Diabetes Prediction Data Model using Big Data Technologies

[1]Naveen Raja S. R [2]Aswin Kumar. V [3]Richard Paul. V [4]Dr. D. Doreen Hephzibah Miriam [5]Ms. Shobana.G
[1,2,3]Student [4]Director [5]Assistant Professor
[1,2,3,4,5]Department of Information Technology Engineering
[1,2,3,4,5]Loyola-ICAM College of Engineering and Technology

Abstract

The idea that the purely phenomenological knowledge that we can extract by analyzing large amounts of data can be useful in healthcare seems to contradict the desire of VPH researchers to build detailed mechanistic models for the need of patients. But in practice no model is ever entirely phenomenological or entirely mechanistic. In Today's world many different people are in need of healthcare at the finest and quickest way possible, to make this possible our application does a critical analysis of the people's profile and comes up with the nearest hospital that could treat the patient the best way possible in terms of Finance, Disease that the patient is diagnosed with. Using this application, the patient enters his basic Information with the added details of the symptoms, allergies or specifically stating to which disease he needs treatment in order to do this, big data technologies must be further developed to cope with some specific requirements that emerge from this application. In turn, the Application does a big data analysis using Hadoop and reduces the result to the best possible solution consisting of hospitals situated the nearest to the patient that meets all his requirements inclusive of his financial status.

Keyword- Big Data, VPH, Healthcare

I. INTRODUCTION

Enormous information is developing exponentially consistently and there is no indication of its development rate going down anytime. These days, vast information volumes are every day created at a phenomenal rate from heterogeneous sources (e.g., well-being, government, interpersonal organizations, promoting, budgetary). This is because of numerous innovative patterns, including the Internet of Things, the multiplication of the Cloud Computing and the utilization of shrewd gadgets which give gushing information. Effective frameworks and appropriated applications are supporting the foundation and numerous hubs. Versatility, adaptability and execution required in Big Data setting. To remove information from Big Data, different models, programs, virtual products, equipment types and innovations have been planned and proposed. They endeavor to guarantee more exact and dependable outcomes for Big Data applications. Be that as it may, in such condition, it might be tedious and testing to pick among various innovations. There exist numerous Big Data overviews in the writing however the greater part of them tend to center around calculations and methodologies used to process Big Data as opposed to innovations Dissimilar to conventional information, the term Big Data alludes to vast developing informational indexes that incorporate heterogeneous organizations: organized, unstructured and semi-organized information. Enormous Data has a mind - boggling nature that require intense innovations and propelled calculations. Along these lines, the conventional static Business Intelligence devices can never again be productive on account of Big Data applications. Volume, Large volumes of computerized information are created constantly from a great many gadgets and applications. Speed, Data are created and ought to be prepared quickly to extricate valuable data and significant bits of knowledge. Assortment, Big Data are produced from conveyed different sources and in numerous organizations (e.g., recordings, archives, remarks, logs). Substantial informational collections comprise of organized and unstructured information, open or private, neighborhood or far off, shared or secret, finish or deficient. What Is Big Data Analytics? Enormous information alludes to colossal informational indexes that are requests of size bigger (volume); more various, including organized, semi-organized, and unstructured information (assortment); and arriving quicker (speed) than you or your association has needed to manage previously. This surge of information is created by associated gadgets —from PCs and advanced mobile phones to sensors, for example, RFID per users and movement cams. In addition, it is heterogeneous and comes in numerous configurations, including content, record, picture, video, and that's only the tip of the iceberg. The genuine estimation of huge information is in the bits of knowledge it produces when broke down — found examples, inferred meaning, pointers for choices, and at last the capacity to react to the world with more prominent insight.

Enormous information examination is an arrangement of cutting-edge advances intended to work with huge volumes of heterogeneous information. It utilizes complex quantitative techniques, for example, machine learning, neural systems, apply autonomy, computational arithmetic, and computerized reasoning to investigate the information and to find interrelationships and example.

