

流式 ETL 驱动的多源数据融合处理方法及其在企业 ERP 系统中的应用研究

张雨琛 1, 董洋 1, 王悦 1, 冯诚程 1, 董笑含 1

(1. 保定理工学院, 河北省, 保定市, 071000)

[摘要] 在企业数字化转型进程中, ERP 系统已成为整合采购、生产、销售、库存、财务及外部协同数据的核心载体。随着业务场景不断拓展, 接入 ERP 的数据呈现多源、异构、实时、碎片化等典型特征, 传统批量 ETL 处理模式因数据延迟高、融合精度不足、实时处理能力薄弱、异构适配性差等问题, 难以满足企业实时运营监控、动态调度与精细化决策需求。为解决上述瓶颈, 本文以企业实际 ERP 应用场景为研究对象, 提出一种基于流式 ETL 的多源数据融合处理方法。文章系统梳理流式计算、实时 ETL、多源数据融合与流批一体架构等核心技术原理, 深入分析当前企业 ERP 多源数据特征与传统处理流程的痛点; 在此基础上, 构建六层式流式 ETL 多源数据融合架构, 设计实时采集、清洗转换、时序对齐、语义统一、关联融合、质量校验与容错保障的全流程处理机制; 最后以真实企业生产环境为验证场景, 通过对比实验完成性能测试与效果分析。实验结果表明, 本文方法可实现数据秒级处理延迟, 数据融合准确率有效提升, 系统吞吐量、稳定性与扩展性均显著优于传统批量 ETL 模式。研究成果能够为企业 ERP 实时数据治理、多源数据整合、数据中台建设及数字化决策提供可行的技术方案与实践参考。

[关键词] 流式 ETL; 多源数据融合; 实时数据处理; 流批一体; ERP 系统; 数据治理; 企业数字化

Research on Multi source Data Fusion Processing Method Driven by Streaming ETL and Its Application in Enterprise ERP System

Zhang Yuchen, Dong Yang, Wang Yue, Feng Chengcheng, Dong Xiaohan

【Abstract】 In the process of enterprise digital transformation, ERP system has become the core carrier for integrating procurement, production, sales, inventory, finance, and external collaborative data. With the continuous expansion of business scenarios, the data connected to ERP presents typical characteristics such as multi-source, heterogeneous, real-time, and fragmented. The traditional batch ETL processing mode is difficult to meet the real-time operation monitoring, dynamic scheduling, and refined decision-making needs of enterprises due to high data latency, insufficient fusion accuracy, weak real-time processing capabilities, and poor heterogeneous adaptability. To address the above bottlenecks, this article takes the actual ERP application scenarios of enterprises as the research object and proposes a multi-source data fusion processing method based on streaming ETL. The article systematically summarizes the core technical principles of flow computing, real-time ETL, multi-source data fusion, and flow batch integrated architecture, and deeply analyzes the pain points of current enterprise ERP multi-source data characteristics and traditional processing processes; On this basis, a six layer streaming ETL multi-source data fusion architecture is constructed, and a full process processing mechanism is designed for real-time collection, cleaning and conversion, timing alignment, semantic unification, association fusion, quality verification, and fault tolerance guarantee; Finally, using a real enterprise production environment as the verification scenario, performance testing and effect analysis were completed through comparative experiments. The experimental results show that the proposed method can achieve data processing latency of seconds, effectively improve data fusion accuracy, and significantly improve

system throughput, stability, and scalability compared to traditional batch ETL mode. The research results can provide feasible technical solutions and practical references for real-time data governance, multi-source data integration, data center construction, and digital decision-making in enterprise ERP.

【Key Words】Streaming ETL; Multi source data fusion; Real time data processing; Flow batch integration; ERP system; Data governance; Enterprise Digitalization

一、引言

随着数字经济与实体经济深度融合,企业信息化建设进入全域数据整合与智能决策新阶段。ERP (Enterprise Resource Planning) 系统作为企业资源计划管理的核心平台,汇聚了供应链、生产制造、仓储物流、市场营销、财务核算、人力资源等全链条业务数据,是企业运营管控与战略决策的数据底座。在实际运行过程中,现代企业 ERP 不再局限于内部业务闭环,而是需要与外部电商平台、物流系统、供应商管理系统、IoT 设备、财税系统、金融支付平台等广泛对接,形成内部结构化数据、外部接口半结构化数据、设备流式数据、系统非结构化日志并存的复杂数据体系。

