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

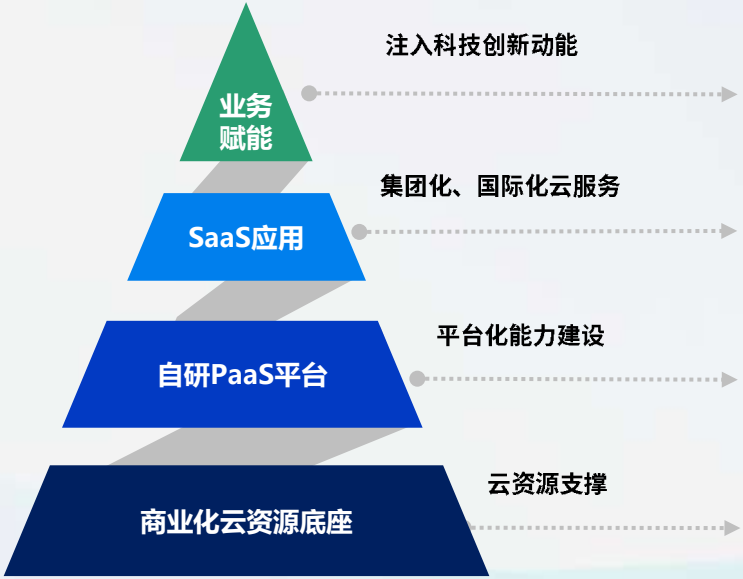
01、国泰海通云数据中心定位

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**。

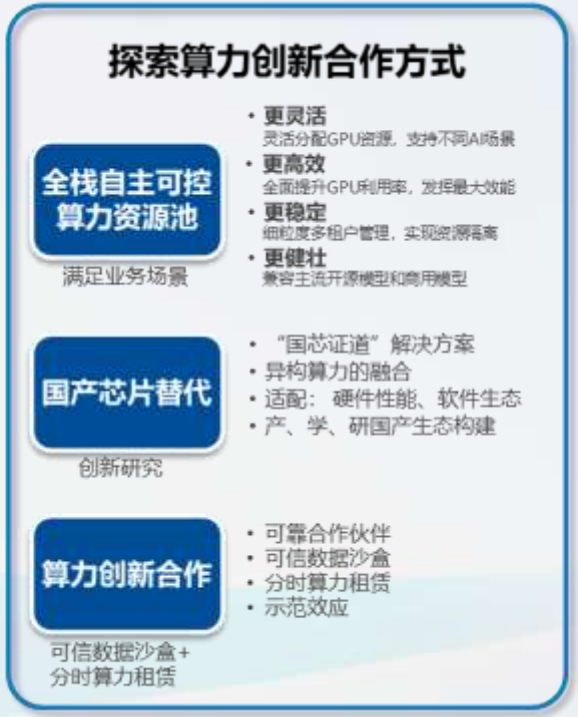
承载金融科技落地和数字化转型，实现非交易类业务**100%云化**，为构建灵活共享的业务中台、融合智能的数据中台、协同高效的云上办公，以及金融场景创新等场景提供**稳定、高效、易用**的云底座。

