Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 6, June-2016

ADOPTION OF HADOOP FOR A REMOTE SENSING REAL TIME BIG DATA ANALYSIS

[1]Mr.Harshaverdana S, [2]Kavitha G B.E,M.TECH

[1]*P.G.Student*

[2]*ASSISTENT PROFESSOR*

Department of Studies in Computer science and engineering

UBDT College of Engineering, Davanagere

Abstract -*In this paper we proposed that How Real time Big data can be managed for remote sensing application using Hadoop as a tool.Managing Big data is one of the biggest problem,its a new turn to the cost oriented companies by managing huge volume of data, velocity and variety of information. The real-time Big data for remote sensing application looks easy at first, but the useful data extraction in a effective manner gives the system towards a enormous computational challenges, they are analyzing, aggregating, and storing, here information are remotely gathered. Keeping in view that the above said points; designing such a system which calls for a both offline, as well as online processing of data needed. That's why, in this paper, we discuss real-time Big Data for remote sensing satellite application using hadoop as a tool. In the Architecture it has three main blocks, they are 1) remote sensing Big Data acquisition unit (RSDU); 2) data processing unit (DPU); 3) data analysis decision unit (DADU).Firstly, RSDU will collect data from the satellite and transmit this data to the Base Station, where starting task process takes place. Later DPU plays an important role in architecture for effective processing of real-time Big Data by providing filtration, load balancing, and parallel processing. Then DADU is responsible for storage of the results, and generation of decision based on the results received from DPU. The Design has the power of dividing, load balancing, and parallel processing of only useful data. Thus, it results in efficiently analyzing real-time remote sensing Big Data using earth observatory system. Finally,a detailed analysis of remotely sensed earth observatory Big Data for land and sea area are provided using Hadoop as tool*

Keywords—*Big Data, data analysis decision unit (DADU), data processing unit (DPU), Hadoop, mapreduce, offline, real time, remote senses, remote sensing Big Data acquisition unit (RSDU).*

I.INTRODUCTION

Now a days lot of research are going on using Big Data which led for the rise of genuine Real time applications and systems. Increase of data day by day creating lots of issues such as large volume of data from online transitions, logs, Scientific data, emails ,videos ,social media, mobile phones, Real time Remote sensors and in new applications .Where each of the data stored in the huge database and grow fastly with a massive amount and becomes complicated to store, process, manage and analyze The advanced research in new technology gives the direction to the remote data, which can be collecting, 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.

In most recently designed sensors provide in the earth and planetary observatory system are generating continuous stream of data. Moreover, majority of work have been done in the various fields of remote sensory satellite image data, such as change detection , gradient based edge detection , region similarity based edge detection and intensity gradient technique for efficient intra prediction .The incredible growth in the data also posing new challenges, such as, how to aggregate massive amount of data? How to store such data in a limited amount of memory allocated for the particular task? Moreover, how to process and analyze these data when there is no such intelligent algorithm is available? Moreover, large-scale data cannot be tackled by standard reduction techniques since their runtime becomes impractical . Having a large-amount of data, all of this has to happen in a mechanized manner since it requires diverse data structure as well as semantics to be articulated in forms of computer readable format. However, by analyzing simple data having one data set, a mechanism is required of how to design a database. There might be alternative ways to store all of the same data. In remote access networks, where the data origin such as sensors can generate an overwhelming amount of raw data. i.e., data acquisition, in which much of the data are of no meaning that can be filtered or compressed by orders of magnitude. With a view to using such filters, they do not discard useful information. The second challenge is by default generation of accurate metadata that describe the composition of data and the way it was collected and analyzed. Such kind of metadata is hard to analyze since we may need to know the source for each data in remote access. Generally, the data gathers from remote areas are not in a format ready for analysis.

Therefore, the second step does data extraction, which pull out the meaningful information from the underlying sources and transfer it in a structured formation suitable for analysis. For instance, the data set is covert to single-class label to facilitate analysis, even though the first thing that we used to think about Big Data as always describing the fact. However, this is far away from reality; sometimes we have to deal with erroneous data too, or some of the data might be not clear. To address the aforesaid needs, a remote sensing Big Data analytical architecture [1], this is used to analyze real time, as well as offline data. At first, the data are remotely preprocessed, which is then readable by the machines. Afterward, this meaningful information is delivered to the Earth Base Station for further data processing. Earth Base Station (EBS) performs two types of processing, such as processing of real-time and offline data. In case of the offline data, the data are transmitted to offline data-storage device. The incorporation of offline data-storage device helps in later usage of the data, whereas the real-time data is directly transfer to the filtration and load balancer server, where filtration algorithm is employed, which drag out the meaningful information from the Big Data. On the other hand, the load balancer balances the processing power by equal distribution of the real-time data to the servers. The filtration and load-balancing server not only filters and balances the load, but it is also used to enhance the system efficiency.

