THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Information Security Techniques to Prevent Malicious Activities like Data Breaches on Big Data Obtained for Analysis Purpose

Dhrumil Malani, Siddharth Lilani, Jimit Modi and Fenil Soni

B.Tech. Information Technology, SVKM's NMIMS

ABSTRACT

Data breach is a cyber-threat which is a successful cyber-attack with the motive of damage or gaining an unauthorized access to the information technology assets such as computer network, sensitive data or any form of intellectual property or any sort of personal data. It mainly focuses on the software bugs and vulnerabilities that are present in an application of which a hacker takes advantage of and avails the personal data.

Index Terms—Data Breach, Big Data, Prevention, Technique

I. INTRODUCTION

DATA BREACH may happen due to an innocent approach, severe damage can be done if the person steals and sells Personally Identifiable Information (PII) with unauthorized access. Data is considered breached even if it is viewed by an unauthorized person. There are various ways and people associated with this activity-

1. An accidental Insider
2. A Malicious Insider
3. Lost or Stolen Devices
4. Malicious Outside Actors

There are also various ways by which one can safeguard their personal social content which will be detailed at later stage.

Considering Big Data present at this moment, data breach may lead to severe consequences which may have adverse effect all over the world. There is a significant increase in the analysis of the Big Data as it has served many businesses and has proved beneficial to them, but it has always questioned about its security standards. Lack of laws and regulation is the major cause of data breaches. Any individual, organization or a state cannot be easily held responsible of such an activity. Data breach not only cause economical damage but also compromises with the reputation of an organization or an individual and may lead to even more serious repercussions. The data is generated in huge amount across countries, any social media or any sort of digital process produces it. Basically whenever analysis of Big Data is taken into consideration, it is bifurcated and defined into 3Vs- Volume, Variety and Velocity.

1) Volume: The growth of the Big Data is exponential in nature. Data being in the form of audio, video, images and text with suitable extensions make data measurable in Terabytes and at times Petabytes also. To handle this huge amount data the architecture and the applications designed to maintain such data must be re-evaluated
2) Variety: Big Data is available in all type of formats, raw data, structured and unstructured data, numeric data from old traditional databases, data generated from the line-of-business applications, unstructured text formats, video, audio, emails, data generated from financial transactions. Their must be proper format followed to merge, manage and govern these diverse natures of data.
3) Velocity: Data is being streamed in an unprecedented speed which must be dealt in timely manner. Sensors and RFID tags (smart barcodes are used as the tracking system to identify items), torrents of data must be dealt. Maintaining the data velocity is the biggest challenge face by most of the organizations.

ISSN:0971-1260

Vol-22-Issue-41-December-2019

There are additional 2 metrics of defining Big Data

1) Complexity: Data being random in nature with different extensions must be correlated with each other for analysis purpose. Data links can be easily spiral out of control.

2) Variability: Increasing velocities and variants of data, the flow of data is abnormal and inconsistent when measured at regular peaks.

This calls for an effective strategy to control flow of Big Data to ensure information security. This paper systemizes the knowledge required for detection and prevention of data breach on Big Data. Discussions of this paper may lead to some entirely different and new research related to protection of private data and will also focus on updated security standards that need to be used in every stage in flow process of Big Data.

STAGES OF BIG DATA

Acquisition of Data: This is the primary stage in which data is acquired from relevant sources where the data is generated exponentially. Most of the data is not suitable for analysis purpose and must be discarded. This kind of sorting may be challenging. Redundant data is either superimposed or merged with other relevant data. Due to strong interconnection between devices through World Wide Web, data is constantly being collected and stored

3) Extraction of Data: As discussed previously the redundant data must be discarded. This act can be challenging. Taking an example of CCTV camera the stable or still footage will consistently have same faces repeated over and over again with is undesirable and must be removed or eliminated from analysis perspective. But when it comes to heart beat reading one must not remove redundant frames as it may prove to be important aspect of analysis. This kind of scrutinization must be done to obtain desired results in analysis report.

