Healthcare Analysis Using Hadoop Framework

MS.Minu[1], Ishan Meena[2], Pratyush[3], R. Aravind[4], Vijayditya Sarker[5]
[1, 2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3,4, 5] SRM Institute of Science and Technology, Chennai, Tamil Nadu - IN 600089

Abstract- *Nowadays, in the healthcare industry there is a significant rise in the number of doctors, patients, drugs and medicines. The analysis and prediction of future health conditions are still in developing stage. The data which is exerted in a little amount has risen greatly from a few bytes to terabytes, not only has the storage increased but also the dataset maintenance. The traditional method of using data mining and diagnosis tools is difficult, therefore the need for big data tools and techniques arises. Big data and Big data analytics are a rising technology. Primary sets of such data are created for the medical and healthcare contexts. There isn't an ideal method to measure the patient satisfaction. This paper presents the ideas and methodology using data mining techniques such as clustering which is acquired from the datasets and put forward into big data tool like HADOOP for effective analysis of healthcare data.*

Keywords- Big data, Healthcare, Clustering, Hadoop.

I. INTRODUCTION

Healthcare Industry is one of the world's greatest and most extensive ventures. Amid, the ongoing years the Healthcare administration around the globe is changing from infection focused to a patient-focused model and volume-based model. Teaching the predominance of Healthcare and diminishing the cost is a guideline behind the creating development towards-esteem based Healthcare conveyance model and patient-focused mind. The volume and interest for huge information in Healthcare associations are developing little by close to nothing. To give successful patient-focused care, it is fundamental to oversee and analyse the huge amount of data sets. The traditional methods are obsolete and are not sufficiently adequate to break down enormous information as assortment and volume of information sources have expanded and a very large rate in previous two decades. There is a requirement for new and creative tools and methods that can meet and surpass the capacity of overseeing such a huge amount of data being generated by the healthcare department.

The social insurance framework of healthcare departments is community in nature. This is since it comprises of a substantial number of partners such as doctors with specialization in different sectors, medical caretakers, research centre technologists and other individuals that cooperate to accomplish the shared objectives of decreasing medicinal cost and blunders and also giving quality healthcare experience. Every one of these partners produce information from heterogeneous sources, for example, physical examination, clinical notes, patients' meetings and perceptions, research facility tests, imaging reports, medications, treatments, overviews, bills and protection.

The rate at which information is being generated from heterogeneous sources from various healthcare department has incremented exponentially on the daily basis. Therefore, it is becoming hard to store, process and break down this inter related information with traditional dataset handling applications. Nonetheless, new and efficient methods and systems are in addition to provide great processing advancements to store, process, break down and extricate values from voluminous and heterogeneous medical information being generated in a continuous way. Henceforth, the medicinal services framework is quick turning into a major information industry.

Generally, medicinal services information has developed enormously in both organized and unstructured way, to a great extent driven by the requests of always extending information parched populace what's more, operational attributes of e-health stages. This dangerous multi-dimensional development has lead scientists, to add numerous more watchwords to portray Healthcare Big Data (HBD). It isn't only the volume however their assortment, specifically the kinds of sources that deliver information and the objective sorts that request them are excessively different and various in Healthcare area. These incorporate medicinal services workforce (doctors, clinical staff, parental figures), benefit giving organizations (counting safety net providers), healing facilities with resources, clinicians, government controllers, drug stores, pharmaceutical makes (with look into groups included), and therapeutic gadget organizations.

II. RELATED WORKS

Today, social insurance industry is experiencing a stage which is seeing a regularly expanding rise in number of specialists, patients, illnesses and pharmaceuticals. The human services information that existed in couple of megabytes in the past has seen an exponential ascent to petabytes and exabytes

as of late and billions of dollars are being spend in the capacity of information as well as in keeping up data sets around the world. Since there is an extensive augmentation in data sets, it is getting to be hard to deal with it with conventional information mining and analysis devices, so it is turning into a need to utilize huge information instruments and procedures in medicinal services industry.

