International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Cloud Enabled Predictive Big data Analytics Framework for Healthcare

Madhuri A Dalal , D.R Ingle

Abstract: ***Big data is one of the recently emerged domain in today's digital age where everything is being digitized. Very large size of data is produced from various organizations all through the globe .This very large size of data is termed as big data..Conventional databases are not able to handle the challenges with enormous data. Detailed examination of large amounts of data requires a lot of efforts at various levels with the aim of finding knowledge , hidden patterns for decision making. Big data analytics plays an essential character in order to attain predictive analytics in healthcare. Cloud systems can be employed for the storage of big data so that users can be able to access it from anywhere . The objective of this paper is to review the concept of big data , healthcare in the context of big data and cloud computing. It also proposes architecture framework model for predictive big data analytics in healthcare area. It also presents technology for big data analytics. This paper also addresses the use of big data analytics along with specifying healthcare applications.***

Keywords : Big data , Analytics, Healthcare, Cloud Computing.

I. INTRODUCTION

Big data exploded on the scene in the first decade of the 21st century. Firms like Google ,LinkedIn, eBay and Facebook were built around big data from the beginning. We all are surrounded by massive amount of data. There are following sources which contributes to the creation of big data: Social media data namely Facebook , Twitter, online shopping, online advertising, Stock Exchange Data ,Search Engine Data, machine generated data etc. generates huge amount of data[1].

As a result , the machines have to generate and keep huge data too. Due to this exponential growth of data , analysis of that data becomes challenging and difficult.

Healthcare industry is extremely huge and there is lot of medical data and it also generates a large volume of data . For example, for genomics data, each human genome requires 200GB of raw data or 125MB if we store just snipes, a single functional MRI i.e. medical resonance imaging is about 300GB in case of medical imaging data. monitoring, temperature, heart rate ,medication dispensing measure at intensive care unit [2].

Healthcare also generates lots of various information such as clinical information including patients demographics ,diagnosis procedure, medication, lab results , medical records of a patient and patient generated health data includes body sensors and other equipments that patients wear and live data sources such as blood pressure measure, blood glucose measure at intensive care unit [2].

In recent years, retrieval based upon the history of data available has become a growing research area even in the medical field. Disease retrieval/prediction based on the patients data history has aided medical practitioners in discovering abnormalities in early stages. Medical diagnosis from the vast availability of dataset remains to be a major challenge for the medical practitioner. This can affect the life of the victim to a greater extent. Hence a solution has to be formulated to measure the medical problems in initial stage of affection itself so that preventive actions can be taken in advance to reduce the severity.

Various approaches were formulated recently for retrieving the disease based upon the historical data available. With the advancement of cloud computing, large amounts of data are stored in cloud which can be effectively retrieved based upon the required queries.

The rest of the paper is organized as follows: Section I contains introduction of big data analytics and cloud computing. Section II illustrates background technology big data analytics. Section III presents proposed architecture framework for predictive big data analytics in healthcare, Section IV explores various healthcare applications. Finally, section V concludes the paper.

A. Big Data Analytics :

Different descriptions of Big data has been recommended by various researchers. Majority of the researchers specifies that Big data is being characterized by five V's (Volume, Variety, Veracity ,Velocity and value) [3]. Volume deals with the size of available datasets which typically require storage in a distributed manner and processing. Velocity deals with the speed with which data is being collected like social networks, mobile devices and Internet of Things (IoT) .Variety deals with different formats such as structured, unstructured as well as semi-structured data like text ,sound, audio, video[4] . Veracity includes the noise, biases, errors, missing data and irregularity in data[5]. Value deals with the method of finding huge unseen patterns from voluminous datasets with various of data types and speedy generation.

