

基于流程控制的国际气象通信数据收集管理系统的设计与实现

顾文静 王鹏 戴晴 韩鑫强

(国家气象信息中心, 北京 100081)

摘要 近年来,天气和气候业务对国外数据种类和时效要求不断提高,为解决大量数据的收集分发压力以及管理问题,设计并实现了基于流程控制的国际气象通信数据收集管理系统。系统应用数据收集器 Logstash、Inotify 监听机制与分析平台 Kibana 等技术,支持多种通信协议,扩展 Scrapy 网络爬虫功能,通过流程控制、任务调度等方式实现大量数据快速高效收集和集成管理,有效提高数据并行处理能力和可扩展性,且设计了全流程监视,保障数据时效。系统已经上线提供服务,为业务/科研用户提供全球观测数据、国外卫星数据和数值预报产品等 300 余种,日收集数据量约 2 TB。

关键词 国际气象通信;国外数据;收集处理;调度管理;Scrapy

中图分类号: P409 **DOI:** 10.19517/j.1671-6345.20210493 **文献标识码:** A

引言

随着天气和气候业务的发展,业务/科研对国内外气象数据和产品种类需求不断增加。中国气象局是世界气象组织全球通信系统的亚洲区域通信枢纽之一,国家气象信息中心国际通信系统负责全球实

时、近实时气象数据和产品的发现获取,为国家级和省级业务单位及相关行业用户提供全球观测数据、国外卫星数据和数值预报产品^[1]。国际通信系统包含全球通信系统(GTS)、Internet 平台和北京 GISC 门户系统,收集网络如图 1 所示。

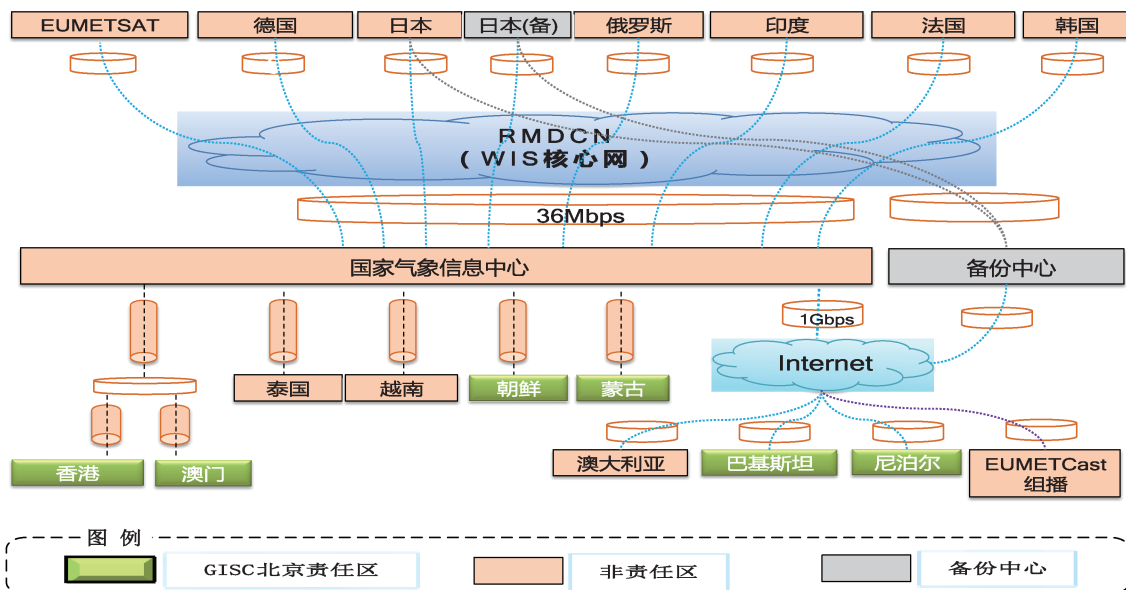


图 1 全球数据收集网络

国际通信 Internet 平台作为国际通信系统的核心组成,通过互联网下载、双边交换,组播线路(EU-METSAT)等多种方式进行数据收集,共获取数据 300 余种。根据数据的特征分别采用 AFD、LFTP、WGET 或者爬取的方式收集,考虑到数据下载频率、数据量,以及国外数据中心对下载流量限制等诸多因素,将收集程序部署到多台异构服务器上。当程序需要修改调整时,技术人员要针对不同的数据,登录不同的服务器操作;且由于缺乏统一的管理调度,庞杂的数据信息清单需要手动维护,收集任务的分配由人工评估,且缺乏集中监视,数据的收集和故障处理效率都非常低下。

为了解决上述问题,国家气象信息中心组织开发国际气象通信数据收集管理系统,针对大量数据的收集和分发压力,基于流程控制的自动化调度处理平台,实现国际数据收集的统一管理和集成,支持 HTTP、FTP、TCP、UDP、SMTP/POP3 等多种通信协议,扩展 Scrapy 网络爬虫框架,通过文件传输、消息通信等方式,按照相应的收集策略,实现快速、高效的数据收集,有效提升系统并行处理能力和可扩展性^[2-3]。并设计开发统一监视功能,有效提升故障发现和处理时效。

