*International Journal of Modern Computation, Information, and Communication Technology*
*2022;5(2):33-39.*
***ISSN: 2581-5954***

# A Data Analytic Model for Remote Sensing Data Using Data Mining Techniques

Anand M[1*], S. Babu[1]

[1]*Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.*

[*]*Corresponding author: am4639@srmist.edu.in*

**Abstract.** "Big Data" is a term used to describe the massive quantity of real-time data generated by the digital world's remote sensing assets. Data from remote sensing is far more complex than it appears at first glance; thus, retrieving relevant information in an effective manner can lead a system to severe computing issues, such as analysis and storage wherever data are remotely acquired. Accordingly, a design specification that allows for both real-time and offline data processing is needed considering the foregoing considerations. So, in this research, we offer a real-time data mining framework for remote sensing satellites. Ubiquitous Big Data collecting, processing, and decision-making units are the three key components of the suggested architecture. Using the suggested architecture, only valuable data will be divided up, balanced out, and processed in parallel. Consequently, real-time remotely sensed Big Data employing Earth observatory systems may be efficiently analysed using data mining approaches. As an additional benefit, the suggested architecture includes the capacity of retaining raw data to undertake offline analysis on huge dumps, if necessary. Big data from Earth observation satellites are analysed using Hadoop for land and water areas.

**Keywords:** Data mining, big data, decision making, remote sensing, data analysis

## INTRODUCTION

There has been a recent surge in interest in Big Data and its analysis, mostly due to the vast number of research difficulties directly connected to real-world applications, such as model-based modelling and processing, query-based querying, data mining, and wide distribution. "Big Data" refers to types of data sets that contain formless data and are stored in the data layer of technological computer applications and the Internet [1]. Large scale data, which refers to the magnitude and the centralized data; computation complexity (e.g. Big Data); sustain harvesting conversion loading (ETL) method from low, raw data to well-thought-out data; and development of uncomplicated applications are all characteristics of the information stored in the substrate surface of all these multi - paradigm numerical computing application scenarios [2].

There are many different types of big data that are created these days. These include online transactions and transactions including video/audio and/or email as well as click streams, logs, postings and data from various social media platforms and scientific databases. It is difficult to manage, create, store, share, process, analyse and display this data using standard database software tools since the databases they are stored in have grown enormously. Remote data collection, processing, analysis, and management have been transformed by advances in Big Data technology and computer technology during the past decade. Sensors in the Earth and Planetary Observatory System, particularly the most recently constructed ones, are continuously generating data streams. Furthermore, the bulk of work has been done in the many disciplines of remote sensing satellite visual information, including pattern recognition, horizontal stripe background subtraction, region resemblance dependent threshold method, and brightness gradient approach [3].

In this study, we referred to "Big Data," which is pushing us into a new world of issues, as "continuous stream of data" or "large volume offline data." It is imperative that scientists comprehend the repercussions of transforming remotely sensed data into scientific knowledge. There is a growing desire from both individuals and businesses for a more effective way to gather, analyse, and retain remote access data because of the rising volume of data. Finding, recognising, analysing, and citing data in the context of Big Data is, in some ways, more difficult. When dealing with enormous amounts of data, it's necessary to automate the process because of the complexity of the data structures and semantics that must be articulated in a computer-readable format [4].

*Received: 06.04.2022 Revised: 27.05.2022 Accepted: 13.06.2022*

**FIGURE 1.** Sample Remote Sensing Image Data

However, in order to construct a database based on basic data, a method must be devised. It's possible that other methods of storing the same data exist. In such circumstances, the design may be superior to others in some processes but may have disadvantages in others. Relational database providers have responded to these demands by providing a variety of analytical platforms. Software-only platforms can be found, as well as services that are hosted by other parties [5]. Sensors, for example, may generate enormous amounts of raw data across remote access networks. Most data are of little interest and may be removed or compacted by magnitudes in the first stage, which we call data gathering.

