大规模异构数据迁移的自适应清洗与智能转换框架*

许文静，安　宁，于　重，刘珠慧
（国务院国有资产监督管理委员会干部教育培训中心，北京　100053）

摘　要：在数字化转型背景下，传统集中式数据库向分布式架构迁移面临异构数据模型语义冲突、业务连续性要求、人工转换低效等核心挑战。提出智能转换框架 AUTO-MIG，其核心创新在于深度挖掘数据内在关联的智能决策机制与适应大规模异构环境的高效执行引擎。AUTO-MIG 创新性地利用图神经网络自动发现隐含于数据库模式中的复杂表间关联，并结合多目标优化模型智能决策最优存储方案，提升跨模型转换的自动化程度。同时，框架设计独特的双模式日志捕获与流批协同清洗管道，实现对海量历史数据与高频实时变更数据的低延迟、高可靠同步与清洗。该框架成功实现了在容器化平台上的部署并以大规模教育培训系统数据迁移为典型应用案例实践验证。结果表明其图神经网络驱动的关联发现显著提升了复杂查询性能，而双模式协同执行引擎则大幅缩短了迁移总耗时并优化了资源利用效率，为企业数字化转型提供了可靠的技术支撑和实践路径。

关键词：异构数据；数据迁移；智能转换框架；元数据感知；图神经网络
中图分类号：TP39　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2025.09.006
引用格式：许文静，安宁，于重，等. 大规模异构数据迁移的自适应清洗与智能转换框架［J］. 网络安全与数据治理，2025，44(9)：35-45.

Adaptive cleaning and intelligent transformation framework for large-scale heterogeneous data migration

Xu Wenjing，An Ning，Yu Zhong，Liu Zhuhui
（SASAC Education and Training Center，Beijing 100053，China）

Abstract：In the context of digital transformation，migrating from traditional centralized databases to distributed architectures presents core challenges including semantic conflicts in heterogeneous data models，business continuity requirements，and inefficient manual conversion processes. This paper proposes an intelligent transformation framework named AUTO-MIG，whose core innovations lie in an intelligent decision-making mechanism that deeply mines intrinsic data relationships and a high-performance execution engine adapted to large-scale heterogeneous environments. AUTO-MIG innovatively employs graph neural networks（GNN）to automatically uncover complex inter-table relationships embedded within database schemas and combines a multi-objective optimization model to intelligently determine the optimal storage strategy，thereby enhancing the automation of cross-model data transformation. Furthermore，the framework incorporates a uniquely designed dual-mode log capture mechanism and a stream-batch hybrid cleaning pipeline to achieve low-latency，highly reliable synchronization and cleaning of massive historical data and high-frequency real-time changes. The framework has been successfully deployed on containerized platforms and validated through a large-scale educational training system data migration case. The results demonstrate that the GNN-driven relationship discovery significantly improves complex query performance，while the dual-mode collaborative execution engine considerably reduces total migration time and optimizes resource utilization efficiency. This provides reliable technical support and a practical pathway for enterprise digital transformation.
Key words：heterogeneous data；data migration；intelligent transformation framework；metadata awareness；graph neural network

*基金项目：国务院国资委干部教育培训中心（中央党校国务院国资委分校）科研项目（25GZW0313）

0 引言

随着数字化转型进程的加速推进，企业信息系统正经历从传统集中式架构向分布式架构转型，传统集中式数据库系统正逐渐被新型混合存储架构所替代[1]。

新旧系统数据迁移工作面临规模性、异构性、时效性三个方面技术挑战[2]。规模性挑战体现在海量历史数据的迁移需求上。传统迁移方法需要较长停机时间，导致无法满足业务系统高可用性的要求。异构性挑战体现在不同数据库系统在数据模型和查询语义等方面的差异。这种差异导致自动化迁移过程中出现各种兼容性问题，特别是在业务逻辑转换方面。时效性挑战体现在迁移过程中的数据一致性保障。由于缺乏有效的增量同步机制会导致业务状态不一致，直接影响用户体验和系统可靠性。这些挑战共同形成数据迁移工作的主要难点是在有限的时间资源下，难以同时保证迁移效率、数据一致性和业务连续性。此外，现有解决方案在异构模型转换和智能化能力方面也存在明显不足，导致成本居高不下。

基于规则的数据转换方法、增量数据同步技术以及分布式事务管理方案为现有研究工作的主要技术方向。虽然这些方法在特定场景下取得了一定成效，但普遍存在明显局限。基于规则的方法需要大量人工干预，难以应对复杂的模型转换需求。基于语义映射的方法虽然提高了转换精度，但面临可扩展性问题。虽然机器学习方法为数据转换提供新的思路，但在实际应用中仍存在训练数据需求大、业务规则处理能力弱等缺陷[3]。

针对异构性、规模性和时效性三大核心挑战，本文提出智能转换框架 AUTO-MIG。该框架的核心创新包括两方面：一是基于图神经网络（Graph Neural Network，GNN）的深度关联发现机制，可自动识别数据库中未明确定义的复杂表间关联，减少对人工规则的依赖，为跨模型映射提供支持；二是面向大规模异构迁移的双模式协同执行引擎，结合全量数据分块并行处理与增量日志流式捕获，在保障一致性的同时提升吞吐量、降低迁移时间。AUTO-MIG 通过元数据驱动的动态适配、自解释模式转换与分布式执行策略等技术实现上述机制。为验证其有效性，本文选取具有海量历史数据、高频更新、复杂网状关联和强领域规则的大规模教育培训系统进行迁移测试，该场景能够充分体现框架的普适性与智能性。

1 相关技术综述

1.1 数据集成框架

数据集成技术的发展经历从静态规则到动态适配的发展过程。早期 ETL（Extract-Transform-Load）工具采用预定义映射模板实现数据转换，其核心缺陷在于无法自动识别不同数据库系统的语法特性。随着分布式架构的普及，新一代框架[4]通过模块化提升扩展性，允许开发者自定义连接器组件。然而，这些方案仍存在显著不足：首先，模式映射过程需要人工定义字段对应关系，当面对包含数百张表的复杂系统时，配置工作量呈指数级增长；其次，基于批处理的架构导致增量数据同步延迟过高，难以满足实时性要求。流式集成方案[5]虽然通过变更数据捕获技术提升时效性，但在智能路由和跨模型存储优化方面仍存在明显短板。现有数据框架在动态语法解析能力、自动化约束迁移机制以及混合负载下的资源调度策略等方面待突破。

1.2 数据清洗技术

数据清洗作为迁移过程中的关键环节，其技术发展呈现出从规则驱动到智能化的演进趋势。传统方法主要依赖正则表达式和预定义规则集进行数据修复，虽然在小规模场景下有效，但面对大规模数据时存在严重的性能瓶颈[6]。分布式计算框架的引入提升了处理能力，但其核心算法（如局部敏感哈希）仍局限于语法层面的去重和标准化，无法解决业务逻辑冲突等深层次问题。近年来，机器学习技术开始应用于该领域。监督学习方法[7]在结构化数据清洗中展现出潜力，但其依赖大量标注数据的特性制约了实用性。无监督方法[8]虽然降低对标注的依赖，但存在误判率高、解释性差等缺陷。当前研究的难点在于如何在保证清洗精度的同时，实现算法在分布式环境下的高效执行。

