

何林,高亮书,冯洁琼,等. 省级多源融合实况产品融入“天擎”关键技术研究与应用[J]. 陕西气象,2024(2):69-75.

文章编号:1006-4354(2024)02-0069-07

省级多源融合实况产品融入“天擎” 关键技术研究与应用

何林^{1,2},高亮书²,冯洁琼²,陈婧²

(1. 陕西省气象局秦岭和黄土高原生态环境气象重点实验室,西安 710016;

2. 陕西省气象信息中心,西安 710014)

摘要:按照“云+端”技术架构对气象数据、算力、算法资源进行集约管理的思路,基于气象大数据云平台(“天擎”)提供的“数算一体”服务能力,研究省级多源融合实况分析系统的业务产品融入“天擎”的关键技术。梳理省级数据产品注册、发布和共享的基本流程,归纳并阐述了通信系统分发入库、算法改造写入、调用标准化接口写入等三种实时数据接入方案与实现方法。通过对比各方案的优缺点,给出不同应用场景下的最优接入策略,为其他各类本地化特色数据产品融入“天擎”提供借鉴。

关键词:实况产品;天擎;通信系统;加工流水线;MUSIC接口

中图分类号:P409

文献标识码:A

获取天气的实时状态是人民群众日常生活中对气象需求最旺盛的服务之一。采用种类繁多、来源丰富的气象观测数据开展传统的气象服务存在数据意义不统一、质量不一致,分散无序发布等缺点,无法实时、精准地反映出大气的状态和结构。近年来气象部门形成了以研发多尺度、高精度、高时效、时空连续的格点化实况分析场(简称“实况产品”)为核心的多源融合实况分析业务(简称“实况业务”)[1]。在当前中国气象局提出的构建以气象大数据云平台(China meteorological administration data as a service,CMADAAS,中文简称“天擎”)为“云”,气象业务系统为“端”的“云+端”新型业务技术体制改革中[2],实况业务也面临“云化”改造和融入“天擎”运行。研究省级多源融合实况分析产品融入“天擎”的主要技术,是实现实况业务“云化”运行的核心与关键环节。

1 基于“天擎”的省级实况业务体系介绍

实况业务是指各级气象部门通过多种手段有序衔接、快速精准的提供不同尺度实时、精准的大气状态、结构、成分、现象等的业务。作为智能网格天气预报和气候预测的“零时刻”起点,实况业务已成为无缝隙预报业务的重要组成部分[3]。“天擎”在国家级与省级布局,建立在基础设施资源池之上,以数据为主线,采用云计算、大数据、分布式存储架构等核心技术,构建了数据交换质控、产品加工、挖掘分析、存储服务等四大子系统[4],其与综合业务实时监控平台(“天镜”)对接,实现“全流程、一体化、可视化”的业务监控[5]。省级实况业务体系的核心是基于“天擎”加工流水线(Data processing line,DPL)容器环境运行的多源融合实况分析系统,充分利用“天擎”对数据、算力、算法集约管理提供的“数算一体”服务能力,实现地面观测、雷达、卫星、数值模式等多源数据的融合应用。系统的基本组成和数据流如图1所示。

收稿日期:2023-03-22

作者简介:何林(1987—),男,汉族,陕西武功人,高工,硕士,主要从事气象信息技术、气象数据管理与应用。

基金项目:陕西省气象局秦岭和黄土高原生态环境气象重点实验室开放基金课题(2022K-6);陕西省重点研发计划(2023-YBSF-191)

