



International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294)

Hadoop Technology: Simplifying the Health-care Systems

BasitMaqboolMattu

M-Tech CSE Student,. (4th semester).

Department of Computer Science & Engineering,

Ganpati Institute of Technology and Management, Bilaspur, Haryana, India

Abstract - The aim of this paper is to study about the Hadoop technology which can be used to simplify the Healthcare Systems. The Healthcare systems are nowadays becoming very complex and it is not easy for the traditional systems to manage all the issues like high cost, scalability and flexibility. So to make our healthcare systems easy to become and get less costly we will use the Hadoop technology. This technology has many applications in the Healthcare systems, thus making all the services easy to work. As the world is becoming explosive with huge amounts of the information and data also known as the Big Data, so to manage such a large amounts of data, we are changing our trend and using only the NOSQL databases nowadays as a very effective tool in managing the Big Data. In our Healthcare application we are dealing with the same issue and companies using Healthcare systems are not able to manage the growing data problem. So in this paper we will analyze how the Hadoop technology will help us to make effective and easy use of the Healthcare systems and how this technology finds its applications in the Healthcare Systems.

Keywords: *Hadoop, NOSQL, Big Data, MapReduce, HBase, Hive, HDFS(Hadoop Distributed File System).*

1. INTRODUCTION

Hadoop was developed by the Apache Software Foundation. Its initial release was in December 10, 2011 which is about six years ago. And its stable release was given on August 25, 2016. Now since it is very active in the market, the lot of work is done using this technology. This technology was written in java language. It gives the cross platform support among all the operating systems. Now we can also mention that its file system is HDFS (Hadoop Distributed File System). The license for this technology was Apache License 2.0. It is commonly known by the name as the Apache Hadoop. Now we can also

say that the Apache Hadoop is actually an open-source software framework. It is basically an open source software framework which can be used for the purpose of the distributed storage and also to manage the large data-sets. It is also used for the processing of these large data-sets using some good technique known as the MapReduce Programming model. This model is designed in a way that it contains many computer clusters which are built from the commodity hardware. Now this model is designed in such a way that all the hardware failures etc. can be managed very easily by the framework. So this framework is actually fault tolerant. The faults and hardware failures are easily handled by this framework. If we talk of the core design of the Apache Hadoop we can explain it in terms of its processing and the storage part. There are two main parts of the core of the Apache Hadoop. These two parts are known as the HDFS called as Hadoop Distributed File System and the processing part which is known by the term MapReduce programming model. Hadoop works by splitting the files into some blocks which are then distributed or allocated to all the nodes in a cluster which then process the data in the parallel fashion. That means all the nodes can access the data by their own without assistance from others. This is main feature of the Hadoop model which makes the work very simpler. The nodes will also manipulate or process the data which can be accessed by them. In this way it makes very simpler to make the datasets to be processed at a very fast rates and very efficient than the traditional architectures.

The core base Apache Hadoop framework is actually made up of the following modules. These are:

1. Hadoop Common: This module consists of all the necessary utilities and libraries which are useful for all the modules present in the framework.
2. Hadoop Distributed File System (HDFS): It is actually a distributed file-system which stores all the data on some hardware called as commodity machines. In this way it has main benefit that it increases the overall bandwidth of all the clusters present.
3. Hadoop YARN: This module is mainly responsible for managing all the resources in the clusters which are useful for the user's applications.
4. HadoopMapReduce: This is simply the implementation of the processing model of the Apache Hadoop. This processing model is known as the MapReduce programming model.

