IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 51-54
www.iosrjournals.org

HADOOP: Solution for Big Data Challenges in Bioinformatics and its Prospective in India

Akshay Ware[1, 2], Ganesh Janvale[1], Faiyaz Shaikh[1], Sanjay Harke[1]
[1](*MGM's Institute of Biosciences and Technology, Aurangabad, India*)
[2](*MGM's Dr. G.Y.P. College of CS and IT, Aurangabad, India*)

Abstract: *Data becomes big data while its volume, variety, and velocity exceed abilities of our systems architecture and algorithm. Data volume is increasingly triggered by recent advancements in high-throughput technologies like Next Generation Sequencing (NGS), the discovery of gene expression and personalized medicine. Hadoop open source platform is magnificent to obtain information from raw biological big data with providing massive storage, advanced security, and fast data processing platform. In this review, comprehensive applicability of Hadoop being an open source platform is discussed with its global status to overcome big data challenges and its perspective in India.*

Keywords: *Bioinformatics, Hadoop, Open Source*

I. Introduction

Big data is unstructured, semi structured or structured large amount of data [1]. Annual data production is increasing day by day. By 2020 rate of data generation will reach ten times greater than the data generated nowadays [2]. The scientific data produced within the past 7 to 8 years is found to be massive when compared to the entire human history [3]. The influence of advanced information technology gives bioinformatics an interdisciplinary nature [2]. High through-put sequencing technology plays a major role in the explosion of biological data [3]. The Gene bank data submission rate is doubling every 14 months and till 2010 more than 100 million data was uploaded [3]. Biology is the complex field where the data processing is much needed because the data is used for mankind. Big Data in the biological field has a transforming power to bring dramatic changes in the biology. The growth of large, unstructured datasets is driving the development of new technologies for finding items of interest in biological data. Because of the tremendous expansion of data from DNA sequencing, bioinformatics has become an active area of research in the supercomputing community. Therefore processing of big data is considered to be most beneficial towards biological problems [4]. Biological databases play an important role in bioinformatics to provide wide information's on one platform. The biological databases and web portal maintenance is the most important job for bioinformatics which include security, submission control, authenticity, retrieval etc. The human genome project revealed billions of not only bases but also gene present. High-throughput next generation sequencing technology plays an important role in providing ultra high speed and lower cost of sequencing [5].

II. Bioinformatics And Big Data

Bioinformatics is an important source of huge information at various levels [6]. (Fig 2) demonstrating few of the data sources. With the digitization of all processes and availability of high throughput devices at lower costs, data volume is rising everywhere, including in bioinformatics research. For example, the size of a single sequenced human genome is approximately 200 gigabytes [7]. The European Bioinformatics Institute (EBI) had approximately 40 petabytes of data about genes, proteins, and small molecules in 2014 as compared to 2013 which is 22 petabytes more and total size of storage doubling every year [2, 8]. The Protein Data Bank (PDB) structural database had 124588 Biological Macromolecular Structures [9]. Data in Bioinformatics is continually growing due to technology being able to generate more molecular data per individual. The increasing amount of data in the biological field has greatly increased the importance of developing data mining and analysis techniques which are efficient to handle Big Data. The high volume of data is helpful for the highly sensitive field of research like bioinformatics. The main issues in bioinformatics are the generation of data challenges to the profound source of storage, data transportation, security, and privacy.

III. Open Source Approach

Open source software's is software's like any other where they get distribute by their license. Open source gives the right to access and modify the source code. Open source software can modify and redistribute without any royalty or another cost [10]. BioPython, BioJava, BioPig are open source tools used in bioinformatics. BioPython is a big open-source application programming interface (API) used in bioinformatics to scripting for common bioinformatics tasks [11]. BioJava is an open-source project that provides a framework for the processing of biological data [12]. Hadoop is also an open source platform use for big data analytics [13].