4) Collection of Data: Data from isolated single source is enough for analysis purpose and making predictions, it must be from various and needs to be of similar type to obtain best possible outcomes. Health report generator must take reading from all possible sources such as thermometer, heart-rate sensor, sphygmomanometer and pedometer accurate results. Convergence of data is always considered an important aspect of processing.

5) Structuring of Data: When all the data is aggregated, it is crucial to present it in a structured format and must be stored for future use. Query processing becomes easier on data which is in structed format. With evolution of data bases, query can be processed of unstructured data using NoSQL which is widely used for Big Data analysis. But following method is inefficient in generating real-time results, here the aggregated structured data is more reliable over unstructured data.

6) Visualization of Data: After structuring the data, query processing is enforced on the data which will draw visual representation of the data in justified format. Instead of showing tables which contains figures the visualization imparts more better understanding of data. For instance, considering the popularity of Facebook and its usage, the image below displays usage pattern of Facebook.

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Fig 1. Big Data Visualization

II. NEED FOR SECURITY IN BIG DATA

For the purpose of research and predictions, many business uses Big Data analysis which lack in fundamental assets majorly from a security perspective. The challenge faced is detection and prevention of advance threats from intruders with malicious intensions, must be solved. There are techniques which are used to detect threats in early stages which include sophisticated patterns analysis and multiple data source analyzing. Apart from the security, there are data privacy challenges for existing industries and for federal organizations which must be solved. With increase in the usage of Big Data in business organizations many firms are affected or are victims of privacy issue. The complaints of data breach is statistically displayed in the graph below-

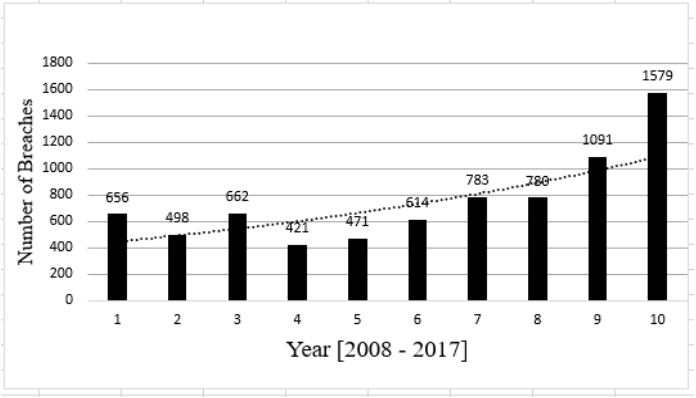

Fig 2. Complaints of Data Breach

Another graph shows the records of data breach exposed (in millions) for the duration of last ten years (2008 -2017).

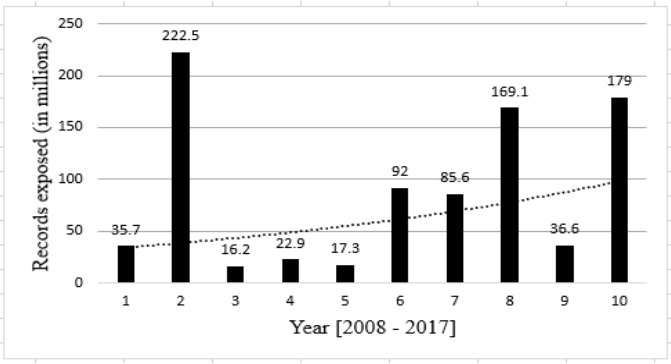

Fig 3. Records exposed (in millions)

ISSN:0971-1260

Vol-22-Issue-41-December-2019

III. TECHNIQUES TO PREVENT BREACHING OF DATA

There are various information security techniques which can be implemented to safeguard the private data. These practices must be always followed while handling the Big Data. These methodologies also avoid the data breach at personal level.

