Aids Detection System Using Big Data Analytics

PACKIYAM.S
Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga.
Email: packyafrnds143@gmail.com
PREMA.A
Department of Computer Science, Raja Doraisingam Govt. Arts College , Sivaganga.
Email:jeyaprem2010@gmail.com
CHELLADURAI.K
Department of Computer Science, Raja Doraisingam Govt. Arts College , Sivaganga.
Email:kchelladurai05@gmail.com

ABSTRACT

The exponential growth of data over the last decade has introduced a new domain in the field of information technology called Big Data. The developing countries like India with huge population face various problems in the field of healthcare with respect to the operating cost, meeting the needs of the reasonably poor people, access to the hospitals, and research in the field of medicine and especially in the time of spreading epidemics. Here, we contain focused on Acquire Immune Deficiency Syndrome (HIV/AIDS) which seems succeed in many women/Children who come to Government Hospital. HIV is a disease which is not simply known. This involves certain tests to be carried out like ELISA, Ora Quick In-Home HIV Test etc. Most of the people do not come onward for HIV test. So, we attempt to develop gives possibility percentage of AIDS in a person using big analytics and Hadoop Cluster.

KEYWORDS- **Big data Analytics, Hadoop, HIV/AIDS.**

I. INTRODUCTION

Big Data is a volume, velocity and variety, Information asset that demand cost-effective, original forums of in sequence processing for improved insight and decision making. Big data is distinct from big existing, database which uses Hadoop framework for data intensive distributed applications. The detection of DNA and its variation is critical for many fields, as well as clinical and veterinary diagnostics, business and environmental testing, rural researches and science. Disease diagnosis and prediction are based on effective detection of disease environment (e.g. Cancer), communicable organisms (e.g. HIV) and inherent markers Amir H. Payberah Presented [1].

Sandrine Dudoit et al described, DNA analysis from original specimens is a complex process involving multiple chemical compositions as well as multistep reactions. The human genome is the complete set of acid sequence for humans (Homosapients), encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual Mitochondria. DNA is the largest human DNA, DNA number 1, consists of around 220 million base pairs a would be 85 mm long if straightened [2].

II. Related works

Rubbin DB presented, The Big Data and Analytics architecture incorporates many different types of data:

- Operational Data – Data residing in operational systems such as CRM, ERP, warehouse management system is typically very well structured. This data, when gathered, cleaned, and formatted for reporting and analysis purposes, constitutes the bulk of traditional planned data warehouse, marts and OLAP cube.
- COTS Data – Custom off-the-shelf (COTS) software is frequently used to support standard business process that does not differentiate the business from other similar businesses. COTS applications often include analytical packages that function as pre-engineered data marts. COTS analytical data, transformed from operational data, can also are incorporated into the data warehouse to support analysis across business process.
- Content – Documents, videos, presentations are typically managed by a content management system. These forms of in order can be linked to other forms of data to support search, analysis and discovery across data types.

Most companies already use analytics in the form of reports and dashboards to help run business. This is mostly based on well structured data from operational systems that conform to pre-determined relationships. Big Data on the other hand doesn't follow this structured model. The streams are all dissimilar and it is difficult to establish common relationships. But with its diversity and wealth come opportunities to learn and to develop new ideas – ideas that can help change the business. MI is considered a sound approach for large datasets [6].

Big Data is a term that describes large volumes of high speed, complex and data that need advanced techniques and technologies to enable tasks like capture, storage, division management, and analysis of the information. It is a computing infrastructure that can take in, authenticate and analyze high volume of data, and analyzing divers' data (structured/unstructured) from multiple sources [9].

A.Hammad Presented, At the United Nations General Assembly Special Session on HIV/AIDS, governments from 189 countries committed themselves to a comprehensive programmer of international and national action to fight the HIV/AIDS. The Declaration established a number of goals for achievement of specific quantified and time-spring targets, including reductions in HIV infection among infants and young adult's improvements in HIV/AIDS educations, health care and treatment; and improvement in soul support [3].

Rugg. D Presented, It is important to analyze the data or each of the NCPI sections include a write-up in the Country Progress Report in terms of progress made in

- Policy and strategy development and
- Implementation of policies and strategies,

In order to tackle the country's HIV epidemic. Commends on the agreements or discrepancies between overlapping questions in Parts A and B should also be included, as well as a trend analysis on the key NCPI data since 2003, where available [4].

