Volume 6, Issue 11, November 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Overview of Hadoop in Remote Sensing Image Processing with Various Algorithms and Techniques in Cloud

**Mayank Verma**
Amity School of Engineering and Technology, Amity University,
Noida, Uttar Pradesh, India

***Abstract— In today's era of big data, with the introduction of high resolution systems, remote sensing image processing is one of the fastest growing fields resulting in a rapid increase in volume of data being generated day by day. To handle such massive volumes of data, a high processing speed has become an indispensable requirement. This is possible with the help of big data platforms such as Hadoop. This paper presents a distributed techniques and algorithms which is based on Hadoop platform to process large-scale remote sensing image processing in cloud . The huge volume of data in the modern world, particularly multimedia data, creates new requirements for processing and storage. As an open source distributed computational framework, Hadoop allows for processing large amounts of images on an infinite set of computing nodes by providing necessary infrastructures. There is an increased in large quantity of data with the super-resolution quality data and increased demand in high quality image data . The high performance computing in the field of remote sensing to address the computational requirement for processing of large remote sensing images.***

***Keywords—Cloud Computing, Hadoop, Big Data, Image Processing, MapReduce***

## I. INTRODUCTION

Cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advent of this technology, the cost of computation, application hosting, content storage and delivery is reduced significantly. Cloud computing is a practical approach to experience direct cost benefits and it has the potential to transform a data center from a capital-intensive set up to a variable priced environment.

The idea of cloud computing is based on a very fundamental principal of „reusability of IT capabilities'. The difference that cloud computing brings compared to traditional concepts of "grid computing", "distributed computing", "utility computing", or "autonomic computing" is to broaden horizons across organizational boundaries.

Forrester defines cloud computing as: *"A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end-customer applications and billed by consumption."*[1]

The amount of image data has grown considerably in recent years due to the growth of social networking, surveillance cameras, and satellite images. However, this growth is not limited to multimedia data. This huge volume of data in the world has created a new field in data processing which is called Big Data that nowadays positioned among top ten strategic technologies.

### 1.1 Remote sensing image processing:

Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. It is composed of Hadoop Common, Hadoop Distributed File System, Hadoop YARN and Hadoop MapReduce. HDFS is an open source implementation of the Google file system (GFS). Although it appears as an ordinary file system, its storage is actually distributed among different data nodes in different clusters. MapReduce, a parallel data processing framework pioneered by Google, has been proven to be effective when it comes to handling big data challenges. [2]

### 1.2 How Does Hadoop Work?

In this system, large data files, such as transaction log files, feed reader of social networks, and other data sources are segmented and then distributed in the network. Sharing, storing, and retrieving large files on a Hadoop cluster is undertaken by its distributed file system called HDFS [3]. To increase the reliability of the system, each part of the file is distributed among multiple compute nodes. Therefore, if a node stops working, its file can be retrieved again. From previous studies, it has been observed that image process consists of following steps:

### *A. Images Upload*

Large number of images are acquired from NASA and stored in file system in compressed format.

### *B. Map Reduce Programs*

The objective of this phase is to extract the features of the test image that will be compared to the features of image for image processing operations [7]. On hadoop distributed file system, we execute set of operations like

1) Duplicate image removal
2) Zoom in or zoom out and
3) Find differences among Images using map reduce programs.

### *C. Resultant Image*

The purpose of resultant image generation phase is to generate the resultant image then uploaded in web server and shown to user through web application depending upon the image processing operation selected.

### *D. Hadoop Distributed File System*

To process a large number of images efficiently this Bundle of images is fed to hadoop distributed file system. It is necessary to divide these higher resolution images into multiple segments and assign each image segment to different slave machines to efficiently compare the images. This can be done in distributed environment [7].

