Development model of remote sensing data management for long-term storage using big data

Riyan Mahendra Saputra[1], Agnes Sondita Payani[2], Ogi Gumelar[2]
[1]Directorate for Policy Formulation of Research, Technology and Innovation, National Research and Innovation Agency (BRIN), Jakarta, Indonesia
[2]Research Center for Geoinformatics, National Research and Innovation Agency (BRIN), Bogor, Indonesia

Article Info

Article history:

Received Oct 30, 2025
Revised Jul 6, 2026
Accepted Aug 28, 2026

Keywords:

Big data
Data distribution
Data management
Hierarchical storage management
Satellite

ABSTRACT

The growing volume of remote sensing data requires efficient approaches for long-term storage and data management. The National Remote Sensing Data Bank (BDPJN) stores multi-resolution satellite imagery acquired since 1993 to support disaster monitoring, climate studies, natural resource management, and other Earth observation applications. However, managing continuously growing archives remains challenging due to manual migration processes, inefficient storage utilization, and limited scalability. This study proposes an integrated remote sensing data management model that combines hierarchical storage management (HSM) with Apache Spark-based parallel processing to support policy-driven data migration and long-term storage optimization. The proposed model enables data migration across multiple storage tiers based on data characteristics and storage policies while improving migration efficiency through distributed processing. Experimental evaluation was conducted using Landsat-8, SPOT-6, SPOT-7, and Pleiades datasets under both 1 Gbps and 10 Gbps network environments. The results demonstrate migration performance improvements ranging from 80.14% to 91.04% compared with the existing manual migration system. These findings indicate that integrating HSM with Spark-based parallel processing can reduce migration bottlenecks, improve storage utilization, and provide a scalable and cost-effective approach for managing large-scale remote sensing archives while supporting long-term data governance and accessibility.

Corresponding Author:

Riyan Mahendra Saputra
Directorate for Policy Formulation of Research, Technology and Innovation
National Research and Innovation Agency (BRIN)
B.J. Habibie Building, 5th Floor, Jl. M.H. Thamrin No. 8, Central Jakarta, 10340, Indonesia
Email: riyan.mahendra.saputra@brin.go.id

1. INTRODUCTION

Remote sensing data have become crucial for long-term environmental monitoring, providing continuous and historical observations of the Earth over decades. Long-term satellite records enable quantitative documentation of environmental change that would be impossible with short-term or ground-based observations alone. Remote sensing offers continuous, wide-area coverage through satellites and other sensors, significantly improving our ability to observe climate variability and ecological changes at global and regional scales over time [1]. Effective long-term monitoring depends not only on data availability but also on the preservation of historical archives, data integrity, and efficient data management and access mechanisms. These extensive time-series datasets form an indispensable foundation for detecting trends, validating models, and devising policies in domains such as climate change adaptation, natural resource management, and disaster monitoring.

As remote sensing archives continue to expand, their long-term value increasingly depends on the ability to manage and preserve large volumes of heterogeneous data in a reliable manner. Modern Earth observation missions generate massive data streams with high volume, velocity, and variety, driven by multi-sensor platforms and increasing spatial, spectral, and temporal resolutions [2]. These characteristics pose significant challenges for data storage, access, and long-term preservation, particularly when datasets remain usable for future scientific analysis and policy formulation. Without appropriate long-term storage strategies, the sustainability and reusability of remote sensing big data would be severely compromised [3]. In operational environments, these challenges are often reflected in manual data migration, inefficient storage utilization, and limited scalability of existing storage workflows.

Cloud-based architectures and data lifecycle management frameworks have been explored to support scalable access and sustained usability of large and heterogeneous remote sensing datasets [4]. To address these challenges, in large-scale data repositories and data-intensive computing environments, hierarchical storage approaches are used to manage data placement across multiple storage tiers in order to balance performance and storage efficiency. Hierarchical storage management (HSM) enables data to be distributed across different storage tiers according to access patterns and workload characteristics, allowing frequently accessed data to reside on high-performance storage while less active data are placed on lower-cost tiers [5]. Recent studies further demonstrate that automated and intelligent tiered storage mechanisms can dynamically adapt to changing access patterns and significantly improve data processing performance and system efficiency in distributed platforms such as Hadoop and Spark-based infrastructures, as well as in extreme-scale scientific computing environments [5], [6].

Big data technologies provide scalable infrastructures for storing, managing, and processing the rapidly growing volumes of remote sensing data. In contrast to traditional centralized systems, big data platforms are designed to handle high data volume, velocity, and variety through distributed storage and parallel processing mechanisms. Distributed computing frameworks such as Apache Hadoop and Apache Spark enable parallel data processing and efficient handling of massive geospatial datasets by executing computations across multiple nodes, thereby improving throughput and reducing processing time for data-intensive workloads [7], [8]. These capabilities make big data technologies particularly suitable for large-scale remote sensing archives, where data ingestion, migration, and access operations must be performed efficiently over continuously expanding datasets. Furthermore, integrating distributed processing with storage systems enables advanced data management strategies, such as parallel data migration and policy-driven data placement. These strategies are essential for improving the scalability and operational efficiency of large and continuously growing remote sensing data repositories.

Existing studies on remote sensing big data management have investigated various approaches, including parallel data processing frameworks, distributed big data processing, and hierarchical storage architectures [7]-[10]. These works emphasize the importance of scalable computation, efficient data lifecycle management, and long-term storage strategies for large remote sensing archives. However, most existing approaches address these aspects in isolation, focusing either on accelerating large-scale data processing or on structuring storage hierarchies, without providing an integrated operational framework that jointly considers HSM and parallel migration mechanisms. Consequently, the lack of unified models limits the effectiveness of large-scale archive management in data-intensive remote sensing environments.

To further clarify the novelty of this study, Table 1 compares representative prior works with the proposed model. The reviewed studies have contributed to different aspects of remote sensing big data management. However, within the reviewed literature, these approaches generally address storage management, data processing, or migration performance as separate concerns. Limited attention has been given to an operational model that combines HSM-based storage tiering, Spark-based parallel migration, and policy driven migration rules for long-term national remote sensing archives. This study addresses this gap by proposing and evaluating an integrated HSM–Spark architecture for scalable long-term management of remote sensing archives.

This limitation becomes more evident in the context of Indonesia’s National Remote Sensing Data Bank (BDPJN), which manages multi-resolution satellite imagery to support national research and policy needs. Currently, data migration across storage media is largely performed manually using age-based rules, which are insufficient to capture diverse data priorities and access patterns. As a result, the existing system lacks automated, policy-driven migration and parallel processing capabilities, constraining its scalability and operational efficiency for managing continuously growing remote sensing archives. Although recent studies have proposed governance and data quality management frameworks for BDPJN [11], issues related to large-scale data migration performance and storage optimization remain insufficiently addressed. These gaps indicate the need for an integrated data management model that combines HSM with parallel processing to support efficient, scalable, and policy-driven migration of large remote sensing archives.

In addition to the core technologies discussed above, recent developments further emphasize the need for integrated storage and processing frameworks. Cloud-based platforms, multi-source data integration, and machine learning-driven analysis have been widely adopted to support environmental monitoring, land use analysis, and decision-making processes [12]-[18]. In parallel, recent studies have highlighted the role of strategic transformation and sustainability in improving the adaptability and efficiency of organizational systems [19]. In the storage domain, HSM has evolved with adaptive and intelligent mechanisms, including reinforcement learning and workload-aware data placement, to optimize storage efficiency and performance in large-scale systems [20], [21]. Recent advancements also explore workload-aware architectures and domain-specific implementations of hierarchical storage, such as block-based management frameworks, large-scale scientific data handling, and hierarchical access control models in distributed storage environments [22], [23]. These developments further confirm the need for integrated frameworks that combine storage management and parallel processing capabilities to support long-term and scalable remote sensing data management.

