Big Data Analysis for Aids Disease Detection System using Clustering Technique

S. Packiyam[1,] A. Prema[2]
[1]*M.Phil Scholar, Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu*
[2]*Assistant Professor, Department of Computer Science, RDM College,*
Sivaganga, Tamilnadu

Abstract - *Big data analysis is the demanding one because it contains large amount of records. In today's world, the massive information in health care is to be processed in order to recognize, diagnose, detect and prevent the various diseases. It is projected to develop a centralized patient monitoring system using big data. In the planned system, large set of medical records are full as input. From this medical data set, it is aimed to extract the required information from the record of AIDS patients using clustering technique. The classification process states whether the patient is normal or abnormal and in the detection step using clustering technique to detect the disease and decrease the dataset. Thus, the proposed system helps to classify a large and complex medical dataset and detect the AIDS disease. Hadoop is the most popular platform for big data analysis. The Hadoop ecosystem is vast and involves many supporting frameworks and tools to effectively run and manage it. This article focuses on the center of Hadoop concepts and its technique to handle data.*

Key words - *Big data, Hadoop, Cluster, AIDS*

I. INTRODUCTION

A large volume of data is available in most of the real time applications. This raw unstructured data is of no use until it is preprocessed into useful in order. It is necessary to analyze this huge amount of data and extract useful in order from it. For that extraction process, clustering technology is needed. Today big data is the emerging technology. Big data means when the data which is to be mined varies from a small data set to a large dataset. It is future to develop the big data analysis for medical application using big data. Cluster plays a fundamental role in big data. A massive volume of both structured and unstructured data and that is so large and it is difficult to store, analyze, process, share, visualize and manage with normal database and software techniques because it is not much capacity to store large data. Nowadays, many tools are available for processing big data like Hadoop, MongoDb, Talend, Tableau, pentaho, Google charts, SAP in Memory etc. In health care field, there are different types of data available like in signals, images etc. In medical field, big data is a booming factor because in this lot of research work are emerging for the classification of diseases. Big data analytics enables organizations to analyze a mix of controlled, semi-structured and unstructured data in search of valuable business information. Big data analytics is the process of examining large data sets containing a variety of data types.

Hadoop tool is an Apache open source framework written in java that allows distributed processing of large datasets across grouping of computers using simple programs. For the proposed work, the hadoop tool is installed using cygwin terminal for user friendly environment. Hadoop tool is mainly used for processing large amount of data. In hadoop, there is distributed storage system for storing of the data. Hadoop is designed to scale up from single server to thousands of systems, each providing local computation and storage.

Praveen Kumar et al, discussed that hadoop tool is also used in enterprise data, and challenges of processing these huge chunk of data and have found that none of the existing centralized architecture could efficiently handle this huge volume of data. Cluster is a tool implemented for managing and processing vast amount of unstructured data in parallel. Programs which are map reduces are programmed to manage the vast amounts of data. This enables parallel processing of the problem and efficient computation was possible[6].

K. Sharmila, et al examined and revealed the benefits of hadoop in the healthcare sector using data mining. The apache hadoop has become a worldwide adoption and it has brought parallel processing in the hands of average programmer for big data. They presented an overview of various data mining techniques and application of diabetic dataset using this platform. This helped us to acquire knowledge about how hadoop can be implemented to predict the diabetics and related disease. In the hadoop tool, they have used Map reduce concept. Map reduce can divide the dataset into multiple chunks, each will be processed in parallel among multiple nodes[9].

Group of independent servers interconnected through a dedicated network to work as one national data processing source. Clusters are capable of performing multiple complex instructions by distributing workload across all linked servers. Clustering improves the system's availability to users, its combined performance, and overall tolerance to faults and module failures. A failed server is automatically shut down and its users are switched instantly to the other servers. The Cluster program run on the Hadoop tool in an Apache open source framework.

