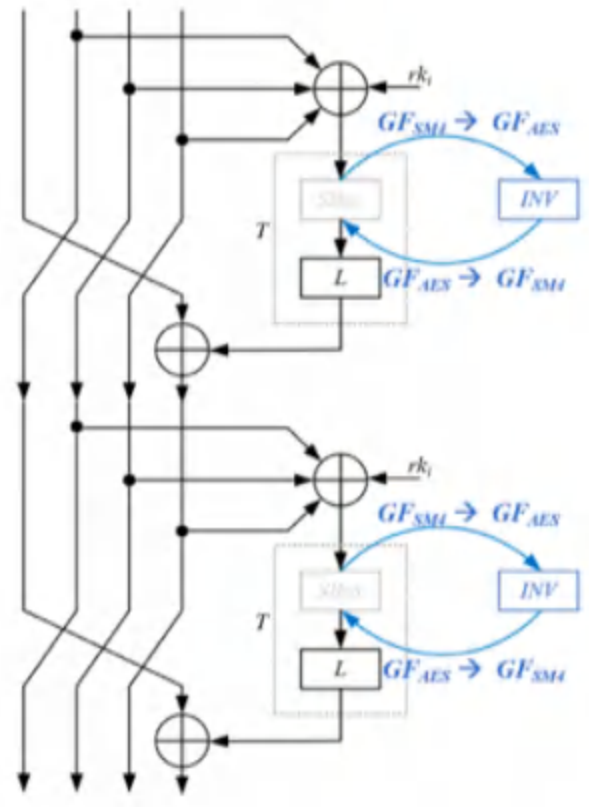
Linux

Mac OS

android

iOS

Intel专利



基于SM4的保留格式加密 (FPE)

加密10亿条手机号，仅耗时20秒

天下密码
唯快不破

炼石独有PCT专利
打破Intel垄断

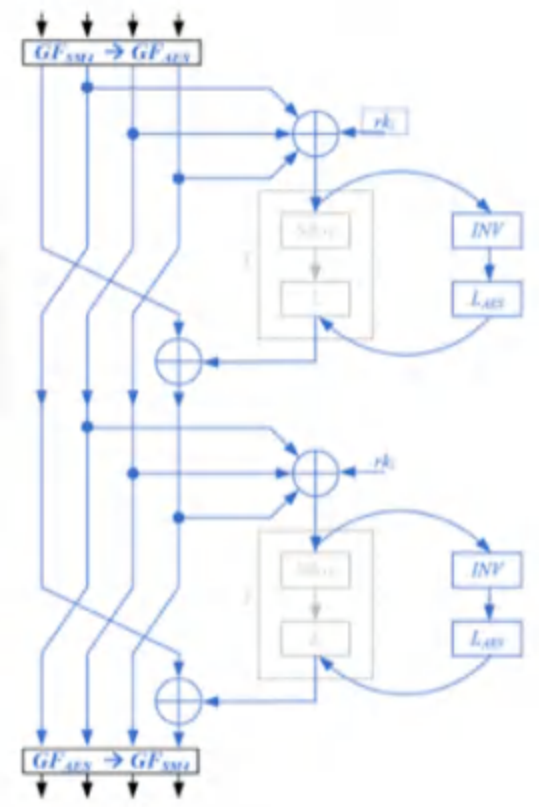
单颗CPU上SM4加解密
突破 140Gbps!

相当于每秒可加密6部3GB大小的高清电影

移动端SM4加解密
突破1.7Gbps!

相当于每秒可加密100多张2MB大小的高清照片

炼石PCT专利 (安全、性能高)



高可靠：专业密钥管理、实时离线解密，保障高可用

KMS密管设备



真随机数

根密钥

- 1、根据门限算法加密导出至3个USBKey，任意2个可还原
- 2、跨设备需人工同步

真随机数

模块密钥

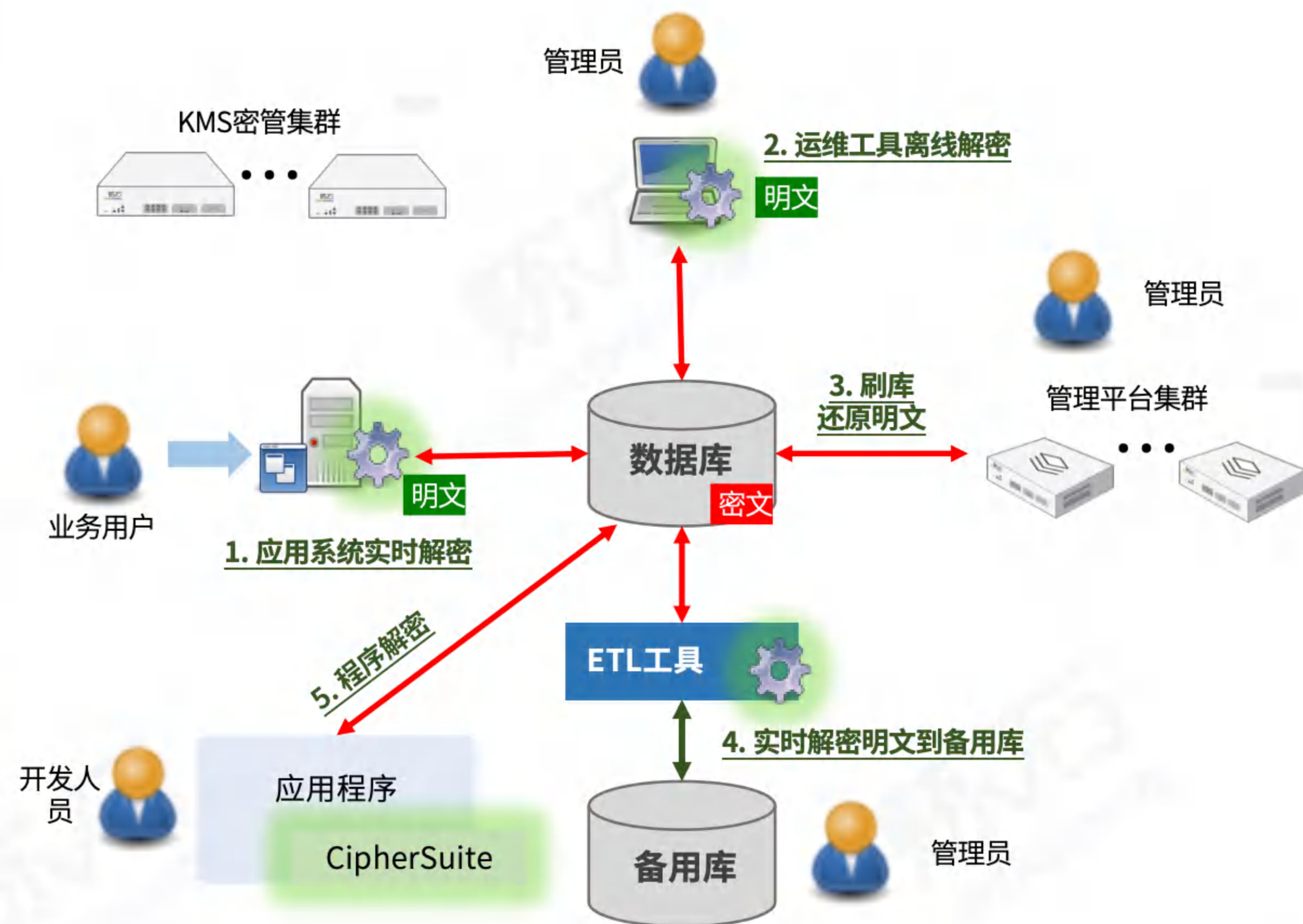
- 1、由根密钥加密落盘存储
- 2、跨设备自动同步

模块密钥派生

工作密钥（数据加密、数据MAC）

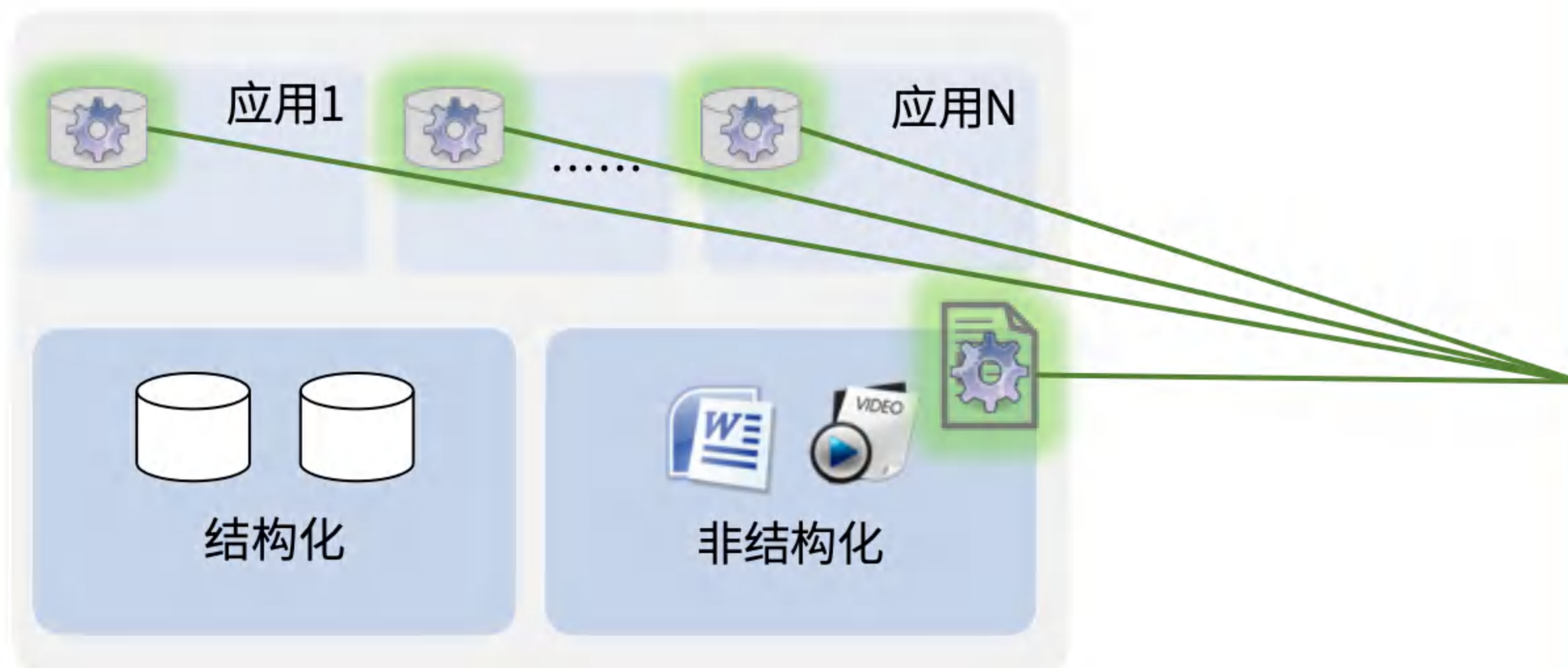
- 1、不落盘存储
- 2、由模块密钥实时派生

- 密管设备**支持双机热备、两地三中心等高可用模式**，根密钥明文不出密码卡，模块密钥明文不出密管设备
- **工作密钥安全从平台分发到数据控制面模块，模块可缓存工作密钥后独立运行**，平台不需要实时在线
- 保守策略下，当重启应用服务后，模块才需从平台更新密钥



