International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I4P125
Volume 67 Issue 4, 128-130 April 2019

Original Article

Big Data Storage using Synthetic DNA

Shyna Sharma[1], Shruti Pathak[2], Taranjot Singh Kathuria[3], Tarun Sharma[4], Malvinder Singh Bali[5]

[1,2,3,4]*Students, Department of CSE, Chitkara University, Rajpura Campus, Punjab, India.*
[5]*Assistant Professor, Department of CSE, Chitkara University, Rajpura Campus, Punjab, India.*

Abstract **-** *Big data generally refers to the enormous amount of data created by people via various digital platforms. Through the help of analytics and big data, one can easily estimate the behavior and area of concern of a process. However, there are a lot of issues, such as cost association and storage of large data, that need to be identified and verified before one can implement the data. So, in this paper, a model has been proposed in which only one base can be changed in DNA, and the rest of the bases must be kept the same. Moreover, instead of mapping one bit to one base pair, bits are arranged in multidimensional matrices, and sets of molecules represent their locations in each matrix. This idea keeps all the important data and resources secure. Also, this method is capable enough to cater the problems like large storage, accessibility, and portability, wherein it also gives a brief idea about the composition of DNA. Hence, synthetic DNA storage can be an effective and innovative way to eradicate the issues. Also, this storage of DNA has very promising results.*

KeyWords **-** *Big Data, Synthetic DNA, Cloud Computing*

I. INTRODUCTION

The dictionary meaning of 'cloud' is a white or grey mass of condensed water vapor floating in the sky. Scientifically, it may be expressed as a large aggregation of objects that visually appear at a remote place, and these objects are not physically supervised, touched, or inspected [21]. High-capacity storage devices are being made for humans and many organizations like banking, life sciences, health care, retail, and government for their storage and database. The requirement for data storage is increasing day by day as a huge amount of data is generated daily. Total information in digital format in 2012 was about 2.7zettabytes [3]. The existing storage capacity is not enough to store such data, and the storage devices available will be exhausted one day. Every device maker is improving their technology to make their devices better, more reliable, portable, and cheaper, but today's technology is at its limit.

Huge data loss can occur due to storage device failure or accidental deletion of the data without backup. Device failure can occur due to manufacturer faults which leads to boot failure, high temperature of the system, which leads to a hard disk crash, power surges which can cause startup failure, and mechanical or internal failure as it is made of many fragile parts moving at a very high speed, corrupted files, human errors which include poor handling of the devices or accidental deletion of the files or accidentally dropping liquid on the device which can cause damage [8]. These devices need very favorable conditions and environments to work consistently, and unfavorable conditions can cause device failure.

The term big data means a vast amount of data collection that is continuously increasing at a rapid pace. The data is so large that an efficient storage system needs to be found. For this, the concept of big data storage in synthetic DNA is proposed. Big data can be categorized into three types: structured, unstructured, and semi-structured. The main characteristics of big data are Volume, Velocity, and Variety. Volume refers to the large amount of data that is generated every second. Velocity means the fast streams of data that travel here and there. Variety of data describes whether the data is a text file, audio, video, graph, etc. Big data is characterized by three aspects: (a) the data is numerous, (b) the data cannot be categorized into regular relational databases, and (c) data is generated, captured, and processed very quickly [9]. There are some major problems in healthcare data management systems like working with sensitive data, distributed data management under security and performance constraints, and specialized analytics to define the "physiological envelope" during each patient's daily life. Scientists have proposed a cloud-based framework that effectively manages health-related big data and benefits from the ubiquity of the internet and social media. The framework facilitates the mobile and desktop users by offering: (a) disease risk assessment services and (b) consultation services with the health experts on Twitter [12].

New problems usually arise with new technologies, as with big data. These problems are related not only to the 3 Vs. of big data but also to privacy and security.

Big data increases the scale of the problems related to privacy and security as faced in the traditional management of security and adds new ones that should be addressed with different techniquesandmeasures[15].

