

MULTI-TIER STORAGE MANAGEMENT AND APPLICATION OF REMOTE SENSING IMAGE DATA

Jin Liu¹, Lei Liu¹, Xuchao Xing^{1*}, Xinyan Zheng¹, Yang Gao¹, Quandong Xu², Jiang Du¹

¹National Geomatics Center of China, 28 Lianhuachi West Road, Haidian District, Beijing 100830, China - (liujin, liulei, xingxuchao, xinyan_zheng, gaoyang, dujiang)@ngcc.cn

²School of Geosciences and Info-Physics, Central South University, 932 Lushan South Road, Yuelu District, Changsha, Hunan, China - 15706403286@163.com

Commission III, WG III/IVb

KEY WORDS: Remote sensing image data, Storage management, Thermal evaluation model, Multi-tier migration, Dynamic flow, Local storage cluster, Cloud storage

ABSTRACT:

With the rapid development of the remote sensing platform and sensor technology, remote sensing image data presents the typical characteristics of the massive complex, multi-source heterogeneous, spatial-temporal intensive, which puts forward higher requirements for data storage management efficiency and real-time online service capability. Combined with the demand for remote sensing image data, the multi-tier migration strategy and approach based on the thermal evaluation model of remote sensing image data are proposed, considering the file size and data activity of remote sensing image data. The linkage between local storage cluster and cloud storage implements multi-tier migration and dynamic flow of remote sensing image data, improves the utilization of storage devices and the rationality of storage resource allocation, and enhances the capability of fast, dynamic, and real-time online service of remote sensing image data.

1. INTRODUCTION

With the rapid development of the remote sensing platform and sensor technology, the scale of high-precision, high-resolution, multi-temporal, and multi-spectral digital remote sensing images with real-time, all-weather, and large-area access to surface information is explosively increasing. The amount of data is often PB or even EB magnitude (Mei et al., 2001) (Hu et al., 2021). Taking the National Geographic Condition Census as an example, from 2015 to 2021, nearly 250,000 remote sensing images were used. They came from dozens of remote sensing satellites. The overall specification results were 4-band 16-bit orthophoto. The data amount of a single scene was 4GB -12GB, and the total data amount was more than 2.3PB. The remote sensing image data presents the typical characteristics of the massive complex, multi-source heterogeneous, spatial-temporal intensive.

The existing storage system in the data center can realize the hierarchical storage management of online, near-line, and offline remote sensing image data. However, it cannot solve the problem of multi-tier migration and the dynamic flow of remote sensing image data. In the face of such large and complex remote sensing image data, how to establish a reasonable and efficient, dynamic, and orderly multi-tier migration strategy based on the use of thermal and life cycle characteristics of remote sensing image data. To improve the efficiency of storage equipment, the rationality of resource allocation, the flexibility of storage space management. To lay the foundation for efficient management and fast, dynamic real-time online service of remote sensing image data. It is a difficult problem to be solved in the online storage management of remote sensing image data.

At present, papers mainly focus on the analysis and research of remote sensing image data organization, storage, and

management based on different methods. For examples, the multi-level grid model (Li et al., 2016), big data architecture (Hu et al., 2016), pyramid model (Yang et al., 2017), cloud data management (Yan, 2017), visualization management (Yu et al., 2017), object storage structure (Lu, 2019), metadata database, spatial database and image compression (Su, 2019), distributed storage (Jing and Tian, 2018) (Tian, 2019) and spatio-temporal data lake (Lu, 2021) were used to study the storage management of remote sensing image data. There were often aimed at online, near-line, and offline hierarchical storage and migration management (Lv et al., 2011) (Wu et al., 2014). However, On the basis of hierarchical storage, there are relatively few analyses and studies on further multi-tier management and dynamic migration of remote sensing image data in online resource pools. This paper proposes to automatically migrate infrequently used data or rarely accessed "frozen" data to a large-capacity and affordable object storage through a migration strategy to cope with rapid data growth, and optimize storage resources.