There are three types of compute nodes in HDFS. Name management node is responsible for sharing the files and storing the address of each part. Periodic review of nodes and determining their being phased out are also the tasks of Hadoop file management system. Data node that encompasses each one of Hadoop member computers contains file blocks. There is a name management node in Hadoop system for each data node set. The third type is the secondary node that there is a copy of name management node data on it. Therefore if the node stops working, the data will not be lost.

Large files are distributed and further divided among multiple data nodes. The map processing jobs located on all nodes are operated on their local copies of the data. It can be observed that the name node stores only the metadata and the log information while the data transfer to and from the HDFS is done through the Hadoop API.. Fig. 1 shows Structure of HDFS file system.

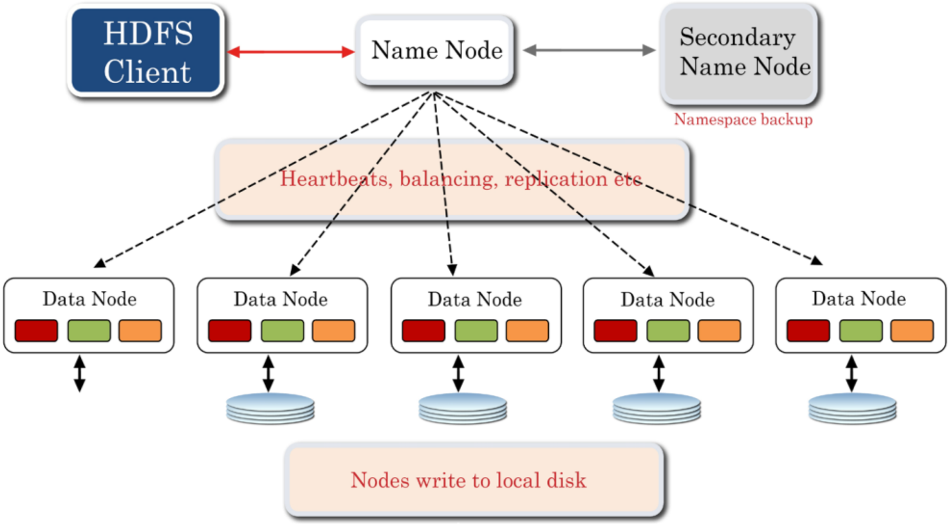

Fig. 1 Structure of HDFS file system.

### 1.3 Hadoop approach to Image Processing

In order to process a large number of images effectively, we use the Hadoop HDFS to store a large amount of remote sensing image data, and we use MapReduce to process these in parallel. The advantages of this approach are three abilities: 1) to store and access images in parallel on a very large scale, 2) to perform image filtering and other processing effectively, and 3) to customize MapReduce to support image formats like TIFF. The file is visited pixel by pixel but accessed whole as one record.

The main attraction of this project will be the distributed processing of large satellite images by using a MapReduce model to solve the problem in parallel. In past decades, the remote sensing data acquired by satellites have greatly assisted a wide range of applications to study the surface of the Earth. Moreover, these data are available in the form of 2D images, which assisted us to manipulate them easily using computers [8].

## II. WORKFLOW EXECUTION IN HADOOP

In the Hadoop environment, processing is executed as a sequence of MapReduce operations, consisting of Map and Reduce phases. Both may have multiple instances executing on different nodes, identified by a key, which is propagated by the data. Data is shuffled between the phases by Hadoop and cached using a built-in caching mechanism. Data is allocated in HDFS in blocks, with adjustable maximum size, and each block is processed in parallel. The simplest approach to Hadoop is the individual processing of input images by assigning unique keys to them. Each input image is loaded to HDFS in one block, and is processed independently. Thus, the execution of the entire workflow can be performed as a single Map phase. Unfortunately, due to the many transformation steps applied within the process (e.g. segments and clusters are created), the analysis of large files may cause performance issues due to memory restrictions.

A more general approach is to partition images into multiple, individually processable blocks, which can be performed by AEGIS beforehand. However, in order to determine the appropriate partitioning methodology, one must examine the properties (input, working set and output) of the algorithms. Consider the following.

