Dps-MuSyQ: A Distributed Parallel Processing System for Multi-Source Data Synergized Quantitative Remote Sensing Products Producing

HONGYI LI1,2 AND PING TANG1

1Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
2School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Hongyi Li (lihy@radi.ac.cn)

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA19080301, and in part by the National Key Research and Development Program under Grant 2018YFA0605500.

ABSTRACT With the development of earth observation technologies and the construction of earth observation systems, an increasing amount of remote sensing data are being obtained. These provide the datasets required for research on remote sensing monitoring across large areas. To compensate for the shortcomings of global and large-area temporal monitoring data, synergized computing using multi-source remote sensing data can improve the accuracy and temporal resolution of remote sensing monitoring. However, remote sensing data are drawn from multiple sources and multiple scales, and have a complex structure and large volume; in addition, the nested system architecture of multi-source synergized remote sensing products makes the design of large-scale multi-source synergized remote sensing monitoring systems difficult. In this paper, we describe the design and implementation of a distributed parallel processing system for multi-source data synergized quantitative remote sensing based on a distributed cluster platform. The system integrates the algorithms normalizing more than 30 kinds of data sources and producing 40 quantitative remote sensing products. The system also connects a number of centers for satellite data, serves for several applications, and implements dynamic expansion integration for highly efficient quantitative remote sensing products. The system has produced approximately 50 TB of quantitative remote sensing products, and the application of these data to agriculture, forestry, the environment, and water conservation has resulted in very positive effects.

INDEX TERMS Multi-source data synergies, quantitative remote sensing, task parallelism, data integration, distributed system.

I. INTRODUCTION

With the development of earth observation technologies and the construction of earth observation systems, the total amount of remote sensing data has increased exponentially, and remote sensing has entered the era of big data. For instance, the earth observing systems operated by the US National Aeronautics and Space Administration (NASA) [1], National Oceanic and Atmospheric Administration (NOAA) [2], and the China High-resolution Earth Observation System (CHEOS) [3] receive and distribute a large number of remote sensing data, particularly in the case of NASA-EOS, which receives daily archiving and distribution data of the order of several TB [4].

The increased number of remote sensing observation methods and the long-term accumulation of remote sensing data provide a database for remote sensing monitoring research across large areas of the world, the growth of remote sensing data brings opportunities and challenges to global change [5]–[9]. At present, many researchers are extending their research scope to the global scale, and the 50 m global land cover mapping based on Landsat satellite data has been completed [10], [11], such as the global forest mapping [12]. Based on Moderate Resolution Imaging Spectroradiometer (MODIS) and NOAA satellite data, various quantitative remote sensing products [13] with a global resolution of 1 km have been developed, such as the case of NASA-EOS, which receives daily archiving and distribution data of the order of several TB [4].
leaf area index (LAI) [14], albedo (Albedo) [15], downward shortwave radiation (DSR) [16], photo synthetically active radiation (PAR) [17], and broadband emissivity (BBE) [18].

And the cloud system platform [19] that can produce a variety of remote sensing products and the framework platform pipsCloud [20], for remote sensing data processing and management. These global and large-scale remote sensing products mainly use single data sources or single-resolution data as input, without considering the cooperative operation of multi-source remote sensing data. Regional remote sensing monitoring and monitoring products are mostly based on a single data source, such as the Feng Yun satellite ground system [21] and harmful algal bloom monitoring in the Yellow Sea using the HJ1-CCD satellite [22]. There has been no deep collaboration on multi-source remote sensing data algorithms, and there is no synergistic use of multi-resolution data. Long time series monitoring by satellite data in global and local regions requires multi-source remote sensing data to enable synergized computations that improve the accuracy and spatial-temporal resolution of monitoring products [23], [24].

The accumulation of remote sensing data provides a database for global and large-scale remote sensing monitoring research. Remote sensing data are generally drawn from multiple sources and multiple scales, have a complex structure and various formats, and can be of very large volume. These characteristics present various challenges to the monitoring of multi-source remote sensing data in global and large-scale regions. Moreover, the structure of the production system of quantitative remote sensing products has the characteristic of hierarchical nesting: for instance, the normalized difference vegetation index (NDVI) [25] and vegetation net primary productivity (NPP) [26] have different production levels. Because of the uncertainty of multi-source raw data in the production system of large-scale multi-source synergized products, the inputs to upper products are not fixed, which makes it difficult for multi-source remote sensing data to be collaboratively monitored. In addition, each data center and application department of satellite remote sensing data stores and manages these data according to their own needs in accordance with their respective industries and the application of various standards. As a result, the data structure is complicated and disorderly, which must be considered in designing a platform for a production system of synergized quantitative products for multi-source remote sensing data.

