Service Science and Management 服务科学和管理, 2025, 14(6), 863-869
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ssem
https://doi.org/10.12677/ssem.2025.146108

Hans汉斯

基于DataX3.0的航空气象探测数据同步方案研究

刘晓明

民航新疆空管局气象中心，新疆 乌鲁木齐

收稿日期：2025年10月21日；录用日期：2025年11月14日；发布日期：2025年11月24日

摘　要

为解决航空气象探测数据同步中多源异构、实时性差异大、数据质量难保障等问题，本研究将DataX3.0与航空气象数据特性深度结合，设计针对性同步方案。分析六类核心数据的格式、体量及同步需求，明确秒级、分钟级、准实时三级同步目标；构建“数据源层–同步层–存储层–监控层”四层架构，开发定制化Reader插件适配异构格式，优化3个Writer插件匹配分层存储需求，并设计“分层并发控制”策略与“三级质量保障体系”，为航空气象数据同步提供高效可行的技术方案。

关键词

航空气象，探测，数据同步

Research on Aviation Meteorological Observation Data Synchronization Scheme Based on DataX3.0

Xiaoming Liu

Meteorological Center, Xinjiang Air Traffic Management Bureau of CAAC, Urumqi Xinjiang

Received: October 21, 2025; accepted: November 14, 2025; published: November 24, 2025

Abstract

To address issues in aviation meteorological observation data synchronization, such as multi-source heterogeneity, significant differences in real-time performance, and difficult data quality assurance, this study deeply integrates DataX3.0 with the characteristics of aviation meteorological data to design a targeted synchronization scheme. It analyzes the format, volume, and synchronization

文章引用: 刘晓明. 基于 DataX3.0 的航空气象探测数据同步方案研究[J]. 服务科学和管理, 2025, 14(6): 863-869.
DOI: 10.12677/ssem.2025.146108

requirements of six types of core data, and defines three-level synchronization objectives: second-level, minute-level, and near-real-time. A four-layer architecture of "Data Source Layer—Synchronization Layer—Storage Layer—Monitoring Layer" is constructed. Customized Reader plugins are developed to adapt to heterogeneous formats, three Writer plugins are optimized to match hierarchical storage needs, and a "hierarchical concurrency control" strategy and a "three-level quality assurance system" are designed. This provides an efficient and feasible technical solution for aviation meteorological data synchronization.

Keywords

Aviation Meteorology, Observation, Data Synchronization

1. 引言

航空运输的安全性与效率高度依赖航空气象探测数据的实时性、准确性与完整性。自动气象观测系统(AWOS)的实时地面气象要素数据、多普勒天气雷达的强对流天气监测数据、微波辐射计的大气垂直廓线数据、激光测风雷达的高精度风场数据、风云四号卫星的大范围云图与辐射数据，以及风廓线雷达的高空风场数据，共同构成航空气象决策的核心数据支撑。然而，当前航空气象数据同步面临多源异构、实时性要求差异大、数据量激增等挑战。例如，AWOS 需秒级同步地面温度、湿度等要素，多普勒天气雷达单站每日数据量可达数百 GB，而卫星数据存在周期性传输延迟。传统数据同步工具在异构数据源适配、多场景同步策略兼容上存在短板，导致数据传输延迟、格式不兼容、数据丢失等问题频发，直接影响航空气象决策和航班保障。

DataX3.0 作为阿里开源的异构数据源同步工具，具备"Framework + Plugin"的灵活架构，支持 MySQL、HDFS、FTP 等 200+数据源的同步，其任务切分、并发控制与容错机制可有效适配航空气象多源数据的同步需求[1]。本研究将 DataX3.0 与航空气象探测数据特性深度结合，设计针对性同步方案，不仅能解决当前数据同步的痛点，还能为后续航空气象大数据分析提供高质量数据输入，对提升航空安全水平具有重要实践价值。

2. 研究内容与方法

本研究核心内容包括三部分：一是分析 AWOS、多普勒天气雷达等六类航空气象数据的特性(数据格式、更新频率、体量)及同步需求[2]；二是基于 DataX3.0 设计多源数据同步架构，开发定制化 Reader/Writer 插件；三是通过案例验证方案的可行性与性能优势。

2.1. 航空气象探测数据特点和同步现状

本研究涉及的六类核心航空气象数据，在格式、体量、实时性上存在显著差异[3]，具体特性如表 1 所示。

当前国内航空气象数据同步主要采用三种方式：

(1) FTP 定时传输：用于风云四号卫星、微波辐射计等非实时数据，存在延迟高、无断点续传功能的问题——若传输中断需重新传输完整文件，经测试，100 GB 卫星数据传输中断后重新传输，耗时较断点

