A NOVEL BIG DATA PROCESSING FRAME-WORK FOR MASSIVE REMOTE SENSING WITH CLOUD COMPUTING TECHNOLOGIES

[1]Patil Manik Rao, [2]Padma Parshapu, [3]C.Nagesh, [4]J GOPAIAH

[1],[2]Associate.Professor, Dept of IT, Guru Nanak Institutions Technical Campus, Hyderabad
[3]Associate.Professor, Dept of CSE, Tadipatri Engineering College, Tadipatri, Anantapuram
[4]Associate.Professor, Dept of CSE, SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY, IBRAHIMPTNAM ,HYDERABAD, TELANGANA, INDIA

Abstract

Consistently an enormous number of Earth perception (EP) spaceborne and airborne sensors from a wide range of nations give a huge measure of remotely detected data. Those data are utilized for various applications, for example, normal danger observing, worldwide environmental change, urban arranging, and so forth. The applications are data-driven and generally interdisciplinary. In light of this it can genuinely be expressed that we are currently living in the time of big remote sensing data. Moreover, these data are turning into a monetary resource and another significant asset in numerous applications. We consider a dispersed system of cloud web administrations for processing satellite data, which gives data processing offices to Earth remote sensing inside the SaaS model. For these Big data investigative engineering is proposed. The design includes three fundamental units, for example, 1) remote sensing Big Data acquisition unit (RSDU); 2) data processing unit (DPU); and 3) data analysis decision unit (DADU). RSDU secures data from the sensors and sends this data to the Base Station. DPU gives productive processing of Data by giving filtration, load adjusting, and parallel processing. DADU is liable for the arrangement, stockpiling of the outcomes, and age of a decision dependent on the outcomes got from DPU.

KEYWORDS:

Big Data, data analysis decision unit (DADU), data processing unit (DPU), high-performance computing, land and sea area, offline, real-time, remote senses, remote sensing Big Data acquisition unit (RSDU).

I. INTRODUCTION

As moving data generators, individuals make data consistently. We are altogether associated by sharing data from interpersonal organizations, smart gadgets, and so forth. Remote sensing gadgets have been generally used to watch our planet from different points of view and to make our lives simpler. It isn't misrepresented to state that the entire Earth has now been made computerized. Along these lines, the digitized Earth in addition to the moving data generators are the principle on-screen characters for big data in remote sensing, which can be utilized to make governments increasingly proficient (e.g., improving administrations like police, medicinal services, and transportation) and additionally for business, i.e., to improve decision making, fabricating, item advancement, customer experience and administration, etcAs revealed by IBM, 2.5 quintillion bytes of data are currently produced each day. As it were, "90% of the data on the planet today has been made over the most recent two years alone."1 We are genuinely living in the big data age, and now government pioneers, ventures, and charitable associations are rapidly realizing that it is essential to gather big data in various settings. Be that as it may, there still exists a typical issue identified with how we can pick up bits of knowledge into big data. This issue is a problem: On one hand, an abundance of big data can bring us big opportunities. Then again, despite everything we don't have a clue how to saddle such a big measure of data with colossal multifaceted nature, assorted variety, and heterogeneity, yet with high potential qualities. This makes the data extremely hard to process and break down in a sensible time. Big data can be predominantly described by three highlights: volume, assortment, and speed characterized as three "V" measurements by Meta Group (presently Gartner) in 2001 [1]. It is significant that "esteem" is a significant nature of big data, however it's anything but a characterizing trademark. Big remote sensing data can be depicted by its measurements (alluded hereinafter as 3Vs).

1) The chronicled data are portrayed by their expanding volume, from terabytes (TB ¼ 1024 GB) to petabytes (PB ¼ 1024 TB), and even to exabytes (EB ¼ 1024 PB). For example, an enormous measure of remote sensing data are presently uninhibitedly accessible from the NASA Open Government Initiative.3 Only one of NASA files, the Earth Science Data and Information System (ESDIS), holds 7.5 PB of data with about 7000 interesting data sets and 1.5 million clients in 2013 [2]. This volume just contains in-space remote sensing data.

2) as far as assortment, we can see since big remote sensing data comprise of multisource (laser, radar, optical, and so on.), multitemporal (gathered on various dates), and multiresolution (diverse spatial goals) remote sensing data, just as data from various controls relying upon a few application areas [3].

3) The speed of big data in remote sensing includes not just the age of data at a quickly developing rate yet in addition the productivity of data processing and analysis. At the end of the day, the data ought to be dissected in an (almost) real or a sensible time to accomplish a given undertaking, e.g., seconds can spare a huge number of lives in a quake.