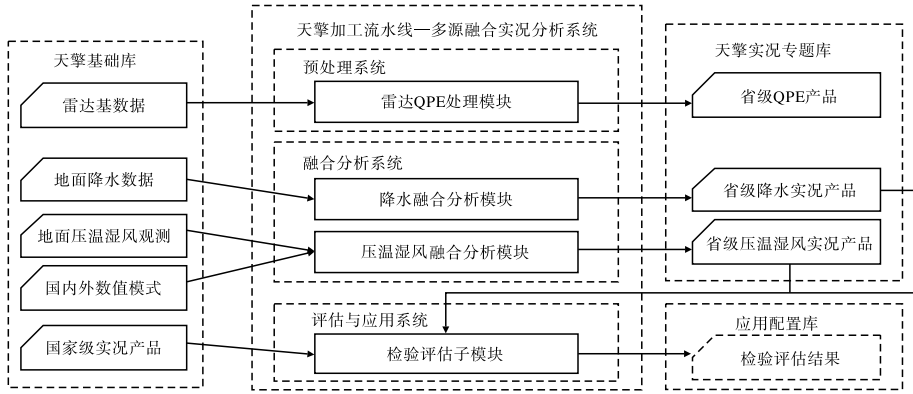


图1 基于“天擎”的省级多源融合实况分析系统

该系统集成雷达 QPE,降水融合分析,气温、气压、湿度、风的融合分析、检验评估等四个功能模块,可以基于“天擎”独立运行,也可以协同使用,支持省级实时生成本地化雷达 QPE 及降水、气温、气压、湿度、风等 3 大类 12 种产品,产品分辨率达 1 km、10 min,时效达 2~3 min。目前,省级多源融合分析系统已实现了数据源接入“天擎”、算力由“天擎”弹性供给、算法基于“天擎”DPL 运行、监控与“天擎”对接的改造,构建了省级实况产品制作共享、检验评估、应用服务的业务全流程,有效推动建成集约化的“国省实况一张

图”。但由于该系统生产的多种本地化实况分析产品并非由国家级下发、国省统一的通用数据产品,其融入“天擎”的标准尚未统一;因此,以此为切入点,将实况产品作为省级特色数据的典型案例开展融入技术研究。

2 实况产品融入总体思路

按照“天擎”数据融入规范,实况产品融入“天擎”的总体技术路线如图 2 所示。其中,前三步主要是准备工作,实时数据接入为数据融入“天擎”的核心,产品发布和接口适配主要是为了实现数据的全局共享。

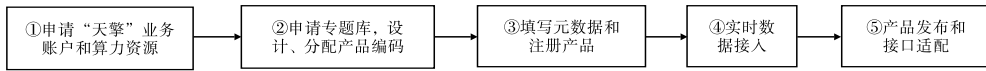


图2 陕西省气象实况分析产品融入“天擎”的技术路线

在准备工作中,首先使用个人证书实名登陆“天擎”门户网站,为多源融合实况分析系统申请业务账户,并按需开通数据存储、MUSIC 接口、算力资源权限。账户和权限审核通过后,在该业务账户下创建实况产品专题库,为每个产品分配标准的四级编码。专题库是“天擎”存储管理系统(Service-oriented data storage system, SOD)采用多租户模式,为应用系统提供具备资源隔离、独立管理的逻辑数据库。包括结构化数据库、历史分析库、实时应用库、文件索引库等类型。按照实况产品的特点,在“天擎”中申请创建省级多源融合实况产品专题库,并采用文件索引库+NAS 存储的专题类型。四级编码是“天擎”按照《气象资料分类与编码》标准规范^[6]建立的气象数据全局目录,其为每类数据分配唯一 ID,并表征该资料

处于其生命周期全流程中的位置。

四级编码体系类似网络 IP 地址,共分为四段,用“.”分割。其中:第一段使用表 1 所示的大写字母,标注气象资料所属 15 大类之一;第二段为国家级和省级编码标识,其中陕西省的数据标识为 8200-8299;第三段为按照服务需求,设立具有独立应用元数据的资料子集;第四段描述血缘关系、进行系统间的数据追溯^[7]。以陕西多源融合降水实况分析产品为例,设计四级编码追溯关系如图 3 所示。

所有实况产品编码设计规划完成后,在门户网站用户空间中注册数据,其核心是填写元数据表单。元数据是描述数据产品基本信息的数据。表 2 以陕西多源融合降水 1 km/10 min 实况分析产品为例,按照“天擎”存储管理系统的要求填写

表 1 四级编码第一段标注的气象资料所属大类

序号	资料大类	字母	序号	资料大类	字母	序号	资料大类	字母
1	地面资料	A	6	数值分析预报产品	F	11	卫星资料	K
2	高空资料	B	7	大气成分资料	G	12	科考和实验资料	L
3	海洋资料	C	8	历史气候待用资料	H	13	气象服务产品	M
4	辐射资料	D	9	气象灾害资料	I	14	空间天气资料	N
5	农业和生态资料	E	10	雷达资料	J	15	其他	Z