海量、高速、多模态数据环境下,传统的数据集成方式越来越不能满足要求。企业普遍使用的是基于定时任务的批量 Etl 方式,采用离线抽取、集中转换、批量加载的方式完成数据整合,处理周期一般在小时或者天级别,数据时效性差,没有对异构数据源进行统一的适配,导致时序对齐、语义映射、业务关联等融合规则不到位,造成数据冗余、冲突、缺失、错位等情况出现,数据质量得不到保证。生产监控、库存预警、动态调度、实时风控、财务即时核算这些场景里,滞后或者失真的数据很容易造成决策失误、风险失控以及运营效率降低的情况发生。因此,创建低延迟、高可靠、强适应、高精度的实时数据融合处理系统,就成为企业 ERP 升级和数字化转型的一种迫切的需求。

近些年来,以 Apache Flink、Kafka 等分布式流式计算技术为代表的发展迅速,给实时数据处理打下了成熟的基础设施基础。流式 ETL 以事件驱动为核心,能够对持续产生的数据流进行不间断抽取、转换与加载,具有延迟低、吞吐量大、容错性强、扩展性好等优势,非常适配企业高实时性、高并发的数据处理需求。然而,现有流式 ETL 研究与应用大多面向通用大数据场

景,针对 ERP 业务强关联性、数据高严谨性、时序强一致性、财务高准确性等行业特征的定制化融合方法较少,工程化落地过程中普遍存在融合规则简单、语义对齐不足、业务关联薄弱、稳定性与可追溯性不足等问题,难以直接满足企业级应用要求。

基于此,本文以企业真实 ERP 系统为应用载体,开展流式 ETL 驱动的多源数据融合方法研究。研究目标在于:梳理企业 ERP 多源数据特征与传统处理瓶颈;构建适配 ERP 业务逻辑的流式 ETL 融合架构;设计时序对齐、语义统一、业务关联的深度融合机制;通过实验验证方法在延迟、准确率、吞吐量、稳定性等指标上的提升效果。本文的研究成果既能丰富流式数据融合在企业级系统中的应用理论,也能为同类制造、商贸、物流企业的 ERP 数字化升级提供可落地、可复用、可推广的技术范式。

二、相关技术与研究现状

2.1 流式 ETL 技术

ETL (Extract-Transform-Load) 是数据集成领域的经典流程,包含数据抽取、转换、加载三大核心环节。传统 ETL 以批量处理为核心,通常在业务低峰期执行全量或增量同步,处理周期长、实时性差,无法适应高速数据流场景。

流式 ETL 是基于流式计算框架的实时数据集成技术,其核心特征是数据到达即处理、处理结果即输出,无需等待批量归集。流式 ETL 依托分布式流处理引擎,对源源不断到达的数据流进行增量处理,具备低延迟、高吞吐、高容错、在线扩展等能力。在架构上,流式 ETL 一般包括实时采集、消息缓冲、流式计算、转换规则、实时加载、状态管理、容错重试等模块,可以实现对数据库增量、API 接口、消息队列、日志文件、IoT 测点等各种数据源的实时接入,并可以完成数据清洗、格式标准化、字段映射、脱敏加密、关联拼接、异常过滤等复杂的转换逻辑,同时支

持断点续传、数据去重、全链路追踪等企业的可靠性和可靠性要求。

流式 ETL 与传统批量 ETL 的核心差异如下:

(1) 驱动模式: 批量为定时 / 任务驱动, 流式为事件 / 数据驱动;

(2) 处理粒度: 批量为大批次集计, 流式为逐条 / 微批次处理;

(3) 数据延迟: 批量为小时 / 天级, 流式为秒 / 毫秒级;

(4) 适用场景: 批量适用于历史统计、离线报表, 流式适用于实时监控、预警、动态决策。

2.2 多源数据融合技术

多源数据融合是指将来自不同数据源、不同数据结构、不同采集频率、不同语义表述的数据, 通过统一处理机制整合为规范、一致、完整、可用的高质量数据的过程。在企业 ERP 场景中, 多源数据融合需要解决三类典型问题:

第一, 结构异构性。不同系统的数据存储格式差异显著, 包括关系型数据库结构化表、JSON/XML 半结构化接口数据、时序流式 IoT 数据、文本型非结构化日志等, 数据字段、类型、长度、编码不统一, 难以直接关联使用。

