主题：人工智能行业应用

交易技术前沿

2024 年 第 2 期 总第 56 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 2 期 总第 56 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

近年来，习近平总书记多次强调，面对新一轮科技革命和产业变革，必须抢抓机遇，加大创新力度。2024年《政府工作报告》首次将“大力开展‘人工智能+行动’”写入其中。面对人工智能新科技浪潮，行业不断深入探索人工智能应用，在智能投顾、智慧经纪、数字员工等方面不断培育新动能，促进行业新发展。本期《交易技术前沿》以“人工智能行业应用”为主题，精选行业关于人工智能技术探索与实践应用方面的优秀文章，为行业落地人工智能，推动数字化转型提供参考。

国金证券的《证券行业大语言模型优化方法与应用示范》针对证券公司大规模使用大语言模型面临的数据治理、数据安全和技术集成等问题，提出了一种结合检索式问答生成模型、提示工程、以及Agent技术的综合技术路径。

上交所的《基于人工智能的动态中点订单研究与建议》介绍了纳斯达克交易所基于AI的动态中点订单，对动态中点订单机制原理和演进历程进行综述，并提出加强相关证券市场微观机制研究的建议。

中信建投证券的《知识图谱：驱动员工能力的引擎与机遇》聚焦证券公司财富管理机构的协同工作等问题，建立员工赋能平台项目，将知识图谱与大语言模型等相关技术结合，弥补了金融领域数据缺少组织结构、价值密度低、难使用的不足。

招商证券的《人工智能大模型在证券行业应用路径与实践》探讨了证券业如何借助大模型技术进行“数智化”转型。文章分析了大模型的应用场景、实践案例和技术路径，并对国内外实践案例进行梳理，展望了行业未来的挑战与趋势。

申万宏源证券的《基于深度强化学习的客户资产均衡建模》将人工智能技术与资产均衡理论相结合，采用深度强化学习对客户资产均衡性进行建模，从客户兴趣、资产多样性、等多个维度打通数据底座到智能化应用通路，为客户提供个性化投资建议。

证券信息技术研究发展中心（上海）
2024年6月7日

目 录

证券行业大语言模型优化方法与应用示范

王洪涛 | 国金证券股份有限公司 | Email：wanghongtao@gjzq.com.cn

摘　要： 大模型在证券行业的核心作用是充分的萃取数据中的信息和知识，提升证券公司的含智量，培养新质生产力。然而，鉴于证券行业的业务独特性以及大模型自身的局限性，将这类模型在证券业中广泛应用面临不少挑战。为了克服这些挑战并有效利用大模型的潜力，本文提出了一种结合检索式问答生成模型（RAG）、提示工程、以及Agent技术的综合技术路径和应用模式。这种综合方案旨在帮助证券公司提高业务效率、更好地控制风险，并优化客户体验。国金证券作为该领域的先行者，采用创新的应用模式不仅为证券行业内大模型的广泛应用提供了实践案例，也展现了结合行业特定知识和先进技术的重要性，为证券行业在大数据时代的转型和升级提供了有力的借鉴和启示。

关键词： 新质生产力；大语言模型；搜索引擎；RAG；Agent

一、引言

ChatGPT的出现打破了现有的人机交互模式，其展现出的强大的语义理解和生成能力引发了人们对其背后的支撑技术的广泛关注。然而，大模型（Large Language Model，LLM）在证券行业的应用尚处于起步阶段，对于如何充分发挥大模型的潜力以及所面临的挑战，业界尚未形成共识。由于金融市场的复杂性和动态性，大模型需要实时更新和学习新的金融知识。其次，大语言模型的性能受到训练数据的限制，如何提高证券场景下大模型生成内容的质量仍有待探索。证券公司大规模使用大语言模型具有以下挑战：

数据治理问题　在现阶段，许多证券公司的数据治理体系尚未完全建立或优化。这意味着数据可能存在分散、不一致或质量不高的问题。由于大型语言模型高度依赖于数据质量和结构，这些问题可能导致模型性能不佳或产生误导性的输出。

数据安全性　鉴于证券公司处理的是高度敏感和机密的财务数据，数据安全成为一个重大关注点。大型语言模型的应用可能涉及将数据传输至外部服务器进行处理，这增加了数据泄露或被恶意利用的风险。

技术集成和兼容性问题　将语言模型集成到证券公司现有的IT架构和工作流程中可能遇到技术挑战。这些挑战包括系统兼容性问题、需要升级或更换现有系统的成本和复杂性，以及确保新技术不会干扰现有工作流程的稳定性和效率。

针对证券公司的业务特点，以及现有金融科技发展的实际情况，我们提出了证券公司优化大语言模型的三种方法：**采用提示词工程优化证券业务流程、通过搜索引擎与大模型结合加工实时财经资讯信息，以及通过Agent的模式外挂证券业务算法。**

我们认为上述方法比采用大量数据训练和微调通用大语言模型更适合证券公司的实际情况。本方法具有以下好处：

更高的效率与准确性　通过精准的提示词工程和特定算法，能够更有效地理解和满足客户特定需求。这种方法可以更直接地针对证券业务的特点，提供更准确的服务，尤其是在处理复杂的金融信息和交易时。

实时信息获取　结合搜索引擎和大模型，使得证券公司能够实时获取和分析市场动态和财经新闻。这种方式比传统的大数据训练模型更灵活，能够快速适应市场变化，为投资决策提供即时支持。

定制化服务与创新　通过Agent模式外挂专门的证券业务算法，可以根据公司和客户的具体需求定制服务。这种方法允许证券公司创新其服务和产品，为客户提供更个性化、高度适应性的解决方案。

成本效益与风险控制　相比于传统的大规模数据训练，这种方法可能更节省资源和时间，因为它专注于特定的业务需求和场景。同时，通过更精确的算法和实时信息，公司可以更有效地管理风险，避免依赖过时或不精确的数据。

总得来说，本文阐述的大模型优化方法使证券公司能够更有效地应对快速变化的市场环境，提供高质量的客户服务，同时控制成本和风险。同时我们也看到大模型的探索与发展又是一个开放的、不断优化前进的过程，随着证券公司数据治理的推进，数据安全的发展，以及交易系统技术兼容性的不断进步，大模型技术会随着证券公司底层技术的进步而不断的向前发展。

二、大模型在证券行业应用面临的问题

当前，证券公司内部有广泛的知识检索需求，是大模型落地的极佳场景。然而，作为一种新兴技术，大模型自身仍有一定的局限性，包括事实错误（幻觉）、缺乏领域知识、信息过时等问题[1]。因此，如何建设具备高专业度、强时效性的证券大模型亟需探索。

2.1 通用大模型的问题

通用大模型基于海量高质量的语料进行预训练，将所学习到的知识存储到模型参数中，展现出优异的内容生成能力，已在多个领域得到广泛应用。但是，通用大模型并不完美，仍存在诸多不足之处：

（1）知识记忆能力有限。大语言模型的“伸缩法则”（Scaling Law）表明，随着参数规模、数据集大小、训练计算量的不断增加，模型的性能将持续提升。尽管如此，大模型无法记住训练语料中的所有知识，尤其是出现频率较低的长尾知识。证券行业的数据安全要求较高，还包含大量的长尾知识，而不同类型的客户有差异化的需求，如何利用大模型提供多样化的服务至关重要。

（2）知识时效性不足。通用大模型难以与外部世界互动，由于知识的快速迭代，模型知识的时效性较差。如果使用微调的方法频繁更新模型参数，其算力消耗仍然不容忽视且容易出现灾难性遗忘问题，对于大部分证券公司而言难以负担。

2.2 挂载知识库的大模型的问题

基于知识库的大模型能够与外部进行有效交互，获取与用户提问有关的额外信息[2]。构建知识库时，首次提交的文档通过非结构化加载器读取文本，根据预定义的规则或语义信息进行文本切分，然后使用Embedding模型将文本块向量化存储到向量数据库中。当用户提交问题，通过向量相似度匹配召回与用户问题最相似的前k个文本作为提示，大模型根据问题和提示做出响应生成回答，如图1所示。

外部知识库能够进一步扩展通用大模型所拥有的知识数量，通过本地化部署证券公司的数据安全性得以保障，员工通过大模型可以针对内部规章制度、非公开研究报告等信息进行提问。然而，多样的非结构化数据（文档、图片、图形表格等）给知识库的构建带来极大困难，并且知识库的时效性依然难以保证。多个存在重复内容的文档构建的知识库，可能产生对大模型产生反作用效果，这是因为特定领域知识被稀释以及文档间相互有影响。

大型语言模型挂载文档库通常是历史数据，这可能导致模型无法反映最新的市场信息和动态，对于快速变化的证券市场来说是一个重大弱点。在证券行业，理解市场趋势和预测未来走势至关重要。大型语言模型可能无法完全捕捉到市场的微妙变化和潜在的投资机会。模型的性能在很大程度上取决于其训练数据的质量和范围。如果文档库中的数据不全面或存在偏差，模型的输出可能会受到影响。

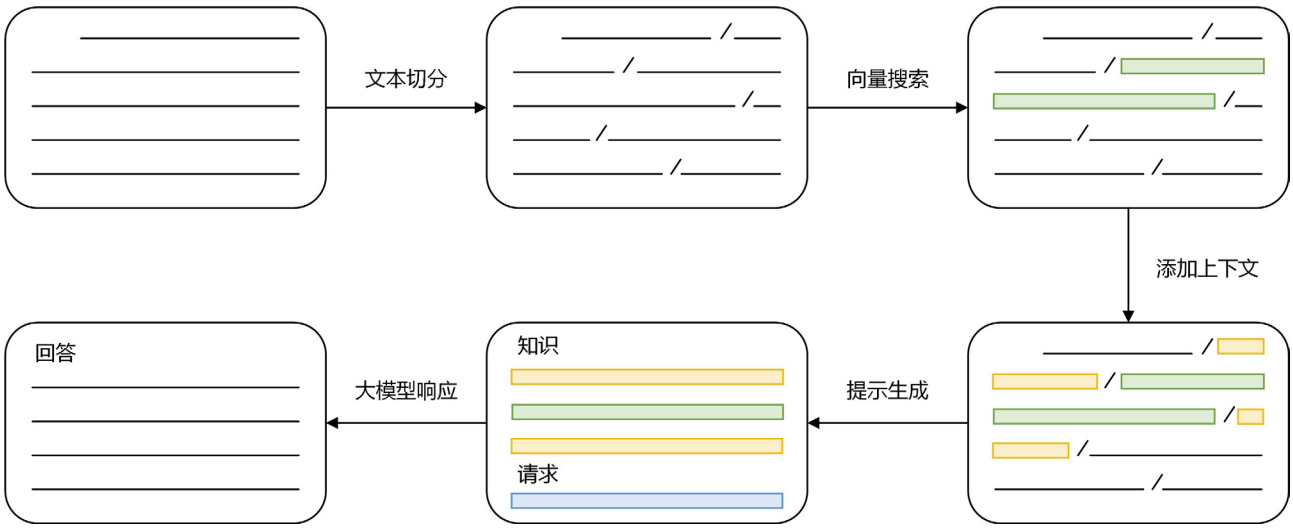


图1：基于外部知识库的大模型的响应流程

三、证券行业大模型性能提升的方法

3.1 优化的方向

在对大型语言模型进行性能优化的过程中，OpenAI采用了一种综合性的优化流程。如图2所示，该流程横跨了上下文优化（Context Optimization）与LLM优化两个关键维度。上下文优化关注于模型需要了解的信息，即为了成功执行任务，模型需要了解的背景知识。而LLM优化则着重于模型的行为方式，即模型采取的方法和行动来解决特定问题。

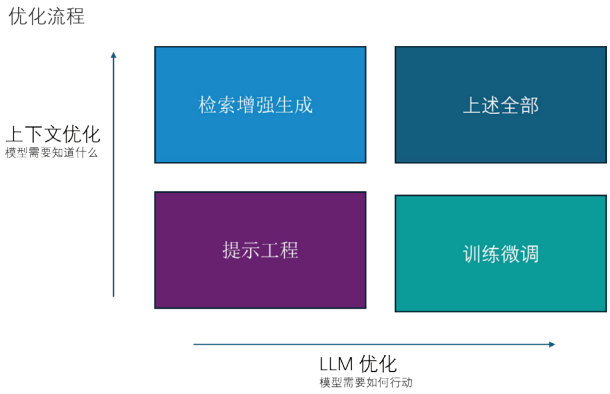


图2：OpenAI的大模型优化流程

在证券行业，可以获取与问题相关的上下文，并进一步通过提示工程、检索增强生成、智能体技术引导大模型的推理方向，以显著提升回答的准确性和即时性。下面分别对这三种技术进行概述。

提示工程（Prompt Engineering）是开始优化的最佳起点，旨在设计和优化指示大模型在进行特定任务时应该采取什么行动或生成什么输出的提示。针对证券公司的业务，可以采用提示工程多次调用大模型的API，并结合RPA等工具多次问答自动生成需要业务的报告，例如：日报、研报摘要等场景。

检索增强生成（Retrieval-Augmented Generation，RAG）适合引入新的信息，以及通过控制内容来减少幻觉。搜索引擎结合大语言模型可以在保证信息时效性的同时，从海量的财经类新闻中抽取需要的信息更加快速和高效。

智能体（Agent）可以视作一种能够自主理解、规划和执行复杂任务的系统。通过利用Agent可以将不同业务算法外挂、内嵌、整合到大模型中。

上述三种优化方法不是互斥的，可以联合使用，多次迭代直至最优。表1总结了大模型优化方法及其适用证券业务场景。

3.2 优化的技术方案

优化方法	业务场景
提示词工程优化	各类问答助手
	业务营运助手
搜索引擎增加实时信息获取能力	业务报告自动生成
	产业链图谱智能生成
	智能研报助手
智能体链接业务算法	量化投资助手
	智能投研助手

表1：优化方法和业务场景总结

3.2.1 提示工程优化业务服务能力

提示工程的优化始于编写清晰的指令，以便于模型可以理解和执行任务。同时，需要将复杂任务分解为更简单的子任务，从而使模型可以对每个子任务做出正确的响应。在这一过程中，给予大模型时间去思考是另一项重要策略，这意味着让模型在生成回答之前有充分的内部处理时间，模型更有可能成功执行任务。此外，设定合理的评估体系是关键环节，系统地测试每次调整对于性能的实际影响，保证提示工程的优化朝着指定方向前进。

我们对针对金融证券领域的特性，重构了金融提示的设计架构，整体架构如图3所示。

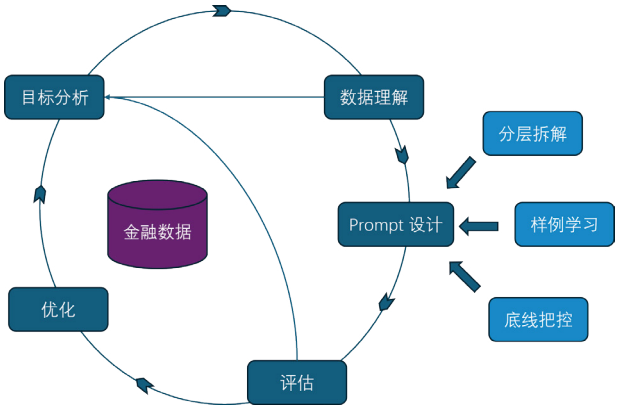


图3：金融提示工程整体架构

在进行金融大模型应用的过程中，首先需要进行目标分析，以明确任务目标和评估相关形势、资源、风险和局限性。紧接着是数据理解阶段，涉及对数据的存储形式、量级、内容进行全面分析，并对初步解决方案进行微调。基于这两个阶段的成果，接下来是提示设计阶段，旨在针对特定任务场景创建有效的提示。评估阶段主要通过指标测试（如精确率、召回率等）来评估提示的性能，并分析模型输出是否满足目标要求，同时识别存在的问题。最后是优化阶段，根据评估结果对提示进行调整，以确保在正式部署前达到最佳状态。

3.2.2 搜索引擎增加实时信息获取能力

当模型需要引入大模型未知的特定信息以回答问题时，无需进行大模型微调，而是通过搜索引擎、向量数据库等外部工具来扩展模型的知识，以推理产生准确的回答，这种方法称为检索增强生成[3]。RAG的工作流程如图4所示。RAG最直接的优势就是能够让大模型利用自身的逻辑推导能力，去理解公司的私有数据，实现问答能力的拓展。尽管模型微调也可以实现类似的效果，但RAG的技术路线更适用于大部分证券公司，这是由于考虑到其特殊的场景需求，即外部的公开数据及其内部的私有数据以一定的频率动态更新，GPU算力尚不充足，且通常要求大模型的回答能够给出引用来源以保证可靠性。

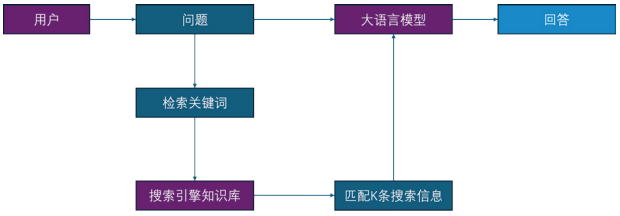


图4：RAG工作流程

在金融领域，RAG模块可用于增强大型语言模型进行金融情绪分析的能力。金融情绪分析是提取、量化和研究金融文本、新闻文章和社交媒体内容中的情感状态和主观信息的重要工具，它可能有助于分析证券市场走势，并为投资者的行为提供有价值的见解。

3.2.3 智能体链接业务算法

大语言模型的浪潮推动了AI Agent 相关研究快速发展，AI Agent 是当前通往通用人工智能的主要探索路线。大模型庞大的训练数据集中包含了大量人类行为数据，为模拟类人的交互打下了坚实基础；另一方面，随着模型规模不断增大，大模型涌现出了上下文学习能力、推理能力、思维链等类似人类思考方式的多种能力。

一个基于大模型的AI Agent系统可以拆分为大模型、规划、记忆与工具使用四个组件部分。AI Agent 可能会成为新时代的开端，其基础架构可以简单划分为 Agent = LLM + 规划技能 + 记忆 + 工具使用，其中大模型扮演了Agent的“大脑”，在这个系统中提供推理、规划等能力。图5展示了基于大模型的AI Agent系统的总体概念框架，由大脑、感知、行动三个关键部分组成。

基于大模型的Agent可以理解人类的自然语言指令并执行日常任务。在面向任务的部署中，Agent遵从用户的高级指令，承担目标分解、子目标规划、环境交互探索等任务，直至实现最终目标。为了探索Agent是否能够执行基本任务，部分学者将它们部署到基于文本的游戏场景中。在这类场景中，Agent完全使用自然语言与世界互动。通过阅读周围环境的文字描述，并利用记忆、规划和试错等技能，它们可以预测下一步行动。然而，由于基础语言模型的局限性，Agent在实际执行过程中往往依赖于强化学习。随着大模型的逐步发展，具备更强文本理解和生成能力的Agent在通过自然语言执行任务方面展现出巨大潜力。

四、国金证券金融大语言模型实践案例

4.1 国金FinGPT设计思路

图6展示了国金FinGPT的设计思路，以大模型规模化应用为目标，面向业务人员、科技研发人员、AI算法人员

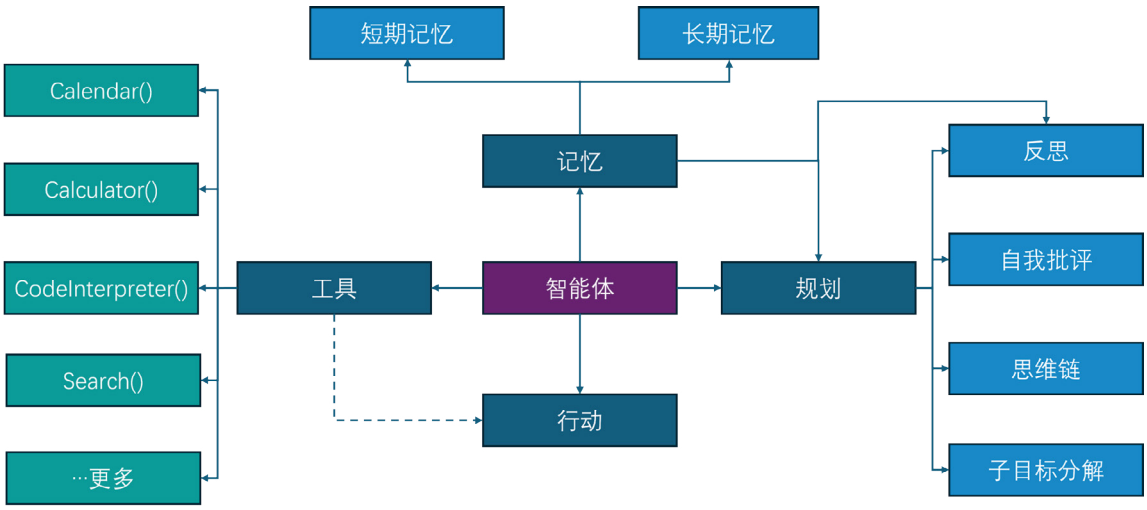


图5：基于大模型的AI Agent系统总体框架

等不同角色，构建流程化大模型研发流水线，建立RAG的大模型及解决方案，打造基于大模型的提示词中心，共同形成大模型共享给共建的应用市场生态，快速赋能数字国金建设。

4.2 国金AI员工助手：基于提示工程构建不同办公场景的应用市场

国金证券科技团队基于大模型技术搭建AI员工助手于2023年11月份全面上线，供公司所有员工使用，极大地提升员工工作效率。AI员工助手集成了多种大模型，支持同一个问题同时问多个大模型，从中择优选择答案。如图7所示，通过提示词工程，构建了不同办公场景的应用助手，包括：技术类、角色类、翻译类、文本类、文案类等，也支持用户根据需求进行个性化配置。

图8展示了AI员工助手2024年的使用次数统计，当前工作日的调用平均超过2000次。AI回答的问题以证券业务为主，通用问答，日常问答，科技类问题为辅。这将有助于培养公司内部的数字化思维和创新氛围。

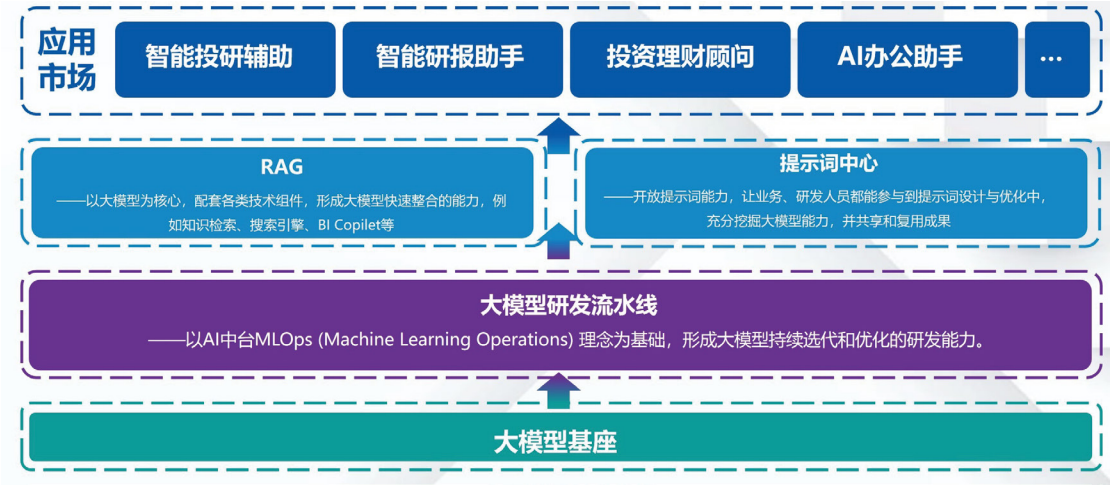


图6：国金FinGPT设计思路

图7：基于提示工程构建不同场景的AI员工助手应用市场

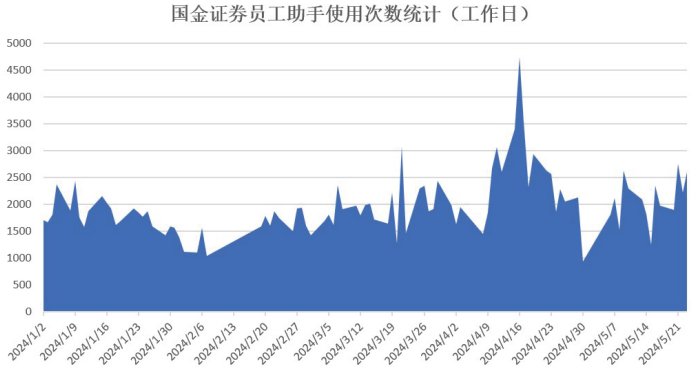


图8：国金证券员工助手使用次数统计

图9所示，通过对员工问题进行词云分析，可以看出在办公场景中大部分员工关注证券市场行情、合规风险、客户满意度等方面的信息，国金的员工已经在使用大模型提升工作中的信息和知识的萃取效率。

图9：员工助手问题词云图

图10统计了最近一周员工助手问题不同类别，大部分员工更关注证券业务，其次分别是通用知识、日常闲聊、科学技术。

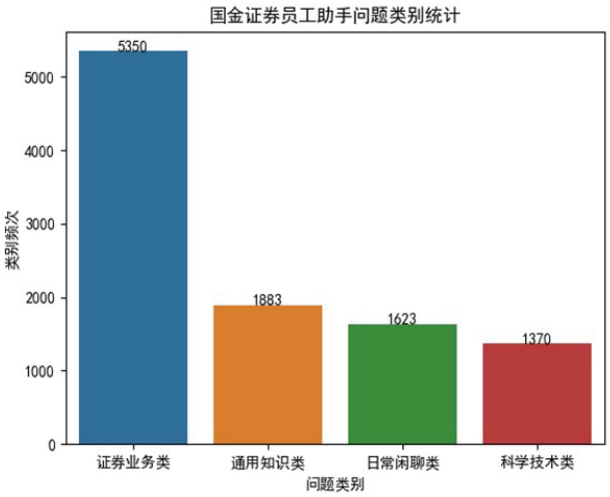


图10：员工助手问题类别统计

4.3 基于提示工程的文档撰写助手

利用具体的业务流程，优化提示词的文档撰写助手提供可灵活配置报告模板，根据不同的需求自动化生成定制报告。目前支持研报摘要、高客建议书、理财月报、员工日报和周报等自动撰写，显著提升员工工作效率和团队管理效能。

系统根据定制化模板自动生成标准报告，比如研报摘要、高客建议书、理财月报、员工日报、周报等，目前文档撰写助手已生成2.5万+份客户服务报告，预计节省4000万页报告的人工撰写成本。而且系统1分钟内就可生成人工需要2天的报告，显著提升员工办公效率。

4.4 基于提示工程和搜索引擎的产业链图谱智能挖掘

大模型产业链挖掘是国金证券首先在业内提出的大模型特色应用场景。国金证券研究所和科技团队充分合作，进行了大模型自动生成产业链图谱的初步尝试，可以根

图11：文档撰写助手：提高文字加工和萃取的效率

据最新舆情挖掘投资标的、产业链上下游、关联度等信息，从而快速认知市场。

大语言模型与搜索引擎相结合，通过分析、整合、萃取、推理新闻舆情中的标的与产业链的频率、频次、正负面及关联关系，非常适合用于智能化挖掘新型的产业链的上下游，并分析标的与产业链的关联度的标准化度量。通过构造以大语言模型为核心的智能体和产业分析提示工程，可以自动化完成产业链梳理和标的关联度分析。

针对较常见的产业链，还可以分析产业链的动态变化，从而分析板块的轮动、舆情对产业链的扩散影响。此外，对比较新的产业链，大模型掌握的相关知识较少，可以基于检索增强生成为产业链智能体配置搜索引擎。检索增强生成包含检索与生成两个步骤，1、寻找与该产业链最相关的已有产业链的信息，2、将新型产业链与已存在的产业链进行整合，基于最新的舆情信息推理分析出最新的产业链，从而推理出新型产业链的上下游以及标的关联度。

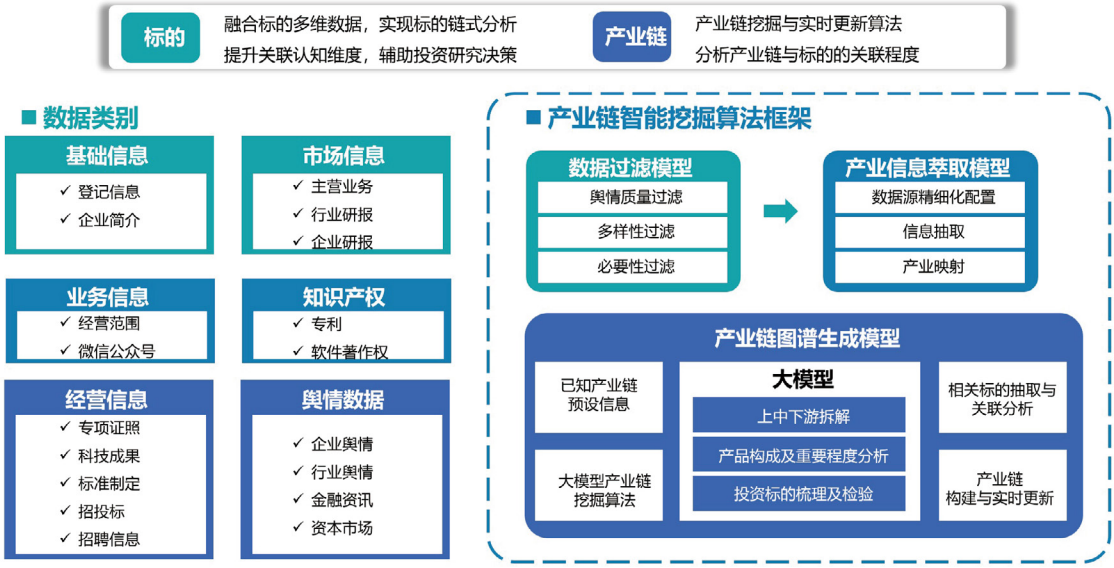


图12：大模型产业链图谱智能挖掘技术框架

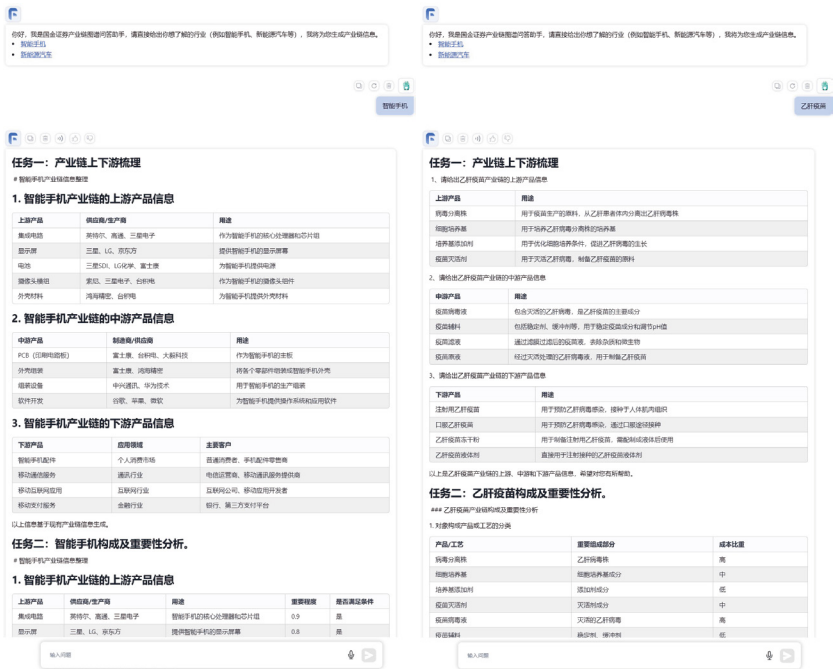


图13：国金证券大模型产业链图谱挖掘对话界面

4.5 基于Agent的大模型量化投资助手

国金证券在大模型量化领域进行了分析和探索，通过利用大模型的数学推理和代码生成能力，能够实现因子代码的自动生成，并且策略代码生成的运行通过率较高，这在一定程度上可以降低量化客户的编写量化策略的门槛。

此外，通过使用大语言模型进行舆情情感分析，可以更全面地理解市场情绪和投资标的表现，从而在量化投资中能够做出更明智的决策。例如：近期国金证券利用大模型的舆情分析能力，对负面股票进行风险提醒，目前已经在影子账户荐股中得到了应用。大模型在量化投资中的应用为投资者提供了一种新的工具和视角，可帮助其更加科学地进行投资决策。

图14：量化投资助手架构

计算方式：

1. 均线：n 日内的收盘价逐日相加后平均

2. 均量线：n 日内的成交量（值）相加后平均

买入策略： 当 5 日均线上穿 10 日均线；5 日均量线上穿 10 日均量线；MACD 指标中的 DIF 红色线上穿 DEA 黄色线。同时发生时进行买入交易

卖出策略： 当 5 日均线向下穿过 10 日均线；5 日均量线向下穿过 10 日均量线；MACD 指标中的 DIF 红色线向下穿过 DEA 黄色线。同时发生时进行卖出交易

```
class Strategy(bt.Strategy):
    params = (
        ('maperiod1', 5),   # 均线周期 1
        ('maperiod2', 10),   # 均线周期 2
        ('volperiod1', 5),   # 均量线周期 1
        ('volperiod2', 10),   # 均量线周期 2
    )

    def log(self, txt, dt=None):
        dt = dt or self.datas[0].datetime.date(0)
        print('%s, %s' % (dt.isoformat(), txt))

    def __init__(self):
        self.dataclose = self.datas[0].close
        self.volume = self.datas[0].volume

        self.sma1 = bt.indicators.SimpleMovingAverage(self.datas[0], period=self.params.maperiod1)
        self.sma2 = bt.indicators.SimpleMovingAverage(self.datas[0], period=self.params.maperiod2)
        self.sma_vol1 = bt.indicators.SimpleMovingAverage(self.volume, period=self.params.volperiod1)
        self.sma_vol2 = bt.indicators.SimpleMovingAverage(self.volume, period=self.params.volperiod2)

        self.macd = bt.indicators.MACDHisto(self.datas[0])

    def next(self):
        if self.position:
            if self.dataclose[0] < self.sma1[0] and self.dataclose[0] < self.sma2[0] and
self.sma_vol1[0] < self.sma_vol2[0] and self.macd.lines.histo[0] < 0:
                self.log('SELL CREATE, %.2f' % self.dataclose[0])
                self.sell()
        else:
            if self.dataclose[0] > self.sma1[0] and self.dataclose[0] > self.sma2[0] and
self.sma_vol1[0] > self.sma_vol2[0] and self.macd.lines.histo[0] > 0:
                self.log('BUY CREATE, %.2f' % self.dataclose[0])
                self.buy()
```

图15：国金证券量化投资助手代码生成示例

五、总结

AI是将数据变为信息和知识的关键途径，是数字化转型的最终形态。本文探讨了一种结合检索式问答生成模型（RAG）、提示工程和Agent技术的先进应用模式。这种综合性的技术路径旨在提高证券公司的业务效率，同时优化客户体验并更加精准地控制风险。国金证券作为该技术应用的先行者，不仅为证券行业内大模型的实际运用提供了一个范例，还突显了将行业特定知识与尖端技术相结合的重要性。这一实践案例为金融领域在大数据时代的转型和升级提供了宝贵的借鉴和启发，展示了金融科技在现代证券行业中的核心作用和广阔前景。

大型语言模型的发展之旅是持续不断、充满创新的过程。随着证券公司在数据治理方面的不断进步、数据安全技术的日益成熟，以及交易系统的技术兼容性持续提升，这些底层技术的发展势必推动大模型技术向前迈进，不断实现新的突破和优化。

参考文献：

[1] 桑基韬,于剑.从ChatGPT看AI未来趋势和挑战[J].计算机研究与发展,2023,60(06):1191-1201.

[2] 车万翔,窦志成,冯岩松等.大模型时代的自然语言处理:挑战、机遇与发展[J].中国科学:信息科学,2023,53(09):1645-1687.

[3] 张维佳.大模型VS搜索引擎：短期内共存[N].中国电子报,2023-12-05(005).