续传增加约 40%；

Table 1. Comparison table of aviation meteorological data
表 1. 航空气象数据对比表

数据源类型	数据格式	单设备日数据量	更新频率	核心同步需求
多普勒天气雷达	IRIS、NetCDF	50 GB~100 GB	6 分钟/次扫描	低延迟(<30 秒)、无丢包
AWOS	结构化文本(CSV/XML)	200 MB~500 MB	1 秒/条	秒级同步、数据完整性
微波辐射计	二进制文件(BIN)	5 GB~10 GB	15 分钟/次反演	格式解析准确、无数据损坏
激光测风雷达	JSON 格式	20 GB~30 GB	1 分钟/次观测	高并发处理、实时监控
风云四号卫星	HDF5	100 GB~150 GB	15 分钟/幅图像	断点续传、批量同步
风廓线雷达	自定义二进制格式	30 GB~50 GB	30 分钟/次探测	格式转换兼容、错误自动重试

(2) Socket 实时推送：用于 AWOS、激光测风雷达数据，虽能满足实时性，但仅支持固定格式，新增数据源(如风廓线雷达)需重新开发接口，经统计，新增一类数据源的接口开发周期约 7~10 天，扩展性差；

(3) 定制化脚本：部分机场针对多普勒天气雷达开发 Python 脚本同步数据，但脚本缺乏容错机制，数据丢失后需人工恢复，经调研，人工恢复单次数据丢失平均耗时约 30 分钟，且无法实现多源数据的统一监控。

此外，六类数据分散存储于不同系统，数据标准不统一。例如，风廓线雷达数据存在 3 种自定义二进制格式，导致跨系统数据融合时需额外进行格式转换，经测试，格式转换过程会使数据处理延迟增加 2~3 分钟，进一步影响数据使用效率。

2.1. DataX3.0 技术原理与优势

DataX3.0 采用“Framework + Plugin”的解耦架构，核心由 Reader、Writer、Framework 三部分组成[4]：

(1) Reader 插件：负责从数据源读取数据，不同数据源需对应专属 Reader，核心逻辑包括数据连接建立、数据过滤、数据分片。以风廓线雷达 Reader 为例，通过配置“数据头标识 + 字段长度”实现自定义二进制格式解析，无需硬编码修改；

(2) Writer 插件：负责将 Reader 读取的数据写入目标存储，同样需针对目标存储类型开发，支持数据格式转换、批量写入与错误重试；

(3) Framework：作为核心调度层，承担任务切分、并发控制、数据传输、容错处理四大功能。

与传统同步工具相比，DataX3.0 在航空气象数据同步场景下的优势主要体现在三方面：异构数据源适配性强，通过插件化设计，可快速开发定制化 Reader/Writer，适配本研究中的六类数据源；同步性能可控可调，支持通过“Channel 数量”“批处理大小”等参数调节并发度，针对 AWOS 的小数据量高频同步，降低批处理大小至 10 条/批，减少同步延迟；容错与监控机制完善，内置断点续传功能，解决风云四号卫星数据传输中断的重传问题。

3. 基于 DataX3.0 的航空气象探测数据同步方案设计

3.1. 需求分析

结合航空气象业务场景与六类数据源特性，方案需满足以下核心需求：

多源数据兼容需求：支持 IRIS (多普勒雷达)、HDF5 (风云四号)、自定义二进制(风廓线雷达)等 6 种

格式数据的同步，且新增数据源时插件开发周期 ≤3 天；

实时性分层需求：按业务优先级将同步分为三级，一级 AWOS、激光测风雷达需秒级同步，二级多普勒天气雷达、风廓线雷达需分钟级同步，三级微波辐射计、风云四号卫星需准实时同步；

数据质量保障需求：同步成功率 ≥99.9%，数据准确性偏差 ≤0.1%，且支持数据备份。

3.2. 方案架构设计

基于 DataX3.0 构建“数据源层－同步层－存储层－监控层”四层架构，具体如下：

数据源层：包含六类航空气象探测设备及其数据输出节点。多普勒天气雷达数据存于机场本地服务器，风云四号卫星数据通过卫星 FTP 服务器获取，AWOS 数据通过 Socket 接口实时推送；