In spite of the fact that the 3Vs can depict big data, we think about that it isn't important for big data in remote sensing to fulfill all the three V measurements. For example, anybody of volume and speed, volume and assortment, or assortment and speed would already be able to characterize a big data issue. With the exception of the basic difficulties of big data described by the 3Vs, there are different difficulties for remote sensing applications, for example, extensibility to incorporating various divergent administration frameworks for various satellites for a remote sensing data focus [4]. Of specific significance is the estimation of the data, a significant quality covered up in the big data. Data processing strategies can be used to find such esteem, and then the estimation of big data can be realized in a real remote sensing application.

In this way, to all the more likely understand big data, three points of view ought to be bound together, i.e., owning data, data applications, and data strategies. In the paper, a trinity system is proposed to all the more likely understand big data with regards to remote sensing applications. Every one of the features of such trinity share normal difficulties and alternate points of view have singular difficulties of its own.

II. RELATED WORK

UNDERSTANDING BIG DATA FOR REMOTE SENSING

From a general viewpoint, we can understand big data as having various undertones in regards to the individuals who possess the big data, the individuals who can process and break down the big data, and the individuals who use the big data. Appropriately, various data techniques might be abused to handle big data difficulties to proficiently infer the estimation of those data. In the accompanying, a trinity (three out of one) is talked about for the understanding of big data (with specific spotlight on remote sensing applications). Here, we recognize three aspects for understanding big data,i.e., owning data, data techniques, and data applications, which contribute together to a solitary big data life cycle.

A. First Facet: Owning Data

This is a significant part of big data dependent on which we can recognize applications and use or plan legitimate data strategies to address a real issue (e.g., a remote sensing issue). The comparing opportunities depend on the way that increasingly various data can be gained by astute gadgets where the vast majority of the people approach the web presently to get both individual and moving data generators. In like manner, data esteems can be gotten from those mind boggling, assorted, heterogeneous, and high-dimensional remote sensing data and other data from the internet. Be that as it may, big difficulties emerge at each progression while getting and sorting out big remote sensing data. For example, remote sensing data are procured from satellites, planes, or other sensing gadgets while different types of data are recovered from the internet. Remote sensing data are preprocessed by geometric and radiometric rectification, georeferencing, commotion evacuation, and so on [18], and the data from the internet ought to be cleaned to lessen blunders and clamor, in which data quality can be improved. Remote sensing data ought to be conveyed from satellites to ground stations, and from ground stations to clients. Other related issues are data pressure, data filing, data recovery, data rights, and assurance, and so forth. We underline that data are of no incentive until they are used for applications. The key contrast between customary data and big data is the means by which to distinguish the correct data sets and how to join them to take care of a difficult or novel issue.

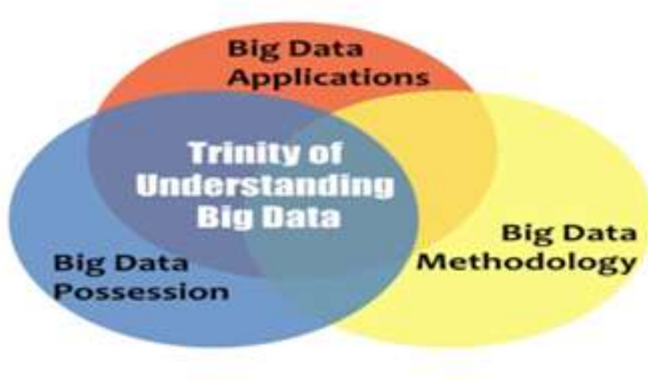


Fig. 1. The trinity idea of big data

There are normal and various difficulties in the individual aspects of understanding big data, which are point by point straightaway.

B. Second Facet: Big Data Methodologies

A big data procedure ought to be intended to efficiently address big data issues from various remote sensing spaces. Such a procedure is utilized to plan new data techniques for big remote sensing data arrangement, data sending, data extraction, data displaying, data combination, data perception, and data elucidation. These viewpoints are especially significant in remote sensing applications, in which preprocessing steps are as similarly significant as data extraction steps. Be that as it may, data processing and analysis speak to a multistep pipeline and data-driven strategies could be fundamentally not quite the same as the perspective of explicit applications and areas. Due to the previously mentioned heterogeneity and high dimensionality of big data in remote sensing, we likewise face significant computational and factual moves identified with processing adaptability, commotion aggregation, fake relationship, accidental endogeneity, and estimation blunders [19], [20]. These difficulties require new computational and measurable strategies to handle big data analysis and processing. The analysis and processing procedures are data-driven and can profit by speculations and techniques from the fields of insights, AI, design acknowledgment, man-made brainpower, data mining, and so on. Area information is another pivotal angle that ought to be firmly connected to data analysis.

