基于元数据分析实现数据运营态势感知智能化的研究

Research of Achieving Intelligent Data Operations Situational Awareness Based on Metadata Analysis

陈新亮，孙而焓，林理直，欧胜昶，王艺斌（中国联通福建分公司，福建 福州 350002）
Chen Xinliang，Sun Erhan，Lin Lizhi，Ou Shengchang，Wang Yibin（China Unicom Fujian Branch，Fuzhou 350002，China）

摘　要：
数据中台全域数据集约化体现为数据规模大、关联关系复杂、服务保障要求高等运营特点，如何高效地运营管理成为一个重要挑战。提出了一种利用元数据分析构建数据生产运营态势全流程智能化感知平台的方法，以实时可视化方式在一个体系中对资源、数据、程序、流程、应用服务全纳管，具备问题预警、自动定位、态势自动恢复等能力，显著提升数据服务连续可用率及使用感知。

关键词：
元数据；智能化；态势感知；实时可视化
doi：10.12045/j.issn.1007-3043.2025.03.012
文章编号：1007-3043（2025）03-0065-06
中图分类号：TN915
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The operational characteristics of full-domain data centralization in the data center include large data scales, complex relationship networks, and high service assurance requirements. Efficient operation and management becomes a significant challenge. A method is proposed for constructing an intelligent situational awareness platform for the full-chain data production operations using metadata analysis. This method provides real-time visualization for comprehensive management of resources, data, programs, processes, and application services within a single system. It has capabilities such as issue prediction, automatic fault localization, and automatic situational recovery, significantly improving constant availability of data service and user perception.

Keywords：
Metadata; Intelligenization; Situational awareness; Real-time visualization

引用格式：陈新亮，孙而焓，林理直，等. 基于元数据分析实现数据运营态势感知智能化的研究[J]. 邮电设计技术，2025(3)：65-70.

0 引言

随着数据要素在金融、电信等各个行业经营管理中的生产力价值凸显，以及大数据平台体系架构的不断成熟，基于数据集约化的中台运营体系也在不断发展。数据中台是企业级公共的、可复用的数据及其衍生能力的组合[1]，规模庞大、关联复杂，传统运营管理在自动化智能化方面面临不少困难和挑战，包括各类数据要素的治理及运营信息未有效组织在一起，缺乏自动化汇聚分析能力；资源/程序/数据/流程/服务等各个要素间的关联关系复杂，缺乏智能化多级血缘识别，无法精准标定要素价值，难以对算力储力等资源占用做减法；生产运营态势实时感知能力弱，无法高质量保障面向应用的数据服务连续可用率、缺乏问题预判及自动定位能力，无法对前端生产应用进行提前预警及提供态势自动恢复处理、状态跟踪等能力。

收稿日期：2025-02-16

1 总体设计

针对数据集约化后生产运营面临的挑战，提出一个利用企业数据中台的高成熟度数据治理方案，从元数据分析的角度，以软件工程方法构建一个包含全量要素静态信息及全链条生产运营实时动态信息的自动化处理体系，实现了数据规范统一(包括数仓及元数据)及软件体系统一(包括云化及容器化)，具有良好的开放性及可移植性[2]。主要研究内容如下。

a) 云端采集数据中台能力开放接口上的各类元数据，并对数据进行自动分类处理，在前端以数据可视化的方式对处理结果进行渲染，形成数据要素基本运营信息聚合。

b) 对元数据映射关系、日志数据等进行实时自动化分析处理，对各要素间的血缘依赖关系进行解构，形成完整的逻辑关联路径识别。该项内容是建立生产运营实时态势感知能力的基础，同时也为后续复杂的从应用服务端溯源进行的全链条的要素使用冷热度识别、价值标定、清算策略等资产资源动态化管理奠定基础。

c) 各类元数据实时分析处理的结果同步用于构建全流程实时态势感知能力，由软件输出数据中台生产运营全流程实时运营信息，自动判断生产运营节点及数据流状态，同时利用血缘路径识别形成问题预判预警及自动定位、自动态势恢复等能力。