This study considers the construction of a robust and efficient production system for multi-source remote sensing products. This system can be used to produce global and large-scale multi-source synergized quantitative remote sensing products, is robust to production problems brought about by the multi-source, multi-scale, complex structure, and diversity characteristics of remote sensing data, and can implement the automatic production of multi-hierarchical nested quantitative remote sensing products. Under the support of China’s National High Technology Research and Development (“863”) Program, which considers systems for satellite-aircraft-ground comprehensive quantitative remote sensing and its applications, we have developed a Distributed Parallel Processing System for Multi-source data with Synergized and Quantitative remote sensing (Dps-MuSyQ).

The proposed system produces 40 kinds of multi-source synergized quantitative remote sensing products and more than 30 normalized products on the global and regional scale. It requires multi-source data to be used collaboratively across multiple data centers and can form a sustained, long-term production capacity. This paper describes the design and construction of Dps-MuSyQ based on the distributed cluster platform, and discusses the implementation of the high-performance production of multi-source quantitative remote sensing products.

II. SYSTEM OBJECTIVES

In view of the characteristics of multi-source data collaboration and the hierarchical nesting architecture in the production system of quantitative remote sensing products, Dps-MuSyQ is intended to achieve the following objectives: (1) implementation of large-scale, rapid processing of remote sensing data; (2) implementation of on-demand and automated production of hierarchical nested multi-source synergized remote sensing products.

To implement the generation of high-performance products, parallel processing is used to solve the geographical dispersion of remote sensing data, as well as the resource dispersion and I/O bottleneck problems caused by the centralized processing of big data. To achieve the on-demand automated production and dynamic scalability of quantitative remote sensing products, the solution is as follows:

- Automatically generate tasks that can be performed by the user;
- Automatically execute a hierarchical nested algorithm for the production of quantitative remote sensing products;
- Integrate large-scale data in a single computing task [27].

III. SYSTEM DESIGN

A. DISTRIBUTED ARCHITECTURE

The overall architecture of Dps-MuSyQ is designed for distributed and high-efficiency processing requirements, and its computing center can expand laterally and dynamically. The system architecture and internal modules of Dps-MuSyQ are shown in Fig. 1. There are six relatively independent components: User Interface (UI), Task Factory, Task Proxy, Computing Center, Metadata Center, and Public Storage. Each component can be deployed independently, and can also be integrated; the six components form a unified whole that accomplishes the multi-source data collaboration of quantitative remote sensing products. The UI is controls the human-computer interaction, and is responsible for receiving the user’s production requirements and returning the data and the product status to the user. The Task Factory is the order and task management center, in which the core function is
to convert the user order requirements into a task that the computer can perform according to the system architecture for multi-source synergized quantitative remote sensing products. The Task Proxy is responsible for receiving the tasks assigned by the Task Factory and scheduling the Computing Center for calculations. The Metadata Center manages the metadata required for the creation of the entire system and provides metadata services. Public Storage is responsible for storing multi-source remote sensing data and providing data services.

**B. COMPONENT DESIGN**

The UI is oriented to the user requirements, providing users with Order Service and Data Service.

Order Service: the users select orders through the UI to generate the type, time range, and space range for the quantitative remote sensing products. These inputs generate production orders that are submitted to the Task Factory. After an order has been submitted, the users can query the status of the order through the UI.

Data Service: the users store the input and query data of the products through the UI, import local or external disks into the system, and export the queried products to local or external storage. The import function requires the location of the data to be specified, and then they are automatically imported into the production process.

The core functionality of the Task Factory consists of Order Manage and Task Manage operations.

Order Manage: the Order Service of the UI is a user-oriented human-computer interaction, but the Order Manage operation in the Task Factory is oriented to order management for system operations. First, orders submitted by the user are stored as XML files; these are subjected to a format check and lexical analysis, and orders that pass the check will be processed by the Task Manage operation. Order Manage is also responsible for tracking the status of the order and constantly updating the order status in the metadata database.

Task Manage: for orders that pass the format check, a Task Parser is responsible for automatically parsing orders into tasks using a product knowledge base. Each task corresponds to a Task Scripts File, which is submitted to the Task Proxy by the Task Submit operation. Task Manage is also responsible for tracking the status of tasks and storing the status in the metadata database.

The Task Proxy is responsible for receiving and scheduling tasks, with the main functions including Task Accept, Top Scheduler, and Computing Center Listen (CC Listen).
Task Accept is responsible for receiving submitted tasks from the Task Factory. CC Listen monitors the state of the Computing Center, and the Top Scheduler is responsible for sending tasks to the Computing Center according to the CC Listen status and completing the top-level task scheduling.

The Computing Center is designed for production operations, and mainly consists of a Task Scheduler, Algorithm Pool, and Data Assembly.

The functions of the Task Scheduler include Task Accept, Down Scheduler, and Task Listen. Task Accept receives the task pushed by the Proxy, the Down Scheduler is responsible for scheduling tasks to every processor node/core, and Task Listen monitors the status of each task.

