面向高质量企业征信数据集的治理体系构建与实践

吴新丽，时俊苓
中国经济信息社，北京 100073

摘要
针对高时效业务场景下企业征信数据存在的多源异构、时效性不足、质量不一致、安全管控分散等问题，本文提出一种融合数据质量、时效性、安全服务与AI赋能的全链路实时数据治理框架。该体系采用基于优先级判决的增量式多源融合引擎，实现企业实体的精准对齐与冲突消解；基于Kappa混合架构，通过实时流处理与增强批处理双链路协同，满足分钟级至小时级的分级时效处理要求；通过构建内嵌安全能力的统一数据服务层实现数据集安全合规供给。以中国经济信息社企业征信数据平台为验证场景，实践表明：该体系构建的高质量企业征信数据集覆盖工商、司法、经营等100余个核心维度，数据处理周期从T+1压缩至3小时以内，关键数据准确率达99%以上，可为金融征信、行业监管、商业决策等场景提供高可信、高时效、高安全的数据支撑，也为同类高质量数据集治理提供了可复用的工程方法论。

关键词
高质量数据集；企业征信；数据治理；多源数据融合；AI赋能；数据标准化；实体对齐

中图分类号：TP274；F83　　文献标志码：A

Construction and Practice of a Governance System for High-Quality Enterprise Credit Datasets

WU Xinli, SHI Junling
China Economic Information Service, Beijing 100073, China

Abstract
To address the challenges of multi-source heterogeneity, insufficient timeliness, inconsistent quality, and scattered security management in enterprise credit data under high-timeliness business scenarios, this paper proposes a full-link real-time data governance framework integrating data quality, timeliness, security, and AI empowerment. The framework adopts a priority-based incremental multi-source fusion engine to achieve precise entity alignment and conflict resolution. Based on a Kappa hybrid architecture, it coordinates real-time streaming and enhanced batch processing to meet graded timeliness requirements from minute level to hour level. A unified data service layer with embedded security capabilities ensures compliant data supply. Taking the enterprise credit data platform of China Economic Information Service as a validation scenario, the practice shows that the constructed high-quality enterprise credit dataset covers more than 100 core dimensions (industry, commerce, judiciary, operations, etc.). The data

processing cycle is shortened from T+1 to within 3 hours, and the accuracy of key data exceeds 99%. The framework provides high-reliability, high-timeliness, and high-security data support for financial credit, industry supervision, and business decision-making, and offers a reusable engineering methodology for the governance of similar high-quality datasets.

Key words
high-quality dataset, enterprise credit, data governance, multi-source data fusion, AI empowerment, data standardization, entity alignment

1 引言

在数据要素市场化配置持续推进、国家大数据战略全面落地的背景下，高质量、高时效、高安全的企业征信数据，已成为智能风控、实时监管、精准授信与商业决策的重要基础。然而，企业征信数据具有多源异构（来自工商、司法、税务等多渠道及不同三方供应商）、时效分层（不同数据源更新时效涉及分钟级、小时级、天级）、质量不一且强合规要求等显著特征。当前业界普遍以单一数据源与T+1批处理为主，数据在更新速度、信息完整性、服务响应效率及安全管控上存在明显短板，难以形成适配业务需求的高质量数据集，一定程度上限制了数据驱动业务创新的效果。

近年来，流批一体计算、多源数据融合、数据安全治理、大模型辅助治理等技术逐步成熟，为高质量数据集建设提供了技术可能。现有研究在数据融合方法、实时计算架构、数据服务平台等方向取得了一定成果，但面向企业征信数据的复杂应用场景，仍存在以下三方面不足：

其一，端到端的高质量企业征信数据集治理框架较为缺乏。现有研究多聚焦于单一技术点（如多源融合或实时计算），未能将数据接入、分级时效处理（分钟、小时、天）、企业主体统一映射、多源融合与高时效API合规输出等环节整合为一体化治理框架[1-2]。特别是在企业征信领域，高吞吐（日均千万级记录）、强实时（风险监控需秒级响应）的业务压力下，缺乏从数据源到服务交付的全链路协同设计，导致治理能力碎片化。

其二，多源异构数据融合的冲突消解机制尚不完善。企业征信数据源之间存在语义异质性、时效差异和可靠性差异等多重冲突。现有研究大多基于单一的数据源优先级规则进行冲突判决，缺乏可扩展的记录级与字段级融合框架。在实体对齐方面，传统方法对名称简称、同义表述等复杂场景的处理精度仍不理想[3]；在多源冲突消解方面，现有工作未能充分融合优先级判决与字段级规则，难以支撑跨源、跨维度的精细化融合治理[4]。此外，由于企业征信数据质量参差不齐，业务主键空值情况多见，依赖单一确定规则的集成方法容易造成数据重复。

其三，AI技术在治理全流程中的系统化嵌入尚处起步阶段。尽管基于大语言模型的数据清洗、错误校正研究已取得一定进展，但现有工作大多聚焦于通用场景，尚未面向企业征信数据的业务特点进行定制化适配[5]。在实体解析、冲突消解、异常检测等关键环节，大模型能力与实时治理流程的衔接尚未解决，鲜有将大模型系统化嵌入多源融合、质量检测和修复补全全流程的工程实践[6]。更深层的问题在于，AI大模型API调用单次秒级响应，难以直

接嵌入企业征信数据高吞吐、强实时的数据处理流程中，需要在异步、批量化、缓存化等方面进行工程适配。

为系统性地填补上述研究空白，本文提出以下三个具体研究问题：

问题1（融合机制）：如何针对企业征信数据多源异构、主键空值多的特点，设计一种兼顾确定性规则与字段级优先级判决的增量式融合算法，以实现企业实体的精准对齐与冲突消解？

问题2（时效架构）：如何设计分级时效处理架构，满足API推送（分钟级）、数据库同步（小时级）、批量文件（天级）的差异化更新需求，同时控制计算成本？

问题3（AI增强）：如何将AI大模型以非侵入、异步方式嵌入治理流程，在保障主链路性能的前提下，提升非标字段标准化和缺失补全的自动化水平？

研究设计采用设计科学与工程实践相结合的方法，首先，基于中国经济信息社企业征信数据平台的真实痛点（如企业ID不一致、税号空置率高、行政区划表述不规范等），提出治理框架与技术架构；其次，通过工程实现与平台部署，在真实生产环境中进行量化评估，检测验证时效性、数据质量、安全等核心指标；最后，根据反馈迭代优化，形成可复用的工程方案。本文遵循“问题分析→框架设计→技术实现→工程验证”的技术路线。

