ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

An Investigation of Big Data in Healthcare using Hadoop

Archita Bhatnagar

Asst. Professor, Dept. of Computer Science Engineering, Subharti Institute of Technology & Engineering, Meerut, U.P., India

ABSTRACT: Healthcare has nowadays become more advance. With this advancement, the data used in it has become vast and is increasing volume continuously. For this much volume of data, proper analysis has to be done to presume the outputs regarding NO. OF PATIENTS, MEDICINE STOCK, DISEASE ATTACKS etc. Its implementation is targeted towards the future use. The big data in healthcare is used to predict epidemics, cure of disease, improve quality of life and avoid preventable deaths. With increasing population, the data of patients is also increasing at pace. Therefore, *we target towards storing, analysing and processing of the big healthcare data using HADOOP technology.*

KEYWORDS: Healthcare, Analysis Big data, Hadoop.

I. INTRODUCTION

The rapidly growing healthcare systems are producing volume of structured, unstructured and unequal data which has to be accumulated, managed and analyzed. The increasing population and awareness among people has led to the growth in data which is placed in multiple places, without any structure and is unmanaged. This creates havoc in searching, storing and processing of the data.

This problem is solved by big data analytics which provide tools for aiding the process of care and disease exploration by making predictions. This technique allows such big data, which is unmanageable even by big database, to get stored, searched, managed and processed in an easy manner. Here, we are using HADOOP for the above said task.

Hadoop is a tool for storing and processing big data efficiently using statistics and probability. It works with processing components like query language, file storage mechanism and cluster computing framework. It is able to search, store, manage and process such a big data using its components which allow it to bring out the meaningful, managed and predicted data.

II. LITERATURE SURVEY

JH Bemmel, M A Musen provides a brief overview of history of the patient record, starting with Hippocrates to be followed by modern views on the structure of patient record, the development and use of computer based patient records, the entry of data into the CPR, coding and standardization, representation of time in CPR and clinical use of CPR [4]. *Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard and Kayvan Najarian* discuss major challenges with a focus on three upcoming and promising areas of medical research: image, signal, and genomics based analytics [1]. *Mukesh Borana , Manish Giri , Sarang kamble , Kiran Deshpande , Shubhangi Edake* mention how the healthcare factor become more advance in modern world. This includes that the health care data should be properly analyzed so that we can deduce that in which group or gender, diseases attack the most [3]. *Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, Maya Sanghati* propose an alternative technique which is Dynamic Hadoop Slot Allocation by keeping the slot-based allocation model. It relaxes the slot allocation and depending on their needs allows slots to be reallocated to either map or reduce tasks [2]. *Jimeng Sun and Chandan K. Reddy* proposed sourses and techniques of big data in healthcare [6]. *Lidong Wang and Cheryl Ann Alexander* discuss Big Data benefits, its applications, opportunities in medical areas and health care, methods and technology progress about Big Data, Big Data challenges in medical applications and health care[7].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Priyanka K, Prof Nagarathna Kulennavar gives a brief introduction about how we can uncover additional value from health information used in health care centers using a new information management approach called as big data analytics [8]. *Sanskruti Patel and Atul Patel* studies the impact of implementing the big data solutions on the healthcare sector, the potential opportunities, challenges and available platform and tools to implement Big data analytics in health care sector [11]. *Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang* provides an overview of recent developments in big data in the context of biomedical and health informatics. An outline of the key characteristics of big data is presented and how medical and health informatics, translational bioinformatics, sensor informatics, and imaging informatics will benefit from an integrated approach of piecing together different aspects of personalized
information from a diverse range of data sources, both structured and unstructured, covering genomics, proteomics, metabolomics, as well as imaging, clinical diagnosis, and long-term continuous physiological sensing of an individual [12].

III. OBJECTIVES

To handle the huge data sets, Hadoop is now used by finding correlations in data sets. In healthcare solutions, the analysis has to be done on latest data as well as to provide high quality care. The following are several characteristics of healthcare data that make it unique:

1. *Data is saved in multiple places*

Healthcare data exist in several places from EMRs to HR systems and different departments depending on the field. This data is then aggregated into central system and make this data accessible and actionable.

2. *Healthcare data also occurs in different formats*

This data is saved in the formats like text, numeric, paper, digital, pictures, videos, multimedia, etc. A data can be present at different locations in different formats, which cannot be joined.

