

汉江数字孪生系统开发的ETL技术应用探索

王 奕, 李水泉, 吴鸿虎, 张成孝

长江水利委员会水文局汉江水文水资源勘测局网络信息中心, 湖北 襄阳

收稿日期: 2025年2月19日; 录用日期: 2025年4月25日; 发布日期: 2025年6月27日

摘 要

近年来, 数字孪生技术作为一种创新的数字化解决方案, 在众多领域展现出巨大潜力。为贯彻水利部推进智慧水利建设的决策部署, 汉江数字孪生系统的开发成为汉江流域管理的重要课题。数字孪生系统开发中需要分析、整合、汇聚多种数据, ETL技术对于解决数字孪生系统开发中的部分痛点具有重要作用, 在数据相关流程中扮演着关键角色。本文介绍了数字孪生系统开发概况, 通过ETL技术应用于汉江数字孪生系统开发的案例, 深入探讨了ETL在数字孪生系统开发中的应用实践及深远影响, 并对未来发展方向进行了探索。

关键词

ETL, 数字孪生系统, 汉江流域, 数据仓库, 数据治理

An Exploration into ETL Application in the Development of Hanjiang River Digital Twin System

Yi Wang, Shuiquan Li, Honghu Wu, Chengxiao Zhang

Network Information Center, Hanjiang River Bureau of Hydrology and Water Resources Survey, Bureau of Hydrology, Changjiang Water Resources Commission, Xiangyang Hubei

Received: Feb. 19th, 2025; accepted: Apr. 25th, 2025; published: Jun. 27th, 2025

Abstract

In recent years, digital twin technology, as an innovative digital solution, has demonstrated huge potential in numerous fields. To implement the decision-making arrangements of the Ministry of Water Resources for promoting the construction of smart water conservancy, the development of the Hanjiang River Digital Twin System has become an important task in the management of the Hanjiang River Basin. ETL technology

作者简介: 王奕, 女, 工程师, 硕士, 主要从事水文水资源工作, Email: 1817525178@qq.com

文章引用: 王奕, 李水泉, 吴鸿虎, 张成孝. 汉江数字孪生系统开发的 ETL 技术应用探索[J]. 水资源研究, 2025, 14(3): 287-294.
DOI: 10.12677/jwrr.2025.143030

plays a crucial role in solving some pain points in the development of digital twin systems and occupies a key position in data-related processes. This paper presents an overview of the development of digital twin systems. Through the case of applying ETL technology to the development of the Hanjiang River Digital Twin System, it delves into the application practices and far-reaching impacts of ETL in the development of digital twin systems and explores the future development directions.

Keywords

ETL, Digital Twin System, Hanjiang River Basin, Data Warehouse, Data Governance

Copyright © 2025 by author(s) and Wuhan University & Bureau of Hydrology, Changjiang Water Resources Commission. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

汉江是长江最长支流，汉江流域是南水北调中线工程重要水源地，流域的有效管理和流域资源的高效利用对于我国经济社会具有重大意义[1]。2003 年，Michael Grieves 提出“与物理产品等价的虚拟数字化表达”概念，成为数字孪生概念模型的雏形[2]。后续随着人工智能、物联网、虚拟现实等各种技术的不断发展和研究的深入，数字孪生的概念也被不断丰富和完善，其应用范围也逐渐扩大，如今已广泛应用于航空航天、工业生产、城市管理等众多领域。数字孪生系统通过构建物理实体的虚拟模型，对物理实体的各项数据指标进行实时监测、分析和优化，成为了连接物理世界和数字世界的桥梁。国家“十四五”新型基础设施建设规划明确提出，要推动大江大河大湖数字孪生、智慧化模拟和智能业务应用建设[3]，汉江数字孪生系统的开发对于汉江流域管理至关重要。

数字孪生系统的有效运行往往依赖于大量的异构数据，这些数据来源广泛(如业务系统、历史数据库、远程设备实时采集等)，数据的复杂性和多样性给数字孪生系统的开发带来了巨大挑战[4]，如通过不同厂商的设备采集获得的水文、气象、地形、工程设施等信息数据结构不一致等。ETL (Extract-Transform-Load)作为一种成熟的数据处理技术，在解决数字孪生系统的数据相关问题方面具有独特优势。它能够从多个数据源抽取数据，对数据进行清洗、转换和集成，确保加载到数字孪生系统中的数据符合系统的要求，为数字孪生系统开发提供了技术基础。