2 关键模块设计

2.1 数据采集及处理

数据中台的实时运营信息聚合主要包括资源、程序、流程、数据(如标签、模型等)、应用服务等信息，需要从能力开放平台、数据运营平台、应用平台等实时采集各类元数据及日志数据进行分析处理[3]，具体如图1所示。

根据数据源的提供方式不同，采集方式可分为API采集、消息中间件采集、文件采集3种。文件方式较为简单，本文重点讨论API及消息中间件这2种效率较高的方式。

2.1.1 API方式

API方式主要获取配置信息，如模型类信息中的表信息、字段信息等。API信息的获取一般包括：请求方式，包含post、get、put等方式；请求地址，完整的url地址信息；请求头部信息header，通常包含一些鉴权的信息、响应的数据格式等；请求参数信息，一般包含json、x-www-form-urlencoded、form-data、xml、raw等；返回结果信息，一般包含返回状态、返回数据。

以采集流程类节点依赖信息元数据为例，该数据使用post请求方式，header包含相应格式为json以及鉴权登录信息；请求参数使用json格式，使用流程编码与节点编码作为查询请求；返回信息为json格式，返回内容包含节点编码、节点名称、节点类型、依赖事件名称、依赖事件其他编码、依赖事件编码、触发事件名称等业务使用信息。

2.1.2 消息中间方式

以KAFKA为例，消息中间方式主要获取实时信息，如流程执行日志、映射执行日志等。可使用Flink+Yarn+Redis+OceanBase作为实时消息落地技术方案，使用实时流开发平台作为底座，主要需要关注KAFKA消息源端信息获取[4]以及落地过程中的关键参数。

以流程执行日志为例，需要获取源端KAFKA地址信息、分组消息(groupsf_mysql_caiji2)、主题信息

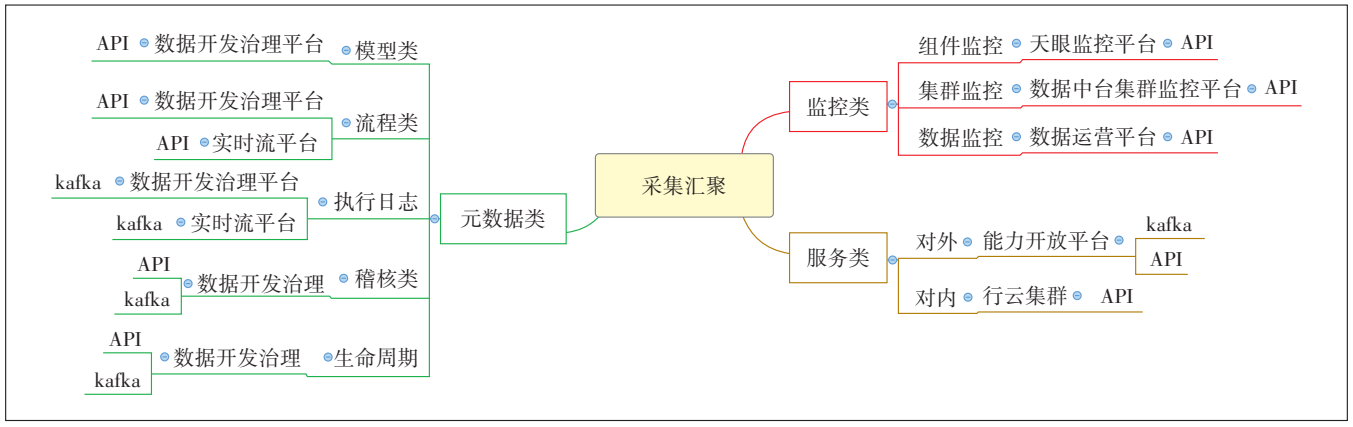

图1 数据采集示意

topic（sf_mapping）、格式信息 format（json）、安全认证 protocol（SASL_PLAINTEXT）、消费偏移量 group-offset（earliest-offset）；目标端连接引擎 connector（jdbc）、地址信息 url、表编码（SRC_ZB_H_DAM_RUN_FLOW_LOG）、缓存记录信息 sink.buffer-flush.max-rows（1条）、缓存时间间隔 sink.buffer-flush.interval（1 s）等信息。最终获取实时消息包含流程名称 WOLKFLOWNAME、映射名称 NODENAME、映射属组 MAPSCHEMA、实例值 INSTANCEVALUE、表名称 TABLENAME、资源名称 DSTRESOURCECODE、开始时间 STARTTIME、结束时间 ENDTIME、任务状态 TASKSTATUS、项目编码 PROJECTID，以为后续数据流程实时状态跟踪提供基础数据支持。

