基于云－边－端的多源异构大数据治理架构研究*

许　政，阮西玥，陈祥浩
（中航机载系统共性技术有限公司，江苏　扬州　225000）

摘　要：随着航空机载产品制造过程中数字化程度的不断提升，多源异构工业大数据高速增长，这些实时与非实时交融的大数据对系统的数据管理能力提出了更高的要求。设计了一种基于云－边－端的多源异构数据治理架构，以提升数据传输和管理效能为目标，重点围绕数据采样同步机制、边缘数据治理、云端数据治理等方面进行功能设计，通过实验验证了架构的可行性和可用性，能支撑云、边、端各类节点的差异化数据应用。

关键词：云－边－端；多源异构；大数据治理；架构设计
中图分类号：TP301.6　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097－1788.2024.12.007
引用格式：许政，阮西玥，陈祥浩. 基于云－边－端的多源异构大数据治理架构研究［J］. 网络安全与数据治理，2024，43(12)：47－53.

Research on multi-source heterogeneous big data governance architecture based on cloud-edge-end

Xu Zheng，Ruan Xiyue，Chen Xianghao
（AVICAS Generic Technology Co.，Ltd.，Yangzhou 225000，China）

Abstract：With the continuous improvement of digitization in the manufacturing process of aviation airborne products，the massive multi-source heterogeneous industrial data grows geometrically，and these real-time and non-real-time intermingled big data put forward higher requirements for the data management capability of the system. In this paper，a multi-source heterogeneous data governance architecture based on cloud-edge-end is designed to enhance data transmission and management performance，focusing on data sampling synchronization mechanism，edge data governance，cloud data governance and other aspects of the functional design，and the feasibility and usability of the architecture is verified through experiments，which can support the differentiated needs of various node data applications in the cloud-edge-end.

Key words：cloud-edge-end；multi-source heterogeneous；big data governance；architecture design

0　引言

随着航空制造业数字化转型进程的高速推进，制造产线数字化程度不断提升以及传感技术广泛应用，海量类别多样、复杂度高、质量参差的实时数据、非实时数据，以及结构化数据和非结构化数据呈几何式增长，数据中心的网络资源、计算资源和存储资源都面临巨大的压力，通信网络和服务器资源的建设进程已难以满足当前航空制造数字化业务的信息联通和算力需求。

云计算严重依赖网络，计算资源过于集中，灵活性不足，面对系统实时性和可靠性要求较高的应用场景时，服务能力不足[1]；边缘计算作为云计算的补充，能够在数据源头完成一定规模的数据分析处理，具有网络时延小、实时性好的特点[2]；因此，云－边－端作为一种新型的数据治理架构被提出[3]，它结合了云计算和边缘计算的优点，运用云－边资源同步[4]技术，以最低的成本满足用户低延迟服务的需求。

目前，云－边－端数据治理架构已广泛应用于电力系统[5]、车联网[6]、物联网[7]、智能交通[8]等领域，相关研究主要聚焦于系统架构优化[9]、传输时延及带宽降低[10]、数据质量提升[11]等方面，主要目标是减少资源同步时延、降低系统能耗以及提高用户体验。

基于此，本文面向航空机载产品制造场景，提出一种基于云－边－端的多源异构数据治理架构，从数据传输、边缘数据治理、云端数据治理等维度给出解决方案，

*基金项目：MJ专项科研项目（MJXX－XXXX）

提升数据传输效率和数据存储质量，并通过实验验证了架构的可行性和可用性。

1　基于云－边－端的多源异构数据治理架构设计

本文以数据为核心，构建基于工业互联网云平台的远程分布式云－边－端一体化多源异构数据治理架构，如图1所示，数据自底向上逐层传输处理，涵盖数据采集、数据存储、数据处理和数据应用全过程，形成多源异构数据采集、边缘数据预处理传输、云数据中心工业大数据存储以及云端业务应用系统等各环节技术与业务架构解决方案。

（1）工厂端：解决制造现场端多类型设备、多类业务系统以及多种监测方案产生的多源异构数据采集问题，采集数据包括设备实时运行数据、工艺数据以及计划、质量等业务数据；

（2）边缘端：采集工厂端设备和本地网络中的数据，通过数据协议解析完成采集，经清洗处理在本地数据库形成备份，实现设备告警、状态监控、数据处理、历史数据查询等功能，同时将设备采集信息传送给数据中心；