本文介绍了 ETL 技术、数字孪生系统及汉江数字孪生系统开发的 ETL 技术应用案例，深入研究了 ETL 在数字孪生系统开发中的应用实践，并对未来发展方向进行了探索，为相关领域的数字化建设提供有价值的参考。

2. ETL 技术

ETL 意为抽取(Extract)、转换(Transform)、加载(Load)，主要用于将数据从来源端经过抽取、转换等步骤，最后加载至目的端[5]。ETL 技术在数据科学、软件工程、自动化等多种领域有着广泛的应用，能够将分散、零乱、标准不统一的数据进行快捷有效的整合，为数据分析、系统开发、决策支持等应用场景提供可靠的数据基础。

2.1. ETL 技术原理

ETL 技术包含三个方面：

抽取是从各种数据源获取原始数据，例如使用结构化查询语言(Structured Query Language, SQL)从关系型数据库中提取数据、通过应用程序接口(Application Programming Interface, API)调用 Web 服务所提供的数据等。其中数据源包括关系型数据库如 MySQL、Oracle 等、非关系型数据库如 MongoDB、Cassandra 等、外部文件如 CSV 文件、XML 文件、JSON 文件等；提取方式包含一次性提取数据源全部数据的全量提取，以及仅提取发生变化

数据的增量提取[6]。

转换是对抽取得到的数据进行处理和加工，使其满足目标系统的要求，例如数据清洗、转换、验证、聚合、计算衍生字段、规范化和去规范化等。其中数据清洗包括对缺失、重复、错误等问题数据的处理；数据转换包括数据类型转换、标准化、计算和汇总[7]。

加载是转换后的数据加载到目标系统中，包含直接加载(将转换后的数据直接插入到目标存储中)、批量加载(将数据分批加载到目标存储中)、增量加载(只加载新的数据或更新后的数据到目标存储)。

2.2. ETL 流程步骤

应用 ETL 技术通常包含以下流程，如图 1 所示。

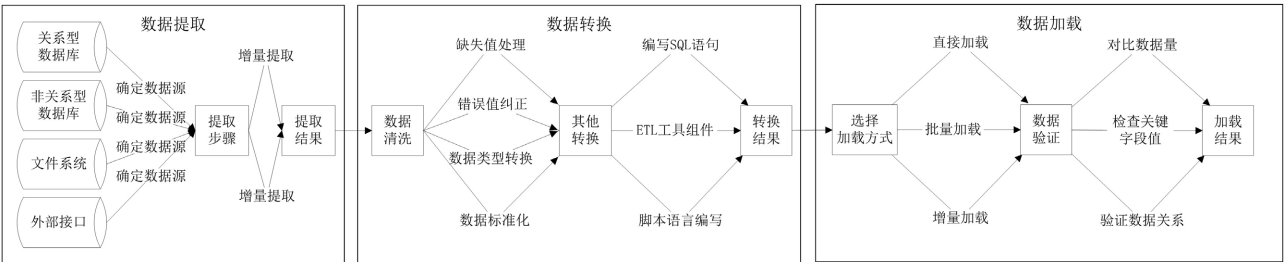


图 1. ETL 流程步骤

在数据抽取步骤中，数据来源需要首先被确定，据此选择抽取策略，在数据量不大且数据更新不频繁的情况下选用全量提取，一次性获取数据源中的所有数据；对于更新频繁且数据量较大的数据源选用增量提取，只提取上次提取之后有变化的数据，从而减少数据传输量 and 处理时间。

