International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 07 | Oct-2015 www.irjet.net p-ISSN: 2395-0072

Healthcare Data Analysis using Hadoop

Mukesh Borana [1], Manish Giri [2], Sarang kamble [3], Kiran Deshpande [4], Shubhangi Edake [5]

[1234]Student, Computer Engineering Department, MMIT, Maharashtra, India
[5]Professor, Computer Engineering Department, MMIT, Maharashtra, India

---

Abstract - ***In this paper we 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. This beneficial outputs which include: getting the health care analysis in various forms. Thus this concept of analytics should be implemented with a view of future use. Beyond improving profits and cutting down on wasted overhead, Big Data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. With the world's population increasing and everyone living longer, models of treatment delivery are rapidly changing, and many of the decisions behind those changes are being driven by data. The drive now is to understand as much about a patient as possible, as early in their life as possible – hopefully picking up warning signs of serious illness at an early enough stage that treatment is far more simple (and less expensive) than if it had not been spotted until later.***

Key Words : *Healthcare, Analysis, Clustering, Hadoop, Big data.*

1. Introduction

In traditional hadoop system, the master assign equal task to all node. where the This technique get fail in heterogeneous environment. where performance of each and every node consider Differently. To avoid this scenario we will consider advance hadoop big data framework. The data explosion i.e. generating large amount of data. And it is very difficult to mange, Retrieve and processing by using traditional base system. This healthcare organization has created by keeping record, and regulatory requirement. This potential will help to improve quality of life. Hadoop consist of basically two Factors ,

1) Map Reduce

2) HDFS (hadoop distributed file system) .

Hadoop is platform which are in distributed manner and deployed in clustering format. And cluster should be homogeneous. This gigantic size of analytics will need large computation which can be done with help of distributed processing, Hadoop. MapReduce, a popular computing paradigm for large-scale data processing in cloud computing. disease and their possible symptoms are group together and send it as input to system which generate cumulative information. After analysis done ,if we provide symptoms then system will generate name of disease. Algorithm will create clear picture of output in graphical format. Age, Gender, Disease, Region, Survival Status, Insurance are some grouping categories based on which analysis and grouping can be done. This will be achieved with the help of Hadoop Framework with the help of which we can do a very fast analysis for big data. It will be a very good impact if the system used by Govt. of India.

This framework consist of two function namely map() and reduce(), each having different parameters. Map function contain two parameters i.e. key and value. By default this framework assigns value 1 to all keys. Hadoop uses a specialized scheduling mechanism for allocating task to every node. Scheduling is an important aspect of Hadoop which ensures fair task allocation and load balancing. In heterogeneous clusters the performance of every node differs from all other nodes. To max imize the performance of such clusters and for better resource utilization, the task scheduling should be adaptive.

In hadoop data will not store on single cluster but it will save on number of clusters. so data Will be proceed in parallel manner to achieve performance. Hadoop is trying to keep backup of data. numbers of times data will get vanished, to avoid this group of clusters will be generated.

2. LITERATURE SURVEY

To enhance the processing of conventional healthcare system, we have a proposed a series of Big Data health Care System by using Hadoop. There are many techniques proposed in order to efficiently process large volume of medical record which has explained below:

1] Aditi Bansal and Priyanka Ghare proposed "Healthcare Data Analysis using Dynamic Slot Allocation in Hadoop ".

In this paper HealthCare System is analysis using Hadoop using Dynamic Hadoop Slot Allocation (DHSA) method. This paper proposed a framewook which focus on improving the performance of MapReduce workloads and maintain the system. DHSA will focuses on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically.

2] Wullianallur Raghupathi and Viju Raghupathi has proposed "Big data analytics in healthcare: promise and Potential "

In this paper author proposed the potential and promise of big data analytics in healthcare. The paper provides a broad overview of big data analytics for healthcare researchers and practitioners. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

3. PROPOSED SYSTEM

The conceptual framework of big data analysis project in health care is similar to that of a traditional health analytics project. The main difference between both lies in how processing is executed. In a regular health analysis project, the analysis can be performed on a stand-alone system, such as a desktop or laptop.