The term Hadoop has a very broader meaning. This term has a meaning which is very different from a simple terminology. Now if we define Hadoop as an ecosystem then we can say that it consists of so many additional software packages within it. So it means that Hadoop is very big collection of many of the software packages. All these packages can be installed on the top of the Hadoop. Some of these packages are Apache Pig, Apache Hive, Apache Hbase, Apache Phoenix, Apache Spark, Apache Zookeeper, Apache flume, Apache Sqoop, Apache oozie, Apache Storm. Apache Hadoop's MapReduce and the HDFS components are very much inspired by the Google research. They wrote some papers on these two models of the Apache Hadoop. We should also mention this point here which is important that Hadoop framework is actually written in the Java programming language. Though Java programming may be used for the MapReduce Programming but other platforms can also be used. Now after giving the introduction we will have a look on the need of Hadoop technology in the Healthcare systems.

II. THE NEED OF HADOOP FOR HEALTHCARE APPLICATIONS

There are usually a lot of crises among the people regarding the issue of the Big Data nowadays. So in a proper manner it does make a sense that this Big Data is also present in the Healthcare applications also. We need to analyze the uses of Big data in the very essential Healthcare applications. Now if we talk of where big data is used in the healthcare applications then we can take an example of the services like curing diseases, growing profits, preventing epidemics, improving the life quality of the people, preventing deaths, reducing the overall costs. All the labs and the hospitals make use of the Big data to reduce the overall costs in the services. We will also have some study of how the Big data is in Healthcare Industry. The New York based firm carried on a research in which it was analyzed that the US Healthcare industry in 2011 has generated about 150 billion gigabytes which is equivalent to 150 Exabytes of the data. The study also indicates that the data was generated by the various means like patient care, record keeping and the other requirements. Since this year there is very big growth in the data which has increased to about \$1.2 trillion in the Healthcare applications. Big data in the Healthcare systems usually come from very large healthcare datasets. Also these datasets are very much difficult to manage by using some traditional systems. It is not easy for them to manage such a large datasets. The use of very old data management techniques, make it very difficult to leverage the data. The Big data in the Healthcare systems is a very important concept as the data not only grows in size but it is also having so many forms like unstructured, semi-structured and it is also increasing at a very fast rate or speed. Now data analytics on this type of the data is somewhat very difficult and also data management is not so easy. So we can say that data analytics has become a very big issue in the Healthcare systems nowadays. It also leads to concept of the Big Data analytic technology since there are many challenges known to it. The data in the Healthcare Industry is nowadays changing into the digital form which is different from the earlier printed forms. So now when the data is getting digitalized it is very important to understand the uses of this Big Data in the Healthcare Industry. The main uses of Big Data in the Healthcare industry are population management, Clinical decisions and the disease surveillance. The Healthcare industry has just now started working on the Big data. Almost 80% of the data in the Healthcare Industry is in the unstructured forms. In this case, it is big challenge for the Healthcare industry to look into it and manage the problem. The Healthcare industries are looking ahead of how to manage this unstructured data in such a way that it becomes useful for them for various purposes in the healthcare industry. The data in the Healthcare Industry is increasing at a very fast rates so in the coming years it is a big challenge for the Healthcare industry to change their trends with the increasing data. They cannot get the sole profit from it but they need to have techniques which will help them to monitor the big data and use best of it that can easily leverage Big data in the Healthcare effectively.

Now let us talk about the main need of Hadoop in the Healthcare Data Solutions. An Information Solutions Architect Charles Boicey said that the Hadoop technology is the best solution for nowadays Healthcare systems. It helps them to store data in their native forms. If there was not Hadoop then it would have been a very big issue to make decisions about storing of the data in the Healthcare systems. Without its existence we couldn't make the right decision about how to store the data in the Healthcare systems. Whether we use the data warehouse or in some electronic forms. But now it is very easy for us to get everything in the Hadoop and start storing the entire data into just one single data store. It has been already predicted that the healthcare records of the patients in the year 2016 will likely increase from the tens of millions to the tens of billions of records. So to manage these very large records we need to think of the infrastructure and the costs associated with it. The main points which we can mention that what is actual need of Hadoop Technology for the Healthcare systems are as follows:

- The first point regarding the need of Hadoop technology for the Healthcare systems is Parallel Data Processing that is usually unconstrained.
- The next point is that Hadoop technology provides storage for the billions and trillions of the data records.
- The Hadoop technology has one more main feature already discussed in this paper above is that it has good fault tolerance and it has very high availability of the system

Hadoop technology is such a good technology which we need to implement for our Healthcare systems. Hadoop technology meets all challenges which are in the Healthcare systems by using the technologies like the MapReduce and HDFS System. Using these technologies Hadoop has the main capability of storing billions and trillions of records of data very easily. So it simplifies the use of Healthcare Systems. It also makes use of very cheap commodity hardware which is less costly to implement also.

