ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

Performance Analysis of a classified data in Healthcare system

Priyadarshini, Karthikayini

P.G. Student, Department of CSE, New Horizon College Of Engineering, Bangalore, Karnataka, India
Assistant Professor, Department of CSE, New Horizon College Of Engineering, Bangalore, Karnataka ,India

ABSTRACT: The MapReduce programming model simplifies large-scale data processing on commodity cluster by exploiting parallel map tasks and reduces tasks. Although many efforts have been made to improve the performance of MapReduce jobs, they ignore the network traffic generated in the shuffle phase, which plays a critical role in performance enhancement. General approach followed is hash function for portioning the data in the reduce tasks. This does not have great support for size of the data associated with each key. In this project performance of the classified data is analysed for reducing the cost of a MapReduce jobs by the implementation of data partioning scheme. The data partioning scheme is implemented by using updated distributed algorithm which helps in the retrieval of the optimized data.

I. INTRODUCTION

One of the popular frame work used for the computation, processing of the huge data is the MAP REDUCE .It is a simple programming model that can automatically manage parallel execution. Leading domains in real world that deal with big data like Healthcare System, E-commerce and social networking applications are using this framework for implementing machine learning ,cyber-security and bioinformatics. The computation in Map Reduce are divided into two phases, map phase and reduce phase, these two phases are performed in different map tasks and the reduce tasks. In the map phase, map tasks are launched in parallel to convert the original input splits into intermediate data in a form of key/value pairs. These key/value pairs are stored on local machine and organized into multiple data partitions, one per reduce task. In the reduce phase, each reduce task fetches its own share of data partitions from all map tasks to generate the final result. There is a shuffle step between map and reduce phase. In this step, the data produced by the map phase are ordered, partitioned and transferred to the appropriate machines executing the reduce phase. The resulting network traffic pattern from all map tasks to all reduce tasks can cause a great volume of network traffic, imposing a serious constraint on the efficiency of data analytic applications For example, with tens of thousands of machines, data shuffling accounts for 58.6 percent of the cross-pod traffic and amounts to over 200 petabytes in total in the analysis of SCOPE jobs.

The healthcare industry has generated large amount of data generated from record keeping, compliance and patient related data. In today's digital world, it is mandatory that these data should be digitized. To improve the quality of healthcare by minimizing the costs, it's necessary that large volume of data generated should be analyzed effectively to answer new challenges. Similarly government also generates petabytes of data every day. It requires a technology that helps to perform a real time analysis on the enormous data set. This will help the government to provide value added services to the citizens. Big data analytics helps in discovering valuable decisions by understanding the data patterns and the relationship between them with the help of machine learning algorithms (1). This paper provides an overview of big data analytics in healthcare and government systems. It describes about big data generated by these systems, data characteristics, security issues in handling big data and how big data analytics helps to gain a meaningful insight on these data set.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

II. METHODOLOGY

The network traffic minimization problem is formulated in this paper. To provide the analysis of the large-scale data, auxiliary graphs with a three-layer structure should be implemented .The given placement of mappers and reducers are applied in the map layer and the reduce layer, respectively. In the aggregation layer, potential aggregator is created in each machine, where aggregation of data is done from all mappers. Single potential aggregator is much more sufficient for each machine, we also use N to denote all potential aggregators. In addition, we create a shadow node for each mapper on its residential machine. In contrast with potential aggregators, each shadow node can receive data only from its corresponding mapper in the same machine. It mimics the process that the generated intermediate results will be delivered to a reduce directly without going through any aggregator. All nodes in the aggregation layers are maintained in set A. Finally, the output data of aggregation layer are sent to the reduce layer. Each edge in the auxiliary graph is associated with a weight, where denotes the machine containing node u in the auxiliary graph.

III. BIG DATA ANALYTICS IN HEALTHCARE AND ITS ADVANTAGES

Health data volume is expected to grow dramatically in the years ahead. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits.
By digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant benefits. Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications; patients at risk for sepsis, MRSA, or other hospital-acquired illness; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible co-morbid conditions. Big data could help reduce waste and inefficiency in the following three areas:

A. Operations in clinic*:*

Comparative effectiveness research to determine more clinically relevant and cost-effective ways to diagnose and treat patients.