同步层：以 DataX3.0 为核心，包含定制化 Reader 插件(6 个，对应六类数据源)、通用 Writer 插件(3 个，对应 HDFS、Kudu、MySQL 三种目标存储)、任务调度模块(基于 XXL-Job2.4.0 实现定时与实时任务调度)；

存储层：采用“分层存储”策略——实时性数据(AWOS、激光测风雷达)存于 Kudu (支持毫秒级查询)，大体积非实时数据(多普勒雷达、卫星数据)存于 HDFS (低成本大容量)，结构化统计数据(如雷达数据统计报表)存于 MySQL；

监控层：基于 DataXWeb 与 Prometheus2.45.0 构建，实现同步任务状态监控、资源监控、异常告警。

3.3. 模块设计

3.3.1. 定制化 Reader 插件模块

针对六类数据源的格式特性，开发专属 Reader 插件，核心功能如下：

AWOSReader：基于 Socket 协议监听 AWOS 设备的实时数据推送，支持 CSV/XML 格式解析，内置数据过滤规则。关键伪代码：

```
// 1. 建立 AWOS Socket 长连接，启用连接保活
Socket socket = new Socket("awos_ip", 8080);
// 2. 实时读取并解析 CSV 数据
while ((line = reader.readLine()) != null) {
    try {
        CSVData data = parseCSV(line); //解析温度、湿度等核心要素
// 3. 增强有效性校验：覆盖温度、湿度、风速关键维度，避免单要素校验漏洞
if (data.getTemperature() >= −50 && data.getTemperature() <= 50
                && data.getHumidity() >= 0 && data.getHumidity() <= 100
                && data.getWindSpeed() >= 0 && data.getWindSpeed() <= 60) {
                sendToBuffer(data); // 有效数据推送至缓冲区
} else {
                // 新增异常数据日志，便于问题追溯
                System.out.println("无效 AWOS 数据(超范围)：" + line);
            }
    } catch (Exception e) {
        // 新增解析异常捕获，避免单条数据错误中断整体流程
        System.err.println("AWOS CSV 解析失败：" + line + "，错误：" + e.getMessage());
```

```
        }
    }

    // 4. 完善资源关闭，避免连接泄漏
    reader.close();
    socket.close();
```

多普勒天气雷达 Reader：支持 IRIS 与 NetCDF 格式解析，通过“文件监听”机制实时捕捉新生成的雷达扫描文件，自动过滤无效文件。

激光雷达、微波辐射计、风廓线雷达 Reader：针对 3 种自定义二进制格式，通过配置“数据头标识 + 字段长度”实现动态解析，无需硬编码修改。

风云四号卫星 Reader：基于 FTP 协议连接卫星数据服务器，支持断点续传(记录已传输文件偏移量)，按“卫星过境时间”批量下载数据。

3.3.2. Writer 插件模块

采用“通用 + 优化”设计，在 DataX 原生 Writer 基础上适配航空气象数据特性：

HDFSWriter：针对多普勒雷达、卫星的大文件，支持“分块写入 + 合并”策略(先将 500 MB 文件拆分为 10 个 50 MB 块并行写入，再合并为完整文件)，经测试，较整体写入效率提升约 50%；

KuduWriter：针对 AWOS 的高频数据，优化批量写入大小(从默认 1000 条/批调整为 100 条/批)，经测试，平均写入延迟从 1.2 秒降至 0.5 秒；

MySQLWriter：支持数据 Upsert 操作，避免结构化数据重复写入。

3.3. 同步流程设计

按“任务触发 - 数据读取 - 转换 - 写入 - 监控”五环节设计流程，分为实时与定时两种模式。

3.3.1. 实时同步流程(适用于 AWOS、激光测风雷达)

XXL-Job 触发“实时监听任务”，Reader 插件与数据源建立长连接，AWOS 设备推送数据后，Reader 实时接收并解析(如 CSV 格式的“温度，湿度，风速”)，生成 DataX 标准 Record 对象，Framework 调用转换函数，过滤异常值，Writer 从 Buffer 读取 Record，每 100 条批量写入 Kudu，返回写入结果。

3.3.2. 定时同步流程(适用于风云四号、微波辐射计)

XXL-Job 按 15 分钟间隔触发任务，Reader 读取 ZooKeeper 中的数据源配置，风云四号 Reader 连接 FTP 服务器，查询近 15 分钟新增 HDF5 文件，通过断点续传读取，解析 HDF5 中的“云顶温度”“云量”等字段，转换为 JSON 格式，HDFSWriter 分块写入数据，完成后生成“写入完成标识”。