The valuable information is not discarded while applying these filters. Is it appropriate to preserve the name of the firm associated with the information in fresh reports? The report isn't essential; do we need only a portion of it to identify this specific individual? The second problem is the production of correct metadata that describes the data's content and how it was gathered and evaluated by default. Because we might have to know the origin of each piece of remote access data, this type of information is difficult to decipher. In most cases, remote-area data is not in a format suitable for analysis. It is thus necessary to collect the relevant information from a variety of sources and offer it in an organised manner that is appropriate for further investigation [6].

Big Data is being used in a variety of ways, including a reduction of the data set to a single-class label for easier analysis. However, this is a long cry from the truth, since we occasionally must deal with inaccurate or incomplete data. This research provides a satellite imagery Big Data analytical framework for real-time and offline data analysis to meet the requirements. The data are initially pre-processed remotely so that the machines can read them. After then, the Earth Base Station receives this important data and processes it further. For example, the Earth Base Station processes both real-time and archived data. It is possible to store offline information on an external storage device. While real-time data is sent straight to the membrane separation and load balancing server, where a filtering algorithm pulls crucial data out from Big Data, an offline backup and recovery device is used to store the data for later use.

The load balancer, on the other hand, distributes the real-time data equally among the servers. In addition to filtering and balancing the traffic, the filtration as well as load-balancing virtual machine is used to improve the efficiency of the system. Parallel servers then process and send the filtered data to the data gathering unit for comparison by the determination and analysing server. Data from sensors and the network may be accessed using the suggested architecture. MapReduce programming is used to build the suggested architecture and methods in Hadoop utilising data from the Earth Observatory's remote sensing network.

## EXISTING WORKS

To detect patterns, categorise information, and retrieve features from big data sets, data mining methods are employed by data scientists and analysts. A statistical agency's massive survey datasets may be analysed using these approaches, which are often used in the commercial world including market analysis, identity verification, and relationship management [7]. Despite the widespread usage of massive datasets in many statistical organisations, data mining methods have not been generally applied to improve official statistics. Data mining techniques may be used to distil or expose hidden information in massive databases [8].

Classifying data into subsets, predicting outcomes, clustering records into comparable subgroups, or assigning predicted values for just some dimension to records are only some of the strategies available. In remote sensing, information about a target is gathered without coming into direct touch with the target [9]. Remote sensing programmes assist official agricultural data in several nations along with the EU-25, Japan, Pakistan, and several emerging economies, Thailand, and North America. These countries all have remote

sensing programmes. Satellite photography, a kind of remote sensing, is used by NASS to acquire unique, timely, and comprehensive land cover categorization [10]. Images of the Earth's surface are captured by an Indian Upgraded Wide Field Spectrometer.

There is a central bank of images used by numerous government agencies, including NASS, that NASS purchases these images from. Research & Design Firm's Spatial Analysis Section Of the research (SARS) employs a commercial software package to handle data, categorise pixels and develop visual products and estimate acreage. Annual geospatial material delivery to stakeholders has been improved thanks to SARS's ongoing remote sensing methodology development [11].

When it comes to improving survey data collecting, processing, estimate, and distribution, these strategies may be quite beneficial when used in new and creative ways. Based on the greatest difference between a group of variables and their target variable, decision tree models utilise an algorithm to separate data. Segmentation can be determined using algorithms like the chi-square automated interaction detection technique. Sequentially, the segregation is carried out [12]. Observations are sent to the appropriate node using the splitting rule, which determines which variables and values should be separated. It will continue this method at each new node, sending the observations in that node down the proper branch for that node's successor. Leaves are the nodes at the end of a tree.

Consequently, a tree-like shape emerges. Forecasting whether someone will practice tennis on any given day is the main goal. A chi-square test or logistic regression may be used as a standard strategy [13]. The decision tree, a data mining tool, may also be used to anticipate the best circumstances for playing golf. In the 2007 Census, reaction likelihood groups were employed to represent grading modification cells in the decision tree models. Census mail list (CML) operations were divided into subgroups with homogenous response propensities based on variables such as employee sex and ethnicity, farm type and size. These groups were then given their own set of non-response weights [14].

