LARGE SCALE SATELLITE IMAGE PROCESSING USING HADOOP DISTRIBUTED SYSTEM

Sarade Shrikant D., Ghule Nilkanth B., Disale Swapnil P., Sasane Sandip R.

***Abstract*- The processing of large amount of images is necessary when there are satellite images involved. Now a day's amount of data continues to grow as more information becomes available. With this increasing amount of surface and recognition, segmentation, and event detection in satellite images with a highly scalable system becomes more and more desirable. In this paper, a semantic taxonomy is constructed for the land-cover classification of satellite images. Whole system is constructed in a Distributed HADOOP Computing platform. This system is divided into two major part Training and Running classifier. The Training part classifies subsequent images such as Vegetation, Building, Pavement, water, snow etc. 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. Training classifier is implemented using HADOOP MapReduce Framework and is based on Google Earth. The Running classifier performs zoom-in, zoom-out and calculates the difference between old and new images.**

***Index Terms*- event detection, hadoop, MapReduce, satellite, segmentation.**

I. INTRODUCTION

Hadoop is an open source framework for writing and running distributed applications that process large amounts of data, the basic for writing a scalable, distributed data-intensive program

Everyday, in this modern era we're surrounded by information in the form of data like people upload videos, take pictures on their cell phones, text friends, update their Facebook status, leave comments around the web, click on ads, and so forth. You may even be reading the book as digital data on your computer screen, and certainly your purchase of this book is recorded as data with some retailer. The exponential growth of data presents the challenges to cutting-edge business such as Google, Yahoo, Amazon, and Microsoft. They needed to go through terabytes and petabytes of data to figure out which websites were popular, what books were in demand, and what kinds of ads appealed to people. Existing tools were becoming inadequate to process

such large data sets. Google was the first to publicize MapReduce- a system they had used to scale their data processing needs. Doug Cutting saw an opportunity and led the charge to develop an open source version of this MapReduce system called Hadoop. Soon after, Yahoo and others rallied around to support this effort. Today, Hadoop is a core part of the computing infrastructure for many web companies, such as Yahoo, Facebook , LinkedIn, and Twitter. Many more traditional businesses, such as media and telecom, are beginning to adopt this system too. Large-scale distributed data processing in general, is rapidly becoming an important skill set for many programmers.

Apache Log Processing with Cascading		
1 Node	Runtime	21m46s
	Sec/MB	0.127
	Sec/MB/Node	0.127
3 Nodes	Runtime	8m3s
	Sec/MB	0.0471
	Sec/MB/Node	0.0157
15 Nodes	Runtime	1m30s
	Sec/MB	0.00878
	Sec/MB/Node	0.000585
Naive Perl	Runtime	42m49s
	Sec/MB	0.251
	Sec/MB/Node	0.251

Table 1.1 Apache Log Processing with Cascading

The events occur on earth surface are possible to detect. For example- The events such as flooding, tsunami and snow storm etc. can be detected from the measurable change in ground surface cover as a result of damage to existing structures.

Satellite images requires very huge amount of data storage. Large Scale Land-cover Recognition System Collects large amount of data of higher resolutions Satellite Images. It provides a collection of training data classifiers and performing subsequent image classification in distributed environment.

MapReduce is a popular parallel model first introduced by Google, which is designed to handle and generate large scale data sets in distributed environment. It provides a convenient way to parallelize data analysis process. Its advantages include conveniences, robustness, and scalability.

II.LITERATURE SERVEY

In the internet service data base is most important part in that database image are stored in the world large data of images so that handling is very difficult just like ,the duplication of image it's increases the data size .We are all known about if that data was big the processing time also more With the proliferation of online photo storage and social media from websites such as Facebook and Picasa, the amount of image data available is larger than ever before and growing more rapidly every day . the billions of images available to us on the web. These images are improve, however, by the fact that users are supplying tags (of objects, faces, etc.), comments, titles and descriptions of this data for us. This information produces with an amazing amount of unprecedented context for images. The idea can be applied to a wider range of image features that allow us to examine and analyze images in a revolutionary way. The current processing of images goes through ordinary sequential ways to accomplish this job. The program loads image after image, The processing of data today is done by using oracle versions such as 9i,10G or by any another DBMS software. But with the increasing usage of internet all over the world the data on net is increasing rapidly. So, the processing of mass data is not possible by using any oracle software or any another existing DBMS software. the report generated after analysis will help the user to know about his usage. For analyzing the data & images we are going to use Hadoop technology.

