IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 2 (Sep. - Oct. 2013), PP 53-56
www.iosrjournals.org

Big Data in Bioinformatics & the Era of Cloud Computing

Prakash Nemade[1], Heena Kharche[2]
[1]*(Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal MP, India)*
[2]*(Department of Computer Science & Engineering, IES-IPS Academy, Indore MP, India)*

Abstract: *With the recent breakthrough in bioinformatics, demand of more storage space is increasing day by day. With this exponentially increasing data i.e. the big data of bioinformatics sector, the data needs to be handled in more flexible and cost effective manner. With this growth in the volume, variety and velocity of data, cloud computing promises to address big data issues and analysis of challenges of big data in bioinformatics. Big Data of bioinformatics consisting of various sequences (nucleotide or amino acid) which are now showing exponential growth because of high throughput experimental technologies. By the use of various bioinformatics tools demands for large and fast data storage technology is increasing. In this paper, security model is re-drafted for flexible use of model by consumers and various cloud based services and techniques are coined up to make an easy approach for implementing big data of bioinformatics using cloud.*

Keywords: *Big Data, Bio Cloud, Bioinformatics, Cloud, Cloud Computing, Secure Cloud*

I. INTRODUCTION

In general bioinformatics can be defined as “An interdisciplinary field that develops and improves upon method for storing, retrieving, organizing and analyzing biological data” [1].With the emergence of highly efficient bioinformatics technologies, the volume, variety and velocity of data is increasing with a great pace. As reported on February 2012, two nanopore sequencing platforms (GridION and MinION) are capable of delivering ultra-long sequencing reads (~100kb) with additionally higher throughput and much lower cost [2]. With this amount of data, it would be increasingly daunting for small scale industries to invest in a large data server that will be kept in back room, to store your data. Moreover big servers demands for highly skilled people to maintain the data which increases their operating costs. At present, a promising solution to address this challenge is cloud computing, which exploits the full potential of multiple computers and delivers computation and storage as dynamically allocated virtual resources via the Internet [3].

Bigdata in bioinformatics face various implementation challenges which hinders customer to shift their data onto the cloud despite of several advantages. Cloud Computing is here to stay, as it is proposed to transform the way IT is deployed and managed, promising reduced implementation, maintenance costs and complexity, while accelerating innovation, providing faster time to market, and providing the ability to scale high-performance applications and infrastructures on demand [4]. Features like Broad network access, resource pooling, pay as you go, rapid elasticity and on demand access has made cloud computing king of computing. With the use of various services and concepts of cloud computing one can easily handle the problems of big data in bioinformatics. With the better use of the services i.e Software as a service (SaaS), Platform as a service (PaaS) and Infrastructure as a service (Iaas) [4] we can easily access the data at affordable costs. Cloud computing basically uses three Deployment models namely Public cloud: Cloud sharing data of various enterprises on single storage server; Private Cloud: A dedicated cloud storage server of a particular enterprise and Hybrid Cloud: Combination of above two deployment models for easy and flexible use of information.

The bioinformatics big data opens up enormous possibilities for discovery, solutions, analysis and even cures. Biological big data are generally challenging for their variety, value, and critical need for veracity. The modern biological data covers diverse collection of omics field like genomics, proteomics, metabolomics, metagenomics, lipidomics and glycomics. The omic data is generated from high throughput experimental technologies. To store and analyze these data requires massive amounts of computational power of databases, data formats, software packages and pipelines [5]. Today, the competition is on for the solutions to analyze and store the data faster and easier and to bring cost down to the point where sequencing becomes easily available for a much wider market. Penta bytes of raw information can be used to provide hints for everything from preventing terabytes to shrinking healthcare costs – if we can figure out how to use the space efficiently.

This paper will be heading towards the constructive approach for saving bioinformatics big data using cloud computing and also the challenges that would hinder bioinformatics bigdata to take a step towards cloud. This paper is organized in the following sections:

In Section II, we give a background of big data in bioinformatics. We explore in more detail with the example of bigdata in bioinformatics.

In Section III, we describe issues of bioinformatics big data.

In Section IV, we describe the ways in which cloud will give the solutions to challenges states in Section III.

Finally in section V, we will give the conclusion showing importance of the suggested solutions and their extensions.

II. BIOINFORMATICS BIG DATA

As we move into the century, we stand at an inflection point in biology how we view and practice biology has forever changed. Biology is becoming data intensive as high throughput experiments are generating data to much greater pace as high throughput experiment technologies are becoming accessible to more number of scientists. One of the inflection points is Human Genome Project (HGP) which is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up human DNA, and of identifying and mapping the total genes of the human genome from both a physical and functional standpoint. [6]

Human Genome Project provided a genetics parts list and catalyzed the development of high throughput measurement tools (e.g. the high speed DNA sequences, DNA arrays, high throughput mass spectrometry, etc) and high throughput measurement strategies (e.g. yeast two hybrid technique for measuring protein protein interactions) as well as stimulating the development of powerful new computational tools for acquiring, storing and analyzing biological information. The HGP has catalyzed the view that biology is an informational science. Therefore, the huge amount of data generated by these high throughput experiments in the genomic and proteomic arenas requires that databases are built in a way that facilitates not just the storage of these data, but the efficient handling and retrieval of information from these huge databases. [7]

The use of DNA sequencing machines, which are smaller in size but capable of generating piles of data faster and at a lower cost, have changed science and medicine in ways never seen before [8]. The current era is beginning to look like the era of “big data”; a term refers to explosion of available information, which is byproduct of the digital revolution [9]. However, with biomedical data accumulating in computers and servers around the world [8], concerns over privacy and security of patient data are emerging.

