Big Data Security and Privacy: A Review on Issues, Challenges and Privacy Preserving Methods

Anupama Jha
Assistant Professor,
VIPS, GGSIPU
New Delhi, India

Meenu Dave, PhD
Professor
Jagannath University
Jaipur, India

Supriya Madan, PhD
Professor
VIPS, GGSIPU
New Delhi, India

ABSTRACT
In recent years the rapid growth of Internet, IOT and Cloud Computing has led to voluminous data in almost every organization, academics and business area. Big data has rapidly developed into a hot topic that attracts extensive attention from such area around the world. Maintaining the privacy and security of Big Data is a very critical issue. The5V characteristics of big data (Volume, Variety, Velocity, Value and Veracity) alleviate the standard of security required for it. In this research paper, we have emphasized several Big Data security and privacy issues and challenges released by CSA (Cloud Security Alliance) that need to be addressed to make data processing and computing infrastructure more secure as well as possible solutions to address such challenges. This paper also gives insights on overview of big data privacy preserving K-Anonymity technique which aims to protect against leakage of individual's identity and sensitive information before releasing the dataset during analysis. Finally, this paper overviews big data security solution application and their features provided by the top companies

Keywords
Big data 5V Characteristics, Security, Privacy, CSA, K-Anonymity

1. INTRODUCTION
With the development of Internet applications, social networks and Internet of Things (IOT), a continuous growth in data generation has been observed, which we called today is big data [1]. As Big data is generated from multiple sources with multiple formats and with very high speed, is also characterized by its 5V's properties, such as Volume, Velocity, Variety, Value and Veracity. Due to its 5V characteristics, Big data is coming with new challenges during its all phases of life cycle [2], which involve privacy and 5 major security aspects such as confidentiality, efficiency, authenticity, availability & integrity. The figure 1 showing all the security aspects corresponding to Big data 5V characteristics.

As confidentiality is the basis of big data security and privacy, we need to protect data from leakage. Once data is leaked, its value will be lost. The value of the big data could be disappeared if hackers attack the data by changing the data or obtaining secret information. Efficiency is especially crucial in big data security and privacy as it require high network bandwidth. Authenticity is necessary to ensure reliable data sources, data processors and authorized data requesters. Authenticity can avoid wrong analysis result and assist achieving high potential value from big data. Big data should be available anytime when we need it. Otherwise, it could lose its value. Integrity is also essential to get valuable and accurate data. With inaccurate or incomplete data, we cannot analyse correct result especially when the lost data is the most sensitive and useful.

Table 1: Security aspects in Big Data Life Cycle

Big data 5V characteristics	Security Aspects				
	Confidentiality	Efficiency	Authenticity	Availability	Integrity
Volume		√		√	
Velocity		√		√	
Variety		√		√	
Value	√		√	√	√
Veracity	√		√		√

Today for analysis purpose, big data is mainly used by different sectors such as Healthcare, Government agencies, businesses, research and other organization. Such analysis frequently requires their data for publishing, investigation and for other purposes. Big data also contains individual's specific information, so directly releasing this data for analysis can pose serious threats to user's privacy. Hence, privacy preserving big data mining techniques are needed which aims to protect against identity disclosure and sensitive information disclosure of the dataset.

2. SECURITY VERSUS PRIVACY
Security and Privacy in big data is an important issue. In order to properly utilize big data, one must address the security and privacy issues. Security focuses on protecting data from pernicious attacks and stealing data for profit [3]. Data privacy focuses on the use and governance of individual's personal data like making policies to ensure that consumers' personal information is being collected, shared and utilized in right ways.

3. BIG DATA SECURITY AND PRIVACY ISSUES & CHALLENGES
Everyday Big Data faces high level of challenges while dealing with the privacy and security of gigantic and heterogeneous data. Data are shared on a large scale by different people such as researchers, scientists, doctors, business officials, government agencies etc. Although the tools and technologies that have been developed till date to handle these huge volumes of data are not efficient enough to provide adequate security and privacy to data. Also, the present technologies have weak security and privacy maintenance capability so they are continuously being breached both accidentally and intentionally. Recently, CSA (Cloud Security Alliance) released the top ten big data security &privacy challenges [4]. The objective of

highlighting such challenges is to bring renewed focus on stimulating big data infrastructures. The top researchers from CSA's Big Data working group compiled such relevant challenges in the context of big data security and privacy which can be categorized into four aspects and further these categories are subdivided into 10 distinct security challenges as follows [5]:

I. Infrastructure Security
II. Data Privacy
III. Data Management
IV. Integrity and Reactive Security

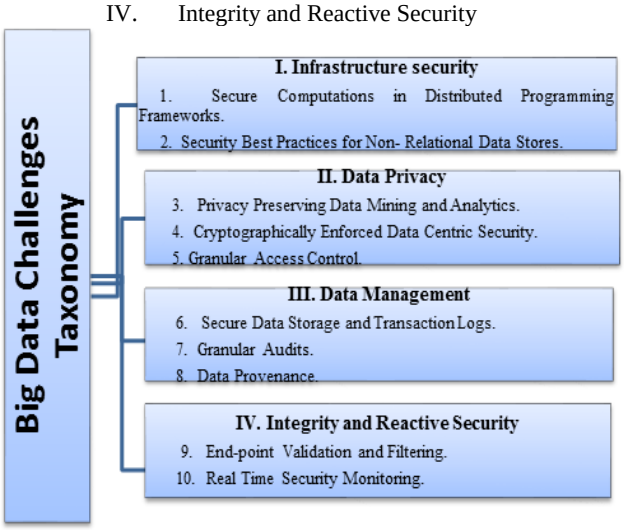

Figure 1: Taxonomy of Top 10 Big Data Challenges

i) Secure Computations in Distributed Programming Frameworks:

Distributed programming frameworks use the concept of parallel computation and storage to process the massive amounts of data. The MapReduce framework is the best example. It splits the input file into multiple chunks and then the mapper reads the chunks, does computations and provides outputs in the form of key/value pairs. The Reducer then combines the values belonging to each distinct key and outputs the result. The two-major attack prone issues here are: securing the mappers and securing the data from a malicious mapper.

ii) Security Best Practices for Non- Relational Data Stores:

NoSQL (Non-relational) databases which is used to store big data, handle many challenges of big data analytics without concerning much over security issues. Developers using NoSQL databases usually embed security in the middleware. NoSQL databases do not provide any support for enforcing it explicitly in the database. An additional challenge to the robustness of such security practices is the clustering aspect of NoSQL databases [6].

iii) Privacy Preserving Data Mining and Analytics:

Big data can potentially allow invasions of privacy, invasive marketing, decreased civil freedoms and increase state & corporate control. User data collected by companies and government agencies are constantly mined and analyzed by inside analysts and possibly outside contractors. A malicious insider or untrusted partner can abuse these datasets and extract private information about customers. It is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures.

iv) Cryptographically Enforced Data Centric Security:

The private data must be encrypted based on access control policies to ensure end to end secured and only accessible to the authorized entities. Specific research in this area such as attribute-based encryption (ABE) should be made richer, more efficient, and scalable. A cryptographically secure communication framework must be implemented to ensure authentication, agreement and fairness among the distributed entities.

v) Granular Access Control:

Access control is about two core things according to the CSA: restricting user access and granting user access. The challenge is to build and implement a policy that chooses the right one in any given scenario.

vi) Secure Data Storage and Transaction Logs:

Data and transaction logs are stored in multi-tiered storage media manually moving data between tiers gives the it manager direct control over exactly what data is moved and when. However, as the size of data set has been and continues to be, growing exponentially, scalability and availability have necessitated auto-tiring for big data storage management. Auto-tiring solutions do not keep track of where the data is stored, which poses new challenges to secure data storage [7].

vii) Granular Audits:

It is a best practice to perform granular audits. With real-time security monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case. To get to the bottom of any attack, we need audit information. This is helpful to understand what happened and also it is important from compliance and regulations point of view. Auditing is not new, but the scope and granularity might be different, e.g. we might need to deal with a large number of distributed data objects.

viii) Data Provenance:

In big data applications, provenance metadata will grow in complexity due to large provenance graphs generated from provenance- enabled programming environments. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

ix) End-point Validation and Filtering:

Many big data applications such as a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software application in an enterprise network. A key challenge in the data collection process is input validation: How can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the Bring Your Own Device (BYOD) model.

x) Real Time Security Monitoring:

Real-time security monitoring has been an on-going challenge in the big data analysis scenario mainly due to the number of alerts generated by security devices. These alerts, may be co-related or may not be, lead to many false positives and due to human being's incapability to successfully deal with such an huge amount of them at such a speed, results in them being clicked away or ignored [8].