Algorithm Pool: each Computing Center stores algorithms for quantitative remote sensing products, and then forms an algorithm pool for their production. Additionally, the input, output, and data flow of these algorithms are stored. Finally, the knowledge base of the hierarchical nested quantitative remote sensing product algorithm is constructed.

Data Assembly: this is designed for large-scale data input of a single task. Most of the multi-source quantitative synergized remote sensing products are synthetic, referring to the desired geographical range and time range. A single task involves a large amount of input data and multiband remote sensing data, and the processes of single-product production do not require all bands. Thus, a Data Extract function is needed to reduce the volume of data before the production process. In addition, input data from different sources have inconsistent widths and scales, so it is necessary to use Data Alignment to facilitate the collaborative use of various data in the same production task.

The core function of the Metadata Center is to provide meta-information services and check the internal consistency of the data. The functions of this module include a Data Catalog Service, Data Interface Service, and Data Consistency Check.

Data Catalog Service: this is the core of the Metadata Center, providing numerous data meta-information inquiry services.

Data Interface Service: this provides the interface for the import and export of original remote sensing data, quantitative remote sensing product data, and metadata information.

Data Consistency Check: this function stores the MD5 value and data version number of the product file as meta-information for the corresponding product database table. It also periodically proofreads the extracted MD5 values and version numbers of meta-information and disk files stored in the database to ensure the consistency of data records and data file entities, and updates the database tables according to this proofreading in real time.

Public Storage is a globally unified data storage space that provides consistent external data mapping, and any Computing Center has the same access rights. The Private Center of Public Storage and Computing Storage makes up the entire product entity data storage space of the system.

C. SYSTEM FLOW
Dps-MuSyQ takes the three input elements of the user’s selection of quantitative remote sensing product, and applies the following workflow steps (see Fig. 2) until the production process is finished.

- Selection of three input factors: in the UI, the type, time range, and geographical coverage of the products are selected, and then the production order is submitted.
- Metadata query: according to the three elements of the order, the system automatically queries the knowledge base, and the input data are determined by the production system and knowledge base according to the desired time and space range.
- Task parser: the system parses the query according to the production system architecture, and then generates a script file to produce the required product.
- Top scheduling: according to the distribution of server resources, server status, and data on the server, the system schedules the task script files among the servers.
- Down scheduling: According to the resource and state of the nodes in the server, the system schedules the task scripts allocated to the server and schedules the scripts to the processor core.
- Data acquisition: according to the data address config in the script, the data required for product production are obtained locally or remotely before the production algorithm is executed.
- Input reconstruction: according to the production rule of the product algorithm, the input data are reconstructed and organized into an easy-to-use structure for quantitative remote sensing products, and then the product is produced.
- Data maintenance: after the implementation of the production algorithm, order information, product information, and task information related to the whole production process are maintained.

IV. IMPORTANT MODULE IMPLEMENTATION
In view of the system objectives, the implementation of important Dps-MuSyQ function modules is introduced for the realization of two kinds of goals. On the business level of automated on-demand production, the Task Parser and Data Integration are implemented. The Task Parser parses user requirements into tasks that a computer can perform. Through design rules, on-demand production is achieved for hierarchical nested quantitative remote sensing products.
The newly added quantitative remote sensing products can be integrated into Dps-MuSyQ according to certain rules, and can be combined with existing quantitative remote sensing products to form a quantitative remote sensing product system framework. Data Integration realizes the collaborative calculation of large-scale data in a single computing task to produce real-time assembled data, and ensures the memory and external storage demands remain within the controllable range. To enhance system performance, efficient task parallelization based on a distributed cluster environment is implemented using a Top-Down Scheduler strategy.

**A. TASK PARSER**

The Task Parser converts user requirements into executable tasks. A comprehensive analysis is made by combining the input and output information of each quantitative remote sensing product, the production demand, and the meta-information knowledge base stored in the database. The user’s production requirements are parsed into task scripts that can be executed by the computer. Firstly, the input parameters of multi-source quantitative synergized remote sensing products are analyzed, then the rules are modeled and a task script is generated.

1) **ALGORITHM INPUT**

The input parameters of the production algorithm for quantitative remote sensing products using synergized multi-source data have the characteristics of hierarchical nesting and a variable number of input parameters. In addition, because there are multiple sources for the input parameters of quantitative remote sensing products, there can be inconsistencies in time resolution, spatial resolution, and data format between input parameters, as well as between input parameters and output products.