III. EXISTING SYSTEM

Current processing of images goes through ordinary sequential ways to accomplish this job. The program loads image after image, processing each image alone before writing the newly processed image on a storage device. Generally, we use very ordinary tools that can be found in Photoshop. Besides, many ordinary C and Java programs can be downloaded from the Internet or easily developed to perform such image processing tasks. Most of these tools run on a single computer with a Windows operating system. Although batch processing can be found in these single-processor programs, there will be problems with the processing due to limited capabilities. Therefore, we are in need of a new parallel approach to work effectively on massed image data.

IV. PROPOSED SYSTEM

From previous studies, it has been observed that image process consists of following steps:

1 Images Upload
2 Hadoop Distributed File System.
3 MapReduce Programs
4 Resultant Image

Images Upload: Large number of images are acquired from NASA and stored in file system in compressed format. Fig.1 shows some sample images stored in file system database.

Fig.4.1 Sample Images

Hadoop Distributed File System: To process a large number of image efficiently this Bundle of images is fed to hadoop distributed file system. The acquired signature image as as shown in Fig3.1. 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.

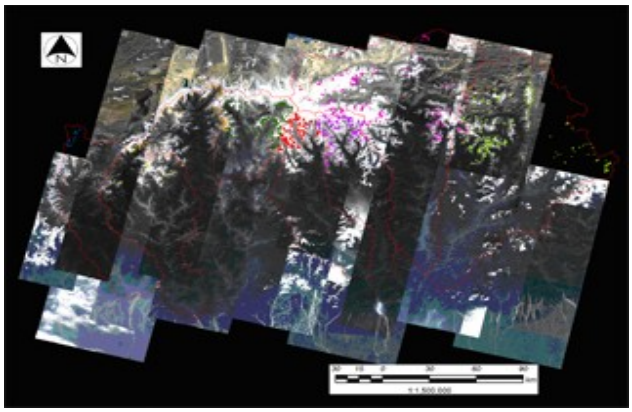

Fig 4.2 Segmented image

MapReduce: 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. On hadoop distributed file system, we execute set of operations like (i)

Duplicate image removal(ii) Zoom in or zoom out and (iii) Find differences among Images using map reduce programs.

Resultant Image Upload: 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.

Fig 4.3 Resultant image

V. SYSTEM ARCHITECTURE

A Large Scale Land-cover Recognition System is essentially a web application supported by a backend database. Large Scale Land-cover Recognition System is programmed in languages such as Java and AJAX.

The basic idea is to implement MapReduce to split the large input data set into many small pieces and assigned small task to different devices. A scalable modeling system implemented in the Hadoop MapReduce framework is used for training the classifiers and performing subsequent image classification.

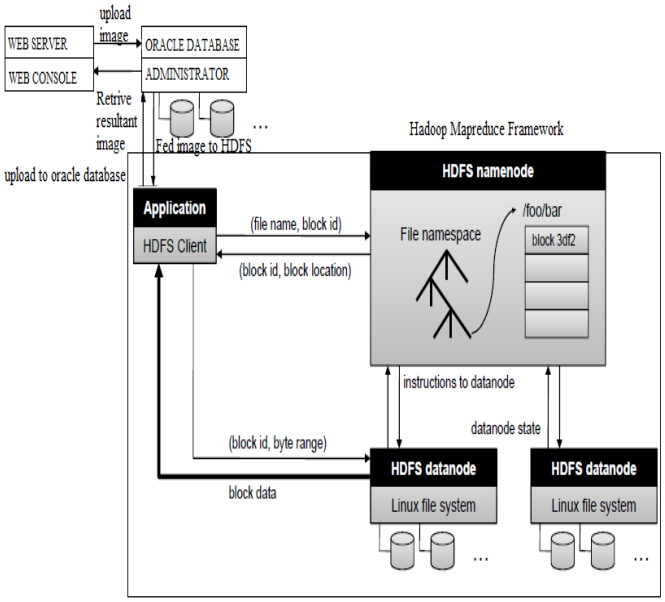

Fig 4.1 System Architecture

Following is the working of system:

1. Large no. of images stored in file system.
2. This Bundle of images is fed to Hadoop distributed file system.
3. On HDFS, we execute set of operations like duplicate image removal, zoom in and find differences among Images, using MapReduce Programs
4. The Result is then uploaded in web server and shown to user through web application.