Healthcare Big Data is more mind boggling than Big Data emerging from some other basic area in light of the fact that an assortment of information sources and techniques are followed in conventional doctor's facility settings and in social insurance organization (e-Health). Keeping in mind the end goal to accomplish their essential objective, which is to improve experience while managing reliable care inside money related practice and regard for government controls, the HBD ought to be dissected to decide the fulfilment level. [5] [1]

A large portion of the Big Data are in a type of unstructured information, significant strides of enormous information, administration in human services industry are information obtaining, capacity of information, dealing with the information, investigation on information, and information representation.

A gigantic measure of information is created day by day by the medicinal associations, which on the whole comprises of patients. Human services focuses, restorative masters and obviously, the sicknesses. The information is gigantic and gives a knowledge into future forecasts, which may keep most extreme restorative cases from happening. Be that as it may, without huge information investigation methods and Hadoop bunch, this information stays futile [3] [4].

Social insurance has expanded its general, an incentive by receiving huge information strategies to break down and comprehend its information from different sources.

The medicinal data sets comprise of huge volumes of information which are typically produced from differing sources, for example, doctors' case notes, clinic affirmation notes, release rundowns, drug stores, insurance agencies, restorative imaging, labs, sensor-based gadgets, genomics, internet-based life and also articles in therapeutic diaries. Healthcare information is anyway exceptionally intricate and hard to oversee. This is because of the galactic development of medicinal services information, the rate at which this information is produced and in addition the different information generated.

The catching, stockpiling, investigation and recovery of healthcare related information are quickly moving from paper-based framework towards digitization. In any case, the immense volume and additionally the intricacy of these information makes it troublesome for the information to be prepared and dissected by conventional methodologies and systems. Thus, advancements, for example, distributed computing and virtualization are presently bit by bit utilized for preparing enormous information adequately and safely in medicinal services. Thus, the data being generated by the medical industry is quickly turning into a major information industry for analysis of patient data.[2] [7].

Notwithstanding the traits of enormous information specified above, it is fundamental that instruments exist for representation and comprehension of the data and relations between the information present in the datasets, which is called, Business Insight (BI). This requires information stockpiling and administration, equipment and programming assets, proper space learning, and new techniques and advances. Joining enormous information with examination can give a significant preferred standpoint to settle on auspicious and efficient choices identified with 1) cost, 2) time, 3) item improvement, and 4) enhancement.

A humongous measure of information is caught and put away in various configurations (organized, semi-organized and unstructured), from various sources (sensors, machines, applications, web, IoT) and put away by the associations. The information is caught, put away, handled in clusters or constant -with the assistance of calculations or mechanical procedures.

Utilization of these strategies fluctuate for various areas, going from avionics, car industry, managing an account and capital speculations, correspondences, vitality, utilities and mining, government, wellbeing industry, protection, retail, innovation, and so on. It is imperative for these businesses to make most out of the frail signs from a few key information sources both organized and unstructured and convey an ongoing effect for a simple, snappy and powerful basic leadership [7] [11].

Huge information gives an extensive variety of chances for the healthcare industry. For example, the Harvard Business Review uncovered that the appropriation of huge information in healthcare has prompted the disentanglement of Information Technology (IT), confirm based and esteem based prescription, better preventive care and customized treatment.

Besides, huge information uncovers examples and patterns in the information which helps during the time spent

on diagnosing and treating the patients. In this way, the organization of enormous information in medicinal services has given rise to the change of patients' mind at a lower cost and also expanded the patients' fulfilment. Be that as it may, one of the real difficulties of enormous information in Electronic Health Records (EHR) is security and protection issues of the data stored. This is on account of Electronic Health Records are exceedingly defenceless to unauthorized access, information breaches.

In this manner, the human services framework is portrayed by high blunder rates and staggering expense which result in a high rate of mortality. Subsequently, this paper looks at the idea of huge information with regards to social insurance, the apparatuses and strategies utilized for the execution of huge information in medicinal services and in addition the advantages and difficulties of huge information in human services [7].

III. IMPLEMENTATION

In order to process a huge amount of health data records at once we need efficient tools and methodologies.

The proposed papers use the Hadoop Framework to handle the data, and the algorithm being used is Map Reduction.