Hadoop holds a gap in the market by effectively storing and providing computational capabilities over substantial amounts of data. It's a distributed system made up of a distributed file system and it offers a way to parallelize and execute programs on a cluster of machines. You've most likely come across Hadoop as it's been adopted by technology giants like Yahoo!, Facebook, and Twitter to address their big data needs, and it's making inroads across all industrial sectors. MapReduce is a programming model and an associated implementation for processing and generating massive data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that combine all intermediate values associated with the same intermediate key. Furthermore, the filtered data are then processed by the parallel servers and are sent to data aggregation unit (if required, they can store the processed data in the result storage device) for comparison purposes by the decision and analyzing server. The proposed architecture welcomes remote access sensory data as well as direct access network data (e.g., GPRS, 3G, xDSL, or WAN). The architecture and the algorithms are implemented in Hadoop using MapReduce programming by applying remote sensing earth observatory data.

II.LITERATURE SURVEY

The increase in the data rates generated on the digital universeis escalating exponentially. With a view in employing current tools and technologies to analyze and store, a massive volume of data are not up to the mark, since they are unable to extract required sample data sets. Therefore, we must design an architectural platform for analyzing both remote access realtime and offline data. When a business enterprise can pull-out all the useful information obtainable in the Big Data rather than a sample of its data set, in that case, it has an influential benefit over the market competitors. Big Data analytics helps us to gain insight and make better decisions. Therefore, with the intentions of using Big Data, modifications in paradigms are at utmost. To support our motivations, we have described some areas where Big Data can play an important role.Understanding environment requires massive amount of data collected from various sources, such as remote access satellite observing earth characteristics [measurement data set (MDS) of satellite data such as images], sensors monitoring air and water quality, metrological circumstances, and proportion of CO2 and other gases in air, and so on. Through relating all the information drifting such as CO2 emanation, increase or decrease ongreenhouse effects and temperature, can be found out.

In healthcare scenarios, medical practitioners gather massive volume of data about patients, medical history, medications, and other details. The above-mentioned data are accumulated in drug-manufacturing companies. The nature

of these data is very complex, and sometimes the practitioners are unable to show a relationship with other information, which results in missing of important information. With a view in employing advance analytic techniques for organizing and extracting useful information from Big Data results in personalized medication, the advance Big Data analytic techniques give insight into hereditarily causes of the disease.

III.SYSTEM REQUIREMENT SPECIFICATION

3.1System Requirements

The system is simulated as a protocol under Hadoop clustering environment for cloud infrastructure in apache server. A detailed overview is shown as below.

A. **Software Requirements**

Operating System: Ubuntu 14.04
Technology: JDK 1.7
Framework: Eclipse
Server: Apache

B. Hardware Requirements

Processor: i3
Speed: 2.1 GHZ
RAM: 2GB
DiskSpace: 5GB

3.2 Proposed System Design Overview

We propose remote sensing Big Data architecture to analyse the Big Data in an efficient manner.Delineates n number of satellites that obtain the earth observatory Big Data images withsensors or conventional cameras through which sceneries are recorded using radiations. Special techniques are applied to process and interpret remote sensing imagees for the purpose of producing conventional maps, thematic maps, resource surveys, etc. We have divided remote sensing Big Data architecture into three parts.

3.2.1Remote Sensing Big Data Acquisition Unit (RSDU)

Remote sensing promotes the expansion of Earth observatory system as a cost effective parallel data acquisition system to satisfy specific computational requirements. The Earth and space science technology approved this solution as the standard for parallel processing in this particular context. Therefore the need for parallel processing of the massive volume of data was required which efficiently analyse the big data. In this case the offline data processing is done in which the earth base station transmits the data to the data centre for storage. This data then used for the future analysis. However, in real-time data processing the data is directly transmitted to the filtration and local balancer server (FLBS), since storing of incoming real-time data degrades the performance of real-time processing.