- 密码算法是标准，密钥管理机制可靠，**只要密文数据还在，一定可解密**
- 常态状态下，支持应用系统解密、支持DBA运维工具直接解密
- 应急情况下，支持刷库还原明文、支持实时解密明文到备用库
- 极端情况下，用运维工具或程序可以兜底解密

高可用：模块健壮可诊断、平台“两地三中心”



【模块可离线运行】

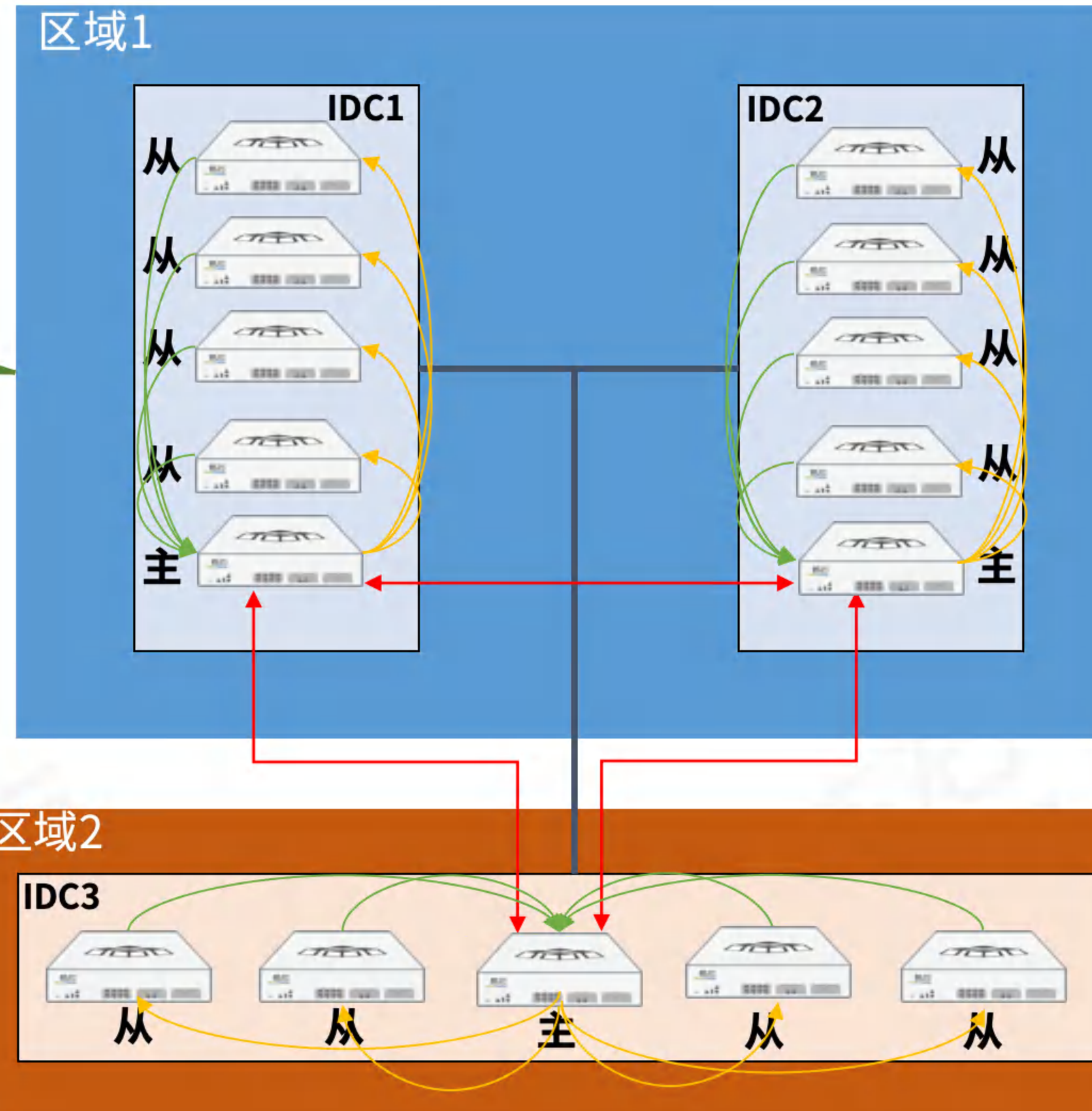
- 模块从平台获取规则后可缓存在内存中，当平台不可用也可正常运行
- 仅当模块所在服务重启时才需要从平台集群重取规则

【模块强化健壮性】

- 以AOE插件为例，本身作为应用的一部分，无单独进程，不会死进程
- 在没有匹配规则状态下，模块默认采用透传模式，最大限度优先业务
- 模块经过多年研发打磨、运行实践，兼容广泛苛刻的企业级应用场景

【模块可诊断】

- 模块自身可记录环境、启动流程、异常事件等日志，以及告警功能
- 针对Java等应用环境支持不停机实时诊断机制，排查潜在故障成因



- 两地三中心各5套安全管理平台（以及KMS）组成集群，三中心的主之间实时双向同步数据，在同一中心内主从间同步数据

支持数据加密后的密文态查询

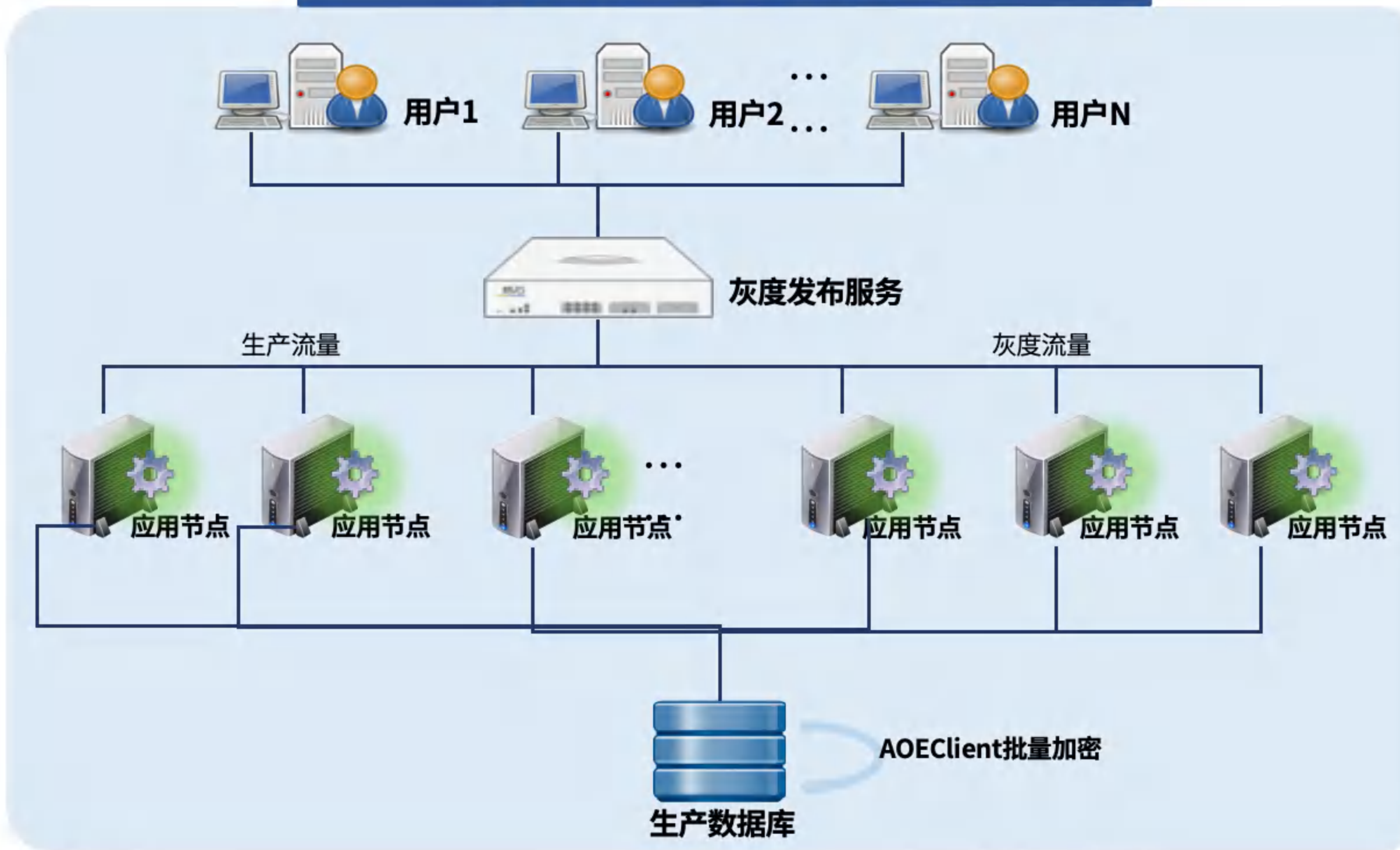
类型	规则	示例
a	手机号型	按照3、4、4拆分加密
b	身份证号型	按照6、4、4、4拆分加密
c	住址型	按照3、3、其余拆分加密
e	其他	

姓名	性别	年龄	手机号	身份证号
张三	男	20	188 1111 2222	110102 1996 0102 3333
n1 n2 n3			p1 p2 p3 p4	ID1 ID2 ID3 ID4 ID5

密钥ID	n1,n2,n3	男	20	p1,p2,p3,p4	ID1, ID2,ID3,ID4,ID5
------	----------	---	----	-------------	----------------------

- 支持密态数据库的模糊查询操作。对加密字段做特定处理，先将加密字段按照规则拆解，然后对拆解后部分分别加密处理，之后将各部分密文合成一个字段（拆解的规则可由应用定制化，通过策略管理来进行配置）
- 限制条件
 - 支持应用对加密策略进行定制，但加密策略确定后原则上不建议更改
 - 为支持模糊查询会对加密字段的长度进行扩张

不停服加密关键应用



前置条件:

