主题：数智驱动下的行业数字金融发展

交易技术前沿

2024 年 第 4 期 总第 58 期

ITRDC 证券信息技术研究发展中心（上海）

上海證券交易所 SHANGHAI STOCK EXCHANGE

内部资料 免费交流
《准印证》编号沪（K）0671

交易技术前沿

2024 年 第 4 期 总第 58 期

运营:
证券信息技术研究发展中心（上海）

主管、主办:
上海证券交易所

刊首语

随着数字技术与实体经济持续深入融合，加快发展数字经济，成为推动经济社会高质量发展之必需。2023年10月，中央金融工作会议提出包括数字金融在内的“五篇大文章”，进一步明确了新阶段下利用数字化转型推动金融高质量发展的蓝图目标。其中，随着ChatGPT为代表的大模型技术横空出世，人工智能技术向世人证明了其强大的数据分析能力、广泛的普适能力和出色的创造能力，凸显其在第四次工业革命中引领创新驱动、变革生产方式的核心驱动作用。2018年，习近平总书记在中央政治局第九次集体学习时强调，人工智能是新一轮科技革命和产业变革的重要驱动力量。近年来，行业机构在《证券期货业科技发展“十四五”规划》的指引下，加大算力、算法和数据建设，积极利用智能技术驱动业务数字化转型，在智能投顾、智能运营、智能风控等领域结出丰硕成果。本期《交易技术前沿》以“数智驱动下的行业数字金融发展”为主题，精选行业在数智驱动转型方面的优秀文章，为行业推进数字金融发展提供参考。

中信建投证券的《数字金融-中信建投证券投顾大模型推动业务创新》应用大语言模型，结合RAG、提示工程与Agent技术，优化了投顾业务服务模式，实现从预测型服务到交互型实时投顾升级，提升主动服务占比。

海通证券的《印章物联助力证券营业网点开启智慧运营新篇章》利用生物识别、红外感应等技术，实现了印章的全生命周期数字化管理和人印分离，提高运营效率，提升风控水平。

西部证券的《数智驱动：西部证券统一风险管理平台实践》将数智化理念融入风险管理，依托大数据、AI、RPA等技术，建立分层数据体系与分布式微服务应用架构，打破了数据壁垒，构建了业务全覆盖的全面风险管理体系。

恒生电子的《基于大模型的证券业务办理助手探索与实践》针对证券公司复杂的业务流程和专业门槛，提出一个结合多Agent和MLLM多模态技术的高效业务办理模型，通过优化业务流程提高业务办理的效率和精度，增强交互体验。

证券信息技术研究发展中心（上海）
2024年11月15日

目录

01 本期热点

02 前沿技术应用

03 实践探索

04 信息资讯采撷

01 本期热点

数字金融-中信建投证券投顾大模型推动业务创新

陶剑峰，张蕾，贾振兴，滕龙启，孙智强 | 中信建投证券股份有限公司 | Email：zhangleixx@csc.com.cn

摘　要： 中信建投证券的投顾业务已实现了基于传统专家系统的“智能投”服务，提供预测型主动投顾服务，但被动型服务占比依然很高，服务能力提升面临挑战。通过践行数字金融发展理念，我们引入最新的人工智能技术，借助大语言模型泛化性强、上下文问答、可以借助外部工具的优势，在金融行业大模型的基础上训练了投顾领域大模型。结合投顾专家经验、检索增强生成、提示工程、智能体技术，开发了投顾业务运营助手、投顾专家助手、智能专家投顾等若干个创新业务，提升“智能顾”的能力，服务内部员工和外部客户。变预测型主动投顾服务为交互型实时投顾服务，大幅提升了主动服务的占比，实现了投顾业务创新。

关键词： 数字金融；投顾；大模型；训练

一、引言

中信建投证券投顾业务处于行业第一梯队，打造了“找好投顾——到中信建投”的行业品牌。形成了人工投顾 + 智能投顾 + 基金投顾三位一体的、涵盖股债基的完整的投顾业务体系。为全体客户提供投前、投中、投后全投资周期内的高质量的投顾服务。建设了包含行情异动、资讯解读、选股、择时、交易、风控、组合、投后分析等完备的投顾产品体系。但目前的服务体系，只做到了通过训练投顾专家系统来主动分析客户画像，基于客户自选、持仓、交易行为、使用行为数据等为广大长尾客户提供预测型主动投顾服务，侧重于“智能投”的能力。存在服务效果需要后评估、货架式被动型服务占比高、主动型服务占比低、交互性差、服务质量难以保证等问题。主要面临如下的挑战：

数据准确性、服务时效性：投顾服务需要根绝最新的行情、资讯等数据来制作投顾策略，数据的实时性、时效性非常重要。

海量、个性化服务能力：智能投顾服务的目标是海量的长尾客户，此类客户在证券公司的占比超过 80%，提供海量、个性化的服务能力，对计算资源、并行计算能力等提出了不小的考验。

服务温度：投顾服务更多的是一种陪伴式服务，整个过程需要体现服务的温度。从投资小白到投资专家，从投资者教育到专业的投资报告，从稳健性投资到激进型投资，每个过程都需要投顾人员专业的引导和陪伴。甚至投资过程中暂时的回撤，都需要投顾人员基于及时的心理安抚，做到理性投资。智能投顾通过机器人提供自动化服务，在意图识别、多轮交互、专业服务等各个环节都需要提升。

投顾专业能力：投顾服务的专业度直接影响客户投资的效果，需要具备千人千面的、专业的服务能力。

实时的风险控制：智能投顾作为自动的机器人投顾，需要保证服务的安全合规，实现实时的风险控制，在风险识别、合规检查等技术上面临诸多难点。

投顾业务闭环：投顾服务是一个覆盖投前需求分析、投中组合再平衡、投后绩效分析进而产生新需求的一个持续投资的过程，需要实现全过程的闭环式服务。

针对以上情况，我们结合数字金融的发展理念，推动金融与数字技术的有机结合，利用最新的人工智能技术——大模型技术提出了实时交互的主动型投顾形式，全面增强“智能顾”的能力来实现业务创新。

新业务具有如下优势：

7*24h 在线自主运营：实现投顾业务 7*24h 在线自主运营，解答投顾产品使用、签约指导、收费咨询、信号推荐等问题。

专业化的投顾服务：训练高水平的 AI 投顾专家抓手，能够学习投顾专家风格、偏好，结合行情、资讯、交易、财报、研报、客户持仓自选等信息，辅助提供高质量的投顾话术生成、投资建议指导。从而协助投顾老师高效的完成投顾产品制作、客户分析、投资咨询服务等。提高投顾老师的服务效率和服务质量下限。

个性化的自主服务：通过学习专家风格、偏好，结合行情、资讯、交易、财报、研报、客户持仓自选等信息，自动提供高质量的投顾话术生成、投资建议指导。可以作为数字人的底层驱动，创造出投顾老师的数字分身，通过实时交互的形式，自动完成客户分析、需求收集、投顾产品匹配、个性化投资咨询服务、形成产品改进建议等等。极大地改善投顾服务体验。

二、大模型在投顾领域应用面临的问题

当前，投顾领域有广泛的实时交互服务需求，是大语言模型落地的理想场景。然而，大模型作为一项新技术，具有幻觉、可控性差、道德风险高、合规控制难、信息不及时、领域知识欠缺等问题。如何建设具备高度专业性、时效性强、风险控制能力强的投顾领域大模型，实现投顾业务创新，是亟待研究的问题。

幻觉：大模型存在生成的答案会存在不基于任何事实数据、一本正经的胡说八道。

可控性：大模型本质上是以预测下一个 token 为核心的概率式模型，会存在生成内容不可控。

专业性：通用大模型因为缺乏专业领域的知识和业务 KNOWHOW，会存在在领域内专业性不足。

时效性：预训练的模型采用的知识语料通常截止到某一个固定的时间点，模型在时效性方面会存在知识更新滞后的问题。

有限的数据：投顾业务涉及行情、资讯、研报、财报、交易、策略、法规等众多数据，特别是专业的投顾经验性数据严重缺失，导致投顾大模型训练难度加大，模型无法充分利用多维数据进行全面分析，预测准确性下降。

安全合规：大模型输出内容的不可控性，增加了数据泄露和滥用的风险，进而引发操作风险、道德风险等各种风险。使得模型的合规控制变得很难。

三、中信建投证券投顾大模型的研发实践

3.1 设计思路

充分利用大模型在海量、非结构化数据处理、人机交互、生成、逻辑推理、多模态等方面的显著技术优势。利用 Multi-agent（多智能体）、RAG（Retrieval-augmented Generation，检索增强生成）等技术弥补大模型在可信、逻辑推理、实时性、风控等方面的不足。同时对大模型能力进行分层，分为 5 层，通过设计不同层次的应用场景和应对措施来确保在应用侧的落地效果，见图 1。

3.2 整体框架

整体开发框架分为算力、适配、模型、能力、应用 5 层，见图 2。

底层由 A100、H100 等高性能计算资源提供强大的算力支持。其上是算法适配与加速层，通过稀疏量化、分布式优化等技术保障大模型的高效运行。大模型层支持灵活调用包括文心一言、LLAMA2、自研投顾 LLM 等模型，灵活适配场景的效果要求。能力层负责将大模型的生成、理解、推理等关键能力进行封装，以支撑应用层答疑、投顾 Copilot 等场景应用的创建。

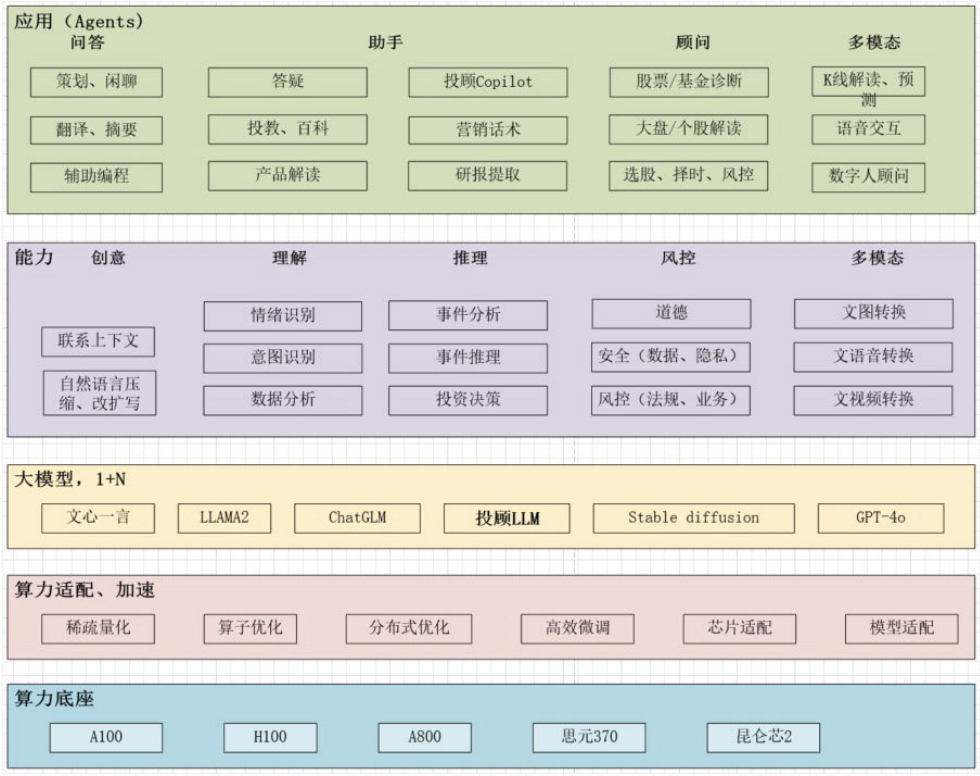


图2：投顾大模型技术架构

图1：投顾大模型能力&场景象限

3.3 实现方法

3.3.1 投顾大模型的训练

在金融模型基础上采用迁移学习的方式实现投顾 LLM 训练和微调，见图 3。

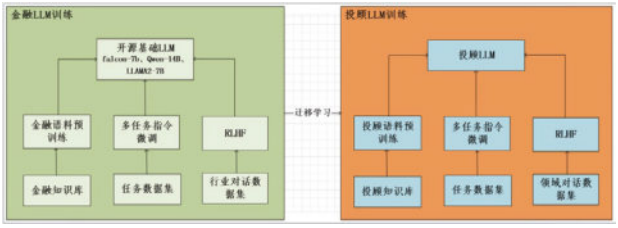


图3：投顾大模型训练过程

首先我们用数千亿 TOKEN 的中文金融语料和数百万的多轮对话语料预训练了一个金融行业大模型。在金融行业大模型的基础上，我们用构建的投顾场景的语料进行 SFT，形成投顾领域大模型。

3.3.2 基于 Agent 技术研发的投顾场景一体化开发平台

基于开源框架自研了 Agent 一体化开发平台，该平台融合了包括函数调用（Functioncall）、检索增强生成（RAG）、自然语言转数据库查询语言（NL2SQL）、知识库管理、工作流（Workflow）等深度 Agent 技术能力，以满足投顾业务复杂性要求。能够实现模型调度、应用编排、插件接入、知识库接入、幻觉控制、风控、协同管理、混合云部署等。平台能够处理复杂的逻辑，通过整合不同大模型的能力、集成调用 python 等，很方便的进行特定业务逻辑的处理。

3.3.3 投顾场景实现过程

以投顾大模型为模型底座，采用 1+N 的模型（1 个投顾主模型 +N 个辅模型）组合形式，采用 Multi-agent 技术，实现投顾场景从问题输入（识别）、问题规划（思维链）、问题实施（拆解、RAG）、到总结与反思，全流程的自动化处理。针对投顾场景需满足合规要求的业务特点，针对性设计了合规风控 agent，见图 4。

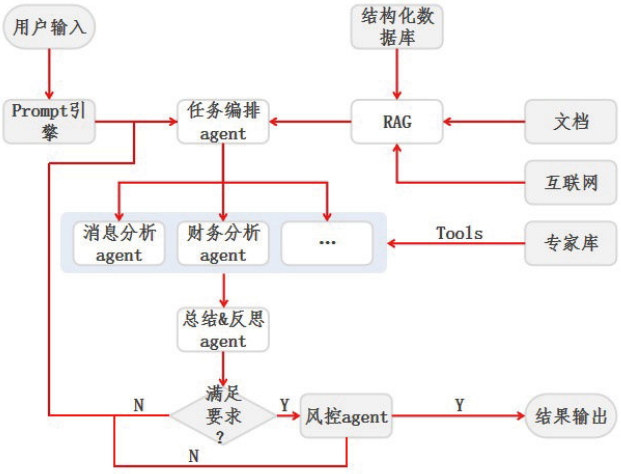


图4：投顾agent开发流程

3.4 评测体系建设

构建了包括基准评测集、智能投顾大模型评测集及投顾场景评测集的三层评测体系，以对齐最终应用场景的落地效果。

3.4.1 基准评测集

基于通用评测网站的公开评测集进行通用能力评测，包括 HELM、MMLU、C-EVAL、BigBench、HumanEval、AGIEVal、SuperCLUE、OpenLLM。

3.4.2 智能投顾大模型评测集 RAVal

加入投顾领域知识进行领域大模型专用能力评测，涵盖金融知识、投顾业务、人文社科、创意、推理、安全风控、道德对齐等。

3.4.3 投顾场景评测集

针对不同的投顾场景，加入场景特有评测数据，进

行场景层面的个性化评测，如运营助手、BI 助手、投顾专家助手、智能专家投顾等特有场景评测。

四、中信建投证券投顾大模型的实践效果

4.1 投顾业务运营助手

4.4.1 投顾答疑机器人

我们构建了一个面向客户经理的企微投顾问答机器人，见图 5。将客户经理关心的投顾产品培训等材料建成知识库，替总部投顾团队的运营专家解答客户经理关于产品的长尾问题，有效的释放了总部投顾团队运营专家的精力。

经过持续优化，目前在构建的评测集上问答准确率做到 97% 以上，问答结果可溯源。且拒绝回答投顾领域外问题，保持业务严谨性。

4.4.2 投顾数据分析机器人

我们构建了一个面向投顾运营人员的投顾数据分析机器人，见图 6。能够自动完成投顾运营数据库的数据分析。运营人员通过自然语言交互的形式来完成报表生成，降低 IT 人员依赖，提高分析效率。

取数准确率超过 90%。查询分析过程可溯源、验证。

4.2 投顾专家助手

我们构建了一个面向投顾老师的投顾专家机器人，见图 7。作为一体化投顾支持平台，能够学习投顾专家的风格、偏好，结合行情、资讯、交易、财报、研报、客户持仓自选、各种投顾策略的真实信号信息，生成高质量的投顾话术、投资建议等。协助投顾老师完成投顾产品制作、客户分析、个性化咨询等工作。分析标的涵盖 A 股全部股票。

实现了一员一专家助手，助手风格、偏好可定制，适应不同的投顾场景。

生成内容时效性强，数据源完整、丰富，且集成了投顾特有高质量策略库，保证了助手的专业度。

生成内容的参考资料可溯源、可验证。通过训练投顾质检模块，确保生成内容依法合规。

图5：投顾答疑机器人

4.3 AI 专家投顾

我们正在构建一个面向客户的专家投顾机器人，通过自主学习投顾老师的专业知识、服务经验，打造 AI 专家投顾机器人。让全体客户均可享受到高质量的 AI 专家投顾服务。7*24h 进行投顾产品使用讲解、图文展示、研报解读、观点生成、资讯解读、大盘 / 个股解读、个股 / 基金诊断。根据客户持仓、自选、行情、资讯、使用行为及各种投顾策略的真实信号信息，自主为客户提供选股、择时、交易、组合等个性化的投资建议。

计划实现一客一 AI 专家投顾，投顾风格、偏好可选，适应不同的顾问需求。

生成内容时效性强，数据源完整、丰富，且集成了投顾特有高质量策略库，保证了专家的专业度。

生成内容的参考资料可溯源、可验证。通过训练投顾质检模块，确保生成内容依法合规。

查看取数过程

任务分析

1. **查询对象和数据类型**:

- 查询对象是"直播品类"，这指向了数据类型C，具体表为 bm_zh_live_category。
- 需要查询的字段是 live_c2（二级品类）和 retail_sales（网络零售额）。

2. **时间**:

- 时间为2023年，没有具体的月份或季度，因此我们考虑全年的数据。
- 对应的时间查询为：time_id IN (SELECT id FROM bm_zh_data_time WHERE year = 2023 AND simplealiasna
'12月累计')。

3. **地域**:

- 地域为德阳市，查询的是数据类型C。
- 对应的地域查询为：area_id = (SELECT id FROM bm_m_area WHERE name = '德阳市')。

4. **排序和限制**:

- 需要按照网络零售额降序排序，并取前五个。
- 使用 ORDER BY retail_sales DESC LIMIT 5。

完整SQL语句

基于以上分析，原问题对应的完整sql语句如下：

```
SELECT live_c2, SUM(retail_sales) AS total_sales
FROM bm_zh_live_category
WHERE time_id IN (SELECT id FROM bm_zh_data_time WHERE year = 2023 AND simplealiasname = '12月累计')
AND area_id = (SELECT id FROM bm_m_area WHERE name = '德阳市')
GROUP BY live_c2
ORDER BY total_sales DESC
LIMIT 5;
```

这个SQL语句将查询德阳市在2023年累计的直播品类中，网络零售额最高的前五个二级品类。

答案:

[{'total_sales': 1.0689325841, 'live_c2': '生鲜食材'}, {'total_sales': 1.0546300794, 'live_c2': '配件'}, {'total_sales': 0.7889328543, 'live_c2': '手机数码'}, {'total_sales': 0.3259002067, 'live_c2': '美容美妆'}, {'total_sales': 0.28963496, 'live_(

图6：数据分析机器人

图7：投顾专家助手

五、总结与展望

本文展示了在数字金融大背景下，在券商行业如何利用大模型技术重塑投顾业务。通过引入大语言模型（LLM）、RAG、提示工程、Agent 等前沿技术，自建大模型评测体系，研发了多款投顾创新应用，显著提升了主动服务的占比，实现了投顾业务的数字化转型。

展望未来，尽管投顾大模型在业务中取得了显著成效，但仍面临幻觉、可控性差、专业性不足等挑战。中信建投证券将持续践行数字金融发展理念，继续优化大模型的专业性和实时性，增强模型的可控性和合规性，推动大模型技术在更多应用场景落地。如数字人投顾、多模态投顾等。通过持续创新和优化，中信建投证券将打造更加智能化、精准化的投顾服务体系，保持投顾业务在行业的领先地位。

参考文献：

[1] 沈艳，人工智能在数字金融中的应用 . https://mp.weixin.qq.com/s/mo9lCnhbvzWHRafS3Kna6w

[2] 技术狂潮 AI，开源 LLM 微调训练指南：如何打造属于自己的 LLM 模型 . https://mp.weixin.qq.com/s/R-6ds1bFmOqPANIgVCs2Gg

[3] 机器之心，大型语言模型综述全新出炉：从 T5 到 GPT-4 最全盘点，国内 20 余位研究者联合撰写 . https://mp.weixin.qq.com/s/7HRr55Md2Wl6EHQMGioumw

印章物联助力证券营业网点开启智慧运营新篇章

陆颂华，任荣，曾利，李田凤，王春，金宗敏，陈善新 | 海通证券股份有限公司
| E-mail：ltf13923@haitong.com

摘　要： 随着金融科技、物联网技术与证券行业深度融合，网点智慧运营模式转型成为金融机构增强核心竞争力的重要发展机遇，公司着力提升整体金融服务质量的同时对营业网点的合规风控水平要求不断提高，智能印章的推广摒弃了传统的印章管理模式，结合物联网+技术，实现了“人印分离”，从而增强风控管理水平，赋能网点智慧运营。本文阐述了网点智慧运营的背景与印章物联赋能智慧运营的意义，介绍了印章物联管理系统的设计思路、硬件功能、审批监控流程和印章全生命周期管理模式等，最后总结了印章物联对于赋能网点智慧运营的成效。

关键词： 物联网；印章物联；智慧运营；数字化；合规风控

一、背景

1.1 市场背景

近年来，以物联网、人工智能、区块链、云计算、大数据等为代表的信息技术飞速发展，在改变人们生活方式的同时也冲击着金融行业的传统经营模式，在这场以智能化、数字化、信息化为核心的网点智慧运营改革浪潮中，多家证券公司将金融科技纳入重点发展战略，并加大信息技术投入，大胆探索新的运营管理模式。因此，推动证券行业的智慧运营转型，通过运用现代信息技术，来优化公司业务流程、提高服务效率、降低运营管理成本、增强合规管理和风险控制能力成为重要改革方向。

1.2 行业政策

在此背景下，国家对证券行业的转型发展也十分关注。早在 2021 年，证监会发布《证券期货业科技发展“十四五”规划》，明确了“十四五”时期科技监管工作的两大主题是“推进行业数字化转型发展”与“数据让监管更加智慧”。2023 年在第十四届全国人大第一次会议上对证券行业提出《关于进一步推进证券行业数字化转型的建议》的议案，中国证监会在就有关问题的答复中提到一直高度重视并持续推动证券期货业数字化转型工作。相关政策的出台加快了行业数字化转型的进程，为资本市场高质量发展注入了强劲动力。

1.3 公司建设

海通证券在“十四五”规划期间紧跟国家与行业发展步伐，以“数字化转型”为重要抓手，加大信息技术投入，推动营业网点智慧运营建设。增强合规管理与优化运营效率是智慧运营的重要环节，但由于公司传统印章用印频繁，管理难度大易存在监管风险，因此，如何对印章的使用情况进行全程监控成为了规划期间亟需解决的难题。

公司通过建设印章物联管理系统实现了传统印章的数字化和纸质表单的无纸化管理，成为行业首家探索分支机构印章物联管理模式的券商，在加强传统印章风控管理方面走在了行业的前列。

1.4 物联印章赋能智慧运营

在数字化经济的新时代，公司把握好智慧运营发展机遇，通过推广使用印章物联管理系统，从而提升客户体验、优化业务流程、提高运营管理效率，增强合规管理和风险控制能力，实现金融科技与业务发展相融合。物联印章助力企业智慧运营的意义主要有以下几点：

1）提升客户体验。印章物联管理系统利用生物识别、红外感应、人工智能等技术满足客户专业化的需求，增强客户归属感和认同感，从而提高公司的市场竞争力。

2）优化运营效率。利用线上审批、设备在线管控、用印数据统计查询等功能，提高营业部印章管理效率，推进网点集约管理，从而提升运营效率，促进利润增长。

3）加强风控管理。通过用印人员监控、违规用印预警、电子围栏等方式，对用印潜在风险全面管控，从而降低监管风险，提高合规风控水平。

二、总体设计

印章物联管理系统是公司强化智慧运营建设、赋能分支机构的重要项目之一，也是严格落实监管要求、提升合规经营能力的重要举措。

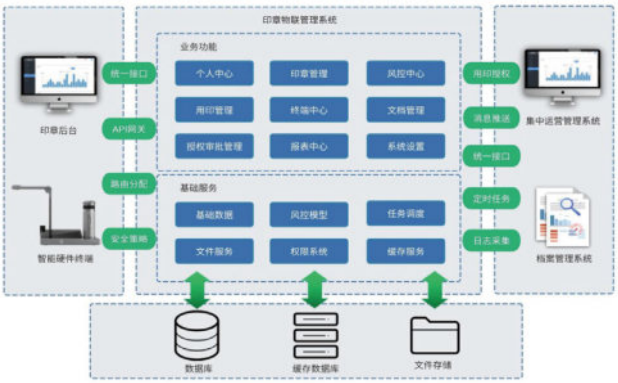


图2.1：印章物联管理系统

印章物联管理系统包含智能硬件设备终端和软件管理系统两部分，软件管理系统可以实现印章登记、授权解锁、用印记录、统计管理、风控预警等功能；智能硬件设备终端包括智能章筒、印控台、摄像头、红外感应等组件，支持印章保护、用印文件电子影像采集、用印人员头像采集、用印区域感应预警等功能。

本系统对业务印章管理进行智能化转型升级，实现实体印章的刻制、领取、使用、更换、销毁等环节的全生命周期管理，严防印章违规使用、丢失等风险；建立了用印分析多维报表，通过实时查询用印详情及各类数据分析，为企业发展赋能; 通过软硬协同实现管理闭环，进一步规范分支机构业务印章管理。印章物联管理系统的推广打通了人、章、文件的互通互联，实现了印章的数字化、精细化管理，标志着证券行业智能、智慧、安全用印时代的到来。

三、印章物联管理系统建设介绍

3.1 智能硬件设备终端

印控实体机由智能章筒和印控台组成。该套设备在兼顾功能性的同时，具备简洁优美的外观和轻巧便携的机身，可实现多种用印监管模式和防盗用功能。

图3.1：智能硬件设备

智能章筒主要是用于密封实体印章，做到人章分离，具体由章筒、实体印章、蓝牙、防拆报警传感器、三轴陀螺仪等构成，主要有以下功能：

1）机身具有高强度防拆外壳，防止暴力拆卸，机顶的指示灯可提示智能章筒蓝牙连接、网络、充电、系统运行等状态，并配备液晶显示屏，触摸可显示电量。

2）通过红外传感技术与三轴陀螺仪结合，可在远程授权状态时检测到章筒是否超出盖章区域，有效防止偷盖印章。

3）内置防拆报警传感器、自动化舵机，可在设备被恶意破坏时强制弹回印章并实时发送预警信息，防止风险扩大。

4）具备 GPS 定位，当印章被外带时可在后台实时跟踪定位。

印控台可实现对智能章筒的近距离预警管控、跟踪用印过程，主要由防遮挡摄像头、人脸摄像头、红外传感器、液晶触摸屏等构成，主要有以下功能：

1）具有大尺寸液晶触摸屏，可以高效便捷地实现对智能章筒的管控、记录用印过程、设置网络和蓝牙连接信息、进行版本升级等功能。

2）采用验证码认证方式解锁智能章筒，提高印章使用的安全性。

3）配备防遮挡摄像头和人脸摄像头，可在用印时对盖章文件、用印人人脸进行拍照归档，在实现表单电子化的同时增加了安全性。

4）内部的红外传感器可设置“电子围栏”，实现安全区域内盖章，脱离范围无法用印。

3.2 印章物联管理后台

3.2.1 系统方案设计

在系统部署之初，结合本公司实际应用，遵循系统设计一致性、系统灵活性等原则，对信息保密程度分级，对用户操作权限分级，对网络安全程度分级（安全子网和安全区域），对系统实现结构分级（应用层、网络层、

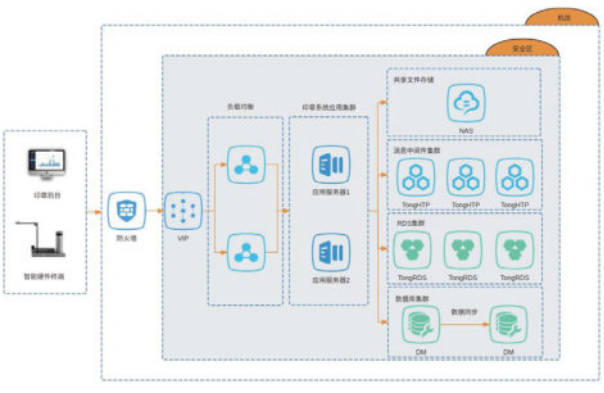


图3.2：印章物联管理系统架构

链路层等），从而实现系统最优方案设计。

印章物联管理系统采用云化部署，已支持全栈国产化，系统应用期间产生的数据全部在本地服务器中进行加密存储。软件平台使用 JAVA 语言开发，B/S 架构，使用达梦数据库实现业务数据存储，TongWeb 作为应用服务器，TongRDS 作为高性能数据存储，TongHTP 作为消息中间件，TongTHS 作为前端服务负载中间件。PC 端支持 360、奇安信、IE、chrome、火狐、Google 等浏览器访问，可实现无客户端模式访问应用。

3.2.2 系统功能概述

印章物联管理系统通过公司内部网络对传统印章进行远程集中管控，可有效的降低传统印章在使用过程因偷盖、漏盖等不规范操作引发的业务风险，同时提高了印章管理人员对实体印章进行盘点和监督的效率。

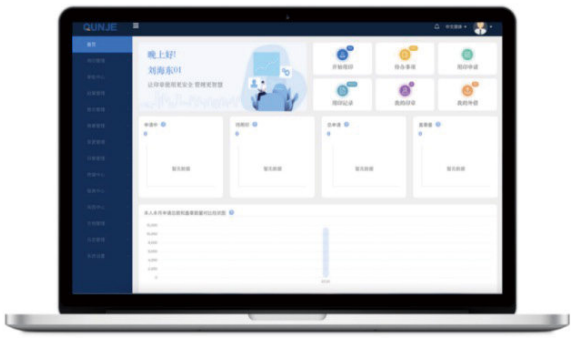


图3.3：印章物联管理后台

后台主要由用印管理、审批授权管理、印章管理、终端中心、报表中心、风控中心、文档管理等业务模块构成，主要功能如下：

1）用印管理，对用印情况进行统计分析，可按照印章名称、所属部门、盖章文件类型、用印时间段等条件进行筛选，记录用印文件、用印人人脸、用印过程日志。

2）审批授权管理，展示当前账号待处理、已经审批的事项列表，可快速查阅审批业务流程。

3）印章管理，对公司所有印章进行统计盘点，并根据设备识别码与蓝牙 MAC 地址绑定唯一的营业部印章与设备，对每个终端独立设置预警机制，并且可查看印章的定位分布信息。

4）终端中心，对智能章筒进行管理，主要用于添加智能章筒，设置区域预警功能；对印控台进行管理，主要用于维护、查看印控台应急密码等信息。

5）报表中心，将所有用印的信息以图表形式直观展现，可快速了解公司印章的使用情况，展示信息包括已申请用印流程数量，印章使用情况，盖章文件占比，流程完成占比，各省用印分布，各省印章数量分布等。

6）风控中心，对印章长按压、异地用印、区域外用印、拆机、应急用印等风险性操作进行预警记录提示。

7）文档管理，对用印流程中产生的用印文件、用印人人脸进行归档查询。

3.3 智能用印审批流程

智能用印审批流程包含常规用印和应急用印两种场景，通过公司集中运营管理系统调用实现用印业务办理，在智能硬件设备终端（智能章筒、印控台）实现表单用印。两种用印场景的设计增加了系统异常兼容性，在满足日常用印需求的同时，也可实现先用印，后审批的应急业务办理流程。

3.3.1 审批流程

1）常规用印

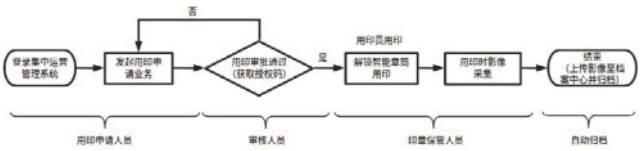


图3.4：常规用印流程

常规用印流程主要用于营业网点的日常用印需求。在有用印需要时，营业部的业务办理人员在登录集中运营管理系统后可发起用印流程，录入印章类型、用印次数、用印文件类型等要素后提交审核，经审核人员审核通过后获取授权码，印章保管人员通过在印控台输入授权码解锁智能章筒用印，用印后影像文件可自动归档。

2）应急用印

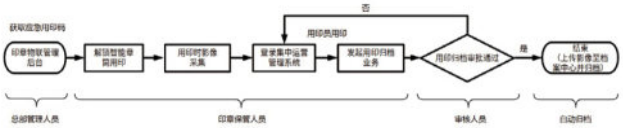


图3.5：应急用印流程

应急用印流程主要用于紧急用印需求。应急用印流程可免去事前申请审批的时间，由总部管理人员登录印章物联管理后台获取应急用印码，印章保管人员拿到应急用印码后可直接解锁智能章筒用印，用印完毕后再登录集中运营管理系统发起应急用印归档业务，经审核人员审核通过后可自动归档相关影像。

3.3.2 用印流程

图3.6：印控台用印流程

用印流程主要指在印控台进行解锁用印的过程。用印人员在集中运营管理系统申请用印授权码之后，由印章保管人到印控台进行用印，用印前需要先选择对应的章筒进行蓝牙连接，连接成功后点击智能用印菜单输入授权码便可解锁智能章筒用印，智能章筒只能在“电子围栏”规定范围内盖章，超出用印范围会触发预警，同时，若发生长按压章筒、强制拆机等行为也会触发预警，用印时印控台会通过摄像头拍摄用印文件和用印人脸，这样系统会对用印过程详细记录留痕，大大降低了违规用印的风险。

3.4 印章全生命周期管理

为加强印章的集中化管控，印章物联管理系统结合公司 OA 系统，对营业网点实体印章的刻制、领用、保管、使用、上缴和销毁进行详细备案，实现印章全生命周期的信息化管理。主要有以下几点：

1）印章刻制。当营业网点新建或迁移时，如需刻制或更换业务印章，可由营业网点相关人员通过公司 OA 系统发起印章刻制专项流程，总部收到申请后统一刻制。

2）印章领用。完成实体印章刻制并装入智能章筒后，由总部通知营业网点领取智能章筒和印控台，做好领用登记后在印章物联管理系统登记相关信息。

3）印章保管。智能章筒和印控台交由营业网点合规总监或综合岗保管，印章保管人负责定期检查设备，保证运行正常。

4）印章使用。用印人员可以通过集中运营管理系统发起用印申请，审核通过后可以在印控台解锁智能章筒用印，用印时拍照记录用印文件与用印人人脸后自动归档到档案中心。

5）印章上缴。当营业网点不再需要使用业务印章后，需要安排专人将智能章筒和印控台上缴总部，总部相关人员做好登记后妥善保管。

6）印章销毁。总部收集停用的业务印章后，统一提交上级管理组织销毁。

3.5 印章使用数据监控大屏

数据大屏主要包含全国各省印章分布地图、事项和用印统计、事项和用印数量 Top3 排行榜、月用印量统计、24 小时内用印量统计、组织排名统计、预警次数统计等功能模块。

数据大屏可以实时监控印章的使用情况，包括营业部用印情况、盖章环比增长量、使用次数、使用人员、使用时间等信息，有助于分析印章使用的趋势和模式，优化印章管理策略；可以帮助识别潜在的印章滥用风险，例如区域外盖章、异地盖章、长按压等异常用印情况，这有助于降低合规风险，增强风控水平；在有统计需求时可以快速获取印章使用的相关信息，减少查询统计时间，提高效率。

四、智慧网点建设成效

2022 年 1 月，海通证券完成业务印章物联管理系统全面上线，实现分支机构全覆盖，推进智慧营业网点建设，全国使用印章总数 600+ 个，上线以来每年用印申请流程超数万次，用印次数超十几万次。

图3.7：印章使用数据大屏（测试数据）

4.1 印章风控体系完善实现网点建设“合规化”

在原始的人工管理印章模式下，易夹杂人情章、越权私盖、偷盖印章等违规行为，印章管控流程制度繁琐低效，分散化管理模式风险较高，且总部难以实现对分支机构用印的合规管理，容易给公司造成巨大经济损失和名誉损失。

智能硬件终端的实时定位、电子围栏、异常预警等功能可以有效地对分支机构用印行为进行规范，防止印章丢失，实现人章隔离，可以对用印过程全纪录并责任到人，从而降低风险，增强实体印章的风控管理水平，实现网点“合规化”建设。

4.2 盖章流程专人审批实现网点用印“规范化”

印章用印专人审批，相当于对用印过程增加安全管控环节，该环节通过在线上发起流程审批，使用印信息公开透明化，可以分散用印权力，降低违规用印风险，增强营业网点用印的“规范化”管理。

4.3 印章全周期监控实现网点管理“精细化”

印章全生命周期管理，是对印章的刻制、领用、保管、使用、上缴和销毁过程进行全面的闭环管理。对印章的线上申请、线下使用进行全面的管控，做到线上申请有记录，线下使用可追溯，对文件用印前进行登记、用印中进行监控、用印后进行存档，确保用印过程中所有可线上查询审计，从而实现网点用印的“精细化”管理。

4.4 用印数据全面监控实现网点管控“数字化”

证券公司在面对诸多营业网点时，在传统管理模式下，无法对营业网点的用印合规性、用印审批过程、用印量等情况全面了解，而印章物联管理系统则实现了印章的集中高效管控，在用印人员无感的情况下实现了用印过程的详细记录，可在后台快速查询追溯，并且可对印章状态实时掌控，结合监控大屏可查询各类报表数据，实现了营业网点的“数字化”管理。

4.5 印章管理模式创新实现网点运营“智能化”

传统的印章管理模式下印章过于分散化，总部对用印监管时效滞后，合规风险较大，在印章物联管理系统推广后摒弃了传统的印章管理模式，实现了“用印审批分离，人章分离”，构建了行业领先的业务印章管理模式，有效地推进了公司智慧网点建设，提升了网点“智能化”水平。

五、展望

用印安全对于提高公司合规风控能力具有重要的意义，而合规风控能力的提升则是网点智慧运营转型的重要一环，印章物联在证券行业的推广使传统印章管理模式得到了创新，增强了印章管理的可靠性和安全性，但是仍需要持续探索优化潜在的风险。

在未来，可以引入 OCR（智能文本识别）技术，自动将文字提取后与电子文件进行比对，若出现文字改动、空白页能触发自动预警，及时发现潜在的违规操作；增加印控台“人脸 + 指纹”双因素认证解锁模式，使用印安全得到更高保障；也可结合移动 APP 应用，通过蓝牙技术实现手机与章筒直连，可以让员工更方便地进行申请和审批操作，为客户和审核人员提供便利。总之，在这个飞速发展的时代，各种前沿科技都有可能成为助力智慧运营发展的一环，智能印章管理作为提升企业运营效率和风险防控能力的重要手段，将来会更加智能化、便捷化、安全化，而性能日益提升的智能印章将拥有更广阔的舞台。