The main contributions of this study are threefold. First, this study proposes an integrated data management model that combines HSM with big data technologies to support scalable and efficient handling of large remote sensing data archives. Second, it introduces policy-driven data migration rules based on multiple data attributes, including data age, file class, file size, priority, and storage capacity thresholds, which are executed within a Spark-based parallel processing environment. Third, this study presents an end-to-end performance evaluation of the proposed framework, covering data migration, data distribution, and data access scenarios under both 1 Gbps and 10 Gbps network infrastructures, and reports performance improvements compared to the existing system. These contributions are expected to support more scalable, efficient, and sustainable management of national remote sensing archives. They are particularly relevant for climate monitoring, disaster management, natural resource monitoring, and long-term scientific research.

Table 1. Related works on remote sensing big data management and archives

Study	Focus	Method / technology	Key findings	Limitations
Chi *et al*. [7]	Big data for remote sensing	State-of-the-art review	Identified challenges in storage, migration, and visualization	Conceptual; no implementation of integrated framework
Chebbi *et al*. [8]	Big data processing for remote sensing	Comparison: Hadoop MapReduce vs. Spark	Spark more efficient for iterative remote sensing tasks	Focused on computation; did not integrate with storage management
Lan et al. [9]	Large-scale remote sensing data processing	Spark-based parallel in-memory processing on cloud platforms	Processed ~107.4 GB of Landsat 8 data in 1,018 s, demonstrating scalable processing of large remote sensing datasets	Storage lifecycle management and governance were outside the scope of the study
Wu and Fu [10]	Remote sensing data management	Metadata-driven file and database management system (DBMS)	Improved retrieval and cataloging of archives	Did not address migration speed or scalable tiered storage
Lee and Kang [24]	Geospatial big data architecture	Three-layer design (data–process–service)	Provided architectural model for integration and analytics	High-level; lacks validation in migration and long-term storage
Huang *et al*. [25]	Cloud-based remote sensing big data management	Cloud computing architecture	Demonstrated the ability of cloud computing to support scalable storage and processing of large satellite imagery datasets	Did not specifically integrate HSM-based tiering with Spark-based parallel migration policies

2. METHOD

This study develops a remote sensing data management model by integrating HSM with a big data processing framework. The research method consists of four main stages: analysis of the existing BDPJN storage workflow, development of the proposed HSM–Spark architecture, formulation of migration policies, and experimental evaluation. The experimental equipment used in this study had the following specifications: (i) one master server with 64 cores and 64 GB of memory, (ii) worker servers with 32 cores and 1 GB of memory, (iii) main storage with a capacity of 1 petabyte, (iv) secondary storage with a capacity of 750 terabytes, (v) network configurations of 10 Gbps and 1 Gbps, and (vi) a database server using PostgreSQL.

2.1. Existing BDPJN storage workflow

The BDPJN system is developed to support activities ranging from data acquisition, processing, storage, distribution, and dissemination of remote sensing data and information. The implementation of BDPJN refers to Indonesian Space Law No. 21 of 2013, which regulates national remote sensing data management. Figure 1 illustrates the Level-1 data flow of the BDPJN system. The system consists of several main entities that interact within the remote sensing data management workflow. Users represent groups of remote sensing image users who can search and obtain data through the dissemination system. The data acquisition system is responsible for obtaining remote sensing imagery through acquisition, data procurement, and downloading from international ground stations. The data management system regulates data storage, including registration, migration, archiving, and data access. The data processing system provides value-added products from the acquired data. The data dissemination system distributes data and information through the data catalog subsystem. The database acts as the central repository that records activities within the BDPJN system. There are three approaches to remote sensing image storage management: file-based storage management, geographical database-based management, and a hybrid approach combining file and database management [10].

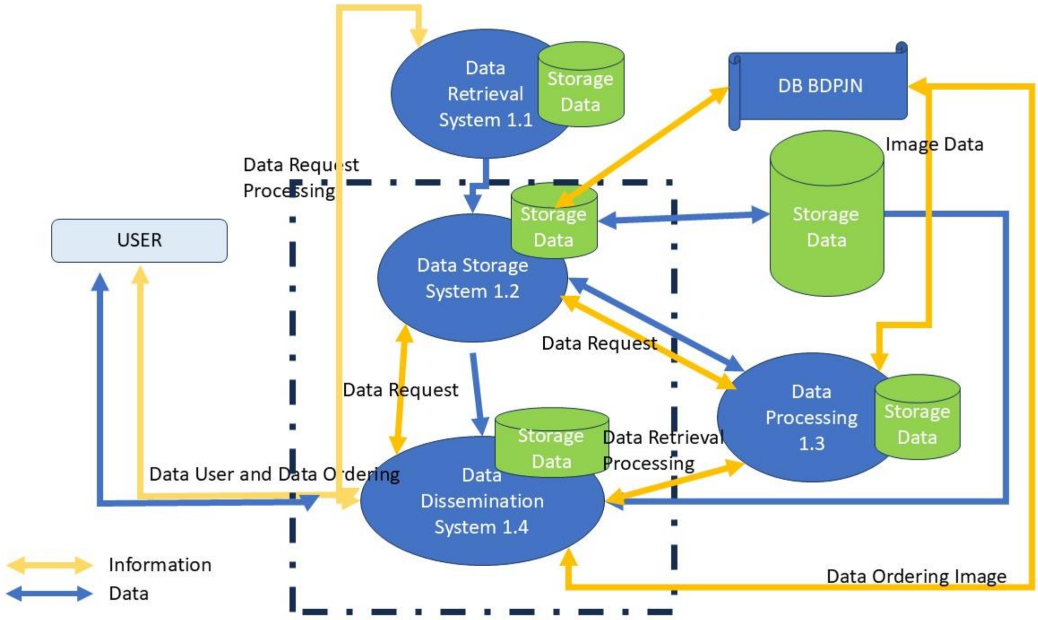


Figure 1. BDPJN system level 1 data flow chart

2.2. Proposed HSM-SPARK integrated architecture

To address the limitations of the existing BDPJN workflow, this study proposes an integrated HSM–Spark architecture to enable scalable remote sensing data migration and policy-driven storage management. Big data geospatial systems are generally organized into three layers: geospatial big data integration and management, geospatial big data analytics, and geospatial big data service platforms. The first layer functions for storing, retrieving, indexing, and searching geospatial data. The second layer serves to interactively analyze geospatial data in real-time and dynamic data. The third layer functions as a service platform that handles user interaction with big data systems [24]. Figure 2 shows the overall system design of a big data geospatial architecture. Referring to this structure, the proposed architecture adapts and modifies the model to suit the operational characteristics of BDPJN. In this context, Apache Spark is used as a distributed processing framework to enable parallel data migration, while the Hadoop distributed file system (HDFS) functions as the underlying storage system for managing large-scale datasets. Metadata is organized through catalog-based mechanisms to support efficient data indexing and retrieval.

