International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319 – 6505, Impact Factor: SJIF: 5.995
Available Online at www.journalijcar.org
Volume 6; Issue 7; July 2017; Page No. 4476-4480
DOI: http://dx.doi.org/10.24327/ijcar.2017.4480.0522

Research Article

BIG DATA ANALYTICS USING HADOOP SOLUTION FOR HIV

Packiyam S[1] and Prema A[2]

[1]Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu
[2]Department of Computer Science, RDM College, Sivaganga, Tamilnadu

A R T I C L E I N F O

Article History:
Received 13th April, 2017
Received in revised form 4th May, 2017
Accepted 17th June, 2017
Published online 28th July, 2017

Key words:
Big data, Hadoop, Cluster, AIDS

A B S T R A C T

Big data refers to the dynamic, large and disparate volumes of data being formed by people, tools and machines it requires new, original and scalable equipment to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to customers, risk, profit, performance, productivity management and enhanced shareholder value. A good understanding of Hadoop Architecture is required to leverage the power of Hadoop. This list primarily includes questions related to Hadoop Architecture, Hadoop and Hadoop Distributed File System (HDFS). Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one a further, but are very dissimilar to objects in other clusters. Clustering is commonly used to search for unique grouping within a data set. AIDS disease is a major health problem and it is the leading causes of death during the world. Early detection of AIDS disease has become an important issue in the medical research fields.

INTRODUCTION

Big data includes information garnered from public media, data from internet-enabled devices (including smartphones and tablets), mechanism data, video and voice recordings, and the constant preservation and logging of structured and unstructured data. Below are few important practical questions which can be asked to a Senior Experienced Hadoop Developer in an interview. I hope you will find them useful. 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 data set. Thus, the proposed system helps to classify a large and complex medical detect the AIDS disease.

Edd Dumbill presented the Big sensing data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity. Cloud computing provide a promising platform for big sensing data processing and storage as it provides a flexible stack of massive computing, storage, and software services in a hasty manner. Current big sensing data processing on Cloud have adopted some data compression techniques [1].

Amir H. Payberah presented that, However, due to the high volume and velocity of big sensing data, conventional data compression techniques lack sufficient efficiency and scalability for data processing. Based on detailed on-Cloud data compression requirements, they have proposed a novel scalable data compression approach based on calculating similarity among the partition data chunks. Instead of compressing basic data units, the density will be conducted over partitioned data chunks [5].

Tudorica, B.G presented the, to restore unique data sets, some restoration functions and predictions will be calculated. MapReduce is used for algorithm implementation to achieve extra scalability on Cloud. With real world meteorological big sense data experiments on U-Cloud platform, they have confirmed that the proposed scalable compression approach based on data chunk similarity can significantly improve data compression efficiency with affordable data accuracy loss [2].

Huizhi Liang eta al presentedthat, Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model Hadoop creates cluster of machines and coordinates work among them. It is designed to scale up from single servers to thousands of equipment, each offering local computation and storage Hadoop consists of two component Hadoop Distributed File System (HDFS) and MapReduce Framework [4].

Jeffrey Dean presented that, Group of independent servers (Usually in closeness to one another) 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 [3].

Corresponding author:* **Packiyam S
Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu

A.Hammad *et al* presented that, 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. The value pair and display the count value as output [8].

A., H. A. Weiss *et al*, presented that, the master nodes, organism unique, have significantly different storage and memory requirements than the slave nodes. They are recommented using dual Name Node servers-one primary and one inferior. Both Name Node servers should have highly reliable storage for their name space storage and edit-log journalist. Multiple vendors sell NAS software. It is significant to check their specifications before you invest in any NAS software [16].

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 salable, 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 [20].

HIV (Human Immune Virus)-AIDS (Acquired Immune Deficiency Syndrome)

Sandrine Dudoit and Robert Gentleman presented that, in the early 1980s, the first recognized cases of the acquired immune deficiency condition (AIDS) occurred among homosexual men in the United States. These men suddenly began to develop rare opportunistic infections and cancers that seemed stubbornly resistant to any treatment. At this time, AIDS did not yet include a name, but it quickly became obvious that all the men were suffering from a common syndrome. By 1983, the etiological manager, the human immunodeficiency virus (HIV), had been identified [6].