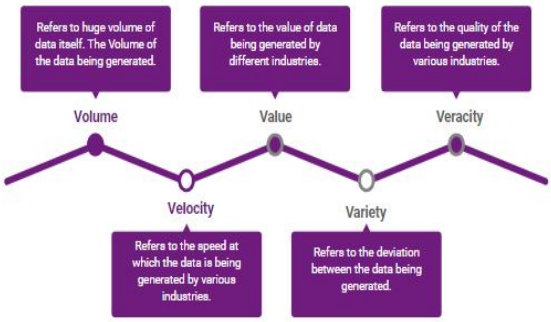


Fig 1: 5V's of Big Data

Hadoop Framework is a collection of open-source software utilities that facilitate using a network of many computers (Clusters) to take care of issues including tremendous amount of data and there analysis.

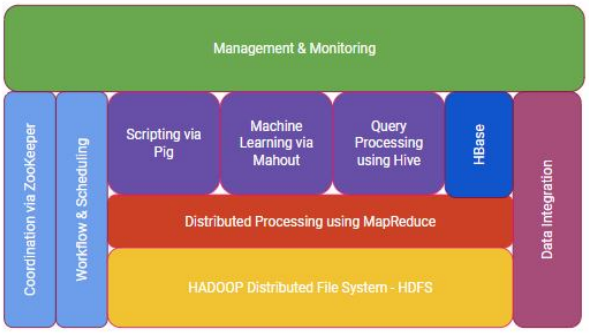


Fig 2: Apache Hadoop Ecosystem

The **Hadoop Distributed File System** (HDFS) is the essential information stockpiling framework utilized by Hadoop applications. It comprises of NameNode/The Master and DataNodes/The Slave design to execute a disseminated record framework called Hadoop Distributed File System to get to information crosswise over exceedingly adaptable Hadoop Clusters in an effective way. Hadoop Framework in total consists of 5 daemon processes namely:-

1. **NameNode:** NameNode is utilized to store the Metadata (data about the area, size of files/blocks) for HDFS. The Metadata could be put away on RAM or Hard-Disk. There will dependably be just a single NameNode in a cluster. The only way that the Hadoop framework can fail is when the NameNode will crash.

2. **Secondary NameNode:** It is used as a backup for NameNode. It holds practically same data as that of NameNode. On the off chance that NameNode falls flat, this one comes into picture.

3. **DataNode:** The actual user files or data is stored on DataNode. The number of DataNode depends on your data size and can be increased with the need. The DataNode communicates to NameNode in definite interval of times.

4. **Job Tracker:** NameNode and DataNodes store points of interest and genuine information on HDFS. This information is likewise required to process according to users' prerequisites. A Developer writes a code to process the information. Processing of data can be done using MapReduce. MapReduce Engine sends the code over to DataNodes, making jobs in multiple nodes running alongside of each other. These employments are to be persistently observed by the Job tracker.

5. **Task Tracker:** The Jobs taken by Job Trackers are in genuine performed by Task trackers. Each DataNode will have one task tracker. Task trackers communicate with Job trackers to send statuses of the undertaken job status.

HDFS bolsters the quick exchange of information between Master and Slaves as it is combined with MapReduce, an automatic system for information handling and to access information at a higher rate.

When HDFS takes in information, it separates the data into partitioned squares and appropriates them to various nodes making the system effective via parallel processing.

In addition, the Hadoop Distributed File System is exceptionally intended to be fault tolerant. The framework reproduces, or duplicates, each bit of information various occasions and conveys the duplicates to singular hubs, putting no less than one duplicate on an alternate server rack than the others. Accordingly, the information on hubs that crash can be discovered somewhere else inside a group. This guarantee preparing can proceed while information is recuperated.

HDFS has an Master & Slave architecture. A HDFS group comprises of a solitary NameNode, A Master server that deals with the data stored and manages access to data by the authorized users in the Hadoop environment.

The Hadoop Distributed File System take it's core from Google File System (GFS), a restrictive document framework laid out in Google technical papers, and in addition IBM's General Parallel File System (GPFS), a configuration that lifts I/O by writing blocks data into disks in parallel to provide efficiency. While HDFS isn't Portable Operating System Interface demonstrate consistent, it echoes POSIX configuration style in a few angles.