2.2 运营信息实时聚合分析

从数据要素生产运营角度出发，建立分域分类型的生产运营数据处理、分析、加工、展现的一站式统揽能力，将复杂的数据中台的各类资源、程序、流程等全量要素信息整合在一起[5]，提供实时信息分析能力，满足高效管理的需求，为构建智能化的生产运营态势感知及自动化态势恢复等能力奠定基础。

2.2.1 资源、程序、数据等资产类信息聚合

本地可被调用的数据类资产包括集群算力储力、数据源、数据湖、模型、标签、实时数仓、程序资产等要素信息，体现B域、O域、M域等的分域信息、数据获取途径及数量等。数据模型可按DWD/DWA/DIM进行分类，数据标签按不同的场景以视图方式进行分域，实时数仓按实时模型、实时标签进行划分，涉及到数据加工的程序资产包括实时处理程序、定时处理程序、工具类程序，需标识出系统来源[6]。

2.2.2 数据流程信息聚合

数据流程信息聚合包括数据入湖以及数据推送，具体包括基础模型加工、萃取模型加工、标签生成以及数据推送4类；前端数据应用部分也会对数据中台的数据流调用信息进行整合，包括所有前端应用服务涉及的数据流；数据流的分析结果需要输出运行状态、延迟预估2类分析数据。

2.2.3 数据服务信息聚合[7]

数据中台提供各类数据库、API调用、SFTP调用等不同类型的服务信息，通过信息聚合可以看到不同的数据服务的调用冷热情况，包括TOP模型、TOP数据服务接口等，可从服务类型（模型、API、文件）、服务时效（实时、离线）以及服务范围（内部，外部）3个维度进行聚合。

2.3 数据要素血缘路径关系识别

资源—数据—程序—流程—应用服务调用这一链条具有紧耦合特性，关联关系复杂，互相嵌套依赖，本文从元数据分析角度构建自动化的数据路径图谱识别能力，解构关联关系。血缘路径识别一是可快速定位运营中可能出现的问题点及关联到的影响点[8]，并对后续态势恢复过程中的全流程节点进行状态跟踪；二是可从应用端调用方面自上而下地形成一个评估价值链，即服务价值→数据价值→程序价值→流程价值→资源价值等，通过传导对全链低效无效要素资产做减法，优先保障高效高价值要素资产，从而实现动态化管理。

血缘依赖关系自动化识别的逻辑是根据统一数据治理平台沉淀的程序/流程/数据/资源/应用服务等元数据映射关系原始信息表，构建一个软件体系，并对这部分元数据映射关系进行解析，形成所有要素间的关联关系表。从图2可以看出，在数据治理平台沉淀的应用—流程—程序—数据—资源的映射关系均可被识别，数据映射可细分到表级或字段级。

元数据采集自动处理后，通过映射关系识别构建关联关系表，并放入云端数据库保存。一个典型的识别路径是，当链条上的任意节点信息出现异常时，自动对全链中受到影响的血缘进行评估（见图3）。当某个日收入指标发生计算异常时，系统则按图3中红色箭头指向字段迅速地评估影响范围和处理路径，并自动监控链条节点的处理状态，通过流程映射关系形成链式的自动恢复处理机制。该项能力为构建全体系的数据中台生产运营态势实时感知奠定了基础。