参考文献：

[1] 杨德胜 . 一种智慧印章管控技术方案及其在电网企业的应用 .
[2] 冯建龙 . 楼天建 . 数字化背景下商业银行“智慧运营”的路径研究

数智驱动：西部证券统一风险管理平台实践

黄裕洋，李海涛，张校，徐国澍，杨登航，张昕妍

西部证券股份有限公司｜E-mail：zhangxiaoo@xbmail.com.cn

摘　要： 证券期货行业行稳致远，需紧抓企业风险管理之舵。在业务开展中，如何建立完善的风险管理体系，及时有效地捕捉、识别、管控风险至关重要。西部证券自2022年启动统一风险管理平台建设，通过大数据、AI、RPA等技术手段，打破各业务系统的数据孤岛，构建以自营投资及做市、信用业务为主的业务全覆盖的全面风险管理体系。通过业务与技术协同建设，已初步建成一套行之有效的数智化风险管理平台。

关键词： 数智驱动；全面风险；统一风控

一、背景

当前背景下，金融机构面临着风险事件多样化、复杂化的问题，要求市场机构在金融科技领域不断探索创新以应对风险的多变性和管理的即时性。此外，监管机构也陆续发布若干全面风险管理相关的规定，为严格遵守行业监管自律要求及提升内部风险管理能力，全面落实《证券公司风险控制指标管理办法》《证券公司全面风险管理规范》等要求，建立完善的证券公司及子公司风险管理体系，强化风险管理系统的建设是证券期货行业风险管理工作的重中之重。

西部证券股份有限公司（以下简称：西部证券）2022年依托公司风险管理部和数字化转型办公室为主体，各业务部门协同，成立专项工作小组，启动全自研的统一风险管理平台（以下简称：统一风控平台）建设，以此来代替分散的外采系统，旨在通过大数据、AI（Artificial Intelligence，人工智能）、RPA（Robotic Process Automation，机器人流程自动化）等技术手段，打破数据孤岛，构建以自营投资及做市、财富信用为主的全覆盖的风险平台，提高公司整体业务风险管理水平，打造差异化的业务风险管理体系，提升公司全面风险管理的数字化和智能化水平。

二、整体架构

统一风控平台作为西部证券风险管理业务和技术双向奔赴创新的典型实践案例，致力于解决公司日益复杂的风险管理问题。

在业务架构上，覆盖公司自营投资及做市、财富信用、投资银行、资产管理、机构业务等五大业务板块以及子公司业务，以数据为素驱动各业务实现事中、事后的数字化管理。在专项风险管理领域，覆盖市场风险、信用风险、操作风险等，实现公司全面风险管理。

在数据架构上，打造分层的数据体系，包括ODS（Operational Data Store）贴源层、DR（Data Risk）风险业务数据层、DM（Data Mart）数据集市层等。利用大数据平台、数字员工（RPA等技术）、AI数据处理能力构建智能化数据采集和加工体系，基于业务场景构建

图1：统一风控平台业务架构

数智大脑，打造数据和风险驱动的 AI 服务体系，实现数智驱动风险管理。

在应用架构上，采用微服务的技术方案，基于业务和综合监控视角，采用模块化的架构设计方案，快速适应不同业务的敏捷迭代，前后端分离显著提升研发效率和扩展性，利用持续集成和部署和自动化测试等手段提质增效。

2.1 业务架构

根据业务风险管理要求，实现实时盯市、压力测试、业务报表及监管报表等功能，落地风险识别、监测、预警体系。以实时盯市为例，该模块应用于融资融券维持担保比例的实时监控、股票质押式回购融资人实时履约保障比例的监控，以及证券投资实时盈亏的止盈止损预警等场景。

在业务风险管理基础上，以信用风险、市场风险为突破点，发力全面风险管理体系建设。综合指标监控以内外规要求为切入点，按使用场景分为外部监管类指标、内部风险容忍度及限额等指标，以及面向特定事项的专项指标。

2.2 数据架构

建设统一数据集市，实现公司风险数据的集中化和标准化管理。

图2：统一风控平台数据架构

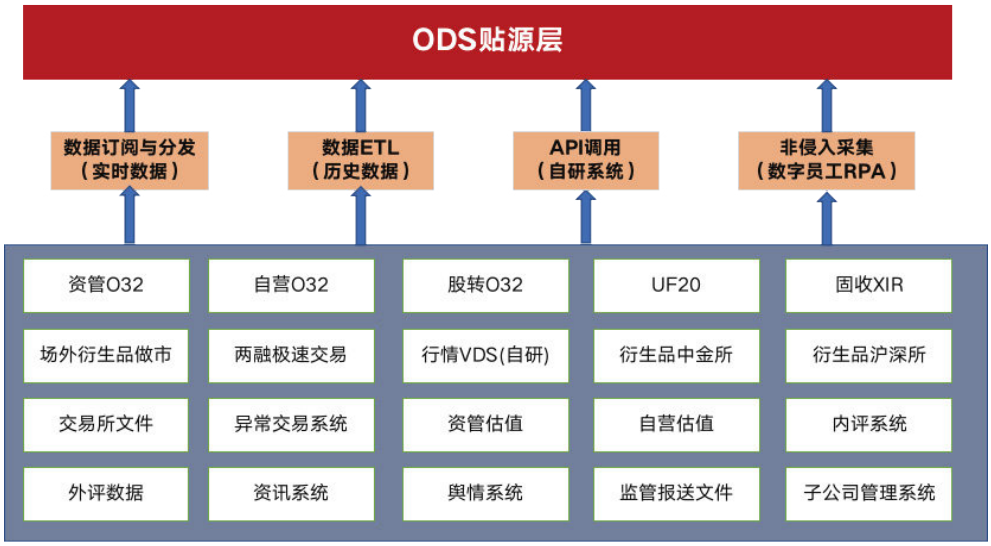


图3：统一风控平台ODS贴源层数据处理

2.2.1 精细化分层：数据模型

针对上游数据源分散、业务逻辑各异的问题，采用分层数据模型设计，提升系统扩展性和维护性，促进数据共享，有效支持实现公司各类业务风险管理建设。

采用数据批量同步、实时分发、RPA 对接等多种技术，解决上游数据分散导致的数据孤岛问题，按照统一标准将业务数据汇聚于 ODS 贴源层，实现对全业务数据的整合。基于数据模型分层设计理念，以业务为导向，通过统一建模、集中清洗、标准化指标模型体系等方式将各种来源的业务数据进行综合和提炼，打造符合研发和业务人员使用习惯的 DR 层。

借助自研的低代码体系搭建风控业务参数管理模块，动态管理 DR 数据明细层中模型转换参数，实现集中管控不同系统中的业务常量定义，增强平台数据处理灵活性。

DM 数据集市层作为数据模型的应用层，为平台的风险管理功能、风险计量引擎和智能化服务中心提供关键数据支撑。

2.2.2 自动化协同：数字员工

数字员工以非侵入的方式，通过模拟员工操作行为，实现平台外延，在数据采集与外发环节提升与外部系统

协同能力，具体包括：

非侵入的数据采集服务：通过 RPA 技术，将部分底层数据黑盒的系统，通过界面操作的方式进行数据采集，以提升数据覆盖范围。如：使用 RPA 打通主体池与慧眼 x-insight 风险智能监测系统，将预定义的事项向统一风控平台同步。

分散数据的预规整：自营业务定期报告中不同交易品种的期初期末数据分散在不同账套中。通过 RPA 技术将衡泰系统中不同账套数据文件中的内容进行定向抽取与汇总，合并成数据汇总表后向统一风控平台填报。

风控关键参数的多系统同步：风控系统自身会维护业务参数，如可投主体的黑、灰、白池。该部分业务参数的实际使用会分散在不同的业务系统中，如衡泰 xIR 等。通过 RPA 技术，实现关键业务参数的多系统间同步，确保风控参数在多个不同系统执行的有效性。

自动化报告生成：RPA 定时提取风控数据，按需生成报告，用于经营状态汇报、监管报送、交易对手对账等。如：使用 RPA 技术生成期权、互换对手方估值表，并定向发送邮件。

监管报送报告的提交：监管数据的报送涉及到专岗与专有设备。RPA 技术能够辅助专岗人员实现报送文件的获取与报送流程的执行，并在处理过程中进行截屏留存等处理，确保整个过程可稽核。

2.2.3 智能化碰撞：西部大脑

自研的 AI 中台“西部大脑”通过整合一系列 AI 工具和智能服务能力，以 API(Application Programming Interface，应用程序编程接口) 的形式为平台提供以下的“数智”处理能力：

文档解析服务：基于 OCR(Optical Character Recognition, 光学字符识别)、pdf 结构化抽取等技术，提供 pdf 文档中表格、特定描述段落等内容进行提取。将申请通知书、规范要求文件等关键事项信息进行提取。

要素提取服务：对于文本中的实体要素、事件要素、指标要素、参数要素、情感要素等进行针对性提取，针对文本的特征选择如正则匹配、本地私有化开源模型、分类器模型等 AI 工具进行相匹配的智能化服务。

参数管理服务：通过自研低代码模块，提供统一的业务参数配置服务，实现一处设置多处应用，提升参数的配置效率。在模型探索、模型生产上线环节中，提供业务参数的一致性保障。

2.2.4 业务化驱动：业技协同

业务驱动数据模型设计，借助数据可视化引擎，助力风险管理人员自主进行风险监控和风险计量，推动风险管理从传统模式向数据驱动的模式转变。

在数据明细层（DR）的设计中，以业务视角实现数据模型设计，持续优化设计以适应业务。通过精简关键业务属性、消除歧义字段、沉淀高频字段，构建简洁、高效、易于理解的模型。基于 DR 数据明细层进行研发，降低研发难度，让业务人员专注于风险业务的研究。

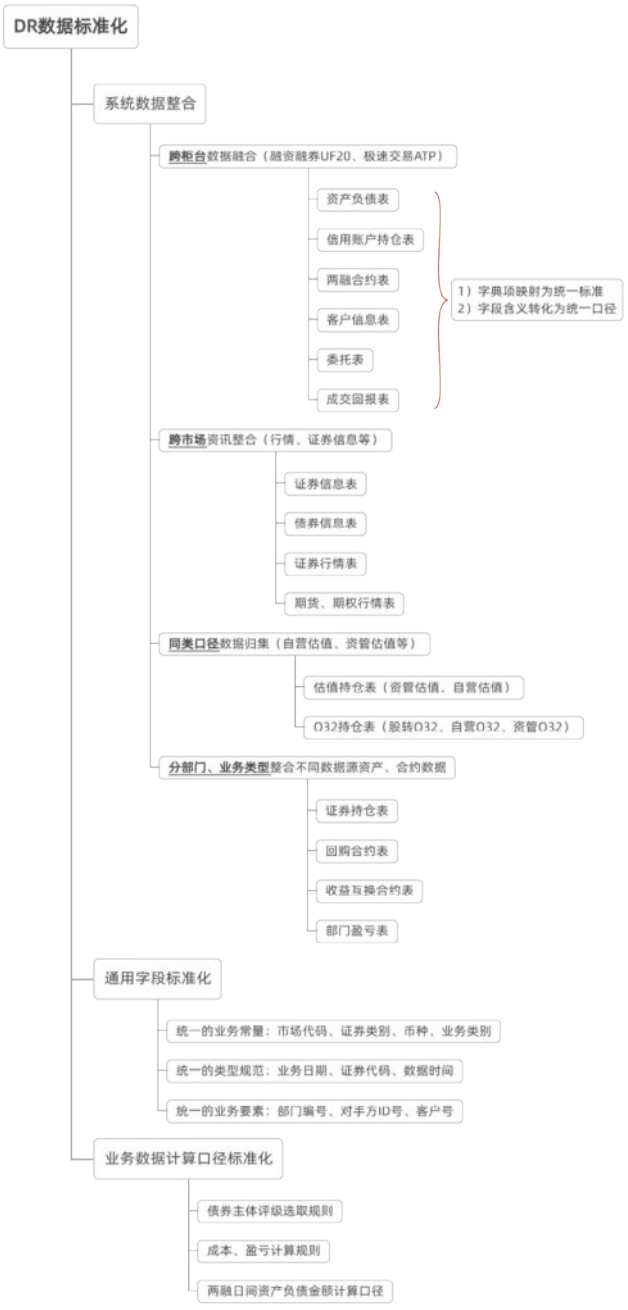


图4：统一风控平台DR数据标准化内容（部分）

平台搭建 DR 风险可视化引擎，基于平台规则引擎可以自动补全 DR 数据明细层表间的字段关联逻辑。风险管理人员能够通过直观的操作界面，自主提取数据进行分析，实现风险管理的敏捷化。

图5：统一风控平台DR市场代码转换参数维护页面

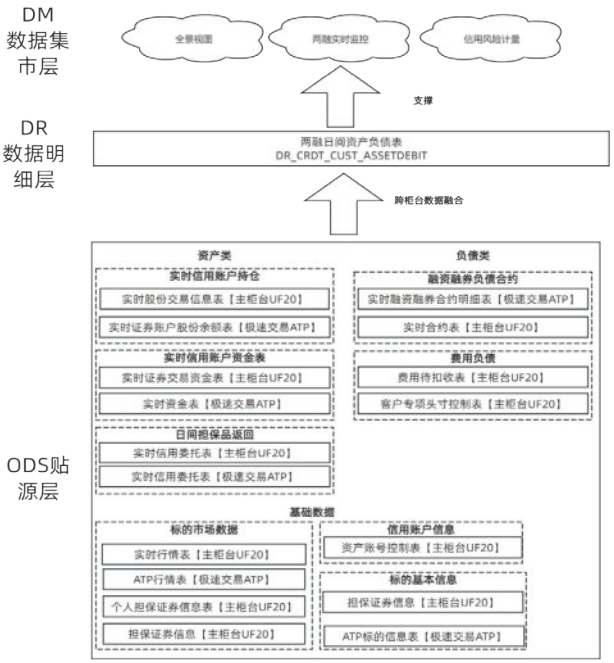


图7：统一风控平台DR层实时两融资产负债表数据逻辑

2.3 应用架构

应用层采用主流的微服务技术，在设计层面注重用户体验和操作的便捷性，在架构上应用系统分前端表现层、后端服务层、数据底座三层。

前端表现层则为用户提供一个直观、易用的界面。为用户提供全景大屏、业务概览、基础查询等丰富的功能。

后端服务层提供包括应用端任务调度、网关转发服务、认证授权服务和系统监控服务等多种服务。

在数据层使用关系型数据库及大数据存储引擎存储业务及管理数据，如业务信息、中间表信息、用户部门信息、角色菜单权限等，使用缓存中间件处理高频数据缓存和登录授权校验等安全相关数据，以解决数据的快速访问，保障数据安全。同时，借助文档解析、实体识别、要素提取、超参优化等前沿技术，构建能够处理复杂场景数据的平台底座。

组件与中间件层集成一系列高效的工具和服务，以确保

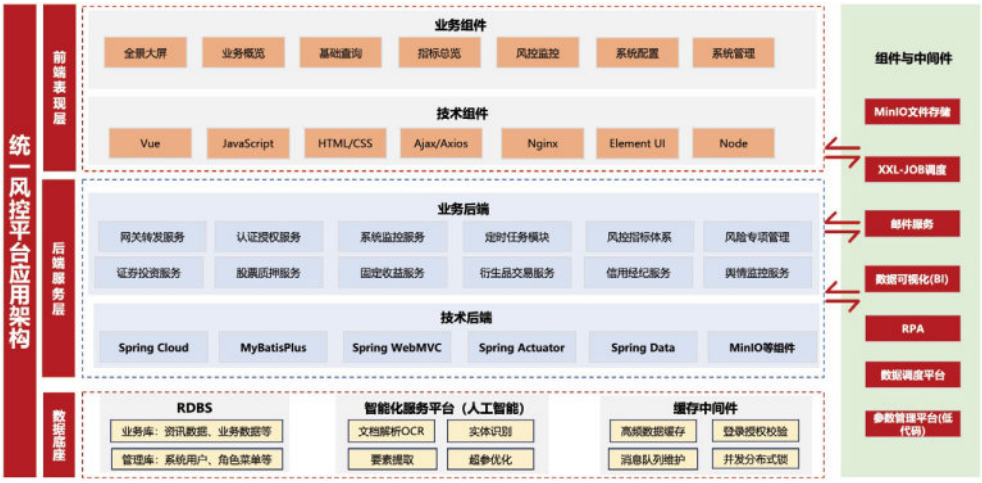


图8：统一风控平台应用架构

系统的高性能和可扩展性。集成 MinIO 文件存储组件，支持大规模数据存储需求；通过 XXL-JOB，平台能够自动化执行周期性的风险评估和监控任务；通过邮件服务已实现风险日报或风险预警信息推送。

三、关键技术与创新点

构建一个集大数据、人工智能、RPA 等前沿技术于一体的平台，紧贴业务要求，差异化利用指标配置、低代码、AI、RPA 等技术打造核心场景支撑能力。

3.1 借助指标引擎：构建风险全景

为敏捷响应业务变化和风险管理要求，构建统一的风控指标库，对风控指标集中化管理，搭建配置化的指标引擎，将指标的计算口径通过脚本、表达式、函数等方式呈现，引擎根据触发器自动计算风险指标。

指标引擎支持多维度计算、链路分析、明细下钻联动功能。引擎支持公司整体、资金账户、证券代码等明细指标计算，支持实时、历史的回溯计算，兼顾公司级综合监控和业务实时明细盯市需要。通过解析指标间依赖关系，引擎实现指标数据血缘依赖的自动分析。引擎设置多种不同样式的下钻明细数据可视化模块，用户可以根据指标属性定制每个指标专属的指标下钻界面。

同时，指标引擎搭配一站式指标配置工具，通过系统固化指标配置方式，简化指标配置流程，通过简单操作即可更新指标逻辑。

3.2 融合 RPA 技术：自动化提质增效

RPA 技术通过模拟人类用户的操作，自动化执行繁琐、重复且高度程式化的任务，如数据收集、报告生成及合规性检查等，显著提高风险管理的效率的同时降低操作风险。借助 RPA 技术，平台全天候不间断地监控业务流程，及时发现和报告异常行为，确保风险监控的连续性和实时性。

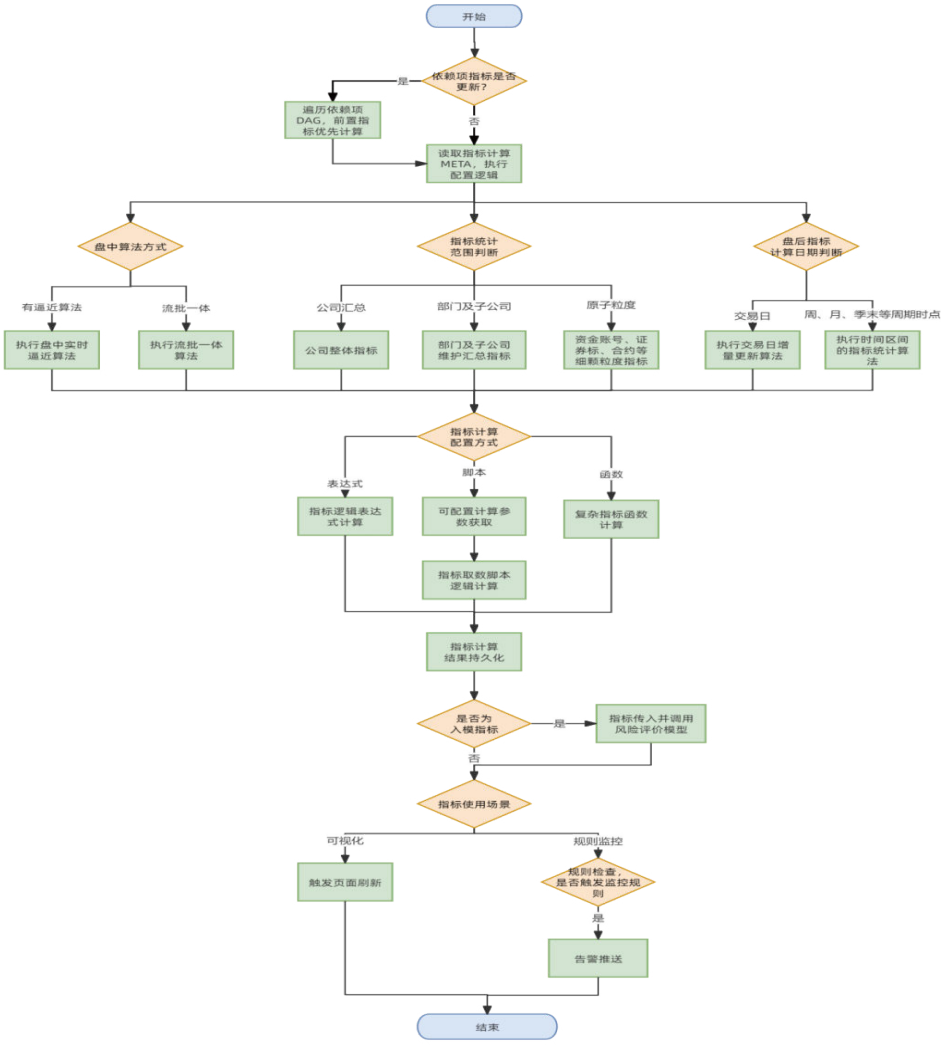


图9：统一风控平台指标引擎处理流程图

图10：统一风控平台指标监控下钻及联动展示
（数据均非真实数据）

图11：统一风控平台指标配置工具

风险管理系统会有大量数据需要从外部系统获取后导入至平台，借助 RPA 技术能够与现有的 IT 系统无缝集成，实现跨系统的自动化操作。随着 RPA 技术在此类场景中的逐步使用，不仅提高数据的一致性和可访问性，而且大大地减轻人工负担，减少人为错误，使风险管理团队能够更专注于策略制定与复杂问题的分析。

3.3 自研低代码引擎：敏捷实现各类功能

搭建基于 SQL 的配置页面功能，提升开发效率和灵活性，常见的数据查询、多 TAB 切换的查询、下钻详情查询等均可通过 SQL 配置实现。在数据呈现上可以设置

数据的显示格式，包括但不限于千分位符、小数位数、日期格式化等，以匹配不同的业务场景和用户偏好。

借助 SQL 配置，系统快速实现大量风险业务功能，涵盖融资融券、自营投资、资产管理、子公司等业务。以衍生品持仓详情、持仓概览、交易流水、业务盈亏功能开发为例，通过 SQL 配置，3 小时即可实现 15 个功能的应用端研发工作，研发效能提升 10 倍以上。

3.4 分布式微服务技术：保障系统高度协同

采用分布式微服务架构，通过服务发现与注册、服务熔断、服务网关、配置中心、链路追踪和负载均衡等技术，实现服务的松耦合、高可用性和弹性。

Nacos 集群作为服务注册和配置管理的核心，通过网关集群来处理外部请求，实现高效的负载均衡和安全认证。业务服务作为平台的关键部分，实现包括信用、自营、资产管理、子公司等核心服务，服务集群通过 Feign 客户端进行通信，以支持复杂的业务逻辑。

在保障系统稳定性和安全性方面，平台集成认证中心和分布式任务调度系统，保证任务顺利执行和用户访问的安全性。数据层采用包括多种数据库技术，结合缓存服务和分布式文件存储，以支持多场景的数据处理和存储。此外，数据调度中心的集成，为平台提供强大的数据处理能力，而集成的企业通讯服务则加强内部信息流通。

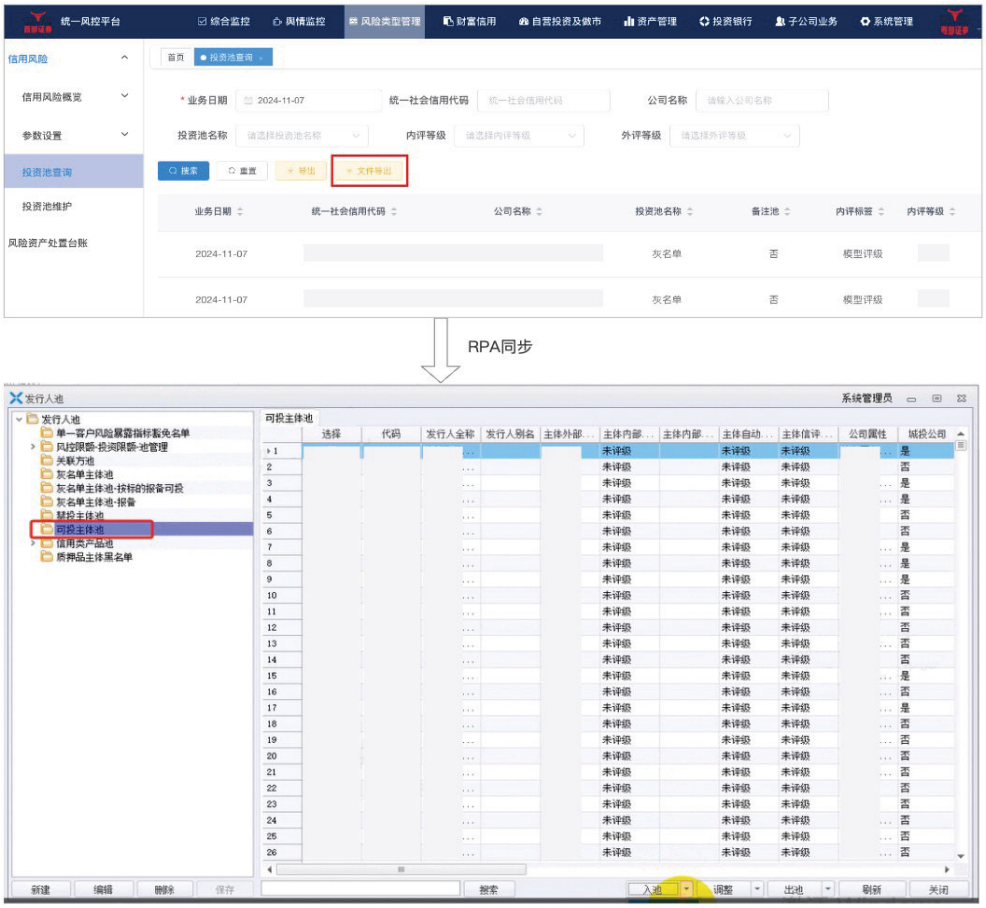


图12：统一风控平台通过RPA向衡泰xIR系统同步主体池

四、平台效果

统一风控平台的建设标志着西部证券在风险管理的数字化和智能化方面迈出了关键一步。

4.1 搭建公司业务风险管理数据主题

在业务风险体系建设中，将客户、持仓、交易、估值以及外部资讯等数据接入平台，融合风险管理场景，按照统一的标准建立风险管理数据集。已完成财富板块的融资融券、转融通、股票式质押，自营板块的证券投资、固定收益、衍生品业务的覆盖，包括客户基本信息、资产负债信息、持仓信息、交易信息、估值信息、关键外部资讯信息等，建立业务数据主题模型 900 余个，为业务风险的系统化管理形成清晰、易用的风险数据积累。

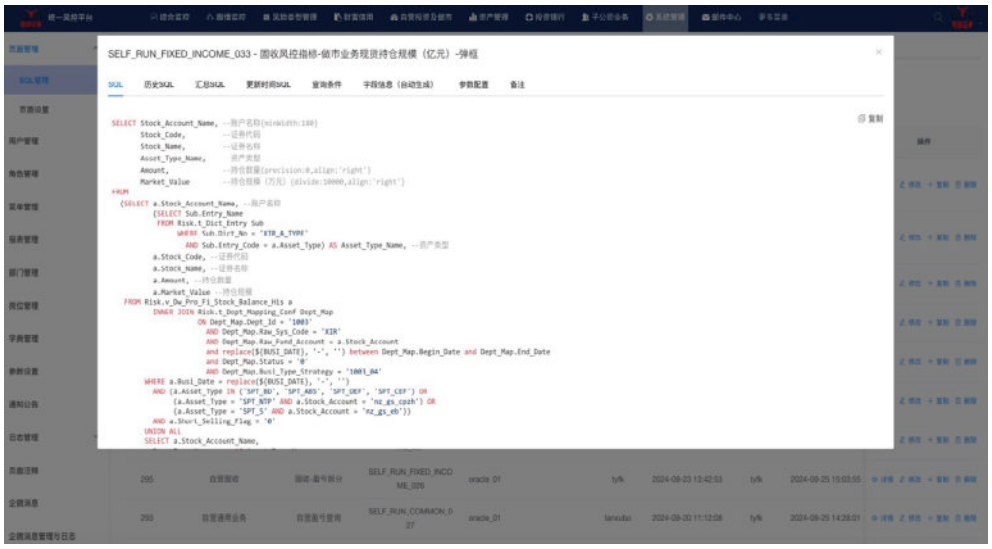


图13：系统功能SQL配置示例

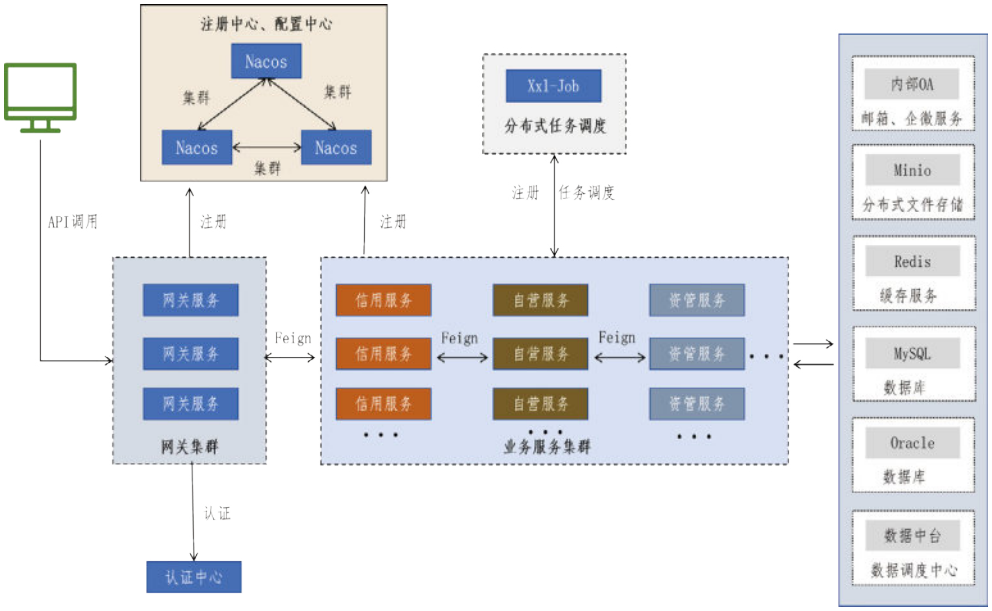


图14：分布式应用架构

4.2 满足业务风险管理时效性要求

时效性是风险管理的关键要素，实时接入融资融券、股票质押式回购、证券投资、固定收益、衍生品等数据，建立和上游业务系统的协同机制，目前已形成稳定的实时、历史数据通道，在业务整合层面，以两融为例，实现跨柜台数据的实时汇聚，有效支持极速柜台两融业务的开展。

在业务风险管理层面，系统形成多样式呈现、多业务场景触达的监控功能体系，以场外衍生品业务为例，通过气泡图直观反映出雪球产品距离敲入、敲出线的分布情况，在出现异常值时可以快速识别并进行风险揭示。另辅以压力测试，支持下钻查询具体合约详细信息，从而实现合约风险快速定位。

4.3 建立公司风险指标监控体系

全面提升公司风险管理制度的系统化支撑能力，依托自研的配置化指标引擎，实现“公司级整体风险容忍度指标”和“部门级风险限额指标”的自动化计算，显著提升风险管理效率。建设风险指标监控中心，支持风控人员自主配置指标，指标明细下钻根据指标计算过程自动生成，指标数据回溯计算。

五、总结和展望

统一风控平台践行了数字化驱动风险管理的理念，为行业提供借鉴的案例，有助于提升行业风险管理数字化水平。

创新驱动的行业赋能：借助平台我司不仅实现了风险管理创新，也为行业提供了案例参考。平台的实践展示了数据驱动风险管理的巨大潜力和应用前景，后期将加大对信用风险、市场风险等全面风险体系的建设，进一步增强风险管理能力。

技术创新的持续驱动：平台采用的多种先进技术，如大数据、RPA 和实时技术等，具有高度的可扩展性和持续创新的潜力。随着技术的不断发展和成熟，平台可以不断引入新的技术和方法，持续提升风险管理能力和效率，保持技术领先和竞争优势。

业务扩展的灵活适配：平台在设计和实施过程中，充分考虑了业务扩展的灵活性和适应性。通过模块化设计和标准化接口，平台可以根据业务和管理的变化，灵活调整和扩展应用范围。

图15：融资融券客户信用资产负债查询功能（数据均非真实数据）

图16：衍生品敲入敲出障碍价格集中度监控（数据均非真实数据）

图17：部门风险限额指标体系

基于大模型的证券业务办理助手探索与实践

赵岩，杨彬，夏杨铭 | 恒生电子股份有限公司 | Email：zhaoyan17768@hundsun.com

摘　要： 近年来，人工智能技术的快速发展正深刻重塑金融行业格局。大参数语言模型和多模态AI技术为金融科技领域提供了强大支撑，优化了传统业务流程，催生了新型金融服务。本文探讨证券金融行业面临的挑战，聚焦如何应对复杂业务流程和高专业门槛。我们将论述如何结合AI工作流程与多个Agent，构建高效业务办理模型，提升效率，降低使用门槛。未来，AI技术将彻底变革金融服务。智能金融助手将提供个性化建议和实时分析，降低从业人员学习成本，提高工作效率和容错率。这一革新不仅让金融服务更普惠高效，也推动了"让金融变得更简单"的愿景实现。

关键词： Agent；大语言模型；MLLM；证券金融；多模态；Transformer

一、引言

Transformer 和大语言模型（LLM）的发展带来了人工智能领域的又一次革新，被认为是第四次工业革命的核心驱动力之一。Transformer 架构的出现被视为自然语言处理领域最重要的突破，不仅提高了模型性能，还改变了处理语言任务的方式。

LLM 技术在金融行业的广泛应用，涵盖了银行、证券、保险等多个领域，为行业智能化转型提供了参考。这些技术为普惠金融提供了新的可能性，使更多人能获得个性化金融服务。恒生电子股份有限公司提出了基于多个 agent 结合、搭载 MLLM 多模态的 AI 工作流探索和实践，为同行提供大模型技术路线和探索思路。

二、基于大模型的业务办理助手

在 AI 行业中，Agent 有三种主流定义：

- **智能助手型 Agent：**基于大语言模型的交互式 AI 系统，如 ChatGPT，能理解和执行用户指令
- **自主决策型 Agent**：具有自主性和决策能力，适用于复杂问题解决，如游戏 AI 或自动驾驶系统
- **多 Agent 系统：**由多个智能 Agent 组成，通过协作或竞争解决复杂问题，如金融市场模拟

本文将探讨如何将传统证券经纪业务通过 AI 工作流和多 Agent 协同完成，使用 MLLM、LLM 和视觉能力处理合规和风险场景。

** 本文聚焦技术发展，不讨论政策合规准则 **

图1：传统业务受办理流程

2.1 场景分析

根据上述场景分析，传统业务流程复杂，几乎全程依赖人工操作。我们不仅期望 Agent 能够如同柜员般高效处理各项流程，并准确理解客户需求，还要求其具备专业的金融知识。因此对流程进行了简化与抽象，如下所示：

图2：新一代业务办理Agent

2.2 技术路线选型

实现一站式服务功能需全面分析用户行为并选择适当技术，以整合业务流程，提升服务质量。从日常的证券公司业务办理中总结用户行为，包括咨询、操作、合规审核等。系统整合了多种 AI 能力，如 LLM 问答、接口调用、意图分析、信息提取、图像分类、语音处理、多模态分析等，有效应对各种复杂证券业务场景，提高服务效率和准确性。

2.3 改善用户交互体验

构建全面协助投资者办理业务的 agent 时，需精准定位具备基本金融知识且明确意图办理业务的投资者。agent 应采用简洁明了的回答方式，优化交互形式，以提升用户体验和业务办理效率。

2.3.1 减少关键 token

图3： 卡片单元技术架构图

图4： 三种常见卡片类示例

优化大模型对话效率的关键是减少 token 使用。主要技术手段包括意图识别、上下文压缩、信息抽取和拆解复杂任务等。此外，还有结构化 prompt 等方法，将在下文详述。

2.3.2 优化对话形式

投资者在证券业务办理过程中，原本是人 - 人的沟通，采用 TTS/ASR 交互模式可以显著提升投资者的对话体验。这种方式不仅使交互更加自然和友好，还能为不便使用文字输入的用户提供便利。其实传统的 TTS 在现如今的场景中应用也很广，但是往往在用户的实际应用端还是给人一种低亲和力，主要原因在于：**1. 缺少上下文的关联 2. 缺少细粒度情感和语言韵律。**

2.3.3 运用卡片式交互

纯对话型大模型虽能提升交互体验，但在处理复杂业务时仍有不足。本文引入创新的卡片式交互 AI 引擎，通过适度约束操作选项，引导用户做出合规、有效决策。这种方法使 LLM 平台能更精准捕捉用户意图和关键信息，协助完成业务办理。此设计简化了 AI 工作流程，减少前端开发工作，提高整体效率。

以上是本文中支撑卡片式交互的技术架构图。基础服务层是由恒生公司多年沉淀的业务系统组成的，向上提供原子功能且包括如下能力：

1）业务组件：将业务组件拆分为细粒度的卡片，支持灵活组合。减少投资者因语言、文字等表述不明确导致的意图识别错误，业务办理阻塞等。

2）动态渲染引擎：基于 LLM 的编码能力，对不同格式的卡片进行二次渲染和加工。

3）数据来源一致：除了大模型加工的对话数据，其他都是以卡片形式面向投资者。其中数据来源都是来自数据交互层，除了 NL2SQL、NL2API 等能力，还配备了自定义 API 数据引擎。

4）对端友好：采用原生 H5 开发，可以嵌入 webview 等终端设备。同时也可以服务于柜面端和 PC。

本文采用多种卡片类型（多选、展示、单选、图标等）构建了高效的业务办理系统。卡片式交互带来了显著的业务价值：快速响应用户需求和减少前端开发工作量。通过业务原子驱动数据生成方法，形成了一个高效的闭环系统。

目前，本项目收录了 50 种卡片类型，支持 20 余种业务流程，将办理效率提升 35%。证明了卡片式交互在金融业务中的优势，为数字化转型提供了强有力的技术支撑。

2.4 快速录入

证券金融业务办理涉及大量材料和档案。为提高资料录入效率和准确性，我们引入了 " 智能分拣 & 录入 " 模块。该模块采用创新的图像文本分类方法，结合 Transformer、end2end 和 LLM 技术，显著优化了传统 OCR+NLP 和视觉算法的不足。

2.4.1 分类模型 | CLIP[1]

传统档案分类模型在处理非标准化文档时效率低下。我们采用基于 OpenAI-CLIP 的创新微调模型，整合图像和文本数据，实现文档内容的深度理解。该模型在处理金融领域复杂格式和专业术语时表现卓越，具备识别文档角度和自动矫正能力。

实际应用中，本系统分类准确率达 97.5%，远超传统 CNN 模型的 85%。关键性能对比如下：

表1：选型数据列表

模型	准确率	召回率	F1 分数	推理时间	训练时间（小时）	GPU 使用（GB）
OpenAI-CLIP	97.5%	96.8%	97.1%	0.15 秒	20	6
CNN(ResNet50)	85.0%	82.5%	83.7%	0.5 秒	30	8
ViT(ViT-B/32)	90.0%	89.0%	89.5%	0.4 秒	25	7
BERT+CNN 组合	92.0%	91.0%	91.5%	0.45 秒	28	8

