计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):112−117 [doi: 10.15888/j.cnki.csa.006534]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Apache Spark 的 MODIS 海表温度反演方法①

刘　欢, 陈能成, 陈泽强

(武汉大学 测绘遥感信息工程国家重点实验室, 武汉 430079)
通讯作者: 陈能成, E-mail: cnc@whu.edu.cn

摘　要: 为应对海量遥感影像快速计算的需求, 通过对影像获取、算法和计算过程优化和改进, 提出了一种基于 Apache Spark 并行计算框架的 MODIS 海表温度反演方法, 实现了海量 MODIS 遥感影像的海表温度快速反演. 应用四轮网络查询请求获取特定的时空范围影像数据, 提高影像获取阶段的效率; 应用简化算法参数、拟合过程变量改进海表温度劈窗算法, 使之适合快速并行计算; 应用弹性分布式数据集 (RDD) 窄依赖关系的优点, 避免并行计算中的数据交换延迟. 通过单机模式与集群模式对比实验, 发现集成了并行计算框架的集群模式影像处理效率约为单机模式的 10 倍. 研究结果表明了融合集群计算技术的海表温度反演过程有效提高了传统单机应用程序的处理效率.

关键词: Apache Spark; MODIS; 海表温度

引用格式: 刘欢,陈能成,陈泽强.基于 Apache Spark 的 MODIS 海表温度反演方法.计算机系统应用,2018,27(9):112–117. http://www.c-s-a.org.cn/1003-3254/6534.html

Retrieving Method for MODIS Sea Surface Temperature with Apache Spark

LIU Huan, CHEN Neng-Cheng, CHEN Ze-Qiang

(State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China)

Abstract: In response to computing problems of massive remote sensing images, a method based on Apache Spark is proposed and implemented in retrieving MODIS Sea Surface Temperature (SST) by optimizing and improving the image acquisition, algorithm, and computing process. It applied four bouts of network requests to acquire user-defined data of specific time and zones to improve the efficiency of image acquisition. For a parallelizable algorithm, improvements that reduce parameters and simplify intermediate models are added to the split window algorithm, thus to adapt to fast parallelized computing. Taking advantage of narrow dependence between Resilient Distributed Datasets (RDD), delays for partitions' interactions are evaded. With comparison between single mode and cluster mode, the latter incorporated with Apache Spark has an efficiency of ten times to the former. This study proves that, comparing with a single machine's, programs that retrieving MODIS SST with cluster computing techniques has a higher efficiency.

Key words: Apache Spark; MODIS; Sea Surface Temperature (SST)

海水表面温度是一种重要的海洋环境参数, 被广泛应用于海洋动力过程、海气相互作用和全球气候变化等研究领域中. 但就传统原位测量方式而言, 其设备成本代价高、站点分布离散、监测范围有限, 无法满足较大区域范围、较长时间序列和连续性观测的需求. 针对以上不足, McMillin[1]最早提出使用遥感影像劈窗算法反演地表温度. 随着遥感技术和卫星反演算法的成熟, 覃志豪等[2]提出采用 MODIS 遥感影像反演海表

① 基金项目: 国家自然科学基金 (41771422); 湖北省自然科学基金 (2017CFB616)
Foundation item: National Natural Science Foundation of China (41771422); Natural Science Foundation of Hubei Province (2017CFB616)
收稿时间: 2018-01-29; 修改时间: 2018-02-27; 采用时间: 2018-03-02; csa 在线出版时间: 2018-08-16

温度产品, 并应用于相关大尺度、长周期的研究中. 同时伴随遥感数据的海量增长和应用算法复杂度的急剧增加, 传统计算方式难以满足日常的影像计算需求. 朱义明、付天新等[3,4]使用 MapReduce 编程模型对遥感影像实现了并行处理. Almeer、常生鹏等[5,6]使用 Hadoop 实现了图像处理中的滤波、锐化、增强和分割等算法. 刘震[7]提出了一种基于五元组的遥感图像并行处理方法. 以上遥感数据并行计算研究均采用离线存储的方式, 数据下载、人工预处理程度高; 算法设计的参数变量较单一, 对于中间输入参数较多的劈窗算法并没有提供相应的解决方案参考; 没有充分利用 RDD 窄依赖特性提高计算的效率. 本文综合相关研究和算法特性, 提出了一种基于 Apache Spark 的 MODIS 海表温度快速反演方法, 融合在线获取、实时解析、弹性分布式数据集处理和适于快速计算的反演算法提高海量遥感数据处理效率.

