A Review on Big Data Privacy and Security

Y Rajeswari, C Sreekanya, Dr. K Venkata Ramana

Abstract— Big data is a convergence of new hardware and algorithms that allow us to discover new patterns in large data sets-patterns we can apply to making better predictions and, ultimately, better decisions. Today, organizations are putting Big Data into practice in such diverse fields as healthcare, smart cities, energy and finance.. Big data can be characterized by 3 V's i.e., Volume, Velocity, Variety. Big data analytics is the term used to describe the process of researching massive amounts of complex data in order to reveal hidden patterns or identify secret correlations. However, there is an obvious challenge between the security and privacy of big data and the widespread use of big data. This paper focuses on privacy and security concerns in big data, differentiates between privacy and security and privacy and requirements in big data. There have been a number of privacy-preserving mechanisms developed for privacy protection at different stages (for example, data generation, data storage, and data processing) of a big data life cycle. The goal of this paper is to provide a major review of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. This paper also presents recent techniques of privacy preserving in big data like differential privacy, privacy preserving big data publishing. This paper refer privacy and security aspects healthcare in big data. Comparative study between various recent techniques of big data privacy is also done as well.

Keywords—Bid data, Privacy and security, Privacy preserving.

1 INTRODUCTION

Big Data [1, 2] which consists of both structured and unstructured data which is in bigger in size is flooded to data servers by organizations in terms of billions of bytes on day-to-day basis. Due to recent technological development, the amount of data generated by internet, social networking sites, sensor networks, healthcare applications, and many other companies, is extremely increasing day by day. All the expansive measure of data produced from various sources in multiple formats with very high speed [3] is referred as big data. The term big data [4, 5] is defined as "a new generation of technologies and architectures, designed to economically separate value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery and analysis". On the premise of this definition, the properties of big data are reflected by 3V's, which are, volume, velocity and variety. Thus, veracity, validity, value, variability, venue, vocabulary, and vagueness were added to make some complement explanation of big data [6]. A common theme of big data is that the data are diverse, i.e., they may contain text, audio, image, or video etc. This differing qualities of data is signified by variety. In order to ensure big data privacy, various mechanisms have been developed in recent years.

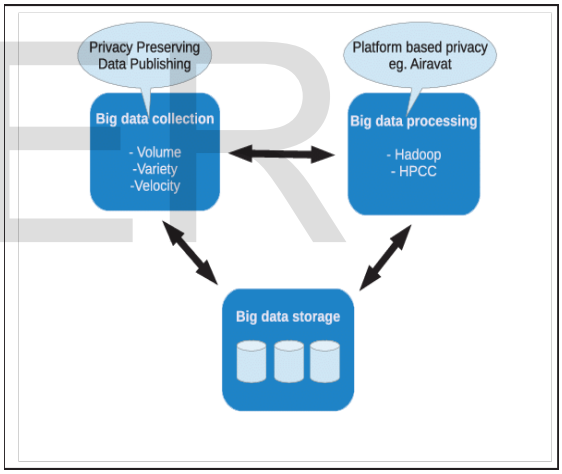


Fig.(1): Big data life cycle stages, i.e., data generation, storage, and processing are shown above

To handle various measurements of big data in terms of volume, velocity, and variety, there is need to design efficient and effective frameworks to process expansive measure of data arriving at very high speed from various sources. Big data needs to experience multiple phases during its life cycle.

Smart energy big data analytics is also a very complex and challenging topic that share many common issues with the generic big data analytics. Smart energy big data involve extensively with physical processes where data

intelligence can have a huge impact to the safe operation of the systems in real-time [7]. This can also be useful for marketing and other commercial companies to grow their business. As the database contains the personal information, it is vulnerable to provide the direct access to researchers and analysts. Since in this case, the privacy of individuals is leaked, it can cause threat and it is also illegal. "Privacy and security concerns" section discusses of privacy and security concerns in big data and "Privacy requirements in big data" section covers the Privacy requirement in big data. "Big data privacy in data generation phase", "Big data privacy in data storage phase" and "Big data privacy preserving in data processing" sections discusses about big data privacy in data generation, data storage, and data processing Phase."Recent Techniques of Privacy Preserving in Big Data" section presents some recent techniques of big data privacy and comparative study between these techniques.