2.4 生产运营态势实时感知

2.4.1 生产运营实时及离线数据流全纳管

数据中台生产运营包含实时数据流及离线数据流，应用端对2类数据流的调用是实时在线的[9]，所以运营态势感知在设计上需要确保实时响应。云端持续扫描各类数据信息变动、数据库日志变化、所有数据流程节点处理状态变化、状态恢复检测、任务基线时间点检测等，识别异常情况后进行秒级响应，再根据血缘识别信息对所有受影响的依赖流程、应用服务等进行预警，在前端实时输出异常监测结果预警、异常数据流节点自动定位、态势恢复状态自动监测及处理进度更新、预估恢复时间点等自动化内容。图4给出了数据要素生产运营实时态势感知输出示例。

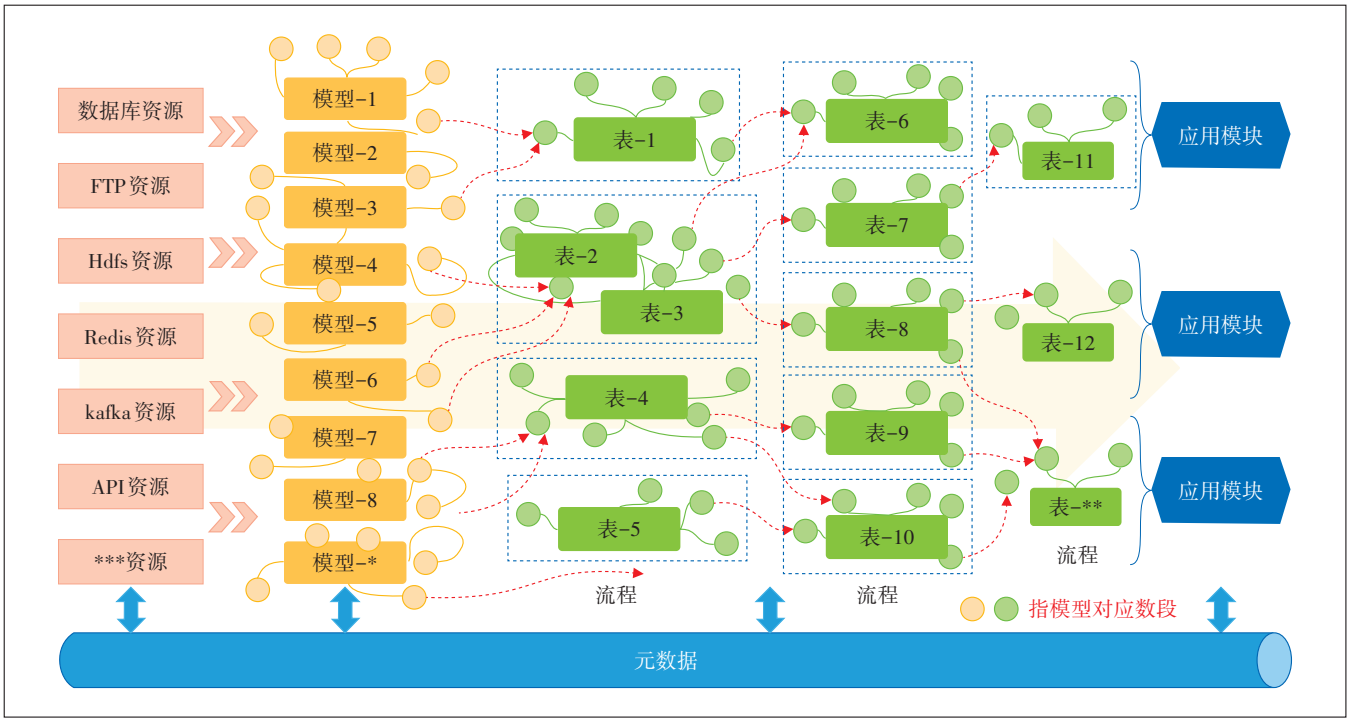

图2　通过元数据映射解构要素间关联关系

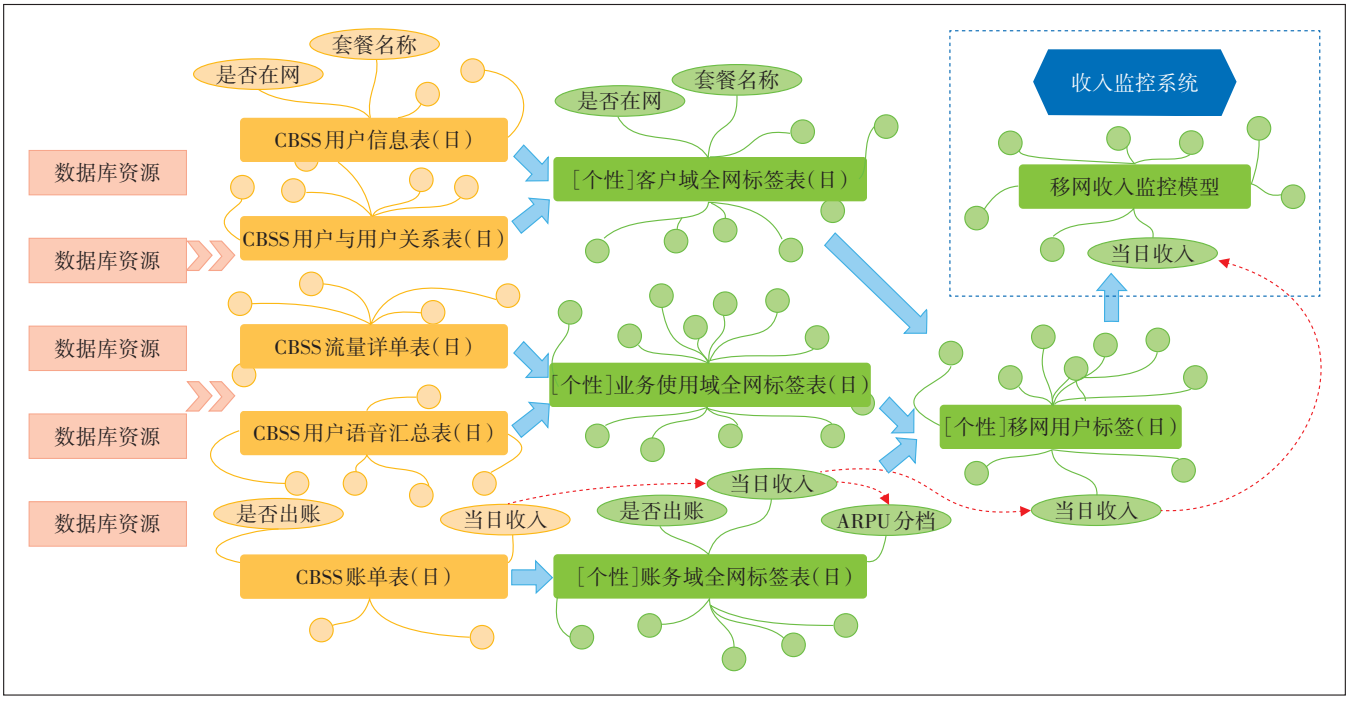

图3　异常节点形成血缘路径遍历示例

2.4.2 模型自动恢复处理预估完成时间计算

态势感知的一个重要内容是预测生产应用端的输出流程节点什么时间能恢复，这项设计将对生产应用端使用数据的感知带来很大提升，须设计一套算法进行估算。依托实时采集的流程执行日志、流程依赖事件，根据态势感知中所有模型清单对应的血缘关系图谱，按照拓扑排序后，形成加权有向无环图（Directed Acyclic Graph，DAG）[10]，调用最长路径算法

图4　数据要素生产运营实时态势感知输出示例

(Longest Path Algorithm)得出到达每个目标模型所需的运行分钟数。对于由某个根节点数据延迟等原因导致的全链运行延迟的情况，有的数据运营平台可能会在恢复时对作业组进行并发控制，此种情况下需要根据历史故障时间日志对实际运行时长进行适当纠偏，以做出更准确的估算。

在此过程中，通过时间计时器实时判断关键路径上各流程的执行状态，结合当前时间与预估完成时间的关系，加入防跳跃机制，从而使预估完成时间呈现倒计时效果。

以图5所示的流程为例，当Flow-2发生异常时，到达目标流程Flow-18的加权最长路径为Flow-2→Flow-5→Flow-7→Flow-12→Flow-17→Flow-18，预估完成时间在Flow-2后的15 min(计算过程中所述的加权值即流程运行分钟数)。如果Flow-2因某个前置节点延迟，预计在10:00完成，则Flow-18将在10:15完成。