II. System architecture

The proposed system architecture is as follows – It consist of Data sources from various devices, agencies, personal, surveys or official government records. The obtained data is pre-processed before loading it, into the warehouse. Data is cleaned and integrated, null values and duplicate data are removed or transformed to match the content. After pre-processing is done, data is loaded into the data warehouse so it could be stored and retrieved whenever it's necessary. The files and records in the data warehouse imported/moved into Hadoop File system (HDFS). The imported data sets are analyzed using HIVE for complex systems, which gives the user more flexibility and ease to analyze data thereby generating more precise data/information. However, on the other hand for Map Reduce algorithm could also be used to analyze data wherein the accuracy of data analysis lies in the efficiency of the code. Analysis can be done either using HIVE alone or a combination of both HIVE and Map Reduce. Hive is a NoSQL database which allows us to query and join varies tables similar to a traditional database but its efficiency and scalability compared to a traditional database is high. It can process queries faster than a traditional SQL system. It can also be used to obtain histographies, determine mean, median and do other statistical analysis. Join operation provides the possibility of combining varies tables with similar data to obtain a holistic view. Followed by the execution of the queries, the results are stored in HDFS or converted into a report format and given to the data analysts or the concerned authority who require it. Sometimes if only one data set is present it can be directly sent to the Map-reduce program and a result can be produced from it, this method is better when there is only one table under consideration. Map reduce algorithm, which scans and fetches data from the table based on the conditions defined in the code. The final report generated might be in text or can be converted into other forms like table or a graph depending on how we want it at the end.

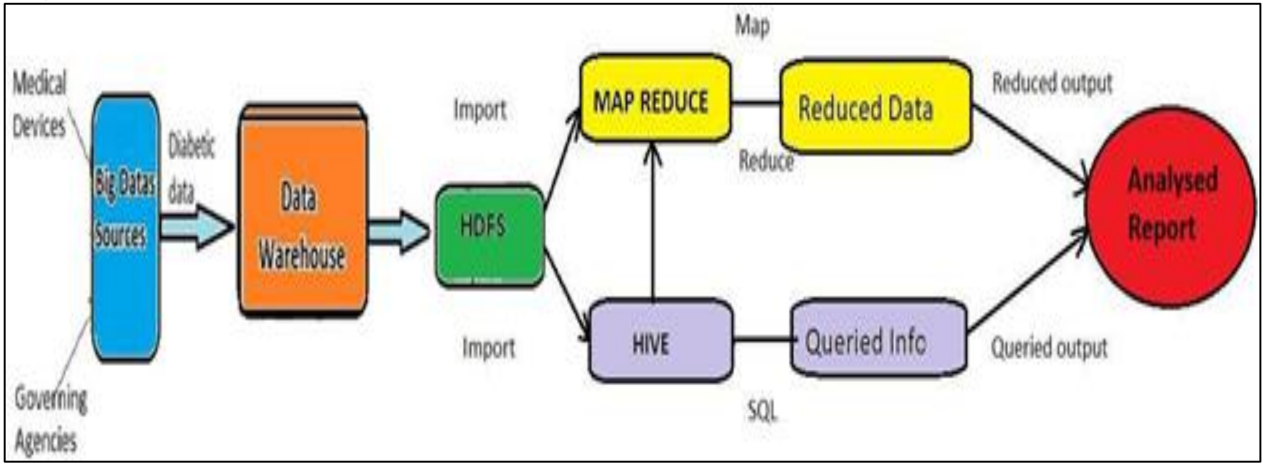

Fig. 1: System Architecture

III. Implementation

The system consists of three main modules:

- Data Sources
- Map Reduce
- HIVE Data

Data sources are obtained in various forms and are from many origins, in this project we have taken an old diabetics data set from a reliable website which provides free datasets for research purposes. The data is cleaned, transformed and integrated so it could be easily queried and analyzed without any discrepancies. If null values are present they are replaced with a constant which is known or arbitrarily assumed. The data is static and already recorded. The data which is considered for the implementation consist of the patient's ID, their race, age range, gender, the doctor who treated them, medicine prescribed and if they took insulin or not. From the various fields it is reduced in way that the utilization of medicine in accordance to region and age group is generated by which the drug industries and doctors get an overview of rate of consumption of drugs and thereby providing an insight of the demand of medicine in a particular region in form of a report. The reduced data after processing it provided as the report. Furthermore, the reduced data can also be processed again to refine the information and narrow down the scope of knowledge gained. Data could be also a streaming data for which HBase could be used. HBase can handle even un - structured data and has its own format. It has row key, column key and timestamp for each column, the timestamp is because the data is temporal in nature. Map-Reduce is a java program which is used to receive a large amount of data, categorize it and reduce it for better understanding. Map reduce in an inbuilt algorithm for Hadoop. The program is made of two parts, the mapper part which scans and gets the values and stores in variables. Then there is the reducer part which reduces the obtained values based on the count and condition given.

The program can be modified based on the data given to us. The data to be reduced is loaded into the HDFS before giving it has input to the code, multiple inputs can be fed at a time. The input can also be an output from another analysis, for instances, it can be processed using pig or hive and then fed into the map-reduce code to obtain any further results. Hive is a NoSQL database which allows us to store large volumes of data in table format in HDFS. The reason for preferring hive is that it allows us to load and process enormous data quickly compared to traditional relational databases. Hive can be utilized to perform histography functions, run statistical algorithms and joint varies tables to form a single entity for better understanding which is not possible in relational databases. Since our dataset consists of only one table it cannot be joined with any other tables but if any similar data sets are present it could be joined together to give a more detailed view. The query is given in the hive console to perform analytical operations on the data set. The obtained result is stored in the HDFS. Similar to Map-Reduce is can also take a result from another operation and process it. The resulting data from the hive is sent to the map - reduce program and reduced.

