Privacy Preservation in Big Data

Vijay laxmi Sharma[1]
[1]Asst.Professor, Computer Science & Eng. Dept.
Jaipur Engineering College & Research Centre
Jaipur, Rajasthan, India

Abstract

Big Data: a gigantic volume of both structured and unstructured data that it's hard to process utilizing customary database and programming methods.. Privacy preservation is one of most concerned issues in Big Data.The Proposed More Efficient and protection saving cosine similitude figuring protocol (PCSF) that can proficiently compute the cosine similitude of two vectors without unveiling the vectors to each other. SHA-3 Hash function" KACCAK" and AES Cryptographic algorithm are used that ensures the Authentication and Integrity while Processing of Data. It provides Privacy Preservation and thus be very useful for privacy-preserving in big data analytics.

Keywords — *Big data , Kaccak ,AES and Cosine similarity*

I. INTRODUCTION

Big data, since it can dig new learning for monetary development and specialized advancement, has recently received considerable attention, and many research endeavours have been coordinated to big data processing. Security and protection issues are amplified by the velocity, volume, variety, Value and Veracity of Big Data**.** Ttraditional security mechanisms, which are customized to securing small-scale, static data, are insufficient . Big data analytics, which is for the most part made out of three sections: multi-source big data collecting, distributed big data storing, and intra/inter big data processing.

The key component of big data analytics is big data processing, Because big data processing proficiency is a vital measure for the accomplishment of big data, the security necessities of big data processing turn out to be all the more difficult. big data processing can be divided into two types: intra big data processing if all data belong to the same organization, and inter big data processing if big data belong to different organizations. Since inter big data processing runs over multiple organizations data, big data sharing is essential, and ensuring privacy in big data sharing becomes one of the most challenging issues in big data processing. Therefore, it is desirable to design efficient and privacy-preserving algorithms for big data sharing and processing

Authentication and Integrity is required during Big Data Processing.In this paper Hash functions and AES **(**Advanced Encryption Standard) are used to provide Integrity and Authentication. Data integrity check is a most common application of the hash functions. It is used to generate the checksums on data files. This application gives affirmation to the client about accuracy of the data. SHA-3 (Secure Hash Algorithm-3) based on KECCAK, the algorithm1 that NIST chose as the champ of the general population SHA-3 Cryptographic Hash Algorithm Competition. KECCAK is based on the sponge construction. After the pre-processing (which divides the message into blocks and provides padding), the sponge construction consists of two phases: Absorbing (or input) phase: - The message blocks *xi* are passed to the algorithm and processed. Squeezing (or output) phase: - An output of configurable length is computed.

AES is a symmetric-key algorithm, which means the same key is used for both encrypting and decrypting the data. AES is based on a design principle known as a substitution-permutation network, combination of both substitution and permutation, and is fast in both software and hardware.

II. LITERATURE REVIEW

Many research efforts have been directed to find privacy requirements and different privacy preservation methods. Lu et. al [1] formalize the general design of big data analytics, identify the corresponding privacy requirements, and introduce an efficient and privacy-preserving cosine similarity computing protocol as an example in response to data mining's efficiency and privacy requirements in the big data era. Sung-Hwan et. al [2] describe the background of big data, data mining and big data features, and propose attribute selection methodology for ensuring the value of big data and concentrate on two things. Firstly, quality pertinence in big data is a key element for extracting information. Secondly, it is difficult to secure every single enormous data and its qualities. They consider big data as a solitary protest which has its own particular traits. They expect that a attribute which have a higher significance is more vital than different attribute es. They consider big data as a single object which has its own attribute.

Xuyun et. al [3] propose an adaptable multidimensional anonymization approach for big data privacy preservation utilizing MapReduce on cloud. In the approach, an exceedingly versatile median-finding algorithm combining the idea of the

median of medians and histogram technique is proposed and the recursion granularity is controlled to achieve cost-effectiveness.

Mahesh et.al [4] propose another strategy to protect the security of individuals' sensitive data from record and attribute linkage attacks. . In the proposed technique, security protection is accomplished through generalization of semi identifier by setting range values and record elimination**.** Golipour et . al [5] ventures the requirement for SHA-3.Because of the weakning of the broadly utilized SHA-1 hash calculation and worries over the comparatively organized algorithmes of the SHA-2 family; the US NIST has started the SHA-3 challenge keeping in mind the end goal to choose a reasonable drop-in replacement. In this paper they give survey of KECCAK hash function;s algorithm and apply a few strategies to enhance the execution as for throughput, frequency and timing.

