基于消息的气象服务数据加工流水线设计与应用

宋瑛瑛　李雁鹏　陈正廷　凌柏　冯宇星　曹宇钊
（中国气象局公共气象服务中心，北京 100081）

摘要：梳理了公共气象服务业务的数据流程，提出了一种基于消息的气象服务数据加工流水线技术，用于提高气象数据服务的时效性与数据资源更新的一致性。着重介绍了消息中间件、流水线架构设计、交互文件标准等技术问题。结合分钟级降水预报数据实例，初步展示了消息驱动数据流应用在气象服务数据支撑系统中的优势。

关键词：气象数据，消息中间件，交互文件标准，系统设计

DOI：10.3969/j.issn.2095-1973.2016.06.008

Design and Application of Message-Driven Data Processing Flow in Meteorological Service

Song Yingying, Li Yanpeng, Chen Zhengting, Ling Bai, Feng Yuxing, Cao Yuzhao
(Public Meteorological Service Center of China Meteorological Administration, Beijing 100081)

Abstract: Based on the analysis of public meteorological service data flow, this paper proposes a message-driven data processing technology in the meteorological service to improve the timeliness of meteorological data services and the uniformity of data resources. The key technologies, such as message middleware, system architecture, interactive file standard, are introduced here. An example of PM-MQPF data is described in detail to express the advantages of data flow, which was driven by messages in meteorological data services system.

Keyword: meteorological data, message-oriented middleware, interactive file standard, system design

0　引言

气象服务是人类生产生活的基础条件，它不仅与公众生活密切相关，同时也直接影响政府决策、国民生产以及社会发展。随着信息网络技术的不断发展，各行各业对气象服务的要求越来越趋于精细化，随之而来，分门别类的气象服务产品层出不穷。为满足各类应用需求，数据信息交互单元也逐渐变得庞杂。当前，在中国气象局信息化建设的大背景下，公共气象服务信息化建设的需求越来越紧迫，集中表现在从分散向集约化发展，提供统一的气象数据服务。因此，本文梳理了国家级公共气象服务业务的数据流，分析了现存问题，提出了基于消息的气象服务数据加工流水线技术，旨在提高气象数据流的时效性和鲁棒性，从而更好地支撑气象服务业务的快速发展。

1　消息中间件

随着气象业务的跨越式发展，当前中国气象局公共气象服务中心应用的气象服务数据系统存在数据处理流程交叉、运行效率低、业务耦合性强等问题。本文提出利用消息驱动业务流程，解决上述问题。具体而言，利用消息中间件高效可靠的传递机制进行平台无关的数据交换，基于数据通信进行分布式系统的集成。发送者将消息发送到消息服务器，消息服务器将消息存放在若干队列中，在满足触发条件后再将消息转发给接收者。消息传输模型分为点对点模型（PTP）和发布/订阅模型（pub/sub），发布/订阅模型具有异步、松耦合、多对多通信等特点。发布者广播发送消息给中间代理，订阅者只需去代理中接收自己感兴趣的消息，发布者并不知道究竟是哪个订阅者接收到了自己发布的消息。

目前，市面上常见的发布/订阅模型消息中间件有：Kafka、ActiveMQ、OpenJMS和RabbitMQ等。国内各气象部门在应用消息中间件传递数据方面做了很多研究，例如：应用ActiveMQ技术缩短了预警信号发布时耗[1]，基于RabbitMQ消息中间件和元数据技术实现了多种气象观测资料的采集、传输、入库和备份的统一处理[2]。通过业务调研，本文以Kafka消息中间件为核心，在集约资源的基础上，兼顾性能，设计了基

收稿日期：2016 年 9 月 21 日；修回日期：2016 年 11 月 25 日
第一作者：宋瑛瑛（1983—），Email：syy0822@126.com
资助信息：中国气象局公共气象服务中心业务基金项目（K2016006）

于消息的气象服务数据加工流水线，用以提高数据流动效率，实现数据内容统一。

2 气象数据流

气象数据流具有多来源、高并发及更新快等特点。针对这些特性，以往的气象数据处理过程，多采用多进程并发技术[3]，基于web服务[4]，实现气象数据实时共享；近年来，多采用消息驱动机制处理气象数据[1-2, 5]，以期提升传输效率。