2 PRIVACY AND SECURITY CONCERNS IN BIG DATA

Privacy and security is an important issue. Big data security model is not suggested in the event of complex applications due to which it gets disabled by default. However, in its absence, data can always be compromised easily. As such, this section focuses on the privacy and security issues.

- *Y Rajeswari is currently pursuing Master of Computer Applications in KMMIPS, Tirupathi, PH-0877-2289100. E-mail:rajeswarirajiy@gmail.com*
- *C Sreekanya is currently pursuing Master of Computer Applications in KMMIPS, Tirupathi, PH-0877-2289100. E-mail:sreekanya78cheluri@gmail.com*
- *Dr K Venkata Ramana, Head of the Department Master of Applications in KMMIPS, Tirupathi, PH-0877-2289100. E-mail:ramanakv4@gmail.com*

Privacy Information privacy is the privilege to have some control over how the personal information is collected and used. Information privacy is the capacity of an individual or group to stop information about themselves from becoming known to people other than those they give the information to.

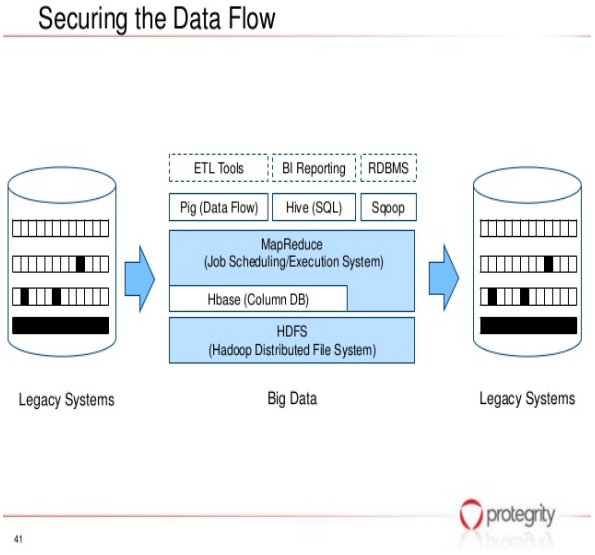
Security Security is the practice of defending information and information assets through the use of technology, processes and training from:- Unauthorized access, Disclosure, Disruption, Modification, Inspection, Recording, and Destruction.

Privacy vs. security Data privacy is focused on the use and governance of individual data—things like setting up policies in place to ensure that consumers' personal information is being collected, shared and utilized in appropriate ways. Security concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit [8]. While security is fundamental for protecting data, it's not sufficient for addressing privacy.

3 PRIVACY REQUIREMENTS IN BIG DATA

Big data analytics draw in various organizations; a hefty portion of them decide not to utilize these services because of the absence of standard security and privacy protection tools. These sections analyse possible strategies to upgrade big data platforms with the help of privacy protection capabilities.

Businesses and government agencies are generating and continuously collecting large amounts of data. The current increased focus on substantial sums of data will undoubtedly create opportunities and avenues to understand the processing of such data over numerous varying domains. But, the potential of big data come with a price; the users' privacy is frequently at danger. Ensures conformance to privacy terms and regulations are constrained in current big data analytics and mining practices. Developers should be able to verify that their applications conform to privacy agreements and that sensitive information is kept private regardless of changes in the applications and/or challenges privacy regulations. To address these, identify a need for new contributions in the areas of formal methods and testing procedures. New paradigms for privacy conformance testing to the four areas of the ETL (Extract, Transform, and Load) process as shown in Fig. 2 [9, 10].



3.1 Pre-hadoop Process Validation

This step does the representation of the data loading process. At this step, the privacy specifications characterize the sensitive pieces of data that can uniquely identify a user or an entity. Privacy terms can likewise indicate which pieces of data can be stored and for how long. At this step, schema restrictions can take place as well.

3.2 Map-reduce Process Validation

This process changes big data assets to effectively react to a query. Privacy terms can tell the minimum number of returned records required to cover individual values, in addition to constraints on data sharing between various processes.

3.3 ETL Process Validation

Similar to step (2), warehousing a logical bases for a course of action should be confirmed at this step for concurrence with privacy terms. Some data values may be aggregated unspecified or excluded in the warehouse if that indicates high probability of identifying individuals.