本文详细介绍了国际气象通信数据收集管理系统的设计实现及业务应用等,依次对系统功能架构、技术实现及业务应用情况进行详细论述。

1 系统设计

1.1 功能设计

国际气象通信数据收集管理系统包含任务调度、数据获取、数据预处理、数据缓存共享、监视信息采集处理和统计查询 6 个后台处理功能模块。如图 2 所示。

(1)任务调度。实现任务在不同的节点进行分布式执行,支持脚本和 JAVA 类库注册、动态加载等形态的集成与管理^[4-5]。任务调度通过 WEB 管理站点统一配置,进行严格的权限控制。

(2)数据获取。对现有收集工具和脚本进行改造,支持的 HTTP/HTTPS,FTP/SFTP,FISH/递归/镜像,REST 接口调用,Scrapy 网络爬虫等方式的脚本集成,实现数据收集统一管理和集成。通过 WEB 站点配置实现对收集数据存储路径、文件名、落地时间、数据量等基本信息的监听识别。

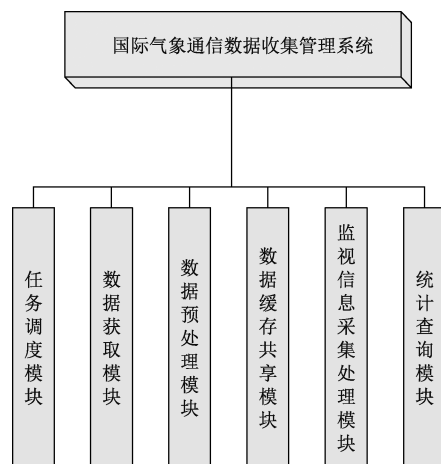


图 2 功能结构

(3)数据预处理。根据气象行业标准规范,对获取的气象数据进行一系列预处理操作,包括文件校验、内容解析识别、文件换名、文件拼接、解包解压缩、编码解码、格式转换等^[6]。

(4)数据缓存共享。按照相关策略实现对获取文件的存储清理和索引入库。通过 FTP 服务及数据接收端文件的规范传输接口^[7],实现对国内通信系统和国外中心数据产品的快速高效的共享分发。

(5)监视信息采集处理。按照综合业务监控的要求,实现数据收集处理、信息分发共享等业务功能,以及系统运行状态信息^[8]的记录并发送到气象综合业务实时监控系统。

(6)统计查询模块。实现数据收集、分发过程中多维的统计、查询功能,并能够以直观、界面化、图形化的形式展示,同时提供文件下载功能。

1.2 架构设计

为保证整个系统的稳定可靠,需要遵循合理的安全保障体系和标准规范与运行监管体系。系统的总体架构可以划分为 4 个层次和 4 个保障体系,如图 3 所示,4 个层次自底向上分为基础设施与服务层、数据服务层、技术支撑层、应用服务层和用户访问层,4 个保障体系分别为标准规范体系、运行维护体系、安全保障体系、技术支持体系^[9-10]。

