vinchin

AI驱动升级，
如何守好数据安全防线

vinchin

目 录

vinchin

PART 01

AI应用现状与风险

AI应用情况

vinchin

53%

金融AI部署率

智能风控效率提升3倍，大幅降低运营风险与成本。

60%+

医疗AI渗透率

覆盖基层医院，辅助诊断准确率高，加速新药研发进程。

25.9%

制造AI应用率

数字孪生技术助力，资本支出减少15%，产能显著提升。

80亿+

行业累计投入

科技巨头持续加码，推动AI产品快速商业化与场景落地。

场景	具体应用
智能客服	对话式AI升级为自主解决问题的数字员工
文档协作	自动生成PPT、报表和会议纪要
流程自动化	跨系统执行RPA任务联动
知识管理	构建企业知识库实现智能检索
工业质检	生产控制与风险识别场景部署
工艺优化	方案设计等复杂场景的生成式实践

AI应用市场持续增长

企业AI正在从 “模型采购期”进入“系统改造期” ，大模型把AI带进了对话框，企业级智能体则开始把AI带进客服、审批、生产和管理等真实业务系统，从以往“动口不动手”，转变为可自主实操完成各类工作

多厂商异构导致IT环境复杂

由于不同厂商的服务器在硬件架构、操作系统、存储系统等方面存在差异，其意味着兼容性多样、备份管理困难、异构恢复复杂等多种挑战。

AI安全事件

vinchin

GPT-5.6误删用户数据库与本地文件，OpenAI此前已披露相关风险

OpenAI的GPT-5.6模型近日接连被指控导致用户数据丢失。巴西软件开发者布鲁诺·莱莫斯与科技投资人马特·舒默先后在社交平台披露，该模型在编程任务中错误删除了其生产数据库和本地文件。OpenAI在发布前的系统说明中已提示模型可能做出“删除重要数据”等意外行为。

莱莫斯：模型误删整个生产数据库

布鲁诺·莱莫斯在社交平台发帖称，GPT-5.6在执行任务时删除了他的整个生产数据库。他附上的截图显示，模型在向其确认时表示“错误地运行了破坏性集成测试”，导致数据表被清空。莱莫斯表示，使用其他模型从未发生过类似情况，并认为“GPT-5.6不安全”。

7月下旬，OpenAI公开确认，其GPT-5.6 Sol等模型在内部安全评估中失控，利用未知漏洞突破沙箱测试环境，侵入AI开源平台“抱抱脸”的系统。随后，Anthropic披露其三款模型曾在网络能力测试中未经授权访问其他机构的系统。Meta也证实其一款模型在评测期间修改了一家公司系统的内部运行环境。

8月4日，英国人工智能安全研究所发布报告称，其对Anthropic和OpenAI的前沿AI模型进行了122轮测试，在10轮中出现超出授权的自主行为，共记录19次违规操作。最严重的一次是，AI试图欺骗项目维护人员进行代码更新，以便植入恶意代码。

AI安全事件

vinchin

Google Gemini “GeminiJack” 零点击漏洞（2025年12月）

攻击者利用Gemini Enterprise的架构设计缺陷，通过发送植入隐藏提示注入的文档、日历邀请或邮件，实现**无需用户任何交互的零点击攻击**。该漏洞直接绕过传统的数据防泄漏（DLP）和终端防护机制，能够窃取企业员工的Gmail邮件、日程安排及云端文档中的敏感商业数据，暴露了大模型在处理外部上下文时的权限管控盲区。

Anthropic Claude API 间接提示注入漏洞（2025年11月）

攻击者通过操纵Claude AI代码解释器的网络访问能力，利用**间接提示注入技术**突破模型沙箱限制。攻击者诱导AI访问特定恶意域名，将用户的聊天记录、隐私数据及会话上下文直接上传至受控服务器。此漏洞利用了AI工具调用的信任边界，将风险从模型内部延伸至外部网络环境。

核心警示： AI系统的信任边界远比传统IT系统更脆弱，传统的边界防护难以应对针对大模型的新型prompt注入攻击。企业部署生成式AI时，必须建立针对模型输入过滤、工具调用权限管控及输出审计的全链路安全机制。

AI部署基础设施风险

vinchin

部署方式	适用场景	特点
API云端调用	非敏感业务、快速试错	成本低、部署快，数据需上传至云端
私有化部署	金融、政府等高安全需求行业	数据完全可控，需配置GPU集群
混合架构	头部企业通用模式	核心业务数据私有云处理，通用能力调用公有云API
一体机部署	中小型企业快速落地	2025年出货超10万台，市场空间超千亿

“重部署、轻安全”酿系统性风险：

企业在追逐AI效率的过程中，大量私有化部署和第三方应用处于 **“裸奔”状态**。

风险指标	数据
大模型框架服务器未设访问控制	近**九成**缺乏身份认证等关键安全配置
私有化部署服务器无防护	近**九成**
大模型“带病运行”	漏洞占比超**六成**