基于上述研究路线，本文预期的主要贡献如下：

（1）提出一种面向企业征信数据的端到端协同治理框架，实现质量、时效、安全与AI增强的一体化。

（2）设计并实现两项关键使能技术：一是基于优先级判决的增量式多源融合算法，实现多源数据的冲突消解与智能补全；二是基于Kappa混合架构的分级时效处理，通过“实时流处理”与“增强批处理”双路径协同，适配征信数据源的差异化更新频率。

（3）构建内嵌安全能力的统一数据服务层，实现企业征信敏感信息的字段级动态脱敏与审计。

（4）引入AI大模型作为异步智能增强插件，以“规则优先—AI增强—知识反哺”闭环提升行政区划标准化、裁判文书实体抽取等任务的自动化水平。

2 相关工作

高质量企业征信数据集的构建是一项融合数据集成、实时计算、服务化与安全治理的综合性工程。本章从企业征信数据集构建现状、数据治理与高质量数据集构建、多源数据融合与质量提升、实时计算架构、数据服务与安全治理、AI增强的数据治理五个维度，梳理相关领域的技术发展与研究现状，作为本文研究与创新的参照基础。

2.1 企业征信数据特征与治理挑战

企业征信数据在来源、处理及合规层面呈现显著的多源异构、阶梯时效、高质量与强监管特征，对传统数据治理体系构成严峻挑战。

数据来源方面，企业征信数据涉及工商、司法、税务、招投标、知识产权、舆情等多个渠道，涵盖结构化（数据库记录/CSV文件）、半结构化（JSON/XML报文）与非结构化（判决文书/招标公告）数据。不同来源的数据，数据质量参差不齐，存在大量空值、异常值及格式不一致问题，而且更新频率与可信度差异巨大。为提升

数据覆盖与可靠性，同一业务维度（如司法涉诉）也会采购多个来源的数据，这些数据源的数据结构、数据质量、更新频率各不相同，为后续融合治理带来挑战[7]。

数据处理方面，数据接入涉及CSV文件、数据库同步、API接口等形式，数据处理需同时支持API接口数据的分钟级处理、数据库同步的小时级处理，以及CSV文件的天级处理。传统单一的T+1离线批处理模式难以满足高时效业务需求，易造成数据积压与价值延迟。此外，由于关键业务主键（如统一社会信用代码、案号等）常存在空值，在多源数据融合过程中，需要根据多种非空字段分类比较，极易产生数据重复与关联失效问题，进一步增加了治理复杂度[8]。

数据服务与安全合规方面，征信数据的采集、处理与服务需遵循《征信业管理条例》《数据安全法》等法规。这要求治理体系必须内嵌数据授权、全链路审计与动态脱敏等能力[9]。

2.2 数据治理与高质量数据集构建

高质量数据集以数据标准、数据质量、数据时效、数据安全为核心要素。近年来，研究者从战略、组织、标准、质量、安全等多个维度构建数据治理理论框架，并梳理了数据治理研究的演进脉络[10]。推动全生命周期数据治理需要关键技术的支撑，以实现从战略框架到生产实践的落地。在高质量数据集的度量方面，准确性、完整性、一致性、时效性和安全性已成为共识性评价维度[11]。然而，现有研究多侧重宏观框架或单一质量维度优化，针对企业征信多源异构、强时效、高合规的特点，缺少贯穿数据全生命周期的实时治理体系工程实践。

2.3 多源数据融合与质量提升技术

多源数据融合是提升数据集覆盖度与质量的核心，关键难点为实体解析与冲突消解。早期方法依赖规则匹配与字符串相似度计算，在企业简称、跨模态文本[12]等复杂场景下精度不足。后期研究转向表示学习与图神经网络，如融合多级特征的实体对齐模型、基于自适应融合技术的多模态对齐方法，以及基于图结构信息的对齐算法等[13-14]。在冲突消解方面，除优先级、投票等传统策略外，开始引入外部知识库辅助决策[15]。

上述方法在特定数据集上效果较好，但应用于企业征信场景时，面临海量数据、动态更新、业务规则复杂的挑战：一是处理效率无法满足实时要求；二是深度学习模型的可解释性弱，与业务确定性规则融合难度大。本文设计的基于优先级判决的增量式融合引擎，兼顾处理确定性、运行效率与业务适配性，更适合工程化落地。

2.4 实时计算架构的演进

为平衡数据处理能力与时效性，大数据处理架构经历了从Lambda到Kappa、并向流批一体演进的历程。Lambda架构通过独立的批处理层与流处理层保障最终一致性，但系统复杂、维护成本高。Kappa架构以流处理系统统一处理所有数据，简化了架构，更适合实时场景，但在历史数据回溯和复杂批处理任务方面存在成本挑战。流批一体是当前主流趋势，旨在通过统一引擎（如Apache Flink）处理实时与历史数据，并已有针对具体平台的设计与优化研究[16]。

然而，在真实业务环境中，不同数据源的时效性要求存在梯度差异（如API推送数据需分钟级，数据库同步数据可接受

小时级，批量文件可接受天级）。现有架构选型往往需要在“纯流处理的高成本”与“纯批处理的滞后性”之间做出妥协。本文针对企业征信场景的混合时效需求，提出Kappa混合架构——并非采用纯流处理，而是通过“实时流处理”与“增强批处理”双路径的协同设计与智能调度，在保障关键数据极致时效的同时兼顾海量数据处理的经济性。这是对现有架构范式的一种务实演进。

2.5 数据服务与安全治理

数据服务化通过API实现数据资产与应用解耦，是数据中台的核心能力之一。与此同时，数据安全治理的内涵从网络安全向数据自身安全与合规拓展，《网络数据安全管理条例》等法规推动安全治理从“外挂式”检查转向与数据生命周期融合的“内生安全”[17]。当前研究聚焦公共数据授权运营安全框架、金融数据安全体系、生成式AI数据安全风险、联邦学习等隐私计算应用、隐私保护数据集构建等方向[18-21]。

现有实践多将认证、脱敏等安全能力以网关形式外挂于服务层，与底层数据模型、处理逻辑的联动较弱，无法实现字段级、行级精细管控。本文搭建的统一数据服务层内置安全防护能力，将安全策略深度融入数据存储至接口输出全链路，实现安全能力自动化执行。