Therefore, in order to strengthen the utilization rate of data storage resources and the effectiveness, this paper carries out the thermal priority ordering of remote sensing image data and formulates the appropriate multi-tier migration strategy. With the help of the linkage between local storage cluster and cloud storage, the dynamic multi-tier storage of remote sensing image data and the flexible delivery of storage resources are realized. It improved the utilization rate of equipment and the rationality of storage resources configuration. It increased the flexibility of storage space management. It reduced the cost of storage equipment and resources. So as to improve the efficiency of fast, dynamic, and real-time online service of remote sensing image data.

* Corresponding author

2. STORAGE REQUIREMENTS OF REMOTE SENSING IMAGE DATA

Remote sensing image data not only have a large amount of data, but also grow rapidly due to the continuous imaging of remote sensing satellites at all times (Zhu et al., 2016) (Li and Huang, 2017). This undoubtedly brings new challenges to the capacity, performance, and cost performance of large-scale image storage systems.

(1) Large capacity: Large-scale image data storage system not only needs to have enough capacity, but also needs to have good horizontal expansion capability to cope with the explosive growth of data volume.

(2) High performance: The improvement of storage system performance is crucial for the whole fast, dynamic real-time online service system of remote sensing image data. With the increasing amount of remote sensing image data and users, the demand for better quality and more agile services of remote sensing image data is greater, and the pressure on the storage system is increasing. It is necessary to pay close attention to the response time and throughput of the storage system.

(3) High cost-effective: How to make full use of the advantages of different types of storage devices? Improving the efficiency of remote sensing image data service under the pursuit of lower unit capacity cost and lower management and maintenance cost (Zhao

et al., 2020). In the face of the terabyte-level growth of remote sensing image data, it is necessary to choose the appropriate storage technologies for differentiated needs.

The higher requirements for data read-write performance and storage capacity have been put forward by the fast, dynamic, and real-time online service of remote sensing images. The original online, near-line, and offline hierarchical storage management architecture has been unable to meet its requirements, in special the storage efficiency, efficient dynamic balance, and elastic share capacity of online storage resources.

3. MULTI-TIER STORAGE ARCHITECTURE OF REMOTE SENSING IMAGE DATA

In order to meet the demand for fast, dynamic real-time online service of remote sensing image data, a multi-tier data storage architecture integrating local storage cluster and cloud storage is designed and integrated. The local storage cluster is composed of the local data cache, the high-speed and low capacity solid-state disk and the large capacity mechanical disk. The cloud storage is composed of object storage. With the effective linkage of local storage cluster and cloud storage, it can meet the storage management requirements of different levels of remote sensing image data in a differentiated way. As well as it realizes dynamic multi-tier storage of remote sensing image data and flexible delivery of storage resources. As shown in Figure 1.

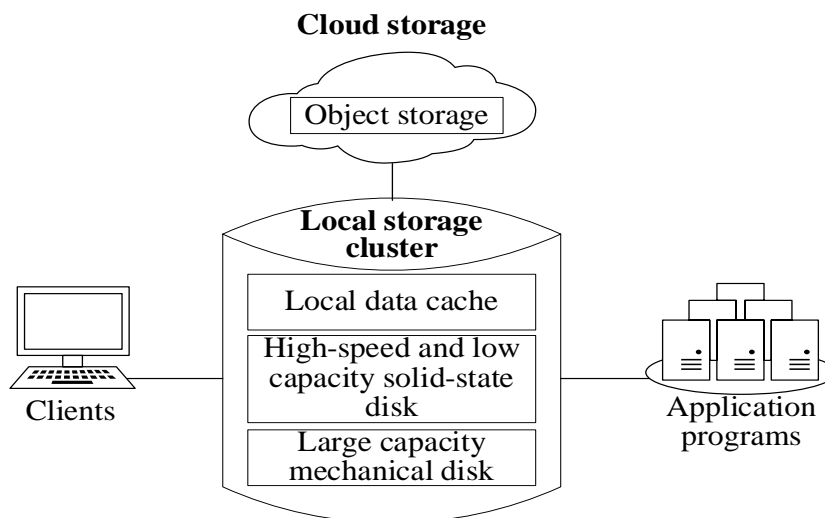


Figure 1. Multi-tier storage architecture of remote sensing image data.

