

信创筑基 智算未来

2025金融业新一代数据中心发展论坛

2025.09.12 | 贵阳

主办单位：



金科创新社

Fintech Innovation in China



基于AI Agent的金融云平台 全场景运维决策机制研究

高坤 | 国泰海通证券

01、国泰海通云数据中心定位

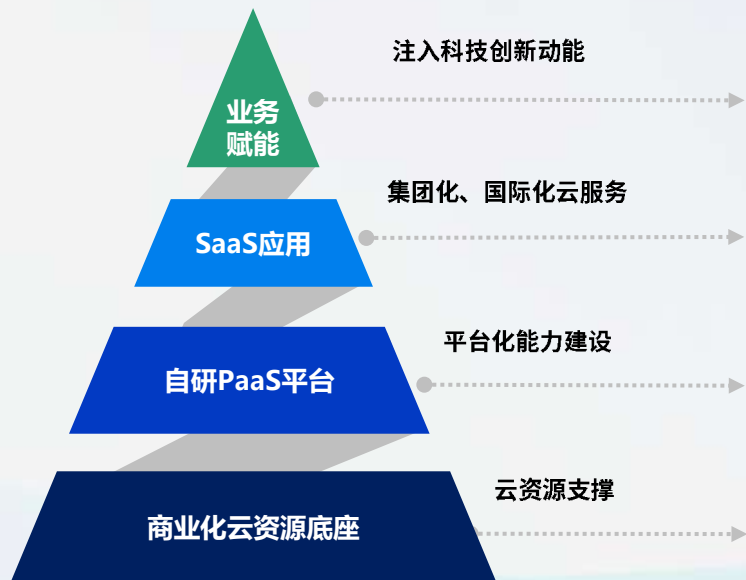


秉承公司“人人数字化、处处数字化”理念，构建全面领先的**自主可控金融云平台**，支撑公司业务**大规模、高速发展**。

承载金融科技落地和数字化转型，实现非交易类业务**100%云化**，为构建灵活共享的业务中台、融合智能的数据中台、协同高效的云上办公，以及金融场景创新等场景提供**稳定、高效、易用**的云底座。

实现从传统业务平移上云到全面云原生**化**，进行分布式应用、微服务、业务智能化

02、构建资源统一、高效协同的科技服务云平台



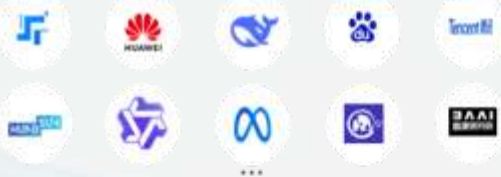
03、ALL in AI构筑1+N的应用策略

构筑“1+N” AI大模型底座

“N”类业务场景落地应用



“1”个行业共享大模型



领先通用大模型

金融行业大模型

场景模型

探索算力创新合作方式

全栈自主可控 算力资源池

满足业务场景

- 更灵活
灵活分配GPU资源，支持不同AI场景
- 更高效
全面提升GPU利用率，发挥最大效能
- 更稳定
细粒度多租户管理，实现资源隔离
- 更健壮
兼容主流开源模型和商用模型