2.6 AI增强的数据治理

人工智能，特别是大语言模型的突破，为数据治理的智能化带来了新机遇。研究集中于利用AI进行数据自动标注、质量评估、元数据生成和自然语言查询等[22]。在数据集成领域，AI被用于提升实体对齐的语义理解能力；在质量治理方面，AI可用于异常模式检测与根因分析，实现数据问题的主动发现[23]。然而，现有工作大多将AI技术应用于独立的、离线的治理任务，将其作为“智能增强组件”系统化、工程化地嵌入在线数据治理流程中的研究仍较为缺乏。特别是针对企业征信数据，大模型在实体解析、冲突消解、缺失补全等环节的实时或准实时嵌入，尚未形成成熟的工程范式。本文尝试将AI大模型以“增强插件”形式嵌入核心治理环节，通过异步、批量的智能分析机制，解决因增量数据不规范导致的质量修复与智能补全问题，是AI赋能数据治理方向的一次积极工程实践。

2.7 小结

综上所述，现有技术在单一领域已有长足发展，但面向企业征信这一特定场景，缺乏一个将多源融合、分级时效处理、内生安全服务与AI智能增强进行系统性整合，并能通过“治理—处理—服务—优化”闭环实现持续运营的端到端治理框架。本文工作旨在填补这一空白，提供一套经过完整工程验证的系统性解决方案。

3 面向高质量企业征信数据集的治理框架

企业征信数据治理框架的构建，在方法论层面借鉴了国际通用的数据治理体系（如DAMA-DMBOK），以确保治理活动的完整性与规范性。针对企业征信数据“多源异构、时效分层、高质量、高安全”的领域特征，本框架对通用体系进行了重构与深化，具体体现为三个方面：一是提出“流批一体Kappa混合”实时治理架

构，精准适配从分钟级到天级的梯度化时效需求；二是强化“主体归一、冲突消解、企业图谱构建”等征信特有治理环节，并将“异议处理与数据修正”内化为质量管理环节，形成质量闭环。三是增设AI智能数据质量增强环节，以实现对非标字段与长文本的自动化清洗与标准化。总之，该框架不仅继承了通用治理的系统性，更通过征信领域化创新，完成了面向征信数据集治理的专用体系构建。

具体框架如图:3-1

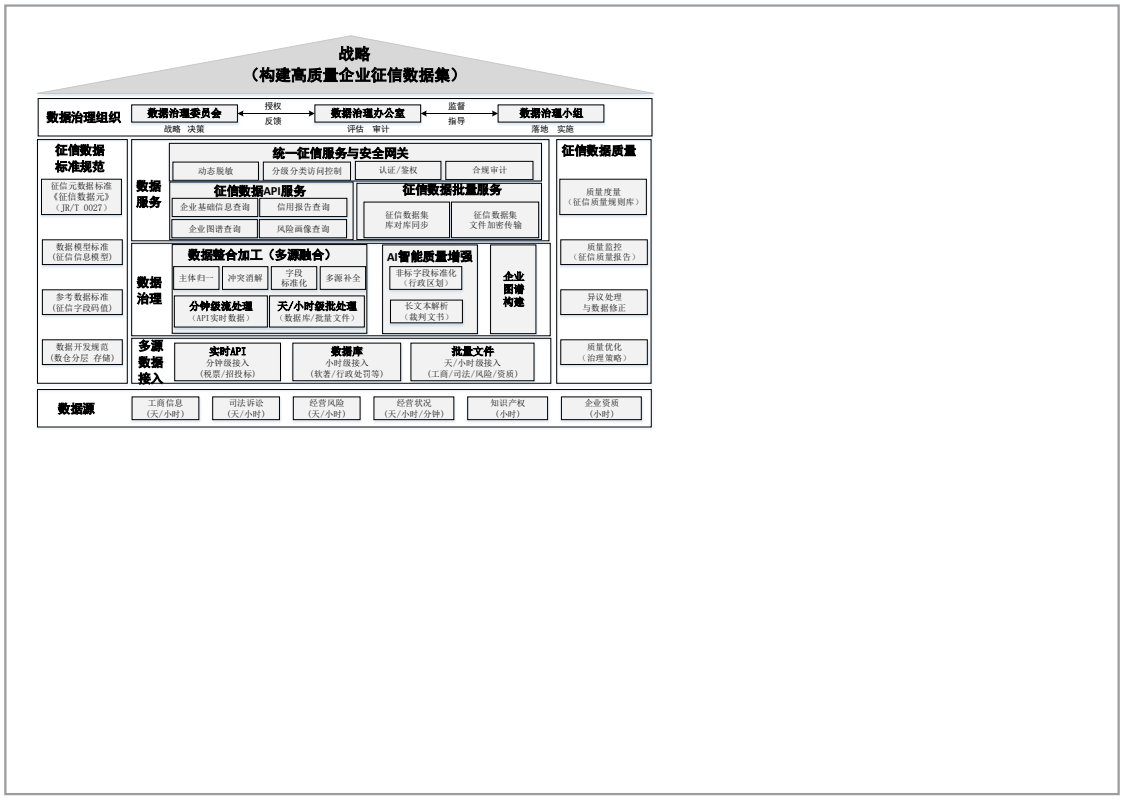

图3-1 企业征信数据治理框架

（1）战略与组织：明确“构建高质量企业征信数据集”的战略目标，设立“数据治理委员会（决策）→ 数据治理办公室（评估/审计）→ 数据治理小组（落地实施）”的三层组织架构，形成决策、协调、执行的闭环。

（2）征信数据标准规范：依据《征信数据元》（JR/T 0027）等标准，制定征信数据元标准、数据模型标准、参考数据标准及数据开发规范，确保字段命名、码值、数仓分层一致性。

（3）多源数据接入：支持API实时接入、数据库同步、批量文件等多种方式接入工商、司法、经营风险等多领域数据源；适配不同数据源的更新频率。

（4）数据治理核心：采用Kappa混合架构，提供实时流式处理与增量批处理双路径，通过差异化调度实现分钟级、小时级和天级的分级时效供给。数据整合加工，集成多源融合引擎、冲突消解算法、AI质

量增强引擎等、完成原始数据到高质量标准化数据集的转换。基于融合后的数据，构建企业知识图谱（股权、投资、担保、上下游关系）。

（5）数据服务：通过统一服务网关提供企业征信API、批量数据服务，内嵌动态脱敏、分级分类鉴权与访问审计，保障数据安全合规输出。

4 关键治理活动的技术实现

企业征信数据既有天/小时级更新的工商、司法数据，也有分钟级处理的税票、招投标数据，面对数据持续增长及更新频繁的现实，数据治理既要满足大批量小时级更新，同时满足分钟级加工计算（贷前报告）的要求。另一方面，数据服务大多以企业为入口，高并发的企业详情点查较多，而企业关联方的多跳查询在关系型数据库中已成为瓶颈，需要图数据库的支持。