Now we have a detailed study into 5 applications of the Healthcare systems where the Hadoop technology is implemented.

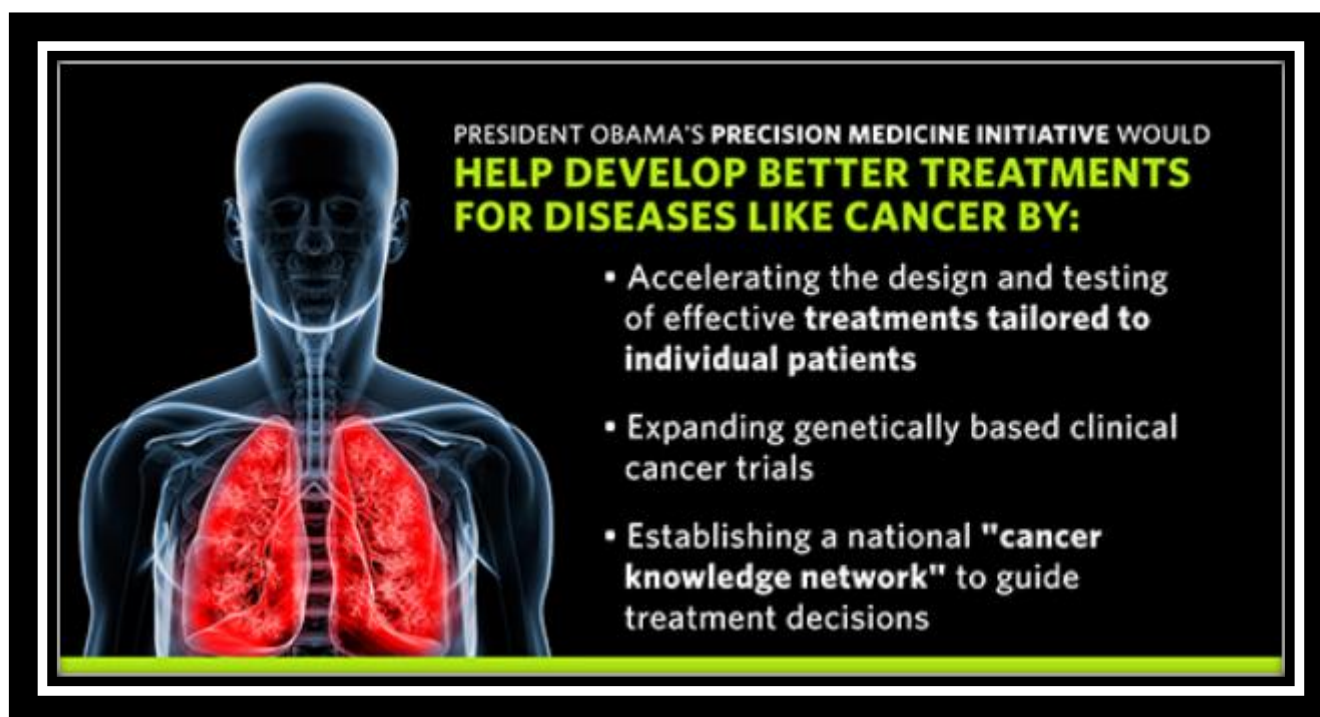
III. CASE STUDY OF HEALTHCARE APPLICATIONS USING HADOOP

Here we will have a look into some of the Healthcare applications which use Hadoop. These are as below:

1. **Hadoop technology in Cancer Treatments and Genomics:** When we talk of how Hadoop technology is to be implemented in Cancer treatments in the Healthcare services, then we will quote an example from the Amazon web services whose product manager Deepak Singh said that we have analyzed the growth of Hadoop technology in the Healthcare services including all the problems of the sequencing and other read mapping. The developers discovered the number of bioinformatics problems were present well in the Hadoop at a very good scale. Coming back to the reports of the Industry, they mention that that in the Human body there are about 3 billion pairs that make the Human DNA. So if we want to cure the cancer disease then the problem is that we need to manage all these DNA pairs in an effective way. To fight cancer we need to think how to find an effective treatment for it. As all of us know that cancer treatment till now is still not easy and cancer is still known as a incurable disease. This is because the Human DNA is very complex and it is not easy for us to fight this disease. Now the researchers of Oncology indicate that cancer usually mutates in some very complex patterns and its reactions are very different. The cancer disease is based on the genetic makeup of the patient. Hence, the oncology researchers now analyze that we need to find some personalized solution to this problem. And the cancer patients need to be provided with some personalized treatments which is based on the individual patient's genetic makeup. Using the Hadoop technology it will provide us a great help in the mapping of all the 3 billion DNA base pairs using the technique called as MapReduce technique.

Again the director of Health and life sciences, KetanParanjape says that they are trying their level best to improve the effects and use of Hadoop Technology in the Healthcare Systems. So that it will be have greater impact on it in the future. Now the goal of his idea was that Hadoop technology is to be used in the Healthcare systems in such a way that it will help us to analyze all the data and it can then provide everything among the millions of people and also finding a particular solution to the specific type of the cancer. David Cameron the PM of UK has announced some funding of about \$300m in August 2014 for a

project. This project was based on a goal of mapping 1,00,000 human genomes. The main goal of the project was to implement the use of Hadoop technology in Healthcare systems. Next one CASI was again developing some system which could generate the petabytes of genetic data of the individuals. This system was very helpful in identifying the gene of the particular individual and then searching out some drug meant for cancer treatment. Now we will analyze by a simple example that how Hadoop technology solves the problem of the genomics data. We take an example of a drug which we say that it is 45% effective against the deadly disease. The drug found may be 100% effective for the people of the same genetic profile or area. But it may be 0% effective for those people having different genetic profile, environment or who are not responding to this particular drug. Now take a case of the traditional database, it cannot the variety of Big Data. We should know that Human genome has 20,000 different genes and if we store this data in traditional systems, it means for each single person we need 1 mn DNA, which can take at least 20 billion rows of data. So we have to analyze and think that why the traditional databases are not capable to deal with the problem.



HADOOP TECHNOLOGY IN CANCER TREATMENTS.

2. **Hadoop technology in Monitoring Patient Vitals:** This is the main use of Hadoop technology in Healthcare systems. Coming to this point we can say that nowadays the Hadoop is used almost everywhere in the Hospitals. The main reason behind this is that the big data in unstructured forms is to be managed only with the support of the Hadoop. That is why we implement the Hadoop in almost all the hospitals now. If we take an example Children's Healthcare of Atlanta, in which more than 6,000 children are treated daily in the ICU units. The duration for which the patient stays in the ICU Unit can vary from some months to a year. What the Children's Healthcare of Atlanta do is that they use some types of sensors behind their beds. The function of these sensors is that it gives the warning signs to the doctors if reported. So it can give results in terms of heartbeats, respiratory rates or even Blood pressure .It gives out the data in very large chunks. This means that the Big data is coming out on the daily basis in the Healthcare of the Atlanta.

This issue can be handled only by the Hadoop. Any changes in the patterns are also reported to the doctors about the patients. This step was achieved by using some of the Hadoop ecosystems like Hive, Flume, Sqoop and Spark.



