Study on Implementation of Medical Image Processing with Big Data Analytics and Map Reduce

D. Peter Augustine
Assistant Professor, Computer Science Department
Christ University
Bangalore, India
peter.augustine@christuniversity.in

***Abstract*- In this paper, we analyze and reveal the benefits of Big Data Analytics and Hadoop in the image processing. Healthcare industry produces the massive volume of data every day. There are huge amount of images produced by various instruments on a patient in various situations. There are various image processing techniques and algorithms evolving to get the best accurate information from the images to lead to a proper diagnosis. The current and fast growing concepts of Big Data and Hadoop can contribute more to the image analysis to yield the best result. This paper gives the involvement of Big Data Analytics and Hadoop in medical image analysis with the help of the HIPI and Map Reduce implemented by different authors and takes the lead to effective implementation of Big Data Analytics (BDA) and Map Reduce in the future.**

Keywords*— *Image Processing, Big Data Analytics, Hadoop, Hadoop Distributed File System, Map Reducing, Image Processing Algorithms.

I. INTRODUCTION

The increasing amount of medical image data produced on a daily basis in present hospitals compels the use of traditional medical image investigation and indexing approaches towards the optimum solutions. During the past 20 years, there is a remarkable increase in the number of images and their dimensionality. Recent development in image processing makes it possible to assist the healthcare professionals in the discovery and classification of vital events in large image series. On the other hand, the process of obtaining complicated features from large datasets of 3D/4D images needs highly advanced software applications, hardware and the recent technologies.

Today, the recent visualization tools and techniques offer enormously precise and high quality 3D/4D images of anatomical structures of human body. Yet, the utilization of those images for effective analysis is not as need of time, because of the intricate construction of medical images of various anatomical organs combined together which cannot give the clear view to the doctors.

Image Processing is usually measured as a very complicated problem due to very large size of datasets, complexity, and disparity of the anatomic organs. The borders of anatomical structures to be imprecise and detached due to noise and low contrast of the image. So it may be a big challenge even for segmenting the images to get the areas of necessity and extract them from the remaining datasets.

There are many algorithms with different approaches for the image processing in literature. But their outcome and the analysis will differ extensively conditional on the specific application, imaging modality (CT, MRI, etc.), and other factors. The algorithm, which provides ideal results for one application, might not even work for another. The common imaging relics such as noise, motion and partial volume effects can have their impact on the efficiency of the algorithms. This variety in requirements creates a challenging problem for the segmentation algorithms. In the current scenario, there are no such 100 percent accurate methods yielding suitable results for any type of medical dataset. Generally, image processing will need to go through various steps with consideration of various external factors to get the accurate outcome.

In this paper, I present a study on the Big Data Analytics with respect to mounted up medical images and the map reduce technique for the bundle of images in the distributed environment to ease the process of image processing to get the accurate result for exact analysis. The study has been enhanced with the help of the HIPI and the results achieved by researchers from implementing map reduce for image processing.

II. BIG DATA AND MEDICAL IMAGE

The mostly considered domains involving Big Data are

- Healthcare
- Public sector
- Retail
- Manufacturing

Big data can produce value in each of the domains mentioned above. According to the statistics if US healthcare were to use big data productively and efficiently, the domain could create more than $300 billion in value every year. US healthcare expenditure can be reduced to 66.66 percent which is approximately 75 percent currently. The same beneficial scenario is observed in the other domains also.

1) Variety: Data variety is exactly as the word depicts. It comes from the mode of imaging techniques such as X-Ray, MRI, CT scans, PET and functional MRI scans, the various formats of images and more. It also comes from the variety of

devices generating images and the various situations they were captured.

2) Velocity: Big Data deals with the rapidity at which data floods in from sources like image capturing machines, networks and human interaction like healthcare professionals discussions etc. The flow of data for the storage and analysis is immense and nonstop. This real-time data can help researchers and healthcare professionals create precious decisions that provide strategic analytic advantages.