There is now clear evidence to prove that HIV causes AIDS. By the mid-1980‟s, it became clear that the virus had multiply, largely unnoticed, throughout most of the world, and because then, the global AIDS epidemic has become one of the greatest threats to human health and development. At the same time, much has been learn about the science of AIDS, as well as how to avoid and treat the disease [9].

This will comprise global statistics, important definitions that distinguish HIV from AIDS, the impact of HIV on the body, modes of transmission, as well as methods of prevention and treatment [10].

This introductory module provides basic information and facts on the current status of the disease globally using the latest available statistics. It targets individuals who are not familiar with the basic facts about HIV and AIDS or who are looking for a useful resource with updated statistics on the global situation [7].

Related works

Dean, J. and Ghemawat, S., presented that, In each country the research project will be different but consistent with the ethos of the People Living with HIV disgrace Index. The number of people interviewed will be diverse, as will be the outreach and composition of responses from different groups (such as men who have sex with men, sex personnel, injecting drug users and other vulnerable groups). The attitude and research design in each country will build on a core commitment to the ethical process and rigor and sensitivity of each individual interview [11].

Guy RJ *et al* presented that, Cascading allows the simplification of data handling along with processing, and provides tools to perform complex queries using the Hadoop framework. The experience acquired during benchmarks shows us that programming Map/Reduce jobs from scratch using Map/Reduce interface is not an efficient expansion scheme.

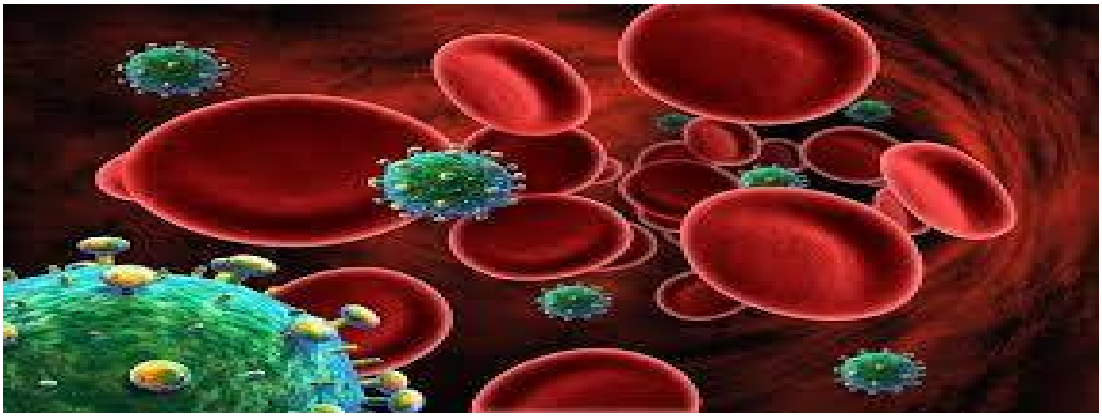

Fig 1 Image of HIV disease

Hive and Pig higher level tools cannot be used because SAGA uses binary Java object as input and not text file based interfaces [18].

Jeffrey Dean presented that, New and different source of data, such as national population-based surveys, are enabling more accurate estimates and more refined understandings of the epidemic's trends. In calculation, the results from the models are being compared with other sources such as data from epidemiological research studies, demographic examination sites, mortality surveys, or early infant diagnosis results to improve the assumptions. Importantly, the roles of national AIDS programmer have changed significantly since the first set of UNAIDS country specific estimates was produced in 1997. Initially, countries were requested to comment on provisional estimates [12].

B., M. Carael *et al.*, Presented that Since along with its partners (including East-West Center, Futures organization, WHO, UNICEF, the US Census Bureau and the US Centers

for Disease Control and Prevention) have carried out a series of regional training workshops in which epidemiologists from over 150 countries were trained in the HIV judgment process. Such efforts have led to much greater involvement by national programmer, national statistics offices and other government and academic organizations in the production of estimates. The result has been better excellence estimates, due to the use of additional data and the application of local knowledge [17].

