文章编号 :1673-8411(2019)02-0094-04

基于流式计算的自动站资料实时处理系统研究

郭捷，乔文文，张金标，寇媛媛

（广东省气象探测数据中心，广州 510640）

摘要：为解决自动气象站资料处理业务中存在的流程分散、处理统计时效性不足等问题，提出一种自动气象站资料实时处理系统的建设思路和方法。将自动气象站资料封装为消息，使用消息中间件进行数据传输，引入分布式流式计算引擎，利用基于现有的数据解码和统计算法，在流式计算框架内进行设计、重构和整合，将处理后的数据存储到分布式关系型数据库。结果表明，统一平台下的自动气象站资料接入和处理方法，能够使构建的系统具有较高的可靠性、低延迟和容错性等特点，满足自动气象站资料实时数据服务的时效性、准确性需求。

关键词：流式计算； 自动气象站资料； 实时统计

中分类号：P49 **文献标识码：**A

Automatic meteorological station data real-time processiong system based on stream computing

Guo Jie,Qiao Wenwen,Zhang Jinbiao,Kou Yuanyuan
(Guangdong Meteorological Observation Data Center,Guangzhou 510640)

Abstract: In order to solve the problems of decentralized and insufficient process in automatic meteorological station data processing business, a construction idea and method of automatic meteorological station data real-time processing system are proposed. The automatic meteorological station data is encapsulated into messages, the message middleware is used for data transmission, and the distributed stream computing engine is introduced. Based on the existing data decoding and statistical algorithms, the design, reconstruction and integration are carried out in the stream computing framework. The processed data is stored in a distributed Relational Database. This study realizes the access and processing of automatic meteorological station data under the unified platform. The system has the characteristics of high reliability, low delay and fault tolerance, and meets the timeliness and accuracy requirements of real-time data service of automatic meteorological station data.

Keywords: stream computing; automatic meteorological station data; real-time processiong

引言

自动气象站经过多年的建设和发展，已建成分布合理、密度适中的观测站网，截至2017年全国共有2400多个国家级自动气象站，5.8万多个区域自动气象站[1]，其中广东省已建设86个国家级自动气象站以及约3000个区域自动气象站，对短时临近预报、精细化预报和公共气象服务等实时业务非常重要[2-6]。自动气象站的观测频次达到分钟级，传输时效要求达到秒级，面对海量的自动气象站资料，如何进行针对性的调整优化资料传输流程[7-8]，合理有效地利用信息技术提高自动气象站资料的处理效率，在尽可能短的时间内完成处理、存储和服务[9]，为气象预报预测业务、科研和公众气象服务提供更快更好的支撑，是气象现代化建设过程中需要思考的问题之一。

因此，本研究以广东省自动气象站资料处理

收稿日期：2018-11-25
基金项目：广东省省级科技计划项目“广东省气象大数据科技协同创新中心(2018B020207012)”
作者简介：郭捷(1986-)，男，硕士，工程师，主要从事气象信息技术工作。E-mail: 313995@qq.com

为例，基于实时计算系统技术、消息中间件技术、分布式数据库存储技术，设计和实现了并行、流式的自动气象站资料实时处理系统，实现自动气象站资料的接入、解码以及变温变压、小时累计降水量、分钟累计降水量和日值统计等处理方法，整合及优化自动气象站资料业务流程，提高自动气象站资料的处理和统计效率，以期更好满足气象现代化业务的需求。

1 现状分析

当前广东省自动气象站资料的处理流程为：国家级和区域自动气象站资料文件通过 FTP 上传到 CTS，经过数据分发平台[10]的转发，到达解码入库服务器，由不同的作业程序进行解码和统计处理，结果写入实时资料数据库并同步到业务应用数据库[11]，为全省的数据业务应用提供数据支撑。

资料处理的实现主要通过文件轮询、数据库存储过程、定时任务等方式，存在一定的缺点：

(1) 文件轮询方式的时效性取决于轮询的间隔，多个环节的轮询会造成延迟的积累和放大。存储过程方式会占用服务端资源，对数据库服务器造成较大的压力。定时任务方式的时效性不足，数据到达情况和任务触发时间对数据处理的时效性和完整性影响较大，任务执行早了会导致晚来的数据没有被处理，执行晚了则会导致统计数据的时效性差，为兼顾时效性和完整性而进行频繁统计则会造成系统资源的浪费。

(2) 各类资料处理应用由不同的开发人员在不同的时期编写，其技术实现、统计算法可能存在差异；各任务之间相互独立，若某个任务对数据进行局部的更新但没有同步到其他环节，或某个任务出现异常时，该类数据与其他任务的处理结果将出现偏差，数据的全局一致性将得不到保障，降低了数据的可用性。基于上述原因，有必要对现有业务流程进行改进。