3) Volume: Data volume is clear that each moment the data is getting accumulated from various sources. The size of the data ranges from kilobytes to petabytes. The data may be generated by machines, networks and human interaction on systems.

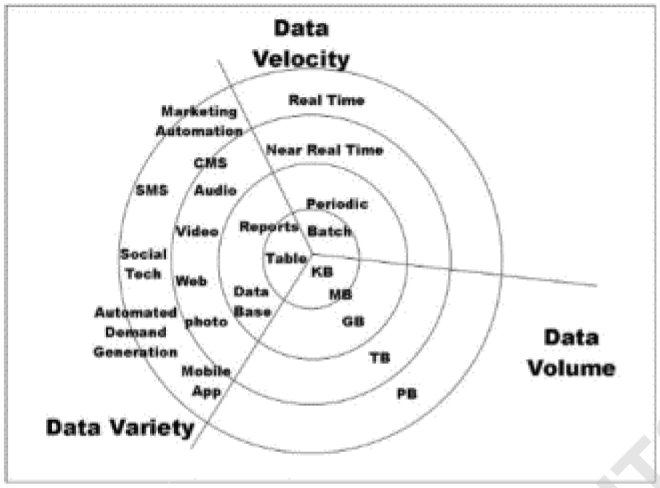

Fig. 1. Big Data's 3 v's

One of the current hottest topics in medicine from both a research and clinical perspective is Big Data. One of the major challenges with medical imaging is the difficulty of discovery of imaging information in the electronic medical record. Our imaging reports are, almost without exception, unstructured and our medical images are rarely tagged in such a way as to be discoverable or useful to data mining efforts. This must change if medical imaging is to play a substantial role in this era of big data and personalized medicine in healthcare. The role of big data is crucial in diagnostic imaging and speculations help in visualization of images and data, as well as diagnosis and treatment.

The influence of big data may be intense in various domains, and it will have extensive inferences for medical imaging as healthcare needs to follow, handle, make the most of, and state relevant patient information. How one can deliberately gather, store up, preserve, and then make the data producing the expected and exact value for the medical professional.

With the above understandings we can affirm that Big Data can be used in several ways, including the following:

- To improve early discovery, analysis, and medical treatment.
- To foresee patient's future health.
- To amplify interoperability and interconnectivity of healthcare so that the medical professional can gain the needed knowledge from anywhere in the world.
- To enhance patient care by means of remote analysis, remote care and remote medicine by the information gathered from home devices

III. APACHE HADOOP

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

As the framework is already being used for large-scale data analysis tasks by many companies such as Facebook and Yahoo, and at the same time is easily adapted for use with any kind of hardware, ranging from a single computer to large data center. Hadoop because of its salient features can be best opted for image processing on the MapReduce model. A typical Hadoop cluster consists of a master node and any number of computing nodes. The purpose of the master is to interact with users, monitor the status of the computing nodes, keep track of load balancing and handle various other background tasks. The computing nodes deal with processing and storing the data.

As the framework is already being used for large-scale data analysis tasks by many companies such as Facebook and Yahoo, and at the same time is easily adapted for use with any kind of hardware, ranging from a single computer to large data center. Hadoop because of its salient features can be best opted for image processing on the MapReduce model. A typical Hadoop cluster consists of a master node and any number of computing nodes. The purpose of the master is to interact with users, monitor the status of the computing nodes, keep track of load balancing and handle various other background tasks. The computing nodes deal with processing and storing the data.

IV. MAP REDUCE

The MapReduce framework is a distributed computing framework and has recently been used for large–scale image description and analysis.

Hadoop's MapReduce is a software framework which supports to develop applications to process parallel huge quantity of data like multi-terabyte data sets in a consistent and fault tolerant way on thousands of nodes working together.