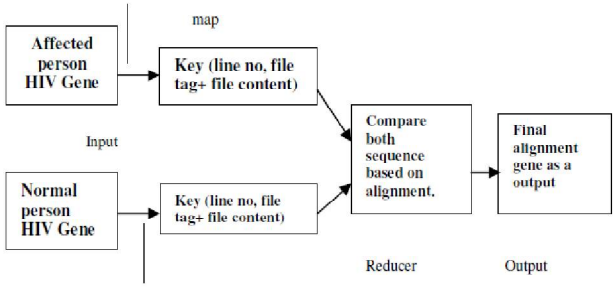

Fig 2 HIV Gene sequence Alignment

Körner N. *et al* presented that, DNA sequencing is the process of determining the precise order of nucleotide within a DNA molecule. It include any method or technology that is used to determine the order of the four source adenine, guanine, cytosine, and thymine-in a strand of DNA. The two sequence files such as query file(HIV affected sequence) and the non-affected DNA sequence are given as input to the mapper where we pass offset as key and value is line of file and file tag is also passed to verify the file[19].

Yandong Mao *et al* presented that, the available evidence suggests that HIV prevalence in East Africa has stabilized and in some settings may be declining. Declines in HIV prevalence reported in Uganda in the past decade appear to have reached a plateau (Wabwire- Mangen *et al*., 2009), although these trends may partly be related to the roll-out of antiretroviral therapy programmed [15].

E. Molina-Estolano *et al* presented that, this system uses the symptoms and HIV DNA sequence comparison result to identify the possibility fraction of AIDS. Here, most common symptoms already mentioned are taken as input. After the alignment received from reducer, it finds the comparison percentage of two sequences. Finally, add the similarity percentage of two sequences and result of symptoms percentage to predict the possibility percentage of AIDS in a person[14].

K. Shvachko presented that, Hadoop is designed to run on a large collection of machines that shares neither memory nor disks. That means that unlike HPC huddle each node serves a dual-purpose: on the one hand it is a computing resource, on the other hand it is a storage unit. The advantages of this software are that it can handle PetaBytes data sets simply, and it provides a framework for distributing processing over a cluster. Its major drawback is the difficulty to handle complex data structure and to perform complex queries on it. Fortunately, other frameworks on top of Hadoop like Cascading exist [13].

Proposed work

The features of HIV data set are analyzed. The analyzed data features are classified to detect the condition and the identify whether it is normal or abnormal. Also, it is aimed to extract the useful information from large volumes of data set collected from various sources. We proposed a new algorithm namely JNK-node algorithm with an integrated concept of Big data analytic and Hadoop to classify the AIDS disease.

Fig 3: Shows the methodology of proposed work.

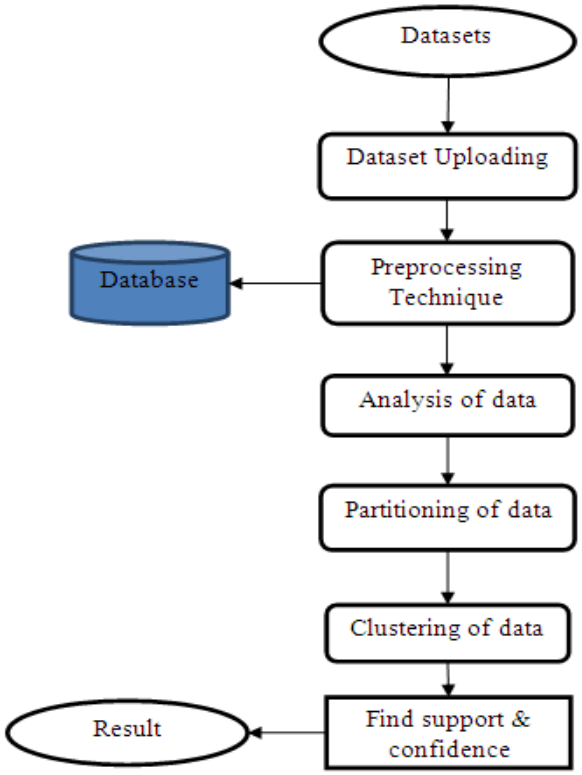

Fig 3 Methodology of proposed work

The impact of AIDS on individuals, families, communities and the nation is increasing drastically. HIV/AIDS has affected individuals and families physically, socially, psychologically, emotionally, and economically. Therefore, every one of us has got responsibility to guard against HIV infection, to fight its spread and to support those infected and affected in different ways.