C. Third Facet: Big Data Applications

The principle objective in big data applications is to distinguish the correct data to take care of the issues at hand, which are hard to be tended to or generally can't be controlled by conventional remote sensing data. At that point, the following issue is the means by which to gather, sort out, and use these big data to manage real remote sensing issues. To distinguish the correct data, we ought to be firmly connected to the principal aspect of understanding big data. As it were, to bridle big data right off the bat one ought to acquire data from the related data operators (or, as a rule, data industry or association). To get to the data, coordinated effort crosswise over areas or associations ought to be effectively considered. This is one of the significant difficulties in remote sensing applications. Subsequent to getting the correct data, for example, remote sensing data, literary data, and pictures from informal organizations, creative data systems ought to be created to find, realize, and exhibit the estimation of big data for remote sensing applications.

The difficulties of big data in remote sensing include not just managing high volumes of data [1]. Specifically, challenges on data acquisition, stockpiling, the executives, and analysis are likewise identified with remote sensing issues including big data. In this area, we especially examine the difficulties of big data in remote sensing which include the various aspects of understanding big data in the past segment.

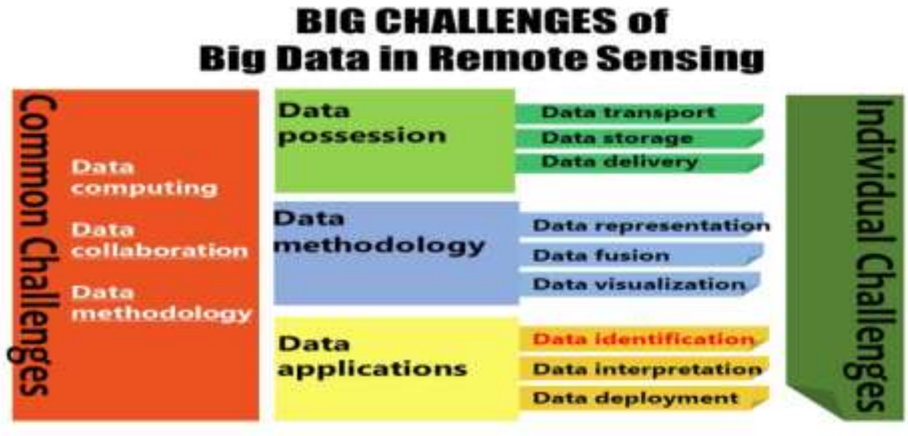


Fig. 2. ***A summary of the challenges introduced by big data.***

From alternate points of view of understanding big data, we are confronting big difficulties in utilizing the worth that data bring to the table. In the three features, similar difficulties are shared, for example, data computing, data joint effort, and data philosophies for various applications; meanwhile, we are confronting various difficulties in the individual aspects of understanding big data. Fig. 2 outlines the normal and various difficulties, which are portrayed in detail in consequent areas.

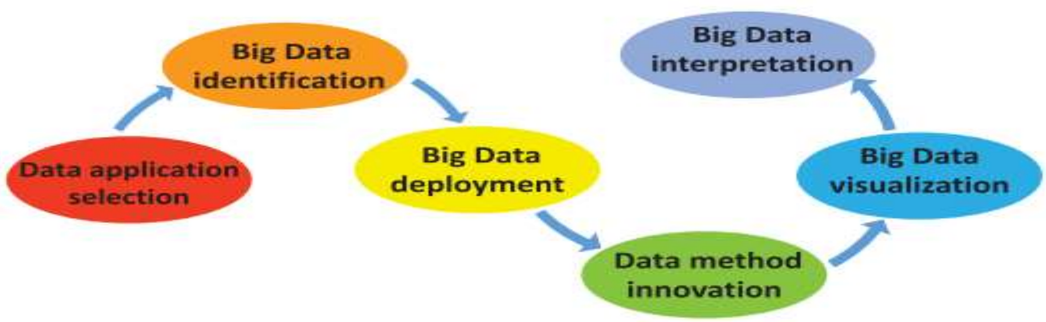


Fig. 3. ***Life cycle to address big data tasks in remote sensing applications.***

a couple of essential difficulties are talked about with regards to planning a big remote sensing data lifecycle (see Fig. 3). Subsequent to understanding the business need(e.g., a remote sensing application including big data), some significant advances are to recognize the correct sort of data crosswise over various orders, to convey the big data, to use or plan imaginative data techniques, and at long last to picture and decipher the got outcomes. Here, data techniques incorporate data analysis, data demonstrating, data processing, and so on.