Manuscript published on 30 September 2019
* Correspondence Author
Madhuri A. Dalal*, Research Scholar, Bharati Vidyapeeth College of Engineering, C.B.D., Belpada, Navi Mumbai, India. (Email: madhuridalal1012@gmail.com)
Dr. D. R. Ingle, Professor, Bharati Vidyapeeth College of Engineering, C.B.D., Belpada, Navi Mumbai, India . (Email: dringleus@gmail.com)

All of this data is said to be unstructured data and so cannot be stored in structured databases. Traditional data management systems and existing tools are facing difficulties to process such a big data[6].

Big Data analytics is an advanced research topic with the availability of enormous with vast amount of data storage and computing capabilities provided by improved and scalable computing infrastructures.

Big data analytics is being divided into following 3 categories:

1. Predictive Analytics -: It uses existing data and trends to predict what might happen in the future.

2. Descriptive Analytics -: It describes what has happened in the past. It alludes to summarization technique of big data.

3c. Prescriptive Analytics -: It makes usage of optimization and simulation algorithms to get recommendation on possible outcomes and answers. It refers to what is to be done.

B. Cloud Computing :

Cloud computing is the usage of distinct privileges, like software development platforms, servers, storage and software, over the world wide web also called internet, often referred to as the "cloud."

As per National Institute of Standards and Technology (NIST) , "It is a representative model for empowering universal, convenient, when ever required network access to a shared pool of configurable computing resources such as storage, applications, networks, servers and other utilities that can be provided very quickly and freed with very little management work or service provider interaction"[7].

2.1 Features of Cloud Computing:

Cloud computing exhibits following five essential features:

1) On-demand self-service -: A user can acquire cloud computing benefits which makes usage of runtime data without need for interaction of human with a service provider.

2) Broad network access -: The user can acquire the data of the cloud or upload the data to the cloud from everywhere only by making use of a device and an internet connection.

3) Resource pooling -: As per users command, several user's request can be serviced from the same physical resources by securely isolating the resources on logical level.

4) Rapid elasticity -: The resources in cloud can be resiliently distributed ,assigned as well as made free. For example, if a web application gets an exceptional traffic, user can design more server in order to support this service.

Thus, the application can elegantly and automatically scale with demand.

5) Measured service -: Cloud systems automatically manage and optimize resource usage by taking advantage of a metering capability. Resource usage can be tracked, handled ,supervised, and noted providing visibility for both the service provider and user of the utilized services.

6) Location independence -: Majority of the cloud services are location free where user can acquire accessibility to their devices, applications or facilities across the world through internet connection despite of their geographical position.

7) Pay as you go -: In cloud computing , the user has to make payment for the facilities ,services or the storage space they have been used. There is no invisible or extra amount which is to be charged. The service is efficient as well as inexpensive and usually , many times some storage space is allotted for free.

2.2 Cloud Computing Service Models:

Cloud computing services can be categorized as infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS)[8].

- ***Infrastructure as a service (IaaS)*** **:** It is a cloud computing service in which service provider makes distribution of computer infrastructure like servers , storage, hardware, network components to the enterprise with the aim to support enterprise operations as per need. Within a service provider's infrastructure ,Enterprises can make use of their own platforms and applications. It authorises enterprises the costs of purchasing and maintaining their own hardware[8].
- ***Platform as a service (PaaS) :*** It is a cloud computing service in which service provider makes distribution of hardware and software tools ,especially tools which are essential for the design of application are allocated to users over the world wide web called internet . It is comprised of easily accessible runtime execution environment, development and deployment tools .The most common example is Google App Engine.
- ***Software as a service (SaaS) :*** It is a cloud computing service in which end users are assigned with computer applications over the world wide web by third party service provider. While making usage of this service, on their local computer or laptop ,users do not install applications on their own ,instead of that the applications reside on a remote cloud network .These applications can be accessed through the web or an application programming interface, API. The amenities seen on an application layer which is an extension of ASP (Application service provider), wherein application is kept in existence, run and confirmed by a service vendor. The most common examples of SaaS model includes Gmail, Hot mail and online banking deployment .