（3）云端：采集边缘端上报数据和工厂端本地数据，根据数据特性选择实时数据库、关系数据库或者分布式存储，经数据处理，形成数据资产，实现制造过程数据云端的统一汇聚，支撑制造业务服务、数字孪生、大数据服务、人工智能算法服务等云端应用。

1.1　数据资源云－边－端采样同步机制设计

云－边－端架构下的数据采样传输机制分为上行数据链路和下行数据链路，具体传输机制见图2。

（1）上行数据链路：制造现场由不同类别制造设备，业务系统，以及传感器、路由器、控制器、执行器等组成，产生设备实时数据、业务过程数据、文本类数据、图片视频类数据等工业大数据，这些海量的数据直接上云会给云端服务器带来网络拥塞、通信效率低等问题。基于云－边－端架构，根据具体业务场景将制造过程数据分割为延时敏感型数据与延时容忍型数据，延时敏感型数据通过 Modbus、Kafka 等高速通信协议，实时通过机载专网将数据传输至附近的边缘系统，经边缘系统传输至云端数据中心；延时容忍型数据以周期的形式，通过 FTP/SFTP、数据库同步等多种方式，直接从制造现场端将数据传输至云端数据中心，云数据中心以数据湖的形式集成了 Doris、MySQL、IoTDB 以及 HDFS 等多种数据存储系统，能够支持结构化和非机构化数据存储，并通过 API 接口等多种灵活形式，供云端的上层应用服务调用。

（2）下行数据链路：云端上层应用服务通过 MQTT 等通信协议，将控制指令和业务数据下发至边缘系统，再经边缘系统下发到具体的生产制造设备和业务系统。

1.2　边缘端数据治理策略设计

云－边－端架构下，边缘端可通过 MQTT、Modbus 等南向通信接口接收网关传来的海量多源异构工业数据，并利用“规则引擎”进行本地数据处理、边缘数据处理、路由处理、边缘数据清洗等动作，对数据去重，过滤掉噪声数据，提炼设备价值数据，主要治理策略如下：