Shi-Wei Lo Presented, the Information Services components provide a service-oriented approach to information consumption. They promote the definition, classification, governance, and reuse of information services using standard interface technologies. These components provide service level abstraction such as message level transformation, mapping, routing and interface and protocol intervention. In addition, they provide a means to classify and catalog services in order to promote reuse of common services and help manage resources in a shared environment. Information Services are a original building block to a Service-Oriented Architecture. An approach is used to automatically view the floods in the particular area by using cyber surveillance systems and image process methods to receive the immediate flooding results [5].

WHO and UNAIDS thank all the people who have provided valuable input and suggestions for this publication. Most of the data in this journal are taken from collected by surveillance systems in developing countries, and we thank the public health professionals who have worked so hard to generate these data. Many of the examples of data use are also based on the growing experience of surveillance professionals and their many partners. We cannot acknowledge every country individually, but we are grateful for the lessons learned and shared from around the world. In keeping with this new metric, they report here in linkage in Georgia within 30 days, a change from previous reports using 90 days [7].

III. Hadoop Framework at a Glance

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. In traditional approach, an enterprise will have a computer to store and process big data. Here data will be store in an RDBMS like Oracle Database, MS SQL Server or DB2 and difficult software can be written to interact with the database, process the compulsory data and present it to the users for analysis purposes.

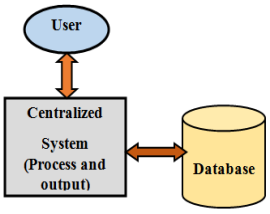


Fig 1: Traditional approach to store and process big data

Sandrine Dudoit presented, A decreasing number of Canadians are able to correctly identify, top of mind, some of the main ways that HIV is transmitted. Most Canadians answer, exclusive of prompting, that the HIV virus is spread through unsafe sexual intercourse between a man and a woman (63 percent, down from 76 percent in 2006). A slight majority, 55 percent, also report that blood to blood contact (i.e., an open wound) is a way to convey the virus. Fewer know that both sharing drug needles (31 percent) and unsafe oral sex (19 percent) can put a person at risk of contracting HIV. Some Candidates (4 percent unprompted and 23 percent prompted) continue to incorrectly believe that HIV can be transmitted through kissing [12].

Andreue-Perez J et al discussed, Hadoop is an open source software framework licensed under Apache Software Foundation, built for supporting data intensive applications running on large clusters and grids, to offer scalable, reliable and circulated computing. Apache Hadoop framework is predominantly designed for the distributed processing of large sets of data residing in clusters of computers using simple programming paradigms. It can be operated from single server or tens of thousands of computers, where each computer is in charge for local computation and storage. Apart from this, Hadoop framework identifies and tackles node failures at the application layer there by offering high availability of the service [15].

Yehia et al presented, Consistent correct use of condoms within non-regular sexual partnerships substantially reduces the risk of sexual HIV program. This is especially important for young people who often experience the highest rates of HIV acquisition because they have low prior exposure to infection and (typically) relatively high numbers of non-regular sexual partnerships. Regular condom use with non-regular sexual partners is important even in circumstances where non-regular relationships are common. Condom use is one measure of protection against HIV/AIDS; delaying age at first sex, reducing the number of non-regular sexual partners, and being faithful to one uninfected partner are equally important [8].

Laney D presented, Big data is really critical to handle as it is rising as one of the faster technologies in current period. The importance of big data is in logical use

which can help in generating informative decision to provide better and quick tune. The big data has three characteristics, known as data volume, velocity and variety, it is also known as 3Vs [10].

IV. Cluster overview

D.Fisher et al discussed that Integration was an important requirement of the design. We needed to take into account the place available in our data center, in which other platforms are already in construction. The technology available today provides the possibility to integrate up to 128 cores in 2U. In our situation, several performance issues would have risen. Allowing for feasible bottlenecks, the selected solution was to integrate four servers, which represents 64 cores, in a 2U chassis. At the end, the operational configuration of the DPCC Hadoop cluster will be made of 400 nodes (with the components explained above) In 100 2U chassis for around 6400 cores and 3PB of raw storage volume. The acquisition of the cluster will be done in five steps, according to the augment in data volume during the cycles [14].

Spileman DA et al presented The vertices with low probability values can either be outside the cluster or inside the cluster but with relatively low significance. Unlike, which involve a sweep operation and a cluster fitness function, they did another round of graph exploring from these unimportant vertices. After the second graph exploring completed, they applied the cluster integration algorithm described in "Cluster merging phase: section [13].

V. Proposed work

The main objective of the proposed work is to build a big data Analytics System that with hadoop technology helps to classify a large and complex medical dataset and detect the disease. In the proposed system, large set of medical records are considered. From this medical dataset, it is aimed to extract the needed information from the record of heart disease patients.

