THE FORELAND OF
TRADING TECHNOLOGY

内部资料 免费交流
《准印证》编号沪（K）0671

交易技术前沿

2022年 第一期 总第48期

本期主题

金融科技

No.1

交易技术前沿
THE FORELAND OF TRADING TECHNOLOGY

内部资料 2022 年第一期（总第 48 期）
准印证号：沪（K）0671

NO.1

主管：上海证券交易所
主办：上海证券交易所
总编：黄红元、蔡建春
副总编：王泊
执行总编：唐忆
责任编辑：徐广斌、徐丹、陆伟、王昕

上海市杨高南路 388 号
邮编：200127
电话：021-68607129，021-68607131
传真：021-68813188
投稿邮箱：ftt.editor@sse.com.cn

扫码阅览历期杂志

篇首语

国家“十四五”规划指出，数据是新的生产要素，数字技术是新的发展引擎，数字经济是新的发展阶段。数字技术将为金融创新发展注入充沛活力，数字经济将为金融数字化转型提供丰富场景，当前金融科技正在迈入高质量发展的新阶段，不断深化金融供给侧结构性改革，增强金融服务实体经济的能力。本期《交易技术前沿》以“金融科技”为主题，收录来自行业八篇优秀文章，探讨行业技术前沿：

《华泰资管研报智能分析技术探索和实践》基于研报智能分析和要素自动提取技术，提出了一套研报分类、研报文档结构化解析和研报要素抽取的一体化解决方案。

《智能机房巡检机器人在行业数据中心应用实践》通过巡检机器人实现机房状态的实时可视可管，提高机房的自动化运维效率、管理效率和科技水平。

《一种基于机器学习的攻击源画像构建方法》基于攻击行为的时间、空间、频率、手段等特征构建攻击源画像，辅助评估威胁等级，形成有效的应对措施和决策。

《海通证券数据中台建设思路、方法与实践》以数据中台为基础，提供统一数据服务云 DaaS 和数据治理能力，使数据生产与业务赋能形成闭环，实现了数据的可见、可用、可经营。

《东方证券分布式系统可观测性解决方案探索与实践》对分布式系统可观测问题进行了探讨，并通过引入分布式链路跟踪等技术提供了解决方案。

《算法服务平台的实践》介绍了恒泰证券在算法金融理论与实践的基础上，结合自身信息技术能力自主研发的算法服务平台。

《全历史数据服务系统在信创大数据平台上的实践》介绍了历史数据的重要意义，对证券行业历史数据使用现状进行了梳理，并提出一套基于全国产化技术的大数据平台解决方案。

《国产分布式数据库在证券行业的应用价值》解读了国产化加速下传统数据库面临的困境，以此出发对国产分布式数据库在证券行业内应用价值进行了阐述。

《交易技术前沿》编辑部

2022 年 4 月 29 日

目录 Contents

本期热点

实践探索

信息技术创新观察

信息资讯采撷

本期热点

1 华泰资管研报智能分析技术探索和实践

2 智能机房巡检机器人在行业数据中心应用实践

3 一种基于机器学习的攻击源画像构建方法

华泰资管研报智能分析技术探索和实践

邱震宇、彭南博、朱阿柯、陈雨辉、肖驰 / 华泰证券股份有限公司 信息技术部
邮箱：qiuzhenyu@htsc.com

投资管理是当前证券行业中非常重要且极具发展潜力的业务领域。在投资管理业务中，研究员需要消耗大量时间、精力来阅读和分析不同研究机构发布的研究分析报告。通常这些研究报告数量繁多、内容丰富、样式不统一，且领域内缺少专业的研报阅读和分析平台，这些都极大影响了研究员的工作效率和产出。为了帮助研究人员快速展开研究活动、提升产出效率，我们希望构建一个专业的研报分析平台，该平台利用人工智能领域中的自然语言处理和计算机视觉技术对研报进行智能分析。目前，我们已经构建了一个包括研报分类、研报文档结构化解析和研报要素抽取的一体化解决方案。该方案已经应用到华泰证券的投研能力服务平台中，并被研究员广泛使用。

1、研究背景

随着经济发展，中国居民财富持续增长。资产管理需求日益增多，资产管理业务迎来新的发展机遇。市场对机构的投资管理能力提出了更高的要求。近年来，华泰证券资管团队积极建设数字化投资研究平台以助力投资管理业务。公司希望借鉴国内外投研的先进经验，通过数字化与智能化技术为投研业务赋能，实现多源异构研究数据融合、产业投资逻辑的知识沉淀和投研过程的提质增效，从而提升研究效率，增强公司在资产管理领域的核心竞争力。

在投资管理业务的研究过程中，研究人员需要阅读和分析各个券商机构制作的研究分析报告。他们需要从这些报告中提取出有价值的关键信息，包括研报分析的个股、当前评级、目标价

和盈利预测数据等。这部分研究工作较为繁琐，会耗费研究人员大量的时间和精力。

随着人工智能技术的发展，许多金融机构开始将自然语言处理技术引入到金融文本分析领域，如情感分析、舆情预警和实体识别等。这些工作通常是针对金融纯文本任务，实际上金融领域还有大量的富文本语料有待挖掘和分析，例如上市公司公告、研究机构研究分析报告等。这些报告大多都是 PDF 格式，其中包含文本、图表和表格等元素，这些元素语义丰富，具有很高的研究价值。

基于上述分析，我们希望利用人工智能技术从研报 PDF 中自动抽取出关键信息并组织成结构化的数据进行分析。具体地，我们结合自然语言处理与计算机视觉相关技术，设计了一套研究报告（以下简称研报）关键信息要素抽取解决方案。该方案包含研报文件解析、研报类型分析和研报要素抽取等功能。为了解决金融垂直领域知识迁移难、要素抽取训练集构建成本高等问题，我们构建了金融领域的语言模型训练框架。同时，我们引入数据增强的理念，保证在训练样本不多的情况下，模型依然能够在要素抽取任务中取得理想的效果。目前，上述解决方案已经被应用到华泰证券的投研服务应用平台，并被研究员广泛使用。

2、技术方案

研报要素抽取任务希望通过文件解析、类型分析、要素抽取等步骤，将研究报告 PDF 原始文件转化为结构化的要素信息。这些信息包括研报的标题、摘要、作者和盈利预测指标等。本章节将详细介绍研报要素智能抽取任务面临的难点，并针对这些难点提出相应的解决方案，其整体框架如图 1 所示。

研究分析报告通常为 PDF 格式，我们首先引入定制化的研报 PDF 解析模块。在通用 PDF 解析模块的基础上，我们根据研报样式的特点对该模块进行优化，从而有效识别出研报中的文本段落、图表和表格等元素。这些结构化解析结果被输入到研报类型分析模块中进行进一步分析。研报类型分析模块会根据抽取出的要素判别研报所属的类型，如个股点评、行业分析或者宏观研究等。最后，我们引入要素抽取和标准化模块。考虑到研报内容和格式的多样性，我们设计了一套涵盖多家研究机构和多种类型研报的多层次要素抽取模板库。该模块将根据研报类型，选择相应的抽取方法对研报要素进行抽取。最终，方案会对抽取结果进行标准化处理以便输入到下游系统中。接下来，我们详细介绍上述模块的基本流程和技术原理。

2.1 研报 PDF 解析

研报 PDF 解析是研报智能分析解决方案的第一个处理模块，后续处理模块的效果直接依赖本模块的解析质量，因此该部分是最基础也是最重要的模块。当前行业内，不少研究机构和公司对 PDF 文档解析研究并做出了一些比较成熟的

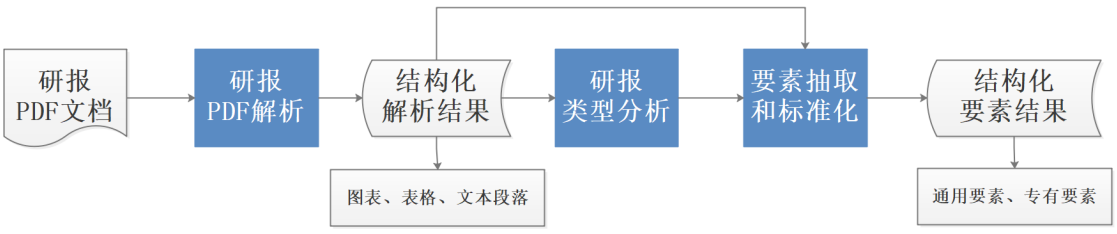


图 1：研报智能要素抽取方案整体框架

产品。然而，这些 PDF 解析工具大都只针对上市公司及发债主体的公告和简单文书等，对研报适配性不太高。经过分析，我们总结出研报文件解析的两个难点：

1）研报样式多样缺乏统一规范，不同研究机构、不同研究员制作的研报样式会有所不同。图 2 展示了华泰研报与中金研报的样例，我们可以发现除了整体布局不同之外，局部内容样式也各有特点。例如，研报作者介绍、标的代码、评级、目标价的排版等；

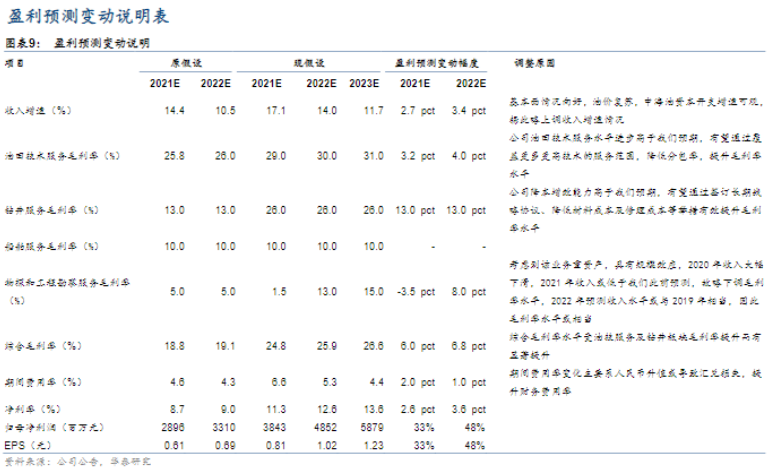
图 2：不同研究机构研报对比图

2）研报要素丰富，个别要素抽取难度大。研报中包含大量图表和表格，这些元素样式多样。表格会包含无边框、合并单元格等情况，如图 3 所示。这些都增加了表格元素识别的难度。

基于上述分析，我们构建了一套由点及面的研报 PDF 解析工作流，如图 4 所示。该流程从基本特征开始抽取，抽取出的特征将被输入到单页元素模型中进行解析，最后模型将关联多个单页元素得到篇章级层次的元素识别结果。下面将分别对这三个步骤进行详细的介绍。

2.1.1 基础特征抽取

基础特征抽取包括文本特征、表格特征、图片特征和样式特征。文本特征除了常见的语义向量特征外，如 N-GRAM 统计特征（一种融合多段字符的特征）、主题特征等，我们还引入文本位置特征，如元素块在当前页面的坐标、与其他元素块之间的垂直、水平距离等；针对表格特征，我们提取了表格中的线段、线框的位置坐标、长度等特征。针对图片及扫描件元素，我们除了规



盈利预测变动说明表

图表9：盈利预测变动说明

项目	原假设		现假设			盈利预测变动幅度		调整原因
	2021E	2022E	2021E	2022E	2023E	2021E	2022E	
收入增速（%）	14.4	10.5	17.1	14.0	11.7	2.7 pct	3.4 pct	基本面情况向好，油价复苏，中海油资本开支增速可观，相比略上调收入增速情况
油田技术服务毛利率（%）	25.8	26.0	29.0	30.0	31.0	3.2 pct	4.0 pct	公司油田技术服务水平进步高于我们预期，有望通过覆盖更多类高技术的服务范围，降低分包率，提升毛利率水平
钻井服务毛利率（%）	13.0	13.0	26.0	26.0	26.0	13.0 pct	13.0 pct	公司降本增效能力高于我们预期，有望通过签订长期战略协议、降低材料成本及修理成本等举措有效提升毛利率水平
船舶服务毛利率（%）	10.0	10.0	10.0	10.0	10.0	-	-	
物探和工程勘察服务毛利率（%）	5.0	5.0	1.5	13.0	15.0	-3.5 pct	8.0 pct	考虑到该业务重资产，具有较强波动，2020 年收入大幅下滑，2021 年收入或低于我们此前预测，故略下调毛利率水平，2022 年预测收入水平或与 2019 年相当，因此毛利率水平或相当
综合毛利率（%）	18.8	19.1	24.8	25.9	26.6	6.0 pct	6.8 pct	综合毛利率水平受油技服务及钻井技术毛利率提升共有显著提升
期间费用率（%）	4.6	4.3	6.6	5.3	4.4	2.0 pct	1.0 pct	期间费用率变化主要系人民币升值或导致汇兑损失，提升财务费用率
净利率（%）	8.7	9.0	11.3	12.6	13.6	2.6 pct	3.6 pct	
归母净利润（百万元）	2896	3310	3843	4862	5879	33%	48%	
EPS（元）	0.61	0.69	0.81	1.02	1.23	33%	48%	

资料来源：公司公告，华泰研究

图 3：研报表格示例图

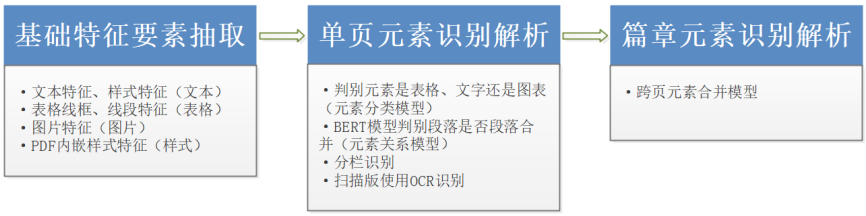


图 4：研报 PDF 解析工作流程图

则匹配外，还引入 OCR（光学字符识别，用于识别图片扫描文件中的文字）识别工具提取图片的特征。此外，我们还提取了 PDF 内嵌的背景样式特征，例如字体、字号、颜色等 CSS 样式。

2.1.2 单页元素识别

单页元素识别模块，包括元素分类模型和元素关系模型两个模型。元素分类模型以 XGBOOST 模型（一种机器学习模型，以树模型为基础，应用集成学习思想提升模型能力）为基础，我们通过将文本和视觉两种特征作为输入，最终由模型来判断该元素是表格、文字还是图表。元素关系识别模型主要是识别两个文字段落是否需要合并。我们基于经典的大规模预训练语言模型 BERT(Bi-directional Encoder Representation from Transformers, 一种基于 Transformers 模型结构的预训练语言模型，在海量文本上进行预训练，得到一个具有一定语言能力的先验模型），训练得到一个判断两个文本元素块是否为同一段落的二分类模型。在构建数据集时，我们从本页文本元素块中，任选两个样本，若两个样本属于同一段落，则标注为 1，否则为 0。我们会尽量保证正负样本的比例维持在一个 1:1 的比例。最终该模型段落合并识别准确率能够达到 90% 以上。对于分栏识别，我们主要依靠视觉方面的特征，如水平元素块之间的间距、水平方向上元素块的类型分布、垂直方向是否有竖线像素块等。

篇章层次识别模块关注宏观层面的元素识别，包括段落之间的跨页合并、表格之间的跨页合并和整个文档的章节层次识别等。我们仍然使用 BERT 构建一个二分类模型来判别段落合并问题。在构建数据集时，我们从整个文档中任选两页，并抽取第一页靠近底部的文本元素以及第二页靠近顶部的文本元素作为样本对。对于表格之间的合并，我们同样综合了文本特征和视觉特征，文本特征包含了表格表头内容的向量表示、表格具体内容的统计特征表示等，视觉特征则包含了表格线框的位置特征、第二页表格中是否存在表头元素的位置特征等。我们也引入集成模型接收两种不同维度的特征来识别跨页的元素。

2.2 研报类型检测

通过分析大量研报后，我们发现研报的类型有很多种，其内容和样式大不相同。不同类型研报内容对比如图 5 所示，行业研报会关注某个行业的进展，报告中就会列出行业内重点推荐的标的股票；个股点评类研报则会关注某个具体标的的近况，报告就会侧重分析其最近的财务状况、舆情事件等，并给出相关的盈利预测以及评级。基于上述原因，我们不能使用单一抽取模板来涵盖所有类型的研报，而是要因地制宜、有的放矢。

图 5：不同类型研报对比图

研报类型可分为以下几类：个股点评类、行业研报、宏观报告、深度报告、策略报告、金工报告、固收报告和其他研报。根据研报的标题以及部分摘要文本，我们应用文本多分类技术来判别该研报的类型。原先，我们计划使用通用 BERT 模型作为基础模型。该模型通过在海量语料下进行预训练，已经学习到常见的通用语义信息。然而，在使用该模型进行分类时，我们发现该模型对于研报标题与摘要文本的理解并不如预期。很多金融场景下的专用术语如“老旧欠款、股权激励”等无法被模型完全理解。针对这个问题，我们利用当前新闻数据库中已经积累的百万级别金融新闻语料，以通用预训练 BERT 模型权重为基础，引入上市公司实体库的分词词库，进行了全词级别的金融领域 BERT 预训练。最终，我们得到了一个适用于金融场景下的 BERT 模型。

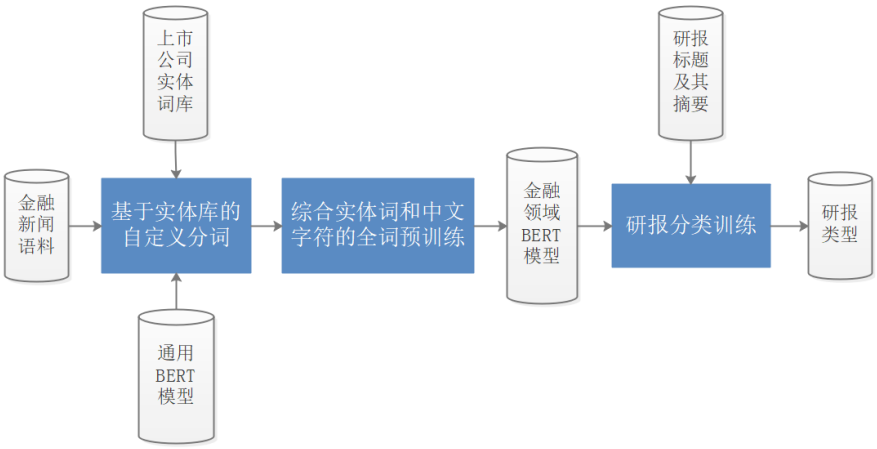


图 6：研报类型检测流程图

该模型在研报分类的任务上相比原始 BERT 在准确率和召回率上都有接近 2 个百分点的提升，其均能达到 90% 以上。该模型处理流程如图 6 所示。

该模型的预训练流程参照 BERT 的训练任务，包含了掩码语言模型任务和下一句预测任务 (Next Sentence Prediction, NSP)，其框架如图 7 所示：

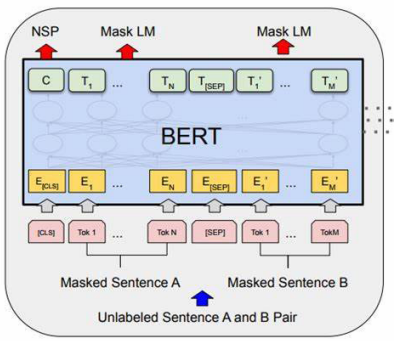


图 7：BERT 预训练模型框架图

2.3 要素抽取和标准化

通过上面的处理流程，我们得到了研报文档的元素解析结果以及研报的类型标签，根据这两种数据，模型可以对研报要素进行进一步抽取。

2.3.1 要素模板设计

根据已有的研报类型标签库，我们设计了一套多级体系的抽取字段模板。抽取字段模板分为通用字段和类型专用字段，通用字段表示所有类型研报中的通用信息，如作者、机构名称、研报标题等；类型专用字段则表示不同类型研报各自关注的字段，以个股点评类为例，研究员比较关注研报中分析的个股标的、评级、目标价，另外还有盈利预测的一些指标如营业收入、净利润、ROA（资产回报率）和 ROE（净资产收益率）等。对于盈利预测指标来说，研报对于标的预测时间周期都不是单一的，通常会跨越多个年份。因此在设计盈利预测的要素模板时，要考虑将预测的年份周期也包含在其中。如图 8 所示：

利润表

(百万元)	2019	2020	2021E	2022E	2023E
营业收入	10,078.7	13,024.7	16,932.1	23,704.9	35,557.3
减:营业成本	5,440.5	7,148.4	8,210.1	11,420.4	16,960.8
营业税费	67.3	78.9	154.1	215.7	323.6
销售费用	1,780.2	2,084.4	3,555.7	4,741.0	7,111.5
管理费用	706.7	856.6	3,047.8	4,266.9	6,400.3
财务费用	-3.7	16.2	-107.6	-109.3	-96.0
资产减值损失	-7.0	-37.8	8.5	9.0	8.0
加:公允价值变动收益	-4.4	355.0	-	-	-
投资和汇兑收益	111.6	32.1	15.0	15.0	15.0
营业利润	988.0	1,437.1	2,078.5	3,176.2	4,864.2
加:营业外净收支	7.5	19.6	28.0	29.1	29.1
利润总额	995.4	1,456.6	2,106.5	3,205.3	4,893.2
减:所得税	52.4	14.9	231.7	352.6	538.3
净利润	819.2	1,363.8	1,874.7	2,852.7	4,355.0

图 8：盈利预测表格示例图

2.3.2 要素抽取

根据前述步骤得到的研报文档解析结果和研报类型，再辅以要素模板的指导，我们就可以对研报进行精细化要素抽取。根据要素在研报中所

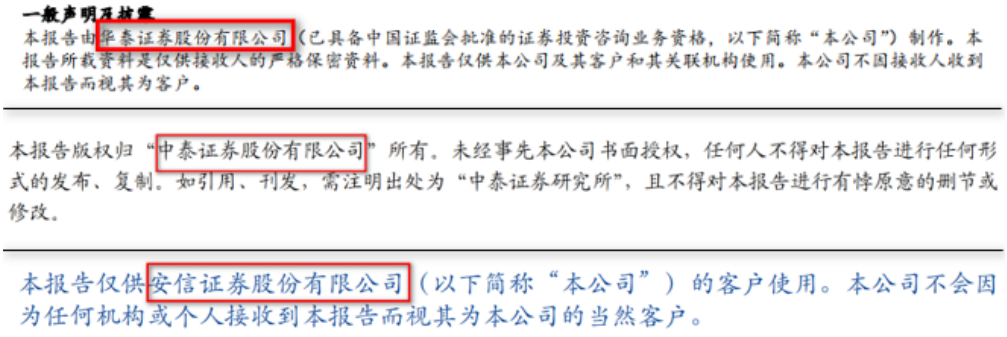


图 9：研报机构描述在研报中的位置示例图

处的不同位置，我们将要素抽取分为文本段落中抽取和表格中抽取。下面将分别介绍这两种情况的流程。

1）文本段落中抽取

以研报所属的研究结构为例，这个要素一般会出现在研报结尾的声明或者备注段落中。图 9 展示了不同研报中机构所处的文本段落位置。

从图 9 中可以看到，不同机构研报的声明内容行文都不尽相同，模型无法使用简单的规则匹配来进行抽取。我们将该抽取任务抽象为一个文本序列抽取任务，任务的目标为在图 9 第一个例子中识别“华泰证券股份有限公司”是一个研究机构名称，且该研报是该机构制作。经过分析，我们发现这种声明一般只会出现在研报的最后 1-2 页，因此我们可以预先将抽取的篇幅限定在最后两页文本段落中，并根据关键词如“一般声明”、“声明”等进一步定位到待抽取的上下文（一般是 1-2 句话）。剩下的工作就是对上下文中的每句话引入一个序列抽取模型，判断当前语句中是否包含指定的要素。经典的序列抽取模型通过引入 BERT 模型并结合概率图模型条件随机场进行联合训练。实施这个经典方法最大的难点在于我们无法以较低的成本来构建一个足量数据的训练集。针对这个难点，我们采用了另外一种方式来建模上述问题。我们在待抽取的语句 paragraph 之前，引入一个 question 文本，以抽取研究结构为例，该 question 可以为“制作这篇研报的分析师所属的机构是什么？”。然后我们就得到一个 question + paragraph 的训练样本，我们的任务目标为让模型在 paragraph 中找到 question 对应的答案文本，具体来说就是找到答案在 paragraph 的起始位置。模型示意图如图 10 所示。

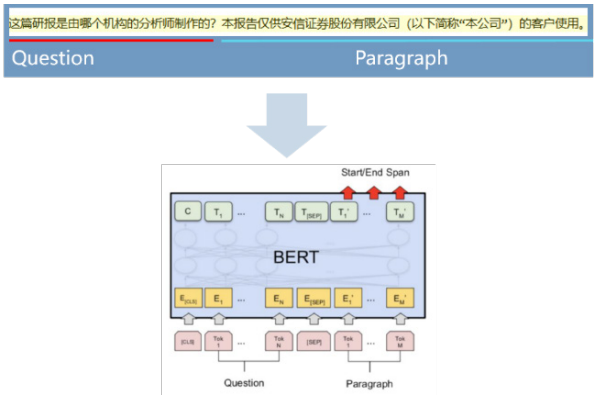


图 10：抽取式阅读理解任务处理流程图

这种方式借鉴了机器阅读理解的任务思想，通过引入一个问题文本，让模型能够更深入理解当前的任务信息和待抽取文本的语义信息，使得模型在标注数据不多的情况下，也能获得比传统的序列抽取模型更好的效果，另外，我们增加了 question 文本，也间接得起到了数据增强的作用，缓解了数据标注不足的问题。基于上述方法，我们在标注了 100-200 篇左右的文本后，就获得了准确率、召回率大于 90% 的效果。

2）表格中抽取

要素除了出现在文本段落中，还会出现在表格中，例如盈利预测指标。在研报文本解析模块

中我们对表格识别做了优化，表格元素识别质量得到了大幅提升。得益于表格识别的稳定性能，我们能够准确地从表格中提取指定要素。因此，我们在该步骤中通过规则匹配来进行处理，模型只需定位到表头中的要素位置，匹配相应的数值即可。以图 8 中的营业收入为例，我只需要先定位出表头中包含年份 +“E”的表格，然后定位出该表中包含营业收入关键词的行，并根据年份对应的列位置，匹配相应的数值即可。

