A Survey on Big Data Analytical Architecture for Real Time Remote Sensing Application

Harshaverdana S

M. Tech student, Dept. of CSE, UBDT College of Engineering, Davanagere, Karnataka, India

ABSTRACT: The assets of remote senses digital world daily generate massive volume of real-time data call that as Big Data, where insight information has a potential significance if collected and aggregated effectively. In today's era, there is a great deal added to real-time remote sensing Big Data than it seems at first, and extracting the useful information in an efficient manner leads a system toward a major computational challenges, such as to analyze, aggregate, and store, where data are remotely collected. Keeping in view the above mentioned factors, there is a need for designing a system architecture that welcomes both realtime, as well as offline data processing. Therefore, in this paper, we propose real-time Big Data analytical architecture for remote sensing satellite application. The proposed architecture comprises three main units, such as 1) remote sensing Big Data acquisition unit (RSDU); 2) data processing unit (DPU); and 3) data analysis decision unit (DADU). First, RSDU acquires data from the satellite and sends this data to the Base Station, where initial processing takes place. Second, DPU plays a vital role in architecture for efficient processing of real-time Big Data by providing filtration, load balancing, and parallel processing. Third, DADU is the upper layer unit of the proposed architecture, which is responsible for compilation, storage of the results, and generation of decision based on the results received from DPU. The proposed architecture has the capability of dividing, load balancing, and concurrent processing of only useful data. Thus, it results in efficiently analyzing real-time remote sensing Big Data using earth observatory system. Furthermore, the proposed architecture has the capability of storing incoming raw data to perform offline analysis on largely stored dumps, when required. Finally, a detailed analysis of remotely sensed earth observatory Big Data for land and sea area are provided using Hadoop. In addition, various algorithms are proposed for each level of RSDU, DPU, and DADU to detect land as well as sea area to elaborate the working of an architecture.

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

I.INTRODUCTION

As of late, a lot of enthusiasm for Big Data has risen, for the most part determined from across the board number of research issues emphatically identified with genuine applications and systems. Day by day the data is increasing very large volume from social media, videos, emails, online transitions, logs, Scientific data, mobile phones, Remote sensors and other applications. These data store in the database and grow rapidly with a massive amount becomes complicated to store, process, manage and analyze. The advanced technology in the big data gives a way to the remote data, which can be collection, managing, analyzing and processing. Recently designed remote sensors that are used for the earth observatory streams the data continuously and generates large amount of data. Many of the work have been done in the different fields of remote sensing data from the satellite, such as gradient based edge detection , change detection and etc. This paper is concentrated on the high speed continuous real time streaming data or large amount of offline data i.e. Big data, this leads to a new challenge. Such consequences for scientific understanding of transformation of the remote sensed data is critical task.

Data is collected from the remote sensors; these remote sensors generates a very large volume raw data this is also called as data acquisition. The collected data has no meaning in it, the sensor simply collects all the information. So the data need to be processed and filtered to extract the useful information from it. The main challenge in this is the data accuracy, the information that are generated by the remote sensors are not in the correct format for analysis. Now the data need to be extracted to pull the useful or meaningful data and converted into to the structured format for best

analysis. Sometimes the data might be not clear or it may be erroneous too. To address the above needs, the architecture is introduced, for the remote sensing big data.

This architecture has the capacity to analyze both type of data, offline data as well as real time data. First, the data has to be remotely processed in the readable format of the machine then the useful data is sent to the base station of the earth for the further data processing. The earth base station processes 2 types of data one is offline data and the other is real time streaming data. The offline data are sent to the offline data storage device incorporation of the later usage of data. Where in the real time data, the data is directly processed to filtering and the load balancing server. Filtering extracts the meaningful or useful data from the big data and the load balancing will balance processing by distributing the real time data equally to the server. These filtering and the load
balancing server will also improve the system efficiency.

Next, the data is directly sent to the data aggregation unit for comparison by analyzing and the decision server. The proposed method is implemented by the Hadoop framework using the map reduce programming by the data of remote sensing.

II. RELATED WORK

This section provides the detail summary of the previous work done in the remote sensing real time big data. The digital world generating the high amount of the data continuously, current technology and the tools to store and analyze the large amount of data not an easy task, since it is not able to extract the needed data sets. So there is a need of an architecture that can analyze both the offline data as well as real time data sets. There is an influential benefit in the business enterprise by obtaining the required information from the Big data than sample data sets. Some of the areas that are described below where big data can play very important role.

