Survey on analyzing and processing of EHR Medical data using Matlab and Hadoop

[1]KAVITA GARG, [2]PRAMILA JOSHI

[1]M.Tech Scholar & Project Lead in IT firm Aviva LIC, [2]Assistant Professor
Birla Institute of Technology Extension Centre Noida
A-7, Sector – 1, Noida - 201301, Uttar Pradesh, India

***Abstract*— In today's world data is evolving at a very high pace which we term as Big Data. Big Data can be described as a large volume of data which can be in the form of structured and unstructured data. It is very difficult to analyze and process this huge data using the existing traditional databases and the software technologies. The medical image data produced by the hospitals increasing exponentially at a higher rate. Due to the new imaging techniques and instruments available, the growth has been accelerated. To analyze the required information from such a vast amount of image data is challenging. To fetch the important information from these images in the form of digital data requires the new innovative soft wares. This survey paper explores the processing of medical data images using MATLAB tool and then processing the output of the tool using Hadoop. MATLAB has a MapReduce functionality which can be used to analyze the big data set. It is fully compatible with Hadoop MapReduce as a result of which MapReduce based algorithms in MATLAB can be run within the Hadoop MapReduce Framework.**

***IndexTerms*— BIG Data, Medical Image, MATLAB, Hadoop, Map Reduce, HDFS, EHR Data, Cloudera**

1. INTRODUCTION: Big Data

Big data is term which implies large data sets which are large in size and complex. The size of Big data can range from few dozen terabytes to many petabytes of data all from different sources like Web, Sales, Customer Contact Center, Social Media, Mobile Data , application logs and so on .

Analysis of the Big data has the potential to lead to better decisions and strategic business moves to help Organizations to improve operations and make faster, more intelligent decisions.

Due to the large size and the complexity of the data it becomes very difficult to perform effective analysis using the traditional data processing applications . Big data can be characterized by the 5V's which are volume, velocity, variety, veracity and value which put forward many challenges.

In dealing with Big date the fundamental issue which we came across are storage issue , management issue and processing issue . Each of these areas have their own technical problems to handle it .

1.1BIG DATA CHARACTERISTICS

1.Volume
It measures the amount of data generated every second . Data can be collected from the emails, twitter messages , photos, medical images and videos that are produced and shared every second.
Storage issue arises due to the large volume of data and as well as there will be issues in analyzing this massive amount of data .

2.Velocity
Velocity refers to how fast the new data is generated and also the speed at which the data is being processed to meet the demand and challenges. It can be described in the terms of the speed of data creation, streaming, and aggregation. Big Data technologies allows us to analyze the data without ever putting it into databases . [9]

3.Variety
Variety refers to the different type of data available . In today's world data is unstructured and it can be represented in various formats like text, images video, audio, etc. Due to the large volumes of data , it becomes the biggest obstacle to analyze the data . And the challenges ahead as per the analytic perspective is due to the incompatible data formats, non-structured data and inconsistent data semantics . [9]

4.Value

Value measures our ability to turn the Big data into value which helps in making decisions. It has been noted that the access to big data is no good unless we can turn it into value . Analytic science encompasses the predictive power of Big data . [9]

5.Veracity

Veracity measures the trustworthiness of the data. The degree of interconnectedness and interdependence in big data structures is so large that a small change in one or a few elements can substantially affect the behaviour of the system . [9]

2. EHR Data

An electronic health record (EHR) are digital records of health information . EHRs are real time, patient-centered records instantly accessible to authorized providers across practices and health organizations. An EHR can be shared across the information network and exchanges with all clinicians and organizations involved in a patient's care like laboratories, pharmacies, specialists, medical imaging facilities, emergency facilities, school and workplace clinics. [2]

An EHR includes following types of data :