1 方法

本文研究涉及的主要方法有 MODIS 影像在线获取与动态解析、适于快速计算的改进劈窗算法和基于 Apache Spark 的遥感影像计算过程优化. MODIS 影像在线获取与动态解析使用自动化的影像查询机制提高影像定制的效率; 适于快速计算的改进劈窗算法通过减少输入参数, 降低模型复杂度的形式提高算法的时空效率; 基于 Apache Spark 的遥感影像计算过程优化通过 RDD 窄依赖关系减少计算过程中的网络数据交换延迟.

1.1 MODIS 影像在线获取与动态解析

针对人工影像数据获取的低效性, 本文采用自动化影像条件查询与多线程下载程序实现海量数据自定义批量下载, 避免了中间过程的人工干预, 提高了影像获取的效率. 美国宇航局以 RESTful API 方式提供了查询 MODIS 影像元数据和服务器存储路径的网络请求方法. 如图 1, 用户可以通过不同的网页请求参数获取不同的影像元数据信息. 如通过 searchForFiles 页面传递产品名称、数据集编号、时空范围和影像状态信息等参数查询研究区影像 ID 列表; 通过 getFileProperties 页面传递影像 ID 列表查询具体影像的状态信息, 并据此下载影像或提交订单; 通过 getFileUrls 页面获取影像的服务器路径, 并据此下载相应的影像数据文件. 本文为了提高处理效率, 避免磁盘 IO, 在数据处理前将影像文件以字节数组的形式缓存在内存中.

针对影像数据提取与解析过程, 本文采用一种适于并行化处理的第三方处理方案. MODIS L1B 数据文件存储格式是一种自描述多对象超文本文件格式——HDF4. HDF4 通过数据描述符和数据元素来管理不同的数据对象. 数据描述符包含了标识符、参照数、数据偏移长度和数据长度. 通过相应的数据描述符可以将文件中不同对象的数据提取出来, 达到解析的目的. Spark 兼容所有 Hadoop 支持的文件格式, 但是对于如 HDF4 这样的遥感数据并不提供原生的支持, 需要通过自定义的方式进行解析. 目前常用的第三方解析方案有美国国家超级计算机应用中心基于 C 语言的解析方案和美国大气研究中心基于 Java 的解析方案. 本文考虑到 Spark 运行于 Java 虚拟机支持的开发环境和平台可移植性的需求, 采用基于 Java 的解决方案. 通过自定义的解析程序, 可以将 Spark 不识别的 HDF4 文件数据, 提取转化为其支持并行化的操作数据.

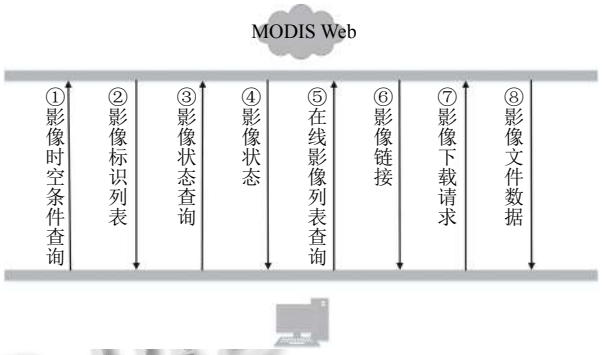

图 1 影像获取流程

1.2 适于快速计算的改进劈窗算法

劈窗算法, 是一种利用相邻两个热红外通道对大气水汽吸收作用的差异特性, 通过波段组合消除大气影响, 最终将海表温度表达为波段亮度温度函数的方法. 本文利用影像元数据、影像值和大气参数拟合公式, 减少反演过程中参数输入和拟合的效率损耗, 避免计算节点间数据交换. 其具体算法流程如图 2.