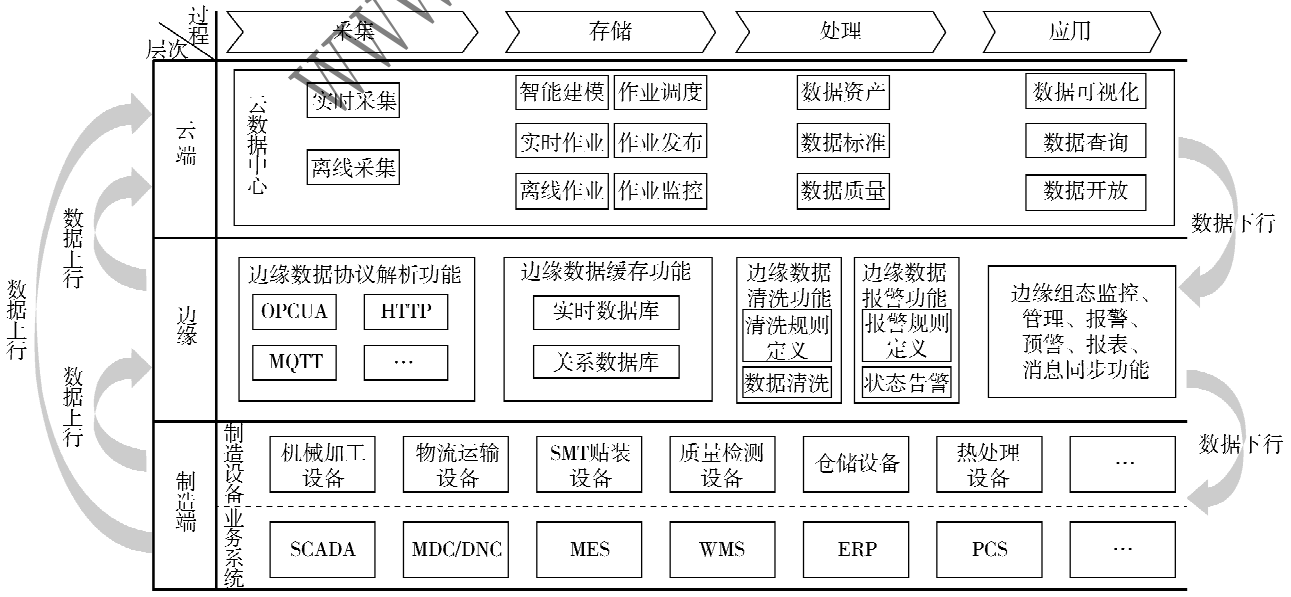

图1　云－边－端一体化多源异构数据治理总体架构

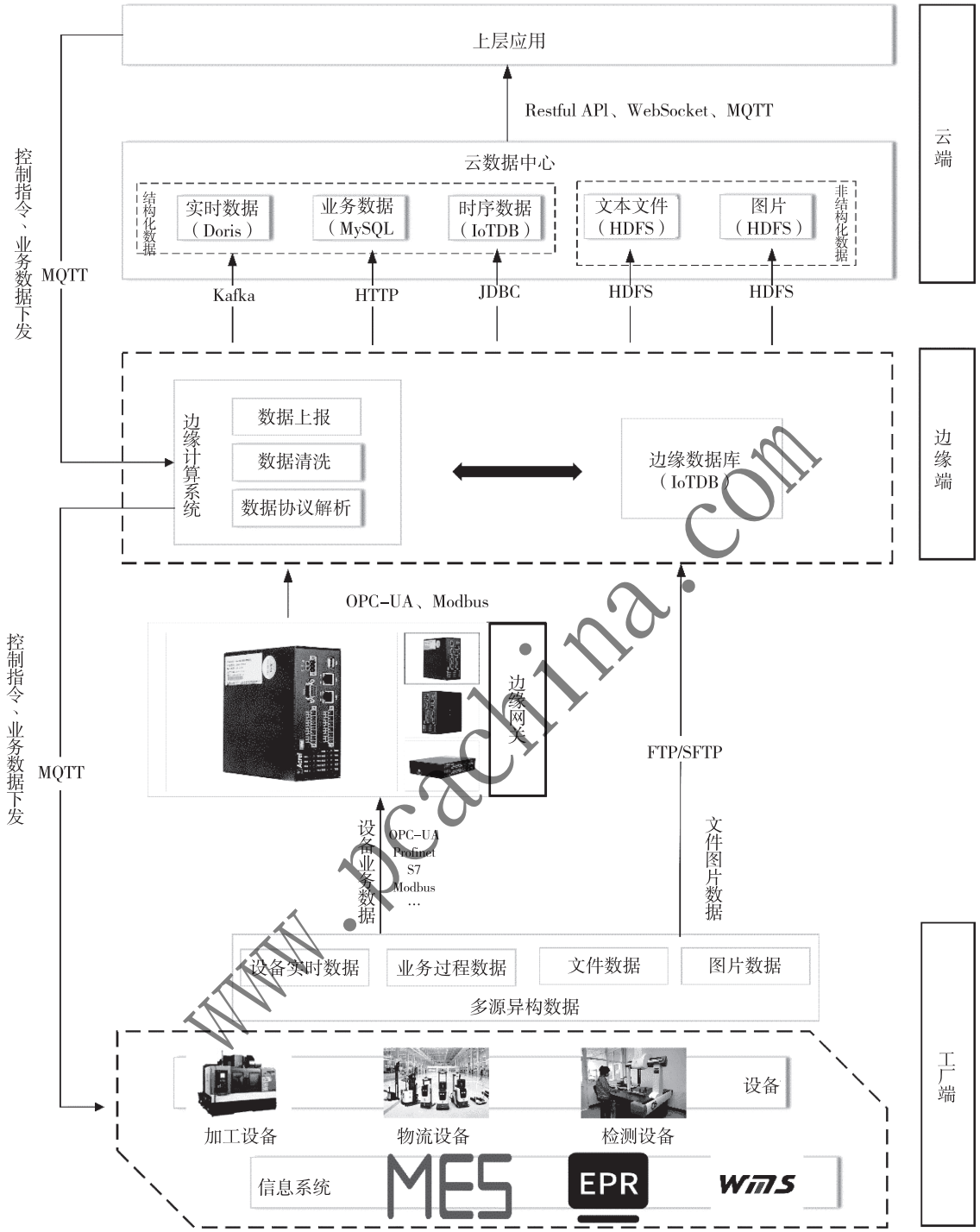

图 2　云－边－端数据采样同步机制

（1）去除/补全有缺失的数据：根据数据需求确认数据缺失的范围和缺失程度，制定缺失数据的去除、补全策略。对于缺失而不需要的内容进行剔除，对于缺失但必要的数据进行填充，如：以业务经验或知识推测缺失值；以同一对象同一指标的总体计算结果（均值、中位数、众数等）、历史数据进行缺失值估计；以不同指标的计算结果填充缺失值，依据同一对象的其他数据对缺失数据进行填补；用统一的关键字填充缺失数据，向用户提供告警信息。

（2）去除/修改格式和内容错误的数据：多源数据由于来源不同，采用的数据计量方式可能有区别，需制定针对数据源的数据格式转换规则，如日期、时间、单位等的转换；以经验数据为基础，结合数据使用方的需求，制定数据模板，对明显错误的数据内容、格式进行剔除。

（3）去除/修改逻辑错误的数据：根据经验设置数据范围，对于超范围的明显异常数据块进行删除；对同一对象相同来源的相同数据项建立交叉对比关系库，对于明显矛盾数据进行标记、删除、修复（删除矛盾数据中的明显偏离数据，保留可能的正确数据）。

（4）数据相互校验验证：对于同一对象不同来源的相同数据项进行交叉比对，通过预设模板判断数据的正确性和内在矛盾，进行错误数据的删除、矛盾数据的修复、难以判断数据的标注。