1. Personal statistics like age and weight.

2. Information about visits to health care professionals.

3. Allergies.

4. Insurance information.

5. Family history.

6. Immunization dates.

7. A list of medications.

8. Treatment plans

9. Laboratory and test results

EHS systems store the data accurately and capture the state of a patient across time. There is no need of previous paper medical records as the data is stored is available in digital format which ensures the accuracy and legibility of the information . It can decrease paperwork as there is only one modifiable file which likely to be updated for any transaction . The benefit of EHS is that the patient health history is available as digital information and can be searched in a single file, EMR's are more effective as it helps in diagnose patients, reduce medical errors and provide safer care . [2]

3. GENERAL OVERVIEW OF MATLAB

MATLAB stands for MATrix LABoratory and it is a powerful tool which is utilized for versatile operations which includes numerical computation, visualization, aerospace research, virtual modelling, finance, graphical user interface & much more. MATLAB has some tool boxes useful for signal processing, image processing, optimization, etc. Matlab is a programming as well as a scripting language and with the help of predefined matlab commands one can easily write the program.[12]

MATLAB features a of family of application-specific solutions called toolboxes. Set of specific functions which are available in the MATLAB toolbox functionality allow to learn and apply specialized technology. MATLAB functions available in toolboxes are stored as M-files that provide more specialized functionality to solve particular classes of problems. MATLAB has tool boxes for signal processing, image processing, optimization, control systems, simulation, neural networks, fuzzy logic, wavelets, and many others. [13]

Example: Excel link functionality of Toolbox allows data to be written in a format recognized by Excel and similarly Statistics Toolbox helps in more specialized statistical manipulation of data (Anova, Basic Fits, etc).

3.1 MATLAB COMPONENTS

Following are the main components of the MATLAB system :

1. The MATLAB language.

MATLAB is a high-level matrix/array language. It has control structures , functions, data structures, input/output, and features of object-oriented programming . It allows rapid creation of throw-away programs as well as create complete large and complex application programs.[13]

2. The MATLAB working environment.

This is the set of tools and facilities to work with as the MATLAB user or programmer. It includes facilities to manage variables and support export and import data across applications. It also includes tools to develop and manage MATLAB files . Profiling M-files and debugging of MATLAB applications are more flexible with MATLAB.[13]

3. Handle Graphics.

This is a subsystem of MATLAB that handle graphics. It has high-level commands for two-dimensional and three-dimensional data visualization. Handle Graphics helps in generating Image processing, animation and presentation graphics. It has low-level commands which allow customizing the appearance of graphics. It allows to build customized Graphical User Interfaces.[13]

4. The MATLAB mathematical function library.

This is a collection of computational algorithms. In this elementary computational functions like sum, sine, cosine and complex arithmetic. And more sophisticated functions like matrix inversion, matrix eigenvalues and Bessel functions. It includes transformation functions like fast Fourier transformation Functions.[13]

5. The MATLAB Application Program Interface (API).

It is a library that allows the user to write C and Fortran programs that interact with MATLAB environment. It include facilities for calling routines from MATLAB using dynamic linking. MATLAB can be called as a computational engine and also used for reading and writing MAT-files.[13]

3.2 IMAGE PROCESSING USING MATLAB AND HADOOP

MATLAB has numerous functionalities for exploring and analyzing big data sets. It provides MapReduce programming technique which is powerful and applies filtering, statistics and other general analysis method to analyze Big data.

MapReduce funcionality analyze data that does not fit into memory. By using MapReduce based algorithms provided by Parallel Computing Toolbox the processing resources of the desktop can be utilized without changing the algorithms .

MATLAB MapReduce is optimized for array-based analysis. These can be executed within the Hadoop MapReduce Framework as it is fully compatible with Hadoop MapReduce.

In this paper, the vast techniques available in MATLAB will be studied to extract features from the images . First need to extract features from images and save it in mat format (matrix format) in Matlab. Then convert the mat file into csv or text format. After that analysis may be done using hadoop .

4. HADOOP

4.1 INTRODUCTION

Hadoop is an open source project which allows for the distributed processing of large data sets across clusters of commodity hardware. It is designed in such a way so as to scale up from single servers to thousands of machines each offering storage and local computation . As it is a open source project , a number of vendors have developed their own distributions either by adding new functionality or improving the code base like Cloudera, MapR and Hortonworks.