Big data has affected several unrelated sectors in society including communications, media, medicine, scientific research among others [9]. However, even though both the computers and the internet have become faster, we have lack of computational infrastructure that is needed to securely generate, maintain, transfer and analyze large scale information in life sciences to integrate omics data with other data sets, such as clinical data from patients. Next Generation Sequencing (NGS) platforms that use semiconductors [10] or nanotechnology [11] have exponentially increased the rate of biological data generation in the last three years.

III. ISSUES IN BIOINFORMATICS BIG DATA

Big Data generation and acquisition gives birth to profound challenges for storage, transfer and security of the information. Even if companies were forced to limit their data collection and the storage space, still the big data analytics would be needed. Big data storage space would be needed by companies to store their data without any limits. Also, the computational time is needed to be decreased with the increase in the data for faster processing and efficient results. Having large storage and computational infrastructure can be difficult to maintain by the company moreover, implementing cost of this infrastructure may touch the sky.

Another challenge is to transfer the data from one location to another; it is mainly done either by the use of external hard disks or by mail. Transfer and accessing of this data becomes time consuming leading to decrease in processing time. Moreover, transfer of data may reduce work efficiency. Big data has to be processed and computed simultaneously so that we can get faster outputs which can be shared and used from any location the user wants to access. Transfer of big data may require large storage space transfer devices which can be time consuming and may sometime lead to inefficient results due to unavailability of updated results in the system.

Security and the privacy of data from individuals is also a concern. In every case the most important issue of handling bioinformatics big data is security of the data whether it is in the storage database or while transferring of data via external hard disks or email, security issue is the worry. Data has to be free from any threats as well as data integrity and security has to be maintained. In this growing era of technologies, authenticity and confidentiality of the data is also necessary for the safety of data.

IV. CLOUD AS SOLUTION TO BIOINFORMATICS BIG DATA

With the increased need to store data and information generated by big projects, an easy and flexible computational solution such as cloud based computing has emerged. Cloud computing is the only storage model that can provide the elastic scale needed for big data in the field of bioinformatics whose volume of output generation is increasing day by day. With the promising technology of cloud nothing is impossible. The issues that are stated above can be solved in one or the other ways described below.

1. Big Data Storage Space: With cloud computing one can use infrastructure that is data is now resided on the outsourced servers. Use of cloud infrastructure which falls under Infrastructure as a Service (one of the

cloud services) [4]. As infrastructure as a service, service of cloud provides you all the necessary resources like server, networking, storage space etc. to be used easily and efficiently, it has become very easy to store the big data. The cost to keep your data on cloud infrastructure is very affordable as cloud uses "pay as you go" feature. It also provides rapid capacity elasticity and easily deliverable whenever and wherever you want. Use of cloud infrastructure allows multi tenancy allowing you to share the data by multiple companies in smart and secure manner. You can automate regular practices—such as provisioning, backup, and replication—and integrate those practices as part of your cloud services, the better your environment will scale. Using this multi tenancy and automation feature a group of people working on the same project may simultaneously get the updated results provided by each other.

2. Data Transfer: Say no to all external hard drives to carry your data, as all your data is on cloud and you can access your data from anywhere anytime or you can share your data by using multi tenant nature of the cloud. Anyone who is authenticated and authorized to access that data will automatically get the updates data which you wanted to transfer. Cloud has reduced the fear of unavailability of data due to hardware failure during transfer to a great extent. Moreover, if you want to switch the data from one cloud service provider to another it can be easily done without any difficulty. Proper service level agreements are signed in order to ensure you the data security at cloud providers' storage space.
3. Security and Privacy: The major issue to move or store your big data in cloud is its security. Security solutions includes the use of better security systems with advanced encryption algorithms and proper signing of Service level agreements may secure your data to a higher level.

As the data is residing on third party storage server, one is very much concerned about his data. In order to trust that the cloud provider has kept your data safe on his premises one should use proper security measures from his side also. Trust can be easily implemented by using the digital certificates [12] with proper transport layer security, to completely rely on the cloud provider, as customers' data is now within their enterprise boundary as shown by Kharche H. et al. Security Model, Public Root CA and Enterprise CA can be implemented to secure the big data efficiently and effectively. Making the model little more flexible and easy to use by each and every individual according to their need the implementation flow has been re-drafted. The flowchart given in Figure 1consists of three phases:

Phase 1: In phase one, according to the needs of the user one can use any encryption algorithm with any key size fulfilling his requirement. Using any encryption algorithm as per user need makes the flow chart more flexible as now the expertise of the encryption algorithm used is in the enterprise boundary.

