The Use of De-identification Methods for Secure and Privacy-enhancing Big Data Analytics in Cloud Environments

Gloria Bondel, Gonzalo Munilla Garrido, Kevin Baumer and Florian Matthes

Chair for Software Engineering for Business Information Systems, Faculty of Informatics, Technical University of Munich, Boltzmannstrasse 3, Garching, Germany

Keywords: Security, Privacy, Big Data Analytics, Cloud Environments.

Abstract: Big data analytics are interlinked with distributed processing frameworks and distributed database systems, which often make use of cloud computing services providing the necessary infrastructure. However, storing sensitive data in public clouds leads to security and privacy issues, since the cloud service presents a central point of attack for external adversaries as well as for administrators and other parties which could obtain necessary privileges from the cloud service provider. To enable data security and privacy in such a setting, we argue that solutions using de-identification methods are most suitable. Thus, this position paper presents the starting point for our future work aiming at the development of a privacy-preserving tool based on de-identification methods to meet security and privacy requirements while simultaneously enabling data processing.

1 PROBLEM STATEMENT

Smartphones, social media, and Internet-of-Things (IoT) are just some of the technical developments which lead to the digitalization of our day-to-day life. The utilization of these new technologies results in the creation of a vast amount of data, so-called big data, which is collected by the organizations providing digital services and products. To generate useful information from the collected data, organizations use different analytics approaches. These analytics approaches enable the testing of hypotheses and the identification of patterns, ultimately allowing organizations to design better products and services, enhance the user experience, or optimize internal processes.

Organizations often use traditional business intelligence approaches to analyze data. Business intelligence approaches transform data into a predefined structure and store it on a central server (i.e., a data warehouse) for future processing. However, for big data, these business intelligence approaches are no longer applicable due to the properties of big data, which are availability in large quantities, requires quasi-real-time processing, and comprises many different types of structured and unstructured data (Laney, 2001). Instead, approaches for big data analytics based on distributed data processing have emerged. Distributed processing frameworks enable faster data processing by storing data in a distributed file system and moving the processing activities into the data nodes, thereby enabling parallelization.

Distributed data processing is storage intensive, making a large computational infrastructure necessary. Therefore, big data analytics and cloud computing, offering efficient management and reduced cost of IT infrastructures, are often associated (Hashem et al., 2015; Liu et al., 2015; Ma et al., 2013). Nowadays,cloud providers even offer special products for distributed big data processing, e.g. Cloudera[1], Amazon EMR[2], Microsoft Azure HDInsight[3], or IBM BigInsights[4].

However, big data processing combined with cloud computing leads to security and privacy concerns (Liu et al., 2015; Stergiou and Psannis, 2017). Even if the data is encrypted while being transferred over a network, the data needs to be stored in the cloud in plaintext to enable data processing. This leads to two major security issues. First of all, the cloud presents a central point of attack for external attackers. For example, a hacker could manage to penetrate the cloud and gain access to sensitive data.

[1] https://www.cloudera.com/

[2] https://aws.amazon.com/de/emr/

[3] https://azure.microsoft.com/de-de/services/hdinsight/

[4] https://www.ibm.com/support/knowledgecenter/SSPT3X_4.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/c0057605.html

Bondel, G., Garrido, G., Baumer, K. and Matthes, F.
The Use of De-identification Methods for Secure and Privacy-enhancing Big Data Analytics in Cloud Environments.
DOI: 10.5220/0009470903380344
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 2*, pages 338-344
ISBN: 978-989-758-423-7

The second security concern originates from within the cloud service provider's (CSP) organization. Administrators of the CSP have certain privileges to the data stored in the cloud, which is necessary to perform maintenance activities as well as to prevent misuse of provided resources (Li et al., 2013). However, these administrative privileges can be abused for personal benefits, as shown in an incident in 2010, when Google had to fire a key engineer after breaking into the Gmail and Google Voice accounts of several children (Krazit, 2010). Furthermore, the cloud provider could give access to cloud resources to third parties, e.g., to government entities for reasons of legal prosecution.