- 应用支持以灰度发布方式更新发布
- 生产所有应用节点提前安装插件，且插件配置为支持明密文共存场景

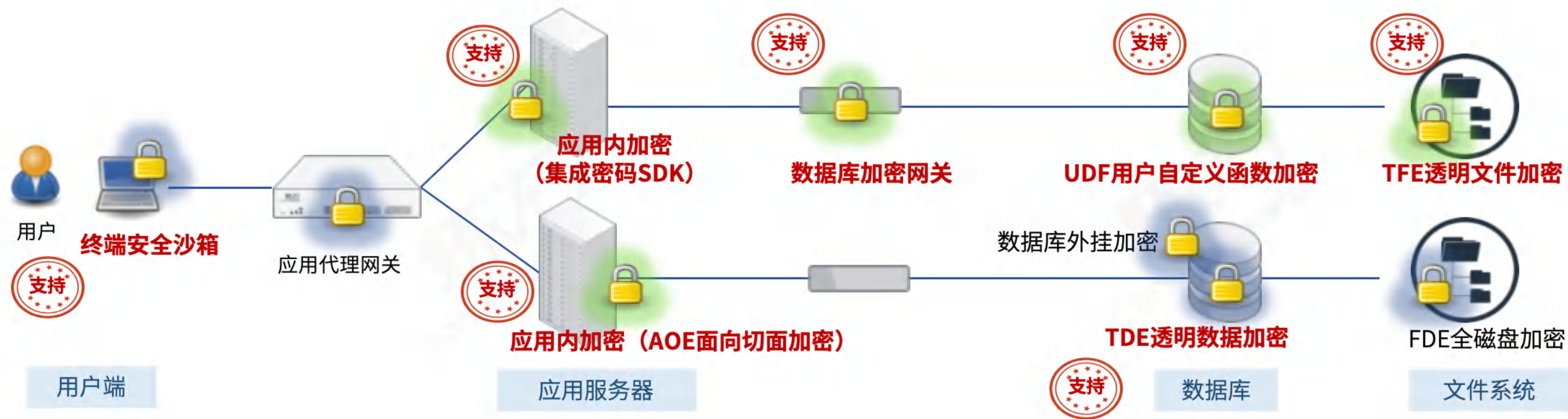
割接步骤:

- 应用滚动更新安装支持明密文共存场景的AOE插件
- 验证业务功能
- 对存量数据执行批量加密任务
- 加密过程中对业务功能进行验证
- 加密完成后应用滚动更新安装正常AOE插件
- 验证业务功能

好用好管：集中式管控，分布式保护，分阶段上线



覆盖齐全：一站式支持多路线，结合业务场景交付最优组合



加密技术	部署位置	加密粒度	性能	防DBA	DB存储过程 密文计算	实施成本	可靠性
终端安全沙箱	终端	文件/虚拟卷	高	/	/	低	高
应用代理网关	终端-应用服务器之间	字段/文件	中	支持	影响	高	低
应用内加密（集成密码SDK）	应用服务器	字段/文件	高	支持	影响	高	高
应用内加密（AOE面向切面加密）	应用服务器	字段	高	支持	影响	低	高
数据库加密网关	应用服务器-数据库之间	字段	中	支持	影响	低	中
数据库外挂加密	数据库	字段	低	支持	影响	中	中
TDE透明数据加密	数据库	字段/表空间	高	不支持	不影响	低	中
UDF用户自定义函数加密	数据库	字段	高	不支持	不影响	高	高
TFE透明文件加密	文件系统	文件	中	不支持	不影响	低	高
FDE全磁盘加密	文件系统	磁盘/卷	高	不支持	不影响	低	高

AOE · 数评整改

事前事中事后全防护

数据安全法律法规位阶（金融行业为例）



立法法规定我国法律效力位阶共分六级，从高到低依次为根本法、基本法、一般法、行政法规、地方性法规、规章，效力位阶依次降低。



合规筛选：国家“两法一条例”与金监局央行“两办法”技术要求



I.识别	P.防护		D.检测	R.响应	R.恢复	C.反制	场景
数据分类分级标识	数据加密	数据去标识化	数据安全风险监测		数据容灾备份	数据访问审计	数据出境安全评估
数据安全风险评估	访问控制	数据匿名化	API变更安全测试			数据转移追溯	数据引入或共享管控
个人信息保护影响评估	漏洞管理	安全认证	数据安全应急响应与演练				人工智能数据安全
	隐私计算	专用线路传输					

序	数据安全能力	“两法一条例”条款	《银行保险机构数据安全管理办法》条款	《中国人民银行业务领域数据安全管理办法》条款
1	数据分类分级	《个人信息保护法》第五十一条；《条例》第五条、第二十九条	第十六~十九条、第二十一条	第六~九条、第十二条、第三十条、第八条、第十九条
2	数据安全风险评估	《数据安全法》第二十九条、第三十条；《条例》第三十一条、第三十三条、第五十二条	第二十二、三十四、六十六	第三十四、三十六、四十二
3	个人信息保护影响评估	《个人信息保护法》第五十六条	第五十八	/
4	数据加密、密码保护	《个人信息保护法》第五十一条；《条例》第九条、第三十三条	第二十八	第三十二、三十五
5	匿名化、脱敏/去标识化、删除	《个人信息保护法》第五十一条；《条例》第二十四条	第二十七、三十八、四十七	第十七、二十六、三十、三十三、四十六、三十八
6	访问控制	《个人信息保护法》第五十一条；《条例》第九条、第三十条	第四十一、四十二、四十九	第二十九、三十五
7	安全认证、身份管理	《条例》第九条	第四十七	第二十九、第三十一
8	漏洞管理	《数据安全法》第二十九条；《条例》第十条	第六十八	第四十一
9	数据安全风险监测		十一、七十二、七十四	三十九、四十、四十二、四十五
10	应急演练	《个人信息保护法》第五十一条、第五十七条；《条例》第十一条、第三十条	十一、五十二、六十四	第四十四
11	数据容灾备份	《条例》第九条	三十七、四十五、四十九	第三十二
12	数据访问审计	《条例》第二十七、五十二	二十八、三十、四十三、四十九	四十五、四十六
10	数据转移追溯、风险溯源	/	第三十三	第三十
11	数据出境安全、数据外部引入或合作共享	《个人信息保护法》第五十九；《条例》第十二条	第二十	第二十四
12	人工智能数据安全	《条例》第十九	五十二	/
13	隐私计算	/	/	二十三、三十六
14	专用线路数据传输	/	/	第三十五
15	API变更安全测试	/	/	第三十六

数评合规： 国标83项407条提炼技术构建实战

ICS 35.030
CCS L 80



中华人民共和国国家标准

GB/T 45577—2025

数据安全技术 数据安全风险评估方法

Data security technology—Risk assessment method for data security

2025-04-25 发布

2025-11-01 实施

国家市场监督管理总局
国家标准化管理委员会

发布

制定目的

为贯彻落实《数据安全法》关于数据安全风险评估的要求，**国家市场监督管理总局、国家标准化管理委员会联合发布《数据安全技术数据安全风险评估方法》（GB/T45577-2025）**，用于指导开展数据安全风险评估工作。

主要内容

该标准描述了数据安全风险评估的**基本概念、要素关系、分析原理**,给出了数据安全风险评估的**实施流程、评估内容、分析评价方法**等；适用于指导**数据处理者、第三方评估机构**开展数据安全风险评估,也可供有关**主管监管部门**实施数据安全检查评估时参考。

8.3 数据安全管理	a) 数据安全管理制度	A.1.1.1 数据安全制度体系 A.1.1.2 数据安全制度落实
	b) 安全组织机构	A.1.2.1 数据安全组织架构 A.1.2.2 数据安全岗位设置
	c) 分类分级管理	A.1.3.1 数据资产管理 A.1.3.2 数据分类分级制度 A.1.3.3 数据分类分级保护
	d) 人员安全管理	A.1.4.1 人员录用 A.1.4.2 保密协议 A.1.4.3 转岗离岗 A.1.4.4 数据安全培训
	e) 合作外包管理	A.1.5.1 合作方管理机制 A.1.5.2 合作协议约束 A.1.5.3 外包人员访问权限 A.1.5.4 第三方接入与数据回收 A.1.5.5 政务数据委托处理
	f) 安全威胁和应急管理	A.1.6.1 安全威胁和事件 A.1.6.2 安全应急管理
	g) 开发运维管理	A.1.7 开发运维管理
	h) 云数据安全	A.1.8 云数据安全 (使用云计算服务) A.1.8 云数据安全 (云计算服务提供者)
8.4 数据处理活动安全	a) 数据收集安全	A.2.1.1 数据收集合法正当性 A.2.1.2 通过第三方收集数据 A.2.1.3 数据质量控制 A.2.1.4 数据收集方式 A.2.1.5 数据收集设备及环境安全
	b) 数据存储安全	A.2.2.1 数据存储适当性 A.2.2.2 逻辑存储安全 A.2.2.3 存储介质安全
	c) 数据传输安全	A.2.3.1 传输链路安全性 A.2.3.2 传输链路可靠性
	d) 数据使用和加工安全	A.2.4.1 数据使用和加工合法性 A.2.4.2 数据正当使用 A.2.4.3 数据导入导出 A.2.4.4 数据处理环境 A.2.4.5 数据使用和加工安全措施
	e) 数据提供安全	A.2.5.1 数据提供合法正当必要性 A.2.5.2 数据提供管理 A.2.5.3 数据提供技术措施 A.2.5.4 数据接收方 A.2.5.5 数据转移安全 A.2.5.6 数据出境安全
	f) 数据公开安全	A.2.6.1 数据公开适当性 A.2.6.2 数据公开管理
	g) 数据删除安全	A.2.7.1 数据删除管理 A.2.7.2 存储介质销毁
	h) 其他数据处理活动安全	A.2.8 其他
	a) 网络安全防护	A.3.1 网络安全防护
	b) 身份鉴别与访问控制	A.3.2.1 身份鉴别 A.3.2.2 访问控制 A.3.2.3 授权管理
	c) 监测预警	A.3.3 监测预警
	d) 数据脱敏	A.3.4 数据脱敏
8.5 数据安全技术	e) 数据防泄露	A.3.5 数据防泄露
	f) 数据接口安全	A.3.6.1 对外接口安全 A.3.6.2 接口安全控制
	g) 数据备份恢复	A.3.7 数据备份恢复
	h) 安全审计	A.3.8.1 审计执行 A.3.8.2 日志留存记录 A.3.8.3 行为审计
	a) 个人信息处理基本原则	A.4.1.1 合法、诚信原则 A.4.1.2 正当、必要原则
	b) 个人信息告知	A.4.2 个人信息告知
	c) 个人信息同意	A.4.3 个人信息同意
	d) 个人信息处理	A.4.4.1 个人信息保存 A.4.4.2 个人信息共同处理 A.4.4.3 个人信息委托处理 A.4.4.4 个人信息转移 A.4.4.5 向他人提供个人信息 A.4.4.6 自动化决策 A.4.4.7 个人信息公开
	e) 敏感个人信息处理	A.4.5.1 通用规则 A.4.5.2 生物特征识别信息安全
8.6 个人信息保护	f) 个人信息主体权利	A.4.6.1 个人信息的查阅、复制、可携带 A.4.6.2 个人信息的更正、补充 A.4.6.3 个人信息的删除 A.4.6.4 其他个人信息权利
	g) 个人信息安全义务	A.4.7.1 个人信息保护措施 A.4.7.2 个人信息保护负责人 A.4.7.3 个人信息保护影响评估 A.4.7.4 个人信息安全应急
	h) 个人信息投诉举报	A.4.8 个人信息投诉举报
	i) 大型网络平台个人信息保护	A.4.9 大型网络平台个人信息保护