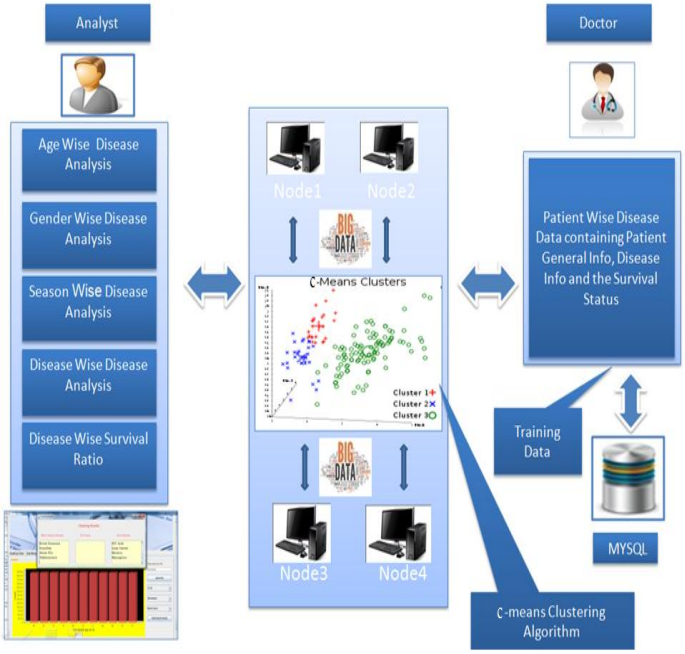

Figure. System Architecture

The concept of distributed processing has existed for decades. Because big data is by definition large, processing is broken down and executed across multiple nodes. What is relatively new is its use in analyzing very large data sets as healthcare providers start to tap into their large data repositories to gain insight for making better-informed health-related decisions. Furthermore, open source platforms such as Hadoop and MapReduce have encouraged the application of big data analysis in healthcare.

All the algorithms and models used for analysis in traditional and the big data healthcare system are similar, Only the user interfaces of traditional analysis tools and those used for big data are entirely different; traditional health analysis tools have become very user friendly and transparent. On the other hand, Big data analysis tools are complex, programming intensive. They have emerged as open-source development tools and platforms, and therefore they are less supportive and user-friendly. As the complexity begins with the data itself.

Big data in healthcare can come from electronic health records, clinical decision support systems, government sources, laboratories, pharmacies, insurance companies often in multiple formats (flat files,

.csv, relational tables, ASCII/text, etc. For the purpose of big data analysis, this data has to be processed or transformed.

The HealthCare System is overwhelming not only because of massive volumes but also diversity of data types and the speed at which it managed.

As shown in the architecture that the Analyst will analysis the massive data and get the desire result in chart or graphical format. The system consists of various users namely Adminstrator, Doctor, Analyst/Researcher. The Adminstrator will be responsible to add/remove the User. The Doctor jobs is to add the disease and their symptons in database with his unique hospital id which after getting authenticated by the admin gets access for entering the medical records. The role of the analyst/researcher is to select the parameters for the analysis. The parameters can be in the form of dates, gender or Age . Once Analyst selects the parameter, he can select the representation method in which the desire output is display. We are sure that this type of health care analysis will surely help the Admin to keep track of their User and Records as well as the analyst will be able to do the analysis in a more organised way. Big data analytics and applications in healthcare are at a nascent stage of development, but advances in platforms and tools can accelerate their maturing process to a greater extent. System architecture consist of various parts described as follows:

We are implementing this project by using Java Technology and MySQL database and Ubuntu server with Hadoop installed on it. The Disease and their possible symptoms are grouped together and analyzed using MapReduce in Hadoop. After the analysis, the result are applied to algorithm which then show the clearer picture of the analysis result.