1.2.1 辐射定标

MODIS L1B 数据文件中保存的波段像元值为 16 位量化级数. 根据 MODIS 用户手册[8]所述, 反演计算中所需的辐射亮度值需要通过以下公式进行计算:

$$Radiance = Scale \times (DN - Offsets) \quad (1)$$

其中, $Radiance$ 为辐射校正后的波段辐射亮度,

Scale 和 *Offsets* 为 MODIS 元数据提供的定标参数, *DN* 为波段量化级数.

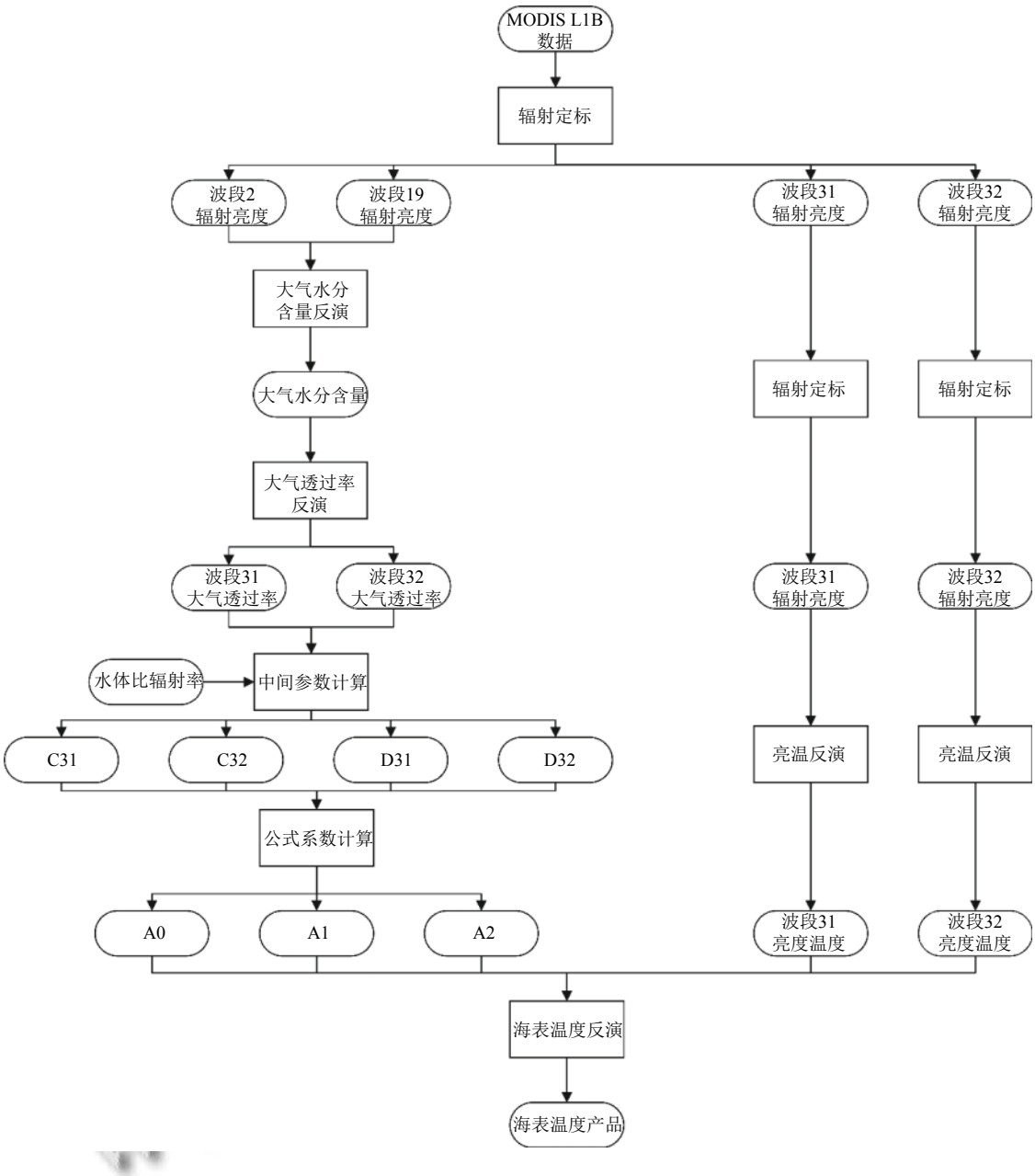

图 2 劈窗算法处理流程

1.2.2 大气水分含量

大气透过率通常是根据与大气水分含量的关系计算而得. Kaufman 和 Gao[9]提出利用 MODIS 相应光谱范围的波段函数组合获取大气水分含量的方法, 相较地面观测数据, 其误差在 5%~10% 之间.