2.1 数据流分析

公共气象服务中心的气象数据服务主要包括专业气象和公众气象两大部分。其中，专业气象为预警信息发布、自然灾害决策等提供数据服务；而公众气象为常规天气预报制作提供数据支撑，主要服务对象有中国气象频道、中国天气网、中国天气通APP等媒体和天气服务软件。在现有气象数据服务系统中，存在以下不足之处：1）两大业务存在数据源重叠交叉，数据重复采集、多次加工，数据样本多重复制，在极端条件下，会出现气象服务产品数据内容不一致的情况。2）关联任务执行时存在任务堆积、耗时长等问题。在部分业务执行时间不确定的条件下，为了保障业务数据的完整性，需延长定制计划任务时间范围，这就增加了后续处理环节的等待时间，造成延迟。3）对超出定制时间范围的数据无法实现自动处理，只能通过手动执行应急处理，无法做到全自动无人值守。4）数据读写操作频繁，增加服务器的并发负载，造成资源浪费。可见，定时计划任务是拖延数据流动的重要一环，而消息即来即走、分布式并发等特性正可以解决这些问题。

2.2 消息驱动数据流水线

基于消息驱动的数据流，将流动过程分解成主题，通过发布/订阅模式，即时发布主题至消息总线，计算集群监听消息，即时处理主题，缩短了数据流动时间。这种分布式、高并发、无延迟的集约型数据流为下游气象产品制作业务提供即时的数据支撑。如图1所示。

1）业务数据流解耦。根据业务逻辑关系，将数据流分解为多个独立执行单元，即消息。同时为每个消息建立主题（topic），例如数据读取主题、数据处理主题、数据发布主题等。消息主题不具有业务意义，仅作为消息总线上的标识使用。

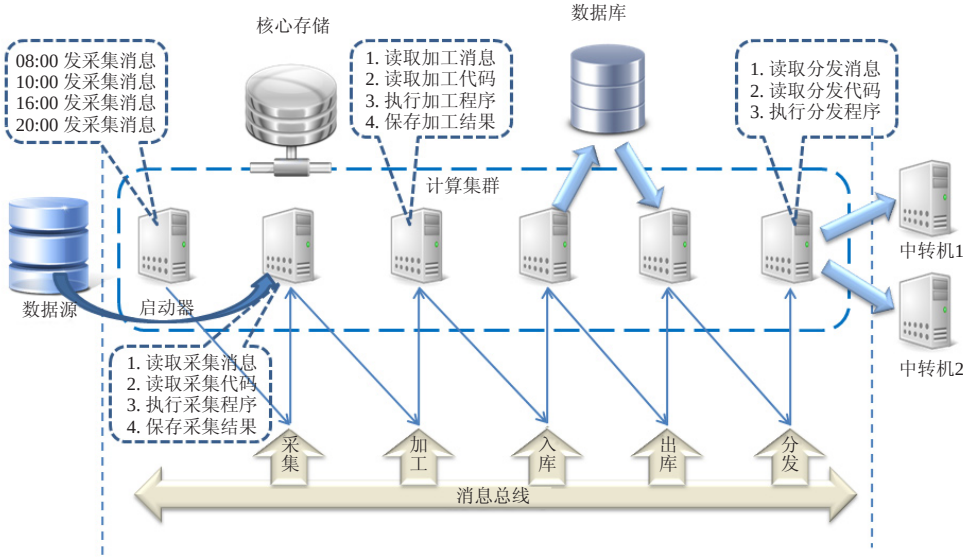

图1 数据流水线
Fig. 1 Data flow

2）数据即时加工处理。生产者（producer）发布主题至消息总线，消费者（consumer）随时监听消息总线上已订阅的主题，一旦发现触发消息，立即执行。本文中的消费者由多台物理机或虚拟机组成基于SSH协议的计算集群实现，集群中的每个节点是无状态的，可并行消费多个主题。

3）数据流分发。处理完成的数据按需分发至下游业务系统。

2.3 两种数据流程特性比较

消息驱动数据流摒弃了定时计划任务模式，采用了发布/订阅消息模式，不断地从消息总线上监听消息，一旦接收到消息即刻触发执行主题任务。由计算集群采用分布式计算完成任务的并发操作，改变了传统数据流集中控制、顺序执行任务的处理过程，极大

地提高了数据处理效率。目前，气象消息总线的主要应用是业务集成和数据共享。从业务集成方面看，可以有效解耦气象数据流处理过程，提高信息交互能力和功能复用水平。从数据共享方面看，消息总线为数据的及时更新和一致性提供了可靠保障。两种数据流程特性比较如表1所示。