2.3 Cloud Computing Deployment Models:

There are 4 deployment models of cloud computing :

- ***Public Cloud:*** It is developed for the general community people where resources, applications and facilities are granted by cloud service providers (CSPs) over the world wide web based on subscription or pay per use model.
- ***Private Cloud***: It is developed for single organizations which is not available publicly but operates within a firewall. It provides greater security. For example, Amazon operates a private cloud as part of its overall (public) Amazon Web Services known as Amazon Virtual Private Cloud.

- ***Hybrid Cloud:*** It is a composition of public and private cloud. This type of architecture demands together off-site server based cloud infrastructure as well as on-site resources [8].
- ***Community Cloud:*** The community cloud is a combination of different types of cloud like public, private or hybrid clouds, which is being agreed by many institution for the same purpose usually security [9].

II. BIG DATA ANALYTICS TECHNOLOGY AND TOOLS IN HEALTHCARE

For the purpose of handling big data set and execution of algorithm to process big data set , we need big data system for good performance. For efficient big data management, following big data frameworks/tools are developed: Hadoop , Spark, Flink,Samza. The different building blocks of Hadoop framework are explained in Figure 1.

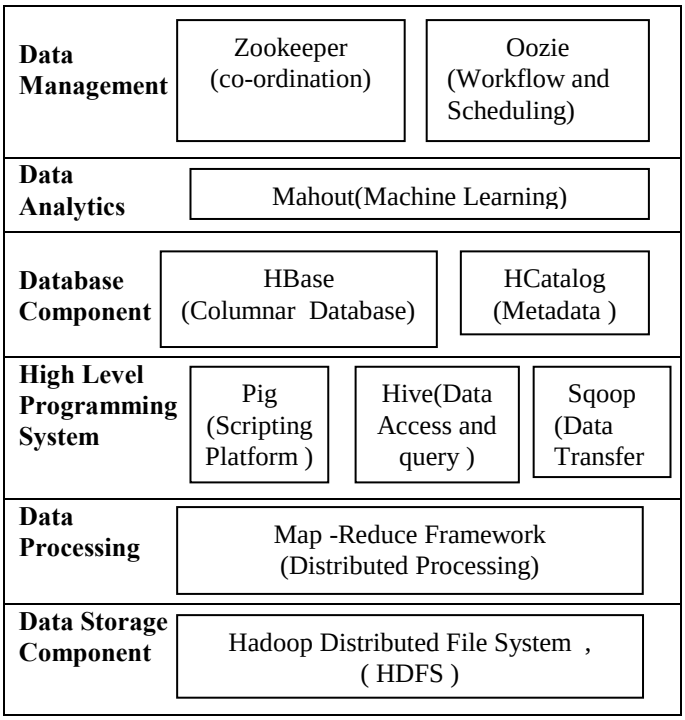

Figure 1 : Hadoop Framework

Apache Hadoop is an open source software framework for storage and processing of big data in a distributed fashion on large clusters of commodity hardware. Basically, it performs two tasks namely massive data storage and faster processing. Hadoop stack is divided into the following components.

A. Data Storage Component:

For data storage purpose, Hadoop manages a component called Hadoop Distributed File system, HDFS[10]. It is a distributed file system used for storage which supplies an interface for executive administration of the file system to enable it to extent and provide high throughput. It has two components namely Name node, known to be Master and Data node, known to be slave. When the loading of a file is being done into HDFS, the same file id being copied across multiple nodes for reliable data protection and divided into 64MB fragments (by default) known as blocks that are stored across the cluster nodes, which are being referred as data nodes. Name node is accountable for storage space and executive administration of meta-data.

B. Data Processing Component:

In Hadoop framework, Map-Reduce software framework is used for data processing. Hadoop Map-Reduce is a programming model and software for writing applications which can process vast amount of data in parallel on large clusters of computers. This framework is used to solve computational problems , analytics on large scale data which is the need of today's business. It works on divide and conquer principle. Map-Reduce uses two primitives namely mapper and reducer for data processing. The input to both the stages is key-value pair. In the map stage, a group of key-value pairs comprises the input and upon each key-value pair, the required operation is executed so as to produce a group of intermediary key-value pair. During Reduce stage, the output of map phase is accumulated[11].