The architecture can automatically match the storage resources with the read-write performance requirements of remote sensing image data, and always write the updated data into the local data cache. According to the specified business rules and data multi-tier migration strategy, periodically and automatically tiered storage of remote sensing image data. Without affecting data protection, application performance, or normal operation time, a large single file and the data that are not frequently accessed are tiered to the solid-state disk, mechanical disk, or object storage. It means that from high-performance storage devices to low-performance storage devices. This migration process is transparent to users and does not affect the user's access to the file, but the access efficiency will decrease. Files archive data from local storage cluster to cloud storage, the high-performance storage space will be available, which can provide services for other high-calorie data. The multi-tier architecture storage realized storage resource optimization and storage cost reduction.

4. MULTI-TIER MIGRATION STORAGE STRATEGY AND APPROACH

According to the characteristics of remote sensing image data and the requirements of real-time online service, considering the file size and data activity of remote sensing image data, the thermal priority ordering of remote sensing image data is carried out. The migration function of tiered storage is used to organize and manage data migration on different levels of storage, so as to improve the utilization rate of storage resources and the efficiency of equipment utilization. As well, it provides strong storage support for the grid scheduling technology of remote sensing image data and improves the fast, dynamic, and real-time online serviceability of remote sensing image data.

Thermal evaluation model of remote sensing image data:

$$TE(X) = (DF = 0) \cap (FS \geq 10MB) \quad (1)$$

where X = a single remote sensing image data file
 DF = the number of updates of a single remote sensing image data file within a specified time period
 FS = actual size of storage space for a single remote sensing image data file

Note: The measurement parameter of 10MB can be adjusted according to the actual situation.

In the specified time period, when the above two conditions are met at the same time, it is considered that the remote sensing image data file heat is low, and the data should be archived from the local storage cluster to the cloud storage, otherwise stored on the local storage cluster.

According to the evaluation results of the thermal evaluation model for remote sensing image data, the multi-tier migration strategy of data is set based on the combination of file matching conditions in the multi-tier storage, and multiple file pools are identified, protected, and controlled. The specific storage location of each remote sensing image data on the local storage cluster is defined, so as to multi-tier management and dynamic flow.

The combination of file matching conditions can include file name, file path, file system object type (conventional file, directory, and other), custom attributes, the time of last modification of the file, the time of last access to the file, the time of last modification of file metadata, the time of file creation and file size. At the same time, any number of file matching conditions can be added to refine the multi-tier migration strategy.

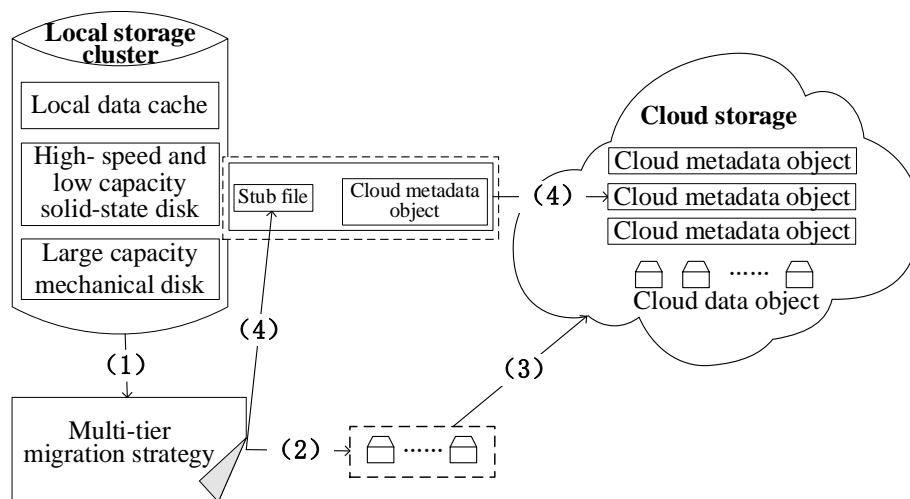


Figure 2. Diagram of data archive operation workflow.