表1 两种数据流特性对比

Table 1 Comparisons between two kinds of data flow

类型	传统数据流	新数据流
系统架构	紧耦合	松耦合
交互模式	计划任务	发布/订阅消息
逻辑处理	集中控制	分布式并行
数据内容	不一致	一致
时间	有延迟	无延迟
资源	重复浪费	统一管理

3 系统设计与应用

3.1 系统架构

为了能更好地集成现有业务数据流处理过程，基于消息的气象服务数据加工流水线设计采用三层架构实现，分为数据层、消息中间层和业务层（图2）。

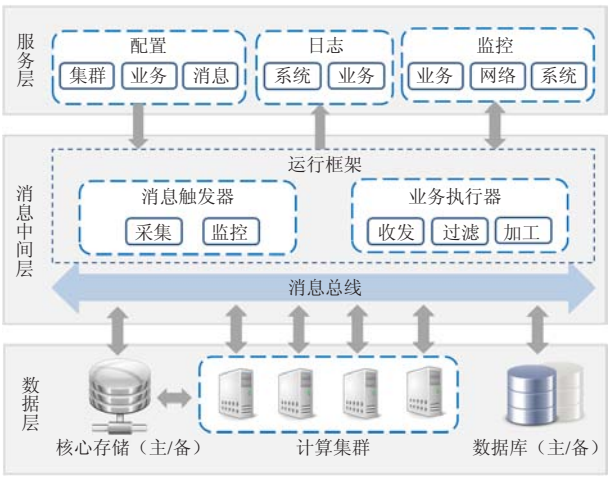

图2 系统架构

Fig. 2 System architecture

数据层是气象服务数据加工流水线的基础。数据文件存储在大型分布式物理设备中，计算集群是数据流各环节的真正执行者，将存储和计算隔离开，实现了数据与计算的解耦。

消息中间层在本架构中具有承上启下的作用，在Kafka消息总线的基础上实现业务运行框架。将一条业务数据流处理过程分解为启动和执行两部分，启动模块是处理过程的起点，是消息总线上的第一条消息，执行模块监听到该消息后分步骤并行处理业务程序代码，将结果消息发送至总线。本文将启动和执行模块封装为消息触发器和业务执行器。

根据应用需求，本文设计了两种消息触发器：1）采集数据：根据业务配置开始时间和间隔，轮询向上游数据提出采集要求，采集完成后即刻发出指令，启动本次业务数据流处理过程；2）监控数据：根据业务配置目录和文件（可多个），监控数据状态，完整到达后即刻向消息总线发出消息，触发执行下一环节消息主题。

业务执行器由消息收发、消息过滤和业务程序执行代码组成：1）消息收发：计算集群各节点监听订阅的主题消息，处理完成后，向消息总线发送本环节消息结果；2）消息过滤：过滤重复消息，防止反复触发造成消息异常堆积；3）业务程序执行代码：分解数据流处理过程，封装业务单元程序执行代码，由计算集群并行处理。

服务层是气象服务数据加工流水线的窗口。配置服务为运行框架提供集群、业务和消息等参数配置；日志服务记录业务运行状态；监控服务为运维提供了可视化用户界面。

3.2 关键技术

为解决数据处理时效性问题，本文设计了一种基于Kafka消息总线的业务运行框架，实现了气象服务数据即时处理；为解决业务层与运行框架间的数据交互问题，制定了规范性xml消息配置，提供了统一的参数标准。

3.2.1 基于 Kafka 的消息驱动

Kafka包括生产者、消费者和代理者三个部分。生产者生产的消息被放在主题中，一个主题可以设置多个分区，推送消息给代理者，等待消费者接收；消费者在想要订阅消息时，向代理者发出请求，告知其主题和分区值，主动拉取消息（图3）。数据流经过解耦，运行在消息总线上，经多个计算集群即时执行处理，迅速将加工数据分发出去。

图3 消息发布/订阅简图

Fig. 3 Message publish/subscribe

以Java代码为例，介绍消息的生产者和消费者。

1） 创建生产者

```
// 创建生产者属性参数
Properties  dps = new Properties();
//指定代理服务器
dps.put("metadata.broker.list",kafkahosts.toString());
//用该属性参数创建生产者
```

```
Producer<String, String> producer = new Producer
<String, String>(new ProducerConfig(MsgController.dps));
//生产者广播发送topic/link消息至消息总线
producer.send(new KeyedMessage<String,
String>(topic, new Date().getTime() + "" ,next-link-id));
```