Records on the original CML which did not reflect farming operations were also identified using decision tree models. Operation locations, past gross revenues, and other characteristics were utilised to identify lower-probability agricultural activities from a list of a set of criteria that included record source, length of time record had been on NASS list frame, location [15], and previous gross receipts. Operations with lesser chances of registering as a farmed in regions with bigger mailing lists than intended were eliminated from the CML. Data collecting expenses were lowered by implementing this programme, which increased total census processing efficiency [16].

## PROPOSED SYSTEM

There has been an exponential rise in the amount of data being created in the digital world. A large volume of data cannot be effectively analysed and stored using present techniques and technologies, because they are unwilling to extract the necessary sample data sets. Because of this, we'll need to create an architectural framework that allows us to examine data from both distant and offline sources simultaneously. An advantage over the competition arises when a company can extract all of the important information included in the Big Data rather than just a sample. Insight and better decision-making are made possible thanks to the use of Big Data analytics.

To use Big Data effectively, paradigm shifts are a must. We've outlined a few situations wherein Big Data could make a difference to back up our arguments [17-18]. Many data sources, such as remote access satellites that supervise earth's qualities [measurement results set (MDS) satellite data such as images], sensors that oversee clean air and water, geological situations, and the proportion of CO2 and other gases in the atmosphere among other things, must be used to fully understand the environment. Finding out about CO2 emission, greenhouse effect changes, and temperature changes may be accomplished by connecting all the scattered data.

Patients' medical histories, prescriptions, and other facts are collected by medical professionals in healthcare settings. Pharmaceutical businesses keep track of all the information listed above. In other cases, professionals are unable to prove a connection between this data and other information, resulting in a lack of critical information. Analytic tools for organising and retrieving relevant evidence from Big Data can provide insight into the hereditary origins of disease, which can be used to develop tailored medicine. A cost-effective parallel data collecting method for Earth observatory system growth is promoted by remote sensing.

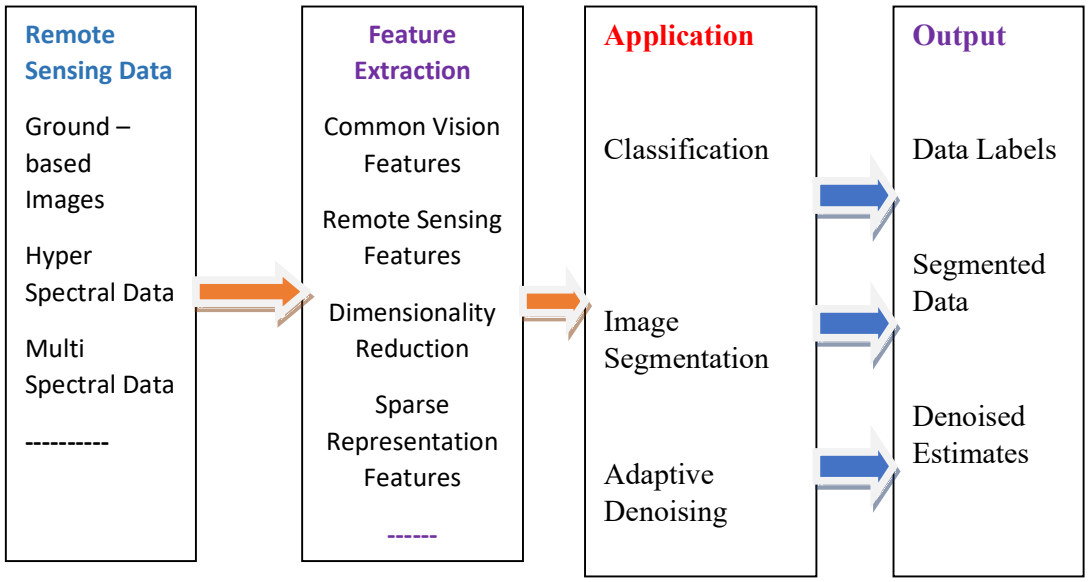


**FIGURE 2.** Proposed System Architecture

The Astronomy and Space Research Society initially certified this method as the norm for parallel computing in this specific context. When Earth observation satellites began to incorporate better capabilities for Big Data collecting, it soon became clear that standard data processing technology couldn't handle the amount of data that was being generated. As a result, vast amounts of data required parallel processing to be analysed effectively. Since the planned RSDU would collect data from several satellites all over the world, it has been integrated into the satellite imagery Big Data architecture.