IV. HADOOP: A Growing Solution

Hadoop is a most significant platform for big data analytics under apache open source. Google had taken steps towards developing Hadoop through MapReduce concept. With the launching of MapReduce algorithm, Google has solved many problems regarding big data. MapReduce algorithm divides the task into small parts and assigns it to different nodes, and collects the result after processing. Using this solution by Google, Dough cutting and his team developed Hadoop platform. The first release of Hadoop was launched on 10th December 2011[13] and the first stable version (2.7.3) came into existence on 25th August 2016[14]. Hadoop is an open source software framework that can be installed on a commodity Linux cluster for large-scale distributed data analysis [15]. Hadoop provides high-throughput access to application data and is suitable for applications that have large data sets. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. Hadoop consists of MapReduce and Hadoop distributed file system (HDFS) on which application can run. Every node can act as master or slave according to the requirement based for MapReduce and HDFS [16]. There is an open source biological project built on top of Hadoop that is BioPig. CloudBurst Hadoop-based tool is used for the alignment of short read on AmezonEC2 [17]. Cloud computing is one of the major aspects to overcome problems of data transportation. It also provides Unified, location- an independent platform for data and processing. Cloud based services in bioinformatics are grouped into Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS) [18]. John Craig Venter Institute (JCVI), Northwest Environmental Business Council (NEBC) Harvard School of Public Health and others provide cloud infrastructure to enable genome analysis on cloud computing platforms with a BioLinux. BioLinux provides 135 and more bioinformatics software which include sequence alignment tool, clustering tool, sequence assembling software’s, visualization tools, genome editing and phylogenetically-based packages [19]. With the presence of numerous bioinformatics clouds, interoperability and standardization between clouds will become important issues [20]. Developing new strategies using Hadoop cluster is an important aspect in bioinformatics. (Fig 1) denoting the position of Hadoop and tools in bioinformatics data process system.

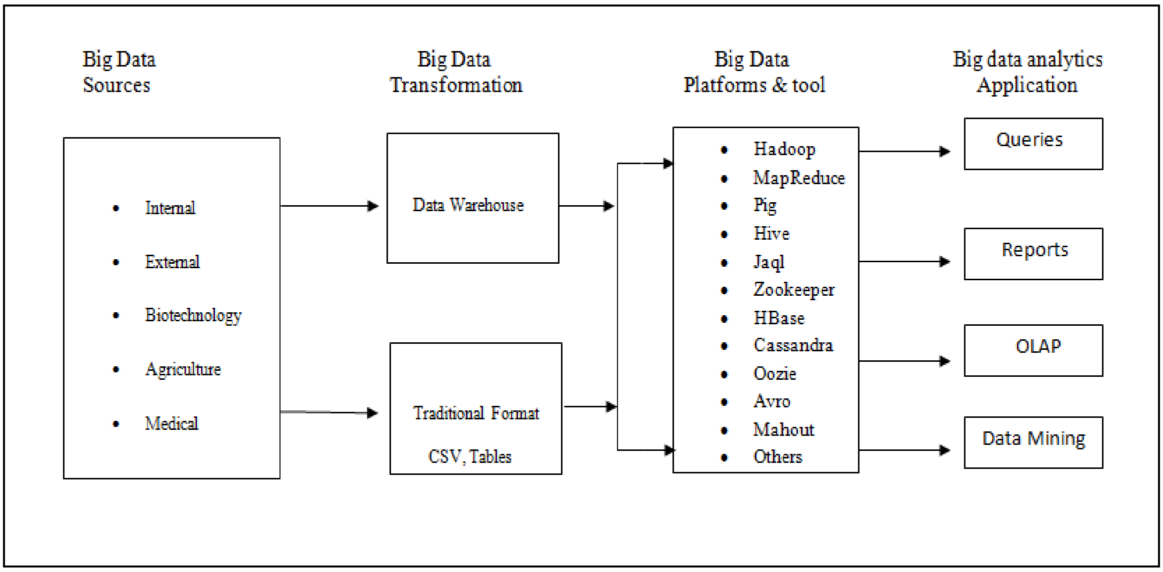
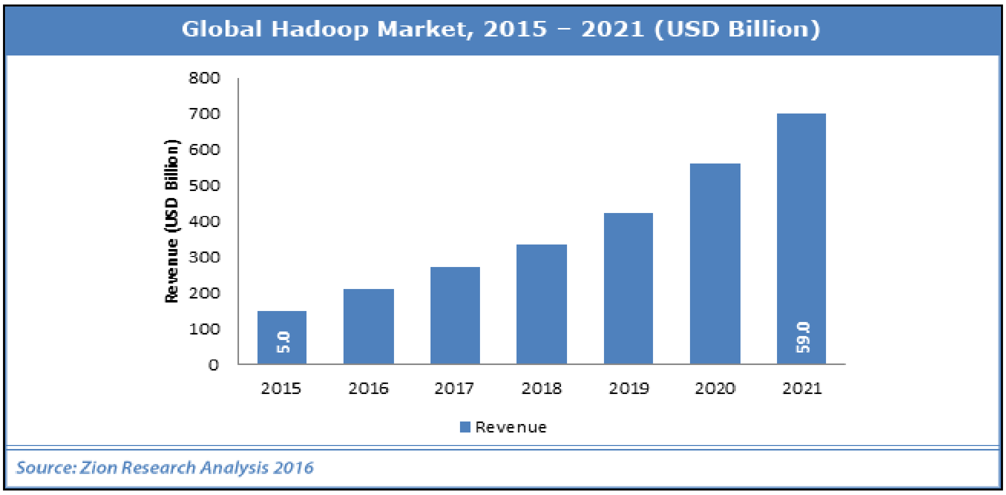