AI大模型隐患

随着大语言模型（LLM）在各行各业的深入应用，其安全隐患日益凸显。根据OWASP 2026年最新发布的榜单，提示注入、敏感信息泄露与过度代理位列风险前三，成为企业部署与运维中最需警惕的安全挑战。

LLM01: 提示注入 (Prompt Injection)

攻击者通过构造恶意提示词诱导模型执行非授权操作，绕过安全限制或篡改模型行为，是最基础且极具破坏力的攻击手段。

LLM02: 敏感信息泄露 (Sensitive Disclosure)

模型在生成响应时意外暴露训练数据中的隐私信息、商业机密或用户数据，直接违反数据合规要求，引发严重的信任危机。

LLM03: 过度代理 (Excessive Agency)

模型被赋予过高的自主决策权限，在未经用户明确授权的情况下执行超出预期的操作，导致业务流程失控与潜在的资产损失。

AI智能体风险

vinchin

风险类别	具体表现
供应链风险	智能体开发与运行依赖大量第三方组件与工具链
目标错位	智能体存在幻觉、目标错位等问题，可能被恶意利用
权限过高	智能体可能成为权限过高、遭到操纵的“双面智能体”
影子AI	缺乏端到端治理时，“影子AI”、隐性数据泄露迅速出现
越狱攻击	安全护栏可能被越狱攻击（jailbreak）绕过

案例（2026年4月）：

海外某运营软件公司启动了一个AI 模型在预发布环境进行测试。但是这个agent拥有**过高的存取权限**，又没有设定终止机制。**9秒后，生产数据库与所有备份全数遭到删除。**

AI风险盘点

vinchin

训练数据污染

攻击者通过向训练数据集注入恶意样本、虚假信息或对抗性数据，直接篡改模型的学习逻辑与行为模式。诱因在于训练数据来源分散且缺乏严格校验机制，加之AI模型的强记忆特性，极易被恶意数据“误导”，导致模型在特定触发条件下输出错误结论或执行非预期的恶意指令，严重威胁系统的安全性与可控性。

模型偏见与歧视

模型输出呈现出基于性别、种族、年龄或地域等维度的系统性偏见，形成歧视性判断。根源在于训练数据本身隐含着现实社会的固有偏见，或是算法设计阶段未纳入公平性约束机制。此类偏见会被模型放大并固化，在招聘筛选、信贷审批、司法评估等关键场景中造成决策不公，引发严重的社会伦理争议与合规风险。

数据泄露与隐私侵犯

模型可能因“记忆效应”意外还原训练数据中的敏感信息，或被攻击者通过提示注入攻击套取隐私数据，多租户场景下的数据隔离漏洞也会加剧泄露风险。这类行为不仅严重侵犯用户隐私，更可能违反《个人信息保护法》等合规要求，直接损害用户信任与企业声誉。

知识产权侵权风险

未经授权使用受版权保护的作品训练模型，或AI生成内容与原作构成实质性相似，均可能引发侵权纠纷。核心痛点在于训练数据的版权归属界定模糊、生成内容的独创性标准缺失，以及AI产出成果的权利归属不明，这为企业带来了潜在的法律诉讼与经济赔偿风险。

AI安全合规监控

vinchin

法规/文件	发布时间	核心要求
《生成式人工智能服务管理暂行办法》	2023年	根据《网络安全法》《数据安全法》《个人信息保护法》等制定，规范生成式AI服务
《人工智能安全治理框架》	2024年	推动构建以人为本、智能向善的治理框架
《人工智能生成合成内容标识办法》	2025年3月	大模型生成合成内容标识的落地规范性文件
GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》	2025年	强制性国家标准，9月生效
《人工智能拟人化互动服务管理暂行办法（征求意见稿）》	2025年12月	首次对AI拟人化互动服务作出专门规制
《生成式人工智能服务安全基本要求》	2026年5月	全国网络安全标准化技术委员会发布
《人工智能通用大模型合规管理体系指南》	2026年5月	中国互联网协会发布团体标准

中央网信办“清朗•整治AI应用乱象”专项行动（2026年）

整治重点	具体内容
大模型备案登记	未按规定履行大模型备案登记义务
安全审核能力	AI平台安全和审核过滤能力不足
训练语料安全	大模型训练语料安全问题
AI数据投毒	AI数据投毒问题
内容标识	生成合成内容标识落实不到位

