Analysis of Spatial Big Data Management Systems

Annette John
annette.john@ucalgary.ca
University of Calgary
UCID: 30224129

ABSTRACT

Spatial data, encompassing geographic information, has emerged as a pivotal component in various domains, playing a crucial role in decision-making, resource management, and technological advancements. Spatial data provides a comprehensive framework for understanding and analyzing the geographical relationships between various entities, offering insights that extend beyond traditional data sets. Its applications span diverse fields such as urban planning, environmental monitoring, healthcare, agriculture, and disaster management. In the era of big data, the sheer volume, velocity, and variety of data generated necessitate innovative approaches to extract meaningful patterns and knowledge. This paper aims to review the existing frameworks and methods in place to analyze big spatial data and provides a comprehensive summary of the same.

KEYWORDS

Geospatial big data, GeoSpark, SpatialHadoop, HDFS, Apache Sedona, SpatialIgnite, HPC

1 INTRODUCTION

In an era dominated by data-driven decision-making, the significance of spatial data management systems has risen to the forefront, transforming how we perceive, analyze, and utilize geographical information. About 80% of world’s data is geospatial data. The relation between spatial data and diverse applications necessitates sophisticated systems capable of efficiently storing, organizing, and retrieving information while accommodating the complexities inherent in geographic datasets, that traditional database management systems cannot provide.

This review embarks on a comprehensive exploration of the vast and evolving realm of spatial data management systems. Recently, there is an unprecedented proliferation of location-based data from sources such as satellites, sensors, and mobile devices, the demand for robust data management systems becomes increasingly pronounced. This paper aims to scrutinize the current state of spatial data management, delving into the architectures, functionalities, and methodologies employed by various systems.

The central focus of this review encompasses an in-depth analysis of the architectures that underpin spatial data management. From traditional relational databases to NoSQL databases, distributed file systems, and emerging technologies, the paper aims to dissect the strengths and limitations of each approach in handling the unique characteristics of spatial datasets. Furthermore, we will also see how geospatial big data is handled in different use-cases like agriculture and smart city enablement.

1.1 Motivation

This review is motivated by the intention to offer a cohesive narrative that bridges existing gaps in understanding. While spatial technologies like databases and indexing techniques have laid essential foundations, the advent of spatial data management frameworks introduces more complex computations to harness big spatial data effectively. This review aims to provide a gist of the existing technologies and frameworks, shedding light on their interoperability and performance under varied conditions. Additionally, the paper emphasizes the role of benchmarks in objectively assessing spatial data management systems, contributing to establishing performance standards crucial for informed decision-making.

2 GEOSPATIAL BIG DATA

Traditional big data analytics primarily focuses on managing and analyzing vast volumes of diverse, structured, and unstructured data to extract valuable patterns and knowledge. In contrast, geospatial big data introduces a crucial spatial context to this paradigm, incorporating location-based information into the analytical process. The integration of geographic data, such as coordinates, boundaries, and spatial relationships, enriches the understanding of complex phenomena and facilitates more nuanced decision-making. By merging traditional big data methodologies with geospatial analytics, organizations can unearth valuable spatial patterns, correlations, and trends, unlocking a new dimension of information that is particularly pertinent in fields like urban planning, environmental monitoring, logistics, and disaster response. This intersection not only broadens the scope of data analytics but also reinforces the recognition that location is a fundamental aspect shaping the intricacies of the data landscape.

2.1 Challenges

The primary challenge spatial big data faces stems from the unprecedented scale, complexity, and heterogeneity of spatial datasets. Managing and processing vast volumes of geospatial information poses significant scalability and performance issues, necessitating innovative approaches for storage and computation that are different from how we handle traditional big data. Existing tools do not natively offer support to process geospatial data, as we will see described in the papers that were reviewed. Over the years, an effort has been made to add extensions to existing database systems to try and fill this gap.