Usually, a file is splitted into one or more block depending upon the size of the file and are put away in an arrangement of DataNodes. The NameNode executes tasks like opening, shutting, and renaming data files and folders. It is also responsible for decides the mapping of data to DataNodes. The DataNodes are also in charge of managing the read and write demands from the authorized users. The DataNodes does the job of data block creation, deletion, and replication when it is given the instruction to do so from the namenode for a particular block.

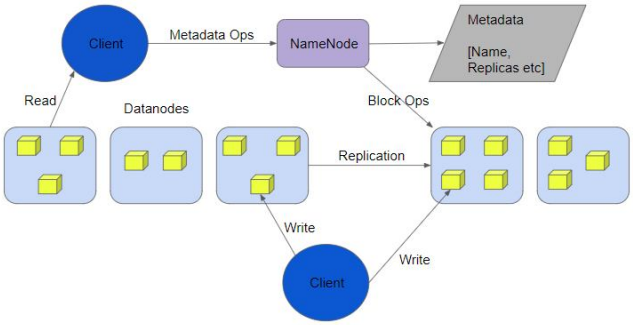


Fig 4: Namenode & Datanode Interaction

Map Reduction algorithm contains two important tasks, namely Map and Reduce.

- Mapping – Attained by Mapper Class
- Reduction – Attained by Reducer Class.

MapReduce utilizes different numerical calculations to separate an errand into little parts and dole out them to various frameworks.

MapReduce calculation helps in sending the Map and Reduce errands to proper servers in a bunch. The tasks are executed in parallel in all the different nodes and finally the result is returned to the user.

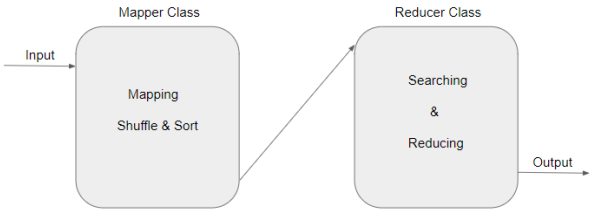


Fig 5: Mapper & Reducer Class

The reduce task is always performed after the map job.

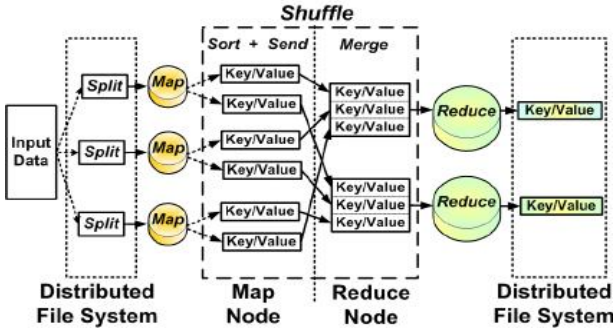
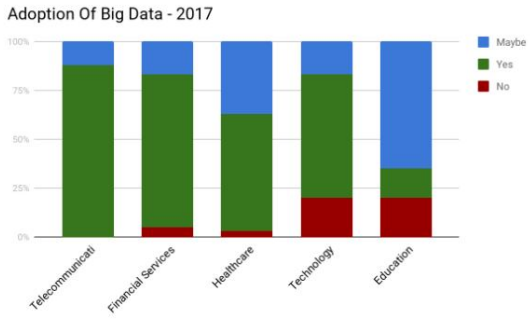


Fig 6: Map Reduction

Big Data in Healthcare Industry

Fig 7: Big data Adoption in Industry



As can be seen above most of the industries are already using or shifting towards Big Data including the Healthcare Industry.

The Healthcare Industry uses it primarily for the following

- Data Warehouse Optimization
- Patient Analysis
- Predictive Maintenance

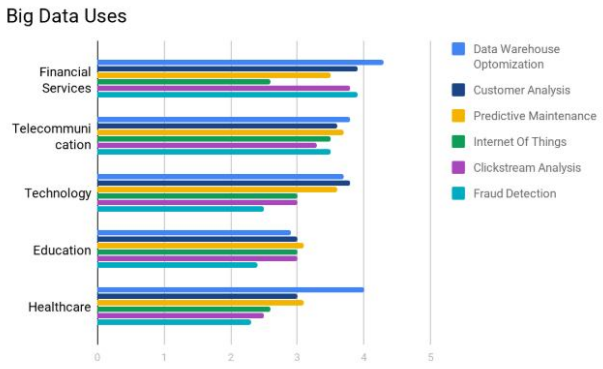


Fig 8: Big Data Uses

Hadoop uses MapReduce algorithm to create tasks, called jobs which can be executed independently on different clusters (DataNodes) while the result is fetched back to a single node (NameNode) for output.