3. *Data is structured as well as unstructured.*
4. *New research and advancements in healthcare practice.*
5. *Complex data*

IV. PROBLEM STATEMENT

We are living in **second most populated country** in the world. As the population is increasing with high pace and the living conditions are changing, the health issues are also moving towards growth. Heath issues are to be judged and cured by the hospitals which are there in every region of country. If we talk about the hospital services of a particular region, the no. of patients depend upon the population of that region and living standard of the people. Sometimes, the patients in a particular region increase so much that the admission facilities to all the patients is not fulfilled at par.

The changing style of living is contributing towards generation of different type of diseases in range of time. With the production of different diseases, it is mandatory for the healthcare units to be vigilant about the facilities needed then.

Also, our country hails to be **seventh largest economy** in the world. This economy highly depends on the health of people who are living and contributing to it. There is an economy cycle with Employment- Food needs- Health-Services- Economy share. Here, HEALTH plays a vital role in the economy cycle and if this module is disturbed, the whole economy of the country will get disturbed.

Additionally, the patients needs to be treated with **MEDICINES** which the hospitals have to order and purchase from the pharmacy and manufacturers. Most of the medicines have a narrow margin between manufacture and expiry and after expiry of the medicines it is of no use to chemists which in turn gives loss to them. This depends upon the no. of patients who are admitted and are in need of the medicines.

1. So, the first problem with healthcare units is the prediction of no. of patients who can be admitted into the hospital.

2. Secondly, since the patient needs full care and attention in the hospital, it is mandatory to make a balance of infrastructure and the staff arrangement to manage the admissions effectively.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

3. Thirdly, the medicine stocks have to be made available in such a way that they get fully utilized; not expired and are not out of stock at any point of time. Also, there is some kind of life saving drug which the healthcare systems need to have all the time as and when needed.

4. Fourthly, there are some facilities which are not provided by the hospitals and have to refer to the specialty hospitals. This brings down the budget of the hospital as the patient who is not treated in a hospital, will not pay anything.

V. PROPOSED SOLUTION

With every problem in the world, it is required to seek a solution. The solutions to above mentioned problems are proposed in this section.

1. Firstly, the no. of patients who can take admission in coming month/year can be predicted depending upon the no. of patients who were in the hospital last month/year. This will be based on the previous statistics of data.

2. With the prediction of this data, the infrastructure and the staff can be arranged accordingly. This will provide ease in the management of human resource as well as infrastructure. There will be no shortage of staff and infrastructure will be available to each and every individual.

3. This prediction will make the chemists and pharmacists decide that how much drug they have to purchase this month/year so that it may not fall short and may not expire due to short usage.

4. Also, this type of prediction will decide the pattern in which patients are coming. It will analyze the patients and the diseases they are facing. This will make the hospital to settle up with new facilities if it is in much demand.

VI. METHODOLOGY

In order to make predictions which are proposed in the previous section, we will be working on APACHE HADOOP 2.7.0. This is a tool used on big data using the concept of statistics and probability. Hadoop uses its modules for storing and processing the data.

We will be working according to an algorithm- Mapreduce. This algorithm is used for searching, sorting, analyzing and predicting of the big data.

MapReduce Algorithm:

This algorithm contains two elements, MAP and REDUCE. ***Map*** takes a set of data and converts it into another set where all the units of data are broken into tuples. ***Reduce*** task take output from the map as input and converts those tuples into reduced format. Reduce task is always done after the Map task.

The Algorithm

Step 1: **Map Stage**

The mapper's job is to process the input data. This data is stored in Hadoop file system (HDFS). After processing the data several small chunks of data is created.

Step 2: **Reduce Stage**

The reducer's job is to produce new set of data which is reduced from the mapped data and get stored in HDFS.

Step 3: After the data gets reduced, the Hadoop sends the data to servers in the cluster for issuing task, verifying task and copying the data around cluster. Here, the unstructured data gets structured. After clustering, the data is sent back to the Hadoop server.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Here in *Fig.6.1*, reduction of data is shown.