Phase 2: In phase two, the company should create and use his own digital certificates in order to sign the encrypted data resulted in phase 1. The private key generated during your own digital certificate generation should be kept confidential as no one else will be able to open your data now.

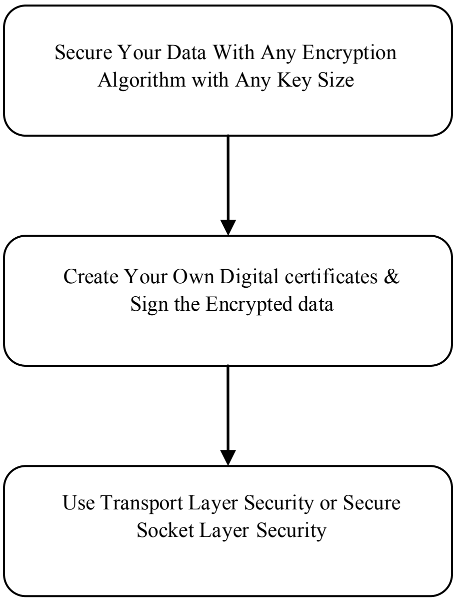
Secure Your Data With Any Encryption Algorithm with Any Key Size
Create Your Own Digital certificates & Sign the Encrypted data
Use Transport Layer Security or Secure Socket Layer Security

Figure 1: Flowchart to Implement Security and Trust

Phase 3: In phase three, the data is sent over the network in secured form. That is, one can use either of Secure Socket Layer (SSL) or Transport Layer Security (TLS) to send the whole data over the network in the secured format which makes the data triple secure.

Above suggested model is very easy and flexible to be used by any company with great ease with all the secrets of the data encryption and the private key in the enterprise boundary. Even though the data is resided at the cloud providers' storage space, he doesn't know about how data is secured at the enterprise level. This makes the above model more secure and easy to use.

With cloud computing, one can easily overcome all the challenges which were hindering the focus of customers to shift to cloud. All the challenges of storage space, transfer of data and security over cloud makes it very affordable for an individual to move towards cloud. Infrastructure maintenance is done at cloud providers' end, which reduces the cost of hiring technical expertise and the maintenance of the high data storage servers.

V. CONCLUSION

Two clear advantages can be gained by using cloud to store bioinformatics bigdata. One, It let companies to analyze massive datasets without capital investment in hardware and host the data internally on the server. Two, hosted model tends to abstract the complexity, enabling more immediate deployment of big data technology. One can easily use the cloud for the bioinformatics big data to large genomic sequences efficiently with proper security, processing speed and affordable costs. Using the re-drafted flow of the model individual can easily gain the trust on cloud. In this era of cloud computing one can easily store enormous amount of data with the easily scalability of resources in cloud Big Data leads to big challenges leading to proper arrangements to data storage in cloud effectively. Bioinformatics big data will need to implement scalable structure to deal with volumes of data generated by omics technologies. Advances in informatics to successfully address the big data challenges that to be faced in next decade should be adapted soon.

REFERENCES

[1] Wikipedia found at http://en.wikipedia.org/wiki/Bioinformatics

[2] Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. *Nat Biotechnol.2010;28(7):*691–693. doi: 10.1038/nbt0710-691.

[3] Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A,Stoica I, Above the Clouds: A Berkeley View of Cloud Computing. Berkeley: EECS Department, University of California; 2009.

[4] Kharche, Ms Heena, and Mr Deepak Singh Chouhan. "Building Trust In Cloud Using Public Key Infrastructure." *International Journal of Advanced Computer Science and Applications* 3.3 (2012).

[5] Higdon et al "Unraveling the complexities of life sciences data". Mary Ann Liebert Inc. Vol. 1, No. 1, Big Data.

[6] Wikipedia found at http://en.wikipedia.org/wiki/Human_Genome_Project

[7] Andreas D. Baxevanis, B.F. Francis Quellette *Bioinformatics: A practical guide to the analysis of genes and proteins*. (NJ: Wiley-Interscience, 2005).

[8] Costa FF ,Big Data In Biomedicine. *Drug Discovery Today*. In press (2013)

[9] *www.nytimes.com/2012/08/12/business/how-bigdat a-* became-so-big-unboxed.html?r=0

[10] Rothberg J. M. et al. "An integrated semiconductor device enabling non-optical genome sequencing": Nature 475(7356): 348–352(2011)

[11] Clarke J. et al. "Continuous base identification for single-molecule nanopore DNA sequencing": Nat Nanotechnol. (4): 265–70 (2009)

[12] Kharche, Ms Heena, and Mr Deepak Singh Chouhan. "Implementing Trust In Cloud Using Public Key Infrastructure." *International Journal of Engineering Inventions* 1.5 (2012).