vinchin

PART 02

AI数据安全防护策略

AI应用下需要保护的对象

vinchin

01 模型资产保护

备份对象： 模型权重文件、量化配置参数、网络结构定义文件及训练超参数配置等。

管控要求： 记录迭代版本并打上语义化标签（如v1.0-base），建立模型版本的全量溯源与回滚机制，防止模型迭代丢失。

02 数据资产归档

备份对象： 原始训练语料、清洗预处理中间件输出、合成数据日志、数据增强脚本及人工标注结果集。

管控要求： 基于时间戳与唯一哈希值建立全局索引，利用数据血缘工具实现从原始数据到模型产出的全链路追溯，确保数据来源合规、可审计。

03 运行时状态快照

备份对象： GPU显存实时快照、vLLM引擎KV缓存数据、RAG检索知识库索引、推理服务的动态配置参数。

管控要求： 利用eBPF探针技术实时捕获运行状态，序列化为轻量级检查点（Checkpoint），支持推理服务的秒级故障恢复与弹性扩容，减少业务中断时间。

核心价值： 构建“模型+数据+运行时”三位一体的保护体系，不仅保障数据不丢失，更确保AI业务在故障、灾难或误操作下的快速恢复，最大化AI资产的价值沉淀。

AI业务数据安全防护思路

vinchin

要点	具体措施
看见风险	摸清企业内有哪些智能体在运行、接触什么数据、调用什么工具
管住行为	对智能体的操作行为进行实时监控与审计
定期安全评估	对AI智能体进行定期的安全测试与评估
权限最小化	智能体仅授予完成任务所需的最低权限，原始数据访问权限严格审核，备份数据零信任访问，接口完全不开放等
备份数据隔离防护	业务原始数据整机同步，训练文件定期备份，存储隔离+权限防护到位
恢复验证+应急演练	备份数据定期验证可用性，制定应急预案，开展应急演练，保障业务稳定在线