Satellite image data continues to grow and evolve as higher spatial and temporal resolutions become available.

With sufficient spatial and temporal resolutions, event detection becomes possible. With this increasing amount of surface and temporal data, recognition, segmentation and event detection in satellite images with a highly scalable system becomes more and more desirable. In this paper, a semantic taxonomy is constructed for the land-cover classification of satellite images.

VI. ALGORITHM

Much of what the layperson thinks of as statistics is counting, and many basic Hadoop jobs involve counting. We can write a MapReduce program for this task. Like we said earlier, you hardly ever write a MapReduce program from scratch. You have an existing MapReduce Counting things program that processes the data in a similar way.

We already have a program for getting the inverted citation index. We can modify that program to output the count instead of the list of citing patents. We need the modification only at the Reducer. If we choose to output the count as an IntWritable, we need to specify IntWritable in three places in the Reducer code. We called them V3 in our notation.

For Example

```
public static class Reduce extends MapReduceBase
implements Reducer<Text, Text, Text, IntWritable> {
public void reduce(Text key, Iterator<Text> values,
OutputCollector<Text, IntWritable> output,
Reporter reporter) throws IOException {
int count = 0;
while (values.hasNext()) {
values.next();
count++;
}
output.collect(key, new IntWritable(count));
}
}
```

By changing a few lines and matching class types, we have a new MapReduce Program. This program may seem a minor modification. We expect a large number of patents to have been only cited once, and a small number may have been

cited hundreds of times. It would be interesting to see the distribution of the citation counts.

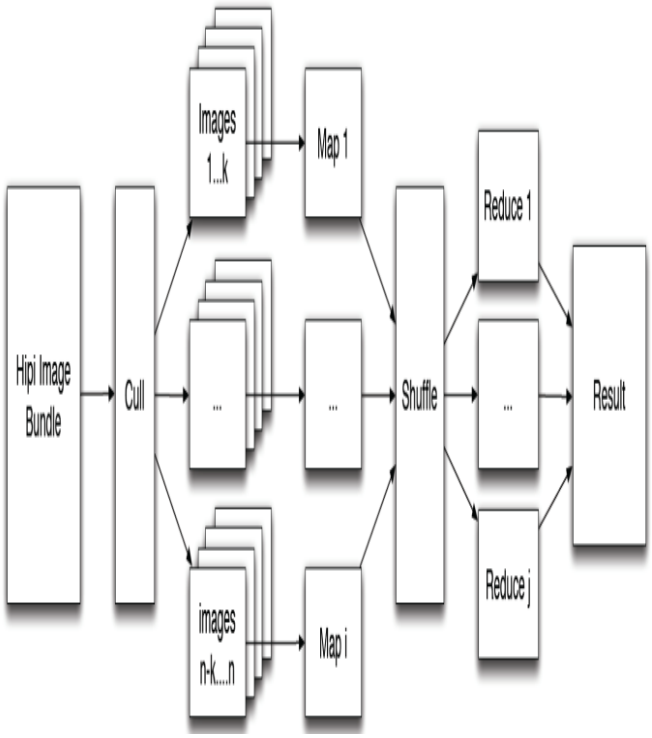

Fig 6.1 Map Reduce Algorithm

VII. SYSTEM FEATURES

This product is independent and self-contained which includes following components:

- Scaling
- Feature Extraction
- Recognition

Scaling: Image scaling is the process of resizing a digital image. Scaling is a non-trivial process that involves a trade-off between efficiency, smoothness and sharpness. As the size of an image is increased, so the pixels which comprise the image become increasingly visible, making the image appears "soft". Conversely, reducing an image will tend to enhance its smoothness and apparent sharpness.

Feature Extraction: MapReduce allows the computation to be done in two stages: the map stage and then the reduce stage. The data are split sets of key–value pairs and their instances are processed in parallel by the map stage, with a parallel number that matches the node number dedicated as slaves. This process generates intermediate key–value pairs that are temporary and can later be directed to reduce stages. Within map stages or reduce stages, the processing is conducted in parallel. The map and reduce stages occur in a sequential manner by which the reduce stage starts when the map stages finishes.