Understanding the earth atmosphere or environment requires large volume of information or data gathered from different sources, such as air and water quality monitoring sensors, amount of oxygen, co2 and the other gases present in the air, remote access satellite for the observing the characteristics of the earth and so on. In the healthcare scenarios, there is large amount of the data about the medications, patients, medical history and other details gathered by the medical practitioner. The above mentioned data is very complex in nature, there is a chances of missing the important data. Day by day the data becoming very large by social networking, online streaming, system logs, mails and remote data, it will be very difficult to compute massive amount of data. Main problem is how to store the large amount of data i.e. big data and what data is to keep and what data is to be discarded, extracting the useful data from the big data is the challenging task.

Most of the data is generated by the streaming data. In data stream model, the data will arrive at a very high speed and the algorithm have to process them. This data stream causes several challenges in design of the data mining algorithms. First, algorithm has to make use of less number of resources. Second, it can deal with data that can change over time. Resources are managed in an efficient and low cost way, by the green computing. Green computing is the process or study to use the computing resources in an efficient way. Here, the problem is not only the scaling issue but also error handing, lack of structure, heterogeneity, privacy, visualization and timeliness. The challenge is to design a high performance computing systems that can be able integrate resources from different location. Even though the cloud computing systems shown high level performance for RS applications, there are challenges still remaining regarding energy and the time consumption. The big challenge emerges when collecting and the managing Remote Sensing (RS) big data.

The RS data are collected form spacecraft, airplanes, satellite and other sensing devices. Remote sensing data growing explosively, we have entered in the period of very high resolution, observation of the earth. Remote sensing data also considered as a "Big Data". With the advance sensors we can take even high spatial resolution images, spectral resolution and also temporal resolution. The advancement in the technology of the computers and the remote sensing devices increases a massive growth remote sensing data. The earth observatory data that is streamed from the spacecraft approximately around 1.7GB, this data is collected by single satellite and increased many terabytes per day. The global records of observatory data of the earth would exceed to one Exabyte, according to the OGC statistics.

Various standard format data sets of remote sensing are stored in structured files, the formats including ASCII, HDF, netCDF and so on. Different organization have different standard format of the data sets, different format of data has its own format libraries and operation interfaces. Huge amount of data need to compute in an efficient way and only the useful information need to be extracted from the big data. So there was a need of the architecture for filtering the data, load balancing, aggregating and the decision analysis.

III.PROPOSED WORK

Even the big data has the variety technologies as in cloud computing. System architecture is considered as the core of any proposed system fetched from analysis and design. The architectural diagram for our system is shown in fig 1 and it contains the following components.

Generally the server acquires the data from the satellite and thus provides a backup at servers connected directly to satellite. These servers are unauthorized and highly preserved from public accessing. Apart from this, the servers also store the images in reliably higher ratio of memory and space. On demand from national or regional servers, the data is optimized and forwarded. These set of data is considered as internet data. In our proposed work we have considered a demanded earth Arial pictorial data.

These main servers are connected with inland servers and also proceed with connection with other primary servers. On request, the data is flowed from main server to inland servers via internet and thus we achieve the input data sets. Future, the samples are connected and stored into the regional servers and databases. The acquired images for Hadoop node is fetched from regional servers in our proposed system. Apart from storage, the data preprocessing and filtering is performed in this stage. Data cleaning is requires as acquired data is accompanied with other relevant and irrelevant attribute set.

Finally we move towards, Map Reduce based data reduction and monitoring under a scalable and most efficient time of processing. The data acquired here is considered as a primary asset for processing and achieving image redundancy. The system also aims to focus on time consummation during processing a huge data sets in Hadoop environment and Core Java environment.

They have divide the architecture of the remote sensing big data into 3 parts, 1) RSDU (Remote Sensing Big Data Acquisition Unit); 2) DPU (Data Processing Unit); 3) DADU (Data Analysis and Decision Unit). The working of these 3 units are described below.

RSDU (Remote Sensing Big Data Acquisition Unit)

Remote sensing encourages the growth of observatory system of the earth as cost efficient parallel data acquisition system to fulfil certain computational requirement. For efficiently analyzing big data there is a need of the parallel processing to process the big data in an efficient way. For this reason, the proposed method i.e. RSDU (Remote Sensing Big Data Acquisition Unit) is introduced in the architecture of remote sensing big data, that collects the data from different satellite from the globe. There is a possibility that raw data received can be distorted by various atmospheric gases and the dust particle.