4. SOLUTION TO ENSURE BIG DATA SECURITY AND PRIVACY

Several techniques that have been developed to ensure that data remains protected are:

- **Access control technology:** Due to the huge no of users and complex authority in big data environment, a new technology must be adopted to realize the controlled sharing of data. The Role-based access control (RBAC) is widely used access control mode. This can be achieved by restricting access to the data by adding access control to the data entries so that sensitive information is accessible to a limited user groups only. The main challenge here is that no sensitive information can be misconduct by unauthorized individuals.
- **Homomorphic Encryption Schemes (HES):** To protect the privacy of big data, even if the data with privacy leak, the attacker cannot obtain the effective value of data. It is a type of encryption technique that allows functions to be computed on encrypted data without decrypting it first. That is, given only the encryption of a message, one can obtain an encryption of a function of that message by computing directly on the encryption.
- **Secure Multi-Party Computation (SMC):** It generally deals with problems of function computation with distributed inputs. Goal of SMC is to compute function when each party has some input. In this protocol, parties have security properties e.g. privacy and correctness. Regarding privacy a secure protocol must not reveal any information other than output of the function
- **Data Anonymization Technology:** Data anonymization or De-identification is a very popular technique used on both distributed and centralized database. With this approach, sensitive information fields should be anonymised so that they cannot identify an individual record. Even if the attacker gets such data, he cannot get the original exact data, because the value of the key field is hidden. The two main privacy objectives that should be achieved when data is anonymized are:
  - Unique identity disclosure: If data is published then there should not be any record that can identify an individual.
  - Sensitive attribute disclosure: Attackers won't be able to learn about sensitive attribute of an individual via disclosed attributes.

In this research paper, our main focus is on data anonymization technique.

4.1 Data Anonymization / De-identification Technique

It refers to the hiding of individual's private and sensitive data [9]. It is a privacy preserve technique which is used when big data is published to third parties. Normally a record in a dataset consists of 3 types of attributes:

- Key attributes are those attribute that uniquely identifies each individual. e.g. ID, Name, address, phone number. They always removed before release.
- Quasi-Identifiers (QI) are those set of attributes that can be linked with other datasets which is publicly available to identify individual's private data. It can be used for linking anonymized dataset with other datasets. e.g. Age, sex, zip code, city etc.
- Sensitive attributes contain some sensitive information which an individual wants to hide from others. e.g. income, salary, disease, medical records etc.

One of the major benefits of the data anonymization based information sharing approaches is that, once dataset is anonymized, it can be freely shared across different parties without involving restrictive access controls. There are 3 privacy-preserving methods of data anonymization that are used to prevent attacks on the privacy of the published d a t a . They are K-anonymity, L-diversity and T-closeness.

4.1.1 K-anonymity

When attributes are suppressed or generalized until each row is identical with at least k-1 other rows then that method is called as a k-anonymity. It prevents definite database linkages and also guarantees that the data released is accurate. But it has some limitation:

- It does not hide individual identity.
- Unable to protect against attacks based on background knowledge.
- K-anonymity cannot be applied to high dimensional data.

4.1.2 L-diversity

Method overcomes the drawbacks of k-anonymity but fails to preserve the privacy against skewness and similarity attacks.

4.1.3 T-closeness:

This method preserves the privacy against homogeneity and background knowledge attacks. It is called as t-closeness when the distance between the distribution of a sensitive attribute in same class and the distribution of the attribute in the whole table is no more than a threshold. A table is said to have t-closeness if all equivalence classes have t-closeness.

5. K-ANONYMITY

K-Anonymity provides a measure of privacy protection by preventing re-identification, which makes highly accurate and secured data analysis. This privacy model is used to prevent linking attacks [10][11]. A dataset is K-Anonymized when a tuple / individual in the published dataset is indistinguishable from at least k-1 other tuples / people in that dataset. Therefore, an attacker who knows the values of quasi-identifier attributes of an individual are not able to distinguish his record from the k-1 other people / records. The two main techniques that have been proposed to reinforcing K-anonymity on a private dataset are [12]:

- Generalization which refers to replacing a value with more generic value. For example, male/female can be replaced with person.
- Suppression which refers to hiding the value by not releasing it at all. The value is replaced by a special character such as *, @.

The above both technique has the property of preserving the truthfulness of data.

Algorithm for K-Anonymity method

Input: A private dataset PT, Quasi-identifier attributes QI, Sensitive values A, anonymity parameter K.