In our example, the map function is in the following format –

<serial_number, name, drug_prescribed, gender, total_expenditure_on_prescribed_drugs>

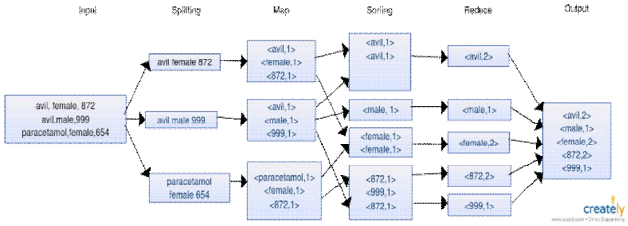


Fig 9: Map Reduction of Patient Data

As can be seen above, the system will group together the items having the same key. Finally, the system provides the requested output.

Patient's data is stored in a centralized repository which makes the system cost effective by reducing number of storage warehouses as well as eliminates any sort of data redundancy, which leads to the system being consistent as well.

IV. FUTURE SCOPE

At present as the medicinal services showcase is developing, it has turned out to be certain that associations which are equipped for utilizing the force of examination are showing a sensible favored outlook in bit of the general business over their opponents. Big Data analysis in medical services has indeed, now turned into the main thrust for making openings for creating ideal treatment pathways, enhancing the restorative setback extent and better managing clinical decision candidly steady systems.

With taking off healthcare services costs and additionally developing administrative weights for both moderateness and changes in clinical results – Analytics has risen as a silver covering for this industry. Examination in medical services has demonstrated to create experiences that not just lower add up to costs, decrease wasteful aspects, and distinguish high hazard populace yet additionally can foresee a patient's future social insurance needs.

Thus, with examination today, emotionally supportive networks are presently being subjected to different measurable and computerized reasoning systems. This is further bringing about advancement of significant bits of knowledge, for example, ID of different patient hazard factors, gathering of patients in light of changed wellbeing conditions, arrangement of noteworthy data to doctors at the purposes of care and above all, quantifiable advancement on healthcare services results. The presentation of examination in medical services hence, has helped in beating a lot of difficulties making genuine incentive for this part. A portion of these ordinary difficulties include:

- Diverse information sources that make it hard to make a solitary hotspot for reality.
- Information not being accessible in an opportune way, so choices are not information driven.
- Lack of an unmistakable vision on how the association can profit by examination.
- Too many manual frameworks sent, bringing about lacking electronic information.
- The culture not being prepared to end up an information driven association.

Anyway, to conquer the greater part of the above difficulties, it is basic that the data being recorded via the patients should be put to good use. Additionally, all clinical data should be put away in their standard information organizations. For example - EHRs must be changed into useful information on which analysis can be done, in order to effectively get significant insight from it and improvements can be achieved over the same to provide personalized healthcare experience to the patient.

To finish up along these lines, it tends to be said that analysis today is undoubtedly an urgent process in medical services that will altogether reshape its scene in the upcoming years. Besides, investigation is likewise being current drive of a move in this industry towards arrangements that are fit for conveying genuine esteem, for example,

- Improving clinical nature of care.
- Improving tolerant security and lessening therapeutic mistakes.
- Improving wellbeing, avoidance and ailment administration.
- Optimizing supply chains and human capital administration.
- Improving hazard administration and administrative consistence.

V. CONCLUSION

This paper has discussed about the substantial use of big data in healthcare, which is based on the need for the big data storage and extracting the information combining HADOOP platform and the data mining technique-cluster analysis. Depending on the vastness of the patient information, MAP REDUCTION algorithm technique can be used to refine the data on the basis of making clusters of each dataset. By using the methodology, we can suggest the better treatment of healthcare diseases and other related problems in less time and with cost effectiveness. This will help the doctors to give a better treatment with more processed information and also help people who aren't available to better treatment due to high end costs.