The HIV Replication Cycle site presents a review of HIV duplication in cells, but linked to extensive web-based resources. Accounts of the HIV proteins are better with movies of HIV protein structures to allow visualization in three dimensions. Numerous network links lead from the site to other resources.

One link allows readers to navigate out on to the human genome and surf approximately, viewing positions of HIV incorporation sites. All images, movies and other materials are available for free download for use by AIDS researchers and educators. Importantly this context, simple explanations of different parts of the HIV replication cycle are linked to large data sets available for study in the Gene Overlapped site.

We have taken the data set of HIV affected workers to implement this algorithm and the HIV data set of is given in the following table.

JNK node Algorithm

Step 1: Develop weight factors for part and machines iw andjw

Step 2: If there are more than weight factors then, convert each weight factor into percentage and then sum it. Else assign the weight to the part numbers

Step 3: Create an n*m matrix bij(binary number for part and machine). Where, n is parts and m is machines

Step 4: Rearrange the parts and machines in descending order based on weights Step 5: For each row of icompute, ij*2m-j

Step 6: Rearrange the rows in descending order based on the computed numbers Step 7: For each row of j compute,ij*2n-i

Step 8: Rearrange the columns in descending order based on the computed numbers

Step 9: Repeat step 1 until there is no change is observed in step 3 and 5

Step 10: Stop

classified for knowing the patient's condition and it displays the type of AIDS disease. 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 constant, 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. In the future work, the data set will be reduced using Cluster technique.

Table 1 : Attacks on AIDS workers: Summary statistics, 2007-2017

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Number of incidents	63	63	74	107	123	165	155	130	152	170	251
Total worker victims	143	125	172	240	675	897	456	345	789	234	765
Total killed	87	90	56	45	87	50	234	523	756	978	543
Total Injured	6	89	348	418	49	82	81	77	934	633	277
Total kidnapped	75	36	94	987	453	265	264	163	125	275	887
International victim	65	24	98	34	86	243	65	657	99	190	100
National victims	46	388	49	300	276	36	90	37	675	342	980

The result of this algorithm is shown in figure 4.