$$\omega = \left[\frac{\alpha - \ln\left(\frac{\rho_{19}}{\rho_2}\right)}{\beta}\right]^2 \tag{2}$$

其中, ω为大气水分含量 ($g\cdot cm^{-2}$), α=0.02, β = 0.632, ρ_i为波段i的地面反射率.

1.2.3 大气透过率

大气透过率与大气水分含量的关系一般通过大气传输模型模拟确定. 毛克彪等[10]针对大气传输模型参数难以实时获取的问题, 利用神经网络拟合 MODTRAN 模型, 获得了了以下统计回归表达式, 其波段反射率平均误差不大于 0.008.

$$\tau_{31} = 2.9 - 1.88 \times e^{\frac{\omega}{21.23}} \quad (3)$$

$$\tau_{32} = 4.6 \times e^{-\frac{\omega}{32.71}} - 3.59 \quad (4)$$

其中, τ_i为i波段的大气透过率, ω为上一步所求大气水分含量. 考虑到反演计算过程中外部模型参数计算与输入会引起节点间数据交换的效率损耗, 本文采用以上统计回归表达式计算大气透过率.

1.2.4 线性组合系数[11]

地表比辐射率主要取决于地表的物质组成, 就海表温度反演而言, 其地表类型单一, 因此直接取相应的水体比辐射率.

$$\varepsilon_{31} = 0.996\,83, \varepsilon_{32} = 0.992\,324 \quad (5)$$

$$C_i = \varepsilon_i \tau_i(\theta) \quad (6)$$

$$D_i = [1 - \tau_i(\theta)][1 + (1 - \varepsilon_i)\tau_i(\theta)] \quad (7)$$

其中, i为 31 或 32 波段, C_i、D_i为计算线性组合系数$A_j(j = 1,2,3)$的中间参数, ε_i是波段i的地表比辐射率, $\tau_i(\theta)$为波段i视角为θ的大气透过率.

$$A_0 = \left[\frac{D_{32}(1 - C_{31} - D_{31})}{D_{32}C_{31} - D_{31}C_{32}}\right] a_{31} - \left[\frac{D_{31}(1 - C_{32} - D_{32})}{D_{32}C_{31} - D_{31}C_{32}}\right] a_{32}$$

$$A_1 = 1 + \frac{D_{31}}{D_{32}C_{31} - D_{31}C_{32}} + \left[\frac{D_{32}(1 - C_{31} - D_{31})}{D_{32}C_{31} - D_{31}C_{32}}\right] b_{31}$$

$$A_2 = \frac{D_{31}}{D_{32}C_{31} - D_{31}C_{32}} + \left[\frac{D_{31}(1 - C_{32} - D_{32})}{D_{32}C_{31} - D_{31}C_{32}}\right] b_{32}$$

其中, a_{31}= –64.603 63 , a_{32}= –68.725 75, b_{31}= 0.440 817, b_{32}= 0.473 453.

1.2.5 亮度温度

亮度温度是与物体在相同频率下具有相同辐射率的黑体的绝对温度, 是海表温度反演的重要参数之一. 利用普朗克黑体辐射方程可以直接求解相应波段的亮度温度.

$$T_i = \frac{k_{i,2}}{\ln\left(1 + \frac{k_{i,1}}{B_i}\right)} \quad (8)$$

其中, T_i为波段的亮度温度, $k_{31,1}$ = 729.541 636 $W \cdot m^{-2} \cdot sr^{-1} \cdot \mu m^{-1}$, $k_{32,1}$ = 474.684 7799 $W \cdot m^{-2} \cdot sr^{-1} \cdot \mu m^{-1}$, $k_{31,2}$ = 1304.413 871 K, $k_{32,2}$ = 1196.978 785 K