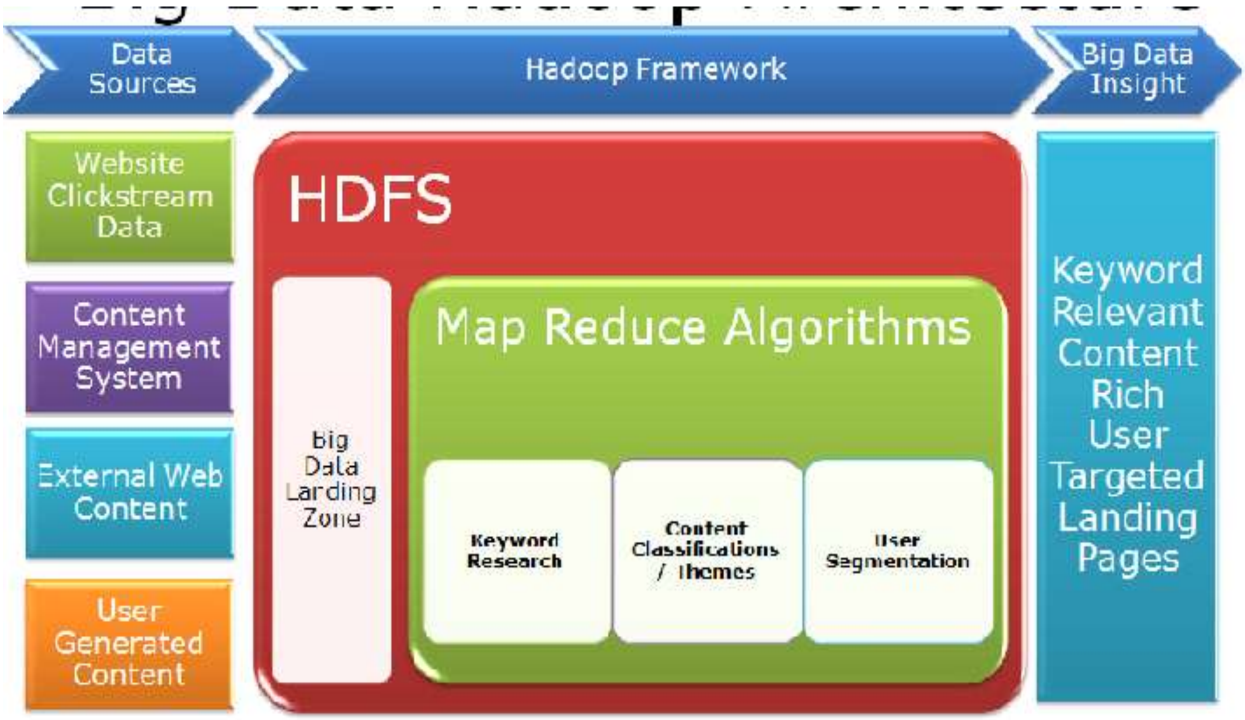

Figure 1

There are following four main modules in the standard open source Hadoop distribution (Apache Hadoop) :

- **Hadoop Common**

This contains Java libraries and utilities needed by other Hadoop modules. The Filesystem and OS level abstractions available in the libraries contains the necessary java files and scripts which are required to start the Hadoop .[6]

- **Hadoop Distributed File System (HDFS)**

A distributed file-system which is used to store the data on commodity machines and provides very high aggregate bandwidth across the cluster as well as high-throughput access to application data . [6]

- **Hadoop YARN**

This is a framework for resource-management platform responsible for managing computing resources in clusters and for job scheduling of users' applications. [6]

- **Hadoop MapReduce**

This is YARN-based programming model for large scale data processing .[6]

4.2 Architecture of Hadoop

The runtime environment of Hadoop consists of mainly five building blocks as shown in the below figure:

Cluster

It is a set of host machines called as nodes which can be partitioned in racks. This is the hardware part of the Hadoop Infrastructure .