看见方能治理：数据流转安全态势感知

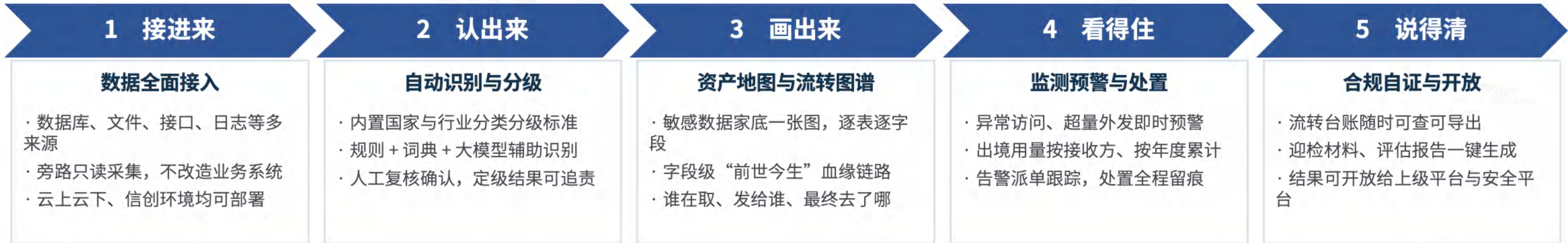
看清每一份敏感数据的来龙去脉 —— 贯通事前预防 · 事中管控 · 事后追溯

家底说不清
哪些系统、哪些字段里躺着敏感数据和重要数据？人工填表申报，报上去的与实际对不上。

流动看不见
数据被谁取走、发给了谁、有没有流到境外，通常事后才知道，甚至一直不知道。

承诺难自证
监管检查、上级审计、客户问询时，“数据未滥用、未出域”拿不出可核查的证据链。

平台做什么 | 五步走通 接进来 → 认出来 → 画出来 → 看得住 → 说得清



一个平台，横跨全周期 同一份资产地图与流转图谱，支撑三个阶段的不同决策

事前 · 防患 · 敏感数据、重要数据自动盘点并定级 · 圈出高风险外发出口与过度授权账号 · 加密、脱敏、访问管控用在刀刃上 回答：我有什么？哪里最危险？	事中 · 管住 · 流转、外发、出境全过程持续监测 · 出境用量逼近年度阈值提前预警 · 与加密网关等管控手段协同处置 回答：此刻在发生什么？要不要拦？	事后 · 说清 · 任一数据可回溯来源与全部去向 · 泄露或变更的影响面一键分析 · 台账与报告支撑检查、通报与问责 回答：谁碰过？影响多大？怎么证明？
--	--	--

看得见，才敢放开用：先把家底和流动摸清楚，保护措施才有落点，合规承诺才有证据。数据安全法 · 个人信息保护法 · 分类分级国标

功能展示： 现在怎么样 / 数据怎么流 / 合规迎检怎么办

图为演示环境，数据均为样例

① 态势大屏 现在怎么样



- 敏感流转、出境流转、未处置告警、综合风险指数，一屏可见，不用四处翻报表
- 全域流转态势实时铺开：红色链路即数据出境，橙色即高敏级流转，点开可下钻到事件
- 血缘热点按「影响面」排序，直接回答「动这张表会波及多少下游、跨几次境」

② 流转图谱 数据怎么流



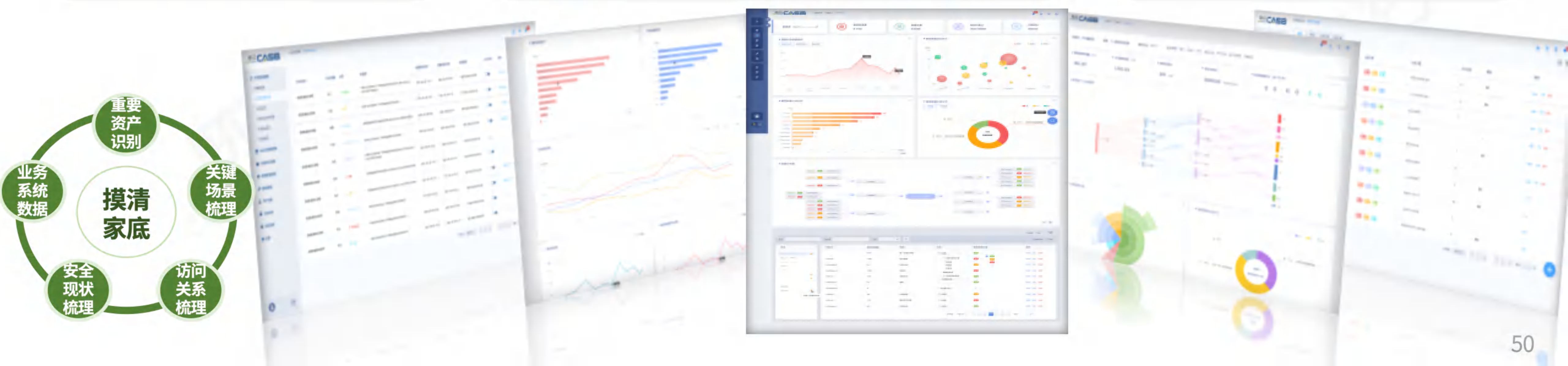
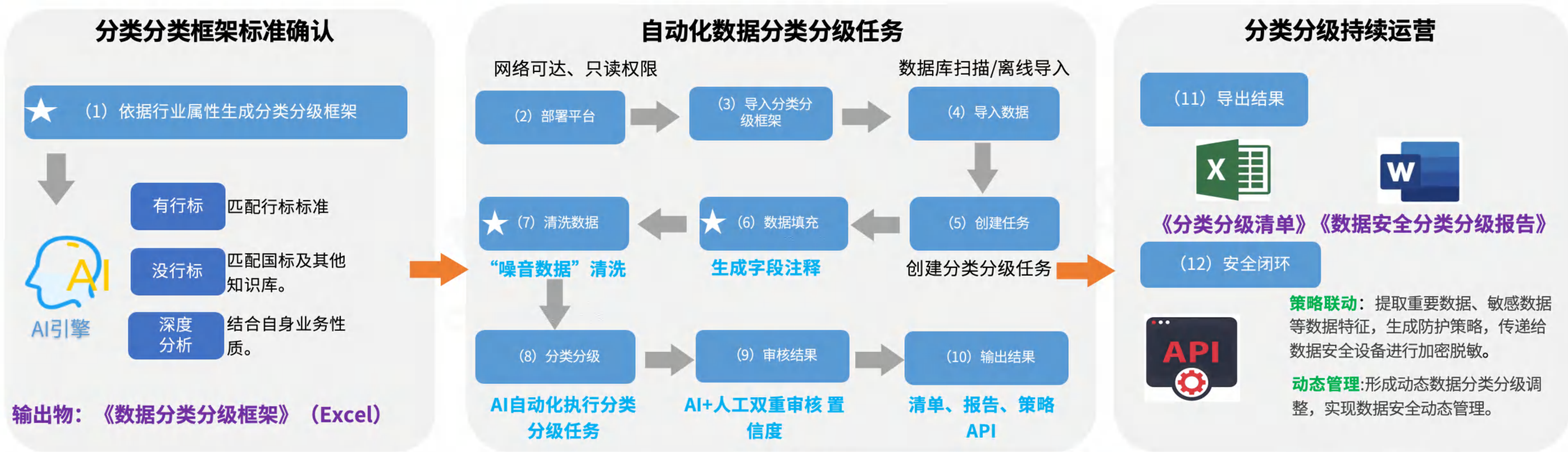
谁在取、发给谁、最终去了哪 —— 逐条链路可追可导出