3.4 Reports Testing

Reports are another form of questions, imaginable with higher visibility and wider audience. Privacy terms that characterize 'purpose' are fundamental to check that sensitive data is not reported with the exception of specified uses.

4 BIG DATA PRIVACY IN DATA GENERATION PHASE

Data generation can be classified into active data generation and passive data generation. By active data generation, we mean that the data owner will give the data to a third party [11], while passive data generation refers to the circumstances that the data are produced by data owner's online actions (e.g., browsing) and the data owner may not know about that the data are being gathered by a third party. Minimization of the risk of privacy violation amid data generation by either restricting the access or by falsifying data.

4.1 Access Restriction

If the data owner thinks that the data may uncover sensitive information which is not supposed to be shared, it refuse to provide such data. If the data owner is giving the data passively, a few measures could be taken to ensure privacy, such as anti-tracking extensions, advertisement or script blockers and encryption tools.

4.2 Falsifying Data

In some circumstances, it is unrealistic to resist access of sensitive data. In that case, data can be twist out of shape using certain tools prior to the data received by some third party. If the data are twist out of shape, the true information cannot be easily revealed. The following techniques are utilized by the data owner to falsify the data:

- A tool Socket puppet is utilized to hide online identity of individual by deception. By utilizing multiple Socket puppets, the data belonging to one specific individual will be regarded as having a place with various people. In that way the data collector will not have enough knowledge to relate different socket puppets to one individual.

- Certain security tools can be used to mask individual's identity, such as Mask Me. This is especially useful when the data owner needs to give the credit card details amid online shopping.

5 BIG DATA PRIVACY IN DATA STORAGE PHASE

Storing high volume data is not a major challenge due to the advancement in data storage technologies, for example, the boom in cloud computing [12]. If the big data storage system is compromised, it can be exceptionally disorderly as individuals' personal information can be disclosed [13]. In distributed environment, an application may need several datasets from various data centres and therefore confront the challenge of privacy protection.

The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes [14]. Responding to the 3V's nature of the big data analytics, the storage infrastructure ought to be scalable. It should have the ability to be configured dynamically to accommodate various applications. One promising technology to address these requirements is storage virtualization, empowered by the emerging cloud computing paradigm [15]. Storage virtualization is process in which numerous network storage devices are combined into what gives off an impression of being a single storage device. SecCloud is one of the models for data security in the cloud that jointly considers both of data storage security and computation auditing security in the cloud [16].

6 BIG DATA PRIVACY PRESERVING IN DATA PROCESSING

Big data processing paradigm categorizes systems into batch, stream, graph, and machine learning processing [17]. For privacy protection in data processing part, division can be done into two phases. In the first phase, the goal is to safeguard information from unsought disclosure since the collected data might contain sensitive information of the data owner. In the second phase, the aim is to extract meaningful information from the data without violating the privacy.

7 RECENT TECHNIQUES OF PRIVACY PRESERVING IN BIG DATA

7.1 Differential Privacy

Differential Privacy [18] is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. The distraction introduced is large enough so that they protect the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some techniques have been used to protect the privacy, but proved to be unsuccessful.

Differential Privacy (DP) deals to provide the solution to this problem as shown Fig. 3. In DP analyst are not provided the direct access to the database containing personal information. An intermediary piece of software is introduced between the database and the analyst to protect the privacy. This intermediary software is also called as the privacy guard.

Step 1 The analyst can make a query to the database through this intermediary privacy guard.

Step 2 The privacy guard takes the query from the analyst and evaluates this query and other earlier queries for the privacy risk. After evaluation of privacy risk.

Step 3 The privacy guard then gets the answer from the database.

Step 4 Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst.

The amount of distortion added to the pure data is proportional to the evaluated privacy risk. If the privacy risk is low, distortion added is small enough so that it do not affect the quality of answer, but large enough that they protect the individual privacy of database. But if the privacy risk is high then more distortion is added.