Hadoop-based big data frameworks are commonly implemented using a master–slave architecture, where the master node coordinates system operations while worker nodes execute distributed processing tasks [24]. In remote sensing environments, big data architectures are often organized into three main components: the remote sensing data acquisition unit (RSDU), data processing unit (DPU), and data analysis and decision unit (DADU). The RSDU is responsible for acquiring remote sensing data from ground stations. The DPU performs data filtering, load balancing, and preprocessing tasks before storing the processed data. The DADU aggregates processing results and distributes them to various applications for further analysis and decision-making processes [26].

Referring to the models described above, this study proposes a modified geospatial big data model for remote sensing data management within the BDPJN environment, as shown in Figure 3. The proposed model consists of three layers:

– Remote sensing big data management layer: this layer manages data storage, metadata registration, indexing, and data transformation. It coordinates storage tier allocation and prepares datasets for distributed processing. Data are stored across multi-tier media, including performance storage (Tier-1), capacity storage (Tier-2), and archive storage (Tier-3).
– Big data analysis and processing layer: this layer implements distributed and parallel processing using Apache Spark. The Spark framework operates in a master–worker configuration, where the master node coordinates job scheduling and resource allocation, while worker nodes execute parallel migration tasks. This configuration enables concurrent data transfer operations across storage tiers, reducing bottlenecks during large-scale migration.
– Service and access layer: this layer provides application modules and application programming interface (API)-based interfaces that allow users and system administrators to perform data registration, migration, monitoring, and access control operations.

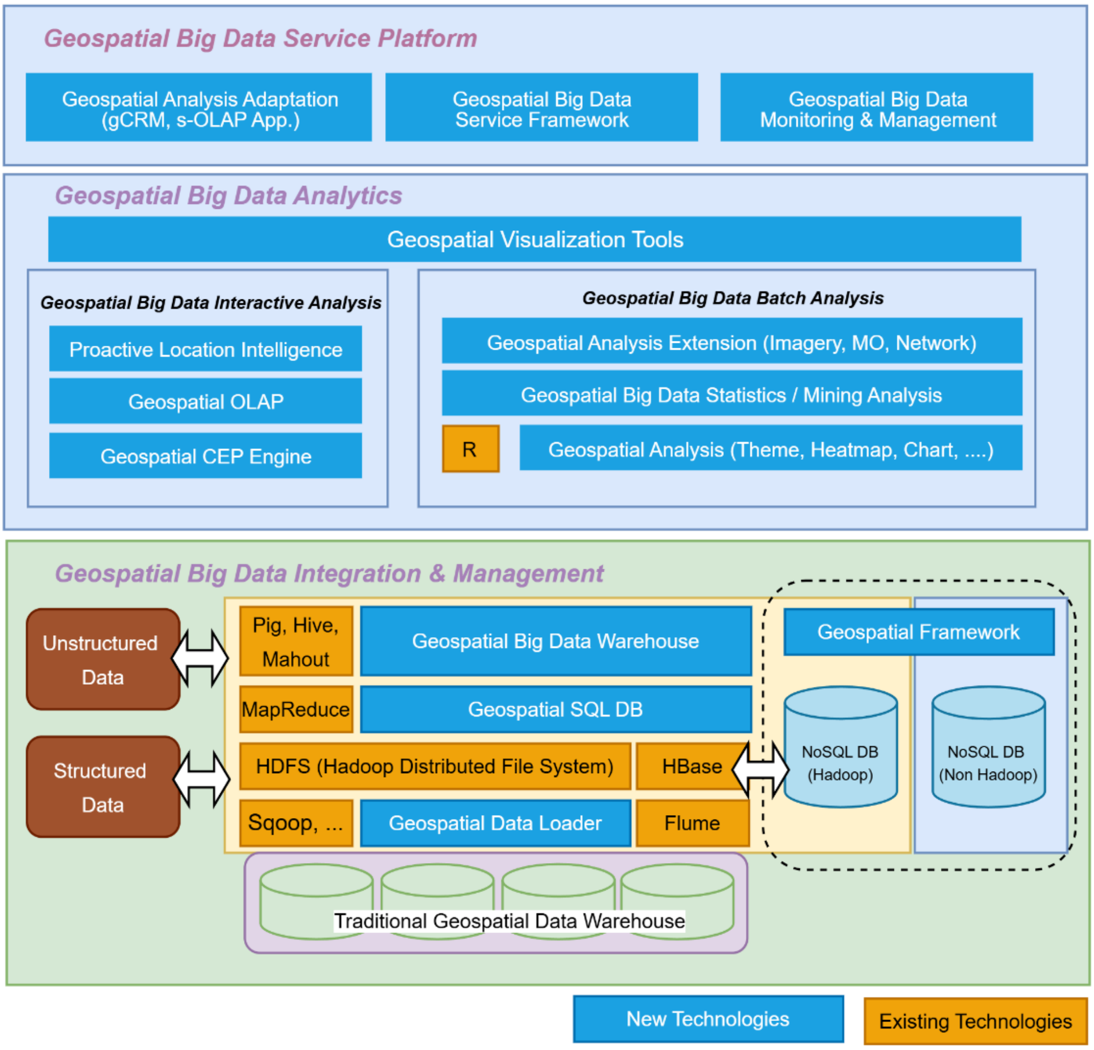


Figure 2. Big data geospatial architecture [24]

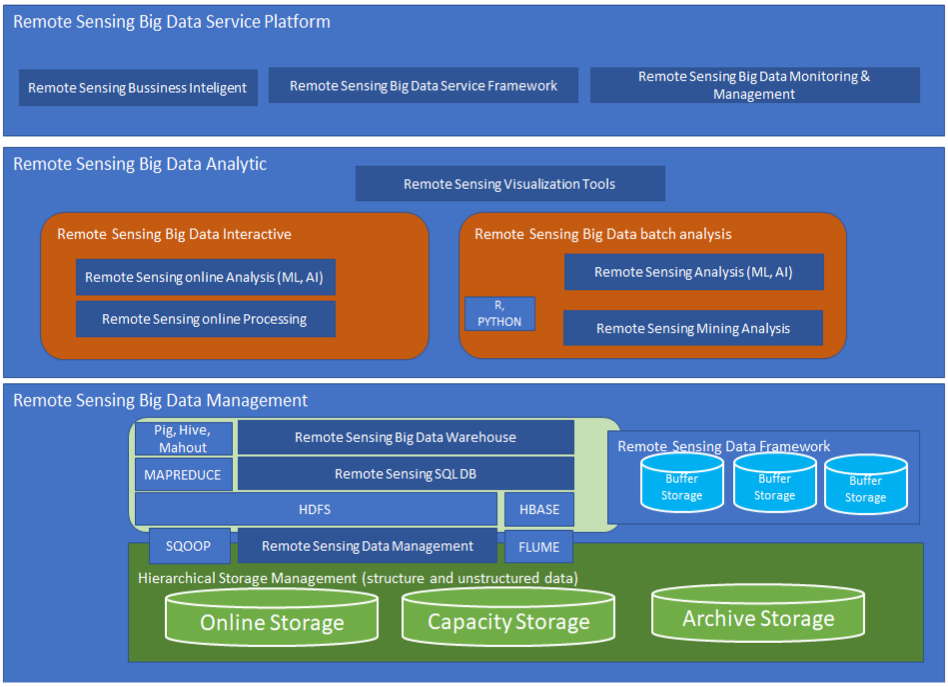


Figure 3. Architecture of the proposed remote sensing data management model

2.3. Migration policy formulation

The migration policy in this study is designed to manage remote sensing data movement across multiple storage tiers. Data stored in the BDPJN system consist of remote sensing imagery obtained from ground station acquisition, processed data products, and procurement data. The data are classified into two categories, namely hot data and cold data. Hot data represent frequently accessed datasets and are typically stored in Tier-1 storage. Cold data represent less frequently accessed datasets and can be migrated to lower storage tiers. Several parameters are used to determine migration priority, including data age, data type, and storage capacity conditions. Data age is used to determine how long data remain in higher-performance storage before being migrated to lower tiers. Migration is triggered when storage capacity reaches a predefined threshold. These parameters enable the system to determine candidate datasets for migration and support efficient storage utilization in large-scale remote sensing data environments.

