ABSTRACT
With the recent advancement in the field of information technology and internet, the amount of data being generated, processed and stored is increasing very rapidly to the scale of Big Data. This data contains sensitive private information of individuals, their privacy needs to be protected from adversaries. Concern for data security and privacy are rising. In this paper, we have tried to describe a list of different privacy methods for big data, differential privacy-preserving methods such as Data Anonymization, Differential Privacy and Notice and consent. Comparative studies of data anonymization methods like k-anonymization, T-closeness, and L-diversity.

Keywords: Big data, Data Privacy, Anonymization, Differential Privacy, Notice, and Consent.

1. INTRODUCTION
By reason of recent technology development the number of data generated sources area is social networking sites, sensor networks, internet, healthcare applications, and many other companies, is drastically increasing day by day. All the massive amount of data generated from completely different sources in other than one format with very high speed is referred as big data. Big data generation rate is growing very quick that it is becoming extremely difficult to handle it using traditional methods or systems.

Big Data refers to datasets whose size or type is beside the capacity of traditional relational databases to capture, manage, and process the data with cheap latencies. This dataset contains a different type of data like structured/unstructured and streamed/batch and different sizes lie between terabytes to zettabytes. These add another challenge when performing data storage and processing task.

Big Data Analytics is the use of advance analytic techniques and algorithms against these datasets. Analyzing big data permit analysts, researchers, and business users to make higher and quicker decision using data that was earlier inaccessible or unusable. This modified the approach data is analyzed as rather than the well-defined and intense processes, the unstructured nature of the data and the flexibility of processing resources make it a lot of easier to adapt to dynamic environments.

Big data has generated from sources like on-line transactions, search queries, mobile phones, emails, videos etc. It has saved by partitioning on different servers. This denotes that big data sources and storage systems square measure detached all around the internet. Because it is broadly supporting the web model, security issues are a key challenge in big data analytics. Not guaranteeing secure big data analytics may cause huge loses to equally public and organizations.

Exposing whole of the big data may signify permanently analytics results though at the equivalent time can produce good security challenges. The worth hidden in big data square measure typically good value to hackers and invaders. Therefore, there is a trade-off between big data accessibility and big data security. One has to guarantee a fitting balance between the two. Therefore, the need is to guard data hidden in big data throughout the complete analytics process.

Traditional security mechanisms fail to handle big data due its massive volume, variety, and velocity. Among very different security aspects of big data, privacy is one of the major important issues [8]. Traditional methods like cryptography square measure typically used they are not persuaded be capable because of complicated nature of data [1].

For the growing issues on the individual privacy and increase in sort of illation attacks, the standard confidential agreements and privacy policy does not guarantee privacy. Organizations have to be compelled to adopt technology-based solutions. It is very difficult to provide protection of individual privacy while not losing the analytical value of the data. It becomes even tougher once it’s challenging when it is Big Data and lots of parties are concerned with integrated data analysis.
A. Privacy concerns

Privacy issues have for some time been impacting enactment. Policies for the security of individual data were made, for instance, inside the world organization (1995), Canada (PIPEDA, 2000) and Australia (Privacy Act, 1988), and few of them cross business sectors. The United States has adopted an organization strategy, instituting separate policies for health care (HIPAA, 1996), finance (Gramm-Leach-Bliley, 1999) and protection of children’s data (COPPA, 1998). Moreover, privacy principles address basic rules concerning how organizations should handle personal information – e.g., the Fair Information Practices (FTC, 2000) and OECD Privacy Principles (2010). Therefore, based on privacy principles and laws, individual’s privacy must be protected.

Within the Big Data era, people themselves offer Personally Identifiable Information (PII) to be ready to like the advantages and conveniences of services brought by internet services and information platforms. PII can be used on its own or at the side of different information to identify, contact, or locate a single person (e.g., name, addresses, social security IDs, credit card numbers, etc.). Individuals privacy can be compromised because once this information is made available it is no longer under their own control regarding how and by who it is used.