The input data set is split into independent large pieces by a MapReduce job and then the pieces are processed by the predefined map tasks in a completely parallel manner. The outputs of the maps are sorted by the framework, and then they are given as input to the reduce tasks. The file system

stores both the input and the output of the job. The Hadoop framework automates of scheduling the jobs, scrutinizing them and executes them again if a task fails.

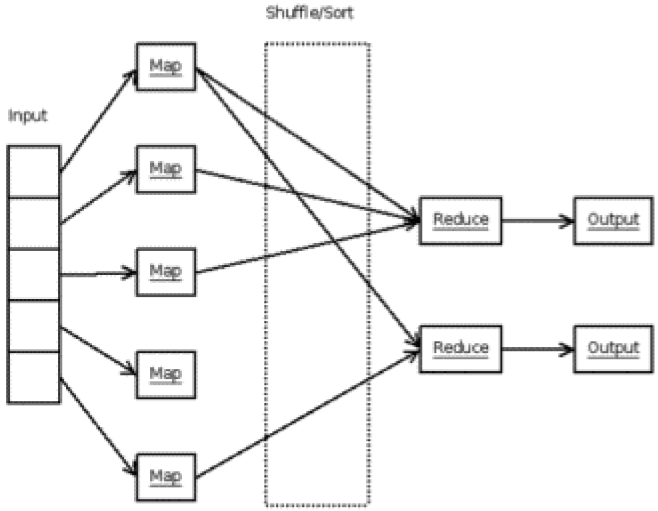

Fig. 2. Hadoop's Map Reduce

Since the MapReduce framework and the Hadoop Distributed File System are running on the same set of nodes, the computing and storing nodes are the same. Because of this configuration, the Hadoop framework successfully arranges odd jobs on the nodes where data is already available. This causes very high cumulative bandwidth across the cluster.

Each cluster node contains a single master JobTracker and one slave TaskTracker in the MapReduce framework. The scheduling of the jobs is done by the Master whereas the slaves work on the component tasks, monitoring them and re-executing the failed tasks. The master is responsible to direct the slaves execute the tasks.

Because of the efficiency of the Hadoop Framework we can achieve the following benefits for the Image Processing in Healthcare.

1)Scalable: Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.

2)Cost effective: Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

3)Flexible: Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data.

4)Fast: Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster.

5)Resilient to failure: A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

V. MEDICAL IMAGE PROCESSING

The field of medicine needs various types of images of a same person in different situations from different devices. The result emerging from the captured images has great impact in the diagnosis. Imaging has become a necessary component in many fields of medical practice to identify, understand and rectify the health problems. X-Ray, MRI, CT scans, PET and functional MRI scans are the instruments for producing various images. Sophisticated computerized quantification and visualization tools are required to analyse of these varied types of images to mine the accurate and investigative result.

The National Electrical Manufacturers Association (NEMA) created the Digital Imaging and Communications in Medicine (DICOM) standard. The aim of DICOM is to support the distribution and viewing of medical images from X-Ray, MRI, CT scans, PET and functional MRI scans. The DICOM format is an extension of the older NEMA standard. A header and the image data are the parts of a DICOM file. The header stores information about the patient's name, the type of scan done to get the respective image, position and dimension of image and lots of other related data. The image data part includes all the image related information. To reduce disk space the DICOM image data can be compressed either lossless or lossy. DICOM is the common standard used for scans from hospitals. Medical Image processing uses real medical images and the supporting environment to demonstrate and explain concepts and to construct perception, imminent and thoughtful.

VI. IMAGE BASED MAP REDUCE

Standard Hadoop MapReduce programs are capable of handling input and output of data very efficiently. But they find difficulty in representing images in a format that is useful for scholars. Currently, the methods take extra burden to get the representation of standard float image. For example, the user has to pass the images as a string to supply a set of images to a set of Map nodes, then the decoding of each image in each map task before being able to get the pixel information. This show not only the inefficiency, but also the inconvenience caused for the user. It creates extra headache for the users and makes the code messy for understanding and debugging. The authors of HIPI say that their library focuses on bringing familiar image-based data types directly to the user for easy use in MapReduce applications.