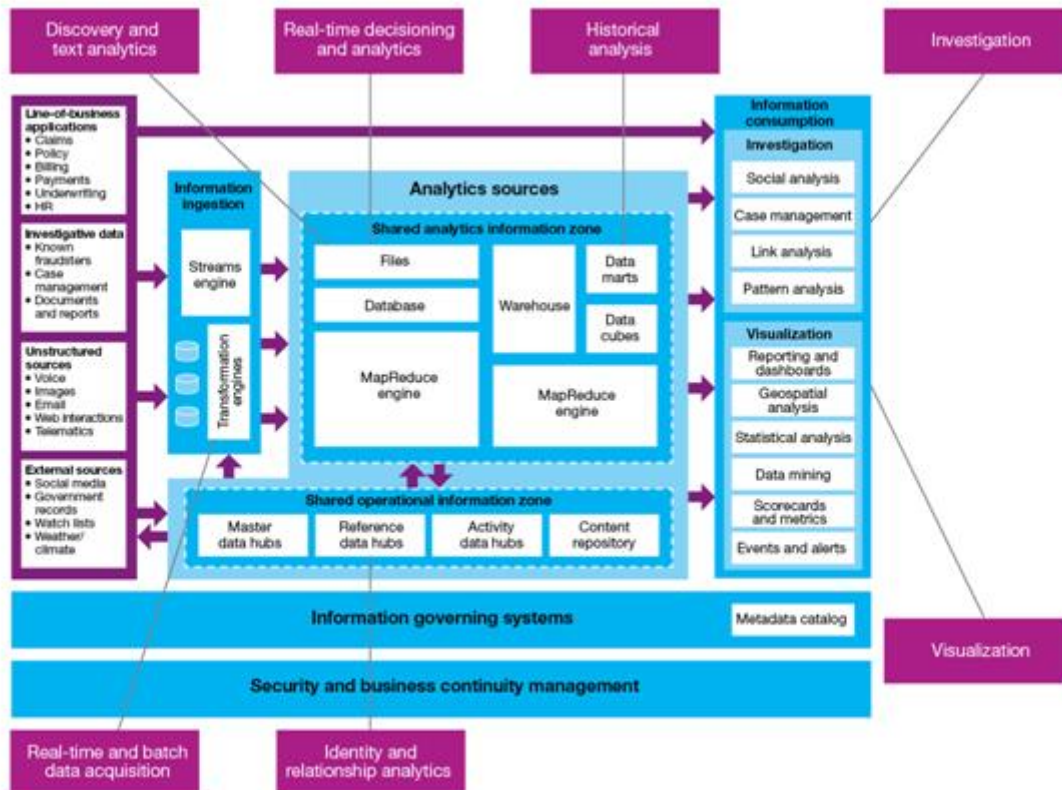
Hadoop technology in monitoring Patient Vitals.

3. **Hadoop technology in the Hospital Network:** The Cleveland clinic Company called as Explorys are making use of the Big data in their healthcare only to provide best clinical facilities. The other important achievements by there is the reduction of the costs and also the proper management of the population of the patients. It is a fact that this company Explorys has made the largest database consisting of the billions of records of data at a single data source only with help of the Hadoop technology. Hadoop technology is being used by this company for the use of their experts who analyze the data which are from different sources like payroll data, or the electronic health records. The tool which was developed by this company can also be used by experts for the purpose of the data mining which can further extract the useful knowledge patterns of the patients and their effective treatments. So when they analyze this the health practitioners have a clear understanding of the drugs on the particular population of the people and also of the best treatments which could be provided to them.
4. **Hadoop technology in Healthcare Intelligence:** Health Insurance companies also sometimes make use of the Hadoop technology in analyzing the risks and costs associated with them. So we can also say that Hadoop has the business applications. The insurers analyze the data in a group and divide it in a proper

way. The healthcare agencies can improve their business solutions in an effective way using the Hadoop technology. Now if we take an example of a Healthcare insurance company which need to understand that why and how a particular area or group of people are prone not prone to certain disease. So when they analyze this thing they will make an insurance policy for all the people of that particular region. This will include a large datasets which comprise of all the information regarding the symptoms, opinions, treatments, geographic regions etc. Only to manage these large datasets they make use of Hadoop technologies like Hadoop pig or Hive.