在数据转换步骤中，首先对抽取出来的数据进行数据清洗，包括缺失值处理(删除记录、采用默认值填充、采用均值、中位数等统计量进行填充等)，错误值纠正(对于格式错误的值可采用格式转换，对于不符合逻辑的值要根据具体业务进行转换)，数据类型转换，数据标准化(将数据的格式和编码进行统一，例如都为大写等)；大量的数据往往需要使用外部工具软件来实现转换，例如使用 SQL 的存储过程和函数；许多 ETL 工具也提供图形化的转换组件，也可以使用编程语言如 Python、Java 等来实现复杂的数据转换规则。

在数据加载步骤中，首先选择加载方式，包括将转换后的数据直接插入到目标中的直接加载、将数据分成若干批次执行加载的批量加载，以及只加载更新后数据到目标的增量加载；然后进行数据验证，包括对比数据量、数据内容验证(检查关键数据字段的值)和验证数据之间的关系是否符合预期。验证通过后，数据才能成功加载。

2.3. 主流 ETL 工具

ETL 技术的实现可以使用 SQL 的存储过程、函数和编程语言等，也可以使用现有的 ETL 工具，用户可根据业务需求选择。表 1 对这些 ETL 进行横向对比。

表 1. 主流 ETL 工具对比

分类	名称	数据支持性	使用便捷性	可扩展性	处理性能
商业	Informatica PowerCenter	★★★	★★	★★★	★★★
商业	IBM InfoSphere DataStage	★★★	★★	★★★	★★★
商业	ODI	★★	★★★★	★★★	★
商业	OWB	★★	★★	★★★	★★

续表

商业	微软 DTS	★★	★★★★	★★★★	★★
商业	Beeload	★★★★	★★★★	★★★★	★★
商业	久其 ETL	★★★★	★★★★	★★	★★
商业	udis 睿智 ETL	★★★★	★★★★	★★	★★
商业	Amazon Web Services Glue	★★★★	★★	★★★★	★★★★
商业	AWS Data Pipeline	★★	★★★★	★	★★
开源	Sqoop	★★★★	★★	★	★★★★
开源	Apache NiFi	★★★★	★★	★★	★★★★
开源	Apache StreamSets	★★★★	★★★★	★★	★★★★
开源	Apache Airflow	★★★★	★★	★★★★	★★★★
开源	KETTLE	★★★★	★★★★	★★	★★

在汉江数字孪生系统项目中，经全面评估，项目对于系统扩展性以及处理性能的要求有限，这一特性成为影响工具选型的关键因素。同时，考虑到引入全新工具所伴生的学习成本，以及严格控制项目整体成本的现实需求，经过综合分析比较，最终选定 KETTLE 作为该项目的适用工具。

3. 数字孪生系统

数字孪生早期主要应用于航空器模拟，随着信息技术的不断发展，数字孪生的概念也被不断丰富和完善，应用范围也逐渐扩大至工业制造、城市管理、医疗保健、航空航天等众多领域[8]。

3.1. 数字孪生系统的多层架构

数字孪生系统通常可以分为如图 2 所示的多层架构。物理实体层是现实世界中实际存在的物理对象，是系统的数据来源和所模拟和反映的对象基础。数据采集层是各种传感设备，用于获取物理实体的状态信息、运行参数、环境数据等。孪生数据层用于存储采集到的海量原始数据以及经过处理和分析后的数据。模型层描述物理实体的静态特征和动态行为，对物理实体的结构、属性、行为等进行建模和仿真。应用层根据不同的行业和领域，开发相应的数字孪生应用场景，实现特定的需求和功能。

数据在数字孪生系统中处于核心地位，数据从物理实体层开始产生，通过数据采集层准确、及时地进入系统，经过孪生数据层进行存储和治理，为模型层的构建提供依据，最后应用层通过对数据的处理和展示提供给用户系统的各项功能[9]。