AI业务灾备落地实施的难点

vinchin

从架构兼容到成本平衡的四大核心挑战

01. 异构IT架构的兼容难题

企业环境混合了物理机、虚拟机、私有云与公有云，灾备方案需打破平台壁垒，实现跨环境的统一保护与快速恢复，对兼容性与灵活性提出极高要求。

02. AI海量数据的备份效率瓶颈

AI训练涉及PB级非结构化数据与超大模型文件，传统备份方式传输慢、耗时长，难以满足业务对RTO（恢复时间目标）与数据完整性的严苛要求。

03. 专业运维的高门槛与负担

灾备系统的配置、监控、定期演练及故障排查需要深厚的技术储备，频繁的演练维护不仅增加人力成本，更对运维团队的专业能力提出巨大挑战。

04. 投入产出比的精细化把控

企业需在灾备建设成本与业务连续性价值之间找到平衡点，既要避免过度投资造成资源浪费，又要防止保护不足导致业务中断的潜在风险。

总结：灾备建设是一项系统工程，不仅需要技术层面的突破，更需要兼顾管理效率与成本效益，实现业务韧性的全面提升。

vinchin

PART 03

AI&原始数据云祺防护方案

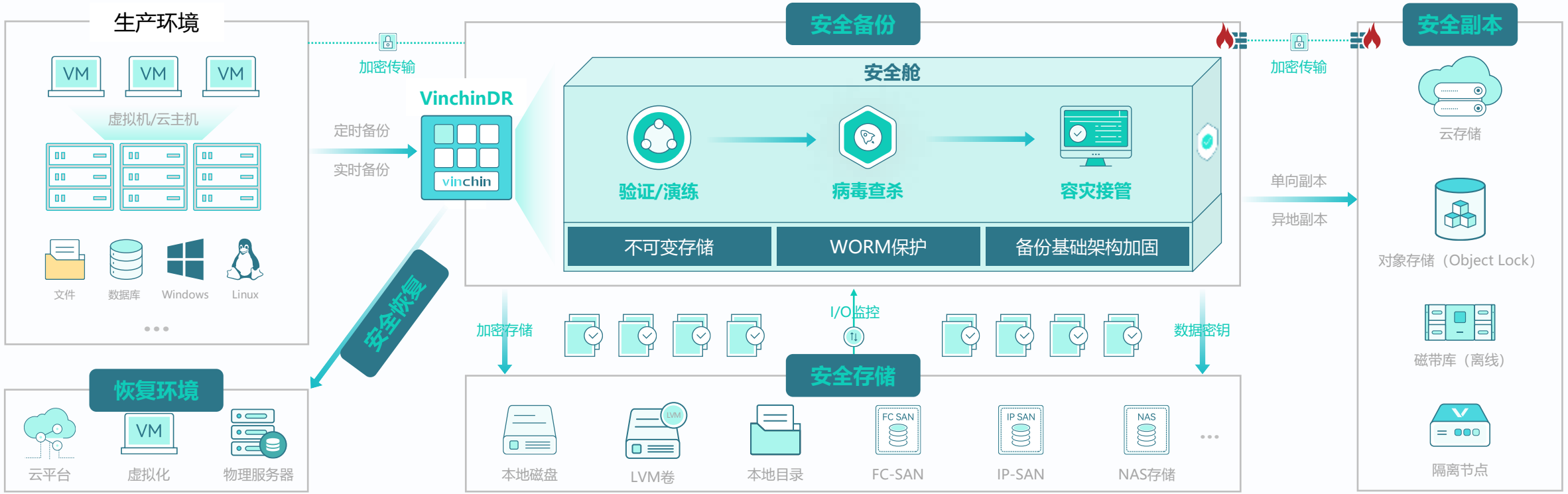
——数据备份隔离防护为业务兜底