③ 合规迎检 合规迎检怎么办



检查项逐条锚定法规条款，不达标可下钻证据、一键转整改

分类分级：结合法规要求，基于AI技术落地





对多种常见文件格式通过内容的提取和分析，采用高效识别语言算法自动判别文件是否疑似涉敏，并保存扫描结果，以备进一步的处理



可对识别出的疑似涉敏文件进行加白处理，将其标注为非涉敏文件；或对识别出的疑似涉敏文件从终端上永久性删除



将客户端的扫描结果以数据看板、数据态势，可视化图表的形式呈现，形成可见数据分析及可导出的报表

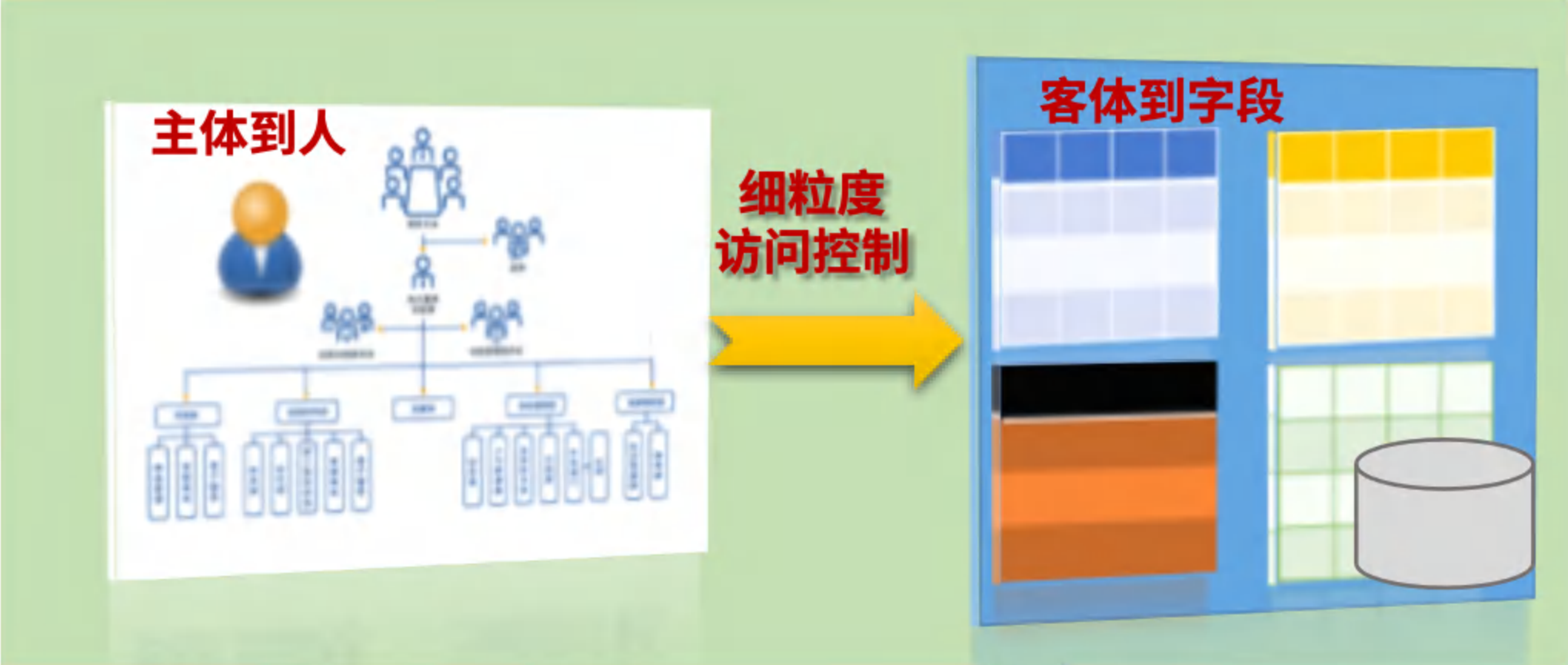


强化后台管理功能，可进行权限配置，自定义扫描范围和规则，覆盖办公环境终端，实现对敏感数据的合规管理

事前防护：最小化敏感数据访问范围



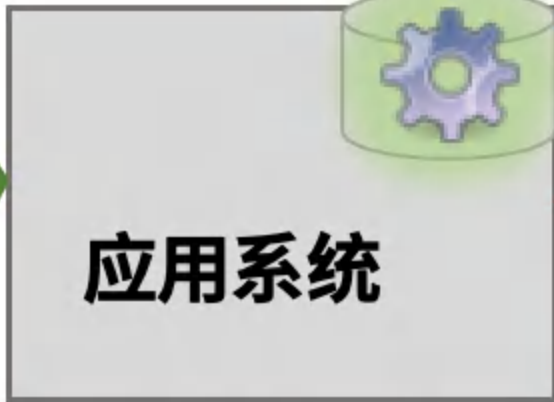
易于实施的“主体到人”，支持与应用身份管理打通



姓名	张珊珊
身份证号	621087335677356125
手机号	13900334901
行程	Pvg-pek-pvg

姓名	张珊珊
身份证号	62108*****56125
手机号	139*****901
行程	p**_***_***

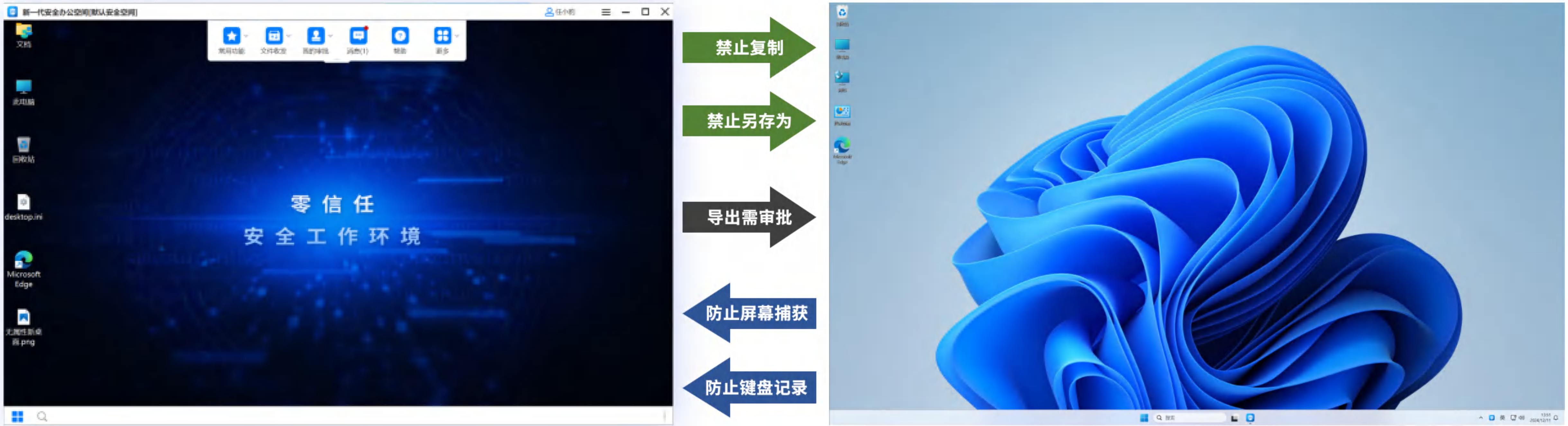
姓名	张珊珊
身份证号	84950084048405048X
手机号	13234540661
行程	3a7v0wCPOHGNabWv



姓名	张珊珊
身份证号	84950084048405048X
手机号	13234540661
行程	3a7v0wCPOHGNabWv

- 可适配应用系统自带身份管理或统一IAM系统，实现结合用户身份的动态脱敏
- 可支持ABAC基于属性的访问控制等细粒度授权

事前防护：终端安全沙箱构建可信工作空间



安全工作空间

普通桌面

采用零信任架构，结合**安全沙箱、轻量可信计算、加密技术**在用户电脑终端创建出一个**受保护的逻辑隔离的安全工作桌面**，与之对应的，把原Windows环境叫做“普通桌面”；“安全桌面”内，所有的文件自动加密，所有的网络访问受控，未经授权任何把数据和文件从里面“拿”出去的行为都会被审计和检查授权，任何试图越权访问的操作都会被拒绝，而用户操作习惯无需改变。

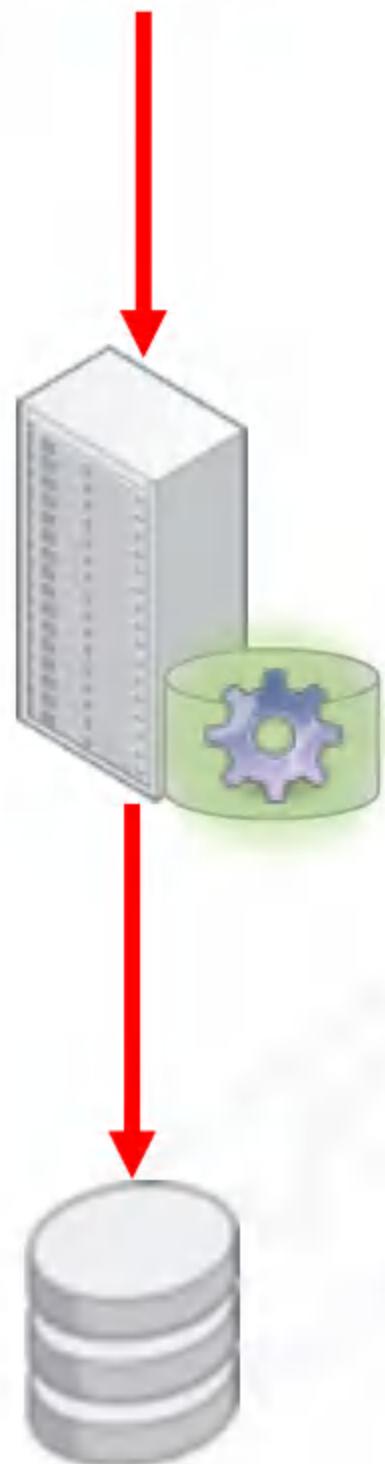
- 注册表隔离
- 文件重定向隔离
- 服务隔离
- 通讯隔离
- 输入输出管控
- 外设接入隔离

事中检测响应：业务人员风险操作的识别处置

存储 → 使用 → 加工 → 传输 → 提供



用户C



应用

数据库

示例

用户C 20161221 23:43:22 下载文件 产品库/
详细设计/ 作动器总体设计图（最终）
行为不符合日常工作时间 行为异常 执行阻断

场景举例

某员工用户C有权使用单位的PLM系统，并能访问到某Z型产品中的设计数据。偶然一次，此人在正常工作时间内，批量下载全套的产品数据，此时数据控制面的实时监控功能捕捉到此异常行为，会立即执行阻断，并将该事件进行记录和报警。

- 免改造控制面对所有流经的业务操作进行监测，结合当前的业务上下文，判断行为是否符合正常的行为模式
- 如果发现是异常操作和不合规操作，则进行告警甚至对操作阻断，以防止恶意操作带来的数据泄露和系统破坏