When remote sensing image data is migrated to cloud storage, the files remain visible on the local storage cluster. After file data has been archived in the cloud storage, the file is truncated to an 8KB file, and the information about where to retrieve data is retained on the local storage cluster. The 8KB file is called a stub file. Each stub file contains cache information, cache data, and mapping information. The cache information is used to record the attributes of the stub file, the cached data is used to retain part of the file data locally, and the mapping information is used to point to the object storage. As shown in figure 3.

When the remote sensing image data file is matched with the data multi-tier migration strategy, it can be tier migrated between the local storage cluster and cloud storage. The tiered storage method based on the thermal evaluation model of remote sensing image data is as follows:

4.1 Data archive operation workflow

Data archive operation is a process of migrating remote sensing image data files from the local storage cluster to cloud storage. This process involves extracting data from files and placing them in one or more cloud data objects, moving objects to cloud storage, and retaining representative stub files on the local storage cluster. More workflow details include the following, as shown in figure 2.

- (1) Based on the use of storage space and the thermal analysis results of remote sensing image data, the data multi-tier migration strategy is set.
- (2) Automatically select remote sensing image data files that match the multi-tier migration strategy. The migrated remote sensing image data files file data is split into chunks of cloud data objects.
- (3) The chunks are sent from the local storage cluster to cloud storage, and a checksum is applied for each chunk to ensure data integrity.
- (4) The migrated remote sensing image data files are truncated into a stub file stored on the local storage cluster. The cloud metadata objects are written into the cloud storage.

The data stored in the cloud is called cloud objects, including cloud data objects and cloud metadata objects. Cloud data object refers to the remote sensing image data that need to be stored, which are split into chunks for storage. The chunk is a 128MB logical container of contiguous space. Chunks are written in an append-only pattern. The append-only behavior means that an application's request to modify or update an existing object will not modify or delete the previously written data within a chunk, but rather the new modifications or updates will be written in a

new chunk. Therefore, no locking is required for I/O and no cache invalidation is required. Cloud metadata object refers to the system metadata and user-defined metadata. The system

metadata can be divided into identifiers and descriptors, encryption keys, internal tags, location information, timestamp, and configuration/tenants information. As shown in figure 3.

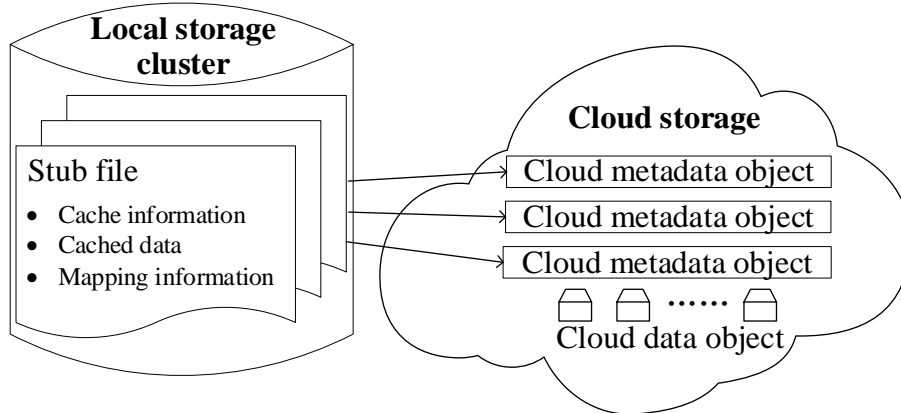


Figure 3. Diagram of stub file and cloud object.

4.2 Data recall operation workflow

Data recall operation is the process of remote sensing image data from cloud storage to the local storage cluster, replacing stub files, and deleting cloud objects from cloud storage. More workflow details include the following, as shown in figure 4.