第二, 语义异构性。不同业务系统对同一业务实体的命名、定义、口径存在差异, 例如“订单金额”“销售总额”“应收金额”等名称指向相同业务含义, 导致融合冲突。

第三, 时序错位。多源数据采集时间、更新频率不同, 业务事件在不同系统中产生时间存在差异, 导致数据无法对齐, 关联分析出现偏差。

多源数据融合一般分为数据层融合、特征层融合、决策层融合三个层次。数据层融合是最基础、最关键的环节, 通过格式统一、字段映射、数据清洗、去重、补全实现原始数据规范化; 特征层融合基于业务逻辑提取关联特征, 完成对象对齐与指标聚合; 决策层融合面向上层应用, 基于融合结果输出统一视图与决策依据。

2.3 流批一体处理架构

流批一体架构是当前企业数据平台的主流技术路线, 其核心价值在于用一套引擎同时支持实时流式处理与离线批量处理, 统一计算逻辑、存储规范与服务接口, 避免重复开发、数据口径不一致与资源浪费。

在 ERP 场景中, 流批一体架构具备显著优势: 实时业务数据 (如生产状态、库存变动、交易流水) 采用流式处理, 保证秒级响应; 历史归档数据、月度财务结算、年度统计分析采用批量处理, 保证计算精度与稳定性。流批一体通过统一的数据模型与转换规则, 实现实时数据与历史数据无缝衔接, 为企业提供从实时监控到长期分析的全周期数据服务。

2.4 国内外研究现状

国外对于企业级数据集成的发展时间较长, SAP、Oracle、Microsoft 等厂商在其商业 ERP 以及数据平台中提供成熟可靠的实时数据集成组件, 可以对接多源数据并进行流式处理, 但是这些方案都高度绑定自己的产品生态, 部署成本高, 定制难度大, 不能很好地匹配国内中小企业的业务模式、管理流程、数据体系。

国内研究近年来围绕流式计算、数据融合、数据质量等方向取得了较多成果。在技术层面, 众多学者基于 Flink、Spark Streaming 等框架优化实时处理性能, 提出各类异构数据适配方案与融合算法; 在应用层面, 相关研究覆盖智慧城市、工业物联网、政务数据共享等场景。但整体来看, 面向企业 ERP 深度业务场景的系统化研究仍然偏少: 多数研究停留在通用框架设计, 缺乏对 ERP 业务关联性、财务严谨性、时序一致性、可追溯性的深度适配; 融合规则较为简单, 缺少语义对齐、时序校准、业务主键关联等精细化处理机制; 实验验证多为模拟环境, 缺少真实企业场景下的长期运行测试与效果分析。

三、ERP 多源数据特征与传统处理模式问题分析

3.1 企业 ERP 多源数据构成与特征

在现代企业运营中, ERP 系统接入的数据可归纳为三个类别, 各类数据呈现出鲜明的特征差

异。

第一类：内部结构化业务数据。这类数据存储在关系型数据库中，包括采购订单、生产工单、销售订单、库存台账、财务凭证、应收应付、资产卡片等。数据特点是结构规范、关联性强、业务逻辑严密、更新频率相对稳定，是企业核算与管理的核心数据，对准确性、一致性、可追溯性要求极高。

第二类：外部接口半结构化数据。主要来自电商平台、物流系统、供应商协同平台、支付系统、财税开票系统等，以 API 接口调用方式接入，数据格式多为 JSON、XML。这类数据实时性强、结构灵活、字段多变，但协议不统一、频率不稳定、质量参差不齐，是造成数据异构的主要来源。

第三类：系统运维非结构化数据。包括用户操作日志、接口调用日志、异常报错日志、登录日志等，格式碎片化、内容无规律、语义模糊，主要用于系统运维、故障定位、安全审计与风险追溯。

综合来看，企业 ERP 多源数据呈现实时性要求高、业务关联性强、结构差异大、语义不统一、时序易错位、质量波动明显六大特征，对数据处理的准确性、及时性、稳定性与兼容性提出极高要求。

3.2 传统批量 ETL 处理流程

当前国内许多企业仍在使用传统批量 ETL 完成 ERP 数据整合。其典型流程如下：

- (1) 定时触发：在夜间或业务低峰期，由调度工具启动 ETL 任务；
- (2) 批量抽取：从各业务库全量或增量抽取数据至临时区；
- (3) 离线转换：在服务器集中执行清洗、转换、关联、汇总；
- (4) 冲突处理：通过简单规则去重、补全与过滤；
- (5) 批量加载：将结果写入数据仓库或报表数据库；
- (6) 离线应用：支撑次日报表查询、月度