- A reflectance and spectral index computation are local operations, and as such the partitioning of images does not influence their computation, as long as each pixel component is available locally. As such, both can be computed in a single Map phase as part of the preprocessing [8].
- Segmentation is a regional image operation, where spatially neighboring pixels are required for evaluation. The examined regions are of irregular shape, and may be different for each algorithm and even each configuration of an algorithm. The result of the segmentation is a segment map, which can be represented by a set of pixels for each segment.
- Clustering is also a regional operation, but in the multispectral space domain of the image, where the location of pixels (spectral vectors) is independent of their spatial location. Either the original image or the segment map can be specified for input, and the output is a cluster map (similar to the segment map).
- The reference based classification relies on the matching of the reference image to the result of the classification. It is also performed on pixel level.

### 2.1 Hadoop Image Processing Interface (HIPI)

HIPI is an image processing library designed to be used with the Apache Hadoop Mapreduce parallel programming framework [7]. HIPI facilitates efficient and high-throughput image processing with MapReduce style parallel programs typically executed on a cluster. It is flexible enough to withstand continual changes and improvements within Hadoop's Mapreduce system. The goal of HIPI is to create a tool that will make development of large-scale image processing and vision projects extremely accessible [9].

## III. ALGORITHMS USED IN REMOTE SENSING IMAGES

### *A. Sobel Edge Detection Method*

The Sobel edge detector is used to detect edges and is based on applying horizontal and vertical filters in sequence. Both filters are applied to the image and summed to form the final result. Edges in images are areas with strong intensity contrasts. Detecting the edges in an image significantly reduces the amount of data and filters out useless information while preserving the important structural properties in an image.

The Sobel operator applies a 2-D spatial gradient measurement to an image and is represented by an equation. It is used to find the approximate gradient magnitude at each point in an input greyscale image. The discrete form of the Sobel operator can be represented by a pair of 3 × 3 convolution kernels, one estimating the gradient in the x-direction (columns) and the other estimating the gradient in the y-direction (rows). The Gx kernel highlights the edges in the horizontal direction; while the Gy kernel highlights the edges in the vertical direction. The overall magnitude of the outputs, |G|, detects edges in both directions and is the brightness value of the output image.

### *B. Laplacian Edge Detection*

The Laplacian Edge Detection method is a second-order derivative of an image and it is applied by convolving the non directional Laplacian filter. The second order derivative an edge will have a zero crossing in the region where there is the highest change in intensity. Therefore the location of the edge can be obtained by detecting the zero-crossings of the second-order derivative of the image and this is known as Laplacian filter which is an effective detector for non-sharp edges where the pixel intensity level change over space slowly. A single filtering kernel of different sizes (e.g. 3x3, 5x5, 7x7, etc.) that has low values (usually negative) in the middle of the kernel surrounded by positive values can be used as Laplacian edge detection. Because the Laplacian is an approximation of the second-order derivative of an image preserving the high frequency components, it is very sensitive to noise and therefore it is usually applied to an image that has first been smoothed using the Gaussian filter in order to suppress noises in the image. [10]

### *C. Canny Edge Detection*

The Canny edge detection is considered as the optimal and standard edge detector. This multi-step method which was developed by the canny, which aims to develop optimal algorithms that satisfies three main criteria. The first one is good edge detection by maximizing the signal-to-noise ratio meaning the method should detect edges to the maximum possibility but with low probability of detecting edges falsely. The second criterion is that detected edges should be as close as possible to the real edges. The third criterion is to have minimal number of response and edges should not be detected more than once.

