ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 6, October 2016

An Approach for Storing the Medical Data of HIV Patients Using Big Data Technologies

[1] Amol Joglekar, [2] Dr. G. Prasanna Lakshmi [3] Mr. Maunash Jani
[1] Research Scholar , Pacific University, Udaipur [2] Guide and Woman Scientist
[3]Member

Abstract: **-- Medical science has a huge amount of data. This data can be collected from different places like patient database, pathology, blood bank , X-ray clinics etc. day by day this data gets increased and hence we need to store it so that same can be retrieved in future within no time. There are many diseases like HIV where one/doctor needs to keep all information of a patient including hobbies, habits, status and other medical parameters. This paper proposes a new approach to store this big data in a systematic format so that treatment can be started in no time and hence there will be an easy access to all medical records with an ease.**

Keywords: **-- Big Data, Hadoop, HIV, data mining**

I. INTRODUCTION

In today's fast life style we are developing different habits, likings and because of tremendous competitions one gets badly affected by diseases. There are many diseases which are related to psychological or health issues. Also people get less time to visit doctor's clinic and this may lead to taking antibiotics or finding a quick solution using internet. Still, in order to detect the correct disease and diagnosis one has to visit doctor. Doctors can ask different data related to medical parameters, habits like smoking, drinking or any other depending on symptoms. As per the advice given by medical experts some data may be collected from different places like X-ray clinic, pathology, history of patients etc. This data may not be in a particular format. Some of them may be images which needs to be recorded for future use. It may happen that patient's may take second opinion from some other medical expert for which he may travel to some other state or country if patient is not satisfied with earlier medical treatment. He has to carry all his documents, files to other place so that medical expert can examine and provide an advice.

Therefore we have to think on some centralized database repository so that from the server one can easily access his own medical history and reports which will be in read-only mode. Here the big data will come into picture in medical and healthcare industry. This medical industry is another largest industry that produces huge amount of data in today's era. Big data is a term which states the large amount of data which can be in any format. In medical science one can observe variety of data containing images, records, medicines etc. sometimes one can find incomplete data on the patient's information form or due to many reasons data may not be available which leads to uncertainty of data known as veracity. As there is a huge volume of data available one has to really analyze the data using an innovative approach and care.

A. Big data and Issues

The big data cycle is comprised of four steps that result in insight and action. The results of each cycle are often used for the next cycle, creating a continuous insight loop.

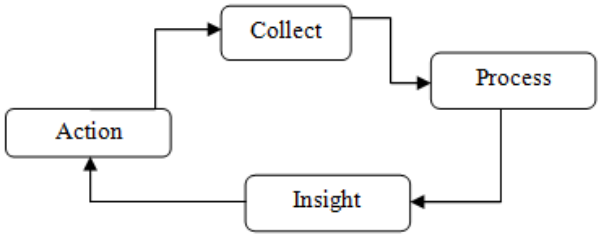

Figure 1: Cycle of Big Data

Data is first collected from devices, log files, third-party data sources. This may include traditional ETL processing but typically on a much larger scale. If data streams generate large bursts of data, it is quickly stored and processed later.

In case of medical science there are big problem which one has to study. There are following factors which one needs to consider before providing any solution or technique.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 6, October 2016

1) Incompleteness

Many times patients do not want to reveal different symptoms or may hide some important information. In case of HIV/AIDS, patient may not want to reveal some essential data which may lead to wrong results or outcome.

2) Privacy

This is a challenging part of any technology. In medical science especially when we want to keep data of HIV, this is a very important issue. There may be several questions which might be very per-sonal but they play a vital role in identifying HIV disease. There has to be some laws or some security mechanism technique to protect users data.

3) Searching/ storing

The large amount of information needs to be kept in a proper database/ data warehouse so that using data mining tools one can dig out information at a proper speed.

4) Analysis

This plays an important role in analyzing data. If data is available in scattered form there may be difficulties in analyzing huge amount of data.

B. HIV and AIDS

"HIV" stands for Human Immunodeficiency Virus. To understand what that means, let's break it down:

- H – Human – This particular virus can only infect human beings.
- I – Immunodeficiency – HIV weakens your immune system by destroying important cells that fight dis-ease and infection. A "deficient" immune system can't protect you.
- V – Virus – A virus can only reproduce itself by taking over a cell in the body of its host.