3 METHODOLOGY

3.1 Features of Geospatial Big Data

The features of geospatial big data seamlessly intertwine with the foundational 5Vs of big data—Volume, Variety, Velocity, Veracity, and Value—bringing a distinctive spatio-temporal dimension to the forefront. The Volume aspect is accentuated by the sheer magnitude of spatial information generated from satellites, sensors, and location-aware devices. Variety takes centre stage with diverse geospatial data formats, encompassing vector maps, raster imagery, and real-time streaming information. The temporal Velocity of geospatial big data, influenced by real-time updates and dynamic events, requires agile processing to extract timely insights. Veracity gains prominence in geospatial contexts, emphasizing the accuracy and reliability of location-based information, crucial for decision-making in applications like navigation and disaster response. Finally, the synthesis of these aspects leads to the creation of Value from geospatial big data. In essence, the unique features of geospatial big data align harmoniously with the 5Vs, presenting a multifaceted analytical challenge that, when navigated adeptly, unlocks a wealth of valuable information.

3.2 Sources of Geospatial Big Data

Geospatial big data is produced from diverse sources, each contributing substantial volumes of information. Earth observations, facilitated by in-situ and remote sensors, generate massive and dynamic datasets, exemplified by the Landsat archive and NASA's Earth Observing System Data and Information System (EOSDIS). The advancements in computing power enable geoscience model simulations, exemplified by climate models from the Intergovernmental Panel on Climate Change (IPCC), producing extensive simulated geospatial data. The Internet of Things (IoT) connects devices worldwide, generating vast amounts of geospatial data, characterized by its unstructured and dynamic nature. Volunteered geographic information (VGI) leverages citizen sensors, producing massive amounts of location-based data through platforms like Twitter and Instagram. These sources collectively contribute to the complexity of geospatial big data, encompassing Earth observations, model simulations, IoT-generated streams, and volunteered citizen data.

3.3 Geospatia data formats

Geospatial data formats serve as standardized structures for organizing and storing geographic information, and they are made available through various means to facilitate accessibility and interoperability. One common method is through open data repositories and government agencies that publish geospatial datasets in widely used formats like GeoJSON, Shapefile, and GeoPackage. These datasets are often shared on online platforms, allowing users to download and utilize the information for research, analysis, or application development.

Additionally, many organizations provide Application Programming Interfaces (APIs) that enable real-time access to geospatial data. These APIs deliver data in standard formats, such as GeoJSON or KML, allowing developers to integrate dynamic and up-to-date geographic information directly into their applications.

Web services, such as Web Map Services (WMS) and Web Feature Services (WFS), play a crucial role in distributing geospatial data over the internet. These services allow users to request and retrieve map images or raw spatial data, respectively, fostering seamless integration with GIS applications.

Cloud-based solutions and web services have also become prominent in providing geospatial data. Platforms like Amazon Web Services (AWS), Google Cloud, and Microsoft Azure host vast repositories of geospatial information, offering data in different formats and providing tools for analysis and visualization.

Overall, the availability of geospatial data in diverse formats through open data platforms, APIs, and web services promotes widespread access, collaboration, and innovation in the field of geographic information systems.

3.4 Comparision between SQL and NoSQL database systems for geospatial data

In geospatial data management, a critical consideration lies in choosing the appropriate database system to effectively handle the challenges posed by massive volumes, diverse formats, and dynamic characteristics of geospatial Big Data. Traditional SQL databases, including Relational Database Management Systems (RDBMS) with spatial extensions, have long been employed for geospatial services. However, their limitations become pronounced when confronted with the substantial volume, velocity, and variety components inherent in geospatial Big Data. A typical SQL-based system brings with it challenges such as the necessity for a mapping layer, which can impact scalability, availability, and overall performance. In contrast, NoSQL databases emerge as a promising alternative, offering a storage data model aligned with application data models, thereby eliminating the need for a mapping layer. The inherent design of NoSQL databases, emphasizing data distribution and scalability in clustered environments, positions them as a robust solution for handling geospatial Big Data efficiently.