B. Research & development:1) predictive modeling for lower attrition and generate a faster ,leaner, and more targeted R & D pipeline in drugs and devices; 2) statistical tools and algorithms are used for improvising clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market; and 3) analysing the patient records and clinical trials for identifying and follow the indications and find adverse effects before products reach the market.

C.Public health: 1) analyzing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response; 2) faster development of more accurately targeted vaccines, e.g., choosing the annual influenza strains; and, 3) turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of populations

- Mobile body network has wireless sensor devices worn by a patient which provide physiological sensing. Data from the mobile body network is transmitted through the emplaced sensors to user interfaces or back-end systems.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

- Emplaced sensor network has devices deployed in the living space to sense environmental quality or conditions, such as temperature, dust, motion, and light. Emplaced sensors maintain connections with mobile body networks as they move through the living space.
- Back-end systems provide online analysis of sensor data and long-term storage of data.
- User interfaces allow any legitimate user of the system to query sensor data.

Wirelessmedical sensor networks certainly improve patient's quality-of-care without disturbing their comfort. However, there exist many potential security threats to the patient sensitive physiological data transmitted over the public channels and stored in the back-end systems.

IV. USE CASES IN BIG DATA

Big data in health-care refers to the patient care data such as physician notes, Lab reports, X-Ray reports, case history, diet regime, list of doctors and nurses in a particular hospital, national health register data, medicine and surgical instruments expiry date identification based on RFID data. Healthcare organizations are depending on big data technology to capture all of these information about a patient to get a more complete view for insight into care coordination and outcomes-based reimbursement models, health management, and patient engagement.

Need for Big Data Analytics in Healthcare: To improvise the quality of healthcare by considering the following:

A. Providing patient centric services: To provide faster relief to the patients by providing evidence based medicine--detecting diseases at the earlier stages based on the clinical data available, minimizing drug doses to avoid side effect and providing efficient medicine based on genetic makeups(1). This helps in reducing readmission rates thereby reducing cost for the patients.

B. Detecting spreading diseases earlier: Predicting the viral diseases earlier before spreading based on the live analysis. This can be identified by analyzing the social logs of the patients suffering from a disease in a particular geo-location. This helps the healthcare professionals to advise the victims by taking necessary preventive measures.

C. Monitoring the hospital's quality: Monitoring whether the hospitals are setup according to the norms setup by Indian medical council. This periodical check-up helps government in taking necessary measures against disqualifying hospitals.

D. Improving the treatment methods: Customized patient treatment---monitoring the effect of medication continuously and based on the analysis dosages of medications can be changed for faster relief. Monitoring patient vital signs to provide proactive care to patients. Making an analysis on the data generated by the patients who already suffered from the same symptoms, helps doctor to provide effective medicines to new patients.

V. BIG DATA ECOSYSTEM FOR HEALTHCARE

It is a complex system that constitutes of components and technologies to handle large scale data processing and analytics on it. It includes getting the data from various sources, store them in PhpMyAdmin, process the data using Hadoop components such as Map-Reduce, perform analysis using PIG and generate Business Intelligence reports such as patient scorecards.

A. Big Data Lifecycle

- **Data Collection:**It involves the collection of data from various sources and storing it in HDFS. Data can be anything such as case history, medical images, social logs, sensor data etc.
- **Data Cleaning:**It involves the process of verifying whether there is any junk data or any data that has missed values. Such data needs to be removed.
- **Data Classification:**It involves the filtering of data based on their structure. For example Medical Big data consists of mostly unstructured data such as hand written physician notes. Structured, semi-structured and unstructured data should be classified in order to perform meaningful analysis.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

- **Data Modelling**: It involves performing analysis on the classified data. For example Government may require the list of malnourished children in a particular location. First it has to classify the data based on the specific location, need to trigger the health report of children, need to identify the children whose families are under poverty line and these data should be processed.
- **Data Delivery:**It involves the generation of report based on the data modelling done. Based on the example after the data is processed it will generate a report based on malnourished children in a particular location. This will help the government to take necessary measures to avoid any further complications. At the all the stages of BDLC (Big Data Lifecycle) it requires data storage, data integrity and data access control.