1.3 云端数据治理策略设计

为了应对数据管理问题，提升数据质量，云端数据中心构建数据资产管理平台提供全生命周期管理服务，实现数据的统一管理，保证业务数据在采集、转换、存储、应用整个过程中的完整性、准确性、一致性和时效性，主要包括数据注册、数据标准和数据质量三个部分，总体架构如图 3 所示。

（1）数据注册

垂直数据是指云数据中心接入存储的工厂端数据，包括设备实时数据和工厂业务数据；按照数据应用业务领域进行主题域划分，例如划分为大数据分析主题域、数字孪生主题域和业务系统主题域，形成各主题域公共数据；在公共数据基础上，根据应用系统具体业务逻辑需求，对公共数据做数据统计、数据汇聚等数据加工处理，落地形成功能数据，做到一次加工多次使用或一对一设计（需求场景—数据）。

元数据是关于数据的数据，主要用来描述数据的上下文信息，从垂直数据、公共数据和功能数据中提取数据表名、表字段、表索引、数据来源等描述性信息形成元数据，构建元数据信息能更方便地检索数据资产，理解数据，发现和描述数据的来龙去脉；主数据是指垂直数据、公共数据和功能数据表单自身数据。将元数据和主数据信息注册为数据资产，为数据质量提供稽核对象。

（2）数据标准

接入的工厂端数据会应用于多个业务域，通过构建通用数据标准，促进业务域内、业务域之间的数据分类，定义和理解一致，保障系统间的数据交换和共享有效。数据标准包括制定统一数据业务含义、数据业务分类、数据类型、数据格式等，为数据提供数据规范，为数据质量稽核提供依据。