C. High Level Programming System :

Pig refers to run-time environment which permits users to carry out Map-Reduce operations on a Hadoop cluster. Pig platform uses scripting language known as Pig Latin which is at high level. Pig works with two modes namely local mode and HDFS mode. It is a tool which can be used to analyze voluminous data representing them as data flow. It works on the top of Hadoop. It is also a high level data flow platform in order to execute Map-Reduce programs of Hadoop[12].

Hive is a tool for data warehouse framework in order to transform structured data which is stored in tables in Hadoop. The language used by Hive is HiveQL, which is SQL- type language basically used for the purpose of querying. It is not suitable for online transaction processing but designed for online analytical processing. Hive is ETL (Extract-transform-load) ,data warehouse tool, an abstraction over Hadoop[13]. Sqoop also called as SQL to Hadoop tool. This tool has the ability to transfer data in both ways between Hadoop and other data repositories of Hadoop like Hive or HBase or relational database servers. It is used to import data from relational databases such as Oracle , MySQL, to Hadoop HDFS, and export from Hadoop distributed file system to relational databases[14].

D. Database Component :

HBase is and non-relational database system in a Hadoop environment.It is, distributed, open source versioned, column-oriented store. It is a NoSQL database that runs along with Hadoop as a distributed and scalable big data store. It is based on columns rather than rows. HBase achieves high throughput and low latency by providing faster Read/Write access on voluminous data sets. Applications which require fast & random access to large amount of data HBase is the preferred NoSQL database[15].

HCatalogue serves as metadata and table storage management facility for HDFS. The main objective is to facilitate interaction of user with HDFS data and grant sharing of data between different tools and execution platform.

E. Data Analytics:

For data analytics purpose, Hadoop introduces component called Mahout. It is scalable machine-learning and data mining library. It ensures scalable and efficient implementation of large scale machine learning applications and algorithms over large data sets. Here scalability of machine learning algorithm refers to the given type of operations they perform, it can be executed as a group of parallel processes. The execution of algorithms in Mahout library can be carried out in a distributed fashion and also written for Map-Reduce. It provides analytical capabilities and optimized algorithms.

F. Data Management :

In Hadoop framework, Zookeeper and Oozie are used for data management. Zookeeper provides various services such as storage, maintenance of configuration information ,naming , synchronization in a distributed manner with master and slave nodes. It also permits communication of distributed processes with each other by means of name space of data registers which are hierarchically shared. This data registers also termed as znodes[[16].

Oozie[16] is a job coordinator and workflow manager for jobs executed in Hadoop . An Oozie workflow is a collection of Hadoop jobs and actions organized in a Directed Acyclic Graph.

III. PROPOSED ARCHITECTURE

The proposed architecture of predictive big data analytics for healthcare has five layers that combines big data analytics along with cloud computing as shown in Figure 2. It has five layers namely Data acquisition layer, Data storage layer, Data Transformation layer, Big data Predictive analytics layer and Visualization layer.

The first and bottom-most layer of proposed architecture is Data Acquisition layer. The main purpose of this layer is to gather data from several devices ,healthcare sensors , in several formats. When patients health is monitored continuously, various types of medical data is being generated. Medical data includes structured data like Electronic Health record (EHR), semi-structured data like data obtained from Body Sensor Networks (BSN) , Lab test report of patients, doctor's prescription, unstructured data like biomedical imagery.

The second and next layer is Data Storage layer. Data storage is the difficult task as it involves very large amount of data. The data collected from various devices is stored in data storage component. As conventional database management system are not suited for the storage of huge data, this component may make usage of cloud infrastructure, Hadoop Distributed File System (HDFS), No-SQL database for the purpose of storing data . The advantage of storing the data on cloud is that data from different sources can be stored at one place and doctors can access it from any place, anytime .The storage also requires to be done in a systematic way so that results can be obtained rapidly.