1.2.6 海表温度反演

覃志豪等[2]综合研究国内外 17 种劈窗算法, 通过比较其参数复杂性和算法精度, 在保证较高反演精度的情况下提出了线性反演公式. 其具有良好的反演精度, 已被应用于农业旱灾监测. 本文基于简化输入参数, 减少像元间计算引起的节点数据交换的需要, 采用以下线性反演公式进行海表温度的快速反演.

$$T_s = A_0 + A_1 \times T_{31} - A_2 \times T_{32} - 273.15 \quad (9)$$

其中, T_s是海表温度 (℃), T_{31}和 T_{32}分别是 MODIS 第 31 和 32 波段的亮度温度, A_0、A_1和A_2是劈窗算法参数.

1.3 基于 Apache Spark 的遥感影像计算过程优化

针对基于 Spark 的海表温度反演, 本文综合考虑了算法时间复杂度和空间复杂度, 采用如图 3 所示计算流程.

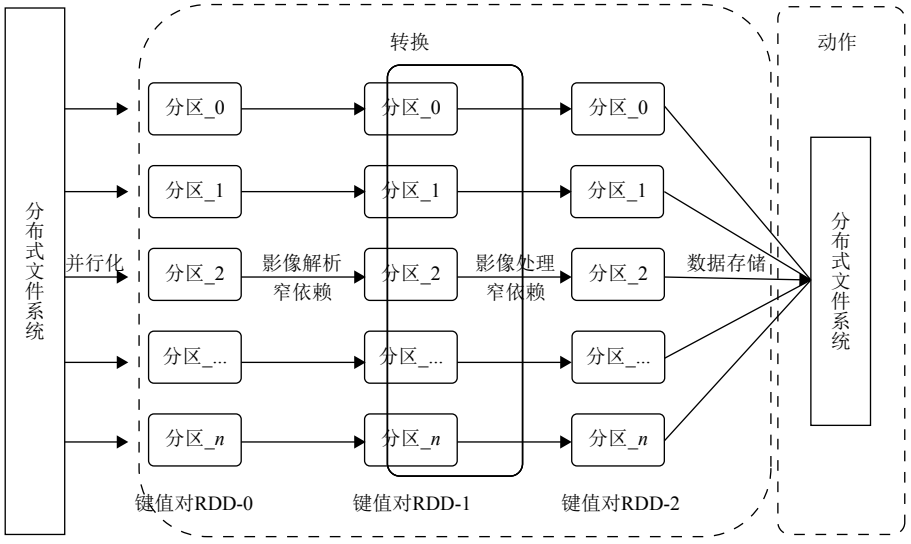

图 3 基于 Spark 的计算流程

网络下载的影像文件以独立影像的形式保存在分布式文件系统中 (本文采用的是 HDFS). 为减少影像处理时节点间的数据传输, 数据块的大小一般按照影像文件的大小分块. 如数据容量最大为 169 MB, 分块大小应至少设置为 169 MB 才能保证不会因文件分块导致跨节点访问. Spark 作业开始时, 影像文件以绝对路径集合的形式传入, 经并行化操作构建包含有 n 个分区的键值对弹性分布式数据集 (以下简称 RDD). 其中, 键名为每个文件的绝对路径字符串, 键值为相应的影像文件. 利用第三方解析方案提取出每个影像文件中的波段数据. 针对每个影像的波段数据, 按照算法设计, 结合 Spark 提供的转换算子和动作算子按照劈窗算法执行相同的波段组合函数, 进行海表温度的反演. 为了避免数据节点间的数据块传输和 RDD 间的组合计算, 从创建键值对 RDD-0 到生成键值对 RDD-2 的过程中, 通过对影像文件分块优化、算法过程优化和计算过程优化实现 RDD 间的窄依赖关系. 最终的反演结果保存到分布式文件系统中.

2 实验

2.1 实验数据

本文实验数据采用 MODIS L1B Terra 星载数据 (https://ladsweb.modaps.eosdis.nasa.gov/search/), 数据空间范围涵盖山东半岛、渤海湾一带, 时间范围为 2015 年 1 月 1 日至 2015 年 12 月 25 日, 共计 6634 景, 总计 1000 GB. 由于本文采用的劈窗算法需要可见光波段参与计算, 因此实验数据均为白昼期间获取的影像.