YARN Infrastructure

It is a Resource Negotiator and this framework is responsible to allocate the computational resources like CPUs, memory etc., which are required for the application execution .

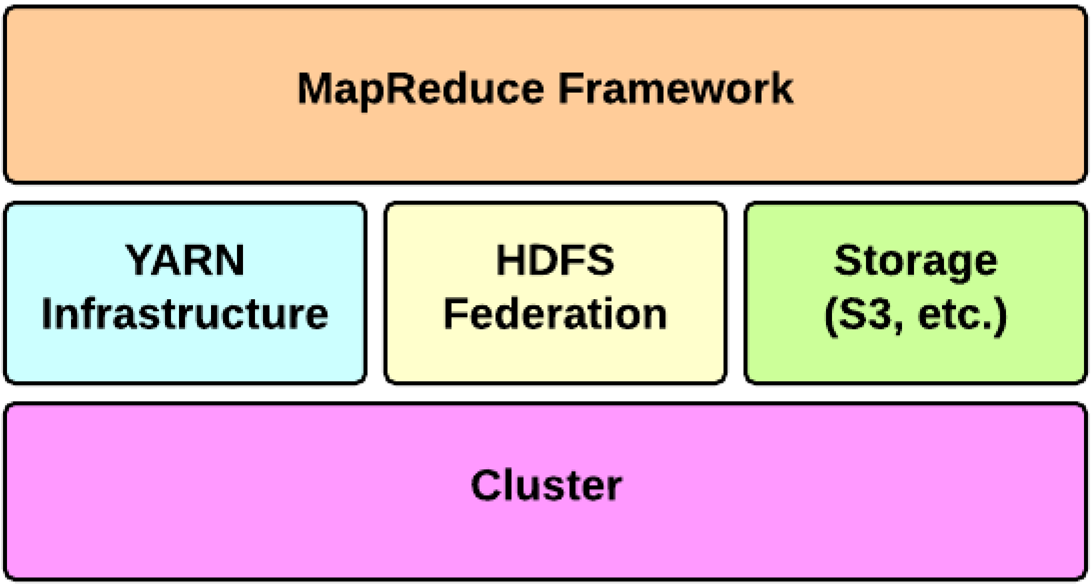

Figure 2

HDFS Federation

This is the framework responsible to provide permanent, reliable and distributed storage. It is used to store the inputs and outputs but not the intermediate data .

Storage

This provides alternative storage solutions .

MapReduce Framework

This is the software layer implementing the MapReduce paradigm .

4.3 Hadoop Distributions

Hadoop Vendor distributions are, of course, designed to overcome the drawbacks and issues with the open source edition of Hadoop and provide additional value to customers. Added functionalities focus on following things :

1.Reliability

Hadoop vendors react faster whenever a bug is detected. They promptly deliver fixes and patches with the intent to make commercial solutions more stable.[8]

2. Support

A variety of companies provide technical guidance and assistance, which makes it easy for customers to adopt the Hadoop for mission-critical and enterprise-grade tasks.[8]

3. Completeness

Very often Hadoop vendors couple their distributions with add-on tools which helps customers customize the Hadoop application to address specific tasks. [8]

Three of the top Hadoop distributions available are following :

- Cloudera
- MapR
- HortonWorks

4.4 COMPONENTS OF HADOOP

1. Pig

It is a data analysis platform that uses a high level data flow language Pig Latin that is optimized, extensible and produces sequences of Map-Reduce programs. Its structure is open to considerable parallelization which makes it easy for handling large data sets and analyzing huge data sets efficiently and easily. Operating System: OS Independent.[1]

2. Hive

It is developed by Facebook and is a data warehouse which promises easy data summarization, ad-hoc queries and other analysis of big data. It provides a simple language known as HiveQL which is similar to SQL for querying, data summarization and analysis .Hive uses indexing which makes query response faster. Operating System: OS Independent.[1].

3. HBase

It is a non-relational column-oriented database. It uses HDFS for underlying storage of data. It supports linear and modular scalability, random reads and also batch computations using MapReduce. Operating System: OS Independent.[1].