Demchenko et. al [6] discuss the challenges that are imposed by Big Data on the modern and future Scientific Data Infrastructure (SDI). The paper proposes the SDI generic architecture model that provides a basis for building interoperable data or project centric SDI using modern technologies and best practices. The paper explains how the proposed models SDLM and SDI can be naturally implemented using modern cloud based infrastructure services provisioning model and suggests the major infrastructure components for Big Data.

Antorweep et. al **[**7**]** propose A system for keeping up security and preserving privacy for analysis of sensor data from smart homes, without compromising on data utility is presented. Nickolas et. **a**l **[**8**]** focuse on KECCAK SHA-3 algorithm and the sponge construction encryption process with iterative permutation. The algorithm utilizes the hashing function which is used for secured message authentication of data, digital signatures and password protection. Leontidies et. al **[**9**]** depict an extensive variety of data analysis operations involves a similarity detection algorithm between user data. Similarity decisions are imperative for various applications such as: online social networks, recommendations systems and behavioral advertisement. In this paper a mechanism is proposed that protects user privacy and preserves data similarity results although encrypted.

Rajan et. al [10] Security and Privacy issues are amplified by Velocity , Volume ,Variety and Veracity of Big Data. In this paper they highlighted Top 10 major big data- specific security and privacy challenges. Hawashin et. al[11] describe During the similarity join process, at least one sources may not permit imparting its information To different sources. For this situation, a privacy preserving similarity join is required. In this paper, they introduce a secure efficient protocol to semantically join sources when the join attributes are long attributes. They give two secure conventions to both situations when a preparation set exists and when there is no accessible preparing set. Moreover, they presented the multi name administered secure convention and the expandable regulated secure convention. They demonstrate that their conventions can proficiently join sources utilizing the long properties by considering the semantic connections among the long string values. In this manner, it enhances the general secure similitude join execution.

III. METHODOLOGY

The proposed algorithm PCSF can productively ascertain the cosine similarity of two vectors without unveiling the vectors to each other, and consequently be exceptionally valuable for security safeguarding in big data analytics. Cosine similarity is an important measure of similarity of two objectives captured by vectors $\vec{a} = (a_1 \ldots a_n)$ *and* $\vec{b} = (b_1 \ldots b_n)$*, respectively*. In big data analytics, *cos* $(\vec{a}, \vec{b})$ has become a critical building block for many data mining techniques. When we consider inter big data processing (i.e., $\vec{a}$ and $\vec{b}$ *do not belong to the same organization), the direct* cosine similarity computation would reveal each other's privacy. To achieve the privacy-preserving cosine similarity computation, we can apply Proposed efficient and privacy-preserving cosine similarity computing protocol (PCSC) [1] to compute $(\vec{a}, \vec{b})$.However, since PCSC lacks Authentication and Integrity, it is inefficient in big data processing. Therefore, we introduce Proposed More Efficient and protection saving cosine similitude figuring protocol (PCSF) for big data processing. For the privacy preservation of PCSF, since each a_i, $i = 1, 2 \ldots n$ is one-time masked with random $C_i = s\,(a_i\, a + c_i) \bmod p$, P_A can ensure that each a_i is privacy-preserving. Adding $a_{n+1} = a_{n+2} = b_{n+1} = b_{n+2} = 0$ is to ensure that at least two random numbers, r_{n+1}, r_{n+2}, are included in D, which can prevent P_A from guessing P_B's vector b = (b1, …, bn).Therefore, adding the random values r_{n+1}, r_{n+2} *is necessary*, which can eliminate the above guessing attack. To provide authentication and integrity SHA-3 KACCAK Hash function and AES is applied. It can ensure that each C_i and D_i is Privacy preserving.As a result, the proposed PCSF protocol should be privacy-preserving. For efficiency, compared with the PCSC -based protocol, the proposed PCSCFprotocol ensures authentication and integrity during big data processing.