1.3 跨模型转换

关系型数据库向文档型数据库的转换涉及深层的模型冲突问题。从理论角度看，这种转换本质上是将二维表结构映射为嵌套文档的过程，需要解决主外键关系的表达、事务一致性的保证以及查询模式的适配[9]三个核心问题。研究者提出了一种数据转换框架，该框架允许用户声明新数据的特征，并能够将数据转换为所需的格式[10]。他们使用 JSON 作为中间数据类型实现了 MySQL 到 MongoDB 的转换。此外，工业界的实践也呈现出多元化特征。一部分方案采用完全反范式化的深度嵌套结构，虽能提升读取性能，却显著增加了更新操作的复杂度；另一些方案则选择保留引用关系，通过在应用层执行 JOIN 操作来维持数据一致性。然而，这些方法存在缺乏对混合工作负载的统一优化和未能充分适配目标数据库固有特性两大局限。

1.4 数据质量

早期的数据质量保障技术基于规则的方案（如正则表达式验证、约束检查）。虽然解释性强，但面临规则维护成本高、适应性差的局限[11]。随着技术进步，机器学

习方法（包括基于统计的异常检测、生成对抗网络的缺失值填补等）显著提升了自动化程度，将人工干预降低，但仍受限于标注数据需求大和业务规则理解不足的问题。一些研究通过半监督学习框架结合少量标注数据，以及构建业务规则知识图谱实现语义级验证，在提升准确率的同时降低标注成本。然而，该领域仍面临复杂业务规则的数字化表达、低质量数据下的模型鲁棒性，以及实时流数据场景的效率优化等挑战[12]。

2 智能转换架构

2.1 总体架构设计思想

针对新旧系统迁移中的异构性、规模性与时效性挑战，本文提出智能转换框架 AUTO-MIG，其设计遵循三大核心原则：一是动态适配原则，通过元数据驱动自动识别源系统特征，消除人工规则配置；二是分层解耦原则，将迁移流程拆分为感知、决策、执行三层，支持独立扩展；三是增量协同原则，历史数据批量处理与增量变更流式同步无缝衔接。架构图如图 1 所示。

2.2 元数据感知层

2.2.1 元数据智能感知器

元数据智能感知器通过动态语法适配实现异构数据库集成。基于语法解析工具构建包含多种结构化查询语言的语法规则库，结合词法解析器与抽象语法树生成器，自动化解析数据库元数据。调用连接接口获取数据库引擎版本、保留字集等特征参数，输入至加权规则引擎进行判定。首层识别方言专属语法标记，次层对比系统函数差异，终层融合元数据模式相似度加权评分。此外，还支持规则库热更新，新增方言时仅需扩展语法定义文件，无需重构核心解析逻辑，提升异构环境适应效率。

业务逻辑的数字化表达依赖约束关系深度提取。通过扫描表结构元数据捕获显式主键与外键声明，同时通过静态代码分析解析存储过程、触发器等对象中的隐式关联（如动态查询生成的临时约束）。构建约束关系图谱以数据表为节点，包含字段类型、索引等属性；以约束关系为边，标注关系类型及级联操作规则。针对无显式外键的数据库，采用列值分布相似性分析推导潜在关联。

2.2.2 增量捕获引擎

为解决全量与增量异构数据环境下低延迟、高可靠的数据同步问题，采用双模式日志采集机制建立增量捕

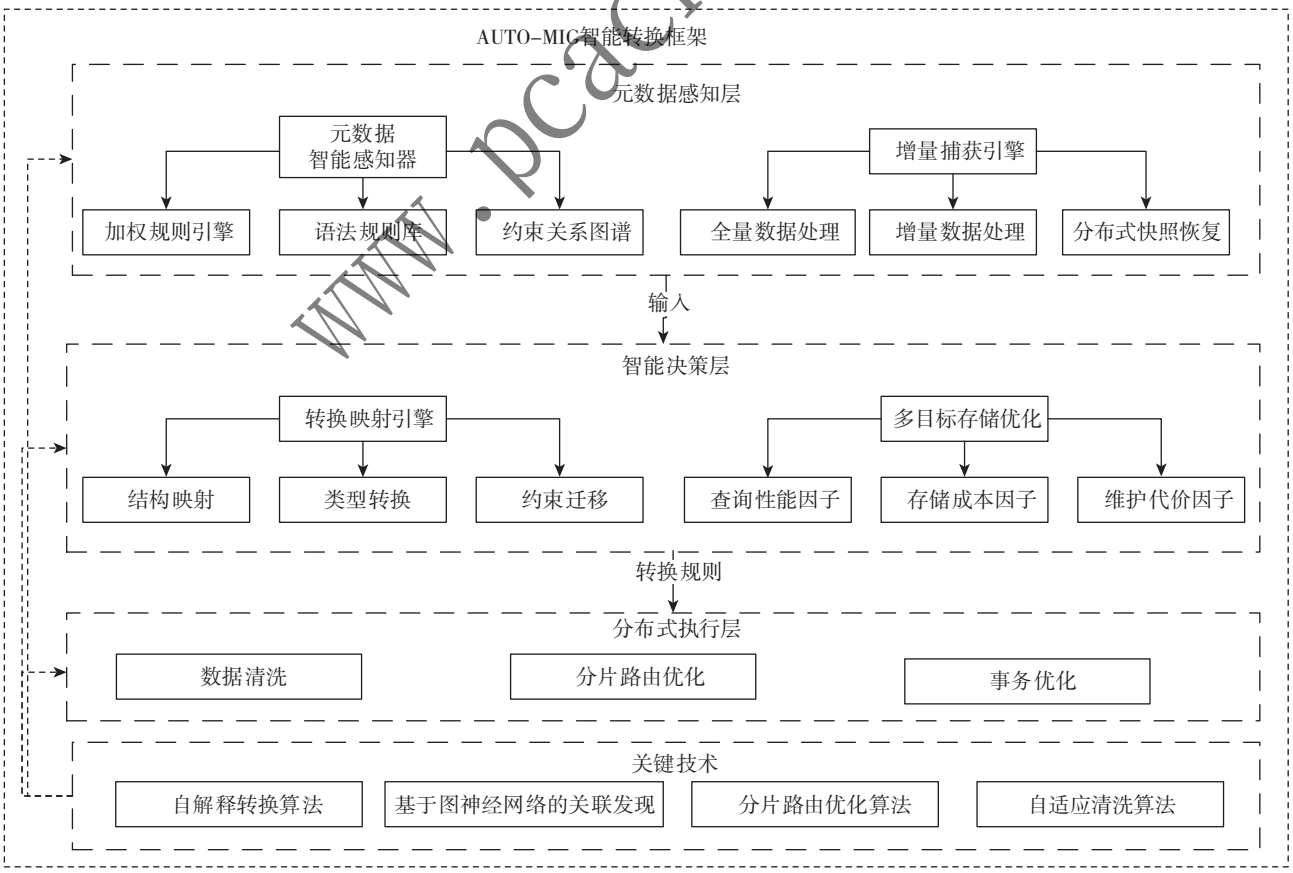

图 1 智能转换架构图

获引擎。针对全量快照数据场景，基于主键区间或哈希分区将表数据拆分为数据分块，通过多线程并发扫描。采用批量获取、传输压缩及磁盘顺序写入优化，确保太字节级数据采集时延低。针对增量数据场景，依托事务日志监听框架捕获数据库变更事件，经由消息队列实现异步流式分发。通过微批处理窗口平衡吞吐与延迟，结合主键分区路由保障事件顺序性，端到端延迟稳定控制在 1 s 内。