The evaluation of database management systems for geospatial Big Data reveals distinct advantages and challenges associated with both relational (SQL) and NoSQL databases. Relational DBMSs, characterized by tables, keys, and SQL, excel in structured data environments where strong consistency is paramount. SQL databases, with their fixed schemas, transaction support, and efficient handling of fixed-schema geospatial data, are well-suited for enterprise GIS systems. However, challenges arise with highly connected data, as joins become computationally expensive, and scalability compromises the normalized relational model. On the other hand, NoSQL databases, embracing diverse models like key-value, document, column-family, and graph, provide scalability and flexibility for large amounts of semi-structured and unstructured data. While each NoSQL type addresses specific use cases, they collectively offer alternatives for achieving scalability and distribution. Document databases, for instance, enable efficient geospatial data management with flexible queries, supporting proximity queries and handling relationships. Column-family databases exhibit high performance in search and data retrieval, making them suitable for GIS applications with heavy data insertion and fast retrieval needs. Graph databases, rooted in graph theory, excel in managing highly connected data, a characteristic beneficial for modeling geospatial data as networks and efficiently handling topological relationships. The choice between SQL and NoSQL hinges on the specific requirements of the geospatial Big Data application, with each type offering unique strengths to cater to diverse use cases.

3.5 Review of NoSQL database systems

The systematic review on spatial data handling in NoSQL databases compares popular systems such as Redis, MongoDB, CouchDB, Cassandra, HBase, and Neo4j. The analysis includes an assessment of native spatial support and available spatial extensions. Redis provides limited support, while MongoDB stands out with comprehensive spatial data types and operations. CouchDB and Neo4j offer native support for simple points, while Cassandra and HBase lack native spatial capabilities. Spatial extensions, proposed primarily for Cassandra and HBase, aim to improve spatial query processing using techniques like geohashing, space-filling curves, and specific spatial indexing methods. The study correlates these findings with spatial application requirements, highlighting variations in ad-hoc query support, interoperability, adherence to standards, data visualization, query processing efficiency, and extensibility among the considered NoSQL databases.

In summary, MongoDB emerges as a strong candidate for spatial data handling due to its robust native support, including

GeoJSON representation and spatial operations. While Redis, CouchDB, Cassandra, and HBase lack certain spatial capabilities, researchers have proposed extensions to enhance their spatial functionalities.

3.6 Tools for Geospatial Big Data Processing and Analytics

The landscape of geospatial big data processing and analytics is rich with diverse tools and frameworks tailored to overcome the unique challenges posed by massive datasets with location information. Extended systems for Hadoop/Spark engines, such as Parallel Secondo, ESRI Tools for Hadoop, SpatialHadoop, GeoTrellis, GeoSpark, Magellan, LocationSpark, and Spatial In-Memory Big Data Analytics (SIMBA), contribute significantly to enhancing the functionality of these engines for spatial data management.

In big data architectures, the discussed frameworks, including Lambda, Kappa, Liquid, BDAS, SMACK, and HPCC, have been developed on integrated infrastructures. However, the focus on spatial data management within these architectures remains limited.

NoSQL databases play a pivotal role in spatial data storage and processing. The Cassandra-Solr-Spark framework enables spatial query processing, while a dedicated NoSQL (Cassandra) based spatial data storage framework offers distributed and scalable APIs for operations like location search, proximity search, and KNN search. GeoSpark and GeoTrellis, spatial extensions for Spark, contribute to spatial analytics applications.

Parallel processing systems, exemplified by Dask-GeoPandas and Apache Sedona, prove crucial in the context of smart city applications, with Apache Sedona showcasing superior performance in read, write, join, and clustering operations. Additionally, emerging solutions such as Apache Ignite and DISTIL address the limitations of Spark-based systems, offering in-memory distributed computing for efficient spatial data processing.