2.4. Experimental setup

The image data used in this study consist of remote sensing imagery obtained from multiple satellite sources. In total, the BDPJN system manages imagery from 22 satellite platforms. However, migration testing in this study is limited to Landsat 8, SPOT 6, SPOT 7, and Pleiades datasets, which represent actively acquired satellite imagery. These datasets were selected because they represent satellite imagery that is still actively acquired in the BDPJN operational environment and have different spatial resolutions and file size characteristics. Landsat 8 represents medium-resolution imagery, SPOT 6 and SPOT 7 represent high-resolution imagery, while Pleiades represents very high-resolution imagery with relatively larger file sizes. The remaining satellite datasets managed by BDPJN are primarily archive data that are no longer routinely acquired and therefore were not selected for the migration performance experiment.

For the migration experiment, datasets were prepared in several sizes consisting of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 image files. The storage infrastructure used in the experiment consists of three storage media tiers. The first storage tier has a capacity of 1 PB, the second tier has a capacity of 750 TB, and the third tier has a capacity of 200 TB. The experimental environment consists of one master server, two worker nodes, one database server, and one application module server. The network configuration supports bandwidths of 1 Gbps and 10 Gbps for performance evaluation.

3. RESULTS AND DISCUSSION

This section presents experimental results and evaluates the performance of the proposed remote sensing data management model. The evaluation focuses on data migration parameters, system implementation, and performance under different network environments (1 Gbps and 10 Gbps). These results demonstrate that the integration of HSM and parallel processing overall improves data migration efficiency and system performance compared to the existing system.

3.1. Analysis of migration parameters

The parameters used to determine data migration include class, age, size, and capacity threshold. These parameters are applied to control how data are distributed across storage tiers in the existing storage environment. The current storage configuration consists of unified storage and storage virtualization with a capacity of 1 PB, where 750 TB is allocated for Tier 1 and Tier 2, and a tape library with a capacity of 200 TB is used as Tier 3. The data stored includes remote sensing imagery from acquisition, processing, and procurement activities. Data are classified into two categories, namely hot and cold. Hot data represent actively acquired imagery that is primarily stored in Tier 1, while under certain conditions they may be moved to Tier 2. Cold data are typically migrated to lower storage tiers based on predefined policies.

Table 2 presents the image data characteristics and corresponding migration parameters used in this study. This table is important because it illustrates how different data types are assigned migration priorities based on their resolution, size, and storage policies. As shown in Table 2, higher-resolution and larger-size datasets tend to have higher migration priorities, indicating their impact on storage management and system performance. The use of multiple migration parameters provides a more structured approach compared to the existing condition, which relies mainly on data age. This enables a more accurate representation of data characteristics in determining migration decisions. In addition, the capacity threshold is defined as 20% of total storage capacity, combining a 10% hardware alert and 10% reserved capacity. This mechanism ensures that only 80% of storage is actively used, allowing early migration to prevent storage overload. This threshold-based mechanism improves system reliability by enabling proactive migration before storage capacity reaches critical limits.

Table 2. Image data and migration parameters

No	Resolution	Satellite	Data class	Data age (year)	Temporary (days)	File size	Migration priority
01	Low resolution	NOAA	H	0-3	100	300 MB	6
02	Low resolution	MODIS TERRA	H	0-3	100	5,5 GB	7
03	Low resolution	MODIS AQUA	H	0-3	100	5,5 GB	8
04	Low resolution	NPP-NPOESS	H	0-3	100	10 GB	9
05	Medium resolution	LANDSAT 5	C	0-4	100	350MB	4
06	Medium resolution	LANDSAT 7	C	0-4	100	560MB	5
07	Medium resolution	LANDSAT 8	H	0-4	100	780 MB	10
08	Medium resolution	ALOS (AVNIR, PRISM)	C	0-4	100	280MB	3
09	Medium resolution	SPOT 2	C	0-4	100	60 MB	1
10	Medium resolution	SPOT 4	C	0-4	100	60 MB	2
11	High resolution	SPOT 5	H	0-5	100	240 MB	11
12	High resolution	SPOT 6	H	0-5	100	300 MB	12
13	High resolution	SPOT 7	H	0-5	100	600 MB	13
14	Very high resolution	IKONOS	H	0-5	100	600 MB	14
15	Very high resolution	QUICKBIRD	H	0-5	100	1,5 GB	16
16	Very high resolution	PLEIADES 1A/1B	H	0-5	100	2 GB	17
17	Very high resolution	WORLDVIEW 1/2/3	H	0-5	100	4 GB	18
18	Very high resolution	GEO EYE	H	0-5	100	4 GB	19
19	Very high resolution	RAPID EYE	H	0-5	100	600 MB	15
20	Radar	RADARSAT	H	0-5	100	410 MB	20
21	Radar	TERRASAR-X	H	0-5	100	1,1 GB	22
22	Radar	ALOS PALSAR	H	0-5	100	750 MB	21