Overall, the risks of cloud computing inhibit the use of big data analytics (Liu et al., 2015; Li et al., 2013), which hinders the realization of high potentials. Therefore, several approaches that address security and privacy concerns in the context of big data analytics in cloud environments exist. However, these approaches reduce processing performance, are often described in a very abstract manner, do not contain precise instructions for action, or are outdated. Thus, it is difficult for organizations to get an overview of current approaches, their advantages, and their disadvantages.

Therefore, a persisting problem in big data analytics can be formulated as follows:

How can data be protected in big data cloud environments while enabling a maximum of processing functionality and minimizing performance constraints as well as utility loss?

In this context, the conflicting goals of security and privacy, on the one hand, and preserving the utility of data to enable analytics, on the other hand, should be evaluated (Tomashchuk et al., 2019; Isley, 2018). As a first step to address this problem, we identified five approaches for secure big data analytics in cloud environments. These approaches are homomorphic encryption, partial encryption in combination with trusted hardware or partial encryption in combination with trusted client/hybrid cloud, de-identification, and privacy-preserving cloud architecture. After analyzing the advantages and disadvantages of these approaches, we argue that de-identification is a promising approach since it enables a multitude of analysis functionalities while simultaneously realizing security and privacy objectives. This assumption is additionally confirmed by the emergence of several commercial tools that aim at implementing de-identification for data processing in cloud environments.

However, to the best of the authors' knowledge, a holistic view of the impact of different de-identification methods on the trade-off between security and privacy versus data analytics capabilities does not exist. Therefore, in our future work, we aim at analyzing this trade-off on the use cases of data generated by wearables and vehicles. Furthermore, we aim at developing a privacy-enhancing tool for the easy application of different de-identification methods.

In the following, we will first present a short definition and delineation of the terms security and privacy in section 2. This is followed by an overview of existing approaches to secure big data analytics in cloud environments including their advantages and disadvantages in section 3. This section also covers a short rational why we deem de-identification methods as most suitable approach. Afterwards, in section 4, we present existing approaches using de-identification methods from research and practice. Finally, we conclude with a conclusion in section 5.

2 FOUNDATIONS

In this section we will introduce the concepts of security and privacy as well as their interrelation.

We adapt the definition of security provided by (Fink et al., 2018), who defines security as *"[...] a set of measures to ensure that a system will be able to accomplish its goal as intended, while mitigating unintended negative consequences"*. Thus, security aims to prevent vulnerabilities of software and hardware, making it resilient against malicious attacks, natural disasters, unplanned disruptions, and the unintended use of computational resources (Hurlburt et al., 2009).

In general, privacy can be defined as *"[...] freedom from observation, disturbance, or unwanted public attention [...]"* (Fink et al., 2018). However, to make the term privacy more actionable in the context of computer science, the threat-based definition of privacy provided by (Wu, 2012) is adopted: *"[Privacy] is defined not by what it is, but by what it is not - it is the absence of a privacy breach that defines a state of privacy"*. Thus, privacy is about identifying and characterizing relevant privacy threats as well as protecting information against these threats (Wu, 2012; Solove, 2015).

Different views on the relationship between the terms security and privacy exist (Hurlburt et al., 2009). In most cases, security and privacy are interpreted as overlapping concepts. The overlapping area between security and privacy is often referred to as information security. Information security aims at protecting different kinds of information and data from destructive forces and unwanted actions (Mukherjee et al., 2015). The three principles confidentiality, integrity, and availability, are known as the CIA triad that characterizes information security (Fink et al., 2018; Domingo-Ferrer et al., 2019). These principles do not only support and shape the theoretical understanding of information security, but they are also often used as a basis for defining privacy rules and for protecting electronic health information (Samonas and Coss, 2014).

However, in other cases, privacy is interpreted as an aspect of security (Hurlburt et al., 2009). This results from the observation that some security methods have a direct effect on privacy (Fink et al., 2018).

Summarizing, both - privacy and security - have in common that they are concerned with the appropriate use and protection of information. However, the concepts vary concerning the scope and rationale of the protection (Fink et al., 2018; Hurlburt et al., 2009).