国产芯片替代

创新研究

- “国芯证道”解决方案
- 异构算力的融合
- 适配：硬件性能、软件生态
- 产、学、研国产生态构建

算力创新合作

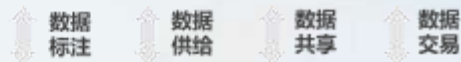
可信数据沙盒+
分时算力租赁

- 可靠合作伙伴
- 可信数据沙盒
- 分时算力租赁
- 示范效应

强化数据治理和语料基础

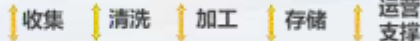
开放证券

战略合作伙伴 行业数据服务商 金融同业



语料中台

数据 >> 高质量语料库 >> 模型训练



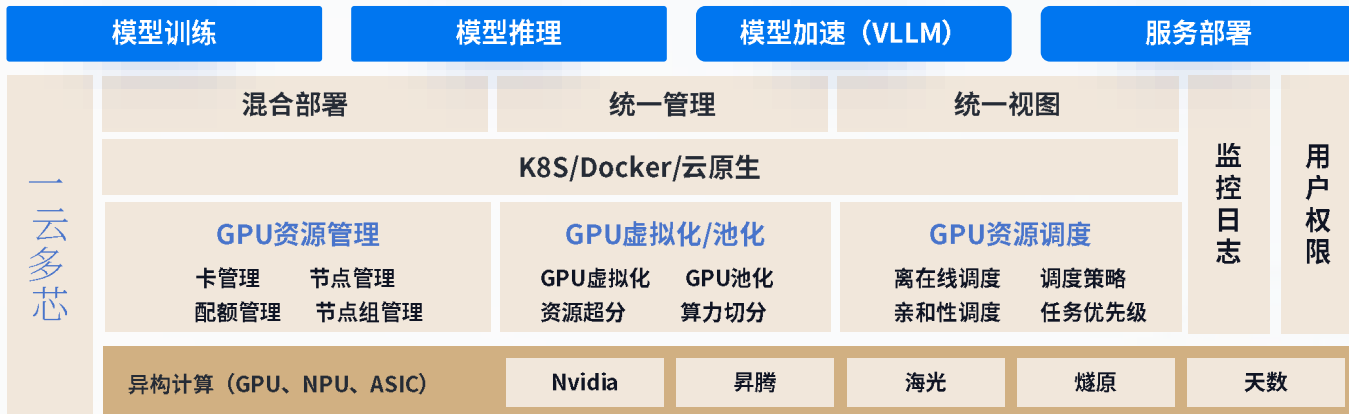
智慧化数据中台

国泰海通采用“通用+垂直”融合的1+N应用策略，基于大模型核心的算法、算力和数据三要素，构建行业大模型、场景大模型的数智化基础，有机协同，满足各类场景应用。

04、AI智算底座采用的建设模式

全栈自主可控
AI算力资源池

场景孵化
模型部署



- 更灵活**
灵活分配GPU资源，支持不同AI场景
- 更高效**
全面提升GPU利用率，发挥最大效能
- 更稳定**
细粒度多租户管理，实现资源隔离
- 更健壮**
兼容主流开源模型和商用模型

国产芯片替代
创新研究

- “国芯证道”解决方案
- 异构算力的融合
- 适配：硬件性能、软件生态
- 产、学、研国产生态构建

算力创新合作
可信数据沙盒+
分时算力租赁

可靠合作伙伴

选择有较强技术实力+算力，且具有可信资质的算力服务商企业进行合作

独享算力



可信数据沙盒

支持建立金融企业的独立数据存储环境，建立可信数据沙盒，满足模型应用需求

弹性共享算力



分时算力租赁

通过数据加密等技术创新手段，支持算力动态调配，满足高峰算力需求，节约成本投入

监管沙盒



示范效应

可形成算力调度、可信数据环境并重的标杆案例

满足大规模算力需求

05、传统运维在AI浪潮下面临的挑战

传统可观测平台的局限性

传统可观测平台在数据处理、故障诊断上存在效率低、难以精准定位等局限，无法满足证券行业日益复杂的业务需求。



部署架构：三层体系

AI Agent在金融云平台采用数据层 - 模型层 - 应用层的部署架构，确保系统高效运行。

从“被动监控”到“主动决策”

引入AI智能体可实现从传统的“被动监控”到“主动决策”的转型。通过将LLM、genetic Workflow等技术与可观测平台结合，实现根因分析智能体自主决策推断故障根本原因，提升运维人员应急处置效率。



协同机制：多Agent协作

监控、决策、执行Agent协同工作，实时采集、智能分析、自动操作，提升运维效能。

算力运维与业务场景深度耦合

随着证券行业数字化程度加深，算力运维与业务场景深度耦合成为趋势。AI智能体能够更好地适应这种趋势，将运维与业务紧密结合，保障业务的稳定运行。



兼容性与安全：适配现有体系

与现有运维体系兼容，同时采取数据安全隔离措施。

07、特定领域的RAG知识库

三步完成一个特定领域文档助手



多种格式文档自动化处理



RAG知识库助手模板应用



生成web应用发布使用

用户上传文件后，通过自动化的RAG workflow，让大型语言模型深入理解企业知识和业务，利用知识库助手模板快速生成一个web应用，点击发布后即可对外使用。

08、构建AI智能体

低代码的构建自主Agent

1、丰富的工作节点

2、自定义工具和公司内部工具接入

3、可视化的运维监控

通过可视化且低代码的流程编排，支持对接公司内部多套工具和个人自定义工具，构建面向特定业务场景的自主智能体并通过可视化运维、监控，持续优化系统。

09、基于AI、大模型驱动的精细化云资源运维能力建设

覆盖虚拟化、云原生、云中间件、云数据库等云内应用，实现对云网络的可观测能力覆盖；结合CMDB等多源数据，依托AI中心和多平台联动，实现海量数据深入挖掘和分析，提供不同运维场景的数据服务。



10、场景1：科技服务云智能助手

科技服务云智能助手

通过对日常运维文档、平台手册以及知识库的系统化梳理和结构化整理，结合大模型和工具调用能力，方便云用户更快捷地使用云服务。

运营数据采集、云资源配置优化更新

集成云平台资源容量接口，实现对运营数据自动化的采集与实时分析，为云资源优化配置提供数据支撑。

01、通过AI能力提升日常效率

云主机磁盘读写速率 (KB/s) : 564.00
云主机磁盘写入IOPS : 90.00
云主机磁盘读取大小 (KB) : 48.00
云主机磁盘读取耗时 (ms) : 9.59
云主机磁盘块速率 (KB/s) : 192.00
云主机磁盘写入IOPS : 8.00
云主机磁盘块大小 (KB) : 26.33
云主机磁盘块耗时 (ms) : 0.22

2.4 宿主机组数据

宿主机的CPU使用率 (%) : 12.30
宿主机的内存使用率 (%) : 40.49
宿主机的网卡接收带宽 (MB/s) : 0.00
宿主机的网卡发送带宽 (MB/s) : 0.00