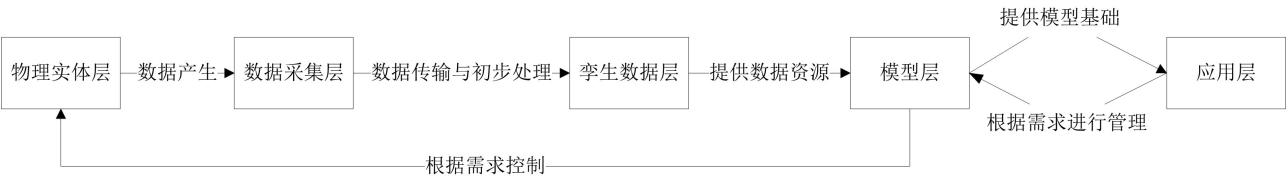


图 2. 数字孪生系统的多层架构

3.2. 汉江数字孪生系统案例

汉江数字孪生系统通过整合多种数据资源，包括水文、气象、地形、工程设施等信息，构建了一个与真实汉江流域一致的虚拟模型，该模型能够实时反映流域的状态变化。系统架构如图 3 所示。

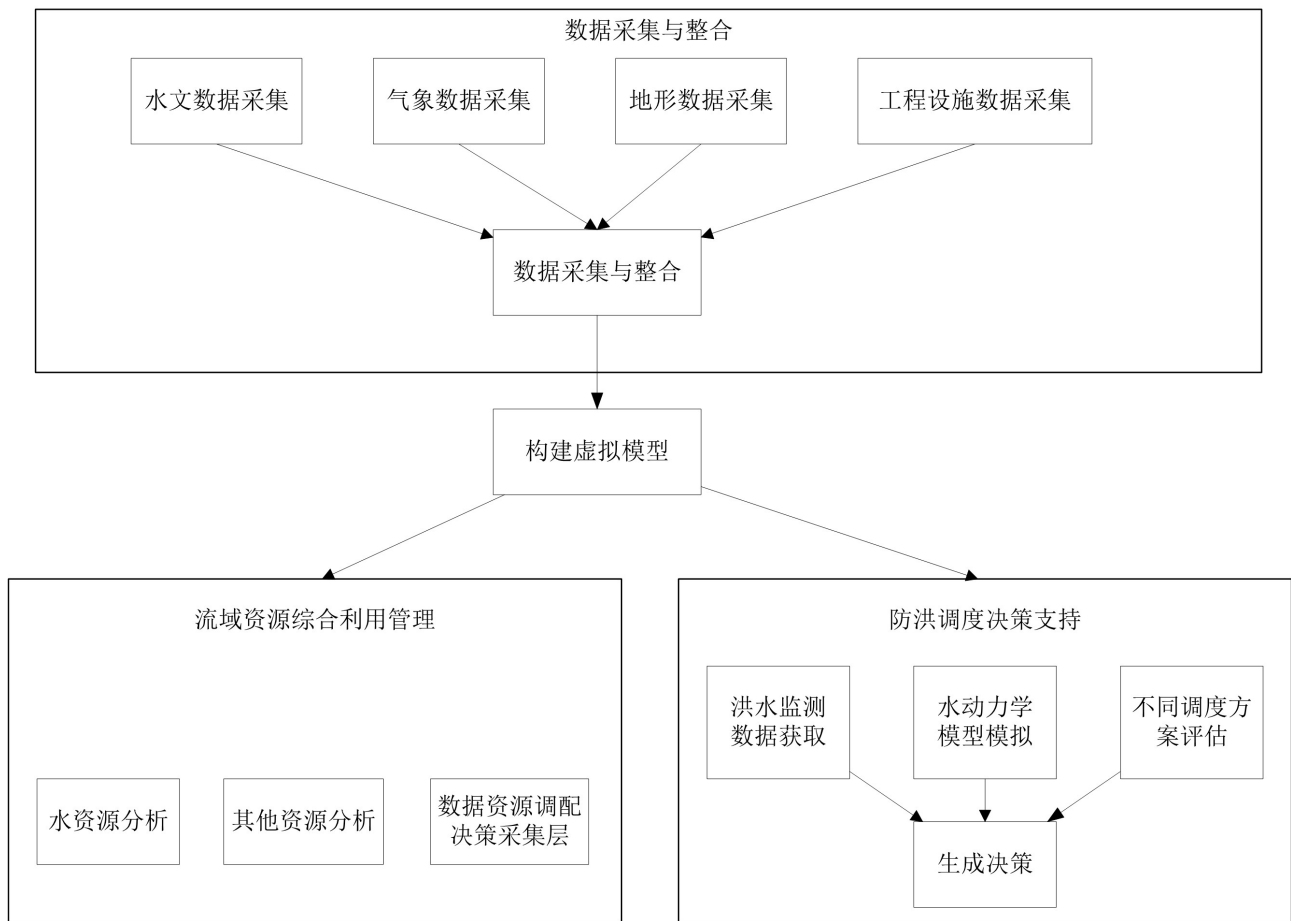


图 3. 汉江数字孪生系统案例

数据采集与整合：利用高精度卫星遥感和航空摄影测量技术构建了流域的三维地形模型，分辨率可达 1 米左右。对于坝、船闸等工程设施，通过激光扫描和建筑信息模型(Building Information Modeling, BIM)技术，构建了精细的设施模型。对于汉江流域各水文站点的水位、流量、泥沙含量等实时监测设备，通过定时数据传输获取相应的数据，精度可达厘米级。同时，系统接入了周边多个气象站和气象卫星的数据，共同为流域内的降雨-径流模型提供输入。

流域资源综合利用管理：流域资源综合利用管理是指以流域为基本单元，对流域内的水资源、生态资源、土地资源等各类自然资源进行统筹规划、协调和管理。汉江流域的水资源利用情况被全面地进行监测和分析。以实时获得计算水库的蓄水量、可供水量等指标，结合流域内的用水需求，系统可以进行水资源的合理分配。流域船闸的运行情况得到实时监测，系统通过传感器获取船闸的水位变化、闸门开合状态、船舶进出闸时间等信息，从而合理调度航运资源。

防洪调度决策支持：在洪水期，系统能够根据实时的水文和气象数据，快速模拟洪水演进过程。例如，当上游来水量增大时，系统通过水动力学模型，计算洪水在水库以及下游河道中的传播时间、水位涨幅等情况。系统可以进行不同防洪调度方案的模拟对比，在不同泄洪流量情况下，对水库水位、下游防洪控制点水位以及淹没范围的影响进行对比分析。