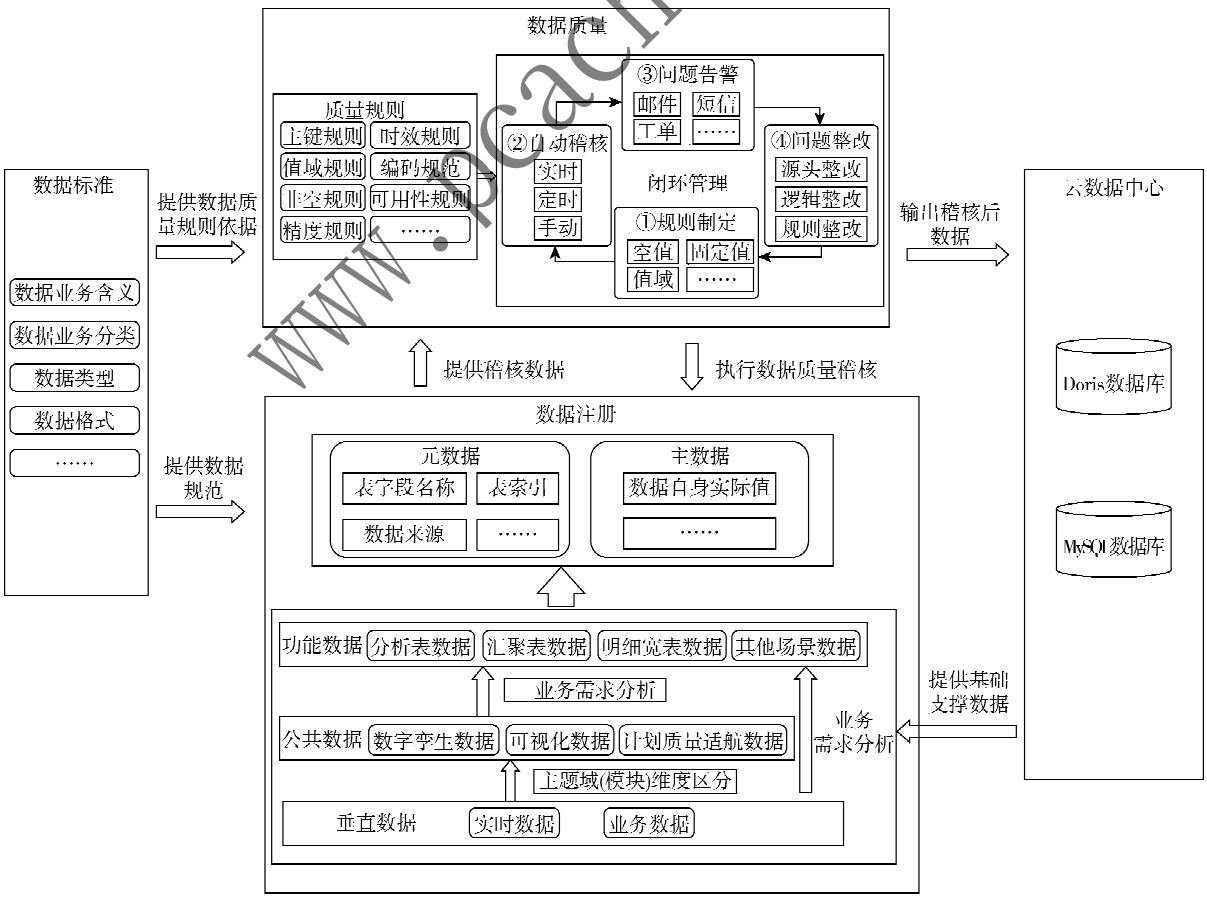

图 3　数据资产管理架构图

（3）数据质量

依据数据标准和数据模型，通过数据质量管理定位分析数据质量问题，提升数据可访问性、可用性、正确性、一致性等。数据质量规则包括主键规则（数据表主键检测）、值域规则（表字段数据范围检测）、非空规则（表字段是否为空检测）、时效规则（数据是否过期检测）等，构建实时或者定时任务对注册数据做稽核，对发现的数据质量问题提供邮件预警和短信预警，根据稽核结论整改输入数据或者优化数据质量规则，最终形成数据质量闭环管理。

2 实验与技术验证

本文面向航空机载航电、机电、飞控系统等多家研制单位生产制造过程，开展了云－边－端一体化数据治理架构性能的实验验证。下面分别从实验室环境下的传输性能验证和生产制造现场应用验证两方面开展技术验证。

2.1 传输性能验证

2.1.1 数据准备及环境配置

本实验数据主要包括图像检测数据、质量检测文本类数据、业务系统数据以及设备实时数据，数据量总计800 MB（包括 26 179 200 个数据点位），边缘设备配置 Kafka、FTP、HTTP、MQTT 传输协议，云端数据中心配置 Doris、MySQL、HDFS、IoTDB 存储系统。