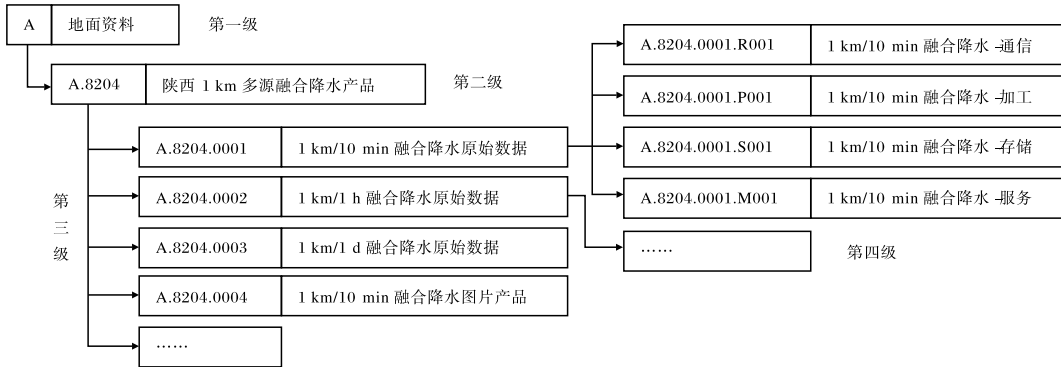


图 3 气象资料四级编码追溯关系示例

表 2 “天擎”中陕西多源融合降水 1 km/10 min 实况分析产品元数据表单示例

序号	元数据	赋值示例
1	资料种类	[A]地面资料
2	数据名称	陕西多源融合实况分析 1 km 分辨率 10 min 降水实时产品
3	数据分级	Ⅲ级数据(内部数据):仅供内部或特定用户使用的敏感数据
4	数据范围	省级发布的特色资料
5	四级编码	存储编码 A.8204.0001.S001;服务编码 A.8204.0001.M001
6	数据集代码	SURF_BEXA_ART_PRE_1KM10MIN_NC
7	数据描述信息	知识产权信息;空间分辨率 1 km;地理覆盖范围为陕西格点责任区...
8	关联表信息	实况分析专题库-文件索引库,表名 SURF_BEXA_ART_PRE_FILE_TAB
9	存储路径与策略	/CMADAAS/DATA/SURF/ART...;存储时长 10 a;空间估算 4 TiB
10	服务信息	发布 MUSIC 接口服务[是];发布目录服务[是];存入对象存储[否]

元数据表单的示例。

完成数据注册并审批通过后,“天擎”将按照填写的元数据信息,自动在后台建立数据表、存储目录,配置相应的存储策略。以上完成了准备工作,管理员可在后台数据表检索数据进行验证,此时,数据表存在但查询结果为空,为实现数据入库,还需进行关键的实时数据接入。

3 实时数据接入

实时数据接入主要目的是将数据写入已建成

的专题库数据表中,对于结构化数据,主要是写数据库操作,对于非结构化文件,包括写索引数据库和文件实体按规划目录落盘。实况产品皆为非结构化文件,按照其特点,提出三种实时数据接入方式。

3.1 通信系统接入

第二代国内气象通信系统(China meteorological administration telecommunication system,CTS2)是从气象卫星综合应用业务系统

(9210工程)建立的基本技术框架下逐步发展扩充,集数据收集、预处理、分发交换、传输监控等功能于一体的应用系统^[8],在整个气象数据全流程中处于“天擎”的上游。CTS2支持文件、消息、流三种数据传输方式,采用CTS2提供的基于FTP协议的文件数据分发策略,按照实况产品的特点设计文件名模板和接口目录^[9],采用目录轮询方式将接口目录中的实时文件分发至下游的“天擎”

解码处理系统(data decode process center, DPC),同时发送消息通知,DPC订阅对应数据的消息队列后,利用常驻内存的数据解码进程对目标目录的文件进行处理,实现文件重命名、索引入库和搬移等步骤。整个接入过程几乎不需编写代码,仅对CTS2和DPC的相关配置文件进行更新后重启解码进程即可,以陕西多源融合降水1 km/10 min实况分析产品为例,核心配置项见表3。