The third and next layer is Data Transformation layer. Processing of raw data without transformation may incur extra computational cost. Therefore it is suggested to transform data properly in order to obtain accurate results . Data transformation involves data cleaning, filtering operations, handling missing value, handling noisy data, data moving, data sorting ,data splitting. In the healthcare area, missing data should be handled with maximum accuracy as invalid decisions may have serious consequences. Many algorithms like Expectation- Maximization (EM) algorithm and multiple Imputation algorithm exist in the field of data mining in order to handle missing values.

The fourth and next layer is Big data Predictive analytics layer. The idea is to create a model which is capable of making predictions for new examination based on previous data. Big data analytics predictive component may make usage of data mining technique namely classification , clustering, machine learning technique or Hadoop map/reduce. During this phase, the relevant features are extracted and selected for building prediction model using feature extraction and selection technique. Here clustering can be used to make a group of similar patients together based on relevant features. Classification can be used to identify the class of a new patient depending on previously available data. Even though there are several data mining tasks like classification, clustering , association rules for information extraction, the goal is to create a predictive analytics model which make usage of machine learning technique with hadoop map/reduce for parallel processing of data as data size is very huge. The last and uppermost layer is Data Visualization layer. Data visualization refers to graphical representation of information and data. The purpose of this layer is to produce visualization report of patients health monitoring and predictive decision report. This layer will be accessible by patients , doctors and end-users. Patterns, trends and co-relations which are undetected in textual data can be easily recognized with data visualization