基础设施层:依托于国际通信 Internet 平台为整个系统提供服务器、存储、网络以及支撑系统运转的操作系统、数据库软件、存储备份系统、安全软件等。

数据服务层:数据资源层是系统信息汇集和数

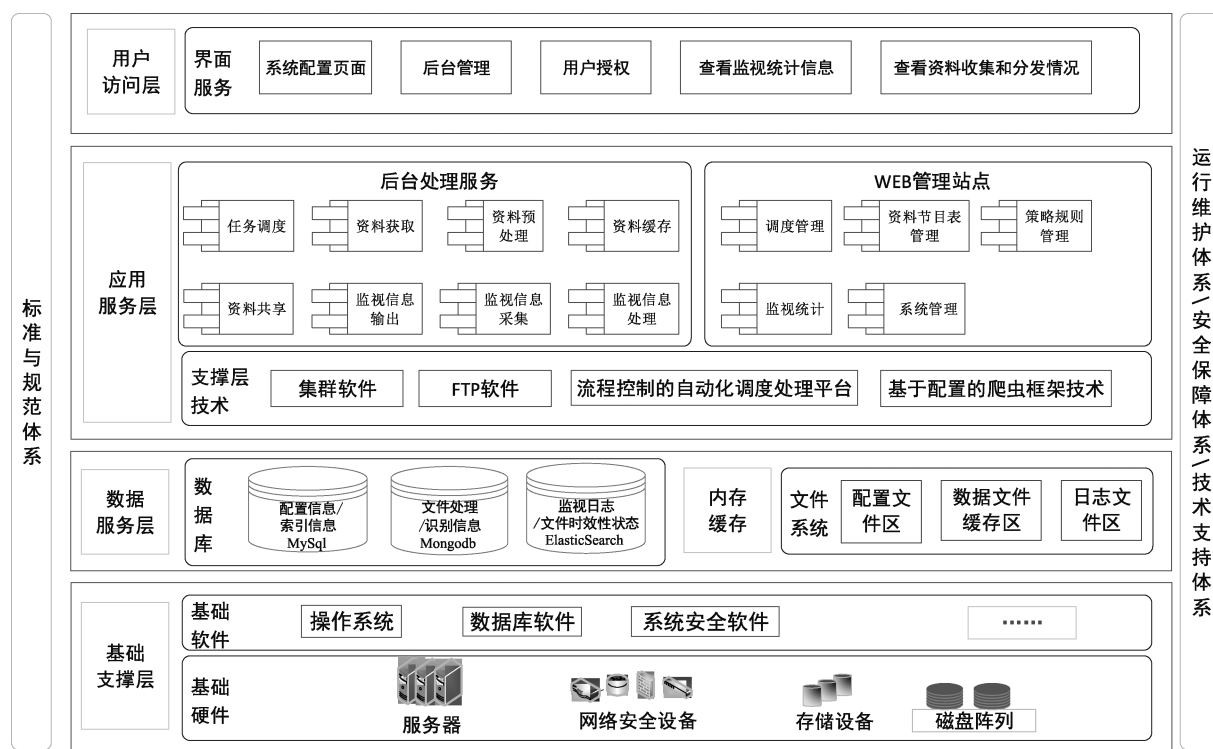


图 3 系统总体架构

据存储与管理的基础。包括文件系统、数据库和内存缓存这 3 种数据的存储方式。其中数据文件和日志文件通过文件系统进行存储；系统和业务监视信息、统计信息、业务策略和配置信息和用户信息、目录信息等通过数据库进行存储；对数据传递过程中的中间态文件以内存方式进行缓存^[11]。

基础支撑层：充分利用自有平台和开源技术对系统进行快速设计与开发，主要平台和技术包括集群软件、FTP 软件、流程控制的自动化调度处理平台、基于配置的爬虫框架技术等。

应用服务层：负责核心业务逻辑，包括任务调度、数据获取、数据预处理、数据缓存共享、监视信息采集处理和统计查询等后台处理服务功能，以及调度管理、数据节目表管理、策略规则管理、系统后台管理、监视信息展示等 WEB 管理站点功能。

用户访问层：主要指人机交互的相关功能，管理人员可通过 WEB 页面进行系统配置、后台管理、用户授权等操作，运维人员可查看监视统计信息，业务用户可查看数据收集、分发情况。

标准规范体系：是支撑系统运转的基础，用以实现应用协同和信息共享，节省项目建设成本、提高项目建设效率，满足系统不断扩充、持续改进和版本升

级的需要。

运行维护体系：为确保国省数据共享系统的协调性和连续性，构建以系统运行保障为基础的运行维护体系，建立了运维管理机制，明确以事件管理、问题管理、变更管理和配置管理等环节为基础的运行维护流程^[12]。

安全保障体系：是保障系统安全应用的基础，包括物理安全、网络安全、信息安全及安全管理等。

技术支持体系：技术支持体系的核心是人，通过科学的服务管理流程建立专业服务体系的标准化，针对国省数据共享系统新的数据需求和功能需求不断推陈出新，优化组合，保障系统的不断优化和完善。

2 系统实现

2.1 关键点解决方案

2.1.1 流程控制的自动化调度处理平台

根据业务需要，系统整合不同协议的多个独立软件，用自研的流程控制的自动化调度处理平台，通过集中调度和分布式执行方式，实现统一的配置管理和集成调度的交换服务微代理模式，实现国外数据并行高效地获取。

调度平台由调度中心和执行器组成,调度中心负责系统管理和任务调度功能,自身不承担业务代码,系统管理功能包括任务管理、执行器管理、日志管理等,任务调度功能从数据中心拉取任务,按照执行时间将新任务投递给执行器;执行器负责接收调度请求并执行任务逻辑,任务模块专注于任务的执行等操作。

调度处理平台以 XXL-JOB 调度器,分布式组件 Zookeeper, RabbitMQ 消息中间件,以及 Redis、MySQL 数据库等作为支撑,XXL-JOB 调度器在 quartz 的基础上修改了任务调度的模式,采用注册

和 Dubbo(RPC 远程服务调用)方式来实现任务调度;Zookeeper 实现任务的分布式协调,并通过 Redis 分布式锁,避免集群中的多个服务器同时调度任务;当任务调度时,发送消息到 RabbitMQ,执行器收到任务消息后,消费任务信息并执行任务^[13-14]。整体架构如图 4 所示。

调度系统与任务解耦,提高了系统可用性和稳定性;支持可视化、简单且动态的管理调度信息,包括任务新建,更新,删除,任务报警等,所有上述操作都会实时生效,同时支持监控调度结果以及执行日志^[15]。