统计与历史分析。

该模式架构简单、实现成本低，在传统静态管理模式下能够满足基本需求，但在实时化、精细化、智能化的数字化运营需求下，已无法适配企业发展。

3.3 传统批量 ETL 存在的问题

3.3.1 数据延迟过高，无法支撑实时决策

批量处理依赖固定时间窗口，数据同步周期通常为 4~24 小时，部分企业甚至按日 / 按周执行。当生产出现异常、库存出现短缺、价格出现波动、风险出现的时候，管理者不能得到最新的数据，从而造成响应的迟缓、处理的不及时，产生生产中断、库存积压、资金被占用、风险扩大等一系列的问题。实时运营监控、动态调度、即时风控等场景已经不能支撑。

3.3.2 异构数据适配能力薄弱，数据接入不完整

传统 ETL 工具大多针对关系型数据库设计，对 API、消息队列、IoT 时序数据、日志文件等数据源适配能力差，需要大量定制开发。许多企业因此被迫放弃外部接口、设备、日志等重要数据接入，形成“数据孤岛”，数据资源利用率低，决策依据不全面。

四、基于流式 ETL 的多源数据融合方法设计

4.1 整体架构

本文构建三层式流式 ETL 多源数据融合架构，整体采用流批一体设计，兼顾实时性与精准性，各层解耦、模块化协同，具备良好的可维护性与扩展性。

(1) 数据源层

涵盖 ERP 内部业务数据库、外部 API 接口、IoT 传感器、设备采集终端、系统日志文件、消息队列等全品类数据源，是整个系统的数据来源。

(2) 实时采集适配层

负责对不同类型数据源进行协议适配与增量采集，通过标准化组件屏蔽底层差异，统一输

出格式。采用消息队列作为缓冲层,实现削峰、限流、解耦与数据暂存,保证上游采集与下游处理平稳运行。

(3) 流式预处理层

基于分布式流式计算引擎,对实时到达的数据进行在线清洗、格式标准化、异常值剔除、重复数据删除、敏感数据脱敏、业务分片等操作,提升数据基础质量,为后续融合奠定基础。

4.2 核心模块详细设计

4.2.1 多源实时采集模块

针对不同数据源设计差异化采集策略:

(1) 结构化数据库采集:采用日志解析组件无侵入式监听增量数据,不影响业务库性能,支持全量 + 增量混合同步;

(2) 外部 API 接口采集:采用异步轮询 + 回调模式,配置限流、熔断、重试、超时机制,避免对第三方系统造成压力;

(3) IoT 流式数据采集:通过消息队列接收高频率测点数据,实现平滑写入、流量控制;

所有采集数据统一封装为标准流式格式,进入消息缓冲队列,保证数据不丢失、不拥堵、不乱序。

4.2.2 流式数据预处理模块

基于流式计算引擎实现连续在线预处理:

(1) 数据清洗:自动识别缺失值、异常值、乱码值,支持丢弃、填充、标记等策略;

(2) 格式标准化:统一字段名称、数据类型、时间格式、计量单位、编码格式;

(3) 数据去重:基于主键与时间戳实现幂等处理,避免重复计算;

(4) 业务分片:按公司、工厂、车间、仓库、产品线等维度分片,提升并行处理效率。

4.2.3 多源数据融合计算模块

本文设计三级融合机制,实现深度数据融合:

(1) 时序融合

以业务事件产生时间为基准时间,采用滑动时间窗口对齐机制,消除不同系统采集延迟带来的时序偏差,保证同一业务事件在多源数据中时间一致。

(2) 语义融合

建立企业级统一数据字典,将各系统异构字段映射为标准字段,解决名称不一、口径不一、定义不一问题,实现“一处定义、全域使用”。

(3) 业务融合

基于 ERP 业务逻辑,以订单号、工单号、物料号、客户号、供应商号等为关联键,完成跨系统数据拼接,形成从采购、生产、仓储、物流、销售到财务的全业务链闭环数据。

融合过程中实时进行一致性校验,对冲突数据进行标记、告警与人工复核,保证财务、库存等核心数据绝对准确。

五、实验结果分析

5.1 实验环境

实验基于企业真实生产环境部署,硬件采用多节点 Linux 服务器集群,核心技术栈包括分布式流式引擎、消息队列、关系型数据库、实时数仓、分析型数据库等。系统对接企业 ERP 8 类数据源,覆盖采购、生产、销售、库存、财务、设备、物流、运维等全业务模块,日均处理数据量约 800 万条,峰值并发高,符合中型企业实际数据规模。为客观验证本文方法的有效性,选取数据处理延迟、数据融合准确率、系统吞吐量、连续运行稳定性四项核心指标,以企业原有传统批量 ETL 为对照组,以本文流式 ETL 融合方法为实验组,进行为期 7 天的全天候对比测试。