Dean J et al presented that Cluster is a programming model, Google has used successfully dealing out its big data sets. A cluster of computing nodes which are built on commodity hardware will scan the batches and summative their data. Then the multiple nodes' output get merged to generate the final result data. HIV has many things in common with Cluster. In HIV, as in Cluster, processing of data is distributed across many calculated nodes, these separate nodes process their data in parallel and multiple output sets are assembled together to produce a final result set. But, for a variety of reasons, HIV are used in rather different scenarios in data sets.[11]

II. LITERATURE SURVEY

Xindong Wu presented that the data-driven model involves demand driven aggregation of information, source, mining and analysis, user interest modeling and security. One of the main characteristics of big data application is independent data sources with distributed and decentralized control. In which the authors analyzed some issues in the tier model as data sharing and privacy, domain and application knowledge. Big data is mainly related for healthcare system[10].

Kiyana Zolfaghar described about big data driven solutions to predict the 30-day risk of readmission for Congestive Heart Failure (CHF) incident. They mainly used HIV (AIDS Health System) data set. First they extracted useful feature from National Inpatient Dataset (NIS) and supplement it with our patient data set from MHS. Then they developed scalable data mining models to predict risk of using the integrated data set. Also they used random forest algorithm because it can work with all types of predictor variables. For taking the dataset as whole like rural area, the data processing may vary [3].

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 (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 these large volume of data, they used hadoop tool[4].

Prashant Chauhan et al discussed, the GMR (Google Map Reduce) was invented by Google back in their earlier days so they could usefully index all the rich textural and structural information they be collecting, and then present meaningful and actionable outcome to users. MapReduce (you map the operation out to all of those servers and then you reduce the results back into a only result set), is a software paradigm for processing a large data set in a distributed matching way. Since Google's MapReduce and Google file system (GFS) are proprietary, an open source Google's MapReduce platform by using thousands of cluster nodes [31].

.

Sathiyavathi R presented that the first step is to collect the data from various sources, prediction attribute will be identified and then the respective algorithm should be applied in the case. It specifies a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key [8].

Saravana N et al presented for improving the above model, used predictive analysis algorithm in hadoop/map reduce environment to predict the diabetes types and the type of treatment to be provided. This algorithm included various phases like data warehousing, data collection, analysis and submitted the analyzed report. For diabetic treatment, it is necessary to test the patterns akin to plasma, glucose concentration, serum insulin, diastolic blood force, diabetes pedigree, Body Mass Index (BMI). This system is used to predict and classify the types of DM and it leads to the improved focus on every individual patient health[7].

A.Pradeepa et al, presented that consequent algorithms corresponding to the Map reduce based on roughest theory, they put forward to deal with the massive data. Rough set theory proposed a new mathematical approach to imperfect knowledge. It explained the topological operations, center and closure called approximations. Rough set has resolved complex problems. It is a powerful mathematical tool to describe the dependencies along with attributes, evaluate the significance of attributes, and obtain decision rules [5].

Agneeswaran VS et al discussed, Big data is not about data, it involves different tools, techniques and frameworks are used to manage the data. There are many big data platforms available with different characteristics, selection of the platform depends on the capability of the platform and various dimensions as listed in data, Technologies used to handle big data play an important role in data analysis which leads in the accuracy of decision making resulting in cost decrease, faster services, considering calculative risks, and gaining operational efficiencies. To manage and process the great volume of data selecting correct infrastructure is very main. The technologies can be used in capturing and storing the big data and analyzing big data.[12]

Laney D Presented that Big data is really critical to handle as it is emerging as one of the fastest technologies in current era. The importance of big data is analytical use which can help in generating informative decision to provide better and fast service. The big data has three characteristics, known as data volume, velocity and variety , which means that the size of data is large, the data is generated very speedy, and the data exists in heterogeneous formats which can be among structured data, semi structured data with unstructured data captured from different sources [13].

LaValle et al presented that Because the current technology enables us to store and query large datasets efficiently. The focus is now on techniques that make use of the whole data set, instead of sampling. This has tremendous implications in areas like machine education, pattern recognition and classification to name a few.

As a result, building multi-disciplinary teams of “Data scientists” is often an essential means of gaining a competitive edge. More than ever, intellectual property and patent portfolios are becoming essential assets. One of the obstacles to widespread analytics adoption is a lack of understanding on how to use analytics to improve the business. The objects to be modeled and simulated are complex and massive, and correspondingly the data is vast and distributed[29].