- **a: NESTED HIERARCHY STRUCTURE**

The input parameters of the 1 km Fractional Vegetation Cover (FVC) [16] products are shown in Fig. 3(b). The most direct input parameters for producing FVC are NDVI/1KM and MCD12Q1. The NDVI/1KM (Vegetation Index) requires eight normalized input data, as shown in Fig. 3(a). Therefore, at the beginning of the normalization process, the production schema levels of NDVI and FVC are defined as 1 and 2, respectively. In particular, direct and indirect input parameters will be more complex, and there are likely to be higher production levels for higher-level products such as the vegetation NPP.

- **b: UNKNOWN THE BEGINNING**

Because quantitative remote sensing products have a nested hierarchical structure, the variability of user demand leads to the storage of different levels of input data at different times. This leads to many unknowns when the product is first produced, as shown in Fig. 4. When user A generates the NDVI product within the R2 area and archives the data, and user B generates FVC products within the R1 area, the beginning points of the product are divided into two categories.

For FVC products in the R2 region, direct production is performed using NDVI and MCD12Q1, and the product is produced at a depth of 1. For R1’s FVC products, the NDVI production algorithms are invoked to produce NDVI, and then NDVI and MCD12Q1 are used to produce FVC, and the product is produced at a depth of 2.

The above covers only FVC products with a production level of 2. When the production level is higher and the production input is greater, the start point of the production process will be more complicated.

- **c: VARIABLE PARAMETER NUMBERS**

When fixed-size quantitative remote sensing products are generated, there are two reasons why the number of input parameters in the production algorithm is not fixed. Firstly, the satellite transit cycles and coverage in different regions are inconsistent, and the amount of data received varies. Secondly, the process of receiving, processing, and storing data leads to different amounts of data at different times, and these data may cover different regions. Thus, the number of parameters is not fixed so that the system can accommodate the...
variable number of input parameters. However, for each of the quantitative remote sensing products, certain input parameter constraints must be satisfied. In the case of NDVI/1KM, these are illustrated in Fig. 3(a), where the quantitative relations among the input parameters are defined below. The total number of input parameters in the six classes A-F must be greater than or equal to 1. The input parameters of G and H must be greater than or equal to 1 for each class of input parameters.

2) USER INPUT PARSER STATIC TASK SCRIPTS

The parsing of user requirements into task script files requires three steps: rules and formal descriptions, modeling, and generation of task scripts. Combined with the algorithm input characteristics, the implementation process of each step is as follows.

a: RULES AND FORMAL DESCRIPTIONS

The hierarchical nesting and number of input parameters of the quantitative remote sensing product algorithm are not fixed, resulting in a high degree of complexity in the production process. Therefore, all quantitative remote sensing products will need to be processed to build rules and formal descriptions.

First, the input of the product algorithm is classified according to the rules of the data source and the product type, as shown in Fig. 3. The NDVI input parameters are divided into eight categories, A-H. The product level of NDVI is defined as 1, and the product level of FVC is defined as 2. The input parameter relationships are constrained for each class of the product to give formal descriptions. The input constraint relationship of the NDVI product is shown in (1), and the input constraint relationship of the FVC product is shown in (2).

\[ \{(A, x_1) \cup (B, x_2) \cup (C, x_3) \cup (D, x_4) \cup (E, x_5) \cup (F, x_6) \cup (G, x_7) \cup (H, x_8) \Rightarrow (VI, 1) \} \tag{1} \]

\[ \{(A, x_1) \cup (B, x_2) \Rightarrow (FVC, 1) \} \{[x_1 \geq 1, x_2 \geq 1] \} \tag{2} \]

b: MODELING

Modeling is used to describe the system rules and formal expressions, and to establish models that computers can recognize and handle. The XML language is used to realize modeling. The input parameters are inducted as eight metadata items (see Table 1) and the output parameters are inducted as six metadata items (see Table 2).

For a constraint in (1), the six input parameters are merged as homogeneous inputs. Thus, after modeling, there are three NDVI input parameters. As the FVC input parameters are not merged, it still has two input parameters.

c: GENERATE TASK SCRIPTS

The process of generating task scripts is illustrated in Fig. 5, and proceeds as follows. The three elements (time range, space range, product type) input by the user are converted into a task unit within the system according to the time resolution reference of the product and the single scene/image space reference. Combining the product modeling rules, data are recursively queried in the Data Catalog Service. Task parsing and the pairing of tasks are performed for the task unit and the queried data in accordance with the modeling rules. This excludes conflicting tasks resulting from different production branches and duplicate tasks that the database has already produced. The script files are stored hierarchically for scheduling according to the product priority, which is based on the modeling rules.
When the level of the production is 2, there are 2 folders for generating scripts, and the folder name is indicated by the level number. The folder named “1” stores the production scripts with all production levels of 1, which contain many different types of the products. The folder named “2” stores the production script at level 2. A hierarchical nested production script with production level 2 is shown in Fig. 6. The production of a product requires a variety of different types of inputs, and there are multiple data for the same type of inputs. Input parameters of different types are separated by spaces, and multiple data of the same type are separated by commas. According to the scheduling rules, the scheduling system first completes the production script in the folder “1”, then executes the production script in this level “2”, so, the hierarchical nested production of multi-input products has been completed.