三、分析结果

云主机资源利用率特征，计算性能和资源。CPU均值仅1.22%，内存均值21.05%，云主机资源使用合理，未过载，网络性能表现正常，带宽使用率为零，PPS较低，无丢包，网络延迟状态良好，存储性能中，磁盘性能优于高性能，即IOPS验证，请关注业务负载需求，确保业务，资源使用效率验证，建议优化配置以提升利用率，同时确保服务质量。



11、场景2：云资源智能巡检

1. 面向场景的运维风险分析（提前发现问题、防微杜渐）

针对云环境，实现对云资源底座、云原生底座、云原生中间件、云原生数据库、科技服务云等资源的运行分析，协助管理员进行异常挖掘和诊断分析，深度巡检时间从小时级缩短到分钟级

云资源智能巡检

智能体数据获取、数据分析、诊断报告

1. 发起任务

2. 数据获取

3. 诊断分析

4. 巡检报告

5. 查询结果

设置每日定时任务或者手动发起



使用工作流获取不同数据源



大模型根据组合信息进行诊断分析



提供诊断报告与分析建议



全连接云平台助手同步查看巡检任务



12、场景3：面向平台管理员的运维风险分析

面向管理员的运维风险分析

加速问题排查，定界是否云资源存在性能问题

从告警触发 -> 视图发现 -> 大模型辅助判断，阶梯式融入AI智能化能力，逐步从+AI，转向AI+。

帮我分析一下10.180.239.45的云主机性能

需要调用 `analyze_vm_performance_status` 工具来查询云主机的性能。需要用户提供符合的 IP 地址。用户已经提供了 IP 地址。接下来调用工具。

我已经用 `analyze_vm_performance_status` ...

I know what

计算资源

该云主机的性能 (CPU使用0.27%/峰值1.24%、内存使用2.615%、网络IO/秒中119.99字节、网络IO/秒中119.99字节、接收PPS及4.61帧/秒、事务处理能力峰值、存储性能数据与网络性能 (113KB/s, IOPS11)、17.63ms的响应时间。需关注网络介质类型或从网络设备合理性。该主机的综合性能 (CPU使用0.24%/内存14.00%) 快速健康检查。当前配置池内测试环境无告警。您在查看网络情况。建议网络业务需求评估网络性能可行性。同时关注存储性能是否影响业务。

帮我查一下10.180.239.45的存储集群性能情况

需要使用 `analyze_storage_performance` 工具来查询存储集群的性能情况。用户已经提供了 IP 地址。接下来调用工具。

我已经用 `analyze_storage_performance` ...

根据 `analyze_storage_performance` 工具查询10.180.239.45的存储集群中有多台云主机的存储性能情况。需关注这些情况并显示给用户。

根据存储集群性能分析，10.180.239.45的存储集群中有多台云主机的存储性能高。具体如下：

1. IOPS 超过 1000;
2. 吞吐量超过 100MB/s;

帮我查一下10.180.239.45的网络流量情况

需要使用 `analyze_network_performance` 工具来查询云主机的网络流量情况。用户已经提供了 IP 地址。接下来调用工具。

我已经用 `analyze_network_performance` ...

网络资源

我已经获得了网络流量情况的报告链接，现在需要将这个链接提供给用户。并简要说明。

10.180.239.45 的网络流量情况报告已生成。您可以查看 [这里](#)。

结合云主机性能问题快速定界、存储问题性能下钻、云网访问链路分析与诊断，实现对计算、存储、网络资源的辅助诊断。

3.2 技术瓶颈与优化方向

智能体自主决策能力有待提升

约30%的复杂场景下，AI Agent处理故障需人工干预，影响自动化运维效率，需提升推理判断与自主决策闭环能力。

跨云、跨系统协同效率不足

多云环境中，资源调度存在延迟，跨系统协同效率低，制约敏捷交付与统一调度，需构建更加智能、动态的协同架构，提升跨域协同的自动化水平。

合规规则适配滞后

监管政策快速更新时，AI Agent对合规规则的识别与动态适配滞后，动态调整能力不足，需建立政策变化的智能感知与规则的自动映射机制，提升系统合规与风险防控能力。

多模态数据融合等技术攻坚

对异构、多源、多模态数据的融合和分析，强化多模态数据融合决策能力，是突破当前技术瓶颈、提升智能运维深度与准度的关键方向。

3.2 全场景运维决策机制深化场景

构建多模态数据融合的智能决策体系

整合日志、指标、链路、业务数据，构建覆盖“基础设施-平台-应用-业务”全景视图，为运维决策提供全面信息。

加强预见性维护能力,优化动态资源智能调度

基于时序预测与趋势分析算法，对关键业务场景的风险预判机制，结合AI驱动的交易负载预测与资源匹配机制，实现按需分配与动态调整，提高资源利用效率。

构建运维知识图谱，强化智能决策知识支撑

沉淀历史故障经验、典型处置流程以及动态更新的合规监管要求，构建覆盖技术、业务与合规的多维度运维知识图谱。提升复杂场景下的判断准确性和处置合理性。



谢谢观看