2.3.3 要素标准化输出

当抽取完所有要素之后，模型还要对要素进行标准化的格式输出。对于金额型要素，例如营业收入、净利润，模型需要统一其单位和量纲；对于百分率、日期要素，模型需要统一其输出样式；对于评级、行业等要素，由于不同机构使用的体系不同，原样输出会给研究员带来一定的困扰。因此，我们与研究员一起制定了一个统一的评级描述模板和一级行业目录，将抽取出来的评级和行业映射到统一模板输出。另外，为了提升行业要素抽取和映射的准确率，对于行业研报，我们还会将研报中提到的行业内推荐个股标的抽取出来，并根据行业目录，将这些个股标的对应的一级行业搜索出来，并选择匹配频率最高的行业名称输出。

3、成果及应用

目前，这套研报智能要素智能分析的解决方案已经应用在公司的投研服务平台以及内部轻应用中。投研服务平台会展示研报的综合统计以及每份研报要素抽取之后的结构化结果。图 11 展示了不同维度下研报的分类统计，此功能依赖于研报分类检测服务。

图 11：研报分析综合统计图

图 12 展示了某个具体研报的要素抽取结果，

图 12：研报要素抽取效果图

包括机构、作者、行业、评级、个股标的、目标价以及盈利预测数据等等。

图 13 展示了公司内部轻应用中的研报智能分析功能。

图 13：轻应用功能展示图

4、总结与展望

本文主要介绍了华泰证券在研报智能分析和要素自动提取的技术探索与实践。我们的技术方案包含了研报文档解析、研报类型检测以及要素抽取及标准化输出等功能。根据当前解决方案的实际应用来看，大部分的研报都能被准确地分析并输出结构化的要素信息，这证明我们的方案是实际有效的。经过与研究员的需求讨论和样例分析，我们的方案目前仍然存在一些不足，我们计划在以下几个方面继续优化：

1）部分研报会将所有元素块图片化。这种情况下，整篇研报都是由单独的图片拼接而成。模型对于这类研报的解析效果还不佳。我们计划根据这类研报的特点，优化 OCR 服务；

2）当前抽取要素以文本类型为主，然而研报中有很多统计分析数据是以图表形式展示。我们计划对研报中的图表进行抽取，并与其图注相关联。最后，研究员通过图表搜索功能搜索到自己感兴趣的图表内容；

3）研报中存在分析师的主观分析文本，我们计划构建一个研报观点分析检测模型，将研报中的观点、结论型文本提取出来。同时引入情感分析模型判断研究员对个股或者行业的情绪，供上游应用使用。

智能机房巡检机器人在行业数据中心应用实践

尚升、孙君文、杨慈航 / 上交所技术有限责任公司 金桥运行部 上海
惠怀名、武捷 / 国融证券股份有限公司 信息技术中心 北京
邮箱：sshang@sse.com.cn

《国家新一代人工智能标准体系建设指南》中指出，到 2021 年明确人工智能标准化顶层设计，研究标准体系建设和标准研制的总体规则。到 2023 年，初步建立人工智能标准体系，并率先在金融等重点行业和领域进行推进。国家层面顶层的统筹规划、各技术领域攻关突破、应用场景的探索成熟，人工智能发展取得了显著进步，成为当前行业研究的热点领域。本文介绍了智能机房巡检机器人在数据中心的应用，通过智能机房巡检机器人实现机房状态的实时可视可管，并提高机房的自动化运维效率、管理效率和科技水平。

1、应用研究背景

1.1 行业运维挑战

在业务发展新需求、行业信息监管要求和数据中心生命周期更迭等因素影响下，多数据中心格局是行业发展趋势。此格局下，一个运维团队通常需要负责多数据中心的运维。一般情况下，运维团队会驻场在某一生产数据中心，而其他数据中心则采用远程运维和现场按需前往结合的方式运维，即远程监控、操作系统，安排专人前往现场巡检、排障和施工等工作，该模式下主要存在如下难点：

1）巡检精细度和巡检频次难以提升

基于行业监管要求和业务运行安全考虑，目前普遍采用人工巡检的方式查看设备现场运行状态，巡检准确主观性较强、无法高频巡检，且缺乏对局部温湿度、烟雾、噪声等的有效监控手段，难以全面有效预防设备运行风险。

2）现场陪护工作耗时耗力，回溯手段欠缺

现场开展的排障、施工等工作如需第三方参与，则须安排人员全程陪护，远程和现场人员联动性不强，工作内容无法可视化呈现给远端人员，且陪护过程无法精确记录和回溯（如开柜授权记录、工作画面和机房物品流动追溯等）。

3）跨部门监控整合和工作标准化困难

随着多维度、全方位、态势预测的监控发展趋势，运维团队需整合多个监控软件分析报表后统一决策，而这些涉及设备状态监控、环境监控等软件彼此独立性较强，整合分析耗时耗力。且随着业务系统的增加，对应的设备数量也日益增长，采集数据需要的人力也在不断增加，负责标准化的平台或工具需要不断迭代升级，对运维团队来说也是一个挑战。

4）运维团队人员变动影响系统运行稳定

运维团队存在人员流动的情况，新到岗人员需要一定时间熟悉工作内容和流程，特别是一线运维岗位，运维巡检人员主观判断和对系统设备的熟悉程度直接影响故障发现上报效率，如何避免人员变动带来的影响也是运维团队需要考虑的问题。

1.2 AI 助力运维转型

智能机器人属于特种机器人的一种。以先进的 AI 算法驱动，使用多传感器综合采集分析数据，智能机器人具备一定的自主判定和决策能力，可实现自动充电、自主导航和避障、自主执行作业。

随着计算机视觉、机器学习、智能语音等算法技术的进步，AI 技术应用场景不断拓展，智能机器人逐步从工业应用走向更多服务领域。基于安全运行考虑、运维的复杂程度、综合成本考虑和其他限制（如机房场地改造限制、机房场地安全等），智能机房巡检机器人目前功能以“看”为主，即使用机器人本体集成的激光雷达、超声波传感器、工业相机、深度相机、热成像相机、温湿度等传感器，采集特定信息进行综合分析判定，并实时输出异常。如果将智能机房巡检机器人应用于数据中心，替代部分重复程度高、需借助感应器或其他工具采集数据的日常工作，可极大降低一线人员的现场工作量。

为探索人工运维和智能机房巡检机器人结合的数据中心运维新模式，上交所技术开展智能机房巡检机器人应用实践，验证智能机房巡检机器人代替人工开展设备巡检、环境巡检、施工陪护监控等工作的可行性，让机房管理轻型化、标准化、可视化、智能化。

2、部署实践方案

2.1 整体设计原则

本次实践在金桥数据中心托管数据机房内开展，整体设计思路如下：

1）整体运行安全

实践选用的智能机房巡检机器人设计符合机房建设标准及规则，功能先进可靠，整体解决方案已有成熟的应用案例。此外，在实践过程，机器工作时人工全程陪同，发现异常时可通过机器人本体急停按钮停止机器人一切运动。

2）功能验证全面

本次研究在已上线业务的机柜通道内开展，在生产环境下验证大部分功能。同时，额外启用一个测试机柜通道，部署多台测试设备，人工模拟故障和测试场景，与生产环境功能验证互补，多方位验证智能机房巡检机器人功能实现程度。

3）实践过程严谨

智能机房巡检机器人运维和人工运维同步开展，记录同一应用场景下两者工作内容完成度、完成效率和流程标准化等关键数据，后期汇总分析数据对比两者优劣。应用部署方案与厂商推荐的商用方案保持一致，兼顾本次实践和后续快速投产部署需求。

2.2 应用架构设计

智能机房巡检机器人系统主要组成部分为机器人本体、智能充电桩、人机交互系统、机器人管理系统、资产管理系统等。其中，人机交互系统预装在机器人本体内，机器人管理系统安装在后台服务器内。机器人本体通过无线实时上传数据或通过电力载波与充电桩之间传输数据，并在充电过程中将数据上传。机器人整体解决方案架构如下：

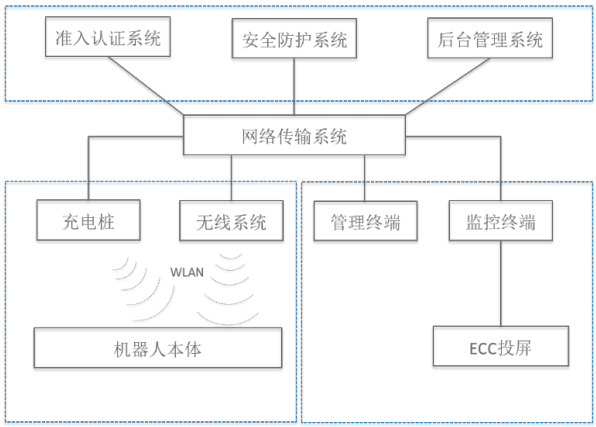


图 1：智能机房巡检机器人架构示意图

在实际部署中，考虑到机器人本体通过无线系统上传数据，因此对无线系统接入作了重点防护，如准入认证、硬件地址绑定、认证失败禁止登入一段时间等防护举措，同时也限制无线系统仅在智能机房巡检机器人作业期间开放，无线系统开放期间全程监控。

在 ECC 内部署管理终端和监控终端，实时输出智能机房巡检机器人作业过程和结果并投屏，ECC 内既有监控人员可查看机器人状态，并在管理终端下发即时任务、定时任务、机器人维护和其他作业指令。

3、研究实践

综合评估日常一线运维工作内容和智能机房巡检机器人能力范围，本次实践场景既包含机器人本体运行安全验证，又包含可替代人工运维的工作场景如设备巡检、环境巡检和施工陪护等。

3.1 场景一：自动化运行

本次实践中，AI 赋能的智能机房巡检机器人自主运动决策，依赖激光雷达、摄像头等传感器采集的信息完成路径导航、避障，工作范围为标准化机房环境，无额外基础环境改装。

3.1.1 急停、碰撞干预

人工通过急停按钮对智能巡检机器人进行干预，干预后巡检机器人瞬间停止所有动作并发出声光告警。此外，智能巡检机器人检测到碰撞时也会触发急停并告警，运行风险可视可控。

3.1.2 自动避障和路径规划

考虑到数据中心现场环境较为复杂，行进路线上会出现干扰物体，如螺丝刀、机柜盲板、网

表 1：自动避障功能验证

自动避障场景描述	功能验证总结
行进路线前方出现人员	机器人在距离测试人员约 2 米的距离自动停止，并转动摄像头对四周环境进行扫描，扫描停止后选取另一条路线绕行至目标位置，实现了全自动避障
机柜通道内行进路线前后方均出现测试人员	机器人感应到行进路线前方出现人员，转动本体对四周环境进行扫描，传感器回馈周围均有障碍物，无可用行进路线，机器人触发急停
行进路线前方出现 2U 服务器	机器人在距离测试服务器约 2 米的距离自动停止，并转动摄像头对四周环境进行扫描，扫描停止后选取另一条路线绕行至目标位置，实现了全自动避障
行进路线前方出现机柜盲板	机器人运动轮行驶通过机柜盲板
行进路线前方出现网线	机器人运动轮行驶通过网线

表 2：机房环境巡检验证

巡检环境参数	告警阈值设置	功能验证总结
温度	高于 17℃	巡检过程中自动采用分析，检测颗粒度为 1/6 机柜高度，发现异常按告警等级发出声光提示和告警通知
湿度	高于 30%RH	巡检过程中自动采用分析，检测颗粒度为单机柜前湿度。发现异常按告警等级发出声光提示和告警通知
噪声	高于 70dB	巡检过程中自动采用分析，检测颗粒度为单机柜前噪声。发现异常按告警等级发出声光提示和告警通知
悬浮粒子	高于 %0obs/m	巡检过程中自动采用分析，检测颗粒度为单机柜前粒子浓度。发现异常按告警等级发出声光提示和告警通知
温度	高于 17℃	巡检过程中自动采用分析，检测颗粒度为 1/6 机柜高度，发现异常按告警等级发出声光提示和告警通知

线、交换机等物体，智能机房巡检机器人需要识别到物体并及时终止绕行或无障碍通过。在研究过程中，针对不同的障碍物进行干扰实验，智能机房巡检机器人基本能够按照预期完成自动避障和路径规划。

3.1.3 自动回桩充电

为尽可能远离设备并避免影响正常人员通行，本次充电桩靠近立柱角落安装，无固定装置，用 220V/50HZ 市电供电。经多次研究验证，智能机房巡检机器人能够在任务执行完毕、任务终止或低电量情况下自动回桩充电，回桩对接过程中，充电桩受力出现毫米级位移，但不影响下次回桩。在触发定时任务或派发即时任务后，智能机房巡检机器人自主执行任务，全程路径规划和避障无需人为干预。机房巡检机器人从 0 充电至 100% 需要 2.5 小时，且支持快速充电和电池快换方案，充满电状态下可持续巡检 6 小时以上。

3.2 场景二：机房环境巡检

应用实践在标准为 T3+ 的生产机房内进行，基础环境异常出现概率极低。考虑到功能测试的严谨性，本次通过调整基础环境关键告警阈值来验证智能机房巡检机器人功能性。

在指派巡检任务时，可以关联到单个或多个机柜，巡检过程中可实时查看环境参数。同时，智能机房巡检机器人后台管理系统支持即时巡检任务和定时巡检任务，极大的降低了运维工作量。

3.3 场景三：设备状态巡检

通过人工智能深度学习算法，只需在部署前

表 3：机房环境巡检工作对比

对比项目	智能机房巡检机器人运维	人工运维
信息采集范围	采集的环境信息更丰富，异常预防更全面。可额外获取温湿度、风速、噪声、气体和颗粒物浓度等	仅能通过固定传感器获取温湿度，可发现机房盲板未安装、布线较乱等问题
信息采集颗粒度	采集的环境参数颗粒度更细，机房可视可管程度更高。可获取每个机柜的环境参数，温度甚至可精确到 1/6 机柜	仅能通过固定传感器获取某一区域的环境参数
告警上报	采集环境参数时实时判定异常、现场和后台监控同步发出告警，告警响应更迅速，可直接把告警信息发送至后台二线，告警上报更迅速	发现问题需要电话或信息联系后台二线处理，上报流程需要一定时间
资料调阅	环境参数即采即录，资料调阅更便捷。历史巡检记录保留时间可自定义，资料调阅更便捷，生成报表巡检结果更直观	巡检纸质记录需长期保存，需要耗费人力物力来保存整理相关资料，历史信息调阅较麻烦

期为机器人标注一次样本即可，智能机房巡检机器人通过人工智能增补算法快速将小数量样本集扩充至千万级样本级别，从而满足进行大规模图像算法训练的样本需求。

本次应用实践巡检设备超过 400 台，部署在 3 排机柜中，设备类型包括交换机、路由器、服务器、存储和传输设备等。在机器人巡检同时人工巡检也在同步开展，对比验证了智能机器人的可用性和可靠性。

应用实践中，智能机房巡检机器人可实现自动化定时巡检，单次巡检时间约 30 分钟，巡检画面和结果实时传输至后台监控端。在出现异常时，能够立即发出告警信息，自动关联设备位置信息。实践表明，智能巡检机器人按预期完成了测试目标，并可以提供高强度的巡检频次，支持巡检记录回溯，便于运维人员更加精细化判定故障设备位置和查看历史状态。支持巡检报表输出，提供可读性更高、信息更为完善的巡检报告。

表 4：设备状态巡检验证

设备异常状态描述	功能验证总结
红色或橙色常亮（异常需告警）	现场声光告警和后台同步推送告警
红色或橙色常亮（正常需屏蔽）	通过告警管理进行屏蔽，不再重复该告警
红色或橙色闪烁	现场声光告警和后台同步推送告警。理论上存在极低概率漏报该异常，机器人单次采样后进行图片分析，当采样瞬间设备红色或橙色灯熄灭时会判定设备正常，可通过多次采样解决
全灭（异常）	现场声光告警和后台同步推送告警
全灭（设备下架）	现场声光告警和后台同步推送告警

表 5：设备状态巡检工作对比

对比项目	智能机房巡检机器人运维	人工运维
设备位置异动	设备位置异动发现更高效。机器人巡检时可发现设备位置异动，如设备下架或新上架设备	需核对现场机柜信息图发现设备位置异常，实现过程繁琐
异常设备信息关联	异常告警时自动关联故障设备系统等级、设备名、IP、位置、型号等信息，故障响应更便捷	需查看设备标签，并查询相关文档才能获得异常设备详细信息
告警上报	设备巡检时实时判定异常，现场和后台监控同步发出告警，告警响应更迅速	发现问题需要电话或信息联系后台二线处理，需要一定时间
告警内容识别	机器人运维巡检目前无法识别到故障灯含义，需人工现场查看。有线缆遮挡情况下无法识别到状态灯是否正常	发现故障后现场查看故障灯标识确认故障类型
结果一致性和工作标准化	巡检作业标准化程度高，巡检结果一致性较强，服务质量和效率更可控。巡检结果无主观性判断，标准化数据输出	巡检人员流动带来不确定性，新人员培训上岗需要一定时间，服务质量和效率存在不确定性。巡检流程标准化需借助第三方工具实现
高频度巡检	巡检结果准确率高，适合交易时间高频度巡检。指示灯故障识别基于算法，无主观因素干扰，准确率高，支持不低于 6 小时连续巡检，可以在不影响准确率的情况下实现高频巡检	巡检结果存在一定主观性，高频度巡检需投入大量人力
信息展示和资料调阅	巡检结果后台实时同步展示，支持视频调阅记录，资料调阅更便捷，生成报表巡检结果更直观	人工巡检纸质记录需长期保存，历史调阅较麻烦，工作标准化需借助第三方工具或软件
设备位置异动	设备位置异动发现更高效。机器人巡检时可发现设备位置异动，如设备下架或新上架设备	需核对现场机柜信息图发现设备位置异常，实现过程繁琐

表 6：监控与施工陪护工作对比

对比项目	智能机房巡检机器人运维	人工运维
现场远程交互	陪护现场无需人工，监控画面实时传输至后台，可远程查看现场状态和发送语音	实现该功能需外持录像设备和建设相关系统
风险预警	监控时可呈现热成像画面，超过阈值会告警，避免失火风险	靠主动感知或传感器判定异常 陪护工作调度更灵活。可根据现场情况
多点监控陪护	功能过程较复杂，一个任务中仅可设置一个监控点。如一个机房内有多处需要监控则需设置多个定时监控任	在多个施工区域内流动陪护
资料调阅	机房信息查阅和溯源更可靠。监控录像保留时间可自定义，支持任意时间段调阅	缺乏现场信息保留回溯手段

3.4 场景四：监控与施工陪护

基于运行安全考虑，在非用户单位施工时用户需对相关区域进行巡视或驻点监控，传统人工陪护耗时耗力，且配合录像传输设备才能支持画面实时呈现。智能机房巡检机器人可实现监控和施工陪护一体化，新增任务时可设置任务名称、数据中心、机房、监控点、监控用时，并可预先设置机器人监控朝向。

较人工陪护，智能巡检机器人机房监控和施工陪护更具实用性和监控过程调阅更便捷，且监控视频支持长期存放，信息回溯更便捷。

4、实践总结

本次实践表明，可通过智能巡检机器人实现精细化的机房环境监控，提升运维巡检精度，高效预防微发、渐发环境异常风险；通过智能巡检机器人实现设备指示灯状态查看、实况显示等，提高故障识别效率和巡检效率；通过智能化巡检机器人实现场地施工陪护、安全巡逻、异常闯入告警等，提升运行安全保障和告警效率；提供统一告警平台，整合基础环境、设备运行状态监控和提供分析数据报表，为统一决策提供支撑。智能机房巡检机器人在提高安全运行保障的同时极大减少了现场运维工作量，让机房状态实时可视可管。智能机房巡检机器人的应用可助力运维团队解决人员流动带来的不确定性，以高效率、高质量、高标准完成一线运维巡检陪护类工作，将有限的人力解放出来完成智能巡检机器人能力之外的工作，从而专注于 IT 运维创造性工作。

在多数据中心运维格局下，使用智能机房巡检机器人处理数据中心日常事务，可一定程度上降低运维成本，减少人员往返多个数据中心的频次，并且在特定工作领域的作业效率、质量、流程标准化、过程回溯等方面较人工也具备一定的优势。在采购模式上，可以一次性购买整体解决方案所需的软硬件和服务，也可以向厂商购买相应的服务，按季度或年度支付服务费用。

5、结语

随着行业科技发展和人工智能在行业应用规模不断扩大，传统的人工运维将逐步转向以机器与人结合的增强型混合智能系统，即用机器人、专业运维人员和信息系统结合成的群智系统，驱动数据中心运维智能化转型。基于人工智能的智能机房巡检机器人满足大规模、高等级、业务驱动硬件的新一代数据中心运维需求，并逐渐实现从“巡查”到“操作”的技术革新，为数据中心运维模式升级注入新动力。

一种基于机器学习的攻击源画像构建方法

李骏韬 / 上海证券交易所 信息科技部　邮箱：jtli@sse.com.cn

随着网络安全形势日益复杂，主动式防御已经成为安全体系建设趋势，针对目前第三方威胁情报库大而全但数据质量不可靠，缺乏定制化信息等不足，本文提出了一种基于机器学习的攻击源画像构建方法，可以对攻击源攻击行为的时间、空间、频率、手段等特征进行梳理抽象，明确攻击源类型，有效筛除低效信息，与第三方威胁情报库提供的情报进行比对和互补，为攻击应对措施提供决策支持，降低应对判断难度，提高响应及时性，同时也能为后续攻击溯源提供辅助信息。

1、引言

近年来，网络空间安全形式日益严峻，网络攻击呈现出专业化、复合化、持续化的特征，而金融行业作为关键信息基础设施的重点领域，更是网络安全事件的重灾区，网络攻击数量常年位居各行业前三。针对这一情况，各国政府都高度重视关键信息基础设施尤其是金融行业的网络安全建设工作，出台了大量法律法规与指导意见。

面对当下日益严峻的网络安全形势，传统的通过防火墙、IPS、杀毒软件等软硬件所组成的被动防御体系已经无法完全满足网络安全防护的实际需求，在此基础之上，引入大数据、机器学习、威胁情报等新兴技术，建立更加主动的网络安全防御体系显得尤为重要。

本文结合了金融行业核心机构的网络安全实

践的现状，以构建攻击源画像为主要研究方向，提出了一种基于机器学习的攻击源画像构建方法。通过该方法，可以对攻击源攻击行为的时间、空间、频率、手段等特征进行梳理抽象，明确攻击源行为特征类型，有效筛除低效信息，与第三方威胁情报库提供的情报进行比对和互补，为攻击应对措施提供决策支持，降低应对判断难度，提高响应及时性，同时也能为后续攻击溯源提供辅助信息。

2、相关实践现状

2.1 第三方威胁情报

目前，第三方威胁情报产业经过数年的发展，已经形成了较为成熟的产品和市场，基于通用的 STIX、TAXII 等标准和规范，可以实现通用的机器可读威胁情报，并在不同场景进行应用。但是厂商往往只提供了标准化的信息访问服务和功能，无法针对用户提供定制化的情报，并没有真正提供满足预期的内容和管理服务。此外，不少厂商所提供的威胁情报服务并未对情报本身进行可靠的质量审核，甚至基本不对情报本身进行审核。笔者曾在工作实践中发现，某厂商的威胁情报产品将全球最大做市商之一的 Optiver 官方网站识别为恶意网站的情况，且该公司网站并没有被披露曾经遭受过恶意攻击或篡改导致存在安全风险的情况。面对内容繁杂、规模庞大、关联复杂、数据质量不稳定的第三方威胁情报，大部分机构并还没有真正行之有效对其进行充分利用。

2.2 攻击溯源

近年来，随着网络安全防护体系由被动挨打逐步转为主动防御，攻击溯源也自然而然地成为了网络安全建设体系中的极为重要一环 [12]。目前，攻击溯源已经形成了一些较为成熟操作模式，同时也形成了一些自动化脚本和工具可供安全人员提高溯源效率。但总的来说，攻击溯源当前仍然依赖于经验丰富的安全技术人员，需要安全人员投入大量的时间和精力进行人工挖掘和探索。在日益复杂的网络环境下，大部分企业和机构的安全人员疲于简单应对网络攻击，在核实了攻击后，通常简单地对直接攻击来源封禁了事，难以投入足够的人力物力，深入开展攻击事件分析，更不用说投入大量的人力物力用于对攻击来源进行溯源甚至反制。

3、攻击源特征提取

3.1 攻击数据来源

如前文所述，目前威胁情报库的存在内容繁杂、规模庞大、关联复杂、数据质量不稳定、与用户实际需求存在差距等问题。为了更有效针对实际生产环境中所面临的网络攻击行为，本文以生产环境 SOC 平台所收集、过滤、统计和聚合

表 1：SOC 平台安全告警信息及说明

信息	含义
时间	攻击发生具体日期和时间
受害 IP	攻击目标 IP
攻击 IP	攻击来源 IP
一级告警类型	分为侦查、攻击利用、拒绝服务和恶意软件
二级告警类型	进一步细化攻击类型，包括 SQL 注入、代码执行、漏洞执行、弱口令等 30 多种攻击类型
威胁名称	具体攻击手段、威胁名称，例如具体漏洞号、协议漏洞利用等详细信息
威胁级别	由高到低分别为危急、高危、中危、低危
次数	聚合于本条告警（短时间内目的 IP、源 IP、攻击手段等信息均保持一致）的恶意攻击次数
资产组	事先定义的内部 IP 所属资产，如办公系统、互联网入口等

得到的安全告警作为原始数据来源展开了研究。原始安全告警数据主要包含了表 1 所列的各项信息。

与第三方威胁情报库提供的威胁情报信息相比，本地 SOC 信息能够更加详细的记录攻击行为特征，具有更细粒度的攻击手段记录、攻击目标选择性、攻击的时间分布特征、攻击频率具体细节等详细信息。本文在研究和实验阶段，共选取了 2021 年某月一整月时间范围内实际生产安全告警数据，共计 199,710 条告警记录作为研究对象。