Acampora G et al described that comprehensive survey of different tools and techniques used in persistent healthcare in a disease-specific manner. It enclosed the major diseases and disorders that can be quickly detected and treated with the use of expertise, such as fatal and non-fatal falls, Parkinson’s disease, cardio-vascular disorders, stress, etc. We have discussed different pervasive healthcare techniques available to address those diseases and many other stable handicaps, like blindness, motor disabilities, paralysis. Moreover, a plethora of commercially available pervasive healthcare products. It provides understanding of the different aspects of pervasive healthcare with respect to different diseases [35].

Bhawna Gupta et al discussed that how big data is analyzed by using the technique of hadoop and why the big data security analytics is important to mitigate the security threats to secure the enterprise data more efficiently. There should be number of opportunities for big data security analytics to enter the enterprise security. They would use the result for securing and implementing preventive measures from threats to enterprise data. Some researchers are using network monitoring tools like Packet pig, Mahout etc. to enhance the security levels [1].

Devi.L.S et al presented that improving the efficiency of cluster functionality, suggested the theory as store manager. Before executing the real computing job task queries the cache manager. In a data alert cache, cache request and cache reply mechanisms are calculated. Implementing cache by extending hadoop it improves the completion time of map reduce jobs. It detects the amount of repeated job in the incremental data process. Also, it stops the constant work and minimizes the processing time so that to provide the optimized usage of Map Reduce nodes. The data aware cache in map reduce framework helps to overcome this problem and provide high efficiency in incremental processing[2] .

Harrison KM et al presented that the Centers for Disease Control and deterrence released HIV Surveillance Supplemental Report. The report provides data by selected jurisdiction on stage of disease at diagnosis of HIV infection and on the HIV Care Continuum (previously called the HIV Care Cascade). These metrics can be used to monitor progress toward the achievement of objectives outlined in the National HIV/AIDS Strategy for the United States (NHAS). Selection of appropriate measures must take into consideration availability and accuracy of data collection ystems, as well as possible uses of the metrics [40].

Demchenko Y et al discussed the main objective of the proposed work is to build a Big data Analysis System that 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. For this extraction, features in the data set are analyzed. The goal is to extract the useful information from large volumes of dataset collected from various sources. In the proposed system, it is aimed to take AIDS disease dataset to classify and detect the various types of heart disease[15].

Dhruba et al presented that the term “big data” has become a word and as such, it is often over used and misunderstood. For this reason, the original step in choosing between big data frameworks is to determine if they are needed. In arrange to do this, it is important to have an understanding of what constitutes big data. This segment provides definitions of big data and discuss the challenges associated with it. In the years since numerous people have proposed additions to this list and many refer to, adding in Value or Veracity[16].

Suman Arora et al presented that group expresses a slightly higher than average preference for information from television shows and the radio. They are somewhat less apt to rely on doctors, the government, an anonymous clinic, or a person living with HIV/AIDS as sources of information about HIV/AIDS. They tend to place greater than average confidence in the information provided by their friends and would be most uncomfortable seeking information from a range of sources including doctors, pharmacist, and other health care providers, and, in particular, a person with HIV/AIDS [25].

Andreu-Perez J et al presented that despite annotation subjectivity we found sufficient agreement between the observers to support our answer, which show how big data themes are identified in biomedical literature. Technology and methods are found fairly frequently in topics. Note that the identification of these themes is facilitated because they can be associated to concrete terms such as machine, cloud, and platform for Technology, or model, infer, and suggest for Methods. From the V’s, volume and velocity were the most recognized themes, which are also easily associated with terms such as large scale, performance, and computability [33].

Vishal S Patil et al discussed, the heart of machine learning is the data that powers the models, and the original era of Big Data is machine learning to the forefront of research and industry applications. The meaning of the term “big data” is still the subject of some difference, but it generally refers to data that is too big or too complex to process on a lone machine. We live in an age where data is growing orders of magnitude faster than yet before[14].

Yehia et al presented that Publication of this statement was made possible with the contributions of the Georgia Core HIV surveillance staff, HIV Case Report Forms submitted by Georgia health care facility staff, HIV infection-related laboratory test results transmitted by laboratory facilities in Georgia, data matches with other public health programs, and the ongoing efforts of multiple individuals from public and private sector organizations dedicated to improving surveillance, prevention, testing, and care of persons living with HIV infection[38].