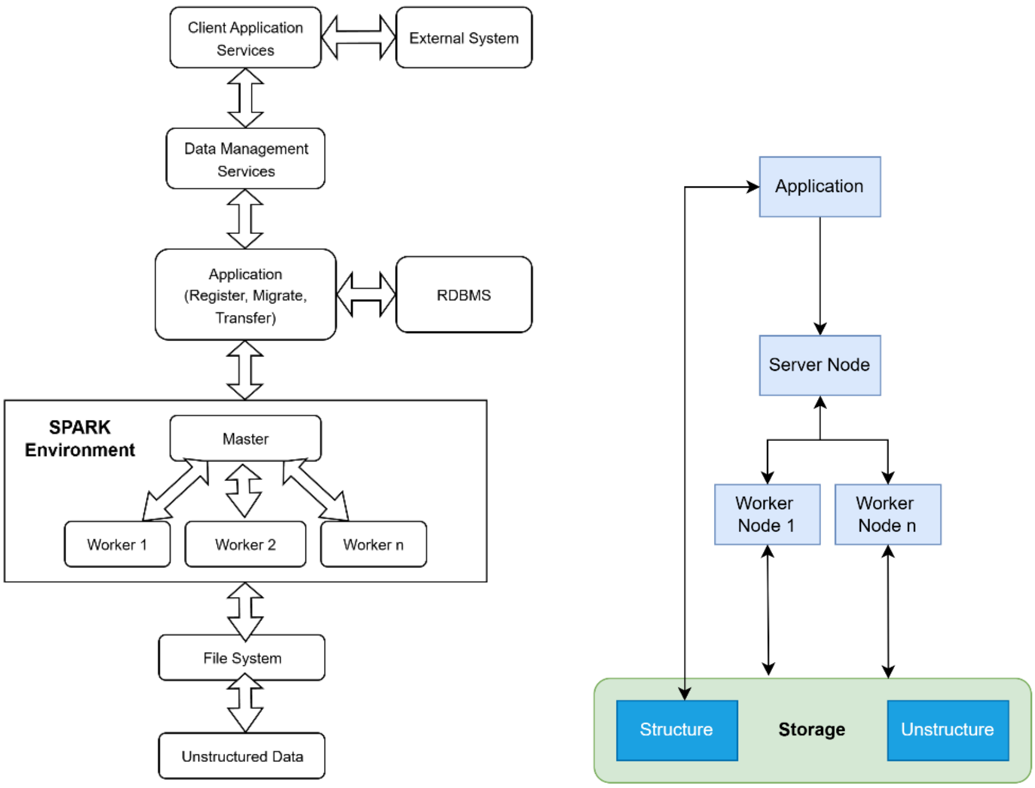
3.2. Model design

The results of the migration parameter analysis are used as the basis for designing the proposed storage management model and migration method. This design is important because it enables policy-driven data migration across storage tiers using Spark-based parallel processing. The proposed model consists of three main components: (i) data and storage management, (ii) process management system, and (iii) application and API gateway. These components work together to manage data storage, execute parallel migration processes, and provide user access to the system. This integration allows migration policies to be executed dynamically within a parallel processing environment, improving coordination between storage management and data processing compared to the existing system.

The migration experiments were conducted using the implementation architecture shown in Figure 4. The experimental environment consists of one application server, one master (server) node, two worker nodes, and the storage subsystem used to evaluate migration performance under 1 Gbps and 10 Gbps network environments. The service layer in the proposed architecture also supports integration with external systems through client application services and an API gateway. This feature enables the data management system to provide remote sensing data services to other applications, such as disaster monitoring platforms, satellite communication-related data services, and internet of things (IoT)-based environmental sensor networks. In this context, scalable storage and parallel migration are relevant not only for geoinformatics, but also for telecommunication and computing domains because they support faster data delivery, distributed processing, and service-oriented access to large-scale remote sensing archives.

Figure 4 illustrates the proposed design architecture. Figure 4(a) shows the application architecture for data registration, migration, and transfer using Spark-based processing and relational database management system (RDBMS) integration. Figure 4(b) shows the experimental architecture, including the application, server node, worker nodes, and storage components for structured and unstructured data. Figure 5 presents the proposed data management system model in more detail. The model shows how client applications access the system through application services and an API gateway, while the core data management application handles registration, migration, and transfer processes. These processes are connected to structured data storage and Spark-based parallel processing, which coordinates the master and worker nodes. The unstructured remote sensing data are managed across online, nearline, and archive storage tiers, enabling policy-driven migration and long-term storage management.

The service layer and API gateway in the proposed architecture also provide opportunities for integration with external systems. Through this layer, remote sensing data services can be accessed by other applications, such as disaster monitoring platforms, satellite communication-related services, and IoT-based environmental sensor networks. This capability broadens the relevance of the proposed model beyond geoinformatics by supporting service-oriented access, distributed computing, and scalable data delivery for telecommunication and computing applications.



(a) (b)

Figure 4. Design architecture application: (a) application architecture for data registration, migration, and transfer using Spark-based processing and RDBMS integration; and (b) experimental architecture showing the application, server node, worker nodes, and storage components for structured and unstructured data

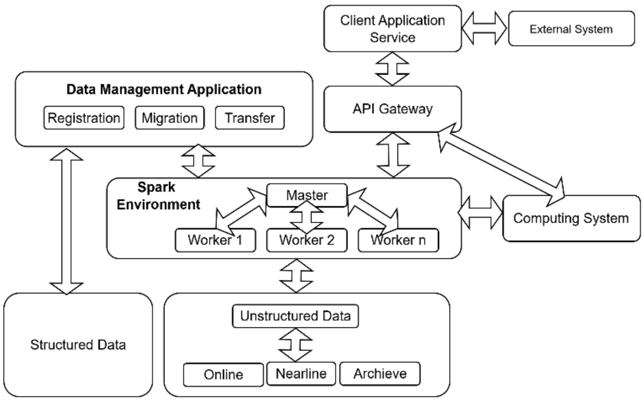
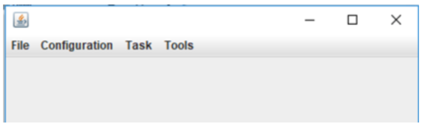


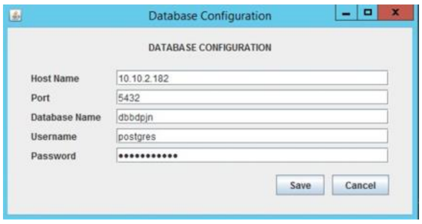
Figure 5. Proposed data management system model design

3.3. Application module design

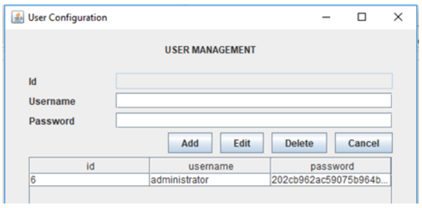
The application consists of several modules, including data registration, migration, data transfer, and storage management. These modules are designed to support the execution of migration policies and enable data movement across storage tiers. The implementation of these modules demonstrates how the proposed model operates in practice, particularly in executing migration processes and managing data across different storage levels. Figure 6 shows the main interface of the application module, including system configuration, user management, and Apache Spark management shown in Figures 6(a) to (d).



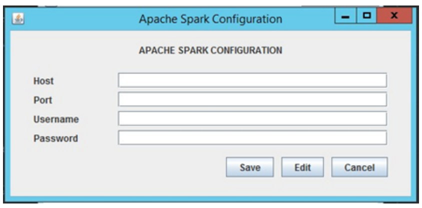
(a)



(b)



(c)



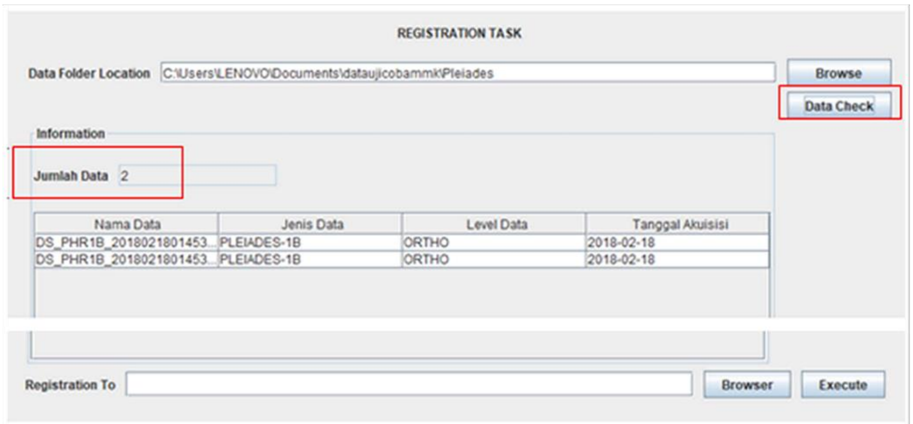
(d)

Figure 6. Application module interface: (a) initial menu, (b) database configuration, (c) user configuration, and (d) Apache Spark management