HIV is a retrovirus that infects cells of immune system such as CD4 cells and it hammers their functions.

In this virus the CD4 cells copies the cells DNA to ensure that it cannot be identified and destroys the immune system. It replicates many times within the cell.

Within 2-4 weeks after HIV infection, many, but not all, people experience flu-like symptoms, often described as the "worst flu ever." This is called "acute retroviral syndrome" (ARS) or "primary HIV infection," and it's the body's natu-ral response to the HIV infection.

II. LITERATURE REVIEW

The proposed research paper is about showing a way to store big data generated from medical science for a disease HIV/AIDS. There are many diseases which have symptoms and they need to be recorded based on their attribute or values. There may be different stages for various diseases. Therefore we need to study all different types of disease, how the data is collected and in what format so that attributes and parameter for symptoms can be taken into consideration. Various papers are being reviewed based on big data and Hadoop so that a new innovative model can be generated to record HIV/AIDS symptoms and treatments.

Remya Panickar[1] had highlighted how big data technology can be useful for many developing countries. Author claims that such technologies can be useful in decision making purpose on which growth of country is depend. She also had mentioned various applications where big data and different tools can be used.

Chanchal Yadav, Shuliang Wang and Manoj Kumar[2] had reviewed different types of algorithms in order to simplify and analyse huge data generated by data mining tools. They stated the general architecture and methodology that can be used to handle large data sets. They compared various algorithms or techniques available for big data. Also they had given brief inputs for problems associated with big data along with various fields where it can be implemented.

Zhendong Ji[3] stated how big data plays an important role in medical science. He claims that medical industry generat-ed huge amount of data and many other industries may use this data for their own purposes. Therefore analysis of such big data is must so that decision making process can be made fast. Customized services can be provided to medical sector and people associated with the field after analysing the data properly.

Dr. K. Rameshkumar [4] developed a model using ARM(Association Rule Mining) to extract valuable information from database. Author has proposed a new algorithm which would take care of missing values for detecting HIV AIDS. With the help of

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 6, October 2016

this proposed algorithm author could able to extract information about CD4 cell counts, RNA levels and treatment given for various patient. The model is lacking of handling data with a very good accuracy.

A. B. Rajesh Kumar, G. V. Ramesh Babu, C. Phani Ramesh, P. Madhura, and M. Padmavathamma[5] proposed a new idea of using centralized database for accessing patient's information in a secured way. There are many diseases like HIV where data is more sensitive and hence it should be kept secret. Therefore security is needed so that misuse of data will not be there and correct data shall be ready for diagnosis. Authors proposed a technique using cloud computing to keep such data in a protected form. They found a decision tree can be used to predict the disease. Expert system was proposed to take some decisions based on patient's information which helps doctors to take decision and provide proper treatment and at the same time data is made secure using cloud computing.

Ronaldo Cristiano Prati, Maria Carloin Monard and Andre C.P.L.F.de Carvalho [6] presented different way to extract knowledge rules from database which is HIV infected. Their main feature was to incorporate exceptions into the representations used by system or machine understandable format. That method had two steps: training of common sense rules and checking exceptions. To implement this there was a need of real world dataset where a viral protease cleaves HIV viral poly protein amino acid remains. A method was to find general rules which were suitable for analysis. It allows more easy way to help people and understand the process.

V.A. Ayma, R.S. Ferreira, P. Happ, D. Oliveria, R. Feitosa, G.Costa, A. Plaza, P.Gamba[9] proposed an approach of integrating Hadoop with data mining with the help of ICP package. They had tested big data and compared results with respect to execution time. They claimed that execution of data with proper classification technique gives best performance.

III. RESEARCH METHODOLOGIES

This research paper proposes a new technique of integrating data mining with the concept of Hadoop technology. In medical science we can collect data from different sources like logs, files, images etc. The data has different form of representation i.e. structured and unstructured and semi-structured. Some data may have missing values or duplicate ones. Therefore we need to clean data and convert data into one format. We can further store data in central repository system. With the help of mining tool we can dig out the required data.