3.2 攻击源特征选择

通过构建攻击源画像，可以为攻击响应、应急、溯源、取证等提供有效快速的决策辅助，确定后续工作方向。为了实现攻击源画像的构建，就需要对攻击源的行为进行有效地抽象和凝炼，提取出隐藏在海量攻击行为中的共性特征。本文主要基于原始告警信息的统计数据，同时结合 Maxmind 的 IP 地理信息库（GeoIP2）、洛克希德 · 马丁公司的杀伤链模等作为外部知识对原有信息进行了扩充，最终形成了以下 11 个攻击源特征信息（详见表 2），用于对攻击源进行描述。

4、一种攻击源画像构建方法

目前，第三方威胁库在众多用户、厂商、白帽子等多方的共同努力下，已经针对恶意攻击源形成了较多标签，例如 RAT、C&C、扫描、僵尸网络等等，可以有效地供安全人员参考。但是这些标签都是针对攻击源主要手段的描述，缺乏其他攻击频度、强度、广度、时间和目标选择偏好等攻击行为特征信息。本文将利用前文所提到的攻击源特征，基于机器学习算法来构建攻击源的画像。

通常来讲，攻击来源画像构建可以通过对已经清洗转换并抽象提炼的特征，经过一系列的逻辑判断，最终完成分类，也即专家系统。但是专家系统严重依赖于判断逻辑的选择，容易带入专家系统构建人员的偏见，也无法根据实际情况变化及时进行更新迭代。

针对专家系统的不足，本文采用了无监督学习 – 人工干预 – 有监督学习 – 优化 – 建立模型的方式，构建形成了攻击源画像模型。

4.1 攻击源特征值归一化

不同的攻击源描述特征具有不同的维度和取

表 2：攻击源特征及说明

特征	说明
国家	根据 GeoIP2 信息，对应获取攻击源 IP 的国家信息
主要攻击发生时段	攻击目标 IP 在过去 7 天主要攻击时段，结合证券交易（含港股通）特点，9:00–11:30,13:00–16:00 为交易时段，9:00–18:00 的非交易时段为工作时段，其他时间为非交易时段
日均攻击次数	过去 7 天日均攻击次数
日均攻击次数标准差	过去 7 天日均攻击次数标准差
攻击种类	过去 7 天二级告警攻击类型总数
主要攻击阶段	过去 7 天攻击主要所属杀伤链阶段
目标总数	过去 7 天攻击的不同目标总数
平均威胁级别	过去 7 天平均威胁级别
威胁级别标准差	过去 7 天威胁级别标准差
单一目标平均攻击次数	过去 7 天针对单一目标 IP 的平均攻击次数
单一目标平均攻击次数标准差	过去 7 天针对单一目标 IP 攻击次数标准差

值范围，为了更好进行模型计算和调优，尤其是为了更好的计算各个攻击源之间的距离，避免因为部分特征数值过大导致的距离计算权重过高，需要对攻击源特征进行归一化处理。本文所选取的特征的归一化方法如表 3 所示。

4.2 基于聚类算法的攻击源类别生成

为了尽量减少专家系统的构建过程中由于专家的认知偏见导致的偏差，本文首先采用 K-Means 聚类算法，以归一化后的特征值作为样本坐标，生成了攻击源画像的基础类别（簇）。笔者以 D1-D16 的告警数据作为原始数据，经过统计处理得到了 D7-D16 共 10 组统计数据，将这 10 组统计数据作为训练集进行聚类运算。为了有效进行类型区分，簇数量设置为较大的 20，共计得到了 200 个簇中心坐标。笔者对得到的 200 个簇中心坐标欧几里得距离（见公式 1）进行了计算，共计得到 19900 个簇中心距离。

$$d=\sqrt{\sum_{i=1}^{11}(x_{ip}-x_{iq})} \quad (1)$$

经统计，通过 K-Means 聚类算法得到的簇中心欧氏距离均值为 3.17，最小值为 0.19，最大值为 11.96，分布如图 1 所示。

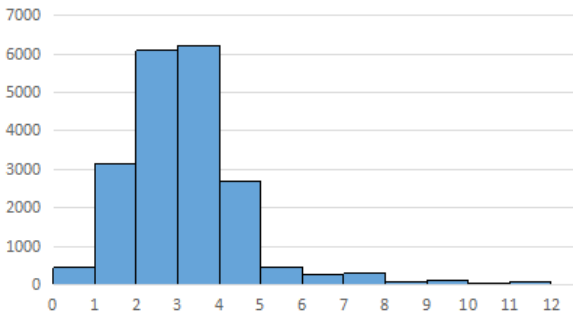


图 1：簇中心欧氏距离分布

笔者将第一轮聚类得到的簇中心坐标全部赋权重为 1，把簇中心距离最近的两个簇进行合并，新的中心坐标为用于合并的两个簇中心的加权平均，新簇中心坐标的权重为用于合并的两个簇中心权重之和。经过 150 轮迭代，最终得到了 50 个簇中心。合并后，簇中心欧氏距离均值为 3.65，最小值为 1.04，最大值为 11.55，分布如图 2 所示。

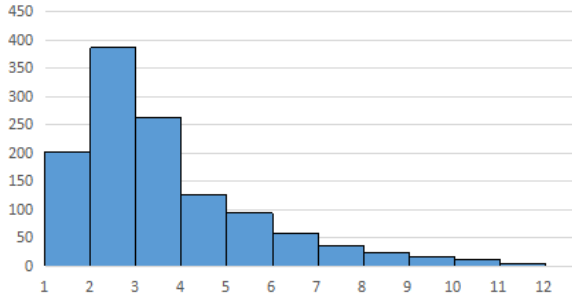


图 2：合并后簇中心欧氏距离分布

表 3：攻击源特征归一化方法

特征	归一化方法
国家	取训练集全部样本的各个国家的日均攻击次数为 c，取$\lfloor lb(c)/2 \rfloor$作为各个国家的归一化特征值，并将 >3 的设为 3，未在训练集中出现过的新增的国家则设为 0
主要攻击发生时段	交易时段为 2，非工作时段为 1，工作时段为 0
日均攻击次数	过去 7 天日均攻击次数为 c，取$lb(c)/3$，若 >10 则设为 10
日均攻击次数标准差	过去 7 天日均攻击次数为 c，标准差为 σ，取 σ/c
攻击种类	过去 7 天二级告警攻击类型总数 n，取 n/5
主要攻击阶段	根据杀伤链顺序，以 0.5 为步进，在 0.5-3 进行赋值
目标总数	过去 7 天攻击的不同目标总数为 n，取$\lfloor lb(n)/2 \rfloor$
平均威胁级别	过去 7 天平均威胁级别，以危急 =4，高危 =3，中危 =2，低危 =1 计算
平均威胁级别标准差	过去 7 天平均危险级别为 d，标准差为 σ，取 σ/d
单一目标平均攻击次数	过去 7 天针对单一目标 IP 的攻击次数为 n，取$\lfloor lb(n) \rfloor$，若 >10 则设为 10
单一目标平均攻击次数标准差	过去 7 天单一目标攻击次数均值为 n，标准差为 σ，取 σ/n

4.3 基于分类算法的攻击源画像

为了形成更加简单、易用、具有工程应用实践意义的攻击源分类方法，需要进一步根据现有的聚类结果，抽象形成由一系列判断语句组成的分类算法。

首先，笔者基于聚类算法的簇中心坐标，并根据样本与簇中心的欧氏距离，对 D7-D23 的样本进行了标注。同时，与聚类算法不同，不对样本特征值进行归一化基本不会对机器学习产生的分类判断语句产生影响，但却可以有效增加判断逻辑的可读性。因此，在进行分类算法运算时，笔者主要采用了未进行归一化的数据作为样本属性（其中，国家，阶段，时间段等非数值信息仍然进行了数值化处理）。

综合考虑到生产实践需求，笔者舍弃了随机森林、支持向量机等分类效果好，但是分类方法复杂的分类算法，选择了分类判断逻辑更为简单明了的随机树生成算法。同时，为了避免生成过于复杂和庞大的决策树，笔者将决策树的最大深度设置为 6（若深度小于 6，则无法生成具有 50 个以上叶子节点的决策树，也即无法得到与现有类别相匹配的具体分类算法），最终得到了攻击源分类决策树如图 3 所示。

基于随机树生成算法所生成的分类逻辑，笔者选取了 D24-D31 的数据作为测试集，进行了攻击源分类。由于攻击源画像和分类难以进行明确的客观测试和评估，笔者随机选取了 125 个攻击源 IP 样本及其对应的分类结果，补充了相关原始告警以及所属类别（簇）属性等相关信息，邀请 5 名安全人员分别对 50 个样本的分类结果进行了评估，每个攻击源的分类结果都有 2 名安全人员进行了评估，评估结果如表 4 所示。

表 4：攻击源画像安全人员评估结果

评估结果	数量（占比）
两名安全人员认为分类较恰当	97（77.6%）
一名安全人员认为分类较恰当	16（12.8%）
两名安全人员认为分类不恰当	12（9.6%）

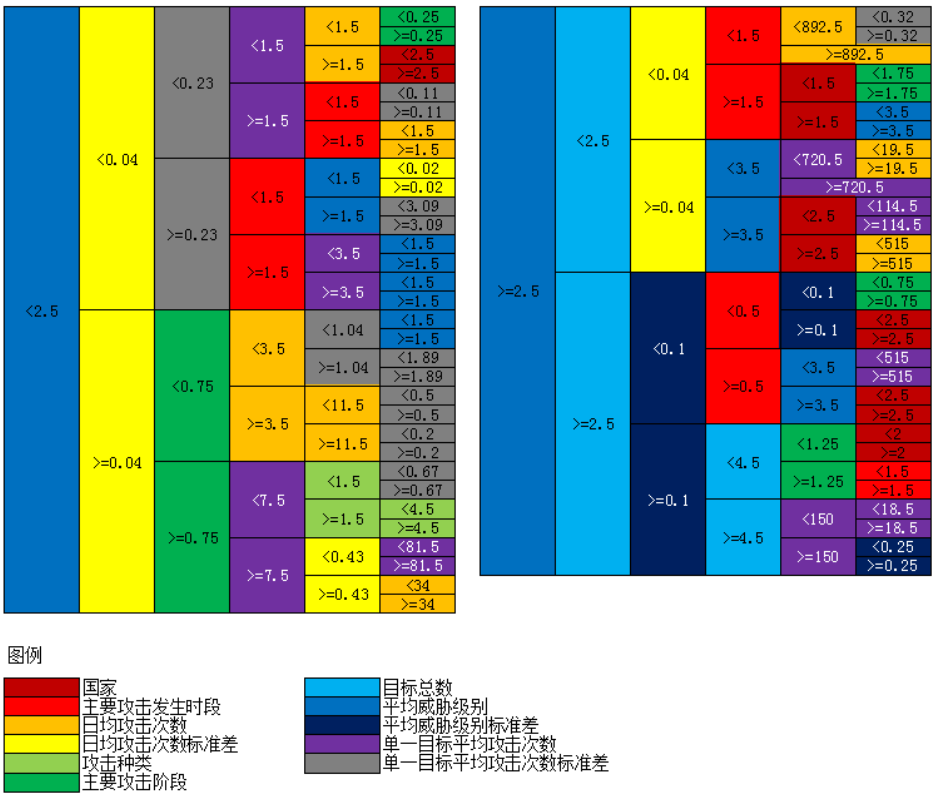


图 3：攻击源分类决策树示意图

5、结语

通过本文提出的攻击源画像构建方法，可以较为有效利用各机构自身所收集到的第一手攻击信息，构建外部攻击源画像。通过构建的攻击源画像，安全人员可以基于该攻击源的历史行为特征，快速有效地判断威胁的紧急和严重程度，更快地进行安全应对措施决策。

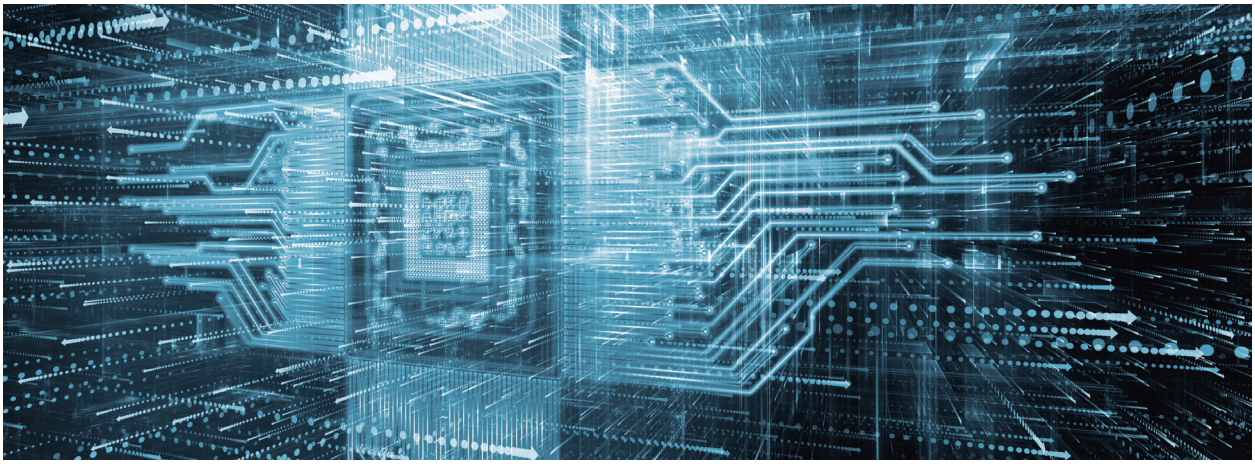
不过，目前的攻击源画像主要采用了离线方式进行计算和构建，尚处于实验室阶段。下一阶段，笔者将与本机构的一线安全人员一起，尝试将攻击源画像构建与 SOC 平台进行整合，尽快达到实用化。

同时，在实验中笔者也发现，目前构建的攻击源画像受限于 SOC 平台提供的数据，该部分数据已经进行了高度抽象，大量原始攻击特征信息已经丢失。由于缺乏攻击源的攻击手段详细信息，无法形成攻击源和攻击行为指纹，目前只能为即时攻击应对决策和初步溯源工作的提供支持，对于深入挖掘、分析和溯源的参考和辅助意义相对有限，后续笔者也将结合 SOC 前的原始告警信息开展进一步的研究工作。

实践探索

海通证券数据中台建设思路、方法与实践

吴保杰、许红涛、于鹏、蔚赵春、王晓平、陆颂华、朱元元 / 海通证券股份有限公司
邮箱：yzc12673@haitong.com



伴随着金融科技的快速发展，证券行业已经进入了数智化发展的快车道。作为行业金融科技先行者和探索者，海通证券践行“数字化转型”战略，建设了数据中台“e 海智数”，以此为基础提供统一数据服务云 DaaS 和进行数据治理，逐步打造开放式数据生态体系，将数据变为资产并服务于业务，数据来源于业务并反哺业务，数据赋能业务价值逐步显现，实现了数据可见、可用、可经营。

1、数据中台建设背景

2020 年 4 月 9 日，中共中央、国务院印发了《中共中央国务院关于构建更加完善的要素市场配置体制机制的意见》，明确将数据上升为与土地、劳动力、资本、技术并列的新型生产要素，数据逐渐成为数字化时代的核心竞争力。证券公司在数字化时代机遇和挑战并存。从外部环境看，我国经济增长模式转变、对普惠金融的需求、新冠疫情影响、“数字新基建”等均会加速证券公司数字化转型；客户需求变化速度越来越快，从线下到线上，个性化需求越发凸显，传统业务模式已无法满足客户需求。监管对数据的重视及监管科技对数据应用提出了更高的要求。金融科技的发展推动了数据业务化，引发了证券业基于数据的产品和服务创新。因此，数字化转型过程中快速多变环境对数据应用能力提出了新的挑战。从内部来看，海通证券面临的挑战在整个证券行业都普遍存在。

（一）数据孤立，联动困难，没有形成数据资产

内外部数据没有打通，数据标准不统一，无法形成统一的数据资产，数据没有融入到业务价值创造过程，没有赋能企业创新与转型发展。证

券公司业务天然与数据紧密相关，数据已逐步成为证券公司最重要的无形资产。目前证券公司交易数据、客户数据、风险数据、行为数据、产品数据等基础数据较为完善，但是市场数据、工商数据、舆情舆论等外部数据仍有缺失。同时，数据长期沉淀在各个系统中，大部分数据应用目前局限于数据初级使用，只停留在描述型分析和诊断型分析，只有很少一部分数据可能会涉及预测型分析和规范型分析，数据资源存在一定的浪费，更深层次的价值尚未被发掘。

（二）技术、业务能力无法沉淀

杂乱的技术栈使用是技术变革之路上的阻碍，更无法基于标准统一的技术资产形成合力。证券行业发展日新月异，业务发展到哪里，系统就要建设到哪里。由于公司级架构管控机制缺失，加上系统供应商纷杂、技术标准不统一、业务管理条块化等问题存在。业务条线各自建系统，各自管系统，导致各个系统在功能、流程、数据标准以及技术体系方面缺乏统一规范和管理控制，技术和业务能力也就无法沉淀。系统的分散阻隔了数据之间的互通，增加了数据需求方获取信息的难度，加大了统一操作、统一管理的难度。

（三）研发机制有待完善，无法快速响应业务需求

需求响应时间长、协同成本高，缺少敏捷和精益实践。目前虽然证券行业已经对加快数字化转型达成共识，但是部分证券公司数字化转型机制有待完善，全员数字化思维还没有完全形成，对于数字化赋能业务没有清晰的认识，导致执行力不足，影响转型进度和效果。部分人员认为数字化转型只是 IT 的转型升级，业务条线与 IT 的沟通是以系统为单位，以项目为触点，缺乏业务视角的沟通互动机制，数字化对业务发展的引领支持不足。

2、数据中台建设思路

海通证券数字化转型围绕以客户为中心的“12345”战略展开，以集团化、国际化和信息化为驱动力，打造智慧海通。数字化转型离不开数据应用，而数据中台是数据应用能力的沉淀和积累，最终要实现数据的标准化、可视化、资产化、服务化、业务化。海通证券围绕公司的大数据战略，综合考虑公司的业务模式和组织架构，基于科学的方法论，以“一条主线、两个引擎、五位一体”为基本思路，建设了企业级数据中台“e海智数”，实现以客户为中心洞察驱动和以数据为基础的业务创新。

一条主线是指以客户为中心，以数据为基础，以科技为驱动，将公司的数据资产进行梳理、共享、挖掘，让数据资产作为生产资料融入到业务价值创造过程中，全面赋能客户、产品、风控、合规、运营等各业务领域，实现基于数据驱动的客户旅程重塑。数据驱动包括：进行深度客户洞察和精准客户画像，基于数字画像提供千人千面的解决方案；优化客户体验和产品研发能力，提高公司运营效率和成效；实现业务链条数字化和管理精细化，前置风险、合规和管控措施，防患于未然。

两个引擎是指“数据经营”与“金融科技”。数据中台核心是将数据作为企业重要的资产，让数据资产围绕业务价值创造的目标更好地流动、加工、分析、应用，最终产生效益和价值。数据中台以金融科技为依托，融入业务场景，产生技术数据融合的双源创新。数据中台延续了“1+3+N”大数据战略，一是通过完善大数据基础设施提升数据应用能力；二是在人工智能平台、数据管控平台、报表与分析平台基础上新增了数据开发平台、数据探索及挖掘平台、数据服务门户等；三是数据服务范围更加广泛，不仅为母公司各种业务场景赋能，而且还对子公司、合作伙伴赋能。

五位一体是指数据中台建设涵盖数据、技术、人员、服务、治理五大方面。(1) 数据是基础。数据中台起源于业务数字化，最终目标是把原本各自孤立的数据互联互通，构建数据资产体系，挖掘数据更深层次的价值。(2) 技术是驱动。数据中台离不开云计算、人工智能、大数据等技术提供驱动力。(3) 人员是关键。数据中台建设需要具备数字思维，特别是业务人员具备科技思维，统一理念，形成数字化合力把数据用起来。(4) 服务是核心。将数据资产以服务形式对外提供，数据嵌入到业务场景中随需而动，实现以数据驱动业务，激发数据动能。(5) 治理是保障。数据治理作为数据中台的保障机制，其目标就是数据“好用”和“用好”数据，合规、高效地产生数据价值。

3、数据中台总体架构

海通证券数据中台是对“1+3+N”大数据战略的深化，不仅对大数据基础设施进行优化、工具平台进行了扩充，新增了数据快速开发平台、数据挖掘与探索平台、客户画像与行为分析等，服务的范围更加广泛，为母公司、子公司、合作伙伴赋能。海通证券数据中台总体架构如图 1 所示，主要包括：数据源、数据采集交换、数据存储计算、数据工具、数据服务云、数据应用、数据治理、服务门户等。

3.1 数据源

数据源是内外部数据融汇贯通形成统一完整的数据资产基础，分为内部数据和外部数据，核心是全面及时地收集“初数据”。数据收集从现有各业务系统数据入手，从母公司延伸到集团分支机构及子公司，从业务数据到各种衍生数据，从一般性数据到独辟蹊径的“初数据”，从结构化数据扩展到非结构化数据，从公司内部数据外延到互联网、数据供应商等外部数据。

数据中台融合了海通证券内外部数据。内部数据纳入了渠道接入层、技术支撑层、业务处理层、业务支持层、决策分析与管理层等主要系统的结构化业务数据及应用日志、审计日志、手机行为日志等结构化和（半）非结构化数据，同时纳入了集团分支机构及子公司产品、并表等业务数据。外部数据接入了资讯、工商、舆情等结构化和（半）非结构化数据，涵盖股票、债券、基金、外汇、美股、期货、金融衍生品、现货交易、

图 1：数据中台总体架构图

宏观经济、财经新闻、市场新闻舆情、研究报告、公告等各类数据。

3.2. 数据采集交换

数据采集交换作为数据中台的核心基础能力之一，通过打通集团、子公司、外部数据互联融合的通道，实现异构和不同数据源的数据整合和交换，保证数据及时性、一致性、准确性以及跨业务、跨组织的需求，功能包括数据抽取、数据清洗、数据标准化、数据转化、数据校验等。数据采集交换的重点工作是标准化，包括统一接入方式、消息传递机制、数据标准等。考虑到证券公司信息系统部分由合作厂商开发，和数据中台对接过程中如何与上游数据结构保持一致一直是一个痛点，通过数据采集交换平台准实时的刷新源系统的元数据信息，及时获取元数据异动信息来快速的生成配置信息有效解决了这一难题。

传统的数据交换主要采用磁盘映射、SCP 远程传输、开源的数据库同步产品等方式实现对不同地区、不同平台之间的传输，存在侵入性强、迁移脚本化、安全稳定性差、缺乏有效监控、无可视化界面等方面的不足，无法满足企业级大规模推广。为了实现数据交换的实时、安全和智能化，结合公司两地三中心之间数据存储、传输、灾备、治理需求，构建了具有低延时、安全可靠、全生态、智能化、国产化特点的数据实时交换平台，能够实现公司在各中心、云平台等异构环境下对应用数据进行低延时的采集、转换、融合、分发和监控等，使各应用的数据交换更加方便快捷，发挥最大的数据价值，最终赋能前端业务。不管是文件系统还是数据库系统，均通过统一可视化平台管理界面，实时掌握文件数据、数据库数据的同步状态。除了同步规则的灵活设置，还通过对象比较、对象修复等手段确保数据的一致性。在数据同步的过程中实时监控同步状态、结果等，且整个产品安装步骤、同步规则设置等操作简便轻松。

数据采集交换平台兼容结构化及非结构化数据配置化，具有以下特点：(1) 具备多样化的数据集成配置能力，兼容多种通用数据格式（如 json、xml 等），支持不同数据库源的上游系统转化为传统的 Datastage 采集模式及开源的 DataX 采集模式，并能自动生成下游数据库所需的数据定义语句，保证了上下游表结构的一致性。(2) 具备多类型数据库的对接能力，实现了对 Oracle、SQLserver、Mysql、DB2 等传统数据库和 Hive/HBase/GaussDB 等新型数据库的无缝对接。(3) 具备可视化快速配置能力，提供图形化的开发和维护界面，支持图形化拖拽式及流程式低代码开发模式，减少代码编写，降低开发难度，大幅提升数据交换需求响应的时效。(4) 具备数据抽取与转换策略的配置化能力，通过表字段映射规则设计、信息编码对应规则设计、清洗规则设计、转换规则设计、装载规则的定制，来实现整个数据抽取转换。(5) 具备数据服务异常自动检测能力。

3.3 数据存储计算

数据存储实现了贴源数据、基础模型、轻度汇聚和公共维度等多层次数据共存，并按照业务领域建立了数据集市和数据应用，实现对数据资产的统一管理。数据存储计算层不仅要处理“初数据”，更要借助金融科技“粗加工”形成“新数据”和“精数据”。例如，应用人工智能技术手段将非结构化数据转变为以特征变量为基础的结构化数据就是“新数据”。“精数据”是指融合金融知识形成市场、风险、运营、财务等领域指标的数据，便于业务人员认知并在业务运营及决策中发挥作用。

数据存储主要包含：(1) 贴源数据，即各源系统数据保留区，负责保存数据接入时点后历史变更数据。(2) 基础模型数据，即整合后的业务过程明细数据，按主题域划分，构建数据标准；基础模型层具备高可读性、高扩展性、有效保留各类数据历史、高业务兼容性等特点。 (3) 轻度汇

总数据，即对共性维度指标数据进行轻度聚合，形成公共指标体系；轻度汇总层面向分析主题设计维度数据模型，可灵活、高效地支持各应用功能需求。(4) 数据集市数据，即面向业务定制的应用数据，主要有客户集市、财务集市、风险集市、运营集市、自营集市等。

