ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# RealBDA: A REAL TIME BIG DATA ANALYTICS FOR REMOTE SENSING DATA BY USING MAPREDUCE PARADIGM

**Yaman Kumar Gunturi*, K. Kishore Raju**
* SRKR Engineering college, Bhimavaram India
SRKR Engineering College, Bhimavaram India

**DOI**: 10.5281/zenodo.260070

## ABSTRACT

Enormous data generated by Satellite sensors, Storage and Processing of Remote Sensing Data is a challenging task due to its variety and volume. This paper studied on real-time Big Data Analytical architecture for remote sensing satellite application. To handle Remote Sensing Data proposed architecture comprises three main units, such as Data Pre-Processing Unit ($D_{PRE}U$), Data Analysis Unit ($D_AU$) and Data Post-Processing Unit ($D_{POST}U$). First, $D_{PRE}U$ acquires the required data from satellite sensors by using filtration, balanced distributed storage and parallel processing using Hadoop environment. Second, $D_AU$ identifies the hidden patterns from data stored in distributed File System using Map functions followed by Reduce functions in Map-Reduce paradigm. Finally, $D_{POST}U$ is the upper layer unit of the proposed architecture, which is responsible for compiling storage of the results, and generation of decision based on the results received from $D_AU$.

**KEYWORDS:** Remote Sensing Data, Big Data, Distributed File System, Map-Reduce, Hadoop Environment.

## INTRODUCTION

In recent days Big Data and its analytics playing predominant role in optimal storage of semi or unstructured data and Decision making by using mining techniques and predictive analytics.Especially Remote Sensing accumulates huge data in the form of multispectral high resolution satellite images. These images contain variety of data in huge volume in the form of pixels. Distributing high volume data into different commodity systems using distributed file system is a major revolution created by Hadoop framework to handle big data with the available hardware and computational capabilities. Map Reduce is a technique which performs Map functions and Reduce functions on the distributed file system.

Mapper functions are split into number of record readers and they will read the data loaded distributes file system by using key-value pair. The output of each Map function is taken by Reducer function for further analysis.

For Recently it's a great deal of curiosity in the field of Big Data and its study has risen, mainly driven from extensive quantity of research tasks strappingly related to bonafide submissions, such as modeling, processing, querying, mining, and distributing large-scale repositories. The term "Big Data" classifies specific kinds of data sets comprising formless data, which well in data layer of technical computing applications and the Web. The data stored in the underlying layer of all these technical computing application scenarios have some precise individuality in common, such as

- Large scale data, which refers to the size and the data warehouse or mart;
- Scalability issues, which refer to the application's likely to be running on large scale (e.g., Big Data);
- Sustain extraction transformation loading (ETL) method from low, raw data to well thought-out data up to certain extent; and
- Development of simple interpretable analytical over Big Data granaries with a view to deliver an intellectual and momentous knowledge for them.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Big Data are usually generated by online operations, email, video/audio, number of clicks, number of views, logs, posts, views, social network data, Advertisements, scientific data, remote access sensory data, Electronic items, and their apps. These data are collected in databases that grow extremely and become complex to confine, form, store, manage, share, process, analyze, and visualize via typical database software tools. We fetch the twits on the server then we are preprocess those steps using proposed architecture. Selection, data load managing, Aggregation and Data Analysis algorithms are achieved on those process. Then using this *RealDBA* we classify the highest and lowest humidity values.

## LITURTURE SURVEY

These section efforts the detail precipitates of the previous workdone 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 andanalyze the large volume of data not an easy task, since it is not able to excerpt the needed data sets. So there is a need of anarchitecture that can analyze both the offline data as well as realtime data sets. There is asignificant benefit in the businessenterprise by attaining the required information from the Bigdata than sample data sets.