We assume that the distorted or the erroneous data, satellite can correct. However, the remote sensing satellite uses the algorithm Doppler or SPECAN to make the raw data into the image format. The data is sent for further processing to the earth base station by direct communication link. In this two types of data processing takes place one is offline data processing and the another one is the real time data processing. In offline data processing, data will be transferred for storage, to the data center by the earth base station. This data is used for the future analysis. In the real time data, the data is directly sent to the FLBS (Filtration and Load Balancing Server

DPU (Data Processing Unit)

In DPU that is Data Processing Unit, it has two responsibilities, such as first, data need to be filtered by the filtration process. Second, balance the processing power by the load balancing server. Filtration recognizes or identifies the useful information, remaining data discarded of blocked. Hence, it improves the results of performance of the system. The load balancing server give the facility to divide the filtered data into parts and each part will be processed by the processing server. This load balancing and the filtration algorithm changes from analysis to analysis; example, if there is a need for only temperature data and the sea wave, then the needed data is filtered out and it is divided into parts.

Every processing server has its algorithm, to process the incoming segments of data from the filtration and the loadbalancing server. The processing servers performs some measurements, statistical calculations and makes other logicalor mathematical operations to create the intermediate results from every segments of data. Since each processing servers executes the tasks in parallel and independently, hence the proposed system dramatically boosts the performance. The results obtained by every processing server are sent to further processing to the aggregation server for organization, compilation and storing..
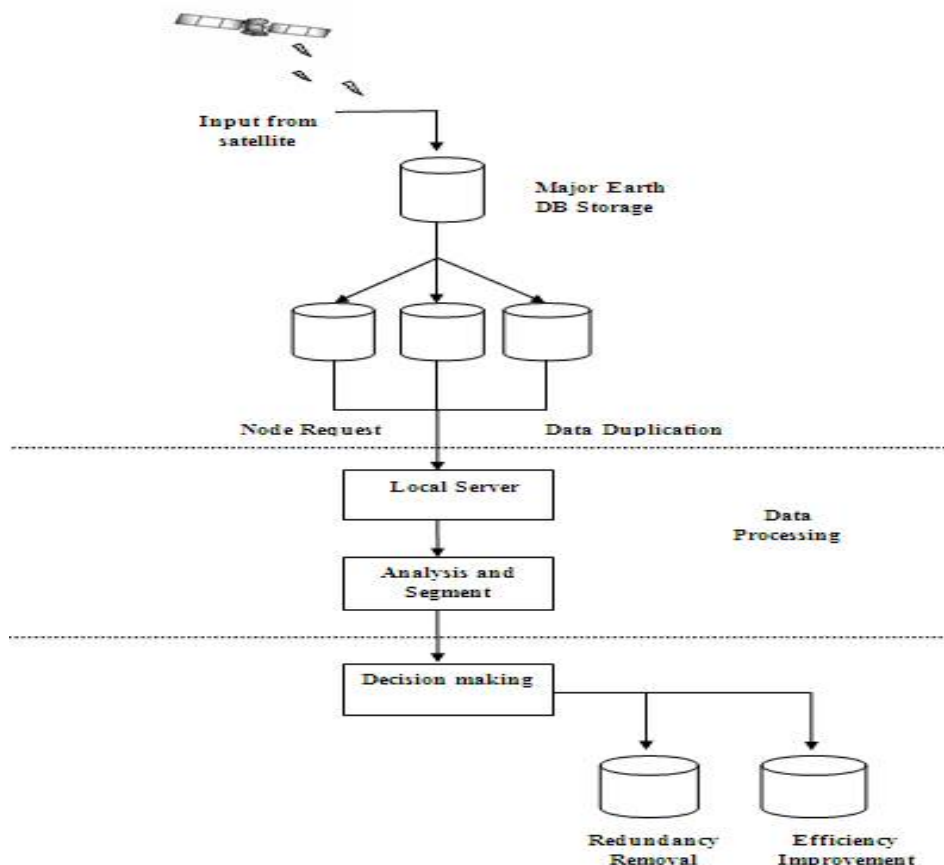

Figure 1: Analytical Architecture of remote sensing Big data

Data Analysis and Decision Unit (DADU)

DADU contains three major portions, such as aggregation and compilation server, results storage server(s), and decision making server. When the results are ready for compilation, the processing servers in DPU send the partial results to the aggregation and compilation server, since the aggregated results are not in organized and compiled form.