基于以上挑战，本方案技术选型需重点考量两大核心因素：一是数据时效性支持能力，涉及处理机制与数据更新能力。二是存储与计算的经济性和扩展性，包括底层存储成本、横向扩展能力与资源弹性。

具体选型依据如下：传统MPP数据库（如Doris、Greenplum）在存储成本与数据湖生态集成上存在瓶颈，故采用Hive构建离线数仓底座，采用spark/flink作为流批一体处理引擎；针对高并发点查与OLAP混合负载，TiDB（HTAP）相比单一OLAP库（如Doris）更具扩展性；针对企业关联图谱查询瓶颈，引入NebulaGraph作为专用图引擎。

表 4-1 技术选型对比与决策依据

需求维度	可选方案	征信场景约束	最终选型	选型依据
离线数仓底座	MPP (Doris/GP) VS Hive	存储PB级海量历史数据,扩展性好,需与数据湖生态集成	Hive	存储成本低 计算和存储的扩展性更好,生态多样性更丰富,与Spark/Flink集成好,适合构建稳定、低成本的离线批处理数仓。
流批处理引擎	Spark Streaming VS Flink	需同时支持分钟级实时流和小时级批量,且需状态管理	Flink+Spark	Flink处理高时效API数据,Spark处理复杂批量融合任务
高并发点查与OLAP	单仓OLAP (Doris) VS HTAP(TiDB)	需支撑日均千万级企业详情点查,同时满足复杂分析	TiDB	HTAP架构天然支持行存点查与列存分析,扩展性强
图谱查询	关系型库+自连接 VS 图数据库	股权穿透、实际控制人挖掘需多跳关联(深度>3)	NebulaGraph	关系型库多跳JOIN性能急剧下降,图数据库毫秒级响应
全文检索	数据库like查询 VS ES	企业名称模糊搜索、地址关键词匹配	ES	倒排索引分词检索,支持高并发模糊匹配

本章系统阐述企业数据治理体系的实现路径。技术架构如图4-1，覆盖从多源异构数据接入、流批协同计算处理，以及内嵌安全能力的数据服务层交付，最终赋能上层业务应用的全流程。

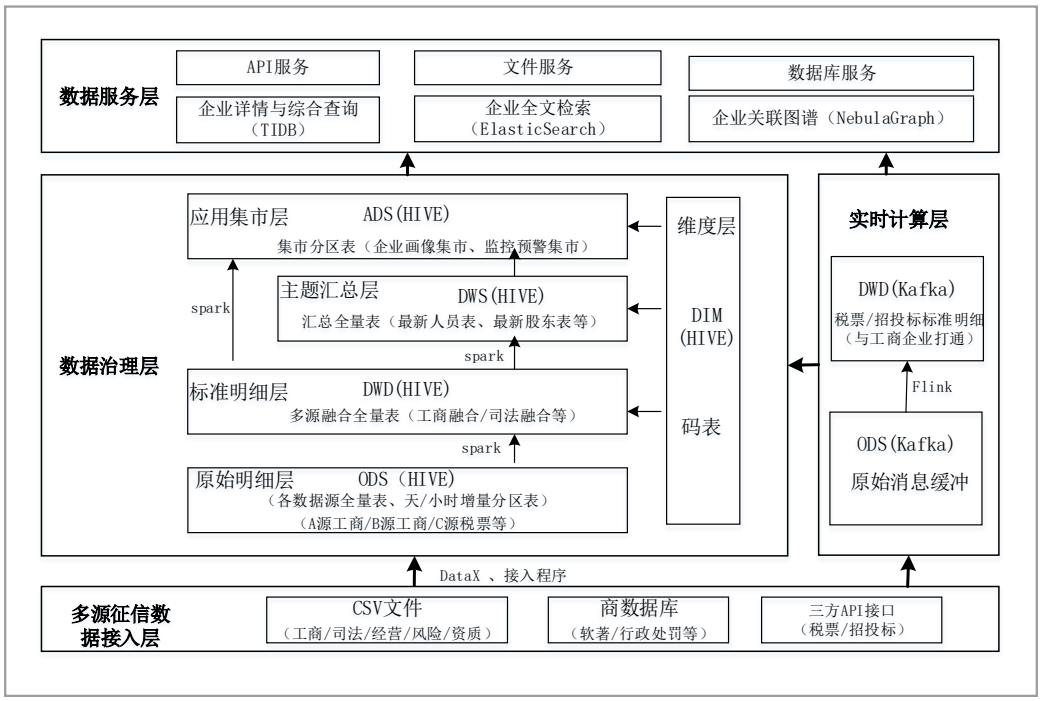

图4-1 企业数据治理体系技术架构

数据源与接入是整个体系的入口环节，融合了API实时推送、关系型数据库同步以及CSV文件批量导入等多类型异构数据源，通过统一接入工具完成数据规范化采集与标准化接入，保障数据入口环节的一致性与可控性。流批一体计算层则采用双链路协同运行模式，以满足差异化时效治理需求。其中，高时效计算链路面向API推送的分钟级高时效数据，采用Apache Kafka与Apache Flink构建流式处理通道，实现低延迟实时计算；准实时计算链路则面向数据库、CSV文件等小时级或天级的批量导入数据，采用Apache Spark构建离线处理通道，严格遵循ODS→DWD→DWS→ADS经典数仓分层架构，保障数据集加工过程规范、结果高质量。统一数据服务层采用多模态存储与统一接口相结合的设计方式，由TiDB支撑高并发点查与复杂分析业务、ES支撑全文检索需求、NebulaGraph支撑关联图谱分析场景，各类治理后的数据均通过标准化接口对外提供，从而保障数据服务能力与上层应用的解耦。在此基础上，新华财经、行业洞察、新华信用和数字经济等各应用通过调用统一数据服务，快速实现各类数据应用功能。

4.1 数据质量治理：多源融合引擎

4.1.1 基于优先级与规则的增量融合

算法

质量治理以准确性、完整性、一致性和唯一性为目标，采用基于优先级与规则的增量式多源融合算法，实现同一数据维度下多源增量数据向全量标准数据集的有序融合。算法流程包括以下五个核心步骤。