4. Zookeeper

Zookeeper earlier known as Hadoop sub-project is "a centralized service which provides simple, fast, reliable and ordered operational services for a Hadoop cluster . It is responsible for distributed synchronization, configuration information and providing a naming registry for distributed systems .Operating System: Linux, Windows (development only), OS X (development only) [1].

5. Ambari

It is RESTful API which provides easy to use web interface for Hadoop management, configuration and testing of Hadoop services and components . It provides step-by-step wizard for installation of Hadoop ecosystem services.[1].

6. Flume

This component is used to gather and aggregate large amounts of data . Apache Flume collects data from its origin and transfers it back to the resting location i.e. HDFS. It is Java-based, robust and fault-tolerant. Operating System: Windows, Linux, OS X[1].

7. Sqoop

This component is used for importing data from external sources into related Hadoop components like HDFS, HBase or Hive . Sqoop allows data imports, parallelized data transfer and mitigates excessive loads . It provides efficient data analysis and copies the data quickly . Operating System: OS Independent[1].

8. Oozie

This is a workflow scheduler designed to coordinate the scheduling of Hadoop jobs and the workflows are expressed as Directed Acyclic Graphs. Oozie runs in a Tomcat which is a Java Servlet container and use database to store the states and variables of all the running workflow instances along with the workflow definitions . It trigger jobs based on data and time dependencies. Operating System: Linux, OS X.[1].

4.5.ADVANTAGES OF HADOOP

1. Fast

Hadoop uses a unique storage methods which is based on a distributed file system which basically implements a mapping system to locate data in a cluster. It has the ability to store and process huge amounts of any kind of data faster as the tools for the data processing are often on the same servers where data is located .[7]

2. Computing Power

Hadoop's processes big data fast due to the distributed computing model which divides tasks in a manner that allows execution of tasks in parallel. As the computing nodes are increased , the processing power increases .[7]

3. Fault tolerance

In a Hadoop System if data is sent to an individual node, that data is replicated to other nodes in the cluster to make sure that in the event of failure the distributed computing does not fail . And so, the data and application processing are protected against in case of hardware failure.[7]

4. Flexibility

Hadoop enables businesses to easily access various new sources of data and operate on different types of data whether structured or unstructured . Hadoop can derive valuable business insights from various data sources such as social media, email conversations or clickstream data . Hadoop has a usage in a wide variety of purposes such as recommendation systems, log processing, marketing analysis , data warehousing and fraud detection.[7]

5. Low cost

The scale-out architecture of Hadoop with MapReduce programming can affordably store all of the organization's data for later use . The cost savings are massive as Hadoop the open-source framework is free and offers computing and storage capabilities for hundreds of pounds per terabyte .[7]

6. Scalability

Hadoop is a platform which is highly scalable because of its ability to store as well as distribute large data sets across plenty of servers that operate in parallel . And by addition of each new node the processing power also increases.[7]

5. CLOUDERA

CDH which stands for Cloudera Distribution including Apache Hadoop, is Cloudera's open source platform distribution . It is built specifically to meet enterprise demands . Cloudera provide services to enterprises to store, process, and analyze the organization's data and empowering them to extend the value of existing investments and providing them new fundamental ways to derive value from the massive data.