3.2.2Data Processing Unit (DPU)

In data processing unit, the filtration and load balancing server have two basic responsibilities such as filtration of data and load balancing of processing power. Filtration identifies the useful data for analysis since it only allows useful information where rest of the data are blocked and discarded, hence it enhancing the performance of the whole proposed system. The load balance system is the part of the server provides the facility of dividing the whole filtered data into parts and assign them to the various processing servers. The filtration and load balancing algorithm varies from analysis to analysis. The result generated by each server is then sent it to the aggregation server for compilation, organization, and storing further processing.

3.2.3Data Analysis and Decision Unit (DADU)

DADU contains three major portions, such as aggregation and compilation server, result storage server (s), and decision making server. When the results are ready to compilation, the processing server in DPU sends the partial result to the aggregation and compilation server, since the aggregated result are not in organized and compiled form. Thus, there is need aggregate the related result and organize them into proper form for further processing and to store them. In the proposed architecture, aggregated and compiled servers are supported by various algorithms that compile, organize and store, and transmit the results.

IV.SYSTEM DESIGN

4.1 System Architecture

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 4.2 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.

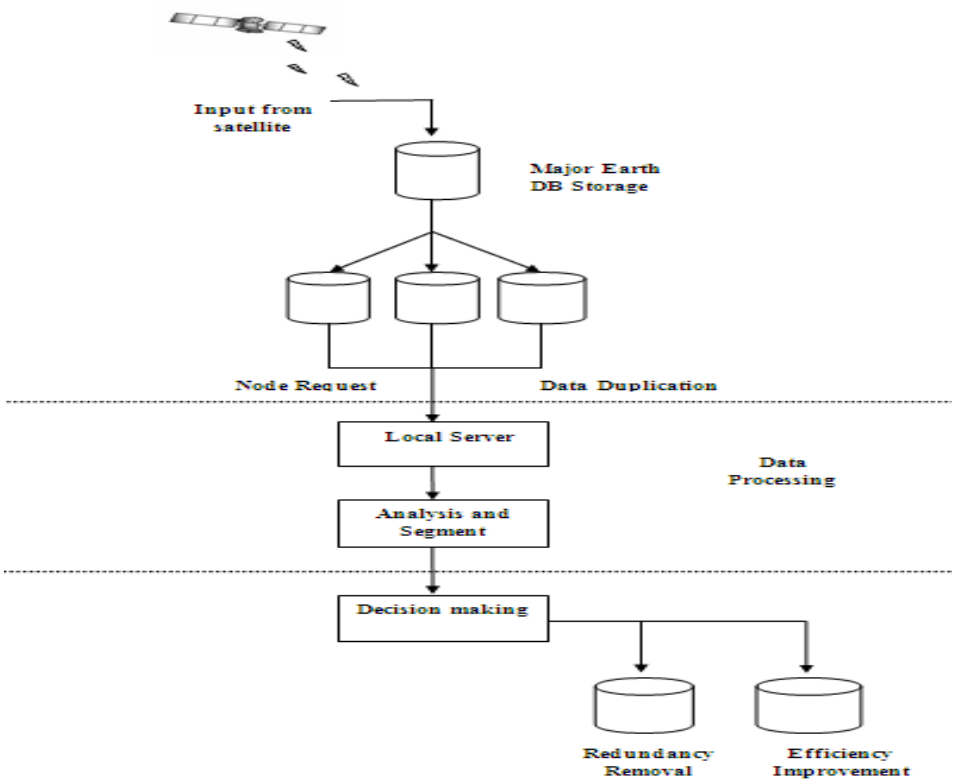


Fig 4.1 Architectural Diagram

4.2 Data Flow Diagram

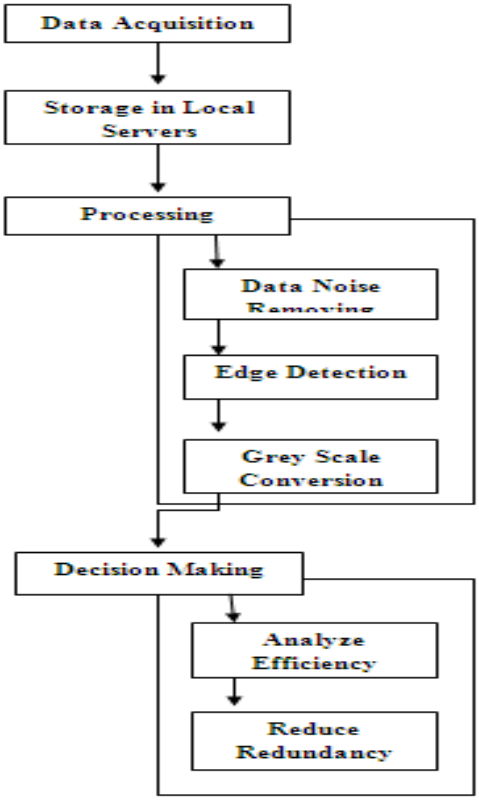


Fig 4.2: Data flow diagram

Data flow diagrams are most important and fundamental design structures in analyzing a problem and its solution. 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.