基于人工智能的动态中点订单研究与建议

丁逸俊、张伟、徐广斌、陆伟 | 上海证券交易所 | Email：gbxu@sse.com.cn

摘　要： 在资本市场领域，AI逐渐成为业务创新和科技监管的重要驱动力.2023年9月，纳斯达克交易所宣布获批推出全球首个用于交易所的AI订单类型--动态中点延时订单，利用AI来实时调整订单的等待期，使得可以降低交易的等待时间并改善结果。根据测试，新订单类型使得订单成交率提高了20%，并有效降低订单交易冲击成本。本文系统回顾动态中点延时订单机制的历史演进，分析其机制原理和对AI模型的应用，并提出在证券市场微观机制研究中加强对中点订单机制和人工智能应用的探索。

关键词： 人工智能；中点订单；波动保护；冲击成本

伴随新一轮科技革命蓬勃发展，前沿技术和产业深度交叉融合已成为推动数字化转型的重要动力。其中，人工智能（AI）技术近年来不断加速迭代和创新，在诸多应用场景取得了重要突破和进展，包括我们熟知的智能驾驶、ChatGpt、Alpha Go、声像识别、自然语言处理等等。在资本市场领域，AI也逐渐成为业务创新和科技监管的重要驱动力。

2023年9月，纳斯达克交易所宣布，美证监会(SEC)已批准其推出全球首个用于交易所的AI订单类型。这种订单类型被称为动态中点延时订单（Dynamic Midpoint Extended Life Order，或DM-ELO）。纳斯达克传统的中点延时订单（M-ELO）使用固定的等待期（Holding Period）来匹配买家和卖家，作为改进，DM-ELO订单利用AI来实时调整订单的等待期，使得可以降低交易的等待时间并改善结果。根据测试，新订单类型使得订单成交率提高了20%，价格滑点减少了11%。Nasdaq还表示，未来将会基于AI增强机制引入更多的动态市价订单类型。

一、背景介绍

1.1 美国证券市场结构与订单类型

根据SEC的分类标准，美国全国证券市场共含三类交易平台：全国性证券交易所（纽交所、纳斯达克等）、另类交易系统（ATS，包括ECN和暗池）以及经纪商内部撮合平台。根据《国家市场系统管理规则》（Regulation NMS），所有交易所采用统一的跨市场交易互联，按照“最优执行”原则执行订单，并在交易时向其客户提供全国最佳买入价和卖出价（NBBO）报价。其中，NBBO是指在一个给定的时间段内每个证券在所有交易平台上的最佳买价和最佳卖价，该价格由中央证券信息处理中心（SIP）计算并对外发布。

美国证券交易所接受的最基本的订单类型为市价委托和限价委托，在此基础上各大交易所提供了种类多样的补充选项。比如，指定该委托所适用的时间，即当日有效、立即成交否则撤销、集合竞价等；指定委托单的价格或数量是否可见，即可见订单、保留订单（又称冰山订单）、隐藏订单等；指定订单是否只能在某一交易所成交或是可以传递（route）至其他市场成交，即路由订单、非路由订单等。根据不同选项组合，美股市场形成了多达几十种的丰富的委托单类型。

1.2 M-ELO订单的提出

2018年，纳斯达克创造性地推出了中点延时订单（M-ELO），以吸引具有长期投资意愿的机构交易者。M-ELO具有以下四点特征：1）是一种隐藏订单；2）挂单价格为NBBO报价的中点，动态变化；3）仅能与反方向的M-ELO订单成交，其成交信息给予披露；4）订单激活前有一段500毫秒（ms）的等待期。随后，经过两年的实际执行效果以及客户反馈，为了进一步提高该订单的效率、降低对市场的影响，纳斯达克于2020年把等待期的时间缩短为10ms。

最新数据表明，2023年10月，纳斯达克平均每天成交2590万股M-ELO订单，环比增加3%。M-ELO订单执行后一秒，NBBO在84%的时间内保持稳定。执行超过1000股的情况下，在92%的时间内保持稳定。以上报价稳定性[1]的数据均好于其他场内或场外成交的中点订单。

1.3 动态M-ELO订单的提出

在将等待期时间从500ms缩短为10ms后，纳斯达克依然收到部分投资者反馈，认为10ms也可能超出了实现M-ELO订单“提高订单效率、降低对市场影响”本意所需的时长。为此，纳斯达克开始尝试进一步优化等待期的持续时间。

* 1.报价稳定性衡量的是NBBO中点在执行前后1秒内没有变化的交易占比。

经过研究，纳斯达克发现较短的等待期可以为参与者在交易成本上实现相同的甚至更好的结果。但在价格波动加剧的时期，M-ELO面临更高的风险，较长的等待期有利于保护M-ELO免受此类风险。最终，纳斯达克认为，再次全面减少静态（即固定）10ms的等待期并不是最佳选择。

为此，纳斯达克要求其人工智能和机器学习实验室（AI Core Development Group）探索是否可以利用AI技术来优化M-ELO订单在不同价格波动状态下的等待期时长，根据结果实时动态改变M-ELO的等待期，DM-ELO由此被提出。

二、机制原理

长期投资者普遍采用拆单方式来为降低其交易成本，减少对市场影响。为便利此类投资者的交易，NASDAQ等交易所向投资者提供了中点订单、冰山订单等订单服务。此类订单不仅可以降低投资者交易成本，而且可以避免投资者暴露交易意图。然而，普通的中点价订单等仍然难以避免长期投资者面临的市场波动和逆向选择风险。对此，NASDAQ交易所推出M-ELO订单，在中点订单的基础上通过增加等待机制来寻找合适的对手方，并降低投资者面临的风险。

2.1 M-ELO订单工作机制

根据NASDAQ的研究，使用M-ELO订单可以较好的帮助长期投资者匹配到相似的对手方并降低交易风险。其具体的工作方式是：M-ELO订单包含一个计时器，该计时器的触发条件是NBBO的中间价不高于M-ELO买单价格，或者不低于M-ELO卖单价格。在计时器触发后，AI计时器会自动给出一个等待期，例如10ms。经过10ms等待时间后，该订单将进入独立的交易系统进行匹配撮合。由于M-ELO订单仅与M-ELO订单进行匹配，所以M-ELO订单不会与扫单[2]成交，而会选择在系统中等待对手方M-ELO订单，当满足条件的对手方M-ELO订单也在等待一段时间后进入系统后，买卖双方的M-ELO订单将进行匹配，成交价格为当时NBBO的中点价。通过延迟交易，一定程度上可以较好的区分出真正的长期投资者，使得买卖双方寻找到合适的对手方进行交易。

2.2 DM-ELO订单工作机制

从实际运行效果看，等待期时长对订单的执行率（Fill rate）和滑点（Markout）有较大影响。此前，NASDAQ曾设置过两个固定长度等待时间，分别为500ms和10ms。但在实践中发现，大部分通过M-ELO订单匹配对手方的时间要小于10ms，过长的等待时间反而会导致投资者因担心市场波动风险而撤单，进而降低了成交率。但在少数市场极端波动的情形下，更长的等待时间又可保护投资者避免波动风险。为此，NASDAQ引入了等待时间的动态调节机制，由AI根据市场情形来给出当前最佳的等待时间，从而使得买卖双方在更稳定的市场条件下达成交易。举例来说，在静态等待时间设定下，有一个买方M-ELO订单的等待时长为10ms，另一个卖方M-ELO

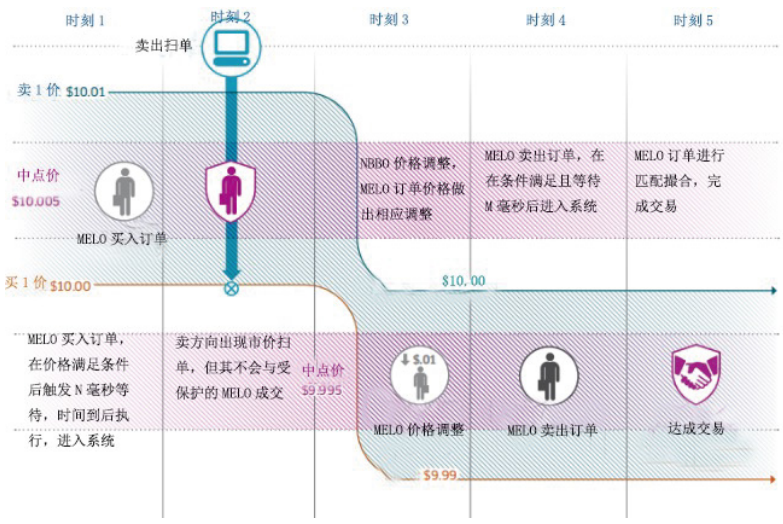


图1： M-ELO订单交易工作机制

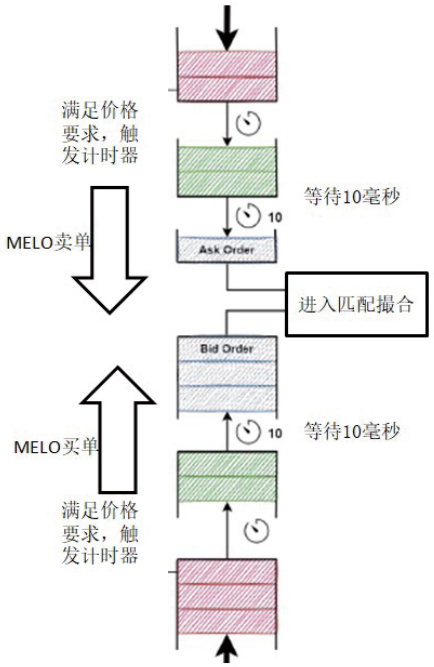


图2： M-ELO订单执行过程

* 2. 其执行改变对手方最优报价的订单。

订单的等待时长也为10ms，结果由于等待时长过长，导致市场进入不稳定阶段，导致投资者成交结果与预期存在较大偏差。同样，在动态等待时间设定下，AI认为当前市场较为稳定，并给出2ms的等待时长，则买卖双方成交在市场稳定阶段，其成交结果与预期接近。

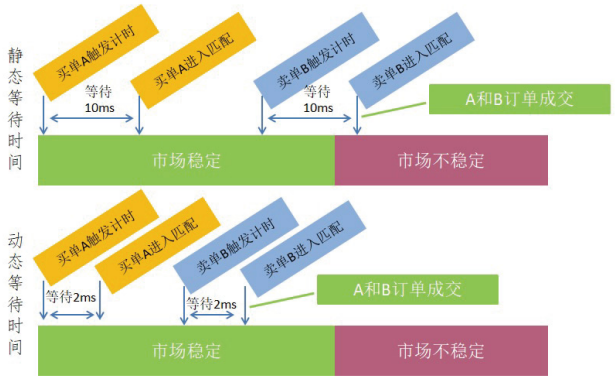


图3： 采用静态和动态等待期的风险比较

2.3 AI动态计时器

NASDAQ使用强化学习方法来训练AI计时器，其学习的目标为提高订单的成交率并降低成交价格的滑点，同时采用深度Q网络来表示动作策略。在竞价阶段，AI机器人会根据过去30秒内的行情（NASDAQ共选取142个特征指标来评估市场情形，包括M-ELO历史订单、NBBO历史行情、过去N秒买卖量等），对等待期在0.25ms至2.5ms范围内进行0.25ms为步长的调节（每30秒调节一次，全天共780次）。并且，会在市场发生极端波动时会启动保护机制，进一步延长等待时间，例如延长至12ms。NASDAQ在模拟环境中对AI计时器进行训练，并使用真实市场数据进行回测，在实践中NASDAQ将定期对深度Q网络进行重新训练以适应不断变化的市场。

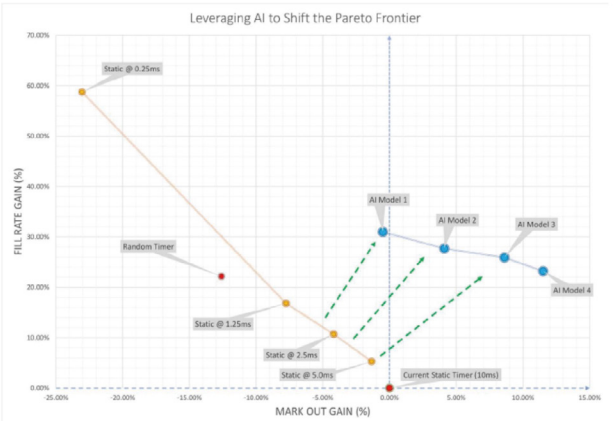


图4： 使用AI进行优化使M-ELO的执行质量得到整体提升

从测试结果来看，DM-ELO与使用固定10毫秒等待期的M-ELO相比，获得31.7%的每股加权平均综合效益提升，其中包括20.3%的订单成交率提升，以及11.4%的价格滑点下降，整体上也要优于使用其他固定等待时长以及随机等待时长的M-ELO订单。如图4所示，使用AI对M-ELO订单的等待时长进行动态优化，使得在指定时长范围内M-ELO订单的综合执行效益均得到明显提升，在整体上实现了更优的多目标帕累托边界，显示出AI应用在求解最优化问题方面的巨大优势。

三、模型实现

根据纳斯达克发布的白皮书，其采用了拥有超过35000个参数的DDQN（Double Deep Q-Network）人工智能模型，根据142个特征指标，每30秒动态计算并更新一次等待时长。该等待时长的变化范围为0.25-2.5毫秒，每次变化的幅度为[-0.5,-0.25,0,+0.25,+0.5]毫秒。

3.1 DDQN基本原理

DDQN算法脱胎于经典强化学习算法Q-learning。强化学习是人工智能技术的一个重要分支，广泛应用于需要根据环境变化做出不同决策的动态策略问题，例如围棋机器人AlphaGo、自动驾驶等领域。其基本原理是训练一个模型（也即agent），该模型通过环境状态（state）做出决策（action）并得到一个奖励或惩罚反馈（reward），通过不断地交互（训练），该模型就学习到了在各类环境下如何做出恰当的决策以得到最大化的奖励。

在Q-learning算法中，智能体建立一个表格（Q table），通过不断地尝试，将不同环境状态下，做出不同动作得到的Q值（可视为某种累计奖励）记录在表格中以完成训练过程。之后就可以通过查表的方式，知道不同环境状态下应该采取何种动作，从而得到最大奖励。

Q-learning算法的核心在于建立相应的表格，但在围棋、股市等复杂场景下，由于参数量太大，构建此类表格异常困难。一种改进方式是利用深度神经网络（Deep Neural Network）来替代该表格，由此产生了DQN（Deep Q-network）算法。深度神经网络是一种具有大量参数和强大函数拟合能力的模型，亦是当前大部分人工智能算法的基础。DDQN算法则是在DQN算法基础上，通过两个结构相同而参数更新频率不同的Q-network，改进了DQN算法的高估问题。

3.2 设计与实现

对于纳斯达克的DM-ELO机制，模型（agent）根据当前市场状态判断如何更新动态等待时长（action），随

后在仿真环境中模拟市场交易行为可能产生的变化，并向模型反馈奖励（reward）和下一个市场状态（next state）。

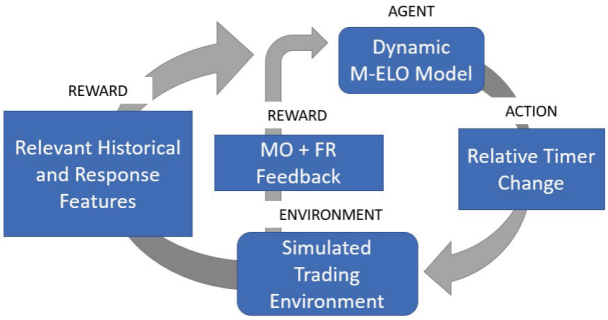


图5： 训练DM-ELO等待时长的强化学习框架

此算法的一个关键在于设置合理的奖励函数，纳斯达克采用的奖励函数计算方式如下：

$$\frac{\lambda}{\sum q_i}\sum_{i=1}^{T} q_i\left(\mathrm{MO}_{\delta,\mathrm{agent},bps}-\mathrm{MO}_{\delta,\mathrm{synthetic@10},bps}\right)+(1-\lambda)\left(\mathrm{FR}_{\mathrm{agent}}-\mathrm{FR}_{\mathrm{sim@10}}\right)$$

该奖励函数可以认为是动态等待时长相对于固定10毫秒在滑点和成交率指标上的边际优势。其中，T是单位时间内的总交易数量；q_i是该时间内第i个交易的股票数量；$\mathrm{MO}_{\delta,\mathrm{agent,bps}}$和$\mathrm{MO}_{\delta,\mathrm{synthetic@10,bps}}$分别是动态等待时长和固定10毫秒情况下的滑点值；$\mathrm{FR}_{\mathrm{agent}}$和$\mathrm{FR}_{\mathrm{sim@10}}$是对应的成交率；λ是加权因子，用于控制成交率和滑点在奖励函数中的权重。

此算法的另一个关键在于设计DDQN采用的深度神经网络。根据纳斯达克披露的信息，其深度神经网络拥有超过35000个参数，但没有进一步披露其具体结构。而该模型的输入为142个特征指标，包括历史统计信息，如一天和五天的M-ELO订单、成交统计信息，历史NBBO价格等，以及即时的成交率、滑点等。

纳斯达克利用2022年一季度的历史数据来训练DDQN模型，并从该时段内6257个交易标的（symbol）中选取了较为活跃的380个（约占使用M-ELO订单标的的67%）进行模拟。同时在训练模型时，每个交易日被简化为780个片段，每个片段代表30秒周期。以30秒为一个周期是通过大量实验，综合模型训练可行性和潜在回报后确定的。

另外，为了使DDQN模型更好适应最新的市场环境，纳斯达克设计了一种基于迁移学习的再训练机制。该机制可以简单理解为每周完整地训练并调整模型的35000个参数，之后冻结其中部分参数，用每日数据训练更新剩余的参数。通过这种方式，DDQN模型可以更好地学习到市场的长期规律和短期趋势，同时得到针对某些特殊市场日的特定模型。

3.3 稳定性与安全保护机制

为了应对特殊市场波动，纳斯达克还设计了一个独立于DDQN模型的保护机制。当市场不稳定（unstable）[3]，即NBBO价格异常波动并触发该保护机制时，M-ELO等待时长将被固定为12毫秒，并在之后的750毫秒时间内维持在12毫秒，不受DDQN模型输出结果的影响。当750毫秒周期结束时，等待时长将被调整为当前的DDQN模型输出结果。

NASDAQ的AI模型和交易系统相对独立，对交易系统唯一的影响就是从AI模型处接收一个动态时长，且其范围有限，对系统的改造较小。即使AI部分宕机，也就是动态等待时长退化为固定时长，市场影响可控。

Algorithm 1 D3QN: Dynamic M-ELO-DDQN

```
for date d in {date, end date} do
    -------- Begin Agent Experience Collection --------
    for symbol s in {symbols} do
        for iteration i in 780 × (30 sec periods / 1 day) do
            Gather and store (state, action, reward, next state) experiences
        end for
    end for
    -------- End Agent Experience Collection --------
    ---------- Begin Agent Training Loop ----------
    for m in number of symbols do
        Sample a batch of (state, action, reward, next state) experiences
        for experience e in {batch} do
            Generate target Q-values
        end for
        Calculate loss L
        Update main model parameters via gradient descent
    end for
    Update target weights to be evolving weighted linear combination of
main model and target model weights
    ---------- End Agent Training Loop ----------
end for
```

图6： DM-ELO使用的DDQN改进算法

四、研究建议

4.1 研究中点订单机制，便利中长期专业机构交易需求

目前在盘中阶段，交易者需要通过券商下单并传输至交易所，然后等待市场行情到达指定价格才能执行。如果行情波动较大或市场状态发生变化，可能造成交易者以不利的价格成交，这对希望在盘中进行大量头寸调整的基金、保险等专业机构者来说具有不利，由此带来的价格冲击也不利于市场的平稳运行。中点订单允许订单根据市场情况实时调整报价，减少订单的价格冲击和执行成本，即使市场行情突然变化，配合限价和保护期等机制可帮助降低订单执行的价格风险，更好满足服务专业机构、大宗交易者在连续竞价阶段的交易需求。

4.2 研究盘中波动保护机制

* 3. 根据Nasdaq相关资料披露，市场发生1%几率的极值波动被视为异常波动。

程序化交易等新型交易方式在提升交易效率、增强市场流动性等方面具有积极作用。但在特定市场环境下，譬如市场发生大涨大跌时，由于可能采取趋同的止盈或止损策略，而导致市场波动的加剧，如何及时、有效地实施干预，是程序化交易趋利避害的关键。对无价格涨跌幅限制的股票，盘中设置临停机制，而对于有涨跌幅限制的股票，未设置其他盘中波动保护机制。参考M-ELO相关做法，可以将盘中发生近期（前一日或一周）波动指标极值的情况，可考虑在后续一段时间内对程序化交易采取限速限流等各种限制措施，缓解特定市场环境下程序化交易对市场的冲击。

4.3 探索把人工智能应用于动态复杂多因子数字监管场景

传统模式下的一线监管和风险监测，大多采用“基于规则的信息系统+专家人工处理”的模式，主要面向静态的有限规则、固定阈值、小数据甚至是人工处理方式，随着市场规模扩大、产品丰富、交易复杂、风险交叉以及数据激增，新的现象、问题和风险不断涌现，对复杂多因子数据进行快速、智能分析方面的需求大增。例如，对于交易监管，可在现有手段上增加对标的的人工智能实时监测机制，对目标标的构建人工智能模型，将近期订单簿、成交、监测指标、异常交易或异常波动发生的特征数据作为经验输入模型训练，在盘中并行对各个实时特征值训作为输出识别对应异常交易或波动，并以之为基础实施报警或处置措施。对于风险监测，利用人工智能的复杂多因子分析能力，可以把目标指数或标的历史交易数据，相关的宏观、中观数据、舆情数据，以及对应的风险指标数据进行训练，每日根据特征值动态更新，对风险进行识别、预警和处置。

参考文献：

[1] Diana Kafkes et al., “Applying Artificial Intelligence & Reinforcement Learning Methods Towards Improving Execution Outcomes,” SSRN, October 19, 2022.

[2] SEC,SR-Nasdaq-2022-079 Amendment 2 ,July 18,2023.

[3] Hasselt, H. Deep Reinforcement Learning with Double Q-Learning. Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16) 23, 2094-2100 (2016). 10.5555/3016100.3016191.

[4] Karpe M , Fang J , Ma Z ,et al.Multi-Agent Reinforcement Learning in a Realistic Limit Order Book Market Simulation[J].Papers, 2020.DOI:10.1145/3383455.3422570.

[5] Kochedykov D ."Introduction to Reinforcement learning with application for trade execution"[J]. 2017.

[6] 徐广斌等。个股动态涨跌幅限制机制：概念、实证及建议。上证研报[2017]001号.2017年2月.

知识图谱：驱动员工能力的引擎与机遇

潘建东、王赵鹏、马张晖、刘国杨、孙冰、尹序鑫、訾顺遥、梁彬

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

摘　要： 针对目前财富管理机构存在的一线员工学习压力大，客户服务针对性弱，专业人才流失，缺乏高效专业的协同工作等问题，中信建投证券提出员工赋能平台项目，将知识图谱与大语言模型等相关技术结合，弥补了金融领域数据缺少组织结构、价值密度低、难使用的缺点。员工赋能平台通过数据智能处理、专家生产工具设计以及认证鼓励机制开发，为员工构建出一个顺畅的生产环境，通过灵活组队服务功能，让一线员工可以随时提问专业信息，直接联系沟通到总部专家，快速响应客户需求。该系统大大提升了员工的工作和合作效率，挖掘出更多的业务机会。

关键词： 综合财富管理；人工智能；知识图谱；知识生产；知识应用

一、引言

财富管理指以客户为中心，设计出一套全面的财务规划，通过向客户提供现金、信用、保险、投资组合等一系列的金融服务，将客户的资产、负债、流动性进行管理，以满足客户不同阶段的财务需求，帮助客户达到降低风险、实现财富保值、增值和传承等目的。经过四十年改革开放，我国国民财富积累迅速增长，同时近年来房地产吸引力下降、资管新规、权益市场吸引力不断提高，居民财富增值的意愿达到了新高度，对金融机构高质量财富管理的需求日益强烈。

当前，证券公司等金融机构在进行财富管理业务时，普遍面临着庞大的客户群体与综合服务能力不匹配的问题：一线员工学习压力大，客户服务针对性弱；个体经验难以持续产生价值，出现专业人才流失现象；缺乏高效专业的协同等。为解决这些问题，中信建投证券积极探索利用科技赋能，提升财富管理赋能能力。通过建设员工赋能平台项目，降低知识生产门槛，实现知识数据可视化、业务规则数字化和自动化，有效链接用户、产品和知识等实体数据，打造开发出一套专业易用的专家知识生产系统。

二、基于知识图谱技术的员工赋能系统

如何整合来自web端、文档、音视频等多源非结构化数据，实现快速检索、多元互联等目标，利用大语言模型构建高价值密度、高利用率的垂直领域知识库显得尤为重要。中信建投证券团队通过建设员工赋能平台，利用先进的实体识别、关系抽取算法构建知识图谱，再基于实体对齐、链接预测技术对图谱进行补全和完善，得到高质量、高可用的垂直领域知识图谱应用于下游任务，助力一线员工和领域专家直接交互，辅助员工和专

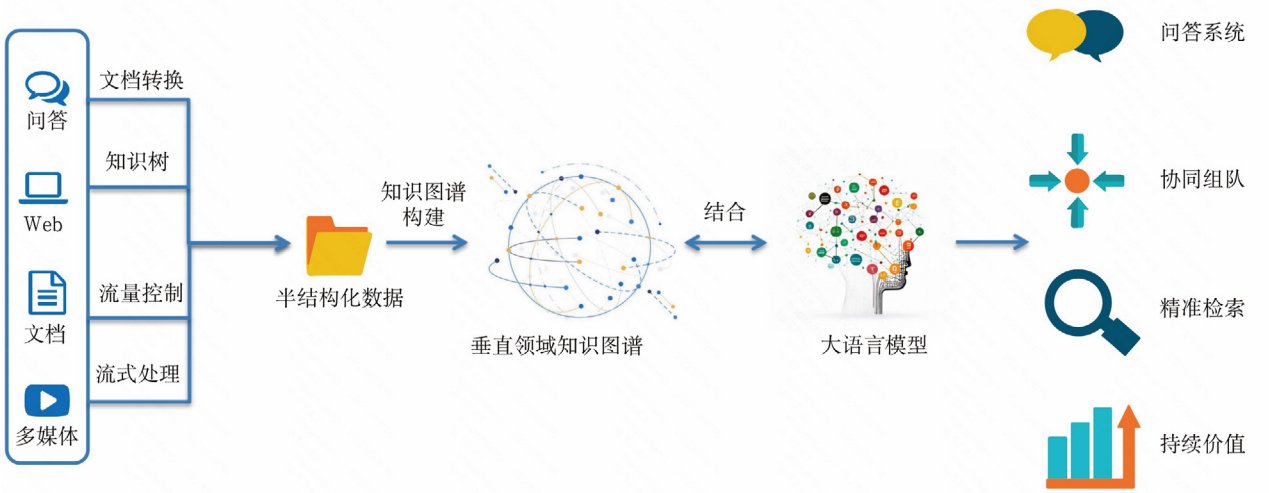


图1：员工赋能平台架构图

家开展合作，实现对客户全方位的服务提升。

如下图1所示，平台整体划分为数据接入层，知识生产层，知识应用层三个单元。数据接入层，负责接入并整合大量分散在数据中心服务器、员工电脑本地的异构非结构化文件等组织知识。知识生产层，进一步将数据构建为知识图谱形式，依靠图谱良好的关联性和高信息量特点，实现快速检索、推理等功能，满足员工作业需求。随后，在知识应用层中，知识图谱数据与大语言模型结合，拓展应用到员工赋能平台中的知识检索、问答系统以及协同组队等功能模块中。通过这种方式，平台将知识信息与专家信息录入知识图谱，以实体和关系的形式进行联合。一线员工不仅可以随时提问所需的专业知识、获取调用学习相关的服务文档、案例、经验等知识内容，还可以根据不同业务通过企微互联直接联系对应专家，灵活组队为客户提供高质量的综合性金融服务。

三、面向员工赋能平台的知识图谱实践

3.1 知识图谱数据处理

大量的组织知识分散在数据中心服务器和员工电脑本地的Word、PPT、音频、视频、图片等格式的文件中，平台需要能够兼容各种格式的数据类型，因此需要人工智能技术将各种数据类型格式的非结构数据进行初步的统一化、半结构化，然后才能支持后续精加工流程的顺畅实现。

员工赋能平台不仅要能“兼收并蓄”，同时也要能“多样绽放”。平台的重要任务之一是为基于高级算法构建的应用提供优质输入（例如基于图计算的推荐任务场景下，首先需要将客户数据生产为知识图谱，然后利用图计算相关算法进行客户分类和推荐），应用场景和算法的不同，要求输入的知识表示形式也不同。而多样化的知识表示形式，要求平台尽可能集成和支持多样化的人工智能算法模型，用于自动化地将数据生产为知识。平台通过使用流处理技术和实时数据集成工具完成实时数据集成，将不同数据源的数据实时整合到一个单一的视图中。实时数据集成技术可以帮助企业更快地做出决策，并提供更准确和实时的数据分析。集成后通过自动化数据准备，使用机器学习和自然语言处理等技术来自动识别、清理和转换数据，减少了手动劳动力，缩短数据准备的时间，并提高数据的准确性和一致性。最后通过无代码/低代码数据接入，使用可视化界面和图形化工具来简化数据接入的过程。通过这种方式，减少对技术专业知识的依赖，使更多人能够参与数据接入和分析，从而提高企业的数据文化。

3.2 知识图谱生成

员工赋能平台通过知识生产层得到高质量的知识图谱数据，用于下层知识应用。我们调研尝试了大量经典基线模型算法，并进行了对比与改进，最终确定了平台的算法实现方向。下表1罗列了经典算法与平台所用算法的对比情况：

数据接入模块输出大量结构化、半结构化或非结构化的数据，再统一利用实体抽取和关系抽取技术，将其转化为生产结构化知识图谱数据，用于下游NLP任务。

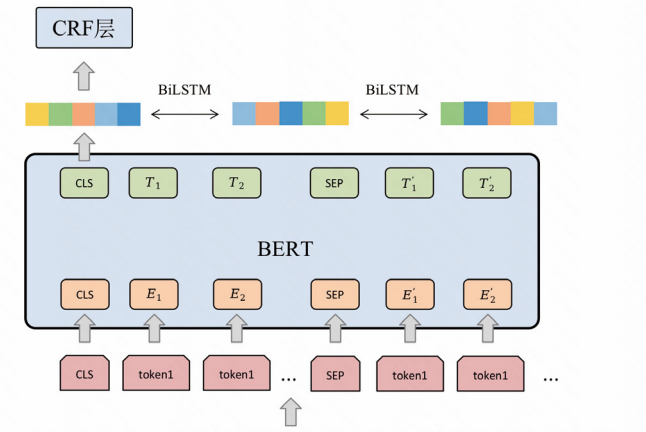


图2： 实体抽取模型结构

实体抽取部分，团队使用BERT预训练模型+BiLSTM+CRF的算法模型。经典的实体抽算法例如Word2vec模型+LSTM+CRF，将实体抽取看作文本序列标注问题，先通过Word2vec模型获得文本的初始嵌入向量，再利用LSTM对向量进行小范围内的二次聚合，最后用CRF替代Softmax，对标注结果进行规则上的限制。类

	经典算法	采用算法	对比总结
实体抽取	Word2vec 模型+LSTM+CRF	BERT+BiLSTM+CRF	更好的扩展性和适应性，支持小样本学习和持续学习
关系抽取	RNN、MultiR	GPT2 大语言模型	抽取更精准
实体对齐	GCN-Align	MuGNN	多粒度，泛化能力强
链接预测	GNN+TransE	T5+prompt	资源开销小，准确率高

表1： 算法对比

似的这类经典算法存在一些问题。首先，Word2vec模型生成的词向量均为静态词向量，扩展性不强，且生成速度较慢。LSTM模型虽然在小范围内可以对文本向量进行再次聚合，但聚合方向为单向，其能力也有待进一步提升。其次，经典算法将实体对齐任务视为序列标注问题，在面对小样本学习和可持续学习的任务上表现乏力。然而，公司的数据在不同业务领域种类较多、流量较大，需要实时更新、持续学习，部分领域还可能存在数据量较小的情况，使用经典算法效果欠佳。

BERT预训练模型+BiLSTM+CRF的算法模型可以在保持轻量级的同时克服上述问题，结构总览如图2所示。首先使用BERT预训练模型替换Word2vec，可以生成句子级别的表示，同时考虑了多个单词之间的语义关系。此外BERT可以通过微调来适应不同的任务和数据集，从而提高模型的性能和泛化能力。赋予了词特征向量灵活性，并减小了系统开销。BiLSTM可以同时考虑前向和后向的上下文信息，从而更好地捕捉序列中的依赖关系。在财报、年报、财经新闻等自然语言信息中，前后文信息对于理解信息的含义非常重要，因此双向性可以提高模型的准确性和泛化能力。对照实验数据见表2，其中准确率、召回率、F1值是考量模型表现的相关指标，越高说明模型精度越好。

	准确率	召回率	F1 值
Word2vec 模型+LSTM+CRF	0.76	0.68	0.71
BERT+BiLSTM+CRF	0.79	0.75	0.77

表2： 实体抽取实验数据表

关系抽取部分，团队使用OpenAI开源的GPT2大语言模型作训练和微调。GPT模型可以通过预训练和微调的方式来完成关系抽取。预训练阶段，GPT模型通过大规模的文本数据训练得到了广泛的语言知识和语义理解能力，这些知识和能力可以在关系抽取任务中得到充分的利用。微调阶段，GPT模型根据不同的关系抽取任务要求，进行微调和优化，从而实现更加精准的关系抽取。在具体实践中，GPT模型可以使用基于文本生成的方法来完成关系抽取。团队首先使用大规模语料文本对GPT模型进行预训练，为模型赋予语义理解、文本生成、结构生成的能力。然后使用财经、金融领域数据集在预训练模型上根据员工赋能平台的需要进行微调。

通过实体抽取和关系抽取，构建出结构化知识图谱数据后，需要对数据进一步进行补全和过滤。通过NLP技术构建的知识图谱，一方面可能存在遗漏的三元组，即两个关联实体间缺少关系链接，另一方面由于汉语一义多词的现象，可能存在重复的同义实体。以上情况都会影响知识图谱的信息准确性，破坏知识图谱的结构化特性，进而影响下游任务。为此团队通过链接预测和实体对齐技术，对知识图谱进行对齐和补全。

实体对齐部分团队采用MuGNN算法，MuGNN（Multi-Granularity Graph Neural Network）是一种用于知识图谱中实体关系抽取的先进的多粒度图神经网络模型。该模型的主要特点是使用了多粒度的图表示学习，它将知识图谱中的实体和关系表示为多层次的图结构。每个层次的图结构都对应一种不同的粒度，可以捕捉不同层次上的语义信息和关系。

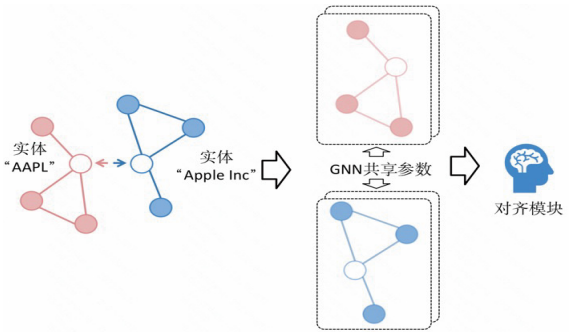


图3： MuGNN实体对齐模型结构

如图3所示，MuGNN模型的输入是员工赋能平台通过实体抽取和关系抽取技术搭建的知识图谱，其中包含实体和关系的信息，模型将知识图谱表示为一个多层次的图结构，每个层次的图结构都对应一个不同的粒度。在每个粒度上，MuGNN模型都使用MG-GCN进行特征提取和表示学习。同时，MuGNN模型还使用自适应注意力机制将不同粒度上的特征融合起来。团队通过使用多任务学习策略训练MuGNN模型，提高模型的泛化能力和效果。这个策略可以在多个任务之间共享模型参数，从而使得模型可以同时处理多个任务。相较于传统的单粒度图神经网络模型，MuGNN模型在员工赋能平台的知识图谱实体对齐任务中表现出色。表3展示了经典算法和团队采取的算法在相关数据集上的测试结果，MRR、Hits@1和Hits@10都高于基线经典算法。