This comprehensive suite of tools and technologies collectively forms a robust ecosystem for geospatial big data processing and analytics, catering to a wide array of applications and use cases.

4 PROPOSED GEOSPATIAL BIG DATA ANALYTICS ARCHITECTURES

In this section, we present summaries of two papers that contribute significantly to the evolving landscape of geospatial big data architecture. The first paper proposes a big data analytics framework tailored for the agricultural domain, addressing the challenges of spatial data management and providing scalable solutions for real-world applications. The second paper explores a framework designed for smart city analytics, emphasizing the integration of key technological drivers such as big data, Geographic Information Systems (GIS), and cloud computing for sustainable urban governance and data-driven decision-making.

4.1 Big Data Analytics Architecture for spatial data in agriculture

The proposed big data analytics architecture focuses on three main components: data preparation, big spatial analytics, and data visualization. The architecture is designed to be in-memory, open source, scalable, flexible, extendible, and cost-effective.

4.1.1 Data Preparation Framework: It fetches consistent data from disparate sources, unifying it through abstraction layers and complex tools, like data fusion algorithms and schema mapping tools.

4.1.2 Big Spatial Data Analytics Framework: Developed on top of the big data stack with Spark as the core processing engine and Cassandra for data storage. It offers distributed and scalable APIs for spatial operations and a web-based REST interface. The architecture allows users to execute ad-hoc queries on Spark or Cassandra, balancing low latency queries on Cassandra and complex queries on Spark

4.1.3 Data Visualization Framework: Implemented to showcase analytical results through Restful ad-hoc APIs and interactive maps. A dashboard application is designed for dynamic visualization of analytical results

Prototype applications in the agriculture domain are developed using this architecture, addressing challenges specific to both developed and developing countries. The implementation involves collecting spatial and non-spatial data on weather, crop, and market from various sources, displaying results through a web-based dashboard.

The big data analytics architecture is designed to manage massive-scale data, particularly spatial data, addressing challenges related to variety and volume. The implementation includes a REST interface for data collection and the development of prototype applications in the agriculture domain.

This comprehensive architecture provides an integrated solution for managing and analyzing large-scale spatial data, making it a valuable case study for applications in diverse domains.

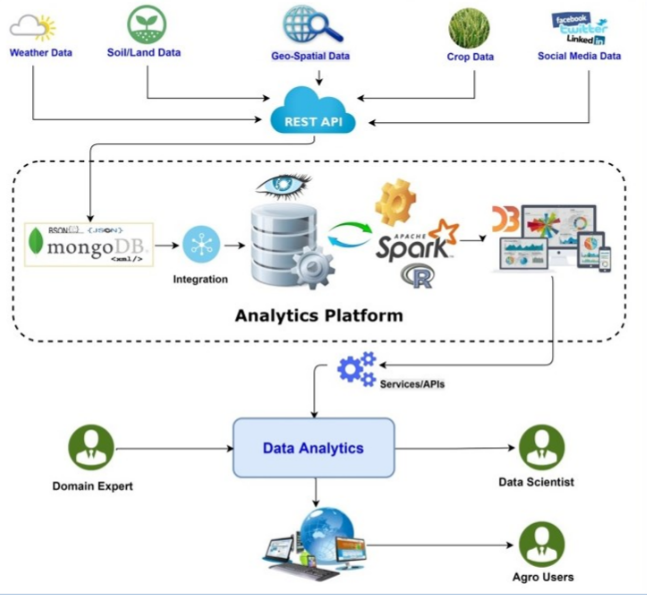

Figure 1: Big data analytics architecture for agriculture data

4.2 Geospatial Big Data Analytics for Sustainable Smart Cities

The proliferation of urbanization globally presents challenges in social, environmental, and economic aspects, necessitating sustainable smart cities leveraging information and communication technologies (ICTs). Geospatial big data management is crucial for smart city services, relying heavily on location-based data. This paper explores the effective handling of geospatial big data for sustainable smart cities, emphasizing storage, visualization, analytics, and analysis stages to foster green building, green energy, and net zero targets. This study advocates for high-performance and scalable infrastructures like parallel processing and cloud computing to address the challenges posed by spatial data.