IV. Related works

As a tremendous measure of data are delivered amid the time it is difficult to administer, process and store them. All of a sudden since its beginning, overall web action will outflank 1 zettabyte (1 billion terabytes) in 2016, as demonstrated by a Cisco look at the paper, having risen fivefold in the past five years. An alternate report assesses that 90% of the world's data was made in the past two years. Not solely are we clicking, informing, going to and taking photos or chronicles more than ever, associations have cottoned onto how data is noteworthy so are securing a consistently expanding number of data. Datasets, for instance, site get to logs and snap data are never again being disposed of – they are being chronicled and mined to deliver productive bits of information. Not at all like conventional information, has the term Big Data alluded to vast developing informational indexes that incorporate heterogeneous configurations: organized, unstructured and semi-organized information. Enormous Data has an intricate nature that requires intense advances and propelled calculations. Along these lines, the conventional static Business Intelligence apparatuses can never again be productive on account of Big Data applications. Volume, Large volumes of computerized information are created persistently from a great many gadgets and applications. Along these lines developing in an exponential way consistently. This is because of the advance and improvement in the developing mechanical patterns, which incorporates Internet of Things (IoT), Artificial Intelligence, Augmented reality, Cloud Computing. Also, the current advancement that is observed to be in the Internet of Things is that it is being incorporated with blockchain innovation to frame another new innovation which is named as Blockchain Internet of Things (BIOT). BIoT can be utilized to track shipments of pharmaceuticals and to make keen urban areas in which associated warming frameworks better controls vitality utilize and associated activity lights better oversee surge hour. And furthermore, blockchain, one of the fundamental advancements for the hot cryptographic money bitcoin, can make IoT gadgets significantly more valuable. It makes computerized record crosswise over hundreds or thousands of PCs, immeasurably lessening the danger of hacking.

V. Drawbacks of existing system

Data analytics systems are not new technologies which came into existence recently, they have been in use for nearly half a decade yet the efficiency and the method of utilization of these systems and technologies have under gone changes every year. This project's or system's idea was inspired from 'A Systematic Review of Type-2 Diabetes by Hadoop/Map-Reduce' [6]. In this they used data analytics and statistical algorithm to find a solution which could prevent the occurrence of Type -2 Diabetes. Our system would scan and provide a report where not only the prevention could solution can be obtained but also understand the drug consumption rate and age wise categorization of diabetes affected people in a specific region which would give a survey like report to the doctors and aid pharmaceutical companies to expand their drug distribution is a specific location there by expanding their business.

VI. Result

Hence, we collect huge information that is created in the Health Industries analyzing and reducing them into meaningful facts and reports that would empower the Pharmaceutical Industries and the doctors to exponentially expand their business and effectively serve patients in an ideal way respectively.

VII. Conclusion

Thus, Big information being the pattern of the period, it is important to implement the necessary algorithms and strategies that could be possibly used to manage and process the data. Subsequently yielding valuable and significant outcomes, which would be the wellspring of rising advancements.

References

[1] A Survey on Big Data Market: Pricing, Trading and Protection Fan Liang; Wei Yu; Dou An; Qingyu Yang; Xinwen Fu; Wei Zhao

[2] Medical big data existence flavors; A review Fadia Shah; Jianping Li; Fazal Rehman Shamil;Mubashir Iqbal 2017 2nd International Conference on Robotics and Automation Engineering (ICRAE)

[3] Big Data Analytics in Industrial IoT Using a Concentric Computing Model Muhammad Habib ur Rehman; Ejaz Ahmed; Ibrar Yaqoob; Ibrahim Abaker Targio Hashem; Muhammad Imran; Shafiq Ahmad

[4] Analysis of Big-Data Based Data Mining EngineXinxin Huang; Shu Gong 2017 13th International Conference on Computational Intelligence and Security (CIS)

[5] Is big data for everyone? The challenges of big data adoption in SMEsS. Shah; C. Bardon Soriano; A. D. Coutroubis 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).

[6] A Systematic Review of Type-2 Diabetes by Hadoop/Map-Reduce. Munaza Ramzan, Farha Ramzan and Sanjeev Thakur; Indian Journal of Science and Technology, Vol 9(32), DOI: 10.17485/ijst/2016/v9i32/100184, August 2016