	MRR	Hits@1	Hits@10
GCN-Align	0.549	0.413	0.744
MuGNN	0.844	0.494	0.844

表3： 实体对齐实验数据表

团队在链接预测的三元组分类问题上使用了T5-large模型和prompt技术，通过将这个问题转化为文本生成问题，使用T5-large模型+prompt模版来进行训练和推理，如图4所示。T5-large模型是一种由Google开发的大型预训练语言模型，可以用于多种NLP任务，包括文本生成、问答系统等。而prompt技术是一种将任务描述（prompt）嵌入到模型输入中的技术，可以帮助模型

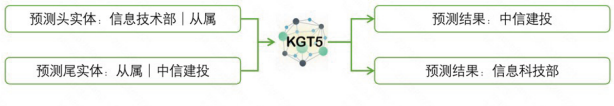


图4： KGT5模型（T5+prompt）推理过程

团队采用编码器-解码器结构，采用T5模型的编码器作为输入层，将输入的实体和关系表示为向量形式。T5编码器是一个具有多层自注意力机制的神经网络，可以将输入的序列编码为固定长度的向量。使用T5模型的解码器作为输出层，将生成的实体和关系表示为向量形式。T5解码器是一个具有多层自注意力和交叉注意力机制的神经网络，可以根据输入的向量生成文本序列。除了编、解码器，为了更好地表示实体的语义信息，团队使用了实体嵌入层和关系嵌入层。该层将每个实体和关系映射到一个低维向量空间中，以便于模型学习实体、关系之间的关联信息。表4中展示了赋能平台在链接预测任务上采用的算法和经典算法在相关数据集上的实验对比，相比于传统的链接预测算法，T5+prompt方法具有更好的扩展性和适应性，能够更好地应对不同领域和数据量的链接预测任务。此外，赋能平台所使用的改进算法还显著降低了参数量，节省资源开销的前提下提升了模型泛化能力和精度，这使得该算法具有很高的实用价值。

	MRR	Hits@1	Hits@3	Hits@10	参数量
TransE	0.253	0.170	0.311	0.392	2,400M
T5+prompt	0.336	0.286	0.362	0.426	674M

表4： 链接预测实验数据表

3.3 知识图谱应用

精准检索是知识应用的基础应用，是指在文本检索任务中，通过各种技术手段，提高检索结果的准确性和相关性，以满足用户的信息需求。在赋能平台的实际应用中，精准检索可以帮助员工快速找到所需的专业信息或文档，协助员工完成、理解任务，并作为基础应用服务于知识问答。团队使用大模型+知识图谱的架构实现精准检索。该方法利用大型语言模型的强大语义理解能力和知识图谱的结构化知识，实现对复杂自然语言查询的准确解析和精准匹配。

赋能平台首先使用清华大学开源的预训练大语言模型ChatGLM对自然语言查询进行编码，得到查询的向量表示。然后，利用知识图谱中的实体和关系信息，对查询进行语义解析，得到查询所涉及的实体和关系。接着，利用知识图谱中实体和关系的语义信息，对查询向量进行扩展，得到更加丰富的语义表示。最后，将扩展后的查询向量与知识图谱中的实体和关系向量进行匹配，得到与查询相关的实体和关系。

此外，该方法还可以支持多种查询类型，包括实体查询、关系查询、属性查询等，具有较好的可扩展性和适应性。知识图谱的结构化数据对于大语言模型对文本的理解和映射有着莫大的帮助，相对于直接使用非结构化文本，知识图谱的结构化数据对大模型的检索速度和精度均有一定程度的提升。

3.4 知识图谱的更新和维护

知识图谱的自动化定期更新和维护主要通过不断从各种网络资源和结构化数据源中抽取新信息来实现。具体来说包括:从日新月异的网络内容中抽取出新的实体、概念、关系和属性,将它们纳入知识图谱；分析新文本中已有实体之间的交集和联系,更新他们之间的关系。利用相似性计算、上下文分析等方法,识别出相同或新的关系,完善知识图谱结构；将抽取出的相似实体进行合并,利用现有知识和规则,对新抽取的信息进行验证和推理,判断其真实性和完整性,从中不断积累新的规则和知识。由于金融信息的时效性和安全性特性，团队使用以下技术对知识图谱进行更新和维护，保证信息有效安全：

1.数据抓取和清洗：在知识图谱的自动化更新和维护中，数据抓取和清洗是非常重要的步骤。团队使用网络爬虫技术从各种数据源中抓取新的数据，并使用数据清洗技术进行数据处理和转换，使其符合知识图谱的格式和要求。数据清洗包括数据去重、数据标准化、数据转换等多个步骤，以确保知识图谱中的数据质量和准确性。

2.知识生成：对于抓取和清洗的新信息，团队使用上文提到的知识生成技术扩展和补充知识图谱中的实体和关系。并对通过实体抽取、关系抽取构建的新知识图谱做质量监控和修正，以保证知识图谱的高质量。

3.定期删除：在知识图谱中，一些信息可能会随着时间的推移而失效或过期，因此需要对这些信息进行删除或标识。但是，并不是所有的信息都可以自动删除，需要根据具体情况进行判断和处理。对于一些时间敏感的信息，例如新闻、股票价格等，可以设置过期时间，超过该时间后自动删除。对于一些长期有效的信息，例如历史事件、基础知识等，应该保留在知识图谱中，以便后续的查询和分析。在员工赋能系统中，团队通过人工审核和自动化模型等方式进行信息删除和标识。使用ARIMA模型，利用时间序列分析技术对某些信息的变化趋势进行预测，从而判断其是否已经过期。该模型可以用于分析时间序列数据的趋势、季节性和周期性等规律，利用已有的时间序列数据，预测未来的趋势。如果发现某些信息的趋势已经不再变化，则可以判断该信息已经过期。

4.可视化和查询接口：为了让用户更加方便地使用和查询知识图谱中的信息和知识，可以使用图形化界面和查询接口。通过图形化界面，用户可以直观地浏览和操

作知识图谱中的实体和关系；通过查询接口，用户可以根据自己的需求查询知识图谱中的信息和知识。同时，也可以通过用户的反馈来自动更新和维护知识图谱，以不断提高知识图谱的质量和价值。

四、建设成效与总结

中信建投证券目前已经完成知识图谱构建及应用的算法设计和落地，实现了无结构化文档自动构建知识图谱，以此作为外部知识库增强大语言模型的下游任务，嵌入到员工赋能平台中，为员工提供出开放接口。目前平台优化完善已迭代3个版本，未来员工赋能平台2-3年的规划目标是：打造证券行业垂直领域的“知乎”，基于时序知识图谱技术和强化学习算法完成实时事件分析系统，针对国内外突发事件开展业务分析并及时响应客户的投资需求变化，挖掘重大事件背后的商机，协助员工展业。

中信建投证券正在以智能化为主导思想，不断深入推进智慧营销平台的建设，旨在打造全周期数字化智能营销服务，以数据驱动为客户提供适宜的服务和产品，从而实现高效、合规、精细的服务。这种以数据为核心构建的智能化体系将成为支撑未来券商发展的关键要素。在金融行业，智能化建设具有广阔前景，并将对未来证券业态发展产生深远影响。

人工智能大模型在证券行业应用路径与实践

邓维、易卫东 | 招商证券股份有限公司 | E-mail：foreverdengwei@163.com

摘　要： 2023年中央金融工作会议提出要做好数字金融这篇大文章，2024年政府工作报告提出要开展“人工智能+”行动。大模型人工智能技术作为新质生产力，证券行业应当积极拥抱大模型进行数智变革。本文详细分析了国内外证券大模型的应用场景、实践案例和技术路径，希望为国内金融机构应用大模型提供借鉴参考。首先，本文简介了国内外通用大模型和金融垂类大模型发展概况；然后，本文基于证券行业场景分类，分别阐述了大模型在外部服务、内部提效和监管科技三类10大业务场景的模式变革和应用路线图；其次，本文梳理了国内外Bloomberg、金融科技企业Broadridge和投行J.P.Morgan等金融机构各类主体在智能客服、智能营销、智能投顾、智能投研、智能风控、智能运营、AI助手等各方面的实践案例；最后，本文对比分析了金融大模型落地的通用VS垂类、开源VS商业、私有VS公开等三种技术路径和系统架构，并对未来挑战与趋势作出简要总结，期待证券行业共同努力建立体系化的大模型建设蓝图，共赴数智金融新时代。

关键词： 大模型；智能投顾；智能投研；智能投资；智能风控；智能运营

一、全球人工智能大语言模型发展现状

2022年底OpenAI发布基于大模型、大数据、大算力的ChatGPT，意味着人工智能（Artificial Intelligence，AI）的发展到了从“弱AI”向“强AI”跃迁的分水岭。

国外通用大模型发展现状。国外基本上是OpenAI和Google等公司主导，顶尖的资本、技术、人才较为集中。尤其是微软所投资的OpenAI每隔几个月就更新产品，2023年3月发布万亿参数多模态大模型GPT4、2023年11月发布GPT4升级版GPT4-turbo、2024年1月发布GPT4应用商城GPT Store、2024年2月发布文生视频大模型Sora，持续引领着大模型发展方向。虽然开源大模型在实力上相较于商业公司还存在一定差距，但是对于大模型生态具有重要意义。尤其是2023年7月Meta开源最新大模型Llama 2、2024年3月马斯克旗下xAI发布旗舰大模型Grok-1，加速大模型生态构建和繁荣。

国内通用大模型发展现状。国内大模型还处在“千模大战”状态，商业和开源群雄并起，目前没有拉开显著差距。2023年3月百度发布“文心一言”，4月阿里巴巴发布“通义千问”，5月科大讯飞发布“星火”，7月华为发布“盘古”大模型，9月腾讯发布“混元”大模型；创业公司智谱和Moonshot也拿到了入场券……但当前技术仍未达到国外顶尖水平。

国内外金融垂类大模型发展现状。金融行业一直是数字化技术最早实践者。金融大模型始于2023年3月BloombergGPT。Bloomberg沉淀了大量高质量数据，利用大模型分析处理，大幅提高金融服务效果，比如可以分析市场情绪和金融资讯，提供精准投资服务。Morgan Stanley等知名金融机构均发布大模型应用。与此同时，中国机构也不甘落后，迅速推动金融大模型的应用实践，而且更加贴近产业端。而国产金融大模型也已分出了明显的两个“流派”，一派来自于券商等传统金融机构，另一派来自于金融科技企业，如恒生电子LightGPT和蚂蚁金融大模型。

二、大模型重塑证券行业十大业务场景

证券行业天然具有数据属性和良好数字化基础，强烈的数智化转型需求与多样化的业务需求，以及对新兴技术的较高接受度与资金支持度，使得金融行业是大模型最佳应用场景之一。

图1：大模型重塑证券业务场景一览

大模型重塑证券业务场景的分为外部服务、内部提效和监管科技三大类：外部服务包括智能投行、智能投

类型	场景	当前模式与问题	大模型 AI 重塑模式
业务服务	智能投行	数据繁多处理难；文字工作量大；核查依赖人工	智能抽取信息；智能生成报告；智能核查底稿
	智能投顾	存在服务同质化、分析不够及时、投顾效率低等问题	精准分析客户；一键创作内容；自助投顾服务。
	智能投研	数据源和数据获取上依赖传统金融数据及结构化数据库，数据处理上依赖于研究人员自身能力	全面收集信息；自动分析数据；自动撰写报告；智能服务客户
	智能投资	主观投资依赖于人的主观判断和经验，而量化交易也存在数据来源有限、分析不及时、因子挖掘困难等	数据自动处理；因子挖掘与量化策略编写；投资指令智能解析；实时风险监测和调仓
	智能营销	客户画像难，服务不精准，营销效果难评估	精准分析客户画像,个性化推荐服务；快速内容创作，生成专业话术；智能服务匹配、智能工单创建；运营策略探索分析
	智能客服	搜索引擎和 NLP 的语言生成能力较弱，机器人对话明显，服务温度大大低于人工客服	提升智能客服的拟人度和人机协同效率；自动生成会话内容摘要；智能分析咨询结构统计与热点识别监测，补充语料
内部提效	智能风控	建模效果有限、小样本数据不足，性能不达标，单点的防御和预测能力	自动构建风控模型；实时风险智能识别；跨风险类型的预警；风控合规助手服务
	智慧运营	数据多、环节多、操作风险大，基于 OCR、NLP 等的标注量大、规则维护成本高、扩展性差等问题	文档智能分片、要素智能提取、合同条款智能匹配、审计内容智能审核
	智能办公	依靠人工和办公软件完成，依靠大量开发、测试和运维人员编写代码、测试案例和运维分析	自动会议纪要、工作日报、发言稿、邮件等任务，提升和改变办公方式。智能研发（代码生成等）智能运维（智能工单等）
监管科技	身份识别、监管数据、风险监测	数据核验来源不完整，无法及时准确判断等	大模型多模态信息处理能力、多模态交叉身份验证、处理多维风险数据，提升准确度

表1： 大模型重塑证券行业十大业务模式

顾、智能投研、智能投资、智能营销、智能客服等6项业务场景；内部提效包括智能风控、智慧运营、智能办公（含IT研发)等3项场景；监管科技这些场景涉及身份识别、监管数据支撑、风险监测等内容。当然这些场景并非全面，仅以点带面作为示例说明，见图1。当前十大场景都存在一些数据不全面、泛化能力弱、人工效率低等问题，大模型AI凭借多模态、泛化性、通用性、生成能力，将重塑现有业务模式，参见表1。

当然这些场景并不都是一蹴而就齐步实现。从技术成熟度和场景价值度维度分析，智能营销、智能客服、智能办公、智能开发运维等场景是相对技术最成熟和容易落地的场景（见图2）。智能投研、投顾、风控是证券业务高需求的价值场景，是大模型有希望尽快落地下来放大业务的场景。而智能投资、智能投行、产品设计这些场景相对技术成熟度较低，未来落地需要一点时间。未来智能代理Agent将独立能够感知环境、进行决策和执行任务，成为大模型的新载体，不断重构金融业态。

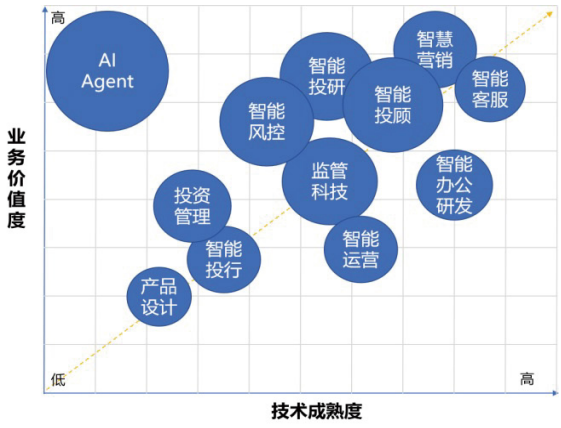


图2：大模型金融行业应用路线图分析

分类	机构	模型	领域	应用	时间
资讯公司	Bloomberg	BloombergGPT	客服、投研	Bloomberg 终端智能问答和资讯分析	2023 年 3 月
金融科技公司	Broadridge 子公司 LTX	BondGPT	投顾	为机构客户提供债券投资建议	2023 年 6 月
	Stratosphere	FinChat	投研	为投资者精准提供上市公司可靠数据的 ChatGPT 工具	2023 年 4 月
投行券商	摩根士丹利	AI@Morgan Stanley Assistant	投顾	在财富管理领域基于知识库赋能投资顾问	2023 年 3 月
	摩根大通	IndexGPT	投顾	分析和选择适合客户需求的证券，提供投资建议	2023 年 5 月
		ChatGPT	投研	“鹰鸽指数”货币政策预测模型	2023 年 4 月
	日本大和证券	ChatGPT	办公	员工写电子邮件、商业计划书、策略提案	2023 年 4 月
资管机构	贝莱德	GenAI	投研、办公	提高客户 Aladdin 信息搜集效率，同时赋能员工研究报告和投资建议	2023 年 12 月
	TwoSigma	ChatGPT	投研	分析财务报告和新闻内容，以识别潜在的投资机会	2023 年 3 月
	Citadel	ChatGPT	办公	用于软件开发和信息分析	2023 年 3 月

表2：国外金融大模型应用案例一览

三、 全球证券行业大模型百项应用案例

正如上文分析，国内外金融机构正围绕着这些业务场景链条，大力推动大模型技术与传统证券业务的融合落地，推出不下上百项创新应用。

3.1 国外金融大模型应用

国外金融大模型应用早于国内，三类机构纷纷布局，以传统金融机构为主：（1）传统投行等金融机构，以摩根大通为代表，其优势是投入资金充足、金融业务理解力强和高质量数据标注能力；（2）金融资讯公司，以Bloomberg为代表，其 IT 能力和积累的海量数据库具有核心优势；（3）金融科技公司，以推出FinChat/BondGPT等代表性产品的科技公司为代表，其优势是创新动能较强，并加速推出开源模型。

从案例上分析，一部分机构是自主研发垂类GPT大模型应用，一部分机构是基于通用ChatGPT等开展应用，场景上则涵盖了智能营销、智能客服、智能投研、智能投顾和智能办公等各方面（见表2）。

具体以摩根大通的智能投顾场景进行说明。摩根大通AI成熟度全球排名第一（询公司Evident Insights），拥有1000多名数据管理人员，900多名数据科学家，600多名机器学习工程师，200多名一流AI研究人员。2023年4月摩根创制一个解码央行的AI模型“鹰鸽指数”Robo-Fedwatchers模型，以25年以来的美联储声明和央行官员们的讲话为训练材料，使其可以预测到潜在的市场变动，进行鹰鸽等级评分，再将具体评分与一系列资产表现挂钩，用于预测政策的变化，并发出可交易的信号。2023年5月摩根注册“IndexGPT”，该系统加载了30年来所关注的所有公司的专有数据，通过分析客户需求并利用人工智能技术，提供更加智能化和个性化符合客户投资需求的证券推荐（见图3），IndexGPT采用的GPT模型在生成关键词数量上是以往软件的2倍多，使得主题的投资代表性更强。

Evident Insights AI Index
（人才、创新、领导力和透明度四维度）

IndexGPT服务（来源美国专利商标申请文件）	
投资服务	提供和更新证券价值指数，证券投资业务咨询
金融分析	对全球计算机网络或数据网络在线提供的金融信息进行分析，供具有见解的金融分析服务
证券投资	为证券投资者提供财务信息、推荐和分析服务
模型	用于金融服务领域的生成式预训练Transformer 模型（与GPT系列模型同款），是一款通过SaaS方式帮助客户分析和推荐证券的软件

图3：摩根大通AI大模型应用情况

3.2 国内金融大模型应用

国内布局金融大模型的机构与国外类似，主要是金融资讯公司如同花顺、金融科技公司如恒生和传统券商等金融机构（见表3）。智能客服、智能投研、智能运营、数字员工、AI助手是最常见的主要应用方向。银河证券与百度合作打造场外衍生品交易服务大模型，广发证券合作研发大模型率先应用于智能客服场景，中信证券债券助手Bond Copilot缓解投行债券全链条承揽、承做、承销三大环节工作，而海通证券“泛海言道”、国泰君安“灵犀布道”、东吴证券“东吴秀财GPT”均在探索智能投顾、智能投研、智能运营、智能风控和办公等场景应用。幻方、兴证、工银瑞信等基金研制GPT大模型用在数字人问答、数据分析师、AI交易员等领域。其中，招商证券在智能运营、智能研究、智能投顾、智能办公和编码助手等多个场景都在探索应用，例如托管大模型客服准确率高于90%，并且支持至少3轮上下文和事后标签自动分析，客户体验大幅提升。

类型	机构	领域	模型	应用	时间
券商	招商证券	运营	助手	托管智能客服、运营助手	2023 年 12 月
	海通证券	多场景	泛海言道	智能问答、智能研报、智能研发、数智人 2.0	2023 年 12 月
	国泰君安	多场景	灵犀布道	智能投顾、智能投研、智能投行、智能运营、智能风控、智能协作、智能运维等	2023 年 12 月
	东吴证券	多场景	东吴秀财 GPT	涨跌分析和盘后总结，企微 AI 客服助手、智能尽职调查、年报问答、基金问答、量化投资等	2023 年 12 月
	国金证券	客服、办公	FinGPT	选用 LangChain 和 ChatGLM2 交互式问答、代码助手、AI 办公助手	2024 年 1 月
	中信证券	投资	Bond Copilot	债券助手缓解投行债券全链条承揽、承做、承销三大环节工作	2024 年 1 月
	银河证券	投资	百度云	百度智能云金融智能场外交易发现平台，场外衍生品交易服务	2023 年 9 月
	广发证券	多场景	自研+合作	智能客服、财富管理、机构交易、投资银行、企业运营等领域	2023 年 11 月
基金	兴证全球	投资	“兴宝”AI 交易员	交易员资金交易场景深度结合，运用机器学习、自然语言处理等人工智能技术，打造意图判断、多轮对话、交易直达等核心能力	2023 年 6 月
	工银瑞信	客服，办公	FundGPT	数字人问答技术、检索增强大模型、数据分析师伴侣	2023 年 9 月
	幻方量化	投资、办公	DeepSeek	数学能力表现突出，指令跟随、编程能力领先	2023 年 11 月

表3：国内金融大模型应用案例一览

四、证券行业大模型落地三大技术路径

当前金融行业应用大模型主要有两种模型类型：一是通用大模型，如GPT4，二是金融垂类大模型，如FinGPT。大模型常见的合作路线包括开源大模型应用、商用大模型采购合作、产学研联合创新大模型研制、完全自主研制大模型等几种方式。大模型部署方式分为私有化部署、行业云部署和公有云部署等。

从上述应用案例和业务趋势可以看出，不同的金融机构选择了不同的技术路径：有的使用通用大模型，有的使用金融垂直大模型，有的应用开源模型，有个与商业模型合作，有的应用部署在公有云上，而大多数采用私有化部署。证券行业大模型技术路径选型可以从三方面进行分析：

(1) 企业规模大小：大型企业通常采用多种模式并行的混合路径。既会投入人力物力自主研发私有化部署的垂类金融大模型，也会商业化采购垂类大模型进行本地化合作，还会使用云端的开源的通用大模型进行尝试探索。中小企业由于缺乏大算力硬件和大算法人才，所以普遍会先采用使用公有云的开源通用大模型和垂类大模型进行探索试用。

(2) 数据敏感程度：数据敏感场景适合采用私有化部署模式，自研更可靠。数据不敏感场景可以采用公有云或者行业专用云模式，可以使用开源和通用大模型。

(3) 业务场景价值：业务价值高场景适合金融垂类模型，私有化部署更适合，自研或商业化合作更有保障。而价值一般通用场景可以采用通用大模型云端模式。

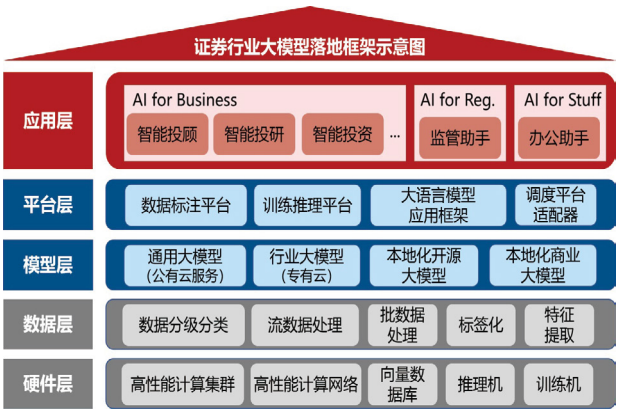


图4：常见金融大模型技术架构

这里也分析一下金融行业大模型落地常见技术架构和路径（见图4）。最底层是基础设施，包括强大的计算和存储的硬件，以及完善的金融大数据语料和数据平台，这是应用落地的一个重要基础。中间层是模型层和平台层，模型层可能既有通用大模型、行业大模型，也会有本地化商业大模型和开源大模型，而平台层包括数据标注、模型机器学习的框架、模型训练和模型调度等能力。应用层则是面向证券业务场景、面向内部员工、面向监管机构的各类赋能场景。

五、证券行业大模型应用挑战以及趋势

尽管大模型在证券行业的应用前景广阔，但目前还处于起步阶段，面临诸多挑战：既有技术方面挑战，大模型在训练过程中存在的大硬件算力、海量内存和高效通信的挑战，以及机器幻觉问题在容错率较低的金融领域可能会影响分析结果的可信度；也有业务方面挑战，金融行业应用场景缺少通用范式、欠缺高质量金融训练数据；还有监管方面的难题，由于金融行业合规要求高，严格保护数据安全和隐私，金融机构和监管机构需要投入大量资源以满足合规要求。为了应对这些挑战，未来金融行业还有一系列工作要做，包括协同共筑AI算力基础设施以及完善算法优化与风险管控，加强数据治理工作和积极构建高质量金融数据集，加强金融应用的最佳实践指导与监督管理等。那么，随着大模型技术不断升级、金融业务场景不断落地探索、监管的不断完善，大模型落地能力更强、场景更广、应用更深、价值更高，重塑金融行业基础设施、组织架构、业务模式、经营模式，金融行业共赴数智金融新时代。

参考文献：

[1]12世纪资管研究院. 大模型重塑金融业态报告[R],2024-2-5.
[2]林建明. AIGC重塑金融：AI大模型驱动的金融变革与实践[M]. 机械工业出版社出版，2024年2月.
[3]中国银保传媒、腾讯研究院、毕马威. 2023金融业大模型应用报告[R].2023-11-9.
[4]吴一凡. 海外AIGC金融有哪些落地进展[R]. 长江证券，2024-2-2.
[5]清华大学、度小满、MIT科技评论.2024年金融业生成式AI应用报告[R]. 2024-1-24.
[6]潘玉蓉. 机构争分夺秒抢滩 金融大模型落地为时尚早[N]. 证券时报网, 2023-11-30.
[7]苏仪. 券商大模型应用进展几何[R]. 中泰证券，2024-2-21.
[8]北京金融信息化研究所. 大模型金融应用实践及发展建议[R]. 2023-11.
[9]参见《境外金融机构AI大模型应用案例：投行券商、资讯公司、金科公司各显神通》，微信公众号：资管业务与科技.最后访问日期：2024年5月25日, 网址：https://mp.weixin.qq.com/s/OXiuXeR5S6BQ8Sx7VIai1g.

基于深度强化学习的客户资产均衡建模

王瑜、褚丽恒、刘敏慧、梁钥、侯立莎、邱子聪、石宏飞、李海英

申万宏源证券有限公司｜E-mail：wangyu4@swhysc.com

摘　要： 财富管理业务在我国迅猛发展，证券行业积极进行财富管理转型，以客户为中心的服务体系基本建立。但目前，主要依靠客户经理提供一对一的投资建议，服务成本高且覆盖有限。为此，申万宏源证券有限公司将人工智能技术与资产均衡理论深度融合，采用深度强化学习对客户资产均衡性进行全方位建模，从客户兴趣、客群挖掘、资产多样性、资产分布、投资组合理论、资讯舆情等多个维度建立个性化、专业化的模型，为每位客户提供智能化、针对性的投资建议，提升客户服务质量。本研究基于证券行业现状及客户数据，打通从数据底座到数据智能化应用通路，为客户提供“千人千面”贴身服务。目前已取得初步业务成果，助力产品销售额达到15.1万，客户购买转化率提升约2倍，客户资产均衡性向良性方向发展。

关键词： 资产均衡建模；深度强化学习；数据智能化

一、我国财富管理市场现状

随着中国财富管理市场的迅速发展，截至2020年末，中国的个人金融资产已达205万亿，互联网财富管理市场达8.2万，同时客户人群也在持续年轻化，互联网财富管理的主要客群为21-35岁，精准营销成为证券公司的重要增长点。为了做好精准营销，证券公司需要关注三大关键要素：1）以服务客户为中心；2）提供丰富优质的金融产品和内容服务；3）利用人工智能技术实现最合适的产品推荐。同时，内容服务已经成为券商APP的重要业务，通过自动化实时获取客户偏好和有效的舆情信息，纳入资产均衡模型，提高资产均衡评估的精准度，是该模型优于已有资产均衡评估方法的重要维度。这些举措有助于建立良性健康的合作模式，增强证券公司的流量能力、顾问和陪伴能力，以及产品与投资能力，为客户实现财富管理目标提供更好的支持。

二、客户资产均衡模型

2.1 客户资产均衡建模问题

资产配置是以投资者的风险偏好为基础，通过定义并选择各类资产类别、评估资产类别的历史和未来表现，来决定各类资产在投资组合中的比重，以提高投资组合的收益-风险比。资产配置的核心，是资产种类和具体投资的多元化，即资产配置均衡。客户资产均衡建模就是从客户需求出发，为客户提供专业化、智能化、个性化的投资顾问服务。

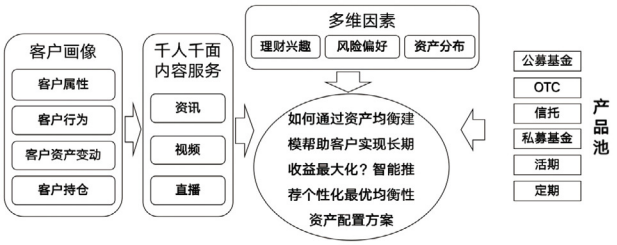


图1：客户资产均衡建模问题定义

如图1所示，一方面，基于完备的客户相关数据实现对客户全面了解，有效识别客户偏好，为客户提供千人千面服务。另一方面，证券公司拥丰富的产品池。客户资产均衡模型解决的问题就是如何在综合考虑多维因素的前提下，对资产均衡性进行多维有效评估，帮助客户实现长期收益最大化。

2.2 建模总体思路

客户资产均衡模型旨在为金融产品精准营销提供智能化、专业化、个性化模型指引。金融产品精准营销倡导理性投资,追求符合投资者预期的资产收益。互联网个性化推荐模型倾向于迎合客户偏好，客户资产均衡模型的目标是客户长期收益最大化，二者存在本质差异。建构客户资产均衡模型时,需要建立完备的数据支撑和理论指导，除考虑客户偏好外,还需对其资产分布、风险能力及投资舆情进行建模。

基于以上差异，选择用深度强化学习技术进行客户资产均衡建模，进而解决精准营销问题，主要有三点原因：第一，强化学习具有长远眼光，注重长期回报，与客户资产均衡模型目标匹配；第二，强化学习解决序列

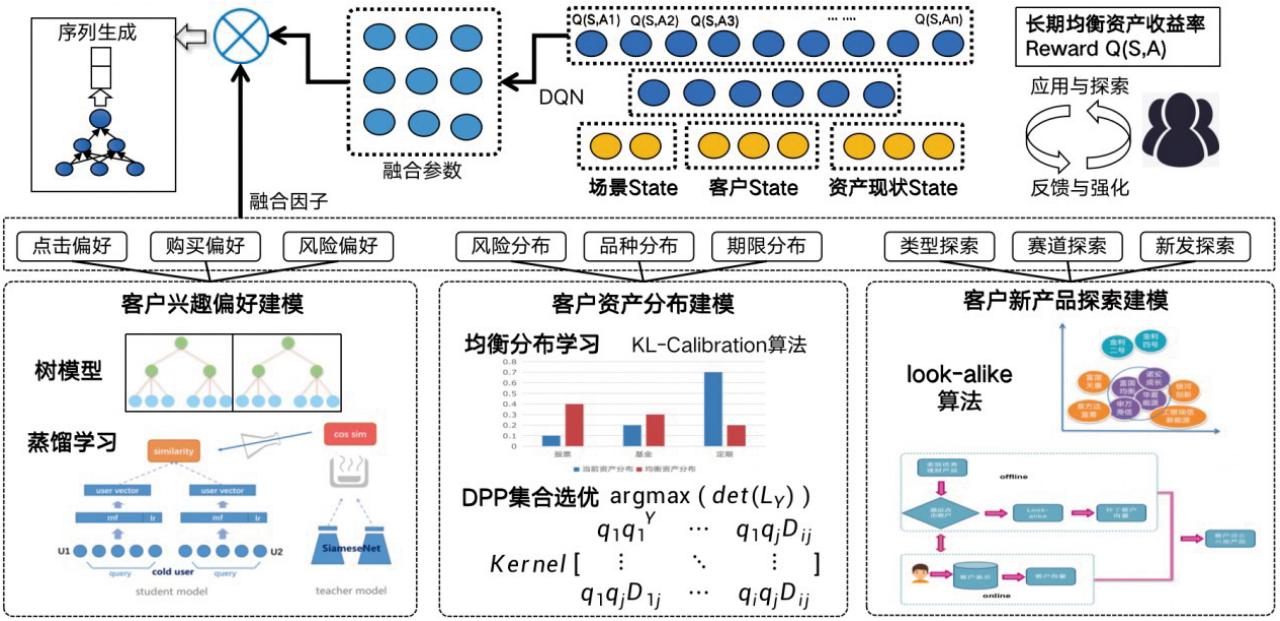


图2：资产均衡模型核心框架

行动的决策优化问题，会从训练数据中不断获得反馈，持续加强对当前环境的精准反应，客户投资组合策略的持续变化就是一个序列决策问题；第三，Model-Free深度强化学习(如DQN)无需建模环境，无需大量标注数据，用Action试错进行自学习。在客户资产均衡建模时，证券市场环境复杂，变幻莫测，难以建模，强化学习可巧妙解决此问题。

为解决精准营销问题，构建如图2的资产均衡模型核心框架。首先，对客户兴趣偏好进行建模，采用树模型、蒸馏学习建模客户偏好，学习客户产品浏览购买倾向，以此加深对客户的理解。其次，对客户的资产分布进行建模，资产分布的平衡性是客户资产均衡性的重要评估指标，这也是金融产品精准营销场景与互联网推荐产品的核心差异之一。再次，需要探索客户的潜在兴趣，探索对客户而言不在其之前偏好范围内的优质产品。

基于这三个维度的建模，应用深度强化学习DQN网络进行多维度因子融合参数自适应学习。DQN强化学习的目标是客户长期资产收益率，以该目标为导向进行持续应用探索、反馈强化，最终通过融合因子和融合参数的结合，进行最终排序，形成客户资产均衡的结果集合推送给业务端。

2.3 客户资产均衡模型详解

客户资产均衡模型主要包括客户兴趣偏好建模、客户资产分布建模、客户潜在兴趣探索三个部分。

2.3.1 客户兴趣偏好建模

不同客户产品偏好不同，传统基于产品维度的营销已经无法满足客户的个性化需求，精准识别客户兴趣成为关键。该问题的解决思路是区分新老客户。首先，对于行为较多的活跃客户，数据基础足够支撑进行GBDT建模，学习活跃客户的点击偏好、购买偏好、风险偏好。另一方面，对于新客户而言，也就是非理财客户，行为和购买信息非常少，需采用蒸馏学习方法。在图3中左侧的Teacher Mode中，训练的是活跃客户的相似度，数据是最底层全部的数据特征，这样进行训练会得到较高的学习精度，计算结果也是所有理财老客户相似度。进而通过蒸馏萃取，输入到中部的Student Model，左侧是一个新客户，右侧是一个活跃客户。这时Student Model并不会采用客户购买行为等数据作为特征，蒸馏萃取以当前新客户为基础，找到与其最相似的老客户，用老客户偏好间接表示新客户偏好，提升对新客推荐精准度。