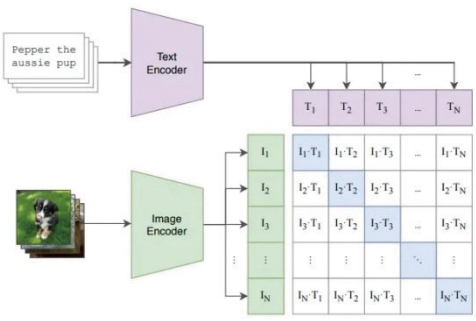


图5：CLIP架构-对比预训练

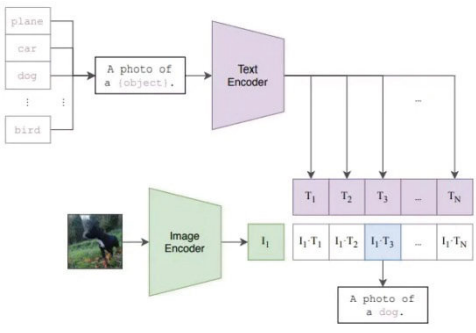


图6：CLIP架构-测试过程

CLIP 模型展现出卓越的迁移学习和 zero-shot 学习能力，无需额外训练即可在各种视觉分类任务中表现优异。本系统在某证券项目中，利用 4 块 NVIDIA A100 GPU 仅用 20 小时完成训练，推理时间低至 0.15 秒 / 图。系统处理大规模数据时表现出色，GPU 内存占用仅 6GB，CPU 使用率 30% 以内。

通过广泛的**预处理、模型训练和全面测试**，我们确保了系统的稳定性、性能和实际应用中的可靠性。

2.4.2 端到端的 OCR + LLM

上文中提到需要一个单元来提取档案信息，有两种方式，一种是通过 NL2API 的方式，另外一种就是表格提取。如果是投资者办理二次业务（非开户），那么 NL2API 或者系统内置流程引擎的情况下，获取数据比较简单。本节主要讲下比较难的表格提取。

智能文档（IDP）早已存在，涉及 OCR、NLP、YOLO 检测和版面表格识别等多学科技术。然而，这种多技术组合在实际应用中存在局限性。证券金融领域主要处理文本表格和协议表格，这些文件格式相似、信息丰富，具有高利用率和敏感性。

本方案采用 end2end 模型算法，适合证券金融行业特殊需求。主要处理表格类档案，虽可解释性较差，但在大数据情况下效果优异。GOT 模型[2] 概述如下：

1）架构：采用编码器 - 解码器范式，高压缩率编码器将图像转为 token，解码器支持长上下文输出。

2）训练：分三阶段：编码器预训练、联合训练和解码器后训练。

3）增强特性：引入细粒度 OCR、动态分辨率策略和多页 OCR 技术。

4）输入输出：输入端支持场景风格和文档风格的图像；在输出端，则能通过简单的提示生成纯文本或格式化结果（如 Markdown、TikZ、SMILES 等）。

图7：分类模型示例图

图8：证券金融业务场景-智能提取

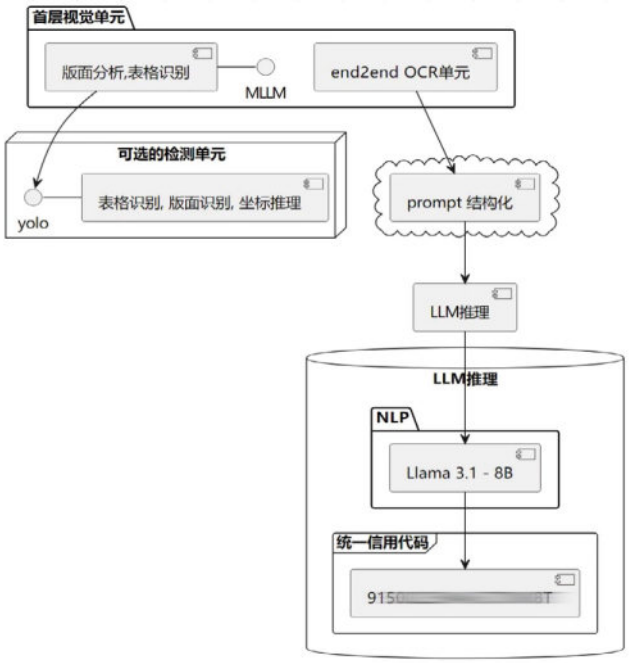


图9：提取单元逻辑图

上图是我们这个单元具体的实现逻辑，其中end2end OCR 识别单元，需要自己准备训练数据，本系统，大概 30 万以上的数据训练出效果就会非常好。无论是版面识别或 OCR 识别准确率都可以达到商用级别。

其中 LLM 推理，采用了 Llama3.1-8B 的模型，在选型过程中针对其推理效率、能力、表格和自然语言对话能力进行了对比，最终选择了该模型。但在实际应用中需要自行训练中文偏好模型。

2.5 实时合规审核

上述步骤实现了所有录入的内容，完成了整体解决方案的一半，因为在证券金融行业，除了信息和档案的有效录入，还有风险合规的监管要求。本文中调研了相关行业内人员和业务系统可提供的指标，整理出大致如下几种：

1) 人像识别：头像中、双录视频中的人脸需要和证件相符，确认是本人办理业务

2) 环境识别：双录视频中的背景，如：需要背景中有证券公司 LOGO

3) 特征识别: 人物穿着，年龄评估，性别识别。例如: 开户职业选择公职人员，但是穿着差异很大的需要警示。特别是性别不符的需要警示

4) 音色识别：投资者在录制时，不能出现其他人的声音进行引导、干预

5) 姿态评估：投资者在观看投教时，需专注屏幕内观看投教视频

6) 注意力评估：投资者在观看投教视频，听投教话术和风险揭示时，需要保持注意力。这里可以引入一个注意力的量化模型，用户注意力系数的判断：结合面部特征和时序信息，利用多模态分析技术评估用户在视频中的注意力水平。例如，通过分析眨眼频率、头部姿态和面部表情变化以及是否接打电话。来推断用户的专注度。

7) 影像资料评估：材料中手写字体是否本人，是否含有印章且清晰

图10：MLLM效果（注：该身份证图片为合成示意图，其中身份信息非真实信息）

8) 攻击性评估：所提供材料是否虚拟构成

2.5.1 为什么是多模态 MLLM

相比于传统的多模态模型，本系统引入一个 MLLM 模型来做视觉方面的审核，主要解决的问题：

1) 视觉组件设计不足：当前的多模态模型在视觉组件方面往往探索不够深入，导致难以准确地将视觉信息与现实世界的场景联系起来。不能做到真正像营业部柜员、审核人员一样去审视图片、视频和视觉信息。

2) 基准测试的局限性：现有的多模态模型基准测试存在一些问题，如结果整合和解释困难，因此需要一个更加注重视觉任务的新基准。

图 10 是 MLLM 的效果，可以针对一些常规照片进行识别，并且根据一些特征断定风险，可能非公职人员，以及证件的类别和住址等信息。效果相对较好，如果是双录视频信息，可以采用 ffmpeg 采集分帧图片的方式，送入 MLLM 进行要素评估。

2.6 持续提升模型能力

在证券金融行业，多模态大语言模型（MLLM）可以极大地提升数据分析、风险管理、客户服务等多方面的能力。关键在于如何有效利用大量的训练数据，将其转化为持续提升 MLLM 性能与实用性的动力。

2.6.1 数据准备与清洗

1) 高质量数据集：确保数据集的质量，包括准确性、完整性和一致性。相对来说证券公司的数据会更加标准和合规，相对数据质量很高。

2) 多样化数据源：结合文本（如基金报告、公司章程）、图像（如图表、图形）、视频（如人物模态分析）等多种类型的数据，以增强模型的多模态理解能力。

2.6.2 模型架构优化

1) 连接器设计：借鉴 Cambrian-1 中的空间视觉聚合器（SVA）等技术，设计高效的连接器来整合不同模态的信息。

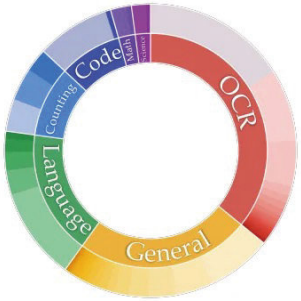
2) 自适应学习：引入自适应学习机制，使模型能够根据不同场景自动调整其权重和参数。

2.6.3 训练策略

1) 分阶段训练：采用两阶段训练策略，首先对连接器进行预训练，然后联合训练整个模型。这样可以提高模型的稳定性和性能。

2) 领域微调：在预训练的基础上，使用金融领域的专业数据进行微调，以提升模型在特定任务上的表现。

3) 强化学习：考虑引入强化学习机制，通过与环境的交互来进一步优化模型的行为和决策能力。



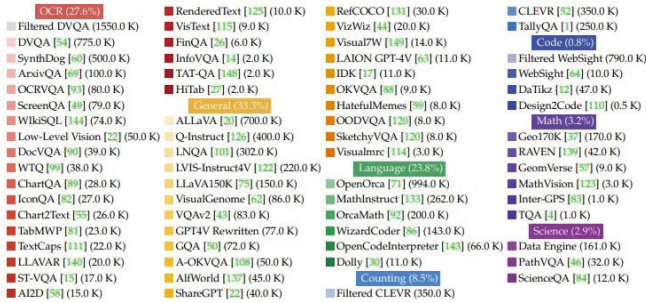


图11：评价指标示例图

上图是 Cambrian-1[3] 在多模态训练评价选择的指标模型，对本系统中可根据分析调整其权重。例如 DocVAQ 被认为本系统中重要评价的一环，原因是大部分流程中的图片信息都包含影像档案等。也可以根据自己数据特征，构建 MLLM 的评价系统。

2.7 知识库的构建

本文介绍了标准业务办理流程，但考虑到投资者知识水平差异，业务办理中需要专业知识问答。通用大语言模型难以满足要求，因缺乏专业金融知识储备且易出现幻觉。解决方案有三种：

2.7.1 prompt 工程

通过调整结构化提示词，使大模型回答更符合场景风格，如模仿专业金融经理人语气。以 CHAT 模型为例，包含 Character（角色）、History（背景）、Ambition（目标）

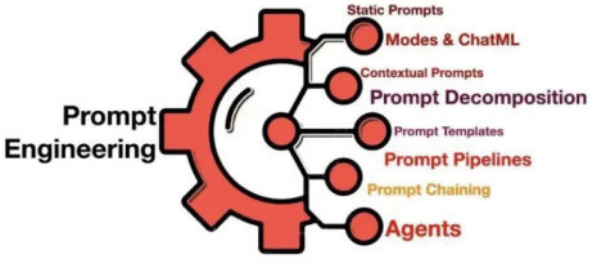


图12：prompt模型

和 Task（行动）四个要素。

通过结构化的 prompt 优化，更明确的得到 LLM 回答，但对提升 agent 专业金融知识影响有限。因此，还需 RAG 知识库和微调来实现目标。

2.7.2 RAG

根据本文描述，RAG 是优选方案。他可以在拥有通用大模型能力的情况下，通过量化数据库创建全新的证券金融知识库。RAG 模型的工作原理可以概括为三个步骤：检索（Retrieval）、增强（Augmentation）和生成（Generation）。

本系统中 RAG 建设架构图如图 13 所示，系统架构分为四层：数据层、模型层、框架层和应用层。可以有

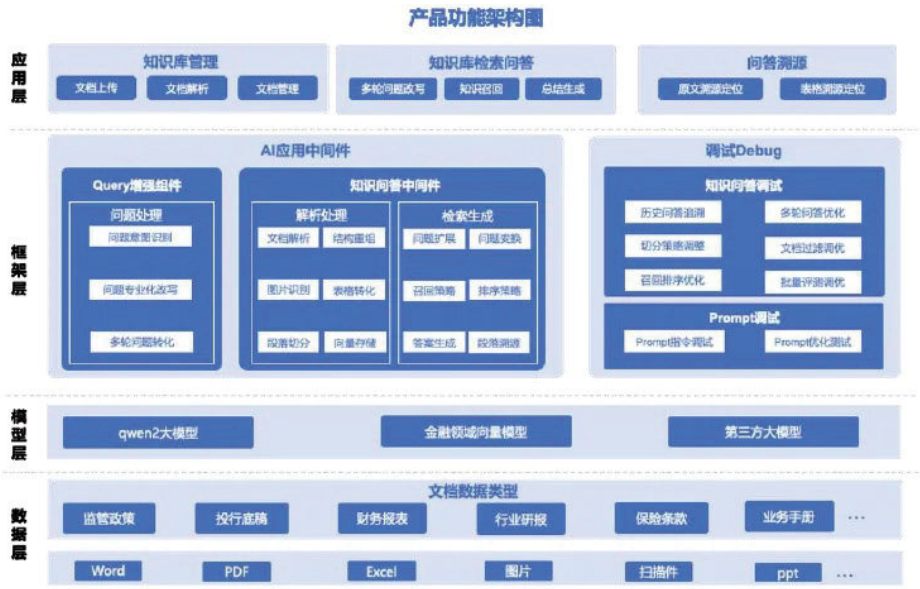


图13：RAG知识库架构图

效地让问答准确性、专业程度更高，避免了大模型幻觉等。

2.7.3 大模型微调

相比于全参数微调的超高成本，本文中采用 LoRA（Low-Rank Adaptation，低秩适配器）这样一种微调方法，主要是把整理的一些常见金融基础知识作为数据，调优我们的预训练大语言模型，例如前面提到的分析端到端 OCR 结果的 Llama3.1-8B 模型，LoRA 通过在原始模型中插入低秩矩阵来适应新的任务，而不需要更新整个模型的权重。这种方法大大减少了微调所需的时间和计算资源，同时仍然能够实现良好的效果提升。

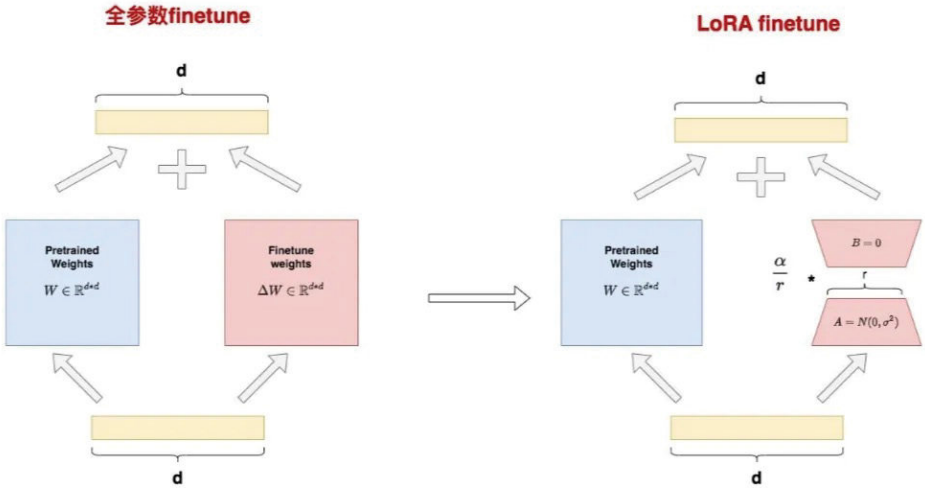


图14：LoRA微调改进示例图

利用 LoRA 方式进行微调，是一个比较简单的方式，大致流程步骤如下：

利用自有数据进行微调的核心关键是数据，本系统中的数据采集互联网证券金融相关网站上的公开数据、恒生公司内部系统操作手册、恒生公司内部业务相关文档、券商客户侧内规文档数据等。

三、结果

在本系统的验证场景中，选择了几个角色：

•金融小白：完全不懂金融证券业务，根据 agent 提示办理业务。没有接触过相关系统

•专业产品经理：具有一定的金融证券知识基础。懂得开户系统的相关知识，但是没有真实为客户办理过业务

•金融投资者：有一定的金融证券知识基础，但是没有操作过系统

•证券公司柜员：采集真实环境为客户办理相关业务的时长数据

以一定采集样本的情况下，可以说明观点：金融小白使用 agent 系统办理相关业务，是可以和专业的柜员相当的，并且办理时长上击败了专业的产品经理和金融投资者。为系统的优化结论和方向提供了证明。说明本 agent 确实可以降低证券金融业务办理的学习成本、并且可以帮助有一定基础的从业人员提高效率。也可以利用 MLLM 等多模态能力提升业务办理的容错能力。

四、合规性与未来趋势

本方案中，总结了一个完整的协助投资者进行金融证券行业的相关业务办理过程。但实际上还有很多未来可以提升的地方。本章节从未来技术发展、合规方面讨论该话题：

表2：实验结果表

	限开通业务 A	权限开通业务 B	信息修改业务 C	开户业务 D
金融小白	4min	6min	3min	21min
专业产品经理	5min	7min	5min	35min
金融投资者	17min	21min	8min	未办理成功
证券公司柜员	5min	5min	4min	15min

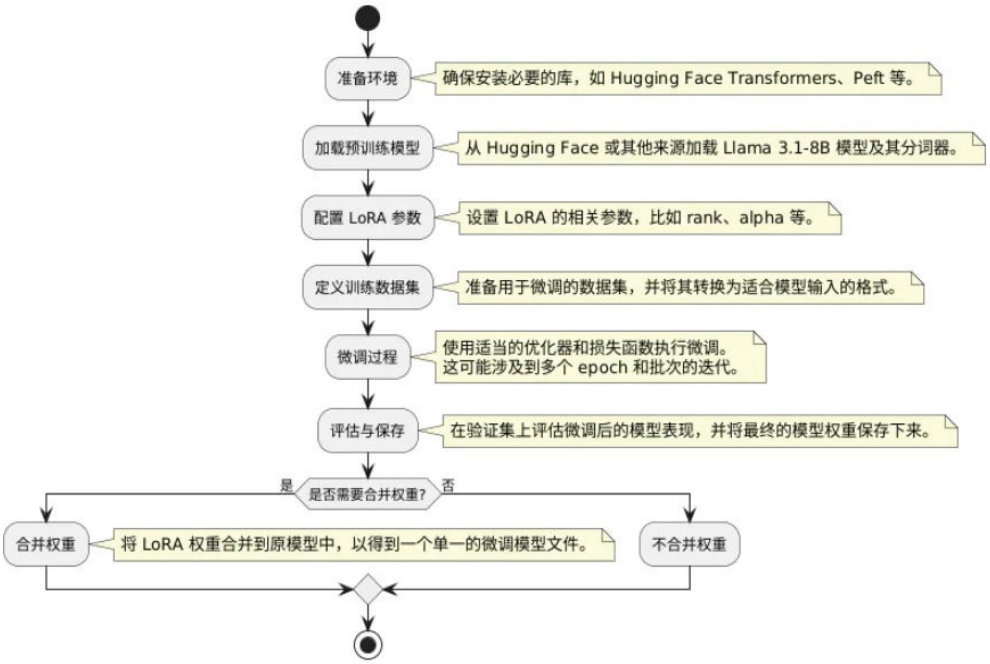


图15：LoRA微调流程图

4.1 合规方向

由于金融行业的特殊性，有强监管和严格要求的特点。上述的 agent 流程到真实的展业中，可以有些特殊业务、特殊群体投资者。但是由于中国金融业发展，未来向数字化、智能化转型的过程中。监管合规方面也一定会围绕着 AI 能力，逐步演进的。故本文中提到的所有要素单元，都可以作为能力单元逐步进入到券商的系统当中。

4.2 视觉模型的发展

首先语言模型在处理句子时，即便是经过了分词处理，每句话的意思仍能较好地保持完整，尤其是在现代神经网络架构下，通过注意力机制等手段加强了词语之间的关联性，使得即使是长距离依赖关系也能得到有效处理。反之视觉模型现如今的技术中，仍然会有 Information Loss 的问题。

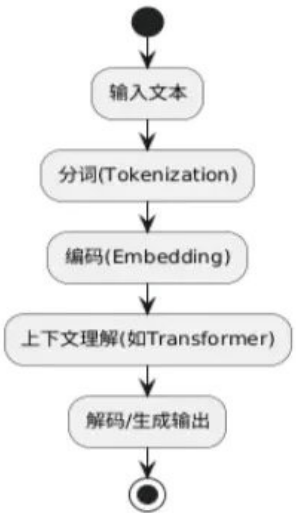


图16：语言模型

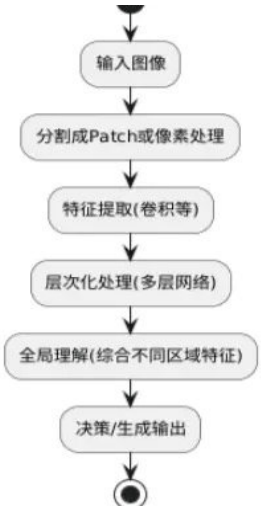


图17：视觉模型

当未来技术发展，能攻克解决这个问题时，金融投资者的服务体验将大幅提升。系统将能整合文本、语音和面部表情，提供个性化智能咨询，精准理解客户需求。身份验证将结合面部和声纹识别，提高安全性并简化流程。投资建议将通过分析交易历史、市场情绪和非言语信号，提供准确的市场洞察和风险管理策略。个性化报告将融合图表、文字和语音摘要，使信息更直观。总之，多模态技术将为金融服务带来无缝智能化体验。

4.3 超长上下文的硬件问题

证券行业建设大模型面临 GPU 硬件资源昂贵的挑战，尤其是多知识库、多 LLM 私有化部署时。随着大模型热度提升，硬件价格水涨船高。

行业内出现新思路，如算法、系统和硬件协同设计（co-design）提高长序列生成效率，解决 LLM 的内存 I/O 问题。通过静态 KV 压缩在 GPU 处理，动态部分在 CPU 存储计算，利用 CPU 内存带宽优势缓解瓶颈，提高整体效率。

技术实现包括静态 KV 压缩（如 H2O 项目）、StreamingLLM[4]、以及 KIVI，LoCoCo 等。未来 LLM 使用成本将降低，通过友好算法模型、协同设计实现模型压缩，或开发精细化领域 LLM 模型服务金融行业。

五、总结

在证券行业，智能化和 AI 赋能一直都是人们话题，为此本文提出一种创新模式通过整合视觉处理、多模态大语言模型（MLLM）和大规模预训练模型，将业务流程中的各个原子功能串联起来，实现更简化、高效且容错性更高的解决方案。利用 CLIP、Transformer 架构及端到端学习等先进技术，该模式革新了文档检测、信息提取和审核流程，应用新技术解决老问题。

同时，基于自然语言处理的大规模模型增强了客户交互体验。证券公司需强化数据治理、建立灵活可扩展的技术架构、培养跨领域人才团队，并持续优化现有方案。以智能业务办理代理作为切入点，将大模型建设纳入 AI 中台战略的核心，不仅短期内提升运营效率和服务质量，也为长期推动行业向数字化、智能化转型打下基础。

参考文献：

[1]Learning Transferable Visual Models From Natural Language Supervision Alec Radford ,Jong Wook Kim, Chris Hallacy
[2]General OCR Theory: Towards OCR-2.0via a Unified End-to-end Model Haoran Wei, Chenglong iu,-Jinyue Chen
[3]Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs Shengbang Tong, Ellis Brown, Penghao Wu
[4]EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS Guangxuan Xiao，Yuandong Tian，Beidi Chen

02 前沿技术应用

基于大模型的数据资产识别方法及应用

苏玎，郭恋，朱一清，苑博，赵泽源，葛青青，刘锦奥，李翔
｜国泰君安证券股份有限公司，华东师范大学｜Email：suding026304@gtjas.com

摘　要： 本文旨在探索基于大语言模型（LLM）的数据资产识别技术。通过构建统一的识别标准，对重点数据资产领域进行了系统化标注，并利用大语言模型进行微调，以实现自动化数据资产识别。研究中采用了多轮迭代优化的策略和机器学习技术，显著提高了数据资产识别的效率和准确率。实验表明，所提出的模型在集团“高价值”数据和个人信息数据领域的识别中表现出色，实现了较高的识别准确率和召回率。研究成果展现了基于大语言模型的数据智能方法在数据治理领域具有广阔的应用前景，能够有效赋能业务应用成效及提高风控合规保障能力。

关键词： 大语言模型；数据治理；数据资产识别；个人信息保护

一、引言

随着数字经济时代的深入发展，数据已成为新型生产要素和国家基础性战略资源，而人工智能技术的不断进步更是为金融行业的数字化转型注入了强大动力。在此背景下，国泰君安坚持“SMART 投行”的全面数字化转型愿景，加快打造数字化转型行业标杆，深入推进全面数字化建设，致力于通过数据赋能和人工智能创新，为公司的高质量发展提供有力支撑。

数据资产的识别是厘清数据内容、挖掘数据价值的重要手段，面向全公司数据的资产化登记不仅有助于提升数据质量、安全和易用性，还有助于合理分配资源，提升数据治理工作效率，并促进数据在全公司范围内高效安全地流转与应用。因此，本文研究拟聚焦集团重点数据资产领域，并基于大语言模型（Large Language Models）技术，探索落地自动化、智能化、高效率数据资产识别的具体技术方案。

1.1 研究内容

在当前项目中，结合相关监管法规指引与司内高价值数据底座建设的背景，重点数据资产将聚焦于能够促进公司业务发展的“高价值数据”以及支撑个人隐私保护的“个人信息数据”两大领域。整体工作遵循“三步走”策略，分为以下三大工作模块逐一展开：

1）标准构建与数据标注：建立一个统一且可落地的评估标准，明确“高价值数据”或“个人信息数据”，基于该标准并结合元数据信息，进行数据集的标注工作。

2）模型训练、评估与优化：结合当前任务场景及数据特性，选择合适的机器学习或深度学习模型来构造一个分类器，落地目标围绕分类效果达标、训练成本可控、预测效率最大化三点展开。

3）场景应用：在多轮迭代优化之后，实现存量数据的资产分类识别以及增量数据的实时预测，并将数据资产识别成果纳入数据资产管理系统，以更好地支撑数据流转、应用与安全保护的数据全生命周期管理。

1.2 问题与挑战

数据处理者在进行数据资产识别任务时，通常应基于如下步骤开展识别工作，具体包括识别需求及目标分析、制定内部识别规则、实施数据资产识别、审核上报目录、动态管理更新。但在海量的数据背景下，基于既定规则开展人工识别的方法面临显著的效率瓶颈，同时需不断兼顾数据的变更和新增问题。为解决上述问题，本项目融合机器学习算法进行数据资产识别任务，综合降低人力成本并在多轮迭代之后实现逐步替代人工识别的效果。

但常规机器学习分类算法对标注数据需求高，处理高维特征的大规模文本数据时面临挑战，难以捕捉复杂关系和深层语义，尤其在语义分析和长期依赖理解上表现不足。因此，本项目采用大规模自然语言模型并进行微调，利用自注意力机制优化长距离依赖的处理，提高文本解析能力，降低对精确标注数据的依赖，减少成本投入。而在实际模型训练中，随着预训练模型参数增大，计算资源成为限制因素。同时，微调和推理效率，以及解决“大模型幻觉”问题，都是高效识别数据资产的关键。本项目针对这些问题提出了解决方案，并解释了常用于“生成式任务”的 LLM 在分类任务中的适配性。

二、标准构建与数据标注

本部分介绍数据标注前的标准构建准则，以及如何基于标准系统化地进行海量数据标注。标准分为高价值数据和个人信息数据的识别制定：前者基于各业务条线的数据资产现状，界定了四大盘点范围；后者遵循法律法规，提炼了个人信息、个人敏感信息、非个人信息三个主类别，并细分为 12 个子类别。

2.1 高价值数据的识别准则

高价值数据资产汇聚是司内高质量数据底座建设的先导项目，而引领高价值数据资产的统一汇聚，实现物理入湖比例显著提升的关键在于能够准确且及时识别什么是高价值数据资产。因此，围绕着实现“统一管理数据”、“打通数据通道”、“保障数据安全”、“确保数据完整”的四大汇聚目标，并结合集团内各业务条线的数据资产现状，界定了如下盘点范围或指引。

1）集市采集：现有数据集市已经按需采集的数据。

2）数据共享：当前已实现共享的数据，包括条线内部、跨条线、母子、母分公司等分享对象。分享方式包括在数据资产平台上共享、在系统内部的用户权限表上共享等。

3）指标计算：涉及关键指标和标签计算的明细数据，该类数据在下游运用中体现了极高的价值。

4）盘点认定：各部门认定的高价值数据，例如采集

图 1：个人信息的定义及整体判定逻辑

表 1：个人信息子类及案例

个人基本资料	个人姓名、生日、性别、民族、国籍、家庭关系、住址、个人电话号码、电子邮件地址等
个人身份信息	身份证、军官证、护照、驾驶证、工作证、出入证、社保卡、居住证等
个人生物识别信息	个人基因、指纹、声纹、掌纹、耳廓、虹膜、面部识别特征等
网络身份标识信息	个人信息主体账号、IP 地址、个人数字证书等
个人健康生理信息	个人因生病医治等产生的相关记录，如病症、住院志、医嘱单、检验报告等
个人教育工作信息	个人职业、职位、工作单位、学历、学位、教育经历、工作经历、培训记录等
个人财产信息	银行账户、鉴别信息（口令）、存款信息、房产信息、信贷记录、征信信息、交易和消费记录等
个人通信信息	通信记录和内容、短信、彩信、电子邮件、以及描述个人通信的数据（通常称为元数据）等
联系人信息	通讯录、好友列表、群列表、电子邮件地址列表等
个人上网记录	指通过日志储存的个人信息主体操作记录，包括网站浏览记录、软件使用记录、点击记录等
个人常用设备信息	指包括硬件序列号、设备 MAC 地址、软件列表、唯一设备识别码等在内的个人常用识别基本情况的信息
个人位置信息	包括行踪轨迹、精准定位信息、住宿信息、经纬度等
其他信息	婚史、宗教信仰、性取向、未公开的违法犯罪记录等

自核心业务系统、可赋能条线关业务、提供前瞻展业帮助的数据等。

2.2 个人信息数据的识别准则

本项目中，个人信息的定义遵循《中华人民共和国个人信息保护法》整体框架与要求，并以《信息安全技术 个人信息安全规范》中的个人信息及个人敏感信息示例为蓝本，同时结合《证券期货业数据分类分级指引》、《金融数据安全 数据安全分级指南》等相关标准，在公司实际数据情况的基础上，进行了个人信息、敏感个人信息及各相关子类的界定。

2.3 训练数据集搭建

在明确数据资产识别标准之后，本节详细介绍了训练数据集的搭建过程，重点关注数据质量的提升和有效标注的实现。最终，构建了覆盖多个业务系统和数据类别的训练数据集，为后续的数据分类和建模提供了坚实的基础。具体的技术链路可见下图所示。

所构建的训练数据整体情况如下所示：对于高价值数据识别，标注了 2 个大类（共计 119514 条数据），覆盖 162 个业务系统；个人信息数据识别总计标注 34897 条数据，分为 3 个大类与 13 个子类，覆盖 142 个业务系统。

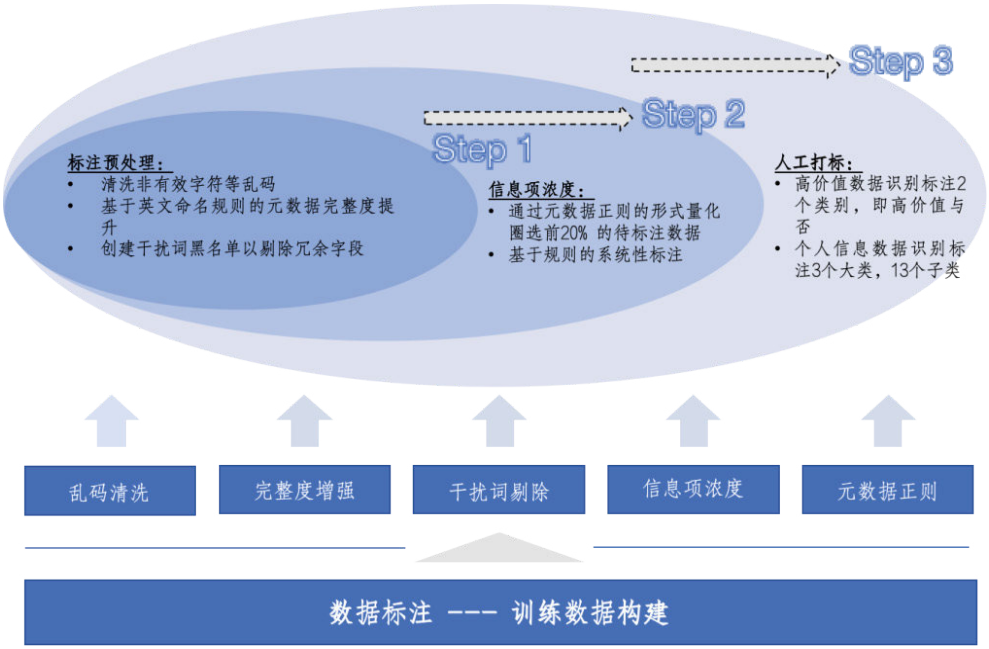


图 2：数据标注整体架构

图 3：训练数据集搭建总览

图 3：训练数据集搭建总览

三、模型训练、评估与优化

本节详细介绍基于大模型技术的智能化数据资产识别框架的设计与实现。该框架旨在高效、准确地识别“高价值数据”和“个人信息数据”，实现在多次迭代的基础上逐步替代人工识别的模式。通过构建通用模型基座，整合了数据处理、模型微调、推理优化和效果评估等各个环节，展示了高效的数据识别能力和良好的泛化性。在效果评估部分，验证了优化技术在提升模型性能和降低成本方面的优势。

3.1 技术框架 - 基于 LLM 的通用模型基座

我们创建了基于 LLM 的通用赋能基座，内在层级包括整合层、训练层、优化层、审核层、推理层、以及应用层。其针对性实现了 LLM 中 prompt 工程的搭建，模型微调架构的实现与加速，推理任务的优化，大模型幻觉问题的缓解，人工及量化评估结果的展示。同时，在整体架构中为进一步考虑对于相似数据治理任务的适用性，允许在完全实现封装的各层级的基础上，仅修改 prompt 工程和 LoRA 参数即可适配相似的分类任务，展现了该基座较好的泛化性。目前，该模型基座于“高价值数据”识别任务中，分类识别准确率已到达 92%。在推理速度层面，以 1000 条数据量为例，推理层实现全量识别约为 8 分钟（推理速度约 3-4 条 / 秒），相较于人工识别近 2 小时的耗时，整体速度提升 15 倍左右。依托于六大层级逐步推进的相互协作，对于数据资产识别任务展现了较高的识别准确率和识别效率。

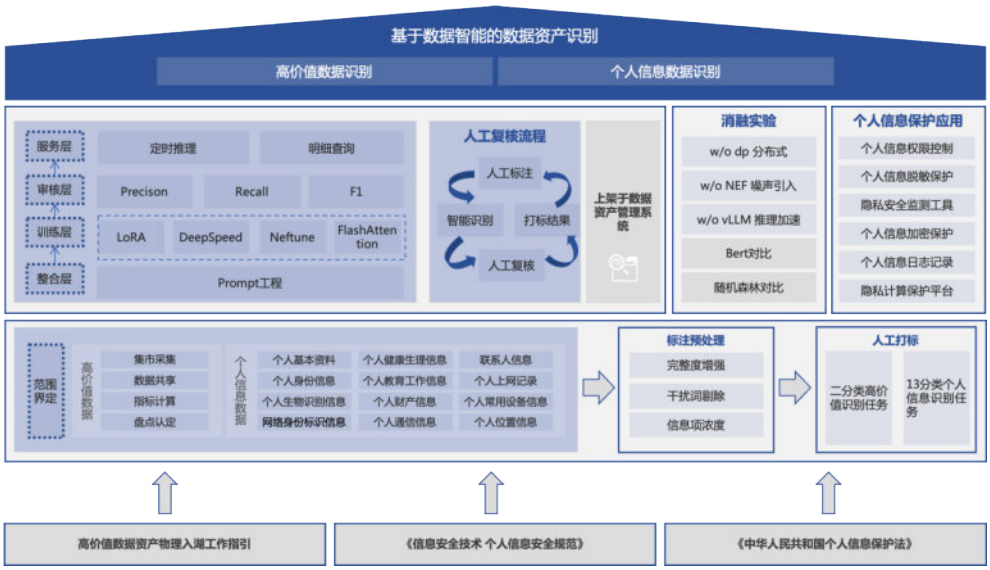


图 4：模型架构图

在下文中，将介绍各层级中主要涉及的技术细节及运作流程，其中整合层中完整的数据处理框架将实现 LLM 自“生成式任务”向“分类任务”的适配性转变；vLLM、DeepSpeed、Flash Attention 三点将针对性解决微调与推理过程中的资源及成本限制；Neftune 技术旨在通过减少模型的过拟合以缓解大模型运用中出现的“幻觉问题”。

3.1.1 数据整合

在该层级，人工标注后的数据将进一步进行整合，以封装成可输入大模型的结构化文本数据格式。其中主要步骤包含：关键语义信息的提取、提示词模板的采样、文本整体构造。

1）表格语义信息提取：将数据标注中收集且标注完成的表格数据提取成文本格式记为 INFO_TEXT。

2）提示词模板采样：为有效降低模型对于提示词的敏感程度，构造了提示词候选集以供每一个样本进行随机采样，采样后的提示词模板记为 PROMPT。

3）文本整体构造：输入文本的构造中，选择性添加示例以提升模型的上下文学习能力，记为 DEMONSTRATE。最终的结构化文本数据格式为 INPUT = {PROMPT} + {INFO_TEXT} + {DEMONSTRATE} (OPTIONAL)。

详细见下图所示：

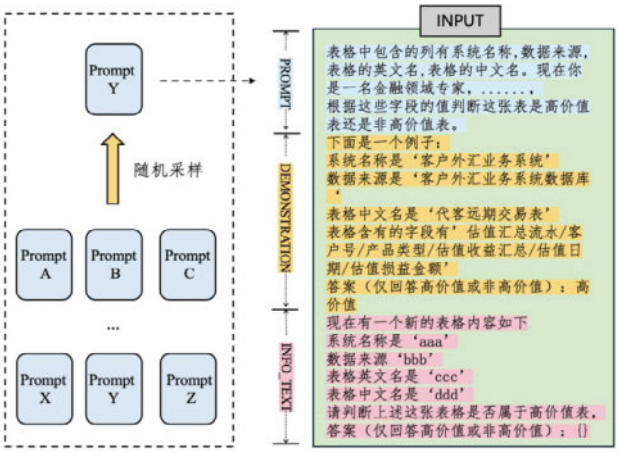


图 5：Prompt工程结构

3.1.2 模型训练及优化

在模型训练阶段采用完全开源的 70 亿参数的通义千问作为基座模型，并使用 LoRA(Low-Rank Adaptation of Large Language Models) 方法在特定任务的数据集上进行参数微调以得到相应的任务驱动型参数。通常 LoRA 参数量仅为预训练模型参数量的 0.8% - 1.5%，这有助于保持整体参数的稀疏性，从而减少计算负担和存储需求。

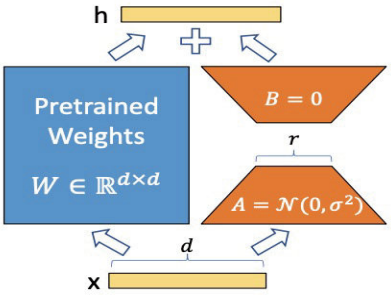


图 6：LoRA微调原理

为进一步控制计算成本以最大化运用计算资源，在训练层中额外使用了 DeepSpeed、FlashAttention 以及 Neftune 等前沿优化技术。其中 DeepSpeed 和 FlashAttention 分别为微软和斯坦福大学开源的优化技术，前者通过分布式框架来加速模型训练和推理，后者通过提高注意力机制的计算效率来加速模型的训练和推理。Neftune 则是在文本向量中引入噪声向量来缓解微调阶段过拟合，从而提升模型的鲁棒性。