Therefore, there is a need to aggregate the related results and organized them into a proper form for further processing and to store them. In the proposed architecture, aggregation and compilation server is supported by various algorithms that compile, organize, store, and transmit the results. Again, the algorithm varies from requirement to requirement and depends on the analysis needs.

Aggregation server stores the compiled and organized results into the result's storage with the intention that any server can use it as it can process at any time. The aggregation server also sends the same copy of that result to the decision-making server to process that result for making decision. The decision-making server is supported by the decision algorithm,which inquire different things from the result, and then make various decisions (e.g., in our analysis, we analyze land, sea, and ice, whereas other finding such as fire, storms, Tsunami, earthquake can also be found). The decision algorithm must be strong and correct enough that efficiently produce results to discover hidden things and make decisions. The decision part of the architecture is significant since any small error in decision-making can degrade the efficiency of the whole analysis. DADU finally displays or broadcasts the decisions, so that any application can utilize those decisions at real time to make their development. The applications can be any business software, general purpose community software, or other social networks that need those findings (i.e., decision-making). The self-explanatory flowchart supporting the working of the proposed architecture is depicted in Fig 2.
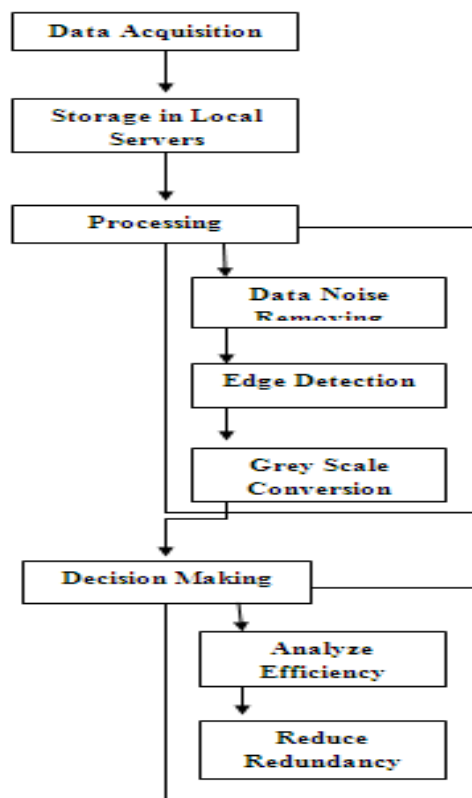

Figure 2: Flowchart for remote sensing big data architecture.

The system consists of a planned server accessing of data images from satellite and is fetched on request. The data acquired is stored in local servers and thus the request is processed in these servers for fetching the data. The remote servers contain data greater than regular size thus processing time and storage is considerably high and thus we propose the system to process such complicated images in a simpler manner under Hadoop Cluster.

Primarily the data is acquired and cleaned under Hadoop environment. The preprocessing step involves data pre-processing with data aligning and refining. The alignment of data is also achieved with filling an unfilled attributes. This is considered to be our primary processing step. In future steps, the data is analyzed and processed under filtering algorithm, the Map Reduce operation is also performed in this step. The data collected and proceed is now stored offline under local regional servers.

Decision making and Analysis is considered to be most difficult and standout step performed in our proposed system and thus fetches generic data sets for acquiring data analysis and controlling. The system also helps the users to achieve image redundancy optimization from regular and trivial storage systems. The system also contributes in this step to identify a distinct difference in processing a time gap from regular JAVA and Hadoop clustering. The major contribution of this system application is to achieve reliable results on selecting Hadoop clustering environment v/s java environment in terms of data processing and efficiency matching. The application also makes use of high data processing record and thus converts the given input data sets into gray images and extracts edges for achieving a secure data redundancy.

IV. IMPLEMENTATION AND RESULTS

In the Hadoop implementation, Map function takes the image block offset as a key and the image block (pixel values) as a value parameter. Since Hadoop MapReduce cannot directly process image blocks, the whole product image data are converted into sequence file to be processed using MapReduce. In such a way, one line of the sequence file contains one image block. Map function performs parameters calculations on incoming block values and finally sends the block number as a key and list of parameters results as a value to the Reduce function. Reduce function uses parameter results for performing decision-making on them. Moreover, when the product size exceeds 200 MB, it produces better results as compared with simple Java implementation. Hence, for smaller size products, the Hadoop implementation is not efficient because of its lots of input and output operations due to Map and Reduce function. In the case of large-size products, Hadoop divided whole products into blocks and performed parallel tasking on them, which resulted in increasing efficiency.