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



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

模型训练 | 模型推理 | 模型加速（VLLM） | 服务部署

一云多芯

混合部署 | 统一管理 | 统一视图

K8S/Docker/云原生

GPU资源管理
卡管理 节点管理
配额管理 节点组管理

GPU虚拟化/池化
GPU虚拟化 GPU池化
资源超分 算力切分

GPU资源调度
离在线调度 调度策略
亲和性调度 任务优先级

监控日志

用户权限

异构计算（GPU、NPU、ASIC）：Nvidia | 昇腾 | 海光 | 燧原 | 天数

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

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

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

示范效应
可形成算力调度、可信数据环境并重的标杆案例

独享算力 + 弹性共享算力 + 监管沙盒 = 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。

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。

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。

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。

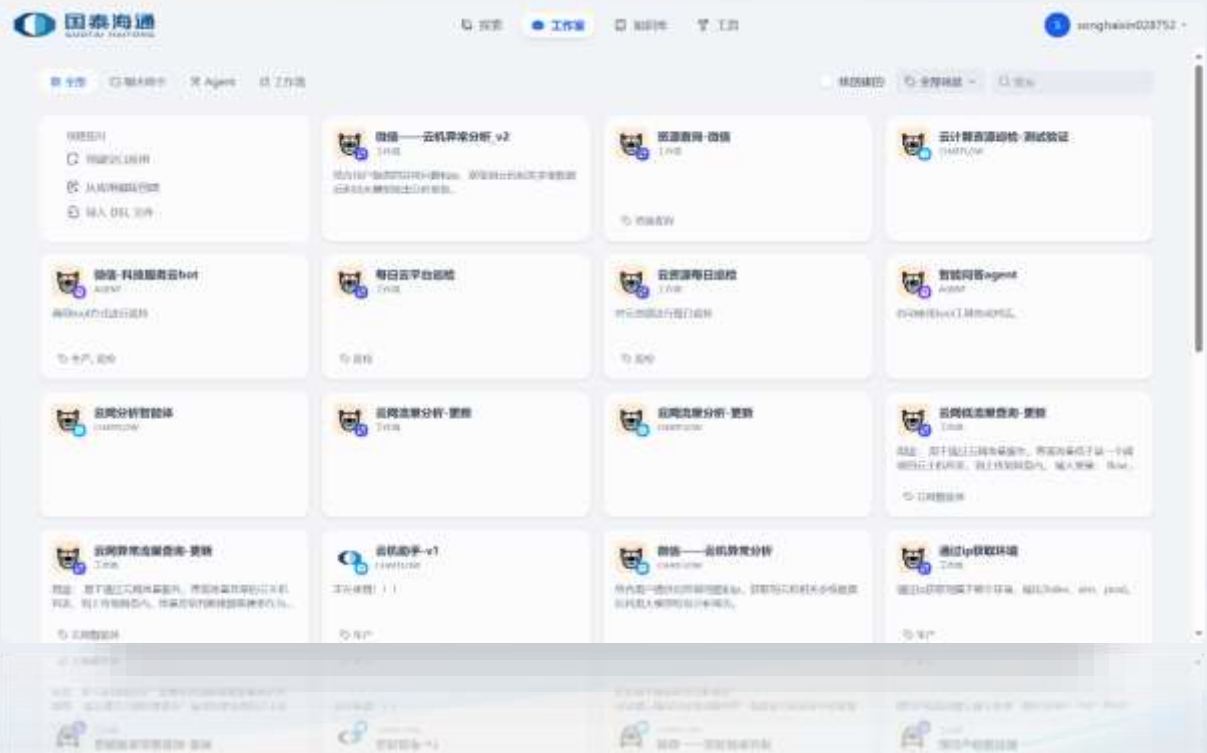
协同机制：多Agent协作

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。

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。

06、自研大模型（LLM）应用开发平台



自研平台支持模型、数据私有化，保障数据安全，促进团队协作。

自无需编程技能，0 代码快速搭建个性化 AI 原生应用

接入公司多套内部系统和提供自定义工具能力，构建出功能强大的大模型应用

多租户管理 团队协作，效率UP

超低使用门槛 低码搭建，智在指尖

容器云部署 性能保障，专注创新

保障数据隐私 守护隐私，安全护航

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工作流，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

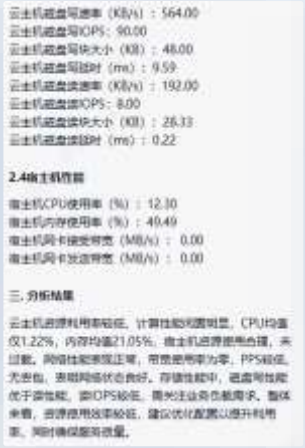
科技服务云智能助手

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



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）

云资源智能巡检

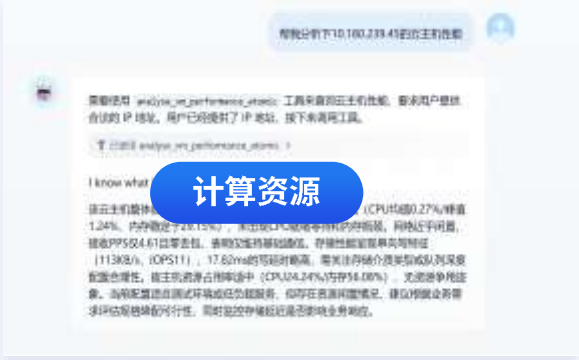
针对云环境，实现对**云资源底座、云原生底座、云原生中间件、云原生数据库、科技服务云**等资源的运行分析，协助管理员进行异常挖掘和诊断分析，深度巡检时间从**小时级**缩短到**分钟级**

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+。



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。

13、场景4： 面向科技服务云用户的AI智能分析

面向科技服务云用户的AI智能分析

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