3 研究成果验证

研究成果于2023年7月在某省联通的数据中台上进行了全量数据要素生产运营全流程纳管的测试运行，验证了该体系在人力效能、数据服务保障能力、资源利用效能提升方面的成效。

3.1 人力效能

由于信息聚合后态势信息分析、监测及流程恢复处理、状态更新等大部分功能均为系统在云端自动化执行，节省了原先大量工作信息的人工交互及处理，数据中台专职生产维护及任务调度人员从3人减少为1人，运营管理整体工效提升超过200%。

3.2 数据服务保障能力

在对3个集群资源、约7 000项标签模型、3万个程序、4 700项数据流程、23类前端生产应用、8 300项接口服务进行全部纳管后，该体系实现了数据流运行节点状态的全自动化检测，可细到每个应用中某个模块，对较大问题的判断及响应从小时级提升至分钟级，对前端生产应用进行预警的能力覆盖率从0达到100%，无需等前端数据应用或使用人员反馈问题后再进行处理；全量基线流程任务完成及时率从96.3%提升至99.1%，数据服务连续可用率(含应用集群、API及SFTP等所有数据服务)从97.2%提升至99.3%。

3.3 资源利用效能

基于数据要素生产应用迭代更新速度快的特点，从应用服务层进行溯源，对全链条的要素使用冷热度识别、价值标定、资源清算进行测试，嵌入血缘映射关系识别[11]。根据对较复杂样本模型的测试验证(某项大业务模型共涉及约280个标签数据应用)，1个识别周期(暂定为6个月)可实现对10%~15%的低效无效数据资产的识别，从而对相关的算力储力资源进行循