High-performance scalable computing

The created system expect the utilization of high-performance SSCC assets for time-expending counts. The high-performance computing subsystem is actualized dependent on the SSCCIP (Siberian Scientific Computing Center – Image Processing) programming created by the creators for coordinating remote high-performance PCs into the processing and analysis of satellite data [6]. SSCCIP comprises of four segments:

- customer segment - the administrator's work environment in the Windows condition (it gives a GUI to designing a solicitation for data processing, following its execution results, and envisioning them.

- server part - Unix-application running on a supercomputer (it gives the processing of the customer segment solicitation and its dispatching);

- correspondence part - arrange trade among customer and server segments (which is executed dependent on the safe SSH convention);- computational segment – in light of the high-performance picture processing libraryParImProLib [7] created by the creators. The library gives estimations on the NKS-30T + GPU SSCC half breed bunch: between hub trade is given by the MPI interface; and computations on the GPU (right now being developed), by Nvidia CUDA innovation. Later on, it is wanted to expand the library with the capacity of processing on the Intel XeonPhi CPU to empower it to perform estimations on the new SSCC group NKS-1P. Note that the SSCCIP framework and the ParImProLib library are likewise worked as a system: their source code is sorted out with the goal that activities basic to a specific kind of processing calculations are isolated from the tasks explicit to specific calculations of this sort. In [6], it was demonstrated that this methodology permits keeping up high performance of the program code while altogether rearranging the way toward adding new processing calculations to the framework. To utilize SSCCIP as a subsystem of a cloud web administration, its customer segment will be improved with the capacity to perform self-sufficient processing with an interface indistinguishable from the interface gave via PlanetaMonitoring. Subsequently, the structure being created will separate the webserver from the exceptional attributes of a specific PC.

III. PROPOSAL WORK

REMOTE SENSING BIG DATA ANALYTICS ARCHITECTURE

The term Big Data covers various advancements equivalent to distributed computing. The contribution of Big Data originates from informal organizations (Facebook, Twitter, LinkedIn, and so on.), Web servers, satellite symbolism, tangible data, banking exchanges, and so on.

Remote Sensing Big Data Acquisition Unit (RSDU)

The RSDU in remote sensing Big Data design accumulates the data from different satellites far and wide as appeared in. The got crude data might be twisted by dissipating and assimilation by different climatic gasses and residue particles. After the pre-processing phase,[9] the gathered data are transmitted to a ground station utilizing a downlink channel. This transmission is straightforwardly or through hand-off satellite with a proper following recieving wire and correspondence connect in a remote climate.