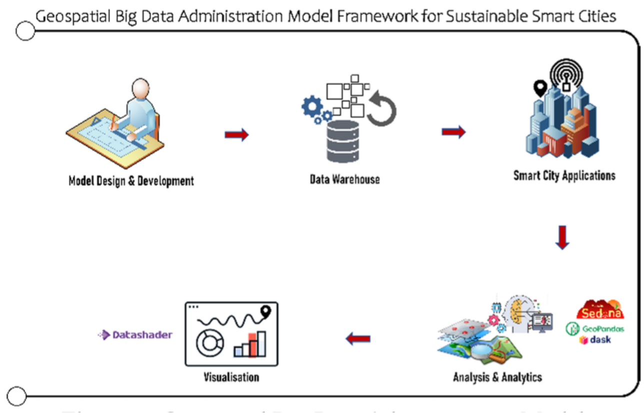
Geospatial Big Data Administration Model Framework for Sustainable Smart Cities
Model Design & Development
Data Warehouse
Smart City Applications
Analysis & Analytics
Sedona
GeoPandas
dask
Visualisation
Datashader

Figure 2: Geospatial Big Data Administration Model Framework for Sustainable Smart Cities

4.2.1 Geospatial Big Data Tools Dask-GeoPandas and Apache Sedona: Two prominent open-source tools, Dask-GeoPandas and Apache Sedona, are highlighted for effective geospatial big data management. Dask-GeoPandas combines spatial capabilities with scalability, offering enhanced parallelism for large geographic data processing. Apache Sedona is a cluster processing system designed to efficiently handle large-scale geospatial data, providing spatial indexing, partitioning, and serialization operations.

4.2.2 Implementation: The study focuses on England and Wales, utilizing open data sources such as Energy Performance Certificates (EPC), Ordnance Survey (OS) Open Unique Property Reference Number (UPRN), and OS Building data. Performance comparisons between Dask-GeoPandas and Apache Sedona are conducted for read, write, and spatial join operations, with Apache Sedona demonstrating superior performance.

4.2.3 Geospatial Big Data Analytics: Spatial clustering analysis, employing k-means clustering, is performed to group buildings based on location and attribute information. The study visualizes building-scale energy efficiency analytics using Datashader, providing insights into regional energy efficiency patterns. Spatial data analytics enables problem detection, prediction, and decision optimization for smart governance.

Geospatial big data tools, such as Dask-GeoPandas and Apache Sedona, are crucial in implementing sustainable smart city components. The study emphasizes the need for effective management of geospatial big data, showcasing the capabilities of parallel processing systems. By leveraging big data analytics in smart city applications, it becomes possible to design and implement components such as "Smart Environment," "Smart Infrastructure," "Smart Energy," "Smart Building," and "Smart Governance," addressing specific city requirements and maturity targets.

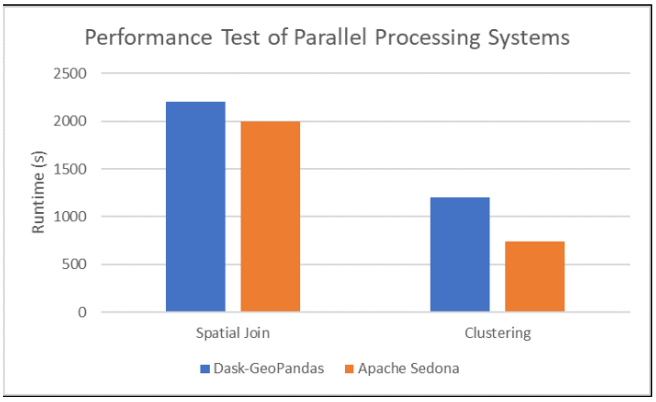

Figure 3: Performance test for comparing geospatial big data parallel processing frameworks.