Table 1 : Attacks on Aids workers: Summary statics, 2003-2013

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Number of incidents	63	63	74	107	123	165	155	130	152	170	251
Total aid worker victims	143	125	172	240	220	278	296	254	309	277	460
Total killed	87	56	53	87	88	128	109	72	86	70	155
Total Injured	49	46	96	87	87	90	94	86	127	115	171
Total kidnapped *	7	23	23	66	45	60	93	96	96	92	134
International victims	27	24	15	26	34	51	75	46	29	49	59
National victims	116	101	157	214	186	227	221	208	280	228	401
UN staff	31	11	27	61	39	65	102	44	91	60	110
International NGO staff	69	69	112	110	132	157	129	148	141	87	130
LNGO and staff	35	43	28	55	35	46	55	47	77	105	191
ICRC staff	8	1	3	10	4	5	9	10	5	3	14

For this extraction, features at the data set are analyzed. The analyzed data features are classified to detect the condition, of the heart to identify whether it is normal or abnormal. Also, it is aimed to extract the useful information from large volumes of dataset collected from various sources. In the proposed system, we have taken AIDS disease dataset to classify and the various types of heart disease.

The Human genome is the complete set of nucleic acid sequence for humans (Homosapiens), encoded as DNA within the 23 chromosome pairs in cell nuclei and small DNA molecule found within individual Mitochondria. DNA is the largest person chromosome, chromosome number 1, consists of approximately 220 million base pairs a would be 85 mm long if straightened.

a. Feature Analysis

At the first stage, data are proposed and some features for disease detection are analyzed. In the feature analysis phase, the mean value for all the intervals is calculated for analyzing the type of heart Diseases.

b. Reducing Phase

In reducing phase, map reduce function is used to merge the values from map function into a single result. It reduces a set of intermediate values which share a key to a smaller set of values.

c. Mapping Phase

In a mapping phase, first mapped tokenizes the document and emits an intermediate key-value pair for every record. After this process each of these elements will then be sorted by their key.

The methodology of proposed work is given below

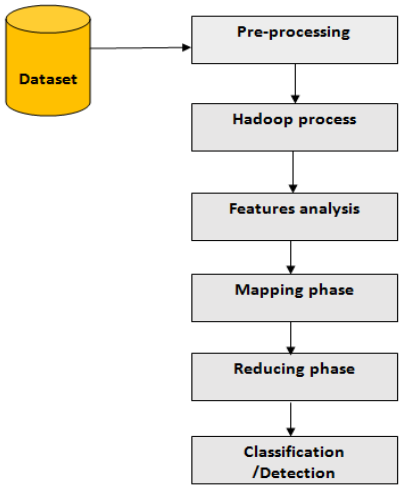
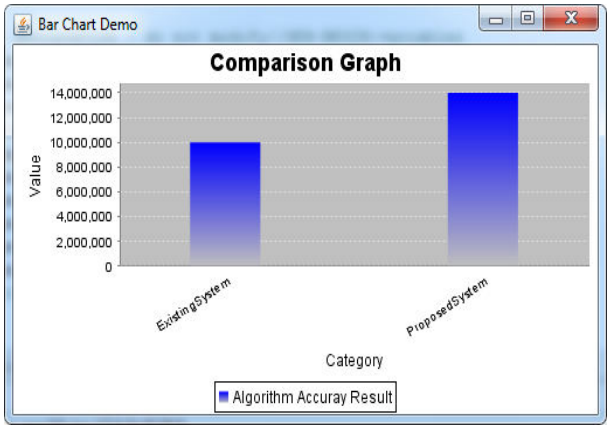