表3 通信系统接入实时数据的CTS2和DPC核心配置项示例

序号	所属系统	配置名称	配置项示例
1	CTS2	通信编码	A. 8204. 0001. R001
2	CTS2	文件名模板	Z_SURF_C_BEXA_%t_P_CMPA_RT_REG_0P01_TMIN-PRE-?????????????.nc
3	CTS2	分发策略	接口目录 SURF[地面资料],文件名检查,格式检查
4	DPC	消息转发	DataRouteConf. xml 配置文件中增加一段 <pre><DataRoute> <DataType>A. 8204. 0001. R001</DataType> <ExchangeName>X. OBS. CMADAAS</ExchangeName> <ExchangeType>topic</ExchangeType> <RoutingKey>SURF_ORI_A. 8204. 0001. R001_001</RoutingKey> <QueueName>SURF_ORI_A. 8204. 0001. R001_001</QueueName> </DataRoute></pre>
5	DPC	入库配置	Table. xml 配置文件中增加一段 <pre><table> <dataType>A. 8204. 0001. R001</dataType> <storeType>9</storeType> <!-- 通过 index. txt 配置入库流程 --> <indexTable>SURF_BEXA_ART_PRE_FILE_TAB</indexTable> <!-- 索引表--> <sodDataType>A. 8204. 0001. S001</sodDataType> <splitRegex>-. _</splitRegex> <!-- 文件名中的分隔符--> <storyPath>~/SURF/BEXA/[9]/[10]/[12(0,4)]/[12(0,8)]</storyPath> </table></pre>
6	DPC	索引规则	Index. txt 配置文件中增加一段 <pre>[A. 8204. 0001. R001] D_FILE_ID=d_file_id D_DATA_ID=sod_type D_IYMDHM=sysdate D_RYMDHM=recvtime D_DATETIME=' \${12}00' D_STORAGE_SITE=newpath</pre>

续表 3

序号	所属系统	配置名称	配置项示例
		索引规则	D_FILE_SIZE= filesize V_CCCC='\${ {3} }'
7	DPC	私有配置	新建 SURF_A. 8204. 0001. R001. prperties # RabbitMQ 消息中间件服务器监听附列配置 RabbitMQ. queueName= SURF_ORI_A. 8204. 0001. R001_001
8	DPC	进程启动	java-jar dpc_general_frename. jar8204 SURF_A. 8204. 0001. R001. properties

3.2 算法改造接入

“天擎”继承于全国综合气象信息共享平台 (China integrated meteorological information service system, CIMISS), 最大特色是新增了算法库和 DPL 子系统, 通过提供基础数据和算力资源, 按需适配多样化的调度策略, 实现对各类气象算法的集约管理和运行^[10-11]。实况产品通过算法改造方式进行接入时, 剥离当前省级多源融合实况分析系统的代码逻辑, 包括数据输入、加工处理、产品输出 (data input, process and output, IPO) 三部分, 重点对算法的产品输出部分代码进行重构。针对非结构化的实况产品的输出, 需进

行文件迁移和索引入库两步, 分别按照表 2“存储路径与策略”和“关联表信息”元数据项确定文件落盘的目标路径和写入索引的目标表名。写索引时可采用直连数据库的方式, 在代码中加载实况产品专题库选型的虚谷数据库驱动开发包, 创建连接、拼接并执行 SQL 插入语句, 最后释放连接即可。对改造后的算法在“天擎”门户网站依次进行注册、上传、加载, 并定义任务、环节和运行策略等操作^[12]。表 4 以陕西多源融合降水 1 km/10 min 实况分析产品写入算法为例, 在网站的算法管理部分填写表单的示例。