5 BENCHMARKS

The performance and scalability of spatial data systems are critical considerations for both research organizations and enterprises seeking effective solutions for their spatial data processing needs. To address this, in this section we will review a benchmarking study that scrutinizes the present landscape of Big Spatial Data systems, offering a comparative analysis of representative systems based on Apache Hadoop and Apache Spark frameworks. The primary objective is to assess the efficacy of these systems in managing large volumes of spatial data efficiently. The benchmarking methodology involves a comprehensive set of spatial join operations, range queries, and spatial analysis functions. Inspired by notable benchmarks such as Jackpine, the evaluation study introduces a new benchmark tailored to Big Spatial Data systems. This benchmark encompasses various OGC-compliant topological relations and spatial analysis functions, providing an assessment of system capabilities.

5.0.1 Representative Systems: The study includes three representative systems: **SpatialHadoop**, a mature Hadoop-based system; **GeoSpark**, an active Spark-based system; and **SpatialIgnite**, a novel distributed in-memory system based on Apache Ignite. Each system is subjected to rigorous testing, employing a diverse range of spatial queries and operations to unveil their strengths and weaknesses.

5.0.2 Results:

- SpatialIgnite: The study shows that SpatialIgnite outperforms both Hadoop and Spark-based systems, specifically GeoSpark and SpatialHadoop, in terms of real-world spatial datasets.
- Focus on Perfomance: The paper emphasizes the importance of performance and scalability as key factors when assessing Big Spatial Data systems. It highlights the limitations of popular frameworks like Spark due to overhead associated with scheduling, distributed coordination, and data movement.
- Benchmark Contributions: The paper introduces a new benchmark for Big Spatial Data systems, aiming to help the research community assess the state of the art. It extends the Jackpine benchmark with additional operations.

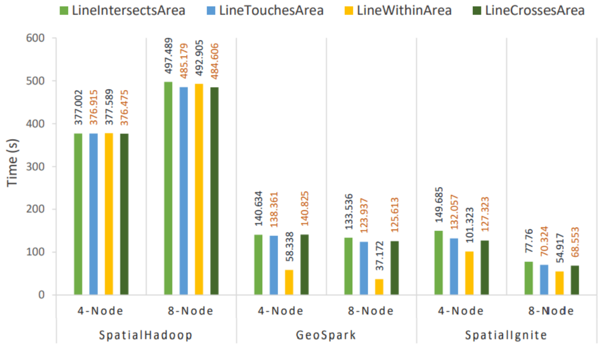

Figure 4: Performance Comparison - Scalability (4-nodes vs 8 nodes)

6 GEOSPATIAL BIG DATA PROCESSING IN HIGH-PERFORMANCE COMPUTING (HPC)

Handling geospatial big data in an HPC environment poses several challenges that need careful consideration. This section discusses key challenges related to data storage, spatial indexing, domain decomposition, and task scheduling. Additionally, it provides insights into existing platforms categorized into general-purpose platforms, geospatial-oriented platforms, query processing systems, and workflow-based solutions. The discussion highlights the importance of efficient spatial indexing, domain decomposition strategies, and task scheduling considerations for optimal parallelization.

6.0.1 Data Storage Challenges: Effective geospatial big data processing demands robust data storage solutions. Traditional architectures like Shared-Everything (SEA) and Shared-Disk (SDA) have their limitations in handling petabyte-scale geospatial datasets. The Shared-Nothing Architecture (SNA) is optimal for its scalability, fault tolerance, and parallel processing capabilities.

6.0.2 Spatial Indexing for Quick Data Access: Efficient data access in geospatial datasets hinges on robust spatial indexing. Tree structures like quadtree, KD-tree, and R-tree play a key role. Platforms such as SpatialHadoop and GeoSpark leverage spatial partitioning for enhanced data access efficiency, laying the groundwork for subsequent analyses.