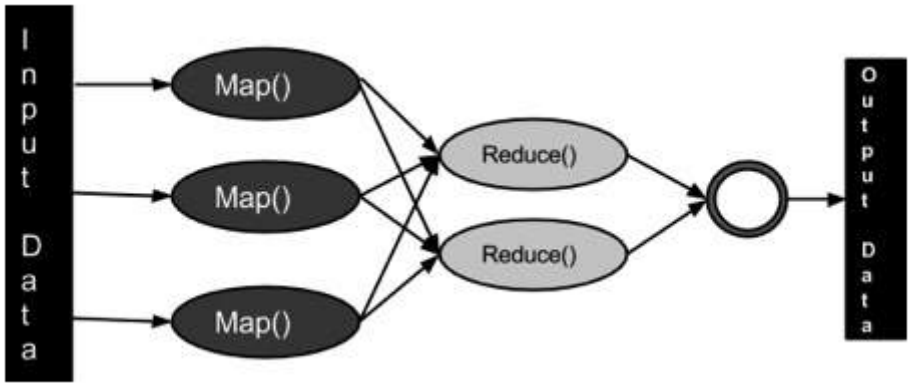

Fig.6.1: Mapreduce

In *Table 6.1*, the input and output value id data is shown.

	Input	**Output**
Map	<k1, v1>	list (<k2, v2>)
Reduce	<k2, list(v2)>	list (<k3, v3>)

Table 6.1 : Mapreduce

VII. MERITS & DEMERITS

Merits:

1. Data Processing in the healthcare unit becomes easy.
2. Easy to make predictions.
3. Accurate with large volume of data.
4. Less time consumption.
5. Management of data in one go.

Demerits:

1. The data has to be really big; else the predictions will not be accurate.
2. Servers for clustering and Hadoop is needed; transfer of data from one server to another is to be managed

REFERENCES

[1] Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard and Kayvan Najarian, Big Data Analytics in Healthcare, Hindawi Publishing Corporation BioMed Research International Volume 2015, Article ID 370194, 16 pages http://dx.doi.org/10.1155/2015/370194, 2015.

[2] Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, Maya Sanghati, Health Care Analysis Using Hadoop , INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 12, DECEMBER 2015 ISSN 2277-8616 279 IJSTR©2015 www.ijstr.org , 2015

[3] Mukesh Borana , Manish Giri , Sarang kamble , Kiran Deshpande , Shubhangi Edake, Healthcare Data Analysis using Hadoop, International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 07 | Oct-2015.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

[4] M.A. Musen and J.H.Bemmel, Handbook of Medical Informatics, HOuten: Bohn Stafleu Van Loghum, 1999.
[5] https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
[6] *Jimeng Sun and Chandan K. Reddy*, Big Data Analytics for Healthcare, Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.
[7] Lidong Wang and Cheryl Ann Alexander, Big Data in Medical Applications and Health Care, *American Medical Journal*, Review Articles, 2015.
[8] Priyanka K, Prof Nagarathna Kulennavar, Survey On Big Data Analytics In Health Care, International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014
[9] Gesundheit Österreich Forschungs- und Planungs GmbH, Study on Big Data in Public Health, Telemedine and Healthcare, Final report EUROPEAN COMMISSION, 2016
[10] Cognizant 20-20 Insights, Big Data is the Future of Healthcare, 2012
[11] Sanskruti Patel and Atul Patel, A BIG DATA REVOLUTION IN HEALTH CARE ECTOR: OPPORTUNITIES, CHALLENGES AND TECHNOLOGICAL ADVANCEMENTS,International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016.
[12] Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong,and Guang-Zhong Yang, Big Data for Health, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015
[13] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, A Review Paper on Big Data and Hadoop, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014
[14] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplifed Data Processing on Large Clusters, OSDI, 2004
[15] Vidyullatha Pellakuri, Dr.D. Rajeswara Rao, Hadoop Mapreduce Framework in Big Data Analytics, *International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 3– Feb 2014*
[16] Abdelrahman Elsayed, Osama Ismail, and Mohamed E. El-Sharkawi, MapReduce: State-of-the-Art and Research Directions, *International Journal of Computer and Electrical Engineering, Vol. 6, No. 1, February 2014*
[17] Shafali Agarwal, Zeba Khanam, Map Reduce: A Survey Paper on Recent Expansion, *International Journal of Advanced Computer Science and Applications, Vol. 6, No. 8, 201*
[18] A.Pradeepa, Dr. Antony Selvadoss Thanamani, HADOOP FILE SYSTEM AND FUNDAMENTAL CONCEPT OF MAPREDUCE INTERIOR AND CLOSURE TROUGH SET APPROXIMATIONS, *International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013*
[19] Silky Kalra, Anil lamba, A Review on HADOOP MAPREDUCE-A Job Aware Scheduling Technology, *International Journal of Computational Engineering Research (IJCER)*
[20] Aaron McKenna, Matthew Hanna, Eric Banks, et al., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.* 2010 20: 1297-1303 originally published online July 19, 2010