2.3.2 客户资产分布建模

客户资产分布学习是客户资产均衡模型的核心评估指标。根据资产组合平衡模型理论，投资者根据自身投资偏好，按照风险收益原则将财富分配给各种可供选择的资产，形成最佳资产组合。一方面通过人工智能算法进行客户资产分布学习，基于KL-Calibration的资产分布学习是以客户历史数据为基础，挖掘判断客群中表现优秀的产品组合，使用客户投资组合的夏普比率来引入对风险的考量，形成理想产品投资分布。另一方面，在进行产品推荐时，在符合客户兴趣偏好前提下，呈现更多与其持仓产品差异大的产品可提升资产均衡性，采用DPP

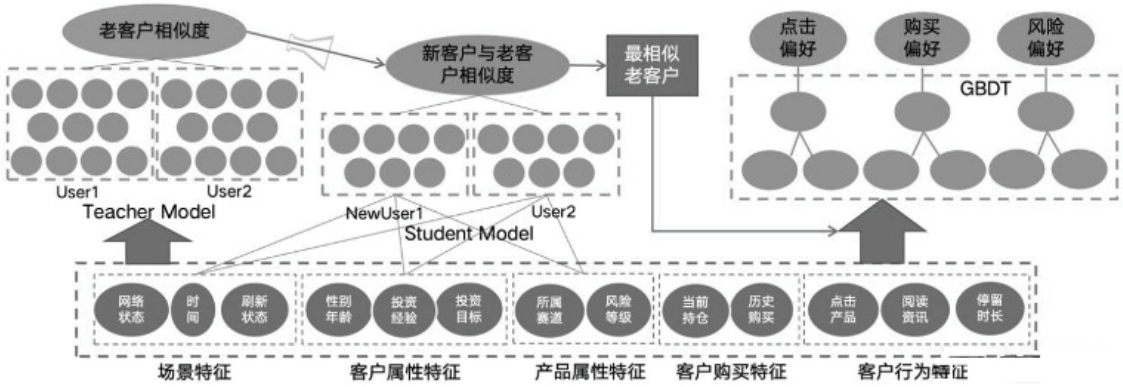


图3：客户兴趣偏好模型

集合选优算法实现此目标。一般有两种提高多样性的方法：1）使用规则控制；2）使用算法控制。常见的算法模型包括有界贪心选择策略（Bounded Greedy Selection Strategy, BGS）、最大边际相关性方法（Maximal Marginal Relevance, MMR）和行列式点过程方法（Determinantal Point Process, DPP），基于DPP具有能兼顾相关性和多样性且效率高的特点。

2.3.3 客户潜在兴趣探索

建设客户资产均衡模型需对客户潜在兴趣进行探索。客户兴趣偏好和资产均衡建模基于客户已有数据进行，客户资产均衡模型需要探索功能对客户未知兴趣进行探索，防止信息茧房。

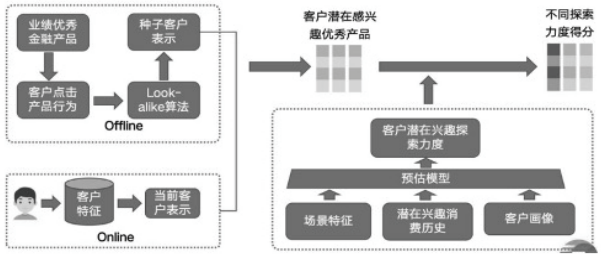


图4：客户潜在兴趣探索模型

如图4所示，通过Look-alike算法，用业绩优秀金融产品对客户进行试探展示。使用业绩优秀的金融产品进行探索，是因为即使客户并不感兴趣，也会被认为是一个好的产品，不会引起客户强烈的反感。首先，找到业绩优秀产品池，通过客户对产品的点击购买行为采用Look-alike算法找到产品的“种子”客户表示。面对某一客户，根据当前客户表示，找到与他距离最相近的种子客户所对应的产品进行兴趣探索。为客户推荐与他相似的人所偏好的优秀产品来作为他的探索产品，探索客户的潜在兴趣，客户兴趣探索是在不同时间、不同场景下，进行的不同力度的探索。

2.3.4基于深度强化学习的多维因子精准融合

完成对客户兴趣偏好、资产分布以及兴趣探索建模后，可得到客户对不同产品的多维度评分，需对资产均衡性评估的多维度因子进行有效融合，才能最终为客户提供符合资产均衡性的资产配置建议，此模型的核心部分为最终采用深度强化学习的DQN技术对多目标得分进行自适应融合，得到客户资产均衡性全面评估。DQN融合技术在图5算法中，S代表当前客户的场景属性以及资产现状等，A代表多个融合因子参数的离散值的结合，R代表客户购买产品后对短期的收益。客户购买产品后，其获得的长期收益就是整个强化学习去学习的目标Q*(S,A)。

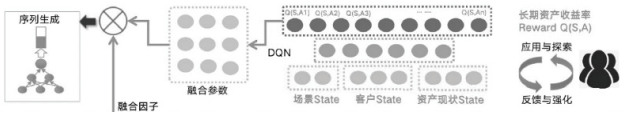


图5：基于深度强化学习的多维因子融合

2.4 智能内容理解模型

客户的资产购买意愿除了对各类资产的直接行为以外，也可以从其他方面得以体现，如客户在APP内对各类资讯的阅读偏好。实时对舆情信息进行智能内容理解，能有效捕获对投资决策有价值信息，辅助客户投资决策。通过建立智能内容理解模型获取客户对资讯文章的偏好，进而发掘客户未购买的偏好产品，增强客户购买意向，一定程度上达到使客户资产趋向均衡的目的。

智能内容理解模型的建设主要包括内容理解和模型建设两部分。内容理解以对资讯内容的打标签、去重和情感分析为主；模型建设通过机器学习和深度学习模型等对客户进行智能推荐，推荐该客户可能感兴趣的资讯内容，智能理解客户需求，辅助客户进行投资决策。

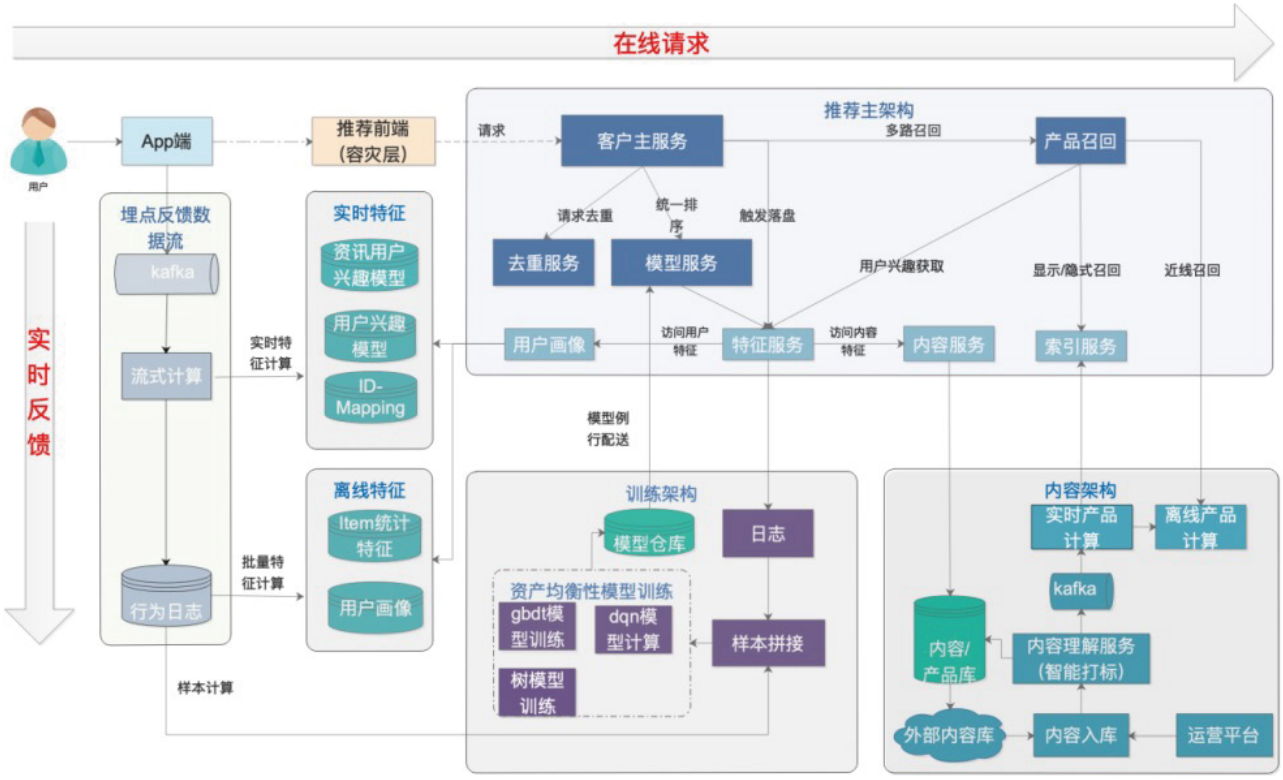


图6：基于资产均衡模型精准营销主服务架构

在对客户进行资讯等内容智能服务时主要用到Xgboost和Wide&Deep等机器学习模型和深度模型。Xgboost模型由多棵树模型组成，原理是通过不断添加新的树模型拟合前一树模型预测的残差值来提升模型训练的精度。树模型对特征特别是缺失值的处理具有天然优势，Xgboost在智能内容理解模型中的使用主要集中在特征选择阶段，筛选出重要度前100的特征，再将特征引入深度模型进行训练和预测。客户资产均衡建模场景和互联网场景不同，需要较高的模型可解释性。纯互联网场景的推荐排序模型目前大多使用如ESSM或MMOE等深度模型，层层网络迭代使推荐越来越精准，但纯向量化的输入和计算也让模型越来越复杂，可解释性不强。结合互联网和金融两类场景的不同特色，资讯智能推荐的模型使用的是Wide&Deep排序模型，此模型结构上由线性模型（Wide）和DNN深度模型（Deep）两部分构成，Wide部分的让模型具有较强的“记忆能力”，Deep部分的让模型具有“泛化能力”，使模型兼具逻辑回归和深度神经网络的优点，除了能够快速处理记忆模型特征外，更重要的是具有较强的表达能力。

2.5 模型业务应用架构

客户均衡模型涉及多类建模方法和不同外部数据源，在架构建设上需要灵活应用不同的场景架构，并支持高并发高可用的服务场景。架构设计理念包含以下两方面：算法驱动，以人工智能算法为核心，给客户推送有价值、高相关并符合资产均衡模型的产品等。高性能稳定架构具备低时延、高可用性(99.9%)、服务自愈、弹性扩缩容特性，同时包括完备的服务容灾、降级、应急方案。离线服务保证数据低丢失率、样本熔断、数据容错、自动恢复等特性。整体架构如图6所示，包括内容架构、客户架构、召回架构、模型架构、融合架构。

此架构共包含两条数据流，请示数据流和反馈流。1）请示数据流：客户请求精准营销服务，经过一层容灾进入推荐主服务，主服务调用历史服务和客户服务获取客户下发历史、ID映射、客户画像等数据。第二步请求召回服务进行个性化召回，获取客户最可能感兴趣、对客户有价值的内容列表和信息，通过去重服务进行去重。第三步根据内容列表请求模型服务得到模型打分进行排序，利用融合技术强插运营干预内容，最终返回给客户进行展现。最后，主服务记录下发数据，防止刷到重复内容。2）反馈流：反馈流技术是应用资产均衡模型建设精准营销系统的核心，决定精准营销效果上限。客户点击、停留、收藏等反馈信息通过APP埋点技术进行反馈日志上传，在系统后端形成反馈数据。一方面，基于反馈数据进行特征建设，包括内容维度的统计；另一方面，反馈数据和客户请求数据结合，按照不同模型建设目标，形成不同正负样本，进行模型训练，建设基于点展行为的CTR模型和基于时长的多目标模型。

三、业务应用效果

申万宏源的业务应用模式是创新性的，采用线上和线下相结合、人工运营和智能算法相结合的方式。线上场景主要依托智能算法，基于海量数据建模的客户资产均衡模型提供精准营销服务，将产品等精准触达到客户。线下场景主要包括营销人员通过营销平台对客户推荐的产品进行配置和营销。同时，为客户展现的推荐产品列表也是资产均衡模型推荐产品和运营产品结合产生的。最后，收集客户反馈数据，基于客户画像和产品画像等数据进行综合分析形成营销分析报告。营销人员获取数据和分析结果加深对客户的理解，进一步通过线下电话或者微信推广方式联系客户。

目前基于资产均衡模型的智能产品精准营销应用模块已经在申万宏源客户APP的首页和理财商城上线，经过多次策略升级迭代，精准营销产品和内容的平均点击率提升20%以上，产品购买转化率提升近2倍，产品成交金额达到15.1亿。此外，C端业务把精准营销的结果进一步扩展到PC端理财商城、企业微信以及APP资讯推荐等模块。智能内容理解模块目前已经在APP资讯相关模块上线，线上点击率提升约20%。智能精准产品营销应用模块目前共上线场景13个，包括APP端和PC端，不同场景均基于客户资产均衡模型预估结果，在金融产品候选集和个性化推荐策略上会存在差异。

同时重点打造B端亮点业务，例如对流失预警客户进行挽留产品推荐、对资产发生异动的客户进行相应产品推荐。通过资产均衡性建模，把股票投资客户向理财客户进行转化，帮助客户实现资产均衡性配置。同时，通过引入业界前沿人工智能技术带来更多年轻群体客户开户。通过对客户持续深入了解，提升客户服务满意度，为公司的财富管理转型提供强有力的技术支持。

四、总结及未来工作展望

基于深度强化学习的客户资产均衡模型，能在金融行业有效智能化建模客户资产均衡性，综合考虑客户偏好、资产分布、潜在兴趣探索、舆情内容理解等维度，个性化评估客户资产，并提供客户资产配置建议。在技术规划上，客户资产均衡模型未来将进行更精准建模，提升客户资产均衡性评估精度，进而提升业务应用效果：第一，采用DDPG升级DQN，由于DQN仅支持离散Action，而目标融合是一个连续的过程，因此后续会升级用DDPG这种更复杂的深度强化学习模型来提升融合精度；第二，采用DNN升级GBDT，随着对业务理解的持续加深采用非线性表达能力更强的DNN深度网络进行客户偏好学习；第三，采用分布学习替代KL-Calibration，即不同客户的理想资产分布是有差异的，升级为采用分布学习进行自适应学习。

云原生GPU虚拟化在证券投行业务的创新实践

谢杨军、王一帆 | 国泰君安证券股份有限公司 | Email：xieyangjun@gtjas.com

摘　要： 随着越来越多的新型容器化应用，例如高性能、深度学习应用，开始依靠GPU，有效支撑GPU在容器云中变得至关重要。同时也更好地支持AI/ML业务的快速发展，虽然 GPU 虚拟化已经广泛，特别是在VM进行了较深研究，而在容器做了有限的工作，现有很多场景停留在使用单一特定的 GPU 虚拟化技术来部署容器，例如GPU直通或API转发，不仅缺乏远程GPU虚拟化优化。且没有释放 GPU 的全部能力。因此本文实践方案采用云原生容器云GPU虚拟化，实现多业务共享，充分利用GPU算力、显存等资源，在资源调度层面可以提供渲染混布和编解码混布，实现了AI 算力、渲染算力、编解码器等 GPU 全部资源的统一调度。

关键词： GPU虚拟化；云原生；容器；投行

一、GPU虚拟化技术及行业该领域浅谈

1.1 GPU共享技术浅谈

GPU虚拟化与OS虚拟化是完全正交，通常我们指的虚拟化为，CPU 虚拟化由 VT-x/SVM 解决，内存虚拟化由 EPT/NPT 解决，而GPU属于PCIe设备，那么GPU虚拟化即是PCIe设备虚拟化。GPU虚拟方案很多：

PCIe 设备直通。虚机等完整使用这个 PCIe 设备，物理机上一样。这种方案，我们称之为 PCIe 直通（PCIe Pass-Through）。很显然，PCIe 直通只能支持 1:1 的场景，无法满足 1:N 的需求，其实并不能算真正的虚拟化，也没有超卖的可能性。

SR-IOV: revisited，要实现 SR-IOV，需要满足两点：硬件资源要容易partition和无状态（至少要接近无状态）。但这类技术适合于网卡等设备，因GPU硬件复杂度极高，远远超出网卡和NVMe设备，GPU 的 SR-IOV 利用封装了 PCIe TLP 层的 VF 路由标识以及类似MPF技术来实现虚拟化。

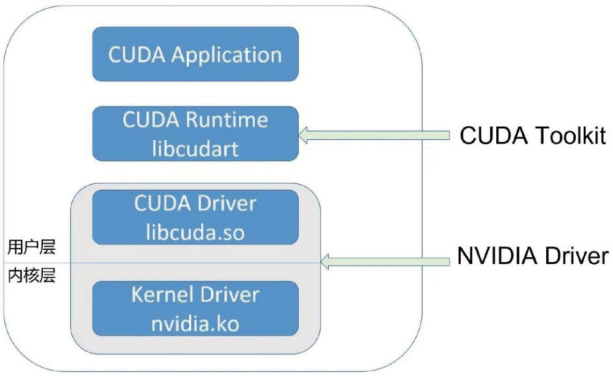


图1： CUDA与NVIDIA driver关系

API 转发，API 层的 GPU 虚拟化是目前业界应用最广泛的 GPU 虚拟化方案。它可以屏蔽底层硬件，灵活，但也面临着其复杂结构带来的弊端。

MPT/MDEV/vGPU，这种方案本质是PCIe的虚拟化，典型就是nVidia GRID vGPU、Intel GVT-g方案，而且必须有一个 pGPU 驱动，负责 vGPU 的模拟和调度工作。驱动也是强依赖不同厂家。

本文讲述的容器GPU虚拟化都是基于NVIDIA 生产的GPU、且只考虑 CUDA 计算场景，GPU天然适合向量计算。加速AI/ML场景落地。

如图用户是完全有可能绕过中间的限制，去直接触达原生CUDA的，进而霸占整个物理GPU。很多业内技术已实现“静态”与“动态”两种分配形式分配，静态很好理解；所谓“动态”，就是在这个容器的生命周期当中，虚拟GPU是可以被调节大小与数量的，一旦任务运行结束（CUDA指令发送完成），那么被占用的虚拟GPU资源立即释放，动态很适合云原生场景下，体现敏捷、灵活等特点。

1.2 GPU虚拟化技术行业浅析

GPU虚拟化近些年在各行业均有发展，GPU并行计算技术以其强大的计算能力，迅速在证券行业的多个领域推广落地，特别是在量化策略交易、人工智能等应用。

阿里的cGPU（container GPU）[1]是最早提出的通过内核劫持来实现容器级GPU共享的方案，向社区贡献了

device plugin[2]和调度器[3]。，通过限制每个容器可下发kernel的时间片来隔离算力资源，2017年收入使用阿里GPUshare技术。但仅仅从显存切片，从而提高整卡利用率。

腾讯提供了两种GPU共享方案。基于CUDA劫持，提供了一套GPU共享解决方案GaiaGPU[4]，GaiaGPU也提供了Device plugin GPU manager[5]和调度器 GPU admission[6]。基于内核劫持，提供了资源隔离方案qGPU[7]方案。

百度的MPS+CUDA Hook的GPU隔离方案也是一种CUDA劫持显存，使用MPS隔离算力。爱奇艺也适用CUDA劫持显存，但其算力使用了空分限制。AWS aws-virtual-gpu使用了Tensorflow框架隔离显存，MPS隔离算力，但仅仅用于Tensorflow领域。 OrionX有别与以上几家方案，最为出色特点为提供GPU池化类解决方案，而且对Device plugin和调度器均有改动。

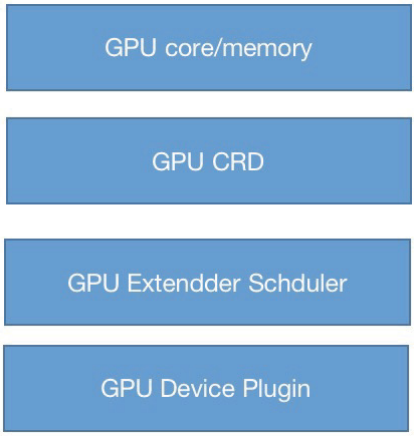


图2： GPU虚拟化

1.3 落地实践经验

在公司容器云GPU虚拟化落地实践前需要具备一定前置条件：使用NVIDIA vGPU功能需要提前准备以下工作：

(1) 确保物理GPU型号支持虚拟化切割。

(2) 确保物理机BIOS中开启Intel VT-d / AMD IOMMU功能，且物理机内核已开启IOMMU支持。

(3) 确保物理机BIOS中开启SR-IOV和Memory Mapped I/O above 4GB功能。

(4) 确保已获取物理GPU驱动和vGPU驱动。相关网卡驱动以及安装方法请联系网卡提供厂商获取帮助。

(5) NVIDIA vGPU功能与操作系统内核存在依赖关系，确保物理机使用h76c、h79c及更高版本ISO。

(6) NVIDIA A系列vGPU需额外确保物理机操作系统Kernel版本升级至4.18，GCC版本升级至8.3.1。

(7)PCI设备热插拔开关：用于设置是否允许云主机热插拔GPU设备，默认为true。若热插拔时出现硬件兼容性错误，或不支持该硬件设备时，建议关闭此功能（设置为false）。

二、投行业务云原生GPU共享实践

2.1 投行智能刷报

随着债券业务的市场不断扩大，监管提出了更严格、更全面、更深入、更精准的信息披露要求，在尽职调查、价值识别、持续跟踪、监管沟通等各方面对债承业务部门都提出了许多务实的要求。监管强调压实中介机构的主体责任，在此背景下，急需快速、高效、专业的项目质量和风控体系建设。时至今日，人工智能等新兴技术为代表的新一轮金融科技革命驱动着投行业务开启新一轮的数字化与数智化变革浪潮，头部券商及部分银行、基金等机构均已早早布局智能撰写系统。利用人工智能技术来对投行文档进行智能撰写、刷数，不仅提升了投行部门的工作效率、解放人力，还有效提高了对外公开信息披露文件的质量，通过技术赋能投行业务，提高执业质量的同时也降低合规的风险。

智能刷报撰写系统作为智能投行建设规划的一部分，主要用于债券业务及财务类数据刷新。国内债券业务排名靠前的大券商投行部门都已采用，极大提升了投

```
containers:
- env:
  - name: GPU_DEVICES
    value: auto
  - name: CUDA_VISIBLE_DEVICES
    value: "0"
  - name: TZ
    value: Asia/Shanghai
  image: xxxx/yyyy/segment:release
  imagePullPolicy: IfNotPresent
  name: ibdatagrand-segment
  ports:
  - containerPort: 8097
    name: tcp-8097
    protocol: TCP
  - containerPort: 3764
    name: tcp-3764
    protocol: TCP
  resources:
    limits:
      aliyun.com/gpu-mem: "6"
      cpu: "6"
      memory: 8Gi
    requests:
      cpu: "6"
      memory: 8Gi
```

图3： 云原生业务GPU虚拟化部署yaml样例

行债券文档的处理效率和信息披露质量，防范了合规风险。智能刷报撰写系统，旨在加快推进公司投行业务转型升级，打造智能化、数字化投行工作平台。运用行业领先的人工智能、自然语言处理、深度学习等技术，搭建一套投行智能撰写系统。以技术赋能、解放人力，系统性提升投行业务效能，加强尽调手段与能力，提高信息披露材料质量，降低合规风险。

业务系统中每一个微服务模块都是可以冗余部署的，全栈容器化形式部署，支持各模块支持水平扩展。系统所需的硬件配置与用户对性能与高可用等需求相关，各个微服务模块都需要 CPU 算力与 GPU 算力，应用层与AI层会打包到容器镜像中。

2.2 投行银行流水核查

随着科技的发展和数字化转型进程的加速，银行流水核查也在不断地发生着变化。传统的银行流水核查方式主要是人工审核，费时费力，且容易出现疏漏。而现在，越来越多的金融机构开始采用自动化银行流水核查技术，如人工智能、大数据分析等，以提高效率和准确性。这些技术的应用使得银行流水核查更加快速、高效、精准，同时也降低了人为因素所带来的风险和不确定性。

银行流水核查对于证券行业投行业务的重要性不言自明。它不仅能够帮助投行更好地了解客户的资金流动情况，还可以帮助客户更好地管理自己的财务状况。同时，银行流水核查也可以提高投行的信用评级和风险控制能力，为金融市场的稳定和发展做出贡献。因此，银行流水核查在证券行业投行业务中的作用和影响是不可忽视的。利用云原生容器云GPU虚拟化技术，不仅提高了GPU利用率，同时能在短时间内完成大量银行流水的核查工作，并且可以实现自动化处理。这不仅提高了效率和准确度，还降低了人工成本，最终为投行部门提供了更加高效、准确、安全的金融服务，同时也促进了证券行业的规范化和标准化。

```
containers:
  - name: grpc-pdf-io
    image: xxxxx/yyyyy/pdfinsight:1.3.74
    ports:
      - name: grpc-pdfio-port
        containerPort: 8000
        protocol: TCP
    env:
      - name: MODE
        value: nio
    resources:
      requests:
        cpu: "4"
        memory: 8Gi
      limits:
        cpu: "8"
        memory: 16Gi
        tencent.com/vcuda-core: "100"
        tencent.com/vcuda-memory: "60"
```

图4：云原生业务GPU虚拟化部署yaml样例

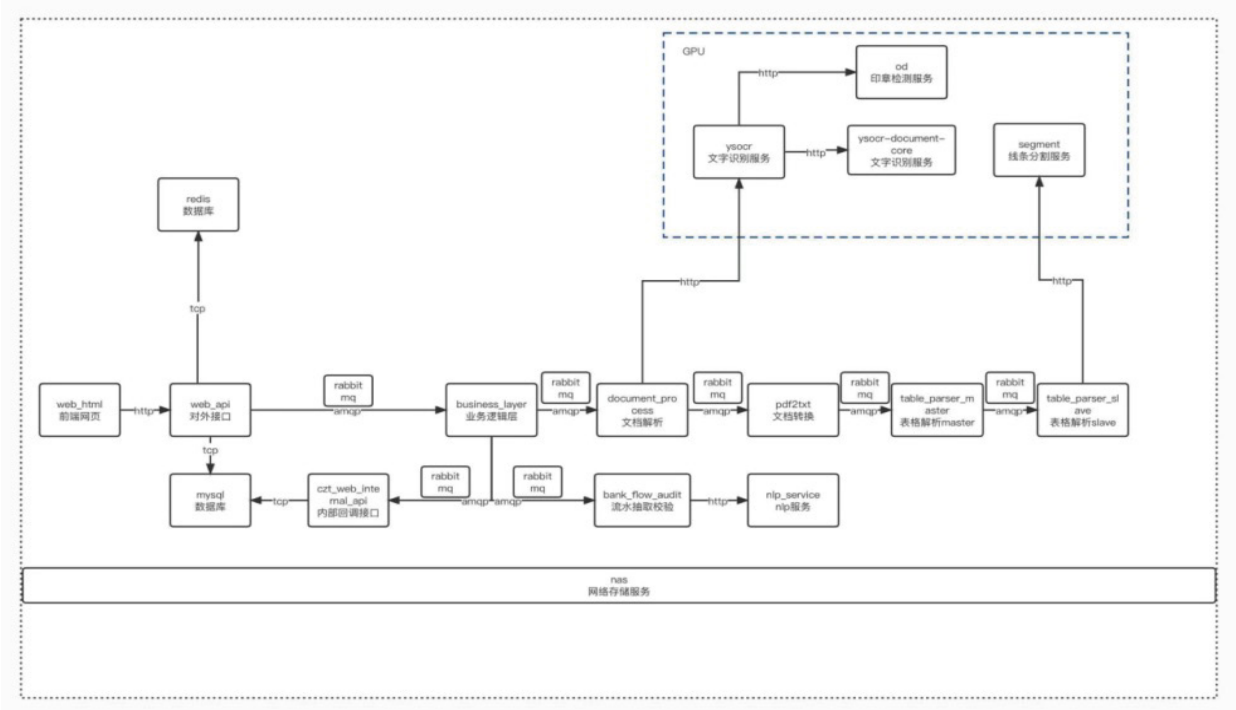


图5：投行业务逻辑架构

三、结论与展望

目前GPU共享已经逐渐开始进入工业落地的阶段了，云原生GPU虚拟化是未来云原生中台能力模块，在解决解决GPU池化能力输出，同时提供其性能，而GPU内存、带宽和延迟等豆影响着整体性能，未来可以尝试从以下方向提高性能：

(1) 能够提供稳定的服务，运维便捷。比如MPS的错误影响是不能被接受的，另外对于带有预测的实现，也需要更高的稳定性。共享工作负载尽量降低。

(2) 更低的JCT时延，最好具有保障部分任务QoS的能力。对于一个已有的GPU集群进行改造时，需要尽量减少对已有的用户和任务的影响。

(3) 不打扰用户，尽量不对用户的代码和框架等做修改，另外也需要考虑框架和其他库的更新问题。

参考文献：

[1]https://www.alibabacloud.com/help/zh/elastic-gpu-service/latest/use-docker-to-install-and-use-cgpu

[2]https://github.com/AliyunContainerService/gpushare-device-plugin

[3]https://github.com/AliyunContainerService/gpushare-scheduler-extender

[4]https://github.com/librabyte/paper-reading/issues/1

[5]https://github.com/tkestack/gpu-manager

[6]https://github.com/tkestack/gpu-admission

[7]https://cloud.tencent.com/developer/article/1831090

东方证券IT研发运营标准化探索与实践

李晨辉、赵泽、王国喜 | 东方证券股份有限公司 | E-mail：zhaoze@orientsec.com.cn

摘　要： 数字化转型对于业务运营和精益管理的优化作用日益凸显，已成为企业降本增效的核心驱动力。在这场浪潮中，金融科技却因其信息化起步早、业务场景多元化、监管隔离要求严等原因，使得数字化转型之路尤为复杂和尴尬。多年来券商业务模式不断革新、技术持续演进，IT部门存量技术架构杂、数据孤岛多等现实因素严重制约了IT项目标准化管理有效实施，历史工作惯性已成为数字化建设的沉重包袱。东方证券研发总部在“一窗口，两平台”的IT数字化系统建设基础上，落地IT标准化管理、实现项目生命周期流程一体化，向公司及业务部门输出需求跟踪和项目数字运营的可视化能力，取得了较好反响。

关键词： 标准化；流程一体化；数字化；DevOps

一、背景

IT研发运营标准化建设的目标是在安全合规的前提下，以获得IT组织资源最优分配、最佳效率为目的，通过项目管理规范的制订、管理系统建设、管理流程一体化等一系列措施，建立统一IT管理标准的实践过程。知易行难，建立一个可以切实落地的，面向全公司的、适合于各条业务线的标准管理是充满难度和挑战的：

首先是规范和效率的博弈。规范的建立会利于组织的效率提升，但流程的细致化可能会影响个体成员的效率。证券行业线上化服务自90年代开始提供并延续至今，IT各团队与各业务部门间已形成了多种协作模式。推动统一的IT项目规范建设过程中，无论是需求的细化评估，还是交付规范化等要求，都不可避免的破坏原有交付习惯，让业务方产生麻烦、不灵活的感受。

其次是组织建设与职责范围的明确划分。项目生命周期自需求评估与商务采购开始，经过研发、测试、部署，到进入监控运营为止，涉及合规、商务、研发、运行多部门多岗位协同工作。证券IT标准化建设的任务不单单是将原以邮件和OA审批为主的协作手段进行线上化改造，更重要的是通过切分工作流，建立各环节权责清晰的标准化流程，把合规风控和管理要求落实在一体化流程中，并推动完成新模式落地。

第三是团队和技术的复杂性。证券行业信息化起步较早，早期以运维为主的环境下往往忽视基础工具的统一和技术更新，存量项目中有复杂的历史架构、陈旧的技术栈和各种烟囱式的定制化工具。想要交付过程规范化，就要先从统一基础工具链建设入手，先具备开放性去满足不同团队、不同项目的技术需求，再通过统一的技术管理要求，作为度量和标准化的抓手。

尽管研发运营标准化工作面临艰巨的挑战，但缺乏标准框架的IT项目管理会因为流程混乱造成沟通不畅、权责不明、质量不稳定等问题。交付团队各行其是的后果是IT部门既无法打通交付上下游流程，更因缺乏一致性的衡量数据而难以定义团队效率。在业务需求和技术人员日益增长的今天，IT部门无法通过项目数据驱动管理决策就会承受混乱带来的成本增加。因此，东方证券研发总部力求通过落地研发运营标准化推动项目技术能力和管理能力的数字化升级，实现部门降本增效，从而更好地去满足各方需求，帮助公司快速发展。

二、顶层设计与实践方案

2.1 顶层设计

标准化体系设计首先要确定范围、内容和方法，保证可操作性和成果的可衡量性。因此，实践设计以解决现实问题为入手点，着力解决当前项目交付过程中IT外部、IT管理、IT内部三个主要矛盾：

1）IT外部矛盾来源于业务部门与IT部门的认知差异。一方面是前台与后台部门间对于成本与收益的权衡角度不同，另一方面技术壁垒与业务复杂性增加了部门间沟通难度。双方对需求价值和技术工作量间的认识差异更会随着项目进度情况的不透明而逐步放大。

2）IT管理矛盾在于行政管理和技术管理的关注点不同。行政管理重视整体，为所有项目提供稳定的运营环境和合规保障，关注部门的整体运营、资源分配；技术管理重视细节，关注项目开发任务的完成度，强调技术架构的合理性。当前双方仅交汇于简单的OA更新上线流程，会让管理人员无法看到项目过程中的问题和风险，失去早期洞察和介入调整的能力。

3）IT内部矛盾在于研发与运维的诉求不同。运维团队

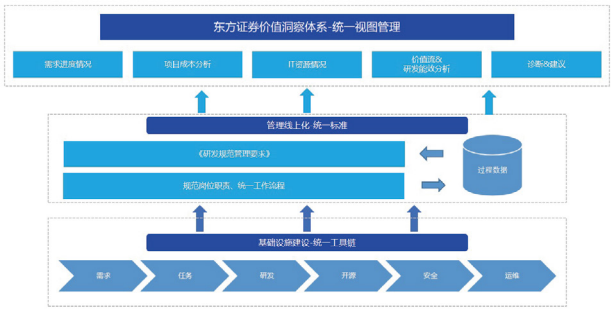


图1：顶层设计

需要更多的标准化和确定性，如功能、非功能需求，配置管理、故障响应计划等来保障系统的安全性和可维护性。而研发团队要不断的面对新需求的拆解与功能开发，更加关注交付效率。缺乏自动化工具和流程混乱，会在交接时产生大量文档对接和人工操作，随着迭代频率的增加，目标不同会让双方产生对立情绪。

基于上述分析，研发运营标准化设计如图1所示：整个体系应具备工作量化和进度可视化能力，通过各项目价值流数据改善IT与业务部门沟通问题；拥有标准化和高效的协作流程，兼顾可规模性管理的项目共性和下钻到具体项目的技术特性，以全局化的管理视角解决IT内部管理问题；通过建设自动化的工具和工作流水线减少重复劳动，提高效率和安全性，解决研发、运维间协作问题。所以，研发运营标准化体系要满足自动化、标准化、可视化三个核心要求，通过流程一体化设计搭建既统一又灵活，可以在实践中快速反馈及不断迭代的体系。

2.2 实践方案

研发运营标准化建设方案是“以统一平台为承载，以管理规范为指导，以流程一体化为抓手”。如图2所示，管理规范是外部监管与部门管理的集中体现，为流程建设制定规则。流程一体化则通过重新整合岗位权责，以流程推动管理规范落地，两者相辅相成。平台建设是一切的基础，按业务功能划分成“一窗口，两平台”统一向内、向外提供IT全流程管理和自动化服务。

图2：建设方案

标准化体系落地会涉及组织、流程等多方变革，让所有项目遵循统一规范是一个长期的过程。因此，建设方案遵循两个原则：

1.“统一开放”。所有项目和岗位应使用统一平台，在流程交接节点执行统一标准，以质量门禁形式保证提交数据、信息一致性。对节点以外的工作形式暂时开放用以兼容历史存续，提高灵活性。