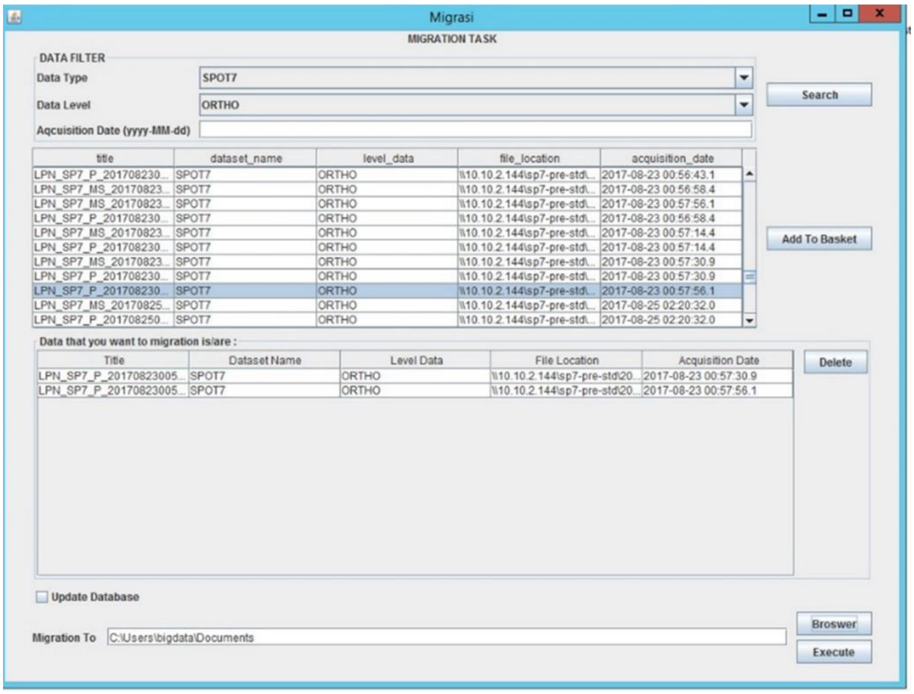
This interface enables users to manage data migration processes and monitor system operations. The migration process is executed through a parallel workflow, where the master node distributes tasks to worker nodes. Each worker processes assigned data and reports execution results back to the master system. In addition, the storage management module provides information on storage tiers and policies, supporting controlled data placement and monitoring across Tier 1, Tier 2, and Tier 3.

Migration performance is influenced by the effective data transfer speed, which depends on network configuration and system setup. In the 1 Gbps environment, the maximum transfer speed ranges from 754 Mbps to 900 Mbps after configuration improvements. In contrast, the 10 Gbps environment achieves an average transfer speed of up to 5 Gbps. This difference in transfer speed contributes to the observed variation in migration performance between the 1 Gbps and 10 Gbps environments.

Figure 7 illustrates the data registration (Figure 7(a)) and data migration interfaces (Figure 7(b)), which support the execution of migration tasks within the system. These implementation results demonstrate that the proposed model is capable of executing policy-driven and parallel data migration. The effectiveness of this approach is further evaluated through migration experiments under different network conditions.



(a)



(b)

Figure 7. User interfaces supporting data management processes: (a) data registration page, and (b) data migration page for managing data transfer between storage tiers

3.4. Migration experiment

Migration experiments were conducted using multiple satellite datasets, including Landsat 8, SPOT 6, SPOT 7, and Pleiades, with varying data volumes ranging from 10 to 100 datasets. The experiments were performed under two network environments, namely 1 Gbps and 10 Gbps. These experiments are important to evaluate how the proposed model performs under different data loads and network conditions. Figure 8 presents the results of data migration experiments under 1 Gbps and 10 Gbps network environments using different satellite datasets.

Figure 8(a) presents the data size comparison. Figure 8(b) illustrates the migration time under the 1 Gbps network environment, where processing time increases almost linearly with the number of datasets. Figure 8(c) presents the corresponding results under the 10 Gbps environment, showing a substantially lower increase in migration time as dataset volume grows. This indicates that the migration process is strongly influenced by data size when network bandwidth is limited. In contrast, in the 10 Gbps environment, migration time becomes less sensitive to increasing data volume. Although processing time still increases, the rate of increase is lower than in the 1 Gbps environment. This behavior suggests that higher network bandwidth reduces data transfer bottlenecks and supports more efficient migration, particularly for larger datasets.

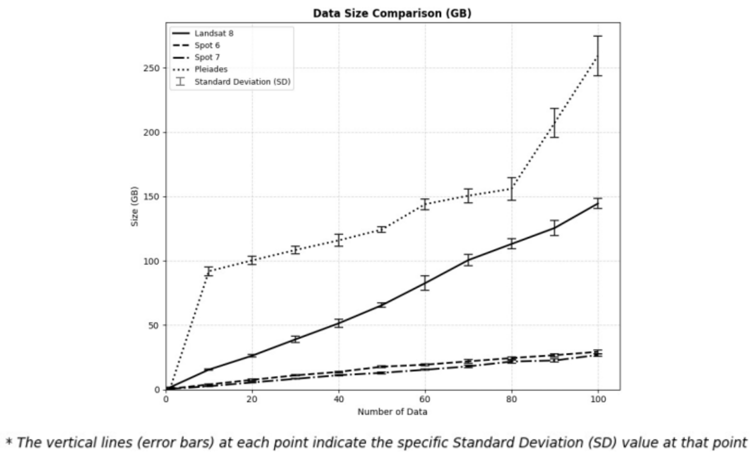
The improvement is also influenced by the Spark-based parallel processing mechanism. In the existing sequential migration process, data transfer tasks are generally executed one after another, so each task must wait until the previous task is completed. In the proposed model, migration tasks are distributed to worker nodes and executed in parallel. This allows multiple files to be migrated concurrently and reduces the total processing time. Therefore, the observed performance gain is not only caused by higher network bandwidth, but also by the ability of the parallel processing environment to reduce waiting time and improve overall throughput.

Figure 9 shows the percentage improvement in migration performance between the 1 Gbps and 10 Gbps environments for different datasets. As shown in Figure 9, performance improvements are observed across all datasets, with improvement percentages ranging from 80.14% to 91.04%. The highest improvement is observed in SPOT-7 (91.04%), followed by Landsat-8 (90.42%). Although larger datasets such as Pleiades also show substantial improvement, the percentage gain is slightly lower. This variation can be explained by differences in file size, data volume, and transfer overhead among datasets. Smaller or medium-sized datasets may benefit more directly from parallel task distribution and higher bandwidth because multiple files can be transferred concurrently with lower I/O overhead. In contrast, very high-resolution datasets such as Pleiades have larger file sizes, so migration performance may still be affected by storage I/O, file system access, and coordination overhead between the master and worker nodes. Therefore, the speed-up does not increase uniformly across all datasets.