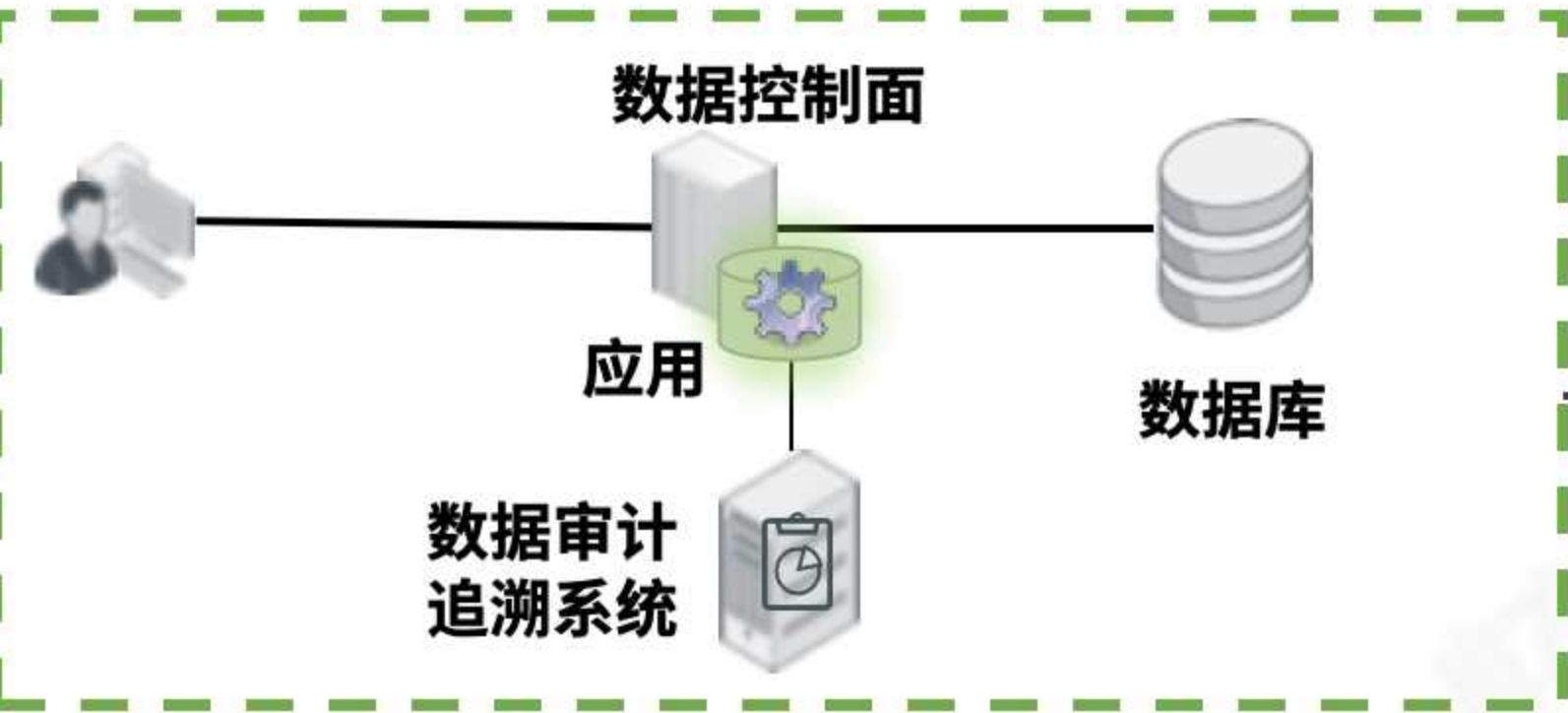
示例

客户信息安全内控增强

- 客服人员每天需要查询高价值的大客户信息，以便提供服务，但存在过量访问的内控风险，而业务应用往往缺失细粒度安全机制
- 通过免改造控制面的安全内控增强，可实现客服人员每天访问大客户信息的阈值，比如100条以内可正常访问，超过100条系统进行告警，超过200条会触发阻断（通过流程审批后才能访问）

事后审计：可定责到人的访问审计追溯

存储 使用 加工 传输 提供



记录所有访问敏感数据的行为

包括账号、时间、IP、会话、操作、对象、耗时、结果等等内容，形成包含终端用户身份的SQL日志。用户可通过日志查询数据的访问情况，在事故发生后精确溯源

操作时间	操作者	行为	操作对象	操作对象ID	操作对象名称	操作对象所有者	状态	是否告警	告警类型	附件属性	客户端IP	日志签名
------	-----	----	------	--------	--------	---------	----	------	------	------	-------	------

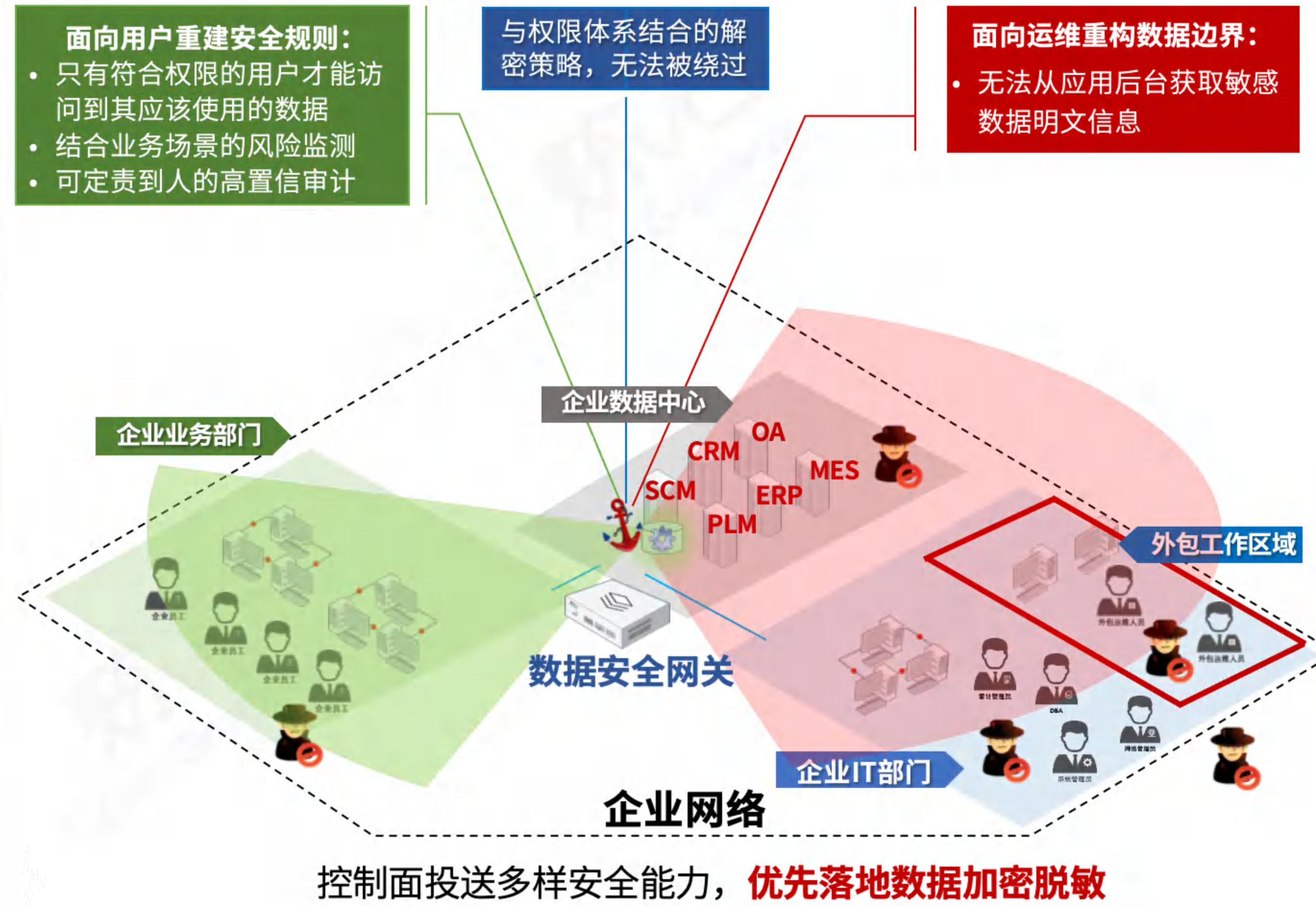
操作时间	操作者	行为	操作对象	操作对象ID	操作对象名称	操作对象所有者	状态	是否告警	告警类型	附件属性	客户端IP	日志签名
2018-02-01 16:13:25	huhu	下载	附件	09_3	09_3	huhu	成功	否	否	279548	172.16.10.98	MEUCIQC/7L...
2018-02-01 16:13:25	huhu	下载	附件	01_1	01_1	huhu	成功	否	否	213008	172.16.10.98	MEUCIApAW...
2018-02-01 16:13:25	huhu	下载	附件	03_3	03_3	huhu	成功	否	否	95760	172.16.10.98	MEUCIGyQ...
2018-02-01 16:13:24	huhu	下载	附件	01_13	01_13	huhu	成功	否	否	96784	172.16.10.98	MEUCI...
2018-02-01 16:13:24	huhu	下载	附件	02_2	02_2	huhu	成功	否	否	376336	172.16.10.98	MEUCIQYk...
2018-02-01 16:13:24	huhu	下载	附件	04_4	04_4	huhu	成功	否	否	163344	172.16.10.98	MEUCIBFG...

事后溯源：隐写水印结合操作上下文实现泄露追溯

在**显示屏幕**、**电子文件**（文档、图片、视频、音频）、**访问页面**（APP页面、业务系统页面）嵌入隐性水印，用于**文件流转路径追踪**、**数据泄露取证**。支持独立软件部署或SDK调用，嵌入的隐性水印不影响用户正常工作。



实战防护：密码安全一体化，免改造打造数据防绕过保护



AOS · AI数据安全

企业落地AI痛点：不敢用 → 不能用 → 用不起来

先敢用，再谈能用、多用 —— 三种“死法”的解药

不敢用



客户的问题

敢不敢让它碰生产/数据？



不解决会怎样

出一次事=一票否决（猝死）



AOS解药

安全合规（国密/等保）

不能用



客户的问题

账号能不能跑得起来？



不解决会怎样

封号 = 当天停摆（急性）



AOS解药

IP 稳定性治理

用不起来



客户的问题

人愿不愿用？旧习惯改没改？



不解决会怎样

野马没人骑 = 温水煮青蛙（慢性）



AOS解药

harness 驾驭工程



安全是敢用的前提

敢用：让企业敢放手让 AI 碰生产和数据

政企不敢用 AI，不是不想，是不敢 —— 安全是“敢用”的前提

商业秘密防泄露

- **痛点：**AI 能读你的源码、凭证、客户数据，可能经 LLM 外发或被诱导泄露
- **我们做：**工具/凭证网关治理：凭证零知识替换、数据不外流 + 国密加密
- **知情权：**看见 agent 跟所有服务器交换了什么（含 AI 厂商的隐性遥测）

出事可举证 · 免责留痕

合规红线 / 免责

- **端侧沙箱四切面 guard：**危险命令/文件操作实时拦
 - rm -rf 误删、git push --force 覆盖远端、读 ~/.ssh 私钥外发 —— 当场挡下
- **全程国密审计：**行为可追溯、不可抵赖，出事能举证免责
- **威胁覆盖：**绝大多数的 AI 实际事故是“误导型”（被诱导犯错），AOS 覆盖这条主线

合规地基 | 国密 + 信创 · 等保 / 密评 / 关基
端侧沙箱（端侧处理 · 数据不出域）

数据出境合规

- **痛点：**用 Claude 等境外 AI，对话、代码、客户数据都在往外发
- **我们做：**明文留在端侧，默认只上“元数据 + 哈希”；出口流量切面管控，该拦的拦
- **合规：**对齐数据出境监管，数据主权可保障 —— 私有化部署可全程留在本地