Output: Releasing Table RT.

Step1: Select Data set PT from a Database.
Step2: Select Key attribute, Quasi-identifier attribute and Sensitive Attribute from given attribute

list.

Step3: Select the set of most sensitive values A from list of all sensitive values that is to be preserve.

Step4: For each tuple whose sensitive value belongs to set A. If t[S] ∈ A then move all these tuples to Table 1.

Step5: Find the statistics of quasi attributes of Table 1 i.e. distinct values for that attribute and total number of rows having that value.

Step6: Apply generalization on quasi-identifiers of Table 1 to make it K-Anonymized, which is an output table RT and ready to release.

Table 2: Original Medical Dataset

Key Attribute	QI Attributes			Sensitive Attribute
Name	Age	Sex	Zipcode	Disease
Ravi	30	Male	12567	Fever
Sam	25	Male	13001	Cancer
Ramesh	26	Male	13001	Flu
Manav	34	Male	76512	Flu
Suhani	23	Female	14599	Viral
Keshav	35	Male	13057	Pneumonia
Anita	35	Female	17000	Fever
Hema	23	Female	32451	Cancer
Reshma	27	Female	14560	Flu

Table 3: Anonymous Dataset (After removing Key Attributes)

Age	Sex	Zipcode	Disease
30	Male	12567	Fever
25	Male	13001	Cancer
26	Male	13001	Flu
34	Male	76512	Flu
23	Female	14599	Viral
35	Male	13057	Pneumonia
35	Female	17000	Fever
23	Female	32451	Cancer
27	Female	14560	Flu

Before releasing the original table, the key attributes must be removed from the table. The resultant table is shown in table 2. But table 3 can be linked with external data which is available to the attacker. Following table shows the external data which is a Voter Registration List available to the attacker.

Table 4: Publicly available Voter Registration Dataset

Voter Id No	Name	Age	Sex	Zipcode
QDT2398452	Ravi	30	Male	12567
SAP2345918	Sam	25	Male	13001
QAC4982107	Ramesh	26	Male	13001
HTR4356723	Manav	34	Male	76512
RAP3412090	Suhani	23	Female	14599
KDE7351898	Keshav	35	Male	13057
WRT2783261	Anita	35	Female	17000
WER254367	Hema	23	Female	32451
KYE3456123	Reshma	27	Female	14560

After comparing Table 3 and Table 4, the attacker will get to know that Ramesh is suffering from Flu. So even after removing the key identifiers, an individual can be re-identified with the help of publicly available data. So, combining data of the released table with the data of the publicly available table is known as Linking Attack. Hence, this privacy model is used to prevent linking attacks. That is if we try to identify an individual from a release, then the only information we have is his/her age, gender and zip code. Table 5 below is an example of 2-anonymous table, where k=2 i.e. at least two tuples have same values in the quasi-identifier attributes. Here it can be found that t2[S]= t3[S]= t6[S] &t5[S]= t9[S]

Table 5: 2-Anonymized Dataset

Age (equivalence class)	Sex	Zipcode (after suppression)	Disease
[20-30]	Male	125*	Cancer
[20-30]	Male	130*	Cancer
[20-30]	Male	130*	Flu
[30-40]	Male	765*	Flu
[20-30]	Female	145*	Viral
[30-40]	Male	130*	Pneumonia
[30-40]	Female	170*	Fever
[20-30]	Female	324*	Cancer
[20-30]	Female	145*	Flu

Attacks on K-Anonymity:

i) Attribute disclosure attack/Homogeneity attack: It occurs when the sensitive attribute lack diversity in values and the attacker is only interested in knowing the value of sensitive attribute. For example, in 2nd, 3rdand 6th tuples of table 4, the values of the sensitive attributes Age, Sex and Zip code are same and can come to a conclusion that Ramesh is also suffering from Cancer or Pneumonia.

i i) Background Attack: This is another kind of attack which k-anonymity cannot prevent. This model assumes that attacker has no additional background knowledge.