Dignity Health another US based Company analyzed that Hadoop technology is growing at a very fast track and it is very useful in the Healthcare services. Since the data in applications is dynamic and changing tremendously, so the only way to analyze this data is to make use of Hadoop technology. Hadoop technology can meet all the challenges which are now faced in the present business or healthcare applications. Let us take an another example of some disease in USA like Sepsis condition which affects almost million people. Now if we analyze the disease properly using Hadoop we can save many lives from this deadly disease. The sooner we predict this disease, it will good as we still can save some lives who can be cured at a right time. The need is that bring all the data at one place so that it can be easily analyzed. It can be properly analyzed to find some permanent solution for a certain disease. This can make the Hadoop a big golden opportunity to work for the Healthcare applications.

5. **Hadoop technology in Fraud Prevention and Detection:** Here we can think that more than 10% data of Healthcare insurance are affected by frauds. This problem is a very big problem in the worldwide now. This issue is increasing at an exponential rate. With the use of Hadoop technology the insurers can make use of some models and patterns with the help of which they can identify the fraudsters and can prevent and detect all the frauds very easily with the use of real time historical data. In this way we make use of MapReduce techniques etc to overcome all the issues of frauds etc and store the large datasets with the help of NOSQL database. With the help of Hadoop technology there is large fallout in the cost of storage of data in the healthcare applications which has fallen from \$1,00,000 to about \$1,200 per year. So we can say that increasing demand of the Hadoop technology in Healthcare systems can make a new trend in it and we can provide only the personalized treatments to all patients and that too at a very less and affordable costs.



Hadoop technology in Fraud Prevention and Detection.

IV. CONCLUSION

After finishing writing this research paper on the topic Hadoop technology we can give the conclusion that the Hadoop technology is best for simplifying the problems in the Healthcare systems. It can be implemented easily in the Healthcare systems as it overcomes all the limitations of the traditional systems. So in this way we can give personalized treatments to all the patients at a very controlled costs and can analyze the data at a single data store to conclude with the permanent treatment for the particular disease. The main point is that Hadoop technology has found the solution to the Big data problem in the Healthcare systems. It will be used in the future also in all the Healthcare systems to simplify the work.

V. ACKNOWLEDGEMENT

I am thankful to all the people who motivated me to the research work on this. I am also thankful to all the research scholars and people who have previously carried on the research work on this or either published some research papers related to this topic, with the help of these people I was able to write this research paper on this particular topic.

VI. REFERENCES

- AmeyaNayak, Anil Poriya, DikshayPoojary, “**Types of NOSQL Databases and its comparison with the relational databases**” March 2013, International Journal of Applied Information Systems, Volume 5, No. 4, www.ijais.org.
- A B M Moniruzzaman and Syed AkhtarHossain, “**NOSQL Database: New Era of Databases for Big Data Analytics-Classification,Characteristics, and Comparison**” International Journal of Database theory and Application.
- Santhosh Kumar Gajendran“**A Survey on NOSQL Databases**”.
- <http://www.google.com>
- <http://www.wikipedia.com>
- <http://www.youtube.com>
- <https://www.ibm.com>
- BDV Association Document “Big Data Technologies in the Healthcare”.
- Anusha B. Dhakite and Prof. Sameer Y. Thakur, “ **A Survey on Hadoop Technology and Its Role in Information Technology**” 2015, Volume 5, Issue 4, www.ijarcsse.com
- Roshani K. Chaudhari and Prof. D. M. Dakhane, “**Contribution of Hadoop to Big Data Problems**” April 2015, volume 5, Issue 4, www.ijarcsse.com









