Health Care Analysis Using Hadoop

Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, Maya Sanghati

ABSTRACT: The Electronic Medical Records (EMRs) are the primary sources to study the enhancement of health and medical care. The rapid development in science and medical technology has produced various methods to detect, verify, prevent and treat diseases. This has led to the generation of big health-care data and difficulties in processing and managing data. To capture all the information about a patient and to get a more detailed and complete view for insight into care coordination and management decisions big data technologies can be used. A more detailed and complete picture about patients and populations can be identified along with patients at risk before any health issue arises. Optimal strategies to commercialize treatments and the next generation of health care treatments can be identified and developed by it.

KEYWORDS: Electronic Medical Records (EMR), Big data, data mining, Hadoop.

I. Introduction

Historically, large amount of data, driven by record keeping, compliance and regulatory requirements and patient care has been generated by the healthcare industry. The current trend is toward rapid digitization of these large amounts of data, while most data is stored in hard copy form. The big data has the potential to improve the quality of healthcare and on the other hand to reduce the costs. It assures to support wide range of medical and healthcare functions, disease surveillance and population health management. Health care needs to be modernized with the new era of big data, this includes the health care data to be properly analysed so that we can deduce in which group or regions or age or gender diseases attack the most. Distributed processing, Hadoop can be used for the computation of this gigantic size of analytics. Map Reduce is a popular paradigm in computing for large-scale data processing in cloud computing. However, the slot-based Map Reduce system can suffer from poor performance due to its unoptimized resource allocation. To solve this problem the resource allocation is optimized by the framework in this paper. Many times slots can be severely under-utilized due to the static pre-configuration of distinct map slots and reduce slots which are not fungible.The reason of this is map slots might be fully utilized on the other hand reduce slots may remain empty, and vice-versa. To overcome this problem we 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. Getting the health care analysis in various forms are multipurpose beneficial outputs which will be provided by the framework. Thus with a view of future use this concept of analytics should be implemented. Analysis of healthcare big data in this way is not yet carried out by any other systems. Companies are realizing that “data is king,” but the question is how to analyze it? The amount of healthcare data that is captured is measured in terabytes and peta bytes. The question “How do we analyze that amount of data?” remains even with NoSQL data stores. Analysis of different patient attributes will be done by the system. This will be achieved with the help of Hadoop Framework using which we can do a very fast analysis for big data. If this system is used by Govt. of India it will be a very good impact.

II. LITERATURE SURVEY

In the following we examine some of the reasons why we need the hadoop technology and his system for the efficient use and management of health care records. In this paper [1] Large amount of data driven by record keeping, compliance & regulatory requirements and patient care is generated by healthcare industries, historically. The current trend is toward rapid digitization of these large amounts of data, though most of the data is stored in hard copy form. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery on the other hand reducing the cost, these huge amounts of data (called as “Big data”) hold the promise of supporting various medical and healthcare functions, including among others clinical decision support, disease surveillance, and management of population health. Reports say in 2011, data from the U.S. healthcare system alone reached, 150 exabytes. Big data for U.S. healthcare will soon reach the zettabyte (1021 gigabytes) scale and , not long after, the yottabyte (1024 gigabytes), at this growth rate. Kaiser Permanente, the California-based health network, which has more than 9 million members, is believed to have between 26.5 and 44 petabytes of potentially rich data from EHRs, including images and annotations. By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficulty (or impossible) to manage with traditional software and/or hardware; nor they can be easily managed with traditional or common data management methods and tools. Not only because of the diversity of data types and the speed at which it must be managed but also because of its volume big data is overwhelming. In this paper [2] A lot of challenges in terms of data transfer, storage, computation and analysis has been brought by the exponential evolution of data in health care. Ample patient information and historical data, which enclose rich and significant insights that can be exposed using advanced tools and techniques