4.3 Sequence Diagram

Under this unit, a modularity of analyzing the proposed system in a sequential manner is performed and thus the same is showcased in Fig 4.3. The system's sequential diagram is aided with server, hadoop cluster and decision.

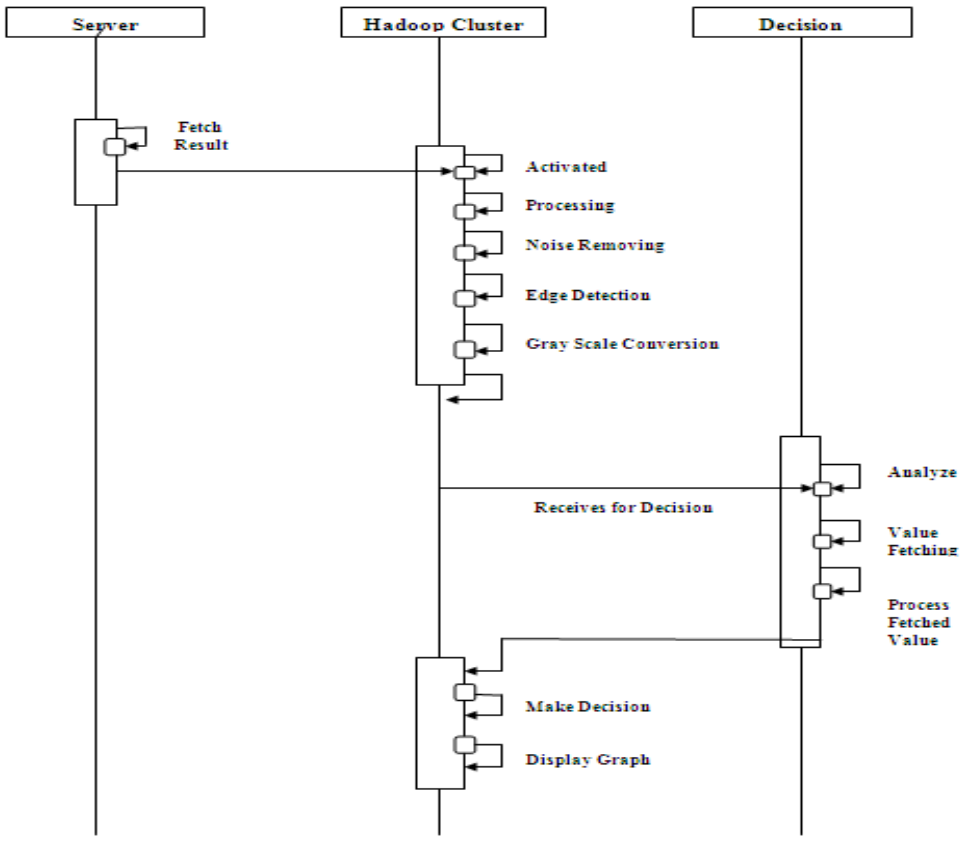


Fig 4.3 Sequence Diagram

V.IMPLEMENTATION

Implementation is the process of fetching real-time modulations for proposed system, this also focus towards design of mathematical model and modulations for achieving trusted results. In our system, we have considered three major modules for implementation as follows.

1. Data acquisition
2. Data Processing and Filtering
3. Decision making and Analysis

5.1 Data acquisition

The data sets collected and processed and acquired from real-time satellite connected sever and a sample of the same is projected below in fig 5.1.

5.2 Data Processing and Filtering

The process of data filtering and processing is monitored with a hyper active server environment for providing a filtering process on missing data and re-aligning the same under norms for data storage and maintained in intermediate serves.

5.3 Decision making and Analysis

Here final decisions and at last analysis are made.