(1) Recalling remote sensing image cloud data objects from cloud storage to the local storage cluster.

(2) Restoring the remote sensing image data file on the local storage cluster, and replacing the corresponding stub files with the original file data.

(3) According to the set retention period, if it expired, the corresponding cloud objects that have been recalled in the cloud storage are deleted asynchronously.

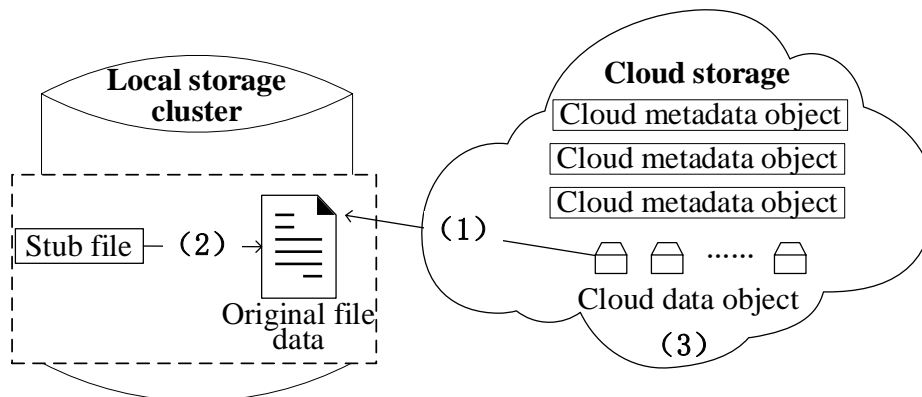


Figure 4. Diagram of data recall operation workflow.

4.3 Data read and update operation workflow

Users and applications can seamlessly access data at any physical location on the local storage cluster and cloud storage through the same network path and protocol. Firstly, the remote sensing image data is searched from the local data cache. If the accessed data is not in the local data cache, it is searched from the high-speed and low capacity solid-state disk. If the access data are not in the high-speed and low capacity solid-state disk, they are searched from the large capacity mechanical disk. If the accessed data are not in the large capacity mechanical disk, they are searched from the object storage stored in the cloud storage.

Users can directly read and update remote sensing image data stored on the local storage cluster. Transparent access to remote sensing image data stored in the cloud storage, the actual access for the user through the stub file on the local storage cluster. When a client opens a file for reading, the chunk will be added to the cache in the associated stub file by default, which can view and edit remote sensing image data files normally. When users

modify and save files, they will save the changed contents in the local cache. As well, it periodically scans the stub files to determine whether there are pending data changes, writes them to the corresponding cloud objects in cloud storage, so as to always maintain the latest version.

Reading the remote sensing image data in cloud storage, more workflow details include the following, as shown in figure 5.

(1) Users access the file through the stub file on the local storage cluster and find the file data in the local data cache.

(2) If the file data has been in the local data cache, file data is directly sent to the client from the local cache on the local cluster.

(3) If it does not exist, retrieve the cloud data objects from cloud storage, but the file data in the local cache on the local storage cluster, and then sent it to the client.

(4) Clearing the expired cache information of the stub file, the space used by the cache is temporary and configurable.

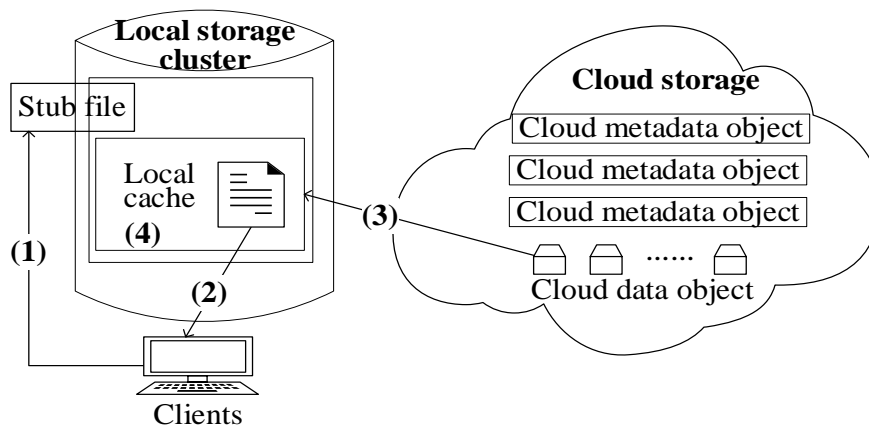


Figure 5. Diagram of data read operation workflow.