Recognition: On hadoop distributed file system, image recognition 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.

VIII. FUTURE ENHANCEMENT

In the future, we might focus on using different image sources with different algorithms that can have a computationally intensive nature. So our application will be beneficial in future to the following sectors:

- Meteorological disaster - Violent, sudden and destructive change to the environment related to, produced by, or affecting the earth's atmosphere, especially the weather-forming processes.
- Military navigation - study of traversing through unfamiliar terrain by foot or in a land vehicle.
- Monitoring around the globe- to extract discriminative information about regions of the globe for which GIS data is not available.

IX. CONCLUSION

In this paper a case study is presented for implementing parallel processing of remote sensing images in TIF format by using the HadoopMapReduce framework. The experimental results have shown that the typical image processing algorithms can be effectively parallelized with acceptable run times when applied to remote sensing images. A large number of images cannot be processed efficiently in the customary sequential manner. Although originally designed for text processing, HadoopMapReduce installed in a parallel cluster proved suitable to process TIF format images in large quantities. Thus we have decide to implementation parallel Hadoop which is better suited for large data sizes than for when a computationally intensive application is required.

X. REFERENCES

[1] Towards Large Scale Land-cover Recognition of Satellite Images Noel C. F. Codella, Gang Hua, ApostolNatsev, John R. Smith

[2] H. Daschiel and M. Datcu. Information mining in remote sensing image archives: system evaluation. IEEE Trans. on Geoscience and Remote Sensing, 43(1):188–199, 2005.

[3] G. M. Foody. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-

sensed data. International Journal of Remote Sensing , 17(7):1317–1340, 1996.

[4] Y. Li and T. R. Bretschneider. Semantic-sensitive satellite image retrieval.IEEE Transactions on Geoscience and Remote Sensing , 45(4):853–860, April 2007.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision , 60:91–110, November 2004.

[6] Y. Li and T. R. Bretschneider. Semantic-sensitive satellite image retrieval. IEEE Transactions on Geoscience and Remote Sensing , 45(4):853–860, April 2007.

[7] P. M. Atkinson and A. R. L. Tatnall. Neral networks in remote sensing. International Journal of Remote Sensing , 18(4):699–709, April 1997

[8] A. Carleer, O. Debeirb, and E. Wolff. Comparison of very high spatial resolution satellite image segmentations. In L. Bruzzone, editor, Proc. of SPIE Image and Signal Processing for Remote Sensing IX , volume 5238, pages 532–542, Bellingham, WA, 2004.

[9] A. Parulekar, R. Datta, J. Li, and J. Z. Wang. Large-scale satellite image browsing using automatic semantic categorization and content- based retrieval. In Proc. ICCV'2005 Workshop on Semantic Knowledge in Computer Vision , pages 1873–11880, Beijing, China, October 2005

[10] G. G. Wilkinson. Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Trans. on Geoscience and Remote Sensing, 43(3):433–440, 2005

[11] W. Messaoudi, I. R. Farah, K. S. Ettabaa, and B. Solaiman. Semantic strategic satellite image retrieval. In Proc. of 3rd International Confer- ence on Information and Communication Technologies: From Theory to Applications, pages 1–6, Damascus, April 2008.

Authors

SaradeShrikant D.
B.E.Computer
University of Pune
Department of Computer Engg.
Vishwabharti Academy's college of Engineering Ahmednagar.
Tal- Ahmednagar,Dist-Ahmednagar. India.

GhuleNilkanth B.
B.E.Computer
University of Pune
Department of Computer Engg.
Vishwabharti Academy's college of Engineering Ahmednagar.
Tal- Ahmednagar,Dist-Ahmednagar. India. .

Disale Swapnil P.
B.E.Computer
University of Pune
Department of Computer Engg.
Vishwabharti Academy's college of Engineering Ahmednagar.
Tal- Ahmednagar,Dist-Ahmednagar. India.

.

Sasane Sandip R.
B.E.Computer
University of Pune
Department of Computer Engg.
Vishwabharti Academy's college of Engineering Ahmednagar.
Tal- Ahmednagar,Dist-Ahmednagar. India. .