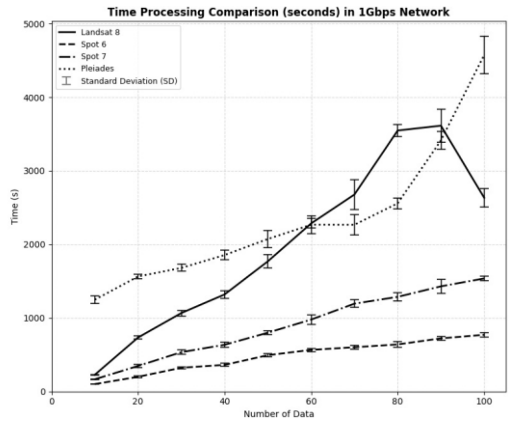
The results also show that the 10 Gbps environment consistently provides shorter migration time than the 1 Gbps environment. However, the performance improvement is not purely linear with the increase in network bandwidth. This is because migration performance is affected not only by network capacity, but also by storage access speed, file size, metadata update processes, task scheduling, and worker coordination. Thus, increasing network bandwidth can significantly improve migration performance, but the overall result still depends on the balance between network, storage, and processing components.

From the storage management perspective, the experiment highlights the trade-off between storage tiers. Tier-1 storage is suitable for active or frequently accessed data because it provides faster access, but it has higher cost and limited capacity. Tier-2 and archive storage are more appropriate for older or less frequently accessed data because they provide larger and more cost-efficient storage capacity, although with slower access performance. In the BDPJN environment, the proposed HSM-based policy helps maximize the use of existing hardware by assigning data to appropriate storage tiers based on data age, priority, file size, and capacity threshold. This approach improves storage efficiency without requiring all data to remain in high-performance storage.

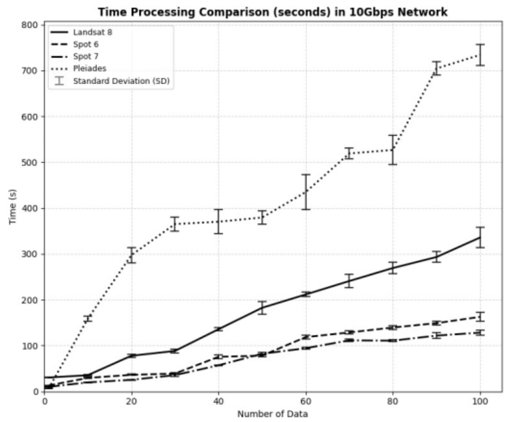
The migration experiment indicates that network bandwidth, storage tier characteristics, and parallel processing configuration jointly influence migration performance. The integration of HSM with Spark-based parallel processing can therefore enhance data migration efficiency, particularly when supported by higher network capacity. In practical terms, this model can support large-scale remote sensing data management by reducing migration bottlenecks, improving access to active datasets, and enabling more efficient use of archive storage. However, further validation using repeated experiments, additional baseline comparisons, and larger heterogeneous datasets is needed to provide stronger statistical evidence and evaluate scalability under broader operational conditions.



(a)



(b)



(c)

Figure 8. Experimental datasets and migration processing time under different network environments: (a) data size of the evaluated satellite datasets (Landsat-8, SPOT-6, SPOT-7, and Pleiades) as the number of data increases. Migration processing time under (b) 1 Gbps and (c) 10 Gbps network environments. Error bars indicate the standard deviation (SD) of repeated measurements

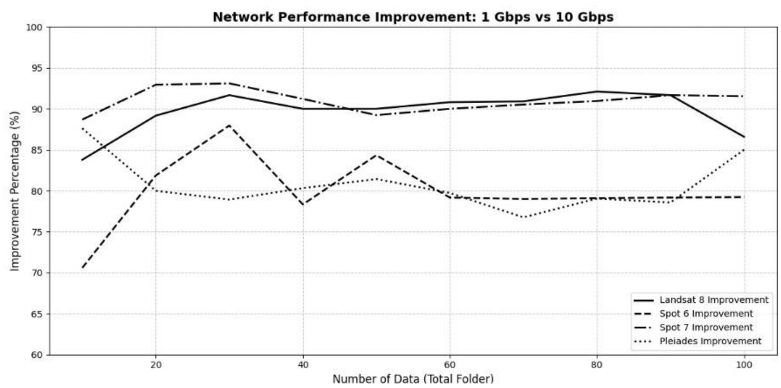


Figure 9. Percentage improvement in data migration performance between 1 Gbps and 10 Gbps network environments for Landsat-8, SPOT-6, SPOT-7, and Pleiades datasets

4. CONCLUSION

The main contribution of this study is the development of an integrated HSM–Spark model for long-term remote sensing data management. The proposed model combines HSM across performance, capacity, and archive tiers with Spark-based parallel migration to support more efficient data placement and migration. Experimental results show that the proposed model improves migration performance by approximately 80.14% to 91.04% compared with the existing manual migration system. These findings indicate that integrating HSM policies with parallel processing can reduce migration bottlenecks, improve storage utilization, and support more scalable management of large remote sensing archives.

The proposed model is also relevant to telecommunication and computing domains because it provides service-oriented access through the application service layer and API gateway. This architecture enables future integration with external systems, such as national disaster early warning systems, climate monitoring platforms, smart city infrastructures, satellite communication-related services, and IoT-based sensor networks. Therefore, the model can support faster data delivery, distributed computing, and long-term accessibility of satellite imagery for broader operational applications.

This study has several limitations and opportunities for future improvement. The experiments were limited to selected operational satellite datasets, namely Landsat 8, SPOT 6, SPOT 7, and Pleiades, and were conducted in a controlled experimental environment. Future research should extend the evaluation to more heterogeneous datasets and larger-scale infrastructures, including hybrid cloud or cloud-native storage environments. Future development may also incorporate AI-driven migration policies, workload-aware storage tiering, and natural language processing interfaces that allow users to submit migration commands through prompts, which can then be interpreted and executed by the data management application.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the National Research and Innovation Agency (BRIN), particularly the National Remote Sensing Data Bank (BDPJN) team, for their technical assistance and support during the data management and migration experiments. The authors also thank colleagues who provided valuable input and feedback throughout the preparation of this manuscript.

FUNDING INFORMATION

This research was supported by the National Research and Innovation Agency (BRIN), Indonesia, through the National Remote Sensing Data Bank (BDPJN) research program. Authors state no external funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Riyan Mahendra Saputra	✓	✓	✓	✓	✓	✓		✓	✓			✓	✓	
Agnes Sondita Payani					✓	✓		✓		✓	✓			
Ogi Gumelar		✓			✓		✓			✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are derived from the National Remote Sensing Data Bank (BDPJN), managed by the National Research and Innovation Agency (BRIN). Due to institutional data policies and confidentiality agreements, these datasets are not publicly available.

REFERENCES

[1] M. A. Wulder *et al.*, "Fifty years of Landsat science and impacts," *Remote Sensing of Environment*, vol. 280, p. 113195, 2022, doi: 10.1016/j.rse.2022.113195.

[2] E. Dritsas and M. Trigka, "Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey," 2025, doi: 10.3390/rs17030550.

[3] Group on Earth Observations (GEO), "Revised GEO Data Sharing and Data Management Principles," Group on Earth Observations (GEO). Accessed: Mar. 06, 2025. [Online]. Available: https://old.earthobservations.org/index.php

[4] R. R. Downs, "Implementing the group on earth observations data management principles: Observations of a scientific data center," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, no. 3W2, pp. 51–55, 2017, doi: 10.5194/isprs-archives-XLII-3-W2-51-2017.

[5] H. Herodotou and E. Kakoulli, "Automating distributed tiered storage management in cluster computing," in *Proceedings of the VLDB Endowment*, Sep. 2019, pp. 43–56, doi: 10.14778/3357377.3357381.