2.“持续反馈”。在实践中解决历史遗留问题，逐步规范非标操作，根据现实状况不断增加规范化管理程度。

三、 技术实现

在“统一开放，持续反馈”的原则下，研发运营标准化体系的整体架构设计采用服务化的形式。对已有的研发运维工具和历史应用进行整合，将其下沉为服务后封装到所属平台。平台间通过流程传递信息和数据，该架构既兼容了历史工具和工作方式，又可以在未来对工具进行灵活替换，避免了大量硬性切换系统造成的震荡和浪费。

3.1 平台服务架构

如图3所示，完整系统以三个平台间松耦合形式组成，需求平台作为对外窗口，接受和评估业务需求，并对业务方提供需求交付状态；IT数字化平台作为项目规划管理平台，从项目里程碑维度对商务、人力、运营等方面执行行政管理；研发运行一体化平台作为交付管理平台，将项目管理能力下钻到研发、测试、运维监控等细节。平台间遵循使用统一的命名规则、数据字典、项目系统表等要求，交互时以项目和版本两个维度为锚点进行关联。统一的流程引擎、待办管理、Jira等工具支撑平台间以流程形式完成任务交接、审批等工作，实现系统的流程一体化。

图3：服务架构

3.1.1 需求管理平台

该平台以业务部门需求为关注点，对外提供需求发起入口，支持公司级部门间的审批要求，对内增加需求分

析、需求拆解等环节，以开发任务维度串连IT管理和开发数据。平台提供数据统计功能，按业务部门的角度生成统计报表，实现需求成本，工作量和交付过程数据可查询，结束交付过程黑盒子的情况，极大增强IT工作可视化能力，是部门间沟通和协作的重要窗口。

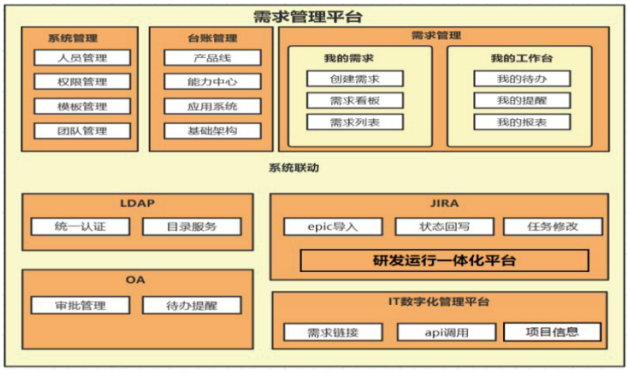


图4：需求平台架构

3.1.2 IT数字化平台

IT数字化平台作为部门管理维度的数字化系统，负责项目里程碑级别管理和IT部门各维度的规划管理工作。

商务管理：预算填报、采购申请、招投标、公司级合同、合同验收、合同支付。

项目管理：项目前评估、立项、更新/上线、项目验收、项目后运营。

人力管理：实习生/外包员工的入职、离职、分摊、系统交接。

部门管理：系统管理、奖项管理、数字化驾驶舱、软件管理。

图5：IT数字化平台架构图

3.1.3 研发运行一体化平台

该平台从技术维度上管理从需求到生产发布的整个交付过程。平台以流水线形式实现持续集成和持续交付的职能，将交付标准化通过平台的流程管控落地，并以流水线形式提供给研发、测试、运维等不同角色所需的自动化能力。架构如图6所示，研发运行一体化平台以三层架构设计分离工具、功能与操作，通过版本管理流程贯通项目交付全流程的数据链路为目的，通过提测单、变更单等交付节点链接软件开发生命周期中涉及的多个阶段和不同岗位工作[1]。

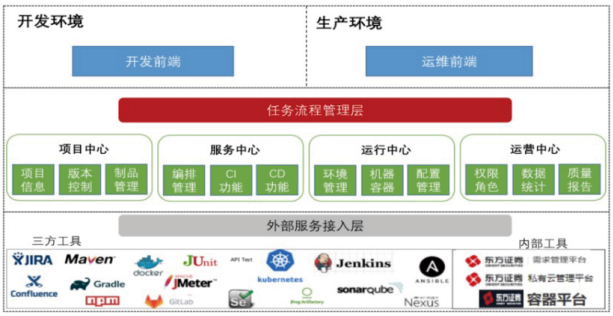


图6：研发运行一体化平台架构图

3.2 技术架构

整个体系在技术实现上均以云原生架构为主，采用微服务形式保障工具和功能可以灵活调整变化的需求。以服务治理框架构维护服务调用关系及拓扑结构，提供服务注册发现、负载均衡、黑白名单等功能。服务均以镜像形式通过研发运行一体化平台部署于公司信创容器平台/云管平台，提供服务的敏捷迭代、动态伸缩等能力。

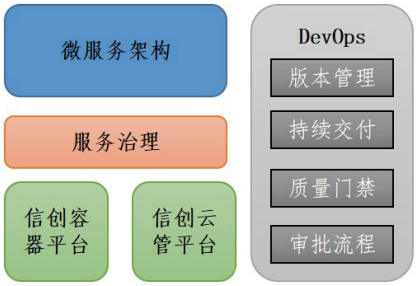


图7：云原生架构

标准化体系自身服务也使用研发运行一体化平台管理交付，如图8所示，该平台基于DevOps方法论，整合研发运维工具链，提供一站式的自动化工具管理和完整的交付审批流程。

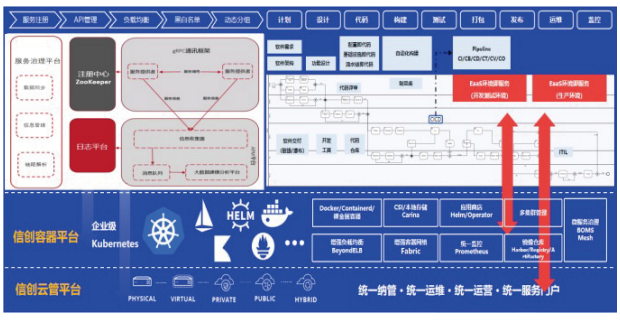


图8：技术架构

四、实践成果

实践至今，部门通过统一平台和流程一体化建设逐步实现项目管理规范化目标。如图9所示，三个平台各自聚焦其业务重点，通过流程一体化实现平台间数据交互，覆盖项目由需求立项到交付全周期，从而获得落地标准化和全局可视化能力。进一步的，为实现数字化运营，抽象项目前、项目中、项目后三个环节作为数据的交汇点，用以追踪需求、项目过程数据，以数据辅助部门项目管理。

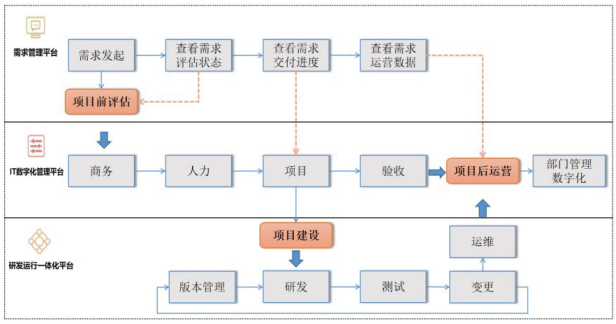


图9：数据与业务流

4.1 管理规范

管理为先，建设为要，研发运营标准化体系以部门管理要求为指导。通过《软件项目管理规范》制订项目管理要求，确保软件项目管理工作流程具有一致性。对各环节准入准出工作的内容、成果、文档格式等进行明确，规范了软件项目实施工作中的管理流程，从质量管理的宏观角度对软件项目生命周期进行节点把控。

图10：项目生命周期规范管理视图

技术方面以《架构决策》建设体系化及分层次的架构评审机制，对项目技术向如数据、工具、中间件等方面分别实行立项前评审和上线前评估。通过架构组织完成企业架构的标准化及规范化落地工作，掌控架构、系统细节，具备遵循架构标准进行研发的能力 [2] 。

东方证券 IT 架构决策

架构决策	服务治理平台接入规范	编号	TA. [illegible]
提出方	[illegible]	提出日期	2[illegible]
架构领域	TA：技术架构	主题	各系统须按照规范接入服务治理平台
决策人员	架构委员会	决策时间	2[illegible]
问题	大量的业务及支撑系统开始上线运营并提供服务，目前公司内部系统间接口调用无统一标准，异构化非常严重，各厂商都有各自私有协议，且存在有 SPX、T2、Web Service、REST、TCP 等各类型异构接口，无法以全局视角对内部服务进行统一治理，进一步增加了系统开发运维优化的难度。		
假设及约束	1、所有新建系统需接入服务治理平台，旧有系统逐步改造； 2、应用通过服务治理框架完成服务注册和发现，通过服务治理平台进行服务治理治		

图11：架构决策示例

4.2流程一体化

流程一体化是指通过重新梳理业务流程、审批流程、数据传输流程等，打造从业务的最前端完整延伸到业务结束的全周期流程，是部门规范管理的应用和体现。研发运营标准化体系通过需求、交付、项目三条业务流设计，以流程覆盖项目管理全生命周期，以自动化工具获取数据，又通过流程要求推动项目管理规范落地。

4.2.1需求全周期管理

需求管理以需求接受-评估-拆分-跟踪-验收等环节组成完整链路。在流程中遵循证券行业“合规先行，保持风险意识的原则”，需求接受和上线审批流要经过相关部门以确保符合监管要求。在工具上提供看板、版本跟踪等功能实时展示需求状态、工作量、排期等信息。在数据上记录需求评估的价值、工作量、实际交付等信息，作为项目交付后运营和年终复盘的数据依据。

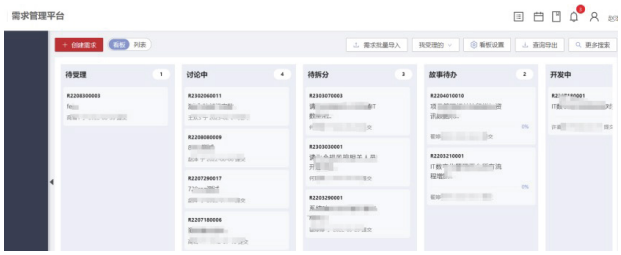


图12：需求接受及处理

图13：需求立项

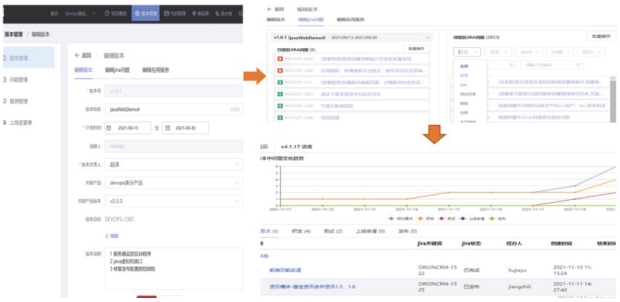


图14：需求交付状态可视化

图17：制品管理

4.2.2 交付全周期管理

项目交付作为IT部门的核心工作，涉及技术、人员、硬件、软件多方面工作，包含复杂的细节。同时，证券行业上到机房建设、网络安全，下到系统更新、配置更改都有严格的合规风控要求。因此，交付管理不但要提供自动化工具满足技术要求，更要通过流程设计保证工作按照既定的管理要求执行，从而实现项目交付过程的完整留痕和有效合规。

1）工具自动化

研发运行一体化平台通过纳管工具链，将散落独立的工具重新组织，以流水线形式提供交付过程中所需的功能。通过唯一可信源和制品晋级制度，管理多网络环境下多机器集群的部署和运维白屏操作。在提高效率的同时，通过逐步引导交付团队使用统一工具来标准化交付过程数据。

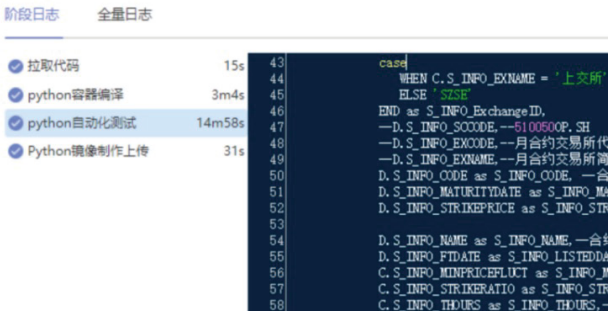


图15：可配置的自动化功能

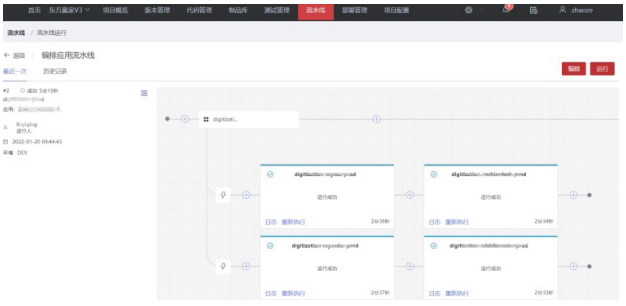


图16：自动化流水线

图18：多环境一键CICD

2）流程管理

交付管理以需求-版本任务-提测单-变更单-管理审批-生产部署等流程建立自动化交付全流程，串连研发、测试、质控、运维等岗位交接工作。通过版本管理、质量门禁、制品管理晋级制度等保障合规等管理标准化监督落地，也为工作量化提供坚实的基础。

图19：版本管理

图20：提测管理

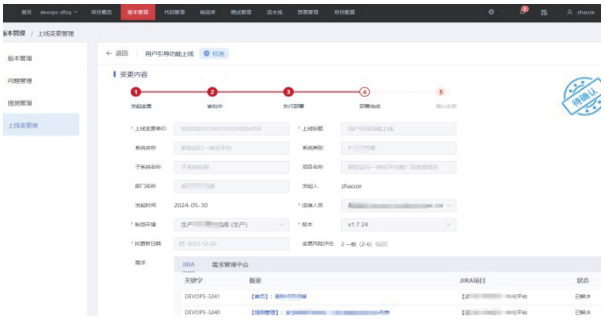


图21：变更管理

4.2.3 项目全周期管理

项目管理以立项-评审-商务-质控-上线更新-验收等流程，流程上在项目里程碑维度进行行政管理，通过流程审批要求落实管理工作。工具上结合部门商务、人力等数据建立项目前评估-项目中跟踪-项目后运营的数字模型，可通过项目流程下钻查看历次交付数据细节，强化部门数字化和可视化能力。

图22：项目里程碑管理

图23：商务流程

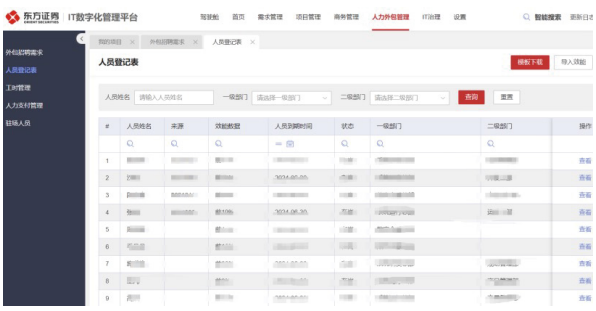


图24：人力资源池

图25：项目交付审批记录

图26：更新审批质量门禁

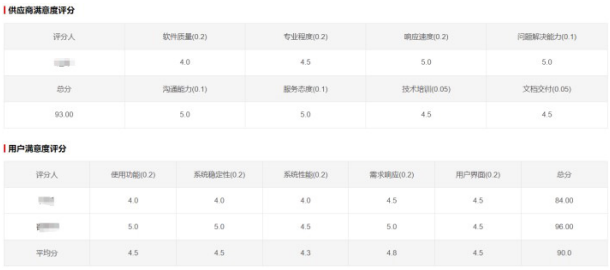


图27：项目后满意度调查

4.3 数字化运营

随着研发运营标准化建设的推进，项目管理逐步规范化。不同项目在需求-评估-执行-交付；立项-评审-更新-验收；研发-测试-质控-运维等多个维度的数据有较高程度的完整性和一致性，基本具备通过数字描述项目过程和状态的条件。当前部门已将成本、价值、质量、交付效率、业务需求历史成功率等数据用于项目前评估和项目后运营，作为部门盘点和效能的参考依据。未来计划建设数据中台汇聚形成部门运营的数据统计。逐步通过提高数据质量和展示能力实现以价值为导向可视化洞悉问题，从而改善局部环节，优化整体管理。

图28：团队效能数据测试环境

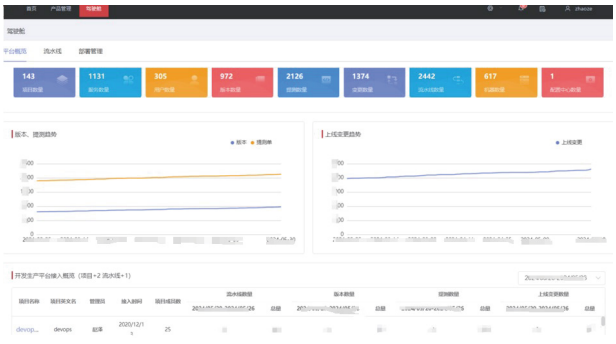


图29：项目交付数字化

图30：部门管理数字化运营测试环境

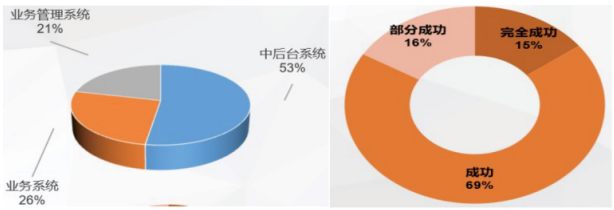


图3： 项目后运营测试环境图1

图32：项目后运营测试环境图2

五、总结及展望

东方证券系统研发总部经过多年的规范建设，通过对部门管理的规划和内部流程的重新设计成功的将零散的IT业务串联起来。以工具自动化提高技术效率，以制度标准化规范IT管理行为，基于一体化流程落地研发运营的标准化管理，初步搭建了IT数字化管理体系。在未来，会进行以下几个领域的进一步工作：

1.以数据驱动决策。利用已经建立的数字化研发运营体系，进一步挖掘和分析数据，通过数据驱动，发现IT部门的交付瓶颈，从而有针对性的改进

2.持续流程优化创新。一方面当前一体化流程为了兼容历史遗留问题，随着这些问题的不断标准化可以进一步完善流程。另一方面随着业务和技术环境的不断变化，也要定期审视和更新流程。

3.拓展数字化应用场景。结合我司已有的“东方大脑”、“RPA”、“繁微”等智能服务、数据分析平台，进一步拓展数字化体系的功能和价值。

东方证券将继续秉承开放共享、合作共赢为原则，以金融科技规划为引领，通过IT数字化管理转型驱动创新，以科技为业务赋能，不断助力公司数字化转型与行业创新发展！

参考文献：

[1]樊建、赵泽.《东方证券研发运行一体化平台探索与实践》.交易技术前沿，Vol.50,2022.12

[2]樊建、严伟富.《企业级证券业务中台探索与实践》.交易技术前沿，Vol.49,2022.09

业务流程治理体系探索及实践

徐鑫鑫、陈心亮、李军林 | 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： 业务流程管理（BPM）系统在中国结算上海分公司具有多年发展历程，承载着登记结算业务的办理入口和业务流转职能，是业务与技术融合的典型应用。随着公司数字化转型进程的不断加速，中国结算上海分公司紧随行业发展趋势，积极开展新一代BPM系统建设，围绕BPM系统探索建立符合司情的流程治理体系，在业务实践中加强经验总结，将数字化、智能化理念融入日常业务流程中，持续加强技术创新引领、激活数字化经营动能，不断提高组织效率，提升运营质量。

关键词： 流程治理；标准建设；数字化；智能化

一、研究背景及意义

1.1 数字化转型要求

2023年中央金融工作会议指出，要做好科技金融、绿色金融、普惠金融、养老金融、数字金融五篇大文章，推动我国金融高质量发展。数字金融是现阶段金融通过数字化转型发展起来的新金融业态，是金融在社会、经济、科技发展潮流下的大趋势。中国证监会组织相关单位编制了《证券期货业“十四五”科技发展规划》，阐明了“十四五”时期证券期货业科技监管工作和行业数字化转型的指导思想、工作原则。在行业数字化从多点突破迈入深化发展新形式下，作为资本市场重要基础设施，我司积极拥抱数字化变革浪潮，驱动业务转型升级，是时代发展的必然要求。

业务流程管理（BPM）系统是职工业务办理的窗口和工作站，处在用户感知的第一线。运行于BPM系统上的各类业务流程，从最初的线上化、电子化逐步向数字化、智能化发展，推动业务办理模式不断优化升级，伴随着公司数字化转型的深入开展，如何进一步开展业务流程改进优化，更好的支持新一代登记结算系统建设，是在工作实践中值得深入研究的重要课题。

1.2流程治理的意义

业务流程治理，是一种包含理论方法、演进规则、评价手段、使用工具和管理实践的结构性方法，确保企业建立起长期有效的流程管理规则和运营机制，保证流程体系的有效执行和持续改进，从而实现企业战略目标，提升业务能力。流程治理可以提高组织的效率和质量，降低成本和风险，加强组织内部的沟通和协作效率。

中国结算上海分公司BPM系统的建设工作于2009年开展实施，累计部署200多个流程，基本覆盖所有操作类业务，有效提升了业务办理的电子化水平和安全性，成为公司业务办理中不可或缺的一部分。但BPM1.0主要实现业务流程的线上化改造，存在流程环节多、流转时间长、办理效率低等问题，重平台、轻标准，流程优化和流程管理手段较为缺失。中国结算上海分公司于2020年启动新一代BPM系统（BPM2.0）研究和建设工作，结合RPA（Robotic Process Automation，机器人流程自动化）、OCR（Optical Character Recognition，字符文本识别）等相关技术，将数字化、智能化理念融入日常业务办理过程，逐步探索建立具有登记结算特色的流程治理体系，以期实现对流程改进和治理能力全方位、系统性重塑，推动业务场景中技术、数据和管理有机协同。

二、流程办理模式发展变迁

中国结算上海分公司业务流程的范围包括登记、回售类等发行人相关业务，结算、账户类等结算参与人相关业务，开户、过户类等投资者相关业务，以及内部财务、管理类等业务，业务办理模式正逐步由电子化向数字化、智能化方向转变。

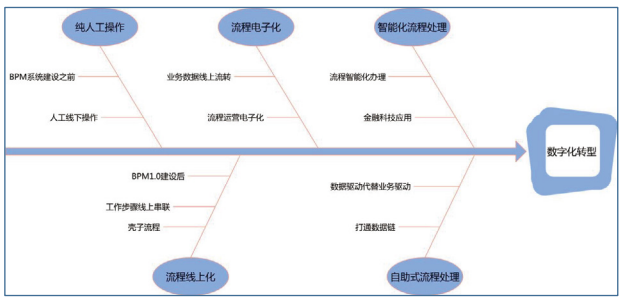


图1：中国结算上海分公司流程办理模式发展变迁

在BPM系统建设之前，业务办理的每一步都需要由人工进行操作，业务材料以纸质材料形式进行传递，以完

成具体单一的某一任务为目标，业务办理效率低、风险高，管理不便。BPM1.0系统的建立，为业务流程线上化打下了基础，流程办理的工作步骤设为独立环节，多个环节串联形成了线上流程，初步实现流转记录和业务统计，解决了流转和时效性问题。伴随着流程线上化工作的深入开展以及BPM1.0系统的不断完善，实现了流程设计、流程建设、流程推行、流程运营全线电子化，目前分公司大部分BPM流程处于电子化阶段，这种模式能够达到业务办理闭环，缺点是办理环节多，对平台模型依赖较重。近几年，分公司流程处理模式逐步开发出了自助式办理和智能化办理等新的形态：自助式流程办理以数据驱动代替业务驱动，自动完成委托受理并实时反馈查看，给予业务申报人更大的自主权；智能化办理是将RPA、OCR等新兴金融科技技术融入流程办理过程，简化用户操作，为业务流程办理模式注入智慧基因。

三、流程治理体系初探

业务规模扩大化、业务规则复杂化、业务数据多元化不断考验着技术与业务的融合能力。单纯的流程由线下搬至线上已无法满足发展需要，围绕BPM系统建立流程治理体系是数字化转型的必然要求。中国结算上海分公司流程治理体系，是在实际工作中逐步提炼深化，以业务需求为牵引，以流程标准化建设+平台技术支撑+金融科技赋能+数字化生态建设为主要内容，锚定“三提两强一防”目标，打造技术引领、业务-数据双驱动运行体系。

3.1 流程标准化建设

流程的标准化，是流程治理体系建设的基础，有了流程标准化，才能更好的实现业务的规范化和自动化管理，更好的助力流程优化反馈和改进。

流程标准化建设，实质上是建立一套业务场景流程化的方法论，以统一的标准、统一的要求实现对各类流程的规范、改进和提升，从而获得更高效更优质的运营效果。中国结算上海分公司结合多年实践经验，建立起流程治理“五步法”，按照流程识别、流程分类、流程设计、流程实施、流程改进步骤对业务场景进行锁定分析，实现场景流程化全生命周期管理，确保流程所经过的每个管理过程都按照标准化要求执行。

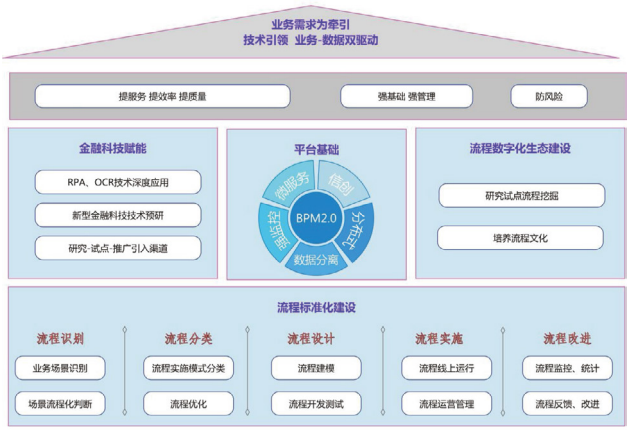


图2：BPM流程治理体系架构图

流程识别，用于确定各种业务流程，并对其进行起底评估。目前流程识别主要有两个来源，一是业务主办部门制定和发布业务流程，根据发布的规章转化为电子流程；二是在日常工作发现的操作制约点和瓶颈，通过试点流程挖掘，确认是否需要进行流程化处理。对于不适用流程办理的业务，建议直接实施系统改造对接，适用流程化改造的业务场景，还要进一步进行流程分类。

流程分类，用于确定流程实施对应的模式，根据流程分类实施标准，对于不同的业务场景采用不同的设计模式：对操作步骤较少、操作环节明确、风险性较高的业务场景，建议采用流程电子化模式；对有外部参与人、数据量大、实时性要求高的业务场景，建议采用自助式流程处理模式；对规则明确、重复性操作多、办理频次高的业务场景，建议采用智能化流程处理模式。

流程设计，对已经识别和分类的业务流程，按照不同设计模式进行设计开发。电子类流程，依据流程梳理表，采用统一标准进行技术建模，在编码开发前使用BPM系统开发工具实现流程数字化设计，流程的设计需要遵从BPM系统架构要求，采用标准化工作交付件，如流程图、流程展示样式、表单排版等，按照统一要求完成和实现；自助式处理类流程，建立样式统一的申报界面和模式统一的系统校验改造规则，以确保其可行、高效和规范；智能化流程则充分借助RPA、OCR系统开发工具和设计规范，确保开发、测试过程不走样、不变形。

流程实施，将完成建模和设计的流程上线，实现业务数据和业务材料线上操作，确保每个环节都能按照规定执行。在流程执行层面，同一业务场景下业务流程只有一个，不管发起人、执行者具体人员是谁，都按照既定流程执行，所有角色参与人员都按照同样标准执行。

流程改进，BPM系统提供多渠道全方位监控手段，通过作业监控、指标统计等方式对流程的执行情况进行跟踪。对流程发起、流转、运行、办理过程等多维度数据进行分析，将系统数据接入大数据平台进行深度分析。建立流程反馈渠道，收集业务人员在流程办理过程中对流程的感受、意见和建议。根据监控结果和用户反馈，对流程进行不断改进和优化，形成流程优化闭环。

主要工作	特色标准
流程识别	业务场景识别标准、业务流程电子化标准、BPM 流程实施规范
流程分类	流程分类实施标准、流程智能化改造标准
流程设计	流程梳理表、流程开发手册、BPM 开发测试规范、RPA 开发设计手册
流程实施	BPM 运行管理规范、RPA 运行管理规范
流程改进	业务访谈报告、流程监控管理指标、流程异常处理手册、流程运营评价指南

表1：流程治理环节相应标准

3.2 平台技术支撑

3.2.1 新一代BPM系统建设

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，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。

3.2.2 完善监控及数据分析手段

流程治理是一个持续的过程，需要对BPM流程运行情况不断进行监测和改进，以适应业务的变化和发展需求。健全的监控方式和数据分析手段是促进治理水平持续提升的有力抓手。新一代BPM系统实现业务办理多维度深层次监控，建立合理高效监控模块，包括作业监控、部门流程查询、流程办理统计等，提供对流程的查询、管理和监控统计，如任务运行日志查询、异常任务处理、全部流程实例查询、流程使用统计等，建立起BPM流程完整的监控功能体系。

监控模块	BPM 作业监控	部门流程查询	流程办理统计
功能名	任务运行日志查询	全部流程实例查询	部门使用统计
	异常任务处理	部门已参与流程查询	个人使用统计
	作业运行监控	部门待办任务查询	应用系统统计
			部门流程完成统计

表2：BPM2.0系统监控功能模块

在数据分析层面，新一代BPM系统将业务数据从BPM流程数据库中分拆，业务办理主数据、办理明细数据、原始明细数据将存储到业务数据库，实现业务数据和流程数据相分离，进行数据分析统计时更加安全、高效，推动BPM流程数据治理、数据流动、数据共享技术实现，形成多层次、多维度的数据治理机制，充分发挥数据的核心要素积极作用。

3.3 金融科技赋能

技术的价值，更多体现在为业务赋能。中国结算上海分公司坚持业技融合创新，通过研究-试点-推广模式，不断推动金融科技在业务领域的深度应用。根据业务实际需要，将RPA、OCR等技术引入到业务办理中，取得了较好的应用效果，为流程治理体系的发展注入充沛动力。

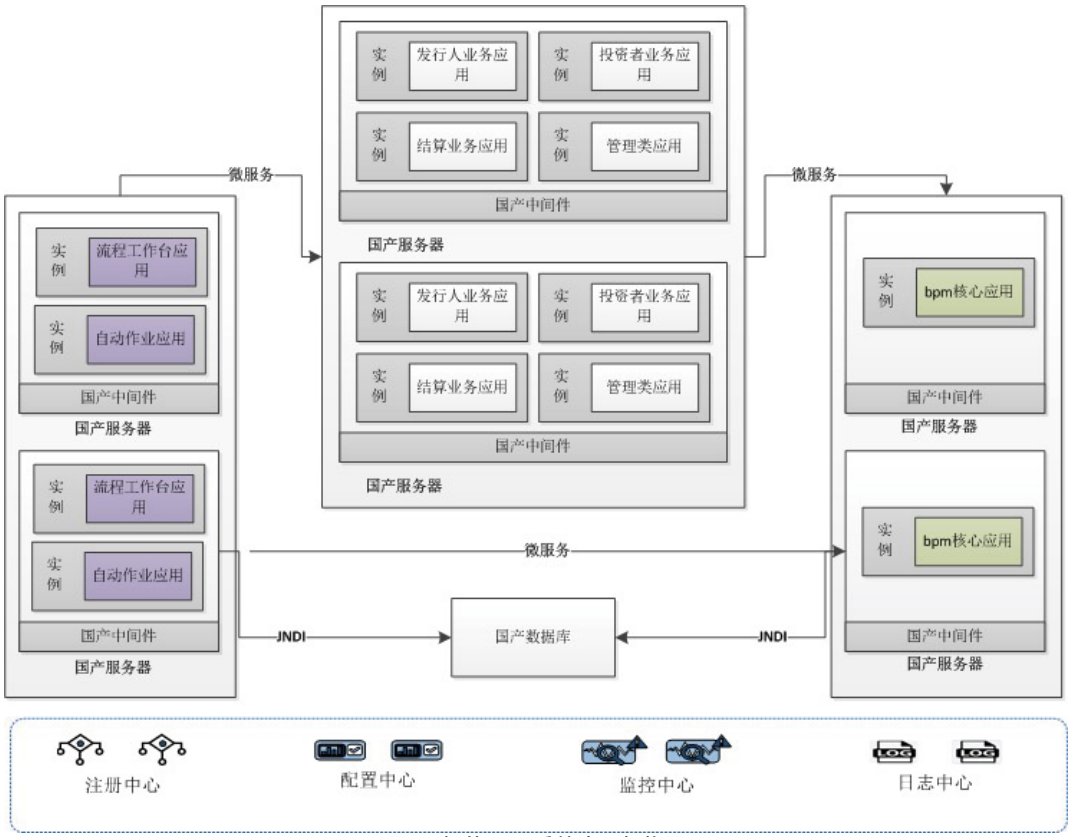


图3：智能BPM系统应用架构图

RPA是一种根据预先设定的业务规则操纵多个不同系统、自动完成特定功能和任务的软件，在处理大批量、重复性业务操作上具有天然优势。RPA业务在实施上具有“短、平、快”特点，不侵入系统数据，对办理效率带来立竿见影的效果，RPA与BPM系统结合，能够有效减轻业务人员操作量，实现跨系统、跨角色、跨时序的业务灵活定制与编排。例如某开户业务，RPA机器人每日可替代人工处理30余笔业务，每半小时自动进行开户业务的接单、审核、流程处理操作，单笔操作时长在3分钟以内，相比人工操作，耗时降低80%以上，同时业务处理更加及时，有效提升客户满意度。

OCR技术能够有效进行文本识别和转换为可编辑、可搜索文字，替代人工识别转换、比对文档中的业务数据，既减轻了业务人员的工作压力，也能提高业务办理效率。例如某对账业务，该业务场景需要从外部系统下载电子单据，将进行分类，根据不同类别的电子单，提取不同业务要素并录入到系统中，用于系统对账。该业务工作量大，准确性要求高，在RPA与OCR协同作用后，办理工作时间由2小时缩减至30分钟左右，业务人员只需要最后阶段确认数据的准确性，有效降低劳动负荷。

聪者听于无声，明者见于未行。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。通过加强对生成式AI、数字原生、隐私计算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上海分公司流程治理体系高质量发展。

3.4 流程数字化生态建设

3.4.1 探索引入流程挖掘

流程挖掘核心原理是从信息系统的事件日志中提取有价值的信息，用数据驱动的方式自下而上来发现、监控、仿真和改进实际流程。流程挖掘技术有效扩充了流程识别体系，为流程治理提供更科学、智能化的工具和方法。新一代BPM系统建设过程中，建立了统计指标体系，其中流程维度指标包括流程办理时长、流程参与部门数、流程参与角色数等；任务维度指标包括任务等待时长（从到达到领取）、任务办理时长（从领取到提交）、任务打回/回收次数等指标；岗位维度指标包括人员参与流程数、办理环节数、环节办理峰值等，同时提供指标的按月比对数据，呈现业务的发展情况，以期辨识业务流程慢点堵点，为业务流程的进一步优化提供数据支持，稳妥推进业务由经验决策型向数据决策型转变。

3.4.2 营造流程持续改进氛围

目前中国结算内部已经建立良好的流程使用氛围，无论是通过业务主办部门制定和发布业务流程还是日常工作的流程发掘，越来越多的业务人员愿意通过使用流程来改进工作过程，提升工作效率。

流程治理不仅仅是一次性的流程标准制定和系统建设，还需保障流程的持续性和灵活性，因此，需要不断强化数字思维、培育数字文化，提升全员数字素养。一是进一步在公司内部树立流程的权威性，打造“重视流程、使用流程、管理流程”的氛围，确保流程得到有效执行，避免出现绕开、避开流程的现象，形成按流程执行的操作习惯。二是建立流程改进机制，及时发现实际业务需求和线上流程之间的偏差，消除可能存在的业务流程与实际执行不符的潜在风险，协助流程主办部门进行端到端流程分析，形成业务人员和技术人员联动机制，积极听取一线流程办理人员反馈意见，根据意见和建议进行流程迭代改进，消除操作堵点和瓶颈，循序渐进培养分公司持续改进的BPM流程机制和流程文化。