假设将某关键水库的泄洪流量设定为每秒 10,000 m³和每秒 12,000 m³两种情况，系统会分别模拟出相应的水位变化过程和下游可能的受灾区域，为决策者提供科学的调度依据，确保在防洪的同时，最大程度地减少对下游地区的不利影响。

4. ETL 技术应用

4.1. 技术选型

平台支持：不同组织的众多用户可能使用多种操作系统，需要选择一款跨平台支持度高的 ETL 工具。由于一些高精度模型数据可能存在保密问题，数字孪生系统暂不考虑云平台兼容性。ETL 工具需支持标准接口以实现系统集成，以适应数字孪生系统未来功能的扩展。

数据源支持：ETL 工具应该能够处理不同类型数据库中的数据，还要能够处理部分文件格式的数据。除了专门服务于某些大型数据库管理系统的 ETL 工具，大部分 ETL 工具都具有广泛的数据源支持性。

转换功能：ETL 工具需要能够处理数据中的错误、重复和不完整的数据，如数据格式、数据类型等。同时，ETL 工具还应该支持复杂的转换，如对数据进行聚合操作(计算平均值、总和等)和分组等操作。目前主流 ETL 工具大部分具有较好的数据转换功能，一些高级的 ETL 工具还可以整合外部数据来丰富内部数据，例如根据用户的出生日期计算用户的年龄等。

运行性能：ETL 工具应该能够对 ETL 任务进行有效的管理，例如记录并展示每个任务的执行状态、按照预定的时间表来执行 ETL 任务等。对于汉江数字孪生系统，还需要支持复杂的调度策略，如根据数据的更新频率来调整任务执行时间；或者在特定事件(如数据量达到一定阈值)发生时触发任务。在处理海量数据时，需要采用并行抽取、批量加载等技术来提高效率。ETL 工具应该尽量减少对业务系统的干扰，例如在业务系统的低峰期进行数据抽取操作，或者采用增量抽取的方式，避免对业务系统的性能造成影响。