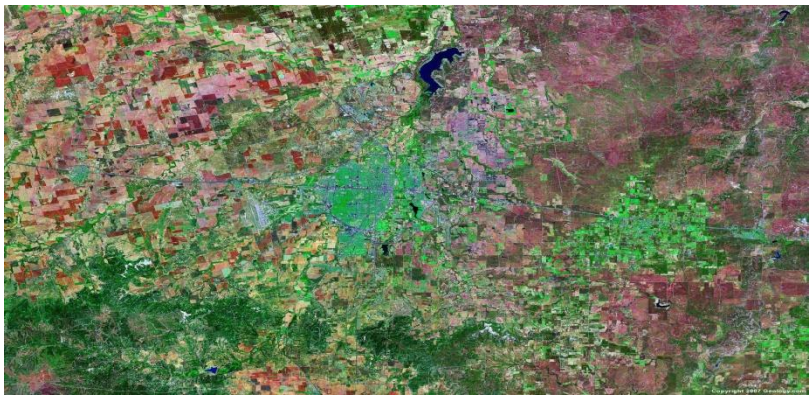


Fig 5.1: Sample of Acquired Data set

VI.TESTING

Testing is the most significant inter technological and maintenance stream of validating and verifying the proposed or designed system for consumer exposure and release. This chapter we have discussed a detailed view on test cases with respect to the occurrence and inbounding scenarios.

Typically the system is performed with white box testing, black box testing and integration testing under overall simulation approach. Testing at an IT field is segregated as an independent domain for work and analysis.

6.1 Test Planning

The main objective of software development life cycle is to produce a product with no errors or very few errors. In the processes of achieving hassle free software we plan testing and test cases.

6.1.1 Hadoop Cluster Initiation Testing

Hadoop clustering is pre-installed and in this case, we personal activate the same. Test cases are designed based on these functionalities.

TEST CASE ID	**TC-1**
NAME OF THE TEST	Cluster Activation
TEST DESCRIPTION	Hadoop activation is seen and a IDE is activated via JAVA Luna environment
EXPECTED OUTPUT	Should be activated and authenticated with clearing and building command
REMARKS	Passed

Table 6.1: Hadoop Initialization Test Case

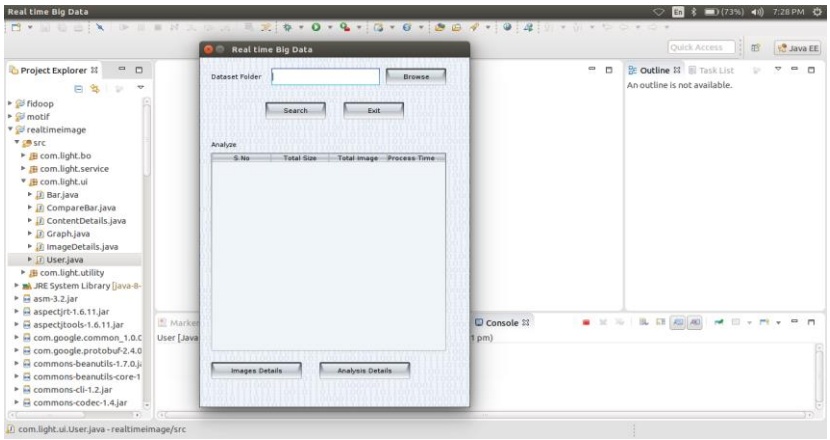
TEST CASE ID	**TC-2**
NAME OF THE TEST	Earth Image Validation
TEST DESCRIPTION	Validation of remote earth images is performed under normal clustering and edge detection technique
EXPECTED OUTPUT	Image quality greater than 600KB
REMARKS	Passed

Table 6.2: Real-Time images validation Test Case

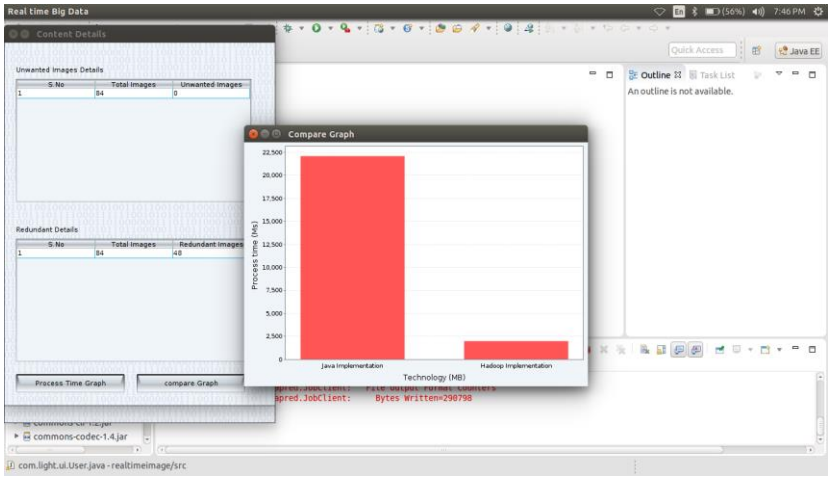
TEST CASE ID	**TC-3**
NAME OF THE TEST	Validation Test
TEST DESCRIPTION	Data processing and redundancy generation is performed with mathematical modeling and system architectural rules
EXPECTED OUTPUT	Fetch calibrated and redundant images
REMARKS	Passed