A. PATCHING AND UPDATING THE SOFTWARE:

Patching vulnerabilities and updating computer software is extremely important, especially taking into considerations the most successful malicious attacks on data exploiting well-known vulnerabilities containing patches. By observing these kinds of attacks IT professionals must enforce the strong patch management strategy that verifies security and functionality when patches are appended to existing operation systems.

Software developers consistently deliver patches of software products in order to apply regular updates to the systems. Introducing patches will result in new functionalities of the system. For instance, Microsoft Windows often introduces updates by stating new functionalities of the existing features. This process is often done to safeguard the operation systems from malicious attacks. Even Hadoop has a strong patch management which promotes safety of the Big Data from getting accessed by unauthorized person or organization. Tools like auto approval, reporting and scheduling are the innovative patch management functions which provides the platform helping companies against malicious attacks.

B. ENCRYPTION OF BIG DATA:

Big Data is majorly obtained from various heterogeneous sources. Encryption is basically associated with unreadability of data. The unauthorized parties cannot read the data even after successfully breaching it. There are various algorithms which are used to convert the data into unreadable form. The algorithms used are as follow:

1) Triple DES (Data Encryption Standards) Algorithm
2) RSA (Public key encryption)
3) Blowfish
4) Twofish
5) AES (Advanced Encryption Standard)

There are methods like Quantum Key Distribution which are emerging methods which sends keys embedded to photons in a fiber optic. There are tools like *Cryptographic tools* which are used to secure a large amount of data, majorly Big Data. When the data has reached to the destination finally decryption process must be done. The steps to secure Big Data through encryption are as displayed in the diagram.

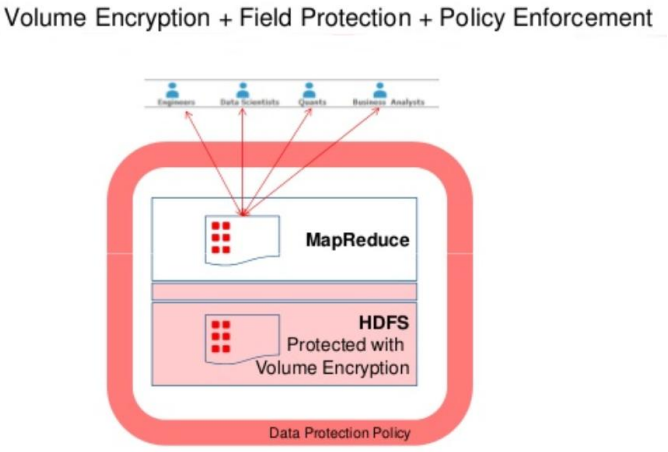

Fig 4. Encryption

ISSN:0971-1260

Vol-22-Issue-41-December-2019

C. USER AUTHORIZATION AND ACCESSIBILITY

Data must be protected right from the beginning than to handing it randomly to anonymous person. Whenever a naive user handle data, one must prove his/her identity to gain access to the data of the system. To prevent malicious intruder from gaining the access to private sensitive data there must be different layer to authentication which must include complex passwords containing all types of the characters present. Implantation of more secure security feature like smart cards and fingerprint should be present. Ultimately these activities lead to elimination of malicious unauthorized parties gaining access to private data.

D. VIRTUAL BARRIERS:

These barriers are designed and developed with the purpose of restricting the data when it moves within the networks. For instance Firewalls acts as the barriers which filters the incoming data and the restricted data is only accessed by the authentic user. Transport Layer Security (TLS) and Secure Sockets Layer (SSL) are also implemented with the purpose of authentic communication throughout the Internet, it reduces eavesdropping by a malicious unauthorized party. It works on the principle of handshaking which is done on both the sides, client and server.

To avoid breaching these technologies must be constantly monitored and must be updated frequently so that unauthorized parties cannot adapt to the existing software.