The user only needs to specify a HIPI Image Bundle as an input, and HIPI will take care of parallelizing the task and sending float images to the mappers. An input specification created using the HIPI Image Bundle data type as inputs, will distribute images in the HIPI Image Bundle across all map nodes. The images are distributed such that there is an attempt to maximize locality between the mapper machines and the machine where the image resides. Typically, a user would have to create InputFormat and RecordReader classes that specify how the MapReduce job will distribute the input, and what information gets sent to each machine. This task is nontrivial and often becomes a large point of headaches for

users. InputFormat and RecordReaders are included to take care of this for the user. The specification works on HIPI Image Bundles for various image types, sizes, and varying amounts of header and exif information. All of these different image permutations are handled behind the scenes to bring images straight to the user as float images. No work is needed to be done by the user, and float images are brought directly to the Map tasks in a highly parallelized fashion.

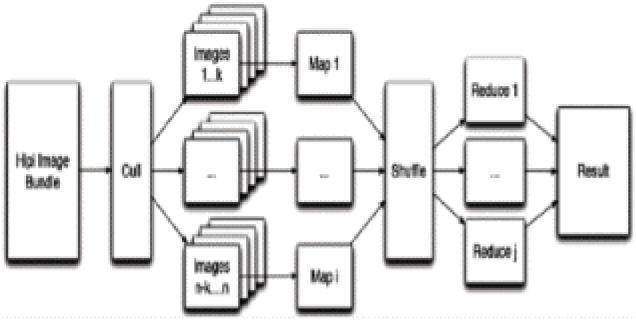

Fig. 3. Hadoop Image Processing Interface

During the distribution of inputs but before the map tasks start we introduce a culling stage to the MapReduce pipeline. The culling stage allows for images to be filtered based on image properties. The user specifies a culling class that describes how the images will be filtered (e.g. pictures smaller than 10 megapixels, pictures with GIS location header data). Only images that pass the culling stage will be distributed to the map tasks, preventing unnecessary copying of data. This process is often very efficient because culling often occurs based on image header information, so it is not required to read the entire image.

Additionally, images are distributed as float images so that users can immediately have access to pixel values for image processing and vision operations. Images are always stored as standard image types (e.g. JPEG, PNG, etc.) for efficient storage, but HIPI takes care of encoding and decoding images to present the user with float images within the MapReduce pipeline. As a result, programs such as calculating the mean value of all pixels in a set of images can be written in merely lines. There are operations such as cropping for image patches extraction. It is often desirable to access image header and exif information without need for pixel information, so authors have abstracted this information from the pixel data. This is particularly useful for the culling stage, and for applications such as im2gis3 that need access to metadata. Presenting users with intuitive interfaces for accessing data relevant to image processing and vision applications will allow for more efficient creation of MapReduce applications.

A. Image Processing with Map Reduce

1) Implementation with Hadoop:

Portioning the images is the first job before taking them into the Hadoop's Map Reduce. Having partitioned the image into pieces that fit into memory, the next step is to design a MapReduce program to operate on this data. Since, by specifying overlaps and fitting the pieces within the HDFS block size during the partitioning phase, we have already ensured that each instance of the algorithm has its necessary data locally available. Also, since we can easily perform all necessary computations in the Map phase, there is no need for a 36 Reducer - Hadoop can be configured to simply write output to storage after the Map phase. Therefore, in this case, the MapReduce program consists only of three definitions: InputFormat, Mapper and OutputFormat. Their respective purposes are straightforward: read blocks from HDFS and convert them to Java objects that contain the name, dimensions and pixel values of the block's contents (one block contains one piece of the complete image), process these pieces with the fast O(1) bilateral filter, and finally convert the resulting objects back to PNG files and write to HDFS. Similarly to the other practical example presented in the previous section, the Key is the filename of the image (the filenames signifying which piece of the full image they represent), and the Value is a Java object containing the image.