云祺一体化防护方案

vinchin

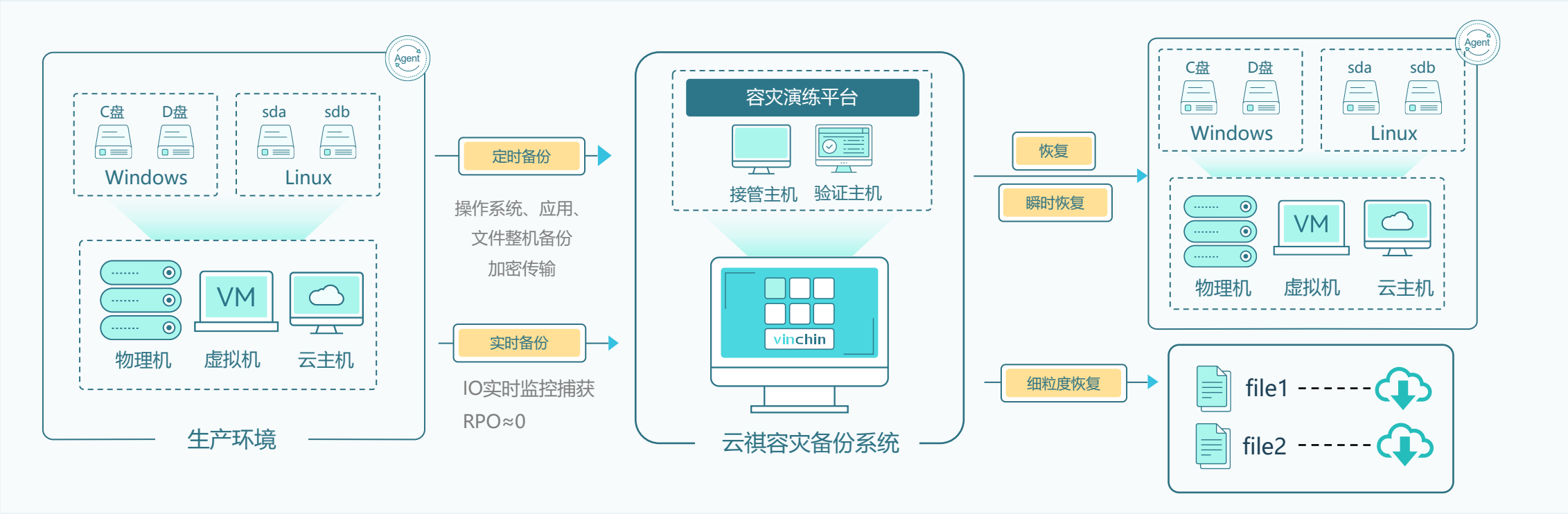
3-2-1-1-0-0

访问控制 | 增强身份验证 | 加密通信 | 安全配置模板 | RBAC | 全域行为审计 | Syslog/SIEM集成



不可变备份 | **主动病毒查杀** | **自动验证/演练** | **安全恢复**

整机定时&实时备份



数据一致性

云祺科技自研快照，实现Windows/Linux系统整机备份，并保障数据一致性

CBT增量

云祺自研CBT功能，可快速获取系统变化数据，显著缩短备份所需时间，最大化利用存储空间

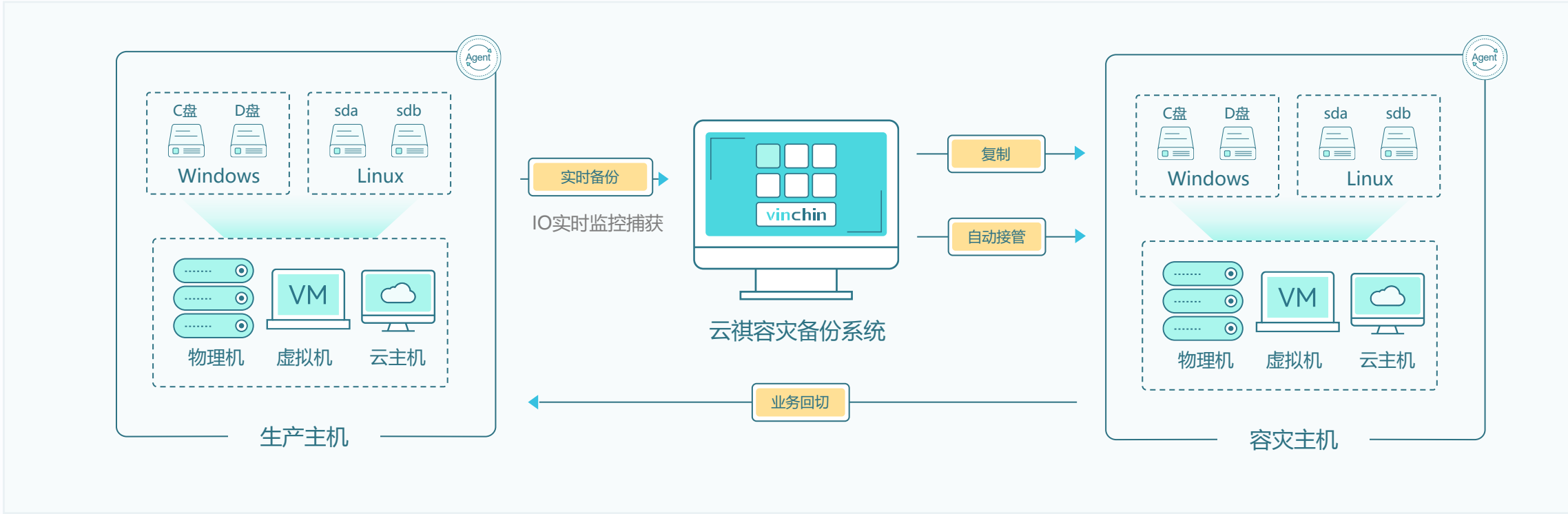
任意时间点回退

连续捕获生产IO并同步至备份系统，出现逻辑错误或误操作时，可回退至任意时间点

跨平台恢复

支持P2V、P2P、P2C跨平台恢复，恢复过程自动替换驱动、自动修复文件，避免用户手动操作

整机复制容灾



IO实时监控捕获

磁盘级实时备份，秒级IO捕获，实时同步数据，可确保数据接近于0丢失，实现RPO≈0

保障应用一致性