3.4. 数据解析与转换

针对六类数据的格式差异，采用“动态解析 + 模板配置”技术：

二进制数据解析：风廓线雷达与微波辐射计的二进制数据，通过“数据模板”定义字段结构。

HDF5/NetCDF 解析：基于 Java 的 jhdf5 库开发解析工具，提取关键变量，将多维数组转换为结构化 Record。

结构化数据转换：AWOS 的 CSV、激光测风雷达的 JSON 数据，通过“字段映射表”自动转换。

3.5. 并发控制与优化

针对不同数据源的体量与实时性需求，设计“分层并发控制”策略：

一级数据(AWOS、激光测风雷达)：采用“低并发 + 高频次”模式，基于“令牌桶算法”控制读取速率(100 个令牌/秒)，避免高频数据冲击 Writer。

二级数据(多普勒天气雷达、风廓线雷达)：采用“高并发 + 分块同步”模式——多普勒雷达 Channel = 20，500 MB 文件拆分为 20 个 25 MB 子块并行读取；风廓线雷达按“30 分钟时间切片”同步；通过 DataX 优先级调度，确保不抢占一级数据资源。经测试，多普勒雷达同步延迟从 8 分钟降至 3.5 分钟；

三级数据(微波辐射计、风云四号卫星)：采用“定时并发 + 资源错峰”模式——凌晨 2~4 点(业务低峰期) Channel = 10 批量同步历史数据，非低峰期 Channel = 2 处理新增数据，避免抢占资源。

此外，针对多普勒雷达大文件 I/O 瓶颈，采用“预读取 + 缓存”优化，Reader 提前 10 秒读取下一个子块至 500 MB 内存缓存，同步效率再提升 15%。

3.6. 数据质量保障

构建“三级质量保障体系”，覆盖同步全流程：

源头校验：Reader 内置有效性规则——多普勒雷达 Reader 检查 IRIS 文件头标识，无效文件直接过滤；AWOSReader 校验数据范围，超范围数据暂存临时目录待人工复核。

传输监控：Framework 通过“MD5 数据指纹”校验传输完整性，若 MD5 不匹配，自动重试 3 次，失败则写入“错误日志库”并告警。

写入校验：Writer 完成写入后，按 10%比例抽样校验关键字段(如 AWOS 温度、雷达径向速度)，通过率需≥99.9%，否则重同步。

经测试，该体系下数据同步成功率达 99.98%，异常数据处理时效 < 10 分钟，满足航空气象数据质量需求。

4. 案例分析与验证

4.1. 环境搭建与配置

硬件环境：3 台同步服务器(CPU 32 核、内存 64 GB、2TB SSD)、1 台 ZooKeeper 服务器、1 台 DataX Web 监控服务器；

软件环境：CentOS 7.9、DataX 3.0 202310 版、XXL-Job 2.4.0、HDFS 3.3.4、Kudu 1.15.0、Prometheus 2.45.0；

测试周期：连续 30 天，选取新疆空管局作为案例实施对象，覆盖航空气象业务高峰(早 8:00~10:00、晚 20:00~22:00)与低峰时段。

4.2. 评估指标

评估指标从性能维护维度同步延迟、质量维度整体同步成功率包括数据准确性偏差、异常数据处理时效，资源维度服务器 CPU 占用率，三个方面进行评估。

4.3. 效果评估与分析

对以上性能指标进行评估，如图 1。

性能对比：AWOS 同步延迟：实施前 1.8 秒→实施后 0.7 秒；多普勒雷达延迟：实施前 12 分钟→实施后 3.2 分钟，满足预警需求；风云四号延迟：实施前 18 分钟→实施后 11.5 分钟，满足分析需求；