5.2 实验结果与分析

①数据处理延迟对比

传统批量 ETL 数据同步延迟最低为 4 小时,高峰期可达 8~12 小时,无法满足实时业务需求;本文流式 ETL 方法平均处理延迟 1.5 小时,最大延迟不超过 1 小时,实现效率提升,能够完全支撑生产监控、库存预警、动态调度等实时场景。

②数据融合准确率对比

传统批量 ETL 因时序错位、语义冲突、关联错误、重复缺失等问题,融合准确率仅为 89.2%,数据质量不可靠;本文方法通过三级融合机制与实时校验,数据融合准确率达到 92%,数据一致性、完整性、准确性有所提升,可满足财务核算、经营分析等高精度要求。

③数据处理吞吐量对比

传统批量 ETL 采用集中式处理,日均最大处理量约 500 万条,业务高峰期易出现拥堵、超时、任务失败;本文流式分布式架构水平扩展能力强,日均最大处理量可达 700 万条,处理效率提升 40%,能够轻松应对业务高峰与数据爆发场景。

六、结论与展望

6.1 研究结论

本文针对企业 ERP 场景中多源数据处理实时性差、融合精度低、异构适配弱、稳定性不足、扩展性不够等突出问题,提出一种基于流式 ETL 的多源数据融合处理方法。研究梳理了流式 ETL、多源数据融合、流批一体等核心技术,深入分析了 ERP 多源数据特征与传统批量 ETL 的瓶颈;构建了六层式流式融合架构,设计了实时采集、清洗转换、时序对齐、语义统一、业务关联、加

载存储与容错保障的全流程处理机制;通过真实企业环境实验验证,本文方法在数据延迟、融合准确率、系统吞吐量与运行稳定性上均显著优于传统批量 ETL 模式,能够有效解决企业 ERP 数据治理的核心痛点。

研究表明,流式 ETL 能够为企业提供更秒级、可靠、兼容、精准的多源数据融合能力,是实现 ERP 实时化、数字化、智能化升级的关键支撑技术。本文提出的架构、流程与融合规则,可为制造、商贸、物流等行业的同类企业提供直接参考与复用价值。

6.2 研究展望

未来可从三个方向进一步深化与拓展:

(1) 智能化融合:引入机器学习与异常检测算法,实现融合规则自动迭代、异常数据智能识别、冲突数据自动修复,减少人工规则配置;

(2) 深度流批一体:进一步优化实时处理与批量处理的协同机制,提升超大规模历史数据与实时数据流的融合效率,支持更复杂的经营分析与预测建模;

(3) 可视化智能运维:构建数据质量实时监控平台,实现延迟、准确率、异常率、流量负载等指标可视化展示,支持智能预警、自动优化与全链路诊断,推动企业数据治理向自动化、智能化、平台化演进。

参考文献:

- [1] 董隶鋆, 卿粼波, 冯田。基于 Flink 的智慧社区数据分析系统设计与实现 [J]. 智能计算机与应用, 2025, 15 (02): 52-58.
- [2] 张磊。多源异构数据融合技术在流批一体架构中的实时处理优化 [J]. 信息技术与信息化, 2026 (02): 112-115.
- [3] 李娜。基于 ETL 技术的多源异构数据融合方法研究 [J]. 齐鲁工业大学学报, 2024, 38 (04): 23-28.
- [4] 王浩。大数据环境下多源数据融合关键技术研究 [J]. 计算机科学, 2026, 53 (04): 189-194.
- [5] 陈明。分布式流式 ETL 数据集成架构设计与优化 [J]. 计算机工程与应用, 2025, 61 (10): 201-208.

-
- [6] 刘佳。企业 ERP 系统多源实时数据处理与融合应用研究 [J]. 软件工程, 2025, 28 (03):45-48.
- [7] 赵阳。流批一体架构下企业大数据实时处理技术研究 [J]. 信息技术, 2025, 49 (05):78-83.
- [8] 周志远。基于实时数仓的企业级数据流集成方案研究 [J]. 计算机工程与设计, 2025, 46 (06):121-127.
- [9] 吴雪梅。多源异构数据语义对齐方法研究 [J]. 数据分析与知识发现, 2024, 8 (03):45-53.
- [10] 林小光。工业物联网环境下实时数据采集与流式处理研究 [J]. 计算机应用与软件, 2025, 42 (07):112-118.