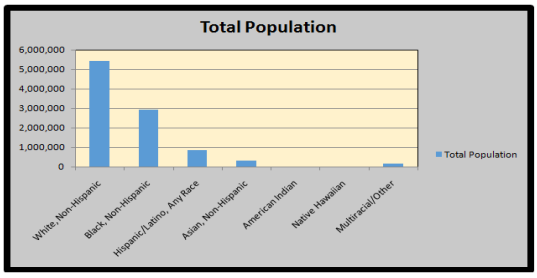
Fig 2: Methodology of proposed work

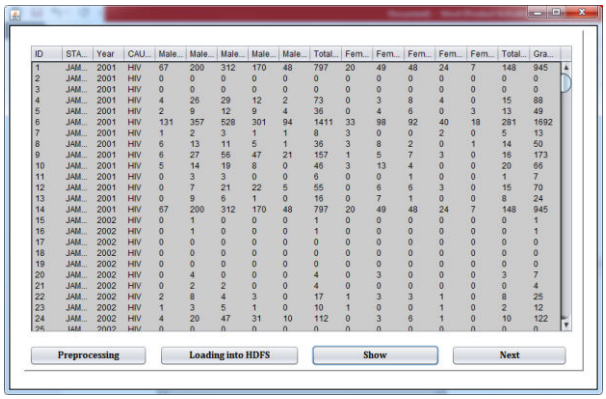
VI. Implementation Results



The GUI of the proposed system of Centralized Patient Monitoring System using Big Data.Muni Kumar N et al, Identified the very big shortage of proper healthcare amenities and addressed how to provide greater access to primary health care service in rural areas of India. Big data processing in real time situation is to turn the dream of (Healthy India) into reality. They analyzed some key factors to make the performance of health centre better and people live healthier. The proposed concept enables doctors, patients and staff to have role-based access to information on electronic health records. They proposed the following seven big ideas to fix rural health care in India and bridge the gap between quality and affordability in government hospitals. For processing this large volume of data, they used hadoop tool [11].The implementation results are given below.

Fig 3: HIV resource for Total population





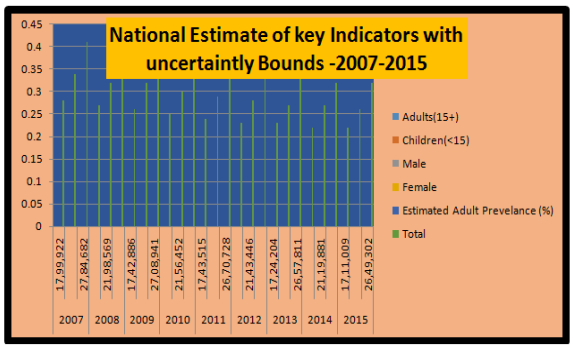


Fig 4: National Estimate of key Indicators with uncertaintly Bounds

VII. Conclusion

Thus the data set are preprocessed and features and analyzed. Using rule based classification, the features are classified for knowing the patient's condition and it displays the type of AIDS disease. In the future work, the data set will be reduced using Cluster technique. This system is expected to be useful in the medical field for the physician to easily analyze the heart disease. It will aid the physicians for taking decision. This system uses Map Reduce technology in hadoop for DNA sequence alignment. Map Reduce framework processes vast amount of data in parallel on large cluster of commodity hardware in a consistent, fault-tolerant manner. This tool is used for detecting HIV/AIDS disease in very effective manner than early approaches. So, it is faster than other existing systems.

VIII. Reference

[1]Amir H. Payberah, "Introduction to Big Data – SICS", April-8, 2014.

[2]Sandrine Dudoit and Robert Gentleman, "Introduction to Genome Biology" 2003.

[3]A. Hammad, A.Garcia, "Hadoop tutorial", September 7, 2011.

[4]Rugg, D.Peersman.G and M. Careal editors (2004). "Global Advances in HIV/AIDS Monitoring and Evaluation New Directions for Evaluation", No.103.Hoboken: Jossey-Bass

[5]Shi-Wei Lo. "health efficiency with medicine"open access sensors 2015,15,23639-2387.

[6]Rubin DB. "Multiple imputation for non response in surveys". New York: John Wiley & Sons; 1987.

[7]Office of "National AIDS Policy, National HIV/AIDS Strategy for the United Ststes": Updated to 2020, July 2015.

[8]Yehia, Baligh R., Fleishman, John A., Metlay, Joshua P., et al. "Comparing different measures of retention in outpatient HIV care". AIDS 2012,

[9]International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcss.com Special Issue on 5th National Conference on Recent Trends in Information Technology 2016 Conference Held at P.V.P Siddhartha Institute of Technology Kanuru, Vijayawada, India.

[10]Laney D. "3D data management: controlling data volume, velocity and variety, META Group, Technology" Rep 2001.

[11]Muni Kumar N and Manjula R (2014), "Role of Big Data Analytics in Rural Health Care- A Step Towards Svasth Bharath". (IJCSIT) International Journal of Computer Science and Information Technologies.

[12]Sandrine Dudoit and Robert Gentleman, "Introduction to Genome Biology", 2003.

[13]Spielman DA, Teng SH. "A local clustering algorithm for massive graphs and is application to nearly – linear time graph partitioning" 2008.

[14]Shi – Wei Lo. "Cyber surveillance for flood disasters" open access sensors 2015.

[15]Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z Big data for health. IEEEJ Biomed Health Inform 2015.