B.Secured Big Data Architecture:

The security challenges faced by the big data processing in distributed environment are as follows:
1. To provide network level security
2. To provide authentication for users, nodes and applications involved in distributed environment.
3. To provide authentication for users, nodes and applications involved in distributed environment.
Data stored in database can be encrypted. Data can be transmitted between nodes can encrypting it with Attribute based encryption method. This is effective in preventing the data from malicious users. Inbuilt logging can be implemented in JVM of Map-Reduce using differential privacy to store the user identity where the map-reduce job is done. This helps to identify who is responsible for the leakage of sensitive data.

VI. RESULTS AND DISCUSSIONS

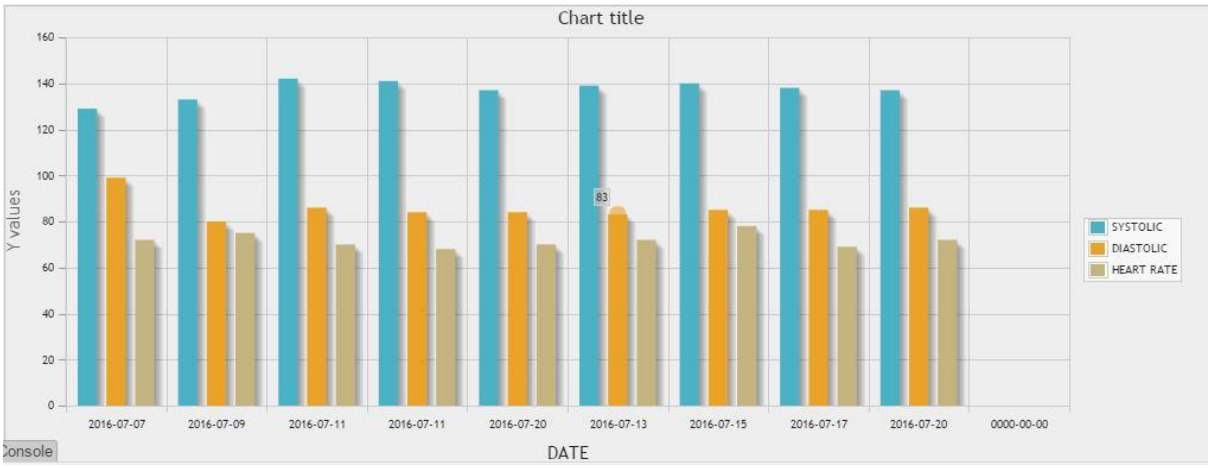


Fig 6.1: Monitoring the health of a patient using Column graph

The above represents the analysis of healthcare data of a particular patient. The Heart rate, systolic and diastolic of a patient is monitored by the Healthcare system for some period of time. This graphical representation makes easy for the physician to come up with the right diagnosis for a patient. By this kind of analysis, quality of human life can be improved.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