The obtained raw data may be affected because of ambient gas and dust particle scattering and absorption. We have faith in the satellite's ability to fix the data that was incorrect. However, the remotely sensed data satellite employs Doppler or SPECAN algorithms to transform the raw data into an image format. Satellite remote sensing pre-processes information in a variety of scenarios to combine data from multiple sources that not only reduces storage costs but also enhances analysis accuracy. Techniques for relational data preparation include data integration and data cleansing.

The acquired data are transferred to a ground control station via the downlink channel after the pre-processing step. In a wireless environment, this message can be made directly or via a relay spacecraft with the proper tracking dish and communication connection. Many factors, such as platform motion, orientation, planet deformation, non-homogeneous lighting, sensor characteristic fluctuations (among others), must be considered while correcting data. Direct communication link is used to send the data to Earth Base Station for additional processing.

There are two types of data processing: genuine Big Numerical computation and offline Big Data analysis when processing data offline, the Planet Base Station sends it to the data centre, where it is stored. This information is then utilised in future research. As a result, the real-time processing server receives the data directly from the filtering and load balancing server, rather than having to store them.

## RESULTS AND DISCUSSION

Earth observatory data may be analysed using the suggested architecture for both offline and online traffic. It is assumed that the data is large and difficult to process on a single server. There is a constant stream of high-speed data flowing from the satellite. Big Data necessitates the use of specialised algorithms to process, evaluate, and make decisions on it. Remote sensing data is used here to identify land, marine, and ice areas. A decision-making algorithm based on the suggested architecture has been developed.

**TABLE 1.** Prediction Accuracy Analysis

| Product | Total Blocks | Detected Blocks | Detected Sea Blocks | TP % | FP % |
|---|---|---|---|---|---|
| 1 | 600 | 544 | 56 | 98.9 | 0.16 |
| 2 | 840 | 840 | 0 | 100 | 0 |
| 3 | 495 | 475 | 20 | 99.54 | 3.21 |
| 4 | 2793 | 2704 | 89 | 93.2 | 0.79 |
| 5 | 3420 | 3340 | 80 | 97.65 | 10.3 |
| Overall | 8148 | 7903 | 245 | 97.858 | 2.892 |

Using satellite-sensed BigData from the European Space Agency, we begin by analysing the ice, land, and sea. In light of these findings, we developed a set of techniques for managing, analyzing, evaluating, and choosing on ocean, terrestrial, and glacier area) utilising our suggested architecture for satellite imagery Big Data pictures. Satellite data were analysed using the BEAM VISAT version6.0 and EnviView software packages. Envisat mission data files can be easily deciphered using Beam VISAT and EnviView.

**TABLE 2.** Sample Analysis

| Product | No. of Blocks Taken | No. of Sample Values (Pixels in Each Block | Minimum Mean | Maximum Mean |
|---|---|---|---|---|
| 01 | 20 | 20 000 – 30 000 | 1604 | 3067 |
| 02 | 20 | 20 000 – 30 000 | 408 | 603 |
| 03 | 20 | 20 000 – 30 000 | 1351 | 3050 |
| 04 | 20 | 20 000 – 30 000 | 928 | 1439 |
| 05 | 20 | 20 000 – 30 000 | 928 | 2372 |

As a result, they may be used for rudimentary statistical analysis. We couldn't utilise these technologies for rigorous calculations and faster functioning of Big Data collections. The implementation of the suggested technique makes use of Apache Hadoop's Map Reduce software running on a single node configuration since Hadoop offers parallel, high-performance computing on many servers. That's why vast amounts of remote sensing picture data may be effectively analysed with it.