[6] D. Medeiros *et al.*, "IO-SEA: Storage I/O and Data Management for Exascale Architectures," *Proceedings of the 21st ACM International Conference on Computing Frontiers 2024 Workshops and Special Sessions, CF 2024 Companion*, pp. 94–100, 2024, doi: 10.1145/3637543.3654620.

[7] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016, doi: 10.1109/JPROC.2016.2598228.

[8] I. Chebbi, W. Boulila, N. Mellouli, M. Lamolle, and I. R. Farah, "A comparison of big remote sensing data processing with Hadoop MapReduce and Spark," 2018, *IEEE*, doi: 10.1109/ATSIP.2018.8364497.

[9] H. Lan, X. Zheng, and P. M. Torrens, "Spark Sensing: A Cloud Computing Framework to Unfold Processing Efficiencies for Large and Multiscale Remotely Sensed Data, with Examples on Landsat 8 and MODIS Data," *Journal of Sensors*, vol. 2018, pp. 1–12, Aug. 2018, doi: 10.1155/2018/2075057.

[10] H. Wu and K. Fu, "A management of remote sensing big data base on standard metadata file and database management system," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, no. 3/W10, pp. 653–657, Feb. 2020, doi: 10.5194/isprs-archives-XLII-3-W10-653-2020.

[11] A. S. Payani, A. M. Mayanda, A. Adiresta, and Y. Ruldeviyani, "Data Quality Management Strategy to Improve Remote Sensing Data Quality : A Case Study on National Remote Sensing Data Bank," *IPTEK The Journal for Technology and Science*, vol. 33, no. 3, pp. 162–174, 2022, doi: 10.12962/j20882033.v33i3.14198.

[12] C. Xu *et al.*, "Cloud-based storage and computing for remote sensing big data: a technical review," *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1417–1445, 2022, doi: 10.1080/17538947.2022.2115567.

[13] X. Wang, Y. Zhang, and D. Yu, "Exploring the Relationships between Land Surface Temperature and Its Influencing Factors Using Multisource Spatial Big Data: A Case Study in Beijing, China," *Remote Sensing*, vol. 15, no. 7, 2023, doi: 10.3390/rs15071783.

[14] G. Yang, C. Su, H. Zhang, X. Zhang, and Y. Liu, "Tree-level landscape transitions and changes in carbon storage throughout the mine life cycle," *Science of the Total Environment*, vol. 905, p. 166896, 2023, doi: 10.1016/j.scitotenv.2023.166896.

[15] G. Li, G. Cheng, G. Liu, C. Chen, and Y. He, "Simulating the Land Use and Carbon Storage for Nature-Based Solutions (NbS) under Multi-Scenarios in the Three Gorges Reservoir Area: Integration of Remote Sensing Data and the RF–Markov–CA–InVEST Model," *Remote Sensing*, vol. 15, no. 21, p. 5100, 2023, doi: 10.3390/rs15215100.

[16] J. Xu *et al.*, "Toward a Novel Method for General On-Orbit Earth Surface Anomaly Detection Leveraging Large Vision Models and Lightweight Priors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–21, 2024, doi: 10.1109/TGRS.2024.3432749.

[17] X. Wang *et al.*, "Municipal and Urban Renewal Development Index System: A Data-Driven Digital Analysis Framework," *Remote Sensing*, vol. 16, no. 3, p. 456, 2024, doi: 10.3390/rs16030456.

[18] T. Dube, M. D. Shekede, and C. Massari, "Remote Sensing for Water Resources and Environmental Management," *Remote Sensing*, vol. 15, no. 1, p. 18, 2023, doi: 10.3390/rs15010018.

[19] J. Martusewicz, A. Wierzbic, and M. Łukaszewicz, "Strategic Transformation and Sustainability: Unveiling the EFQM Model 2025," *Sustainability (Switzerland)*, vol. 16, no. 20, p. 9106, 2024, doi: 10.3390/su16209106.

[20] T. Zhang, A. Hellander, and S. Toor, "Efficient Hierarchical Storage Management Empowered by Reinforcement Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5780–5793, 2023, doi: 10.1109/TKDE.2022.3176753.

[21] X. Liu *et al.*, "Crame: Hierarchical data management framework for cloud-edge-device collaboration," *Information Sciences*, vol. 740, p. 123197, 2026, doi: 10.1016/j.ins.2026.123197.

[22] V. Hayot-Sasson and T. Glatard, "Hierarchical Storage Management in User Space for Neuroimaging Applications," *Neuroinformatics*, vol. 24, no. 1, 2026, doi: 10.1007/s12021-025-09760-3.

[23] Y. Hong, M. S. Gil, S. Lee, and Y. S. Moon, "HBAC: Hierarchical-Based Access Control Model for Storage Management in Data Lake Environments," *Software - Practice and Experience*, vol. 56, no. 4, pp. 445–461, 2026, doi: 10.1002/spe.70055.

[24] J.-G. Lee and M. Kang, "Geospatial Big Data: Challenges and Opportunities," *Big Data Research*, vol. 2, no. 2, pp. 74–81, 2015, doi: 10.1016/j.bdr.2015.01.003.

[25] Y. Huang, P. Gao, Y. Zhang, and J. Zhang, "A Cloud Computing Solution for Big Imagery Data Analytics," in *2018 International Workshop on Big Geospatial Data and Data Science (BGDDS)*, 2018, pp. 1–4, doi: 10.1109/BGDDS.2018.8626847.

[26] M. M. Rathore, A. Ahmad, A. Paul, and A. Daniel, "Hadoop based real-Time Big Data Architecture for remote sensing Earth Observatory System," 2016, *IEEE*, doi: 10.1109/ICCCNT.2015.7395242.

BIOGRAPHIES OF AUTHORS

Riyan Mahendra Saputra received his Master's degree in Computer Science from Universitas Indonesia, Indonesia, in 2012. He has been a researcher in the field of remote sensing technology and data management since 2013, previously at the Remote Sensing Technology and Data Center, National Institute of Aeronautics and Space (LAPAN), Indonesia, which is now integrated into the National Research and Innovation Agency (BRIN). He is currently affiliated with the Directorate for Policy Formulation of Research, Technology and Innovation, National Research and Innovation Agency, Indonesia. His current research interests include remote sensing data management, big data systems, geospatial data infrastructure, and spatial information systems. He can be contacted at email: riyan.mahendra.saputra@brin.go.id.

Agnes Sondita Payani received her Bachelor's degree in Mathematics from Universitas Padjadjaran, Indonesia, in 2015, and her Master's degree in Information Technology from Universitas Indonesia, Indonesia, in 2024. She is currently an Assistant Researcher at the Research Center for Geoinformatics, National Research and Innovation Agency (BRIN), Indonesia. Her research interests include information systems, data governance, remote sensing data management, geospatial data infrastructure, and big data implementation for earth observation systems. She can be contacted at email: agne004@brin.go.id.

Ogi Gumelar received his Bachelor's degree in Mathematics from Universitas Gadjah Mada, Indonesia, in 2004, and his Master's degree in Remote Sensing and Geographic Information Systems from Beihang University, China, in 2018. He is currently affiliated with the Research Center for Geoinformatics, National Research and Innovation Agency (BRIN), Indonesia, and is pursuing a doctoral degree at Beihang University, with research focusing on satellite image-based carbon modelling. His research interests include remote sensing, geographic information systems, satellite image analysis, carbon modelling, and geospatial data applications. He can be contacted at email: ogig001@brin.go.id.