2 系统设计

自动气象站资料实时处理系统结构如图 1 所示，系统划分为数据源、数据处理、数据存储三个部分。

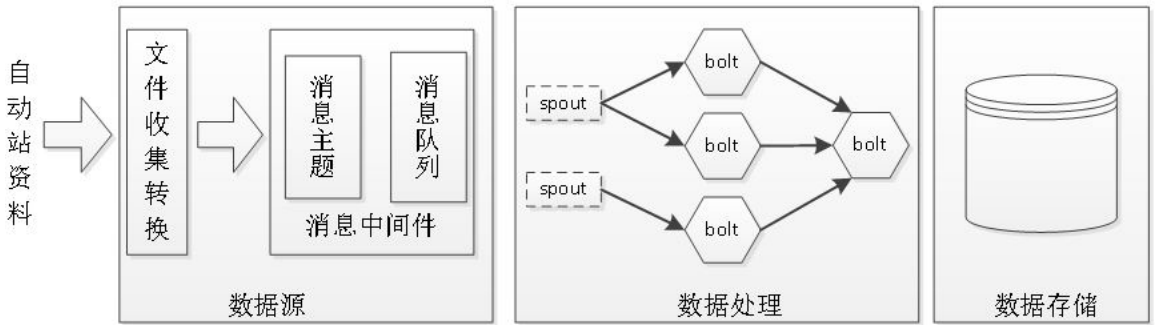

图 1 自动气象站资料实时处理系统结构图

(1) 数据源负责向实时处理系统提供自动气象站资料数据。文件收集转换模块集成 FTP Server 服务，将接收的自动气象站资料文件封装成消息后发送到 RabbitMQ 消息中间件。利用 RabbitMQ 的消息主题和队列路由控制功能，为系统提供数据交换服务，数据流转不再通过文件系统落盘中转，有效减少中间环节、降低耦合度和提高传输时效。

(2) 数据处理模块基于 JStorm 分布式实时计算引擎，利用 JStorm 提供的接口和框架，构建自动气象站资料处理的逻辑组件，各组件分别实现从消息中间件接收数据、数据解码以及进行变温变压、小时累计降水量、分钟累计降水量和日值统计等功能，将组件按数据流向和计算逻辑进行组合和定义形成任务拓扑，提交到系统运行。

(3) 数据存储采用分布式关系型数据库服务(Distributed Relational Database Service，简称 DRDS)，利用其分布式架构、高可用性、低延迟、高吞吐的特点，作为系统的数据基础存储平台，实现自动气象站资料原始数据和统计数据的快速更新和查询。

3 关键技术

3.1 自动气象站资料处理拓扑设计

如图 2 所示为自动气象站资料处理拓扑图。首先，两个数据源组件分别从消息队列中获取待处理的国家级或区域自动气象站资料，分别发送到对应的资料解码组件。然后，资料解码组件经过格式检查、字段拆分、要素转换等处理步骤，将资料数据转换为统一标准的观测要素字段名和观测值的键值对集合，分发到变温变压统计、小

时累计降水量统计、分钟累计降水量统计和日值统计组件进行气象要素统计。完成处理的数据写入 DRDS 数据库，结束一份资料的统计过程。

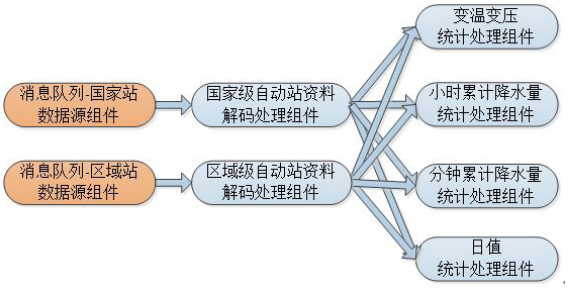

图 2 自动气象站资料处理拓扑图

3.2 要素统计方法设计

根据《地面气象资料实时统计处理业务规定》进行气象要素统计，设计合理的统计处理流程，提高统计效率。

资料处理需要保证无数据丢失，每一份数据都被正确处理，以保证数据的完整性。定义了数据处理状态标识，若自动气象站资料处理逻辑组件运行的过程中出现错误，如进程异常退出、数据库无法连接、数据更新查询失败、网络中断等情况时，将返回处理失败的数据处理状态标识，利用 JStorm 的消息跟踪机制，将处理失败的情况反馈到数据源组件，数据消息被重新放回 RabbitMQ 消息队列等待再次处理。