2.1.2 结果分析

首先开展单数据源传输性能测试，云－边－端架构整个链路可以实现最大 115 MB/s 的数据传输，完成 26 179 200个数据点写入云端数据中心，耗时约 7.8 s，见图 4。

进一步测试云－边－端架构的并发传输性能，在 18 路数据源同时并发传输的情况下，服务器 CPU 占用率全程小于 15%，内存占用在 50% 以下且无波动，测试时 CPU 和内存的资源使用情况见图 5。将并发数据源提升至 42 路，在大概 11 min 时，10.51.127.2 服务器的 CPU 占用率约 60%，10.51.127.3 服务器的 CPU 占用率约 40%，内存占用全程无波动，见图 6。

2.2 制造现场应用验证

2.2.1 数据准备

基于航空某制造单位机加生产线实际制造过程数据，开展云－边－端数据治理架构的应用论证。自动化产线包含加工机床、清洗设备、测量设备和传感器等终端设备，实时产生设备告警数据（样例见表 1）、机床实时状态数据（样例见表 2）、环境数据（样例见表 3）等多类数据。

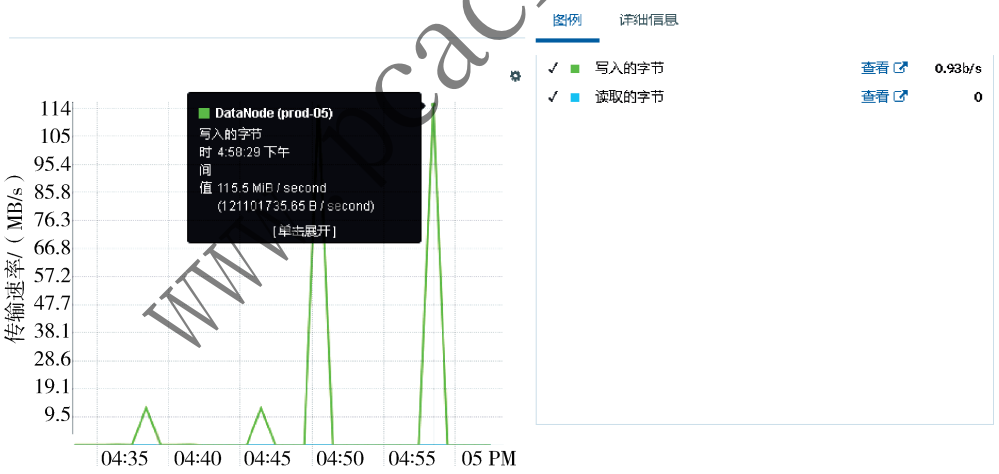

图 4 数据传输速率测试图

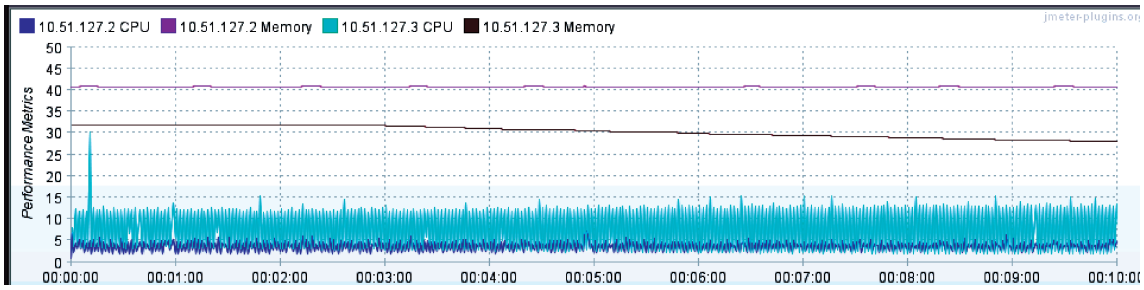

图 5 并发 18 路数据源传输资源使用图

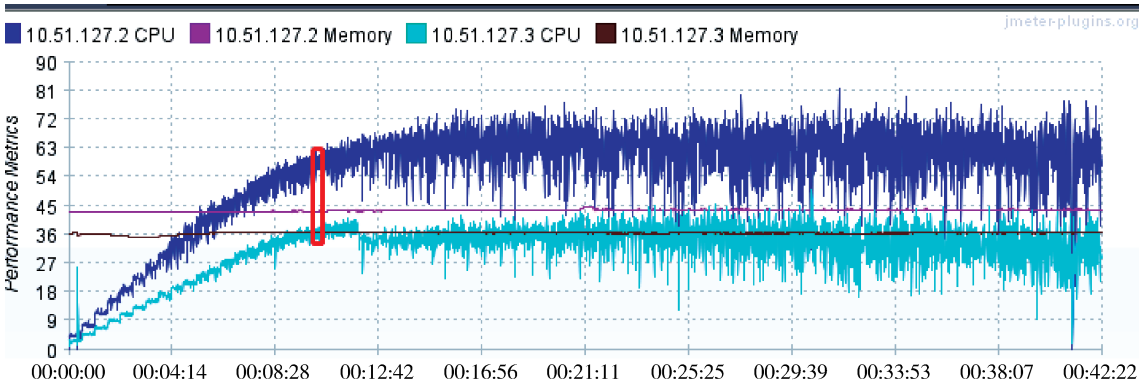

图 6　并发 42 路数据源传输资源使用图

表 1　生产设备告警数据表

	维修工单编号	报警类型/故障码	报警时间	故障设备	…	故障参数	故障原因
样本 1	1	警告	XX－XX－XX 14：30：05	RXX41	…	1 000	机器人程序错误
样本 2	2	警告	XX－XX－XX 14：30：00	RXX42	…	1 000	机器人程序错误

表 2　机床实时状态数据表

	时间	设备编码	设备运行状态	主轴进给倍率/%	主轴进给/（mm/min）	主轴倍率/%	主轴负载	…	主轴转速/rpm	设备利用率/%