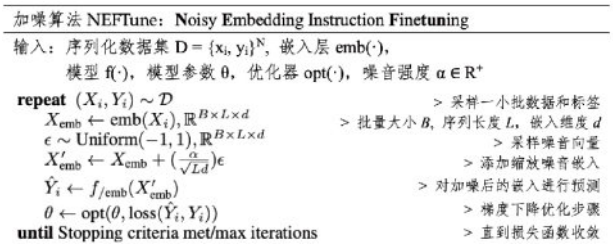


图 7：NEFTune算法

3.1.3 审核及推理

数据整合层和模型训练优化层实现了数据资产识别模型的落地，但在实际应用前，需要全面评估模型效果并确保定期优化。基于此，数据识别模型采用量化指标与人工审核双重评估，量化指标包括精确度、召回率和 F1 分数；人工审核则由业务负责人在模型识别结果推送至系统后进行复核，审核通过后结果同步至元数据管理模块。若审核错误，结果返回训练层进行优化迭代。上述技术架构中各层次的流转链路如下：

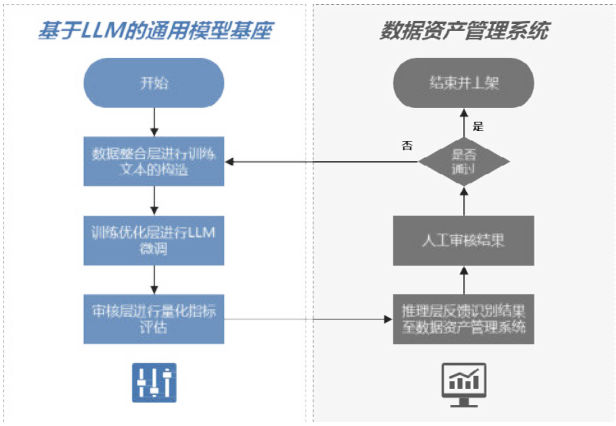


图 8：模型预测结果迭代审核机制

3.2 效果评估与展示 -LLM 可行性及优势

3.2.1 数据集划分

对于高价值数据识别和个人信息数据识别任务，我们分别基于基座模型 Qwen 微调构建了数据识别分类器。在两个任务中，训练集与测试集的比例均大约为 4:1。

3.2.2 模型效果分析

1）模型效果的量化展现：以高价值任务为例，精确度衡量是否有误判，即被预测为高价值的样本中实际为高价值的概率；召回率衡量是否有遗漏，即实际为高价值的样本中被预测为高价值的概率；F1 分数是基于前者的平均判定标准，更能综合衡量在目前数据不平衡场景下模型的综合性能。从下表中可见模型在多个评价指标上都有较好的表现。

2）以个人信息识别任务为例，Qwen、Bert、随机森林模型对比结果如下图所示。Qwen 在小训练集时即达到了较高的 F1 分数，随训练集增大性能提升显著。Bert 需大量数据才达最佳性能，而随机森林的 F1 分数随数据集的提升比较有限，最高约 70%。这一现象说明 LLM 具备更强的语义理解能力，使其能够更好地捕捉和建模复杂的上下文信息。此外，LLM 预训练阶段丰富的“先验知识”使得在下游任务中，即便缺乏大量的标注数据，模型依然表现出色。

表 2：模型效果展示

任务名称	精确度	召回率	F1 分数
高价值判别	92.85%	87.47%	90.08%
个人信息判别	88.21%	86.73%	87.40%

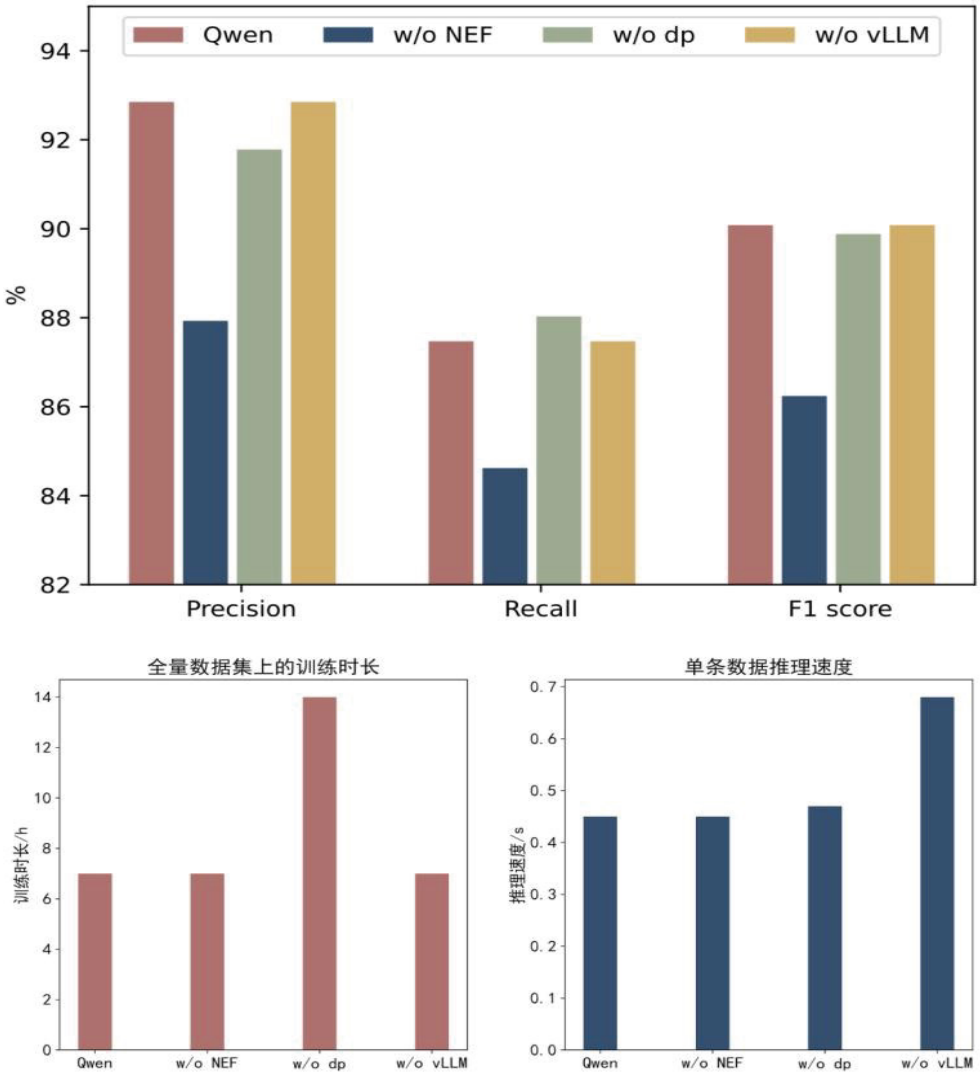


图 10：各维度效果比对展示

3）进一步展示在实际训练过程中针对 LLM 微调所采取的计算成本和推理速度优化措施的效果：我们量化评估了 Neptune、DeepSpeed 和 vLLM 带来的性能提升。具体采用控制变量法设计以下消融实验。Qwen 代表最终模型；w/o（without）NEF 代表不使用 NEFTune 引入噪声；w/o dp 代表不使用 deepspeed 进行分布式训练；w/o vLLM 代表不使用 vLLM 进行加速推理。以高价值任务为例，实验结果如下所示。

整体上，Qwen 始终保持最佳效果，使用 DeepSpeed 和 vLLM 对模型效果影响不大，主要优势在于缩短微调时长和提升推理速度；但不使用 NEFTune 会显著下降模型效果，表明 NEFTune 在数据增强和防止过拟合方面提升了模型性能。图 10 显示，不使用 NEFTune 时，三大评估指标均有明显下降，体现了其在提升模型泛化能力和稳定性方面的关键作用。使用 DeepSpeed 进行分布式训练时，模型训练时长缩短一半，14B 模型需多卡才能运行。未使用 vLLM 时，单条数据推理速度几乎翻倍。整体优化显著提升了性能，解决了数据资产识别中的效率问题，增强了链路稳定性与可靠性。

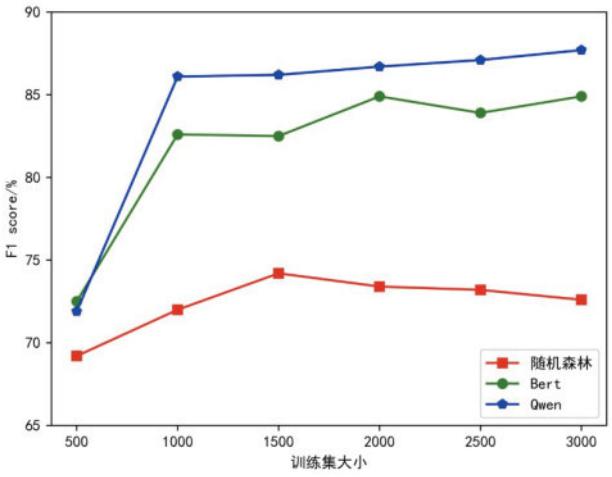


图 9：F1分数 VS 训练集大小

四、场景应用

高价值数据资产识别和个人信息数据资产识别在企业的数据管理和保护中具有至关重要的作用。高价值数据资产的识别和物理入湖，不仅可以打破数据孤岛，构建高质量的数据底座，还能促进数据的高效流转和利用。同时，精准识别个人信息数据资产，有助于企业在遵循法律法规的前提下，有效保护用户隐私，提高数据安全防护能力。

4.1 高价值数据资产识别在数据资产流转中的应用

高价值数据资产识别在数据资产流转中具有重要作用，精准且高效地识别高价值数据资产是构建高质量数据底座的先导项目。通过对应条线高价值数据资产的物理入湖，实现物理入湖比例显著提升以建设高质量数据资产底座。这有助于将公司内外部的数据汇聚在一起，对数据进行重新组织和联接，让数据有清晰的定义和统一的结构，并在尊重数据安全与隐私的前提下，让数据更易获取，最终打破数据孤岛和垄断。通过数据汇聚及其管理，主要可以实现如下目标：

1）统一管理结构化、非结构化数据。将数据视为资产，能够追溯数据的产生者、业务源头以及数据的需求方和消费者等。

2）打通数据供应通道，为数据消费提供丰富的数据原材料、半成品以及成品，满足公司自助分析、数字化运营等不同场景的数据消费需求。

3）确保公司数据完整、一致、共享。监控数据全链路下的各个环节的数据情况，从底层数据存储的角度，诊断数据冗余、重复以及“僵尸”问题，降低数据维护和使用成本。

4）保障数据安全可控。基于数据安全管理策略，利用数据权限控制，通过数据服务封装等技术手段，实现对涉密数据和隐私数据的合法、合规地消费。

图 11：高价值数据识别意义

4.2 个人信息数据资产识别在个人信息保护工作中的应用

个人信息保护外部监管要求日趋严格，随着相关法律法规的不断深化，企业对个人信息保护的重视程度持续提升。个人信息保护需要从数据生命周期的各阶段出发，对数据安全风险进行有效识别，并提升企业自身的管理及技术能力，而个人信息数据资产的高效、精准的识别则是贯穿整个数据生命周期保护的基础。

企业需按照统一的个人信息识别方法，依据自身业务特点对产生、采集、加工、使用或管理的数据进行分类，

并采用规范、明确的方法区分数据的重要性和敏感度差异，根据数据的不同敏感性等级，确定数据在其生命周期的各个环节应采取的数据安全防护策略和管控措施，进而提高企业的数据管理和安全防护水平，确保个人信息的完整性、保密性和可用性。

同时，在个人信息处理活动中，诸多合规性要求亦需要以个人信息的精准识别作为基础。例如，依据国家法律法规要求，在企业涉及处理敏感个人信息、利用个人信息进行自动化决策、委托 / 提供 / 公开处理个人信息、向境外提供个人信息等场景时，应当事先进行个人信息保护影响评估。

五、总结与展望

通过对高价值数据资产及个人信息识别的应用，我们认为基于 LLM 的数据智能方法在数据治理领域具有广阔的应用前景，能够有效赋能业务应用成效及提高风控合规保障能力。高价值数据资产识别模型的研究旨在提供一个智能化、自动化、可解释、高效率的方法识别高价值数据资产，促进高价值数据资产汇聚、集成及应用工作，对实现集团内数据“有、存、通、管、用”全数据价值链的高效运作，发挥数据资产的核心要素作用具有重要作用。基于 LLM 的个人信息识别有助于提升公司对个人隐私数据的理解与认知，精细化把控数据保护需求，提升公司对个人隐私数据的分析与管理能力，有助于确定更细粒度的个人隐私数据安全控制策略及措施。

此外，当前项目在技术架构上仍有一定改进空间，主要分为更广泛的预训练模型探索，以及通过引入检索增强生成技术 (RAG) 元数据的完整性与准确性。基于项目中实际的数据现状，对于由英文字母缩写或拼音缩写构成的元数据信息，可构建一个中英文双语的对照表，并使用 langchain 的离线技术将其转为向量数据库。引入 RAG 可有效提升特定场景下大模型的应用效果，提升 LLM 整体的适配性和高效性，进一步提升模型总体性能。

参考文献：

[1]Dao, T., Fu, D. Y., Ermon, S., et al. (2023). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv. Available at: https://arxiv.org/abs/2205.14135

[2]Jain, N., et al. (2023). NEFTune: Noisy Embedding Instruction Fine Tuning. arXiv. Available at: https://arxiv.org/abs/2310.05914

[3]Kwon, W., Zhu, et al. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv. Available at: https://arxiv.org/abs/2309.06180

[4]Rajbhandari, S., Rasley, J., Ruwase, O., He, Y. (2020). ZeRO: Memory Optimization Towards Training A Trillion Parameter Models. arXiv. Available at: https://arxiv.org/abs/1910.02054

[5] 平安银行智能化数据安全分类分级实践分享 .(2023). https://www.secrss.com/articles/65669. 2023

[6] 兴业银行数据安全分类分级实践 . 中国电子银行网 . https://www.cebnet.com.cn/20240617/102957776.html). 2023

人工智能驱动的知识中台在证券行业的应用探索

潘建东，马张晖，王赵鹏，刘国杨，尹序鑫，孙冰，訾顺遥，梁彬

中信建投证券股份有限公司｜Email：mazhanghui@csc.com.cn

摘　要： 党的二十大报告提出的完善分配制度和规范财富积累的新要求，强调了财富管理在提升居民财产性收入和效率方面的作用，以及在增加低收入者收入和扩大中等收入群体中的重要职责。我国证券财富管理业务的发展目标转向为广大民众提供服务，而非仅针对富裕阶层，面临的挑战是如何高效、规模化地提供综合、定制化和个性化的服务。本文针对我国证券行业综合财富管理业务的问题和挑战，探讨了人工智能技术支撑下的知识中台特点及其架构设计，详细介绍了中信建投证券在该领域的实践案例，展示了其在促进业务发展中的实际效益和潜力。

关键词： 综合财富管理；人工智能；知识中台；知识生产；知识应用

一、引言

财富管理起源于近百年前，起初是为高净值个人提供区别于普通客户的服务。金融巨头如高盛和摩根士丹利奠定了这一行业的基础。随着业务发展，财富管理形成了综合模式，通过线上线下沟通全面了解客户需求，提供个性化的投资管理和财务规划。服务内容广泛，包括投资分析、财务规划、法律协助等，核心是以客户为中心的定制化服务，如图 1 所示。

当前，中国财富管理行业面临转型挑战，需提升服务全面性、解决人才短缺、增强组织协作和服务质量。国内服务尚未完全满足中级资产配置要求，缺乏高级综合服务能力。为适应市场变化，需扩展资产配置选择，提高从业人员专业水平，加强内部合作，突破传统架构限制，实现以客户为中心的服务模式。通过监测客户需求、积累经验、提高服务效率，中国财富管理业务有望有效应对挑战，实现高质量发展。

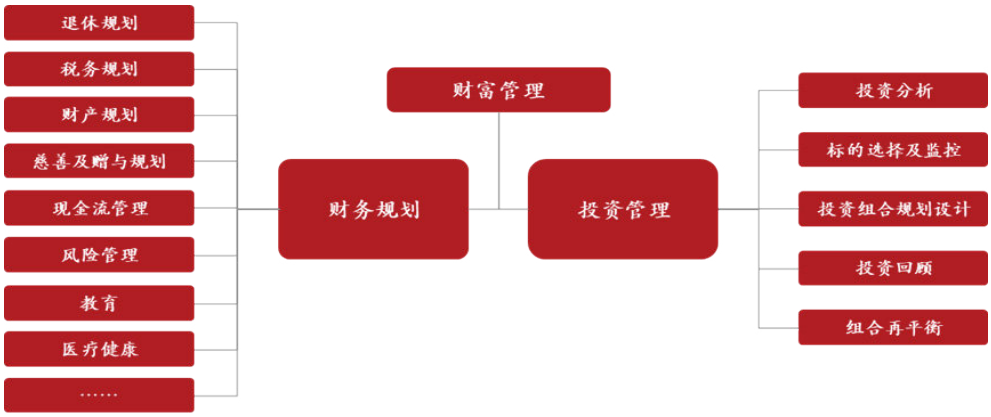


图1：财富管理服务内容总结

二、基于人工智能技术的知识中台

在数字化转型的推动下，实现以“数据驱动”为核心的业务模式，关键在于推动知识生产的自动化和连续性，即自动化知识工程的实施。这不仅响应了综合财富管理业务对客户情况的实时、持续监测需求，也是组织内部知识经验更新和演进的必要趋势。利用人工智能技术构建的知识中台，为综合财富管理业务提供了全面的技术解决方案，成为数字化转型中提升服务能力的关键基础设施。通过这一平台，组织能够更有效地利用数据，实现对客户洞察的即时性和准确性，从而推动业务的持续增长和创新。

知识中台技术架构的设计应以综合财富管理的业务需求和人工智能驱动的知识中台的技术特性为基础，如下图 2 所示。

图2：基于人工智能技术的中台架构设计

知识中台是一个集数据深化处理与应用连接于一体的平台，它位于数据资源与上层应用之间，将组织内外的数据资源转化为知识生产的原材料。平台的“精加工能力”使其产出的知识可以直接供业务端查询、调用，或作为中间产品支持高级分析和决策辅助，同时也支持智能化助手等应用，广泛应用于员工培训、市场研究、客户洞察等多个场景。知识中台的架构包括六大模块：

数据接入模块：负责整合组织内外的广泛数据资源，包括数据仓库、数据中台、第三方数据库等，通过 API 机制持续向知识生产模块输送数据。

知识定义模块：决定知识生产的表示方式、算法模型和参与生产的数据源。它还涉及知识生产的具体配置，如图知识的属性和关系设置。

模型管理模块：集成多种 AI 算法模型，支持通过图形化界面进行模型参数配置和优化。它不仅包含机器学习和深度学习模型，还能整合经济、金融领域的统计计量模型和业务规则模型。此外，该模块还包含标注数据集管理功能，便于业务人员更新和维护训练数据集，以优化或训练新模型。

知识生产模块：根据配置指令执行知识生产任务，并通过图形化界面进行生产控制和状态监测。它能够自动启动增量式知识生产，实现数据驱动的应用目标，并提供生产过程的可视化交互功能。

知识管理模块：负责对已生产的知识进行管理，包括增删改查操作、知识发布共享、知识审核质检以及用户权限管理。

知识应用模块：为上层应用提供知识输出接口，支持自然语言问答式的知识搜索和高级分析及决策辅助应用。

三、关键技术应用

在综合财富管理的应用背景下，人工智能驱动的知识中台技术要点主要集中在两个方面：①实现知识持续生产的算法模型的应用；②实现自然语言问答的知识随需而用。

3.1 实现知识持续生产的算法模型

知识中台系统在财富管理领域的应用主要涉及三种技术：图表示、向量表示和产生式规则知识生产。

图表示知识生产：利用自动化知识工程技术，结合规则、句法分析和实体识别，从各种数据中提取实体、属性和关系，辅助 NLP 模型自动化标注数据，加速知识生成。

向量表示知识生产：将知识转化为多维特征向量，便于分析和决策支持。依赖图表示学习技术，如随机游走和图神经网络，将复杂图数据转换为低维、稠密向量，便于特征表达和应用。

产生式规则知识生产：关注 "If-Then" 规则的因果关系表示，支持基于规则的推理和决策。从 Rete 算法到基于模型的规则提取，产生式规则在可解释性和稳定性上具有优势，对财富管理领域具有重要应用价值。

3.2 实现自然语言问答的知识随需而用

知识中台系统通过“基于知识图谱的问答（KBQA）”和“机器阅读理解（MRC）”技术实现自然语言问答。

KBQA 技术：利用知识图谱理解问题并检索或推理答案。方法分为基于语义解析（将问题转化为逻辑形式）和基于信息检索（通过向量相似度检索实体和信息），都需要数据标注和训练。

MRC 技术：训练模型预测文本中的答案。方法包括基于规则、基于经典机器学习（人工定义特征）和基于深度学习（自动学习特征，使用 RNN 和 CNN 等框架）。深度学习是主要交互方式，但初期结合规则和机器学习方法，尤其是人工答案标注，对解决数据集不足问题至关重要。预训练模型和元学习模型的进步为 MRC 在小样本垂直行业应用提供了新可能，支持知识中台的自然语言问答。

四、应用实践

近年来，中信建投证券积极推进财富管理和数字化的双轨转型战略。财富管理的转型为数字化提供了具体的业务方向，而数字化则赋予了财富管理转型强大的技术支持。通过构建以人工智能为核心的知识中台，中信建投证券为综合财富管理业务注入了新动力，并在知识的连续生成与更新、知识生产成果的高效检索及利用、高级分析与决策支持等关键领域取得了突出成就。

4.1 知识的连续生成与更新

中信建投证券通过建立基于人工智能的知识中台，克服了传统知识生产和管理的分散性、重复性、数据碎片化和人工依赖等局限。该中台实现了组织内数据的高效整合和自动化知识生产，显著提升了综合财富管理业务的效率和效果。

知识中台系统的数据接入模块已成功整合了多个业务部门和分公司的数据资源，将不同格式的数据转化为可用知识，支持高级分析和决策辅助应用。目前，系统能够处理 Word、Excel、PPT、PDF 等文档和主流视频格式的数据。

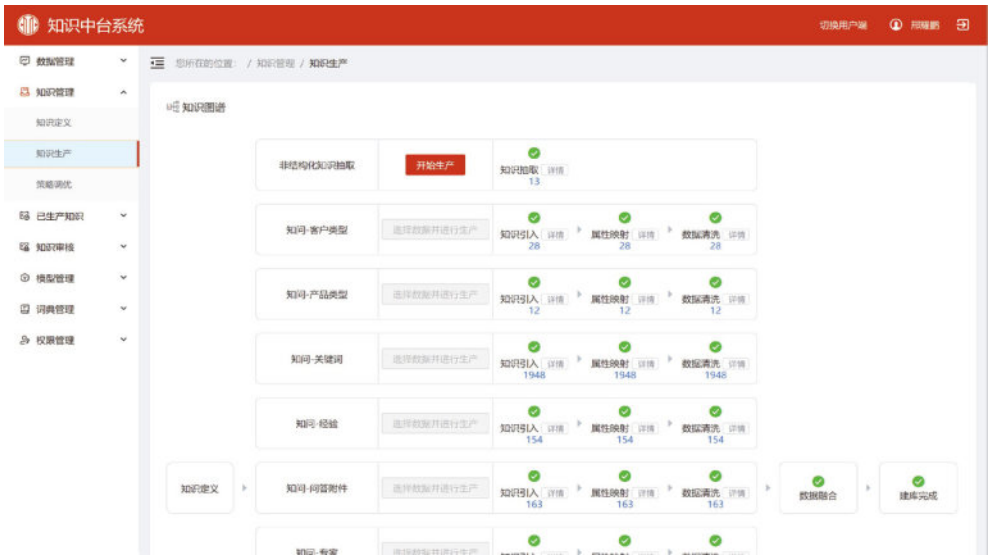


图3：知识中台知识生产模块操作界面

知识定义模块支持图表示、问答对表示和向量表示等多种知识形式，满足不同业务需求。模型管理模块集成了深度学习算法，促进了数据的自动化知识生产。业务人员可以通过图形化界面进行个性化知识定义和模型管理，灵活适应商业环境的变化。如图 3 所示，知识生产及管理模块负责监控知识生产状态，并对已生产的知识进行管理。这一模块将传统模型的“黑盒”处理过程可视化，提高了资源利用效率，加速了业务敏捷性。

此外，中信建投证券还培养了领域服务专家，建立了知识运营团队，并设立了知识合规审核，确保知识生产的质量和合规性。知识中台的稳定运行为公司提供了强大的数据处理能力，主要集中在客户通讯、投资资料、培训分享、操作规范、合规文件和市场分析报告的知识生产。知识中台的数据应用层为上层应用提供了高效检索和决策辅助工具的数据接口，激活了各业务线的多源异构数据。中台系统的持续知识生产能力为公司带来了实质性进展，特别是

在知识高效检索利用和高级分析及决策辅助工具支持方面。

4.2 知识生产成果的高效检索及利用—智能助手

知识中台支持各种智能助手的应用。智能助手系统通过结合自然语言处理、语音识别和搜索推荐技术，为员工和客户提供实时在线支持。这些系统能够实时提供话术建议、流程指导和产品知识，提高服务质量，同时进行通话质检以降低合规风险。此外，它们能够快速回应客户问题，减少人工成本并提升客户满意度。在人机协作模式中，智能助手扮演着关键角色。智能助手的效果依赖于其知识库的质量。由于不同智能助手的知识库在结构和内容上具有相似性，知识中台能够统一为这些应用提供知识生产服务，降低运营成本并支持智能助手的持续演进。

中信建投证券开发的智能助手应用专为财富管理顾问设计，可集成到资产配置平台和会话系统中，根据场景提供实时帮助和提示，如图 4 所示。利用知识中台生成的知识及 NLP 技术，智能助手能够实时分析会话内容，识别关键信息并推荐相关知识，实现知识的即时应用。知识中台还具备上传标注数据集的工具功能，使业务人员能够构建针对特定业务场景的语料库，并在中台重新训练模型以提高关键业务信息抽取的准确性。这些信息可以实时更新到客户知识图谱中，为下一轮会话提供提示，引导客户进行投资组合的调整。

此外，智能助手能够基于识别出的关键信息实体，触发并向前端推送最相关和最新的知识内容，包括专家经验、FAQ 和资料链接。这种实时推荐功能确保了财富管理顾问能够快速、专业地响应客户需求，显著提升客户服务体验和信任度。

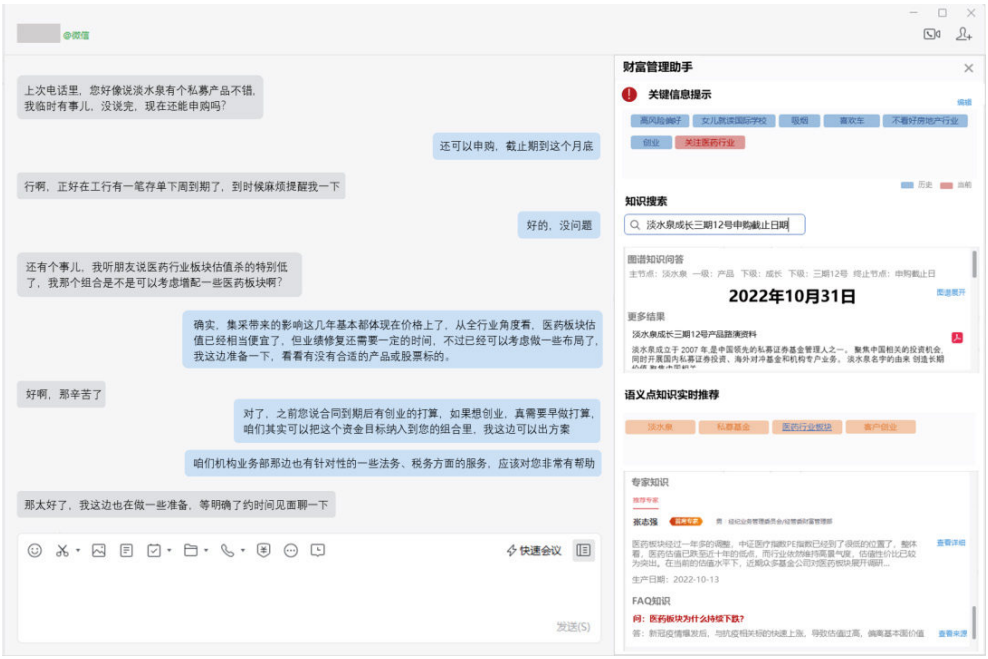


图4：财富管理助手操作界面示例

4.3 高级分析与决策工具支持

4.3.1 客户类分析及决策辅助应用

（1）全景式客户综合画像

如图 5 所示，中信建投证券通过知识中台整合了员工的专业知识，创建了面向内部员工和客户的 FAQ、经验分享和文档资料库。此外，通过数据中台，公司还汇集了来自不同业务系统的数据，经过整理和提炼，形成了包含客户资源、产品详情、市场资讯、营销资料和投资策略等在内的丰富信息资源。这些数据和信息为公司提供了全面的

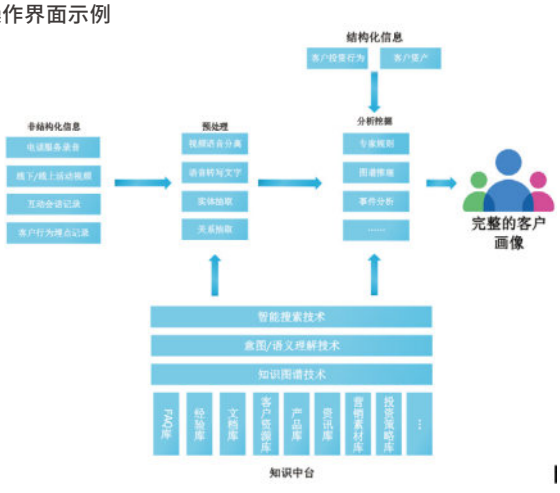


图5：知识生产的客户动态画像系统架构图

客户洞察，使得公司能够运用多种技术手段构建详尽的客户画像。

（2）潜在客户挖掘

知识中台利用多渠道客户数据创建的客户图谱为潜在客户的识别提供了丰富的数据基础。通过分析已成功转化的客户数据，建立标注的客户转化图数据集，并运用无监督学习如聚类和标签传播算法，或有监督学习如分类算法，精准高效地发掘与现有综合财富管理客户具有相似特征的潜在客群。

（3）客户流失风险评估

综合财富管理中，客户流失风险的评估同样依赖于全面的客户画像。借鉴潜在客户识别的方法，通过分析已流失客户的资料建立客户流失图谱数据集，并应用图表示学习的分类算法，无论是无监督还是有监督，以实现对客户流失风险的精确且高效的评估。

4.3.2 业务类分析及决策辅助应用

（1）事件影响分析系统

如图 6，中信建投证券运用知识中台处理多源数据，通过 NLP 技术构建产业链知识图谱和上市公司图向量，分析突发事件与公司的文本相关性。结合舆情数据和股价走势，运用统计方法计算事件的数学相关度，并将两种相关度综合评估，形成事件影响分析。系统提供交互式可视化操作，能够定量分析事件对市场、行业及个股的影响。财富管理顾问可利用此系统快速分析历史事件影响，针对不同客户资产状况进行精准沟通，指导客户进行风险管理和投资组合优化。

（2）预测因子挖掘系统

中信建投证券利用知识中台构建的产业链知识图谱，并结合统计计量模型，开发了一套图形化交互工具平台，旨在为财富管理顾问提供决策支持。

平台的因子挖掘模块通过统计和文本相关性指标评估宏观、行业和市场数据对上市公司业绩的预测潜力。统计相关性指标利用格兰杰因果检验等方法，分析财务数据与其他数据的时间序列关系；文本相关性指标则通过比较产业链图谱向量与关键词向量的相关度来构建。这两个指标帮助识别出具有预测价值的因子。时间序列预测模块将这些上述因子导入并运用多种集成的预测算法进行模型训练和实验，所有操作如特征选择、模型构建和参数调整都可通过可视化工具完成，无需编程技能，如图 7、8。

(a)　　(b)

图6：知识中台产业链关系知识生产结果查询界面

图7：因子挖掘操作界面

图8：时序预测操作界面

这些高级分析和决策辅助应用增强了财富管理顾问方案的差异化竞争力和说服力，同时帮助顾问利用分析结论与客户进行深入沟通，增加互动频次，从而提升客户忠诚度。

4.3.3 协同类分析及决策辅助应用

中信建投证券基于人工智能技术的知识中台持续不断将涉及客户、专家（员工）、标的及专业服务的多源异构数据以图表示方式进行加工生产，基于已生产的图数据，实现基于客户特征的团队协同组队推荐（见图 9）。知识生产及推荐算法不仅涵盖了客户、员工、产品、服务的知识管理，还实现了知识中台与企业微信的无缝连接。当财富顾问需要专业支持时，系统能够智能地识别并推荐符合条件的组织内专家，并通过企业微信实现即时沟通，显著提高了团队沟通的效率。此外，系统还支持跨区域的资源方与服务专家组队，共同为特定客户提供服务，并根据预先设定的比例分配收益。这不仅促进了专家个体经验的跨区域共享和价值创造，也实现了一个集知识中台、协同组队以及企业微信通讯于一体的综合员工赋能平台和机制。

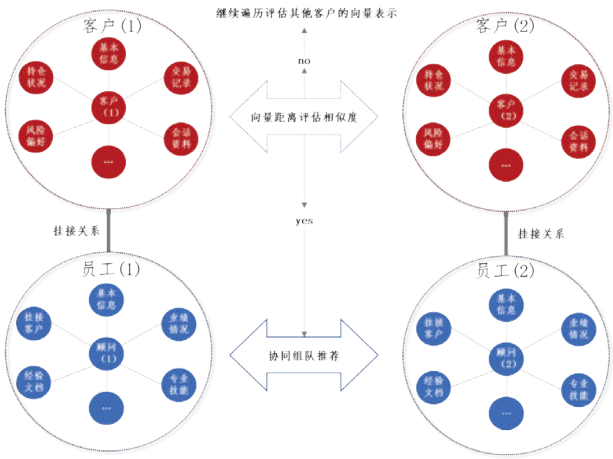


图9：协同组队推荐原理示意图

五、总结

中国证券业协会在 2021 年 3 月提出的《证券行业文化建设十要素》中的第九条强调了“崇尚专业精神”的重要性。该原则鼓励证券行业高度重视并不断提升自身的专业化水平，培育专业精神和专业主义，通过增强专业附加值来营造一个尊重专业、学习专业、并按照专业精神行事的文化环境。

知识中台的建设和基于其上的应用实践，正是对这一原则的具体落实。通过构建和利用知识中台，证券公司能够促进员工之间的知识共享和学习交流，从而持续提升整个组织的专业化水平。知识中台作为一个技术和数据驱动的平台，不仅支持员工获取和更新专业知识，还鼓励他们在实际工作中运用这些知识以实现卓越的业务成果。此外，知识中台作为一个平台，让员工在专业精神的指导下，进行更高效的沟通协作，共同解决复杂的业务问题。这种以专业精神为核心的职业氛围，有助于提升员工的工作满意度和忠诚度，同时也能够提高客户服务质量，增强公司的市场竞争力。因此，知识中台的建设和应用实践不仅是技术进步的体现，更是证券行业文化建设的重要支撑。

参考文献：

[1]Chhabra A B. Beyond Markowitz: a comprehensive wealth allocation framework for individual investors [J]. The Journal of Wealth Management, 2005, 7(4): 8-34.

[2]Jahnke W. The comprehensive wealth management doctrine[J]. Journal of Financial Planning, 2000, 13(9): 50.

[3]Zhang C, Lai Y, Feng Y, et al. A review of deep learning in question answering over knowledge bases [J]. AI Open, 2021.

[4]Lan Y, He G, Jiang J, et al. A survey on complex knowledge base question answering: Methods, challenges and solutions[J]. arXiv preprint arXiv:2105.11644, 2021.

[5]Baradaran R, Ghiasi R, Amirkhani H. A survey on machine reading comprehension systems[J]. Natural Language Engineering, 2020: 1-50.

[6]Zhang N, Hui L I, Tang J. An Approach to Answer Extraction in Question Answering Based on Semantic Concept[J]. journal of xihua university(natural science edition), 2009.