观测数据与统计结果必须保持一致，以保证数据的一致性。业务中常会出现更正报或者补传某个时次的缺失数据的情况，为了保持原始数据和统计数据的一致性，所有要素统计的逻辑中实现了关联更新功能，即统计时除了要完成当前时次的要素统计外，还需要判断出该时次的数据变更是否会影响其他相关时次的统计结果，如有则需要对其他时次进行重新计算，保证数据变更后所有关联的统计结果均被更新。

如图 3 所示为小时累计降水量统计关联更新示意图，以统计某站的小时累计降水量为例，处理程序在完成数据解码和统计当前时次的过去 2h、3h、6h、12h、24h 累计降水量之后，查询未来 23h 的时次是否有统计记录，若有记录并且当前观测时间包含在该时次的累计时间范围内，则需要重新统计该时次对应时段的累计降水量。

4 性能测试和分析

为检验系统的处理能力，将自动气象站资料处理任务的工作节点数分别设置为 1、2 和 4，并行度设置为工作节点数的 1、5、10、15、20 倍进

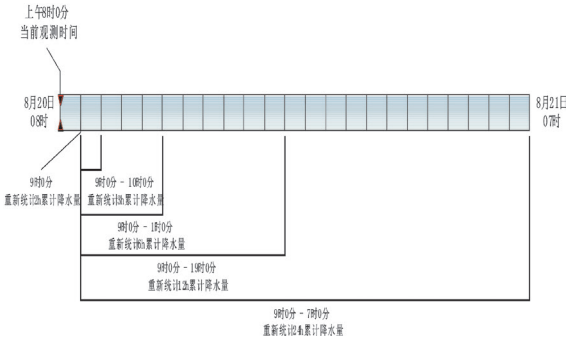

图 3 小时累计降水量统计关联更新示意图

行组合测试，根据单位时间内处理的报文数量计算得出自动气象站资料的处理速度。

测试环境共使用 5 台虚拟服务器，每台配置为 8 核 CPU、主频 2.4GHz 和 16GB 内存，测试数据为广东省一个月的国家级自动气象站 (75 万条，973MB) 和区域自动气象站 (1240 万条，12.1GB) 数据。

图 4 为不同工作节点数和并行度的自动气象站资料处理速度统计。从图中可以看出，在相同的工作节点数下，随着并行度与工作节点数的比值从 1 增加到 10，处理速度基本呈线性增加；当并行度与工作节点数的比值进一步增加到 15 和 20 时，处理速度继续增长但幅度逐渐减缓。在相同的并行度与工作节点数的比值下，随着工作节点数的增加，处理速度成倍增加。当工作节点数为 4、并行度与工作节点数比值为 20 时，处理速度为 2517 条 /s，即 1.5h 内可完成广东省一个月的自动气象站资料 (1315 万条) 入库和统计，实时数据在 1s 内即可完成消息读取、解码和统计的处理，满足自动气象站资料实时和统计数据的时效性需求。

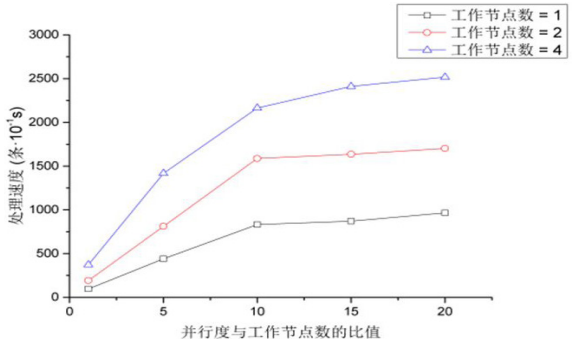

图 4 不同工作节点数和并行度的自动气象站资料处理速度统计

5 结语

针对当前广东省自动气象站资料处理业务中存在的问题，采用消息作为数据传输手段，简化自动气象站资料从接收端到资料处理端之间的流

程，提高传输的可靠性和时效性。引入分布式流式计算引擎，构建自动气象站资料实时处理组件，实现自动气象站资料解码、变温变压统计、小时累计降水量统计、分钟累计降水量统计和日值统计等功能，通过数据处理状态跟踪机制和关联更新机制，保障处理数据的完整性和一致性。测试结果表明，系统的处理效率可以满足气象业务时效性要求。

系统完成了基本功能的开发，但是也存在一定的问题，例如系统的部署和维护不够方便、对平台整体运行状态的监控不足、数据处理效率还存在进一步提升的空间等等，离业务化应用还有距离。在后续的研究和应用中，可以利用容器技术对平台进行改造，将平台的各模块进行容器化改造并部署到容器管理平台，提高应用部署效率，方便管理维护，利用容器平台提供的接口实现系统资源的监控；在处理组件的重要业务逻辑中设置埋点，采集资料处理的性能监视信息；与基于消息传输的标准格式 (BUFR) 自动气象站资料进行对接，进一步提高自动气象站资料业务的时效性。

参考文献：

[1] 孙超，霍庆，任芝花，等．地面气象资料统计处理系统设计与实现 [J]．应用气象学报，2018,29(5):630-640.