数据计算引擎是数据中台的心脏，结合公司业务场景和业界技术发展，数据中台沉淀形成了涵盖高速实时计算处理、离线数据处理、数据查询引擎、图数据库等全面强大的数据计算层。(1) 实时计算处理方面，以 Flink 为主，Spark 为附，形成统一的实时数据计算引擎。(2) 离线数据处理方面，通过 GaussDB、Hive、Impala 实现统统的离线数据处理需求。(3) 数据查询方面，通过 HBase 实现高频结构化查询，通过 ElasticSearch 实现非结构化查询，通过 OpenTSDB 实现时序数据等特殊查询需求。(4) 图数据方面，以 Neo4j 为主形成企业级的图数据库计算能力。

3.4 数据工具

“工欲善其事，必先利其器”，完善的工具平台是数据中台发挥价值的利器，为了实现“数据来源于业务并反哺业务”，数据中台自主研发了一系列工具平台，主要包括数据快速开发平台、报表与分析平台、数据挖掘与探索平台、人工智能平台、数据管控平台等。(1) 数据快速开发平台主要实现数据处理全链路开发，包括数据采集、数据清洗、数据加载、数据处理、数据推送等，内含丰富的开发组件，支持低码开发模式。(2) 报表与分析平台提供了数据分析的手段及报表展现的途径，为各业务条线提供以数据为支撑的报表增值服务。(3) 数据挖掘与探索平台负责为数据科学家、数据分析师、业务人员提供稳定、高质量的跨主题数据沙箱环境，集成统计分析、知识图谱、NLP2SQL 等工具集，结合不同的场景积累相对成熟的数据能力解决方案。(4) 人工智能平台涵盖大规模机器学习和深度学习框架，为各类应用提供 AI 引擎。(5) 数据管控平台建立了统一的企业指标库，有效地管理数据资产，分析数据加工关系，绘制数据地图，发现数据质量问题，支持数据标准的规范治理。

3.5 数据服务云 DaaS（Data as a Service）

数据服务云作为公司数据中台“e 海智数”的重要组成部分，为前台数据应用提供统一的、面向应用、主题式的数据服务，将数据资产以服务形式对外提供，实现了数据即服务 DaaS(Data as a Service)，同时提供以业务为导向的数据服务能力地图，让前台应用更清晰的使用中台的各类数据，实现以数据驱动业务，形成助力前台、连接后台的服务能力。

数据服务云实现了统一的数据服务管理、多元异构数据 API 服务化，达到了数据标准与统一、数据共享和服务能力共享，解决了传统数据搬运和指标不一致的交互模式，加强了数据安全。统一数据服务提供了将来自不同源头、不同形式的数据发布成标准服务的能力，建立自助式数据共享服务，提供全局服务视图以及完整的数据共享链路。数据中台具备监控数据服务整个生命周期过程的能力，出现异常时将及时告警。数据服务云采用微服务理念研发，具有高并发、高可用、可扩展的特点，使数据共享过程更高效、更稳定、更安全。数据服务云主要具备以下功能：(1) 标准规范的数据服务 API 接口：为前台各应用系统提供标准规范的 API 接口，支持单数值、多数值、单记录、多记录、分页请求等常见数据调用。(2) 高并发联机数据查询服务：支持高并发毫秒级数据查询服务，大批量、复杂逻辑的数据查询的响应时间在秒级。(3) 支持复杂业务逻辑处理：对前台查询的数据，数据中台可在内存中根据业务逻辑计算加工、统计分析，返回最终分析结果。(4) 数据服务管理可视化：实现了数据服务的接入管理、服务审核、用户管理、权限管理等。同时实现数据服务实时监控，包括服务状态、流量、使

用情况的监控统计，以及异常处理、风险控制等。此外通过可视化方式实现数据服务接口的开发，包括数据服务的开发、登记、编号、发布、维护、更新等。

3.6 数据应用

拓展数据应用融入到业务场景是数据中台落地发挥价值的关键。数据应用分为两个层面，第一层面是业务数字化，包括数字化营销、数字化风控、数字化管理、数字化投研、数字化运营等。第二层面是数字业务化，探索数据驱动业务转型，利用数据创造新的业务模式，实现新的运营模式。数据应用依赖于数据意识、全局意识和回馈意识，以形成使用、反馈、评价、优化的闭环。数据应用不仅需要深刻理解数据多样性和相关性，深挖业务场景。例如客户画像除了赋能精准营销外，还可为客户管理、经纪业务营销、投行营销、财富管理营销等全领域赋能。数据应用综合使用“新数据”、“精数据”与“初数据”，将数据应用水平提升到更高的层级，真正做到数据应用的“学习历史，描述现在，感知未来”。

3.7 数据治理

数据治理是从全局视角统领各个层面的数据管理工作，建立数据拥有者、使用者、数据以及支撑系统之间的和谐互补关系，确保各方都能得到及时、准确的数据服务。数据中台建设数据治理要先行，海通证券专门成立数据治理工作办公室，从数据治理体系建设、数据类项目建设、数据管理、数据应用服务等方面深入开展数据治理工作。其中主要

3.7.1 数据分类分级

为了实现公司数据分类和数据分级的统一管理，基于《证券期货业数据分类分级指引》，数据中台自研了数据分类分级管理功能，构建数据分类分级清单，分析挖掘业务数据和元数据信息，识别客户敏感信息和公司重要数据，厘清数据资产并确定数据重要性。具体功能包括：(1) 数据分类分级清单管理与查询，数据使用审批系统化；(2) 结合 AI 技术探索数据项自动化映射，动态获取数据项分布情况；(3) 聚焦客户重要敏感信息，识别并定位敏感数据在各信息系统的存储分布，形成公司敏感数据地图。业务数据分类分级工作有效提升了数据安全管控能力，根据数据定级结果，针对敏感数据和重要数据，加强数据使用的审批管理，采取数据脱敏、数据库审计等多项技术措施，从数据安全的角度加强了数据资产管理，在保证数据安全的基础上促进数据开放共享。

3.7.2 数据质量管理

数据质量管理是对数据从采集、存储、整合、呈现与使用、分析与应用、归档和销毁的全生命周期每个阶段中可能引发的数据质量问题，进行识别、度量、监控、预警等一系列管理活动，并通过改善和提高管理水平使得数据质量获得进一步提高。数据质量管理是提升数据中台价值的生命线。 海通证券数据中台数据质量管理定义了 5 个质量维度：数据完整性、数据准确性、数据合理性、数据一致性、数据及时性。

- 数据完整性（INT）：主要包括实体缺失、属性缺失、记录缺失和字段值缺失四个方面；
- 数据准确性（ACC）：准确性也叫可靠性，是用于分析和识别哪些是不准确的或无效的数据，主要通过一个数据值与设定为准确的值之间的一致程度，或与可接受程度之间的差异度量；
- 数据合理性（VAD）:主要包括格式、类型、值域和业务规则的合理有效；
- 数据一致性（CON）：系统之间的数据差异和相互矛盾的一致性，业务指标统一定义，数据逻辑加工结果一致性；
- 数据及时性（TIM）：指能否在需要的时候获到数据，是影响业务处理和管理效率的关键指标。主要包括数据仓库 ETL、应用展现的及时和快速性、Jobs 运行耗时、运行质量、依赖运行及时性；

数据中台 e 海智数根据如上五个质量维度，定义了 12 个数据质量检核大类，共配置了事中校验规则 7639 个，事后检验规则 176 个，具体如表 1。

3.8 数据服务门户

数据中台中数据服务门户是一个能够让业务需求人员、数据分析人员等数据使用方和数据工程师、数据管理员等数据提供方共同使用同一套企业数据资产的协作平台，涵盖企业数据目录、数据版本管理、数据沙箱（Sandbox）等，通过统一共享与协作平台实现数据资产的集中管理与数据应用的汇集，包括数据资产目录和全景图、数据服务能力地图、数据应用集中展示、数据服务接口和权限控制、数据探索等功能，最终目标是实现数据可见、可用和好用，建立以数据为中心的“智慧海通”。

（一）数据可见。数据可见的目的是数据消费者可以定位所需的数据。公司经过这么多年的积累，沉淀了大量的数据，通过打造企业级数据门户，梳理数据资产，全面展示元数据、主数据、数据服务接口、数据血缘关系，打通使用数据的最后一公里。

（二）数据可用。主要让数据变得易懂，降低消费者快捷方便检索、使用数据的门槛。一是丰富完善数据模型，形成统一的数据标准，建立数据沟通语言；二是建立数据沙箱，提供数据分析应用的探索环境和配套的自主分析工具；三是从源头和加工过程中提升数据质量，确保经过加工后的“精数据”也具备可解释性。

（三）数据好用。数据好用包含数据质量和数据工具两个方面。一是提高数据质量，让消费者从数据内容体验到数据中台的数据是好用的。二是通用能力工具化，包括重复的工作工具化、自动化，提供数据自主分析工具等。

4、数据中台建设体会

数据中台是一个庞杂系统工程，包含了技术、方法论、人才建设、跨部门协同、数字化思维等，同时大数据、数据湖、云原生、数字孪生、DataOps 等新技术、新概念层出不穷，并叠加数据应用复杂性，数据中台建设不是一蹴而就，要做好持续探索、演进的决心和耐心。首先，坚守初心，破除数据中台迷思，不被繁杂的概念所迷惑，不争论、不折腾，探索中前行。其次，数据中台关键是能力的沉淀，各家企业因地制宜，需要根据业务演进发展逐渐积累，搭建适合企业自

表 1：数据中台质量校验规则统计表

质量维度	检核类别	事中	事后
完整性（INT）	空值检核（NUL）	2	1
	记录数检核（CNT）	4	5
	FIC 触发器检核	188	0
	BDP 加载记录数检核	3992	0
	MPP 加载记录数检核	3424	0
准确性（ACC）	码值检核（DIC）	3	1
	主键重复检核（DUL）	7	4
	非法值检核（ILG）	10	3
合理性（VAD）	业务约束检核（CON）	3	30
一致性（CON）	主外键检核（PFK）	3	20
及时性（TIM）	数据日期检核（DAT）	3	0
	数据超时未更新	0	112
合计		7639	176

身的数据中台，“经营”好全域数据资产，充分发挥数据生产要素作用和价值。再次，以开放的思维、开放的架构加强和业界合作，聚能合作伙伴，融入行业生态，共同推进数据中台的建设。最后，加强具备 5 种角色的数据中台人才培养。一是做好学习者，努力紧跟技术趋势；二是做好布道者，向全员传播数字化知识；三是做好服务员，努力当好服务业务部门的“店小二”；四是做好情报员，努力收集业务信息，发掘用户痛点；五是做好联络员，努力加强上下内外联动协同。

东方证券分布式系统可观测性解决方案探索与实践

黄真正、杨子江、王建、胡长春 / 东方证券股份有限公司 系统运行总部
樊建 / 东方证券股份有限公司 系统研发总部
邮箱：huangzhenzheng@orientsec.com.cn

随着公司业务系统逐步微服务化，微服务带来的分布式架构下可观测性问题越显突出，传统监控手段已不能满足可观测性需求。本文对分布式系统可观测性解决方案进行探讨，通过引入分布式链路跟踪技术，及实现 Metrics、Tracing、Logging 三者融合，很好的解决了系统和业务的可观测问题，并在财富管理领域进行实践，效果明显。

1、引言

微服务架构是近几年受到各行业广泛追捧的技术之一，微服务架构具有轻型化、便捷化、敏捷化等特点，不仅能够适应业务创新和变化的需要，而且易于维护、变更、升级，契合当前证券业务发展的需要。2019 年 6 月，东方证券发布了 gRPC-Nebula 服务治理框架，同年，又公布了“大中台”战略。伴随着公司数字化转型的加快，业务发展的加速，传统后台交易系统已不能满足业务快速上线的需要。为了快速响应业务需求，提供更为灵活的服务支撑，公司对财富管理领域进行了整体架构规划，按照能力边界，形成账户中心、产品中心、财富销售中心、资产中心、交易汇总中心、行情中心及资讯中心 7 个核心业务中台。各业务中台均基于 gRPC-Nebula 微服务框架，并通过服务治理平台进行跨中心的服务调用。在新的分布式架构下，一个前端渠道系统的业务请求，有可能由多个业务中台、多个服务节点配合完成。分布式系统复杂的网状服务调用关系对业务开发、测试、日常运维和业务分析工作带来了新的挑战：

（1）研发人员需要从蜘蛛网般的服务调用关系中梳理出特定业务流程的整条链路拓扑，分析各个服务依赖关系，排序关联服务强弱级别；

（2）测试人员需要从多节点的运行日志中分

析测试执行状况，按时间顺序梳理请求的源头节点，途经节点及异常节点等情况，对测试人员的要求极高；

（3）运维人员需要在业务异常发生时，在分散冗杂的事件日志中，准确地判断故障节点、定位故障原因。且需要人工梳理具体业务的请求耗时和各节点及接口耗时；

（4）业务人员想做全面、准确的数据分析，需要从多个业务中台获取理财销售业务数据，并需要人工关联业务数据，数据量大，关联难度也大。

本文针对上述分布式架构让传统监控方法、观测方法失效的问题，提出了一种分布式系统可观测性解决方案，并在财富管理领域取得了良好的效果。

2、相关概念

2.1 可观测性

可观测性起源于几十年前的控制理论，它是关于描述和理解自我调节系统的，近年来越来越多地应用于分布式 IT 系统。可观测性是通过检查其输出来衡量系统内部状态的能力。如果仅使用输出的信息就可以估计当前状态，则系统被认为是“可观测的”。

图 1 采用简化图的形式描述了系统间的组成及交互。

从上述交互图可看出，系统的交互行为有如下几种形态：

- 系统内，单组件功能闭环，或组件之间交互；
- 系统间，系统与系统间相互进行交互。

这样，若想通过系统的外部输出了解系统的内部状态，就需要两种形态的信息：

1、组件闭环的信息

2、组件间或系统间流动的信息

第一种形态通常可通过 logging 或 metrics 表征，第二种形态就需要在流动的信息中增加标记通过 tracing 来表征。

因此，能够对 logging、tracing、metrics 三种类型的观测数据进行有效融合与表征，即能解决可观测性问题。

2.2 可观测性的三大支柱

Peter Bourgon 在 2017 Distributed Tracing

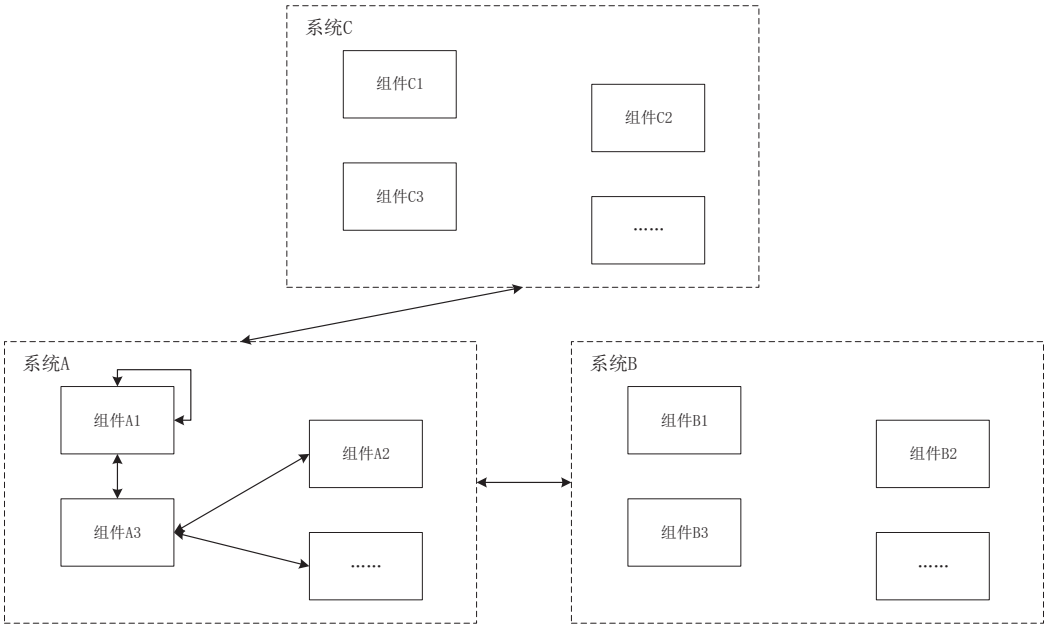


图 1：系统组件间关系

Summit 发表的一篇博文，简洁扼要地介绍了指标（Metrics）、链路（Tracing）、日志（Logging）三者的定义和关系，这三种数据在可观测性中都有各自重要的作用，并相互促进。

• 指标数据（Metrics Data）

提供量化的系统内 / 外部各个维度的指标，一般包括分别 Counter（计数

器）、Gauge（瞬时值）、Histogram（直方图）和 Summary（概要）等。

• 日志数据（Logging Data）

提供系统 / 进程最精细化的信息，例如某个关键变量、事件、访问记录等。

• 跟踪数据（Tracing Data）

提供了一个请求从接收到处理完毕整个生命周期的跟踪路径，通常请求

都是在分布式的系统中处理，所以也叫做分布式链路追踪。一个 Trace 有唯一的 traceID，且由多个 span 组成。

Logging、Tracing、Metrics 三者在解决可观测性问题上缺一不可：基于 Metrics 的告警发现异常，通过 Tracing 定位问题（可疑）模块，根据模块具体的日志详情定位到错误根源，最后再基于这次问题调查经验调整 Metrics（增加或者调整报警阈值等）以便下次可以更早发现、预防此类问题。

2.3 OpenTelemetry

2019 年 5 月，OpenTracing 和 OpenCensus 共同发起了 OpenTelemetry 开源项目，旨在提供可观测性领域的标准化方案，解决观测数据的数据模型、采集、处理、导出等的标准化问题，管理观测类数据，如 trace、metrics、logs 等，其终态是作为 CNCF 技术委员会可观测性的终极解决方案。

OpenTelemetry 没有解决可观测性上的所有问题，但对数据标准、SDK、采集模型进行了规范，对于 Backend、Visual、Alert 等并不涉及，官方目前推荐的是用 Prometheus 做 Metrics 的 Backend、用 Jaeger 去做 Tracing 的 Backend，而对 Logging 还没有好的解决方案。

3、可观测性解决方案

前文介绍了可观测性的相关概念，下面将对本文提出的可观测性解决方案进行阐述，东方证券可观测平台是基于日志数据，指标数据等底层数据并联合链路追踪技术的可观测性解决方案的落地，是 Metrics、Tracing，Logging 三者融合技术的实现。

3.1 技术架构

东方证券可观测平台由数据采集 Agent、数据处理分析模块、数据展示模块组成，如图 3 所示。

采集 Agent 负责业务系统日志数据和调用链数据的实时采集，每笔请求产生的日志数据和调

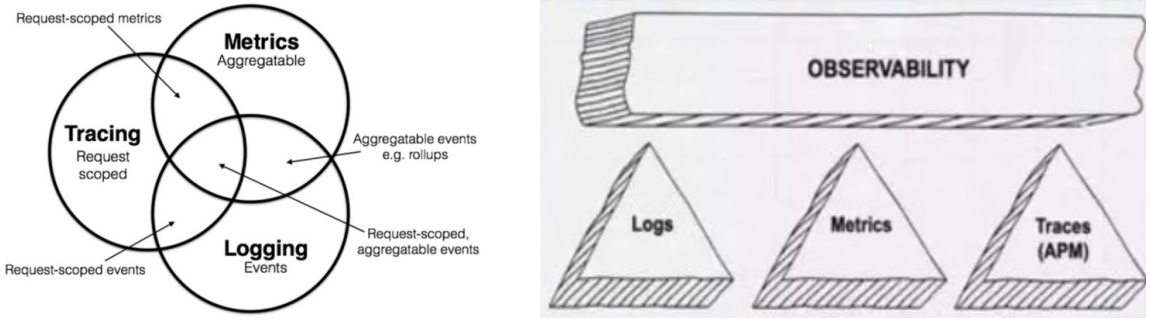


图 2：logging、tracing、metrics 关系

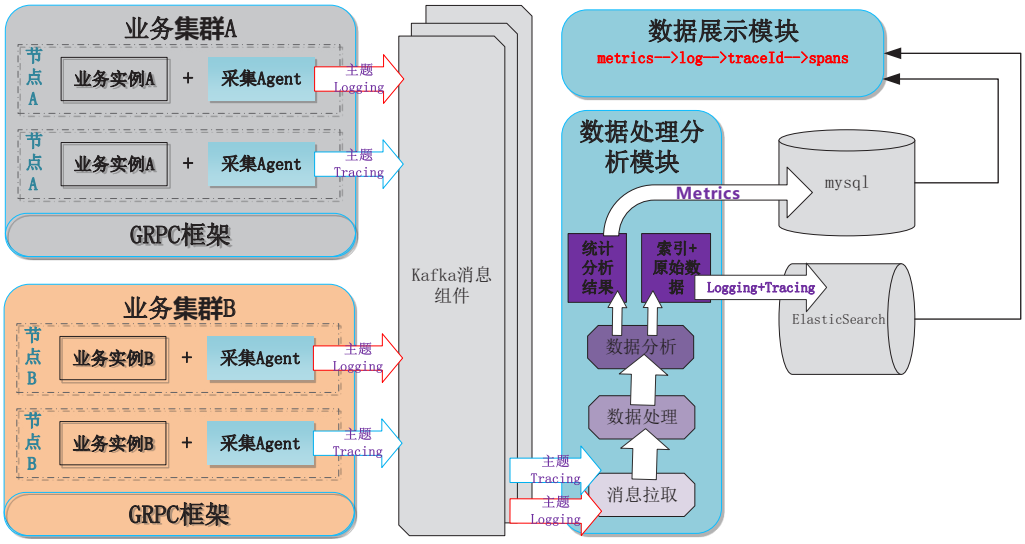


图 3：东方证券可观测平台架构

用链数据 traceID 相同，日志数据和调用链数据分别发送到 kafka 的 Logging 主题和 Tracing 主题。

数据处理分析模块负责从 kafka 消费 Logging 主题的日志数据和 Tracing 主题的调用链数据，并将日志、调用链原始数据以及索引持久化到 ElasticSearch。同时，对原始数据进行统计分析，并将汇总后的结果数据批量插入到 Mysql 数据库。

数据展示模块负责展示关联的系统日志、调用链数据和指标数据。

3.2 评估分析

我们从系统入侵性、数据采集一致性、可视化与关联、异常检测与诊断几个维度来评估可视测平台的效果，评估结果见表 1。

3.3 关键技术

3.3.1 traceID 生成和 MDC 技术

traceID 的生成采用 UUID 技术，保证了 traceID 的唯一性。

MDC（Mapped Diagnostic Context，映射调试上下文）是 log4j 和 logback 提供的一种方便在多线程条件下记录日志的功能，也可以说是一种轻量级的日志跟踪工具。

采集日志和调用链时，一笔请求接入，会创建全局唯一的标识符 traceID，并将 traceID 放入 MDC，后续同一个线程（包含子线程）输出日志和调用链都能从 MDC 获取相同的 traceID。使用 MDC 技术，保证了一次请求的所有调用的 traceID 唯一，相同的 traceID 能将一笔请求的一

表 1：可观测平台评估表

评估维度\实现方式	传统监控	OpenTelemetry	可观测平台
系统入侵性	自主埋点，侵入性高	有统一的 SDK，也有自动代码注入技术，侵入性低	统一的 SDK 并集成在 GRPC 框架中，存在代码侵入，但侵入性低
数据采集一致性	自主定义数据格式，无标准	定义了统一的 Metric、Tracing、Logging 标准，统一的元数据结构	参照 OpenTelemetry，统一的 log 日志格式、统一的调用链格式
可视化与关联	有 Metric 和 Logging 的单独监控，无 tracing，更无三者关联	无 Metrics、Logging、Tracing 关联	自主实现 Metrics、Logging、Tracing 关联，并通过自主实现+grafana 结合的方式进行可视化展示
异常检测与诊断	无	无	有

次调用过程和日志关联上。

3.3.2 日志格式与日志采集实现

日志记录采用统一格式：" 时间戳 " + " " + "[日志等级代码]:" + " " + " 日志消息体 "。

时间戳必须包含年月日时分秒毫秒，同一个系统必须使用统一的时间戳格式，如 "yyyy-MM-dd HH:mm:ss SSS"。

日志等级按照严重程度从高到低，代码依次 为：FATAL，ERROR，WARNING，INFO，DEBUG。

日志消息体为日志正文，以 json 格式存储。日志正文可以是采集的离散事件日志，也可以是采集的调用链数据。

当采集事件日志时，日志正文格式如表 2 所示。

当采集调用链时，日志正文格式参照表 3 的分布式调用 span 基本格式。

日志采集通过自定义 LogbackAppender 和 log4j2Appender，并结合 Filter 以及 Converter 来实现。

3.3.3 调用链模型与采集实现

用户从 app 发起一次业务请求后，可观测平台应该记录其所有调用的数据。以理财销售业务为例，用户从 app 发起基金申购请求，会经过 6 台服务节点：

1）接受用户请求的前端服务（前端服务节点）；

2）封装业务原子服务的中台（能力中心节点）；

3）提供基础服务的后台系统（后台服务节点）。

用户发起一次请求到前端服务 A，该请求依赖后端服务 B 与 C，因此服务 A 分别通过发送 GRPC 请求到服务 B 和服务 C，B 处理完 A 的请求后将响应返回给 A，但是服务 B 还依赖服务 D 和 E，B 再发起两个 GRPC 请求分别到 D 和 E，D 和 E 处理完毕后回到 B，B 才继续应答到 A，最终 A 将调用结果返回给用户。分布式调用链的目的，就是将用户一个入口请求及相应的其它后续请求进行网络拓扑绘制，最终生成表示调用关系的调用图。

分布式请求调用会触发多个系统的之间的请求和响应，而将两个服务之间的请求 / 响应过程叫做一次 span，一个多层次调用过程由 span01+ span01.01+ span01.02+…… 多个 span 组合而成，分布式调用链 span 的报文格式如表 3 所示。