2) Testing

Parameters of the m1.small and m2.xlarge instance types according to the Amazon EC2 official web page. One EC2 Compute Unit can be thought of as the equivalent of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Instance type	m1.small
Memory	1.7 GiB
CPU	1 virtual core with 1 EC2 Compute Unit
Local storage	160 GB
Platform	64-bit
Instance type	**m2.xlarge**
Memory	17.1 GiB
CPU	2 virtual cores with 3.25 EC2 Compute Units each
Local storage	420 GB
Platform	64-bit

Fig. 4. Parameters of instance types for testing images with Map Reduce

All tests were run on a Hadoop cluster with one m1.small virtual machine as the master node and m2.xlarge virtual machines as computing nodes on the Amazon EC2 cloud. With regard to the configuration of the Hadoop cluster, most parameters remained set to the default values both in runs with version 0.20.2 and 1.0.3. The only exceptions were setting the HDFS block size to 64 megabytes, and setting the maximum memory for Map and Reduce tasks to 15 000 megabytes. The choice of m2.xlarge instances for computing nodes was directly influenced by the requirements of the algorithm - all attempts to run the tests with 37 m1.small instances failed because there was not enough memory available. Using m2.xlarge eliminated these problems and due to the number of cores, also allowed simultaneous processing of two images. In the following, the author will present the results of testing the fast O(1) bilateral filter algorithm in various configurations.

The principal results of testing can be seen in figure 5. In order to best compare the MapReduce adaptation of the algorithm to its performance as a stand-alone ImageJ plug-in, author wrote a shell script which started an ImageJ macro to sequentially process all the parts of the original image in a

m2.xlarge instance. Since the technical parameters of the instance were identical to the computing nodes', this gives us a good estimate of how much the Hadoop framework affected the speed of the computations. As can be seen from the chart in figure 5, the decrease in speed is noticeable, but small.

Considering that Hadoop also provides fault-tolerance, load balancing and handles the distribution of data all by itself, it can be argued that this sort of approach to image processing has justified itself, and could reliably used as a solution for similar problems.

The results of comparing performance between Hadoop version 0.20.2 and 1.0.3 can be seen in figure 5.

Number of nodes	Wall time, 0.20.2 (s.)	Wall time, 1.0.3 (s.)
8	3031.2	3286.9
16	1557.2	1693.4

Fig. 5. comparing performance between Hadoop version 0.20.2 and 1.0.3

Comparison in processing time between a cluster running on Hadoop version 0.20.2 and 1.0.3. In the latter case, the result is an average of five test runs.

VII. Conclussions

While big data has already been used successfully in consumer markets, challenges remain to its implementation in Medical Image Processing. The most important challenge in shift to big data advances is the measureless amount of data in existing systems don't relate with each other and also the data existing in different file formats. The following challenge for image data in the healthcare is to maintain the privacy of the patient while storing and sharing the information interconnected without proper connectivity. It is a burdensome task for institutions that develop applications involving Big Data and Hadoop accordance with acts amended by the National Indian Health Board.

Overcoming these realistic challenges involve the government and its policies, doctors, medical professionals and importantly the technical developers of applications using technology. It is sure that bringing the effectiveness of Big Data Analytics and Hadoop's Map Reduce into the Medical Image Processing will surely increase the efficiency of the algorithms to yield the accurate results in a better way.

Acknowledgment

My thanks to the expert who has contributed towards development of the template. I would like to thank Dr. Pethuru Raj for providing various online resources. The immense experience of Dr. Pethuru Raj in developing and implementing Cloud applications, Big Data and Hadoop applications is an eye-opener for me to see how to reap the benefits of Big Data Analytics and Hadoop for the Healthcare especially with the image processing.