Updating the remote sensing image data in cloud storage, more workflow details include the following, as shown in figure 6.

(1) When the user updates the file through the client and those changes are stored in the local cache first.

(2) Periodically sends the updated file data to the cloud storage from the local data cache.

(3) Clearing the expired cache information of the stub file, the space used by the cache is temporary and configurable.

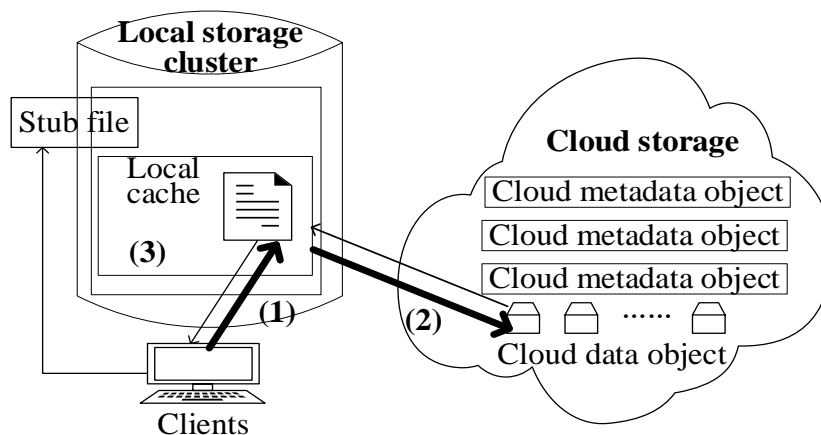


Figure 6. Diagram of data update operation workflow.

5. MULTI-TIER MIGRATION INSTANCE

At present, the data center integrates and builds the local storage cluster of 18 nodes and the cloud storage of 5 nodes. The available space of the local storage cluster is more than 2PB, the access speed of a single node is 1GB/s, and the available space of cloud storage is more than 4PB. Nginx service forwarding is used to realize the linkage between the local storage cluster and cloud storage, and the data settlement and promotion speed can reach 48TB per day. According to statistics, there are 1.81 billion cloud data objects migrating from the local storage cluster to the cloud storage in the existing remote sensing image data. The response time of the 50th object in every 100 cloud data objects is not more than 4ms, and the response time of the 99th object in every 100 cloud data objects is not more than 110ms.

In order to ensure the real-time computing service of national remote sensing image data, the remote sensing image data of each issue need to be stored in the online local storage group, and the storage space of high-speed and low capacity solid-disk or large-capacity mechanical disk is up to 300-400TB. Nowadays, the remote sensing image data migration application is carried out based on the thermal evaluation model. The migration strategy is

set according to the number of updates and the actual size of the storage space of a single remote sensing image data file within a specified time period. The remote sensing image data files with low heat are automatically migrated from the local storage cluster to the cloud storage. After calculation, this method only needs to store the hot data of 100-130TB remote sensing images in each issue in the online local storage group, such as the data of the eastern coastal area, developed areas, relatively new time phase, and new large infrastructure. The remote sensing image data with a long time and remote area are migrated to the cloud storage, saving more than 80% of the online local storage space. At the same time, the real-time computing service performance of hot data remains unchanged, and the access time reaches 200ms per tile in parallel. Non-hot data can also meet the real-time computing service performance requirements, access time up to 500ms per tile in parallel, better guarantee the quality of image service. At present, the remote sensing image results of the National Geographic Conditions Census from 2016 to 2021 have all realized hierarchical storage management, which improves the utilization efficiency of online storage resources in the data center.