Understanding the earth atmosphere or environs requires large volume of data or data gathered from different sources, such as air and water quality monitoring sensors, amount of oxygeno2, co2 and the other gases present in the air, remote contact satellite [1] for the observing the characteristics of the earth and so on. In the healthcare scenarios, there is enormous amount of the data about themedications, patients, health history and other details congregated by the medical practitioner. The above acknowledged data is very complex in environment, there is a chances of lost the significant data.

Present days the data becoming very big by social networking, online flowing, system logs, mails and remote data, it will be very difficult to compute massive amount of data. Main problematic is how to store the bulky amount of data i.e.big data and what data is to keep and what data is to be rejected; extracting the useful data from the big data is the interesting assignment [2].

Most of the data is made by the spilling data. In data stream model, the data will arrive at a precise high speed and the algorithm has to process them. This data stream causes numerous challenges in design of the data mining algorithms. First, algorithm has to make use of fewer amounts of resources.Second, it can deal with data that can change over time.Resources are managed in an efficient and small cost way, by the green calculating [4]. Green computing is the process or study to use the computing possessions in an efficient way. Here, the problem is not only the scaling issue but also error control, lack of structure, heterogeneity, privacy, visualization and timeliness.

The challenge is to design a high enactment computing systems that can be able to integrate resources from dissimilar location. Even though the cloud computing systems shownhigh level presentation for RS applications, there are challenges still remaining concerning energy and the time depletion.The big challenge emerges when collecting and the managing Remote Sensing (RS) big data [3]. The RS data is unruffled form airliners, spacecraft, satellite and any other sensing devices. Remote sensing data is increasing explosively; we have entered in the period of actual 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 pictures, spectral resolution and also sequential resolution. The progression in the technology of the computers and the distant sensing devices increases a massive development remote sensing data. The earth laboratory data that is streamed from the spacecraft approximately around 2(1.7) GB, this data is collected by single satellite and increased more number of terabytes per day. The total records of observatory data of the earth would be exceed to one *Exa*byte, as per the OGC statistics [10].

Various standard format data sets of remote sensing arestored in structured files, the formats including ASCII, HDF, net CDF 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[5] 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 cleaning the data, load balancing, aggregating and the decision analysis.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Big Data analysis is one of the challenging task for locating, identifying, understanding, and screening data [6]. Having a large-scale data, all of this has to happen in a computerized manner since it requires diverse data structure as well as semantics to be expressed in forms of computer-readable format. However, by analyzing the 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 information. In such conditions, the declared design might have an improvement over others for certain process and possible disadvantages for some other purposes. In order to address these needs, various logical platforms have been provided by relational records vendors [7].

Any platform comes in frequent natures from software only to analytical services that run in third revelry hosted type. In remote entree networks, where the data source such as devices can produce hug amount of raw data. We refer it to the first one, i.e., data preprocessing, in which much of the data are of no interest that can be filtered or compressed by orders of enormousness. With a view to using such selections, they do not discard useful information. The challenge is by default compeers of precise meta data that describe the configuration of data and the way it was composed and studied. Such kind of metadata is hard to scrutinize since we may need to know the source for each data in remote access [1].

Normally, the data composed from remote extents are not in a format ready for investigation. Therefore, the first one refers us to data processing, which drags out the useful info from the fundamental sources and delivers it in a organized formation suitable for analysis. For instance, the data set is reduced to single-class label to simplify analysis, even though the first thing that we used to think about Big Data as always relating the fact. However, this is far away from authenticity; sometimes we have to deal with meaningless data too, or some of the data might be inexact.