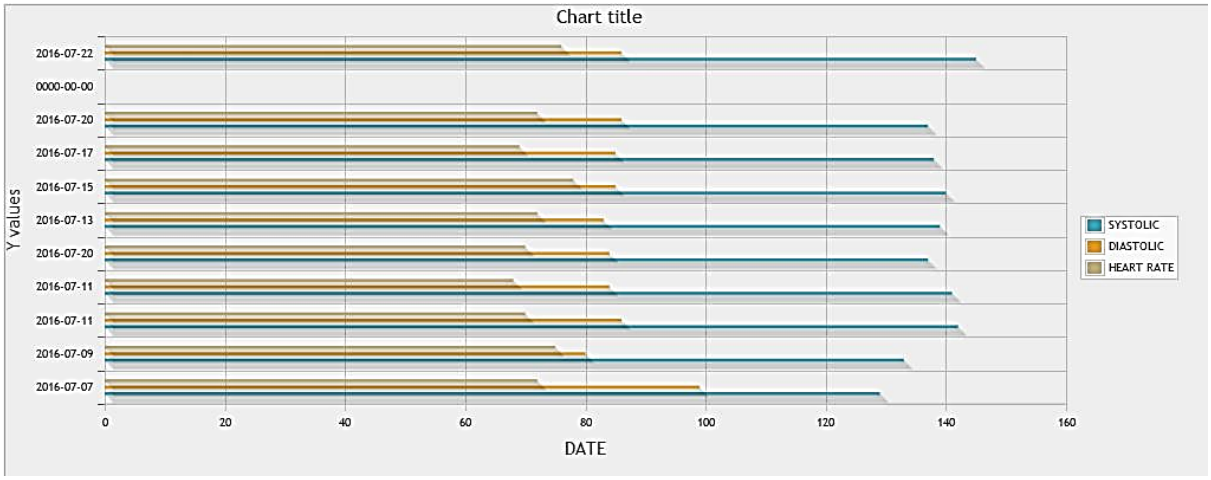


Fig 6.2: Monitoring the health care data using Bar graph

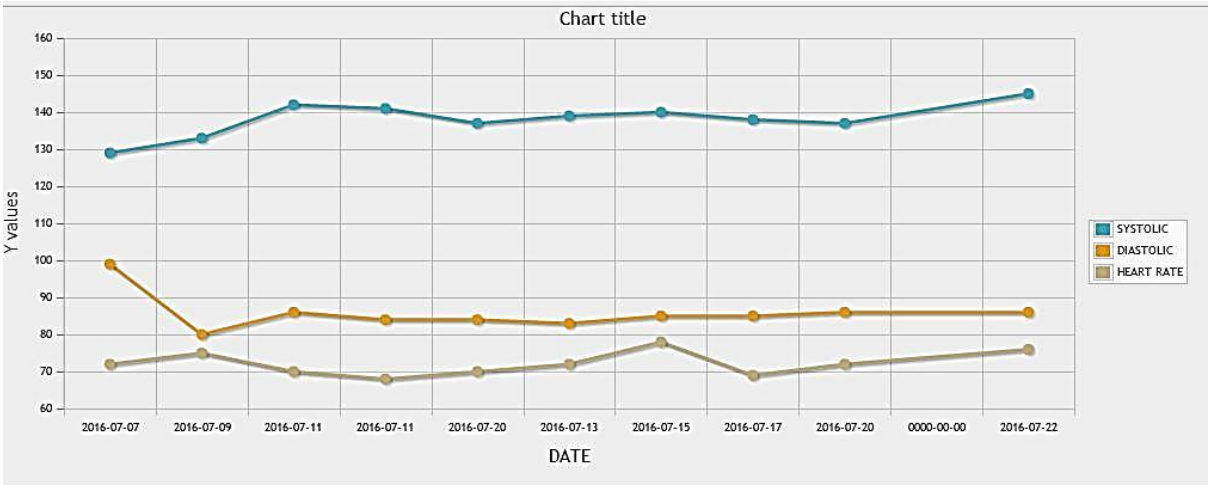


Fig 6.3: Patient data analysed using line graph

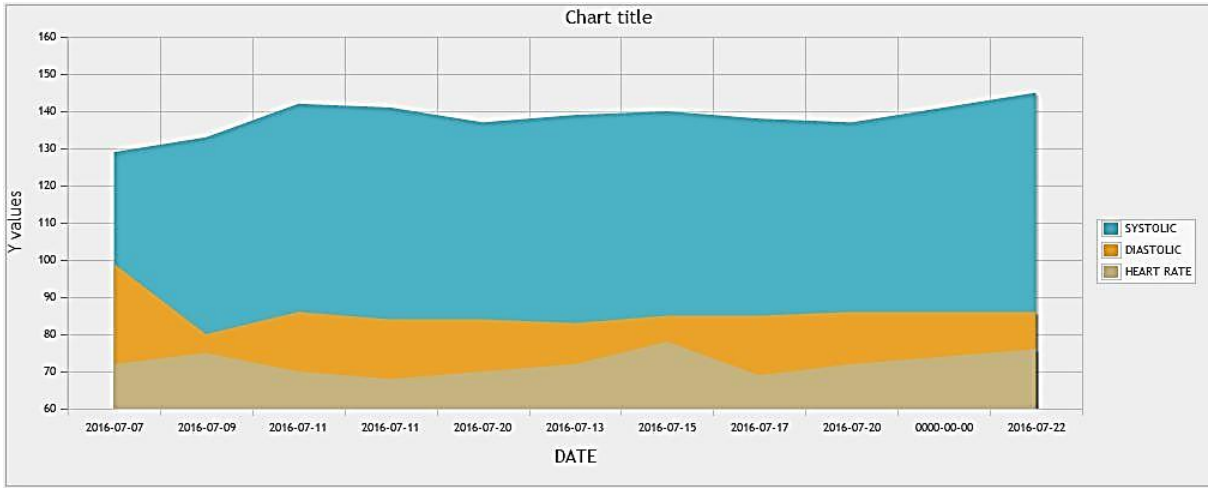
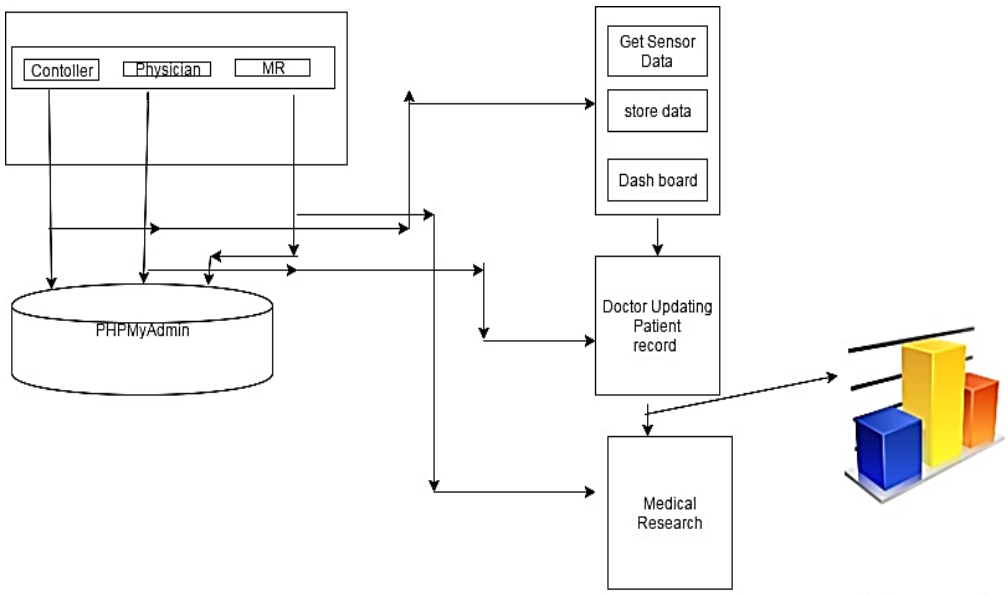


Fig 6.4: Analysing the sensor data of a patient using Area graph

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017



The above system architecture is implemented in the project. The Architecture basically kicks starts the application by the PhpMyAdmin server which is the core part of the application. In this application controller, physician and Medical researcher are constantly interacting with server for inserting, reading the data from server.