B. DATA INTEGRATION

After the static task script has been generated, the production task is scheduled to run on the execution node by the scheduling system. Multi-source and multi-temporal synthetic products are produced by Dps-MuSyQ with the following characteristics:

- A single multi-source synthetic product involves multiple sensors or multiple types of data for simultaneous collaborative computations;
- Multi-temporal synthesis results in multiple sets of input data from a single sensor, resulting in a large amount of data for a single product.

As multi-source sensor data, multi-type data, and duplicate data that are repeatedly covered by the same sensor need to be collaboratively produced in the same production task, the production reference of the product is required to determine the subdivision specification of the product (resolution, size, starting point of the subdivision, etc.) [28], [29]. Different data are then organized according to the same subdivision specification. Taking NDVI/1KM five-day products as an example, the process of Data Integration is shown in Fig. 7. First, the input of multiple 1 km sensor data is processed by Data Extraction and Geometric Alignment, organized into Memory Cubes, and then combined with other data sources. Data Extraction takes the required datasets from the source data, and Geometric Alignment organizes data from different sources in a geographically aligned manner.

Data Integration requires multiple remote sensing data for a single synergized production task. The Dps-MuSyQ system uses a parallel task mode to implement high-performance distributed processing. To achieve high task parallelism within a node, multi-level cache optimization is employed for a single production task, and the total memory usage of a single task is controlled to be less than 2 GB.
TABLE 4. List of synergized and quantitative remote sensing products.

| No | Products name                          | Abbreviation       | No | Products name                          | Abbreviation |
|----|----------------------------------------|--------------------|----|----------------------------------------|--------------|
| 1  | Downward Longwave Radiation at 5 km resolution | MuSyQ.DLR.5km      | 21 | Cloud Intensity Indices at 1 km resolution | MuSyQ.CII.1km |
| 2  | Photosynthetically Active Radiation at 5 km resolution | MuSyQ.PAR.5km      | 22 | Enhanced Vegetation Index at 1 km resolution | MuSyQ.EVI.1km |
| 3  | Downward Shortwave Radiation at 5 km resolution | MuSyQ.DSR.5km      | 23 | Net Primary Productivity at 1 km resolution | MuSyQ.NPP.1km |
| 4  | Land Surface Temperature at 5 km resolution | MuSyQ.LST.5km      | 24 | Fractional Vegetation Cover at 1 km resolution | MuSyQ.FVC.1km |
| 5  | Net Radiation at 5 km resolution        | MuSyQ.NRD.5km      | 25 | Sea Ice Temperature at 1 km resolution   | MuSyQ.SIT.1km |
| 6  | Impervious Surface Cover at 1 km resolution | MuSyQ.ISC.1km      | 26 | Evapotranspiration at 1 km resolution    | MuSyQ.ET.1km |
| 7  | Atmospheric Water Vapor Content at 1 km resolution | MuSyQ.TCWV.1km    | 27 | Surface Reflectance at 30 m resolution  | MuSyQ.REF.30m |
| 8  | Surface Reflectance at 1 km resolution | MuSyQ.REF.1km      | 28 | Land Surface Albedo at 30 m resolution | MuSyQ.LSA.30m |
| 9  | Land Surface Temperature at 1 km resolution | MuSyQ.LST.1km      | 29 | Fraction of Photosynthetically Active Radiation at 30 m resolution | MuSyQ.FPAR.30m |
| 10 | Land Surface Emissivity at 1 km resolution | MuSyQ.LEL.1km      | 30 | Normalized Difference Vegetation Index at 30 m resolution | MuSyQ.NDVI.30m |
| 11 | Land Surface Albedo at 1 km resolution | MuSyQ.LSA.1km      | 31 | Silicification Anomaly Index at 30 m resolution | MuSyQ.SAI.30m |
| 12 | Fraction of Absorbed Photosynthetically Active Radiation at 1 km resolution | MuSyQ.FPAR.1km | 32 | Atmospherically Resistant Vegetation Index at 30 m resolution | MuSyQ.ARVI.30m |
| 13 | Normalized Difference Vegetation Index at 1 km resolution | MuSyQ.NDVI.1km | 33 | Aerosol Optical Depth at 30 m resolution | MuSyQ.AOD.30m |
| 14 | Sea Ice Distribution at 1 km resolution | MuSyQ.SID.1km      | 34 | Hydroxyl Anomaly Index at 30 m resolution | MuSyQ.HAI.30m |
| 15 | Atmospherically Resistant Vegetation Index at 1 km resolution | MuSyQ.ARVI.1km | 35 | Normalized Difference Water Index at 30 m resolution | MuSyQ.NDWI.30m |
| 16 | Aerosol Optical Depth at 1 km resolution | MuSyQ.AOD.1km      | 36 | Soil Brightness Index at 30 m resolution | MuSyQ.SBI.30m |
| 17 | Normalized Difference Water Index at 1 km resolution | MuSyQ.NDWI.1km | 37 | Leaf Area Index at 30 m resolution | MuSyQ.LAI.30m |
| 18 | Soil Brightness Index at 1 km resolution | MuSyQ.SBI.1km      | 38 | Enhanced Vegetation Index at 30 m resolution | MuSyQ.EVI.30m |
| 19 | soil moisture index at 1 km resolution | MuSyQ.SMI.1km      | 39 | Fraction of Vegetation Cover at 30 m resolution | MuSyQ.FVC.30m |
| 20 | Leaf Area Index at 1 km resolution      | MuSyQ.LAI.1km      | 40 | Net Primary Productivity at 300 m resolution | MuSyQ.NPP.300m |