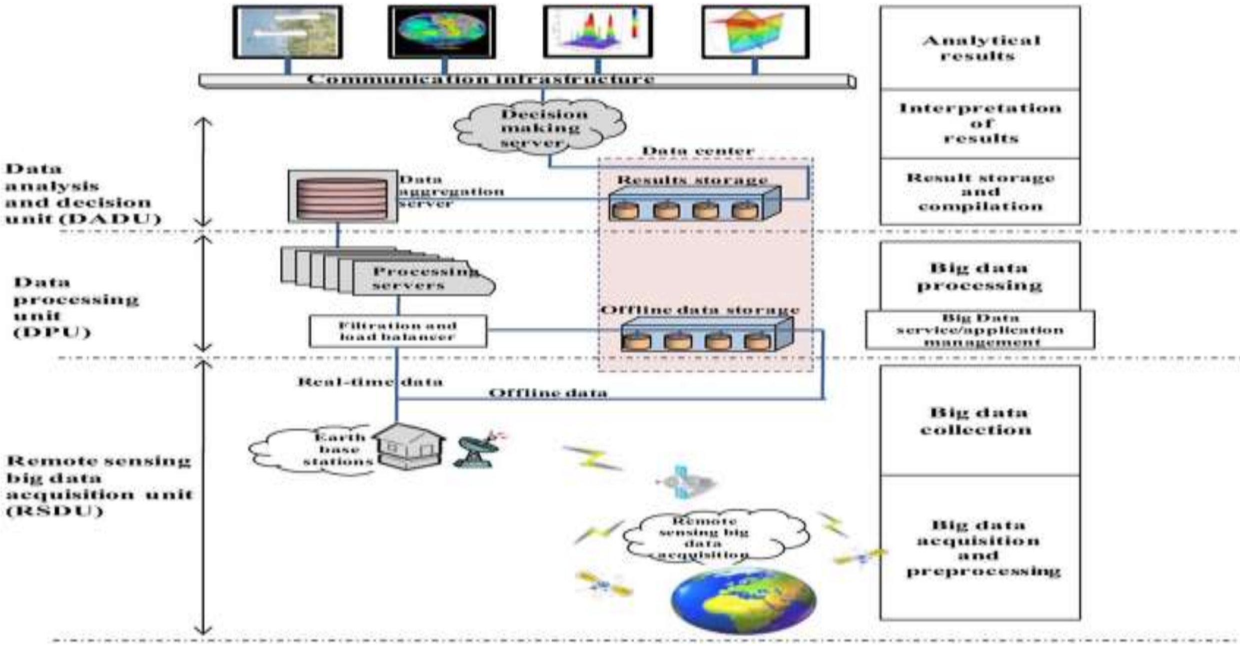


Fig .4.Communication of RSDU

We separated the data processing technique into two stages, for example, real-time Big Data processing and disconnected Big Data processing. On account of disconnected data processing, the Earth Base Station transmits the data to the data community for capacity. This data is then utilized for future examinations. In any case, in real-time data processing, the data are legitimately transmitted to the filtration and burden balancer server (FLBS), since putting away of approaching real-time data corrupts the performance of real-time processing.

Data Processing Unit (DPU)

In the data processing unit (DPU), the filtration and burden balancer server have two fundamental duties, for example, the Filtration of data and burden adjusting of processing power. Filtration distinguishes the valuable data for analysis since it just permits helpful data, while the remainder of the data are blocked and are disposed of. Thus, it brings about upgrading the performance of the entire proposed framework. The heap adjusting some portion of the server gives the office of partitioning the entire separated data into parts and allocate them to different processing servers. The filtration and burden adjusting calculation change from analysis to analysis; e.g., if there is just a requirement for analysis of sea wave and temperature data, the estimation of these depicted data is sifted through and is divided into parts. Each processing server has its calculation usage for processing approaching sections of data from FLBS. Each processing server makes factual estimations, any estimations, and performs other scientific or sensible errands to create moderate outcomes against each portion of data. Since these servers perform undertakings autonomously and in parallel, the performance proposed framework is significantly upgraded, and the outcomes against each fragment are created in real-time. The outcomes created by every server are then sent to the total server for aggregation, association, and put away for further processing.

Data Analysis and Decision Unit (DADU)

DADU contains three significant parts, for example, conglomeration and assemblage server, brings about capacity server(s), and decision-production server. At the point when the outcomes are prepared for assemblage, the processing servers in DPU send the incomplete outcomes to the collection and accumulation server, since the totaled outcomes are not in composed and arranged structure. Hence, there is a need to total the related outcomes and sorted out them into a legitimate structure for further processing and to store them. Total server stores the aggregated and sorted out outcomes into the outcome's stockpiling with the expectation that any server can utilize it as it can process whenever. The conglomeration server likewise sends a similar duplicate of that outcome to the decision-production server to process that outcome for settling on a decision.

IV. RESULTS ANALYSIS

a. Accuracy Evaluation

We first show in Fig 5 outwardly the dish honing results by utilizing the proposed dispersed parallel usage of the DNN-based strategy. We play out the DNN-based dish honing process by utilizing the proposed system that joins both MapReduce component and planning technique to investigate the ideal distributed computing arrangement. We likewise actualize the sequential variant of the DNN-based technique, in which all subtasks of the dish honing process are executed in sequential on a solitary gauge machine. Fig. 6(a) and (b) shows the HR multispectral picture and LR panchromatic picture, individually. Fig. 6(c) and (d) shows the intertwined picture by the sequential form DNN-based technique and the proposed strategy, separately. For correlation reason, the pseudocolor picture of the reference multispectral picture is likewise given in Fig. 6(e). We can see that the DNN-based strategy performs well in remote detecting picture combination, both in the sequential adaptation and the conveyed variant.