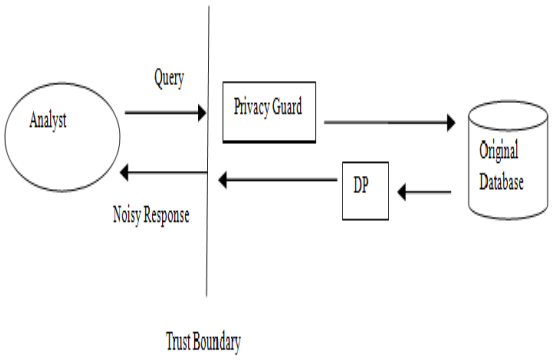


Fig. 3 Differential privacy big data differential privacy (DP) as a solution to privacy-preserving in big data is shown

8 PRIVACY PRESERVING BIG DATA PUBLISHING

The publication and distribution of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time. Privacy-preserving models broadly fall into two different settings, which are referred to as input and output privacy. Much of the work in privacy has been focused on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently [19].

9 PRIVACY AND SECURITY ASPECTS OF HEALTHCARE IN BIG DATA

The new wave of digitizing medical records has seen a paradigm shift in the healthcare industry. As a result, healthcare industry is witnessing an increase in absolute volume of data in terms of complexity, diversity and timeliness. The term "big data" refers to the a large group of many different things collected and complex data sets, which exceeds existing computational, storage and communication capabilities of conventional methods or systems. In healthcare, several factors provide the necessary propulsion to tackle the power of big data. The harnessing the power of big data analysis and complete research with real-time access to patient records could allow doctors to make informed decisions on treatments. Big data will oblige security to amend their predictive models. The real-time remote monitoring of vital signs through embedded sensors (attached to patients) allows health care providers to be alerted in case of an inconsitency. Healthcare digitization with integrated analytics is one of the next big waves in healthcare Information Technology (IT) with Electronic Health Records (EHRs) being a crucial building block for this vision. With the introduction of HER incentive programs, healthcare organizations recognized EHR's value proposition to facilitate better access to complete, accurate and sharable healthcare data, that eventually lead to improved patient care. With the ever-changing risk environment and introduction of new emerging threats and vulnerabilities, security violations are expected to grow in the coming years [20].

10 CONCLUSION

Big data [2] is analysed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analysed. In this paper, we have investigated the privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy preserving techniques are sufficient for big data processing. Privacy challenges in each phase of big data life cycle [7] are presented along with the advantages and disadvantages of existing privacy-preserving technologies in the context of big data applications. This paper also presents traditional as well as recent techniques of privacy preserving in big data. In terms of healthcare services as well, more efficient privacy techniques need to be developed.

REFERENCES

[1] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream management. VLDB J. 2003;12(2):120–39.

[2] Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015;2(4):155–65.

[3] Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.

[4] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next

frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1–137.

[5] Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.

[6] sai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.

[7] Hu J, Vasilakos AV. Energy Big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid. 2016;7(5):2423–36

[8] Jing Q, et al. Security of the internet of things: perspectives and challenges. Wirel Netw. 2014;20(8):2481–501.

[9] Han J, Ishii M, Makino H. A hadoop performance model for multi-rack clusters. In: IEEE 5th international conference on computer science and information technology (CSIT). 2013. p. 265–74.

[10] Gudipati M, Rao S, Mohan ND, Gajja NK. Big data: testing approach to overcome quality challenges. Data Eng. 2012:23–31.

[11] Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access. 2014;2:1149–76.

[12] Liu S. Exploring the future of computing. IT Prof. 2011;15(1):2–3.

[13] Sokolova M, Matwin S. Personal privacy protection in time of big data. Berlin: Springer; 2015.

[14] Cheng H, Rong C, Hwang K, Wang W, Li Y. Secure big data storage and sharing scheme for cloud tenants. China Commun. 2015;12(6):106–15.

[15] Mell P, Grance T. The NIST definition of cloud computing. Natl Inst Stand Technol. 2009;53(6):50.

[16] Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.

[17] Xu K, et al. Privacy-preserving machine learning algorithms for big data systems. In: Distributed computing systems (ICDCS) IEEE 35th international conference; 2015.

[18] Groves P, Kayyali B, Knott D, Kuiken SV. The 'big data' revolution in healthcare. New York: McKinsey & Company; 2013.

[19] EHR incentive programs. 2014. [Online]. https://www.cms.gov/Regulations-and Guidance/Legislation/EHRIncentivePrograms/index.html.

[20] First things first—highmark makes healthcare-fraud prevention top priority with SAS. SAS; 2006