C. TOP-DOWN SCHEDULER

The scattered nature of remote sensing data, the decentralization of processing resources, and the specialty of remote sensing data processing mean that the relevant algorithms must be developed by professionals. Taking into account the efficiency and feasibility of product generation, a parallel task model based on distributed clusters is designed.

To achieve resource scheduling and load balancing optimization for a distributed computing environment, a two-level Top-Down task scheduling architecture has been designed (see Fig. 8).

Top Scheduler: For all computational tasks that are committed to the server, the overall schedule is performed according to the total/idle amount of computing and storage resources and the data storage distribution in the distributed cluster. Scheduling proceeds from low to high according to the production level of the product. The overall strategy is as follows:

Data placement priority: the calculated data are sorted according to the data distribution in different clusters. The larger the portion of task data in the cluster, the higher the priority of that computing task.

Computational resource priority: all computing resources are arranged in the cluster, and tasks are assigned to the corresponding cluster in conjunction with the estimated computation time of each task. The computation $C_j$ of the $jth$ cluster is given by:

$$C_j = \frac{P_j \times F_j}{\sum_{i=1}^{N} P_i \times F_i} \quad (3)$$

where $P_i$ and $P_j$ denote the number of cores in the $ith$ and $jth$ clusters, respectively, $F_i$ and $F_j$ denote the $ith$ and $jth$ dominant frequencies, respectively, and $N$ denotes the number of clusters. Ignoring network latency, the Top Scheduler mainly uses computational resource optimization for scheduling.

Down Scheduler: For internal computing tasks assigned to the cluster, business and open source task scheduler software are used to perform low-level scheduling between nodes to ensure load balancing within the Cluster. Because the storage inside the cluster is transparent, there is no need to consider the optimal placement of computing data within the cluster.

V. SYSTEM PERFORMANCE AND PRODUCTS

The proposed Dps-MuSyQ system integrates the normalization and production of more than 30 data sources (see Table.3) at scales of 30 m, 1 km, and 5 km, and produces 40 quantitative remote sensing products (see Table.4) covering...
FIGURE 9. Remote sensing products examples.

(a) Daily total downward shortwave radiation in 5 km resolution

(b) 5 days composite LAI in 1 km resolution

(c) 1 month composite maximum fractional vegetation coverage in 30 m resolution
vegetation, radiation, ice, snow, and water. Data from the National Satellite Meteorological Center, National Satellite Ocean Application Service, China Centre for Resources Satellites Data and Application, Institute of Remote Sensing and Digital Earth, and Computer Network Information Center can be integrated within the system. Each part of the system is deployed on the server of the Chinese Academy of Sciences. The Task Proxy, Computing Center, and Public Storage components are deployed in the distributed cluster of the Gigabit Internet connections in the Computer Network Information Center, the Task Factory and Metadata Center components are deployed in the server of the Institute of Remote Sensing and Digital Earth, and UI components are deployed in users’ personal terminals.

The single server configuration for the distributed cluster in the Computer Network Information Center is as follows:
- CPU: Intel(R) Xeon(R) CPU E5645 @2.40 GHz with 24 CPUs;
- Memory: 64 GB;
- Storage: 32 TB;
- OS: Centos 6.6 64-bit.

The server configuration at the Institute of Remote Sensing and Digital Earth is as follows:
- CPU: Intel (R) Xeon (R) CPU E5-4607 V2 @2.60 GHz with 32 CPUS;
- Memory: 32 GB;
- Storage: 144 TB;
- OS: Windows Server 2008, R2 Enterprise 64 bit.

Fig. 9 shows three remote sensing products from Dps-MuSyQ system, which are 5 km daily total downward shortwave radiation (DSR), the 1 km/5-day composite leaf area index (LAI), and 30 m/1 month composite maximum fractional vegetation coverage (FVC).