6.0.3 Domain Decomposition Strategies: A geospatial big data processing strategy is the divide-and-conquer approach, dissecting complex problems into manageable subsets by abstracting geospatial data into a five-dimensional tuple informs 1D, 2D, and 3D spatial decompositions.

6.0.4 Task Scheduling for Parallelization: Efficient task scheduling is pivotal for successful parallelization. Load balancing algorithms, both static and adaptive are useful. However, in the context of geospatial big data platforms, customized load balancing mechanisms with a focus on data locality is imperative to minimize movement costs.

6.0.5 Review of existing HPC Platforms: Diverse platforms cater to geospatial big data processing, falling into general-purpose, geospatial-oriented, query processing, and workflow-based categories. Open MPI, HTCondor, CUDA, HadoopGIS, GeoSpark, and PostGIS offer unique strengths and use cases.

There are also several advantages of cloud-based HPC in geospatial big data processing. On-demand resource provision and high scalability are key factors enabling the realization of the full potential of geospatial analyses. The adaptive nature of cloud-based HPC environments provides insights into the future of efficient and scalable geospatial big data processing.

7 LIMITATIONS

While the review provides a comprehensive exploration of geospatial big data management systems, certain limitations should be acknowledged. The scope of the review primarily focuses on existing frameworks, methodologies, and architectures, and it may not capture the latest developments in this rapidly evolving field. Additionally, the emphasis on certain technologies and tools may introduce a degree of bias, and emerging solutions might not be fully represented.

The benchmarking study, while insightful, is based on a specific set of representative systems, and the results may vary with different datasets and use cases. It's important to note that the evaluation is not exhaustive and may not cover all possible spatial data processing scenarios.

8 CONCLUSION

In conclusion, the analysis of geospatial big data management systems reveals a dynamic landscape with both opportunities and challenges. Embracing cloud-based high-performance computing proves instrumental in unlocking the full potential of geospatial analyses, with on-demand resource provision and scalability being key catalysts.

The limitations identified underscore the need for continuous research and innovation in spatial data processing. Overcoming scalability concerns, refining spatial indexing methodologies, and optimizing task scheduling mechanisms are crucial for advancing the efficiency and applicability of geospatial big data platforms.

As technology evolves, addressing these limitations will pave the way for more robust geospatial big data processing, fostering innovation across domains such as urban planning, environmental monitoring, agriculture, and disaster management. The adaptive nature of cloud-based HPC environments provides a glimpse into a future where geospatial analyses are not only efficient but also seamlessly scalable to meet the evolving demands of a data-driven world.

9 REFERENCES

(1) Developing Big Data Analytics Architecture for Spatial Data https://ceur-ws.org/Vol-2399/paper03.pdf
(2) GEOSPATIAL BIG DATA ANALYTICS FOR SUSTAINABLE SMART CITIES https://isprs-archives.copernicus.org/articles/XLVIII-4-W7-2023/141/2023/isprs-archives-XLVIII-4-W7-2023-141-2023.pdf
(3) A Performance Study of Big Spatial Data Systems https://www.cs.unb.ca/~sray/papers/BigSpatialBenchmark_BigSpatial2018.pdf
(4) Geospatial Big Data Handling with High Performance Computing: Current Approaches and Future Directions https://arxiv.org/ftp/arxiv/papers/1907/1907.12182.pdf
(5) Evaluation of Data Management Systems for Geospatial Big Data https://core.ac.uk/download/pdf/297024646.pdf
(6) Geospatial Big Data: Challenges and Opportunities https://www.iqytechnicalcollege.com/GIS610%20arge%20spatial%20datasets%20analysis.pdf

(7) http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m16c/2021/12.09.12.04/doc/Goncalves_Spatial.pdf