基于 Flink 的车联网大数据平台研究与应用

刘力岩 满国晶

哈尔滨信息工程学院

摘 要: 目前汽车产业正加速向智能化、网联化与电动化演进,“软件定义汽车”已成为行业发展的核心趋势。各大车企在日常运营中积累了海量的车辆运行相关数据,但传统的数据存储架构与计算模式难以满足车辆原始数据高效存储、实时运算及智能服务的实际需求。对此,本文设计并搭建了一套基于 Flink 技术的车联网大数据处理平台。该平台以车载 TBOX 终端采集的车辆原始数据作为输入,通过嵌入式程序解析转化为 TSP、DCS、充电、HU 等多种类型的业务数据,经 ETL 数据清洗转换流程后,存储至分布式文件系统与非关系型数据库;依托 Flink 框架对 Kafka 消息队列中的数据流开展实时计算处理,最终通过 API 接口向外界提供数据服务,并以可视化报表和数据大屏的形式呈现数据分析结果。

关键词:实时计算;车联网大数据;分布式处理平台;车辆远程诊断;实时故障监测

DOI:10.65976/3078-8145.2026.03.013

引言

随着物联网与人工智能技术的深度融合发展,车联网已成为智能交通体系的核心支撑环节。作为连接车辆、道路基础设施与云端服务的关键枢纽,车联网通过实时的数据交互流转,实现了交通状态的全域感知与动态调度。随着车载传感器数量的不断增加、自动驾驶技术的逐步落地以及车路协同应用场景的持续拓展,车联网系统单日产生的数据量已达到 TB 级别,部分场景甚至突破PB级别,呈现出典型的大数据特征。如何高效完成海量异构车辆数据的采集、稳定存储、深度分析及价值挖掘,成为车联网技术从研发阶段走向规模化实际应用的核心瓶颈。

针对上述存在的问题,本文构建了基于 Flink 的车联网大数据平台。该平台能够实现车辆运行状态的全天候实时监测,当车辆发生故障时,可快速推送故障编码及详细信息,同时对车辆电源系统进行集中统一管控。借助实时驾驶行程分析功能,以数据驱动的模式优化驾驶行为、提升车队运营效率、保障行车安全。该平台可提前识别车辆潜在的故障风险,增强车辆防盗监测能力,为后续车辆工况分析、故障预警等功能的开发提供坚实的技术支撑。

1 核心技术阐述

1.1 Apache Flink

Apache Flink 是一款可实现无界与有界数据流统一处理的分布式计算框架,既能够部署在 YARN、Mesos 等资源调度平台上,也可独立搭建集群运行,能够适配海量数据处理、数据源不稳定等复杂应用场景。本平台选择 Flink 作为核心计算框架,核心原因在于车联网数据同时具备流式实时数据与批量离线数据的双重特征,而 Flink 可实现流批一体的高效处理,将批量数据视为特殊的流式数据,全流程采用流处理模式执行,具备极低的数据处理延迟,能够满足车联网实时计算需求。

1.2 Kafka

Kafka 是一款基于 Zookeeper 协调管理的分布式消息中间件,拥有高吞吐量、低时延、高可靠性、强容错性等显著优势。其低时延特性可实现毫秒级的数据传输,单节点每秒能够处理海量消息;高吞吐能力即便在普通商用服务器上,也可实现每秒 10 万级别的消息转发。与此同时,Kafka 支持数据本地持久化存储与多副本备份机制,可有效避免数据丢失,能够完全适配车联网海量数据的传输场景。

1.3 Phoenix

Phoenix 是构建在 HBase 之上的 SQL 查询中间层,支持通过标准的 JDBC API 完成 HBase 数据表的创建、数据写入及查询操作,无须依赖原生的 HBase 客户端接口。该组件采用 Java 语言开发,以嵌入式 JDBC 驱动的形式集成于 HBase 中,可将 SQL 查询语句转化为 HBase 扫描操作,并编排执行生成标准的 JDBC 结果集。本平台通过 Phoenix 组件,实现原始数据的实时 ETL 处理与高效写入 HBase,确保车辆状态实时监测的时效性。