为解决断点续传与状态一致性问题，基于分布式快照算法实现故障恢复。在消息消费者组中设立检查点，定期持久化分区偏移量及模式版本至协调服务。系统重启后从最近一致性快照重放日志，依托全局事务标识符实现重复操作过滤。针对分布式事务场景，引入两阶段提交协议保障跨分区操作的原子性，杜绝部分失败导致的状态不一致。

2.3　智能决策层

2.3.1　转换映射引擎

模型转换映射引擎通过结构映射、类型转换、约束迁移三层抽象实现异构数据模型的无损转换。第一层结构映射。基于改进的命名相似度算法自动化匹配表或字段，支持跨数据库重命名，当相似度阈值大于 0.85 时触发自动映射。第二层类型转换。构建动态类型兼容矩阵库，涵盖多种数据类型转换规则。第三层约束迁移。将主键与唯一约束转换为目标数据库的唯一索引机制；将外键约束拆解为应用层完整性校验微服务，解决非关系型数据库原生约束缺失问题。

2.3.2　多目标驱动的存储优化决策

多目标驱动的存储优化决策围绕查询性能、存储成本、维护代价三维度构建加权效用函数。多目标决策模型为：

$$\max U(S) = \alpha \cdot Q_{\text{perf}} + \beta \cdot C_{\text{storage}} + \gamma \cdot M_{\text{maintain}} \tag{1}$$

查询性能因子采用非线性响应函数量化。当查询延迟低于 100 ms 临界值时维持高评分（$\geqslant 0.9$），超过后每增加 50 ms 评分指数衰减 40%。查询性能因子计算公式为：

$$Q_{\text{perf}} = \frac{1}{1 + e^{-k(T_{\text{latency}} - T_0)}} \tag{2}$$

其中，T_{latency} 表示实际查询延迟，即系统处理查询所需时间；T_0 为延迟阈值（临界值），当 $T_{\text{latency}} = T_0$ 时，性能因子 $Q_{\text{perf}} = 0.5$；k 为斜率系数（敏感度参数），控制函数在 T_0 附近的陡峭程度。

存储成本因子通过压缩效益对数函数计算，以原始数据量与压缩后数据量比值为输入。存储成本因子计算公式为：

$$C_{\text{storage}} = \log\left(\frac{\text{Size}_{\text{raw}}}{\text{Size}_{\text{compressed}}}\right) \tag{3}$$

其中，Size_{raw} 为原始数据大小，$\text{Size}_{\text{compressed}}$ 为压缩后数据大小。

维护代价因子则加权求和历史模式变更操作，形成综合运维成本预期，突破传统单目标优化局限，实现多约束条件的协同量化。维护代价因子计算公式为：

$$M_{\text{maintain}} = \sum w_i \cdot \Delta_{\text{schema}} \tag{4}$$

其中，Δ_{schema} 为模式变更的复杂度，量化单次变更的复杂程度（例：新增字段 = 1，重构分区 = 5）；w_i 为权重系数，表示不同变更类型的成本权重（例：索引变更权重 = 1.5，字段类型变更权重 = 0.8）。

三因子权重系数（α、β、γ）通过信息熵权法实现场景自适应。首先归一化各因子原始值（性能延迟、压缩比、变更频次）；其次依据数值离散程度计算信息熵（离散度越高则熵值越大）；最终基于熵值差异性生成动态权重。

决策流程如图 2 所示，输入表结构特征、查询模式及成本约束，通过分层判定输出最优存储方案。验证阶段采用三重保障：一是帕累托前沿分析筛选效用空间中的非支配解集；二是代价敏感性测试验证鲁棒性；三是基准测试对比综合效用。关系型特征强度综合外键依赖数量、范式化程度等多项指标评估；嵌套层级深度统计半结构化字段的嵌套层数；混合存储策略高频访问字段存储于关系型数据库，大文本及二进制数据存储于文档数据库。

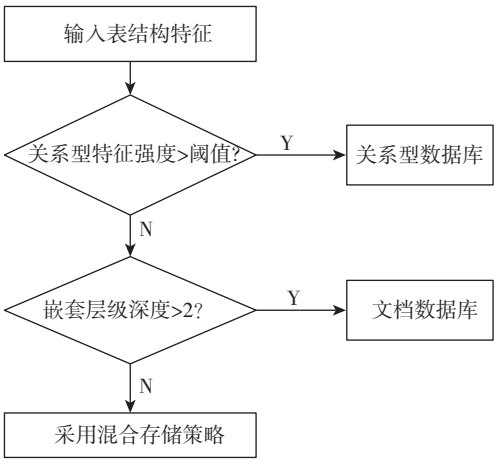

图 2　决策流程图

通过构建量化模型综合权衡查询性能、存储成本与维护代价三维目标，突破传统单维优化局限；采用信息熵权法动态计算各因子权重系数，依据数据离散度自主生成最优权重分配方案，实现多约束条件下的智能决策闭环。

2.4 分布式执行层

2.4.1 数据清洗

针对实时数据处理场景中面临的时效性与质量管控挑战，建立智能流批协同处理方法。批处理阶段通过分布式计算引擎执行多层级清洗。在语法层面运用正则表达式统一数据格式并实施差异化空值填充策略（数值型采用均值填补、类别型采用高频值填补）；在逻辑层面通过分布式连接验证数据关联完整性，识别并处理未关联主键的异常记录，同时对超出合理范围的数值依据统计学原则进行修正。流处理阶段基于实时计算引擎构建高效处理流水线，从消息队列持续获取增量数据后，通过动态规则引擎执行业务逻辑校验，利用底层硬件指令级并行技术实现百万级吞吐量的质检处理，最终根据数据访问特征智能分发至不同类型的存储系统。通过元数据感知机制自动调度处理模式，在保证 85% 以上场景实现秒级实时处理的同时，智能切换至批处理模式应对数据积压或历史回溯需求，形成完整的自适应数据处理体系。

2.4.2 自适应分片路由与事务优化

针对分布式系统中数据分布不均和跨节点事务一致性的核心挑战，建立自适应分片路由与事务保障机制。在路由层面，采用智能分片管理策略，基于一致性哈希算法实现数据键值与分片的基础映射，并实时监控各分片负载状态。当检测到分片负载超过预设阈值时，自动启动负载均衡流程，通过动态检索低负载分片并执行智能重路由，确保集群资源的高效利用。引入时序预测模型，通过分析历史负载特征实现分片压力预判，提升热点规避的响应效率。

在事务处理层面，通过可靠的分段式事务管理协议，将复杂事务分解为可独立执行的原子子任务，每个子任务均配备对应的补偿操作。通过建立完善的逆向执行链和持久化日志机制，确保在部分子任务失败时能够有序回滚，维持系统的最终一致性。