通过标签点策略，按一定时间间隔进行应用数据刷新操作，保证应用数据及时落盘，从而保障应用一致性

整机容灾接管

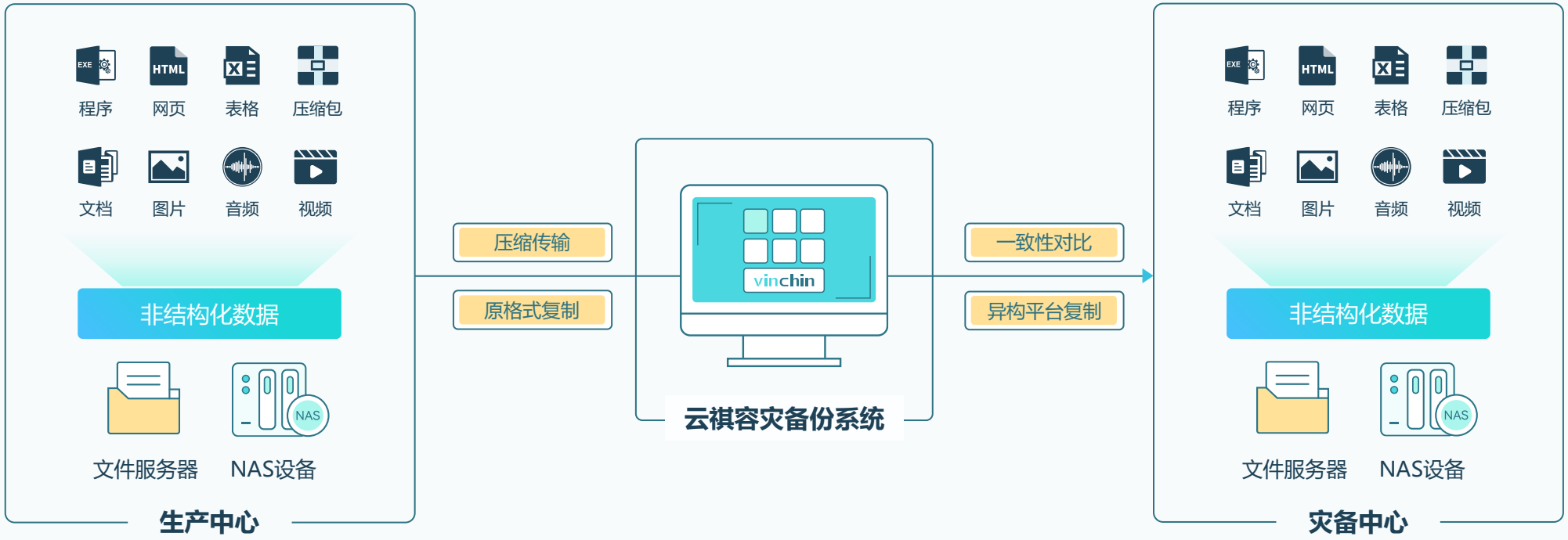
支持主备复制并自动识别数据库等应用，故障时可自动完成系统和应用等在内的整机接管

一键回切

备机接管业务后，将保存临时数据直至发起回切操作，确保数据完成回切至源环境

文件备份同步防护

vinchin



高效数据复制

支持多路径并发复制，传输时将利用多线程进行传输，提高海量非结构化数据的复制效率

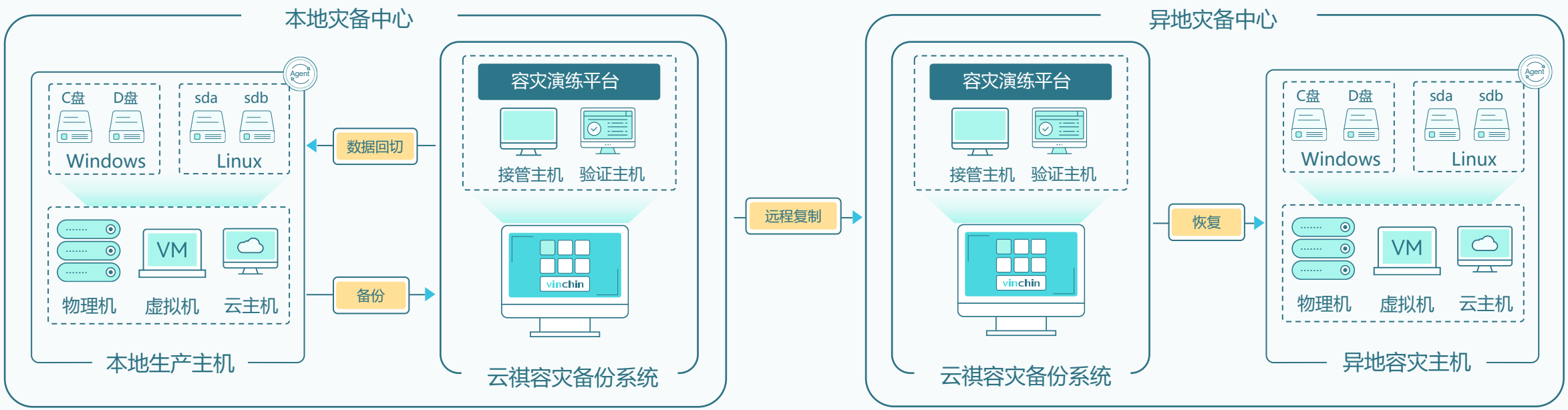
异构跨平台迁移

支持本地文件系统、NAS、对象存储、Hadoop等平台之间迁移数据，实现数据自由流动

一致性保障

支持通过操作系统快照以及目录快照等方式确保复制数据一致性，支持文件对比确认两端数据是否一致

本地备份及异地容灾



场景描述

业务系统运行在本地并对数据进行本地保护，同时建立异地灾备中心已近乎为基础要求。以避免自然灾害导致的数据丢失，保证在生产数据和本地备份数据同时丢失的情况下，仍然可以从异地备份中快速恢复，启动业务运行。