Whitmore SK et al presented that the 18 pregnancies that resulted in mother-to-child (MTC) HIV transmission, 12 had received at least one prenatal visit (range 5-10), 4 had received no prenatal care, and the prenatal care status of 2 was unknown. Of the 12 women receiving prenatal care, all except two were diagnosed with HIV infection before or during pregnancy. One woman's diagnosis timing was unknown. One woman receiving prenatal care was HIV-negative early in pregnancy, was subsequently diagnosed with HIV after birth and faced extenuating social circumstances, including IV drug use and homelessness[39].

Bekkerman R et al described the today the problem of big data collections is often solved through distributed storage system, which are planned to carefully control access and management in a fault-tolerant anner. One solution for the problem of big data objects in machine learning is through parallelization of algorithms. This is typically accomplished in one of two ways data parallelism, in which the data is divided into more manageable pieces and each subset is computed simultaneously, or task parallelism, in which the algorithm is divided into steps that can be performed concurrently [18].

Cox et al presented that among the first authors in scientific literature to discuss big data in the context of modern computing. Their job focused on data dream, but their observations about the big data problem can easily be extrapolated to general data analytics and machine knowledge. The big data problem, according to them, consists of two distinct issues: In 1997[17].

White T presented it is not uncommon to encounter big collections of big objects as data grows and becomes more widely available. This coupled with unprecedented access to computing power through more affordable high performance machines as well as cloud army, is opening up many new opportunities for machine learning explore. Many of these new directions utilize increasingly complex workflows which require systems built using a combination of state-of-the art tools and techniques. One choice for such a system is to use projects from the Hadoop Ecosystem[19].

A Review Hadoop cluster includes a single master and multiple employee nodes. The master node consists of a JobTracker, TaskTracker, NameNode also DataNode. A slave or worker node acts as together a DataNode and TaskTracker, though it is practical to have data-only worker nodes and compute-only worker nodes. These are usually used only in nonstandard applications. Hadoop requires Java Runtime setting (JRE). The standard start-up and shutdown scripts require Secure Shell to be between nodes in the cluster[32].

P.V.P. Siddhartha et al presented that Big Data is a term that describes large volumes of high velocity, complex and data that require advanced techniques and technologies to enable tasks similar to capture, storage, distribution, management, and analysis of the information. It is a computing infrastructure that can take in, confirm and analyze high volume of data, and analyzing divers data (structured/unstructured) from multiple sources[30].

Satyanarayana.A presented Sampling and compression are two representative data reduction methods for big data analytics because reducing the size of data makes the data analytics computationally fewer expensive, thus faster, especially for the data coming to the system fast. In addition to making the sampling data represent the original data effectively , how many instances need to be selected for data mining method is another research issue because it will affect the performance of the sampling method in most cases [37].

Kaiser Permanente et al described, "Inside the early on stage of HIV virus, the the majority regular symptoms are nobody". This tool is based on the symptoms and the values are assigned based on the advice of doctors. The details of HIV symptoms are collected from government hospitals. Within a month or two of HIV entering the body, people experience the following symptoms known as acute retroviral syndrome[27].

According to Amir H et al, Big Data is a high volume, high velocity and high variety in order to asset that demand cost-effective , innovative forums of information dispensation for enhanced insight and decision making. Big data, a buzzword in the business aptitude can handle petabytes or terabytes of data in a reasonable amount of time. Big data is distinct from large accessible database which uses Hadoop framework for data intensive distributed applications. Disease diagnosis and prospects are based on effective detection of disease setting (e.g. cancer), infectious organisms (e.g. HIV) and genetic markers. Still, DNA study from original specimens is a complex process involving multiple chemical compositions as well as multistep reactions[22].

Spielman DA et al presented that 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 health function, we do another round of graph exploring from these trivial vertices[36].

Vavilapalli VK et al presented the addition of AIDS to the Hadoop and Cluster were tightly coupled, with responsible for both cluster resource management and data processing. Big data has now taken over the resource running duties, allowing a separation between that infrastructure and the programming model. With AIDS, if an application wants to run, its client has to request the launch of an application manager process from the store manager, which then finds a join manager. The node manager then launches a container which executes the application course. [20].

