A REVIEW: PRIVACY PRESERVATION IN BIG DATA

Victor Sodeinde & Emmanuel Ayodele
Department of Computer Science, Federal Polytechnic, Ilaro, Ogun state, Nigeria
victor.sodeinde@federalpolyilaro.edu.ng; emmanuel.ayodele@federalpolyilaro.edu.ng

Abstract

An immense measure of significant information is created and is opening up for examination because of headways in smart digital actual frameworks. The information comes from different sources, like medical services, shrewd homes, and wise vehicles, and frequently incorporates private, possibly delicate data that requirements suitable disinfection prior to being delivered for investigation. Distributed computing gives adaptable framework and high stockpiling limit with regards to Big Data applications. The gradual and quick nature of information age in these frameworks requires adaptable protection safeguarding systems with high security and utility. Notwithstanding, security conservation frequently comes to the detriment of information utility. The three significant privacy preservation techniques are information anonymization, notice and assent, and differential protection. In this paper, we examine these security protection techniques for huge information and how differential protection is an improved and answer for large information protection.

Keyword: Anonymization, Big data, Data privacy, Differential privacy, Notice and consent

Introduction

One of the cutting-edge technologies bringing about transformation in the field of data analytics is big data. It has the ability to reveal information about the hidden facets of data analysis. When a significant amount of diverse data is produced quickly, it is referred to as "big data." It is now a resource that may be exploited to improve the intelligence and profitability of enterprises. Big data is produced from a variety of sources, including internet transactions, search queries, mobile devices, emails, and movies. It is kept through partitioning among different servers. This suggests that there are numerous massive data sources and storage systems spread throughout the internet. Security concerns are a significant problem for big data analytics because it heavily relies on the web model. Big data analytics that are not secure could result in significant losses for both individuals and corporations. While exposing all of the big data may produce useful analytics, it can also present significant security risks. For hackers and intruders, the value buried in massive data might be quite valuable. Therefore, there is a trade-off between big data security and availability. One must make sure that the two are properly balanced. Therefore, it's important to safeguard big data analytics during the entire process.

Traditional Data Privacy Preservation Methods

The term "cryptography" refers to a collection of methods and formulas used to secure data. Various encryption algorithms are used in cryptography to transform plaintext into cipher text. This methodology is the foundation for many different techniques, including public key cryptography and digital signatures. Common cloud computing and big data services demand privacy, which cryptography cannot enforce (Van Dijk & Juels, 2010). This is due to the three V's of big data, which set it apart from conventional large data collections (velocity, variety, volume). Large data architecture differs from conventional information architectures because of these characteristics of big data.

Cryptography and conventional encryption techniques are no longer scalable to meet the privacy requirements of big data due to these changes in architecture and its complicated nature. The issue with encryption is the all-or-nothing nature of data retrieval (Karle & Vora, 2017). The user is not permitted access to the less sensitive data that can be valuable in big data analytics because it is likewise encrypted. Data becomes inaccessible to individuals without access to the decryption key. Additionally, if data is stolen before encryption or cryptographic keys are utilized improperly, privacy may be violated. The privacy of huge data can also be protected with attribute-based encryption (Kim et al, 2013). This approach to big data security is built on associations between big data attribute values. Depending on the nature of big data and corporate policies, it is determined which qualities need to be secured. To put it simply, encryption or cryptography cannot stand on its own.

Privacy Preservation for Big Data

Big data privacy has sparked significant concerns, making effective privacy preservation techniques necessary. The methods of data anonymization, notice and consent, are covered in this section

Data Anonymization

The process of modifying data that will be used or released so that the identification of important information is prevented is known as data anonymization (Sedayao, 2012). It's also known as data de-identification at times. This technique conceals crucial secret data in a way that protects data privacy. Data is anonymized before dissemination to the public by organizations. In this context, anonymization often refers to the concealment of identifying attributes (attributes that specifically and individually identify individuals), such as full name, license number, voter id, etc. The primary issue with data anonymization is that although data may appear anonymous, re-identification is simple when it is connected to other external data (Sweeney, 2002).

K-Anonymity

K-anonymity is a property possessed by certain anonymized data. Given person-specific field- structured data; produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals. K-anonymity can be achieved by using suppression and generalization. Suppression: In suppression, an asterisk (*) is used in place of some property values. A column's values could be substituted with * whole or in part. Generalization: In generalization, a broader category is used to replace an attribute's specific values.

L-Diversity

L-diversity is a form of group-based anonymization. L variety is used to preserve. By using generalization and suppression approaches, the k-anonymity model, an extension of the l-diversity model, decreases the granularity of data representation so that each record in the data maps onto at least k other records. The k-anonymity approach, which maintains that protecting identities down to the level of k-individuals does not necessarily equate to protecting the associated sensitive values that were suppressed or generalized, especially when the sensitive values within a group exhibit homogeneity, is also partially addressed by the l-diversity approach.