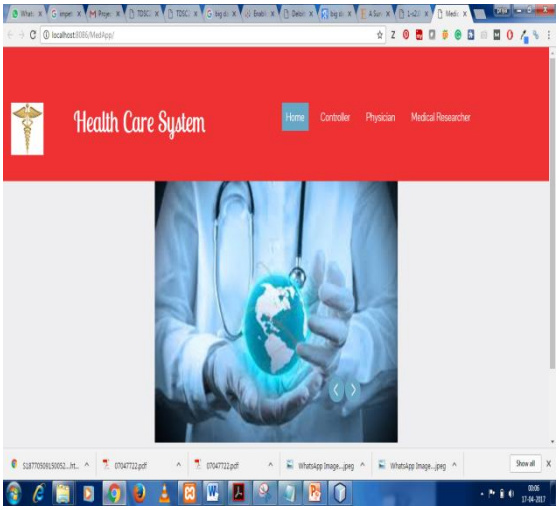


Fig 6.5. Home screen of the application

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 4, Issue 3 , March 2017

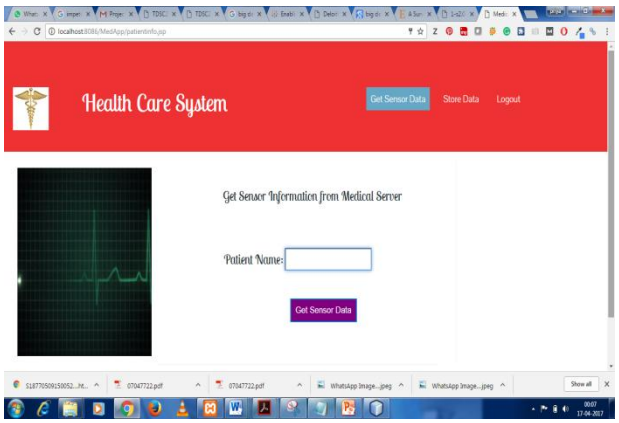


Fig 6.6. Controller can get the sensor data of the patient

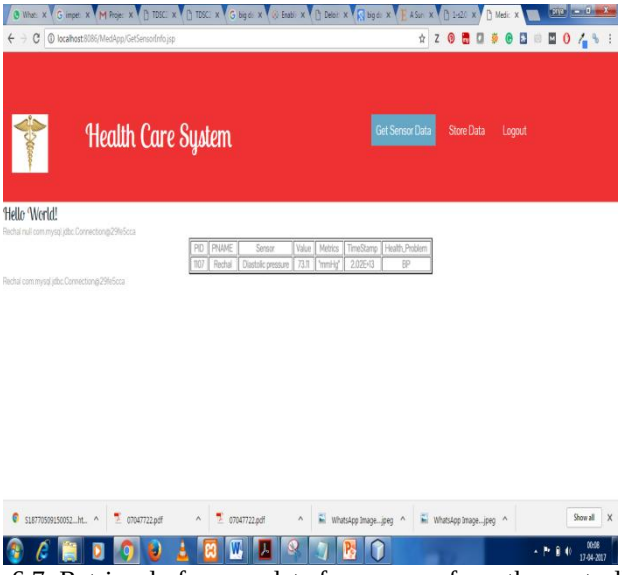


Fig 6.7: Retrieval of sensor data from server from the controller

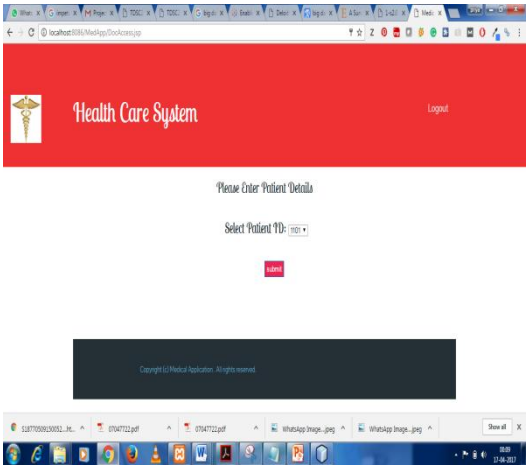


Fig 6.8: Physician can retrieving the patient details from server and updating the patient details