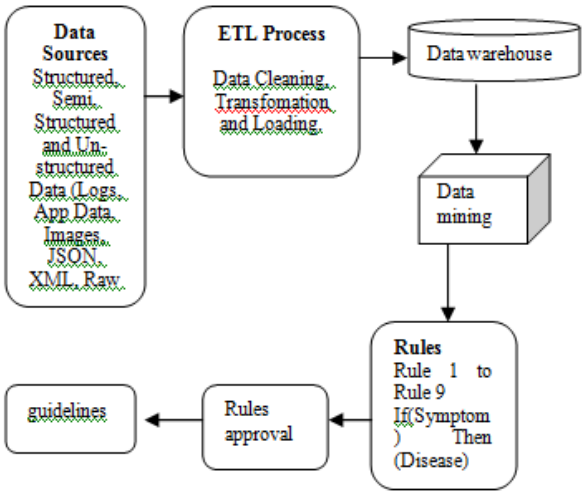

Figure 2: ETL Process Architecture

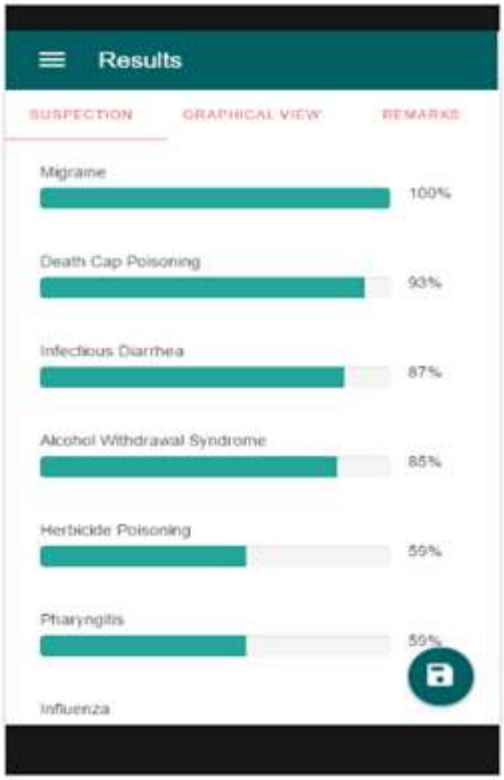

Figure 3: GUI for Exporting Data In JSON format

Proposed architecture using Hadoop technique

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 1, Issue 6, October 2016

The proposed architecture may have following components:

1) Data Collection
2) ETL and data transformations
3) Platform/tool Selection (HIVE)
4) Conceptual Model
5) Analytics Techniques

Queries, Reports, Data Mining (Association Rules, Classification, Fuzzy Logic)

IV RESULTS & DATA INSIGHT

For the purpose of big data analytics, this data has to be pooled from various sources. In the second component the data is in a 'raw' state and needs to be processed and trans-formed. The approach of data warehousing where data from various sources is aggregated and made ready for processing. Via the steps of extract, transform, and load (ETL), data from diverse sources is loaded and cleansed. Depending on whether the data is structured or semi struc-tured or unstructured.

In this next component in the conceptual model, several decisions are made regarding the data input approach, tool selection and analytics models. Finally, on the three typical applications of big data analytics in healthcare/medical science are represented. These include queries, reports and data mining.

The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform). Hadoop can serve the twin roles of data organizer and analytics tool.

Data Generated through the App is in JSON, a semi struc-tured format. For processing this data in the Big Data Analy-sis framework Hadoop we will be uploading the JSON data to Hive, a platform used to develop SQL type scripts to do MapReduce operations on Hadoop.

When using Hive, users typically perform the following functions or workflow steps:

1. Create tables
2. Load data from source system in to Hive table
3. Analyze/Process data by writing Hive QL

JSON format data:

{ "patientid": "1", "results":[
{"disease":"Migraine","percentage":"100"},
{"dis-ease":"Death Cap Poisoning","percentage":"93"},
{"dis-ease":"Infectious Diarrhea","percentage":"87"}
...]}

1. Create Table on Hive

CREATE EXTERNAL TABLE patient results(json_body string)
LOCATION '/json/results.txt ';
(Location is the path of file stored on the file sys-tem on Hadoop)

2. Load data from a file into the patient results table as below:

LOAD DATA LOCAL INPATH '/json/results.txt' OVERWRITE INTO TABLE patient results;

3. Analyze/Process data by writing Hive QL

SELECT p.patientid, r.disease, r.percentage FROM patientresults pr LATERAL VIEW JSON_TUPLE(pr.json_body, 'patientid') p AS pa-tientid, results LATERAL VIEW json_tuple(p.results, 'disease', 'percentage') r AS disease, percentage;

Here,json_tuple: Takes a set of names (keys) and a JSON string, and returns a tuple of values.