IV. ACKNOWLEDGEMENT

The author is very grateful to Prof. Kapil Nagwanshi, Department of Information Technology, Mukesh Patel School of Technology Management and Engineering, Shirpur for his encouragement in research work in the field of Big Data and its security and his guidance in process of paper writing.

V. CONCLUSION AND ITS FUTURE SCOPE

Changing needs of the business usage and analysis of Big Data is increasing exponentially and it is believed to have precise outcomes and predictions which are realistic in nature. Due to excessive usage of social media across the world the data produced is exponentially high, by 2022 the data generate per hour will be quadrupled. The data will be available will be measured in Exabytes and may be even in Zettabytes. The databases containing this large amount data will face new problems related to the security. There must be upgrade in the policies and the existing method must be re-modified to handle such a large amount data. Introduction of smart devices in the world would also accelerate the data generation process. Redundant data will also increase which must be handled and sorted properly. Excessive usage of devices will also add on to data generation which is raw data. Filtering of this data is must for effective usage of available data. More data generation will be boon to the simulation and analytics industries as they will be able to draw more accurate conclusion out of it, ultimately reducing the cost of production of certain good or service. By 2025, the analytics tools will get doubled due to immense use of big data throughout the world. There will also be reduction product development cost. The applications involving artificial intelligence will be boosted at cheaper approach. Even the software industries developing security application will get a boost and privacy issues will keep on increasing day by day with increase in data generation due to excessive use of social media in upcoming years. Proper bifurcation and aggression of similar kind of data is will be biggest challenge faced by the industries dealing with big data. Structuring algorithms must be modified or upgraded to handle such large amount of data. These modifications are must to deal with large amount data in future. Data analytics, data science, machine learning, artificial intelligence courses will be added as predefined courses in the curriculum of Computer Science and Information Technology branches. Taking future in consideration Big Data analysis is immense scope of research and will also open various types of employment in the computer science and information technology fields.

VI. REFERENCES

[1] Charles Schmitt, “Security and Privacy in the Era of Big Data (A RENCI/National Consortium for Data Science WHITE PAPER),” iRODS, a Technological Solution to the Challenge of Implementing Security and Privacy Policies and Procedures

ISSN:0971-1260

Vol-22-Issue-41-December-2019

[2] Shashank,S.K.Saravanan, Information Security in Big Data using Encryption and Decryption, Department of Computer Applications, Valliammai Engineering College, SRM Nagar, Kattankulathur-603203 .

[3] Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath, "Big Data Security Issues and Challenges" International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015)

[4] Dona Sarkar, Asoke Nath, "Big Data – A Pilot Study on Scope and Challenges", International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS, ISSN: 2371-7782), Volume 2, Issue 12, Dec 31, Page: 9-19(2014)..

[5] Ariel Hamlin, Nabil Schear, Emily Shen, Mayank Varia, Sophia Yakoubov, Arkady Yerukhimovich "Cryptography for Big Data Security", Book Chapter for Big Data: Storage, Sharing, and Security (3S) published in December 17,2015.

[6] Conrel Bradford, "5 Common Encryption Algorithms and the Unbreakables of the Future," article July 31.

[7] Liu, S., Schulze, J. P., Herr, L., Weekley, J. D., Zhu, B., VanOsdol, N., Plepys, D.M., & Wan, M. (2011). CineGrid Exchange: A workflow-based peta-scale distributed storage platform on a high-speed network. Future Generation Computer Systems, 27 (7), 966-976

[8] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. SP-12 Proceedings of the 2012 IEEE Symposium on Security and Privacy, pp. 300-314. Washington, DC, USA: IEEE Computer Society [Accessed July 22, 2013]

[9] Venkata Narasimha Inukollu, Sailaja Arsi, Srinivasa Rao Ravuri SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[10] https://www.sisense.com/glossary/big-data-security/

[11] http://airccse.org/journal/nsa/6314nsa04.pdf

[12] https://safenet.gemalto.com/data-encryption/big-data-security-solutions/

[13] https://i-sight.com/resources/data-breach-prevention/