Proposed More Efficient and protection saving cosine similitude figuring protocol (PCSF)

P_A P_B

$a\rightarrow = (a_1, a_2 \ldots a_n) \in F_q^n$ $b\rightarrow = (b_1, b_2, \ldots, b_n) \in F_q^n$

Step1: **(performed by P_A)** given security parameters k_1, k_2, k_3, k_4, choose two large primes α, p such that $|p| = k_1$, $|\alpha| = k_2$, Set $a_{n+1} = a_{n+2} = 0$.

Choose a large random number $s \in Z_p$, and n + 2 random numbers c_i, i = 1,2,...,n + 2, with $|c_i| = k3$

For each a_i, i = 1, 2... n + 2

$C_i = s(a_i \cdot a + c_i) \bmod p, \quad a_i \neq 0$

$C_i = s \cdot c_i \bmod p, \quad a_i = 0$

End for

Step2: **(performed by P_A)** Compute h (m) using SHA-3 KACCAK hash function. Where m=$C_1 \ldots Cn$.

Compute E (K, h (m)) using AES algorithm.

Compute $A = \Sigma a_i^2$ where i=1…n, keep s^{-1} mod p secret, and send (α, p, $C_1 \ldots C_{n+2}$) and E (k,h(m)) to P_B

α,p, C_1,.....C_n,E(K,h(m))-------->

Step3: **(performed by P_B)** set $b_{n+1} = b_{n+2} = 0$

Perform Decryption D (K, E (K,h(m)) and get h(m)

Compute h (m) Kaccak hash function using $C_1 \ldots C_n$

Compare h (m) to Decrypted h(m) if both are same

Compute D_i

For each b_i, i = 1, 2... n + 2

$D_i = b_i \cdot \alpha \cdot C_i \bmod p, \quad b_i \neq 0$

$D_i = r_i \cdot C_i \bmod p, \quad b_i = 0$

Where r_i is a random number, with $|r_i| = k4$

End for

$B = \Sigma b_i^2$ and $D = \Sigma D_i \bmod p$ for i=1 to n+2,

Compute h(m)

Where m=$D_1 \ldots D_n$. Compute E (K ,h(m))

send (B, D) back to P_A B, D

Step4: **(performed by P_A)** Perform Decryption D (K,E(K,h(m)) and get h(m)

Compute h (m) Kaccak hash function using $D_1 \ldots D_n$

Compare h m) to Decrypted h (m) if both are same

Compute $E = s^{-1} \cdot D \bmod p$

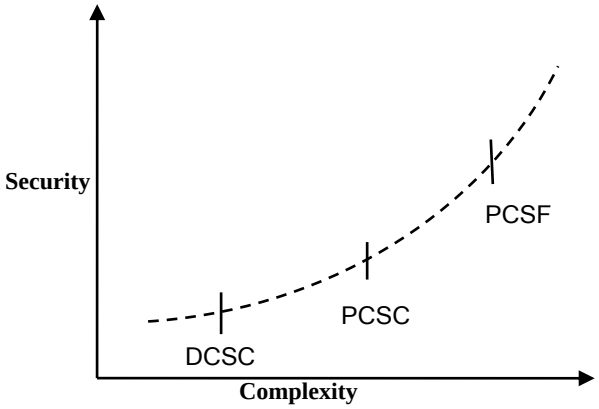
Compute $a^{\rightarrow} \cdot b^{\rightarrow} = \Sigma a_i \cdot b_i = (E - (E \bmod \alpha^2)) / \alpha^2$, for i=1 to n

$\text{Cos}(a^{\rightarrow}, b^{\rightarrow}) = (a^{\rightarrow} \cdot b^{\rightarrow}) / (\sqrt{A}\sqrt{B})$

IV. RESULTANALYSIS

To evaluate the proposed More Efficient Proposed More Efficient and protection saving cosine similitude figuring protocol (PCSF) protocol, we compare it with PCSC. We first implement the More Efficient PCSF protocol, proposed efficient and privacy-preserving cosine similarity computing protocol (PCSC) [1] with Java. We run them with the same input on a PC. Concretely, with the parameter settings q = 128, k1 = 512, k2 = 200, k3 = k4 = 128, n = {50, 100, 150, 200, 250}, each time we first randomly generate a vector $a^{\rightarrow}$=(a1… an) ∈ F_q^n and read 1000 vectors $b_i^{\rightarrow} = (b_{i1}, \ldots, b_{in}) \in F_q^n$ with *i* = 1, …, 1000, from an ,existing big data set, and use three ways to respectively calculate cos ($a^{\rightarrow}$, $b^{\rightarrow}$), with i = 1, …, 1000 for performance evaluation in terms of total complexity and security .For each parameter setting, We run the experiments 10 times, and the average performance results over 10 runs are reported. In Fig. 3, we plot the complexity and security of DCSC, PCSC, and PCSF varying with different parameter *n*. For each parameter setting, We run the experiments 10 times, and the average performance results over 10 runs are reported. In Fig. 1, we plot the complexity and security of DCSC, PCSC, and PCSF varying with different parameter *n*.