Lateral View: First applies the UDTF(User Defined Table) to each row of base table and then joins resulting output rows to the input rows to form a virtual table having the supplied table alias.

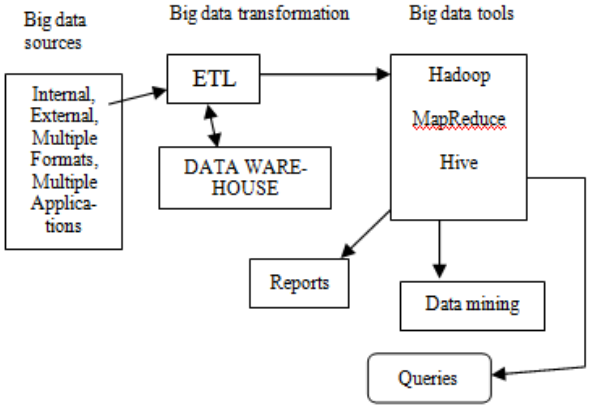

Figure 4: Integration of Hadoop and Big Data Technologies

V CONCLUSION

Big data analytics has the potential to transform the way healthcare and medical providers use

sophisticated technologies to gain insight from their clinical and other patient data repositories and make informed decisions and reports. Big data analytics and applications in medical industry are at a emerging stage of development and the rapid advances in platforms and tools can accelerate their growing process.

REFERENCES

[1] Remya Panickar "Adoption of Big Data Technology for the Development of Developing Countries" Proceedings of National Conference on New Horizons in IT - NCNHIT 2013 ISBN 978-93-82338-79-6.

[2] Chanchal Yadav, Shuliang Wang and Manoj Kumar "Algorithms and approaches to handle large data- A survey" IJCSN International Journal of Computer Science and Net-work, Vol.2 ,Issue 3, 2013. ISSN 2277-5420.

[3] Zhendong Ji "Applications analysis of Big data analysis in medical industry" International Journal of database theory and application. Vol. 8, No. 4 (2015), pp 107-116.ISSN 2005-4270 IJDTA.

[4] Dr. K Rameshkumar "Association Rules Mining from HIV/AIDS patient's case history database with missing val-ues" International Journal on Data Mining and Intelligent Information Technology Applications" Vol.2 , No. 1,pp. 18-24, March 2012.

[5] A. B. Rajesh Kumar, G. V. Ramesh Babu, C. Phani Ra-mesh, P. Madhura, and M. Padmavathamma "Medical Diagnosis Expert System as Service in Cloud" International Journal of Computer and Communication Engineering, Vol. 2, No. 4, July 2013; DOI: 10.7763/IJCCE.2013.V2.211.

[6] A. B. Rajesh Kumar, G. V. Ramesh Babu, C. Phani Ra-mesh, P. Madhura, and M. Padmavathamma "Medical Diagnosis Expert System as Service in Cloud" International Journal of Computer and Communication Engineering, Vol. 2, No. 4, July 2013; DOI: 10.7763/IJCCE.2013.V2.211.

[7] Ronaldo Cristiano Prati, Maria Carolina Monard and Andre C.P.L.F de Carvalho. "Looking for exceptions on knowledge rules induced from HIV cleavage data set." In-ternational Journal Genetics and Molecular Biology, Vol.27 , Issue.4, pp.637-643,2004.

[8] V.A. Ayma, R.S. Ferreira, P. Happ, D. Oliveria, R. Fei-tosa, G.Costa, A. Plaza, P.Gamba "The International Arc-hives of the Photogrammery, Remote Sensing and Spatial Information Science, Volume XL-3/W2, 2015. PIA15+HRIGI15- Joint ISPRS conference 2015,25-27 March, Munich, Germany .

[9] www.avert.com

[10]www.health.com

[11]www.aidsprogramme.ukzn.ac.za

[12] Han J Kamber "Data Mining Concepts and Techniques" Morgan Kaufman Publishers 2006.

[13] Big Data for Dummies, J. Hurwitz, et al., Wiley, 2013.

[14] Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data, Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, McGraw-Hill, 2012.

[15] Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Michael Minelli, Wiley, 2016