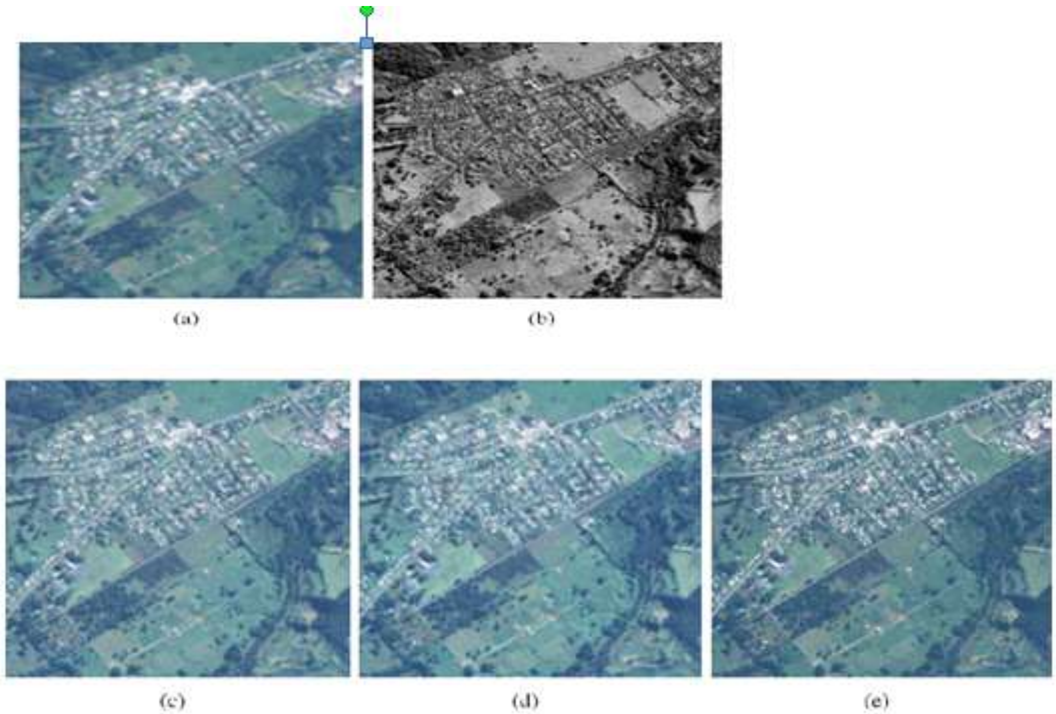


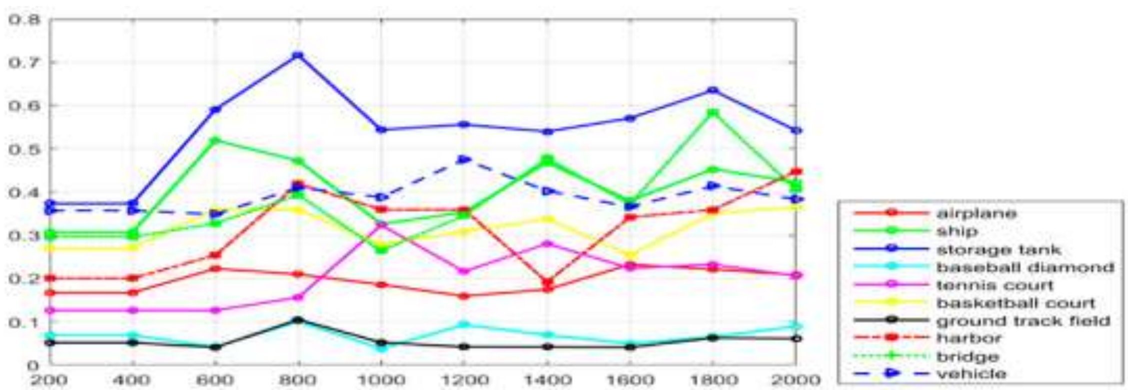
Fig. 5. QuickBird images and experimental results by the DNN-based pan-sharpening method. (a) LR multispectral image. (b) HR panchromatic image. (c) Fused image by the serial implementation of the DNN-based method. (d) Fused image by the proposed cloud computing framework. (e) Pseudocolor image of the reference multispectral image.

b. Computational Efficiency Evaluation

Having confirmed the picture combination exactness by dispersed handling of the container honing strategy, we further assess the computational productivity of the proposed streamlining structure. One key commitment of this paper is to improve computational productivity by settling on choices on both undertaking apportioning and errand assignments. We initially check that the proposed planning approach outflanks conventional sched-ulers by lessening the all out execution time. In light of a similar improvement system, we assessed the execution time of the DNN-based container honing stream by the proposed QEA-based booking calculation and contrast it and the outcome by cooperative calculation, which is a broadly utilized planning technique in existing distributed computing structures. The correlation brings about Fig. 7 show that for different quantities of laborers just as information measures, the execution time for the entire dish honing stream on the first QuickBird informational index has been diminished to various degrees. Note that he number of laborers on the flat hub indicates the greatest number of every single accessible specialist sent on Ten-sorFlowOnSpark. As announced in Fig. 7, the decrease in execution time ranges from 9.26% to 24.22%. The explanation is that in the cooperative technique, undertakings are haphazardly dispensed onto laborers, while the QEA-based planning calculation is fit for allocating a fitting number of laborers for partitionable assignments as per their calculation loads. Note that for every arrangement of laborer hubs, we rehash the cooperative booking methodology a few times and present the normal estimation of execution time.