Table 6.3: Data Validation and decision making Test Case

VII.SCREENSHOTS



This picture shows how the process starts , at first java process will get started within few seconds the hadoop process will also get started.



Finally I can show the difference between java and hadoop. Our paper main intension was to prove that adopting hadoop is way better than java. This picture depicts the advantage by adopting hadoop over java.

VIII. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, the architecture for real-time Big Data analysis for remote sensing application is proposed. The proposed architecture efficiently processed and analyzed real-time and offline remote sensing Big Data for decision-making. The proposed architecture is composed of three major units, such as RSDU, DPU and DADU. These units implement algorithms for each level of the architecture depending on the required analysis.

This proposed system has successfully achieved predominant results on processing time and efficiency with respect to processing time in Hadoop Cluster and Core Java. Apparently, we have also observed a high redundancy optimization of earth images with an overall input processed with datasets. The system is successfully performed a processing of 60MB data under an interval of 11,000 sec under Java and less than 2500 sec in Hadoop cluster. The system also showcases a new edge on understanding a Hadoop MapReduce operation for redundancy.

Enhancement in terms of direct server dataset processing can be proposed in the upcoming version of this system. For future work, there is a plan to extend the proposed architecture to make it compatible for Big Data analysis for all applications, e.g., sensors and social networking. We are also planning to use the proposed architecture to perform complex analysis on earth observatory data for decision making at realtime, such as earthquake prediction, Tsunami prediction, fire detection, etc.

ACKNOWLEDGEMENT

I am highly obliged to Department of computer science and engineering, UBDT college of engineering. And I am highly grateful and thankful to our guide Assist Prof . Kavitha G for her valuable instructions,guidance,corrections in my project work and presentation.

REFERANCES

[1] 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.

[2] 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.

[3] A. Paul, J. Wu, J.-F. Yang, and J. Jeong, “Gradient-based edge detection for motion estimation in H.264/AVC,” IET Image Process., vol. 5, no. 4, pp. 323–327, Jun. 2011.

[4] A. Paul, K. Bharanitharan and J.-F. Wang, "Region similarity based edge detectionformotionestimationinH.264/AVC,"IEICEElectron.Express, vol. 7, no. 2, pp. 47–52, Jan. 2010.

[5] 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.

[6] S.Kalluri,Z.Zhang,J.JaJa,S.LiangandJ.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.

[7] A. 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.

[8] EnviSat, A. S. A. R. Product Handbook, European Space Agency, Issue 2.2, Feb., 2007.

[9] R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd ed. New York, NY, USA: Academic Press, 1997.

[10] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ, USA: Wiley, 2003.

[11] C.-I. Chang, Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Norwell, MA, USA: Kluwer, 2003.

[12] J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York, NY, USA: Springer, 2006.

[13] 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.

[14] A. Paul, J. Wu, J.-F. Yang, and J. Jeong, "Gradient-based edge detection for motion estimation in H.264/AVC," IET Image Process., vol. 5, no. 4, pp. 323–327, Jun. 2011. [15] A. Paul, K. Bharanitharan, and J.-F. Wang, "Region similarity based edge detectionformotionestimationinH.264/AVC,"IEICEElectron.Express, vol. 7, no. 2, pp. 47–52, Jan. 2010.

[15]Efficient Analytical Architecture in Real-time Big Data for Remotely Sensing Application Using Hadoop Framework Ajay Katware1, Pankaj Patil2, Yogesh Lokare3,*1D.Y. Patil School Of Engg. Academy, Ambi.*

[16] Real-Time Big Data Analytical Architecture for Remote Sensing Application, Muhammad Mazhar Ullah Rathore, Anand Paul, *Senior Member, IEEE*, Awais Ahmad, *Student Member, IEEE*, Bo-Wei Chen, *Member, IEEE*, Bormin Huang, and Wen Ji, *Member, IEEE*