In this paper, we have tendency to describe varied measures, which will facilitate to make sure privacy in big data. This paper containing as follows: section II offers a big data privacy issues and challenges. Section III We have discussed privacy techniques for big data. Sections V conclude our work.

2. BIG DATA PRIVACY ISSUES AND CHALLENGES

The big data could be a current innovate technology, which is rapidly adopt by numerous industry to anticipate showcase patterns and client behavior. Since the essential part of big data tools is the capacity and procedure of enormous volumes, these tools have not however sufficiently saves space to conceal some relating to security and privacy shield.

A. Big Data Privacy Issues

These applications are shows how to privacy issue is creating in these applications. Following are the application of privacy preserving in big data:

1) Privacy issue in Mobile Data

Everything is available on mobile nowadays. People are sharing a lot of information on mobile phones. Mobile sends data to the service provider without user’s knowledge. Identifying the person using his mobile data and the details provided by the service provider is very easy. Therefore, privacy in mobile data is very important. Text message analysis is an example of unstructured big data analytics in mobile. This method is used in WhatsApp to identify the mobile number.

2) Privacy issue in Health Care Data

Big data analytics and genome research having real-time access to the patient record to take a decision. Electronic Health Record (EHR) helped a lot to digitize the health care system make accurate and complete EHR. EHR have personal information. Therefore, the privacy-preserving analysis is required and data need to be anonymized before data analysis. Ex. Pathology report

3) Privacy issue in Social Media Data

Social media is the biggest revaluations in past decade. -Lots of information is being shared by people on social media. Sometimes, people close to you share some information about you, which you don’t want disclose on social media. This may lead to privacy violation of an individual. Ex-Facebook, Twitter etc.

4) Privacy issue in Web Usage Data

Intel wants to make its internal website dynamic based on web usage data of all the users of the website. With browser information and IP address from web usage data any user be identified and whatever activities he is performing on line may be detected. Thus, privacy is required.

B. Big Data Privacy Challenges

A gathering of privacy and security issue must be considered before building a big data situation. In this, we have following most significant challenges, when managing with big data.
1) Random Distribution
Big data analytics in view of parallelism the extensive data is store and processed in a different cluster, which is gathering of disseminated servers around the world. We have the propensity to do not capable to distribute storage and processing with the directions and data sensibility [4].

2) Privacy
Current big data analytics care for all data with the similar need and do not combine unique activities, similar encryption or impaired processing [10]. Thus, programmer and malicious node gain access to the cluster.

3) Computations
The main objective of big data is to separate helpful bits of knowledge performing the particular calculation. It is significant to secure and protect this computation.

4) Integrity
Big data contain a substantial volume of a substance. For creating decision base on big data, it’s necessary to make sure the validity and therefore trust level of that data in ordered to avoid on the suspect or compromised records.

5) Communication
Big data is put away data in a few hubs, which are distributed around in the world. All communication linking nodes and cluster is complete by the ordinary public and private network [4].

6) Access control
In big data context, storage access organizes should mange access data. Any modification in cluster’s state such expansion or erasure of nodes ought to checked by confirmation mechanism to shield the system [4].

3. PRIVACY-PRESERVING TECHNIQUES FOR BIG DATA
Privacy in big data has raised significant issues transfer into seeing the need for proficient privacy preservation methods. In this segment, we need to examine three privacy-preserving techniques: Data Anonymization, Differential Privacy, Notice, and Consent.