T-Closeness

The t-closeness strategy was developed to further l-diversity by retaining the distribution of sensitive fields in addition to the distribution of values given the availability of attacks where sensitive attributes may be deduced based on the distribution of values for l-diverse data. If the difference between the distribution of a sensitive attribute in this class and the distribution of the attribute throughout the entire table is less than a threshold t, the equivalence class is said to have t-closeness.

Notice and Consent

The most common privacy preservation strategies for web services are notice and consent. Every time a user accesses a new application or service, a privacy warning is displayed. With this approach, a person can better safeguards his own rights. It makes the individual accountable for protecting their privacy (Cate & Mayer-Schönberger, 2013). When dealing with enormous amounts of data, this strategy has several drawbacks. The majority of the time, when notice and consent are given, applications of enormous amounts of data are either unexpected or unknown. Because of this, whenever big data is used for a new purpose, the notification needs to be updated.

Differential Privacy Differential

Privacy is a technique that gives researchers access to databases holding personal data while providing robust individual privacy protections (Salido, 2012). When searching the data, it seeks to reduce the possibility of individual identification. Figure 1 depicts the method of differential privacy. Data is not changed under differential privacy, unlike anonymization. The database is not accessible to users directly. An interface is used to compute the results and

add the desired errors. It act as a firewall. These errors are significant enough to safeguard privacy, but little enough to maintain the value of the information provided to analysts and researchers.

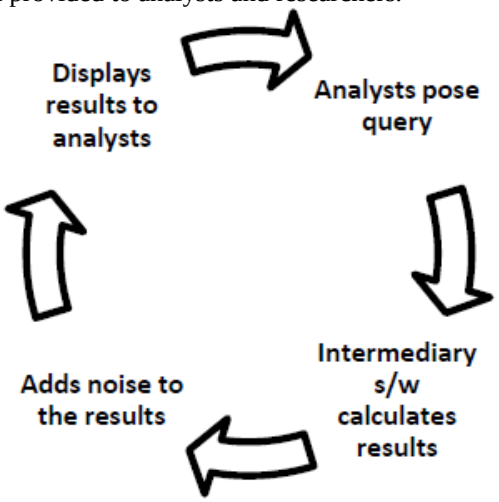


Fig. 1. Differential privacy process
Note. This model was produce in 2014, summarizing the differential privacy process. From "Privacy preservation in big data" by Gosain, A., & Chugh, N. (2014). Privacy preservation in big data. International Journal of Computer Applications, 100(17).and c

Conclusion

Big data privacy has grown in importance as a result of its direct connection to consumers. Today, a company must guarantee privacy when using big data analytics. Instead of concentrating on data collection, privacy protections should now concentrate on data use. Regarding the size and unanticipated uses of big data, they should be changed. When used with huge data, techniques like anonymization have limited promise. Customers are also responsible for guaranteeing privacy using the notice and consent method. Differential privacy may be considered a workable big data privacy option. Without really altering the data, it might guarantee privacy when changed and used to huge data.

Reference

Cate, F. H., & Mayer-Schönberger, V. (2013). Notice and consent in a world of Big Data. International Data Privacy Law, 3(2), 67-73.

Chakravorty, A., Wlodarczyk, T., & Rong, C. (2013, May). Privacy preserving data analytics for smart homes. In *2013 IEEE Security and Privacy Workshops* (pp. 23-27). IEEE.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. In *2013 International conference on collaboration technologies and systems (CTS)* (pp. 48-55). IEEE.

Gosain, A., & Chugh, N. (2014). Privacy preservation in big data. *International Journal of Computer Applications*, *100*(17).

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013, June). A discussion of privacy challenges in user profiling with big data techniques: The EEXCESS use case. In *2013 IEEE international congress on big data* (pp. 25-30). IEEE.

Karle, T., & Vora, D. (2017, February). PRIVACY preservation in big data using anonymization techniques. In *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)* (pp. 340-343). IEEE.

Kim, S. H., Eom, J. H., & Chung, T. M. (2013, June). Big data security hardening methodology using attributes relationship. In *2013 International Conference on Information Science and Applications (ICISA)* (pp. 1-2). IEEE.

Liu, W., Uluagac, A. S., & Beyah, R. (2014, April). MACA: A privacy-preserving multi-factor cloud authentication system utilizing big data. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 518-523). IEEE.

Salido, J. (2012). Differential privacy for everyone. White Paper, Microsoft Coproration.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

Sedayao, J., & Enterprise Architect, I. I. (2012). Enhancing cloud security using data anonymization. *White Paper, Intel Coporation*, 17.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, *10*(05), 557-570.