3 EXISTING APPROACHES

In this section, we will first present existing approaches for big data analytics as well as their advantages and disadvantages. Fig. 1 presents a schematic overview of the different approaches. Due to the tremendous potentials of big data for organizations, several approaches to solving security issues of big data analytics in the cloud have been proposed in the past:

Homomorphic Encryption. This approach describes the use of an encryption scheme, which allows the processing of encrypted data (see Fig. 1, a). Fully homomorphic encryption would allow arbitrary processing of encrypted data, but as of today, no such encryption scheme exists. However, scientific literature presents several partially homomorphic algorithms, allowing to perform limited functionality on encrypted data, e.g., data aggregation (Paillier, 1999; Castelluccia et al., 2005; Lu et al., 2012), cosine similarity (Lu et al., 2014), and order-preserving search (Agrawal et al., 2004). Due to the limited functionality as well as losses in processing performance, homomorphic encryption schemes are not relevant in practice.

Partial Encryption in Combination with Trusted Hardware. The approach is based on the integration of a trusted hardware device into the public cloud infrastructure. The trusted device runs as an autonomous compute element that the cloud administrator cannot access (Bajaj and Sion, 2014) (see Fig. 1, b). First, the user of this approach splits the data into sensitive and not sensitive data, which is uploaded to the public cloud in ciphertext and plaintext, respectively. The processing of not sensitive data is done directly in the public cloud. However, if the sensitive data needs to be processed, it is first transferred onto the trusted hardware, where it is decrypted, processed, encrypted, and sent back to the cloud. The public cloud passes the resulting plaintext and the ciphertext on to the client, who has to decrypt the processing results of the sensitive data and merge it with the results of the not sensitive data. This way, the sensitive data is never in the public cloud without being encrypted. Since the trusted hardware is integrated into the cloud infrastructure, this approach can only be implemented by the CSP, which requires trust that the trusted hardware is not compromised. Furthermore, splitting a dataset into two parts and processing each part individually leads to performance losses. Again, several examples are presented in research (Bajaj and Sion, 2014; Eguro and Venkatesan, 2012; Arasu et al., 2013; Pires et al., 2016), but few implementations exist in practice, e.g., Intel SGX[5].

Partial Encryption in Combination with Trusted Client/Hybrid Cloud. Similarly to partial encryption in conjunction with trusted hardware, the data set is split into sensitive and non-sensitive data. However, this time, the sensitive data is not sent to the public cloud but stored and processed in a private cloud (Hacigümüş et al., 2002; Zhang et al., 2013) (see Fig. 1, c). Disadvantages of this approach are that the client has to maintain one or more local servers after all. Moreover, performance issues due to inter-site communication arise.

De-identification of Data. De-identification methods are approaches that make it difficult to restore the link between an individual and his or her data by removing or transforming specific data points (Kushida et al., 2012). Before uploading the data to the public cloud, it is sanitized using different de-identification methods, e.g., pseudonymization, generalization, character masking, or suppression (see Fig. 1, d). The risk of using de-identification techniques is that ”re-identification attacks” can be launched to identify specific individuals (Kushida et al., 2012). Privacy models provide a means for measuring the likelihood of re-identification attacks and thus for defining different levels of privacy (Tomashchuk et al., 2019). Privacy models are, for example, k-anonymity (Sweeney, 2002), l-diversity (Machanavajjhala et al., 2006), t-closeness (Li et al., 2007), and differential privacy (Dwork, 2006). However, re-identification attacks can still circumvent privacy models by linking more data to the anonymized dataset, creating uniqueness of each entry, and therefore establishing a link

[5] https://www.intel.de/content/www/de/de/architecture-and-technology/software-guard-extensions.html

back to an individual's identity.