2）创建消费者

```
// 创建消费者属性参数
Properties cps = new Properties();
//指定消费组
cps.put("group.id", group);
//消费者zookeeper 配置
cps.put("zookeeper.connect", zookeeperhosts.
toString());
//用消费者属性参数创建消费者
ConsumerConnector consumer = kafka.consumer.
Consumer.createJavaConsumerConnector(new
ConsumerConfig(cps));
```

3.2.2 消息配置格式

xml文件提供统一的方法来描述结构化数据，通常作为异构系统间数据交换的格式[6]。参数配置采用xml文件描述，一个业务对应配置一个xml文件，设计xml文件命名规则为业务种类（2位数字）+业务序列（5位数字）.xml，例如：1010001.xml。

消息配置内容由三部分组成：

1）service：业务基本信息。name：业务名称；type分两类：采集（collection）和监控（notify），与消息触发器类型对应。

2）type描述：分为采集和监控标签，为消息触发器提供所需参数信息。①采集标签：description：描述采集业务详情；start-time：开始时间；period：时间间隔。②监控标签：description：描述监控业务详情；path：要监控的文件目录；files：监控的数据源文件。

3）link：主题参数。id：标识主题单元号，由业务种类（2位数字）+业务序列（5位数字）+环节号（3位数字）共10位数字组成；description：描述该环节主要操作内容；order：描述主题执行命令代码；next-link：指定下一topic/link对。

3.3 业务应用

实时天气预报是网站和移动终端用户的重点选择气象服务，具有广泛的用户群体和极大的需求。本节以雷达分钟级降水系统（PM-MQPF）数据为例，阐述业务数据流实际应用。分钟级降水数据来源于214部雷达站逐6分钟实时观测产生的基数据，经过计算处理，形成临近降水预报服务产品。在现有读取基数据、前处理、主处理、拼图4个处理程序基础上，通过Kafka消息驱动机制进一步将这四部分分解剥离，形成独立单元流水线作业模式，通过单元返回值触发后续单元主题。将在整个数据计算加工过程中耗时较长的前处理部分，运用计算集群并行处理，以提高计算效率。

1）分解PM-MQPF数据处理过程。封装为基数据、前处理、主处理和拼图四个部分，为每部分增加返回值。

2）配置PM-MQPF数据消息参数。为监控指定目录和文件，为业务环节分配主题，T1（前处理），T2（主处理），T3（拼图）。

3）消息总线监听文件，有文件到达即发送消息。

4）计算集群监听消息，即时处理主题。T1分布式并行对每个数据文件进行前处理；T2统计每6分钟前处理完成的情况，处理量完成80%以后发送消息启动主处理，主处理完成之后再发送消息；触发T3加工生成拼图，完成后通知下游环节进行产品分发。

PM-MQPF分钟级降水数据流水线采用消息机制和计算集群，分解封装处理单元，降低了各处单元的耦合度，提高了系统运行效率。因此，气象信息总线的应用不仅降低了系统开发和运行成本，也提高了业务应用的实时性，同时为后续拓展应用服务提供了空间。

4 结束语

基于消息的气象服务数据加工流水线降低了业务耦合度，优化了气象数据流程，为进一步实现气象数据实时同步、提高数据流时效性奠定了必要基础。当前，系统在功能、性能和稳定性方面都有优异的表现，不但能够满足下游各气象业务的实际需求，同时为业务拓展提供了空间。在后续的研究计划中，计算集群的高并发、负载均衡等问题将被进一步探索。

参考文献

[1] 钱峥, 胡亚旦, 黄旋旋. 基于“消息中间件”技术的气象信息总线. 气象科技, 2016, 44(2): 217-222.
[2] 韩笑, 王力, 王吉滨, 等. 一种地市级气象数据库的设计与应用. 气象科技, 2015, 43(6): 1053-1059.
[3] 胡英楣, 沈文海, 宋之光. 多进程并发在国内气象通信系统的应用. 应用气象学报, 2007, 18(6): 877-884.
[4] 王甫棣, 林润生, 胡英楣. 基于Web服务的气象数据服务. 计算机工程, 2009, 35(8): 280-282.
[5] 王力, 韩笑, 刘培宁, 等. 基于MQ的气象数据采集与监控系统设计. 气象科技, 2015, 43(3): 451-457.
[6] 曹卫. XML技术在气象信息发布系统中的应用. 福建电脑, 2009, 25(4): 110-111.