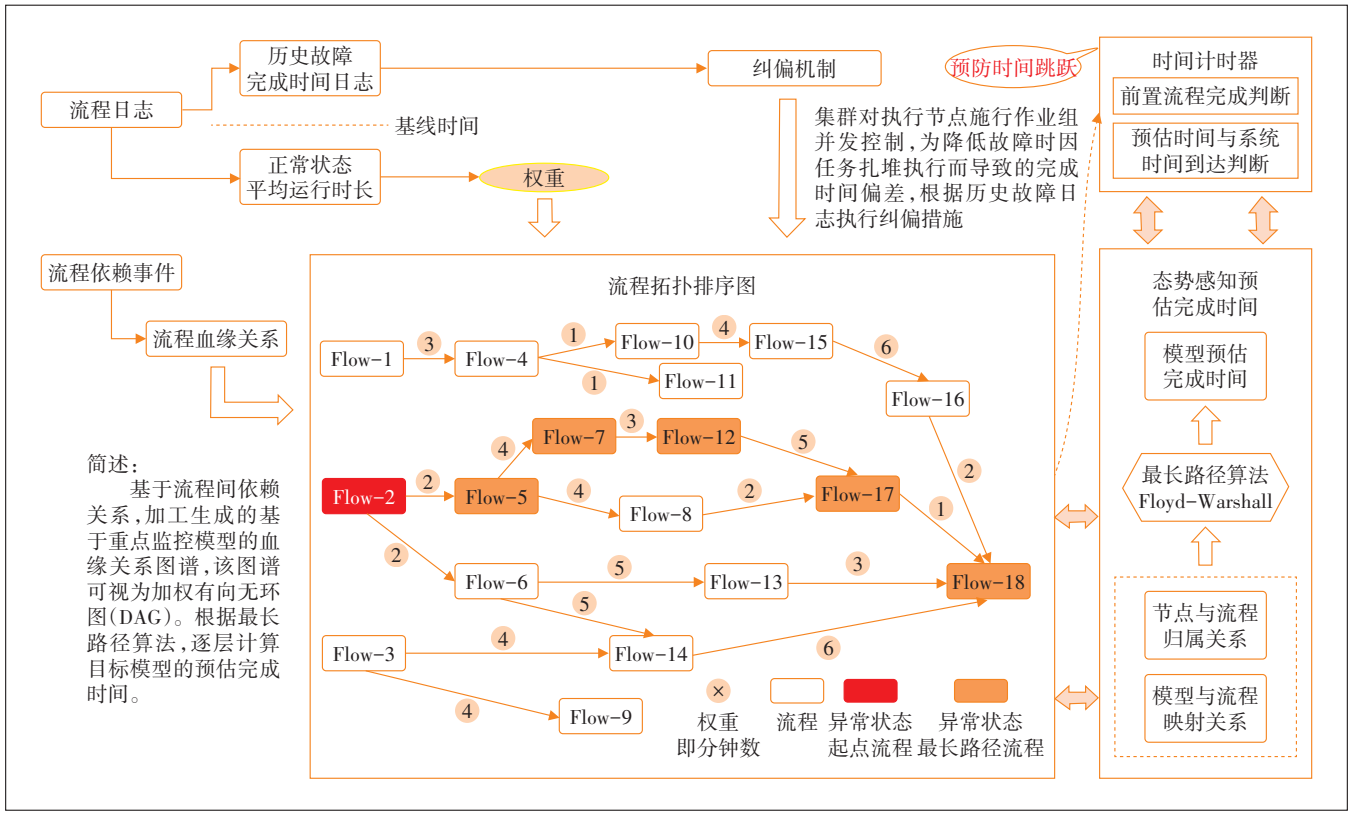

图5　模型自动恢复处理预估完成时间的计算逻辑示例

环再利用，后续将在企业数据要素资源动态管理方面发挥重要作用。

4 结束语

基于元数据分析的数据运营态势感知研究，由于其实时性、自动化、智能化等特点，对当前通信运营商及其他行业中正在大规模推进数据中台集约化运营，特别是对希望建立体系化的数据要素生产运营全流程纳管、数据要素资源动态管理、生产运营态势实时感知、实现资源/人力高效能、数据服务高质量、运营管理高度自动化的企业或项目，具有一定的参考价值。未来将不断完善态势预判及自动恢复、数据价值分析等方面的相关算法，持续提升智能化能力。

参考文献：

[1] 张洁，许建宏，肖伟. 关于数据中台建设思路的探讨[J]. 邮电设计技术，2021(8)：74-79.
[2] 孙舒颖，郭树行. 面向中台架构的数据资产运营模型研究[J]. 中国高新科技，2021(23)：43-44.
[3] 奇点云. 浅谈元数据采集[EB/OL]. [2024-10-21]. https://blog.csdn.net/StartDT/article/details/125911882.
[4] 吴文炤，李炳森，聂玲，等. 基于人工智能的元数据关系研究[J]. 电力信息与通信技术，2022，20(9)：43-50.
[5] 徐葳. 大数据技术及架构图解实战派[M]. 北京：电子工业出版社，2022.
[6] 宋春涛，张帆，王勇，等. 电信运营商的数据资产综述：数据、内联及外延[J]. 邮电设计技术，2019(9)：20-24.
[7] 李柯. 基于Flume、Kafka的日志采集系统分析研究[J]. 电子技术与软件工程，2022(10)：255-258.
[8] 朱青. 基于数据血缘分析的电网企业数据资产价值评估[J]. 信息与电脑（理论版），2023，35(13)：91-93.
[9] 唐雪飞，樊治强. 基于元数据映射关系的结构化数据血缘分析方法[J]. 现代电子技术，2022，45(16)：67-70.
[10] 黄凯，章铖. 一种基于大数据的可视化数据治理平台的研究[J]. 电子制作，2022，30(6)：36-38，23.
[11] 吴孟泽. 有向无环图在构建Logistic预测模型中的应用研究[J]. 科学技术创新，2023(3)：63-66.

作者简介：
陈新亮，工程师，学士，主要从事数据中台相关规划及设计研发工作；孙而焓，工程师，学士，主要从事大数据体系架构相关设计开发工作；林理直，工程师，学士，主要从事数据治理及数仓相关设计开发工作；欧胜昶，高级工程师，硕士，主要从事数据运营体系相关工作；王艺斌，工程师，学士，主要从事数据运营体系相关工作。