Privacy Preserving Cloud Architecture. The cloud architecture presented by (Li et al., 2013) removes control rights of the provider, ensuring that the CSP can not access any dataset stored in the cloud (see Fig. 1, d). However, CSPs want to keep their control rights to prevent misuse of their cloud infrastructures. Similar approaches are proposed by (Pacheco et al., 2017; Jr. et al., 2016)

All the solutions presented have certain advantages and disadvantages. One of our underlying assumptions is that the CSP is not trustworthy. Thus, we can exclude the approach of partial encryption in combination with trusted hardware. Also, we omit the approach of a cloud architecture that preserves data privacy, since to the best of the authors' knowledge, currently no cloud provider implements such an architecture. The limited functionality and reduced processing performance of homomorphic encryption hamper the exploitation of big data analytics. Thus, we do not consider homomorphic encryption for secure and privacy-enhancing cloud computing any further. Although the approach of partial encryption in combination with a Trusted Client / Hybrid Cloud does not realize many of the advantages of using public clouds, this approach is not completely excluded. For example, small amounts of data, such as keys used in encryption, can be stored on on-premise databases.

In summary, in our future work, we will focus on de-identification methods to implement privacy for big data analytics in cloud environments. This approach enables a multitude of functionalities while simultaneously realizing security as well as privacy objectives and does not require the client to trust the CSP. Besides, depending on the security and privacy requirements, a higher or lower level of security and privacy can be achieved by the selected de-identification methods, which makes it possible to implement different use cases. Finally, it is also possible to assess the risk of re-identification attacks using privacy models, and thus to evaluate the choice of de-identification methods and, if necessary, adjust them accordingly.

4 RELATED WORK

In this section, we present existing approaches to security and privacy for big data analytics in cloud environments, as presented in research and practice.

Many scientific publications emphasize the relevance of secure big data analytics in cloud environments, e.g., (Liu et al., 2015; Stergiou and Psannis, 2017; Zissis and Lekkas, 2012; Neves et al., 2016). However, focussing on solutions based on de-identification methods, the number of publications is significantly smaller. The use of encryption for secure big data analytics in the cloud, which is not homomorphic encryption, is limited to the use of the Advanced Encryption Standard (AES) (Sachdev and Bhansali, 2013). There are also examples for generalization of data, such as (Prasser et al., 2017; Wan et al., 2015; Dankar et al., 2012; Prasser et al., 2016), that mainly deal with health data and are affiliated with the product ARX Data Anonymization Tool.

A concrete context in which the de-identification of data plays a vital role is in the area of personal health data (PHI) in the USA. In 2002, the revised version of the HIPAA Privacy Rule was adopted, which sets national standards for the protection of medical records and other personal health information (U.S. Department of Health and Human Services, nN). Under HIPPA, personal health information, if de-identified as required, may be used and disclosed for any purpose.

The Safe Harbor Model, not to be confused with the entirely different and currently non-existent EU Safe Harbor, specifies what data must be removed or generalized. Several initiatives automate HIPAA de-identification, such as ZIPpy Safe Harbor De-Identification Macros for SAS (Chatfield and Parker, 2018). However, HIPAA Safe Harbor de-identification methods are limited and have been shown to be insufficient for protection against re-identification attacks (Sweeney et al., 2017).

In the meantime, however, other tools emerged that were originally developed for the purpose of implementing HIPAA Safe Harbor, but now also offer more extensive de-identification methods. These include Privacy Analytics Eclipse[6], ARX Data Anonymization Tool[7], IQVIA[8] or the Google Cloud Healthcare API[9]. However, all of these products still have a strong focus on health data and therefore are often based on standards for interoperability of health data (e.g. FHIR[10], DICOM[11]).

Other products that offer the de-identification of data commercially are Anonos[12], IBM InfoSphere