For an executable task, after parsing, Dps-MuSyQ dispatches the production tasks to each cluster according to the product level, and then completes the Down Scheduler within the cluster. Dps-MuSyQ optimizes the memory and external storage usage of a single production task. Compared with a single production task, parallel task scheduling achieves speedup that is consistent with the number of CPU cores in the cluster, using the hardware performance of the system to improve the speed of remote sensing monitoring responses. The system rules and models for the input and output parameters are handled in the Task Parser, which makes the program complexity of the system robust to the type of inherited multi-source data and quantitative remote sensing product. A configuration file must be added according to the modeling rules without increasing the amount of coding, so as to achieve the horizontal dynamic expansion of the product type.

**VI. CONCLUSION**

A distributed parallel processing system for multi-source data with synergized quantitative remote sensing products has been designed based on a distributed cluster environment. The Dps-MuSyQ framework integrates more than 30 kinds of standard products and 40 quantitative remote sensing production processes, and finally realizes the automatic production of multi-source synergized quantitative remote sensing products according to user requirements (time, space, and product type). The system consists of important components such as a Task Parser, Data Integration, and Top-Down Scheduler to solve the challenges of multi-source, multi-scale data with complex structures, diverse formats, large volumes, and hierarchical nested architectures. The result is a production system for synergized quantitative remote sensing products.

To guarantee the compilation of the “China - ASEAN regional ecological environment” report in the National Remote Sensing Center of China’s 2014 annual report, and for the joint experiments of Gansu, Zhangye, China, the Dps-MuSyQ system has produced about 50 TB of products containing approximately two million records for Southeast Asia and Gansu, Zhangye, China. The products have been applied in fields such as agriculture, forestry, the environment, and water conservation, and have achieved good results.

In future work, we will further design and implement a scheduling strategy combining data placement and computing resources to further improve the production efficiency of Dps-MuSyQ and achieve a greener computing environment. Moreover, the Dps-MuSyQ will be further adapted and deployed to the cloud computing platform to support large-scale remote sensing applications and more open application models.

**REFERENCES**

[1] H. K. Ramapuroriyan, J. Behnke, E. Sofinowski, D. Lowe, and M. A. Esfandiari, “Evolution of the earth observing system (EOS) data and information system (EOSDIS),” in *Standard-Based Data and Information Systems for Earth Observation*. Berlin, Germany: Springer, 2010, pp. 63–92.

[2] S. P. Center. (2012). *National Oceanic and Atmospheric Administration*. Seattle, Washington, DC, USA. [Online]. Available: https://www.noaa.gov/

[3] D. R. Li, X. Shen, H. C. Ma, and G. Zhang, “High-resolution earth observation system is imperative,” *Geomatics Inf. Sci. Wuhan Univ.*, vol. 39, no. 4, pp. 386–389, 2014.

[4] X. Lu, C. Cheng, J. Gong, and L. Guan, “Review of data storage and management technologies for massive remote sensing data,” *Sci. China Technol. Sci.*, vol. 54, no. 12, pp. 3220–3232, Dec. 2011.

[5] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, “Remote sensing big data computing: Challenges and opportunities,” *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, Oct. 2015.

[6] M. M. U. Rathore, A. Paul, A. Ahmad, B.-W. Chen, B. Huang, and W. Ji, “Real-time big data analytical architecture for remote sensing application,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, pp. 4610–4621, Oct. 2015.

[7] H. Zhang, Y. Yuan, W. Li, and B. Zhang, “A real-time precipitable water vapor monitoring system using the national GNSS network of China: Method and preliminary results,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1587–1598, May 2019.

[8] W. Jing, S. Huo, Q. Miao, and X. Chen, “A model of parallel mosaicking for massive remote sensing images based on spark,” *IEEE Access*, vol. 5, pp. 18229–18237, 2017.

[9] L. Wang, Y. Ma, A. Y. Zomaya, R. Ranjan, and D. Chen, “A parallel file system with application-aware data layout policies for massive remote sensing image processing in digital Earth,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1497–1508, Jun. 2015.

[10] P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, and S. Liu, “Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data,” *Int. J. Remote Sens.*, vol. 34, no. 7, pp. 2607–2654, 2013.
[11] C. Jun, Y. Ban, and S. Li, “Open access to Earth land-cover map,” Nature, vol. 514, no. 7523, p. 434, Oct. 2014.

[12] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, “High-resolution global maps of 21st-century forest cover change,” Science, vol. 342, no. 6160, pp. 850–853, Nov. 2013.

[13] X. Zhao, S. Liang, S. Liu, W. Yuan, Z. Xiao, Q. Liu, J. Cheng, X. Zhang, H. Tang, X. Zhang, Q. Liu, G. Zhou, S. Xu, and K. Yu, “The global land surface satellite (GLASS) remote sensing data processing system and products,” Remote Sens., vol. 5, no. 5, pp. 2436–2450, 2013.