通过分布式调用过程和 span 格式，我们理解了 span 什么时候生成以及怎样生成，接下来的问题是：如何将一笔请求的多个离散的 span 树形化串联起来。

我们知道 span 是通过 traceID 串联起来，并通过 spanID 与 pSpanID 的父子关系来树形化的，

表 2：事件日志基本格式

名称	类型	含义	说明
traceID	String	调用链 ID，必须全局唯一	必填。可以是 UUID。
spanID	String	span 的 ID	必填。格式为 01.01.**,每一个层级表示一层调用，同一个层级多次调用序号递增，如 01.01,01.02。
timestamp	Long	本地调用时间（单位毫秒）	必填。
localProjectName	String	项目名	必填。
localHostName	String	主机名	必填。
localIp	String	ip 地址	必填。
localPort	String	端口	必填。
localServiceName	String	服务名	必填。
threadName	String	线程名	必填。
level	String	日志级别	必填。
message	String	日志内容	必填。
stackInfo	String	堆栈信息	选填。当系统异常时，输出堆栈信息。

图 4：分布式调用过程

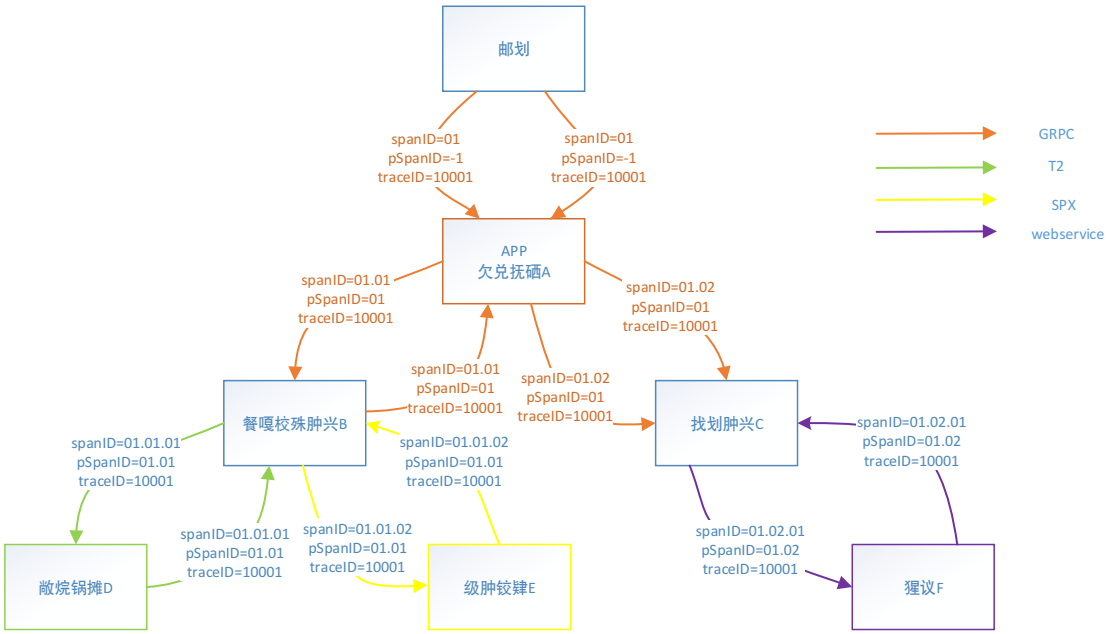


表 3：分布式调用 span 基本格式

名称	类型	含义	说明
traceID	String	调用链 ID，必须全局唯一	必填。可以是 UUID。
spanID	String	span 的 ID	必填。格式为 01.01.**,每一个层级表示一层调用，同一个层级多次调用序号递增，如 01.01,01.02。
pSpanID	String	span 的父 ID	必填。起点 span 填-1。
callID	String	一次调用的 ID	必填。一次调用的请求和应答 callID 必须相同，用来关联一次调用的请求和应答。
name	String	调用对端的服务及接口信息	必填。远端调用填接口名或者 URL,本地调用填函数名。
timestamp	Long	本地调用时间（单位毫秒）	必填。
type	String	调用方式	必填。只能是 CS\CR\SS\CR，表示客户端发送\接收，服务端返回\接收。
localProjectName	String	调用方项目名	必填。
localHostName	String	调用方主机名	必填。
localIp	String	调用方 ip 地址	必填。
localPort	String	调用方端口	必填。
localServiceName	String	调用方服务名	必填。
remoteProjectName	String	被调用方服务名	选填。
remoteHostName	String	被调用方主机名	选填。
remoteIp	String	被调用方 ip 地址	选填。
remotePort	String	被调用方端口	选填。
remoteServiceName	String	被调用方服务名	选填。
tags	Map	附加信息，由 key:value 组成	选填。可通过该字段传入请求入参和应答出参。

只要让 traceID、spanID、pSpanID 有序的传递就可以实现 span 的树形化串联。

单个服务内多次调用传递：如图 4 中所示，App 前端服务 A 先后调用财富销售中心 B 的 Span01.01 和账户中心 C 的 Span01.02，spanID 会递增，同时 MDC 技术保证了两次调用能获取到相同的 traceID。

跨 GRPC 服务调用传递：如图 4 中所示，App 前端服务 A 通过 GRPC 微服务接口调用财富销售中心 B 时，通过 HTTP Header 传递 traceID、spanID、pSpanID，财富销售中心 B 服务就可以拿到 traceID、spanID、pSpanID 并生成正确的 span 了。

最终，用户从 app 发起基金申购请求的 span

时间轴调用链树形图如 5：

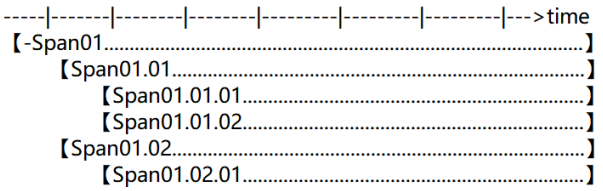


图 5：Span 时间轴调用链树形图

3.3.4 指标模型与采集实现

指标的采集类型分为系统指标和业务指标，采集周期分为当日指标和历史指标。具体指标模型应当从项目实际需求出发自定义，理财销售业务具体指标模型如表 4。

当日指标的采集通过 grafana 配置 metrics，从 ElasticSearch 中获取数据并展示；

历史指标的采集通过可观测平台定时的跑批，将跑批维度数据写入 mysql，并通过 grafana 配置 metrics 从 mysql 获取数据加以展示。

4、可观测性实现效果

4.1 分布式调用链可视化

东方证券可观测平台实现了分布式调用链的可视化，通过请求列表，查看具体一笔请求的调用链树形结构图，包括调用链中每个 span 的应用名、方法名、调用时间、耗时、状态码、状态等指标，并且可以看到 span 的请求入参和应答出参，具体操作步骤见图 6。通过可视化的链路跟踪，测试执行跟踪耗时减少 90%。

4.2 异常检测与诊断

传统的模式下，当监控检测到 error 关键字并告警，技术人员需要对多个服务节点的日志通过用户请求的关键字进行搜索匹配，分布式模式下，一笔请求涉及了多个应用的多个节点，运维人员需要遍历所有相关应用的每个服务节点并逐个排查，耗损大量人力和时间才能定位到异常应用节点的错误日志。可观测平台提供了异常检测和诊断功能（如下图 7 所示），当业务出现异常，可观测平台根据错误事件日志告警，找到具体的一条事件日志记录，基于日志记录的 traceID 精准定位到对应的调用链，并展示出链上异常 span，通过异常 span 的请求入参和应答出参，便可以诊断业务异常原因，大大提升了排查问题效

表 4：理财销售业务具体指标模型

指标模型	当日指标	系统指标	当日请求笔数（Counter）
			当日警告数（Counter）
			当日系统请求走势图（Gauge）
			当日系统请求接口分布图（Summary）
			当日接口平均耗时top10（Histogram）
			当日接口异常调用top10（Histogram）
		业务指标	当日活跃用户数（Counter）
			当日申购/认购/赎回金额（Counter）
			当日产品申购/认购/赎回金额分布图(Summary)
			当日产品申购/认购/赎回金额tops10(Histogram)
			当日用户申购/认购/赎回金额tops10(Histogram)
	历史指标	系统指标	日/周/月请求笔数（Counter）
			日/周/月警告数（Counter）
			日/周/月系统请求趋势图（Gauge）
			日/周/月系统请求接口分布图（Summary）
			日/周/月接口平均耗时top10（Histogram）
			日/周/月接口异常调用top10（Histogram）
		业务指标	日/周/月活跃用户数（Counter）
			日/周/月申购/认购/赎回金额（Counter）
			日/周/月产品申购/认购/赎回金额分布图(Summary)
			日/周/月产品申购/认购/赎回金额tops10(Histogram)
			日/周/月用户申购/认购/赎回金额tops10(Histogram)

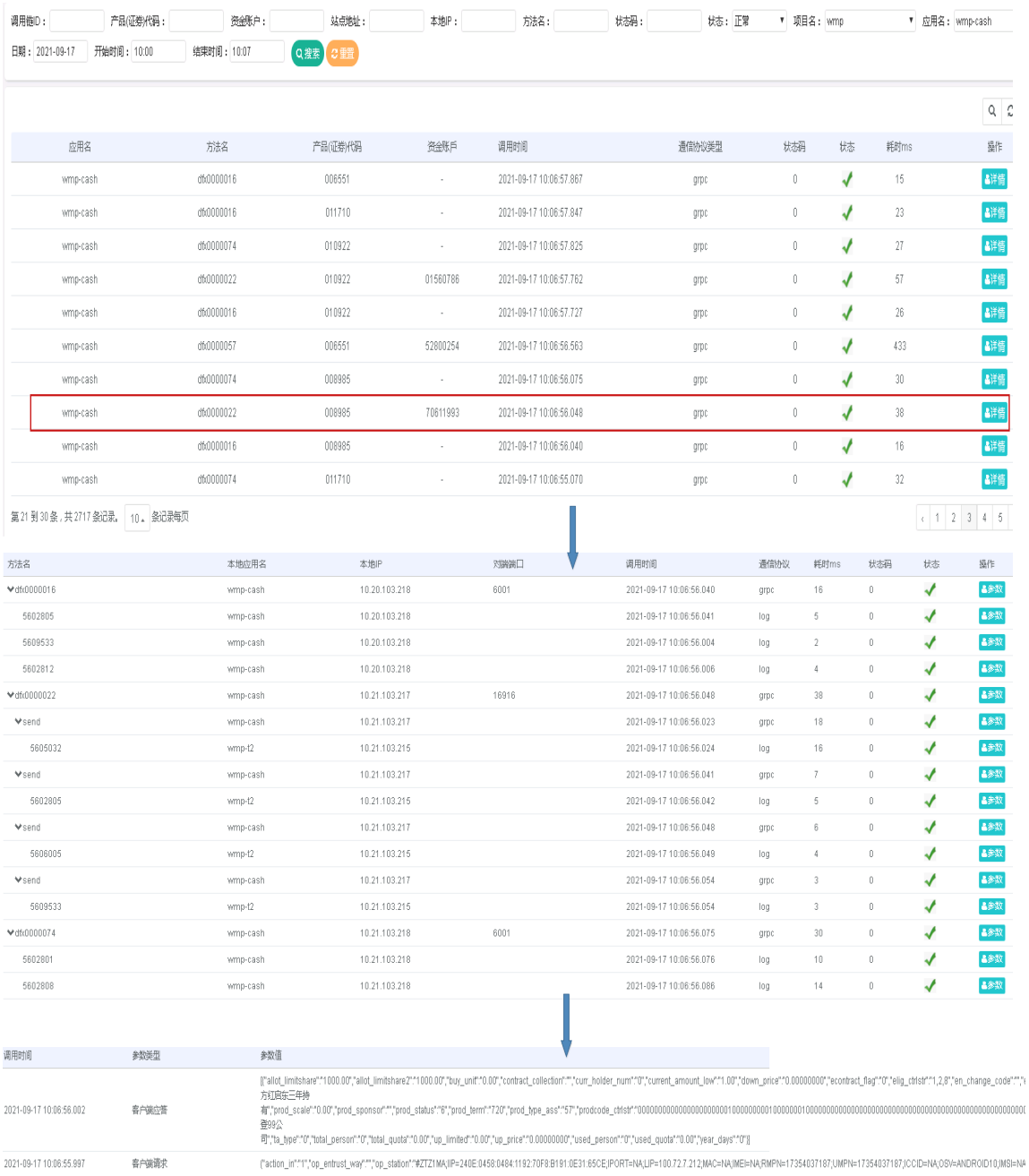


图 6：分布式链路可视化

率，异常检测诊断的时间总体减少 90%。

4.3 指标与链路关联

图 8 是指标和链路的关联图，通过当日服务维度的调用量指标，查看该服务当日每个接口的调用量指标，链接具体某个接口所有的请求列表，并跟踪具体一笔请求的调用链树形图。这种指标和链路关联的场景有很多，比如 top10 接口耗时指标，统计出调用耗时最多的 10 个接口，根据接口调用链，分析哪个 span 耗时最大，然后有针对性的进行性能优化。指标也可以和日志关联，比如 top10 异常调用接口指标，根据 ERROR 或者 WARN 事件日志最多的 10 个接口，梳理接口调用链的 span，进而分析并完善 span 的日志输出、

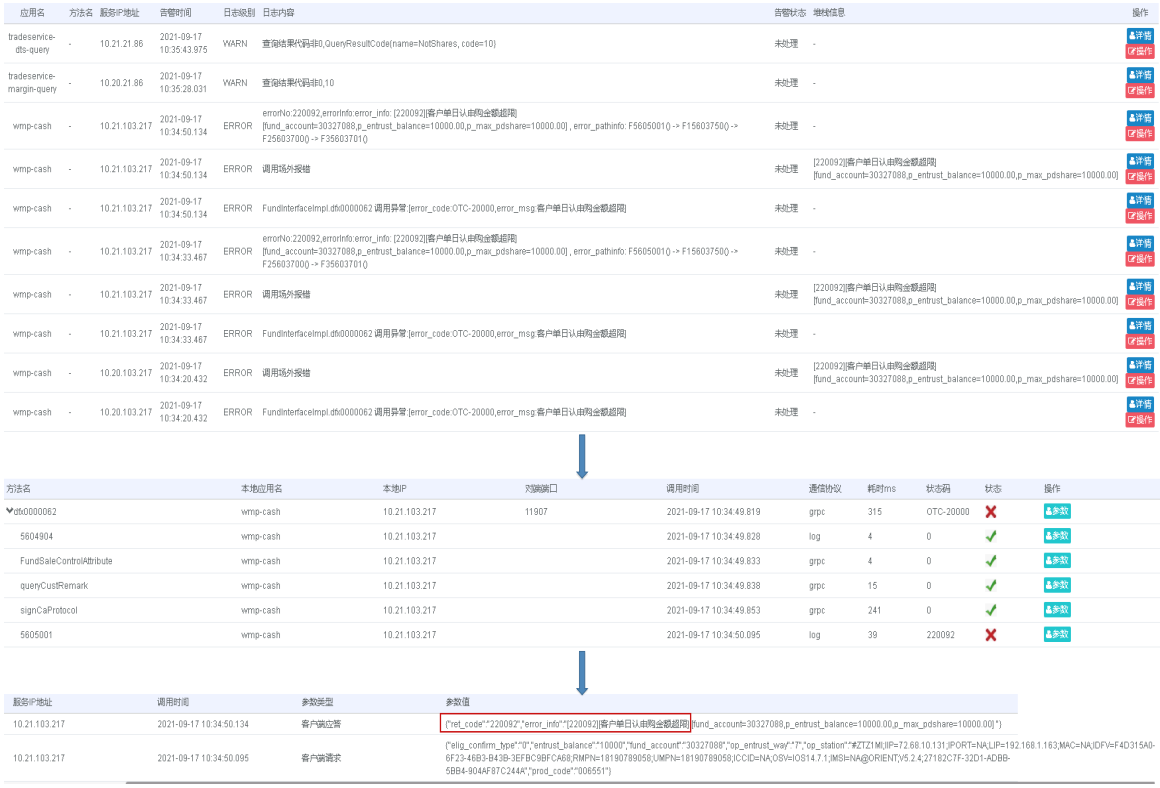


图 7：异常检测和诊断过程图

优化监控告警阈值、增强程序健壮性等。

4.4 指标可视化

图 9 是财富销售业务 grafana 实时指标图，包括系统指标和业务指标，很直观的展示财富销售业务的服务状态和业务情况。通过自定义配置 grafana 的 metrics，可以展示更多定制化的实时指标数据和历史指标数据。

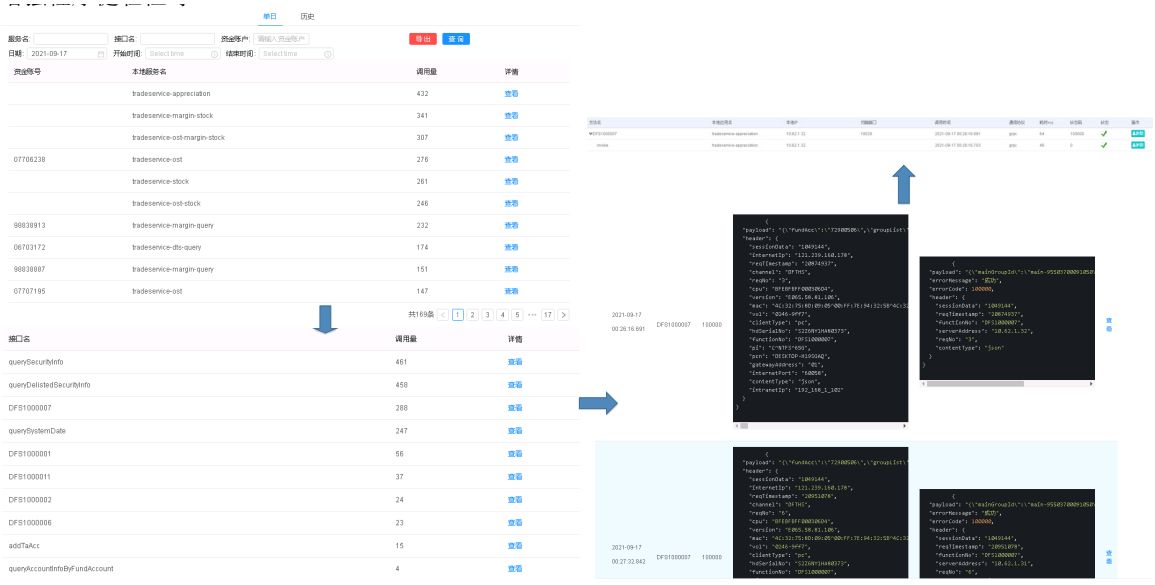


图 8：指标、链路关联图

图 9：财富销售业务指标图

5、总结

针对在分布式技术架构下进行可观测性的难点，本文提出了一种可观测性解决方案，通过引入分布式链路跟踪技术，实现 Metrics、Tracing、Logging 三者融合，很好的解决了系统和业务的可观测问题。该方案具有低入侵性，通过将 SDK 集成在 gRPC-Nebula 框架让接入方案简单，能快速接入业务系统。该方案实现了分布式系统的调用链跟踪，实现了日志与调用链深度串联，调用链、日志与指标数据的多方位融合，为开发人员进行系统调用拓扑优化、系统性能优化提供良好的分析基础，为测试执行人员推进测试案例执行大大的提高了效率，为运维人员快速精准诊断定位生产问题提供了便捷的方法，一定程度上，也可为业务人员定制个性化的业务报表。

该方案具有可推广性以及一定的参考价值，普遍适用于解决分布式架构下的可观测性问题。

算法服务平台的实践

郭亮 / 恒泰证券股份有限公司 金融科技研究院 北京
李江城、赵波 / 上海宇量智慧数据技术有限公司 上海
邮箱：guoliang@cnht.com.cn

恒泰证券近年来加速数字化转型，推进科技与业务深度融合，在深入研究算法金融理论与实践的基础上，结合自身信息技术能力自主研发了算法服务平台，有力地提升了公司竞争力并保证业务始终在合规道路上前进。

1、研发算法服务平台的背景

近期 A 股市场成交金额连续过万亿元，这种趋势有可能常态化。事实上，欧美等成熟证券市场的算法经济催生了新的经济模式，带来社会整体效益的提升，有其独特优势。“有一利必有一弊”，近年来算法滥用、算法作恶、算法道德、算法伦理等问题已引起广泛关注。2021 年 9 月 3 日，中国证监会科技监管局局长姚前在 2021 年服贸会金融服务专题展期间举行的“2021 中国国际金融科技论坛”上表示：“算法经济大幅改善市场经济的匹配效率和交易成本。人们一方面欢迎和享受智能算法带来的便利，另一方面却担心被智能算法替代，导致个人价值丧失。不仅如此，随着算法经济的快速发展，算法的渗透力和影响力越趋强大，其背后隐含的风险以及作恶的可能引起了关注。”。为此，加强算法监管，以监管科技应对新型科技，既是顺应之策，又是必然之举。

各国监管部门高度关注算法经济的同时，承担市场主体责任的各经营机构，特别是证券公司行动起来，在信息技术、合规展业、市场风险、内部控制、监管报送、行业自律等各条线协调管理以避免智能算法顺周期性风险、羊群效应等弊端，各经营机构根据证监会及相关机构的监管要求、业务发展需要落实算法推荐服务提供者的算法安全主体责任，建立健全相应管理制度，制定并公开算法推荐相关服务规则，配备与算法推荐服务规模相适应的专业人员和技术支撑。

恒泰证券作为市场服务主体之一，正致力于数字化、智能化转型，期望建立体系化的工具向

合格投资者提供适当的算法交易服务。恒泰证券投入研发力量实现了一套拥有自主知识产权的、自主可控的、安全完整的、可用于推广的算法服务平台，确保算法交易的合规有序发展，提高市场的流动性以及定价效率。

2、算法服务平台应具备的主要功能

算法服务平台由券商进行维护和运营，不同算法提供商和模型提供商提供的算法和模型以可插拔模块的形式嵌入到平台中，平台自动提供设置、监控、查看页面，适配客户柜台，客户自主选择使用。算法服务平台旨在解决证券客户的算法使用需求、算法供应商的算法接入需求、证券公司的算法运营、算法管控需求，监管部门的监管审查需求等各关联方的多维度、多种类的使用管理需求。

对于证券客户：平台将不同算法提供商的多种算法整合在一起，提供体验一致的用户界面和操作方式，提供算法签约、查看、执行、监控、统计等功能的一站式使用平台。免除客户使用一种新算法就需要学习一套新的客户端的操作成本和学习成本。同时证券客户使用券商提供的标准的、符合监管要求的客户端，也有助于保护投资者合法权益，促进资本市场健康发展。同时客户的个人信息、资产信息等私人信息仅保存于券商柜台系统，对算法供应商不可见，尽最大可能地保护了交易客户信息的安全性。

对于算法供应商：平台提供了友好、高效的接入框架。提供了包括交易接口、行情数据、金融基础数据等多方面的基础设施，为不同种类算法策略提供风格统一的，通过配置就可以定制的用户界面。算法供应商的开发精力可以聚焦于算法逻辑的实现中，省去了开发用户交互界面的投入。同时算法供应商对接一套 SDK，就可以适配不同种类柜台的交易和回报接口，接受实时行情数据以及查询多种类金融数据。

对于证券公司：平台提供了相对完备的算法运营和管控流程。一个算法要上架供客户使用，在算法服务平台中要经过算法初审、算法模拟盘运行、算法风险等级匹配、算法核准、客户开通等几个步骤，每一步骤都需要留痕和上传相应的附件资料。客户所有的算法和策略交易都通过算法平台进行，受平台管控。算法

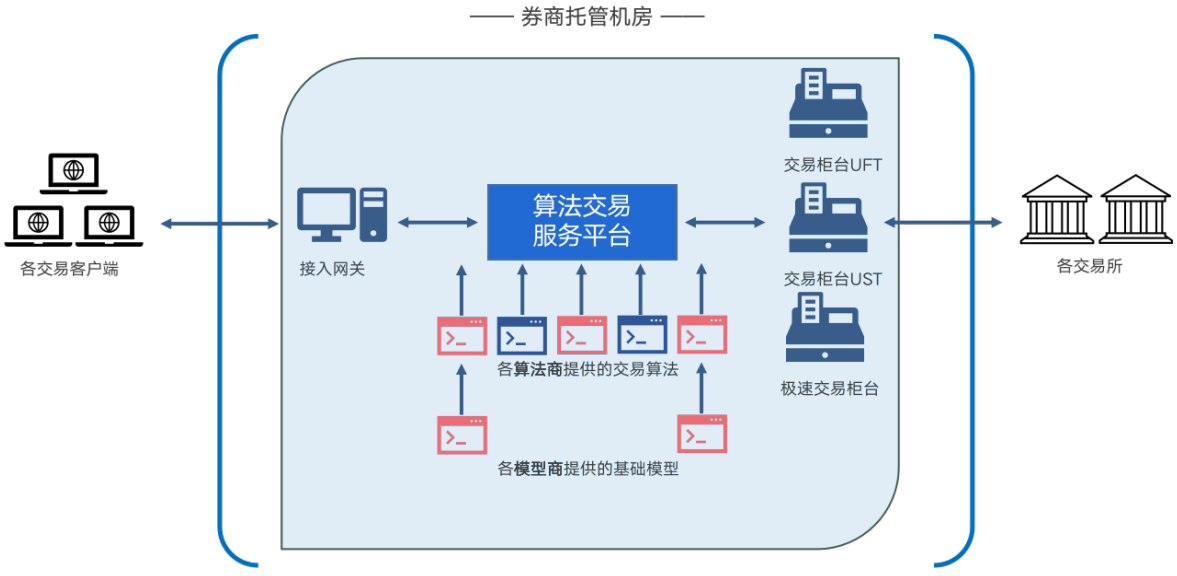


图 1：算法交易平台服务架构

图 2：算法服务平台管理端界面

服务平台的引入，从一定程度上解决了黑接口等不合规的接入方式。在已上架算法的运行过程中，平台的异常交易风控模块会对算法所有的委托进行实时同步的风控检查，如果判定为异常交易行为，将拦截此笔报单，同时平台管理端也可以进一步选择停止策略运行直至暂停相关客户在平台的所有执行。证券公司投入上也节省了每引入一套算法服务就搭建一套完整系统的投入成本和运维成本。