2.5 协同智能机制

AUTO-MIG 框架的智能本质体现为三层自主决策闭环。一是感知层智能通过动态语法适配器实现元数据自感知，其多级加权判定引擎支持异构数据库的零配置解析，消除人工规则定义需求。二是决策层智能依托图神经网络的关联发现实现规则自学习，基于图卷积层提取表结构特征与注意力层量化关联权重，自主发现业务系统中的隐式关联逻辑。三是执行层智能借力分片路由优化达成资源自适配，通过负载监控、热点预测、收益评估和分片重组的响应链动态优化资源分配。智能闭环逻辑可总结为感知层消化异构环境，经决策层生成迁移知识，驱动执行层动态调优资源，其运行数据反馈至决策层驱动策略进化，最终形成自进化智能中枢。

3 关键技术实现

3.1 自解释转换算法

自解释 Schema 转换算法致力于解决异构数据库系统之间的数据类型兼容性问题。算法的核心在于建立智能化类型转换机制，能够在不同数据库管理系统之间自动完成数据类型的映射与转换。算法流程如图 3 所示。

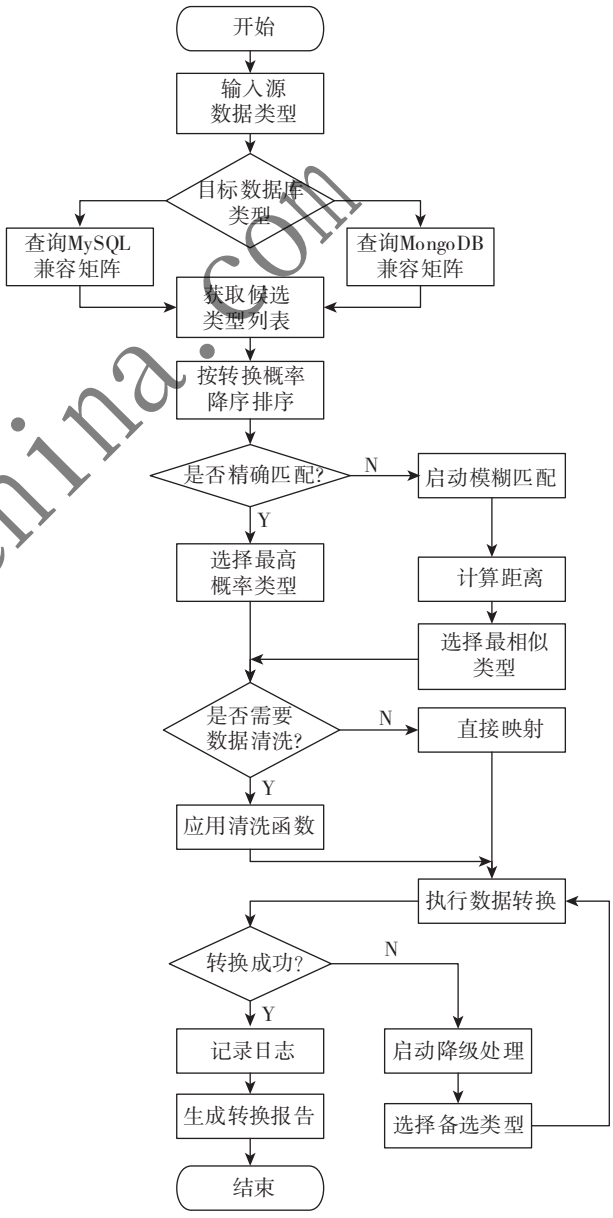

图 3 自解释 Schema 转换算法流程

算法采用三级处理架构来实现。构建一个动态更新的类型兼容矩阵，该矩阵记录各种源数据类型到目标数据类型的转换规则及对应的转换成功率。矩阵中的每个条目不仅包含类型映射关系，还存储转换函数的引用和

历史转换成功概率。当需要进行类型转换时，会优先选择成功概率最高的转换路径。为实现多层次转换决策流，先检查是否存在精确匹配转换规则，如果没有则启动模糊匹配机制，通过计算类型名称的相似度来寻找最接近转换方案。在转换过程中，算法会实时评估数据特征，例如数值范围、字符串长度等，以防止数据溢出或截断。转换成功后生成详细的转换日志，日志记录每个字段的转换决策过程，包括考虑过的备选方案、选择最终方案的理由以及转换过程中出现的任何警告信息。这种自解释特性大大提升可审计性和可维护性。

本算法通过历史成功率驱动的动态类型映射（基于实时更新的兼容矩阵优选高成功路径）、模糊匹配机制（计算语义相似度适配非标类型）及数据特征实时评估（扫描样本范围/长度防溢出）实现三重智能决策，降低人工干预并提升转换效率。

3.2 基于图神经网络的关联发现

数据库模式中隐含表间关联的自动发现，本质上是挖掘数据实体（表）之间未通过外键明确定义但存在潜在语义或结构依赖性的关系。传统方法主要依赖预定义的规则模板或基于统计相似性的启发式方法。这些方法存在局限：一是规则模板难以覆盖复杂多变的关联模式，且高度依赖领域专家的先验知识，难以泛化到新场景；二是基于统计相似性的方法通常只考虑表或列的局部特征，忽略数据库模式作为一个整体图结构的全局拓扑信息，难以捕捉表之间通过多跳路径或共享邻居形成的复杂、非直接的依赖关系；三是如何有效融合表的结构特征、统计特征和语义特征以形成全面且具有区分度的表表示。针对上述挑战，基于图神经网络的关联发现方法被提出。图神经网络是专门用于处理图结构数据的深度学习模型，其核心优势在于能够同时建模图数据中的节点属性（特征）和图结构（拓扑关系），使其成为解决数据库模式隐含关联发现问题的最佳选择。

3.2.1 算法思想

针对数据库系统中隐含表间关联的自动发现问题，本文提出一种基于深度学习的智能解决方法。该方法突破传统基于规则方法的局限性，通过图神经网络架构实现数据驱动的关联发现。具体而言，首先将数据库模式转化为属性图结构，其中节点对应数据表，边表示显式定义的外键关系[13]。在此基础上，通过多层次图神经网络模型来捕获表间复杂的关联模式。

为实现数据库隐含表间关联自动发现，本文设计了一个端到端的图学习框架。首先，采用图神经网络架构，结合图卷积网络和图注意力机制，既保留局部结构信息，又能动态学习不同关联的重要性权重。其次，引入多源特征融合机制，将表的结构特征（如字段数量、主键信息）、统计特征（数据类型分布）和语义特征（字段命名模式）有机整合，形成全面的表表示。最后，通过路径正则化约束将领域知识融入模型优化过程。框架采用多层消息传递架构来处理属性图。第一层网络负责提取局部特征，捕获单个节点的属性信息，如字段数据类型、约束条件等。第二层网络则专注于聚合邻域信息，通过分析节点之间的连接模式来推断潜在的关联关系。

训练过程采用半监督学习策略。利用已知的外键关系作为正样本，同时通过负采样技术生成负样本。损失函数设计不仅考虑分类准确性，还加入关联路径长度的惩罚项，以鼓励发现直接而紧密的关联关系。为了提高泛化能力，训练过程中还采用数据增强技术，主要包括随机删除部分已知边来模拟不完整的信息，以及添加噪声节点来增强模型的鲁棒性。