1.4 MongoDB

MongoDB 是一款基于分布式文件存储的开源 NoSQL 数据库,采用 C++ 语言开发,专门为 Web 应用提供高可用性、易扩展性的数据存储解决方案。凭借其文档型数据模型、实时数据处理能力以及与大数据生态的良好兼容性,MongoDB 成为车联网“端 – 边 –

云”数据流转的核心载体，能够适配数据模型快速迭代、地理围栏预警、流批一体分析等各类业务场景。

2 车联网大数据平台架构设计

本平台重点围绕工业大数据的存储、整合与分析需求展开设计，为企业多类业务场景提供决策支撑与数据服务，整体架构分为数据源层、消息队列层、数据分析层、可视化分析层、系统集成层五大核心模块。

2.1 数据源模块

平台的数据源主要分为静态基础数据与实时流式数据两大类，具体数据采集方式如下。

1. 静态系统数据：由专业人员负责采集车辆生产日期、车型、设备编号等基础信息，整理为标准化的 Excel 表格，可直接接入平台进行使用；

2. 车载终端数据：通过车企自主研发的新能源汽车智能终端，采集车辆实时运行数据，并及时上传至平台的数据处理单元；

3. 系统日志数据：主要记录平台软硬件运行状态及各类系统事件，为平台的运维管理与故障排查提供支撑；

4. 业务静态数据：车辆使用过程中会产生多种格式的业务静态数据，通过 Sqoop 数据转换工具完成格式标准化处理后，接入车联网大数据平台。

2.2 消息队列模块

消息队列作为数据传输过程中的临时缓存载体，能够有效应对车联网数据类型繁杂、数据体量庞大的问题。其核心作用在于缩短系统响应时间、提升系统运行稳定性，保障数据传输的有序性与安全性，同时实现系统各组件的解耦与数据的异步传输。

本模块选用 Kafka 作为消息队列中间件，通过 Flink 框架将实时流数据与批量数据写入消息队列。其中，Flink 作为数据生产者，持续生成各类数据消息并推送至 Kafka；Kafka 作为数据消费者，接收 Flink 推送的消息，并为后续的数据分析计算环节提供稳定的数据输入。

2.3 数据分析模块

该模块以 Flink 为核心计算引擎，同时支持实时大数据处理与离线批量数据计算，根据数据类型进一步分为实时处理子模块与批量处理子模块。在实际业务场景中，人工操作失误、数据采集设备受现场环境干扰等因素会导致采集的数据出现失真问题，此类数据若直接入库，不仅会降低平台数据查询的准确性，还会大幅影响平台的运行效率。因此，该模块通过 Flink 框架完成无效数据、重复数据及高缺失率数据的清洗与过滤，保障数据质量。

驾驶行程指用户在一定时间段内的连续驾驶行为，由载重行程与空驶行程两部分组成，以停车时长超过 15 分钟作为行程划分的标准。驾驶行程分析流程为：原始数据→ Kafka 消息队列→ Flink 实时处理（ETL 清洗 + 业务逻辑解析）→行程数据入库。针对车辆数据上报过程中出现的乱序问题，采用 Flink 的水位线（Watermark）机制，结合时间戳分配与水位线生成，确保数据按照事件时间有序处理。

2.4 可视化分析模块

平台采用开源工具 Grafana 搭建可视化分析层，支持 IoTDB、MySQL、InfluxDB 等多种类型数据库的接入，提供折线图、数据表、统计图等多种数据展示形式，让用户无须了解后台运行逻辑，即可直观获取所需数据。同时，该模块具备设备阈值预警功能，可通过 Slack、钉钉、邮件等多种渠道推送告警信息，实现设备异常状态的精准提醒。

2.5 集成过程模块

数据源层通过 Flume 组件采集系统运行过程中产生的日志数据，或直接从传感器获取原始数据，由专业人员整理为 Excel 格式。Flink 框架贯穿平台全流程数据处理，针对 MySQL 数据库中的数据采用 DataSet API 进行处理，针对 InfluxDB 数据库中的数据采用 DataStream API 进行处理。可视化展示层依托 Grafana 组件，实现各类数据的增删改查与实时监测，当数据超出预设阈值时，自动触发邮件预警机制。