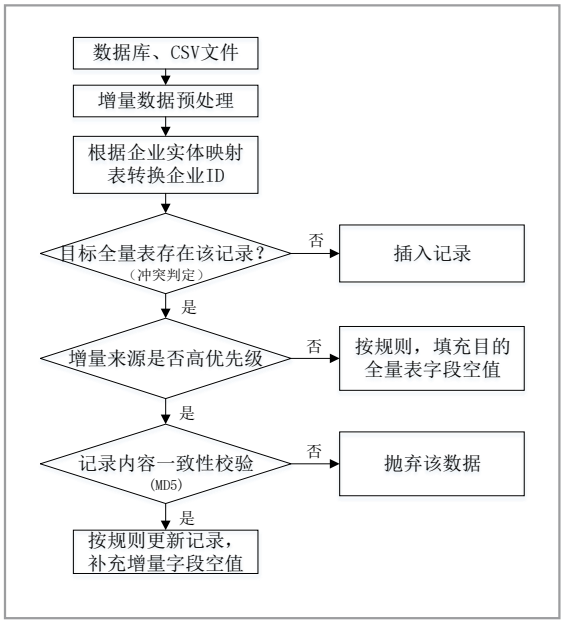

图 4–2　增量融合算法

（1）　实体解析与全局ID映射

以统一社会信用代码、企业名称、组织机构代码、工商注册号等增量企业基本信息为核心特征，与全量标准库中的现有记录进行匹配，建立全局企业唯一ID映射。若判定为同一企业主体，则建立增量企业ID与全量库企业ID的映射关系；若判定为新的企业实体，则为之生成新的全局唯一企业ID。通过这一环节，有效解决跨源主体不一致问题，为数据集标准化奠定基础。

（2）　跨源记录关联

依据企业ID映射表，将司法、招投标、经营等各类业务主题的增量数据关联至统一主体，将原始企业ID转换为全量标准库中的企业ID，完成跨源业务记录与统一企业主体的准确关联。

（3）　冲突检测与消解

基于数据源优先级与字段规则，分别执行记录级与字段级的精准更新判决。当增量业务主键关联到全量标准库中已存在的对应记录时，即触发冲突检测。在记录级更新判断层面，基于预定义的数据源优先级规则，若增量数据源的优先级不低于全量库数据源优先级，则允许更新目的全量库相应记录。在字段级更新判决层面，逐字段应用业务规则，规则库包含“非空值不覆盖”（增量空字段不更新目的全量库非空字段）、“人工校验锁定”（人工确认过的字段禁止自动更新）以及“高准确性字段保护”（置信度高的基准字段不被低置信度增量覆盖）等策略，实现字段粒度的精准治理。

（4）　双向空值智能补全

在冲突消解的同时，算法执行智能补全逻辑以最大化数据完整性。一方面，利用增量数据中的非空值补全基准库中对应的空值；另一方面，在规则允许范围内，用全量基准库中的信息反向补全增量数据中的空值，从而最大化提升数据集完整性。

（5）　新实体与记录的注入

未匹配记录作为新实体或者业务记录插入全量标准库，完成数据集的增量扩充。

4.1.2　企业工商信息融合案例

以企业工商信息的融合数据治理场景为例。假设全量基准来自高准确性数据源A，新增量来自高时效性数据源B，且优先级A > B。如下表1所示，融合后的企业状态和注册资本因时效性采纳数据源B的最新值；法定代表人方面，利用数据源B补全了数据源A的空值，提升了完整性；

经营范围则因数据源A权威性更高而得以保留。这一案例清晰展示了优先级判决与字段级更新策略在实践中的协同作用。

表4-2 企业工商信息多源融合示例

字段名称	全表融合前(源A)	增量信息(源B)	治理规则	全表融合后(源A&源B)
企业状态	在营	注销	时效优先	注销(数据源B)
法定代表人	空值	张**	权威补空	张** (数据源B)
注册资本	1000万元	2000万元	时效优先	2000万元 (数据源B)
经营范围	文本A	空值	互相补空	文本A(数据源A)
注册地址	地址A	地址B	非空即更新	地址B(数据源B)
统一社会信用代码	92	92	无	92
企业ID	110	001	实体对齐	000

4.2 大模型增强数据治理

在企业征信数据治理实践中，传统依赖人工预定义转换规则的治理模式，常因规则覆盖不全导致数据标准化转换失败，进而导致大量字段出现空值，严重影响数据集完整性。为破解该难题，本文将AI大语言模型作为智能增强插件，无缝嵌入现有自动化数据处理流程，以异步方式解析数据并动态生成新的映射规则，进一步提升数据标准化的准确率。

以“企业行政区划标准化”为例，AI与规则引擎的深度协同如图3-3所示。

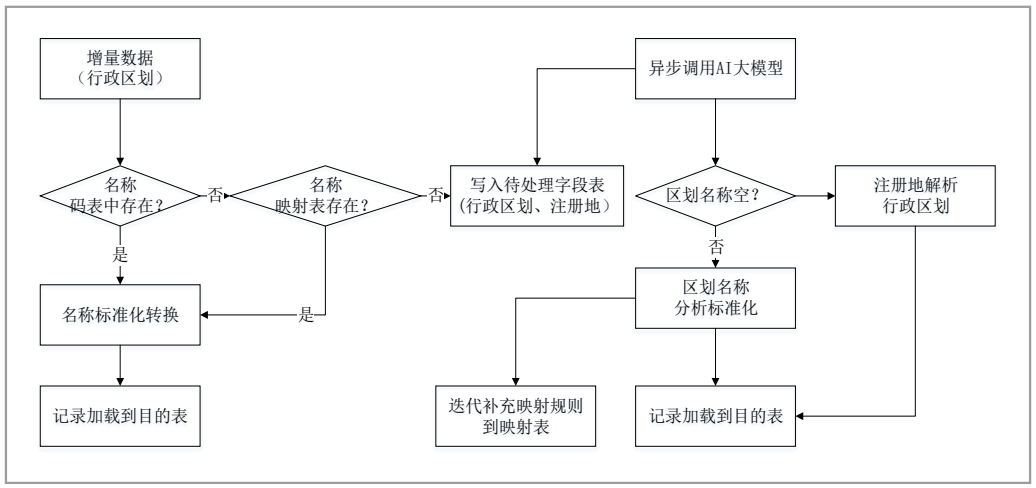

图4-3 AI增量数据标准化流程

（1） 增量数据触发与规则匹配

新的企业注册地址等增量数据触发流程，系统首先执行高效的确定性规则匹配：

● 一级匹配（码表匹配）：首先查询国家标准行政区划码表。若命中，则直接完成数据标准化转换并加载至目标表，此流程处理绝大部分规范数据。

● 二级匹配（映射表匹配）：若未能在国家标准行政区划码表命中，则查询“人工映射表”。该表存储了人工预定义映射规则，以及通过AI处理沉淀下来的映射规则，命中后可处理部分非规范化数据。