3.2.2 算法模型

给定数据库模式 $S=T_1, T_2, \cdots, T_n$，其中每个表 T_i 包含列集合 $C_i=c_{i1}, c_{i2}, \cdots, c_{im}$，元数据 $\mathcal{M}_i$（主键、索引、约束等）。构建属性图 $G=(V, E)$，其中顶点集 V 表示每个表 T_i 对应一个顶点 v_i；边集 E 表示已知的外键关系 $e_{ij}=(v_i, v_j)$；顶点特征为 $\boldsymbol{x}_i \in \mathbb{R}^d$（表结构特征向量）；边特征为 $\boldsymbol{e}_{ij} \in \mathbb{R}^k$（关系强度等）。目标函数为学习映射函数 $f: V \times V \rightarrow [0, 1]$，预测顶点对 (v_i, v_j) 存在关联的概率 $P(y_{ij}=1 \mid G)=f(v_i, v_j; \boldsymbol{\Theta})$，其中 $y_{ij}=1$ 表示存在关联。

图神经网络模型采用双层级联结构，专门用于发现数据库中的隐含关联关系。第一层为图卷积网络，其作用为捕获表节点的局部结构特征，通过聚合直接邻域节点的信息，形成每个表的初级嵌入表示，将原始特征压缩到隐藏维度空间并输出。对于顶点 v_i，第 l 层表示更新：

$$h_i^{(l)}=\sigma\left(\boldsymbol{W}^{(l)} \cdot \mathrm{AGGREGATE}^{(l)}\left(h_j^{(l-1)}: j \in \mathcal{N}(i)\right)\right) \tag{5}$$

其中，$\mathcal{N}(i)$ 是 v_i 的邻域顶点；σ 是 ReLU 激活函数；$\boldsymbol{W}^{(l)}$ 是可学习权重矩阵；AGGREGATE 函数为：

$$\mathrm{AGGREGATE}=\frac{1}{|\mathcal{N}(i)|}\sum_{j \in \mathcal{N}(i)} h_j^{(l-1)} \tag{6}$$

第二层是图注意力网络，其作用是学习不同邻接表的重要性权重。通过引入注意力机制，使模型能区分关键关联和次要关联。例如在学员－订单－培训班场景中，自动赋予“学员↔订单”比“订单↔银行流水”更高的注意力权重。注意力系数计算规则为：

$$\alpha_{ij}=\frac{\exp(\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\mathrm{W}h_i \,\|\, \mathrm{W}h_j]))}{\sum_{k\in\mathcal{N}(i)}\exp(\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\mathrm{W}h_i \,\|\, \mathrm{W}h_j]))} \tag{7}$$

顶点表示更新为：

$$h_i'=\sigma\Big(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}\mathrm{W}h_j\Big) \tag{8}$$

顶点对（v_i，v_j）的关联分数为：

$$\hat{y}_{ij}=\sigma\ (\boldsymbol{u}^{\mathrm{T}}\ [h_i \,\|\, h_j \,\|\varphi\ (x_i,\ x_j)]) \tag{9}$$

其中，φ 是顶点特征交互函数，$\boldsymbol{u}$ 是分类权重向量。

区别于传统规则方法，基于图神经网络的关联发现将数据库模式建模为属性图，利用多层图神经网络自动学习图结构与表/列属性中的复杂关联模式，仅需少量已知外键即可高精度预测多跳隐含依赖，实现无需人工规则的智能关联挖掘。

3.3 分片路由优化算法

为解决最小化分布式查询处理中的跨分片操作问题，建立分片路由优化算法。先构建一个加权的表关联图，其中边的权重反映表间连接操作频率和代价。这个关联图综合来自系统日志的实际查询模式和数据库设计时的外键约束信息。基于该关联图，采用发现算法来识别密切相关的表集合。通过优化模块度指标，将高度互联的表划分到同一个分片中，从而确保大多数频繁发生的连接操作都能在单个分片内完成。为了实现动态调整机制，通过持续监控查询负载的变化，当检测到某些表之间的连接频率显著增加时，会自动触发分片重组过程。这种重组操作在保证数据分布均衡的前提下，尽可能地将频繁连接的表对放置在同一个分片上。为了平衡分片调整的开销和收益，采用预测模型来预估调整后的性能改善。只有当预测的性能提升超过预定阈值时，才会实际执行分片调整操作。这种谨慎的策略有效避免不必要的分片重组带来的系统开销。分片路由优化算法流程图如图 4 所示。

分片路由优化算法实时监控分片负载与跨片连接代价，基于加权表关联图动态迁移高频关联表至同分片；结合时序模型预测热点趋势，仅当性能收益显著超越迁移成本时触发调整，实现负载感知的智能调度与资源优化。

3.4 自适应清洗算法群

3.4.1 数据清洗算法模型

针对多源异构数据的复杂清洗，本文提出统一分层数据清洗算法模型。首先，采用分层解耦机制将清洗流程分解为预处理、关联映射、特征工程和业务转换四个独立而又协同的层级；其次，采用动态安全协议，智能地结合国密 SM4 与 AES 加密算法；第三，引入基于 KL 散度的数据分布漂移检测机制；最后，建立自适应学习

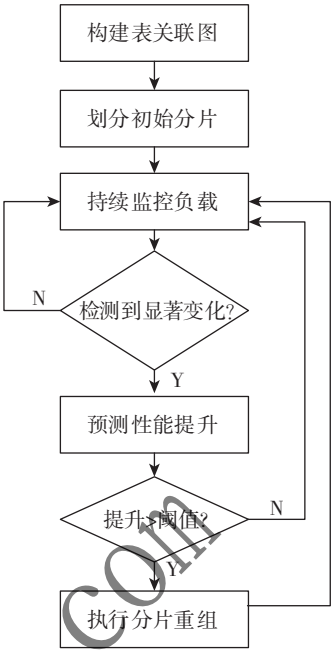

图 4 分片路由优化算法

框架。模型将整个数据清洗过程形式化为一个优化问题，旨在最小化清洗过程中的数据损失，同时满足信息安全性和处理效率的约束条件。具体而言，模型通过四个有序的处理层级来实现这一目标。预处理层负责基础数据净化，关联映射层处理实体解析，特征工程层构建高级数据特征，业务转换层确保数据符合业务规则。

给定异构数据集 $D=\bigcup_{k=1}^{4}D_k$，数据清洗任务可表述为寻找最优转换函数$\mathcal{F}^*$：

$$\mathcal{F}^*=\underset{\mathcal{F}\in\mathcal{H}}{\operatorname{argmin}}\Big[\sum_{k=1}^{4}w_k Q\ (\mathcal{F}\ (D_k))+\lambda\varepsilon\ (\mathcal{F})\Big] \tag{10}$$

其中，Q（·）为数据质量评估函数，ε（·）为计算复杂度度量，$\mathcal{H}$ 为候选函数空间（含加密、映射、特征工程等操作）。最优清洗函数可分解为四层复合函数：

$$\mathcal{F}\ (d)=\Gamma\circ\Phi\circ\Psi\circ\Omega\ (d) \tag{11}$$

其中，Ω 表示预处理层，Ψ 表示关联映射层，Φ 表示特征工程层，Γ 表示业务转换层。

预处理层通过正则表达式验证和空值检查来过滤无效数据。其次，根据数据敏感级别实施差异化加密策略，对高敏感数据采用国密 SM4 算法，标准敏感数据则使用 AES 加密结合 Base64 编码。分层加密操作可定义为：

$$\mathrm{Enc}\ (a_k)=\begin{cases}\mathrm{SM4}\ (a_k,\ K_{\mathrm{high}}), & a_k\in\mathcal{A}_{\mathrm{lv3}}\\ \mathrm{AES}\circ\mathrm{B64}\ (a_k,\ K_{\mathrm{std}}), & a_k\in\mathcal{A}_{\mathrm{lv2}}\end{cases} \tag{12}$$