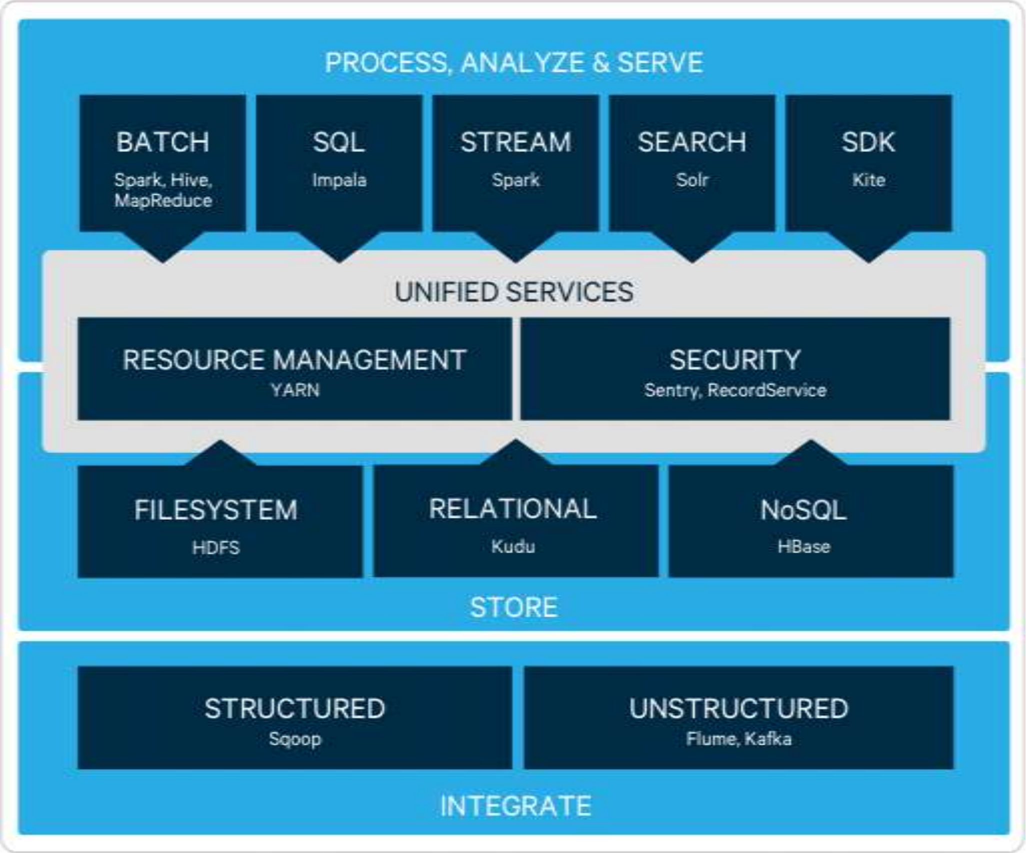

Figure 3

Cloudera was first founded in 2008 and is currently, the leading provider and supporter of Apache Hadoop for the enterprise sector. It offers differentiated software, support, training, professional services for business critical data challenges which includes storage, access, management, analysis, security, and search.

6. CHALLENGES

1. First challenge is in the storage of the massive data . The analysis of such a huge data with the available limited computational facility may become a barrier .

2.Hadoop is an open source framework and so advantages and limitations of open source platforms apply. [15]

3.In healthcare systems the data is Real-time and massive . There needs to be a requirement to address the issue of lag between data collection and processing. [15]

4.Health care data is rarely standardized and generated in legacy IT systems as structured or unstructured formats. The issue needs to be addressed to process different type of formats .15]

7. CONCLUSION AND FUTURE SCOPE

Due to the various new technologies the medical data is increaing at a huge speed in the healthcare industry. So, in the future we will see the rapid and widespread implementation and use of in the field of big data analytics.

The challenges needs to be addressed to guarantee privacy of the health records of the patients . So there will be need to improve the tools and technologies for the efficient storage systems .

In this paper , we explored the way to analyze and process the medical image using MATLAB and Haoop Framework . There is a scope for research to find out the way to search on tools which can be used for massive storage of data. And to check the tools which can process the real time medical images faster.

REFERENCES:

[1] Pramila Joshi,"Analyzing Big Data Tools and Deployment Platforms ", International Journal of Multidisciplinary Approach and Studies, ISSN NO:: 2348 – 537X, Volume 02, No.2, March – April, 2015.
[2] https://en.wikipedia.org/wiki/Electronic_health_record
[3] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N , " Analysis of Bidgata using Apache Hadoop and Map Reduce ", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 4, Issue 5, May 2014.
[4] Lalit Malik, Sunita Sangwan " MapReduce Framework Implementation on the Prescriptive Analytics of Health Industry ", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, IJCSMC, Vol. 4, Issue. 6, June 2015, pg.675 – 688
[5] Kenli Li, Wei Ai, Zhuo Tang, Fan Zhang, Lingang Jiang, Keqin Li, and Kai Hwang, Fellow, IEEE, " Hadoop Recognition of Biomedical Named Entity Using Conditional Random Fields", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26, NO. 11, NOVEMBER 2015.
[6] https://en.wikipedia.org/wiki/Apache_Hadoop
[7] http://www.sas.com/en_us/insights/big-data/hadoop.html
[8] http://www.networkworld.com/article/2369327/HYPERLINK "http://www.networkworld.com/article/2369327/software/comparing-the-top-hadoop-distributions.html"software/comparing-the-top-hadoopHYPERLINK "http://www.networkworld.com/article/2369327/software/comparing-the-top-hadoop-distributions.html"-distributions.html
[9] http://www-01 .ibm.com/software/data/bigdata
[10] Pramila Joshi, "Cloud Architecture for Big Data ", International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 4 Issue 6 June 2015
[11] Peter Bajcsy, Antoine Vandecreme, Julien Amelot, Phuong Nguyen, Joe Chalfoun, Mary Brady, ・Terabyte-sized Image Computations on Hadoop Cluster Platforms・2013 IEEE International Conference on Big Data.
[12] http://www.uniqtechnologies.co.in/matlab%20projects%20ieee-2013.html
[13] http://cimss.ssec.wisc.edu/wxwise/class/aos340/spr00/whatismatlab.htm
[14] K. AshaRani and S.Subbalakshmi, "Image Recognition System in Hadoop", International Journal of Advance Research in Computer Science and Management Studies, ISSN: 2321-7782 (Online), Volume 3, Issue 6, June 2015
[15] Dimitrios Markonis, Roger Schaer, Ivan Eggel, Henning M¨uller, Adrien Depeursinge, "Using MapReduce for Large–scale Medical Image Analysis", University of Applied Sciences Western Switzerland (HES–SO), Business Information Systems, Sierre, Switzerland
[16] Peter Bajcsy, Antoine Vandecreme, Julien Amelot, Phuong Nguyen, Joe Chalfoun, Mary Brady, "Terabyte-sized Image Computations on Hadoop Cluster Platforms ",2013 IEEE International Conference on Big Data.
[17] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu, Member, IEEE, "i2MapReduce: Incremental MapReduce for Mining Evolving Big Data ",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015.
[18] D. Peter Augustine, " Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services " , International Journal of Computer Applications (0975 – 8887) Volume 89 – No 16, March 2014.
[19] Zhigao Zheng, Ping Wang and Jing Liu and Shengli Sun, "Real-Time Big Data Processing Framework: Challenges and Solutions ", Applied Mathematics & Information Sciences, An International Journal, Appl. Math. Inf. Sci. 9, No. 6, 3169-3190 (2015).

[20] Seyyed Mojtaba Banaei, Hossein Kardan Moghaddam, " Hadoop and Its Role in Modern Image Processing ",Open Journal of Marine Science, 2014, 4, 239-245
Published Online October 2014 in SciRes. http://www.scirp.org/journal/ojms
http://dx.doi.org/10.4236/ojms.2014.44022.
[21] Suresh Lakavath, Ramlal Naik L, "A Big Data Hadoop Architecture for Online Analysis ",IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol. 4, No.6, December 2014.
[22] Sarade Shrikant D., Ghule Nilkanth B., Disale Swapnil P., Sasane Sandip R., " LARGE SCALE SATELLITE IMAGE PROCESSING USING HADOOP DISTRIBUTED SYSTEM", International Journal of Advanced Research in Computer Engineering & Technology(IJARCET), Volume 3 Issue 3, March 2014.
[23] Sreekanth R, Dr. Gondkar RR, " MapReduce Program to Efficiently Analyse Big Data Electronic Health Records Database using Hadoop Cluster on Amazon
Elastic Compute Cloud ", International Journal of Advanced Research in
Computer Science and Software Engineering, ISSN: 2277 128X, Volume 5, Issue 8, August 2015.