数据驱动式投资者智慧服务链建设

徐鑫鑫，陈心亮，张津铨

｜中国证券登记结算有限责任公司上海分公司｜E-mail：xinxinxu@chinaclear.com.cn

摘　要： 始终把用户需求放在首位是建设高质量业务系统的根本前提。中国证券登记结算有限责任公司坚持理论创新、实践创新，坚持把金融服务实体经济作为行动指南，充分发挥业务技术合力，从工作实践中抽取业务核心链条，通过建应用、搭系统、筑平台等方式，借助金融科技创新驱动力，打造数据驱动式投资者智慧全景式服务链条，为登记结算业务投资者提供场景化、专业化、智能化一站式服务体系。

关键词： 数据驱动；微服务；服务链；智慧赋能

一、服务链建设背景

2023 中央金融工作会议强调，金融是国民经济的血脉，要加快建设金融强国，全面加强金融监管，完善金融体制，优化金融服务，防范化解风险，坚定不移走中国特色金融发展之路。中国证监会党委传达学习贯彻中央金融工作会议精神中指出，要紧紧围绕加快建设金融强国的目标，增强金融报国情怀和政治担当，扎扎实实办好资本市场的事情。作为资本市场重要的基础设施之一，中国证券登记结算有限责任公司（以下简称我司）始终行金融为民理念，从业务发展和技术进步两手抓，紧随技术发展趋势，持续推进投资者条线各类型业务办理流程和操作的优化工作，通过建设数字化系统、优化券商代理渠道、开展一柜通办等多项措施，持续为机构投资者和个人投资者提供高质量服务。

登记结算投资者业务办理过程是将各类业务材料、要素的信息汇总处理后获得投资者期望的业务效果，本质上是业务数据的变更。我司在多年投资者服务过程中，围绕业务数据生命周期提炼出“申请 - 办理 - 日终 - 反馈”服务主链条，基于主链条不断推动投资者服务拓展和升级。从最初的纸质材料提交 - 人工审核 - 纸质凭证反馈等业务驱动式办理模式，逐步过渡到由数据在相关系统之间的流转数据驱动模式，是中国结算不断提升服务质量，在实践中总结改进而形成的服务新形态。而随着技术尤其是金融科技的不断发展进步，给投资者服务由“办成”向“办好”提供了源源不断的助力，能够更好的满足不同投资者多元化需求。

二、业务发展及技术趋势

中国证券登记结算有限责任公司是中国证券交易所市场的集中统一登记结算机构。投资者业务是公司登记存管核心业务之一，其用户类型可分为个人投资者和机构投资者，所涉及业务类型包括账户类业务、存管类业务、名册类业务等，不同于银行业、证券业等通用性投资者业务场景，我司的投资者业务具有资本市场特殊性质，相关业务只有我司具有办理权限，因而对我司技术系统提出了更高的要求。

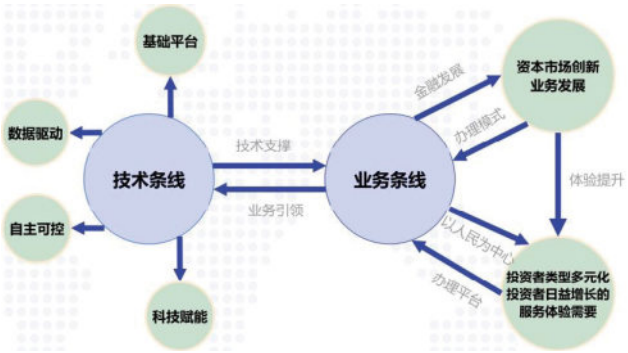


图1：投资者业务及技术发展趋势背景

我司投资者服务链建设，是在实践中不断发展前进的，从业务和技术两个条线不断吸取市场意见，借力技术发展，多措并举完成建设画卷。

2.1 业务条线

中国结算投资者服务建设过程中，始终深刻把握金融工作的政治性、人民性，为经济社会发展提供高质量的金融服务。

资本市场创新带来的业务类型及业务规则变化，推动基础设施服务体系持续升级改造。通过全面梳理投资者业务子类别，修订京沪深三地投资者业务指南及有关业务操作规定，在确保安全合规的前提下，提高业务通柜办理效率，减少业务办理中非紧要环节，不断适配资本市场业务创新；在三年疫情防控期间，我司允许券商代理投资者业务的全部子类别，市场各方均更加适应通过券商代理渠道办理投资者业务，投资者可就近券商网点办

理业务，对柜台业务分流效果较为明显，投资者免除奔波之苦，满意度得到大幅提升。

投资者类型多元化以及投资者日益增长的服务体验需要，对业务办理模式和办理平台持续不断提出考验。中国结算上海分公司积极响应市场投资者需求，解构业务逻辑，提供定制化业务应用，改善用户体验，提升系统服务精准度。借助好差评、问卷调研等多种模式认真听取投资者改进意见，引导投资者向采用预填单、受理系统等互联网渠道进行业务提交和办理，“让客户少跑腿，让数据多跑路”，全方位立体化满足客户多维需求，让更多的客户感受到我司的专业和诚意。

2.2 技术条线

中国结算在建设投资者服务相关的技术系统时，坚持立足当前、着眼长远发展，推动整体服务链式发展，同时牢牢守住不发生系统性风险的底线，强化创新渠道，建立高水平自强自立投资者综合服务应用及平台。

一是基础平台建设。从最初的手工记账结算，到核心系统建设，再到业务办理平台 BPM 系统建设，技术创新有效支持了业务创新，对业务发展产生了长远的推动力。当前我司面对数字化转型时代趋势，采用微服务、多应用组件式开发模式开展系统建设，沉淀通用业务能力，实现集约建设。

二是数据驱动探索。业务数据是最核心要素和资产。通过充分发掘投资者业务办理中数据的关键作用，自动进行数据联通、采集和整理，让数据成为串联服务链各环节的丝线。

三是技术自主可控。针对面临的国际形式，我司积极响应国家号召，推动技术自主可控，打造基于国产化技术的业务办理平台，实现从底层芯片、硬件设备到操作系统、业务应用的全方位自主掌控，以加强服务稳定性，降低供应链风险和技术风险。

四是金融科技赋能。金融科技作为技术驱动的金融创新，是系统不断发展向前重要引擎。通过加强对人工智能、流程挖掘、大模型等新兴金融科技的的研究，不断引入新技术、新路径，借助外脑推动中国结算投资者服务体系高质量发展。

通过业务条线和技术条线协同发力，建立了投资者服务链雏形并不断发展完善，打好业务整合和技术赋能组合拳，从实践中来，到实践中去，以投资者实际需求为出发点不断丰富和扩充服务链内容。

三、智慧服务链建设实践

3.1 建设目标

中国结算投资者智慧服务链建设目标，是抓住业务数据流转核心主线，以一条主链，多重支链，双向服务，应用集成为内容，以金融科技为抓手，配套数字化系统建设及管理，充分实现业技融合，建立投资者服务的智慧全景式服务链条，为投资者提供场景化、专业化、智能化一站式服务体系。

3.2 服务链总体架构

在服务建设层面，采用内外两条线支撑业务创新需求，着力于外部一站式服务和内部智能化建设，打造多渠道融合智慧营业厅，借助金融科技构建微服务平台，丰富服务供给，确保技术路径与业务需求高度匹配。

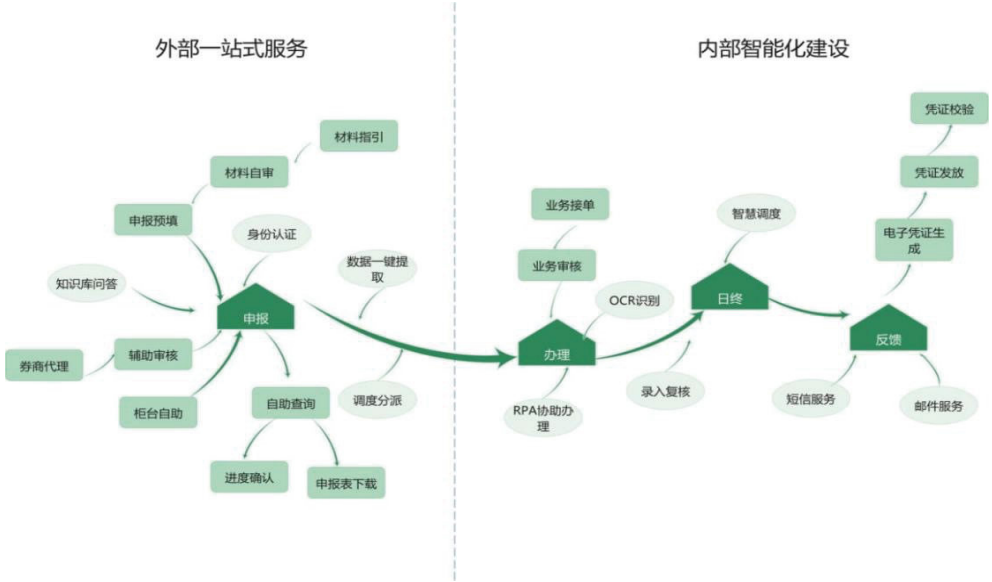


图2：投资者服务全景图

在投资者智慧服务链体系中，申报环节是投资者服务第一站，我司利用互联网和移动端，通过门户网站、移动端应用、微信公众号、PROP 客户端、智能柜台机等多种方式，建立多渠道融合服务模式，支持投资者在不同的业务场景下灵活组合使用。面对多元化投资者客群，允许投资者在线自助申报、券商代理申报和柜台人工申报，尽最大能力满足不同客户的需求。同时，在主链基础上进行支链建设，如在线预填支链，通过智能材料指引和清单生成，提供客户材料自审功能，减少材料错误缺失等可能性，避免业务反复影响流转效率。在链条建设基础上，加大数据协同共享力度，提高信息流转效率，通过进度展示功能实时反馈办理进度和办理状态，让客户看得见，放下心。

在内部智能化工作中加大内功建设，推动中国结算“一网通办”和“一柜通办”深入开展，将数字化理念融入日常业务，采用微服务架构体系，全力构建新型在线服务平台，促进线上线下一体化建设，加强系统自主可能建设，以开放的姿态积极响应绿色金融号召，建立分公司电子凭证系统，实现凭证在线生成和数字化存储，有效减少纸张使用量，助力绿色转型。

投资者智慧服务链整体应用架构中，以标准化接口提供可扩展技术服务，支持应用 - 系统 - 环节 - 服务链体系建设，采用微服务多应用建设模式，将系统拆分为多项子应用，以应用组件化进行系统组装扩展和升级，打造低耦合、高内聚技术底座，通过数据流动串联预填单系统、在线业务受理系统、BPM 系统、日终核心系统等，实现技术赋能业务。

3.3 主要建设内容

中国结算核心系统目前正在进行新一代建设，以技术架构转型和技术自主可控为基本任务，有力支持公司业务整合和提升市场服务质量。围绕核心系统升级，响应市场需求，除核心系统外，服务链各环节所涉及应用和系统建设等任务，根据作用效果及服务对象分为三类，一是面向投资者一线的相关应用体系组成了智慧营业厅，二是内部业务解耦型基础平台，三是金融科技赋能提升的相关应用系统。

3.3.1 投资者一站式智慧营业厅

服务渠道多媒体化、轻量化和交互化，推动金融服务向线上办、掌上办转型，智慧营业厅直面投资者一线，是投资者业务办理门户和窗口，为投资者提供面对面服务。系统作为对接投资者的桥头堡，把主要工作在线前置，提高用户体验的同时能够减轻办事人员负担，提升业务流转效率。智慧营业厅融合网站、APP、微厅、专用客户端等渠道满足不同类型投资者多元化需求，打通了数据链，通过数据流转消除渠道壁垒，能够方便获得所需数据，办理流程一目了然，在尊重用户意愿的前提下引导用户进行智能化、自助化办理，打造多维无边界的业务服务能力。

中国结算“业务预填单”以微厅和网厅为入口，为临柜办理业务的投资者提供集材料指引、业务预填、表单生成、进度反馈为一体的线上线下一体化服务。系统将用户常用业务竖井化分类，覆盖证券质押、非交易过户、查询业务、其他业务全类型投资者柜台业务，投资者填

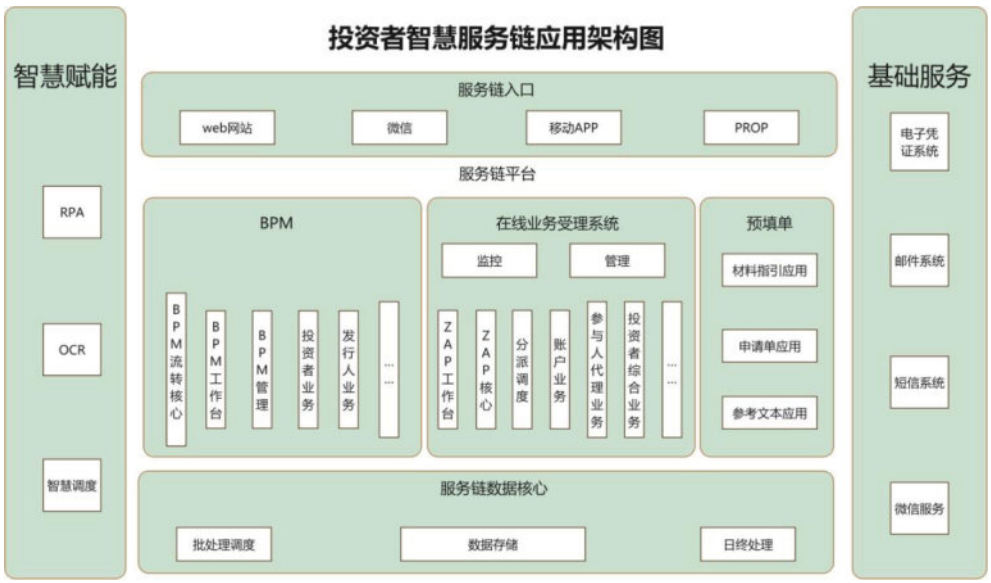


图3：投资者智慧服务链应用架构图

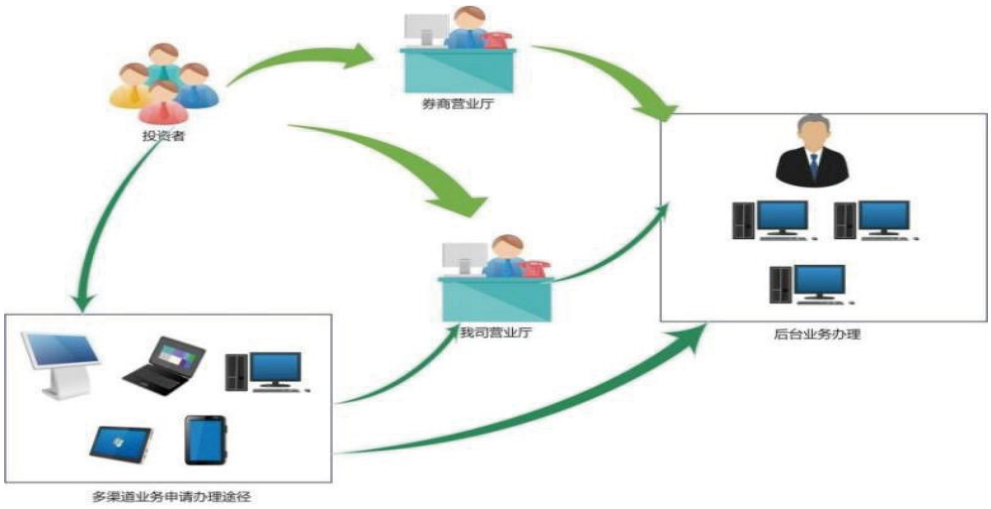


图4：智慧营业厅访问办理过程

图5：预填单网厅业务展示图

单后，数据可一键导入办理系统，系统具备参考文本填单功能，方便生成制式表单，如授权委托书、合同文本等，为投资者提供进一步便利服务。

考虑移动渠道的普及性，“业务预填单”应用在移动端通过微信和专用 APP 对外提供服务。探索“一网通办”，推进业务流程整合，构建掌上“一站式办理”的服务模式，形成全链条打通的业务开展形式，提升市场用户体验。

“在线业务受理系统”是券商代理渠道重要途径，提供网站和 PORP 专用客户端两种方式，采用分布式架构设计，包含 ZAP 工作台、ZAP 核心、ZAP 调度等多个平台子应用，以及投资者业务类子应用，通过券商代理服务投资者，券商提交后系统通过调度自动对接内部办理平台，我司业务人员处理后直接反馈结果，全程在线，无需投资者临柜处理。

3.3.2 通用业务办理平台

在建设通用业务办理平台过程中，坚持平台共用、业务解耦原则，扩大系统和应用的使用范围。

BPM 系统是流程建设和运营的支撑平台。中国结算上海分公司新一代 BPM 系统建设项目搭建了适配国产化软硬件环境、自主可控新型 BPM 平台，通过微服务化实现核心引擎应用与业务分离。BPM 系统在设计时，不仅考虑投资者业务，也希望实现其他业务类型的办理。通过微服务应用拆分，将基础功能和业务功能解耦，不同业务类型封装成业务应用，实现能够快速推广部署于相关业务领域；业务流程运行于一套平台之上，支持业务流程竖井化开发和部署，形成结构化、一体化、系统化流程管理和运营工具，实现工具和方法论的有机统一。

电子凭证系统是支持绿色金融、推动可持续发展的有效手段。通过构建统一的、通用的、标准化的对内及对外的电子业务凭证生成、查询等服务机制满足对内对外需求，提高我司业务办理效率，优化客户业务办理体验，提高市场的满意度。电子业务凭证服务系统实现了电子业务凭证生成的入口唯一、管理统一，具备凭证生成、加盖电子印章、凭证查询、发送电子邮件、凭证失效 / 无效、手工上传、指令预约、批量处理等多项功能。

3.3.3 金融科技赋能

通过引入机器人流程自动化（RPA）和字符识别技术（OCR）进行投资者业务的办理，对办理效率带来立竿见

图6：预填单微厅业务展示图

图7：在线业务受理系统投资者证券账户业务模块

影的效果。中国结算上海分公司建立了企业级 RPA 服务平台，形成具有公司特色的高频流程自动化解决方案，目前已在上海分公司多个部门部署了 10 余个数字机器人，其中 2023 年度处理业务超过 1 万笔，综合办理效率提升 200% 以上。在开户业务中，通过 RPA 与 OCR 结合，自动进行开户业务的接单、审核、流程处理操作，业务处理更加及时，有效提升客户满意度。同时，我司也在加大对金融科技的研究力度如 AIGC、金融大模型等，持续探索金融科技对服务链的赋能和提升作用。

3.3.4 服务链建设成效

我司投资者智慧服务链体系运转高效，服务于京沪深三地登记结算业务投资者，为投资者提供多元化办理渠道和服务体验。目前上海分公司已实现所有投资者日常业务在线平台办理，每年可办结约 7 万笔投资者业务，其中 99% 的账户新开类业务已全面通过在线平台发起办理，97.6% 修改类账户业务通过在线平台办理，每年可生成近 7 万份电子凭证，节省纸张近 10 万张，有力贯彻数字金融和绿色金融理念。

四、总结与展望

大厦之成，非一木一材也；大海之阔，非一流之归也。中国结算投资者智慧服务链建设，是一项长久实施的大工程，与资本市场业务发展以及金融科技发展均紧密相关。我司始终站稳人民立场，倾听人民的意见，持续跟踪登记结算投资者业务活动中的新需求新问题，及时跟进行业内应用新型技术手段开展情况，适时制定服务链更新升级计划，坚持核心自主、安全可控导向，从业务技术两条线推进服务链优化升级，确保时时有进展，按期见成效，让投资者始终能够享受到高质量登记结算服务。下一步，中国结算将持续探索知识图谱、大数据、大模型等一系列新技术在投资者服务中的应用，不断扩充服务链的服务范围和服务广度。新故相推，日生不滞，让服务链更广泛、更深入、更有效的服务广大登记结算投资者，支持资本市场高质量发展。

参考文献：

[1] 杨舒，苏放．基于微服务的分布式数据安全整合应用系统 [J]. 计算机工程与应用 ,2021,57(8):238-247.
[2] 徐鑫鑫 , 陈心亮 , 李军林 . 业务流程治理体系探索与实践 [J]. 交易技术前沿 ,2024,56(2):42-45.
[3] 陈宛 ．数字要素时代证券公司的金融科技转型探索与实践 [J]. 人工智能 2023,(02):62-70.

国泰君安灵犀一语达
—— 跨平台语音文字全能助手

周素珍，于三川，王睿楠，张孟 | 国泰君安证券股份有限公司 | Email：zhousuzhen@gtjas.com

摘　要： 在数字化时代发展下，如何帮助用户在复杂的系统中迅速定位所需功能成为亟待解决的问题。“灵犀一语达”通过统一管理和训练软件操作数据，采用语音识别和模型推理技术，实现用户指令（如银证转账、会议总结）的即时响应和无感交互，消除功能认知盲点和操作不便。方案支持跨平台应用，以国泰君安托管服务PC平台和员工全连接APP为例，有效提升员工效率和用户体验，助力企业数字化转型。

关键词： 混合模型架构；语音识别；无感交互；意图识别；菜单训练与管理平台

一、概述

国泰君安作为一家大型金融服务机构，拥有一系列自主研发的软件应用，包括服务于零售客户的国泰君安君弘、富易和锐智，以及面向机构客户的道合和君极等。此外，还为内部员工提供全连接等应用支持。随着业务的发展和平台的不断完善，这些应用程序的功能日益丰富，例如托管管理人服务平台的功能菜单已超过 200 个，全连接 PC 端的功能更是达到 760 多个以上。然而，功能的增多也带来了一些挑战。首先，用户面对如此繁多的功能菜单，尤其是新用户，查找所需功能的入口变得困难。以托管管理人服务平台为例，用户可能需要经过三级菜单的切换才能找到管理费率调整的选项。其次，当用户的需求涉及多个系统时，业务流程可能会变得繁琐且耗时。例如，一名内部员工需要预约腾讯会议并进行会议纪要总结，他需要在全连接平台上找到腾讯会议按钮，然后进行预约、录制、下载转写记录等一系列操作。如果他还需要发送邮件，则需要再找到邮箱入口并进行发送。整个过程需要至少 6 次操作，且需要在腾讯会议、邮箱和全连接三个系统之间进行切换。

因此，当前面临的一个关键业务痛点是如何帮助用户在功能繁多的 APP 和 PC 中快速找到所需功能，以及如何将复杂的操作指令自动化执行，以提高用户的操作效率和体验。

2024 年 9 月 5 日至 7 日，2024 外滩大会在上海举行。大会期间，支付宝发布 AI 生活管家 App“支小宝”1，提出：“只要一句话，生活就 AI”与我们的想法不谋而合，“灵犀一语达”旨在探讨一个高效、安全且用户友好的跨平台应用系统的设计和实现。其具体目标是统一管理不同应用的菜单功能，根据不同场景利用用户真实的操作数据进行训练，并借助语音识别、模型推理、向量库、实体识别等技术，实现用户指令（如银证转账、总结会议等）

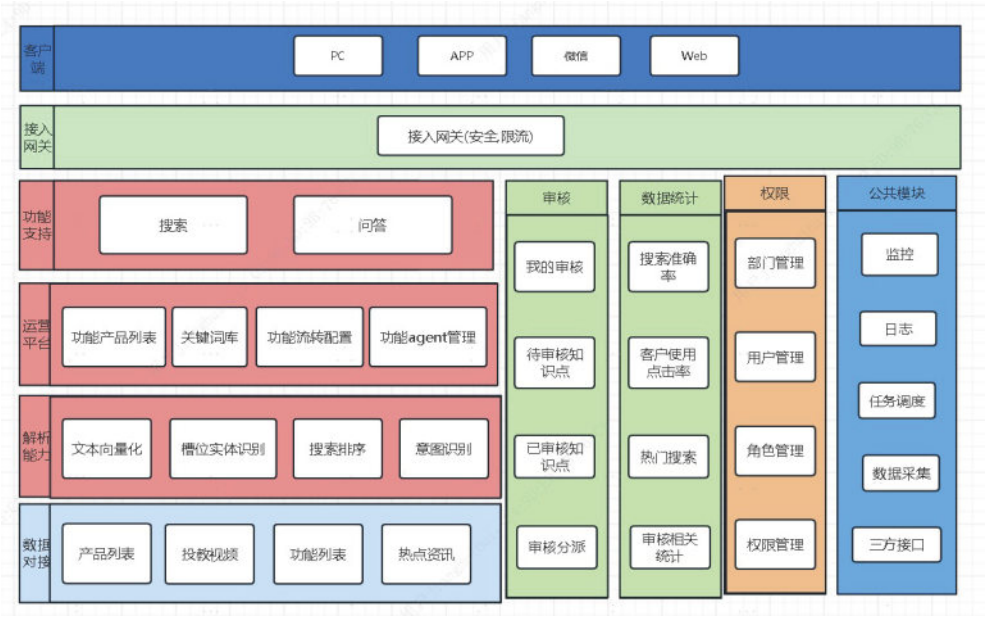


图1：整体功能框架图

的一语即达、无感交互，解决用户对功能认知的盲点和操作的不便。该方案支持跨平台落地应用，并将以托管服务平台和全连接为应用示范进行推广使用。

二、建设方案

2.1 整体功能框架

整体功能框架如图 1 所示。该系统可支持跨平台客户端的接入，包括个人电脑（PC）、移动应用程序（APP）、微信和 Web 端，为用户提供无缝的接入体验。网关的设计不仅确保了用户访问的安全，还通过限流机制保证了系统的稳定性。

一旦用户成功接入系统，他们将享受到一个由多个功能模块组成的综合服务平台。这些功能模块包括但不限于：

1）审核模块：确保用户提交的内容符合规定标准，维护平台内容的质量和合规性。

2）数据统计：通过收集用户行为数据，为系统优化和决策提供支持。

3）权限管理：精细化控制不同用户或角色的访问权限，保障信息安全。

4）核心模块 提供搜索、问答支持等核心技术服务，增强用户体验。

5）监控系统：实时监控系统性能和状态，及时发现并解决问题。

6）运营模块：包括关键词库管理和功能流转配置，提高服务效率和质量。

7）日志记录：详细记录用户行为和系统事件，便于问题追踪和数据分析。

此外，系统还具备与第三方接口对接的能力，能够与外部服务和数据库进行交互，实现数据共享和功能扩展。统计模块对系统使用情况进行综合分析，为持续改进提供数据支持。

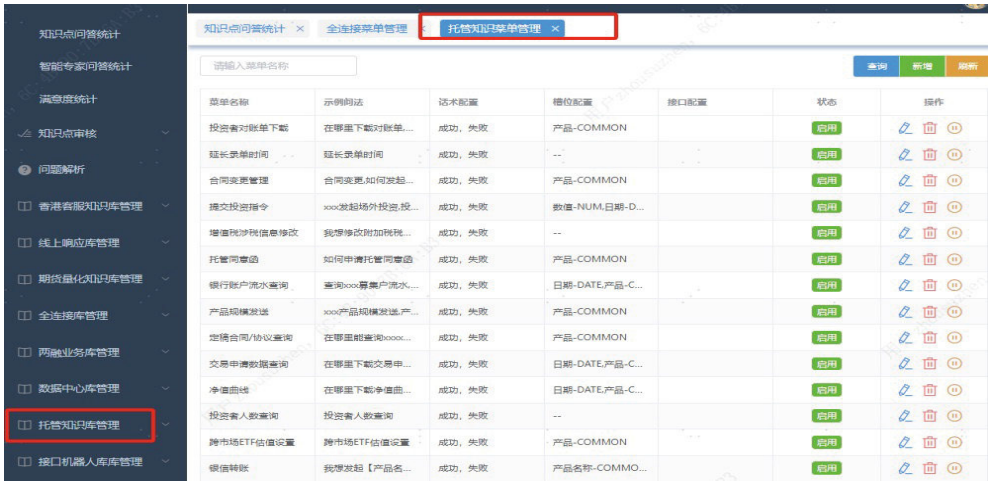


图 2：菜单列表

在本文中，我们将深入探讨上述功能模块的设计原理、实现方法以及它们如何协同工作，以构建一个高效、安全且用户友好的跨平台服务应用系统。我们将详细分析每个模块的角色和功能，以及它们对系统整体架构的影响，从而提供一个全面的系统理解。

2.2 重点模块建设

2.2.1 菜单列表管理

在构建本系统时，我们特别关注了用户界面的直观性和操作的便捷性。系统菜单列表的设计旨在提供一个清晰的导航结构，使用户能够快速找到并访问所需的服务和功能。不同的应用管理各自不同的菜单，以下是托管库中管理的菜单列表截图：

2.2.2 插件接口管理

实现与外部系统和服务的高效集成是提升用户体验和系统功能的关键。为此，我们的系统采用了插件接口管理机制，以实现灵活、安全的第三方服务接入。插件接口管理允许系统开发者和管理员通过标准化的接口与外部应用进行交互，这些接口遵循预定义的协议和数据格式。通过这种方式，系统能够扩展其功能，利用外部资源，同时保持核心服务的稳定性和安全性。

在我们的系统中，插件接口管理包括以下几个关键方面：

1）接口代码与名称：为每个接口提供一个独特的标识符和易于理解的名称，以便于管理和识别。

2）描述：提供接口功能的详细描述，包括其用途、输入参数和预期输出。

3）状态：标识接口当前是否启用，以及其在系统中的可用性。

4）操作：允许管理员对接口进行启用、禁用或更新等操作。

5）输入参数：标准接口的输入参数列表，便于其他地方配置使用。

2.2.3 槽位实体配置

本系统的核心技术是理解和准确解析用户的查询。为此，我们采用了先进的自然语言处理技术，通过槽位实体配置来识别和提取用户查询中的关键信息。槽位代表用户查询中需要提取的信息类型，而实体则是具体的数据值。

如图 4，在预订线上会议的场景中，系统需要识别以下槽位及其对应的实体：

1）员工：需要参加会议的人员名单。

2）会议主题：用户希望预订的会议的主题。

3）数值：会议预定的持续时长。

4）日期：会议预定的开始时间。

系统还提供了灵活的配置选项，如是否支持反问以适应不同的用户查询和业务需求。而槽位实体配置允许系统通过预定义的规则和模式来识别和提取这些关键信息。在我们的系统中，每个槽位都与特定的操作相关联，

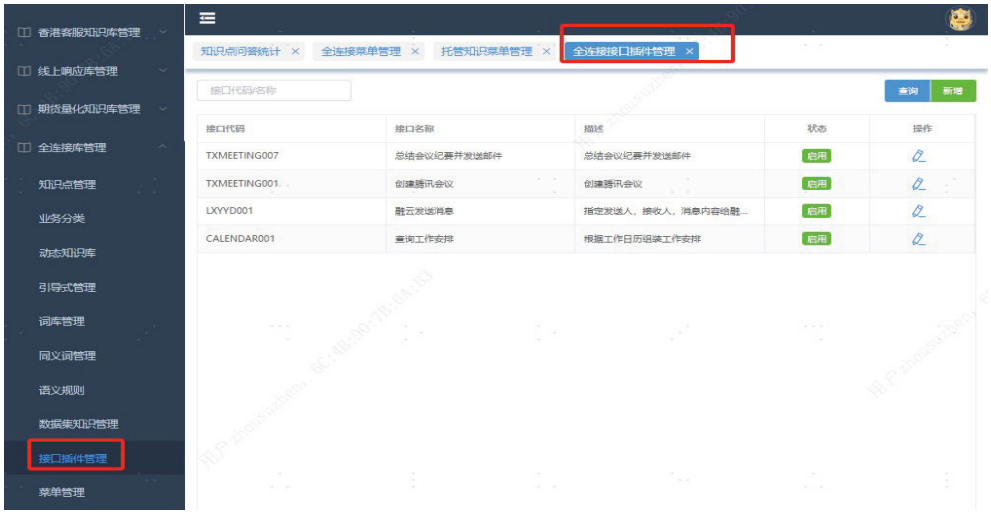


图 3：接口插件列表

图 4：槽位实体配置

一共支持以下三种配置：

1）调用接口：系统通过调用相应的接口来查询该槽位的待选项，如根据姓名关键词查询候选项。

2）是否实体识别：系统需要在实际检索流程中判断是否需要对该槽位进行实体识别，是否需要进行槽位到实体的转换。

3）是否必须：标识该槽位的实体是否为必须提供的信息。

最终，通过第三方接口对接，系统能够与外部服务和数据库进行交互，实现数据共享和功能扩展，比如在槽位提取和实体识别之后自动化地处理腾讯会议预订。

2.2.4 核心功能检索

本系统的核心是实现快速、准确地响应用户需求。通过一系列高级算法和数据处理技术，确保用户能够迅速找到所需的信息或服务。

涉及到的关键技术有：

1）关键词匹配与意图识别：系统首先通过关键词匹配和意图识别技术，快速理解用户的查询意图。这涉及到使用向量检索技术 + 大模型推理技术来分析用户的输入，并将其与预定义的意图进行匹配。

2）槽位提取与实体识别：一旦识别了用户的意图，系统将进入槽位提取阶段，其中系统将识别和提取用户查询中的关键信息（槽位），如日期、时间、产品名称等。实体识别算法用于将槽位提取到的信息与数据库中的准确数据进行模糊识别、精准转换。

3）会话缓存技术：会话缓存技术用于记录当前会话，支持系统处理槽位混取情况，并在必要时向用户提出反问，以获取缺失的信息。

4）大模型离线微调：使用大型预训练模型进行离线微调，以适应特定的业务需求和语言模式。本次创新

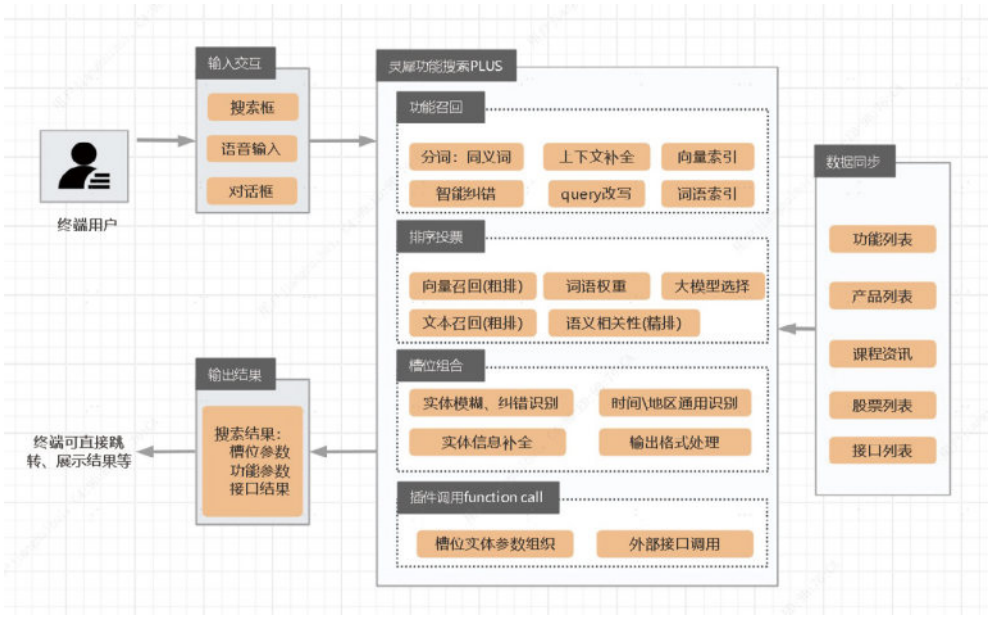


图 5：用户交互设计图

应用中托管服务平台是收集了用户操作的真实数据进行模型微调，以实现私有化部署模型应用的要求。

三、应用案例

3.1 国君托管智能助手：国君托管服务的智慧大脑

托管外包业务涉及私募运营过程中的方方面面，运营过程复杂且分散；私募运营人员查找菜单、办理业务有一定专业性要求。对于某项业务可以直接通过平台办理还是线下办理存在疑问需要咨询。作为业务办理平台，对于高频业务场景存在重复办理诉求，客户希望有极简的操作体验。

"灵犀一语达"创新地将大模型解决方案融入托管服务管理平台，为用户提供智能化的问答服务。用户只需通过提问的方式，系统即可根据问题匹配相应结果，智能地提供解答或推荐相关菜单，并引导用户完成业务办理。此外，用户还可以直接输入指令，系统会将其转化为平台操作，经用户确认后即可更便捷地执行。这种智能交互方式不仅有效减轻了客服和运维人员的工作负担，更提升了客户的平台使用体验和工作效率。针对核心功能 1——菜单识别本文准备了 586 条测试语料，核心功能 2——槽位提取准备了 189 条测试语料，分别进行了以下实验：

通过实验对比，本文设计的多级接力融合算法，巧妙地结合了关键字提取、向量搜索重排以及大模型提示

表1：意图分类实验数据表

	纯关键词匹配法	纯向量搜索算法	文心一言	微调大模型	本方案：多级接力融合算法
通过率（单位：%）	48.1	83.4	62.5	96.6	98.6

表2：槽位提取实验数据表

	开源大模型	文心一言	本方案：微调大模型
准确率（单位：%）	48.1	62.5	96.6

图6：自动化会议总结效果图

工程等技术，并且在本地微调训练了一个 ChatGLM6b-2 的大模型，显著提高了复杂逻辑推理任务的意图识别准确率，使准确度高达 97.6%。基于数据敏感性要求，托管服务平台采用微调后的 ChatGLM2-6B 模型作为小模型进行槽位提取。准确率也达到了９６.６%。

3.2 全连接一语即达：多模态交互无感体验

国泰君安全连接平台经过多年建设与公司内的大范围推广，已经成为员工办公的核心工具，其中全连接 APP 端在 23 年员工共计启动超 1700 万次，承载着 30 个核心办公功能和各个业务系统、IT 系统和行政管理系统的 200 多个入口菜单。与此同时，随着技术的发展和工作模式的转变，企业和个人都在寻求能够提升工作效率、减少手动输入和提高办公效率的工具。

" 灵犀一语达 " 的全连接功能实现了类似于车载导航的智能化体验，为公司员工提供了四个高频且无感交互的应用场景。无论是上下班打卡、发送消息、播报工作日程还是拨打员工电话，都能通过语音交互轻松实现，显著提高了员工的工作效率和便利性。此外，全面支持从会议预订到自动生成并发送会议纪要邮件的全流程，为企业提供了高效的会议管理解决方案。这一功能大大减少了人工整理会议纪要所需的时间和精力，让会议管理更加智能和便捷，助力企业提升协作效率。

四、创新性与应用性

4.1 技术创新

4.1.1 语音识别与 NLP 技术的集成实现多模态交互体验

“灵犀一语达”集成了先进的语音识别技术和自然语言处理（NLP）技术，能够精准地捕捉和理解用户的语音指令。将语音信号转换成文本形式，并进一步解析用户的意图。通过针对应用场景中常用的用户操作指令进行语音模型训练，有效提高了语音识别的准确率和抗噪性能，为用户提供更佳的交互体验。

4.1.2 多级接力融合算法提高意图识别准确性

多级接力融合算法通过结合关键词提取、向量搜索重排和大模型 prompt 意图识别等多种技术，构建了一种多级融合架构，有效地提升了复杂逻辑推理任务的意图识别准确率。这种算法能够更准确地理解用户的意图，从而为智能助手、对话系统等应用提供更智能化、更准确的服务。

4.1.3 混合模型架构整合大小模型优势

基于数据敏感性要求，托管服务平台采用微调后的 ChatGLM2-6B 模型作为小模型进行槽位提取。通过 Prompt 工程引导大模型执行特定任务，灵活调整模型行为以适应不同场景。在意图识别和会议摘要等多任务中，

通过设计 Prompt 规范模型输出结构，提高响应质量。同时，大模型具备上下文理解能力，可补全多轮对话内容，提供连贯的服务。

4.1.4 实体识别算法实现精准实体匹配

针对用户口语化、非标准的槽位信息进行实体识别、模糊匹配以及歧义纠正，将其转化为后台自动化执行所需的标准化槽位数据。同时，进行完整性校验以确保数据调用的准确性。

4.1.5 自动化操作与执行引擎

“灵犀一语达”还内置了一个强大的自动化操作与执行引擎，它能够根据用户的指令自动执行复杂的操作流程。通过工程化的方式打通 APP 内的功能和数据壁垒，系统能够将用户的声音指令转化为具体的行动步骤，并在后台无缝地执行这些步骤，最终达到用户想要的结果。

4.2 应用前景

4.2.1 提高员工办公效率

通过全连接的落地实践，我们验证了＂灵犀一语达＂在内部员工应用方面具备广阔的前景，比如：1）个人办公助理：员工可通过语音指令快速执行各种办公任务，如日程安排、文件搜索、数据查询等，省去繁琐的手动输入操作。2)团队协作与信息共享: 作为团队协作的纽带，＂灵犀一语达＂可促进员工之间的信息快速共享和任务协同处理，显著提升团队工作效率。3）工作流程优化：借助自动化操作和智能推荐，＂灵犀一语达＂能够优化工作流程，减少冗余环节和步骤，使员工能够更专注于高价值工作。

通过以上方式逐步减轻员工在繁琐任务上的时间消耗，从而提升整体工作效率。

4.2.2 提升用户服务体验

托管服务管理平台的实践也证明了＂灵犀一语达＂在客户服务方面的能力。未来，该应用可落地于君弘app、富易终端、道合平台等，有效解决功能繁多导致的使用者不了解、找不到功能的问题，从而提高软件使用者的用户体验。同时，通过对语音交互记录的分析，不断改进服务质量，增强用户粘性。

4.2.3 推动数字化转型

通过智能化技术解决信息流通不畅的问题，促进数字化转型进程。构建高效协同的数字职场，为业务战略的实施提供有力支持。“灵犀一语达”能够为企业带来高效的运营模式、优质的用户体验和战略竞争力的提升。

参考文献：

[1] 机器之心
Pro.https://baijiahao.baidu.com/s?id=1809412375602596056&wfr=spider&for=pc.（网络文献）

基于多运行时的弹性云服务在证券行业场景下的应用探索

李银鹰，卢勇辉，张明，沙烈宝 | 国投证券股份有限公司 | Email：luyh3@essence.com.cn

摘　要： 云原生时代，以“弹性伸缩”的方式来处理波动负载并提升资源利用率已是一种普遍方案。区别于应用服务单元较为成熟的弹性扩展能力，经典云服务所采用的中间件和数据库技术往往并不具备弹性，而是保留传统静态资源分配方案，难以良好地匹配波动负载业务场景。本文提出一种基于多运行时的弹性云服务构建方案，以中间层的形式动态、弹性地管理底层云服务中的实例，云服务的各个实例可如同微服务实例一样实现弹性、高效的横向扩缩容。而区别于微服务弹性扩展，其可额外保障云服务实例之间的状态一致性。方案结合撮合交易这一典型的高并发、高波动应用案例进行验证，验证了方案对弹性云服务的良好弹性扩展支持能力，使得应用对云服务的波动访问需求可灵活、自动得到满足。