四、总结与展望

中国结算上海分公司通过标准化、搭平台、引技术、建生态四驾马车带动流程治理体系的发展，随着业务规模不断扩展，系统数字化迭代进程加快，流程治理体系也紧随趋势不断调整、优化、创新。基于业务实践的流程治理探索取得了一定的成果，重塑了证券登记结算业务流程，建设经验可复制、可推广，形成了一套业务流程数字化转型标准体系和规范，夯实了数字根基。对业务流程的严格管理和监控措施也有助于预防和减少潜在的风险隐患。

下一步，中国结算上海分公司将持续推进流程治理体系建设，充分进行金融科技的前沿探索，不断深化流程挖掘技术的研究和应用，采用流程自动发现、复杂事件处理等技术和工具，开展流程挖掘和流程再造，发挥基础技术能力对于推动业务办理高水平运转的核心支撑作用，为公司及行业数字化转型持续贡献力量。

参考文献：

[1]权国志.BPM业务流程平台通用化研究[J].信息与电脑(理论版),2021,33(13):114-116.
[2]张春阳,李晓丹.以数字化流程再造为突破，打造中小银行科技管理新生态[J].中国金融电脑,2023,(12):11-13.
[3]曹善文.基于流程挖掘视角下的数据要素利用研究[J].信息通信技术与政策,2023,49(04):59-64.
[4]戎力.机械性重复劳动可交给机器人[J].《计算机与网络》.2018,44(07):32-33

基于《证券期货业信息系统压力测试指南》的集中交易系统压力测试实践

王岐、王晓龙、李鑫、赵晓红、刘震、于召洋 | 中信建投证券股份有限公司
E-mail：wangqixx@csc.com.cn

摘　要： 近年来证券期货行业交易系统宕机事件时有发生，究其根本原因其中多数为系统性能问题。压力测试作为信息系统性能保障最重要的环节，越来越受到监管机构和证券期货公司的重视。中信建投证券股份有限公司交易系统压力测试工作开展多年来，对标2023年《证券期货业信息系统压力测试指南》（意见征求稿）中各项要求，提出“补短板、锻长板”的改进要求，经过半年多的完善，相关工作效果取得明显提升。本文通过对中信建投证券集中交易系统压力测试工作的介绍，使读者对相关工作有全面的了解。同时，本文阐述了如何通过对标“指南”中的要求，使公司信息系统压力测试工作的规范性、标准性和科学性得以进一步完善。

关键词： 交易系统性能；压力测试指南；压力测试指标

一、引言

2023年中国证券监督管理委员会科技监管司、上海证券交易所、上交所技术有限责任公司、深圳证券交易所、大连商品交易所、上海期货交易所、中国证券登记结算有限责任公司上海分公司及多家证券、基金、期货公司，交易系统软件开发商共同编写了《证券期货业信息系统压力测试指南》（意见征求稿）（以下简称“指南”）。指南对证券期货业信息系统压力测试的原则、内容、目标、流程等几个方面进行了详细的阐述，旨在对证券期货业压力测试进行全面规范，提升行业信息系统压力测试能力，提高压力测试质量，控制压力测试实施风险，从而提升行业信息系统运行的稳定性和可靠性。

中信建投证券股份有限公司自交易系统大集中以来，由于客户数量合并增多，交易行为趋于活跃，单个交易节点吞吐量成倍增加。根据《证券期货业网络和信息安全管理办法》第四十八条之规定，证券期货业关键信息基础设施运营者应当对关键信息基础设施的安全运行进行持续监测，定期开展压力测试，发现系统性能和网络容量不足的，应当及时采取系统升级、扩容等处置措施，确保系统性能容量在历史峰值的三倍以上，交易时段相关网络带宽应当在近一年使用峰值的两倍以上。鉴于此，中信建投自主开发了压力测试工具，实现模拟客户发送不同业务请求的功能，并能够对业务处理的延时、吞吐率、CPU等指标进行实时统计和展示。压测目标主要是针对综合处理性能进行评估，业务样本取自生产系统实际运行过程中占比较大、排名靠前的若干功能号，并依据实际占比控制各功能号每秒发送请求的数量，达到与生产运行状态基本一致的效果。客户样本是脱敏系统内实际持仓的客户，随机抽取，保证客户样本的可用性和随机性。综合性能评估以来，在实际升级变更过程中确实提前发现过一些功能号耗时过大的系统问题。

但是由于近年来系统复杂度增加，客户使用程序化交易等因素影响，交易系统瞬时并发量突增的情况越来越多。跨系统调用、爆款产品定时销售等突发事件导致的系统综合性能下降情况时有发生，当前系统性能容量方面的风险逐渐显现。原有的综合性能压力测试方案存在测试环境、工具配置、功能号入参、报告编制等方面的不足，对压力测试工作的效率和结果造成了一定的影响。“指南”基于证券期货业信息系统技术特点和业务类型、监管要求等因素考量，对交易系统压力测试工作进行了全面指导，在测试流程、测试管理方面提出要求并给出落实方法。尤其在压力测试指标及度量方面提出统一标准，使行业信息系统压力测试工作有据可依。

二、中信建投证券交易系统压力测试工作

本文将结合中信建投证券压力测试工作实践，在组织架构、测试环境、测试工具、测试流程、测试内容、测试报告六个方面进行说明。

2.1 压力测试工作组织架构

中信建投为支撑及保障压力测试有效开展，进行了合理的职能设置并建设了完善的治理架构。在组织架构方面，压力测试相关团队包括基础架构组、交易系统运维组和交易系统开发组。基础架构组组织牵头并负责维护测试环境、压力测试工具、落实具体压测工作，交易系

统运维组对压力测试报告进行确认并依此评估交易系统升级变更风险，交易系统开发组提供应用系统和数据库升级变更的技术支持。

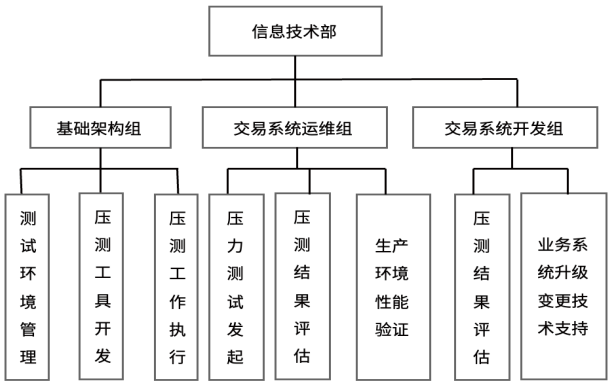


图1：压力测试工作组织架构

在交易系统升级变更之前，由系统开发组发起压力测试申请并提供应用程序、数据库升级变更脚本给基础架构组，后者准备压力测试环境进行多场景压力测试，形成压力测试报告后评估升级风险并反馈给开发组和运维组。交易系统升级窗口期，运维组会在生产环境进行最后的验证测试，确保性能没有问题后留存相关资料并正式投产。

2.2 压力测试环境管理

为确保压力测试的有效性和准确性，需要对压力测试环境进行有效规划、配置、维护和管理。中信建投证券压力测试环境对标生产环境，全部使用物理机部署。数据库采用4路服务器，应用中间件采用2路服务器。根据系统各组成模块对于硬件资源的不同要求，分配不同配置的机器，如交易中间件一般配置主频高服务器、通讯中间件配置低延迟服务器。考虑到降本增效，压力测试环境服务器硬件基本来自生产环境利旧的服务器，在计算能力方面略低于生产环境。此外，为了更好的与生产环境对标，压测环境网络均为万兆接入。

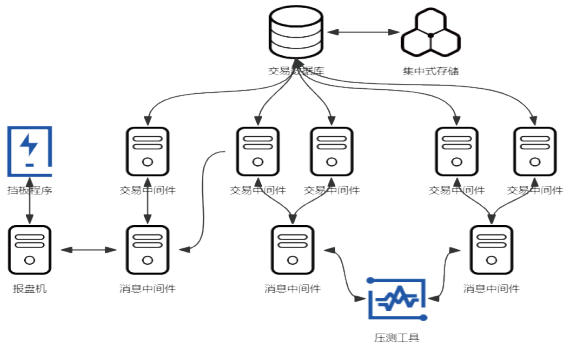


图2：压力测试环境

压力测试环境由专人维护，如日常的监控、巡检等。在发现故障时可以及时响应和处理，避免影响压力测试相关工作。此环境被定义为敏感测试环境，与其他开发测试环境网络物理隔离。访问压力测试环境通过双因素认证的虚拟桌面进行，有效防止非授权访问、数据隐私泄露、网络攻击等。

2.3 测试工具

基于交易系统高吞吐量、低延迟、注重稳定性等要求，中信建投证券自研了交易系统压力测试工具。该工具可以对交易系统进行全链路压力测试和评估，评判交易系统在高负载条件下的性能和稳定性，用于系统变更或者升级后上线前的检测，提前发现潜在的问题和风险并及时反馈，减少业务中断风险。

2.4 测试流程

“指南”对压力测试流程提出了明确的要求，中信建投证券交易系统压力测试可以从测试规划、测试设计、测试执行、测试总结四个方面说明。

2.4.1 测试规划

首先分析测试需求，根据升级变更文档明确测试对象、测试指标、测试方法、测试资源、测试风险等。根据测试需求制定测试计划，并通过DEVOPS工具分配测试任务，明确任务负责人、时间要求，并通过流程审批跟踪任务进度和结果。

2.4.2 测试设计

压力测试设计主要包括业务场景设计，如业务场景有混合压测、基准压测、专项压测等；测试环境设计，如扩容应用中间件数量、应用新硬件新驱动、更新操作系统补丁等；压力参数设计，如增大某个功能号占比、调整并发数等。根据每次升级变更内容，针对可能影响性能的业务变化，需要设计补充新的测试用例。

2.4.3 测试执行

根据测试方案执行测试用例，记录测试结果进行实时分析。基线压力测试通常运行20分钟左右，对交易系统整体性能进行观测和评估；摸高压力测试将并发和吞吐量调整到测试环境的上限，运行5到10分钟左右，观察系统承压能力；稳定性测试在固定压力情况下运行2小时以上；写入压力测试则是选取包括委托、成交、登录等写

入功能号进行压测。执行压力测试过程中要开启测试环境监控，通过CPU使用率、网络使用率、输入输出延迟等指标实时判断压力测试异常情况，并及时记录压力测试过程中各项数据。

2.4.4 测试总结

每次压力测试结果会以报告的形式作为测试资产留存，为后续性能评估、系统扩容等工作提供参考。生产环境实际运行中的异常性能表现也会对压力测试起到正向的反馈作用，为分析现有压力测试管理工作中的缺陷提供了现实依据，同时在改进压力测试工作后也能避免更严重的系统故障的发生。

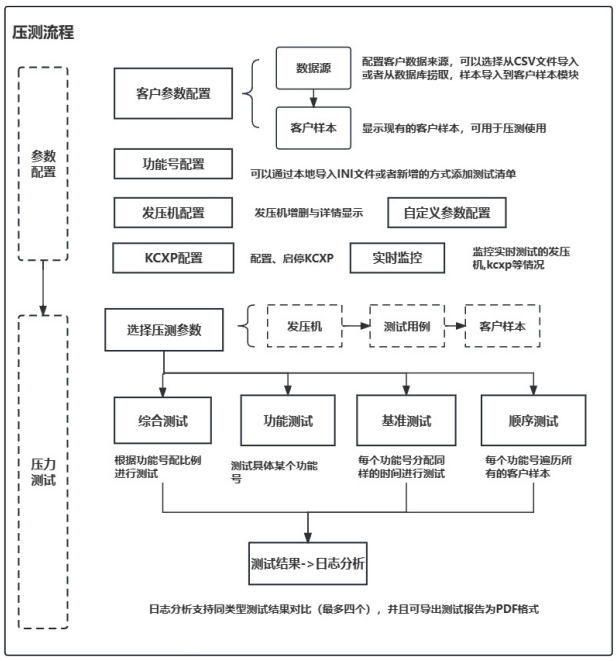


图3：压力测试流程（KCXP为消息中间件）

2.5 测试内容

测试内容包括配置业务场景、选取压测样本、置备测试数据、记录测试数据。

为了保证压力测试的有效性，选取的压力测试场景必须真实反映实际生产业务调用情况。中信建投证券压力测试的业务场景参考真实的生产业务数据，采样近一个月交易系统的调用情况，根据业务调用的实际占比调整压力测试模型。

压测样本由若干条测试用例组成，每个测试用例包含必要的入参，如订单功能需要客户号、股东代码、股票代码、买卖类型、委托金额、委托手数等。压力测试平台会顺序选择测试用例并根据业务比例循环发起系统调用。

不同业务场景对应不同的数据置备方法，如模拟系统初始化后测试需要将订单表和成交表清空，而模拟业务高峰期时的测试则要根据实际生产情况对订单表和成交表进行装载后再开始压测。

测试数据包括吞吐量、延迟、资源消耗、稳定性等方面，根据“指南”的标准分类，中信建投将压力测试指标进行了规范，具体内容见下表。

指标类型	指标名称	度量方法
吞吐率	订单峰值吞吐率	X=A/T，A=发送订单并接收响应订单确认响应的订单总数,T=持续报单时间。
	成交峰值吞吐率	X=A/T，X=A/T，A=成交回报的订单总数，T=持续报单时间。
	订单持续吞吐率	X=A/T，A=发送订单并接收响应订单确认响应的订单总数；T=持续报单时间。
	成交持续吞吐率	X=A/T，A=成交回报的订单总数；T=持续报单时间。
	并发查询处理能力	X=A/T，A=发送查询并接收查询确认响应的查询总数；T=持续查询时间。
	并发登录处理能力	X=A/T，A=发送登录并接收登录确认响应的登录总数；T=持续登录时间。
延迟	订单处理时延	$X=[\sum_{n=1}^{n}(T_{1k}-T_{0k})]/n$，T0=订单到达被测系统的时间戳；T1=订单回报离开被测系统的时间戳；n=订单数量。
	订单成交时延	$X=[\sum_{n=1}^{n}(T_{1k}-T_{0k})]/n$ T0=订单到达被测系统的时间戳；T1=成交回报离开被测系统的时间戳；n=订单数量。
	查询响应时延	$X=[\sum_{n=1}^{n}(T_{1k}-T_{0k})]/n$ T0=查询请求到达被测系统的时间戳；T1=查询响应离开被测系统的时间戳；n=查询请求数量。
	登录响应时延	$X=[\sum_{n=1}^{n}(T_{1k}-T_{0k})]/n$ T0=登录请求到达被测系统的时间戳；T1=登录响应离开被测系统的时间戳；n=登录请求数量。
处理容量	日订单处理容量	$C=\sum_{n=1}^{n}H_t$，t=单位时间；Ht=单位时间内成功处理订单笔数；n=日处理时间。
	日成交处理容量	$C=\sum_{n=1}^{n}H_t$，t=单位时间；Ht=单位时间内成功处理成交笔数；n=日处理时间。
	交易单元登录容量	C=MAX(A_1,A_2,…,A_n)，An 为第 n 次测试时系统交易单元登录数量；n=验证次数。
基础数据容量	系统帐户容量	P=A/B，A=目前实际需要支持的最大账户数；B=系统设计时可支持的最大账户数。
	系统产品容量	P=A/B，A=目前实际需要支持的最大产品数；B=系统设计时可支持的最大产品数。
	系统持仓总数	P=A/B，A=目前实际需要支持的最大持仓数；B=系统设计时可支持的最大持仓数。
	系统成交金额	P=A/B，A=目前实际需要支持的最大成交金额；B=系统设计时可支持的最大成交金额。
	系统交易单元容量	P=A/B，A=目前实际需要支持的最大交易单元数量；B=系统设计时可支持的最大交易单元数量。
资源利用性	CPU 峰值占用率	CPU 使用率百分比
	内存峰值占用率	内存使用率百分比
	磁盘 IO 峰值带宽	单位 MB/秒
	网络带宽峰值占用率	网络带宽使用率百分比
成熟性	订单或数据丢失率	P=A/B，P=A/T，A=错误订单或数据的数量；T=系统故障时段订单或数据总量。
易恢复性	系统恢复时间 RTO	X=T2-T1，T2=系统恢复至可以支持业务正常运营的时间点；T1=系统发生故障导致业务停顿的时间点。
	数据恢复时间 RPO	X=T2-T1，T2=系统发生故障导致业务停顿的时间点；T1=故障发生后系统可以恢复到的最近时间点。
稳定性	处理正常业务量时长	定义的交易时段内稳定运行的时间

表1：测试指标

应用系统变更升级前会对所有功能号进行四种不同的压力测试，分别为全功能测试、主要功能测试、单功能号基准测试、业务突发模拟测试。全功能测试即是将实际生产中发生的所有业务进行全覆盖测试，通常会涉及到上百个功能号，该测试优点是尽量模拟生产系统的所有业务情况。主要功能测试则是将生产环境占比前99%的功能号纳入到测试范围，将调用功能号个数减少到40以内，该测试基本可以反应出实际生产情况。单功能号基准功能测试为每个功能号分配固定的压测时长，根据配置的用例，逐个压测功能号，用于测试每个功能号的基准性能情况。业务突发模拟测试是在主要功能测试的基础上，逐个将某一业务量配比提高3倍，观察其对系统性能的影响。

除了常规压力测试之外，日常压力测试工作还包括若

干专项测试内容，如业务系统配置变更、新版本报盘软件变更、操作系统版本或杀毒软件版本变更等。无特殊要求的情况下，此类测试内容与常规压力测试基本相同。

2.6 测试报告

每次压力测试之后都会生成压力测试报告，记录压力测试的全过程，主要包括压力测试环境、用例、过程、结果、异常等信息，并且给出本次压力测试的结论。压力测试报告会备份并保存，以备比对分析、检查审计、跟踪回顾及经验总结。

三、总结

证券基金行业信息系统对性能要求越来越高，关键信息系统的压力测试已经开展了许多年。然而行业级别的相关标准和制度相对缺乏，相关机构各自为战，隐藏着诸多风险。“久旱逢甘霖”，2023年中国证券监督管理委员会科技监管局联合多家机构发布了《证券期货业信息系统压力测试指南》（意见征求稿），对于行业的指导意义巨大。中信建投证券开展压力测试工作以来，有效提升了交易系统测试质量，控制了业务量增大等情况带来的性能风险，防患于未然，保障了交易系统运行的稳定性和可靠性，但是在规范化、科学性方面存在一定的缺陷。本着“真学真用，先行先试”的思想，压力测试项目组深入学习和参考“指南”，对信息系统压力测试工作进一步完善，尤其在测试流程、测试指标度量等方面更加规范、科学、有效，切实保障信息系统平稳高效运行。

参考文献：

[1]中国证券监督管理委员会，《证券期货业网络和信息安全管理办法》，2023年2月
[2]中国证监会科技局，上海证券交易所等，《证券期货业信息系统压力测试指南》（意见征求稿），2023年10月

交易全链路追踪监控实践

应国力、李健舒、王海兵、张贺龙、刘军 | 上海金融期货信息技术有限公司
E-mail：yinggl@cffex.com.cn

摘　要： 目前全链路追踪在互联网行业微服务体系架构下应用非常广泛，但在金融行业特别是证券期货交易系统上使用案例甚少，主要原因为金融行业对系统安全稳定运行要求较高，系统的升级频率低，交易系统引入这种注入式链路追踪的风险较大。针对该现状，中国金融期货交易所（以下简称为“中金所”）自主设计研发了基于网络镜像技术的交易全链路追踪方法，该方法在不影响主交易系统实时交易的情况下，通过解析核心网和接入网的镜像流量，实现对交易系统逐笔委托及行情纳秒级精度的延时测量和监控，具有对系统无侵入、实时性强、精度高、易扩展、可视化等诸多优点。

关键词： 全链路追踪；交易系统；逐笔监控；无侵入

一、整体架构及部署方案

交易全链路追踪监控主要由4大模块组成，分别为SPAN镜像模块，数据采集分析模块，大数据可视化模块，主要架构如图1所示：

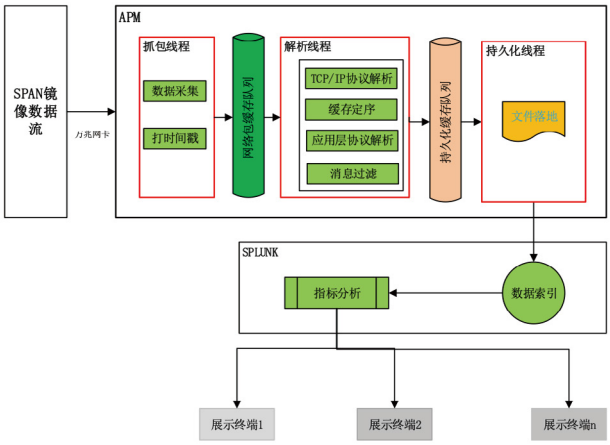


图1：交易全链路追踪监控架构

SPAN镜像模块 负责将接入网与核心网交换机的全量网络数据包实时镜像到数据采集模块，其通过端口镜像方式采集数据包，不会占用交易主机资源与交易核心网带宽，且对应用无侵入、无感知，符合交易链路场景中的安全要求。

数据采集分析模块 通过应用性能管理工具（以下简称“APM”）的网络抓包单元捕获接入网与核心网的全量网络数据包，并对全量交易网络数据进行逐笔分析。其具体功能包括数据过滤、数据预处理、协议解析、字段解析、数据管理、数据分析、格式化落地、并行加速等，上述功能完全自主开发，单实例性能可达10万笔报单/秒实时解析。其支持横向扩展，理论上限为机器资源上限。

大数据可视化模块 负责展示分析结果与检索历史分析结果。其基于SPLUNK与云原生技术，可以在Web端便捷访问，在以亿为单位的数据记录数的数据库中，检索性能可低至10毫秒级。其支持海量图表 ，通过可视化与拖拽式操作可快速制作数据展示视图与告警监控视图。

部署方案如图2所示，当有交易数据产生并以网络包形式进入交换机时，SPAN镜像模块将数据从交换机镜像到目标网卡。数据采集模块将原始网络包同步至数据分析模块，数据分析模块对网络进行解析与过滤，并根据需求进行各种数据分析，并将结果进行格式化落地。大数据可视化模块对落地数据建立索引，并对可索引的海量数据提供可视化服务，最终用于分析与展示。

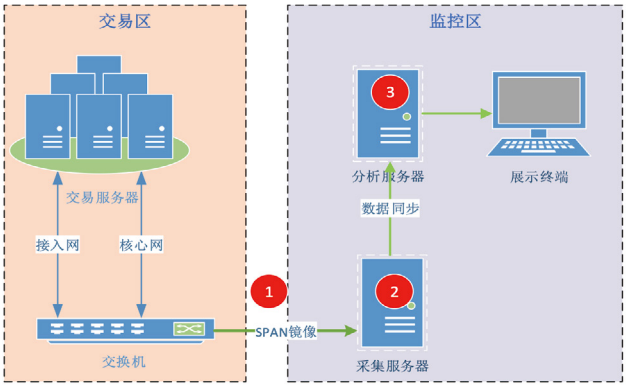


图2：交易全链路追踪监控部署方案

二、关键技术

2.1 SPAN镜像

SPAN（全称Switched Port Analyzer）技术是一种网络设备的镜像技术，主要用来监控网络交换机上的数据流，使用SPAN镜像可以在不干扰现有业务的情况下，

将需要监控的网络流量通过旁路的方式，发送到本地或者远端设备上，在网络监控排障、网络抖动延时分析、网络流量异常等方面使用非常广泛。金融期货交易系统特点是高可靠、高并发、低延时、网络流量集中（行情瞬间），采用常规的基于日志采样的监控采用频率低，导致监控精度低；而基于注入式的链路追踪监控，则需要侵入改动系统，带来非常大的升级成本，对系统的稳定性也有较大影响。而基于SPAN技术的数据分析，可以很好的满足金融期货交易系统的监控需求。本实践使用SPAN技术，可以在不影响报文正常处理流程的情况下，将镜像端口的报文复制一份到观察端口，再通过APM分析模块来分析复制到观察端口的报文，进行网络及应用的监控和故障排除，具有对系统无侵入、实时性强的优点。

为了提升监控精度，本实践支持两种精度模式。如果网络设备是低延时网络设备，支持ERSPAN（Encapsulted Remote Switch Port Analyzer），网络报文的时间戳可以直接使用ERSPAN包头时间戳，精度可以达到百纳秒级，如果网络设备是普通的网络设备，只支持常规的SPAN，网络报文的时间戳可以采用APM主机的时间戳，精度为微秒级。目前金融期货交易系统领域的延时通常在百微秒量级，在资源有限的情况下，采用APM主机的时间戳基本满足业务需求。

交易系统的前置、撮合、行情等所有模块进出网络设备的数据流都会通过SPAN汇集到APM主机接收端网卡，特别需要注意的是，SPAN流量镜像时要考虑交换机入口的带宽，如果入口带宽是千兆，可以多个口汇聚到一个口上；如果入口带宽是万兆，只能一对一SPAN，否则可能会出现较高的丢包率，影响监控精度。

2.2 抓包

通过配置网络设备，可以把SPAN镜像流量推送到指定的主机网卡，数据到达网卡后，还是原始的网络报文，原始报文的目的IP和端口并不是APM主机所在的网卡IP和端口，无法使用操作系统提供的SOCKET函数库获取原始报文。本实践基于libpcap函数库在APM中设计了一个网络抓包单元，用于捕获原始网络数据包。网络抓包单元的核心功能是捕获网络数据包，支持各种参数配置。APM所在主机需要接收各个模块的流量，通常会采用多块网卡作为不同业务镜像流量的目标端。所以抓包单元需要支持配置指定要捕获数据的网卡接口。另外，对于不同的镜像流量，采用的传输层的协议可能不同，同一个流量内部可能有需要关注的数据，也有需要过滤掉的数据，所以抓包单元需要支持指定要捕获的协议类型，从而减少不必要的数据量。此外，抓包单元也需要支持指定端口号，以便于只捕获特定端口上的数据流。

金融期货交易系统在行情推送瞬间通常有非常大的网络流量（微爆现象），即使是万兆网卡也存在瞬间流量满载的情况，所以抓包单元的瞬间吞吐能力是整个链路追踪系统的关键性能指标，APM单个进程可以配置多个抓包单元，每个抓包单元由单独的线程驱动，通过配置线程池的方式管理抓包单元可以实现抓包单元的灵活伸缩，满足不同硬件条件和业务监控场景下的抓包需求，同时不影响系统的稳定性和响应时间。根据测算，单个抓包单元可以达到8万笔报单/秒，全链路最高30万笔报单/秒的抓包性能。

捕获的数据通常会被存放在缓存队列中，等待进一步的处理。这个队列可以是内存中的数据结构，也可以是磁盘上的文件。存放数据的方式可以根据具体需求进行配置，以确保数据不会丢失，并且能够高效地被后续的数据分析程序处理。

2.3 消息解析

解析线程从缓存队列中读取数据后，对原始的链路层消息进行解析，包括TCP/IP协议层转换、缓存定序、应用层解析、消息过滤。消息解析完毕后，会实时落地成解析后的文件。

TCP/IP协议解析 APM的输入是交易系统数据链路层消息，需要完整解析整个TCP/IP协议栈。libpcap库提供了基础的抓包接口，APM通过调用libpcap的抓包接口，从网卡抓取原始的链路层报文，然后从下往上依次解析MAC层、链路层、IP层、TCP/UDP层的协议。

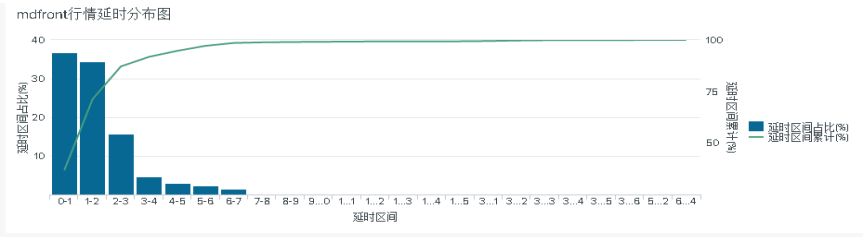
缓存定序 镜像流量无法保证数据可靠传输，所以原始报文中可能存在丢包、重传、乱序、数据不完整等现象，APM为了保证数据的正确解析，底层创建了一个定序缓存，用于保存接收的数据，所有收到的TCP消息，都需要通过定序缓存，如果收到的的包是一个完整的业务包，才进行解析，如果出现包有缺失，则丢弃。

应用层协议解析 SPAN的流量有接入网流量和核心网流量，其中接入网流量网络层使用的是TCP协议，应用层使用的是FTD协议；核心网网络层使用的是UDP协议，应用层使用的是XTP协议（XTP为中金所自定义消息协议）。APM可以根据输入协议类型动态选择解析应用层协议。

消息过滤 随着期权的推出，每天交易和行情的数据量可以达到百GB级别。如果对所有的消息都进行索引和归档，资源的需求是巨大的。所以为了节省硬件资源，APM支持对消息进行过滤，过滤规则包括：

（1）消息种类过滤：APM支持业务消息种类级别的过滤，用户可以通过IGNORE_TID选项忽略掉不想落地的消息种类，比如资金持仓切片、合约状态切换等消息；

（2）字段过滤：APM也支持字段级别的过滤，考虑到某些字段可能是敏感字段，或者不需要的字段，可以在配置文件中把



mdfront行情延时分布表

延时区间(ms)	延时区间占比(%)	延时区间累计(%)
0-1	36.76 %	36.76 %
1-2	34.54 %	71.30 %
2-3	15.85 %	87.15 %
3-4	4.68 %	91.83 %
4-5	2.91 %	94.74 %

图3：行情推送延时分布图

该字段设置成FALSE，从而避免该字段的输出；

（3）IP过滤：SPAN操作会把交换机指定端口上的所有流量都进行转发，有些IP的数据可能不是用户关心的，可以通过过滤操作，去掉不关心的数据；

（4）行情采样过滤：目前TCP行情是轮询发送，如果行情席位有1000个，那么同样的一笔行情通过SPAN以后就会有相同的1000份数据。APM支持配置采样间隔，从而大大减少行情消息数量。

pcap包分析和落地 APM支持直接分析通过其他抓包工具（如tcpdump）抓取的pcap包，也可以通过添加参数落地pcap包。

三、系统建设成效

系统无侵入 传统的链路追踪的方式通常是通过日志采样或者字节码注入等方式追踪订单。采用日志的方式，系统性能开销大，采样频率低，很难观测到系统的微观层面的运行状态，特别是对于期货交易这种周期性微爆流量的场景，采样很难真实的反映订单在系统中的实际延时分布情况。基于字节码注入的链路追踪方案需要预定义注入逻辑，修改代码进行适配，如果分析逻辑或者统计方式发生变化，就需要改动被监测系统的代码。而交换机SPAN镜像的方案，直接通过交换机的SPAN功能，从网络层面镜像一份数据，被监测系统无侵入、无感知。

追踪精度高 基于日志或者字节码注入的追踪方式，通常使用的是主机时钟，精度单位为微秒，不同主机之间，使用NTP时钟校准，如果出现时钟校准失败，对产生的数据精度会有很大的影响。 基于网络镜像的方式，有两种精度模式，如果采用交换机的时钟戳，精度可以达到百纳秒级，如果采用主机的时间戳，精度可以达到微秒，两种方式可以同时工作，相互佐证，有很高的容错性。

监控精细化 数据实时同步、实时分析、实时展示，相比与基于采样日志的追踪方式，全链路追踪方案能从更细粒度的对金融交易系统的服务质量进行度量，基于全链路追踪技术，引入了业务指令网络分段耗时、服务质量抖动概率密度、行情驱动抢单分析、前置分发时延差异等数十个细粒度监控指标，图3为SPLUNK中的行情推送延时分布图。

方案易扩展 本实践支持会员、席位、客户等多级别的订单和行情追踪，从一笔订单进入交易系统到排队、撮合、以及产生行情等进行全链路追踪。系统支持单实例多端口和多实例多端口两种部署模式，单实例网络抓包峰值速率可以达到10万笔报单/秒，支持水平扩展，在资源充足的情况下采用多实例多端口模式，处理速度快，实时性高。如果计算资源有限可以采用单实例多端口模式，节省资源。

四、总结与展望

本文对交易全链路追踪监控的架构和关键技术分别进行了介绍和阐述。为了解决金融期货交易系统性能和业务实时监控侵入性强、精度低、延时高、可扩展性差的问题，本实践采用全链路跟踪的思想，在不影响主交易系统实时交易的情况下，提出了一种更加轻量化、精细化的监控方式。目前，该方案已在生产和内部开发测试环境投入使用，至今已在多个应用场景中落地，并发现多个常规监控手段无法监测的现象，见表1。

应用场景	解决问题
订单与行情数据关联分析	首次获取了行情触发订单的精确延时分布
全链路监控	发现订单拖尾现象和防火墙性能不佳导致的重传率高的问题
组播和 TCP 行情数据延时分布	获取了组播和 TCP 行情的精确性能差距
分段延时分析	协助定位测试环境延时双波峰和低频交易下毛刺率高的问题
各托管机房订单和行情延时分布概率密度分析	定位了托管机房性能瓶颈

表1：APM应用场景

随着中国金融市场的发展，交易所业务量和业务复杂程度必然会不断上升，如何更好地进行全链路追踪监控、提早发现隐患显得愈发重要。中金所后续将继续精进，持续探索优化，为金融市场稳定运行做出贡献。

参考文献：

[1]Dapper, a Large-Scale Distrubuted Systems Tracing Infrastructure [ER/OL].http://research.google.com/pubs/-pub36356. 2010.

[2]Cisco Systems Inc, Catalyst 3550 Multilayer Switch Software Configuration Guide [M]. 2006.

[3]戴昆，约翰逊，默克,等.Splunk智能运维实战[M].机械工业出版社，2015.