## *RealBDA* ARCHITECTURE OF REAL-TIME BIG DATA ANALYZER for REMOTE SENSING BIG DATA

The architecture for processing real time big data generated by remote sensors is a challenging task. To handle variety and high volume data efficiently and effectively *RealBDA* architecture is proposed. The data generated by Remote Sensors called Raw data contains bulk amount of information in unstructured, semi structured and structured format. It is tedious task for data analysts to identify hidden patterns from large amount of data having noisy data. To handle this problem *RealBDA* architecture is implemented in three levels

Level-i: Data Pre-Processing Unit ($D_{PRE}U$)
Level-ii: Data Analysis Unit ($D_AU$)
Level-iii: Data Post-Processing Unit ($D_{POST}U$)

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

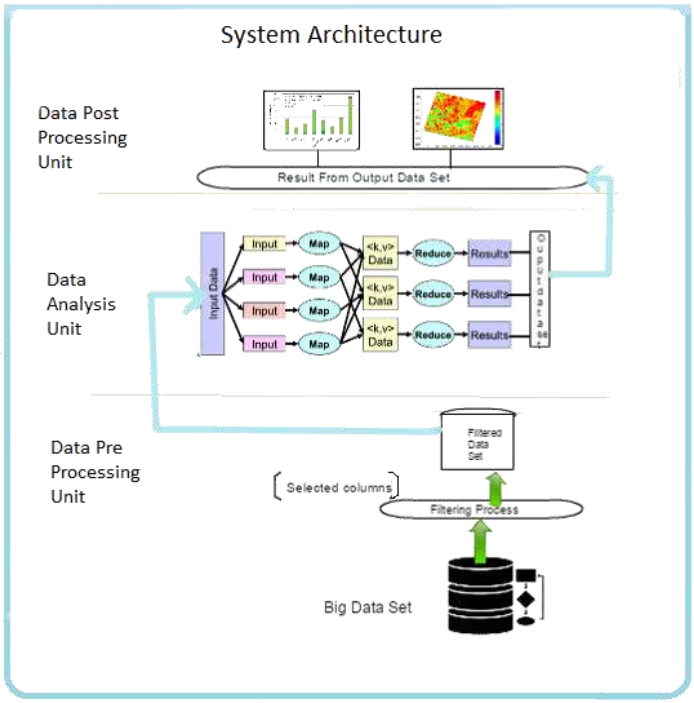


***Figure 3.1: Architecture of RealBDA for remote sensing data Mainly this three units working on explained on below***

***Data Pre-Processing Unit ($D_{PRE}U$)***

In $D_{PRE}U$[12] that is Data Pre-Processing Unit, It is a derelict but important step in structured or semi structured data set before mining process. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus the representations and quality of data is first and foremost before running for analysis. Data Pre-Processing is one of the most critical steps in a data processing making it in to meaning full data which supports to research and revolution of the initial dataset Data Pre-Processing has method like Data Cleaning, Data Reduction, Data Alteration, Data Amalgamation .Generally the data has been taking from data set which we preferred for Analysis in that Big data set it goes in to Selection process based on number of columns selected columns goes to HDFS.That Selected Data set from the Distributed system taken as Filtered Data set for the Map Reduce process.

***Algorithm Data preprocessing***

**Input** : Raw Data of D = {P1,P2,… Pn} .
**Output :** Data available in HDFS.

Step1: A load Raw Data
Step2: Set the Parameters P = {Px,Py,Pz}
Step3: Filter required parameters P from Raw Data
Step4: Load Filtered Data in HDFS.

***Data Analysis Unit ($D_AU$):***

In **$D_AU$**[12] that is Data Analysis Unit, it has responsibility,such as first data need to be filtered by the selection process.Then balance the processing power by the load balancingsystem. Filtration recognizes or identifies the usefulinformation, remaining data discarded of blocked. Hence, itimproves the results of performance of the system. The loadbalancing interconnected give the facility to divide the selected data intoparts and each part will be processed by the processing system.This load balancing and the filtration algorithm changes fromanalysis to analysis; example, if there is a need for only temperature data and the sea wave, then the needed data isfiltered out and it is distributed into parts.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Every processing system has its algorithm, to process the incoming segments of data from the filtration and the load balancing system. The processing server performs some measurements, statistical controls and makes other logical or mathematical calculations to create the intermediate results from every segment of data. For that data applying Map reduce Paradigm so it starts the process as below.