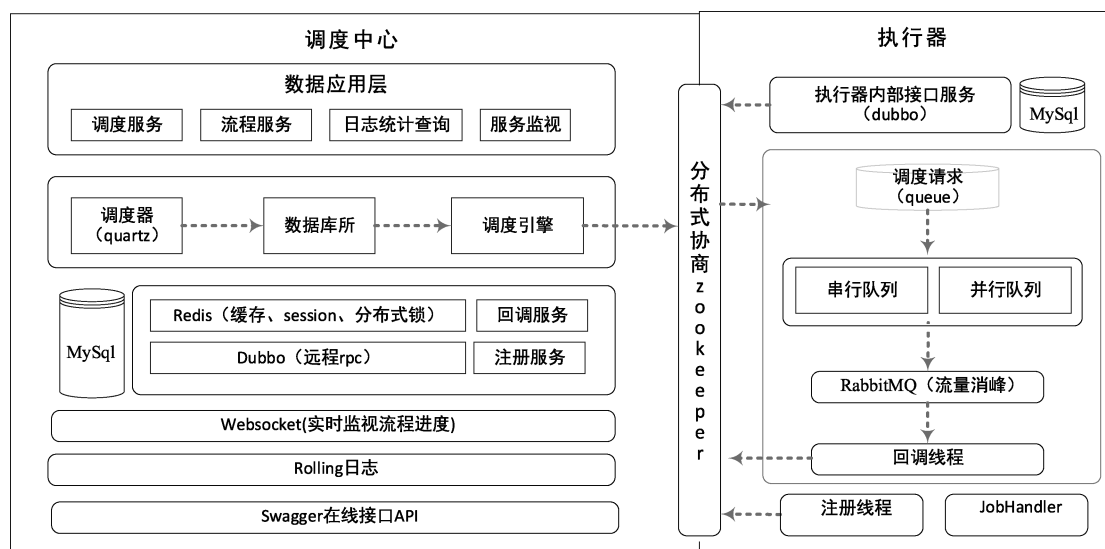


图 4 调度平台整体架构

2.1.2 基于配置的 Scrapy 爬虫框架技术

根据业务需求,从互联网上获取各类气象数据,面对网页数据爬取的复杂性,基于配置的 Scrapy 爬虫框架技术可以灵活配置获取互联网气象数据和产品,对 HTTP/HTTPS 协议有良好的支持,适应不同网页数据内容的灵活获取,并依托于流程控制的自动化调度处理平台实现动态加载服务、界面可配置、结果可展示、流程可监控。

2.2 技术实现

系统实现紧密结合现代气象业务发展和实际业务需求,统一软件技术框架,面向接口实现服务化开发。通过数据库与文件库的优化配置、动态调整作业优先级、动态调整执行作业数、作业负载均衡、硬件合理的冗余配置等措施提高系统的处理时效^[16-17]。

系统主要涉及到任务调度、数据获取、处理分发和监视四个方面技术点,如图 5 所示,数据获取模块封装 FTP/SFTP、HTTP/HTTPS、FISH/递归/镜像以及 Scrapy 等工具,实现快速、高效地文件、消息收集及爬取等数据获取。应用基于事件驱动机制的 Inotify 程序进行数据目录的监听,实现对事件处理的实时响应,且没有轮询造成的系统资源消耗。当被监听文件落地后,将文件路径、文件名、文件落地时间等信息写入 MongoDB 库。任务调度模块按照配置的路由策略和时间触发识别任务,任务调度模块采用 Redis 作为注册中心,可以动态发现和检查执行器的状态。任务调度模块将任务路由指定执行器后,启动识别任务,调用 dubbo 服务获取数据的识别配置(如文件路径、文件名规则等),使用 MongoDB 提供的正则匹配功能快速匹配文件,识别数

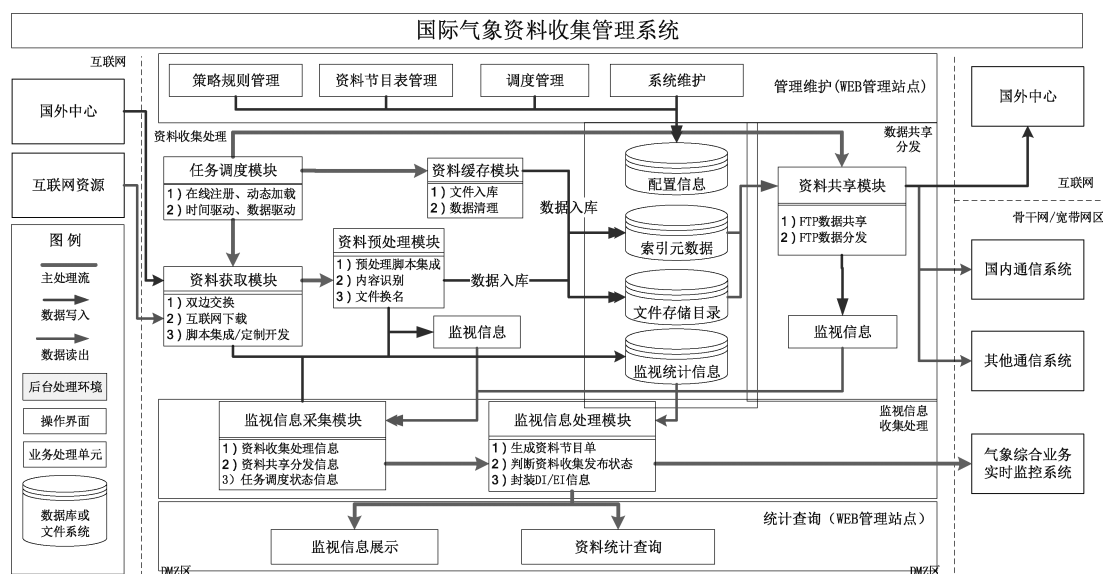


图 5 系统功能模块交互关系

据时次、预报时效等信息,并将处理状态和识别后的信息更新到 MongoDB 库,实体数据文件通过链接方式发布到用户目录下,实体数据文件仅存储一份,减少文件存储空间。

各个环节均根据规范要求输出详细日志,监视信息采集模块通过 Logstash 实现数据收集日志信息、FTP 下载服务日志信息等监视日志实时抽取并存储到 Elasticsearch 库表中。统计查询模块读取各类数据的监控配置,判断文件的时效性状态,将其更新到 Elasticsearch 逐日文件明细表中,按照数据进行汇总,更新逐日数据汇总信息到 MySQL 库表中。并封装成 DI 发送至气象综合业务实时监控系,异常情况则根据告警策略封装成 EI 发送气象综合业务实时监控系。

2.3 数据库结构

数据收集管理系统包含了数据收集、数据识别预处理和缓存、监视统计、共享分发以及调度框架、权限认证 5 部分基础信息表,共创建 29 张数据表。对数据识别预处理和缓存表以及监视统计相关表进行统计预处理,生成数据收集日志索引和节目单索引表;其中索引表的可统计属性均与基础信息表建立主外键关系^[18]。

3 业务应用

国际气象通信数据收集管理系统已实现美国、欧洲、日本等国家 300 余种观测数据、卫星高分和数

值高分数据的收集处理,接收 EUMETSAT 的 METOP、GOES、Sentinel 等卫星高分数据,获取数据量约 10 TB/d。其中日本 Himawari-8 卫星数据、全球预报系统 GFS、GEFS 产品、全球数据同化系统 GDAS 产品、欧洲中心大气模式产品、日本高分辨率数值预报产品(GSM)等备受用户关注的数已经通过共享分发提供国家局和省局天气/气候业务使用,提升预测、预报精准度^[19-20]。

3.1 任务调度

3.1.1 任务调度配置

对任务调度进行统一管理,根据业务需要注册新的脚本或类库进行可视化任务调度配置,并支持动态修改任务状态、暂停/恢复任务,以及终止运行中任务,即时生效。

通过【数据管理】的【关联程序】/【关联任务】分页面进行脚本注册和任务配置。脚本注册配置包括脚本名称、执行类标识、脚本类别、阻塞处理策略、脚本类型、脚本全路径以及数据分类等信息。新增任务配置包括任务调度策略、任务执行策略、任务种类、选择脚本、优先级和调度异常处理等信息(图 6)。

3.1.2 任务调度管理

【任务管理】的【调度日志】页面展示所有调度任务的名称、调度时间、调度结果、执行结果、执行状态及详细日志信息,并对近一个月的调度信息进行汇总形成运行报表,如图 7 所示。

脚本注册

1 基本参数
2 输入参数
3 部署节点

脚本名称: CFSv2收集

脚本类别: 资料收集脚本

脚本类型: sh

执行类标识: 脚本集成处理类

阻塞处理策略: 单机并行

脚本全路径: /nas6/ftp/nccep/etc/DID_20200709_

所属资料种类: 数值预报产品

美国NCEP再分析

请选择三级分类

请选择资料

上一步
下一步

保存
取消

新增任务

1 任务参数

任务名称: CFSv2收集任务

任务调度策略: 选择CPU空闲的节点执行

任务种类: 资料收集任务

优先级: 4

所属资料种类: 请选择一级分类

任务执行策略: 每10分钟执行一次

选择脚本: CFSv2收集

调度异常处理: 失败告警 + 报警邮件 + 多个邮件间隔

请选择二级分类

请选择三级分类

请选择资料

上一步
下一步

保存
取消

脚本管理

序号	脚本名称	执行类标识	脚本路径	操作
1	CFSv2收集	ordinaryHandler	/nas6/ftp/nccep/etc/DID_20200709_001.sh	编辑 删除 新增
2	mongodb备份	mongoBackupHandler		编辑 删除 新增
3	推送程序	pushHandler		编辑 删除 新增
4	数据清理程序	clearExpireDataHandler		编辑 删除 新增
5	数据同步程序	followHandler		编辑 删除 新增
6	节目单汇总程序	programmeStatScrollHandler		编辑 删除 新增
7	节目单更新程序	programmeScrollHandler		编辑 删除 新增
8	逐日节目单生成程序	logiProgrammeHandler		编辑 删除 新增
9	inotify事件配置	reloadInotifyConfigHandler		编辑 删除 新增

任务管理

任务名称	下次执行时间	任务执行策略	状态	操作
mongodb备份程序	2021-11-03 17:00:10	每小时执行一次	ONLINE	执行一次 暂停 删除 任务
推送任务	2021-04-21 15:48:00	每分钟执行一次	ONLINE	执行一次 暂停 删除 任务
数据清理任务	2021-11-04 02:00:10	每日18时执行一次	ONLINE	执行一次 暂停 删除 任务
数据同步任务	2021-11-03 16:30:00	每小时执行一次	ONLINE	执行一次 暂停 删除 任务
节目单汇总任务	2021-11-03 16:25:00	每分钟执行一次	ONLINE	执行一次 暂停 删除 任务
节目单更新任务	2021-11-03 16:30:00	每小时执行一次	ONLINE	执行一次 暂停 删除 任务
逐日节目单生成任务	2021-11-04 08:00:10	每日18时执行一次	ONLINE	执行一次 暂停 删除 任务
inotify事件配置 (配置新路径后手动触发)	2021-05-30 02:00:10	每日18时执行一次	ONLINE	执行一次 暂停 删除 任务

图 6 任务调度策略页面

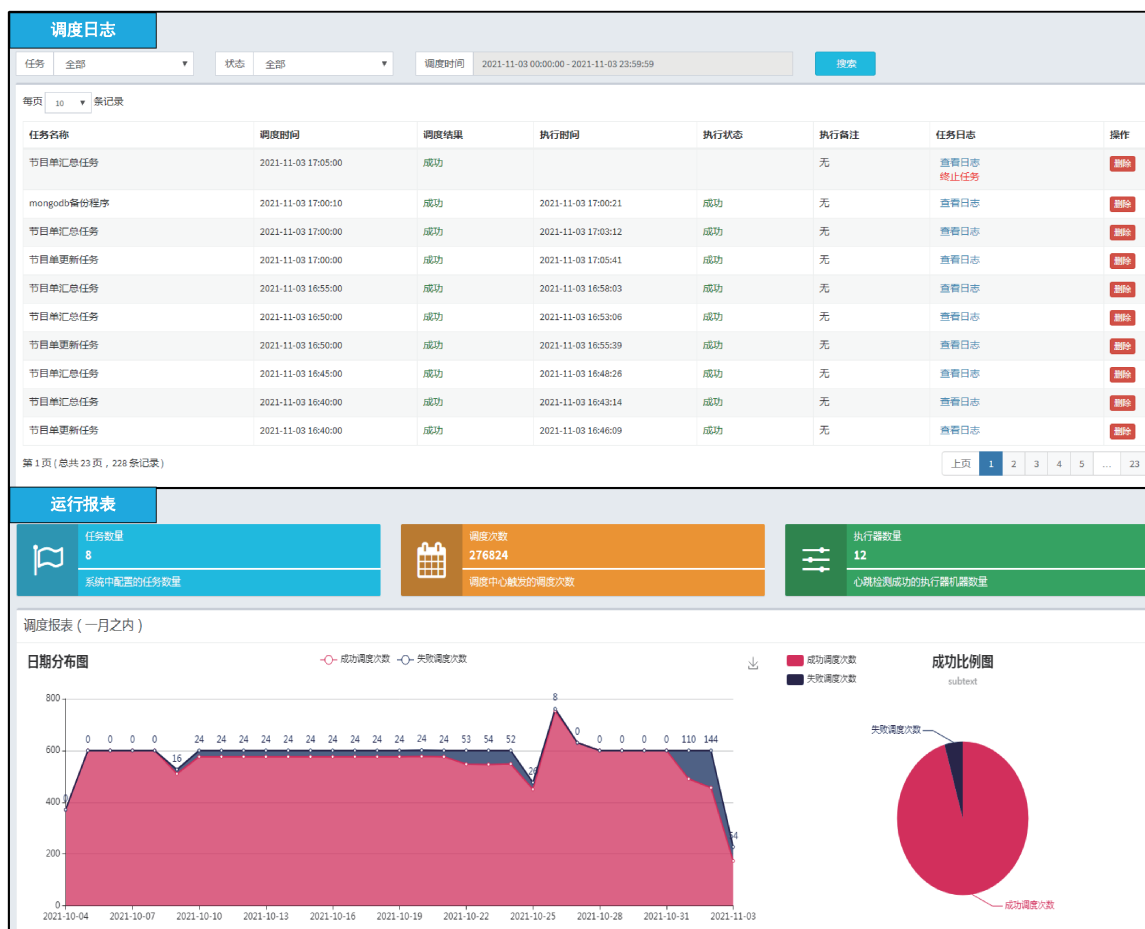


图 7 任务调度管理页面

3.2 监视统计

根据调度策略定时启动的节目单更新和汇总任务,统计各业务时次数据的收集情况并在【监视统计】页面按照业务时次粒度(年、月、日、时、分)展示,

根据节目表将数据文件划分为及时(已到文件数量)、不影响业务迟到、影响业务迟到和缺收 4 种状态,用不同的颜色表示,如图 8 所示。

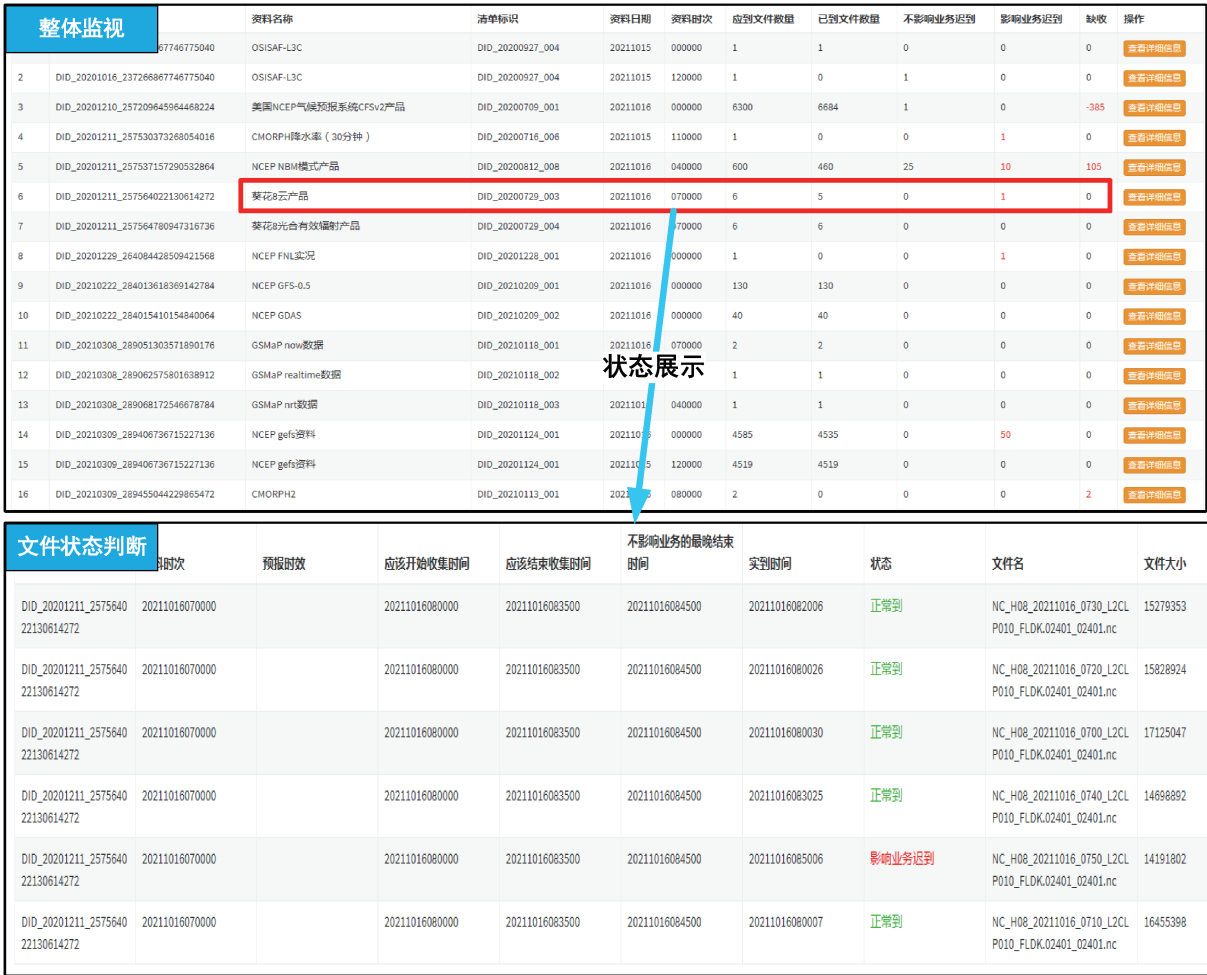


图 8 数据监视统计页面

3.3 时效统计

2021 年 10 月,针对用户重点关注的 10 种数据进行收集时效性统计,如表 1 所示,及时率均在 99%以上,根据业务需求,文件大小小于 50 MB 的文件收集时效小于 30 s,文件大小大于 50 MB 小于 1 GB 的收集时间小于 2 min,文件大小大于 1 GB 小于 5 G 的收集时间小于 10 min,如表 2 所示,10 种数据均满足业务要求。

4 结论

本文设计实现了国际气象通信数据收集管理系

统,在搜索引擎数据库 Elasticsearch、数据收集器 Logstash、可视化搜索与分析平台 Kibana 的基础上,扩充并完善了任务调度、数据获取、识别、解析、预处理和监视具体方法,并在业务应用中不断完善,先后进行了数据和调度配置功能改进、节目表页面展示改进、数据识别配置和监控配置改进,逐渐形成稳定、可靠的调度、管理、监视系统。系统的实际运行效果较好,配置灵活方便、易于使用,对流程控制的自动化调度处理平台设计和实现具有参考价值,对国际气象数据收集、管理具有重要意义。

表 1 重点数据到报情况统计(2021-10)

序号	文件名称	获取方式	月应收 总数	及时		迟到		缺收	
				个数	百分率	个数	百分率	个数	百分率
1	欧洲中心环境预报模式产品	FTP/SFTP	60016	60016	100.00	0	0.00	0	0.00
2	全球降水卫星制图数据集(GSMaP)临近数据(now)	FTP/SFTP	1488	1475	99.13	13	0.87	0	0.00
3	日本全球谱模式预报产品	FISH/递归/镜像	109523	109498	99.98	25	0.02	0	0.00
4	美国 GFS 全球预报系统产品($0.5^{\circ} \times 0.5^{\circ}$) 全球预报系统 GFS 产品($1.0^{\circ} \times 1.0^{\circ}$)	FISH/递归/镜像	32240	32238	99.99	2	0.01	0	0.00
5	Himawari-8 全圆盘数据(HSD 格式)	HTTP/HTTPS	704320	704318	100.00	2	0.00	0	0.00
6	欧洲中心大气模式集合预报产品(C3E)	ECMWF 推送	27900	27900	100.00	0	0.00	0	0.00
7	欧洲中心大气模式确定性预报高分辨率产品(D1D)	ECMWF 推送	12276	12276	100.00	0	0.00	0	0.00
8	MetOp-B/AVHRR 1B 数据 MetOp-A/AVHRR 1B 数据	EUMETSAT 推送	29760	29752	99.97	8	0.03	0	0.00
9	ERA5 地面再分析产品 ERA5 等压面再分析产品 ERA5 模式面再分析产品	网页数据获取及 解析	372	367	98.66	5	1.34	0	0.00
10	ECMWF 观测站点数据收集情况图	网页数据获取及 解析	2668	2668	100.00	0	0.00	0	0.00

表 2 单个文件传输时效(2021-10)

序号	文件名称	获取方式	单个文 件大小	平均 时效	百分率			
					<30s	<2min	<10min	>10min
1	欧洲中心环境预报模式产品	FTP/SFTP	1.2 MB	5.2 s	100	0.00	0.00	0.00
2	全球降水卫星制图数据集(GSMaP)临近数据(now)	FTP/SFTP	1.4 MB	6.4 s	100	0.00	0.00	0.00
3	日本全球谱模式预报产品	FISH/递归/镜像	4 MB	15.2 s	100	0.00	0.00	0.00
4	美国 GFS 全球预报系统产品($0.5^{\circ} \times 0.5^{\circ}$) 全球预报系统 GFS 产品($1.0^{\circ} \times 1.0^{\circ}$)	FISH/递归/镜像	152 MB 44 MB	52.4 s 25.5 s	0.00 100	100 0.00	0.00	0.00
5	Himawari-8 全圆盘数据(HSD 格式)	HTTP/HTTPS	3 MB	11.8 s	100	0.00	0.00	0.00
6	欧洲中心大气模式集合预报产品(C3E)	ECMWF 推送	479 MB	89.7 s	0.00	100	0.00	0.00
7	欧洲中心大气模式确定性预报高分辨率产品(D1D)	ECMWF 推送	1.2 GB	110.5 s	0.00	100	0.00	0.00
8	MetOp-B/AVHRR 1B 数据 MetOp-A/AVHRR 1B 数据	EUMETSAT 推送	10 MB	19.8 s	100	0.00	0.00	0.00
9	ERA5 地面再分析产品 ERA5 等压面再分析产品 ERA5 模式面再分析产品	网页数据获取 及解析	125 MB 1.1 GB 1.4 GB	56.3 s 7.8 min 10.8 min	0.00	0.00	96.70	3.30
10	ECMWF 观测站点数据收集情况图	网页数据获取 及解析	0.13 MB	3.6 s	100	0.00	0.00	0.00

参考文献

- [1] 王鹏. WIS 服务监视门户的设计与实现[J]. 气象科技, 2018, 46(3): 497-502.
- [2] 秦运龙, 王迎迎, 张冰松, 等. 省级外网气象大数据服务平台研究与实现[J]. 气象科技, 2020, 48(6): 823-829.
- [3] 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.
- [4] 王甫棣, 赵希鹏, 王帅. 基于 SOA 的任务调度框架设计与实现[J]. 气象科技, 2020, 48(3): 363-367.
- [5] 何佳. 公共气象服务数据传输软件设计与实现[J]. 气象科技, 2019, 47(3): 426-432.
- [6] 李兴宝, 侯方. 基于 C/S 模式的气象资料自动下载系统[J]. 气象科技, 2010, 38(5): 609-612.
- [7] 马渝勇, 徐晓莉, 宋智, 等. 省级气象信息共享系统的设计与实现[J]. 应用气象学报, 2011, 22(4): 502-512.
- [8] 杨立苑, 胡佳军, 邓卫华, 等. 省级区域站数据直传监控系统设计与实现[J]. 军事通信技术, 2019, 47(5): 773-779.
- [9] 付更丽, 曹宝香. SOA-SSH 分层架构的设计与应用[J]. 计算机技术与发展, 2010, 20(1): 74-77.
- [10] 冯新扬, 范颖, 崔凯, 等. 利用设计模式改进分层架构[J]. 计算机工程与设计, 2007, 28(15): 3686-3689.
- [11] 李兴宝, 侯方. 基于 C/S 模式的气象资料自动下载系统[J]. 气象科技, 2010, 38(5): 609-612.
- [12] 陈京华, 邓莉, 王舒, 等. 国家气象业务内网 WebGIS 数据服务系统设计与应用[J]. 气象科技, 2020, 48(4): 497-502.
- [13] ZooKeeper Documentation 3. 4. 10 [EB/OL]. [2020-11-26]. <https://zookeeper.apache.org/>.
- [14] xxl-job 1. 8. 0 Documentation [EB/OL]. [2017-7-9]. <http://www.xuxueli.com/xxl-job/en/#/>.
- [15] 田熙清, 党延忠. 基于工作流技术的项目管理系统的分析和设计[J]. 计算机工程与应用, 2013(8): 131-134.
- [16] 詹利群, 霍庆, 张志强, 等. 地面气象资料一体化统计加工系统设计与实现[J]. 气象科技, 2018, 46(3): 503-508.
- [17] 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.
- [18] 胡争光, 魏丽, 薛峰, 等. 世界气象中心(北京)全球预报服务共享平台设计与实现[J]. 气象科技, 2019, 47(4): 581-590.
- [19] 贺雅楠, 高嵩, 薛峰, 等. 基于 MIPCAPS4 的精细化天气预报平台设计及应用[J]. 气象科技, 2018, 46(1): 200-206.
- [20] 邱贵强, 武永利, 赵桂香, 等. 基于 MIPCAPS4 的灾害性天气历史个例库应用系统的设计与研发[J]. 气象科技, 2019, 47(4): 592-599.

Design and Implementation of International Meteorological Communication Data Collection and Management System

GU Wenjing WANG Peng DAI Qing HAN Xinqiang

(National Meteorological Information Center, Beijing 100081)

Abstract: In recent years, meteorological and climate services have increased their requirements for foreign data types and timeliness. An international meteorological communication data collection and management system based on process control is designed and implemented to solve the collection and distribution pressure and management problems of a large amount of data. The system applies the technology of Logstash, Inotify and Kibana, supports a variety of communication protocols, expands the network crawler function of Scrapy, and achieves the fast and efficient collection and integration management of a large amount of data through process control and task scheduling, effectively improving the data parallel processing ability and scalability. And the whole process of monitoring is designed to ensure data timeliness. The system has been online to provide services, providing operational/scientific research users with more than 300 kinds of global observation data, foreign satellite data and numerical prediction products; the daily data collection is about 2 TB.

Keywords: international meteorological communication; foreign material; collection and processing; dispatching management; scrapy