For complex analysis, algorithm creation, and testing, the suggested architecture makes use of a comparable load balancing mechanism: Hadoop. Advanced synthetic reconfigurable radar (ASAR) and intermediate resolution spectrometry (MRIS) devices or sensors were used to examine ENVISAT mission data sets (e.g. products). Because the ENVISAT lunar project has been continually delivering worldwide measurements of the globe, including sea, land, ice, and forest, since 2002, the primary focus is on ENVISAT ASAR data sets. It can detect the earth's surface using 10 different instrument data sets.

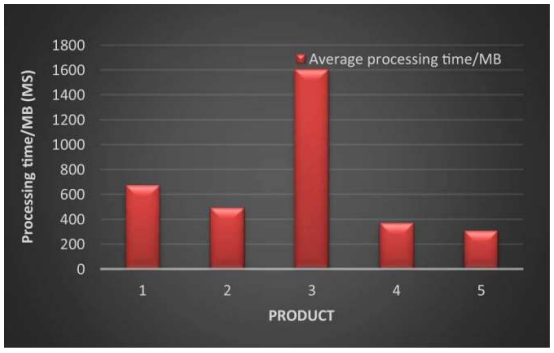


**FIGURE 3.** Data Processing Analysis

Only two instruments, namely ASAR and MERIS, have been considered thus far. MERIS, ASAR, and a few additional ENVISAT sensors provide more data for analysis, but here we focus on the five most common ASAR data sets provided by ESA. Earth's surface was covered by five remote sensing data sets that included sea, desert; forest; beach; and city areas. The data came from all over the world, including China, Ukraine, Italy, the Western Sahara, Morocco, and Sudan, as well as South Africa and Spain.

They were all collected at various points in history. APM (altering polarisation medium-resolution picture), WSM (wide swath medium-resolution image), and GMM (global monitoring mode) are three examples of these products (GMI). Various ASAR versions are utilised to collect data. This product's ID/name, kind, sensing time, version number, purpose, SPH descriptors and covered region are all listed in the complete product description.

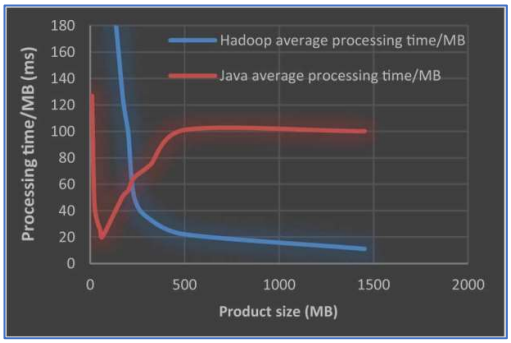


**FIGURE 4.** Performance Analysis

In contrast to the other data sets, the MPH as well as SPH information are in ASCII format. There is a series of recordings and one or maybe more categories of data for every session in MDS, ADS, and GAD. Many entries in our MDS have a correspondingly large number of fields. During our analysis, each row of the satellite picture corresponds to a recording in the MDS.

To test the suggested method, we employ three well-known datasets of remote sensing images that can be accessed by the public. Both models' average error rates are compared on a class level, in addition to comparing their overall average mistake rates. Based on these results, each model's efficiency may be evaluated in terms of its ability to correctly classify a given dataset's common subsets. This model's improved accuracy can be attributed to the additional information provided by the images that were used to create it.

## CONCLUSION

Big Data processing for sensor technologies is the focus of this research. For decision-making, the suggested architecture processed and analysed remote sensing Big Data in real-time and offline using data mining approaches. Three primary components make up the suggested architecture. Depending on the analysis needed, various units execute algorithms at various levels of the architecture. A general, application-independent real-time Big Data architecture may be utilised for any kind of distant sensor Big Data processing. In addition, any unnecessary data is discarded to execute filtering, division, and parallel processing on just the most important information. Using these methods, real-time Big Data analysis of remote sensing data is much more efficient. Data from remote sensing sensors may be analysed with the aid of the methods introduced in this study for every unit and component. The data mining approach proposed in this article handles the image data obtained from remote sensing data set as explained above and outperforms the existing approach.