路线对比：一刀切禁用不是答案，管得住才敢用

路线A · 一刀切禁用

全面禁用先进AI + 推自研替代
= 主动放弃头部生产力

路线B · 裸用放任不管

编码 agent 外泄是行业性风险；
开源编码 agent 过滤可被绕过
= 红头文件点名的风险仍存在于核心网段

路线C · AOS管住继续用

升级安全版本,继续用头部 AI
外联全见 · 域名级拦截 · 审计链存证 · 判级脱敏 · 私有化不出域
= 监管要求逐条变成部署清单

监管连续加码：2021 《网络产品安全漏洞管理规定》 >>> 2024 ————— 2025 ————— 2026

建NVDB

数据出境评估常态化

CSTIS通报DeepSeek部署风险

NVDB点名境外AI编程工具

监管坐实了我们一直在管的事 —— 管得住，才敢用；管得细，才用得好

来源 [NVDB 通报](#) · [模型对战榜\(量子位\)](#) · [GitHub 泄露\(The Register\)](#) · [开源 agent 过滤绕过\(THN\)](#)

基于AOS的AI生产力方案，让企业AI 敢用·能用·多用

自下而上四层：安全层让企业敢用，稳定层让服务能用，应用层让全员多用，最终由使用者获得可衡量的产出。



中立执行层不要求 AI 配合，只在它与外界对话的接触面上设置关卡，无需改造现有工具——账号管稳，全员用透，企业才敢放手

案例-政府海量数据：某部委14亿人口核心民生数据保护

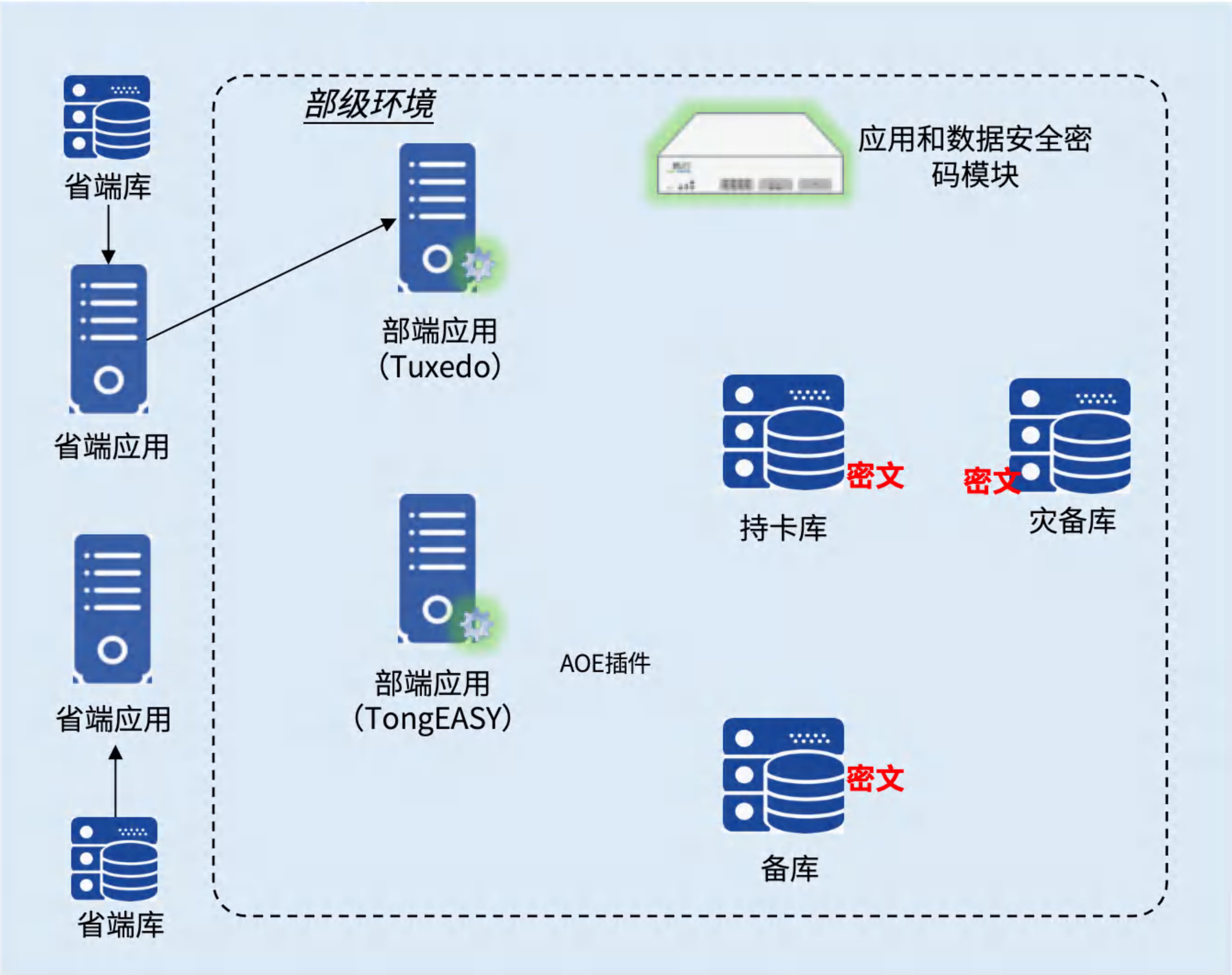
项目背景介绍

客户行业：政府
项目名称：某部委持卡库数据安全建设项目

项目需求：
某部委持卡库拥有全国的某民生卡数据，均为涉及公民隐私的敏感数据，受《网络安全法》、《密码法》、《数据安全法》以及《个人信息保护法》等法律法规的合规要求，需要对持卡库数据采用密码技术进行保护。

- 项目建设目标及实施痛点：**
- 目标1：通过免应用改造模式，实现部级持卡库敏感数据的加密保护。
 - 目标2：基于国密算法实现加密，满足合规要求。
 - 目标3：适配省端数据通过接口上传至部级库的业务场景，通过本方案实时加密后，以密文形式存入数据库中。
 - 痛点1：海量数据，部级库中已有历史数据数百亿条，加密处理，包括集中库，以及灾备库。
 - 痛点2：实时性，方案实施后，生产库和灾备库之间海量数据的实时同步机制不变。

部署和实施方案



案例-金融关键应用：某交易所重要数据保护

项目背景介绍

客户行业：金融

项目名称：某大型证券交易所数据安全项目

项目需求：

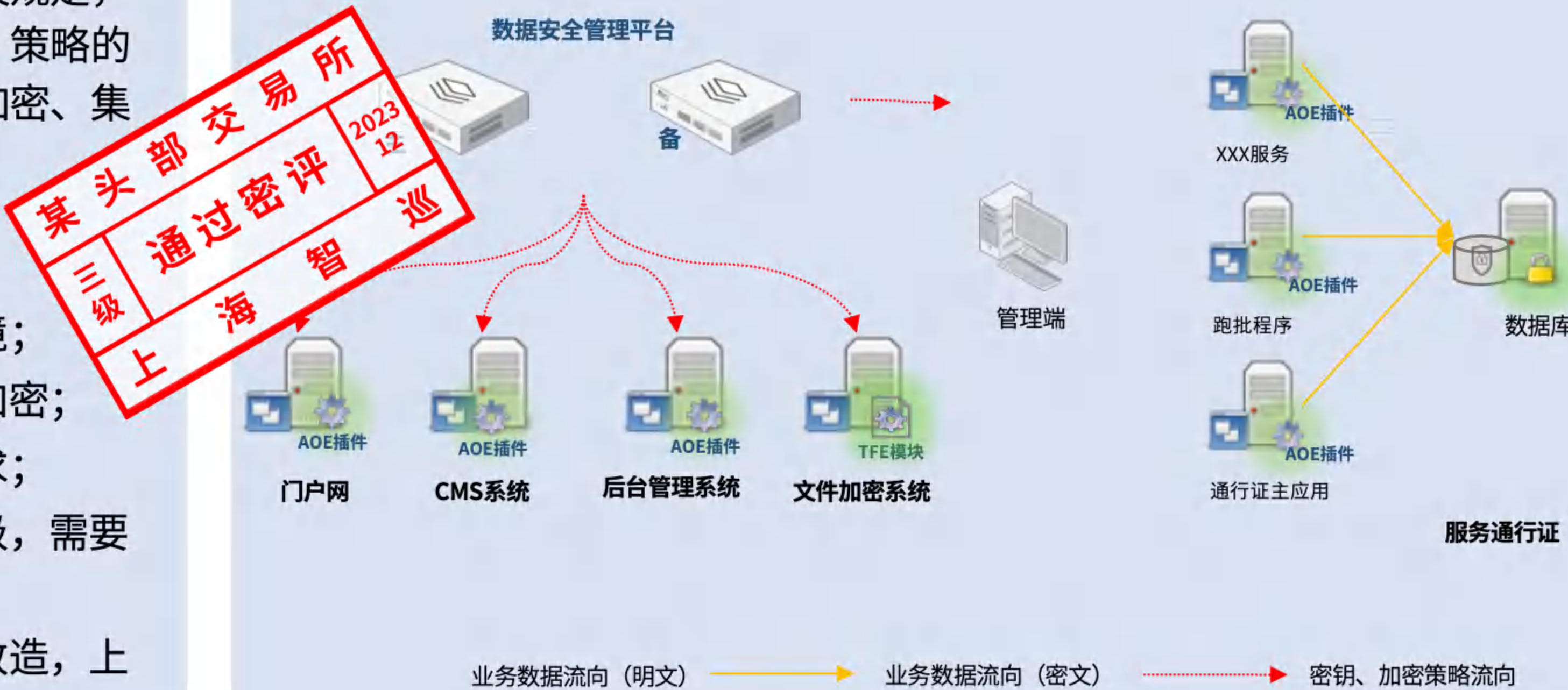
行情经营、网站等相关的多个系统中均存在大量的会员相关信息，如会员账号、股东账号、个人邮箱、真实姓名等敏感数据，数据需要在日常业务中高速流转。依据行业及国家相关规定，需要对敏感数据进行保护。需要实现密钥的统一管理、策略的统一管理、加解密元数据的统一管理，实现“分布式加密、集中式管控”的保护体系。

项目实施痛点：

- 生产应用使用信创环境，需要完全适配当前信创环境；
- 需要使用国密算法，对系统中存放的敏感数据进行加密；
- 需要通过国家密码管理部门核准，符合密评审核要求；
- 应用存在批量数据导入场景，单次导入数据千万量级，需要在执行批量导入时加密数据不影响系统使用；
- 新增解决方案在部署过程中无需应用系统进行代码改造，上线变更需要操作快，周期短；

项目建设方案

在数据中心机房部署两套数据安全平台，平台旁路部署，使用高可用部署模式。在应用侧部署免改造加密插件实现对业务系统中结构化(AOE 插件)和非结构化(TFE 插件)数据的存储加密，提供有效且易于实施的数据安全保护。