2.2 硬件环境

整个集群构建在相对封闭的千兆局域网中. 集群采用主从式架构, 使用商用台式机作为节点硬件基础.

主节点负责整个集群的任务提交、分配、管理与汇总, 不参与数据计算过程. 从节点负责数据的分布式存储与实际计算. 集群及节点详细硬件配置如图 4 和表 1.

2.3 软件环境

为了保持节点行为的一致性, 所有节点均采用相同的软件配置. 操作系统选用的是 CentOS,在其上构建程序运行所需的 Java 和 Scala 运行环境. Hadoop 框架提供计算平台所需的分布式文件系统 HDFS 和统一的资源调度机制 YARN. 软件环境配置如表 2.

3 结果及分析

实验结果按照是否使用 Spark 集群计算框架, 分为单机结果和集群结果. 单机结果在单一计算节点上本地运行获取, 集群结果在 YARN Cluster 模式下获取. 每种结果均采用不同的数据量作为变量进行测试: 10 GB (66 景)、50 GB (329 景)、100 GB (658 景)、500 GB (3347 景) 和 1000 GB (6634 景). 每项数据均在相同实验条件下重复三次, 取平均值. 具体实验结果如表 3.

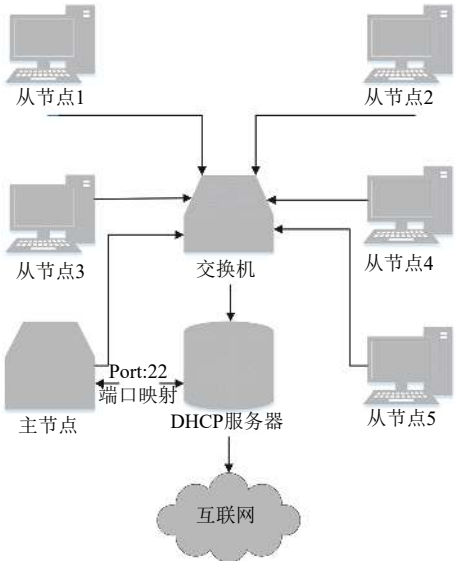

图 4 Spark 集群硬件体系架构

表 1 集群节点主要硬件配置

集群及节点	硬件	配置
主节点	中央处理器	Intel® Core™ i7-4790 CPU @ 3.60 GHz 4×2
	内存	SK Hynix DDR3L 1600 MHz 4 GB*2
	网卡	Intel Corporation Ethernet Connection I217-LM(rev04)
	硬盘	WD Blue WD10EZEX 1 TB 7200 rpm
从节点	中央处理器	Intel® Core™ i7-6700 CPU @ 3.40 GHz 4×2
	内存	SK Hynix DDR4 2133 MHz 4 GB*2
	网卡	Intel PRO/1000 Network Connection
	硬盘	Seagate ST1000DM003 1 TB 7200 rpm
网络设备	交换机	TP-Link TL-SF1008+
	DHCP 服务器	Huawei AR101W-S

表 2 集群节点主要软件配置

软件	版本
系统软件	CentOS 7.3 64 bit
开发语言包	Java Developement Kit 1.8.0.121 Linux 64 bit Scala 2.11.8
大数据处理软件	Hadoop 2.7.3 Spark 2.1.0Pre-built for Hadoop 2.7 & Later
开发工具软件	IntelliJ IDEA Ultimate 2017.2.4 for Linux 64 bit

在单机模式下, Scala 程序能够实现约每秒 0.04 GB (或每秒 0.29 个影像文件) 的处理能力. 在加入 Spark 并行处理框架后, Scala 程序能够实现约每秒 0.40 GB (或每秒 2.67 个影像文件) 的处理能力. 总体上, 在集群模式下影像处理的效率约为单机模式的 10 倍. 就不同的数据量而言, 当数据量较小时, 由于集群启动、并行化和任务调度等的开销, 处理效率低于平均水平. 当数据量在 50~500 GB 时, 影像处理效率呈现略微上升的态势. 但当数据量达到 1000 GB 时, 影像的处理效率出现小幅度回落. 可能在集群硬件资源达到饱和时, 在相同分区数量下, 单一分区的处理粒度加大, 处理任务间的间隙、资源调度与任务分发延迟加大, 导致了效率的下降.