Fig 1: Conceptual architecture of big data analytics in Bioinformatics

V. HADOOP: Global Status

Hadoop plays a major role for enterprises over 5-6 years in handling the increasing data. Hadoop could be occupying in various sectors including distributed computing and security issues. Nonetheless, continuous development and a massive investment in the Hadoop technologies are expected to unfold new opportunities during the future years [21]. The main major global Hadoop market includes Amazon Web Services, Teradata Corporation, Cisco Systems, IBM Corporation, Cloudera, Inc., Datameer, Inc., Oracle Corporation, Hortonworks, Inc., Karmasphere, Inc., etc. many others. Fig 2 predicting the global Hadoop market will increase about $700 US billion by 2021.


Global Hadoop Market, 2015 – 2021 (USD Billion)

Source: Zion Research Analysis 2016

Fig 2: Global Hadoop market prediction.

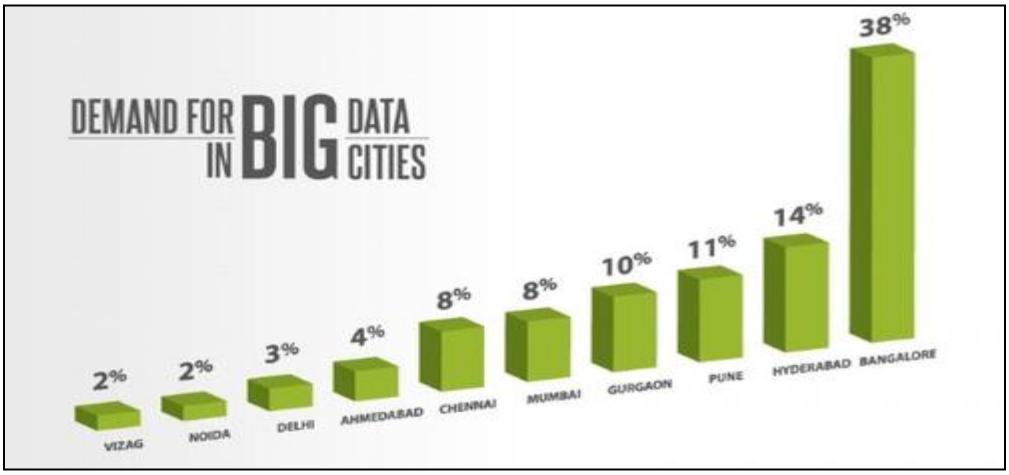


Fig 3: Demand for big data in Indian cities. [22]

VI. HADOOP Status In India

The potential growth for big data in India is mainly due to the growing number of industries trying to get meaningful results from their enormous data generation in their businesses. For Indian IT industry, the rising big data challenges can be overcome using Hadoop platform and could be future strategy to deal with big data complication. Bangalore is popularly known as the silicon valley of India having an excessive demand for big data analysis (fig 3). India ranks second a worldwide for farm productivity. E-Agriculture service data can be considered as a Big Data because of its variety in data with enormous volumes flowing with high velocity. Currently, HDFS, MapReduce, Hadoop, and storm are important tools solution use for e-agricultural data [23].