样本 1	XX－XX－XX 15：47：19	RXX41	1（运行）	100	200	100	1. 185	…	1 440	1. 146
样本 2	XX－XX－XX 15：47：06	RXX42	1（运行）	100	100	100	0. 945	…	1 000	1. 147

表 3　环境数据表

	时间	温度/℃	湿度/%	…	压力/kPa	洁净度/（mg/m^3）	电磁/mG
样本 1	××－××－×× 15：47：19	33. 5	33. 8	…	95. 7	2. 44	0. 7

2. 2. 2　结果分析

工厂端负责采集、解析各类设备工业协议（如 Modbus、OPCUA），将二进制编码数据转换成明文可读数据，并通过 MQTT 消息缓存中间件技术将数据传输至边缘端，相比于传统 API 接口的 HTTP 协议传输方式，不需要考虑 API 接口的 HTTP 三次握手时延、对向接口的业务处理耗时带来的响应延迟和接口输入数据类型限制，极大地提升了设备端至边缘层的数据传输速度，传输耗时由 10 s 左右降低至 1 s 以内。边缘端负责消费 MQTT 数据并上传至云端 Kafka，在传输过程中建立数据清洗规则，对数据进行预清洗工作。针对告警数据，剔除告警码不正确的记录，针对设备实时状态类数据，剔除状态码不可识别的记录，针对明显异常的环境数据，修正为车间标准值，数据清洗过程中，会去除重复、错误、不完整或冗余的记录，从而降低数据量，提升数据质量。经过数据清洗和处理操作，数据削减率达 20%，数据可信度达 85%。云端负责消费 Kafka 数据，采用轮询的方式读取数据，进行数据质量稽核，对上传的各类数据规定数据标准，例如数据类型、具体的时间格式、数据阈值范围，定时地对比、修正数据库数据，经过稽核处理后的数据可信度达 99. 999 999%。

总体而言，运用本文提出的云－边－端架构，经过 1 年多时间的运行，通过调取云端数据中心工厂生产产线加工设备、测量设备运行数据，对比原先采用的云－端架构，引入边缘处理方案显著减少了数据在网络中的传输量，降低了网络带宽的消耗，进一步降低了数据传输延迟，在数据传输及存储方面整体效率提升 20% 以上。