Apache Hadoop is a distributed totaling framework modeled after Google MapReduce to process huge amounts of data in parallel. Once in a while, the first thing that comes to data about distributed computing is EJB. EJB is a component model with remote capability but short of the critical features being a distributed computing framework that include computational parallelization, work distribution, and tolerance to unreliable hardware and software. Hadoop Distributed File System (HFDS) modeled on Google GFS is the underlying file system of a Hadoop cluster. HDFS works more efficiently with a few big data files than numerous small files. A real-world Hadoop job classically takes minutes to hours to complete, therefore Hadoop is not for real-time analytics, but rather for offline, batch data processing. Recently, Hadoop has undergone a complete overhaul for improved maintainability and manageability. One major objective of Hadoop is MapReduce paradigm to accommodate other parallel computing model

The code hereby includes a Map and a Reduce class. Put simply, a Map class does the heavy-lifting job of data filtering, transformation, and splitting, a Mapper instance only processes the data clusters on the same data node, a concept termed data locality (or data proximity). Mappers can run in parallel on all the available data nodes in the cluster. The outputs of the Mappers from different nodes are shuffled through a particular algorithm to the appropriate Reduce nodes. A Reduce class by nature is an aggregator. The number of Reducer instances is configurable to result.

**Mapper Algorithm in Data Analysis**

**Input :** Data Record ( $Id_r$, r ) , r € D,

**Output:** Key Value Pairs (block id ←$B_i$, Humidity ←$H_r$)

**Step1:** For each attribute value Bi in r find its specialization in $B_{i-n}$

$B_i \sum^4_{i-1} H_r$

$B_i (H_r)$ , r = n

**Step2:** For each Bi value count

$B_i \sum^4_{j <= 1} H_r$ , count.

**Reducer Algorithm in Data Analysis**

**Input :** Value pairs **(**$B_i \sum^4_{j <= 1} (H_r)$ ,count**)**
**Output: S**imilar id humidity value
(spec $B_i$, spec(count) for all serialization)

**Step 1:** For each $B_i$ ,
← Sum $\sum^4_{j<=1} B_i (H_r)$, j <= r ;

**Step 2:** For each D = $\sum^4_{i<=1}$ , update count

|[$B_i(H_r)$,Count ++]| ← sum;

**Step 3:** Emit ( spec $B_i$, $B_i \sum^4_{i<=1} H_r$) where r = n records

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Data-Post Processing Unit has the Output data result of MapReduce (similar id humidity value). In this variations of map reduce data in graphical representation end-user understandable and predictive format. This similar data of result has to represent as graph. Humidity aggregate average value is taking as X-axis and the maximum count of Humidity representing in the Y-axis as per the average number of the Humidity value Bar Graph representation and Line graph is showing in below figure 4.1