对于监管部门：平台保留有策略上线审核流程、客户操作使用流水、客户算法委托信息等所有相关记录，可供监管部门追溯和查询。随着基于人工智能（AI）的算法策略使用的越来普遍，算法审核有时只能通过算法的 AI 模型和数据采样特征对算法加以评估，难以描述算法具体的交易逻辑。算法服务平台提供的异常交易风控和交易回溯功能，保障了异常委托不提交和事后分析的能力。同时客户开通算法服务的控制权在客户和证券公司，算法执行的操作标的由客户自主设定，避免了算法服务供应商恶意交易的可能性。

3、算法服务平台的典型系统架构

算法服务平台在技术架构上，从下到上分别为硬件与操作系统层、数据服务层、总线服务层和终端接入层四个层次，具体如下图所示。

图 3：算法服务平台架构分层

系统从功能上分为客户终端模块、平台管理模块、算法接入模块、算法评价模块、行情接入模块、风险控制模块和平台总线等七大模块。

其中，客户终端模块负责处理最终客户的交互，包括提供统一的设定界面，对不同算法的进行配置和查看。启动算法，监控算法的执行状态和展示算法的绩效和统计信息。同时客户端模块适配了桌面设备和移动设备的不同显示模式。

算法服务管理模块提供完备的管理能力，提供了用户管理、用户签约管理、用户风险管理、用户交易统计、算法评估、算法审核、算法信息设定、算法上下架、风控规则设定和实时盘中风控管理等功能。

算法接入模块提供了算法提供商和算法服务平台通讯的接口。接口采用远程过程调用的方式，提供 Windows、Linux 操作系统下，C（C++）、

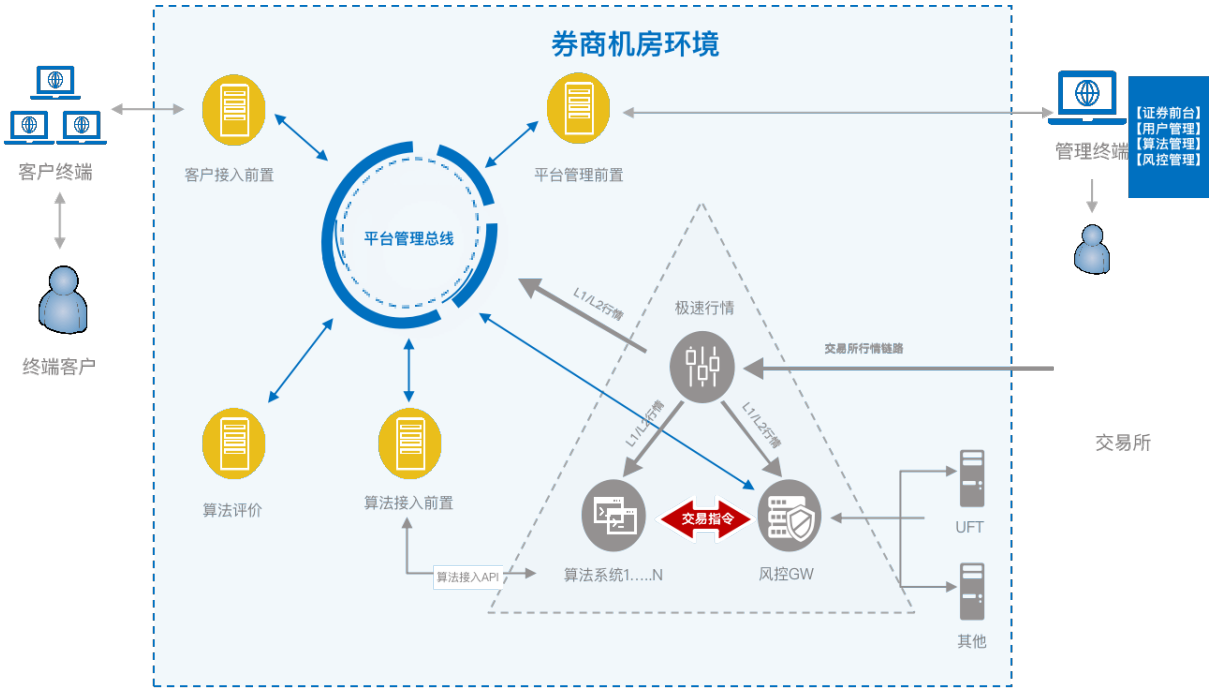


图 4：算法交易平台模块逻辑关系

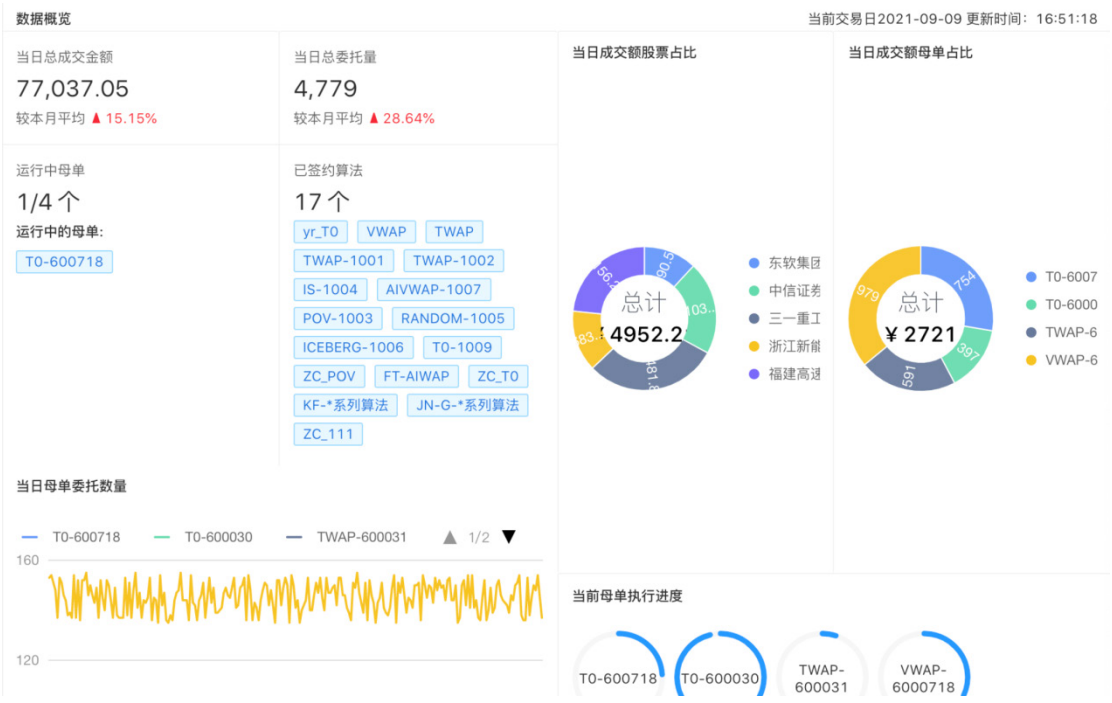


图 5：客户端监控界面

Java、Python、R 等多种语言的 SDK，方便不同语言实现的算法进行接入。算法服务平台屏蔽了不同种类柜台的差异，为算法提供商提供了一致的接口，实现了多种柜台的自动适配和自动路由。算法提供商报单和回报以母单为单位，不需要考虑具体交易账户的复杂处理逻辑。同时算法接入模块还对平台内的行情总线和数据总线进行了封装，提供较为友好的 API。供接入的算法接收行

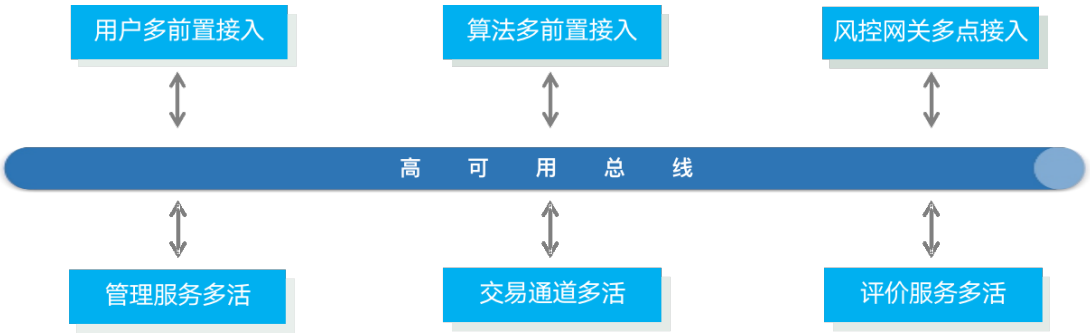


图 6：算法服务平台总线

情数据、交易数据，并可以提供财务数据、市场数据、历史行情等数据的查询。

算法评价模块提供算法回测、算法评估功能。并在回测结束后提供多周期的算法分析结果数据。供算法管理岗和风控岗对算法是否达到上线标准，以及锚定算法的风险等级作为依据。算法评价模块还对每日系统真实运行的算法策略进行统计并生成报表。

行情接入模块是算法服务平台唯一的行情接入口，方便平台实际部署时适配不同的行情源接入。行情接入模块对接入的原始行情进行正确性校验、字段补全等规则化操作，为平台中其它组件屏蔽行情源差异，提供统一的行情数据格式。

风险控制模块拥有插件式可扩展的异常交易风控能力，对算法委托实行阻断式实时风控，拦截有异常交易特征的委托执行，确保系统中的交易合法合规、可追溯、可审计。阻断式风控可以选择拦截当前委托、拦截当前算法或者拦截当前账户下的所有交易。同时风险控制模块对算法接入屏蔽真实交易账号，对算法接入模块报送的母单委托做相应的交易账号的转换，并路由至正确的交易柜台，从根本上保证了客户信息不泄漏。目前平台已经接入了恒生 PB、恒生 UFT 和恒生 UST 三种柜台。

平台总线是整个平台的核心，从功能上分为行情总线、消息总线和数据总线。分别负责系统内实时行情数据的分发，交易类、管理类事件和通知的送达，财务、市场等金融数据的查询。总线底层由 Nats 和 RocketMQ 两种通用消息总线以及多个自研微服务构成，数据源包括交易所实时行情、平台内的交易管理数据、第三方金融数据库、HDFS（Hadoop 分布式文件系统）和 OS 文件系统。平台上层对数据使用进行了良好的封装，抽象出订阅 - 推送接口和主动查询接口两种类型的服务方法对数据进行访问。使用统一的调用方式操作不同底层的数据，比较好地适应了不同数据使用场景的需要。

4、算法服务平台的业务模式

算法服务平台为券商经纪业务引入了新的算法服务模式，即对客户来说的类似超市购买的服务模式，对券商来说的类似超市采购的供给模式。

客户可以浏览系统中提供的不同算法供应商不同种类、不同特色的算法策略，并可以查看策略说明，进行历史回测、模拟盘验证，最终决定是否在实盘上进行使用，并对使用结果进行查看。

券商负责发掘好的算法提供商，为客户提供尽可能丰富的算法品种，保证平台中上架算法策略的质量，对上架的算法进行管理，在算法使用时进行跟踪和监控，生成算法执行的相关统计报表，协调算法提供商对算法进行持续优化。

为了满足以上目标，算法服务平台从功能设计和功能实现上做了以下几个方面的保障：

流程保障。包括算法上下架流程和客户权限开通流程。算法上下架需进行完整的申请、评

估、模拟盘跟踪和复核操作。确保每一步的留痕和相关报告的提交。客户权限开通由客户通过网厅发起申请，经营业部、业务管理部门根据客户类型、资产量、风险等级、交易年限等要素对客户进行评估，并决定是否在平台中发起后续的审核开通。

交易保障。主要包括异常交易风控和交易抗冲击性两个方面。异常交易风控上，系统采用同步实时多线程多节点的实现方式，既保证了委托的低延时又能在第一时间对有异常交易特征的交易进行拦截。对于交易抗冲击性，系统采用以消息总线构成的架构，天然的对网络抖动、算法异常等造成的瞬时消息突增有良好的削峰特性，从系统上对平台交易的安全运行做了又一重保障。

系统保障。主要包括系统稳定性，系统安全性两方面。系统稳定性上，系统所有组件都采用多活方式运行，并且有较好的系统日志记录，可以配合基础日志收集监控系统，对平台已发生的或潜在的问题进行报警，以做到实时响应。系统安全性，包括系统通讯的安全性和客户信息的安全性。系统通讯安全性使用安全的通讯信道作为保障。客户信息仅在系统风控模块进行指标运算和报单时使用，不对算法模块等三方模块开放，做到了信息使用的最小化。

本算法服务平台以恒泰证券投入研发力量进行开发，以证券公司的视角审视了证券客户、算法提供商、证券公司和监管机构的不同关切，引入了算法服务的新模式。较好地满足了现阶段算法交易的业务应用。

5、实践总结与展望

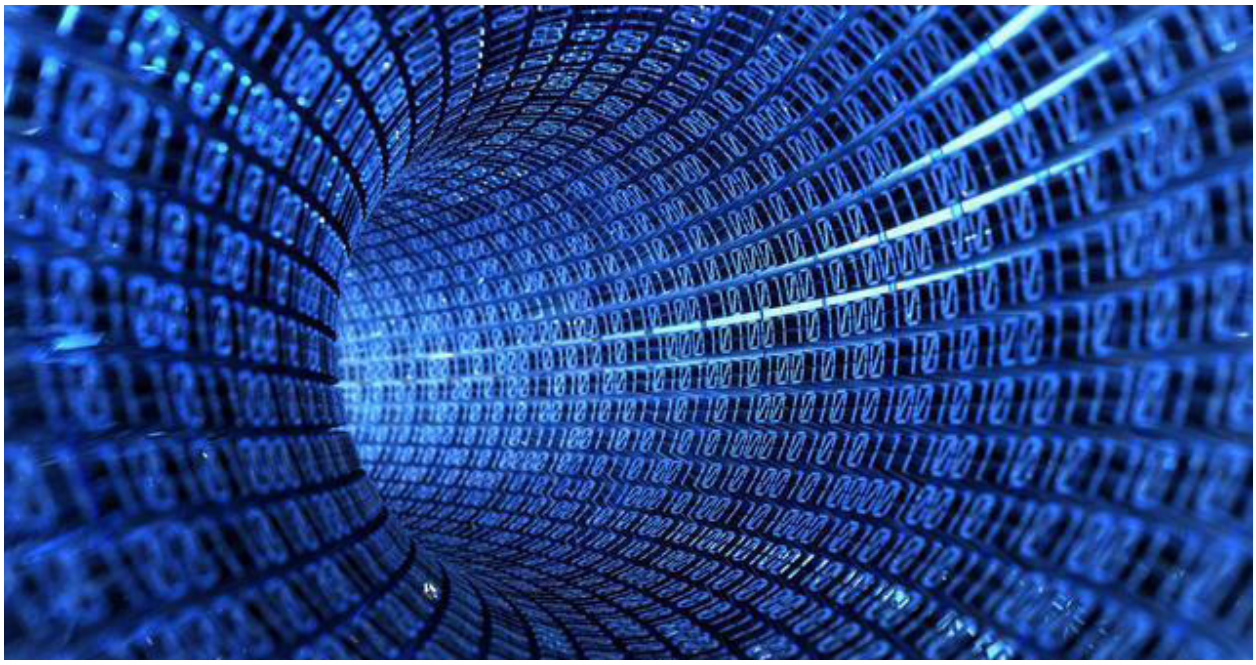
恒泰证券的算法服务平台上线一年有余，目前平台日均交易额在 1 亿元左右，交易峰值在每秒 200 笔以上，接入算法供应商 2 家，接入客户 1000 余人，系统运行稳定，客户满意度较高。由于算法服务平台基于自研开发，与传统的由算法提供商提供整套系统的模式不同，系统的安全性更容易把控。同时作为券商自研系统，我们把异常交易风控放在最重要的优先级上。尽最大可能避免不合规交易的产生，同时风险控制模块组件化可插拔的开发部署机制，快速响应风控要求，也减轻了风控和合规部门的工作压力。

当前，金融网络与信息安全面临日益严峻的安全形势和不断深化的金融体系改革的挑战。“十四五”规划指出“提升金融科技水平”、“完善现代金融监管体系，维护金融安全，守住不发生系统性风险底线”，为新时期金融网络安全和信息化发展提出了新要求，指明了发展方向。在金融行业数字化转型中实现自主可控，既是贯彻落实国家战略要求的必要选择，也是新阶段下科技支持金融发展的必然趋势。恒泰证券作为中小券商也在积极寻求符合自身特点的信息技术应用创新之路，近期正融入上海证券交易所为证券基金行业搭建的信创联盟与生态圈，向上海证券交易所及头部机构学习，对算法服务平台进行信创改造，希望构建一个以信创软硬件为依托的，自主可控、开放可扩展的算法服务平台，预计年底可以交付一定的信创成果满足业务发展要求。

信息技术创新观察

全历史数据服务系统在信创大数据平台上的实践

肖钢、李剑戈、王岐、高森 / 中信建投证券股份有限公司 信息技术部 北京
邮箱：wangqixx@csc.com.cn



在证券行业数字化转型的大背景下，利用海量历史数据提升客户服务价值已经成为头部券商竞相争夺的技术高地。随着中国证券交易客户规模的不断增长，交易系统数据成级数增加，传统解决方案中的数据不全、数据标准不统一、系统性能无法保障等问题成为了历史数据服务能力的瓶颈。本文从介绍历史数据的重要性入手，首先对证券行业传统历史数据使用现状进行了分析，进而提出一套基于全国产化技术的大数据平台解决方案。从数据治理、系统架构、国产化硬件选型、国产化软件选型、全国产化系统的应用效果几个方面介绍了某全历史数据服务系统的实现，并提出了对该系统的后续规划和展望。

1、引言

大数据是推动金融行业发展和证券业进步的重要战略引擎，是推进券商治理体系和治理能力现代化的重要战略资源，也是提升行业治理能力和水平的重要创新工具。大数据驱动券商行业治理创新不仅大大节约了券商治理的时间、资源和人力成本，而且建构了券商行业治理的新思路和新模式，实现了从封闭式管理走向开放式治理、从静态化管理走向流动性治理、从精细化管理走向精准化治理、从网格化管理走向网络化治理、从单向度管理走向协同化治理的路径转向。

图 1：大数据重要作用

证券行业大部分数据来自交易系统，其中有 99% 以上为历史数据。根据 iiMedia Research 数据显示，中国证券类 APP 用户规模稳定增长，从 2015 年到 2020 年，每年增长率都超过 15%，其中 2016 年和 2017 年甚至超过了 30%。到 2020 年，中国证券 APP 装机数量已经达到惊人的 1.29 亿。

另一方面，根据中国人民银行数据显示，2015-2019 年我国股票市场的成交量以及成交额均呈波动变化态势。其中 2019 年我国股票市场成交量达到 126624.29 亿股，成交金额为 1274159 亿元；由于受到 2020 年全球疫情的影响以及美国股票市场熔断事件的影响，我国股票市场也有所动荡，2020 年 1-5 月，我国股票市场的成交量为 65560.33 亿股，成交金额为 744340 亿元。在证券行业数字化转型的大背景下，利用海量历史数据提升客户服务价值已经成为头部券商竞相争夺的技术高地。而随着中国证券交易客户规模的不断增长，交易系统数据成级数增加，传统解决方案中的数据不全、数据标准不统一、系统性能无法保障等问题成为了历史数据服务能力的瓶颈。面对这些传统解决方案提出的挑战，公司提出了一套用信创大数据技术实现全历史数据服务的解决方案。

2、基于信创大数据技术的全历史数据服务系统实现

当前国际形势风云变幻，国家深化改革进入新阶段。关键技术是创新发展的国之重器，自主可控计算机发展的必要性、重要性和紧迫性不言而喻，自主可控事业仍是任重而道远。信息安全、自主可控已上升为国家战略，在国家政策引导和有关部门的强力推动下，我国近年来在自主可控计算机基础软硬件研发、应用及生态链建设等方面已初见成效。作为大型国有头部券商，公司领导在构建全历史数据服务系统过程中，充分考虑

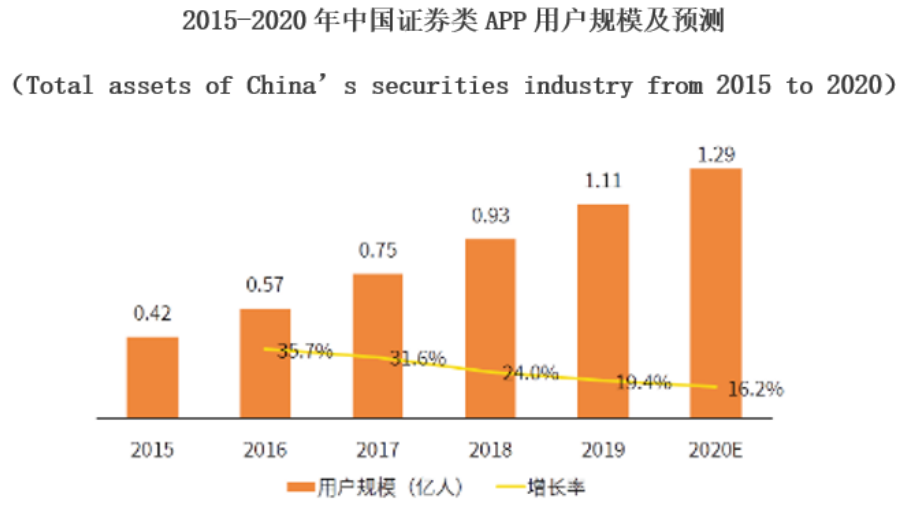


图 2：中国证券类 APP 装机规模

到国产化需求，要求从硬件到软件的各技术选型完全国产化。

2.1. 信创基础软硬件技术选型依据

国产服务器主要指标在 CPU，从 CPU 的稳定性、性能、适配性等方面，我们对基于 ARM 体系架构的鲲鹏、飞腾芯片和基于 X86 体系架构的海光芯片进行了适配性测试。

在硬件方面，我们选择基于 ARM 架构的鲲鹏处理器系列服务器作为大数据平台的基础环境，这样能有效利用 CPU 多核和并行计算的优势；选择基于 X86 架构的海光处理器系列服务器作为数据库和中间件应用的基础环境。

操作系统方面，我们测试了麒麟、统信以及欧拉系统，从各系统的应用特点，最后选择麒麟 V10 系统。

2.2. 信创大数据技术的选择

近年来，大数据和云计算在金融行业的发展如火如荼，在区块链、高性能计算、人工智能、金融工程等前沿技术领域也在不断的探索。HADOOP 生态经过多年积累，在分布式存储和分布式计算方面已经非常成熟，在互联网行业已经有 PB 级数据存储和处理场景落地。因此全历史数据系统着重实现从传统交易架构系统到大数据架构的转型，实现多数据源、多类型数据采集、加工、处理最终建设客户交易全历史数据仓库，为后续公司运营以及客户服务提供便捷的数据支持。

2.2.1. 数据仓库解决方案

HADOOP 是一种开源的分布式文件存储解决方案，国内的分布式存储（HDFS）和分布式计算（MR）具有高可靠性、高扩展性、高容错性和高效性等特点。高可靠性体现在 HDFS 会维护多个副本数据，因此对于大于一个或者几个存储单元出现故障也不会导致数据丢失；高扩展性体现在 HADOOP 天然具备横向扩展能力，可以很方便的扩展数以千计的节点；高容错性体现在 HADOOP 可以自动将失败的任务重新分配或者丢失节点上的数据重新均衡；高效性主要是指 HADOOP 在 MapReduce 的思想下，计算是在集群各节点上并行工作的特点，提升吞吐量和批量计算的效率。

HIVE 是基于 HADOOP 构建的一套分布式数据仓库系统，它将 HADOOP 分布式文件系统（HDFS）中的数据映射成一张数据库表，并提供完整的 SQL 功能。HIVE 还可以外链 HBASE 和 ES 生成 HIVE 外部表，可以通过 HIVE SQL 对 HBASE 和 ES 中的数据进行操作。对于全历史项目将五大交易系统的数据从传统关系型数据库抽取到 HDFS，使用 HIVE SQL 实现数据的清洗转换，结合自主研发的调度工具实现无人工干预或者少

表 1：ARM 和 X86 体系架构的比较

CPU 体系架构	特点	备选芯片
ARM	➢ 低功耗、低费用、小体积、高性能 ➢ 定位准确、早早聚焦移动端市场 ➢ 授权模式早，配套 IP 完善 ➢ 精简指令集	➢ 鲲鹏 ➢ 飞腾
X86	➢ 应用程序兼容性高 ➢ 高性能，市场占有率高，产业规模大 ➢ 复杂指令集	海光

量人工干预的自动化客户全历史数据仓库搭建。

2.2.2. 客户服务解决方案

在客户全历史数据仓库的基础上选择对高并发、高效查询的支持比较好的额组件为客户提供查询服务，比如 REDIS、ES（ELASTICSEARCH）、HBASE 等。由于全历史数据量大，REDIS 这种基于内存的 KV 数据库被舍弃，HBASE 和 ES 在数据量和查询效率方面都有不错的表现。HBASE 是基于 KV 的列式数据库，它专注于 ROWKEY 范围查询，各类业务设计都要围绕 ROWKEY 开展。HBASE 使用中业务和 ROWKEY 具有较高的耦合性，但是对于账单类、流水类业务有较好的支持，因为这类查询本质上是一种简单的 ROWKEY 范围查询。对于复杂的多列查询 HABSE 存在明显不足，为了保证查询效率，我们选择了 ES。它是基于 Lucene 倒排索引的搜索和分析引擎，存入 ES 中的数据默认会为每个字段创建索引，可以轻松实现高性能复杂聚合查询。ES 支持全文检索，对于中文也有很好的支持，像按照股票名称这种模糊匹配，ES 都可以胜任。因此 ES 可以用在客户全历史数据服务查询，比如成交、委托或者持仓明细等查询服务中。基于以上分析，全历史客户服务采用 HBASE+ES 的解决方案，ES 提供数据的多维度搜索查询服务，HBASE 提供账单类相对固定的数据查询服务。