表 3 对比实验结果

数据量 (影像数量)		1000 GB (6634 景)	500 GB (3347 景)	100 GB (658 景)	50 GB (329 景)	10 GB (66 景)
单机实验	运行耗时 (s)	25 379.0	12 291.0	2386.0	1227.0	231.3
	吞吐量	0.26 景/s; 0.04 GB/s	0.27 景/s; 0.04 GB/s	0.28 景/s; 0.04 GB/s	0.27 景/s; 0.04 GB/s	0.29 景/s; 0.04 GB/s
集群实验	运行耗时 (s)	2791.0	1252.0	257.7	130.7	36.3
	吞吐量	2.38 景/s; 0.36 GB/s	2.67 景/s; 0.40 GB/s	2.55 景/s; 0.39 GB/s	2.52 景/s; 0.38 GB/s	1.82 景/s; 0.28 GB/s

4 结论

本文提出了一种将 Spark 大数据并行计算技术应用于海表温度反演的方法. 通过将自动化影像获取过程、简化算法流程和 RDD 窄依赖并行计算结合, 实现了一种海表温度快速反演的机制. 通过对比单机模式和集群模式的处理效率, 验证了基于 Apache Spark 的海表温度反演方法的快速计算能力. 通过将大数据并行计算技术应用于遥感图像处理, 可以有效提高数据产品的更新频率, 为大范围、长周期和高分辨率的海洋、气象观测提供良好的支持. 本文采用的改进简化劈窗算法和 RDD 窄依赖特性也可以应用于其他遥感产品生产和海量数据并行计算过程.

参考文献

1 McMillin LM. Estimation of sea surface temperatures from two infrared window measurements with different absorption. Journal of Geophysical Research, 1975, 80(36): 5113–5117. [doi: 10.1029/JC080i036p05113]

2 覃志豪, 高懋芳, 秦晓敏, 等. 农业旱灾监测中的地表温度遥感反演方法——以 MODIS 数据为例. 自然灾害学报, 2005, 14(4): 64–71. [doi: 10.3969/j.issn.1004-4574.2005.04.011]

3 朱义明. 基于 Hadoop 平台的图像分类. 西南科技大学学报, 2011, 26(2): 70–73. [doi: 10.3969/j.issn.51-1660/C.2011.02.015]

4 付天新, 刘正军, 闫浩文. 基于 MapReduce 模型的生物量遥感并行反演方法研究. 干旱区资源与环境, 2013, 27(1): 130–136.

5 Almeer MH. Cloud hadoop map reduce for remote sensing image analysis. Journal of Emerging Trends in Computing and Information Sciences, 2012, 3(4): 637–644.

6 常生鹏, 马亿昨, 蔡立军, 等. 一种基于 Hadoop 的高分辨率遥感图像处理方法. 计算机工程与应用, 2015, 51(11): 167–171. [doi: 10.3778/j.issn.1002-8331.1403-0121]

7 刘震, 朱耀琴. 一种基于 Spark 的高光谱遥感图像分类并行化方法. 电子设计工程, 2017, 25(12): 19–22, 26. [doi: 10.3969/j.issn.1674-6236.2017.12.005]

8 Toller G, Isaacman A, Leader M. MODIS Level 1B Product User's Guide. https://mcst.gsfc.nasa.gov/sites/mcst.gsfc/files/file_attachments/M1054D_PUG_083112_final.pdf. [2012-07-12].

9 Kaufman YJ, Gao BC. Remote sensing of water vapor in the near IR from EOS/MODIS. IEEE Transactions on Geoscience and Remote Sensing, 1992, 30(5): 871–884. [doi: 10.1109/36.175321]

10 毛克彪, 唐华俊, 李丽英, 等. 一个从 MODIS 数据同时反演地表温度和发射率的神经网络算法. 遥感信息, 2007, (4): 9–15, 8. [doi: 10.3969/j.issn.1000-3177.2007.04.002]

11 高懋芳, 覃志豪, 徐斌. 用 MODIS 数据反演地表温度的基本参数估计方法. 干旱区研究, 2007, 24(1): 113–119.