4.2. 应用实践

数据抽取：针对 BIM 模型，抽取其中关键的几何结构、材质特性、空间位置等数据信息。由于各类水文要素的原始数据具有异构性特征，具体表现为：① 数据存储位置分散，可能分布于不同物理主机的数据库系统中；② 数据库结构存在显著差异，包括表结构设计、字段命名规范等方面的不一致性。基于此，汉江数字孪生系统中的文监测数据的抽取过程需针对不同水文要素(如水位、流量等)分别进行。对于水文监测站点的实时监测设备，需要考虑不同设备的数据传输协议和存储方式进行数据抽取，例如从汉江上下游的多个水位站通过卫星通信和光纤网络等多种通信方式，按照每五分钟一次的固定频率抽取实时水位数据；利用水平式 ADCP 进行在线流量测量，在每次测量数据有更新时进行数据抽取。系统接入了周边多个气象站和气象卫星的数据，以获取全面的气象信息用于流域管理。过程详见图 4。

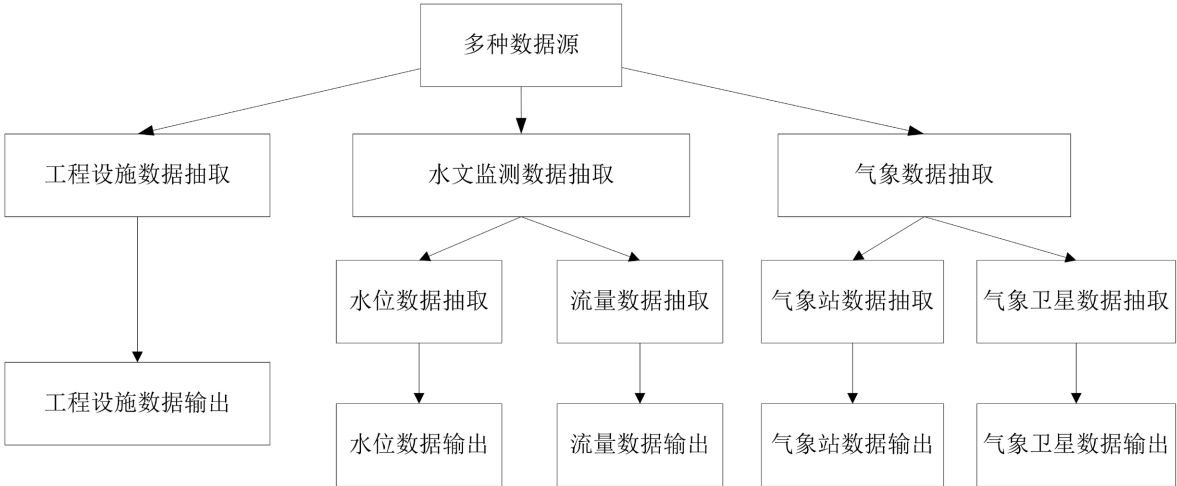


图 4. 汉江数字孪生系统数据抽取

数据转换：采集到的数据可能存在一些不准确的数据点或异常值，需要对其进行清洗来去除，同时对剩余数据进行预处理，确保后续分析和模拟的准确性。对水文监测数据，需要清洗掉异常水位值、流量值，同时对正常的数据进行预处理，为水资源的合理分配、洪水演进模拟等提供准确的数据基础。对气象数据，由于气象站设备故障或局部气象条件异常导致的过高或过低的温度、降雨量等数据也需要去除后对剩余数据进行预处理，确保气象数据的准确性和可靠性。为了使不同数据源能够在汉江数字孪生系统中进行有效的整合和分析，需要对各类数据进行格式统一与标准化转换。将通过卫星遥感获取的地形地貌特征数据、通过激光扫描和 BIM 技术获取的工程设施数据等，按照系统设定的通用地理信息数据格式进行转换，使得不同来源的数据能够进行准确的叠加、分析和模拟。详见图 5。

数据加载：将经过抽取、转换后的各类数据，需要按照系统数据库的结构和存储要求加载到汉江数字孪生系统的数据库中。为确保数据的完整性和准确性，根据不同类型的数据库设置不同的存储表和索引，以便于后续快速查询和调用。系统还需要进行数据关联与整合加载，使得系统数据库中的数据形成一个有机的整体，为系统的流域资源综合利用管理、防洪调度决策支持等各种功能的实现提供全面、准确的数据支持。数据加载详见图 6。