REFERENCES

[1] Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam S, Big Data in Healthcare Management: A Review of Literature, American Journal of Theoretical and Applied Business, Vol. 4, No.2, 2018, pp.57-69.

[2] Delia Ioana Dogaru, Ioan Dumitrache, Holistic Perspective of Big Data in Healthcare, Grigore T. Popa University of Medicine and Pharmacy, Siania, Romania, June 22-24,2017.

[3] KaiYu Wan, Vangalur Alagar, Analyzing Healthcare Big Data for Patient Satisfaction, 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017).

[4] Manpreet Singh, Vandan Bhatia, Rhythm Bhatia, Big Data Analytics: Solution to Healthcare, in 2017 International Conference of Intelligent Communication and Computational Techniques(ICCT), Manipal University Jaipur, Dec 22-23, 2017.

[5] Priyank Dhaka, Rahul Johari, HCAB: HealthCare Analysis and Data Archival using Big Data Tool, Univrsity School of Information and Communication Technology, G.G.S. IPU, New Delhi-78, India.

[6] Fadoua Khennou, Youness Idrissi Khamlichi, Nour El Houda Chaoui, Designing A Health Data Management System Based Hadoop Agent, Sidi Mohamed Ben Abdellah University, Fes.

[7] Iroju Olaronke, Ojerinde Oluwaseun, Big Data in Healthcare: Prospects, Challenges and Resolutions, FTC 2016 – Future Technologies Conference 2016, 6-7 December 2016, San Francisco, United States.

[8] Situated Big Data and Big Data Analytics for Healthcare, Mark Sterling.

[9] IEEE Access special section editorial: Big data analytics for smart and connected health.
[10] Prasan kumar sahoo, Suvendu kumar Mohapatra, Shih Lin WU, Analyzing Healthcare Big data with Prediction for Future Health Condition, supported by Ministry of Science and Technology, Taiwan.
[11] Pankaj goel, Aniruddha Dutta, M.Sam Mannan, Application of big data analytics in process safety and risk management, Texas A&M University, College Station, USA.
[12] Lekha Narra, T.S., Peta Stapleton, "Clinical Data Warehousing a Business Analytics approach for managing health data", in Proceedings of the 8th Australasian Workshop on Health Informatics and Knowledge Management, Sydney, Australia
[13] Fabrizio Marozzo, D.T., Paolo Trunfio, "P2P-MapReduce: Parallel data processing in dynamic Cloud environments". Elsevier - Journal of Computer and System Sciences, 2011: p. 1382 - 1402.
[14] Hui He, Z.D., Weizhe Zhang, Allen Chen, "Optimization strategy ofHadoop small file storage for big data in healthcare". Springer Science Business Media New York - The Journal of Supercomputing, 2015.
[15] Mike P. Papazoglou, W.-J.v.d.H., "Service oriented architectures: approaches, technologies and research issues". The VLDB Journal, July 2007. 16: p. 389-415.
[16] Ralph Mietzner, T.U., Robert Titze, Frank Leymann, "Combining Different Multi-tenancy Patterns in Service-Oriented Applications".Enterprise Distributed Object Computing Conference, 2009. EDOC '09. IEEE International Sept. 2009 p. 131 – 140.
[17] Alex Homer, G.M., and Masashi Narumoto, "Developing big data solutions on Microsoft Azure HDInsight" R. Corbisier, Editor. June 2014.
[18] Amogh Pramod Kulkarni, M.K., "Survey on Hadoop and Introduction to YARN" International Journal of Emerging Technology and Advanced Engineering, 2015. 4(5): p. 303.
[19] Hyuck Han , Y.C.L., Seungmi Choi , Heon Y. Yeom , Albert Y. Zomaya, "Cloud-Aware Processing of MapReduce-Based OLAP Applications" Proceedings of the Eleventh Australasian Symposium on Parallel and Distributed Computing, 2013. 140.
[20] Linlin Dinga, G.W., Junchang Xina "ComMapReduce: An improvement of MapReduce with lightweight communication mechanisms." Elsevier - Data & Knowledge Engineering, 2013. 88: p. 224–247.