关键词： 多运行时框架；弹性云服务；撮合交易

一、多运行时技术与弹性云服务

1.1 云服务的弹性化

云服务作为支撑云原生应用的重要部分，提供数据存储、缓存、消息队列等多种服务能力。应用服务单元（如微服务）可将状态相关的存储、访问支持卸载至专用的云服务组件，专注于无状态的业务逻辑处理，从而实现高弹性的计算扩展。在常见场景中，云服务往往以单实例的形式供微服务访问，若微服务实例发生扩展，则单一的云服务实例可能成为访问瓶颈所在，限制了微服务弹性实例的效果。因此，往往也需要根据负载水平的变化特点对云服务实例的规格进行扩展。

通常，云服务实例扩展可通过两种基本方式实现，即纵向扩展和横向扩展。二者虽然模式不同，但均可使云服务的聚合带宽、吞吐量实现相应倍数的扩展。纵向扩展通常方式简单，但可扩展性差，难以突破单服务器可提供的资源上限；横向扩展模式通常技术复杂，但具有较好的可扩展性。然而，无论何种扩展模式，都面临资源过量供应问题，若集群中普遍存在资源过量供应问题，将拉低集群的整体资源利用率，提升企业 IT 成本。其根本原因在于，在资源分配及实例管理上采用了静态方法，而负载通常具有时间上的波动特征，二者难以良好地匹配。

云原生提倡通过“弹性计算”的方式来处理波动负载并提升资源利用率，针对云服务也提出了弹性云服务或 Serverless 云服务的模式，然而相较于计算组件（如微服务），云服务的弹性化更具挑战。近年来，云厂商推出了各具特色的弹性云服务，如阿里 RDS MySQL Serverless[1]、AWS Aurora Serverless[2]、AWS ElasticCache for Redis[3] 等，其共性在于相应的云服务实例可根据负载自动弹性伸缩，从而避免资源闲置浪费，并降低云服务运维成本。这也使得弹性云服务适用于：有明显业务波峰波谷的场景、间歇性定时任务的场景、不确定负载的场景（例如物联网、边缘计算）、期望降低运维成本并提升运维效率的场景等。

1.2 多运行时技术

多运行时框架提供了一种新型的云原生应用构建架构，使微服务应用的开发在云基础设施上可得到进一步简化。它提出了一种微服务单元的构造方法，即由“微逻辑”与“Mecha-Runtime”多运行时构成，将微服务运行所需要的包括状态管理、消息通讯、资源绑定、事件驱动、可观测、密文、配置等外移至多运行时，使微服务的开发过程可更专注于面向业务的“微逻辑”，而云上分布式计算需求的支撑能力均可借助多运行时透明地实现。

当前，微软 Dapr[4]、蚂蚁 Layotoo[5] 等框架是具有代表性的多运行时框架，为实践多运行时提供了重要支撑。以 Dapr（Distributed Application Runtime）为例，其率先提出了多运行时能力抽象的关键概念，将构建微服务应用的最佳实践设计成开放、独立和模块化的方式，每个构建块都是完全独立的，可以采用其中一个、多个或全部来构建应用。

多运行时框架提出的对中间件的抽象，可作为轻量化弹性云服务实践的重要基础。Dapr 框架构建了状态管理、发布订阅、输入输出绑定等多个与云服务紧密相关的基块，其中，状态管理可提供缓存服务支持，发布订

阅可提供消息队列访问支持，输入输出绑定可提供关系型数据库访问支持。

二、基于多运行时的弹性云服务架构

云服务的弹性伸缩不仅仅是简单地部署或回收服务副本，一方面，需保障应用无感使用动态变化的云服务集群，另一方面，还需要考虑在多实例间合理的分配数据（如关系型数据库表、KV 键值对），并保持必要的一致性（如强一致性）。因此，构建弹性云服务应至少考虑以下三方面的保障：

1）应用无感的开发运行支持。应用在开发、部署、执行的各个阶段均无需针对云服务弹性伸缩的特点进行感知及编程支持，从而实现无感透明地使用弹性云服务。

2）负载感知的弹性扩展。弹性云服务受实时的负载水平驱动，自动地调整云服务的实例数目，完成实例的自动部署及实例间的协同管理，并提供分布式化（如负载均衡等）的请求分发支持。

3）可靠的数据管理。云服务的实例数目等发生调整时，其存储的状态数据可得到有效的迁移，避免数据丢失、不一致性等情况的发生。

为此，本方案提出基于多运行时框架扩展弹性云服务能力，通过特定的运行时基块实现弹性访问及弹性控制支持，有效支撑上述需求，如图 1 所示。多运行时框架作为介于应用与云服务之间的弹性服务中间层，向上以 sidecar 形式与应用微服务模块交互，向下与底层的云服务集群进行交互。

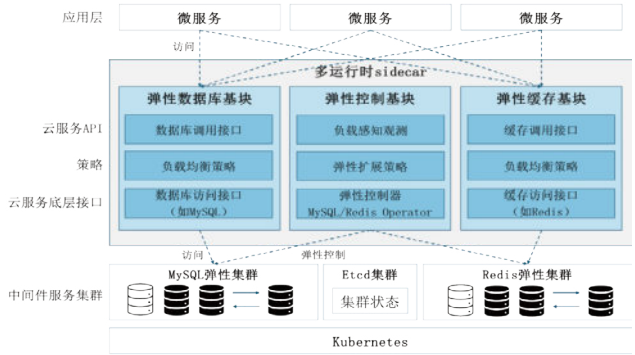


图1：面向弹性云服务的多运行时基块架构图

本方案选取最具代表性的云服务，即数据存储和缓存，构建了对数据库、缓存服务的轻量化支持，形成了多运行时的 3 个基块，包括数据库及缓存弹性服务基块支持，以及支持数据库及缓存云服务弹性伸缩的弹性控制基块。

数据库及缓存弹性服务基块采用了相同的层次结构，包括云服务 API 层、能力封装层、策略层、底层接口层等。基于此，应用服务访问弹性云服务的完整流程可划分为以下 5 个层次：

1）应用层：应用微服务通过编程语言提供的 HTTP 或 gRPC 编程库，可对多运行时层能力进行访问，多运行时层根据能力访问类型和层次的不同，提供了声明式、命令式等不同类型的访问支持。

2）云服务 API 层：云服务 API 层作为运行时模块的网关，提供对应用层云服务访问请求的解析支持，并将请求下发至相应的运行时基块进行处理。

3）策略层：根据基块的不同，该层可归纳为两类策略。一是云服务弹性访问策略，根据云服务分布式、弹性化访问的特点，提供多种类型的云服务请求处理策略，如读写分离、负载均衡、尾延迟消除等。二是云服务弹性控制策略，提供负载感知的弹性策略，支持基于负载波动情况实现实例的横向扩展，并负责在各实例之间实现状态数据的迁移同步。本方案提出一种无重分布的一致性哈希数据管理算法，可使各实例在弹性伸缩中快速实现数据状态的同步，从而提升访问效率。

4）云服务底层接口层：该层中提供了不同类型云服务的对接支持，包括关系型数据库、缓存服务等基础云服务，使应用无需面向特定云服务进行开发，便于形成统一的标准 API，也可促进策略层中形成通用的弹性服务策略以适配多种云服务。

5）云服务集群层：为保障弹性伸缩的高灵活性并维持云服务集群的轻量化，底层的各云服务实例采用无共享架构（Nothing-Shared Architecture）组成服务集群。云服务集群采用国投证券中间件平台为支撑，其可提供云服务实例的扁平化管理。多运行时中间层提供的弹性控制能力，可进一步支持不同实例组成特定的服务集群并实现数据动态维护。

三、基于多运行时的弹性云服务关键技术点

3.1 弹性云服务基块原语扩展

在多运行时弹性云服务基块中，需要针对弹性云服务提供标准化的基于 HTTP/gRPC 协议的 API 接口。本方案在维持 API 设计标准化、通用化的基础上，对关系型数据库及缓存基块的基本原语（各类 Read/Write 操作）进行了面向弹性服务的适配优化，并新增扩展了数据库 ACID 场景（如 Begin、Commit、Rollback）、缓存分布式锁（Lock）等场景中的 API。基于这些 API，无论底层云服务集群的当前构成为单实例或多实例的，应用均可平滑地在其集群上实现数据分片、数据复制、自动分库分表、读写分离等，无感实现云服务集群弹性扩展的访问。这些具体策略的执行需要结合弹性控制基块极其弹性伸缩框架进行协同设计。

3.2 云服务弹性伸缩框架

弹性伸缩框架是多运行时中负责底层云服务弹性伸缩控制的核心基块，其构成也分为云服务 API、策略层、底层云服务接口等 3 层。

以 MySQL、Redis 的弹性控制器为例，如图 2 所示，在运行过程中，数据库或缓存服务基块的 API 层通过劫持分析，产生请求负载的数据日志，并通过主动调用形式提交至弹性控制基块，弹性控制基块依据时序特征对负载数据进行管理，交由弹性控制基块策略层处理。该策略层根据根据负载监控指标（RPS/ 执行时间），利用既定的负载驱动的弹性伸缩规则进行决策，若弹性伸缩条件得到满足，策略层则形成包含弹性控制指令及数据迁移指令的弹性伸缩指令，下发给中间件接口层；接口层按指令执行，通过相应云服务的 Kubernetes Operator 驱动云服务集群伸缩，并计算数据分布位置拓扑的更新方法，进行必要的数据迁移及同步，并将集群的结构变化元数据提交至 etcd 集群。各运行时实例的弹性服务基块可通过读取 etcd 信息，实现对底层集群变化的感知，动态更新基块内的服务访问信息，包括动态调整数据库配置、动态重建数据库连接等，实现无停机的弹性扩展访问支持。

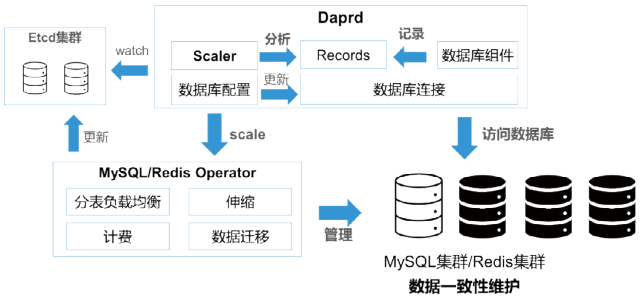


图2：弹性伸缩框架

3,3 弹性实例的访问策略

在弹性云服务的访问中，根据数据库及缓存中间件的特性，本方案主要提供了以下两个层面的支持：

数据读写分离：为各个云服务实例设定不同的读写服务职责，使得云服务可根据应用负载的读写特征进行动态调节。在数据库或缓存集群中，可以相配合地设置主从数据库 / 缓存，自动分库分表等。

负载均衡：根据云服务集群实例间关系（对等、读写分离）的不同，负载均衡策略也可进行针对性的设置，目前本方案可支持通过请求类型的识别，在不同中间件实例之间执行随机、轮询、加权轮询、最少连接数等负载均衡策略，以更好地根据数据访问的分布（如均衡分布、非均衡分布）特征来在多实例间发挥负载均衡的效果。

结合以上访问策略以及弹性控制策略的特点，本方案提出一种无重分布的一致性哈希机制，使用一致性哈希和重定向表实现数据分布管理，可高效实现数据分布计算，以轻量化的模式，使弹性服务基块可高效命中云服务实例以实现数据读写访问，弹性控制基块可在弹性伸缩发生时有效调整实例间的数据分布。

表 1 及表 2 分别展示了弹性控制器对弹性伸缩的中间件节点进行动态加入、删除管理的方法，其在实现节点管理的同时也提供了数据的一致性维护操作。具体而言，其对每个云服务实例进行标号，基于一致性哈希机制计算数据（如数据库的分库分表、缓存的数据分片等）的实例位置。当实例数目发生变动而需要调整数据分布时，则将调整要求以重定向记录的形式保存在重定向表中。

表1：基于一致性哈希策略的节点扩展算法

Algorithm 3 ADD_NODE(n1)

```
Input: New Node n1
Output: Updated redirection table
for each hash slot s managed by n1 do
    Identify the current node n0 responsible for s
    redirection_table[n1] ← n0
end for
return updated redirection table
```

表2：基于一致性哈希策略的节点收缩算法

Algorithm 4 DELETE_NODE(n2)

```
Input: Node to be deleted n2
Output: Updated redirection table
Mark n2 as read-only
Determine the new node n3 that will take over n2's hash slots
redirection_table[n3] ← n2
for each entry n2 → nx in the redirection table do
    Update to n3 → nx
end for
for each key-value pair (k, v) in n2 do
    n3.set(k, v)
    n2.delete(k)
end for
if all data is migrated from n2 to n3 then
    Remove n2 from the system
    Remove all n2 related entries from the redirection table
end if
return updated redirection table
```

四、证券行业撮合交易场景应用实践

2017 年 6 月，央行在《中国金融业信息技术“十三五”发展规划》中明确指出要稳步推进系统架构和云计算技术应用研究，支持实力较强的机构独立或联合建设金融云服务平台，面向同业提供云服务，提高行业资源使用效率，拓展云服务应用领域。近年来，国投证券以云原生领域已有实践成果为基础，开展基于多运行时的弹性云服务技术探索，并形成了以撮合交易场景为代表的弹性云服务应用实践。

撮合交易是当今市场上最为常见的交易方式之一，它利用计算机系统对买卖双方的报价进行匹配，以促成

交易。撮合交易的原理基于价格优先、时间优先的原则，这种交易方式的优势在于，它能够快速、准确地匹配买卖双方的报价，减少交易成本，提高市场的流动性。同时，撮合交易对系统并发处理能力、弹性负载处理能力有着较高要求，适用于测试云服务弹性服务能力。

结合现有撮合交易架构，研究中主要模拟高并发、波动并发交易场景，利用微服务实现交易等业务模块，利用存储中间件作为后端业务服务状态的存储支持，从而通过动态利用云端集群资源提升服务资源的利用率及吞吐性能等保障能力，借此观察本研究所形成的中间件弹性云服务在相应场景下的可行性，并基于此推演判断在其他相似场景下的技术应用潜力。具体而言，在验证中将撮合交易微服务模块的云服务访问（包括数据库及缓存）实现使用本方案所提出的多运行时框架进行支持，如图 3 所示，以负载发生器模拟用户客户端行为，进行随机波动的负载请求，撮合交易微服务将根据云端弹性扩展器的管理，进行实时的弹性扩展，而同时其通过多运行时驱动，使得相关中间件服务同步弹性扩展，实现面向高波动负载压力下计算及存储的协同弹性服务。

实验表明，本方案使得撮合交易微服务在波动负载下的吞吐获得了显著提升，如图 4 所示。首先，当撮合交易微服务副本数为 1 时，允许 Redis 实例弹性扩展，服务响应的 RPS 维持在 3000 左右，说明单微服务实例的吞吐上限为 3000RPS，如图 4(a) 所示；当固定 Redis 实例数，允许撮合交易的微服务副本弹性扩展，服务响应的 RPS 仍维持在 3000 左右，这说明单 Redis 实例可支持的吞吐上限为 3000RPS，如图 4(b) 所示；当允许撮合交易微服务副本及 Redis 弹性实例均进行弹性扩展时，服务响应 RPS 呈现出线性增长的趋势，在基准测试场景中，可达到 15000RPS，这说明在微服务及中间件弹性伸缩的配合下，使得撮合交易系统的吞吐实现了显著提升，如图 4(c)

结合本方案在撮合交易场景下的验证效果，我们认为在以微服务及中间件服务为核心技术要素的经典金融应用云服务场景下，基于多运行时的弹性云服务技术可作为提升应用服务可扩展性及其效率可参考技术形式。

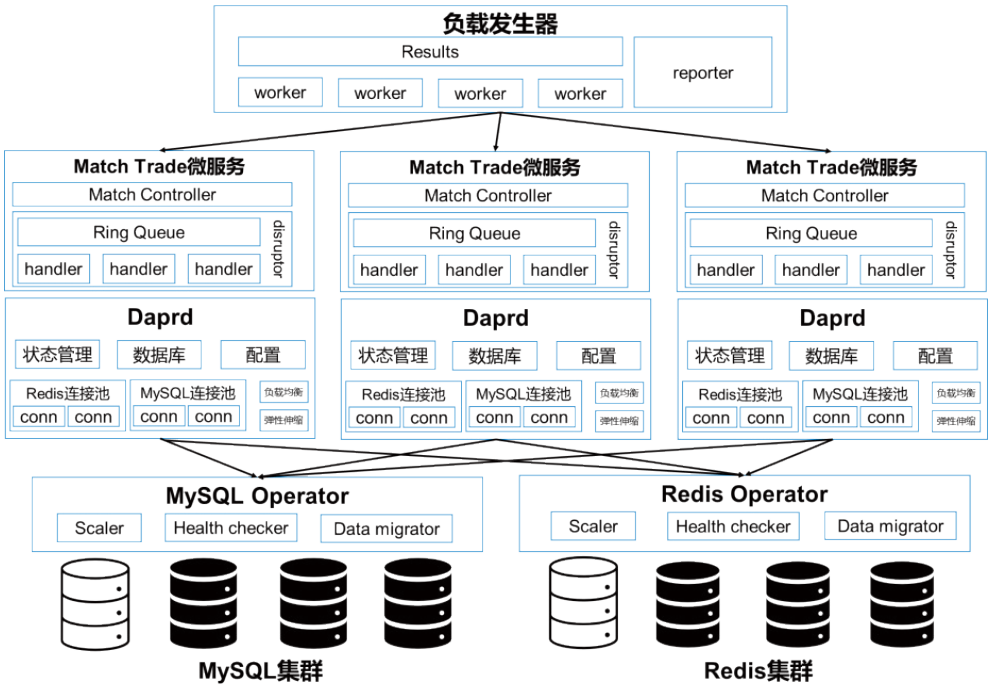
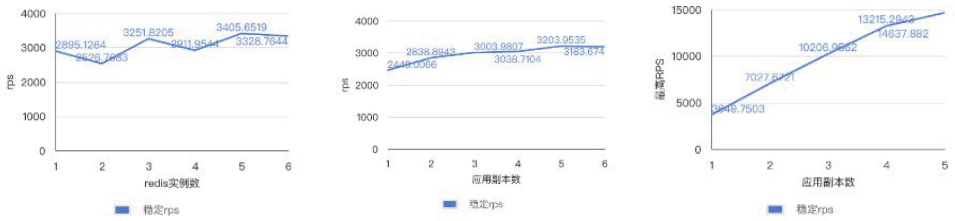


图3：基于弹性云服务的撮合交易架构示意



(a)仅弹性扩展Redis / (b)仅弹性扩展撮合交易微服务 /
(c)同时弹性扩展中间件及撮合交易微服务

五、总结与展望

本研究面向弹性云服务的构建，提出基于多运行时框架的弹性云服务实现方案，并针对数据库及缓存服务实现了弹性支持。在框架中，构建 3 组新的运行时基块，包括面向数据库及缓存的云服务基块，以及负载感知的运行时基块，提出了一种五重分布的一致性哈希算法驱动弹性实例访问及数据分布策略。在基于撮合交易的实验验证中，本方案使得撮合交易微服务在波动负载下的吞吐获得了显著提升。

基于此次探索，我们意识到基于多运行时框架实现对中间件弹性服务的构建有着巨大的技术及成本收益，同时多运行时框架也是以云原生形式在应用中接入一些其他服务特性的有效技术路线。在后续探索中，我们将围绕中间件弹性服务能力建设等场景，继续扩展多运行时中间层，形成更广泛、深入的云原生服务能力的探索和实践。同时，扩展潮汐弹性伸缩方法、冷启动方法，进一步提升云服务弹性扩缩容效率。

参考文献：

[1] 阿里云 . 云数据库 RDS 简介 [R/OL]. 2024.10. https://help.aliyun.com/zh/rds/product-overview/what-is-apsaradb-rds.
[2] 亚马逊云科技 . AWS Aurora Serverless 产品技术介绍 [R/OL]. 2024.5. https://aws.amazon.com/cn/rds/aurora/serverless/.
[3] 亚马逊云科技 . AWS ElasticCache for Redis[R/OL]. 2024.5. https://aws.amazon.com/cn/elasticache/redis/.
[4] Dapr. https://dapr.io/. 2024.5.
[5] Layotoo. https://cloudnative.to/blog/mosn-layotto-intro/. 2024.5.

创新压力测试技术 筑牢系统安全防线
—— 广期所研发建设高性能压力测试平台

苏恒志，董琳 | 广州期货交易所 | Email：suhz@gfex.com.cn

摘　要： 随着资本市场的发展，证券期货业对交易系统的性能和容量要求不断提高。为应对这一挑战，广州期货交易所按照证监会要求，全面加强系统压力测试，优化测试方法和技术，研发了一套高性能压力测试平台。该平台具有真实高效的测试场景设计、微秒级阶梯压力控制、高精度可视化性能监控等特点，有效保障了交易所的安全生产。本文介绍了平台的建设情况和应用成效，分享了系统建设经验，供业界交流参考。

关键词： 交易系统；压力测试；性能；容量；安全运行

一、引言

近期，证券期货市场交易活跃，日交易量和瞬时交易笔数屡创新高，核心交易系统的安全运行面临严峻挑战。常态化开展压力测试、加强系统性能容量分析是防范交易系统运行风险的重要手段。广州期货交易所（以下简称广期所）自 2021 年成立以来，一直高度重视系统压力测试工作，从 1.0 业务系统建设之初就一并引进了时延监控平台和性能测试工具，对业务系统的性能和容量等核心技术指标进行评估和监测，并且按照《证券期货业网络和信息安全管理办法》《证券期货市场交易结算核心机构信息科技管理暂行办法》的要求，定期开展生产系统压力测试。2024 年，广期所为更好开展压力测试工作，不断优化压力测试方法、改进压力测试技术，规划建设了一套自主可控的高性能压力测试平台。2024 年 7 月，压力测试平台一期建成上线运行，测试效率和效果得到大幅提升，已能够满足广期所目前全部信息系统的性能测试需要。

二、压力测试平台建设需求

广期所压力测试平台的总体建设需求是建成一套符合我所业务发展需要、自主可控、技术先进的压力测试平台，其中平台一期主要实现性能测试需求，二期主要实现可靠性测试需求。平台一期的具体需求如下：

1. 建设一套同时支持接口、网页端、客户端压力测试的测试框架，并集成广期所现有性能测试工具，能够支持交易系统、结算系统、监查系统、会员服务系统、电子仓单系统、门户网站等广期所全部信息系统的性能测试。

2. 实现压力测试场景设计、压力构造引擎、性能指标监控、性能结果分析、性能瓶颈定位等功能，以支持验证《证券期货业信息系统压力测试指南》《资本市场交易结算系统核心技术指标》中要求的各项核心技术指标。

3. 压力测试平台满足信创要求，支持信创服务器、操作系统、数据库、中间件的部署，支持信创版本业务系统的性能测试。

4. 压力测试平台提供对外接口，能够通过接口集成或调度其他性能测试工具或功能插件，支持二次开发，易于维护和扩展。

三、压力测试平台核心功能

3.1 全类型系统的测试覆盖

压力测试平台一期的设计可以全面支持广期所各类型信息系统的性能测试需求，确保这些系统在高负载和复杂业务场景下能够稳定运行。典型涵盖的业务系统包括：交易系统（后台实时服务系统），包括验证订单处理、成交确认等核心流程的性能；结算系统（C/S 架构），确保资金结算过程在高负载情况下的稳定运行；监查系统（C/S 架构），验证相关数据处理能力，以及超过三倍历史峰值数据量下数据整理、盘后评分、初始化时长；会员服务系统（WEB 架构），评估会员管理和高峰时段的服务表现；电子仓单系统（WEB 架构），测试仓单生成、存储和查询效率。

3.2 真实高效的测试场景设计

广期所压力测试平台具备灵活高效的场景设计能力，允许用户根据实际业务流程创建复杂的测试案例，特点包括：

1. 灵活的场景设计。用户可以基于交易所实际业务需求，设计出符合真实业务逻辑的测试场景，如独立设置期货期权报单比例、成交比例、撤单比例、套利比例等。

应用场景涵盖从订单生成到成交确认的整个交易流程，确保了测试场景贴近真实环境，提高测试结果的准确性和实用性。

2. 矩阵式压力配置生成算法。用户能够模拟生产实际会员席位登录负载均衡的场景，用自研的线性方程组系数矩阵算法精确计算各类复杂的业务参数，并通过压力引擎报单，有效保证测试场景的真实性和准确性，从而能够更精确地模拟实际业务中的交易行为和系统负载。

3. 精准的压力控制。平台压力引擎采用多线程发送机制，在持续施压过程中可以动态调整压力值，观测被测系统实时状况，直至探测出性能瓶颈。各线程可对每秒报单速率实施微秒级的压力控制，通过时间切片技术进行一秒内的压力分配，可均匀分布也可按指定压力频率分段集中报入，模拟各会员在实际交易中的真实操作。

3.3 全方位的性能指标监控

为了提供更为详尽准确的测试反馈，广期所压力测试平台引入了先进的监控机制，涵盖从底层硬件到上层应用软件的各个层面，具体包括：

1. 核心系统应用链路监控。涵盖交易前置、前总线、资金、撮合总线、撮合引擎等关键模块的性能情况，通过实时采集日志并动态展示相关指标，及时发现并解决潜在问题。

2. 关键基础软硬件监控。实时跟踪并记录被测系统中所有服务器及进程的资源使用情况，及时优化资源配置，提高系统运行效率。监控内容包括 CPU 利用率、内存使用率、磁盘 I/O 速率等。

3. 可视化实时监控大屏。提供直观的图表和仪表盘，展示关键性能指标，大屏内容包括系统整体运行状况、各项性能指标的变化趋势、核心模块的吞吐量以及响应时间等关键指标。同时提供历史数据分析，支持对过去一段时间内的性能数据进行回顾和对比。

3.4 灵活开放的可扩展接口

压力测试平台可以灵活地集成各种实用工具和技术组件，包括查询交易状态、场上入金、启停交易系统服务等测试工具。压力测试平台还集成了 JMeter 等开源工具，扩展了平台的功能和适用范围，提升了测试的多样性。

四、压力测试平台技术架构

广期所压力测试平台一期的技术架构主要包括三个子系统：平台管理服务子系统、压力测试引擎子系统以及监控子系统。整体架构设计采用了微服务架构，各子系统之间通过 API 接口进行通信。

4.1 平台管理服务子系统

平台管理服务子系统提供了高效、易用且可扩展的测试平台管理解决方案，确保测试过程的顺利进行和测试结果的有效分析。平台管理服务子系统的架构如图 1 所示。

该子系统的功能可分为实施管理、基础管理和环境管理三个类别，具体如下：

1. 实施管理相关。用户通过图形化界面定义复杂的测试场景，提供参数化配置和预定义模板，简化复杂场景的创建。允许用户创建测试任务，选择测试场景，可自动收集响应时间、吞吐量等指标数据，并通过折线图、柱状图和自定义报表展示。

2. 基础管理相关。支持多角色分配，每个角色具有不同的权限集，如创建测试场景、执行测试任务和查看测试结果。提供集中式的配置界面，用户可以配置各种系统参数。

3. 环境管理相关。支持创建和管理多个测试环境，

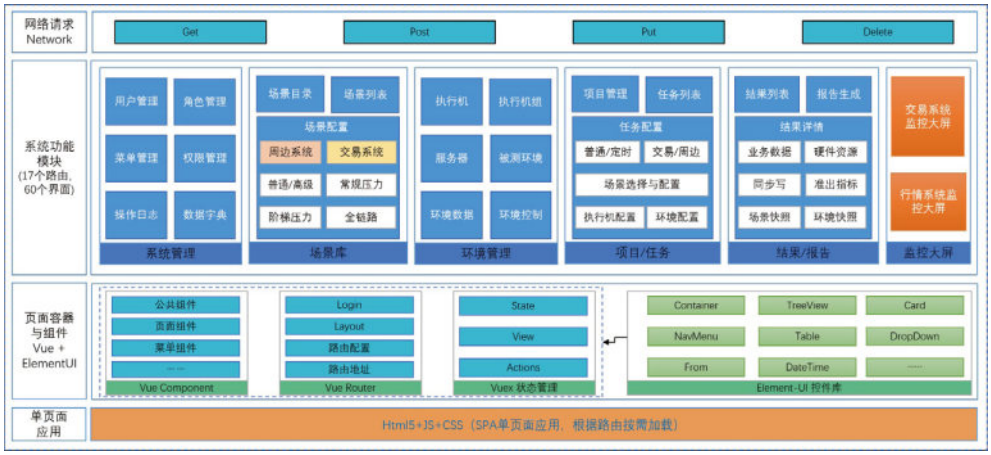


图1：平台管理服务子系统架构图

允许用户部署和重置执行机，选择不同的压力引擎版本，并支持自动化部署和批量操作，提高管理效率。

4.2 压力引擎子系统

为了满足交易系统在高负载和复杂业务场景下的性能测试需求，压力引擎子系统被设计成能够模拟大规模并发交易，支持高吞吐量的订单生成、处理和确认。同时，该子系统采用自研的线性方程组系数矩阵算法，精确计算各类复杂的业务参数，能够真实地模拟各种交易场景。压力引擎子系统设计如图 2 所示。

压力引擎采用多线程和异步 I/O 技术，这不仅加速了测试进程，减少了等待时间，还提升了整体测试效率。压力引擎运行主流程如图 3 所示，具体如下：

1. 实例化压力引擎实例。创建并初始化压力引擎对象，设置其基本参数和配置。该实例负责管理整个压力测试过程，包括订单生成、处理和确认。

2. 获取全局数据对象。加载并初始化全局数据对象，该对象包含所有必要的配置信息和其他共享数据。全局数据对象确保各个模块可以方便地访问和使用这些数据，

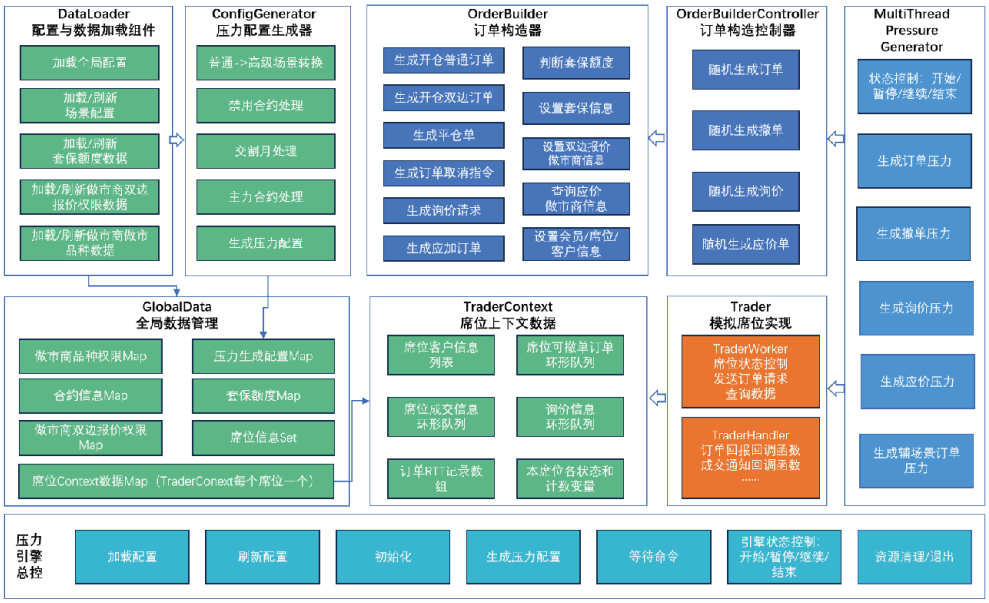


图2：压力引擎子系统架构图

从而提高系统的整体性能和一致性。

3. 加载全局配置文件。读取并解析全局配置文件，该文件包含了压力引擎的所有配置参数。配置文件通常包括测试场景、压力级别、并发数、订单类型等关键参数，确保压力引擎按照预期的配置进行工作，提供灵活的测试环境设置。

4. 轮询检查状态控制文件。持续监控压力引擎的状态控制文件，该文件用于动态控制压力引擎的行为。一旦文件的修改时间发生变化，读取文件的第一行内容，并根据指令执行相应的操作。支持的操作指令包括开始测试、停止测试、暂停测试、恢复测试等，使得用户可以实时控制压力测试的过程。

图3：压力引擎运行流程图

4.3 监控子系统

压力测试平台的监控子系统作为用户与监控数据的交互界面，监控大屏以图表、仪表盘等形式直观展示关

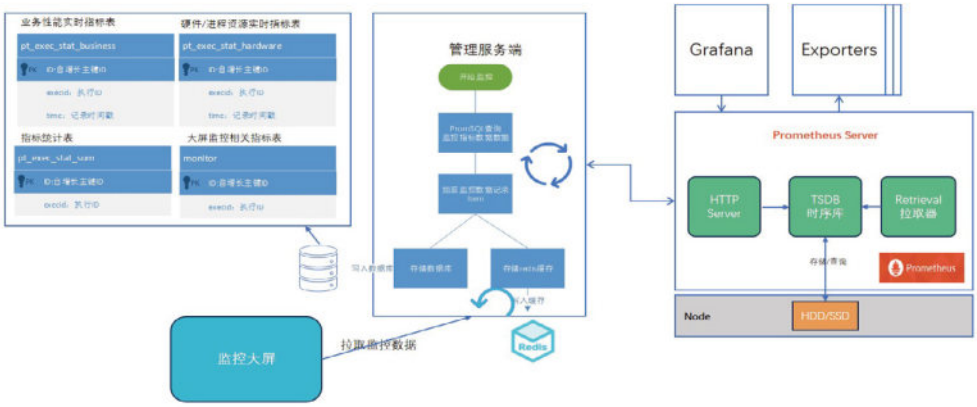


图4：监控子系统架构图

图5：监控大屏（交易系统）效果图

键性能指标，实时监控系统运行状态。监控子系统的核心架构由 Prometheus 和定制 Exporter 组成，系统架构如图 4 所示。

监控大屏界面展示了各种性能指标，通过实时采集并动态展示相关指标，可以及时发现并解决潜在问题，效果如图5所示。监控大屏（交易系统）主要监控内容包括：交易前置的实时响应时间、订单请求速率、成交通知速率；前总线的每秒接收的数据包数量、缓存区长度；资金模块的每秒订单请求个数、成交通知个数；撮合引擎的订单请求速率、成交通知速率等。

通过上述设计，压力测试平台能够全面覆盖从测试场景的创建到测试任务的执行，再到测试结果的分析，形成一个完整的闭环。监控子系统不仅确保了压力测试平台运行的稳定性和可靠性，还为用户提供了实时、详细的性能反馈，有效支持了广期所各业务系统的稳定运行。

五、压力测试平台应用效果

广期所压力测试平台一期建成上线后，在新品种上线、业务系统改造等各信息系统项目中扮演了至关重要的角色，并取得了显著的成效。本章将介绍压力测试平台的部署情况和实际应用情况。

5.1 平台部署方案

广期所压力测试平台采用轻量化部署方式，目前分别部署在系统测试环境和类生产测试环境上。部署架构如图6。

压力测试平台由一台数据库服务器、一台文件服务器、一台管理 / 监控服务器和一组压力机集群构成。平台使用东方通作为中间件，被测系统服务器部署了监控和传输引擎探针，用于收集并回传服务器监控指标和相关日志至数据库，以支持后续的结果分析。压力机集群同样部署了监控和传输引擎探针，用于监控各压力机状态、可用资源，便于灵活调度压力引擎。

平台数据库包括关系数据库和时序数据库。关系数据库使用达梦数据库用于存储常规数据，如用户管理、权限管理、测试场景等等；时序数据库用于处理监控管理模块中对实时处理要求较高的数据。数据库需要支持定期备份保证数据可靠。

平台支持全栈信创服务器部署，所有用于部署的服务器均可使用虚拟机，硬件成本较低，可通过增加虚拟机的方式动态扩展压力机资源。

5.2 实际应用效果

压力测试平台已经在多个重要项目中得到广泛应用，并取得了显著的成效。以下是几个典型的应用案例：

1. 核心交易系统压力测试。近期，为了维护资本市场的安全、平稳和高效运行，科技监管司根据交易所信息系统的特点和压力测试有关规定，组织各交易所进行核心交易系统的压力测试。广期所使用压力测试平台顺利完成了测试任务，通过模拟大规模并发用户访问和交易操作，检验了核心交易系统在高负载下的处理能力，验证了系统在极端条件下能够稳定运行。

2. 拟上市新品种上线压力测试。期货交易所新品种上市、新业务上线前，必须对技术系统承载能力进行评估。广期所压力测试平台针对新品种上市设计了专门的测试模型，能够模拟新品种上市的交易活动，包括期货期权报撤单、套利、双边报价、交易、闭市、结算等核心业务场景。通过这些场景测试，能够客观评估技术系统对新品种上市的承载能力。

3. 支持业务系统信创改造性能测试。随着行业信创工作的深入推进，广期所 2024 年需要完成核心业务系统信创改造，因此需要在国产软硬件环境下对改造后的业务系统进行性能测试，并验证系统的兼容性和稳定性。广期所压力测试平台支持信创服务器、操作系统、数据库和中间件，不但验证了交易、结算、会员服务等多个信创业务系统的性能和容量，还发现了国产关键软件的一些性能问题和缺陷。

六、总结与展望

广期所压力测试平台建设是信息系统建设五年规划中的重要内容，对于保障交易所安全生产、支持品种上市、促进业务创新、推动技术创新有着重要意义。利用压力测试平台，每年例行开展 1-2 次生产系统压力测试，检验系统的性能、容量和稳定性，确保交易所系统安全稳定运行；在新品种新业务上线前，开展业务系统承载能力的技术评估，保障新品种新业务平稳上线；在系统信创改造、新一代系统研发过程中，通过不断的压力测试验证系统及关键组件的技术指标，保障技术创新质量，提升技术创新效能。

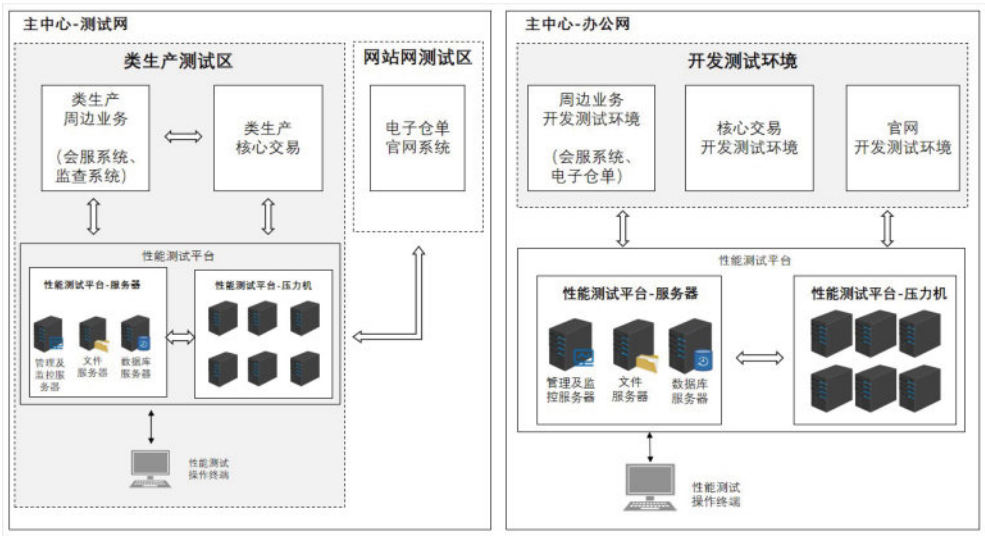


图6：压力测试平台部署架构图

为了实现上述目标，广期所压力测试平台在以下三个方面进行了技术创新：一是真实高效的测试场景设计，基于生产数据设计测试模型，通过矩阵式压力配置算法模拟实际业务场景。二是微秒级阶梯压力控制，通过微秒级时间切片技术和阶梯压力控制技术，进行精准压力构造，模拟真实压力和负载场景；三是高精度可视化性能监控，通过集成实时监控组件，实时监测、采集和展示关键性能指标，为用户提供详细的性能报告和图表分析结果。以上特点使得压力测试平台不但能够满足目前广期所生产系统的测试需要，也能支持信创版本业务系统及未来新一代业务系统的测试需要。