表 4 “天擎”门户网站算法管理的核心表单项示例

序号	步骤名称	表单项	内容
1	算法注册	基本信息	算法名称: 陕西多源融合降水 1km/10min 实况分析产品写入; 开发语言: java; 版本号: V1.0; 所属分类: 通用算法; 发送 DI: 否; 输入输出:
		知识产权	贡献者、注册人、维护人、算在单位、联系方式等, 按实际填写
		运行环境	操作系统: redhat7.5; 计算框架: 普通计算; 计算资源: 2 核 CPU/4GB 内存
2	算法加载	算法上传	可执行文件、依赖包、源码(可选项)上传及说明; 运行命令、运行参数
		部署信息	选择部署节点; 日志保留天数: 默认 30 d; 加载方式: 自动加载, 其它默认
3	任务定义	环节信息	环节名称、关联算法模型; 触发方式: 数据源感知-输入数据到报率大于 95% 自动触发; 前置环节: 无; 失败重试: 是; 告警规则配置
4	任务管理	可操作项	启动、停止、人工触发、修改、删除、查看日志

3.3 MUSIC 接口写入

气象数据统一服务接口 (meteorological unified service interface community, MUSIC) 遵循《气象数据服务接口规范》行业标准, 基于“天擎”数据环境, 面向气象业务和科研, 提供了统一、标

准、丰富、具有多种调用形式的数据访问服务和应用编程接口, 为各级应用系统提供直接支撑的数据接入服务^[13]。MUSIC 提供了数据读写、统计、格点解析等标准化接口, 支持主流的编程语言和 REST、客户端、脚本等调用方式。气象资料通过

MUSIC 接口接入“天擎”,本质是将实时数据解译组装、拼接 SQL 语句、写入数据库表等操作整合,以应用编程接口 API 方式提供给用户,实现

气象资料入库^[14]。以陕西多源融合降水 1 km/10 min 实况分析产品通过 MUSIC 写入接口实现接入为例,基本流程如图 4 所示。

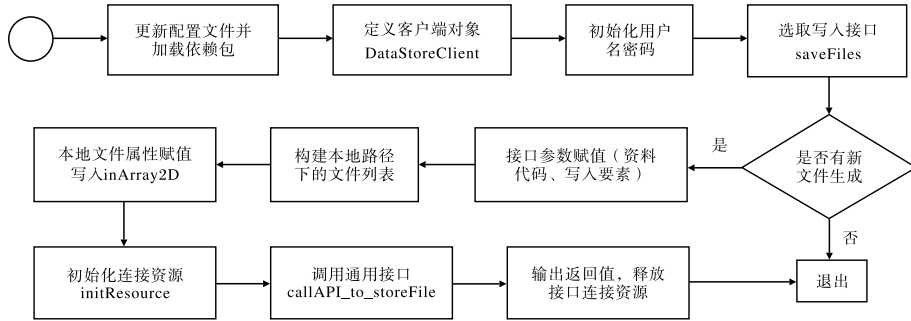


图 4 使用 MUSIC 写入接口接入实况产品流程图

3.4 三种接入方式对比

对三种实时数据接入方式从时效性、开发代

码量、文件规范性要求等指标进行对比,如表 5 所示。

表 5 三种接入方式重点指标对比

指标项	通信系统接入	算法改造接入	写入接口接入
时效性	取决于目录轮询配置	最高	适中
代码量	少	多	适中
文件名	标准长文件名	不要求	标准的长/短文件名
配置操作	多	适中	少
通用性	低	适中	高

通信系统接入的最大优势是无需编程,按照实况产品的文件规则添加配置项,复用现有解码进程实现接入;但仅适用于通信系统能够识别的标准化长文件名的资料,且配置环节较多,任何一个环节出问题都会影响数据的入库和服务。算法改造接入需要对原有代码逻辑进行剥离和修改,可以通过直连数据库即时生产和写入,时效最高;但由于对非结构化文件需要同时写 NAS 和索引库,因此开发代码量最多。适用于加工与生产逻辑与算法强绑定,且对时效要求非常严格的数据产品。MUSIC 接口写入方式需调用标准化的编程接口,相当于通过封装好的算法实现写 NAS 和索引库的步骤,灵活性相对较差,且调用接口时需要组织本地文件的元数据作为写入接口参数赋值,因此要求文件名能够体现数据的基本信息,即标准的长/短文件名,综合而言,通用性较高,且开发量适中。