We at that point assess the presentation of the proposed optimiza-tion system for handling remote detecting information of various sizes. For this reason, we mosaicked the first QuickBird informational index (a HR panchromatic picture and a LR multispectral picture) to deliver huge scale pictures. It is advantageous to bring up that because of the high computational intricacy of DNN preparing, even a little scale informational index would produce a lot of preparing information. To be explicit, for the first QuickBird informational collection, the measure of preparing information may arrive at 158 MB during the preparation organize and up to 556 MB during the reproduction arrange. Following a similar combination process, the measure of preparing information additionally develops relatively to the extents of mosaicked panchromatic and multispectral pictures, bringing about enormous scale preparing information of sizes 884 MB, 1.8, 3.6, 7.2, and 14.4 GB, individually. When playing out the DNN-based skillet honing stream, it would be an especially provoking errand to manage remotely detected large information on a solitary machine.

This perception is because of the way that with not many laborer hubs, the real errand calculation time rules the correspondence time among hubs. Be that as it may, in the event that we continue including more specialists, the speedup would be undermined. For instance, when the size of preparing information is 1.8 GB, utilizing 32 laborers prompts a 16.08× speedup, and the speedup has been improved to 18.64% by multiplying the quantity of laborers. Comparative speedup patterns concerning laborer check can be watched for all informational collections. The explanation is that with an expanding number of laborers, the correspondence overhead between the PS hub and specialist hubs additionally increments extensively. The measure of time spent in refreshing hub parameters and synchronizing hub employments is never again irrelevant contrasted with the measure of real calculation time. This perception additionally approves the significance of the advancement methodology in the proposed edge work to completely use the registering assets. Having considered the reliance of speedup upon the quantity of laborers, we further examine the connection between computational proficiency and information size.



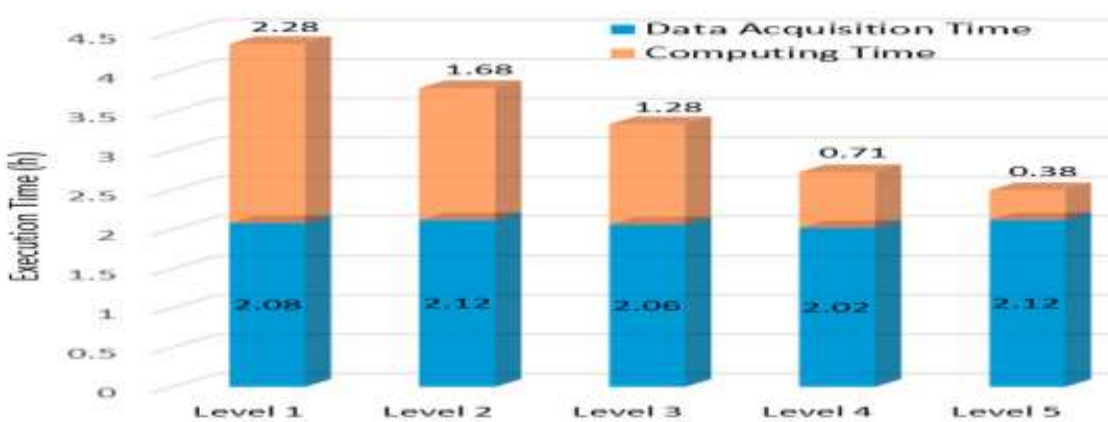


Fig. 6. Speedup achieved by the proposed method for various data sizes with different numbers of workers.

Fig. 7. Execution time comparison between the proposed scheduling algorithm and round-robin algorithm for different data sizes with 64 workers.

At long last, we check the appropriateness and adaptability of the proposed streamlining system when preparing huge scale remote detecting information. With the quantity of laborers fixed at 64, we further watch the adjustment in execution time for playing out the entire dish honing stream as the preparation information size increments. Fig. 9 exhibits the execution time for handling 884-MB, 1.8-, 3.6-, 7.2-, and 14.4-GB preparing information, separately. For correlation reason, the execution time insights for these informational indexes by cooperative methodology are additionally given. By utilizing QEA-based planning system, the execution time for preparing 884-MB information is 145.10 s. At the point when the information size increments by multiple times (3.6 GB), the execution time is 590.65 s, and the execution time for 14.4 GB information is 1663.93 s. The outcomes confirm the versatility of the proposed strategy when handling huge scale remote detecting information. All the more critically, contrasted and cooperative booking methodology, the QEA-based planning calculation can deliver top notch planning results, prompting up to 23.17% decreases in execution time.