4. ALGORITHMS

In this paper we are using Fuzzy C-means Clustering Algorithm and ID3(Iterative Dichotomiser 3) Classification Algorithm.

Fuzzy C-means Clustering Algorithm

In this paper to generate the output in the form of clustering we are using Fuzzy C-means Algorithm.This algorithm is centroid-based clustering algorithm in which dataset is group into n-cluster based on certain degree of belonging.

Let X = {x1, x2, x3 ..., xn} be the set of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

Randomly select 'c' cluster centers.

Calculate the fuzzy membership 'μij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

Repeat step 2) and 3) until the minimum j value is achieved or $|| U^{(k+1)} - U^{(k)}|| < \beta$.

where,

k is the iteration step.

β is the termination criterion between [0, 1].

U = (μij)n*c is the fuzzy membership matrix.

j is the objective function.

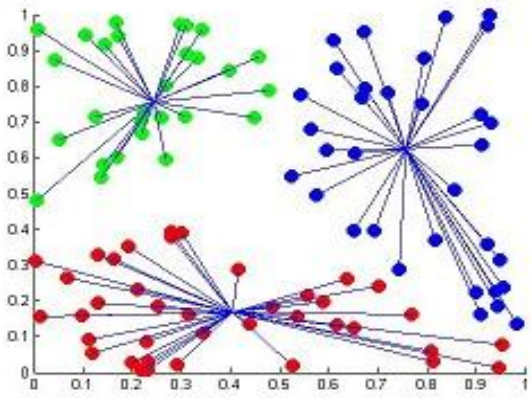

Figure. C-means Clustering

Iterative Dichotomiser 3 (ID3) Classification Algorithm :

ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node.

STEPS :

1) Establish Classification Attribute (in Table R)

2) Compute Classification Entropy.

3) For each attribute in R, calculate Information Gain using classification attribute.

4) Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).

5) Remove Node Attribute, creating reduced table R_S.

6) Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table.

ENTROPY:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

INFORMATION GAIN:

For Set S , Attribute A
Where S is split into subsets based on values of A
$\subset_S^A$ = Subset A of S

$$I_E = Entropy\ , \mathrm{p}(\subset_S^A) = \frac{\mathrm{size}(\subset_S^A)}{\mathrm{size}(S)}$$

$$I_G(S, A) = I_E(S) - \sum^{n} (\mathrm{p}(\subset_S^{A_n}) * I_E(\subset_S^{A_n}))$$

5. CONCLUSION AND FUTURE WORK

The capability of big data will transform the way today's healthcare providers operate the sophisticated technologies to get knowledge from clinical records and make good decisions. In the nearby future we will see implementation of big data analytics in health care industry.Big data provides security and privacy. This paper proposes a framework which is aiming that it will improve the performance of MapReduce workloads and at the same time will maintain the fairness.

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deepest gratitude to all those who have supported us and helped us to make this paper a reality. We are highly indebted to Prof. Shubhangi Edake for her guidance and constant supervision and also for her support in completing the paper. In conclusion, we would like to thank our parents and colleagues for their kind co-operation and investing their time, and willingly helping us out with their abilities to make this paper a reality.

REFERENCES

[1] Shanjiang Tang, Bu-Sung Lee, Bingsheng He, "DynamicMR: A Dynamic Slot Allocation Optimization Framework for MapReduce Clusters ", IEEE Transactions,Vol 2 No.3 Sep 2013,pp.333-345.

[2] Wullianallur Raghupathi and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems ,pp.1-10, 2014.

[3] Aditi Bansal, Balaji Bodkhe, Priyanka Ghare, Seema Dhikale, Ankita Deshpande, "Healthcare Data Analysis using Dynamic Slot Allocation in Hadoop" International Journal of Recent Technology and Engineering , Vol-3 Issue-5, November 2014, pp. 15-18.

[4] Divyakant Agrawal, UC Santa Barbara, Philip Bernstein, Microsoft Elisa Bertino, Purdue Univ. "Big Data White pdf", from Nov 2011 to Feb-2012.