（2） 疑难数据分流与缓冲

若上述两级规则匹配均无法完成标准化，则将该数据判定为未知疑难样本。系统自动提取记录ID、字段名称与字段值等关键信息，写入待处理字段表（结构如表3-1所示），进入AI异步增强队列，避免阻塞其他处理流程。

表4-3 待处理字段表

表名	记录ID	字段名称	字段值	描述
企业基本信息表	1	行政区划名称	省市县JSON串	保存注册地址

（3） 嵌入式异步智能增强

AI增强插件周期性地从待处理字段表中批量提取任务。针对每条记录，流程自动判断并选择最优智能策略。若原始“省市县”信息完整，则异步调用AI大模型对不规范的名称直接进行“区划名称分析标准化”。若存在字段为空，则利用大模型的上下文理解与空间语义推理能力，自动解析出省、市、区县级标准行政区划。

（4） 结果回填与知识反哺

经AI增强处理后的结果自动回填至目的数据表，同步生成非标准表述与标准编码的映射规则，存入智能映射表供后续使用。由此形成“规则处理—智能增强 —知识沉淀—持续优化”的自治治理闭环：历史疑难样本经AI处理后转化为确定性规则，从而使治理体系具备了持续进化能力，不断提升自动化处理覆盖率与准确率。

4.3 数据时效性治理：Kappa混合架构

针对企业征信数据多源异构、更新频率不同的特点，本文设计了一种高时效与准实时双路径协同的kappa混合架构，对于秒级/分钟级数据，通过实时链路处理实现分钟级延迟；对于批量数据，则通过准实时链路处理实现小时级可见，最终在数据服务层达成高质量数据集状态的统一与高效查询。

4.3.1 高时效链路处理流程

面向API推送的高时效数据，端到端延迟控制在秒级/分钟级，可支撑税票类、招投标类等商机洞察与风险管控需求。主要流程如下：

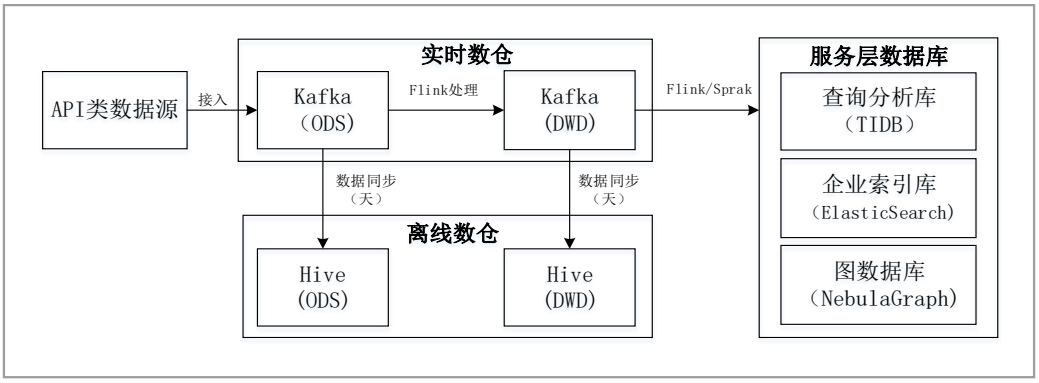

图 4-4 高时效链路处理流程

（1）实时接入：对 API 接口推送的数据进行实时采集，接入 Apache Kafka 消息队列，利用消息队列完成数据缓冲与上下游系统解耦。

（2）流式处理：基于流式计算框架搭建实时数仓分层结构，依次完成数据清洗、格式标准化、企业主体统一映射与关联打通。

（3）实时 ODS：对原始数据格式进行解析，按标准结构完成结构化存储，完整保留数据原始字段与形态，不做业务加工。

（4）实时 DWD：在流式链路中开展实时清洗与标准化转换，采用维度建模方法，将数据加工为可直接使用的业务明细数据。

（5）实时加载：将处理完成的结果数据写入 TiDB 与 NebulaGraph 存储引擎，支撑秒级指标查询与计算。

（6）异步离线同步：通过定时离线任务，把实时处理结果同步至离线数仓对应增量分区，并执行全量数据合并操作，确保全量数据集的完整性与口径一致。

4.3.2 准实时链路处理流程

面向数据库及 CSV 文件数据源，遵循“批量接入、增量处理、优先发布”的原则，在保证数据质量与处理经济性的前提下，实现小时级数据可见性。具体流程如下：

（1）统一批量接入与原始存储：通过 DataX、CDC 工具，将来自数据库或 CSV 文件的原始数据同步到 Hive ODS 层，按小时或天分区进行组织，形成原始的、未加工的“全量原始库”。

（2）增量提取与多源融合治理：借助 Spark 作业，从 Hive ODS 层分区表中提取增量数据，经过字段映射、转换、校验后存入 Hive TMP 层。对于同一维度的多源表，按数据源优先级进行串行处理，与 Hive DWD 层中的“全量标准库”进行比对融合，对于单源表完成标准化后，异步与全量表合并。

（3）增量发布与敏捷供给：从单源增量表、多源融合全量表获取增量数据，去除敏感字段后发布到 HIVE ADS 层增量表。依次按小时和天和分区存储，以满足不同应用的时效性要求。