V. CONCLUSION

The remote senses produce a lot of real-time data from the Satellite or the Aircraft with the assistance of the sensors. Presently a day there is an extraordinary demand added to the real-time big data for remote sensing applications, these data must be prepared and extricate the helpful data can prompt computational difficulties. In this paper, we talked about the proposed explanatory engineering of real-time big data for remote sensing applications. The proposed engineering is structured so that; it can dissect both offline just as real-time data in a productive manner. The proposed engineering for the remote sensing application, The three principle units includes the proposed design the three units are First, Remote sensing data acquisition unit (RSDU) takes the data from the satellite and sends to the Base Station, where processing begins in this unit. Second, Data processing unit (DPU) is the primary job in the engineering, the real-time data will process productively by separating, load adjusting and parallel processing and Third, Data Analysis and Decision unit (DADU) this unit is liable for the putting away the outcomes and creates the decision dependent on the aftereffects of the data processing unit. In this paper, the remote sensing data examined by each proposed unit to settle on the better decision making. For future work, this proposed design can use to figure for progressively complex data for decision making at a real-time of earth observatory, for example, fire location, tidal wave expectation, seismic tremor forecast, and so on. The engineering needs to make good with all applications for big data analysis.

REFERENCES

1. J. Sun et al., "An Efficient and Scalable Framework for Processing Remotely Sensed Big Data in Cloud Computing Environments," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 7, pp. 4294-4308, July 2019.
2. V.V. Asmus, A.A. Buchnev, V.A. Krovotyntsev, V.P. Pyatkin, G.I. Salov, Optoelectronics, Instrumentation, and Data Processing 54(3) 1–9 (2018)
3. P.A. Kim, P.A. Kalantaev, V.P. Pyatkin, Pattern Recognit. Image Anal. 25(2), 220–222 (2015)
4. A.A. Buchnev, P.A. Kim, V.P. Pyatkin, G.I. Salov, Journal of Siberian Federal University. Engineering and Technology 10(6), 741–746 (2017)
5. D. Laney, “3D data management: Controlling data volume, velocity, and variety," Gartner, Tech. Rep., Feb. 2001.
6. H. Ramapriyan, J. Brennan, J. Walter, and J. Behnke, “Managing big data: NASA tackles complex data challenges,” Earth Imaging J., 2013.
7. M. Chi, J. Shen, Z. Sun, F. Chen, and J. A. Benediktsson, “Oil spill detection based on web crawling images,” in Proc.IEEE Int. Geosci. Remote Sensing Symp.Que´bec, QC, Canada, Jul. 2014, pp. 1–4.

8. Srinivas, A., J, S. B., Kumar, A. P., & Gupta, K. G. (2015). PARALLEL PRECEDENCE CONSOLIDATION FOR SIMILAR WORKLOAD IN CLOUD. IJRDO - Journal of Computer Science Engineering (ISSN: 2456-1843), 1(7), 54-62.
9. Anantharam, Banoth, and Gurnadha Guptha. "Security Issues in Various Cloud Computing: Solution and Occur-Rent Solutions." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN (2016): 2456-3307.
10. Becker-Reshef, C. Justice, ‚NASA Contribution to the group on earth observation (GEO) global agricultural monitoring system of systems,' The Earth Obs., vol. 21, pp. 24–29, 2009.
11. Russell G. Congalton, A review of assessing the accuracy of classifications of remotely sensed data, Remote Sens. Environ. 37 (1) (1991) 35–36.
12. F.F. Sabins, Remote Sensing: Principles and Interpretation, Freeman, 1978.
13. C.L. Philip Che, Chun-Yang Zhang, Data-intensive applications, challenges, techniques, and technologies: A survey on big data, Inform. Sci. 275 (2014) 314–347.
14. Gorton, Software architecture challenges for data-intensive computing, in Software Architecture, 2008. WICSA 2008. Seventh Working IEEE/IFIP Conference on, Feb. 2008, pp. 4–6.
15. Labrinidis and H. V. Jagadish, ‚Challenges and opportunities with Big Data,' in Proc. 38th Int. Conf. Very Large Data Bases Endowment, Istanbul, Turkey, Aug. 27– 31, 2012, vol. 5, no. 12, pp. 2032–2033.
16. P. Chandarana and M. Vijayalakshmi, ‚Big Data analytics frameworks,' in Proc. Int. Conf. Circuits Syst. Commun. Inf.Technol. Appl. (CSCITA), 2014, pp. 430–434.
17. Wikibon Blog. (Oct. 14, 2014). [2310]. Big Data Statistics [Online]. Available: wikibon.org/blog/big-data-statistics.