3　典型应用场景

本文提出的云－边－端架构典型应用场景有：

（1）高实时云架构数字孪生

高实时云架构数字孪生是指利用云技术实现数字孪生的一种架构，需要以极低的延迟处理数据，快速响应物理世界的变化。采用基于工业互联网的云－边－端数据治理架构，应用数据高效传输、数据驱动模型运行等技术，解决数据实时性问题，提升航空机载产品制造数字孪生运行效率以及远程协同能力；云架构数字孪生突破地理位置的限制，让产线管理人员能够随时随地、360°监控生产过程，及时监测故障，大大降低生产运维成本，促使生产数据透明化、价值化。

（2）基于人工智能的图像实时检测

航空典型模块电装过程中小阻容件缺失难以检测，人工检测耗时长、准确率难以保障。面对这一需求，引入人工智能技术，依托云－边－端架构采集、存储的航电典型模块电装过程图像大数据，采用人工智能图像检测技术，云端完成检测模型的训练，以 Docker 的形式发布到边缘侧，能够高效和实时地实现快速图像检测和缺件识别标注，相较传统人工逐个检测标注方式（历史数据统计平均单个缺件检测耗时秒级），智能检测耗时毫秒级，检测准确率约 95%，效率提升 10 倍以上。

4 结论

本文面向航空机载产品制造过程中海量多源异构工业大数据传输和管理需求，提出了一种基于云－边－端的多源异构数据治理架构，设计了数据采样同步机制、边缘端数据治理策略、云端数据治理策略。通过实验室环境下的传输性能实验和生产制造现场应用验证了架构的可行性、可用性，为后续复杂业务的应用奠定了技术基础。

本架构具有如下 3 方面优点：

（1）数据处理在云端和边缘之间进行分布式处理，实现了数据的快速传输和处理，提高了数据处理的效率和速度。

（2）云－边－端架构可以根据实际需求将数据存储和处理分配到云端或边缘设备，实现了资源的灵活配置和调度，提高了系统的灵活性和可扩展性。

（3）通过在边缘设备上进行数据处理和分析，可以减少数据传输的成本，提高了系统的响应速度和性能。

参考文献

[1] YU S, CHEN X, ZHOU Z, et al. When deep reinforcement learning meets federated learning: intelligent multitimescale resource management for multiaccess edge computing in 5G ultra-dense network [J]. IEEE Internet of Things Journal, 2021, 8 (4): 2238－2251.

[2] 李彬，贾滨诚，曹望璋，等. 边缘计算在电力需求响应业务中的应用展望 [J]. 电网技术，2018，42（1）：79－87.

[3] KAI C H, ZHOU H, YI Y B, et al. Collaborative cloud-edge-end task offloading in mobile-edge computing networks with limited communication capability [J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7 (2): 624－634.

[4] MAZAYEV A, CORREIA N. A distributed CoRE-based resource synchronization mechanism [J]. IEEE Internet of Things Journal, 2020, 7 (5): 4625－4640.

[5] 许鹏，何霖. 新型电力系统下 5G＋云边端协同的源网荷储架构及关键技术初探 [J]. 四川电力技术，2021，44（6）：67－73.

[6] 张巍，王丹. 基于云边协同的电动汽车实时需求响应调度策略 [J]. 电网技术，2022，46（4）：1447－1458.

[7] 蒲世亮，袁婷婷. 基于云边融合的物联网智能服务架构探讨 [J]. 智能物联技术，2018，1（1）：1－6.

[8] 周超，林湛，李樊，等. 城市轨道交通视频监控系统云边协同技术应用研究 [J]. 铁道运输与经济，2021，42（12）：106－110.

[9] YAO X X, KONG H F, LIU H, et al. An attribute credential based public key scheme for fog computing in digital manufacturing [J]. IEEE Transactions on Industrial Informatics, 2019, 15 (4): 2297－2307.

[10] NING Z L, KONG X J, XIA F, et al. Green and sustainable cloud of things: enabling collaborative edge computing [J]. IEEE Commun. Mag., 2019, 57 (1): 72－78.

[11] 李大鹏，李立新，杨清波，等. 云边协同的调控云数据质量优化 [J]. 电力系统及其自动化学报，2022，34（3）：11－19.

（收稿日期：2024－08－25）

作者简介：

许政（1986－），男，硕士，高级工程师，主要研究方向：民机机载嵌入式系统研发及智能制造关键技术。

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