（4）数据销毁：定期清理临时数据，优化存储。

4.4 安全与服务治理：高质量数据集安全合规交付

构建统一的数据服务层，并将安全治

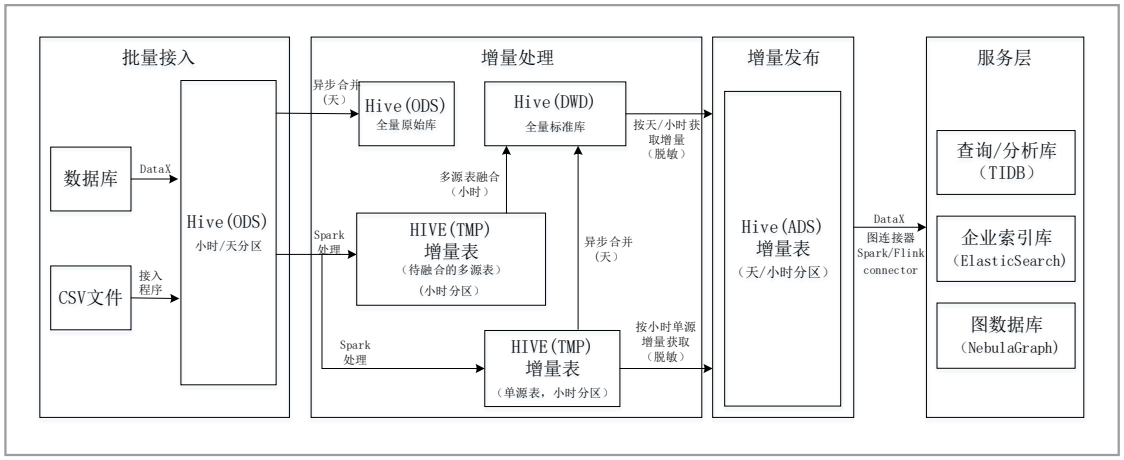

图 4-5 准实时链路处理流程

理能力深度内嵌，是保障高质量数据集安全合规交付的关键环节。

（1）高性能混合存储底座：采用 Elasticsearch + TiDB + NebulaGraph 异构混合存储架构，实现多场景查询能力协同。其中，Elasticsearch 支撑全文检索与模糊匹配场景，TiDB 支撑高并发点查询与复杂 OLAP 分析，NebulaGraph 支撑企业关联、股权穿透等图谱类查询。三类存储按需分工、协同调度，可满足征信数据多元化、高并发、低延迟的消费需求；

（2）服务化封装与内嵌安全治理：构建多层次服务接口与一体化安全管控相融合的统一数据服务能力。服务封装方面，基于领域特定语言 DSL（Domain Specific Language）实现参数化模板查询，支持多表关联与多维分析，具备良好的扩展性与复用性，同时对外提供标准化 API 以满足上层应用的常规查询需求。安全治理方面，通过 API 网关集成全链路安全治理能力，结合数据分级分类策略实现接口与字段级细粒度访问控制，对企业联系方式等敏感信息实施动态脱敏，全量记录数据访问日志以支持审计追溯，并通过限流、熔断与降级机制实现流量管控，保障服务稳定运行。

（3）服务与治理闭环优化：建立“治理–处理–服务–优化”的闭环机制。将数据服务层的调用日志、访问热点、性能指标与质量反馈回流至治理层，用于数据质量评估、热点需求识别及规则有效性验证，进而迭代优化多源融合策略、时效处理逻辑与安全管控规则，持续提升数据集质量与服务效能。

5 实施成效评估

本治理体系在中国经济信息社企业征信数据平台完成部署并稳定运行，通过对数据时效性、数据质量、安全与服务能力等核心指标的持续监测与对比分析，取得了显著的工程实践成效。

5.1 实验设置与数据集

实验数据来源于中经社企业征信数据平台的实际生产环境。平台每日处理数据记录数超过5000万条，覆盖工商注册、司法诉讼、经营行为、风险事件、资质认证、

知识产权等100余个核心维度，数据源超过10个，涵盖API实时推送、数据库同步和文件批量导入三类接入方式。实验对比基线为原平台采用的T+1离线批处理模式，评估指标包括端到端处理延迟、数据准确性、完整性、和一致性等。

5.2 时效性评估

表5-1呈现了本文设计的实时治理体系对不同类型数据源在时效性方面的优化效果。对于API实时推送的税务、招投标等高频动态数据，经过高时效流式处理链路优化后，端到端处理时延从原先12小时以上缩短至10分钟以内，达到了准实时更新标准；对于数据库同步类的商标公告、软件著作权等信息，采用准实时处理链路后，数据可见性由传统T+1模式的次日更新，提升至小时级可见，平均时延控制在3小时以内；对于批量文件导入的工商、司法等周期性数据，经增量批处理优化后，处理时延同样由T+1降至3小时以内。

实验结果表明，基于Kappa混合架构的分级时效处理策略，可有效适配不同数据源的更新频率与业务特征，在兼顾计算资源经济性与治理稳定性的前提下，实现企业征信数据集时效性的全面、显著提升。

表5-1 时效性对比

数据源类型	典型数据示例	原有延迟	新治理体系延迟	提升效果
API实时推送	税务、招投标动态信息	> 12小时	≤10分钟	延迟从天级/小时级压缩至分钟级
数据库同步	商标公告、软件著作权	T+1次日可见	<3小时	延迟从天级压缩至小时级
批量文件导入	工商变更、司法公告	T+1	<3小时	延迟从天级压缩至小时级

5.3 数据质量评估

在数据集质量方面，经多源融合引擎与AI大模型智能治理后，数据在完整性、准确性、一致性、三项核心指标上均得到显著提升。首先，关键字段空值率显著下降，基于优先级判决的增量式多源融合引擎充分发挥多源数据互补优势，对统一社会信用代码、税号、经济类型等核心字段实施双向智能补全，有效解决单一数据源信息缺失问题，例如：税号空值率由治理前的47.37%，下降到11.51%，提升了35.86%。其次，数据准确性实现跨越式提升，针对行政区划、注册地址等存在大量不规范表述的字段，AI大模型异步增强与规则沉淀机制有效完成语义归一与标准化校正，大幅提升字段规范性与正确性，以重庆市变更最新行政区划为例，准确性由原来的81.8%，提高到100%。再次，数据覆盖范围有效扩展，通过引入差异化数据源，系统补全了事业单位、社会团体等非企业主体信息，显著提升数据集覆盖广度与业务适用范围，非企企主体记录完整性提升了16.4%。最后，主体一致性得到强化，依托多源冲突检测与消解机制，系统可精准识别并归一化同一企业主体，确保跨源数据主体唯一、信息权威、口径统一，统一社会信用代码全局唯一且一致，一致性性达到100%。具体质量提升对比如表