- *Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, Maya Sanghati*
- prashantsdhotre@gmail.com
- shimpisayali.s@gmail.com
- poosuryawanshi@gmail.com
- mayasanghati@gmail.com
- *Department of Computer Engineering, SITS, Narhe.*

as well as latest machine learning algorithms for healthcare usage and applications. Though, new big data analytics framework is required for the size and rapidity of such great dimensional data. To show the impact of big data this paper introduces the thought of data in healthcare and the results of various surveys. Some case studies of big data analytics in healthcare are presented. The term “Big data” become popular in last few years, as it represents the hard work of researchers to achieve business intelligence by processing tremendously large amount of data. For typical dataset software tools, it is very difficult to collect, store, manage and analyse. Of course big data is too large to load into memory and store on a hard-drive and fit in a standard database. In this paper [3] Now-a-days, For large-scale data processing in clusters and data centers MapReduce has become a popular high performance computing paradigm. Hadoop, an open source implementation of MapReduce, has been deployed in large clusters containing thousands of machines by companies such as Yahoo! and Facebook to support batch processing for large jobs submitted by multiple users (i.e., MapReduce workloads). In this paper [4] Functionality to a Consultant Physician in Geriatric Medicine in terms of storing and analysing high quality clinical patient data for the purpose of more informed and accurate decision making is provided by the Patient Data Analysis Information System (PDA-IS). To support the Consultant Physician in improving the quality of healthcare delivery is the system’s general aim.

III. FIGURES AND GRAPHS

1. K-means Clustering:

K-means clustering is a method of vector quantization, originally from signal processing. This is popular for cluster analysis in data mining. K-means clustering partitions observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. Given a set of observations(x1, x2, ..., x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S={S_1,S_2, ...S_k} so as to minimize the within-clustser sum of squares (WCSS):

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where, μ_i is the mean of points in S_i.

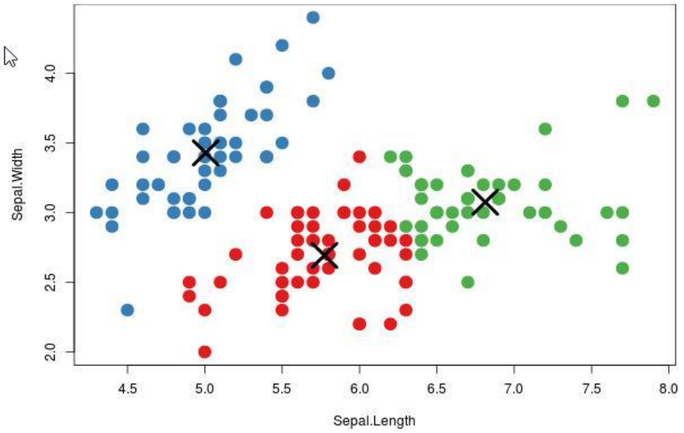

Fig. 1 [5] *K-means Clustering*

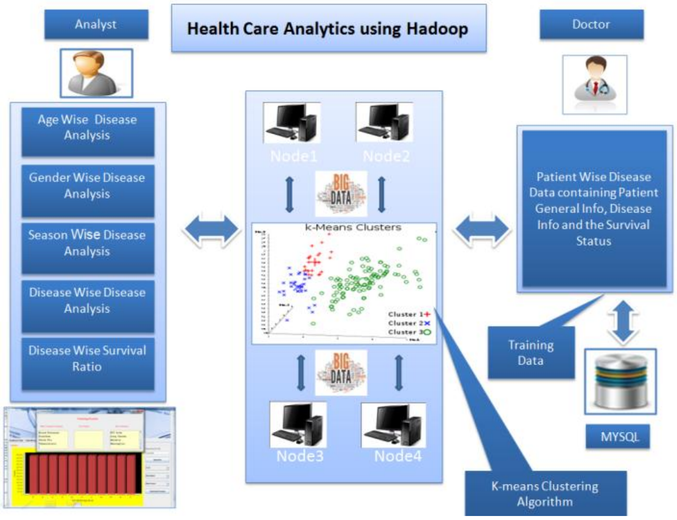
Analyst
Health Care Analytics using Hadoop
Doctor
Age Wise Disease Analysis
Gender Wise Disease Analysis
Season Wise Disease Analysis
Disease Wise Disease Analysis
Disease Wise Survival Ratio
Node1
Node2
BIG DATA
k-Means Clusters
Cluster 1
Cluster 2
Cluster 3
Node3
Node4
Patient Wise Disease Data containing Patient General Info, Disease Info and the Survival Status
Training Data
MYSQL
K-means Clustering Algorithm

Fig. 2 *Architecture Diagram for Health-care Analysis*

The medical records are gathered from various organizations and doctors. The data gathered is send for pre-processing from which patient’s wise disease, patient’s general information, information about the disease and the survival status is extracted. This data is known as training data. The pre-processed data is given for data mining, where for mining the data mining technique is used. After pre-processing is done data is send to perform analysis clustering. To show clearer picture of the analysis algorithm is applied to the resultant data once the analysis is done. Then the statistics can be represented in various forms and groups.