VII. Conclusion

Recent progress in molecular biology and genomics has led to a huge growth of digital biological information. Bioinformatics studies currently require processing of huge amounts of data with heavy computation. Hadoop is a versatile framework that can easily handle both approaches with high efficiency.

Acknowledgements

The authors would like to show his immense gratitude towards to mentors of Mahatma Gandhi Mission's Institute of biosciences and technology for giving such motivation and support.

References

[1] Vivekananth.P, Leo John Baptist.A. An Analysis of Big Data Analytics Techniques. *International Journal of Engineering and Management Research*. October-2015,*Volume-5, Issue-5,*

[2] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. *JOURNAL OF LATEX CLASS FILES*, Sepetember 2014, *Vol. 13, NO. 9*.

[3] Lin Dai, Xin Gao, Yan Guo, Jingfa Xiao and Zhang Zhang. Bioinformatics clouds for big data manipulation. *Biology Direct 2012*.

[4] EMDAD KHAN. Addressing Bioinformatics Big Data Problems using Natural Language Processing: Help Advancing Scientific Discovery and Biomedical Research. *Modern Computer Applications in Science and Education*. ISBN: 978-960-474-363-6.

[5] Divya Kumari,Ravi Kumar**.** Impact of Biological Big Data in Bioinformatics **.***International Journal of Computer Applications,* September 2014,(0975 – 8887).*Volume 101– No.11*.

[6] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald. A review of data mining using big data in health informatics . *Journal of Big Data*, 2014 1:2.

[7] R. J. Robison, How big is the human genome? *Precision Medicine*, January 2014.

[8] EMBL-European Bioinformatics Institute, *EMBL-EBI annual scientific report* 2013, 2014**.**

[9] *http://www.rcsb.org/pdb/home/home.do*. 26/11/2016.

[10] An Introduction to Open Source Software for Government IT.*UK Government ICT Strategy*.

[11] Peter J. A. Cock et.al. BioPython: freely available Python tools for computational molecular biology and bioinformatics. *Vol. 25 no. 11 2009*, pages 1422–1423.doi:10.1093/bioinformatics/btp163.

[12] Richard Holland et.al.BioJava: An Open-Source Framework for Bioinformatics. *Article in Bioinformatics* October 2008,*24(18):2096-7*.

[13] Hadoop release. *Apache.org Apache software foundation*. Retrieved 2016-11-27.

[14] Welcome to apache Hadoop. *hadoop.apache.org*. retrieved 2016-11-27.

[15] Ronald C Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *Taylor BMC Bioinformatics* 2010, 11(Suppl 12).

[16] Simone Leo, Federico Santoni, Gianluigi Zanetti. Biodoop: Bioinformatics on Hadoop. 2009. *International Conference on Parallel Processing Workshops .1530-2016/09*.

[17] M. C. Schatz. Cloudburst. Bioinformatics, June 2009*, 25(11):1363–1369*.

[18] Dai et al. Biology Direct 2012, 7:43. http://www.biology-direct.com/content/7/1/43.

[19] D. Field, B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston. Open software for biologists: from famine to feast. *Nature biotechnology,* 2006, *24(7):801–804*.

[20] Dillon T, Wu C, Chang E: Cloud Computing: Issues and Challenges. *Int Con Adv Info Net*. 2011, 27-33.

[21] *http://www.marketresearchstore.com/report/hadoop-market-z59712*. retrieved 2016/11/28

[22] https://www.linkedin.com/pulse/hadoop-jobs-india-it-skills-training-services. retrieved 2016/11/28.

[23] Rupika Yadav, Jhalak Rathod,Vaishnavi Nair, Big Data Meets Small Sensors in Precision Agriculture. *International Journal of Computer Applications,* 2015,*(0975 – 8887)*.

.