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 4, Issue 3 , March 2017

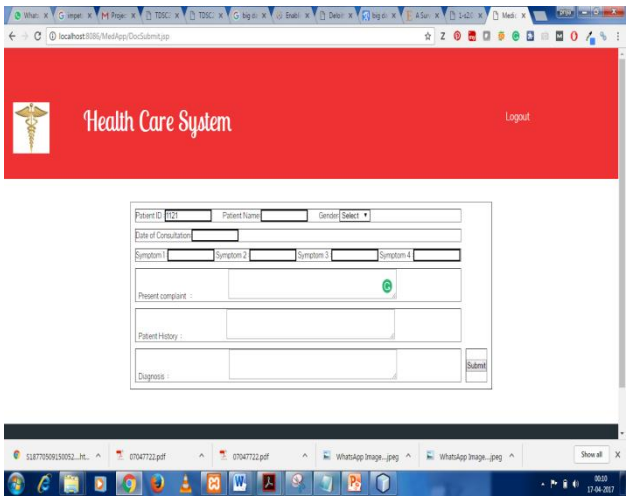


Fig 6.9. Final updating of data of a patient to the server

VII. CONCLUSION

The problem is not the lack of data but the lack of information that can be used to support decision-making, planning and strategy. The entire Health care system can realize benefits from utilizing big data technologies. To successfully identify and implement big data solutions and benefit from the value that big data can bring, government need to devote time, allocate budget and resources to visioning and planning. With the help of Hadoop the goal of effective citizen care management can be achieved by providing an effective data driven services to citizens by predicting their needs based on the analysis of survey conducted among different classes of citizens. Secured BDA can be implemented by using Hadoop in a security enabled Linux environment where access control is provided by the system itself

REFERENCES

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," ACM, vol. 51, no. 1, pp. 107–113,2008.
[2] W. Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang, "Map task scheduling in MapReduce with data locality: Throughput and heavy trafficoptimality," in Proc. IEEE INFOCOM, 2013, pp. 1609–1617.
[3] The_big_data_revolution_in_healthcare.
[4] P. Belsis and G. Pantziou.A k-anonymity privacy-preserving approach in wireless medical monitoring environments. Journal Personal and Ubiquitous Computing, 18(1): 61-74, 2014.
[5] D. Bogdanov, S. Laur, J. Willemson. Sharemind: a Framework for Fast Privacy-Preserving Computations. In Proc. ESORICS'08, pages 192-206, 2008.
[6] S. Chen and S. W. Schlosser, "Map-reduce meets wider varieties of applications," Intel Res., Pittsburgh, PA, USA, Tech.Rep.IRP-TR-08-05, 2008.
[7] H. Lv and H. Tang, "Machine learning methods and their application research," IEEE Int. Symp. Intel Info Process Trusted Computing.(IPTC), pp. 108–110, Oct. 2011.
[8] S. Venkataraman, E. Bodzsar, I. Roy, A. AuYoung, and R. S. Schreiber, "Presto: Distributed machine learning and graph processingwith sparse matrices," in Proc. 8th ACM Eur. Conf. Computer. Syst., 2013, pp. 197–210.
[9] A. Matsunaga, M. Tsugawa, and J. Fortes, "Cloudblast: Combining MapReduce and virtualization on distributed resources for bioinformatics applications," in Proc. IEEE 4th Int. Conf. eScience, 2008, pp. 222–229.
[10] J. Wang, D. Crawl, I. Altintas, K. Tzoumas, and V. Markl, "Comparison of distributed data-parallelization patterns for big data analysis: A bioinformatics case study," in Proc. 4th Int. Workshop Data Intensive Computer. Clouds, 2013, pp. 1–5.
[11] R. Liao, Y. Zhang, J. Guan, and S. Zhou, "Cloud: A MapReduceimplementation of nonnegative matrix factorization for large scalebiological datasets," Genomics, Proteomics Bio information, vol. 12, no. 1, pp. 48–51, 2014.
[12] G. Mackey, S. Sehrish, J. Bent, J. Lopez, S. Habib, and J. Wang, "Introducing map-reduce to high end computing," in Proc. 3rdPetascale Data Storage Workshop, 2008, pp. 1–6.
[13] W. Yu, G. Xu, Z. Chen, and P. Moulema, "A cloud computing based architecture for cyber security situation awareness," in Proc. IEEE Conf. Communications Network Security, 2013, pp. 488–492.
[14] J. Zhang, H. Zhou, R. Chen, X. Fan, Z. Guo, H. Lin, J. Y. Li, W. Lin, J. Zhou, and L. Zhou, "Optimizing data shuffling indata- parallel computation by understanding user-defined functions," in Proc. 9th USENIX Conf. Network System Implementation. (NSDI '12), Berkeley, CA, USA: USENIX Association, 2012, pp. 295–308.