Porambage P et al presented that privacy and security in terms of big data is a main issue. Big data security model is not suggested in the event of complex applications due to which it gets disabled by evasion. However, in its absence, data can always be compromised simply. As such this section focuses on the privacy and security issues. Information privacy is the capacity of an individual or group to stop information about themselves from becoming known to people other than those they give the in order to. One serious user privacy issue is the identification of personal information during transmission over the Internet [34].

Fernández A et al presented that offers an Application Programming Interface (API) that abstracts the traditional keys and values into tuples with field names and offers a number of operations on the tuples that help developers build complex applications more easily and in less time. Cascading primarily supports programming in Java, but too offers Predictive Model Markup tongue. It also supports easy integration of a large number of different data sources [21].

Wang, F. et al presented that characterized by the lowest knowledge about HIV/AIDS by future. They also have the second highest level of rated discomfort around people living with HIV/AIDS. Map Reduce is a software framework for distributed processing of large data sets on computer clusters. This group is likely to distance themselves from the issue of HIV/AIDS, believing that it is a disease found mostly in third world countries, and among the gay population and drug users[26].

Sandrine Dudoit et al presented 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 in a small DNA molecule found within individual Mitochondria. DNA is the largest human chromosome, chromosome number 1, consists of about 220 million base pairs a would be 85 mm long if straightened [23].

Parmeshwari P et al discussed about the NameNode records all of the metadata, attributes, and locations of records and data blocks in to the DataNodes. The attributes, it records are the clothes like file permissions, file change and contact times, and namespace, which is a hierarchy of files and directories[28].

According to A.Hammad et al**, DNA sequencing** is the process of determining the precise order of nucleotides within a DNA molecule. Mapper reads each line as input sequences. Finally, the united sequences are obtained from reducer. The symptoms percentage calculated from user data and the similarity percentage of gene sequences are added and the result is the percentage of a person affected by HIV [24].

III. CONCLUSION

The proposed approaches for collecting and storing Big Data for analytics presented in this paper show how important it is to select the technology migration. 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. Big data involves the data produced by different devices and applications. This paper proposes a big data integrated framework to assist with prevention and control of HIV/AIDS, TB and silicosis in the removal industry. Hadoop MapReduce is a great scale, open source software framework devoted to scalable, distributed, data-intensive computing.

IV. FUTURE WORK

For future work, we will detect many type of diseases using the big data set and by analyzing the data set, the required data will be predicted from the data set as easily using Cluster concept. The task queries the cache manager before executing the actual computing work.

V. REFERENCES

[1] Bhawna Gupta and Dr. Kiran Jyoti (2014), "Big data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data", (IJCSIT) International Journal of Computer Science and Information Technologies.

[2] Devi.L and S.Gowri (2015), "Optimizing Cluster functionality in bigdata using Cache Manager", ARPN Journal of Engineering and Applied Sciences.

[3] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin and Brain Muckian(2013) "Big Data Solutions for Predicting Risk-of-Readmission for Congestive AIDS", IEEE International Conference on Big Data.

[4] 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.

[5] Pradeepa A, Dr. Antony Selvadoss Thanamani (2013), " Hadoop File System and Fundamental Concept of Cluster Interior and Closure Rough Set Approximations", International Journal of Advanced Research in Computer and Communication Engineering.

[6] Praveen Kumar and Dr. Vijay Singh Rathore (2014), " Efficient Capabilities of Processing of Big Data using Hadoop Cluster", International Journal of Advanced Research in Computer and Communication Engineering.

[7] Saravana N, M Ramachandran and S. Lavanya Kumar (2015) , " Predictive Metodology for Diabetic Data Analysis in Big Data", ScienceDirect-Procedia Computer Science.

[8] Sathiyavathi R (2015), " A Survey: Big Data Analytics on Healthcare System", HIKARI Ltd Contemporary Engineering Sciences.

[9] K. Sharmila and Dr. S.A.Vethamanickam (2015), "Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform", International Journal.

[10] Xindong Wu, Fellow, Xingquan Zhu, Gong-Qing Wu and Wei Ding (2014) "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering.