采用分布式哈希技术消除重复记录，确保每个数据分片中的记录都具有唯一性。分布式去重定义为：

$$\mathcal{D}_{\text{pre}} = \{ d \in \mathcal{D} \mid \text{rank}\ (d \mid \mathcal{K}) = 1 \}, \quad \mathcal{K} = \{ k_1, k_2, k_3 \} \tag{13}$$

其中，$\mathcal{K}$ 是复合键集合，分别表示用户标识字段、手机号字段、系统时间戳。

关联映射层主要解决实体解析问题，通过计算数据向量与参考表记录之间的欧氏距离来寻找最佳匹配。实体解析模型为：

$$\Psi\ (d) = \underset{d' \in \mathcal{T}_{\text{ref}}}{\text{argmin}} \left\| v\ (d) - v\ (d') \right\|_2 \tag{14}$$

属性同步机制采用加权平均方法，将相关联的子表属性值按预定权重合并到主表记录中，所有权重之和严格保持为 1 以确保数据一致性。属性同步定义为：

$$d_{\text{sync}}^{(m)} = \sum_{j=1}^{J} w_j \cdot \mathcal{T}_{\text{sub}}\ [j]\ (\text{GID}),\ \sum w_j = 1 \tag{15}$$

特征工程层构建多维特征空间，包含语义特征、时间特征和结构化特征三大类。语义特征通过 BERT 模型处理解密后的文本数据获得，动态解密函数 $\varphi_{\text{sem}}\ (x)$ 为：

$$\varphi_{\text{sem}}\ (x) = \text{BERT}\ (\text{Dec}_{\text{SM4}}\ (x_{\text{enc}})) \tag{16}$$

时间特征采用指数衰减函数计算，衰减系数控制着历史数据的衰减速率，时间衰减 $\varphi_{\text{time}}\ (t)$ 为：

$$\varphi_{\text{time}}\ (t) = \text{e}^{-\lambda \cdot (t_{\text{now}} - t_{\text{event}})} \tag{17}$$

业务转换层基于有限状态机原理，定义标准状态集合，以及触发状态转移的事件类型。

$$\delta\ (q, \sigma) = \begin{cases} q_1, & \sigma = 1 \\ q_2, & \sigma = 0 \end{cases} \tag{18}$$

最后，通过约束函数的乘积运算，确保输出数据严格满足所有预设的业务规则要求。

统一分层数据清洗算法完整流程图如图 5 所示。

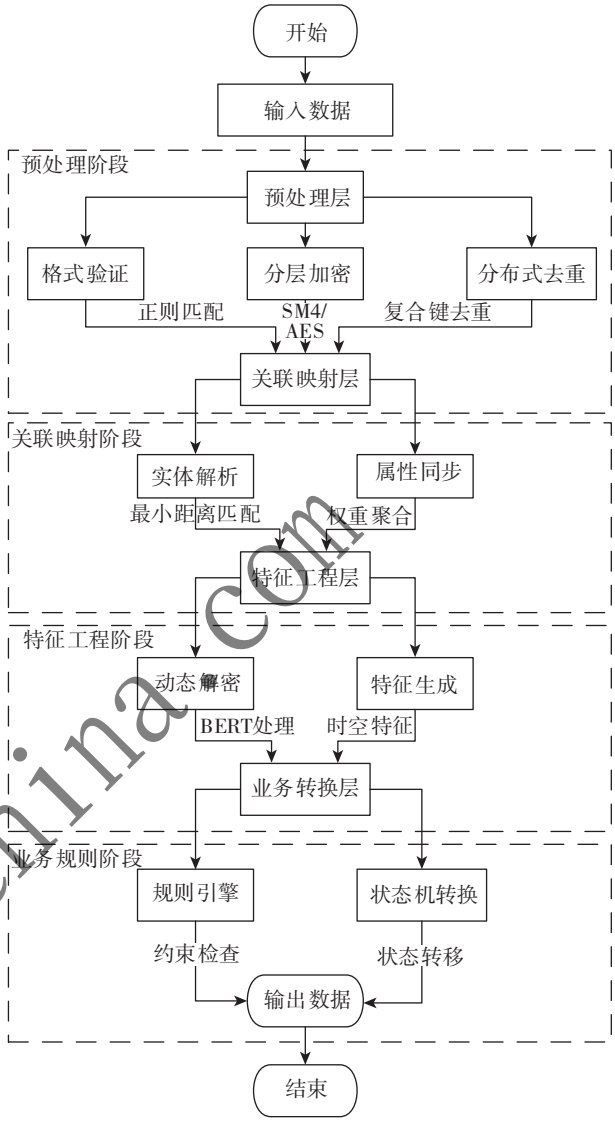

图 5　统一分层数据清洗算法流程图

3.4.2　面向大规模数据的异构数据动态迁形算法

针对大规模数据清洗的核心挑战，特别是涉及亿级规模以上的高频时序数据处理，提出异构数据动态迁形算法（TransMorph）。算法包括预处理、双阶段映射、规则引擎和数据校验四个主要部分。数据处理流程始于多维数据预处理阶段。在此阶段通过添加辅助字段实现模式扩展，并采用索引优化技术构建复合索引以提升操作效率。随后进入双阶段映射过程：第一阶段通过结合正则匹配与模糊哈希技术将原始用户标识解析为标准手机号，输出覆盖率达到较高水平；第二阶段则通过构建布谷鸟哈希映射表将标准手机号精确映射到全局唯一的用户编号，并引入渐进式重映射机制以支持增量更新。为加速用户与课程信息的关联查询，采用二级缓存策略。接下来由类型驱动的规则引擎处理学习进度逻辑，该引擎基于预定义的课程类型到进度规则的映射元组运作，能够动态生成进度规则标识并注入数据流水线，当规则变更时利用轻量级状态机精准标记受影响批次，将需重处理的数据比例控制在合理范围内，同时通过并行优化进度校验过程，大幅压缩海量数据的验证时间。在输出最终结果前，通过三重数据校验以确保质量：首先应用外键约束保证数据的引用完整性，其次进行非空验证以过滤无效记录，最后实施统计抽样校验确保数据分布一致性，此过程采用索引优化的嵌套循环连接策略。算法流程如图 6 所示。

3.4.3　小结

本部分技术的核心体现在分层优化模型的自适应处理能力、安全策略的动态选择机制及规则引擎的敏捷响应特性。统一分层清洗算法模型将清洗任务形式化为带约束的优化问题，支持各处理层依据输入数据的实时特征与上下文语义，动态选择最优处理策略。这种基于数据特征库与策略库的分层决策机制，实现了清洗过程的自适应闭环优化。TransMorph 规则引擎依托预定义的映射