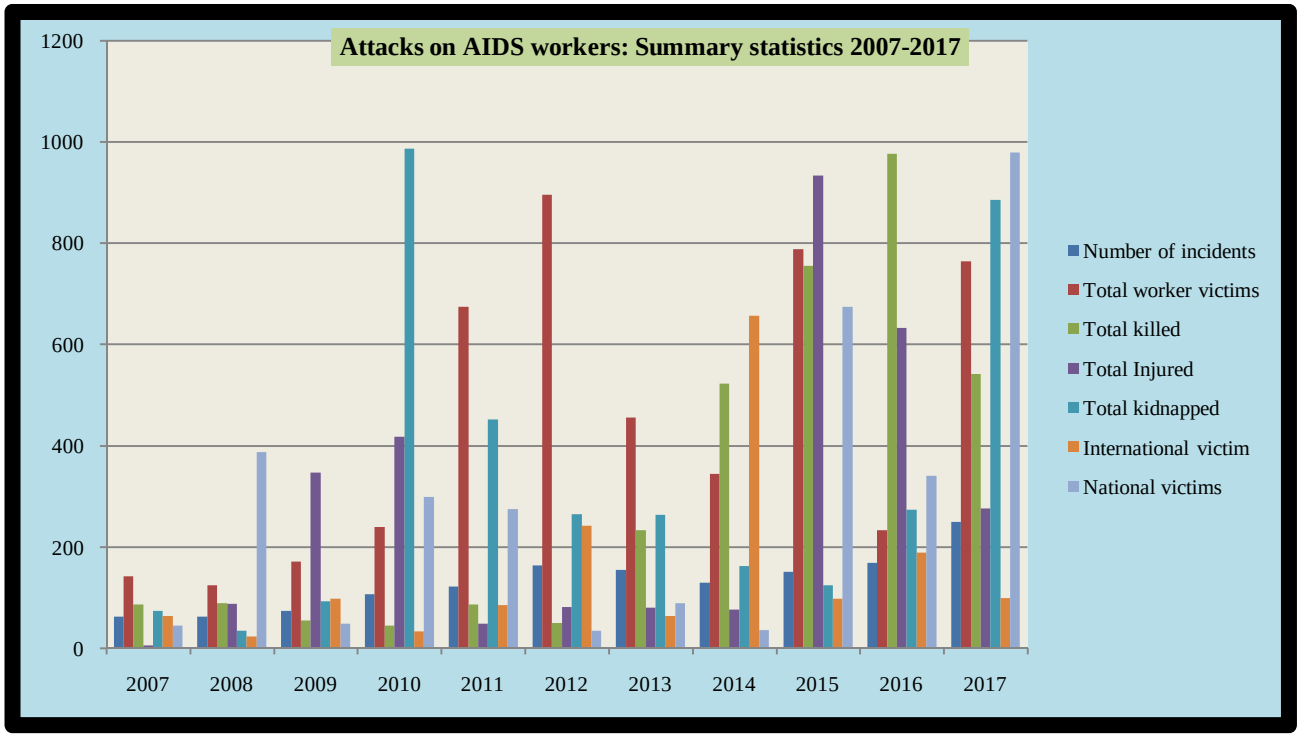

Fig 4 Attacks on AIDS Workers

CONCLUSION

We have recommended an integrated modern approach of Big Data analytic, Hadoop Map reduce with JNK node algorithm to obtain an optimal solution for getting better decision-making Thus the data set are pre-processed and features and analyzed. Using rule based classification, the features are

References

1. Edd Dumbill, Planning for Big Data, O'Reilly Media, 2012.

2. Tudorica, B.G. "A comparison between several NoSQL databases with comments and notes", Roedunet International Conference (RoEduNet), 2011.
3. Jeffrey Dean, "MapReduce: a flexible data processing tool", Communications of the ACM, Volume 53 Issue 1, January 2010.
4. Huizhi Liang, Hogan, J., Yue Xu. "Parallel User Profiling Based on Folksonomy for Large Scaled Recommender 2012.
5. Amir H. Payberah,' Introduction to Big Data - SICS', April-8, 2014.
6. Sandrine Dudoit and Robert Gentleman, 'Introduction to Genome Biology', 2003.
7. World Health Organization,' Early detection of HIV infection in infants and children'
8. A.Hammad, A.Garcia,'Hadoop tutorial', September7, 2011.
9. www.health.com,'16 Signs You May Have HIV'.
10. The download and shutup.us,'Hiv DNA sequence download', [Dec 16, 2014].
11. Dean, J. and Ghemawat, S, 'Hiv dna sequence download'.
12. Jeffrey Dean, "MapReduce: a flexible data processing tool", Communications of the ACM, Volume 53 Issue 1, January 2010.
13. K. Shvachko, H. Huang, S. Radia, and R. Chansler. "The hadoop distributed File system2010.
14. E. Molina-Estolano [, M. Gokhale, C. Maltzahn *et al*., "Mixing Hadoop and Hpc Workloads on Parallel Filesystems," 2009.
15. Yandong Mao, Robert Morris, M. Frans Kaashoek, "Optimizing MapReduce for Multicore Architecture" 2010.
16. A., H. A. Weiss, M. Laga, E. Van Dyck, R. Musonda, L. Zekeng, M. Kahindo, S. Anagonou, L. Morison, N. J. Robinson, R. J. Hayes, and Study Group on Heterogeneity of HIV Epidemics in African Cities, "The Epidemiology of Gonorrhoea, Chlamydial Infection and Syphilis in Four African Cities," 2001
17. B., M. Carael, A. Buve, B. Auvert, M. Laourou, L. Kanhonou, M. De Loenzien, E. Akam, J. Chege, and F. Kaona, "Comparison of Key Parameters of Sexual Behavior in Four African Urban Populations with Different Levels of HIV Infection," 2001.
18. Guy RJ *et al*. HIV diagnoses in Australia: diverging epidemics within a low-prevalence country. 2007
19. Körner N. Late HIV diagnosis of people from culturally and linguistically diverse backgrounds in Sydney: the role of culture and community 2007.
20. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang G-Z Big data for health. *IEEEJ Biomed Health Inform* 2015.

How to cite this article:

Packiyam S and Prema A *et al* (2017) 'Big Data Analytics Using Hadoop Solution For Hiv', *International Journal of Current Advanced Research*, 06(07), pp. 4476-4480. DOI: http://dx.doi.org/10.24327/ijcar.2017.4480.0522

****