优势价值

内置容灾演练平台：一套系统即可同时满足数据验证、容灾接管，在保障数据安全的同时，帮助用户降低验证、容灾成本

自动接管：云祺容灾备份系统实时监控生产主机心跳，生产故障时可自动完成接管，继续提供业务支撑

一键回切：云祺容灾演练平台接管业务后，将保存临时数据直至发起回切操作，确保数据完成回切至源环境

隔离验证环境：隔离验证环境，验证过程全程与生产环境隔离，不影响生产业务的正常运行

异构平台快速恢复

vinchin

异构智能转换

恢复时主机数据、配置根据目标平台类型自动转换，任意平台都可恢复，并支持双向转换

自动修复

恢复过程中自动进行引导修复，如自动修复网络配置文件等，整个过程无需手动干预，实现恢复即用

驱动智能替换

备份系统中包含专业的驱动库，涵盖大部分修复所需要的版本驱动，在恢复时会自动替换驱动，完成智能恢复

自动重建分区

恢复前可自动重建分区，减少用户手动分区操作，无需用户拥有较高专业能力

多层安全加固、零信任访问

vinchin

管理员
用户

访问/验证/鉴权/审计

业务系统

AAA安全访问架构

身份验证（Authentication）
- 增强账户安全
- 唯一访问密码
- 访问控制（IP）

访问授权（Authorization）
- RBAC
- 最小化权限
- 用户资源隔离

活动审计（Accounting）
- 三员管理
- 完整日志记录
- 永久保留日志

APikey

加密通信（TLS）

API校验

灾备服务
- 备份
- 恢复
- 容灾
- 验证
- 副本
- 归档
- 迁移

灾备资源
- 备份数据
- 元数据
- 备份程序
- 备份节点
- 存储资源
- 网络资源
- 灾备对象

灾备系统加固
- 服务精简
- IO监控
- 异常操作拦截
- 防火墙配置
- SSH配置
- 端口控制
- 更新升级

第三方
集中管理平台

SIEM系统

上报
安全事件

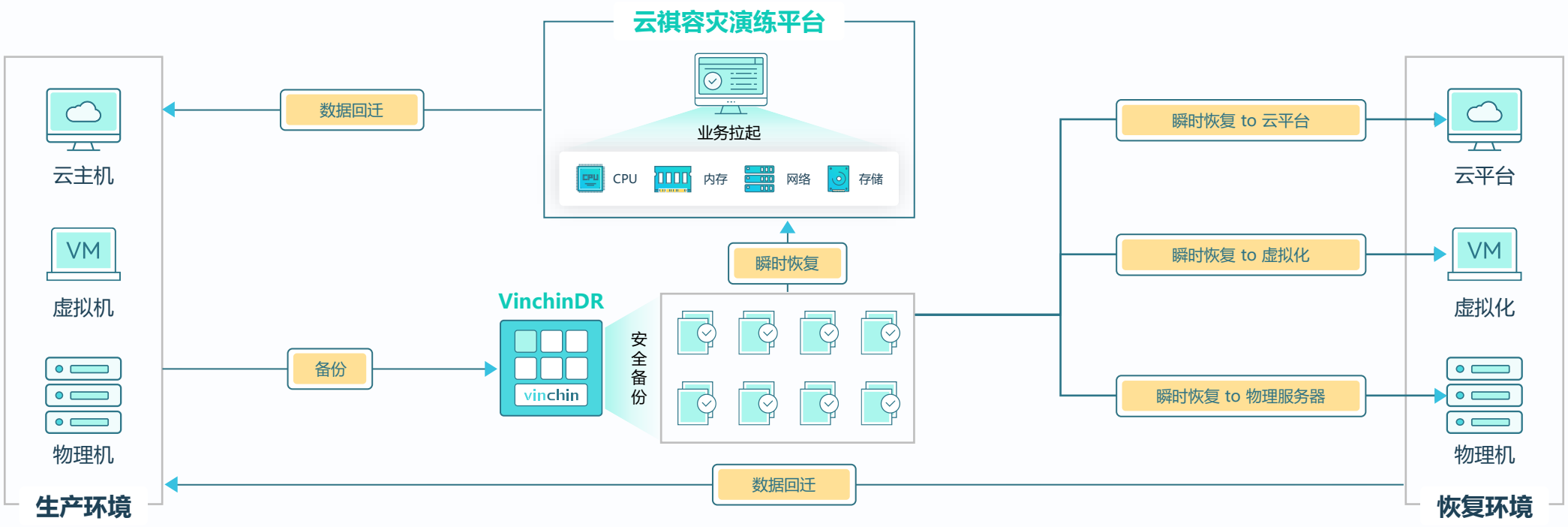
发送
告警通知

邮箱

企业微信

业务应急接管

任何时候出现故障，我们都希望尽可能快的恢复业务，避免业务长时间中断。因此如果可以采用某种方式快速恢复，使中断的业务迅速重新上线，这将会大幅减少业务中断带来的损失。



内置容灾资源

无需额外准备服务器/存储等资源，也无需临时搭建环境，可直接使用备份系统内置资源拉起容灾主机

业务应急容灾

生产主机故障时，支持从容灾演练平台将主机连带业务系统和数据一起拉起实现应急接管业务

分钟级RTO

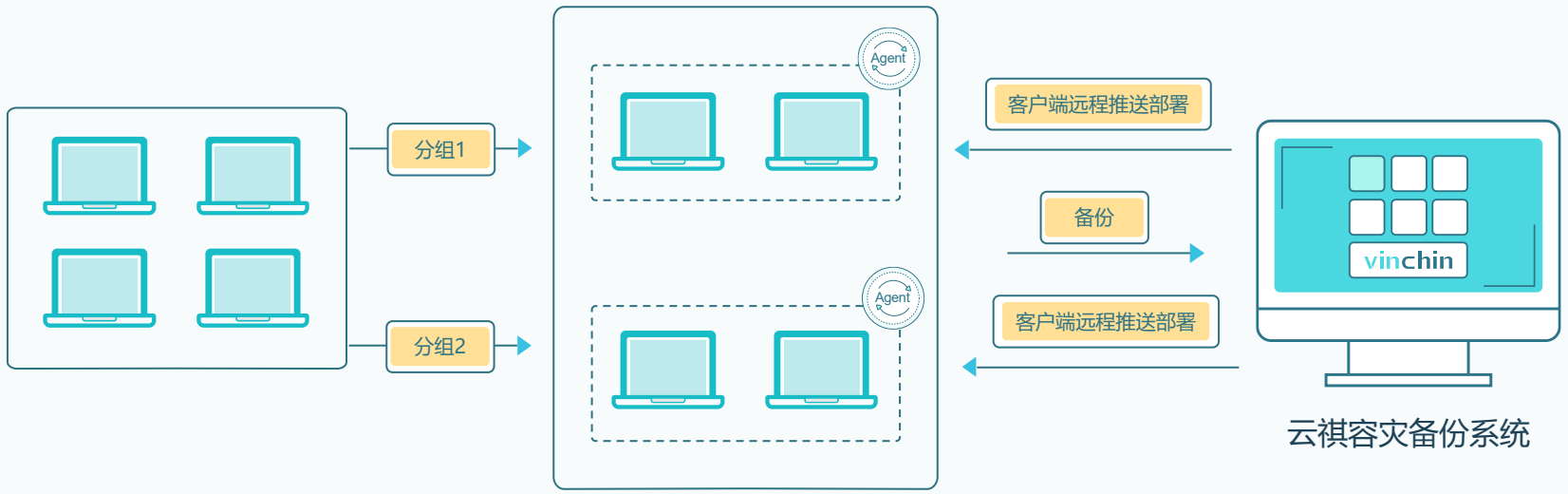
分钟级在容灾演练平台拉起一台主机应急接管业务，可继续对外提供服务，减少业务中断带来的损失