## WORK FLOW OF *RealBDA* FOR REMOTE SENSING BIG DATA ARCHITECTURE

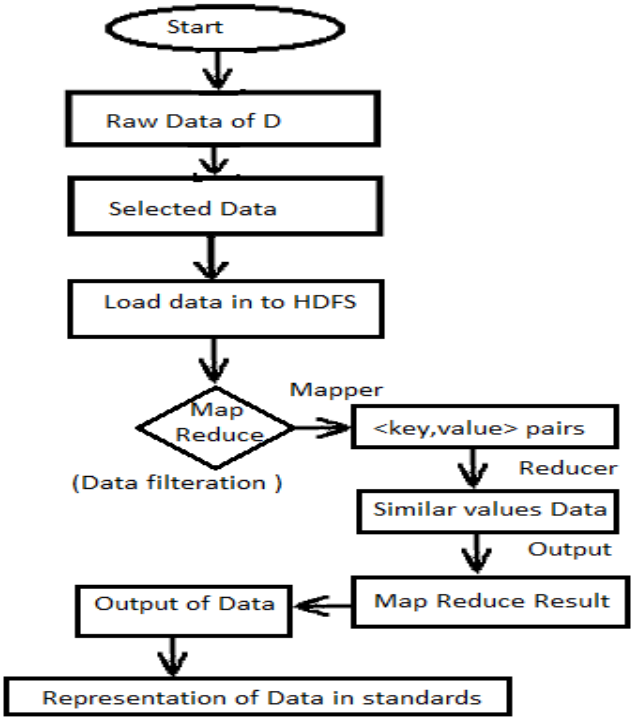


***Figure 3.2: Flowchart for remote sensing big data architecture.***

Aggregation server stores the results into the results storage this helps any other server to use it process at any time. DM(Decision making) server for making the decisions. The decision making server has decision algorithm, to make the various decisions. So any applications make use of these decisions to make their development at real time. The application can be any general purpose software, other social networks or any business software that need decision making. The Figure shows the flowchart for the proposed architecture.

## RESULT

The Proposed architecture is implemented in java using eclipse 7.0 version. As a part of Data preprocessing, Distributed file system is implemented using HADOOP-1.6.0 for storing large amount of data and stored in different nodes.

| Employee serial number | Employee block numbers | Employee max humidity values | Employee min humidity values | Constant values |
|---|---|---|---|---|
| 2 | 1 | 45.9 | 27.95 | 0 |
| : | : | : | : | : |
| 5040 | 4 | 46.75 | 23.03 | 0 |
| 5041 | 4 | 46.72 | 23.05 | 0 |

Key value

Text Value

***Figure 4.1: Selection of required data from Raw data collected from Remote sensor***

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

To implement Data Analysis unit, Map functions are designed using java language by taking large data which was distributed in different nodes as input. In Hadoop, Map function takes the data set column offset as a key and the humidity as a value parameter. Since Hadoop MapReduce cannot directly process Id, the whole product data are converted into sequence file to be processed using MapReduce. In such a way, one line of the sequence file contains location based humidity. Map function performs parameters calculations on incoming block values and finally sends the block number as a key and list of parameters resultant as a value to the Reduce function.