证券核心交易系统代码审计平台建设实践

华仁杰，唐淑艳，华焰，施爱博 | 东吴证券股份有限公司 | E-mail：shiaib@dwzq.com.cn

摘　要： 核心交易系统是券商业务开展的重要核心系统，其软件质量关系到所有投资者的交易安全。本文详细分析了券商开展此项工作所面临的问题和应对思路，在此基础上研发了代码审计平台，实现落地应用，把代码审查能力赋能到测试运维之中。平台实际投入使用后，取得了良好的使用效果，有效提升了代码审核质量与效率，全面加强了代码审核体系建设，形成了一道行之有效的代码保护屏障。

关键词： 代码质量保障；代码审核；证券核心交易系统

一、概述

核心交易系统是券商业务开展的重要核心系统，其软件质量关系到所有投资者的交易安全。随着证券业务和技术发展需要，核心交易系统进入信创改造与日常运维系统升级的叠加，因此软件产品质量保障是一项十分重要工作。

软件代码是构建系统软件的逻辑基础，从代码层面进行审查、杜绝系统安全隐患是软件工程常规方法。代码审核是通过编码以外的人员检查代码是否存在安全隐患，以此来保持代码和产品质量的一个过程。通过代码审查能够尽早地发现软件的缺陷，找出动态测试难以发现或隔离的软件缺陷软件。

通常代码审核都是基于软件开发团队内部的一项工作，适用于熟悉项目的人进行审核。券商作为软件采购商，并没有参与到具体的软件研发过程中，而且券商系统代码变更频繁，审核工作量较大，这种作为第三方的代码审核模式往往事倍功半，因此经了解，鲜有券商会采用人工审核的方式来进行软件质量把关，并且业内也没有成熟有效的方案来适用于第三方审核。那么，随之带来的困境就是在系统出现问题的时候，券商严重受限于对供应商的依赖，往往无法第一时间找到问题根源，延误了处理问题的时机。

如何进一步有效提高生产安全是券商重点关注的问题。在建立健全系统软件质量保障体系，规范系统建设升级流程的基础上，东吴证券多年以来坚持将代码审核作为保障系统软件质量的一道门槛，尤其在核心交易系统信创改造阶段，进一步加强探索适合券商端代码审核的工作思路，形成了规范化、自动化、工具化、数据化、智能化的代码审核方法，并以此打造了东吴证券代码审计平台。

本文将从工作思路入手，分析券商在核心系统代码的质量保障方面所面临的问题，并提出应对方法，在此基础上着手构建代码质量保障体系，并研究开发代码审计平台，将体系思想融入到平台中，通过工具来辅助推动保障体系的形成和发展。

二、工作思路

2.1 面临问题

通常，代码审核会作为软件开发流程的一个必选项。常规的代码审核工作会在软件开发过程中进行，通过检查代码来确保其质量和可靠性。同时在审核过程中辅以模糊测试等安全检测手段，可达到及时发现软件缺陷及漏洞的效果。[1]但是券商作为软件使用方，通常不直接参与软件研发，因此券商对代码审计过程与供应商软件研发过程是相互独立的，这就无法有效使用传统的代码审核方式，需要另辟蹊径，找到符合实际情况的代码审核措施。经过长期观察及经验总结，在审核实施过程中遇到了以下问题：

1）不规范的编码风格成为阻碍代码理解的障碍

软件编码需要遵循安全编码规范是业界共识。由于软件研发人员的能力水平和编码习惯，会形成不同的编码风格，而这些风格化的代码就制造了旁人阅读理解的屏障。

2）审核人员与软件开发人员之间缺少有效沟通机制

沟通是帮助代码审核人员理解代码，了解软件设计思路，从而能够进一步结合已有经验去发现软件缺陷的基础。但事实上，券商作为软件使用方，很难与供应商开发人员保持稳定的沟通，无法及时准确地获取软件程序的相关信息，从传统代码审核方法角度来讲，已经失去了这项工作的开展基础。因此，从券商代码审核角度出发，审核人员与软件开发人员之间的有效沟通机制是缺失的。

3）券商审核工作量较大且集中是代码审核质量的一大困境

由于无法详尽知道开发计划，券商审核人员一般等待供应商发布升级包，确认升级项后，再查找对应代码，对变动代码进行审核。因此，会在升级前夕聚集大量的审核任务。在有限时间的情况下进行代码审核，必然会影响代码审核质量。普遍认同的观点是，要控制单次代码审核的时间，否则审核效率及质量会急剧下降。同时，人员配备不足也会造成工作量的囤积。

4）缺乏专业领域软件的针对性审核

采用传统方法对特定专业领域软件进行代码审核，只能关注代码本身的逻辑问题，与专业领域知识相隔离，不易发现与功能逻辑相关的有价值的问题。此外，在实际工程中，由于未对特定专业领域软件的代码审查方法进行归纳总结，从而造成了同一专业领域的软件的不同的测试项目的代码审查工作之间的可借鉴性不强，效率不高。[3]券商端代码审核必然是要充分结合业务，有针对性地进行业务逻辑方面的审核。

5）无法在代码层面快速定位问题

在实际运维过程中，当出现问题的时候，只能定位到故障模块，很难进一步查找到错误原因。如果此时依赖供应商远程解决，增加了沟通成本和处理问题的时效性。如果能够快速找到问题所关联的代码，将代码与问题查找相结合，必然能够提高实时解决问题的能力。目前，还未有相关产品能够把代码引入到运维阶段，是一个值得尝试的方向。

2.2 应对思路

目前核心系统处于持续优化阶段，同时，证券业务发展迅速、需求变化较大，系统代码修改是一个常态化的过程。券商作为使用方，代码安全审计可用于软件升级上线前的审核检查，降低错误发生风险。可以看到，券商有必要把好升级软件的最后一道关卡，从使用、运维的角度去审计升级软件。

结合当前软件升级质量管控现状，制定了如下的一些代码审核措施方向，希望能够提高审核效率，提高代码质量保障能力。主要从以下几个方面入手：

1）采用多人审核制度

每个人保持对单独一块业务的持久专注，能够解决业务纷繁复杂并不断增长变化的现实情况，保证审核人员对业务和代码有对应的深刻了解，同时借助平台实现信息共享，加强审核人对软件的整体把控。

2）记录评审结果，积累评审经验

通过参照前期评审结果，能够加深理解代码逻辑，掌握代码编写特征，锁定易错点。将经验固化，转化为辅助评审的重要参考。

3）集成辅助工具提高效率

在审核过程中快速查询相关信息，包括数据库结构、开发接口、数据字典、业务需求等定义，节省各种外部工具的查询环节，专注并辅助理解代码逻辑，确认逻辑的一致性。

三、平台建设

3.1 架构设计

代码审计平台聚焦于升级模块代码的查阅、审核，围绕核心目标建立了一套完整的工作流程：从升级包分析、对应代码数据采集、多维度多方法审核、分析数据归档、审核报告生成等过程。实现了审核周边流程的全自动化，以及代码审核的辅助决策支持。

在技术架构方面，平台采用前后台分离模式，web端设计相较于客户端更加具备灵活性和可扩展性。通过与制品库、SVN服务器、Git服务器对接，完成了升级模块及源码的整理归档。采集后的源码进入核心分析阶段，平台提供多种辅助决策、智能化功能配合代码审核人员进行代码审阅、记录、风险判断等工作。

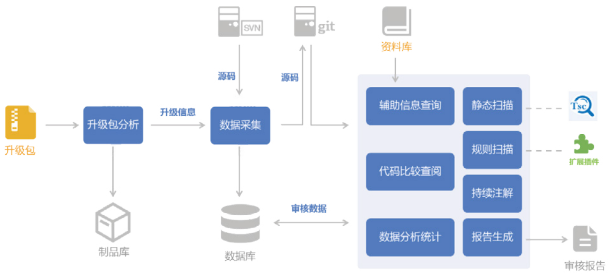


图1：代码审计平台功能架构图

3.2 平台功能

代码审计平台设计的功能围绕实际代码审核需要设计而成，用于辅助人工代码审核，提高审核效率和质量，快速了解版本更迭变化，定位问题，查找潜在风险点等。

图2：代码审计平台功

1）代码比较查阅

代码比较查询用于查找同一模块不同版本之间的代码差异。通过了解代码变更情况，掌握软件升级源码级变动点，以此对照判断代码是否符合变更逻辑，是否满足修改需求。为了提升审核效率，适配审核方法，在源码上进行了智能化处理，增加了特色功能。主要包括基于软件版本的代码合并、便于版本追溯的版本比较、推动代码理解的持续注解等。

2)代码静态扫描

静态检测技术涉及的主要方法包括静态分析和程序验证。其利用二进制比对技术、词法分析、形式化验证技术或者手工测试技术，对被测软件的源程序或二进制代码进行扫描，然后从语法、语义上理解程序的行为，分析程序的特征，找出可能导致程序异常的漏洞。[1]该技术具有简单高效自动化的优势，缺点是仅对代码本身的特征进行检查，对漏洞间复杂的逻辑关联检测力弱，存在大量的误报和漏报。[2]

平台集成静态扫描工具，后台查找相应代码，自动进行扫描，把扫描结果集成显示到了代码显示界面，能够更加直观地定位可疑点。可结合代码特点筛选规则，减少误报条目。

3）智能工具栏

智能工具栏的设计用于在审核过程中，进行相关信息的查询和一些特定功能的快捷应用。主要包括代码注解、数据字典、数据库结构、功能码定义的查询。

其中，代码注解用于查询持续注解的列表内容，能够快速加载当前代码所对应的注释，查看历史备注要点，理解代码内容。数据字典是用于解释代码中字典项的内容。数据库结构是很多上层应用程序设计的基础，平台集成了数据库结构信息。功能码是交易系统中间件基本功能服务单元的接口描述，是构成系统功能的重要组成部分。

智能工具栏后台能够持续更新相关信息，使得在代码审核过程中获取最新的内容。

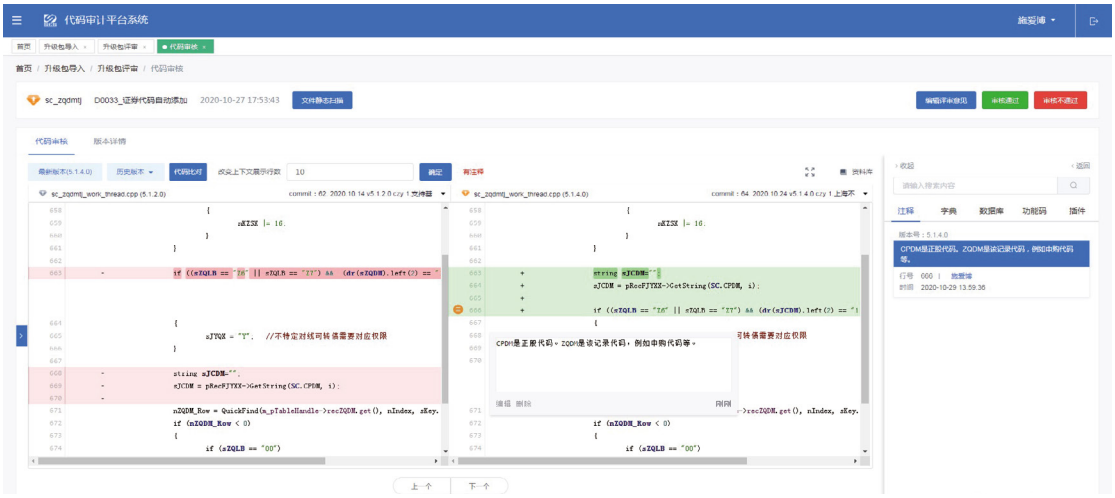


图3：代码审计平台审核界面

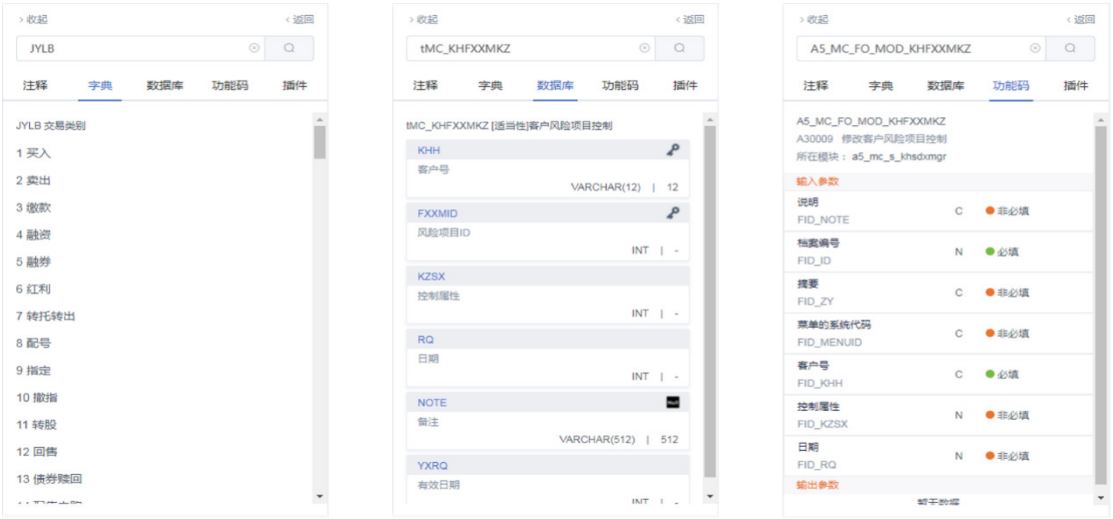


图4：智能工具栏

4）辅助信息查询

辅助信息查询是辅助代码审核过程中进行各种信息查询，包括资料文件的汇总、数据库、功能码等各种信息。相比智能工具栏，这是一个全局的查询入口，可以进行成体系的信息查询。辅助信息查询背后所形成的知识库是长期积累而成。例如资料文件可包括交易所接口文件、成体系的代码说明、功能设计文档等。并可以将文件生成超级链接，插入到代码注解或审核意见中，方便参考。

5）审核报告生成

审核报告集合了审核人员代码审核意见、版本说明、代码变动说明，对每一个升级模块的审核情况进行了描述。审核报告可以生成导出成word文档，用于归档管理。

四、关键技术及创新点

4.1 多维度代码审核体系

核心交易系统的代码质量保障工作有多种方法途径可展开，平台充分融合各种技术，从多个维度对代码进行审核，采取多重措施来保障代码质量。平台设计了静态扫描、人工审查、关联分析、接口审查等。

接口审查可以确保代码接口符合预定义的标准和规范，确保代码质量和稳定性。并且可以检查代码间的兼容性和整合性，防止引入新的错误，是保证代码软件质量的关键步骤。

4.2 插件式自动化审查

除了代码语义层面的静态扫描之外，平台通过插件加强自动化审查。插件可以获取代码内容，结合后台知识库，对代码进行分析。可以从业务层面出发，有针对性地检查代码风险点。例如，对代码中的功能码调用入参、出参进行扫描，比较接口定义，判断参数类型、必传项等是否一致，减少代码错误的发生。

插件式设计可以灵活地扩展平台自动化审查能力，从各个角度多重叠加，降低代码漏洞风险。

4.3 代码结构关联性分析

核心交易系统作为一个大型复杂系统，采用了模块化架构设计，主要将系统划分为菜单模块、中间件模块、数据库表及存储过程。从图5的模型可以了解到整个系统模块间的分层调用关系：菜单调用中间件功能，中间件既可以调用中间件其他功能，也可以访问数据库，同时一个菜单功能封装在一个菜单动态库中，而一个中间件动态库则封装了多个中间件功能。

截至最新的统计结果，整个集中交易系统包含了菜单1628个，中间件功能4456个（封装在144个动态库中），涉及到的数据库表或存储过程636个，存在各模块直接的调用关系18170条。面对拥有如此庞大复杂功能的系统，理清其内部的逻辑关系是十分重要且关键的工作。为此，可以将菜单、中间件、数据库、动态库这些元素视作不同类型的功能节点，根据它们之间的调用、包含关系，形成了类似图状的“功能网络”拓扑结构。

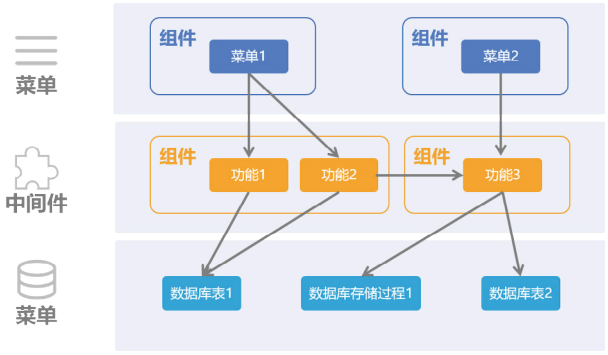


图5：系统模块化设计

平台利用代码中功能调用和数据库访问的基本特征，解析抽取了用于构建“功能网络”的各种要素及要素之间的关系，并最终存储到了图数据库中，进而可以快速地挖掘某一特定功能的网络关系。

除了一些基础的关系查询，可以从整体网络的角度做一些数据统计。例如进行模块扇入扇出、计算网络直径、紧密中心性、介数中心性等网络分析，后续的工作中将对“功能网络”进行更多方面的网络分析，以及通过社区发现等算法查找关联度紧密的菜单、中间件功能或数据库表等，从而理清系统功能的逻辑脉络，为系统测试、代码审核提供参考。

4.4 智能化代码辅助理解

根据数据分析发现，从代码行数来看，超过400行的代码审核，缺陷发现率会急剧下降；从审核速度来看，超过500行/小时后，审核质量也会大大降低，一个高质量的代码审核最好控制在一个小时以内。[5]具体数据可能因人而异，但总体情况确实如此，评审人对持续代码审核工作会产生效率下降的状况。[6]但券商端代码审核工作又比较集中，为了解决矛盾，需要辅助提升审核人员对代码的理解能力。

平台集合了持续注解、智能工具栏等信息参考工具，为审核人员在审核过程中提供了强大的信息参考支撑。在代码审核过程中，尝试引入该工具辅助代码理解。经过试验，AIGC（生成式人工智能）可以有效地添加用于

辅助理解的注释，显示出了AI对于编程以及代码审核这一领域的能力，可以视为代码审核未来发展的一个方向。从平台角度考虑，多了一道智能审核渠道。但是依据目前的AI发展现状，尚不能完全依赖AI进行注释，会有一定的错误率。同时，AIGC技术可能涉及侵犯知识产权，存在代码隐私泄露等问题，在使用过程中还需加以甄别，目前的AI发展对代码理解仅可作为参考。代码审计平台将保持一个开放的态度，积极拥抱探索新技术所带来的能力，促进代码质量保障发展。

4.5 审核人员智能推荐

核心交易系统软件规模庞大，涉及大量的代码，需要多人审核。不同审核人员的经验以及对功能点的了解程度不尽相同，需要合理地指派审核人员审核代码。代码审核人员指派的方法主要有：

1） 基于规则的方法

通过设置规则，系统能够自动根据代码的特征和审核人员的专业知识，为代码选择合适的审核人员。代码审计平台能够根据软件模块功能，预先设置指定相应的审核人员，当某一软件模块需要审核，则自动指派给对应的人员。

2） 基于审核历史的方法。

通过分析代码的审核历史情况，系统能够自动学习审核人员的指定变化，根据审核变化情况选择合适的审核人员。系统会根据历史审核时间远近、审核次数等参数，加权重计算审核匹配度。

3） 基于关联模型的方法

代码审计平台参考代码结构构建了“功能网络”，查找升级包中各升级模块之间关联关系，通过节点间路径长度确定模块关联升级紧密度，以此将关联度紧密的模块指定给同一审核人，提高代码审核的整体质量和效率。

五、实际应用

自代码审计平台上线投入使用以来，已累计审核升级包161个，涉及软件模块865个，功能模块4869个，承接了大量的代码审核工作。2023年单月最高审核软件升级包18个，有效应对高频升级下的质量把控问题。在审核过程中，发现接口应用错误、代码笔误、代码逻辑错误、升级包功能模块版本不一致、内存泄漏、编码不规范等情况30多次，有效减少软件升级的潜在风险，提高了质量把控能力和效率。

重新构建了与版本匹配的代码库，形成了审核意见以及注解，保持最新的数据库表结构定义以及功能接口定义。增加了强化审核的插件，查找接口应用及文档错误50多项。

六、总结

东吴证券在坚持自身系统软件质量保障方针的情况下，总结代码审核经验，提出适应券商的代码审核思路，在体系架构上建设而成的代码审计平台能够有效地进行审核工作的实施，并在测试、运维环境提供代码审查能力。

代码审核平台针对券商进行供应商代码审核的实际情况，解决过程中各种问题，符合特定的应用场景，是代码质量保障体系重要的工具平台。平台按照实际审核过程进行逻辑设计，引入了自动化、智能化技术，形成了提升券商代码审核质效的有效方法途径，并基于归集数据进行分析、反馈完善，对系统代码质量保障起到了一个重要作用。

未来，将进一步完善核心交易系统代码质量保障体系，优化代码审计平台建设，引入智能化、数智化能力，挖掘新的行之有效的代码审核模式。同时丰富扫描插件，多角度保障代码质量。

参考文献：

[1]罗琴灵. 基于静态检测的代码审计技术研究[D]. 贵州大学, 2015.
[2]周景科. 专业领域软件的代码审查方法研究[J]. 软件可靠性与评测技术, 2019, 29 (3): 1-7.
[3]李晓峰, 刘宏伟, 王建民. 基于静态分析的软件安全漏洞检测技术综述[J]. 计算机科学与探索, 2019, 13(2): 191-210.
[4]张志强, 张亚军. 基于范根检查法的软件代码审查方法[J]. 计算机工程与设计, 2017, 38(6): 1534-1538.
[5]蒋春华. 软件代码审查在敏捷开发中的应用研究[D]. 南京理工大学, 2016.
[6]王慧娜. 基于结对编程的软件质量保证方法研究[D]. 南京理工大学, 2015.

基于eBPF技术的无侵入云原生可观测性系统研究

曾东明、沙烈宝、段苏隆、李银鹰 | 国投证券股份有限公司 | Email：zengdm@essence.com.cn

摘　要： 随着证券行业数字化转型的加速，云原生技术如容器、微服务架构等已经广泛落地应用。然而，这些新技术带来的技术复杂性，也给业务监控和问题诊断带来了新的挑战。本文总结了基于eBPF技术进行可观测性系统构建的研究和探索，主要包括以下几个部分：首先，介绍了云原生技术下可观测性技术的挑战和研究目的；接着，简要介绍了eBPF技术，包括原理和开发流程；然后，详细阐述了采用eBPF技术进行相关指标的采集和数据分析的方法和实践；最后，对整个研究进行总结和展望。通过本文的探索，希望能为证券行业提供新的监控和诊断思路。

关键词： eBPF；可观测性；无侵入式； Kubernetes

一、研究背景和目的

目前云原生技术已经在金融行业内广泛落地，以容器和Kubernetes编排为底座的容器云平台已经成了数字化转型的行业标配，并且还在逐步扩大云原生技术应用范围，微服务架构、中间件容器化也都在逐步落地过程中。这些新的技术趋势带来了资源利用率的提升，更高的开发发布效率，更好的运维弹性，但是这些优点都是以服务整体技术栈变得更加复杂为代价的。因此，在新的云原生技术栈下，传统的监控、排障方法已经无法满足需求，更复杂的技术栈对于业务提出了更高的可观测需求。

eBPF技术近年来成为业界一个备受关注的用于可观测性的一个解决方案。它能在内核中动态执行代码，深度监测内核和业务逻辑，无需侵入业务代码。相较传统的监控手段，eBPF以其无侵入性大大降低了业务接入难度。同时，对于一些无法修改源码的系统，通过eBPF在内核增加的探针能力，同样可以获取到服务之间互相访问的可观测数据。

基于上述分析，希望通过本次基于eBPF技术的可观测性系统的研究成果，实现更好的系统和应用程序的可观测性，降低运维成本和风险；同时，也可以为证券行业提供一种高效、可靠的可观测性解决方案，促进eBPF技术在证券行业的推广和应用。

二、eBPF技术简介

2.1 概述

eBPF技术是一种源于传统BPF（Berkeley Packet Filter）技术的新型可编程内核技术，是Linux内核的一项革命性技术。它提供了能够在运行时动态安全地修改Linux内核行为的机制，而无需改变内核代码或者额外加载内核模块。并且整个eBPF程序会运行在沙箱中，不会对内核的稳定性造成影响。

目前eBPF技术目前主要用于三个方面：可观测性数据采集、网络技术增强以及安全审计和安全阻断。在本次研究课题中，使用eBPF技术做可观测性数据的采集，并基于采集的数据，构建可观测系统。

2.2 技术原理

1） eBPF虚拟机及指令集

在Linux内核中，eBPF虚拟机用于执行eBPF指令/字节码。虚拟机内建有一个校验器，用以确保eBPF字节码的安全性，防止可能导致内核崩溃的操作，例如死循环、无效跳转和内存越界访问等。eBPF虚拟机基于寄存器架构，其指令集包含常见的算术逻辑运算、位操作和跳转等指令。

2） eBPF程序类型

eBPF程序都是内核事件触发的，不同的hook点能够执行不同类型的eBPF程序，且从hook点获取的入参/出参也各不相同。在可观测性领域，本文关注的主要是kprobe类型，即BPF_PROG_TYPE_KPROBE。

3） eBPF程序互操作与数据交互

eBPF提供了一组预定义的helper函数和一组特定的内核函数，实现与内核的安全交互。在可观测性领域，通常需要将eBPF程序采集到的数据进行统计和记录，然后导出以供外部程序进一步处理和分析。为处理结构化数据，eBPF提供了统一的Map能力的抽象，实现用户态控制进程与eBPF执行代码之间的双向通信。

4） eBPF程序开发脚手架

eBPF程序的开发和加载，需要一系列的脚手架工具，包括编译、加载、内核交互获取数据等。目前业内有传统的bcc以及比较新的libbpf。本文选择了bcc作为开发脚手架框架。

2.3 可观测性eBPF程序的开发和运行流程

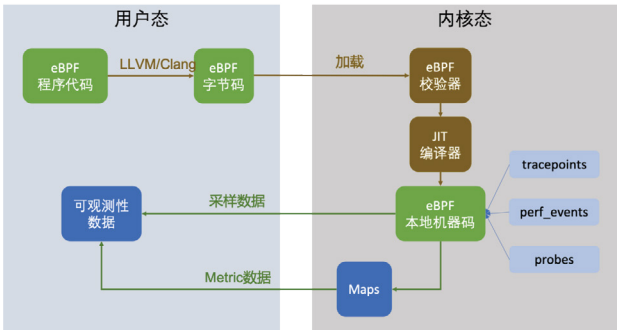


图1：开发和运行eBPF程序大致流程

如上图所示，开发者在完成eBPF程序代码编写后，通过Clang编译器将其编译成字节码。这些字节码随后会被加载到内核中。在加载过程中，内核中的eBPF校验器会对执行指令的安全性进行检查。校验通过后，字节码将被即时编译（JIT），并在事件触发时被调用。当相关的eBPF代码因事件触发而被执行时，其逻辑可以执行数据采集、统计、采样和记录等操作。在用户态，通过查询Map来获取eBPF采集到的数据，从而进一步处理和分析这些信息。

三、研发实践

3.1 需求分析

基础设施的三大基石是计算、存储和网络，其中网络问题最为常见，也是分布式系统故障排查的难点。除了网络问题之外，其余的问题通常被归类为系统问题。本文的需求是识别那些需要增强采集的网络和操作系统指标。下面是一张关于可观测数据项的整体概览图。

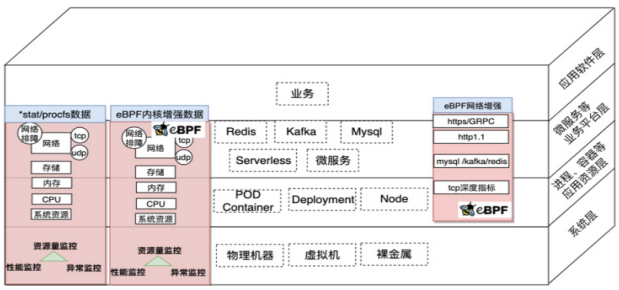


图2：可观测数据项的概览

在典型的容器网络方案中，当访问Kubernetes集群外部网络时，通常在节点出口进行SNAT（源地址网络地址转换）。这种方法会导致在集群外部网络出现异常时，难以精确定位到具体的业务Pod。此外，在进行性能分析时，业务通常需要细粒度的点对点网络监控。常用的解决方案是在应用层使用APM（应用性能管理）工具，通过在业务代码中埋点实现对业务访问情况的监控。然而，这种方法对业务具有侵入性，并增加了业务的复杂性。

综上所述，一个高效的网络可观测性系统应满足以下需求：首先，数据采集需要细粒度，收集的数据应能追溯到最原始的业务容器，并实现不同层级的数据聚合。其次，网络指标应得到增强，相比传统的网络监控，系统应额外采集常见错误和异常场景的指标。此外，网络协议解析应做到无侵入性，应利用eBPF技术对HTTP、Redis、Kafka等网络协议进行解析，而无需修改业务代码。

2） 系统观测需求

系统问题主要可以分为以下几类：CPU调度、内存分配管理、文件操作及IO访问。对于CPU调度，随着现代机器中的CPU核数逐渐增加，CPU架构日趋复杂，需要对CPU的调度情况进行监控。在内存分配管理方面，需要对内存缓存的使用情况、内存申请延迟以及内存回收过程进行细粒度的监测。对于文件操作，需要采集相关文件操作的延迟数据。在容器场景下，还需特别监控两个特定的文件访问延迟：overlayfs的copy up延迟和PVC远程文件系统访问延迟。综上所述，可观测性系统在系统层面的监控应涵盖CPU调度、内存分配管理、文件操作及IO引起的问题，以帮助业务迅速定位问题。

3.2 整体系统设计

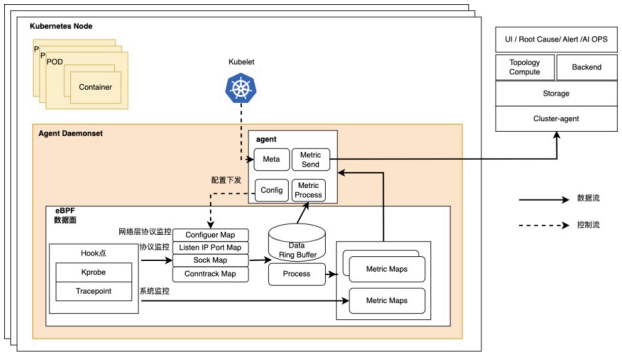


图3：采集设计框图

上图是容器云场景下的整体数据采集设计框图：

·Agent、Cluster、Storage是数据流的核心组成部分；

·Agent以Daemonset的方式部署到每台机器上，负

采集的指标	hook 的内核函数
tcp 带宽和收发数据包统计	tcp_sendmsg/tcp_recvmsg
tcp 延迟和连接和重传	tcp_set_state/ tcp_update_pacing_rate / tcp_retransmit_skb / retransmit_skb
tcp 监听 ip 和 port 获取	inet_csk_accept / inet_csk_listen_stop

表： tcp采集的指标以及对应的hook点

责采集指标以及监听的 IP 和端口的元信息；

·Cluster Agent以集群的方式部署，负责整合指标和监听的IP和端口元信息，形成容器到容器的访问Metric；

·Storage负责存储Metric；

·Topology Compute是一个存储的消费端，负责消费Metric产生拓扑图数据；

3.3 技术实现

网络指标包含四层网络指标和七层应用层协议指标。四层网络指标涵盖了TCP的收发带宽、TCP RTT平均和最大延迟以及TCP异常关闭连接等。而七层应用层指标涉及HTTP、Redis、Kafka、MySQL等协议的请求、错误和延迟指标。四层指标主要用于基础设施的网络监控。通过细粒度的网络指标可以感知容器之间互访的网络质量，进行网络带宽审计，识别并绘制出容器之间的业务依赖关系，形成业务依赖拓扑。此外，这些指标还可以记录业务容器的网络访问痕迹，帮助发现潜在的安全风险并进行影响评估。七层应用层指标则用于业务监控，通常用于观测业务的稳定性以及无侵入的业务调用拓扑分析。

1）四层指标采集实现

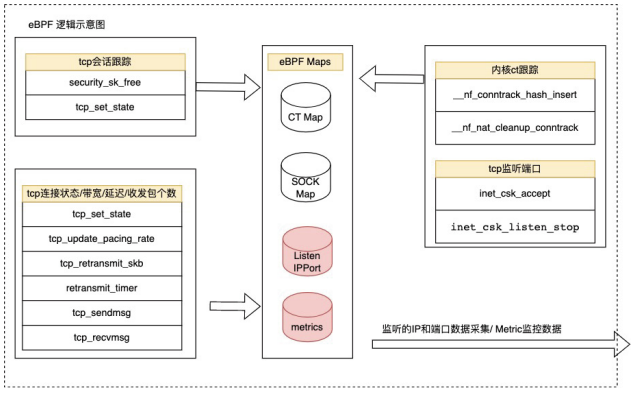


图4：四层指标采集逻辑示意图

Agent负责从内核的eBPF程序的Metric Map中提取具体的TCP指标，以及监听的IP和端口等原始信息。这些原始数据需要进一步与业务数据进行关联处理。在原始数据采集过程中，涉及的技术要点如下：

·TCP连接会话跟踪。主要维护一个以socket为键的Map，用于记录与socket对应的监控元数据。

·内核conntrack跟踪。主要用于将经过NAT转换后的IP地址和端口还原为真实的地址和端口，确保数据正确关联到相应的业务。

·指标跟踪。Metric Map主要用于记录点对点粒度的网络监控数据，包括各种TCP指标。

·TCP端口监听跟踪。主要用来记录机器监听的IP和端口以及容器的对应关系，这有助于更好地进行元数据分析，包括分辨是否为客户端或服务端。

指标采集需要在不同的hook点上实现，具体的指标项和对应的hook点如下表所示。采集到的原始数据将由Agent根据业务元信息进行进一步增强，添加更多的业务信息到指标的标签中，以方便业务的检索和分析。

2）七层指标采集实现

针对不同场景下的需求不同，七层应用协议指标采集提供了精简模式和精细模式。精简模式聚焦于应用粒度的请求QPS、错误率和请求延时等指标，不对协议进行更细粒度的分析。精细模式则将这些指标精确到应用的请求路径粒度。例如，HTTP协议精细到URL级别，MySQL协议精确到SQL语句模板级别。如下图所示，七层应用协议指标采集和四层TCP指标采集的实现方式大致相同，差异点在于七层协议指标计算的hook点在tcp_sendmsg和tcp_recvmsg这两个地方。此外，精细模式的额外协议解析将消耗更多的资源。

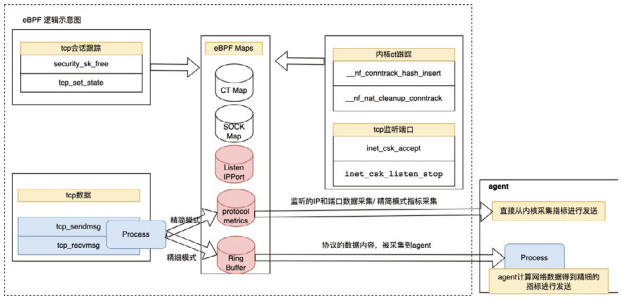


图5：七层指标采集逻辑示意图

含义	hook 的内核函数	作用
cpu 饱和度	ttwu_do_wakeup、wake_up_new_task、finish_task_switch	显示容器在 CPU 调度队列中排队的时间。某些进程在 CPU 排队时间过长，进而引发业务的延迟增大
内存饱和度	alloc_page	用来判断内存申请延迟。如果内存压力比较大，申请内存会触发内存回收，进而引发业务出现延迟抖动
远程文件系统 ceph	ceph_read_iter、ceph_write_iter、ceph_open、ceph_mdsc_do_request	该指标用来判断 ceph 的读写 BPS 和 IOPS，用来从客户端发现远程存储的饱和情况
vfs 虚拟文件系统监控	vfs_read、vfs_write、vfs_unlink、vfs_create、vfs_open	用来判断容器对文件系统 fd 的读写行为，判断容器业务是否是由于访问文件系统导致的延迟
page cache 访问命中率	add_to_page_cache_lru、mark_page_accessed、account_page_dirtied、mark_buffer_dirty	应用访问 pagecache 情况，定位由于 Cache 不命中，导致的业务性能问题
dirty page 监控	mark_buffer_dirty	脏页产生速度监控
tcp rst 监控	tcp_syn_ack_timeout、tcp_v4_send_reset、tcp_reset、tcp_send_active_reset	通过监控 tcp rst 的发送和接收，帮助定位与 tcp reset 相关的问题
fd/pid	do_sys_open、do_fork	通过监控 fd/pid 用来发现系统资源申请失败，这种往往是系统配置不合理导致的问题
磁盘延时	block:block_rq_issue、block:block_rq_complete	每个本地磁盘读写延时，用来判断磁盘的故障情况，业务的磁盘读写延迟等

表2：主要的系统深度指标

3）业务访问拓扑实现

如果业务规模较大，将会产生大量指标数据。在构建拓扑时通常需要进行各种指标聚合。因此，如果每次都实时从时序数据库中获取数据，将会给时序数据库造成较大压力。为解决上述问题，本文提出了一个解决方法，即通过预先计算并缓存结果，来有效缓解压力。如图6所示，图计算服务会周期性地从指标存储中获取所有指标进行计算，然后将计算好的图形结构存储到缓存中。图计算服务可对外提供API，用于实现拓扑图的过滤、筛选和查询限制等功能。经过实践证明，拓扑图在包含10,000个图节点和100,000个图线的情况下，可以实现秒级响应，大幅提升用户体验。

4） 系统指标采集

在容器原生采集指标的基础上，本文额外采集大量的系统深度指标，这些指标在业务层面具有较高的价值，表2是主要采集的系统深度指标的说明：

3.4 性能影响分析

eBPF程序在内核中基于事件触发运行，不可避免地带来一些开销，从而影响业务性能。为了深入了解eBPF程序对业务性能的影响，本文针对多种场景进行了详细测试。根据测试结果，通常情况下，业务延迟增加在1%-10%之间，CPU的额外消耗在5%-15%之间。当然，如果开启更精细的采集，开销也会相应增加。