3 实验验证与结果分析

3.1 实验环境与数据来源

平台集群由 7 台物理机构建而成，其中 1 台作为主节点，其余 6 台作为从节点，主机名分别为 Master、Slave1~Slave6。单台物理设备配置 32 GB 内存与 1 TB 硬盘，操作系统采用 CentOS 7.9 64 位；Flink 集群选用 1.9.3 版本，JobManager 主进程部署在 Slave1 节点，配置文件分发至各个从节点，实现主节点对从节点的免密登录与启动。为保障系统兼容性，选用 Kafka 2.2.0 与 Zookeeper 3.4.10 组合；前端可视化工具采用 Grafana 6.7.2 版本，数据库选用 MySQL 5.5 与 InfluxDB 1.7.3。

实验数据来源于新能源汽车 TBOX 终端上报的运行信息，数据采集频率为每 5 秒 1 次，单辆车每 30 秒上报 6 条数据，小时数据量达到 720 条 / 辆；在万辆车同时运行的场景下，每分钟产生 12 万条数据，15 分钟内的数据规模突破千万条，能够充分模拟实际车联网数据场景。

3.2 平台功能实现

为验证基于 Flink 的车联网大数据平台的可行性

与有效性，对平台各模块进行了全面的功能测试，具体实现流程如下：将批量数据与实时数据导入 Kafka 消息队列，其中批量数据量较小，耗时 3 分 10 秒完成“企业名单”数据的导入；实时数据量较大，耗时 17 分钟完成“设备实时事件统计”数据的导入。随后通过 Flink 框架读取 Kafka 中的数据，执行实时 ETL 清洗与数据分类操作，将处理后的各类数据分别存入对应数据库。

用户通过浏览器访问 localhost:3000 即可登录平台，完成数据库账号密码验证后，可创建所需的业务数据库，并选择折线图、表格、文本等数据展示样式。例如，用户需查询 MySQL 数据库中天宁区的企业名单，可通过 Table 组件输入 SQL 查询语句，即可以表格形式输出查询结果；针对实时数据，可快速展示各类设备的运行状态及对应发生时间，同时支持通过编辑界面选择目标数据库与筛选条件，实现相同设备 ID 查询、同一时间段设备上线量统计、故障设备数量统计、预警阈值与结束时间展示等多种功能。

4　结语

本平台以车联网核心技术为依托，完成了智能终端、服务端与移动端的数据互通及远程控制，具备实时性强、适配范围广、扩展能力突出等特点。在车辆故障监测方面，平台可及时发现车辆潜在故障并推送相关信息，降低故障造成的损失，让车主能够直观掌握车辆运行状态；在车辆防盗方面，平台具备高效的实时性与可靠性，可全天候监测车辆防盗状态，提升防盗预警的准确率，有效降低误报率。车载终端兼容多种 OBD 标准协议，数据读取精准，进一步强化了人车信息交互能力。该平台在汽车安全保障领域具有较高的实用价值，拥有广阔的落地应用与优化升级空间。

参考文献：

[1] 周磊，顾云．集度汽车 Flink on native k8s 实时计算平台实践 [J]. 大数据技术与应用 ,2023,9(2):45-52.

[2] 石静猛．长安汽车基于云器 Lakehouse 的车联网大数据平台建设 [J]. 计算机工程与应用 ,2024,60(6):112-119.

[3] 王悦，谢满刚．信息年龄优先的车载网络异构数据传输策略 [J]. 计算机工程 ,2025,51(8):156-163.

[4] 魏然，李平凡，万钧冉．新能源汽车电子数据第三方取证平台构建探索 [J]. 中国司法鉴定 ,2024(5):27-33.

[5] 朱伟．智能网联汽车实时数据分析破局之道 [J]. 汽车工程学报 ,2026,6(1):78-85.