## REFERENCES

[1]. P. Chandarana and M. Vijayalakshmi, 2014, "Big data analytics frameworks," *In 2014 Int. Conf. on Circuits, Systems, Comm. and Information Tech. Applications (CSCITA),* pp. 430-434.

[2]. S. N. Kalluri, Z. Zhang, J. JaJa, S. Liang and J. R. Townshend, 2001, "Characterizing land surface anisotropy from AVHRR data at a global scale using high performance computing," *Int. J. of Remote Sensing*, **22(11)**, pp. 2171-2191.

[3]. J. Dittrich, J. A. and Quiané-Ruiz, 2012, "Efficient big data processing in Hadoop MapReduce," *Proc. of the VLDB Endowment*, **5(12)**, pp. 2014-2015.

[4]. X. Li, F. Zhang and Y. Wang, 2013, "Research on big data architecture, key technologies and its measures," *In2013 IEEE 11th Int. Conf. on Dependable, Autonomic and Secure Computing*, pp. 1-4.

[5]. S. Marchal, X. Jiang, R. State and T. Engel, 2014, "A big data architecture for large scale security monitoring," *In 2014 IEEE Int. Congress on Big Data*, pp. 56-63.

[6]. M. Mayilvaganan and M. Sabitha, 2013, "A cloud-based architecture for Big-Data analytics in smart grid: A proposal," *In 2013 IEEE Int. Conf. on Computational Intelligence and Computing Res.,* pp. 1-4.

[7]. A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri and M. Marconcini, 2009, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, **113**, pp. 110-122.

[8]. L. Ramaswamy, V. Lawson and S. V. Gogineni, 2013, "Towards a quality-centric big data architecture for federated sensor services," *In 2013 IEEE Int. Congress on Big Data*, pp. 86-93.

[9]. Wikibon Blog. 2014, [2310]. Big Data Statistics [Online]. Available: wikibon.org/blog/big-data-statistics/

[10]. X. Yi, F. Liu, J. Liu and H. Jin, 2014, "Building a network highway for big data: architecture and challenges," *IEEE Network*, **28(4)**, pp. 5-13.

[11]. M. Yang, H. Mei, Y. Yang and D. Huang, 2017, "Efficient storage method for massive remote sensing image via spark-based pyramid model," *Int. J. of Innovative Computing, Information and Control*, **13(6)**, pp. 1915-1929.

[12]. X. Zhao, J. Guo, Y. Zhang and Y. Wu, 2021, "Memory-Augmented Transformer for Remote Sensing Image Semantic Segmentation," *Remote Sensing*, **13(22)**, pp. 4518-4524.

[13]. D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du and B. Zhang, 2020, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. on Geoscience and Remote Sensing*, **59(5)**, pp. 4340-4354.

[14]. J. Geng, J. Fan, H. Wang, X. Ma, B. Li and F. Chen, 2015, "High-resolution SAR image classification via deep convolutional autoencoders," *IEEE Geoscience and Remote Sensing Letters*, **12(11)**, pp. 2351-2355.

[15]. B. Zhao, Y. Zhong, G. S. Xia and L. Zhang, 2016, "Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high spatial resolution remote sensing imagery," *IEEE Trans. Geoscience and Remote Sensing,* **54(4)**, pp. 2108-2123.

[16]. B. Pattanaik and S. Murugan, 2017, "Cascaded H-Bridge Seven Level Inverter using Carrier Phase Shifted PWM with Reduced DC sources," *Int. J. of MC Square Scientific Res,* **9(3)**, pp. 30-39.

[17]. M. U. A. Ayoobkhan and L. A. K. S. Ali, 2022, "Web page recommendation system by integrating ontology and stemming algorithm," *Int. J. Adv. Sig. Img. Sci,* **8(1)**, pp. 9–16.

[18]. R. S. Kadurka and H. Kanakalla, 2021, "Automated Bird Detection in Audio Recordings By A Signal Processing Perspective," *Int. J. Adv. Sig. Img. Sci,* **7(2)**, pp. 11–20.