There are some common terms used in privacy field of this method:
- **Identifier attributes** include information that uniquely and directly distinguishes individual such as full name driver license, social security numbers.
- **Quasi-identifier attributes** a set of information such as birth date, gender, age, zipcode. That can be combined with other external data in order to re-identify individuals.
- **Sensitive attributes** are private and personal information.
- **Intensive attributes** are the general and the innocuous information.
- **Equivalence classes** are sets of all records that consist of the same value on the quasi-identifiers.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>City</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daksh</td>
<td>24</td>
<td>M</td>
<td>Delhi</td>
<td>Cancer</td>
</tr>
<tr>
<td>Jay</td>
<td>25</td>
<td>M</td>
<td>Gurgaon</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>Vivek</td>
<td>28</td>
<td>M</td>
<td>Gurgaon</td>
<td>Dengue</td>
</tr>
<tr>
<td>Maulik</td>
<td>24</td>
<td>M</td>
<td>Delhi</td>
<td>TB</td>
</tr>
<tr>
<td>Kajal</td>
<td>26</td>
<td>F</td>
<td>Delhi</td>
<td>No illness</td>
</tr>
<tr>
<td>Shreya</td>
<td>27</td>
<td>F</td>
<td>Delhi</td>
<td>Viral infection</td>
</tr>
<tr>
<td>Neel</td>
<td>26</td>
<td>M</td>
<td>Delhi</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Krupa</td>
<td>30</td>
<td>F</td>
<td>Delhi</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Bhavin</td>
<td>32</td>
<td>M</td>
<td>Delhi</td>
<td>Viral infection</td>
</tr>
<tr>
<td>Shruti</td>
<td>39</td>
<td>F</td>
<td>Gurgaon</td>
<td>TB</td>
</tr>
<tr>
<td>Krishna</td>
<td>32</td>
<td>F</td>
<td>Gurgaon</td>
<td>Jaundice</td>
</tr>
<tr>
<td>Nirav</td>
<td>40</td>
<td>M</td>
<td>Delhi</td>
<td>Jaundice</td>
</tr>
</tbody>
</table>
A. Anonymization

Data Anonymization is that the method of fixing information which will be used or revealed during an approach that stops the identification of key information. It is typically referer as data de-identification. These are method utilized in this: Key items of confidential data are until obscured during approach that maintains data privacy and unleashes data publically by Anonymization. Example, a hidden attribute like full name, license number, voter id etc. The main drawback with data anonymization is that data might look anonymous however re-identification is often done simply by linking by linking it to other different external data [11]. It is shown that re-identification of anonymous medical records is done exploitation external constituent list data. Example attribute as if genders, date of birth, the zip code that can join with outside data to the re-identify individual call quasi-identifier attributes.

There is anonymization method also known as a De-identification. There are privacy-preserving methods of anonymization: k-anonymization, T-closeness, L-diversity

a) K-Anonymization

K-anonymization is a framework for constructing and evaluating algorithm also system that releases information. It is to allow sharing such data without compromising the privacy of the user. A data set called k-anonymization if for any k tuple with the same quasi-identifier in the dataset there are at least k-1 other records that match those attributes. K-anonymization also was known as De-identification.

Table 3 is a non anonymized database. There are four attributes along with twelve records in this data. There are two regular techniques for achieving k-anonymity for some value of k.

1) Suppression in this method, certain values of the attributes is replaced by an asterisk ‘*’. In the anonymized Table 4 replaced all the values in the ‘Name’ attribute and each of the values in ‘City’ attribute by a ‘*’.

2) Generalization in this method, individual values of an attribute is replaced with a broader category. For instance, the attribute ‘Age’ the value ‘24’ replaced by broader category ‘21 <Age ≤30’.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>City</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>M</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>M</td>
<td>*</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>M</td>
<td>*</td>
<td>Dengue</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>M</td>
<td>*</td>
<td>TB</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>F</td>
<td>*</td>
<td>No illness</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>F</td>
<td>*</td>
<td>Viral infection</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>M</td>
<td>*</td>
<td>Heart related</td>
</tr>
<tr>
<td>*</td>
<td>21&lt;Age≤30</td>
<td>F</td>
<td>*</td>
<td>Heart related</td>
</tr>
<tr>
<td>*</td>
<td>31&lt;Age≤40</td>
<td>M</td>
<td>*</td>
<td>Viral infection</td>
</tr>
<tr>
<td>*</td>
<td>31&lt;Age≤40</td>
<td>F</td>
<td>*</td>
<td>TB</td>
</tr>
<tr>
<td>*</td>
<td>31&lt;Age≤40</td>
<td>F</td>
<td>*</td>
<td>Jaundice</td>
</tr>
<tr>
<td>*</td>
<td>31&lt;Age≤40</td>
<td>M</td>
<td>*</td>
<td>Jaundice</td>
</tr>
</tbody>
</table>