4 数据发布和使用

数据接入后,为告知用户新资料的相关信息及获取方式,需在门户网站进行数据发布。用户一般可通过 MUSIC 接口和直连数据库两种方式获取数据,其中,后者仅适用于 DPL 中的算法调用。在元数据已经注册完毕的前提下,数据发布的主要工作是实况产品适配相应的 MUSIC 读取接口。实况产品分属地面资料和数值模式资料大类,按照常见的应用场景,可选择关联“按时间获取地面文件产品”和“按时间范围获取地面文件产品”接口,对于 GRB2 格式的数据,追加关联“按时间、预报要素获取格点分析场”的接口^[15]。用户申请使用数据时,可通过门户网站浏览已发布的相关资料元数据,并在用户空间申请资料访问权限,包括勾选资料服务代码、填写绑定 IP 地址、访问时间范围等信息,审核通过后可通过 MUSIC 接口获取数据^[16]。此外,可申请存储和算法权

限,在 DPL 中集成算法,通过连接专题库获取文件索引并下载文件实体。

5 结语

通过将省级多源融合实况产品作为本地化新增资料,按照规范流程融入“天擎”,实现了实况产品的集约管理和标准化服务。弥补了省级多源融合实况分析系统在产品共享端的短板,搭建形成“云+端”技术架构下省级实况业务“数据直传、产品直算、服务直通”的管理闭环。可作为典型案例,对其它各类数据产品融入和应用系统“云化改造”具有一定参考价值。

参考文献:

- [1] 师春香,潘旸,谷军霞,等. 多源气象数据融合合格点实况产品研制进展[J]. 气象学报,2019,77(4):774-783.
- [2] 戚云枫,曾小团,梁苑苑,等. 广西网格预报系统融入“天擎”的实践与思考[J]. 气象研究与应用,2022,43(2):111-116.
- [3] 周自江,曹丽娟,廖捷,等. 水文气象信息概述:观测、融合与再分析[J]. 气象,2022,48(3):272-283.
- [4] 国家气象信息中心. 气象大数据云平台设计方案[R]. 北京:国家气象中心,2018:56-57.
- [5] 孙超,肖文名,陈永涛,等. 气象综合业务实时监控系统的设计[J]. 气象科技进展,2018,8(1):153-157.
- [6] 中国国家标准化管理委员会,气象资料与分类编码:GB/T 40153—2021[S]. 北京:中国标准出版社,2021.
- [7] 贺俊彦,刘然,陈永涛,等. 基于元数据管理的气象数据广播精细化监控[J]. 计算机技术与发展,2023,33(1),88-94.
- [8] 刘然,贺俊彦,谭小华,等. 单机版国内气象通信系统 CTS1.0 省级备份方案设计与实现[J]. 气象科技,2019,47(6):885-892.
- [9] 姜娜娜,王凡,崔巍,等. 宁夏国内气象通信系统 CTS2.0 单机版建设与实现[J]. 气象水文海洋仪器,2022,39(2),102-104.
- [10] 赵芳,熊安元,张小缨,等. 全国综合气象信息共享平台架构设计技术特征[J]. 应用气象学报,2017,28(6),750-758.
- [11] 黄志,黄珩,梁维亮,等. 基于“天擎”DPL 的业务融入设计与应用初探[J]. 气象研究与应用,2022,43(1):73-77.
- [12] 余永城,王笑,魏夏潞. 福建气象综合业务平台融入“天擎”技术方案设计及实现[J]. 气象科技,2022,50(5):653-659.
- [13] 何林,吉庆,燕东渭,等. “秦云工程”气象数据对接的设计与实现[J]. 陕西气象,2020(1),47-50.
- [14] 曾行吉,李涛,詹利群,等. 基于 MUSIC 的特色数据与产品回写 CIMISS 方法研究[J]. 气象研究与应用,2018,39(1):111-114.
- [15] 郑波,李湘,何文春,等. 基于 CIMISS 全国精细化格点预报业务数据环境系统设计与实现[J]. 气象科技,2018,46(4),670-677.
- [16] 王娟娟,王华,邢丽平. 天擎湖北 Portal 网站及用户使用概述[J]. 数字技术与应用,2022,40(3),189-191.