[6] https://privacy-analytics.com/software/privacy-analytics-eclipse

[7] https://arx.deidentifier.org

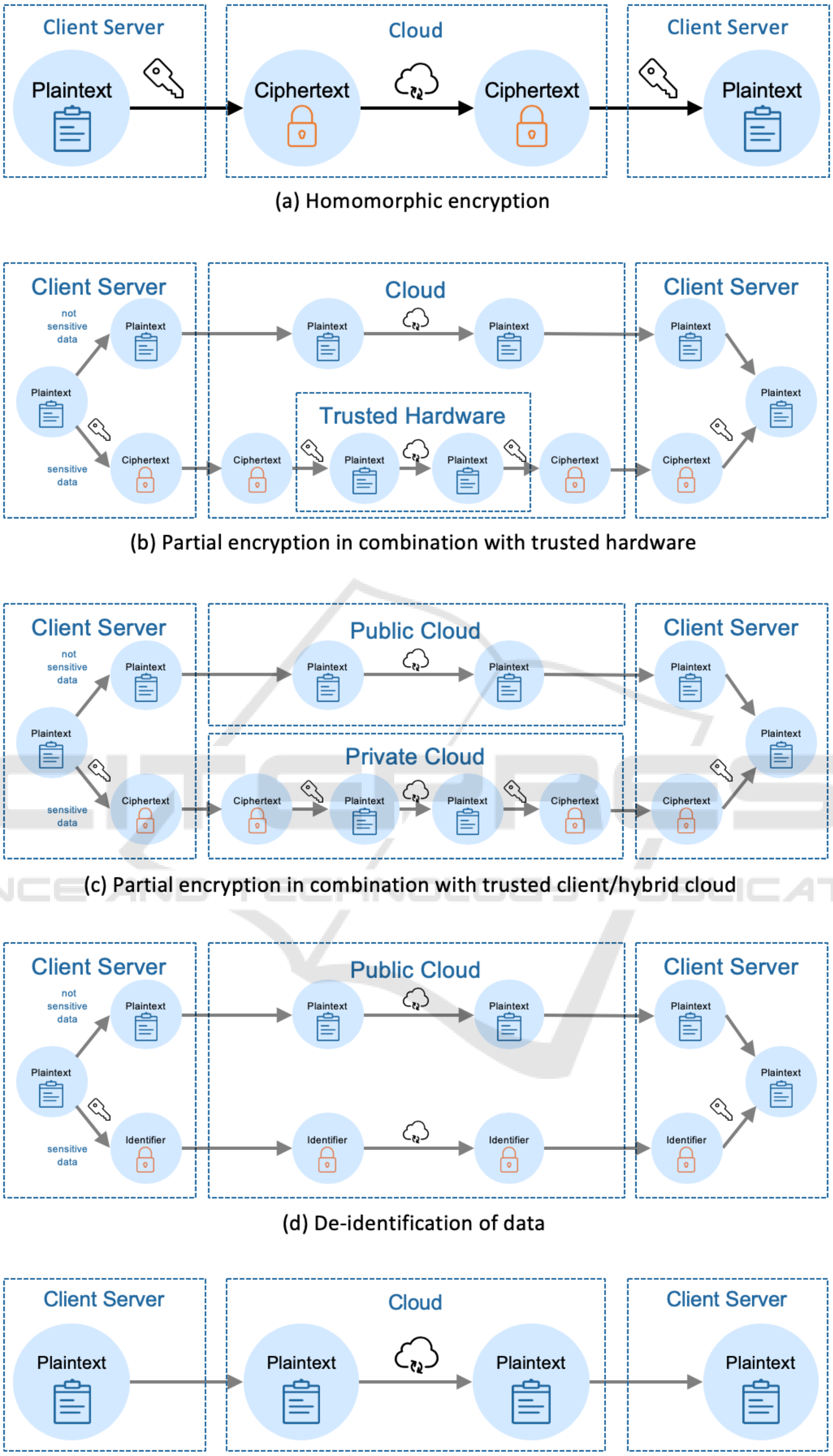
[8] https://www.iqvia.com/solutions/real-world-value-and-outcomes/privacy-preservation-and-data-linkage

[9] https://cloud.google.com/healthcare

[10] https://www.hl7.org/fhir/index.html

[11] https://www.dicomstandard.org/dicomweb/restful-structure

[12] https://www.anonos.com



(a) Homomorphic encryption

(b) Partial encryption in combination with trusted hardware

(c) Partial encryption in combination with trusted client/hybrid cloud

(d) De-identification of data

(e) Privacy preserving cloud architecture

Figure 1: Overview of existing approaches for privacy-securing big data analytics in cloud environments.

Optim Data Privacy[13], Data Sunrise[14] and Privitar[15]. However, these products offer only a single, or a limited set of de-identification methods (e.g., dynamic pseudonymization of Anonos, character masking of Data Sunrise) and only selected databases can be used as Big Data infrastructure.