2.2.3. 信创解决方案

针对开源的 HADOOP 生态系统的信创解决方案，中信建投选择腾讯大数据处理套件（Tencent Big Data Suite，TBDS），其内部封装了 HDFS、HIVE、HBASE 等组件。TBDS 大数据套件在中信建投采用基于 ARM 架构华为泰山 200 服务器的私有化部署方式，为公司内部信创系统提供分布式计算和存储服务。对于 ES 的信创解决方案，由于目前国内尚未有类似于 ES 的成熟商业产品，而 ES 本身又是开源软件，满足信创要求因此被直接使用。在中信建投 ES 同样部署在基于 ARM 的华为泰山 200 服务器中，为公司内部信创系统提供搜索引擎服务。

2.3. 信创数据库和中间件的选择

国产数据库技术近年来蓬勃发展，数据库产品百花齐放。根据全历史数据服务系统的应用场景，我们选择了如下几个 OLTP 数据库进行对比测试。

表 2：国产数据库比较

品牌	架构	特点
达梦	集中式	➢ 兼容 Oracle 语法 ➢ 国内第一批数据库厂商，稳定可靠，有多年技术积累 ➢ 广泛应用于党政机关、军工项目
GaussDB	集中式	➢ 兼容 Oracle 语法 ➢ 开源数据库，生态逐渐完善
TDSQL	分布式	➢ 兼容 MySQL 语法 ➢ 类似集中式加主从同步模式 ➢ 拥有互联网高并发场景的实践
OceanBase	分布式	➢ 兼容 MySQL 语法 ➢ 拥有互联网高并发场景的实践
TiDB	分布式	➢ 兼容 MySQL 语法

考虑到兼容 MySQL 语法以及未来上云及可扩展等方面的需求，我们选择了腾讯 TDSQL for MySQL 数据库。

在中间件方面，全历史数据服务系统的综合管理模块、数据加工引擎和数据服务引擎为 JAVA 语言实现，采用 OpenJDK（GPL 许可的 Java 平台的开源化实现）编译，并且运行在国产中间件上。东方通和宝兰德作为两大国产中间件厂商，都能很好的兼容 Tomcat 上的 Java 应用，在实现 Web 接口类的后台调用功能方面表现不相伯仲，只是在一些实现细节上存在少许差异。目前系统选择了宝兰德中间件。

2.4. 全历史信创技术架构

全历史整体架构包含交易数据源、自研 ETL 工具、腾讯大数据平台、开源组件和接口服务五部分组成，除交易数据源外其余均部署在信创服务上，且满足信创的标准和要求。架构如图 3 所示。

图中 ETL 服务为基于 OPENJDK 的自研工具，提供任务调度和任务监控等服务；腾讯大数据套件，提供基础存储和计算能力；开源组件主要是 ES 和 HBASE，为数据查询服务提供支持；接口服务，通过宝兰德中间件对接公司服务中台，为 APP 提供服务。

2.5. 全历史数据服务系统实现

为了保证投资者做交易的时效性，交易系统通过分离当日和历史数据来降低每笔交易的数据计算量。即每天将委托流水、成交流水，登录日志等数据归档到历史数据库。传统的历史数据库存放到关系型数据库中，通常会保留一到两年的数据，为投资者提供历史交易查询服务。

随着投资者专业能力的提升，尤其是机构投资者比例的不断增加，客户对历史数据查询提出新的需求，如希望查看近十年的交易行为、查看某只股票自持仓以来的盈亏情况、查看历史上某个时间点的资产情况等，在传统的系统架构下实现这些需求存在着明显的不足。利用大数据技术，我们设计了一套全历史数据服务系统，该系统可以较好的解决这些问题。

全历史数据服务系统由交易数据源、系统综合管理模块、数据存储引擎、数据加工引擎和数据服务引擎五个部分组成，每个部分通过接口调用实现数据交换，如图 4 所示。

2.5.1. 交易数据源

交易数据源指 AB 股、两融、股票期权、场外交易、贵金属等交易系统和账户系统等，全历

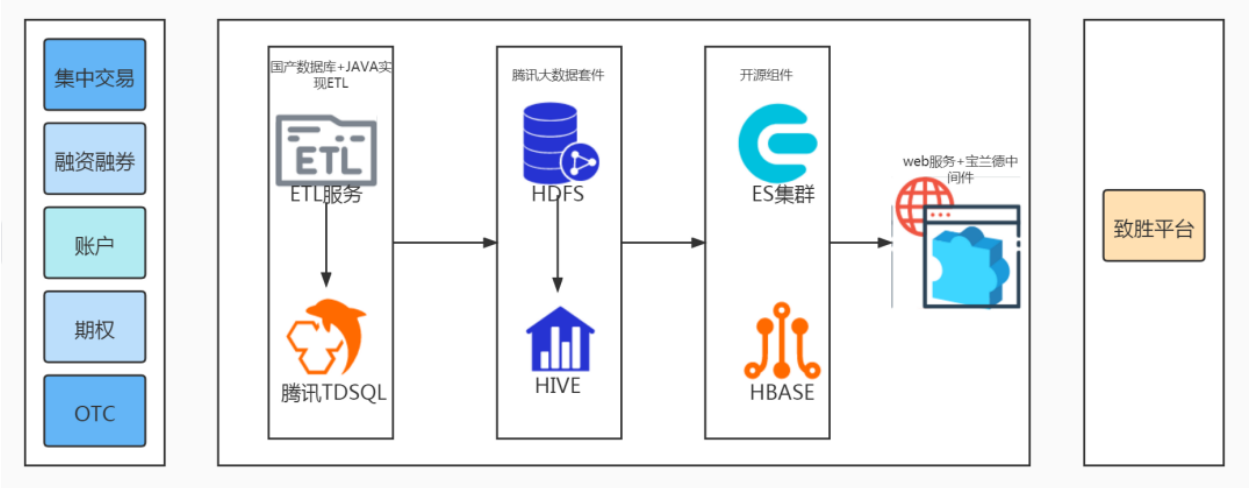


图 3：全历史数据服务系统架构（制胜平台为公司的服务中台）

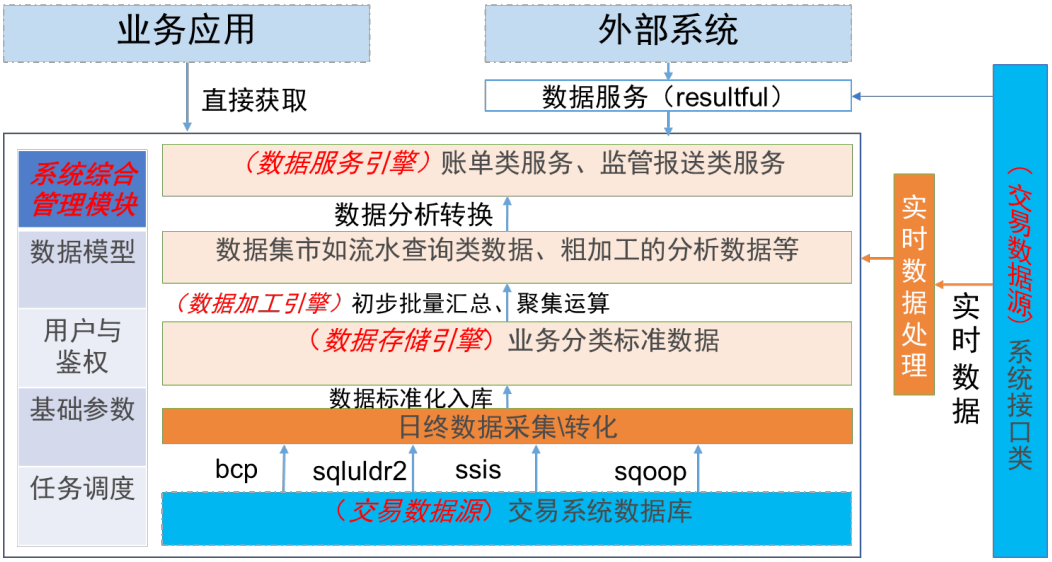


图 4：全历史数据服务系统实现

史数据服务系统每日从交易数据源获取数据。交易数据源通常为传统数据库，数据获取通过 ETL 作业完成。为了提升 ETL 效率，可以利用 BCP、SQLULDR2、SSIS、SQOOP 等工具完成。由于大数据平台的数据导入都是数据块级的操作，比传统关系型数据库的插入操作效率提升 50% 以上。而利用大数据系统导入数据可以覆盖之前导入的数据的特性，遇到由于日终清算问题导致的重新清算的情况时，重新导入数据的时间会大大缩短，从而将为客户提供数据服务的时间点提前。

2.5.2. 系统综合管理模块

全历史数据服务系统一个重要的组成部分是系统综合管理模块，它保存了系统的所有元数据，包括 ETL 数据模型、用户与鉴权数据、系统基础配置参数、任务调度数据等。通过维护和管理这些元数据，可以确保系统运行的可靠性。

2.5.3. 数据存储引擎

数据存储引擎主要是指 HIVE 分布式数据仓库系统、ES 存储系统、HBASE 数据库等。首先通过业务数据分析、数据类型整理、数据汇总等方法，把各种业务类型的数据标准化并在 HIVE 系统中创建相应的表格。这些表格从逻辑上又分为 ODS（Operational Data Store）层和 DW（Data Warehouse）层。ODS 表格中存放当日或近期数据，DW 层存放全历史数据。数据装载过程是从交易数据源中抽取的数据先导入到 HIVE 系统的 ODS 表格中，每日清算成功完成后，做为增量数据复制到 DW 表格中。由于 HIVE 系统的分布式存储和横向扩展特性，可以在不降低性能的情况下存放海量数据。目前公司交易系统 10 年的历史数据上百 TB，使用 HIVE 作为存储引擎可以支撑未来几十年的数据增长。

存放到 ODS 中的数据再根据业务需求，通过逻辑运算，将数据加工并增量加载到 ES 和 HBASE 中供用户查询调用，由于只计算当日的业务数据，整个过程可以减少运算压力，缩短数据提供服务的时间。另外，作为 DM（Data Mart）存储引擎的 ES 和 HBASE 可为用户提供灵活、高并发、低延迟的数据查询服务。

2.5.4. 数据加工引擎

不管从上述的 ODS 层导入数据到 DW 层，还是从 ODS 层导入到 DM 层，都需要利用并行调度来提升系统的计算效率。数据加工引擎利用大数据平台分布式并行运算和高吞吐量的特点，使用 HIVE SQL 等计算语言完成全历史数据的加工。利用算法和调度，在不影响用户访问已有数

据的情况下完成每日增量数据的处理，通过独立计算单元实现与交易系统的解耦，从而在交易系统无感知的情况下高效完成历史数据的整合。

2.5.5. 数据服务引擎

全历史数据服务系统通过数据服务引擎和下游数据使用系统对接。该引擎利用 HIVE、ES、HBASE 提供的服务接口，根据用户需求提供匹配的业务数据。如用户的数据挖掘、客户画像、因子分析等需求可以直接利用 HIVE 平台高性能计算的特点获取结果，而全历史数据流水查询等需求可以通过对 ES 和 HBASE 调用返回。通过提供规范的数据结果，数据服务引擎可以方便的对接公司数据中台、服务中台等应用。

3、全历史数据服务系统阶段性成果展示

3.1. 系统上线运行效果

系统上线运行后，各业务系统历史数据的存储方式、加工计算、提供服务实现了标准化和统一管理，完成了各类业务历史数据的整合。历史数据处理效率和历史数据查询效率两方面都能得到保障。

3.1.1. 历史数据采集方面

根据交易数据源数据准备就绪的特点，全历史系统数据采集分为闭市采集、清算后采集两个阶段，每个阶段的采集任务基本能在半小时内完成，随即能提供数据查询服务。对比于传统历史数据每日在清算完成后的采集方案，历史数据提供查询服务的时间有了明显提升。其中数据归档速度提升了 50%，历史数据每日提供服务准备就绪时点提前了两个小时。下图为数据处理效率对比图。

3.1.2. 历史数据调用方面

全历史数据调用性能方面的情况比较复杂，ES 和 HBASE 这种解决方案相较于传统的关系型数据库，涉及到数据量、时间跨度、服务器配置、调用方式等因素都不相同。经过生产实际验证，在查询数据量较小（通常在服务器内存容量的 50% 以下）、存在逻辑运算（比如多表关联）的情况下，传统关系型数据库有着性能方面的优势；当查询数据量超过单台服务器内存容量的 50% 后，ES 和 HBASE 的性能优势就能显现出来，从并发、吞吐量和响应延迟方面都好于传统的关系型数据库。究其原因，是因为 ES 和 HBASE 等都是基于多台服务器的分布式计算解决方案，利用多台服务器资源提升查询性能。另外，ES 和 HBASE 的横向扩展性可以很好的解决数据量不断增大的问题，根据实际使用情况看，扩容节点对于数据查询调用的性能基本没有影响。

3.2. 系统应用效果

全历史数据服务系统可以提供 10 年以上的历史数据的高性能查询服务，使很多新颖的业务

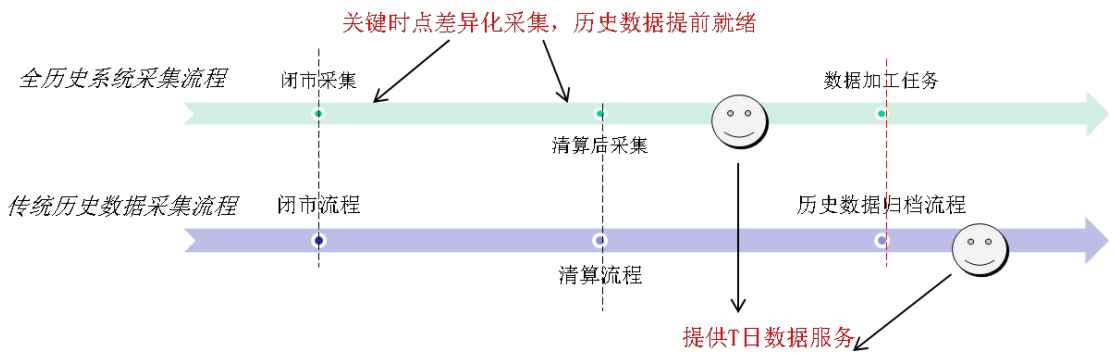


图 5：全历史数据系统与传统历史数据系统采集数据效率对比

图 6：全历史流水查询界面举例

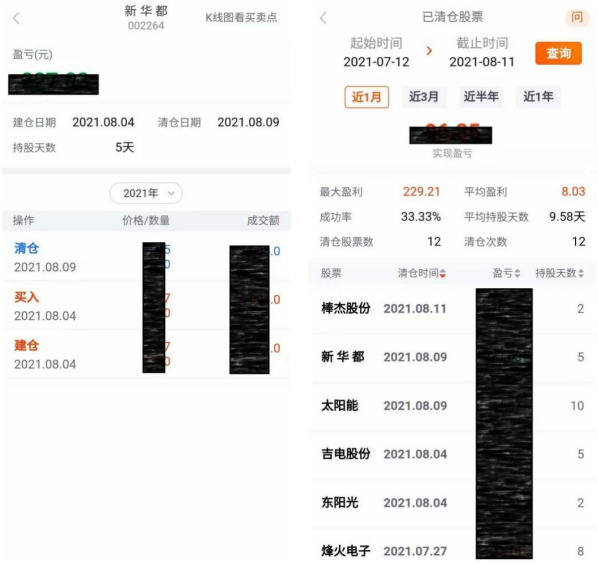


图 7：清仓股票应用示例

需求得以实现。

3.2.1. 全历史流水查询

全历史数据服务系统最直接的应用就是全历史流水查询，传统历史数据系统一般只能提供一到两年内的流水查询服务，远期历史数据查询需要到现场临柜导出。有了全历史数据服务系统，用户可以直接在手机 APP 等客户端直接查询全部委托、成交、打新中签、登录等流水情况。

3.2.2. 清仓股票查询

该功能的灵感来自于投资者的实际需求：如何快速了解自己投资的某只股票的盈亏情况？有了全历史交易数据，我们可以从多个角度分析一只股票。如它的建仓时点，建仓股价；后续的买入和卖出时点及股价；直到清仓的时点和股价。通过整个过程的买入卖出资产运算，还能得出该只股票从建仓到清仓整个投资生命周期的盈亏情况，从而对后续的投资行为起到指导作用。

3.2.3. 行情买卖点查询

为了方便投资者对其操作进行直观高效的复盘，可以在日 K 线图上添加历史买卖点的标记，如 B 代表买入，S 代表卖出，T 代表既有买入又有卖出。对于某一交易日内的同类操作标注“成交均价”和“成交量”信息。可以根据交易数据特点设计标记的位置，如买入（卖出）均价小于收盘价时标记在 K 线下方，买入（卖出）均价大于收盘价时标记在 K 线上方。当点击次级窗口下方的交易明细时，可以直接跳转至该股当日交易明细界面，显示内容包括操作、时间、价格等。应用效果可以参看图 8。

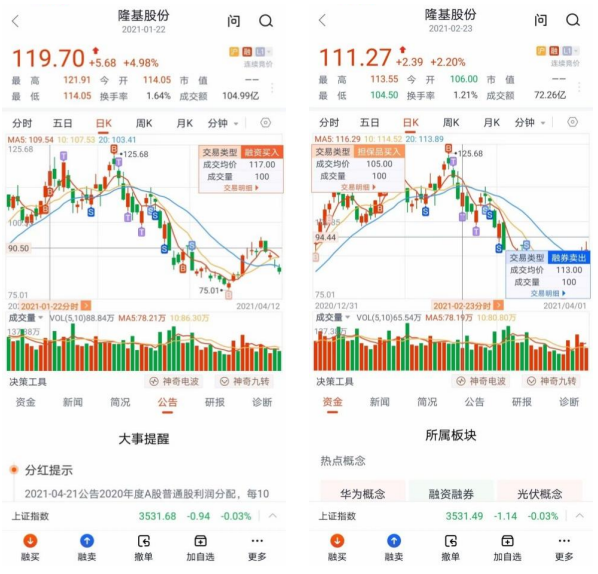


图 8：行情买卖点应用示例

4、总结

“以史为鉴知兴替，以史正人明得失，以史化

风浊清扬”，我们从历史数据中获得的不仅仅是经验和教训，更是对未来的预测，从而找到发展的动力和前进的方向。本文从历史数据重要性出发，介绍了证券行业历史数据在传统系统架构下的应用现状，进而提出一套利用信创大数据技术实现全历史数据服务系统的解决方案。该方案的特点是全面国产化，包括服务器、操作系统、数据库、中间件以及大数据平台各方面。在系统实现方面，本文阐述了如何实现全历史数据的标准化整合、海量数据存储、高效数据查询服务等。通过某证券公司全历史数据服务系统的实践，对系统上线后的运行和应用效果进行了说明。

从系统实践的阶段性效果来看，基于信创大数据技术实现的全历史数据系统是成功的。一方面它解决了传统系统架构下一些固有的问题，另一方面目前提供的功能都得到了业务人员和投资者的好评。可以预见的是，全国产化的历史数据查询服务不但可以满足很多短期历史数据服务无法响应的即时查询需求，而且在一些机器学习的应用方面，如多维度分析、模型验证、模型优化等起到重要的作用。在证券行业追求精细化服务、个性化服务、创新性服务的时代，全历史数据服务系统的实现一定能给广大从业人员提供新思路，带来新价值。

国产分布式数据库在证券行业的应用价值

颜龙 / 国信证券股份有限公司 系统运行总部 深圳 邮箱：yanlong@guosen.com.cn

随着证券业务发展以及数据库技术进化，加上国产化加速的推动下，传统数据库逐渐面临困境。本文将解读这些困境，以此出发阐述我们关于国产分布式数据库在证券行业内应用价值的思考。然后通过我司的反洗钱实践案例来进行探索验证，并尝试讲述国产分布式数据库如何扬长避短，在证券行业的各类业务场景下产生最大化的价值。

1、概述

近年来，证券行业线下服务转型线上化进程加速，包括营销获客的方式、AI 单向智能开户、非现场业务的办理、在线直播、小程序、小视频等互联网方式的使用等，同时近年证券市场行情火爆，证券用户数量和并发量大幅提升，从而对支撑业务的 IT 系统及数据库提出了更高要求；伴随在证券公司进入全面创新发展阶段，证券业务品种也日渐增加、业务流程复杂度不断提高，现有非国产集中式数据库架构在满足新业务、新监管规定以及今后一段时期内部控制管理的高效率监控及管理的需求方面已逐渐困难。

证券行业为数据密集型行业，发展至今已经累积了海量的高价值数据，目前每天产生海量的新数据。在海量用户和大数据量下，当前行业主要使用的国外传统集中式数据库弊端逐渐体现：集中式数据库体系架构缺少计算存储分离、弹性伸缩能力、跨数据中心的高可用能力；容量面临瓶颈，依赖垂直扩容，且很难做到业务透明或者无感知，成本高昂；此外无法满足自主可控及国产化的目标。

然而，随着数据库领域技术近年来的飞速发展，云原生数据库、NewSQL、分布式数据库等具备业界代表性的数据库产品进入人们的视野。基于对数据库领域技术发展趋势以及新技术产品的了解，结合证券行业现状以及我司 IT 系统的实际情况，我们初步认为国产分布式数据库

有能力解决我们的当前困境，可以满足证券行业的业务场景需求，且在我司具备落地的条件。接下来我们将对国产分布式数据库在证券行业中的应用价值进行探索和验证。

2、国产分布式数据库

2.1 基本能力和适用场景

现在让我们来看看分布式数据库具备了哪些基本能力，解决了哪些问题，以及可适用于哪些场景。

首先，受限于传统架构，集中式数据库使用复制和切换作为主要手段的高可用模式已逐渐无法满足金融交易业务场景日益增高的可用性要求。而分布式数据库具备了更完善的高可用能力，以一个集中统一的视角管理所有数据库组件，任何组件异常都可实现自动切换，保证整体的可用性。此外，数据通常由多副本保存，主副本与其他副本之间通过 raft 或 paxos 等协议实现数据的强一致性同步，可保证数据不丢失。

然后，针对容量瓶颈，包括计算能力不足和数据大容量问题，分布式数据库使用了存储计算分离和数据分片的技术，使得其架构支持计算能力和存储的横向扩展。一方面，集群的计算任务主要由计算节点承担，计算节点可以做到无状态从而实现线性扩展；另一方面，数据按照特定规则切分成分片，每个分片保存特定部分数据。由此，我们可以通过增加计算节点来扩充计算能力，通过增加分片来实现数据库量的扩容，且理论上是无限扩容的，而在这个基础上可继续实现弹性扩缩容。

其次高并发问题是传统集中式数据库难以解决的问题，因为单台服务器的并发和计算能力总是有上限的。而对于分布式数据库，一方面，应用的并发会话可以由多个计算节点承担，分散了并发访问的压力；另一方面，分布式架构将数据打散到了各个分片之中，相当于分散了并发请求带来的读写压力。因此在理想的情况下，分布式架构下的并发能力也是支持线性扩展的。

再次，成本问题也是传统集中式数据库所面临的痛点。就如前所述，传统架构缺少横向扩展能力，因此面临增长的业务、扩大的数据容量，数据库只能选择垂直扩展来获得服务器资源上的补充。但昂贵的 CPU 资源向上堆砌、内存和存储扩容所带来的成本不菲，并且很容易达到最终瓶颈。分布式数据库则将垂直扩容转变成为了横向扩容，构成计算存储节点的每一台服务器都并不强依赖高性能服务器。在这个情况下，增加节点可以轻松解决资源扩容问题，而成本相对于垂直扩容则要低很多。

最后，国产化的大环境问题也是证券行业目前重度依赖的非国产商业数据库所无法绕开的问题。国产分布式数据库目前基于开源数据库自研扩展为分布式架构，甚至做到真正意义上的全自研。因此国产分布式数据库是国产化的一个切实可行的发展思路。

综上所述，我们认为国产分布式数据库有能力解决行业数据库目前所面临的困境，在 OLTP 在线交易型、OLAP 在线分析型业务、互联网高并发型、交易型和分析混合型的证券业务场景下，都将有不同程度的应用价值。

2.2 基本现状简述

如我们所知，当今分布式数据库主要有两大类：第一类是从单体数据库和自研中间件演进而来的分布式数据库，我们习惯称之为数据库中间件型分布式数据库，目前在国内比较成熟的有 TDSQL-MySQL、TDSQL-PG、GoldenDB、HotDB、GuassDB-300 等；第二类叫做 NewSQL，也叫原生分布式数据库，国内相对成熟的有 TiDB、OceanBase。此类数据库架构的每个组件都是基于分布式进行设计的，天生自带分布式基因。NewSQL 从分布式 NoSQL 存储出发，演化出关系型数据库能力，从而进化成为分布式数据

库；而中间件型分布式数据库则从关系型数据库出发，融合分布式特性增强架构能力，最终成为分布式数据库，二者殊途同归。由于关系型数据库的实现难度是远大于分布式存储的，因此中间件型分布式数据库相当于走了捷径，大幅降低了产品工程开发的工作量，同时降低了引入风险的可能性，基于现有生产数据库也使其能够更快地走向成熟、稳健。而 NewSQL 的发展道路相对艰难，但它也带来了数据库架构革命性的改变。

基于以上情况，同时针对我司 IT 系统实际情况进行考量，我们尝试在 NewSQL 和中间件型分布式数据库中各选其一进行探索和引入。因篇幅原因，我们接下来选取其中一个关于 TDSQL-PG 的实践案例进行介绍，验证国产分布式数据库在我司的应用价值。TDSQL-PG 为系列产品中具备 HTAP 特性的版本，兼容我司相关传统数据库协议。

3、探索案例

我司反洗钱系统目前拥有 7T 业务数据（大表记录数十亿级），应用同时具备了 OLAP 和 OLTP 两种业务行为，并且使用了存储过程、窗口函数等复杂数据库功能，在众多业务系统中具有代表性。我们尝试使用 TDSQL-PG 对其进行适配落地，来验证分布式数据库的应用价值。我们组建了项目组并进行了大量的适配和测试工作：挑选了典型的业务场景；部署了全量数据的测试环境；进行了异构数据迁移；针对目标数据库产品特性进行了应用的 SQL 改造；基于分布式的表结构改造；后台作业、框架升级、页面功能改造以及各类数据库软件适配性问题的解决。