[14] L. Pasolli, S. Asam, M. Castelli, L. Bruzzone, G. Wohlfahrt, M. Zebisch, and C. Notarnicola, “Retrieval of leaf area index in mountain grasslands in the Alps from MODIS satellite imagery,” Remote Sens. Environ., vol. 165, pp. 159–174, Aug. 2015.

[15] Y. Liu, Z. Wang, Q. Sun, A. M. Erb, Z. Li, C. B. Schaaf, X. Zhang, M. O. Román, R. L. Scott, Q. Zhang, K. A. Novick, M. S. Bret-Harte, S. Petroy, and M. SantClements, “Evaluation of the VIIRS BRDF, albedo and NBAR products suite and an assessment of continuity with the long term MODIS record,” Remote Sens. Environ., vol. 201, pp. 256–274, Nov. 2017.

[16] T. Wang, G. Yan, X. Mu, Z. Jiao, L. Chen, and Q. Chu, “Toward operational shortwave radiation modeling and retrieval over rugged terrain,” Remote Sens. Environ., vol. 205, pp. 419–433, Feb. 2018.

[17] M. Yaseen, M. Singh, U. B. Singh, S. Singh, and M. Ram, “Optimum planting time, method, plant density, size of planting material, and photo synthetically active radiation for safed musli (Chlorophytum borivilianum),” Ind. Crops Products, vol. 43, pp. 61–64, May 2013.

[18] V. A. G. Rivera, M. El-Amraoui, Y. Ledemi, Y. Messaddeq, and E. Marega, “Expanding broadband emission in the near-IR via energy transfer between Er³⁻–Tm³⁺ co-doped tellurite-glasses,” J. Lumin., vol. 145, pp. 787–792, Jan. 2014.

[19] Y. Yan, Y. Ma, L. Wang, K.-K.-R. Choo, and W. Jie, “A cloud-based remote sensing data production system,” Future Gener. Comput. Syst., vol. 86, pp. 1154–1166, Sep. 2018.

[20] L. Wang, Y. Ma, J. Yan, V. Chang, and A. Y. Zomaya, “PipsCloud: High-performance cloud computing for remote sensing big data management and processing,” Future Gener. Comput. Syst., vol. 78, pp. 353–368, Jan. 2018.

[21] N. M. Lv, C. H. Dong, Z. D. Yang, J. M. Shi, and P. Zhang, “Ground segment of the new general of fengyun popoar orbit meteorological satellite (FY-3) and its data application,” Eng. Sci., vol. 14, no. 9, pp. 10–19, Sep. 2012.

[22] C. Q. Wu, W. D. Ma, X. L. Xue, Y. J. Yao, and D. Wu, “Remote sensing monitoring HAB in yellow sea by HJ1-CCD,” Environ. Monit. Assess., vol. 31, no. 3, pp. 166–170, 2015.

[23] B. Zhong, A. Yang, A. Nie, Y. Yao, H. Zhang, S. Wu, and Q. Liu, “Finer resolution land-cover mapping using multiple classifiers and multisource remotely sensed data in the heihe river basin,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 10, pp. 4973–4992, Oct. 2015.

[24] R. Li, X. Li, J. Yuan, and J. Yao, “A triangle-oriented Spatial–Temporal phase unwrapping algorithm based on rotational constraints for time-series InSAR,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 12, pp. 10263–10275, Dec. 2019.

[25] X. Fan and Y. Liu, “A global study of NDVI difference among moderate-resolution satellite sensors,” ISPRS J. Photogramm. Remote Sens., vol. 121, pp. 177–191, Nov. 2016.

[26] G. Pachavo and A. Murwira, “Remote sensing net primary productivity (NPP) estimation with the aid of GIS modelled shortwave radiation (SWR) in a southern african savanna,” Int. J. Appl. Earth Observ. Geoinf., vol. 30, pp. 217–226, Aug. 2014.

[27] F. Van Den Bergh, K. J. Wessels, S. Miteff, T. L. Van Zyl, A. D. Gazendam, and A. K. Bachoo, “HiTempo: A platform for time-series analysis of remote-sensing satellite data in a high-performance computing environment,” Int. J. Remote Sens., vol. 33, no. 15, pp. 4720–4740, Aug. 2012.

[28] T. Ma, C. Zhou, Y. Xie, B. Qin, and Y. Ou, “A discrete square global grid system based on the parallels plane projection,” Int. J. Geographical Inf. Sci., vol. 23, no. 10, pp. 1297–1313, Oct. 2009.

[29] M. Zhou, J. Chen, and J. Gong, “A pole-oriented discrete global grid system: Quaternary quadrangle mesh,” Comput. Geosci., vol. 61, pp. 133–143, Dec. 2013.