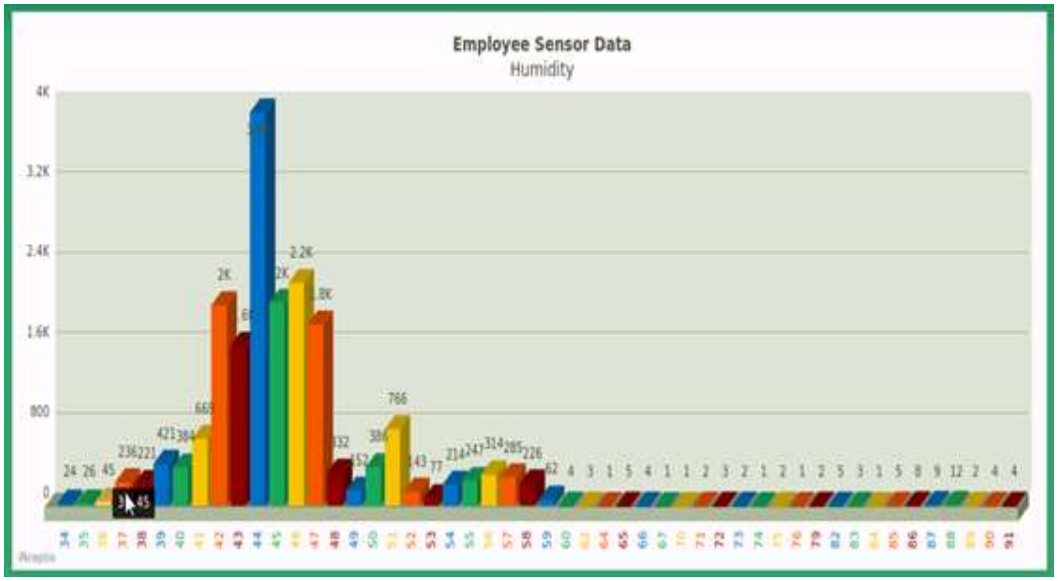


***Figure 4.2: Result for the sensor data set.***

Efficiency measurements are taken by considering the average handling time to process 1-MB data in data set. MapReduce implementation of the analysis algorithm takes less than 1 s the average processing time for various records, which takes 4 s the average processing time. This processing time among various Records varies due to the usage of different bands and image modes, depending on data set. The average processing time for various products is shown in above Figure 3.1. Finally, a comparison is made between the Hadoop MapReduce implementation in the proposed algorithms using average processing time measurements.

The graph shown in Figure 4.2 makes the comparison obvious. For small-size data set, i.e., less than 325 MB, the Hadoop implementation takes more time processing to compute 1 MB data of the records, while a simple Java application is efficient in this case. However, when the product size is become increasing, the average process time starts decreases in MapReduce implementation. Moreover, when the product size exceeds more than 325 MB, it produces better results as compared with the Java implementation. Hence, for smaller size data sets, the Hadoop implementation is not need because of its lots of input and output operations due to Mapper and Reducer function.In the case of high-size dataset, Hadoop divided whole data into blocks and performed parallel tasking and clustering on them, which resulted in increasing productivity.

## CONCLUSTION

The three main units comprises the proposed architecture the three units are First, Data Pre-Processing Unit ($D_{PRE}U$) takes the data from the research site, where processing starts in this unit. Second, Data Analysis Unit ($D_AU$) is the main role in the architecture, and Third, Data Post-Processing Unit ($D_{POST}U$) this unit is responsible for representation. The proposed architecture worked on real time data set which is unclassified and bulky amount of Data which we cannot processed using light weight technologies (java,.net) etc. That raw Data we are taking in to Hadoop Frame work environment in that we are processing the Data set in HDFS and applying MapReduce on that data, in mapper function Clustering happens and the processed data summarizes. That summarized data set will undergo in the process of Reducer it does aggregation on it and it makes all the records in needed manner as understandable format that resultant data set from Hadoop putting in graphical representation. In the graph it specifies the highest humidity and lowest humidity of area specifies in simple way.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

For future work, we are planning 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 real-time, such as earth quake prediction, Tsunami prediction, fire detection, etc.

## REFERENCES

[1] Muhammad MazharUllahRathore, Anand Paul, Bo-Wei Chen,Bormin Huang, and Wen Ji, "Real-Time Big Data AnalyticalArchitecture for Remote Sensing Application," IEEE journal of selected topics in applied earth observations and remote sensing,2015.
[2] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloudcomputing: Current state and future opportunities," in Proc. Int.Conf. Extending Database Technol. (EDBT) , 2011, pp. 530–533.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] P. Chandarana and M. Vijayalakshmi, "Big Data analytics frameworks," in Proc. Int. Conf. Circuits Syst.Commun. Inf. Technol. Appl. (CSCITA), 2014, pp. 430–434.
[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, 15pp., Apr. 2014.
[9] Yan Ma, Haipingwu, Lizhewang, Bormin Huang, Rajiv Ranjan, Albert Zonmaya, weijie, "Remote sensing big data computing: Challenges and opportunities," Article in press, October 2014.
[10] Raghavendra,* Ashwinkumar U M, "A Survey on Analytical Architecture of Real-Time Big Data for Remote Sensing Applications" in Proc. Literature Survey journal.
[11] 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.
[12] D. Rajya lakshmi; K. Kishore Raju; G. P. Saradhi Varma "Taxonomy of Satellite Image and Validation Using Statistical Inference", 2016 IEEE 6th International Conference on Advanced Computing (IACC), Year: 2016, Pages: 352 -361, DOI: 10.1109/IACC.2016.72, IEEE Conference Publications.