Contribution ID: **19** Type: **not specified**

面向大型粒子加速器的 AI-Ready 数据集生成平台的设计与实现

Monday, 25 August 2025 17:20 (20 minutes)

随着人工智能技术在科学装置领域的深入应用，如何高效实现大型粒子加速器的智能调节、智能故障诊断及预测性维护，正在成为研究热点。但是在实际应用中，算法与数据之间存在不容忽视的挑战：加速器系统数据普遍存在异构性强、多模态耦合、时序特征复杂、因存储分散导致的数据孤岛效应、以及数据协议与标准的缺失，成为智能算法模型有效部署与迭代的核心障碍。
针对上述挑战，本文面向高能同步辐射光源（HEPS）和中国散裂中子源（CSNS），设计并实现了一个遵循 FAIR 原则的 AI-Ready 数据集生成平台——FARAD。该平台采用任务驱动的微服务架构，集成了数据清洗、多源时序对齐、特征工程及融合等核心功能模块，打通加速器数据从原始采集到算法训练的全链路流程。本平台基于 MongoDB 与 Kafka 构建高效的数据采集与存储体系，结合 Pandas 与 NumPy 实现数据清洗与特征工程。在数据融合方面，系统通过 RESTful API 打通异构系统，结合统一数据模型、元数据驱动和语义对齐，提升多模态数据集成效率，支持标准化与血缘追踪。平台采用 Docker 容器化部署，具备模块解耦与协议兼容能力，支持弹性扩缩容和快速迭代，能够适配 HEPS、CSNS 等加速器装置及主流 AI 框架，具备良好的扩展性与持续演进能力。FARAD 从根源上提升了数据质量与可用性，降低人工智能模型在加速器场景中的应用门槛。

Summary

Primary authors: 鲍, 伟; 卢 (LU), 晓含（Xiaohan）(高能所); CHENG, Sinong (Institute of High Energy Physics); JIAO, Yi (高能所); HE, Yongcheng (高能所); 黄蔚玲, Weiling (高能所); ZHANG, Yuliang (Institute of High Energy Physics)

Presenter: 鲍, 伟

Session Classification: 人工智能与应用

Track Classification: 人工智能与应用