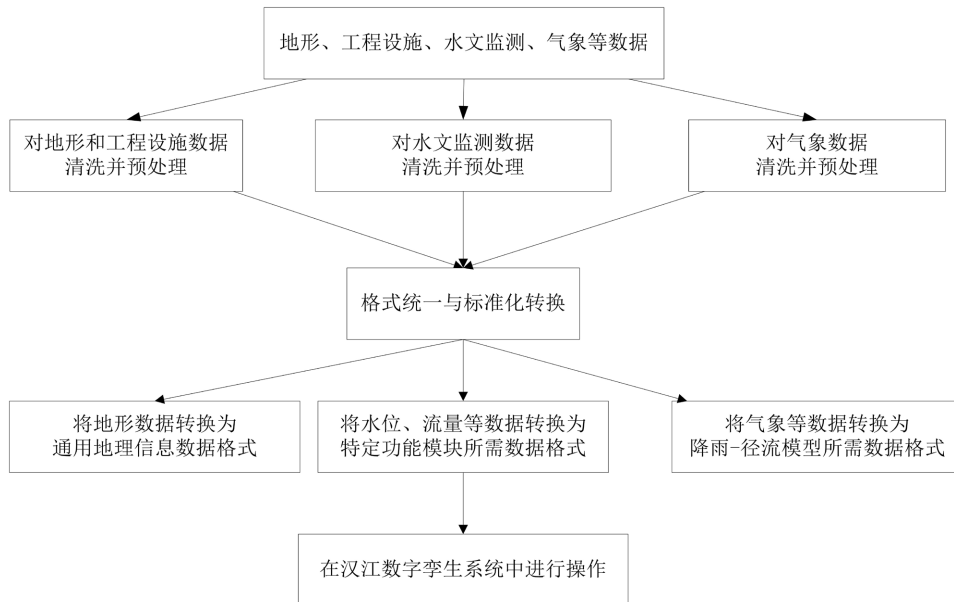


图 5. 汉江数字孪生系统数据转换

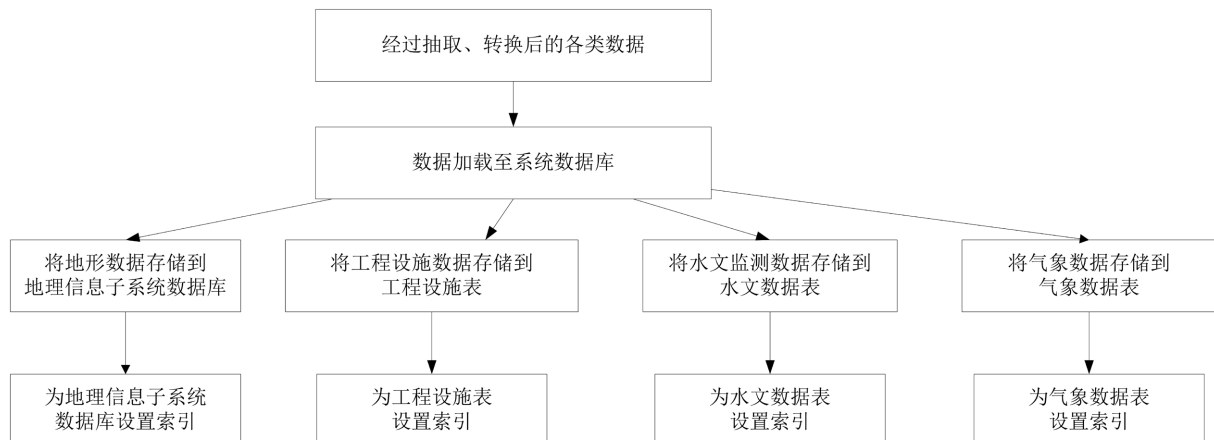


图 6. 汉江数字孪生系统数据加载

4.3. 未来方向

自动化与智能化：ETL 工具将能够自动发现数据源、自动完成数据的 ETL 过程，减少人工干预，提高数据处理效率和准确性。例如，通过机器学习算法自动识别数据中的模式和异常，智能地进行数据清洗和转换，以及自动生成 ETL 脚本和工作流。ETL 工具将借助自然语言处理技术，实现与用户的自然语言交互，增强对现实世界的感知和理解能力，应用于历史水情数据挖掘、实时水情监测报告智能解析、水利工程巡测报告智能处理等典型场景中。