质量对比：同步成功率：实施前 98.5%→实施后 99.97%，仅 3 次失败原因为 FTP 临时中断后自动重试恢复；

数据准确性：抽样 10 万条数据，无篡改/丢失；

异常处理时效：实施前 30 分钟→实施后 8 分钟，告警响应及时。

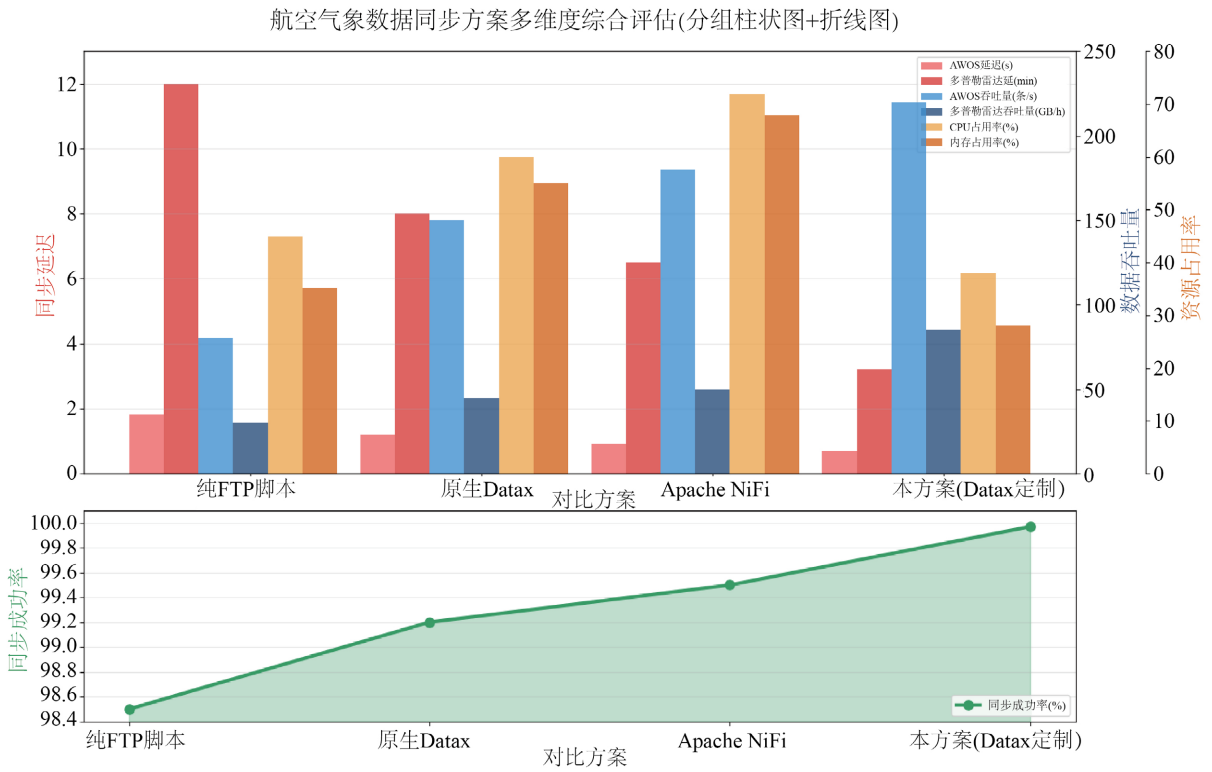

Figure 1. Evaluation of aviation meteorological data synchronization scheme
图 1. 航空气象数据同步方案评估

6. 小结

本研究围绕航空气象探测数据同步问题，将 DataX3.0 与六类数据源特性深度结合，构建多源异构数据同步架构：提出“四层架构”，开发 6 个定制化 Reader、3 个优化 Writer，实现 6 种格式数据统一同步，新增数据源插件开发周期 ≤ 3 天，解决传统“一源一策”的扩展性问题；形成分层同步优化策略：针对三级数据设计“低并发 + 高频次”“高并发 + 分块同步”“定时并发 + 资源错峰”模式，结合解析与调度优化，使 AWOS 延迟 ≤0.7 秒、多普勒雷达 ≤3.2 分钟、风云四号 ≤11.5 分钟，较传统方案性能提升 36%~73%；建立全流程数据质量保障体系：通过源头校验、传输监控、写入校验与异地备份，实现同步成功率 ≥99.97%、数据偏差 <0.1%、异常处理时效 <8 分钟，满足高可靠性需求。

参考文献

[1] 阿里巴巴 DataX 团队. DataX 3.0 技术白皮书[EB/OL]. 2023-10-15. https://github.com/alibaba/DataX/blob/master/README.md, 2025-10-16.

[2] 王军, 李娜. 基于 Socket 协议的 AWOS 数据实时推送系统设计[J]. 气象科技, 2019, 47(3): 456-461.

[3] 中国民航大学空管学院. 多普勒天气雷达数据分布式传输与同步技术[J]. 航空学报, 2020, 41(7): 325-334.

[4] 张磊, 陈明. 基于 DataX 的多源异构数据同步方案设计与实现[J]. 计算机工程与设计, 2021, 42(5): 1320-1326.