案例-运营商BOSS应用：某省移动BOSS数据保护

项目背景介绍

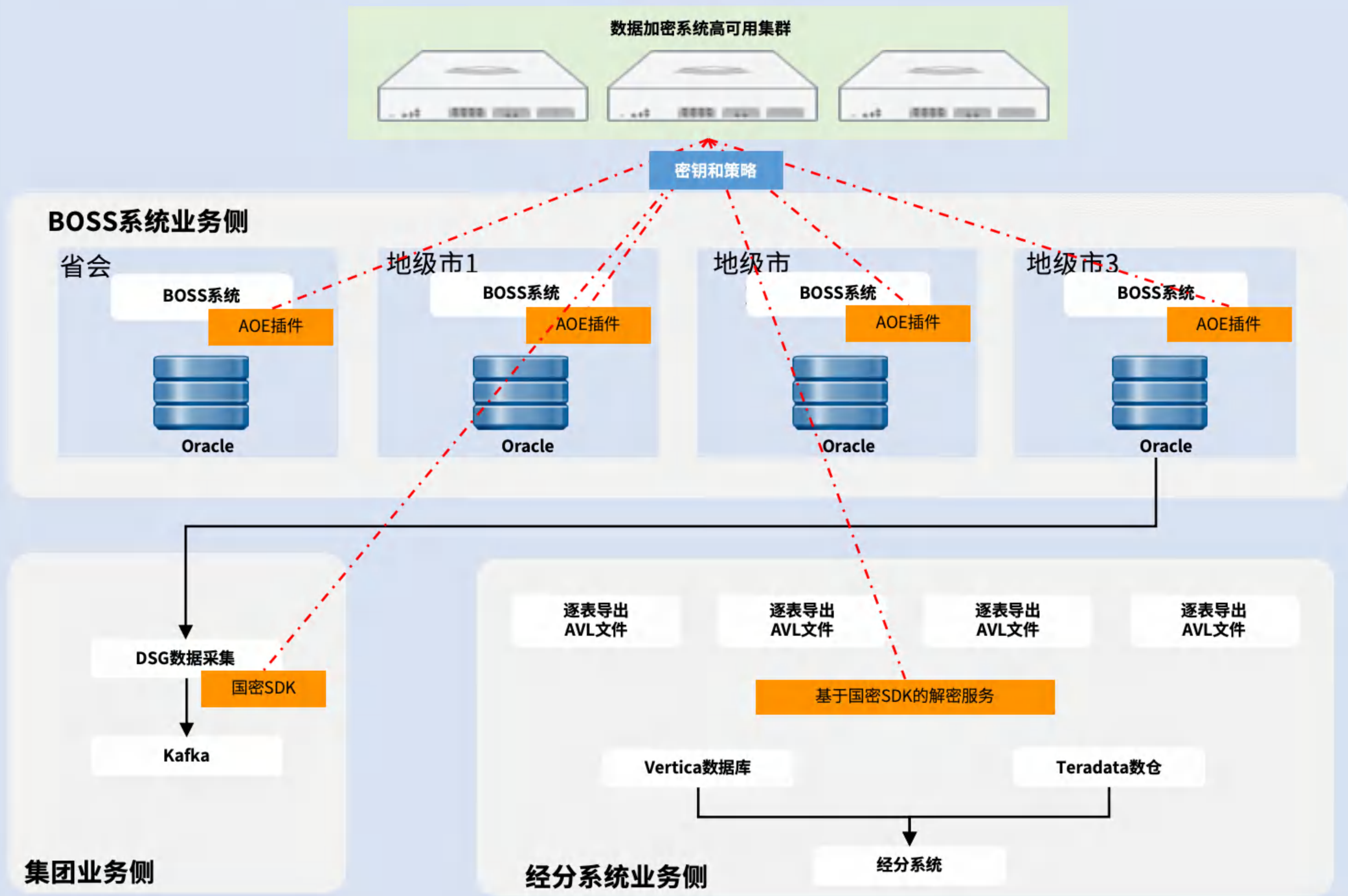
客户行业：运营商
项目名称：BOSS个人信息保护

项目需求：
针对BOSS客户资料表进行加密保护，在实现数据安全的同时，也符合两部委对于商用密码应用的考核要求。该方案已正式通过集团基线评审，获得集团大力支持。

项目实施痛点：
BOSS客户资料表加密，影响所有访问资料表的业务，涉及三方面协同改造（改造范围）：集团全网客户中心、省内CRM/BOSS/经分系统。

解决方案：
BOSS系统：1、应用安装AOE插件，应用写入数据实时加密，应用读取数据实时解密；2、历史数据可通过AOE客户端工具进行全量加密
集团：在省端的数据采集程序集成SDK，调用SDK解密接口，实现采集数据解密
经分：BOSS侧导出的密文AVL文件解密成明文文件，按照原逻辑解析明文文件，把数据写入vertica数据库和TD数据库

项目实施部署方案



案例-工业关键应用：某大型国有车企数据保护

项目背景介绍

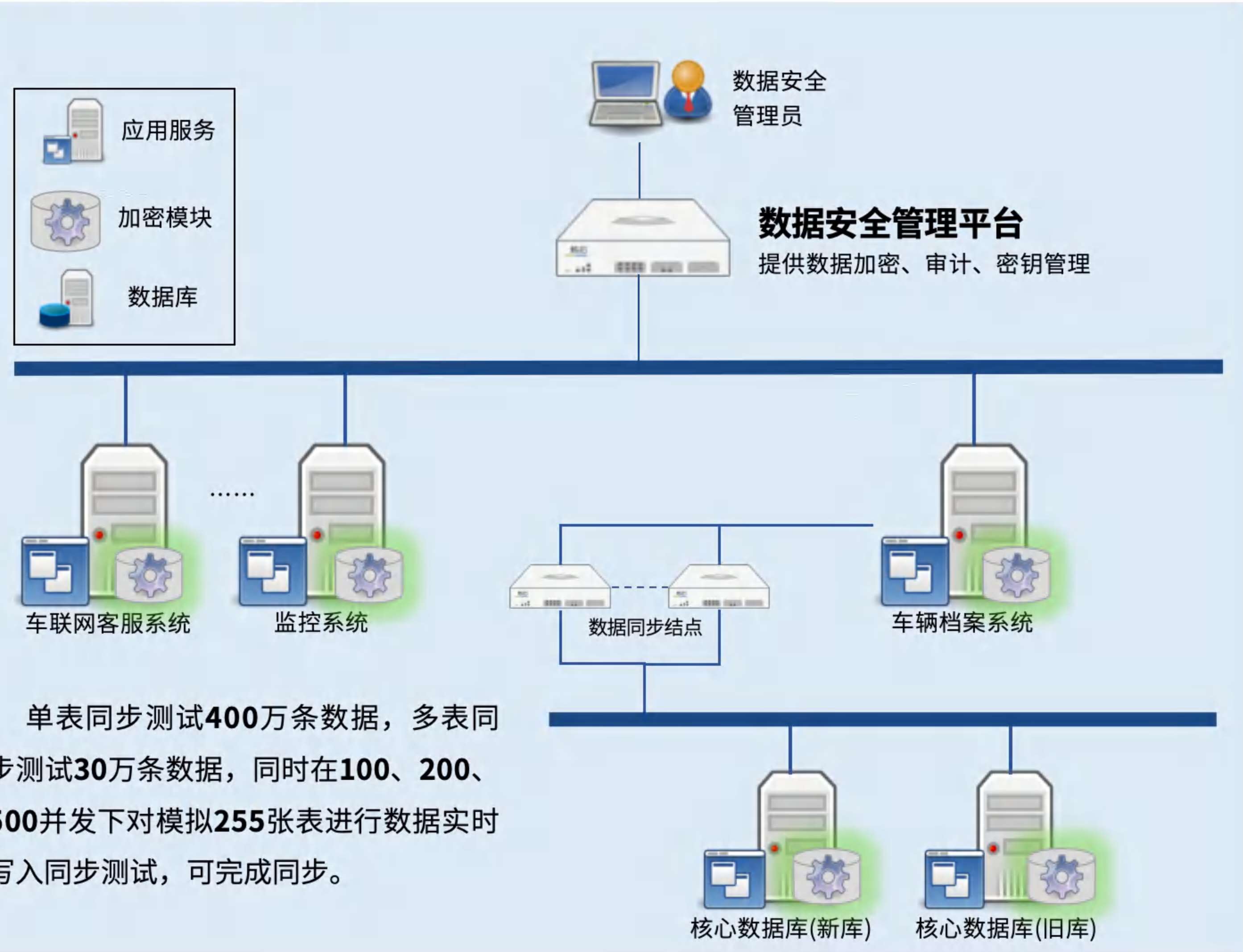
客户行业：制造业
项目名称：某大型现代车企车联网系统数据加密

项目需求：
企业拥有大量车主和车辆的敏感数据，且受国内外GDPR以及国内《网络安全法》、《密码法》等法律法规的合规要求，需要应用密码技术，建设密码平台使得车辆档案系统、车联网客服系统、监控系统，云上+云下数据得到保护。

项目建设内容：
对车联网carbond库10个高敏感字段1400多万条数据进行国密算法加密，同时不影响上下游业务流向，不影响数据维护和使用。

项目收益：
1、实现核心系统的应用数据加密需求，保护用户隐私数据安全，提升系统安全水平。
2、满足国家关于个人隐私数据保护、商用密码信息系统密码应用规范要求。

部署和实施方案



公司定位：炼石就是数据安全

炼石是以**免改造为特色**的AI数据安全厂商，**国家级专精特新“重点小巨人”**，自研高效投送安全能力的免改造产品，聚焦AI数据安全与国密卡脖子难题破解，帮各行业政企客户打造领先AI数据安全保护体系，敏捷交付密评密改合规。



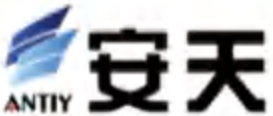
实战型AI数据安全与密改厂商





威胁检测 系统安全

强化炼石系统安全能力



网络安全民企国家队



中国科学院
CHINESE ACADEMY OF SCIENCES

强化炼石安全算法能力



中科院直投基金



腾讯云 腾讯安全
Tencent Security

产品共研深度融合



产业互联网巨头



LH VENTURES
领航新界

拓展市场战略布局

重庆科技成果转化基金

国资



朗玛峰大学
以管理推动科技企业发展的

运营赋能增效发展



科创赛道专业基金



蚂蚁集团
ANT GROUP

蚂蚁生态资本



网安专业基金

炼石免改造数据安全



 www.ciphergateway.com

 010-88459460

 sales@ciphergateway.com