2025 年，广期所将进一步完善压力测试平台：一是引入逐笔订单监控技术，实现更细粒度的性能分析；二是引入生产流水反演技术，实现更真实的压力测试场景；三是引入混沌工程技术，实现可靠性测试功能；四是增加会员端压力测试子系统，完善全市场压力测试机制。

03 实践探索

上海证券面向自营业务的投资管理平台建设实践

牟大恩，宋娜 | 上海证券有限责任公司 | E-mail：moudaen@shzq.com

摘　要： 自营业务作为证券公司的重要收入来源之一，在证券公司发展中扮演着重要的角色。面对投资环境的日益复杂化和业务需求的多元化，如何实现投资管理的系统化与精细化，提升投资管理效率等问题，已经成为业务发展的核心关注点。为此，上海证券立足公司自营业务，建设了新一代投资管理平台。平台旨在为投资管理提供全链路追踪、深入的风险绩效分析、全面的风险管理和详细的投资复盘等能力，提升工作效率，辅助优化投资决策流程，为自营投资业务的数字化转型升级提供强有力的平台支撑。

关键词： 自营业务；投资管理；全链路追踪；绩效分析；风险管理

一、引言

2023 年中央金融工作会议首提“数字金融”，倡导数字技术融合传统金融，驱动金融转型升级。数字经济时代，证券业机遇与挑战并存，数字化转型成为关键驱动力。

自营业务是证券公司推动业务创新与拓展的核心板块，其发展对于推动公司业务多元化、差异化布局，提升公司市场竞争力具有举足轻重的作用。随着资本市场的蓬勃成长，投资品种日益繁多，投资环境也愈发复杂多变，业务需求随之呈现多样化趋势。在此背景下，如何实现投资管理的系统化、精细化和智能化，有效提升管理效率，已成为业务推进中的关键议题。与此同时，金融市场的复杂性与风险性不断攀升，监管要求也日益严格。面对市场风险、信用风险、流动性风险及操作风险等多重考验，投资风险管理工作正面临着前所未有的挑战与更高要求。因此，自营业务的风险管控已成为公司稳健运营中不可或缺的关键环节。

二、项目背景及意义

当前自营业务开展中，业务人员往往面临着多重挑战：繁琐的跨平台手动操作，缺乏自动化投资分析工具及风控提醒机制，报表的编制也大多依赖手工操作，数据指标口径不统一等，缺乏线上化、数字化、智能化的平台来支撑统一投资管理。

1) 投资经理通常需跨多个系统获取数据，手工对数据进行处理，导致在投前决策、实时监控和组合管理等流程管理方面工作效率偏低，同时缺少投资全流程的串联和留痕。

2) 在投资交易系统内完成交易指令下达后，盘后缺少对成交指令、投资目标和投资策略层面的复盘分析。

3) 事中、事后的触发风控事件需风控人员线下进行通知投资经理，且业务端和风控端指标口径有时存在不一致的情况，内控管理亟需拉齐。

4) 在指标及报表的统计方面当前多依赖手工，耗时、低效且准确性难以保证，开发和维护难度大，难以持久应用。

为解决以上问题，上海证券科技团队立足自营投资痛点问题和需求，建设了面向自营投资业务的投资管理平台。此平台旨在增强投资流程的主动管理能力，促进投资业务的高效运行，并为证券投资管理的各个环节、各种场景提供全面的科技支撑。

三、平台建设方案

平台基于云原生微服务架构，运用人工智能、大数据、云计算等技术，围绕投研、投资、风控和业绩归因等关键环节，实现了完备的数据整合能力、金融工程建模能力及灵活敏捷的数据展示能力。平台贯穿投资管理过程中投前准备、投中管理和投后分析的全链路环节，实时监控投资全链路风险，提升投资管理工作效率，客观归因投资组合的业绩收益，辅助支持投资决策，助力实现更高效、稳定、科学的自营投资业务管理。

3.1 平台架构

上海证券投资管理平台的总体逻辑架构如图 1 所示：

在数据处理层，平台实现对数据采集及加工、数据调度及监控和数据校验的统一管理。通过将投资交易数据、行情数据、资讯数据、产品数据、金融法规数据等进行融合，为金融模型构建、组合分析、组合画像、合规试算、绩效归因等投前、投中和投后管理提供强有力的数据支撑。

在微服务层，平台基于 Spring Cloud Alibaba 的微服务架构，构建了公共基础服务、数据服务、金融指标计算服务、实时估值服务、投资分析服务、风险绩效分

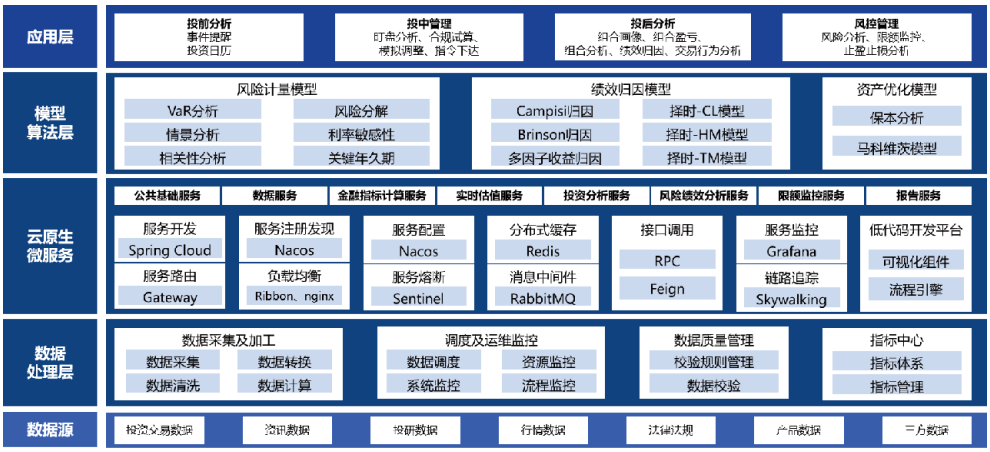


图1：投资管理平台逻辑架构图

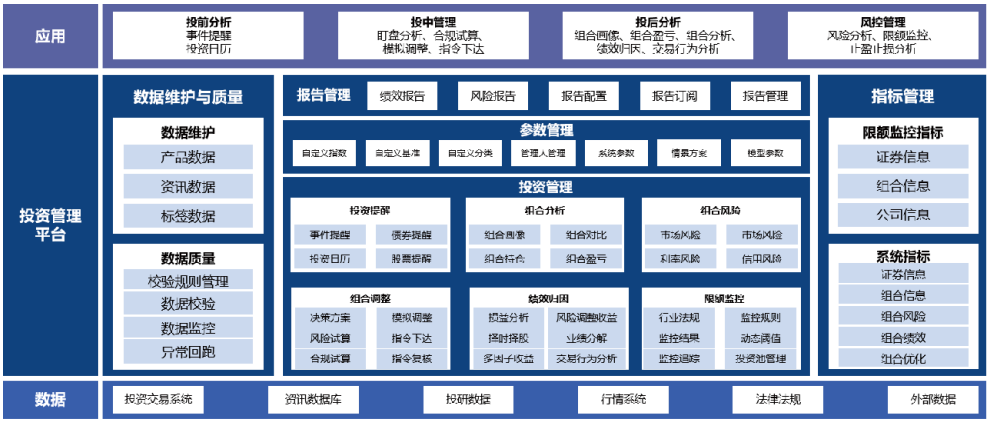


图2：投资管理平台业务架构图

析服务、限额监控服务、报告服务 8 大核心服务。灵活实现了服务拆分与独立部署、弹性伸缩与负载均衡、智能路由与负载管理、服务监控与日志管理等能力。

在模型算法层，平台构建了智能算法模型为投资决策和风险管理提供智能算法支撑，涵盖绩效分析模型、风险计量模型等。基于平台采集整合的市场资讯数据、持仓交易数据等，经过数据预处理、特征工程、模型构建、模型评估的完整模型构建流程，进而实现模型在投资决策及风险管理各场景的具体应用。

在数据层、服务层和模型层之上，以投资全链路管理为核心构建投资管理平台应用，覆盖投前、投中、投后全链路的投资分析及风控管理，高效赋能投资决策与风险管理，实现投资流程的全链路风险可控。

3.2 平台能力

投资管理平台实现的主要业务能力涵盖以下三大方面：平台基础服务、投资管理分析服务、风控管理服务，业务架构如图 2 所示。

3.2.1 平台基础能力

平台基础能力主要提供数据处理及个性化操作控制台管理等相关服务。

1) 数据服务

数据服务提供投资分析、风险管理等应用场景下的数据加工、分析、管理等，将投资交易数据、行情数据、资讯数据等投资场景数据采集整合，实现多源异构数据的加工、计算及集中存储。通过数据调度系统，及时追踪监控数据的采集加工结果。为保证数据的可靠性，平台定义了数据校验规则，实现了数据自动稽核功能，从多个维度对转换后的数据进行核对，包括数据索引、数据间的勾稽关系、数据完整性等，并针对异常数据分类进行实时提醒警示，对于部分异常数据人工确认调整后进行任务回跑，自动执行。

2) 指标管理

平台构建了三层金融指标体系，覆盖证券信息、组合信息、组合风险、组合绩效、组合优化等一级大类，共计 1000+ 指标，每个指标都有明确的定义和清晰的计

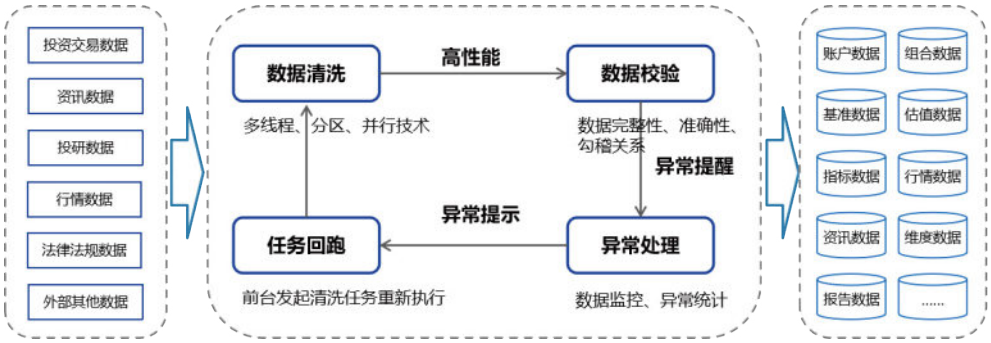


图3：数据处理流程逻辑图

算口径，为投资分析提供了完备详尽的指标参考，业务人员可以基于具体场景选择合适的指标数据进行统计分析，辅助投资决策和风险管理。

3) 报告管理

报告管理支持监管所需各类报表、风险管理所需的标准报表以及自定义个性化报表，实现对报表的自动化、系统化、规范化管理。报告支持灵活的分析视角，提供时间段、业务类型、投资经理、账户、组合等不同维度的风险报告和绩效分析报告。报告关联的指标口径由统一管理入口维护，保证了相关报表数据口径的一致性。

4) 个性化视图

利用低代码开发平台实现业务单元模块化和组件化管理，极大地简化了开发流程，增强了业务使用体验。业务人员只需通过拖、拉、拽操作即可构建个性化分析视图，满足多样化的数据分析需求。每个组件都具备高度的自定义能力，用户可根据实际需要设置特定指标进行展示。同时，内置的工作流引擎，支持用户根据实际业务流程需要，灵活配置审批、通知、任务分配等自动化流程，以确保了业务处理的准确性和合规性。

3.2.2 投资管理及分析能力

投资管理及分析能力为投资人员提供了投资提醒、投资分析、风险试算及交易行为管理等服务。

1) 投资提醒

结合证券资讯信息，平台为投资经理提供组合持仓证券的业务事件提醒。支持债券业务、股票业务、回购业务等不同业务类型。事件内容包括投资机会、投资风险、限额流通、公司行为等关注内容，如新股申购、新债申购、转债申购。在每个业务事件提醒下，可查看同一证券在不同组合中的关联持仓情况。同时，平台采用日历视图形式直观展示持仓证券每日事件提醒的汇总及概览，便于快速捕捉关键信息。事件提醒支持消息盒子和邮件订阅等不同订阅选项，确保投资经理能够及时有效的接收投资提醒。

2) 组合调整及风险合规试算

平台提供组合的决策构建及实时模拟调整。通过合规、风险试算提前监测交易风险，并通过对比持仓模拟调整前后的收益、风险指标变化，提供资产配置和风险管理策略的优化方案，进而最大程度实现预期收益并控制风险。其中，模拟调整功能展示单组合全部持仓明细，调整持仓的同时可即时查看调整对组合的影响，例如久期、凸性、偏离度等。投资决策功能支持外部导入和手动新增来构建用户权限组合的投资决策，并可对决策进行查看、保存、修改等操作。调整试算包括合规试算和风险试算。合规试算调用外部合规计算引擎（如交易系统风控、合规中心风控等），计算调整后是否触发合规限制。风险试算对选定投资组合的敏感性或风险指标进行调整策略前后的试算分析，并支持拆解至个券维度，以进一步细化风险分析。

3) 多维度组合分析

组合分析模块实现投资与风控人员共享同一个金融

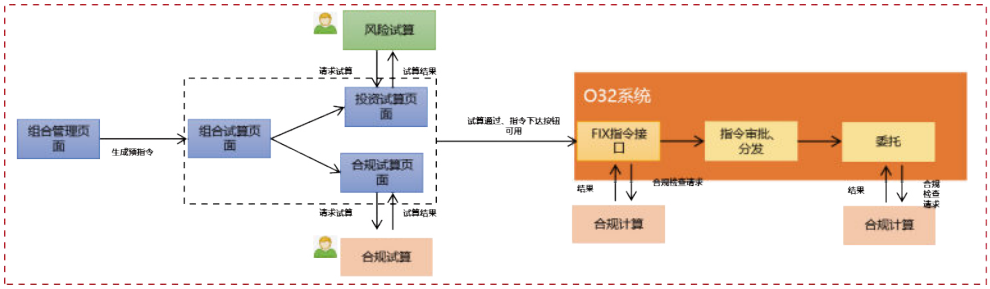


图4：组合调整及风险合规试算流程逻辑图

指标库和投资组合分析功能，支持组合风险、组合绩效、情景分析等多维度组合分析。组合画像从组合概览、资产配置、持仓分析、流动性分析等维度，刻画单组合或多组合的概况。组合对比提供不同组合间的概览对比、持仓对比信息等，捕捉组合间差异信息。组合查询提供组合资产配置集中度和仓位变化、资产、组合盈亏情况分析功能，组合绩效分析支持调仓收益分析、业绩分析、择股择时能力、Campisi 归因绝对收益、Campisi 归因相对收益、Brinson 归因、Campisi 即期、多因子收益归因等不同归因模型的分析，实现投资组合的多维度、穿透式管理分析，辅助投资策略的迭代优化，提升投资管理的工作效率。

4) 交易行为管理

平台支持交易行为分析，提供指令成交分析、止损分析、策略回测等不同维度分析。通过策略指令维度分析验证交易策略的有效性，为投资策略优化提供决策支持，实现交易指令下达与指令分析等交易行为闭环管理。通过从交易系统组合层提取交易指令，在投资管理系统内对指令流水添加交易策略标签，并将组合和策略标签关联，实现交易策略的多维分析。同时，结合自定义模拟组合构建功能进一步实现跨组合的灵活交易行为分析。

3.2.3 风控管理能力

投资风险全链路管理功能涵盖了风险识别、计量、分解、监控及预测等多个环节，针对组合中不同资产类别从多个维度进行深入的风险分析与控制。

1) 组合风险分析

平台提供投资风险全链路管理的功能，包括识别风险、计量风险、风险分解、风险监控、预测风险。针对组合不同资产类别从市场风险、流动性风险、利率风险等角度进行风险分析与控制。同时，平台支持自定义维度对绝对风险或相对风险在任意时间段内进行计算和分解。风险分析包括 VaR 分析、利率敏感性、风险分解、情景分析、关键年久期、相关性分析、历史风险回顾、VaR 后验、变现分析等，并支持用户根据自定义维度对风险指标进行计算和分解，为投资风险管理提供全方位指标透视和决策依据。

2) 限额监控管理

限额监控支持分类、分层、分级的多层级穿透式风险限额管理，用于监管合规、风险限额、止损提示等场景。平台支持用户设定各类合规指标、风险预算指标、自定义指标等各类关键指标，支持指标动态计算、监控指标设置动态阈值，通过系统消息推送或者自动化邮件形式及时预警提醒，以及时采取有效控制和防范措施。同时，平台支持对预警场景的统计分析和持续跟踪，实现风险事件处理的监测落实和追踪管理。

图5：组合分析功能介绍

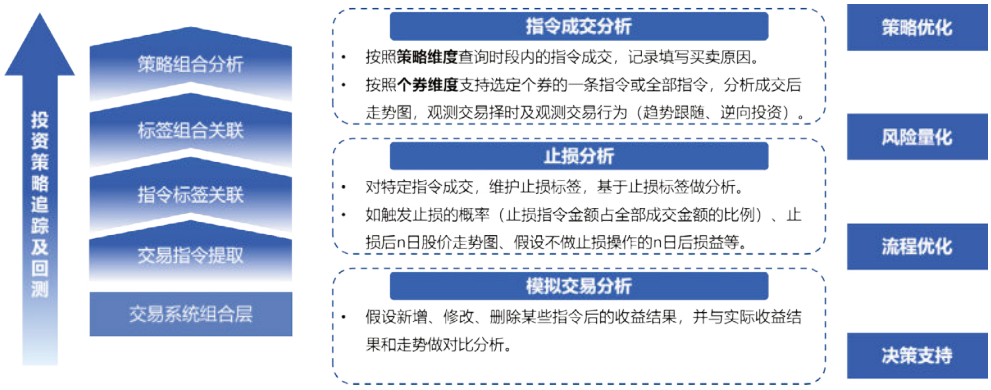


图6：交易行为管理功能介绍

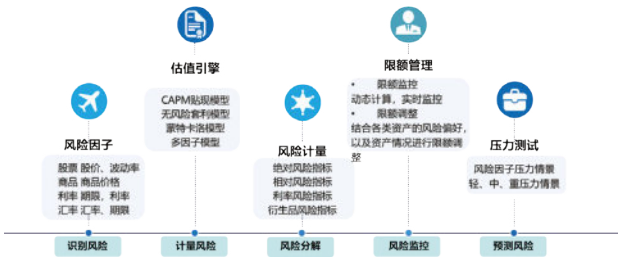


图7：投资风险管理功能介绍

四、平台特点及价值

平台基于大数据和人工智能相关技术，助力自营投资业务实现数据驱动的科学投资决策和全面风险管理，提高投资效率并降低投资风险。

1) 多维度可视化的投资场景管理高效赋能业务全流程追踪与监测

紧密结合投资业务流程，平台整合投资决策的完整链路，贯穿整个投资管理过程的投前准备、投中风控和投后绩效归因的投资管理全链路，支持投资决策过程的全链路追踪和持续优化。在投前分析阶段，平台帮助投资经理评估市场潜力和风险，提供决策支持。在投中管理阶段，平台实时监测和提醒投资风险，及时发现问题并提供风控警示。在投后复盘阶段，平台通过组合绩效归因，促进投资策略的持续优化和迭代。平台实时监控投资风险，客观归因组合的业绩收益，提升投资管理工作效率和投资决策准确性，有效控制投资风险，助力实现更高效、稳定、科学的自营投资业务全链路管理。

图8：投资决策一体化全流程管理示意图

2) 丰富的指标库及高效的计算引擎助力多维度立体化组合分析

平台构建了丰富的金融指标体系，实现对指标的统一管理。金融指标体系为三层指标体系，覆盖证券信息、组合信息、组合风险、组合绩效、组合优化等一级大类，共计 1000+ 指标，为投资分析提供了完备详尽的指标参考。同时，平台打造了高效的计算引擎，通过采用分布式、缓存和异步并发等先进技术，实现多维度实时金融指标的计算。丰富的金融指标库及高效的计算引擎，为投资决策提供了低时延、高并发、高可靠的指标服务，助力实现投资组合的多维度立体化分析。

3) 智能算法模型支撑数据驱动的科学投资决策与风险管理

平台搭建了智能算法模型引擎，如图 10 所示，为投资决策和风险管理提供智能支持，涵盖绩效分析模型、风险计量模型等。基于平台采集整合的市场资讯数据、持仓交易数据等，经过数据预处理、特征工程、模型构建、模型评估的完整模型构建流程，进而实现模型在投资决策及风险管理各场景的具体应用。绩效分析模型包括 Brinson 归因模型、Barra 模型、Fama-French 因子模型等，从择时择股能力、风险调整收益、业绩归因等不同角度归因评估组合，为投资策略的迭代提供科学准确的参考。风险计量模型包括蒙特卡罗模型、无风险套利模型、流动性风险模型、市场风险模型等，从个股、组合、行业等不同维度监测风险，提供组合风险指标的动态计算和动态止盈止损，并进行实时监控和预警提醒。

图10：智能算法模型引擎示意图

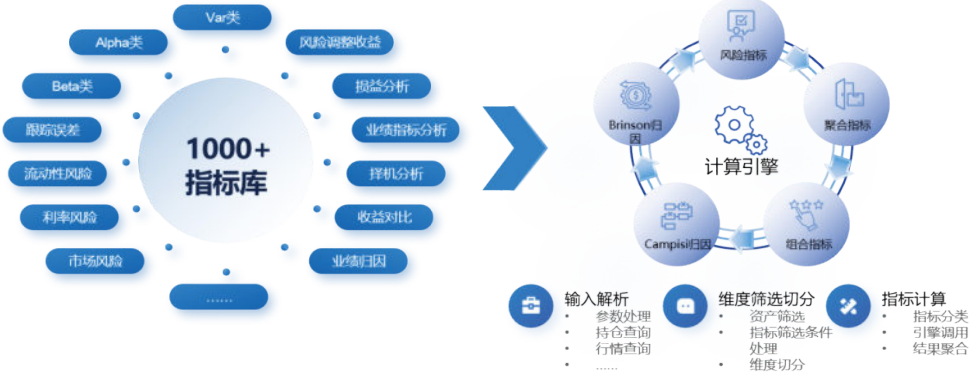


图9：金融指标计算引擎能力示意图

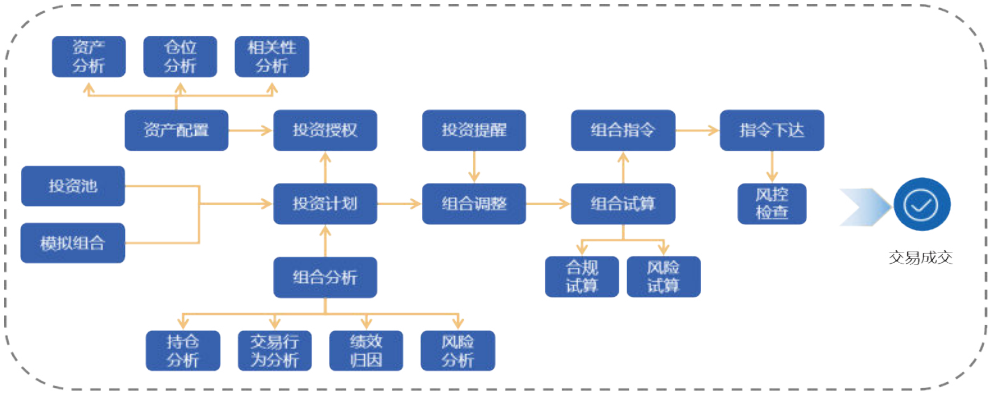


图11：投资决策功示意图

面向投资决策场景，支持投前分析、投中决策与管理和投后分析，覆盖投资管理全链路。盘前的投资提醒提供持仓相关重要事件提醒，盘中支持组合调整试算与指令下达到交易系统，盘后提供组合多维度分析，包括绩效归因、交易行为分析、持仓分析等，实现投资全流程一体化管理，如图 11 所示。

4) 自动化可视报表尽览投资全貌，大副提升工作效率

平台构建投资管理驾驶舱，提供个人投资全貌的概览，便于快速把握投资整体情况。同时以可视化报表的形式展现组合概览、资产配置、风险提示、订阅消息等，支持自定义组件与页面布局，以满足个性化需求。从组合、自定义基准等不同角度提供灵活分析视角，提供按时间区间、业务类型、账户、组合等不同维度的风险报告和绩效分析报告，通过自动化报告极大减少了业务人员工作量，有效提升工作效率。

五、总结与展望

本文对证券自营业务在技术支撑方面的痛点问题进行了分析，并详细介绍了上海证券面向自营业务的投资管理平台的建设成果与实践经验。作为公司数字化转型的重要实践，投资管理平台实现了线上化、数字化、智能化的业务管理。投资管理平台已经应用于上海证券自营业务的日常工作中，覆盖沪深 A 股、港股通、可转债、股指期货、新三板股票等不同业务品种，提供自动化报表、限额监控、组合绩效分析、投资行为分析、模拟组合分析等关键业务功能，为业务开展提供了有力的数据支持和工具支撑，有效提升业务人员工作效率，实现系统化、精细化、智能化管理。

伴随业务的发展，未来我们将不断完善和优化平台功能，持续增加对更多业务品种和业务场景的支持，不断完善对交易策略如套利、对冲、价值投资等的支持和分析。在现有平台基础上，进一步提升数据整合处理效率，增强数据分析和支持能力，强化监控预警能力，为投资服务更加完备和智能的技术平台支持。

券商行业资讯数据微服务设计的探索与实践

刘军，肖航，张赫麟，蔡世界 | 中信建投证券股份有限公司 | E-mail：liujunzgs@csc.com.cn

摘　要： 金融资讯数据在证券行业有着广泛的应用，充分发掘资讯数据价值，提供个性化服务是业内的一个重要的方向。当前在证券行业，券商通过的业务终端软件为客户提供越来越多的资讯类服务。金融资讯类服务相比纯数据类服务的应用场景更为复杂，系统和微服务的设计难度更大，导致在系统服务架构成本、以及服务性能方面难以均衡。本文通过深入研究、技术改造和应用实践探索出一套较优的技术方案，使得针对资讯类服务在经济性、高性能、安全性达到了一个合理的均衡。本文通过分享解决方案和实践经验，抛砖引玉，助力证券行业金融科技的发展。

关键词： 资讯；微服务； 高性能；HTTP

一、概述

1.1 背景

金融资讯数据在证券行业有着广泛的应用，充分发掘资讯数据价值，提供个性化服务是业内的一个重要的方向。

资讯数据类服务在“蜻蜓点金”APP 的服务矩阵中扮演重要的角色。然而，在提升用户活跃度和用户人数的同时，给自身的服务也带来了具体的挑战；当前，通过横向“暴力”扩展服务器，纵向加强服务器性能和网络带宽确实能提升券商 APP 微服务性能，解决一部分性能问题。但当前券商 APP 功能繁多，众多微服务采用独立分布式独立隔离部署，中间层通过中台系统实现前后端分离，我们且把这种架构方式定义为“分布式多级烟囱式”式架构，这种架构方式也需占用了大量资源，形成了显著的成本; 此外，服务器、机房、机架、操作系统、数据库等中间件的成本也明显增加。这些成本也成为券商“降本增效”重点考虑的对象，针对微服务本身的性能提升势在必行!

1.2 需求与难点

本文从一些典型的需求出发，分析这些需求的难点与困难，分享经验。

1）资讯类微服务典型需求

需求一：APP 前端查询当前日期的历史股票相关资讯信息；

需求二：APP 前端根据客户自选股列表查询股票相关资讯信息；

2）需求的难点分析

需求一：用户访问进去将刷新显示基于历史股票相关资讯信息的一个加工和非加工类指标数据，因此后台微服务实际是需要提供历史交易日的数据。

该需求存在几大挑战，一：股票资讯类信息数据的字段通常较长，目前本需求的每条资讯数据字段长度约为 0.55KB，以一天数据平均为 268 条计算，服务接口包为 150KB。数据载荷包超过 100KB 将严重影响微服务的并发响应性能；二：实际上该类型的每日资讯数据上限在 2 万条左右，在开发过程中，观察到每日资讯数据超过三千条，粗略估计包的大小为 1.65MB; 三：股票资讯数据源根据实时盘口信息和舆情信息产生的，数据数据在交易时段实时到达服务的数据库，因此微服务端没有预处理的冗余时间。而业务要求数据的实时性较强。

需求二：公司的拥有千万级别的客户，客户的自选股股池和数量都是不一样的，也存在实时变更的情况。这样，后端服务是无法预处理。此外，该需求同时具备需求一所描述的所有困难和挑战。

难点综述：

股票资讯数据较普通结构化、规范化的数值型数据相比，具有数据包较大，导致 HTTP 传输时延较大，难以获得较高的 TPS 指标。此外，由于资讯类数据是实时产生的，数量不定，有时候每天特别巨大；另外，客户端用户量大，自选股股池各不一样，存在随时变化的可能，要盘中快速地使得资讯数据能够精确地匹配用户的自选股池，因此对微服务的性能和前后端交互方式的设计提出了较大难度的挑战。

二、实施方案

2.1 实施的原则与约束

1）控制微服务分布式系统的服务器节点数量，充分发挥单点服务器的性能

在降本增效的大背景下，充分发挥每台服务器的资源，不能随意扩展集群的规模。由于大型金融机构的业务复杂，系统庞多，如果没有这个约束，将来会发现机房、机位需求成规模的增长，整体的运维难度、复杂度及其成本大大增加。

2）高性能

高性能包括高并发、高可用。在有限的硬件投入下，提高性能意味着节省成本。同时，性能也反映了用户体验，响应时间分别是 100 毫秒和 1 秒，给用户的感受是完全不同的。券商公司 APP 具备非常明显的并行访问的特点，特别在开市前 10 分钟，并发数量急剧攀升，因此高并发的访问特性对 TPS（每秒事务：Transaction per Second）要求较高。

3）安全性和合规性

安全性概念较为广泛，体现在系统安全、框架安全、流程安全、容灾性好，微服务本身的安全以及内容的安全。本文的安全主要关注微服务的多活以及内容的安全性。

合规性在本文中主要体现在内容和操作的合规。

2.2 资讯类微服务设计

2.2.1 总体的互联网服务的框架

互联网服务整体框架设计，如下图。

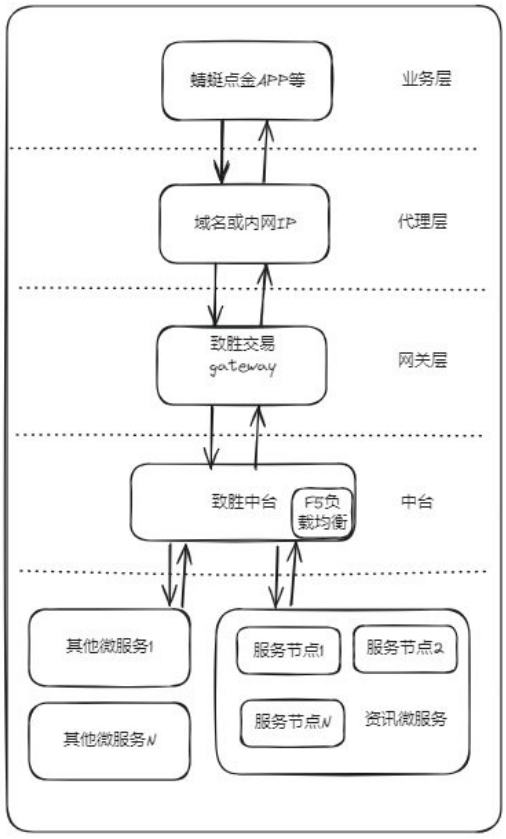


图1：总体互联网服务框架层次

大量的用户都是通过 APP 端进行业务操作并接受相关金融服务的。用户通过“蜻蜓点金”浏览相关 H5 页面，前端的 HTTP 数据请求，经过代理层和致胜交易网关到致胜中台（备注：中信建投证券统一的数据中台系统），最终转发到资讯微服务，资讯微服务根据前端的各种请求，查询相关数据库，组织数据，进行相关逻辑计算，融合数据，最终组成 HTTP 返回包，对前端请求进行应答。

2.2.2 项目实施的关键环节

1）微服务的两地三中心部署及数据同步

由于“蜻蜓点金”APP 的客户量约为千万级别，根据相关要求，需要满足两地三中心的部署要求。

此外，根据信创相关需求，微服务系统采用国产数据库 TDSQL（TDSQL for MySQL，腾讯的一款分布式数据库产品），TDSQ 内核可以做到秒级主备延迟，数据极速同步到备机。共享内存、数据恢复、快速预热，这些都是加速性的。采用主从数据实时同步，保证两地三中心的数据一致性要求及备份要求。

2）资讯服务中心的可扩展框架

如图 1 所示意，致胜中台为的资讯微服务提供统一的接口注册和管理服务，同时提供了负载均衡服务。资讯微服务采用框架部署，实现各服务统一运维管理，通过管理页面实现快速部署服务并观察相关节点资源，增加服务异常报警机制，同时支持服务横向扩展，当微服务性能不足时，可实现快速扩展节点，一键迁移服务到新节点。因此，该架构微服务器横向扩展提供了较强的灵活度，同时服务节点 1 与服务节点 N 之间保持了一定的物理和逻辑的解耦。

该设计带来的益处如下：

A）服务节点数量可以灵活扩展，根据微服务的压力情况灵活增减服务节点数量；

B）解耦的设计方案，是一种冗余设计，保证了微服务的相对安全。

C）流量的负载均衡作用。

3）多重缓存的设计使用

提升资讯类微服务的 TPS 性能有很多方法，目前本文中在服务的多重缓存方面做了较多的实践探索，取得了较好的效果。

本文采用了三级缓存机制的设计，如图 2 所示，提升了微服务性能，而且保护了底层数据库的安全。

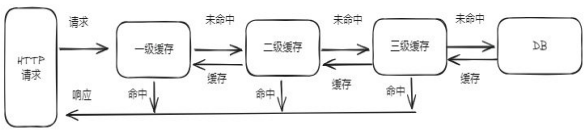


图2：3层缓存逻辑示意

A）一级缓存: 致胜中台提供缓存机制，参数可设置。根据各个 HTTP 接口微服务的实时性要求，设置不同时长的缓存参数，比如 10 秒、30 秒。

B）二级缓存：微服务本地缓存，单节点微系统服务采用 SpringBoot 的缓存框架。

C）三级缓存：内存数据库做为三级缓存。

如图 2 所示：前端 HTTPS 请求过来，如在一级缓存命中，则直接返回数据；否则依次通过二级和三级缓存进行查询，命中的话返回数据，如果所有缓存均未命中，则由微服务进行数据查询相关数据，融合后返回，并依次建立相应缓存。

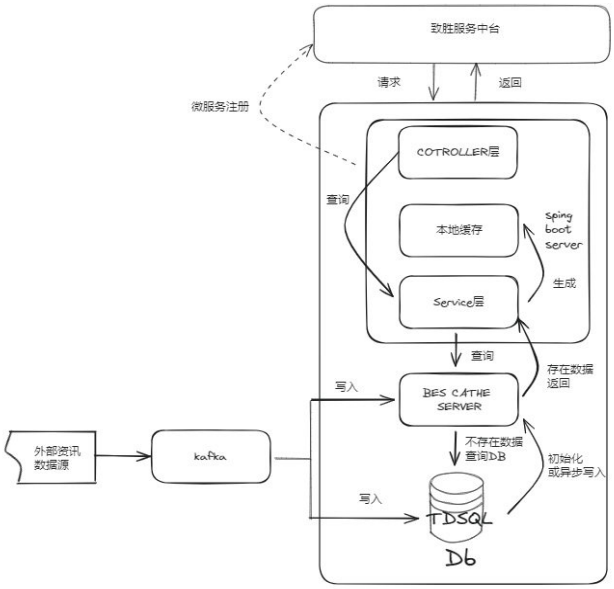


图3：资讯微服务架构及3层缓存应用示意

如图 3 所示，在微服务设计中本地缓存（二级缓存）效率最高，其次是分布式内存数据库，最差的是每次查询数据库。排比情况：本地缓存 > 分布式缓存数据库 > 直接查询数据库。

设计中需要注意几个要点：

A）分布式内存数据库（如 REDIS）一般采用集群式部署，成为一个公共组件。由于微服务通过 TCP/IP 协议的方式访问分布式内存数据库 . 因此 , 微服务是否与分布式内存数据库在同一网段尤为重要。

B）一级缓存和二级缓存的缓存时间长度的考虑。缓存的时间越长越有利于提升 TPS，但是会增加内存的消耗，同时太长的缓存时间会影响数据的更新，从而影响数据的实时性，所以要结合项目的实际情况和资讯数据的实时性要求合理设置缓存的时间长度。

C）由于分布式缓存数据库的数据具有非持久化，因此需要支持数据从持久化的数据库进行异步写入的功能，保障数据的有效性。

4）前端任务与后端任务处理之”争”

当前的互联网服务架构主流采用前后端隔离的方式，分工明确，但实际的项目开发过程中，有一些任务在前后端都能做，但是效果相差较大，需要谨慎和全面考量。

A）当前端的请求参数总体比较固定，变化较少的情况下，优先采用后端统一处理，返回数据给前端。

B）当前端请求参数多、变化较快，服务的并发量较大的情况，一般建议后端统一返回原始数据，前端根据不用的参数处理相关逻辑。因为后端面对大量的变化性的入参，本文前面的多重缓存机制的性能将大大降低，因为变化的数据会使得索引命中率快速下降。缓存的索引命中失效的情况下，微服务的计算逻辑不得不启动，会明显增加计算量，增加内存消耗，增加 I/O 访问次数，这将明显降低性能，因为服务的处理时间增加，TPS 性能明显下降；甚至导致数据库的连接数过多，访问超时，进一步可能可能带来的严重后果就是数据库宕机，并影响其他的依赖数据库的服务全部异常，这个后果很严重，应尽量避免。

5）裁剪和压缩 HTTP 传输包的大小

在处理资讯类这种比较大载荷的 HTTP 数据包，进行一些精细化的处理，有助于降低数据包的大小，从而提升微服务的性能。举例如下：

A）返回前端需要的最少资讯数据

比如: 前端某些地方显示的只是资讯索引，并非详情，此时，微服务没必要返回全部的资讯数据，而是做部分截断。

表1：截断前后的情况

截断前	查询三个交易日数据返回，包括某资讯字段，返回数据量 805 条，返回数据包大小 450k 左右
截断后	查询三个交易日数据返回，对某资讯字段进行截取处理，截取前 100 字符，数据量 805 条，截取后的数据包为 300k 左右 比截断前减少 1/3。

B）长的资讯负载，先压缩，再传输。

考虑压缩数据包所在的环节，是在 HTTPS 传输过程前在 Web server 中压缩，还是微服务接口中压缩?