## IV. ADVANTAGES AND DISADVANTAGES OF HADOOP

- The most important advantage of Hadoop is the ability to process and analyze large amounts of unstructured or semi-structured data which have been impossible to process efficiency (cost and time) so far [4].
- The next advantage of Hadoop is its simple expansion and horizontal scalability. Data can easily be analyzed up to Exabyte level and there is no need for companies to work on sample data and a subset of the original data. With the help of Hadoop, the possibility of checking all types of data is provided.
- Another advantage is its low set up cost, mainly because it is free and there is no need for expensive and professional hardware. In particular, with the spread of cloud computing and its reasonable prices for case processing of data as well as private clouds, it takes only a few hours to set up a Hadoop system [5].

- On the other hand, Hadoop and its subsets are all in the early stages of development and they are unsteady and immature. This will lead to permanent modification of this framework that imposes costs of continuous training on organizations.
- On the other hand, because of novelty of this software model, a few people have the necessary skills for establishing and working on Hadoop-based systems. Lack of expert manpower is the most important challenge of many companies in using this system.
- Also the novelty of this technology causes the lack of valid standards and benchmarks for evaluating different algorithms in this area. Bajcsy *et al*. attempted to assess four different methods of Hadoop-based image processing on cluster [6]. This is one of the few efforts in this area and still we are far from establishing comprehensive benchmarks which are acceptable to academic community.
- Another Hadoop's problem which has an inherent nature is lack of the ability of real-time data processing. The request tracker must wait for each compute node in the system to finish the work, and then it can deliver the final answer to the user. However, this problem will be solved to some extent by the rapid growth of NoSQL databases' technologies and its combination with Hadoop. Moreover, frameworks such as Storm [11] and Samza [12] can also be used for real-time processing of high volume data.

## V. CONCLUSION

In this paper, we presented a case study processing of remote sensing images using the Hadoop framework, algorithms and techniques. The image processing algorithms can be effectively parallelized with acceptable run times when applied to large remote sensing images. Hadoop installed on a cluster of computer in a parallel way to proved suitable to process images in large quantities. We have observed that this Hadoop implementation is better suited for large data sizes, when a computationally remote sensing application is required. In the future, we might focus on using different image sources with different algorithms that can have a computationally intensive nature.

## REFERENCES

[1] http://www.thbs.com/thbs-insights/cloud-computing-overview.

[2] Wang, C et al. **"***ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,"* USA, II.4 , pp. 63-66, 2015.

[3] S. M. Banaei1, H . K. Moghaddam, *"Hadoop and Its Role in Modern Image Processing,"* in Open Journal of Marine Science, 4, pp. 239-245, 2014.

[4] Bakshi, K. (2012) Considerations for Big Data: Architecture and Approach. Aerospace Conference-Big Sky, MT, 3-10 March 2012.

[5] Kelly, J. (2012) Big Data: Hadoop, Business Analytics and Beyond. Wikibon Whitepaper, 27 August 2012. http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond.

[6] Bajcsy, P. et al. "Terabyte-Sized Image Computations on Hadoop Cluster Platforms," in 2013 *IEEE International Conference on Big Data*, Silicon Valley, pp. 729-73 , 6-9 October 2013.

[7] Sarade Shrikant, Ghule, "Large scale satellite image processing using Hadoop distributed system" in *International journal of advance research in computing engineering and technology (IJARCET)* vol. 3, issue 3, march 2014.

[8] Roberto Giachetta, Istv´an Fekete, "A Case Study of Advancing Remote Sensing Image Analysis in *Acta Cybernetica*," 22 , pp. 57–79, 2015.

[9] Helly M. Patel et al, "Large Scale Image Processing Using Distributed and Parallel Architecture" in *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 6 (6) , pp. 5531-5535, 2015.

[10] Rupal Yadav, M. C. Padma, "Processing of Large Satellite Images using Hadoop Distributed Technology and Mapreduce : A Case of Edge Detection," in *International Journal on Recent and Innovation Trends in Computing and Communication*, ISSN:2321-8169, vol. 3, issue 5, 2015.

[11] Storm Project. http://storm.incubator.apache.org/

[12] Samza Project. http://samza.incubator.apache.org/