[2] 张蕾，王明洁，李辉．短时强降水临近预报相对准确率的探讨 [J]．广东气象，2015,37(2):1-6.

[3] 李磊，张立杰，力梅．深圳降水资料信息挖掘及在气候服务中的应用 [J]．广东气象，2015,37(2):48-51.

[4] 熊文兵，叶海宁，吴凤莹，等．基于移动互联的智慧气象为农服务系统研究 [J]．气象研究与应用，2018,39(3):63-65+91+132.

[5] 蒙炤臻，林奕桐，李仕强，等．自动站温度、雨量数据的质量控制方法和应用研究 [J]．气象研究与应用，2014,35(1):99-103

[6] 何健，王潜梅，钱光明，等．广东省区域自动气象站资料的质量控制与评估 [J]．广东气象，2011,33(3):37-40.

[7] 詹利群，黄祎萱，陈德诚．自动气象站中心站资料传输流程优化实践 [J]．气象研究与应用，2013,34(3):68-71.

[8] 张来恩，王鹏，韩鑫强．CTS2.0 消息封装及交换控制策略设计及实践 [J]．气象科技进展，2018,8(1):271-273.

[9] 赵文芳，刘旭林．Spark Streaming 框架下的气象自动站数据实时处理系统 [J]．计算机应用，2018,38(1):38-43.

[10] 张恩红，李高洁，乔文文，等．广东省气象数据通信系统的架构优化及应用分析 [J]．广东气象，2017,39(4):73-76.

[11] 寇媛媛，王晓明，杨玉红，等，分区技术在气象数据库优化中的应用 [J]．广东气象，2018,40(5):73-76.

（上接第 93 页）

[5] 李文娟，郦敏杰．基于探空资料的雷暴潜势预报方法 [J]．广东气象，2011,33(3):28-30.

[6] 覃晓玲，黎洁波，韦丽英．如何正确使用探空高度代替斜距的方法 [J]．气象研究与应用，2007,28(1)：85-86.

[7] 马佩强，张运林，李茂等．L 波段探空雷达跟踪异常成因及应对措施 [J]．广东气象，2008,30(2):89-90.

[8] 翁锦辉，罗建平，王凡，等．气象探空数据动态比对中误差计算方法研究 [J]．气象研究与应用，2010,31(3):74-76.

[9] 姚日升，曹艳艳，涂小萍．插值方法在提高热带气旋路径预报时效分辨率中的应用 [J]．广东气象，2011,33(1):13-15.

[10] 王超球，许嘉玲．区域自动气象站质量控制参数值高度订正的方法 [J]．气象研究与应用，2011,32(3)：=64-66.

[11] 翟盘茂．中国历史探空资料中的一些过失误差及偏差问题 [J]．气象学报，1997,55(5):563-572.

[12] 郭艳君．高空大气温度变化趋势不确定性的研究进展 [J]．地球科学研究进展，2008,23(1):24-29.

[13] Paule.Ciesielski,WEN-Ming,SHAO-Chin,et al. Quality-Controlled Upper-Air Sounding Dataset for TiMREX/SoWMEX:Development and Corrections. [J].American MeteorologicalSociety,2010,46(2):330－351.

[14] Paule.Ciesielski,WEN-Ming,SHAO-Chin,et al. Quality-Controlled Upper-Air Sounding Dataset for DYNAMO/CINDY/AMIE:Development and Corrections [J].American Meteorological Society,2014,78(3):443-462.

[15] 许小勇，钟太勇．三次样条插值函数的构造与 Matlab 实现 [J]．自动测量与控制，2006,25（11）：1006-1576.

[16] 朱亚玉，宋丽莉，姬兴杰．基于分段三次样条函数逐时气象资料模拟方法研究 [J]．气象与环境学报，2017,33(2):44-52.

[17] 周冰，李玉立．GPS 掩星大气探测中数据的平滑处理方法分析 [J]．城市勘测，2018,4(3):88-96.

[18] 李平，徐枝芳，范广洲，等．探空温度资料质量控制技术研究 [J]．气象，2013,39(12):1626-1634.