在服务的请求入参单一或相对比较固定的情况下，在微服务接口中压缩比较好。前端的 HTTP 的第一次请求后将首先建立本地缓存，一旦形成的本地缓存，那么微服务将直接返回已经缓存好的 HTTP 的包，本微服务接口的执行体里面的复杂计算（包括压缩）将省略，在本地缓存一直生效的生命周期里，前端的请求过来，都不需要另外进行计算和压缩，这样就大大节省了压缩的时间，提升了微服务性能。

C）压缩算法

压缩算法比较多：有 GZip、LZ7、deflate 算法等。不同的算法在压缩比、压缩时间、CPU 占用率、压缩耗时等指标有不同的表现。项目组需根据数据的特点选取适合的算法。

本案中采用了 GZIP 算法：

表2：压缩情况

压缩效率与压缩比	压缩前后比约为：4.6 ：1
压缩时间	约为 3ns（纳秒）
CPU 占用率	占用率较低

由表 3 可见，通过截断和压缩后，数据包缩小了 86%，TPS 性能达到了 5641T/S（单节点性能）

表3：截断前后及压缩后微服务性能比较

情况	表现	TPS 性能表现（单节点）
截断前，未压缩	数据量 805 条，返回数据包大小 450k 左右	数据包太大，超时验证 错误率较高
截断后，未压缩	截取前 100 字符，数据量 805 条，截取后数据包 300k 左右，比截断前减少 1/3。	TPS=798
截断后，压缩后	数据包 65k 左右，相比未截断未压缩前缩小了 86%。	TPS=5641

5）对数据源数量的约束

由于资讯类数据是供应商提供，在一定交易日内随着市场行情的波动，基于股票的资讯数量也会波动。

当市场波动较大时，股票相关资讯数据量过大，一个交易日数量可达到三千多条，甚至更多。那么微服务在查询过程中会因数据包过大而导致性能降低，返回超时，这样后果很严重。解决方案之一是：数据供应商控制个股资讯的数量不超过某上限。

6）安全性、合规性

做资讯类的服务需要考虑内容的安全性、合规性与审计要求。

渠道合规性：资讯类数据的渠道合法合规。

内容的合规：信源方的资讯类数据必须合规，比如无敏感词，有版权，无法律纠纷。通过合规部门建立了敏感词库，最后形成敏感词接口服务，所有入库和发布的资讯类数据通过敏感词接口服务进行过滤和报警。

资讯数据的审批与留痕：资讯数据的发布需要有审批流程和留痕要求，因此对所有的资讯数据入库时候需要相应角色的人员进行审核，并在数据库留痕。保留时间根据合规要求，根据不同级别的要求保留相应长的时间。

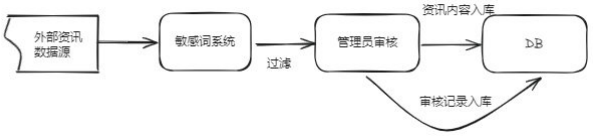


图3：资讯数据的审批与留痕

7）测试方法

测试方法要贴近生产应用实际：对于券商类的互联网接口微服务的测试，特别要贴近生产实际，比如开盘后的高并发的访问，用户有些接口方法的入参基本一致，有一些接口的请求参数千差万别。如果只是一味的固定参数对微服务进行压力测试，这将与实际不符合，测试的性能将不能用于生产运行的评估，否则会误判，导致生产事故。因此，在做压力测试的方案时候，特别注重尽量贴近生产实际情况仿真压力测试。

针对不同的微服务，主要涉及的参数有：微服务入参数量和值的变化、打压的并发数量以及打压时间。

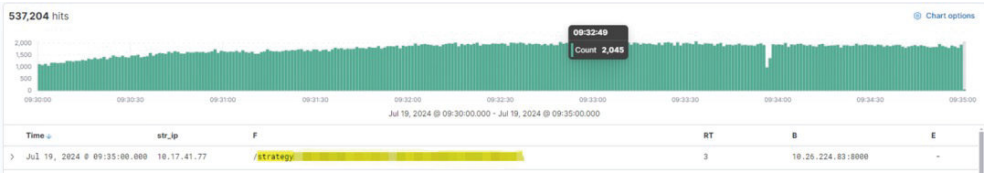


图4：微服务线上运行情况

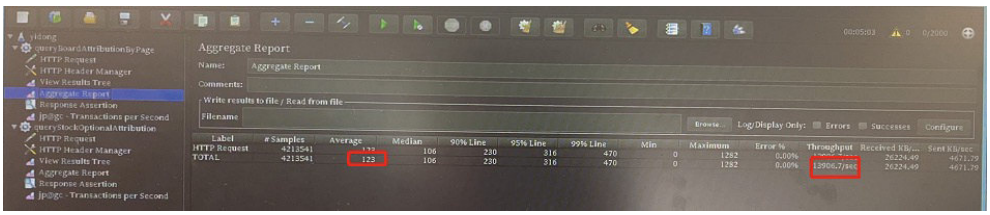


图5：某典型资讯类微服务优化后的单节点性能

2.4 效果说明

在采取整体优化方案之前，该资讯服务占用 8 个计算节点（指资讯的微服务器，不包括中台框架及网关等数据库，中间件公共服务器节点），仅能提供 800TPS 的服务性能。目前仅仅 3 个计算节点，一些微服务的单节点性能就达到了 5K TPS 和 10K TPS 以上，3 个服务节点的集群能提供大约 3 倍的性能，即能到达 15K TPS 和 30K TPS，该性能已能完美地满足峰值时期的客户访问需求。监控显示微服务线上运行十分稳定，如图 4 所示。

如图 5 所示，经过优化后，某典型资讯微服务单节点的压力测试下达到13k tps的性能，相比优化前大大提升。

三、总结与展望

上述方案解决了当前的证券业资讯服务的一些典型问题，带来了明显的安全和性能提升。

两地三中心的分布式服务的框架部署提升了服务的安全性能；采用了多级缓存技术，提升了并发处理能力，提升了数据库的安全性能，大大减少了微服务的分布式系统节点数量；采用裁剪和压缩技术等一系列精细化设计降低了 HTTP 包的大小，提升了微服务的处理能力。方案整体上在经济性、高性能、高安全达到了一个较好的均衡。为同业提供一个宝贵了经验分享！

经过通过项目的实践，探索和积累了不少经验，不过仍然有些问题有待探讨和研究。

1）比如资讯类服务由于本身的载荷数据包大的本质特征，尽管使用了压缩技术及缓存技术，其载荷数据包的大小仍然比单纯传输少量数据型的包大很多，大量的此类互联网传输将占用机房服务中心上下行的带宽。如不通过有效的方式进行限流，将会对挤占总体的通信带宽，影响别的业务正常进行。因此监控并预警资讯类服务的带宽情况尤为重要，作为控制手段的互联网访问限流也是备用手段之一。

2）流量和带宽挤占的压力使得资讯类服务与重要的交易类服务做隔离成为必要考虑。因此可以考虑资讯类服务采用公有云的方式进行部署，交易类的服务由于安全性和数据的隔离性要求，仍然采用私有化本地部署。

参考文献：

[1]2023 年 9 月证券 APP 月活跃用户规模盘点 . https://www.analysys.cn/article/detail/20021123 2023，10.
[2] 一文带你了解 SLB、F5、Nginx 负载均衡 -CSDN 博客 . https://blog.csdn.net/Jiao1225/article/details/122733116 2021，1.
[3]HTTP 协议超级详解 . https://blog.csdn.net/ros275229/article/details/132224059 2023，8.

上交所业务管理系统平台在自主可控上的探索与实践

孙长昊，周秋萍 | 上交所技术有限责任公司 | E-mail：chsun@sse.com.cn

摘　要： 近年来证券金融行业信息化建设水平不断提升，网络安全形势愈发严峻，信息系统的自主可控能力成为重要能力。上交所业务管理系统平台作为支撑业务办理、审核与流转的自研关键业务系统，为进一步落实自主可控要求，同时保障系统安全运行及研发连续性的基础上，克服多方面困难，完成了信创化改造，成为最早一批实现全栈信创改造的关键业务系统，本文以应用系统开发视角，将过程中的经验进行整理与分享，为后来系统的改造工作提供参考。

关键词： 自主可控；业务管理系统平台；信创；全栈单轨

一、引言

长期以来，证券金融行业信息系统和技术较多依靠外部供应商建设维护。近年来随着证券金融行业信息化建设水平不断提升，网络安全形势愈发严峻，信息系统的自主可控能力的重要性与日俱增。

信创，即“信息技术应用创新”，旨在实现信息技术自主可控，其涉及产业链包括 IT 基础设施、基础软件、应用软件、信息安全等方面。信创产业是数据安全、网络安全的基础。

2022 年到 2023 年期间，上交所业务管理系统平台（下文简称“业管平台”）经过 2 年努力分步实现了全栈并轨与全栈单轨改造，成为最早一批实现全栈单轨信创化的关键业务系统，本文以应用系统开发的视角，对过程中的经验进行整理与分享，为后来系统的改造工作提供参考。

二、系统介绍

经过近二十年的发展，到 2018 年时，上交所已经建成了几十个框架各异的业务系统。为进一步提升自主掌控力度，提升整体质效，于 2018 年开始规划与建设了业管平台，最初就本着自主可控原则选用了主流开源技术框架与中间件进行搭建，并由技术公司负责建设与运行。从规划定位看，业管平台定位为上交所业务系统建设的重要平台，建成后再重构整合存量多种架构的业务系统，同时为业务创新提供支撑。业管平台作为上交所业务办理、审核与流转的关键系统，主要提供了包括股票、债券、基金、衍生品、会籍、参与人等各类业务办理的功能，支撑所外机构用户的业务一网通办及相应的所内业务审核与流转。

至 2022 年初，业管平台已完成十余个存量系统的重构整合，上线十余个业务模块，承接二十多种业务，系统体量庞大，约为 20 个中等业务系统大小。系统用量方面，业管平台主要面向交易所外部机构用户，整体平均日活用户超 500，平均业务功能每日使用超 5000 次，日均请求量超 40 万次。运行与研发方面，业管平台线上运行超 100 个技术服务、组件，近年要承接几百项业务需求的开发。业管平台具有用户多外部性强、业务多影响大、功能多体量大、技术组件多、关联系统众多、需求多迭代快的特点，近两年也频繁在全面注册制等重要工作中承担了重要的开发任务。业管平台的平稳运行影响所外用户的业务办理、所内用户的业务审核以及业务流转运行，安全平稳的运行以及连续高效的需求研发成为工作的底线原则。

在保障运行与研发的同时，完成完全自主可控改造这一对系统的全面翻新，无异于要给高速运行的列车进行整体翻新，难度巨大、挑战众多。最终随着业管平台于 2023 年 12 月 23 日完成最后一个改造版本的发布上线，完成了系统整体信创单轨改造任务。改造工作基于信创私有云平台，使用麒麟操作系统、东方通中间件，海量数据库，实现了服务器、芯片、操作系统、中间件、数据库的全栈信创替换。

三、整体策略

新的信创体系相比以往开源技术体系，虽然各领域均有多种候选产品，但在 2022 年上半年信创私有云仍在选型建设中，其所提供的各类组件技术选型尚未明确，系统技术改造方案存在较大不确定性。为降低改造风险，在改造开始确立了分步改造、先试点后推广、持续改进优化的大原则，分两年实现全栈单轨的最终目标。第一

步实现系统全栈并轨，期间完成对新中间件与运行时环境的验证，实现与几十个关联系统的解耦准备（下文简称并轨改造阶段）；第二步实施工作量大时间周期长的工作，完成系统整体全栈单轨改造，下线非信创技术组件，达到最终目标（下文简称单轨改造阶段）。

随着改造的推进，挑战与难点也浮出水面：首先，信创环境与非信创环境相互独立，网络架构差异很大，网络环境以及应用系统间交互的复杂度陡增；其次，业管平台承载了众多业务功能，与交易所内外几十个技术系统通过文件、MQ 及 RESTful API 多种方式交互，系统内部的众多微服务模块之间的 API 调用更是错综复杂，系统关联关系的梳理以及解耦技术组件的搭建至关重要；此外，信创私有云于 2022 年底明确了所提供的各类技术组件，包括华为容器云、东方通或宝兰德中间件以及海量数据库等（具体产品型号见表 1）。其中数据库的替换给项目组又带来了额外难题，从原有 MySQL 数据库换成海量数据的 Vastbase 数据库产品，技术路线的改变需要对原有 SQL 语句进行全面改写与复测，工作量较大，对人力投入以及全过程的计划协调带来了非常高的要求，成为单轨改造阶段最大的难题。

确定了整体的改造路线，明确了关键技术点，项目组针对性制定了改造方案并开始了紧锣密鼓的实施，小步快跑最终达成目标（关键时间点如图 1 所示）。

表1：关键技术组件型号

产品类型	选型方案
虚拟服务器	信创私有云平台（华为 HCS）
虚拟机操作系统	EulerOS
数据库	私有云 RDS（Vastbase）
中间件	东方通 Tongweb，东方通 TongHttpServer

图1：改造关键时间节点

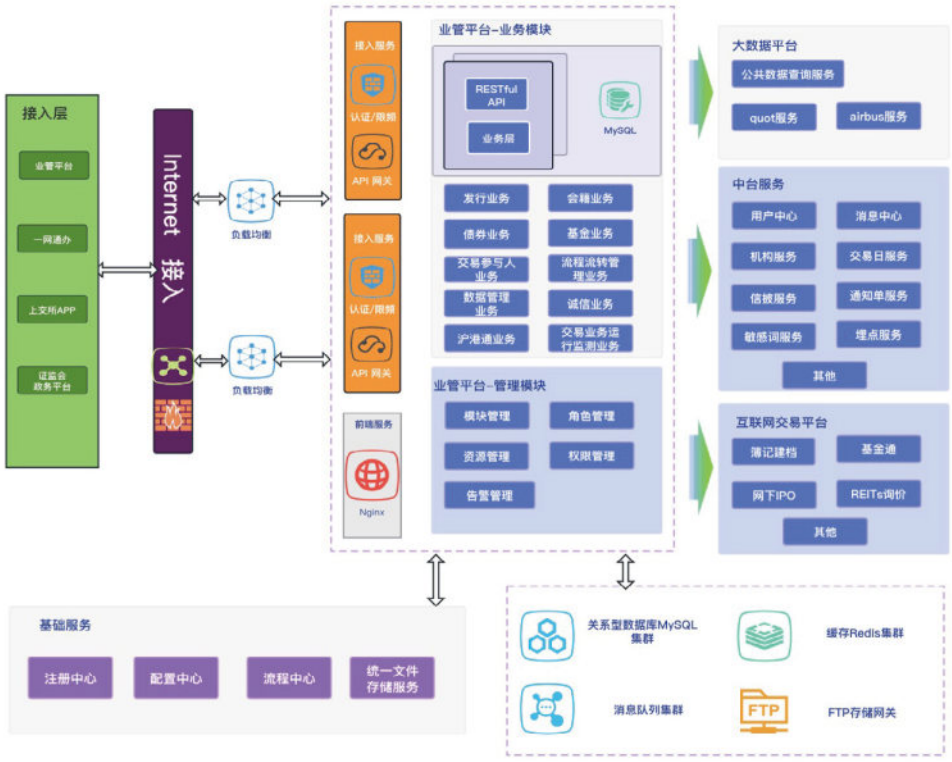


图2：业管平台系统架构

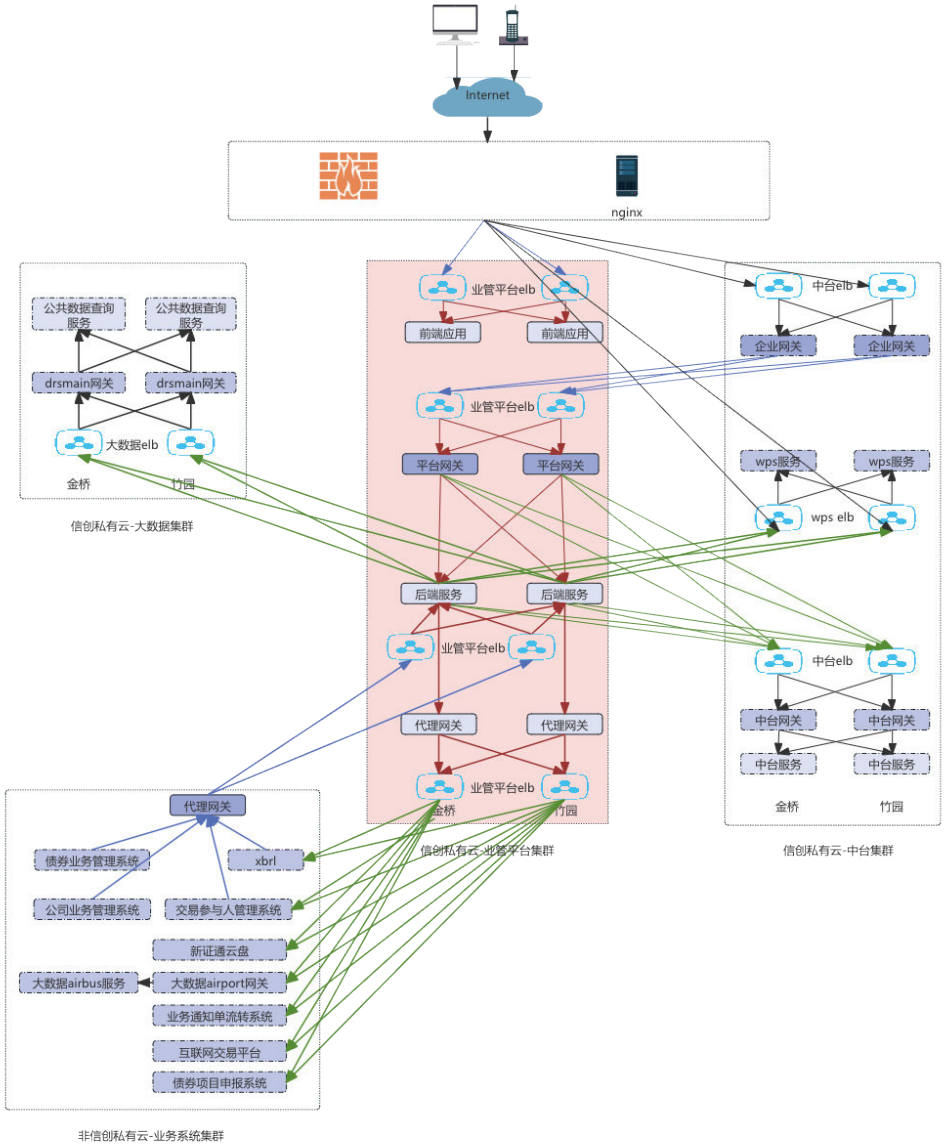


图3：目标部署架构示意图

四、改造方案

改造方案主要基于业管平台的架构情况，具体方案从应用、数据以及代码重构几方面展开：

4.1 系统架构

业管平台系统架构如图 2 所示。

业管平台主要使用 JAVA 语言开发，后端使用 SpringBoot 框架，前端使用 Vue 框架，采用 SpringCloud 框架实现微服务治理，是一个前后端分离的微服务 B/S 架构应用系统，并提供移动端功能，软件的交付与运行使用 Docker 容器，运行在上交所私有容器云环境，并使用私有云所提供的关系型数据库服务和对象存储实现结构化数据与文件的持久化。与业管平台存在数据交换以及接口调用关系的各系统，其信创改造计划也各不相同，基本都会晚于业管平台完成。系统自身以及关联系统均处于动态改造过程中，也给改造的实施带来了更多挑战。

4.2 应用迁移方案

非信创私有云上业务系统群部署在同一组集群服务器中，各系统的微服务都注册到公共的 Eureka 注册中心进行服务发现和服务调用。

信创私有云上为不同业务系统分配独立的 VPC，在提高了系统间的隔离性和安全性的同时，也要求所使用的公共的 Eureka 注册中心适配改造，改为通过私有云提供的弹性负载均衡 ELB（Elastic Load Balance）软件实现跨 VPC 访问。

改造后业管平台将整体迁至信创私有云运行（目标部署架构示意图如图 3 所示），独立部署在分配给业管平台的 VPC 下。同时在业管平台 VPC 下新搭建 Eureka 注册中心服务，为业管平台提供内部微服务的服务发现。

和外部系统服务的相互调用方面，则分为两种方式：

（1）对于已部署在信创环境的系统，业管平台直接访问其 ELB 地址；

（2）对于未部署至信创环境的系统，业管平台新增代理网关组件，供非信创私有云集群和信创业管平台的调用时使用。

4.3 数据库改造方案

数据库改造方面，需要完成 MySQL 数据库迁移到海量 Vastbase 数据库。借助数据迁移工具平台完成数据的迁移实施，方案要点包括 5 方面：

（1）Vastbase 数据库配置就绪：包括字符集、兼容模式、编码、参数的配置。

（2）Vastbase 函数适配就绪：完成对标 MySQL 的内置函数的准备。

（3）数据库表就绪：完成 Vastbase 的 DDL 建表语句开发与执行。

（4）借助数据迁移工具完成 DML 数据的迁移。

（5）索引适配改造就绪：由于 Vastbase 不支持跨表同名索引，需要对索引做重命名和迁移。

4.4 代码重构方案

所选用海量数据库 Vastbase 产品可以对 Oracle、MySQL 等 SQL 语法提供一定程度兼容性，但其本身采用 PostgreSQL 的 SQL 语法，仍需要原本基于 MySQL 开发的 SQL 语句进行适配改写。从数据类型、语法，到内置函数，与 MySQL 都有很大差异。经过深入分析比较，比对 Vastbase 和 MySQL 的具体差异，SQL 语句的重构主要涉及几类：

（1）Vastbase 不支持原 SQL 语句且可以进行转换的情况，通过相应语法转换或关键字替换进行兼容；

（2）Vastbase 不支持原 SQL 语句且没有可行的转换方式，需进行 SQL 语句重写；

（3）Vastbase 不支持的函数，由在数据库级别补充适配的函数定义；

（4）数据处理方面，由于 Vastbase 对数据大小写敏感，分析所有 SQL 的 where 语句，对于需要大小写不敏感的判断条件左右操作数加上 upper 或 lower 函数，保证逻辑不变。

五、挑战与应对

5.1 应用解耦迁移

应用迁移过程中，挑战来源于需要保证在改造全过程对系统用户的使用、业务需求研发上线以及关联系统交互三方面的透明无感：

首先，结合关联系统的数量，对系统整体进行调用关系梳理以及影响评估，结合关联系统改造计划，梳理形成调用解耦清单，新增信创与非信创代理网关，并将相关调用切换通过代理网关完成，后续仅需要对代理网关等组件的调用配置进行调整即可。

其次，结合研发任务紧迫性，将信创改造任务穿插安排其中，针对性安排上线版本计划，在不影响业务正常开展的情况下，分批次上线，在上线后持续监控系统运行状况，确保改造前后对业务无影响。

5.2 数据库改写与复测

数据库改造方面，则存在三大挑战：

一是 SQL 语句重写时，数据库类型的兼容性、查询语法的差异性、索引功能、性能优化、事务管理以及锁定机制差异性与 MySQL 的实现差异虽然不大，但在众多业务逻辑的场景下仍会出现功能与预期不符的情况。需要大量的测试验证与全面的回归测试，要相关的改造版本提前预留足够的测试时间。

二是 Vastbase 数据库函数的重写与适配时，函数数据类型转换的兼容性、方言语法与逻辑调整以及性能方面与 MySQL 存在差异，需在数据库函数的单元测试、业务逻辑的复测的基础上加强对系统性能的测试验证。

三是对于在生产运行过程中暴露出的兼容性问题，主要包括存储过程和和视图的使用、数据库连接池和连接保活以及并发性能几方面问题，需要提前进行预案准备，出现问题及时响应与分析处置。

5.3 信创网络调用复杂度陡增

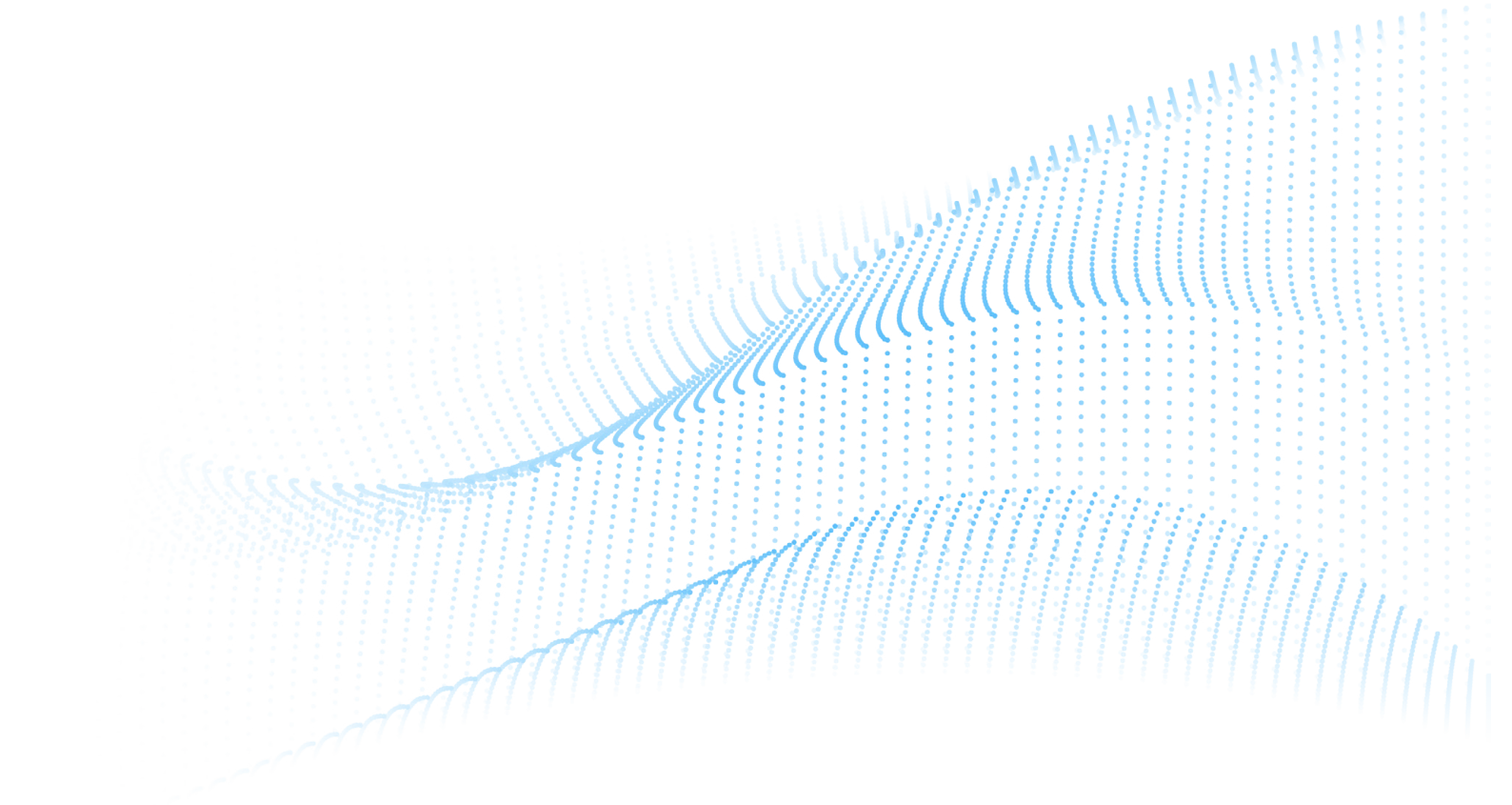
改造前，业管平台运行在非信创私有云环境，这是一个基于 Mesos 与 Marathon 的集群管理框架的容器 PaaS 平台，服务器部署于同一集群下，微服务系统使用统一的注册中心实现服务发现与调用，应用间调用无需网络策略，调用链路环节少，业务应用容器共用宿主机内存资源。改造后，信创私有云作为新一代私有云平台，使用基于 Kubernetes 和服务网格的集群管理，容器进程部署在 pod 上。为提升安全性，新增了相互隔离的 VPC 网络，对外通过 ELB 提供服务调用并实现弹性负载均衡，整体调用链路变长，网络策略复杂度也陡增。在

初期上线运行时，问题的定位与分析难度明显变高，后期在强大的运维辅助工具平台就绪后才逐步缓解。

六、总结与展望

自 2022 年启动信创改造至 2023 年 12 月 23 日的最后一个版本的上线，业管平台在完成信创改造的同时完成了 1700 项常规需求的开发，上线 140 多个版本，保障了研发任务的连续实施；通过多个接口技术组件以及多达几十次配置操作，最终完成了信创单轨改造目标，达成了全过程对用户及关联系统无感的目标，成为最早完成信创改造的重要业务系统，大大提升了自主可控水平。

展望未来，新的信创私有云也带来了新的服务网格能力，提供了更多可能性，业管平台未来将在新的应用进一步对接服务网格及新开发框架，逐步完成应用框架的迭代更新，进一步提升云原生能力。



04 信息资讯采撷

监管科技全球追踪

8 月，证券业协会发布《证券业区块链电子数据存证应用规范》。该规范内容涵盖技术要求、存证平台、应用示例等，旨在为行业提供统一、明确且具有可操作性的标准与指导，以规范区块链技术在证券业电子数据存证领域的应用。

8 月 1 日，欧盟《人工智能法案》正式生效。该《法案》旨在确保在欧盟开发和使用的人工智能是可信赖的，并有保障措施保护人们的基本权利。

8 月，英国央行（BoE）表示，计划进行一系列关于分布式账本技术（DLT）和批发中央银行数字货币（wCBDC）的实验，以跟上支付领域的变化，并评估金融科技发展中的机遇和风险。

8 月 9 日，泰国证券交易委员会（SEC）推出了数字资产监管沙盒，旨在促进新数字资产服务的实验和开发。SEC 指出，沙盒参与者必须将他们的创新融入泰国资本市场数字资产服务的发展中，或者必须参与货币市场监管机构的沙盒。

8 月 13 日，科技部印发《“创新积分制”工作指引（全国试行版）》，将企业创新能力转化为投资机构看得懂的“财务数据”，适用范围扩展至全国，助力具有核心竞争力的“硬科技”“好苗子”企业脱颖而出。

8 月 30 日，国务院常务会议审议通过《网络数据安全管理条例》。条例指出，要对网络数据实行分类分级保护，明确各类主体责任，落实网络数据安全保障措施。要厘清安全边界，保障数据依法有序自由流动，为促进数字经济高质量发展、推动科技创新和产业创新营造良好环境。

9 月 5 日，上海市副市长解冬在“2024 年外滩大会”开幕式上表示，近期将出台推进金融科技中心建设的行动方案，进一步发挥上海在资源集聚、应用场景、营商环境等方面的优势，朝着成为全球有引领性金融科技中心的目标加快迈进。

9 月 9 日，全国网络安全标准化技术委员会发布《人工智能安全治理框架》1.0 版，推动各方就人工智能安全治理达成共识，促进人工智能安全有序发展。

9 月，美国证券交易委员会（SEC）、财政部等九大联邦金融监管机构根据《金融数据透明度法案》提议建立新的联合数据标准，以增强监管机构间金融数据的一致性、可访问性和互操作性。

9 月 13 日，证监会发布《证券发行人信息披露文件编码规则》金融行业标准。证监会表示，将继续推进资本市场信息化数字化建设，着力做好基础标准制定工作，促进通用基础领域标准研制，不断夯实科技监管基础。

9 月 25 日，国家发改委、国家数据局等六部门联合印发《国家数据标准体系建设指南》。《指南》计划到 2026 年底，制修订 30 项以上数据领域基础通用国家标准，初步建立国家数据标准体系。此举为我国国家数据标准体系建设提供了指引。

9 月 29 日，美国加州州长加文 · 纽森以“不应只考虑模型成本和算力”等理由，否决了此前饱受争议的《前沿人工智能模型安全创新法案》（SB 1047）。该法案曾于 8 月 28 日经州议会投票通过，其要求 AI 开发人员对其技术可能造成的任何严重损害负责，任何模型必须具备“可立即全面关闭能力”。

10 月 3 日，马来西亚证券委员会（SC）将引入监管沙箱，并加强其监管框架，以鼓励证券代币化，以帮助刺激资本市场的创新。SC 主席拿督穆罕默德 · 法伊兹 · 阿兹米表示，引入的框架将为测试创新产品和服务提供一个受控的环境，同时确保投资者得到保护。

10 月，2024 年度诺贝尔物理学奖获奖名单公布，计算机人工智能领域科学家“爆冷”获奖。获奖者约翰 · 霍普菲尔德和杰弗里 · 辛顿分别以发明了具有“联想记忆”功能的神经网络和一种可以自主查找数据属性的方法而著称，诺贝尔奖表彰他们“利用人工神经网络进行机器学习的基础发现和发明”。

10 月 16 日，央行、科技部联合印发《关于做好重点地区科技金融服务的通知》，指导和推动北京、长三角、粤港澳大湾区等科技要素密集地区做好科技金融服务。其中提到，建立科技金融数据共享平台，加强信息技术运用，提升科技公共信息共享和使用水平。

10 月 21 日，中国证监会副主席李超在 2024 金融科技大会上表示，证监会高度重视金融科技安全问题。李超指出，要坚持总体安全观，既要继续强化行之有效的传统安全保护措施，又要结合新情况研究新措施、新方法，综合运用法律、技术、管理等手段，防范化解因新技术深入应用而带来的相关技术和业务风险，推动金融科技在资本市场安全、健康、有序发展。

10 月 22 日，华为技术有限公司在深圳正式发布原生鸿蒙系统 HarmonyOS NEXT。这是我国首个实现全栈自研的操作系统，标志着中国在操作系统领域取得突破性进展。

10 月 25 日，韩国财政部表示，韩国计划从 2025 年下半年开始监管加密货币等虚拟资产的跨境交易，引入注册和报告要求。韩国财政部表示，根据新规定，处理虚拟资产跨境交易的企业将被要求事先向当局登记，并每月向韩国银行报告交易情况。

11 月 4 日，香港证监会行政总裁梁凤仪在合规科技论坛上表示，过去三年，在主要打击洗钱程序中，香港的合规科技应用率均有所提升。在 50 家参与调查的企业当中，逾八成确认合规科技有助它们加强打击洗钱的能力。

《交易技术前沿》征稿启事

《交易技术前沿》由上海证券交易所主管、主办，主要面向全国证券、期货等相关金融行业的信息技术管理、开发、运维以及科研人员。近期重点征稿主题如下:

一、云计算

（一）云计算架构

主要包含但不限于：云架构剖析探索，云平台建设经验分享，云计算性能优化研究。

（二）云计算应用

主要包含但不限于：云行业格局与市场发展趋势分析，国内外云应用热点探析，金融行业云应用场景与实践案例。

（三）云计算安全

主要包含但不限于：云系统下的用户隐私、数据安全探索，云安全防护规划、云安全实践，云标准的建设、思考与研究。

二、人工智能及大模型技术

（一）应用技术研究

主要包含但不限于：大语言模型/AIGC的数据处理和治理、可解释的人工智能及大语言模型、用于大语言模型/AIGC的神经网络架构、训练和推理算法、多模态AI等。

（二）应用场景研究

主要包含但不限于：基于人工智能或大语言模型的智能客服、语音图像文本等数据挖掘、柜员业务辅助等。

主要包含但不限于：金融预测、反欺诈、授信、辅助决策、金融产品定价、智能投资顾问等。

主要包含但不限于：金融知识库、风险控制等。

主要包含但不限于：机房巡检机器人、金融网点服务机器人等。

三、数据中心

（一）数据中心的迁移

主要包含但不限于：展示数据中心的接入模式和网络规划方案；评估数据中心技术合规性认证的必要性；分析数据中心迁移过程中的影响和业务连续性；探讨数据中心迁移的实施策略和步骤。

（二）数据中心的运营

主要包含但不限于：注重服务，实行垂直拓展模式；注重客户流量，实行水平整合模式；探寻数据中心运营过程中降低成本和提高服务质量的途径。

四、分布式账本技术（DLT）

（一）主流分布式账本技术的对比

主要包含但不限于：技术架构、数据架构、应用架构和业务架构等。

（二）技术实现方式

主要包含但不限于：云计算+分布式账本技术、大数据+分布式账本技术、人工智能+分布式账本

技术、物联网+分布式账本技术等。

（三）应用场景和案例

主要包含但不限于：结算区块链、信用证区块链、票据区块链等。

（四）安全要求和性能提升

主要探索国密码算法在分布式账本中的应用，以及定制化的硬件对分布式账本技术性能提升的作用等。

五、信息安全与IT治理

（一）网络安全

主要包括但不限于：网络边界安全的防护、APT攻击的检测防护、云安全生态的构建、云平台的架构及网络安全管理等。

（二）移动安全

主要包括但不限于：移动安全管理、移动互联网接入的安全风险、防护措施等。

（三）数据安全

主要包括但不限于：数据的分类分级建议、敏感数据的管控、数据共享的风险把控、数据访问授权的思考等。

（四）IT治理与风险管理

主要包括但不限于：安全技术联动机制、自主的风险管理体系、贯穿开发全生命周期的安全管控、安全审计的流程优化等。

六、交易与结算相关

（一）交易和结算机制

主要包含但不限于：交易公平机制、交易撮合机制、量化交易、高频交易、高效结算、国外典型交易机制等。

（二）交易和结算系统

主要包含但不限于：撮合交易算法、内存撮合、双活系统、内存状态机、系统架构、基于新技术的结算系统等。

投稿说明：

1、本刊采用电子投稿方式，投稿采用Word文件格式（格式详见附件），请通过投稿信箱ftt.editor@sse.com.cn进行投稿，收到稿件后我们将邮箱回复确认函。

2、稿件字数以4000-6000字左右为宜，务求论点明确、数据可靠、图表标注清晰。

3、不设固定截稿日期，常年对外收稿。收齐一定数量的稿件后将尽快组织专家评审。

4、投稿联系方式021-68602496, 021-68607129欢迎金融行业的监管人员、科研人员及技术工作者投稿。稿件一经录用发表，将酌致稿酬。

附件：投稿格式（可通过电子邮件索要电子模版）

标题（黑体 二号 加粗）

作者信息（姓名、工作单位、邮箱）（仿宋GB2312 小四）

摘要：（仿宋GB2312 小三 加粗）

关键字：（仿宋GB2312 小三 加粗）

一、概述（仿宋GB2312 小三 加粗）

二、一级标题（仿宋GB2312 小三 加粗）

（一）二级标题（仿宋GB2312 四号 加粗）

1、三级标题（仿宋GB2312 小四 加粗）

（1）四级标题 （仿宋GB2312 小四）

正文内容（仿宋GB2312 小四）

图：（标注图X. 仿宋GB2312 小四）

正文内容（仿宋GB2312 小四）

表：（标注表X. 仿宋GB2312 小四）

正文内容（仿宋GB2312 小四）

三、结论/总结（仿宋GB2312 小三 加粗）

四、参考文献（仿宋GB2312 小四）

电子平台

欢迎访问我们的电子平台 http://www.sse.com.cn/services/tradingtech/transaction/。我们的电子平台不仅同步更新当期的文章，同时还提供往期所有历史发表文章的浏览与查阅，欢迎关注！

联系电话：021-68602496
021-68607129
投稿邮箱：ftt.editor@sse.com.cn

ITRDC证券信息技术研究发展中心（上海）

上海证券交易所 SHANGHAI STOCK EXCHANGE

中国上海市杨高南路388号
邮编：200127
公众咨询服务热线：4008888400
网址：http://www.sse.com.cn

内部资料 免费交流

本资料仅为内部交流使用，本期印200册，编印单位为上海证券交易所，面向证券期货行业发送，印刷时间为2024年11月，印刷单位为上海华顿书刊印刷有限公司。