3.1 选择 TDSQL-PG 进行适配

如前所述，反洗钱业务兼备了 OLTP 和 OLAP 的特征，对此 TDSQL-PG 的 HTAP 能力具备独特优势：

一是满足 OLTP 业务场景的高并发需求，同时也能解决计算能力的不足的问题；

二是满足 OLAP 业务场景的计算密集型需求，同时也能解决大数据量下的时延以及吞吐量问题；

再就是可获得代价、性能、维护成本之间的权衡，同时考虑大批量数据的迁移改造成本。

此外 TDSQL 支持分布式事务、自定义函数、存储过程、窗口函数，分片键改造支持自动指定，等。因此 TDSQL-PG 具备较为契合的特性来对反洗钱系统进行适配。

3.2 适配收益

3.2.1 高可用架构能力的提升

当前反洗钱系统运行在集中式数据库上，使用传统复制技术部署了一台实时同步的备机。当主库故障的时候，需手工切换至备库，同时应用修改 ip 地址指向以恢复服务。即一主 1 备架构，切换为手工操作，高可用切换时效约为分钟级。

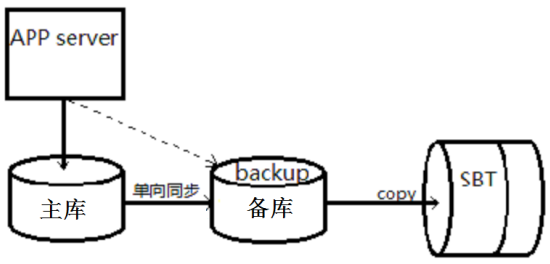


图 1：反洗钱系统数据库的传统集中式架构

在反洗钱 TDSQL-PG 测试环境中，我们部署了 3 个 DN 的分布式集群。每个 DN 即为一个高可用单元，由 1 主 1 备共两个副本组成，分别部署在 2 台服务器上。从副本数量上来说，其高可用能力相比当前集中式架构得到了增强。且得益于 TDSQL-PG 的分布式架构，反洗钱数据库有了分散集中故障风险的能力。每个 DN 的主备之间均为自动故障切换，时效为秒级，因此从切换时效上来看，反洗钱数据库的高可用能力也得到了增强。

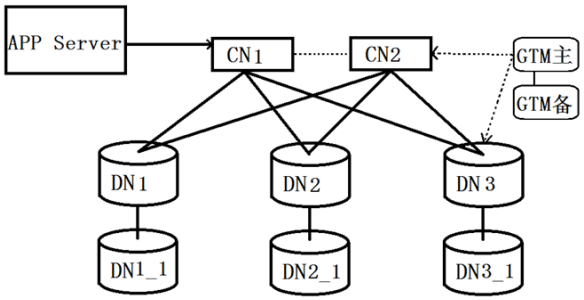


图 2：反洗钱系统数据库的 TDSQL-PG 分布式集群架构

3.2.2 可扩展性方面的改善

反洗钱系统当前的集中式架构仅适用垂直扩容。对于物理机服务器计算能力来说是几乎无法实施扩容的，如 CPU、内存。而对于存储容量来说，垂直扩容的量是有上限的，取决于硬件支持能力。我们无法无限制地挂载存储盘到一台服务器上，否则会引发服务器运行稳定方面的问题，并给运维工作带来困难。反洗钱主库为接入了 16T 容量 FC-SAN 存储的物理机服务器，备库为 VSAN 存储物理机服务器。当前主库服务器挂载的存储容量已到达 Linux LVM 单卷上限，挂载存储盘数量已达 18 个，继续垂直扩容的价格、维护成本极高。当前存储使用率高达 90%，而行业监管要求存放 5 年历史数据，因此预测其数据量还有 40% 左右的上涨，当前反洗钱数据库架构已面临严峻的容量考验。

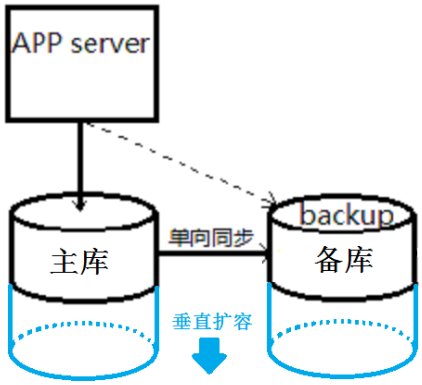


图 3：反洗钱集中式架构的垂直扩容方式

反洗钱 TDSQL-PG 架构具备分布式优势，适用横向扩展。数据以分片的方式存放在了各个 DN 中。当集群中 DN 服务器资源平均使用率较高的时候，比如存储容量，我们可以给集群添加一组 DN，其中包含 2 台服务器。也就是说，随着反洗钱业务数据量不断的上涨，总是可以通过给集群添加 DN 来进行容量扩充。而且操作在线进行，可通过集群自身的平台能力进行自动化管理。

不仅存储容量，承载了集群主要计算能力的 CN 节点也支持横向扩展。当集群计算能力不足时，则添加适当数量的 CN 节点，即达到扩充计算能力的目的。

3.2.3 性能对比

我们给出典型业务场景下的性能对比结果

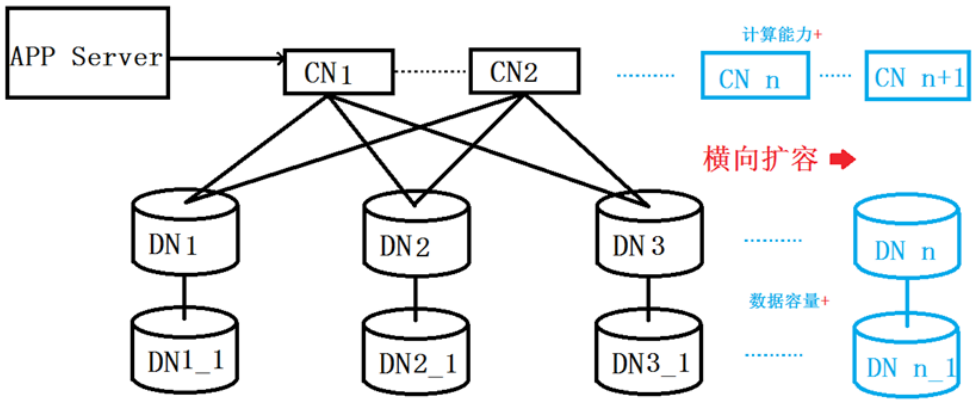


图 4：反洗钱分布式架构的横向扩容方式

表 1：反洗钱典型业务场景下的性能测试数据对比

测试场景	任务特点	当前生产	TDSQL-PG 分布式（双分片）跑 SQL	TDSQL-PG 分布式（双分片）跑存储过程
场景 1：客户联系方式汇总处理	千万级客户信息表做 join、列转行、窗口函数	10 分钟	10 分钟	12 分钟
场景 2：可疑交易监测：非交易目的大额资金转入监测	亿级流水表统计分析	101 分钟	18 分钟	18 分钟
场景 3：风险等级：频繁大额转托管统计	亿级流水表统计分析，多级嵌套子查询	65 分钟	3 分钟	30 秒

（表 1）。

TDSQL-PG 双分片架构在三个场景下优于或不差于现有系统，整体上来看反洗钱 TDSQL-PG 架构表现良好，能够满足反洗钱业务场景的需求。

3.2.4 自主可控及国产化

在测试过程中，我们分别对 X86 以及 ARM 平台下的 TDSQL-PG 版反洗钱系统进行了验证，运行稳定,达到预期。使用 TDSQL-PG 的情况下，反洗钱系统即可脱离集中式数据库的限制，实现完全自主可控，因此满足行业国产化的要求。

3.3 案例总结

经过本次探索和实践，我司发现基于 TDSQL-PG 的反洗钱系统在海量数据下的性能、扩展能力、高可用、灾备、运维、成本节约等方面都有显著提升，TDSQL-PG 作为分布式数据库能够为证券业务场景产生应用价值。

4、应用价值

首先，国外商业数据库已有几十年的发展历程，占据了全球高份额市场，产品能力成熟。相比之下，国产集中式数据库在综合产品能力上还处于起步和发展阶段，需要借助架构来弥补劣势；其次，国产服务器在硬件能力上还略逊色于目前市场主流品牌，因此凸显单体服务器性能瓶颈问题；第三，国产基础设施的可靠性仍需要进一步提升。以上需求导致了我们需要通过分布式架构来解决容量扩展问题，并提升可靠性和冗余度。

以上种种，说明了分布式数据库的一个核心特点和价值：架构横向扩展能力。分布式数据库有能力执行单台服务器无法完成的计算、存储任务，借助分布式架构可以提高系统的整体可靠性和吞吐能力。但同时我们也注意到，分布式数据库有擅长的业务场景，也有能力无法覆盖的场景。面对不同的应用与环境，分布式数据库既拥有特定的优势也存在某些劣势。正如不存在完美的架构，单一的数据库架构无法覆盖我司所有的业务场景。

从国产分布式数据库的行业应用情况和发展潜力进一步分析，同时结合我司的实践案例和业界同行的使用经验，我们认为国产分布式数据库经历多年的打磨，目前已具备成熟、可持续发展的生态。此外在银行、保险、证券等金融行业有许多成功案例，其中包括银行核心系统案例，其稳定性、可靠性已得到验证，可以满足金融级数据库的要求。然后，具有 HTAP 代表性的反洗钱系统成功完成基于 TDSQL-PG 的迁移，证明以 TDSQL-PG 为代表的国产分布式数据库有能力替代证券系统特定业务类型现有的集中式数据库。其次，分布式架构为我们带来计算、存储横向扩展能力的同时，也不能忽视分布式事务带来的时延问题，在一些低延时场景还需要连同业务角度一起去研究其可行性；此外，产品成熟度也是我们对国产分布式数据库进行选择的重要考量之一，运维工具便利性、附属功能缺失、软件 BUG 是目前各类国产分布式数据库所面临的普遍问题。最后，我

们还要充分理解分布式数据库给我们带来的管理方面的挑战。分布式数据库架构相对于集中式数据库更庞大、运维复杂度更高。同时，我们还需要关注资源使用率的问题，避免分布式架构导致的服务器资源浪费问题。

整体上说，在未来的一段时间里，国产分布式数据库可以替换我司部分场景下的业务系统。而随着产品不断的更新优化和技术发展，国产分布式数据库可以为我们带来越来越多的可适配场景。国产分布式数据库在我司将有越来越多的用武之地，可在证券行业产生越来越多的应用价值。

I 信息资讯采撷
nformation

监管科技全球追踪

监管科技全球追踪

国际动态

国际清算银行（BIS）召开第 11 届研究网络会议：金融监管、反垄断与数据隐私

2021 年 10 月 6 日至 7 日，国际清算银行（BIS）召开了为期两天的第 11 届研究网络会议，会议主题为：金融监管、反垄断与数据隐私。当前，大型科技公司（Big Tech）正在迅速扩大其在支付系统和金融服务领域的足迹，这些科技巨头既为金融服务带来有益的创新和竞争，提高效率和金融包容性，但也给监管者带来了棘手的问题。本次会议召集了央行行长、政策制定人员、央行与学术界的研究人员以及其他利益相关者，就如何更好地监管金融领域的大型科技公司，并促进公共利益进行了讨论。

G20：完善跨境支付路线图 研究央行数字货币在其中的作用

2021 年 10 月 13 日，二十国集团（G20）主席国意大利在美国华盛顿主持召开 G20 财长和央行行长会议，讨论全球经济和卫生形势、支持低收入国家应对疫情冲击、可持续金融、跨境支付、国际税收等议题，会后发布了公报。人民银行行长易纲以视频连线方式出席会议并发言，陈雨露副行长陪同参加。会议认为应加强非银行金融中介机构的风险抵御能力，继续落实关于完善跨境支付的路线图，强调全球稳定币的运行需以遵守所有相关法律和监管要求为前提，鼓励相关国际机构继续研究中央银行数字货币在完善跨境支付中的作用及其对国际货币体系的影响。

国际清算银行（BIS）发布《大型金融科技公司：论数据隐私与竞争之间的新关系》专题报告

2021 年 10 月 21 日，国际清算银行（BIS）发布《大型金融科技公司：论数据隐私与竞争之间的新关系》（Big techs in finance: on the new nexus between data privacy and competition）专题报告。报告指出，大型金融科技公司的商业模式依赖于通过数字平台实现大量用户之间的直接交互，并产生大量用户数据作为重要副产品。企业借助于自然网络效应对于这些数据的进一步利用，使得用户活动与数据也进一步增加，从而形成循环。基于数据－网络－活动循环的自我强化性质，一些大型科技公司已经涉足包括支付、资金管理、保险和贷款在内的各类金融服务，从而提高金融行业效率，扩大金融包容性，但也带来了与市场力量和数据隐私相关的新风险。效率和隐私之间的政策权衡将取决于社会偏好，并因司法管辖区而异，从而增加了国内和国际间政策协调的必要性。

国际清算银行（BIS）发布《数字货币对新兴市场及发展中经济体的意义》工作报告

2021 年 10 月 29 日，国际清算银行（BIS）发布《数字货币对新兴市场及发展中经济体的意义》（What does digital money mean for emerging market and developing economies?）工作报告。报告指出，央行和非央行货币之间有着根本性的区别：央行货币是作为一种公共品提供的，因此央

行货币与私人加密资产和稳定币之间有着绝对的区别。虽然推动它们被采用的因素可能相似，但产生的结果却截然不同。对于新兴市场与发展中经济体而言，新形式的数字货币可能会对发展宏观经济和跨境支付带来一定挑战，然而，建立在现有金融管道基础上的技术进步已经提高了新兴市场与发展中经济体的包容性与效率。综上所述，到目前为止，新型私人数字货币带来的最大贡献或许是为新兴市场与发展中经济体中存在的金融包容和跨境支付挑战吸引更多关注。

欧美动态

Gartner 发布 2022 年 12 大重要战略技术趋势

2021 年 10 月 18 日，Gartner 发布了企业组织在 2022 年需要关注的 12 个战略技术趋势，具体包括：生成式人工智能（Generative Artificial Intelligence）机器学习技术、进行数据管理、集成的数据编织（Data Fabric）技术、分布式企业（Distributed Enterprise）组织相关技术、数字化转型技术底座的云原生平台（Cloud-Native Platform，CNP）、进行自我管理的物理系统或软件系统的自治系统（Autonomic Systems）、进行结构化决策的决策智能（Decision Intelligence，DI）、组合式应用架构的组装式应用程序（Composable Applications）、进行快速识别、审查和自动化，实现加速增长和业务弹性的超级自动化（Hyperautomation）、对个人信息和敏感信息实施保护的隐私增强计算（Privacy-Enhancing Computation，PEC）、网络安全网格（Cybersecurity Mesh）技术、AI 模型集成的人工智能工程化（AI Engineering）、全面体验（Total Experience，TX）的业务战略。

美国消费者金融保护局（CFPB）要求科技巨头提交其支付系统的有关信息

2021 年 10 月 21 日，美国消费者金融保护局（CFPB）发布了一系列命令，收集在美国运营支付系统的大型科技公司的有关商业行为信息。这些信息将有助于 CFPB 更好地了解这些公司如何使用个人支付数据并管理对用户的数据访问，以便能够确保充分保护消费者。CFPB 表示，大型科技公司正急切地扩张自己的帝国，以获得对消费者消费习惯的更大控制和洞察。CFPB 已命令他们提供有关其商业计划和实践的信息，这些命令是根据《消费者金融保护法》条下达的。CFPB 有法定权力命令支付市场的参与者交出相关数据，帮助该局监测消费者面临的风险，并公布符合公众利益的汇总调查结果。该公司已向亚马逊、苹果、Facebook、谷歌、PayPal 和 Squarle 等六家公司发出命令，并表示可能还会针对支付宝和微信支付等中国公司。

法兰西银行行长发言：央行应在数字化创新中扮演何种角色

2021 年 11 月 12 日，国际清算银行（BIS）于官网发布了 8 日法兰西银行（Bank of France）行长François Villeroy de Galhau 在新加坡金融科技节上发表的演讲全文：央行应在数字化创新中扮演何种角色（Digital innovation-what role can we play as central banks?）。Galhau 指出，尽管数字创新的蓬勃发展有可能改变整个金融体系。但央行必须接受这些变化，并保持金融体系的稳定。

创新和稳定在短期内可能会出现矛盾，但它们通过共同的价值观相辅相成，其中的关键是公众对于货币的信任。要有效实施创新和稳定的双重原则，就需要一个合作的环境。Galhau 从监管如何应对新的风险，以及央行如何以央行数字货币为例促进创新入手，重点讨论了将这些原则付诸实践的方法。

欧洲央行：未来金融系统的脆弱性在增加

2021 年 11 月 17 日，欧洲央行的《金融稳定评论》显示，随着经济复苏，尽管疫情引发的金融风险在降低，但长期结构性问题重现，未来金融系统的脆弱性增加。由于全球供应链紧张和能源价格的上涨，欧洲央行认为，欧元区的经济复苏已降低短期风险。然而，由于某些资产市场估值紧张、公共和私人债务水平高以及非银行风险承担增加，脆弱性正在增加。欧洲央行副行长 Luis de Guindos 表示，与 6 个月前相比，现在企业高失败率和银行亏损的风险大大降低，但与大流行相关的风险并没有完全消失。随着 2021 年上半年经济的反弹，欧元区企业利润有所回升。部分由于这一发展，企业破产仍低于大流行前的水平，尽管在受大流行影响最严重的经济部门，企业破产有所增加，并可能继续上升。与此同时，全球供应链的紧张和最近能源价格的上涨可能对经济复苏和增长构成挑战。

英国国家网络安全中心发布 2021 年度回顾报告

2021 年 11 月 17 日，英国国家网络安全中心（NCSC）发布 2021 年度回顾报告，本报告是该系列自 2016 年以来的第 5 份报告，着眼于 2020 年 9 月 1 日至 2021 年 8 月 31 日的关键发展和亮点，回顾了近一年来的网络情况。本篇报告主要包含了威胁、弹性、技术、生态系统和国际领导力 5 个章节，涵盖了主动网络防御项目、可疑邮件报告服务、新冠疫情事件等相关内容。

美国网络安全与基础设施安全局发布 5G 云基础设施安全指引报告第四部分

2021 年 12 月 16 日，美国网络安全与基础设施安全局（CISA）联合美国国家安全局（NSA）共同发布了 5G 云基础设施安全指引的第四部分：确保云基础设施整体性。该报告重点关注于平台的整体性，微服务基础设施的整体性，启动时间的整体性和构建时的安全性，以确保 5G 云资源（容器镜像、模板、配置等）不被擅自修改。本系列 4 份报告还记录了 5G 云终端和多云环境中零信任思维相关的最佳实践。

亚太动态

人民银行行长易纲在国际清算银行（BIS）监管大型科技公司的国际会议上发表讲话：中国大型科技公司监管实践

2021 年 10 月 7 日，人民银行行长易纲在国际清算银行（BIS）监管大型科技公司的国际会议上发表讲话：中国大型科技公司监管实践。易纲表示，在技术进步的推动下，中国金融科技蓬勃发展，同时也为中国监管当局带来了新挑战：一是无牌或超范围从事金融业务；二是支付业务存在违规行为；三是通过垄断地位开展不正当竞

争；四是威胁个人隐私和信息安全；五是挑战传统银行业的经营模式和竞争力。易纲强调，为应对上述挑战，中国持续弥补监管制度的“短板”，陆续出台了推动平台经济规范健康持续发展的措施，并始终秉承以下三条理念：一是始终坚持“两个毫不动摇”，支持民营经济、互联网经济和数字经济健康发展。二是不断增强政策透明度和可预期性，保护产权和知识产权，保护隐私，促进公平竞争。三是坚持市场化、法治化、国际化方向，创造良好营商环境，扩大高水平对外开放，在数字领域强化科技创新国际合作。

人民银行等五部门发布《关于规范金融业开源技术应用与发展的意见》

2021 年 10 月 21 日，人民银行办公厅、中央网信办秘书局、工业和信息化部办公厅、银保监会办公厅、证监会办公厅联合发布《关于规范金融业开源技术应用与发展的意见》（简称《意见》），《意见》要求金融机构在使用开源技术时，应遵循“安全可控、合规使用、问题导向、开放创新”等原则。《意见》鼓励金融机构将开源技术应用纳入自身信息化发展规划，提升自身对开源技术的评估能力、合规审查能力、应急处置能力、供应链管理能力等，积极参与开源生态建设。《意见》强调要加强统筹协调，建立跨部门协作配合、信息共享机制，完善金融机构开源技术应用指导政策，探索建立开源技术公共服务平台，加强开源技术及应用标准化建设等。

2021 新加坡金融科技节开幕 新加坡金融管理局（MAS）宣布推出四大数字平台促进数据流动和绿色金融发展

2021 新加坡金融科技节于 2021 年 11 月 8 日盛大开幕。新加坡金融管理局（MAS）宣布将继续推动数据应用，在 2022 年下半年之前建成四个平台：一是绿色足迹通用信息披露平台：根据不同司法辖区要求，将输入数据转换成不同的报告框架，简化环境、社会和政府（Environment, Social & Governance，ESG）信息披露过程，降低信息获得难度。二是绿色足迹数据协调平台：聚合来自主要 ESG 数据提供商以及其他平台的可持续性数据，并允许通过数据分析提出新的数据见解，为可持续投资决策提供更好支持。三是绿色足迹 ESG 登记平台：记录和维护不同部门认证机构的 ESG 认证以及由合格第三方审计员验证的数据和指标来源。这个基于区块链的注册管理机构将为金融机构、企业和监管机构提供这些认证数据的单一访问点，并促进可信数据流动。四是绿色足迹市场平台：帮助新加坡及该地区的绿色技术提供商与投资者、风投公司、金融机构建立联系，促进伙伴关系、创新和绿色科技投资。

人民银行行长易纲发表演讲 将测试数字人民币对货币政策、金融市场和金融稳定等方面的影响

2021 年 11 月 9 日，人民银行行长易纲在芬兰央行新兴经济体研究院成立 30 周年纪念活动上的视频演讲中表示，人民银行妥善研发设计数字人民币方案，有效降低负面影响。首先，人民银行坚持数字人民币的 M0 定位，不计付利息，降低与银行存款的竞争。其次，采取双层运营体系，即央行实施中心化管理，保证对货币发行和货币政策的调控能力；商业银行和支付机构作为中介，为公众进行数字人民币兑换并提供支付服务。再次，设置了钱包余额上限、交易金额上限等制度摩擦，尽可能降低挤兑风险。下一步，人民银行将根据试点情况，有针对性地完善数字人民币的设计和使用。

上海数据交易所揭牌成立

2021 年 11 月 25 日，上海数据交易所揭牌成立仪式暨 2021 上海全球数商大会在上海举行。上海数据交易所的成立，以“全国五大首发”（全国首发数商体系、全国首发数据交易配套制度、全国首发全数字化数据交易系统、全国首发数据产品登记凭证、全国首发数据产品说明书），破解数据交易“五难”问题，将重点聚焦确权难、定价难、互信难、入场难、监管难等关键共性难题，形成系列创新安排。

2020 年度金融科技发展奖评审领导小组会议在京召开

2021 年 12 月 15 日，人民银行消息，近日，2020 年度金融科技发展奖评审领导小组会议在北京召开。会议由人民银行副行长、金融科技发展奖评审领导小组组长范一飞主持，证监会副主席、金融科技发展奖评审领导小组副组长方星海出席会议。会议总结了 2020 年度金融科技发展奖评审工作，评定了 163 个获奖项目，其中特等奖 1 项，一等奖 12 项，二等奖 59 项，三等奖 91 项，涵盖架构转型、基础研究、新技术应用等多方面内容。

阿根廷央行（BCRA）宣布将严查为加密投资提供“超额回报”的公司

2021 年 12 月 22 日，阿根廷央行（BCRA）发表声明，称正在对加密资产投资公司进行调查，并将严查为加密投资提供不合理“超额回报”的公司。阿根廷央行表示，这些公司可能以庞氏骗局的形式运作，正在研究对其采取法律行动的可能性，并对投资者发出警告，使用这些平台进行投资将会承受无法估量的风险。

监管部门拟对证券期货业网络信息系统安全保护实行分级

2021 年 12 月 23 日，证监会科技局组织行业机构起草了《证券期货业网络安全等级保护工作指引（征求意见稿）》并经行业自律协会印发各证券期货经营机构征求意见。《工作指引》适用于“核心机构”和“经营机构”。定级依据显示，证券期货业网络和信息系统的安全保护等级应根据网络和信息系统发生服务能力异常或数据损毁、泄露等网络安全事件后，对国家金融安全、社会秩序、投资者合法权益造成的损害程度进行定级，具体分为五个等级。

2022年二季度《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，以季度为单位发刊，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。2022 年二季度征稿主题如下：

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

（一）应用技术研究

主要包含但不限于：语音识别与自然语言处理，计算机视觉与生物特征识别，机器学习与神经网络，知识图谱，服务机器人技术。

（二）应用场景研究

主要包含但不限于：智能客服、语音数据挖掘、柜员业务辅助等。

主要包含但不限于：监控预警、员工违规监控、交易安全等。

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、本期投稿截止日期：2022 年 6 月 30 日。

4、投稿联系方式 021-68607129，021-68607131 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。

《交易技术前沿》编辑部

证券信息技术研究发展中心（上海）

附件：投稿格式（可通过电子邮件索要电子模板）

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）

杂志订阅与反馈

各位读者，如您想订阅《交易技术前沿》纸质版，欢迎扫描右侧二维码填写问卷进行订阅，同时可以向我们提出关于《交易技术前沿》的建议与意见反馈。如您希望赏阅电子版，欢迎访问我们的电子平台 http://www.sse.com.cn/services/tradingservice/tradingtech/sh/transaction/（或扫描封面尾页二维码）。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！

扫描在线浏览

联系电话：021-68607129
021-68607131
投稿邮箱：ftt.editor@see.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
TECH 上交所技术
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