实时数据处理能力增强：数字孪生系统对实时数据的需求日益迫切，未来 ETL 将具备更强的实时数据处理能力，能够对流式数据进行实时抽取、转换和加载，确保数字孪生系统能够及时获取最新的数据洞察，并基于实时数据做出快速决策，提升系统的实时性和响应能力。随着物联网技术在水利监测中的普及，传感器产生的数据量呈指数级增长，边缘计算通过将计算能力下沉至数据源头(如传感器、网关设备)，通过将 ETL 功能前移至数据采集端，为水利监测系统提供了低延迟、高能效的实时数据处理能力。

数字孪生技术协同发展：ETL 将作为数字孪生系统的数据预处理环节与数字孪生的建模、仿真、分析和优化等技术紧密结合，共同推动数字孪生系统的发展和应用，并不断适配各领域的特定需求，如在智慧城市、智能制造、医疗健康等领域，提供针对性的数据处理解决方案，促进数字孪生技术在各行业的落地和推广。

5. 结论

数字孪生已成为连接物理与数字世界的桥梁，数字孪生系统运行依赖的大量异构数据带来了挑战。通过对汉江数字孪生系统开发中 ETL 应用案例的研究，实现了单传感器增量数据 1 秒内抽取完成、全量数据以每秒千条的速率抽取。为流域管理数字孪生系统开发提供了有益借鉴，也为其他相关领域数字化建设在数据处理与系统构建方面提供了参考，有助于推动数字孪生技术在更多场景下的有效应用与持续创新。ETL 在汉江数字孪生系统案例中的应用体现了一定的普适性，其整合异构数据的方式适用于大部分流域；但对于一些跨国境的流域，其数据归属复杂难以整合，部分基础设施的投资也不适用于一些较小流域。未来还需进一步探索 ETL 技术与数字孪生系统深度融合的更多可能性以及应对新兴技术挑战的策略。

基金项目

SWJ-24CJX19 基于 WebGIS 技术的暴雨洪水分析计算系统研发。

参考文献

- [1] 王欣贝, 彭涛, 由星莹, 等. 丹江口水库蓄水前后汉江中游水文干旱演变特征分析[J]. 长江流域资源与环境, 2023, 32(7): 1479-1490.
- [2] GRIEVES, M. W. Product life cycle management: The new paradigm for enterprises. International Journal of Product Development, 2005, 2(1): 71-84. <https://doi.org/10.1504/ijpd.2005.006669>
- [3] 李国英. 建设数字孪生流域推动新阶段水利高质量发展[J]. 水资源开发与管理, 2022, 8(8): 3-5.
- [4] 冯薇玺, 李清, 黄萍. 基于数字孪生的多信息源输电网数据融合方法[J]. 自动化与仪器仪表, 2023(9): 96-99.
- [5] 游骏鹏. 基于数据仓库的企业决策支持系统的设计与实现[D]: [硕士学位论文]. 北京: 北京交通大学, 2020.
- [6] 陈卓, 汪灏, 平佳伟, 等. 面向火箭数据分析和故障诊断的数据仓库设计[J]. 计算机测量与控制, 2024, 32(1): 51-56.
- [7] 崔有文, 周金海. 基于 KETTLE 的数据集成研究[J]. 计算机技术与发展, 2015, 25(4): 153-157.
- [8] 李志强, 金国胜, 杜王特. 基于数字孪生技术的变电站机电设备状态检测系统[J]. 自动化技术与应用, 2024, 43(11): 43-47.
- [9] 陈炫江, 黄火柴. 以数字孪生为目标的 BIM 数智设计与成果质量管理方法研究[J]. 智能建筑与智慧城市, 2024(11): 12-14.