6. CONCLUSION

Facing the demand for fast, dynamic real-time online service of remote sensing image data, with the help of the local storage cluster and cloud storage linkage, based on the multi-tier storage method of remote sensing image data thermal evaluation model, the dynamic tiered storage of remote sensing image data and the elastic delivery of storage resources are realized, which provides strong storage support for the grid scheduling technology of remote sensing image data, and improves the ability of remote sensing image data fast, dynamic real-time online service.

With the deepening of the application of deep learning, the deep learning classification algorithm can be used in the next research. On the basis of considering the characteristics of remote sensing image data and production characteristics, the thermal evaluation effect can be further improved and the real-time and dynamic management efficiency of remote sensing image data storage can be improved.

REFERENCES

- Hu X D., Zhang X., Qu J S., 2016. Resource storage and management method of massive remote sensing data supported by the big data architecture. *Journal of Geoinformation Science*, 18(5), 681-689.
- Hu S S., Huang Y., Huang C X., et al., 2021. Development Status and Future Prospect of Multi-source Remote Sensing Image Collaborative Application. *Radio Engineering*, 51(12), 1425-1433.
- Jing W P., Tian D X., 2018. An improved distributed storage and query for remote sensing data. *Procedia Computer Science*, 129, 238-247.
- Lv X F., Cheng C Q., Gong J Y., et al., 2011. Review of data storage and management technologies for massive remote sensing data. *Sci China Tech Sci*, 54, 3220-3232.
- Li S., Cheng C Q., Tong X C., et al., 2016. A Study on Data Storage and Management for Massive Remote Sensing Data Based on. *Acta Geodaetica et Cartographica Sinica*, 45(S1), 106-114.
- Li G Q., Huang Z C., 2017. Data Infrastructure for Remote Sensing Big Data: Integration, Management and On-Demand Service. *Journal of Computer Research and Development*, 54(02), 267-283.
- Lu Y., 2019. Research on Massive Image Tile Storage Method Based on Object Storage Structure. Zhejiang University, Zhejiang, China.
- Lu S Q., 2021. Research and application of spatio-temporal data lake. *Land and Resources Informatization*, (03), 28-33+53.
- Mei A X., Peng W L., Qin Q M., et al., 2001: *Introduction to Remote Sensing*. Higher Education Press, Beijing, China.
- Su Y P., 2019. Practical Analyses on Mass Remote Sensing Image Data Management. *ZHEJIANG ARCHIVES*, (05), 60-61.
- Tian D X., 2019. Research on Massive Remote Sensing Image Classification Method Based on Distributed Storage. Northeast Forestry University, Heilongjiang, China.
- Wu H Q., ZHANG J B., He W., et al., 2014. Studies on Hierarchical Storage and Migration Rules for Multi-Source RS Image Data. *Land and Resources Informatization*, (03), 21-26.
- Yan J N., 2017. Research on the Key Techniques of Cloud-based Remote Sensing Data Management and Production across Multi-data Centers. University of Chinese Academy of Sciences, Beijing, China.
- Yang M., Mei H., Yang Y., et al., 2017. Efficient storage method for massive remote sensing image via spark-based pyramid model. *International Journal of Innovative Computing Information & Control*, 13(6), 1915-1928.
- Yu H J., Li S Y., Zhang T., 2017. Study on the Organization and Storage of Remote Sensing Data Oriented to Visualization Management. *Information Science and Cloud Computing(Volume 1).Ed.*, 2017, 209-217.
- Zhao Z Y., Yang D., Liang X H., et al., 2020. Study on Storage and Migration Strategy of Massive Remote Sensing Data. *Journal of Information Engineering University*, 21(01), 115-119.
- Zhu J Z., Shi Q., Chen F E., et al., 2016. Research status and development trends of remote sensing big data. *Journal of Image and Graphics*, 21(11), 1425-1439.