5 CONCLUSION

In this position paper, we address a persisting problem in big data analytics, which is concerned with the trade-off between protecting security and privacy while at the same time enabling analysis functionality. We present existing approaches for preserving the security and privacy of big data analytics in cloud environments and argue that de-identification provides the most promising approach. This assumption is supported by the emergence of commercial tools for enabling security for cloud environments applying de-identification methods. However, to the best of the authors' knowledge, no holistic analysis of the trade-off between different de-identification methods and analysis functionality currently exists. Therefore, our future work aims at analyzing this trade-off in a use case focusing on data generated by wearables as well as on vehicle-generated data. These results will provide the basis for the implementation of a privacy-enhancing tool applying different de-identification methods.

ACKNOWLEDGMENTS

This work was sponsored by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 / X-DACE. The responsibility for the content of this publication lies with the authors.

REFERENCES

Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2004). Order preserving encryption for numeric data. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 563–574, New York, NY, USA. ACM.

Arasu, A., Blanas, S., Eguro, K., Kaushik, R., Kossmann, D., Ramamurthy, R., and Venkatesan, R. (2013). Orthogonal security with cipherbase. In *6th Biennial Conference on Innovative Data Systems Research (CIDR'13)*.

Bajaj, S. and Sion, R. (2014). Trusteddb: A trusted hardware-based database with privacy and data confidentiality. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):752–765.

Castelluccia, C., Mykletun, E., and Tsudik, G. (2005). Efficient aggregation of encrypted data in wireless sensor networks. In *The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, pages 109–117.

Chatfield, A. and Parker, Jessica ans Egeler, P. (2018). Zippy safe harbor de-identification macros. In *SAS Conference Proceedings: SAS Global Forum 2018*.

Dankar, F. K., El Emam, K., Neisa, A., and Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making*, 12(1):66.

Domingo-Ferrer, J., Farràs, O., Ribes-González, J., and Sánchez, D. (2019). Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges. *Computer Communications*, 140-141:38–60.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag.

Eguro, K. and Venkatesan, R. (2012). Fpgas for trusted cloud computing. In *22nd International Conference on Field Programmable Logic and Applications (FPL)*, pages 63–70.

Fink, G. A., Song, H., and Jeschke, S., editors (2018). *Security and privacy in cyber-physical systems: Foundations, principles, and applications*. Wiley IEEE Press, Hoboken, NJ, first edition edition.

Hacigümüş, H., Iyer, B., Li, C., and Mehrotra, S. (2002). Executing sql over encrypted data in the database-service-provider model. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, SIGMOD '02, pages 216–227, New York, NY, USA. ACM.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47:98 – 115.

Hurlburt, G. F., Miller, K. W., Voas, J. M., and Day, J. M. (2009). Privacy and/or security: Take your pick. *IT Professional*, 11(4):52–55.

Isley, P. (2018). Iso/iec 20889 first edition 2018-11: Privacy enhancing data de- identification terminology and classification of techniques. Standard.

Jr., E. C. B., Monteiro, J. M., Reis, R., and Machado, J. C. (2016). A flexible mechanism for data confidentiality in cloud database scenarios. In *Proceedings of the 18th International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 359–368. INSTICC, SciTePress.

Krazit, T. (2010). Google fired engineer for privacy breach. https://www.cnet.com/news/google-fired-engineer-for-privacy-breach/. website, online, accessed 13 July 2018.

[13] https://www.ibm.com/us-en/marketplace/infosphere-optim-data-privacy

[14] https://www.datasunrise.com/data-masking

[15] https://www.privitar.com

Kushida, C., Nichols, D., Jadrnicek, R., Miller, R., Walsh, J., and Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50 Suppl:S82–101.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group, Garnter.

Li, M., Zang, W., Bai, K., Yu, M., and Liu, P. (2013). Mycloud: Supporting user-configured privacy protection in cloud computing. In *Proceedings of the 29th Annual Computer Security Applications Conference*, ACSAC '13, pages 59–68, New York, NY, USA. ACM.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115.