According to the US National Institute of Standards and Technology (NIST), ―Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or cloud provider interaction" [10]. Cloud computing combines a technology platform that provides hosting and storage service on the internet [10]. The most important benefit of cloud computing is that customers don't need to buy the resource from a
third-party vendor. Instead, they can use the resource and pay for it as a service, thus helping the customer save time and money. Cloud computing can provide infinite computing resources on demand due to its high scalability, which eliminates the need for Cloud service providers to plan far ahead on hardware provisioning. Many companies, such as Amazon, Google, Microsoft, and others, accelerate their paces in developing cloud computing systems and enhancing their services, providing to a larger number of users.[17]

The perception of different experts, providers, and professionals about cloud computing slightly differ. The National Institute of Standards and Technology (NIST) defined cloud computing as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [11].
The relationship between big data and cloud computing is based on integration. The cloud represents the storehouse, and the big data represents the product stored in the storehouse since it is impossible to create storehouses without storing any product in them [13].
The community cloud aspires to combine distributed resource provision from grid computing, distributed control from digital ecosystems, and sustainability from green computing with the use cases of cloud computing while making greater use of self-management advances from autonomic computing [18].

II. SYSTEM MODEL

Scientists and researchers have been trying to develop a different technology to solve the low storage problems, and hence, synthetic DNA is the new technology for data storage; it is extremely dense and can store data up to 1 exabyte/mm^3, and no electricity is required to operate and has a half-life of 500 years in the harsh environment [6]. The information stored in DNA can be recovered even after thousands of years, as long as the DNA is stored in dry, dark, and cold conditions. The storage process is done by encoding and decoding binary data to and from synthesized short DNA strands, and the data is stored in a long virtual DNA molecule. Short strands will allow for manipulating data easily. The four nucleotides of DNA used in the model are Adenine, denoted as A, Cytosine as C, Guanine as G, and Thymine as T [6].

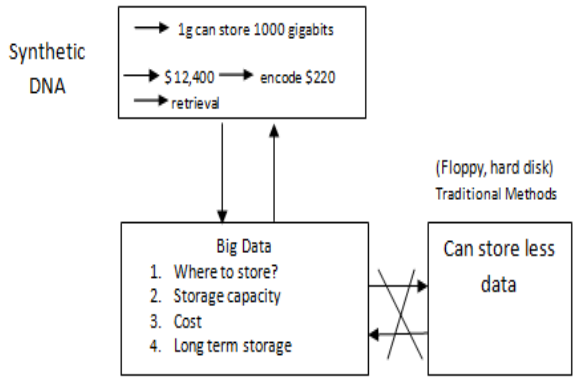

Fig. 1 Big Data in Synthetic DNA

The model in figure 1 shows the advantages of storing big data in synthetic DNA compared to other traditional methods.
For encoding and decoding, Huffman coding is being used. We have also figured out that the cost for encoding data in synthetic DNA is 1.55$ approximately per 1000 Gigabits (125 Gigabytes) of data. Also, the cost for decoding data from synthetic DNA is 0.0275$ approximately per 1000 Gigabits (125 Gigabytes) of data.
This model depicts how beneficial it is to store data in synthetic DNA.

III. CONCLUSION

Big Data is becoming a tremendous challenge for today's rapidly changing traditional markets when performing in-depth analysis. During each stage of the data life cycle, big data management is the most demanding issue. It consists of generation, acquisition, storage, and analysis [22]. The areas where big data is available are social media, mobile phone details, transactional data, financial statements, insurance forms, medical records, customer correspondence, RFID tags, weather information, Internet of Things (IoT), traffic patterns, communication events, etc. Cloud computing is another area in the IT field where different services like software, infrastructure, storage, etc., are offered online services [20].

This idea of storing a vast amount of data in synthetic DNA strands has proven to be a wonderful invention for humanity. It is incredible even to imagine that we can store the entire world's information in such a small space. It is possible to access this DNA stored data from a technological point of view but not economic. Moreover, the extraction of data stored in DNA is comparatively slower. Research is being done on how to resolve these issues. Besides this, DNA can survive in critical environmental conditions. Thus, this idea of big data storage in DNA will be of magnificent use to humankind.

Many key challenges in this domain, including automatic resource provisioning, power management, and security management, are only starting to receive attention from the research community. Therefore, we believe there is still tremendous opportunity for researchers to make groundbreaking contributions in this field and significantly impact their development in the industry [16].

IV. FUTURE SCOPE

This technology will make a drastic change within storage devices as it can fulfill the needs of big data storing companies. If we somehow increase the data transfer speed through synthetic DNA, it will help companies to overcome the problem of exhaustible storage devices. The latency rate between data transfers will also decrease if we can somehow increase the speed with which synthetic DNA will work and help us store a huge amount of data. Cloud computing is a new paradigm of computing utilities that promises to provide more flexibility, less expensive, and more efficiency in IT services to end-users [14].