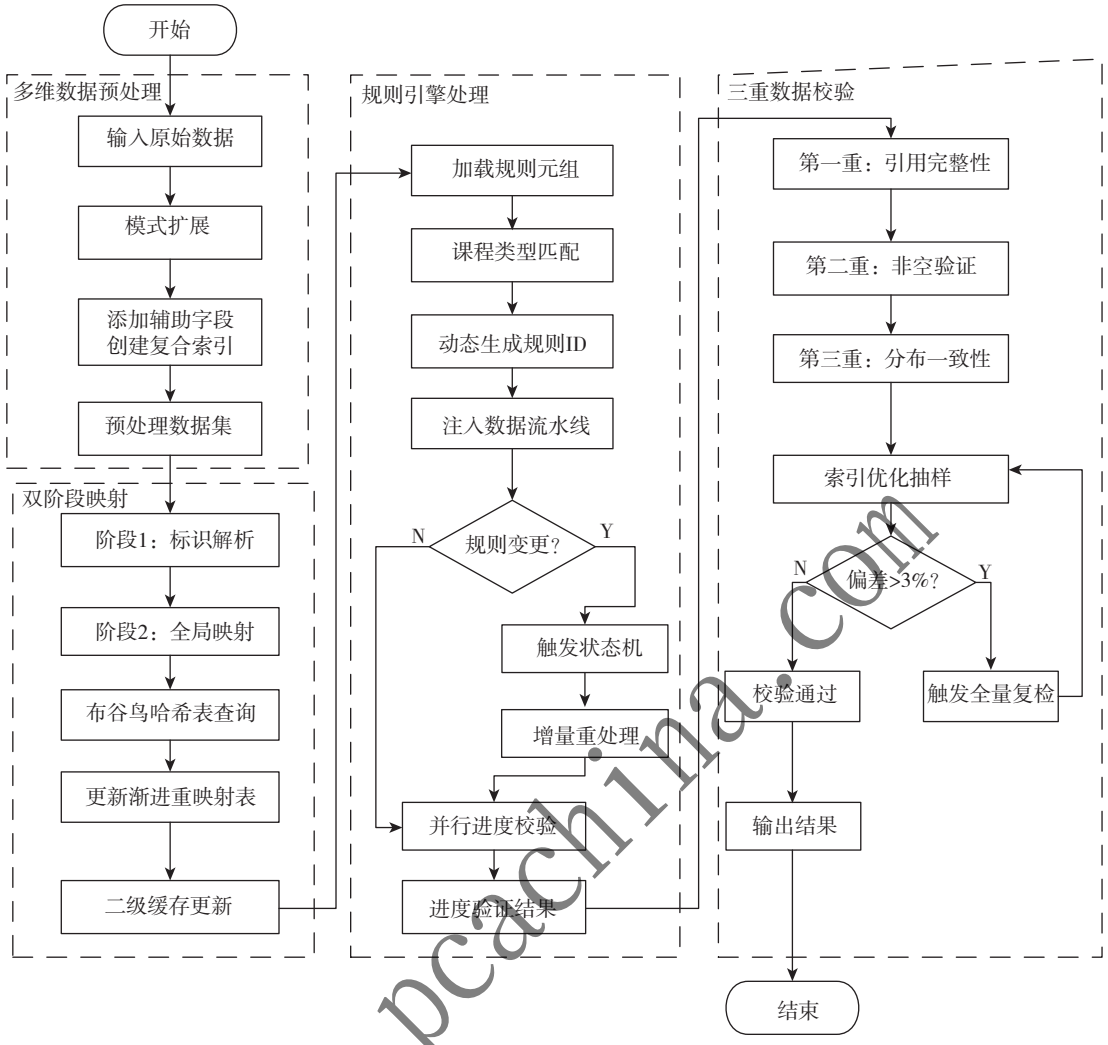

图 6　TransMorph 算法流程图

元组，动态生成数据类型相关的进度校验规则并注入数据流水线；当业务规则变更时，通过轻量级状态机精确标记受影响的数据批次范围，将重处理数据量压缩至合理阈值内，提升对业务需求变化的智能响应。

4　实践验证

4.1　实践环境

本实践采用某大型在线教育平台的真实业务数据，核心数据包括证书信息、订单信息、学习记录、培训班学员关联信息等，数据规模超 3 亿条。该数据集典型特征包括数据规模庞大且增长迅速；学习记录为高频更新的时序数据；存在复杂的多对多关联关系；包含特定的业务规则约束。这些特征使其成为验证异构数据迁移框架处理规模、时效、关联、规则等能力的理想案例。

为解决新旧系统割接数据大规模数据迁移问题，构建含 14 个节点 Kubernetes 集群的容器化环境，采用分布式架构处理异构数据迁移任务。实践环境包含源数据库集群、目标存储集群和数据处理平台三大部分。源端为关系型数据库实例，目标端由分片式关系型数据库集群和文档型数据库集群组成，两者采用不同的数据分片策略。数据处理平台部署于容器编排系统之上，各组件间通过高速网络互联。系统集成元数据管理服务实现数据血缘追踪，并配备完善的监控体系保障任务执行的可观测性。环境操作真实业务场景，包含多种数据类型和关联关系，验证系统在复杂业务条件下的稳定性。实践表明，该架构能够有效支撑异构数据的高效转换与分发，满足大规模数据迁移的性能要求。

4.2　关键指标定义

本文重点考察三个维度的关键指标，即数据处理质量、系统吞吐能力以及资源利用率。

数据处理质量方面，定义数据一致性与数据完整率，其中数据一致性为成功完成的跨数据库/集群事务数除以发起的跨数据库/集群事务总数，衡量在分布式环境下保

证数据 ACID 属性的能力；数据完整率为成功迁移的目标端记录数除以源端有效记录总数，衡量迁移过程是否丢失有效数据。异常数据自动修复率是衡量数据迁移系统智能化水平的关键指标，它反映系统自动识别并修正数据异常的能力。其计算公式为系统自动修复的异常记录数除以检测出的总异常记录数。

系统吞吐能力方面，数据处理速率是单位时间内成功迁移并验证通过的记录数（records/sec），反映系统核心处理能力。峰值吞吐为在最大持续负载压力下达到的最高稳定数据处理速率。总耗时指系统完成端到端数据迁移全流程所需的总耗时，包括数据抽取、转换、加载及最终一致性校验四个阶段的累计时间，该指标直接反映系统处理效率，单位为小时。

在资源利用率方面，通过细粒度的监控数据采集，验证系统在 CPU、内存等维度的资源消耗情况。CPU 利用率为系统处理任务期间，所有相关容器/Pod 的 CPU 使用率平均值及峰值。内存消耗为系统处理任务期间，所有相关容器/Pod 的内存占用平均值及峰值。

4.3 实践效果分析

为全面评估本研究的系统性能优势，选取三大主流 ETL 工具作为对比组，分别是 Informatica PowerCenter（IPC）、IBM InfoSphere DataStage（ISDS）和 Microsoft SQL Server Integration Services（SSIS）稳定版本，性能对比如表 1 所示。

表 1　本文工作与主流 ETL 工具性能对比

关键指标	本文工作	IPC	ISDS	SSIS
总耗时/h	3. 8	14. 2	12. 5	15. 7
平均处理速率/（rec/s）	14 600	3 900	4 400	3 500
CPU 利用率/%	<8	22	25	18
内存消耗峰值/GB	<32	95	85	110
数据一致性/%	99. 98	99. 85	99. 88	99. 82
异常自动修复率/%	92. 1	65. 5	70. 2	60. 8
复杂关联查询/%	87. 9	40. 2	35. 7	30. 1
存储空间/%	46. 3	15. 8	12. 5	10. 2