以下是一个具体测试场景的数据，供参考。该场景选择了最常见的Nginx服务，使用wrk工具模拟不同的请求压力情景，对比了开启和关闭eBPF探针情况下，Nginx服务的请求延迟和整机CPU平均使用率的变化情况。

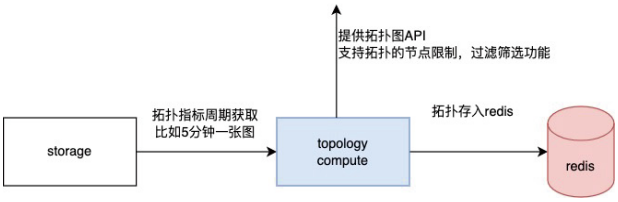


图6：图计算服务说明

图9：网络层监控指标界面

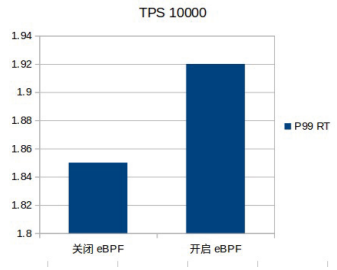


图 7：开关eBPF的TPS性能对比测试

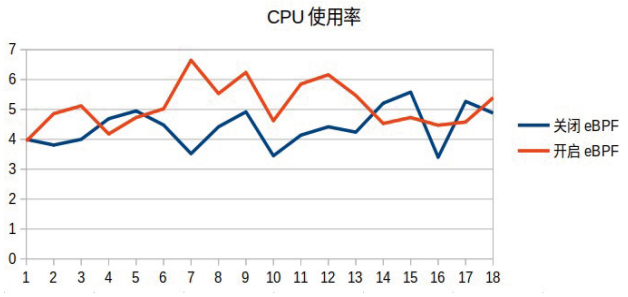


图8：开关eBPF的CPU使用率性能对比测试

从测试数据可以看出，P99时延从1.85毫秒增加到1.92毫秒，时延增加约3.7%。整机CPU平均使用率从4.4%上升到5.1%，大约增加了15%的CPU消耗。由于Nginx本身对时延敏感性较高，这些数据可以作为其他业务评估时的参考基础。实际上，对于普通业务，影响通常不会达到如此显著的程度。

3.5 应用落地效果展示

上图展示了一个网络层监控指标，可以帮助业务进行实时监控、可视化和优化网络性能，其带来的价值包括：

·提高业务性能：通过对监控数据的分析利用，业务可以用来优化网络连接，提高业务性能。

·减少故障时间：可以帮助业务快速发现和解决网络问题，减少故障时间，提高业务连续性。

·优化资源利用：通过分析网络流量数据，可以帮助业务优化资源分配，提高网络设备的利用率。

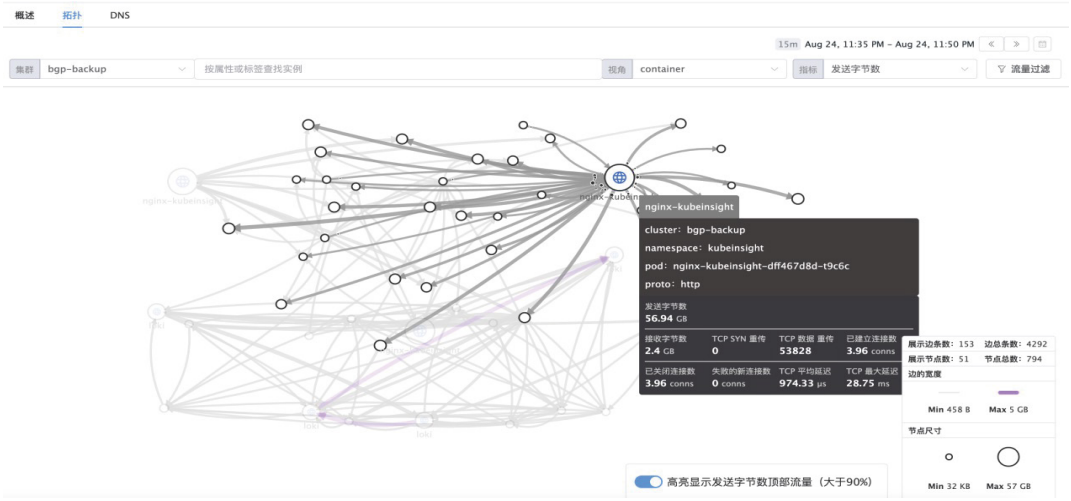


图10：网络拓扑界面

图11：应用层指标界面

·提高安全性： 可以监控网络流量中的异常行为，帮助业务及时发现和应对安全威胁，提高网络安全性。

上图展示了一个网络拓扑，旨在基于收集到的网络流量数据显示服务和应用之间的依赖关系和通信状况。对业务而言，其带来的价值有：

·提高服务可视化：以图形化的方式展示服务和应用之间的依赖关系，帮助业务直观地了解服务间的通信状况。

·故障排查：可以实时显示服务之间的通信性能指标，帮助业务快速定位故障服务和连接。

·性能优化：通过分析网络拓扑中的流量和性能数据，业务可以发现服务间的瓶颈和优化点，从而优化服务性能。

·规划与扩展：可以帮助业务了解现有服务的规模和复杂性，为服务规划和扩展提供有价值的参考信息。

·提高协作效率：网络拓扑可以作为团队间沟通的基础，帮助开发人员、运维人员和管理者更好地协作，共同解决服务问题。

·安全管理：网络拓扑可以帮助业务发现未授权的服务和连接，提高服务安全性。

上图展示了对应用层指标进行无侵入式采集，成功实现了指标采集与图表绘制。对于一些老旧系统，无侵入式采集方式通常是唯一可行的应用访问数据观测手段。

四、总结与展望

4.1 总结

本文顺应业界趋势，探讨了使用eBPF技术来增强业务可观测性的方法。通过对网络和系统的无侵入式监测，获得了细粒度和深度的指标，并构建了业务访问拓扑。此外，本文对eBPF技术在指标采集过程中对性能的影响进行了定量分析，为大规模部署基于eBPF的可观测系统提供了可行性研究和宝贵经验。本文的实践表明，基于eBPF的可观测技术是对传统监控手段的有力补充，有助于更精准地定位和解决问题。然而，由于eBPF技术涉及对内核的增强，对于一些对延迟敏感且请求量大的业务，仍可能产生一定的性能影响。

4.2 展望

首先，当前可观测性系统在特定内核版本和操作系统上使用bcc作为开发框架。鉴于移植性和内核版本/架构的适配问题，未来计划将开发框架从bcc转换到libbpf，以实现“一次编译，到处运行”的目标。

其次，目前的观测数据采集中大量使用了kprobe和kretprobe，但在测试中发现对业务性能造成了一定影响。因此，需要持续优化，包括重新寻找更合适的观测点和优化代码逻辑，以提升系统性能。

最后，eBPF采集到的观测数据只是可观测性系统的一部分，而且eBPF技术本身具有广泛的应用潜力。因此，可以结合其他可观测数据，打造一个统一的观测系统；同时，eBPF技术的使用也可以扩展到增强网络功能、安全阻断等多个领域，从而发挥更大的价值。

参考文献：

[1]刘畅.(2020).基于 eBPF 的容器网络可观测性方法与实践.(硕士论文).浙江大学

[2]高巍. (2022). 基于操作系统 eBPF 在云原生环境下的技术研究. 电子技术与软件工程.

[3]Cassagnes, C., Trestioreanu, L., Joly, C., & State, R. (2020, April). The rise of eBPF for non-intrusive performance monitoring. In NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium (pp. 1-7). IEEE.

[4]Miano, S., Chen, X., Basat, R. B., & Antichi, G. (2023). Fast In-kernel Traffic Sketching in eBPF. ACM SIGCOMM Computer Communication Review, 53(1), 3-13.

[5]https://blogs.oracle.com/linux/post/from-kernel-to-user-space-tracing

[6]https://ebpf.io/what-is-ebpf/

持续测试助力业务价值高质量交付

陈洪炎、吴鑫、屠栽镝、胡跟旺、徐子祺 | 上交所技术有限责任公司
E-mail: hongyanchen@sse.com.cn

摘　要： 随着国家“十四五”规划的发布和2035年远景目标的明确，金融科技行业迎来了前所未有的机遇和挑战。中国信通院牵头制定的《研发运营一体化（DevOps）能力成熟度模型》覆盖了需求、开发、测试、运营等软件全生命周期，为证券期货业机构提供了研运能力提质增效的指导方向，持续测试是DevOps体系的重要一环，秉承了DevOps的思想和理念，以实现价值流顺畅、快速流转为目的，形成了适用于测试领域的方法论和最佳实践。本文从持续测试的产生、价值、体系和实践等维度展开探索和研究，以期为保障业务价值高质量交付提供参考。

关键词： 持续测试；价值流动；度量反馈；持续改进；DevOps

一、引言

近些年，证券期货行业运用金融科技赋能业务发展，实现业务创收和发展降本增效已经成为行业关注的焦点。如何在满足研发运营过程安全、合规的前提下，通过提升研发交付效率，快速交付有价值的产品和服务变得越来越重要。目前证券期货业在数智化转型中面临以下痛点：1）技术债较重，个性化需求难以高效满足；2）迫于业务压力上线，系统建设缺乏整体规划；3）系统架构复杂，维护成本高；4）系统建设周期长，失败成本高。本文旨在帮助行业机构在研发IT软件及相关服务过程中，将需求、设计、开发、测试和运营等关键环节统筹考虑，落地覆盖端到端软件研发生命周期全流程持续测试，助力实现业务价值高质量交付。

二、持续测试的产生与价值

2.1 持续测试的产生

软件研发相关的测试活动主要经历了瀑布型测试、敏捷型测试和持续测试。随着DevOps思想不断发展和传播，持续交付提倡价值流动，测试与持续交付流水线相互融合，测试不断向需求侧、开发侧和运维侧移动，形成覆盖软件研发全生命周期的持续测试闭环。

2.2 持续测试的价值

1.持续反馈

持续反馈是构成持续测试闭环的重要阶段，是对于现有阶段的测试结果、测试流程、测试质量和效率的反馈，有助于更早地找到问题而减少修复问题的成本。

2.持续改进

持续改进是通过对产品、系统、流程等进行持续不断地测试以发现潜在的问题和改进点。持续测试避免了问题在后期积累，促进软件研发过程的持续改进。

3.提升研发效能

持续测试帮助组织通过提高测试效率，增强交付能力，保障交付质量，支撑组织实现业务价值的持续交付，提升整体研发效能。

4.助力数智化转型

持续测试能帮助研发团队的测试转型，软件全生命周期活动均需要开展测试活动。持续测试强调自动化测试的能力，能帮助组织将传统手工测试转型为自动化测试，持续测试支持了预交付流水线的协同，使得全流水线中不同测试活动之间的衔接变得更加紧密。

5.全面质量管理

基于全面质量管理的思想，建立稳定、高可用性的系统架构，通过持续测试加强质量管理和业务连续性管理，保障业务持续稳定运行。

三、持续测试体系

持续测试是在整个软件研发生命周期中持续执行测试以提供有关被测系统质量高效反馈的过程，是软件交付流水线中的一种可以随时开展且具有连续性的测试活动。持续测试体系包含持续测试组织与文化、持续测试能力和持续测试成熟度模型。

3.1 持续测试组织与文化

持续测试鼓励测试活动尽早介入到软件研发过程中，及早发现质量与业务风险，测试尽早介入可以提前发现并纠正在需求设计、产品规划和架构设计中的问题，持续测试能有效降低项目在快速增长期产生的风险，减少技术债务的堆积，持续测试倡导流动文化、反馈文化和合作型组织。

传统研发团队按照职能和专业划分团队职责。业务人员、产品经理、设计人员、开发，测试，运维，运营等均隶属于不同部门。部门之间“沟通壁垒”问题普遍存在，项目研发过程中存在较多无效等待时间。

持续测试提倡合作型组织，将组织划分为管理部落，产品部落和研发部落（图1）。管理部落聚焦战略目标，业务定位，推进并统筹整体事项。产品部落服务于业务客户，作为业务和研发人员的枢纽促进高效协作，实现业务价值高质量交付。研发部落负责实现产品研发并完成产品的交付。

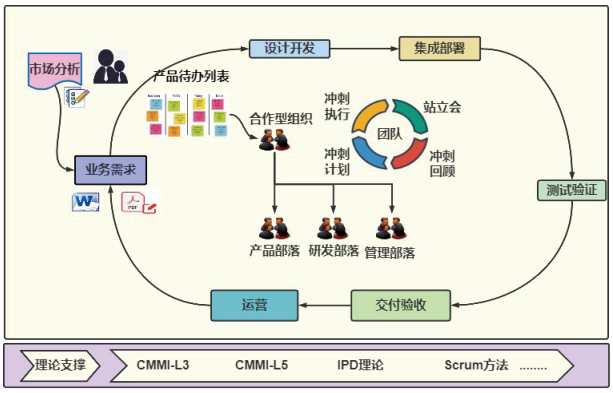


图1：持续测试合作型组织的工作模式

3.2 持续测试工作流程

持续测试倡导测试人员在软件研发过程中持续参与需求分析、需求评审、设计开发评审、代码评审、测试设计、测试用例/脚本开发、冒烟测试、集成测试、验收测试、线上测试等活动，持续测试使得产品经理、开发、测试和运维之间有效协作，极大减少沟通成本和无效等待时间。持续测试在实际过程中的工作流程如图2所示。

3.3 持续测试成熟度模型

持续测试能力成熟度模型（图3）分为过程域、改进域、平台能力域和度量域四个维度。

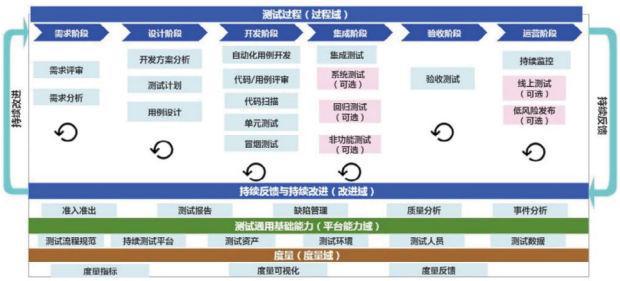


图3：持续测试能力成熟度模型

续测试能力成熟度分为1-5级（表1），主要分级指标包括规范制度、测试左移、测试右移、自动化测试、效能度量、先进技术、智能化测试、过程反馈与改进、精准测试、缺陷预防、辅助决策、科研能力等。持续测试能力成熟度模型设计了递进式的要求，结合最佳实践与先进技术发展趋势，能够用于评估组织各阶段的持续测试能力。

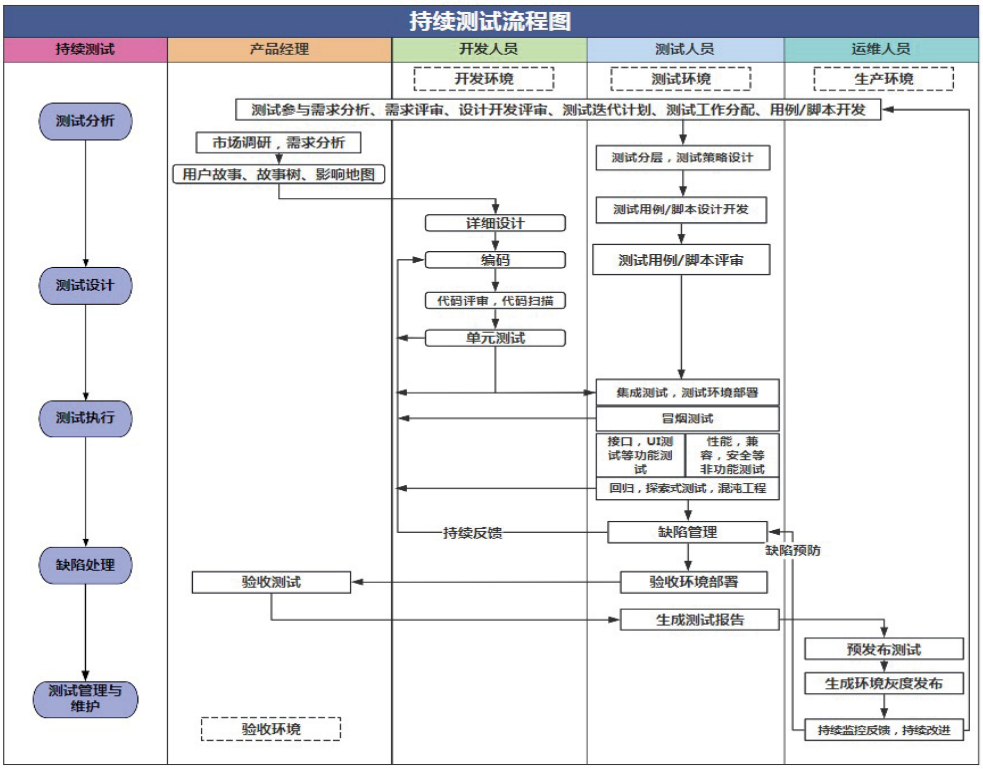


图2：持续测试工作流程

级别	级别名称	级别定义
一级	初始级	组织具备基本测试环节和制度，开展简单持续测试实践。
二级	基础级	组织遵循统一规范和流程，测试流程各阶段高度实现测试自动化，开展测试质量效率的监控和度量，持续测试对交付的质量提升起到一定程度的影响，形成基本持续测试闭环。
三级	全面级	组织初步实现了软件全生命周期的持续测试，测试全面左移和初步右移，具备缺陷预防、精准测试、模糊测试能力，具有全流程多维度度量和质量评估体系，实现辅助决策。
四级	优秀级	组织实现了持续测试平台，可以随时开展自动化测试，测试全面左移和右移，支持全流程持续交付，测试初步智能化，实现测试全流程持续反馈和优化。能向行业输出持续测试的最佳实践。
五级	卓越级	组织实现全流程精准测试、模糊测试和增量持续交付，组织能够实现全生命周期的自主测试，持续推动 DevOps 全流程提质增效，实现精准辅助决策。能参与行业持续测试标准制定。

表1： 持续测试能力成熟度分级表

四、持续测试实践

持续测试实践依据持续测试能力成熟度模型分为过程域实践、改进域实践、平台能力域实践和度量域实践。过程实践体现持续测试在软件生命周期全流程参与的体现。改进域实践是持续测试实践的优秀结果。平台能力域实践是持续测试的基础能力建设。度量域实践用于检验持续测试实践的结果。

4.1 过程实践域

4.1.1 需求阶段测试

精益需求分析通过分析需求与功能点估算对用户故事进行细粒度的拆分，实现业务语言与技术语言之间的转化，用领域模型等方法进行业务建模，帮助完成良好的技术实现。实例化需求是需求分析拆解的一种方法，用具体的例子或可视化的效果将用户故事描述得更明确。测试通过实例化后的故事点进行测试场景设计和测试用例开发，并估算测试工作量，形成测试迭代计划。

持续测试倡导以用户故事方式管理需求（图4），便于后续每个需求的实现和验收。一般将精益需求分析和实例化需求后形成的用户故事按照目标进行分层管理，形成树形结构的故事树，故事树将离散的需求组合起来，使需求汇聚成一个完整的架构，团队能获得项目的全局观和产品研究的方向感，并能清晰地看到产品发展路线图。

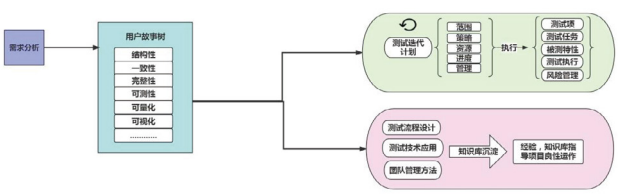


图4：需求阶段持续测试活动

4.1.2设计阶段测试

设计阶段的持续测试一般与实际工作场景相互结合，主要用于对齐合作型组织中各角色认知的一切活动，包括对齐开发方案、测试用例、测试计划等相关活动。通过各角色对齐认知后的产出判定该环节的成熟度，一般衡量设计阶段持续测试能力的成熟度的指标有：测试用例对需求、代码的覆盖情况、开发方案与需求的吻合度、单测断言密度、用例有效性等。

4.1.3开发阶段测试

开发阶段的持续测试通过频繁且精准的单元测试和冒烟测试，有效提高软件编码的质量，逐渐实现持续测试驱动开发。开发阶段的持续测试活动包括代码扫描、代码质量管理、代码分支管理、版本管理、架构设计评审、代码评审，配置管理等，这些持续测试活动能帮助尽早发现研发早期阶段中引入的故障。

4.1.4 集成阶段测试

集成阶段测试是在软件作为制品交付并部署到测试环境后，进行针对性测试、集成测试、系统测试、非功能测试、回归测试等一系列测试活动（图5）。

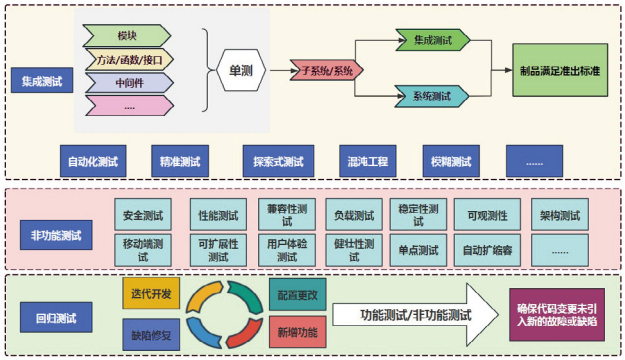


图5： 集成阶段测试阶段的测试活动

1.自动化测试

持续测试提倡应自动化均自动化，包括自动生成用例以及自动提交缺陷，下钻分析等，自动化测试能节省测试执行的时间成本和人力成本，快速有效地反馈产品质量问题，自动化测试能记录故障发生的日志和截图，减少“人为”因素带来的测试漏测和结果误判，节省开发和测试人员沟通成本，提高测试精准度。自动化框架的选择在一定程度上决定了自动化测试后续的维护成本和在项目中产生的价值。自动化框架的选择需基于团队的特点和能力，提倡根据自身业务特点定制开发自动化测试平台，解决通用自动化框架无法满足的痛点和不足。

2.精准测试

精准测试建立在业务功能点（测试用例）和业务代码相互关联的基础之上，获取功能点的覆盖率，进行测试精准覆盖和缺陷精准定位。精准测试通过实时感知代码变动，实现代码变动分析，准确定位代码变动范围和变动影响范围，提供深入准确的测试决策分析依据，从而准确确定测试范围。精准测试通过对代码的量化分析，得到测试用例执行后的代码变更的覆盖率，并以此为依据促进补充和完善测试用例。精准测试依据代码的变更范围，并追踪服务和方法调用链路关系，获得明确的测试范围。

3.性能测试

证券期货业的信息系统因业务的特殊属性，具有数据规模大、耦合性强、复杂性高、安全要求高等特点，对于实时性和安全性的要求较高，性能测试在证券期货业中有着举足轻重的作用。一般常见性能测试类型如：压力测试、负载测试、容量测试、基准测试等。全链路压测是链路化压测的一种最佳工程实践，一般分为调用链路压测和业务链路压测两种形式。调用链路压测是从请求出发到结果返回途经的各层应用、服务、代理所产生的路径，调用链路压测主要验证单一业务场景的应用、服务、缓存、数据库等各层的性能表现。业务链路压测是多个有业务关联场景组合所产生的调用链路的组合，用于评估组合业务场景下系统的性能表现。流量回放技术丰富了全链路压测的手段，在此基础上可根据业务场景对录制的数据进行修正和验证。流量回放技术适合于无状态的系统。

4.其他非功能测试

在实践中，除性能测试外，应定期开展其他非功能测试，如：安全测试、架构测试、兼容性测试、可维护性测试、高可用测试等。其他非功能测试应具备测试规范，约定非功能实施的时机、整体流程和各环节细节。通过手工、自动化工具结合的方式完成测试。推荐的非功能测试实践还包括：精准测试、混沌测试、模糊测试等。

4.1.5 验收阶段测试

验收测试一般由用户或者独立测试人员根据测试计划和最终产品进行接收，确认产品是否满足合同或用户需求的测试。验收测试是正式投入实际生产使用前的最后一次全面质量检验活动，验收测试通过与否将最终决定软件是否合格。业务全链条测试是验收测试中的一种具体实践，当多个系统将共同发生同一新业务上线或重大业务变更时需要在类生产环境中进行联测，一般需要多个机构或组织之间协调配合，共同约定业务场景，数据流转和测试响应服务等。验收测试一般分为设定验收标准、设计验收场景/用例、执行验收测试用例和验收结果这几部分、验收测试工作流程如图6所示。

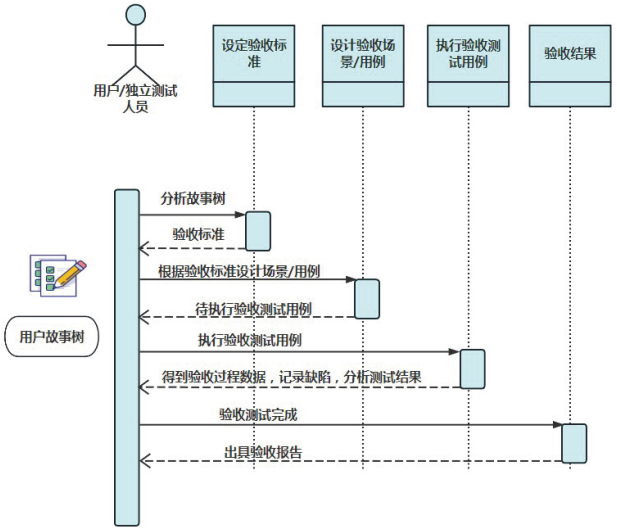


图6：验收测试工作流程

4.1.6运营阶段测试

运营阶段测试是在正式发布上线后对产品进行测试相关的活动，是测试运营一体化的演进。持续测试打破测试和运营的边界，促进测试右移到生产运营过程中，运营阶段持续测试主要是在产品上线后持续对产品线上运行状态、稳定性、高可用、安全性和系统快速恢复进行持续验证，主要包含的活动是线上测试，持续监控和低风险发布。

4.2 改进域实践

改进域实践是构成持续测试闭环的重要阶段，对软件全生命周期的各个阶段进行测试的结果、流程、效率和质量进行反馈，帮助快速定位问题，预防缺陷故障，提升软件研发效能。改进域实践包含准入准出、测试报告、缺陷管理、质量分析和生产事件复盘。

4.3 平台能力域实践

平台能力域是持续测试全流程的支持能力，包括流程制度规范、测试人员管理、测试环境管理、持续测试平台建设、测试资产和数据管理等方面。持续测试平台建设能有效提高持续测试效率，支持在各阶段中持续开展自动化测试、生成测试报告、实现质量门禁和持续交付流水线，产生

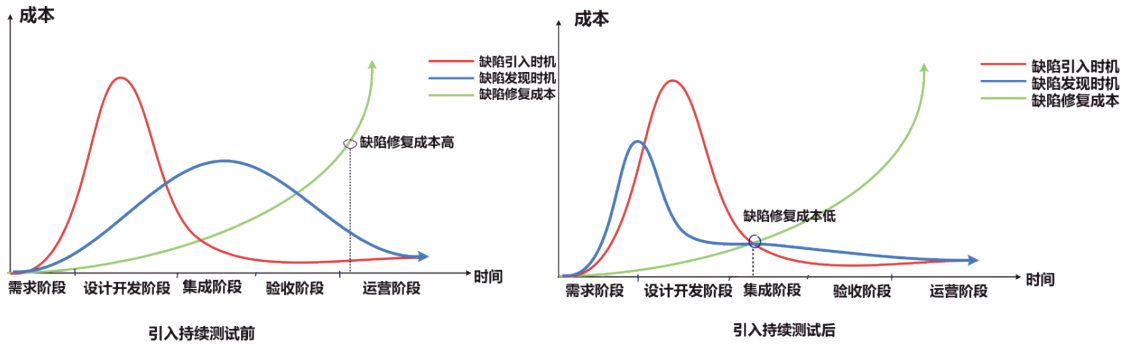


图7：验收测试工作流程

的测试过程数据可作为效能度量的基础分析数据。持续测试平台按版本迭代维护测试计划、测试用例、缺陷、测试结果、测试报告等实体，并建立各实体间的关联关系，形成可视化视图和各维度统计报表，持续测试平台产生和维护的数据可作为度量的基础数据，持续测试平台有助于提高测试效率和产品质量。

4.4 度量域实践

度量域实践包含持续测试效能度量指标，度量可视化和度量反馈，对持续测试所产生的效果和影响进行度量评估有利于持续改进。

4.4.1 持续测试效能度量指标

有效的度量指标应基于业务的价值收益来制定，度量指标指令包括指标定义，指标的权重分配，度量数据采集分析与维护，制定度量指标时需考虑度量结果的指导意义。度量有时候是把“双刃剑”，用得好可以促进组织持续改进，用得不好会产生内耗内卷，反而阻碍发展和创新，减弱组织的生命力和活力。度量结果的计算需要考虑度量指标数据的时效性以及数据覆盖的范围。

4.4.2 度量可视化

度量可视化是借助度量平台展示度量指标，度量平台是将持续测试过程中的离散数据进行有机地组合计算，形成各种客观指标数据，再配合多重展现方式，形成直观数据看板和自驱性的数据分析的能力。设计度量平台可分为数据采集、计算、分析、展现四个部分。

4.4.3 度量反馈

基于度量结果提供快速准确直观的趋势和质量数据，指导研发团队优化改进，推动组织数字化智能化转型与决策的进程，实现数据驱动持续改进。通过度量来反馈问题，从中产生洞察，落实改进行动，通过度量指标体系建立度量分析模型，系统性地分析问题，有效地帮助持续改进闭环。

五、持续测试结果检验

实践结果是持续测试理论的试金石，经过反复实践探索，有效证明持续测试实践有效助力证券期货业信息系统的大力发展，及早发现缺陷，极大减少了缺陷修复的成本，可观测性实践也增强了缺陷预防的能力。持续测试有效降低研发投入成本，图7对比了引入持续测试的前后缺陷修复成本和故障暴露时机，持续测试在软件研发中的实践提前了缺陷暴露的时间，加快了业务价值的流转速率，大幅降低了缺陷修复的成本，有效提升信息系统健康度。

六、总结与展望

持续测试是持续交付质量和效率的核心保障，持续测试倡导不断提升自动化测试比例，并强调覆盖软件全生命周期测试，实现随时执行测试并支持随时反馈，各类开箱即用测试工具组件也让持续测试成为可能。本文重点研究如何在满足安全和合规前提下，通过持续测试实践来提升研发交付效率，旨在解决证券、基金、期货行业相关金融机构进行数智化转型过程中的金融科技痛点问题，以满足现有基础和未来目标的发展需求，实现持续测试的最佳实践共享，助力提升运营效能。

参考文献：

[1]茹炳晟，张乐 《软件研发效能权威指南》 [M]. 北京：电子工业出版社，2022.
[2](JRT0175-2019)证券期货业软件测试规范[S].
[3](YDT 3763-2021)研发运营一体化（DevOps）能力成熟度模型 系列标准[S].

监管科技全球追踪

1月9日，美国金融业监管局（FINRA）发布2024年度监管报告。报告较以往新增两个主题：一是加密资产开发，报告对开展加密资产相关业务的公司提供了指导。二是信息公开，报告调研结果包括网络安全、反洗钱、反欺诈和制裁、最佳利益规则（Reg BI）和共同申报准则（CRS）表格以及合并审计跟踪等。

1月16日，欧洲银行管理局（EBA）宣布将其《反洗钱和反恐怖融资风险因素指南》（ML/TF Risk Factors Guidelines）扩展至加密资产服务提供商（CASPs）。新指南强调CASPs需要考虑的ML/TF风险因素和缓解措施。

1月24日，清华大学等联合发布《2024年金融业生成式人工智能应用报告》，提出生成式AI将赋能银行数字化转型，重塑金融业格局，预计有望带来3万亿增量商业价值，并可能改变交易、投资管理和风险评估方式。

1月30日，俄罗斯央行行长纳比乌琳娜表示，2024年俄方将推动金砖国家互相承认评级和建立反洗钱通用平台。她强调评级互认对贸易投资的重要性，并指出超国家评级机构面临复杂问题。同时，俄方愿分享反洗钱经验，简化金砖国家企业间合作。

2月5日，北京首都国际机场和大兴国际机场境外来宾支付服务示范区正式启用。中国人民银行副行长张青松表示，按照“大额刷卡、小额扫码、现金兜底”的总体思路，将在3到6个月内实质性进一步改善境外来华支付服务。

2月20日，上海市委网络安全和信息化委员会举行会议指出，要持续强化信息化的支撑和引领，加快培育发展新质生产力，让信息化数字化更好赋能现代化。抢抓人工智能发展机遇，推进关键核心技术突破，加快创新企业孵化培育。

2月27日，上海市委书记陈吉宁、市长龚正与中国人民银行行长潘功胜座谈。潘功胜表示，将支持上海高质量发展和国际金融中心建设，深化跨境贸易和投融资便利化，加强金融市场基础设施，优化人民币跨境支付系统，深化数字人民币试点，共同做好“五篇大文章”。

3月1日，中国银行业协会发布《银行业数据资产估值指南》。该指南构建了全面而实用的数据资产估值框架，涵盖数据资产的识别、评估、管理到价值提升等关键环节，为全面构建我国金融领域数据资产估值体系提供了有益借鉴，有助于完善数据要素资源体系，推动数据要素市场科学有序发展和数据资产估值走向规范化、市场化，助力行业数字化转型。

3月7日，香港金管局宣布展开全新的批发层面央行数字货币（wCBDC）项目“Ensemble”，以支持香港代币化市场发展。香港金管局将成立由本地及跨国银行、数字资产行业主要参与者、科技公司及CBDC专家小组组成的wCBDC架构工作小组，以推动制定业内标准及与时俱进的策略。

3月8日，迪拜国际金融中心（DIFC）宣布出台全球首部数字资产法，并对原有证券法和相关修正案进行更新。这些立法旨在确保DIFC法律跟上技术发展带来的国际贸易和金融市场的快速发展，并为数字资产的投资者和用户提供法律确定性。

3月13日，欧洲议会正式批准《人工智能法案》，对AI技术实施全面监管。其将AI分为不同风险等级并规定相应监管措施，违规者将面临至高7%营收的处罚。该法案旨在跟上科技发展，塑造跨国公司数据管理和AI使用方式，适用于27个欧盟国家的企业及在欧盟使用的他国AI系统。

3月27日，韩亚金融集团宣布了一套道德准则，旨在利用快速发展的AI技术，提供更安全、以客户为导向的金融服务。这套AI道德准则优先考虑五组价值观，包括包容性和公平性、安全性和责任性、透明度、数据管理和隐私保护。

4月17日，人力资源社会保障部等九部门发布《加快数字人才培育支撑数字经济发展行动方案（2024－2026年）》，紧贴数字产业化和产业数字化发展需要，用3年左右时间，扎实开展数字人才育、引、留、用等专项行动，增加数字人才有效供给，形成数字人才集聚效应。

4月26日，香港金融管理局推出FiNETech平台，汇集约100家银行、证券公司、保险公司以及科技企业，共同发掘在财富科技、保险科技、绿色科技、人工智能和分布式分类帐技术的进阶合作安排。

4月29日，国家发展改革委、国家数据局印发《数字经济2024年工作要点》。该工作要点提出适度超前布局数字基础设施，加快构建数据基础制度，深入推进产业数字化转型等9方面落实举措。

5月15日，中国证监会发布2023年执法情况综述。其中在2024年执法工作重点提出：强化线索发现。加大科技监管应用，不断提升线索发现的敏感度和精准度，强化跨部门跨领域跨市场监管协作，加强现场监管与非现场监管联动、信息披露与交易监管联动、现场检查与稽查调查联动，坚决消除监管空白和盲区。

5月20日，澳大利亚政府宣布审议通过《数字身份法案（2024）》和《数字身份（过渡和相应条款）法案（2024）》。该立法将建立一个更广范围的数字身份系统，并允许金融组织和相关服务提供商申请和加入政府数字身份平台，为用户提供一个更强大、更安全的在线生态系统。

《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