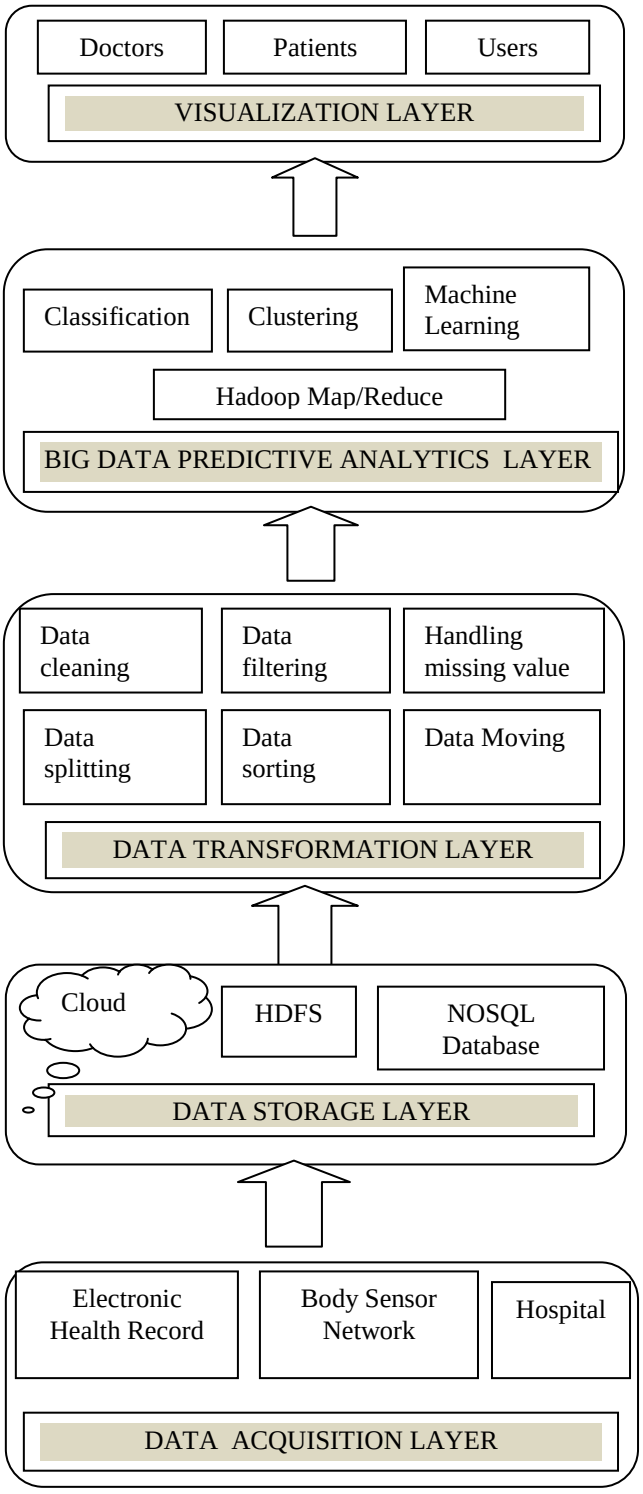

Figure 2: Proposed Architecture for Predictive Big Data Analytics in Healthcare

IV. HEALTHCARE APPLICATIONS

There are following application areas of big data in the arena of healthcare:-

A. Predictive Modeling :

It is the process of representing historical data to view the model for prediction of future outcomes or incidents.
The First challenge in using predictive modeling is that in the context of healthcare, there are many more sick persons in the hospital who comes for the treatment and we want to examine their diagnosis information, medication information and so on .
The Second challenge is that there are many more predictive models to be built. Predictive modeling is not only one algorithm but a chain of computational jos like cohort construction, feature construction, cross validation, feature selection , classification and so on . Every step in this pipeline has many different options. As a result, all of those combined gives us many pipelines to be evaluated and compared.

B. Computational Phenotyping :

It is the process of turning raw patient data (for example, messy electronic health record) into meaningful clinical concepts known as phenotypes. In order to extract phenotypes from raw data, user should deal with some waste products such as missing data, duplicates ,irrelevant data, redundant data.

C. Patient Similarity:

It is about simulating the doctor's case based with computer algorithm. Instead of depending on one doctors memory, it would be nice if we can leverage all patient data in entire database. So the idea is that when the patient comes in, doctor does some examination on patient ,then based on that information ,we can do a similarity search through the database and find those potentially similar patients . Then doctor can provide some supervision on that result to find those truly similar patient to the specific clinical context. Then we can group those patient based on what treatment they are taking and look at what outcome they are getting. Then recommend the treatment with the best outcome to the current patient .

V. CONCLUSION AND FUTURE SCOPE

In this paper, big data analytics along with cloud computing have been reviewed. Meanwhile , an overview of big data system namely Hadoop framework have been provided. We have proposed an architectural framework of predictive big data analytics in healthcare which includes five layers namely data acquisition, data storage, data transformation, big data predictive analytics and data visualization. We have also addressed healthcare applications namely predictive modelling , with specifying challenges ,computational phenotyping and patient similarity.

Future scope of this work would be the implementation of all the above layers for prediction of healthcare big data to build prediction model above map-reduce on hadoop framework.

ACKNOWLEDGMENT

We would like to thank management of Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai, India for providing the infrastructure to carry out the proposed research work.