IV.EXPECTED RESULTS

Following are some grouping categories based on which grouping can be done.

1. **Disease Prediction**
   To provide cumulative information disease and their possible symptoms data is grouped together. By this system will predict the disease according to the symptoms provided by the user.
2. **Disease Wise Treatment Determination**
   A suitable treatment can be given to the patient according to the disease predicted by symptom.

3. **Disease Wise Survival Ratio**
   To make it possible to take decision for the patient and his family about the treatment, system can show the disease wise survival ratio.
4. **Gender Wise Disease Statistics**
   Though many diseases affect both men and women, there are some diseases that occur in women at a higher frequency.
5. **Region Wise Disease Count**
   Some diseases may get spread very fast and can influence the whole region. So with the help of these statistics people leaving in a particular region can take precaution to avoid the spread of that disease and hence reduce death ratio.
6. **Identifying and Developing the next Generation of Health Care Treatments.**
   By studying the different statistics, diseases can be prevented by taking certain precautions wither for a particular age group or a specific gender or for a particular region people. This may lead to the next generation of the Health Care Treatments.

V. CONCLUSION

In this paper we are going to give a clear view of the medical record, the pattern extracted from the EMRs, and the results generated by extracting the patterns. These patterns are represented in the graphical and tabular list format. To provide cumulative information, disease and their possible symptoms data is grouped together and analysed. The system will predict the disease for the symptoms which is provided to the system by us after the analysis is done over it. To show the clearer picture of the analysis algorithm can be applied to the resultant and the grouping can be done. Some grouping categories based on which grouping can be done are Age, Gender, Disease, Region, Survival Status, etc.

VI. FUTURE SCOPE

More refined techniques for data pre-processing, in order to extract required information efficiently comprises future work on this study. In order to generate substitute methods and various clustering techniques which can be investigated for further enhanced analysis, other algorithms for pattern detection shall also be incorporated in the system.

VII. REFRENCES

[1] Shanjiang Tang, Bu-Sung Lee, Bingsheng He, “DynamicMR: A Dynamic Slot Allocation Optimized Framework for MApReduced Clusters:, IEEE Transactions, 2013.

[2] Wullianallur Raghupathi and Viju Raghupathi, “Big data analytics in healthcare: promise and potential”, Health Information Science and Systems 2014.

[3] Divyakant Agrawal, UC Santa Barbara, Philip Bemstein, Microsoft Elisa Bertino, Purdue Univ. “Big data White pdf”, from Nov 2011 to Feb 2012.

[4] International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 7, July 2014).

[5] W. Hersh, “Health-care hit or miss?” Nature, vol. 470, pp.327-329, Feb. 2011.

[6] M.A. Musen and J.H.Bemmel, Handbook of Medical Informatics, HOuten: Bohn Stafleu Van Loghum, 1999.

[7] E.F. Codd, “A relationalmodel of data for large shared data banks” Column. ACM, vol.13(6), pp. 377-387, 1970.

[8] NoSQLDatabases,Available: http://www.nosql-database.org/

[9] 10gen. MonogDB, http://www.mongodb.org/

[10] https://en.wikipedia.org/wiki/K-means_clustering

[11] National Health Insurance Research Database, Available: http://nhird.nhri.org.tw/en/index.htm

[12] National Health Insurance Administration, Available: http://www.nhi.gov.tw/english/index.aspx

[13] P.A.Bemstein, “Future directions in DBMS research-the Laguna Beach Participants” ACM SIGMOD Record, vol. 18(1), pp. 17-26, 1989.

[14] A. Silberschatz and S.Zdonik, “Strategic directions in database systems-breaking out of the box”ACM Comput. Surv, vol. 28(4), pp.764-778, Dec. 1996.

[15] G. DeCandia, “Dynamo: amazon’s highly available key-value store” ACM SIGOPS, vol.41(6), pp. 205-220, Dec. 2007.

[16] F. Chang, “Big table: a distributed storage system for structured data” ACM T.Comput. Syst., vol. 26, no. 2, art. 4, 2006.