In the future, the challenges are to overcome and make way for even more efficient use of the big data by the user in a cloud computing environment. It is very much needed that the computer scholars and IT professionals cooperate and make successful and long-term use of cloud computing and explore new ideas for using the big data in a cloud environment [19].

This will help the world in a fruitful way in which we will be able to store our big data in very less space.

REFERENCES

[1] Skourletopoulos, G., Mavromoustakis, C.X., Mastorakis, G., Batalla, JM, Dobre, C., Panagiotakis, S., and Pallis, E., 2017. Big data and cloud computing: a survey of the state-of-the-art and research challenges. In Advances in mobile cloud computing and big data in the 5G Era (pp. 23-41). Springer, Cham.

[2] Mukherjee, S. and Shaw, R., 2016. Big data–concepts, applications, challenges, and future scope. International Journal of Advanced Research in Computer and Communication Engineering, 5(2), pp.66-74.

[3] Church, G.M., Gao, Y. and Kosuri, S., 2012. Next-generation digital information storage in DNA. Science, 337(6102), pp.1628-1628.

[4] Schmidt, M. and Giersch, G., 2011. DNA synthesis and security. DNA microarrays, synthesis, and synthetic DNA, pp.285-300.

[5] Shah, S., Limbachiya, D. and Gupta, M.K., 2014. DNAcloud: A tool for storing big data on DNA. In 11th Annual Conference on Foundations of Nanoscience: Self-Assembled Architectures and Devices (FNANO14).

[6] Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G. and Strauss, K., 2016. A DNA-based archival storage system. ACM SIGARCH Computer Architecture News, 44(2), pp.637-649.

[7] Alam, J.R., Sajid, A., Talib, R. and Niaz, M., 2014. A review on the role of big data in business. International Journal of Computer Science and Mobile Computing, 3(4), pp.446-453.

[8] https://www.stellarinfo.com/blog/6-worst-reasons-of-hard-disk-failure/

[9] Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, M., Kamaleldin, W., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal, 2014.

[10] Nazir, M., 2012. Cloud computing: overview & current research challenges. IOSR Journal of computer engineering, 8(1), 14-22.

[11] Adamuthe, A.C., Salunkhe, V.D., Patil, S.H. and Thampi, G.T., 2015. Cloud Computing–A market Perspective and Research Directions. International Journal of Information Technology and Computer Science (IJITCS), 7(10), 42-53.

[12] Sabarmati, G., Chinnaiyan, R. and Ilango, V., 2016. Big data analytics research opportunities and challenges: a review. Int. J. Adv. Res. Comput. Sci. Softw. Eng, 6(10), pp.227-231.

[13] Zanoon, N., Al-Haj, A. and Khwaldeh, S.M., 2017. Cloud computing and big data are related to the two: a study. Int. J. Appl. Eng. Res, 12(17), pp.6970-6982.

[14] Subashini, S. and Kavitha, V., 2011. A survey on security issues in service delivery models of cloud computing. Journal of network and computer applications, 34(1), pp.1-11.

[15] Moreno, J., Serrano, M.A., Fernandez-Medina, E. and Fernandez, E.B., 2018. Towards a Security Reference Architecture for Big Data. In DOLAP.

[16] Zhang, Q., Cheng, L. and Boutaba, R., 2010. Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 1(1), pp.7-18.

[17] Kumar, S. and Goudar, R.H., 2012. Cloud computing-research issues, challenges, architecture, platforms, and applications: a survey. International Journal of Future Computer and Communication, 1(4), p.356.

[18] Mell, P. and Grance, T., 2011. The NIST definition of cloud computing.

[19] Venkatesh, H., Perur, S.D. and Jalihal, N., 2015. A study on the use of big data in the cloud computing environment. Int. J. Comput. Sci. Inf. Technol. (IJCSIT), 6(3), pp.2076-2078.

[20] Terzi, D.S., Terzi, R. and Sagiroglu, S., 2015, December. A survey on security and privacy issues in big data. In 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 202-207). IEEE.

[21] E. Bala Krishna Manash, T. Usha Rani "Cloud Computing - A Potential Area for Research ."International Journal of Computer Trends and Technology (IJCTT) V25(1):10-17, 2015. ISSN:2231-2803. www.ijcttjournal.org.

[22] K.Indira Gandhi, Sri.C.Sreedhar "Survey on Big Data: Management and Challenges ."International Journal of Computer Trends and Technology (IJCTT) V20(1):33-36, Feb 2015. ISSN:2231-2803. www.ijcttjournal.org.