REFERENCES

1. S. Thanekar , K. Subrahmanyam , A. Bagwan, "Big Data and mapReduceChallenges, Opportunities and Trends," *International Journal of Electrical and Computer Engineering* 6(6): p 2911-2919,2016.
2. W. Raghupathi, V. Raghupathi, "Big data analytics in healthcare : Promise and potential," *Journal of Health Information Science and Systems* 2014.
3. 3. A. Oguntimilehin and E. Ademola., "A review of big data management,benefits and challenges," *Journal of Emerging Trends in Computingand Information Science,* vol. 5, no. 6, p. 433– 438, 2014.
4. 4. D. Madhuri, "A Novel Approach for Processing Big Data," *International Journal of Database Management Systems (IJDMS)* ,vol. 8, no. 5, pp. 15-24, October 2016.
5. 5. W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, E. Nguifo, "An experimental survey on big data frameworks," Available: https://arxiv.org/pdf/1610.09962.pdf [Accessed: May 12,2019].
6. 6. C. Tsai, C. Lai, H. Chao and A. Vasilakos, "Big data analytics: a survey", *Journal of Big Data, Springer Open Journal* , pp. 1-32, 2015.
7. 7. Fang Liu et al. , "NIST Cloud Computing Reference Architecture," Recommendations of the National Institute of Standards and Technology, Special Publication pp 500-292, 2011.
8. 8. Sabia ,S. Kalra, "Applications of big Data : Current Status and Future Scope," *International Journal on Advanced Computer Theory and Engineering* , vol. 3, issue 5, 2014.
9. 9. B. Jadhav, A. Patankar, "Opportunities and Challenges in integrating Cloud Computing and Big Data Analytics to E-governance," *International Journal of Computer Applications(* 0975 – 8887) Vol. 180, No. 15, pp. 6-11, January 2018.
10. 10. A. Oussous , F. Benjelloun , A. Lahcen, S. Belfkihh, "Big Data technologies: A survey",*Journal of King Saud University – Computer and Information Sciences* , pp. 431-438.
11. 11. S. Kolte and J. Bakal, "Big Data Summarization: framework, Challenges And Possible Solutions**,** *Advanced Computational Intelligence: An International Journal (ACII)*, Vol.3, No.4, October 2016.
12. 12. C. Olston, B. Reed, U. Srivastava, R. Kumar and A. Tomkins, "A Pig Latin :a-not-so-foreign-language for data processing"(2008)SIGMOD International conference on Management of data, pp. 1099-1110, ACM.
13. 13. http:// www.tutorialspoint.com/hive/hive.pdf [Accessed: March 15,2019].
14. 14. http://www.tutorialspoint.com/sqoop/sqoop.pdf [Accessed : March 15,2019].
15. 15. V. Bobade, "Survey Paper on Big Data and Hadoop", *International Research Journal of Engineering and Technology(IRJET)*," Vol. 03 , Issue: 01, Jan-2016.
16. 16. A. Agrahari, D. Rao," A Review paper on Big Data: Technologies, Tools and Trends," *International Research Journal of Engineering and Technology (IRJET)*, Vol. 04, Issue 10 , Oct - 2017.
17. 17. N. Aboudi, L. Benhllima ,"Big Data Management for Healthcare Systems: Architecture ,Requirements and Implementation," Hindawi, Advances in BioInformatics, Vol. 2018, Article Id 4059018 Available: https://www.hindawi.com/journals/abi/2018/4059018/, [Acccessed : April 15, 2019].
18. 18. *J. Archenaa and E. Mary Anita ,"* A Survey of Big Data Analytics in Healthcare and Government**,"** *Elsevier, Procedia Computer Science* 50 , pp. 408 – 413 (2015)

AUTHORS PROFILE

Mrs. Madhuri Dalal is a Ph.d student at Bharati Vidyapeeth college of Engineering , University of Mumbai, India . She received his Graduate degree in Computer Technology from Nagpur University , India in 2000 and Post Graduate degree in Computer Engineering from University of Mumbai, India in 2011. She is a life member of ISTE since 2008. Currently, she is working in big data analytics. She has total 14 years of teaching experience. .

Dr. Dayanand R. Ingle pursued Bachelor of Computer Engineering from Walchand College of Engineering , Sangli in 1996 and M.Tech in Computer Engineering from Dr. Babasaheb Ambedkar Technological University , Lonere, India in 2003.He has pursued Ph.D in Computer Science and Engineering, in 2015 from Sant Gadge Baba Amravati University, India . He is working as Professor and Head Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, University of Mumbai, India .He is a member of CSI , ISTE and Associate member of ACM. He has published more than 75 research papers in reputed international journals and conferences. His main research work focuses on Web Information System, Web services, data mining, Big Data and Cyber Seurity. He has more than 23 years of teaching experience. He has guided more than 35 ME students and guiding 7 Ph.D students