References

[1] C.S. Lindquist, T.S.C. Lindquist, and T.V. Lindquist, "New image processing algorithms requiring almost no a priori design information," Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on , vol.2, pp. 1-4, November 1998.

[2] D. Peter Augustine, "Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services", International Journal of Computer Applications, vol. 89, no. 16, pp. 36-40, March 2014.

[3] "Big data: 5 major advantages of Hadoop." [Online]. Available: http://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/ [Mar. 17, 2014]

[4] [Online]. Available: http://www.sabanciuniv.edu/en [Mar. 17, 2014]

[5] In Kyu Park; Singhal, N.; Man Hee Lee; Sungdae Cho; Kim, C.W., "Design and Performance Evaluation of Image Processing Algorithms on GPUs," Parallel and Distributed Systems, IEEE Transactions on, vol. 22, pp. 91,104, January 2011.

[6] Welch, E.; Patru, D.; Saber, E.; Bengtson, K., "A study of the use of SIMD instructions for two image processing algorithms," Image Processing Workshop (WNYIPW), 2012 Western New York , vol. , pp. 21, 24, November 2012.

[7] Okuhata, H.; Imai, R. Ise, M.; Omaki, R.Y.; Nakamura, H.; Hara, S.; Shirakawa, I., "Implementation of dynamic-range enhancement and super-resolution algorithms for medical image processing," Consumer Electronics (ICCE), 2014 IEEE International Conference on , vol. , pp. 10-13 January. 2014

[8] U. Catalyurek, S. Hastings, K. Huang, V. S. Kumar, T. Kurc, S. Langella, S. Narayanan, S. Oster, T. Pan, B. Rutt, X. Zhang, and J. Saltz. Supporting large scale medical and scientific datasets. In Parallel Computing: Current & Future Issues of High–End Computing, Proceedings of the International Conference ParCo 2005, pages 3–14, 2005.

[9] B. White, T. Yeh, J. Lin, and L. Davis. Web–scale computer vision using MapReduce for multimedia data mining. In Proceedings of the Tenth International Workshop on Multimedia Data Mining, July 2010

[10] Xu Zhengqiao; Zhao Dewei, "Research on Clustering Algorithm for Massive Data Based on Hadoop Platform," Computer Science & Service System (CSSS), 2012 International Conference on , vol., no., pp.43,45, 11-13 Aug. 2012

[11] Hao, Chen; Ying, Qiao, "Research of Cloud Computing Based on the Hadoop Platform," Computational and Information Sciences (ICCIS), 2011 International Conference on , vol., no., pp.181,184, 21-23 Oct. 2011

[12] "Hadoop Study Reveals Usage Stats, Benefits, and Challenges." http://architects.dzone.com/ [Feb. 9, 2014]

[13] http://hadoop.apache.org/. 2013

[14] White, T. Hadoop: the Definitive Guide (2nd Edition) [M]. O'Reilly Media, 2010.

[15] Markonis, Dimitrios, Roger Schaer, Ivan Eggel, Henning Muller, and Adrien Depeursinge. "Using MapReduce for Large-Scale Medical Image Analysis", 2012 IEEE Second International Conference on Healthcare Informatics Imaging and Systems Biology, 2012.

[16] Weiyi Shang. Zhen Ming Jiang. Hemmati, H. Adams, B. Hassan, A.E. Martin, P. 2013. Assisting developers of Big Data Analytics Applications when deploying on Hadoop clouds, Software Engineering (ICSE), 35th International Conference. hipi.cs.virginia.edu/ 2014/ 2014

[17] Mukherjee, A. Datta, J. Jorapur, R. Singhvi, R. Haloi, S. Akram, W. 2012. Shared disk big data analytics with Apache Hadoop, High Performance Computing (HiPC), 19th International Conference.

[18] Karl Potisepp, "Large-scale Image Processing Using MapReduce", Thesis, Tartu University, 2013.