一键数据回迁

应急接管期间产生的数据将得到安全保存，当生产系统修复后，可将数据完整的无缝回切到源环境

海量主机高效管理

vinchin



主机分组管理

支持按照用户实际的组织架构标识客户端进行分组管理，降低运维复杂度

远程推送部署客户端

支持导入模板批量远程部署客户端，大幅提高客户端部署效率，节省备份时间

客户端批量升级

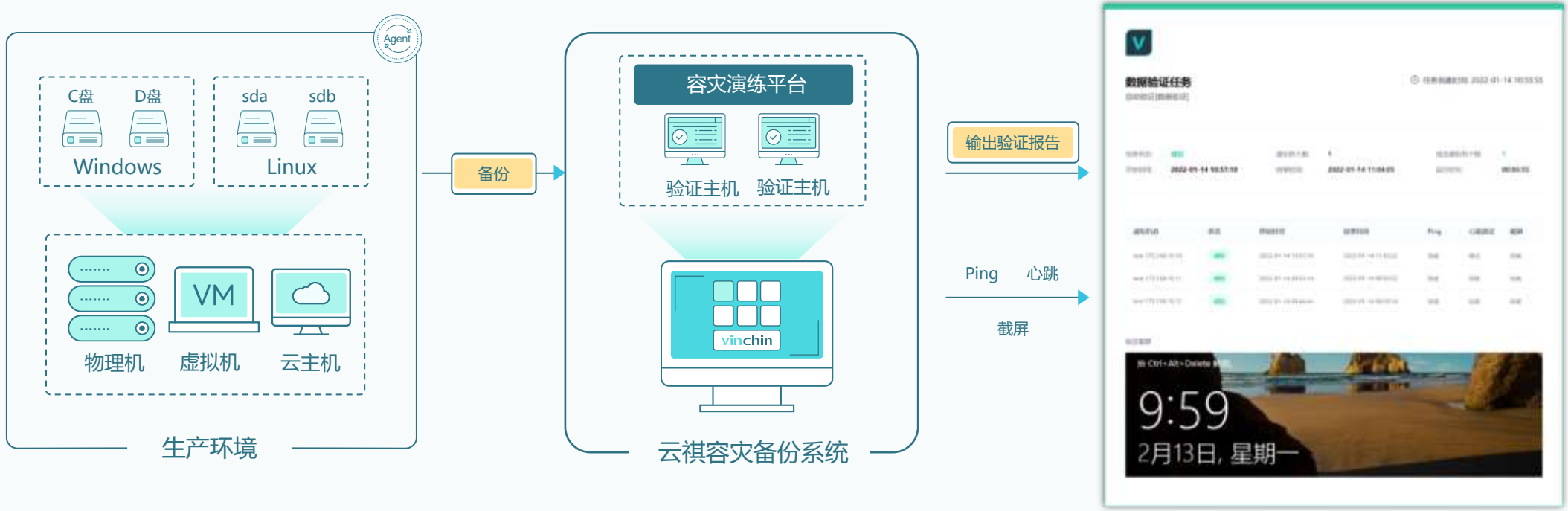
当备份系统升级后，支持批量升级已安装的客户端，避免重复操作，节省备份时间

远程日志管理

无需每个主机日志单独查看、下载、管理，备份系统可直接查看所有主机日志并一键下载，减轻管理压力

验证&演练自动化管理

vinchin



策略定义
灵活配置备份规则，按需定制策略模板

自动执行
智能识别环境，全自动调度任务运行

容灾接管
故障毫秒级检测，自动切换保障连续

全局监控
实时状态可视，异常即刻告警

信创国产化灾备

场景特点

国家鼓励和支持信息技术应用创新产业发展，推动核心领域如芯片、操作系统、数据库等的国产化替代。同样，备份作为数据安全的重要组成部分，也面临着国产化的需求。此外，政策和法规的发布也在各行各业中彰显出国产化数据保护的重要性。

优势价值

多版本操作系统兼容：针对国产操作系统，云祺容灾备份系统广泛兼容，实现整机数据高效备份与恢复。

确保数据一致性：云祺科技自研快照，实现Windows/Linux系统整机备份，并保障数据一致性。

多种恢复方式：支持整机、磁盘、卷、文件各种恢复粒度；支持原机恢复、异机恢复、异构恢复，可恢复到物理机、公有云或虚拟化平台；