表 1 展示的 AUTO-MIG 框架整体性能优势是其内部各核心创新组件协同作用的结果。虽然本次验证侧重于端到端的业务场景效果，但通过深入分析性能数据与框架工作机制，可以识别关键模块的主要贡献。

在复杂关联查询方面，本文方法实现了高达 87.9% 的性能提升，这主要得益于基于图神经网络的关联发现机制。该机制能够自动识别隐含的表间关联，为后续的智能决策提供支持。决策层依据所发现的关联生成了更优的数据分布策略，例如将高频关联实体划分至同一分片，并合理采用反范式化存储，从而显著减少了迁移后跨节点 Join 操作的发生，成为提升复杂查询效率的关键。

在数据迁移吞吐与时效性方面，系统实现了总耗时 3. 8 h、平均处理速率约 14 600 rec/s 的性能。这反映了双模式增量捕获机制与流批协同执行架构的有效性。该机制通过全量并行扫描和传输优化高效处理海量历史数据，并借助日志流实现增量数据秒级延迟捕获与处理，在保障业务衔接时效性同时，满足了严格的时间要求。

异常数据处理方面，系统取得了 92. 1% 的自动修复率，体现了统一分层清洗模型及其核心算法 TransMorph 的有效性。该模型通过将业务规则转化为可计算状态机与约束函数，实现了对教育领域特定数据冲突的精准识别与自动修复，显著减少了人工干预并确保合规性。TransMorph 算法通过双阶段映射、规则引擎优化与索引策略等措施，实现了对大规模时序数据的高效清洗与转换。

在资源使用方面，系统实现了 CPU 平均利用率低于 8%、内存峰值低于 32 GB，并节省了 46. 3% 的存储空间。这一效果得益于自适应分片路由机制依据负载监测与预测实现动态负载均衡，避免热点产生；同时，多目标存储优化模型在查询性能、存储成本与维护开销之间进行权衡，智能选择混合存储方案（如关系型与文档库结合）并应用压缩技术，显著提升了资源分配效率和存储经济性。

上述分析表明，AUTO-MIG 框架的核心设计（动态分片、规则引擎、流批协同、多目标优化）有效化解了培训数据在规模性、时序性、关联复杂性及规则特定性方面带来的独特挑战。其在迁移效率、查询性能、规则处理及资源利用率上的显著提升，验证了该框架在处理具有类似复杂特征的大规模异构数据迁移任务时的先进性与实用性。

5　结论

本文针对系统迁移过程中面临的异构性、规模性与时效性挑战，提出了智能转换框架 AUTO-MIG。其核心创新在于基于图神经网络的深度关联发现机制与面向大规模异构迁移的双模式协同执行引擎。在架构层面，元数据感知层通过动态语法适配机制支持多源异构数据的自动化解析，结合双模式增量同步技术实现历史数据与实时变更的高效协同；智能决策层建立多目标优化模型，综合权衡性能、成本与运维因素驱动存储方案决策，并构建自动化模型转换引擎解决跨数据库模型转换问题；分布式执行层采用流批融合的数据清洗框架保障质量，

结合自适应分片路由与事务优化机制确保集群负载均衡与状态一致性。在算法层面，基于自解释 Schema 转换方法实现智能化的类型兼容处理，基于图神经网络的关联发现模型深入挖掘数据间的隐性关系，并针对复杂业务场景建立统一分层数据清洗算法模型实现端到端的数据治理。针对大规模异构数据的 TransMorph 算法创新双阶段编号映射与规则驱动，使历史数据清洗效率大幅度提升。经过大规模实践验证，该框架显著提升迁移效率，优化了资源利用率，并在数据完整性、事务可靠性等方面达到工业级标准，较传统方案展现出较高优势。

未来研究将沿五个维度持续深化：一是在技术普适性方面，将进一步扩展对新型数据库的兼容能力，增强算法对非结构化数据与无模式系统的适应性；二是实时性优化将探索增量同步的极致延迟压缩，结合硬件加速技术突破流处理瓶颈；三是智能化升级拟引入强化学习实现决策模型的动态调优，构建预测性分片机制以主动规避系统热点；四是生态化集成重点研发自然语言驱动的规则生成接口，降低使用门槛，深化与云原生架构的融合，支持弹性扩缩容，同时建立端到端的安全防护体系保障敏感数据迁移；五是设计并执行系统的消融实验并进行模块级对比验证。通过持续完善技术边界与生态能力，推动 AUTO-MIG 从迁移工具向智能数据中枢演进，为数字化转型提供自主安全的技术支撑。

参考文献

[1] 荆一楠，张寒冰，李智鑫，等. 面向金融场景的下一代数据库测试基准研究 [J]. 中国工程科学，2022，24（4）：121－132.

[2] 朱蔚. 大型企业 PLM 系统迁移的难点与对策 [J]. 软件工程与应用，2024，13（6）：759－764.

[3] CÔTÉ P O，NIKANJAM A，AHMED N，et al. Data cleaning and machine learning：a systematic literature review [J]. Automated Software Engineering，2024，31（2）：54.

[4] WNEK K，BORYŁO P. A data processing and distribution system based on apache NiFi [C]//Photonics. MDPI，2023，10（2）：210.

[5] EMMA O T，PEACE P. Building an automated data ingestion system：leveraging Kafka connect for predictive analytics [J]. 2023.

[6] 张晨雪，肖祥慧，顾阳. 数据清洗方法研究综述 [J]. 北京印刷学院学报，2025，33（3）：49－55.

[7] WANG D，LI S，FU X. Short-term power load forecasting based on secondary cleaning and CNN-BILSTM-attention [J]. Energies，2024，17（16）：4142.

[8] YUAN Q，ZHU Y，XIONG G，et al. ULDC：unsupervised learning-based data cleaning for malicious traffic with high noise [J]. The Computer Journal，2024，67（3）：976－987.

[9] SALAHUDDIN M，MAJEED S，HIRA S，et al. A systematic literature review on performance evaluation of SQL and NoSQL database architectures [J]. Journal of Computing & Biomedical Informatics，2024，7（2）.

[10] DANG T K，HUY T M，DANG L H，et al. An elastic data conversion framework：a case study for MySQL and MongoDB [J]. SN Computer Science，2021，2（4）：325.

[11] 吴信东，董丙冰，堵新政，等. 数据治理技术 [J]. 软件学报，2019，30（9）：2830－2856.

[12] 秦之涓，张会平，王斌，等. 智慧治理中的数据质量 管理困境及对策研究 [J]. 大数据，2024，10（5）：151－167.

[13] 黄跃珍，杨芬，田丰，等. 基于图的异构数据集成方法研究 [J]. 大数据，2025，11（1）：21－35.

（收稿日期：2025－07－14）

作者简介：

许文静（1990－），通信作者，男，博士，讲师，主要研究方向：人工智能、大数据智能、数字教育、信息安全等。E-mail：xuwenjing@ sasac. gov. cn。

安宁（1990－），男，硕士，高级工程师、高级经济师，主要研究方向：软件工程、云计算、人工智能、大数据、信息安全、数字经济等。

于重（1999－），男，硕士，助教，主要研究方向：通信工程、人工智能、数字教育、信息安全等。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部