[11] Dean, J. and Ghemawat, S. 2008. MapReduce: simplified data processing on large clusters. Communication of ACM 51, 1 (Jan. 2008), 107-113.

[12] Agneeswaran VS, Tonpay P, Tiwary J (2013) Paradigms for realizing machine learning algorithms. Big Data 1(4):207–214.

[13] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001.

[14] Vishal S Patil, Pravin D. Soni, **"**Hadoop skeleton & fault tolerance in hadoop clusters**"**, International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 - 4847

[15] Demchenko Y, Grosso P, de Laat C, Membrey P. Addressing big data issues in scientific data infrastructure. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, 2013. IEEE, pp 48–55.

[16] Dhruba, jssarma, jgray, kannan, Nicolas, hairong, krangana than dms, aravind, menon, rsh, Rodrigo, animated. "Apache HAdoop Goes Realtime at Facebook".

[17] Cox M, Ellsworth D. Managing big data for scientific visualization. In: ACM Siggraph '97 course 4 exploring gigabyte datasets in real-time: algorithms, data management, and time-critical design, August, 1997

[18] Bekkerman R, Bilenko M, Langford J. Scaling up machine learning: parallel and distributed approaches. Cambridge: Cambridge University Press; 2011.

[19] White T. Hadoop: The Definitive Guide, 3rd edn. Sebastopol, CA:O'Reilly Media, Inc.; 2012.

[20] Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S, Saha B, Curino C, O'Malley O, Radia S, Reed B, Baldeschwieler E. Apache Hadoop : Yet Another Resource Negotiator. In: Proceedings of the 4th annual Symposium on Cloud Computing; 2013.

[21] Fernández A, del Río S, López V, Bawakid A, del Jesus MJ, Benítez JM, Herrera F. Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. Wiley Interdiscip Rev Data Min Knowl Discov. 2014;4(5):380–409.

[22] Amir H. Payberah,' Introduction to Big Data -SICS', April-8, 2014.

[23] Sandrine Dudoit and Robert Gentleman, 'Introduction to Genome Biology', 2003.

[24] A.Hammad, A.Garcia,'Hadoop tutorial', September7, 2011.

[25] Suman Arora, Dr.Madhu Goel, "Survey Paper on Scheduling in Hadoop" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[26] Wang, F. et al. Hadoop High Availability through Metadata Replication. ACM (2009).

[27] Kaiser Permanente, in Oakland'16 Signs You May Have HIV'.

[28] Parmeshwari P. Sabnis, Chaitali A.Laulkar, "SURVEY OF MAPREDUCE OPTIMIZATION METHODS", ISSN (Print): 2319- 2526, Volume -3, Issue -1, 2014

[29] LaValle et al: Big Data, Analytics and the Path From Insights to Value, (Dec 2010)

[30] International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: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.

[31] Prashant Chauhan, Abdul Jhummarwala, Manoj Pandya, ―Detection of DDoS Attack in Semantic Web‖ International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4-No.6, December 2012

[32] A Review on HADOOP MAPREDUCE-A Job Aware Scheduling Technology ISSN(e): 2250 – 3005 Vol, 04 Issue, 5 May – 2014 International Journal of Computational Engineering Research (IJCER)

[33] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z. Big data for health. IEEE J Biomed Health Inform. 2015;19(4):1193–208.

[34] Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.

[35] Acampora G, et al. Data analytics for pervasive health. In: Healthcare data analytics. ISSN:533-576. 2015.

[36] Spielman DA, Teng SH. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning 2008.

[37] Satyanarayana A. Intelligent sampling for big data using bootstrap sampling and chebyshev inequality. In: Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, 2014. pp 1–6.

[38] Yehia, Baligh R., Fleishman, John A., Metlay, Joshua P., et al. Comparing different measures of retention in outpatient HIV care. *AIDS* 2012, 26:1131-1139.

[39] Whitmore SK, Patel-Larson A, Espinoza L, et.al. Missed opportunities to prevent perinatal human immunodeficiency virus transmission in 15 jurisdictions in the United States during 2005-2008. Women Health 2010 Jul;50(5):414-25.

[40] Harrison KM, Kajese T, Hall HI, Song R. Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach. Public Health Rep 2008;123:618–27.