5-2所示。

表5-2 数据质量对比

评估维度	核心指标	核心字段与测试范围	治理前(单源)	治理后(多源融合+AI增强)	提升效果
数据完整性	字段空值率	统一社会信用代码	13.30%	12.29%	空值率下降1.01%
		税号	47.37%	11.51%	空值率下降35.86%
		经济类型	0.34%	0.17%	空值率下降0.17%
		经纬度	37.44%	31.03%	空值率下降6.41%
	记录完整性	事业单位等非企业数据	5610981个非企业主体	6530981个非企业主体	完整性提升16.4%
		裁判文书	73380371条	162447684条	完整性提升121%
		股东冻结	1448715条	2614027条	完整性提升80%
		失信信息	5159025条	5275191条	完整性提升2.25%
		终本案件	9723219条	9534253条	完整性提升1.98%
		行政处罚	4758126条	7313300条	完整性提升53.7%
数据准确性	字段准确性	企业基本信息表的行政区划 重庆市企业(抽样258万条)	81.80%	100.00%	准确率提升18.2%
		行政处罚文书号准确率	7.4%	100%	92.6%
数据一致性	主体信息一致性	终本案件统一社会信用代码	96.30%	100.00%	一致性提升3.7%
	吊销时间一致性	企业基本信吊销时间与注销吊销表的吊销时间	82.6%	100%	一致性提升17.4%

5.4 安全与服务评估

在安全管控方面，本治理体系实现了数据服务接口的统一认证、鉴权与审计，覆盖率达到100%。同时，通过字段级访问控制和动态脱敏策略，企业联系方式、股东身份信息等敏感信息的安全防护能力显著增强。在服务效能方面，基于标准化API的数据需求交付周期缩短50%以上，数据服务平台日均API调用量超过10万次，平均响应时间保持在300ms以内，系统可用性达到99.95%。

6 结论与展望

本文以高质量企业征信数据集构建为目标，提出并实施了一种融合数据质量、时效性、安全服务与AI赋能的全链路实时数据治理框架。该框架依靠多源融合技术确保数据集的准确性，基于Kappa混合架构以维护数据的时效性，通过内嵌安全管控实现数据全生命周期的合规性，同时借助AI大模型增强治理的智能水平，系统性地解决了在高时效业务场景中，高质量数据集的构建与供给的核心挑战。

实践结果表明，所提出的框架能够稳定地构建高可信、高时效、高安全的企业征信数据集，该数据集覆盖100+核心维度，数据处理周期从T+1压缩至3小时以内，关键数据准确率达99%以上，可为金融征信、行业监管、商业决策等典型场景提供坚实的数据基础与技术支撑。

后续研究将聚焦两个方向：一是更深度地融合大模型，让数据质量异常的自动诊断、修复以及治理规则的生成更加自动化，减少人工干预；二是推动架构向流批一体的数据湖仓升级，引入新的流处理框架和数据湖技术，进一步简化治理链路、提升处理效率。同时，我们也会继续扩展数据集的维度和覆盖范围，为数据要素的合规高效流通提供更好的基础。

参考文献：

[1] Guerreiro L, Martins J, Rosário Bernardo M, et al. Impact of a master data management framework to trigger data governance maturity: a systematic literature review[J]. IEEE Access, 2025, 13: 190644-190667.

[2] 混合数据处理架构：对比与演进[J]. 计算机工程与应用, 2025, 61(7): 1-12.

[3] 赵晓娟, 贾焰, 李爱平, 等. 多源知识融合技术研究综述[J]. 云南大学学报(自然科学版), 2020, 42(3): 459-473.

[4] 张伟, 李华, 王明. 基于自适应融合的多模态实体对齐方法[J]. 计算机学报, 2024, 47(8): 1823-1838.

[5] Wu Y Y, Yang C, Zhu M Y, et al. A Zero-Training Error Correction System with Large Language Models[C]//Proceedings of the 41st IEEE International Conference on Data Engineering (ICDE 2025). IEEE, 2025: 2949-2962.

[6] Yan M Y, Wang Y S, Jiang X H, et al. Towards uncertainty-calibrated structural data enrichment with large language model for few-shot entity resolution[J]. Frontiers of Computer Science, 2025, 19(11): 1911376.

[7] 王晓东, 徐梓原, 段志飞, 等. 数字经济视域下征信数据质量监管研究[J]. 征信, 2025, 43(2): 48-55.

[8] 张晶. 数字经济视域下征信数据治理的趋势与机制[J]. 征信, 2023, 41(2): 35-39.

[9] 林钧跃. 我国企业征信业的发展方向及路径研究[J]. 征信, 2025, 43(11): 1-8.

[10] Hassani H, Huang X, MacFeely S. Mapping the evolution of data governance scientific research[J]. Data & Policy, 2025, 7: e51.

[11] 韩璐, et al. Integrated Multivariate Segmentation Tree for Heterogeneous Credit Data Analysis in Small- and Medium-Sized Enterprises[J]. Expert Systems With Applications, 2025. DOI: 10.1016/j. eswa. 2025.130389.

[12] 王丽, 郑伟. 跨模态文本图像融合的实体对齐技术[J]. 计算机研究与发展, 2025, 62(3): 601-615.

[13] 刘洋, 孙强. 融合多级特征的图神经网络实体对齐模型[J]. 软件学报, 2025, 36(2): 512-528.

[14] 陈思, 赵磊. 利用图结构信息的企业实体对齐算法研究[J]. 中文信息学报, 2024, 38(5): 89-98.

[15] 李泽霖, 等. 基于多跳信息融合的实体对齐模

型[J]. 计算机工程, 2024, 50(9): 142-152.
[16] 周涛, 李勇. 基于Flink的流批一体实时数据仓库设计与优化[J]. 大数据, 2024, 10(4): 78-92.
[17] 国家互联网信息办公室.《网络数据安全管理条例》解读与合规实践[J]. 网络安全与数据治理, 2025, 44(2): 1-8.
[18] 黄科满, 许多, 杜小勇. 数据社会化视角下的数据流通安全治理: 五位一体框架[J]. 大数据, 2024, 10(6): 5-15.
[19] 李晖, 朱辉, 张玉清. 生成式人工智能数据安全风险与治理综述[J]. 信息安全学报, 2025, 10(1): 1-18.
[20] 陈钟, 刘哲理, 李晖. 数据安全治理:框架与实践[J]. 信息安全研究, 2020, 6(9): 771-777.
[21] 刘越, 张敏. 面向隐私保护的数据集构建与大模型推理评估[J]. 通信学报, 2025, 46(4): 112-126.
[22] 赵鑫, 孙茂松. 大语言模型在数据治理中的应用综述[J]. 中文信息学报, 2025, 39(2): 1-15.
[23] 周志华, 陈松灿. 基于机器学习的异常检测与根因分析[J]. 计算机科学, 2025, 52(1): 1-10.

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: