引文格式：高微,杨婷,杨溯,等. 轻量级地理数据分布式处理框架关键技术研究[J]. 地理空间信息, 2024, 22(1): 66-68. doi:10. 3969/j.issn. 1672-4623.2024.01.015

轻量级地理数据分布式处理框架关键技术研究

高　微[1]，杨　婷[1*]，杨　溯[1]，吴浩伟[1]
（1. 自然资源部第三大地测量队，四川 成都 610100）

摘　要：按需定制的仅支持单机运行的GIS应用程序，在面对海量数据处理时存在耗时长、效率低等问题。为提升个性化GIS应用程序处理海量地理数据的效率，通过研究分布式任务调度技术和轻量级虚拟化技术，搭建了基于Docker的轻量级地理数据分布式处理框架。该框架以接口的方式支持个性化GIS应用程序进行地理数据分布式处理，轻量灵活、可跨平台。以影像数据裁切为例，开展了对比测试与分析。结果表明，相较于传统单机作业，该框架处理效率有所提升。

关键词：分布式；Docker；轻量级；地理数据处理
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2024）01-0066-03

Research on Key Technologies of Lightweight Geographical Data Distributed Processing Framework

GAO Wei[1], YANG Ting[1], YANG Su[1], WU Haowei[1]
(1. The 3rd Geodetic Team of the Ministry of Natural Resources, Chengdu 610100, China)

Abstract: The on-demand GIS applications which only support single-machine operation have problems such as time-consuming and low efficiency processing in the face of massive data processing. In order to improve the efficiency of personalized GIS application in processing massive geographical data, we built a lightweight geographical data distributed processing framework based on Docker by studying distributed task scheduling technology and lightweight virtualization technology. This framework supports personalized GIS application for geographical data distributed processing by means of an interface, which is lightweight, flexible and cross-platform. Taking clipping image data as a example, we conducted comparative test and analysis. The result shows that this framework is more efficient than traditional single-machine operations.

Key words: distributed, Docker, lightweight, geographical data processing

近年来，地理空间数据的数据量呈现爆炸式增长[1]，数据密集、计算密集、并发密集和时空密集已成为GIS软件技术与应用面临的挑战[2]。密集的数据尤其是遥感影像数据[3]，需耗费大量的人力物力和时间进行处理；且在实际测绘生产中，自主研发的针对数据处理的GIS应用程序往往仅支持单机执行，时间消耗长、处理效率低，局域网内的计算资源也得不到充分利用。为高效解决测绘生产中个性化GIS应用程序地理数据处理效率低的问题，本文充分利用局域网内的计算资源，通过融合Docker容器技术[4-5]与分布式任务调度技术，搭建了基于Docker的轻量级地理数据分布式处理框架，以接口的方式为地理数据处理用户提供分布式处理能力，轻量灵活、可跨平台。通过实际测试发现，该框架能有效提高作业机器资源利用率，提升地理数据处理效率。

1　关键技术

1.1　分布式任务调度技术

分布式任务调度技术可实现在集群环境下多节点任务的灵活分配与处理，是数据分布式处理的核心，包括资源调度、任务分割、任务分配、状态探测、负载均衡等内容。本文搭建的是一个基于.NET Core的地理数据分布式处理框架，支持将作业分割后分发到多台服务器并行处理，可成倍提升海量数据的处理速度。地理数据分布式处理框架作业处理程序包括分割任务、执行任务、合并任务3部分，包含节点和管理界面两个核心程序，其中节点程序是控制台程序；管理界面程序是网页端程序，负责管理操作作业，包括查看集群节点、上传作业包、监控作业运行、下载作业结果等。

收稿日期：2021-12-20。
第一作者简介：高微（1990—），硕士研究生，工程师，主要从事GIS研发工作，E-mail：gaoweisc@qq.com。
通信作者：杨婷（1990—），硕士研究生，工程师，主要从事GIS研发工作，E-mail：626101874@qq.com。

1.2　轻量级虚拟化技术

针对常规分布式地理数据处理框架与专业软件集成度高、跨平台处理能力弱、普适性弱等问题，本文在充分分析地理数据处理流程的基础上，提出了分割、分发、处理、汇聚的任务处理基本流程，并研究了一种轻量级虚拟化技术——Docker容器技术[6]，可完成虚拟机能完成的所有功能[7]。由于在创建时Docker容器不会虚拟独立的系统内核，而是共用主机的系统内核，因此为应用节约了虚拟环境系统部署的资源，具有轻量、资源利用率高、启动时间短、迁移速度快、维护成本低和扩展性优良等优点[8-10]。同时，Docker容器在任何操作系统上都是一致的，将应用程序运行在Docker容器上即可实现跨平台、跨服务器的操作[11]。基于Docker容器的特性，本文将GIS应用程序运行在各轻量且相互隔离的Docker容器上，以实现分布式计算与资源的高效利用。

2　地理数据分布式处理框架

2.1　基本原理

本文构建的地理数据分布式处理框架的基本原理为将地理数据处理作业在逻辑上划分为若干个分片，并通过作业分片调度器将作业分片指派给特定的执行节点；执行节点通过任务调度状态触发执行作业，执行时将分片序号和参数作为参数传入（图1）。作业的实施逻辑需分析分片序号和分片参数，并以此为依据调用具体的实现，如一个批量处理影像作业，可划分0号分片处理1~10号影像数据，1号分片处理11~20号影像数据。

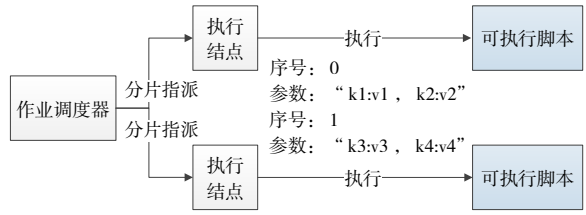


图1　框架基本原理

2.2　分布式任务调度架构

地理数据分布式处理框架包括调度中心和执行器两大部分，调度中心是一个网页端的管理系统，是整个调度系统的大脑，主要负责管理调度信息、按照调度配置发出调度请求、将作业任务分配到各执行器，支持可视化、简单且动态的管理调度信息，包括任务管理、执行器管理、日志管理等功能；执行器则负责接收调度请求并执行任务。本文构建的分布式任务调度总体架构见图2。

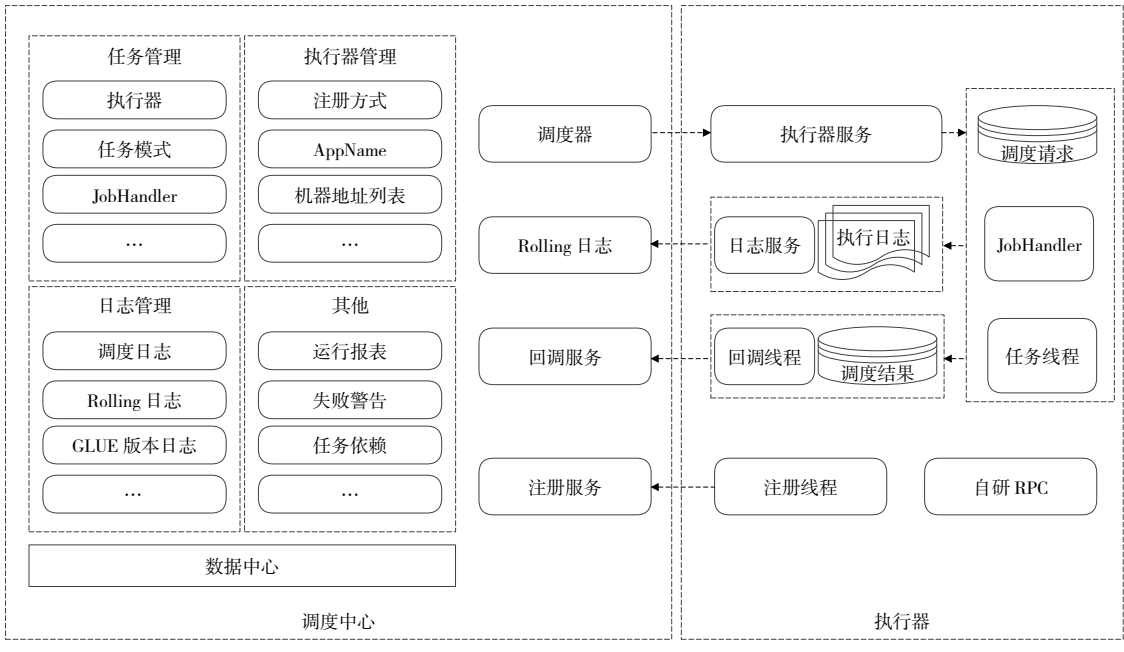


图2　分布式任务调度总体架构

基于Docker的地理数据分布式处理平台调度中心与任务解耦，自身不执行任务，而是将任务以作业包的形式分发至各执行器，性能不再受限于任务模块，可提高系统的可用性和稳定性。平台调度中心与各执行器之间的对应关系见图3。

2.3　总体技术流程

地理数据分布式处理框架为GIS应用程序提供了标准的数据接口与处理接口，用户根据自身业务需求，利用接口实现地理数据处理应用，并结合管理应用程序实现地理数据分布式处理。总体技术流程为：

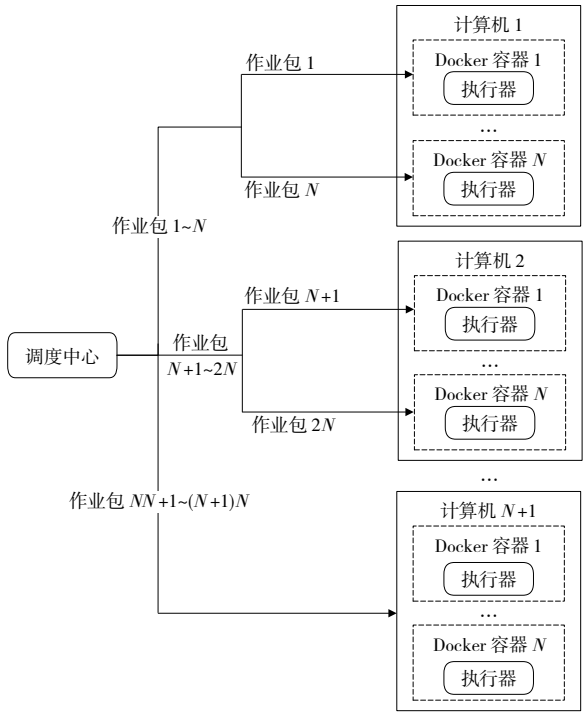


图3 调度中心与各执行器的关系图

①规范程序，GIS应用程序需严格按照地理数据分布式处理框架的接口规范编写，主要包括分割任务、执行任务、合并任务3部分，分别按照框架规定的方法名称和输入输出类型，实现应用程序内部逻辑；②部署环境，在Docker宿主机器上部署GIS应用程序所需的运行环境，部署成功后启动执行器监测任务；③分发任务，将GIS应用程序编译后进行打包，以作业包的形式分发至Docker集群进行处理；④结果下载，将集群各节点处理情况汇聚成用户的最终结果，用户按需进行资源下载。

3 案例分析

以栅格数据裁切为例，本文对比测试了地理数据分布式处理框架作业模式和传统单机作业模式的效率。案例采用6组栅格影像，数据量约为50 GB，裁切为6组规整的图斑，分3组测试，每组测试6次。测试期间测试电脑除进行栅格数据裁切外不做任何其他操作，测试组环境、执行时间等信息见表1所示，可以看出，相较于单机作业，该框架能减少数据处理时间，提升地理信息数据处理效率。

表1 测试环境、执行时间信息统计

测试组	测试环境	平均执行时间	备注
测试组1	计算节点1：CPU为Intel Core i7 2.6 GHz，内存为16 GB 计算节点2：CPU为Intel Core i5 2.3 GHz，内存为8 GB	3 min 55 s	地理数据分布式处理框架
测试组2	CPU为Intel Core i7 2.6 GHz，内存为16 GB	5 min 51 s	单机
测试组3	CPU为Intel Core i5 2.3 GHz，内存为8 GB	7 min 5 s	单机

4 结 语

基于Docker的轻量级地理数据分布式处理框架，以接口的方式为地理数据处理用户提供分布式处理能力，用户不再考虑分布式地理程序的设计与实现，只需按需定制数据处理算法，即可快速实现一个地理数据分布式处理应用，从而充分利用闲置的计算机资源，并发式进行地理数据自动化处理，有效提高数据处理效率。在实际测绘生产中，该框架为提升个性化GIS应用程序处理大数据量地理数据的效率提供了可行的思路和方法。

参考文献

[1] 李德仁. 展望大数据时代的地球空间信息学[J]. 测绘学报，2016,45(4):379-384

[2] Yang C, Goodchild M, Huang Q, et al. Spatial Cloud Computing: How Can the Geospatial Sciences Use and Help Shape Cloud Computing[J]. International Journal of Digital Earth,2011,4(4):305-329

[3] 陆冬华,赵英俊,张东辉,等. 单机多核计算机环境下的遥感图像并行处理技术[J]. 地理空间信息,2018,16(11):55-59

[4] NigelP. 深入浅出Docker[M]. 李瑞丰,刘康,译. 北京:人民邮电出版社,2019

[5] 李在弘. Docker基础与实战[M]. 武传海,译. 北京:人民邮电出版社,2016

[6] 郑军,聂榕,王守信,等. 基于Docker容器故障恢复的属性权重快照选择策略[J]. 信息网络安全,2021,21(5):12-18

[7] 漆雪西,陈长瑜,李维. 基于Docker的ctf赛题设计[J]. 网络安全技术与应用,2021(3):15-16

[8] Potdar A M, Narayan D G, Kengond S, et al. Performance Evaluation of Docker Container and Virtual Machine[J]. Procedia Computer Science,2020,171(4):1 419-1 428

[9] 朱波,童秋英,郑楚,等. 基于微服务的国土资源执法监察信息平台研究[J]. 地理空间信息,2019,17(4):113-117

[10] 董博,王雪,索菲,等. 基于Docker的虚拟化技术研究[J]. 辽宁大学学报(自然科学版),2016,43(4):327-330

[11] 张浩,孙毓忠,肖立,等. RainbowD:一种异构云环境下高效的Docker镜像分发系统[J]. 计算机学报,2020,43(11):2 067-2 083