From the figure, we can see that by increasing *n*, the complexity and security of the PCSC protocol increase hugely which is much higher than that of the direct cosine similarity computation. Therefore, it is not efficient in big data processing. Although the complexity of our proposed PCSF protocol also increases when *n* is large, it is still close to the complexity of PCSC . But PCSF provides more security than DCSC and PCSC. Therefore, the experiment results show that our proposed More Efficient PCSF protocol is not only privacy-preserving but also efficient .We use Kaccak hash function and AES algorithm. Both increases the complexity and computation time. But fully ensures the privacy preservation during the processing of data that is most important. It is particularly suitable for big data analytics.



V. CONCLUSIONS

In this paper, I have investigated the privacy challenges in the big data era and then discussed whether existing privacy-preserving techniques are sufficient for big data processing. I have also introduced More Efficient and protection saving cosine similitude figuring protocol(PCSF) in response to the efficiency and privacy requirements of data mining in the big data era. Although I have analyzed the privacy and efficiency challenges in general big data analytics to shed light on the privacy research in big data, significant research efforts should be further put into addressing unique privacy issues in some specific big data analytics.

REFERENCES

[1] Rongxing, Z. Hui, L. Ximeng, J. K. Liu, and S. Jun, "Toward efficient and privacy-preserving computing in big data era," Network, IEEE, vol. 28, no. 4, pp. 46-50, 2013.

[2] K. Sung-Hwan, K. Nam-Uk, and C. Tai-Myoung, "Attribute Relationship Evaluation Methodology for Big Data Security," in International Conference on IT Convergence and Security (ICITCS) 2013, pp. 1-4.

[3] Z.Xuyun, Y.chi,S. Nepal, L.Chang, D.Wanchun, and C. Jinjun, "A MapReduce Based Approach of Scalable Multidimensional Anonymization for Big Data Preservation on Cloud," in Third International Conference on cloud and Green Computing (CGC) 2013, pp. 105-112.

[4] R. . Mahesh and T. Meyyappan, "Anonymization technique through record elimination to preserve privacy of published data," in International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013, pp. 328-332

[5] A. Gholipour , S.Mirzakuchaki, " High-Speed Implementation of the KECCAK Hash Function on FPGA, " in International Journal of Advanced Computer Science, Vol. 2, No. 8, Pp. 303-307, Aug., 2012.

[6] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey " Addressing big data issues in Scientific Data Infrastructure ," in International Conference on Collaboration Technologies and Systems (CTS) 2013, pp.48-55.

[7] C. Antorweep, W. Tomasz, and R. Chunming, "Privacy Preserving Data Analytics for Smart Homes," in Proceedings of the IEEE Security and privacy workshops: IEEE Computer Society, 2013.

[8] D.B.Nickolas, A. Sivasankar, " High-Speed Implementation of the KECCAK Hash Function on FPGA, " in International Journal of Engineering Trends and Technology(IJETT)-Volume4 Issue6- June 2013.

[9] Iraklis Leontiadis, Melek O¨ nen, Refik Molva, M.J. Chorley, G.B Colombo "Privacy preserving similarity detection for data analysis," in Proceedings of the IEEE Security and Privacy Workshops: IEEE Computer Society, 2013.

[10] S. Rajan , W. Ginkal ,N. Suderesan , "Top Ten Big Data security AND Privacy Challenges, "in Cloud Security Aliance,2012,pp. 1-11.

[11] Bilal Hawashin , Farshad Fotouhi, Traian Marius Trut,William Grosky "Efficient Privacy Preserving Protocols for Similarity Join, " Network,IEEE, 29, no. 9, pp. 55-59, 2013.