Liu, C., Yang, C., Zhang, X., and Chen, J. (2015). External integrity verification for outsourced big data in cloud and iot: A big picture. *Future Generation Computer Systems*, 49:58 – 67.

Lu, R., Liang, X., Li, X., Lin, X., and Shen, X. (2012). Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications. *IEEE Transactions on Parallel and Distributed Systems*, 23(9):1621–1631.

Lu, R., Zhu, H., Liu, X., Liu, J. K., and Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4):46–50.

Ma, M., Wang, P., and Chu, C. (2013). Data management for internet of things: Challenges, approaches and opportunities. In *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pages 1144–1151.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24.

Mukherjee, J., Datta, B., Banerjee, R., and Das, S. (2015). Dwt difference modulation based novel steganographic algorithm. In Jajodia, S. and Mazumdar, C., editors, *Information systems security*, Lecture Notes in Computer Science, pages 573–582. Springer, Cham and Heidelberg and New York and Dordrecht and London.

Neves, P. C., Schmerl, B. R., Cámara, J., and Bernardino, J. (2016). Big data in cloud computing: Features and issues. In *IoTBD*.

Pacheco, L., Alchieri, E., and Solis, P. (2017). Architecture for privacy in cloud of things. In *Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 2: ICEIS,*, pages 487–494. INSTICC, SciTePress.

Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. In Stern, J., editor, *Advances in Cryptology — EUROCRYPT '99*, pages 223–238, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pires, R., Pasin, M., Felber, P., and Fetzer, C. (2016). Secure content-based routing using intel software guard extensions. In *Proceedings of the 17th International Middleware Conference*, Middleware '16, pages 10:1–10:10, New York, NY, USA. ACM.

Prasser, F., Kohlmayer, F., and Kuhn, K. A. (2016). The importance of context: Risk-based de-identification of biomedical data. *Methods of information in medicine*, 55(04):347–355.

Prasser, F., Kohlmayer, F., Spengler, H., and A Kuhn, K. (2017). A scalable and pragmatic method for the safe sharing of high-quality health data. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1.

Sachdev, A. and Bhansali, M. (2013). Enhancing cloud computing security using aes algorithm. *International Journal of Computer Applications*, 67:19–23.

Samonas, S. and Coss, D. (2014). The cia strikes back: Redefining confidentiality, integrity and availability in security. *Journal of Information System Security*, Volume 10(3):21–45.

Solove, D. J. (2015). The meaning and value of privacy. In Roessler, B. and Mokrosinska, D., editors, *Social Dimensions of Privacy*, pages 71–82. Cambridge University Press, Cambridge.

Stergiou, C. and Psannis, K. E. (2017). Efficient and secure big data delivery in cloud computing. *Multimedia Tools and Applications*, 76(21):22803–22822.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., and Brody, J. G. (2017). Re-identification risks in hipaa safe harbor data: A study of data from one environmental health study. *Technology science*, 2017.

Tomashchuk, O., van Landuyt, D., Pletea, D., Wuyts, K., and Joosen, W. (2019). A data utility-driven benchmark for de-identification methods. In Gritzalis, S., Weippl, E. R., Katsikas, S. K., Anderst-Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Trust, Privacy and Security in Digital Business*, volume 11711 of *Lecture Notes in Computer Science*, pages 63–77. Springer International Publishing, Cham.

U.S. Department of Health and Human Services (n.N.). Summary of the HIPAA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html. accessed 16. Juli 2019.

Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., Ganta, R., Heatherly, R., and Malin, B. A. (2015). A game theoretic framework for analyzing re-identification risk. *PloS one*, 10(3):e0120592.

Wu, F. T. (2012). Defining privacy and utility in data sets. *84 University of Colorado Law Review 1117 (2013); 2012 TRPC*, pages 1117–1177.

Zhang, J. Y., Wu, P., Zhu, J., Hu, H., and Bonomi, F. (2013). Privacy-preserved mobile sensing through hybrid cloud trust framework. In *2013 IEEE Sixth International Conference on Cloud Computing*, pages 952–953.

Zissis, D. and Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3):583 – 592.