6. BIG DATA SECURITY SOLUTION AND KEY FEATURES PROVIDED BY VARIOUS COMPANIES

With this research paper, we have overview the top companies' big data security solution application and their features which are listed in Table 6 below:

Table 6: Top companies Big Data Security Solution Applications and their Key Features

COMPANY NAME	APPLICATION	KEY FEATURES
IBM	IBM QRadar Security Intelligence Platform [13]	• A comprehensive, integrated approach that combines real time correlation for continuous insight, custom analytics across massive structured and unstructured data and, forensic capabilities for deep visibility. The entire combination can help to address advanced persistent threats, fraud and insider threats. • High-speed querying of security intelligence data. • Graphical front-end tool for visualizing and exploring big data.
INFOSYS	IIP-The Infosys Information Platform [14]	• An open source data analytics platform. • Enables businesses to operationalize their data assets and uncover new opportunities for rapid innovation and growth. • Provides an end-to-end data platform that leverages open source innovations and internal enhancements to seamlessly integrate into enterprise landscapes in a way that it can operate as a standalone big data solution or as an add-on to existing proprietary tools.
MICROSOFT	Big Data and Business Intelligence Solutions [15]	• Provides a modern data management layer that supports all data types of data i. e. structured semi-structured and unstructured data at rest or in motion. MS makes it easier to integrate, manage and present real-time data streams, providing a more holistic view of business to drive rapid decisions. • An enrichment layer that enhances our data through discovery, combining with the world's data and by refining with advanced analytics. • An insights layer that provides insights to all users through familiar tools like Office's Excel & PowerPoint. • HDInsight, MS's new Hadoop based service that offers 100% compatibility with Apache Hadoop. It enables the customers to gain business insights from variety of data with any size and activate new types of data irrespective of its location.
HP	HPE Security- Hewlett Packard Enterprise [16]	• Security provides best-in-class data encryption and tokenization for structured and unstructured data. • Cost-effective PCI compliance, scope reduction, and secure analytics. • Used by leading companies worldwide, reducing risk and protecting brand.
ORACLE	DBSAT- The Oracle Database Security Assessment Tool [17]	• Quickly identify security configuration errors in the databases. • Promote security best practices. • Improve the security posture of Oracle Databases. • Reduce the attack surface and exposure to risk.
VORMETRIC	Vormetric Data Security Platform [18]	• Enable companies to maximize the benefits of big data analytics. It offers the granular controls, robust encryption, and comprehensive coverage that organizations need to secure sensitive data across their big data environments • Enables security teams to leverage centralized controls that optimize efficiency & compliance adherence. • It offers capabilities for big data encryption, key management and access control.

7. REFERENCES

[1] Jha A, Dave M. and Madan, S. 2016. A Review on the Study and Analysis of Big Data using Data Mining Techniques, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol6, Issue 3.

[2] Jha A, Dave M. and Madan, S. 2016. Quantitative Analysis and Interpretation of Big Data Variables in Crime Using R, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), Vol5, Issue7.

[3] Q, etal, Jing. 2014. Security of the internet of things: perspectives and challenges. 20(8):2481–50.

[4] https://cloudsecurityalliance.org/media/news/csa-big-data-releases-top-10-security-privacy challenges/.

[5] A Cloud Security Alliance Collaborative research, Expanded Top Ten Big Data Security and Privacy Challenges. 2013.

[6] Okman, L., Gal-Oz N., Gonen Y, Gudes E. and Abramov J. 2011. Security Issues in NoSQL Databases in TrustCom IEEE Conference on International

Conference on Trust, Security and Privacy in Computing and Communications, pp 541-547.

[7] Apparao, Yannam and Laxminarayanamma, Kadiyala. 2015. Security Issue on Secure Data Storage and Transaction Logs In Big Data" in International Journal of Innovative Research in Computer Science & Technology (IJIRCST).

[8] Singh, Reena and Kunver Arif Ali. 2016. Challenges and Security Issues in Big Data Analysis, IJIRSET, Vol 5. Issue 1.

[9] Sedayao, J. Enhancing cloud security using data anonymization, White Paper, Intel Corporation.

[10] Kenig Batya and Tassa Tamir. 2011. A practical approximation algorithm for optimal k-anonymity, Data Mining Knowledge Discovery, Springer.

[11] Sweeney, L. 2002. K-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty Fuzziness Knowledge based Systems.

[12] Samarati. P. 2001. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Eng., 13:1010–1027.

[13] https://www-01.ibm.com/software/se/security/bigdata/

[14] http://www.experienceinfosys.com/iip-overview

[15] https://msdn.microsoft.com/en-us/library/dn749804.aspx

[16] www.hpe.com/hpe/jrit?

[17] https://docs.oracle.com/cd/E76178_01/

[18] https://www.thalesesecurity.com/products/data-encryption/vormetric-data-security-platform