V. CONCLUSIONS AND FUTURE WORK

The remote senses generate very large amount of real time data from the Satellite or from the Aircraft with the help of the sensors. Now a day there is a great demand added to the real time big data for remote sensing applications, these data has to be processed and extract the useful information can lead to computational challenges. In this paper, we discussed the proposed analytical architecture of real time big data for remote sensing applications. The proposed architecture is designed in such a way that; it can analyze both the offline as well as the real time data in an efficient way. The proposed architecture for the remote sensing application, The three main units comprises the proposed architecture 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 starts in this unit. Second, Data processing unit (DPU) is the main role in the architecture, the real time data will process efficiently by filtering, load balancing and parallel processing and Third, Data Analysis and Decision unit (DADU) this unit is responsible for the storing the results and generates the decision based on the results of the data processing unit. In this paper, the remote sensing data analyzed by each proposed unit to make better decision making..

For future work, this proposed architecture can use to compute for more complex data for decision making at real time of earth observatory, such as fire detection, tsunami prediction, earthquake prediction, etc. The architecture need to make compatible for all applications for big data analysis.

REFERENCES

1. Muhammad Mazhar Ullah Rathore, Anand Paul, Bo-Wei Chen, Bormin Huang, and Wen Ji, "Real-Time Big Data Analytical Architecture for Remote Sensing Application," IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, 2015.
2. J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, "Mad skills: New analysis practices for Big Data," *PVLDB*, vol. 2, no. 2, pp. 1481–1492, 2009.
3. D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in *Proc. Int. Conf. Extending Database Technol. (EDBT)*, 2011, pp. 530–533.
4. E. Christophe, J. Michel, and J. Inglada, "Remote sensing processing: From multicore to GPU," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. , vol. 4, no. 3, pp. 643– 652, Aug. 2011.

5. K. Michael and K. W. Miller, "Big Data: New opportunities and new challenges [guest editors' introduction]," *IEEE Comput.*, vol. 46, no. 6, pp. 22–24, Jun. 2013.
6. "C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: Mc Graw-Hill, 2012.
7. A. Cuzzocrea, D. Saccà, and J. D. Ullman, "Big Data: A research agenda," in Proc. Int. Database Eng. Appl. Symp. (IDEAS'13) , Barcelona, Spain,Oct. 09–11, 2013.2
8. J. Shi, J. Wu, A. Paul, L. Jiao, and M. Gong, "Change detection in synthetic aperture radar image based on fuzzy active contour models and genetic algorithms" *Math. Prob. Eng.*, vol. 2014, 15 pp., Apr. 2014.
9. A. Paul, J. Wu, J.-F. Yang, and J. Jeong, "Gradient-based edge detection for motion estimation in H.264/AVC" *IETImage Process.*, vol. 5, no. 4, pp. 323–327, Jun. 2011.
10. A. Paul, K. Bharanitharan, and J.-F. Wang, "Region similarity based edge detection for motion estimation in H.264/AVC," *IEICE Electron. Express*, vol. 7, no. 2, pp. 47–52, Jan. 2010.
11. A.-C. Tsai, A. Paul, J.-C. Wang, and J.-F. Wang, "Intensity gradient technique for efficient intra prediction in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 694–698, May 2008.
12. R. A. Dugane and A. B. Raut, "A survey on Big Data in real-time," Int. J. Recent Innov. Trends Comput. Commun., vol. 2, no. 4, pp. 794– 797, Apr. 2014.
13. S. Kalluri, Z. Zhang, J. JaJa, S. Liang, and J. Townshend, "Characterizing land surface anisotropy from AVHRR data at a global scale using high performance computing," Int. J. Remote Sens. , vol. 22, pp. 2171–2191,2001.
14. L. Angrisani, M. D'Arco, G. Ianniello, and M. Vadursi, "An efficient pre-processing scheme to enhance resolution in band-pass signals acquisition," *IEEE Transa. Intrum. Meas.*, vol 61, no. 11, pp. 2932–2940, Nov. 2012.
15. Yan Ma, Haiping wu, Lizhe wang, Bormin Huang, Rajiv Ranjan, Albert Zonmaya, wei jie, "Remote sensing big data computing: Challenges and opportunities," Article in press., October 2014.

BIOGRAPHY

Harshaverdana S is a MTech post graduate student in computer science and engineering, UBDT college of engineering, Davanagere ,Karnataka, India. He received his Bachelor of Engineering in computer science from Malnad College of Engineering, Hassan, Karnataka, India in 2014.