Table 4 has 2-anonymity with respect to the attributes ‘Age’, ‘Gender’ and ‘Disease’ amount of these any attributes found in any row of the table there are always no less than tow rows with those exact attributes. The attributes are available to an adversary are called ‘quasi-identifier’. Each ‘quasi-identifier’ tuple occurs in at least k records for a dataset with k-anonymity.

b) L-Diversity

It is a form of group-based anonymization that is utilized to safeguard privacy in data sets by reducing the granularity of data representation. This decrease is a trade-off that results in outcomes in some loss of viability of data management or mining algorithms for gaining some privacy. The L-diversity model is an extension of the k-anonymity model which diminishes the granularity of data representation utilizing methods including generalization and suppression in a way that any given record maps onto at least k different records in the data. The L-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting to corresponding sensitive values that were generalized or suppressed, particularly when the sensitive values in a group exhibit homogeneity. The L-diversity model includes the promotion of intra-group diversity for sensitive values in the anonymization mechanism. The problem with this method is that it depends upon the range of sensitive attribute. If want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. Also, L-diversity method is subject to skewness and similarity attack and thus can’t prevent attribute disclosure.

c) T-closeness

It is a further improvement of an l-diversity group based anonymization that is used to preserve privacy in data sets by decreasing the granularity of data representation. This reduction is a trade-off that result in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model extends the l-diversity model by treading the values of an attribute distinctly by taking into account the distribution of data values for that attribute.
An equivalence class is said to have $t$-closeness if the distance between the conveyance of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold $t$. A table is said to have $t$-closeness if all equivalence classes have $t$-closeness. The main advantage to $t$-closeness is that as size and variety of data increases, the odds of re-identification too increase. The brute-force approach that examines each possible partition of the table to find the optimal solution takes $n^{o(n)} m^{o(1)} \approx 2^{o(n \log n)} m^{o(1)}$ time.

**B. Differential Privacy**

Differential Privacy could be technique enable analysts to remove helpful answer as of database containing individual information while donation solid individual privacy protection. The aim of this method is to limit the chances of individual distinguishing proof while querying data. The phases of differential privacy are personating in the figure.

![Fig. 2 Differential privacy](image)

**C. Notice and Consent**

The foremost common privacy preservation method for web services is notice and consent. Whenever individual access a new application or services, a notice stating privacy issues is displayed. The end user needs to consent the notice before exploitation service. This technique empowers a private to make sure his privacy rights. It puts the burden of privacy preservation on the individual. Once applied to big data, this technique poses various challenges.

**4. CONCLUSION**

This paper focused on technology for big data confidentiality. Big data analytics should incorporate confidentiality preservation technologies core elements. Existing Privacy preserving techniques are able to address any one V of 3Vs. All outlined methods have restricted potential when applied to big data. In future, we focus on rectifying the challenges of big data with changes of confidentiality technology for the profit of all involved organizations.

**5. ACKNOWLEDGMENT**

I would like to offer my special thanks to Mr. G B Jethava. The advice given by him has been a great help in my dissertation work. Many people have helped, provided direction as well as technical information and I take pleasure in acknowledging the role of people involved, directly or indirectly in the development of this dissertation work. I also express my gratitude to Mrs. Sumitra Menaria Head of the Department, Division of Information Technology for providing me with adequate facilities, ways, and means by which I was able to complete this dissertation work. I express my immense pleasure and thankfulness to all the teachers and staff of the Division of Information Technology, PIET for their cooperation and support. Last but not the least I thank all others and especially my classmates and my family members who in one way or another helped me in the successful completion of this work.

**6. REFERENCES**


